



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-LST_KQ0,20.htm

50 Best Jobs in America

This report ranks jobs according to each job's Glassdoor Job Score, determined by combining three factors: number of job openings, salary, and overall job satisfaction rating.

Employers: Want to recruit better in 2017? [Find out how.](#)

United States 2017 11K Shares    

1 Data Scientist



4.8 / 5
Job Score
\$110,000
Median Base Salary

4.4 / 5
Job Satisfaction
4,184
Job Openings

[View Jobs](#)

2 DevOps Engineer



4.7 / 5
Job Score
\$110,000
Median Base Salary

4.2 / 5
Job Satisfaction
2,725
Job Openings

[View Jobs](#)

3 Data Engineer



4.7 / 5
Job Score
\$106,000
Median Base Salary

4.3 / 5
Job Satisfaction
2,599
Job Openings

[View Jobs](#)

4 Tax Manager



4.7 / 5
Job Score
\$110,000
Median Base Salary

4.0 / 5
Job Satisfaction
3,317
Job Openings

[View Jobs](#)

5 Analytics Manager



4.6 / 5
Job Score
\$112,000
Median Base Salary

4.1 / 5
Job Satisfaction
1,958
Job Openings

[View Jobs](#)

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 by 2018.





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

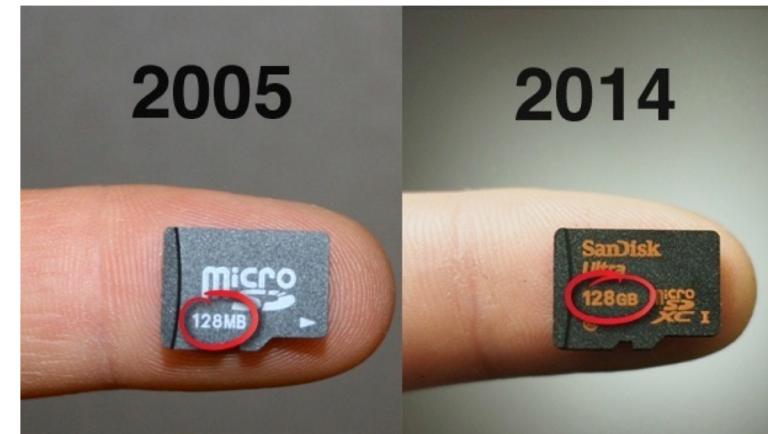
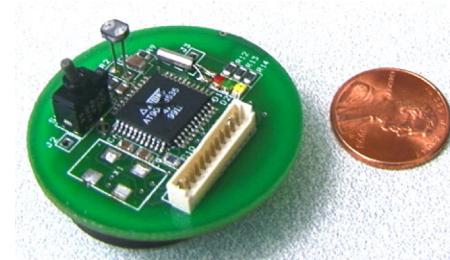
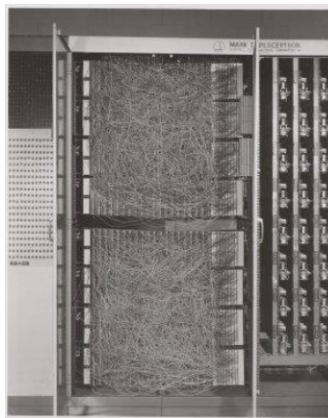
- “There was 5 exabytes of information (5×10^{18} bytes) created between the **dawn of civilization through 2003**, but that much information is now created **every two days**.”



- 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 fundamentals of Data Science, focusing on the main *algorithms* and *tools* in Data Processing, Machine Learning, Data Analytics, Big Data Manipulation, and Large-Scale Predictive Analytics (as time permits!).
- 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



Data Science at Cal State LA

CS4661: Introduction to Data Science (Fall)

CS5661: Advanced Topics in Data Science (Spring)

CS4662: Advanced Machine Learning (Spring)



Course Overview

- **Instructor:** Mohammad Pourhomayoun
 - **Email:** mpourho@calstatela.edu
 - **Class:** Friday, 11:00AM – 1:30PM, ASCB 132
 - **Office Hours:** Tuesday 10:00AM-12:00PM , **Office:** ET, A408
- **TAs:**
 - **Erika Estrada Medina:**
 - **Email:** eestra57@calstatela.edu
 - **Office Hours:** Monday & Thursday 10:30AM-12:00PM, **Office:** ET, A310
 - **Daniel Chang:**
 - **Email:** dchang33@calstatela.edu
 - **Office Hours:** Fridays 8:30AM - 11:00AM, **Office:** ET, A310
- **CSNS:** All course material including slides and homeworks will be provided through CSNS.



Course Overview

- **Textbooks:** There will be no required textbooks, though we suggest the following optional references if you are interested:
 1. Abu-Mostafa, Yaser, etc. *Learning from Data*. AML Book.
 2. Bishop, Christopher M. *Pattern Recognition and Machine Learning*.
 3. Ian H. Witten, etc., *Data Mining: Practical Machine Learning Tools and Techniques*.
 4. Conway, Drew. *Machine Learning for Hackers*. O'Reilly Media.
 5. Foreman, John W. *DataSmart: Using Data Science to Transform Information into Insight*. Wiley.
 6. Siegel, Eric. *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die*. Wiley.
 7. Howe, Bill. *Introduction to Data Science*. Coursera/Washington University.
 8. Andrew, Ng. *Machine Learning*. Coursera/Stanford University.



Evaluation

- Homework Assignments: 40%
 - Theoretical Problems, Implementation, Programming
 - **Late submissions will not be accepted!**
 - Copying homework/project from others is considered as cheating and is not tolerable!
- Final Project (group project): 20%
- Final Exam: 40%
- Participation Bonus: 5-10%
- In this course, we work with **Python 3.x** and its libraries.



In this class...

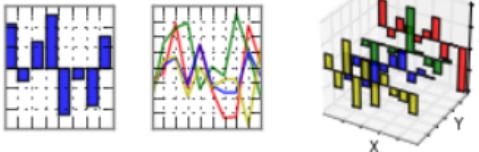
- Feel free to participate in class discussions ..., There is nothing to be shy about! **WE ARE ALL FRIENDS!** 😊
- Your question is important! Please feel free to ask!
- Don't hesitate to interrupt when you have a question.
- Please just let me know if you want me to repeat or clarify something.
- Your 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>
 - Python for Informatics: www.pythontutorial.net
 - ...





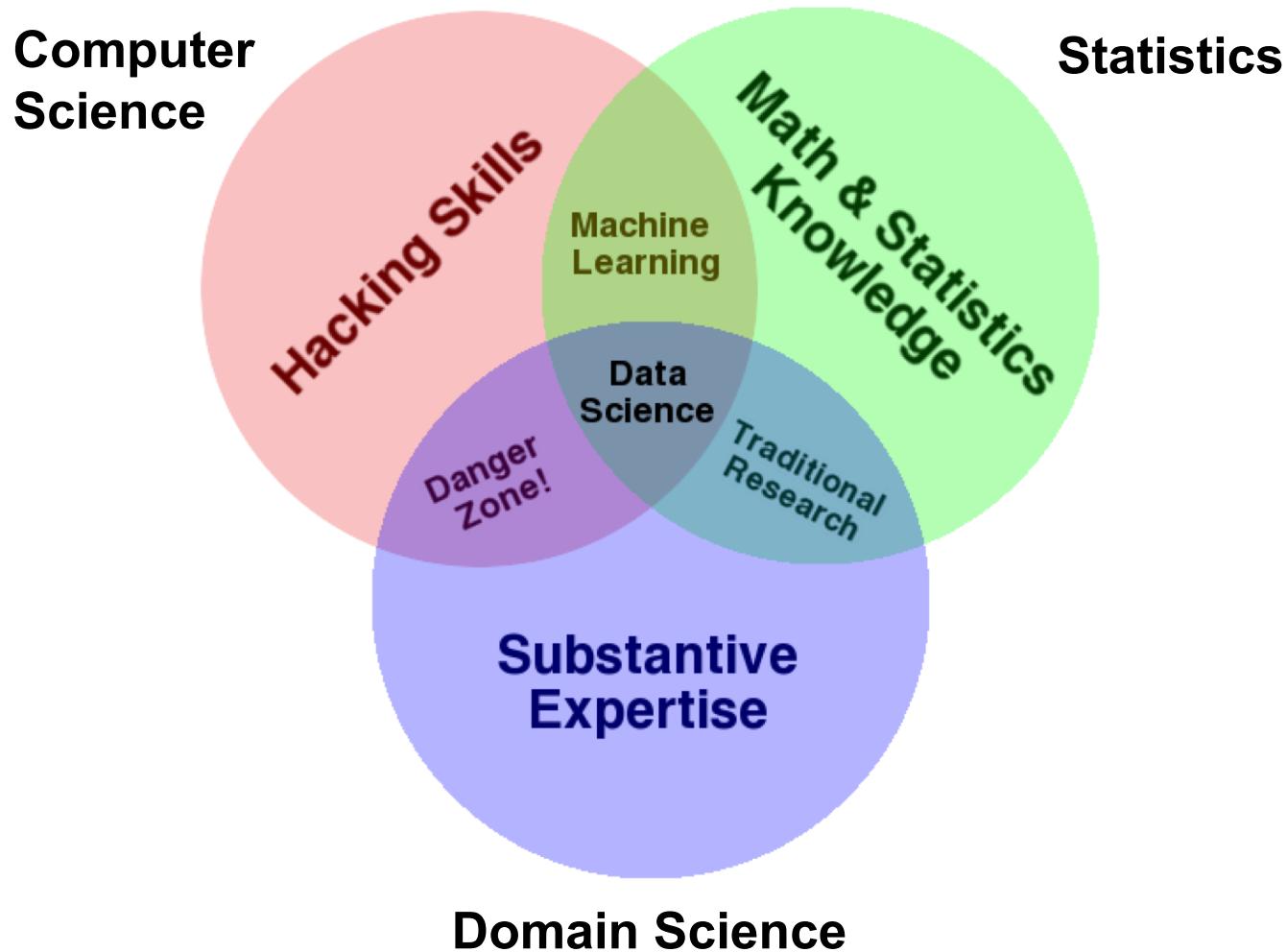
Data Science: Introduction and Applications

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.

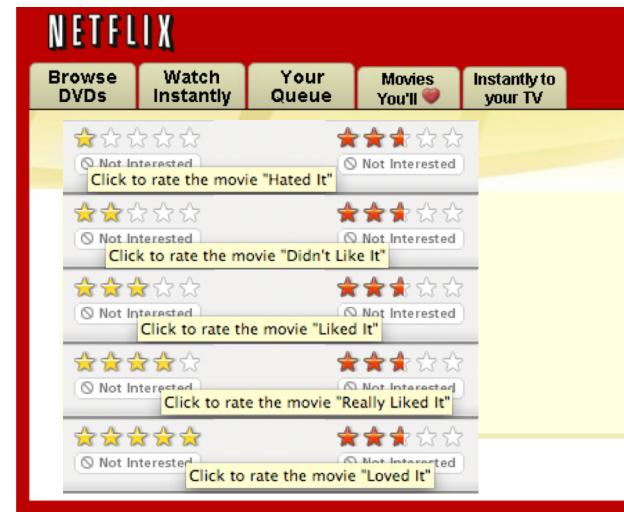


- Drew Conway, CEO and founder of Alluvium, Venn Diagram:



Example: Recommender System

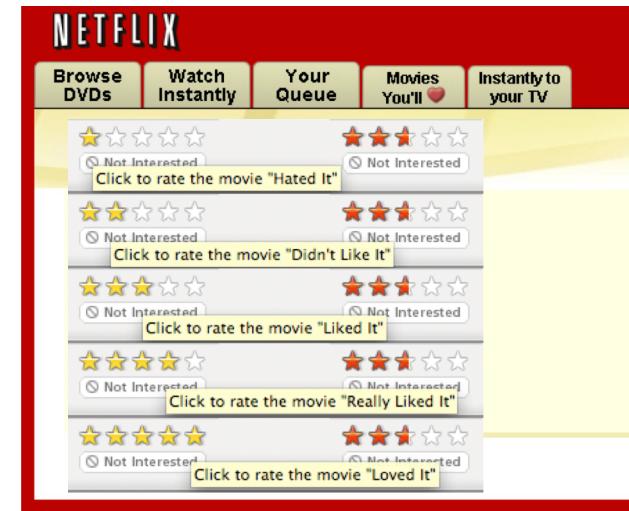
Netflix Prize: \$1,000,000 in an open competition for the best algorithm to predict user ratings for films, based on previous ratings without any other information about the users or films.



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



- **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)



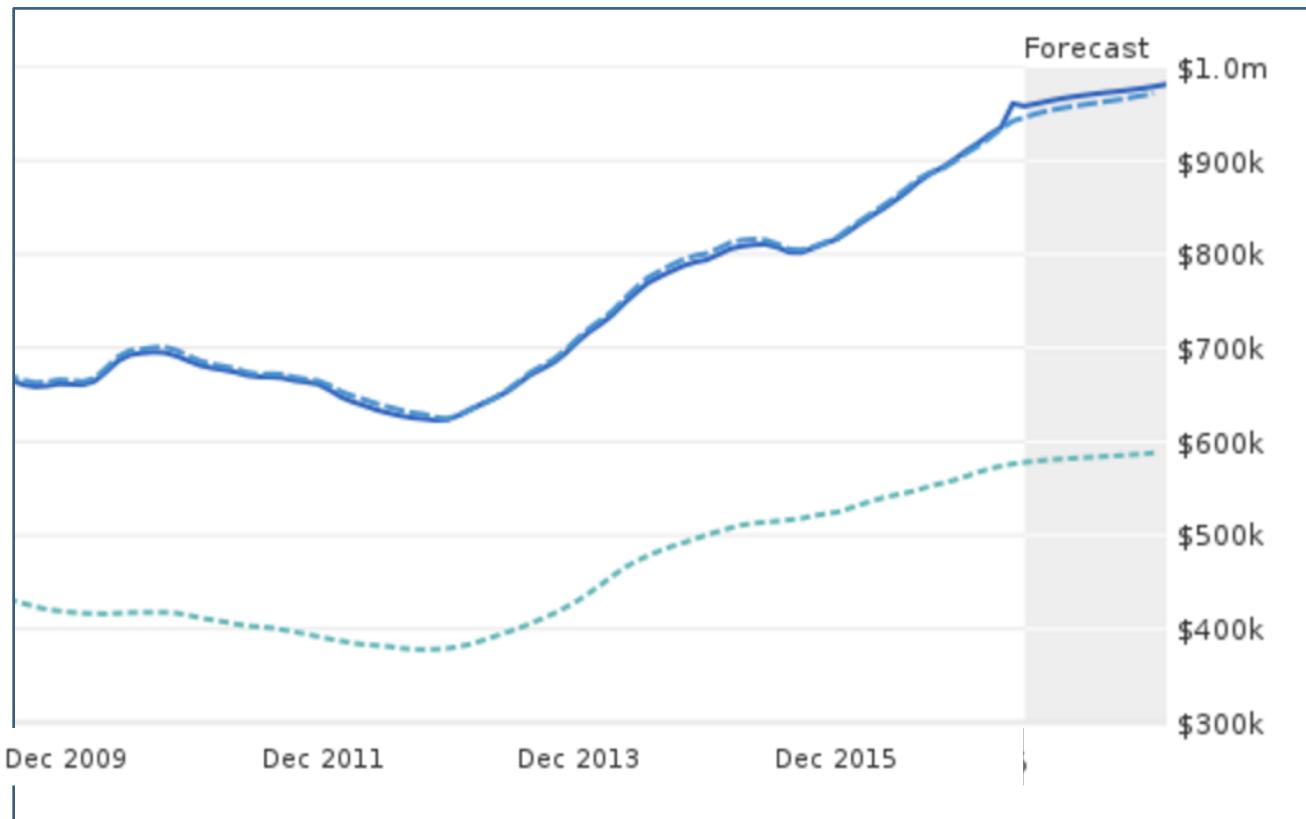
Example: Stock Market Prediction



Example: Real State Prediction

REDFIN™

 Zillow®



Example: Robotics

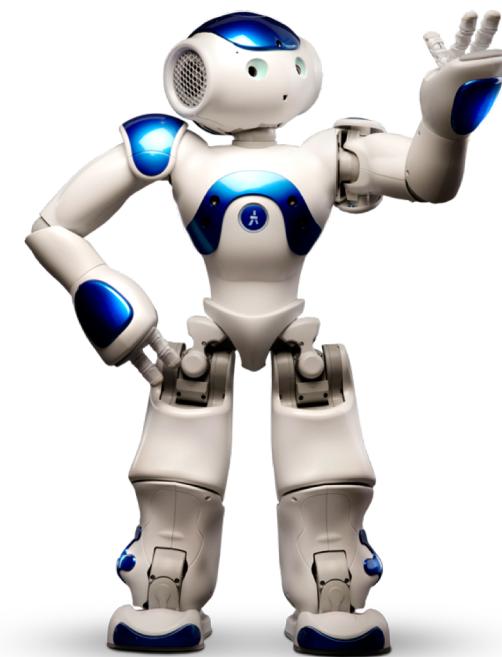
- Every Robot needs a Brain, and this brain is a Computer with data science capabilities that make it act like a human. It can see, hear, learn, and perform many of the human brain functionalities such as automatic decision making!



Cozmo



Dash-Dot



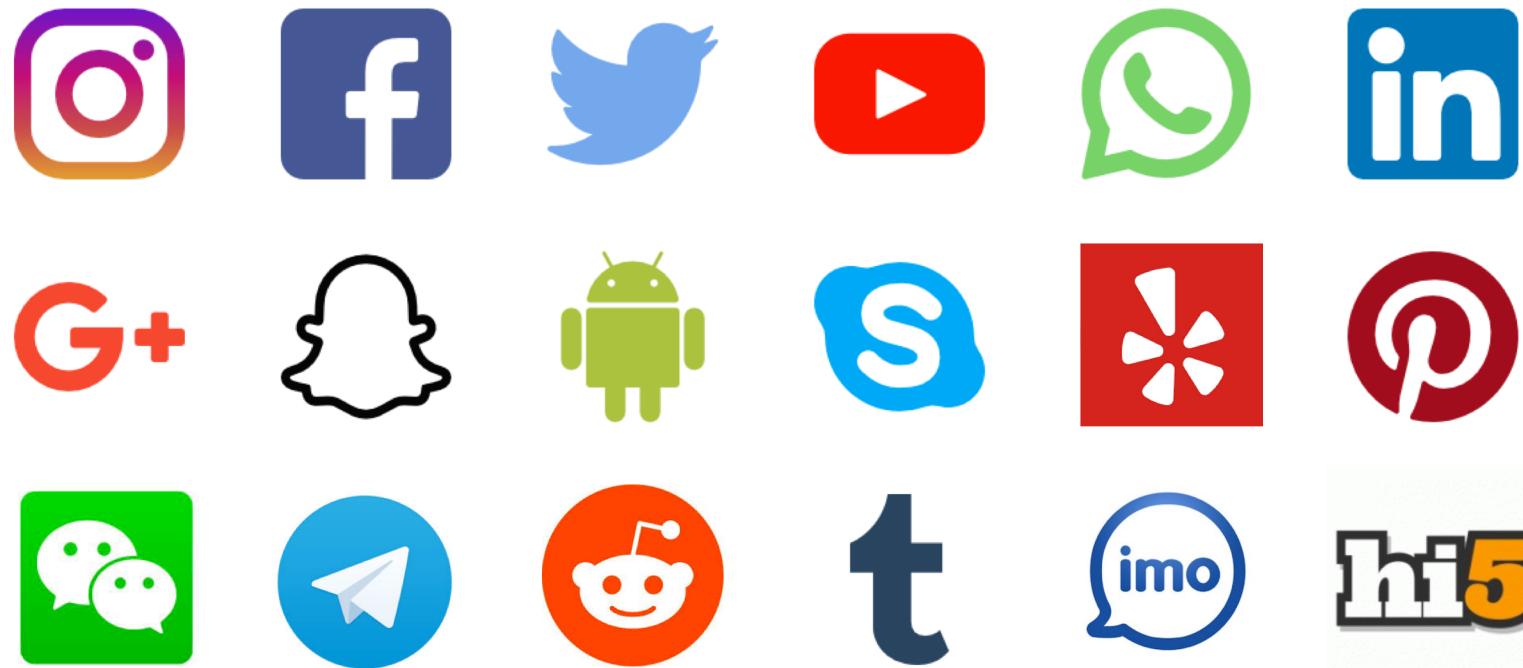
NAO

Example: Spam Detection

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



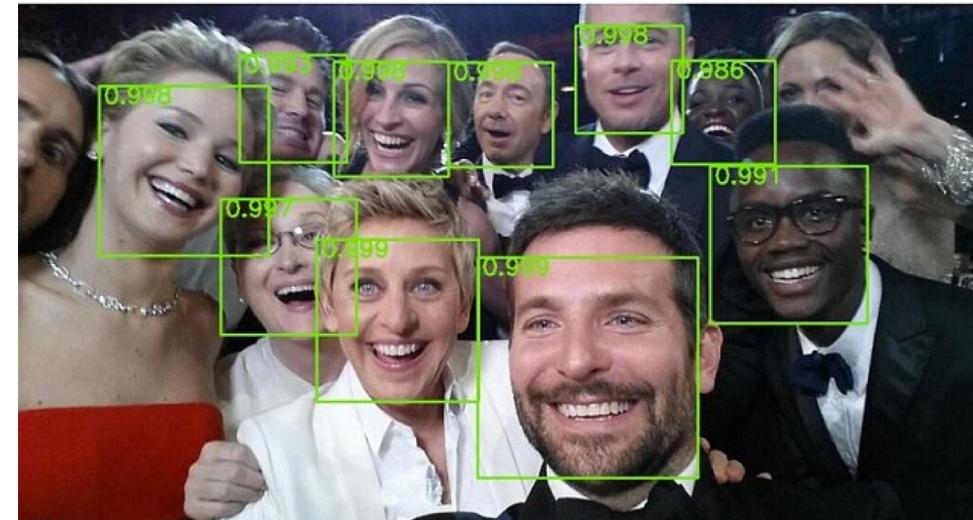
Example: Social Media



Example: Object Recognition



Example: Object Recognition, Face Recognition

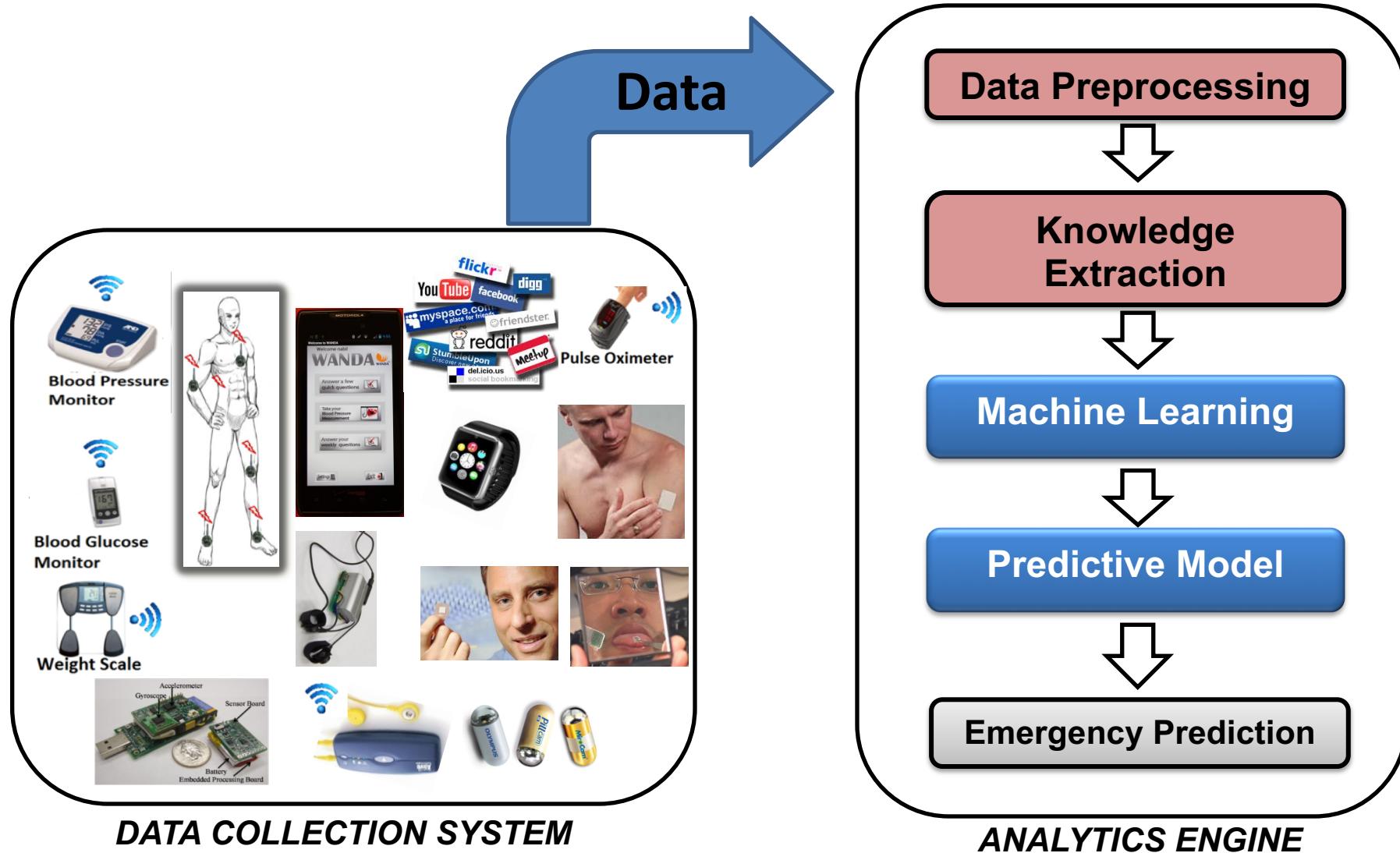


Example: Self-Driving Cars



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



The screenshot shows a web browser displaying the Data Science Research Group page from the Cal State LA website. The URL in the address bar is www.calstatela.edu/research/data-science. The page features a yellow header bar with the Cal State LA logo and navigation links for STUDENTS, FUTURE STUDENTS, FACULTY & STAFF, ALUMNI & GIVING, ACADEMICS, ATHLETICS, and APPLY ONLINE. The main content area has a white background with a large blue heading "Data Science Research Lab". Below it, in blue text, are the names of the College and Department: "College of Engineering, Computer Science, and Technology" and "Department of Computer Science". At the bottom left, there is a section titled "RESEARCH" with a bulleted list of research areas: Data Science, Predictive Analytics/Big Data Analytics, Artificial Intelligence, Machine Learning, and Data Mining, Artificial Neural Network and Deep Learning, Risk Prediction for Healthcare and Medical applications, Health Analytics, Smart and Connected Health, mHealth, and Machine Learning in Image and Video Processing.

Data Science Research Lab

College of Engineering, Computer Science, and Technology
Department of Computer Science

RESEARCH

- Data Science
- Predictive Analytics/Big Data Analytics
- Artificial Intelligence, Machine Learning, and Data Mining
- Artificial Neural Network and Deep Learning
- Risk Prediction for Healthcare and Medical applications
- Health Analytics
- Smart and Connected Health
- mHealth
- Machine Learning in Image and Video Processing



Data Science at Cal State LA

CS4661: Introduction to Data Science (Fall)

CS5661: Advanced Topics in Data Science (Spring)

CS4662: Advanced Machine Learning (Spring)



Prediction

- It seems that the **Prediction** of the future or something unknown is one of the most important applications of Data Science!
- Is there anything that you would like to predict?





Thank You!

Questions?