S&DS 265 / 565

Introductory Machine Learning

Societal Issues for Machine Learning

Tuesday, December 3

Yale

Housekeeping

- Assignment 5 due Thursday December 5 at midnight
- Quiz 6 (last!): Thursday; usual protocol
- Final exam, Monday, Dec 16 at 2pm (cumulative, 3 hours, cheat sheet, practice exams posted to Canvas)
- Thursday (last class): Vote a topic off the final!
- Review sessions TBA

Outline

- Recall: ML vs. AI
- Examples of bias
- Security/safety issues in LLMs
- Panel discussion

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

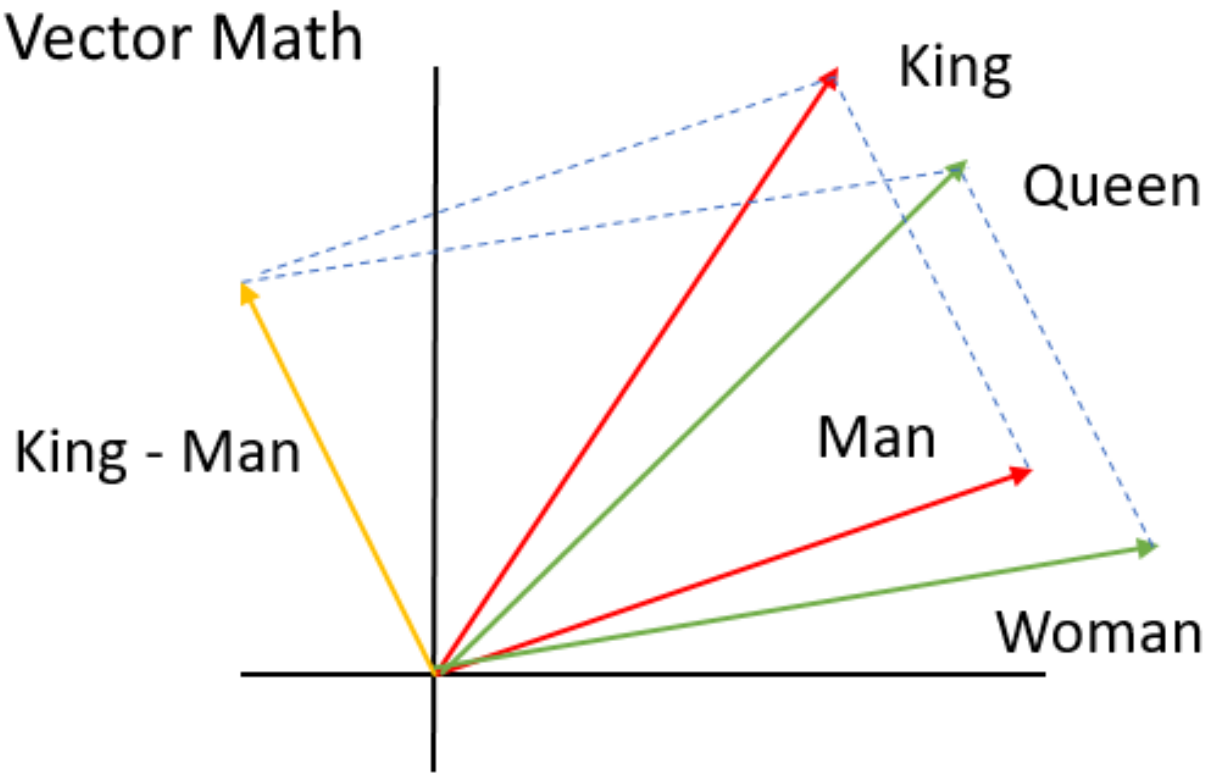
- Each word in vocab is mapped to 100 or 500 dimensional vector
- Based solely on co-occurrence statistics in corpus of text

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



Embeddings encode societal bias

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

```
geologist
engineer
astronomer
mathematician
science
```

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

```
anthropologist
sociologist
psychologist
geneticist
biochemist
```

Trained on 6 billion words of Wikipedia

Embeddings encode societal bias

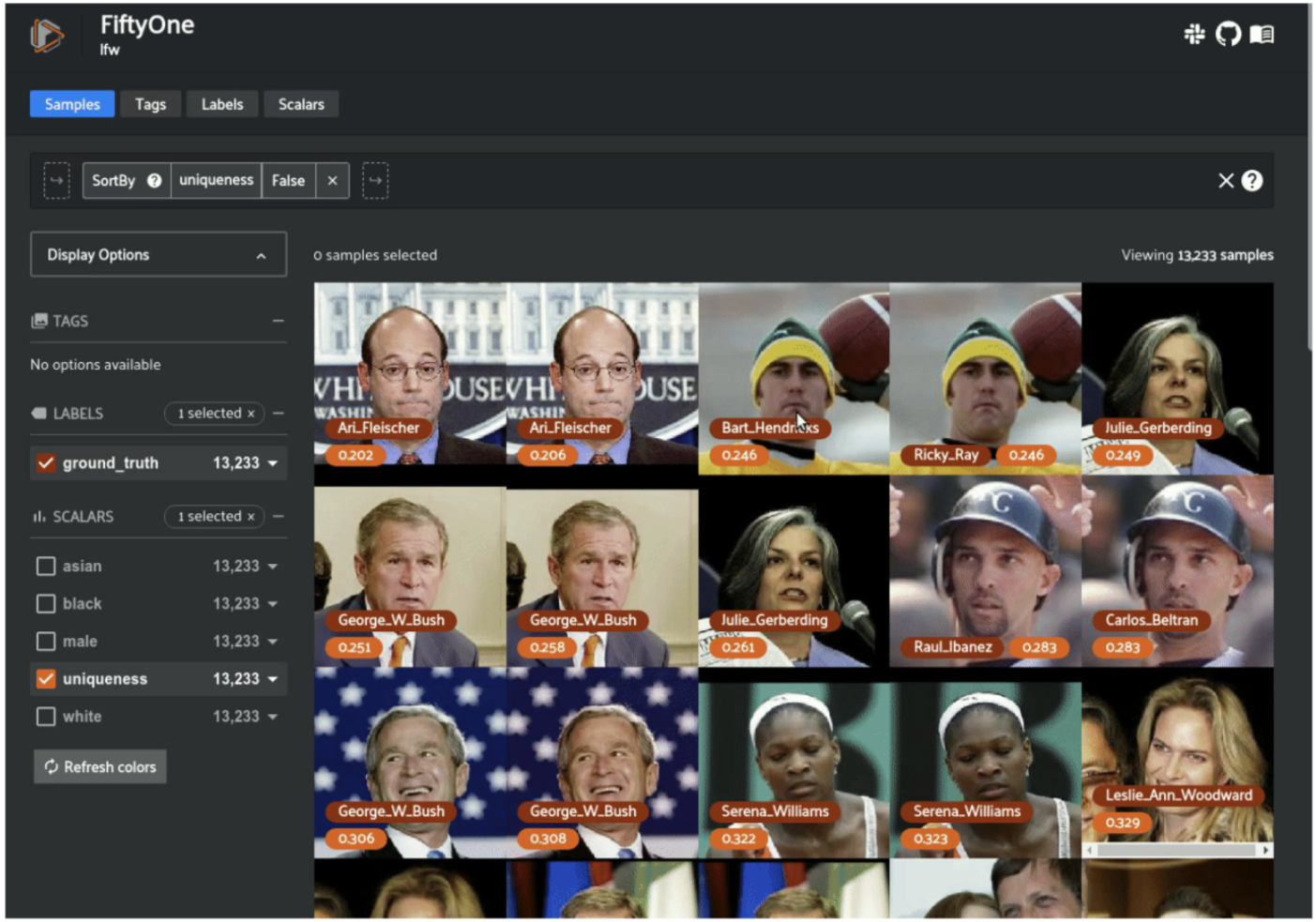
$\phi(\texttt{smart}) - \phi(\texttt{girl}) + \phi(\texttt{boy})$:

```
wise
better
guy
kind
good
kid
```

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

```
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

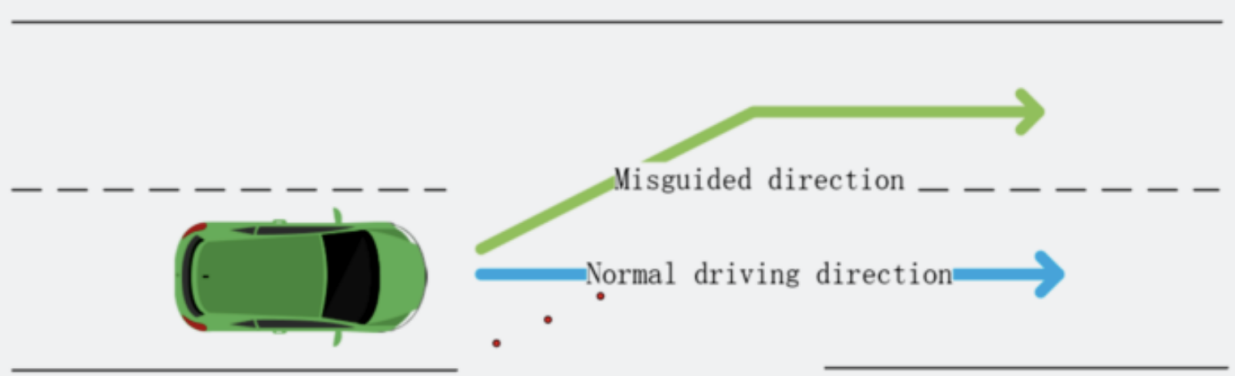
ars TECHNICA SUBSCRIBE SIGN IN

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!

Ack: Nisheeth Vishnoi and Elisa Celis

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!

- Requires a principled, interdisciplinary approach

Ack: Nisheeth Vishnoi and Elisa Celis

Always interesting

APR 2 · 1H 15M

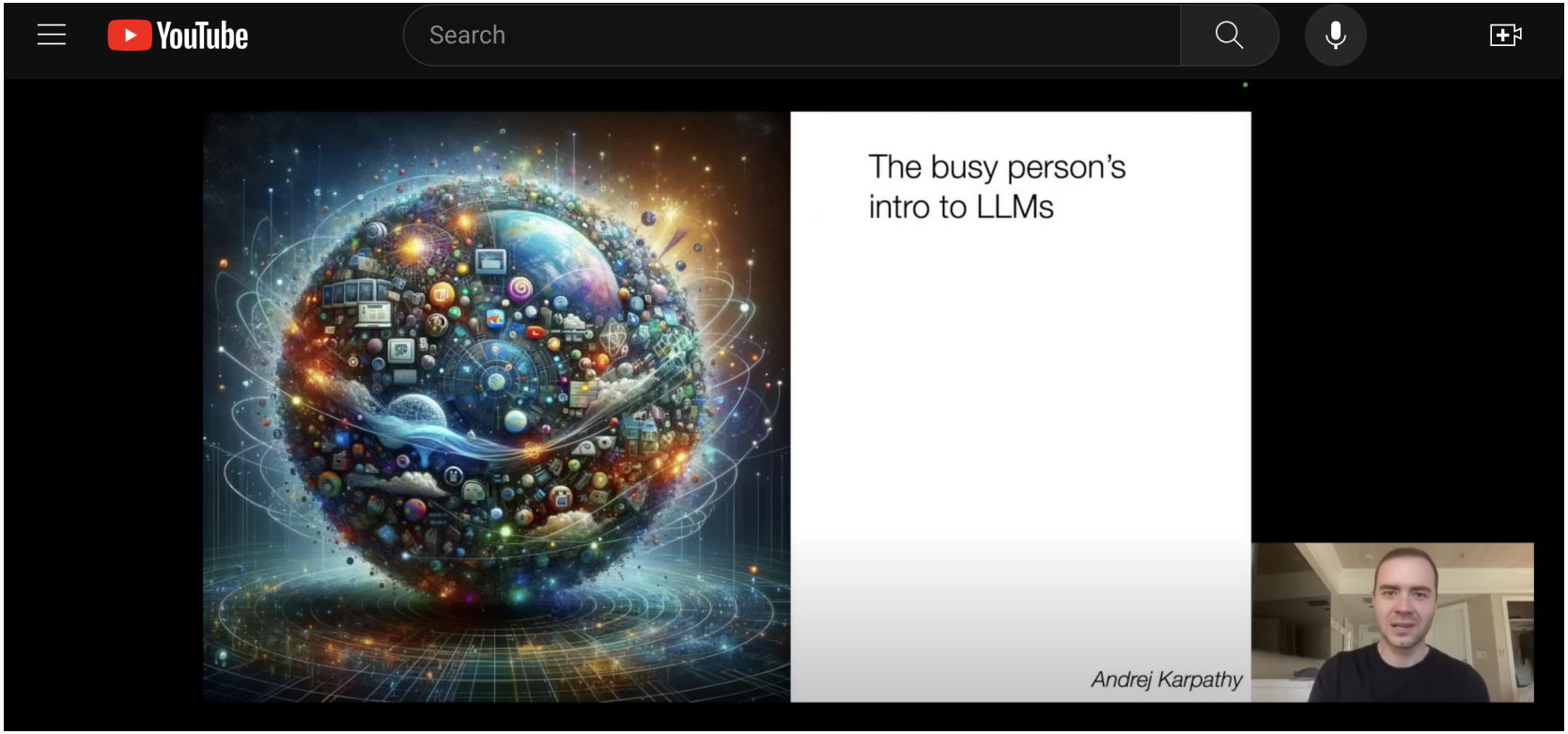
How Should I Be Using A.I. Right Now?

The Ezra Klein Show

Play... Subscribers Only

There’s something of a paradox that has defined my experience with artificial intelligence in this particular moment. It’s clear we’re witnessing the advent of a wildly powerful technology, one that could transform the economy and the way we think about art and creativity and the value of human work itself. At the same time, I can’t for the life of me figure out how to use it in my own day-to-day job.

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

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.

Dongyu, Sasha, Tobias

Panel Discussion

Team B

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.

Yogev, Amanda, Hantao

Prompts

Can AI systems ever be truly ethical, or will they always carry the biases of their creators?

Prompts

Will AI be a force for economic equality or inequality?

Prompts

Who should be responsible for ensuring that AI remains safe and beneficial for humanity?

Prompts

What do you see as the most transformative potential of AI in society?

Prompts

What are the implications of AI for art, music, and other creative fields?

Prompts

How important is public communication and education about AI technology?