

S&DS 265 / 565
Introductory Machine Learning

Societal Issues in ML

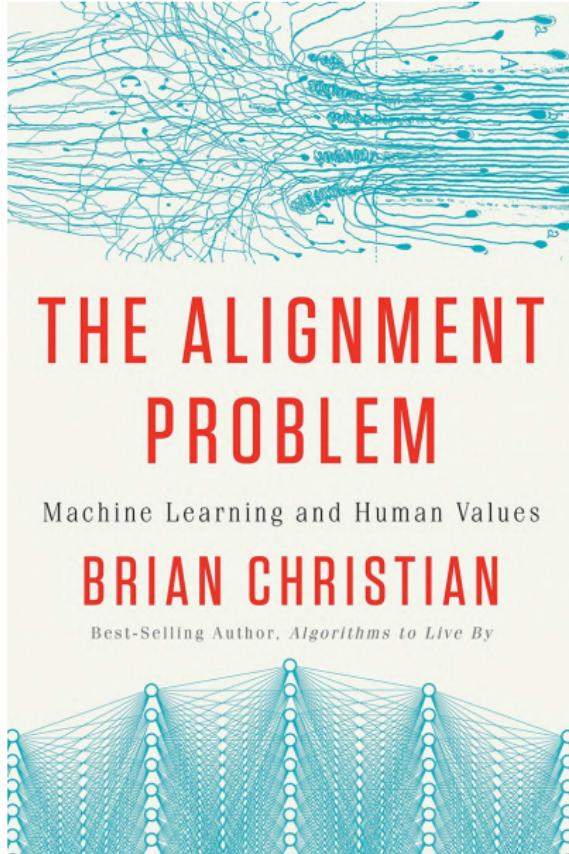
December 6

Yale

Reminders

- Quiz 6 posted today at 10:30—available until December 10
- Assn 6 out; due next week
- Final exam: Monday Dec 19, 7pm in Davies
- Practice exams are posted
- Review sessions TBA

Today: Alignment



It's everywhere: 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
- 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

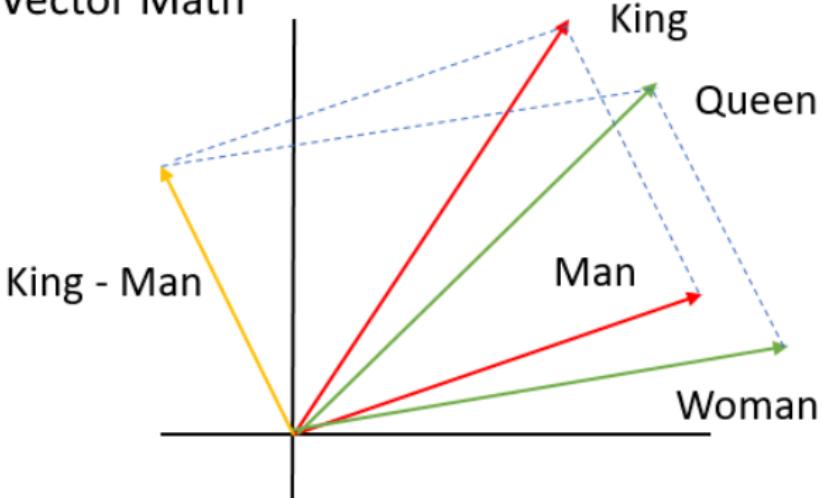
Example of representation learning: Word embeddings

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

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi¹, Kai-Wei Chang², James Zou², Venkatesh Saligrama^{1,2}, Adam Kalai²

¹Boston University, 8 Saint Mary's Street, Boston, MA

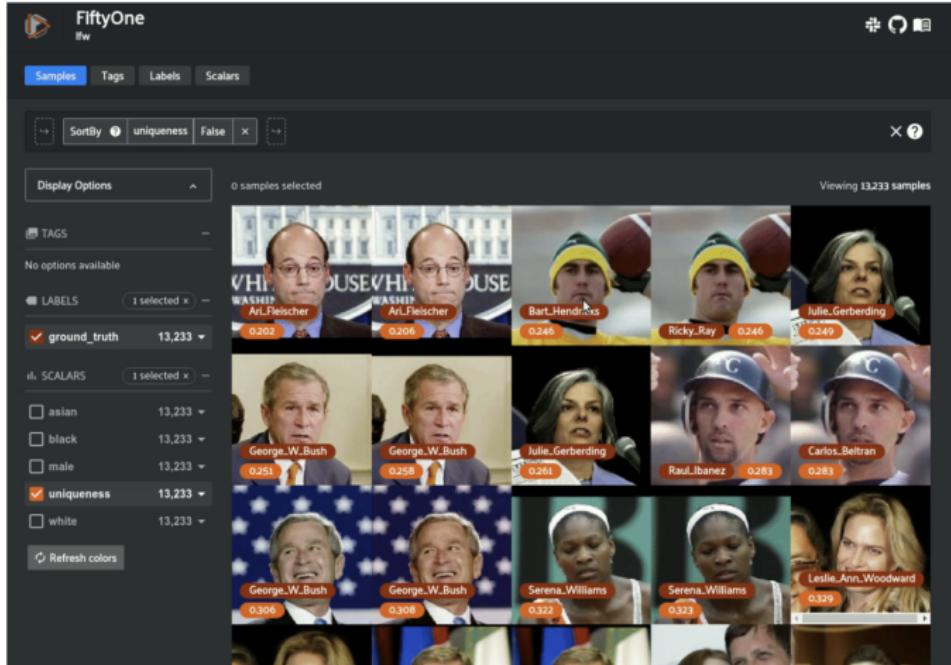
²Microsoft Research New England, 1 Memorial Drive, Cambridge, MA

tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

Abstract

The blind application of machine learning runs the risk of amplifying biases present in data. Such a danger is facing us with *word embedding*, a popular framework to represent text data as vectors which has been used in many machine learning and natural language processing tasks. We show that even word embeddings trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent.

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

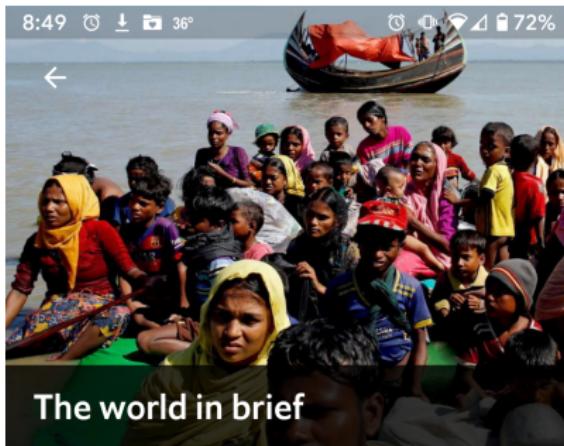
Keen Security Lab



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

From recent news feed



Rohingya refugees, members of an ethnic minority forcibly driven from **Myanmar**, filed a class-action lawsuit against **Facebook** (through its parent, Meta) claiming damages worth \$150bn. Their lawyers say the social-media giant neglected to prevent incitements of **violence** against them. Facebook has said it was “too slow to prevent misinformation” in Myanmar, but also argued it is not liable for the effects.

Descriptions in press

The Scientist and the A.I.-Assisted, Remote-Control Killing Machine

Israeli agents had wanted to kill Iran's top nuclear scientist for years. Then they came up with a way to do it with no operatives present.



Descriptions in press

Jerusalem Post > Middle East > Iran News

Iran denies NYT Mossad assassination report

The New York Times published a report detailing the assassination of Iran's leading nuclear scientist by a Mossad-operated AI machine gun.

By JERUSALEM POST STAFF Published: SEPTEMBER 19, 2021 20:36



Descriptions in press

NONFICTION

A Robot Wrote This Book Review



Elliot Ulm

Descriptions in press

By **Kevin Roose**

Nov. 21, 2021

THE AGE OF AI

And Our Human Future

By Henry A. Kissinger, Eric Schmidt and Daniel Huttenlocher

One of the great promises of technology is that it can do the work that humans find too boring or arduous.

In the 19th and 20th centuries, factory machines relieved us of repetitive manual labor and backbreaking farm work. In this century, artificial intelligence has taken care of a few more tasks — curating Spotify playlists, selecting the next YouTube video, vacuuming the floor and so on — but many more mind-numbing activities remain ripe for the picking. The experts promise us that someday, all of our least favorite chores — including complex cognitive ones, like interviewing job candidates or managing global supply chains — will be outsourced to machines.

But that day has not yet arrived. Or has it?

Corporate (de)initiatives

DeepMind → Safety & Ethics

We want AI to benefit the world, so we must be thoughtful about how it's built and used

On this page ↓

- OVERVIEW
- TECHNICAL SAFETY
- ETHICS & SOCIETY

Corporate (de)initiatives

Elon Musk Has Fired Twitter's 'Ethical AI' Team

As part of a wave of layoffs, the new CEO disbanded a group working to make Twitter's algorithms more transparent and fair.



Corporate (de)initiatives

TOM SIMONITE

BUSINESS 12.02.2021 08:00 AM

Ex-Googler Timnit Gebru Starts Her Own AI Research Center

The researcher, who says Google fired her a year ago, wants to ask questions about responsible use of artificial intelligence.



ILLUSTRATION WIRED STAFF; GETTY IMAGES

Corrigibility



Corrigibility

"If we use, to achieve our purposes, a mechanical agency with whose operation we cannot efficiently interfere once we have started it...then we had better be quite sure that the purpose put into the machine is the purpose which we really desire and not merely a colorful imitation of it."

Norbert Wiener

Moral and Technical Consequences of Automation, 1960

Further listening



PLAY ►

Is A.I. the Problem? Or Are We?

The Ezra Klein Show

Society & Culture

[Listen on Apple Podcasts ↗](#)



If you talk to many of the people working on the cutting edge of artificial intelligence research, you'll hear that we are on the cusp of a technology that will be far more transformative than simply computers and the internet, one that could bring about a new industrial revolution and usher in a utopia — or perhaps pose the greatest threat in our species's history.

Others, of course, will tell you those folks are nuts.

Our intrepid panelists:

Anjali Gupta, Shankara Abbineni, Clark Fisher, Iman Jaroudi,
Aparajita Kaphle, Corin Katzke, Luke Reynolds, Matthew Shu, Kelly
Wang

Questions for reflection

Should machines be held to a higher standard than humans?

- ▶ A self-driving car kills a pedestrian
- ▶ A drunk driver kills a pedestrian

Questions for reflection

Will AI create jobs or destroy jobs?

- ▶ A robot replaces a factory worker
- ▶ A robot supervisor is needed

Questions for reflection

How will AI affect the global balance of power and wealth?

- ▶ A rich country adopts an AI research program
- ▶ A poor country adopts open technology

Questions for reflection

How important is communication of machine learning technology?

- ▶ Public misconceptions are endemic
- ▶ People eventually gain a basic understanding

Questions for reflection

How should history guide development of AI?

- ▶ Technology is neither inherently good or bad (nuclear fission)
- ▶ This is completely new in history

Questions for reflection

How likely is the development of a superintelligent AGI within the next century? What are some risks of an AGI unaligned with human values?