Lancaster University

SCC.460 Data Science Fundamentals

Big Data Technologies

Big Data

- 4 v's
  - Volume (sheer scale)
  - Variety (of forms and formats)
  - Velocity (relentless streams)
  - Veracity (uncertainty)
- Take away point: data are produced at increasingly rapid rates.
  - Challenges raised? How to deal with them? Opportunities?

Big Data

- Big data would not have been big without accompanying technologies.
- Main enabler…
  - Cloud computing
- Other key technologies
  - e.g. Visualization (more on this in Lecture 5)
    - Immersive VR
      - Patrick Millais, et al. 2018. **Exploring Data in Virtual Reality: Comparisons with 2D Data Visualizations**. In CHI 2018. DOI:https://doi.org/10.1145/3170427.3188537
    - Physical visualisation approaches
      - Taher, et al. 2015. **Exploring Interactions with Physically Dynamic Bar Charts**". In CHI 2015. DOI: https://doi.org/10.1145/2702123.2702604

Enablers of Big Data

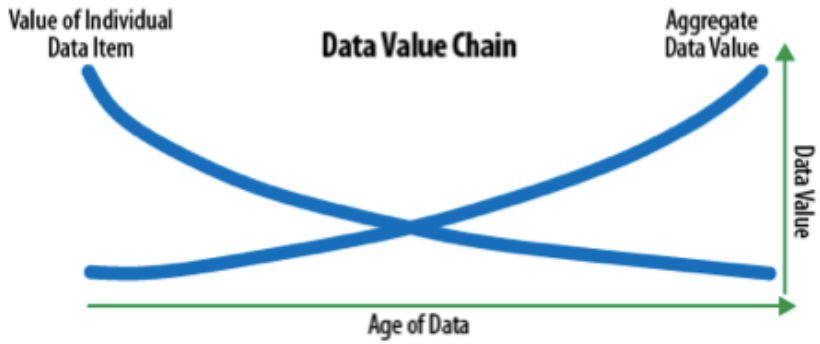
- Data science has always been there.
- What's new: technological advances to work with data in better ways.
  - Collect:
    - Extremely wide participation radius (web)
    - Multimodal data sources (smartphones, IoT)
    - Rich graph dynamics (online social networks)
  - Analyse:
    - Elastic computation, PAYG resources (cloud computing)
    - Solutions to do the heavy lifting
      - e.g. streaming frameworks that can deal with the velocity of data
        - » such as Apache Storm

Streaming Frameworks

- Apache Storm
  - Version 2.2.0 released June 2020:
  - Capitalises on cheap RAM memory (Low latency)
    - Enabling real-time analytics
    - Creating insights in extremely short time (milliseconds)
  - http://storm.apache.org/talksAndVideos.html
  - Constrast with Map Reduce that utilises multiple distributed (potentially cheap) machines with slow writing to disk
    - More on this (and the Hadoop Distributed File System) in SCC.411

Data Value

- Data has the greatest value as it enters the pipeline.
- Realtime analytics can power business decisions.
  - customer engagement
  - fraud prevention
  - resource utilisation optimization



Interactive	Realtime Analytics	Record Lookup	Historical Analytics	Exploratory Analytics
Milliseconds	Hundredths of seconds	Second(s)	Minutes	Hours
• Place trade • Serve ad • Enrich stream • Examine packet • Approve trans.	• Calculate risk • Leaderboard • Aggregate • Count	• Retrieve click stream • Show orders	• Backtest algo • BI • Daily reports	• Algo discovery • Log analysis • Fraud pattern match

What is cloud computing?

Lancaster University

"Cloud computing is a model for enabling **convenient**, **on-demand** network access to a shared pool of **configurable** computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with **minimal management** effort or service provider interaction."

Mell and Grance. "*The NIST Definition of Cloud Computing*". NIST Special Publication 800-14. September 2011. National Institute of Standards and Technology, U.S. Department of Commerce.

- Other definitions exist!
- Typical example: Online Greeting Cards Retailer.

Cloud Computing

- There has been a lot of change in application platforms over the years.
  - Mainframes
  - Stand alone (PC)
  - Client-Server
  - Super computers
  - Distributed (P2P, grid)
- Cloud computing is yet another step on the quest for computing holy grail.

"If computers of the kind I have advocated become the computers of the future, then computing may someday be organized as a public utility just as the telephone system is a public utility... The computer utility could become the basis of a new and important industry."

-- John McCarthy, speaking at the MIT Centennial in 1961

Every cloud…

- Clouds come in different forms:
  - Provisioning level
  - Deployment model

- Different costs attached to each.

Provisioning Levels

	What is Provided	Usage	Example
SaaS	Turn-key Application	Application transactions: run jobs, manipulate data, etc.	Microsoft Office Live, Google Docs, YouTube
PaaS	Runtime environment: software packages and storage support	Develop code/applications	Google AppEngine, Microsoft Azure
IaaS	Barebones hardware resources (CPU, memory, disk, networking) + OS	Customise runtime environments	Amazon EC2, Google Compute Engine

Provisioning Levels

- How do I decide?
  - Identify your skillset; real core competency.
  - How much want to / can you spend (🕑/$£) on each layer?
  - How much flexibility can I live with (or without)?
  - Other questions: e.g. legal + privacy concerns.

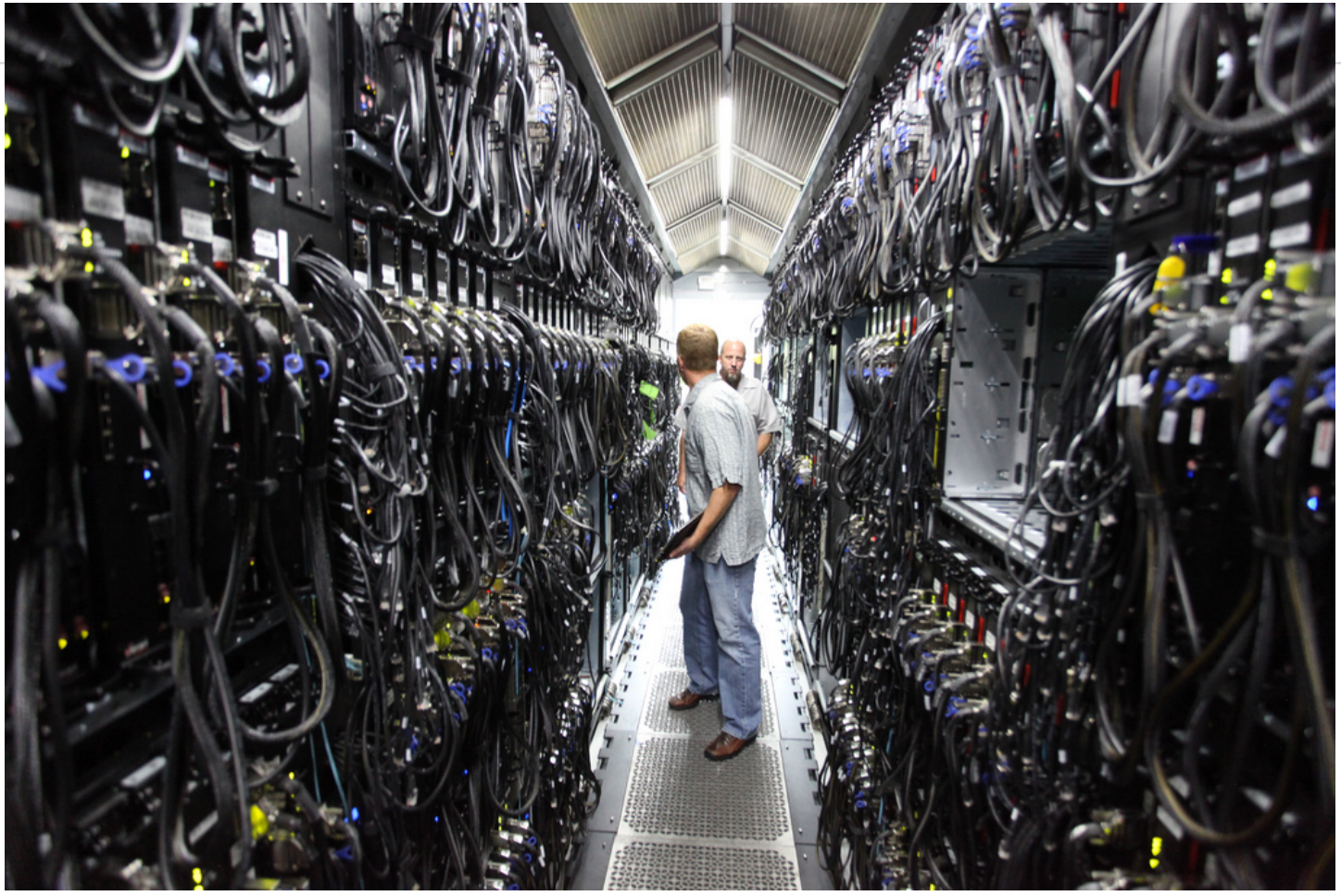
- Applies to both individuals and organisations.

Deployment Models

- Public clouds
  - Third party service providers offer services to the general public.
- Private clouds
  - Organizations (or possibly third parties) set up and maintain cloud services for their own internal use.
- Hybrid clouds
  - A mix of both strategies.

Public Deployments

Lancaster University



Example Public Deployments

Provider	SaaS	PaaS	IaaS
Google	Google Apps	App Engine	Compute Engine
Amazon	Prime Video	S3	EC2

- Others also from Microsoft, IBM, etc.

What deployment model?

	Upfront Cost	Time to Build	Security Risk
Public	Low	Low	High
Private Outsourced	High	Medium	Low
Private Internally Managed	High	High	Medium