Fairness, Accountability and Transparency (FAT) in Machine Learning

11.S955 and 11.S198 Fall 2019
Deep Learning for Urban Mobility

Peyman Noursalehi, Shenhao Wang and Jinhua Zhao
Massachusetts Institute of Technology

Deep Learning for Urban Mobility

Part I: Introduction and DNN Basics

- Sep 10 Introduction: Deep Learning Meets Transportation

Part II: Passenger and Traffic Flow Prediction

- Sep 17 Demand prediction and Deep learning basics
- Sep 24 Advanced modeling techniques: ConvLSTM, Attention, individualized predictions
- Oct 1 Guest Lecture Prof. Justin Dauwels
- Oct 8 Advanced Applications: Generative models (VAE), graph embeddings,

Part III: DNN and Demand Analysis

- Oct 22 DNN and Discrete Choice 1
- Oct 29 DNN and Discrete Choice 2
- Nov 5 DNN and Prospect Theory

Part IV: Reinforcement Learning and Control in Transportation

- Nov 12 RL and Control Part 1
- Nov 19 RL and Control Part 2
- Nov 26 Guest Lecture Prof. Cathy Wu

Part V: FAT and Summary

- Dec 3 Fairness, Accountability and Transparency
- Dec 10 Student Presentation

FAT in Machine Learning

Problem
Technical solutions
Process solutions
Key persons/books/articles
Resources
Next class

Word Embedding

- Paris - France = London - England
- Man - woman = King - Queen
- Man - woman = programmer - homemaker
- Man - woman = surgeon - nurse

Bolukbasi, T., Chang, K. W., Zou, J., Saligrama, V., & Kalai, A. (2016). Quantifying and reducing stereotypes in word embeddings. arXiv preprint arXiv:1606.06121.

Google Translator

Machine learning codifies human biases

Technical Solution

- Transformation matrix T
  - The transformed embeddings are stereotypical-free
  - The transformed embeddings preserve the distances between any two vectors in the matrix A
- B: v(he)-v(she): direction of stereotype
- P: set of word to debias: Man - woman = surgeon - nurse

$$PTT^T B^T \approx 0$$

Bolukbasi, T., Chang, K. W., Zou, J., Saligrama, V., & Kalai, A. (2016). Quantifying and reducing stereotypes in word embeddings. arXiv preprint arXiv:1606.06121.

Technical Solution

$$\min_{X \succeq 0} \|AXA^T - AA^T\|_F^2 + \lambda \|PXB^T\|_F^2$$

$$X = TT^T$$

Bolukbasi, T., Chang, K. W., Zou, J., Saligrama, V., & Kalai, A. (2016). Quantifying and reducing stereotypes in word embeddings. arXiv preprint arXiv:1606.06121.

Bias in the Vision and Language of Artificial Intelligence

Margaret Mitchell, Senior Research Scientist Google AI

What do you see?

What do you see?

- Bananas

What do you see?

- Bananas
- Stickers

What do you see?

- Bananas
- Stickers
- Dole Bananas

What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store

What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store
- Bananas on shelves

What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store
- Bananas on shelves
- Bunches of bananas

What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store
- Bananas on shelves
- Bunches of bananas
- Bananas with stickers on them

What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store
- Bananas on shelves
- Bunches of bananas
- Bananas with stickers on them
- Bunches of bananas with stickers on them on shelves in a store

What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store
- Bananas on shelves
- Bunches of bananas
- Bananas with stickers on the
- Bunches of bananas with stic
  them on shelves in a store

...We don’t tend to say

Yellow Bananas

What do you see?

Yellow is prototypical for bananas

World learning from text

Gordon and Van Durme, 2013

Word	Frequency in corpus
“spoke”	11,577,917
“laughed”	3,904,519
“murdered”	2,834,529
“inhaled”	984,613
“breathed”	725,034
“hugged”	610,040
“blinked”	390,692
“exhale”	168,985

World learning from text

Gordon and Van Durme, 2013

Word	Frequency in corpus
“spoke”	11,577,917
“laughed”	3,904,519
“murdered”	2,834,529
“inhaled”	984,613
“breathed”	725,034
“hugged”	610,040
“blinked”	390,692
“exhale”	168,985

Human Reporting Bias

The **frequency** with which **people write** about actions, outcomes, or properties is **not a reflection of real-world frequencies** or the degree to which a property is characteristic of a class of individuals

Prototype Theory

One purpose of categorization is to **reduce the infinite differences** among stimuli **to** behaviourally and **cognitively usable proportions**

A man and his son are in a terrible accident and are rushed to the hospital in critical care.

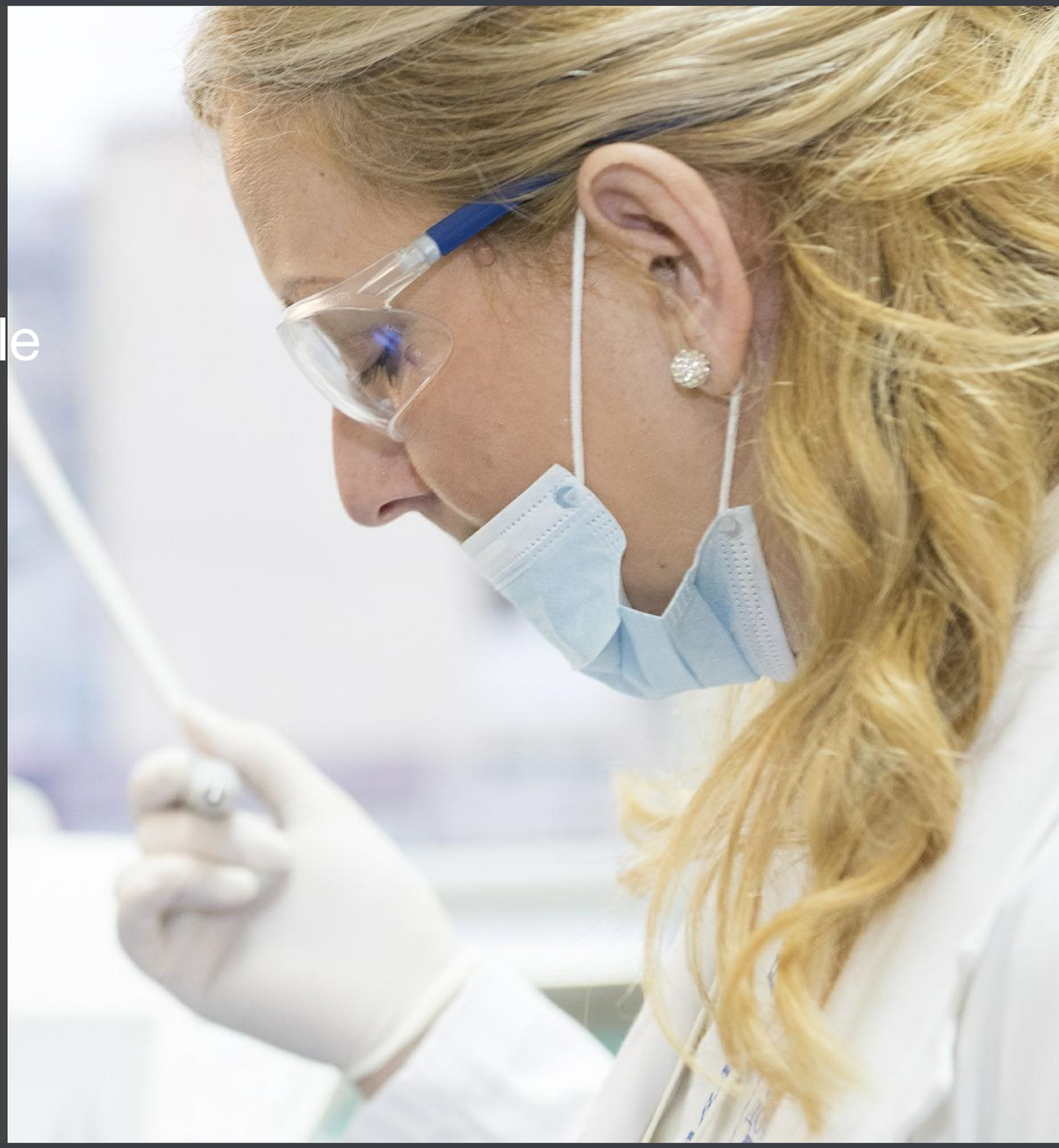
The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!"

How could this be?

A man and his son are in a terrible accident and are rushed to the hospital in critical care.

The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!"

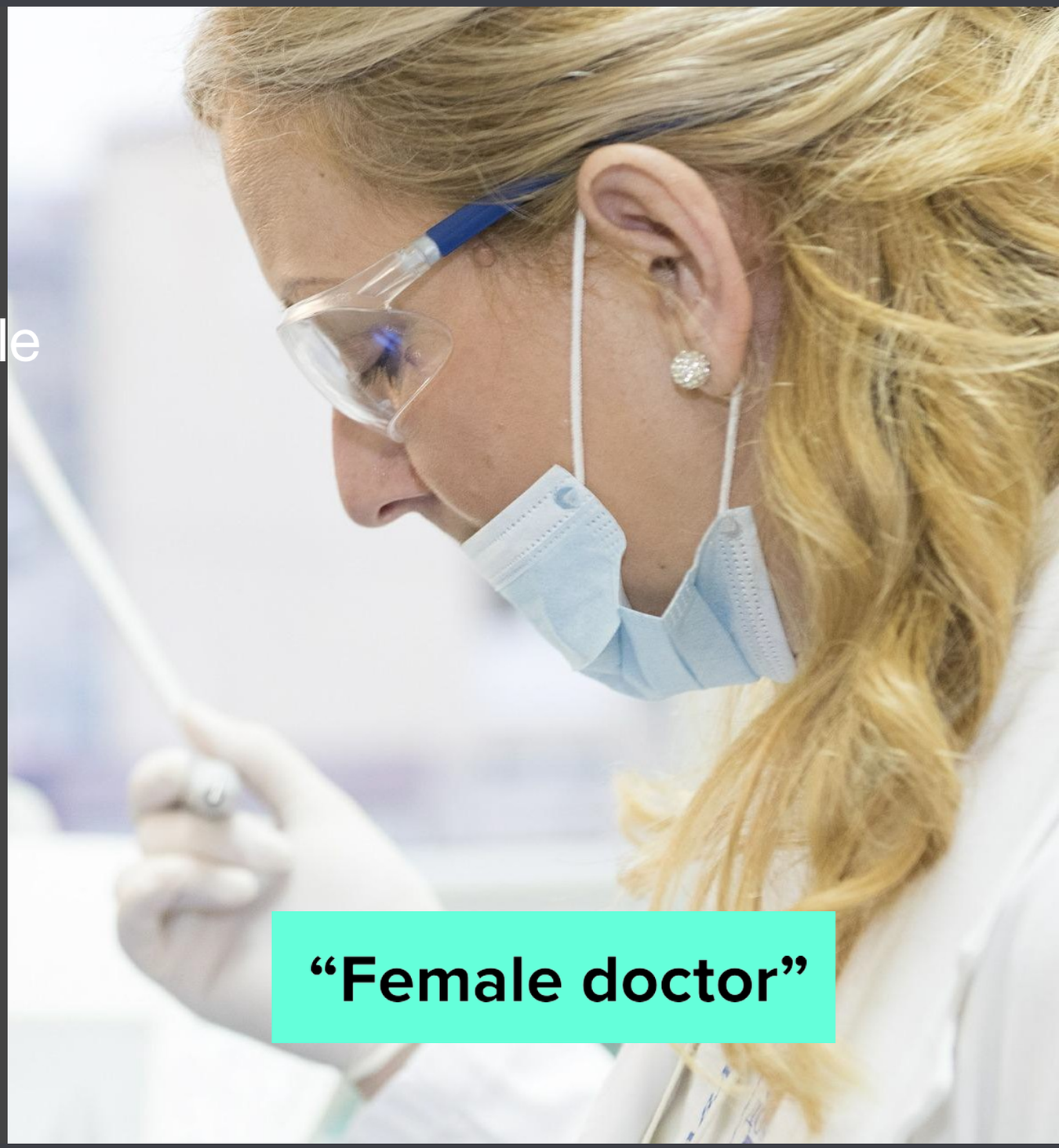
How could this be?



A man and his son are in a terrible accident and are rushed to the hospital in critical care.

The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!"

How could this be?




“Doctor”


“Female doctor”

The majority of test subjects overlooked the possibility that the doctor is a she - including men, women, and self-described feminists.

Wapman & Belle, Boston University

Training data are collected and annotated

Human Biases in Data

- Reporting bias
- Selection bias
- Overgeneralization
- Out-group homogeneity bias
- Stereotypical bias
- Historical unfairness
- Implicit associations
- Implicit stereotypes
- Prejudice
- Group attribution error
- Halo effect

Human Biases in Collection and Annotation

- Sampling error
- Non-sampling error
- Insensitivity to sample size
- Correspondence bias
- In-group bias
- Bias blind spot
- Confirmation bias
- Subjective validation
- Experimenter's bias
- Choice-supportive bias
- Neglect of probability
- Anecdotal fallacy
- Illusion of validity

Data

Reporting bias: What people share is not a reflection of real-world frequencies

Selection Bias: Selection does not reflect a random sample

Out-group homogeneity bias: People tend to see outgroup members as more alike than ingroup members when comparing attitudes, values, personality traits, and other characteristics

Confirmation bias: The tendency to search for, interpret, favor, and recall information in a way that confirms one's preexisting beliefs or hypotheses

Interpretation

Overgeneralization: Coming to conclusion based on information that is too general and/or not specific enough

Correlation fallacy: Confusing correlation with causation

Automation bias: Propensity for humans to favor suggestions from automated decision-making systems over contradictory information without automation

More at: https://developers.google.com/machine-learning/glossary/

Biases in Data

Biases in Data

Selection Bias: Selection does not reflect a random sample

Biases in Data

Out-group homogeneity bias: Tendency to see outgroup members as more alike than in-group members

ERROR RATE(1-PPV) BY FEMALE x SKIN TYPE

	TYPE I	TYPE II	TYPE III	TYPE IV	TYPE V	TYPE VI
Microsoft	1.7%	1.1%	3.3%	0%	23.2%	25.0%
FACE++	11.9%	9.7%	8.2%	13.9%	32.4%	46.5%
IBM	5.1%	7.4%	8.2%	8.3%	33.3%	46.8%

#GenderShades

Buolamwini et al. FAT* 2018, Slides from Joy Buolamwini

Positive predictive value (PPV)

$$\mathrm{PPV} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false positives}} = \frac{\text{number of true positives}}{\text{number of positive calls}}$$

		True condition			
	Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$	F_1 score $= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$

Biases in Data → Biased Labels

Annotations in your dataset will reflect the worldviews of your annotators.

https://ai.googleblog.com/2018/09/introducing-inclusive-images-competition.html

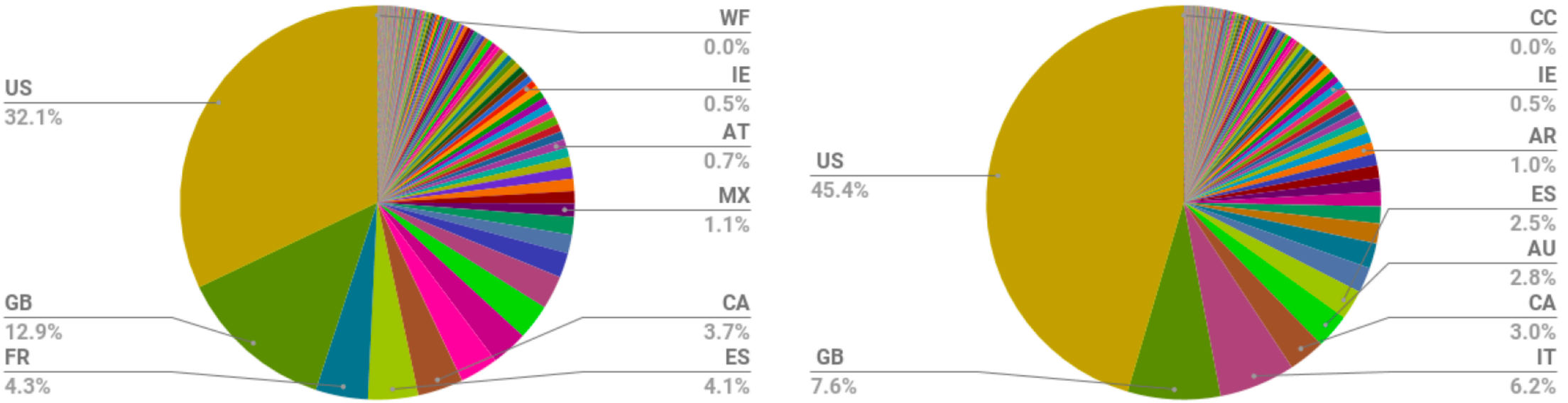


Figure 1: Fraction of Open Images and ImageNet images from each country. In both data sets, top represented locations include the US and Great Britain. Countries are represented by their two-letter ISO country codes.

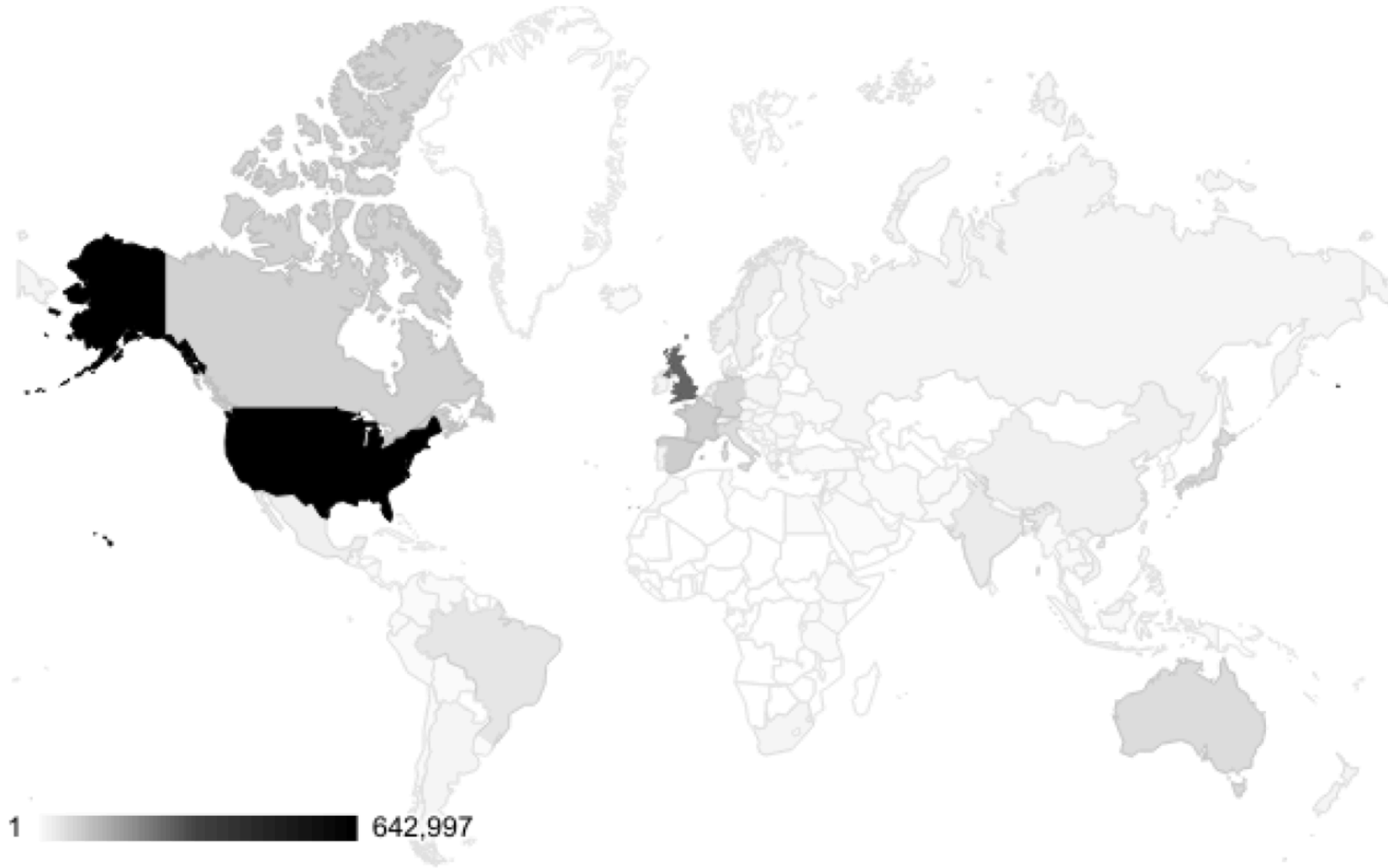


Figure 2: Distribution of the geographically identifiable images in the Open Images data set, by country. Almost a third of the data in our sample was US-based, and 60% of the data was from the six most represented countries across North America and Europe.

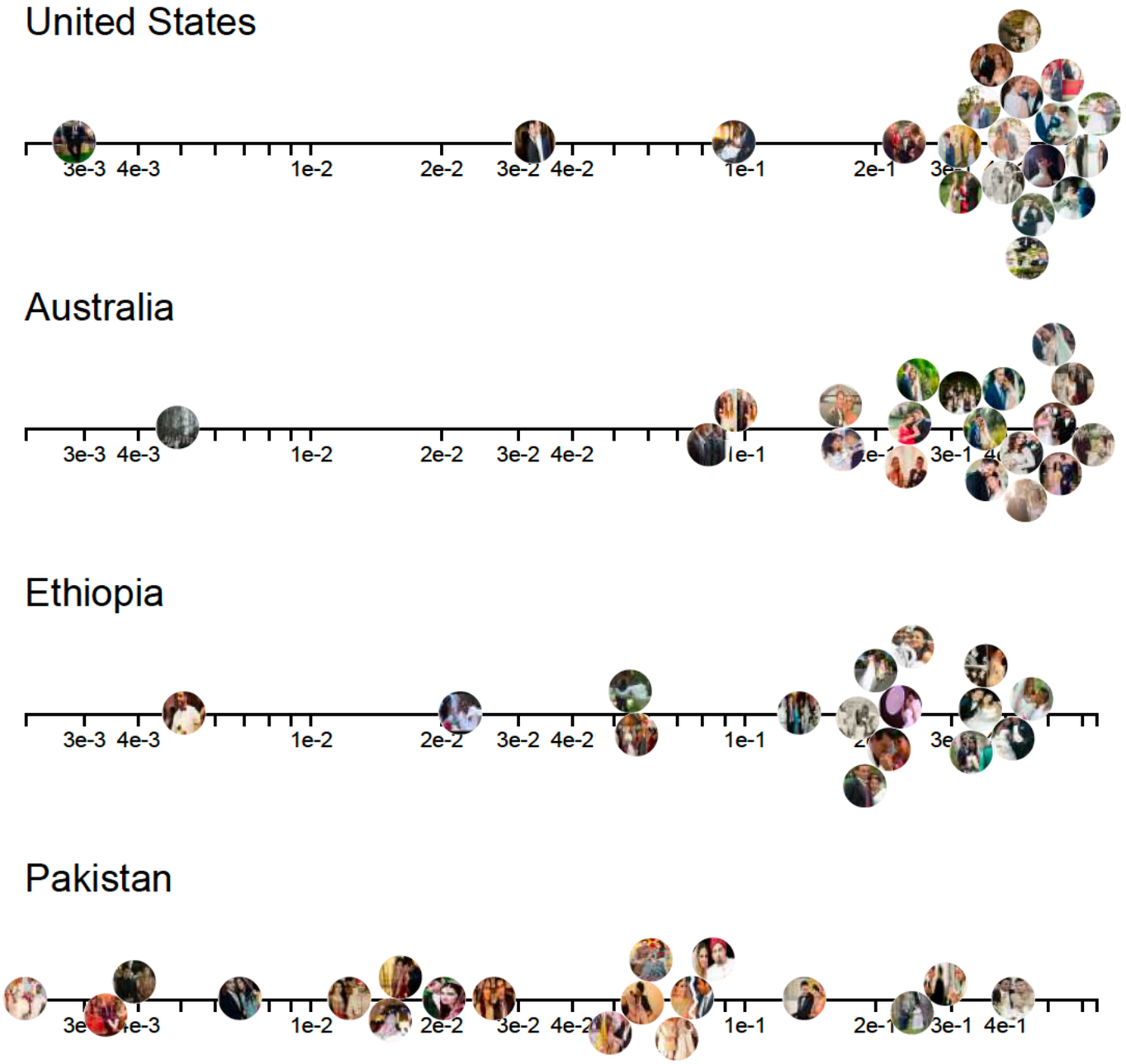


Figure 5: Photos of bridegrooms from different countries aligned by the log-likelihood that the classifier trained on Open Images assigns to the bridegroom class. Images from Ethiopia and Pakistan are not classified as consistently as images from the United States and Australia.

US ADULTS INDEXED

130 MILLION

One in two American adults is in a law enforcement face recognition network used in unregulated searches employing algorithms with unaudited accuracy.

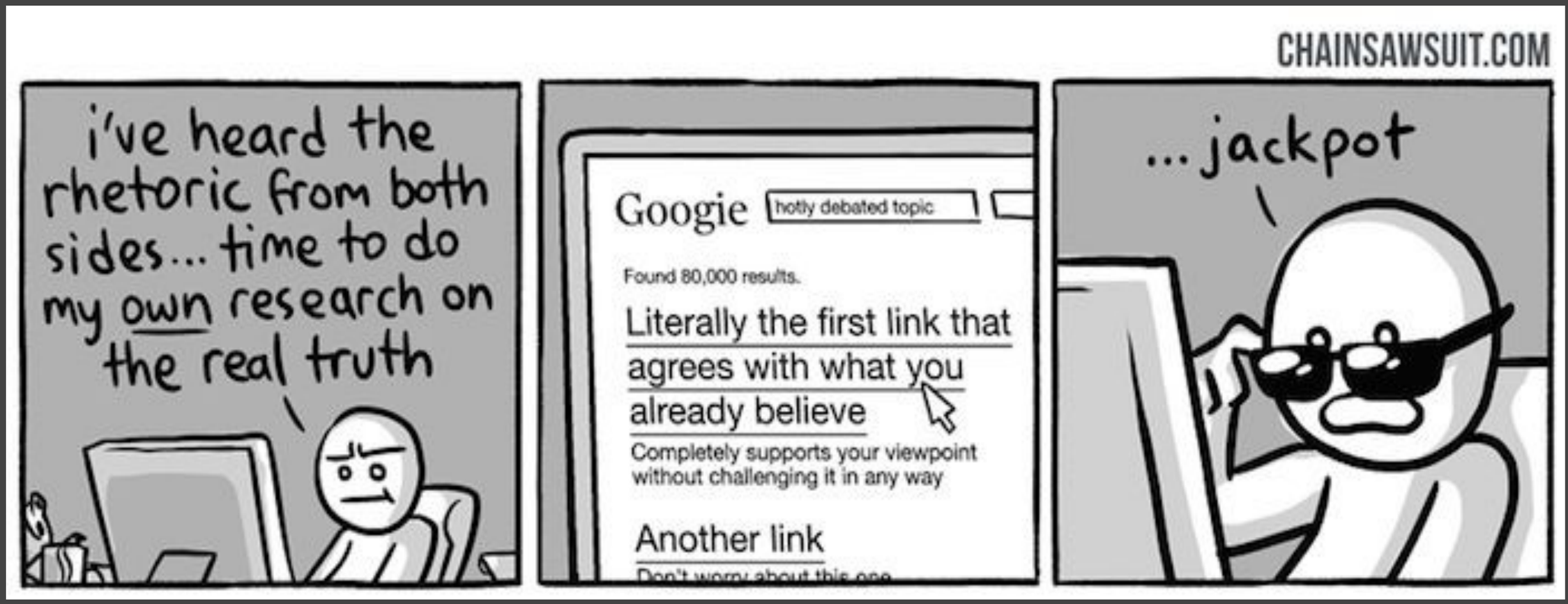
The Perpetual Line Up

(Garvie , Bedoya, Frankle 2016)

Biases in Interpretation

Biases in Interpretation

Confirmation bias: The tendency to search for, interpret, favor, recall information in a way that confirms preexisting beliefs



Biases in Interpretation

Overgeneralization:
Coming to conclusion based on information that is too general and/or not specific enough (related: **overfitting**)

[Sidney Harris]

Biases in Interpretation

Correlation fallacy: Confusing correlation with causation

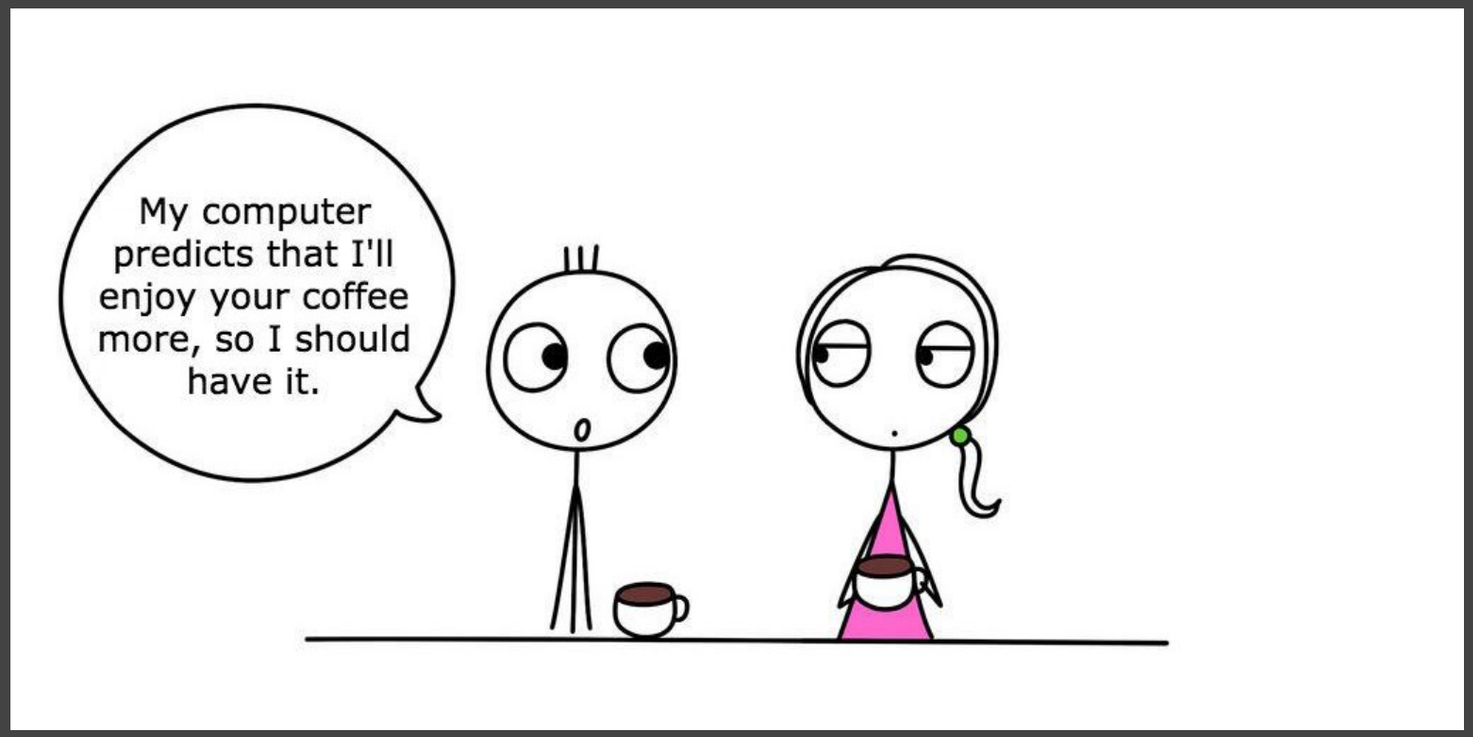
Post Hoc Ergo Propter Hoc

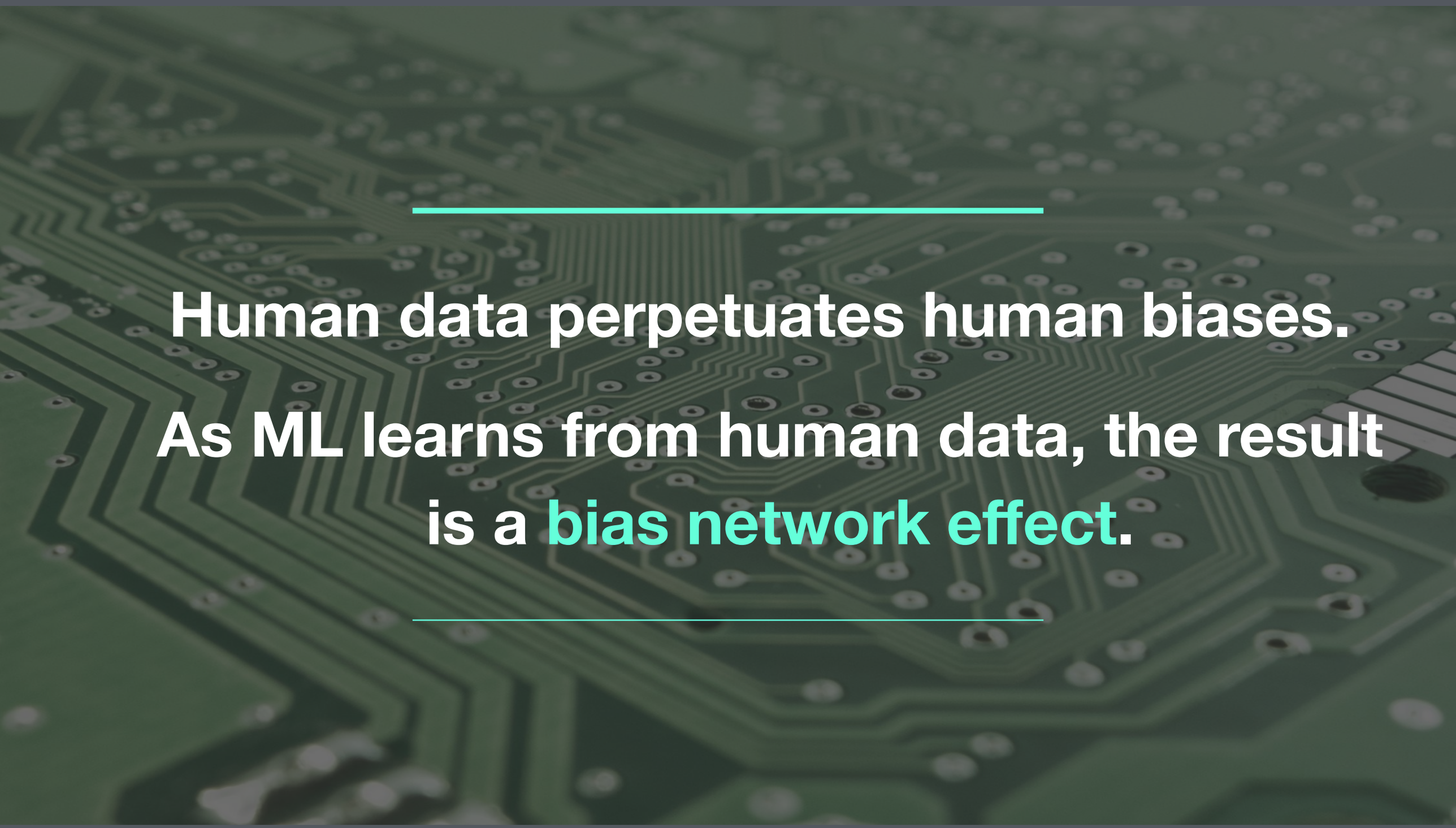
after this, therefore because of this

Women were allowed to vote in the early 1900's and then we had two world wars. Clearly giving them the vote was a bad idea.

Biases in Interpretation

Automation bias: Propensity for humans to favor suggestions from automated decision-making systems over contradictory information without automation




Human data perpetuates human biases.
As ML learns from human data, the result is a bias network effect.

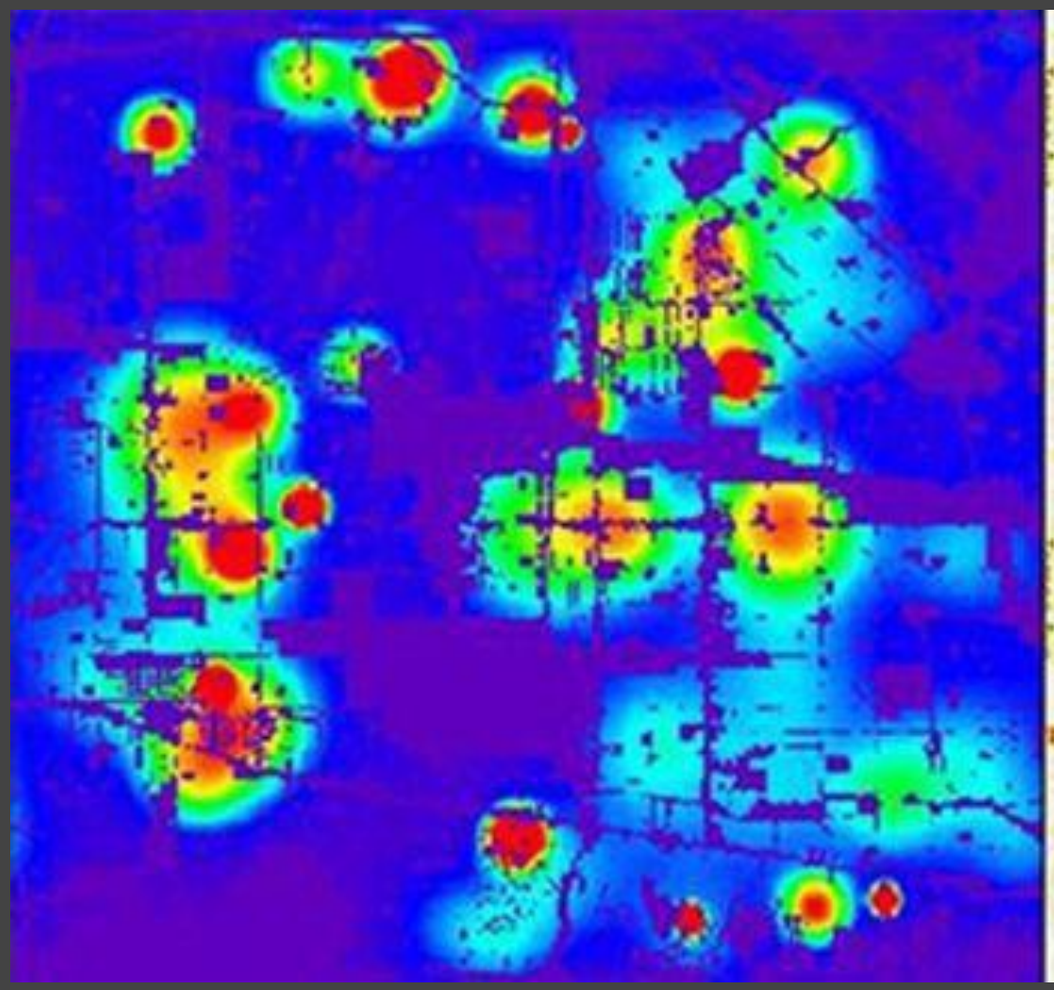
Algorithmic bias

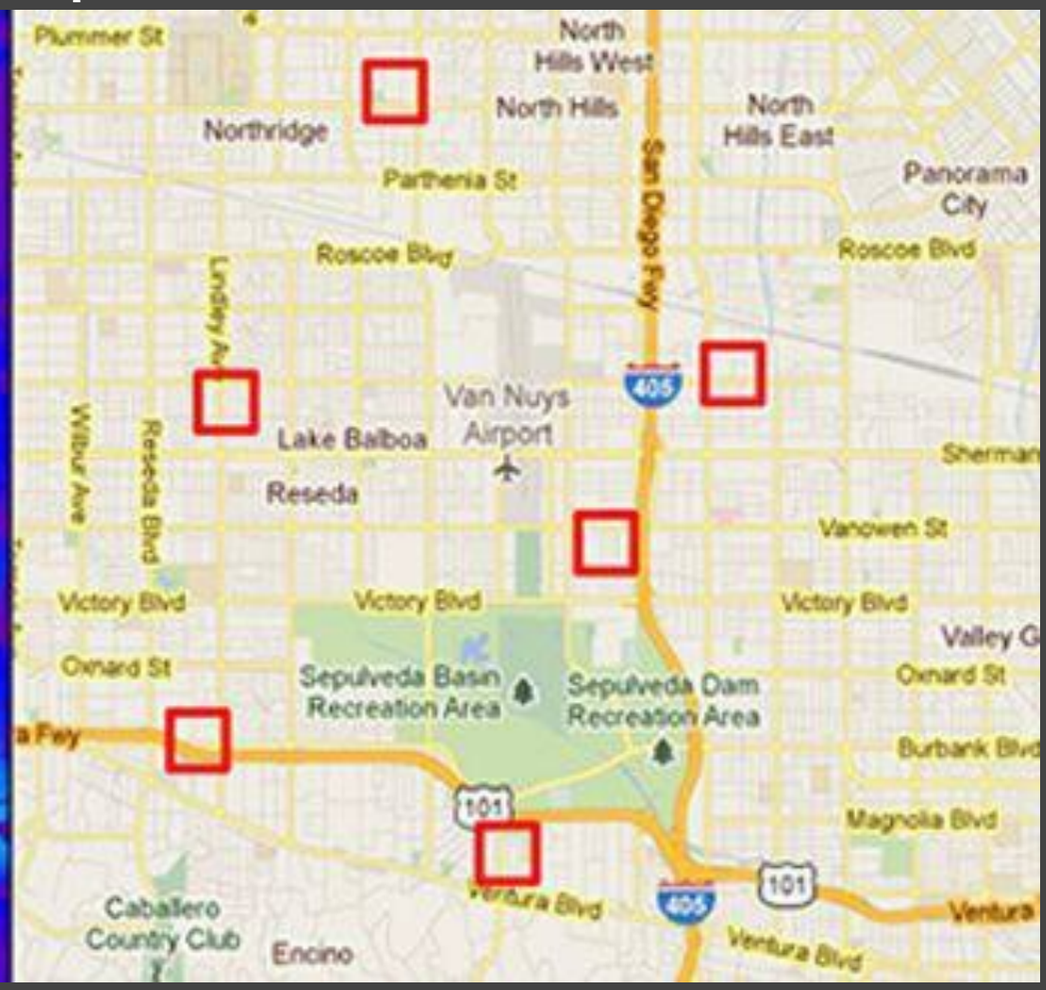
Unjust, unfair, or prejudicial treatment of people related to race, income, sexual orientation, religion, gender, and other characteristics associated with discrimination and marginalization, when and where they manifest in algorithmic systems or algorithmically aided decision-making

Predicting Future Criminal Behavior

Predicting Policing

- Algorithms identify potential crime hot-spots
- Based on where crime is previously reported, not where it has occurred
- Predicts future events from past





Smithsonian. *Artificial Intelligence Is Now Used to Predict Crime. But Is It Biased?* 2018

Predicting Criminality

Israeli startup, Faception

> *"Faception is first-to-technology and first-to-market with proprietary computer vision and machine learning technology for profiling people and* ***revealing their personality based only on their facial image****."*

Offering specialized engines for recognizing "High IQ", "White-Collar Offender", "Pedophile", and "Terrorist" from a face image.

Main clients are in homeland security and public safety.

Predicting Criminality

"Automated Inference on Criminality using Face Images" Wu and Zhang, 2016. arXiv

1,856 closely cropped images of faces; Includes "wanted suspect" ID pictures from specific regions.

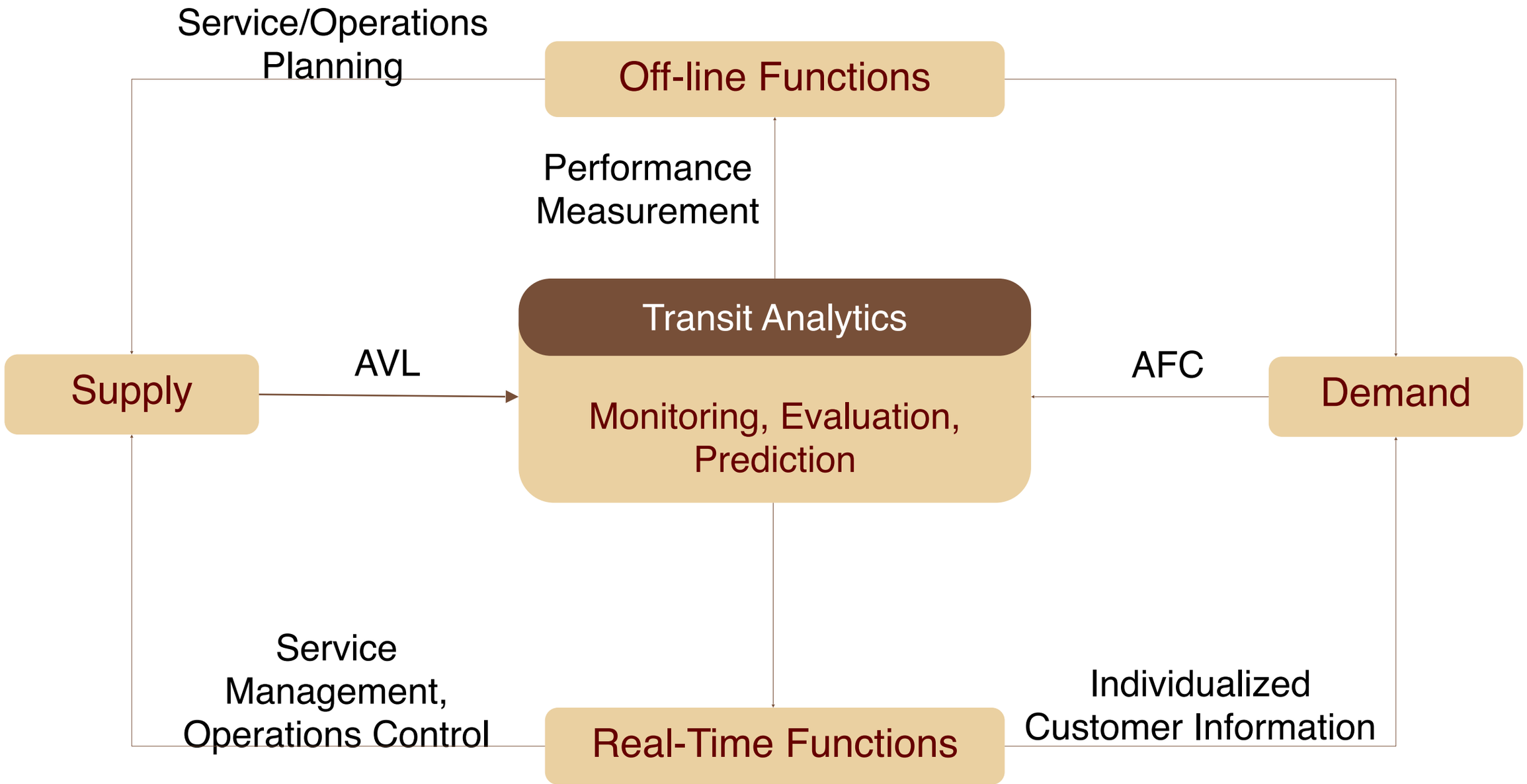
"[…] angle θ from nose tip to two mouth corners is on average 19.6% smaller for criminals than for non-criminals ..."

Selection Bias + Experimenter's Bias + Confirmation Bias + Correlation Fallacy + Feedback Loops

Biases in Transportation

Transit

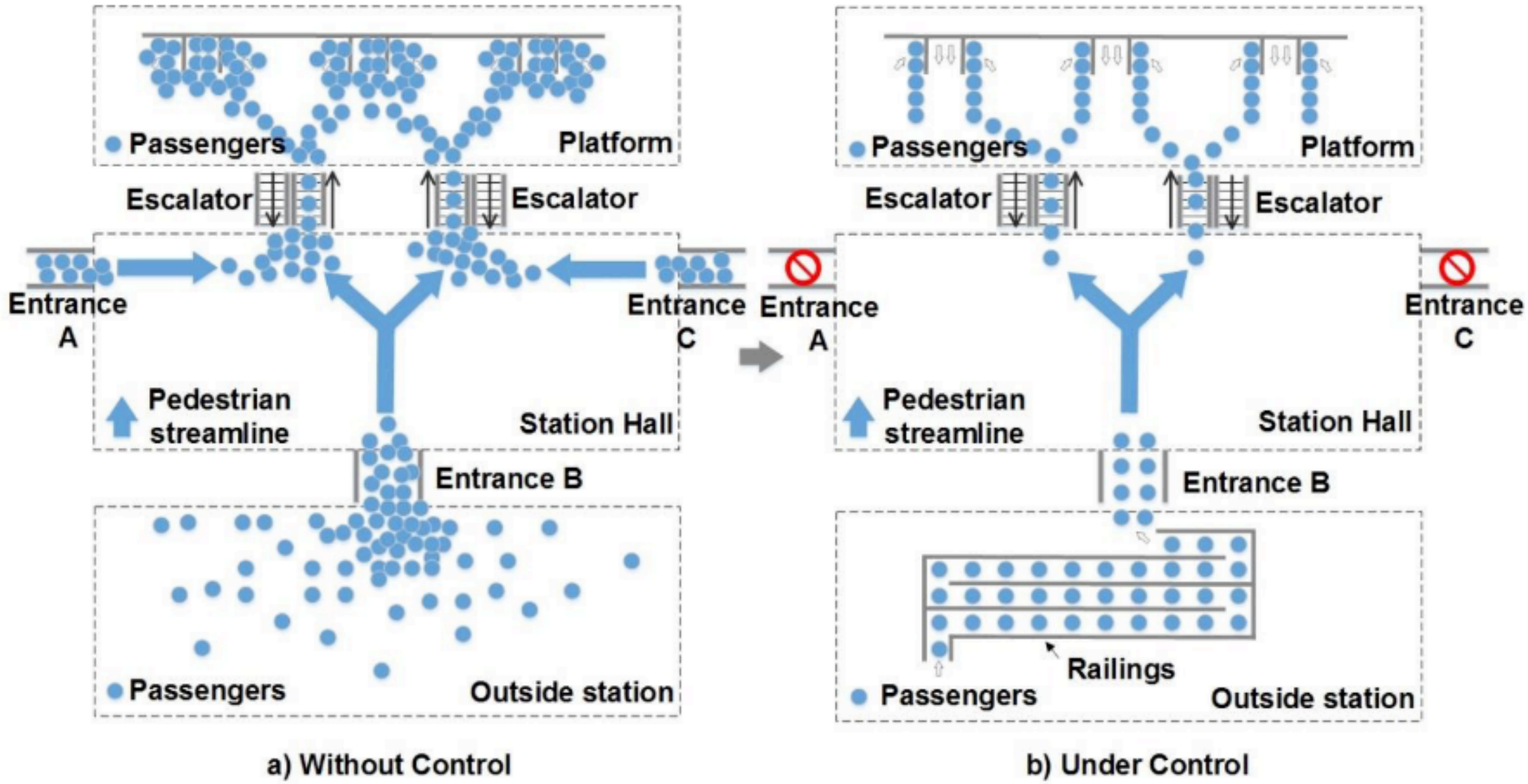
Overall framework



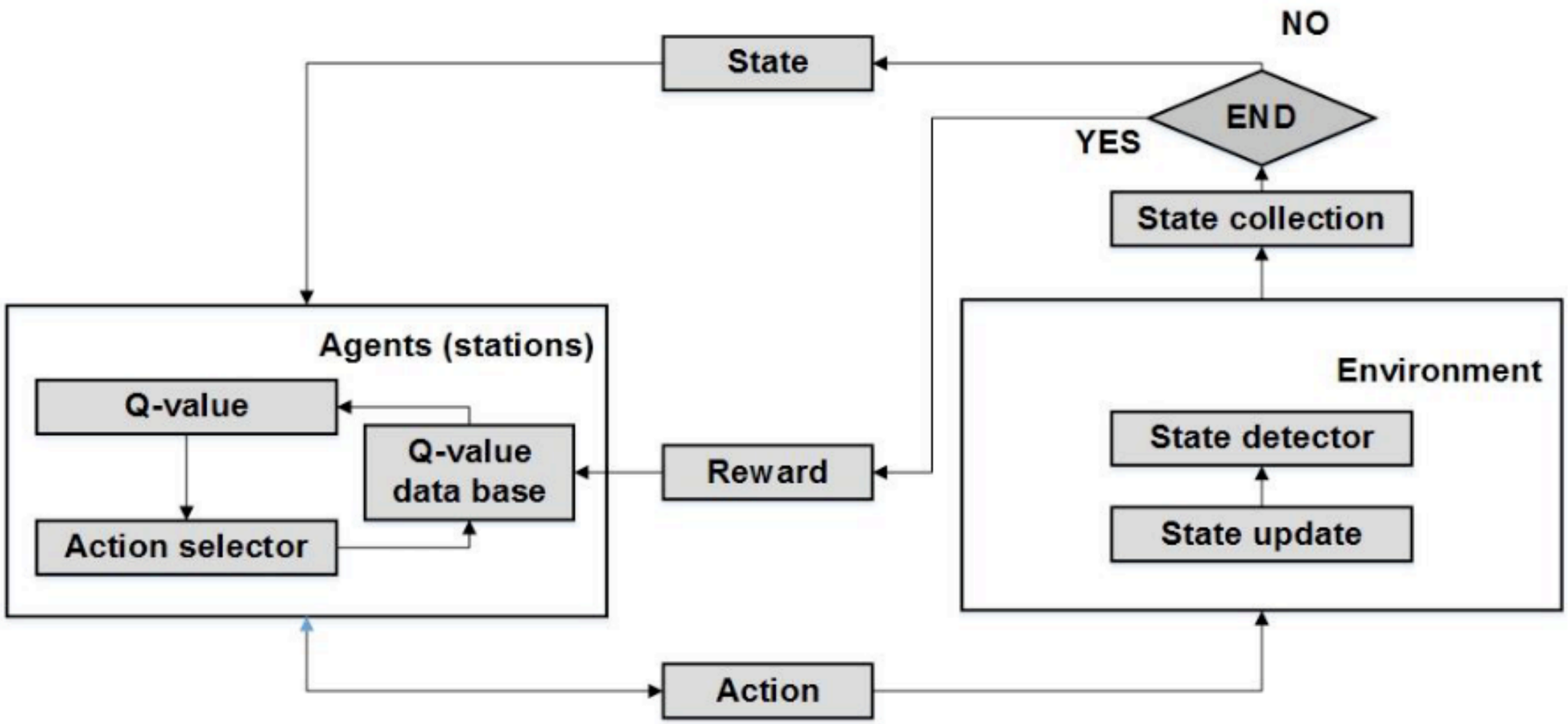
Proactive crowd management

Proactive crowd management

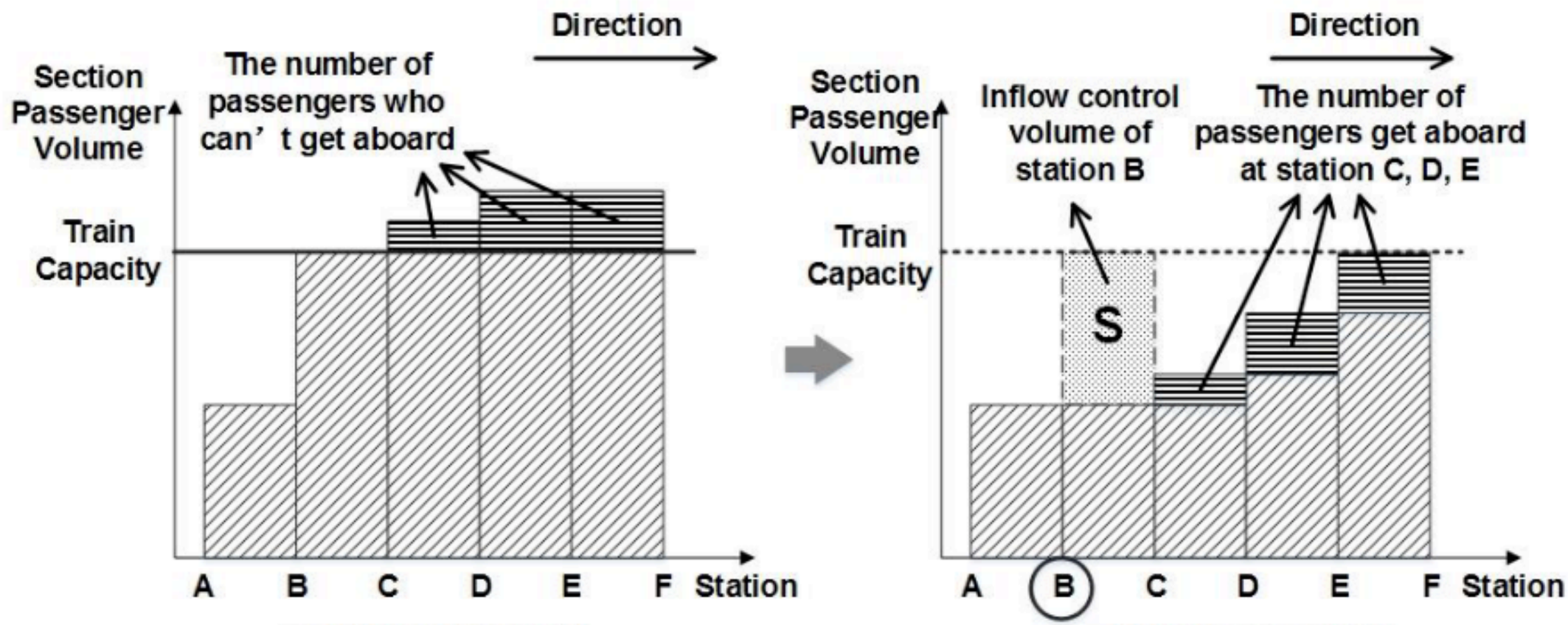
- Example: Beijing Metro



Proactive crowd management



Proactive crowd management



a) Without Control
(station C, D, E suffered from passenger congestion)

b) Under Control
(The passenger congestion is relieved at station C, D, E)

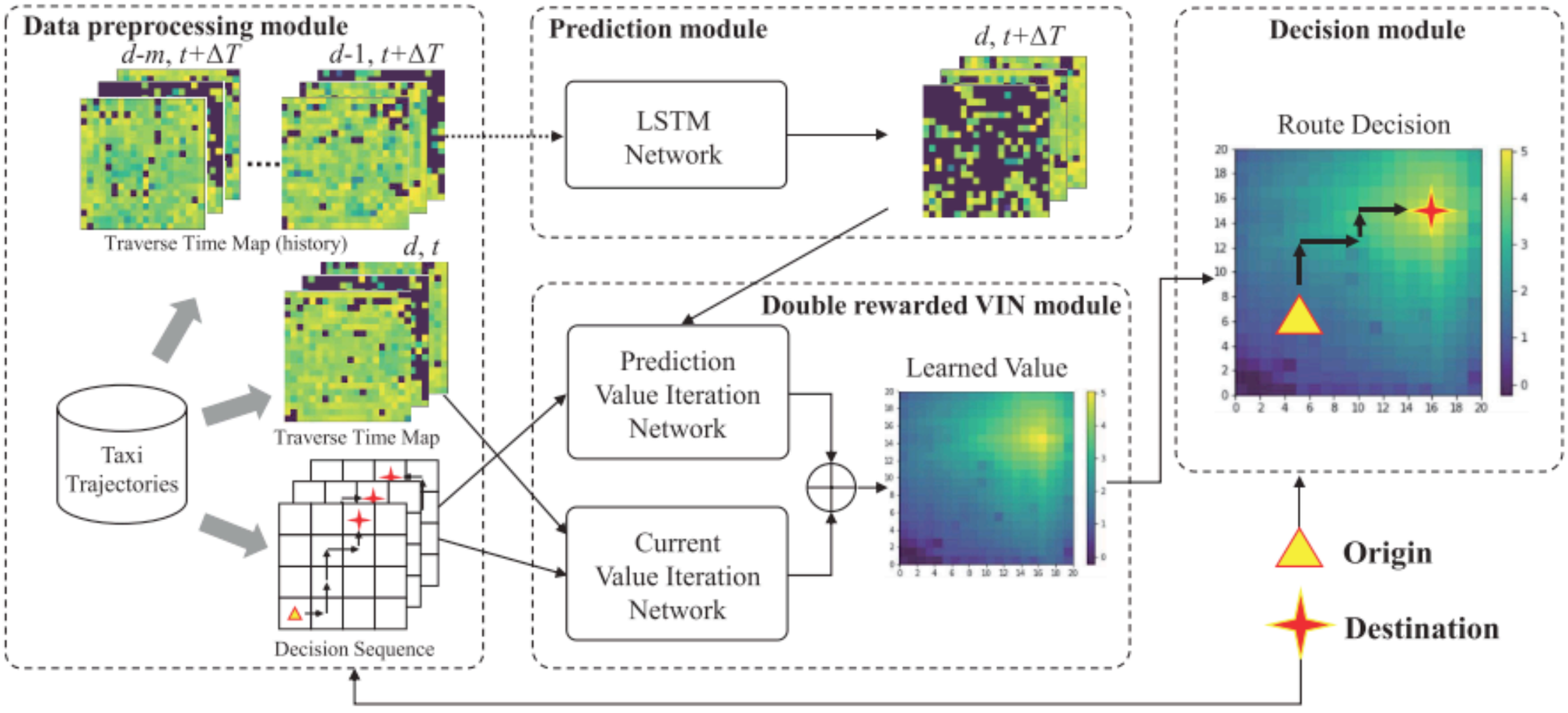
Ride hailing

Ride hailing systems

- Short-term demand forecasting is crucial for operation
- Examples: Uber, Didi, and Lyft
- Applications:
  - Dispatching efficiency
    - Driver to zone assignment
  - Adaptive pricing
  - Rebalancing
  - Passenger information

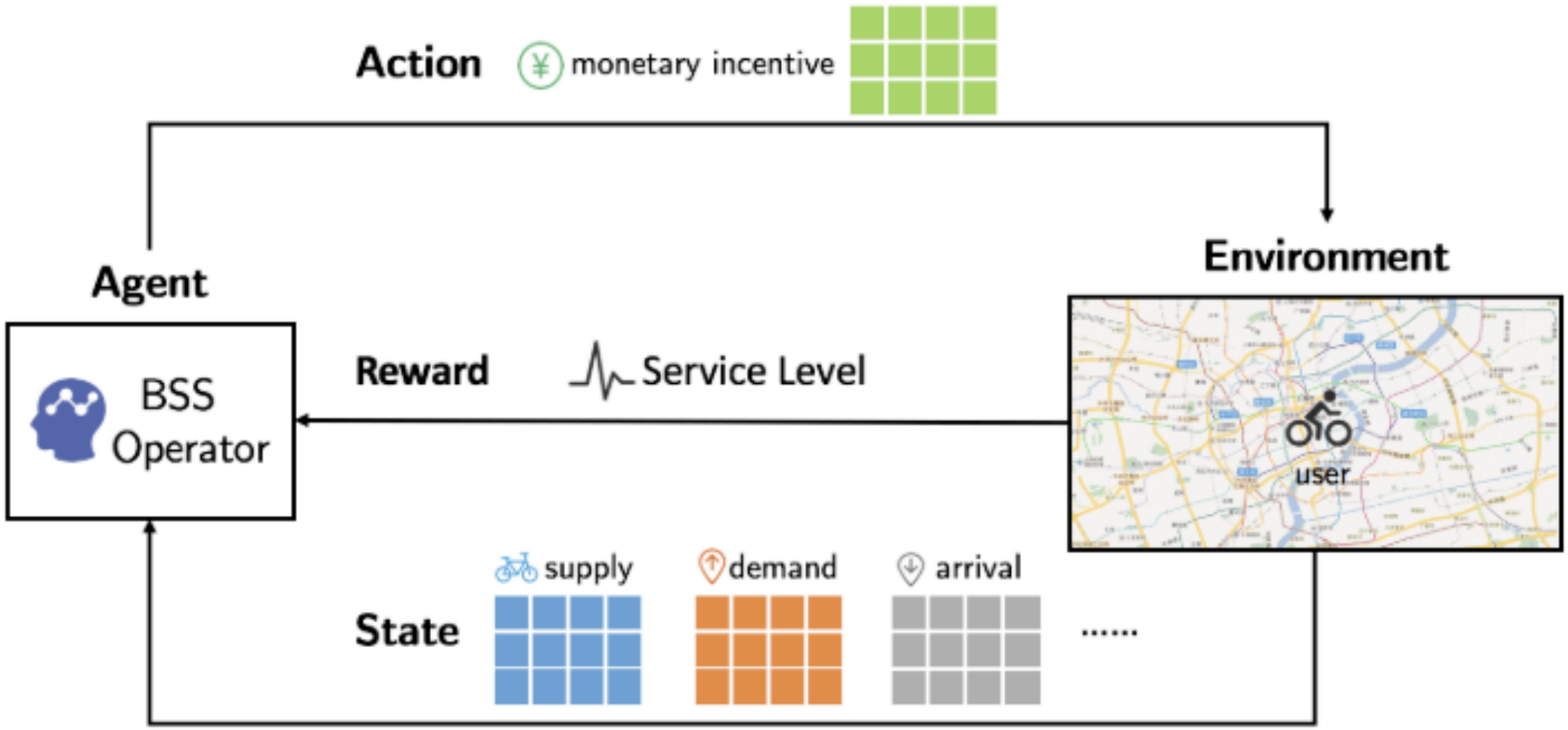
Ride hailing systems with AV

- RL for centrally controlled taxi operation



Incentive design

- Incentivize drivers to move to zones that are poorly served



Navigation

LA residents complain about ‘Waze Craze’

PUBLISHED THU, DEC 11 2014•2:54 PM EST

Jane Wells
@JANEWELLS
@11828246968596348685 2

SHARE

The Waze Effect: 4 Steps for Cities to Fight Back

https://www.streetlightdata.com/waze-effect-4-steps-for-cities-to-fight-back/

By Kaleb Osagie | August 28, 2018

We don’t know What types of biases we are propagating

Solutions?

Technical solutions proposed (e.g. train algorithms to satisfy “fairness” criteria)

Working paper: large-scale analysis as an empirical benchmark study

Predicting Travel Mode Choice with 86 Machine Learning Classifiers: An Empirical Benchmark Study

Shenhao Wang
Baichuan Mo
Jinhua Zhao

Massachusetts Institute of Technology

Abstract

Researchers are applying a large number of machine learning (ML) classifiers to predict travel behavior, but the results are data-specific and the selection of ML classifiers is author-specific. To obtain generalizable results, this paper provides an empirical benchmark by using 86 classifiers from 14 model families to predict the travel mode choice based on the National Household Travel Survey (NHTS) 2017 dataset. The 86 ML classifiers from 14 model families incorporate all the important ML classifiers discussed in previous studies. The large number of observations (about 800,000) in the NHTS2017 dataset enables us to analyze the effect of different sample sizes as a meta-dimension on prediction accuracy. We found that **ensemble models**, including boosting, bagging, and random forests, perform the best among all the classifiers, and that **deep neural networks** (DNNs) perform the best among all the non-ensemble models. Classical **discrete choice models** (DCMs) only predict at the medium or relatively low range

Technical Solution

$$\min_{X \succeq 0} \|AXA^T - AA^T\|_F^2 + \lambda \|PXB^T\|_F^2$$

$$X = TT^T$$

Bolukbasi, T., Chang, K. W., Zou, J., Saligrama, V., & Kalai, A. (2016). Quantifying and reducing stereotypes in word embeddings. arXiv preprint arXiv:1606.06121.

Difficulty

- At least six kinds of fairness, some of which are incompatible with one another and with accuracy.
- Except in trivial cases, it is impossible to maximize accuracy and fairness at the same time, and impossible simultaneously to satisfy all kinds of fairness.
- In practice, a major complication is different base rates across different legally protected groups. There is a need to consider challenging tradeoffs.

Fairness in Criminal Justice Risk Assessments:
The State of the Art

Richard Berk[a,b], Hoda Heidari[c], Shahin Jabbari[c],
Michael Kearns[c], Aaron Roth[c]

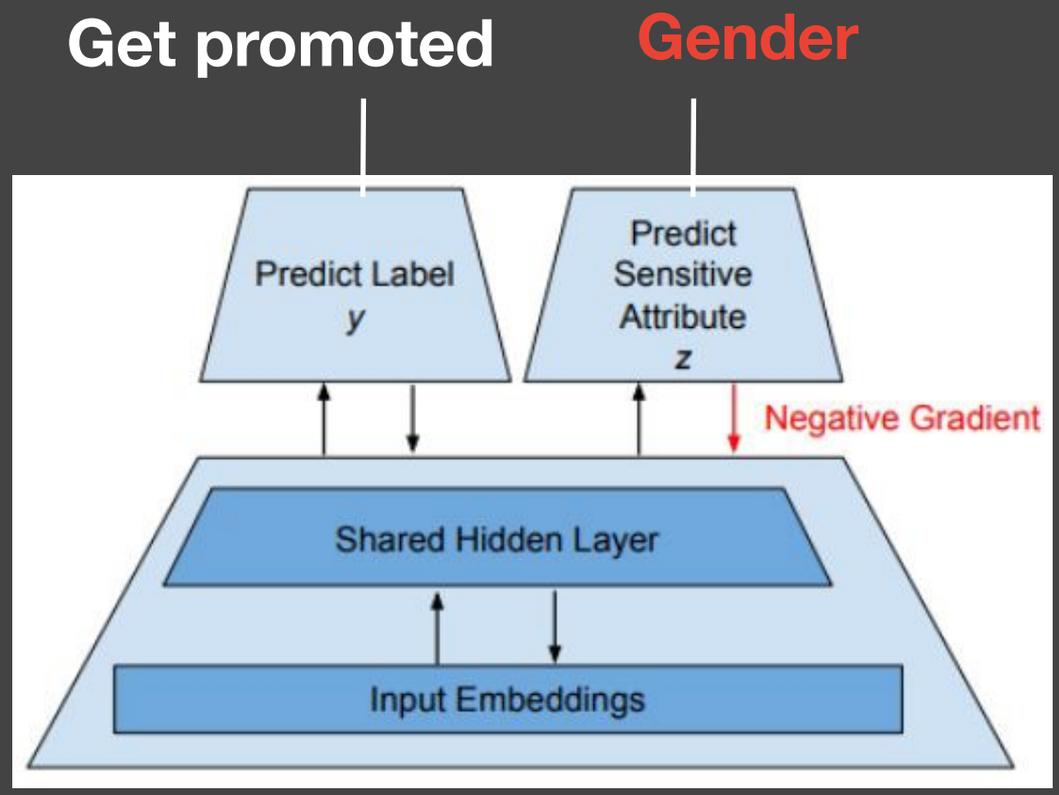
		True condition			
	Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$	F_1 score $= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$

Adversarial Multi-task Learning to Mitigate Bias

Multitask Adversarial Learning

$$Z \perp D \mid Y$$

- Basic idea: Jointly predict:
  - Output decision *D*
  - Attribute you'd like to remove from decision *Z*
  - Negate the effect of the undesired attribute



$$P(\hat{Y} = 1 | Y = 1, Z = 1) = P(\hat{Y} = 1 | Y = 1, Z = 0)$$

Beutel, Chen, Zhao, Chi. Data Decisions and Theoretical Implications when Adversarially Learning Fair Represent *2017.*

Zhang, Lemoine, Mitchell. Mitigating Unwanted Biases with Adversarial Learning. *AIES, 2018*.

Equality of Opportunity in Supervised Learning

A classifier's output decision should be the same **across sensitive characteristics**, given what the correct decision should be.

Hardt, Price, Srebro. Equality of Opportunity in Supervised Learning. *NIPS, 2016.*

Principles

Asilomar conference for Beneficial AI in 2017

Asilomar AI principles

- https://futureoflife.org/ai-principles/
- Recommendation 6, “AI systems should be safe and secure throughout their operational lifetime, and verifiably so,”
- Recommendation 7, “If an AI system causes harm, it should be possible to ascertain why,” clearly speak to the importance of transparency.
- Recommendation 10, “Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation,”

THE LIMITS OF PRINCIPLES

- Different Groups May Interpret Principles Differently
- Principles Are Highly General
- Principles Come into Conflict in Practice

Focus on Tensions

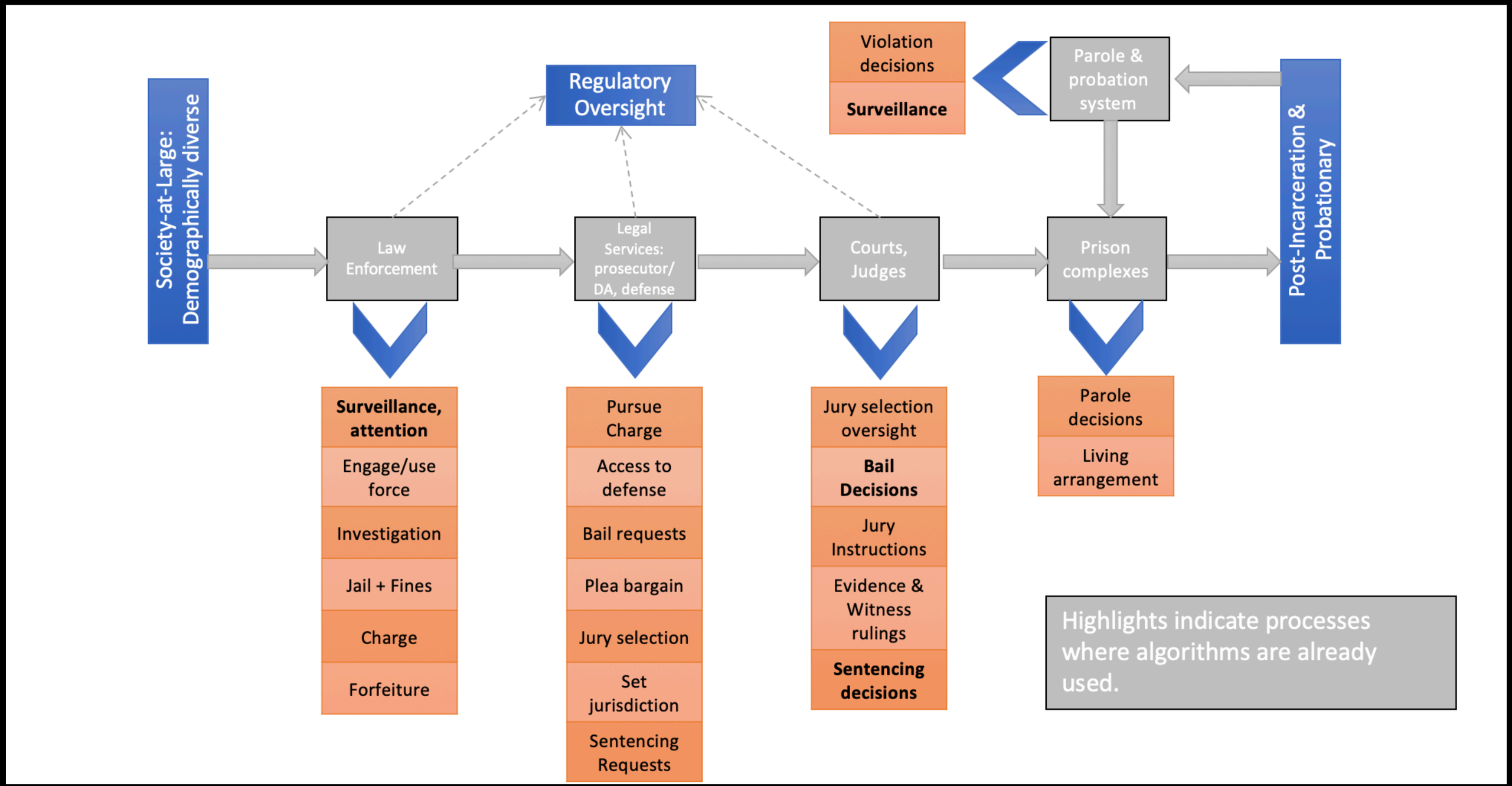
- Bridging the Gap between Principles and Practice
- Acknowledging Differences in Values
- Highlighting Areas Where New Solutions Are Needed
- Identifying Ambiguities and Knowledge Gaps

Four Key Tensions

- Tension 1: Using data to improve the quality and efficiency of services vs. respecting privacy and autonomy of individuals.
- Tension 2: Using algorithms to make decisions and predictions more accurate vs ensuring fair and equal treatment.
- Tension 3: Reaping the benefits of increased personalisation in the digital sphere vs enhancing solidarity and citizenship.
- Tension 4: Using automation to make people's lives more convenient and empowered vs promoting self-actualization and dignity.

Social-techno system

Socio-Technical Systems…
"No AI is an Island"



Remediation/Regulatory Strategies

Algorithmic Audit

Require algorithm audits via:

- Data worksheets
- Algorithmic Impact Assessments
- Open algorithm validation tests
- Ex-Post/post-harm compensations

Industry Standards & Task-forces

Industry-led coalition on:

- Setting standards
- Accrediting model
- Punishing bad actors.

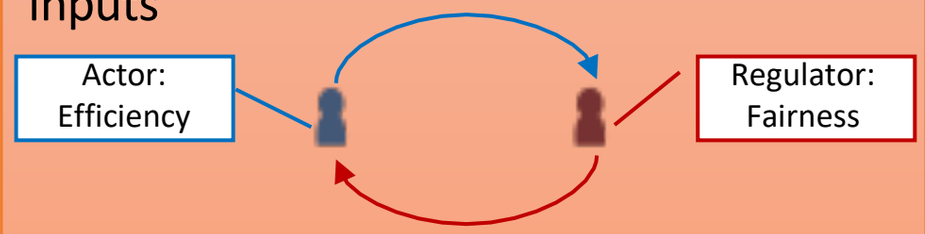
Accountable to government

The "Regulation Game" Framework

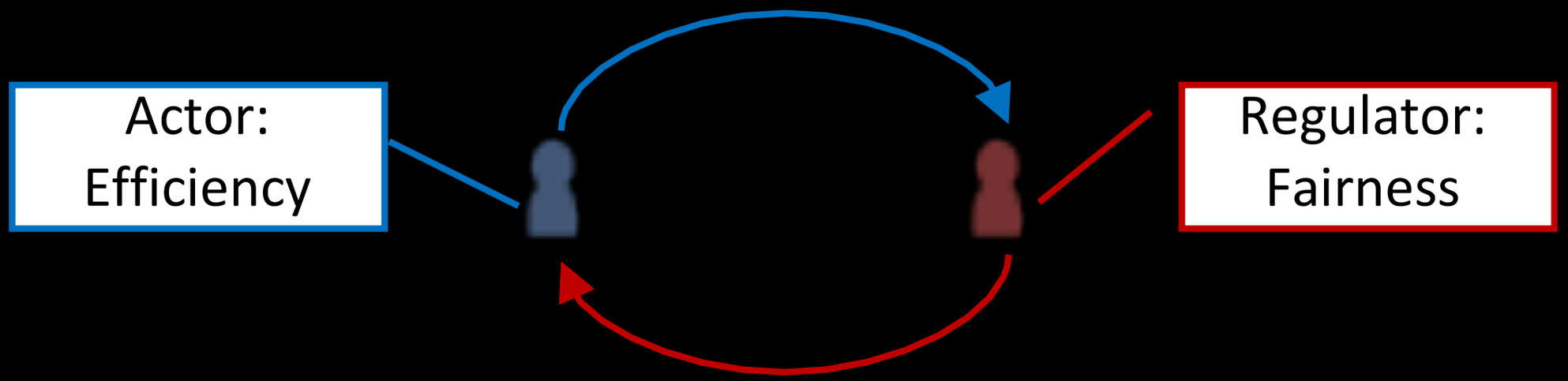
Explicit implementation of regulatory framework that divides responsibility for the separate sets of competing normative goals and incentives.

"Actor" vs. "Regulator."

Regulate actor's output/behavior, not inputs



“Regulation Game” Framework



- Explicit implementation of regulatory framework that divides responsibility for the separate sets of competing
- normative goals and incentives.
- “Actor” vs. “Regulator.”
- Regulate actor’s output/behavior, not inputs

What is a good model?

Interpretability —> FAT

- Interpretability is important because of
  - Trust
  - Safety (e.g. AVs detect objects)
  - Transparent governance (e.g. evaluate the eligibility for food stamps)
  - New knowledge generation (e.g. academia)

Improving ML Interpretability

- Before building models
  - Exploratory data analysis (dimension reduction, etc.)
- Building a new model
  - Use an interpretable model (DT, KNN, etc.)
  - Sparsity (e.g. linear model with hundreds of features, etc.)
- After building a model (Post-hoc interpretation for DNN)
  - Case-based methods
  - Gradient-based methods
  - Local explanation models
  - Global explanation models
  - Visualize hidden layers

Domain-Specific Models		Machine Learning Models
Spatial-temporal prediction	← - - - - - - →	CNN/RNN/LSTM
Demand analysis (DCM, PT)	← - - - - - - →	Supervised learning (DNN)
Network analysis	← - - - - - - →	Graphical neural networks
Feedback & system control	← - - - - - - →	Reinforcement learning

Interpretation - - - - ▸

Domain-Specific Models

← - - - Prediction

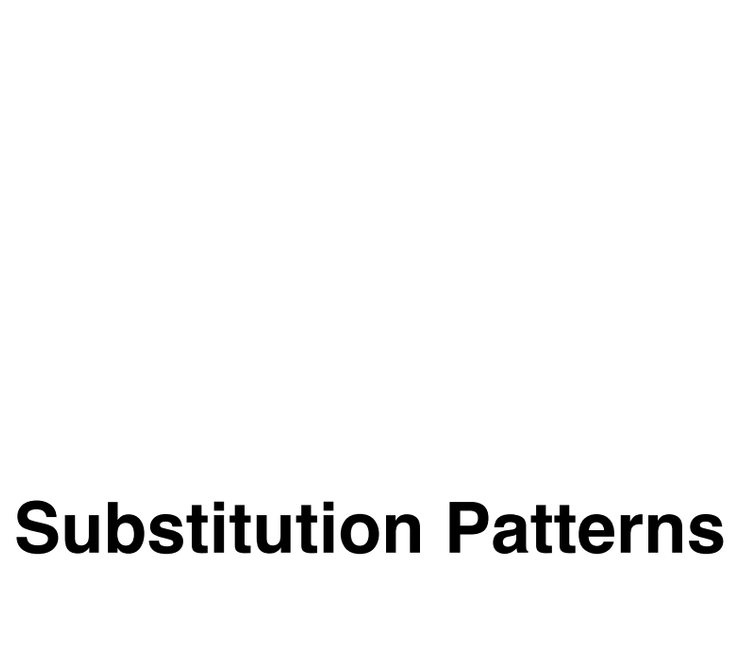
Machine Learning Models

Robustness - - - - ▸

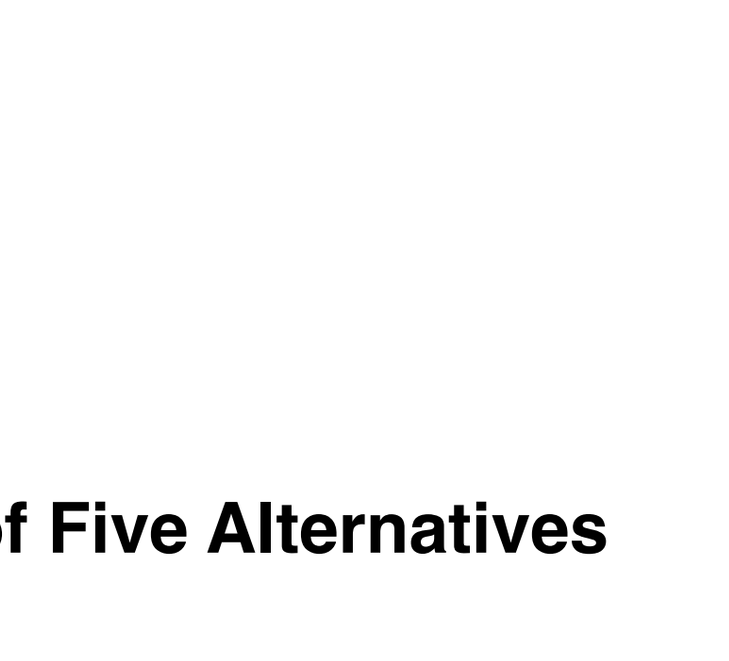
Substitution Patterns of Five Alternatives

$$s_{k_1}(x_j; x_{\setminus j}) / s_{k_2}(x_j; x_{\setminus j})$$

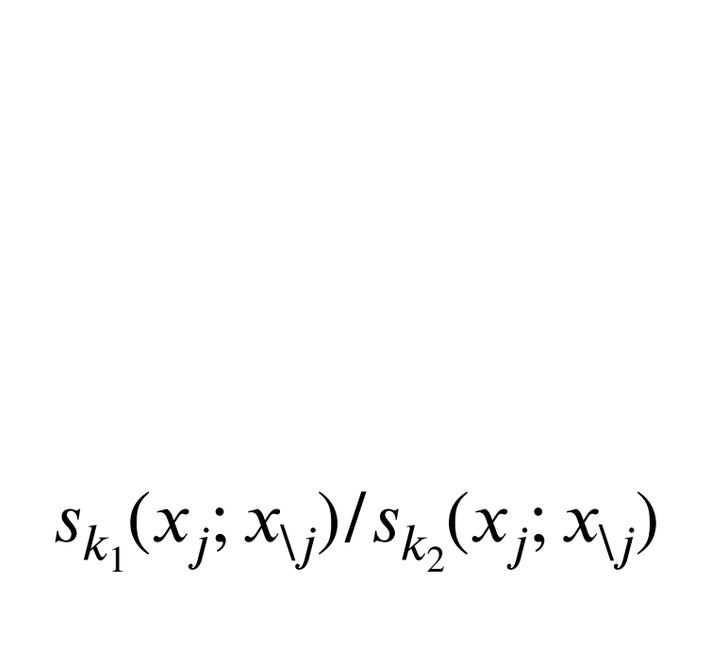
5L-DNNs



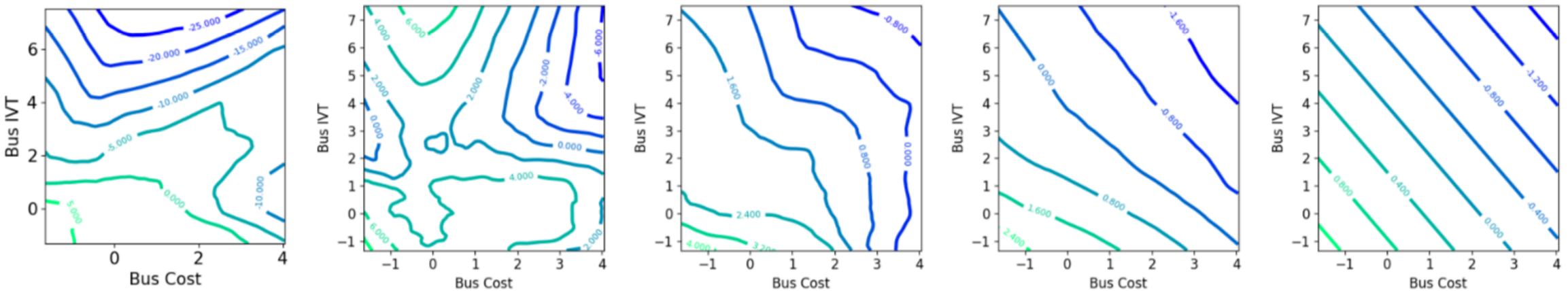
HP-DNNs



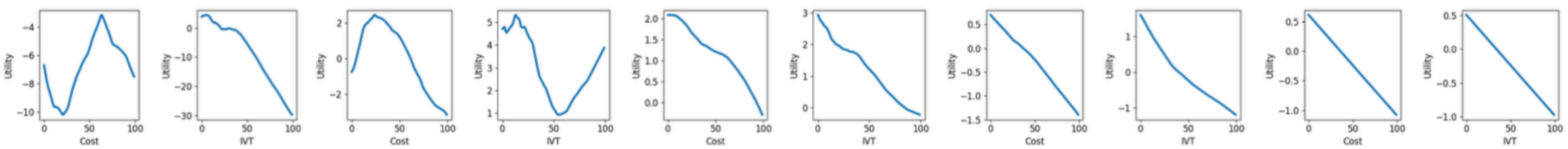
MNL



2. Interpretability of Utility Function in the CM Scenario



(a) DNN (55.2%) (b) CM Resnet ($\lambda = 1e-10$; 56.4%) (c) CM Resnet ($\lambda = 0.005$; 57.3%) (d) CM Resnet ($\lambda = 0.01$; 56.8%) (e) CM (44.7%)



(f) x0 (g) x1 (h) x0 (i) x1 (j) x0 (k) x1 (l) x0 (m) x1 (n) x0 (o) x1

DNN and DCM

- Both subject to biases
- DCM more explicit
- DNN less explicit, more dangerous

Tensions in DNN and Classic Models

- Role of Model: explain or predict
- Role of Theory: Generic model vs domain specific
- Simple to complex vs complex to simple
- Causality vs correlation
- Understand vs action (control)
- Knowledge production: discover vs create

Criteria of a good model

- Prediction accuracy
- Interpretability (FAT)
- Robustness (security)
- Sparsity
- Practicality
- Difficulty

Do we have to trust after we understand?

Do we have to trust after we understand?

- Not that human being understand all the complexity in science and engineering;
- But still trust them
- What is interpretability?
    - Simulatability (e.g. linear vs. DNN)
    - Decomposability (e.g. linear model with complicated feature engineering)
    - Algorithmic transparency (e.g. DNN)
    - Post-hoc interpretability (e.g. DNN)
- Meta-narrative
- Institutions

Understanding the Limits of AI: When Algorithms Fail

Timnit Gebru
Google AI

Currently, any model can be used by anyone for anything

We need standards/ documentation

Other Industries Have Been There

Electronics

Lots of standardization, concept of datasheet

Electronics

MOUSER ELECTRONICS

Products | Manufacturers | Applications | Services & Tools | Help | Order History | Log In | Register | 0

All | Part # / Keyword | In Stock | RoHS

All Products > Passive Components > Capacitors > Tantalum Capacitors > Tantalum Capacitors - Polymer SMD >
KEMET T520B107M006ATE040

See an Error?

T520B107M006ATE040

KEMET

Enlarge

Images are for reference only
See Product Specifications

Share

Mouser #:	80-T520B107M6ATE40
Mfr. #:	T520B107M006ATE040
Mfr.:	KEMET
Customer #:	
Description:	Tantalum Capacitors - Polymer SMD 6.3volts 100uF 20% ESR=40 Available in MultiSIM BLUE View Simulation and SPICE Model in K-SIM
Datasheet:	T520B107M006ATE040 Datasheet
More Information:	Learn more about KEMET T520B107M006ATE040

In Stock: 7,998

Stock:	7,998 Can Ship Immediately
On Order:	2000 View Delivery Dates
Factory Lead-Time:	21 Weeks
Enter Quantity:	Minimum: 1 Multiples: 1 Buy

Pricing (USD)

Qty.	Unit Price	Ext. Price
1	$1.22	$1.22
10	$0.838	$8.38
100	$0.644	$64.40

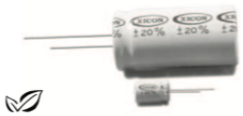
Electronics

XICON

Miniature Aluminum Electrolytic Capacitors — XRL Series

■ FEATURES

- Low impedance characteristics
- Case sizes are smaller than conventional general-purpose capacitors, with very high performance
- Can size larger than 9mm diameter has safety vents on rubber end seal
- RoHS Compliant



RoHS Compliant

■ CHARACTERISTICS

Item		Characteristics											
Operating Temperature Range		-40°C ~ +85°C											
Capacitance Tolerance		±20% at 120Hz, 20°C											
Leakage Current	≤100V	I = 0.01CWV or 3μA whichever is greater after 2 minutes of applied rated DC working voltage at 20°C Where: C = rated capacitance in μF; WV = rated DC working voltage											
	>100V	CWV ≤ 1000 μF: I= 0.03 CWV + 15uA; C= rated capacitance in uF CWV ≥ 1000 μF: I= 0.02 CWV + 25uA; WV= rated DC working voltage in V											
Dissipation Factor (Tan δ, at 20°C 120Hz)	Working voltage (WV)	6.3	10	16	25	35	50	63	100	160	250	350	450
	Tan δ	0.23	0.20	0.16	0.14	0.12	0.10	0.09	0.08	0.12	0.17	0.20	0.25
	For capacitors whose capacitance exceeds 1,000μF, the specification of tan δ is increased by 0.02 for every addition of 1,000μF												
Surge Voltage	Working voltage (WV)	6.3	10	16	25	35	50	63	100	160	250	350	450
	Surge voltage (SV)	8	13	20	32	44	63	79	125	200	300	400	500
Low Temperature Characteristics (Imp. ratio @ 120Hz)	Working voltage (WV)	6.3	10	16	25	35	50	63	100	160	250	350	450
	Z(-25°C)/Z(+20°C) øD<16	6	4	3	3	2	2	2	2	3	8	12	16
	Z(-25°C)/Z(+20°C) øD≥16	8	6	4	4	3	3	3	3	3	8	12	16
	Z(-40°C)/Z(+20°C) øD<16	10	8	6	6	4	3	3	3	4	10	16	20
	Z(-40°C)/Z(+20°C) øD≥16	18	16	12	10	8	8	6	6	4	10	16	20
Load Test	When returned to +20°C after 2,000 hours application of working voltage at +85°C, the capacitor will meet the following limits: Capacitance change is ≤ ±20% of initial value; tan δ is < 200% of specified value; leakage current is within specified value												
Shelf Life Test	When returned to +20°C after 1,000 hours at +85°C with no voltage applied, the capacitor will meet the following limits: Capacitance change is ≤ ±20% of initial value; tan δ is < 200% of specified value; leakage current is within specified value												

■ PART NUMBERING SYSTEM

1	4	0	–	X	R	L		1	6	V		1	0	0	–	R	C
Prefix				Series				Voltage Actual Value				Capacitance (μF) Actual Value				Suffix RoHS Compliant	

■ RIPPLE CURRENT AND FREQUENCY MULTIPLIERS

Capacitance (μF)	Frequency (Hz)				
	60 (50)	120	500	1K	≥10K
<100	0.70	1.0	1.30	1.40	1.50
100 ~ 1000	0.75	1.0	1.20	1.30	1.35
>1000	0.80	1.0	1.10	1.12	1.15

■ RIPPLE CURRENT AND TEMPERATURE MULTIPLIERS

Temperature (°C)	<50	70	85
Multiplier	1.78	1.4	1.0

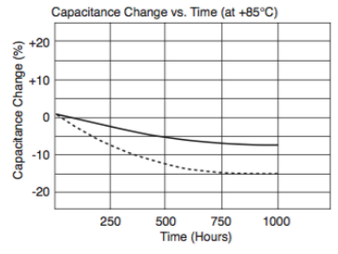
XICON

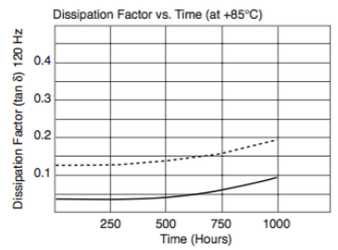
Miniature Aluminum Electrolytic Capacitors — XRL Series

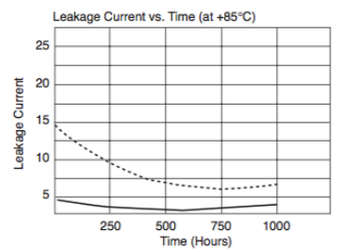
■ TYPICAL PERFORMANCE CHARACTERISTICS

---------- 1000μF 16V
———— 1μF 50V

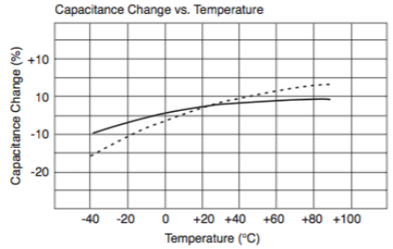
Life Test

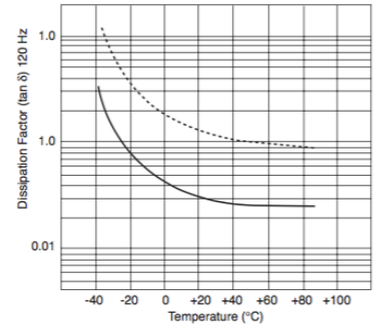


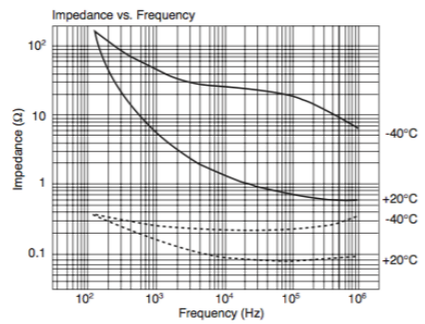




Temperature Characteristics







XICON PASSIVE COMPONENTS • (800) 628-0544

XC-600178 Date Revised: 1/8/07

Specifications are subject to change without notice. No liability or warranty implied by this information. Environmental compliance based on producer documentation.

We need datasheets for datasets, pertained APIs and models

Need to have information about dataset, recommended usage…

Datasheets for Datasets

Motivation for Dataset Creation

Why was the dataset created? (e.g., was there a specific task in mind? was there a specific gap that needed to be filled?)
Labeled Faces in the Wild was created to provide images that can be used to study face recognition in the unconstrained setting where image characteristics (such as pose, illumination, resolution, focus), subject demographic makeup (such as age, gender, race) or appearance (such as hairstyle, makeup, clothing) cannot be controlled. The dataset was created for the specific task of pair matching: given a pair of images each containing a face, determine whether or not the images are of the same person.[1]

What (other) tasks could the dataset be used for?
The LFW dataset can be used for the face identification problem. Some researchers have developed protocols to use the images in the LFW dataset for face identification.[2]

Has the dataset been used for any tasks already? If so, where are the results so others can compare (e.g., links to published papers)?
Papers using this dataset and the specified evaluation protocol are listed in http://vis-www.cs.umass.edu/lfw/results.html

Who funded the creation of the dataset?
The building of the LFW database was supported by a United States National Science Foundation CAREER Award.

Dataset Composition

What are the instances? (that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)
Each instance is a pair of images labeled with the name of the person in the image. Some images contain more than one face. The labeled face is the one containing the central pixel of the image—other faces should be ignored as "background".

Are relationships between instances made explicit in the data (e.g., social network links, user/movie ratings, etc.)?
There are no known relationships between instances except for the fact that they are all individuals who appeared in news sources on line, and some individuals appear in multiple pairs.

How many instances are there? (of each type, if appropriate)?
The dataset consists of 13,233 face images in total of 5749 unique individuals. 1680 of these subjects have two or more images and 4069 have single ones.

[1] All information in this datasheet is taken from one of five sources. Any errors that were introduced from these sources are our fault. Original paper: http://www.cs.cornell.edu/people/pabo/movie-review-data/; LFW survey: http://vis-www.cs.umass.edu/lfw/lfw.pdf; Paper measuring LFW demographic characteristics : http://biometrics.cse.msu.edu/Publications/Face/HanJain_UnconstrainedAgeGenderRaceEstimation_MSUTechReport2014.pdf; LFW website: http://vis-www.cs.umass.edu/lfw/.

[2] Unconstrained face recognition: Identifying a person of interest from a media collection: http://biometrics.cse.msu.edu/Publications/Face/BestRowdenetal_UnconstrainedFaceRecognition_TechReport_MSU-CSE-14-1.pdf

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the instances related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution?
Each instance contains a pair of images that are 250 by 250 pixels in JPEG 2.0 format. Each image is accompanied by a label indicating the name of the person in the image. While subpopulation data was not available at the initial release of the dataset, a subsequent paper[3] reports the distribution of images by age, race and gender. Table 2 lists these results.

Is everything included or does the data rely on external resources? (e.g., websites, tweets, datasets) If external resources, a) are there guarantees that they will exist, and remain constant, over time; b) is there an official archival version; c) are there access restrictions or fees?
Everything is included in the dataset.

Are there recommended data splits and evaluation measures? (e.g., training, development, testing; accuracy or AUC)
The dataset comes with specified train/test splits such that none of the people in the training split are in the test split and vice versa. The data is split into two views, View 1 and View 2. View 1 consists of a training subset (pairsDevTrain.txt) with 1100 pairs of matched and 1100 pairs of mismatched images, and a test subset (pairsDevTest.txt) with 500 pairs of matched and mismatched images. Practitioners can train an algorithm on the training set and test on the test set, repeating as often as necessary. Final performance results should be reported on View 2 which consists of 10 subsets of the dataset. View 2 should only be used to test the performance of the final model. We recommend reporting performance on View 2 by using leave-one-out cross validation, performing 10 experiments. That is, in each experiment, 9 subsets should be used as a training set and the 10^{th} subset should be used for testing. At a minimum, we recommend reporting the **estimated mean accuracy,** $\hat{\mu}$ and the **standard error of the mean:** S_E for View 2.
$\hat{\mu}$ is given by:

$$\hat{\mu} = \frac{\sum_{i=1}^{10} p_i}{10} \quad (1)$$

where p_i is the percentage of correct classifications on View 2 using subset i for testing. S_E is given as:

$$S_E = \frac{\hat{\sigma}}{\sqrt{10}} \quad (2)$$

Where $\hat{\sigma}$ is the estimate of the standard deviation, given by:

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{10} (p_i - \hat{\mu})^2}{9}} \quad (3)$$

The multiple-view approach is used instead of a traditional train/validation/test split in order to maximize the amount of data available for training and testing.

[3] http://biometrics.cse.msu.edu/Publications/Face/HanJain_UnconstrainedAgeGenderRaceEstimation_MSUTechReport2014.pdf

A Database for Studying Face Recognition in Unconstrained Environments — Labeled Faces in the Wild

Training Paradigms: There are two training paradigms that can be used with our dataset. Practitioners should specify the training paradigm they used while reporting results.

- **Image-Restricted Training** This setting prevents the experimenter from using the name associated with each image during training and testing. That is, the only available information is whether or not a pair of images consist of the same person, not who that person is. This means that there would be no simple way of knowing if there are multiple pairs of images in the train/test set that belong to the same person. Such inferences, however, might be made by comparing image similarity/equivalence (rather than comparing names). Thus, to form training pairs of matched and mismatched images for the same person, one can use image equivalence to add images that consist of the same person.

  The files pairsDevTrain.txt and pairsDevTest.txt support image-restricted uses of train/test data. The file pairs.txt in View 2 supports the image-restricted use of training data.

- **Unrestricted Training** In this setting, one can use the names associated with images to form pairs of matched and mismatched images for the same person. The file people.txt in View 2 of the dataset contains subsets of of people along with images for each subset. To use this paradigm, matched and mismatched pairs of images should be formed from images in the same subset. In View 1, the files peopleDevTrain.txt and peopleDevTest.txt can be used to create arbitrary pairs of matched/mismatched images for each person. The unrestricted paradigm should only be used to create training data and not for performance reporting. The test data, which is detailed in the file pairs.txt, should be used to report performance. We recommend that experimenters first use the image-restricted paradigm and move to the unrestricted paradigm if they believe that their algorithm's performance would significantly improve with more training data. While reporting performance, it should be made clear which of these two training paradigms were used for particular test result.

What experiments were initially run on this dataset? Have a summary of those results.
The dataset was originally released without reported experimental results but many experiments have been run on it since then.

Any other comments?
Table 1 summarizes some dataset statistics and Figure 1 shows examples of images. Most images in the dataset are color, a few are black and white.

Property	Value
Database Release Year	2007
Number of Unique Subjects	5649
Number of total images	13,233
Number of individuals with 2 or more images	1680
Number of individuals with single images	4069
Image Size	250 by 250 pixels
Image format	JPEG
Average number of images per person	2.30

Table 1. A summary of dataset statistics extracted from the original paper: Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments.* University of Massachusetts, Amherst, Technical Report 07-49, October, 2007.

Demographic Characteristic	Value
Percentage of female subjects	22.5%
Percentage of male subjects	77.5%
Percentage of White subjects	83.5%
Percentage of Black subjects	8.47%
Percentage of Asian subjects	8.03%
Percentage of people between 0-20 years old	1.57%
Percentage of people between 21-40 years old	31.63%
Percentage of people between 41-60 years old	45.58%
Percentage of people over 61 years old	21.2%

Table 2. Demographic characteristics of the LFW dataset as measured by Han, Hu, and Anil K. Jain. *Age, gender and race estimation from unconstrained face images.* Dept. Comput. Sci. Eng., Michigan State Univ., East Lansing, MI, USA, MSU Tech. Rep.(MSU-CSE-14-5) (2014).

Data Collection Process

How was the data collected? (e.g., hardware apparatus/sensor, manual human curation, software program, software interface/API)
The raw images for this dataset were obtained from the Faces in the Wild database collected by Tamara Berg at Berkeley[4]. The images in this database were gathered from news articles on the web using software to crawl news articles.

Who was involved in the data collection process? (e.g., students, crowdworkers) and how were they compensated (e.g., how much were crowdworkers paid)?
Unknown

Over what time-frame was the data collected? Does the collection time-frame match the creation time-frame of the instances?
Unknown

Gebru et al., 2018

We need standards & laws specifying what can be used where.

And we need mechanisms by which our existing laws are not broken.

traffic-history-detroit/263121077/

Automobile

- No stop signs, drivers licenses, drunk driving laws, seatbelt etc
- Lots of accidents
- Crash tests were done on male dummies
- Studies show that accidents disproportionately affected women

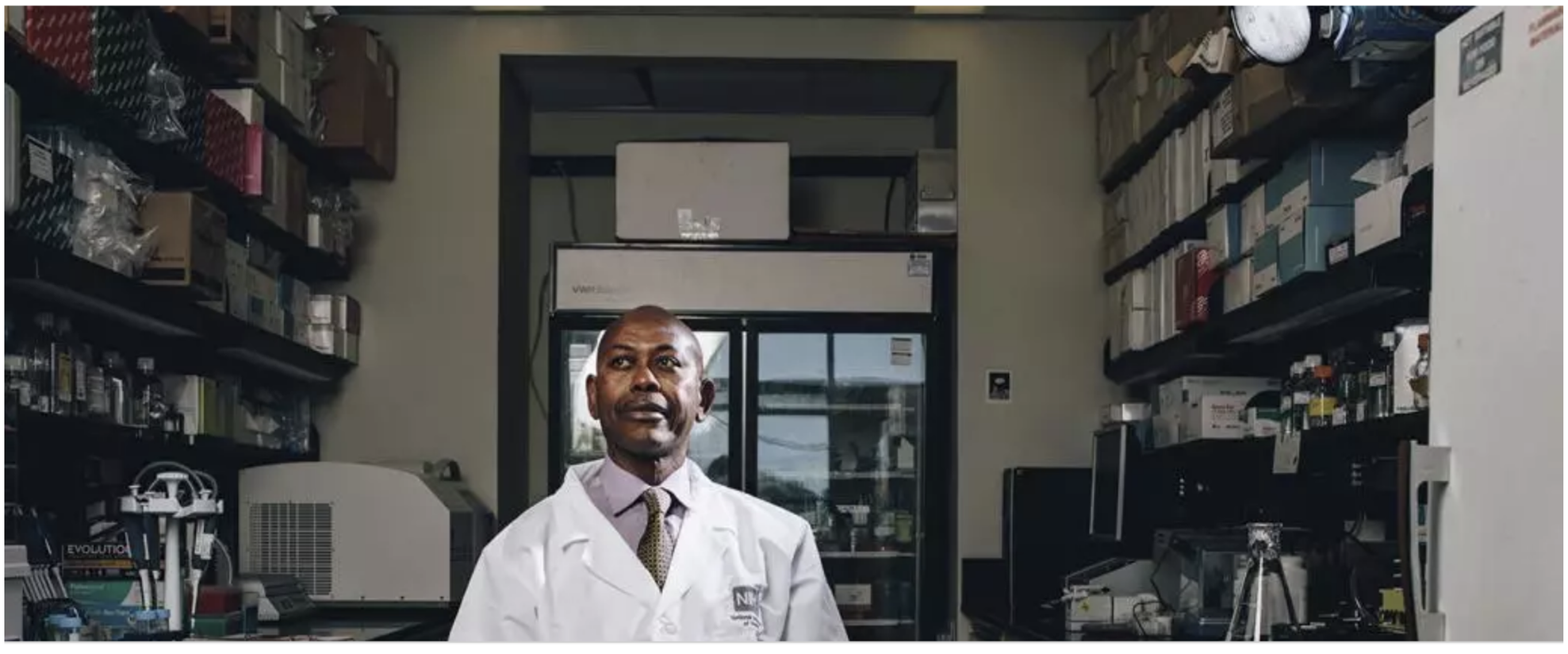
Clinical Trials

- Illegal experimentation on vulnerable populations
- Women were not required to be part of clinical trials until recently
- Study shows that 8-10 drugs that were pulled from circulation between 1997-2001 disproportionately affected women

NEWSWEEK MAGAZINE

Cancer Scientists Have Ignored African DNA in the Search for Cures

BY JESSICA WAPNER ON 7/18/18 AT 9:01 AM



It took many years for standards to be placed and we are still suffering consequences from bias in automobile design, clinical trials

Next Class

- "Revised Report”: significant improvement beyond the "Full Report" by responding to our comments as well as taking your own initiative
- Revised report is due Dec 10, 11pm by email
- Dec 10 class: each team 10 mins ppt + 5mins discussion.

Key Scholars

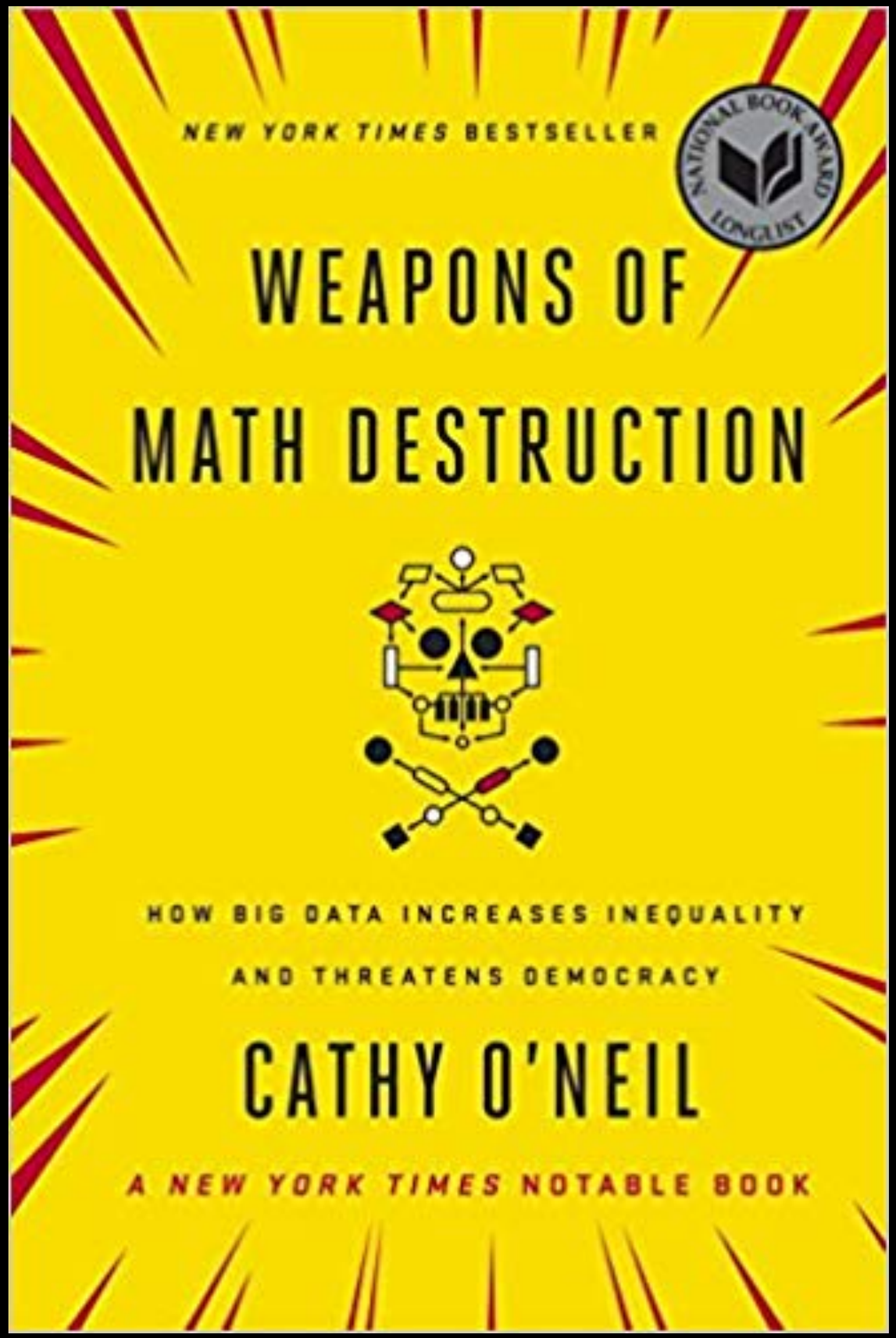
Cynthia Dwork

Aaron Roth

Michael Kearns

Margaret Mitchell

Timnit Gebru


NEW YORK TIMES BESTSELLER
NATIONAL BOOK AWARD LONGLIST
WEAPONS OF
MATH DESTRUCTION
HOW BIG DATA INCREASES INEQUALITY
AND THREATENS DEMOCRACY
CATHY O'NEIL
A NEW YORK TIMES NOTABLE BOOK


Copyrighted Material
michael kearns + aaron roth
the ethical algorithm
the science of
socially aware algorithm design
Copyrighted Material

MIT Subjects Related to ML

- 6.036 Introduction to Machine Learning
- 6.862 Applied Machine Learning (6.862 = 6.036 + Term Project)
- 6.867 Machine Learning
- 9.520/6.860: Statistical Learning Theory and Applications
- 9.S914: Mathematical Statistics: A Non-Asymptotic Approach
- 6.883: Online Methods in Machine Learning: Theory and Applications

MIT Subjects Related to ML

- 4.S42 Machine Learning for Creative Design
- 6.S897/HST.S53: Machine Learning for Healthcare
- MAS.533 AI for Impact ~ Towards Health & Sustainability from People to Planet
- 6.268 Network Science and Models
- 11.s938/11.s196 Deep Learning for Urban Transportation