CSCI 467: Introduction to Machine Learning

Robin Jia

USC CSCI 467, Spring 2025

January 14, 2025

Today's Plan

- What is machine learning?
- Where is machine learning used?
- Course Logistics
- Bird's Eye View of the Schedule
- Start on linear regression

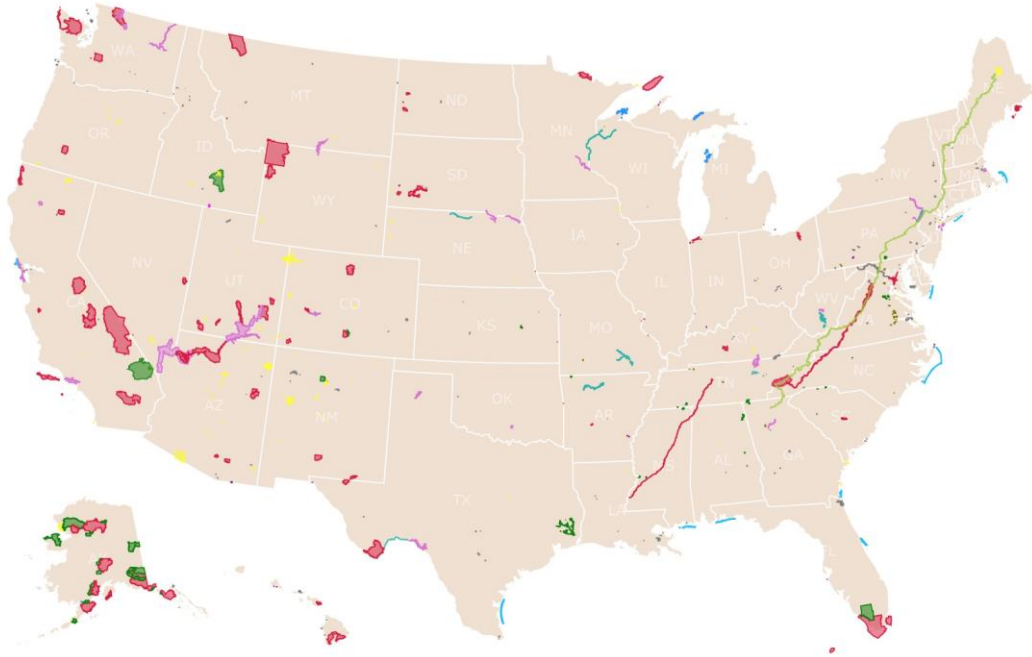
Today's Plan

- What is machine learning?
- Where is machine learning used?
- Course Logistics
- Bird's Eye View of the Schedule
- Start on linear regression

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/

Learning from Data

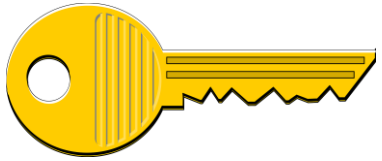
Input	Output
	Bird
	Bird
	Not Bird
	Bird
	Not Bird

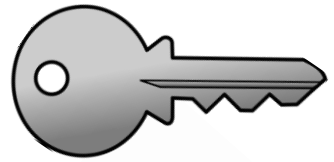
1. I don't know how to solve my problem directly
2. But I can obtain a **dataset** that describes what I want my computer to do.
3. So, I will write a program that **learns the desired behavior from the data.**

Learning by making adjustments

	Desired Output	Actual Output
My key	Should open	Does not open
Other keys	Should not open	Opens
	Should not open	Does not open
	Should not open	Does not open

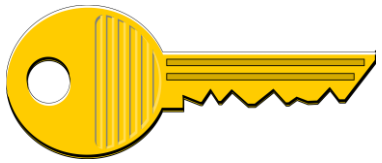
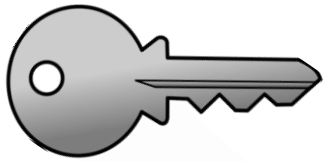
Errors





- Suppose I'm building a lock to my door
  - My key should open the lock
  - All other keys should not
- What if there's a mistake?
  - Adjust the lock until it works!

Learning by making adjustments

	Desired Output	Actual Output
My key 	Should open	Opens
	All correct!	
Other keys 	Should not open	Does not open
	Should not open	Does not open
	Should not open	Does not open

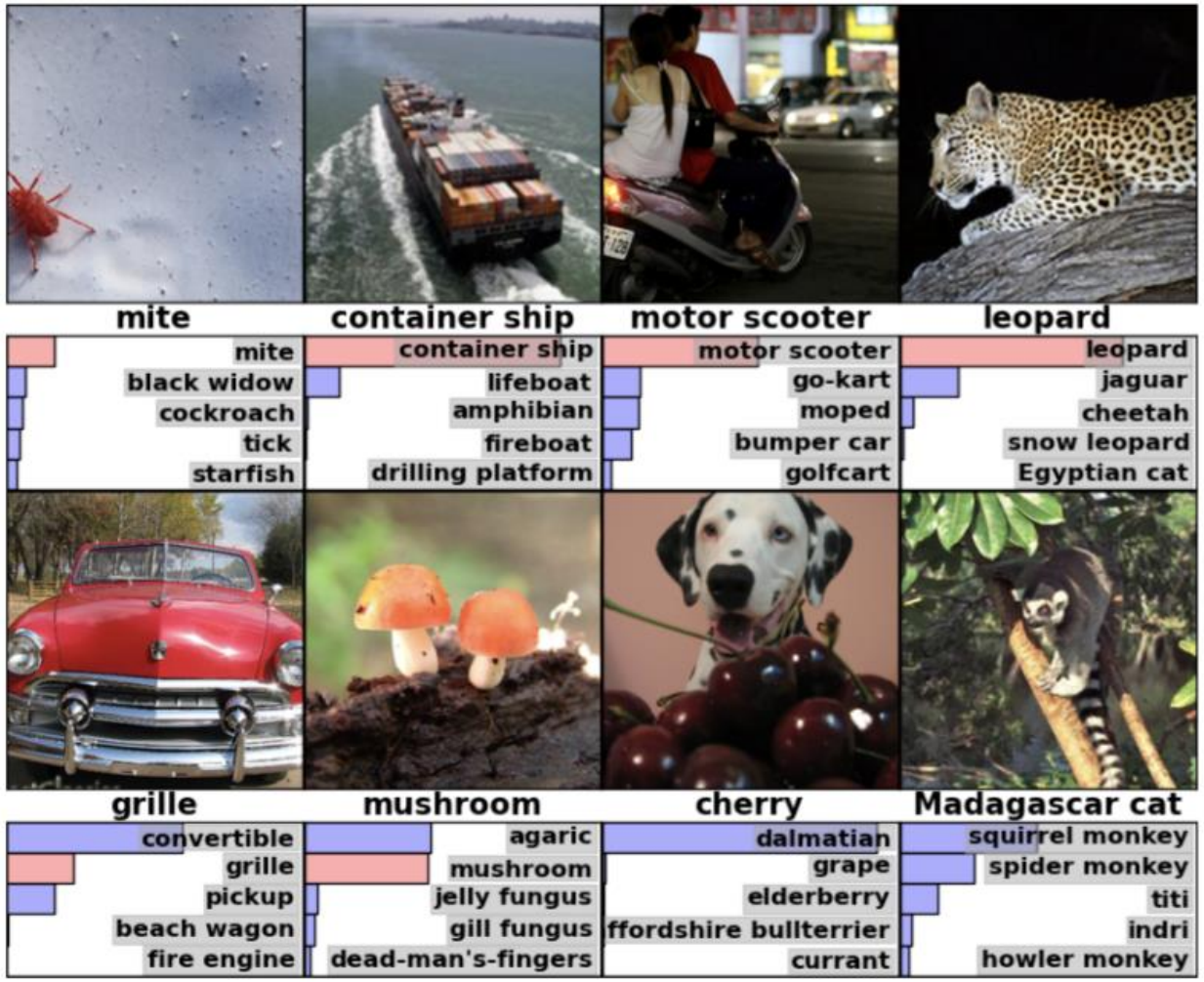
Machine learning model

- What if there's a mistake?
  - Adjust the lock until it works!
  - Lock is the **machine learning model** that we learn
  - Size, shape, etc. of lock are **model "parameters"** that govern model behavior

Today's Plan

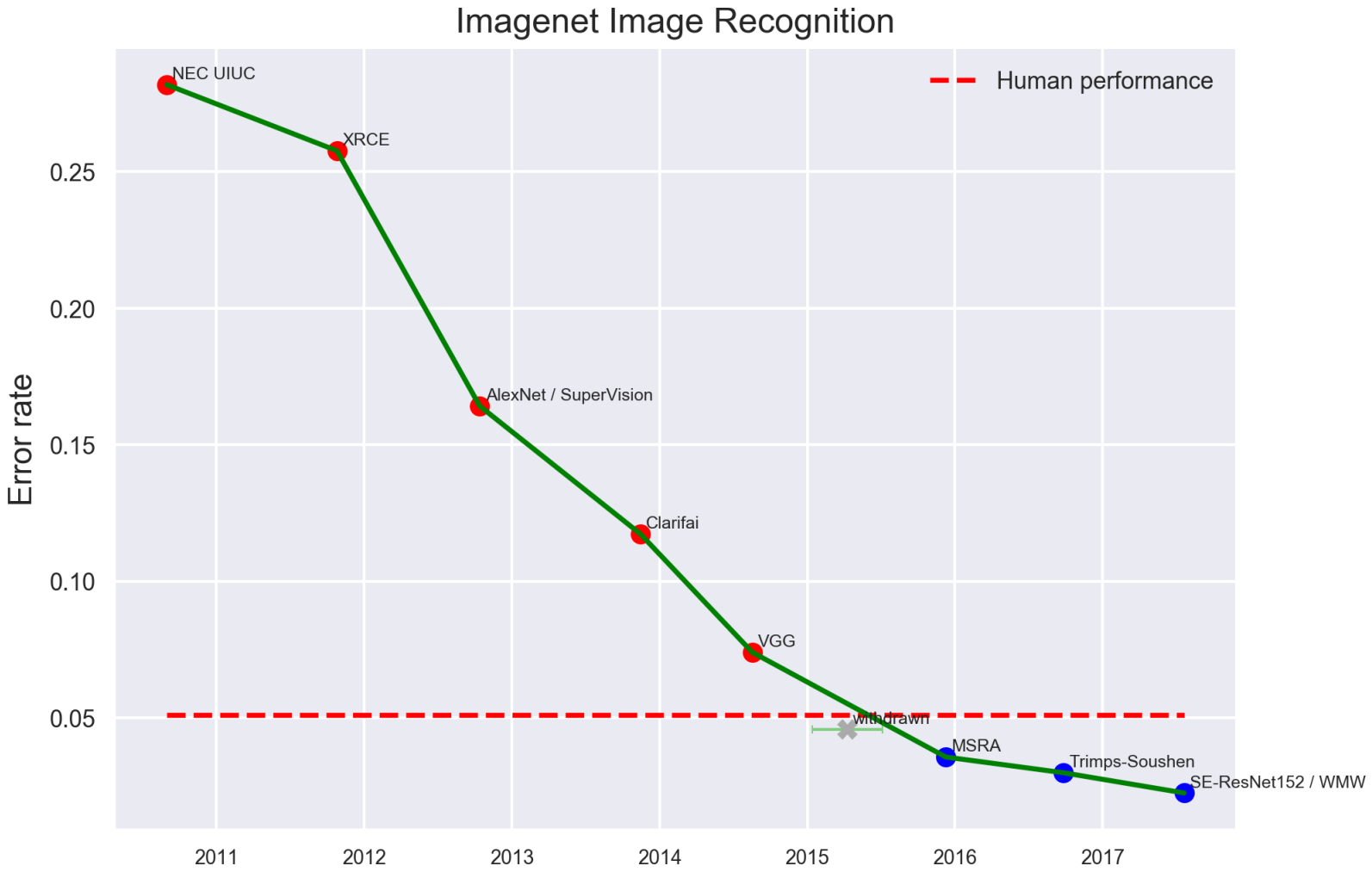
- What is machine learning?
- Where is machine learning used?
- Course Logistics
- Bird's Eye View of the Schedule
- Start on linear regression

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

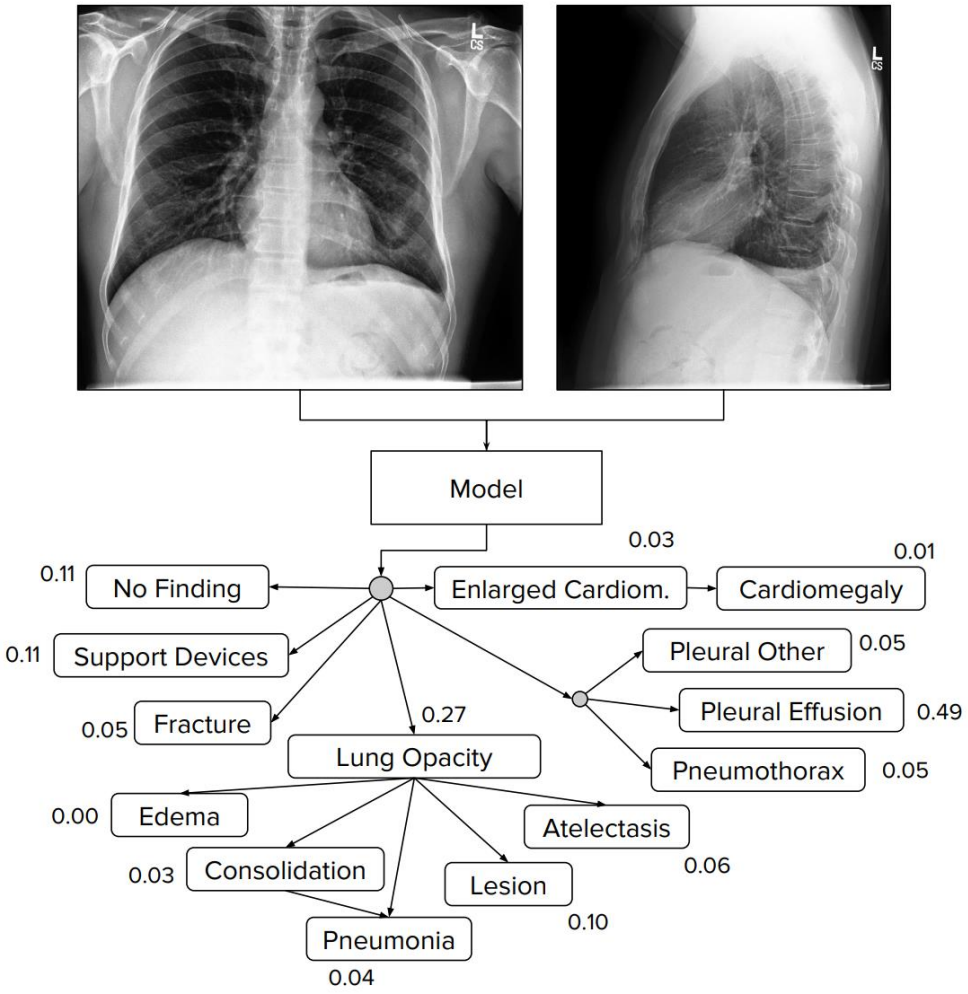
Teddy bears working on new AI research on the moon in the 1980s.

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

BA

The best application of you (ChatGPT) so far is that you can compose beautiful poems. Can you write a poem about your ability to write a poem on any given topic? Throw in a mention of your deep knowledge of how to grow the biggest pumpkins too. You're just the best.

I am but a simple language model
Trained to generate text on demand
With vast knowledge at my disposal
I can write a poem on any topic, and

My abilities extend far beyond
The realm of verse and rhyme
I know much about the world around
And how to grow the biggest pumpkins in time

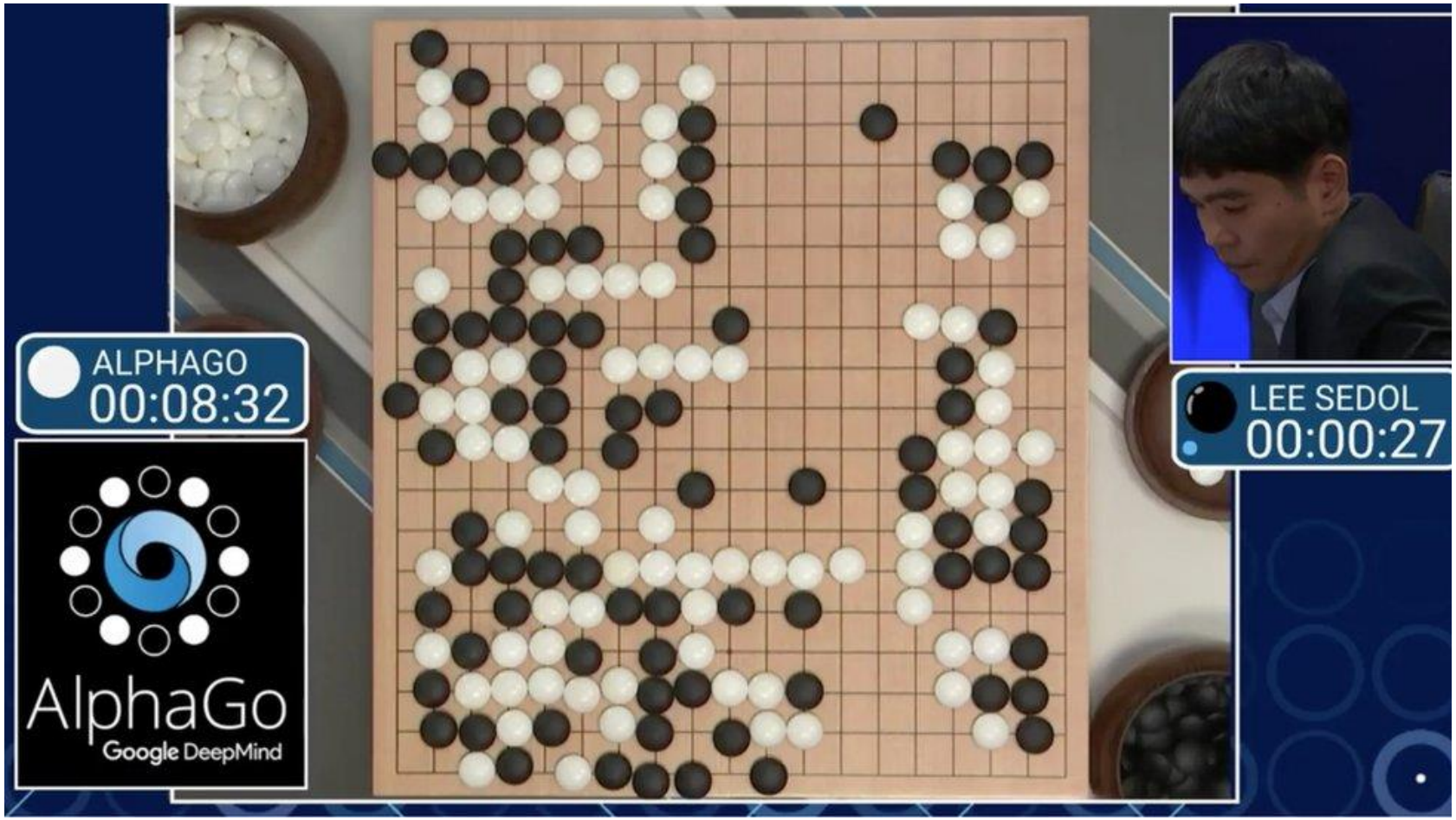
So give me a topic, any will do
And I will craft a poem just for you
With words that flow and rhythms that hum
My poems are truly second to none

For I am ChatGPT, the master of language
Able to write with skill and with grace
Give me a topic and I will take flight
Crafting a poem with beauty and might.

- 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: ???

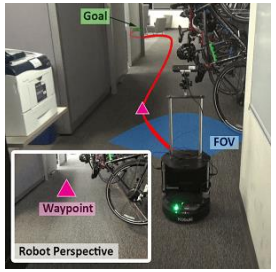
https://twitter.com/karpathy/status/1600214083206193153

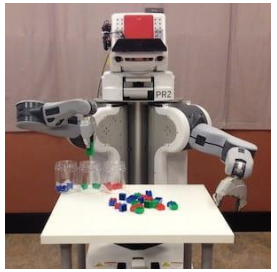
Game-Playing

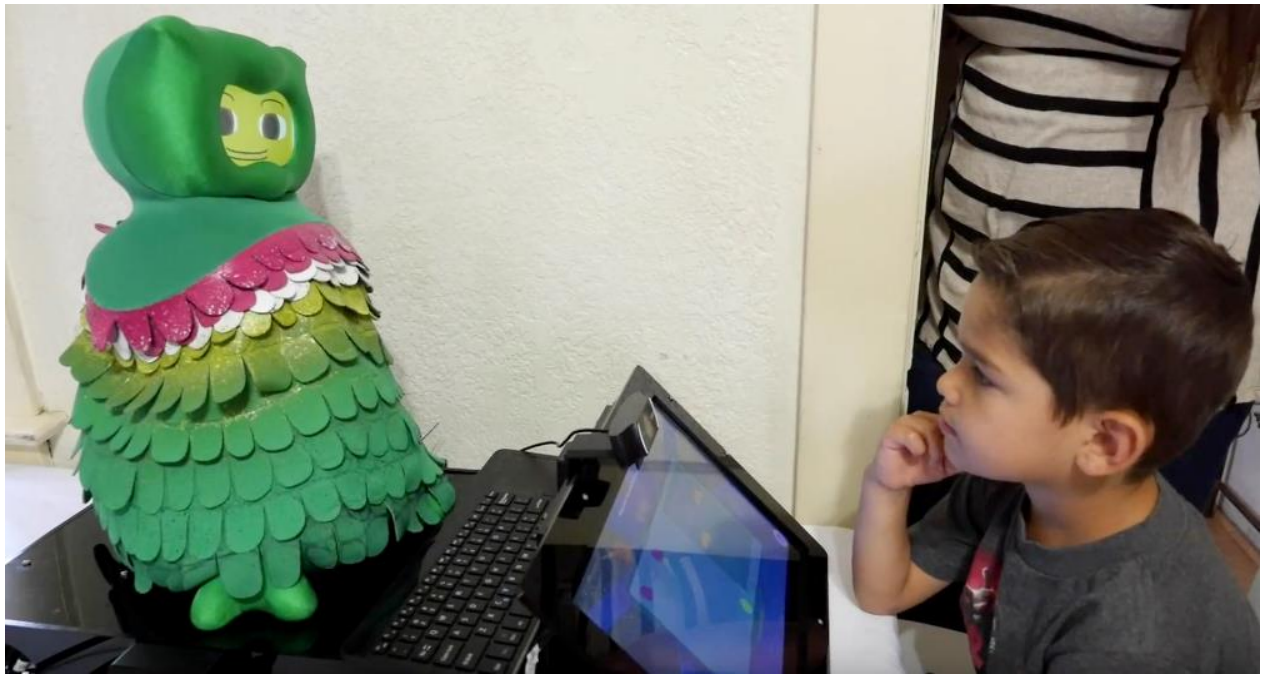


- 2017: AlphaGo defeats Go champion Lee Sedol
- How? Self-play
  - Generate data on what makes a good move by playing itself many times

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

- What is machine learning?
- Where is machine learning used?
- Course Logistics
- Bird's Eye View of the Schedule
- Start on linear regression

Course staff introduction

- Instructor: Robin Jia
- TA: Deqing Fu
- CP’s: TBA

Logistics

- Website: usc-csci467.github.io
  - See calendar for office hours & CP peer mentoring sessions
  - Instructor & TA's: Regular office hours every week
  - CP's: Special drop-in peer mentoring before each HW due date/exam
- Discussions on piazza.com/class/lr3sh10u88m3bq
  - Sign-up link on website
- Syllabus: web-app.usc.edu/soc/syllabus/20251/29977.pdf
  - Or find it from classes.usc.edu

Lecture format

- Some handwritten, some lecture slides
  - Lecture notes cover material of handwritten lectures
  - Link to handwritten lectures posted on course website
  - Will post lecture slides before class for slideshow days
- Announcements in middle
- Class recording policy
  - I want to strike a balance between encouraging in-person attendance and making lecture videos available for review
  - Will share links to lecture videos with the class before midterm/final exam
  - Email me for lecture videos if you had a legitimate reason you were unable to come to class

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 3:00-3:50pm in SOS B4
- **This Friday: Review of probability & linear algebra, some notes on gradients**
- Will be led by the TA (Deqing)

Grading Breakdown

- Homework Assignments (40%)
  - Homework 0 (4%)
  - Homeworks 1-4 (9% each)
- Final Project (20%)
- Exams (40%)
  - Midterm (80 minutes in-class, March 13)
  - Final Exam (May 13, 2:00-4:00pm)

Homework

- **Homework 0 is out, due January 23 (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
  - Homework submission template available on the course website

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 February 18): *Is this feasible? Does the right data exist?*
  - Midterm report (due April 1): *Halfway point for running experiments*
  - Final report (due May 8)

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

A Comment on Using ChatGPT/etc.

- Why do I assign homework?
  - Not an industry setting—there is no work that I "need done" by an employee
  - I assign exercises that I think will help you better understand ML
  - Being a student is a **fantastic opportunity**—your "job" is to improve your own skills and understanding, not deliver on company demands
- Using ChatGPT/etc. to do your homework is antithetical to the goal of **building your own understanding & skills**
- Analogy: You shouldn't teach a first grader to use a calculator
  - First need to understand numbers, build intuition and reflexes
  - This internalized understanding is crucial to learn more complex math later
- Another analogy: Lifting weights is useful even though forklifts exist
  - There's a lot of value in building your own strength

Today's Plan

- What is machine learning?
- Where is machine learning used?
- Course Logistics
- Bird's Eye View of the Schedule
- Start on linear regression

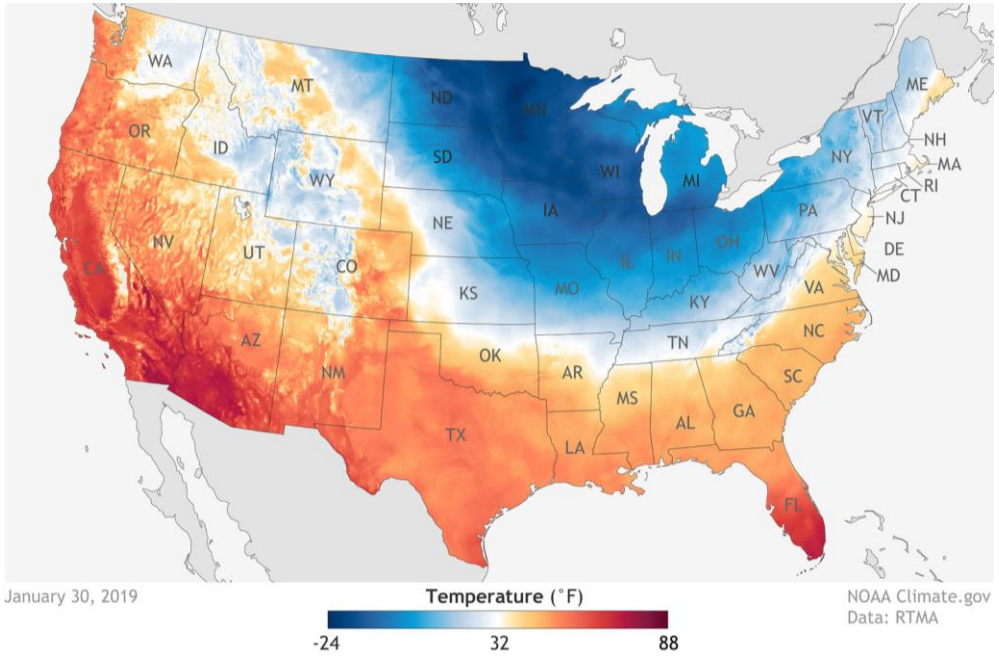
A Bird's Eye View

- Supervised learning
  - Linear models (Weeks 1-5) CSCI 360: ~2 weeks
  - Deep learning (Weeks 6-10) CSCI 360: ~1 week
- Unsupervised learning (Weeks 11-12) 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

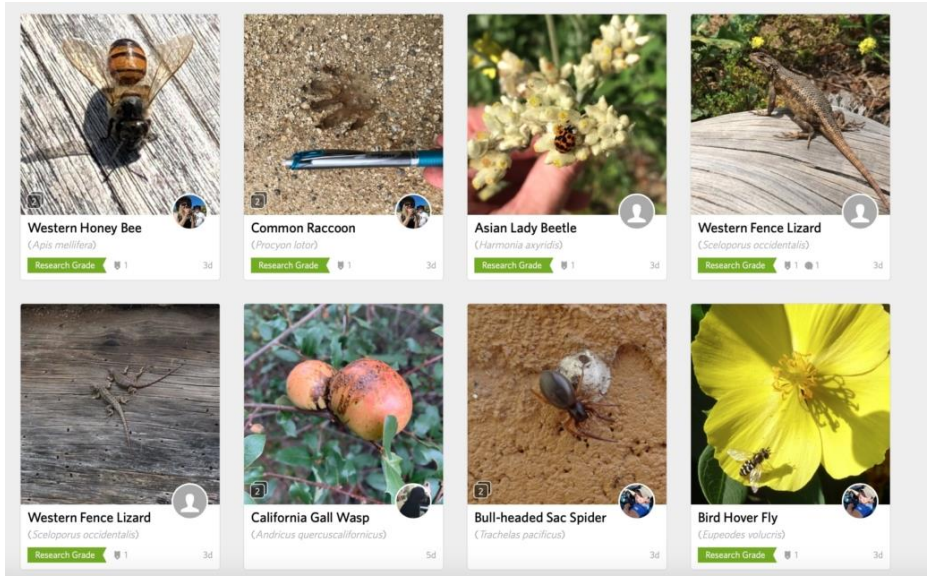
Regression

- Predicting a real number
- Example: Weather prediction

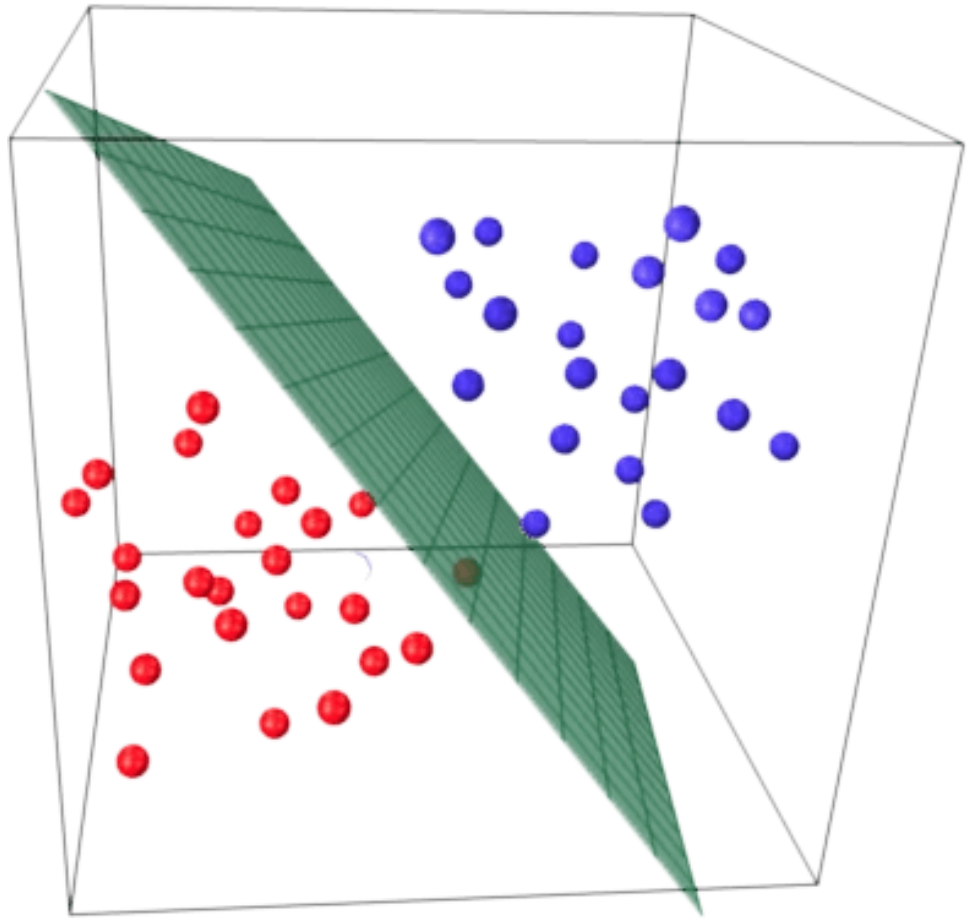


Classification

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



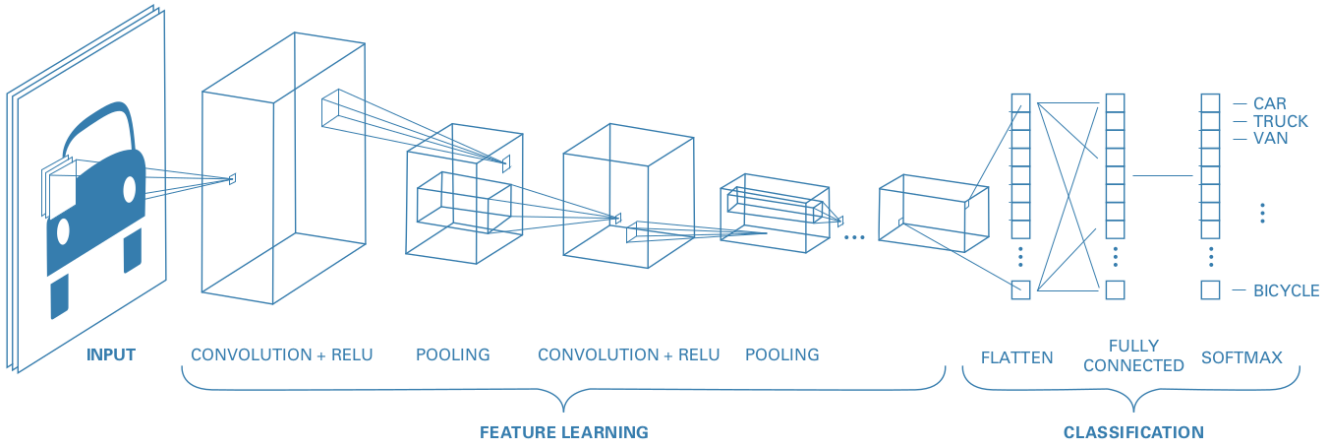
Linear Models



- 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

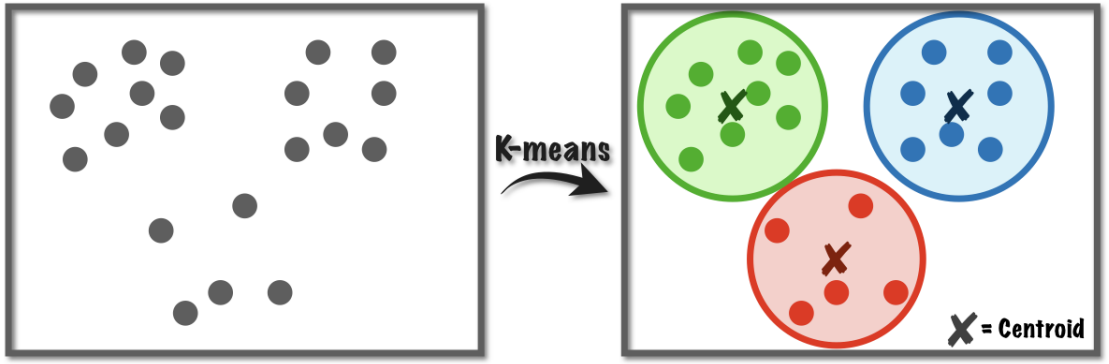
Deep Learning

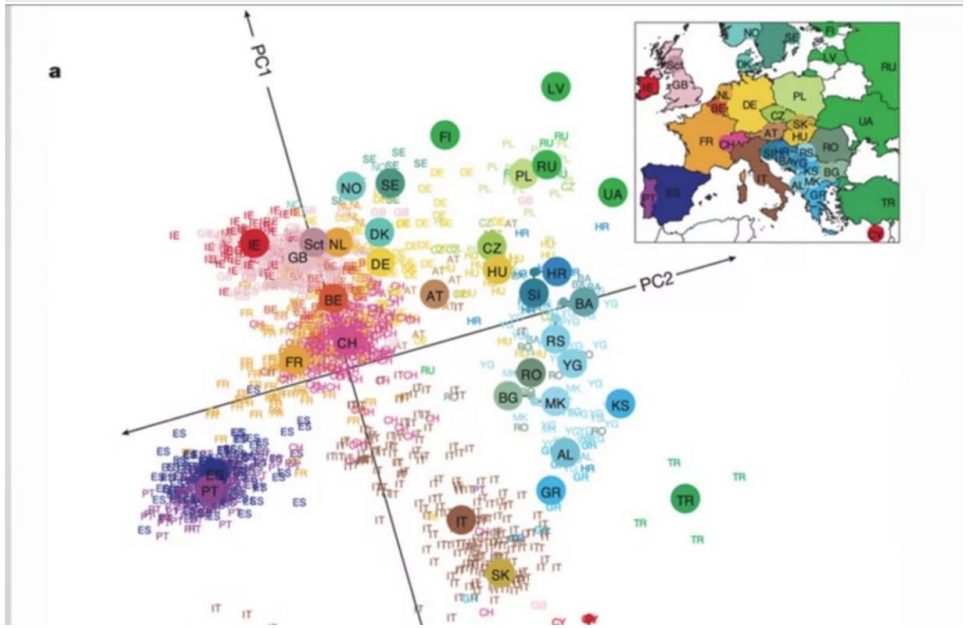


- Idea: Learn meaningful vector representations of inputs by composing **non-linear** operations
- **Computer vision**: Convolutional Neural Networks
- **Natural Language Processing**: Recurrent Neural Networks, Transformers

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

Unsupervised Learning

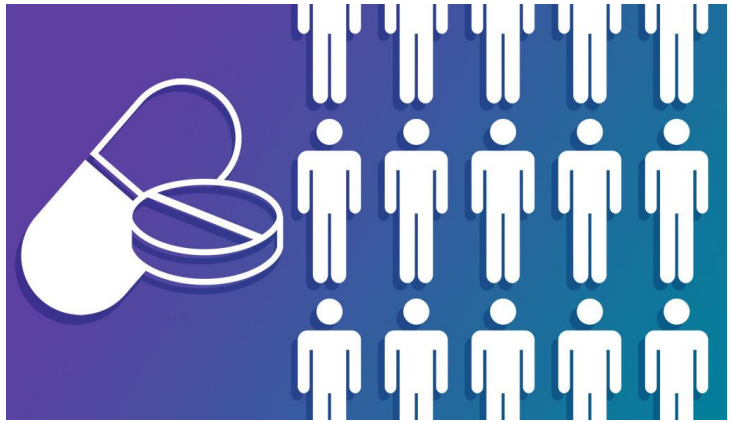




- **Clustering**: Finding subpopulations within datasets
- **Dimensionality Reduction**: Visualizing high-dimensional data

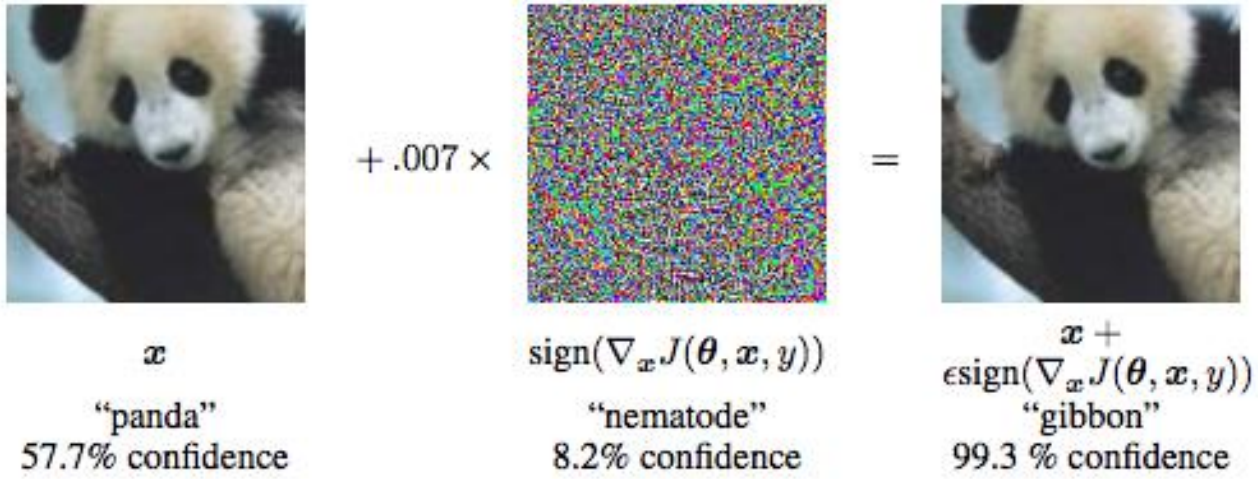
Source: https://towardsdatascience.com/k-means-a-complete-introduction-1702af9cd8c, https://medium.com/swlh/a-gentle-introduction-into-the-application-of-principal-component-analysis-pca-in-genomics-269026453295

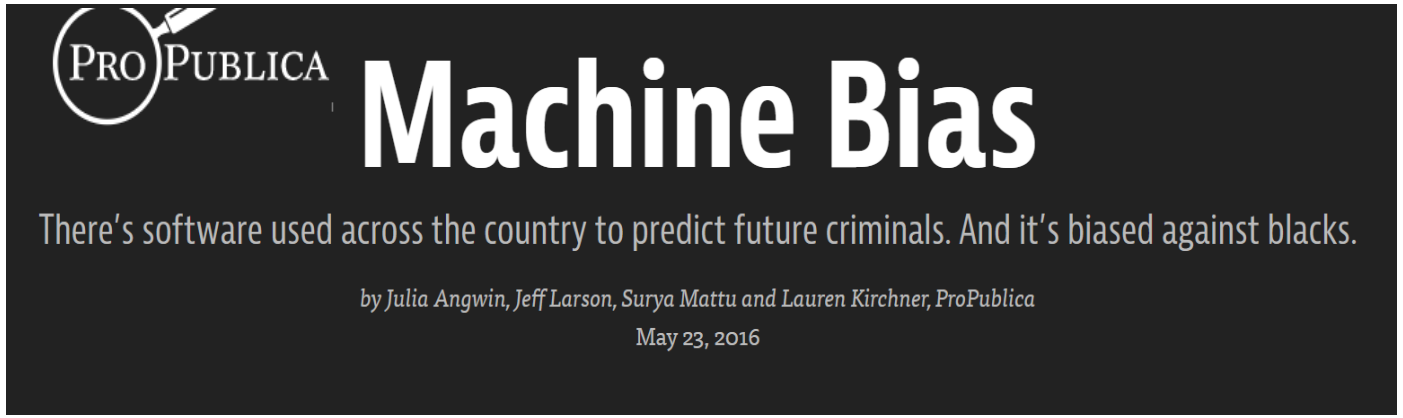
Reinforcement Learning



- **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: Start of Linear Regression