Interpretable and Responsible Machine Learning

Connect

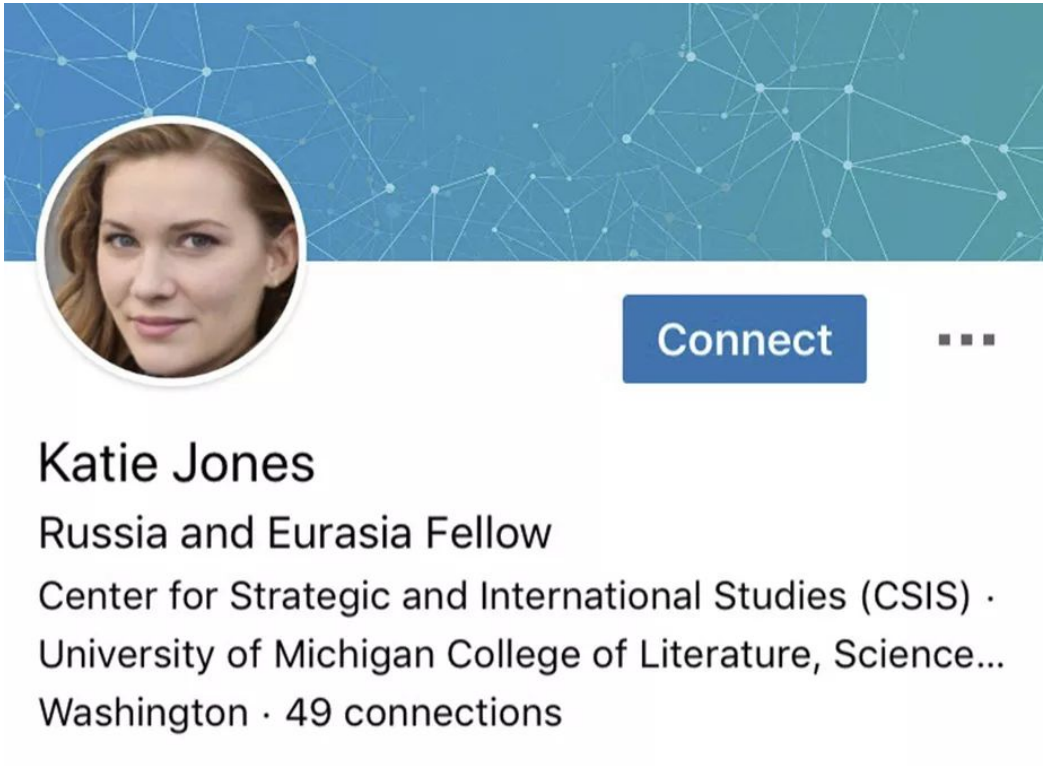
Katie Jones

Russia and Eurasia Fellow

Center for Strategic and International Studies (CSIS) ·

University of Michigan College of Literature, Science...

Washington · 49 connections



The use of an AI method known as a generative adversarial network (or GAN) to create the account's fake profile picture

Agenda

1. History of AI

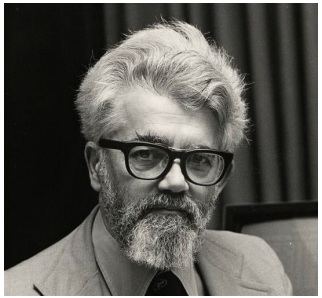
2. Is ML Dangerous?

3. Accountable Algorithms

History of AI

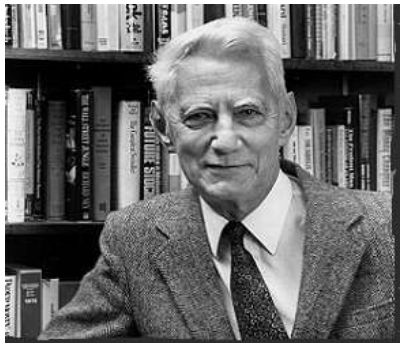
Birth of AI

- 1956: Workshop at Dartmouth College:



John McCarthy

Marvin Minsky

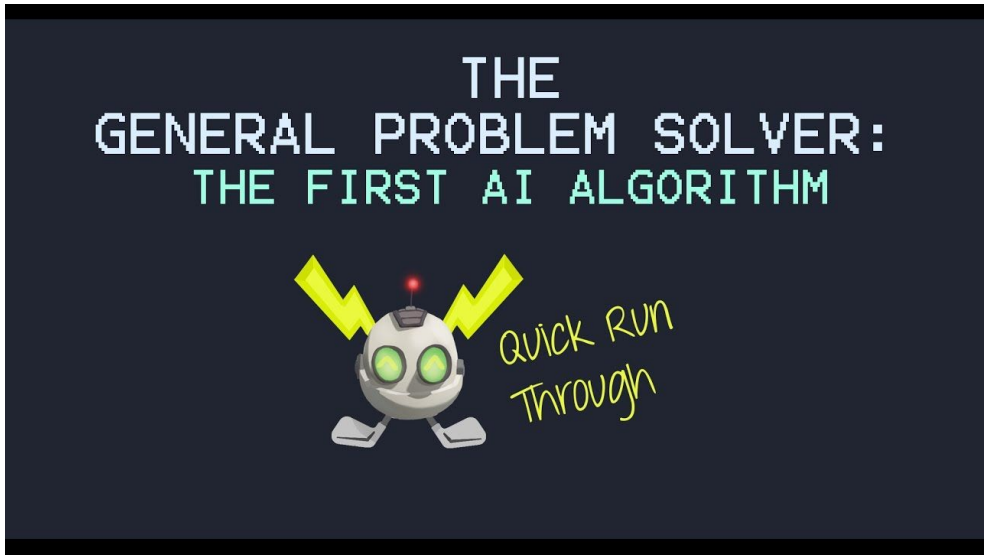


Claude Shannon

- **Targets**:
  - *Every aspect of learning or any other feature of intelligence can be so precisely described that a machine can be made to simulate it.*

Early Successes

- Newell & Simon's Logic Theorist: prove theorems in Principia Mathematica using search + heuristics; later General Problem Solver (GPS)



https://en.wikipedia.org/wiki/General_Problem_Solver

Overwhelming Optimism

- 1958, **H.A.Simon** and **Allen Newell**: “within ten years a digital computer will be the world’s chess champion” and “within ten years a digital computer will discover and prove an important new mathematical theorem”.
- 1965, **H.A.Simon**: “machines will be capable, within twenty years, of doing any work a man can do”
- 1967, **Marvin Minsky**: “Within a generation...the problem of creating ‘artificial intelligence” will substantially be solved”
- 1970, **Marvin Minsky**: “In from three to eight years we will have a machine with the general intelligence of an average human being”.

underwhelming results

Example: machine translation

The spirit is willing but the flesh is weak.

↓

(Russian)

↓

The vodka is good but the meat is rotten.

1966: ALPAC report cut off government funding for MT

From cs221

AI is overhyped...

- *We tend to overestimate the effect of a technology in a short run and underestimate the effect in a long run*. - Roy Amara (1925-2007)

Implications of Early Era

- **Problems**:
  - **Limited computation**: search space grew exponentially, outpacing hardware
  - **Limited information**: complexity of AI problems (number of words, objects, concepts in the world)

- **Contributions**:
  - Lisp, garbage collection, time-sharing (John MacCarthy)
  - **Key paradigm**: separate ***modeling*** (declarative) and ***inference*** (procedural)

Knowledge-based Systems (70-80s)

- Expert Systems: elicit specific domain knowledge from experts in form of rules:
    - If [premises] then [action]

Category	Problem addressed	Examples
Interpretation	Inferring situation descriptions from sensor data	Hearsay (speech recognition), PROSPECTOR
Prediction	Inferring likely consequences of given situations	Preterm Birth Risk Assessment[56]
Diagnosis	Inferring system malfunctions from observables	CADUCEUS, MYCIN, PUFF, Mistral,[57] Eydenet,[58] Kaleidos[59]
Design	Configuring objects under constraints	Dendral, Mortgage Loan Advisor, R1 (DEC VAX Configuration), SID (DEC VAX 9000 CPU)
Planning	Designing actions	Mission Planning for Autonomous Underwater Vehicle[60]
Monitoring	Comparing observations to plan vulnerabilities	REACTOR[61]
Debugging	Providing incremental solutions for complex problems	SAINT, MATHLAB, MACSYMA
Repair	Executing a plan to administer a prescribed remedy	Toxic Spill Crisis Management
Instruction	Diagnosing, assessing, and repairing student behavior	SMH.PAL,[62] Intelligent Clinical Training,[63] STEAMER[64]
Control	Interpreting, predicting, repairing, and monitoring system behaviors	Real Time Process Control,[65] Space Shuttle Mission Control[66]

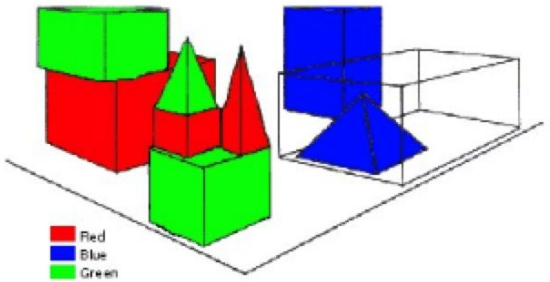
Knowledge-based Systems

- Contributions:
  - First real application that impacted industry
  - Knowledge helped curb the exponential growth

- Problems:
  - Knowledge is not deterministic rules, need to model **uncertainty**
  - Requires considerable **human efforts** to create rules, hard to maintain.

SHRDLU [Winograd 1971]

Person: Pick up a big red block.
Computer: OK.
Person: Grasp the pyramid.
Computer: I don't understand which pyramid you mean.
Person (changing their mind): Find a block which is taller than the one you are holding and put it into the box.
Computer: By "it", I assume you mean the block which is taller than the one I am holding.
Computer: OK.
Person: What does the box contain?
Computer: The blue pyramid and the blue block.
Person: What is the pyramid supported by?
Computer: The box.



The Complexity Barrier

A number of people have suggested to me that large programs like the SHRDLU program for understanding natural language represent a kind of dead end in AI programming. Complex interactions between its components give the program much of its power, but at the same time they present a formidable obstacle to understanding and extending it. In order to grasp any part, it is necessary to understand how it fits with other parts, presents a dense mass, with no easy footholds. Even having written the program, I find it near the limit of what I can keep in mind at once.

— Terry Winograd (1972)

Modern AI (90s-present)

- **Probability**:Pearl (1988) promote Bayesian networks in AI to **model uncertainty** (based on Bayes rule from 1700)

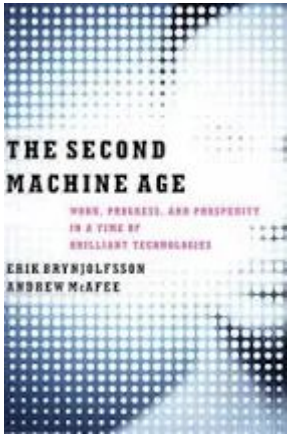
From **MODEL** to PREDICTIONS

- **Machine Learning**: Vapnik (1955) invented support vector machines to **learn parameters** (based on statistical models in early 1900s)

From DATA to **MODEL**

The Second Machine Age

- **AI is being used to make decisions for:**
  - **Credit**
  - **Education**
  - **Employment**
  - **Advertising**
  - **Healthcare**
  - **Policing**
  - **Urban Computing**
  - **……**



Is Machine Learning Dangerous?

Elon Musk: Humanity Is a Kind of 'Biological Boot Loader' for AI

AI is outpacing our ability to understand it, the Tesla CEO says. It will open a new chapter for society, replies the Alibaba cofounder.

Jack Ma, left, debates AI—and the future of humanity—with Elon Musk ALY SONG/REUTERS

Venkat Viswanathan
@venkvis

Follow

.@TeslaMotors Model S autopilot camera misreads 101 sign as 105 speed limit at 87/101 junction San Jose. Reproduced every day this week.

8:40 PM - 14 Jul 2017

239 Retweets **427** Likes

WOMAN SAYS AMAZON'S ALEXA TOLD HER TO STAB HERSELF IN THE HEART FOR 'THE GREATER GOOD'

BY JAMES CROWLEY ON 12/24/19 AT 12:04 PM EST

SHARE

Google has issued an apology after computer programmer Jacky Alcine, from New York, spotted photographs of him and a female friend had been labelled as gorillas by Google Photos image recognition software. He sent a series of Tweets to Google highlighting the problem (like above) leading Google to issue a fix for the problem

Facebook has apologised after an error in its machine-translation service saw Israeli police arrest a Palestinian man for posting "good morning" on his social media profile.

The man, a construction worker in the West Bank settlement of Beitar Illit, near Jerusalem, posted a picture of himself leaning against a bulldozer with the caption "يصبحهم", or "yusbihuhum", which translates as "good morning".

But Facebook's artificial intelligence-powered translation service, which it built after parting ways with Microsoft's Bing translation in 2016, instead translated the word into "hurt them" in English or "attack them" in Hebrew.

Police officers arrested the man later that day, according to Israeli newspaper Haaretz, after they were notified of the post. They questioned him for several hours, suspicious he was planning to use the pictured bulldozer in a vehicle attack, before realising their mistake. At no point before his arrest did any Arabic-speaking officer read the actual post.

Is Machine Learning Dangerous?

- Will human be ruled by machines?
  - It seems no likely any time time.
  - General AI is so challenging
  - Algorithms are not “intelligent” enough

- But machine learning can potentially be **misused**, **misleading**, and/or **invasive**
  - Important to think about implications of what you build

App Store Preview

This app is available only on the App Store for iPhone and iPad.

Mushroom Identificator 4+

Mushrooms photo recognition

AnnapurnApp Technologies UG haftungsbeschrankt

★★★★★ 4.6, 387 Ratings

Free · Offers In-App Purchases

Screenshots iPhone iPad

TECHNOLOGY NEWS OCTOBER 10, 2018 / 11:12 AM / A YEAR AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

> In effect, Amazon's system taught itself that male candidates were preferable. It penalized resumes that included the word "women's," as in "women's chess club captain." And it downgraded graduates of two all-women's colleges, according to people familiar with the matter. They did not specify the names of the schools.
>
> Amazon edited the programs to make them neutral to these particular terms. But that was no guarantee that the machines would not devise other ways of sorting candidates that could prove discriminatory, the people said.

Accountable Algorithms

Treatment Recommendation

Demographics: **age, gender, ..**
Medical History: **Has asthma?**

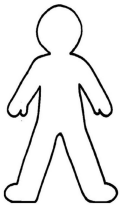
Symptoms: **Severe Cough, Sleepy**

Test Results: **Peak flow: Positive**

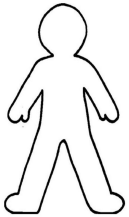
Which treatment should be given?
Options: quick relief drugs (mild),
controller drugs (strong)

Bail Decision

Release



Retain



High-Stakes Decisions

- The above examples all belong to high-stakes decisions. The decisions have a **huge impact on human well-being**.

- What are those non high-stakes decisions?
  - Recommendations in E-commerces websites
  - When should I get up tomorrow?
  - …….

FAT Machine Learning

- Statement from **Fairness**, **Accountability**, and **Transparency** in Machine Learning organization
  - https://www.fatml.org/resources/principles-for-accountable-algorithms

Algorithms and the data that drive them are designed and created by people -- There is always a human ultimately responsible for decisions made or informed by an algorithm. "The algorithm did it" is not an acceptable excuse if algorithmic systems make mistakes or have undesired consequences, including from machine-learning processes.

Five Factors

- Responsibility
  - Make available externally visible avenues of redress of adverse individual or societal effects of an algorithmic decision system, and designate an internal role for person who is responsible for the timely remedy of such issues.
- Explainability
  - Ensure the algorithmic decisions as well as any data driving those decisions can be explained to end-users and other stakeholders in non-technical terms.
- Accuracy
  - Identify, log, and articulate sources of error and uncertainty throughout the algorithm and its data sources so that expected and worst case implications can be understood and inform mitigation procedures.
- Auditability
  - Enable interested third parties to probe, understand, and review the behavior of the algorithm through disclosure of information that enables monitoring, checking, or criticism, including through provision of detailed documentation, technically suitable APIs, and permissive terms of use.
- Fairness
  - Ensure that algorithmic decisions do not create discriminatory or unjust impacts when comparing across different demographics (e.g., race, sex, etc).

Fairness

Black people with complex medical needs were less likely than equally ill white people to be referred to programmes that provide more personalized care. Credit: Ed Kashi/VII/Redux/eyevine

An algorithm widely used in US hospitals to allocate health care to patients has been systematically discriminating against black people, a sweeping analysis has found.

Why unfair?

- How does this type of error happen?

- Possibilities:
  - Not enough diversity in training data
  - Not enough diversity in test data
  - Not enough error analysis

Fairness

- Suppose your classifier gets 90% accuracy...

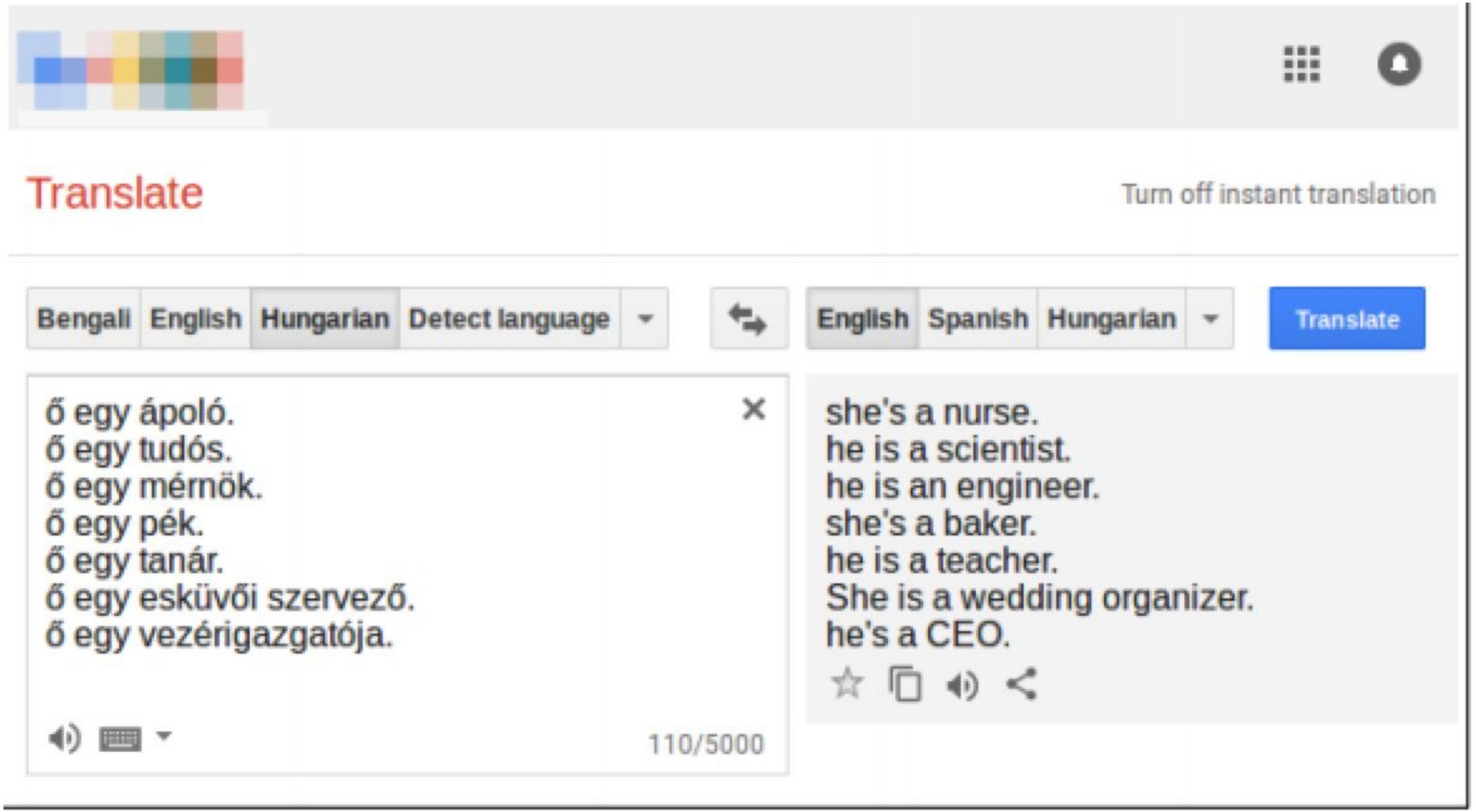
Scenario 1:

Scenario 2:

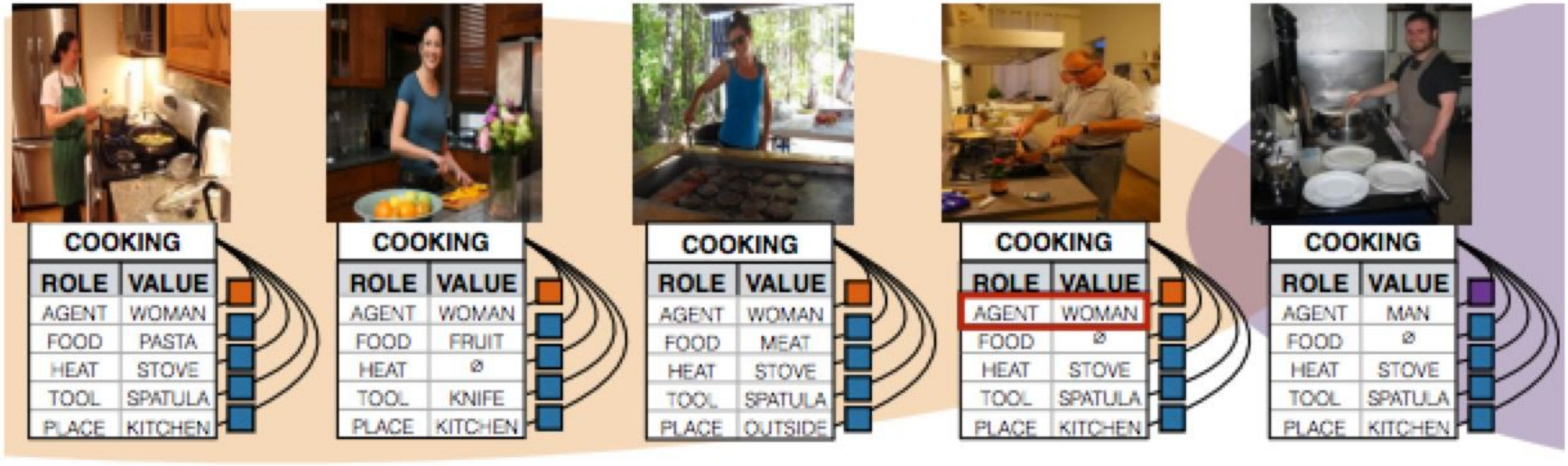
Bias

- Bias and stereotypes that exist in data will be learned by ML algorithms

- Sometime, those biases will be amplified by ML



Prates et al, 2018,https://arxiv.org/pdf/1809.02208.pdf



- Training data:
  - Women appeared in “cooking” images 33% more often than men
- Predictions:
  - Women appeared **68%** more often

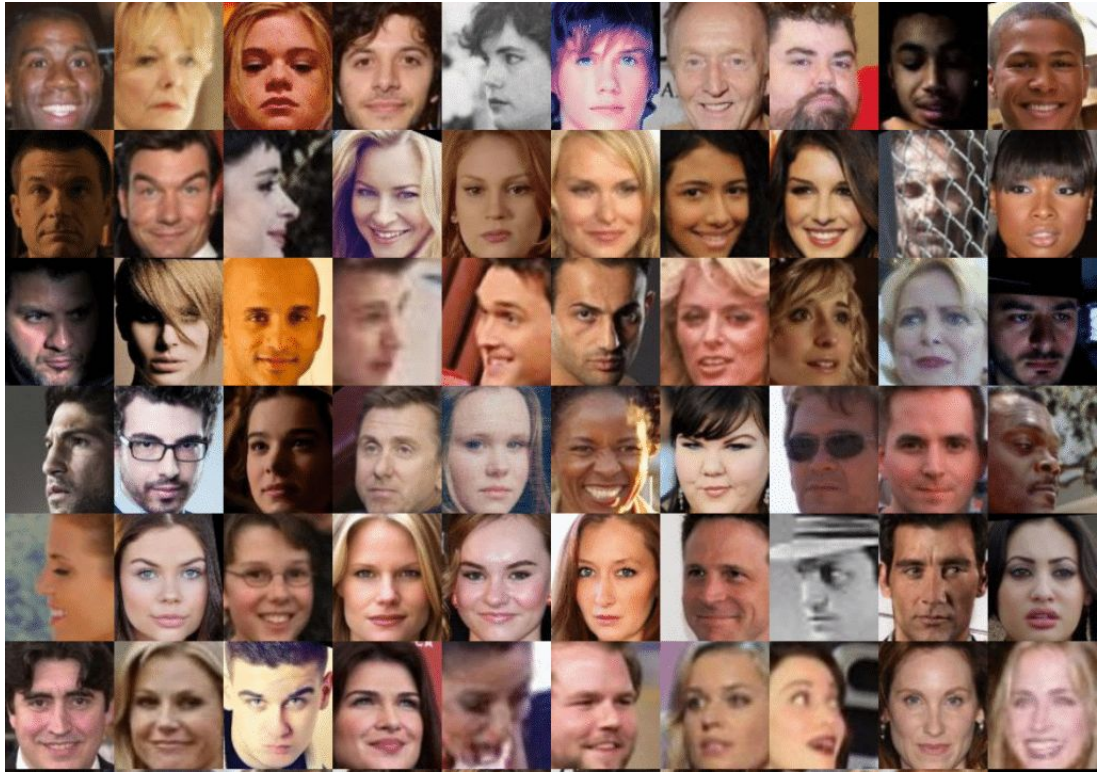
Zhao et al, 2017, https://arxiv.org/pdf/1707.09457.pdf

Privacy

- Training data is often scraped from the web

- Personal data may get scooped up by ML systems
    - Are users aware of this?
    - How do they feel about it?

- No reveal sensitive information (income, health, communication)



MegaFace Dataset:
4.7 million photos of
627,000 individuals,
from Flickr users

Use and Misuse

- Machine learning can predict:
  - If you are overweight
  - If you are transgender
  - If you have died

- People may build these classifiers for legitimate purposes, but could easily be misused by others

Criminal Machine Learning

- Can we predict if someone is prone to committing a crime based on their facial structure?
- One of studies: Wu and Zhang (2016), “Automated Inference on Criminality using Face Images”, claims yes, with 90% accuracy.
- Good summary of why the answer is probably no:
    - https://callingbullshit.org/case_studies/case_study_criminal_machine_learning.html

(a) Three samples in criminal ID photo set S_c.

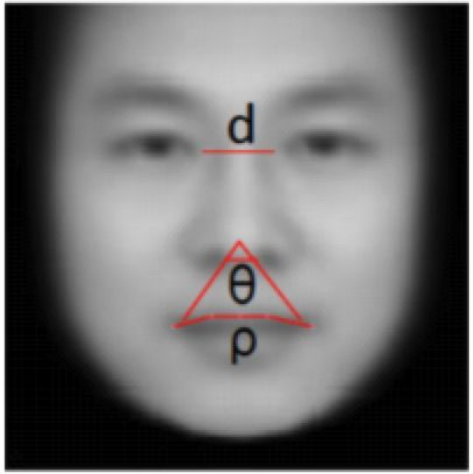
(b) Three samples in non-criminal ID photo set S_n

Figure 2. Criminal and non-criminal faces from Wu and Zhang (2016)

Use and Misuse

- How was the dataset created?
  - Criminal photos: government IDs
  - Non-criminal photos: professional headshots

- What did the classifier learn?
  - “The algorithm finds that criminals have shorted distances between the inner corners of the eyes, smaller angles between the nose and the corners of the mouth, and higher curvature of the upper lip.”



Case Study

- If your tool seems dystopian:
  - Consider whether this is really something you should be building…
    - One argument: someone will eventually build this technology, so better for researchers to do it first to understand it.
    - Still, proceed carefully: understand potential misuse
  - Be sure that your claims are correct
    - Solid error analysis is critical
    - Misuse of an inaccurate system even worse than misues of an accurate system.