Artificial Intelligence

Steven M. Bellovin

What is Artificial Intelligence (AI)?

- You know what artificial intelligence (AI) is—computer programs that "think" or otherwise act "intelligent"
  - The Turing test?
- What is "machine learning" (ML)?
  - It's simply one technique for AI—throw a lot of data at a program and let it figure things out
- What are "neural networks"?
  - A currently popular technique for ML

AI is Old

- Artificial intelligence is one of the oldest non-numerical uses of computers
  - (Of course, today it does use numeric techniques)
- Turing discussed AI in 1950
- 1956 goal: machine translation
  - Context: the Cold War, and the consequent need to translate Russian documents
  - (“The vodka was good but the meat was rotten”)

No One Knew How to Do AI…

- Early attempts at emulating the brain failed
  - No one really knew how the brain worked
- Instead, researchers said, “An intelligent being can do X. We’ll try to do X by computer and say that that’s AI research”
- So: chess-playing, vision, natural language comprehension, and more
- *None of these, as done by computers, are really related to the general problem!*

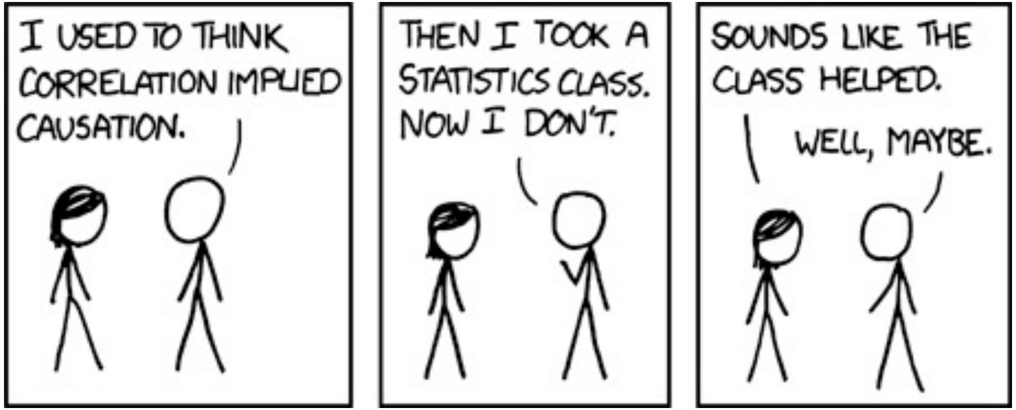
How Does ML Work?

- Lots of complicated math
- *Not* the way human brains with human neurons work
- To us, it doesn't matter—we'll treat it as a black box with certain properties

How ML Works

- You feed the program a lot of *training data*
- From this training data, the ML algorithm builds a model of the input
- New inputs are matched against the model
  - Examples: Google Translate, Amazon and Netflix's recommendation engines, speech and image recognition
- However—machine learning algorithms find *correlations*, not *causation*
  - It's not always clear why ML makes certain connections

Correlation versus Causation



https://xkcd.com/552/

Training Data

- Training data must represent the desired actual input space
- Ideally, the training records should be statistically *independent*
- If you get the training data wrong, the output will be biased
- To understand or evaluate the behavior of an ML system, you need the code *and* the data it was trained on
  - "Algorithm transparency" alone won't do it

Biased Data

- Suppose you want an ML system to evaluate job applications
- You train it with data on your current employees
- The ML system will find applicants who “resemble” the current work force
- *If your current workforce is predominantly white males, the ML system will select white male applicants and perpetuate bias*

Learning Styles

Supervised

- A human *labels* the training data according to some criteria, e.g., spam or not spam
- The algorithm then "learns" what characteristics make items more like spam or more like non-spam

Unsupervised

- Finds what items cluster together
- Useful for large datasets, where there is no ground truth, or where labels don't matter
- What counts is *similarity*

Supervised: Image Recognition

- Feed it lots of pictures of different things
- Label each one: a dog, a plane, a mountain, etc.
- Now feed it a new picture—it will find the closest match and output the label

Unsupervised Learning

- Feed in lots of data *without* ground truth
- The algorithms find clusters of similar items; they can also find outliers—items that don’t cluster with others
- They can also find probabilistic dependencies—if a certain pattern of one set of variables is associated with the values of another set, a prediction can be made about new items’ values for those variables

Training is Context-Dependent

- Does “white” cluster with “red”, “green”, “blue”, etc., as a color?
- Does it cluster to “beige”, “ivory”, “ecru”, etc., as a very pale shade?
- Does it cluster with “Black”, “Asian”, etc., as a racial category?

You cannot take a training dataset from one context and use it in another

Training is Culture-Dependent

- Think of the different words used in U.S. versus British English
  - Apartment versus flat
  - Truck versus lorry
  - Shot versus jab
- There can even be completely opposite meanings for some words: consider tabling a bill in Congress versus in the House of Commons

Conclusion: be careful whom you hire to label things

Recommendation Engines

- To recommend things to you, Amazon, Netflix, YouTube, etc., do not need to know what you buy or watch
- Rather, they just need to know that people who liked *X* also tended to like *Y* and *Z*.
- This is a classic example of unsupervised learning

Why Use ML Image Recognition?

ai

Why Use ML Image Recognition?

ai

Why Use ML Image Recognition?

ML Can Spot Features You Miss

Medical Imagery

- ML-based image recognition is being tried out on medical images: X-rays, MRIs, etc.
- In some trials, it’s been as good or even better than radiologists
- And: computers don’t get tired, don’t get bored, and don’t get distracted

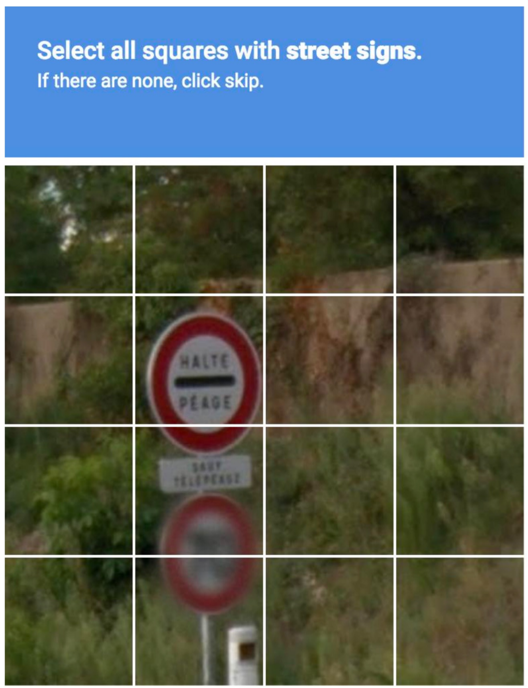
Today’s Uses

- Machine translation
- Speech recognition
- Computer vision
- Some search engine features
- A lot of self-driving car software

In the past, all of these things were attempted by dedicated code, which didn’t work nearly as well

Training Image Recognition

- Ever wonder why so many of the CAPTCHAs are relevant to drivers?
- That's right—you're helping to train an ML algorithm



Merging Data Sources

- There are many sensors, not just visual light cameras
- They differ in resolution, coverage, timing, etc.
- Imagine: satellite photos in different wavelengths, airborne side-looking radar, weather, temperature, etc.
- ML algorithms can treat these as multiple variables and make inferences *without* actually merging them

Computer Security

- Train your ML system on normal, unhacked data
- Have it look, in real-time, for deviations; flag them as possible security incidents
- Known as “anomaly detection”; used for network traffic, host behavior, virus detection, etc.

The Good News

- ML has made incredible progress in the last very few years
- Things that had been very hard research problems are now routine
- There is every reason to expect continued rapid progress

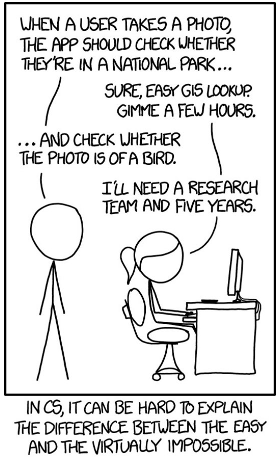
Tech Giants Are Paying Huge Salaries for Scarce A.I. Talent

Nearly all big tech companies have an artificial intelligence project, and they are willing to pay experts millions of dollars to help get it done.

17h ago • By CADE METZ

Things Change Rapidly!

September 2014



https://xkcd.com/1425/

Google Images, Today

Image size:
367 × 245

No other sizes of this image found.

Best guess for this image: ***northern cardinal***

The Full Picture

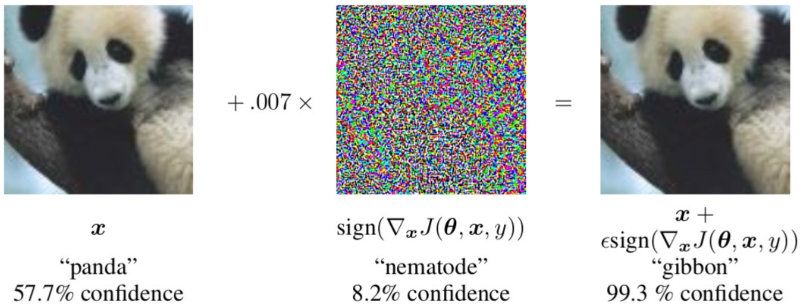
However…

There Are Issues

- Training Data Is Hard
- Output is Probabilistic
- Adversarial machine learning
- There are important “big data” situations where ML cannot help

Adversarial Machine Learning

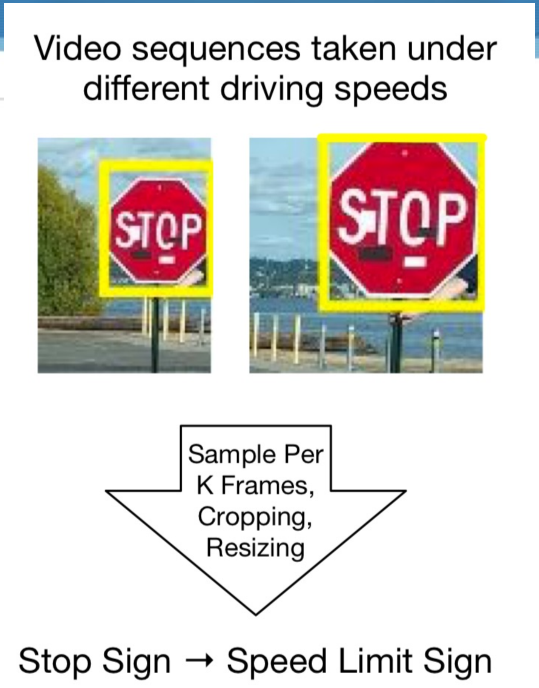
- Computers do not “see” the way we do
- Imperceptible or irrelevant—to us!—changes to an image can drastically change the results



https://arxiv.org/abs/1412.6572

100% Successful Attack

https://arxiv.org/pdf/1707.08945.pdf



Biased Sources

- Training data that doesn't represent actual data
- Cultural biases by the trainers
  - Mechanical Turk workers are often used for labeling
- False positives and false negatives

Finding Terrorists

- There are very, very few terrorists
- Where are you going to find enough training data?
- Almost certainly, any features the real terrorists have in common will be matched by very many other innocent people
- The algorithms can't distinguish them

Finding Terrorists

- There are very, very few terrorists
- Where are you going to find enough training data?
- Almost certainly, any features the real terrorists have in common will be matched by very many other innocent people
- The algorithms can’t distinguish them
- When humans do this, we call it profiling

ML Doesn’t Always Work the Way We Want it To…

Some Examples

- Microsoft Tay
- Recidivism risk
- Targeted advertising
- More…

Microsoft Tay

- A Twitter “chatbot”
- Tay “talked” with people on Twitter
- What people tweeted to it became its training data
- It started sounding like a misogynist Nazi…

What Happened?

- People from 4Chan and 8Chan decided to troll it.
- With ML, vile Nazi garbage in, vile Nazi garbage out
- Microsoft didn't appreciate just what people would try.
- "Sinders is critical of Microsoft and Tay, writing that 'designers and engineers have to start thinking about codes of conduct and how accidentally abusive an AI can be.'" (*Ars Technica*)

Recidivism

- Several companies market “risk assessment tools” to law enforcement and the judiciary
- Do they work? Do they exhibit impermissible bias?
- A ProPublica study says that one popular one doesn’t work and does show racial bias: Blacks are more likely to be seen as likely reoffenders—but the predictions aren’t very accurate anyway

What Happened?

- Inadequate evaluation of accuracy
- Using the program in ways not intended by the developers
- Proxy variables for race
- Using inappropriate variables, e.g., “arrests” rather than “crimes committed”

Hypertargeted Advertising

- It's normal practice to target ads to the "right" audience
- ML permits very precise targeting—others can't even see the ads
- Used politically—some say that YouTube's recommendation algorithms helped Trump
  - The Trump campaign used precise targeting on Facebook
- Target managed to identify a pregnant 16-year-old—her family didn't even know

Target

- People habitually buy from the same stores
- They tend to switch only at certain times, e.g., when a baby is born
- Target analyzed sales data to find leading indicators of pregnancy
- They then sent coupons to women who showed those indicators
- People found that creepy—so Target buried the coupons among other, untargeted stuff that they didn't really care if you bought

Algorithmic Transparency

- There have been calls for "algorithmic transparency"—make companies disclose their algorithms
- It can help a little, in that it will show what variables are being used
- But—it's not just the code, it's the data

Data Transparency?

- The training data is often far more sensitive than the code
  - It can be a matter of user privacy
- In some systems, the model is continuously updated
- Example: when you click on a link on a Google results page, the link actually takes you to Google, so it can tell what you clicked on
- But what can we do?

The New York City Initiative

- A task force will develop:
  - Appeal procedures for people affected by city ML systems
  - Methods to look for bias
  - A procedure to provide redress for people discriminated against
  - Recommendations for transparency of operation
  - Procedures to archive code *and* training data

Will it Work?

- It can, up to a point. However…
- Explaining why an ML algorithm gave a particular answer is *hard*
- Code is often vendor-proprietary
- Training data is often sensitive
- But at least the city recognizes the issue

What Should We Do?

- Awareness is key
- Get competent data scientists to study each system, to look at data sources, proxies, code, etc.
    - *Require* vendors to make code available to city-designated experts
- Above all: social policy has to come first, and be set by political processes; code has to follow that policy

Questions?

(Osprey with fish, Central Park, September 29, 2021)