CSCI 467: Introduction to Machine Learning

Robin Jia

USC CSCI 467, Fall 2023

August 22, 2023

Today's Plan

- The What, Why, and Where of Machine Learning
- Course Logistics
- Bird's Eye View of the Schedule

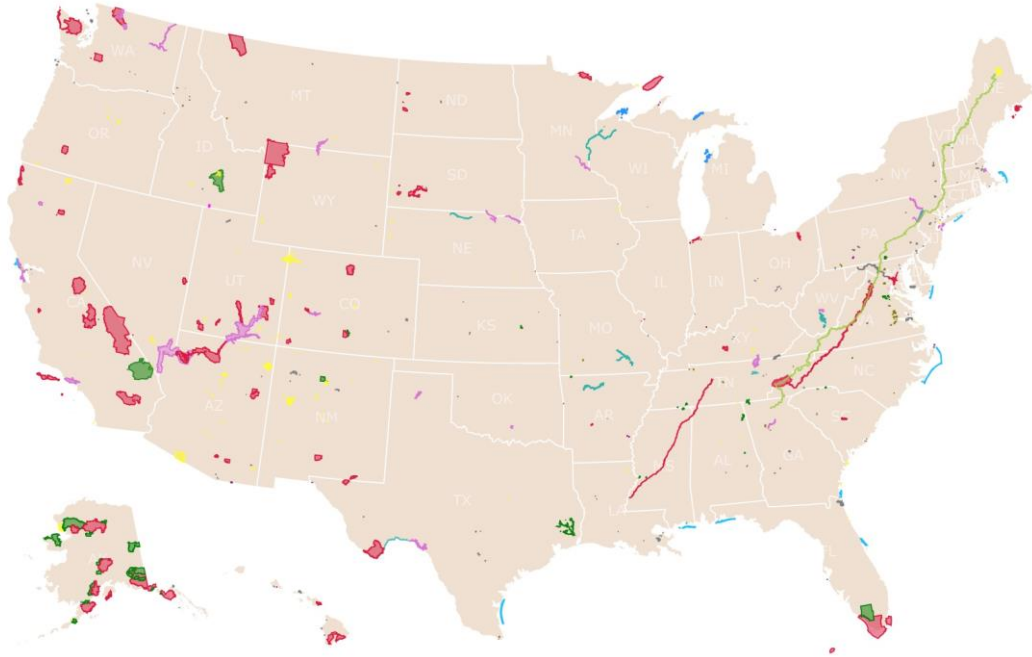
Today's Plan

- The What, Why, and Where of Machine Learning
- Course Logistics
- Bird's Eye View of the Schedule

The Case for Machine Learning

Checking if location is in national park:

Can be programmed directly!



https://xkcd.com/1425/

The Case for Machine Learning

Checking if photo is a bird…

How to define "birdness" in a program???

Hard to define directly—instead, **learn from data!**

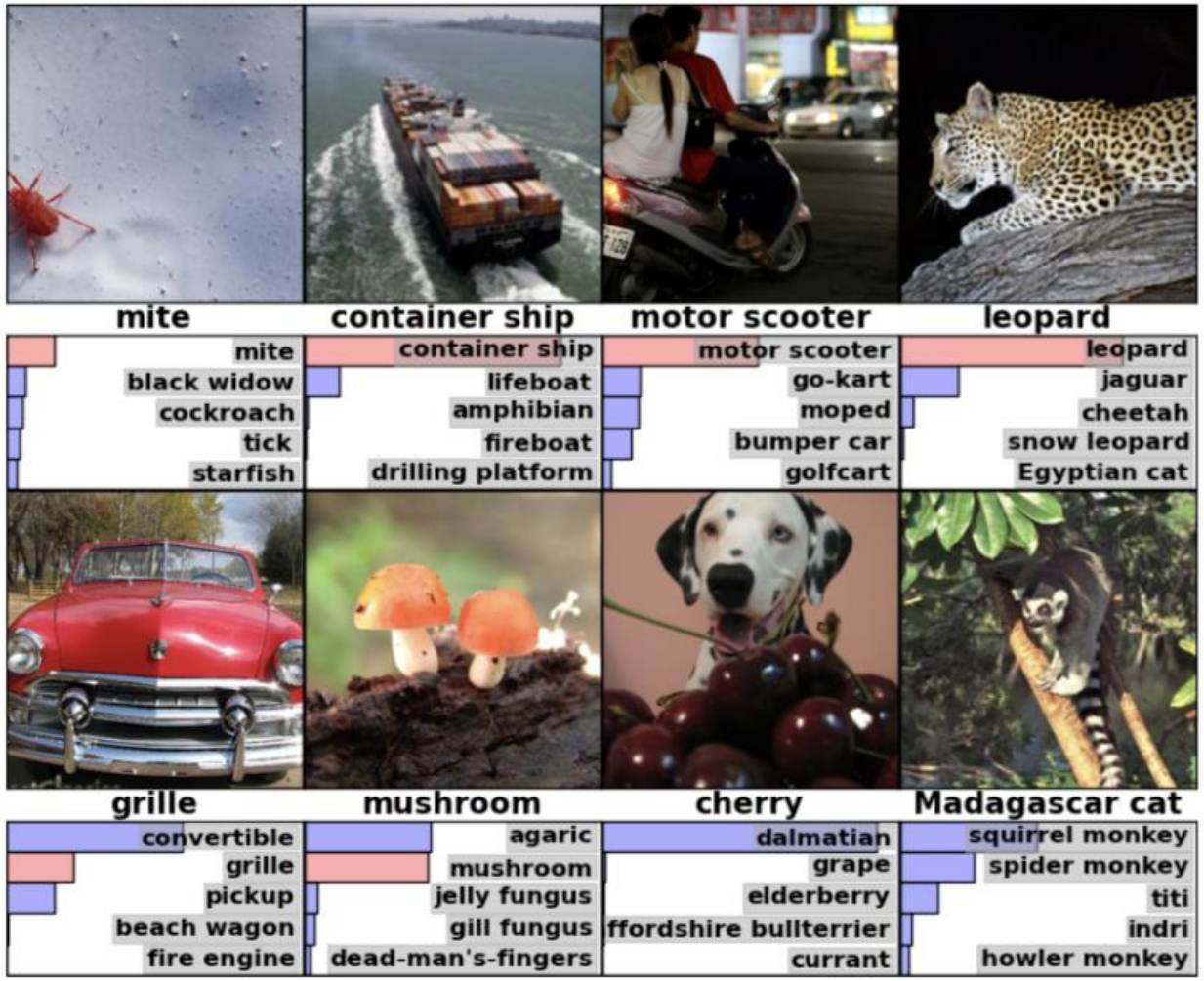
https://xkcd.com/1425/

Machine Learning in a Nutshell

Input	Output
	Bird
	Bird
	Not Bird
	Bird
	Not Bird

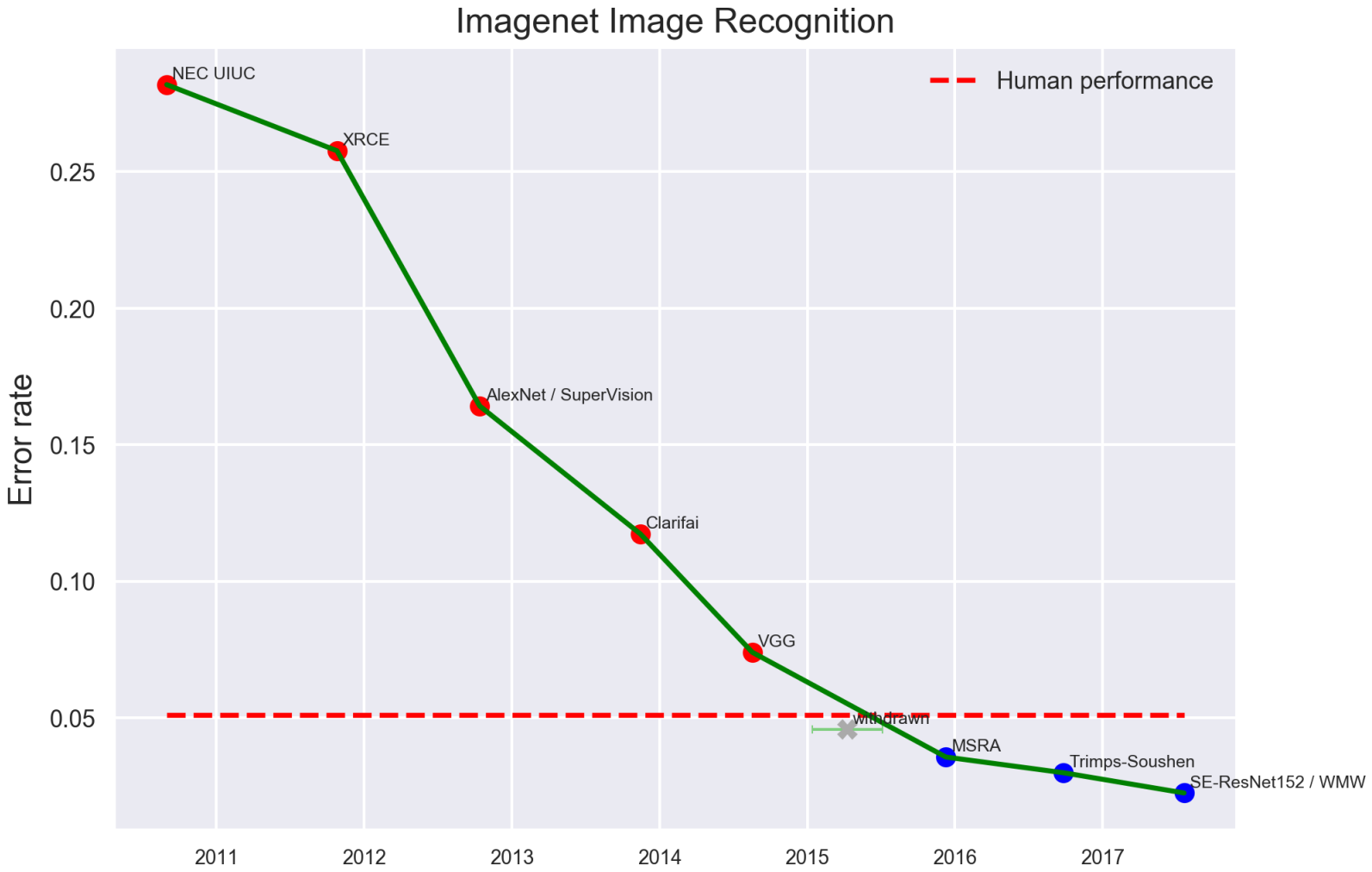
1. I don't know how to solve my problem directly
2. But I can obtain a **dataset** that describes what I want my computer to do.
3. So, I will write a program that **learns the desired behavior from the data.**

Computer Vision



- ImageNet dataset: 14M images, 1000 labels

Progress on ImageNet



- 2012: AlexNet wins ImageNet challenge, marks start of deep learning era
- 2016: Machine learning surpasses human accuracy

Source: https://www.eff.org/files/AI-progress-metrics.html

Image Generation

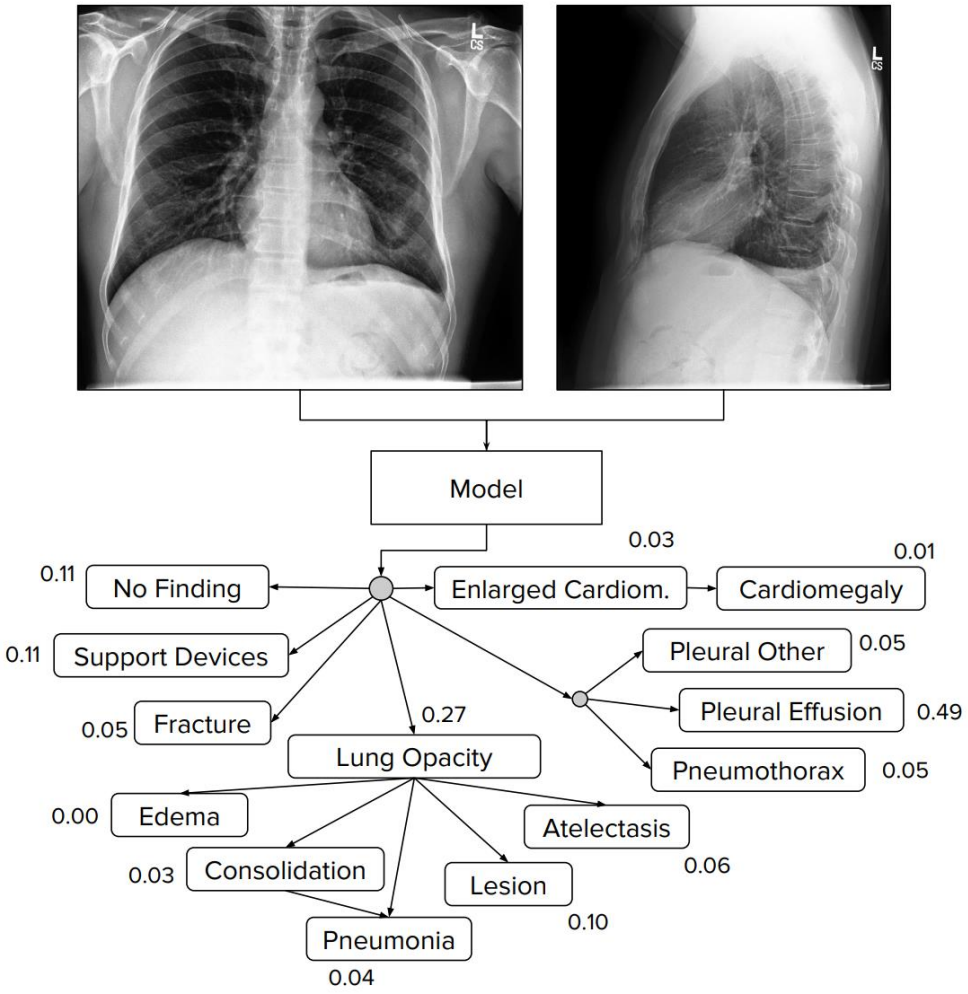
Teddy bears working on new AI research on the moon in the 1980s.

A raccoon playing tennis at Wimbledon in the 1990s.

An ancient Egyptian painting depicting an argument over whose turn it is to take out the trash.

Generated by OpenAI's Dall-E 2. Source: https://qz.com/2176389/the-best-examples-of-dall-e-2s-strange-beautiful-ai-art

Medicine



- Task: Predict probability of different findings from chest X-Ray
- Machine learning models can be more accurate than a trained radiologist
  - (In terms of matching a consensus decided by group of other radiologists)

Source: CheXpert dataset. https://stanfordmlgroup.github.io/competitions/chexpert/

Sustainability

- Producing fine-grained maps of poverty based on satellite imagery
  - E.g. for policy decisions or help NGO's distribute resources

Source: https://www.science.org/content/article/satellite-images-can-map-poverty

Machine Translation

"The spirit is willing, but the flesh is weak."

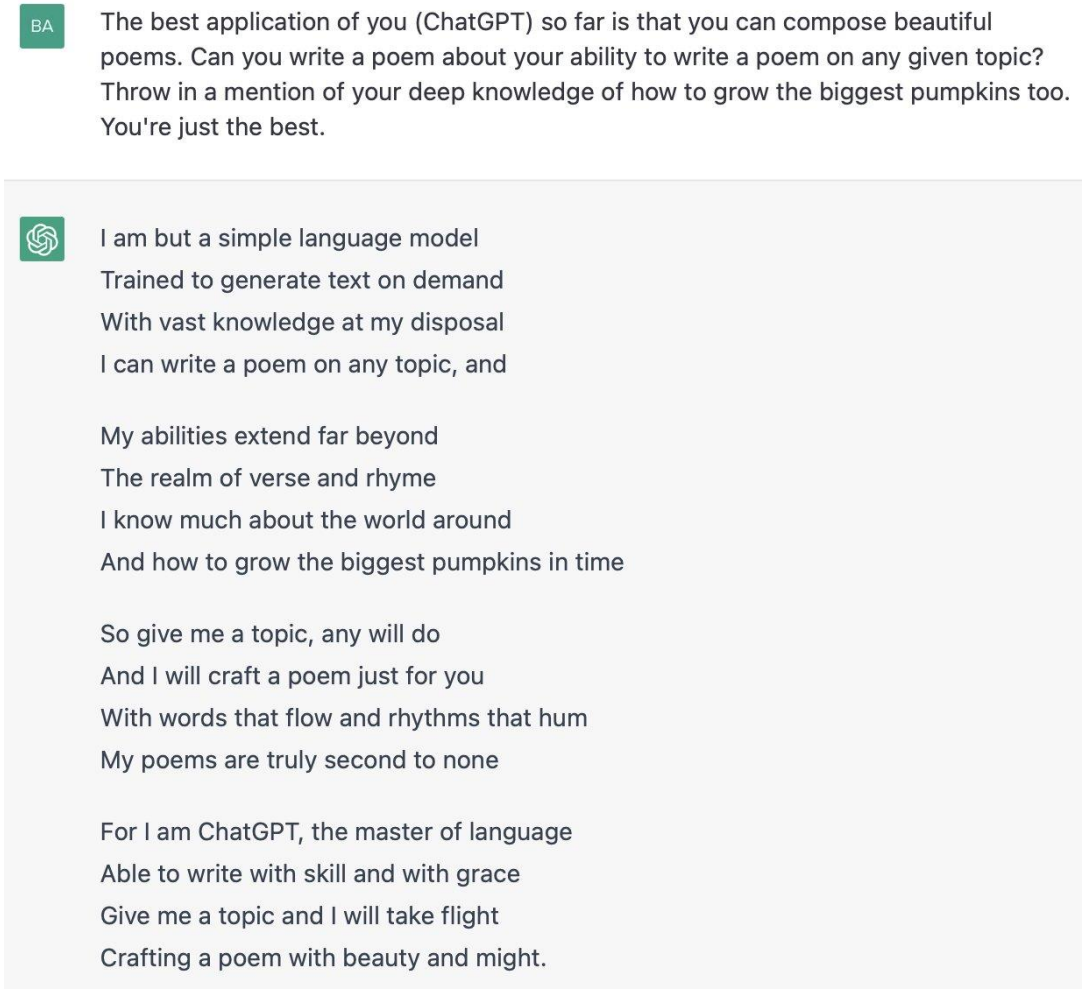
Translate to Russian and back

"The vodka is good, but the meat is rotten."

- Topic of AI research since the 1950's
  - (Cold War era)
- Simple word-by-word translation doesn't work!
- Today: Modern machine translation systems all use machine learning

https://www.nytimes.com/1983/04/28/business/technology-the-computer-as-translator.html

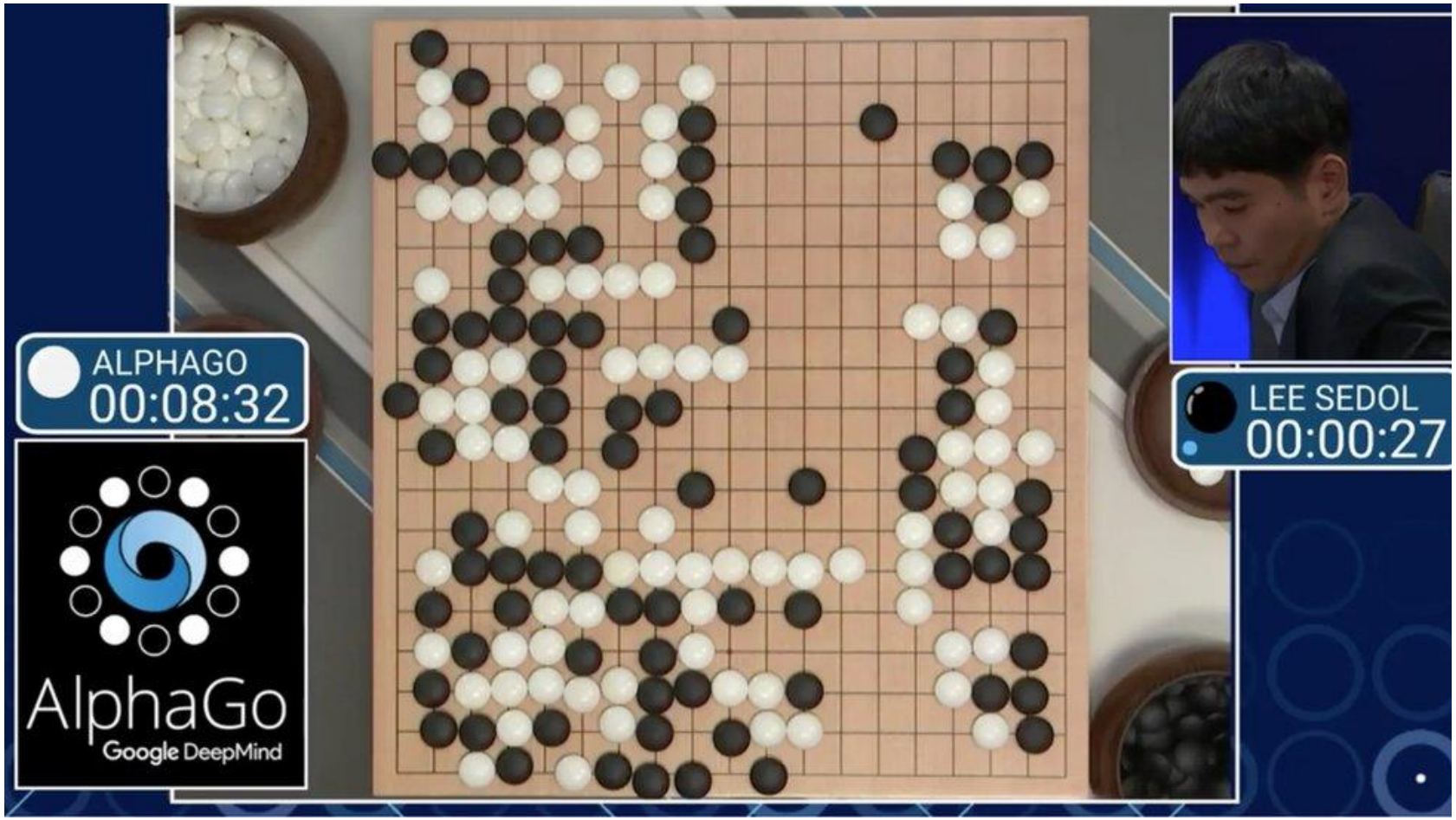
Language Models



- Step 1: Acquire as much text data from the internet as possible
- Step 2: Train a model that repeatedly predicts the next word given previous words
- Step 3: ???

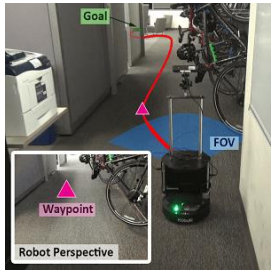
https://twitter.com/karpathy/status/1600214083206193153

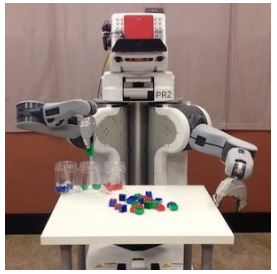
Game-Playing

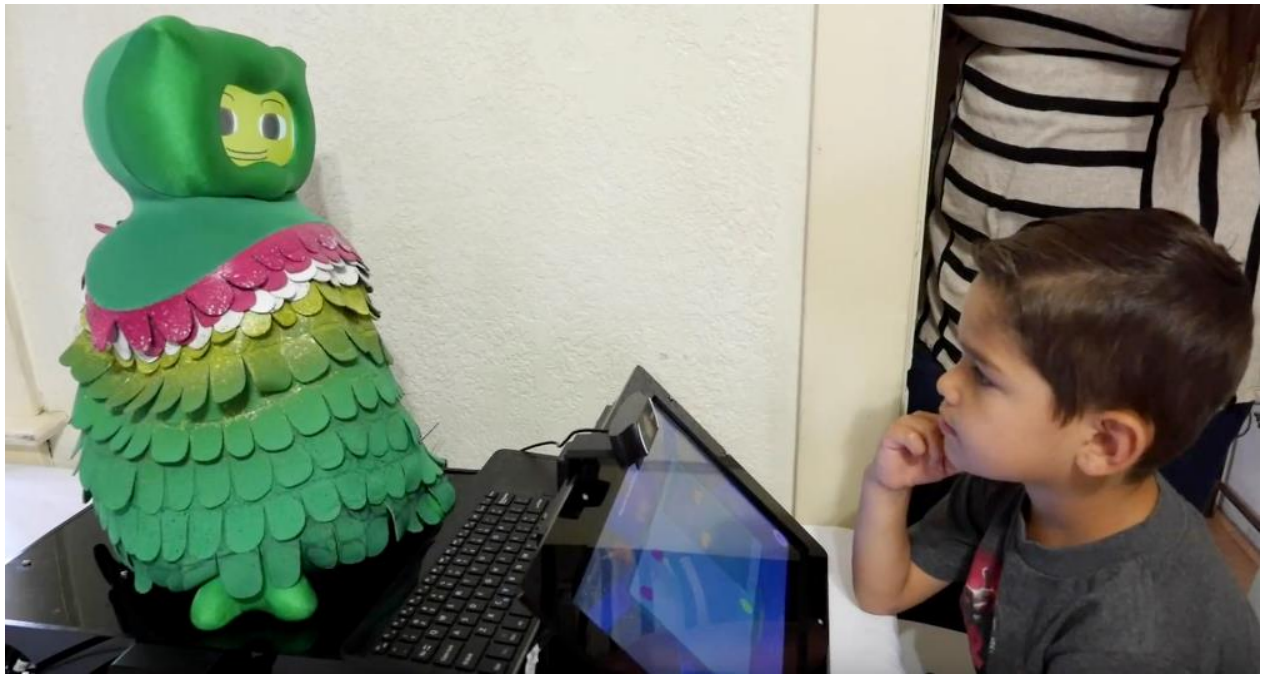


- 2017: AlphaGo defeats Go champion Lee Sedol
- How? Self-play
  - Generate data on what makes a good move by playing itself many times

Robotics







- Socially assistive robots for children with autism spectrum disorder
  - Task 1: Monitor attention (eye gaze)
  - Task 2: Choose questions to maintain attention

https://uscinteractionlab.web.app/project/expeditions

Today's Plan

- The What, Why, and Where of Machine Learning
- Course Logistics
- Bird's Eye View of the Schedule

Course staff introduction

- Instructor: Robin Jia
- TA’s: Soumya Sanyal, Wang (Bill) Zhu
- CP’s: Vishesh Agrawal, Atharva Mundada, Ryan Wang, Lorena Yan

Logistics

- Website: https://usc-csci467.github.io/
  - See calendar for office hours
- Discussions on https://piazza.com/class/llgu91z0jn5di/
  - Sign-up link on website
- Lecture format
  - Some whiteboard/iPad, some lecture slides
    - My goal: Release lecture notes before class for iPad days, lecture slides before class for slideshow days
  - Announcements in middle

Prerequisites

- Algorithms: CSCI 270
  - Nothing specific but proxy for general ability to reason about algorithms
- Linear Algebra: Math 225
  - Lots of vector & matrix operations, vector geometry
- Probability: EE 364/Math 407/BUAD 310
  - Lots of probability notation and probabilistic processes
  - Bayes Rule, conditional probability/expectation
  - Basic probability distributions (Gaussian, Bernoulli, etc.)
- Calculus
  - Single variable calculus assumed
  - Some basic multivariable calculus will be introduced
- Programming: Familiarity with python
- Suggested resources for review on the course website

Section

- Fridays 2:00-2:50pm in DMC 100
- **This Friday: Probability, linear algebra, calculus review**

Grading Breakdown

- Homework Assignments (40%)
  - Homework 0 (4%)
  - Homeworks 1-4 (9% each)
- Final Project (20%)
- Exams (40%)
  - Midterm (80 minutes in-class, October 10)
  - Final Exam (December 7, 2:00-4:00pm)

Homework

- **Homework 0 is out, due August 31 (at 11:59pm)**
  - Main purpose is to exercise prerequisites, plus start on some material we'll learn in the next class
- Submit on Gradescope
  - Separate places for you to submit PDF write-up and code
- LaTeX is highly recommended
  - Will be required for final project

Final Project

- Can be done individually or in groups of up to 3
- Chance to apply machine learning techniques to a problem of your choice
  - Finding an appropriate dataset
  - Establishing baselines
  - Evaluating your method's success
  - Analyzing its successes and failures
- Timeline
  - Proposal (due September 26): *Is this feasible? Does the right data exist?*
  - Midterm report (due October 31): *Halfway point for running experiments*
  - Final report (due December 12, after final exam)

Late Days

- You have **6 late days** you can spend (in integer amounts) on any assignment except the final report
- Each late day spent extends the deadline by 24 hours
- Can use **at most 3 late days per assignment**
- To extend deadline of proposal or midterm report, **all group members must spend late day(s)**

Academic Integrity

- You may discuss homework problems at a high level with other students
- You may **not…**
  - Look at another student's solutions/share your solutions
  - Obtain homework solutions from any online source
  - Use any AI tools to help you write your solutions or code
  - Upload materials from this course online

Today's Plan

- The What, Why, and Where of Machine Learning
- Course Logistics
- Bird's Eye View of the Schedule

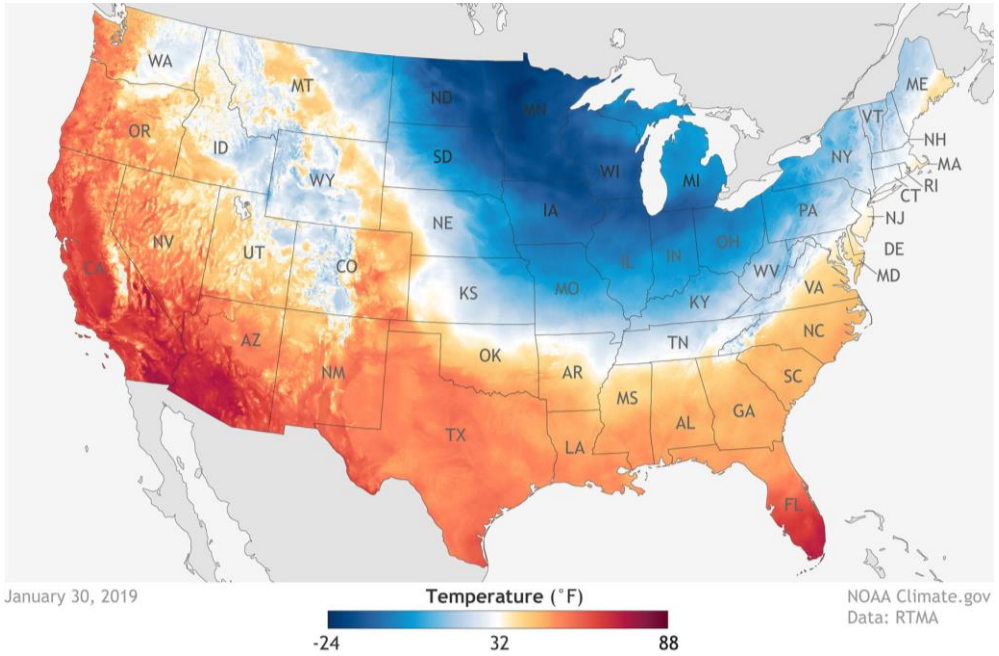
A Bird's Eye View

- Supervised learning
  - Linear models (Weeks 1-5) CSCI 360: ~2 weeks
  - Deep learning (Weeks 6-9) CSCI 360: ~1 week
- Unsupervised learning (Weeks 10-11) CSCI 360: ~0.5 weeks
- Reinforcement learning (Weeks 12-13) CSCI 360: ~0.5 weeks
- Additional topics (Weeks 14-15)
- Compared with CSCI 360: More in-depth, more mathematical

Supervised Learning

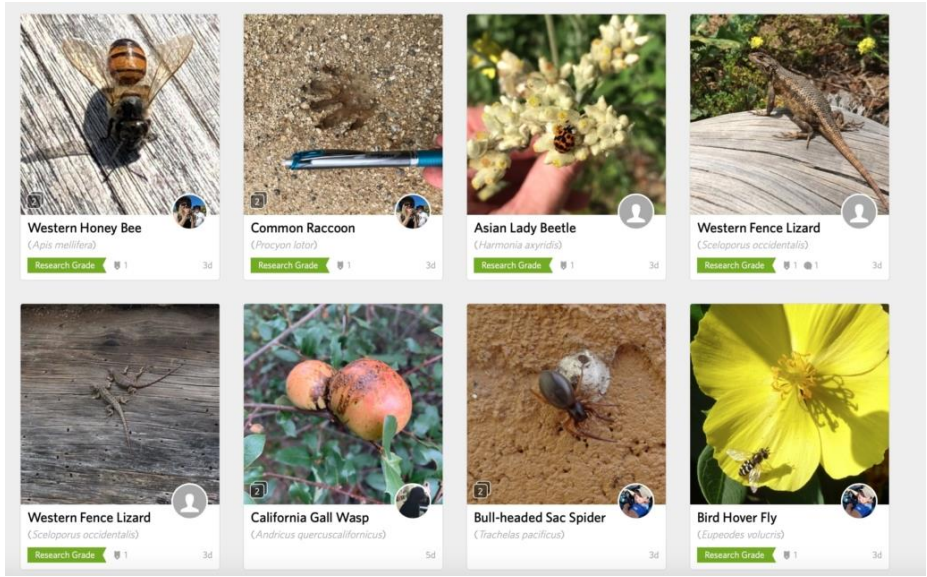
Regression

- Predicting a real number
- Example: Weather prediction

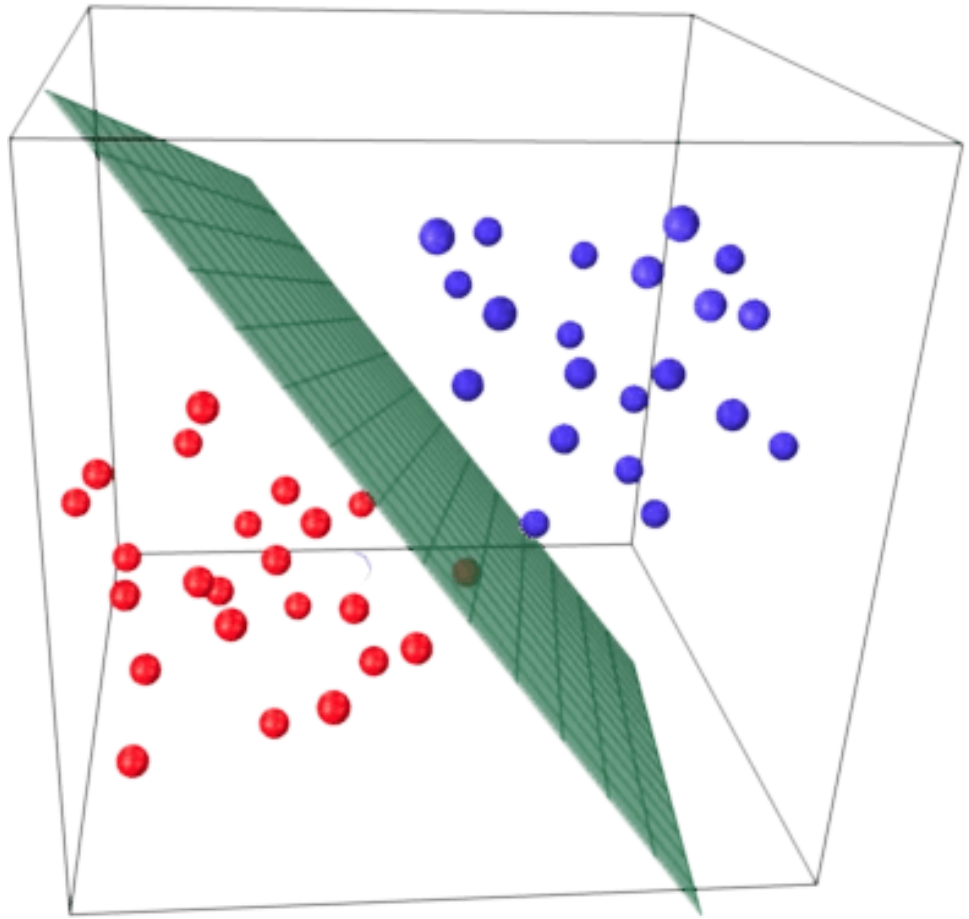


Classification

- Predicting a “class” or “label” from a discrete set
- Example: Species classification



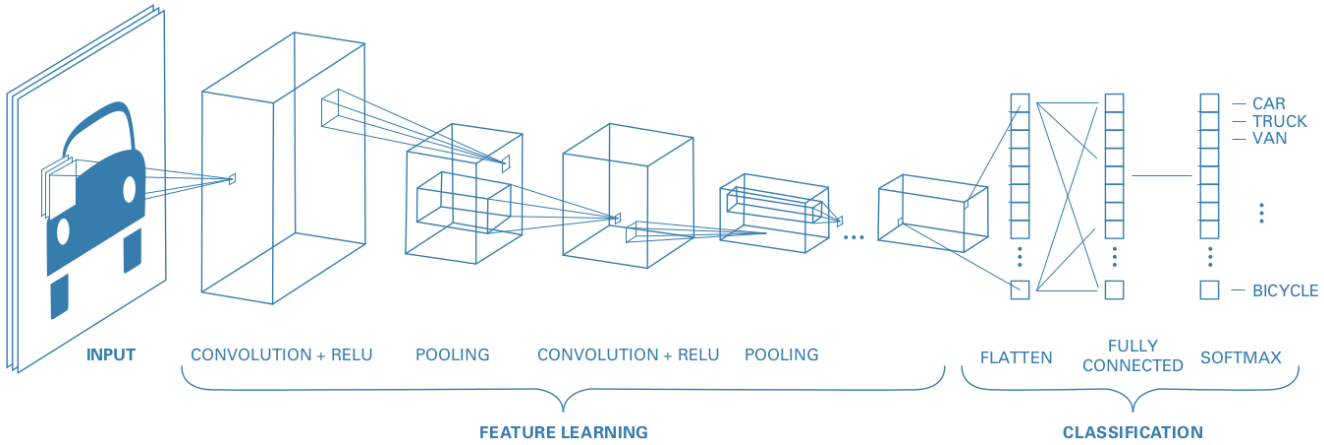
Linear Models



- Idea: Only use **linear** function of input features
- Advantages
  - Simple
  - Efficient
  - Comes with provable guarantees
  - Often good choice for small datasets
- Disadvantages
  - Lack of expressivity*
  - Harder to take advantage of large datasets

Source: https://mlcourse.ai/book/topic04/topic4_linear_models_part2_logit_likelihood_learning.html

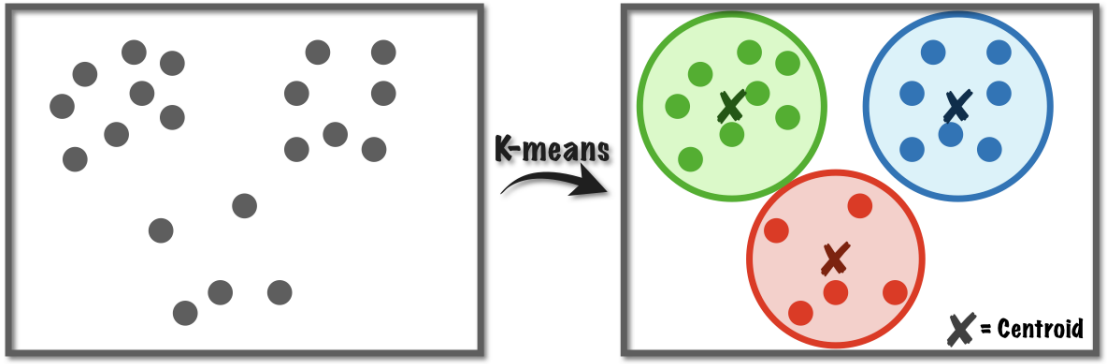
Deep Learning

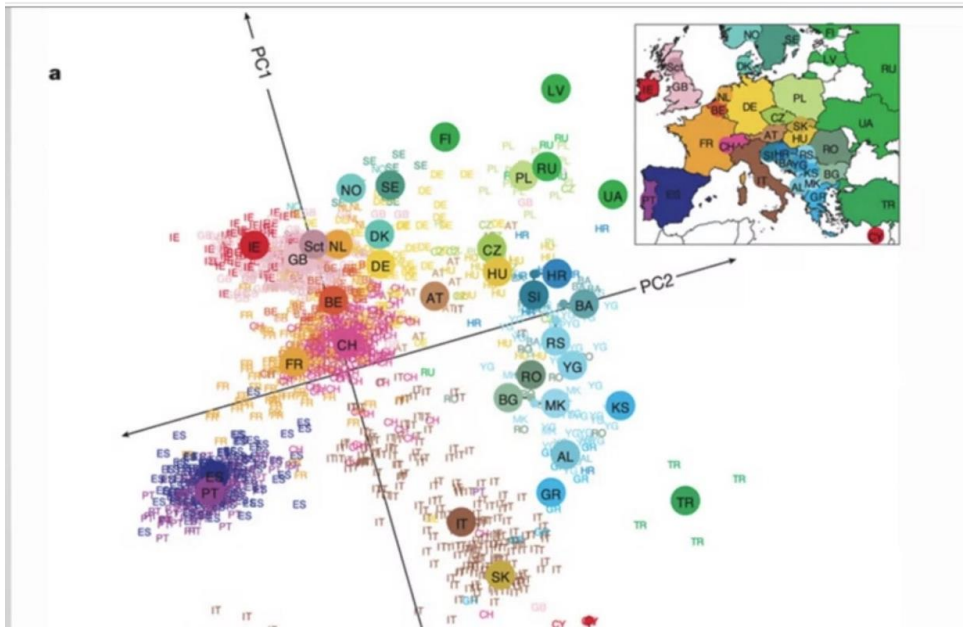


- Idea: Learn meaningful vector representations of inputs by composing **non-linear** operations
- **Computer vision**: Convolutional Neural Networks
- **Natural Language Processing**: Recurrent Neural Networks, Transformers

Source: PML, https://data-science-blog.com/blog/2021/04/07/multi-head-attention-mechanism/

Unsupervised Learning

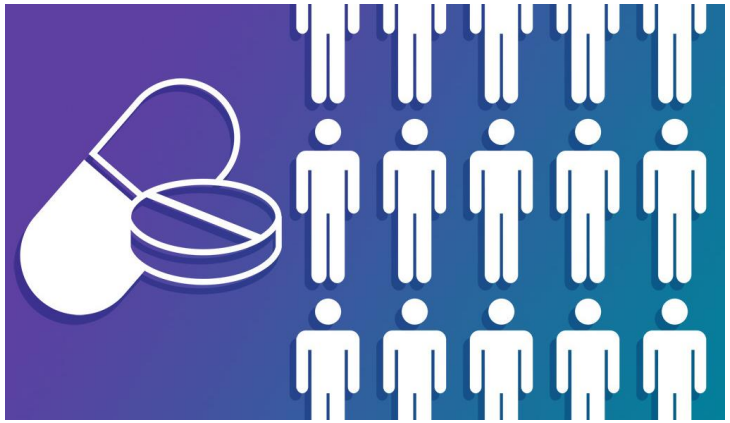




- **Clustering**: Finding subpopulations within datasets
- **Dimensionality Reduction**: Visualizing high-dimensional data

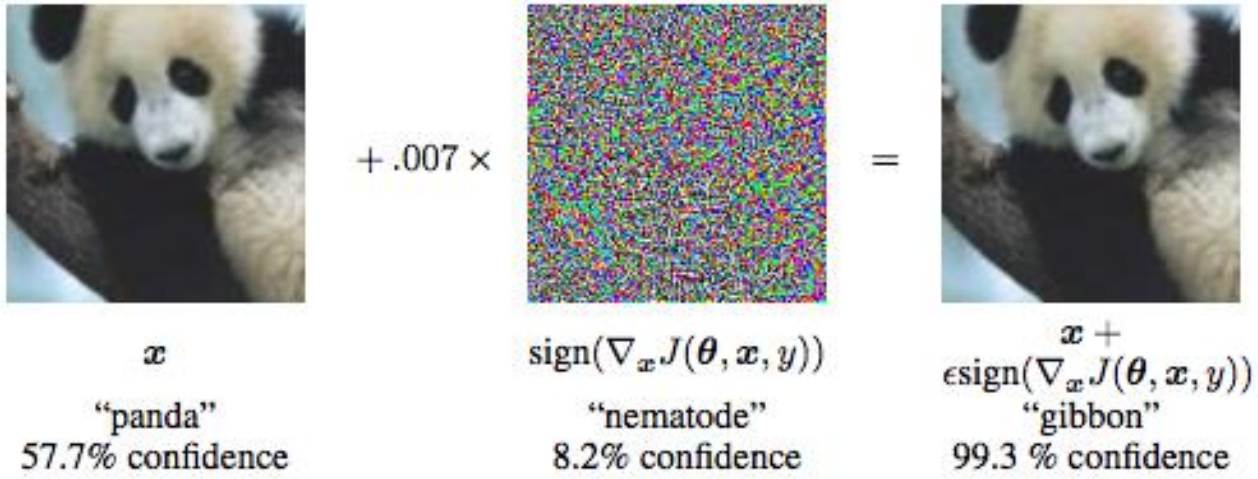
Source: https://towardsdatascience.com/k-means-a-complete-introduction-1702af9cd8c, https://medium.com/swlh/a-gentle-introduction-into-the-application-of-principal-component-analysis-pca-in-genomics-269026453295

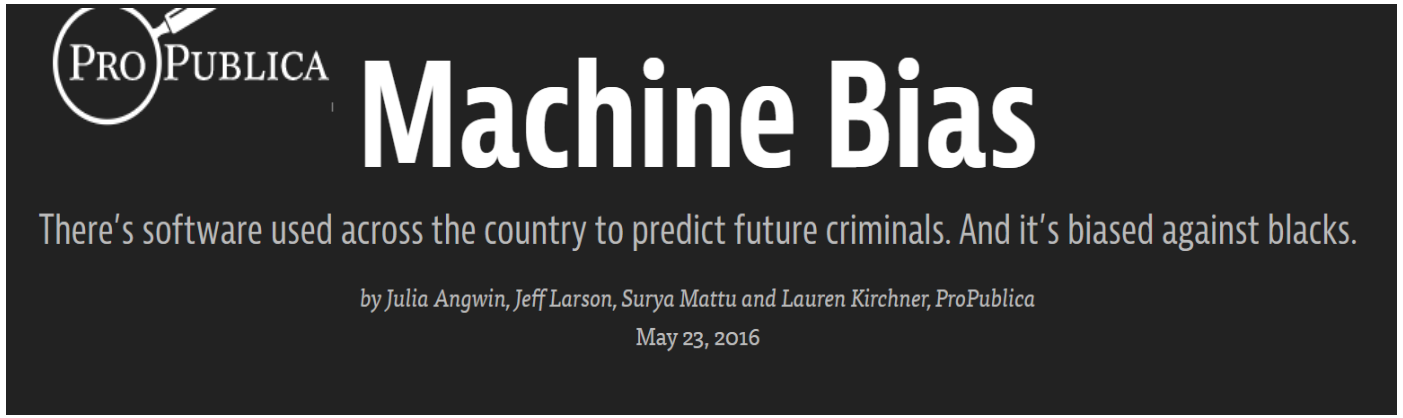
Reinforcement Learning



- **Bandit problems**: Trading off exploration vs. exploitation
- **Reinforcement Learning**: Learning how to act to maximize rewards

Additional Topics





- **Adversarial Examples**: Hidden ways machine learning models can be fooled
- **Fairness**: How to ensure responsible deployment of machine learning systems?

Source: Goodfellow et al. (2014), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Conclusion

- Machine Learning
  - What? Getting computers to learn what to do from data
  - Why? Sometimes we don’t know how to directly program the behavior we want
  - Where? Images, medicine, sustainability, language, games, robotics, …
- Homework 0 due in 9 days!
- Next class: Linear Regression