CS4225/CS5425 Big Data Systems for Data Science

Introduction to Data Science

Bryan Hooi
School of Computing
National University of Singapore
bhooi@comp.nus.edu.sg

School of Computing

Learning Objectives

- What is (big) data science?
- Why (big) data science?
- *Infrastructure* for big data

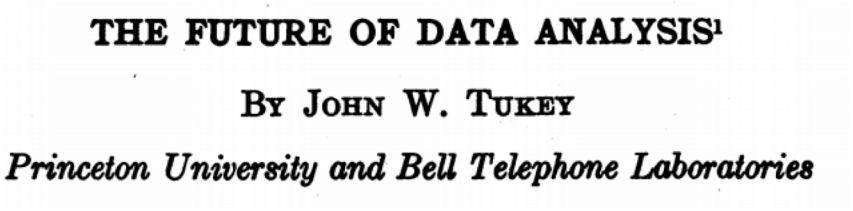
What is Data Science?

Optional

- **Wiki definition**:

"Data science is an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured, which is a continuation of some of the data analysis fields such as statistics, machine learning, data mining, and predictive analytics."

- **Historical view**: in 1962, statistician John Tukey described a field called "data analysis":



THE FUTURE OF DATA ANALYSIS[1]

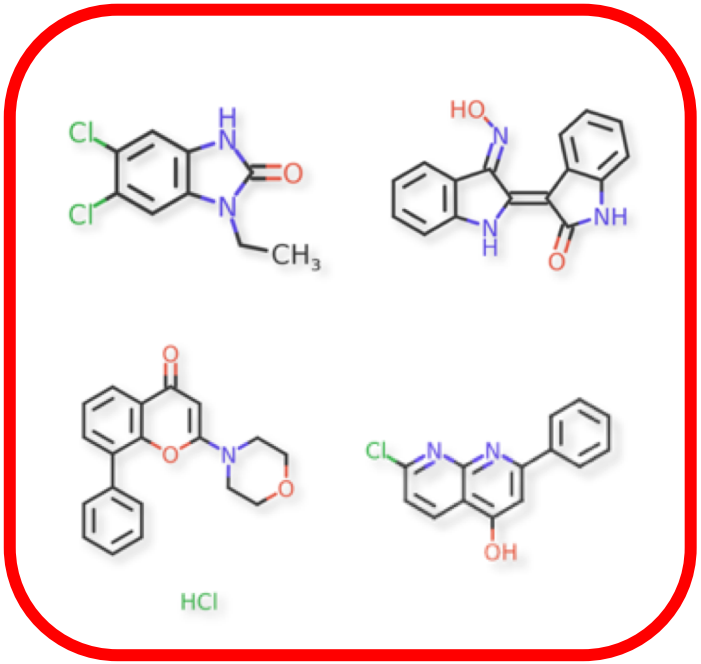
By John W. Tukey

Princeton University and Bell Telephone Laboratories

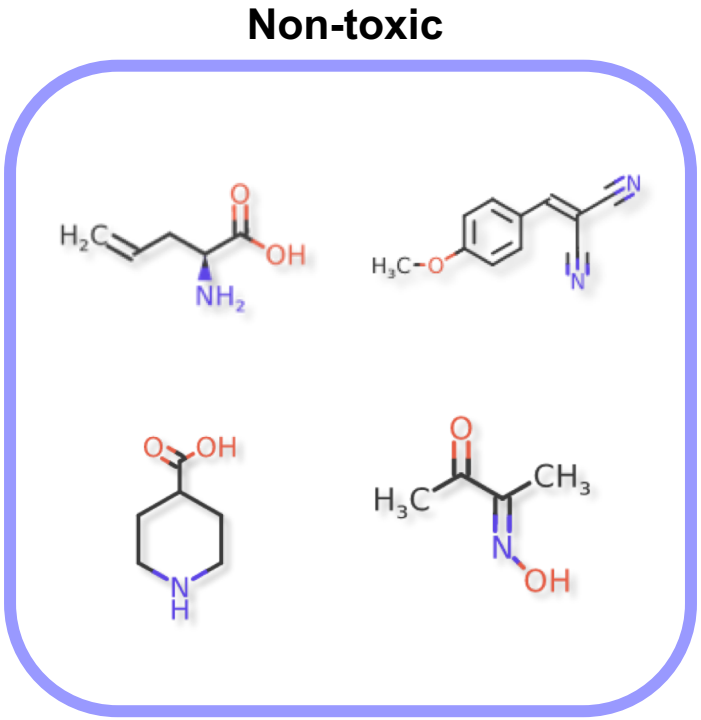
"For a long time I have thought I was a statistician… all in all, I have come to feel that my central interest is in *data analysis*, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate…"

Q: What rule characterizes toxic molecules?

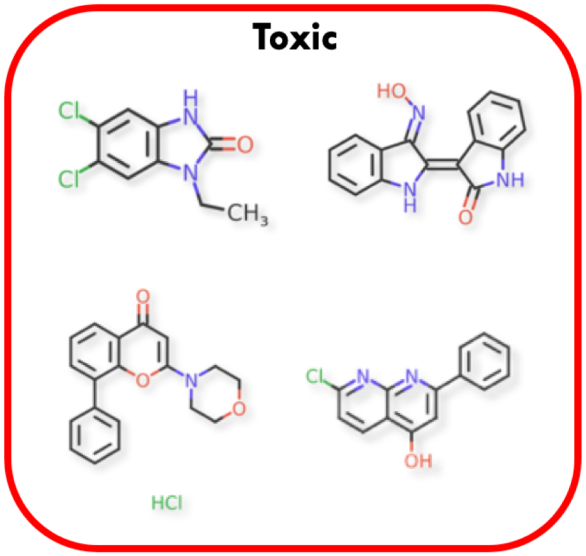
Toxic

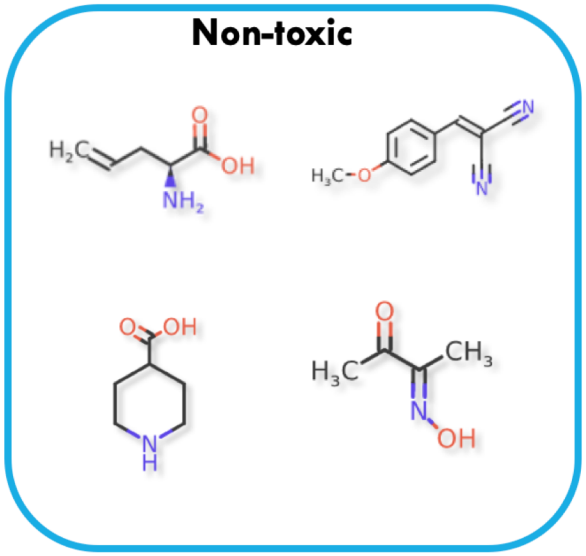


Non-toxic



Q: What rule characterizes toxic molecules?



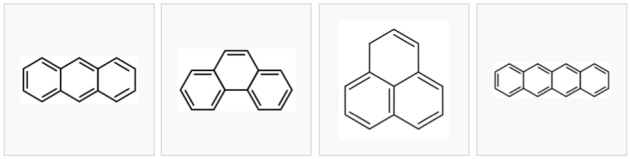



```
if ⬡ > 1:
    toxic
else
    nontoxic
```

Polycyclic aromatic hydrocarbon

From Wikipedia, the free encyclopedia

Principal PAH Compounds

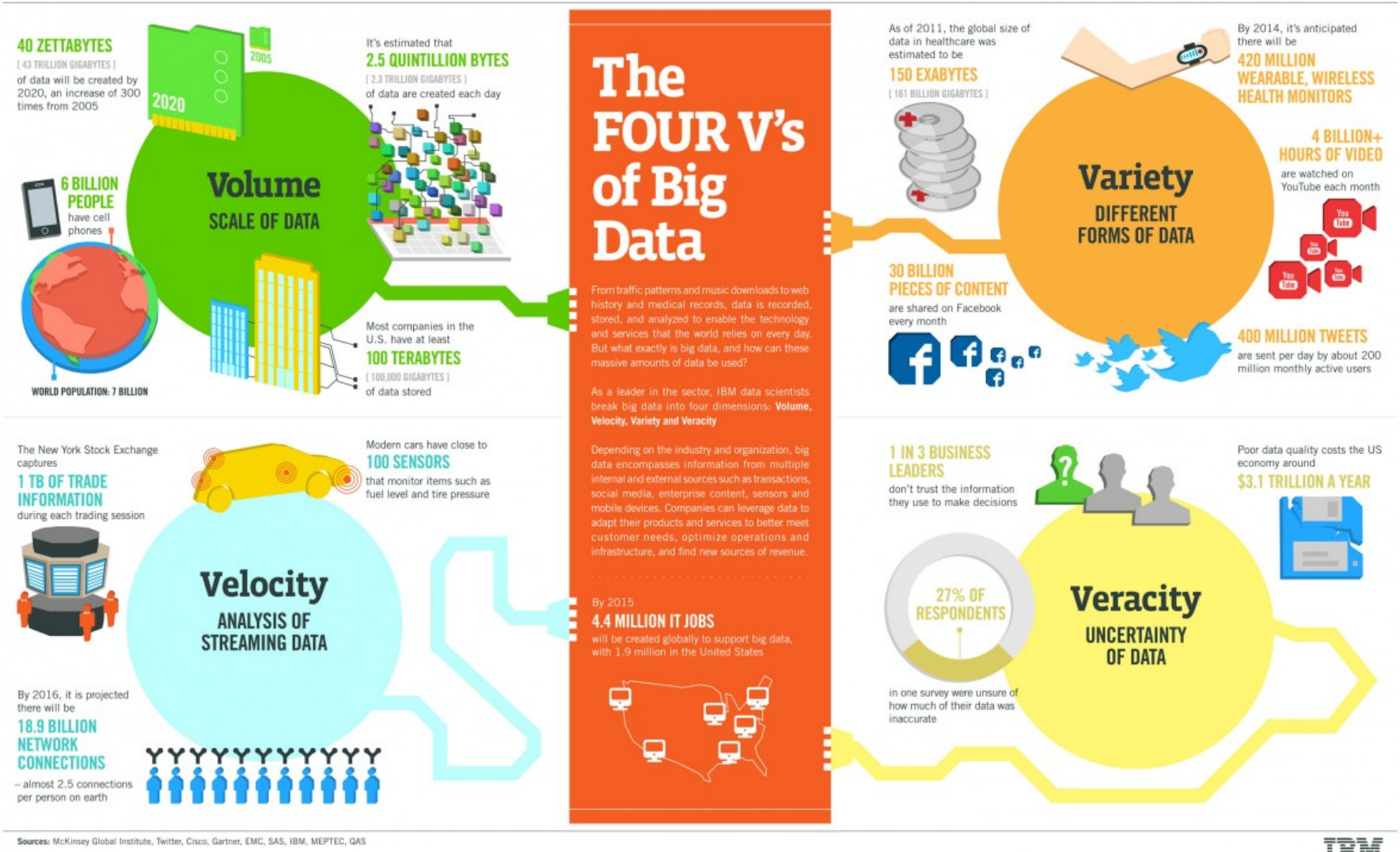


Cancer [edit]

PAHs have been linked to skin, lung, bladder, liver, and stomach cancers in well-established animal model studies.[72] human carcinogens are identified in the section "Regulation and Oversight" below.

Data Contains Value and Knowledge

The 4V of Big Data from IBM



Veracity: ~3.3% of image data in ImageNet (a popular image dataset) are mislabelled

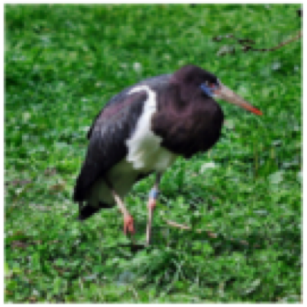
Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks

Curtis G. Northcutt* ChipBrain, MIT
Anish Athalye MIT
Jonas Mueller Amazon


given: cat
corrected: frog


given: lobster
corrected: crab

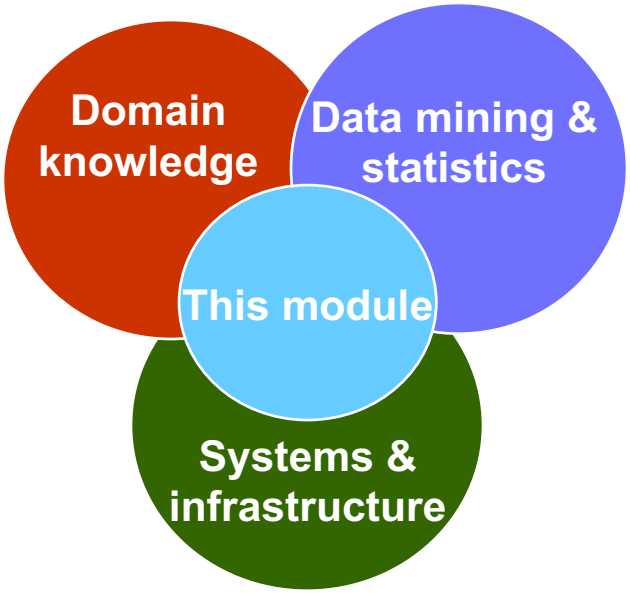

given: ewer
corrected: teapot


given: white stork
corrected: black stork

[1] Northcutt, Curtis G., Anish Athalye, and Jonas Mueller. "Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks." NeurIPS 2021

Big Data Systems and Techniques

- Massive data science == Big data
- Techniques
  - Data mining and statistics
  - Systems and infrastructure
- Domains
  - Web
  - Social network
  - …



- This course talks about the **interplay** among techniques and domains.
  - We will learn how to support large-scale data mining with scalable systems and infrastructure.
  - We will learn how the domain impacts the design of systems and infrastructure.

What will we learn?

- **We will learn to process/mine different types of data:**
  - Data is high dimensional
  - Data is a graph
  - Data is infinite/never-ending
- **We will learn to use different models of computation:**
  - MapReduce/Spark
  - Streams and online algorithms
  - Large graph processing engines

What will we learn?

- **We will learn to solve real-world problems:**
  - Recommender systems
  - Spam detection
- **We will learn various emerging systems:**
  - MapReduce/Hadoop/Spark
  - NoSQL systems
  - Graph engines
  - Stream processing systems

This course is introductory, more on breadth, rather than depth.

Google™ Processes 20 PB a day (2008)
Crawls 20B web pages a day (2012)
Search index is 100+ PB (5/2014)
Bigtable serves 2+ EB, 600M QPS (5/2014)

INTERNET ARCHIVE 400B pages, 10+ PB (2/2014)

YAHOO! Hadoop: 365 PB, 330K nodes (6/2014)

JPMorganChase 150 PB on 50k+ servers running 15k apps (6/2011)

ebay Hadoop: 10K nodes, 150K cores, 150 PB (4/2014)

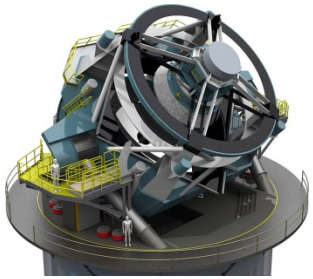
300 PB data in Hive + 600 TB/day (4/2014)

facebook

amazon web services™ S3: 2T objects, 1.1M request/second (4/2013)

LHC: ~15 PB a year

CERN



LSST: 6-10 PB a year (~2020)

SKA: 0.3 – 1.5 EB per year (~2020)

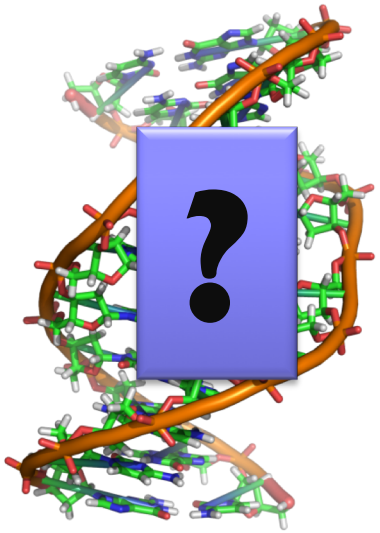
How much data?

Why big data?

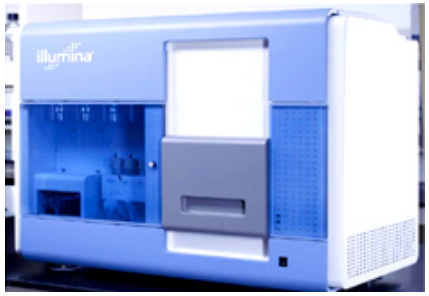
Science
Engineering
Commerce

Source: Wikipedia (Everest)

Genome is Big Data



Subject genome



Sequencer

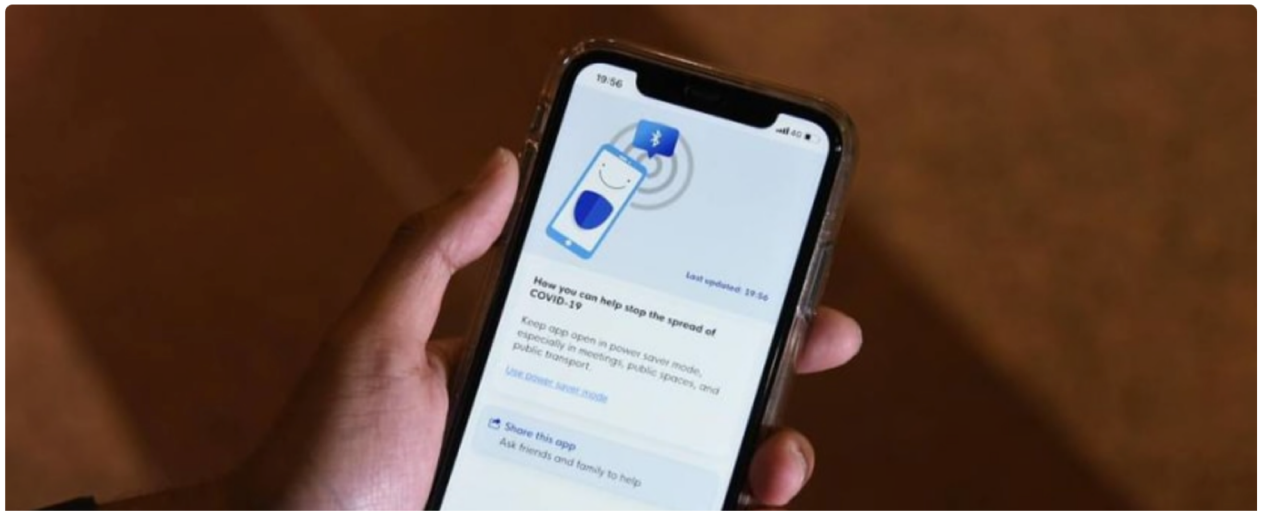
GATGCTTACTATGCGGGCCCC
CGGTCTAATGCTTACTATGC
GCTTACTATGCGGGCCCCTT
AATGCTTACTATGCGGGCCCCTT
TAATGCTTACTATGC
AATGCTTAGCTATGCGGGC
AATGCTTACTATGCGGGCCCCTT
AATGCTTACTATGCGGGCCCCTT
CGGTCTAGATGCTTACTATGC
AATGCTTACTATGCGGGCCCCTT
CGGTCTAATGCTTAGCTATGC
ATGCTTACTATGCGGGCCCCTT

Reads

Human genome: 3 gbp
A few billion short reads
(~100 GB compressed data)

Precision medicine
Health
Insurance
…

Bill restricting police use of TraceTogether data introduced in Parliament, with tougher penalties for misuse



Aqil Haziq Mahmud
@AqilHaziqCNA

01 Feb 2021 01:45PM
(Updated: 01 Feb 2021 02:37PM)

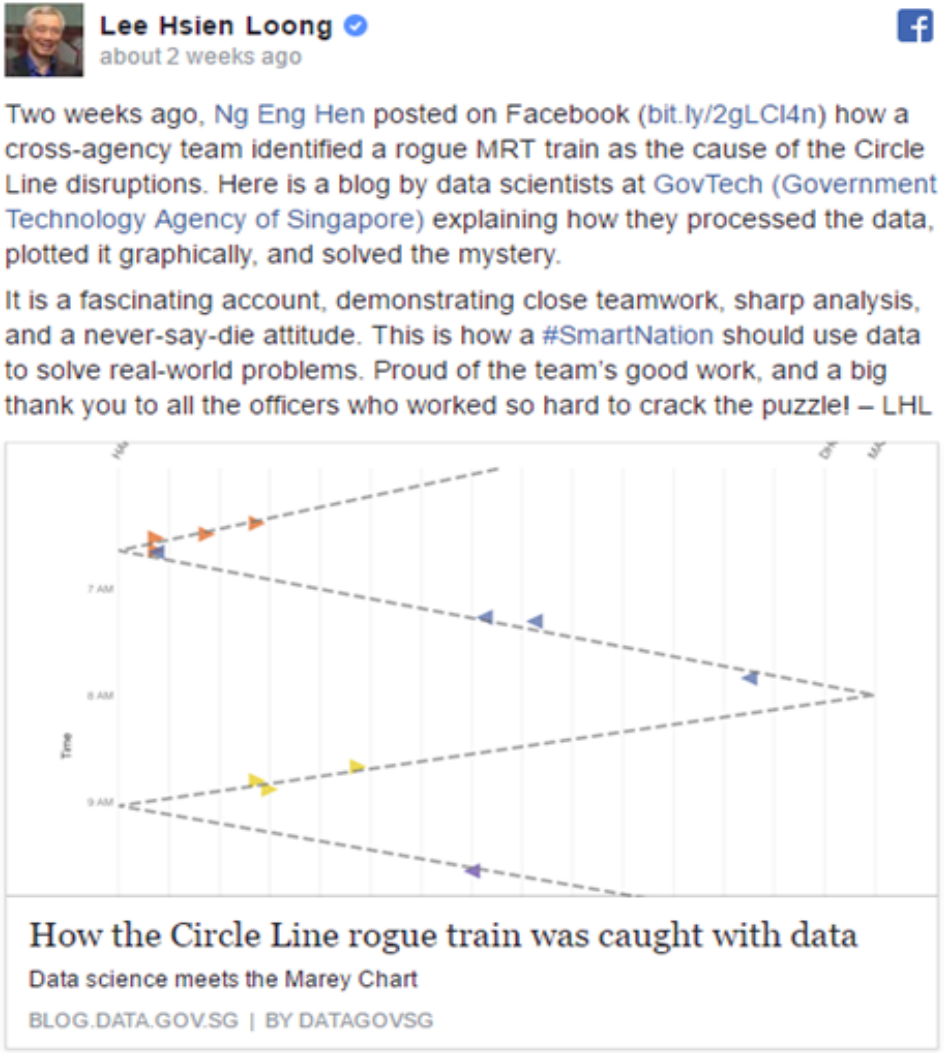
Engineering

The unreasonable effectiveness of data

Count and normalize!

Source: Wikipedia (Three Gorges Dam)

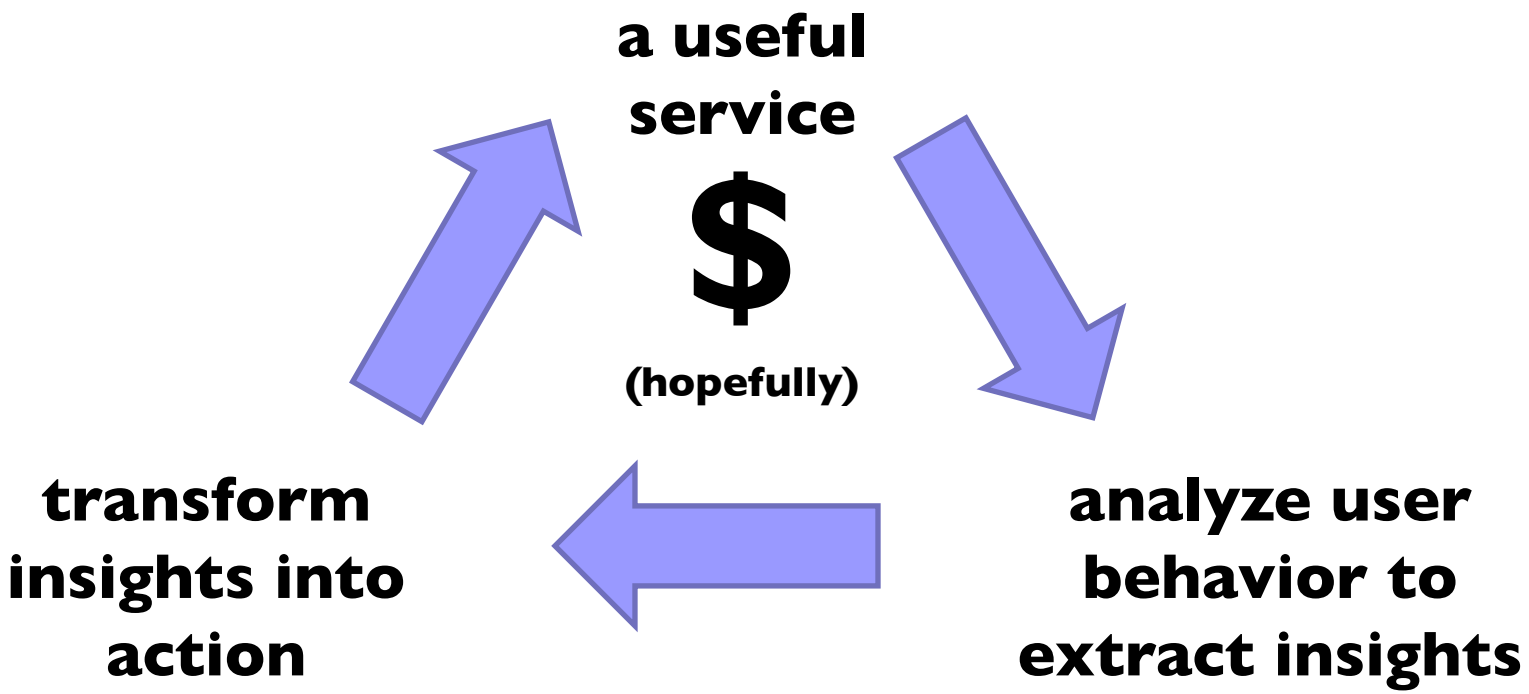
How the Circle Line rogue train was caught with data?



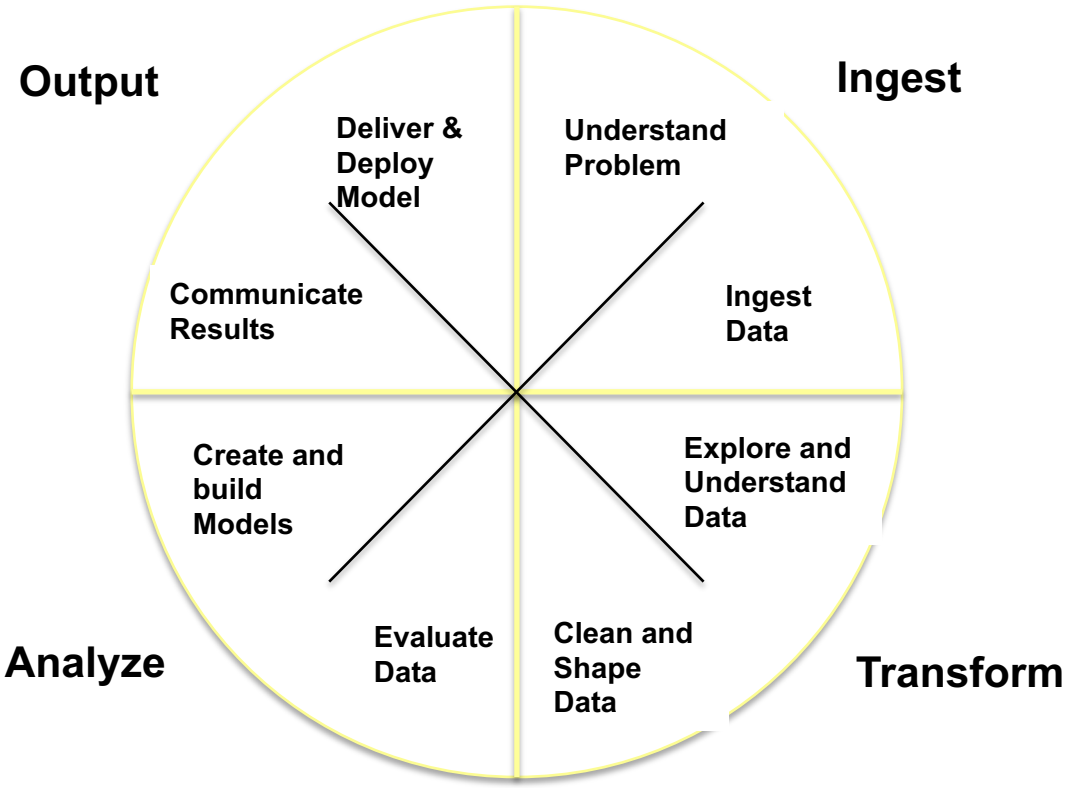
https://blog.data.gov.sg/how-we-caught-the-circle-line-rogue-train-with-data-79405c86ab6a#.5sehgyfsq


Know thy customers
Data → Insights → Competitive advantages
Commerce
EPSON
PEARL BAR
Source: Wikiedia (Shinjuku, Tokyo)

Virtuous Product Cycle



Data Lifecycle



Source: Wikipedia (Noctilucent cloud)

Interlude: Cloud Computing

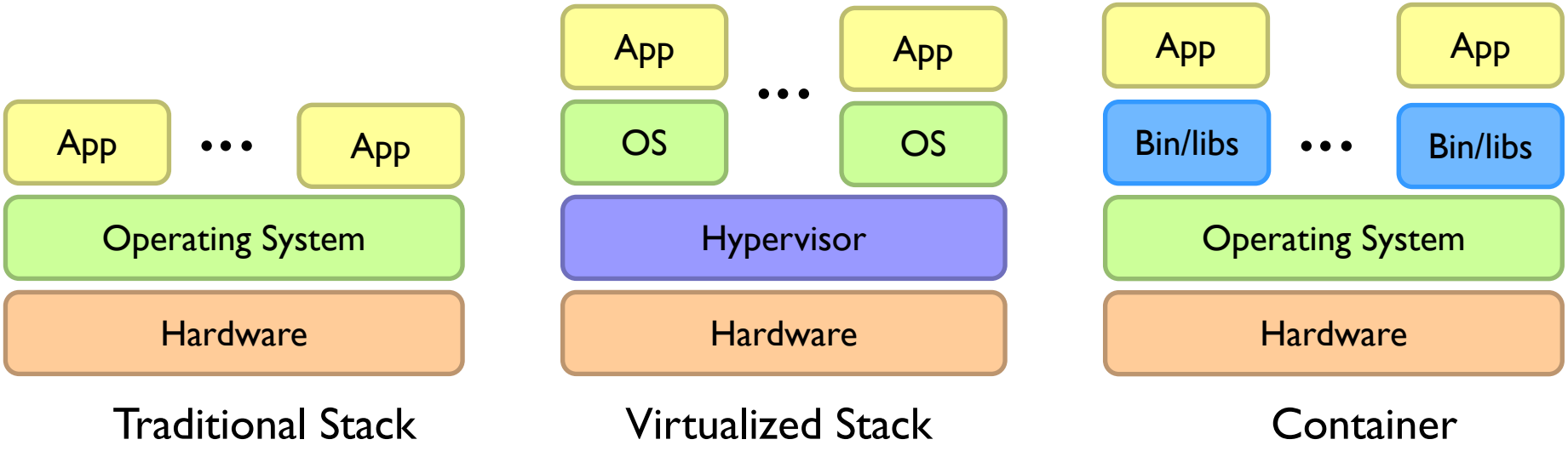
Source: Wikipedia (Clouds)

Utility Computing

- What?
  - Computing resources as a metered service ("pay as you go")
  - Ability to dynamically provision virtual machines
- Why?
  - Cost: capital vs. operating expenses
  - Scalability: "infinite" capacity
  - Elasticity: scale up or down on demand
- Does it make sense?
  - Benefits to cloud users
  - Business case for cloud providers

Thomas J. Watson (attributed?)

Enabling Technology: Virtualization and Container



Everything as a Service

- Utility computing = Infrastructure as a Service (IaaS)
  - Why buy machines when you can rent cycles?
  - Examples: Amazon's EC2, Rackspace, Google Compute Engine
- Platform as a Service (PaaS)
  - Provides hosting for web applications and takes care of the maintenance, upgrades, …
  - Example: Google App Engine
- Software as a Service (SaaS)
  - Just run it for me!
  - Example: Gmail, Dropbox, Zoom

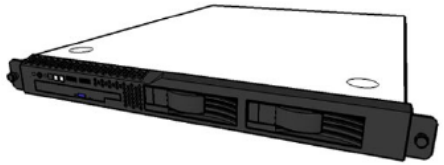
Who cares?

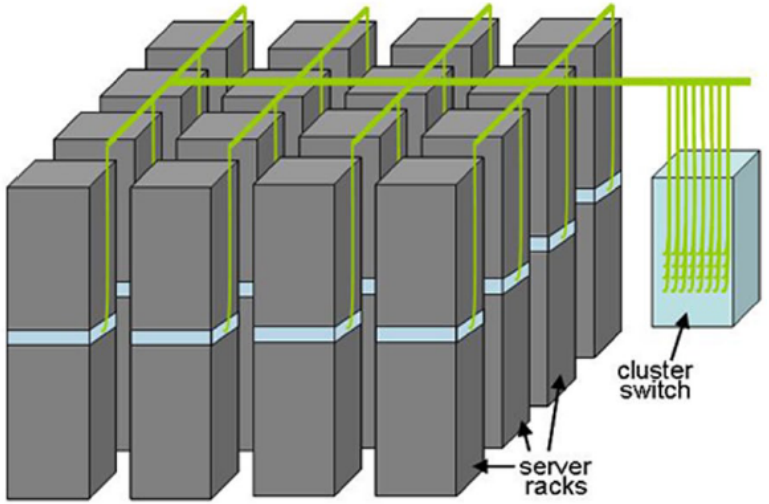
- A source of problems…
  - Cloud-based services *generate* big data
  - Clouds make it easier to start companies that *generate* big data
- As well as a solution…
  - Ability to provision analytics clusters on-demand in the cloud
  - Commoditization and democratization of big data capabilities

The datacenter *is* the computer!

Source: Google

Building Blocks





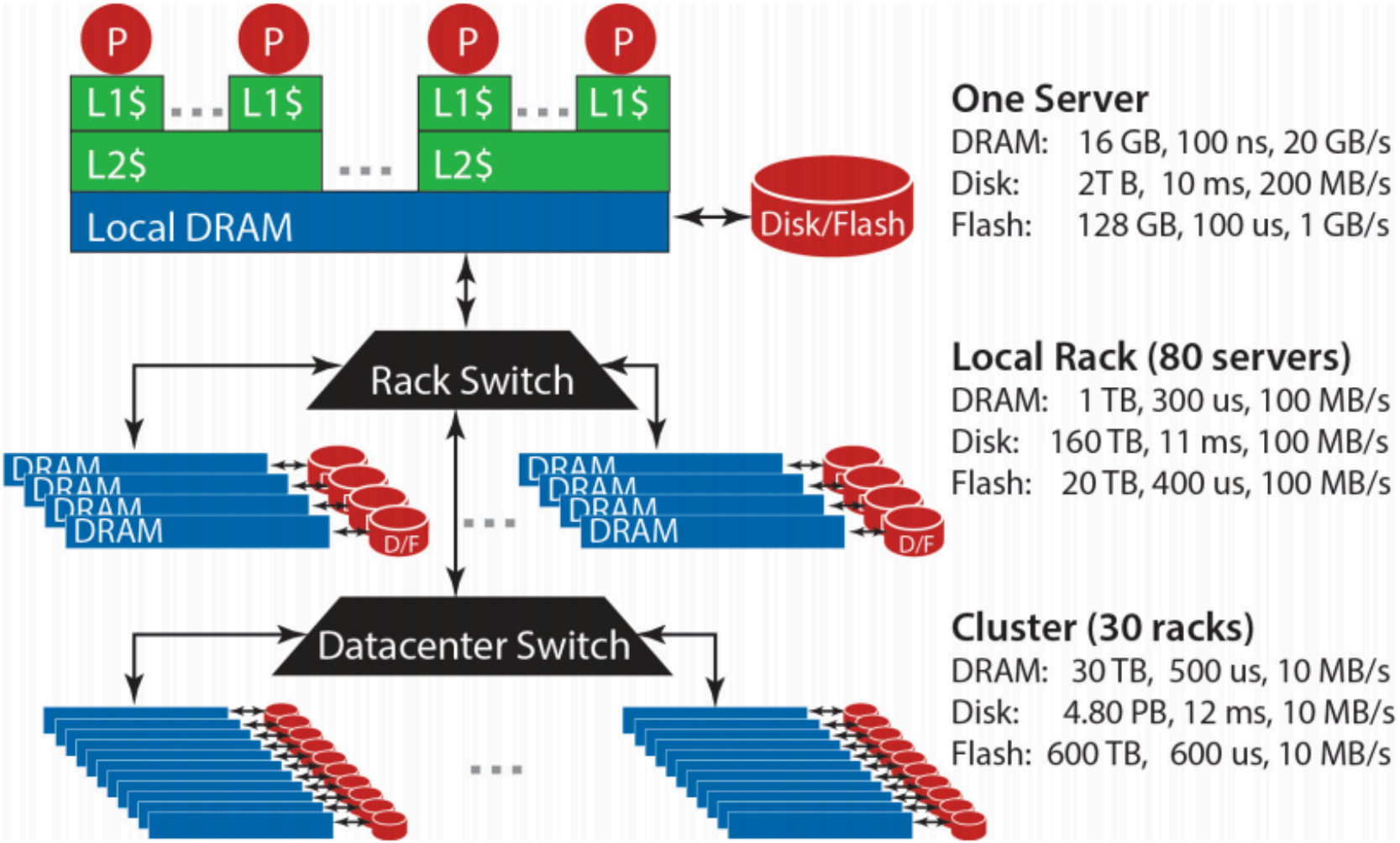
Source: Barroso and Urs Hölzle (2009)

Source: Google

Source: Google

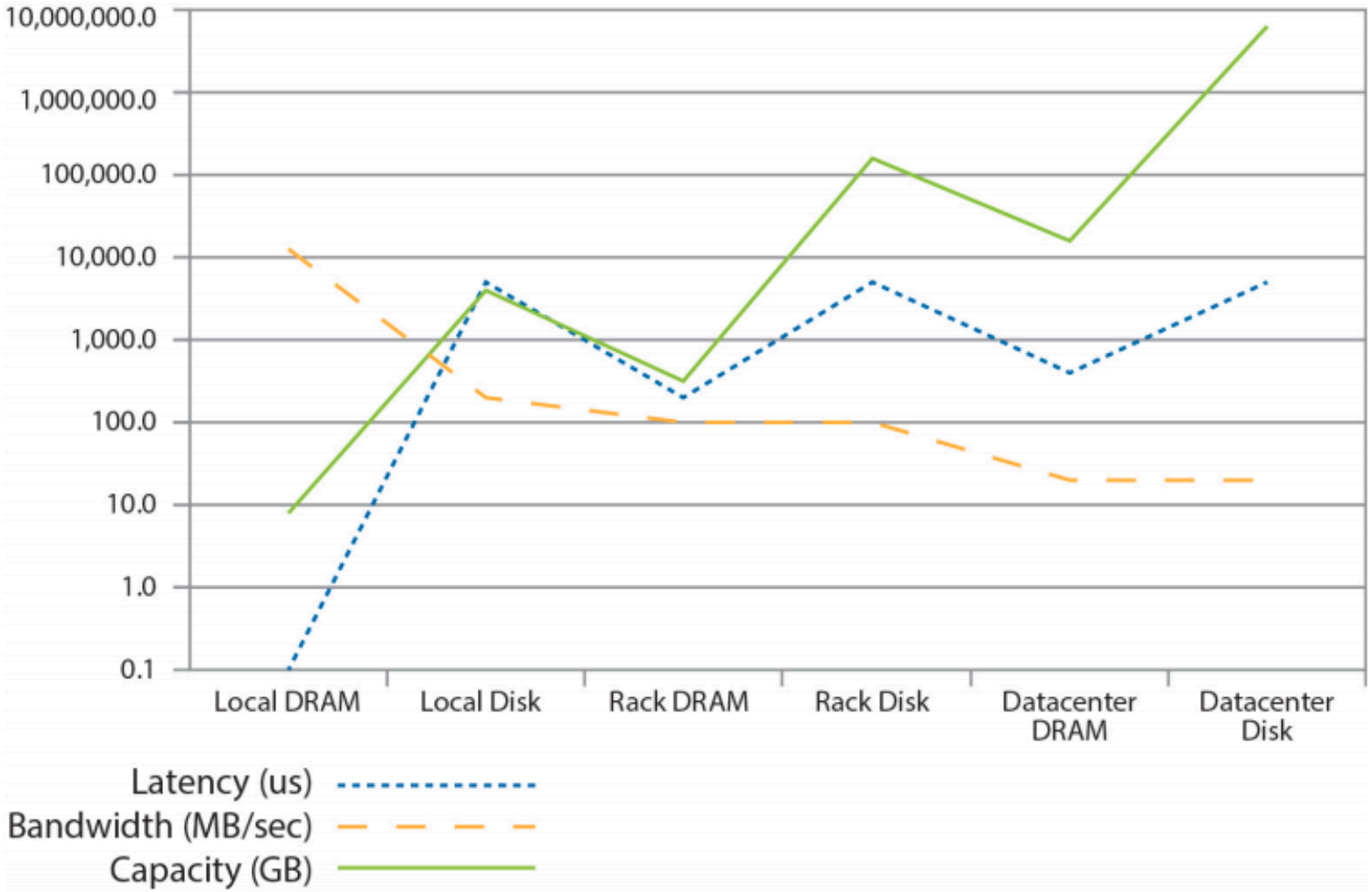
Source: Facebook

Storage Hierarchy



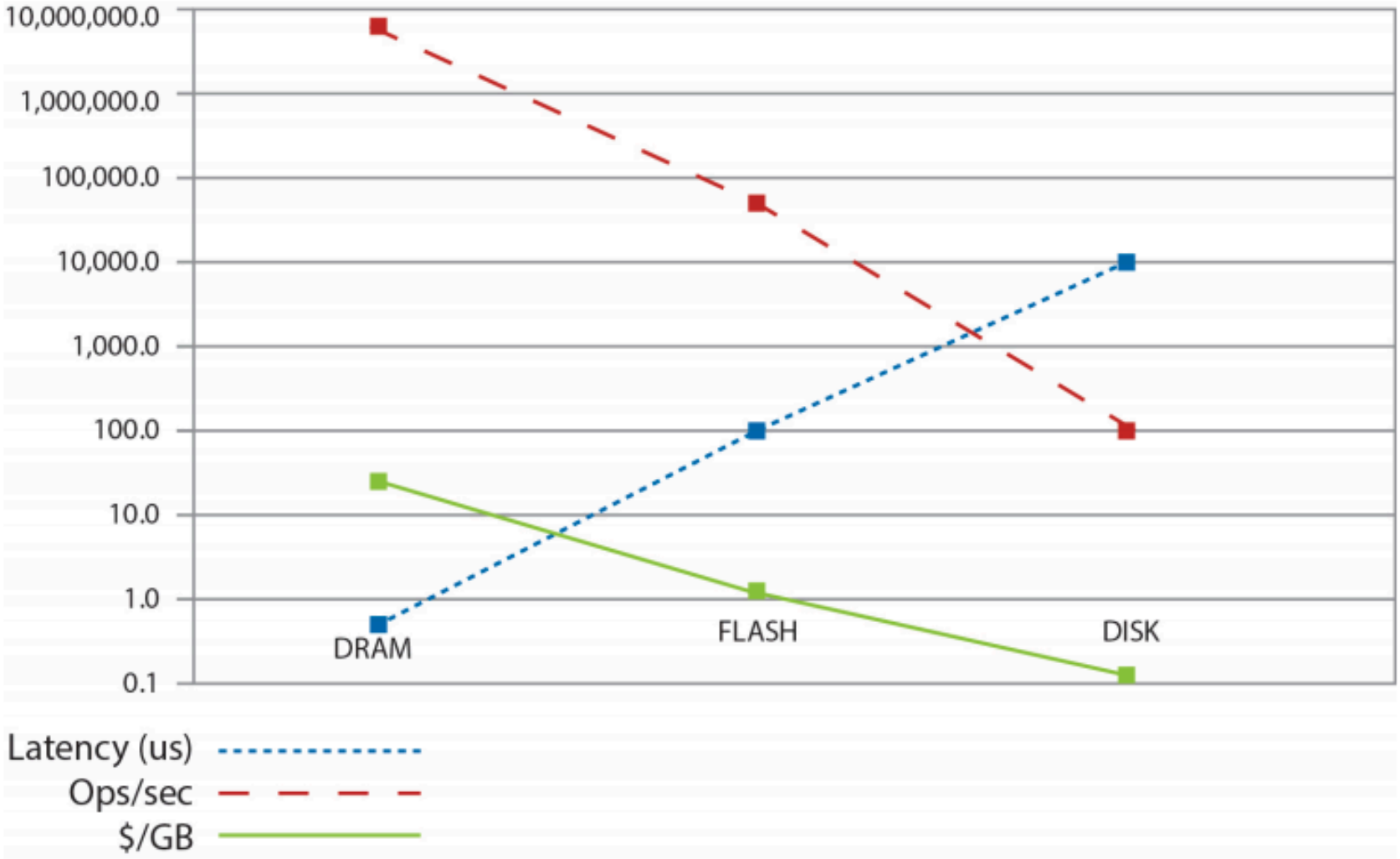
Source: Barroso and Urs Hölzle (2013)

The Cost of Moving Data Around Data Center



Source: Barroso and Urs Hölzle (2013)

The Cost of Moving Data Within a Server



Source: Barroso and Urs Hölzle (2013)

"Big Ideas" of Massive Data Processing in Data Centers

- Scale "out", not "up"
  - scale 'out' = combining many cheaper machines; scale 'up': increasing the power of each individual machine
  - Also called 'horizontal' vs 'vertical' scaling
- Move processing to the data
  - Clusters have limited bandwidth: we should move the task to the machine where the data is stored
- Process data sequentially, avoid random access
  - Seeks are expensive, disk throughput is reasonable
- Seamless scalability
  - E.g. if processing a certain dataset takes 100 machine hours, ideal scalability is to use a cluster of 10 machines to do it in about 10 hours.

Take-away

- Data contains value and knowledge.
- Data science is a cross-disciplinary and emerging research area with interesting applications in science, engineering and commerce etc.
- Cloud computing and data centres are natural infrastructures for data science.
- Further readings:
  - Chapter 1. Jimmy Lin and Chris Dyer. 2010. Data-Intensive Text Processing with Mapreduce. Morgan and Claypool Publishers. https://lintool.github.io/MapReduceAlgorithms/MapReduce-book-final.pdf
  - Vasant Dhar. 2013. Data science and prediction. Commun. ACM 56, 12 (December 2013), 64–73. https://dsmilab.github.io/assets/file/reading_list/data_science_and_prediction.pdf

Questions?

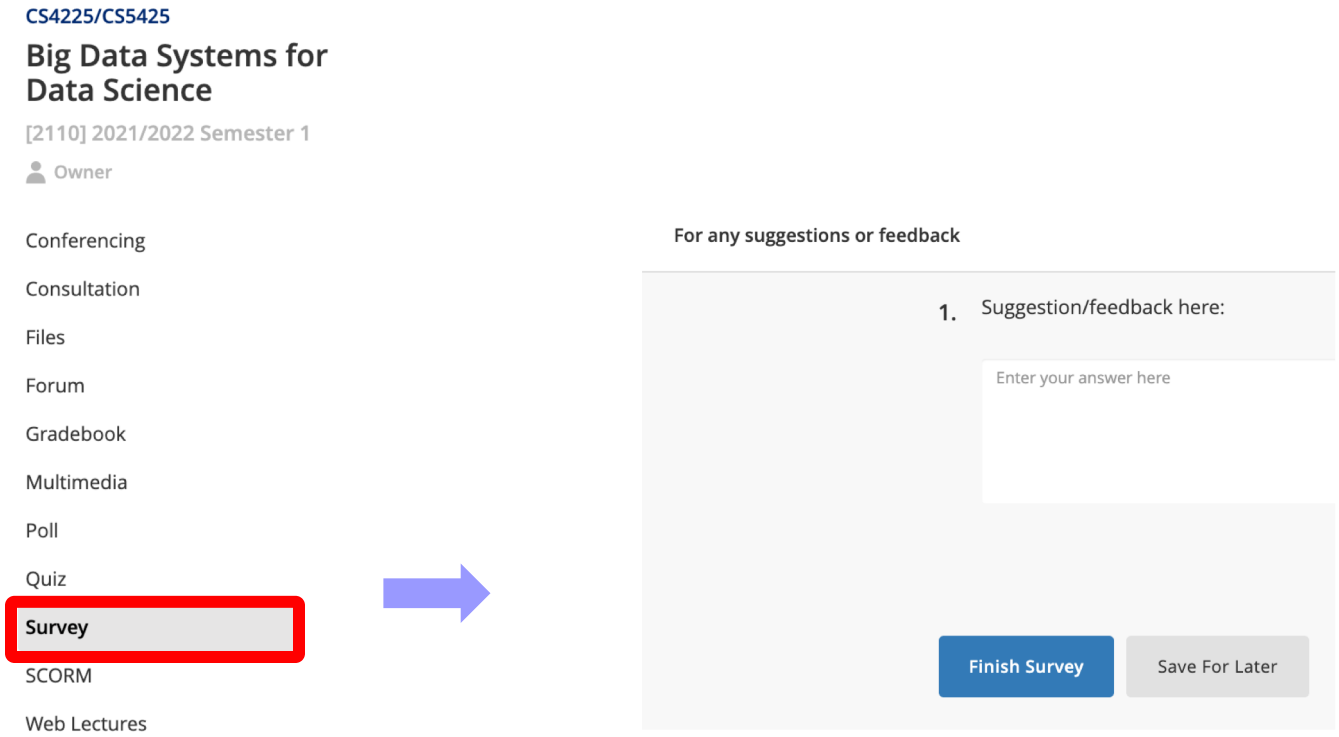
How do you want that data?

Q: What type of application domains are you most interested in?

1. Business
2. Scientific
3. Biomedicine
4. Engineering
5. Finance / Economics
6. Web / E-commerce
7. Education
8. Other

Suggestions / feedback

- I greatly appreciate suggestions or feedback, as they help to make the class better.
- This includes topics you'd really like the class to cover: please let me know.



Acknowledgement

- Slides are adopted/revised from
  - He Bingsheng
  - Jimmy Lin, http://lintool.github.io/UMD-courses/bigdata-2015-Spring/
  - Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. 2014. Mining of Massive Datasets (2nd ed.). Cambridge University Press. http://www.mmds.org/

Supplementary Slides on Data Science

Data Scientists in Singapore

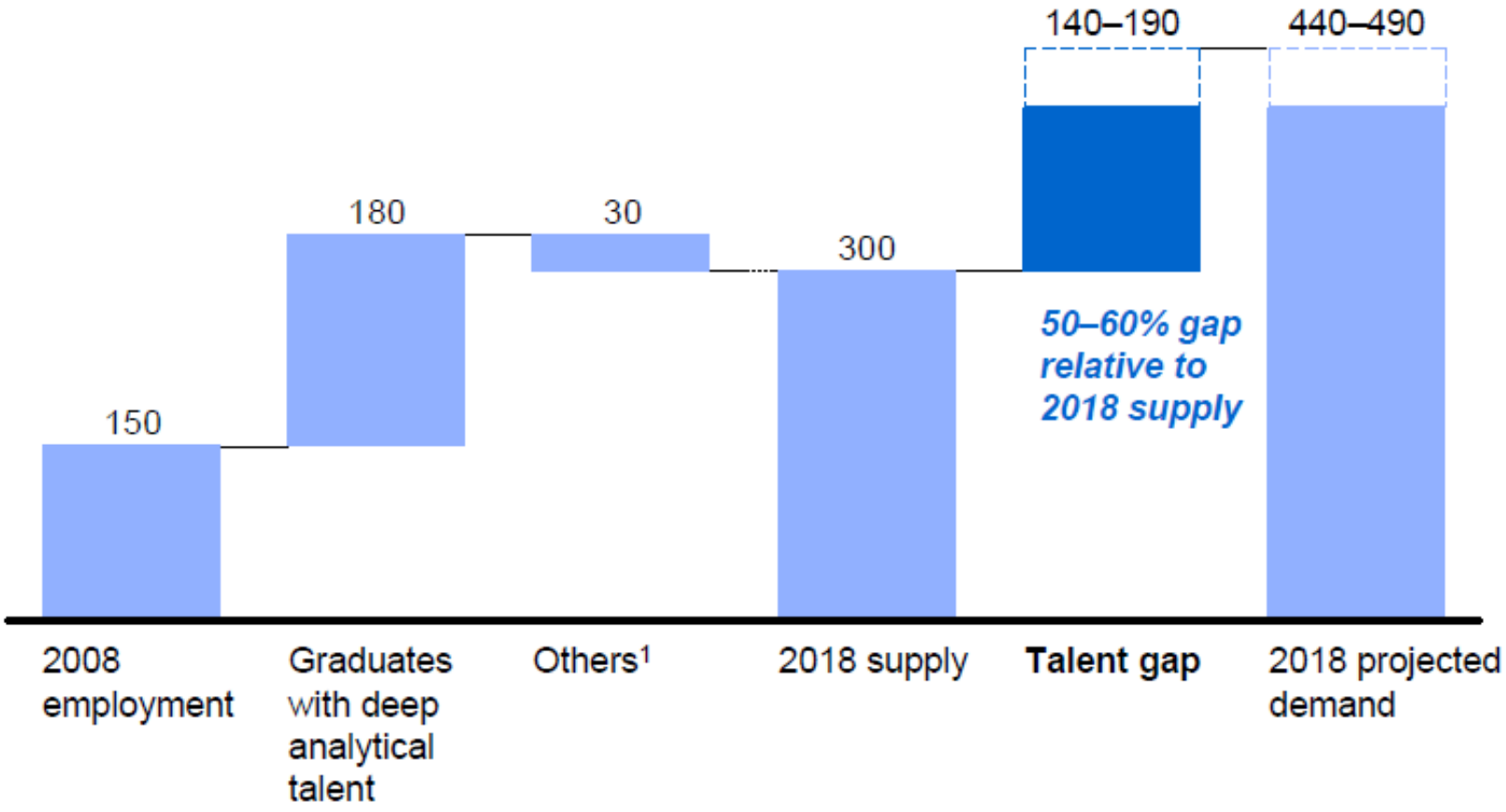
- Healthcare: Every public health cluster (Alexandra Health group, Eastern health alliance, National Healthcare Group, National University Health System, Jurong Health) in Singapore has a data team. Healthcare analytics includes traditional fields like medical research, as well as recent developments in applied operations research in areas like forecasting, population health analytics, logistics, patient care, etc.
- Startups: The Singapore startup scene is growing quickly, here's a non-exhaustive list of startups with established data teams: Grab, Lazada, Zalora, Redmart, Propertyguru, DataRobot, Honestbee.
- Local agencies & Finance: Singtel, IDA, E&Y, KPMG, DBS, AXA Data Innovations Lab, Aviva Digital Innovations.
- Technology / regional headquarters: MapR, Tripadvisor, PayPal, Nielsen, AirBnB, Facebook, Google, Twitter, IBM, Microsoft, Oracle, SAS, Pivotal, Bytedance.
- Research Institutes: A*star, i2r, NUS, SMU, NTU, SUTD, MIT SENSEable city labs, Rakuten Labs.

https://www.quora.com/What-are-the-best-data-science-teams-in-Singapore

Good news: Demand for Data Mining

Demand for deep analytical talent in the United States could be 50 to 60 percent greater than its projected supply by 2018

Supply and demand of deep analytical talent by 2018
Thousand people



1 Other supply drivers include attrition (-), immigration (+), and reemploying previously unemployed deep analytical talent (+).

SOURCE: US Bureau of Labor Statistics; US Census; Dun & Bradstreet; company interviews; McKinsey Global Institute analysis

Growth of Cloud Computing