Laboratory for Interdisciplinary Breakthroughs, TIFR-ICTS, May 2022

Engineering for Research (e4r™)

A ThoughtWorks Initiative To Accelerate Scientific Discovery

ThoughtWorks®

Harshal G. Hayatnagarkar
(Head Scientist, Engineering for Research)

We collaborate with research organizations on challenging computational problems such as -

Thirty Meter Telescope, Hawaii, USA

It would be the **world's largest optical telescope,** built by Consortium of partnering countries **Canada, China, India, Japan, and USA**.

In collaboration with **Indian Institute of Astrophysics, Bengaluru**, ThoughtWorks is working on the open source software to control telescope, including its user-interfaces and data management. Architected on the principles of reactive systems. Source Code on Github; Talks on YouTube.com, SlideShare.net

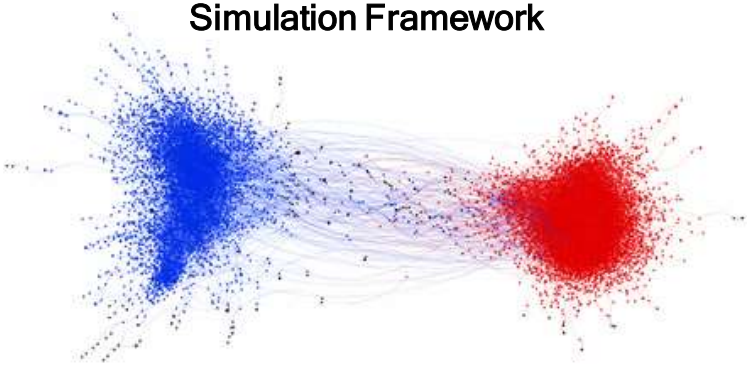
SKA Radio Telescope, Australia and South Africa

Square Kilometre Array (SKA) would be the **world's largest radio telescope** built by consortium of countries such as **Australia**, **China**, **India**, **and South Africa**.

In collaboration with the **Inter University Centre for Astronomy and Astrophysics, Pune** ThoughtWorks is building an image generation pipeline from the radio spectrum data. Prototype version 1.0 has already made a discovery.

In collaboration with **TIFR-National Centre for Radio Astrophysics, Pune** ThoughtWorks is exploring ultra-large-scale data storage, machine learning, and use of accelerators such as GPUs and FPGAs for data processing.

BharatSim: A Large-scale Epidemic Simulation Framework

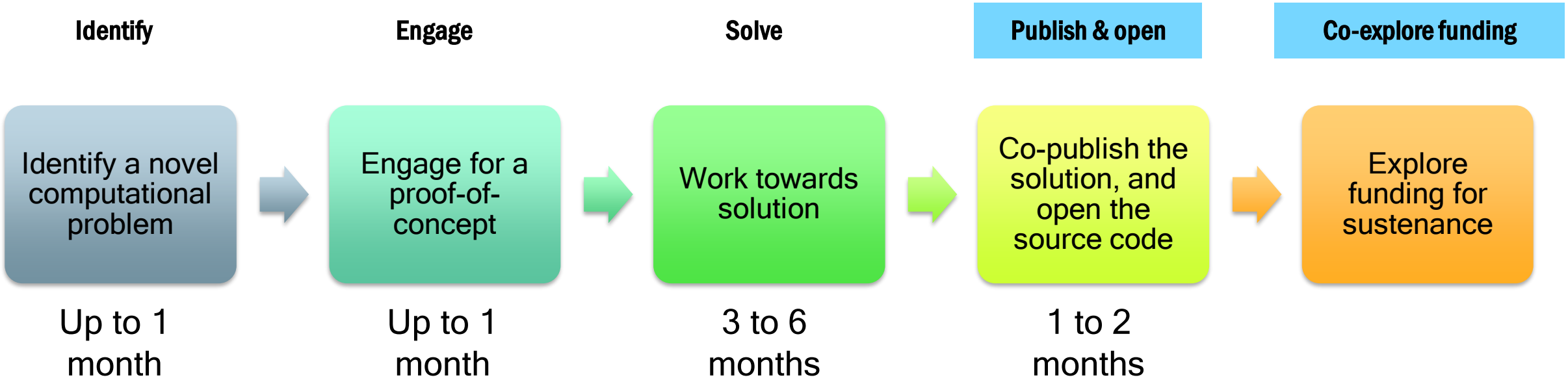


It would be the **India's first ultra-large-scale open-source agent-based epidemic simulation framework.**

A collaboration between **Ashoka University and ThoughtWorks** is co-developing this framework to help researchers simulate COVID19 spread and related policy making in the short term. In the long term, the BharatSim would aid the researchers and the policy makers in **epidemic**, **economic** and **climate change research and policy making**.

This collaboration was **funded by the Gates Foundation.**

How do we collaborate?



How did e4r™ begin?

Have the humans taken the progress for granted?

Laws of Nature

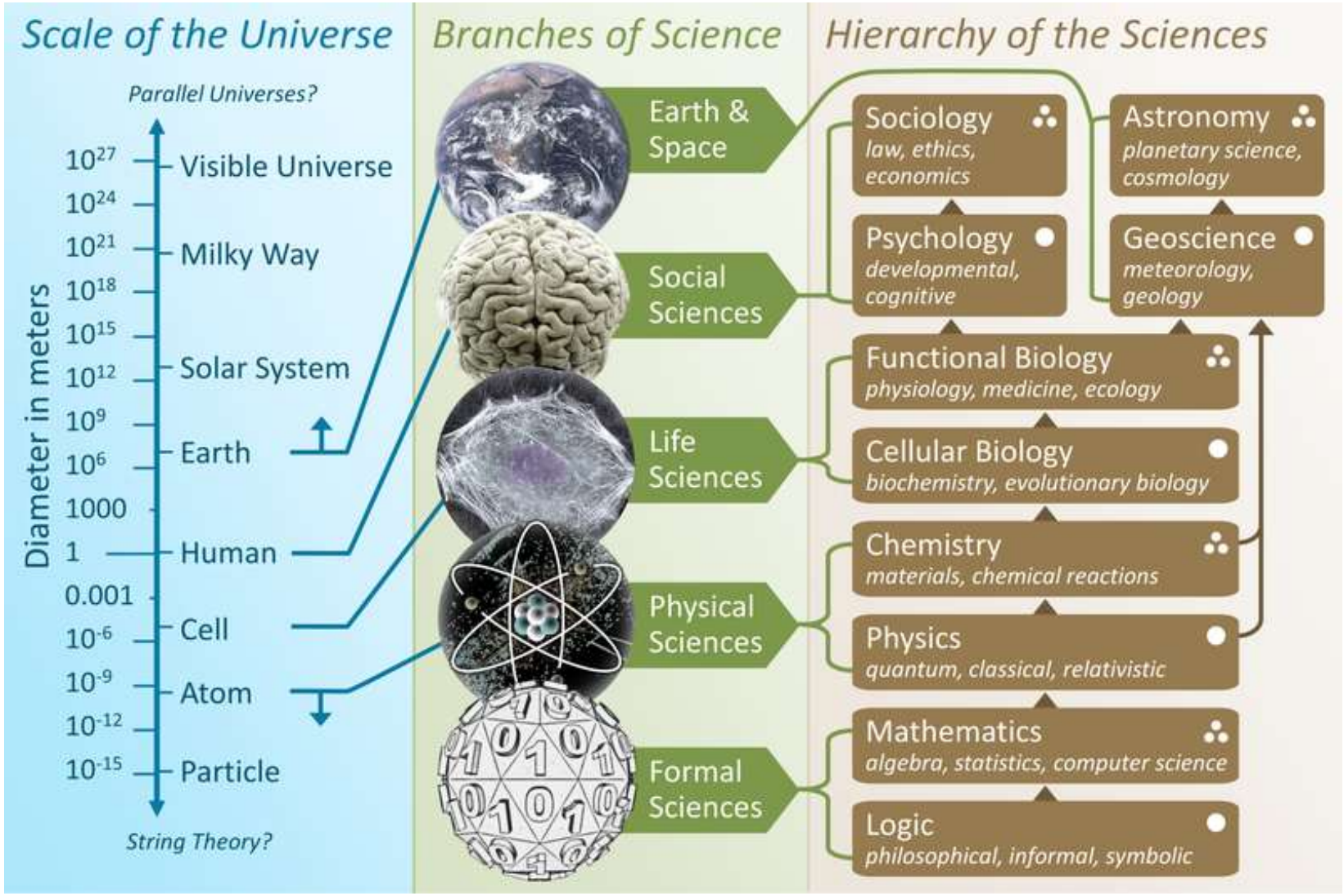
Seeds
Materials
Medicines
Tele-communication
Computing hardware and software

Health policies
Economic policies
Ecological policies
Educational policies

Scientists and policy makers today face enormous **computational challenges**.

The Nature is Complex[1] !

Multiscale emergence, from micro-to-macro ad infinitum



Drivers of complexity[2]

- Scale
- Diversity
- Network
- Dynamics

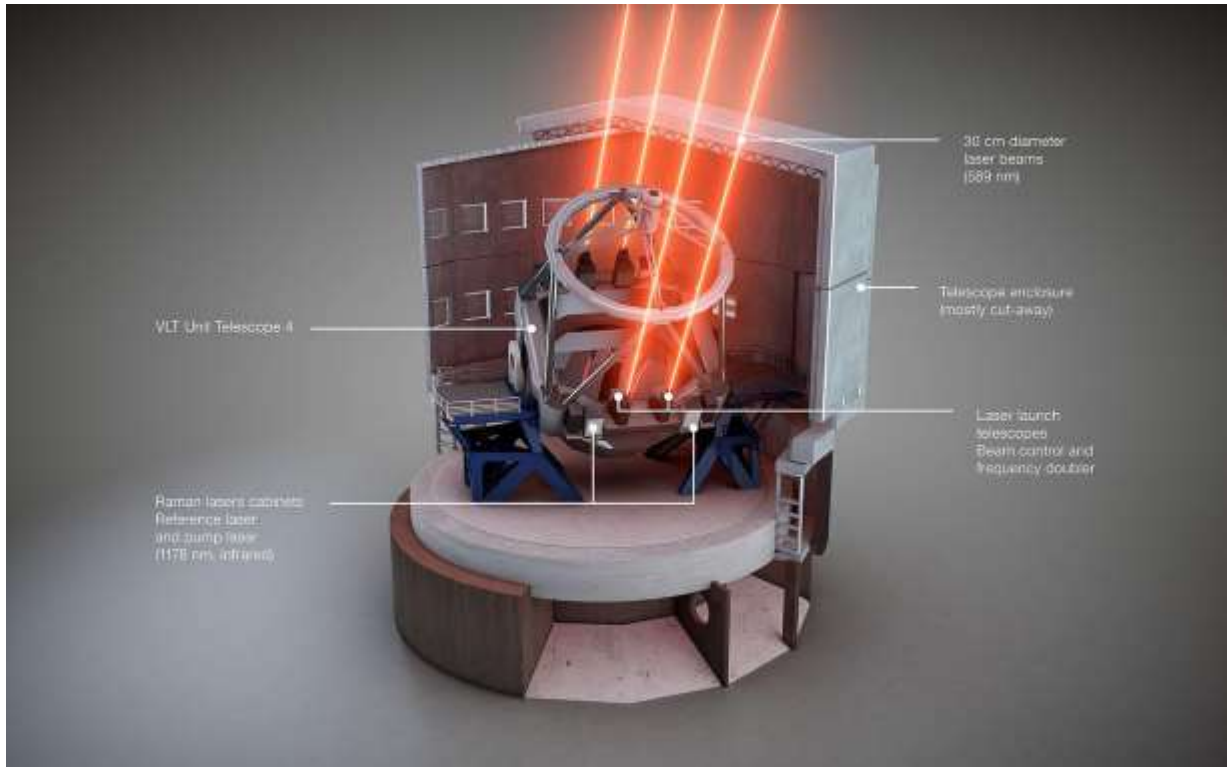
Sources:

1. Image crecit: https://en.wikipedia.org/wiki/File:The_Scientific_Universe.png
2. Hayatnagarkar, H.G., 2018, July. A compositional lens on the drivers of complexity. In IXth International Conference on Complex Systems (p. 99).

Evolution of Scientific apparatus

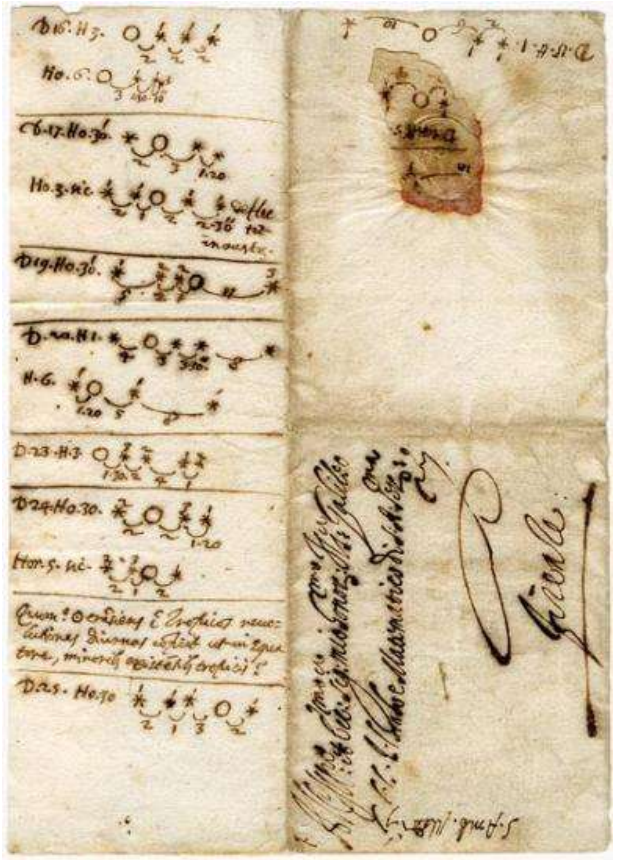
Galileo's telescope[1]
(circa 1610)

ESO VLT, Chile[2]
(Largest optical telescope, first light 1998)



1. https://astronomynow.com/2016/04/27/quadruple-laser-system-heralds-sharper-images-for-esos-very-large-telescope/
2. https://voices.nationalgeographic.org/2011/05/25/brief-history-of-the-astronomical-telescope-i-galileo-galilei/

Documenting observations



Data grid at CERN

Role of computational science

Nature-data
(empirical and hypothesized)

Nature-models
(empirical and hypothesized)

Courtesy:

- https://en.wikipedia.org/wiki/File:The_Scientific_Universe.png
- https://en.wikipedia.org/wiki/File:Computer_monitor.jpg

Evolving scope of computational sciences

Example: Great Irish Famine 1845[1]

1. Emergence of an infection
2. Infecting potato crops
3. Causing famine and deaths
4. Pushing mass migration to USA

--- Future ? ---

Transdisciplinary science

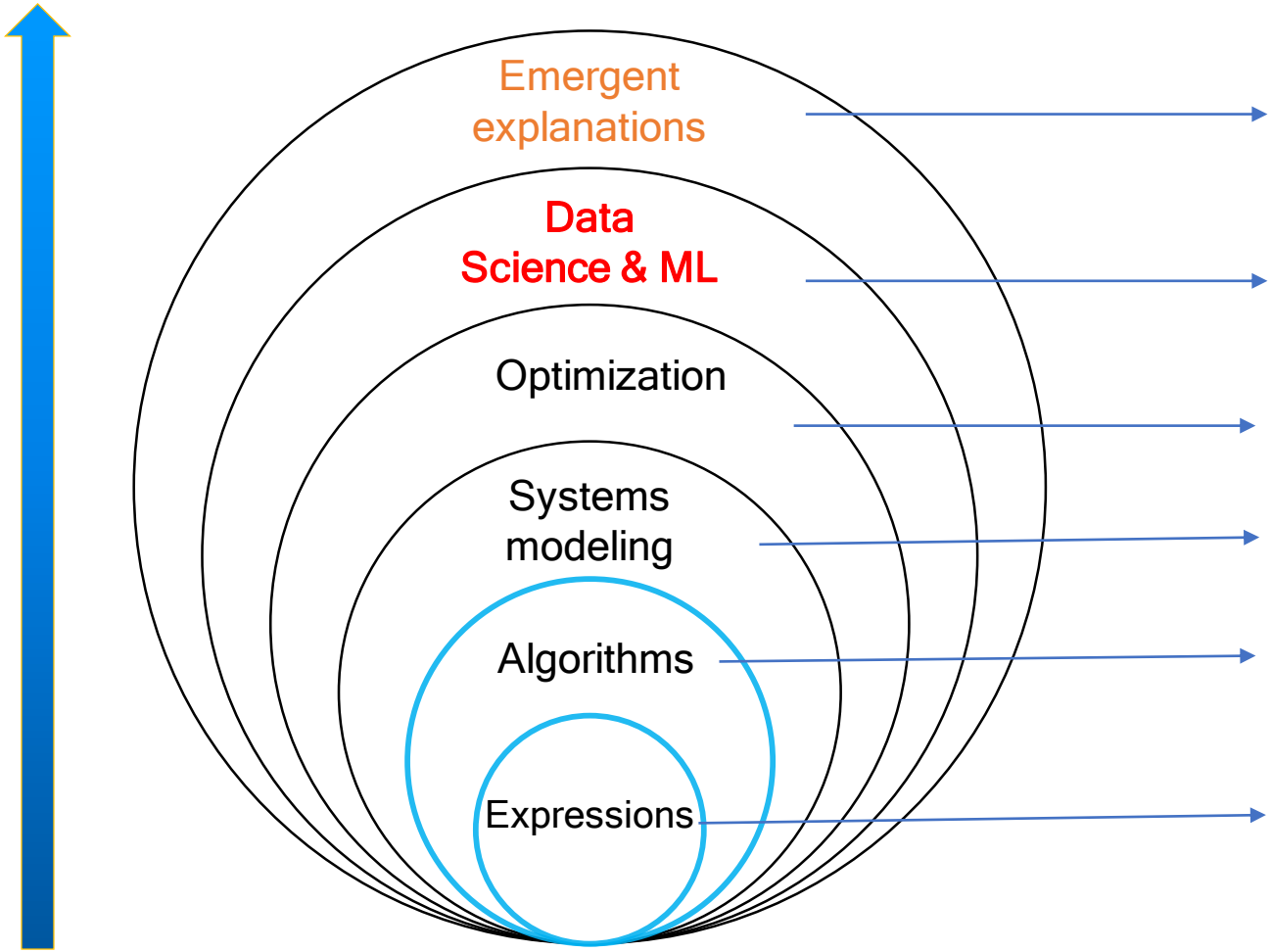
Patterns and anomalies detection

Model selection

Network dynamics

Dynamical systems

Fundamental laws, and building blocks



Multi-scale, multi-path, trans-disciplinary, and network models, and collaboration
(Explain phenomena in a bottom-up fashion)

Data + metadata management, lineage, data mining, machine learning
(Discover and manage patterns and insights)

Constrained optimization, evolutionary computation, heuristics
(Optimize for certain constraints)

Systems engineering, graph theory
(Integrate dynamics of multiple behavioral models)

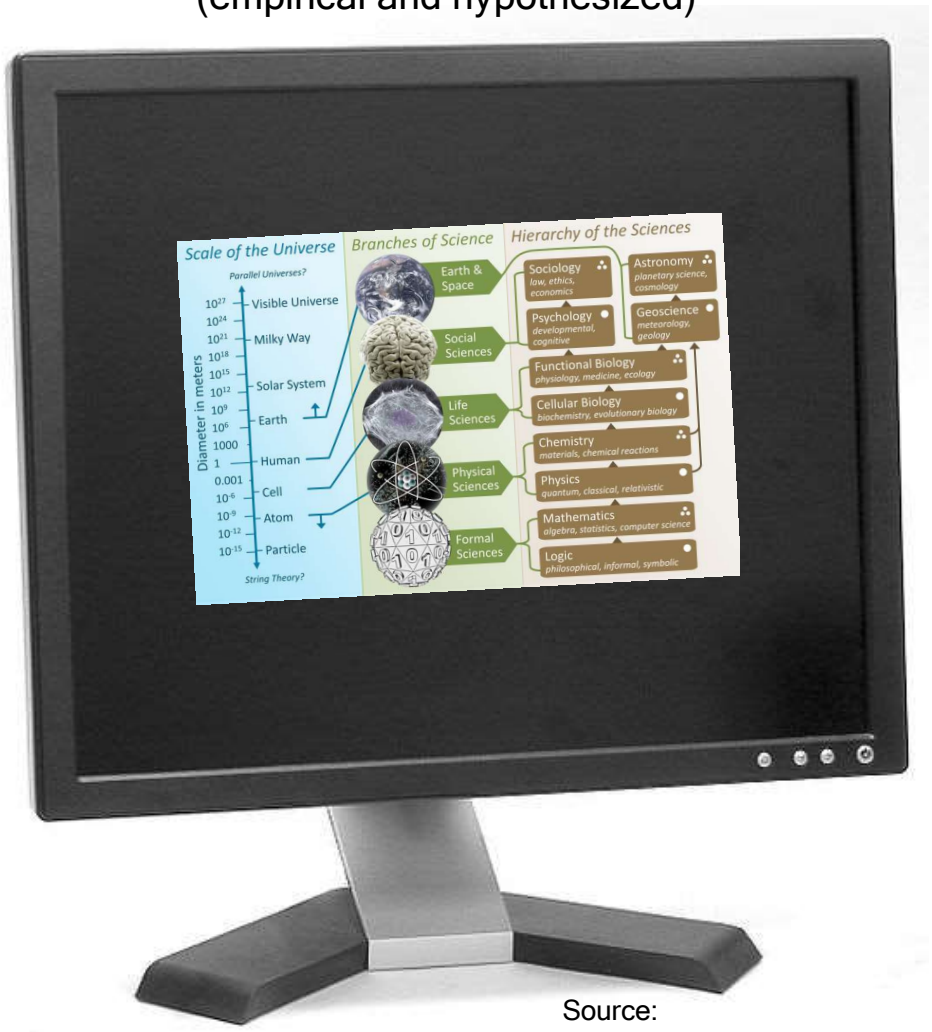
Serial, parallel, distributed
(Dynamics of single behavioral model, data analysis)

Statistics, calculus, linear algebra, system of equations, signal processing
(Behavioral model)

1. https://en.wikipedia.org/wiki/Great_Famine_(Ireland)

Complexity of computational science

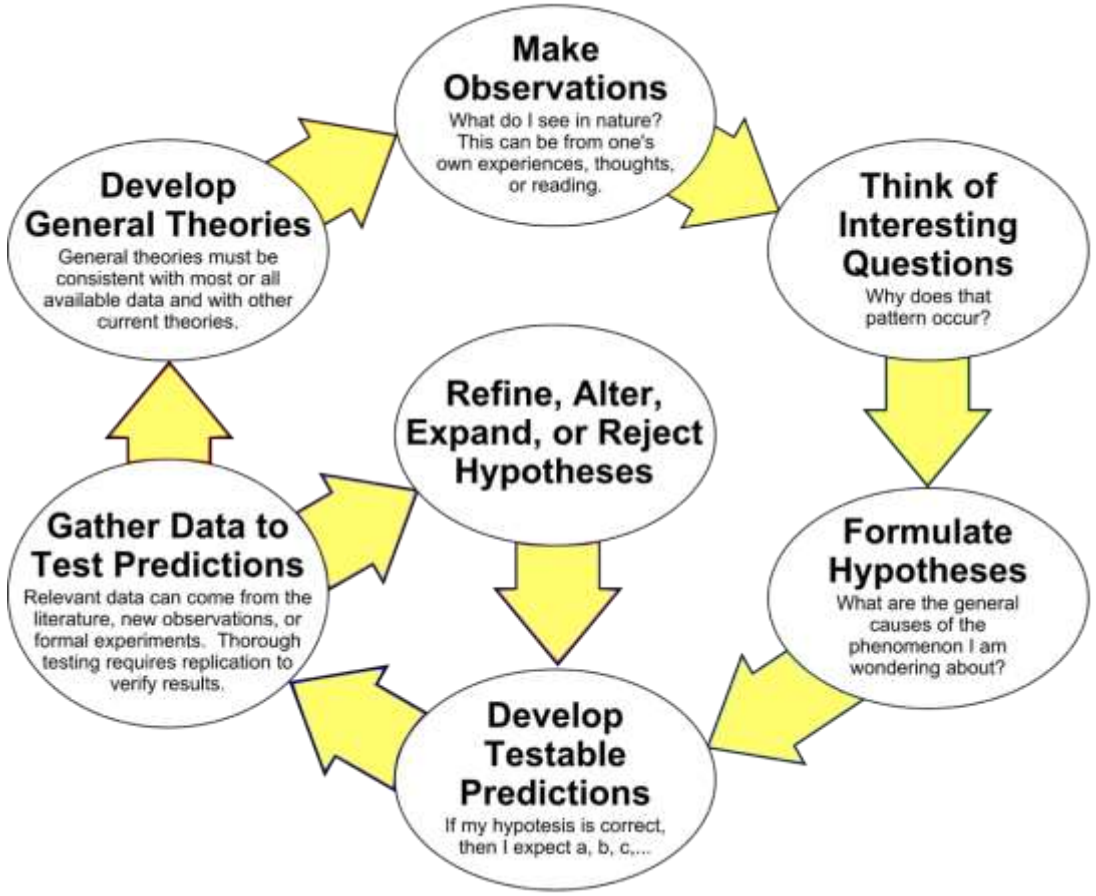
Complexity of nature-models
(empirical and hypothesized)



Complexity of modeling

Source:
- https://en.wikipedia.org/wiki/File:The_Scientific_Universe.png
- https://en.wikipedia.org/wiki/File:Computer_monitor.jpg

The scientific method



- Systematic, yet exploratory
- The ways/paradigms of doing science
  1. Observations and experiments
  2. Theory building
  3. Computer simulations
  4. **Data-intensive (unified)[2]**

Critical **to adapt computational tools**,
from control systems
to data processing
to collaboration
to methods

1. Wikipedia: Scientific Method
2. Book: The Fourth Paradigm of Science - Data-driven Scientific Discovery

Scientists ...

- Domain expert
- Mathematician
- Author
- Speaker
- Academician

Credit: Simeon Jacobson @ Unsplash

- Salesman
- Project manager
- Accountant
- **Programmer**
- **Data scientist**
- **Systems designer**

ThoughtWorks
intends a symbiotic partnership via
Engineering for Research (E4R)

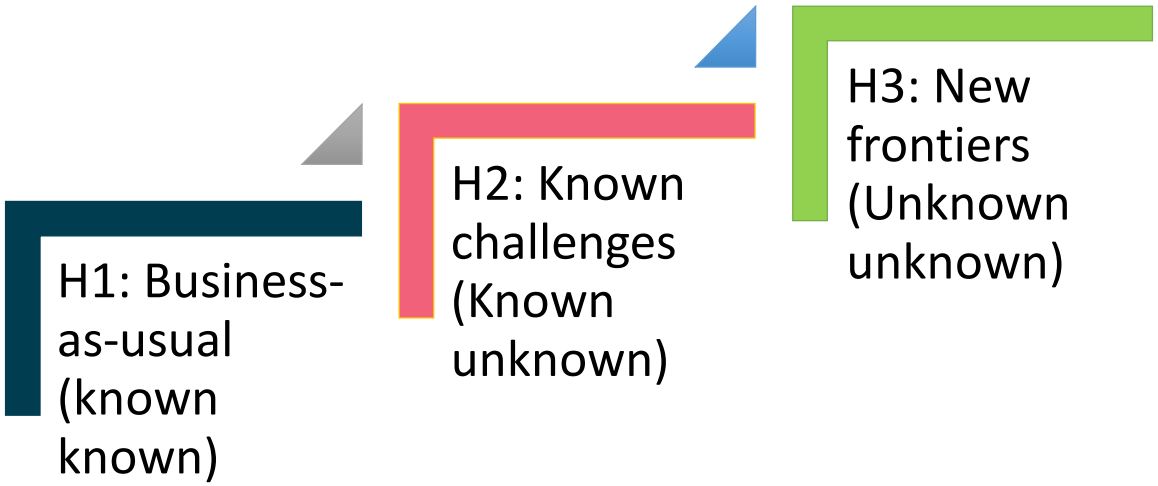
Observations

1. Ambitious big science projects are facing computational challenges at par or greater than that of faced by the Internet giants.

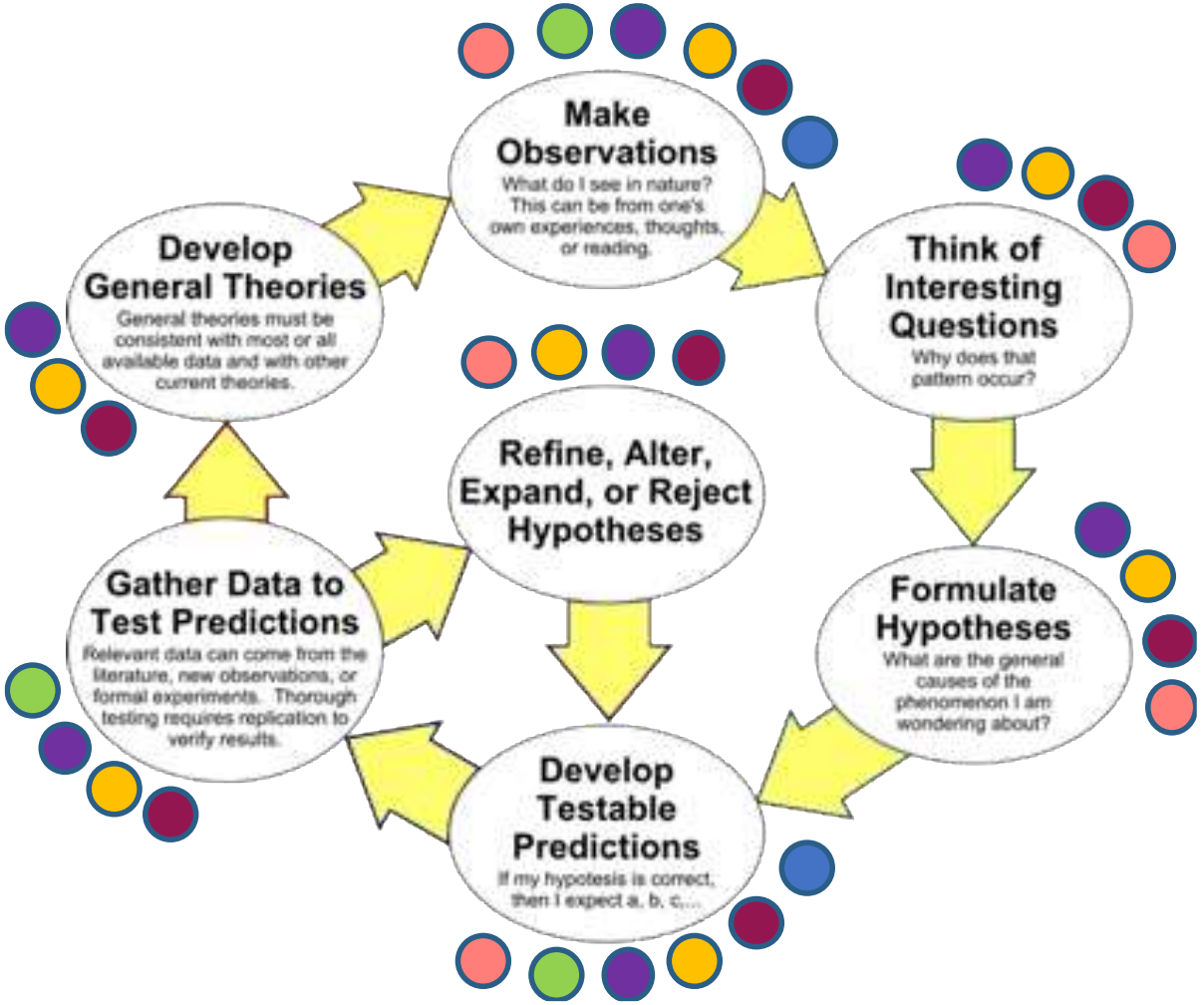
2. Advances in computer science & software engineering can significantly contribute to the progress of science.

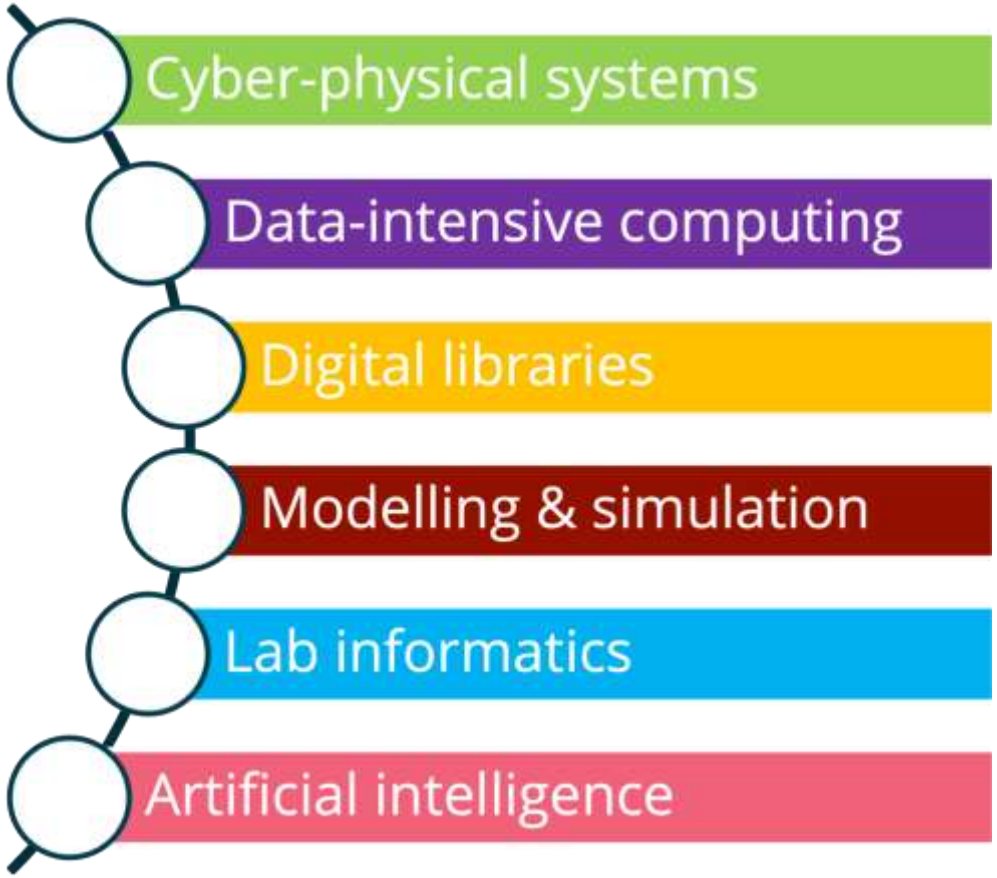
Engineering for Research (e4r)

***Our Vision** is to build a community, working exclusively with research organizations for building tools for scientific exploration, that will enable us to discover computer-science of the third horizon.*

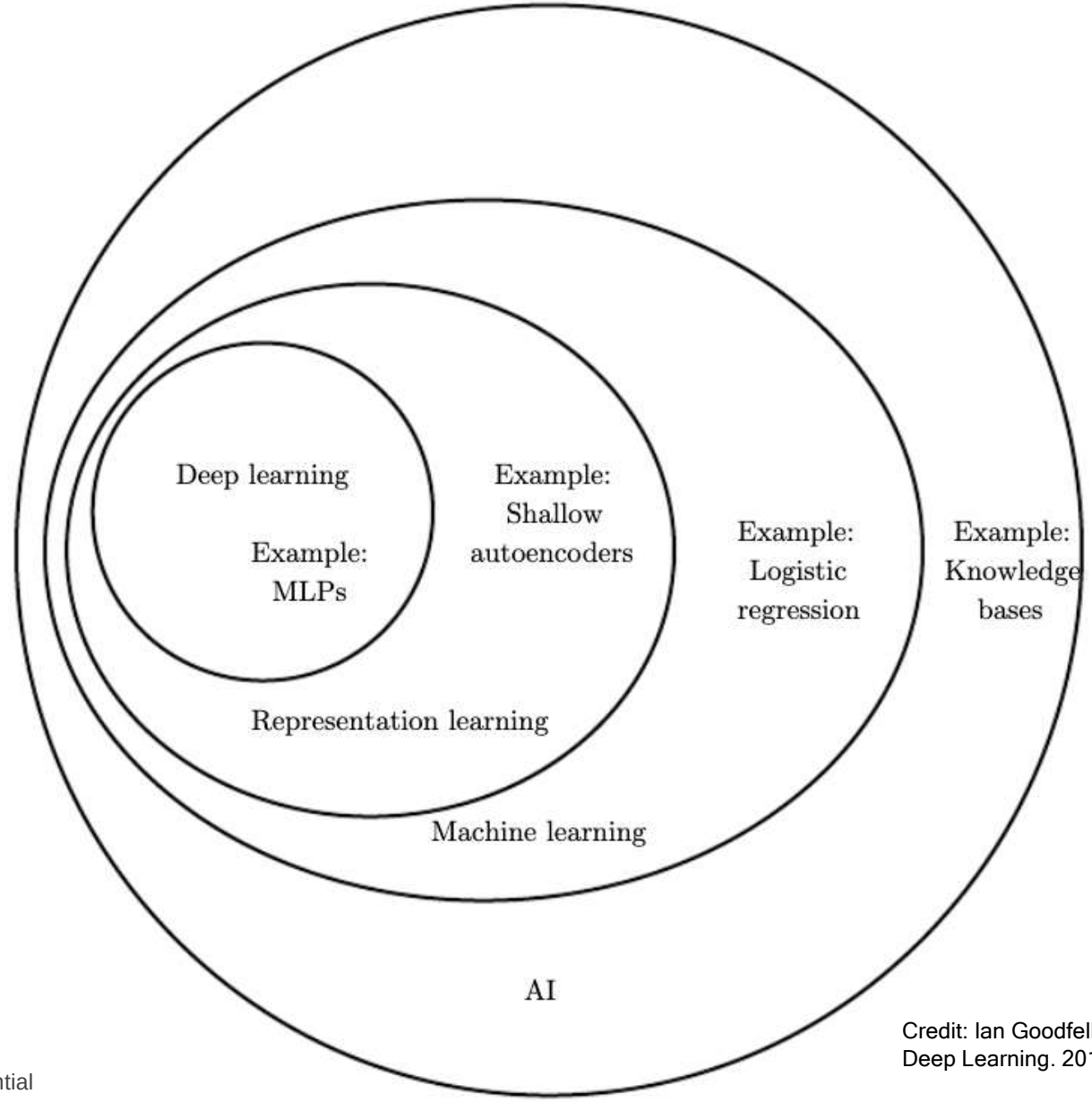


The e4r thrust areas: A holistic approach





Artificial Intelligence v/s Machine Learning



Credit: Ian Goodfellow, Yoshua Bengio, and Aaron Courville.
Deep Learning. 2016. The MIT Press.

AI-augmented Scientific Discovery: Projects @ e4r

✓ Radio Solar Imaging (Astronomy)
Large-scale ML-based data mining of solar data collected from an advanced radio telescope
TIFR-National Centre for Radio Astrophysics, Pune

✓ Star Formation Histories (Astronomy)
Deep learning-based solution to understand galactic properties
TIFR-National Centre for Radio Astrophysics, Pune

Weather Data Analysis (Meteorology)
Machine learning-based solution to understand patterns in weather data
IIT-Bombay

₹ Funded

✓ Completed

AI-augmented Scientific Discovery: Projects @ e4r

✓ Pandora GAN
(Drug Discovery)

Generating antiviral peptides using **Generative Adversarial Network machine-learning technique**

Drug Discovery Hackathon (Govt of India)

₹ Reinforcement Learning for Protein-Ligand Interaction
(Drug Discovery)

Using **machine learning to discover fitment of drug molecules with target proteins**

Drug Discovery Hackathon (Govt of India)

✓ Drug Induced Liver Injury
(Drug Discovery)

Machine learning-based solution to understand side effects of medicines on liver

Drug Discovery Hackathon (Govt of India)

✓ Classification of Anti-microbial Peptides
(Drug Discovery)

Using **machine learning to classify amino acid chains** for anti-microbial properties

Flame University, Pune

₹ Funded

✓ Completed

Cyber-physical Systems, Digital Libraries

TMT CSW ₹ ✓
(Astronomy)

Common services **control system** software

IIA Bengaluru;
TMT Organization, California

TMT DMS ₹
(Astronomy)

Large-scale scientific and engineering **data management system**

IIA Bengaluru;
TMT Organization, California

Sandwich Bot 2.0
(Space Engineering, Smart Cities)

Development of a **generic rover from scratch**, for planetary missions, disaster response, smart cities, etc.

TW-E4R

Large-scale Management of Experiments
(Biology)

Design and development of a framework to manage large number of experiments

A university in Europe

₹ Funded

✓ Completed

Data-intensive Computing, Accelerated Computing

ARTIP
(Astronomy)

₹ ✓

Large-scale **data processing pipeline** for MeerKAT Radio Telescope

IUCAA, Pune

PERC
(Computer Hardware Architecture)

Posit Enhanced Rocket Chip, a **microprocessor design** for **improved number processing**

TW-E4R

RISKA SoC
(Computer Hardware Architecture)

A **system-on-chip** specifically designed for **SKA data processing**

TW-E4R, Raman Research Institute, Bengaluru

Optimizing Data Processing Algorithms
(Radio Astronomy)

Analysis and optimization of data mining algorithms for CPU and GPU architectures

CSIRO, Australia

Storage for A Bio Archive
(Infra Consultancy)

✓ ₹

Consulted to design an exascale storage to archive multiscale biological datasets

European Molecular Biology Laboratory (EMBL), UK

₹ Funded

✓ Completed

Modeling & Simulation

EpiRust (Public Health Policy Making)
Large-scale **epidemic simulations** (100+ million agents population)
TW-E4R

BharatSim ₹ ✓ (Public Policy Making)
Large-scale agent-based framework for **epidemic**, **economic**, and **climate change simulations**
Ashoka University (funded by the Gates Foundation)

Simulation of Planetary Exploration (Space Engineering)
Using multiple rovers for **simulating planetary exploration**
TW-E4R

₹ Funded

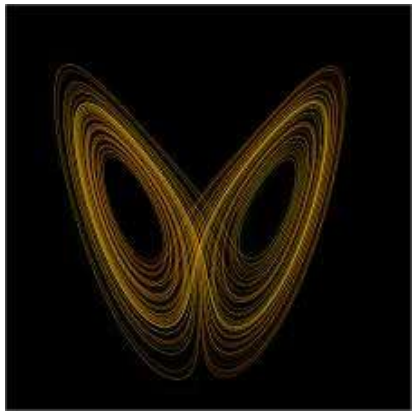
✓ Completed

Formidable trends and Symposia

The Data Deluge

The 4th Paradigm of Science

First TW-E4R Symposium (2018)



Complexity

From Micro to Macro

2nd TW-E4R Symposium (2019)

Artificial Intelligence

Towards A Logic of Discovery: Will AI Ever Win A Nobel Prize?

3rd TW-E4R Symposium (2020)

Fading of Moore's Law

Towards A New Golden Age of Computer Architecture[5]

4th TW-E4R Symposium (2022)

Courtesy

1. https://en.wikipedia.org/wiki/File:Cloudburst_on_phoenix.jpg
2. https://en.wikipedia.org/wiki/Attractor#/media/File:Lorenz_attractor_yb.svg
3. https://www.publicdomainpictures.net/en/view-image.php?image=251903&picture=moores-law-installation
4. Brain Photo by Unknown Author is licensed under CC BY
5. https://cacm.acm.org/magazines/2019/2/234352-a-new-golden-age-for-computer-architecture/fulltext

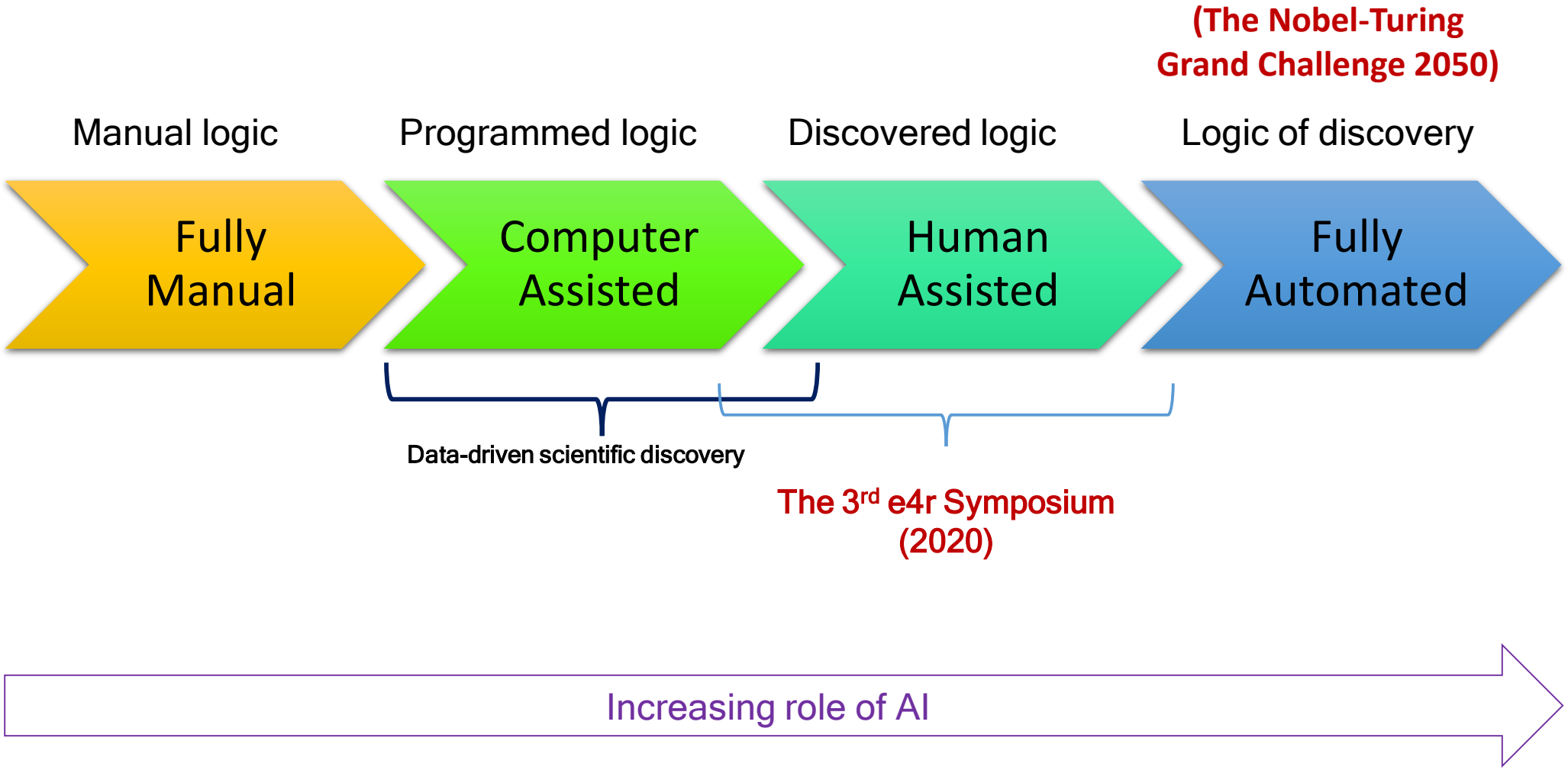
Spectrum of AI-based automation

Human Nobel Laureates ?? ?? AI Nobel Laureates?

Fully Manual → … → … → Fully Automated

Increasing role of AI

Searching for hope amid hype…



A grand scientific pursuit, and
a grand engineering challenge…

The 14th Engineering Grand Challenge in the 21st Century

NATIONAL ACADEMY OF ENGINEERING

ROYAL ACADEMY OF ENGINEERING

CAE

Make solar energy affordable
Provide energy from fusion
Develop carbon sequestration methods
Manage the nitrogen cycle
Provide access to clean water
Restore and improve urban infrastructure
Advance health informatics

Engineer better medicines
Reverse-engineer the brain
Prevent nuclear terror
Secure cyberspace
Enhance virtual reality
Advance personalized learning
Engineer the tools for scientific discovery

1. http://www.engineeringchallenges.org/
2. See also: 10 Big Ideas for Future NSF Investments (https://www.nsf.gov/about/congress/reports/nsf_big_ideas.pdf)

ENGINEERING FOR RESEARCH (e4r™)

e4r@thoughtworks.com

thoughtworks.com/engineering-research

A few ML-related challenges

- Too much data → Scaling of ML methods (samples as well as dimensions), labeling efforts, skewness
- Too less data → Few-shot, zero-shot learning
- Missing data/lower signal-to-noise
- Incorporation of domain knowledge
- Interpretation of results
- Reproducibility
- Explanability
- Logic of discovery → New physics, biology, materials, etc.

Moonshots

BHARATSIM (2020 - 2025)	• India's first ultra-large-scale agent-based simulation framework • Epidemiology and beyond into socio-economic policy making
RISKA (2020-2025)	• Designing an open source system-on-chip for processing SKA Radio Telescope data
PLANETARY EXPLORATION (2020-2025)	• Simulation of multiple rovers collaborating on a planetary mission • Design and construction of actual exploration vehicles
DRUG DISCOVERY (2020-2025)	• Using ML, knowledge graph, and evolutionary computing • Knowledge graphs for protein structural coverage • Accelerated computing • Towards neuro-symbolic computing

Select publications and talks

RISKA: Towards an Open-source RISC-V based Domain-specific System-on-Chip for SKA Data Processing **(CARRVW at ISCA 2021)**

A machine-learning based algorithm to study compact features in the solar image plane **(ADASS 2020)**

Simulating Re-configurable Multi-Rovers For Planetary Exploration Using Behaviour-based Ontology **(WinterSim 2020)**

Alphabet reduction and distributed vector representation based method for classification of antimicrobial peptides **(IEEE BIBM 2020)**

EpiRust: Towards a framework for large -scale agent based epidemiological simulations using Rust language **(SIMS 2020)**

Posit Enhanced Rocket Chip **(CARRV Workshop at ISCA 2020)**

Predicting star formation histories of galaxies using deep learning **(MNRAS 2020)**

Machine Learning for Scientific Discovery **(ADASS 2019)**

Detection of OH radical **(APJ 2018)**

A compositional lens on the drivers of complexity **(ICCS 2018)**

Revealing HI gas in a galaxy **(APJ 2018)**

Large Scale Simulation of Heterogenous Multiple Rovers – An Experience Report **(GLEX 2020; Cancelled due to COVID19)**

Modeling at the Speed of Thoughts **(Conference on Complex Systems 2019)**

Actor Based Architecture for World's Largest Telescope **(Reactive Summit 2018)**

Automated Data Processing for uGMRT and MeerKAT Absorption Line Surveys **(H1 Absorption Conference 2018)**

Service Discovery using CRDTs **(Reactive Summit, Austin, 2017)**

Scholarly contributions: 24 talks + papers;

ThoughtWorks Insights: 4 insight articles

In addition, a white paper with BFSI service line

E4R Team

Computer science

- Data structures & algorithms
- Distributed systems
- Operating systems
- Programming languages
- Database systems

Domains

- Astronomy/Cosmology
- Bioinformatics
- Computer hardware
- Mechatronics/Robotics
- Economics & finance

Data science/AI

- Machine learning
- Knowledge graphs
- Data mining
- Reinforcement learning

Scientific computing

- Supercomputing/high performance computing
- Data intensive computing
- Accelerated computing
- Modelling & simulation

Misc.

- Tech journalism/writing
- Behavioral sciences