CS4225/CS5425 Big Data Systems for Data Science

Introduction to Data Science

Bingsheng He
School of Computing
National University of Singapore
hebs@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?

- **Wikipedia definition**:

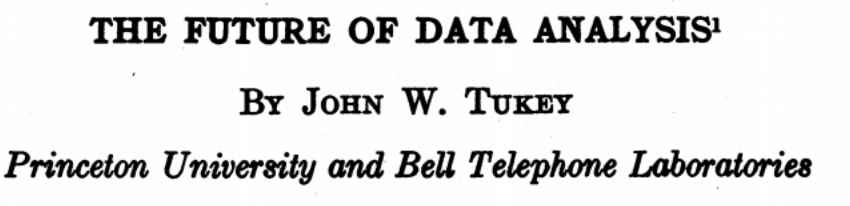
"Data science is an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured, which is a continuation of some of the data analysis fields such as statistics, machine learning, data mining, and predictive analytics."

What is Data Science?

- **Wikipedia 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**: one of the first references to "data analysis" was in 1962, when statistician John Tukey described a field called "data analysis":



THE FUTURE OF DATA ANALYSIS[1]

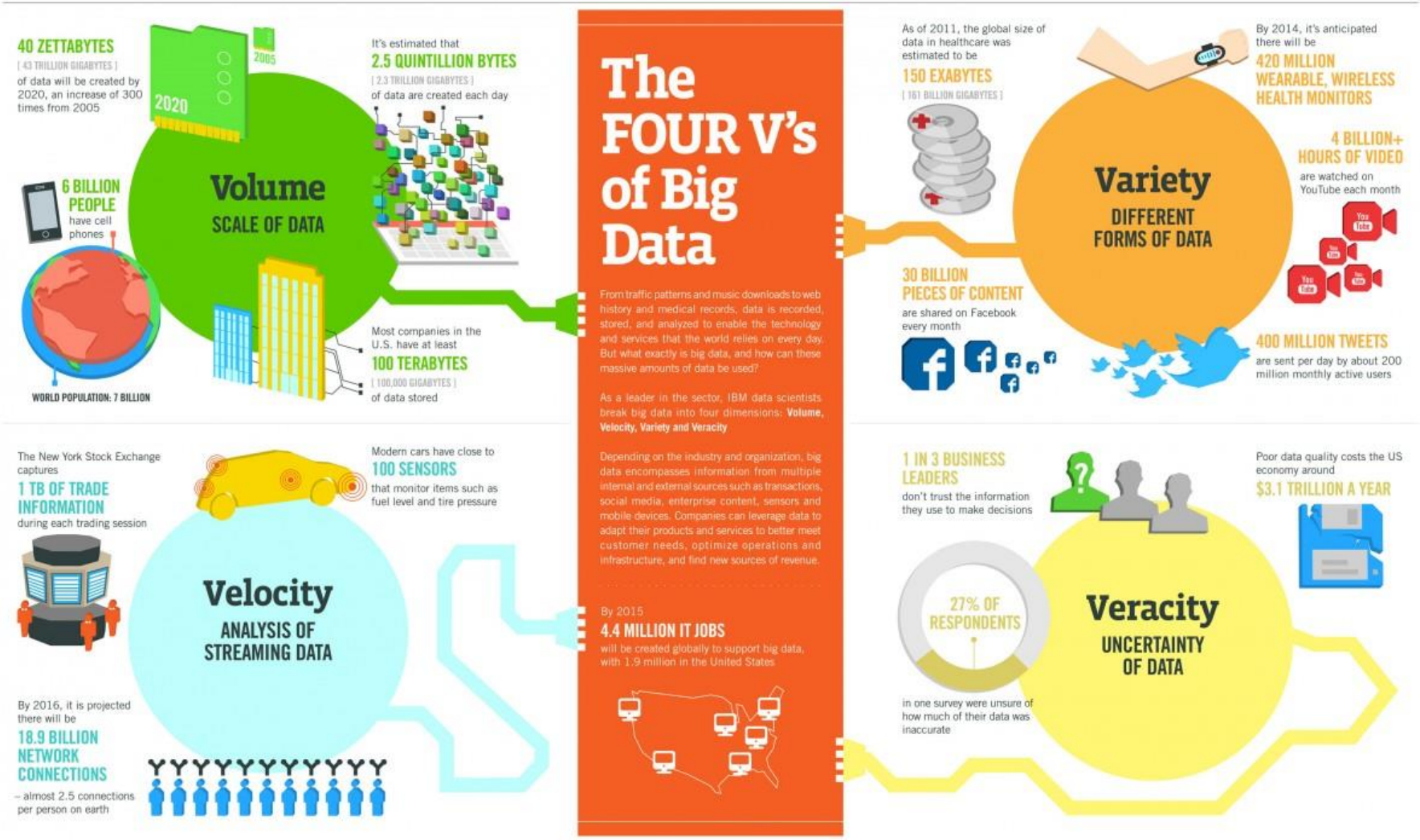
By John W. Tukey

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… procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data…"

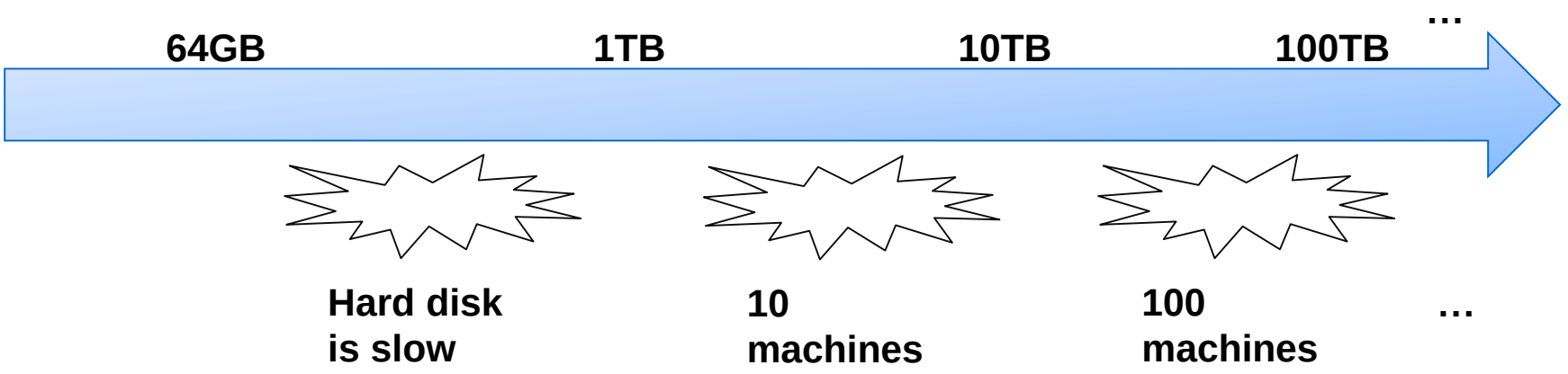
Data Contains Value and Knowledge

Challenges of Big Data: the 4 ‘V’s



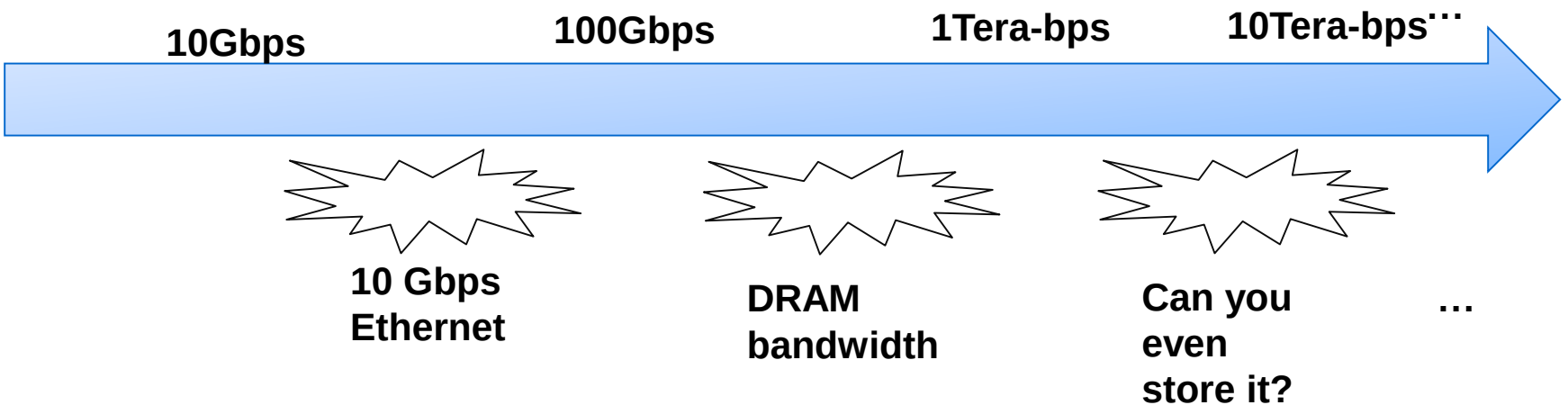
Volume

- Volume: The scale of data
- The challenges of large volume
  - Performance
  - Cost
  - Reliability
  - Algorithm design complexity



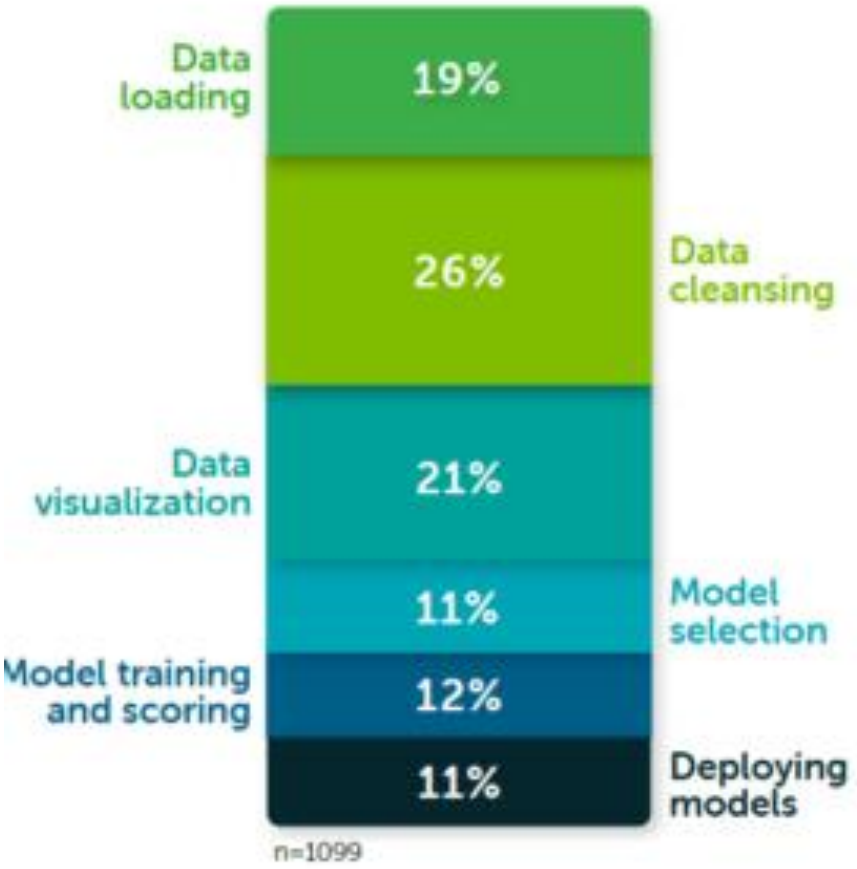
Velocity

- Velocity: the speed of new data (streaming data)
- The challenges of high velocity
  - Performance
  - Cost
  - Reliability
  - Algorithm design complexity



Variety and Veracity

- Why Variety matters?
  - "One size does not fit all"
  - Data integration
  - Multi-modal learning
- Why Veracity matters?
  - Dirty and noisy data
  - Data provenance
  - Data uncertainty



https://www.datanami.com/2020/07/06/data-prep-still-dominates-data-scientists-time-survey-finds/

Veracity: ~3.3% of data in some of the most popular datasets (e.g., ImageNet) 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

Veracity: Importance of Data Quality

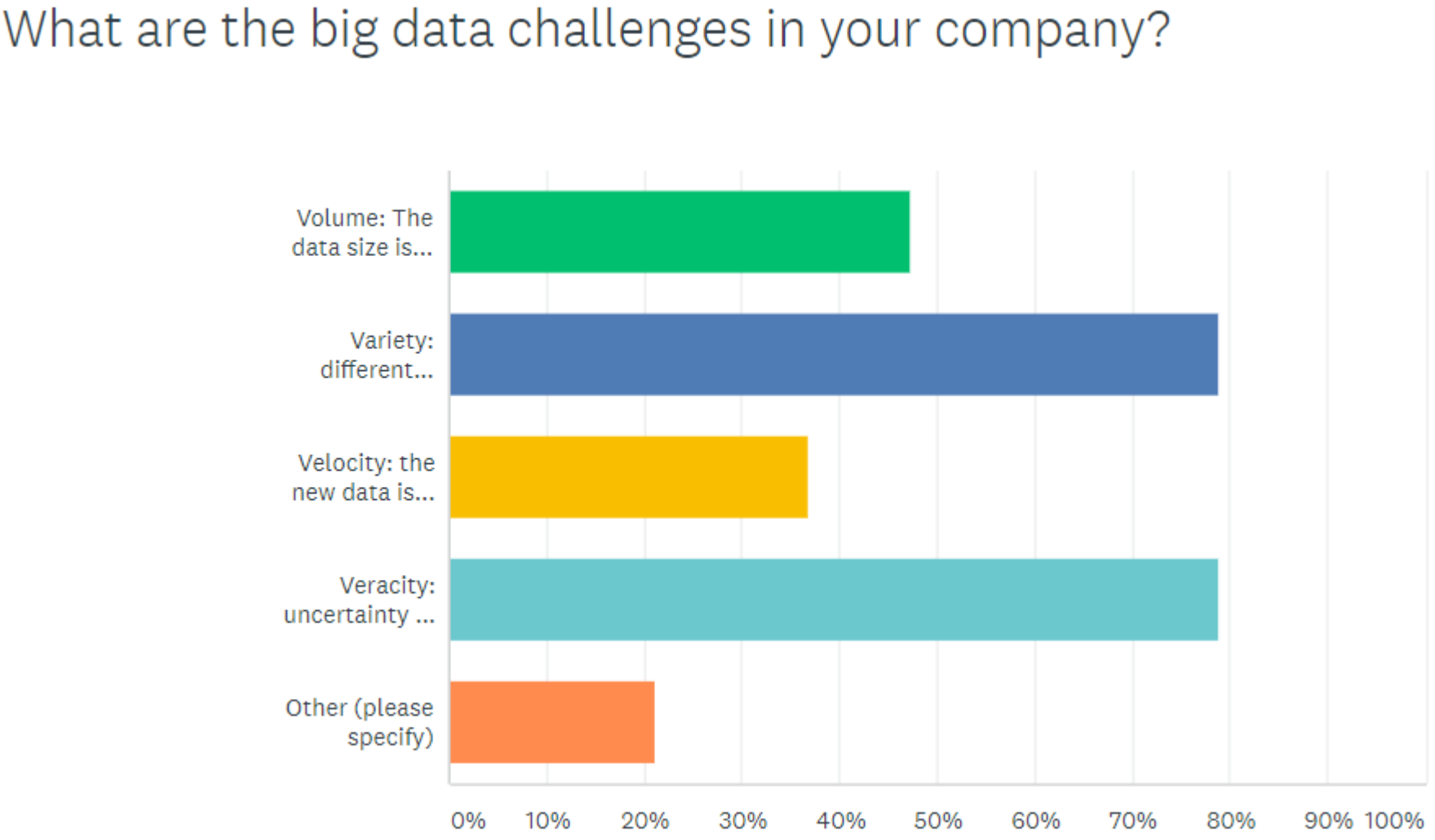
Forbes

ENTERPRISE & CLOUD

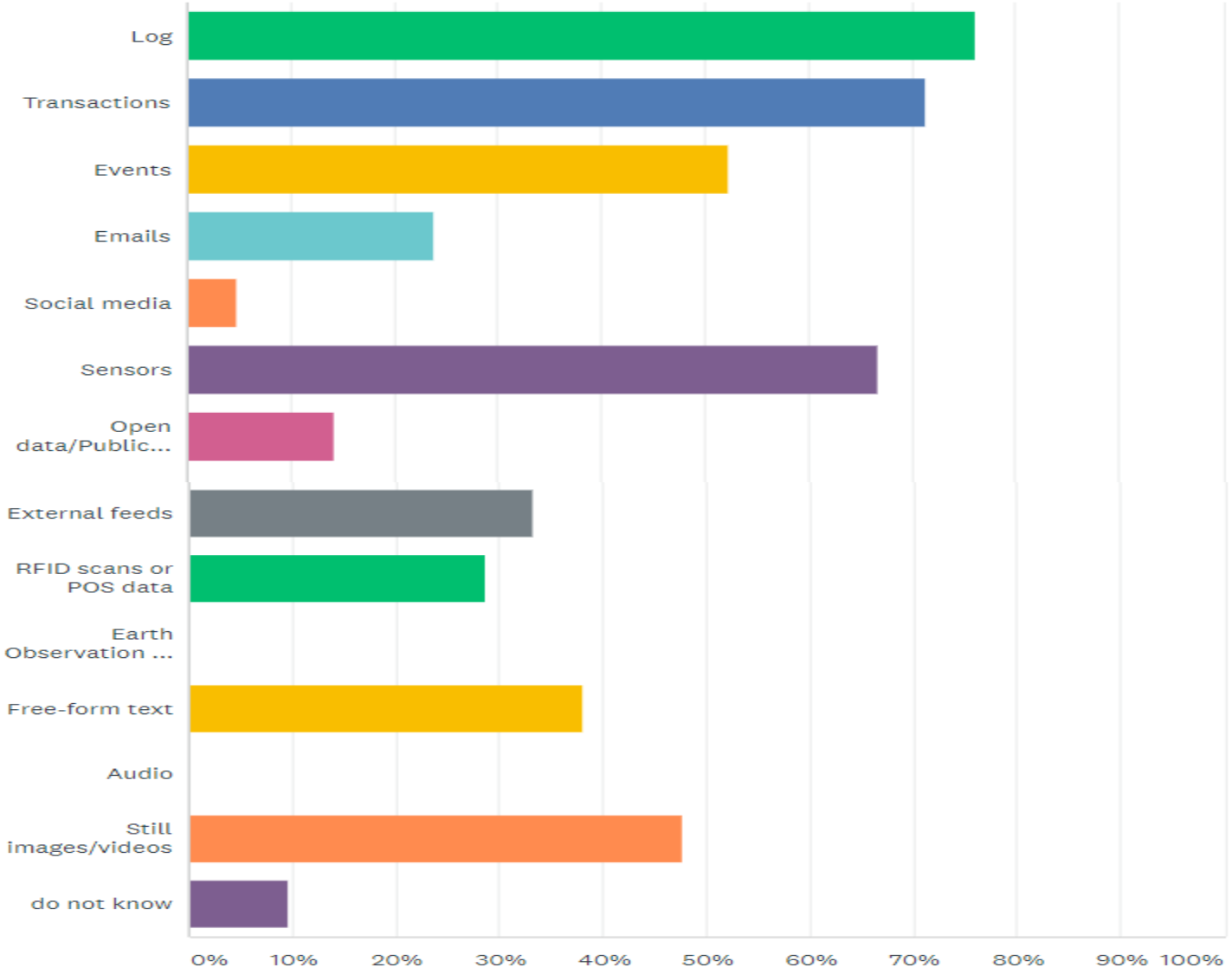
Andrew Ng Launches A Campaign For Data-Centric AI

Data is eating the world so Andrew Ng wants to make sure we radically improve its quality. "Data is food for AI," says Ng, and he is launching a campaign to shift the focus of AI practitioners from model/algorithm development to the quality of the data they use to train the models.

Our Survey

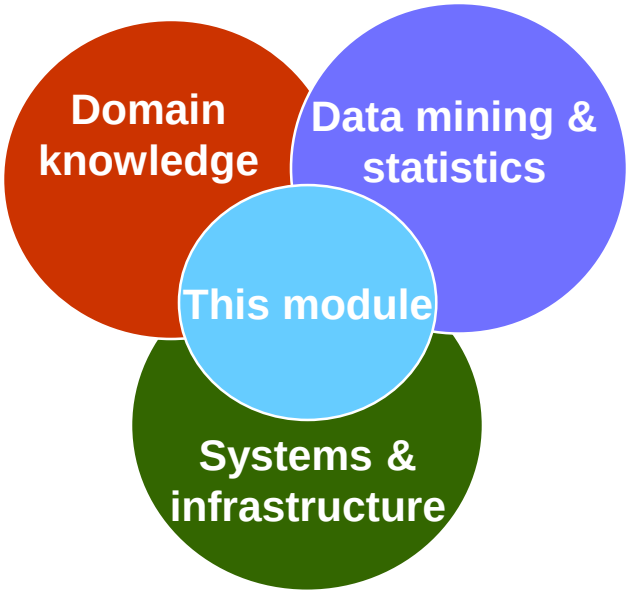

What are the big data challenges in your company?
Volume: The data size is...
Variety: different...
Velocity: the new data is...
Veracity: uncertainty ...
Other (please specify)
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%

From what sources does your organisation collect, or expects to collect, data?



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 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)

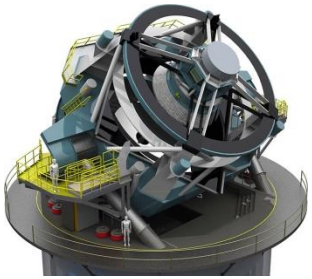
facebook

LHC: ~15 PB a year

CERN

amazon web services™

S3: 2T objects, 1.1M request/second (4/2013)



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

SKA: 0.3 – 1.5 EB per year (~2020)

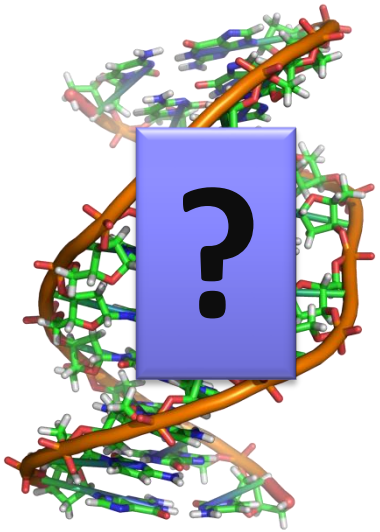
How much data?

Why big data?

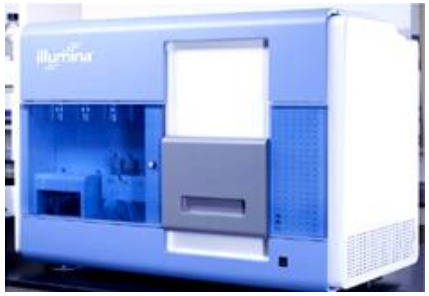
Science
Engineering
Commerce

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)

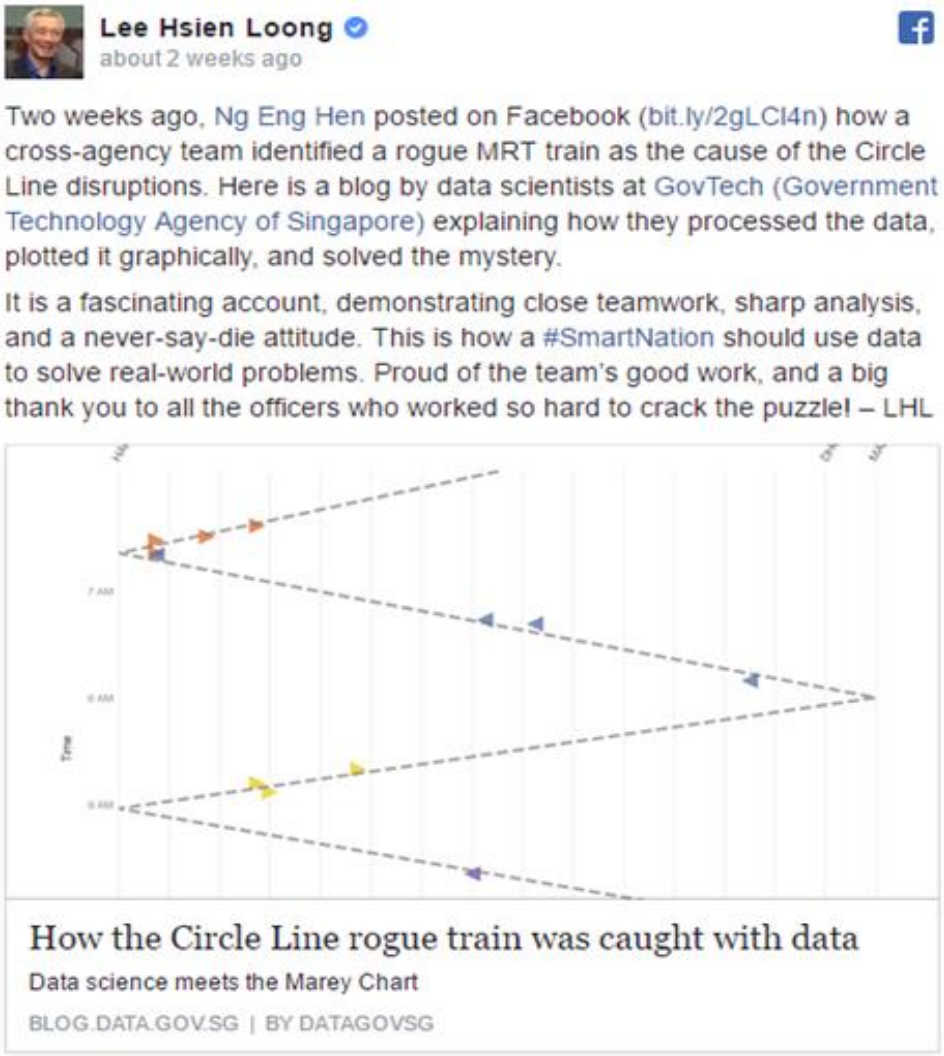
Precision medicine
Health Insurance
…

Engineering

The unreasonable effectiveness of data

Source: Wikipedia (Three Gorges Dam)

How the Circle Line rogue train was caught with data?

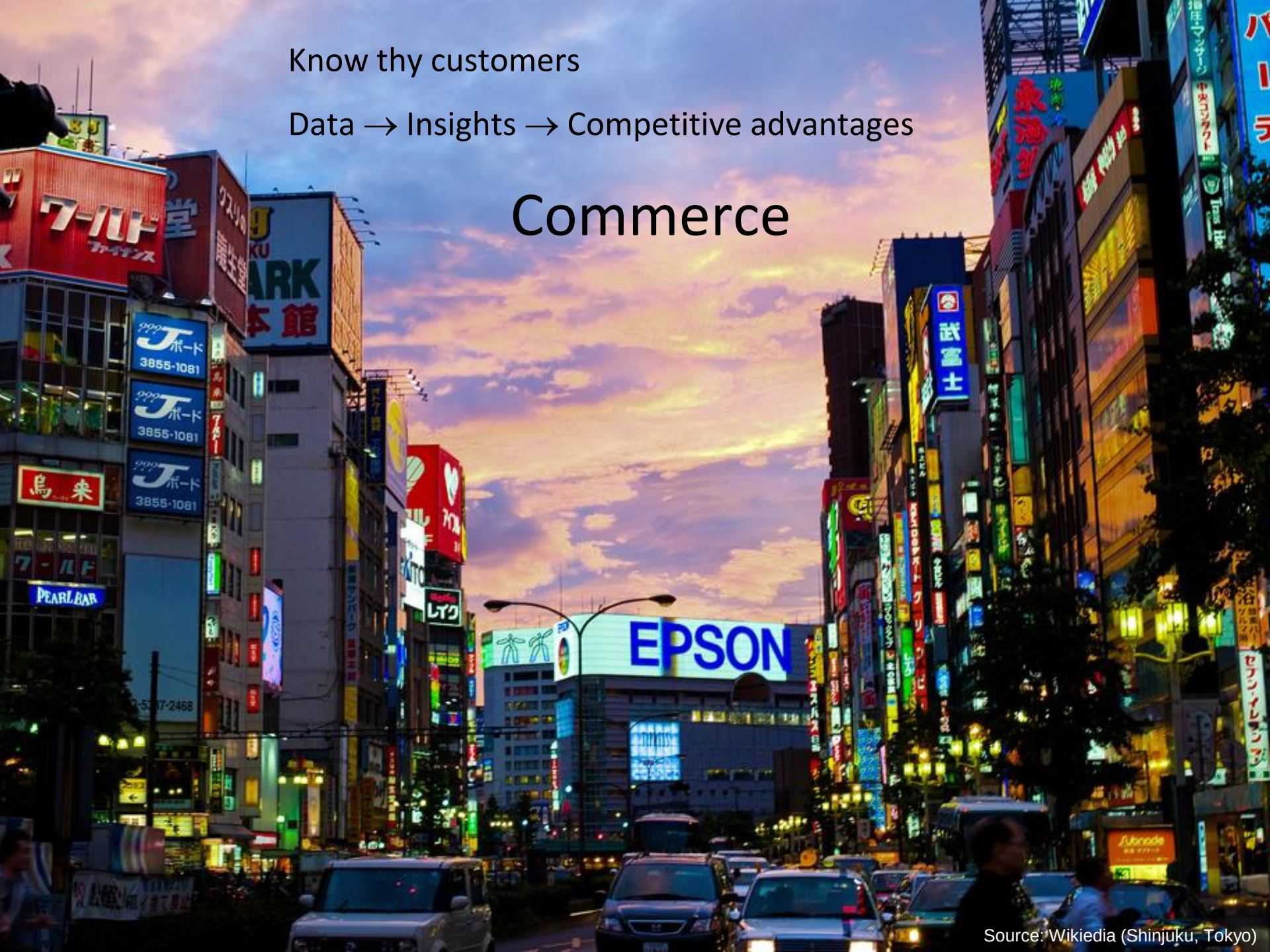


Lee Hsien Loong
about 2 weeks ago

Two weeks ago, Ng Eng Hen posted on Facebook (bit.ly/2gLCl4n) how a cross-agency team identified a rogue MRT train as the cause of the Circle Line disruptions. Here is a blog by data scientists at GovTech (Government Technology Agency of Singapore) explaining how they processed the data, plotted it graphically, and solved the mystery.

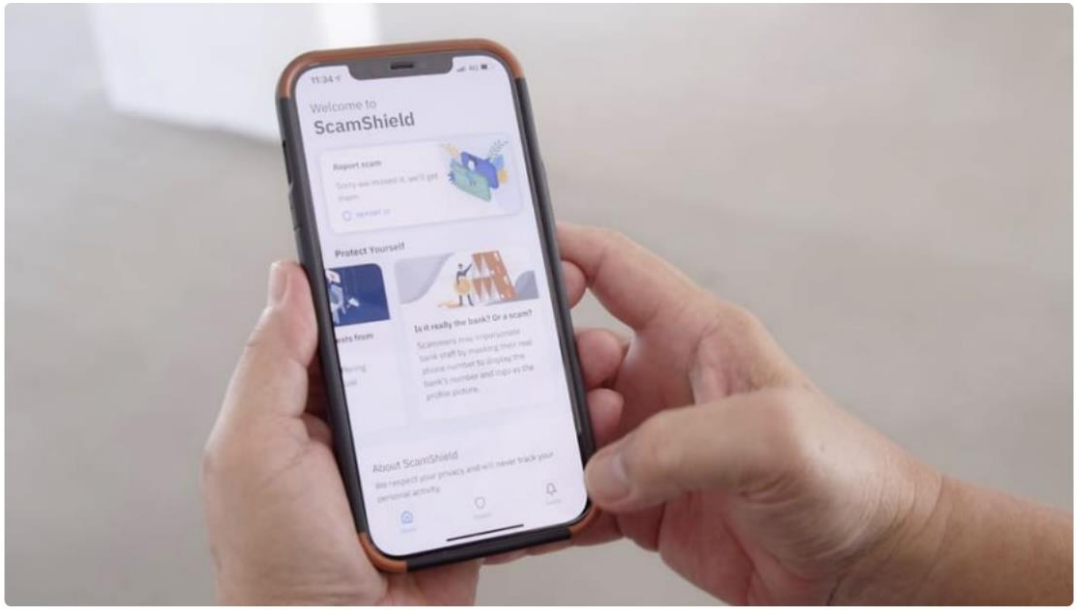
It is a fascinating account, demonstrating close teamwork, sharp analysis, and a never-say-die attitude. This is how a #SmartNation should use data to solve real-world problems. Proud of the team's good work, and a big thank you to all the officers who worked so hard to crack the puzzle! – LHL

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


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

Singapore

More than 722,000 SMSes reported on ScamShield app over last six months

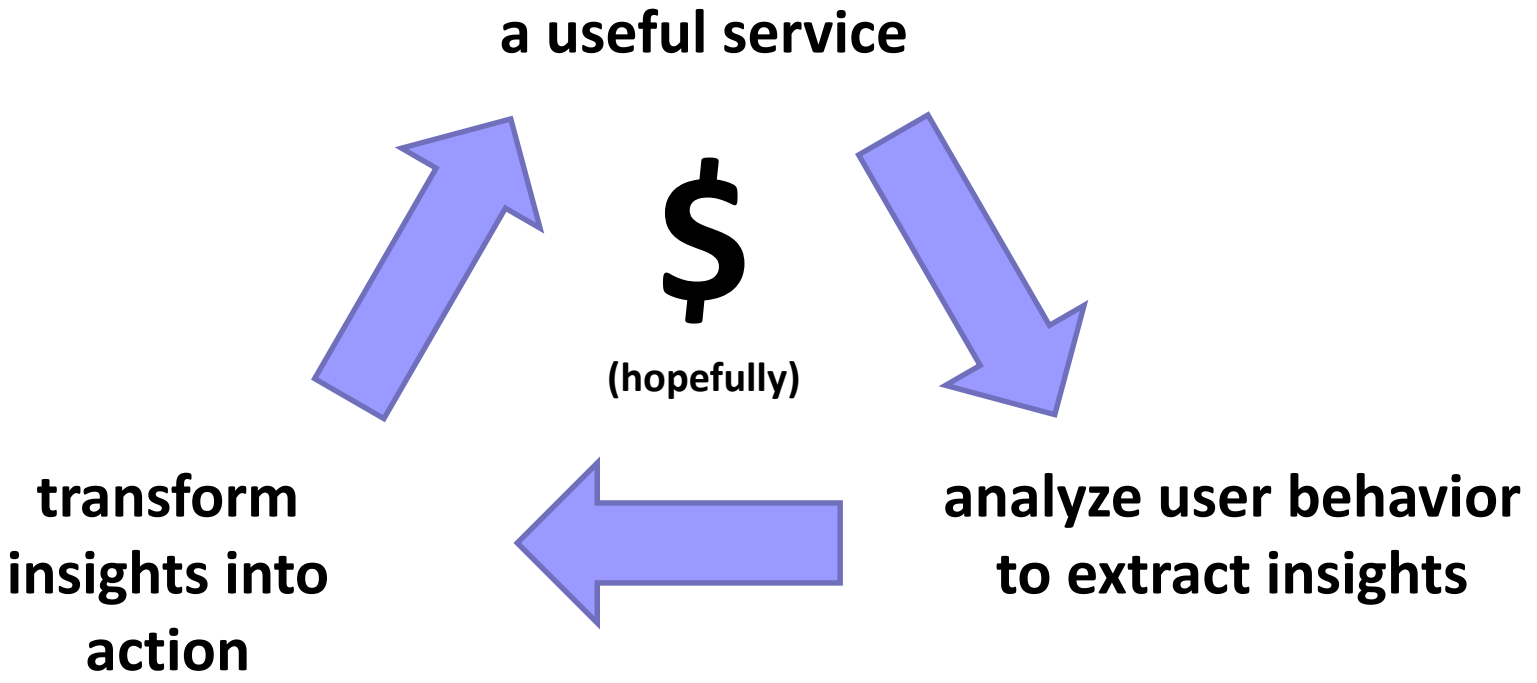


The home page of ScamShield. (Screenshot from a video by the Singapore Police Force)

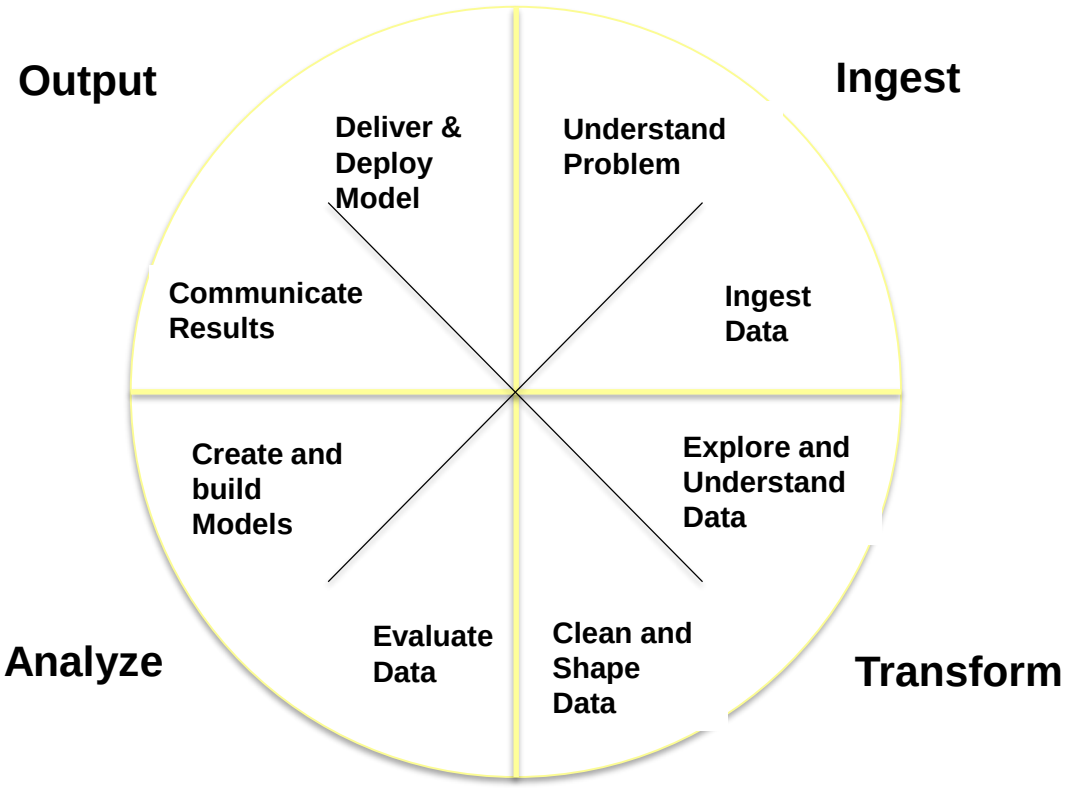
Grace Yeoh
@GraceYeohCNA

29 May 2021 06:14PM
(Updated: 29 May 2021 06:20PM)

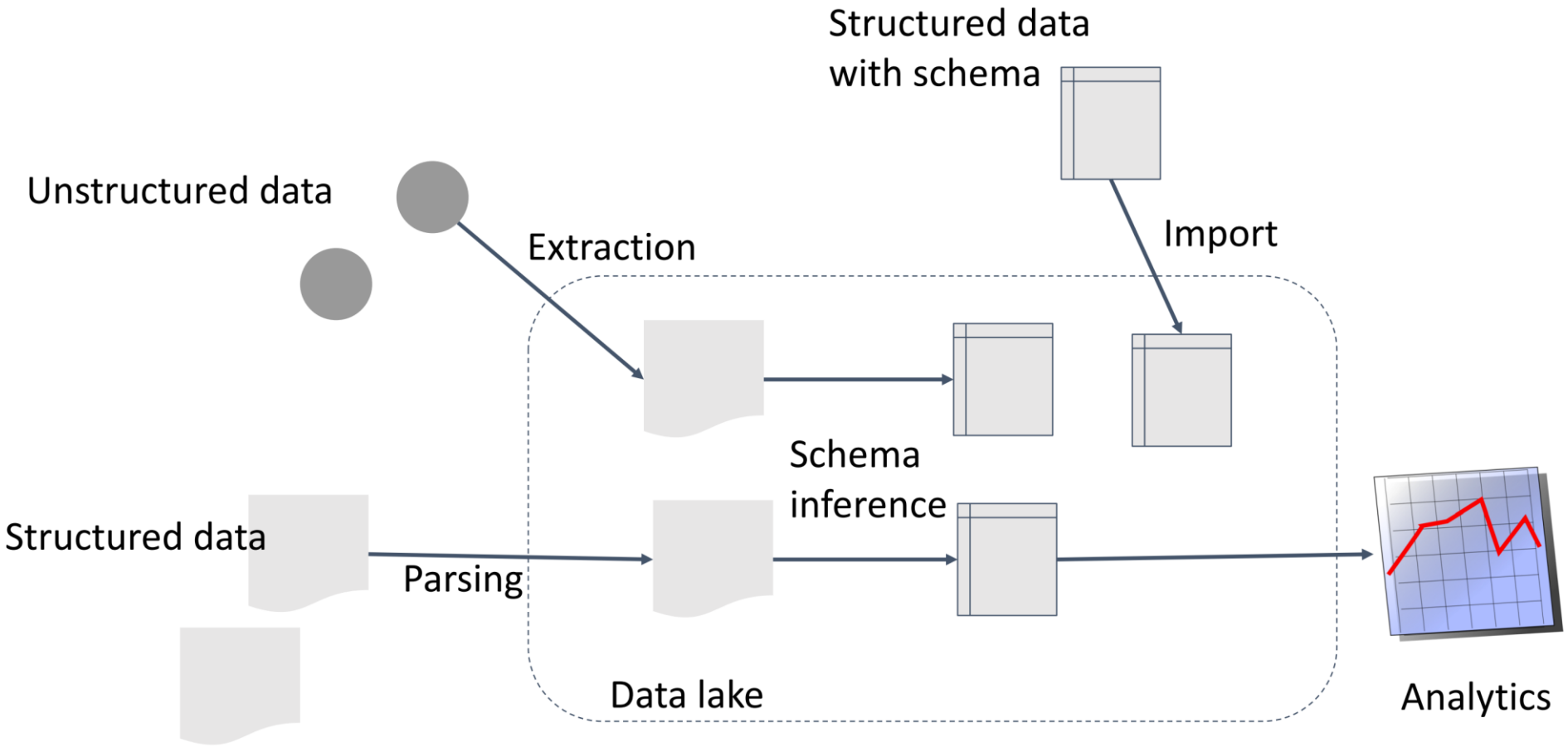
Virtuous Product Cycle



Data Lifecycle



Data Lake: The Next Gen of Big Data?



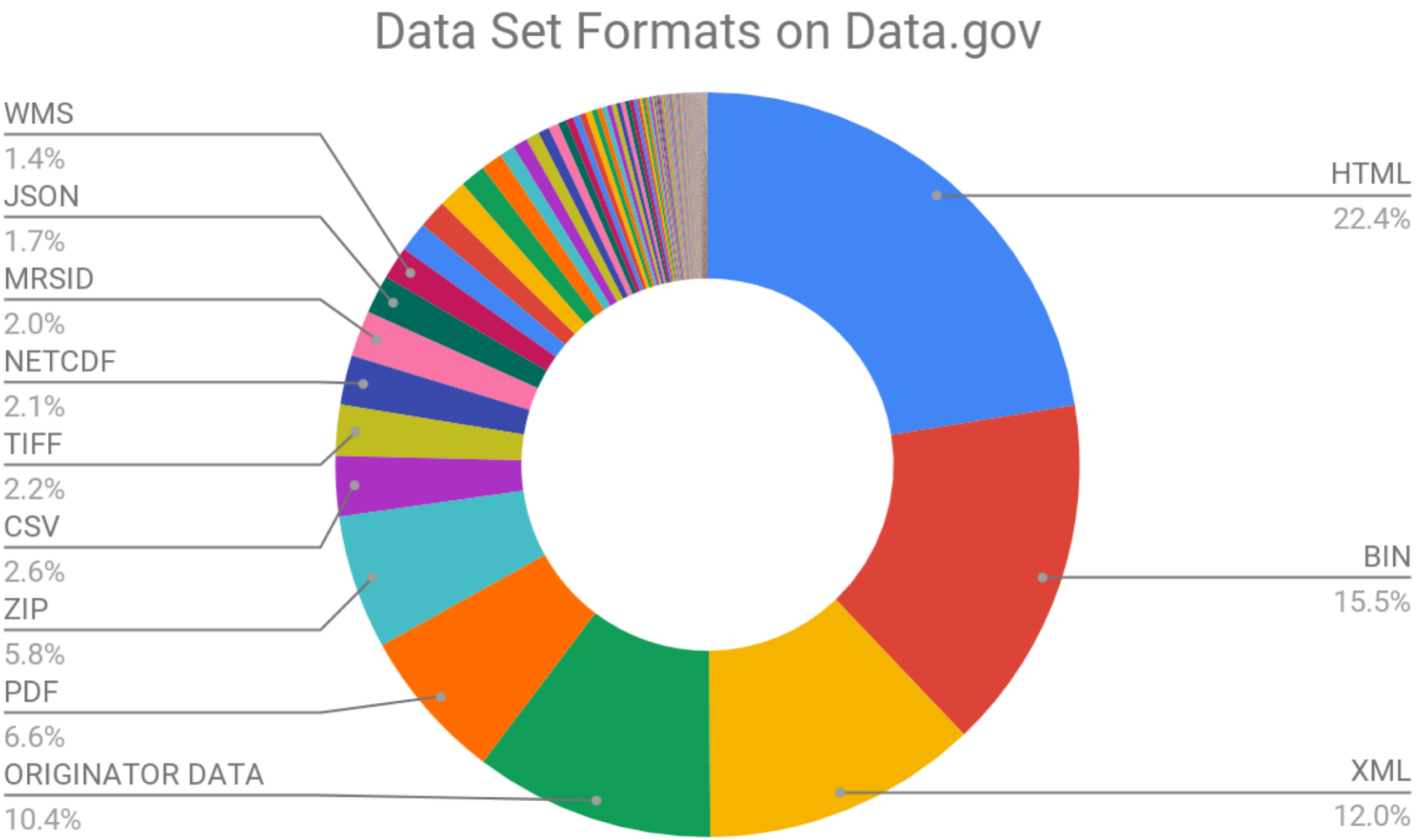
Data Lakes ≥ Lots of (Big) Data

- **Scenario 1:** A global investment bank (>10k employees). More than 100k datasets.
  - During 2008 financial crisis, regulator asked for risk exposure to Lehman Brothers
  - Worked day-and-night scrambled to find relevant datasets scattered across the company to compile the report
- **Data Discovery** is the task of finding relevant datasets for analysis.

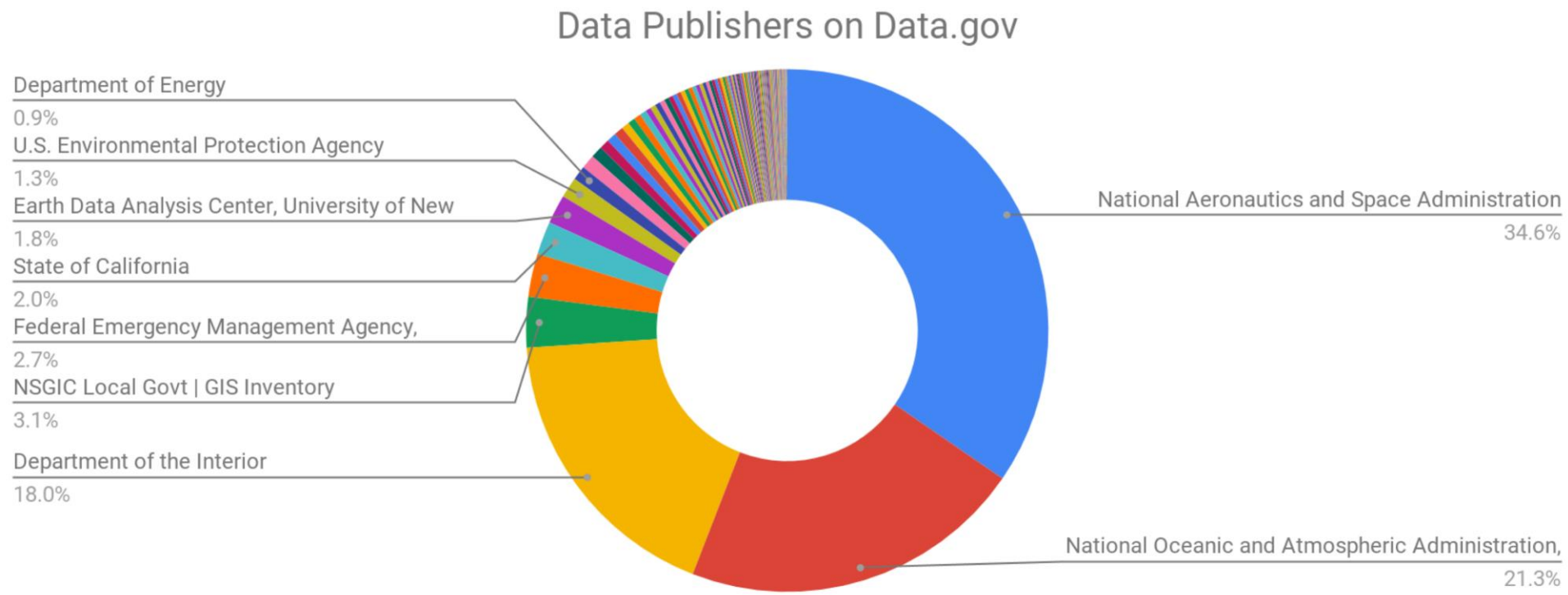
Data Lakes are Evolving

- **Scenario 2:** a data science research institution (~100 employees). 1000-10k datasets.
  - Datasets are stored in HDFS directories
  - Many duplicates as datasets are often being copied for new project
  - Datasets are constantly being updated, having their schema altered, being derived into new ones, and disappearing/reappearing
- **Dataset Versioning** is to maintain all versions of datasets for storage cost-saving, collaboration, auditing, and experimental reproducibility.

Data Lakes are Diverse

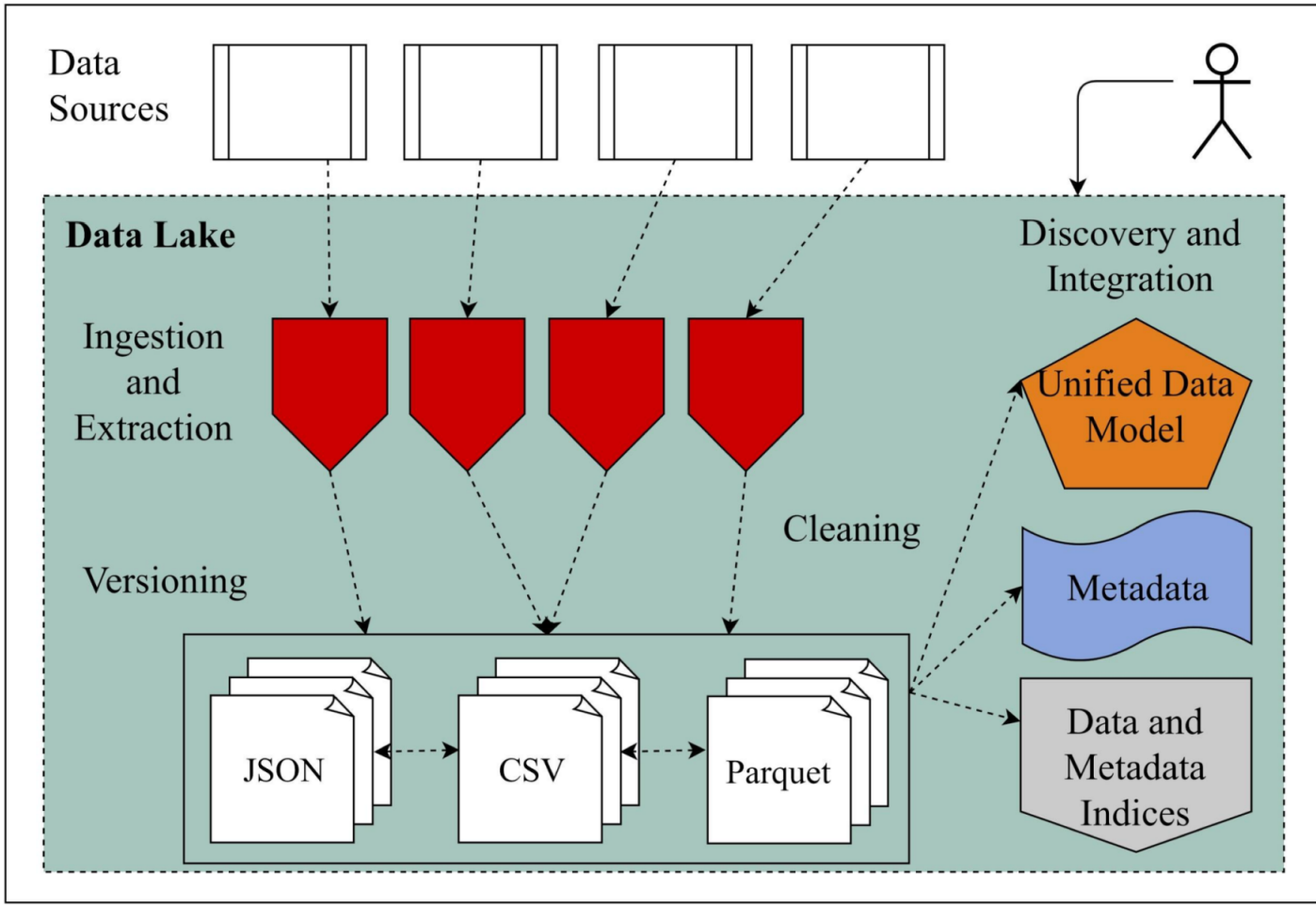


Data Lakes are Diverse



Common Tasks in Data Lakes

- Metadata Management
- Ingestion
- Extraction
- Cleaning
- Integration
- Discovery
- Versioning



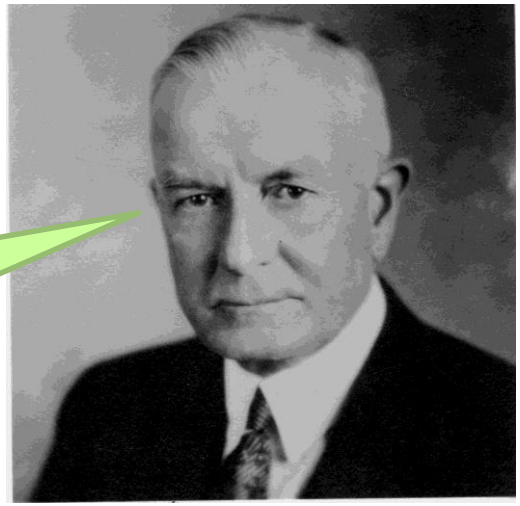
Source: Wikipedia (Noctilucent cloud)

Interlude: Cloud Computing

Source: Wikipedia (Clouds)

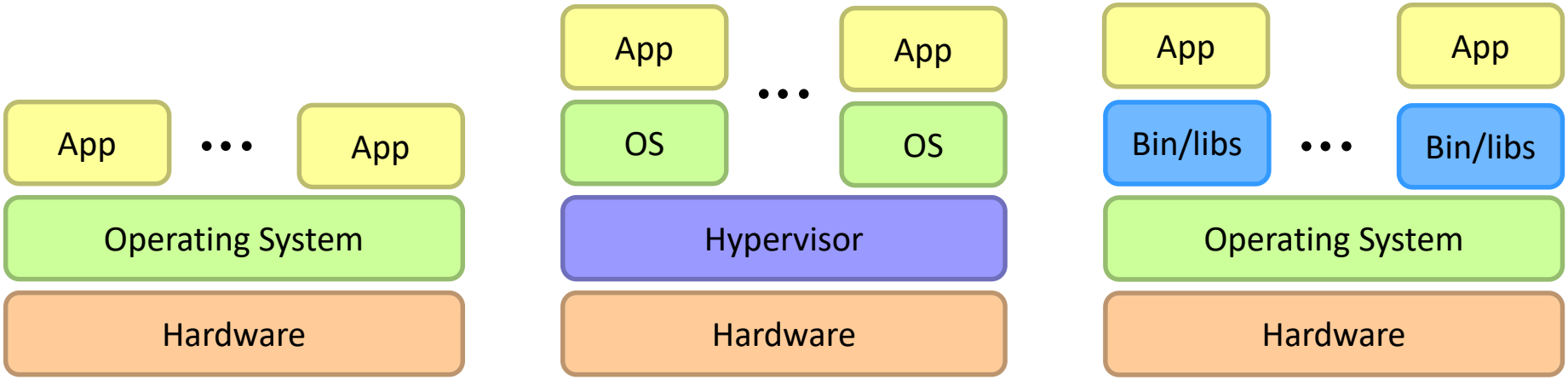
Utility Computing

- 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

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.
- Data lake could be the new era of big data.
- Clouds are natural infrastructures for data science.
- Further readings:
  - Chapter 1. Jimmy Lin and Chris Dyer. 2020. 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?

Acknowledgement

- Slides are adopted/revised from
  - Bryan Hooi
  - 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

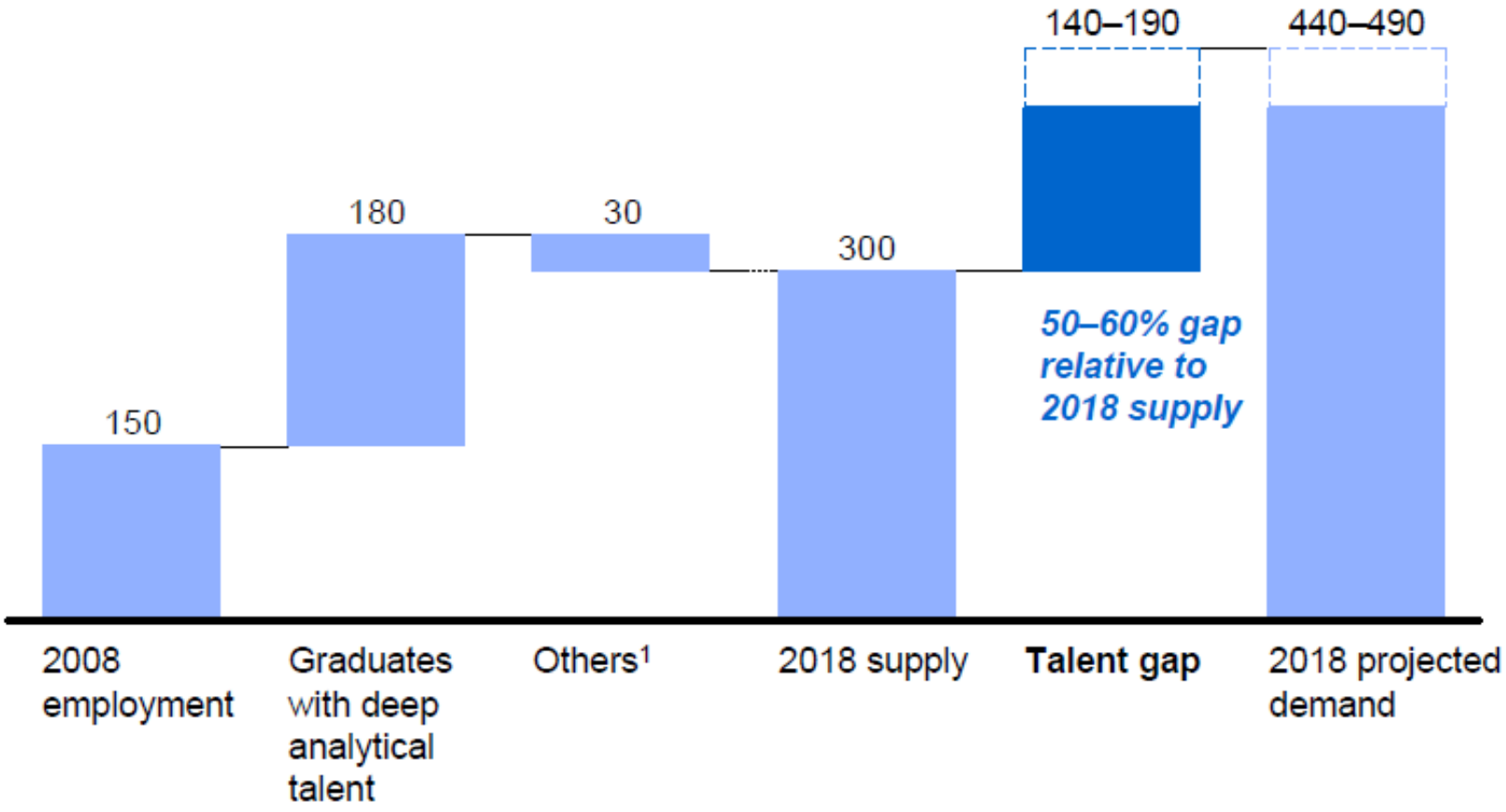
- Healthcare: Every public health cluster (Alexandra Health group, Eastern health alliance, National Healthcare Group, National University Health System, Jurong Health) in Singapore has a data team. Healthcare analytics includes traditional fields like medical research, as well as recent developments 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