Introduction to Data Science

(Lecture 1)

Dr. Mohammad Pourhomayoun

Assistant Professor

Computer Science Department

California State University, Los Angeles

What is Data Science?

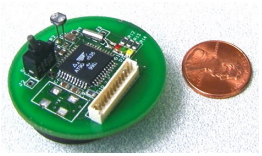
What is Data Science?

- **Data Science** is an interdisciplinary field of research that aims to design and develop automated or semi-automated techniques to **extract knowledge (information) from large-scale data** and use it for **future purposes** such as **prediction, decision making,** or **recommendation**.
- It can be an integration of machine learning, statistics, big data processing, predictive analytics, and computing.

- **Question**: What is the difference between “knowledge” and “data”?





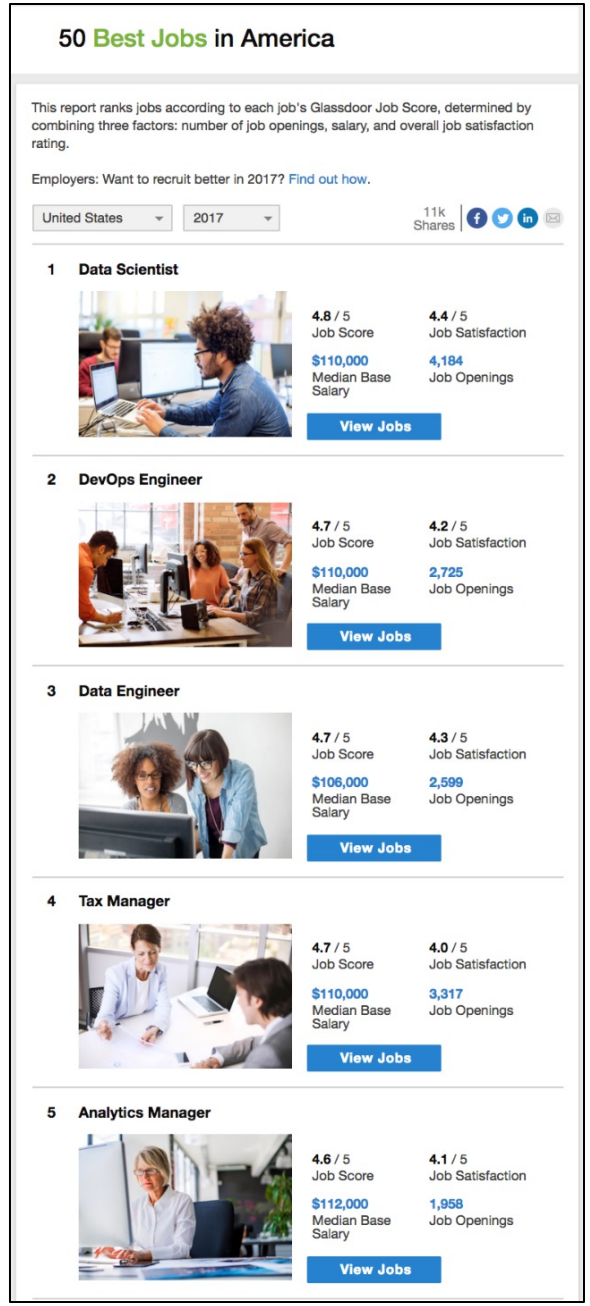

What is Data Science?

- Fortune Magazine:
  - "The Hot New Gig in Tech!"
- The New York Times:
  - "This hot new field promises to revolutionize industries, from business to government, healthcare to academia."
- Fortune Magazine
  - "Companies that want to make sense of all their bits and bytes are hiring so-called **data scientists** – **if they can find any!**"

Who is a Data Scientist?

- **Glassdoor:**
  - "**Data Scientist**" has been rated **#1 in the list of Best Jobs in America since 2016**
  - In this list, the jobs are determined by combining three key factors:
    - **number of job openings**
    - **salary**
    - **career opportunities rating**

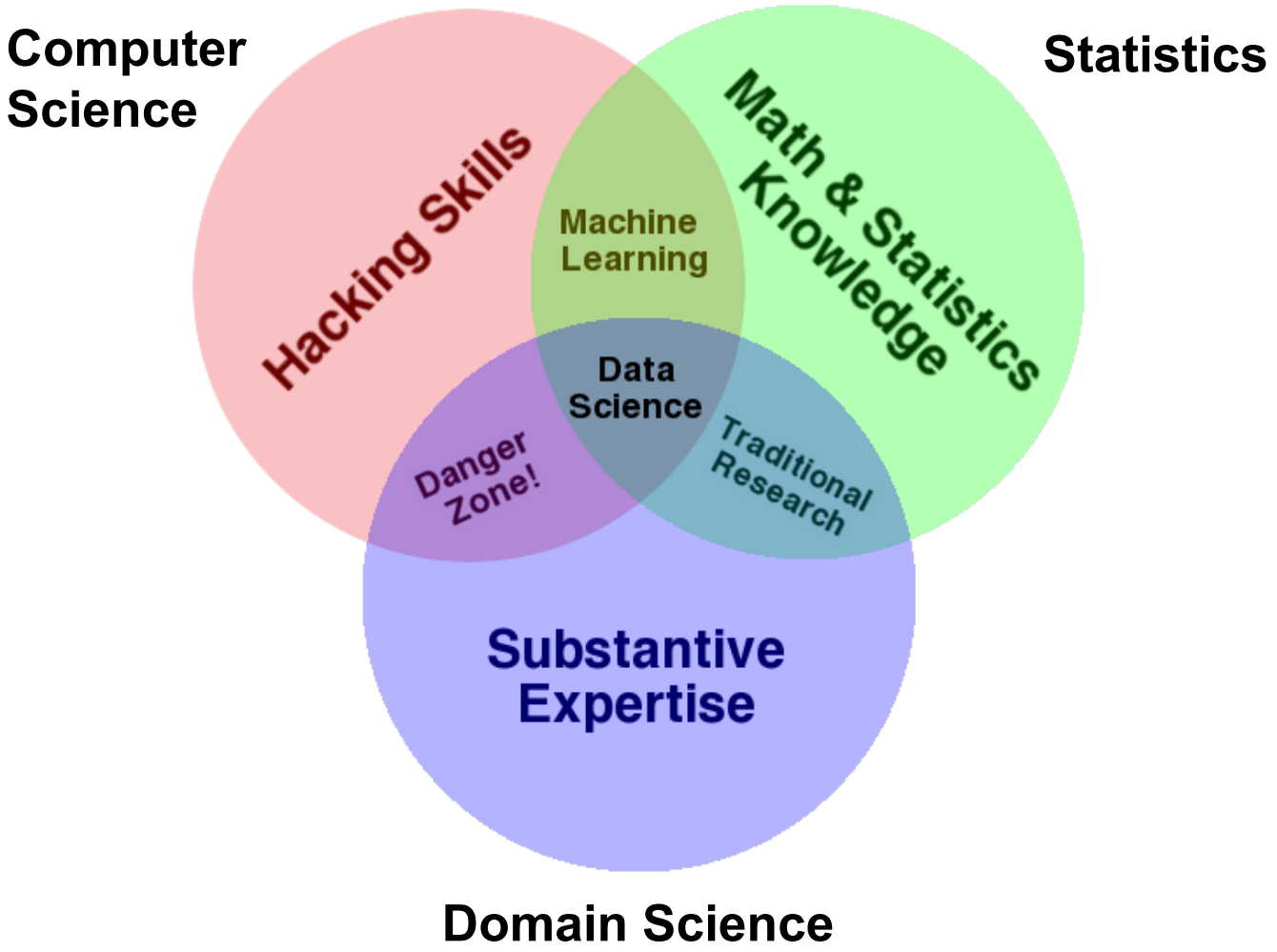
[Ref]: www.glassdoor.com/List/Best-Jobs-in-America-2019-LST_KQ0,25.htm



Who is a Data Scientist?

- McKinsey Global Institute:
  - The United States alone could face a shortage of 1.5 million managers and analysts with the knowledge of how to use the analysis of big data to make effective decisions.
  - The report estimates that there will be 4 to 5 million jobs in the U.S. requiring data analysis skills.

- Drew Conway, CEO and founder of Alluvium, Venn Diagram:



Why Now?

Why Is Data Science So Important Now?

- Why is Data Science an important topic these days?

  (why didn't anyone talk about it 10 years ago?)

- Because now we have:
  1. New Sources of Data that did not exist before.
  2. New Capabilities to acquire, store, and process data.
  3. New Algorithms and Methods to analyze data.

New Sources of Data

- Social Networks: Facebook, Twitter, …
- World Wide Web
- Online Activities: Amazon, ebay, …
- Smart Phone Activities
- Electrical Health Records (EHR)
- Body and wearable sensors
- …

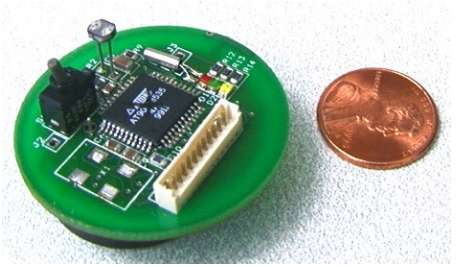
New Sources of Data

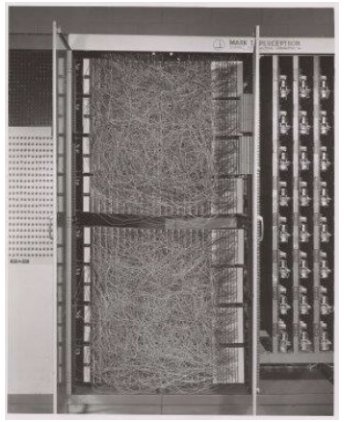
- "There was 5 exabytes of information ($5x10^{18}$ bytes) created between the **dawn of civilization through 2003**, but that much information is now created **every two days**."

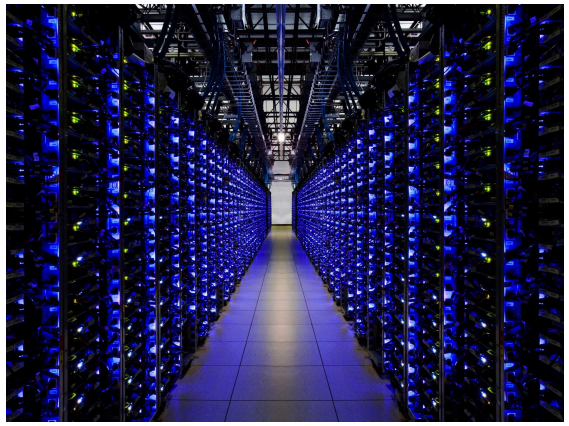
- Eric Schmidt, Google, Alphabet

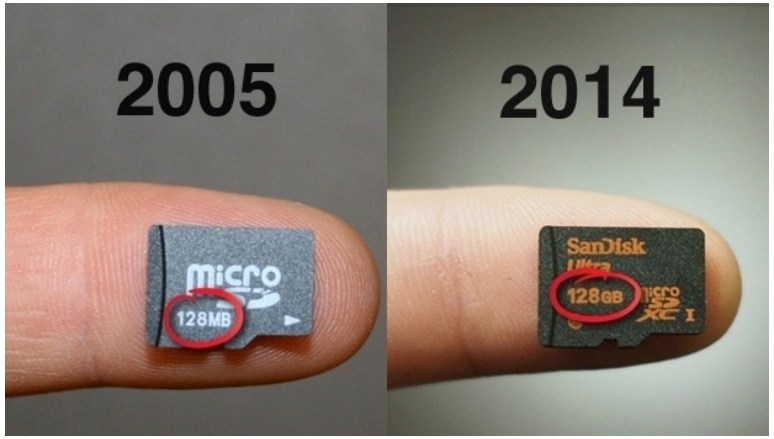
New Capabilities to Acquire, Process & Store Data

- The cost of data acquisition, processing, and storing data has dropped incredibly fast thanks to advances in Electronics and other technologies.
- Do you know how Google works?









CS4661: Course Overview

- **Course Overview**: In this course, we will cover the fundamental topics of Data Science, focusing on the main ***algorithms*** and ***tools*** in Data Processing, Machine Learning, Data Analytics, Big Data Manipulation.
- We will present both theoretical and practical aspects of these methods.
- **The course consists of three main topics:**
  1. Introduction and Applications of Data Science
  2. Machine Learning and Data Analytics
  3. Big Data Processing and Manipulation

Course Overview

- **Instructor**: Dr. Mohammad Pourhomayoun
  - **Email:** mpourho@calstatela.edu
  - **Class:** Section1: Fri, 12:00PM ; Section2: Thu, 1:40PM.
  - **Office Hours:** Thu 12:30PM-1:30PM, and Fri 11:00AM-12:00PM via zoom (link in Canvas).

- **TAs:**
  - **Juya Ahmadi:** jahmadi2@calstatela.edu
    - **Office hours: 10:00 - 12:00AM on Wednesday.**
  - **Amir Ebrahimi:** aebrahi9@calstatela.edu
    - **Office hours: 12:00 - 1:00PM on Monday.**
  - **Ryan Dunning:** ryandunning57@gmail.com
    - **Office hours: 4:00 - 5:00PM on Tuesday.**
- **Canvas:** All course material including slides and homeworks will be provided through Canvas.

Evaluation

- Homework Assignments: 30%
  - Theoretical Problems, Implementation, Programming
  - **Late submissions will not be accepted**!
  - Copying homework/project from others is considered as cheating and is not tolerable!
- Quiz: 25%
- Final Project (group project): 20%
- Final Exam: 25%
- Participation Bonus: 5-10%

- In this course, we work with **Python 3.x** and its libraries.

- **Textbooks**: There will be **no required textbooks**, though we suggest the following optional references if you are interested:

1. Abu-Mostafa, Yaser, Malik Magdon-Ismail, and Hsuan-Tien Lin. Learning from Data, 2012.
2. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, An Introduction to Statistical Learning, 2014.
3. Ian H. Witten, etc., Data Mining: Practical Machine Learning Tools and Techniques, 2011.
4. Barber, David. Bayesian Reasoning and Machine Learning. Cambridge University Press, 2012.
5. Downey, Allen B. Think Bayes. O'Reilly Media, 2013.
6. Foreman, John W. DataSmart: Using Data Science to Transform Information into Insight. Wiley, 2013.
7. Han, Jiawei, Micheline Kamber, and Jian Pei. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2011.
8. Murphy, Devin P. Machine Learning: A Probabilistic Perspective. MIT Press, 2012.
9. O'Neil, Cathy, and Rachel Schutt. Doing Data Science: Straight Talk from the Frontline. O'Reilly Media, 2013.
10. Rajaraman, Anand, and Jeffrey David Ullman. Mining of Massive Datasets. Cambridge University Press, 2012.
11. Ratner, Bruce. Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data. CRC Press, 2011.
12. Richert, Willi, and Luis Pedro Coelho. Building Machine Learning Systems with Python. Packt Publishing, 2013.
13. Siegel, Eric. Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die. Wiley, 2013.
14. Ian Goodfellow , Yoshua Bengio , et al., Deep Learning (Adaptive Computation and Machine Learning series), 2016.
15. Joel Grus, Data Science from Scratch: First Principles with Python, 2019.
16. Ani Adhikari, John DeNero, Computational and Inferential Thinking: The Foundations of Data Science, 2019.
17. Jake VanderPlas, Python Data Science Handbook: Essential Tools for Working with Data, 2016.

In this class…

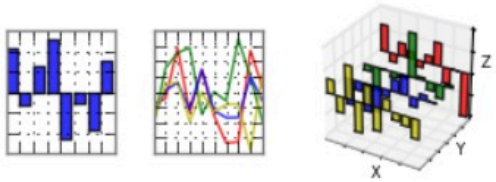
- Feel free to participate in class discussions …, There is nothing to be shy about! **WE ARE ALL FRIENDS!** ☺
- Your question is important! Please feel free to ask!
- Don’t hesitate to interrupt when you have a question.
- Please just let me know if you want me to repeat or clarify something.
- Your Feedback is precious to me!

Python Programming

IP[y]: IPython
Interactive Computing

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



scikits learn
machine learning in Python

NumPy

SciPy.org Sponsored By ENTHOUGHT

matplotlib

Why Python?

- **Python** is very powerful and highly popular for Data Science purposes. We can name it the main programming language for Data Science!
- **Python** is an easy-to-learn, widely used, and general-purpose programming language.
- **Python** includes unique powerful libraries designed for data science.
- **Python** supports both object-oriented and procedural programming styles (*both styles are acceptable in this class*).

Python Programming

- For those of you who don't know Python:
  - No Worries!!! ☺
  - Python is easy-to-learn and user friendly!!
  - We need python just as much as you can do your homework and projects.
  - In this class, we briefly review the basics of python programming for beginners.
  - There are hundreds of excellent free references and resources to learn more.

Python Programming

- In this class, we will briefly review the basics of python programming for beginners, and then we will more focus on python libraries (such as scikit-learn machine learning library, …).
- For Python beginners, we highly recommend to boost your programming skills using at least one of the following references or hundreds of other free tutorials available online:
  - Python Doc: https://docs.python.org/3.7/tutorial/index.html
  - Google's Python class: https://developers.google.com/edu/python
  - …

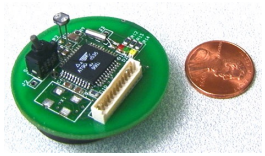
Data Science:
Introduction and Applications

What is Data Science?

- **Data Science** is an interdisciplinary field of research that aims to design and develop automated or semi-automated techniques to **extract knowledge from large-scale data** and use it for **future purposes** such as making **prediction, decision,** or **recommendation**.

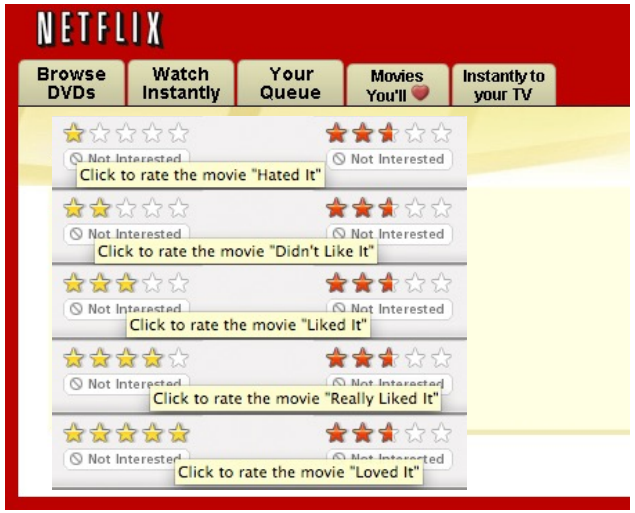
- It can be an integration of statistics, machine learning, big data processing, predictive analytics, and computing.

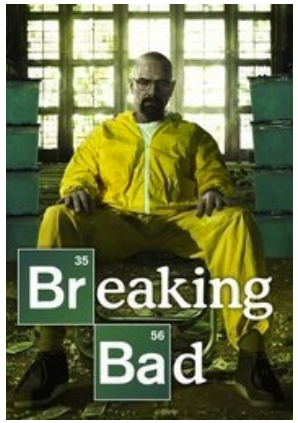





Example: Recommender System

Netflix Prize: $1,000,000 in an open competition for the best algorithm to predict user ratings for films, based on previous ratings without any other information about the users or films.





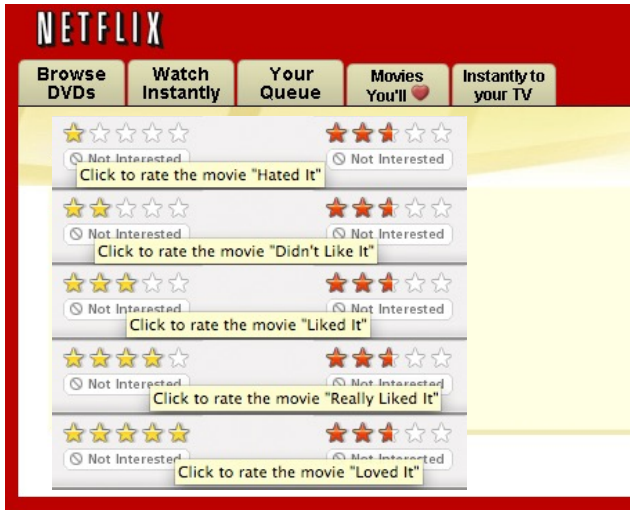
+

+

?????

Example: Recommender System

Netflix Prize: $1,000,000 award for the best algorithm to predict user ratings for movies, based on previous ratings.



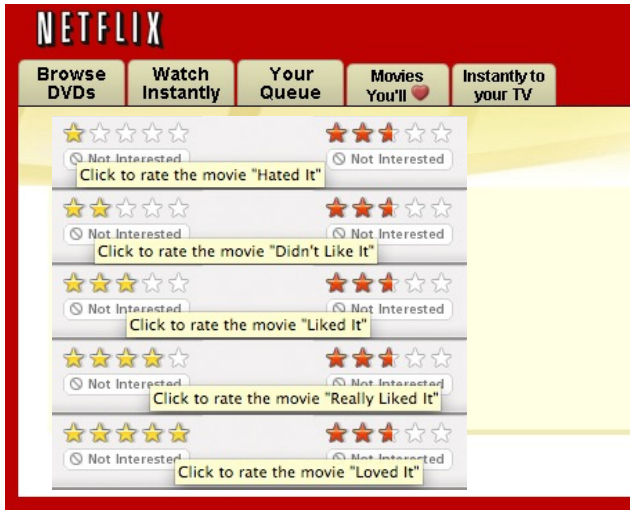
How does it work!!?

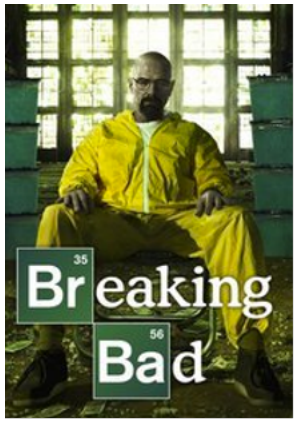
Challenges:

- Massive Data (100M ratings from 480K users for 18K movies)
- High Dimensionality: extremely complicated set of factors that affect people's ratings such as actors, directors, genre, ...
- Missing Data: 99% of data missing.

Example: Recommender System

- The grand prize of $1,000,000 was given to the *BellKor's Pragmatic Chaos* team, which could improve Netflix's original prediction algorithm by 10.06%.





+

+

?????

Example: Recommender System

amazon

- **Customers Who Bought This Item Also Bought:**

+

+

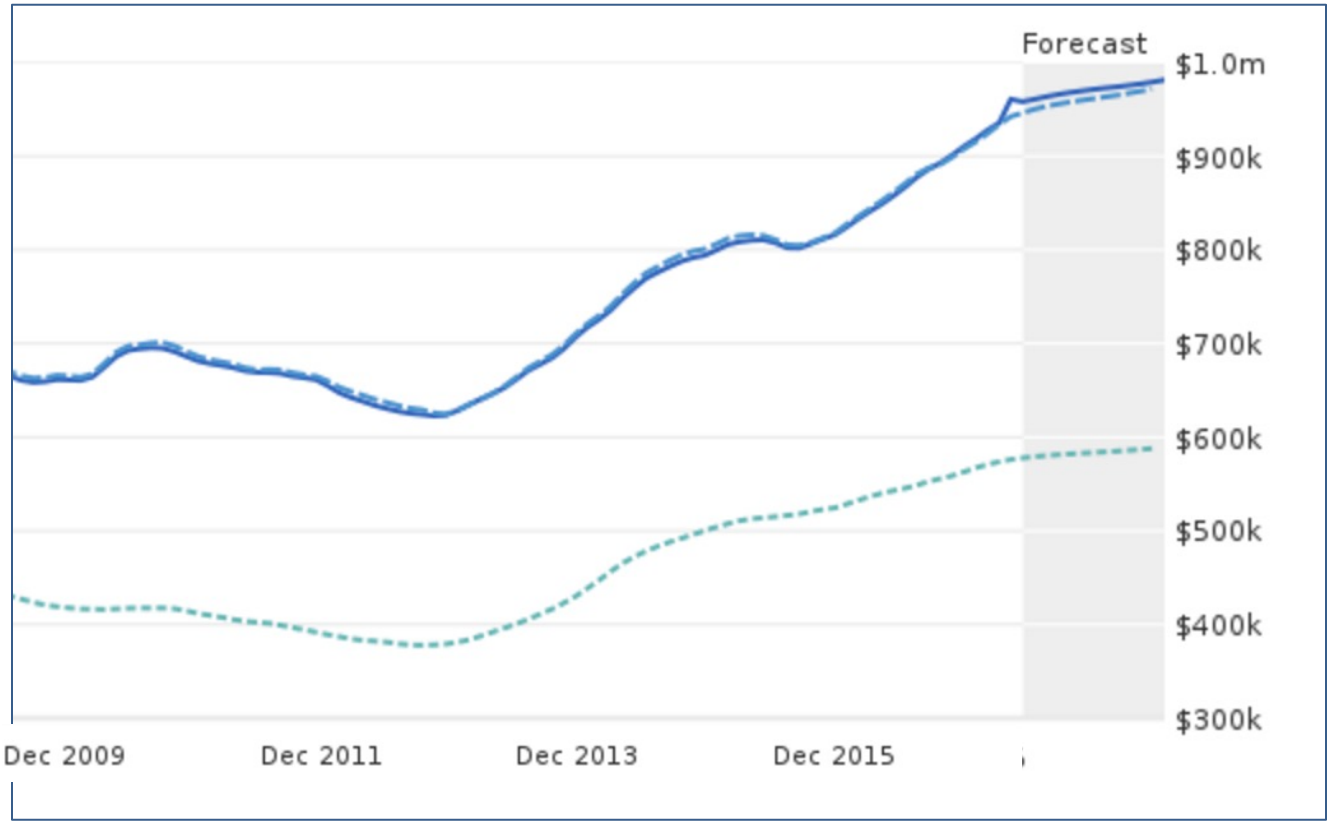
Example: Speech Recognition and Natural Language Processing (NLP)

- Apple Siri
- Amazon Alexa (www.youtube.com/watch?v=xenOYWVwkGY)
- Google Home
- Google Duplex (www.youtube.com/watch?v=D5VN56jQMWM&t=86s)

Example: Stock Market Prediction

Example: Real State Prediction

REDFIN™ Zillow®

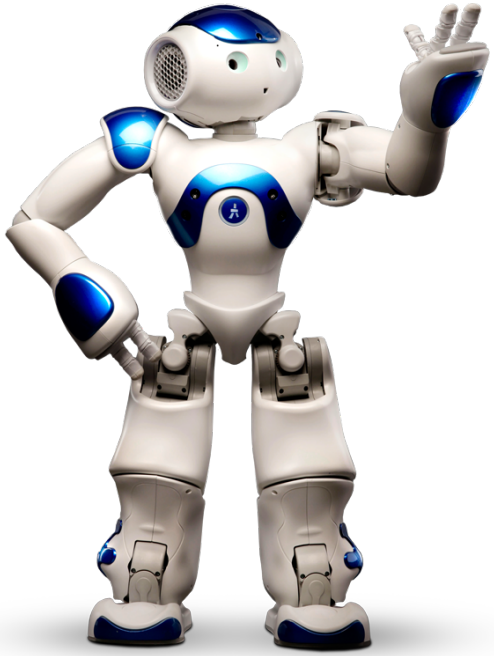


Example: Robotics

- Every Robot needs a Brain, and this brain is a Computer with data science capabilities that make it act like a human. It can see, hear, learn, and perform many of the human brain functionalities such as automatic decision making!

Cozmo

Dash-Dot



NAO

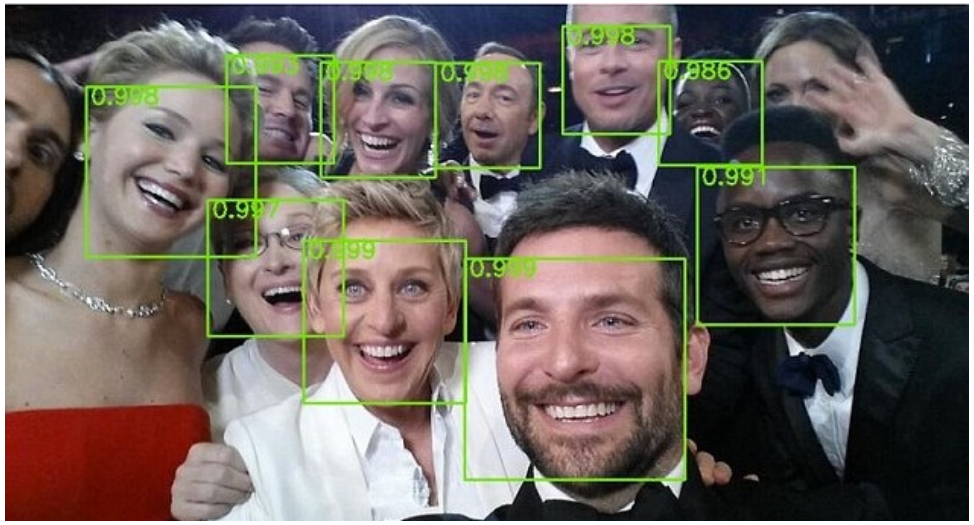
Example: Spam Detection

- The algorithm will learn from previous data (previous emails) that what combination of words indicate suspicious emails.

Example: Social Media

Example: Object Recognition

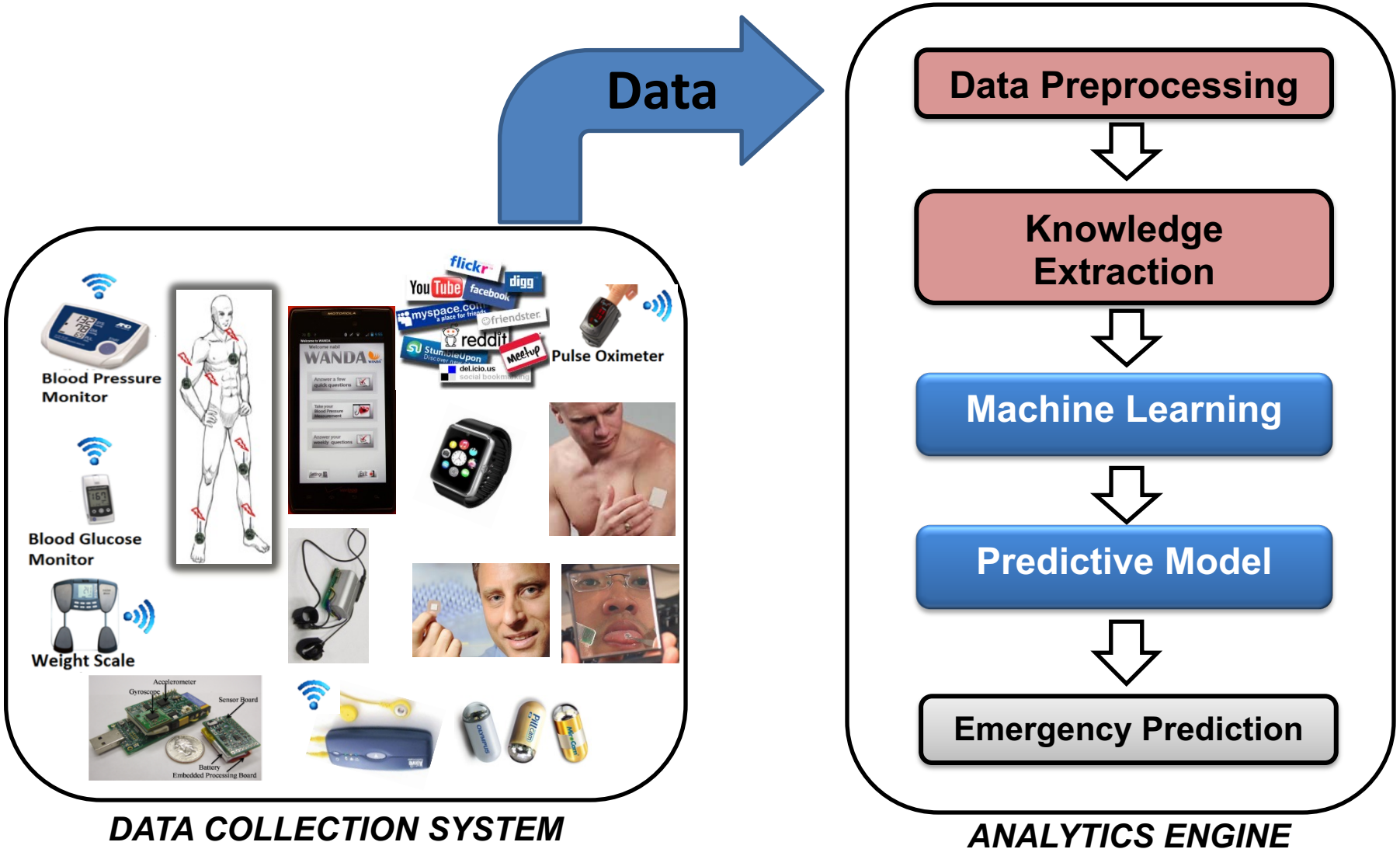
Example: Object Recognition, Face Recognition



Example: Self-Driving Cars

Watch this:
www.youtube.com/watch?v=B8R148hFxPw

Example: Medical Emergency Prediction



Data Science at Cal State LA

Visit our Data Science Research Lab at:

www.calstatela.edu/research/data-science

My Important Advice

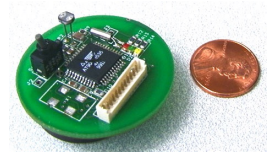
- **My Advice: Use Data Science, Machine Learning, and AI to help and benefit people and society for good!**

[Figure Ref]: UN.

Ingredients

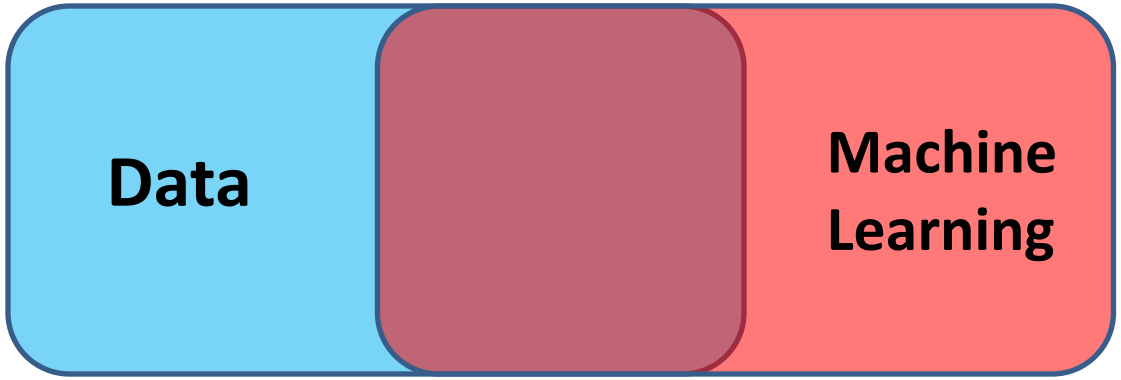
Ingredients

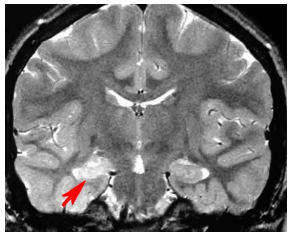
- **Data:** Rapid growth of massive datasets
  - E.g. WWW, Social Networks, Online Activities, Smart Phone, Wearables, Sensor networks, Science, …

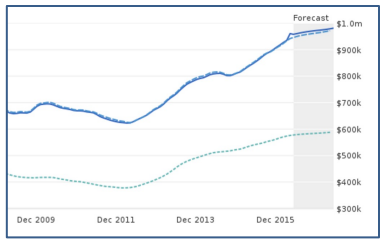


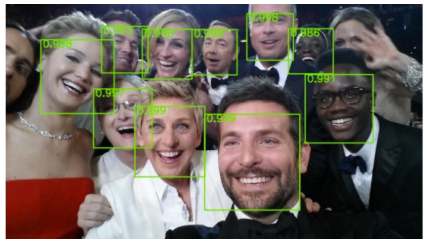
Ingredients

- **Machine Learning:** It is applied Everywhere:
  - E.g., recommendation system, market prediction, speech recognition, Face detection, Fraud detection, Spam filtering, vehicle control, Medical diagnosis, ...



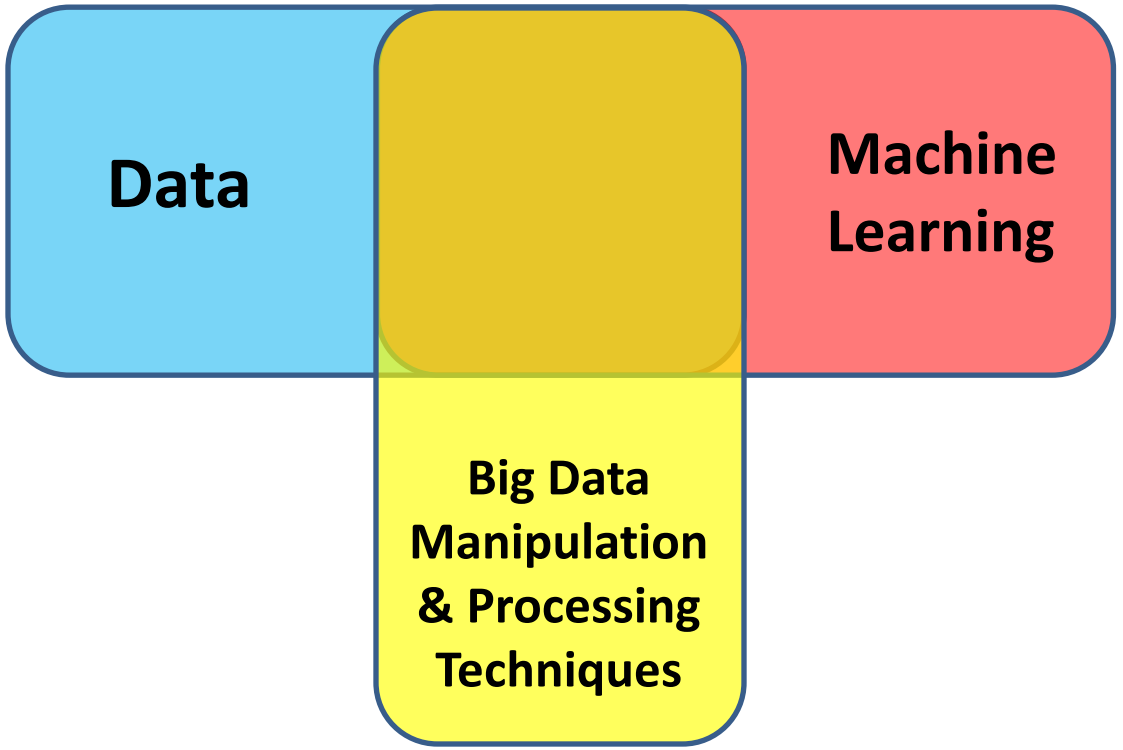






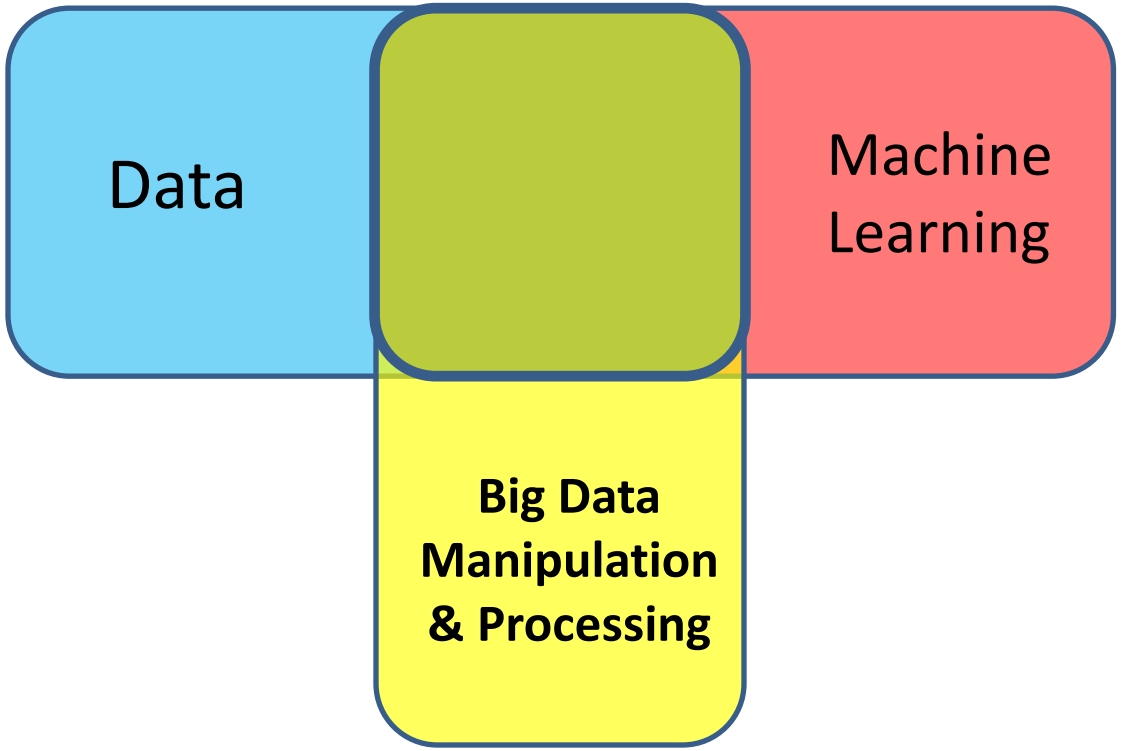
Ingredients

- **Big Data Manipulation & Processing:**
  - Large-Scale Data Processing, Distributed Computing, Cloud Computing



Ingredients

- Data, Machine Learning Algorithms, Big Data Manipulation Techniques



Thank You!

Questions?