COMP 1433: Introduction to Data Analytics &
COMP 1003: Statistical Tools and Applications

Lecture 1 - Data Analytics: An Introduction

Dr. Jing Li
Department of Computing
The Hong Kong Polytechnic University

Jan 11 & 13, 2022

Why you are here?

Dreaming to become part of them?

Google

Find us on Facebook

Baidu 百度
www.baidu.com

Tencent 腾讯

Alibaba.com

amazon web services | Partner Network
CONSULTING PARTNER

bing

Why you are here?

- Do you want to understand what data is and explore the secretes behind?
- Do you want to understand the infrastructure and techniques of data analytics?
- Do you want to know more about the area of data mining and machine learning?

Play the data with mathematics and computing techniques ☺

Roadmap

- A quick look at this course
- A tour of data analytics
  - Introduction
  - Big Data and Data Mining

Roadmap

- <u>**A quick look at this course**</u>
- A tour of data analytics
  - Introduction
  - Big Data and Data Mining

Our Promise. We will…

- provide a positive, respectful, and engaged learning environment inside and outside the classes;
- attend classes at regularly scheduled times without undue variations, and provide before term-end adequate make-ups of classes that are canceled due to leave of absence of the instructor;
- provide the course schedule (in Content File on Blackboard: learn.polyu.edu.hk);
- meet with students with a mutually convenient appointment if necessary (preferably in office hours or other time by appointment);

Our Promise. We will (cont.)

- have reasonable access to University facilities and equipment for assignments and/or objectives
- have access to guidelines on University's definition of academic misconduct, e.g., plagiarism. (*They are strictly forbidden in this course*.)
- have reasonable access to grading instruments and/or grading criteria for assignments, tests, quizzes, or exams and to review graded material
- release the latest materials and announcements on Blackboard and engage in interactions with students, online or offline.

We hope you will…

- provide a positive, respectful, and engaged learning environment inside and outside the classes;
- appear for class meetings timely;
- appear at the mutual appointments for official teaching and learning matters;
- have full attendance at lectures, quizzes, in-class tests, and tutorials;
- get prepared for class, appearing with appropriate materials and having completed assigned readings and homework;

We hope you will (cont.)

- full engagement within the classes, including focus during lectures, appropriate and relevant questions, and class discussion participations;
- cover missed material during subsequent classes if having to miss a class due to emergent issues;
- act with integrity and honesty.

Learning Outcomes

- To understand **data analytics concepts**.
- To capture how to **manipulate**, **analyze**, and **visualize** data with analytics tools.
- To understand and apply related **mathematics operations**.

- **Keywords**: *probability, statistics, mathematics, R language programming*

Textbook

- NO official textbook. Course slides will be enough to work for assessments only.
- Recommended books (*for those want to learn more*):
  - Beecher, K., *Computational Thinking*, BCS, 2017.
  - Teetor, P., R Cookbook, O' Reilley Media, 2011.
  - Wickham, H. and Grolemund G., *R for Data Science*, O' Reilley Media, 2017.
  - Boyd, S. and Vandenbergerghe, L., *Introduction to Applied Linear Algebra*, Cambridge University Press, 2018.
  - Stewart, J., *Calculus: Early Transcendentals*, 8th Edition, Cengage Learning, 2015.

Instructor

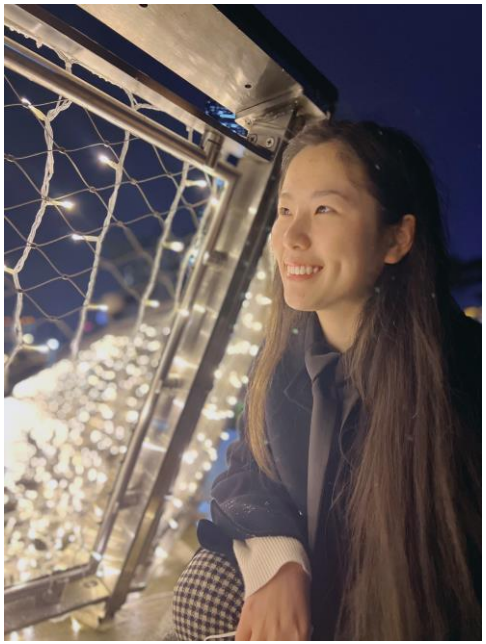
- Dr. Jing Li (Amelia)
  - Assistant Professor
  - Department of Computing
  - Homepage: http://www4.comp.polyu.edu.hk/~jing1li/
  - Email: jing-amelia.li@polyu.edu.hk
    - My connect email (jing1li@connect.polyu.hk) is a fake one!
  - Office: PQ 714
  - Office hour: Every Thursday 2:00 – 4:00 pm (or other time by appointment)

Teaching Assistants

- Mr. Da Ren
  - Office: QT 404
  - Email: da-cs.ren@connect.polyu.hk
- Mr. Chunpu Xu
  - Office: QT 415
  - Email: chun-pu.xu@connect.polyu.hk

Teaching Assistants

- Mr. Hanzhuo Tan
  - Email: hanzhuo.tan@connect.polyu.hk
- Ms. Yuji Zhang
  - Office: PQ 509
  - Email: yu-ji.zhang@connect.polyu.hk



Teaching Assistants

- Mr. Feiteng Mu
  - Office: PQ 723
  - Email: feitengmu.mu@connect.polyu.hk
- Mr. Xiaoyang Zhang
  - Office: PQ 503
  - Email: xiaoyang.zhang@connect.polyu.hk

Time and Venue (*Same content will be taught in the two classes!*)

- Two classes:
  - Class 1: Tuesday 12:30 pm – 3:20 pm
    - Lecture: 12:30 pm – 2:20 pm
    - Tutorial: 2:30 pm – 3:20 pm
  - Class 2: Thursday 8:30 am - 11:20 am
    - Lecture: 8:30 am - 10:20 am
    - Tutorial: 10:30 am - 11:20 am
- Location:
  - QR 403 (Class 1)
  - PQ 306 (Class 2)

Prerequisites

- Better if you have the following background*:*
  - *Probability and Statistics*
  - *Calculus*
  - *Linear Algebra*
  - *Programming* (even with simple languages)

No worries. *We'll provide the background with all the needed math and programming knowledge.*

Grade Assessment Scheme (1433)

- Quiz 1 (5%): Feb 8 (Tue) and Feb 10 (Thu)

  Quiz 2 (5%): Mar 29 (Tue) and Mar 31 (Thu)
- Assignment (20%): out Mar 24 and due Apr 7
- In-class test (25%): Mar 8 (Tue) and Mar 10 (Thu)
- Final Exam (45%): in the examination period

Grade Assessment Scheme (1003)

- Quiz 1 (15%): Feb 10 (Thu)
- Assignment (30%): out Mar 24 and due Apr 7
- In-class test (55%): Mar 10

What we will learn?

- Mathematical weapons for data analytics
  - Probability and Statistics
  - Calculus (differentiation and integration)
  - Linear Algebra (vector and matrix basics)
- Programming with R language
  - Basics: to get started!
  - Data Input and Manipulation
  - Statistics
  - Data Analytics

What we will learn? (cont.)

- Advanced Data Analytics
  - Monto-Carlo Simulation
  - Regression
  - Time-Series Analysis
  - Machine Learning

1	Jan 11 & 13	***Data Analytics: An Introduction***	Hybrid
2	Jan 18 & 20	***Probability Basics for Data Analytics***	Hybrid
3	Jan 25 & 27	***Statistics Basics for Data Analytics***	Onsite
4	Feb 8 & 10 (*quiz 1*)	***Linear Algebra Basics***	Onsite
5	Feb 15 & 17	***Calculus Basics***	Onsite
6	Feb 22 & 24	***Programming with R: Basics, Data Input and Manipulation***	Onsite
7	Mar 1 & 3	***Programming with R: Statistics***	Onsite
	Mar 8 & 10	***In-class test***	Onsite
8	Mar 15 & 17	***Data Analytics with R***	Onsite
9	Mar 22 & 24 (*assignment out*)	***Monto-Carlo Simulation***	Onsite
10	Mar 29 & 31 (*quiz 2*)	***Regression and Time-series Analysis***	Onsite
11	Apr 7 (*assignment due*)	***Machine Learning: An Introduction***	Onsite
12	Apr 12&14	***Review and Exam Q&A***	Onsite

Course Structure

Simulation

Regression

Advanced Data Analytics (3 lectures)

Machine Learning

Mathematical Basics (4 lectures)

R programming (4 lectures)

Probability

Statistics

Calculus

Linear Algebra

Environment

Data Manipulation

Data Analytics

Roadmap

- A quick look at this course
- **A tour of data analytics**
  - Introduction
  - Big Data and Data Mining

Roadmap

- A quick look at this course
- A tour of data analytics
  - **Introduction**
  - Big Data and Data Mining

What is Data?

- Individual units of information.
- A single quality or quantity of some object or pheno menon. In analytical processes, data are represented by variables (*why?*).
- Data is employed in scientific research, businesses m anagement (e.g., sales data, revenue, profits, stock pr ice), finance, governance (e.g., crime rates, unemploy ment rates, literacy rates), and in virtually every other form of human organizational activity.

--- From Wikipedia

Data is Everywhere!

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

Your First Impression

- Popular word/phrase choices
  - Statistics
  - Algorithms
  - Models
  - …

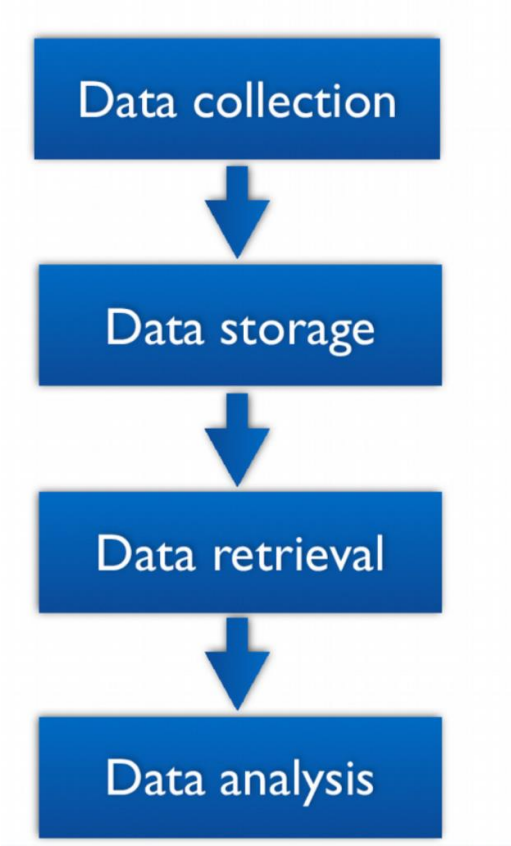
  Methodology

  - Big data
  - Data Mining

  Subcategory

What is Data Analytics

- **Data analysis** is a *process of inspecting, cleaning, transforming, and modeling* of data with the **goal** of *discovering useful information, informing conclusion, and supporting decision-making*.



--- From Wikipedia

What is Data Analytics

- **Characteristics**:
  - Data Driven (the more the better, *why*?)
  - Interdisciplinary (mathematics + computer science)
  - Discover Knowledge/Information from data

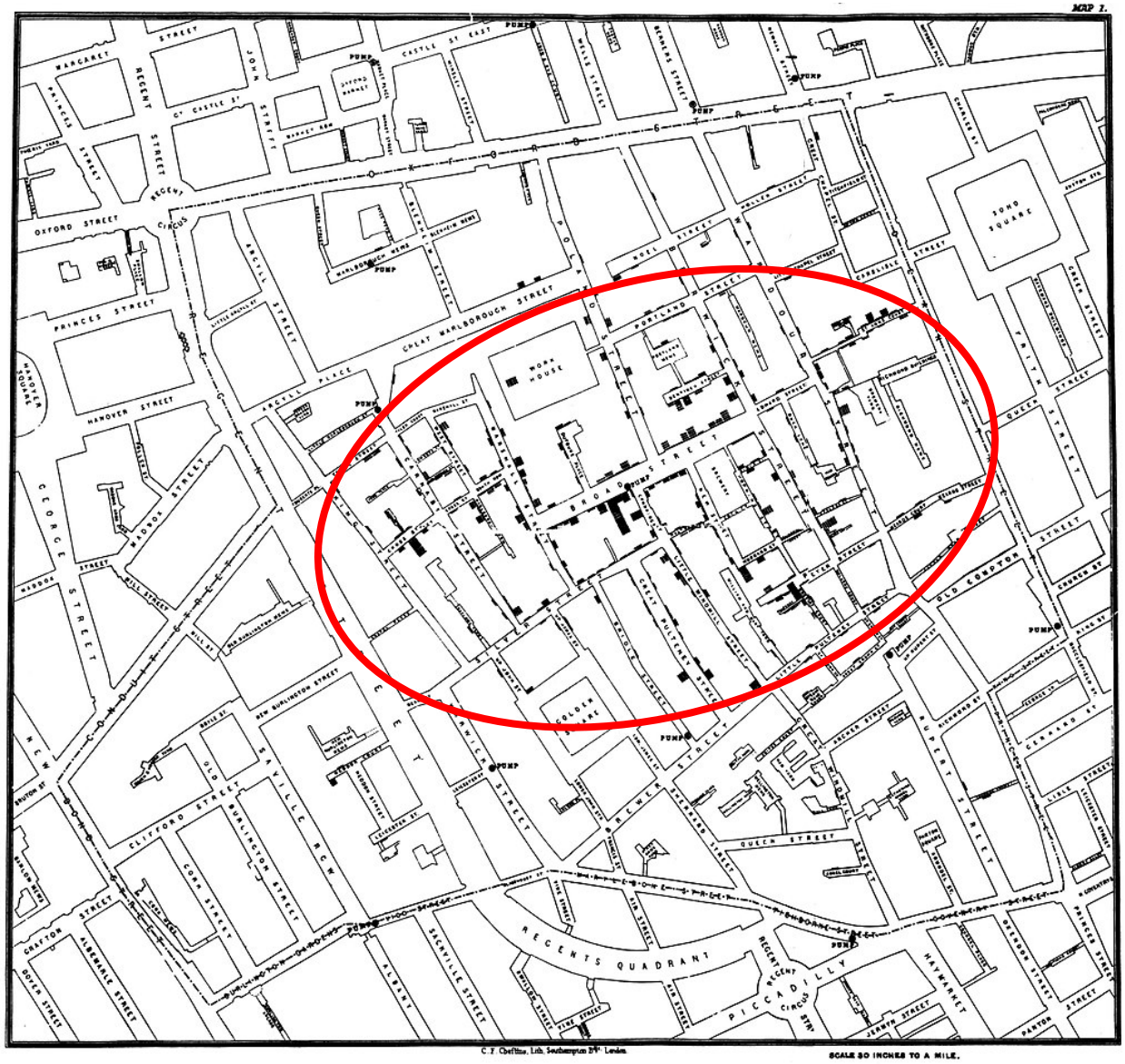
Small Sample Effects

- Subject results of COMP ????:
  - Three students
  - Continuous Assessments (CA) (40%): (D+, C+, A)
  - Final Exam (60%): (C, B, B)
- What is the mean of CA?
- What is the mean of final exam?
- What is the mean of overall grades?
- (**Hint**: all results should be converted to letter grades)
- Letter Grades vs. Grade Points:
  - A+: 4.5; A: 4.0; B+: 3.5; B: 3.0; C+: 2.5; C: 2.0; D+: 1.5; D: 1.0; F: 0.0

Discuss

A Real Story: Data Analytics Saves People's Life

A COURT FOR KING CHOLERA.


MAP I.
OXFORD STREET
REGENT STREET
REGENT CIRCUS
PRINCES STREET
HANOVER STREET
GREAT MARLBOROUGH STREET
ARGYLL PLACE
ARGYLL STREET
WORK HOUSE
BROAD STREET
POLAND STREET
PORTLAND STREET
GREAT PULTENEY STREET
GOLDEN SQUARE
SOHO SQUARE
GEORGE STREET
MADDOX STREET
NEW BURLINGTON STREET
CLIFFORD STREET
OLD BURLINGTON STREET
SAVILE ROW
CORK STREET
QUEEN STREET
REGENTS QUADRANT
PICCADILLY
REGENT CIRCUS
HAYMARKET
OLD COMPTON STREET
QUEEN STREET
PUMP
C.F. Cheffins, Lith. Southampton Bldgs. London.
SCALE 30 INCHES TO A MILE.

Lessons from Data Analytics

- Spawn new data analytics projects
  - Weather prediction
  - Physics research (supercollider data analytics)
  - Astronomy images (planet detection)
  - Medical research (drug interaction)
  - ***Can you show me more examples?***
- Businesses latched onto its techniques, methodologies, and objectives

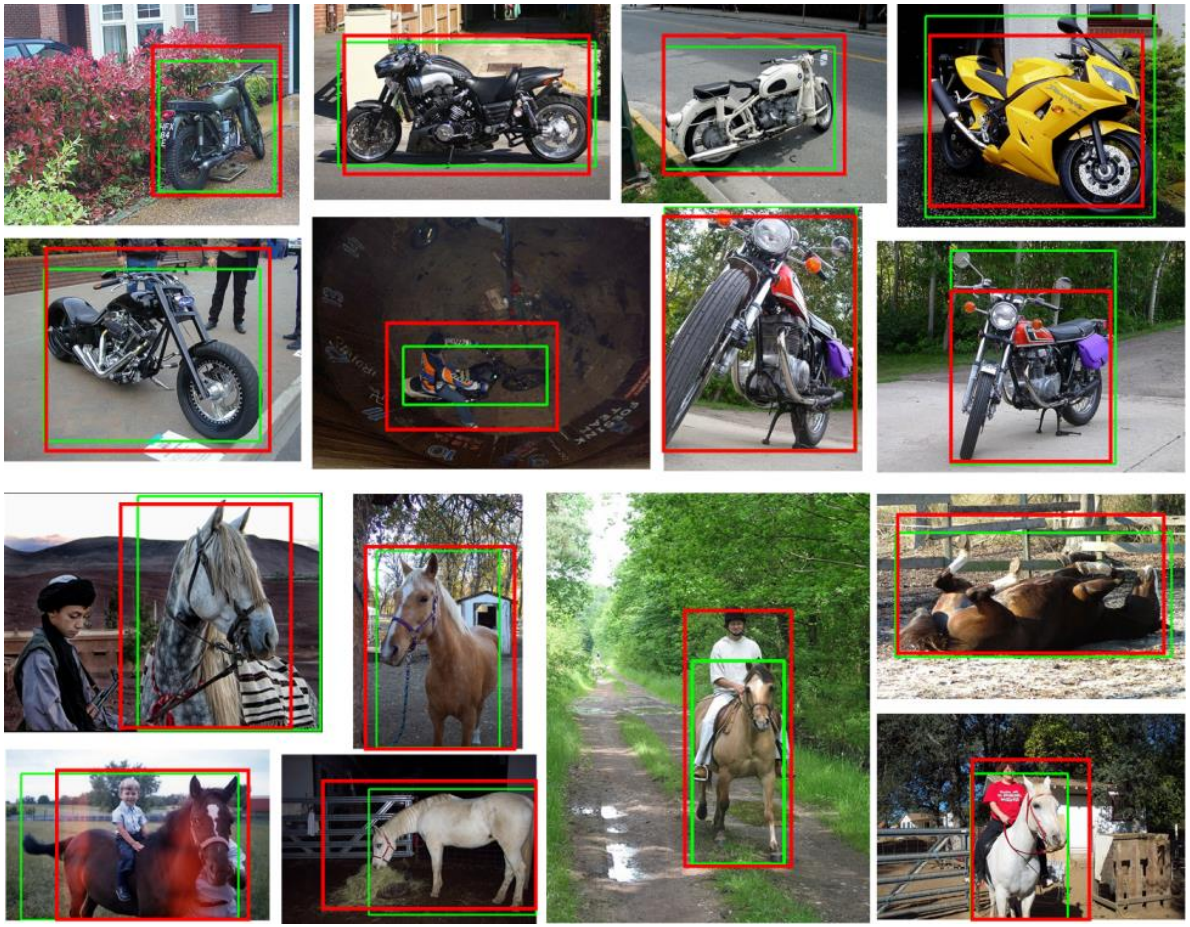
Discuss

Types of Analytics at eBay

- Basically measure anything possible - A **few** examples:

Marketing	Buyer Experience	Finance	Trust & Safety
Technology Operations	Customer Service	Loyalty	Information Security
Infrastructure	Finding	User Behavior	Seller Experience

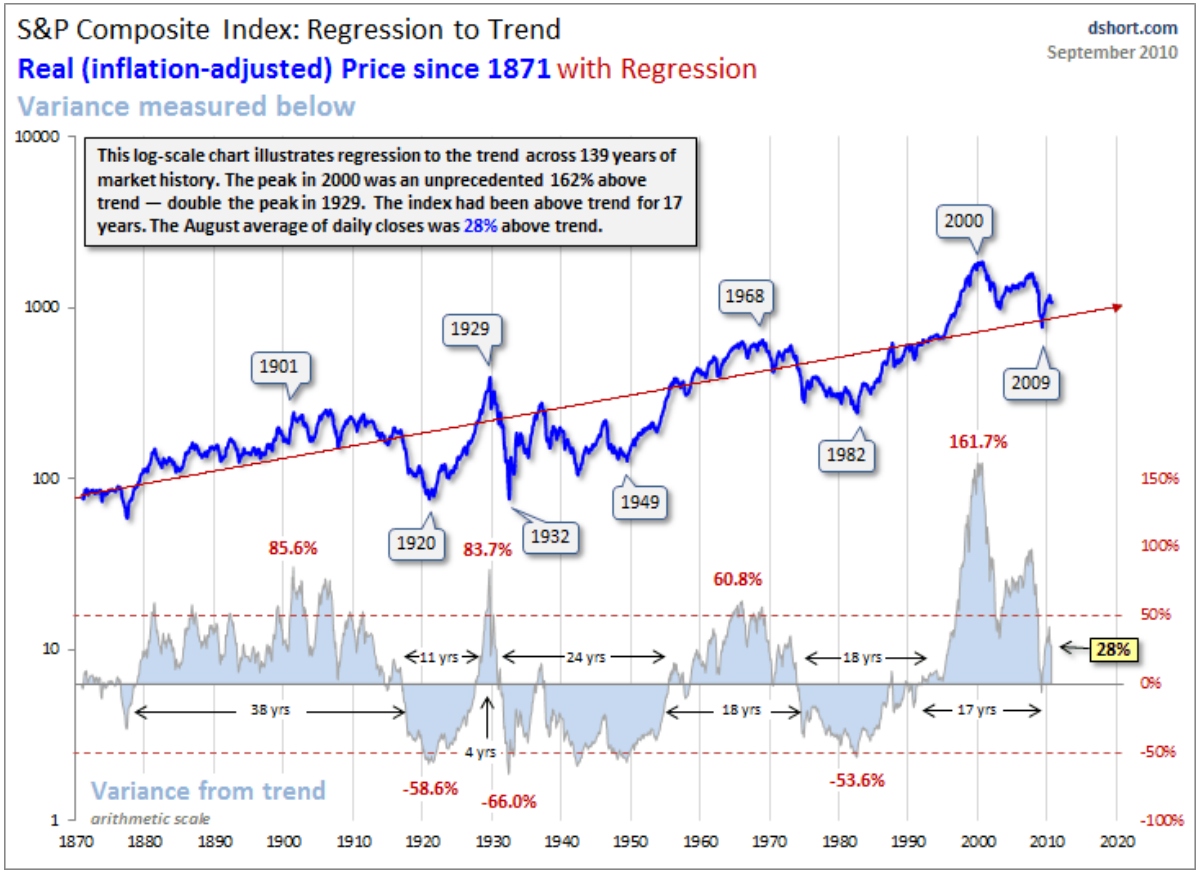
Examples: Classification



What are these objects?

Horses or *Vehicles*?

Example: Regression

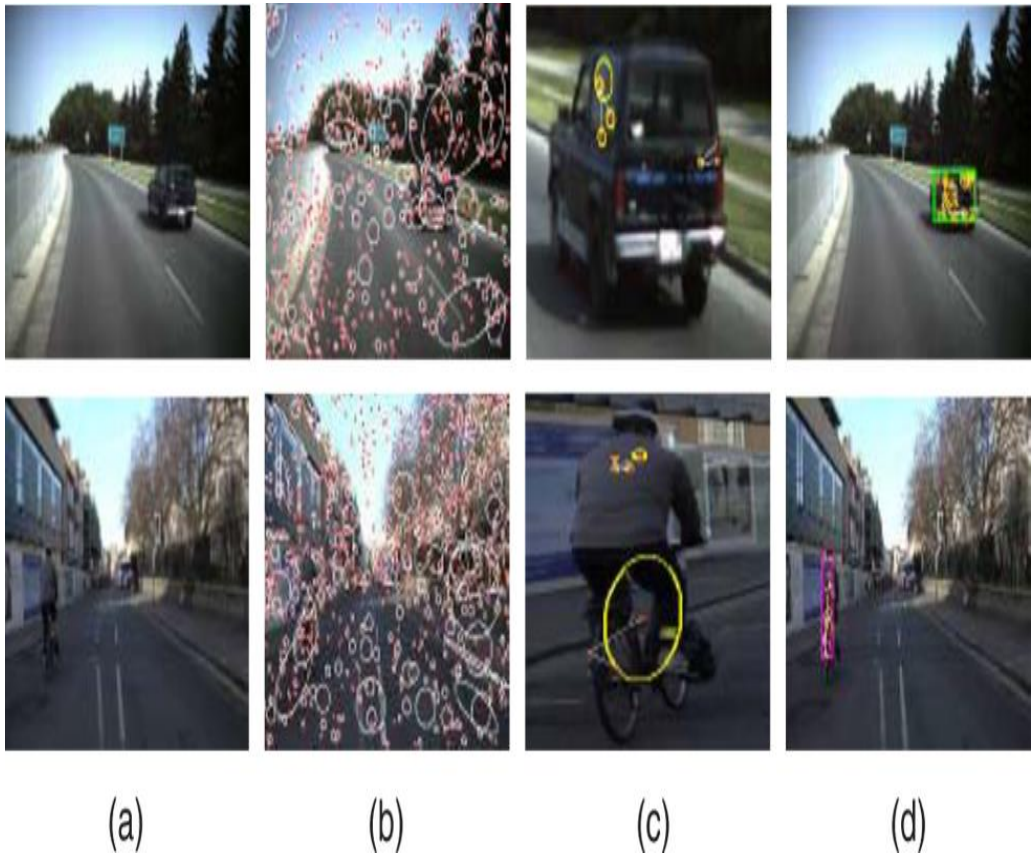


How do we know the trend of price?

Increase or *Decrease*?

How fast? Comparison?

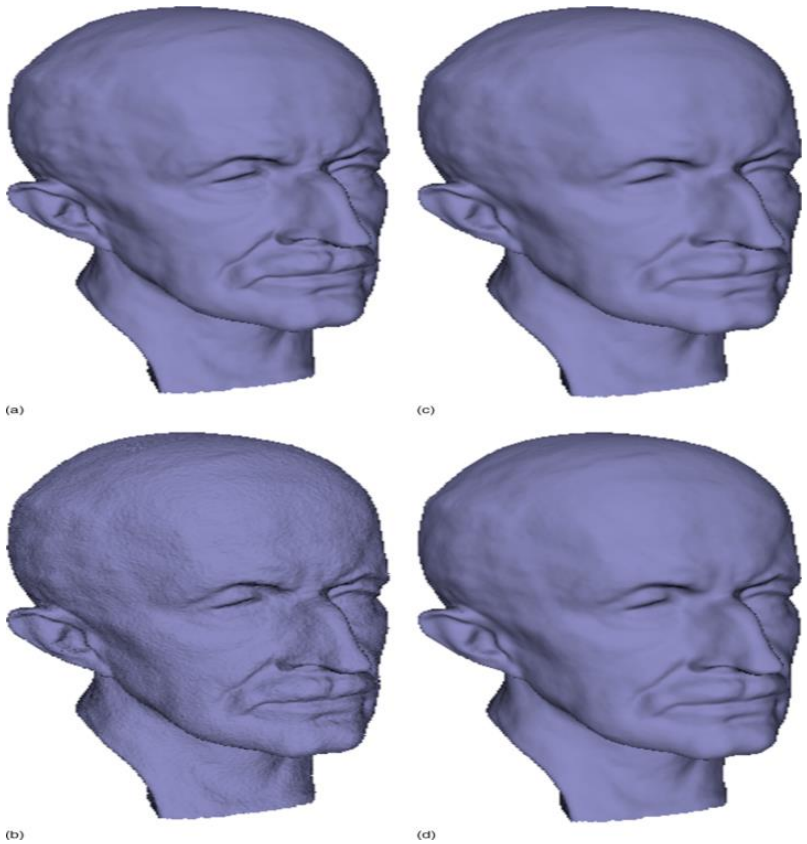
Example: Clustering



How to segment regions?

Pixels in similar colors with their neighbors (e.g., to segment *sky*).

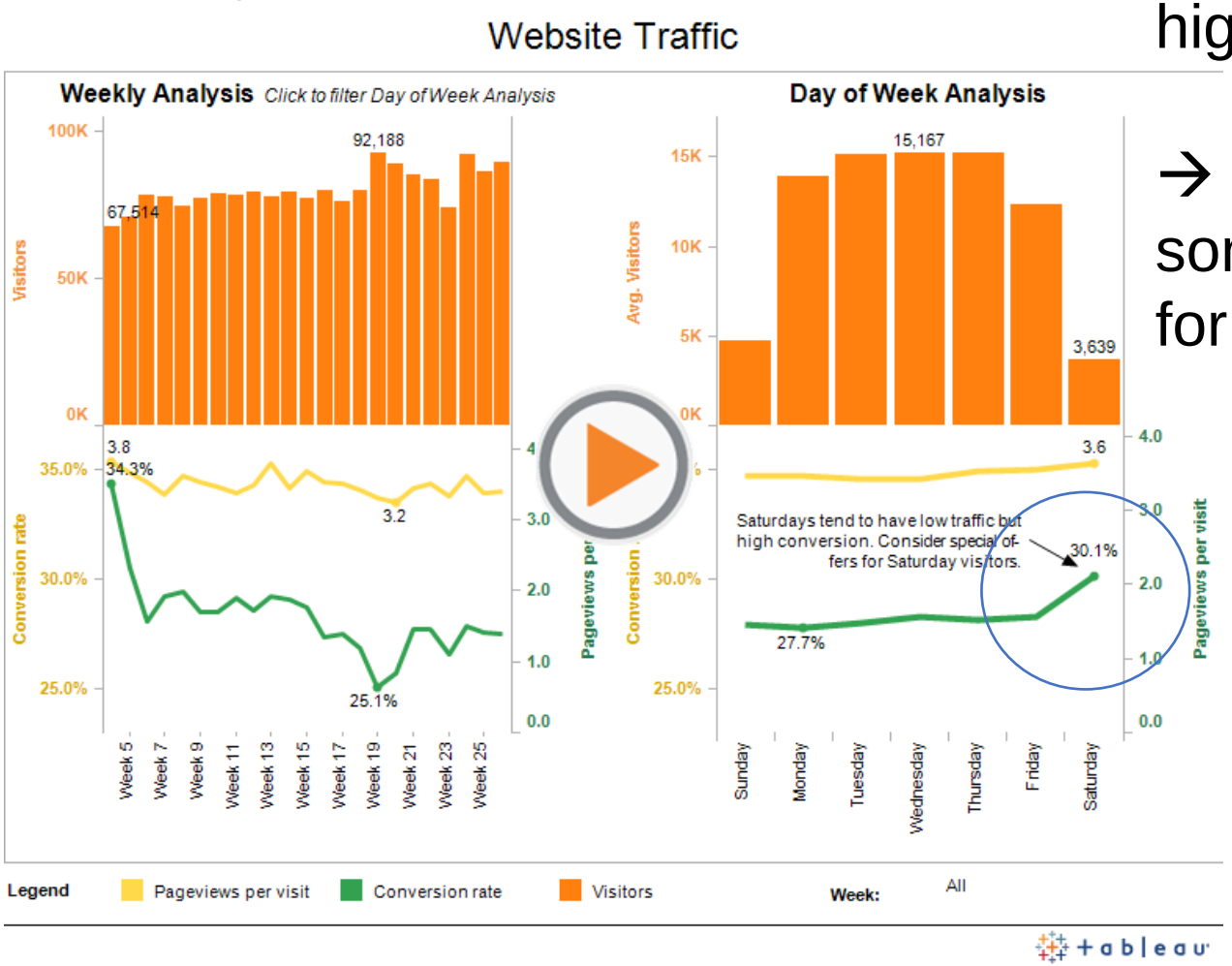
Example: Similarity Matching



Are they collected from the same guy?

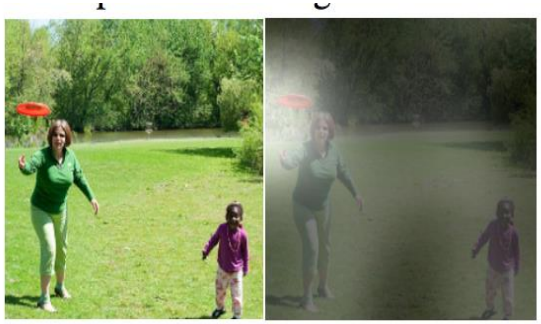
Example: Time Series Analysis

Saturday tends to have low traffic but high conversion.



→ Provide something special for Saturday visitors.

Example: Image Captioning



A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

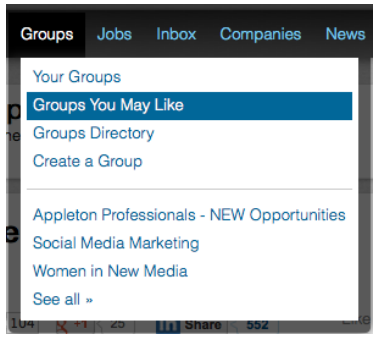
A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

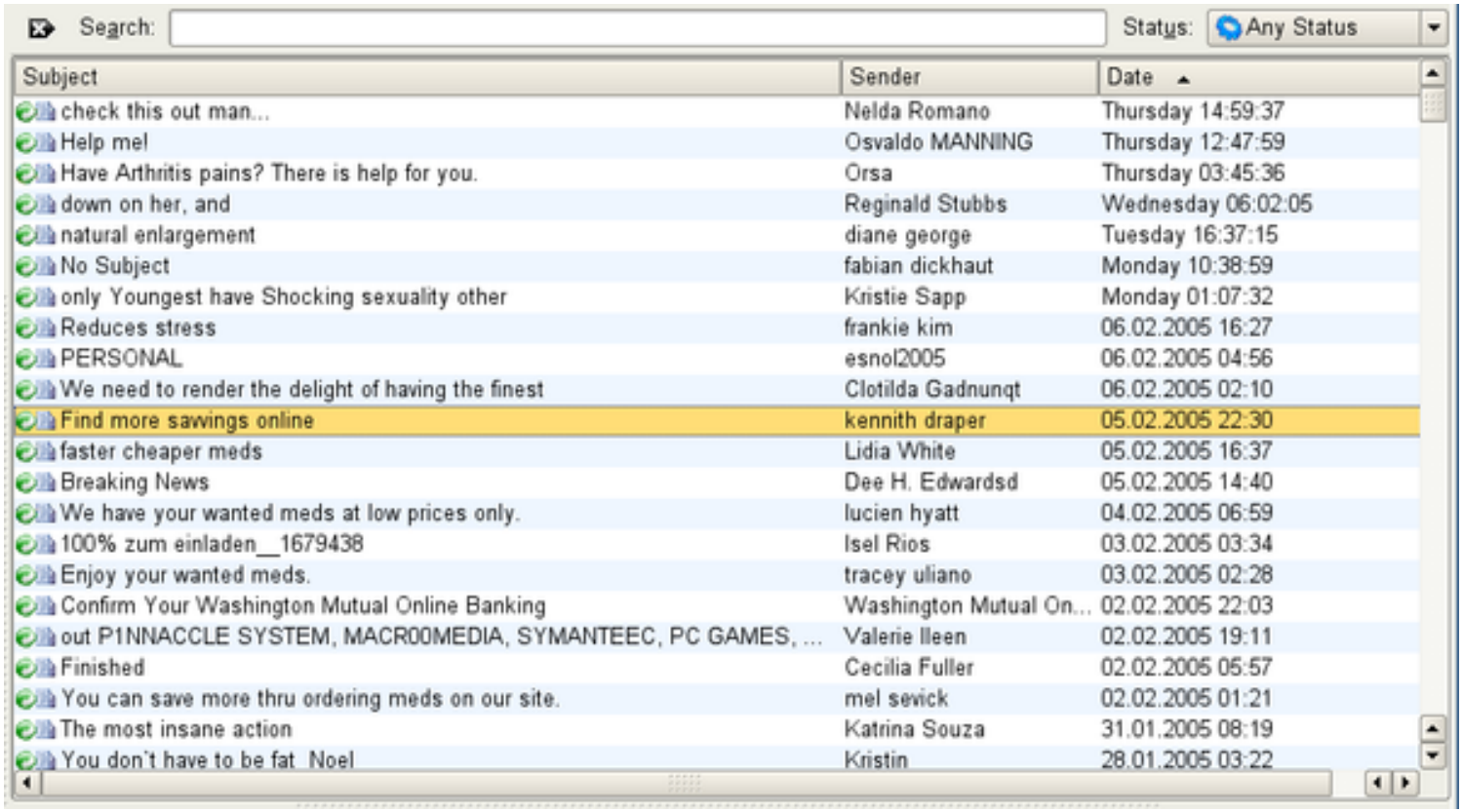
A giraffe standing in a forest with trees in the background.

Example: 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: Spam Detection



Search: Status: Any Status

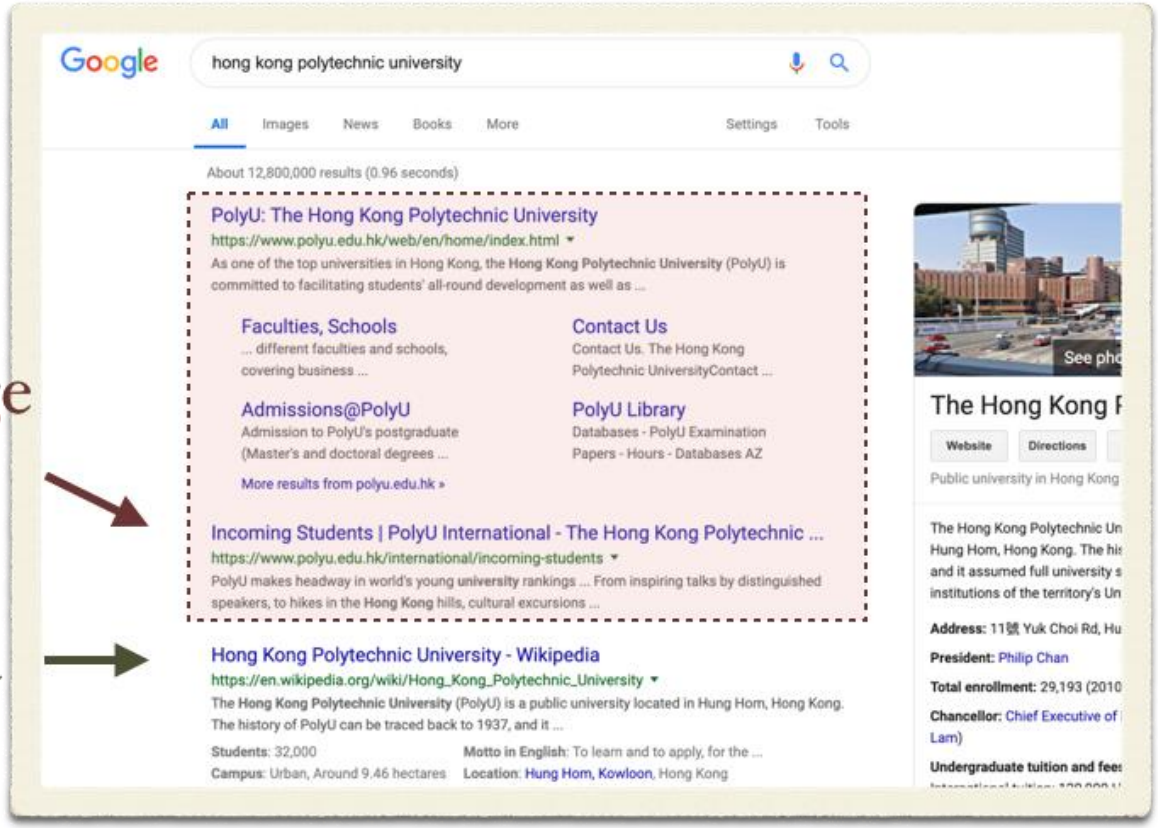
Subject	Sender	Date
check this out man...	Nelda Romano	Thursday 14:59:37
Help me!	Osvaldo MANNING	Thursday 12:47:59
Have Arthritis pains? There is help for you.	Orsa	Thursday 03:45:36
down on her, and	Reginald Stubbs	Wednesday 06:02:05
natural enlargement	diane george	Tuesday 16:37:15
No Subject	fabian dickhaut	Monday 10:38:59
only Youngest have Shocking sexuality other	Kristie Sapp	Monday 01:07:32
Reduces stress	frankie kim	06.02.2005 16:27
PERSONAL	esnol2005	06.02.2005 04:56
We need to render the delight of having the finest	Clotilda Gadnunqt	06.02.2005 02:10
Find more savings online	kennith draper	05.02.2005 22:30
faster cheaper meds	Lidia White	05.02.2005 16:37
Breaking News	Dee H. Edwardsd	05.02.2005 14:40
We have your wanted meds at low prices only.	lucien hyatt	04.02.2005 06:59
100% zum einladen__1679438	Isel Rios	03.02.2005 03:34
Enjoy your wanted meds.	tracey uliano	03.02.2005 02:28
Confirm Your Washington Mutual Online Banking	Washington Mutual On...	02.02.2005 22:03
out P1NNACCLE SYSTEM, MACR00MEDIA, SYMANTEEC, PC GAMES, ...	Valerie lleen	02.02.2005 19:11
Finished	Cecilia Fuller	02.02.2005 05:57
You can save more thru ordering meds on our site.	mel sewick	02.02.2005 01:21
The most insane action	Katrina Souza	31.01.2005 08:19
You don't have to be fat Noel	Kristin	28.01.2005 03:22

Example: Ranking of Webpages

- Computing importance of webpages.

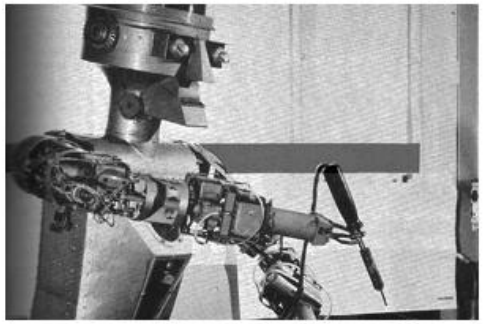
Homepage

Wikipedia

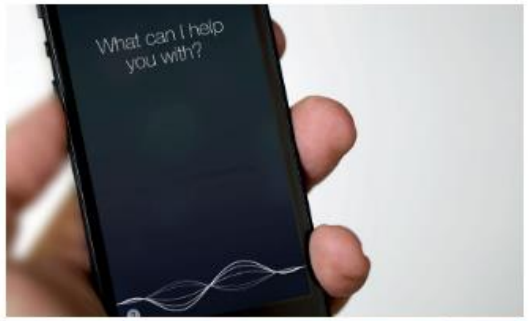


Example: Artificial Intelligence (AI)

1950s-1990s



2000s-2010s



? ? ?

Roadmap

- A quick look at this course
- A tour of data analytics
  - Introduction
  - **Big Data and Data Mining**

Why data science achieves huge success in recent years?

- 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!

How Much Data?

- IDC reports
  - 2.7 billion terabytes in 2012, up 48 percent from 2011
  - 8 billion terabytes in 2015
- Sources
  - Structured corporate databases
  - Unstructured data from webpages, blogs, social networking messages, …
  - Countless digital sensors
- Volume
  - Google processes 20 PB (10^{15}) a day of user-generated data
  - Facebook
    - 2.5B - content items shared
    - 2.7B - 'Likes'
    - 300M - photos uploaded
    - 100+PB - disk space in a single HDFS cluster
    - 105TB - data scanned via Hive (30min)
    - 70,000 - queries executed
    - 500+ TB (10^{12}) - new data ingested

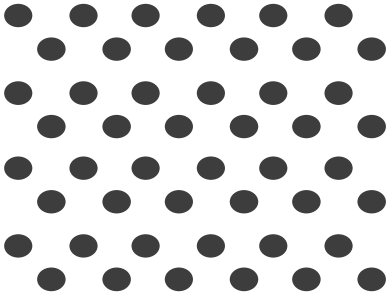
Related Concept: Big Data

- Big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.

--- From Wikipedia

Characteristics of Big Data: 4V

Volume

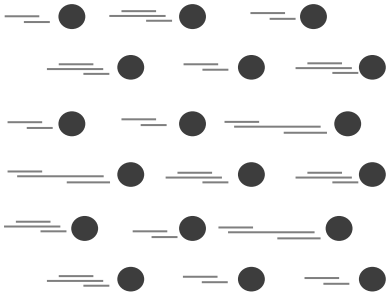


From terabytes to exabyte to zetabytes of existing data to process

IDC Analyze the Future

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

Velocity

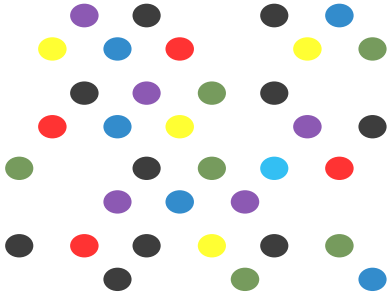


Batch data, real-time data, streaming data, milliseconds to seconds to respond

facebook

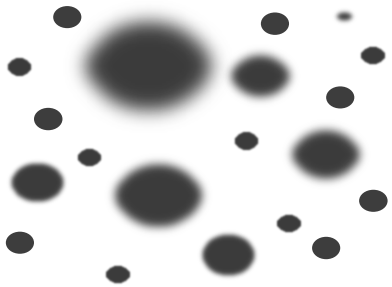
New sharing over 2.5 billion per day
new data over 500TB per day

Variety



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

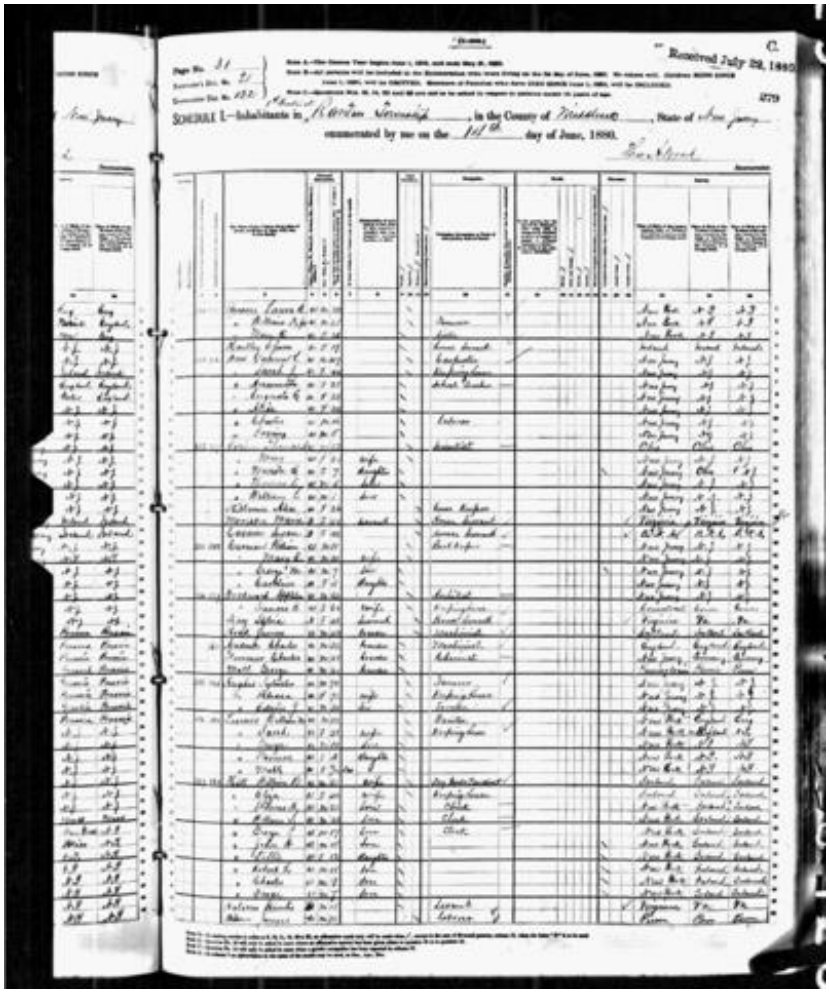
Veracity



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

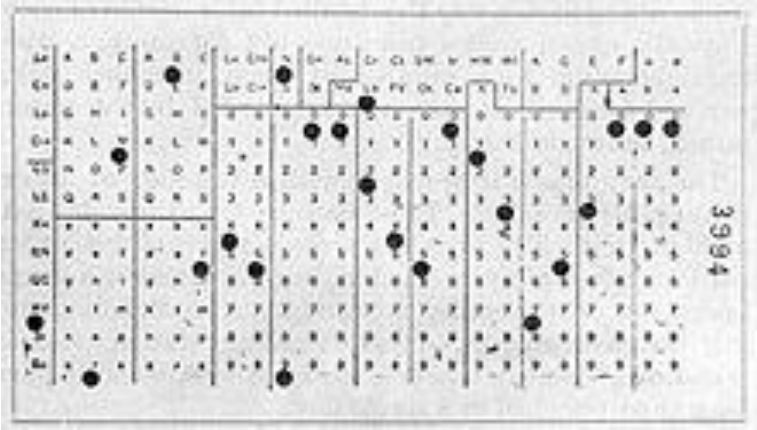
The First Big Data Challenge

- 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

What is Data Mining?

- Discovery of patterns and models that are:
  - Valid: hold on new data with some certainty
  - Useful: should be possible to act on the item
  - Unexpected: non-obvious to the system
  - Understandable: humans should be able to interpret the pattern
- A particular data analytic technique

Data Mining Tasks

- Descriptive Methods
  - Find human-interpretable patterns that describe the data
  - E.g., Beers and Diapers
- Predictive Methods
  - Use some variables to predict unknown or future values of other variables
  - E.g., weather prediction.

Relation between Data Mining and Data Analytics

- Analytics include both data analysis (mining) and communication (guide decision making)
- Analytics is not so much concerned with individual analyses or analysis steps, but with the entire methodology

One Slide to Takeaway

- What is the structure of this course?
- What are data analytics?
- Examples of data analytics?
- What is big data?
- What is data mining?