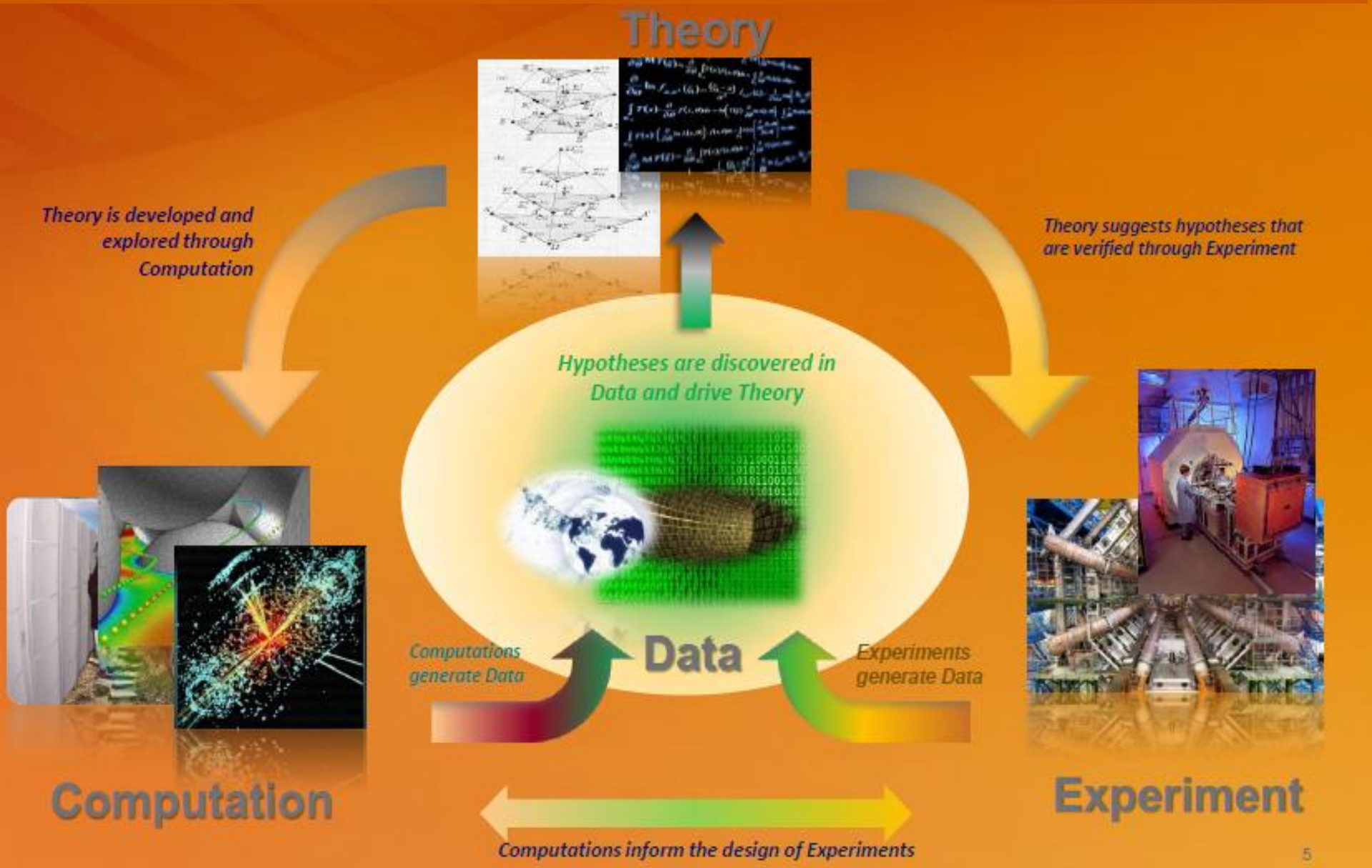
Current Status and Future Trends in Data Centric Science

Prof. Yike Guo

Director， Data Science Institute
Imperial College London

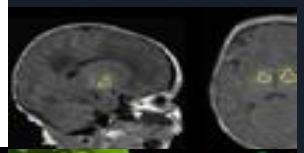
Science Today Is Data Centric



Datafication Turns Sciences into Data Science

Medical

Systems oncology

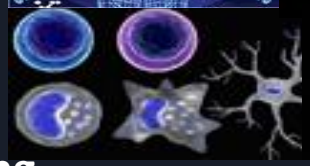


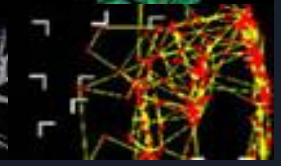
Imaging

Realtime metabolic profiling

Infection/epidemiology

Cardiovascular science

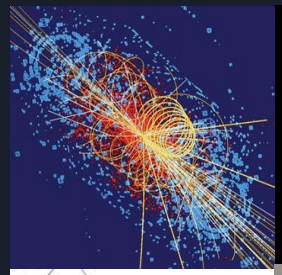




Neuroscience

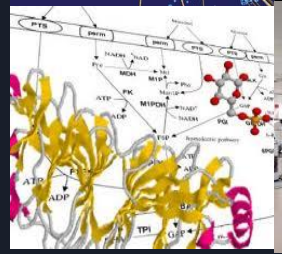
Natural Sciences

Particle physics



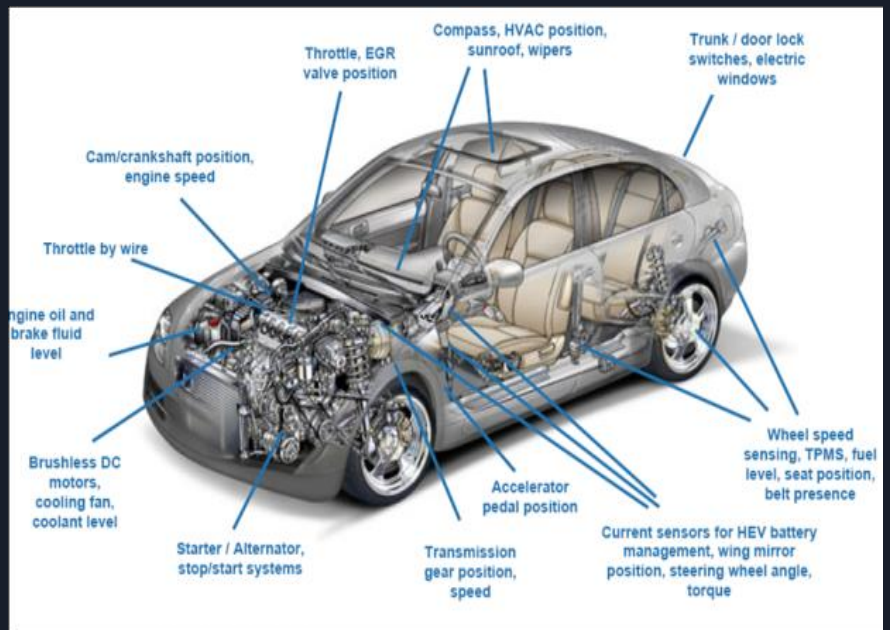
Complex systems & network data analytics

System biology



High-throughput screening

Engineering

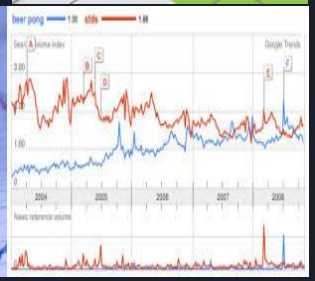


Business

Social media & new data business

Digital cities & urban life

Algorithmic trading



Public health management

Digitalisation vs Datafication

Digitalisation: is a process that has been active within society since the late 1950s, with the birth of the emiconductor industry. It refers to the conversion of pieces of information into digital formats, for example text into HTML pages, music into MP3s, images into JPEG or similar. As the process of digitalisation has progressed, the amount of data that could be processed has increased exponentially. Digitalisation, therefore, from a simplistic perspective may be viewed as the embodiment of idea creation – it is capturing human ideas in digital form for transmission, re-use and manipulation

Datafication relates to the use of digital technologies to unembed the knowledge associated with physical objects by decoupling them from the data associated with them. Datafication is manifesting itself in society in a variety of forms and is often – but not always – associated with sensors/actuators and the emerging Internet of Things (IoT). Datafication may take many forms and in many cases a mobile device is enough to create unembedded knowledge of a person, a thing or a piece of infrastructure.

Ericssion White Paper : Impact of Datafication on the Strategic Level

Example: Dataficating Life

Technology

_ Wearable sensors: capturing personal physiological and behavioural information

_ Cloud: data analysis

Impact

_ Enabling **real-time** health monitoring and behaviour characterisation

_ The foundation: **personalisation of products and services**

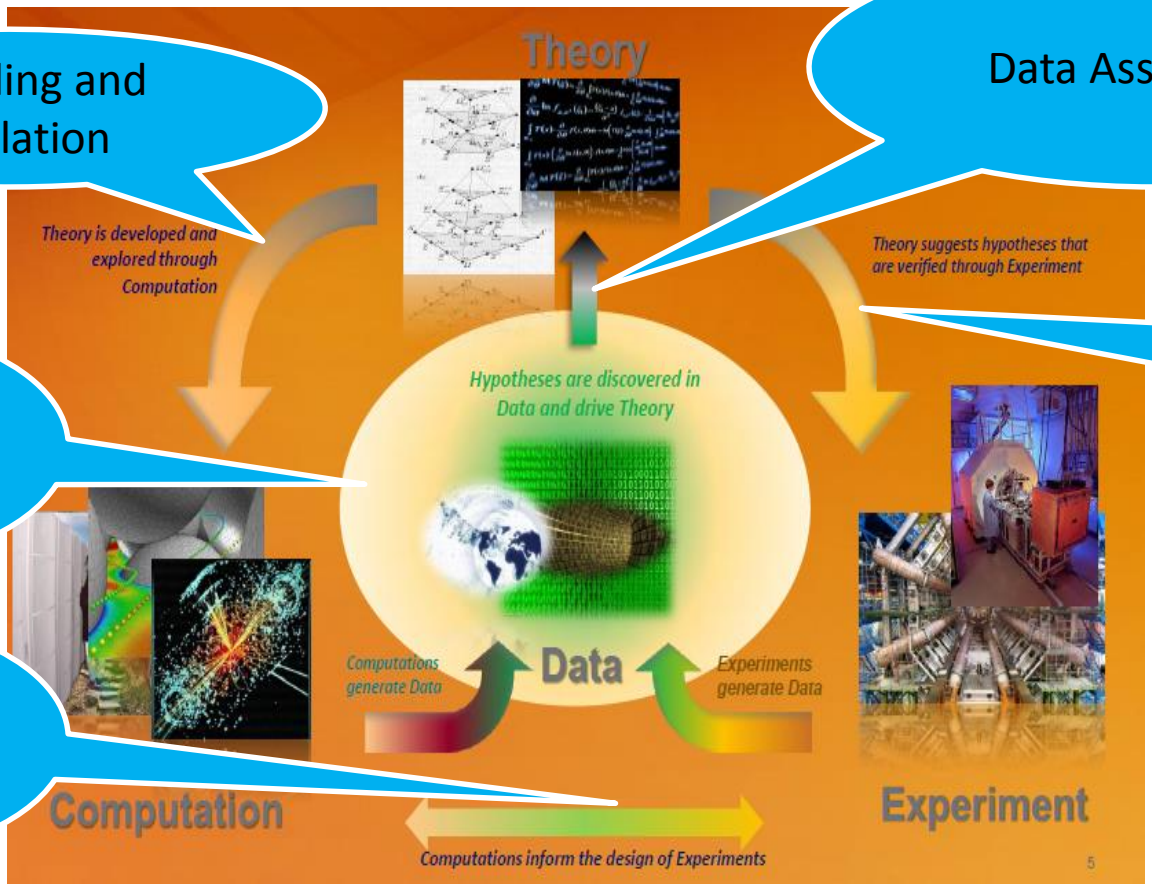
Future trends

_ Ecosystem: **personalised services**

_ Integrated data products: combined with personal biological data for **personalised medicine**

_ Real-time decision support: **mobile health monitoring**

Technologies for Data Centric Science



Modelling and Simulation

Data Assimilation

Experimental Design

Big data Management

Analysis & Machine Learning

4I of Data Centric Science

Integration: Integrating data for system analysis

Intelligence: Machining learning for deep understanding and prediction

Interaction: Integrating data with models and physical systems for adaptive analysis

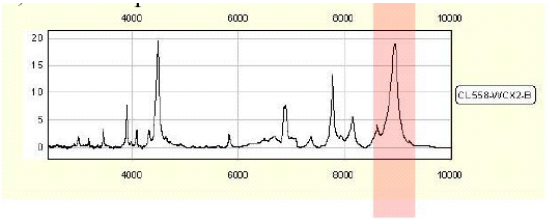
Inter-discipline: Understanding complexity by cross-disciplinary study with data as glue

Being Data Centric => Integrative Analysis

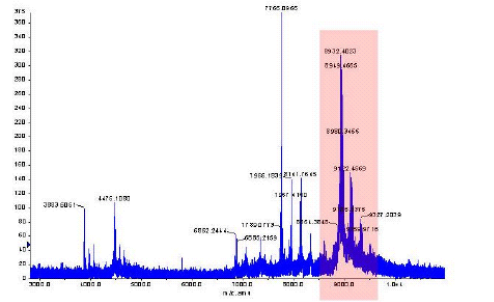
- Data centric research requires to collecting data measuring the various aspects of a physical system
- Collected measurements are required to be calibrated and meaningfully integrated.
- The meaningful integration explores the inherent relationships of different modalities of data
- Exploration the relationship requires deep analysis and curation
- Data integration is the core of “ Web Science”

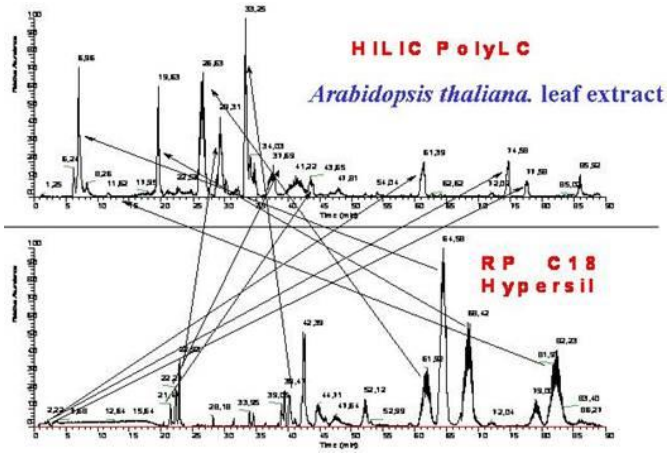
BIOLOGY AS DATA SCIENCE

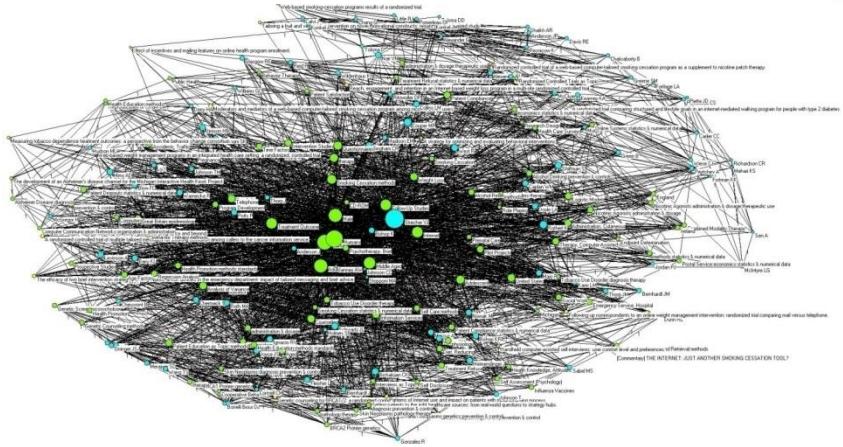
- Genomics
- Proteomics
- Metabolomics
- Phenome
- ……



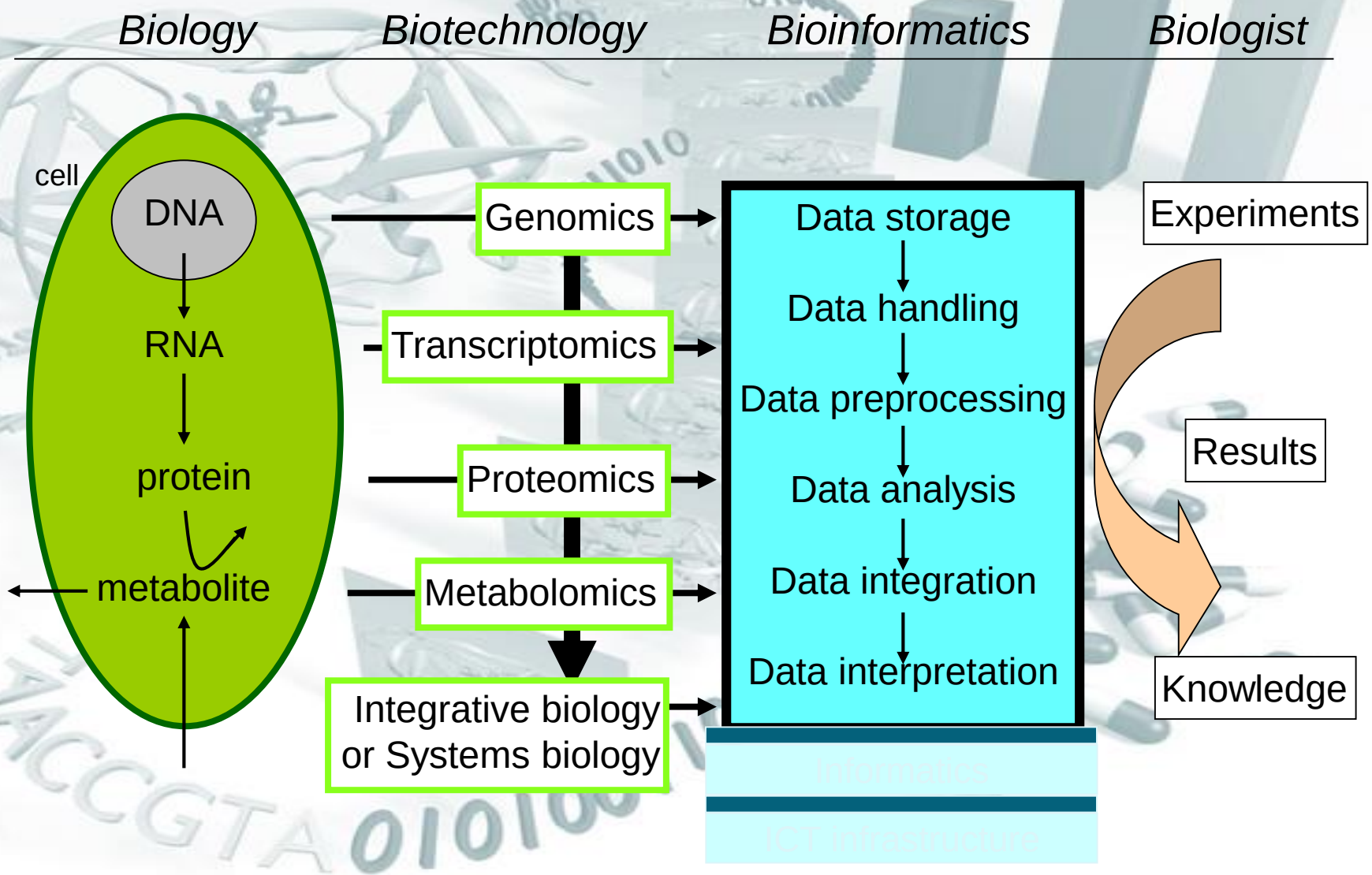
The ABI QSTAR® Pulsar Hybrid LC/MS/MS System is a high performance hybrid quadrupole time-of-flight mass spectrometer





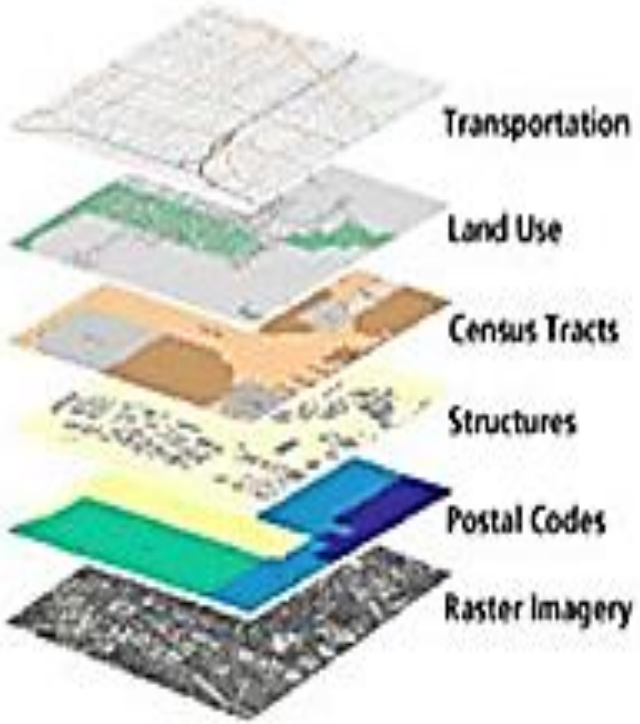


Biology is now a data science

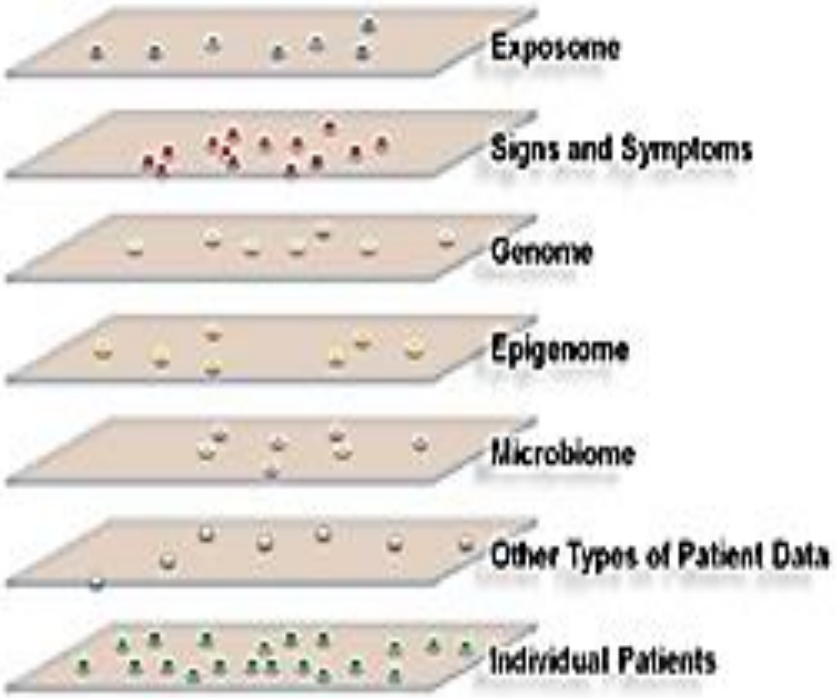


Medicine is now a data science

Google Maps: GIS layers
Organized by Geographical Positioning

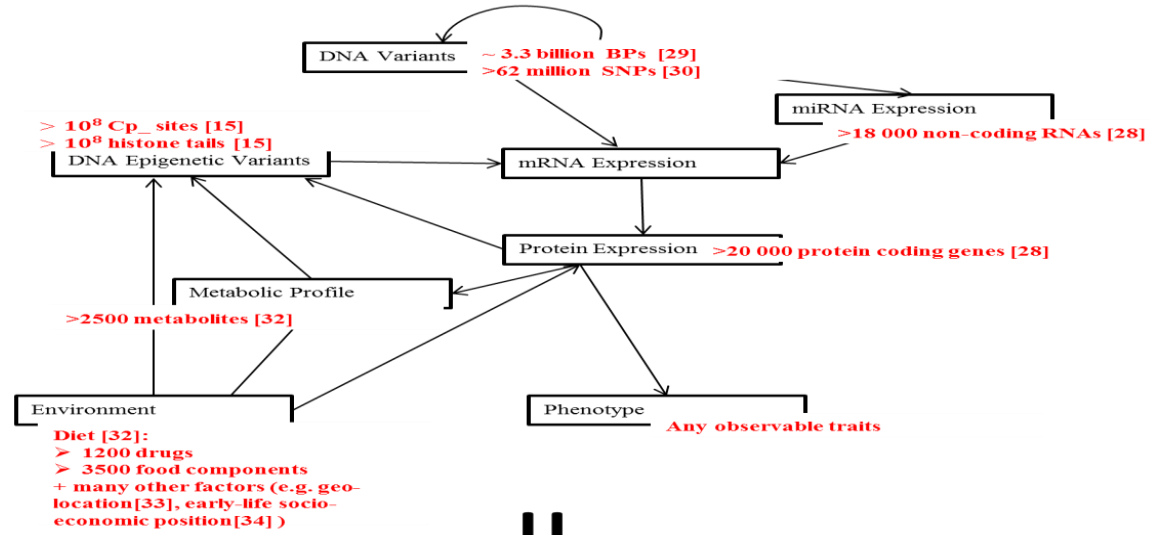


Information Commons
Organized Around Individual Patients



Translational Research : Computing Correlations

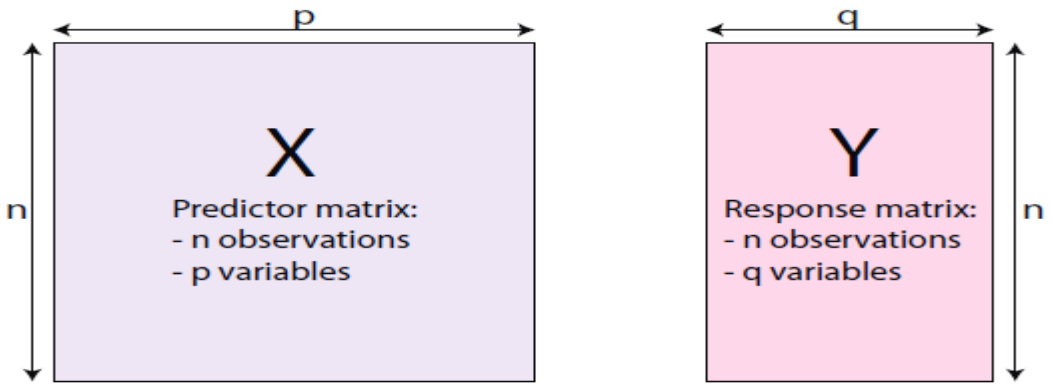
Human Data Size



X n individual samples

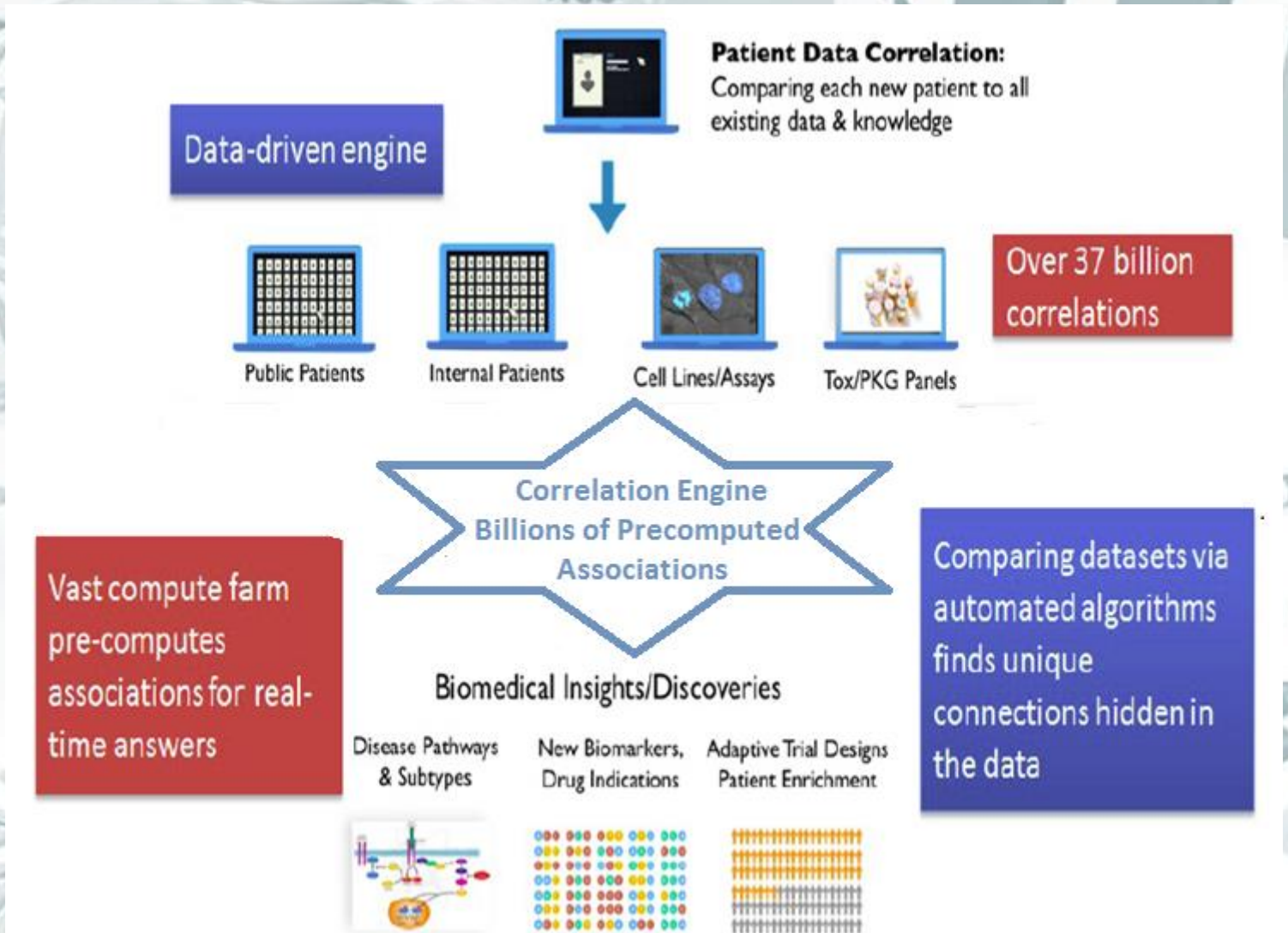
=

Combinatorial Explosion



Aim: identify which of the *p* variables in *X* are significantly associated with the outcome *Y*

Data Driven Medicine: Searching associations

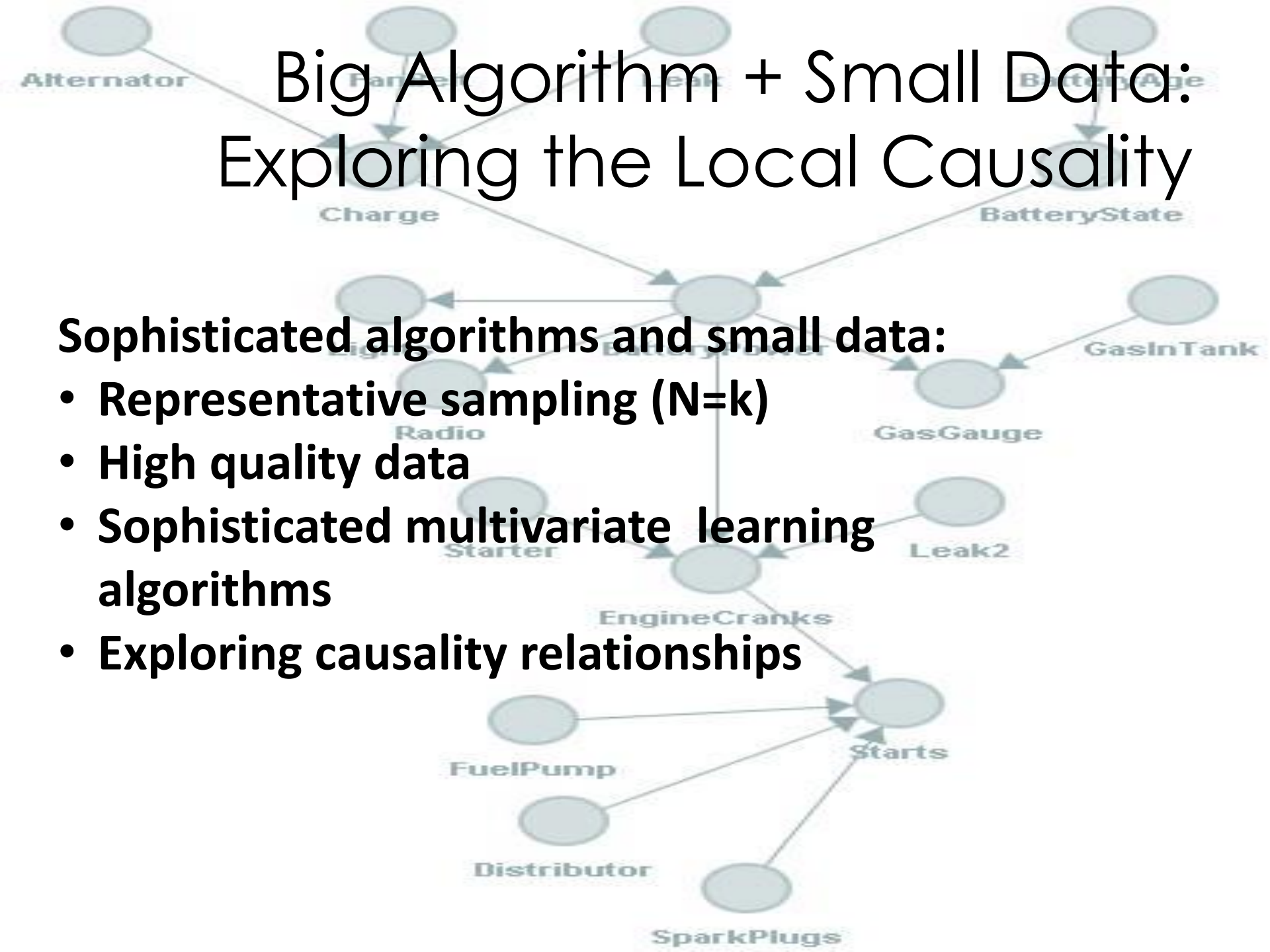


Being Data Centric => Intelligent Analysis

- Data Centric means discovery and predictive
- Discovery requires machine to discover patterns and trends beyond statistical analysis
- Predictive requires machine to build models exploiting the insight from data
- The era of “Machine Science” is coming

(king et al, “The Automation of Science”, Science 3 April 2009, Schmidt M and Lipson H, “Distilling Free-Form Natural Laws from Experimental Data”, Science 3 April 2009)

Big Algorithm + Small Data: Exploring the Local Causality



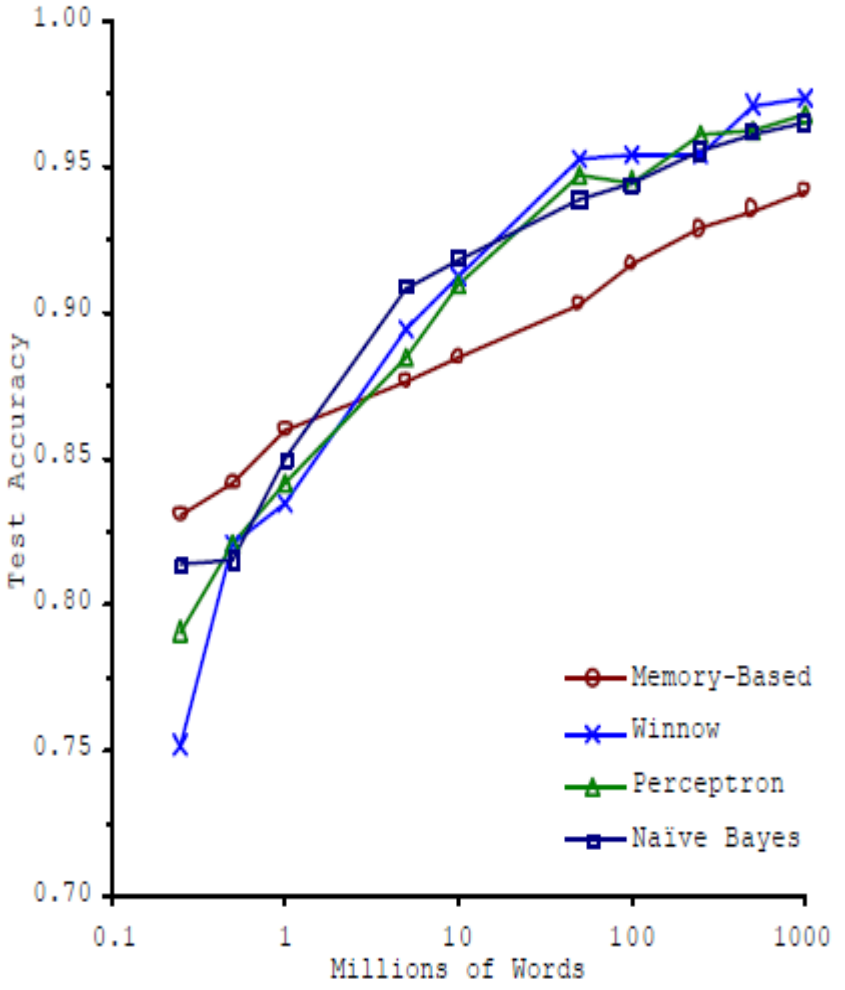
Sophisticated algorithms and small data:

- **Representative sampling (N=k)**
- **High quality data**
- **Sophisticated multivariate learning algorithms**
- **Exploring causality relationships**

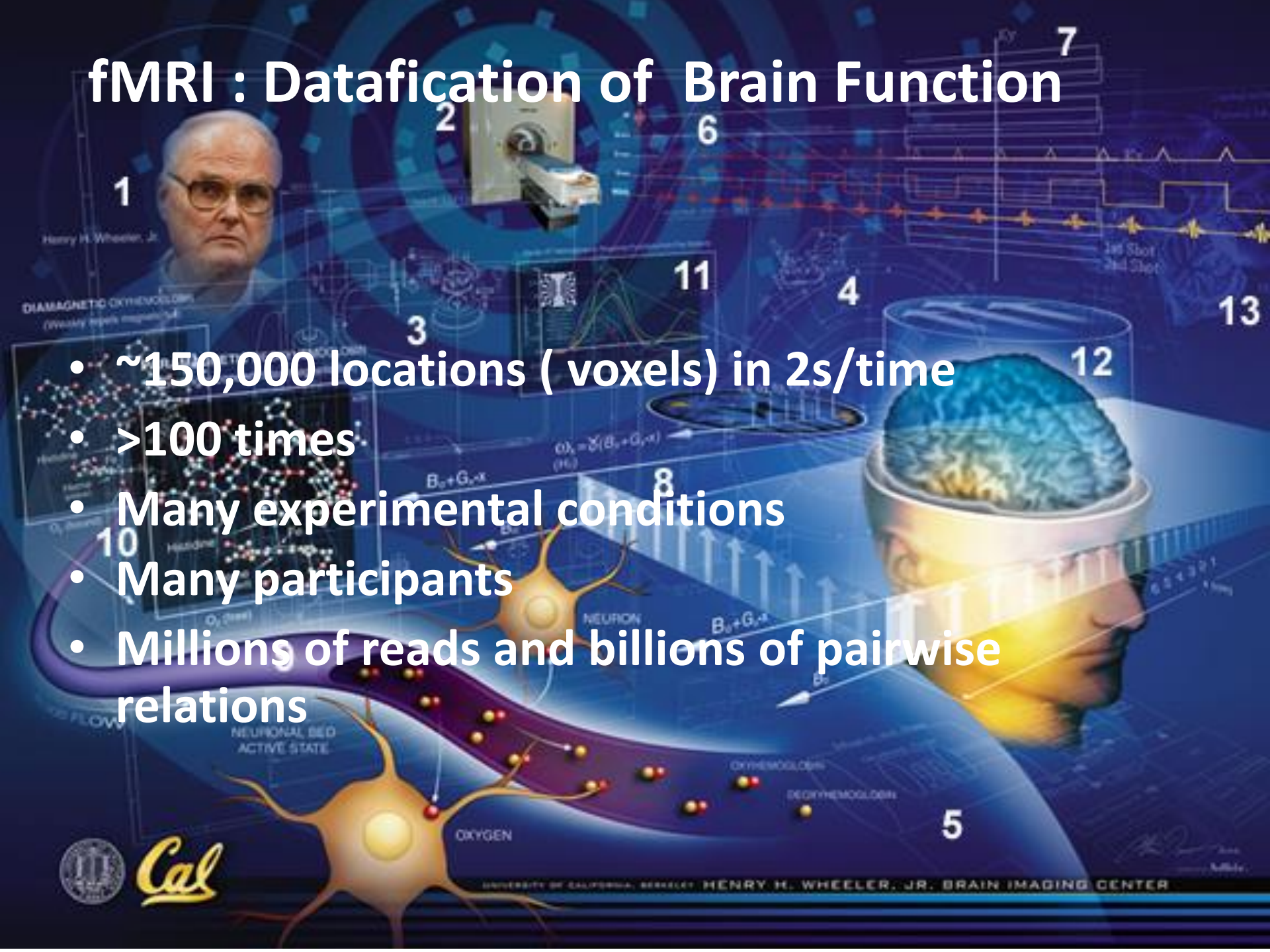
Small Algorithm + Big Data: Exploring Global Correlation

Simple algorithms and big data:

- **Taking all the data (N=all)**
- **Messy data tolerance**
- **Simple/scalable learning algorithms**
- **Exploring correlation relationships**

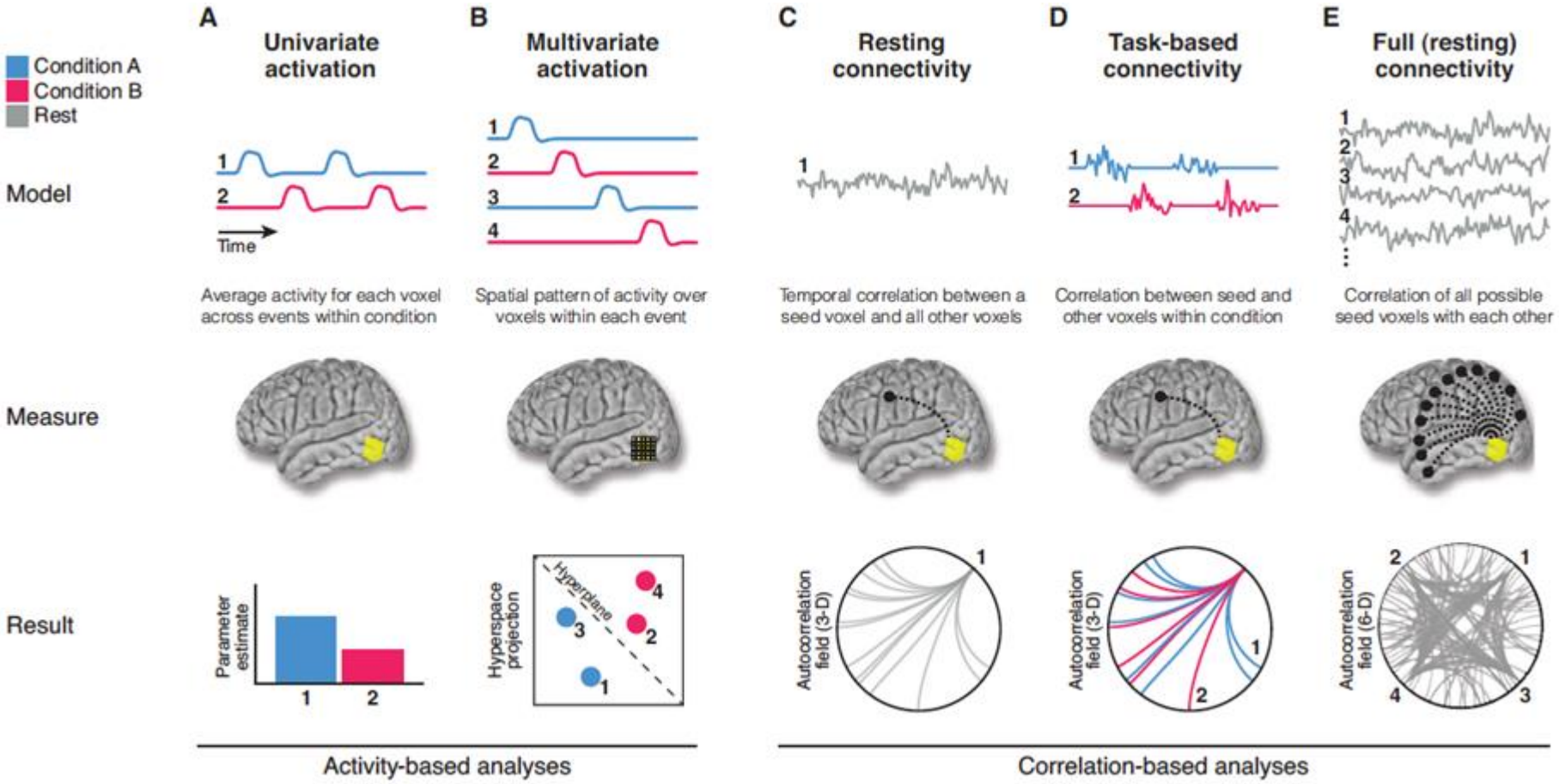


fMRI : Datafication of Brain Function



- ~150,000 locations (voxels) in 2s/time
- >100 times
- Many experimental conditions
- Many participants
- Millions of reads and billions of pairwise relations

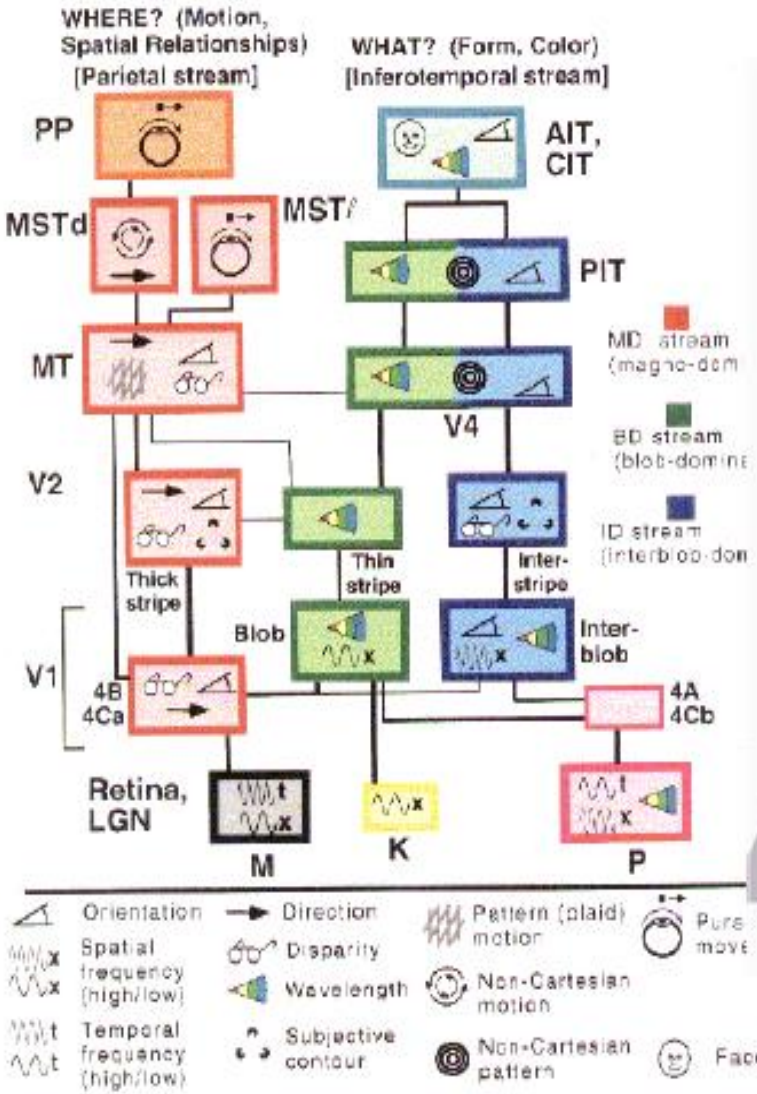
fMRI Analysis

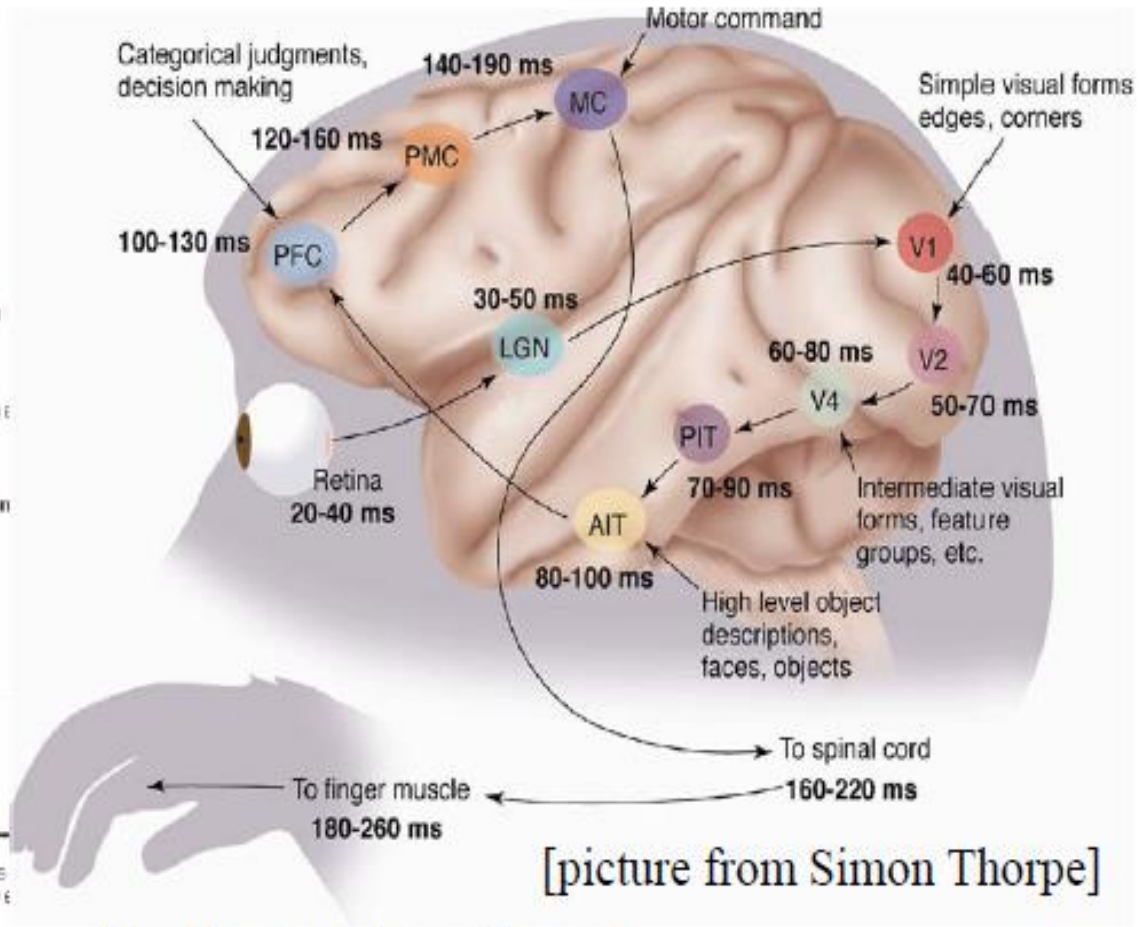


Small Data → **Big Data**

Big Algorithms + Big Data : Cognition

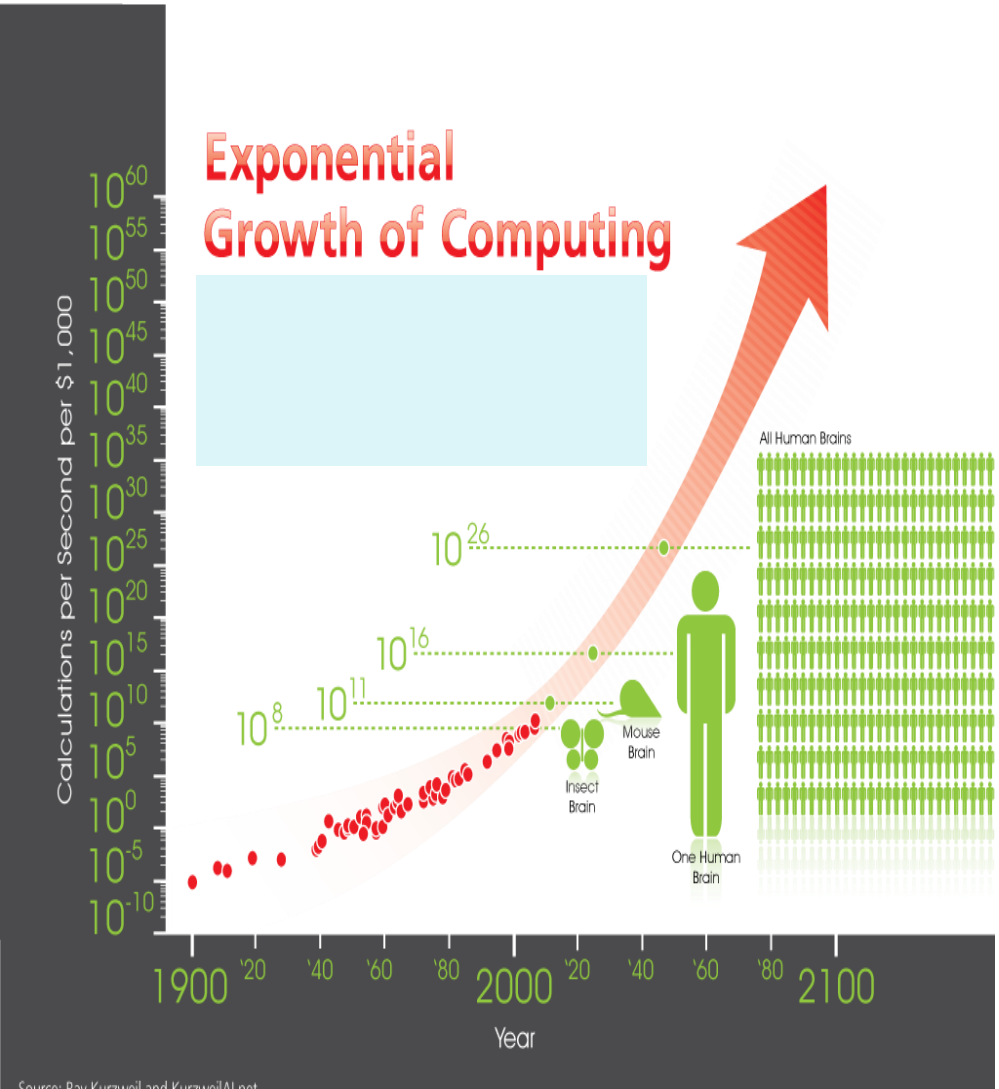
Deep Learning: From Vision to Cognition





[Gallant & Van Essen]

Machine Intelligence Drives the Analytical Technology towards Cognition



Technology

_ Deep learning, Secure learning and NLU

Impact

_ **Unlimited, centralised analytics/computing capacity**

- Knowledge discovery in real time with big data

Future trends

_ Machine cognition

_ Machine Science

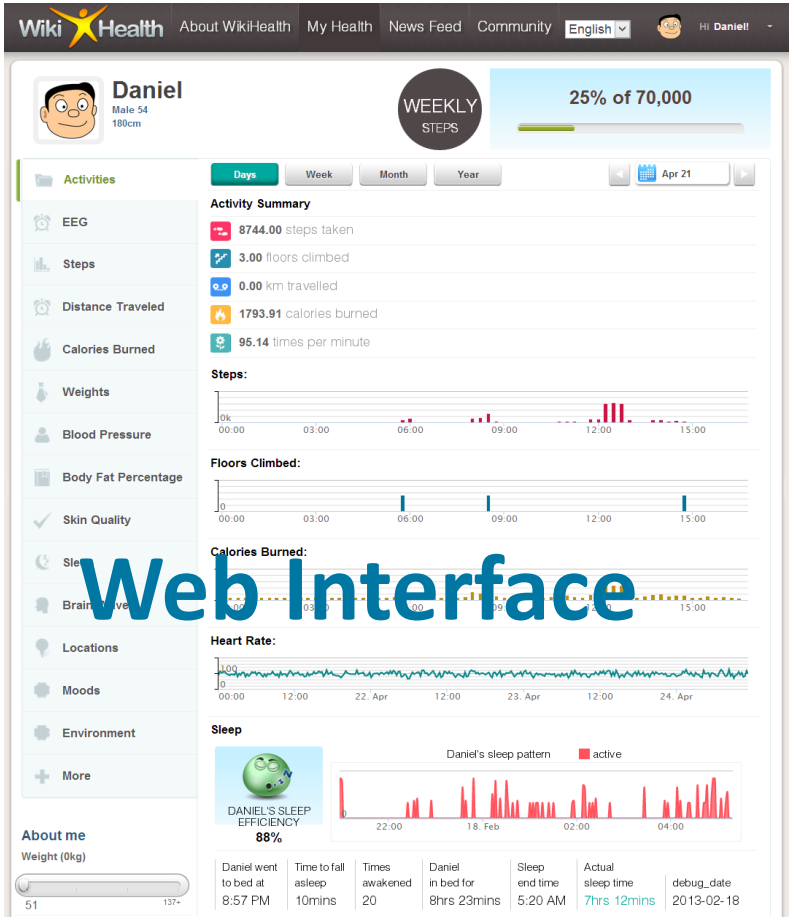
-- Real time discovery

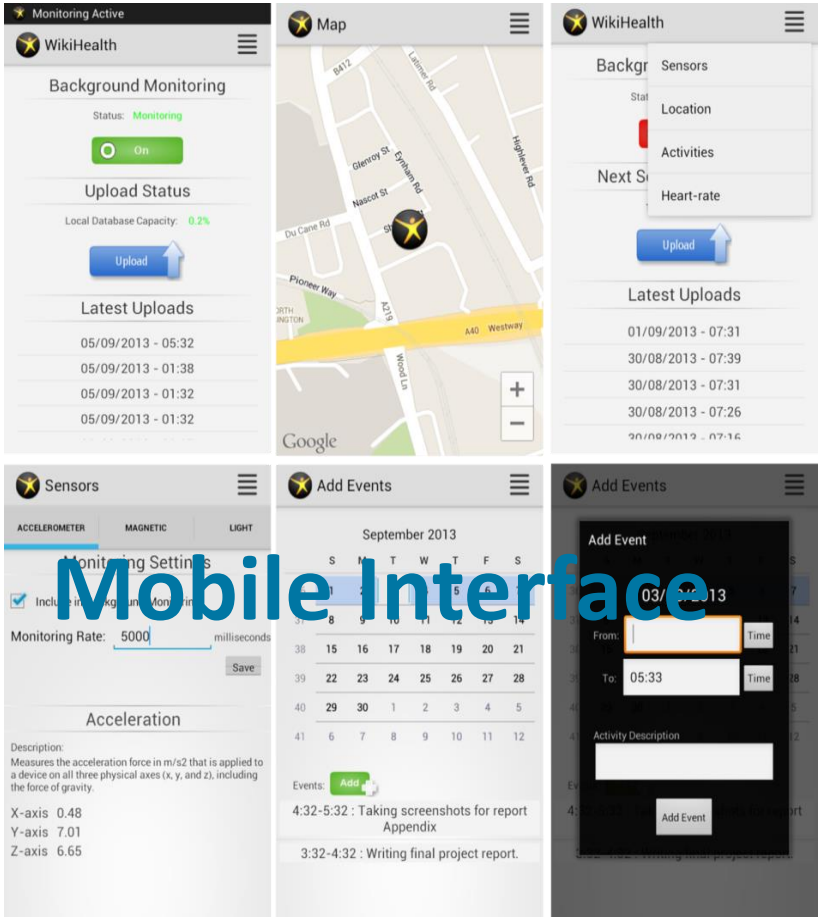
-- Model based knowledge economy

Being Data Centric => Let Data Speak

- Data Centric requires data to be interactive with other entities in the research (models, physical world and human)
- Such interaction enables adaptive decision making
- The adaptions include : sampling strategies, model parameters, visual understanding
- Interaction suggests a “ Data Chemistry” !

WIKIHEALTH : AN INTEGRATED PLATFORM FOR WEARABLE SENSOR INFORMATICS

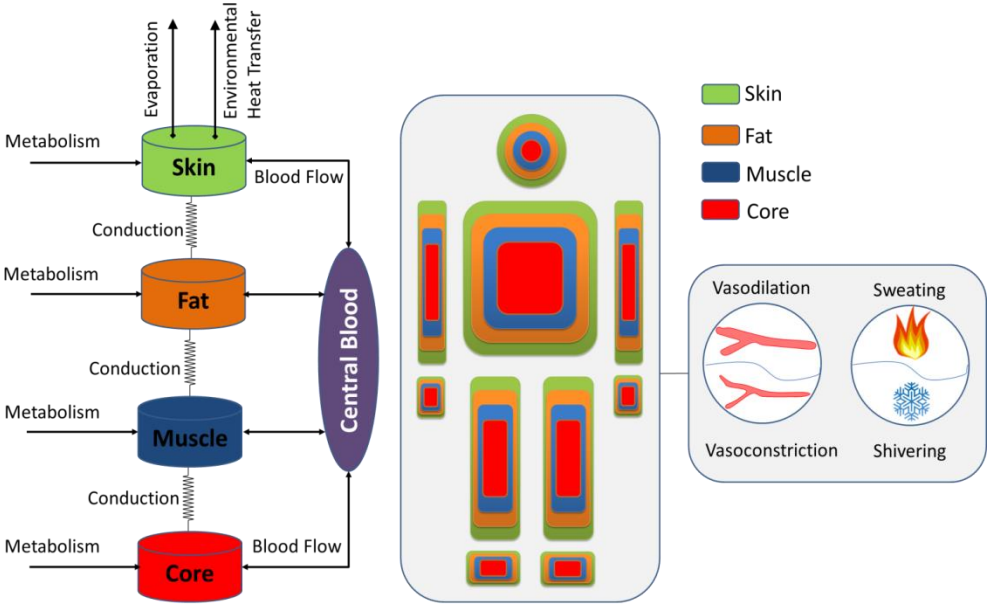


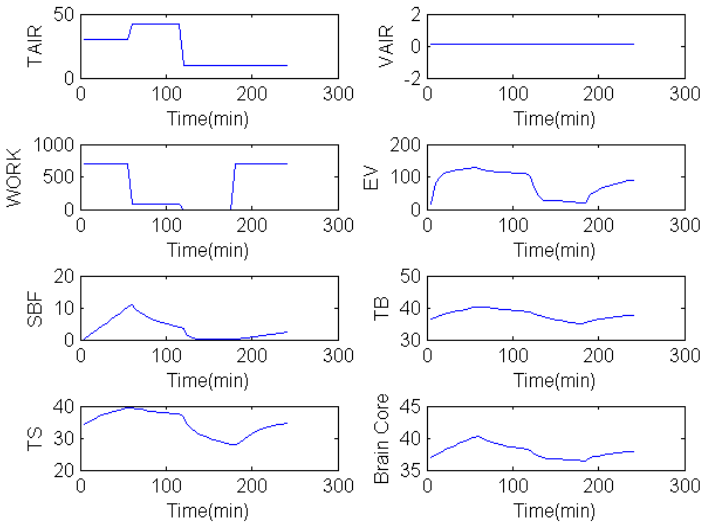


Web Interface **Mobile Interface**

Wiki-Health: Assimilation with human physiology models

- Example of simulating full body temperatures and energy transformation
- Core body temperature & Safety
  - Too high: heat exhaustion, heatstroke
  - Too low: hypothermia





- Age: 27, Weight：80kg
- 1- 60 min running at a speed of 6mph with air temperature 30°C
- 60-120 min sit still with air temperature 42°C
- 120-180 min sit still with air temperature 10°C
- 180-240 min running at a speed of 6mph with air temperature 10°C

Chemistry and Data Chemistry

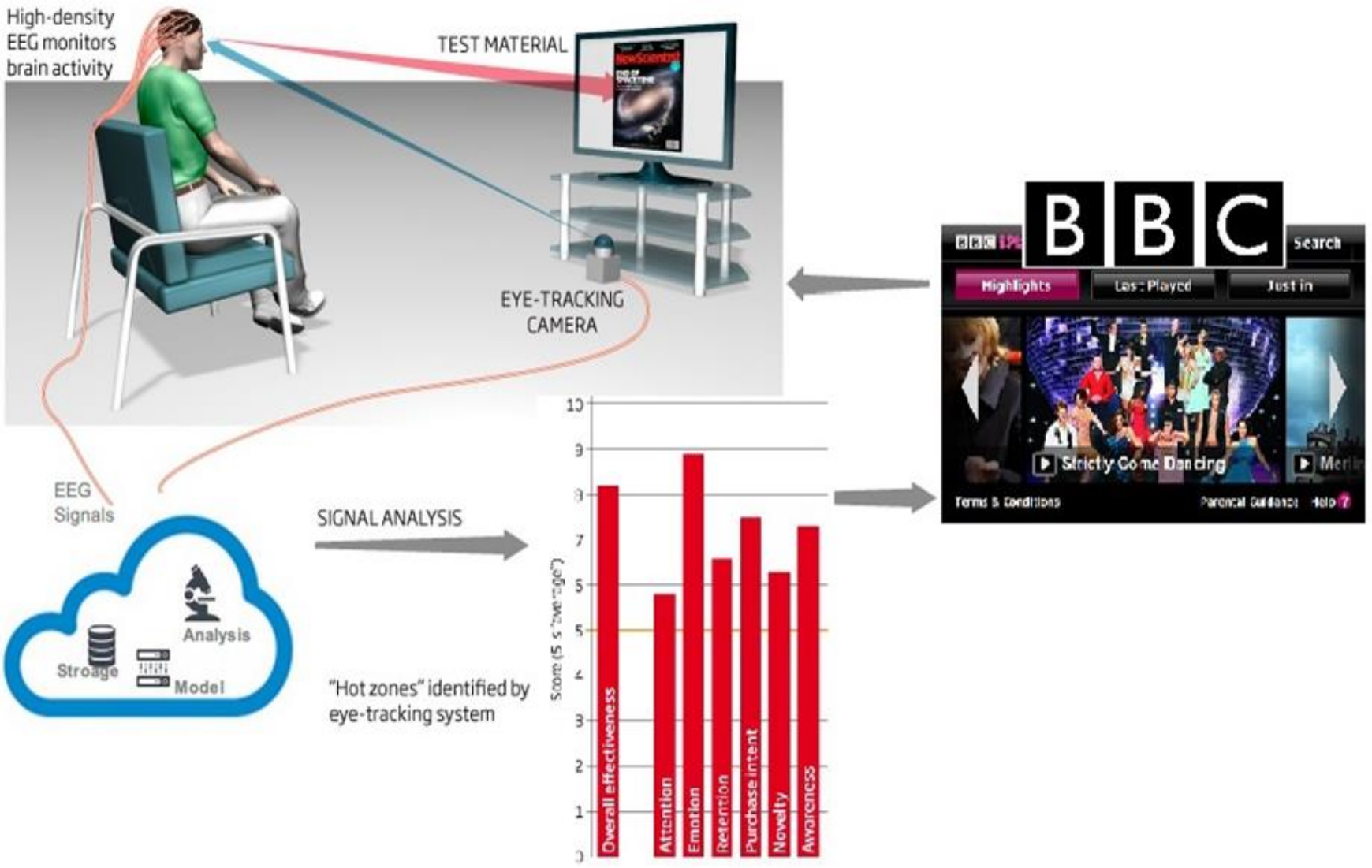
n.

The study of the composition, structure, properties, and reactions of ~~matter~~ data (everything that makes up the ~~universe~~ knowledge).

The Basic Concept of Data Chemistry

- Element types : data, model
- Properties:
  - Semantics
  - Provenance
- Structures:
  - Representation
  - Relation
  - Distributionn (Statistics)
- Reaction:
  - Data-data reaction : integration
  - Data-model reaction : assimilation
  - Model-model reaction : knowledge network
- Derivatives :
  - Data exhausts
  - Models

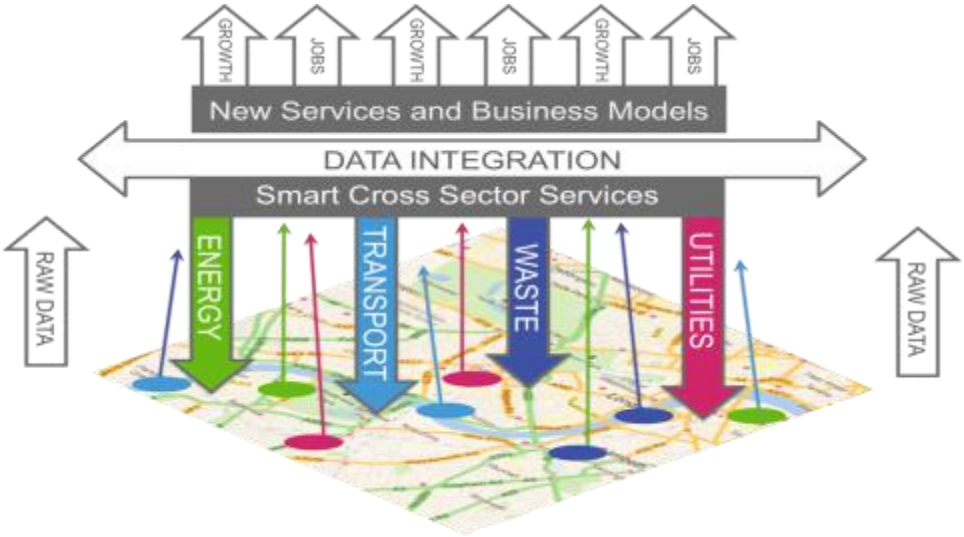
Data/Model Interaction Is Essential in Data Product Innovation



Being Data Centric => Data as Glue

- Datafication made the interaction and integration of scientific disciplines easier
- Data enables a systematic research of a complex system through integrative analysis
- System to system level integration can be achieved via data/model interaction
- Inter-disciplinary research is changing the research organisation structure.

Digital City Exchange : Exploring Data Economy of Smart City



transport
energy
water
waste

harnessing city data

+ new analytics →

Digital City Exchange open data platform

new insights into managing city services

new integrated city services

infrastructure operators

city authorities

developers

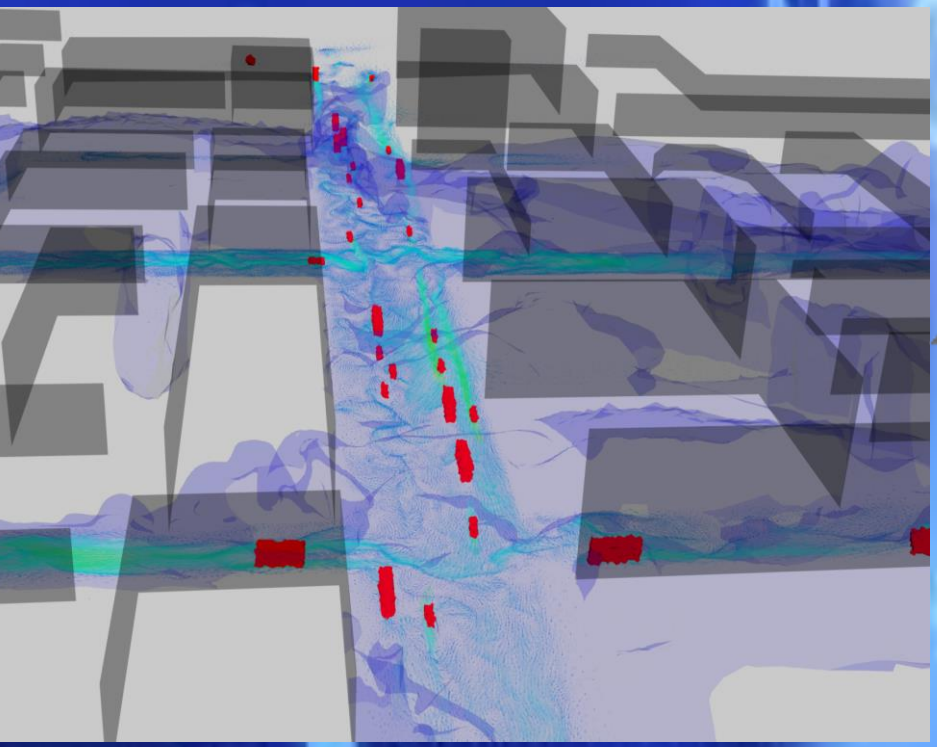
Imperial College London

Digital City Exchange

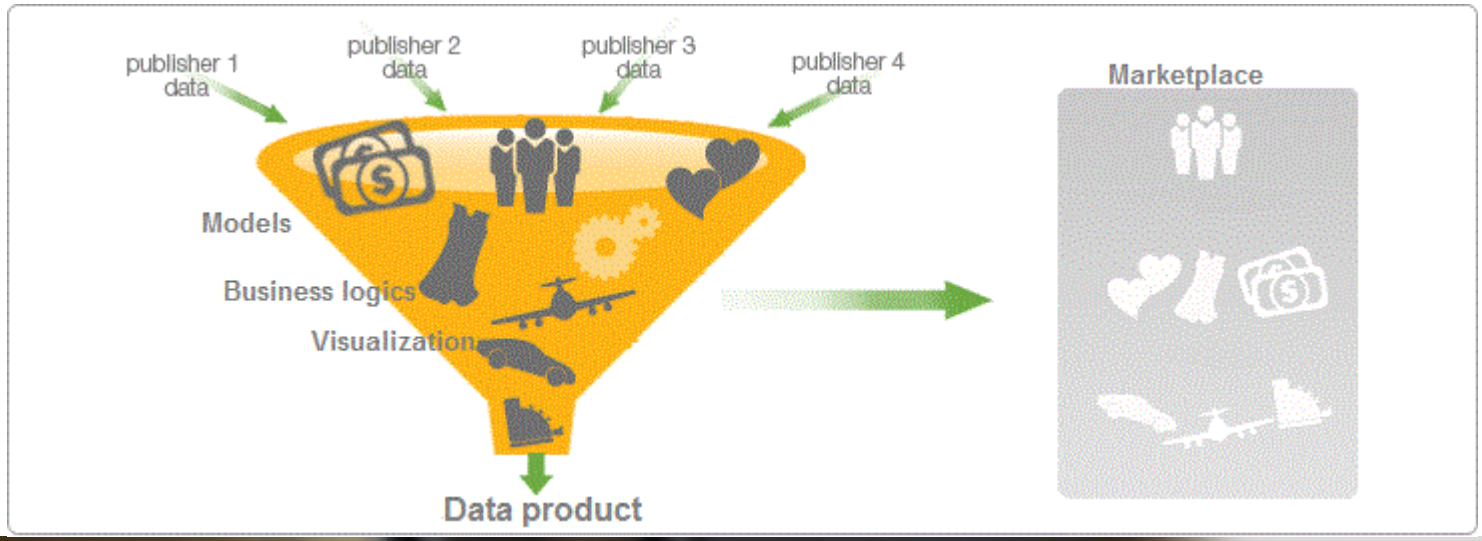
Research Councils UK Digital Economy Transforming Business and Society

City Air Quality Research

Traffic flow, car emission data and weather condition will generate a dynamic map for the air quality of a city.

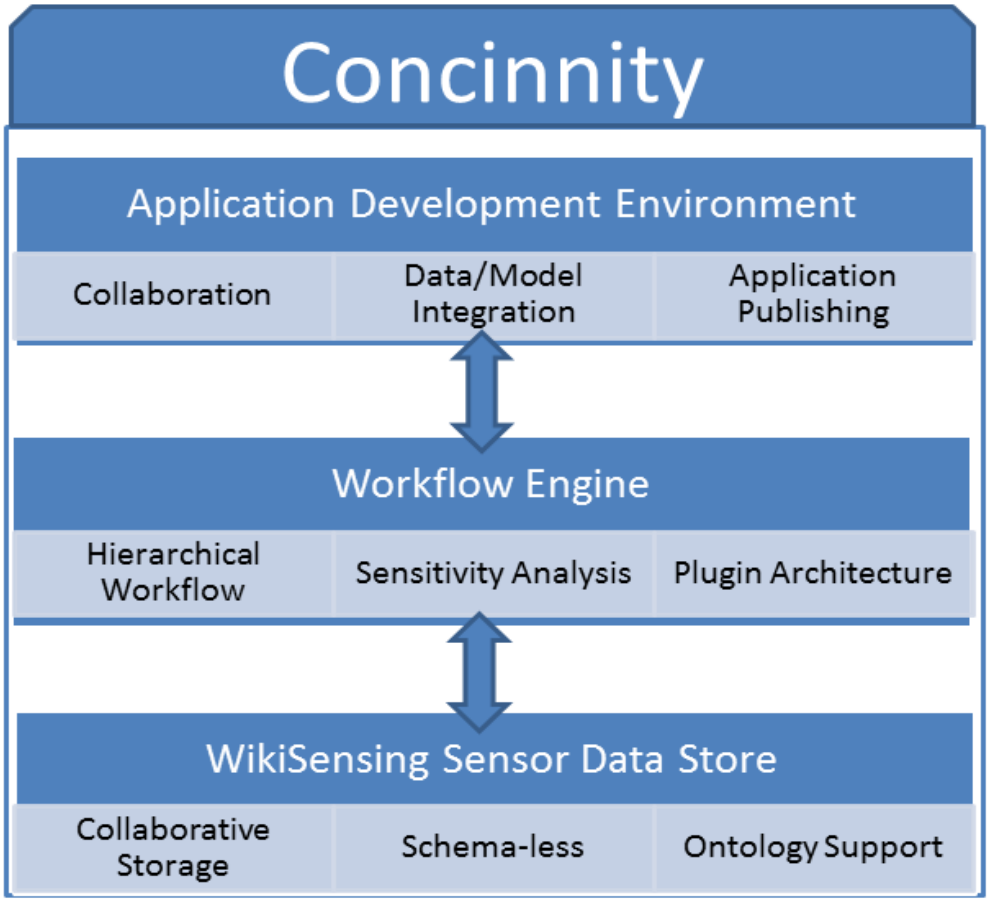


Building Data Products with Urban Informatics structure



Concinnity: The Digital City Exchange platform

- A generic sensor data management platform “Concinnity” [1] built on Wikisensing [2] datastore and Wikimodelling [3] model integration workflow engine
- 3 layers targeting different stages of the data lifecycle
- Data products built using the application development environment



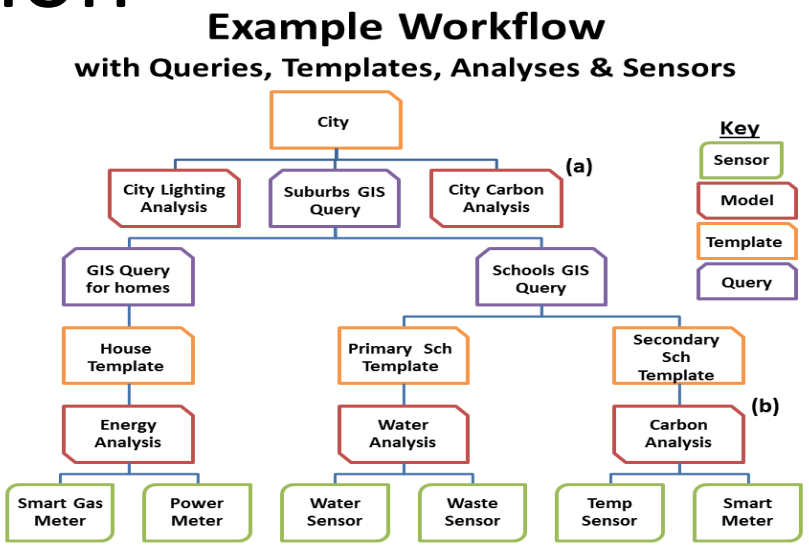
[1] Lee, C-H, David Birch, Chao Wu, Dilshan Silva, Orestis Tsinalis, Yang Li, Shulin Yan, Moustafa Ghanem, and Yike Guo. “Building a Generic Platform for Big Sensor Data Applications.” *Proceedings of the IEEE Big Data conference* (2013), Santa Clara, CA, USA.
[2] Silva, Dilshan, Moustafa Ghanem, and Yike Guo. "WikiSensing: An Online Collaborative Approach for Sensor Data Management." *Sensors*, 12, no. 10 (2012): 13295-13332.
[3] Birch, David, Paul HJ Kelly, Anthony J. Field, and Alvise Simondetti. "Computationally unifying urban masterplanning." *Proceedings of the ACM International Conference on Computing Frontiers*, p. 32. 2013.

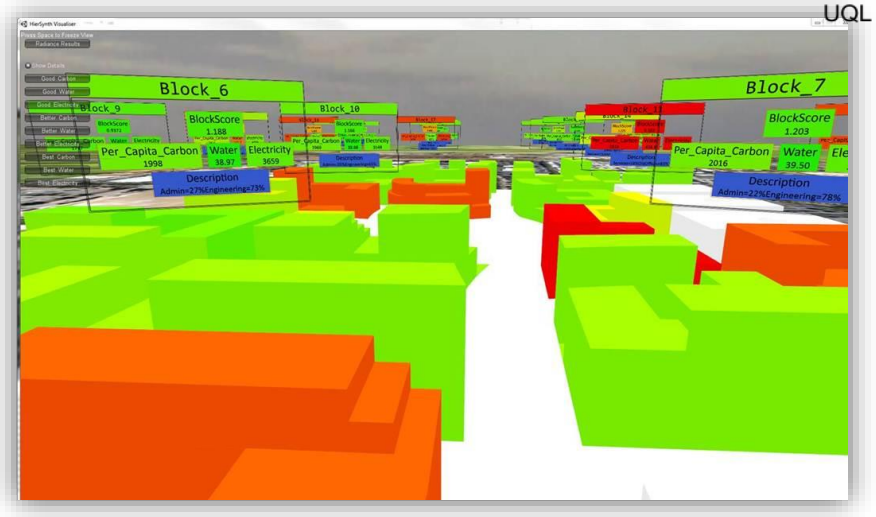
Semantics Engine : Model and Data integration

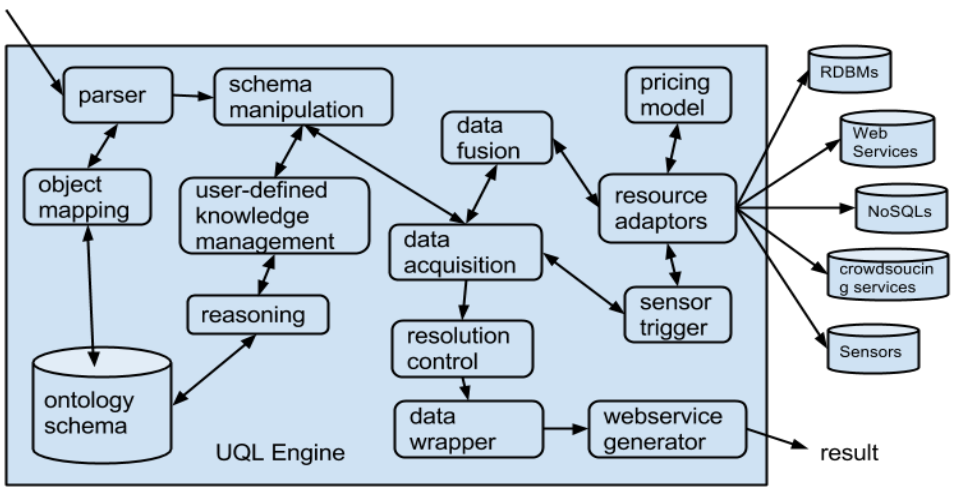
Allows workflows to reflect the hierarchies

- Geographically in the city
- Temporally in data streams

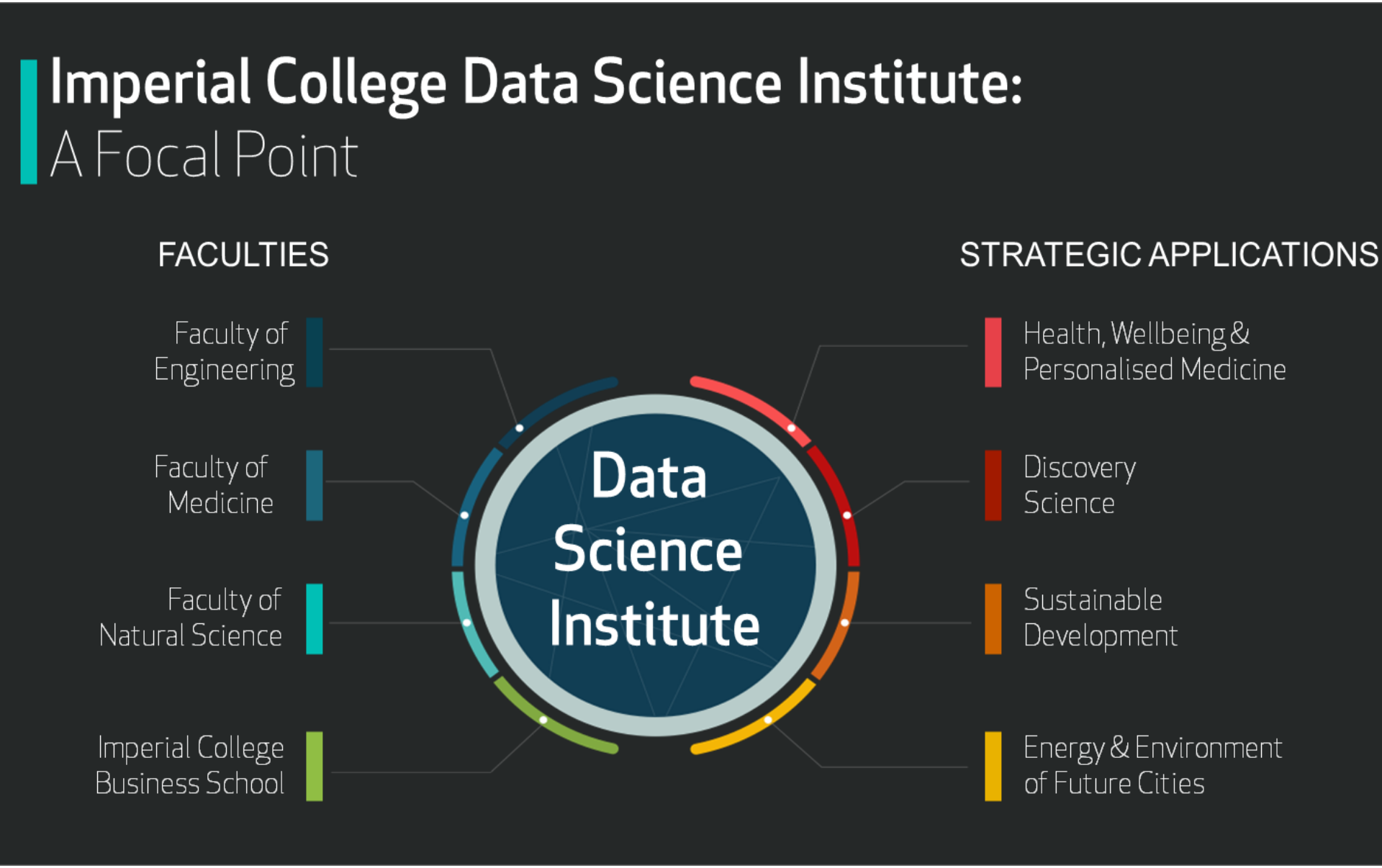
Multi-scale analysis and data integration is enabled







Imperial College Data Science Institute:
A Focal Point



Conclusion

- Datafication drives the era of data centric science
- Data centric science has the 4I characters : Integrative, interactive, intelligent and inter-disciplinary
- Efforts are being made on the development of the 4I technology
- 4I of data centric science are not only technical issues but impact to the research organisation structure
- Data Science Institute of IC aims to develop 4I to explore Big Data for Better Science