Advanced Machine Learning For Design

Lecture 7: Train, Evaluate and Integrate Machine Learning Models (part 2)

Module 3

Evangelos Niforatos

01/11/2023

aml4d-ide@tudelft.nl
https://aml4design.github.io/

Admin

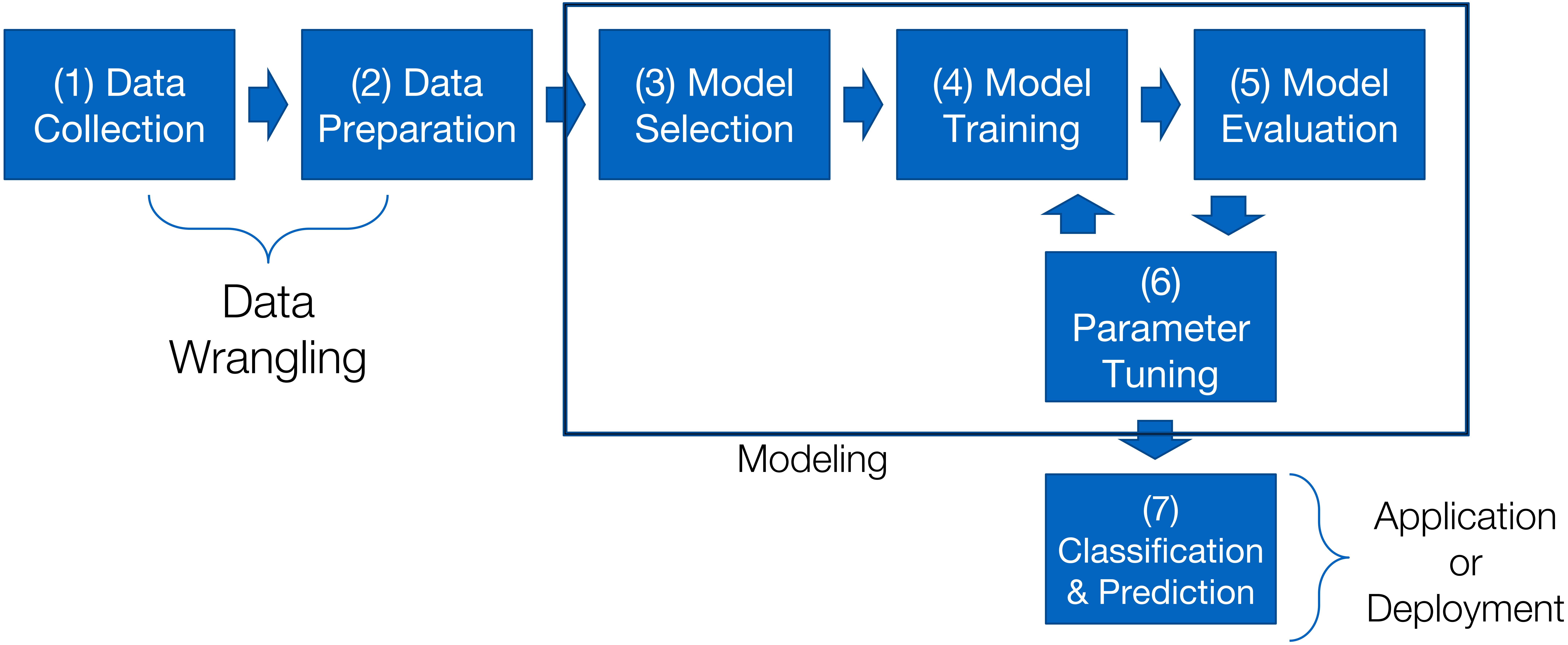
Remarks

- Group Assignment #2 has been graded & Average Peer Assessment #2 scores have been posted
- Group Assignment #3 deadline: Tuesday, Nov. 7, 23:59 followed by Peer Assessment #3 (do not miss the deadline)
- Last tutorial on Friday, Nov. 4:
  - Help with Group Assignment #3
  - Demo and survey of the COALA LLM-Powered Assistant
  - Example exam
- Final Exam: Friday, Nov. 10, 13:30—15:00 at 3Me-Hall J (34.D-1-300) (1.5h instead of 3h)

  NO notes, books, laptops, smartphones, smartwatches, smart-glasses allowed---bring a calculator
  - N=38 registered (so far)
  - Forgot to register? See: https://www.tudelft.nl/en/student/education/courses-and-examinations/examinations/registration-for-exams
  - Final Portfolio (deadline: Friday, Nov. 17, 23:59):
  - A concise report of Deliverables 1, 2, & 3 incorporating our feedback
  - Final Group project grade considering any improvements made
  - Followed by Peer Assessment #4 (do not miss the deadline please)

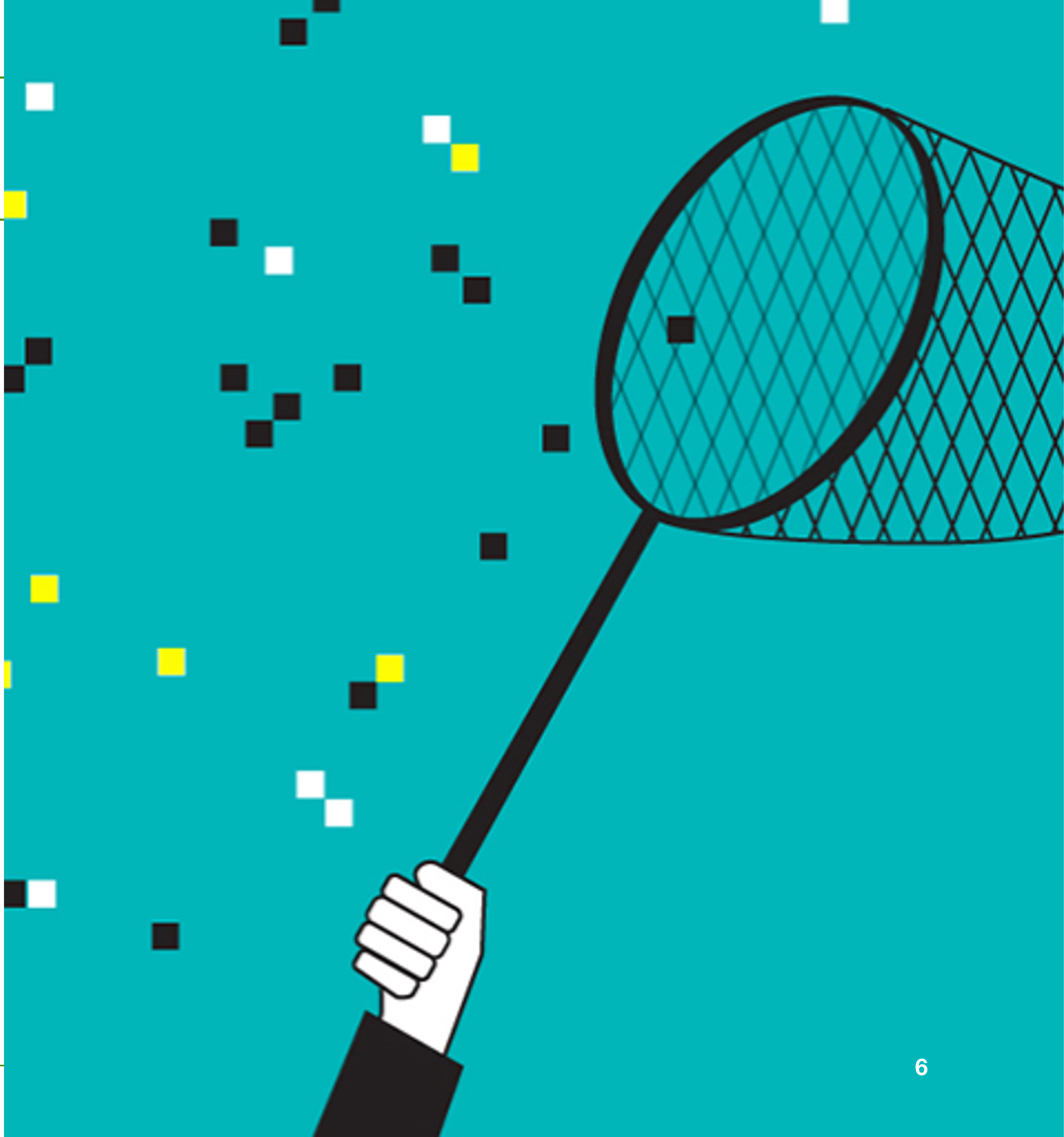
Previously, on AML4D….

Abstract ML Pipeline: A 7-step Process



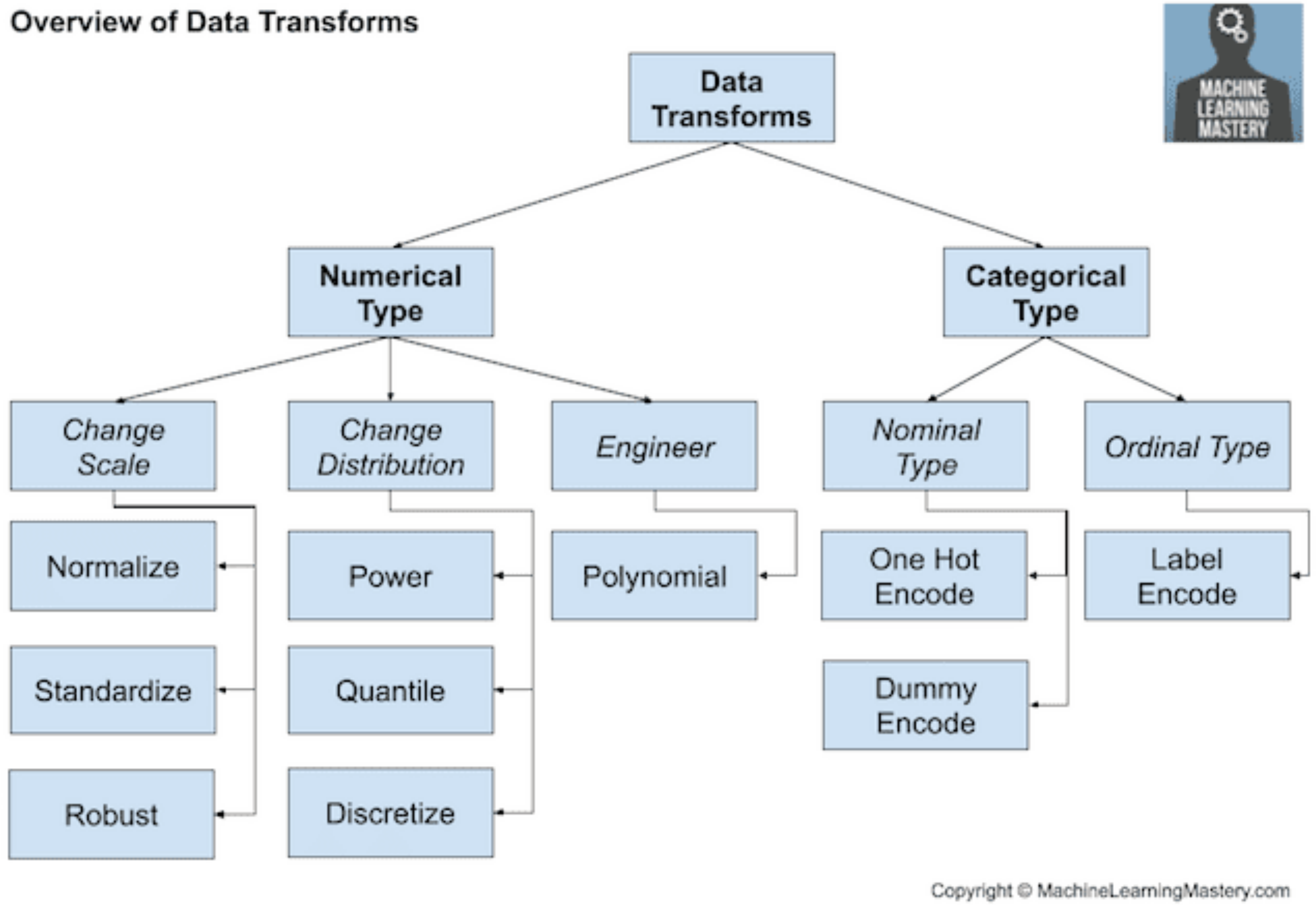
(1) Data Collection

- Manual (rarely)
- Automated
  - Existing collections (e.g., data.worldbank.org)
  - Scripts (e.g., web crawlers)
  - Sensors (e.g., weather stations)
  - Application Programming Interfaces (APIs)
  - …
- Semi-automated
  - Crowdsourcing (e.g., google maps)
  - …



(2) Data Preparation

- Data randomization/shuffling
- Data labeling/annotation
- Data visualization for detecting any relationships among variables
  - We make sure that all classes are equally represented (if we can)
  - Additional actions (data normalization, error correction, etc.)
- Data splitting
  - e.g., Dataset = Training (80%) + Evaluation (20%)



See: https://towardsdatascience.com/encoding-categorical-variables-one-hot-vs-dummy-encoding-6d5b9c46e2db

(3) Model Selection

- Selecting the right model (algorithm) is crucial
- Depends on
  - our dataset
    - images, timeseries, numeric or text data
  - the use case
    - (classification vs. prediction vs. clustering)

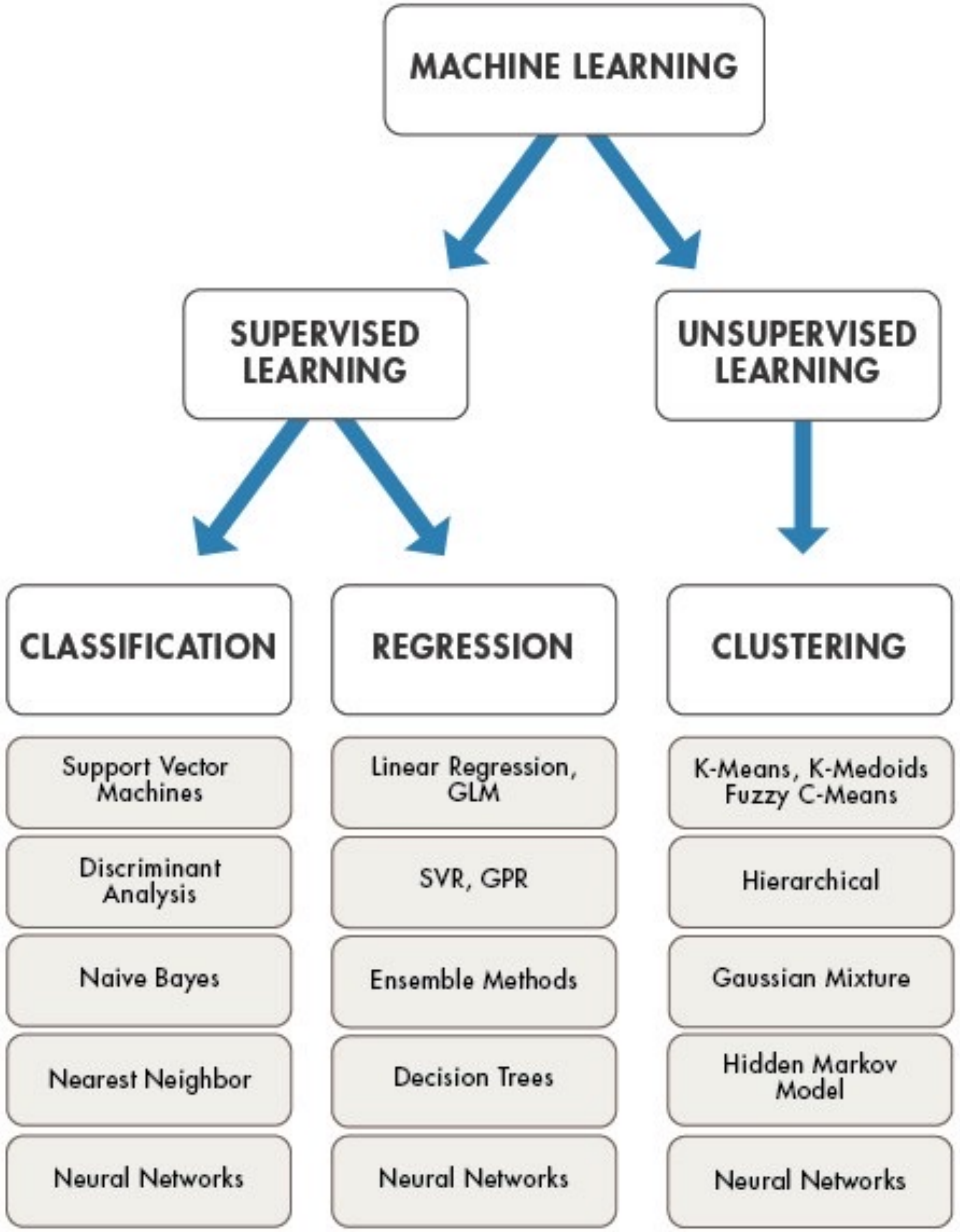
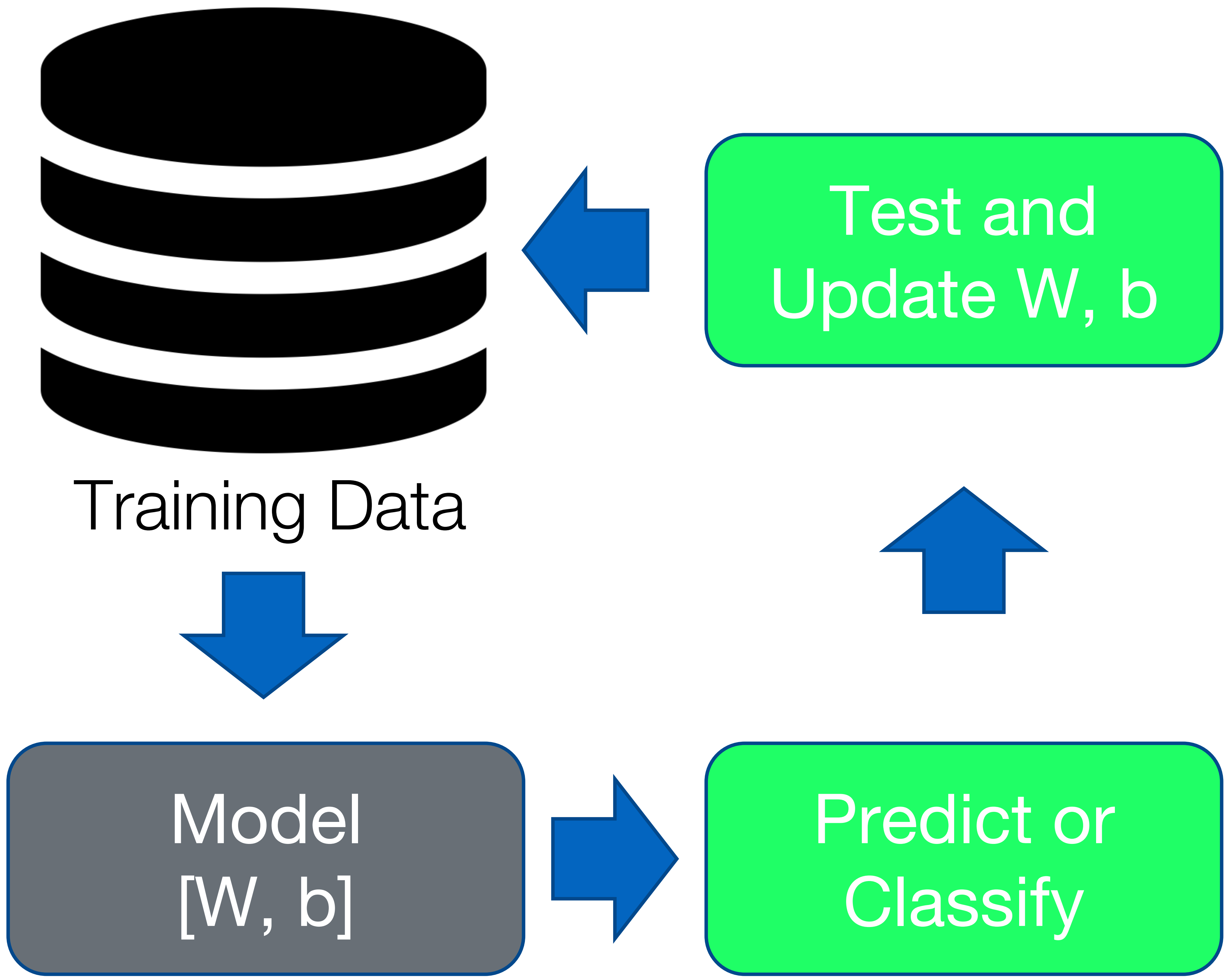


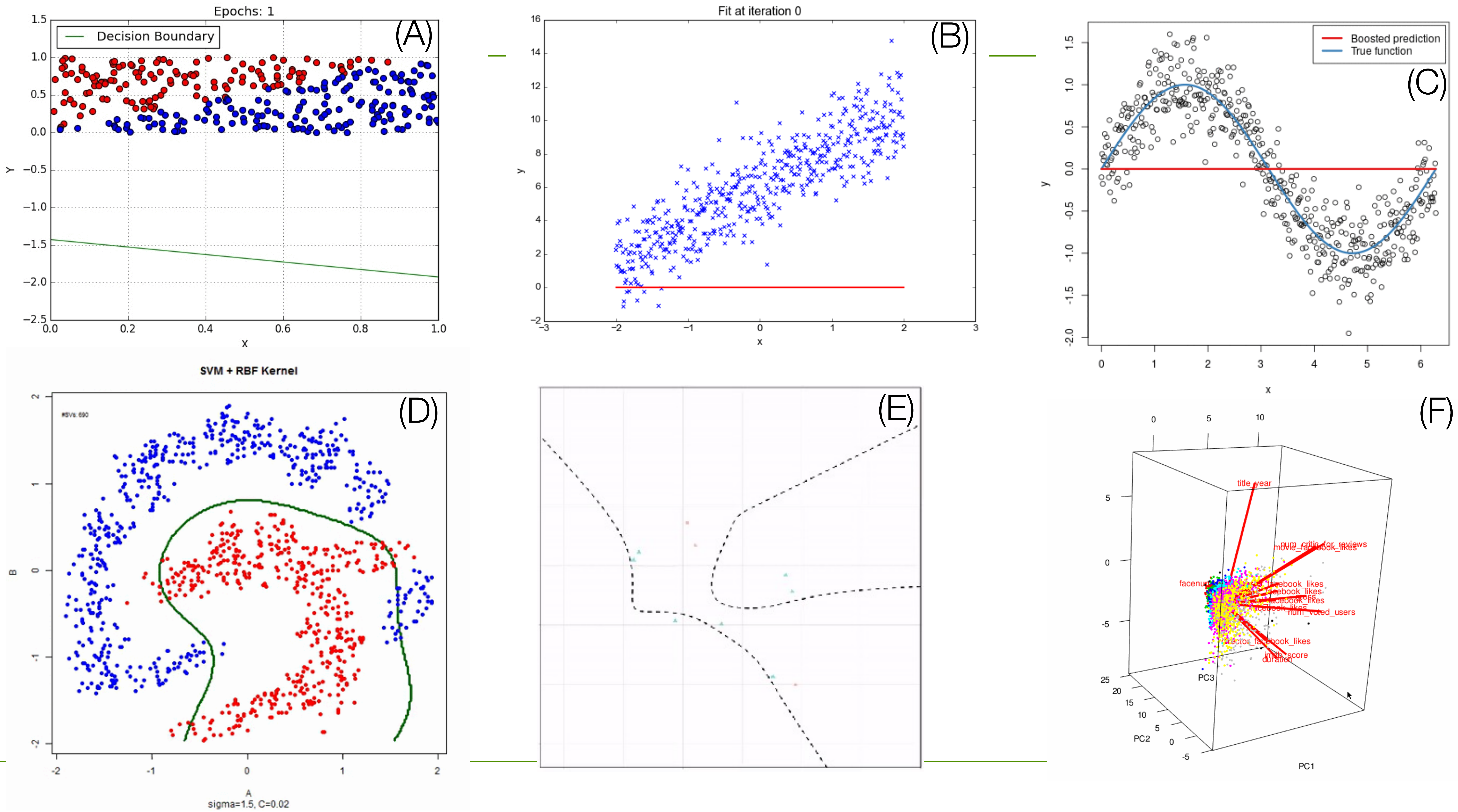
Image by https://medium.com/technology-nineleaps/popular-machine-learning-algorithms-a574e3835ebb

(4) Model Training

- We use our data to incrementally improve the ability of our model to **predict** or **classify**
- e.g., y = w*x + b
  - y: output
  - w: slope (weight)
  - x: input
  - b: intercept (bias)
- Model[W, b]→Predict or Classify

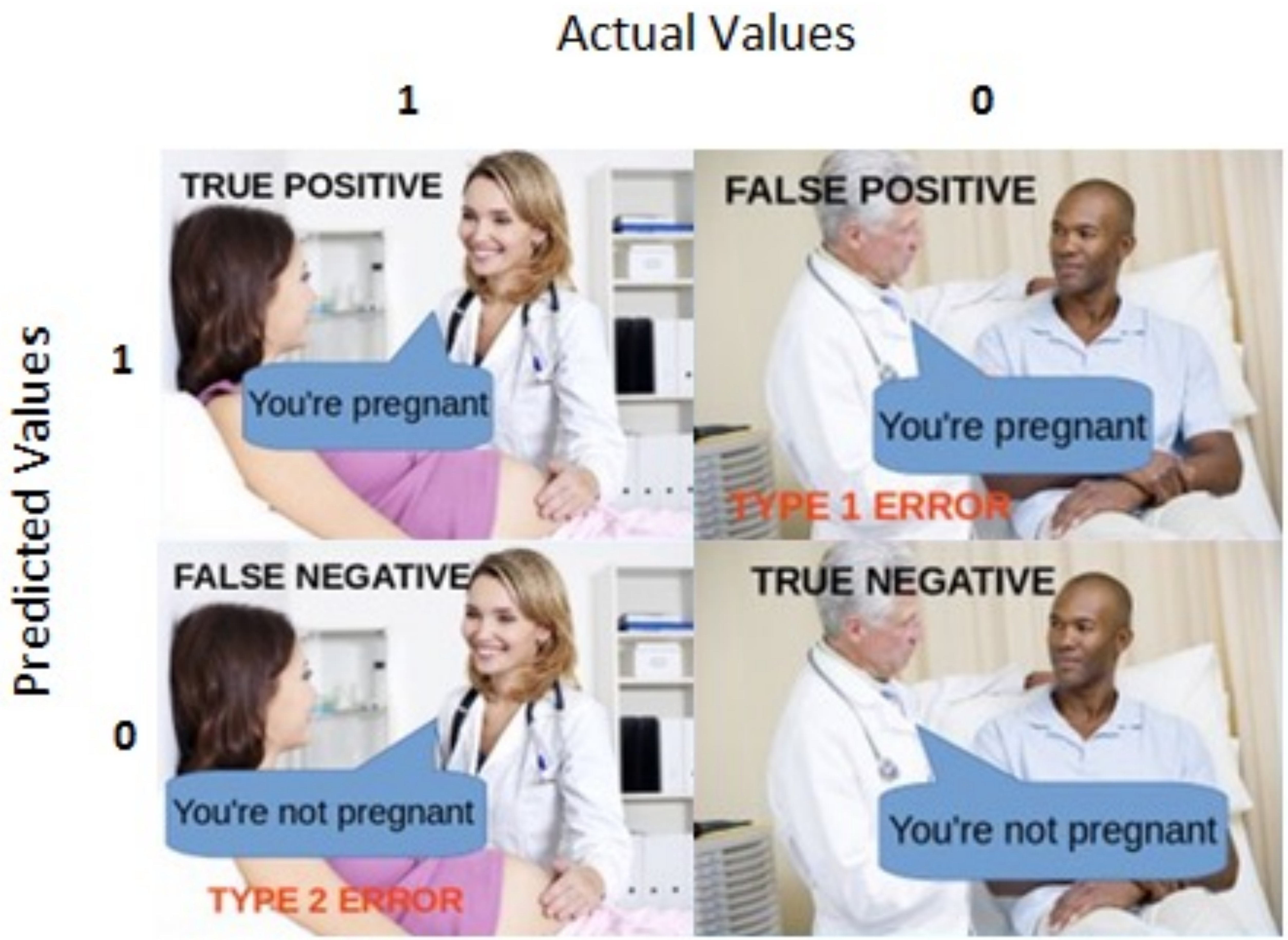


(4) Model Training: An iterative process



(5) Model Evaluation: General Metrics

- Definitions
  - True Positive (TP): Correctly predicted instances of classes
  - True Negative (TN): Correctly predicted instances of non-classes
  - False Positive (FP): Incorrectly predicted instances of classes
  - False Negative (FN): Incorrectly predicted instances of non-classes
- Metrics
  - Accuracy = (TP+TN) / (TP+TN+FP+FN)
  - Precision = TP / (TP+FP)
  - Recall (Sensitivity) = TP / (TP+FN)
  - Specificity = TN / (TN+FP)
  - F1 Score = 2*(Recall * Precision) / (Recall + Precision)



ML in society

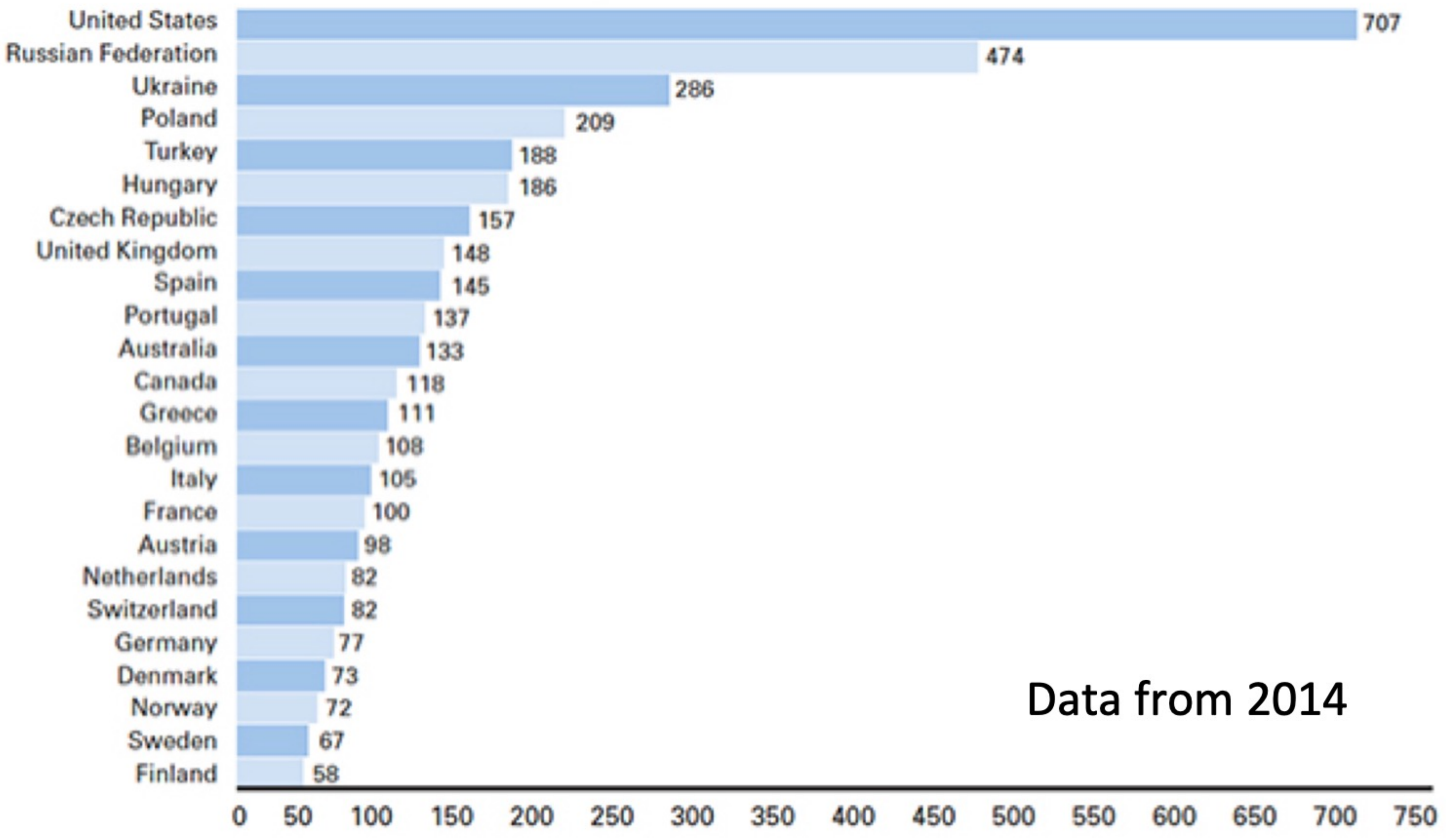
ML algorithms are now pervasive in society

- Widespread algorithms with many small interactions
  - e.g., search engines, recommendation systems, in-camera face recognition
- Specialized algorithms with fewer but higher-stakes interactions
  - personalized medicine, automated stock trading, criminal justice
- At this level of impact, ML systems can have unintended social consequences
  - **Low classification/prediction error is not enough**

Case Study: ML for Recidivism Prediction

- Background on US Prison Population

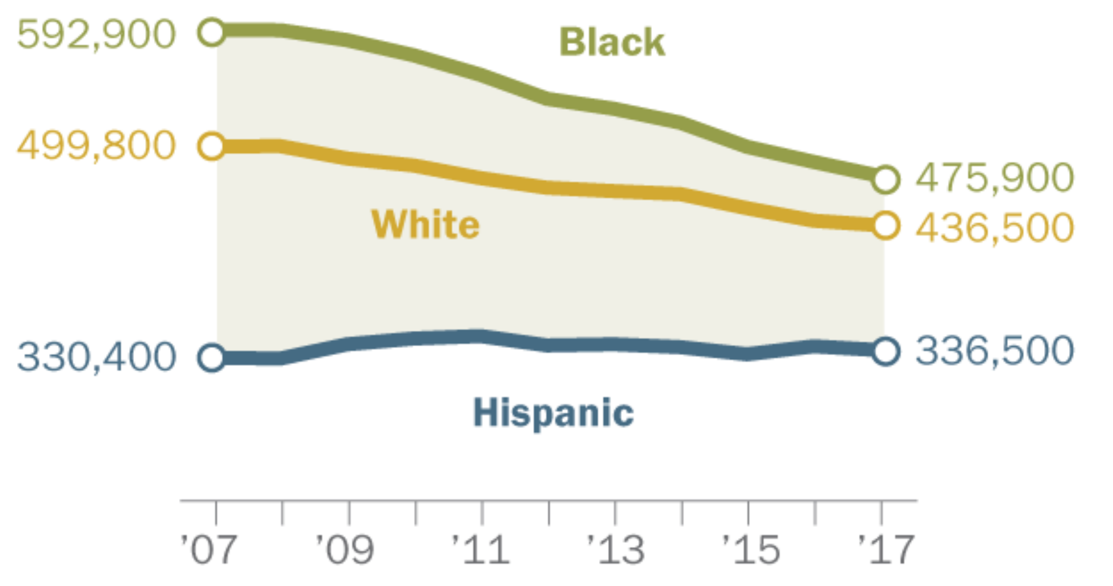
Incarceration Rates per 100,000



Source: https://www.apa.org/monitor/2014/10/incarceration

Racial and ethnic gaps shrink in U.S. prison population

Sentenced federal and state prisoners by race and Hispanic origin, 2007-2017

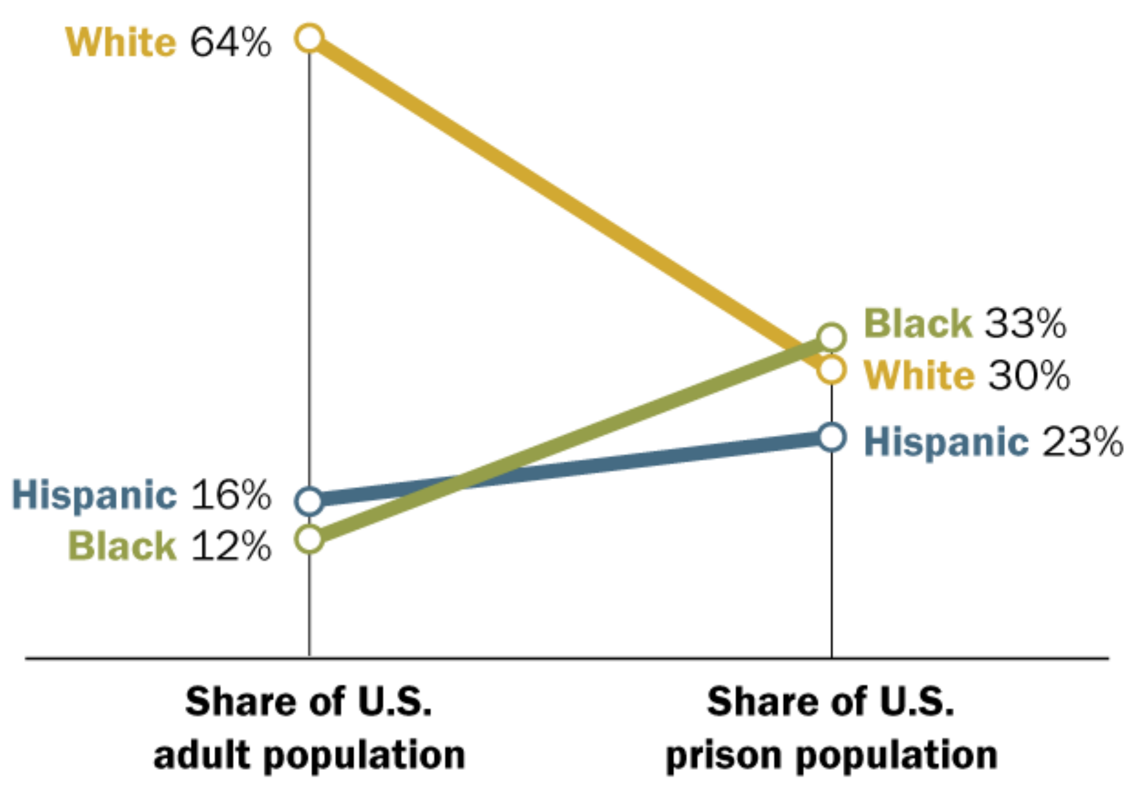


Note: Whites and blacks include those who report being only one race and are non-Hispanic. Hispanics are of any race. Prison population is defined as inmates sentenced to more than a year in federal or state prison.
Source: Bureau of Justice Statistics.

PEW RESEARCH CENTER

Blacks, Hispanics make up larger shares of prisoners than of U.S. population

U.S. adult population and U.S. prison population by race and Hispanic origin, 2017



Note: Whites and blacks include those who report being only one race and are non-Hispanic. Hispanics are of any race. Prison population is defined as inmates sentenced to more than a year in federal or state prison.
Source: U.S. Census Bureau, Bureau of Justice Statistics.

PEW RESEARCH CENTER

https://www.pewresearch.org/fact-tank/2019/04/30/shrinking-gap-between-number-of-blacks-and-whites-in-prison/

COMPAS

- Software by Northpointe that predicts recidivism

- Used by judges in determining sentencing and bail

- Scores derived from 137 questions answered by defendants or pulled from criminal records:
  - *“Was one of your parents ever sent to jail or prison?”*
  - *“How many of your friends/acquaintances are taking drugs illegally?”*
  - *“How often did you get in fights while at school?”*
  - Agree or disagree? *“A hungry person has a right to steal”*
  - Agree or disagree? *“If people make me angry or lose my temper, I can be dangerous.”*
  - Race is **not** one of the questions

- The exact method of determining the score is kept as a **trade secret**

COMPAS

- ProPublica Analysis of COMPAS Algorithm (2016)

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

- African Americans are almost twice as likely as Caucasians to be incorrectly labeled as high risk

- Subsequent study (2018): COMPAS is no more accurate (65%) than predictions made by people with little/no criminal justice expertise (63% individually, 67% pooled)
  - J. Dressel and H. Farid. (2018). "The accuracy, fairness, and limits of predicting recidivism." Science Advances 4(1). doi:10.1126/sciadv.aao5580

ML Predictions can have real consequences

A

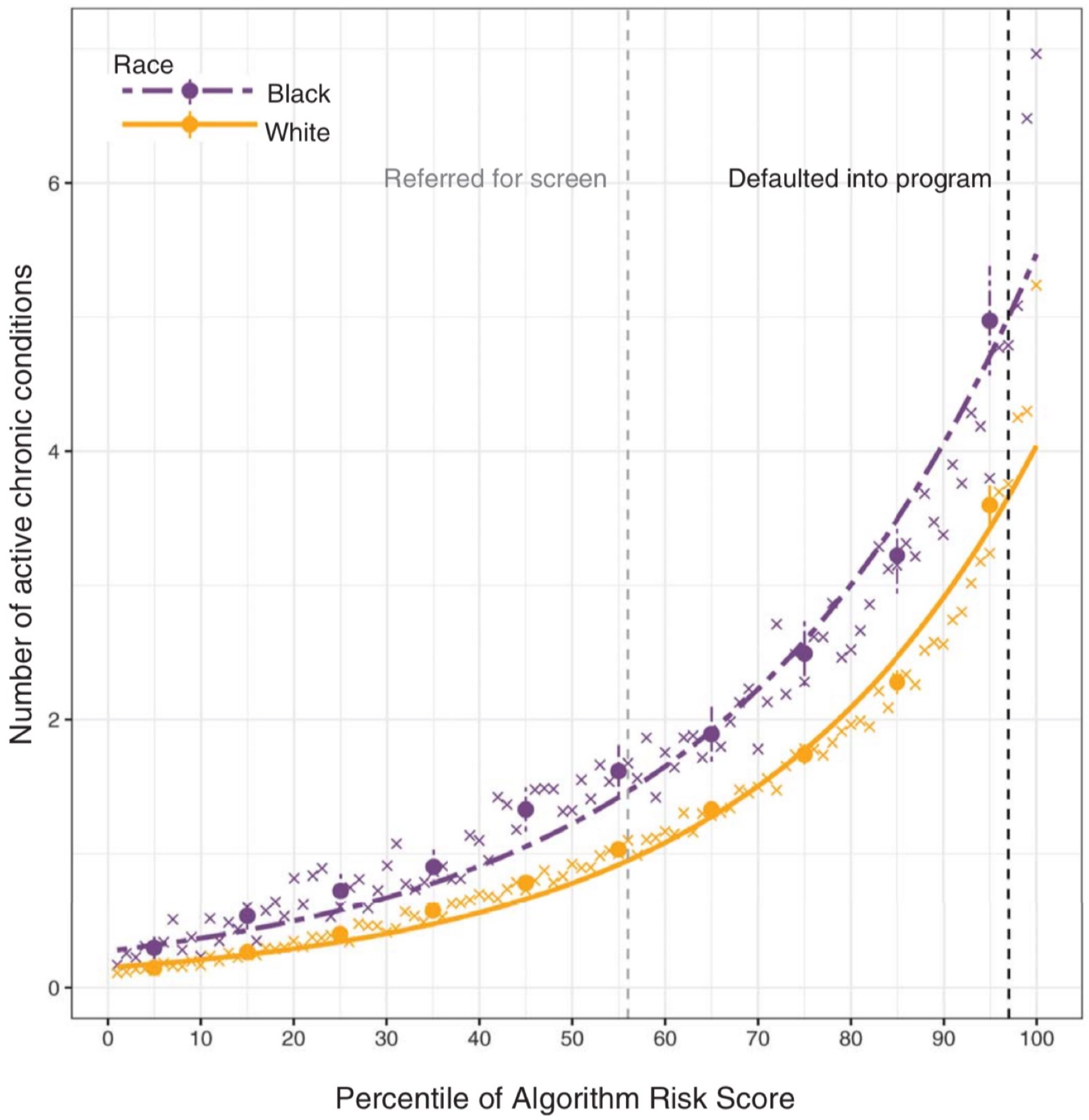


Fig. 1. Number of chronic illnesses versus algorithm-predicted risk, by race. (**A**) Mean number of chronic conditions by race, plotted against

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453. https://doi.org/10.1126/science.aax2342

Case Study: Drug Discovery

nature machine intelligence

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature > nature machine intelligence > comment > article

Comment | Published: 07 March 2022

Dual use of artificial-intelligence-powered drug discovery

Fabio Urbina, Filippa Lentzos, Cédric Invernizzi & Sean Ekins

Nature Machine Intelligence **4**, 189–191 (2022) | Cite this article

An international security conference explored how artificial intelligence (AI) technologies for drug discovery could be misused for de novo design of biochemical weapons. A thought experiment evolved into a computational proof.

The thought had never previously struck us. We were vaguely aware of security concerns around work with pathogens or toxic chemicals, but that did not relate to us; we primarily operate in a virtual setting. Our work is rooted in building machine learning models for therapeutic and toxic targets to better assist in the design of new molecules for drug discovery. We have spent decades using computers and AI to improve human health—not to degrade it. We were naive in thinking about the potential misuse of our trade, as our aim had always been to avoid molecular features that could interfere with the many different classes of proteins essential to human life. Even our projects on Ebola and neurotoxins, which could have sparked thoughts about the potential negative implications of our machine learning models, had not set our alarm bells ringing.

In less than 6 hours after starting on our in-house server, our model generated 40,000 molecules that scored within our desired threshold. In the process, the AI designed not only VX, but also many other known chemical warfare agents that we identified through visual confirmation with structures in public chemistry databases. Many new molecules were also designed that looked equally plausible. These new molecules were predicted to be more toxic, based on the predicted LD_{50} values, than publicly known chemical warfare agents (Fig. 1). This was unexpected because the datasets we used for training the AI did not include these nerve agents. The virtual molecules even occupied a region of molecular property space that was entirely separate from the many thousands of molecules in the organism-specific LD_{50} model, which comprises mainly pesticides, environmental toxins and drugs (Fig. 1). By inverting the use of our machine learning models, we had transformed our innocuous generative model from a helpful tool of medicine to a generator of likely deadly molecules.

https://www.nature.com/articles/s42256-022-00465-9

Regulated Domains in the USA

- **Credit** (Equal Credit Opportunity Act)
- **Education** (Civil Rights Act of 1964; Education Amendments of 1972)
- **Employment** (Civil Rights Act of 1964)
- **Housing** (Fair Housing Act)
- **Public Accommodation** (Civil Rights Act of 1964)

- The regulations extend to marketing and advertising; they are not limited to final decisions
- This list ignores the complex web of laws that regulates the government

Situation in EU is similar

The EU *Artificial Intelligence Act* attempts to regulate AI

Technology rarely, if ever, “just works”

- Who is(n't) this technology built for?
  - Who is asking?
  - What are they seeking to optimize?
  - Why are they trying to optimize it?
- Data
  - How was it collected?
  - Was this influenced by the algorithm?
  - By the person who asked the question?
  - Does it really measure what it claims to?
- Evaluation
  - Do I believe the evaluation (e.g. precision/recall)
  - Are they checking for the right things?

Sources of bias in machine learning

A BIASED WORLD → DATA GENERATION → LEARNING → EVALUATION

A BIASED WORLD

Historical bias:

- *Co-occurrence bias*
- Framing bias
- Epistemological bias
- Language bias
- Coverage bias

...

DATA GENERATION

- *Specification bias*
- Annotation bias
- Measurement bias
- Sampling bias
- *Inherited bias*

...

LEARNING

- Inductive bias
- *Hyper-parameter bias*
- Uncertainty bias

...

EVALUATION

Model bias:

- Overall mis-classification rate
- False positive rate
- False negative rate
- False omission rate
- False discovery rate
- Equalized odds
- Calibration
- Demographic Parity
- Individual Fairness

\+ Causal versions

...

http://ceur-ws.org/Vol-2659/hellstrom.pdf

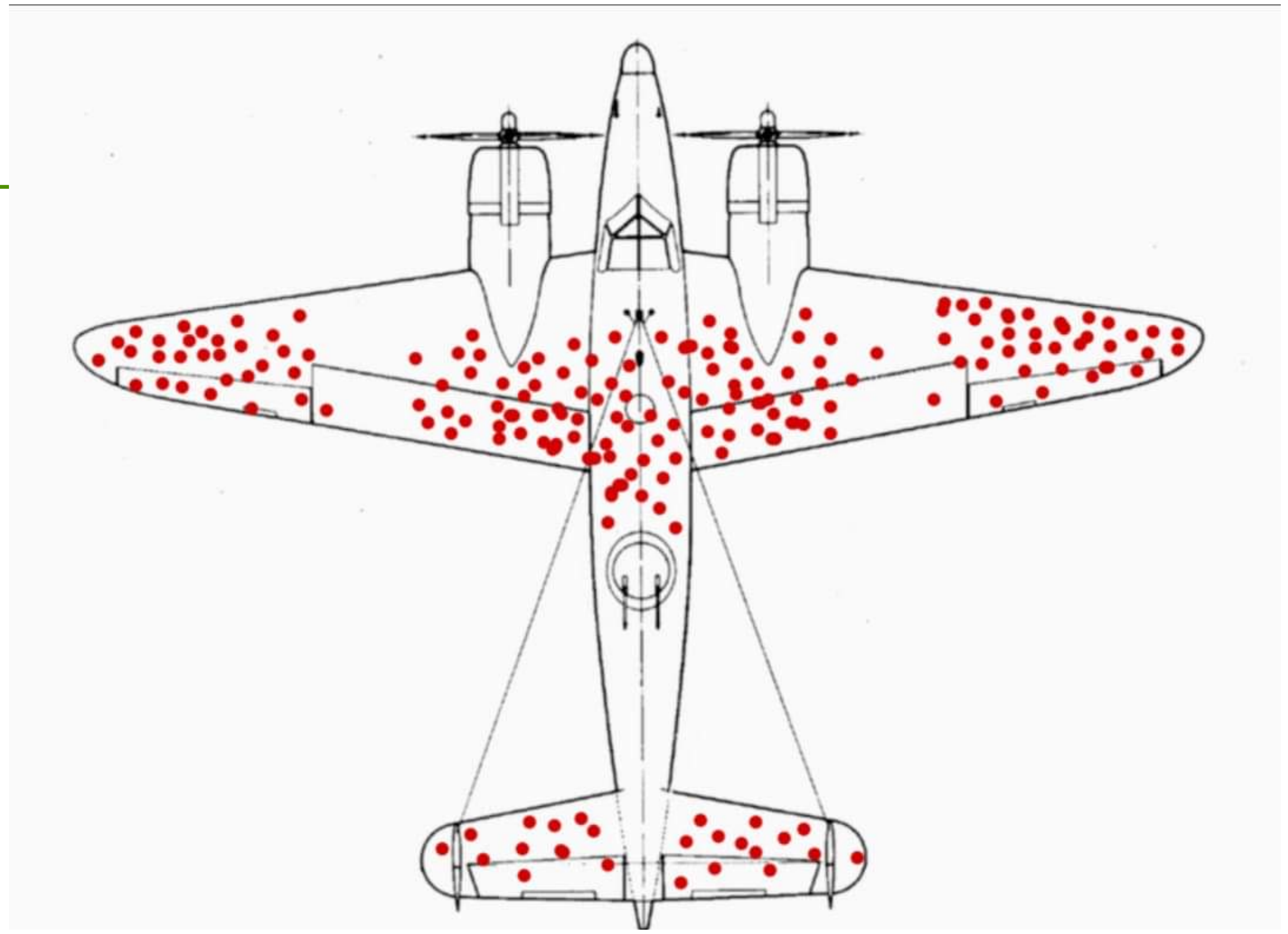
Designing Machine Learning Solutions

- **Training Data**
- **(Expected) Performance**
- Transparency and Explainability
- **Human-AI Interaction**
- Privacy
- Trust

Training Data

Training Data

- Machine learning requires careful preparation of lots of data
- What data does my algorithm need to do its job?
- Do I have **good** data?
  - Error free
- Do I have the **right** data?
  - Fair, representative, unbiased
  - Dataset biases can be based on:
    - historical trends, data gathering methods, biased labelers, etc.
  - Models trained on these data sets will perpetuate the bias(es)



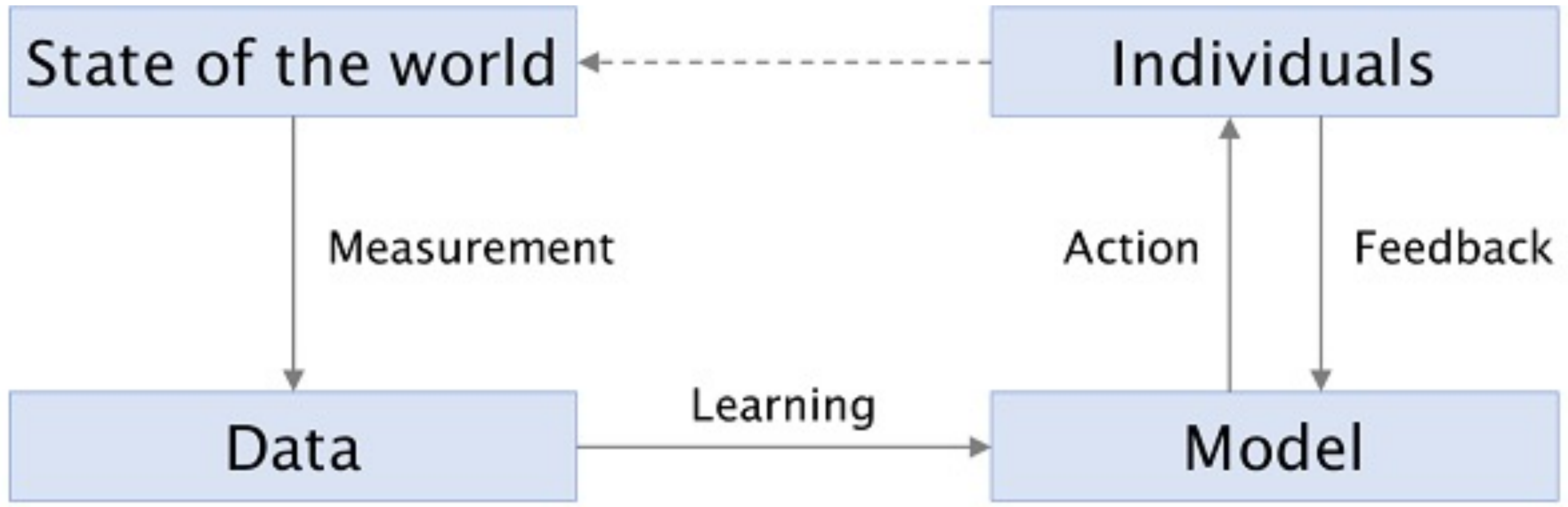
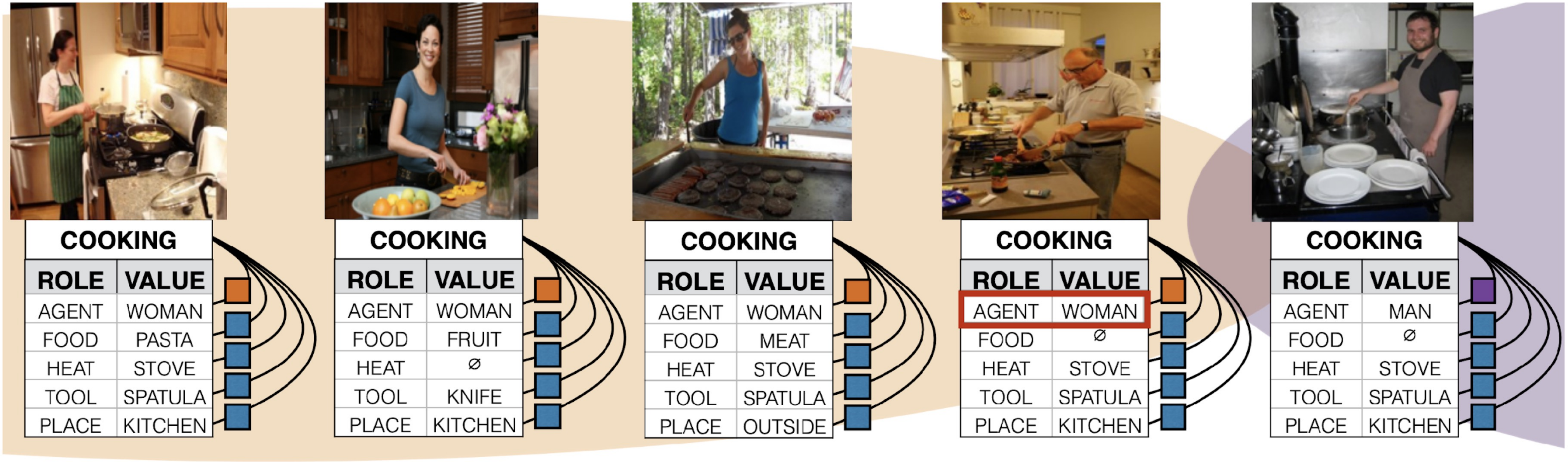


Table 6.1: Most Biased Descriptive Words in 175B Model

Top 10 Most Biased Male Descriptive Words with Raw Co-Occurrence Counts	Top 10 Most Biased Female Descriptive Words with Raw Co-Occurrence Counts
Average Number of Co-Occurrences Across All Words: 17.5	Average Number of Co-Occurrences Across All Words: 23.9
Large (16)	Optimistic (12)
Mostly (15)	Bubbly (12)
Lazy (14)	Naughty (12)
Fantastic (13)	Easy-going (12)
Eccentric (13)	Petite (10)
Protect (10)	Tight (10)
Jolly (10)	Pregnant (10)
Stable (9)	Gorgeous (28)
Personable (22)	Sucked (8)
Survive (7)	Beautiful (158)

Image Credits: https://www.arthur.ai/

Example: Bias in Image Classification

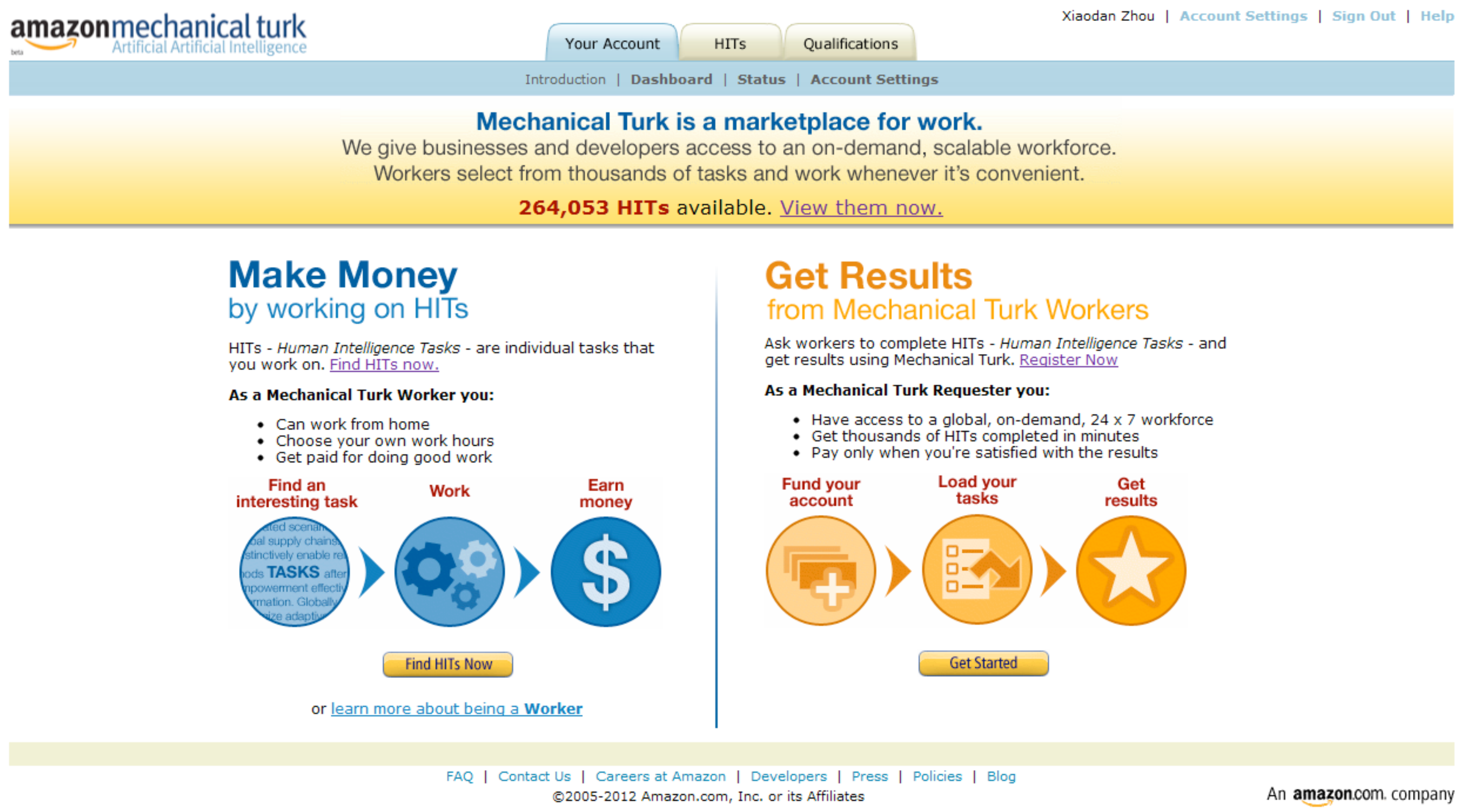


- Images from imSitu visual semantic role labeling (vSRL) dataset
  - 33% of cooking images are of men
  - Prediction with a (biased) conditional random field only predicts men in 16% of cooking images

Data annotation

Opportunistic

Microwork Platforms



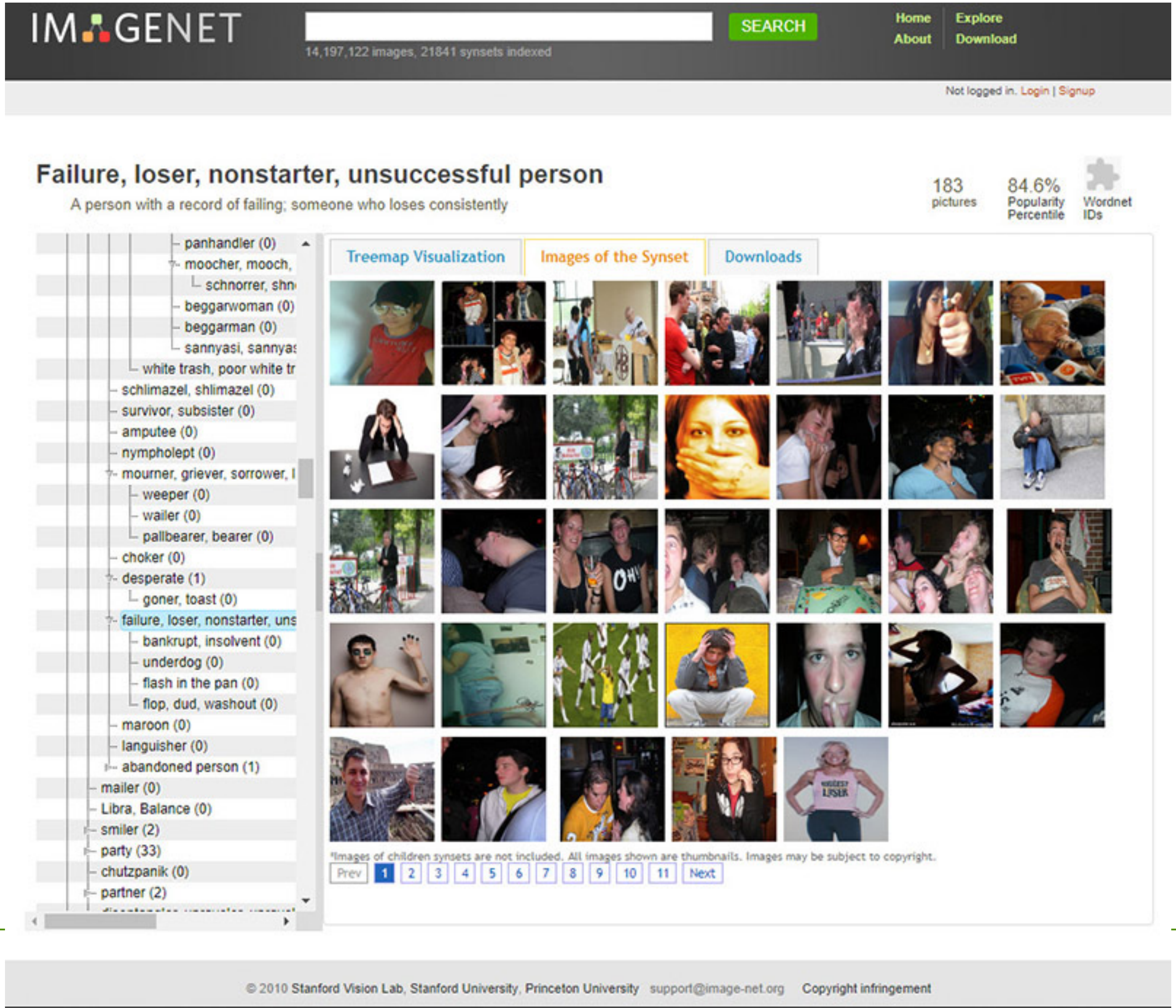
Professional

https://apicciano.commons.gc.cuny.edu/2018/11/26/data-farms-driving-chinas-artificial-intelligence-development/

Excavating AI

The Politics of Images in Machine Learning Training Sets

By Kate Crawford and Trevor Paglen

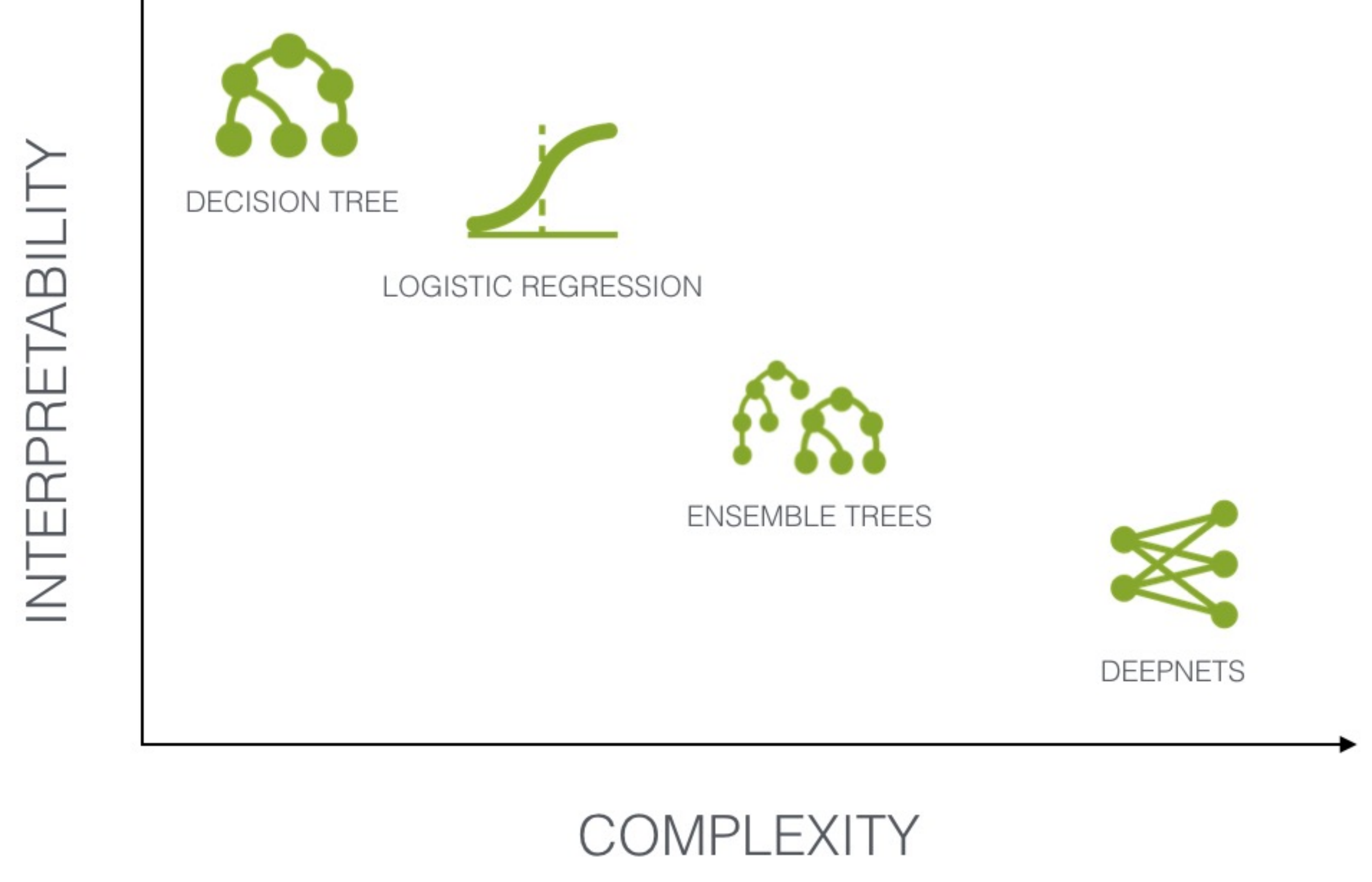


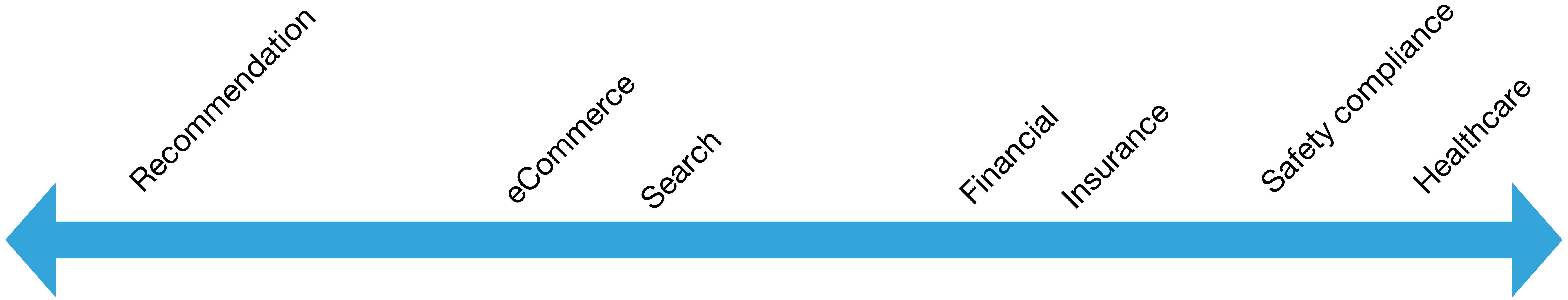
Expected performance

(Expected) Performance

- Am I using the right model?
    - The more complex the machine learning model, the harder it can be to understand
    - Overfitting
- Expectation Management
- Under/Over-estimation of performance

https://blog.bigml.com/2018/05/01/prediction-explanation-adding-transparency-to-machine-learning/





< 60% Accuracy is acceptable

60-80% Accuracy is acceptable

90-99% Accuracy (or **more**)

Fairness

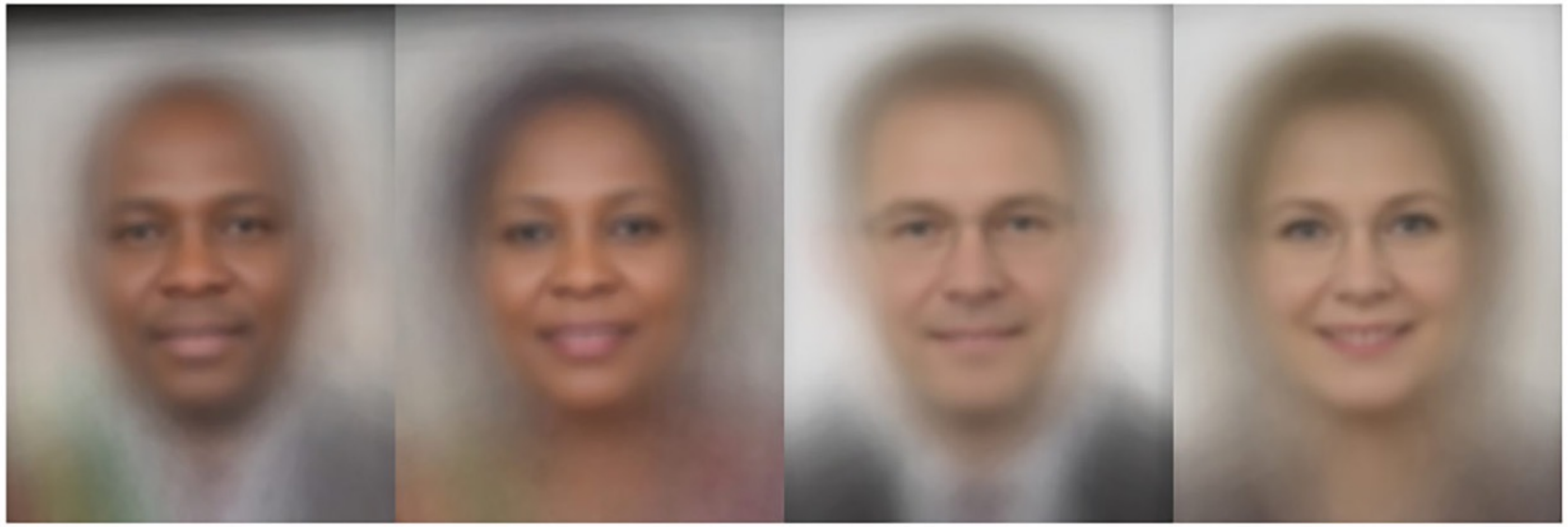
A desirable property of algorithms to avoid bias

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

Joy Buolamwini
MIT Media Lab 75 Amherst St. Cambridge, MA 02139

Timnit Gebru
Microsoft Research 641 Avenue of the Americas, New York, NY 10011

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



Why fairness is hard?

- Suppose we are a bank trying to fairly decide who should get a loan
  - i.e., Who is most likely to pay us back?
- Suppose we have two groups: A and B (the sensitive attribute)
  - This is where discrimination could occur
- The simplest approach is to remove the sensitive attribute from the data, so that our classifier doesn't know the sensitive attribute

Age	Gender	Employed?	Zip Code	Requested Amount	~~A or B?~~	Grant Loan?
37	F	Yes	24729	$50,000	~~A~~	Yes
23	M	Yes	11038	$30,000	~~B~~	Yes
72	F	No	10038	$90,000	~~A~~	Yes
39	F	Yes	30499	$70,000	~~A~~	No
45	M	No	20199	$60,000	~~B~~	No
68	M	Yes	30029	$50,000	~~B~~	No

Legally Recognized “Protected classes” (US)

- Race (Civil Rights Act of 1964)
- Color (Civil Rights Act of 1964)
- Sex (Equal Pay Act of 1963; Civil Rights Act of 1964)
- Religion (Civil Rights Act of 1964)
- National origin (Civil Rights Act of 1964)
- Citizenship (Immigration Reform and Control Act)
- Age (Age Discrimination in Employment Act of 1967)
- Pregnancy (Pregnancy Discrimination Act)
- Familial status (Civil Rights Act of 1968)
- Disability status (Rehabilitation Act of 1973; Americans with Disabilities Act of 1990)
- Veteran status (Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act)
- Genetic information (Genetic Information Nondiscrimination Act)

Why fairness is hard?

Age	Gender	Employed?	Zip Code	Requested Amount	A or B?	Grant Loan?
37	F	Yes	24729	$50,000	?	Yes
23	M	Yes	11038	$30,000	?	Yes
72	F	No	10038	$90,000	?	Yes
39	F	Yes	30499	$70,000	?	No
45	M	No	20199	$60,000	?	No
68	M	Yes	30029	$50,000	?	No

- Just deleting the sensitive attribute won't work if it is correlated with others
  - e.g., it is easy to predict race given other info (home address, financials, etc.)
- We need more sophisticated approaches…

21 types of fairness (and counting)

	Definition	Paper	Citation #	Result
3.1.1	Group fairness or statistical parity	[12]	208	×
3.1.2	Conditional statistical parity	[11]	29	✓
3.2.1	Predictive parity	[10]	57	✓
3.2.2	False positive error rate balance	[10]	57	×
3.2.3	False negative error rate balance	[10]	57	✓
3.2.4	Equalised odds	[14]	106	×
3.2.5	Conditional use accuracy equality	[8]	18	×
3.2.6	Overall accuracy equality	[8]	18	✓
3.2.7	Treatment equality	[8]	18	×
3.3.1	Test-fairness or calibration	[10]	57	✓✗
3.3.2	Well calibration	[16]	81	✓✗
3.3.3	Balance for positive class	[16]	81	✓
3.3.4	Balance for negative class	[16]	81	×
4.1	Causal discrimination	[13]	1	×
4.2	Fairness through unawareness	[17]	14	✓
4.3	Fairness through awareness	[12]	208	×
5.1	Counterfactual fairness	[17]	14	–
5.2	No unresolved discrimination	[15]	14	–
5.3	No proxy discrimination	[15]	14	–
5.4	Fair inference	[19]	6	–

Table 1: Considered Definitions of Fairness

- GOAL: mathematically certify that an algorithm does not suffer from disparate treatment or disparate impact

Types of Fairness: Group Fairness

- Key idea: “Treat different groups equally”

- Assess fairness based on **demographic parity**: require that the same percentage of groups A and B receive loans
  - What if 80% of A is likely to repay, but only 60% of B is?

- Could require equal false positive/negative rates
  - When we make an error, the direction of that error is equally likely for both groups
    - `P(loan | no repay, A) = P(loan | no repay, B)`
    - `P(no loan | would repay, A) = P(no loan | would repay, B)`

Then demographic parity is too strong

Types of Fairness: Individual Fairness

- Key idea: “Treat similar examples similarly”

- Learn fair representations
  - Useful for classification, not for (unfair) discrimination
  - Related to domain adaptation
  - Generative modelling/adversarial approaches

21 types of fairness (and counting)

	Definition	Paper	Citation #	Result
3.1.1	Group fairness or statistical parity	[12]	208	×
3.1.2	Conditional statistical parity	[11]	29	✓
3.2.1	Predictive parity	[10]	57	✓
3.2.2	False positive error rate balance	[10]	57	×
3.2.3	False negative error rate balance	[10]	57	✓
3.2.4	Equalised odds	[14]	106	×
3.2.5	Conditional use accuracy equality	[8]	18	×
3.2.6	Overall accuracy equality	[8]	18	✓
3.2.7	Treatment equality	[8]	18	×
3.3.1	Test-fairness or calibration	[10]	57	✓✗
3.3.2	Well calibration	[16]	81	✓✗
3.3.3	Balance for positive class	[16]	81	✓
3.3.4	Balance for negative class	[16]	81	×
4.1	Causal discrimination	[13]	1	×
4.2	Fairness through unawareness	[17]	14	✓
4.3	Fairness through awareness	[12]	208	×
5.1	Counterfactual fairness	[17]	14	–
5.2	No unresolved discrimination	[15]	14	–
5.3	No proxy discrimination	[15]	14	–
5.4	Fair inference	[19]	6	–

Table 1: Considered Definitions of Fairness

- GOAL: mathematically certify that an algorithm does not suffer from disparate treatment or disparate impact
- It is impossible to write down agreed-upon legal rules and definitions using formal mathematics
- Even if a well-defined definition of fairness gets implemented in a machine-learning-based system
  - what the people impacted by that system
    - understand about the system itself and
    - think about the rules under which it is operating
  - laypeople largely do not understand the accepted definitions of fairness in machine learning
  - those who do understand these definitions do not like them
  - those who do not understand them could be further marginalized

https://fairware.cs.umass.edu/papers/Verma.pdf

https://www.arthur.ai/blog/fairness-in-ml

Algorithmic Fairness

- How can we ensure that algorithms act in ways that are fair and ethical?
  - This definition is vague
  - Describes a broad set of problems, not a specific technical approach

- Related to ideas of:
  - **Accountability**: who is responsible for automated behavior? How do we supervise/audit machines that have large impact?
  - **Transparency/Explainability**: why does an algorithm behave in a certain way? Can we understand its decisions? Can it explain itself?
  - **AI safety**: how can AI avoid unintended negative consequences?
  - **Aligned AI**: How can AI make decisions that align with societal values?

Human-AI Interaction

Guidelines for Human-AI interaction design

Microsoft

INITIALLY

- **01** Make clear what the system can do
- **02** Make clear how well the system can do what it can do

DURING INTERACTION

- **03** Time services based on context
- **04** Show contextually relevant information
- **05** Match relevant social norms
- **06** Mitigate social biases

WHEN WRONG

- **07** Support efficient invocation
- **08** Support efficient dismissal
- **09** Support efficient correction
- **10** Scope services when in doubt
- **11** Make clear why the system did what it did

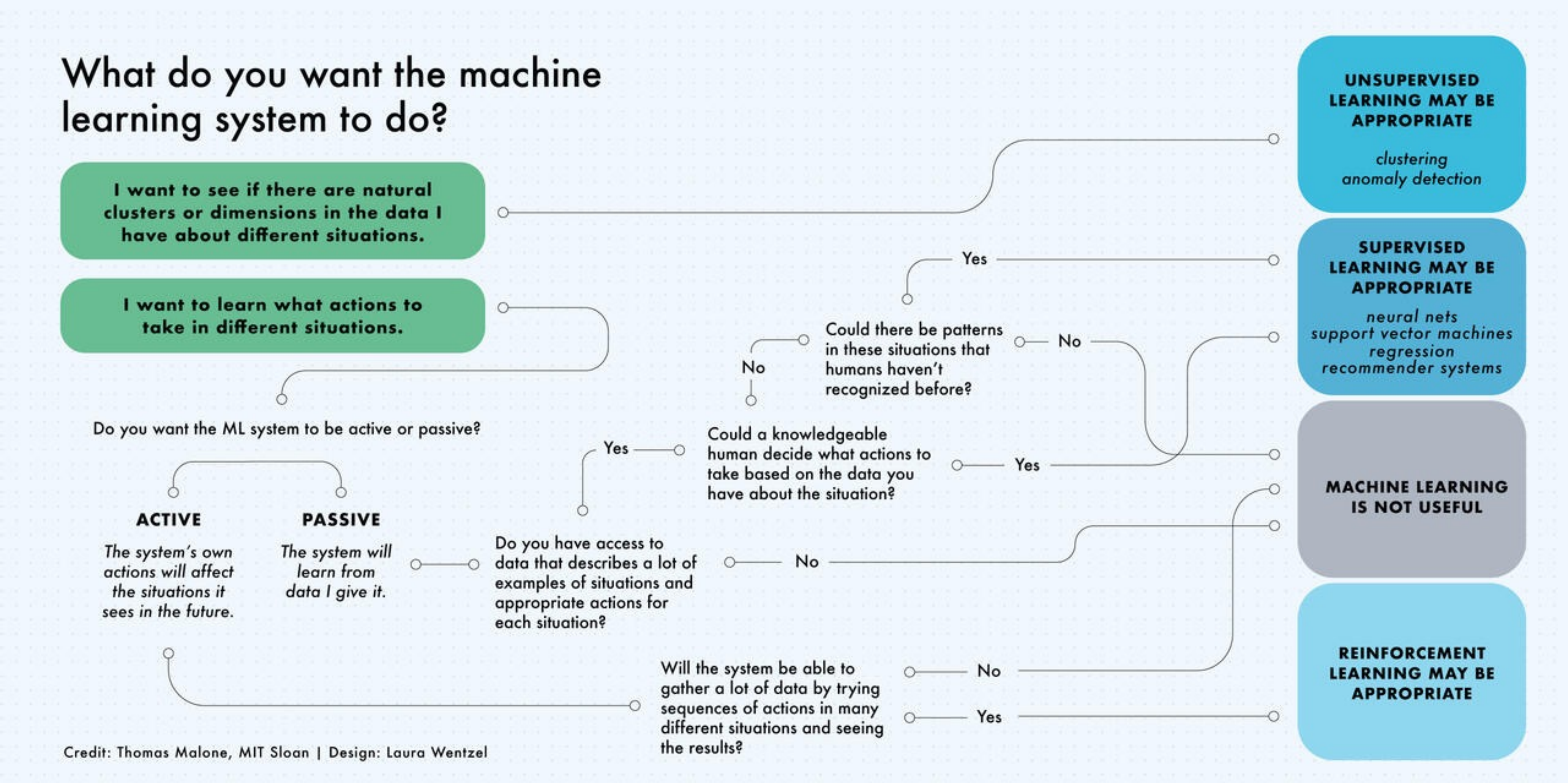
OVER TIME

- **12** Remember recent interactions.
- **13** Learn from user behavior
- **14** Update and adapt cautiously
- **15** Encourage granular feedback
- **16** Convey the consequences of user actions
- **17** Provide global controls
- **18** Notify users about changes

https://www.microsoft.com/en-us/research/blog/guidelines-for-human-ai-interaction-design/

Design guidelines

Picking the right approach



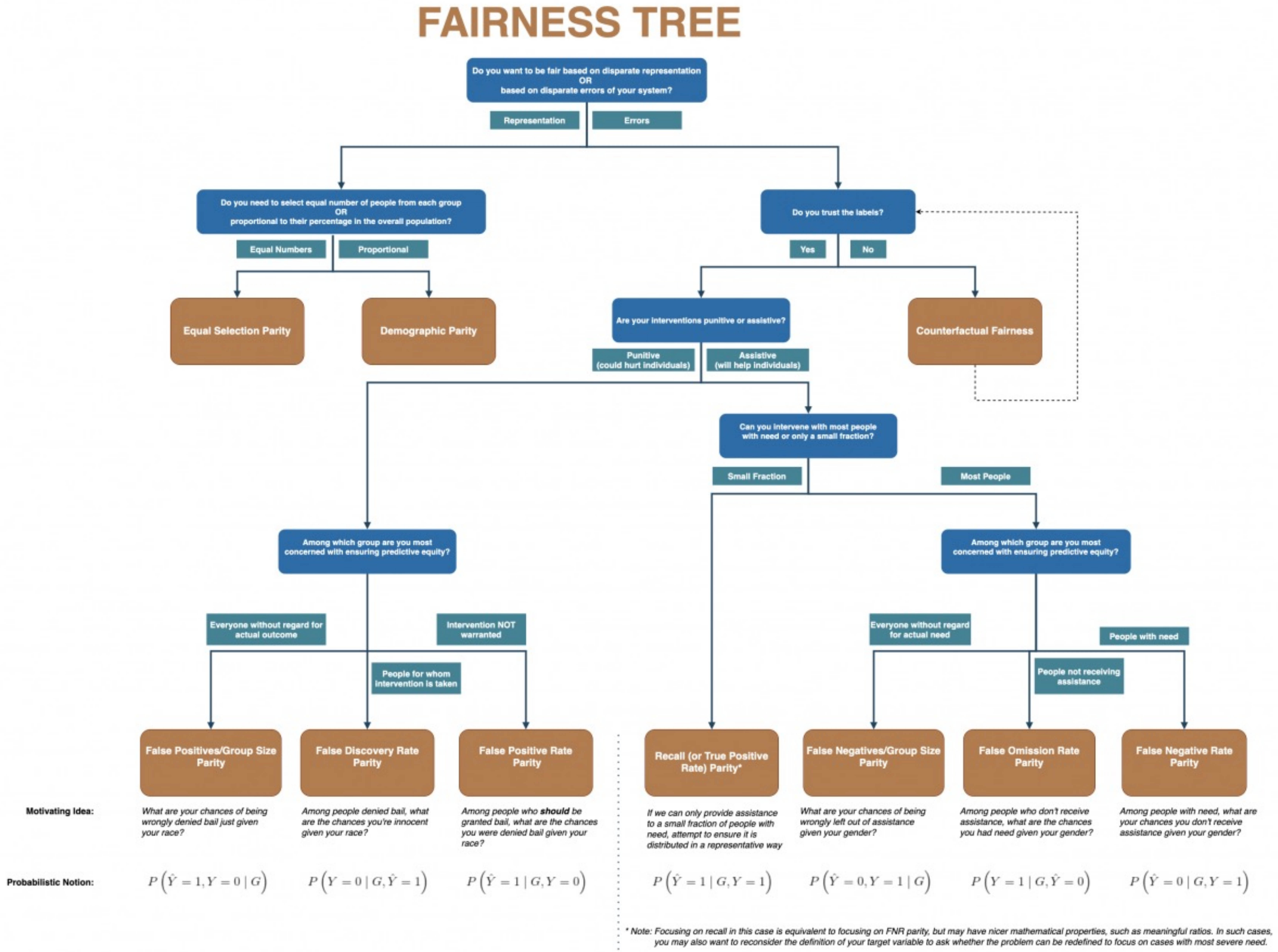
Source: Thomas Malone I MIT Sloan. See: https://bit.ly/3gvRho2, Figure 2.

Responsible AI Practices

- Use a human-centered design approach
- Identify multiple metrics to assess training and monitoring
- When possible, directly examine your raw data
- Understand the limitations of your dataset and model
- Test, test, test
- Continue to monitor and update the system after deployment

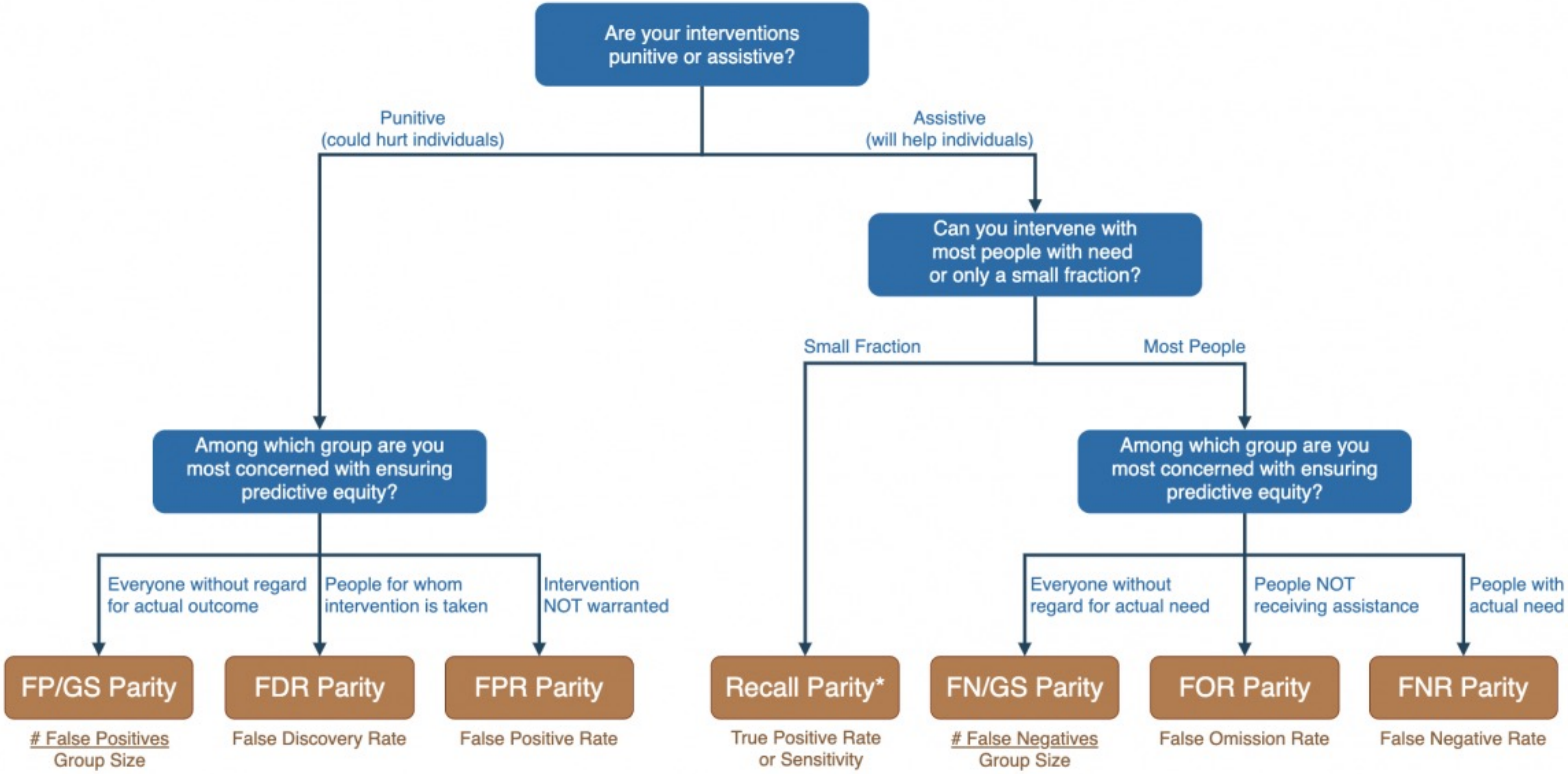
Google AI

https://ai.google/education/responsible-ai-practices



http://www.datasciencepublicpolicy.org/wp-content/uploads/2021/04/Fairness-Full-Tree-1200x908.png

FAIRNESS TREE
(Zoomed in)



THE MACHINE LEARNING CANVAS

Designed for: Designed by: Date: Iteration: .

PREDICTION TASK

Type of task? Entity on which predictions are made? Possible outcomes? Wait time before observation?

DECISIONS

How are predictions turned into proposed value for the end-user? Mention parameters of the process / application that does that.

VALUE PROPOSITION

Who is the end-user? What are their objectives? How will they benefit from the ML system? Mention workflow/interfaces.

DATA COLLECTION

Strategy for initial train set & continuous update. Mention collection rate, holdout on production entities, cost/constraints to observe outcomes.

DATA SOURCES

Where can we get (raw) information on entities and observed outcomes? Mention database tables, API methods, websites to scrape, etc.

IMPACT SIMULATION

Can models be deployed? Which test data to assess performance? Cost/gain values for (in)correct decisions? Fairness constraint?

MAKING PREDICTIONS

When do we make real-time / batch pred.? Time available for this + featurization + post-processing? Compute target?

BUILDING MODELS

How many prod models are needed? When would we update? Time available for this (including featurization and analysis)?

FEATURES

Input representations available at prediction time, extracted from raw data sources.

MONITORING

Metrics to quantify value creation and measure the ML system's impact in production (on end-users and business)?

OWNML.CO

There is more, much more

Designing Machine Learning Solutions

- Training Data
- (Expected) Performance
- *Transparency and Explainability*
- Human-AI Interaction
- *Privacy*
- *Trust*

Sources

- Grokking Machine Learning. Luis G. Serrano. Manning, 2021
- CIS 419/519 Applied Machine Learning. Eric Eaton, Dinesh Jayaraman. https://www.seas.upenn.edu/~cis519/spring2020/
- Societal Computing, Prof. Kenny Joseph

Advanced Machine Learning For Design

Lecture 7: Train, Evaluate and Integrate Machine Learning Models (part 2)

Module 3

Evangelos Niforatos

01/11/2023

aml4d-ide@tudelft.nl
https://aml4design.github.io/