Current Status and Future Trends in Data Centric Science

Prof. Yike Guo

Director, Data Science Institute
Imperial College London
Science Today Is Data Centric

Theory suggests hypotheses that are verified through Experiment

Hypotheses are discovered in Data and drive Theory

Theory is developed and explored through Computation

Computations inform the design of Experiments

Computations generate Data

Experiments generate Data

Experiment

Computation
Datafication Turns Sciences into Data Science

**Medical**
- Systems oncology
- Realtime metabolic profiling
- Cardiovascular science

**Natural Sciences**
- Imaging
- Particle physics
- Infection/epidemiology
- System biology
- Neuroscience
- Complex systems & network data analytics
- High-throughput screening

**Engineering**
- Social media & new data business
- Algorithmic trading
- Digital cities & urban life
- Public health management

**Business**
Digitalisation vs Datafication

**Digitalisation**: is a process that has been active within society since the late 1950s, with the birth of the semiconductor 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.

Ericsson White Paper: Impact of Datafication on the Strategic Level
Example: Dataficationg 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**
The most critical choice for a scientist is what problems to work on. – Judea Pearl (Turing Award 2011)

Technologies for Data Centric Science

- Modelling and Simulation
- Data Assimilation
- Big data Management
- Experimental Design
- Analysis & Machine Learning
- Computation
- Experiment
- Theory
- Hypotheses are discovered in Data and drive Theory
- Computation generates Data
- Experiments generate Data
- Theory suggests hypotheses that are verified through Experiment
- Theory is developed and explored through Computation
- Computations inform the design of Experiments
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
- ......
Biology is now a data science

<table>
<thead>
<tr>
<th>Biology</th>
<th>Biotechnology</th>
<th>Bioinformatics</th>
<th>Biologist</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA</td>
<td>Genomics</td>
<td>Data storage</td>
<td>Experiments</td>
</tr>
<tr>
<td>RNA</td>
<td>Transcriptomics</td>
<td>Data handling</td>
<td>Results</td>
</tr>
<tr>
<td>protein</td>
<td>Proteomics</td>
<td>Data preprocessing</td>
<td>Knowledge</td>
</tr>
<tr>
<td>metabolite</td>
<td>Metabolomics</td>
<td>Data analysis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Integrative biology or Systems biology</td>
<td>Data integration</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data interpretation</td>
<td></td>
</tr>
<tr>
<td>cell</td>
<td></td>
<td>Informatics</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>ICT infrastructure</td>
<td></td>
</tr>
</tbody>
</table>
Medicine is now a data science.
Translational Research: Computing Correlations

Human Data Size

- DNA Variants
  - ~3.3 billion BPs [29]
  - >62 million SNPs [30]
- mRNA Expression
  - ~18,000 non-coding RNAs [28]
- Protein Expression
  - ~20,000 protein coding genes [28]
- Metabolic Profile
  - >2500 metabolites [32]
- Environment
  - Diet [32]:
    - >1200 drugs
    - >3500 food components
    - many other factors (e.g., geo-location [33], early-life socio-economic position [34])
- Phenotype
  - Any observable traits

Combinatorial Explosion

Predictor matrix:  
- n observations
- p variables

Response matrix:  
- n observations
- q variables

Aim: identify which of the p variables in X are significantly associated with the outcome Y.
Data Driven Medicine: Searching associations

- Data-driven engine
  - Public Patients
  - Internal Patients
  - Cell Lines/Assays
  - Tox/PKG Panels

Patient Data Correlation: Comparing each new patient to all existing data & knowledge

- Over 37 billion correlations

Correlation Engine
- Billions of Precomputed Associations

Biomedical Insights/Discoveries
- Vast compute farm pre-computes associations for real-time answers
- Comparing datasets via automated algorithms finds unique connections hidden in the data

- Disease Pathways & Subtypes
- New Biomarkers, Drug Indications
- Adaptive Trial Designs, Patient Enrichment
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
Sophisticated algorithms and small data:
• Representative sampling \((N=k)\)
• High quality data
• Sophisticated multivariate learning algorithms
• Exploring causality relationships
Simple algorithms and big data:
• Taking all the data (N=all)
• Messy data tolerance
• Simple/scalable learning algorithms
• Exploring correlation relationships

Small Algorithm + Big Data: Exploring Global Correlation
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**

- Condion A
- Condition B
- Rest

**Model**

- Univariate activation
- Multivariate activation

- Average activity for each voxel across events within condition
- Spatial pattern of activity over voxels within each event

**Measure**

- Brain images with highlighted areas

**Result**

- Parameter estimate
- Hyperspace projection

Activity-based analyses

**Correlation-based analyses**

- Resting connectivity
- Task-based connectivity
- Full (resting) connectivity

- Temporal correlation between a seed voxel and all other voxels
- Correlation between seed and other voxels within condition
- Correlation of all possible seed voxels with each other

**Big Data**
Deep Learning: From Vision to Cognition

[picture from Simon Thorpe]

[picture from 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

Exponential Growth of Computing

Calculations per Second per $1,000

Source: By: Kamesh and Kumar West
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 adaptations include: sampling strategies, model parameters, visual understanding
• Interaction suggests a “Data Chemistry”!
WIKIHEALTH: AN INTEGRATED PLATFORM FOR WEARABLE SENSOR INFORMATICS
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
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
  - Distribution (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
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

- 3 layers targeting different stages of the data lifecycle
- Data products built using the application development environment

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

FACULTIES
- Faculty of Engineering
- Faculty of Medicine
- Faculty of Natural Science
- Imperial College Business School

STRATEGIC APPLICATIONS
- Health, Wellbeing & Personalised Medicine
- Discovery Science
- Sustainable Development
- Energy & Environment of Future Cities
Conclusion

• Datafication drives the era of data centric science
• Data centric science has the 4I characters: Integrative, interactive, intelligent and interdisciplinary
• 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