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?

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!

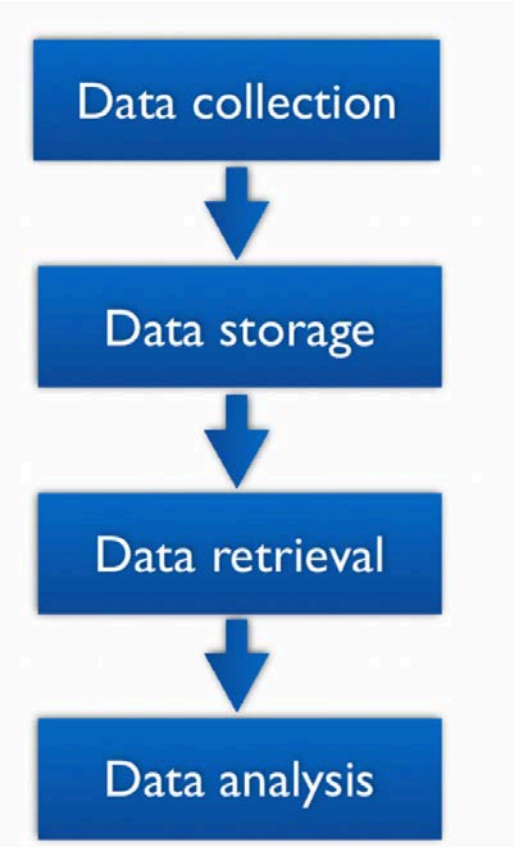
- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - Purchases at grocery stores
  - Bank/Credit Card transactions
  - Social Media
  - …
  - Find COVID-19 patients
  - COVID-19 vaccine?!

Roadmap

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

What is Data Analytics (or Analysis)?

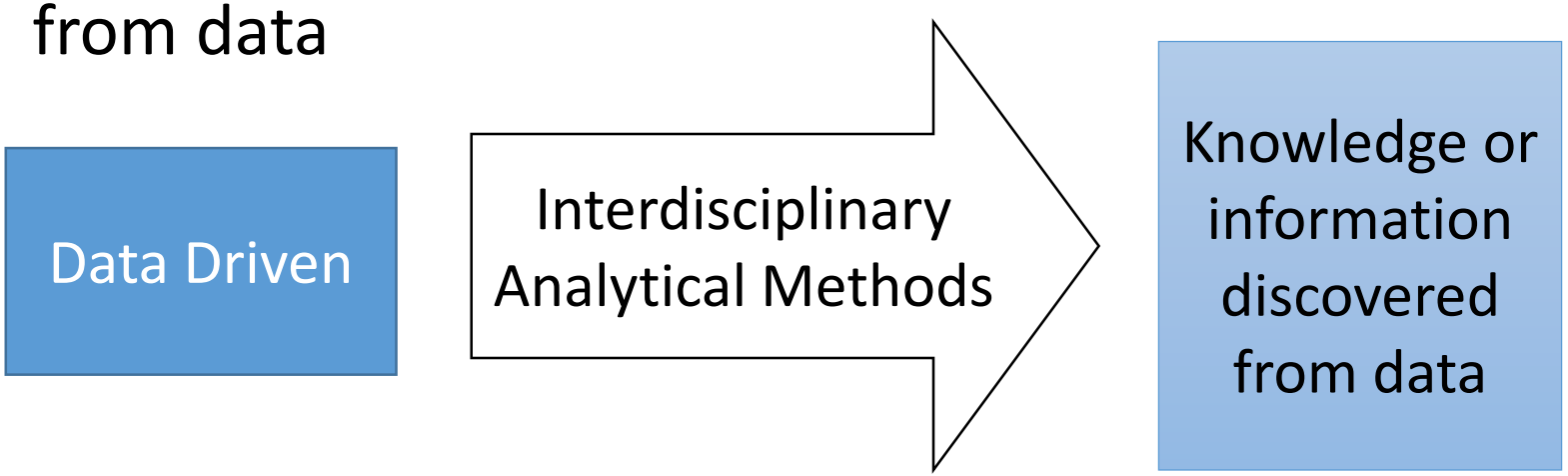
- **The process of data analysis:**
  - The process of inspecting, cleaning, transforming, and modelling of data
- **The goal of data analytics:**
  - To *discover useful information, informing conclusion, and to support decision-making*.



--- 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

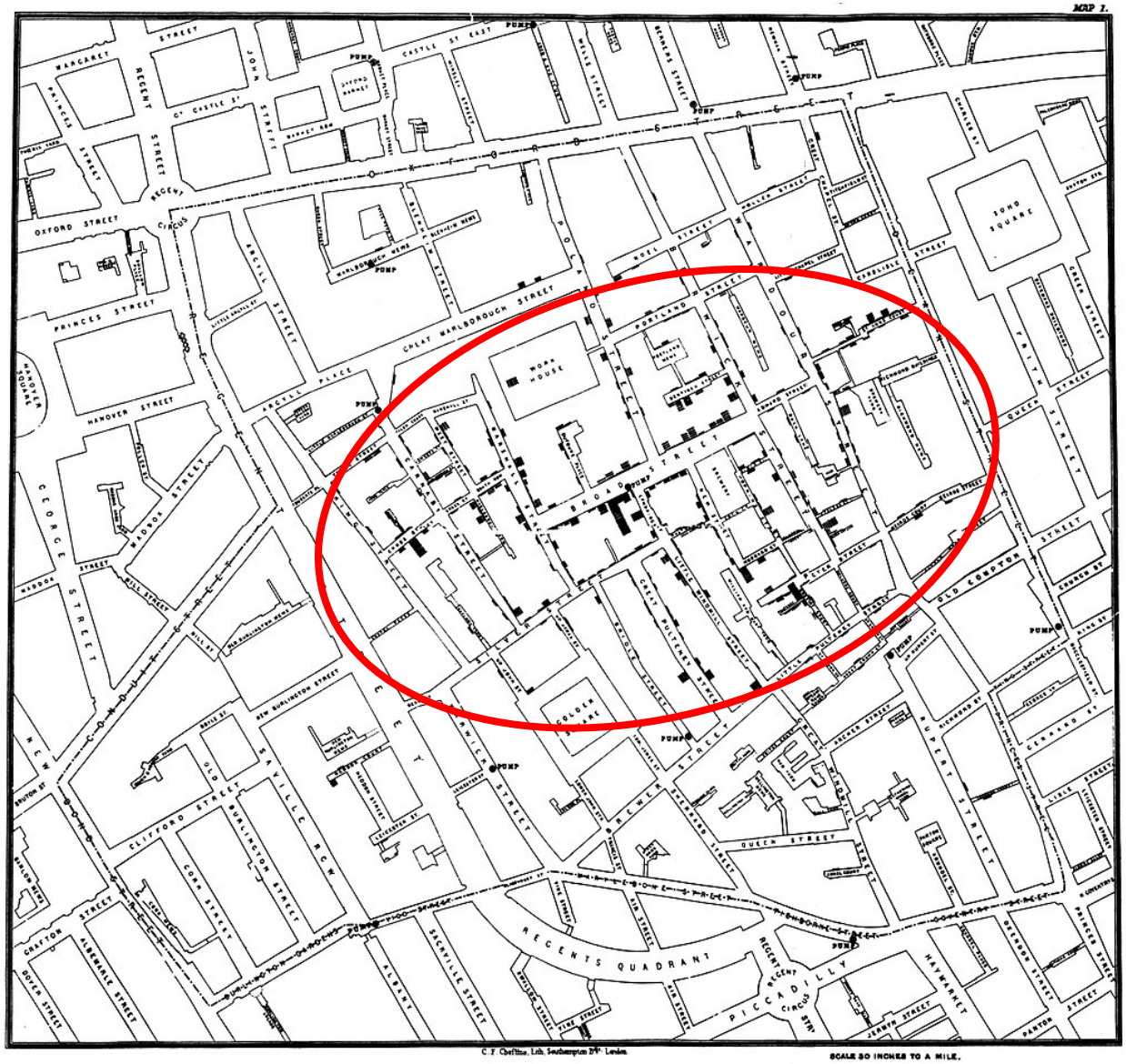


Roadmap

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

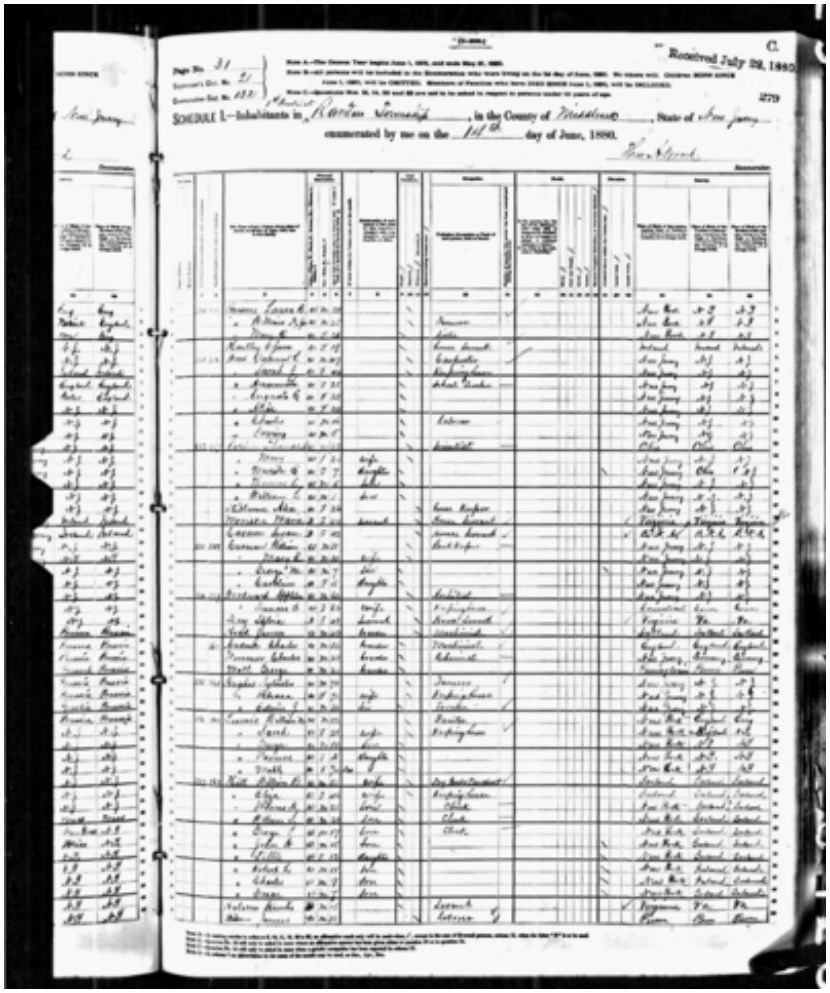
A Real Story: Data Analytics Saves People's Life

A COURT FOR KING CHOLERA.


MAP 1.
BROAD STREET
WORK HOUSE
REGENTS QUADRANT
PICCADILLY
SCALE 30 INCHES TO A MILE.

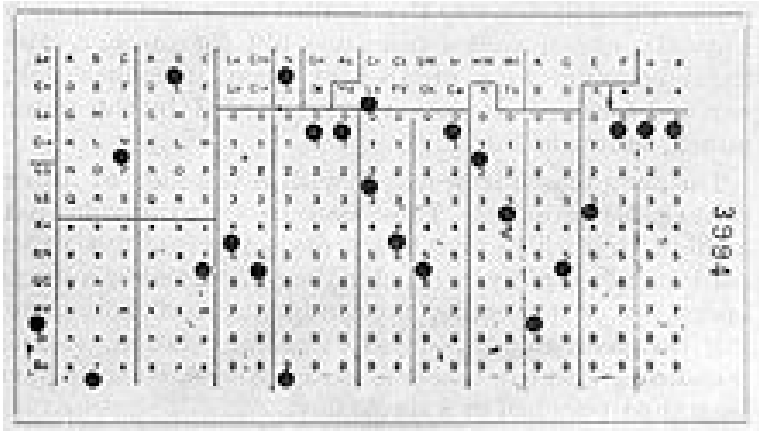
The First Big Data Challenge

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



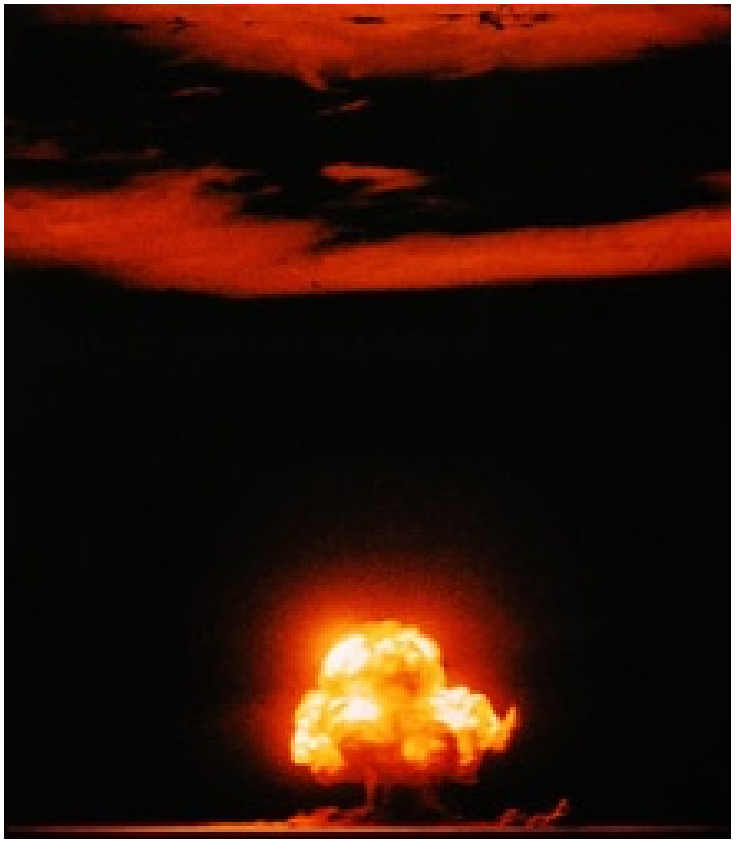
The First Big Data Solution

- 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)

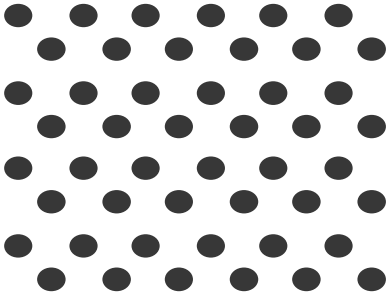
- Began in late 1950s
- An active area of big data nowadays

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

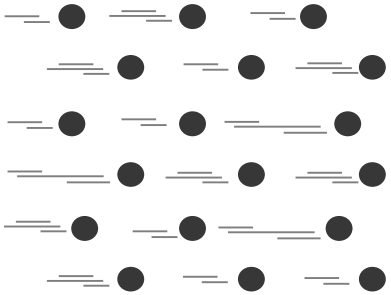


From terabytes to exabyte to zetabytes of existing data to process

IDC Analyze the Future

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

Velocity

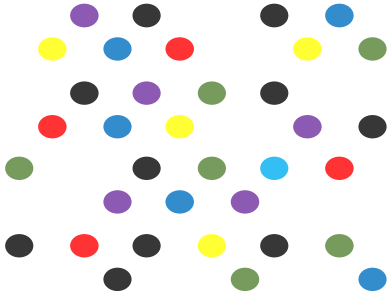


Batch data, real-time data, streaming data, milliseconds to seconds to respond

facebook

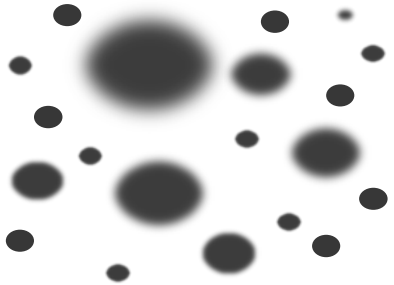
New sharing over 2.5 billion per day
new data over 500TB per day

Variety



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

Veracity



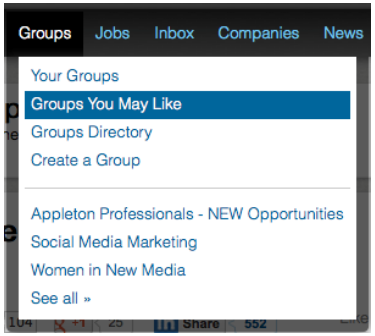
Uncertainty due to data inconsistency & incompleteness, ambiguities, deception, model approximation

Roadmap

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

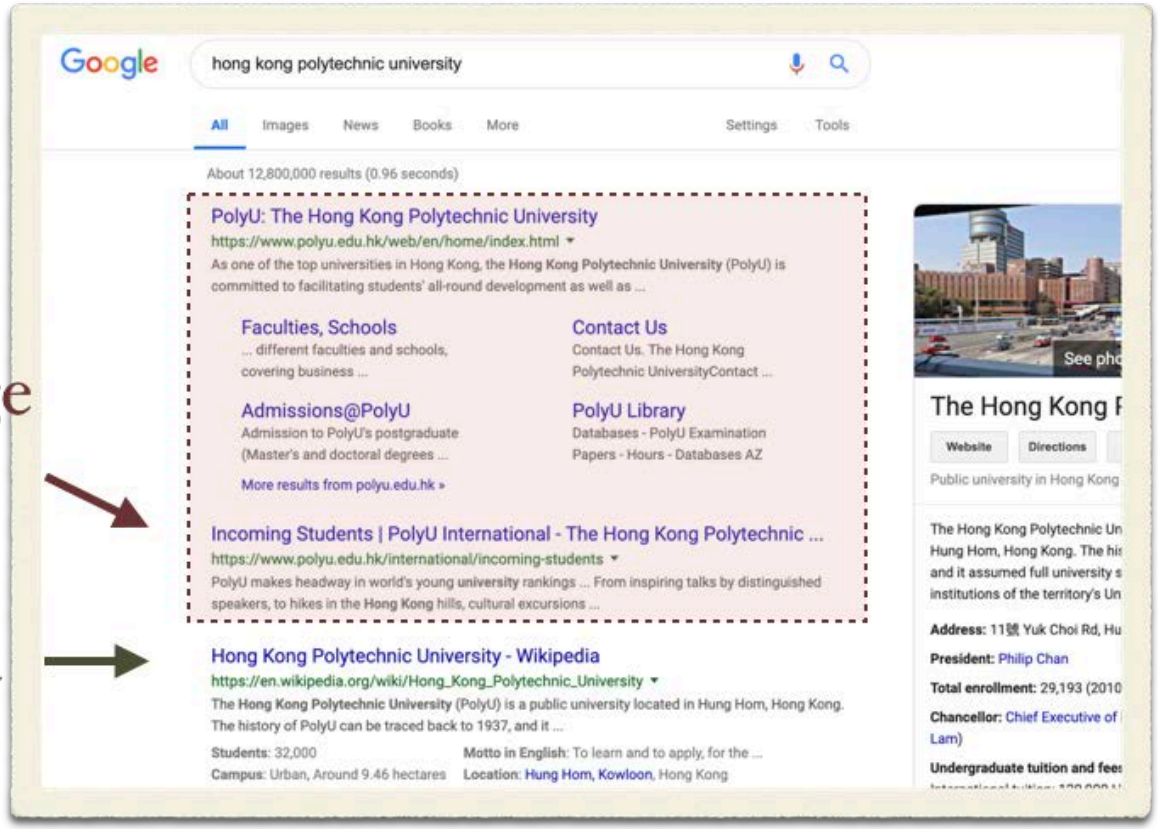
Application: Product Recommendation

- 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



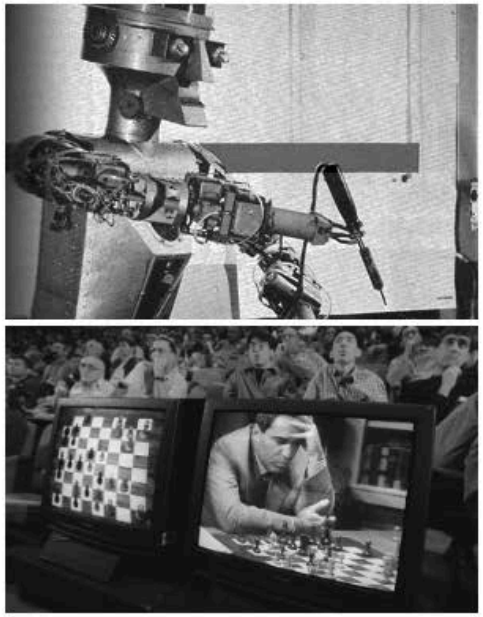
Example: Ranking of Webpages

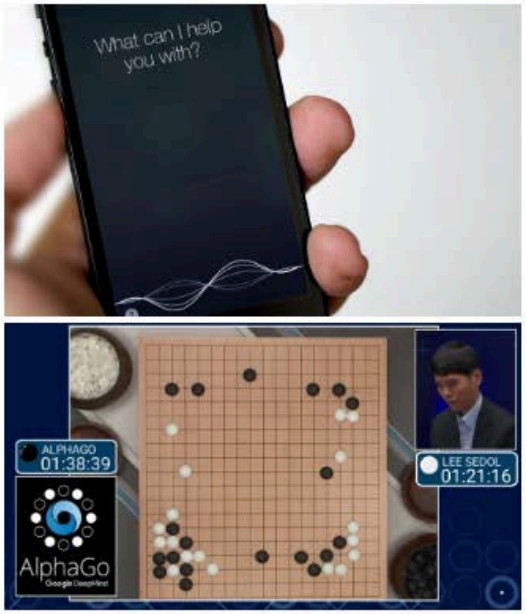
- Computing importance of webpages.



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*.