

Yale:

Societal Issues for Machine Learning

Yale

Housekeeping

- Assignment 5 due today at midnight
- Quiz 6: Thursday; usual protocol
- Final exam, Friday, Dec 12 at 2pm in SSS 114 (cumulative, 3 hours, cheat sheet, practice exams posted to Canvas)
- Thursday (last class): Review for final exam

Outline

- Recall: ML vs. AI
- Examples of bias
- Recent discussion online
- Panel discussion

Today: Home assistants



Pricing and recommending homes

THE WALL STREET JOURNAL.

Subscribe Now | Sign In

\$1 for 2 months

Home World U.S. Politics Economy Business Tech Markets Opinion Arts Life Real Estate 



CIO JOURNAL



Zillow Develops Neural Network to 'See' Like a House Hunter

Granite or stainless steel countertops? Zillow's visual recognition effort can recognize the difference

By **SARA CASTELLANOS**

Nov 11, 2016 3:29 pm ET

Data scientists at Zillow Group are developing complex computer programs that detect specific attributes in photographs of homes, which could aid in estimating their value. Advances in deep learning, big data and cloud computing have converged to allow the online real estate database firm and others to develop technology that mimics how the human brain [...]

Recommended Videos

1. Film Clip: Pirates of the Caribbean: Dead Men Tell No Tales'



2. What to do in your 40s to retire a millionaire



<https://blogs.wsj.com/cio/2016/11/11/zillow-develops-neural-network-to-see-like-a-home-buyer/>

AI vs. ML

Machine learning focuses on making predictions and inferences from data.

AI combines machine learning components into a larger system that includes a decision making component.

An AI system exhibits a behavior, resulting from the collective decisions that are made.

Machine learning frameworks

- Supervised, unsupervised, semi-supervised
- Reinforcement learning
- Generative vs. discriminative models
- Representation learning

Example of representation learning: Word embeddings

- Each word in vocab is mapped to 100 or 500 dimensional vector
- Based solely on co-occurrence statistics in corpus of text

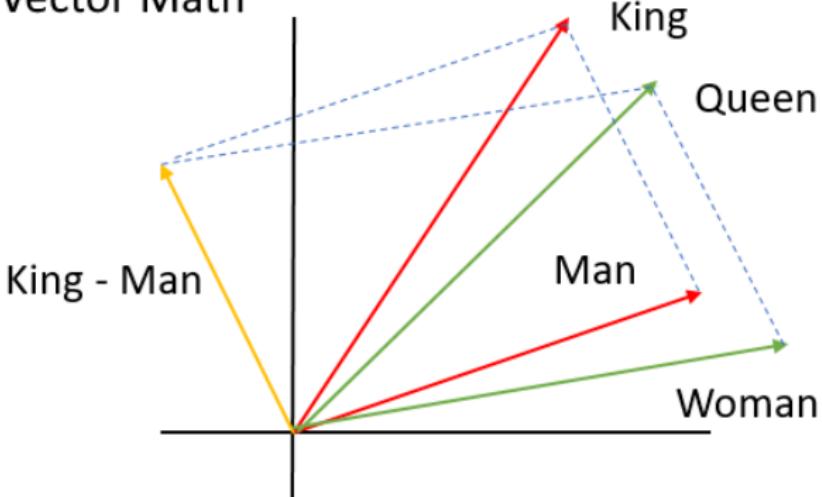
Example of representation learning: Word embeddings

Yale:

```
[ 0.78310001, 0.51717001, -0.38207 , -0.23722 , -0.31615999, 0.30805001, 0.76389998, 0.064106 , -0.74913001,  
 0.60585999, -0.23871 , -0.16876 , -0.25634 , 1.07270002, -0.29967999, 0.020095 , 0.54500997, -0.17847 , -0.26675999,  
 -0.11798 , -0.48692 , 0.22712 , 0.017473 , -0.4747 , 0.44861001, -0.084281 , -0.30412999, -1.13510001, -0.14869 , -0.11182 ,  
 -0.32530001, 1.0029 , -0.35742 , 0.35148999, -1.10679996, -0.064142 , -0.72284001, 0.14114 , -0.41247001, -0.16184001,  
 -0.54576999, -0.12958001, -0.88356 , -0.089722 , 0.10555 , -0.12288 , 0.92851001, 0.50032002, 0.1349 , 0.21457 ,  
 0.35073999, -0.73132998, 0.39633 , -0.43239999, -0.38815999, -1.34669995, 0.37463999, -0.79386002, 0.11185 , 0.18007 ,  
 -0.75142998, 0.24975 , -0.094948 , -0.36341 , 0.24869999, -0.22667 , 0.32289001, 1.29489994, 0.42658001, 1.29120004,  
 -0.13954 , 0.68976003, 0.21586999, 0.13715 , -1.00919998, 0.028827 , 0.11011 , -0.1912 , -0.073198 , -0.52449 , 0.49199 ,  
 0.14463 , -0.18844 , -0.75536001, -0.28704 , 0.019113 , 0.30349001, -0.74425 , -0.072221 , -0.40647 , 0.26899001, -0.28318  
, 0.72409999, 0.50796002, -0.37845999, -0.13008 , -0.13808 , 0.098928 , 0.16215999, 0.16293 ]
```

Word geometry

Vector Math



Embeddings encode societal bias

$$\phi(\text{scientist}) - \phi(\text{woman}) + \phi(\text{man}):$$

geologist
engineer
astronomer
mathematician
science

$$\phi(\text{scientist}) - \phi(\text{man}) + \phi(\text{woman}):$$

anthropologist
sociologist
psychologist
geneticist
biochemist

Embeddings encode societal bias

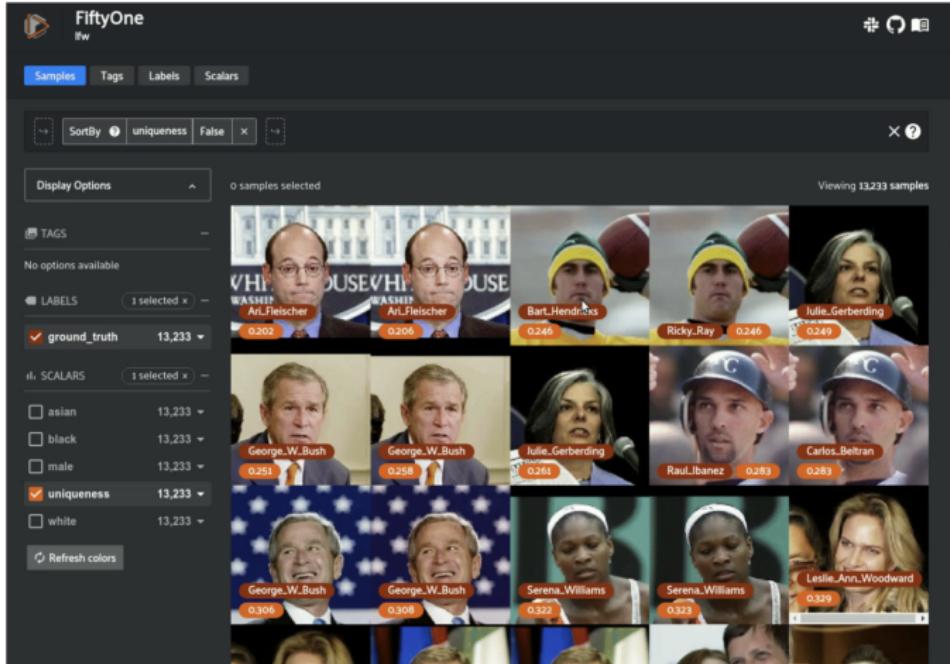
$$\phi(\text{smart}) - \phi(\text{girl}) + \phi(\text{boy}):$$

wise
better
guy
kind
good
kid

$$\phi(\text{smart}) - \phi(\text{boy}) + \phi(\text{girl}):$$

sexy
pretty
incredibly
cute
exciting
funny

Bias in LFW dataset



Sorting by the least unique images to find duplicates and incorrect labels

Hacking AI systems



SUBSCRIBE



SIGN IN ▾

TESLA AUTOPILOT —

Researchers trick Tesla Autopilot into steering into oncoming traffic

Stickers that are invisible to drivers and fool autopilot.

DAN GOODIN - 4/1/2019, 8:50 PM

Keen Security Lab



Machine learning at a large Internet company

- Typical project lifetime: 6 months to 1 year
- Ads projects involve thousands of software engineers
- Often adding new “feature” to existing black box model
- No single person understands entire model
- Not interpretable
- Security issues with data

Important directions

Reliability

- Design AI systems robust to noise and/or adversaries

Fairness

- Design AI systems that can check and control human-biases

Accountability

- Design AI systems that can be held accountable/liable

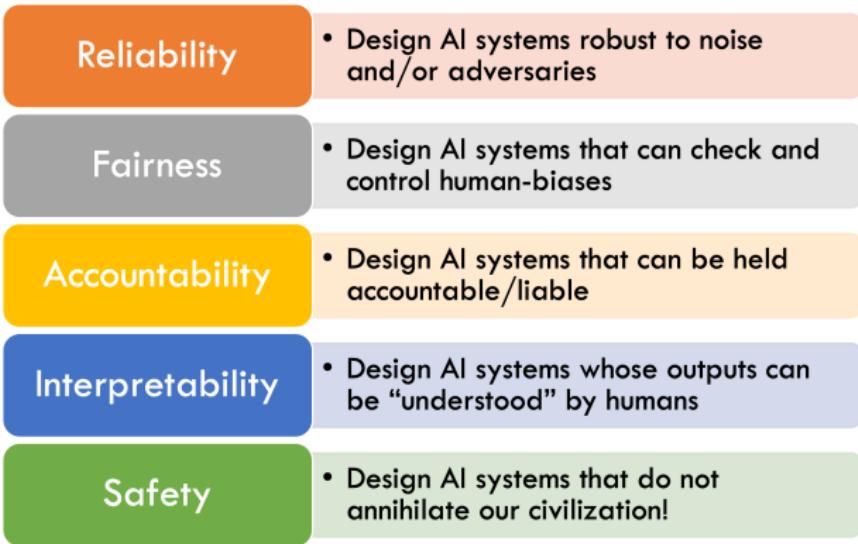
Interpretability

- Design AI systems whose outputs can be “understood” by humans

Safety

- Design AI systems that do not annihilate our civilization!

Important directions



- Requires a principled, interdisciplinary approach

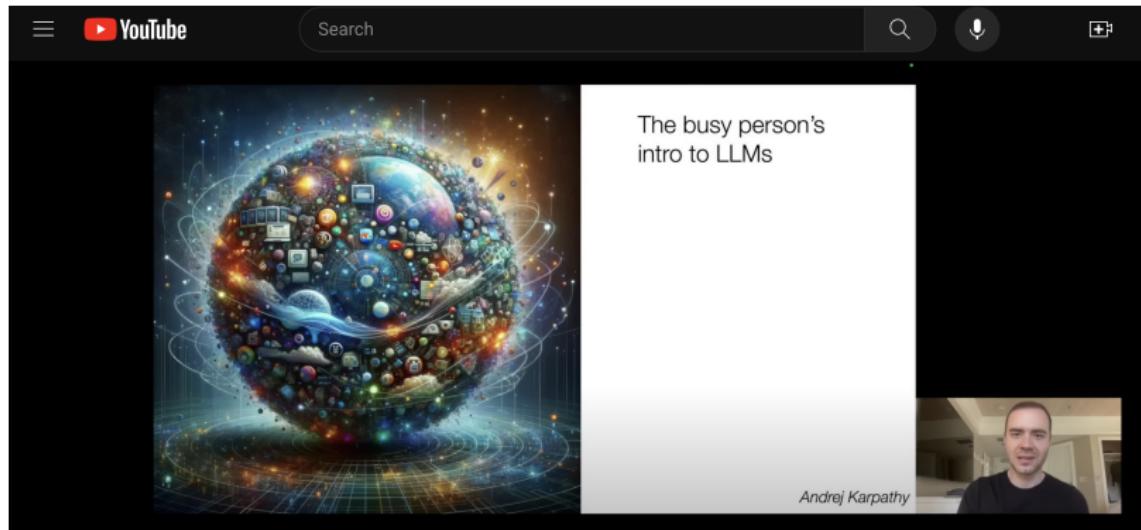
Recommended listening

The screenshot shows a podcast episode page. At the top right, the date "Dec. 1, 2023" is displayed. The main title of the episode is "A Lot Has Happened in A.I. Let's Catch Up." Below the title, a subtitle reads: "The tech journalists Kevin Roose and Casey Newton discuss developments in A.I., a year after the launch of ChatGPT." To the right of the subtitle is a "Transcript" button. At the bottom left, there are playback controls: a circular icon with a double arrow, a circular icon with a play button, and another circular icon with a double arrow. Next to these icons is the word "Listen" followed by the duration "1:13:04".

<https://www.nytimes.com/2023/12/01/opinion/ezra-klein-podcast-casey-newton-kevin-roose.html>

Posted December 1, 2023

Recommended viewing



https://www.youtube.com/watch?v=zjkBMFhNj_g

Posted November 22, 2023

Recommended viewing



https://www.youtube.com/watch?v=zjkBMFhNj_g&t=45m43s

Jailbreaks: https://www.youtube.com/watch?v=zjkBMFhNj_g&t=46m15s

Prompt injection: https://www.youtube.com/watch?v=zjkBMFhNj_g&t=51m30s

Data poisoning: https://www.youtube.com/watch?v=zjkBMFhNj_g&t=56m23s

Panel Discussion

Team A

Dystopian view: AI and ML are going to result in major societal inequities. The harm may outweigh the benefits unless we are very careful, including corporations and government entities.

Debbie, Raphaëlle, Helen, Halli, Ben

Panel Discussion

Team B

Utopian view: AI and ML are going to result in major societal benefits. The technology is unprecedented, and people will adapt to harness it in ways that will improve the human condition.

Luke, Taylor (MC), Anmol, Selma, Joanne