Reflections on Data Science, Machine Learning, AI

Raymond Ng
Director, Data Science Institute, UBC
Canada Research Chair on Data Science and Analytics
Professor, Computer Science, UBC

Data Science in the News…

Data Science Word Cloud

What is Data Science?

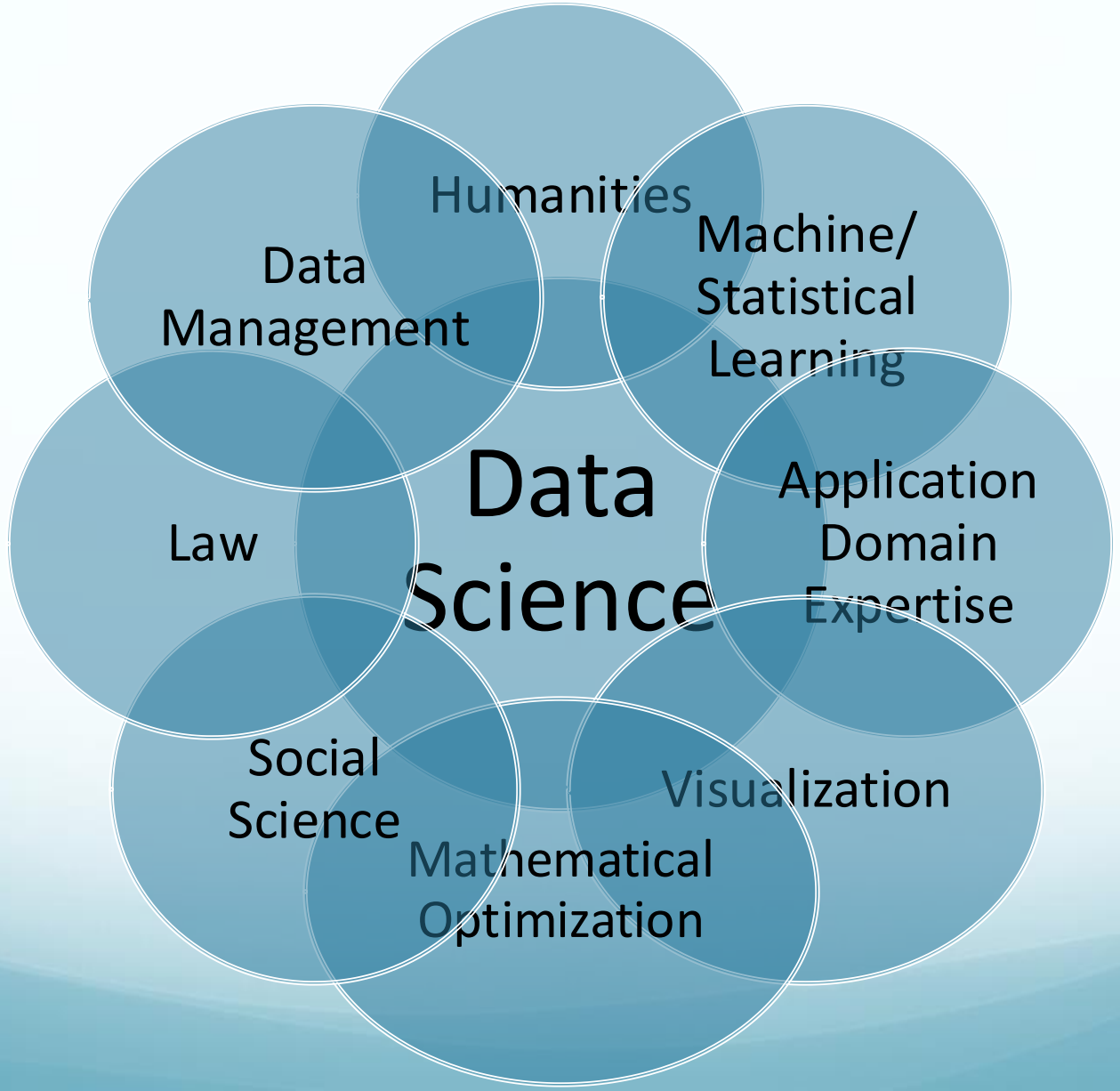
- Refers to methods, processes and tools that allow a user to extract useful information from complex data collections
- An interdisciplinary field:
  - Data mining, machine learning, statistical inference, predictive modeling, databases, visualization, high performance computing, data privacy, security, etc.
- Is critical for organizations and companies to make decisions and drive innovation

Data Science = Big Data??

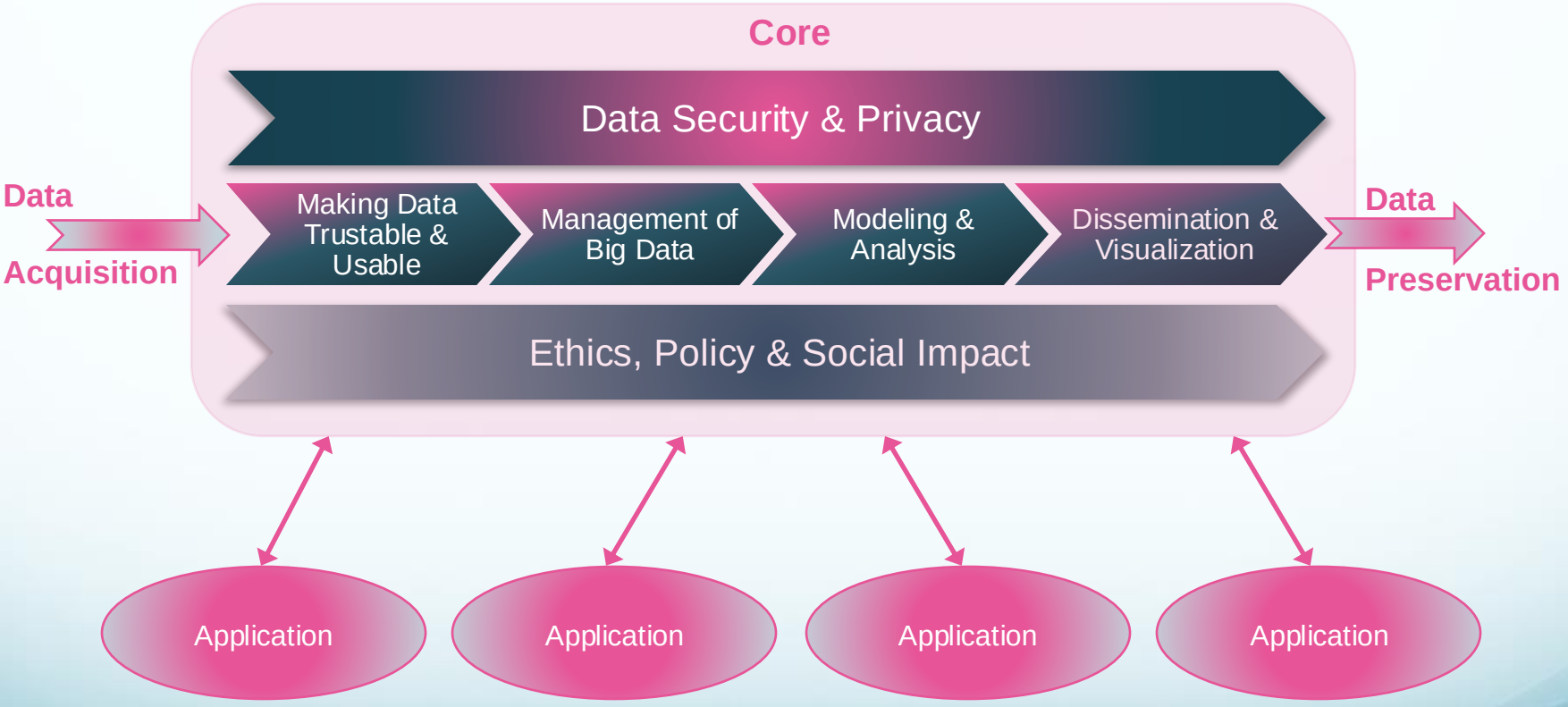
- Not the “same thing”
- If Big data = crude oil
  - Big data is about extracting “crude oil”, transporting it in “mega tankers”, siphoning it through “pipelines”, and storing it in “massive silos”
- Data science is about refining the “crude oil”, i.e., building on top of Big Data

Carlos Samohano
Founder, Data Science London

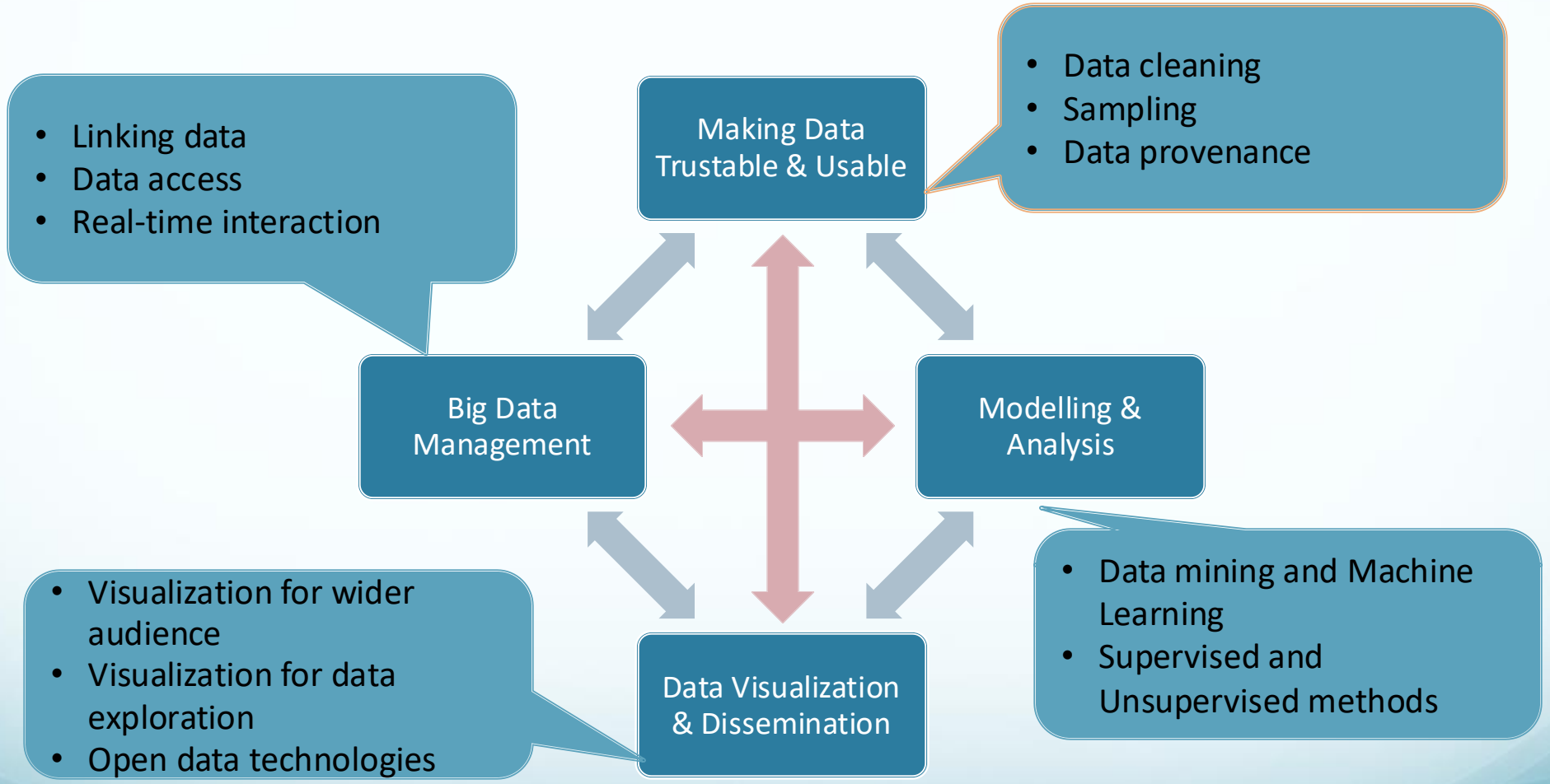
Data Science as a Unifier



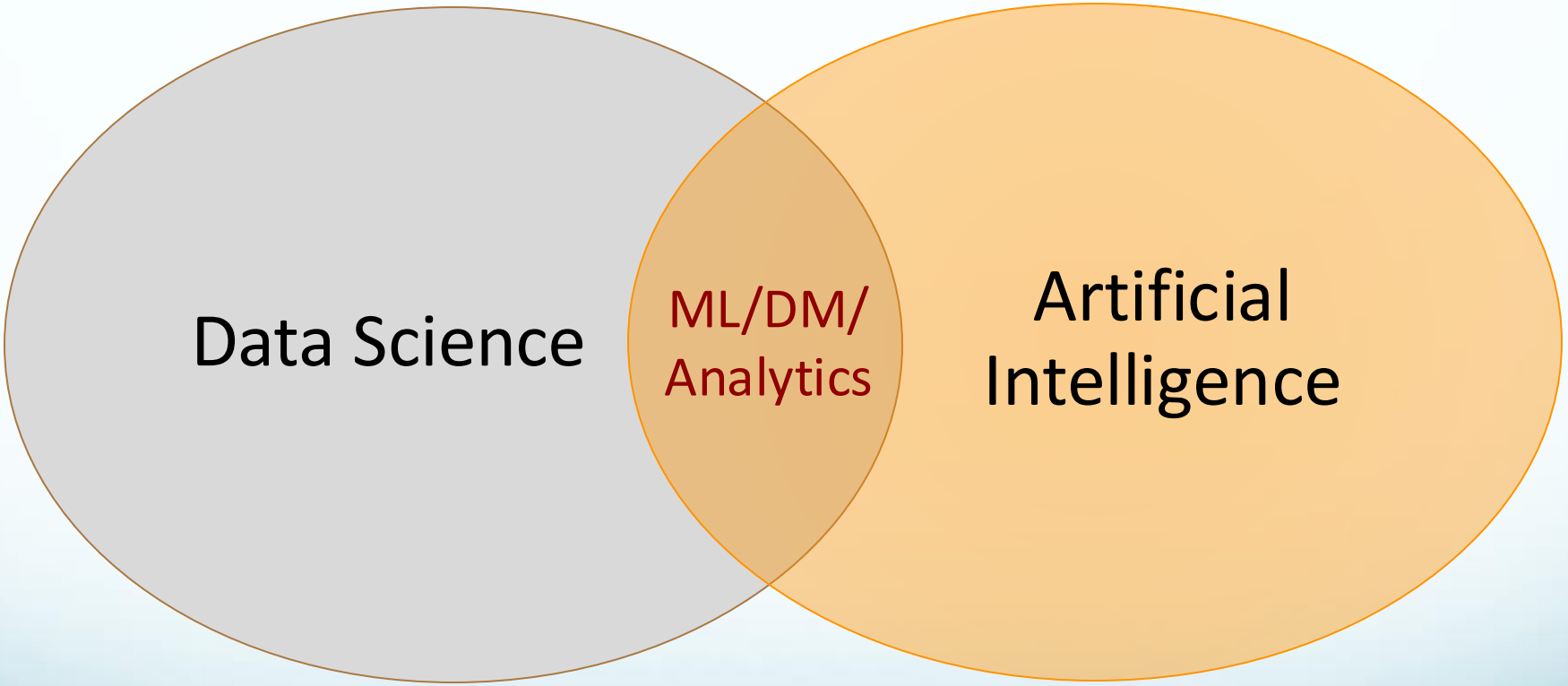
Holistic Approach to Data Science



Core Issues


Making Data Trustable & Usable
• Data cleaning
• Sampling
• Data provenance
• Linking data
• Data access
• Real-time interaction
Big Data Management
Modelling & Analysis
• Visualization for wider audience
• Visualization for data exploration
• Open data technologies
Data Visualization & Dissemination
• Data mining and Machine Learning
• Supervised and Unsupervised methods

Data Science vs Machine Learning vs AI



"Data science systems produce **insights**;
Machine learning systems generate **predictions;**
AI systems make **decisions** and take **actions"**

Data Science Meets Health Science:
Info Tech Meets Bio Tech

Advances in:

- CPU/storage capacity
- data capture
- searching and database mgmt
- Cloud, Hadoop
- algorithms
- statistics
- mathematics, etc.

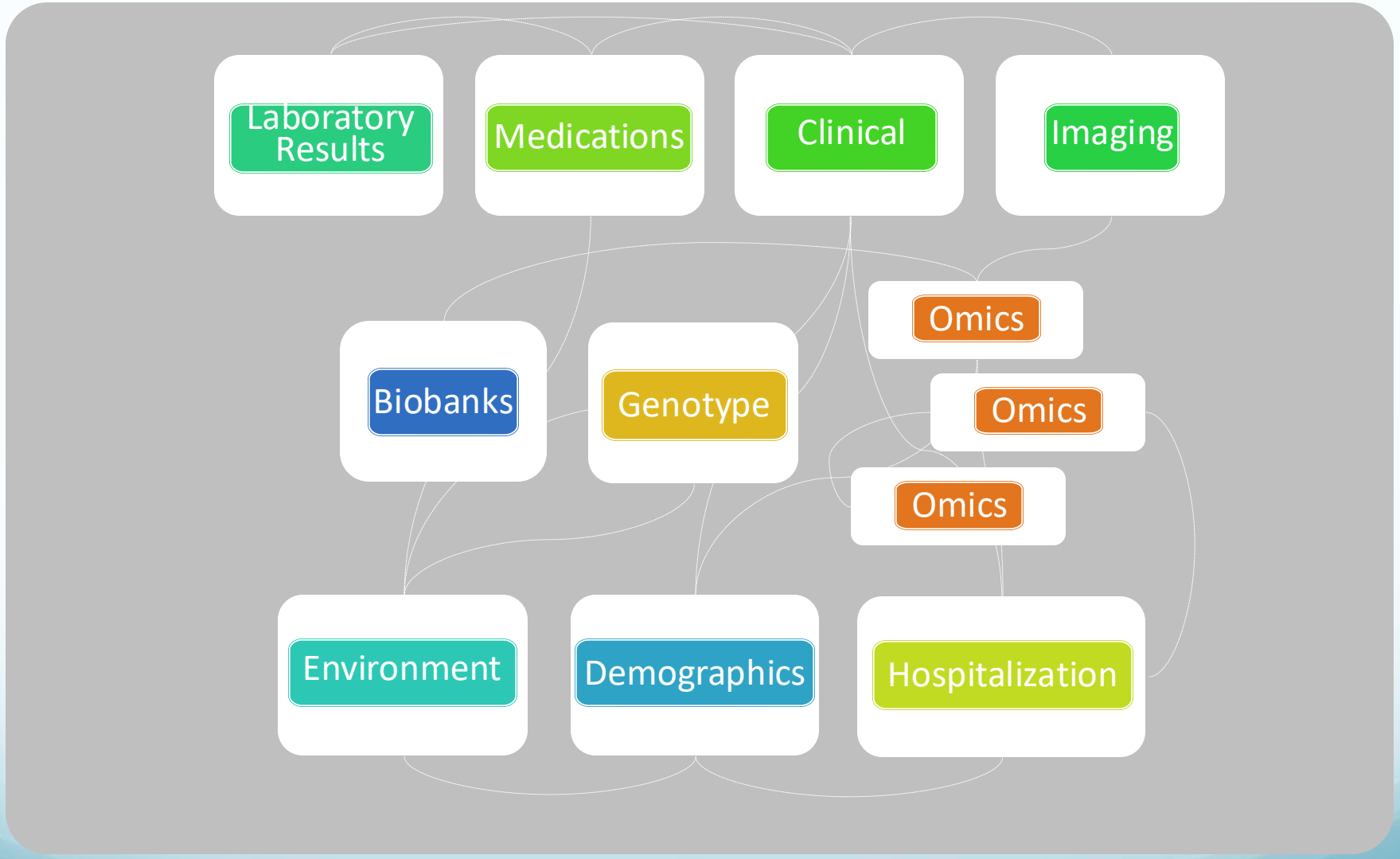
Very Exciting Times!

How the 4V's Challenge Health Care and Health Research?

#1 Volume:

- Rapid advances in biotech, e.g., sequencing of all flavours, mass spectrometry, point-of-care diagnostics
- Electronic Health Records: governments (e.g., MoH) making longitudinal data (e.g., months, years) on whole population available
- Data sharing across jurisdictions, collaborators also increase the dataset size
- Many new ways of collecting data, e.g., health apps

#2 Variety



Potential of modelling across these modalities is HUGE

#3 Velocity

- The speed that data are collected or generated, e.g., health monitoring apps
- E.g., Text data, e.g., whatsapp, clinical trials
  - Patients describing their own sentiments – a "window" into their psychological states, their cognitive states, etc.
  - *Longitudinal* text – capturing changes over time -can be the basis of a powerful predictive model
  - Building a joint model using *multi-omics and text* to monitor chronic disease management (e.g., even covid recovery) is not far away

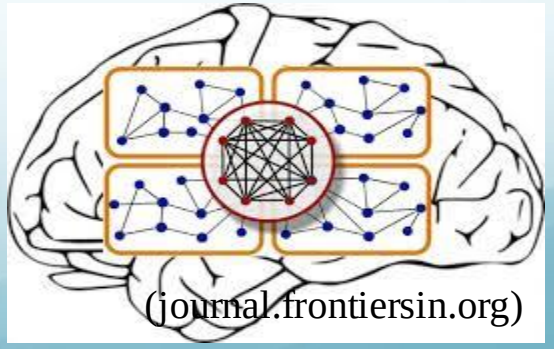
#4 Veracity

- The truthfulness/reliability/quality of data and data sources, e.g.,
  - public repositories vs MoH data
  - Missing values or errors in EHR
  - Self-declared survey data, sentiment data
  - Accuracies of devices

Neural Network Learning for Biomedical Research

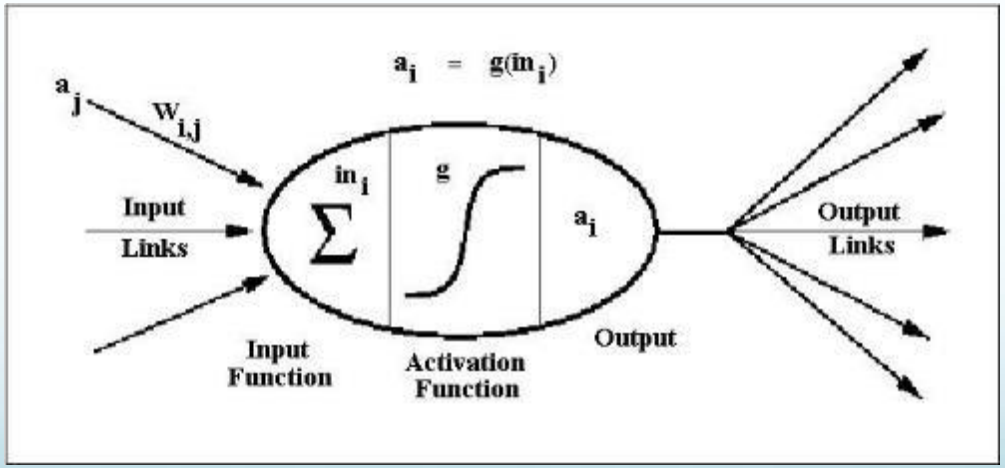
Neural Networks: Simulating Human Brains?

- To build a predictive model, why not do what a human brain can do?
- In particular, neuroscience reveals the existence of massively connected networks of neurons
- In 1943, McCulloch and Pitts proposed a simple mathematical model for neurons, now known as neural networks



Basic Structure of a Network Unit

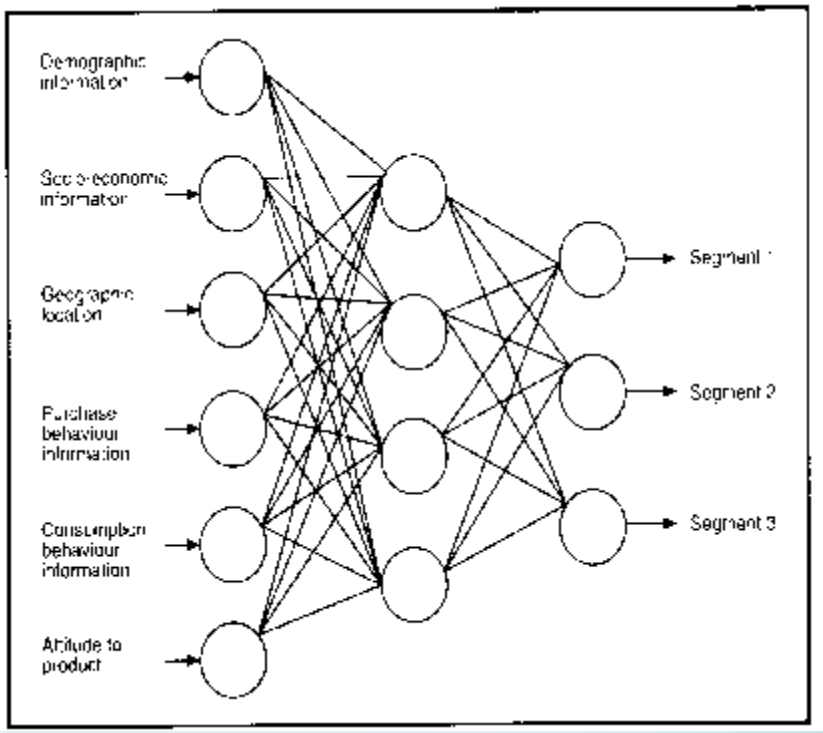
- A Neural network composed on nodes/units connected by links
- A link from unit i to unit j propagates an activation a_i to possibly trigger an activation a_j



(en.wikibooks.org)

Feed-forward and Recurrent Networks

- This network determines to which segment a customer belongs
- This is an example of a feed-forward network
  - No loops, forming an acyclic graph
- A recurrent network has loops and may take time to reach a steady state



Hidden Layers and Deep Networks

- Our example has a hidden layer of units
  - allow us to model quantities not directly measured
  - *features* are being “learned”
- Common to have units arranged in layers, i.e., inputs from the preceding layer and outputs to the next layer
- ”*Deep*" in "deep learning" refers to the number of layers through which the data are transformed
  - called the *credit assignment path* (CAP) depth
  - for a feed-forward network, the depth is 1 + the number of hidden layers
  - ”deep” means CAP > 2
  - extra layers are intended to ”extract/engineer” better features

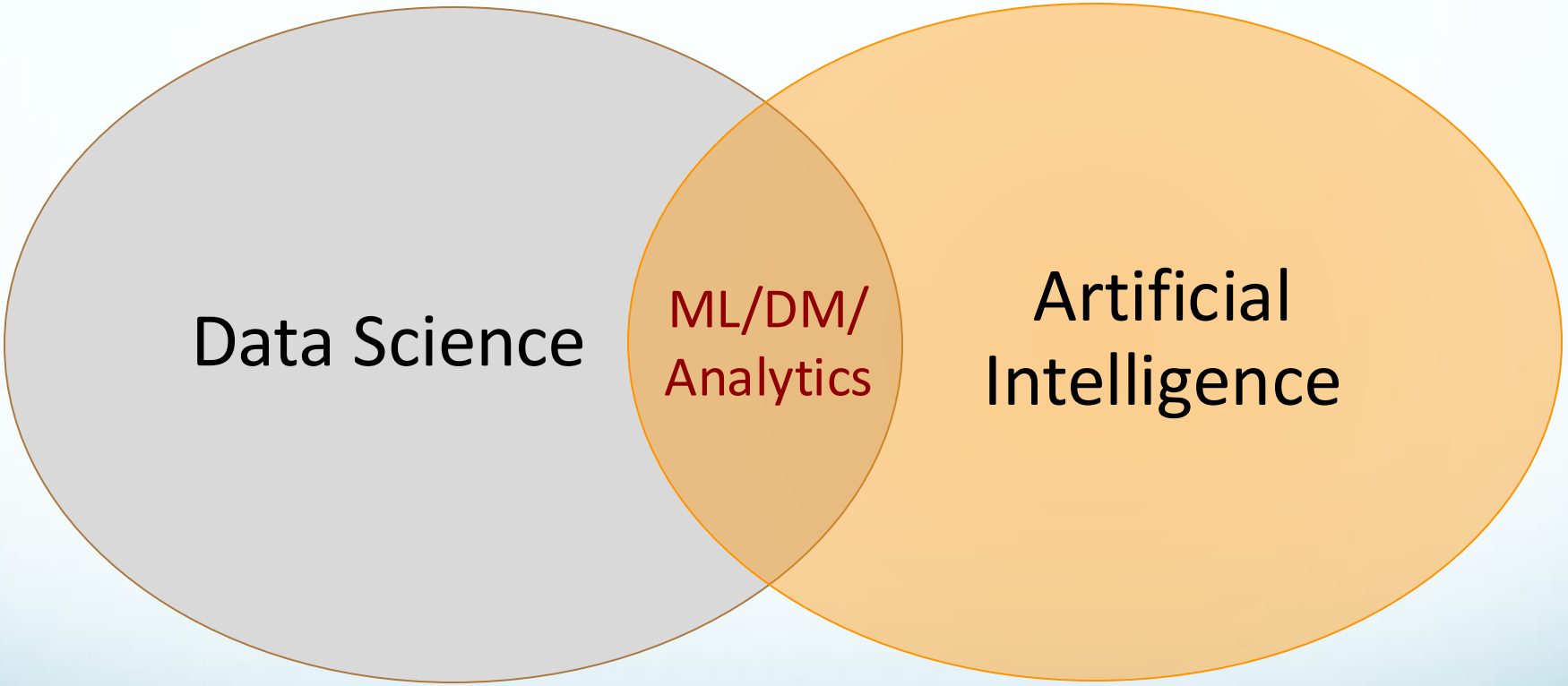
Some Deep Learning Applications

- **Image recognition**: computational radiology
  - E.g., Imperial College to provide image-based assessments of traumatic brain injuries
- **Drug design and toxicology**
  - E.g., AtomNet for virtual drug screening and target discovery
- **Modeling of EHR data**
  - E.g., Children's Hospital Los Angeles to improve pediatric intensive care diagnosis
- **Bioinformatics**
  - E.g., regulatory genomics (Montgomery et al., Nature 2010)

But Buyers Beware…

- Neural networks have many parameters to “estimate”
  - Deeper networks have even more parameters to fit
  - Paramount to have a lot of high quality (training) data
  - Much easier to overfit than many other methods
- The structure of a network is not known a priori
  - Yet results typically sensitive to the structure
- Interpretation is hard
  - A key strength of a deep network is feature engineering
  - But the features may not have natural ”real world” meanings

Summary: Machine Learning, AI Opportunities and Boundaries



Take home: Remember that high quality data is the pre-requisite of ML and AI methods

Thank You!

rng@cs.ubc.ca

Summary

- The new “crude oil”, Big Data present great opportunities
- Data Science “refines crude oil” to bring great value:
  - More jobs: lots of work to do to exploit these data assets
  - New insights leading to smarter decisions/policies
  - When applied to health, better care and more advanced research
  - This program covers all these aspects and more