Introduction to Big Data Computing

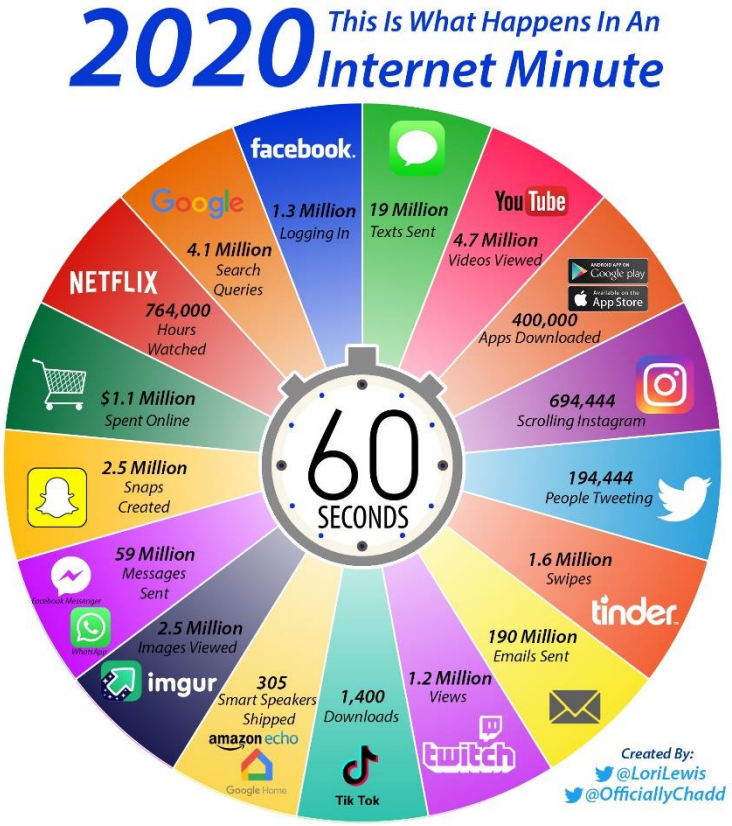
Song Guo
Department of Computing
The Hong Kong Polytechnic University

Roadmap

- What is Big Data?
- What is Big Data Analytics?
- Big Data System
- Applications of Big Data

Where Big Data Come From

- Posts to social media sites
- Digital pictures and videos
- Software logs, cameras
- Microphones
- Sensing data
- Scans of government documents
- GPS trails
- Purchase transaction records
- Cell phone GPS signals
- Traffic
- ...



What is Big Data

- Definition from Wikipedia:
  - “Big data” is a field that treats ways to analyze, systematically extract information from, or deal with data sets that are too **large** or **complex** to be dealt with by traditional data-processing application software.
  - The challenges include include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating, information privacy and data source.
  - Tends to refer to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data.

Characterization of Big Data

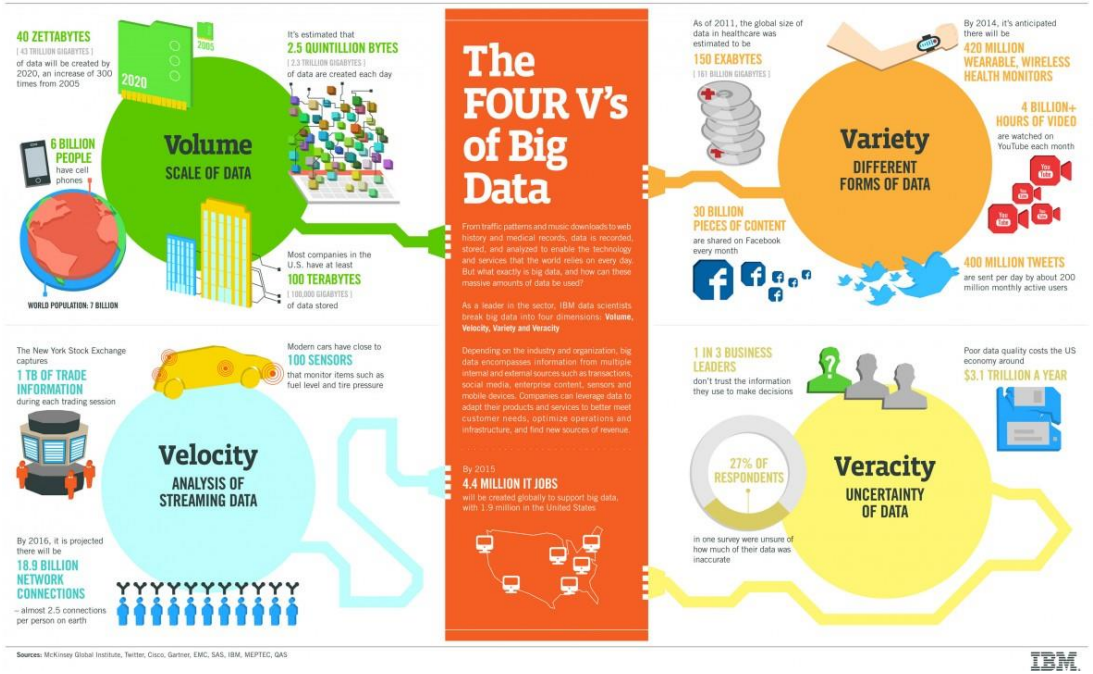
- "Big data is high-**volume**, high-**velocity** and high-**variety** information assets that demand **cost-effective**, **innovative** forms of information processing for **enhanced insight and decision making**." -- Gartner

- "While enterprises struggle to consolidate systems and collapse redundant databases to enable greater operational, analytical, and collaborative consistencies, changing economic conditions have made this job more difficult. Ecommerce, in particular, has exploded data management challenges along three dimensions: volumes, velocity and variety. In 2001/02, IT organizations much compile a variety of approaches to have at their disposal for dealing each." – Doug Laney

Characterization of Big Data

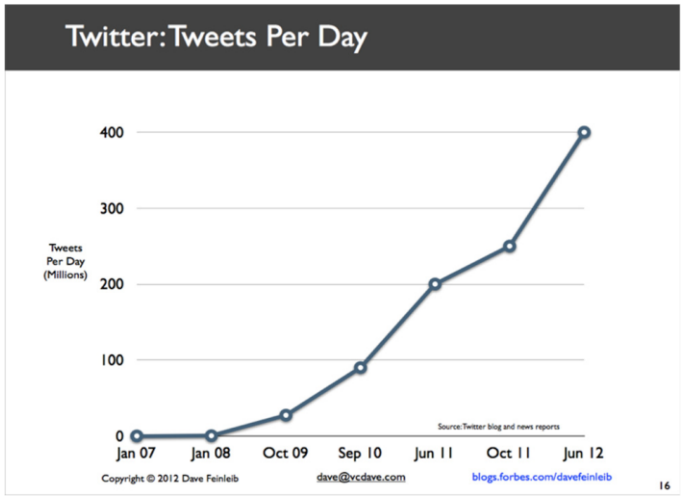
- "Big data is high-**volume**, high-**velocity** and high-**variety** information assets that demand **cost-effective**, **innovative** forms of information processing for **enhanced insight and decision making**." -- Gartner

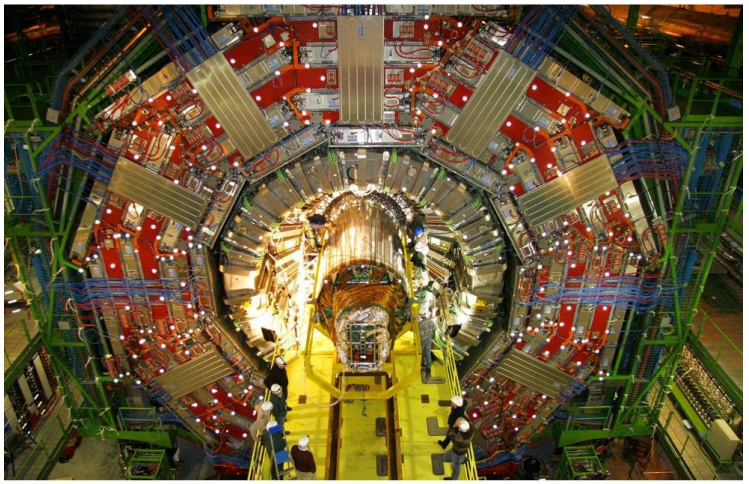
- **V**s of big data
  - **Volume**
  - **Velocity**
  - **Variety**
  - **Veracity**



Volume: scale of data

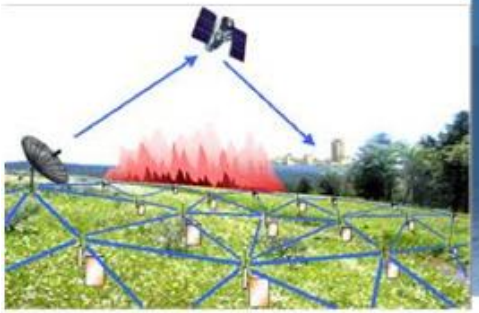
- Data volume is increasing exponentially
- Generated by huge number of devices and sensors
  - 5 billion people have mobile phones
  - A modern car has 100 sensors Value
  - CERN's Large Hydron Collider (LHC) generates 15 PB/year





Velocity: speed of data generation

- Data is generated fast
  - e.g., every 60 seconds, there are 11 million instant messages, 168 million emails sent

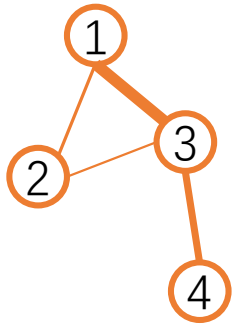


Velocity: speed of data processing

- Data need to be processed fast
  - Online Data Analytics: late decisions means missing opportunities
  - E.g. 1: Based on your current location and your purchase history, send promotions right now for store next to you
  - E.g. 2: Sensors monitoring your activities and body, notify you if there are abnormal measurements

Variety: data in many forms

John likes to watch movies. Mary likes movies too.



Text

Tuple	x_1
1	John also likes to watch football games.
2	John likes to watch movies. Mary likes movies too.

Signal (Voice, Audio)

signal1	signal2
13.58	7.24
12.11	12.50
13.49	8.66
11.25	10.98
14.57	13.75
13.22	9.02

Image

125	200	225
105	150	255
15	75	175

Graph

A	1	2	3	4
1	0	1.2	4.3	0
2	1.2	0	0.8	0
3	4.3	0.8	0	2.6
4	0	0	2.6	0

E	Node1	Node2	Weight
1	1	2	1.2
2	1	3	4.3
3	2	3	0.8
4	3	4	2.6

Variety: data in many forms

- A single application may generate/collect many types of data, e.g., types of data are stored in emails
  - Tabular data: attributes like subject, to, from
  - Text (in email body)
  - Image (in attachment)
  - Hyperlinks
- Types of data
  - Relational Data (e.g., Tables)
  - Text Data (e.g., comments)
  - Semi-structured Data (e.g., XML)
  - Graph Data (e.g., social network)
  - What else?

Veracity: uncertainty of data

- Is the data accurate?
  - Measurement error
  - Human errors like typos in names/addresses
- Does the data come from a reliable source?
- What if data from different sources are not consistent?

Lumia 650 Case, CoverON [HexaGuard Series] Slim Hybrid Hard Phone Cover Case for Microsoft Lumia 650 - Black & Black

Visit the CoverON Store

5 ratings

Currently unavailable.

We don't know when or if this item will be back in stock.

Color: **Black & Black**

Material	Silicone, Rubber, Polycarbonate
Brand	CoverON
Color	Black & Black
Compatible Phone Models	Microsoft Lumia 650

About this item

- CoverON [HexaGuard Series] protective hard hybrid case for the Microsoft Lumia 650
- Hard polycarbonate outer layer combined with a soft flexible rubber silicone inner layer
- Enhanced "basketball" grip on the back
- Shock absorbing and protective yet thin and lightweight
- 90 day manufacturer warranty

Roll over image to zoom in

Fake, Paid-For Reviews in Amazon

This case is really good, it looks very nice and doesn't add bulk ...

By [redacted] on January 16, 2016

Color: Black & Black

The case is lightweight and slick. This case is really good, it looks very nice and doesn't add bulk or excessive size to my phone.It is so easy to get in and out of my pocket and the color is nice. It fits snuggly and provides good protection. It's so pretty and it seems durable. I receive this product to give honest review.

2 comments | 0 of 8 people found this helpful. Was this review helpful to you? Yes No Report abuse

It protects well and feels good on my phone

By [redacted] on January 20, 2016

Color: Black & Black

This case fits my phone well and is very durable. It protects well and feels good on my phone. I received this for free for doing this review.

2 comments | 0 of 8 people found this helpful. Was this review helpful to you? Yes No Report abuse

Full phone protection.

By [redacted] on January 17, 2016

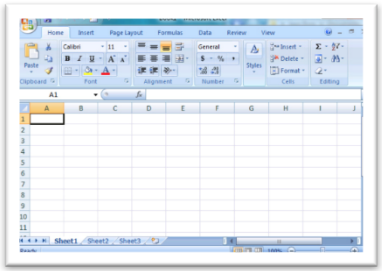
Color: Black & Black

I receive this case yesterday and give this CoverON HexaGuard series hybrid case for the Microsoft Lumia 650 for a family member. This case really protects the phone and this case is an excellent choice for those who want the protection of a 2-layer hybrid case, however it does add some bulkiness to the phone. The tittle said its slim, but for me it's a little thick but to whom I've given it to, he said it's great. This case has a flexible rubber silicone skin as inner shell with thicker extra padding on the corners. Then a tough polycarbonate outer shell. I got this at a discounted price.

Evolution of Data Analytics

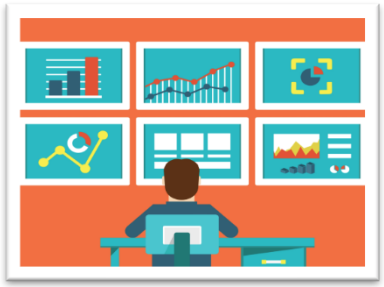
1990s → **2000s** → **2015 and beyond**

What Happened?



Excel

What's Happening?

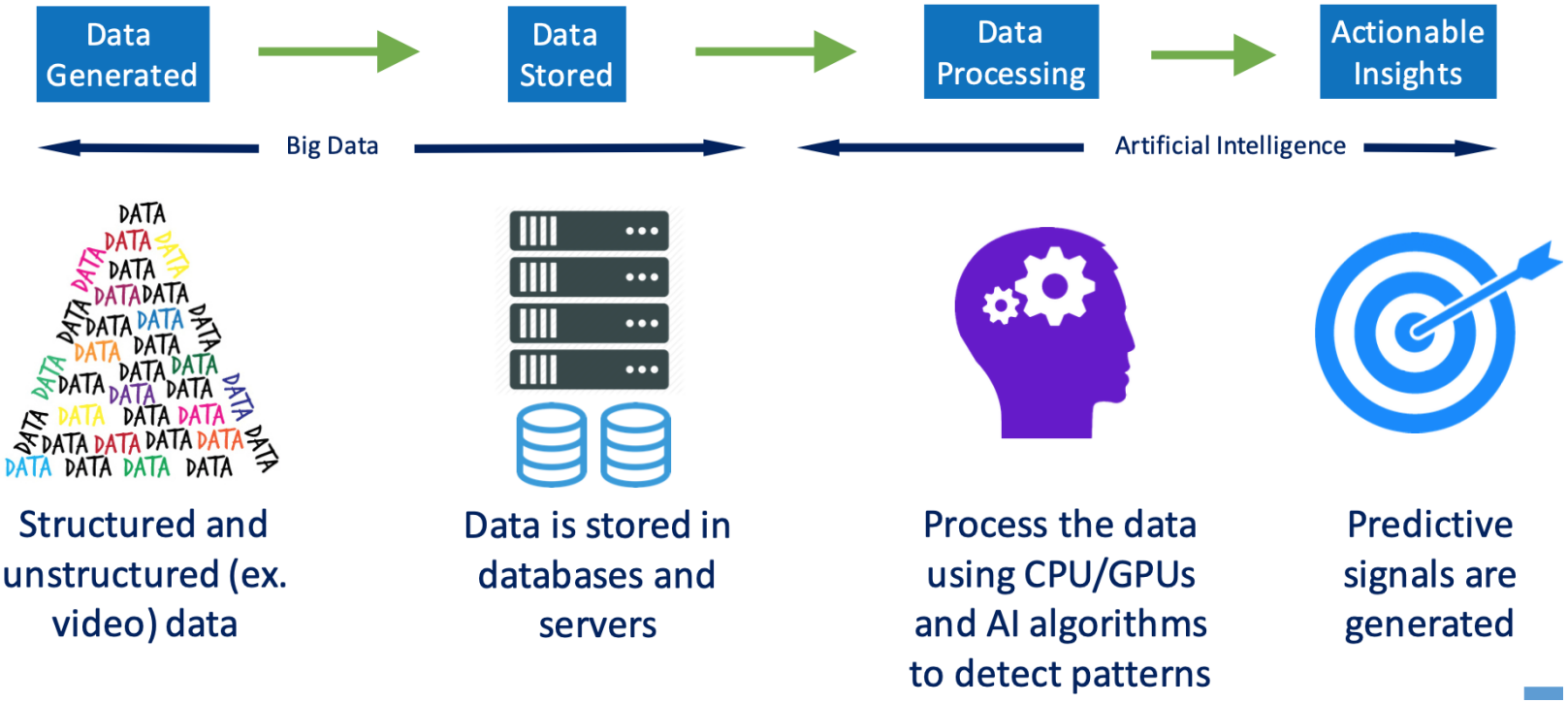


Business Intelligence (BI) Dashboards

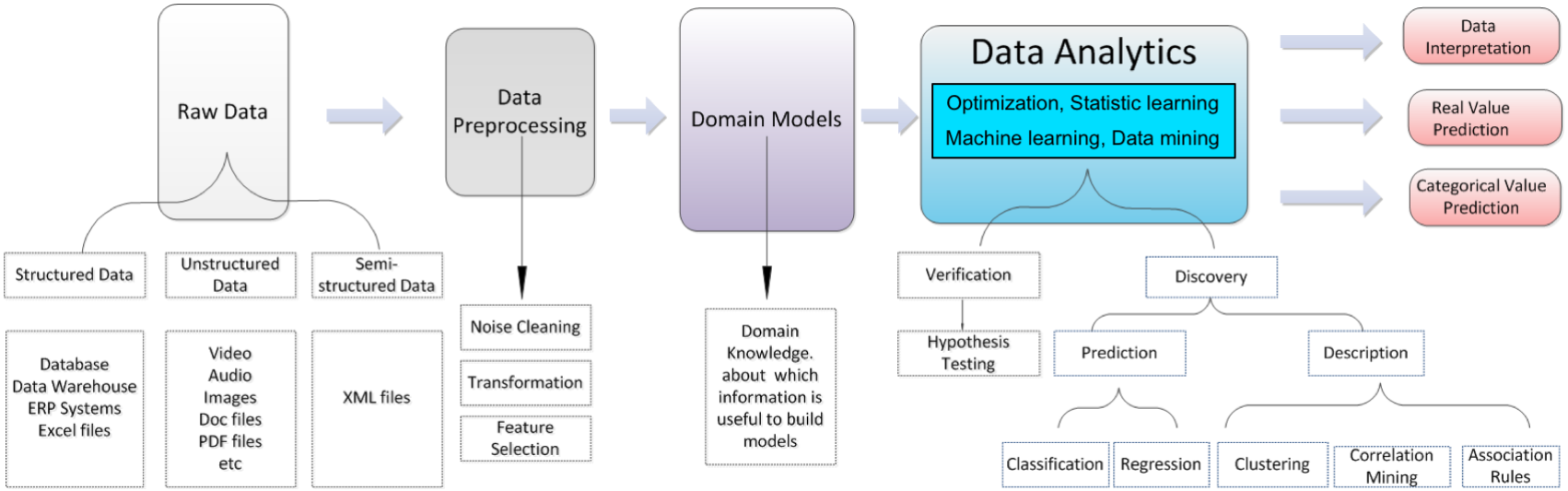
What Will Happen?

Actionable Insights

Process of Data Analytics



Big Data Driven Approach



Big Data Processing Platforms

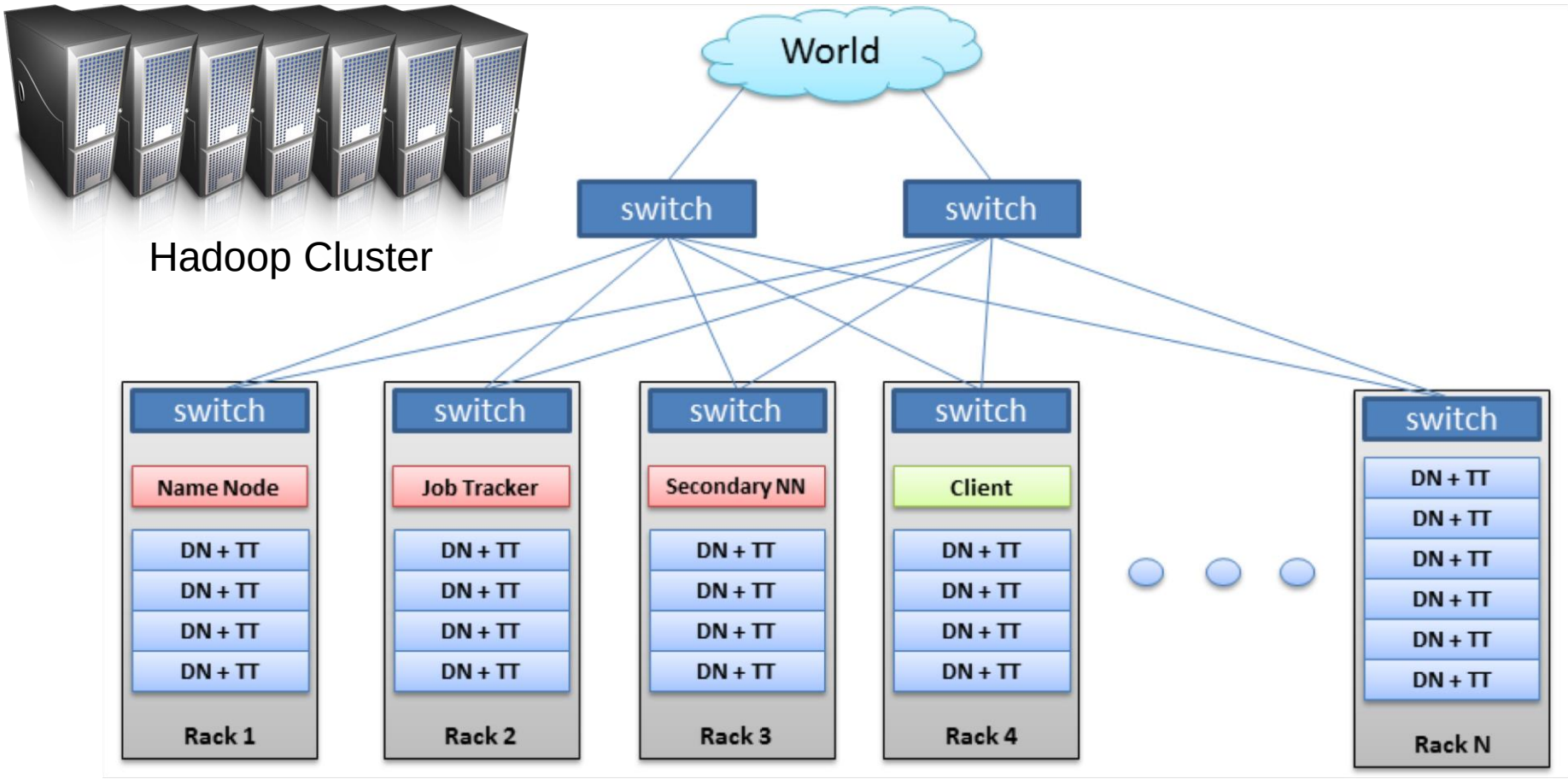
Big Data System

- Data Collection and Analytic on
  - Clusters of machines
  - Program/Query from a single machine perspective
- Data Processing (NoSQL)
  - Traditional Database (e.g., Microsoft SQL Server, Oracle) could not keep up with Big Data (Google, FB) companies need
  - What they needed? Extreme high insert throughput (e.g., tweets)
  - What they did not need? Transactions (ACID)
  - They developed NoSQL “database” themselves
- Analytic Platform (Hadoop and Spark)

Techniques towards Big Data

- Massive Parallelism
- Huge Data Volumes Storage
- Data Distribution
- High-Speed Networks
- High-Performance Computing
- Task and Thread Management
- Data Mining and Analytics
- Data Retrieval
- Machine Learning
- Data Visualization

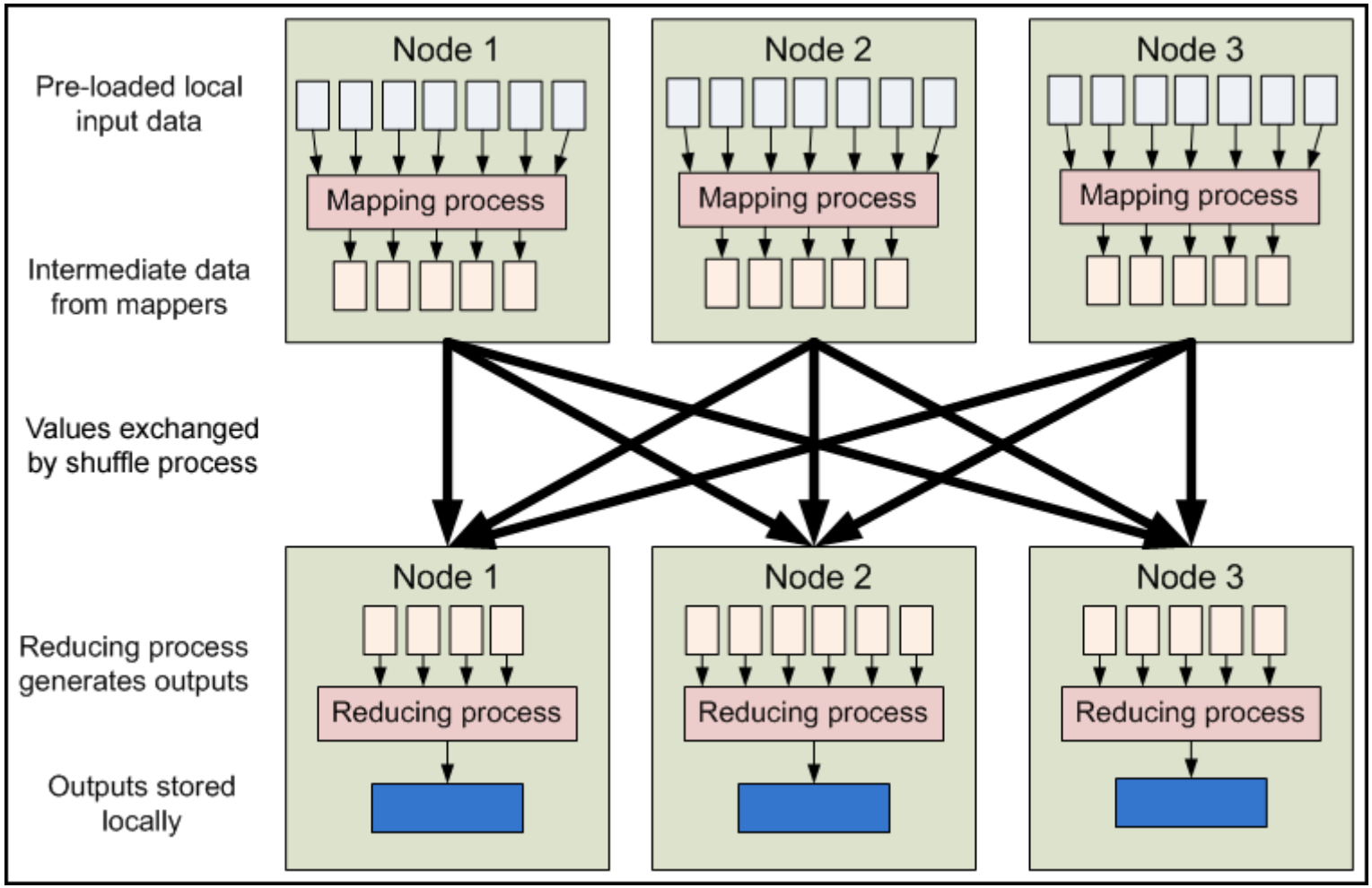
Big Data Analytics Platform



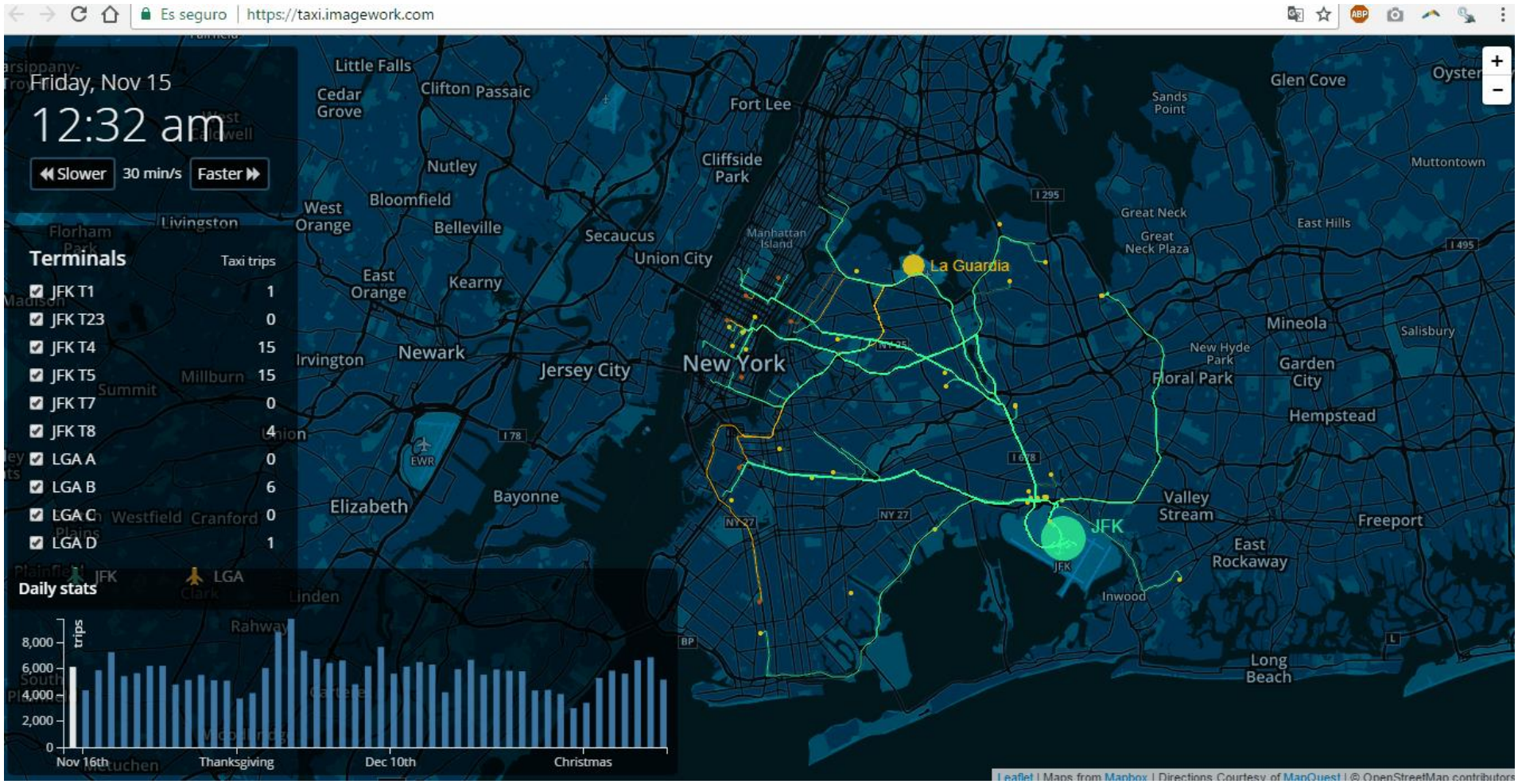
Big Data Programming - MapReduce

- Map: a **mapping** that is responsible for dividing the data and transforming the original data into key-value pairs
- Shuffle: the process of further organizing and delivering the Map output to the Reduce
  - the output of the Map must be sorted and segmented
  - then passed to the corresponding Reduce
- Reduce: a **merge** that processes the values with the same key and then outputs to the final result

MapReduce Pipeline



Data Visualization



Taxi trajectories in New York City from Nov 15th to December 31st, in 2013

Big Data Application - AlphaGO

- AlphaGo learns from 30 million moves of 160 thousands games played by experts (5-9 dan) → **Big Data**
- AlphaGo uses deep learning and neural networks combined with Monte-Carlo tree search to decide the moves → **Analytics**

Go, a complex game popular in Asia, has frustrated the efforts of artificial-intelligence researchers for decades.

ARTIFICIAL INTELLIGENCE

Google masters Go

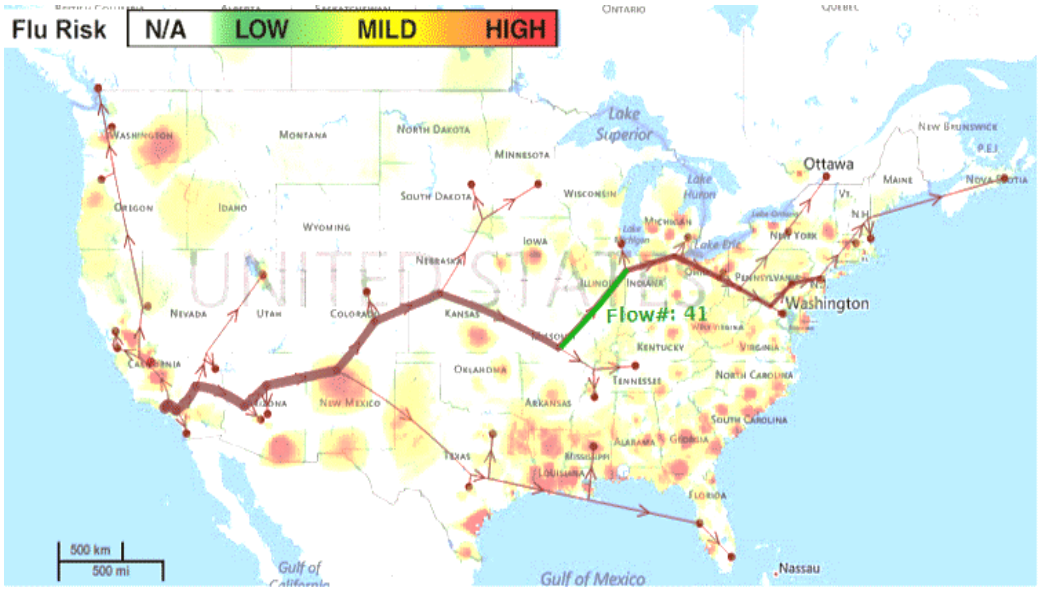
Deep-learning software excels at complex ancient board game.

BY ELIZABETH GIBNEY

A computer has beaten a human professional for the first time at Go — an ancient board game that has long reveals in research published in *Nature* on 27 January[1]. It also defeated its silicon-based rivals, winning 99.8% of games against the current best programs. The program has yet to play the Go equivalent of a world champ- famously beat grandmaster Garry Kasparov in 1997, was explicitly programmed to win at the game. But AlphaGo was not preprogrammed to play Go: rather, it learned using a general-purpose algorithm that allowed it to interpret

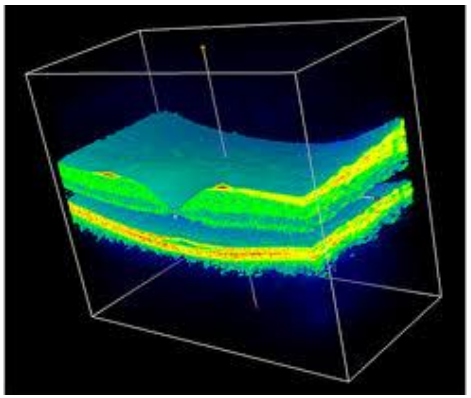
Big Data Application - GTF

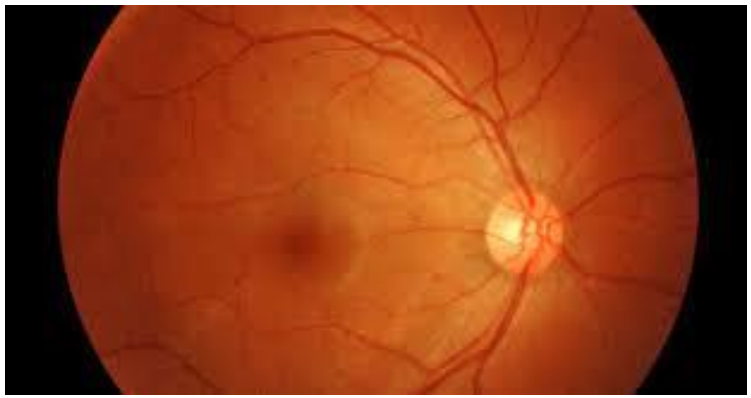
- Google Flu Trends (GFT) was once held-up as the prototypical example of the power of big data
- By leveraging search term data, a group of Data Scientists with little relevant expertise were able to predict the spread of flu across the continental United State
- More accurate than the “experts” at the Centre for Disease Control with their models built from expensive survey data



Big Data Application - DeepMind

- Disease Treatment: Joint research between Google DeepMind and Moorfields Eye Hospital
  - Eyecare professionals diagnose eye conditions by using optical coherence tomography (OCT) scans (over 1,000 a day at Moorfields alone)
  - Achieving expert error rate 5.5% comparably to the two best retina specialists (6.7% and 6.8% error rate)





Big Data Application - CityBrain