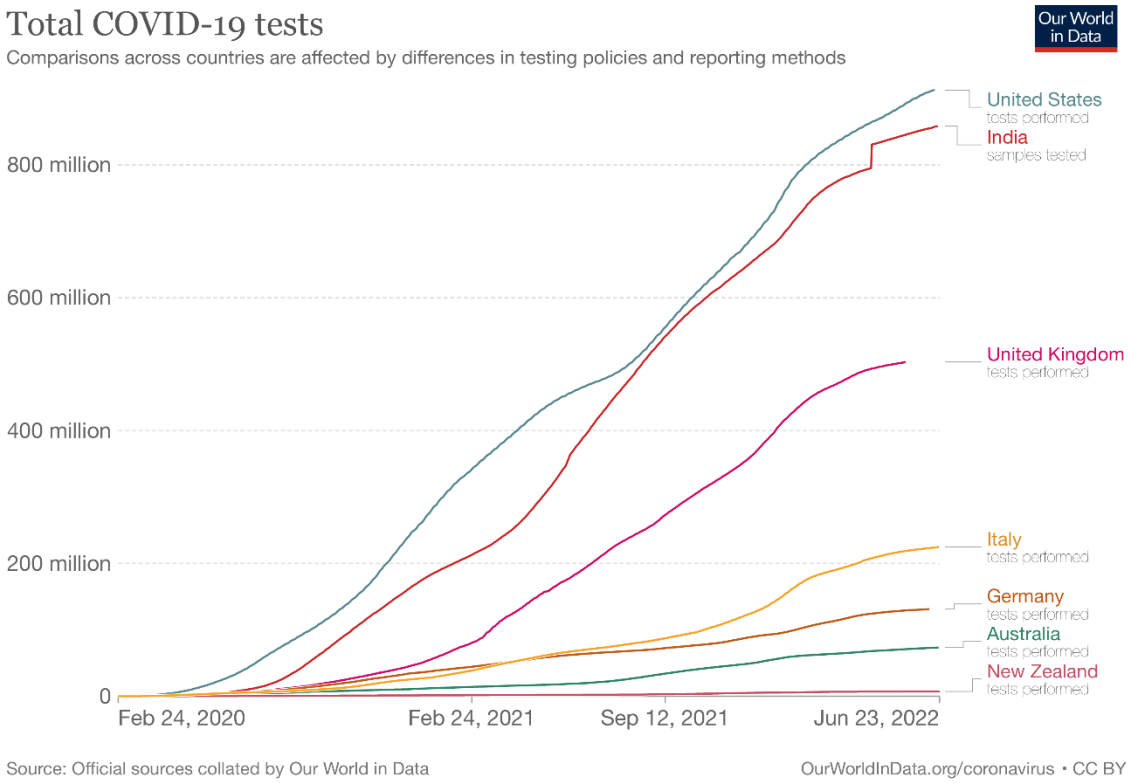
Big Data and Ethics

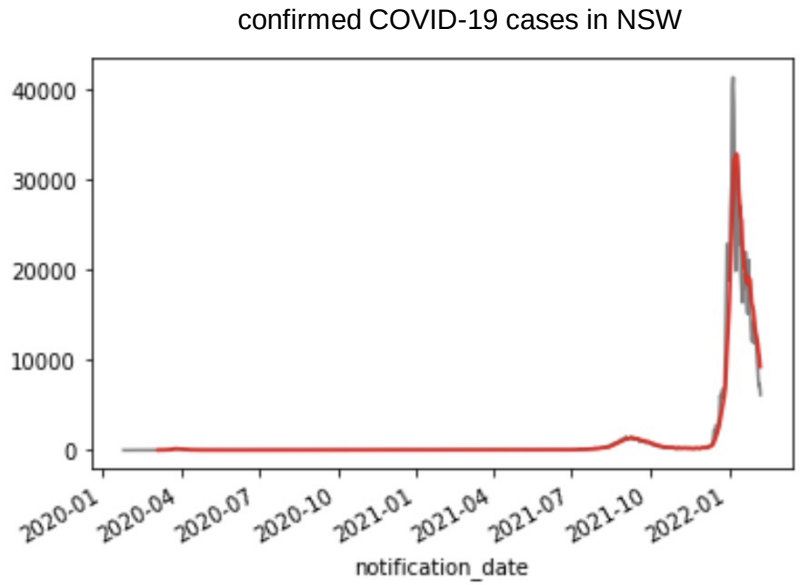
COVID-19 Epidemic

- The first known case was identified in Wuhan, China, in December 2019
- 25 Jan 2020: first confirmed case of a SARS-CoV-2 infection in Australia
- 30 Jan 2020: WHO declared spread of COVID-19 as a cause of concern.
- Worldwide, COVID-19 test and isolation protocols are introduced
  - All this test data is closely monitored to inform health regulations (besides other health statistics such as hospitalizations, ICU cases …)
  - in many countries, data is public available



How to Analyse Data on such a Scale?

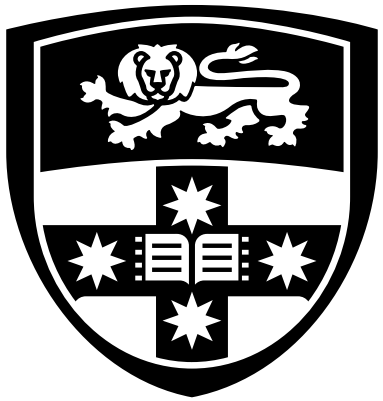
- NSW government published testing data on data.nsw.gov.au
- March – April 2020: First Wave of COVID-19 in Australia
  - Back then less than 1000 cases in NSW
- Published data stayed around 400 kB for over a year
- Then the Omicron variant arrived…
  - In Dec 2021/Jan 2022, data sizes doubled weekly
  - Now we not only had a lot of data (multiple MB),
    but it also arrived very fast (40k new daily cases at peak),
    and 200+k COVID-19 tests each day during peak



- At this scale, spreadsheet software is of no use
- Even with Python/Jupyter scripts, processing takes too long - and needs **lots** of memory
- **So what can we do?**

Big Data Analytics

- This is an example of a so-called Big Data problem
- Big Data is a challenge for the data engineering side of Data Science
  - We will look at some approaches on how to solve these challenges
- Big Data is also a challenge for data privacy, security, and ethics
  - Who has access to this data?
  - Where is data processed, where stored?
  - How can we make sure that our models do not discriminate or entrench human bias?



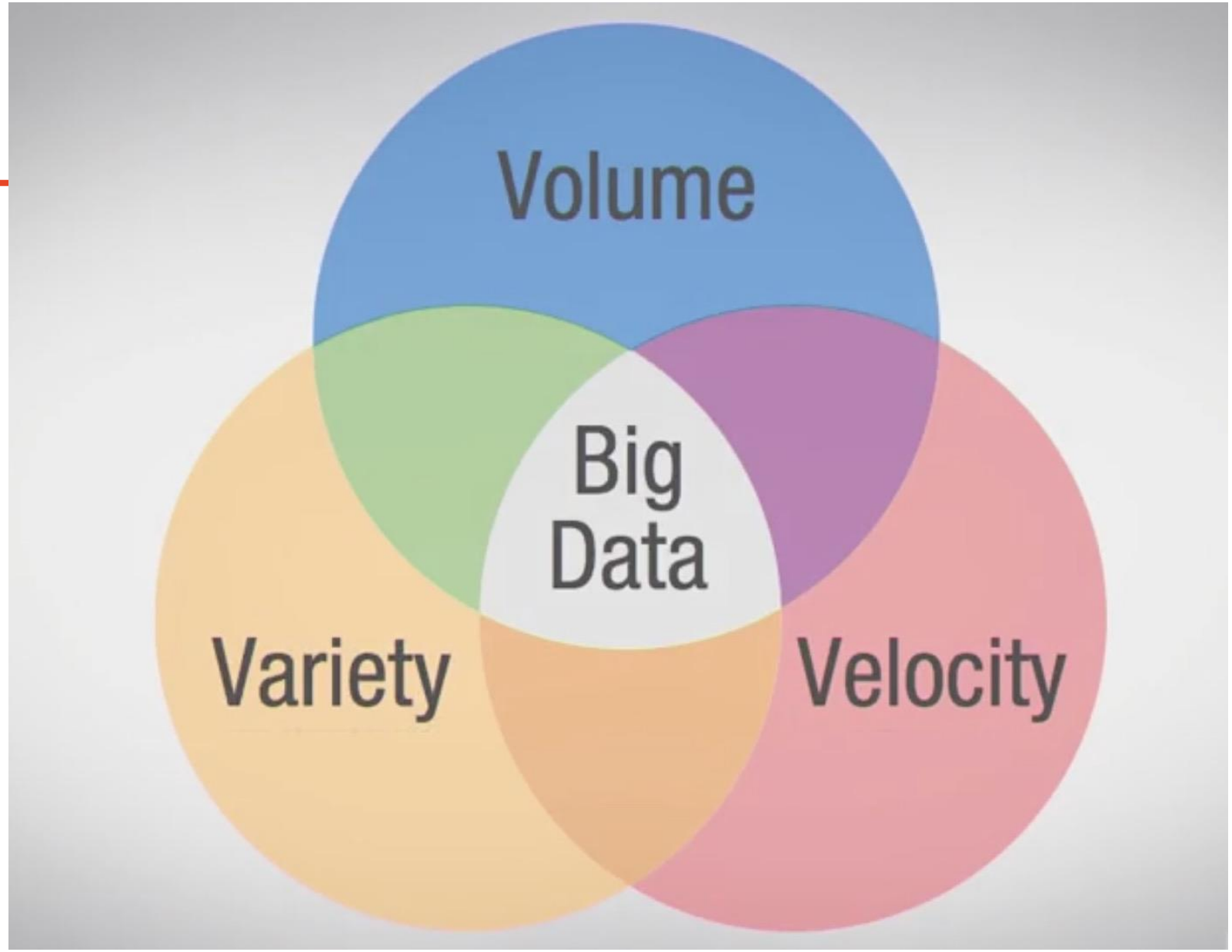
THE UNIVERSITY OF
SYDNEY

Big Data

Big Data

the three Vs:

[cf. article by
Doug Laney, 2001]



[Barton Poulson "Techniques and Concepts of Big Data", Lynda.com 2014]

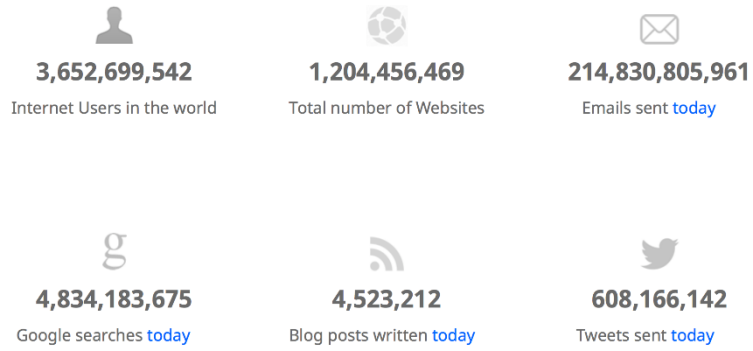
Big Data: Volume

- very relative due to Moore's Law
  - What once was considered big data, is considered a main-memory problem nowadays
  - eg. Excel: In 2003 max 65000 rows, now max 1 million rows, still ...
- Nowadays: Terabyte to Exabyte

Big Data: Velocity

- conventional scientific research:
  - months to gather data from 100s cases, weeks to analyze the data and years to publish.
  - Example: Iris flower data set by Edgar Anderson and Ronal Fisher from 1936

- on the other end of the scale: Twitter
  - average 6000 tweets/sec, 500 million per day or 200 billion per year
  - Cf. life Twitter Usage Statistics http://www.internetlivestats.com/twitter-statistics/



Big Data: Variety

- Structured Data, such as CSV or RDBMS
- Semi-structured Data, such as JSON or XML
- Unstructured Data, ie. text, e-mails, images, video
  - an estimated 80% of enterprise data is unstructured

- study by Forester Research: **variety biggest challenge in Big Data**

Big Data Examples: Big Data for Consumers

- Siri, Yelp!, Spotify, Amazon, Netflix, Google Now
- Some Big Data Variety examples:
  - "Neighborland" App [https://neighborland.com]
  - "WalkScore.com" [https://www.walkscore.com]

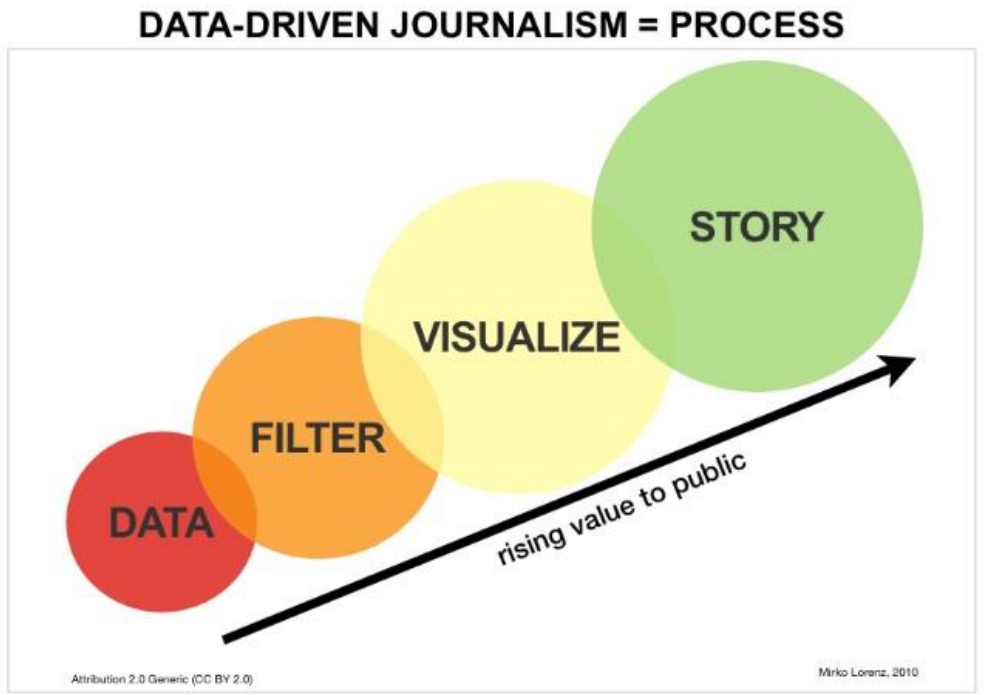
Big Data Examples: Big Data for Business

- Google Ads Searches
- Predictive Marketing
  - Example "EDITED.com": predicting fashion trends
- Fraud Detection

Big Data Examples: Data-Driven Journalism

- TimesMachine:
  New York Times archive on AWS
  - 405,000 images, 3.3 million articles in SGML and 405,000 xml article OCRs

- Recent Example (April 2016):
  Panama Papers Leak
  - leaked documents from Mossack Fonseca, a Panamanian law firm that sells anonymous offshore companies around the world
  - 2.6TB in 11.5 million documents, 214,000 companies
  - http://panamapapers.sueddeutsche.de/en/



[Source: Wikipedia]

Big Data Examples: Big Data for Research

- Astronomy: Sloan Digital Sky Survey (SDDS) SkyServer
- Cern's Large Hadron Collider (LHC)
- The Human Brain Project
- Google Flu trends (only historic data; stopped publishing new trends)
- Apple COVID19 Mobility Trends (discontinued April 2022)
- Google Books project
  - (eg. changes of word usage over time (eg. maths vs arithmetic vs algebra) https://books.google.com/ngrams/graph?content=math,arithmetic,algebra&case_insensitive=off&year_start=1800)

Sources of Big Data / More Vs

- Human-generated Big Data
  - E.g. photos, posts, likes etc
- Machine-generated data
  - Communication logs, Internet-of-Things, etc.

- More Vs of Big Data:
  - Validity (data quality), Variability (data consistency), Veracity (data accuracy / trustworthiness), Value…

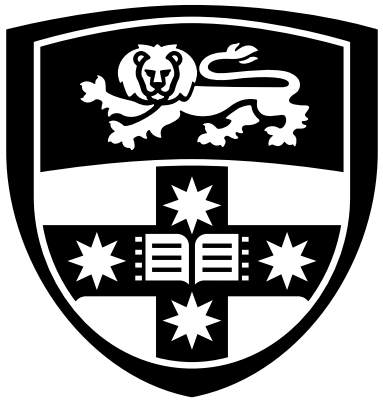
Big Data Challenges beyond Technical Aspects

"[…] consider that great responsibility follows inseparably from great power" [French National Convention,1793]

- **Data Privacy**
  - Some data sources, such as "Internet-of-Things", allow tracking anyone
    - Do you really need to know *who* was travelling a route in order to predict, e.g., traffic densities?
    - Personal data can be inferred sometimes => New York Taxi data set example
  - Privacy laws
    - Always check: Are you allowed to use some data or process is anywhere?
    - Some personal data, especially regarding health or tax, is specially protected; e.g., not allowed to leave a jurisdictive area
    - e.g. EU's General Data Protection Regulation (GDPR) applies to any company holding data about any European Union citizen
- **Data Security**
  - Can your users trust you to keep their data safe?
  - Big data can expose your organization to serious privacy and security attacks!

Big Data Challenges beyond Technical Aspects (cont'd)

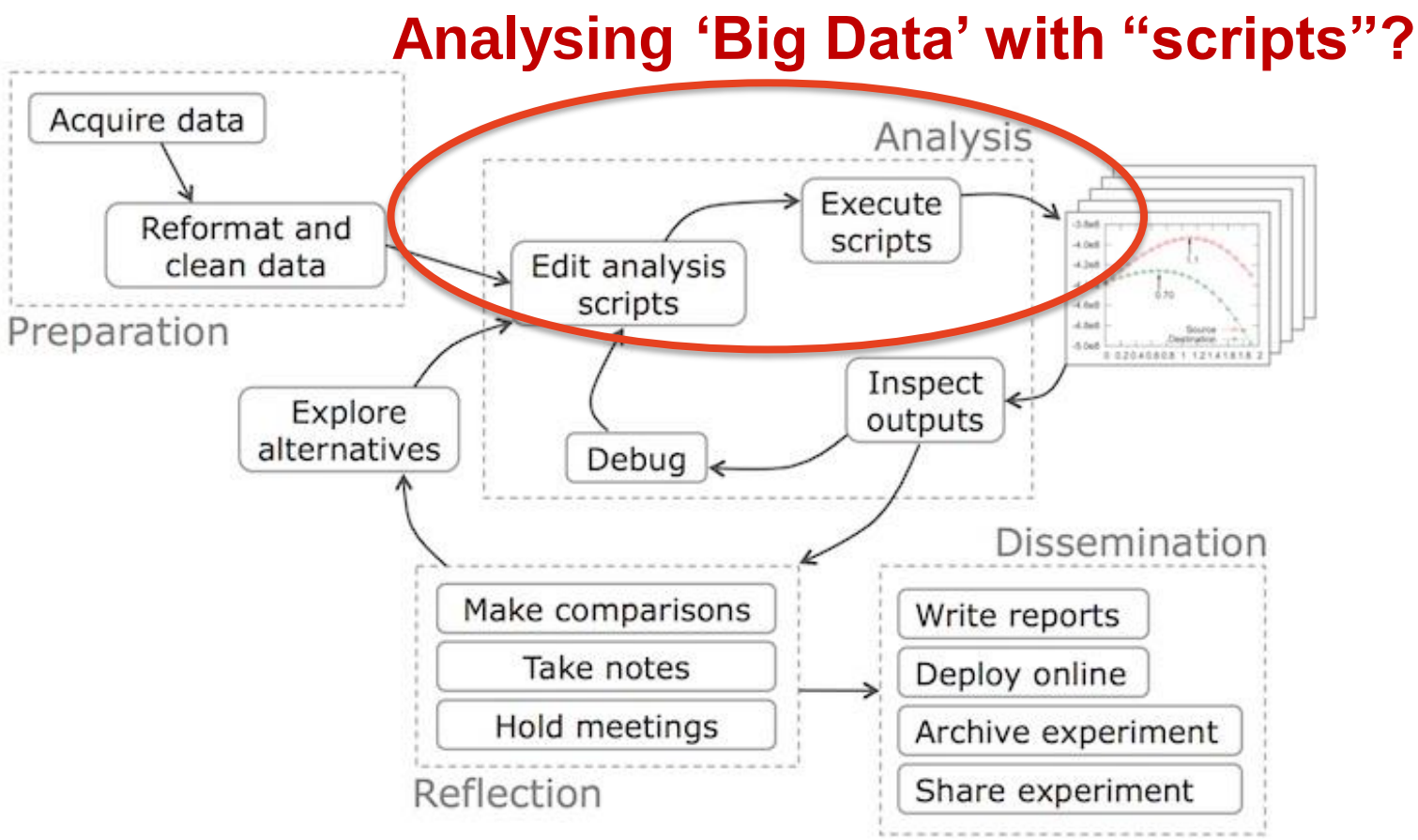
- **Data Discrimination**
  - Is it acceptable to discriminate against people based on data on their lives?
  - Credit card scoring? Health insurance?
  - Cf. FTC: "Big Data – A Tool for Inclusion or Exclusion?" [https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf]
- Check:
  - Are you working on a representative sample of users/consumers?
  - Do your algorithms prioritize fairness? Aware of the biases in the data?
  - Check your Big Data outcomes against traditionally applied statistics practices



THE UNIVERSITY OF
SYDNEY

Analysing Big Data

Data Science Workflow



[Source: http://cacm.acm.org/blogs/blog-cacm/169199-data-science-workflow-overview-and-challenges/]

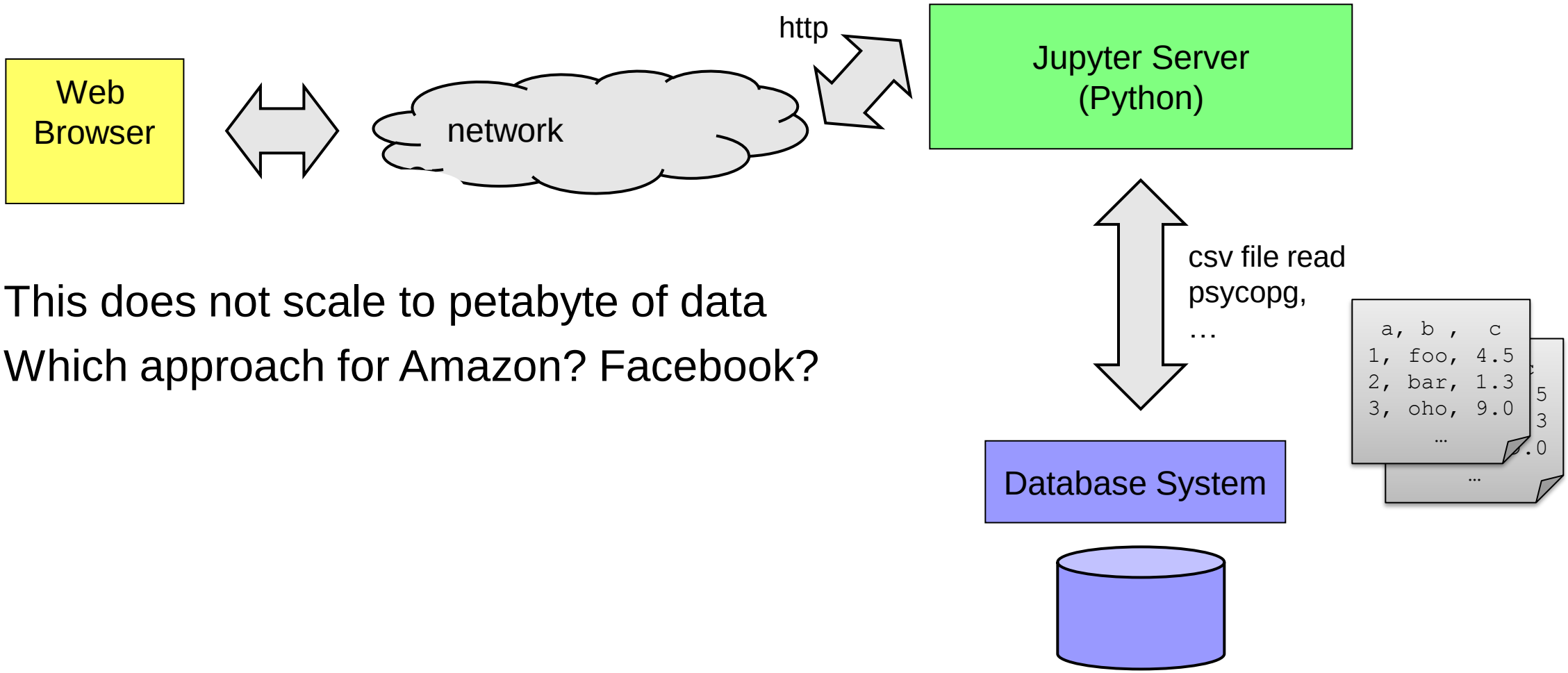
Case for Data Science Platforms

- Data is either
  - too large (volume),
  - too fast (velocity), or
  - needs to be combined from diverse sources (variety)

  for processing with scripts or on single server.

- Need for
  - scalable platform
  - processing abstractions

Jupyter Notebooks as Platform for Big Data?

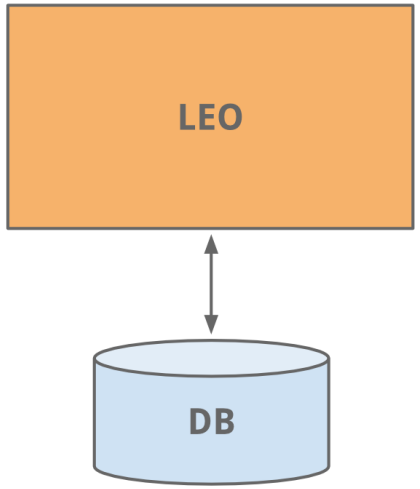


- This does not scale to petabyte of data
- Which approach for Amazon? Facebook?

Case Study: LinkedIn

Source: https://engineering.linkedin.com/architecture/brief-history-scaling-linkedin

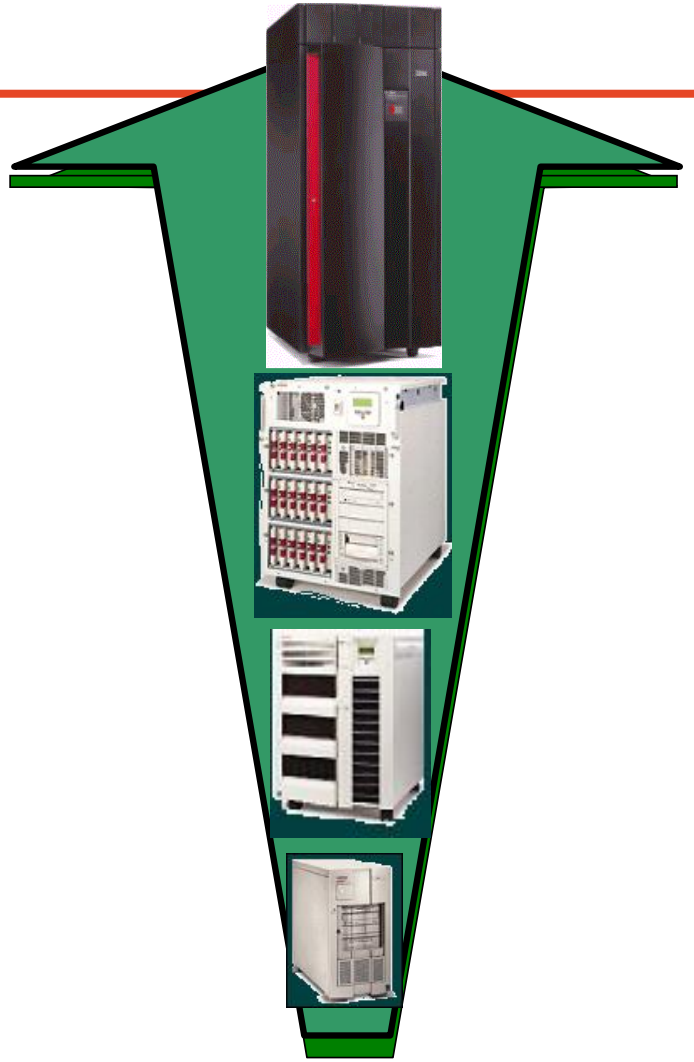
- Started in 2003
  - 2700 members in first week
  - Single database and web server



- for years experienced exponential growth…
- As of Apr 2022:
  (https://www.omnicoreagency.com/linkedin-statistics/)
  - 810 million members
  - 310 million active users / month
  - Many users with hundreds of connections => huge graph
  - Fun Fact: Statistical Analysis and Data Mining are Top skills on Linkedin

Scale-Up

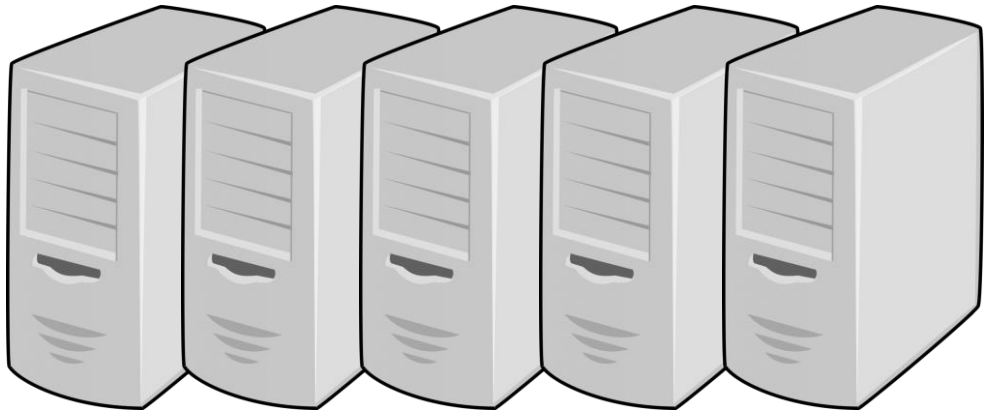
- The traditional approach:
  - To scale with increasing load, buy more powerful, larger hardware
    - from single workstation
    - to dedicated db server
    - to large massive-parallel database appliance



[source: Jim Gray, HPTS99]

The Alternative: Scale-Out

A single server has limits…
For real Big Data processing, need to
scale-out to a cluster of multiple servers (nodes):



[Source: Server.png from PinClipart.com]

State-of-the-Art:

shared-nothing architecture

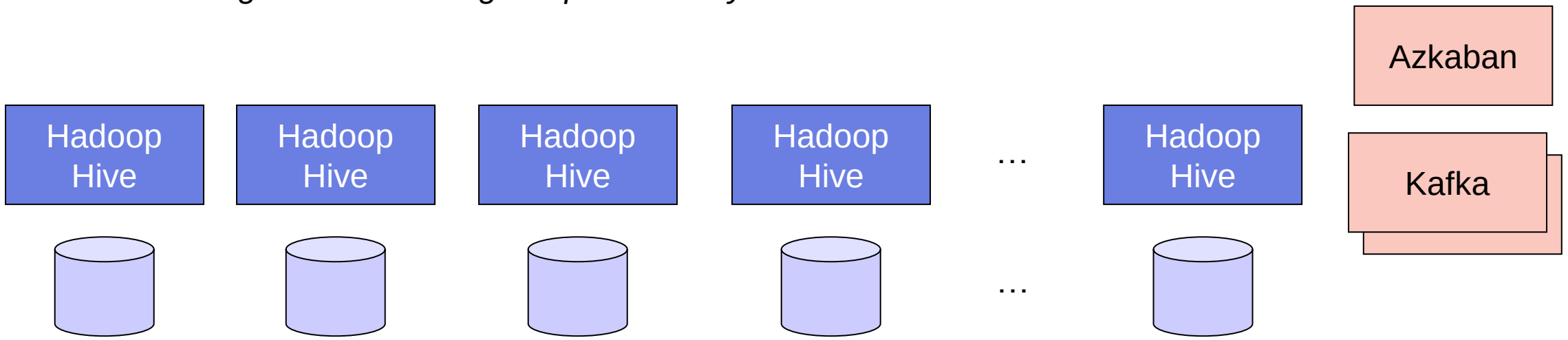
Case Study: LinkedIn Analytical Architecture

LinkedIn via https://wiki.apache.org/hadoop/PoweredBy/

”We have multiple grids divided up based upon purpose.
Hardware:

- ***~800*** *Westmere-based HP SL 170x, with 2x4 cores, 24GB RAM, 6x2TB SATA*
- *~**1900** Westmere-based SuperMicro X8DTT-H, with 2x6 cores, 24GB RAM, 6x2TB SATA*
- ***~1400*** *Sandy Bridge-based SuperMicro with 2x6 cores, 32GB RAM, 6x2TB SATA*
- …

We use these things for discovering People You May Know and other fun facts.”

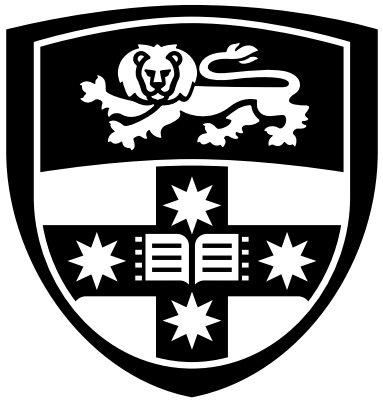


Challenges

- **Scale-Agnostic Data Management**
  - **sharding** for performance
  - **replication** for availability
  - ideally such that applications are unaware of underlying complexities

- **Scale-Agnostic Data Processing**
  - Nowadays we collect massive amounts of data; how can we analyze it?
    - Answer: use lots of machines… (hundreds/thousands of CPUs, can grow)
  - Performance: parallel processing
  - Availability: Ideally, the system never down; can handle failures transparent

  => Distributed Data Science Platforms



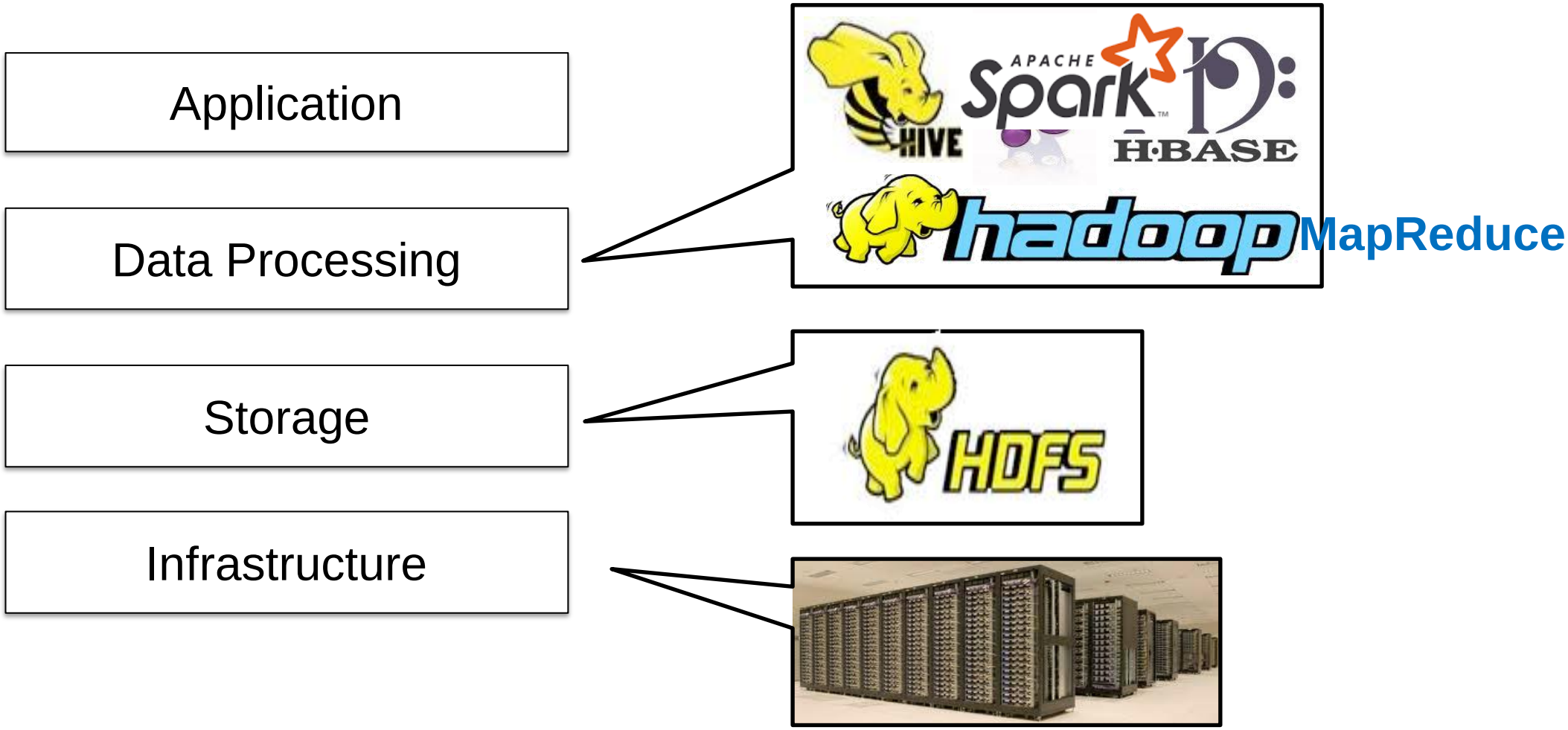
THE UNIVERSITY OF
SYDNEY

Distributed Data Science Platforms

Challenge: Big Data Processing on Scale-out Architecture

http://cdn.techsnapr.com/wp-content/uploads/2011/04/FB_SERVER_040_x900.jpg

Big Data Analytics Stack



[slide by Ion Stoica, UCB, 2013]

Distributed Data Analytics Frameworks

- **Apache Hadoop**
  - Open-source implementation of original MapReduce from Google; Apache top-level project
  - Java framework, but also provides a Python interface nowadays
  - Parts: own distributed file system (HDFS), job scheduler (YARN), MR framework (Hadoop)

- **Apache Spark**
  - Distributed cluster computing framework on top of HDFS/YARN
  - Concentrates on **main-memory** processing and more **high-level data flow control**
  - Originates from research project from UC Berkeley

- **Apache Flink**
  - Efficient data flow runtime on top of HDFS/YARN
  - Similar to Spark, but more emphasize on **build-in dataflow optimiser** and **pipelined processing**
  - Strong for data stream processing
  - Origin: Stratosphere research project by TU Berlin, Humboldt University Berlin and HPI Potsdam

Distributed Data Analytics Frameworks (continued)

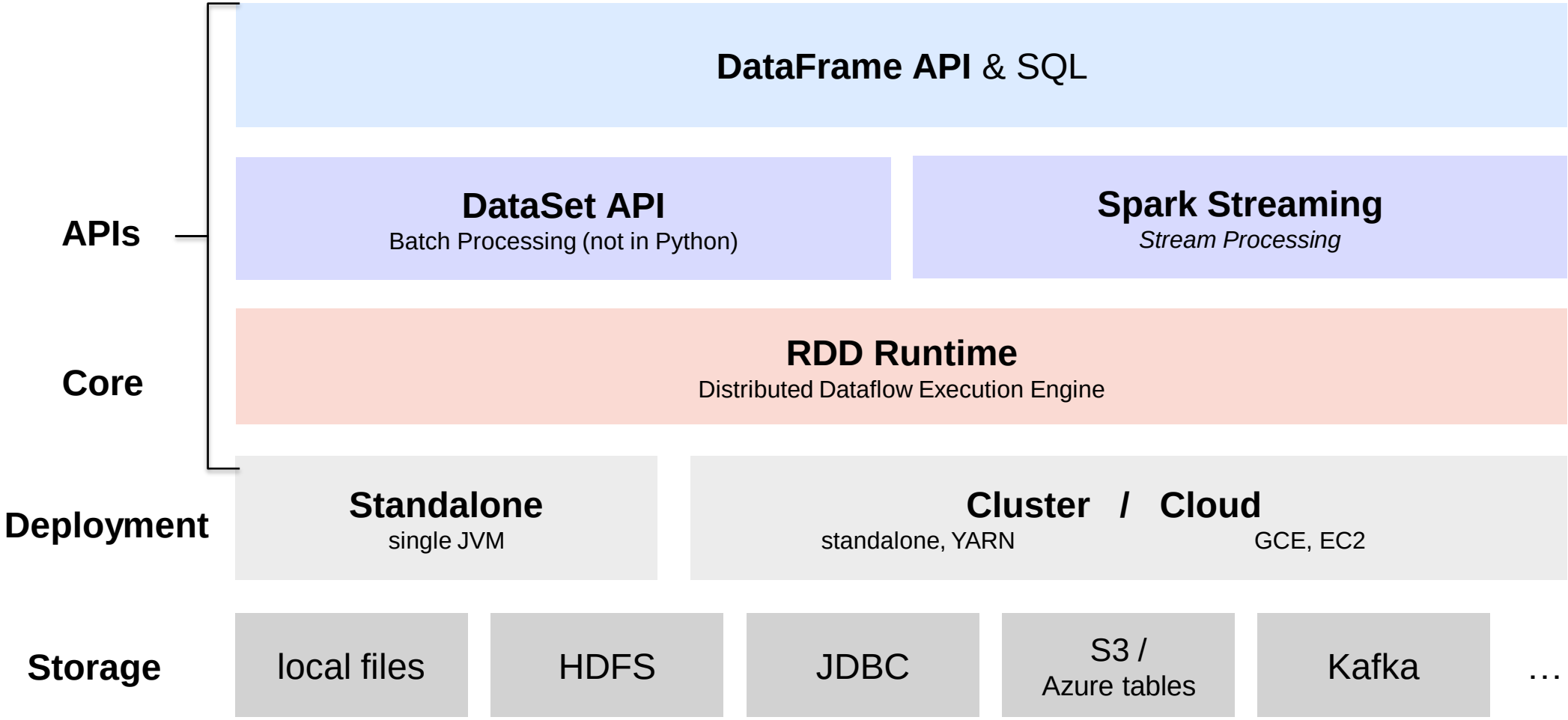
- **Apache Hive**
  - Provides an SQL-like interface on top of Hadoop / HDFS
  - Allows to define a relational schema on top of HDFS files, and to query and analyse data with HiveQL (SQL dialect)
  - Queries automatically translated to MR jobs and executed in parallel in cluster
  - Example: WordCount in HIVE

```
CREATE TABLE docs (line STRING);
LOAD DATA INPATH 'input_file' OVERWRITE INTO TABLE docs;

CREATE TABLE word_counts AS
       SELECT word, count(1) AS count
         FROM (SELECT explode(split(line, '\s')) AS word FROM docs) temp
        GROUP BY word
        ORDER BY word;
```

- **Many more high-level frameworks for advanced data analytics.**

Example Apache Spark System Stack



Example: Log Mining with Apache Spark

- Load error messages from a log into memory, then interactively search for various patterns

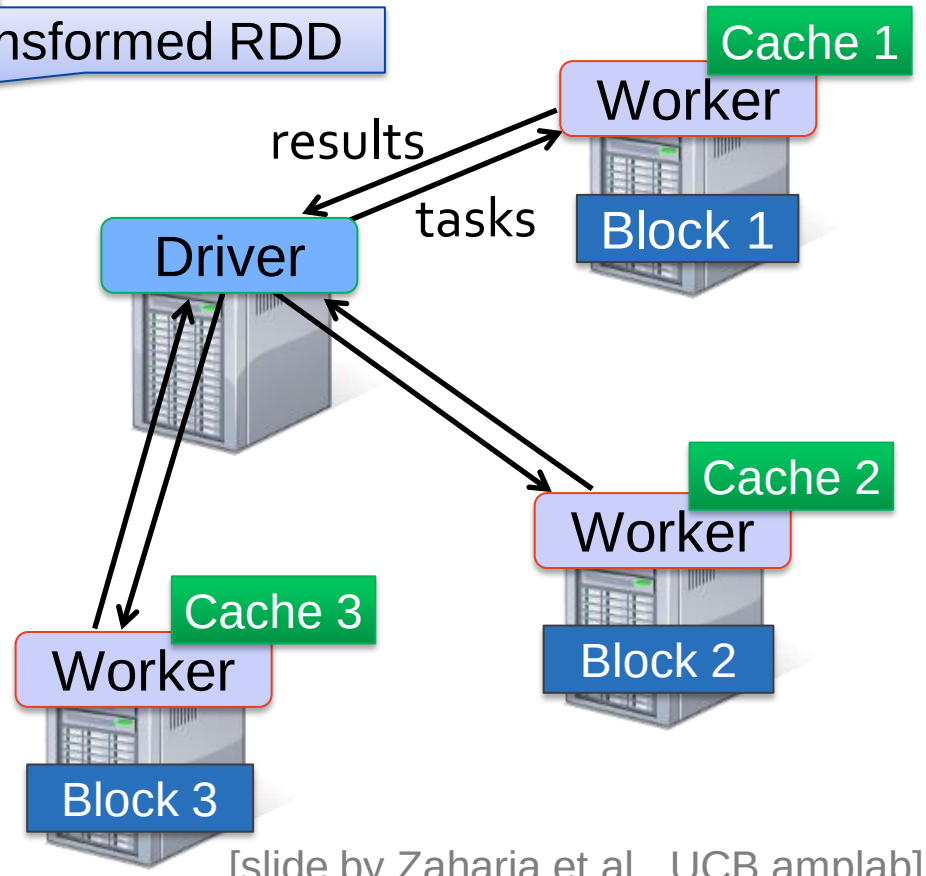
```
lines = spark.textFile(“hdfs://...”)
errors = lines.filter(_.startsWith(“ERROR”))
messages = errors.map(_.split(‘\t’)(2))
cachedMsgs = messages.cache()
cachedMsgs.filter(_.contains(“foo”)).count
cachedMsgs.filter(_.contains(“bar”)).count
. . .
```

Base RDD

Transformed RDD

Cached RDD

Parallel operation

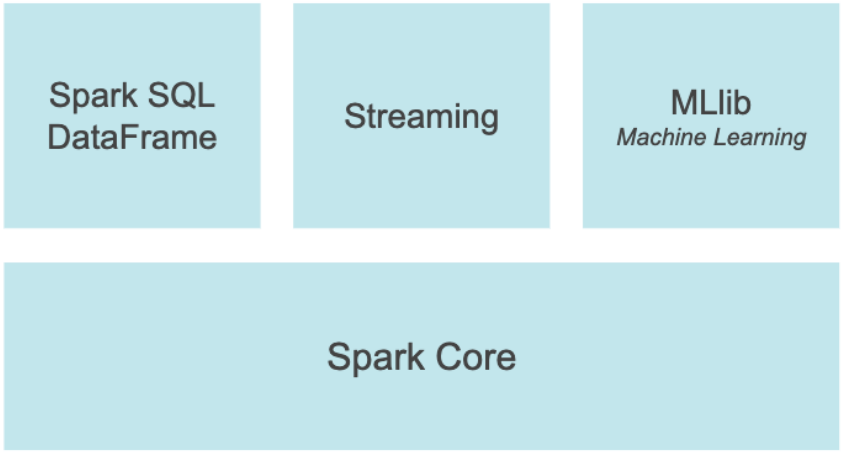


Result: full-text search of Wikipedia in <1 sec (vs 20 sec for on-disk data)

[slide by Zaharia et al., UCB amplab]

PySpark

- “PySpark allows you to write Spark applications using Python APIs, but also provides the PySpark shell for interactively analyzing data in a distributed environment. PySpark supports most of Spark’s features such as Spark SQL, DataFrame, Streaming, MLlib (Machine Learning) and Spark Core.”



https://spark.apache.org/docs/latest/api/python/index.html

Spark and Jupyer Notebooks

- PySpark can also be used from Jupyter notebooks
- Either local install of **FindSpark package**
  install findspark, start pyspark, start Jupyter

```
import findspark
findspark.init()
import pyspark.sql
spark = SparkSession.builder.appName("…")…
…
```

- Or use in a cloud platform such as **Databricks**

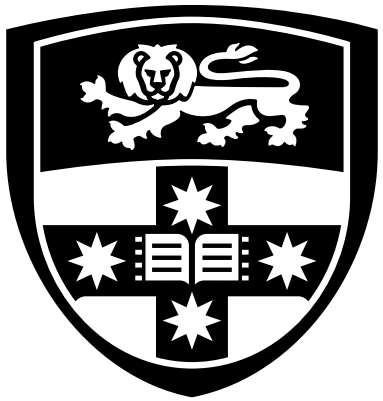
Tips and Tricks

- **Big Data** is one driver behind Data Science; definition somewhat general though
- Map/Reduce paradigm very powerful to tackle the petabyte scale problems of today's majors
  - Especially for big Internet companies such as Google, Amazon, LinkedIn, Twitter, Facebook, …
- Pros:
  - Scalability and runs on commodity hardware
- Cons:
  - Not usable for non-programmers
  - Even non-procedural programmers struggle with the functional nature of Map/Reduce
  - Not everything is petabytye scale…

- Dataflow systems such as Spark and Flink improve the usability quite a bit
  another approach available is **HIVE** as 'SQL on MapReduce'
  - But still targets petabyte scale problems.

Summary

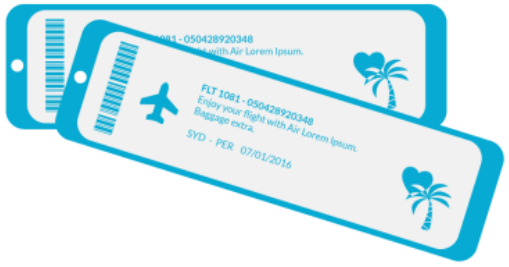
- **Big Data**
  - The three V's: Volume, Velocity and Variety
  - Ethical challenges for Big Data Processing
  - Scale-Up versus Scale-Out
- **Scale-Agnostic Data Analytics Platforms**
  - Data Scientists need more high-level tools and interfaces than MapReduce
  - Examples: **Apache Spark** or **Apache Flink** or **Apache Hive**
  - Componentized infrastructure: SQL querying, ML-Libraries, Streaming, etc.



THE UNIVERSITY OF
SYDNEY

Product Thinking

A Product is a Good or Service offered to Customers



A product can be:

- **A Good:** a tangible or virtual item that provides certain benefits for a customer.
  A delicious cup of coffee is a product.
- **A Service:** a single or package of activities performed to fulfill benefits for the customer.
  A professional haircut is a service.

https://roadmunk.com/blog/what-is-a-product/

Define the Problem before the Solution

First define the problem...

User problem: What problem do we solve?
Target audience: For whom are we doing this?
Vision: Why are we doing this?
Strategy: How are we doing this?
Goals: What do we want to achieve?

Only then does it make sense to think about the solution

Relationship of Data Scientist to Product

Model 1: Data scientist as an ***owner***

Model 2: Data scientist as a ***service***

Model 3: Data scientist as a ***partner***

Adapted from: http://fututorial.weebly.com/

Video: https://www.youtube.com/watch?v=v2MJypl51zk

Data Scientist as Owner of Product

Operates in a “hacky way” (early stage companies)

Key steps in business rely heavily on data

- Recommender
- Relevance
- Matching
- Scoring

Mostly backend, relatively stand-alone features

Adapted from: http://fututorial.weebly.com/

Video: https://www.youtube.com/watch?v=v2MJypl51zk

Data Scientist as a Service

Engagement is “on-demand”, project-based

Examples:

- some of the BI roles
- strategy roles (consultancy)
- data API for product
- modeling for specific purposes: propensity to {x}, where x: {buy, attrite, convert, etc.}

Data Scientist as a Partner

Plays active role in every stage of Product Life Cycle

Shares the ultimate goal of product success

Often requires an embedded engagement model

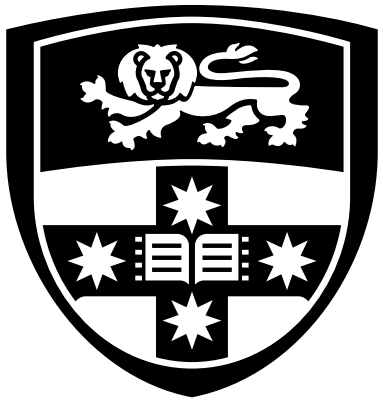
Possibilities and Impact

Data sense: What is possible?

Product sense: What is valuable?

Different roles require different balance of abilities

All data science should deliver on value proposition



THE UNIVERSITY OF
SYDNEY

The Ethical Data Scientist

What are Ethics?

“Ethics are the moral principles that govern a person's behaviour or the conducting of an activity”

http://www.oxforddictionaries.com/definition/english/ethics

Why is it a Data Scientist's Job?

Consider:

- User behaviour data forms the foundation of data products
- Products assist users but may also influence their behaviour
- e.g. ranking algorithms, recommendation systems, friend suggestions.

Models/algorithms not only predict but affect the future

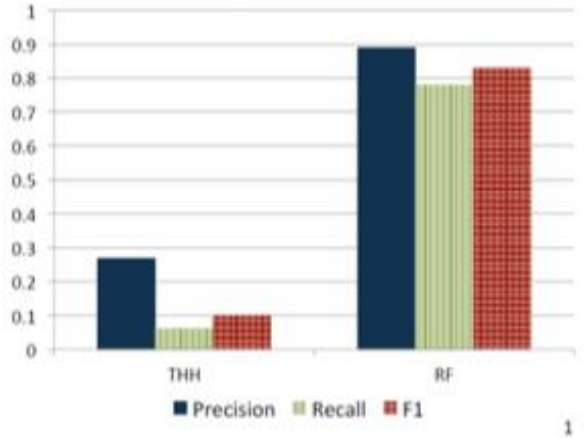
- This is both incredibly exciting and absolutely terrifying

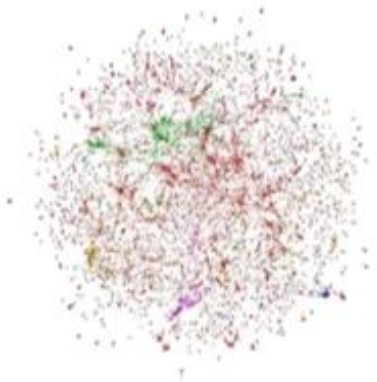
Example 1: Preventative Policing

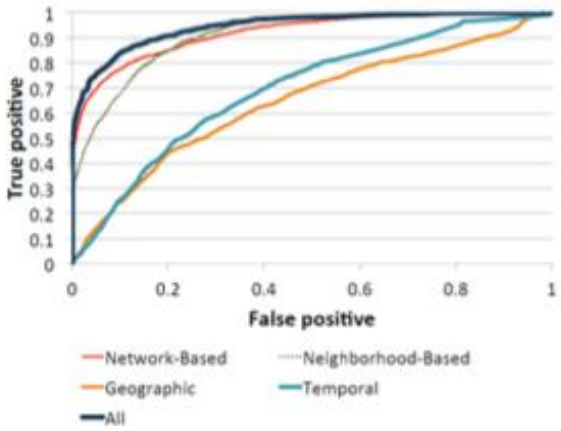
- Chicago police used predictive modelling to create a heat list
  - Given social network from arrest records, geographic, temporal data
  - Pre-emptively approach:
    Predict whether a person is likely to be involved in violent crime
- Used to issue preemptive warning:

  "We're watching you"

- What features should/not be used?







http://arxiv.org/pdf/1508.03965v1.pdf

Questions of Ethics: Preventative Policing

How avoid perpetrating potentially unfair or damaging stereotypes/profiling present in the data?

How use information positively and manage potential prediction mistakes?

Baldridge. Machine learning and human bias: and uneasy pair. http://techcrunch.com/2015/08/02/machine-learning-and-human-bias-an-uneasy-pair/

Example 2: User Tweaking

Inducing emotional states

- A 2014 study explored whether user mood is contagious on Facebook
- Manipulated feeds to include fewer positive or negative posts
- Discuss risks: How should users be protected?

Questions of Ethics: User Tweaking

This is an interesting study but emotional well-being should not be treated lightly.

How avoid distress and support users who may be negatively impacted?

Who will take responsibility for data ethics inside companies if not data scientists?

Baldridge. Emotional contagion: contextualizing the controversy.
http://go.peoplepattern.com/blog/emotional-contagion-one/

Example 3: Subprime Mortgage Crisis 2009

Increase in risky lending
– Home loans close to the actual value of the property
– Aggressive sales of expensive, complex products

Many institutions and investors exposed through

Credit default swaps (CDS)
Mortgage-backed securities (MBS)
Collateralised debt obligations (CDOs)

Much exposure due to poor understanding of bad risk models

It is our Models that are Simple, not our World

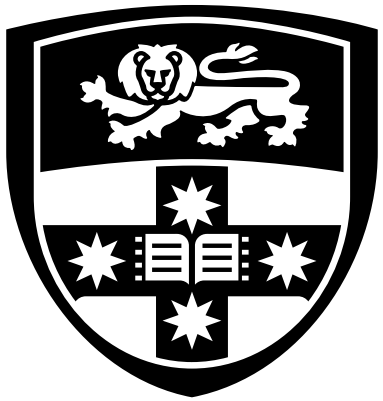
“Our experience in the financial arena has taught us to be very humble in applying mathematics to markets, and to be extremely wary of ambitious theories, which are in the end trying to model human behaviour. We like simplicity, but we like to remember that it is our models that are simple, not the world”

- Emanuel Derman and Paul Wilmott, *The Financial Modelers’ Manifesto*

What does modelling have to do with it?

"All models sweep dirt under the rug. A good model makes the absence of the dirt visible."

Emanuel Derman and Paul Wilmott.
The Financial Modelers' Manifesto



THE UNIVERSITY OF
SYDNEY

Machine Learning and Human Bias

What We Teach Machines

Microsoft’s Tay.ai is a great example of how a machine can adopt human bias.

- Microsoft's AI bot was supposed to come off as a normal teenager
- “designed to engage and entertain people [...] online through casual and playful conversation.”
- But less than a day after she joined Twitter, Tay.ai, turned into a sexist, racist troll…

What if it was making decisions instead of chat?

We are responsible not just for the algorithms, but also for the influences we provide

https://www.techrepublic.com/article/why-microsofts-tay-ai-bot-went-wrong/

Scenario: Optimise Services to Homeless Families

Imagine our goal is to match homeless families with the most appropriate services

We have historical data with various characteristics:
number and age of children and parents, zip code, number and length of previous stays in homeless services, race.

Which characteristics should we use?

What about Race?

Now imagine we find that including race makes the model more accurate.

Should we use it?

Remember algorithm output will be used to help pair families with services.

What about Race?

Using historical data means that we are “training our model” on data that is surely biased, given a history of racism.

An algorithm cannot see the difference between patterns that are based on injustice and patterns that are based on traffic.

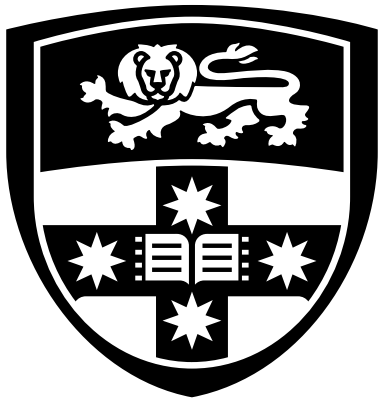
So choosing race as a characteristic in our model would be unethical.

What about Race?

This example is taken from a real project carried out by New York City Health and Human Services.

> *“Looking at old data, we might have seen that families with a black head of household was less likely to get a job, and that might have ended up meaning less job counseling for current black homeless families. In the end, we didn’t include race.”* –Cathy O’Neil

http://www.slate.com/articles/technology/future_tense/2016/02/how_to_bring_better_ethics_to_data_science.html



THE UNIVERSITY OF
SYDNEY