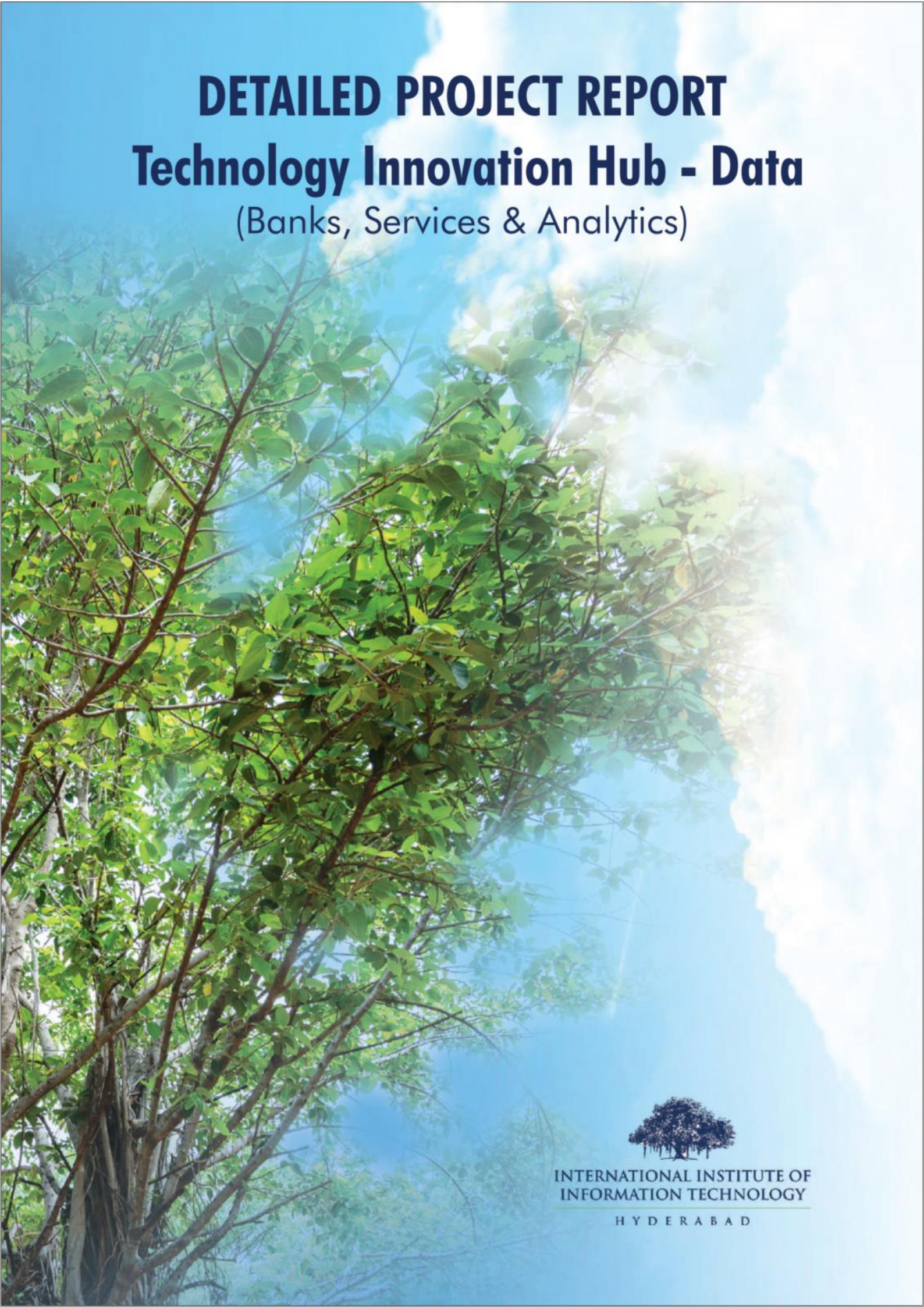
DETAILED PROJECT REPORT

Technology Innovation Hub - Data

(Banks, Services & Analytics)



INTERNATIONAL INSTITUTE OF
INFORMATION TECHNOLOGY
HYDERABAD

Contents

Executive Summary

This Detailed Project Report is to establish a Technology Innovation Hub at IIIT Hyderabad in the areas of Data Banks, Data Services, and Data Analytics, herein referred to as the **TIH-Data,** under the National Mission on Interdisciplinary Cyber Physical Systems executed by the Department of Science and Technology (DST). The Hub will help coordinate, integrate, and amplify basic and applied research in broad Data-Driven Technologies as well as its dissemination and translation across the country. The world has seen explosive growth in computers, mobile phones, communication, information dissemination, etc. in the past two decades, driven largely by the advances in hardware given by Moore's law. The next revolution will be spurred by the advances in "intelligence" that can be applied to the vast amounts of data and information that is available today. Artificial Intelligence (AI) has given reliable and stable algorithms and applications using Machine Learning (ML) in different hues. A critical ingredient of modern ML-based AI is the abundance of reliable and representative data on specific problem areas. Data is today recognized as a critical enabler, being referred to as the "new oil[1]" and the "new electricity[2]" by experts based on its foundational value in growth of different disciplines.

IIIT Hyderabad has become a major player in the research landscape on India. The institute has 30 faculty members and over 200 research students in related areas. It has created several data sets including those for OCRs, Speech, and Language Translation. The IIIT 5K-word data set released in 2012, the IIIT STR data set released in 2013, the Indian Driving Dataset (IDD) released in 2018 are used extensively by global academics and have been downloaded thousands of times. The Indian Brain Atlas created by IIIT using the scans of 100 healthy individuals will become the benchmark for diagnosis of neural anomalies of the Indian population. The institute has also established a highly productive incubation and startup effort, which formed as the foundational model to establish T-Hub by the Government of Telangana. All of this was achieved while staying completely self-reliant in finances. The institute has created strong connections with all national funding agencies, the Government of Telangana, as well as the top companies like Intel, Google, Microsoft, TCS, Amazon, Flipkart, and others.

Data is at the core of the TIH Hub proposed to be established at IIIT Hyderabad. This Hub differs from other Hubs for its mandate to prepare a critical resource for the future use by researchers, startups, and industry. Collecting, collating, and distributing useful data from multiple domains for use by the national and global community will be a significant effort at TIH-Data. This requires not only the expertise of computer scientists and data scientists, but

[1] Attributed to British Mathematician Clive Humby
[2] Attributed to Stanford Professor Andrew Ng

also the cooperation of experts from multiple domains, such as Healthcare, Transportation, Smart Building, Systems, etc. The **Data Foundation** of the TIH-Data will coordinate the data collection, storage, distribution, processing, etc.

Research and Technology Development to exploit such data is another key activity of the proposed Hub. This will be carried out in collaboration with other academic research groups and domain experts. Broad directions for the activities will be laid out by the Hub based on which proposals will be called, evaluated, and selected for execution. Effective ways to monitor and guide such projects will also be set up as part of the Hub.

Reaching out to the broader society to enhance societal expertise in data-driven technologies will be another important activity of the Hub. This would be done through support to undergraduate and post-graduate students and attractive fellowships to PhD students. TIH-Data will also boost the research ecosystem through a number of competitive post-doctoral fellowships and faculty fellowships established across the country. Workshops and short courses for industry professionals, startup employees, and others are planned to spread the expertise in data-driven technologies widely in the country.

Data science and associated areas are highly active internationally. The Hub will have an intense emphasis on collaborating with international experts in associated areas. This will include joint research, visiting schemes for international experts, involvement in the processes and problems of the Hub, etc. A series of workshops, conferences, and meetings are also planned to bring global experts in direct contact with the Indian researchers and students in associated areas.

Start-ups provide the most promising and speedy conduit for advanced technologies to be taken directly to potential users. Start-ups rely on innovations not only in the underlying science and technology, but also in their adaptation to novel application areas and users. Promoting innovations and startups in the area of Data-Driven Technologies and broader AI will be an important activity of the Hub. It will have access to the Centre for Innovation and Entrepreneurship of IIIT Hyderabad, one of the largest academic incubators in the country.

Research should leave the lab and reach the land. This is a challenge in Indian context as the academia is ill-prepared to do the translation. India has significant research capacity in most areas, but only a few success stories exist of the research leading directly to technologies that can be taken forward by startups or can be transferred to industry. The primary endeavour of academia is to develop basic research, not to apply it on real-world scenarios. The startups are often stymied by the gap between academic output and what they need. Institutions need to create explicit channels to translate research output into prototypes and field-ready products. TIH-Data proposes to establish an Applied Research and Translation team of technical experts, domain experts, and product developers, coordinated by a product management team. This team will help translate the research output to the needs of the industry and startups. IIIT-H has very positive experiences from a Product Lab it established in 2017. The Hub will build on and scale up this experience for this purpose.

Societal applications of data-driven technologies is of special interest to the proposed TIH-Data. It is proposed to establish a group focussed on technology for societal applications within the Applied Research wing. This requires close collaboration of the technologists with experts in Healthcare, Transportation, and other domains. It also needs the participation of user agencies such as Government Departments, NGOs, etc., to keep the systems practical and useful. The results are expected to reach the society directly through relevant agencies. IIIT is already in close to finalizing a few projects in transportation safety and public health along with the Government of Telangana, independently of the Hub.

Structurally, the proposed TIH-Data will be a Section 8 company that manages the Data Foundation, Research & Technology Development, Applied Research & Translation, HRD & Skill Development, Innovation & Start-ups, and the International Collaborations directly. A Hub Governing Body will be in charge of policy setting, execution, and monitoring of the Hub. Activities will be envisioned and detailed through the Technical Advisory Council and executed by the Executive Committee. A Hub Academic Head will oversee all aspects and a CEO/COO assisted by a suitable management team will help carry out the activities.

TIH-Data will be set on a high-speed path by the initial funding from NM-ICPS. The activities need to continue and flourish for a long time as they are of fundamental value and import to the country. Generation of resources from other governmental and private sources is necessary to ensure this. Several promising avenues will be explored for this including sponsored research from governmental & private sources, technology transfer & licensing, short courses & executive education, etc. IIIT Hyderabad has great experience in not only being self-sufficient for its activities, but also in growing steadily in education, research, infrastructure expansion, etc. The institute got land and some initial buildings from the state government in 1998. All operational and expansion expenses were raised by the institute through tuition fees, competitive research grants, executive education, CSR donations, etc. This experience puts the TIH-Data at a comfortable pedestal to grow into a world-recognized entity in various data resources and data-driven technologies.

The Grand vision for the Hub is to become the pre-eminent reference for datasets for AI researchers world-wide and also develop solutions using these datasets to address population scale challenges resulting in wide societal impact.

1 TIH-Data: The Context and Background

Introduction

IIIT Hyderabad proposes to establish a Technology Innovation Hub for Data Banks, Data Services and Data Analytics (TIH-Data), to coordinate, integrate and amplify the research in the larger areas of Data Banks, Data Services and Data Analytics under the National Mission on Interdisciplinary Cyber-Physical Systems (NM-ICPS) scheme. The world has seen explosive growth in computers, mobile phones, communication, information dissemination, etc. in the past two decades, mainly driven by the advances in hardware given by Moore's law. Data has become the new oil of the world. It is now fuelling the technology revolution. The ongoing technology revolution is spurred by the advances in "intelligence" that can be applied to the vast amounts of data and information that is available today. The hoary past of Artificial Intelligence has given way to reliable and stable algorithms and applications involving machine learning and analytics. A concerted effort to effectively exploit the vast seas of data in textual, audio, voice, image, video, and other modalities will bring unprecedented advances to humanity in partnership with the digital and communication technologies. All these raw sensor outputs are now labelled as data from which algorithms learn to solve problems. The fundamental developments in Data (Banks, Services and Analytics) can be extended to real-life, impactful technologies and solutions in the socially-relevant domains such as Healthcare, Mobility, Sustainable Buildings, Agriculture, disaster management, and so on. Also, top-class research will result in top-class education, with the graduates and researchers building solutions on their own for society. This also leads to a battery of technology trained engineers who contribute to the technology transformation.

Intellectual Focus

The Hub at IIIT Hyderabad proposes to take some significant steps towards realising these promises in partnership with others. There is a tremendous scope for building large teams spanning multiple institutions in the country to realise the benefits of studying and researching data. The institute has significant research, education, product development, and outreach activities in related areas including Data Science, Machine Learning, Natural Language Processing, Speech Processing, Computer Vision, Robotics, Speech Processing, and Cognitive Sciences.

The institute also has several collaborative projects with national institutions such as:

- IISc, IITs in Bombay, Delhi, Kharagpur, Kanpur, Dharwad, and Hyderabad.
- IIITs in Delhi, and Sri City.
- University of Oxford, INRIA, University of Utah, University of Pennsylvania, Carnegie Mellon University.
- DRDO laboratories such as CAIR, RCI, and ARDE.

- Industrial research teams in Intel, Bosch, TCS, MathWorks, Microsoft, Nissan, and several others.

The Hub will help coordinate and enhance national research and solution development efforts in the Data (Banks, Services and Analytics) areas and take them to the highest global academic standards. Applications towards the betterment of the society around us will have a special focus at the Hub, along with the translation of technologies into viable products and promoting startups based on those technologies.

The Hub will be an entity that both strengthens existing groups working in Data (Banks, Services and Analytics) across the country and facilitates new activities in all related areas. In the long run, the Hub can also play as a force multiplier in attracting projects and funding from other entities in the government or industry sectors.

The Hub aims to combine:

- Technology development activities in related areas across different institutions in the country.
- Applications to socially-relevant domain areas in collaboration with NGOs, domain-experts, Government of Telangana, and other entities in the country.
- Research translation and product development with significant Data (Banks, Services and Analytics) components using dedicated development and management teams.
- Facilitate startups that use broad Data technologies with technical support, customer connection, VC connections, and so on.
- Be the focus point of technical conferences of all sizes, workshops, summer schools, and so on.
- Be a conduit for fruitful international collaborations in all directions such as visits and sabbaticals.
- Disseminate the potential and message of Data to the broader society consisting of industry, government, doctors, journalists, bureaucrats, college students, school children, and so on.

Proposed Activities

The activities will broadly include, but will not be limited to, the following:

- High impact research and technology development in Data (Banks, Services and Analytics) encompassing core Machine Learning, Natural Language Processing, Speech Processing, Computer Vision, Robotics, Cognitive Sciences, etc. Research projects will be selected based on proposals invited from research teams far and wide, following sound and fair evaluation processes.
- Solution development using Data-driven technologies in socially impactful areas will be a high priority activity of the Hub. We propose the domains of Mobility/Transportation, Healthcare, Smart Buildings, India specific problems etc., as the initial phase of activities. These efforts will be carried out in collaboration with domain-agencies and experts. IIIT-H already has a special relationship with the

Government of Telangana, the Public Health Foundation of India, and Intel to work on Healthcare area. Strengthening this activity and building more such campaigns in collaboration with other academic institutions and domain institutions will be done by the Hub.

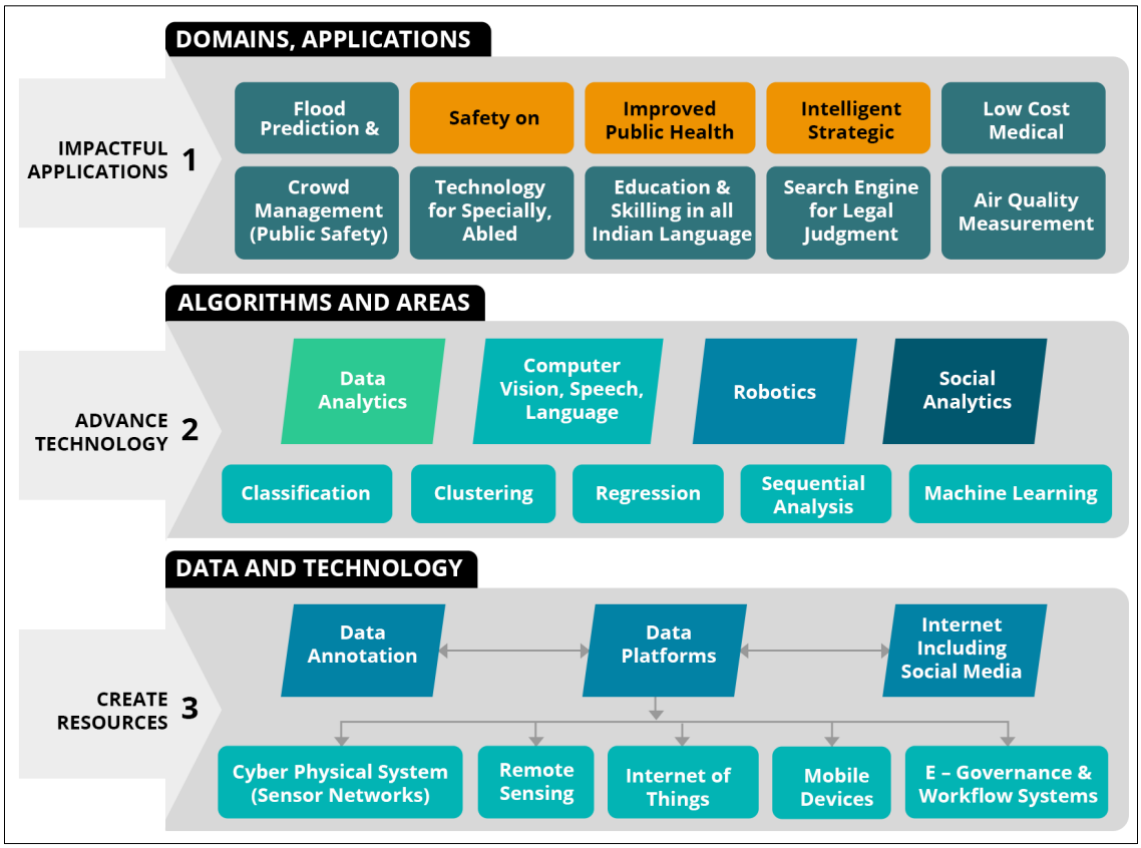
- Research should leave the lab and reach the land. This is a challenge in the Indian context as the academia is ill-prepared to do the required. A specialised team of technical developers and product managers will be set up to translate the research output to the needs of the industry and startups. IIIT-H has very positive experiences from a Product Lab it established in 2017 for this purpose. This experience will be scaled to include the whole of the Hub's activities.
- Startup and entrepreneurship promotion will be worked on by the Hub. IIIT Hyderabad already has the largest academic incubation facility in the country with over 100 companies. Startup companies are the best means to transform the research in the labs to the benefit of the society.
- High-quality teaching programs including those for post-graduate students, at IIIT-H and other institutions, executives in Hyderabad and elsewhere, as well as UG students in the engineering colleges around the country. Establishing national Doctoral and Post-Doctoral Fellowships will be a special focus.
- Data-driven technology is an area that has caught the imagination of the industry, government, and public at large. The Hub should continue to have technical events, targeted workshops, popular sessions, and so on, aimed at specific sections of the society to enable them to understand and exploit the fruits of Data and exploit the utilities of data.
- The focus of the Hub majorly includes applications with significant societal impact. Teaming up with Central and State governments is essential for achieving this. IIIT-H has a strong partnership with the Government of Telangana on several fronts, including Health and Mobility. We intend to leverage on that as well as expand it to other governments so that the fruits of the technologies developed by the Hub will be useful to the larger society.
- AI, Learning and Data-driven technologies have been heavily researched areas recently. The Hub will strive to integrate different entities with significant stakes in the core as well as application areas so that useful and impactful solutions and products emerge from the effort.

This detailed project report explains the plans, strategies, technical details, expected outcomes.

1.1 Role of Data and Data-Driven Technologies

Data has taken the central stage now. Data is impacting the technology development process beyond the human skills and expertise. This demands special attention on data, and data-driven solutions in the context of cyber physical systems. The role and impact of the data is only to increase from here.

Figure 1: A Three-Layered Architecture of AI and Role of Data in India-Specific Sectors



Modern AI could be explained with a three-layer architecture as shown in the figure, Three-Layered Architecture of AI and Role of Data in India-Specific Sectors. In the Indian setting, AI makes impact only when it can meet the basic use cases of common man. This needs special effort in (i) training people and developing research culture around the proper use of data (ii) defining problems that are local and fundamental and indeed come up with technologies that solve the problem(iii) do sufficient engineering and market strategies to make the solution accessible and affordable to the common man.

We will start by explaining the 3-layered architecture.

Layer 1: Data Layer

The foundation layer of modern AI systems is data. Data could come from any different sources. It could be captured with sensors; it could be captured from transactions or operational logs of people associated with such systems. It could be created by machines; it could be created by humans. In general, most of these data are unstructured, multimodal and multidimensional. If you look at the Indian setting, a number of infrastructure projects and missions, from the past, led to creation of huge data. Examples include remote sensing, e-Governance etc. A number of sensors that we carry (e.g., in mobile phones) or deploy (say, in IoT devices) also generate huge quantity of data.

Layer 2: Algorithms

The second layer contains algorithms that process the data. Machine Learning and its variations (unsupervised, supervised, reinforcement etc.) come in this layer. There are also a lot of related areas like computer vision, natural language processing that clearly fits into this space. They all process the data and help in solving problems. This layer focuses on how new problems (e.g., captioning) could be solved or how a specific problem (e.g., classification) could be solved in a better way.

Layer 3: Domains and Applications

However, in Indian setting, impact of the AI is in creating solutions that influence the domains that touch the lives of common man. A number of these areas have associated data, but often not captured, curated or even understood. These domains indeed have burning problems. But the AI and smart technologies are yet to penetrate. We consider this space also opens up a number of technology and business opportunities for researchers and entrepreneurs.

1.1.1 Data and Use Cases

Data has become an integral part of many technology solutions. It is not just some information that human consumes. Over the last decades, data has helped in many different ways. Here we summarise some of the important ones

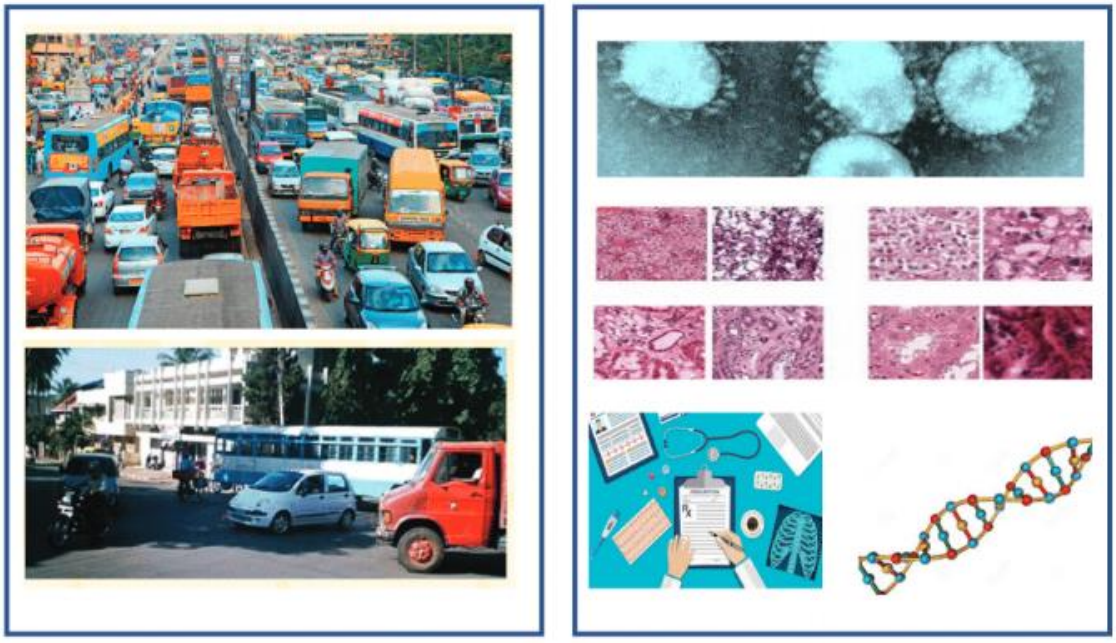
- **Data for Problem Discovery**: A number of problems that needs attention are getting discovered by careful analysis of the data. It could be Frequent or rare patterns of interest. Or it could be the cause and effect relationships. Info-graphics and data visualization. are now common at the level of non-technical people in discovering and presenting new problems.
- **Data for Problem Definition**: Data helps in defining the problem properly. A number of perception and reasoning tasks are now getting defined in terms of data/examples. A human understandable definition in natural languages is not feasible in many of these situations.
- **Data for Driving Research**: Data also has been in the forefront of driving research in new and emerging areas. Many grand challenges and global tasks have been defined in terms of the real-life data. This attracts people who, otherwise do not work in these areas. This also push the performance envelops.
- **Data for Sensing the Progress:** Well curated data, examples and dynamic leader boards have become the integral part of sensing the progress of the community in solving fundamental tasks. This is especially true for fundamental and ill posed tasks like perception.
- **Data for Creating Solutions**: Needless to say, machine learning and data-driven solutions extensively use data for developing solutions, for example input-output mapping in many complex problem situations. Most deep learning solutions of today follow this approach.

- **Data for Validating and Testing the Technology**: Data has also taken a new role of providing statistical evidence of successful testing. Testing on more data implies more reliability of the results that are reported. Traditional testing and verification protocols in engineering and design are getting adapted to the new data age.
- **Data for Creating Services, APIs and Solutions**: Data has now the basis of providing a number of solutions that are provided as services. Cloud services have systematically emerged into a combination of storage, compute and APIs. Data controlling the backstage has enabled this.
- **Data for Personalisation and Adaptation**: Since the applications and services are built over data, it has become relatively simple to provide solutions that are personalized, even if the number of users are large. This is possible, since the solutions can be adapted or fine tunes to a specific user of a class of users with ease. Popular learning algorithms are also incremental, making this feasible.

1.1.2 Two Examples

We now look at two specific example situations where the data-driven solutions will have important role. These are picked as exemplary situations to demonstrate how the domain needs to be aligned with the data-driven technologies.

Figure 2: Indian Roads and Driving Patterns & Indian Genes and Health Parameters are Very Likely to be Different from those in the West



Indian Driving Situations

India witness a large number of accidents on our roads. Many of them are also really fatal. How can technology help in improving the road safety using technology? A natural step is to absorb the solutions that are in the west to India. This is not going to be practical or

effective. Our road and traffic infrastructure are very different. Our vehicles and driving patterns are very different. Our society and culture is different from other. This demands:

- Our own data that capture the driving situations, surroundings.
- Our own data that capture driving patterns, driver behaviours and the characteristics of other trac participants.

What can we do with such data? This data can help in developing solutions and validating them for Governmental agencies (e.g., Trac departments), Vehicle (e.g., driver assistance systems) and Planners. Such a data can help in (examples):

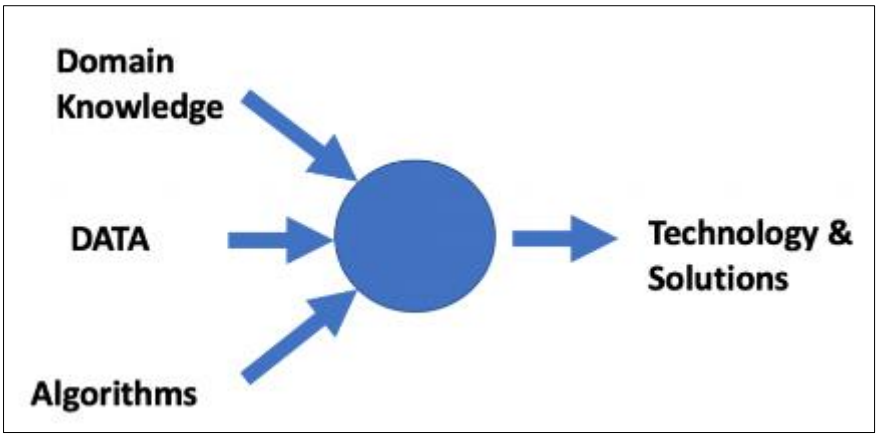
- Creating solutions that improve safety.
- Suggestions from simulations, for planning infrastructure.
- For energy efficiency, less pollution and better environment.

Medical Diagnosis in Indian Setting

Another domain is medical and healthcare. India needs many more doctors and healthcare professionals to meet its needs. Scarcity is more severe in rural India. A number of AI and technology driven solutions are emerging.

- Indian data will be important for designing solutions for the Indian population due to the genetic and environmental variations that we will see in our healthcare system.
- Solutions that are acceptable to the community, whether it is socially, culturally or economically, needs Indian flavour. And data is going to be a big player for validating this.

Figure 3: How does Data Make Deeper Impact?



1.1.3 Data + Domain + Algorithms=>Technology, Services and Solutions

We now briefly discuss how data along with other players like domains and algorithms can lead to a number of technology solutions. Also, how such solutions can lead to services and deployable solutions.

Raw Data + Algorithms=> Knowledge Discovery: In many situations data is “raw” and a class of algorithms (especially unsupervised learning based) lead to discovery of knowledge, important events and patterns.

Raw Data + Domain Knowledge=> Smart Useful Data: Though Big Data is welcome, it is also of interest to use the domain knowledge for choosing the right data and dimensions for creating smart and useful data. This is practically important since we work under a fixed budget of resources such as storage and compute.

Raw Data + Human Knowledge=> Annotated Data: Raw data is not enough for the most popular class of learning algorithms, supervised learning. Domain experts need to add value in terms of tags, labels and knowledge so that the data becomes annotated and well-curated data.

Annotated Data + Learning Algorithms=>Models: Parametric computational models are obtained with the help of learning algorithms from the annotated data. These models have captured the knowledge from the data.

Models + New Situations=> Results to new problems: Computational models of the data-driven solutions, when deployed in new situations, generalize to new results. Models are now usable in practice.

Models + Engineering=>Deployable Solutions/APIs: Finally, additional engineering efforts on these computational models lead to robust implementations, solutions that run on dedicated or constrained hardware or software platforms.

1.2 About IIIT-H

International Institute of Information Technology Hyderabad (IIIT-H) is an autonomous university, founded as a not-for-profit public private partnership (N-PPP) in 1998, and is the first IIIT in India under this model. It was established with the vision of becoming a research-led university that focuses on solutions for the larger society. Over the past 22 years, the institute has developed into a leading entity in the fields of Artificial Intelligence, Machine Learning, Robotics and data-driven technologies. A robust 40-member faculty team comprises of active researchers in leading technologies and their applications.

Figure 4: IIIT Hyderabad

Over the years, the institute has evolved strong research programmes in various areas, with an emphasis on technology and demonstrated applications in technology for industry and society. The institute facilitates interdisciplinary research and a seamless flow of knowledge. Several world-renowned centres of excellence are part of IIIT-H's research portfolio. It has established various joint-collaboration and co-innovation models with an industry outreach spanning significant national and multinational companies. Its innovative curriculum allows students the flexibility of selecting their courses and projects. Apart from academics, the institute provides students with a comprehensive environment that promotes art and culture, sports, societal contributions and self-governance. Even undergraduate students get to participate in ongoing research and technology development - an opportunity unprecedented in India. As a result, a vibrant undergraduate programme co-exists along with a strong postgraduate programme.

1.2.1 Experience in the Proposed Area

Ranking of Research

CSRankings, a metrics-based ranking of top computer science institutions around the world has consistently ranked IIIT Hyderabad India's best institute in the fields of AI and Robotics research. The Asia-wide rankings are available here.

IIIT-H also ranks high in the fields of Computer Vision and Robotics across the globe for its research and publications at top conferences and reputed journals.

Kohli Center on Intelligent Systems (KCIS)

A Unique Cluster Model: IIIT Hyderabad has already created an umbrella organisation, called Kohli Center on Intelligent Systems (KCIS) with the support of TCS foundation to bring together researchers in the larger areas of AI and Intelligent Systems. It is a cluster of research labs that work together to enable interdisciplinary collaborations between faculty

and other researchers in the field of AI, leading to solutions for practical and socially relevant problems.

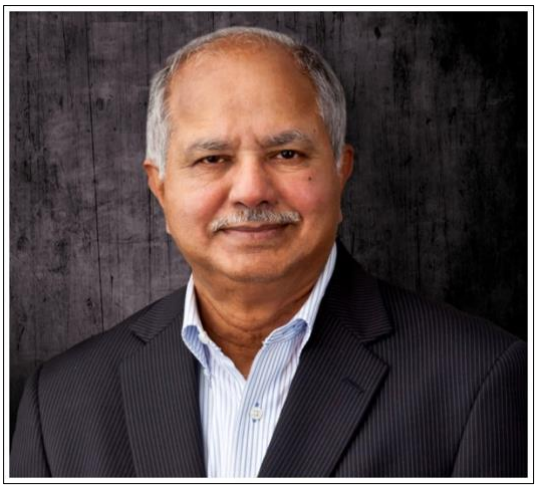
Figure 5: Faculty at IIIT Hyderabad

Governing Council and Advisors

IIIT Hyderabad's governing board consists exclusively of eminent academicians and industrialists along with representatives from the government.

Prof. Raj Reddy, chairman of the governing council since its inception in 1998, as well as GC advisor Dr Vincent Cerf, are both recipients of the Turing Award, considered to be the Nobel Prize in computing.

Figure 6: Raj Reddy, Winner of the Turing Award, Chairman, Governing Council, IIIT-H



The current governing council includes, among others, Debjani Ghosh, President of NASSCOM as well as Dr Jayesh Ranjan, Principal Secretary of IT, Industries and Commerce, Govt of Telangana.

Team: The Largest AI Research Group

IIIT Hyderabad has a dedicated team of over 30 faculty and 200 students actively involved in research, technology, development and publications, as well as the largest faculty and student group working in the area of AI.

Figure 7: Launch of Executive programme in AI/ML in Bengaluru

Domain Focus: Mobility, Robotics, Healthcare, Sciences

In addition to the core areas of AI, Machine Learning and Data Science, IIIT-H has substantial competencies in applied areas such as Robotics, Mobility, Healthcare and Computational Sciences. The research centres on Robotics and Computational Natural Sciences are two of the strongest groups in the institute in terms of research output.

Figure 8: Robotics Research Center (RRC)

Figure 9: RRC Showcasing Research Projects at R&D Showcase 2019 Hosted by IIIT-H

Following are the descriptions of some of the domain research centers that are working in the areas of the interface of computer science with other areas.

Center for Computational Natural Sciences and Bioinformatics

The Center for Computational Natural Sciences and Bioinformatics is involved in world class research in the areas interfacing computing and the natural sciences, physics, chemistry and biology. The center applies advances in computer science (artificial intelligence, machine learning, high performance/distributing computing, graph theory, etc.) to fundamental problems in sciences. The faculty members of the center come with diverse backgrounds such as high energy physics, systems biology, chemical physics, physical chemistry, biophysics, bioinformatics, etc, which enables research in the inter-disciplinary areas. The center has been in the forefront in this area in the country.

Research Areas

- AI/ML for Natural Sciences.
- Data systems and applications in Metabolomics.
- RNA structural bioinformatics.
- Protein dynamics and folding.
- Genomics and proteomics.
- Mathematical modelling of biological networks.
- Algorithms for image diagnostics.
- Computer aided drug design.

Figure 10: Workshop on Machine Learning for Science

Cognitive Science Lab

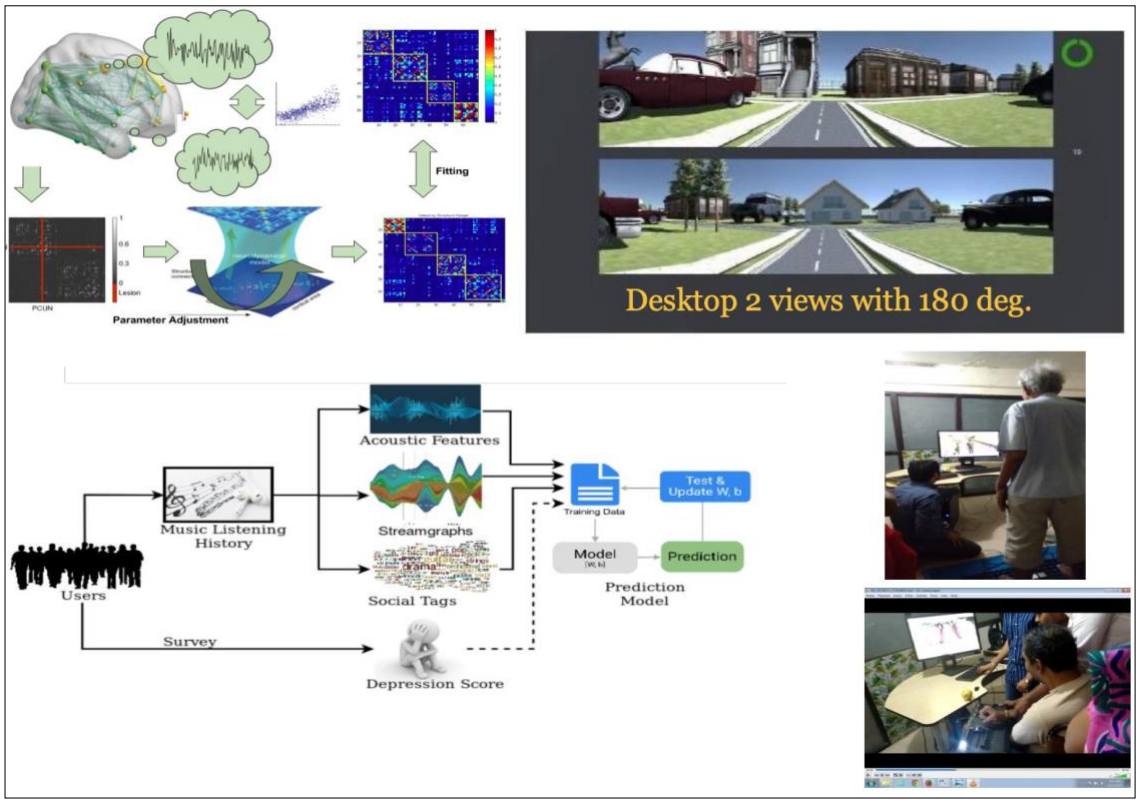
Cognitive Science Laboratory (CSL) is engaged in interdisciplinary endeavour to understand the brain, mind and behaviour using computational models, developing intervention aids and studying mental well-being.

Research Areas

- Neuroimaging: dementia, depression, stroke, autism.
- Virtual Reality (VR-VE) human centered design: 360-degree vision, adaptability.
- Music cognition.

- Perception and Creativity: perception, creative thinking, intelligence.
- Assistive Systems, Games/Simulations: Neuro-habilitation, stroke/brain plasticity.
- Health and Well-being: Autism, depression.

Figure 11: Cognitive Science Lab



Center for IT in Building Science

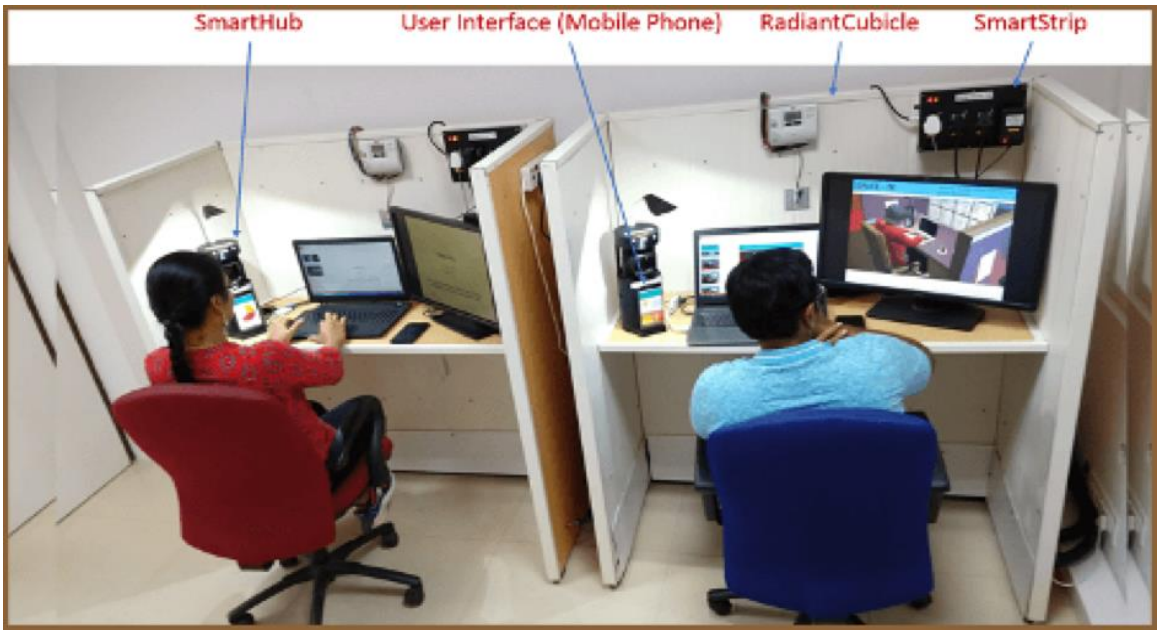
The Center for IT in Building Science (CBS) was established in the year 2000 with the aim of carrying out research for the applications of IT to improve the performance of buildings in the aspects of health, safety and security. With a focus on energy, the major research of CBS revolves around building simulation, building energy monitoring and automation and controls.

Research Areas

- Building energy modelling and simulation.
- Smart Home Energy Management System.
- Life cycle energy optimisation.
- Early design tools for buildings.
- Building fault detection and diagnostics.
- Cool roof monitoring.
- Weathering and ageing of cool roof.
- Task controls and communications.

- Dynamic windows.
- Non-intrusive load monitoring.

Figure 12: Data Collection at the Center for IT in Building Science



Earthquake Engineering Research Center

EERC contributes towards construction of safe building infrastructure and addresses the large and rapidly growing urban seismic risk in India. The center provides technical solutions for designing seismically safe new structures and retrofitting of seismically unsafe existing structures. The vision of the center is to become a National Center of Excellence in research on earthquake engineering and to provide solutions for disaster management in general.

Research Areas

- Seismotectonic modelling of Indian plate.
- Fault motion analysis & wave propagation.
- Nonlinear analysis of structures.
- Seismic evaluation of heritage structures.
- Sustainable construction technologies.
- Earthquake awareness and preparedness.
- Post-earthquake studies.
- Earthquake safety assessment using AI.

Figure 13: Researchers of EERC

Lab for Spatial Informatics

Lab for Spatial Informatics (LSI) at IIIT Hyderabad has been at the forefront of Geospatial science research and applications across multiple and innovative domains. The lab provides a unique platform for research at postgraduate level, with students/researchers working on problems with a multi-disciplinary approach. LSI also works extensively with other National and International organizations like Survey of India, ISRO-NRSC, IIRS, OSGeo, DRDO Labs, etc.

Research Areas

- Remote sensing.
- Geospatial data sciences.
- Modelling and simulation of land, water and environmental changes.
- GIS techniques for agriculture, forestry, urban studies.
- Algorithm and system development for Remote sensing and GIS.

Figure 14: Researchers from LSI at the IIIT-H R&D Showcase

Industry Connect

In keeping with IIIT-H's vision of working on socially relevant and practical research problems, the institute undertakes various comprehensive collaborations with the industry in various forms including industry researchers teaching courses within the institute, working on research problems with the industry and fostering a vibrant startup ecosystem within the institute to take the research from labs to land.

Figure 15: Exchange of MoUs

Figure 16: Exchange of MoUs

Government/PSU Collaborations

As IIIT Hyderabad was started by the state government, it maintains a close relationship with the state government and other PSUs such as DRDO, ISRO, DRL, BHEL, etc. The institute continuously engages with government research agencies such as DST, DBT and MEITY and is actively involved in several government initiatives. IIIT-H has also conducted several workshops for government officials in the fields of AI/ML, Disaster Management and e-governance.

Mutual Working Relationship with State Government

IIIT Hyderabad was started by the then state government of Andhra Pradesh and continues to enjoy strong support from the Telangana government. However, as the institute does not receive any financial support from the state or central government, it functions as an independent organisation under the governing council.

Figure 17: IT Secretary, Government of Telangana, at the launch of BioNest in IIIT-H

International Collaborations

IIIT Hyderabad works closely with several research groups internationally, such as CMU, Oxford, UPenn, UMass, INRIA, and UCSD. IIIT-H has also been involved in several collaborative research initiatives, including UKIERI and CEFIPRA/Indo-French research collaborations.

Figure 18: Delegation of an International Academic Institution Visits IIIT-H

Hosting International Conferences

Hyderabad has hosted several large premier international conferences in AI/ML, with IIIT-H playing a critical role. These include:

- IJCAI 2007 (The top conference in AI).
- WWW 2011.
- ACCV 2006.
- Interspeech 2018
- ICSE 2014.
- Several others such as IJCNLP, IPDPS, FSTTCS, COMAD, ICON, ICDCN, ISEC, CSEET and IBPSA.

Figure 19: India HCI 2019 Hosted by IIIT-H

Expanding Outreach to Industry, Beyond Research

In addition to its research record, IIIT Hyderabad is also actively involved in training industry professionals and engaging with the industry in order to elevate the level of industry products and solutions. These include:

AI/ML Programme

A 14-weekend AI/ML programme for industry professionals to learn and apply AI/ML technologies in their projects and products. Till date, over 2000 industry professionals have been trained in this programme with intensive classroom lectures and labs.

Figure 20: Executive Programme in AI/ML for Industry Professionals Conducted by IIIT-H

Post Graduate Student Status (PGSSP) Programme

The PGSSP programme allows working industry professionals to enrol and attend any of the 100+ courses conducted by IIIT-H at any time, without having to secure admission to a degree programme. At any stage, more than 100 working professionals take courses in IIITH. This not only brings value to the IIIT-H classrooms, but also adds value to the industries around.

The institute also engages with the industry in many other forms such as industry workshops as well as including adjunct faculty from the industry.

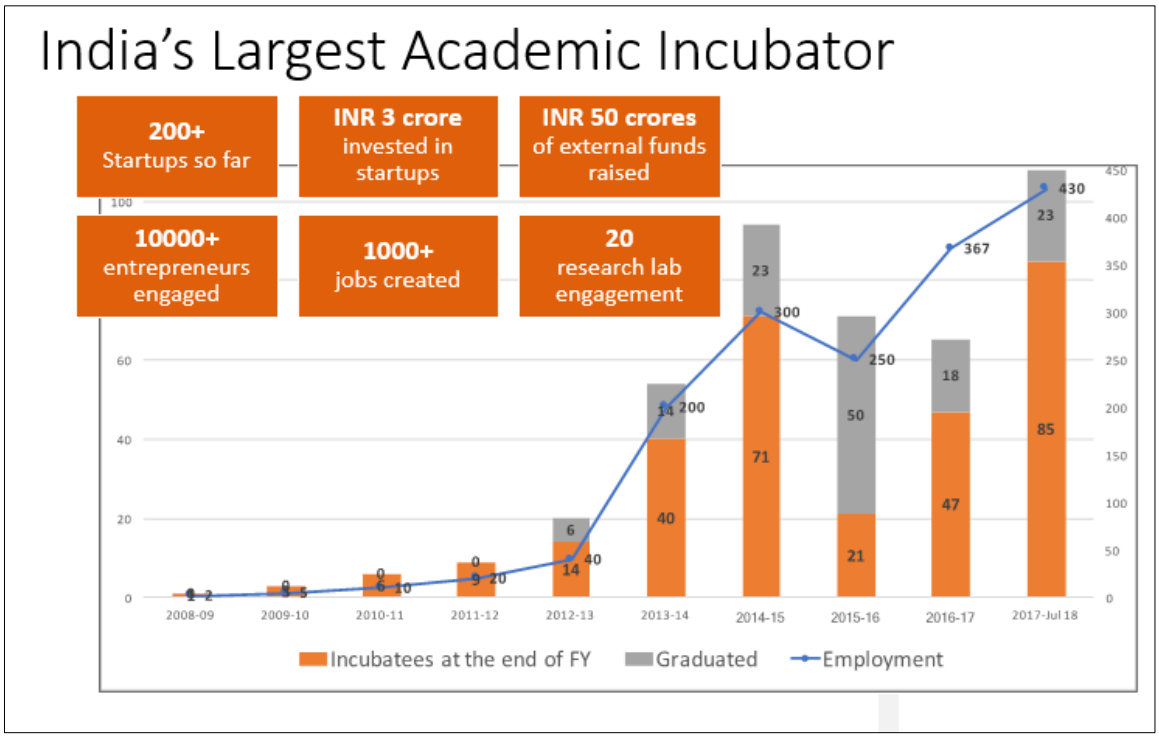
Innovation and Entrepreneurship

IIIT Hyderabad has a strong and vibrant startup incubator, which is possibly the largest academic incubation system in the country.

Center for Innovation and Entrepreneurship

The Center for Innovation and Entrepreneurship (CIE), which is a part of IIIT-H fosters several deep-tech startups and connects them with relevant faculty for mentorship in technology, business and legal aspects.

Figure 21: Center for Innovation and Entrepreneurship



T-Hub

T-Hub, a large incubation effort started by the Telangana government, is situated within IIIT-H campus. T-Hub interacts closely within the institute in terms of overall guidance as well as helping individual startups.

Figure 22: T-Hub

Other Programmes

We are also part of several incubation initiatives such as the BIRAC under the BioNEST scheme, AIC and NIDHI-PRAYAS to support innovation and entrepreneurship.

Productising Research

A special Makers Lab within the institute assists in productising research and reaching out to end users. There is also the Entrepreneur in Residence (EIR) program, where an enthusiastic person can take up the solutions developed in the research labs to convert to a product.

Figure 23: Makers Lab

Solutions Built and Deployed

IIIT Hyderabad has created several solutions that have been transferred to the industry and created a positive societal impact. These include medical solutions such as diabetic retinopathy detection, transportation solutions such as road and city audit mechanisms, crowd monitoring systems for large gatherings such as Kumbh Mela. Also, one of the most significant efforts in improving agriculture through technology was conceived and implemented under the eSagu project and was shown to enhance the product returns for farmers significantly. The institute has also developed several Indian Language Technology solutions including OCR, Machine Translation and Text-to-Speech systems.

Centres of Excellence

Centre of Excellence in Signal Processing: Established in 2012 under the MHRD FAST scheme to conduct fundamental and applied research in the field of signal processing.

Kohli Centre of Excellence on Intelligent Systems (from TCS Foundation): Established to foster collaborative research and take up socially relevant and practical problems in the field of AI and ML.

Ripple Block Chain Centre of Excellence (Ripple): Established to conduct fundamental and applied research in the areas of Blockchain and impart training to industry professionals.

BDL Centre of Excellence (from Bharat Dynamics Limited): Established with the support of Bharat Dynamics Limited, the Centre of Excellence (CoE) takes up projects relating to AI activities in missiles, manufacturing, inspection and allied areas.

CIE-IIITH PRIF Social Impact Incubator (from Pernod Ricard India Foundation): The centre established to address social and environmental problems through the application of AI technologies.

Centre of Excellence (CoE) on IoT for Smart Cities: The centre was established with India-EU collaboration on ICT standardisation, Telecommunications Standards Development Society, India (TSDSI) and European Telecommunications Standards Institute (ETSI).

Academic Achievements

The institute, faculty and students have won several awards for their research and work over the past two decades some of the specific awards include:

Table 1: Academic Achievements of IIIT-H

Publications/Mentions/Fellowships Received	Details
Fellowships	Prof. PJ Narayanan (INAE-2016), Prof. CV Jawahar (INAE-2018, IAPR2016), Prof. Rajeev Sangal (INAE)
Best Paper Honourable Mentions	CVPR Best Paper Honourable Mention (2018)
Best Paper Awards	VLSID-2019, BMVC-2018, ICB-2018, FG-2018, DAS-2018, NCVPRIPG-2017, ACPR-2015, HiPC-2013, ICVGIP-2010, and DAS-2008
Best Paper Award Runner-Ups, Finalist	IROS 2018 and ICVGIP 2014
Industry Fellowships	Qualcomm Innovation Fellowships (2016, 2017)
PhD Fellowships	Our students have won several PhD fellowships, including multiple TCS Fellowships, Intel Fellowships, Google India Fellowship, and MSR India Fellowship.
Publications in the last five years in the areas of AI/ML and its applications **Should be in Data-related areas.**	Number of publications in journals: 80 Number of publications in conferences: 500 Number of Master's Theses: 400 Number of PhD Theses: 64 Major Journals: IEEE Transactions on PAMI, IP, IFS, Multimedia, Computational Imaging, Medical Imaging; IJCV, MVA; IEEE Signal Processing, etc. Major Conferences: CVPR, ICCV, ECCV, NeurIPS, IJCAI, ACL, AAAI, ICRA, IROS, and so on.

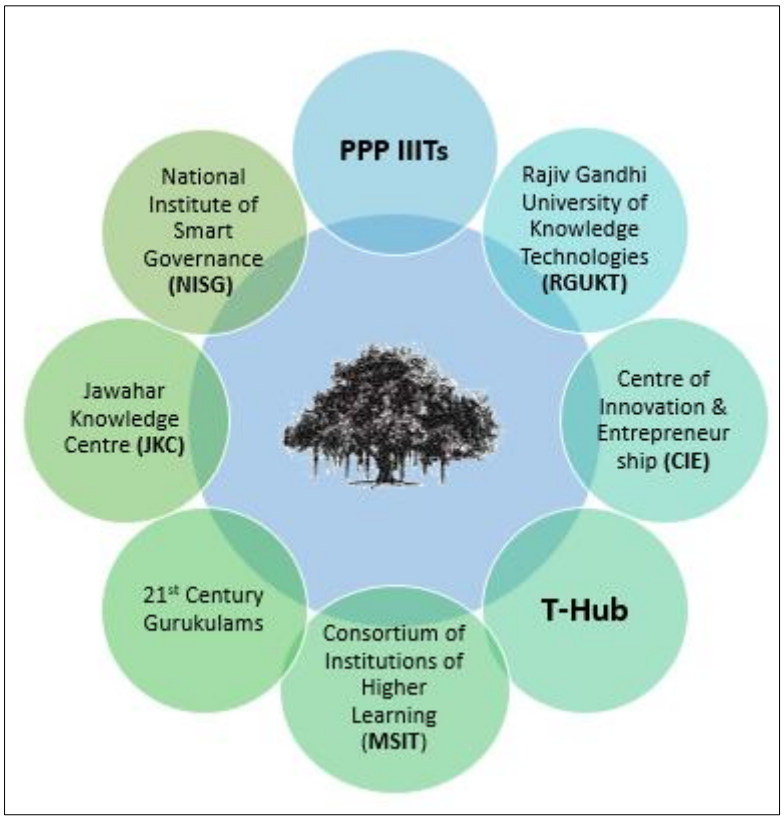
Extended Education Connect: In addition to regular courses and the executive education program, IIIT-H conducts several courses and workshops including the Student Technology Education Programme (STEP)for school students, faculty development programmes and summer- and inter-schools in Machine Learning, Computer Vision, NLP, Speech etc.

Incubating Institutions: IIIT Hyderabad played a critical role in incubating and nurturing many new generation institutions. These include:

ISB, RGUKT, NISG, and T-Hub: IIIT Hyderabad was involved in hosting, nurturing or advising them in several capacities, which in turn enabled them to create significant impact.

Twenty New IIITs: Encouraged by the success of IIIT Hyderabad, the government adopted our model to in creating 20 new IIITs across India. IIIT Hyderabad was hence selected to both create the model and mentor several of those institutions.

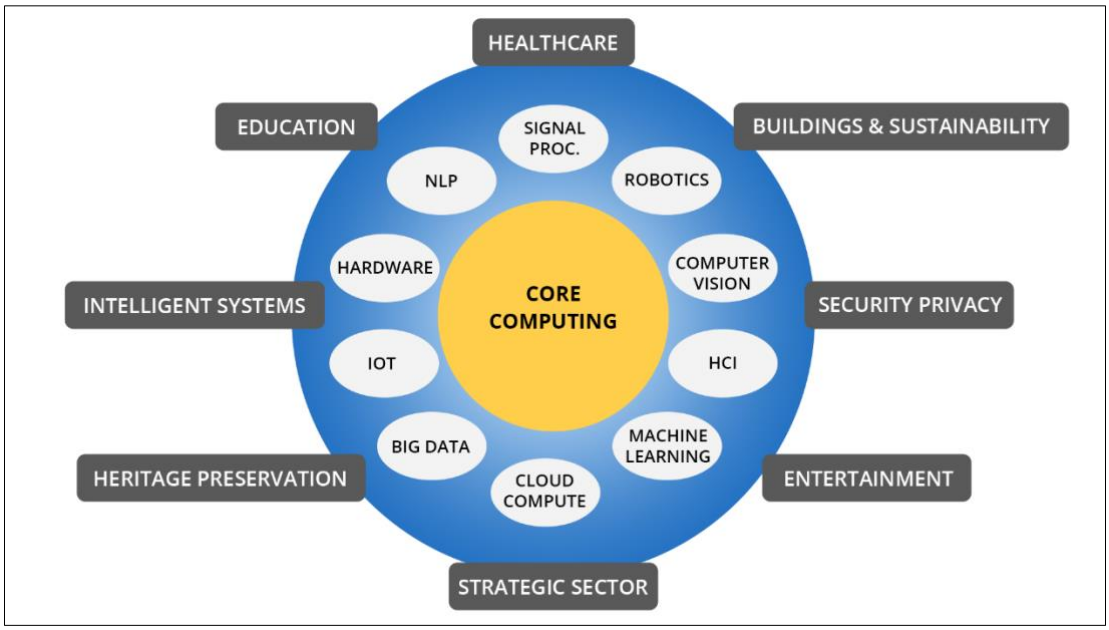
Figure 24:Incubators of Ideas and Institutions



Creation of Public Resources: IIIT Hyderabad has been actively involved in the creation of several resources for Research and Problem Solving. These include:

- Several data sets, corpora and resources in the area of language technology (NLP, Speech and OCR)
- India Driving Dataset (IDD) for road safety and research (also used by NITI for hackathon)
- Several other data sets, open-source codes, applications in the fields of AI/ML, Biometrics, Spatial Informatics, Building Science, Disaster Management and Education.

Figure 25: Computer Science Meeting New Domains



National Research Initiatives: IIIT Hyderabad has been an active and productive member of several national research initiatives, including:

- Consortium and national mission projects such as NIMT, OCR, ILMT, and Bahubhashak.
- Digital Heritage (DST), involving several technical and cultural organisations.
- Virtual Labs (MHRD) for creating teaching tools for technical education at scale.

National and Social Impact: Following the mandate to engage in socially relevant research, the institute has focused on research and developing solutions for the benefit of the society, specifically for India. These efforts include Indian language technology solutions (ILOCR, ILMT), solutions for Indian farmers (eSagu), Crowd Monitoring and Road Audit systems.

1.3 About Technology Innovation Hub – Data (TIH-Data)

IIIT Hyderabad proposes to establish a Technology Innovation Hub for Data Banks & Data Services and Data Analytics (TIH-Data), to coordinate, integrate and amplify the research in the larger areas of Data Banks, Data Services and Data Analytics under the National Mission on Interdisciplinary Cyber-Physical Systems (NM-ICPS) scheme.

1.3.1 Structure of TIH-Data

The structure of TIH-Data has been decided to create a foundation and a seamless ecosystem from education, human resource, R&D, technology development, innovation to commercialization. The structure will help TIH-Data achieve the overall mission of NM-ICPS by generating significant improvement in research, development and higher education in Science, Technology and Engineering disciplines. It will facilitate the development of home-

grown applications for national needs and capability to solve multi-disciplinary Grand Challenge problems and create a conducive environment for Scientific breakthroughs and Innovations.

The TIH-Data will have six major activities: (1) Technology Development, (2) Data Foundation, (3) Applied Research & Translation, (4) Human Resource & Skill Development, (5) Innovation, Entrepreneurship & Startup Ecosystem and (6) International Collaborations. The overall structure to achieve the core-activities is shown in figure, Organisational Structure of TIH-Data below.

Figure 26: Organisational Structure of TIH-Data

The broad approach that shall be used to implement TIH-Data will be:

- Cluster Based approach – The TIH-Data will have cluster(s) based approach for its Research & Technology Development. A cluster is defined first to limit the scope of the projects and activities. For example, some of the clusters are healthcare, mobility, etc. Every Cluster will be led by a faculty or researcher who will be coordinate the research and technology development in this space.
- Data Foundation & Applied Research & Translation – The TIH-Data will have two components for Data-Driven Technologies: (a) a Data Foundation that deals with the collection, curation, validation, management, and access to data-sets in several chosen domain areas and (b) an Applied Research & Translation Wing that performs translational research and development to take research from the academic groups to make it closer to technology that can be transferred to startups or other corporate entities.
- A collaborative network: The TIH-Data will have strong focus on bringing in various stakeholders, for example, Academic Institutions, Industry, startups, aspiring

entrepreneurs, students, researchers together through fellowship programmes, Grand challenges, student exchange, visiting faculty etc.

Some of the other key aspects of the structure of the TIH-Data are:

1. **Legal structure**: The TIH-Data shall be a section-8 company registered in Hyderabad and with an independent board of directors known as Hub Governing Body (HGB).
2. **Management structure**: The two key personnel of the TIH-Data will be Hub Academic Head and Chief Executive/Operating Officer who will be reporting to the HGB and shall be responsible for the overall activities.
3. **Governance**: The HGB will be the apex governing body to take decisions on all statutory matters. There will be Technical Advisory Council (TAC) to guide on everything related to technology and Review Team who shall be evaluating the progress of the all the Hub's activities and give recommendations to the HGB.

Activities & Programmes

The activities and programmes shall be designed and implemented keeping the impact and targets of the TIH-Data as the foundation. These will broadly include, but will not be limited to, the following:

a. High impact research in Data (Banks, Services and Analytics) encompassing Machine Learning, Natural Language Processing, Speech Processing, Computer Vision, Robotics, Cognitive Sciences, etc. Research projects will be selected based on proposals invited from research teams far and wide, following sound and fair evaluation processes.
b. Solution development using Data (Banks, Services and Analytics) technologies in socially impactful areas will be a high priority activity of the TIH-Data. We propose the domains of Mobility/Transportation, Healthcare, Defence/Strategic Sector, as the initial phase of activities. These efforts will be carried out in collaboration with domain-agencies and experts. Strengthening the existing IIIT-H's relationships and building more campaigns in collaboration with other academic institutions and domain institutions will be done by the TIH-Data.
c. Research should leave the lab and reach the land. A specialised team of technical developers and product managers will be set up to translate the research output to the needs of the industry and startups. Through licensing, technology transfer, grand challenges, etc., the applied research and translation efforts will be realised. Startup and entrepreneurship promotion will be worked on by the Hub. IIIT Hyderabad already has the largest academic incubation facility in the country with over 100 companies. Startup companies are the best means to transform the research in the labs to the benefit of the society. A fellowship for aspiring entrepreneurs and seed-fund for early stage startups will facilitate seeding new startups.
d. High-quality teaching and skilling programmes shall be launched to produce highly skills students/researchers/faculties and professionals. Establishing PG, Doctoral and Postdoctoral fellowships, faculty fellowships and chair-professorships will be a special focus.

e. Data (Banks, Services and Analytics) is an area that has caught the imagination of the industry, government, and public at large. The TIH-Data will continue to have technical events, targeted workshops, popular sessions, and so on, aimed at specific sections of the society to enable them to understand and exploit the fruits of Data (Banks, Services and Analytics).

The focus of the Hub majorly includes applications with significant societal impact. Teaming up with Central and State governments is essential for achieving this. IIIT-H has a strong partnership with the Government of Telangana on several fronts, including Health and Mobility. We intend to leverage on that as well as expand it to other governments so that the fruits of the technologies developed by the Hub will be useful to the larger society.

The TIH-Data will strive to integrate different entities with significant stakes in the core as well as application areas so that useful and impactful solutions and products emerge from the effort.

1.3.2 Data Foundation/Applied Research & Translation

Data-Driven Technologies has two components that perform two distinct functions: (a) a Data Foundation that deals with the collection, curation, validation, management, and access to data-sets in several chosen domain areas and (b) an Applied Research & Translation Wing that performs translational research and development to take research from the academic groups to make it closer to technology that can be transferred to startups or other corporate entities.

This gap between domain experts and computer scientists is hard to bridge as these are separate highly specialised branches of knowledge, practiced with very little interaction between these two types of institutions. It is highly implausible for the average domain expert (e.g. a doctor in the Health domain) to be well versed in computer science, or for an average computer scientist to be well versed in the domain (e.g. medicine). Inter-disciplinary efforts are the way to address this, but happens very rarely. Their organisational structures and ways of functioning does not offer any natural means for this to happen. The Applied Research & Translation Wing aims to establish such teams to facilitate development and transfer of the technology developed by the TIH. This confluence methodology has the potential to drive innovation in domains of national interest, which neither medical experts alone, nor computer scientists alone, nor those with sole access to data, can achieve in isolation.

1.3.2.1 The Data Foundation

One major objective of the TIH is to be the custodian of data relating to different domains. This involves collection, creation, curation, annotation, hosting, management, and providing services centred around the data belonging to several chosen domains. We propose to perform this function in the **Data Foundation (DF)**.

To the world-visible front end, the Data Foundation hosts the data on a suitable platform and provides mechanisms and APIs to access this data by authorized personnel. Depending

on the nature of the data, suitable privacy-preservation measures are to be adopted such as anonymization, randomization, encryption, etc. It is proposed to make the data available to researchers free on a need-to-know basis. Data is also an important raw-material for AI-based startups. Data elements suitable for this will be made available to qualifying startups with an intent to promote them. It is also envisaged that product companies and commercial solution providers will be interested to use the resources of the Data Foundation for their activities. We envisage establishing appropriate licensing terms and conditions to facilitate commercial exploitation of selected data through the Data Foundation.

At the backend, the DF performs several functions from the Data Science repertoire of activities. At the most basic form, data is made available to the DF from outside world in a raw and unstructured form. Such data needs to be organised using a suitably powerful and flexible schema that can adapt to current and future uses of the data, including many unforeseen ones. The raw data needs to be curated and organised using the schema. Validation of the data needs to be performed to ensure its quality and suitability for different applications.

Tools to analyse the data in different useful ways are essential for its effective exploitation. Developing such tools and APIs that provide access to both the raw data and its analysis also needs to be performed to ensure the data is usable to a wide range of potential users, ranging from domain experts with little computational expertise to AI/CS researchers interested in novel data processing techniques and algorithms.

One major activity of the Data Foundation is the collation, collection, and generation of data for several application domains. Several of these domains have little or no public data usable data today. Development of AI-based solutions in those domains is greatly stunted by the lack of data. The domains the TIH has in its sights include Healthcare including public-health and individual health, transportation including road data and vehicle data, buildings data including services and occupancy, and several others.

Collection of data is a tedious endeavour with long-term impact and requires active cooperation and help from the domains. For instance, collecting Covid-19 patient data requires first access to a suitable hospital that has the data and is authorized to share it under strict privacy safeguards. The data from such hospitals is likely to be raw and often from non-digitized sources. Services of suitably trained medical professionals are required to transcribe such data into a computer processable form. Effective tools can make such data transcription smooth and painless. The data needs to be validated subsequently with the help of medical doctors or technicians. To make the data useful to ML algorithms, it needs to be annotated using the vocabulary and practices of the underlying domain. This requires significant time of appropriate experts, by itself a difficult task. At this point, the data can reside in the storage of the DF to be analysed, accessed, and processed!

It is reasonably obvious that the process of creating and collating data is time-consuming, tedious, and expensive. The Data Foundation of our TIH has its target the creation of 20-25

high-impact data sets in varied domains, by employing a team of transcriptionists, curators, annotators, etc., from the domain of each. In the end, the Data Foundation will be a library of digital data that are curated and validated, a major resource for the technology community and application developers.

Application challenges and hackathons are important channels to motivate and energize the application developers including startups. The DF aims to conduct several such challenges each year. This requires collaboration with domain experts in identifying problems that are both worth solving and suitable for such engagements. Thereafter, suitable data needs to be made available for the community to start developing and testing their solutions on. Each system is tested on an unrevealed set of data to rank them with respect to one another. A leader board of top-performing groups can be updated periodically to keep the participants suitably excited. The development of easy to use and efficient mechanism to manage the whole process is central.

1.3.2.2 Applied Research & Translation Wing

The TIH has as its major focus the creation and application of data-driven technologies to several application domains of societal interest, with a special emphasis on the Indian scenario. Academic research often stops with an early proof-of-concept. The development of robust and usable demonstrations or implementations are neither of interest to the academic community nor in its realm of expertise.

One of the primary objectives of TIH-Data is to transfer the technologies developed by it to society through startups, established companies, or directly to the users through an app for the same. Towards this end, a dedicated team and effort are planned. This effort will focus, nurture and help to mature the knowledge available in the research labs to products and services. This involves (i) Technology development (ii) Translation and Productization of the technologies using a high-intensity development team, and (iii) Interfacing the industry and other agencies and managing the processes related to them.

The much-needed translation layer to bridge the gap between domain researchers and computer scientists, and achieve a confluence, is clearly missing in the present setup in India. This is a major cause for the dearth of meaningful and technology-based startups from India. IIIT Hyderabad has worked on this problem for the past two years, establishing a productization team consisting of project managers, developers, and student interns. About a dozen results from research labs of the institute have been or are being developed to a from that can be transferred to a startup or a company so far.

1.3.3 Collaborations and Partnerships

TIH-Data will build on all collaborations and partnerships of the institute listed above. Further, build an extended ecosystem that further amplifies and accelerates the process of applying research to solve problems in the society and enterprises.

The various types of partnerships

i. Academic research institutions.
ii. Technology enterprises.
iii. Government departments and ministries.
iv. NGOs and Not for Profits.
v. Grassroot support organisations in sectors like agriculture and healthcare.
vi. Technology Business Incubators in different institutions.
vii. Startup support ecosystem (like networks like TiE, investors, etc.)

A mutual win-win value models will be structured to ensure value for all partner in the process realising some serious technology innovation in the field.

1.3.4 Technology Projects

All of the activities of the Hub will be in Data and Data-driven technologies relevant in few selected domain areas.

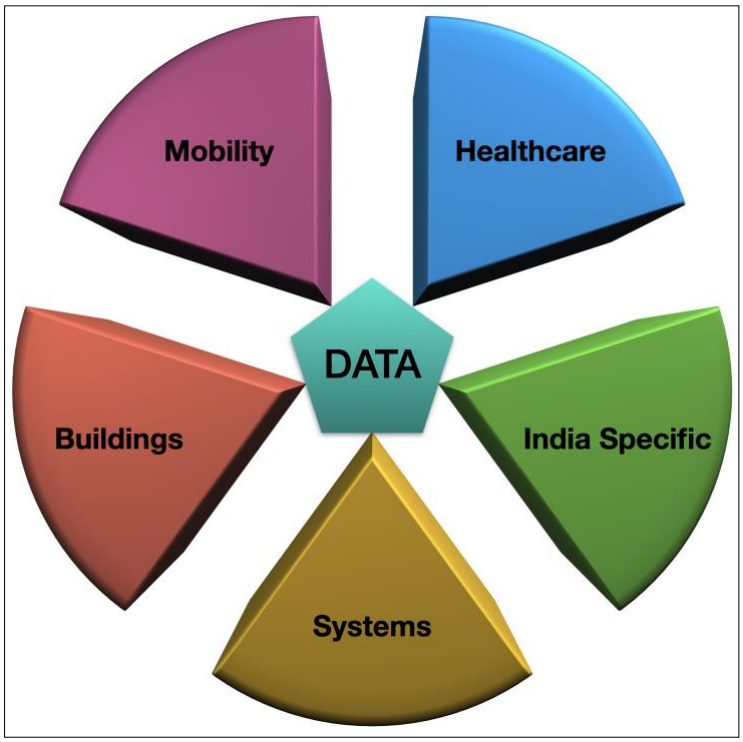
1.3.4.1 Data

Data is central to all of the work done in applied AI/ML. The algorithms only helps us bring out the information/patterns/knowledge already present in the data. Hence, availability of relevant, clean and well formatted data is one of the major bottlenecks in applied AI world.

The Hub will spend a lot of effort in data collection, curation, annotation and access as part of its activities. The hub will also follow the best practices in anonymization of the collected data, data privacy and security. For instance, medical data is ridden with huge volume and variety, and needs privacy and anonymization.

At this stage, five different broad areas in which the Hub will undertake research and development work have been identified: Mobility, Healthcare, Buildings, Systems and India specific problems. These have been selected by keeping in mind three aspects: (a) Needs of the country, (b) Importance of the area and the potential of data-driven technologies to transform the field, and (c) expertise and strengths of the host institute.

Figure 27: TIH-Data: Five Focus Areas



Next question is what kind of data will be collected? The Hub will work with domain experts to identify important problems areas that need immediate attention and has the potential for larger applicability. For example, we have initiated collaborations with Public Health Foundation of India, Telangana Government hospitals, CSIR labs and few startup companies in the area of healthcare. We have identified some problems related to public health starting from the public health centres, diagnosis and treatment protocols for COVID19, and cancer detection/classification based on histopathology images among others.

Once data sets are created and available in proper formats for algorithm development, the Hub will popularise and enable application work by conducting contests/challenges.

1.3.4.2 Domain Areas

Five different domain areas have been identified in which development and implementation of data-driven technologies are perceived to bring about a marked transformation. These areas also have been chosen carefully so that the research done at the Hub will have an international impact, the Hub has local expertise & prior experience and that the problems address the needs of the society.

Mobility

Mobility sometimes taken for granted is fundamental to society, and has a significant connection to the economy, wellbeing, comfort, etc. There are several aspects to the mobility area some of which include smart transportation (telematics, infotainment, communications, driver assistance, autonomous operations), mobility within indoor spaces

such as within warehouses and hospitals, etc. One of the major focus subareas of the Hub in this space is to 'improve safety'. It is a fact that the number of lives that are lost in India due to road accidents is significantly higher than in the developing countries. How can we minimize road fatalities? It is well known that commuters around the world including in India spend a lot of time on road due to traffic congestions. How can we make traffic more efficient? Recent covid19 pandemic has shown that a large number of healthcare workers are being infected. Can we use autonomous robotic technology to sanitize the place to make the environment safe for the healthcare professionals? Availability of data sets developed for Indian conditions for example is crucial to think of solving these problems. Most of the work, for example, in the area of autonomous driving is based on the data sets from the west, and that the technologies developed using these data sets will fail on Indian roads would be an understatement!

Healthcare

There is no better time to talk about the need of the improvement of healthcare than today when the world is reeling under the covid19 pandemic. Even the most robust and sophisticated healthcare systems in countries like Italy, Spain and the USA are being exposed in terms of their inability to handle the present crisis. There is already a number of proof of concepts available that shows that data-driven technologies will make certain processes within the healthcare system more efficient, for example, diagnostics. On the other hand, AI based methods are shown to alter the way in which new drugs and diagnostic protocols are being discovered. One of the main objectives of this activity is to identify areas in which the country benefit in and collect medical data sets based on which implementable technology solutions can be developed. For example, collection of image data related to cancer, collection of genomic data related to CoV 2 infection, etc. have been already initiated as part of the activities of the Hub. How do we make a difference at the level of the public health centres? State of the art research in the pure academic setting is related to demonstrating proof of concept and working demonstrations at best. However, these solutions need several years before it can reach the mass, when it does very rarely! For example, is it possible to have a quick and low-cost diagnostic protocol for detecting TB, malaria, dengue, etc.? To identify problems that are being faced at that level, we are plan to collaborate with the Public Health Foundation of India and work with other Government hospitals (district level) to identify problems and develop implementable solutions. How to optimize the minimal resource so that healthcare availability reaches maximum population? Can we develop data-driven technologies for supporting clinical decisions? Can we predict disease outbreaks?

Buildings

Buildings are an integral part of human life. Be it in an office, home, school, or a shopping mall, people spend most of their time inside them. The primary requirements of a building are to provide safety, comfort and enhance productivity. However, buildings consume a lot of resources such as energy, water, land, air, material, and so on. They also generate a lot of waste and pollution. With rapid urbanization, the number of buildings is also growing

rapidly which will only increase the stress on natural resources (such as water and non-renewable energy sources). Buildings is an important area of research in the Indian setting for several reasons. One of the prime reasons is because there is an exponential increase in the residential and commercial built up area in India currently. This leads to a series of other problems such as energy consumption, resource availability, waste management, maintenance, safety, etc. Smart buildings are seen as solutions to some of these problems where data-driven technologies, IoT sensors, etc., are seen to contribute. For example, why and how should the building adjust to occupancy, light, temperature, humidity and radiant heat? How can we maximize comfort (ambient lighting, cooling/heating) with minimal energy at the individual level? Integration of data, AI algorithms, IoT devices and traditional computing will be used as a strategy to develop solutions for smart buildings. Indicative problem areas that will be undertaken as part of this activity are given later.

Systems

Systems play an integral role in data-driven solutions from collecting data, developing technologies and technology implementation platforms. Some of the important sub areas include edge computing, edge-fog-cloud computing/optimisation, optimisation of data bases, distributed computing, optimization of systems in general, data for systems, etc. Many application areas such as healthcare, automotive industry, smart warehouses, smart cities, smart buildings, smart power grid, intelligent transportation, smart agriculture include the use of edge devices for a variety of tasks. These edge devices usually have meagre available compute, memory and storage resources. Mostly, they are powered by batteries and sometimes these are even irreplaceable due to physical constraints, like a pace maker implanted inside a human body. The network connectivity could be intermittent due to lack of signals. Along with these constraints, certain applications demand privacy and require real-time action or response. Hence, transmitting data to a fog node or cloud is not possible. Hence, computation needs to be done and actionable information has to be inferred at the edge node itself. How do we optimise the storage solutions and the hardware for data storage, distribution/hosting, algorithm development and to implement solutions? On the other hand, how can data contribute to systems? For example, can we use data to predict system behaviour/performance? Can we detect anomalies and predict system failure in advance? Can we use data-driven technologies to optimise resources which may include efficient performance and energy?

India-Specific Research Initiatives

The above discussed four areas and the data itself will cater to two major needs. One is based on what would bring visibility and recognition to the Hub and its activities at the national and international level among the peer researchers. Second is based on what is of immediate need in the Indian context. From time to time, problem areas will be chosen to understand the needs and to identify those which will impact a larger population. Some of the fundamental questions that are planned to be addressed are: Can we develop technologies for predicting crop yield? Can we develop technologies that can automatically detect weed and remove them? Can we develop technologies that can detect pests and

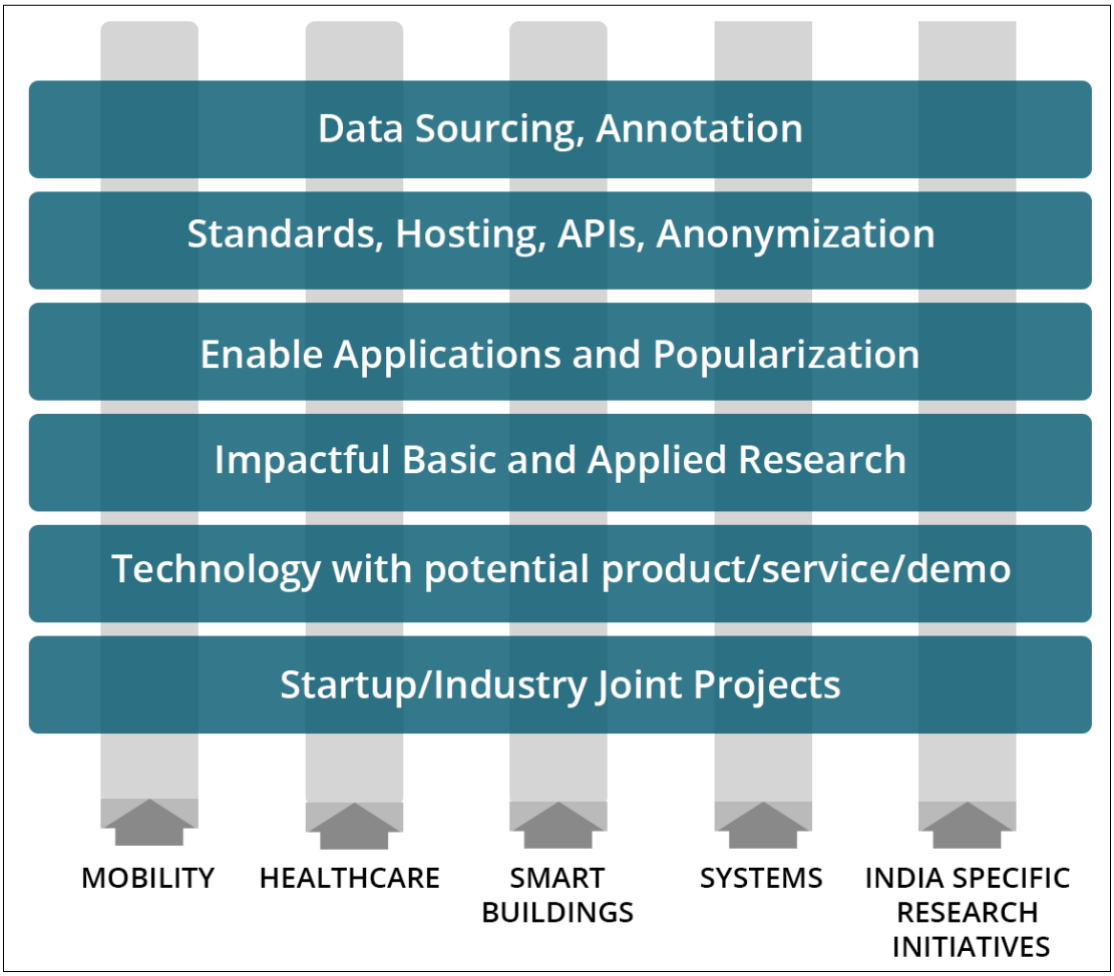
localize treatment so that overuse of pesticides be avoided? Can we minimize resource usage (water, energy, pesticides, fertilizers) and maximize production using data-driven methods? How do we device low-cost diagnostic methods for diseases such as TB, dengue, malaria, etc., that are more common in countries like India but not in the western world? How can we use technology to optimize the use of resources to reach maximum population? Similarly, there are a lot of problems in education, culture, national heritage, especially abled, energy, disaster prediction and management, finance, food security, water security, etc., where developing technology solutions will find a large value in the country due to immediate applications. Hence, in addition to working on data and four other domain areas, the Hub will spend a lot of effort in India specific problems.

2 Challenges to be Addressed

The Hub seeks to address technology gaps and needs for solving problems for society and industry in the areas of data services and analytics.

The activities of the Hub comprise the Data Foundation, applied research/translation, technology development and innovation/startups. All these components will work in the intersection of data banks, services & analytics and domain areas such as mobility, healthcare, smart buildings/habitat, systems & India specific research initiatives shown in the figure below:

Figure 28: Proposed Activities of the Hub in the Interface of Data and Other Domain Areas



2.1 Data

The Data Foundation will host a library of digital data, and will encompass the technology-platform, infrastructure and manpower to collect, create, curate, annotate, secure and deploy it. It will be a major resource for the technology community, researchers and application developers who need such data for developing solutions driven by AI and analytics in socially-relevant domains such as Healthcare, Mobility, Buildings, Systems, and their application in the Indian context.

State of art technology in data analytics and AI available nationally and globally has proven its capabilities and benefits in several domains. Yet, critical gaps remain, as they require large, curated datasets, annotated specifically for each domain problem that needs to be addressed. Identifying and solving the problems that will streamline the building of valuable datasets is a critical need of the day. Some interesting and difficult challenges that present themselves in this context are described below.

2.1.1 Technical Challenges

(i) Universality

The technology-platform to be built for the Data Foundation should be usable *across* domains.

Today, data analytics along with modern machine learning is a mature technology with several open and commercial implementations that are fast, accurate and scalable when designed for any particular application.

However, scaling data analytics across domains remains a major challenge for the data analytics community, both nationally and globally. The bottle-neck in this endeavour lies in developing a deep understanding and abstraction of the needs of the domain, along with the knowledge of possibilities of modern analytics and machine learning solutions. Such abstraction and identification of needs and possibilities requires the continuous engagement of collaborating inter-disciplinary researchers, and is not a task that can be easily packaged and off-sourced to the industry.

The platform to be built needs to capture the aspects that are common in such engagements, while allowing flexibility and configurability in aspects that are diverse across domains.

(ii) Scalability

The volume, velocity and variety of data captured in domains like healthcare and mobility pose unique challenges. Traditional issues such as scaling up (by adding more storage) or scaling out (by adding additional servers) need to be re-considered in the context of multiple varieties of requirements across domains. Additionally, novel issues in a platform view point arise. Resources need to be allocated, commissioned, decommissioned and recommissioned appropriately depending on the storage and computation need, which can change drastically across project lifetimes for each dataset.

Data model design needs to be flexible to cater to the variety of different sources of data that may come with different conventions, fields, labels, formats and missing values. Potential queries and use-cases need to be estimated, and used to optimize data model design, while ensuring scalability. Solutions would lie in the plethora of new NoSQL solutions present today, or may need to be built/added on/integrated. Navigating the NoSQL solution landscape for this context is itself mind-boggling, with 350+ technology alternatives as of September 2021 (https://dbengines.com/en/ranking).

These are interesting challenges, as solving them paves the way to solve common pain-points across projects and domains with common solutions and methodologies.

(iii) Privacy and Security

Data is expected and mandated by national and international standards to abide by, protect and advance citizen's privacy rights. Data that is collected comes with a range of privacy and security requirements. Healthcare data is of particular concern, having been considered worthy of protection ever since the hippocratic oath.

On the other hand, there is a global need to be able to study and analyze healthcare data, in order to advance the field, to cure and prevent disease. Depending on the nature of data, suitable privacy- preservation measures need to be adopted such as anonymization, randomization, encryption, role-based access, etc. Additionally, appropriate usage agreement guidelines and protocols need to identified and implemented to allow the use of such data for research purposes, while following mandated ethics agreements.

Data also needs to be protected for other reasons. Often, researchers engage in building datasets with a keen interest in having the first opportunity to be able to study and analyze them and attain credit for novel research. Mechanisms need to be implemented to make datasets available in secure ways with role-based access to only limited groups of people, and allow the datasets to be published later, in possibly multiple versions.

(iv) Automation in Data Collection, Curation and Annotation

Collection, curation and annotation of data is a tedious endeavour. The challenges in this arena are two- fold:

(1) To build, maintain and manage teams or crowds of skilled data-workers, allocating them a steady supply of skill-appropriate tasks, maintaining quality control and motivation.

(2) To build a toolkit that enables automation of those data tasks that can be automated. This includes domain-specific and domain-agnostic tools such as data scrapers, format convertors, annotation tools, etc. There is ample scope for developing several novel machine learning based methods for such automation.

Automation requires the data to be validated subsequently with manual expert or non-expert help. To make the data usable, it often needs to be annotated using the

vocabulary and practices of the underlying domain. This requires significant time of appropriate experts, by itself a difficult and costly task.

It is reasonably obvious that the process of creating and collating data is time-consuming, tedious, and expensive. For managing teams of data-workers, a project management platform built for such purposes needs to be built and/or deployed. The Data Foundation has its target the creation of 20-25 high-impact data sets in varied domains, by employing a team of data workers for transcription, curation, annotation, etc.

(v) Algorithms for Core Data Analytics and Machine Learning

Part of the Data Foundation activity is to promote overall research to produce novel algorithms and improve existing algorithms for data analytics and machine learning. Of specific interest is the effective application of these algorithms to socially relevant applications. Another interesting research problem is to analyze data that is distributed or federated, and cannot be brought into a centralized place due to privacy/security considerations, or due to sheer volume.

While deep neural networks (DNNs) have achieved high performance on a variety of tasks, there is still scope for improvement, especially in applications where data is scarce, costly or restricted. Research of interest includes generating good quality synthetic datasets, data augmentation methods, attacks on DNNs and identifying deep fakes, making DNNs robust to attacks and identification by augmenting datasets, building interpretable methods, etc.

2.1.2 Non-Technical Challenges

Enforcing and following professional software development practices is not easy in a research context because incorporating rigid development guidelines would hinder the flow of creative thought necessary for fluid abstraction, which is required for interdisciplinary research.

Requirements are often detected in later stages, requiring extensive changes in design and re-implementation. Experts in different domains are often unfamiliar with the needs and capabilities of the other, and may not have the right expectations until familiarity dawns. Further, tasks done in other disciplines may seem either overly trivial or overly difficult, even when it is not so.

Additionally, common problems that plague the development world include junior developers leaving every 2-3 years after being trained; unreadable and unmaintainable code needing re-development; difficulty of actually integrating separate pieces of code developed by several developers; lack of an open channel to raise issues; not following coding standards; and incomplete testing.

A university/academic/research setting with teams of students, junior developers and multiple disciplines is especially vulnerable to all these kinds of problems. Building the right culture and practices to overcome these problems is a major challenge worth solving due to the immense potential impact of the endeavour, that would not be

possible in any other non-research setting, when it may take years of collaboration and abstraction to identify concrete underlying needs.

2.2 Application Areas

The Hub will undertake research related to datasets applied to two major domains, Mobility and Healthcare. Other areas that the hub intends to do technology development and productization are Smart Buildings, Systems and India Specific Initiatives.

2.2.1 Smart Mobility

(i) Enabling Data Driven Mobility Solutions for Indian Roads

Background: Technology and solutions for road safety and navigation rely heavily on AI. Modern AI has become data driven. However, it is well known that many of the state of the art solutions for road safety have not percolated into the Indian market. One major reason, is the failure of the perception and planning modules in these solution. They are not able to cope up with the Indian setting. We need active research and technology development to make this happen. Research and development in this space need systematically collected and well curated data sets. This is seriously lacking.

Challenges and Directions:

(a) **Unavailability of Indian Data**: Autonomous navigation has received lots of attention in the last decade, especially in the western countries. Many data sets from USA and Europe fuelled the research and development in this space. Beyond the goal of full autonomy, the active research in this space has resulted in improving the technology for safety on roads. However, many of these solutions are not usable in India since we do not have appropriately curated datasets. Our goal is to create such datasets and make them available for research and development widely.

(b) **Lack of Data under "all" Diverse Conditions**: Today's technology needs "similar" data while development of the solutions for a specific setting. Generalization to similar situations (that it had seen earlier during the development) is what these solutions primarily demonstrate. As a result, there is always a concern over the performance in novel situations (like driving under rain). This also creates serious problem for Indian road situations where novel situations, novel objects and novel movement patterns can be expected at any time. Collecting and curating data in all these environments is going to be a paramount problem. Our plan is to use techniques from machine learning, computer vision, computer graphics, computational sensing and allied area to create situations and aid transfer of solutions across diverse practical situations that one could anticipate across Indian roads.

(c) **Diversity in Data, Tasks, Knowledge and Representations**: Data related to smart mobility in Indian roads is highly diverse and multimodal. Even the sound signals

con- tribute a lot to the reasoning process on Indian roads. Not just the static world or the objects on the road side are diverse, Indian roads are unique in driving patterns, and driving strategies. This diversity necessitates a design of the data and representations that accommodate diversity in modality, tasks and knowledge.

(d) **Lack of AI-Technology product evaluation frameworks**: Evaluating data driven technologies at scale is still challenging. Proper evaluation protocols are still emerging. Differing with the traditional software testing paradigms, these AI solutions need large and diverse data for testing. One of our goals is to design such benchmark frameworks.

Figure 29: Focus Areas of the Hub in Mobility

(ii) Reduction of Fatality on Indian Roads

Background: India reports around 150,000 deaths every year due to road accidents. This is very very high compared to other countries. This is also very high compared to other types of accidents and deaths that India reports.

Challenges and Directions:

(a) **Research, Technology and a Catalyst**: One of the strategies of the hub is to establish itself as a catalysis for developing technologies and absorbing into the vehicles. Beyond auto manufacturers, there are a number of players who can use technology and help in reducing the road fatalities. This also demands nurturing an ecosystem that innovates and develops technologies for Indian roads.

(b) **Field Trials and Validation**: "Seeing is believing". Hub believes in demonstrating the role and effectiveness of the technology in improving the road safety with city scale deployments of the technologies. This validates the role of technology, bring many stake holders into one platform, and give many practical lessons from the field.

(c) **Scalable Detection of Activities on Roads**: Many governmental agencies have established large surveillance camera networks for improving the safety of the cities. Cost of establishing such networks is very high, unaffordable at this scale. We propose to develop scalable solutions that can be deployed or transported to newer areas with ease. Ego centric cameras mounted on vehicles can effectively survey wider areas. They also complement the existing surveillance strategies.

(d) **Enhancing the Awareness**: Educating the drivers and the larger society is a key component in improved the safety. Over the years, quality of the process for many driving tasks have been improved. There is more to be done to make an impact.

(iii) Technologies for Two Wheeler Safety in India

Background: India adds new 15 to 20 M two-wheelers every year. Two wheelers are cost effective and easy to navigate on most Indian roads. Cars still remain as a vehicle for "rich". Nation has local expertise in design and manufacturing two wheelers. Local neighbourhoods have expertise in maintenance, alterations and even introducing local innovations. Cost is still the primary factor that drives the choice of vehicle. It is not just the vehicle cost. It is also the mileage (km/litre) that people carefully compare. In the middle of focussing on cost and affordability, this segment pays less attention to the safety of the vehicles. At the same time a larger percentage of the number of road accidents are directly connected to two wheelers.

Challenges and Directions:

(a) **Safe Driving**: A number of problems that are identified for safe driving in cars need their counter parts for two wheelers. For example (i) driving quality (ii) driver attention (iii) collision avoidance (iv) rider safety (v) crash detection and alerts. All these require adaptation to two wheeler space.

(b) **Cost Effectiveness**: A primary concern in two wheelers is the cost effectiveness. This limits the type of sensors that can be used. Also the type of computing and communication devices that can to be part of the driver assistance unit.

(c) **Efficiency, Economy and Sustainability**: Beyond the safety, technology can play important role in improving the efficiency, economy of the vehicles, reduction in pollution and encourage sustainable transportation.

(d) **Support for Three Wheeelers**: One of the uniqueness of Indian transportation sector is the omnipresent three wheelers. Technology for this class is similar to the two wheelers in many respect.

(iv) Scalable Road Infrastructure Inspection

Background: India has a huge road network of around 6M kilometers. This is the second largest in the world. These roads are diverse in quality and transportation infrastructure. They are also maintained by different departments, authorities. This naturally leads to uneven maintenance of this infrastructure. Knowing the present status of these road segments itself is a challenging task for planning the maintenance required and allocating the necessary budgetary resources.

Challenges and Directions:

(a) **Need of Scale**: Given that our roads span Millions of Kilometers, we need solution that can work at this scale, and efficient.
(b) **Need of Solutions that are Cost Effective and Ease to Use**: Such infrastructure inspection and logging effort is presently human labour intensive. Human will have to inspect and log the status. For example, regions that need maintenance. Often this log does not come with proper evidences and prone to human errors. Given that it is done by humans, it is not feasible to do it at a very regular intervals. High-Tech solutions will find it difficult to penetrate to local roads and bodies since they could be costly and more over may demand expertise to use.
(c) **Participatory Process**: Working at this scale need participation of all stake holders, people, local bodies, state and central governments and of course a number of semi-govt companies that gets involved. A solution to this need to be “transparent” and “participatory”. Every participant should be able to contribute.
(d) **Variations in Style of Artifacts**: A vast amounts of public spaces in India are centered around roads. Not only the diversity of the type of objects (animals, vehicles) that move on roads, but also the road side artifacts are very diverse in appearance. Many standards are difficult to follow due to local reasons. They are adapted with exceptions. Objects of interest around the roads, start from trees, building, build boards, bus sheds and many other objects around. This extends upto road surface, surface quality, material quality and markings on the roads and traffic signs.
(e) Lack of Technologies This space is today distributed across departments and monitored by humans at the local level. The goal is to create automatic and semi-automatic tools that can help analyze the data centrally collected and tag/characterize the failures or works that need maintenance.

2.2.2 Healthcare

(i) 100% coverage of basic health services for a rural Public Health Centre (PHC)

Background: According to the Lancet Global Burden of Disease Study (2016) (updated in 2019), non-communicable diseases (NCDs) contributed to 61.8% of all deaths, while the communicable diseases contributed to 27.5% of all the deaths. Three of the top five causes of death were NCDs (ischemic heart disease, chronic obstructive pulmonary diseases, cerebrovascular disease), and the remaining two were

communicable diseases (diarrhoea and lower respiratory infections). In all these cases, early detection and diagnosis are critical. Over 70% of the population in India live in villages and majority of the people (80-90%) in a rural public health centre (PHC) do not even get basic health services. Accredited Social Health Activist (ASHA), Anganwadi Worker (AWW), and Auxiliary Nursing Midwife (ANM) staff are the main pillars of healthcare delivery to the rural homes. The challenge is to empower the healthcare workers using data-driven solutions to bring basic healthcare awareness, monitoring and follow-up services to the last mile.

Challenges and Directions:

(a) **Lack of Skilled Personnel:** According to the 15th Finance Commission Report (2021), the ratio of allopathic doctor-to-population ratio in India is 1: 1,511, much higher than the World Health Organization (WHO)'s norm of one doctor for every 1,000 people. Similarly, nurse-to-population ratio of India is 1:670 against the WHO norm of 1:300. The only way to bridge this gap is to utilize the services of ASHA, AWW and ANM staff. While they have their traditional roles in rural health sector in immunization, nutrition, child health, etc, the big opportunity is utilizing their presence to enhance the delivery of basic health services. Technology-enabled solutions such as *apps* for early diagnosis and prevention is an important direction.

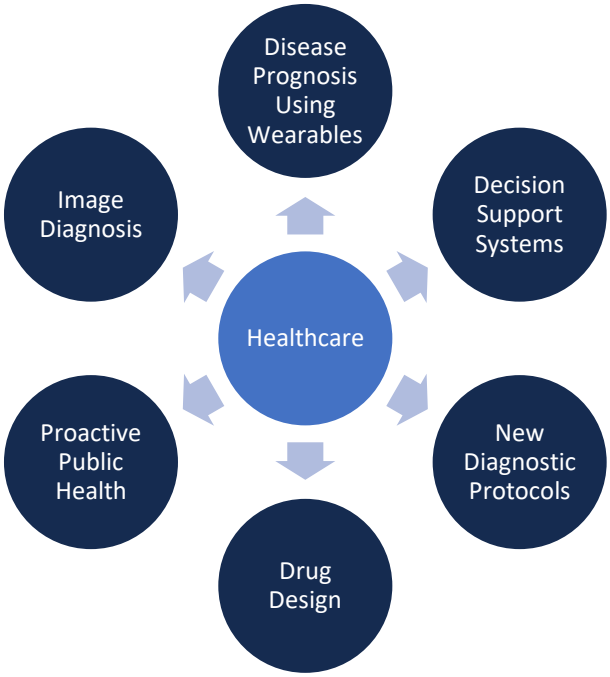
(b) **Trust and Knowledge gap:** Because of limited knowledge of the auxiliary medical staff and meagre infrastructure at PHC, there is tendency to flock to District and City Hospitals even for treatable and common conditions. This results in loss of valuable time, loss of daily wages in some cases, escalated costs of treatment, and wastage of resources. Rural health workers lack the expertise to direct the patients to appropriate medical facility at the right time. Technology-enabled solutions would come in handy in building the trust and bridging the knowledge gap.

(c) **Coverage and Scalability issues:** Since unskilled or semi-skilled health workers would be expected to provide assistance in diagnosis and advice related to the steps to be taken, the technological solution needs to cater to wide range of medical conditions and cater to disparate geographic locations and their disease peculiarities. The solution could be initially rule-based and eventually AI-driven when there is enough data available. Also the UI/UX needs to be appropriate for the skill level of ASHA/ANM. All these solutions need to be compatible with the National Digital Health Management (NDHM) framework.

(d) **Disease Surveillance Challenge:** Pandemics like SARS-CoV-2 happen in centuries to challenge the human existence with over-whelming pressure on the existing healthcare system. The challenge has been pervasive across the globe and in most cases, it caught the scientific community unaware, although there have been other coronaviruses. The COVID-19 has been characterized by diversity of symptoms with majority being asymptomatic. Within symptomatic also, there has been mild, moderate, and severe categories with a fraction resulting in mortality. There has been significant role of Genomics in taking forward the Genomic Surveillance for

SARS-CoV-2 with more than 1.25 million genomes being submitted to GISAID. It has highlighted the strength of Genomics in knowing the pathogen, underlying mutations, mutations linked to disease outcome and genomics surveillance. Genome surveillance through a network of Microlabs to track mutations and variants of concern, developing cheaper diagnostic panels, and helping with policy making for future outbreaks are the key solutions. For holistic surveillance plan, it is important to address challenges faced due current and future diseases caused by pathogens such as SARS-COVID-19, TB (Mycobacterium Tuberculosis Mtb), Dengue, Japanese encephalitis, Cervical Cancer (HPV), Female genital tuberculosis, etc.

Figure 30: Focus Areas of the Hub in Healthcare



(ii) Early Screening and Prevention of Top-5 Cancers in India

Background**:** One in 10 Indians will develop cancer during their lifetime (WHO report). The six most common cancer types in India are: Breast cancer, Oral cancer, Cervical cancer, Lung cancer, Stomach cancer, and Colorectal cancer (cancerindia.org). India sees an incidence of more than 1 million new cases of cancer every year, and early detection and management can be crucial in an optimum cancer treatment regimen across the country. Connecting to primary healthcare settings for screening common cancers is very crucial.

Challenges and Directions:

(a) **Early Screening and Digital Pathology:** India has barely 2,000 pathologists experienced in oncology, and less than 500 pathologists who could be considered an expert onco-pathologists. The knowledge and expertise gap can be mitigated

by AI-driven assistive systems for diagnosis and treatment. Digital Pathology is an important direction to work on. Digital Pathology needs to pave the way for facing the challenge of diagnostic uncertainty, improving the diagnostic accuracy, and disease staging, etc. Part of the diagnostic uncertainty is caused by significant inter-pathologists' variability in cancer staging. Staging is very critical as the treatment protocol depends on the grade (grade 2 versus 3 have different treatment protocol, for example, for Lupus).

(b) **Data availability:** AI solutions require annotated datasets of appropriate coverage of different types of cancer at various stages. It is important create a large repository of pathology datasets along with clinical notation. Curating such repository requires bringing together various pathology departments to contribute their existing collection of clinical data, annotating them, and then building an AI-based solutions. Coordinating with pathology departments at various hospitals and negotiating the ethics protocols for such retrospective studies offer immense challenges.

(c) **Multimodality challenge:** The pathology datasets comprise images from multiple modalities: Pathology images (cytology/histology), Radiology images (CT/MRI/X-ray/Mammograms/ultrasound/nuclear medicine scan), Endoscopic imaging/video (Colonoscopy, Gynoscopy), and Photographs/Others (skin, oral etc). AI solutions need to be appropriately configured to deal with multimodal data comprising not only images but also clinical observations in the form of textual data in natural language.

(d) **Cultural Impediments:** While cervical screening is important for early detection and prevention of Cervical Cancer, in the rural settings there is often resistance or taboo in undergoing repeated examinations. In such low resource settings, technology can make the task easy by using cytopathology-based screening rather than colposcopy. Cytopathology based screening based on AI-driven algorithms is a possible way forward in overcoming cultural impediments to invasive screening.

(iii) Prevention of Stroke Recurrence and Optimal Rehabilitation

Background: Stroke is the leading cause of disability and mortality worldwide. Sleep-disordered breathing or Sleep apnea is an independent risk factor for stroke, with an estimated two-fold increase in the risk of stroke. Sleep is window into Stroke through investigation of sleep stage disturbance and sleep apnea. Almost one-third of our lives are spent sleeping. Three key elements which are responsible for sleep regulation are circadian rhythm, ultradian rhythm (for example, sleep-wake cycle), cognitive and behavioral factors. Insufficient sleep duration and sleep problems such as insomnia and sleep apnea are very common. About 33% suffer people from insomnia and about 9% to 38% suffer from sleep-disordered breathing or sleep apnea. Although sleep is an important physiological function and its role in memory consolidation, synaptic homeostasis is well studied, sleep dysfunction is less well documented in routine clinical practice. This is due to the inherent difficulty in undertaking the gold standard for assessing sleep disorders using the overnight polysomnography (PSG). We need solutions for easy and continuous monitoring of parameters of sleep. The other issue is appropriate post-stroke rehabilitation protocol for optimal recovery outcomes.

Especially in tertiary health institutions where the patients are predominantly from rural and remote locations, compliance with the rehabilitation protocol is very poor. Thus, loss to follow-up in rehabilitation leads to poor post-stroke outcomes.

Challenges and Directions:

(a) **Shortage of Therapists:** Stroke therapy utilizes various professionals such as Physical Therapist (PT), Occupational Therapist (OT), Neuropsychologist, Speech Therapist, based on the conditions of affliction due to stroke. There is severe shortage of trained therapists and especially so in tertiary institutions (Super Speciality Centres such as NIMHANS). A technology-driven solution that triages therapy based on health monitoring and vitals is critical for optimal use of available therapists as well as for sending alerts to patients about their therapy protocol at appropriate times.

(b) **Remote Monitoring:** Most patients are from rural locations and are geographically dispersed. So remote monitoring becomes critical to identify early signs and for rehabilitation advice. Wearables offer an excellent solution, but they are not clinically validated. Immediate challenge is assessing the suitability of available wearable solutions for monitoring as well as to investigate their clinical validity.

(c) **Data annotation for fast workflows:** Polysomnography (PSG) test is the gold standard to for assessing sleep. PSG records the brain activity through electroencephalography (EEG) signals, heart activity through electrocardiography signals (ECG), the oxygen level in the blood, breathing, eye movements through electrooculogram (EOG), and muscle movements through electromyogram (EMG) during the test. A PSG test for an individual is done in a controlled environment in a lab (typically lasts for 8 hours during the night) performed by a trained sleep expert. During the analysis of the PSG data, a sleep expert divides it into individual epochs and manually annotates the sleep stages in those epochs according to the American Association of Sleep Medicine (AASM) rules. It typically takes around 4-6 hours to perform sleep staging on the whole PSG data, thus making it time-consuming and expensive. Deep learning based automated sleep stage annotation algorithms would enable fast workflows. For this, annotated PSG datasets are required. Curating such repository is an important step in this direction.

(d) **Clinical relevance:** Cognitive outcomes due to sleep disturbances could be more clinically relevant than simply knowing the quantum of different sleep stages. Cognitive consequence of sleep disturbances and similarly how sleep apnea leads to disturbed sleep which in turn leads to cognitive deficits is still not clear in the Indian cohort. A prospective study that investigates the relation between sleep disturbances and its relation to cognitive abilities is warranted. If in the process, various surrogate measures from wearable devices are also investigated for their clinical relevance, we would have found an easily implementable, portable, and scalable solution to the problem.

(iv) Impact of Long Covid on Cognitive and Mental Health

Background: The 'Minding our Future' report from Universities, UK [https://www.universitiesuk.ac.uk/sites/default/files/field/downloads/2021-07/] says '50% of mental health problems are established by age 14 and 75% by age 24.' Students form a significant group within this young adult population. There is urgent need to estimate the prevalence of mental disorders in young adult population in India and initiate measures for early detection and prevention. The recent pandemic has made this problem more pressing with the relative isolation from peers and mentors.

In the ensuing pandemic, the medical fraternity is fighting a heroic battle to save lives through immunization, home-based treatment of those tested positive for SARS-CoV-2 and intensive care for the critically ill. Going forward, India needs to prepare for the aftermath and the consequences of the disease. One such consequence that is widely reported is a detrimental effect on cognition. There is some evidence that the SARS-CoV-2 virus may enter the central nervous system. Mild to severe neurological symptoms have been reported in patients infected with SARS-CoV-2. However, more subtle effects have been reported both in peer-reviewed literature and anecdotally by doctors treating these patients such as detrimental effects on sustained attention. The case of cognitive and mental health issues post-recovery from the attack of coronavirus needs immediate attention.

Challenges and Directions:

(a) **Prevalence of mental health conditions in the young adult population**: It is important to understand the scale of prevalence of mental health problems in the young adult population. Like several countries (USA, UK, etc), India too has done a systematic assessment [http://indianmhs.nimhans.ac.in/nmhs-results.php]. Access to the young population and the social stigma attached to any mental health debate are challenges in this front. Technology could play a role in mitigating these challenges. Such studies could take advantage of the advances in the *Experience Sampling Methodology* (ESM) by employing surveys, mood samplers, mini cognitive tasks via smart mobile devices. ESM increases ecological validity of the study, is convenient for experimentation without having to bring the participants to the lab and allows near-continuous monitoring of the enrolled study-participants. Similarly, social media offers unprecedented opportunity for studying the interaction patterns, habits (Music listening, etc) and opportunity to infer the dynamic mental states from textual analysis of social media interactions (on Twitter, Instagram, Facebook, etc).

(b) **Cognitive difficulties during the post-recovery from Covid**: There is anecdotal evidence as well as evidence from small studies abroad that the patients who recovered from Covid-19 are facing cognitive and mental challenges post-recovery. There is urgent need for wider cognitive testing among the Covid-19 patients who recovered from mild, moderate, and severe conditions. Web-based

delivery of survey questionnaire and cognitive testing would be a way to go for wider and safer access to the patients in the ongoing pandemic scenario. A set of standard cognitive tasks could be easily administered through a webpage that participants can access from a laptop or an iPad. Using such a setup, cognitive data can be collected at a large scale, such that the sample includes covid-recovered patients with different levels of experienced severity of COVID-19 as well as healthy controls. Such a study allows to estimate the severity as well as initiate appropriate measures for rehabilitation.

(c) **Multimodal Data Fusion**: Several efforts are taken up recently at collecting data of various kinds such as, genomics, endophenotyic measurements of MRI, fMRI/fNIRS, EEG, eye movements, along with clinical observations including neurocognitive assessments on patients with mental disorders as well as from closely related family members of the patients [https://adbsnimhans.org/]. In line with the efforts at collecting multimodal data related to Mental disorders, data fusion approaches and discovery of patterns in the multimodal data need to be initiated. The effort might require modern methods of data representation using graphs and learning associations and knowledge extraction from multi-layer graphs.

(v) Ageing and Neurodegenerative Disorders

Background: Longevity has increased, resulting in a rapid increase in the aged population and the associated problem of increase in number of patients with neurodegenerative disorders such as dementia, parkinsoniasm etc. According to the Longitudinal Ageing Study of India (LASI) released in 2021, there will be over 319 million elderly by 2050 (this is threefold the number identified by the Census in 2011). Brain structure and function are known to decline with ageing. Age-stratified normative brain data is important to understand the neurodegenerative disorders that afflict the ageing individual. Reversing or delaying some of these conditions is of urgent need in the Indian setting.

Challenges and Directions:

(a) **What is normative brain structure for Indian population?** Magnetic resonance imaging (MRI) atlas is important for many neuroimage analysis tasks both in health and disease. An atlas with a standard coordinate system is mandatory for spatial normalization of a brain MRI and for initiating any comparative studies within the population and across atlases. Further, *age-stratified brain template* would facilitate the study of ageing-related changes in the average Indian brain. Such age-stratified template would play a key role in clinical assessment of neurodegenerative disorders. Major challenges are data collection from a representative population across India, data curation and annotation of the data by expert radiologists. This requires a large-scale, pan-India effort.

(b) **Annotation of Large Brain Data sets**: It is impossible to annotate slice-by-slice all the brain volumes of the cohort. We need to transfer the annotations from small datasets that are carefully, manually annotated by expert radiologists to larger corpora. Recent advances in deep learning approaches need to be exploited for automatic segmentation and label transfer. Validation by experts is another crucial step after the automatic label transfer.

(c) **Relation between Structure and Function of the Brain**: It is evident that the two epochs in life that result in massive changes in the brain structure and function are during the development (early part of the life) and during ageing (later part of life). It is still not clear how function remains relatively stable in face of declining structure due to ageing. Modern machine learning and deep learning methods would be useful in characterizing changes in the brain structure and function both during development and due to ageing. Advances in graph-based representations and methods of finding patterns of association among multimodal data (brain imaging, behavioral and cognitive scores, etc) are key drivers of innovation for this challenge.

(vi) Reducing the lead time in Drug Discovery and Design using Data-driven approaches

Background: The drug discovery process is expensive and time consuming due to various factors including the need for assembling a large interdisciplinary team, extremely high attrition rates, etc. Computational methods are crucial in the initial stages as they are inexpensive compared to experiments and they can estimate binding affinity (drug action) as well as predict ADMET properties (absorption, distribution, metabolism, excretion, and toxicity). The challenge is to utilize the data driven methods to make more accurate estimates and prediction as well as in attempting approaches along the whole pipeline of the drug discovery process and not just in molecular design!

Challenges and Directions:

(a) **Data set challenge**: Although there are several popular datasets available, there are several limitations with the existing datasets such as less target diversity and ligand diversity, size of the datasets being small (AI algorithms typically require large datasets), low transferability to all drug targets, lack of high energy data (both structural & thermodynamic), and experimental errors in the data (measurement differences across labs and across protocols). Creation of ultra large and unbiased dataset is the need of the hour for effective deployment of AI algorithms. The dataset needs to incorporate protein/ligand diversity, multiple binding poses for all protein-ligand pairs, and dynamic factors. Creation of such an ultra large dataset is beset with *computational challenge* where in computational efficiency of the pipeline needs to be optimized and needs to cover several drug targets. Such end-to-end application of data driven technologies for drug discovery and design need to be supplemented with validation using biochemical assays, cellular assays, and in animal models.

(b) **OpenOmics platform**: Health and biological data are being generated at unprecedented rates. Traditional analytical approaches are no longer efficient or scalable. Various domain specific tools have been developed covering areas such as genomics (single cell sequencing, transcriptomics, epigenomics, whole-genome sequencing (WGS), targeted whole-exome sequencing (WES), genome-wide association studies (GWAS), etc), proteomics, metabolomics, gut microbiome, and metagenomics. There is need for exhaustive integration of the existing tools and workflows into an *OpenOmics platform*. Such a platform could incorporate new methods, high-performance computing architectures, modeling tools, workflow design, and data visualization schemes. The platform could facilitate accelerated application of AI-based advances to research across life sciences impacting health, agriculture, synthetic biology etc. The platform could incorporate standardization for data sharing, privacy, ethics, high-performance federated learning for intelligent data parsing, multimodal and/or multiscale information aggregation and integration of health data. The scope could include regulatory genomics, functional genomics, structural genomics, pharmacogenomics, drug discovery, etc to advance understanding of omics and combining omics and traditional healthcare data for improved understanding of human health and disease.

2.2.3 Smart Buildings

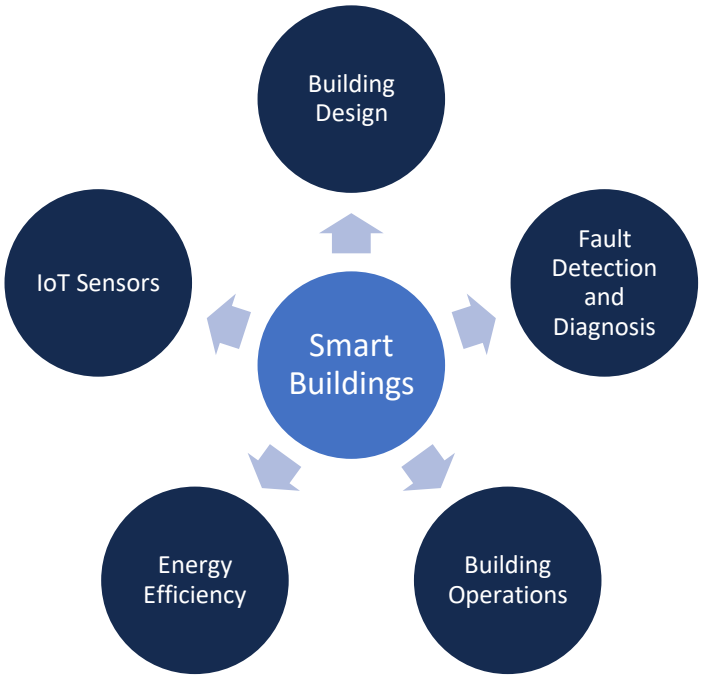
The country is witnessing a rapid growth in the gross built-up area and there is a significant change in the design and operation of buildings over the years. This uncontrolled rapid urbanization and ever-increasing per-capita energy consumption is widening the energy supply and demand gap. Issues such as climate change are adding to the strain of increasing energy consumption. This situation provides an opportunity for the country to leapfrog from the conventional building design and operation to smart buildings. A smart building can optimize the operations of buildings while ensuring comfort to the occupants in the changing outdoor and indoor environment. Adaptation of smart techniques can be enhanced by understanding the local challenges and developing and deploying appropriate solutions with potential scope for replicating in emerging tropical countries.

Design and operation of smart buildings is extremely complex and requires a multi-disciplinary and inter-disciplinary approach. Information technology sits at the center and acts as a critical enabler.

Buildings are an integral part of human life. Be it in an office, home, school, or a shopping mall, people spend most of their time inside them. The primary requirements of a building are to provide safety, comfort and enhance productivity. However, buildings consume a lot of resources such as energy, water, land, air, material, and so on. They also generate a lot of waste and pollution. A common theme in most of the smart building definitions is on the significant use of IT components in making the building smart. Examples of such IT based components include sensor networks, software engines, and network enabled actuators. In a typical smart building, several sensors gather information regarding different parameters related to the building environment such as occupancy, light, temperature, humidity and

radiant heat. Based on this data, software engines can not only learn the current state of the building environment but also anticipate the needs of its user by taking into account user behaviours and usage patterns. This information can then be used to modify different actuators such as air conditioners (AC), lights, and shades to maximize the comfort and safety of the users while minimizing the resource consumption and wastage. In short, the main focus of smart building is towards sustainable resource usage and minimise the side effects with the use of technology, information, and data.

Figure 31: Focus Areas of the Hub in Smart Buildings & Habitats



Three aspects required for enabling smart buildings are design process, operation of building and policy interventions. The following paragraphs give the details of the kind of data required to improve these processes.

Design: At the design stage CAD models are prepared to enable building visualization and construction cost estimation. These models need geometrical information and construction material properties such as visual, thermal, acoustics, embodied energy, availability, maintenance, price, etc. for analysing the performance of the building. For evaluating the building performance with respect to acoustics, fire and structural safety, water consumption, indoor air quality (IAQ), access control, energy, lighting, indoor environment, environmental impact, etc., different models are created and simulated. To have a complete building performance analysis it is required to understand the occupancy pattern, appliance usage pattern, weather data, appliance performance, IAQ requirements. Since the above mentioned data sets are not available, there is a need to prepare these data sets and publish them in a format which is easily compatible with CAD and simulation/analytic tools. It is very important to have the datasets from both vendors/manufacturers and the users on various parameters mentioned above to do the building performance evaluation.

Operations: It is the operations that deliver the design performance of the building. For understanding the building performance, benchmarking is a critical aspect. Benchmarking can be done for energy, water, Indoor Environment Quality (IEQ), space utilization, open area available, etc. Benchmarking can be done in two ways: (i) Benchmarking with similar buildings (ii) Benchmarking with the digital twin. For enabling benchmarking at the country level tremendous amounts of data is required from existing buildings and their performance parameters which are dependent on the type of building, operation hours, days of operation, location, climatic zones, age of the building, etc.

Policy: Policy is an important aspect to enable smart buildings. It can be voluntary or mandatory depending on the type of policy. It is important to estimate the impact of policy on new or existing buildings. Typical reference building models for different categories of buildings need to be created to estimate the energy consumption of different buildings in a city/ country. There is a need in India to create such reference models for commercial and residential buildings. With the availability of reference building models and data of the number of buildings in a region, the total impact of buildings in the region and policy implication can be estimated after simulation and analysis using these models.

Examples of such IT based components include sensor networks, software engines, and network enabled actuators. In a typical smart building, several sensors gather information regarding different parameters related to the building environment such as occupancy, light, temperature, humidity and radiant heat. Based on this data, software engines can not only learn the current state of the building environment but also anticipate the needs of its user by taking into account user behaviours and usage patterns. This information can then be used to modify different actuators such as air conditioners (AC), lights, and shades to maximize the comfort and safety of the users while minimizing the resource consumption and wastage. It is important that these modifications are automated since human interventions may involve significant efforts, cost, and inefficiency. To summarize, the main focus of smart building is towards sustainable resource usage and minimize the side effects with the use of technology, information, and data.

Smart buildings also acquire importance because they are going to be the essential components in the concepts of smart cities and smart grids. In 2015, Government of India has started the smart cities initiative under which 100 cities across the country will be developed as smart cities. One of the enablers of smart cities is smart grids for energy, water, information and so on. The smartness of the cities depends on the smartness of buildings which are nodes on these grids.

2.2.4 Systems

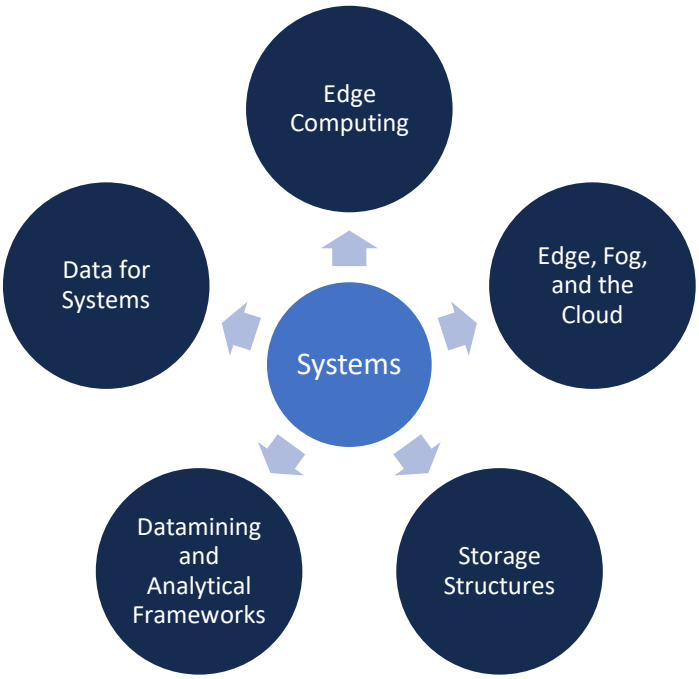
Over the last decade, there is an explosion of data sources resulting in the generation of voluminous data of different varieties at varying velocities. These data sources could be sensors, mobiles, edge devices, vehicles, machines, satellites and humans as summarized in the following table:

Table 2: Sources Contributing to Data Explosion

Source	Examples
World Wide Web	HTML pages, blogs, etc.
Social Media	Facebook, Twitter, Instagram, LinkedIn, etc.
Multi Media	YouTube, Netflix, Prime Video, etc.
Enterprises	Financial transactions, e-mails, etc.
Mobile Apps	Uber, Crowd Sourcing apps, etc.
e-governance	Legal documents, weather, agriculture, etc.
Machine Logs	Data centres, network switches, etc.
Remote Sensing	Geo-spatial maps
Internet of Things	Smart healthcare, smart cities, smart transport, smart grid, smart factories, etc.
Financial & Business Analytics	Stock markets
Block Chain	Block Chain Ledgers, Smart Contracts, etc.

The data generated from these sources may need to be processed partially at the collection end-point (edge), transferred to an intermediate node (fog node) where its further processing and aggregation may happen. Finally, the data will be transferred to the data centre where it will be stored using an appropriate storage structure in a centralized or distributed fashion. Overall, in order to collect, transfer, store and analyse data, the whole systems stack needs to be engineered. All such problem areas fall under the category of Systems for Data Science. On the flip side, if we consider large and medium data centers, where huge amount of data gets generated in the logs, we can do mining and predictive analytics, for performance engineering, pro-active fault detection and fault isolation. Hence, the problem areas can be classified into two main areas. (1) Systems for data and (2) Data for systems. Following are some of the problem areas that will be covered in the Hub.

Figure 32: Focus Areas of the Hub in Systems



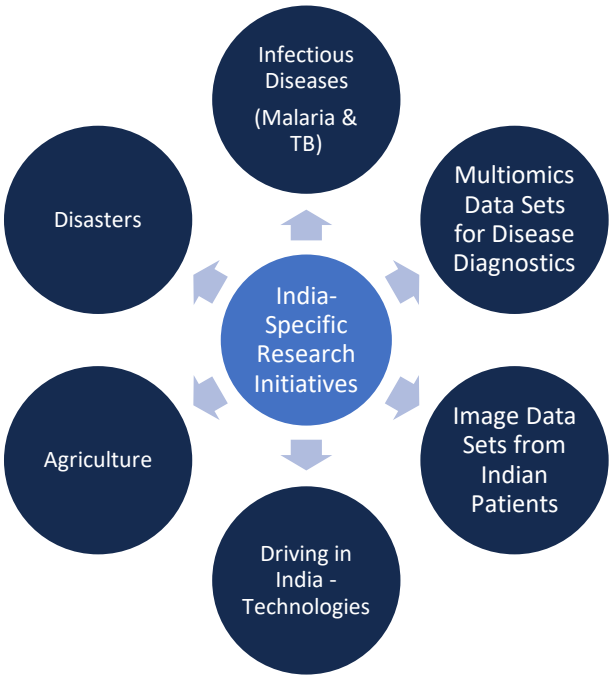
- Edge Computing.
- Edge-Fog-Cloud Optimisation: Data quality, Data aggregation, privacy issues, etc.
- Database Optimisations: Application specific features that enhance the functionality and performance of the database. Time, space, streaming, batch, linked, (Volume, Velocity, Variety and Veracity).
- Distributing Computing Frameworks and their Optimisation: Optimise the frameworks dynamically on a per application basis. Examples: optimise MapReduce node placement, etc. Theoretical guarantees. Performance models. Use of ML/Control theoretic approaches.
- Algorithms for Distributed Frameworks.
- Data Mining and Analytical Frameworks: Data visualization, spatial aggregation vs Temporal aggregation, extracting information from data, information models.
- Data Science for Systems: Predictive vs Diagnostic, behavioural analytics for better SOPs, Automobility of HCI systems.

2.2.5 India-Specific Research Initiatives

It is important for the TIH-Data to work on problems which are relevant at the international level and that resonate with the peers at the national and international levels. It is also equally or more important that the Hub is involved in research and translation activities that will make immediate impact in India specific problems. Data-driven technologies are capable of providing solutions to such issues. This is also of great importance to the Hub activities since there will be limited efforts that will be put by the international community. Some of these areas include infectious disease diagnostics and treatment such as

tuberculosis and malaria, India specific medical data sets (histopathology images, brain fMRI images, genomic data, other omics data, etc.), agriculture, and natural disaster management.

Figure 33: Focus Areas of the Hub in India-Specific Research Initiatives



Agriculture: Data-driven technologies can have a major impact on agriculture. In the last few years, several research papers and prototypes have shown the assistance of AI/ML based methods in aiding in several aspects of agriculture. Some of them include (a) What would be the yield of a given crop? (b) Can we predict the requirement six-months from now to meet the demand and maximize farmer income? (c) Given the difficulties faced by farmers to get/afford daily labour, can we automate certain aspects such as weed identification and extraction? (d) Can we minimize pesticide use and hence maintain soil/crop quality by having autonomous technology to localize pesticide applications? (e) Can we develop technologies that can offer microplanning advice?

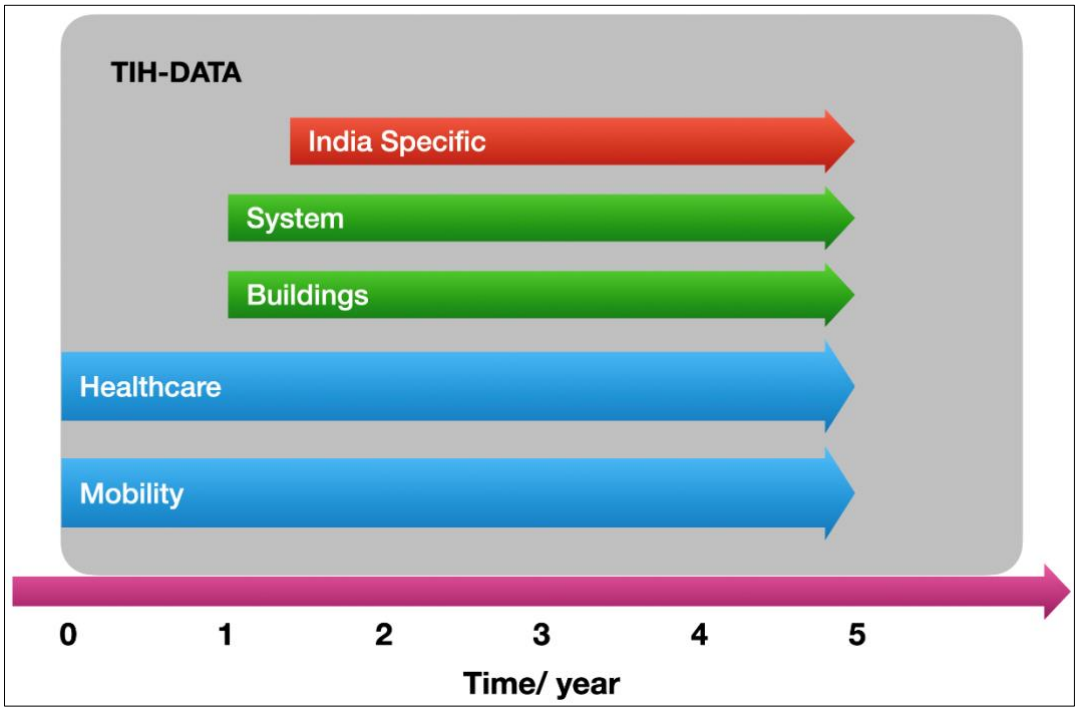
Natural disasters: Natural disasters such as floods, earthquakes and cyclones have a strong negative effect on the livelihoods of everyone. Data-driven technology solutions can be developed for many aspects related to this in terms of prediction of disasters to offering assistance/recovery after disasters. Prediction could help us minimize the damages to human lives and property, and assistance would help us recover to normal faster. (a) Develop data-driven models for disaster prediction? (b) Solutions for rescue (find and rescue human from debris in case of earthquake/cyclone, autonomous vehicles, including airborne, to find life in distress in conditions where the sites are not accessible to the emergency respondents, (c) Technology for rehabilitation of affected population.

India-specific health: Developing treatment and diagnostic protocols for Infectious and vector-borne diseases like TB, malaria, dengue, chikungunya, etc., typically are not of interest to the giant pharmaceutical companies in the west since the disease is more prevalent in developing countries. The Hub will also take up India specific problems in healthcare in areas such as these. (a) Health records from Indian PHCs and hospitals, (b) Prediction of disease outbreaks in advance, (c) Development of clinical decision support systems, (d) Cheap and simple diagnostic methods for these diseases.

In addition, the Hub in consultation with domain/government, will identify important problems in Culture/National Heritage, Environment, Energy, Education, Food/Water Security, Railways, etc., where data-driven technologies will make an impact and work on them.

All activities of TIH-Data in the five domains as explained above will be initiated in a staggered manner. Initially, the Mobility and Healthcare areas will be taken up in the first year. Research & innovation in the other areas will be initiated in the following years (see below).

Figure 34: Domain-Specific Activities of TIH-Data



3 Aims and Objectives

3.1 Overview

The TIH-Data aims to:

1. To be a global leader in research outcomes that also impacts our society with the technologies that are translated to local industries and governmental agencies.
2. To play a central role in development and penetration of data-driven technologies by taking a proactive strategy in curating and creating data banks and data services.
3. To catalyse, nurture, and enable the growth of an ecosystem with researchers, technologists, practitioners and entrepreneurs in the area of data-driven solutions to the local problems.

TIH-Data differs from other Hubs for its mandate to prepare a critical resource for the future use by researchers, startups, and industry. Collecting, collating, and distributing useful data from multiple domains for use by the national and global community will be a significant effort at TIH-Data. This requires not only the expertise of computer scientists and data scientists, but also the cooperation of experts from multiple domains, such as Healthcare, Transportation, Smart Building, Systems, etc. The Data Foundation of the TIH-Data will coordinate the data collection, storage, distribution, processing, etc.

Mobility and Healthcare will be the key initial focus verticals for TIH-Data.

To drive the Hub activities with these larger plans, we find that there has to be at least four different fronts to focus:

- **Research and Technology Development**: India needs a number of technologies to solve the local problems. Our solutions could be very different from that in the west in many dimensions including cost. What also makes these unique is the need for these technologies to work in the local conditions. This requires local data, and technologies that can consume this data. In different timescales, this requires, both basic and applied research.
- **Technology Translation and Working with End-Users Research Should Leave the Lab and Reach the Land**: This is a challenge in Indian context as the academia is ill-prepared to do the translation. India has significant research capacity in most areas, but only a few success stories exist of the research leading directly to technologies that can be taken forward by startups or can be transferred to industry. We propose a set of special strategies to bridge the technology readiness levels on the ground with dedicated productization and engagement strategies.
- **Training Manpower, Improving the Skills and Impacting the Larger Academic Network**: Another broad focus of the Hub is training, and skilling people in the area of data-driven technologies. We support deeply research focused PhD/Postdoc programmes and short courses/workshops for students from wider academic

network to get trained. We propose to connect to an academic network of at least 100 institutions within the country.

- **A Collaborative Research Hub with Minimal Boundaries across Disciplines, Institutions and Countries**: Data science and associated areas are highly active internationally. The Hub will have an intense emphasis on collaborating with international experts in associated areas. This will include joint research, visiting schemes for international experts, involvement in the processes and problems of the Hub, etc. A series of workshops, conferences, and meetings are also planned to bring global experts in direct contact with the Indian researchers and students in associated areas.

3.2 Data Foundation

Data Foundation will deal with the collection, curation, validation, management, and access to data-sets in several chosen domain areas.

a. Establishing the necessary infrastructure/equipment to manage the data storing, annotation and access processes.
b. Data set creation, annotation & hosting.
c. Challenges around data sets to put the data sets to social use.

The Grand vision for the Data Foundation is to become the pre-eminent reference for datasets for AI researchers world-wide.

3.3 Applied Research & Translation

Applied Research & Translation Wing performs translational research and development to take research from the academic groups to make it closer to technology that can be transferred to startups or other corporate entities.

a. Translate the data sets centred research into usable technology. Specifically, the algorithms developed using modern machine learning methods for applications.
b. Identify market needs and opportunities for the resets.
c. Develop prototypes to validate the market need and the feasibility of translating the research into solutions.
d. Design contests and grand challenges around the data sets created
e. Create products that are used by customers.
f. License the products and create new startups.
g. Nurture and grow startups that leverage the data sets and research coming thru the TIH-Data.

3.4 Technology Development

The Technology Development activity of the Hub will work towards promoting ground breaking fundamental research to development of technologies with immediate applications. Depending on the level, projects will be funded at three levels: (a) Knowledge generation: state of the art research that can be published in top conferences and journals, and that will have long term benefits, (b) Development of products/prototypes: based on the existing knowledge, projects intending to develop viable products/prototypes that have potential/important (in terms of business viability and/or in terms of national importance) will be supported, and (c) Technology that have been demonstrated to be complete as assessed by experts in the domain will be supported for full blown product delivery. Thus, technology development aspect will be handled at the full cycle of discovery to technology/product delivery.

The overall objectives of the 'Technology Development' aspects of the Hub are:

a. To provide a platform and necessary resources for knowledge generation and technology development via funding of projects.
b. To identify and support projects related to development of new products/prototypes from already existing knowledge.
c. To identify already existing proof of concepts/prototypes and support the process of technology/product delivery.
d. To provide opportunity to researchers from tier-II and tier-III institutes/universities to involve in high quality research via consortium-based projects.
e. To identify partners and research groups to conduct research in areas where there is specific needs from the industry/society.
f. To support solutions based on proof of contents and demos for further translation and implementations.
g. To fund projects that are of different levels of maturity that will enable knowledge generation all the way to immediate applications in industry.

The sub-research areas under the five domains for 'call for proposals' will be identified after consultation with domain experts and after analysis. Decisions will consider various aspects industry needs, societal needs, needs of the government, etc. Decisions will be taken keeping in mind fairness, transparency, sensitivity towards weaker sections of the society, encouraging next generation researchers and those from tier-ii and iii level institutions and universities.

3.5 HR & Skill Development

Development of technology and knowledge generation in any areas of research involves strengthen the number of researchers at all levels and to empower the next generation with necessary skill set. This could be next generation scientists/engineers/entrepreneurs or skilled workforce. The hub will undertake necessary steps and activities in HRD and skill development that will continue to engage with us to make the initiative highly successful.

i. **Undergraduate research programmes**: IIIT Hyderabad has a strong UG research culture where our BTech students are involved in state-of-the-art research and are part of papers published in many top-level conferences/journals. The host institute also has a strong internship programmes that trains undergraduate students from other institutes. The Hub will strengthen efforts in this direction in the data and domain areas.
ii. **Masters fellowships**: Students who are pursuing postgraduate degree and who are interested in working in the area of data-driven technologies will be encouraged to take up well planned research activities and will be supported.
iii. **PhD fellowships**: Generously supporting PhD students goes a long way in making these researchers a force to reckon with in the future. The doctoral fellowships offered will encourage PhD students to make important contributions and they form an important source of leaders for the future.
iv. **Postdoc fellowships**: The Hub will also support postdoctoral fellowships who will work with senior members of the Hub and institute on problems relevant for technology development in the Hub.
v. **Tenure track faculty fellowships**: These fellowships will be provided to motivated and independent researchers that will allow them to jump start their career quickly with necessary support.
vi. **Chair professorships**: The professorships will enable senior faculty members to dedicate more time and effort to provide direction, undertake research activities and to train young researchers in the areas of the Hub.
vii. **Summer/Winter Schools**: The Hub also will conduct summer/winter schools, long term teaching and training (hybrid: physical and online), and programmes for professionals such that training can be provided to a much wider audience than what is already done directly. Several such programmes tailored towards specific groups of students/teachers/researchers/working professionals will be conducted by the Hub from time to time.

All the fellowships and opportunity to programmes organised by the Hub will be advertised in advance and applications will be invited periodically. Selection will be made in an unbiased manner following transparency and at the same time being sensitive to give chances to underrepresented regions/community/gender.

3.6 Innovation, Entrepreneurship and Startup Ecosystem

Research translated into solutions that help society, commercially or socially, is a key aspiration of The HUB. Creating solutions that can be used in the markets and society demands a very structured approach towards understanding the market/societal needs and fitting the research into a solution to fill those needs. Innovation will happens through the creative thinking to identify the problems and needs a certain research can address, which in most cases is not obvious. And then entrepreneurship will be the avenue to take this solution to market in a gainful and efficient manner.

Key objectives to target around innovation & entrepreneurship will be:

1. **Provide research advisory on market needs**: Tracks relevant technology domains and their application areas, looking for opportunities that current research can address, and capturing and sharing these with the research groups. Research groups to consider tweaking their research towards possible market needs.
2. **Develop proof of concepts (POCs) of product ideas (post research)**: Based on the understanding of market needs that a particular research can address, explore the markets through initial prototypes- a Proof of Concept of the product idea. To understand the solution requirements, and work with the research groups to create a basic prototype of the product. That can be used to better understand the solution needed, feasibility of building it, and validate the market for this solution
3. **Create market prototypes through the EIR programme**: Create market versions of the prototype, once the need is validated. A version that can be used by potential customers to understand the value the product brings. This is done by the institute along with a potential entrepreneur, through the Entrepreneur in Residence (EIR) programme.
4. **Seed new startups**: With market validation in place, the startup will be formally initiated: incorporate the startup, execute the technology licensing agreements and secure initial customers or seed funding for the startup. The research group may choose to be a part of the startup along with the entrepreneur.
5. **Nurture and grow startups**: The startups coming out of, or with the support from, research will be taken into the incubator. The needed support will be made available. From operations to market access, strategy guidance, market creation and raising growth capital.

3.7 Nurturing the Ecosystem

As a technology innovation hub, our project aims to:

- Setup and nurture an academic ecosystem that drives and promotes the data-driven technologies and data banks.
- Catalyse and accelerate an industrial ecosystem that can use data and solve problems that are locally relevant.

We keep the above two aims in mind while working out the detailed plans of the Hub.

Academic Ecosystem

There are a number of aspects that are to be kept in mind to encourage and enable growth of an academic ecosystem. We ask the following questions:

1. How do we establish a continuous and reliable stream of people into the research and development of data-driven technology?
2. How do we encourage collaboration among researchers and bring in fresh ideas and out of the box thinking?

Objectives

We set the following objectives:

- Set up fellowships (PG, PhD, Postdoc) to bring in more trained manpower into the system.
- Encourage research and technology projects, nurture the ideas with continuous feedbacks, fund associated expenses.
- Enable a rich and open collaborative forum and lab that hosts visiting students, researchers.
- Also enable joint and collaborative technology development.
- Provide simple process and mechanisms and model to take the technology to products and startups.
- Expand the academic network with regular research collaboration programmes with a wide range of academic institutions within India.

Industrial Ecosystem

To nurture this, we need efforts and practices that are aligned with the local models. We ask the following questions:

1. How do we package and convert the academic research into solutions/APIs and model that world is looking for?
2. How the relationship and engagement models differ between an industry pilot and a pilot in a government setting?

Objectives

We set the following objectives:

1. Setup databanks and enable data-driven services for researcher, public and local industries inappropriate areas.
1. Setup interaction forums, events and dedicated programmes to develop this community.
2. Setup Systems, Expertise, Policies and Practices that can enable translation, productization, piloting and deployment of the technologies and solution.
3. Encourage more people and entrepreneurs to be in this space and solve local and important problems.

3.8 Outputs/Deliverables

Table 3: TIH-Data Outputs and Deliverables

Components	Activity	Minimum Target
Technology Development	No. of Technologies (IP, Licensing, Patents, etc.)	30
	Technology Products	20
	Publications, IPR and other Intellectual activities	100
	Increase in CPS Research Base	75
Entrepreneurs hip & Start-ups	CPS-Technology Business Incubator (TBI)	1
	CPS-Start-ups & Spin-off companies	38
	CPS-GCC - Grand Challenges and Competitions	4
	CPS-Promotion and Acceleration of Young and Aspiring technology entrepreneurs (CPS-PRAYAS)	1
	CPS-Entrepreneur In Residence (CPS-EIR)	28
	CPS-Dedicated Innovation Accelerator (CPS-DIAL)	1
	CPS-Seed Support System (CPS- SSS)	1
	Job Creation	9550
HRD & Skill Development	Graduate Fellowships	300
	Post-Graduation Fellowships	50
	Doctoral Fellowships	37
	Post-Doctoral Fellowships	10
	Faculty Fellowships	5
	Chair Professors	5
	Skill development	750
International Collaborations	International Collaborations	6

4 Strategies

TIH-Data will have strategies in place to progress in the right direction with focussed activities, which will make it possible to make high impact at the national and international level. An exercise of this big magnitude needs well structure organisation that will absorb large number of activities in the broad spectrum of basic research to product delivery. Secondly, the operational procedures within each of the organisational entities of the TIH-Data should work and contribute towards the overall goals of the Hub. Thirdly, the success of the Hub will also largely depend on how it interfaces with the outside world (researchers from India and abroad, industry, students, startups, society, etc.). It is important to hence consider how to encourage collaborations at different levels to connect with the outside world. This chapter discusses the strategic aspects that will help TIH-Data's success.

4.1 Organisational Design Strategy

Our TIH-Data has done a number of design choices that make the Hub work in a focused manner:

Figure 35: Functional Organization of TIH-Data Activities

Data Foundation: The Data Foundation creates, curates, hosts and provides services centred around the data. Focus of this activity will be on the foundation's interaction with domain experts.

Applied Research and Translation: One of the primary objectives of TIH-Data is to transfer of technology. This effort will focus, nurture and help to mature the knowledge available in the research labs to products and services. This involves (i) technology development effort,

(ii) translation and productization and (iii) interface and management with the industry and agencies.

Research & Technology Development: Another major focus of the Hub is to initiate research and projects that lead to technologies demanded by the society around. A number of short-, medium- and long-term projects will be planned and executed. Formal framework for project selection and management will be aligned with the models followed by other hubs and DST. The budget available for this will be also available for external collaborators and researchers within India.

HRD/Skill Development: TIH-Data also aims to make impact in the academic space with fellowships. We propose to have fellowships for (i) PG students, (ii) PhD students, (iii) Postdocs, (iv) Faculty Fellowships and (v) Faculty chairs.

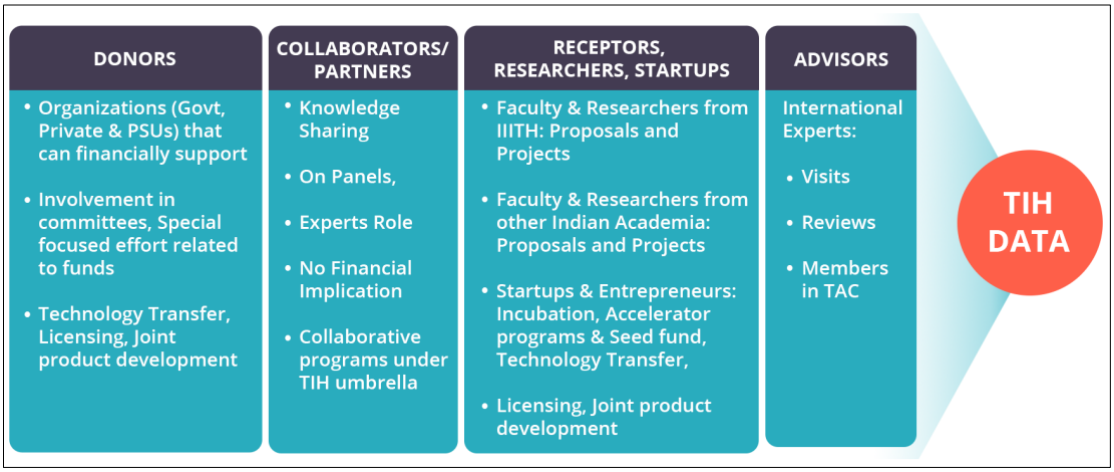
Costing for the fellowship is done based on the guidelines provided in the DPR published by the National Mission on Interdisciplinary Cyber-Physical Systems (NM-ICPS).

Hub Management: TIH-Data, as a Section 8 Company, has its own light weight administrative and managerial team. In addition, there are operational expenses and associated contingencies and capital.

Visitors and Programmes: The unique character of TIH-Data is the ease with which it interacts with experts within and outside the country. We also propose to host a large number of visiting researchers and students in the Hub so that there is strong synergy across research groups.

Innovation and Startups: A route to make impact in the industry is by developing EIRs and incubating startups. TIH-Data has explicit plans in carrying out these through a dedicated effort.

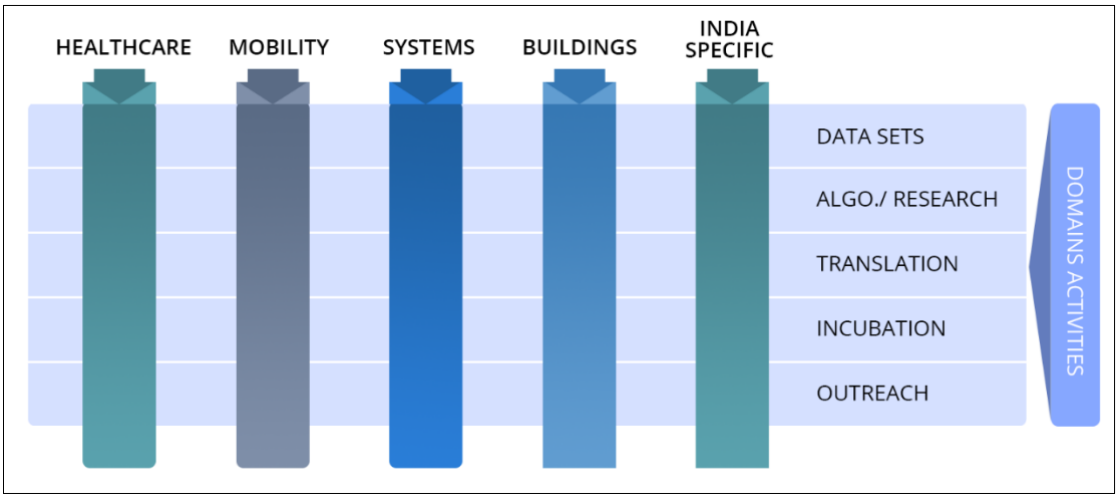
Figure 36: TIH-Data Ecosystem- Interface with External World



4.1.1 Activities and Domains

TIH-Data is engaged in diverse activities related to the theme of data-driven technology.

Figure 37: TIH-Data's Activity-wise Scope

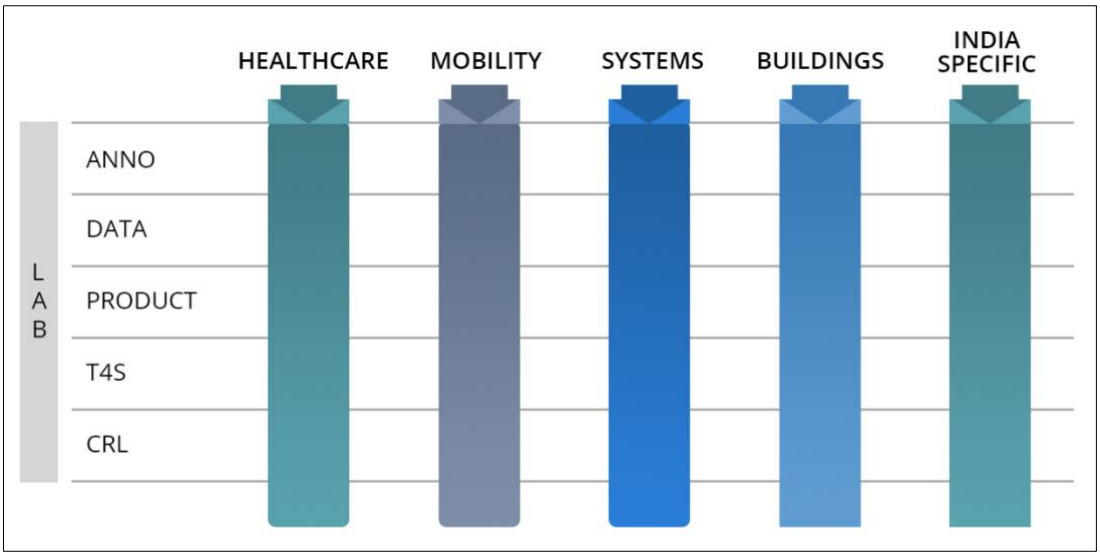


1. **Data Set Creation:** This include creation of the data, working with agencies and groups that have data, curation of the data, annotation and tagging of the data, and so on.
2. **Data Hosting and Services**: Available data is made accessible to the researchers and startups through systematic delivery of data banks and services that can be enabled based on the data.
3. **Algorithms and Systems**: Data-driven technologies consist of a number of AI algorithms. Technologies will be demonstrated by systems and prototypes in most cases.
4. **Translation**: The technology available with the Hub will be made accessible and useful for a wide range of end users through a number of translation efforts. This include productization efforts and setting up a focused effort for technology for society.
5. **Incubation:** From the pilots, technology needs to go further with the help of startups and end users. Incubation and nurturing ecosystem is key for this.
6. **Academic Exchanges**: New ideas and collaborations are key to the success of any research. A vibrant collaborative research programme is planned with participation from national and international experts.
7. **Outreach**: The knowledge generated at the Hub will be made accessible to Tier II and Tier III institutions through systematic programmes. Also, this helps in hub connecting to the technology need at the ground and address them.

4.1.2 Labs and Subsystems

A number of subsystems are planned as part of the activities within the Hub. They are:

Figure 38: Laboratories within the Hub to Improve the Attention to the Specific Goals



- **Annotation Lab** curation and tagging with domain experts. Annotation lab looks into the labour-intensive problems related to data capture,
- **Data Lab** indexing and access (iii) standardisation (iv) interoperability (v) privacy and security (vi) backup Data lab focuses on technology issues related to lab in terms of (i) representation and (ii) maintainability and so on.
- **Product Lab**: Product lab is a key missing component in today's technology translation framework. Additional engineering Product lab converts the technology into demo systems, pilots and services. These efforts are often required to successfully take the technology forward.
- **T4S Lab Technology**: Users could be society who may not demand the need. Also, local and line ministries are often the users of the technology. The interface models with these entities will have to be different from that of interfacing with startups and industries. This lab focuses on these aspects.
- **Collaboration Hub**: A vibrant research space of the Hub is the collaboration hub. This space is filled with like-minded national and global researchers. Simple and attractive models for visits and collaboration will be set up.

4.1.3 Notion and Role of Clusters

The area of interest of the Hub is very broad and we would like to focus on a set of impactful problems. This led to the following strategy:

Figure 39:Technology Development in Focused Areas with Demand and Need

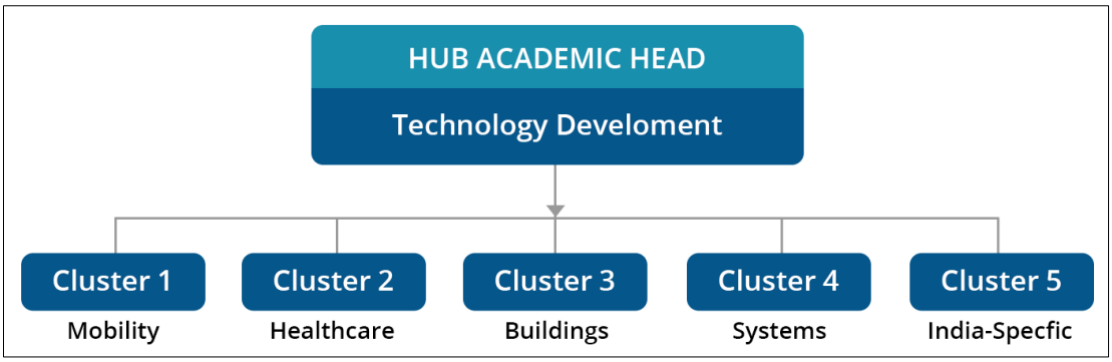
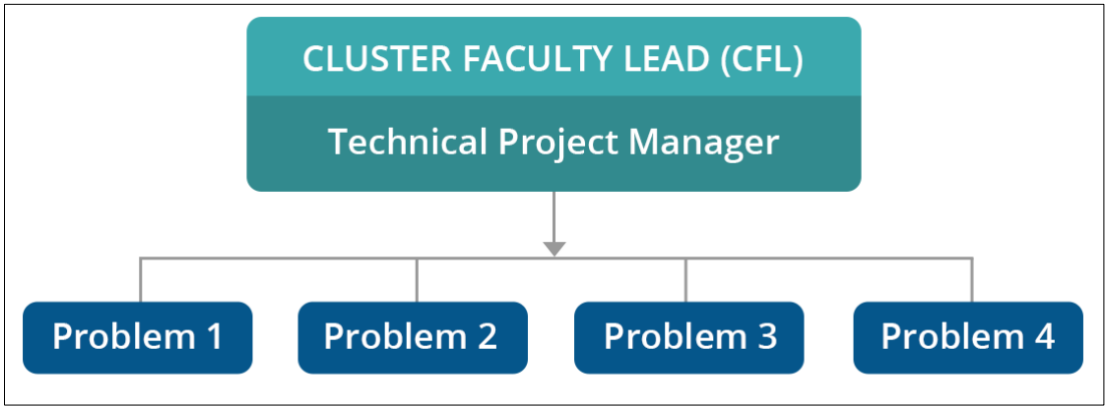


Figure 40: A Cluster Working on 4 Problems



The salient points of the strategy are:

- **Scoping**: Technology development is driven by the need of the technology and also the impact that it will have on society. We consider:
    - The need and impact in Indian space.
    - Expertise available in the team.
    - The readiness of the area to absorb the results.
    - Feasibility of the solution.
    - Possibility of complementing and cooperating with other hubs.
- **Cluster**: A problem cluster is defined first to limit the scope of the projects and activities. An example is *Road Safety*. This itself could have many potential problems.
- **Cluster Lead**: A lead faculty or researcher will coordinate the research and technology development in this space.
- **Problem**: A specific research problem that needs attention is defined with the possibility of inviting multiple proposals and if required awarding multiple projects. For example, "*how to improve the safety of two-wheeler*?" is a specific problem that can have multiple views and angles to think and investigate. One could look at this as a problem under certain weather conditions. One can approach this from the point

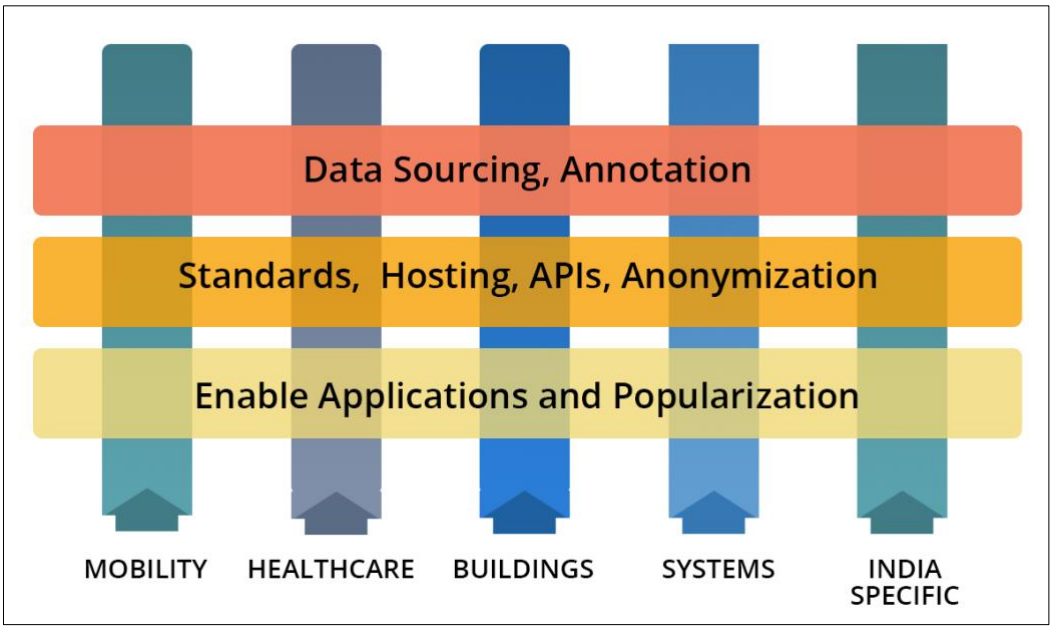
of view of a low-cost driver assistance system. Both may be valid and technically and strategically sound research directions.

- **TPM**: Technology Project Manager (TPM) be the single point contact for the clusters to the Data Foundation and the technology translation.

4.1.4 Scoping the Projects

TIH-Data supports projects and technology development exercise at different stages.

Figure 41: Projects and Technology Development at Various Stages



- **Research Problems with Future Prospects**: Focus of the Hub is on developing technology. We would like focus on some of the deep technological problems with potential deep impact that can critically help in the self-sustenance of the Hub.
- **Technology Development with Potential Immediate Impact**: A number of socially relevant problems are identified within the scope of the Hub. Many of them lead to technology and technology demonstrations (in laboratory setting).
- **Problems with immediate users**: We propose to support academic research groups to work on specific technology development problem with immediate use case. Funding in this space will be small amount and fast processing time.

We propose to keep a right balance between the three models in identification of the specific projects. We also leave possibility of doing:

- **Burst Projects**: Short burst projects that can demonstrate the feasibility of some impressive results. Typical duration for such projects will be six to eight months. It is expected that the PI has some initial results and starting points. Very close project monitoring is planned. Such project proposals are decided and supported also within a month of receiving proposals.
- **Short Duration Projects**: We also propose initiate projects of 1 to 1.5 Years duration to support clear technology development ideas.

- **Full Regular Projects**: Regular projects of 3-year duration is planned for deep investigations, basic and applied research followed by the technology developments and translation.

4.2 Operation Strategy

The strategies behind the Hub's operation are:

- The Hub will spearhead research in the area of data-driven technologies in the country starting from basic research that will result in high quality publications and have long term returns all the way to providing immediate solutions for society/industry.
- The operating protocols of the Hub would be designed and finetuned to support each of these activities efficiently.
- The philosophy within which the operational procedures and policies will be designed are covered below.

TIH-Data

As mentioned above, the overall direction of the Hub will be by provided by the governing board by design. The day to day administration of the Hub will be managed by the CEO/COO, the program managers and the academic faculty members including the Academic Head. The main activities of the Hub are:

i. Data foundation & Applied research and translation
ii. Technology development,
iii. HRD and skill development,
iv. Innovation and startups, and
v. Visitor programmes.

4.2.1 Data Foundation/Applied Research and Translation

a. **Establishing the necessary infrastructure/equipment**: The Data foundation has to establish necessary hardware and higher manpower to maintain the equipment so that Data can be sourced collected created annotated and stored. The Data Foundation team will constitute a committee comprising of experts in the area of systems and high-performance computing to formulate a plan for establishing a data center. After the committee's recommendations the CEO he will follow the best practices and the institutes/DST's policies in procuring the necessary hardware.
b. **Data set creation, annotation & hosting**: The Data Foundation team has to identify areas and problems based on which data will be collected. In this aspect, it is important to keep considering ethics, privacy and security details of the data. The Data Foundation team will constitute ethics committee for identifying potential risks and weighing them up against the potential benefits of the proposed research. Researchers would be asked to propose to the Data Foundation team for potential plans of data collections which will be scrutinized by an evaluation committee and by

the ethics committee before the data collection exercise can start. Standard practices of anonymization of data, taking consents whenever human subjects are involved, confidentiality, incentivising the participants, etc., will be followed. These become extremely crucial especially when dealing with medical data sets such as electronica health records. Domain experts such as medical practitioners, annotators, etc., will be hired to curate, annotate and clean up the data.

c. **Applied research**: Based on the data collected by the Data Foundation (above), based on data collected by collaborators/partners and based on publicly available data, one needs to come up with algorithms developed using modern machine learning methods for applications. Prototypes and systems will be developed as proof of concept which my further be taken out for translation. The data developed as part of the exercises of the Data Foundation may also be used to design contests and grand challenges so that it can be popularized and people around the globe can work on it.
d. **Translation**: Significant effort will be undertaken in translating the research done within the Hub, by collaborators/partners, and by investigators as part of other projects sponsored by the Hub. The necessary framework such as committees for evaluation, admin staff and technical manpower is being planned for successful translation. IIIT Hyderabad also has a product lab which will play a large role in translation.

4.2.2 Technology Development

The technology development activities will be mostly centred around sponsoring research projects in the area of data-driven technologies. In this activity, the Hub will put in place an operational mechanism by which activities of calling for proposals, evaluating them, sponsoring them, conducting monitoring workshops, and so on.

a. **Call for proposals**: The Hub will constitute expert committee for each subject area to decide on the problem areas and each of the application domains for sponsoring projects. The hub will reach out to the larger academic and industry community calling for proposals from time to time in specific problem areas.
b. **Evaluation and decision making**: The committee consisting of experts from members within the Hub and experts from outside will scrutinize the proposals. Best practices will be followed to select deserving proposals for funding.
c. **Monitoring workshops**: After funding, workshops will be held in the Hub for the investigators to come and present their work. This will provide an opportunity for both the investigators to get feedback on for the Hub to assess the progress of the projects. Additionally, the Hub can also identify projects in which translation is possible. Investigators of such projects will be encouraged to further continue their work for applications in society/industry.
d. **Translational research**: Those projects that have been successfully completed and have results that are amenable for translation will be taken up for further

development. The hub will provide the necessary support for the investigators in this process.

e. **Infrastructural facilities**: All the projects sponsored by the Hub will be supported financially to cover the recurring and non-recurring costs of the project. However, the investigators may need to access high-performance computing or may need facilities to host their data sets. The hub will facilitate the use of the computing infrastructure developed as part of its activities.
f. **Weaker section of the society**: The Hub will take steps to reach out to the weaker section of the society. When the projects are evaluated and sponsored, system will be put in place to make sure that the scheme is sensitive to how the applications from women, socio-economically disadvantaged applicants.
g. **Encouraging Young researchers and those from tier-II/III Institutes**: One of the ways the Hub will expand its reach to wider communities is by encouraging investigators from tier-II and tier-III institutions and universities. Additionally, and researchers will be encouraged to apply for different schemes for sponsored research within the Hub. The hub also will reach out to such investigators for continued support during the tenure of the grant.

4.2.3 HR and Skill Development

The Hub will be actively involved in HRD and skill development by offering fellowships at multiple levels. This is expected to encourage students and Young researchers to work in this area of research and gain expertise. The hub will offer these fellowships at the postgraduate PhD and postdoctoral levels to be mentored by scientists and entrepreneurs in the area of data-driven technologies. At the senior level, faculty fellowships and chair professorships will be offered do motivated researchers to encourage their contribution to the Hub in terms of their expertise and time. Committees will be set up to select candidates for each of these fellowship types.

Another major focus in HRD/skill development is conducting symposia and summer/winter schools. Currently, we organise regular summers schools in research areas including AI/ML, computer vision, quantum computing/information, etc. We also train working professionals in the area of AI/ML and blockchain technologies. The Hub will take the experiences from these exercises and conduct events catering to different target groups in the area of data-driven technologies.

4.2.4 Innovation and Startups

One of the major focusses of the Hub is innovation/entrepreneurship/startups:

IIIT Hyderabad is home to the largest academic incubator in India that has been in existence for the past twelve years. The incubator has supported more than hundred startup companies. This incubator, Center for Innovation and Entrepreneurship (CIE), has been supported by DBT and NITI Ayog. IIIT Hyderabad has also played a pivotal role in guide and host T-Hub, one of the largest incubators in India.

The Hub will use the existing setup of the CIE in most of its innovation activities. Some of these activities include grand challenges, startup seeding, market connect, accelerating startups, Data national showcase, etc. Our prior experience in playing a main role in building incubator and nurturing startup culture will help us to successfully develop the Hub activities in this direction.

4.2.5 Visitor Programmes

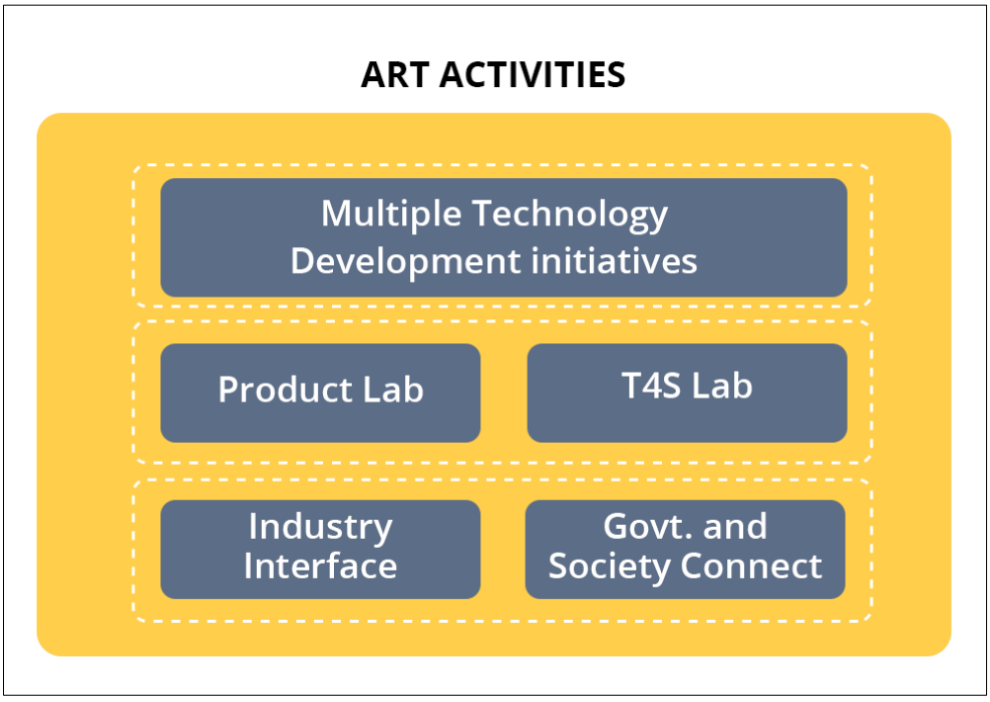
The Hub will encourage visiting researchers at all levels starting from undergraduate students to experts at the highest level. The hub will have mechanisms in place for attracting national and international experts working in the area of data-driven technologies. Necessary arrangements for the visitors to live on campus and a support system to take care of the constant flow of visitors to the Hub will be arranged.

4.3 Collaboration and Growth Strategy

Our TIH-Data has planned a number of specially curated activities that make the Hub work in a focused manner:

4.3.1 Team and Team Focus on Translation

Figure 42: Applied Research & Translation (ART) Activities



There are three major fronts in which the activities are organised:

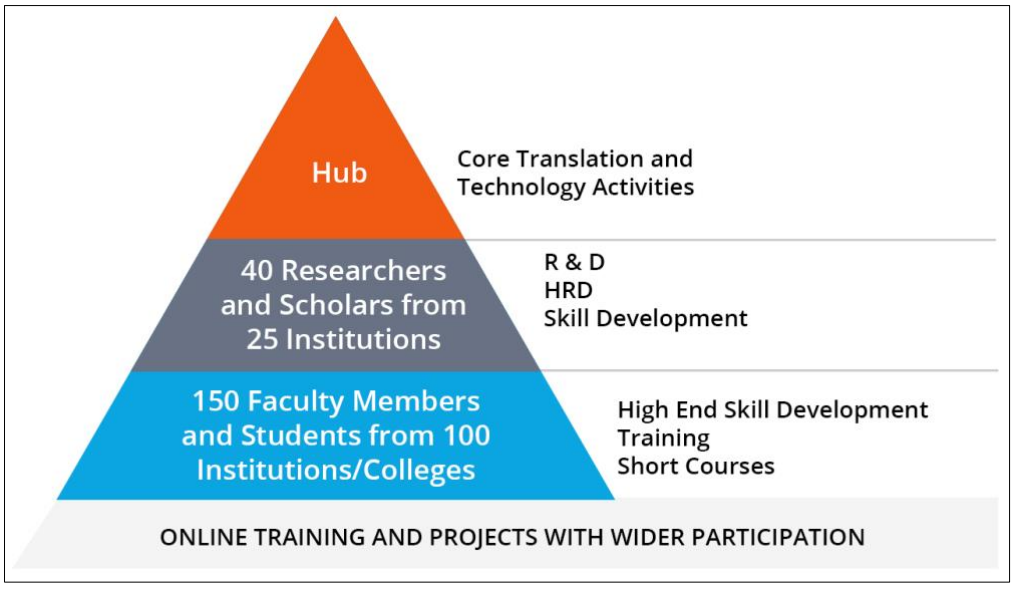
1. **Applied Research and Technology Development**: User requirements are often immediate. The Hub needs to nurture the skill of rapid research prototyping and converting the available results into something that is usable. Focus will be on:
    - User centric applied research

- Encourage system building and demo building exercises

2. **Productization and Industry/Startup Engagement**: Dedicated effort in converting the technology to usable solutions and pilots will be carried out by the product lab. Lab:
    - Creates working prototypes, interfaces and APIs
    - Works with startups, industries and use cases.
    - Provide feedback to the applied researchers and bias the research process to meet the user needs.
3. **Technology for Society and Government Engagement**: Interface with governmental agencies and NGOs are often very different from that with startups or industries. Focus will be on:
    - Work with likeminded NGOs and agencies to impact the local problems.
    - Develop and implement financially viable interface models with Governmental agencies, especially local governments and line ministries.
    - Work with other initiatives (e.g., CSR) of industries, including multinationals, in making impactful contributions.

4.3.2 Academic Network

Figure 43: Hub Proposes to Work with Various Academic Institutions



Hub, which is a Section 8 company hosted and owned by an academic institution has a clear vision to work with a large number of educational institutions, faculty members and students in a systematic manner. We briefly summarize the interface with academic network and ecosystem in the country:

- **Host Institution**: The Hub works very closely with the host institution in:
    - designing dedicated course content, teaching material.
    - running short courses and planning events.

- **Peer Institutions and Peer Researchers**: A set of approximately 25 peer institutions and 40 peer researchers will actively participate in the Research and Development activities. Engagements will involve:
    - Research and Technology Development.
    - HRD, Fellowships and Skill development.
    - Visiting student programmes that enable ease in movement of students across institutions.
    - Joint events, workshops and conferences.
- **Connected Network of Academic Institutions**: Hub proposes to connect to 100 selected institutions/colleges in a long-term manner. Faculty members with research interests that align with the activities of the Hub will be the contact point. Collaborative activities are planned with minimal formal definitions:
    - High end training programmes for students and faculty–Training, events and refresher courses.
    - Attending Short courses and workshops–Student Projects.
    - Specialized hybrid courses.
    - Visiting student programmes.
- **Open Collaborations and Programmes for Students**: Beyond the well curated network of institutions, hub also envisage to initiate training and learning modules in online form. This include:
    - Talks, webinars and online events.
    - Online courses.

4.3.3 Dedicated Hybrid Courses

Hub envisages to work with a set of students and emerging research scholars in a systematic manner by designing a long duration (1 year or more) hybrid (online+offine) programme with different components:

- Online courses.
- Laboratories to be done over the cloud infrastructure.
- Mentored projects with research focus.
- Potential credit transfer mechanism with the universities where ever feasible.
- Visits to the Hub and interaction with the researchers.
- Summer and short duration internships.
- Research projects and intensive engagement for the selected students.

4.3.4 International Collaborative Research

To bring new ideas and energy into the Hub, a vibrant international research collaboration is planned.

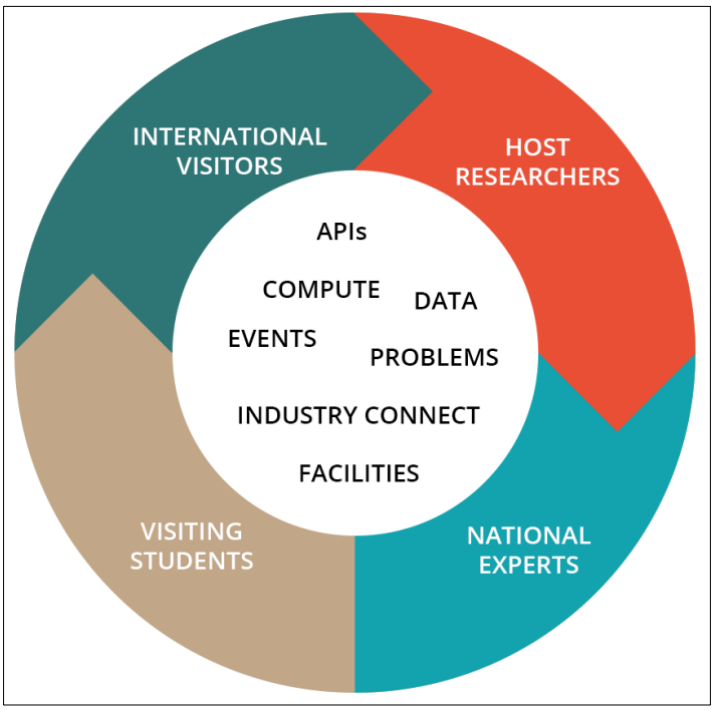
- Enabling joint research between Indian researchers and foreign researchers. Wherever tight collaboration is feasible, complementary support to the Indian

researchers for supporting the PhD students in India will be provided. Some additional research funding is also planned.

- Hosting international researchers in the Hub for sabbatical, to work on special problems such as when the Hub has some unique data or resources.
- Hosting international experts with seminars, review of the projects and programmes, etc.

4.3.5 Collaborative Research Hub for Visiting Researchers

Figure 44:Hub's Framework for Collaboration across Geographies



Success of research is rich collaborations, idea sharing and helping each other in today's world. Our Hub has an explicit lab and model to facilitate this.

The Hub is a rich in experience. A number of things integral to the research is made available to the visitors. This include:

- Data, annotations, APIs.
- Compute, storage and resources.
- Industry connect, use cases, access to the incubation system.
- Edge in research with first timer advantage.

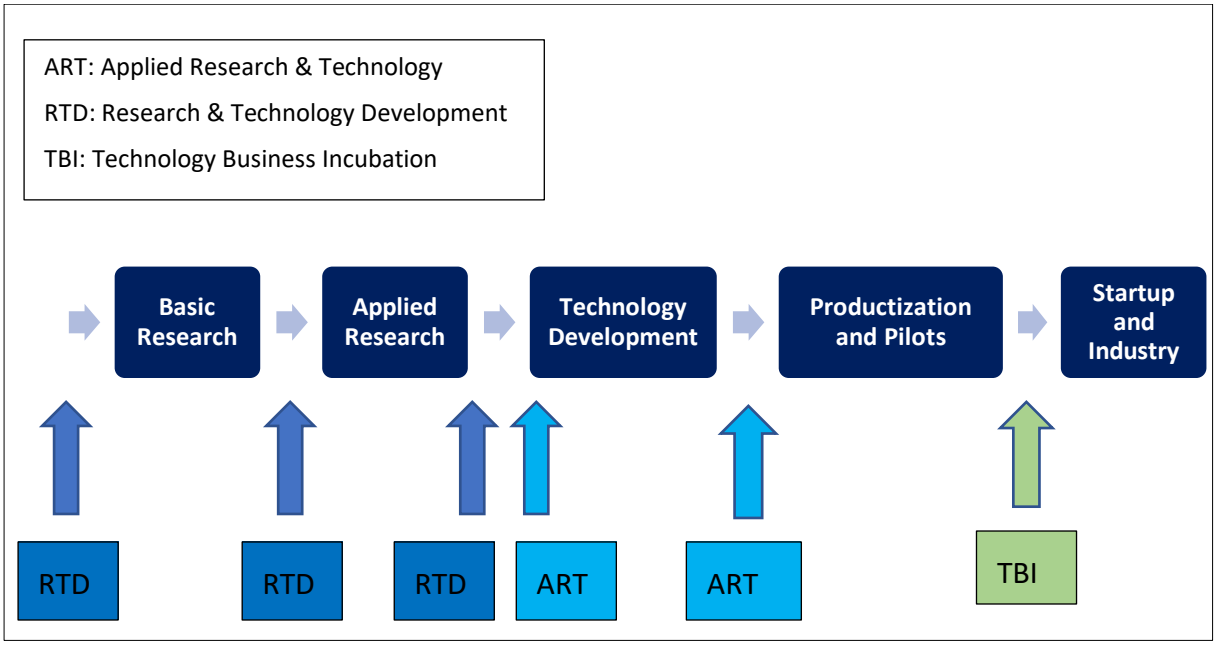
Visitor Researchers and Collaborators: The Hub will design an inviting model for collaboration for working from the Hub and tightly collaborating with the network of peer researchers. Hub envisages to take this catalyst role.

- International experts, NRIs, Indian PhD students studying abroad, participants of the international collaborative research programme.

- Researchers and Research scholars from the national peer research network.
- Visiting students and project students from wide range of institutions within the country.

4.3.6 Process and Activities

Figure 45: Life Cycle: Knowledge Generation through Product Deployment



A key challenge is how to shorten the length/duration.

The full life-cycle of knowledge development to industrial or societal use case is represented in the figure above, Life Cycle: Knowledge Generation through Product Deployment. In a typical research cycle this could be quite long delay to see the fruits of the research and plan for self-sustainability. Two important questions arise in such a long process cycle:

- Q: How do we cut down the long time line/path and see the technology in the market within the first five year itself?
    - We address this by funding and driving at different levels of technology readiness, as can be seen in the figure.
- Q: How do we establish a reliable pipeline beyond the easy technologies or opportunities?
    - We achieve this by also encouraging basic and applied research with different ranges of technology impact horizon.

Considering this we have carefully planned the funding and inputs to the system at different stages of the process to encourage different types of activities.

Research & Technology Development (RTD) Technology development process encouraged projects of different visibility and prospects. Some of the projects could be with clear

technology path while some others are with future prospects and with certain amount of risks.

Applied Research & Translation (ART): Technology transfer team encourages applied research with clear and user driven problem descriptions. Users have visibility of the research and development process and technology transfer in many such cases could be facilitated with lesser effort.

Technology Business Innovation (TBI): The incubation and innovation ecosystem directly encourages startups and EIRs to take the technology from these and go to the market. Mentoring, funding and connects are provided by the incubation ecosystem.

4.3.7 Programmes

IIIT-TIH propose to initiate a number of programmes that strengthens the Hub and connects to the people and organisations around.

- **International Visitor Programme**: We expect a number of strong international collaborations to be nurtured through this.
    - A minimum of ten (10) international experts will collaborate with TIH leading to papers and visible outcomes.
    - We expect to host twelve (12) or more international experts at TIH. Experts interact with the researchers, provide advises and help in connecting the Hub to wider international initiatives.
    - We expect to have researchers from the Hub also visiting universities and research labs abroad to strengthen the collaborations.
- **National Visitor Programme**: We expect a number of researchers in the country visit and collaborate with TIH in various forms. The programme includes:
    - Twenty-five (25) national experts who work closely with the Hub including specific projects.
    - Thirty-five (35) national experts who will visit briefly, give seminars/lectures, interact with the students and other researchers.
    - Partnerships in co-organizing events such as conferences, symposiums and workshops.
- **Visiting Scholar Programme:**
    - Hub propose to host a number of researcher scholars and students from different universities to work on research problems. This will help in strengthening the research collaboration across institutions.
- **Short Educational Programme:**
    - We plan to organise regular summer and winter schools in related areas for the benefit of students and working professionals.
    - We also plan to organise special short courses in emerging and related areas as part of the Hub activities.

- **Long Term Hybrid Courses for Students**: We propose to design and implement long duration, intensive and immersive courses for students from a wide range of institution. Such courses:
    - Strengthen the foundations with intensive and well-designed course content.
    - Practical and immersive training on visits in the sand-witch programme design.
    - Connection to research and technology development.
- **Semester Long Courses for Working Professionals**: Dedicated courses at IIIT will be opened for working professionals so that the working professionals reskill in the new domains.
- **PG Diploma**: IIIT Hyderabad also consider establishing a PG Diploma programme in the related areas to impart training at wider scale.

4.3.8 Collaboration with Other TIHs

The National Mission on ICPS has a collective objective that is distributed across multiple institutions with each hosting one of the several TIHs planned. A natural collaboration across these TIHs is expected. Further, given that the TIH-Data is chartered with curating and hosting data sets, and enable research and solutions around these data sets, there will be a stronger need to engage with the other TIH Hubs. The engagement can be around identifying data sets they may need, collaborating with them to source and curate the data sets, and also work with them for downstream research and productization of research.

4.3.9 Industry Connect

IIITH already has active collaboration programs with major companies (Intel, Google, Microsoft, TCS, Amazon, Flipkart, etc.). Specifically, Intel India will be a major Industry partner for this effort.

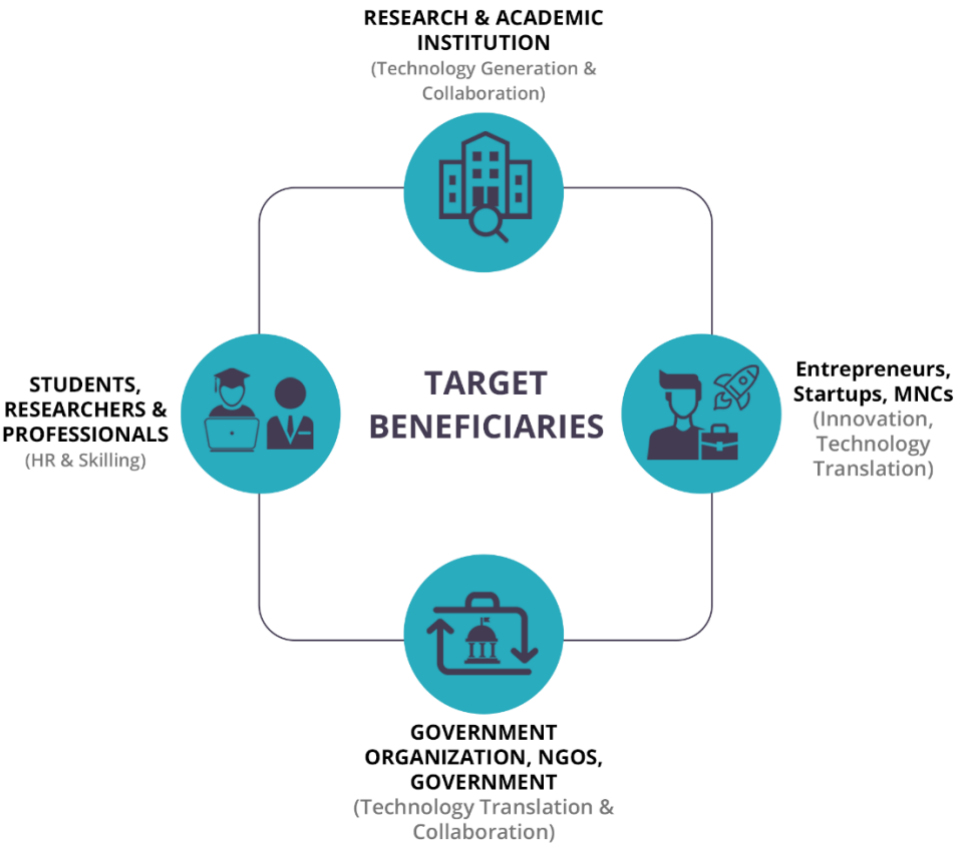
5 Target Beneficiaries

The Hub is multi-dimensional in its framework and has multiple beneficiaries associated at local, national & international levels.

5.1 Target Beneficiaries of TIH-Data

The following are the potential beneficiaries of the Hub's activities:

Figure 46: Target Beneficiaries of TIH-Data



5.1.1 Research Ecosystem - Academic Institutions (National & International)

One of the primary focus of the Hub is to catalyse the research networks within the country and contribute to the research ecosystem at national and international levels.

- Academic Collaborations at International & National levels - One of the activities of the Hub will be to establish reliable and collaborative research connections with national & international experts. This shall be enabled via - (i) Exchange visits of faculty and students between the collaborating institutions, (ii) Advisory Roles for

eminent faculty from India and abroad creating roadmaps for the research ahead, iii) Cooperation in offering joint courses and summer school and (iv) Joint Research and Technology Development between the collaborating institutions.

- A Focused Research Network - A focussed network of researchers working together is clearly a benefit for strengthening the co-operations and activities.
- Resources- This Hub will generate resources (data, APIs, code, etc.) that could help to strengthen the research activities in the country and worldwide.
- Support for Translation- Another benefit will be the common facilities and models for productization and translation.
- Multi-institution projects – The Hub will call for multi-institution projects at national/international level. This will enable brains from different institutions to collaborate, cross-fertilisation of the ideas and contribute towards a common goal and produce research outcomes of international level.

5.1.2 Students, Researchers and Professionals

Another important dimension of the Hub's activities is to establish deep impact through training & upskilling people in associated areas. This initiative will strengthen the Government's Skill India initiative and its outcomes.

- **Training Programmes**: Data-driven technologies need expertise within the country. It being a relatively new area with very less expertise in the country, there is a clear lack of skilled professionals and graduates to meet the requirement of industry. The Hub will establish a strong training programmes of different types enabling interested individuals from academia, corporates to participate. These programmes will be aligned with the educational programmes but with minimal formal connections. This will ensure participation from students, researchers, faculties even from tier-2/3 institutes or cities. Training programme with focus on the faculty will ensure upskilling of the faculty and be able to train in turn the young students & researchers in their respective institutions and local eco-system.
- **Short Courses/Workshops with Internationally Renowned Experts**: In these rapidly changing areas of technology, the Hub will organise regular symposiums, workshops and summer/winter schools to benefit the students in India. A number of international experts will participate in such programmes.
- **Fellowships**: It will benefit researchers at various levels – undergraduate, postgraduate and postdoc- by awarding fellowships in institutions across India. Through these fellowships, the researchers would be able to participate in industry/society relevant research projects. The fellowships would also help increase the research capacity of various institutions across India. The competitive fellowships, PhD, postdoc, UG, PG, will help to nurture the culture of research in this space. It is not common to have postdoc in these areas at this moment and a focused effort will be carried out to increase the statistics.

5.1.3 Technology Beneficiaries

One of the key objectives is to take technology from development to translation into a product/service or commercialisation through various avenues like startups, EIRs, small scale companies or Multinationals. The initiatives listed below will be in sync with the Make in India mission of the nation.

- **Potential Entrepreneurs**: Potential entrepreneurs will benefit with the EIR program. The EIR programme will be focussed in identifying individuals who want to be entrepreneurs and train them in the programme to take off with technology and allied skills. The EIRs will be eligible to pitch for seed-funding to if it decides to take off as a startup and during the EIR tenure, a fixed monthly stipend shall be given to every selected EIR.
- **Startups**: Through the incubator, various initiatives with focus on startups will be implemented. These initiatives will enable technology mentoring, access to data, business mentoring, opportunity to raise fund, connect with investors and market. Startups will be able to engage with the Hub through variety of models made available to work together in specific technology.
- **Small Companies in India**: Small scale industries could engage with hub in terms of specific technology for defined use-cases.
- **Multinationals**: The technical scope of the Hub overlaps with the interests of many multinational companies. Hub will stay focused to establish technical collaborations with at least three multinationals through the Centre of excellence and `other models of engagements.
- **Sectoral Beneficiaries**: The healthcare, mobility, infrastructure, agriculture, etc., will be the domain focus of the Hub and for data, field-trials, use-cases, pilot the technology, through collaboration with government and private organisations in these sectors will be benefitted with the resulting solutions and its implementation.

5.1.4 Data Beneficiaries

Data Beneficiaries: This Hub will generate resources (data, APIs, code, etc.) that could help to expedite the technology development activities in the country and worldwide. The TIH-Data will partner with domain-specific industry, hospitals, PSUs, government organisation & government (local & central) to bring-in data knowledge from various domains and diversified data-sets from across the nation. Given the future-technologies are data-driven, this effort will benefit Government, MNCs, startups, aspiring entrepreneurs, PSUs, etc. IIIT Hyderabad has already existing working relationship with few such organisations and through the activities at TIH-Data, variety of data-sets for specific objectives shall be collected and annotated.

5.2 Cost Sharing and Beneficiary Participation

The focus of the Hub is the promotion of data-driven technologies in the Indian context. This has two parts: (a) create relevant data resources for the use of the wider community of

researchers, developers, startups, decision makers, etc., and (b) develop solutions using machine-learning techniques for several social problem areas using the data so collected.

Machine Learning and Artificial Intelligence are transforming almost all spheres of human activities, from agriculture to high-culture. Data is today descried variously as the new electricity or the new oil, given its potential to fundamentally transform how we do things in the future. The Hub will concentrate in the initial years on creating the required expertise, frameworks, procedures, and processes require to create meaningful and adequate data in several domains of activities, ranging from healthcare to buildings to the society. Data-driven solutions will be beneficial to different areas. The Hub will endeavour to engage different users in the government, society, and industry in developing solutions relevant to them in later years after the Hub's activities stabilize sufficiently.

The Hub will work on socially important problems. It will identify & source socially relevant problem through discussions with Governments, NGOs and local communities. These could be achieved by conducting theme-based conferences, round tables and hackathons. The beneficiaries would include the local citizens, urban & local bodies and government. The Host Institute, IIIT Hyderabad, has strong relationships with the Government of Telangana and has been working with them on a few projects like the Blockchain District, the Year of AI, etc. We are already in discussions with the Departments of Agriculture and Transportation about improving those sectors using data-driven technologies. We will seek and expand similar engagements with other state governments as well as the line ministries of the central government. We have already initiated significant data collection related to the Covid-19 epidemic in the past 3 months by teaming up with the Gandhi Medical College & Hospital in Hyderabad as well as a few CSIR laboratories across the country. We expect to raise significant funding from the engagements with different departments and ministries.

Industry is also in need of expertise in data-driven technologies to stay current in their areas. IIIT Hyderabad has much direct experience in this dealing with multiple industries through sponsored projects, executive education, industry roundtables, etc. The Hub will explore funding from different companies ranging from IT companies to automobile companies to sustain and accelerate the activities of the Hub.

5.3 Impact on Weaker Sections of the Society

The Hub will have multiple of facets of impact on society, particularly the weaker sections of society. The most significant impact will come through the benefits accruing from the innovations, products and solutions coming out of the Hub. Through startups or through solutions built for government and NGOs.

Healthcare is a key focus area, with a key goal being to effect research led innovating cost-effective diagnostics, treatment technologies, public health services' optimization and probably wellness solutions. Innovations that bring down the cost of care, and also significantly increase mass access to good quality care.

India problems in agriculture, rural welfare and livelihood maybe another area where the Hub could make a difference through AI leveraged solutions. Government departments and agencies in the state and centre will be a key avenue for these solutions. Both to identify the problems needing to be solved, and also avenues to take these solutions to the target user groups.

The startups and businesses coming out of the Hub through the entrepreneurial initiatives will further aid society through economic activity and job creation. Through the EIR program, acceleration & incubation program, the Hub will promote inclusive and sustainable industrialisation and foster innovation at multiple levels and domains. A specific emphasis will be the MSME sector. The technological solutions, products and services through MSME will be reaching to the bottom of the society.

The Hub may also provide financial waiver, fellowship, etc., to benefit technology aspirants, students, researchers, entrepreneurs. This will ensure inclusive, quality education and promote lifelong learning opportunities for all in the society.

6 Legal Framework

This chapter describes the legal framework of the TIH. The structure, governance and the various relationship models and IP policies.

6.1 Legal Structure

6.1.1 Section 8 Company

As per the guidelines and support from DST, IIIT-H will set up a Section 8 company. There will be a tripartite agreement between Mission, IIIT-H and Section-8 Company clearly depicting the roles and responsibilities of participating three parties. IIIT-H has already initiated an approval from its governing council for the establishment of the new section-8 company.

6.1.2 Name of the Hub

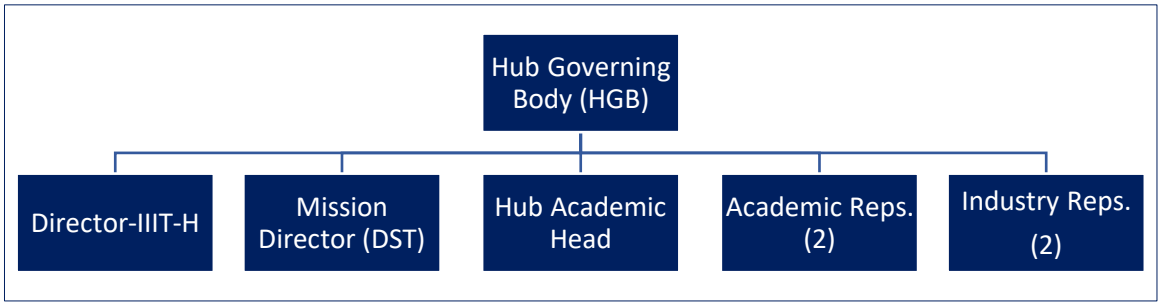
For the naming of the TIH-Data, IIIT Hyderabad shall comply with the nomenclature set by the DST.

6.1.3 Governance and Board Structure

The Hub (Sec 8 company) will be governed by the Hub Governing Board (HGB), and advised by the Technical Advisory Council.

The overall structure of the Hub governance and leadership:

Figure 47: Hub Governing Body: Structure



Hub Governing Body (HGB)

The Hub Governing Body (HGB) is the apex decision making entity of the Hub. It will also be the Board of the Sec 8 company as expected by the Registrar of Companies. It is responsible for the Hub as a whole, sets policy directions, establishes sub-structures, etc., and meets every quarter.

The structure of HGB is as follows:

1. Director, IIIT-H (Chair).
2. Hub Academic Head/Project Director (Member Secretary).
3. Mission Director, NM-ICPS.
4. Academic Representatives (Not less than 2).
5. Industry Representatives (Not less than 2).

The Hub Governing Board will be assisted by a Technical Advisory Council, for domain and technology guidance.

Operationally, the HGB will delegate the executive decisions to an Executive Committee that will include the Academic Head and the CEO/COO of the Sec 8 company.

Technical Advisory Council

The Technical Advisory Council will be a committee of experts constituted by the HGB to guide the board on various technical aspects and structure of various programmes and activities. Will decide on broad academic directions that the Hub undertakes. Can form sub-committees to manage projects, data collection, etc. Meet every two months or as often as necessary.

Executive Committee (EC)

The Executive Committee, to be composed by the HGB, will be an oversight committee to help the executive leadership with day to day matters and providing operational guidance and advise. Working with the executive leadership, will decide on the day-to-day matters of the Hub and Coordinate various activities. This EC will meet every two weeks or as often as required.

Leadership

The executive leadership of the Hub will comprise of two functionaries who will oversee all operations and successful functioning of the programmes.

1. Hub Academic Head (or Project Director): Full responsibility of the Hub as an academic unit. Envisioning and planning of academic activities. Represent the Hub academically outside.
2. CEO/COO: A technical person to lead all activities of the Hub. Plan and execute programmes, calls, drives, etc., as decided. Business development for the Hub.

6.2 Relationship of TIH-Data with the Host Institute (IIIT-H)

TIH-Data is being setup by IIIT-H as the host Institute, there will be a formal operating relationship between TIH-Data and the IIIT-H through an agreement. The key aspects shall be:

- Institutes Director will be the Chairman of the Hub Governing Board. The Hub Academic Head of TIH-Data will be a senior faculty member of the institute. Other

faculty - members may have various roles in operations, committees and the governing board.
- To leverage the experiences of the existing Technology Business Incubator and the Product Labs (Research translation). Maybe even explore models to formally engage.
- Research groups within the Institute will respond to the Call for proposals and undertake research where possible .
- Space and infra will be provided by the Institute. Some maybe as a grant and some maybe for fees as applicable.
- IIIT-H as the host Institute is the sole creator of TIH-Data and in the unlikely eventuality of the Sec 8 company setup for TIH-Data were to shut down, all assets of the TIH-Data will belong to the Institute.

6.3 Governance, Oversight and Reviews

The overall oversight structure will be as below at various levels, from strategy thru to day to day operations.

- Hub Governing Body: Meet every quarter. Define overall policies, strategies, structure, etc.
- Technical Advisory Council: Meet every 2 months. Provide guidance on Technology, Research, etc.
- Executive Committee: Meet bi-weekly. Oversee the Day-to-day operations.
- Hub Academic Head (Project Director): Part-time. As the overall in charge of the Hub.
- CEO/COO: Full time. Executive ownership of all activities and outcomes as the In-charge of the operations.

The overall management structure and entities to assist the Hub Governing Body is described in Chapter 9, Management.

6.4 Partnerships

The Hub by design will work with several partners. Specifically, arounds research collaborations, technology enterprises support, market access and entrepreneurship. Each will be governed by best practices in their respective sectors, captured in the specific agreements and MOUs to be undertaken with the partners. IP related considerations are described in the IP policy section of this document.

6.5 Intellectual Property (IP) Policy

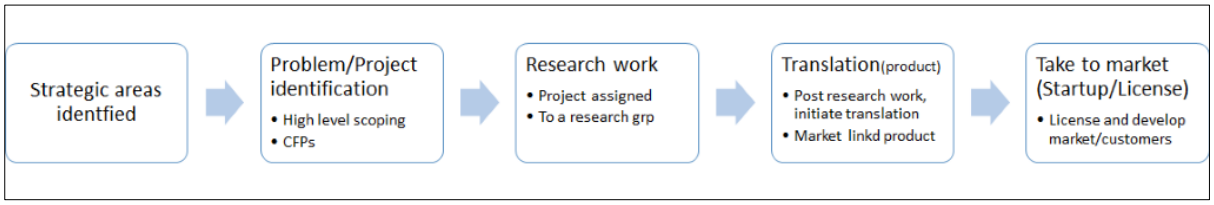
This section describes the Intellectual Property stipulations of the Hub.

6.5.1 What is IP?

Research leads to tangible knowledge that can be leveraged to create products or solutions The Hub will engage in research that will result significant knowledge creation and

technologies that can be leveraged to build the products and solutions. These will broadly constitute the Intellectual Properties being generated by the Hub. The typical stages in the IP creation process are shown below:

Figure 48: Stages in the IP Creation Process



The specific IP artefacts will be, 1) Patents, and 2) Code built based on research.

6.5.2 Stakeholders in IP Creation

The Hub will function through key stakeholders that each have a role the process of creating IP: The research groups from various institutions that undertake the research; the Hub that manages the process of identifying focus areas, calling for proposals, awarding projects, and managing the research process; the Host institute that governs the HGB; the translation group that converts the research into tangible market solutions; and, the startup that licenses the technology and taking it forward as a business.

The role of the various stakeholders in each of the phases of IP creation, from problem definition through creating the startup, is listed in the table below:

Table 4: Role of Stakeholders in IP Creation

Stakeholders	Problem Definition	Research	Translation	Startup
TIH-Data (Sec 8 company)	Overall strategy & directional thinking. & CFP	Track & Oversee progress; dependencies across projects	Initiate the translation project	Approve licensing
IIIT-H (The HI)	Exec Foundational work. And, also respond to CFP	Apart from foundational work, May also take up some of CFP projects	Product Labs could be engaged as a service provider	TTO office could be engaged as a service provider
Research Institute (executing project)	Propose projects against CFP	Execute Projects that have been awarded	Assist in the process of creating product	Assist with initial customer deliberations.
Translation group (in TIH)	Validate Solution possibility and prerequisites needed.	Track progress and start prep work. Identify EIRs	Using engineers and product managers, create the product	Actively handhold. After initial validation, execute licensing agreement
Startup/EIR	NA	If any relevant startups, identify to cocreate (instead of EIR)	Engage the startup/EIR actively	Aggressively validate, and take over the product and its biz

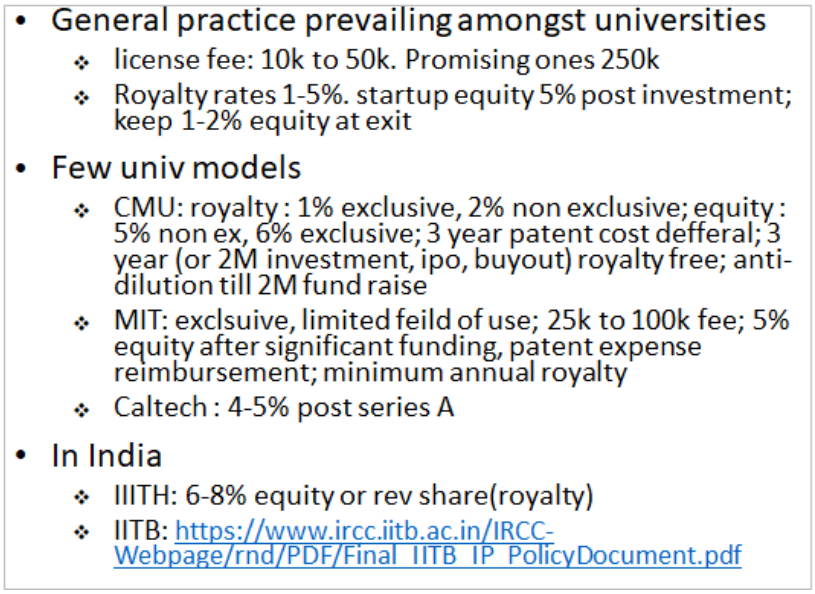
6.5.3 IP Models

The starting point for IP consideration is to define who **owns** the IP generated in the process where there are multiple stakeholders.

6.5.4 Who Owns the IP?

While the broad ownership will rest with the institution that undertakes the research, the project approval will mandate additional conditions on the IP, to apportion suitable ownerships to the other stakeholders, specifically the Hub and the Translation team.

Figure 49: IP Models



The IP licensing models practiced globally include licensing the technology to a company or startup, for equity, revenue share or outright purchase. Further licensing may be done exclusively to one licensee or may be licensed non-exclusively to more than one licensee. The usual equity or royalty models award equity of royalty between 6-10%.

6.5.5 IP Sharing Model for Collaborating Partners

Universities usually keep their licensing stake at 6-10% in equity or royalty payment in the gross revenues of startups. The collaborating partners have a role to play in the process of taking the research into markets.

The activities that each partner plays a role is factored into the overall equity model is captured in the table below:

Table 5: IP Sharing Model

Allocation	Definition	Research	Translation	Mkt Validation	Startup Functional	Equity (%)
TIH-Data	X	X	X	X		2
IIIT-H						-
Research Institute		X	X			6
Translation group (TIH)			X	X	X	2
Faculty (as cofounder)		X	X	X		If, comes from startup's 90%

The faculty involved in the research or translation will have an option to be a part of the startup as either advisor or co-founder. And this will be beyond what the university will get as equity per above table. Specific terms to be agreed up on by the faculty and the startup. The Hub may help facilitate the process.

7 Environmental Impact

The TIH-Data is focussed on Data-Driven Technologies. Its operations are dominantly on data and algorithms that reside in the networked cyberworld and in software. This kind of activities will not have any adverse environmental impact. The Hub will be operating from the existing immovable infrastructure located inside the IIIT-H premise. A space measuring 30,000 square feet has been identified and earmarked for this purpose. Thus, there will be no requirement of new constructions, land acquisitions, dislocations and other environmental clearances.

Consequently, the environmental impact in the establishment of the TIH-Data in IIIT-H is zero and stands as follows:

a. Land acquisitions are not required as this is not a greenfield project.
b. Environmental clearances are not involved as the existing campus of IIIT-H is based on green technologies.
c. Forestry clearances are not required as there is no clearance of forest land or acquisitions are involved.
d. Wildlife clearances are not required as the project is being implemented at the existing infrastructure inside IIIT-H. As a result, there is no direct or indirect impact on wildlife.

8 Technology

This chapter describes the technologies to be worked on in TIH-Data, and how these will be translated into products and solutions. At the base will be the Data Foundation, which will work on sourcing, curating, annotating and hosting the various data sets that will be used by research groups. The ensuing research has several project and solution possibilities in various domains, specifically, in Mobility, Healthcare, Buildings, Systems and India specific initiatives (like Agri). These domains are enumerated in Section 8.2. The targeted products and solutions will come forth only when research is translated into functioning systems. This translation is realised thru the technology and Product Development process described in Section 8.3. The technologies and products so developed, need to be taken to market through startups and other means. The various initiatives and processes around this Innovation & entrepreneurship is discussed in Section 8.4.

8.1 Data Foundation

Several research initiatives can be conceived in the identified vertical domains of (1) Health, (2) Mobility, (3) Buildings, (4) Systems and (5) India specific research initiatives. Such initiatives typically have common requirements:

- To collect, annotate and store valuable data.
- To gain access to data related standards and APIs relevant to the respective domains.
- To host, deploy and promote any further enhancements / annotations to the data and developed applications.

The Data Foundation has three horizontal layers to cater to these aspects, so that these common efforts are isolated and conducted methodically with rigour, thereby serving the needs of a multitude of research initiatives in the above vertical domains, catalysing innovation. The Data Foundation is poised to provide for these common needs across domains.

It will enable the collection of data into a central repository, and also to link to related data in other repositories that cannot be reproduced here. It will include a data annotation layer to perform domain-agnostic simple annotations (such as marking regions in general images and text). Individual research initiatives then only need to plan for and focus on performing annotations that require specific domain expertise (such as in medical documents).

Hosting a central repository of high-quality data, the Data Foundation will enable data access through standardized APIs to governmental and/or research organisations that satisfy suitable criteria which identify them to be stakeholders of that data for national importance.

It will also enable domain experts to define challenges and host it on the platform for registered data scientists to solve. Data released in challenges would consist of data that has been deemed suitable for publication by domain experts and will be suitably anonymised.

The Data Foundation will also serve as a repository to list, link to and publicize solutions and related research initiatives in the respective vertical domains.

In summary, the current proposal is loosely analogous to the Kaggle platform (https://www.kaggle.com), applied on India-centric data sets and targeted at domains of national interest. If developed, this will be a unique national facility.

Most of the activities in the Data Foundation will be:

- Design, develop and deploy a secure, API-accessible databank for effective and efficient storage and retrieval of the volume and complexity of data in the domains of (1) Health, (2) Mobility, (3) Buildings, (4) Systems and (5) India specific research initiatives.
- Enable hosting of data science challenges in the above domains.
- Enable validation and ranking of solutions of challenges on hosted data.

8.2 Projects (Applied Research/Technology Development)

As mentioned above and in chapter 2, the domain areas where the technology development primarily happens are healthcare, mobility, buildings (habitat), systems and India specific research initiatives. Following are some of the specific problem areas within each of these verticals in which applied research, translation and technology development may be undertaken. The annexure A Deep Dive into the Challenges to be Addressed provides an expanded version of the research areas and the list of problem areas in these five fields in which the Hub will undertake focused activities.

8.2.1 Mobility

City Scale Road and Infrastructure

Mapping India has a huge road network, which are managed and maintained by different departments. Proper monitoring of the condition (say the quality of the roads) or state of the traffic infrastructure (e.g., sign boards) is practically impossible to do at scale on a regular, repeatable manner.

Technology to Improve Road Safety and Efficient Mobility

Indian roads witness a number of accidents and deaths every day, whether it is highways, urban or rural roads. Technology can play an important role in making our roads safer. Sensors and cameras connect the physical world to the digital compute. Algorithms and solutions that make use of the sensor data and detect road anomalies such as presence of potholes and ditches, missing traffic signs and signals are the need of the day.

Resource Allocation and Management for Mobility

Our roads and infrastructure utilization can be improved with scientific studies, planning and appropriate technology strategies. This is another area of focus. Several works have looked at this problem from the optimization viewpoint, where an optimal offloading strategy is derived considering various objectives such as total energy consumed in executing an application and time taken to perform the execution. These have been performed under various constraints. The offloading strategy between vehicle, fog and cloud was explored in terms of the stability of LKA and performance. These strategies also perform the task of resource allocation on the fog/cloud architecture in order to offload the computation. Typically, the resources allocated are the virtual machines and the memory resources on the fog and the cloud nodes.

Mobility Solutions for Warehouses

Distribution centers are getting bigger and bigger, especially in e-commerce. Mapping these facilities using vision sensors takes considerable time and computing resources. How can this be done rapidly and efficiently? Warehouse environments are dynamic. There are several workers and other material handling equipment like forklifts operating alongside mobile robots. How can localization and dynamic obstacle avoidance be made more robust to prevent accidents? Modern distribution centers handle very high volumes of products and orders. This would necessitate a large fleet of mobile robots, which in turn could increase congestion and likelihood of deadlocks in certain areas. How can we model this and develop intelligent traffic and fleet management capabilities in robots? Bulk of the work in warehouses involves order fulfilment. Workers manually identify items, pick and pack them. This requires high levels of dexterity and is still incredibly challenging for robots. Current technology is either too slow or too expensive. How can we automate this cost effectively? How can implementation of robotics in warehouses be made simple and fast? How can we improve robot worker interface and cooperation?

Mobility for Surveillance, Search and Rescue

Enhanced mobility systems capable of mapping, surveillance, anomaly detection for campus like areas and wider city scale areas find immense applications. Such mobility systems also include drones and outdoor autonomous vehicles capable of safe driving, passenger and payload transport. Enhanced Mobility Solutions are critical during pandemics, disaster, surveillance, crowd monitoring and many such applications. Such solutions enhance the safety of smart cities. Proposals are sought in such areas of using autonomous/guided agents for delivery, urban infrastructure surveillance, surveillance of campuses and communities. Interesting problems in this space include cooperative payload and delivery by cooperating agents. Cooperation between agents with multiple modes of locomotion for task accomplishment, human in loop systems deserve lot of attention.

Advanced Driver Assistance Systems

Driver assistance systems that can provide both inward and outward looking states find imminent applications. Inward systems estimate driver states while outward systems provide for integrated estimation of on road states including vehicles, pedestrians, state of

the road. Data sets that capture driver states, driver gaze and solutions and algorithms based on such data are sought in this proposal.

One Major objective for the work in the Mobility Domain is to help develop solutions that will contribute towards the VISION ZERO goal of NO life lost to road accidents.

8.2.2 Healthcare

Efficient Disease Diagnostics from Medical Images

Diagnosing diseases is a complex task, is that of an expert who undergoes several years of training and is a time-consuming exercise. AI/ML potentially can process the images and identifies the presence/absence of a disease as accurately as radiologists and pathologists hence making the process more efficient. AI systems for diagnosing diseases based on histopathology and radiology images are required.

Novel Methods for Early Stage Detection of Diseases

The advent of next generation sequencing techniques (NGS) has led to the creation of valuable resources on molecular characterization of diseases. Tumour sample, for example, can be profiled at multiple levels (genomic, epigenomic, transcriptomic and proteomic levels). New methods of diagnosis by integrating multiple data sets will be developed. Some of the diseases that will be addressed are neuropsychotic disorders, autism, cancer and stroke.

Drug Design and Discovery

Drug design and development is a complex process that takes a long time and is expensive. AI/ML methods are capable of cutting down the cost and time by making the search of new chemical entities more efficient. Development of novel and accurate drug design methods and pipelines based on AI/ML algorithms that would make identification of new chemical entities, which strongly inhibits the target proteins, more efficient, thus cutting down time and cost. Development of explainable AI/ML models to predict reliable pharmacokinetic properties that would enable spotting failures early.

AI for Pharmaceuticals

Once the patent of the original drug expires, generic drugs are manufactured which share the same active component of the drug, but are different in other features such as formulation. Given that India hosts top generic pharma companies, the Hub will work with Indian pharmaceutical companies towards developing AI/ML methods for predicting polymorphs and its effects on solubility profiles of new formulations that will help in development of new biosimilars more efficiently.

Wearable Sensors

The future of healthcare is addressing health issues before one even is aware of a condition. Modern AI enabled wearables are pushing the boundaries within the healthcare system in several aspects such as remote health monitoring, remote diagnosis, tracking/improving

fitness, imaging, telemedicine, etc. One of the focus areas of the Hub is development of AI enabled wearables for health monitoring and diagnosis (cardiovascular, diabetic, etc.).

Proactive Public Health:

- Risk profiling & 360-degree view of data.
- Data based targeted preventive measures.
- Predicting outcome of health initiatives.
- Contributory effects of broader Ecosystem.

Health Service Optimisation:

- Ayushman/Arogyasri insurance service optimisation.
- People & Facility use optimizations.
- Supply chain optimizations (medicines).
- Asha workers' efficacies.

One Major objective in the Healthcare Domain is to develop solutions that will help deliver 100% Primary Health Coverage and also Predict onset of Life-style diseases.

8.2.3 Buildings

Data Management: Fusion, Curation, Detection Estimation, Visualisation, Storage, Mining

Smart buildings generate a large amount of data on a regular basis through a large number of sensors for various purposes. To mine potentially useful patterns from this data, various prediction and decision tasks can be modelled as classification and regression problems. It is important to design classifiers and regression algorithms that can be applied in diverse situations with minimal modification such as plug load identification.

- Identify the core prediction and decision tasks of smart buildings that can be automated.
- Design classification and regression algorithms to solve the identified problems.
- Experiment with available options and standard algorithms for the tasks proposed to recommend the best methods for each task.

Securing Wireless Sensor Networks for Smart Building

Wireless Sensor Networks (WSNs) have various potential applications ranging from civilian to military applications including smart building. The sensors are deployed densely in a building and the sensors send important sensing information to the nearby base station(s). In order to protect the sensing information, it is needed to design and implement various security protocols, such as key management, access control, and user authentication protocols. Due to resource limitations and vulnerability to physical capture, these problems become more challenging.

- Identify the protocols that can generate quality and critical feedback on the operation of the secure smart building with the deployment of WSNs in the building.
- Develop various secure protocols such as key management, access control, and user authentication in order to make WSNs secure.
- Theoretical and simulation analysis of the designed protocols, which involves the rigorous security analysis, overheads analysis and functionality features analysis, and simulation through NS2 and AVISPA tool.

Fault Detection, Diagnostics and Prognostics (FDD&P)

Various physical phenomena that affect the building also produce signatures and their impact on the operation of the building. To create the right feedback that can be acted upon either manually or automatically, one can simulate various physical phenomenon which can affect the physical system and the signature of the cause and the effect. Once such a database of cause-and-effect signatures of various physical phenomena are captured, the acquired real-time data can be checked to see if any of the signatures are matching. This can lead to better fault diagnosis and fault localization.

- Identify critical phenomenon that can impact the energy usage of a building. To begin with we will start with the HVAC subsystem.
- Develop a database of signatures of the critical phenomenon.
- Provide querying framework to verify real-time signatures generated.

Water Conservation and Management for Smart Buildings using Wireless Sensor Network

Efficient monitoring and management of water quality with an energy efficient water distribution system is vital for any building. The proposed work will address the optimal water management by considering various sources such as rainfall, surface water, municipal water and treated water and various applications such as water-cooled chillers, evaporative cooling, drinking, washing, irrigation, and so on. The water balance has to be achieved while minimizing total cost of water, treatment, pumping and energy consumption in air conditioning. The present work will adopt wireless sensor network to monitor the water flow rate, leakages, pollution levels and the possible restorative measures at operational level.

- To develop on-line monitoring of water distribution system using wireless sensors network to facilitate efficient management and operation.
- To develop a novel water quality monitoring system for direct, rapid and remote, for detection of contamination in a water distribution system of a building using wireless sensor networks.
- To develop adaptable water distribution network system to bypass faults and prevent contamination.

Health Monitoring of Buildings using High-Precision Building Vibration Sensors and Forecasting of Probable Damage during Earthquakes

At a time when around 56 percent of the land area in India is earthquake prone, it is imperative for India to develop strategies for risk assessment and mitigation because 82

percent of the population is living in 56 percent of the land areas. Risk assessment results in a quantitative index (it does not have any physical significance) that gives a qualitative feel of the level of severity of the problem. The actual process of risk assessment is a detailed exercise. It requires monitoring of building performance during actual earthquake events.

- Identification of dominant building typology in urban area.
- Installation of building vibration sensors. Acquisition of data of ambient vibration and also earthquake events.
- Dissemination of information the moment earthquake occurs.
- Framework for simulating damage to a building after earthquake
- Produce web maps for the use of government and public.

Better Use of Resources

India is still in the infancy of construction of buildings. We expect to see lot more buildings getting constructed in the next two decades. Given that our resources are constrained, we need these buildings to be smart. This requires the tighter integration of AI, IoT and Information technology. Develop technologies that lead to intelligent operation of buildings by efficient and effective use of resources. Resources could include natural resources, energy, etc. We need to meet the potentially contradicting demands of the users, economies and sustainability. AI can help in making the right and optimal decisions automatically and provide intelligent automation by combining signals/observations from different sources and reasoning by meeting the complex optimization requirement.

Buildings Interfacing with Smart Grids

Buildings require external inputs (such as power, water, etc.) on a regular basis. They could be rare resources or could be costly. Smart grids allow the participants to contribute to the demand supply equilibrium. Each player has different objectives to optimize (e.g., comfort vs cost). Buildings could be buying or selling resources to the grids, and have complex objective to optimize and the collective optimization of the performance across a set of buildings (or a city or even country) is far more involved. AI, Multi Agent Systems, Game Theory, Optimization, etc., are critical technologies that will have impact in this space. Many of the decisions to be made here are dynamic and time varying. This naturally forces us to use machines and technology.

8.2.4 Systems

Edge Computing

Many application areas such as healthcare, automotive industry, smart warehouses, smart cities, smart buildings, smart power grid, intelligent transportation, smart agriculture include the use of edge devices for a variety of tasks. For example, in the healthcare, patch recorders and pace makers are example edge devices which continuously sense and record body vitals such as heart rate, blood pressure and ECG signals. Another example of an edge device is a smart water meter which measures the flow of water into a specific area of a distribution network or into a house hold. A strategically placed collection of smart water

meters in a distribution network can help address many critical problems such as identifying the sources of pipeline leaks and any contamination, understand water usage patterns, etc.

These edge devices usually have meagre available compute, memory and storage resources. Mostly, they are powered by batteries and sometimes these are even irreplaceable due to physical constraints, like a pace maker implanted inside a human body. The network connectivity could be intermittent due to lack of signals. Along with these constraints, certain applications demand privacy and require real-time action or response. Hence, transmitting data to a fog node or cloud is not possible. Hence, computation needs to be done and actionable information has to be inferred at the edge node itself. This leads to the following problem areas falling under the broad umbrella area AI/ML at the Edge.

1. Hardware Architectures: Novel processor architectures and special purpose neural accelerators (ASICs) for expediting inference using the deployed deep neural networks (DNNs) at the edge.
2. AI/ML Algorithms on Accelerators: Deployment strategies for fast and low power inference algorithms on devices such as low power CPUs, GPUs and FPGAs.
3. Rethinking AI/ML Algorithms: The state-of-the-art DNNs require enormous amount of computations per inference. DNNs which can give accurate results while performing substantially less computations at a very low precision are required for all edge applications.

Edge, Fog and the Cloud

While some applications computing on the edge, there are many applications where it is profitable to use a intermediary fog node which has more resources than an edge node. For example, in a car there could be tens of Electronic Control Units (ECUs), which sense many parameters. Each of them may not have enough information to infer and take a suitable action. At the same time, it does not make sense for each of the ECUs to stream data to a cloud where all the data streams are analysed in a integrated fashion. This requires lot of unnecessary data transmission as most of the time there may not be actionable information. Even if actionable information is there, the latency will be too high and hence may lead to disastrous consequences. However, we can envision a fog node on the car which collects data from all the ECUs, processes locally, takes necessary action and transmits relevant, aggregated data to the cloud. On the flip side, it is possible for a centralized cloud to control all the edge nodes through the fog nodes in a hierarchical fashion. For example, secure over-the-air software updates are possible through the fog nodes. The following are some major problem areas we envisage in this space.

Storage Structures

Over the two to three decades, due to the expansion of data sources from business processes to world wide web, social media, multi-media and the like, data has become more unstructured and harder to store in traditional relational databases. Also, due to the volume of the data, the underlying storage mechanisms have become inherently distributed. The accesses to these storage structures are highly concurrent as most of front-end applications using them could e-commerce kind of applications. This leads to consistency and availability

issues. In this context, many distributed NoSQL databases and Key-Value stores have been proposed with different consistency, availability and partition tolerance properties. Novel Alongside, novel graph databases such as Neo4j and InfiniteGraph has come up. There are many research problems in this space related benchmarking and performance evaluation, query languages and optimization, database security and privacy, hardware accelerators for databases, graph data management, information retrieval and text mining, scientific databases, semi-structured data, spatio-temporal data, Storage and indexing structures, streaming data and complex event processing, transaction processing, data base optimization using machine learning, optimizing machine learning algorithms using databases and the like.

Distributed Computing Frameworks and Their Optimization

Systems for data science are ubiquitous with examples ranging from domains such as smart cities, energy efficient buildings, social network analytics, and the like. These systems collate, refine, filter, fuse, analyse, and present data and operate on a multitude of devices with wide form factors ranging from handhelds to the clouds. These systems are naturally distributed in nature and hence face all the challenges of distributed systems such as the data partitioning, synchronization, mutual exclusion, failures, checkpointing, and the like. Nevertheless, these systems are expected to work seamlessly under a variety of operating environments while still making optimal and efficient use of resources such as storage, compute, power, and the like.

In light of the above, the following high-level questions emerge in the context of systems for data science.

- Optimize Existing Frameworks: How can applications aim to optimize the behaviour of existing framework to build systems for data science? What are some potential approaches do this optimization? How can one incorporate learning-based techniques to explore the space of optimization in an effective manner?
- Prepare application-specific and architecture-specific frameworks for various systems for data science.
- Analyse algorithms used in systems for data science with respect to parameters such as round complexity, data partitioning and its impact, appropriateness to the under-lying architecture, and the like.
- Identify appropriate performance models that capture the characteristics of the application and the system and compare the measured performance with an ideal performance.
- What are some data storage mechanisms for handling the volume and variety of data that is produced by applications? In particular, how do the existing SQL, and NoSQL based data storage mechanisms such as BigTable, Neo4j, and the like cater to the richness of applications from domains such as the sciences, geospatial informatics, and energy-efficient buildings?

Resource Optimisation

Many of these systems consume a variety of resources and it is necessary to continuously evaluate them for improved performance, develop appropriate metrics, and develop models for incorporating them. For example, in large data centers, there is a need to optimize on power by efficient mapping of jobs to compute resources through analytics.

Data-Driven System Diagnosis

Increased use of sensors/IOTs, software systems and other monitoring mechanisms built-in provide a wealth of data at frequent intervals and/or at critical phases or events both in real-time and as log files. Diagnostic models based on a multitude of these parameters can help build early warning systems, improve alerts, and also help predict failures, if any, in mission critical operations.

Predictive analysis of System behaviour based on its performance analysis across a range of scenarios or operational phases. This might also help detect and predict violations of the Service Level Agreement, degeneration of system, and such other deviations.

Behavioural Analytics of SOPs

For effective functioning of many large complex systems and workflow systems where interactions between/within systems and humans and systems is necessary, standard operating procedures (SOPs) are developed to help assess and predict the behaviour and outcome at each level of such interactions. Post-facto data analysis of such procedures with the information collected can help assess its effectiveness and provide inputs to improve these SOPs.

8.2.5 India-Specific Research Initiatives

In addition to the four domains given above, namely, Mobility, Healthcare, Buildings and Systems, the TIH-Data will also identify important problems that are crucial from the country needs' point of view. Below are some of the indicative problems that the hub may work take it up from the 2nd to 3rd year onwards. Six months before the Hub is ready to take up problems in this area, a brainstorming session involving experts from Academia, Industry, Government and Social Organizations may be arranged to identify specific problem areas with the aim to develop technology solutions that will make a large impact at the grassroots level.

8.2.5.1 Smart Agriculture

Technological advances from Remote Sensing to Drones to IoT devices throw up many opportunities to collect, collate and report on the crop status at different stages and at multiple scales. While a farmer-based system of crop monitoring is for crop diagnosis and crop management, data aggregation and reporting at administrative levels help fine tune policy and enable appropriate interventions in both normal course of agricultural development to disaster mitigation and response. Can Data-driven approaches, across spatial and temporal scales, help assess the Crop suitability? Can patterns or pathways of agricultural development (or decay or stagnation) be traced based on the input and

agricultural management practice data? How can AI/ML based monitoring systems help integrate data across scales for improved Crop performance understanding in a changing environment? Can data-driven models help in early-warning and predicting crop life-cycle?

Many factors drive the crop performance across its life cycle, from providing the right inputs at various stages of the crop to identifying/diagnosing diseases/pests to actual crop yields. While a range of approaches do exist from crop-cutting experiments to surveying, etc., they are too local and the insights they provide are limited. Can new models of performance based on large scale data analysis help provide some key actionable insights? Can it help re-assess farmer incomes based on the cropping patterns across seasons and regions? How does crop suitability fare against crop economics?

8.2.5.2 Garbage Management using Data Analytics

The volume of garbage/trash that is generated by the world is increasing everyday, and India is contributing to this more rapidly with the growing economy. Garbage management is an important problem and technology solutions are expected to mitigate this.

Apart from biodegradable and non- biodegradable wastes, many of the materials in the garbage can be recycled, recovered or reused in whole or parts. Can Life cycle assessments of garbage generated in the country help improve parts of this chain, say in terms of waste reduction? Can such data-driven models provide for improved resource mobilizations and logistical handling? Can a technology driven (use of satellite images, drone imagery, IoTs, Smartphones) approach be taken for improving the data collection mechanisms and location specific garbage dumps. Can data-driven models be developed for improved behavioural response and better compliance to garbage disposal norms?

8.2.5.3 Health Issues/Public Health

Timely prediction of future outbreaks of diseases such as SARS, dengue, chikungunya, etc. is expected to prepare the country and the world to manage other dependent systems effectively. Currently, the effect of the outbreak of covid19 pandemic on human lives, economy/jobs, healthcare system is being witnessed clearly. We also are witnessing the ability of data-driven technologies to bring solutions to a number of problems from simple social distancing, contract tracing to discovery of drugs/vaccine.

Every year around 2.5 million people succumb due to TB and malaria mostly from developing/resource-constrained countries. For business reasons, not much effort go into finding new methods for diagnostics and treatment of these diseases. One of the areas that the Hub will be working on will be in collecting data and develop technologies for these purposes.

8.2.5.4 Disaster Management

The number of disasters that the world is witnessing is on the rise due to several factors with climate change being the primary one. A large population is constantly threatened by disasters such as floods, earthquakes, cyclones, tsunami, etc. Data-driven technologies can

provide solutions to certain problems associated with these disasters. Additionally, the key is to develop solutions that are proactive and not just reactive. Drones and robots today can survey affected areas that are too dangerous for access by humans. Sensors on these autonomous devices can detect life forms in debris where human access is not possible, for example, after an earthquake. Such technology will also minimize the rescue time and allows the rescue workers to maximize their efforts in rescue and minimize their time in searching. Can we develop technology that can assess vulnerability levels of population toward each type of disaster based on multiple factors? For example, if data-driven technologies can predict the most vulnerable area accurately with respect to the expected point hit by a cyclone, it will greatly help in the process of evacuation. In general, data-driven technologies can play a great role during the toughest times of natural disaster, and the Hub may choose problems related to this.

8.2.5.5 Public Safety

Diverse aspects fall into the category of public safety. Some of them include traffic control, policing, criminal investigations, combating terrorism, etc. Data-driven technologies can make a number of processes within these areas very efficient and hence improve public safety in general. Can we develop technologies that can use a large amount of data from CCTV videos and offer assistance in a criminal investigation? Can technology assist in traffic control and in large gatherings in people? For example, in crowds of tens of thousands of people, can data-driven technologies identify those with odd/tense behaviour? Can technology help in preventive policing? Can we develop technology that can help in controlling movement of crowds in religious/political gatherings, sporting events, etc. so that fatal stampedes can be avoided?

Other areas may include food security, railways, Indian culture, Strategic sector, etc.

8.3 Technology and Product Development

The TIH-Data will utilise and promote/proliferate data-driven applications by providing the tools, technologies and resources created by research labs to the startups and industry.

The technologies from labs may be offered in the form of API, Service or Resources as per the agreed technology transfer processes. The startup/industry may use these to offer to citizens applications. The centre will help in integrating and maintaining the services in running condition. This support can be extended to integrators in the form of training, resources availability awareness workshops, Roadshows, related scheme workshops, FAQ, Tech-support, etc.

The centre would drive it’s productization efforts in line with the proposed — AI/ML platform capabilities and specifications and engage in existing Tools and Technologies Enhancement.

The proposed centre will be an applied technology and product engineering entity that will focus on applied research problems on industry, government and social solutions that require language technologies at the core. The centre will be staffed by applied research

scientists and engineers that will focus on research to product translation. The centre will take up startup and corporate innovation initiatives to build products; prototypes of AI/ML research-based solutions.

8.3.1 Objectives of Productization

The major objectives of productization centre are as follows:

- Undertaking technologies enhancement in terms of accuracy, performance and engineering efforts for taking the developed technologies from lab to land.
- Enable cloud-based delivery of AI/ML tools, services for facilitating Startups/software developers.
- Leveraging the emerging technologies like cloud and mobile for reach out to the un-reached sections of the society.
- Creating repository for show casing of tools and technologies to facilitate single window information delivery.
- Innovative Product Designs and engagement with MSME and local markets access mechanisms for bottom of pyramid markets.

8.3.2 Outcomes

The centre intends to deliver on the following outcomes:

- Create market assets: APIs, Services, Solutions from research as per platform plans and guidelines.
- License out the technologies (#Licenses).
- Create language tech startups (#Startups).
- Demonstrate social good (#SocialInnovations).
- Develop solutions for government initiatives (# Solutions).

8.3.3 Approach

The centre would follow a three-step approach for each use case that the helps solve the issues identified through market analysis.

Step 1

Development of core algorithms and technologies relevant to the use case – Research teams working on representative data sets provided.

Step 2

Conducting field trials and pilots in real life scenarios – product engineering labs translating research to application space and working with researchers to refine the technologies.

Step 3

Productization and deployment of robust solution – Product labs and CIE sourcing development and licensing partners (startups, industry or govt agency).

This approach would ensure delivery of tangible results and demonstrate impact in manageable packets. This also help the teams and agencies to deal with changes in environment or priorities or learnings from real life feedback.

8.3.4 Translating Research Deliverable to Deployable Components

The technologies developed as part of this programme need to be validated in real-life situations and be brought to a usable state that a startup or an industry partner can deploy.

The team comprising of Business Analysts, Design Engineers (Human factors) and Product Engineers (referred to as Product Labs) shall receive the research from the labs and translate the same into “application” and “end user” relevant solutions with which the agencies can conduct valuable real-life pilots.

The team also ploughs back to the research teams, the market insights, the user insights by conducting user studies, ethnographic studies, market and business studies so that technologies developed are relevant at all points in time.

8.3.5 Market Connect

A central element of the TIH-Data’s innovation initiative will be to connect markets with research on both ends, markets defining the research needs and translating research through startups. TIH-Data will ensure a strong connect between research and markets. This will be shared with startups and also researchers, to set broad directions.

8.3.6 Researcher Startups

Researcher startups programme will encourage researchers to consider entrepreneurship. This is an active programme to encourage researchers to become entrepreneurs and support them in their journeys to translate their research to create market impact.

8.3.7 Corporate Innovation Co–creation

This is a programme to engage corporate innovation initiatives to leverage startup sand research to cocreate specific innovation corporates need, leveraging the research and knowledge in TIH-Data.

8.4 Innovation, Entrepreneurship and Start-up Ecosystem

8.4.1 A Vibrant Innovation and Startup Ecosystem at IIIT-H

TIH-Data will build on the incumbent startup ecosystem at IIIT-H campus. IIIT-H houses the largest academic incubator in the country, the eleven-year-old Centre for Innovation and Entrepreneurship (CIE).

- CIE has the base DeepTech incubator, which was initially funded by DST & MEITY (in 2009-2015) and is now financially self-sustaining supporting over 100 tech startups. Additionally, in 2018 a MedTec BioNEST funded by DBT was commenced. And end of

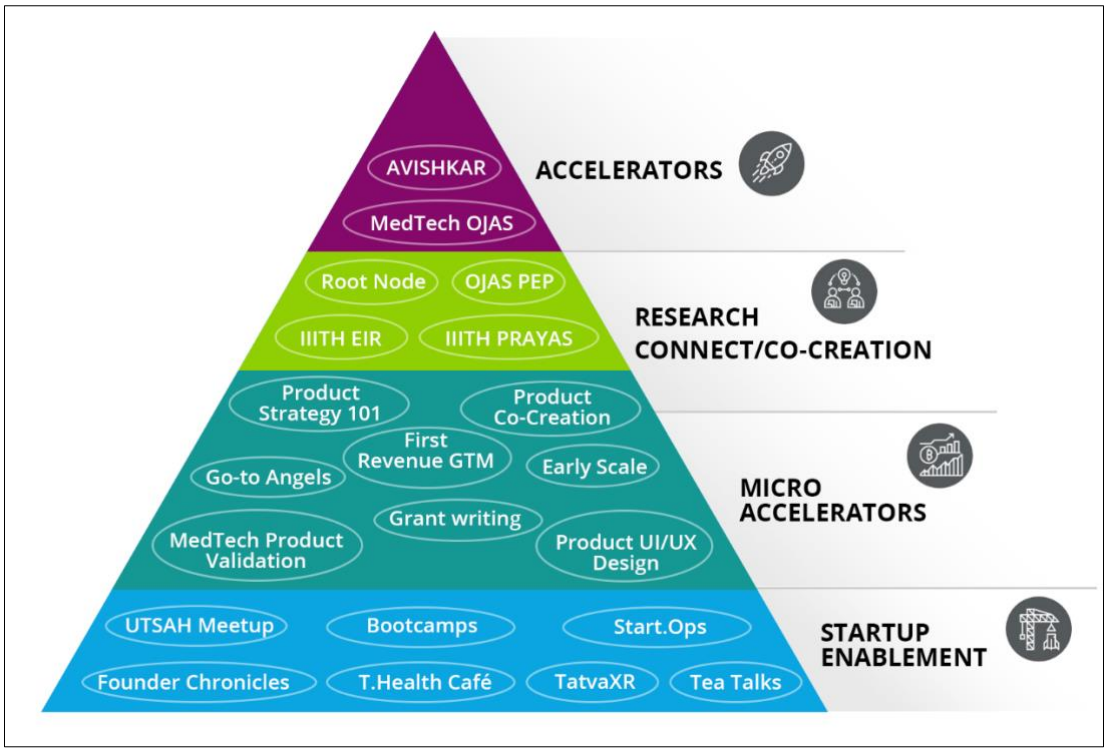
2019, a social incubator (AIC) funded by NITI Ayog was started. In all CIE has capacity to support 150 startups now.

- The primary emphasis at CIE has been around models to create deep IP startups, startups from research, startups based on tech licensed from research, startup s cocreated with research, startups nurtured by research.
- In addition to the normal strategy, business and market mentoring. Apart from incubating 100+ startups, CIE now runs three rich accelerators that spot promising very early stage deep technology startups, connect them with research, and steer them towards growth.
- Being the oldest incubator in the city, CIE has played a strong ecosystem enablement role in the city and state here. Apart from creating strong networks of incubators and enablers, also has actively helped other incubators start in the city.
- Specifically, CIE has worked with the state government to structure, start and host T-Hub, now the largest incubator in the country working with growth stage startups. T-Hub actively works with corporate innovation initiatives, connecting startups with multinational corporations. IIIT-H has also helped the state government to start the State Innovation Cell. All these experiences will be very relevant, for the TIH-Data. IIIT-H will build on all experiences so far to structure a very strong model that will enable startups from the research happening in the TIH hub and spokes:

8.4.2 TIH-Data Startup Initiatives

TIH-Data will support Data intensive Deeptech startups through its incubation and accelerator programmes. The programmes will focus on Deep-Tech startups working in domains of Machine Learning, Computer Vision, Robotics, Natural Language Processing, etc. The research labs of IIIT Hyderabad (Technology Support), investment partners like IIIT-H Seed Fund, strategy advisors like CoCreation Consulting (Startup Coaching & Mentoring) and other key-players of Hyderabad startup ecosystem will be leveraged for the programmes.

Figure 50: TIH-Data Startup Initiatives



Collectively, the incubator will provide support for startups through

- Tech and Business Mentoring
- Large working/ office space for startups
- Funding support through various accelerator programmes
- Industry Connects
- Access to community events
- Demo Days and Pitch days (Investor connect)
- Internet & Cloud Credits
- Branding opportunities

TIH-Data will run some flagship 6-month accelerator programmes in Deeptech and Medtech. Each programme will run two cohorts in a year, expecting over 250 applications across country. Along with these, the centre runs Micro accelerator series, Research connect and Entrepreneurship programmes as explained below:

8.4.3 Product Go-To Market

Upon successful completion of the pilots for each set of use cases it is imperative that these are deployed as robust products, maintained and upgraded from time to time. This can be enabled through startups or industry partners.

To enable the discovery and nurturing of go to market partners, it is required to do the following:

- Scan/scout for reliable partners.
    - Nurture selected partners to be able to successfully deploy the technologies.
    - It is proposed that we conduct proactive outreach programmes to attract entrepreneurs or existing startups or corporates.
- The applies a robust evaluation to ensure relevant partner is selected which CIE at IIIT Hyderabad has been doing for a decade now.
- The partners not only productize the technology from pilot stage, they are also provided support to be successful.
- These partners in some cases may step in right after lab finishes work or after the product lab finishes pilots or may bring in their product directly into the mix to accelerate the programmes. The goal is to make centre self-sustaining after the five-year period in the following manner:

Nurturing startups and capturing the value from Equity shared by startup to the centre (we have a proven model at IIIT CIE developed over a decade that we will leverage).

- Licensing out the market relevant components to large companies.
- Provide services on to of data sets corpus collected to create revenue opportunities especially with large corporate (i.e., custom services on top of data sets we collect).
- Providing Applied technology productization services to different entities. (i.e., we build custom components and products on top of the mature technologies available from research groups as per company requirements.)

Startups Connecting to Market Needs

A central element of the TIH-Data's innovation initiative will be to connect markets with research on both ends, markets defining the research needs and also on translating research through startups. TIH-Data will ensure a strong connect between research and markets. This will be shared with startups and also researchers, to set broad directions.

8.4.4 Startup Seeding

Startup Challenges

TIH-Data will find startups in related areas and enable them to work actively with research. This will address a key gap today wherein it is hard to find startups that are leveraging emerging and deep technologies.

Programmes to Seed New Startups

The incubator will run startup seeding programmes that will enable startups to work actively with research. The incubator will invite and inspire entrepreneurs to create startups out of research happening in TIH-Data. The focus of seeding will be on translating research in to new startups and transferring research knowledge into existing startups.

8.4.5 Research Startup Accelerator

This programme spots and nurtures startups in the Data related areas and actively enables them and connect them with downstream enablers and investors.

Researcher Startups

Researcher startups programme will encourage researchers themselves to also consider entrepreneurship. This is an active programme to encourage researchers to become entrepreneurs and support them in their journeys to translate their research into market impact.

8.4.6 Corporate Innovation Co-Creation

This is a programme to engage corporate innovation initiatives to leverage startup sand research to cocreate specific innovation corporates need, leveraging the research and knowledge in TIH-Data.

8.4.7 TIH-Data Incubators Network and Annual Showcase

Other incubators in the spoke institutions of TIH-Data will be encouraged to join this network. All the above programmes will be available for the startups in those institutions as well.

TIH-Data National Startup Showcase

A large format annual showcase event will be conducted to showcase startups coming out of all the TIH-Data incubators network. Connect the startups with large enterprises, MNC tech companies, investors and more.

8.5 HR and Skill Development

Per the mission mandate, HRD is a pivotal part to harness the full potential of CPS and extend its benefits to a broader group of stakeholders, enabling the humans who will be at work creating new knowledge, converting knowledge into technology, technology into products, diffusing products as a consumer good/service, and also, as knowledge seekers. Development of human resources in a focused area requires infrastructure, learning platforms, knowledge/information resources, teachers, capital, policy framework, etc. Central to all human resource development activities will be research and applying research to create products and solutions.

8.5.1 Research Fellowships

HR and Skill Training includes, (i) postdoc fellowships and other potential short-term fellowships, (ii) organising short courses and educational programmes, and (iii) evaluation of programmes.

8.5.2 Technology & Translation Programmes (for Industry)

The skill development part of the proposed work will include training students, working professionals, domain experts, faculty and entrepreneurs. These training programs will appropriately include lectures on theory, practical considerations and hands-on sessions to make the recipients capable of utilizing the information received for developing practical solutions as well as training others. The specific plans include:

1. *Training domain experts in AI/ML*: The field of AI and ML is a highly applied area. Most effective solutions emerge out of the optimal combination of domain expertise and knowledge of AI/ML technologies. For example, the use of computations for new drug discoveries will include knowledge of sciences, medical considerations as well as AI/ML tool. An important aspect of the success of such a solution is the ability of domain experts (e.g. Science and Health experts) to understand the potentials and limitations of AI/ML technologies. The goal of this training is to impact sufficient background to domain experts so that they can either develop simpler solution on their own or collaborate better with AI/ML and other domain experts in advancing the state of the art.
2. *Training Entrepreneurs and Potential Entrepreneurs*: Entrepreneurs are people who need to develop practical solutions to real problems and do so quickly. To enable this, we will conduct AI/ML training (both empirical science and practices) for such people, where the process of solution development will be given higher importance.
3. *Short Courses for Professionals and Practitioners*: Several of the current jobs in IT and related areas are going to be significantly affected by the advances in tools and solutions that use AI/ML. A Gartner survey has predicted that AI/ML projects are likely to double this year and an estimated 90% of all projects in 2025 will include some component of AI or ML. The training for practitioners and professionals will be designed to enable them to cope with this new landscape of IT solutions.
4. *Teacher Training Programs*: To ensure that the larger set of institutions across India are enabled to train their students in AI/ML, we will conduct teacher training programs for wider impacts in engineering institutions. This will also equip them with material for teaching their students about the theoretical and practical aspects of these technologies.
5. *Training Students and Fresh Graduates*: These training programs will be aimed at enabling the students to develop practical solutions using hand-on session through projects and mini-projects in the area.

9 Management

9.1 Introduction

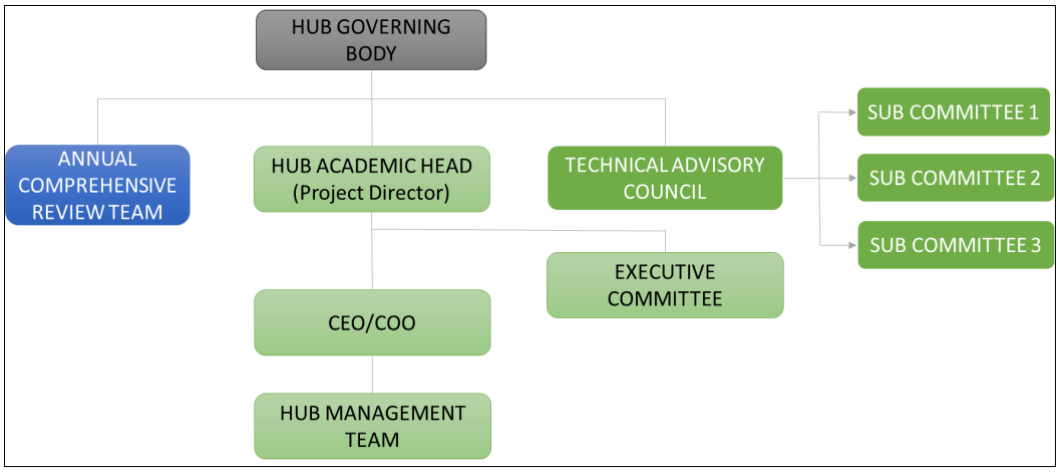
The Technology Innovation Hub is of very high value to IIIT Hyderabad. Since the project is guided by the National Mission on Interdisciplinary Cyber Physical Systems (NM-ICPS) and involves a variety of stakeholders spread across the country, government, industry, startups, academic institutions and international collaboration, it impacts every citizen of the country.

The targets set by the National Mission are ambitious, and thus, it requires the best of the expertise be involved at every level to manage this complex and impactful project. It is proposed to create a Hub Governing Body (HGB) to guide the entire TIH-Data's activities related to technology development, HRD and Skill Development, Entrepreneurship & startups and international collaborations.

9.2 Organisation Summary

At the apex level, Hub Governing Body (HGB) will be the final authority on all activities and compliances related to the Hub. The HGB will be supported by the Executive Committee and the Technical Advisory Council.

Figure 51: Hub Governing Body: Support Structure



9.2.1 Management Plan

Plans for the administration of the Hub with the details of the organisation structure are given below:

9.2.1.1 Hub Governing Body (HGB)

Hub Governing Body (HGB) is the apex governing body and is the statutory Board of Directors of the TIH Section- 8 company. The HGB shall have:

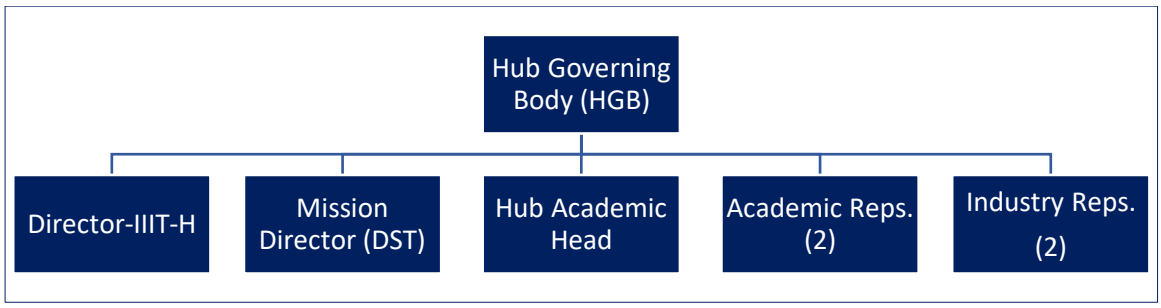
- Full autonomy in devising the internal processes and procedures for achieving the targets and deliverables, subject to the general directions of Mission Governing Body (MGB).
- Full financial and administrative powers including but not limited to:
    - Re-appropriating the budget within the ceiling of the sanctioned budget.
    - Hiring the appropriate human resources as per industry standards.
    - Signing Memorandum of Understanding (MoU) with national and international institutions and industry.
    - Approving foreign visits and collaboration.
    - Partnering with industry.
    - Receiving and award projects in the domain areas of Hubs to academia, R&D institutions, industry and other funding agencies.
- HGB will evolve specific targets and timelines for the Hub in consonance with the MGB directions and monitor the progress and report to MGB.
- HGB meets once every quarter or as often as necessary to take necessary actions/decisions.

The HGB will be assisted by a Technical Advisory Council (TAC) for guidance on strategic directions and an Executive Committee for operational oversight of activities and compliances.

The following structure will be followed to identify the members of HGB:

1. Director, IIIT Hyderabad (Chairman).
2. Mission Director from DST.
3. Hub Academic Head (Project Director) (Member Secretary)
4. Academic Representatives (Not less than 2).
5. Industry Representatives (Not less than 2).

Figure 52: Members of the Hub Governing Body



9.2.1.2 Technical Advisory Council (TAC)

The TAC's expertise will aid the expansion of the TIH beyond local, bring-in perspectives from industry, international academia in structuring the programmes.

The Hub Academic Head shall chair the TAC. The role of TAC is to advice on technology and research focused initiatives.

The role and responsibilities of the TAC include:

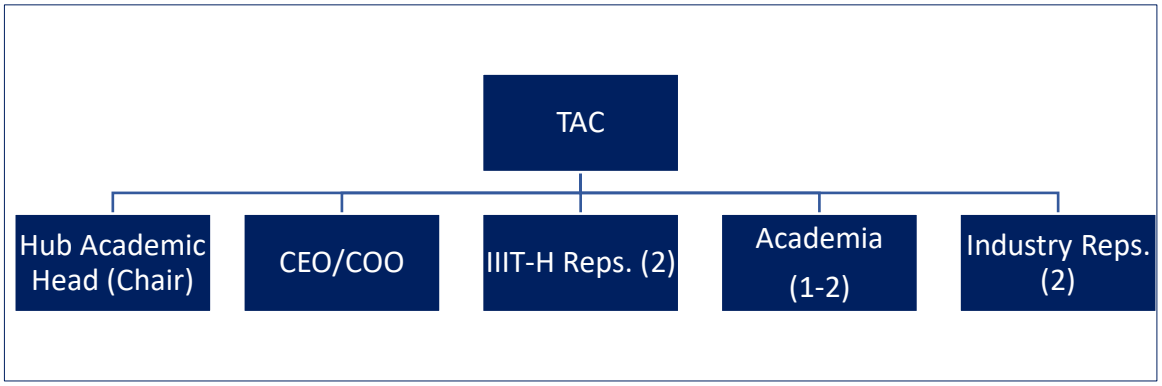
- Providing high-level guidance.
- Providing connects with the external world.
- Raising funds.
- Bringing visibility to TIH-Data.
- Monitoring other activities like outreach.
- Meeting once every 2 months or as often as necessary.

The Hub Academic Head chairs the TAC.

The Technical Advisory Council (TAC) consists of:

1. Hub Academic Head (Chair).
2. Chief Executive Officer/Chief Operating Officer (Secretary).
3. Two members from IIIT-H.
4. One - two members from the academia (international & national).
5. Two members from the industry.

Figure 53: Organisational Structure of the Technical Advisory Council



9.2.1.3 Executive Committee (EC)

The Hub Academic Head shall chair the EC. The role of EC is to make decisions on day-today operational activities, ensuring planned activities are going as per the proposed plan.

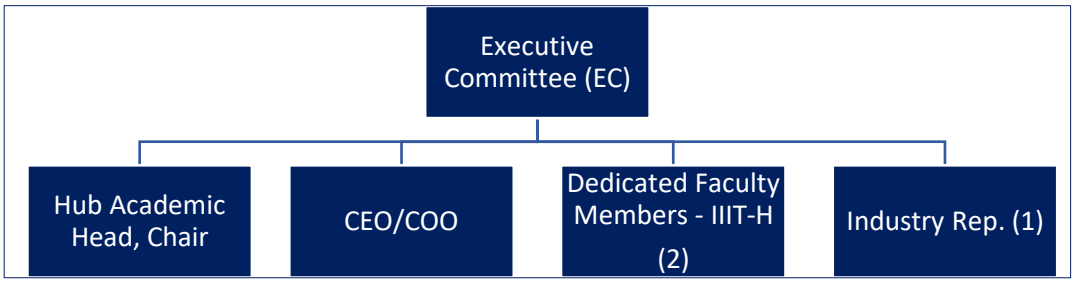
The role and responsibilities of the executive committee include:

- To take decisions on the operational day-to-day activities.
- Ensuring the planned activities are on track.
- Help the Hub Management team to address challenges.
- Meeting twice in a month or as often as necessary to evaluate status updates of programmes, and initiatives.

The executive committee includes:

1. Hub Academic Head (Chair).
2. CEO/COO (Secretary).
3. Two dedicated Faculty members of IIIT Hyderabad.
4. One industry representative.

Figure 54: Organisational Structure of the Executive Committee (EC)



9.2.1.4 Hub Key Personnel

Hub Academic Head (or Project Director)

Hub Academic Head is a faculty from IIIT-H and is overall in-charge of the Hub. His/her expertise shall be essential and critical in anchoring all the technology related activities.

The role and responsibilities include:

- Full responsibility of the Hub as an academic unit.
- Envisioning and planning of all academic activities including but not limiting to Technology development, Research Translation, Skilling & HRD initiatives, collaborations, and so on.
- Represent the Hub academically outside

Chief Executive Officer/Chief Operating Officer

The CEO/COO shall be an accomplished technical person with administrative experience. S/He will be the operational head of the TIH and will be responsible for the execution of all the initiatives under the TIH.

The role and responsibilities include:

- To implement all activities of the Hub.
- Plan and execute programmes, calls, drives, etc., as decided.
- Business development, Partnerships for the Hub.
- Outreach.

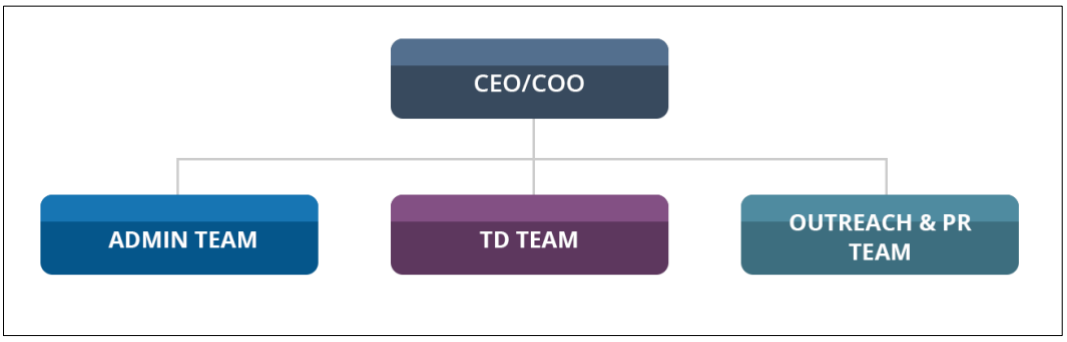
9.2.1.5 Hub Management Team

The CEO/COO will be head of the management team, which will help with operations, finance, administrative, management, outreach of all the activities of the Hub.

The management team includes:

1. CEO/COO.
2. Administration team.
3. Technology Development team.
4. Outreach and PR team.

Figure 55: The Hub Management Team



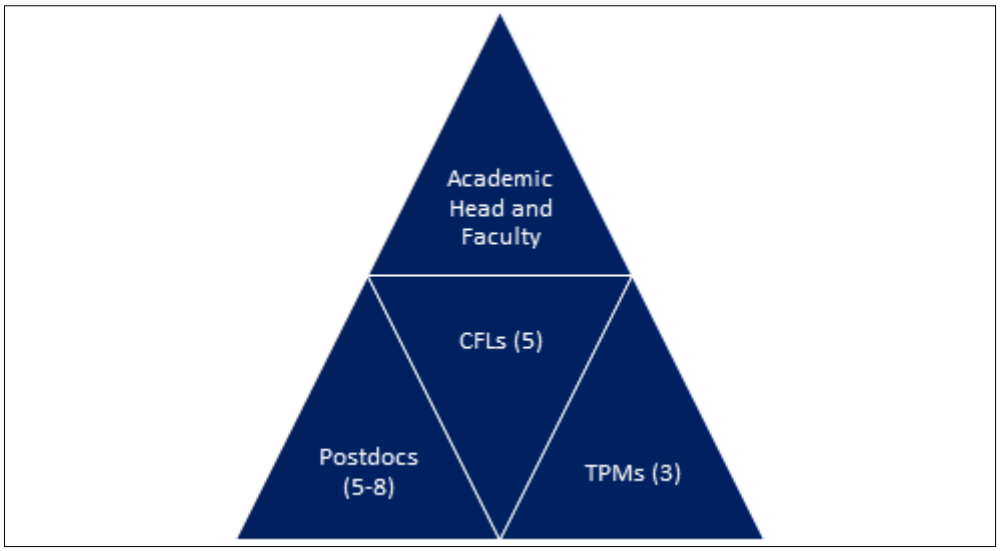
9.2.1.6 Applied Research & Translation

The translation responsibilities of the team include translation of the research output from the Hub into possible market solutions and products. A team of researchers, technical developers, product designers, and product managers will carry out and oversee these activities. An administrative team shall support day-to-day administration, finances, proposal process, international collaborations events, and so on.

The management will have academic and research team corresponding to each cluster. Each cluster will include:

- Hub Academic Head and Faculty.
- Cluster Faculty Lead (CFL).
- Five to eight postdocs.
- Technical Programme Managers (TPMs).

Figure 56: The Academia and Research Team



9.2.1.7 Annual Comprehensive Review

Given that the Hub is a national mission and involves technology life-cycle starting from knowledge until commercialisation, the TIH involves several initiatives, several stakeholders and several beneficiaries associated. To keep the momentum and focus of the TIH-Data aligned to the set targets, it is important to have an Annual Comprehensive Review. The review will include evaluation, assessment and monitoring the progress of overall TIH-Data and give its recommendation to the HGB. The review recommendations shall help the TIH-Data on assessing its impact and future directions. The review will include:

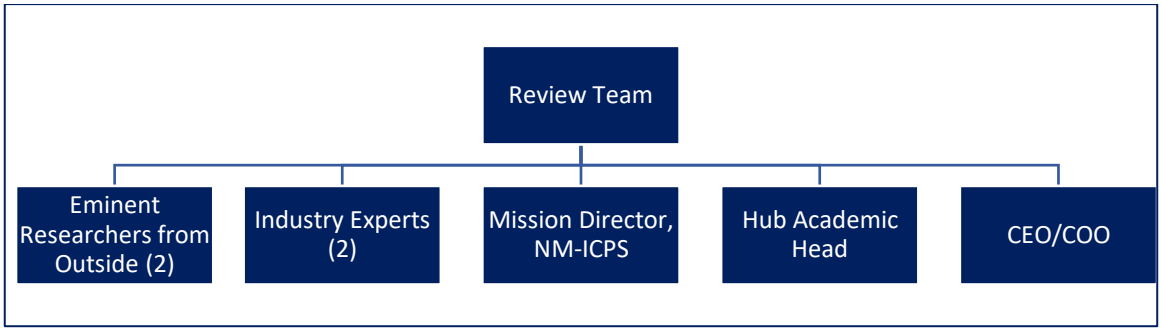
- Ensuring activities are towards meeting the big goals.
- Suggesting course corrections to the HGB.
- Frequency: once in a year.

This review shall be done by a team, comprising of experts, external to the TIH-Data, which will include:

1. 2 Eminent researchers from outside.
2. 2 Eminent industry experts.
3. Mission Director, NM-ICPS.
4. Hub Academic Head.
5. CEO/COO.

It shall be chaired by industry/academic expert from outside.

Figure 57: Review Team



9.2.2 Meeting Schedules of HGB, TAC, EC and Other Teams

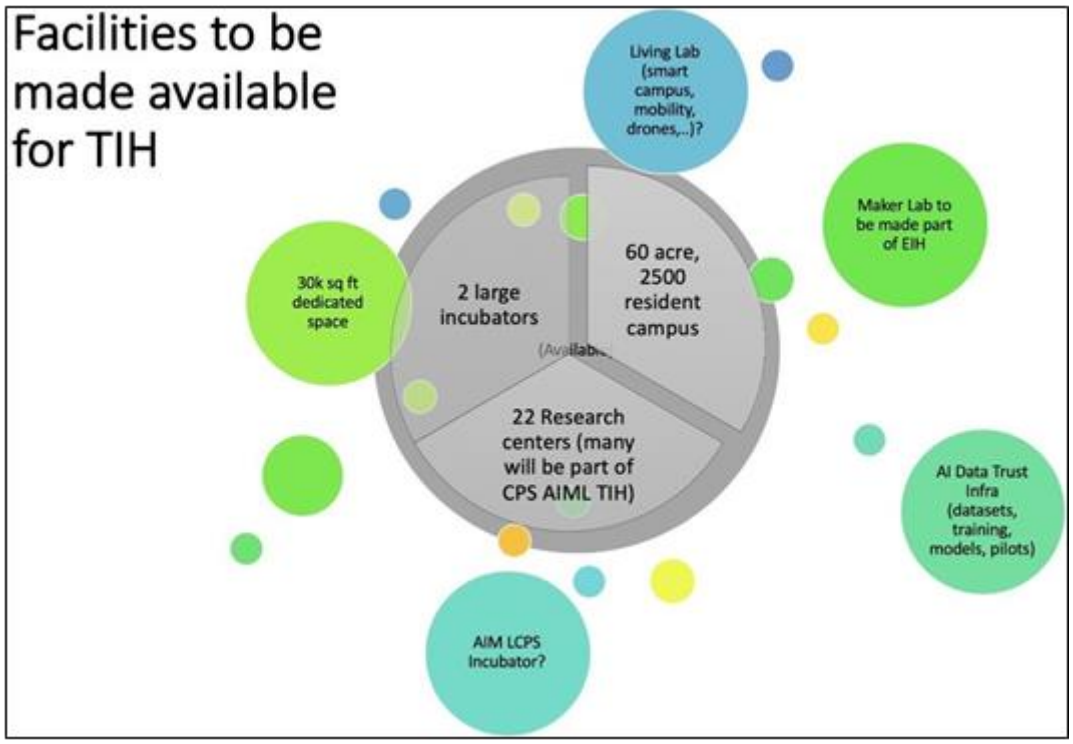
- Hub Governing Body meets twice in a year.
- TAC meets once in every two months.
- Review team meets once in a year or as necessary.
- The Executive Committee meets once in a quarter.
- Other teams meet as frequently as necessary to run the Hub smoothly.

9.3 Infrastructure

The activities of the centre would require significant infrastructure in terms of space, computing resources as well as network connectivity. IIIT Hyderabad has the following resources that may be used shared with the Hub on a cost-sharing basis as is decided in the contract between IIIT Hyderabad and TIH-Data". This will help in ensuring the smooth functioning of any activity and keep the expenses to a minimum.

The facilities available for TIH-Data are shown below:

Figure 58: Facilities Available for TIH-Data



- **Space**: IIIT Hyderabad will provide up to 30,000 sq. ft. of built-up area to host the centre. Additional furnishing of the space may be done through the support of DST using the budget allocated to ensure smooth functioning of the activities of the centre.
- **A computational cluster (part-time)**: IIIT-H has a computing cluster with over 40 servers and 160 GPUs. This resource may be shared with the centre for its research activities. This is in addition to the computing facilities that are proposed as part of the budget in this proposal.
- **Lecture halls and labs**: IIIT-H will share its lecture halls and labs for any of the meetings and courses that are conducted by TIH-Data. This will be a significant resource for the proposed Human Resource Development activities.
- **Network and communication facility**: The institute will provide access to its internal network as well as its dedicated internet connection for any of the activities of the centre.
- **Hostel rooms and guest house**: To facilitate the meetings, workshops and short courses that are conducted as part of the centre, IIIT-H will provide hostel rooms and guest house rooms as per the availability. This will help in ensuring the smooth functioning of such events and keep the expenses to a minimum.
- **Housekeeping, security and similar support**: In addition to the physical, computational and networking resources, IIIT will also assist in terms of security, housekeeping and other personnel assistance for many of the activities of the

centre. As the demand for most of such resources is highly variant based on the activities at any time, it is not practical for the centre to have dedicated resources of this nature. The institute will provide appropriate support to supplement the resources that are available to the centre.

- **Library**: IIIT Hyderabad will provide complete access to a fully equipped library with electronic and physical access to the latest books, journals and magazines in all related areas.

9.4 Names, Roles and Responsibilities of Team (Faculty from IIIT-H)

As described above, TIH-Data encompasses a large number of activities including knowledge generation, research translation, outreach activities, technology development, startups, product development/delivery, and HRD development in the data-driven technologies areas. The Host Institute in its all capacities will assist the Hub in each of these activities. This means involvement of key people/faculty members working in the research areas related to the activities of the Hub. The Hub will start off its activities in the core Data and domain areas of Mobility and Healthcare along with activities starting in Innovation/startups. The above five faculty members who are active in these areas will dedicate a significant amount of time.

The table below has the names, roles and responsibilities of the five identified faculty members who are active in their respective fields listed below. They will dedicate a significant amount of time towards the Hub's initiatives:

Table 6: Names, Roles & Responsibilities of TIH-Data Team

S. No.	Name of Team Member	Area of Expertise	Roles/Responsibilities
1	Prof. C. V. Jawahar	Computer Vision, Machine Learning, Applications	AI/ML for Mobility, Data Sets, Applications
2	Prof. Bapi Raju	AI, Cognitive Science, Machine Learning	AI Applications in Healthcare, Interface with Domain Experts (Doctors)
3	Dr. Vikram Pudi	Data Mining, Machine Learning	Data Foundation, Machine Learning and Applications, System Prototypes and Field Trials
4	Prof. Ramesh Loganathan	Incubation and Entrepreneurship, System and Solution Building, Technology Transfer	Technology Translation, IP Creation, Sustainability
5	Prof. Deva Priyakumar	Computational Biology, AI for Healthcare	Project director; AI applications in Heathcare

10 Finance and Budget

This chapter of the DPR outlines the cost estimates for initiating and running the Hub for the first five years. Most of the activities within the initial phase of the Hub will be supported by the Department of Science and Technology, Government of India. As described in other chapters. The estimates and the split of estimates to support different activities of the hub provided in this chapter are based on several brainstorming discussions. Initially, the overall year-wise budget for each of the activity is presented followed by detailed presentation of the splits.

10.1 Budget of The Proposed Hub for Major Activities

Budget is split across the following major activities:

Note: All numbers mentioned in this section are in Indian rupees unless specified otherwise.

Table 7: Overall Headwise and Activity Budget of the Hub

S. No.	Activity	Y1	Y2	Y3	Y4	Y5	Total
1	Recurring	12.25	18.00	24.00	18.00	17.40	89.65
2	Non-Recurring	8.00	5.00	6.00	1.00	0.35	20.35
3	Total in Rs Crores	20.25	23.00	30.00	19.00	17.75	110.00

Activity Wise Split up.

S. No.	Activity	Y1	Y2	Y3	Y4	Y5	Total
1	Data Foundation	10	4	6	2	2	24.00
2	Applied Research and Translation	2.3	4.5	6.7	5	4.5	23.00
3	Technology Development	4.5	7	9	6	5.5	32.00
4	HRD And Skill Development	1.45	3	3	2.5	2.05	12.00
5	Visitors and Programmes	0.2	1.2	1.2	0.6	0.8	4.00
6	Hub Management	1	1.5	1.5	1.5	1.5	7.00
7	Innovation and Startups	0.8	1.8	2.6	1.4	1.4	8.00
	Total	20.25	23.00	30.00	19.00	17.75	110.00

Recurring Expenses Budget:

S. No.	Activity	Y1	Y2	Y3	Y4	Y5	Total
1	Data Foundation	3.5	3.5	2.5	2	1.7	13.2
2	Applied Research and Translation	1.8	2.5	5.45	4.5	4.45	18.7
3	Technology Development	3.5	5	8	5.5	5.5	27.5
4	HRD And Skill Development	1.45	3	3	2.5	2.05	12
5	Visitors and Programmes	0.2	1.2	1.2	0.6	0.8	4
6	Hub Management	1	1	1.25	1.5	1.5	6.25
7	Innovation and Startups	0.8	1.8	2.6	1.4	1.4	8
	Total	12.25	18	24	18	17.4	89.65

Non-Recurring Expenses Budget:

S. No.	Activity	Y1	Y2	Y3	Y4	Y5	Total
1	Data Foundation	6.5	0.5	3.5	0	0.3	10.8
2	Applied Research and Translation	0.5	2	1.25	0.5	0.05	4.3
3	Technology Development	1	2	1	0.5	0	4.5
4	HRD And Skill Development	0	0	0	0	0	0
5	Visitors and Programmes	0	0	0	0	0	0
6	Hub Management	0	0.5	0.25	0	0	0.75
7	Innovation and Startups	0	0	0	0	0	0
	Total	8	5	6	1	0.35	20.35

Data Foundation: The focus will be to create, curate, host and provide service centred around the data. As explained elsewhere the Data Foundation interacts with domain experts in this process.

Applied Research & Translation: One of the primary objectives of the Hub is transfer of the technology. Towards this, a dedicated team and effort is planned. This effort will focus, nurture and help to mature the knowledge available in the research labs to products and services.

i. Technology development effort.
ii. Translation and Productization.
iii. Interface and management with the industry and agencies.

Technology Development: Another major focus of the Hub is to initiate the research and projects leading to technologies that are demanded by the society around. A number of short-, medium- and long-term projects will be planned and executed. Formal framework for project selection and management will be aligned with the models followed by other hubs and DST. The budget available for this will be also available for external collaborators and researchers within India.

HRD/Skill Development: Our Hub also aims to make impact in the academic space with fellowships. We propose to have fellowships for:

i. PG students,
ii. PhD students,
iii. Postdoctoral students,
iv. Faculty fellowships, and
v. Faculty chairs.

Costing for the fellowship is done based on the guidelines provided in the mission DPR

Hub Management: Hub as a Section 8 Company has its own light weight administrative managerial team. In addition, there are operational expenses and associated contingencies and capital.

Visitors & Programmes: A unique character of the Hub is the ease with which it interacts with experts within and outside the country. We also propose to host a larger number of visiting researchers and students in the Hub so that there is very good synergy across research groups.

Innovation and Startups: A route to make impact in the industry is by developing EIRs and incubating startups. Hub has explicit plans in carrying out these through a dedicated effort.

10.2 Data Foundation

Table 8: Budget Outlay for Data Foundation

S. No.	Head	Y1	Y2	Y3	Y4	Y5	Total
A	**Recurring**						
1	Manpower	**1.95**	**1.95**	**1.65**	**1.2**	**0.93**	**7.68**
2	Consumables	**0.20**	**0.20**	**0.10**	**0.10**	**0.10**	**0.70**
3	Consultants	**0.40**	**0.40**	**0.20**	**0.20**	**0.20**	**1.4**
4	Utilities	**0.60**	**0.60**	**0.30**	**0.30**	**0.30**	**2.4**
	Sub-Total	**3.15**	**3.15**	**2.25**	**1.8**	**1.53**	**11.88**
5	**Overhead (10% Recurring)**	**0.35**	**0.35**	**0.25**	**0.20**	**0.17**	**1.32**
B	**Non-recurring**						
1	Equipment	**6.5**	**0.50**	**3.5**	**0**	**0.30**	**10.8**
	Total	**10**	**4**	**6**	**2**	**2**	**24**

Equipment: The Data Foundation has to support some of the fundamental activities of the Hub such as data hosting, data services and data analytics. These activities demand hosting necessary hardware that primarily are storage servers and computing machines. Availability of state-of-the-art solutions is crucial for running the Data Foundation.

Workforce: Data foundation needs good system design and implementation skills which in general is lacking and costly workforce to acquire. Workforce in this space will be spent on software architecting and development effort. Some manpower for domain-specific expertise is also accounted for beyond that of domain experts as consultants.

Consumables & Contingency: The Data Foundation is an entity which brings together a large number of people, hosts computational resources for both storage and computing. The consumable/contingency heads will be used to provide the essential office stationery, minor computer peripherals such as network switch, storage, cables, batteries, etc., minor software essential for office/technical needs, and costs related to procuring services such as cloud storage and computing.

Consultants: As mentioned above, the Data Foundation from time to time needs human resources at the technical level both for advising in the data aspects. Secondly, the

foundation will need advice from domain experts like doctors, automobile engineers and architects. While hiring such professionals at the full-time basis will be difficult to hire as part of the Data Foundation, it is proposed to get their services as consultants on a regular basis.

Utilities for Data Center: The equipment that are being planned as part of the Data Foundation will be housed in a data center that needs uninterrupted electricity supply, and cooling. A facility of this magnitude is estimated to incur an electricity charges up to 1 crore rupees per year. To partially offset the cost, utility charge is being requested.

Overhead: The equipment for Data Foundation will need large space and incur a large cost in electricity, water, etc. Hence, overhead of 10% on the recurring costs partially defray the indirect costs.

10.3 Applied Research and Translation

Table 9: Budget Outlay for Applied Research and Translation

S. No.	Head	Y1	Y2	Y3	Y4	Y5	Total
A	**Recurring**						
1	Manpower	1.12	1.70	3.71	3.05	3.11	12.69
2	Travel	0.10	0.10	0.30	0.15	0.15	0.80
3	Consumables	0.10	0.15	0.30	0.15	0.15	0.85
4	Contingency	0.30	0.30	0.60	0.70	0.60	2.50
	Subtotal	1.62	2.25	4.91	4.05	4.01	16.84
5	Overhead (10% Recurring)	0.18	0.25	0.54	0.45	0.44	1.86
B	**Non-Recurring**						
1	Equipment	0.20	2.00	1.25	0.50	0.00	3.95
C	Capital	0.30	0.00	0.00	0.00	0.05	0.35
	Total	2.30	4.50	6.70	5.00	4.50	23.00

Equipment: The applied research/translation is responsible for the applied research and its translation for applications in industry and in society. The computing needs for this purpose will be supported from the equipment grant as part of this part of the budget. The equipment requested primarily are GPU compute servers, and the associated hardware such

as networking, UPSs, etc. In addition, this budget head will be used to provide front end computing machines (desktops/laptops), printers, electronic devices for IoT projects, etc.

Capital: The facility will host a large number of technical manpower and associated support system (admin support, space, etc.), which is expected to occupy about 15,000 sq. feet space. The Capex will be used to develop the space to provide good working environment and facilities such as meeting rooms, furniture/furnishings/visitors' working space, etc. Minor construction activity may also be allowed if necessary.

Manpower: For the applied research and for translation, a dedicated team of human resource at multiple levels is necessary. Engineers, scientists, and admin staff including managers who would establish industry connects will be hired particularly for this activity. This team will also include masters, doctors and postdoctoral fellows who may be part of the translational research activity.

Travel: One of the major activities of the facility is creation and curation of data, which needs conversation with a number of experts both from the core data aspects and from the domains such as healthcare and mobility. It is also necessary for the research staff part of the Data Foundation to attend conferences, expositions, etc. to widen their knowledge, to present the work done, and to reach out/connect to right experts. The travel head will be used for all travel related to these activities.

Consumables: Given that a large number of people are part of the applied research/translation team, there will be costs associated with office stationery such as print cartridges, paper supplies, etc., small storage devices, purchase of low-cost software, cloud storage, computing, etc. This will be used for these purposes.

Contingency: The facility will have visitors from industry/academia, will organise events/meetings, will host experts, committee members, conduct proposal presentations, proposal reviews, etc. All these events will require hosting visitors in guest house/hotels, provide local hospitality, provide refreshments during meetings, etc. The requested contingency grant is necessary to take care of such expenses and other exigencies.

Overhead: The applied research/translation will host a number of engineers/scientists/ and other researchers. This also hosts a medium sized compute facility (independent of the Data Foundation) to cater to the needs of these researchers. Financial support to at least partially cover the cost associated with running costs is necessary.

10.4 Technology Development

Table 10: Budget Outlay for Technology Development

S. No.	Head	Y1	Y2	Y3	Y4	Y5	Total
A	**Recurring**						
1	Manpower	**2.10**	**2.80**	**4.60**	**2.90**	**2.90**	**15.30**
2	Travel	**0.15**	**0.30**	**0.80**	**0.35**	**0.35**	**1.95**
3	Contingency	**0.60**	**0.90**	**1.10**	**1.00**	**1.00**	**4.60**
4	Consumables	**0.30**	**0.50**	**0.70**	**0.70**	**0.70**	**2.90**
	Subtotal	**3.15**	**4.50**	**7.20**	**4.95**	**4.95**	**24.75**
5	**Overhead (10% Recurring)**	**0.35**	**0.50**	**0.80**	**0.55**	**0.55**	**2.75**
B	**Non-Recurring**						
1	Equipment	**1.00**	**2.00**	**1.00**	**0.50**	**0.00**	**4.50**
	Total	**4.50**	**7.00**	**9.00**	**6.00**	**5.50**	**32.00**

The technology development will be completely for funding projects via the Hub. The Hub will advertise for project proposals from partner, collaborating and host institutions to involve in knowledge generation and translation in the area. In addition to these, projects may also be granted to investigators from other institutes with good proposals in data-driven technologies. All the projects will be evaluated by a committee and those that are recommended for funding will receive funding via this head. In the DST style, funding will be provided under the recurring and nonrecurring heads that may include manpower, travel, contingency, consumables, overheads and equipment as the sub-heads. These components will be necessary for the PIs of the approved projects to carry out the projects successfully. At the end of the project, further funding to encourage translation of the research may be provided in deserving cases.

10.5 HRD & Skill Development

Table 11: Budget Outlay for HRD & Skill Development

S. No.	Fellowship Category	Deliverables	Y1	Y2	Y3	Y4	Y5	Total
1	PG Fellowship	50	0.05	0.80	0.80	0.50	0.25	2.40
2	Doctoral Fellowship	30	0.90	0.90	0.90	0.90	0.90	4.50
3	Postdoctoral Fellowship	10	0.20	0.40	0.40	0.30	0.30	1.60
4	Faculty Fellowship	5	0.15	0.45	0.45	0.40	0.25	1.70
5	Chair Professor	5	0.15	0.45	0.45	0.40	0.35	1.80
	Total		1.45	3.00	3.00	2.50	2.05	12.00

One of the major activities of the Hub is to nurture talents at multiple levels and to train them in the state-of-the-art emerging area. One of the major activities of the Hub towards meeting this objective is via CHANAKYA fellowships offered at several levels. The fellowship levels are kept at the levels suggested by the DST in its DPR for each of the following categories.

Postgraduate fellowships: One- to two-year (1 to 2 year) fellowships for students in their master's programme (MS or MTech) programmes will be supported via this program. Students from within IIIT Hyderabad or from other Institutes/Universities will be selected based on applications to receive the fellowship. The fellows who have been selected part of this scheme will be encouraged to visit/work at the Hub enabling them better training in the area by the in-house engineers/scientists.

PhD fellowships: Select number of PhD students working in the area of data-driven technologies in general will be selected for a 4-year competitive fellowship. The research may also be undertaken in the Hub or at the host institution or from other Universities/Institutions in the country. They will also have opportunities to spend part of their time at the Hub to further their skills with the help of the team at the Hub.

Postdoctoral fellowships: A limited number of fellows who after finishing PhD are looking to strengthen their capabilities and gain more experience in the given area will be selected for a postdoctoral fellowship. They will be supported for a maximum of two years that will enable them to pursue their careers in the area whether in academia or industry. All the fellows will be assigned to mentors of their choice (to the extent possible) during their stay at the Hub.

Faculty fellowships: This scheme is meant for talented and motivated young researchers who have completed their PhD and have adequate background in the subject. The faculty fellows will have an opportunity to spend dedicated time to engage in research in the area. They will also be supported by adequate research funds to build a lab and jump start their work. Up to five fellows will be supported by this scheme and depending on their performance during the time of the fellowship, they may be eligible for a regular position in the Institute.

Chair professorships: The Hub will employ a number of people at the technical level, admin staff, scientists, managers, etc. The hub will also take the services of industry experts, experts from academia, other consultants, etc. In addition to these, it is essential that there exists a strong core team from the host institute that give enough time to successful implementation of the plans of the Hub. Chair professorship position and associated research fund will be helpful to encourage and incentivize the involvement of faculty members to dedicate significant amount of their time.

10.6 Visitors and Programmes

Table 12:Budget Outlay for Visitors and Programmes

S. No.	Activity	Y1	Y2	Y3	Y4	Y5	Total
1	Education Outreach	0.03	0.10	0.10	0.10	0.10	0.43
2	Workshops & Programmes	0.03	0.10	0.10	0.10	0.10	0.43
3	International Research Collaboration	0.09	0.90	0.90	0.20	0.40	2.49
4	Visitors	0.05	0.10	0.10	0.20	0.20	0.65
	Total	0.20	1.20	1.20	0.60	0.80	4.00

A vibrant visiting researchers programme and researchers to be able to take part in conferences/workshops/summer schools is an essential part of a world class research activity. The Hub proposes to enable events such as these and to host visitors both from within India and from outside India.

Education Outreach: The Hub will take part in diverse educational outreach activities enabling high end skilled human resource development. Short and long courses that allow for a select number of participants to undergo hands-on training along with foundation lectures will be organised from time to time. In the long-term course, the participant may undergo training partially online and partially in physical presence. The courses organised will cater to the needs of fulltime research students and for the working professionals.

Workshops and Programmes: Various programmes such as short symposium style events and one-week short courses (summer/winter schools) will be organised by the Hub. Symposium like meetings will be meant for experts, from student researchers to senior investigators, to come together share ideas and discuss their work. Researchers who have significantly contributed to the area of research will be invited to provide enrich the discussions. On the other hand, the short courses will be primarily for early student researcher who are already working in the area or for those students who want to work in the area. Experts from the area will be invited to deliver introductory lectures and tutorial style lectures, and several hands-on sessions will be conducted to get practical training.

International Research Collaboration: The Hub will enable strong international collaborations both with experts in the domain areas as well as in the core data science areas. As part of this collaboration, the Hub proposes to host about 15 International experts during the first five years who will visit the Hub and involve in several activities. These may include closely working with the researchers in the Hub, providing directions and advise to the Hub, giving lectures in educational outreach programmes, etc. The visitor may also visit other partners and collaborators of the Hub for collaboration/advice or for training purposes.

Visitors: The hub proposes to host students and early career scientists so that collaboration with others in the Hub is enabled. These collaborations will also help the visitors to use the facilities available in the Hub for conducting their research all the way to taking their research to the society/industry. Visiting of senior researchers/experts also will be accomplished via this program.

10.7 Hub Management

Table 13: Budget Outlay for Hub Management

S. No.	Head	Y1	Y2	Y3	Y4	Y5	Total
1	Workforce	**0.40**	**0.40**	**0.70**	**0.70**	**0.70**	**2.90**
2	Operational	**0.10**	**0.20**	**0.30**	**0.40**	**0.40**	**1.40**
3	Contingency	**0.05**	**0.20**	**0.15**	**0.20**	**0.20**	**0.80**
4	Capital/Infra	**0.40**	**0.00**	**0.00**	**0.00**	**0.00**	**0.40**
5	Utilities	**0.05**	**0.20**	**0.10**	**0.20**	**0.20**	**0.75**
	Total	**1.00**	**1.00**	**1.25**	**1.50**	**1.50**	**6.25**
	Non-Recurring	**0.00**	**0.50**	**0.25**	**0.00**	**0.00**	**0.75**
	TOTAL	**0.00**	**1.05**	**1.50**	**1.50**	**1.50**	**7.00**

Workforce: The Hub involves activities starting from basic research to its applications in industry/society. Hence it is necessary to have the right manpower to manage the administrative aspects. The hub will hire a chief executive officer, programme managers, and other staff for the administration of the Hub. A number of positions are required to manage the different aspects of the Hub (the Hub, innovation, startups, visitor programme, technology development, research proposals, etc.)

Operational: The Hub is going to be housed in a 30,000 sq. feet space that will comprise offices, meeting rooms, cubicles, etc. The operational expenses will cover the minimal rent for the space, housekeeping, security, etc.

Contingency: The hub will host a number of events and will host a number of visitors and each of these are associated with certain expenditure. The contingency funds will be used to provide local hospitality to visitors, pay temporary-contract work, refreshments expense related to visitors, etc.

Capital: The hub will be a new entity and hence the space needs to be developed. This space will be occupied by the central management team, other admin staff, engineers/personnel related to innovation/entrepreneurship/startup, visitors of the Hub, etc. The capital cost of 1 crore rupees will be used for creating/renovating the space and for minor construction.

Utilities: Since there are no overheads for running the Hub, and maintaining the space requires running costs. Support to cover the electricity costs is requested.

10.8 Innovation, Entrepreneurship, and Startups

Table 14: Budget Outlay for Innovations & Startups

Innovations & Startups							
S. No.	**Major Component**	**Y1**	**Y2**	**Y3**	**Y4**	**Y5**	**(Cr)**
1	EIR (100)	**0.20**	**0.70**	**1.00**	**0.50**	**0.50**	**2.90**
2	Grand Challenges	**0.10**	**0.20**	**0.30**	**0.10**	**0.10**	**0.80**
3	Startup/Spin Off	**0.40**	**0.70**	**1.00**	**0.70**	**0.70**	**3.50**
4	TBI	**0.10**	**0.20**	**0.30**	**0.10**	**0.10**	**0.80**
	Total	**0.80**	**1.80**	**2.60**	**1.40**	**1.40**	**8.00**

The EIR programme is the first step towards taking a research technology into the market. The programme provides an entrepreneur with a fellowship to explore taking the research into market. Typically, it is a fellowship of 30-50K per month, for 6-12 months. Once market opportunity is validated through a prototype and market pilots, the startup is incorporated and technology licensed.

Grand Challenges

The challenges are alternate to EIR programme to seed new products from research. These will invite existing startups and tech groups to build solutions based on research done. The challenges will fund an incentive for the POC to be built and then a grant for the final winning team to take the product into market. Typically, these maybe about Rs.10-25 Lakhs per challenge.

Startups/Spin offs

The startups and startup products coming out from the EIR and Grand Challenges will be nurtured and supported as they foray into the markets and onboard their initial set of customers. This is done through seed investment in the startup. Typically, one of 3-4 EIR startups will make a successful entry into the markets. Each of these may be supported to the tune of 10-25 Lakhs.

Technology Business Incubator

All the startup seeding and nurturing activities will be taken up in the Technology Business Incubator. An operations team that will organise and manage all the events and campaigns, once cohorts (of EIRs or startups) in place, will manage the cohort. Will onboard resident mentors and advisors. And organise events both prior to the programmes and also on successful completion of each cohort. And specifically organise showcases and discovery events to uncover interesting startups coming out in the Data led domain realm. These activities will be separately funded.

11 Timelines

Many administrative actions are needed to achieve the big picture set by the National Mission for the success of TIH-Data. We group this into two sets of activities.

- First is a set of actions related to setting up and initialising the activities
- Second is more of regular functioning with a five-year visibility.

11.1 Activity List 1: Initiation

- Set up MoA and Section 8 Company.
- Establishment of Hub management/admin team including CEO.
- Hiring technical, admin and operational team.
- Setup Data/compute infrastructure.
- Refinement of the plans and low-level details. Working out low level details for the implementation/operation.
- Cataloguing the list of technologies available, technologies ready for translation and technology products.
- Launching the fellowship programmes at multiple levels. Phase 1.
- Launching the fellowship programmes at multiple levels. Phase 2.
- Launching Tech-focused, Go-To-Market accelerator, Entrepreneurship-In-Residence programme and seed-fund scheme.
- Initiating projects by call for proposals and review/selection process.
- Initiating EIR, startup and innovation programmes.
- Selection of international and industry collaborators.
- Making data banks and data APIs available to the public.

11.2 Activity List 2: Regular

- Call for proposals that are complementary and supplementary to the existing proposals.
- Awarding Projects and project reviews.
- Fellowship cycle.
- Educational events and skill development programmes.
- Organising annual showcases for technologies, products and startups and workshops with wider hub communities, potential end-users, industries and startups.
- Conducting quarterly reviews of all the programmes and fellowships.
- Preparing and publishing annual reports.
- Conduct high-skill training & development programmes periodically.
- Conducting grand challenges and competitions.
- Yearly review of the Hub activities and programmes.
- Regular technology translation and review meetings.

- Incubation, accelerator and EIR programmes.
- Workshops/Summer-Winter schools.
- Hub Governing body and advisory body meetings.

11.3 TIH-Data Timeline

Please refer to the tables, *Activities of the Hub: Year 1-Year 5* and *Activities of the Hub in Year 1 & 2* at the end of this section.

11.4 Year-Wise Details

11.4.1 Year 1

- Set up infrastructure, hiring, team formation.
- Set up compute, Data and technology infrastructure.
- Brainstorm with domain experts and stake holders to structure the programmes; Refinement of the programmes and work out fine details for the implementation.
- Call for Requests for Proposals (RFPs) for Research Grants. Evaluation and Sanction. Establish the project cycle.
- Identify the list of prospective international collaborators, industry collaborators and commence outreach.
- Technical activities: On the technical front we propose to:
    i. initiate the project selection and formal sanction.
    ii. to formalize the data API and engagement models with the users.
- Translation activities:
    i. Collate a number of industrial use cases that can be connected to the Hub mission.
    ii. work with PIs of the projects on the potential technologies and results that will come out of this effort.
    iii. Align the above two.
- HR and Training activities:
    i. Initiate part of PhD and Postdoc Fellowships.
    ii. Organise short courses and educational programmes.
- Annual review of the Hub.

11.4.2 Year 2

- Compiling available technologies and prototypes with potential impact in the domains.
- Identifying five technologies for translation every year.
- Strengthen the Entrepreneur-in-Residence (EIR) programme by recruiting 20 EIRs/year.
- Bringing domain experts to fine-tune project and research directions.

- Interfacing with government agencies and Line Ministries in further strengthening the Hub.
- Launching technology- and theme-focused accelerators with dedicated boot camps, hackathons and identification of startups with interest in technologies that come out of the Hub.
- Technical activities:
    i. More projects out and making good progress.
    ii. Identification of missing important aspects/problems and initiating the second set of projects.
- Translation activities:
    i. Have multiple successful translation.
    ii. Smooth advance in EIRs and Startups and the associated programmes.
    iii. Work closely with potential use cases based on the technology development taking place.
- HR and Training activities:
    i. Furthering the PhD and Postdoc Fellowships.
    ii. Organise short courses and educational programmes.
    iii. Evaluation of fellowship programmes.
- Annual review of the Hub.

11.4.3 Year 3

- Technology products startups showcase/workshop with wider hub community and potential end-users, industries and startups.
- Launch of grand challenges and competitions
- Strengthening international collaboration through the student exchange programme, visiting faculty programme, and so on.
- Transnational effort in full intensity. Clearly visible outcomes.
- Two to three technologies and technology-products to be deployed and field trial be carried out with user agencies. Five patents from the research.
- Technical activities:
    i. Visible technical outcomes in terms of Papers, Demos and technology.
    ii. Work on next list of projects to be released.
- Translation Activities:
    i. Collate a number of industrial use cases that can be connected to the Hub mission.
    ii. Work with PIs of the projects on the potential technologies and results that will come out of this effort.
    iii. Align the above two.
- HR and Training activities:
    i. Initiate part of PhD and Postdoc fellowships.
    ii. Organise short courses and educational programmes.
    iii. Evaluation of fellowship programmes.

- Mid-term review of the Hub and making long term plans.
- Annual review of the Hub.

11.4.4 Year 4

- Technology evaluation and business plans for the initial set of technologies.
- Strengthen the connect with user agencies, understand the feedbacks about the technologies and use cases.
- Technical Activities:
    i. Final list of projects sanctioned.
    ii. Meeting 75% of the targets in publications and technology outcomes.
- Translation Activities:
    i. 75% of the targets in EIR programme.
- HR and Training Activities:
    i. Postdoc Fellowships and other potential short-term fellowships.
    ii. Organise short courses and educational programmes.
    iii. Evaluation of programmes.
- Work on self-sustainability of the Hub. Convincing models for self-sustainability.
- Annual review of the Hub.

11.4.5 Year 5

- Twenty socially-relevant technology deployments. Case studies and lessons learnt from the technology.
- Creation of technology roadmaps and plans; refinement of the scope of the Hub.
- Creation of roadmap for sustainably operating the Hub; formation of alliances and partnerships.
- Technical Activities: On the technical front we propose to:
    i. Initiate the project selection and formal sanction.
    ii. To formalise the data API and engagement models with the users.
- Translation Activities:
    i. Collate a number of industrial use cases that can be connected to the Hub mission.
    ii. Work with PIs of the projects on the potential technologies and results that will come out of this effort.
    iii. Align the above two.
- HR and Training Activities:
    i. Postdoc Fellowships and other potential short-term fellowships.
    ii. Organise short courses and educational programmes.
    iii. Evaluation of programmes.
- Annual review of the Hub.
- Continuation plans and moving forward.

Table 15:Activities of the Hub: Year 1-Year 5

Major Activity	YEAR 1				YEAR 2				YEAR 3				YEAR 4				YEAR 5			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
PG Fellowships																				
PhD Fellowship																				
Post Doc Fellowship																				
Call for Proposals																				
Project Sanctions																				
Project Review																				
EIR Programmes																				
Accelerator Programmes																				
Educational Events																				
Research and Technology Review																				

Major Activity	YEAR 1				YEAR 2				YEAR 3				YEAR 4				YEAR 5			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Technology Translation Planning & Review																				
Technology Workshops																				
Technology Exhibition																				
Hub Review																				
Incubation																				

Table 16: Activities of the Hub in Year 1 & 2

Activity	YEAR 1& MONTHS												YEAR 2 & MONTHS											
	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12
Setup MoA and Section 8 Company																								
Hiring Management team																								
Hiring Technical & Admin Team																								

Activity	YEAR 1& MONTHS												YEAR 2 & MONTHS											
	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12
Fine details of Programmes and plans																								
Setup Data/Compute Infrastructure																								
Establishing Data Foundation																								
Technology Catalogue & Translation Details																								
Launch of Fellowships phase 1																								
Launch Fellowship Phase 2																								
Call for Proposals																								
First set of Projects																								
International Programmes																								
EIR Programmes																								
TBI & Startups																								
Data for Public Use																								

12 Cost Benefit Analysis

The proposed TIH-Data has the following functions that need financial resources:

1. **Knowledge and Technology Generation (KTG):** Envision directions of potential research advances as well as their application opportunities in the broad area of Artificial Intelligence and Machine Learning. The technology development activities will be carried out by research groups/teams in different parts of the country including in the host institution.
2. **Translation and Promoting Innovation (TPI):** Technologies developed by research teams often need to be worked on by a translation team to make it ready for licensing to companies or to startups. This involves creating a development team of engineers and product managers who liaise with the technology creators and potential customers.
3. **TIH-Data Coordination and Administration (TCA):** The hub will manage the process of calling for proposals, coordinate their evaluations, disbursing of funds according to the project approvals, coordinating their progress evaluations, and facilitate the translation of technologies for wider use whenever appropriate. The Hub also coordinates events, visits by renowned researchers, publicity, and other support activities.

The type and quantum of funds required for the above are:

1. KTG activities are scalable based on the level of interest and the fund requirements depend on the specific projects underway. KTG activities are to be funded from DST's funds for the first 5 years. After that duration, we intend to raise funds from different sources to continue the research in related areas, leveraging the results of the Hub model in the previous years. Such hubs can attract funds for mission mode activities from government sources such as MEITY, DBT, DST/SERB, and DRDO. We will also evolve models to attract funds from private companies and foundations for multi-institution projects coordinated by the Hub. This will be a pioneering research-support model, but one that needs to be established for the larger benefit of research in India.
2. TPI activities can also be at different intensities, depending on the number of projects underway. Research-translation can generate funds when some of the technologies are licensed profitably. We expect to raise a medium amount of funds from this source by the end of 5 years itself, with fund-flow increasing in later years. Research translation, startup facilitation, academic events, etc., can also attract funds from different governmental and private sources.
3. TCA funding is for the indirect-expenses or to facilitate the Hub's activities. This needs to be funded from overheads of funds raised from different sources. This

model is well understood globally, but is challenging in the Indian context where research overheads are trivial ordinarily. We need special models to make it work.

The different possible sources of funding for the Hub beyond the 5-year period include:

1. Funds from various funding sources. This needs novel mechanisms for the Hub to work as are search coordinator. We expect to attract funds from MEITY, DST, DBT, DRDO, DAE, and other governmental research funding agencies.
2. Funds from companies and foundations that are interested to run large coordinated projects. There are enough precedents for this already in IIIT-H of engaging with industry. This experience as well as the experience accrued during the first five years of engaging with the industry will be carried forward into the sixth year and beyond.
3. Technology transfer and licensing of technologies can bring considerable amount of revenues to the Hub. These need to be shared in appropriate proportions, but a good fraction can come to the Bub for its activities.
4. Startups that emerge from the projects can also be a source of considerable revenue in the long term.
5. The Hub will be an intellectual power-centre in the areas of AI and ML. The Hub can raise revenue by leveraging this through consultancy, executive education, and other means.

12.1 Sustainability/Revenue Generation

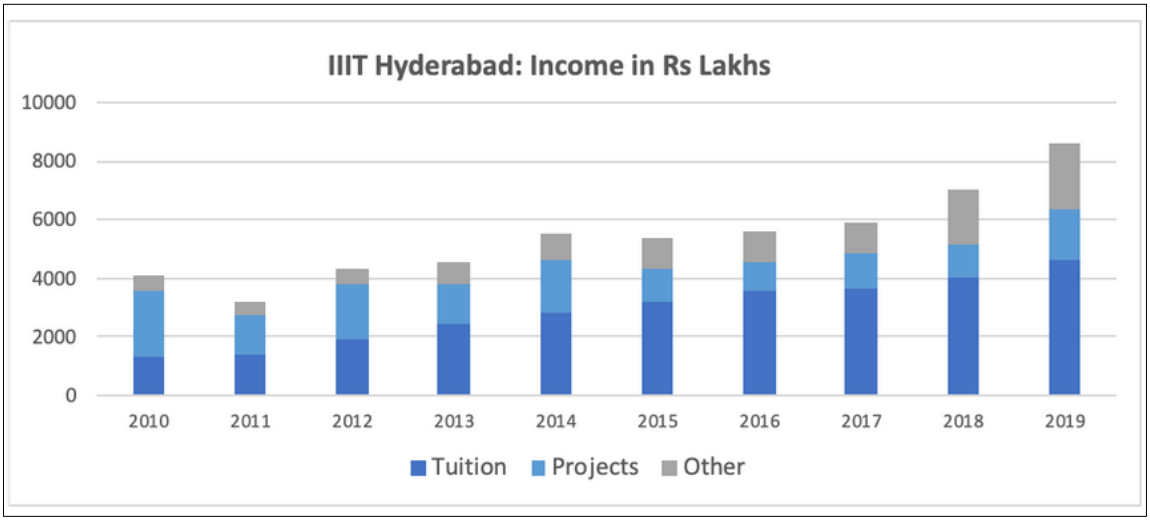
The Technology Innovation Hub (TIH) is being established in the International Institute of Information Technology (IIT), Hyderabad to promote and seed activities in the area of broad Data-Driven Technologies. The importance of rich and varied data from different spheres of human activities is certain to increase several folds in the coming years, as Machine Learning based applications of Artificial Intelligence expands further. Research and development in these areas will expand in scope and extend as a result.

The present Hub is established with substantial support from the DST for a time period of five years. This can only be viewed as the seed funding to establish a focus point in the country to promote data-driven technologies. The hub and its activities must grow significantly in the coming years and extend beyond the initial period of five years. Creating financial and technological models for the Hub is thus very important.

IIIT Hyderabad was established in 1998 by a grant of the land and initial set of buildings by the then state government. The institute has not received any funds as grants from any government right from its very start. All funds required to pay salaries of faculty and staff, to run the institution, and to expand its research have been raised the IIIT-H on its own from day one. In the meanwhile, the institution grew from 50 students in 1998 to over 1800 students in 2020, with over 500 students in research programmes like PhD and MS by Research. This institutional experience will be transferred fully and will be a big asset to the to-be-

established TIH. The flow of finances of the IIIT-H in the past 10 years can be found in the table below, with none of the revenue coming from a blanket government grant.

Figure 59: IIIT-H: Annual Income 2010-2019



The Hub will have multiple sources for future revenues to sustain and grow, during the initial five years and beyond, given that the technologies of Data Science, Data Analytics, Machine Learning, and Artificial Intelligence are slated to become critical technologies for the coming decades. Their journey towards serious societal impact has barely started and a long and profitable road lies ahead.

The primary requirement of funds is for the activities involved with (a) Generation, curation, annotation, and facilitation of different types of data relevant to the country, (b) applied research that exploits such data for societal applications in collaboration with several users, and (c) translating the fruits of the research to industry, startups, and other user agencies. Promoting data-driven innovation and startups is another activity of the Hub that needs to be funded well. The scope and extent of HRD and Skills development also needs to grow. However, that line of activities can generate considerable funds for itself and other activities of the Hub as we outline below.

We envisage 4 sources of revenue for the TIH in the future. Some of it can start in year 3 of establishing the TIH itself. These will continue to take place and grow as time goes by. A small business-development team will be established in the Hub to contact and coordinate potential users to raise the required funds.

Projects from different government sources: The TIH will establish itself as a major force in the areas of data-driven technologies, combining the strengths of IIIT Hyderabad and other partner institutions. Several solutions will be developed in the first few years itself, based on which the Hub will actively seek out funds in terms of large projects from different line ministries of the Government of India and other state governments. Given the importance of this area, funding is likely to be available in ample quantities. In addition, the Hub will partner

with NGOs and other organisations working in the societal sector to implement projects with large-scale impact, with help from governments. For instance, IIIT Hyderabad has already started discussing with the Public Health Foundation of India (PHFI) and the Government of Telangana on undertaking a few projects in the public-health area in Telangana, which will create lots of data and will be the basis for several solutions built on them. Ministries of Health, Transportation, Agriculture, Industry, Labour, etc., are the most promising targets to start these activities with.

Projects from industry sources: The industry builds solutions for the society and are in immediate need of data, solutions, and experts in data-driven technologies. Global industry leaders have established large R&D groups in India and many of their cutting-edge solutions have a significant Made-in-India component. These industries have been reaching out to the academia for help with solutions to some of their problems. TCS, Nissan, Google, Flipkart, Microsoft, Qualcomm, and Honeywell have all had worked on AI-related sponsored projects with IIIT Hyderabad in the past three years. Industry can be a good source of funds for generating data-sets, providing solutions, etc.

Technology licensing and transfer: The TIH effort will result in the generation of several technologies and intellectual properties. The Applied Research and Translation team is focused intensely on transforming output from research teams to solutions that industry can directly adopt through technology transfer. The technology can be transferred to startups also though the returns from them will be in the form of equity for 3 or more years. However, great benefits can ensue in the time scale greater than 5 years even if a single startup makes it big.

Short courses and workshops: Data Analytics, Data Science, Machine Learning, and Artificial Intelligence are very closely related concepts. There is plenty of demand for manpower trained well in one or more of the above in the industry, startups, and even the government. IIIT Hyderabad has run a large-scale executive education programme in AI/ML from 2018 onwards and have trained about 3000 professionals through a 3-month programme that is primarily based on contact. The market is full of such offerings by different academic institutions. The Hub will be in a unique position to do this with its strength in the gestalt of data-driven technologies. Short courses tuned to specific industry segments can also be created.

13 Risk Analysis

Research and Development (R&D) plays an important role in the development of any society.

When R&D gets encouragement from the governments by creating different avenues, it opens the door to more opportunities for academicians, startups, incubators (private & government), corporates and so on to play a vital role in building the nation. It increases the risk appetite not only among the investors but also among the researchers & businesses. The other known benefits are employment generation, enthusiasm and participation by the targeted audience, push to the economy, encouragement to imbibe out-of- box thinking, acceptance to try and fail attitude.

However, to take any research technology from lab to land requires one to face challenges and uncertainties at every step. Thus, the assessment of associated risks is an important exercise.

Risks can be categorised into many types, for example, risks associated with market readiness of any technological innovation. Several times, these risks involved are beyond ones' control. For example, policy changes, environmental changes, political changes are some factors.

Many a time, the effort to convert a research technology into a product or service is so high and is hard to envisage at the commencement of such innovative projects. This, at times, leads to the closure of such translational efforts. Thus, R&D in any organisation is always seen as an investment and many a times even gigantic organisations with deep-pockets in R&D fail to bring innovative products to the market. A product may have all the top-notch technology in it, but the supporting eco-system may not be in existence.

In the current era, the rate at which the technological advancements are happening is phenomenal, thus, the timing & entry to the market is a key factor in determining success. Hence, the risks are high and failure in real sense becomes dependent upon one's definition of success.

Given that the associated risks are so high, it is of utmost importance for the Hub to involve expertise of high quality in the Hub Governing Body (HGB), so as to guide on the technology roadmap with its expertise. The Hub cannot afford to take a risk on managing the initiatives and its execution. Thus, special attention shall be given by the Hub to ensure hiring of most appropriate TIH-team members.

IIIT-H will be identifying the required faculty members, who will be spending 80% of their time, energy and focus on the TIH activities. Other focus will be on bringing meaningful partnership at local, national and international level. This will enable cross-fertilisation of ideas across organisations and incorporation of best practises.

Investments in research & commercialisation may not yield results always, because by the time the product is ready, they need may change. To counter these risks, the Hub will identify challenges which have long-term insights and select partners who are leaders in technology and have long-term roadmap.

There is always a possibility that even the experts may not be able to predict the risks involved. But the Hub will try to periodically review all the activities at different levels through its executive committee and with Annual Comprehensive Review team. These reviews will help to maintain an undivided focus of the TIH on its expected targets and well-defined outcomes.

14 Outcomes

The proposed TIH-Data has the following key functions:

- **Technology Development**: Envision directions of potential research advances as well as their application opportunities in the broad area of Data. The technology development activities will be carried out by research groups/teams in different parts of the country including the team in the host institution.

  Key outcomes of TD over the first five years shall be:

  1. 700+ people trained under HRD & Skill development programme.
  2. 390+ budding & eminent researchers to be awarded fellowships (UG, PG, PhD, Postdoc, Faculty & Chair Professor).
  3. 27+ No. of Technologies (IP, Licensing, Patents, etc.).
  4. 95+ Publications, IPR and other Intellectual activities.
- **Entrepreneurship Development:** Entrepreneurship developed by research teams often need to be worked on by a translation team to make it ready for licensing to companies or to start-ups. This involves creating a development team of engineers and product managers who liaise with the technology creators and potential customers.

  Key outcomes of ED over the first five years shall be:

  1. 18+ Technology Products to be developed.
  2. 25+ EIR under pre-incubation programme.
  3. 36+ start-up's & spin-off companies.
  4. Grand challenges & competitions, as per the theme annually.
- **International Collaboration (IC)**: International collaboration with other leading academic institution & industries will help in bringing in new perspective into the Hub.

  Key outcomes of IC shall be:

  1. Bring in 5 + meaningful collaboration with funds and other resources being involved from both the sides.
  2. Launch international exchange programme for students, professionals, researchers, EIRs, entrepreneurs, etc., with international collaborators.

TIH-Data will support the following across a wide variety of educational institutions in a transparent manner.

HR Development

- **PhD Fellowships**: The Hub will establish special PhD fellowships in the area of Data and its applications such as mobility, healthcare and smart cities. 37 PhD candidates (125 Man years) will be distributed across educational institutions.
    a. Fellowships will be aligned to the larger theme and goals of the Hub.
    b. Fellowships will also be aligned and supported by the industrial research groups we collaborate.
- **Post-Doctoral Research Fellowships:** This Hub will also help in supporting ten postdoctoral fellowships at a time in the areas of interest. Postdocs will help in taking the academic research further and make deeper impact. They also technically interact with the Industry and end users.
- **Encourage High Quality Publications**: Top tier publications and high-quality academic outcomes from partner institutions will be encouraged. While the Hub maintains the deep technology focus, hub also aims to make deep academic impact with high quality publications. They are specially encouraged to strengthen the IP of the Hub/projects.
- **Small to Medium Research Grants:** A number of small to medium size projects will be supported by the Hub for:
    - Encouraging new institutions to do research in Data-specific areas.
    - Specific short-term research problems.

  Typically, these projects will be of 10-15 L per annum.

- **Technology Development Projects:** Research that lead to impactful technology will be supported based on the peer reviewed proposals and managed through regular evaluations. Each of these projects (worth 40-50L per annum) will be for 2-3 years, with process to sanction extensions.
- **Common Infrastructure for Research**: The Hub will establish common infrastructure for research to help wide range of collaborations who individually could struggle to establish and maintain separate infrastructure.

Data Foundation – Data set for Research & Translation – accessible to researchers, startups, industries, etc.

There will be a special effort in creating data sets with focus on technology verticals across healthcare, mobility, building, India-specific use cases, disaster management, agriculture, etc., in the Hub through the Data Foundation.

The short- and long-term goals are summarised below.

Table 17: Short- and Long-Term Goals of TIH-Data

S.No	Activity	Target					
		1st Yr	2nd Yr	3rd Yr	4th Yr	5th Yr	Total
1	**Technology Development**						
a)	No. of Technologies (IP, Licensing, patents, etc)	0	3	5	10	12	30
b)	Technology Products	0	3	3	6	8	20
c)	Publications, IPR and other Intellectual activities	10	10	20	25	35	100
d)	Increase in CPS Research Base	5	10	15	20	25	75
2.	**Entrepreneurship Development**						
a)	Technology Business Incubator (TBI)	1	-	-	-	-	1
b)	Start-ups & Spin-off companies	0	5	8	10	15	38
c)	GCC – Grand Challenges & Competitions	0	1	1	1	1	4
d)	Promotion and Acceleration of Young and Aspiring technology entrepreneurs (PRAYAS)	1	-	-	-	-	1
e)	CPS – Entrepreneur in Residence (EIR)	2	3	5	8	10	28
f)	Dedicated Innovation Accelerator (DIAL)	1	-	-	-	-	1
g)	CPS-Sees Support System (CPS-SSS)	1	-	-	-	-	1
h)	Job Creation	50	500	1500	3000	4500	9550
3.	**Human Resource Development**						
a)	Graduate Fellowships	20	40	60	80	100	300
b)	Post Graduate Fellowships	10	10	10	10	10	50
c)	Doctoral Fellowships	5	8	8	8	8	37
d)	Postdoctoral Fellowships	1	2	2	2	2	10
e)	Faculty Fellowships	0	3	2	0	0	5
f)	Chair Professors	0	3	2	0	0	5
g)	Skill Development	80	100	120	200	250	750
4.	**International Collaboration**	0	1	1	2	2	6

15 Evaluation

15.1 Why Evaluation

Regular evaluation is key to the systematic progress of the Hub and achieving the desired goals. Evaluation involves the overall process of looking at the data and providing constructive feedbacks for regular growth of the Hub. This also helps in properly documenting the state of the Hub and comparing with what is projected and finding the discrepancy if any. It helps the larger hub activities in -

- Providing feedback on the performance and suggest any corrections required.
- Identification of problems and challenges, uncertainties early on the project and explore for the possible alternatives.
- Monitoring the achievements and goals of project
- Absorbing the users and stakeholders' suggestions and views into the project process leading to better participation and success.

An evaluation process also helps in:

- **Accountability**: and transparency of the operations to the Governmental and funding agencies.
- **Proper implementation of the operational**: protocols Regular light weight evaluation protocols help in proper implementation of the operational models.
- **Strategic planning and management:** Evaluation helps in strengthening the strategic planning and management process.
- **Improving the capacity**: Capacity building, self-reliance and confidence building in the team takes place through this.

15.2 What to Evaluate

Different parts of the Hub activities need evaluation and monitoring, to name the major ones:

- **Input and Scheduling**: This involves aspects related to regular operational schedules, timely processing, finances, etc.
- **Process and Models**: The Hub also follows the process and models described in its mandated models. However, it is important to monitor the adherence of the process.

- **Outputs**: Many objective targets are mentioned and planned. This monitors and verifies the objective matches.
- **Outcomes**: Objective outcomes need to be interpreted and seen how it connects to the outcomes and more conceptual goals that are set.
- **Impact**: Finally, the impact of the Hub needs to be measured and evaluated. This is often long term.

Evaluation of the schemes and the respective details are listed in detail in the Chapter 5 of the master DPR provided by the DST. Here we show only one example situation to explain. Consider a component of PhD Fellowship. This is the programme that we are evaluating. Inputs involve the funds, availability and streamlined delivery, and the quality of the PhD fellowship owners. Outputs involve the young professionals trained in the domain of the Hub, PhD degree holders leading the technology in India Outcomes also involve the technologies, IP, Publications and knowledge that these fellowship holders bring to the community.

15.3 How to Evaluate

Evaluation is a regular activity. This implies that the project team is regularly ready for the consolidation of the activities, presentations and receiving feedback. To help that:

- **Internal Almanacs/Schedules**: Internal schedules are prepared and refined at time to time to monitor the smooth functioning of the Hub.
- **Blogs and Activity Notes**: Individual activity holders create simple blogs and logs for summarising the activities.
- **Monthly Activity Summary**: A curated monthly progress report is created from the different notes.
- **Quarterly Reports**: Quarterly reports are prepared with the following objectives:
    i. To internally review the Hub activities by comprehensively looking at the state of the Hub.
    ii. To refine schedules and identify corrective actions.
    iii. To give the status of the Hub to all internal participants.
- **Annual Report**: Annual report of the Hub is publicly available. This also helps in the interface with external stake holders and collaborators.
- **Project Review**: Report Administrative review reports covering activities, outputs, achievements and finance are created at regular intervals for this.

These notes or reports along with formal meetings and reviews provide feedback to the Hub team on the directions, pace and models. Some of these also help in introspection for the Hub leadership team.

15.4 Evaluation Team

Detailed information about the various committees are described in Chapter 9 along with the management of the Hub. Here we list the models in which evaluation and monitoring takes place in the system and the type of expertise that we propose to bring in.

- **Regular Processes Checks**: A number of regular evaluation protocols are integrated to the day to day working models of the Hub by introducing the process documents, policies and guidelines.
- **Internal Reviews**: The management and academic leadership of the Hub meets on a regular manner to discuss the specific issues, evaluations and decision making. This is a regular activity to build cohesion in the team and introducing the regular corrective feed backs.
- **Administrative Reviews**: Audits and administrative reviews are also followed for adherence to the policies and guidelines, more specifically in operation.
- **Academic and Technology Reviews**: External views and inputs are brought in through academic experts from India and abroad.
- **Translation and Technology Reviews**: To improve the impact of the Hub and making sure that there is a direct visible impact in the technology space, views from industrial experts, practitioners and entrepreneurs are brought.
- **Governance and Mandatory Reviews**: The governance model of the Hub also sets up a number of committees and regular evaluations to look at the process of strategic
- **Hub-Project Reviews**: There are also regular review from the funding agencies that is planned for evaluation and guidance of the activities.

Summary

The Hub governing board will provide overall directions of the Hub and will have full financial and administrative powers. The technical advisory council will assist the HGB on technical matters, and the executive committee will assist in the operational/administrational aspects of the Hub. Several committees will also be set up primarily for other evaluations and monitoring.

- **Performance of the Hub**: As mentioned in Chapter 9 (Management), a committee will be formed that will monitor and evaluate the activities of the Hub periodically (at least once in a year).
- **Project evaluation**: The technology development activities of the Hub involves ‘call for proposals’ from researchers in the country. These proposals need to be evaluated, yearly progress need to be monitored and finally the projects has to be assessed when closing. Different committees comprising domain experts (both from Industry and academia) will be setup and will meet at regular intervals.

- **Others committees**: The Hub plans to encourage research at different levels starting from basic research to product delivery. This requires identifying and supporting technology at diverse levels of maturity. Committees will be constituted that can assess these technologies at appropriate levels for further development.

16 Summary and Conclusions

This DPR laid out a proposal and plan to establish a Technology Innovation Hub on Data (TIH-Data) at IIIT Hyderabad. Data will continue to be at the core of the modern AI-dominated world. The NM-ICPS Mission and several hubs can benefit from varied data and tools created by the Hub. Academia, research organisations, government, and startups are among the sectors that can exploit good quality data about the society around us that this hub creates.

The design of TIH-Data includes all components necessary for successful outcomes. It includes a large effort to collect, collate, validate, annotate, and host a variety of data from domains like Health, Transportation, Buildings, and others through the Data Foundation.

TIH-Data will also undertake applied and translational research and development so that the data-driven technologies created by the mission can be licensed or transferred to industry and startups. TIH-Data has as its largest component a strong programme for research and technology development, which will be based on open calls and evaluation of proposals from the larger community. The Hub also proposes to promote expertise by supporting different segments of individuals: students, PhD scholars, postdocs, faculty, etc. Promoting innovation and supporting startups working in data-driven technology areas have received adequate attention in the proposal. To become a globally recognised powerhouse in these areas, strong collaboration programmes are planned with internationally renowned institutions and experts.

The DPR outlines effective management and oversight mechanisms to ensure the proposed TIH can achieve its objectives smoothly. The academic head will lead the effort and will be assisted by a Hub Management team led by a professional CEO/COO. A Technical Advisory Council including outside experts will lead the technical directions for technology development, collection of data, etc. We also propose an annual comprehensive review of the Hub by a team dominated by outside experts to advice the Hub Governing Board on the progress as well as on any course corrections that may be needed.

With the above plan and the necessary ingredients in place, the TIH Data can become a leading entity in the space of data-driven and ML-based research and technology development, playing a truly enabling role within the National Mission on ICPS and for other AI-related efforts in the country. It is the fond hope of IIIT Hyderabad that the TIH Data and the institute will have extraordinary international visibility for these efforts. The strength of its research team in related areas, experience with translational research and startup seeding, and the self-sufficient financial model make IIIT Hyderabad uniquely placed for success in this endeavour.

Annexure A: A Deep Dive into the Challenges to be Addressed

This annexure includes an in-depth discussion about some of the challenges to be addressed.

Mobility

The space of mobility offers numerous challenging problems that still need solutions. This space includes the outdoor world such that involves modern and intelligent transportation systems and the indoor spaces such as offices, warehousing industry and hospitals. Data-driven technologies with data gathered from diverse sensors when perceived by an autonomous agent and acted upon by such agents offers immense scope to improve the quality of life.

Consider for example city-scale transportation that causes numerous accidents and fatalities that causes immense anxieties and despair to families as well as to the nation's economy. Technologies and data sets that map the urban roads and traffic and urban infrastructure in general can go a long way in reducing accidents and fatalities. Advanced driver assistive systems that include technologies that recognize state of the driver and data sets relating to these aspects are critical for enhanced road safety and paves way towards safe mobility in the transportation space.

While being the causative mechanism of numerous accidents it is undeniable that the transportation industry has become increasingly important for the rapidly growing world economy, and it has strong connections with our lives. Among such technologies are driverless cars that pose as perfect models for promising technology engines. These cars have received much attention and appreciation from high-tech companies. Studies have shown that they are statistically much safer than traditional vehicles. As we move towards a futuristic scenario involving automated vehicles as a mode of transportation, a space of problems that involve centralized as well as distributed control of such vehicles over V2V communication engine becomes inevitable.

On the indoor front the explosive growth in e-commerce, changing consumption patterns and rapid adoption of omnichannel distribution models by companies, especially in the backdrop of the Covid-19 outbreak, are driving demand for warehousing and logistics space across the country. Several global firms are also likely to shift manufacturing to India to de-risk their supply chains. And Government initiatives like Make In India, GST, FDI policy, improved transportation infrastructure are expected to accelerate this trend. Some estimates predict that the warehousing and logistics industry in India will grow at 35% in 2021 compared with the earlier industry estimate of 25%.

However, there are several challenges the industry must overcome. In the short term they need to bounce back to full operating capacities while ensuring compliance to safety requirements. Hiring and retaining labour is getting tougher and costlier. Processes have to be re-engineered for contactless deliveries. Rising customer demands for shorter delivery times and lower costs have to be met. And so on. All these factors are likely to further squeeze their already stressed operating margins!

The COVID crisis has also showcased the need for automated payload transfer and delivery mechanisms. Autonomous agents capable of such delivery that limits human intervention during pandemics is also the need of the hour. Such agents could be effective in both outdoor spaces as well as in indoors such as hospitals.

Problem Areas

Connected Vehicle Data Set: To build technologies that provides for V2V (Vehicle2Vehicle), V2C(Vehicle2Cloud), V2P(Vehicle2Pedestrian) communication. Build data sets using these technologies that provides for various intrinsic and extrinsic data such as state, velocity, acceleration, angular velocity and jerk (intrinsic) and inter vehicle distance, relative velocity between vehicles (extrinsic). Leverage such data for analysis of traffic flow, safety, congestion, cooperative navigation and control, intersection control of platoons, etc.

Resource Allocation and Management for Mobility: Computation offloading is a very important function for connected vehicles, whereby an entire application or a part of the application execution is offloaded using V2I/V2C sup- port to a fog node/cloud. The results of the execution are then sent back to the vehicle.

Several works have looked at this problem from the optimisation viewpoint, where an optimal offloading strategy is derived considering various objectives such as total energy consumed in executing an application and time taken to perform the execution. These have been performed under various constraints. There are very few works, which have looked at the computation offloading of an ADAS application such as Lane Keep Assist (LKA). The offloading strategy between vehicle, fog and cloud was explored in terms of the stability of LKA and performance. These strategies also perform the task of resource allocation on the fog/cloud architecture in order to offload the computation. Typically, the resources allocated are the virtual machines and the memory resources on the fog and the cloud nodes.

Most of the previous works on computation offloading made the assumption that the execution of an application and data transfers complete before the connected vehicle moves out of the coverage area of the fog node. However, in reality this cannot always be ensured. This makes it essential to have a mechanism to migrate the execution instance from one fog node to another as soon as the vehicle moves from the coverage area of the former to the latter. We look forward for methods that derive an optimal approach to perform the above task that will explore an efficient combination of virtual machine migration with new virtual machine creation on the target fog node. To add to the above problem, a connected vehicle may encounter several spots on its path with no connectivity

to the fog node/cloud. This seriously impacts the performance because of the inability to offload computation or send/receive data We envisage to solve the problem by using vehicle resources to perform the computation and secondary infrastructure such as unmanned aerial vehicles (UAVs) to send/receive data.

Development of Connected Autonomous Vehicles (CAV): Prototype minimally a pair of connected vehicles that showcase drive by wire/full autonomy even as they exchange data between them on various intrinsic parameters. Analyse and glean extrinsic statistics from the intrinsics exchanged between vehicles. Scale up to a platoon of CAV by transfer of technology. Although lot of theoretical concepts have been developed on Vehicular Ad-hoc Network (VANET) for autonomous vehicles, there is dearth on proof of concept demonstrations on the ground, especially in Indian conditions. This is essential as the theoretical concepts make too many assumptions for simplifying analysis, which might not work in real conditions. For example, channel models assumed may be too simplistic while the actual conditions can be too harsh.

Developing Heterogenous Vehicular Network (HetVNET): In autonomous driving systems, various types of messages or information have varied Quality of Service (QoS) requirements, such as low latency and high reliability for safety messages, and high data rate for non-safety multimedia applications. Again, due to high mobility and the dynamic change of network topology, it is difficult to provide satisfactory services through only a single wireless access network, such as Dedicated Short Range Communication (DSRC) or LTE-V. Each technology has its own shortcomings. For example, DSRC is originally designed for short-range communications without considering the infrastructures already deployed. Although cellular network covers a large area, it might not effectively support real-time information exchange because of the delay caused by the centralised control. Therefore, it is important to work on Heterogenous Vehicular Network (HetVNET), which will integrate different types of wireless access technologies and support real-time message dissemination to meet various communication requirements for autonomous driving. Similarly, a single vehicle has issues such as infeasibility, unreliability, and efficiency because of limited perception. Therefore, a cooperative autonomous driving framework based on HetVNETs is needed, where multiple vehicles can share information both locally for traffic safety and globally for traffic efficiency.

Development of Self Driving Technologies: Prototype algorithms and proof of concept (POC) demos for self-driving such as state estimation, sensor integration, sensor fusion, mapping, navigation, sensor suite instrumentation and collision avoidance. This phase involves both hardware, instrumentation and algorithm development for autonomous vehicles, which will then form the backbone for the development of Connected and Automated Vehicle (CAV) technologies.

Development of Cooperative Driving Technologies/Cooperative Driving Stack: Making use of existing or prototyped Vehicle-to-Vehicle (V2V) technologies we develop the cooperative navigation stack that enables self-driving vehicles to seamlessly navigate in a space without collisions. This stack of algorithms can run distributed on each vehicle or centralized over a

cloud. To drive a fleet of vehicles autonomously or in a distributive sense is the ultimate aim and the holy grail of this effort. The stack of problems includes but is not limited to:

1. **Cooperative control with engine limitation**: An important issue in cooperative control is engine limitation. Therefore, there is a need for a control strategy optimising engine capacity and performance.
2. **Cooperative control with vehicle dynamics uncertainty:** Cars are essentially underactuated systems, while state-of-the-art do not consider such system dynamics in platooning problem. In addition, lack of precise knowledge of system dynamics as an important source of uncertainty. Hence, an adaptive cooperative control solution of driverless cars addressing underactuated dynamics is of high importance.
3. **Heterogeneity in vehicles:** Most of the state-of-art platooning problems rely on homogenous vehicles. However, one cannot deny heterogeneity due to user preference. For example, a family car in a group of sports cars or vice versa. Such situation leads to a host of problems such as dissimilar vehicle dynamics, engine capabilities, speed limits, and so on. Therefore, a comprehensive control strategy is required to deal with such situations.
4. **Distributed algorithms for reducing traffic congestion:** Time efficient and collision free driving schedule and corresponding vehicle trajectory among a group of vehicles. While many simulation frameworks for the same have been developed their portability to a real vehicle platoon is an unknown. We seek proposals that enable and are able to showcase portability of such distributed algorithms on an experimental autonomous vehicle fleet.
5. **Centralised cooperative vehicle intersection control:** This is a centralised solution for safe trajectory for each vehicle while crossing an intersection. It assumes perfect communication model. While literature poses it as a constrained nonlinear optimisation in the presence of perfect communication porting, such algorithms on a real vehicle fleet are immensely challenging. Indeed, it is the holy grail of this effort to centrally/distributively control an autonomous vehicle fleet at an intersection.

Development of Security Stack: Driverless cars heavily rely on information sharing over wireless medium. This makes such connected devices susceptible to cyber-attacks, where gaining access to sensitive information may cause severe consequences not only to platoon structure but also to traffic. Cyber security is, therefore, an extremely big problem for driverless cars which is a less explored area of research.

Fog Level Decision Support for Vehicle Platooning: Vehicle Platooning is an application where vehicles cooperate and move in groups. This has shown its benefits in reducing fuel consumption and gas emissions, while improving the safety and efficiency of transportation. Each vehicle keeps a safe distance from the vehicle in front and behind it and also is sent messages from the leader vehicle regarding manoeuvring the platoon. The stability and efficiency of platoon depends on the speed and reliability with which messages are exchanged among the vehicles in the platoon. A lot of research is ongoing in order to explore the communication technology for exchange of the messages.

However, one important problem that has not been explored well is the system-level support for forming platoons. It is critical to make decisions for the formation of efficient platoons. A vehicle may have many choices in merging into a platoon and thereby reap the benefits of a platoon. There are certain system-level factors such as overall reduction in fuel consumption and throughput of traffic, which need to be considered while making decisions regarding the target platoons a vehicle must merge into. We seek proposals attempt to solve this problem of system level platoon formation strategies given a set of vehicles, their path and other relevant vehicle parameters. To derive an efficient platoon formation, the decision module must be privy to a lot of data in a local area of the network. Therefore, we will leverage the advantage of a fog node in order to quickly run the algorithms to make decisions on the platoon configurations and finally communicate the configuration to the vehicles.

City Scale Road and Infrastructure Mapping: India has a huge road network, which are managed and maintained by different departments. Proper monitoring of the condition (say the quality of the roads) or state of the traffic infrastructure (e.g. sign boards) is practically impossible to do at scale on a regular, repeatable manner.

Technology to improve Road Safety and Efficient Mobility: Indian roads witness a number of accidents and deaths every day, whether it is highways, urban or rural roads. Technology can play an important role in making our roads safer. Sensors and cameras connect the physical world to the digital compute. Algorithms and solutions that make use of the sensor data and detect road anomalies such as presence of potholes and ditches, missing traffic signs and signals are the need of the day.

Advanced Driver Assistance Systems: Driver assistance systems that can provide both inward and outward looking states find imminent applications. Inward systems estimate driver states while outward systems provide for integrated estimation of on road states including vehicles, pedestrians, state of the road. Data sets that capture driver states, driver gaze and solutions and algorithms based on such data are sought in this proposal.

Mobility Solutions for Warehouses: Distribution centres are getting bigger and bigger, especially in e-commerce. Mapping these facilities using vision sensors takes considerable time and computing resources. How can this be done rapidly and efficiently? Warehouse environments are dynamic. There are several workers and other material handling equipment like forklifts operating alongside mobile robots. How can localisation and dynamic obstacle avoidance be made more robust to prevent accidents? Modern distribution centres handle very high volumes of products and orders. This would necessitate a large fleet of mobile robots, which in turn could increase congestion and likelihood of deadlocks in certain areas. How can we model this and develop intelligent traffic and fleet management capabilities in robots? Bulk of the work in warehouses involves order fulfilment. Workers manually identify items, pick and pack them. This requires high levels of dexterity and is still incredibly challenging for robots. Current technology is either too slow or too expensive. How can we automate this cost effectively? How can

implementation of robotics in warehouses be made simple and fast? How can we improve robot worker interface and cooperation?

Towards a Warehouse Data Set: While there exists a variety of problems to solve in the warehousing space involving perception as well as navigation not all researchers have access to real warehouses. A warehouse data set in this regard is extremely fruitful and useful. Data-driven warehouse technologies include the problem of free space estimation in racks, item counting and binning, pose and depth estimation to rack space items, scene understanding in warehouses, and so on.

Developing unified solutions to solve these problems typically requires capture and study of multi-modal sensor data in real warehouse settings. Given the logistical difficulties involved, an alternative strategy could be to study these problems in a synthetic yet realistic warehouse setting and transfer the learnings from this controlled setting to the real-world counterpart.

The approach of developing solutions in synthetic worlds has found success in many other domains such as development of autonomous driving controllers and engineering machinery control. We follow the same approach. However, our challenge is compounded by the fact that no readymade synthetic environments exist. To address this, we solicit proposals as well as propose to create realistic synthetic warehouse environments in various configurations and containing various warehouse items, with various associated attributes (lighting, flooring, human presence, machinery, etc.). In addition, we plan to have embodied sensor-equipped synthetic robot agents navigating these environments. Completing these requirements (warehouse generation and embodied agents) sets the stage for tackling the scenarios mentioned previously.

Mobility for Surveillance, Search and Rescue: Enhanced Mobility Solutions are critical during pandemics, disaster, surveillance, crowd monitoring and many such applications. Such solutions enhance the safety of smart cities. Proposals are sought in such areas of using autonomous/guided agents for delivery, urban infrastructure surveillance, surveillance of campuses and communities. Interesting problems in this space include cooperative payload and delivery by cooperating agents. Cooperation between agents with multiple modes of locomotion for task accomplishment, human in loop systems deserve lot of attention.

Mobility Based Delivery Solutions in Pandemics: With Covid-19 like situation prevailing in India and the rest of the world, autonomous robots can significantly improve availability of services 24x7 while adhering to lockdown, distance and safety requirements. Some basic services that are required during these periods are availability of medicines, dispatching groceries to houses/apartment complexes and spraying disinfectant along the roads, etc. These activities can be done by autonomous vehicles with safety as the paramount concern. In this proposal one key activity is the use of Unmanned Aerial Vehicle (drone) and unmanned ground vehicle (UGV) for delivery of essential and non-essential items under a certain load requirement. In order to enable such mechanism, the UAV and UGV need to cooperate on deploying locations, delivery locations, coverage, and take-off/landing.

Moreover, as the use is increased one has to schedule these vehicles to meet the demand. The service may depend on the demand from different regions and hence the vehicles needs to be scheduled accordingly. A potential approach for schedules and also for robust landing of the UAV on the UGV (stationary or moving) can be based on data-driven techniques.

Data-driven approaches for Bathymetry mapping: Bathymetry mapping of water bodies is essential for sustained maintenance of water resources that have significant socio-economic implication on the society. This becomes intensive in rural region where the water bodies are neglected and hardly few measurements are carried out in years. Moreover, the bathymetry of these water bodies changes with climate and it is essential continuously monitor them. In order to obtain the bathymetry, one has to sample the water body which is time consuming and given the fact that Indian has hundreds of water bodies in a city, the task is very tedious and expensive. One potential mechanism is to employ an autonomous surface vehicle (ASV) like a boat with the echo sounder for the mapping. However, mapping water bodies are in tens of square kilometres is also quite difficult. To cater scalable and low-cost bathymetry mapping solution, one needs to employ data-driven sampling techniques such that with fewer samples one can approximately map the waterbody. These adaptive sampling techniques are highly scalable and hence will allow largescale mapping of water bodies in few days. Also, with rains, the bathymetry changes and hence the water capacity of the water body which implies periodic sampling needs to be performed. As data-driven techniques are scalable and low cost, they can be employed periodically to obtain accurate maps.

Healthcare

The field of healthcare is undergoing significant changes with the accumulation of large volume of biological and physiological data due to the technological advancement. A key challenge is to improve the patient care, disease prevention and understanding of human health and disease by learning from these data. This requires development of computational tools for analysing and interpreting data. Machine learning approaches are showing great potential in learning the patterns from the complex biomedical data. The incorporation of the data-driven models into the clinical care is expected to transform the healthcare sector.

A traditional approach to clinical decision-making by physicians depends on symptoms, lab tests, medical imaging, etc. There is considerable need to make the clinical decision-making more accurate, timely, and personalized. Interpretation of data like medical images are based on manually designed features obtained from domain-specific knowledge and requires involvement of specialists like radiologists/pathologists. The timeframe involved in interpretation of data like histological images remains longer and there is inter-observer discordance. Complex morphological information remains unutilised in unaided approach. The early diagnosis/prognosis of diseases like cancer are important for appropriate clinical management of patients. The recent COVID-19 situation is showing how AI can be used to detect COVID-19 from X-rays and computed tomography (CT) images. There is a need for

developing technologies to automate medical image analysis of different data sets like CT, magnetic resonance (MR), positron emission tomography (PET), mammography, ultrasound, X-ray, microscopic images, etc. Deep learning methods for computer vision tasks (recognition, localisation and segmentation) on natural images and video are being applied to medical image analysis without handcrafting of features. This is helping in the detection of anatomical/cellular anomaly detection and segmentation tasks and in diagnosis of diseases. However, the availability of medical image data is limiting compared to natural images. There is a need for creation of Indian-centric data sets and in histopathology, creation of digital image library is required by scanning the whole slides. There is also need for clinician-computer interaction to improve the accuracy of prediction.

Interpretation of physiological time-domain data (1-D signals) such as electromyogram (EMG), electrocardiogram (ECG), electroencephalogram (EEG), etc., are not easy due to noise, artefacts, varies between individuals and for same person over time. Visual inspection of processed data to detect anomalies is long and tedious. The standard computational pipeline for analysis of physiological data signals are signal processing, feature extraction and classification using regression models. The deep learning methods is providing opportunity for end-to-end learning of physiological data signals to predict events like seizures and cardiac arrest. Data from continuous monitoring of health and well-being using wearable and implantable sensors is also accumulating which also requires technologies to detect abnormal human activities and alert the user. Further, recent digitalisation of healthcare data has led to creation of electronic health records (EHR), which raises the hope to improve the quality of healthcare by predictive modelling. Natural language processing (NLP) approaches are helping to overcome issues such as high dimensionality, sparseness and confounding factors present in EHR and helping in extracting useful information from EHR. Deep learning model trained on patient characteristics, demographics, diagnoses, medication, lab tests, etc., can predict disease severity, hospital readmission and treatment outcome.

Deep learning is also being used to address the problems in genomics with application in medicine. Genomic data presents both opportunities and challenges in terms of data storage and processing it. The analysis of genomic data can help to determine how variations of DNA influence disease risk. It is helping us to move away from ‘trial and error’ approach in traditional medicine to precision medicine by taking into consideration how genetic variations affect drug response. The paring of EHR with genetic information will become a standard. The emergence of such data sets at the population-level will boost the power of machine learning approaches. Further, high throughput measurements of cell variables (DNA, RNA, proteins, protein binding to DNA etc.) by next generation sequencing (NGS) and proteomics are providing data at multiple levels, which can be integrated using deep learning framework to interpret the genome and model the genotype to phenotype relationship in heath and disease. Further, it can help in early diagnosis and prognosis, and identification of new disease subtypes from the perspective of genes and proteins. With exponential decrease in sequencing costs, assessments using NGS data is becoming more and more routine.

The cost of drug development is increasing with higher failure rate during clinical trials. There is a need to develop technologies that transform the drug development process. The emergence of publicly available databases such as ChEMBL (a manually curated database of bioactive molecules with drug-like properties), PubChem BioAssay and BindingDB (a public, web-accessible database of measured binding affinities) are increasing the scope for compound property and activity prediction by data mining. The prediction of the toxicity is also becoming feasible with development of the high throughput toxicity assay. Machine learning approaches to generate new molecules with specific property, to predict products given a set of reactants and to predict reaction steps to produce products are potential alternatives to traditional rule-based approach. Further, assessing the interaction between drug and target is another area where the deep learning methods have potential to speed-up the process lead compound identification. These areas of applications suggest that there is a need to leverage AI in healthcare for improving the patient care and treatment. Further, there is also a need to improve the interpretability of models and benchmarking methods to evaluate the performance of these methods.

Objectives

AI/ML has the potential to significantly aid in healthcare and medicine and is perceived as the one to change the way diseases are diagnosed and new drugs are discovered/developed. Therefore, one of the focus areas of the Hub is Healthcare and Medicine. General areas include disease diagnostics, drug design/development and personalised treatment.

Problem Areas

India-specific medical data sets: Creation of India-centric medical data sets are essential for developing solutions that works efficiently. Collaborations with hospitals, diagnostic centres and healthcare will be required to collect clinical data including medical image data sets that can be made available to other researchers. Further the collaboration with CSIR Institutes like CCMB (Centre for Cellular and Molecular Biology), will help to generate genomic data set that can be used for disease diagnosis.

Efficient disease diagnostics from medical images: Diagnosing diseases is a complex task, is that of an expert who undergoes several years of training and is a time-consuming exercise. AI/ML potentially can process the images and identifies the presence/absence of a disease as accurately as radiologists and pathologists hence making the process more efficient. AI systems for diagnosing diseases based on histopathology and radiology images are required.

Novel methods for early stage detection of diseases: The advent of next generation sequencing techniques (NGS) has led to the creation of valuable resources on molecular characterisation of diseases. Tumour sample, for example, can be profiled at multiple levels (genomic, epigenomic, transcriptomic and proteomic levels). New methods of diagnosis by integrating multiple data sets will be developed. Some of the diseases that will be addressed are neuro-psychotic disorders, autism, cancer and stroke.

Drug design and discovery: Drug design and development is a complex process that takes a long time and is expensive. AI/ML methods are capable of cutting down the cost and time by making the search of new chemical entities more efficient. Development of novel and accurate drug design methods and pipelines based on AI/ML algorithms that would make identification of new chemical entities, which strongly inhibits the target proteins, more efficient, thus cutting down time and cost. Development of explainable AI/ML models to predict reliable pharmacokinetic properties that would enable spotting failures early.

AI for pharmaceuticals: Once the patent of the original drug expires, generic drugs are manufactured which share the same active component of the drug, but are different in other features such as formulation. Given that India hosts top generic pharma companies, the Hub will work with Indian pharmaceutical companies towards developing AI/ML methods for predicting polymorphs and its effects on solubility profiles of new formulations that will help in development of new biosimilars more efficiently.

Wearable sensors: The future of healthcare is addressing health issues before one even is aware of a condition. Modern AI enabled wearables are pushing the boundaries within the healthcare system in several aspects such as remote health monitoring, remote diagnosis, tracking/improving fitness, imaging, telemedicine, etc. One of the focus areas of the Hub is development of AI enabled wearables for health monitoring and diagnosis (cardiovascular, diabetic, etc).

Proactive public health

- Risk profiling & 360-degree view of data.
- Data based targeted preventive measures.
- Predicting outcome of health initiatives.
- Contributory effects of broader Ecosystem.

Health service optimisation

- Ayushman/Arogyasri insurance service optimisation.
- People & Facility use optimisations.
- Supply chain optimisations (medicines).
- Asha workers' efficacies.

Person/village-specific 360-degree health & social strategy

- Culture/geo specific interventions.
- Equitable benefits distribution.
- Vulnerable groups.
- Insights from seemingly unconnected data sets (poverty, NFHS, census, DLHS, literacy).

Collaboration

One of the key areas of research focus of the Hub is in healthcare technologies. For such research to be effective and useful it is extremely important that we closely collaborate with hospital and medical centres. This is important in several aspects:

i. Acquiring diverse and representative medical data that form the basis of learning and evaluation AI/ML solutions,
ii. Getting critical inputs from domain experts in terms of the problems that could be faced in practice while developing healthcare solutions, and
iii. Doing field-trials of the solutions that are developed and collecting feedback to improve the solutions. We will be collaborating with hospitals of several specialities including radio-diagnosis, ophthalmology and oncology.

Buildings

The country is witnessing a rapid growth in the gross built-up area and there is a significant change in the design and operation of buildings over the years. This uncontrolled rapid urbanisation and ever-increasing per-capita energy consumption is widening the energy supply and demand gap. Issues such as climate change are adding to the strain of increasing energy consumption. This situation provides an opportunity for the country to leapfrog from the conventional building design and operation to *smart buildings*. A smart building can optimize the operations of buildings while ensuring comfort to the occupants in the changing outdoor and indoor environment. Adaptation of smart techniques can be enhanced by understanding the local challenges and developing and deploying appropriate solutions with potential scope for replicating in emerging tropical countries.

Design and operation of smart buildings is extremely complex and requires a multi-disciplinary and inter-disciplinary approach. Information technology sits at the centre and acts as a critical enabler.

Why smart buildings?

Buildings are an integral part of human life. Be it in an office, home, school, or a shopping mall, people spend most of their time inside them.

The primary requirements of a building are to provide safety, comfort and enhance productivity. However, buildings consume a lot of resources such as energy, water, land, air, material, and so on. They also generate a lot of waste and pollution.

According to a 2011 report by Center for Science and Environment (CSE), New Delhi, buildings can account for 40% energy use, 30% of raw material use, 20% of water use and 20% of land use. At the same time, they generate 40% of carbon emission, 30% solid waste, and 20% of water effluents.

With rapid urbanisation, the number of buildings is also growing rapidly which will only increase the stress on natural resources (such as water and non-renewable energy sources). This, in turn will result in significant increase in waste generation and pollution which will eventually have an adverse and irreversible impact on the global climate. *Therefore, it is important to treat a building not only as a structure but also as an operational unit. This operational unit needs to be managed smartly so that the resource consumption and waste generation are minimized while providing desired comfort level and services to its residents. This is where the concept of smart buildings appears.*

At present, there is no universally accepted definition of smart building. It means different things to different people. However, a common theme in most of the smart building definitions is on the significant use of IT components in making the building smart. Examples of such IT based components include sensor networks, software engines, and network enabled actuators. In a typical smart building, several sensors gather information regarding different parameters related to the building environment such as occupancy, light, temperature, humidity and radiant heat. Based on this data, software engines can not only learn the current state of the building environment but also anticipate the needs of its user by taking into account user behaviours and usage patterns. This information can then be used to modify different actuators such as air conditioners (AC), lights, and shades to maximize the comfort and safety of the users while minimizing the resource consumption and wastage. It is important that these modifications are automated since human interventions may involve significant efforts, cost, and inefficiency. To summarize, the main focus of smart building is towards sustainable resource usage and minimize the side effects with the use of technology, information, and data.

Smart buildings also acquire importance because they are going to be the essential components in the concepts of smart cities and smart grids. In 2015, Government of India has started the smart cities initiative under which 100 cities across the country will be developed as smart cities. One of the enablers of smart cities is smart grids for energy, water, information and so on. The smartness of the cities depends on the smartness of buildings which are nodes on these grids.

IT Research for Smart Buildings

Major Area	Topics
Wireless and IoT Communications	Sensor network and IoT data fusion algorithms
	Transactional Networks such as VOLTTRON
	Energy harvesting sensors
	Key management algorithms for data security
	Cognitive Radio networks deployment for dynamic and efficient spectrum utilization
	Indoor localization
	Power-line Communications (PLC) for communications and control
	System-wide, light weight communication protocols
Fault Detection and Diagnosis	Fault detection using time series data
	Regression strategies for fault diagnostics
	Real-time simulation and summarization
	Rule based fault detection and diagnostics
	Light-weight embedded agents for efficient, on-device Fault detection and diagnosis
Computing	Frameworks for enhancing programmer productivity including parallelization
	Optimization problems and algorithms
	Clustering and Decision Support Systems
	Distributed computing and multi-agent systems
	Image/Video processing and summarization
	Natural language/speech-based systems
Big Data	Storage strategies for acquired structured and unstructured data
	Visualization support with range queries, dashboard, human- and mobile-friendliness
	Recognition, Mining, Synthesis
	Predictive analytics including what-if studies

Building Design

Major Area	Topics
Early design optimization	Extending and deploying of early design optimization tool (eDOT)
	Inclusion of complex building geometry types in eDOT
	Seamless integration of eDOT to OpenStudio
Simulating smart aspects	Development of intelligent devices and control algorithm modules in EnergyPlus
Integrated design	Supporting BIM for holistic optimization of building design taking into account technical, commercial, sustainability and regulatory factors
Calibrated simulation	Real-time simulation and actual building performance comparison
	Certified software tool reducing the gap between simulated and actual building performance.

Smart Building Operations

Major Area	Topics
Water	Real time leakage detection
	Automatic Meter Reading (AMR)
	Real time quality management
	Smart grid in water distributed systems
Structures	Flexible building components
	Monitoring Liquefaction-Induced settlement of building for different soil system
	Real time structural health monitoring of buildings
Energy	Integrated controls of task and ambient lighting
	Impact of colour tuneable lighting on comfort, productivity and health
	Integrated HVAC controls for energy efficiency
	Model predictive controls for Chilled Beam/Radiant cooling
	Controls for air side economizer/night cooling
	Controls for desiccant based dehumidification using solar/waste energy for recharging
	Enthalpy based cooling optimization
	Cooling control for optimizing total cost of energy and water
	Optimize ground source heat pump operations
	Chiller sequencing optimization
	Controls for switchable glazing
	Plug loads monitoring/identification
	Controls for dynamic Facades
	Smart adaptive occupancy sensors
	Open loop /closed loop self-commissioning daylight sensors
	Fault detection, diagnostics and prognostics for HVAC system
	Mixed mode building operations management
	Managing Thermal storage
	Efficient management of indoor air quality
	Adaptive comfort-based space conditioning
	Interactive passive controls
	Advance display power management system
	Weather prediction for optimizing storage and operation
Safety	Real time fire modelling and evacuation
Security	Open-door hands-free access controls
Flexible spaces	Design and operation of flexible space
Comfort and productivity	Adaptive comfort-based space conditioning
	Controls for Personalized comfort
IAQ and health	Optimize selection of filtration media
Transportation	Controls for elevators and escalators
On-site renewables	Controls for optimize use of onsite renewable energy sources

Major Area	Topics
On-site storage	Controls for use of electric cars for onsite storage
Pollution	Controls for nocturnal light pollution reduction
Waste management	Tools for Waste management and disposal
	Tools for energy saving in treatment plants
Gardening	Drip irrigation control based on real time moisture content measurement
	Automate watering cycles, to monitor and manage nutrients, pH for vertical gardens

Smart Building Operations – Whole Building Level

Major Area	Topics
Integrated Operations	Integrated controls for building management
Grid interactive	Smart grid ready
Climate resilience	Development of climate resilience system
Building energy distribution	Controls optimization for DC grid
	Controls and optimized algorithms for prosumer of energy

Problem Areas

Dynamic Façades - **Illumination-Aware Façades using Environment Maps**: Automatic control of window shades is one of the most effective ways through which a smart building can utilize natural light and heat to conserve power for lighting and cooling. However, effective control of transparency of façades requires the measurement of light reaching every part of the building. This requires a large array of sensors deployed across the building. This work aims to replace such a network with a single camera that captures the illuminating environment from the top of a building and a computational model, which calculates the natural light that reaches each part of the building. Besides a detailed sky map, this computation will require a 3D model of the building and status of the shades. This problem involves two primary challenges that will be aimed to address: a camera that can capture a hemispherical illumination map having very high dynamic range, and an efficient computation model that can predict the illumination at every point in the building.

i) Develop a high-dynamic range fish-eye sensor that can capture environmental maps.
ii) Develop methods for easy installation and calibration of such a sensor on the roof-top.
iii) Collect illumination maps of buildings with internal illumination measurements.
iv) Simulate the illumination inside using ray tracing and verify against measured values.
v) Develop efficient methods for computing illumination with various façade conditions.
vi) Extend the method to a network of sensors on multiple buildings.

Securing Wireless Sensor Networks for Smart Building: Wireless Sensor Networks (WSNs) have various potential applications ranging from civilian to military applications including smart building. The sensors are deployed densely in a building and the sensors send important sensing information to the nearby base station(s). In order to protect the sensing information, it is needed to design and implement various security protocols, such as key management, access control, and user authentication protocols. Due to resource limitations and vulnerability to physical capture, these problems become more challenging.

i) Identify the protocols that can generate quality and critical feedback on the operation of the secure smart building with the deployment of WSNs in the building.
ii) Develop various secure protocols such as key management, access control, and user authentication in order to make WSNs secure.
iii) Theoretical and simulation analysis of the designed protocols, which involves the rigorous security analysis, overheads analysis and functionality features analysis, and simulation through NS2 and AVISPA tool.

Computing and Algorithmic Frameworks for Smart Buildings: Smart buildings require feedback to periodically update their operating conditions. This feedback has to be generated by computing on the data collected about the building in use. Depending on the nature of the computation, one requires the design, development, and implementation of efficient algorithms on a variety of computational platforms.

1. Identify computations that can generate quality and critical feedback on the operation of the smart building.
2. Develop algorithmic and computational frameworks that can improve programmability and productivity.
3. Deploy parallelism where needed to improve real-time response and feedback generation.
4. Understand the nature of the data and its impact on the algorithm design and implementation process.

Indoor Localization for Smart Building Applications: Indoor localisation provides location of people within a building. Using this information, smart buildings can provide interesting application such as personalized services and environment, augmented reality experience, and evacuation planning in case of an emergency.

1. To study and develop an indoor localisation system based on existing Wi-Fi in a building.
2. To develop an indoor localisation system based on infrastructure requirements - Bluetooth low energy beacons and modulated LEDs (visible light communication).
3. Integrate the two localisation techniques developed into potential smart building applications.
4. Conduct a workshop on indoor localisation.

Cognitive Wireless Sensor Networks (CWSN) for Smart Buildings: With the emergence of WSN and IoT devices, it is expected that a smart building will have thousands of sensors. Most of these sensors need to operate in free but crowded ISM bands with no guarantee of service and high interference. Spectrum scarcity is going to be a key bottle-neck for deploying sufficient sensors in smart buildings. A cognitive radio (CR) can solve this issue by monitoring even the licensed bands (such as TV whitespace) in addition to unlicensed bands to use the frequencies efficiently, reliably and dynamically.

1. Development and hardware prototyping of cognitive radio mote (CRM) for smart building applications. The proposed CRM will have the capability of sensing the RF environment and adapting its operation parameters including frequency, waveform, and power.
2. Development of algorithms for exploration of spectrum opportunities and its efficient utilisation using heterogeneous sensors.
3. Implementation and testing of algorithms developed on the designed CRM.

Wireless Sensor Networks (WSNs) and Internet of Things (IoT) for Smart Buildings: Wireless Sensor Network (WSN) and Internet of Things (IoT) are going to be key enablers for Smart Buildings (SBs). Buildings become smart only when they start learning from their environment for which awareness provided by these devices is crucial. Although lot of research has gone in WSN and IoT, the number of practical applications and products for SBs are very few. In this project, indicative application of WSN and IoT for SBs, such as monitoring of ambient building environment will be performed initially. Later, these applications will be extended and customized for other SB applications.

1. Developing proof of concept demos using state-of-the-art WSN and IoT technologies for the SBs to measure physical quantities such as temperature, humidity, and CO_2.
2. Developing and implementing new algorithms for WSNs for increasing reliability, robustness, coverage and battery-life of the network using multiple heterogeneous sensors.
3. Piloting, prototyping and technology transfer of IoT based SB applications.

Can a Society of Smart Buildings be Smarter than Sum-of-parts? Many tasks are inherently collaborative in nature. Popular examples include democratic elections, supply-demand economy, and emergence of intellect from network of neurons to name a few. It is therefore conceivable that a collaborative society of smart buildings can be fundamentally more useful in a variety of ways (compared to isolated smart buildings).

There are two natural ways in which the whole can be greater than sum-of-parts, namely: pervasiveness and robustness. These require sharing of private data among components, which requires secure protocols.

1. ***Pervasive Resources***: The workload, memory, input/output, and the system in general can be transparently distributed leading to an *available in abundance, anytime, anywhere*!

2. ***Robustness:*** It is possible to amortize the reliability in the components leading to solutions that are *functional despite failures* to parts of the system.
3. ***Privacy/Security:*** Can mutually distrusting components successfully collaborate? Perhaps they can. Quantum mechanics makes it simpler too.

Monitoring Liquefaction-Induced Settlement of Building for different Soil System:
Earthquake-induced soil liquefaction can cause significant damage. All the failures in engineering structures result in loss of money, and potentially lives. Therefore, predicting the risk of earthquake-induced soil liquefaction is one of the most important tasks for any geotechnical engineer when working in seismically active regions. The aim of this research is to develop and implement a new technique for monitoring liquefaction-induced settlement of building and mitigation techniques.

1. Nonlinear FE Simulation of Building Foundation Soil System.
2. Development and Calibration of Sensors for monitoring accelerations, pore water pressures and displacements.
3. Development of Centrifuge Test Set-up.
4. Development of Liquefaction Hazard Mitigation Techniques. Comparison of Numerical and Experimental Results.
5. Technology Transfer of the Developed System.
6. Monitoring of liquefaction-induced settlement of building during shaking covering operational time aspects.

VOLTTRON™ Transactional Network (Pilot, Deploy and Enhance for Indian Conditions)

Smart buildings and supporting intelligent systems need to be aware of energy cost to better optimize energy performance. In the US, a Transactional Network concept is being developed to enable operational, energy, and financial transactions between building systems and the electric power grid. This vision will allow millions of sensors, meters, smart appliances, connected equipment, loads, and distributed generation assets to seamlessly communicate and coordinate, to the benefit of both building owners and the grid. VOLTTRON™ is a DOE-BTO funded open source platform for transactive energy applications that allows sensing and control actions to take place as close to devices as possible. Designed to support modern control strategies, including use of agent-based and transaction-based controls, VOLTTRON™ enables mobile and stationary software agents to perform both information gathering, processing, and control actions.

1. Develop agents that reside either on the equipment, on local building controllers suitable for Indian conditions.
2. Piloting, prototyping and technology transfer.

RF Energy Harvesting for WSNs in Smart Buildings

Smart buildings need wireless sensors for gathering feedback to periodically optimize their operating conditions. These sensors will be deployed in large numbers and some of them may be placed in hazardous or inaccessible parts of the building. As such providing wired power to them or changing their batteries frequently may be inefficient and tedious. Energy

harvesting (EH) circuits linked to a receiving antenna captures and converts ambient RF transmissions into usable DC voltage. Use of EH is potentially attractive as replacements for primary batteries in low power wireless sensor nodes.

1. Design and development of EH circuits for WSN in SBs.
2. Research on algorithms for efficient harvesting and expending the energy in WSNs.
3. Integration of developed algorithms and hardware.
4. Proof of concept demos.

Health Monitoring of Buildings using High-Precision Building Vibration Sensors and Forecasting of Probable Damage during Earthquakes

At a time when around 56 percent of the land area in India is earthquake prone, it is imperative for India to develop strategies for risk assessment and mitigation because 82 percent of the population is living in 56 percent of the land areas. Risk assessment results in a quantitative index (it does not have any physical significance) that gives a qualitative feel of the level of severity of the problem. The actual process of risk assessment is a detailed exercise. It requires monitoring of building performance during actual earthquake events.

1. Identification of dominant building typology in urban area.
2. Installation of building vibration sensors. Acquisition of data of ambient vibration and also earthquake events.
3. Dissemination of information the moment earthquake occurs.
4. Framework for simulating damage to a building after earthquake.
5. Produce web maps for the use of government and public.

Water Conservation and Management for Smart Buildings using Wireless Sensor Network

Efficient monitoring and management of water quality with an energy efficient water distribution system is vital for any building. The proposed work will address the optimal water management by considering various sources such as rainfall, surface water, municipal water and treated water and various applications such as water-cooled chillers, evaporative cooling, drinking, washing, irrigation, and so on. The water balance has to be achieved while minimizing total cost of water, treatment, pumping and energy consumption in air conditioning. The present work will adopt wireless sensor network to monitor the water flow rate, leakages, pollution levels and the possible restorative measures at operational level.

1. To develop on-line monitoring of water distribution system using wireless sensors network to facilitate efficient management and operation.
2. To develop a novel water quality monitoring system for direct, rapid and remote, for detection of contamination in a water distribution system of a building using wireless sensor networks.
3. To develop adaptable water distribution network system to bypass faults and prevent contamination.

Data Management: Fusion, Curation, Detection Estimation, Visualisation, Storage, Mining

Smart buildings generate a large amount of data on a regular basis through a large number of sensors for various purposes. To mine potentially useful patterns from this data, various prediction and decision tasks can be modelled as classification and regression problems. It is important to design classifiers and regression algorithms that can be applied in diverse situations with minimal modification such as plug load identification.

1. Identify the core prediction and decision tasks of smart buildings that can be automated.
2. Design classification and regression algorithms to solve the identified problems.
3. Experiment with available options and standard algorithms for the tasks proposed to recommend the best methods for each task.

Fault Detection, Diagnostics and Prognostics (FDD&P)

Various physical phenomena that affect the building also produce signatures and their impact on the operation of the building. To create the right feedback that can be acted upon either manually or automatically, one can simulate various physical phenomenon which can affect the physical system and the signature of the cause and the effect. Once such a database of cause-and-effect signatures of various physical phenomena are captured, the acquired real-time data can be checked to see if any of the signatures are matching. This can lead to better fault diagnosis and fault localisation.

1. Identify critical phenomenon that can impact the energy usage of a building. To begin with we will start with the HVAC subsystem.
2. Develop a database of signatures of the critical phenomenon.
3. Provide querying framework to verify real-time signatures generated.

Early Stage Design Optimisation

Buildings interact with their immediate environment in myriad ways and can impact the lifetime or the operating conditions of a building. It is therefore important to understand the design space and perform simulations to find the optimal values for various parameters.

1. Adding drawing of complex shape building geometry feature to eDOT.
2. Develop Life Cycle Cost optimisation techniques
3. Integrate eDOT to OpenStudio to filter solution that meet energy code compliance criteria
4. System modification based on feedback from industry
5. Final testing and technology transfer to industry

An intelligent Systems for Personalized Adaptive Comfortable Environment (iSPACE)

Using technological advances, the buildings can provide personalized comforts, such as adaptive and automatic adjustment of lighting, temperature, and air movement, while at the same time reducing energy costs.

The overall objective of the project is to research & develop an intelligent Systems for Personalised Adaptive Comfortable Environment (iSPACE). Also, the research involves developing adaptive applications leveraging advancements in communication protocols between the buildings and the electric grid (VOLTTRON – Transactional Network), machine learning and data mining. Moreover, the effectiveness and the energy saving potential of iSPACE would be studied.

1. Develop and test open architecture hardware platform.
2. Develop and test smart applications.
3. Develop and test transaction-based controls.
4. Build a controls testing lab.
5. Conduct workshops.

Systems

Over the last decade, there is an explosion of data sources resulting in the generation of voluminous data of different *varieties* at varying *velocities*. These data sources could be sensors, mobiles, edge devices, vehicles, machines, satellites and humans as summarised in the table, Big Data Sources, below:

Table 18: Big Data Sources

Source	Examples
World Wide Web	HTML pages, blogs, etc.
Social Media	Face Book, Twitter, Instagram, LinkedIn, etc.
Multi Media	YouTube, Netflix, Prime Video, etc.
Enterprises	Financial transactions, e-mails, etc.
Mobile Apps	Uber, Crowd Sourcing apps, etc.
e-governance	Legal documents, weather, agriculture, etc.
Machine Logs	Data centres, network switches, etc.
Remote Sensing	Geo-spatial maps
Internet of Things	Smart healthcare, smart cities, smart transport, smart grid, smart factories, etc.
Financial & Business Analytics	Stock markets.
Block Chain	Block Chain Ledgers, Smart Contracts, etc.

The data generated from these sources may need to be processed partially at the collection end-point (edge), transferred to an intermediate node (fog node) where it is further processed and aggregation may happen. Finally, the data will be transferred to the data

centre where it will be stored using an appropriate storage structure in a centralized or distributed fashion. Overall, in order to collect, transfer, store and analyse data, the whole systems stack needs to be engineered. All the problem areas fall under the category Systems for Data Science and are expanded in sections 2.3.6.1 Edge Computing through 2.3.6.2 Data Science for Systems below. On the flip side, if we consider large and medium data centres, where huge amount of data gets generated in the logs, we can do mining and predictive analytics, for performance engineering, pro-active fault detection and fault isolation. The problems here fall under the area Data Science for Systems which we present in Section 8. The following list summarizes the rest of the proposal.

1. Edge Computing.
2. Edge-Fog-Cloud Optimisation: Data quality, data aggregation, privacy issues and so on.
3. Database Optimisations: Application-specific features that enhance the functionality and performance of the database. Time, space, streaming, batch, linked (Volume, Velocity, Variety and Veracity).
4. Distributing Computing Frameworks and their Optimisation: Optimise the frameworks dynamically on a per application basis. Examples: optimise MapReduce node placement, etc. Theoretical guarantees. Performance models. Use of ML/Control theoretic approaches.
5. Algorithms for Distributed Frameworks.
6. Data Mining and Analytical Frameworks: Data visualisation, spatial aggregation vs Temporal aggregation, extracting information from data, information models.
7. Data Science for Systems: Predictive vs Diagnostic, behavioural analytics for better SOPs, Automobility of HCI systems.

Edge Computing

Many application areas such as healthcare, automotive industry, smart warehouses, smart cities, smart buildings, smart power grid, intelligent transportation, smart agriculture include the use of edge devices for a variety of tasks. For example, in the healthcare, patch recorders and pace makers are example edge devices which continuously sense and record body vitals such as heart rate, blood pressure and ECG signals. Another example of an edge device is a smart water meter which measures the flow of water into a specific area of a distribution network or into a house hold. A strategically placed collection of smart water meters in a distribution network can help address many critical problems such as identifying the sources of pipeline leaks and any contamination, understand water usage patterns, etc.

These edge devices usually have meagre available compute, memory and storage resources. Mostly, they are powered by batteries and sometimes these are even irreplaceable due to physical constraints, like a pace maker implanted inside a human body. The network connectivity could be intermittent due to lack of signals. Along with these constraints, certain applications demand privacy and require real-time action or response. Hence, transmitting data to a fog node or cloud is not possible. Hence, computation needs to be

done and actionable information has to be inferred at the edge node itself. This leads to the following problem areas falling under the broad umbrella area AI/ML at the Edge.

1. **Hardware Architectures**: Novel processor architectures and special purpose neural accelerators (ASICs) for expediting inference using the deployed deep neural networks (DNNs) at the edge.
2. **AI/ML Algorithms on Accelerators:** Deployment strategies for fast and low power inference algorithms on devices such as low power CPUs, GPUs and FPGAs.
3. **Rethinking AI/ML Algorithms**: The state-of-the-art DNNs require enormous amount of computations per inference. DNNs which can give accurate results while performing substantially less computations at a very low precision are required for all edge applications.

There are many high impact edge devices from different application areas which require end-to-end system design. All this fall under the following problem area.

End-to-End Design of Novel Edge Devices Smart water meters, plant health monitors, devices to measure nutritional content in animal fodder, milk testing kits, biomechanical load sensors, etc.

Edge, Fog and the Cloud

While some applications computing on the edge, there are many applications where it is profitable to use an intermediary fog node which has more resources than an edge node. For example, in a car there could be tens of Electronic Control Units (ECUs), which sense many parameters. Each of them may not have enough information to infer and take a suitable action. At the same time, it does not make sense for each of the ECUs to stream data to a cloud where all the data streams are analysed in an integrated fashion. This requires lot of unnecessary data transmission as most of the time there may not be actionable information. Even if actionable information is there, the latency will be too high and hence may lead to disastrous consequences. However, we can envision a fog node on the car which collects data from all the ECUs, processes locally, takes necessary action and transmits relevant, aggregated data to the cloud. On the flip side, it is possible for a centralized cloud to control all the edge nodes through the fog nodes in a hierarchical fashion. For example, secure over-the-air software updates are possible through the fog nodes. The following are some major problem areas we envisage in this space.

1. **AI/ML on the Fog Nodes** Fog nodes are more powerful than the edge devices. Hence it is possible to handle multiple data streams and use more complex machine learning algorithms for better inference. Hence the algorithmic optimisation points for fog nodes are different when compared to edge nodes.
2. **Edge to Fog Networks** The edge nodes can communicate with a subset of their peers and some fog nodes. There are many communication protocols possible at the physical layer and they differ from one another in terms of range, power and latency/bandwidth. The choice at the physical layer have implication on the network, transport and application layers of the protocol stack. Different network topologies

such as star, multi-hop networks are possible and the associated routing protocols also vary. The data transmission and aggregation strategies depend on the underlying network protocol stack.

3. **Internet of Vehicles** The fog nodes from different vehicles can communicate with one another and with road side units (RSUs). The RSUs can communicate with one another and with a central cloud. This opens the scope for a wide array of interesting problems like trac estimation, over-the-air software updates, and so on.
4. **Service Oriented Architectures** The applications which span across edge, fog and cloud are intrinsically distributed. Hence, they are amenable to be designed and developed using modern service oriented architectures. Different micro-functionalities within an application can be encapsulated cleanly as micro-services in containers with appropriate RESTful communication mechanism among them. These large collections of containers can be managed using container managers.

Storage Structures

Over the two to three decades, due to the expansion of data sources from business processes to world wide web, social media, multi-media and the like, data has become more unstructured and hard-to-store in traditional relational databases. Also, due to the volume of the data, the underlying storage mechanisms have become inherently distributed. The accesses to these storage structures are highly concurrent as most of front-end applications using them could e-commerce kind of applications. This leads to consistency and availability issues. In this context, many distributed NoSQL databases and Key-Value stores have been proposed with different consistency, availability and partition tolerance properties. Alongside, novel graph databases such as Neo4j and InfiniteGraph have come up. There are many research problems in this space related benchmarking and performance evaluation, query languages and optimisation, database security and privacy, hardware accelerators for databases, graph data management, information retrieval and text mining, scientific databases, semi-structured data, spatio-temporal data, Storage and indexing structures, streaming data and complex event processing, transaction processing, data base optimisation using machine learning, optimizing machine learning algorithms using databases and the like.

Distributed Computing Frameworks and their Optimisation

Systems for data science are ubiquitous with examples ranging from domains such as smart cities, energy efficient buildings, social network analytics, and the like. These systems collate, refine, filter, fuse, analyse, and present data and operate on a multitude of devices with wide form factors ranging from handhelds to the clouds. These systems are naturally distributed in nature and hence face all the challenges of distributed systems such as the data partitioning, synchronisation, mutual exclusion, failures, checkpointing, and the like. Nevertheless, these systems are expected to work seamlessly under a variety of operating environments while still making optimal and efficient use of resources such as storage, compute, power, and the like.

In order to achieve this, systems for data science have to navigate a complex maze of a diverse set of architectures, theoretical guarantees, a set of generic frameworks for enabling distributed programming, and a potentially huge set of optimisations that are driven by application characteristics. The present architectural space is populated with a variety of devices ranging from multicore CPUs, many core accelerators such as GPUs, FPGAs, and other hybrid models. Distributed programming frameworks such as Map-Reduce, Pregel, Apache Spark, and the like provide a useful abstraction that eases the task of the programmer in building distributed systems. Theoretical guarantees are fundamental to ensure the algorithms used in building these systems are optimal in terms of their resource usage.

Despite the availability of feature-rich and programmer-friendly frameworks for distributed computing, application specific concerns still have a huge impact on the performance of systems built using these frameworks. For example, when one uses the Map-Reduce framework, specific to an application, there is usually ample scope for optimisation of the number and location of the Mappers and Reducers to use.

In addition, several researchers have developed other less feature-rich but more application and architecture specific frameworks for distributed programming. These operate at various scales such as those that accommodate applications across CPU-GPU clusters. Such localized frameworks allow for a more fine-grained optimisation across a set of applications.

Distributed systems for data science often orchestrated several distributed algorithms to achieve the given computation. For concerns of efficiency, it is pertinent that the individual algorithms optimize the number of communication rounds needed, data placement and partitioning is fine-tuned to the needs of the application.

In the light of the above observations, the following high-level questions emerge in the context of systems for data science.

- Optimise Existing Frameworks: How can applications aim to optimize the behaviour of existing framework to build systems for data science? What are some potential approaches do this optimisation? How can one incorporate learning-based techniques to explore the space of optimisation in an effective manner?
- Prepare application-specific and architecture-specific frameworks for various systems for data science.
- Analyse algorithms used in systems for data science with respect to parameters such as round complexity, data partitioning and its impact, appropriateness to the underlying architecture, and the like.
- Identify appropriate performance models that capture the characteristics of the application and the system and compare the measured performance with an ideal performance.
- What are some data storage mechanisms for handling the volume and variety of data that is produced by applications? In particular, how do the existing SQL, and NoSQL based data storage mechanisms such as BigTable, Neo4j, and the like cater to

the richness of applications from domains such as the sciences, geospatial informatics, and energy-efficient buildings?

Many of the questions listed above can form the basis for several projects in this area. One can think of suitably elaborating on these questions based on specific application verticals.

Algorithms for Distributed Frameworks

Different distributed computing frameworks such as map-reduce, Giraph (vertex-centric computing model), Akka (asynchronous distributed computing), etc. require different algorithmic approaches. There is a scope for investigating appropriate algorithms for a wide array of fundamental computing problem from different application domains.

Data Mining and Analytical Frameworks

From traditional systems of data collection like health records, financial transactions, satellite imagery and others with well-defined schema, structured data, to the fast emerging world of next-generation devices like IoT, wearables, etc. to crowd-sourcing models, unstructured data, the data scene is exploding not just in terms of volume and rate of data collection, velocity, but also in terms of variables. The scales of deployment of these and the range of data acquisition, throw up a range of challenges in terms of visual and analytical frameworks needed to processes, filter, visualise, extract the relevant or significant parameters and support decision making systems. This leads to a need to develop appropriate systems for data science, as listed here, that address some of these key challenges:

1. **AI/ML based Data Visualisation Frameworks for Large Data Sets** for intelligent models for extracting domain relevant features and dimensions; visualisation and data aggregation systems for a range of data assimilation approaches ranging from multi-source data sets to problem-dependent scaling of the multi-dimensional data sets.
2. **Data Mining Techniques or Models** for checking data veracity, handling the outliers; identifying and extracting the relevant patterns at a generic level or domain specific level. One such example could be models that help detect device deterioration or sensor failure detection (by combining domain-specific knowledge with the sensor data behaviour).
3. **Data Analysis and Information Models** that can help transform the raw data into sensible or actionable information leading to improved business intelligence and its presentation.
4. **Spatio-Temporal Data Analysis** that takes into account the challenges that spatial and temporal scales pose while handling multi-dimensional data with such features. One example can be developing models that move from a user controlled/ interactive analytical approaches based on hypothesis to a more data-driven and context relevant analysis of the data.
5. **Privacy Preserving Models** for data aggregation and presentation.

Data Science for Systems

Over the last few decades, there is a rise of large complex systems that bring together various engineering and/or technical systems including information systems, cyber-physical systems, etc., with socioeconomic systems leading to an integration of a range of functions.

Many of these systems are designed and organised as either workflow-based systems or complex, interconnected, multi-level systems. An example of a workflow-based system can be the medium and large scale private, public and hybrid clouds that host thousands of machines connected by high speed network switches and cables. Tens of thousands of jobs, some periodic and others aperiodic, run on these clouds on a daily basis, mostly organised as workflows. On the other hand, while design and analysis of cyber-physical systems like urban mobility, smart grids, smart agriculture and others require specialised approaches, their performance monitoring and feedback to improve the overall system behaviour, and its impact on safety and well-being of the public and environment necessitates the need for application of various analytical, mathematical and scientific principles and evaluate them. Here we envisage that the Data Science for Systems will provide some of these key insights to develop models for such large complex engineering systems to business processes evaluation and understanding to performance models. This may be over a mix of engineering and other disciplines and across a system, product or service.

Some of the key areas that Data Science for Systems may address are:

1. **Resource optimisation**. Many of these systems consume a variety of resources and it is necessary to continuously evaluate them for improved performance, develop appropriate metrics, and develop models for incorporating them. For example, in large data centres, there is a need to optimize on power by efficient mapping of jobs to compute resources through analytics.
2. **Data-driven system diagnosis**. Increased use of sensors/IOTs, software systems and other monitoring mechanisms built-in provide a wealth of data at frequent intervals and/or at critical phases or events both in real-time and as log files. Diagnostic models based on a multitude of these parameters can help build early warning systems, improve alerts, and also help predict failures, if any, in mission critical operations.
3. **Predictive analysis of System behaviour** based on its performance analysis across a range of scenarios or operational phases. This might also help detect and predict violations of the Service Level Agreement, degeneration of system, and such other deviations.
4. **Behavioural analytics of SOPs**. For effective functioning of many large complex systems and workflow systems where interactions between/within systems and humans and systems is necessary, standard operating procedures (SOPs) are developed to help assess and predict the behaviour and outcome at each level of such interactions. Post-facto data analysis of such procedures with the information collected can help assess its effectiveness and provide inputs to improve these SOPs.

5. **Automating HCI systems**. As computer systems become more intelligent including AI based systems, many systems are transiting from a manual operational mode to a semi-automated or fully automated mode. This transition can be assisted and aided by understanding the contributions of various sub-systems including humans, products and devices in such systems. Hence, there is a need to develop appropriate models for Data analysis of the various HCI systems and its components along with simulation and modelling of these.

India Specific Research Initiatives

In addition to the four domains given above, namely, Mobility, Healthcare, Buildings and Systems, the TIH-Data will also identify important problems that are crucial from the country needs' point of view. Below are some of the indicative problems that the hub may work take it up from the 2nd to 3rd year onwards. Six months before the Hub is ready to take up problems in this area, a brainstorming session involving experts from Academia, Industry, Government and Social Organizations may be arranged to identify specific problem areas with the aim to develop technology solutions that will make a large impact at the grassroots level.

Smart Agriculture

Agricultural field monitoring and reporting systems: Technological advances from Remote Sensing to Drones to IoT devices throw up many opportunities to collect, collate and report on the crop status at different stages and at multiple scales. While a farmer based system of crop monitoring is for crop diagnosis and crop management, data aggregation and reporting at administrative levels help fine tune policy and enable appropriate interventions in both normal course of agricultural development to disaster mitigation and response. Can Data-driven approaches, across spatial and temporal scales, help assess the Crop suitability? Can patterns or pathways of agricultural development (or decay or stagnation) be traced based on the input and agricultural management practice data? How can AI/ML based monitoring systems help integrate data across scales for improved Crop performance understanding in a changing environment? Can data-driven models help in early-warning and predicting crop life-cycle?

Resource Assessment Mapping: Agricultural fields, in terms of land holdings to crops grown and resources used vary vastly across the country. The statistical data, currently available at an aggregated level of district or state do not provide the right insights for supporting, enhancing, training and evaluating the access to resources and its performance. Also, major decisions done at a macro-level, like all across the state or agro-ecological zones, do not lead to the desired results in many areas. This means the farmer depends heavily on own experience and his social circle to get the inputs. There is a need to develop mechanisms to

effectively collect and collate such information, and also keep it as current as possible. And develop dissemination models based on visual and non-visual modes, to reach out to all the players in the agricultural domain.

Crop Performance Analysis: Many factors drive the crop performance across its life cycle, from providing the right inputs at various stages of the crop to identifying/diagnosing diseases/pests to actual crop yields. While a range of approaches do exist from crop-cutting experiments to surveying, etc., they are too local and the insights they provide are limited. Can new models of performance based on large scale data analysis help provide some key actionable insights? Can it help re-assess farmer incomes based on the cropping patterns across seasons and regions? How does crop suitability fare against crop economics?

Agro Supply Chain: There are multiple ways for the farm produce to reach the consumers table, ranging from raw produce like vegetables and fruits to processed products. There is a need to quantify, identify and provide an understanding of this pipeline, which has both the organised and unorganised sector as players. Can a data rich environment be developed that addresses not just crop price for the farmer, but provide innovative ways of information collection across or over parts of this supply chain? Can such understanding help reduce farm produce wastage?

Garbage Management using Data Analytics

Generation of baseline spatio-temporal distributions of garbage: Can a technology driven (use of satellite images, drone imagery, IoTs, Smartphones) approach be taken for improving the data collection mechanisms and location specific garbage dumps. Can data-driven models be developed for improved behavioural response and better compliance to garbage disposal norms?

Design Optimization models for specific needs: Apart from biodegradable and non-biodegradable wastes, many of the materials in the garbage can be recycled, recovered or reused in whole or parts. Can Life cycle assessments of garbage generated in the country help improve parts of this chain, say in terms of waste reduction? Can such data-driven models provide for improved resource mobilizations and logistical handling?

Scale dependent integrated Spatio-temporal Systems model: Many localities and industries may have the resources or capabilities to either reduce the waste generation or provide mechanisms or linkages to effectively reuse/recover/recycle some of the wastes thus reducing the load on the city-wide systems. How can this be evaluated and solutions developed? Can these models also integrate specific environmental or occupational health hazards? How to use a data-rich environment to continuously learn and improvise the outcomes?

User-specific communication: Once the garbage is disposed off by a locality, they are completely unaware of what happens to it and how it might come back to affect their neighbours or even themselves. Can a data-driven approach provide feedback to the

people? Can better communication models engage the community and encourage them to adopt Good practices such as garbage segregation, reduce garbage?

Health Issues/Public Health

Prediction of Disease Outbreaks: Timely prediction of future outbreaks of diseases such as SARS, dengue, chikungunya, etc. is expected to prepare the country and the world to manage other dependent systems effectively. Currently, the effect of the outbreak of covid19 pandemic on human lives, economy/jobs, healthcare system is being witnessed clearly. We also are witnessing the ability of data-driven technologies to bring solutions to a number of problems from simple social distancing, contract tracing to discovery of drugs/vaccine.

India Specific Diseases: Every year around 2.5 million people succumb due to TB and malaria mostly from developing/resource-constrained countries. For business reasons, not much effort go into finding new methods for diagnostics and treatment of these diseases. One of the areas that the Hub will be working on will be in collecting data and develop technologies for these purposes.

Disaster Management

The number of disasters that the world is witnessing is on the rise due to several factors with climate change being the primary one. A large population is constantly threatened by disasters such as floods, earthquakes, cyclones, tsunami, etc. Data-driven technologies can provide solutions to certain problems associated with these disasters. Additionally, the key is to develop solutions that are proactive and not just reactive. Drones and robots today can survey affected areas that are too dangerous for access by humans. Sensors on these autonomous devices can detect life forms in debris where human access is not possible, for example, after an earthquake. Such technology will also minimize the rescue time and allows the rescue workers to maximize their efforts in rescue and minimize their time in searching. Can we develop technology that can assess vulnerability levels of population toward each type of disaster based on multiple factors? For example, if data-driven technologies can predict the most vulnerable area accurately with respect to the expected point hit by a cyclone, it will greatly help in the process of evacuation. In general, data-driven technologies can play a great role during the toughest times of natural disaster, and the Hub may choose problems related to this.

Public Safety

Diverse aspects fall into the category of public safety. Some of them include traffic control, policing, criminal investigations, combating terrorism, etc. Data-driven technologies can make a number of processes within these areas very efficient and hence improve public safety in general. Can we develop technologies that can use a large amount of data from CCTV videos and offer assistance in a criminal investigation? Can technology assist in traffic control and in large gatherings in people? For example, in crowds of tens of thousands of

people, can data-driven technologies identify those with odd/tense behaviour? Can technology help in preventive policing? Can we develop technology that can help in controlling movement of crowds in religious/political gatherings, sporting events, etc. so that fatal stampedes can be avoided?

Other areas may include food security, railways, Indian culture, Strategic sector, etc.

Annexure B: List of Acronyms

Abbreviation	What does it stand for?
AI/ML	Artificial Intelligence/Machine Learning.
CAV	Connected Autonomous Vehicles.
CFL	Cluster Faculty Lead.
CL	Cluster Lead.
CPU	Central Processing Unit.
CT	Computed Tomography.
DST	Department of Science and Technology, Government of India.
EC	Executive Committee.
EH	Energy Harvesting.
EIR	Entrepreneur-in-Residence.
GPU	Graphics Processing Unit.
HCI	Human Computer Interaction.
HGB	Hub Governing Body.
IIIT-H	IIIT Hyderabad.
IoT	Internet of Things.
iSPACE	Intelligent Systems for Personalised Adaptive Comfortable Environment.
MEITY	Ministry of Electronics, Information Technology, Government of India.
MGB	Mission Governing Body.
MoA	Memorandum of Association.
NASSCOM	National Association of Software and Services Companies.
NGS Techniques	Next Generation Sequencing Techniques.
NM-ICPS	National Mission on Interdisciplinary Cyber Physical Systems (NM-ICPS).
PHFI	Public Health Foundation of India.
PoC	Proof of Concept.
RF	Radio Frequency.
SOP	Standard Operating Procedures.
TAC	Technical Advisory Council
TIH-Data	Technology Innovation Hub, IIIT-H. Also known as the Hub and TIH for Data.
TPM	Technology Project Manager.
UGV	Unmanned Ground Vehicle.
V2C Technologies	Vehicle-to-Cloud technologies.
V2P Technologies	Vehicle-to-Pedestrian technologies.
V2V Technologies	Vehicle-to-Vehicle technologies.
WSN	Wireless Sensor Networks.