Welcome

University of Colorado Boulder

CSCI 4622: LECTURE 1

Slides adapted from Chenhao Tan & Jordan Boyd-Graber

Outline

Basic Information

Who will be supporting this class:

- Instructor: Yoshinari Fujinuma
- Teaching Assistant: Saumya Sinha
- Graduate Student Staff: Vignesh Karthikeyan

Course-Related Websites:

- Canvas: `https://canvas.colorado.edu/courses/71729`
- Piazza: `https://piazza.com/class/kjwy1mjifop501`
- Github: `https://github.com/akkikiki/CSCI-4622-Machine-Learning-sp21`

About Myself

- Yoshi or Yoshinari
- 5th year CS PhD student
- Full time search engineer at Amazon for 2 years
  - Worked on product search ranking
- `http://akkikiki.github.io/`
- Research: Natural Language Processing, focusing on cross-lingual/multilingual learning

Remote Lecture Etiquette

- Feel free to use the chat for questions or discuss among peers
- I may stop to answer questions too

Prerequisites

- Programming in Python 3
  - E.g., prepossessing data
- Basic math skills (linear algebra, probability, calculus)
  - E.g., take derivatives, prove things

Communication

- Piazza for all Q&A
  - One factor in participation grades
- Quizzes and announcements on Canvas
- Emails
- Office Hours

Grading Weights

- Homework: 30%
- Quizzes: 30%
- Final Project: 30%
- Participation: 10%
  - Attending lectures, Piazza.

Final Project

Projects can be either of the followings:

- Replicating a research paper
- Come up with your own project idea
- Participate in a public Kaggle competition

More concrete announcement will be made towards the end of semester

Final Project Schedule (Temporary)

This is subject to change and more concrete announcement will be made towards the end of semester

- March 3rd: Team formulation due
- March 19th: Proposal due
- April 19th: Midpoint check-in
- May 1st: Final project presentation due
- May 3rd: Final project report due

Outline

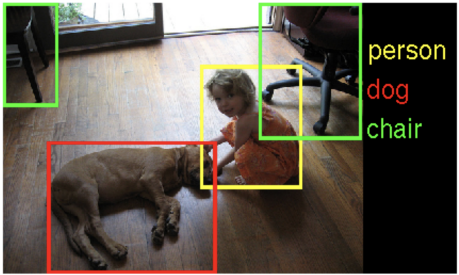
What is Machine Learning?

"Field of study that gives computers the ability to learn without being explicitly programmed" Arthur Samuel (1959)

- Example: Given an array of N integers, return the array sorted in ascending order
- Input: $\{5, 3, 6\}$, Output: $\{3, 5, 6\}$
- Explicitly programmed solution: Use any sorting algorithm
- ML solution: Gather inputs and outputs, let the computer learn how to sort

Machine Learning is Everywhere

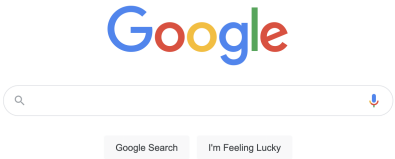
Image from `https://www.microsoft.com/en-us/translator/blog/2018/03/14/`



Machine Learning in a Little More Formal Way

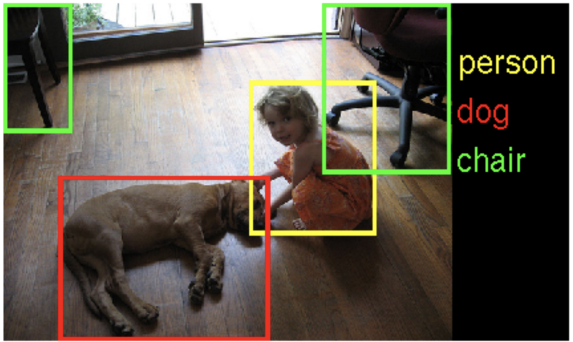
Data	Hidden Structure	Answer/Labels
X	Z	Y

Search Ranking



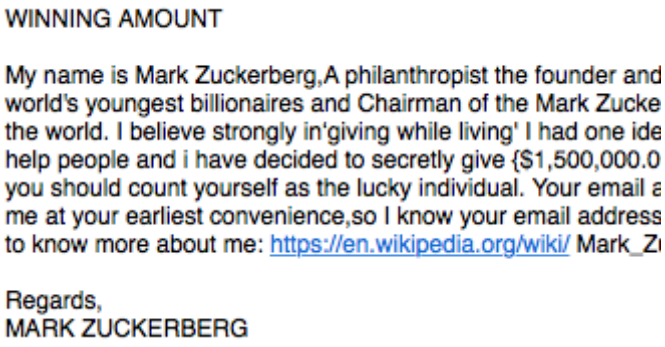
- X: Features of the websites
  - e.g., how many incoming links
- Y: Ranking of the websites
  - e.g., User click-through logs

Object Detection



- X: images (or its pixels)
- Y: labels of objects (person, dog, chair)

Text Classification (Spam or not?)



WINNING AMOUNT

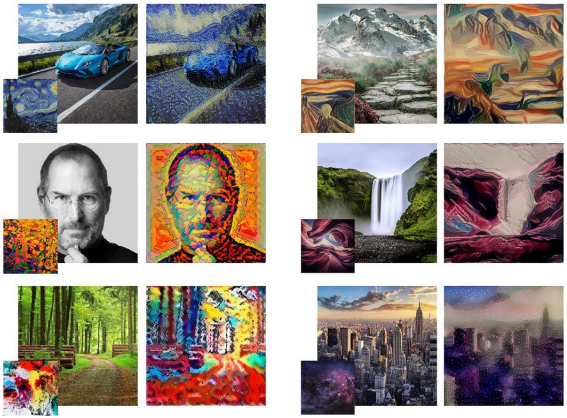
My name is Mark Zuckerberg,A philanthropist the founder and
world's youngest billionaires and Chairman of the Mark Zucke
the world. I believe strongly in'giving while living' I had one ide
help people and i have decided to secretly give {$1,500,000.0
you should count yourself as the lucky individual. Your email a
me at your earliest convenience,so I know your email address
to know more about me: https://en.wikipedia.org/wiki/ Mark_Z

Regards,
MARK ZUCKERBERG

Image from `https://www.ezcomputersolutions.com/blog/hilarious-spam-email-examples/`

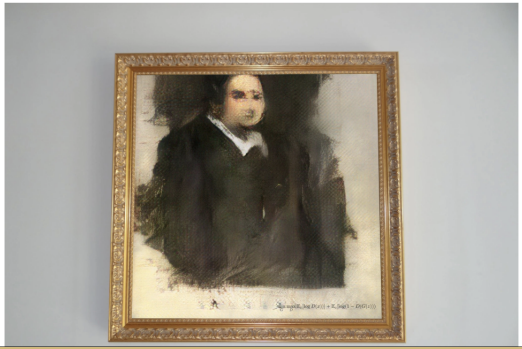
- X: e.g., words
- Y: labels of documents (spam or not)

Machine Learning Applied to Art



Machine Learning Applied to Art

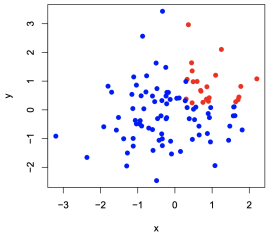
AI Art at Christie's Sells for $432,500



Some Key Concepts in Machine Learning

- Supervised learning
- Unsupervised learning
- Data representation or feature engineering

Supervised Learning



Find patterns in **fully observed** data and then try to predict something from **partially observed** data

- E.g., Spam email classification
  - Fully observed data: Collection of spam and non-spam emails
  - Partially observed data: New incoming emails

Unsupervised Learning

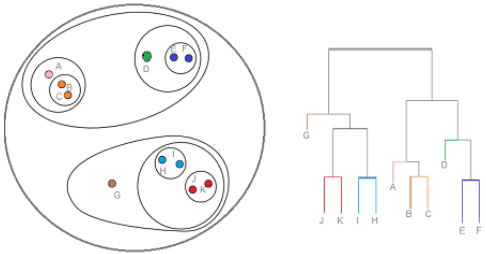


Image from `https://www.statisticshowto.com/hierarchical-clustering/`

Find hidden structure in data, structure that we can never formally observe.
Unsupervised learning is more difficult to evaluate than supervised learning.

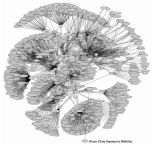
- E.g., Hierarchical clustering

Data Representation / Feature Engineering

$\longrightarrow \langle 1.5, 3.2, -5.1, \ldots, 4.2 \rangle$

Republican nominee George Bush said he felt nervous as he voted today in his adopted home state of Texas, where he ended...

$\longrightarrow \langle 1, 0, 0, 0, 5, 0, 9, 3, 1, \ldots, 0 \rangle$



$$\longrightarrow \begin{bmatrix} 1 & 0 & 1 & \ldots & 0 \\ 0 & 1 & 1 & \ldots & 0 \\ 1 & 0 & 0 & \ldots & 1 \\ & \ldots & & & \\ 0 & 0 & 0 & \ldots & 0 \end{bmatrix}$$

Outline

Learning Objective of this Class

Goal: fluency in thinking about modern data analysis problems

- Assumptions we make on data
- Provide tools to analyze and solve problems in a data-driven
- Capabilities and limitations

Survey

Email me

- Which online ML courses have you took?
- What are you expecting to learn from this class?
- What problems/applications are you willing to solve?
- Any random comments