

# **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](mailto:bhooi@comp.nus.edu.sg)



# Learning Objectives

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

# What is Data Science?

- **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<sup>1</sup>**

By JOHN W. TUKEY

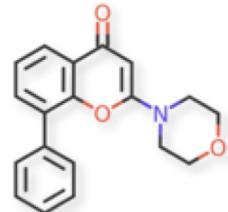
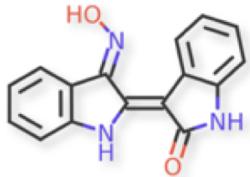
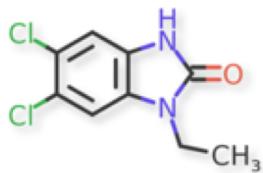
*Princeton University and Bell Telephone Laboratories*

“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...”

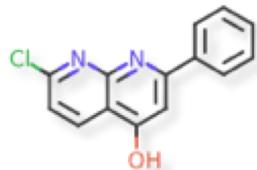
# Q: What rule characterizes toxic molecules?



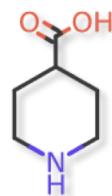
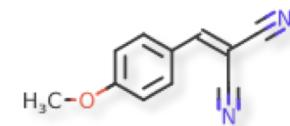
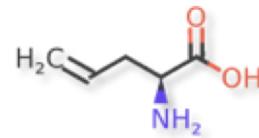
Toxic



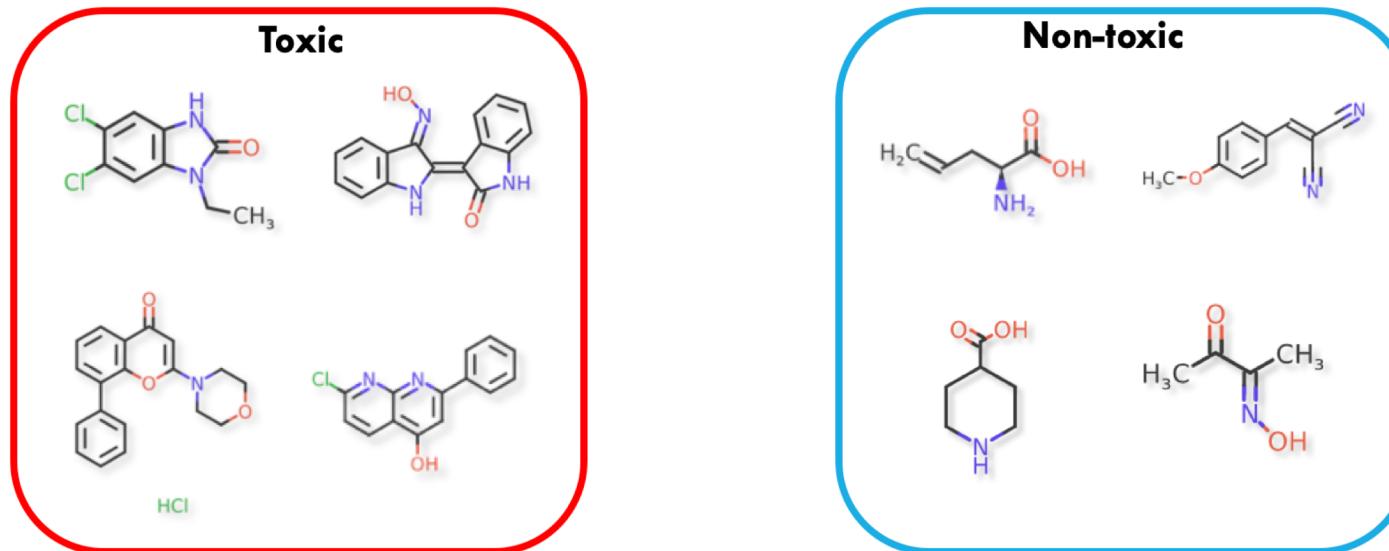
HCl



Non-toxic



# Q: What rule characterizes toxic molecules?

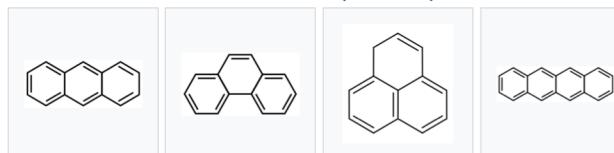


if > 1:  
toxic  
else  
nontoxic

## Polycyclic aromatic hydrocarbon

From Wikipedia, the free encyclopedia

Principal PAH Compounds



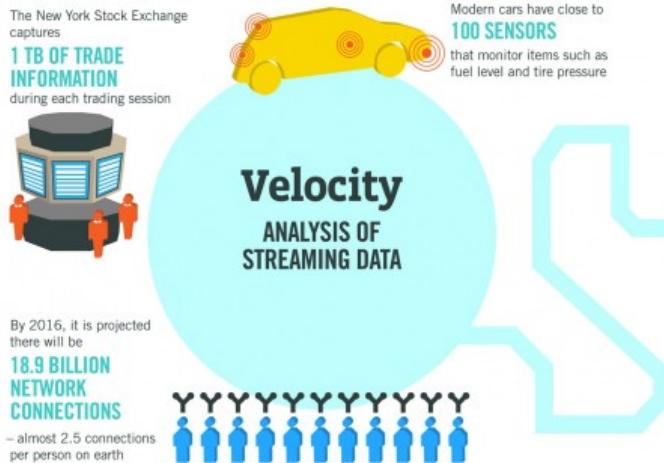
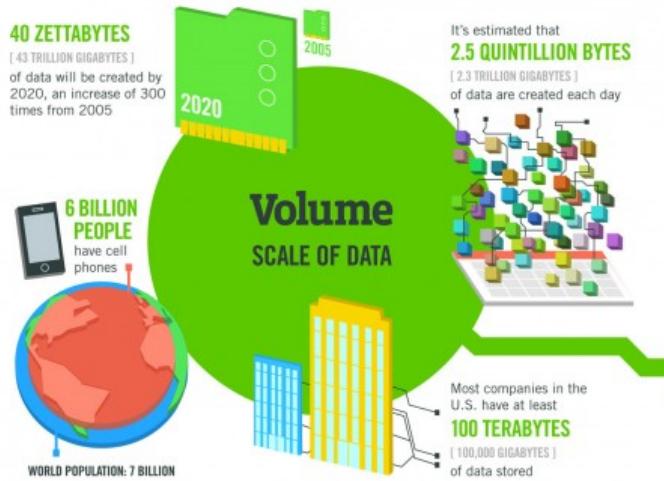
### Cancer [edit]

PAHs have been linked to [skin](#), [lung](#), [bladder](#), [liver](#), and [stomach](#) cancers in well-established animal model studies.<sup>[72]</sup> Human carcinogens are identified in the section "Regulation and Oversight" below.

# Data Contains Value and Knowledge



# The 4V of Big Data from IBM



## The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume**, **Velocity**, **Variety** and **Veracity**.

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015 4.4 MILLION IT JOBS will be created globally to support big data, with 1.9 million in the United States



As of 2011, the global size of data in healthcare was estimated to be

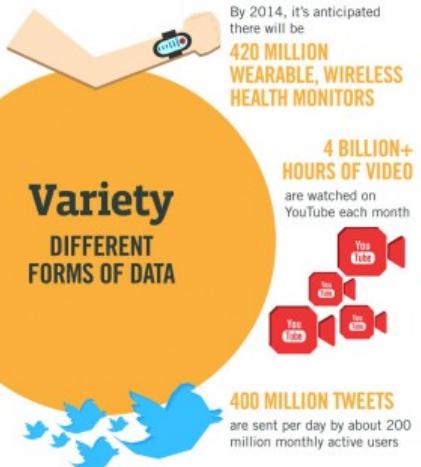
150 EXABYTES

[ 161 BILLION GIGABYTES ]

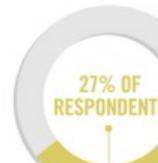


30 BILLION PIECES OF CONTENT

are shared on Facebook every month



1 IN 3 BUSINESS LEADERS don't trust the information they use to make decisions



27% OF RESPONDENTS in one survey were unsure of how much of their data was inaccurate



Poor data quality costs the US economy around \$3.1 TRILLION A YEAR



# 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

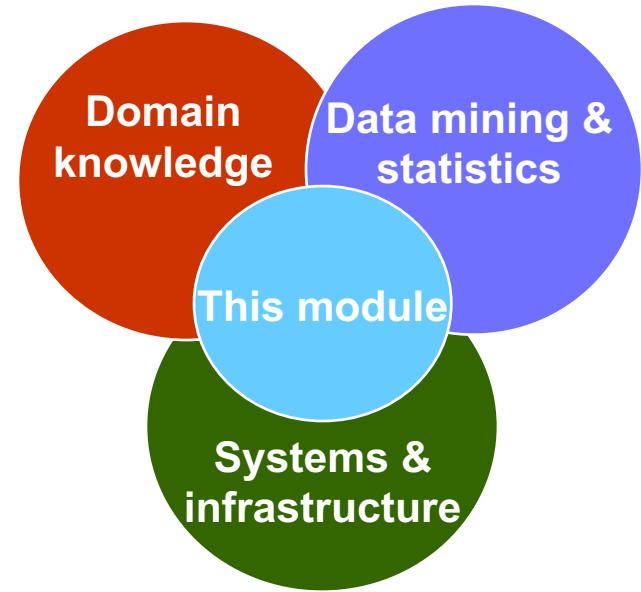


given: white stork  
corrected: black stork

[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.



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)



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



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



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

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



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



640K ought to be  
enough for anybody.

JPMorganChase

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



LHC: ~15 PB a year



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



SKA: 0.3 – 1.5 EB  
per year (~2020)

**How much data?**

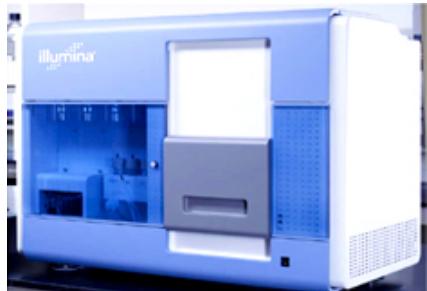


**Why big data?** Science  
Engineering  
Commerce

# Genome is Big Data



**Subject genome**



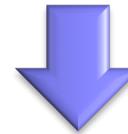
**Sequencer**



GATGCTTACTATGC~~GGGGCCCC~~  
CGGTCTAAC~~TGCTTACTATGC~~  
GCTTACTATGC~~GGGGCCCTT~~  
**AATGCTTACTATGC~~GGGGCCCTT~~**  
**TAATGCTTACTATGC**  
**AATGCTTAGCTATGC~~GGGG~~**  
**AATGCTTACTATGC~~GGGGCCCTT~~**  
**AATGCTTACTATGC~~GGGGCCCTT~~**  
**CGGTCTAGATGCTTACTATGC**  
**AATGCTTACTATGC~~GGGGCCCTT~~**  
**CGGTCTAAC~~TGCTTAGCTATGC~~**  
**ATGCTTACTATGC~~GGGGCCCTT~~**

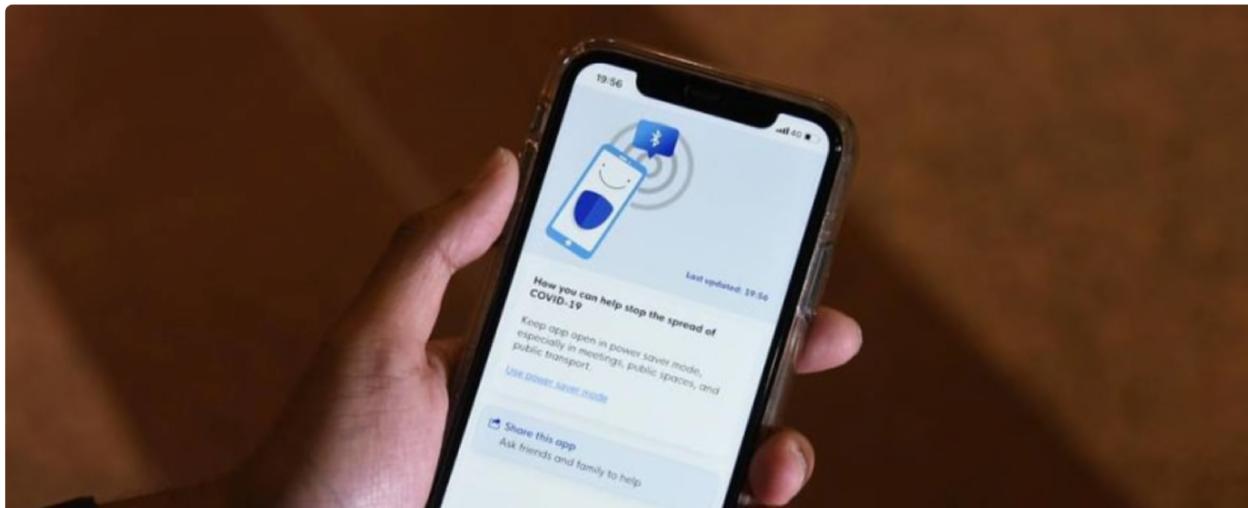
**Reads**

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!



# How the Circle Line rogue train was caught with data?

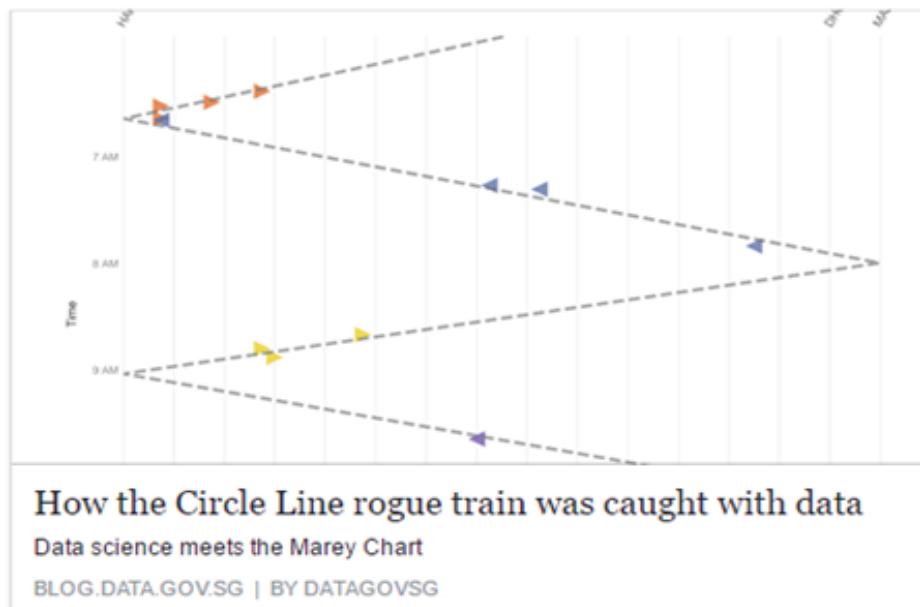


Lee Hsien Loong about 2 weeks ago



Two weeks ago, Ng Eng Hen posted on Facebook ([bit.ly/2gLCI4n](https://bit.ly/2gLCI4n)) 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#.5sehgyfsq>

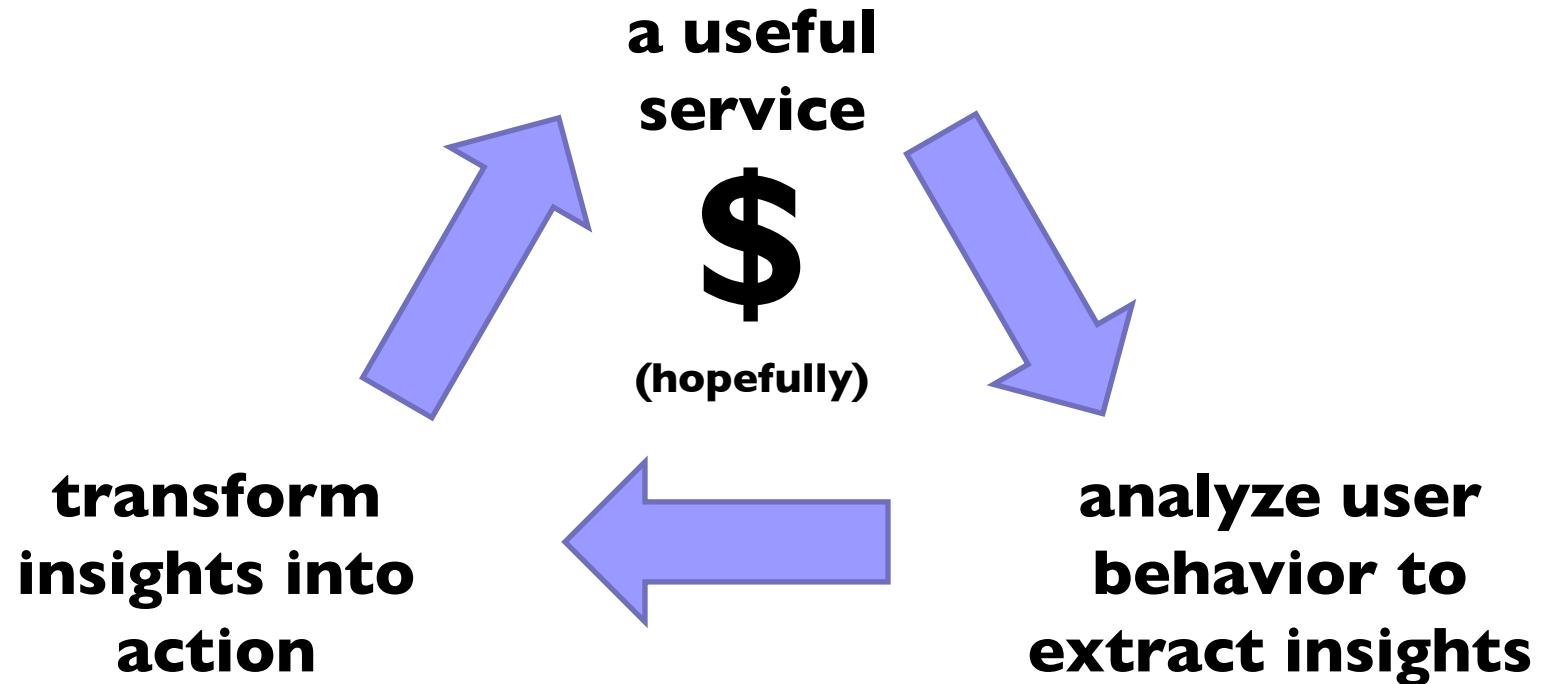


Know thy customers

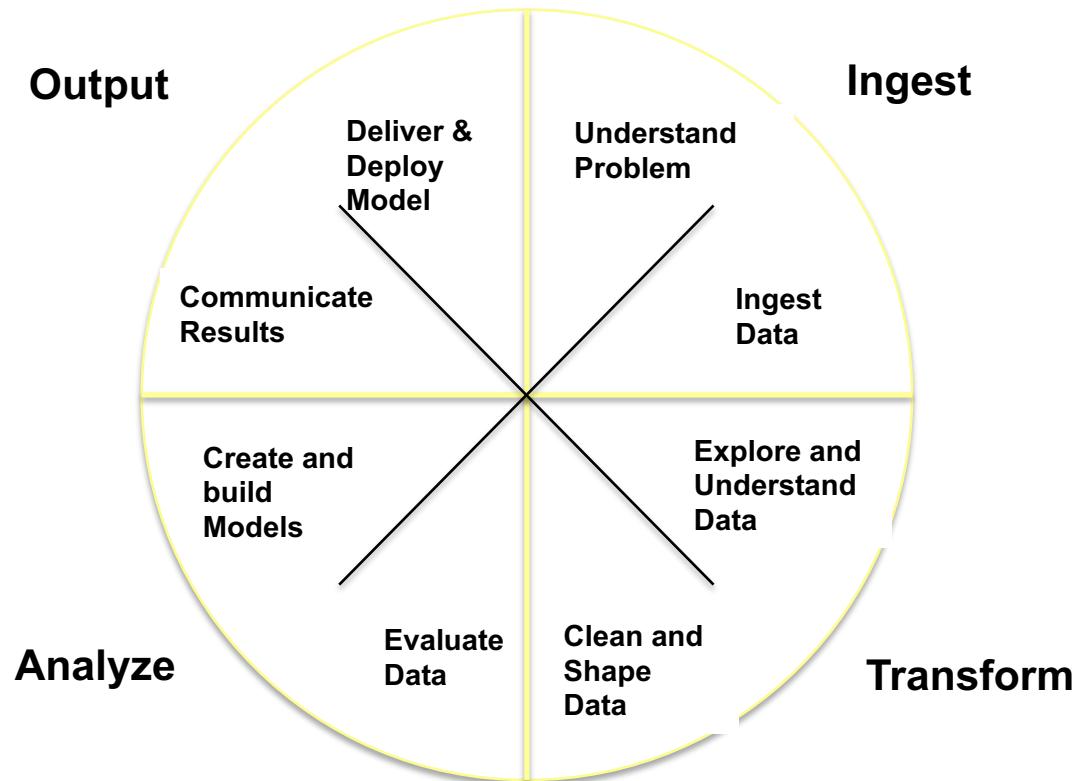
Data → Insights → Competitive advantages

# Commerce

# Virtuous Product Cycle



# Data Lifecycle





Why big data?  
Infrastructure for big data

The background of the image is a vast, fluffy landscape of white and light gray cumulus clouds against a clear blue sky. The clouds are dense and layered, creating a sense of depth. In the lower right quadrant, there are darker, more solid-looking clouds, possibly indicating a front or a different atmospheric layer.

# **Interlude: Cloud Computing**

# Utility Computing

- What?

- Computing resources as a metered service (“pay as you go”)
- Ability to dynamically provision virtual machines

- Why?

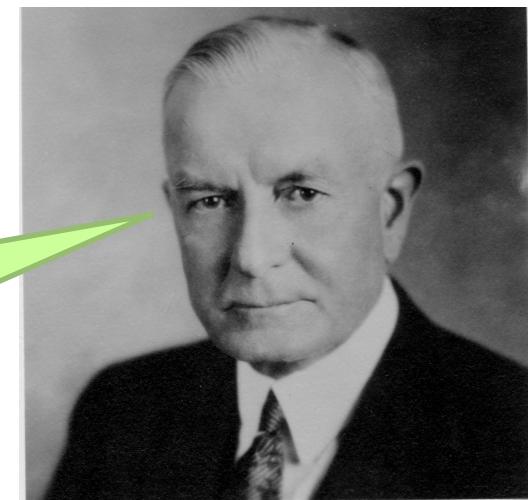
- Cost: capital vs. operating expenses
- Scalability: “infinite” capacity
- Elasticity: scale up or down on demand

- Does it make sense?

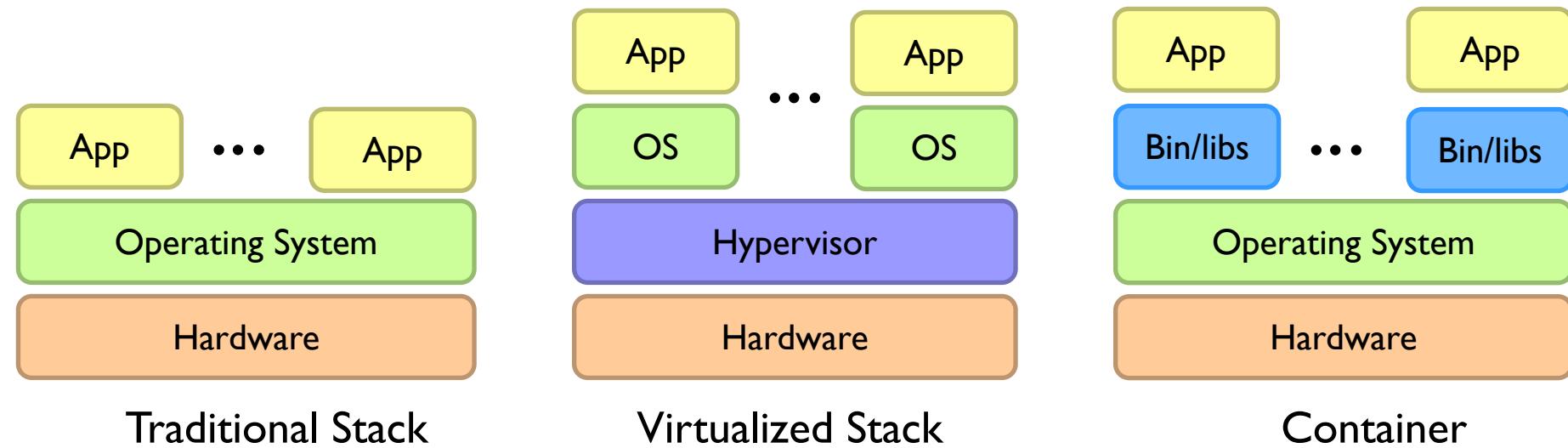
- Benefits to cloud users
- Business case for cloud providers

I think there is a world market for about five computers.

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

An aerial photograph of a large datacenter complex during sunset. The sky is a vibrant orange and yellow. In the foreground, there's a large building with a white roof, several smaller buildings, and a parking lot filled with white shipping containers. A road or highway runs through the middle ground. The background shows a vast, green, agricultural landscape stretching to a distant horizon under a hazy sky.

# The datacenter *is* the computer!

ORGAN & CLAYPOOL PUBLISHERS

## Datacenter Computer

*Introduction to the Design  
of Large-Scale Machines*  
Second Edition

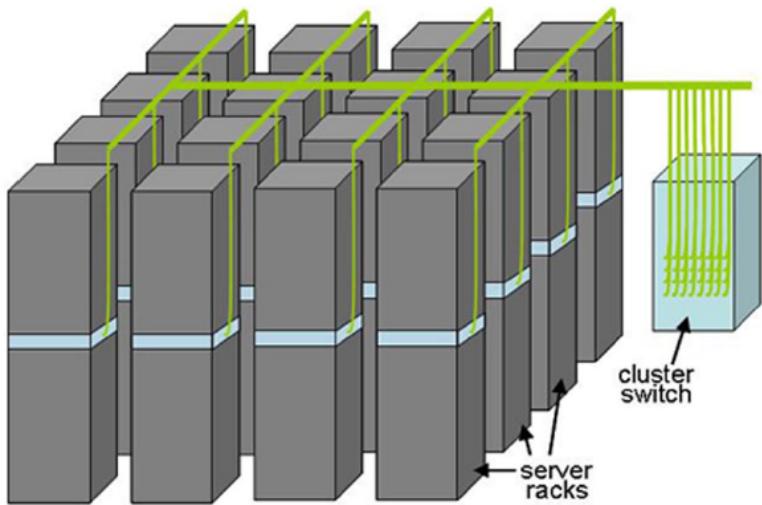
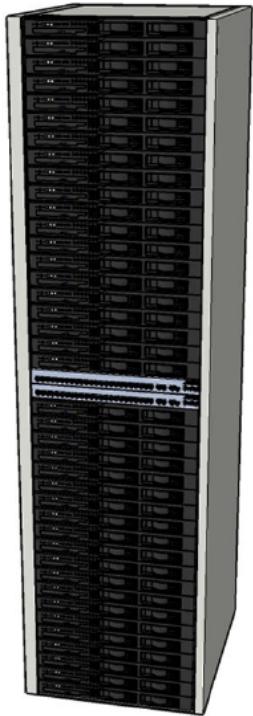
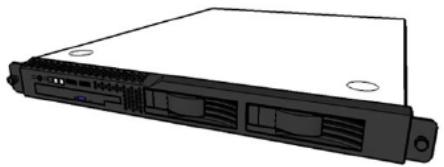
José Barroso  
Ricardo  
de Souza

LECTURES ON  
DATA CENTER  
ARCHITECTURE

Series Editor

Source: Google

# Building Blocks

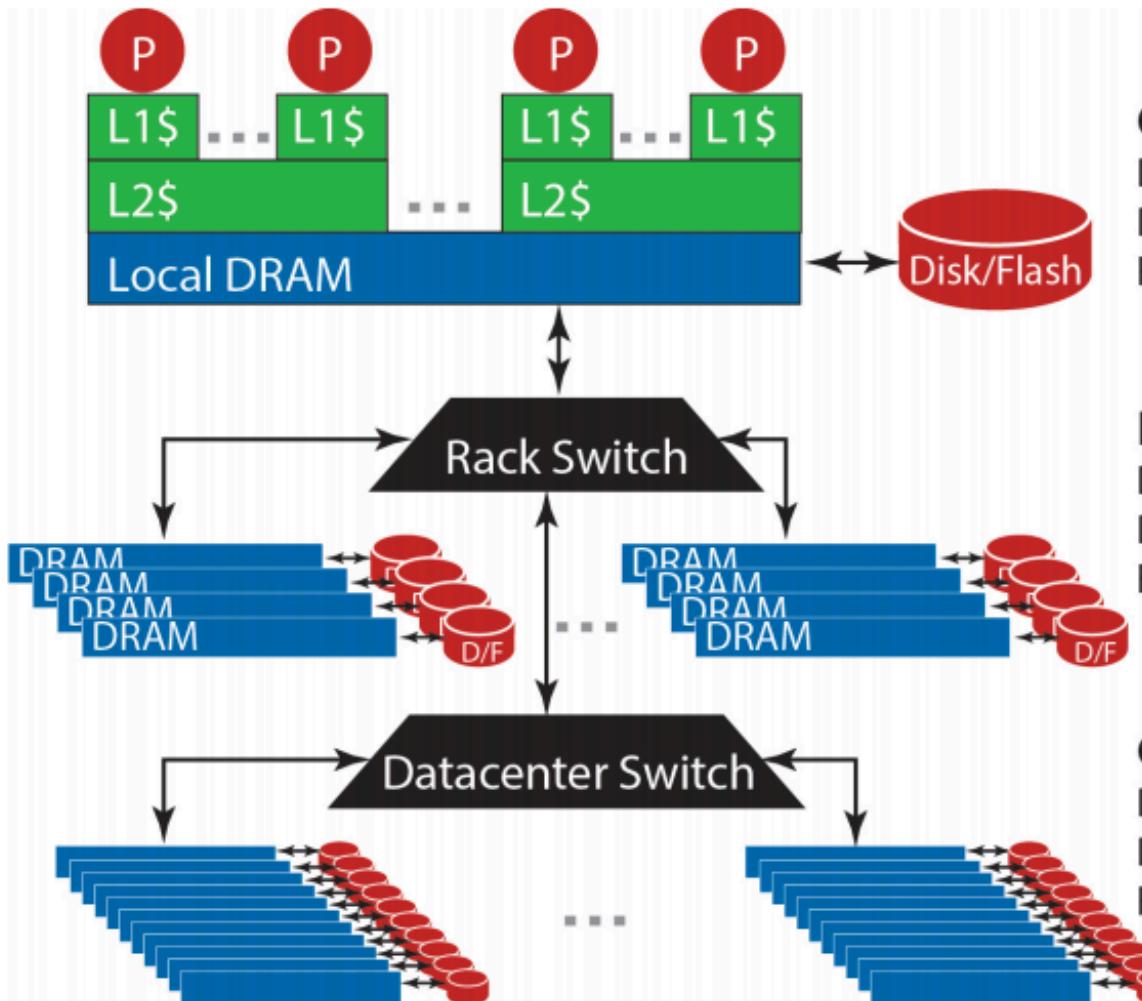








# Storage Hierarchy



## One Server

DRAM: 16 GB, 100 ns, 20 GB/s  
Disk: 2TB, 10 ms, 200 MB/s  
Flash: 128 GB, 100 us, 1 GB/s

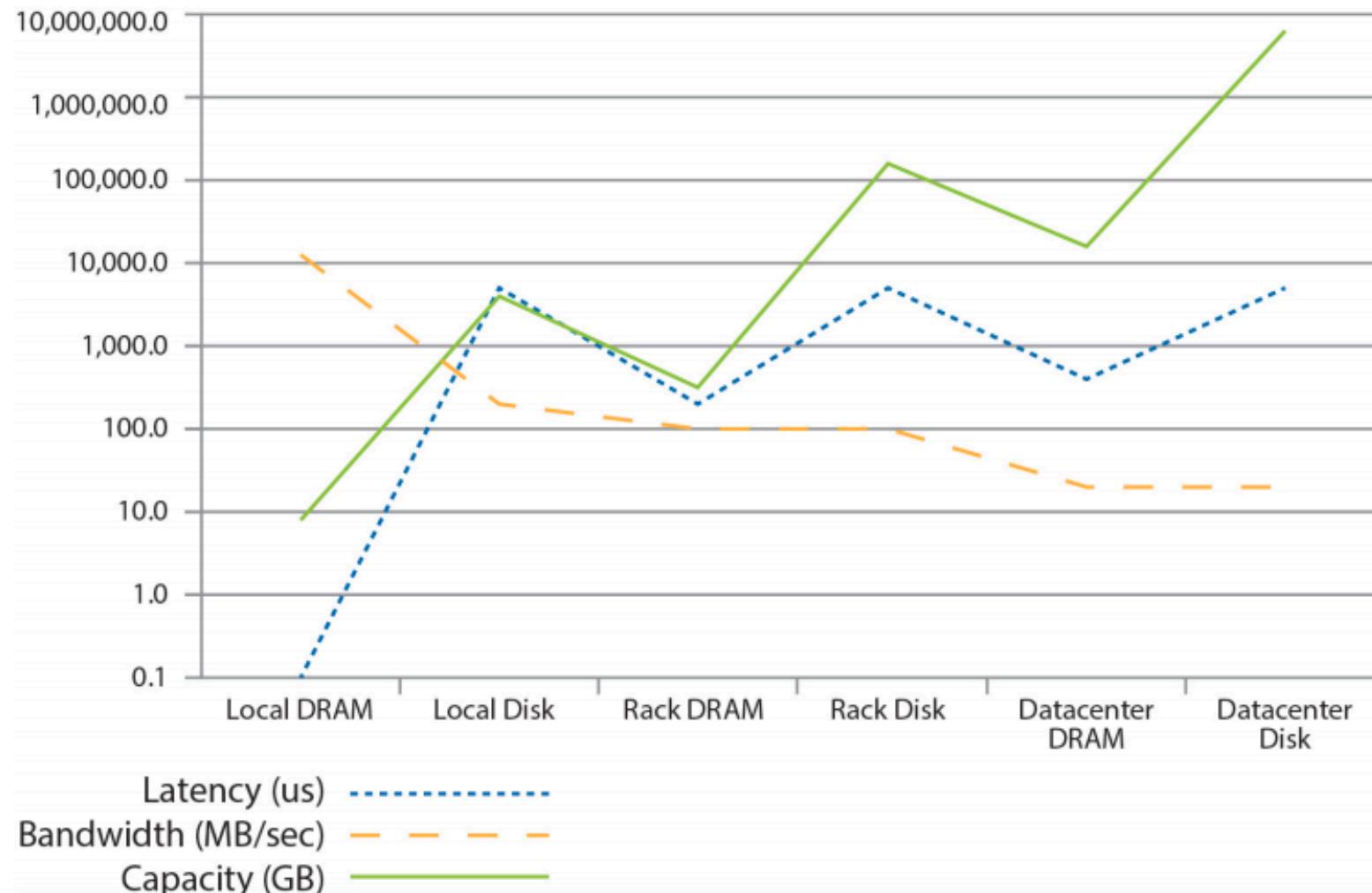
## Local Rack (80 servers)

DRAM: 1 TB, 300 us, 100 MB/s  
Disk: 160 TB, 11 ms, 100 MB/s  
Flash: 20 TB, 400 us, 100 MB/s

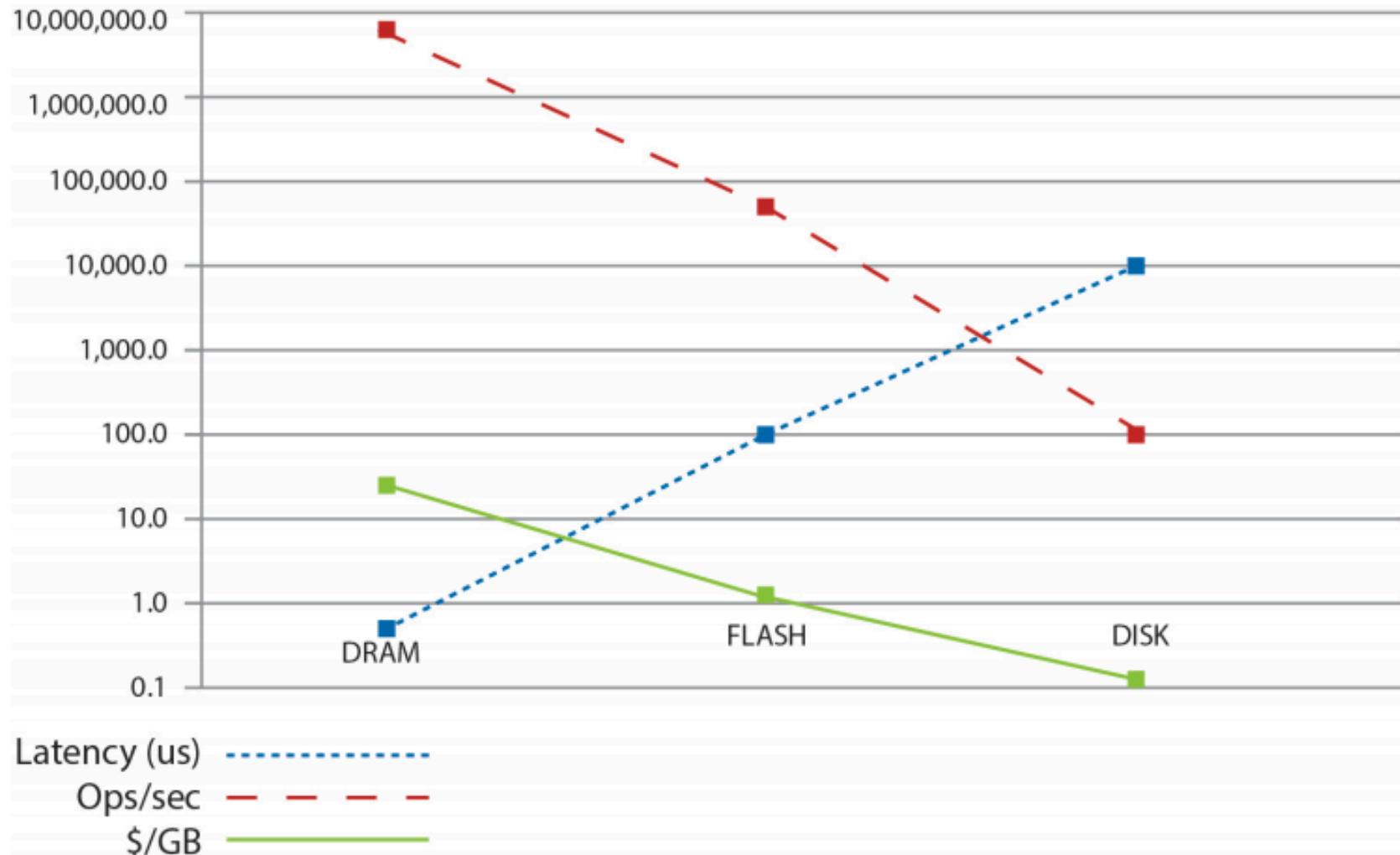
## Cluster (30 racks)

DRAM: 30 TB, 500 us, 10 MB/s  
Disk: 4.80 PB, 12 ms, 10 MB/s  
Flash: 600 TB, 600 us, 10 MB/s

# The Cost of Moving Data Around Data Center



# The Cost of Moving Data Within a Server



# “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
  - Seek operations 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 I. 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](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.

CS4225/CS5425

Big Data Systems for  
Data Science

[2110] 2021/2022 Semester 1

Owner

Conferencing

Consultation

Files

Forum

Gradebook

Multimedia

Poll

Quiz

Survey

SCORM

Web Lectures



For any suggestions or feedback

1. Suggestion/feedback here:

Enter your answer here

Finish Survey

Save For Later

# 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

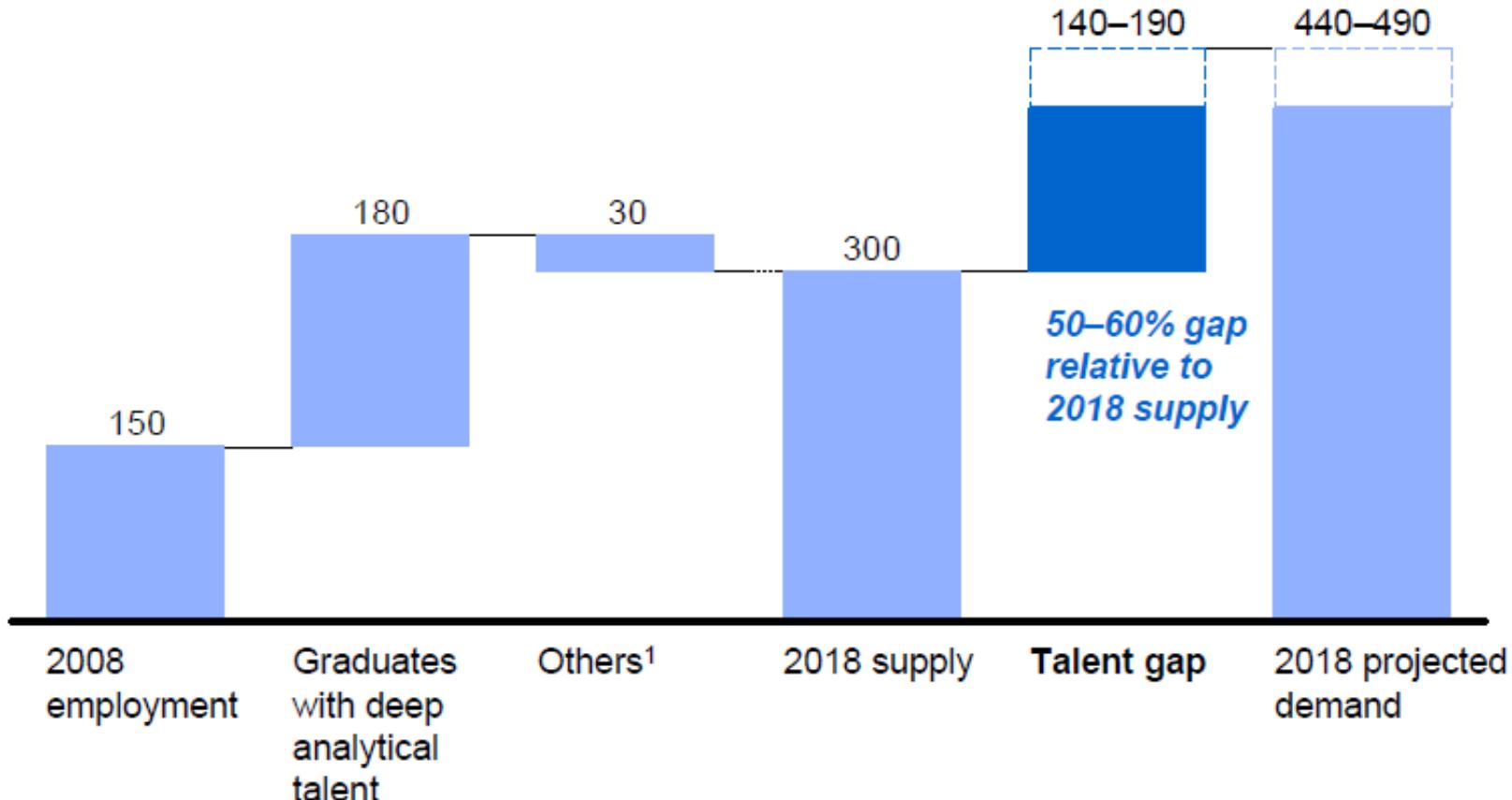
- Healthcare: Every public health cluster (Alexandra Health group, Eastern health alliance, National Healthcare Group, National University Health System, Jurong Health) in Singapore has a data team. Healthcare analytics includes traditional fields like medical research, as well as recent developments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

# 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



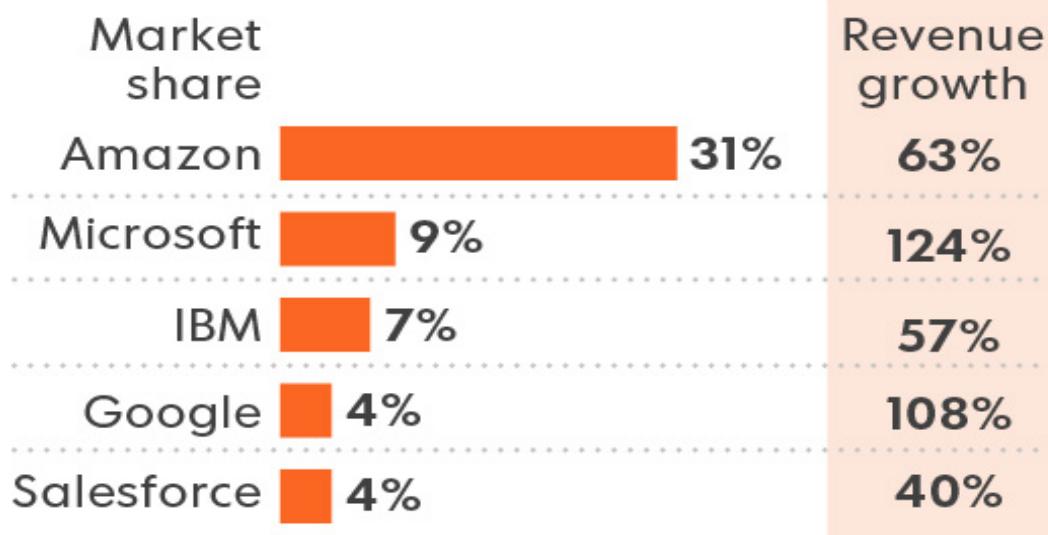
<sup>1</sup> Other supply drivers include attrition (-), immigration (+), and reemploying previously unemployed deep analytical talent (+).

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

# Growth of Cloud Computing

## CLOUD INFRASTRUCTURE VALUE

Worldwide market share and year-over-year revenue growth for cloud infrastructure services for 2015:



Combination of IaaS, PaaS, private and hybrid

SOURCE: Synergy Research Group, February 2016

George Petras, USA TODAY

