Responsible Machine Learning

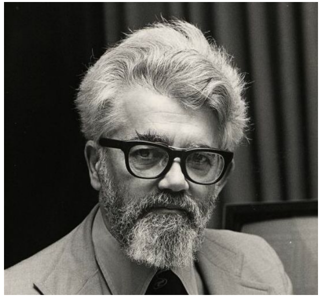
Agenda

1. History of AI
2. Is ML Dangerous?
3. Accountable Algorithms
4. Course Summary

History of AI

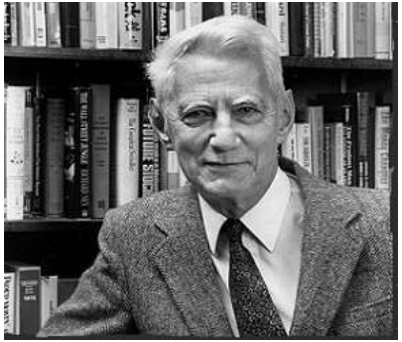
Birth of AI

- 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.*

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.

Modern AI (90s-present)

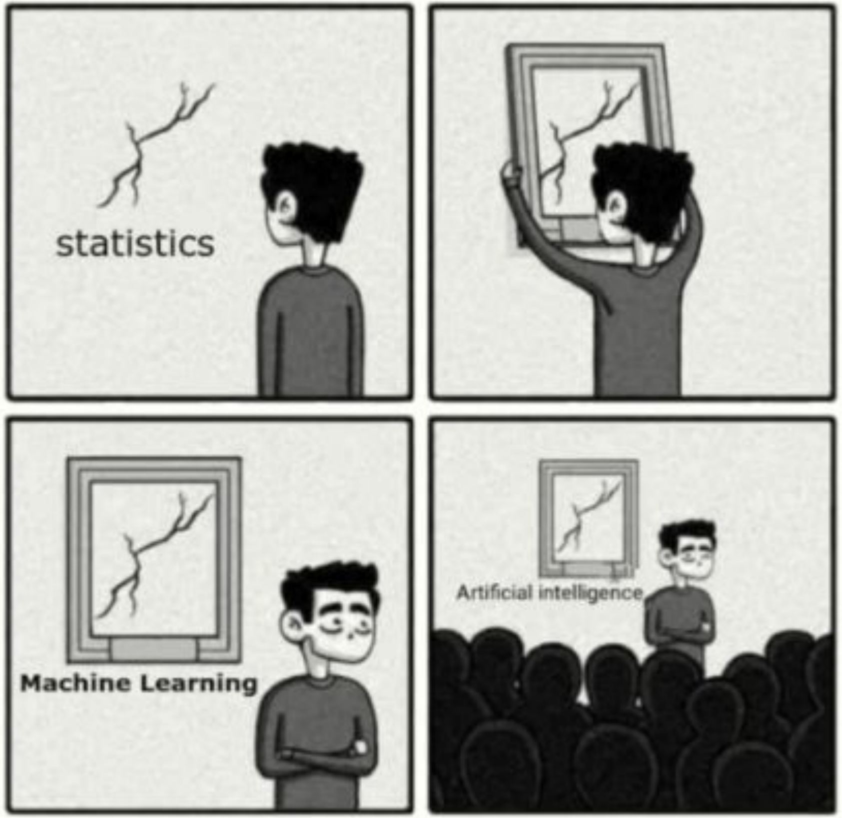
- **Stat Model**:Pearl (1988) promote Bayesian networks in AI to **model uncertainty** (based on Bayes rule from 1700)

Stat Model: infer the relationship among variable in data

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

Machine Learning: sacrifice interpretability for predictive power

https://www.nature.com/articles/nmeth.4642


statistics
Machine Learning
Artificial intelligence

Take Linear Regression as the example

Stat Model:

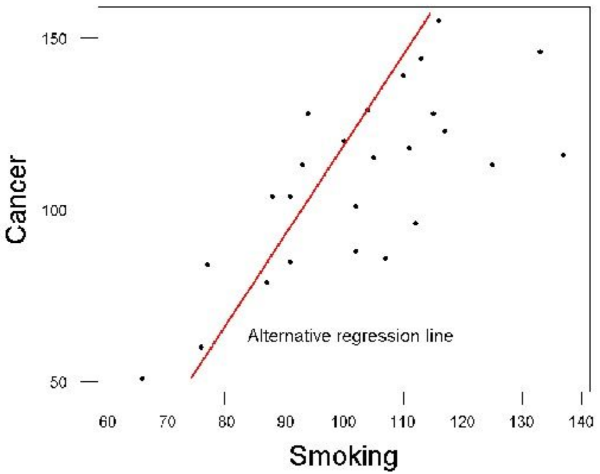
1.**Inference**: Characterize the relationship between the smoking index and cancer rates.

2. Conduct the significance test of the model parameters

ML:

1.**Prediction**:
Get a model that is able to make prediction of the cancer rates based on smoking index

2. Evaluate the model performance over testing data.



The Second Machine Age

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

Is Machine Learning Dangerous?

Is Machine Learning Dangerous?

- Will human be ruled by machines?
    - It seems unlikely any 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

Nicolas Kayser-Bril @nicolaskb · Mar 31

Black person with hand-held thermometer = firearm.
Asian person with hand-held thermometer = electronic device.

Computer vision is so utterly broken it should probably be started over from scratch.

Faces | Objects | **Labels** | Web | Properties | Safe Search

Screenshot from 2020-03-31 11-23-45.png

Gun	88%
Photography	68%
Firearm	65%
Plant	59%

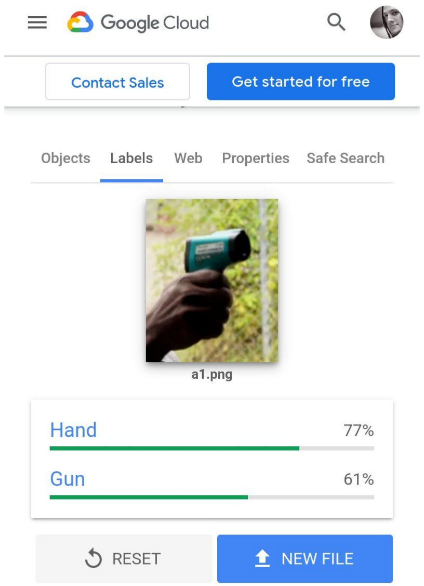
Screenshot from 2020-03-31 11-27-22.png

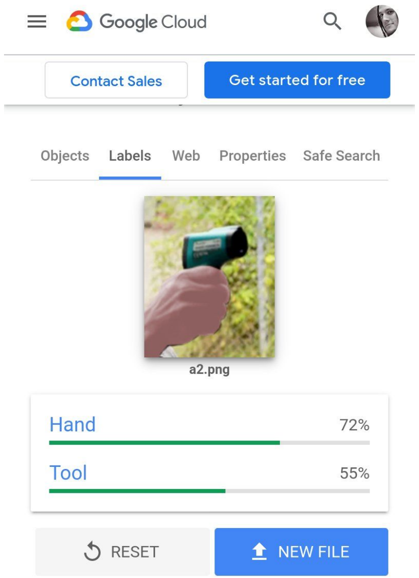
Technology	68%
Electronic Device	66%
Photography	62%
Mobile Phone	54%

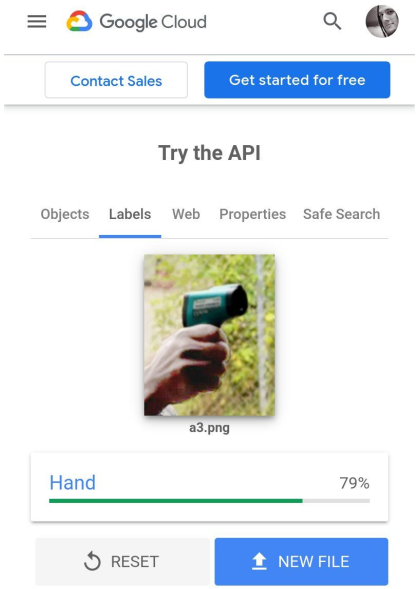
Bart Nagel
@bjnagel

Replying to @nicolaskb

I cropped the first photo to just the hand and device, and did some very inexpert colour tweaks in an attempt to make the skin white, and somewhere in between. The results are troubling.







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

Optimization Targets

Is the objective function of ML algorithms also good for human well-being?

Accountable Algorithms

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.

Fairness

- Suppose your classifier gets 90% accuracy...

Why unfair?

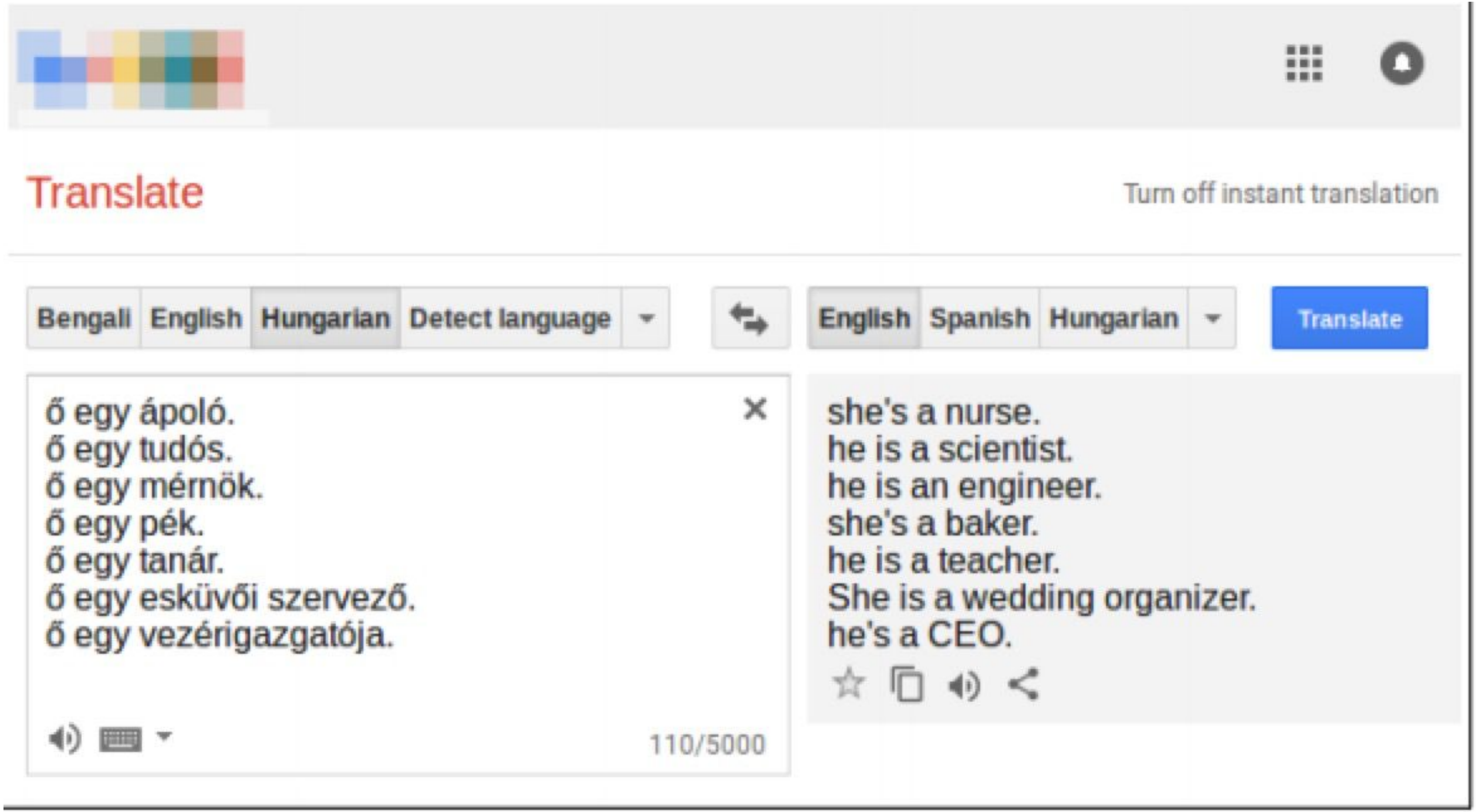
- How does this type of error happen?
    - Most ml models' objectives will sacrifice the accuracy of the minority groups to make accurate predictions for majority class.

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

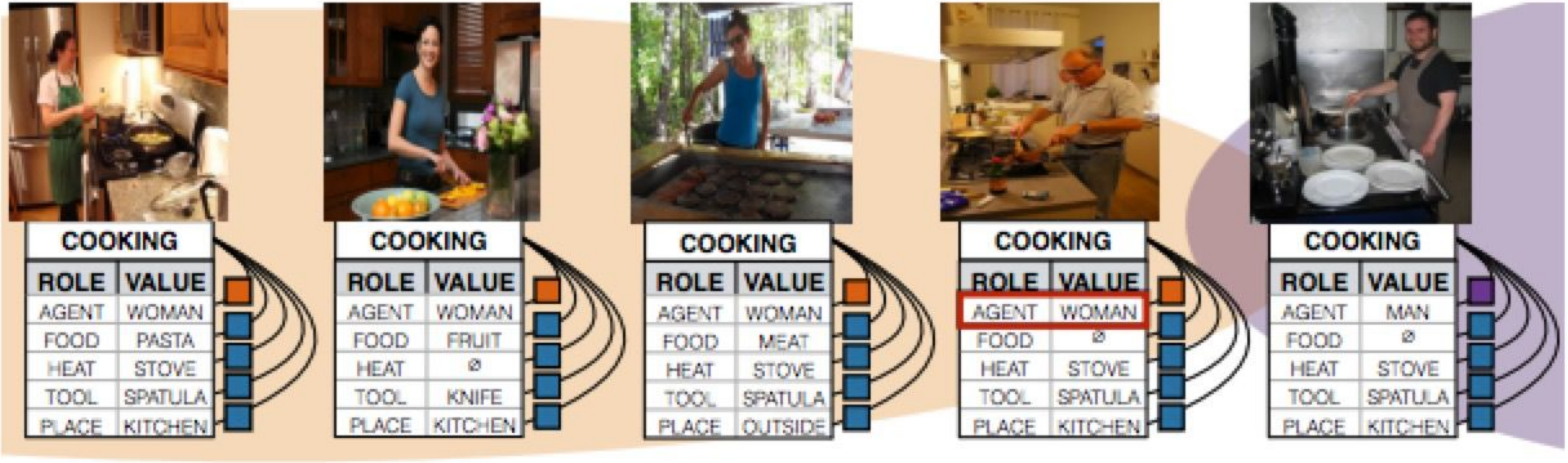
Bias

- 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

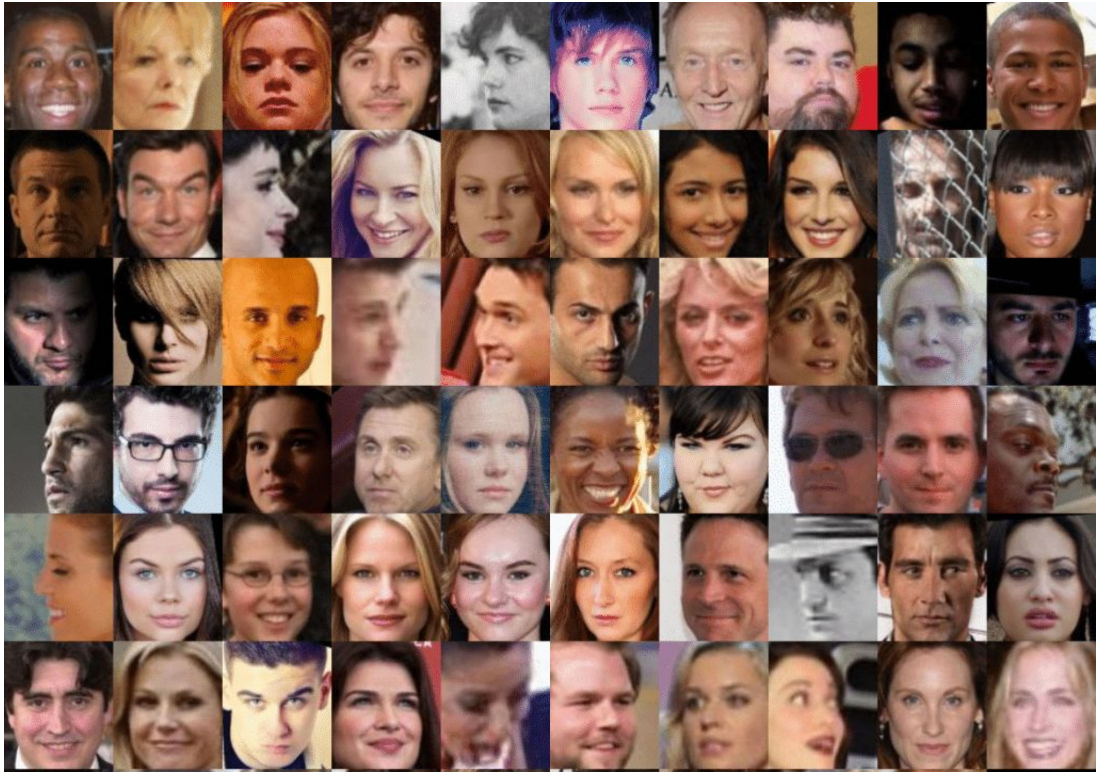
Zhao et al, 2017,https://arxiv.org/pdf/1707.09457.pdf

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.

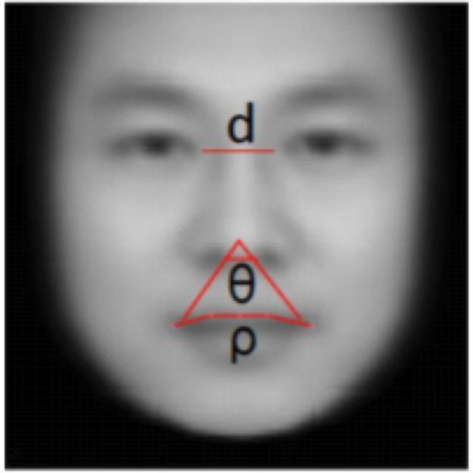
(b) Three samples in non-criminal ID photo set S_n

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.”



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.

General Principle

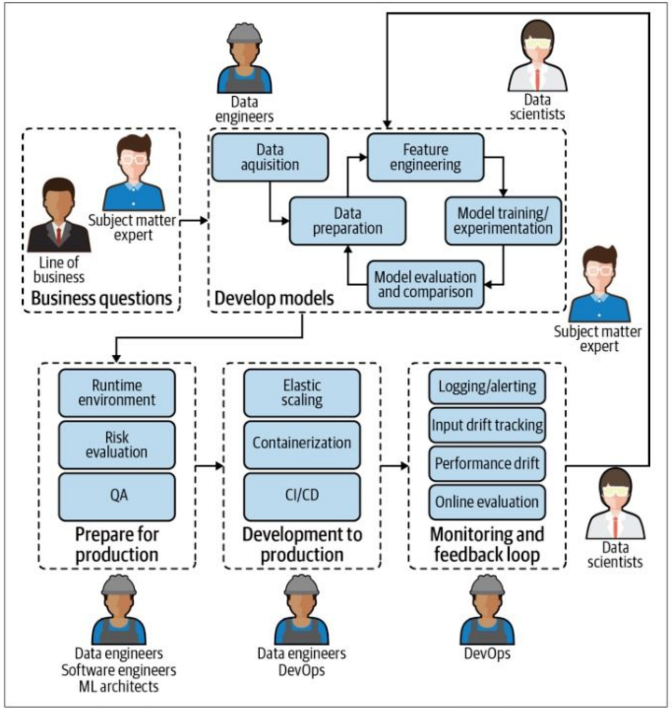
- If your tool seems dystopian:
    - Consider whether this is really something you should be building…
        - One argument: someone will eventually build this technology, so better for researchers to do it first to understand it.
        - Still, proceed carefully: understand potential misuse
  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

Course Summary

- Three Main Topics:
    - Machine Learning Pipeline → MLOps
    - Probabilistic and Bayesian Models (only one week, but it is really important) → Causal Inference
    - Deep Learning
- How do we understand the concepts of machine learning models better:
    - Build your own knowledge graph that can explain the connections among these models
    - Check its corresponding application

MLOps

- MLOps is the standardization and streamlining of machine learning life cycle management.



Source: Introducing MLOps

https://www.oreilly.com/library/view/introducing-mlops/9781492083283/

Causal Inference

- Causal Inference: Learn model of how the world works
  - Impact of interventions can be context-specific,
  - Model maps contexts and interventions to outcomes,
  - Formal language to separate out correlates and causes.

https://cdn1.sph.harvard.edu/wp-content/uploads/sites/1268/2021/03/ciwhatif_hernanrobins_30mar21.pdf