

# Introduction to Data Analytics

Dr. Jing Li

Department of Computing

The Hong Kong Polytechnic University

# Why you are here?

*Dreaming of becoming part of them?*



CONSULTING PARTNER

# Roadmap

- What is Data?
- What is Data Analytics?
- The History of Data Analytics.
- Applications of Data Analytics.



# Roadmap

- What is Data?
- What is Data Analytics?
- History of Data Analytics.
- Applications of Data Analytics.



# What is Data?



- Individual **units of information**.
- A single *quality or quantity* of some object or phenomenon.
- In the *analytical processes*, data is represented by **variables**. (*Why?*)
- Data appears in *various forms of human activities*.
  - Scientific Research (e.g., COVID-19 research)
  - Business Management (e.g., sales data, revenue, profits, stock price)
  - Government (e.g., crime rates, unemployment rates, literacy rates)

--- From [Wikipedia](#)

# Data is Everywhere!

- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - Purchases at grocery stores
  - Bank/Credit Card transactions
  - Social Media
  - ...
  - Find COVID-19 patients
  - COVID-19 vaccine?!



# Roadmap

- What is Data?
- What is Data Analytics?
- The History of Data Analytics.
- Applications of Data Analytics.



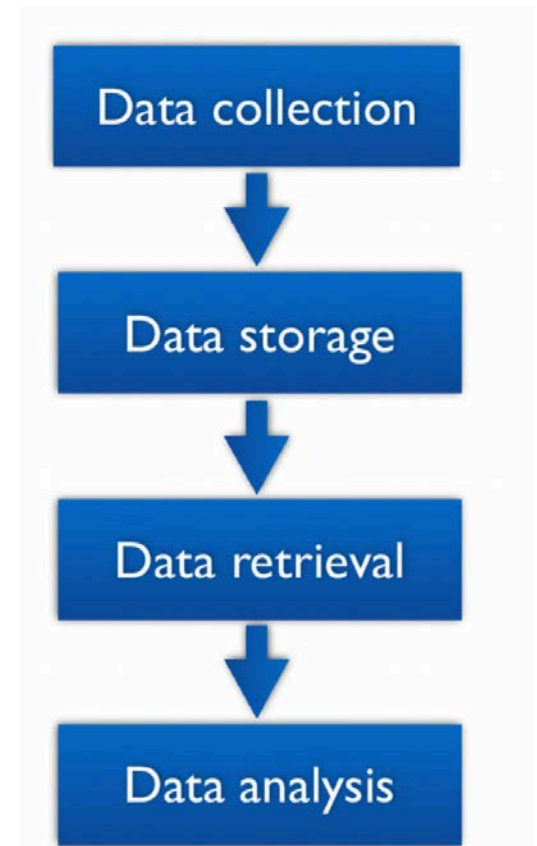
# What is Data Analytics (or Analysis)?

- **The process of data analysis:**

- The process of **inspecting, cleaning, transforming, and modelling of data**

- **The goal of data analytics:**

- To *discover useful information, informing conclusion, and to support decision-making.*



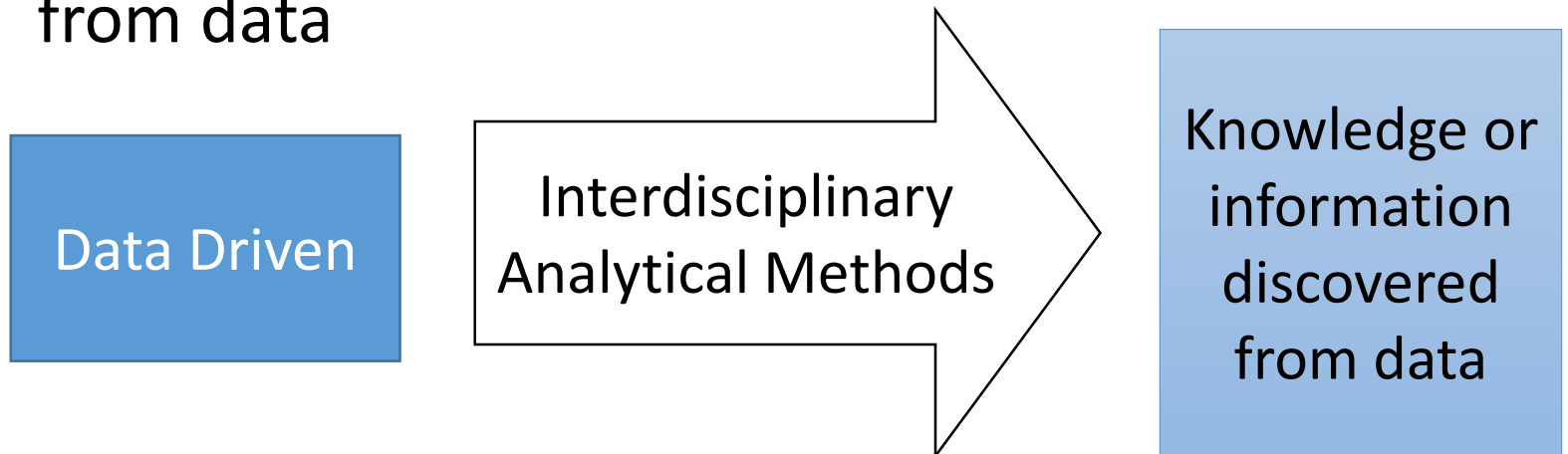
--- From [Wikipedia](#)



# What is Data Analytics

- **Characteristics:**

- **Input:** Data Driven (the more the better, *why?*)
- **Methods:** Interdisciplinary methods (mathematics + computer science)
- **Output:** Discover Knowledge or Information from data



# Roadmap

- What is Data?
- What is Data Analytics?
- The History of Data Analytics.
- Applications of Data Analytics.



# A Real Story: Data Analytics Saves People's Life



A COURT FOR KING CHOLERA.

SCALE 30 INCHES TO A MILE.

# The First Big Data Challenge

- 1880 census
- 50 million people
- Age, gender (sex), occupation, education level, no. of insane people in household

Received July 28, 1890  
 C

-2-4084-

Page No. 31

Secretary's Dist. No. 21

Division No. 122

State A-Other Districts Year Ending June 1, 1890, and until May 31, 1890

June 1, 1890, and the UNITED STATES, MEMBERS OF PARLIAMENT HAVE BEEN Elected SINCE June 1, 1890, with the following:

Members of Parliament who are not yet elected to the United States Senate are not yet elected to the United States Senate.

279

John Jones

**SCHEDULE I-Inhabitants in, Raton, Colorado, in the County of Medicine, State of New Jersey**

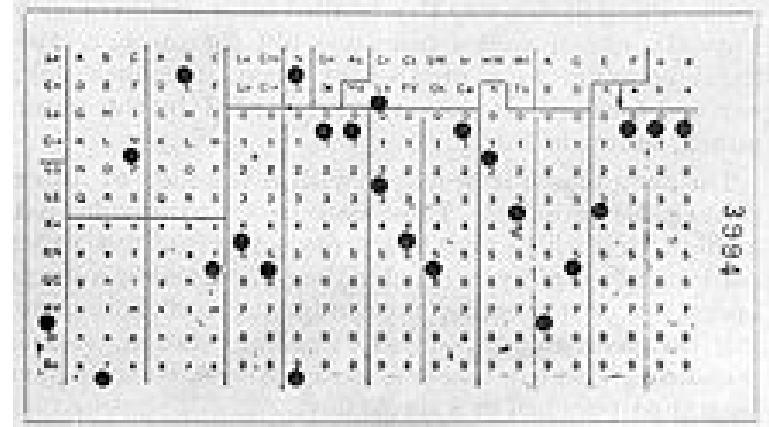
enumerated by me on the 1st day of June, 1890.

John Jones

Name		Age	Sex	Color	Marital Status	Occupation	Place of Birth	Place of Residence	Place of Nativity	Place of Birth	Place of Residence	Place of Nativity
1	John Jones	45	M	W	Married	Farmer	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey
2	John Jones	45	M	W	Married	Farmer	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey
3	John Jones	45	M	W	Married	Farmer	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey
4	John Jones	45	M	W	Married	Farmer	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey
5	John Jones	45	M	W	Married	Farmer	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey
6	John Jones	45	M	W	Married	Farmer	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey
7	John Jones	45	M	W	Married	Farmer	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey
8	John Jones	45	M	W	Married	Farmer	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey
9	John Jones	45	M	W	Married	Farmer	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey
10	John Jones	45	M	W	Married	Farmer	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey
11	John Jones	45	M	W	Married	Farmer	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey
12	John Jones	45	M	W	Married	Farmer	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey
13	John Jones	45	M	W	Married	Farmer	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey
14	John Jones	45	M	W	Married	Farmer	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey
15	John Jones	45	M	W	Married	Farmer	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey
16	John Jones	45	M	W	Married	Farmer	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey
17	John Jones	45	M	W	Married	Farmer	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey
18	John Jones	45	M	W	Married	Farmer	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey
19	John Jones	45	M	W	Married	Farmer	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey
20	John Jones	45	M	W	Married	Farmer	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey
21	John Jones	45	M	W	Married	Farmer	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey
22	John Jones	45	M	W	Married	Farmer	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey
23	John Jones	45	M	W	Married	Farmer	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey
24	John Jones	45	M	W	Married	Farmer	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey
25	John Jones	45	M	W	Married	Farmer	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey
26	John Jones	45	M	W	Married	Farmer	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey
27	John Jones	45	M	W	Married	Farmer	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey
28	John Jones	45	M	W	Married	Farmer	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey
29	John Jones	45	M	W	Married	Farmer	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey
30	John Jones	45	M	W	Married	Farmer	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey
31	John Jones	45	M	W	Married	Farmer	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey
32	John Jones	45	M	W	Married	Farmer	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey
33	John Jones	45	M	W	Married	Farmer	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey
34	John Jones	45	M	W	Married	Farmer	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey
35	John Jones	45	M	W	Married	Farmer	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey	New Jersey

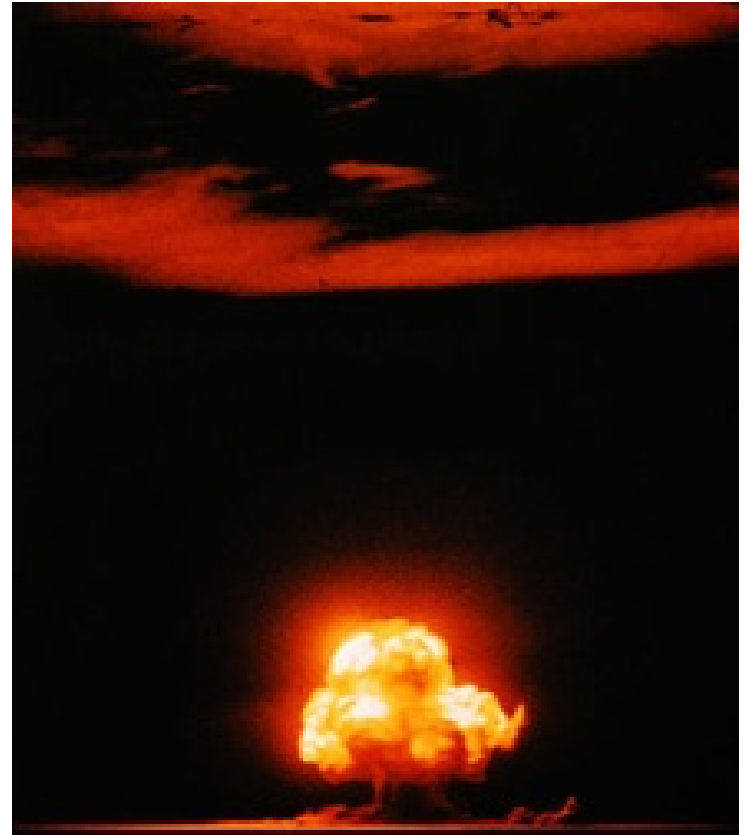
# The First Big Data Solution

- Hollerith Tabulating System
- Punched cards – 80 variables
- Used for 1890 census
- 6 weeks instead of 7+ years



# Manhattan Project (1946 - 1949)

- \$2 billion (approx. 26 billion in 2013)
- Catalyst for “Big Science”



# Space Program (1960s)

- Began in late 1950s
- An active area of big data nowadays



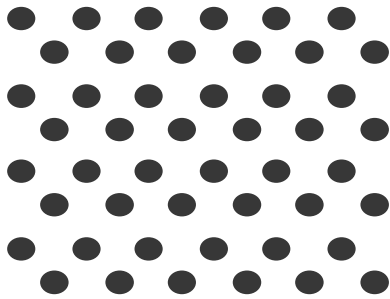


# Now, we are in the big data era!

- Better models?
  - With more variables to fit data!
  - Rule-based -> Statistical -> Deep Learning
- Better computing resource?
  - More powerful RAM, CPU, GPU, etc.
- Also importantly, more data!
  - Huge volume of data is available to do analytics and discover valuable information from it!

# Characteristics of Big Data: 4V

## Volume

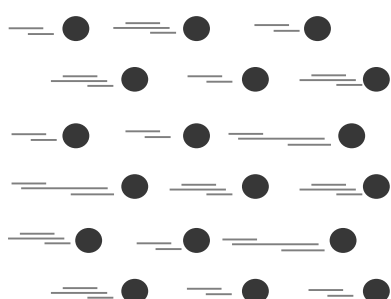


From terabytes to exabyte to zetabytes of existing data to process



8 billion TB in 2015,  
40 ZB in 2020  
5.2TB per person

## Velocity

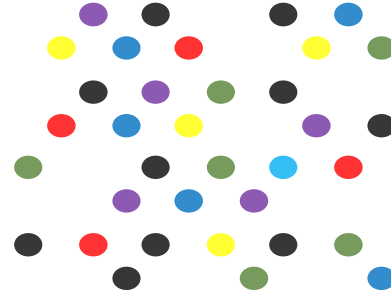


Batch data, real-time data, streaming data, milliseconds to seconds to respond

facebook

New sharing over 2.5 billion per day  
new data over 500TB per day

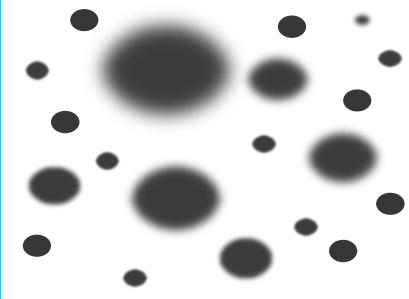
## Variety



Structured, semi-structured, unstructured, text, pictures, multimedia



## Veracity



Uncertainty due to data inconsistency & incompleteness, ambiguities, deception, model approximation



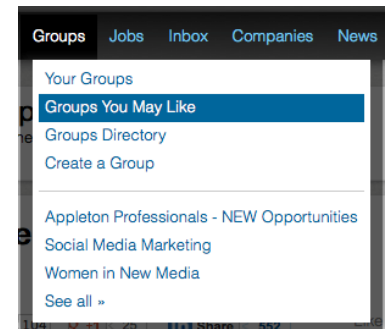
# Roadmap

- What is Data?
- What is Data Analytics?
- The History of Data Analytics.
- Applications of Data Analytics.



# Application: Product Recommendation

- **Main idea:** Recommend items to customer  $x$  similar to previous items rated highly by  $x$
- Example:
  - **Movie recommendations**
    - Recommend movies with same actor(s), director, genre, ...
  - **Websites, blogs, news**
    - Recommend other sites with “similar” content



# Example: Ranking of Webpages

- Computing importance of webpages.

Homepage

Wikipedia

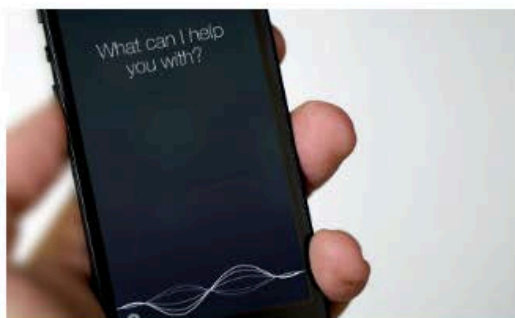
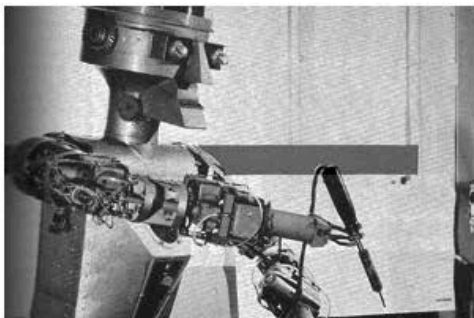
The image shows a Google search interface with the query "hong kong polytechnic university". The search results are displayed on a light yellow background. The first result is "PolyU: The Hong Kong Polytechnic University" with the URL "https://www.polyu.edu.hk/web/en/home/index.html". This result is highlighted with a dashed red border. Below it is a result for "Incoming Students | PolyU International - The Hong Kong Polytechnic ..." with the URL "https://www.polyu.edu.hk/international/incoming-students". The third result is "Hong Kong Polytechnic University - Wikipedia" with the URL "https://en.wikipedia.org/wiki/Hong\_Kong\_Polytechnic\_University". To the left of the search results, the word "Homepage" is written in a large, dark red font, and the word "Wikipedia" is written in a large, dark green font. Two arrows point from these words to the corresponding search results: a red arrow points from "Homepage" to the PolyU result, and a green arrow points from "Wikipedia" to the Wikipedia result. The search results page also includes a sidebar on the right with a photo of the university building and a list of links: "Website", "Directions", and "Public university in Hong Kong".

# Application: Artificial Intelligence (AI)

**Past:** 1950s – 1990s

**Now:** 2000s – 2020s

**Future:** ????



# A slide to take away

- **Data**, individual **units of information**, is usually represented by **variables** in the analytics processes.
- **Data analytics** is *driven by the data* (generally in large scale), processed with *interdisciplinary methods (math + CS)*, and with the goal *to discover knowledge and information from data*.
- There's a *long history for data analytics* and it becomes trendy these years because of the availability of *effective analytical models, rich computing resources*, and *large-scale data*.