Trustworthy Machine Learning

Kush R. Varshney

krvarshn@us.ibm.com
@krvarshney

Kush R. Varshney

krvarshn@us.ibm.com
@krvarshney

Kush R. Varshney

krvarshn@us.ibm.com
@krvarshney

> "If you want to make a difference, you have to learn how to operate within imperfect systems. Burning things down rarely works. It may allow for personal gains. But if you care about making the system work for many, you have to do it from the inside."
>
> —Nadya Bliss, computer scientist at Arizona State University

Kush R. Varshney
krvarshn@us.ibm.com
@krvarshney

Trustworthy Machine Learning

Contents

6 1. Introduction
7 Establishing trust
13 ML lifecycle
16 Safety

23 2. Data
24 Sources and biases
28 Privacy and consent

37 3. Basic modeling
38 Detection theory
46 Supervised learning
56 Causal modeling

66 4. Reliability
74 Distribution shift
95 Fairness
114 Adversarial robustness

121 5. Interaction
123 Interpretability and explainability
135 Transparency
144 Value alignment

162 6. Purpose
164 Ethics principles
169 Lived experience
174 Social good

179 7. Foundation models

Part 1

Introduction and preliminaries

Chapter 1

Establishing trust

“A decision aid, no matter how sophisticated or ‘intelligent’ it may be, may be rejected by a decision maker who does not trust it, and so its potential benefits to system performance will be lost.”

—Bonnie M. Muir, psychologist at University of Toronto

“The toughest thing about the power of trust is that it’s very difficult to build and very easy to destroy.”

—Thomas J. Watson, Sr., CEO of IBM

Trust is the willingness of a trustor to be vulnerable to the actions of a trustee based on the expectation that the trustee will perform a particular action important to the trustor, irrespective of the ability to monitor or control the trustee.

Trusted vs. trustworthy

The trustee has certain properties that make it trustworthy.

The qualities by which the trustor can expect the trustee to perform the important action referred to in the definition of trust.

Being trustworthy does not automatically imply that the trustee is trusted.

The trustor must consciously make a decision to be vulnerable to the trustee based on its trustworthiness and other factors including cognitive biases of the trustor.

Understandably, potential trustors who are already vulnerable as members of marginalized groups may not want to become even more vulnerable.

Attributes of trustworthiness

	Source	Attribute 1	Attribute 2	Attribute 3	Attribute 4
trustworthy people	Mishra	competent	reliable	open	concerned
	Maister et al.	credibility	reliability	intimacy	low self-orientation
	Sucher and Gupta	competent	use fair means to achieve its goals	take responsibility for all its impact	motivated to serve others' interests as well as its own
trustworthy artificial intelligence	Toreini et al.	ability	integrity	predictability	benevolence
	Ashoori and Weisz	technical competence	reliability	understandability	personal attachment

A trustworthy machine learning system is one that has sufficient:

1 basic performance

2 reliability

3 human interaction

4 aligned purpose

Exercise

Spend 3 minutes writing down your positionality.

Description of your identity in society, including privileges and experiences, especially as it relates to developing and deploying artificial intelligence-based technologies.

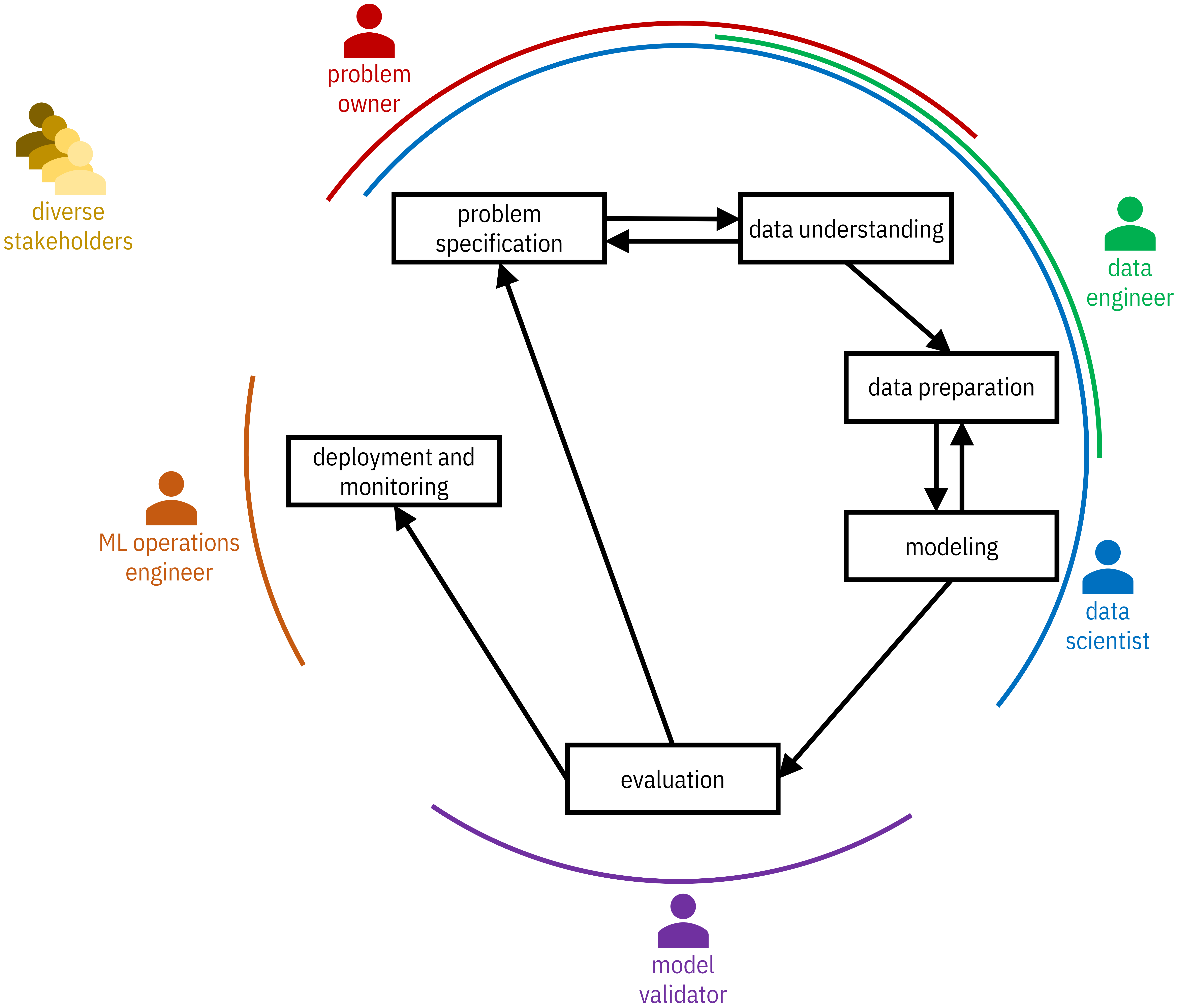
Chapter 2
Machine learning lifecycle

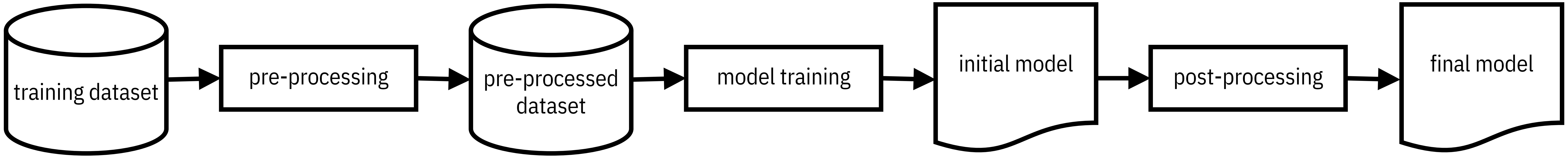
“People are involved in every phase of the AI lifecycle, making decisions about which problem to address, which data to use, what to optimize for, etc.”

—Jenn Wortman Vaughan, research scientist at Microsoft

“We all have a responsibility to ask not just, ‘can we do this?’, but ‘should we do this?’”

—Kathy Baxter, ethical AI practice architect at Salesforce


problem owner
diverse stakeholders
problem specification
data understanding
data engineer
data preparation
modeling
data scientist
deployment and monitoring
ML operations engineer
evaluation
model validator



training dataset → pre-processing → pre-processed dataset → model training → initial model → post-processing → final model

domain adaptation (distribution shift)	domain robustness (distribution shift)	
bias mitigation pre-processing (unfairness)	bias mitigation in-processing (unfairness)	bias mitigation post-processing (unfairness)
data sanitization (adversarial attacks)	smoothing/adversarial training (adversarial attacks)	patching (adversarial attacks)
disentangled representation (lack of explainability)	directly interpretable models (lack of explainability)	post hoc explanations (lack of explainability)

Chapter 3
Safety

“Not knowing the chance of mutually exclusive events and knowing the chance to be equal are two quite different states of knowledge.”

—Ronald A. Fisher, statistician and geneticist

Safety is the reduction of both aleatoric uncertainty (or risk) and epistemic uncertainty associated with harms.

Harm

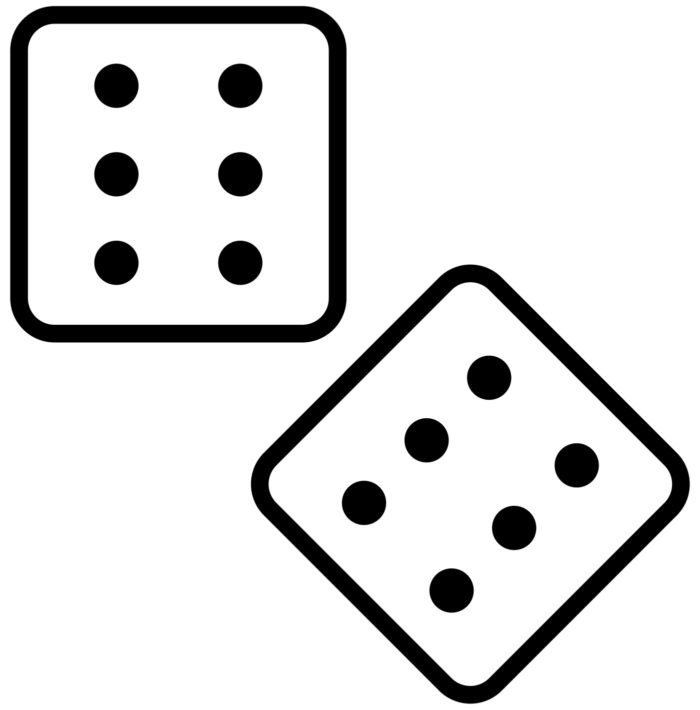
An undesired outcome is a harm if its cost exceeds some threshold.

Image credit: https://www.nytimes.com/2023/02/15/well/live/personal-care-products-chemicals.html

Two types of uncertainty

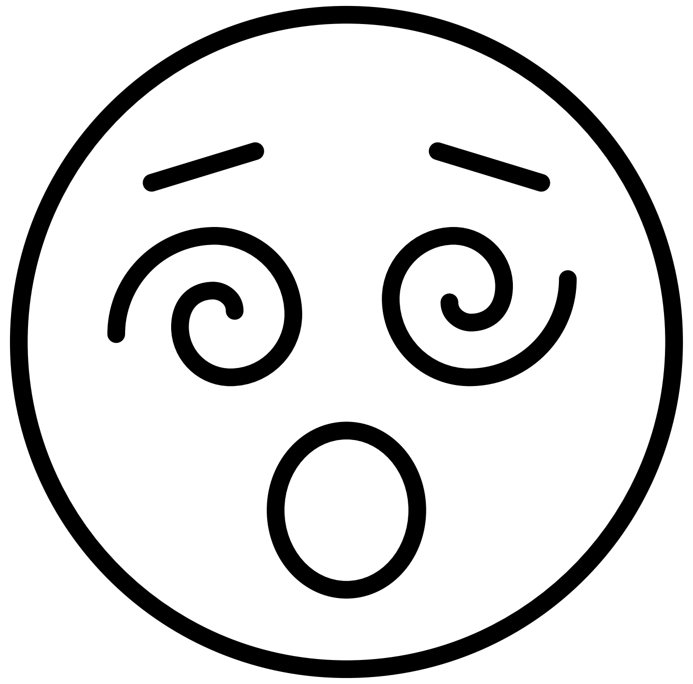
Aleatoric

- Inherent randomness
- Quantified through probability theory
- Relates to basic performance attribute of trustworthiness



Epistemic

- Observer-dependent lack of knowledge
- Quantified through possibility theory
- Relates to reliability attribute of trustworthiness



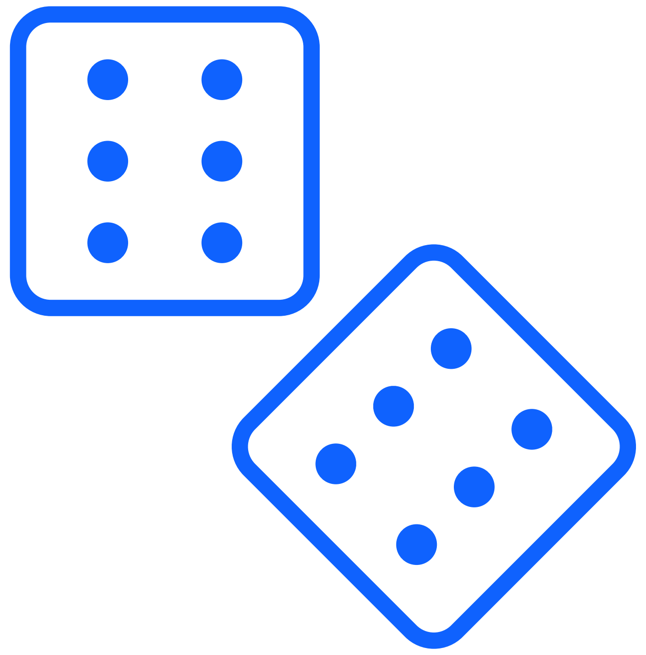
Probability

Properties of the probability function:

$$P(A) \geq 0,$$

$$P(\Omega) = 1,$$

if A and B are disjoint events (they have no outcomes in common; $A \cap B = \emptyset$), then $P(A \cup B) = P(A) + P(B)$.



Possibility

Properties of the possibility function:

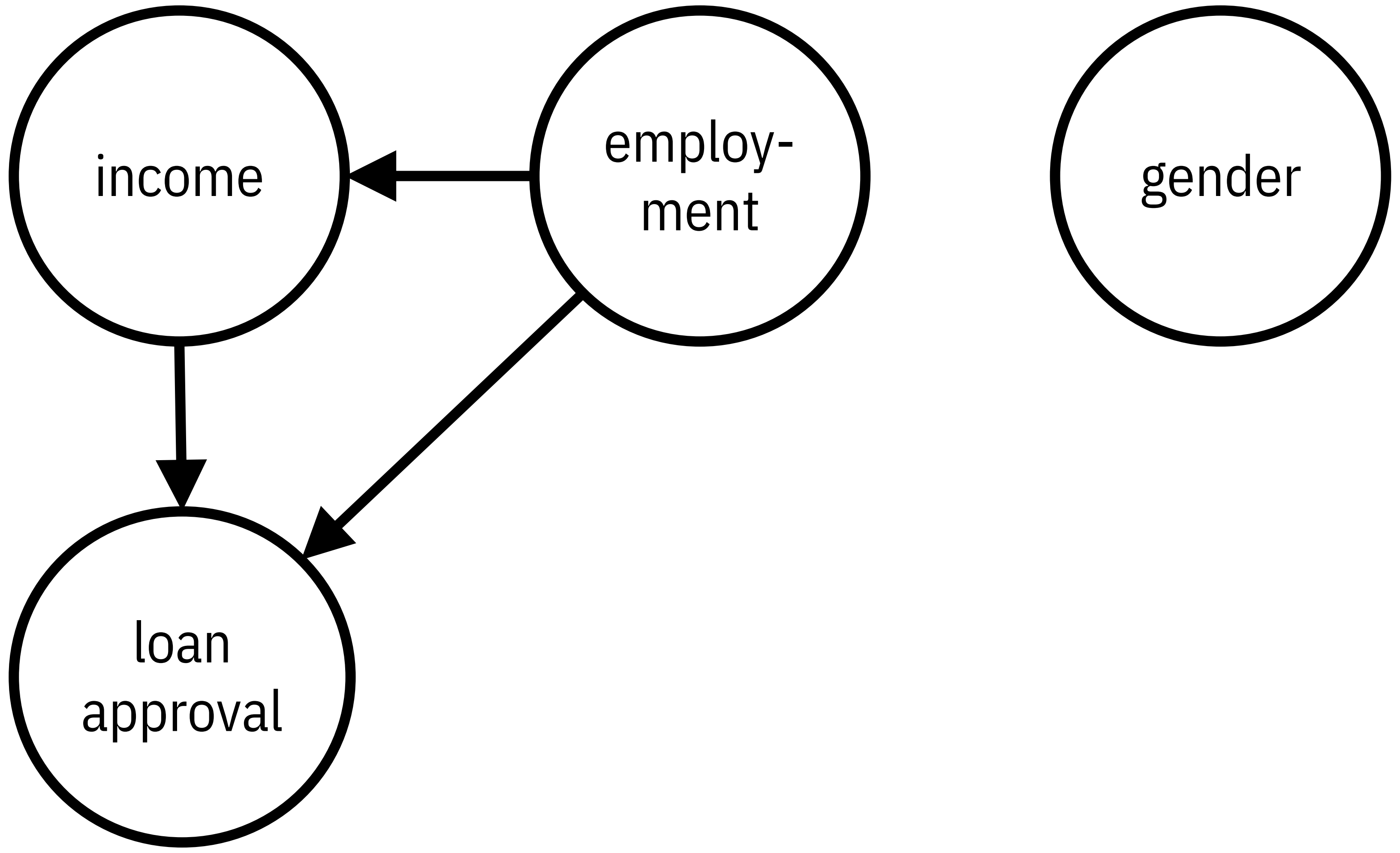
$$\Pi(\emptyset) = 0,$$

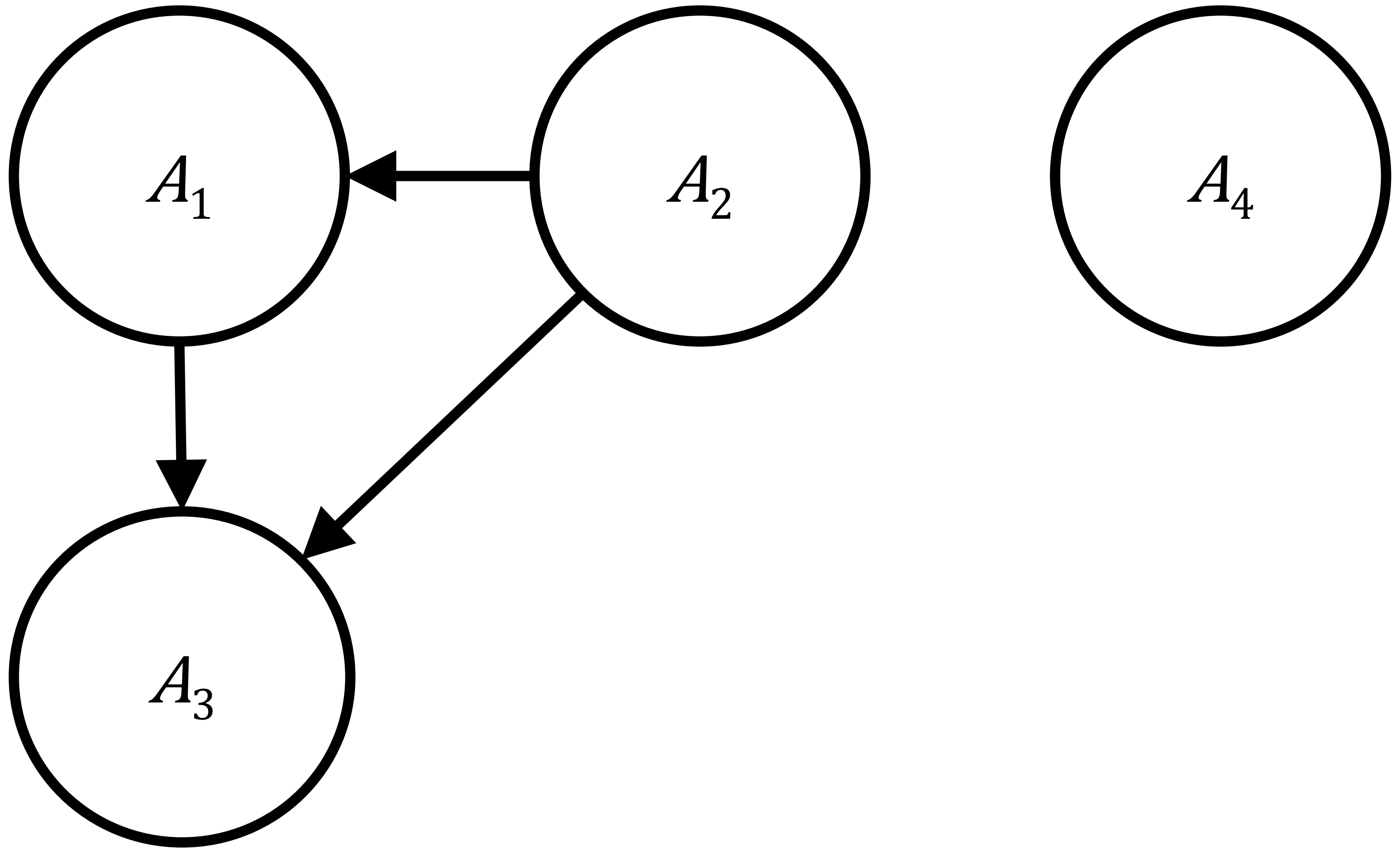
$$\Pi(\Omega) = 1,$$

if A and B are disjoint events (they have no outcomes in common; $A \cap B = \emptyset$), then $\Pi(A \cup B) = \max\big(\Pi(A), \Pi(B)\big)$

Bayesian networks

(credal networks in possibility theory)





$$P(A_1, \dots, A_n) = \prod_{j=1}^{n} P\big(A_j \mid pa(A_j) \big)$$

$$P(A_1, A_2, A_3, A_4) = P(\, A_1 \mid A_2 \,)P(A_2)P(\, A_3 \mid A_1, A_2 \,)P(A_4)$$

Safety is the reduction of the probability of expected harms and the possibility of unexpected harms.

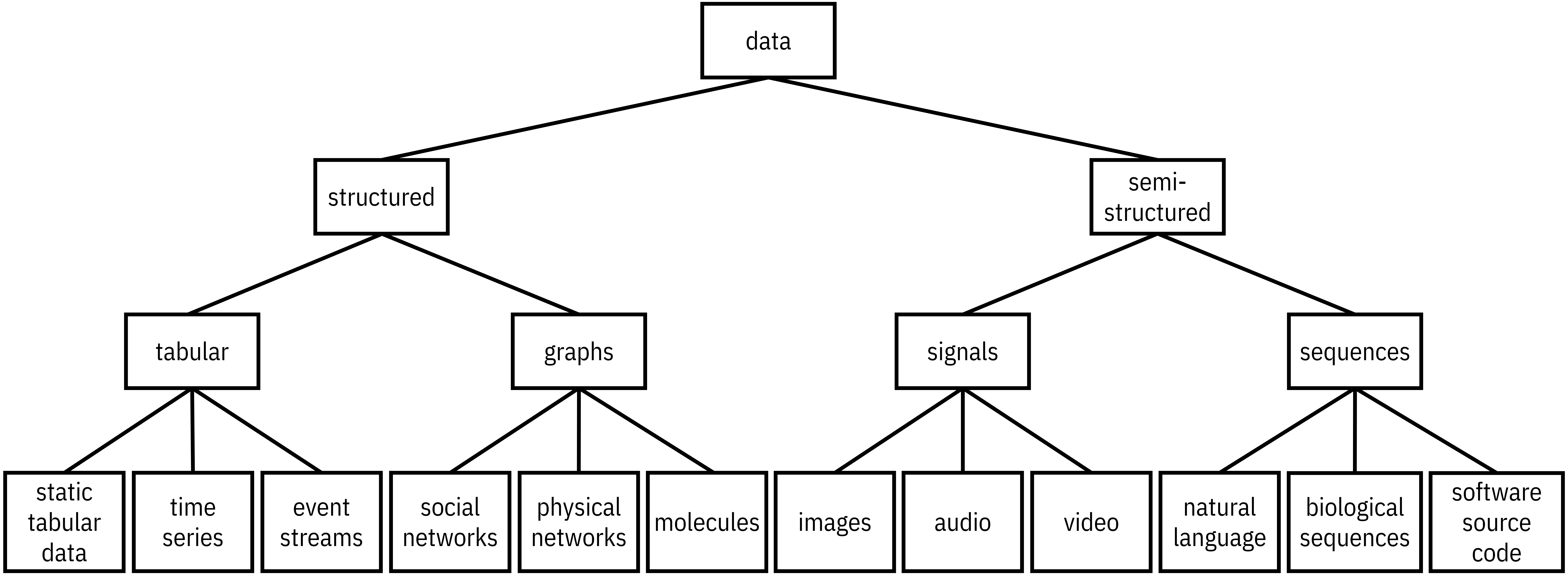
Use both in problem specification.

Part 2
Data

Chapter 4

Data Modalities, Sources, and Biases

“Garbage in, garbage out.”
—Wilf Hey, computer scientist at IBM


data

structured

semi-structured

tabular

graphs

signals

sequences

static tabular data

time series

event streams

social networks

physical networks

molecules

images

audio

video

natural language

biological sequences

software source code

Data sources

Purposefully collected data

Includes surveys and censuses as well as the results of scientific experiments. These sources have the veneer of being well-designed and with minimal bias, but this might not always be the case.

Administrative data

Collected by organizations about their routine operations for non-statistical reasons. Can contain traces of historical prejudices. Might not exactly match the problem you are trying to solve. Usually correct.

Social data

Includes user-generated content, relationships between people, and traces of behavior. Might not match the problem. Reliability may be much less than administrative data. Marginalized populations may be invisible in some types of social data.

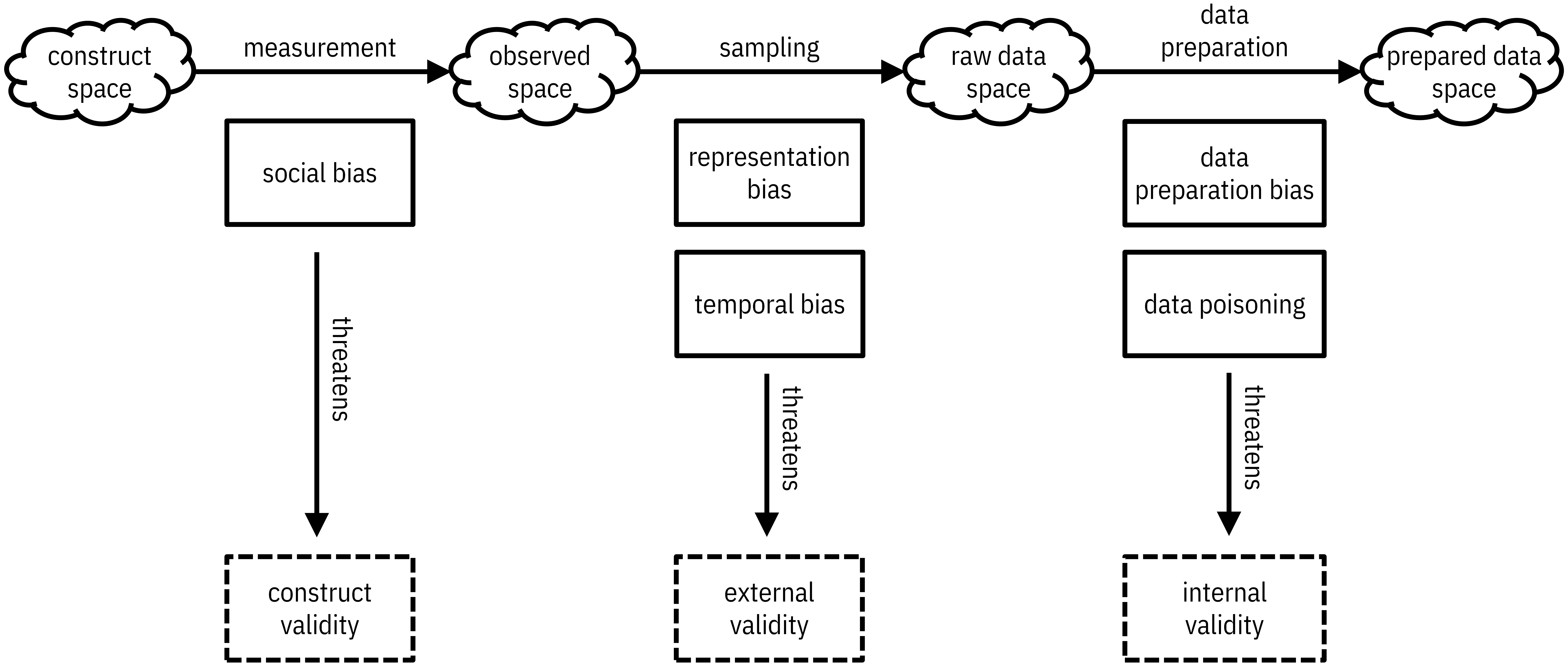
Crowdsourcing

In many crowdsourcing platforms, the workers are often low-skill individuals. They may be unfamiliar with the task or the social context of the task, which may yield biases in labels. Their wages may be low, which raises ethical concerns.

Data augmentation

Performing various transformations of the given dataset may be used to increase data set size without actually collecting additional data. Another way is through *generative* machine learning trained on the given dataset.

Different data sources are useful in addressing various problem specifications, but all have biases of one kind or another.



construct space

measurement

observed space

sampling

raw data space

data preparation

prepared data space

social bias

representation bias

temporal bias

data preparation bias

data poisoning

threatens

threatens

threatens

construct validity

external validity

internal validity

Chapter 5
Privacy and consent

> “Data is the new oil.”
>
> —Clive Humby, data science entrepreneur at dunnhumby

Data is power

Data is a valuable commodity.

It reveals a lot about human behavior at a gross level, but also about the behavior of individual people.

Just like other natural resources, it can be extracted from the vulnerable without their consent and furthermore be exploited for their subjugation.

Data used in machine learning is often fraught with power and consent issues because it is often repurposed.

For example, many large-scale image datasets used for training computer vision models are scraped from the internet without explicit consent from the people who posted the images.

Although there may be implicit consent through vehicles such as Creative Commons licenses, a lack of explicit consent can nevertheless be problematic. Sometimes copyright laws are violated in scraped and repurposed data.

Why does this happen? It is almost always due to system designers taking shortcuts to gather large datasets and show value quickly without giving thought to power and consent.

Problem owners and data scientists should not have any calculus to weigh issues of power and consent against convenience in data collection.

Image by Alan Warburton / © BBC / Better Images of AI / Virtual Human / CC-BY 4.0

Privacy is the ability of individuals to withhold information about themselves.

Privacy is considered an essential human right in many value systems and legal frameworks.

identifier quasi-identifier sensitive attribute

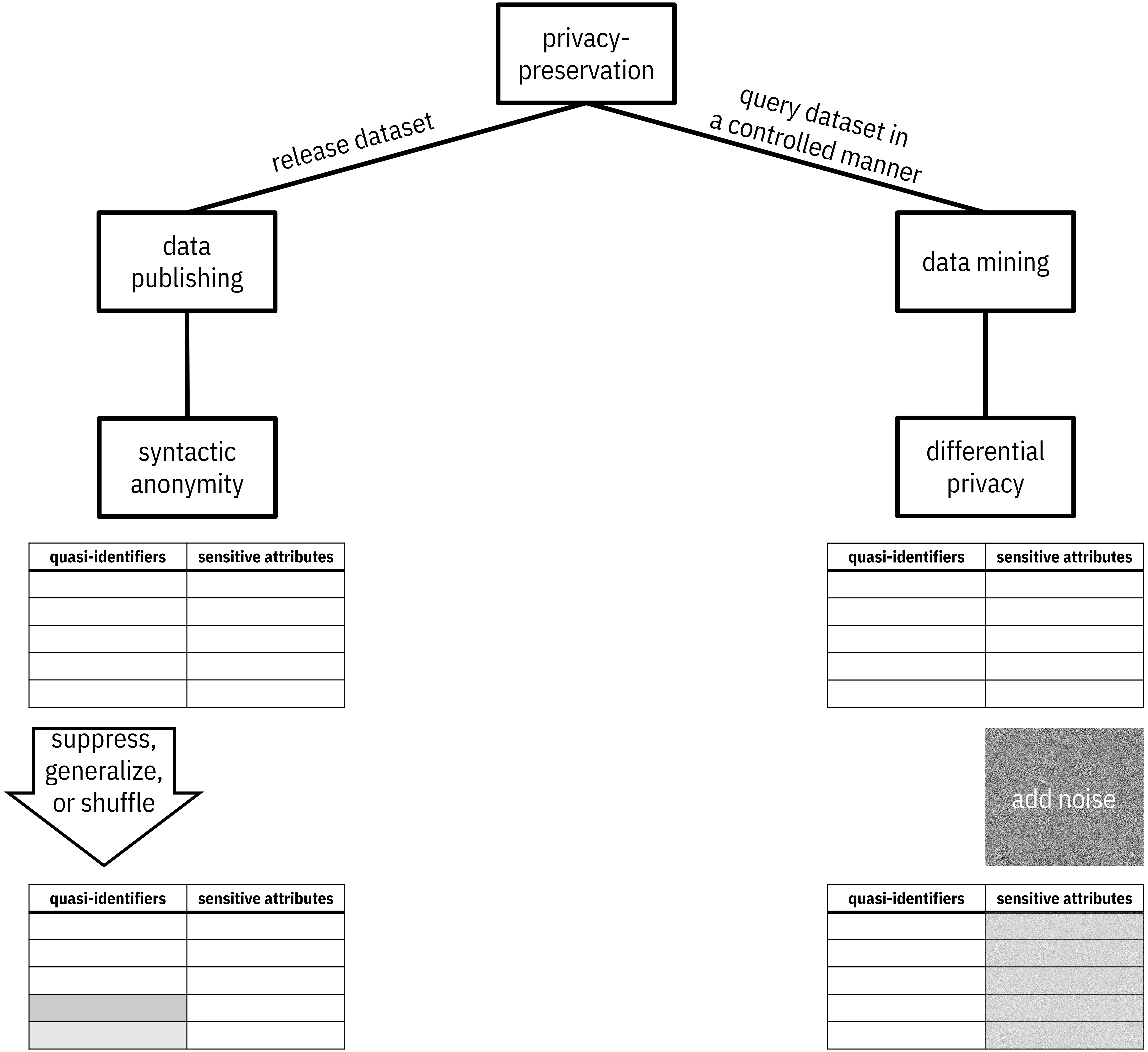
Name	Department	CT Value
Joseph Cipolla	Trustworthy AI	12
Kweku Yefi	Neurosymbolic AI	20
Anjali Singh	AI Applications	35
Celia Sontag	Compute Acceleration	31
Phaedra Paragios	Software-Defined Architecture	19
Chunhua Chen	Thermal Packaging	27

k-anonymity

Organization	CT Value
AI	12
AI	20
AI	35
Hybrid Cloud	31
Hybrid Cloud	19
Hybrid Cloud	27

differential privacy

Name	Department	CT Value
Joseph Cipolla	Trustworthy AI	13.5
Kweku Yefi	Neurosymbolic AI	12.8
Anjali Singh	AI Applications	32.7
Celia Sontag	Compute Acceleration	35.9
Phaedra Paragios	Software-Defined Architecture	22.1
Chunhua Chen	Thermal Packaging	13.4



privacy-preservation

release dataset

query dataset in a controlled manner

data publishing

data mining

syntactic anonymity

differential privacy

quasi-identifiers	sensitive attributes

quasi-identifiers	sensitive attributes

suppress, generalize, or shuffle

add noise

quasi-identifiers	sensitive attributes

quasi-identifiers	sensitive attributes

Demo

https://aip360.res.ibm.com/data

Part 3
Basic modeling

Chapter 6
Detection theory

> “The predictability ceiling is often ignored in mainstream ML research. Every prediction problem has an upper bound for prediction—the Bayes-optimal performance. If you don't have a good sense of what it is for your problem, you are in the dark.”
>
> —Mert Sabuncu, computer scientist at Cornell University

X

feature vector
random variable X
sample space $\mathcal{X}$
sample value x

Y

label
random variable Y
sample space $\{0,1\}$
sample value y

$y = 0$ negative
$y = 1$ positive
$p_0 = P(Y = 0)$
$p_1 = P(Y = 1) = 1 - p_0$

$X|Y$

likelihood functions
$p_{X|Y}(x \mid y = 0)$
$p_{X|Y}(x \mid y = 1)$

$\hat{Y}$

decision function
$\hat{y}\colon \mathcal{X} \to \{0,1\}$
predicts a label from the features

Confusion matrix

	$Y = 1$	$Y = 0$
$\hat{y}(X) = 1$	TP	FP
$\hat{y}(X) = 0$	FN	TN

Confusion matrix

$P(\hat{y}(X) \mid Y)$	$Y = 1$	$Y = 0$
$\hat{y}(X) = 1$	p_{TP}	p_{FP}
$\hat{y}(X) = 0$	p_{FN}	p_{TN}

“Reversed” confusion matrix

$P(Y \mid \hat{y}(X))$	$Y = 1$	$Y = 0$
$\hat{y}(X) = 1$	p_{PPV}	p_{FDR}
$\hat{y}(X) = 0$	p_{FOR}	p_{NPV}

The Bayes risk and the optimal decision function

$$R = (c_{10} - c_{00})p_0 p_{FP} + (c_{01} - c_{11})p_1 p_{FN} + c_{00}p_0 + c_{11}p_1$$

$$R = c_{10}p_0 p_{FP} + c_{01}p_1 p_{FN}$$

$$\hat{y}^*(\cdot) = \arg\min_{\hat{y}(\cdot)} E\left[c\left(Y, \hat{y}(X)\right)\right]$$

The best you can ever do

$$\hat{y}^*(\cdot) = \begin{cases} 0, & \Lambda(x) \leq \eta \\ 1, & \Lambda(x) > \eta \end{cases}$$

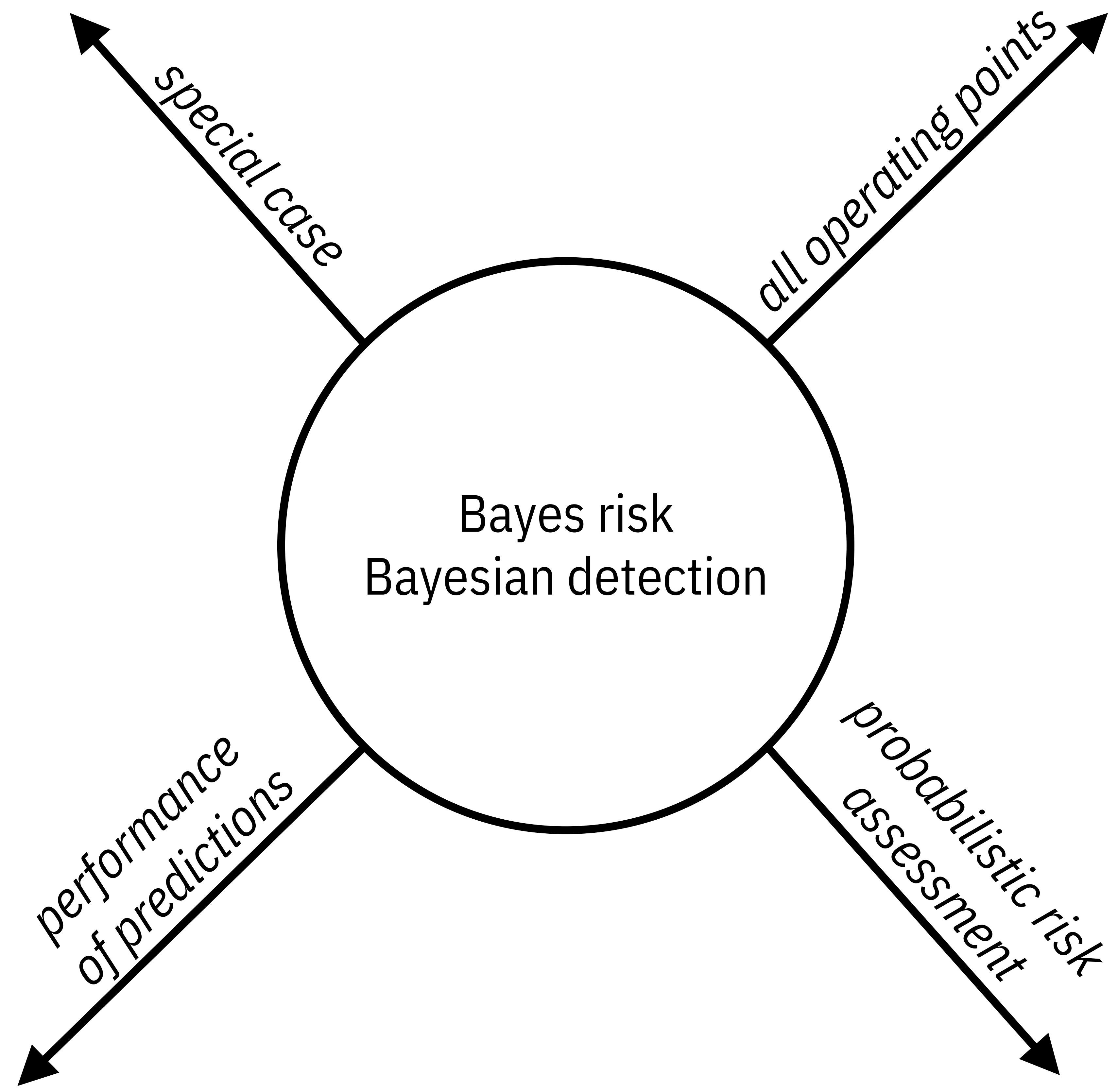
likelihood ratio test

likelihood ratio

$$\Lambda(x) = \frac{p_{X|Y}(\, x \mid Y = 1\,)}{p_{X|Y}(\, x \mid Y = 0\,)}$$

threshold

$$\eta = \frac{c_{10}p_0}{c_{01}p_1}$$



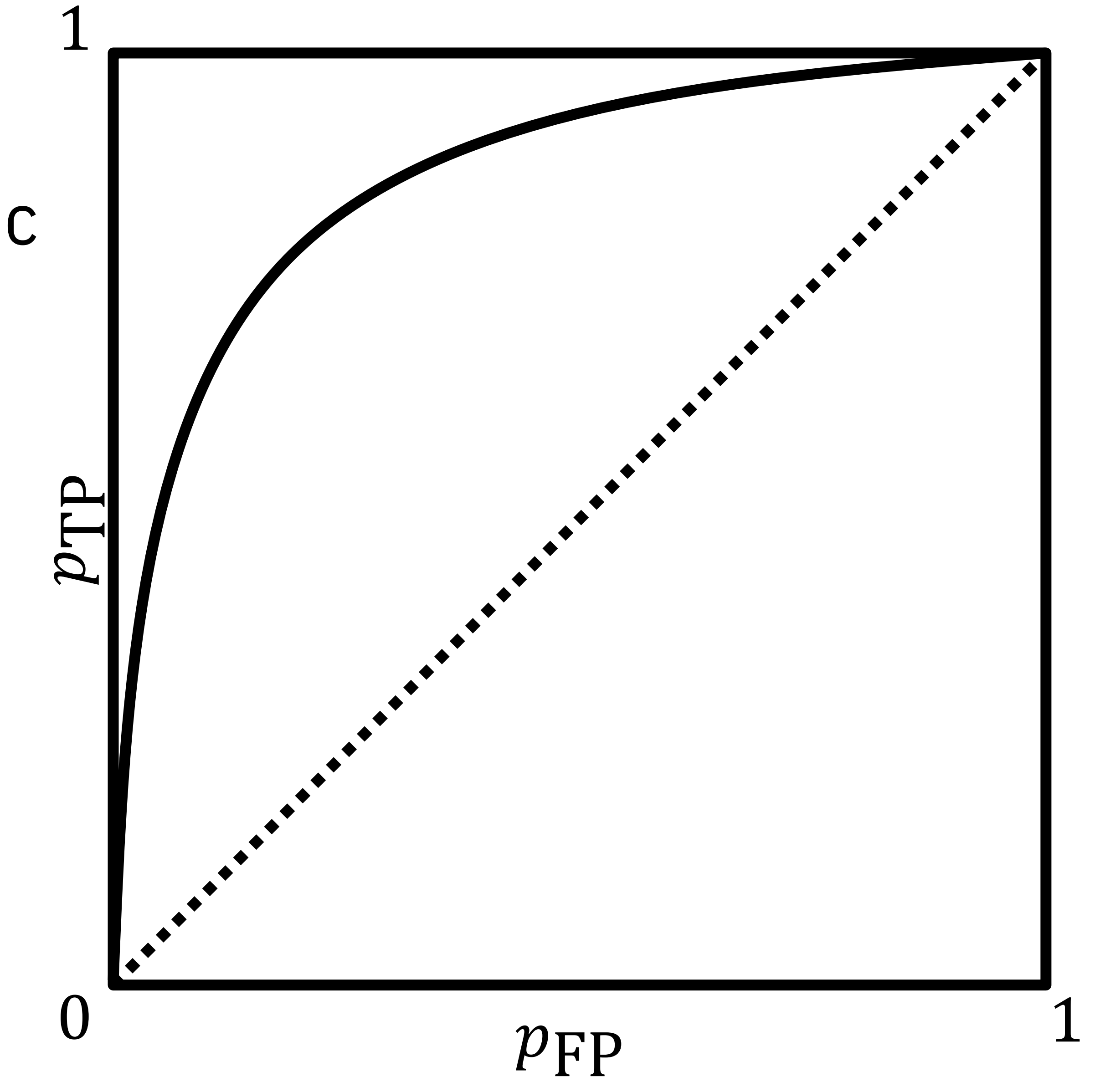
Bayes risk
Bayesian detection

special case

false positive rate
false negative rate
error rate
accuracy

all operating points

receiver operating characteristic
recall-precision curve
area under the curve

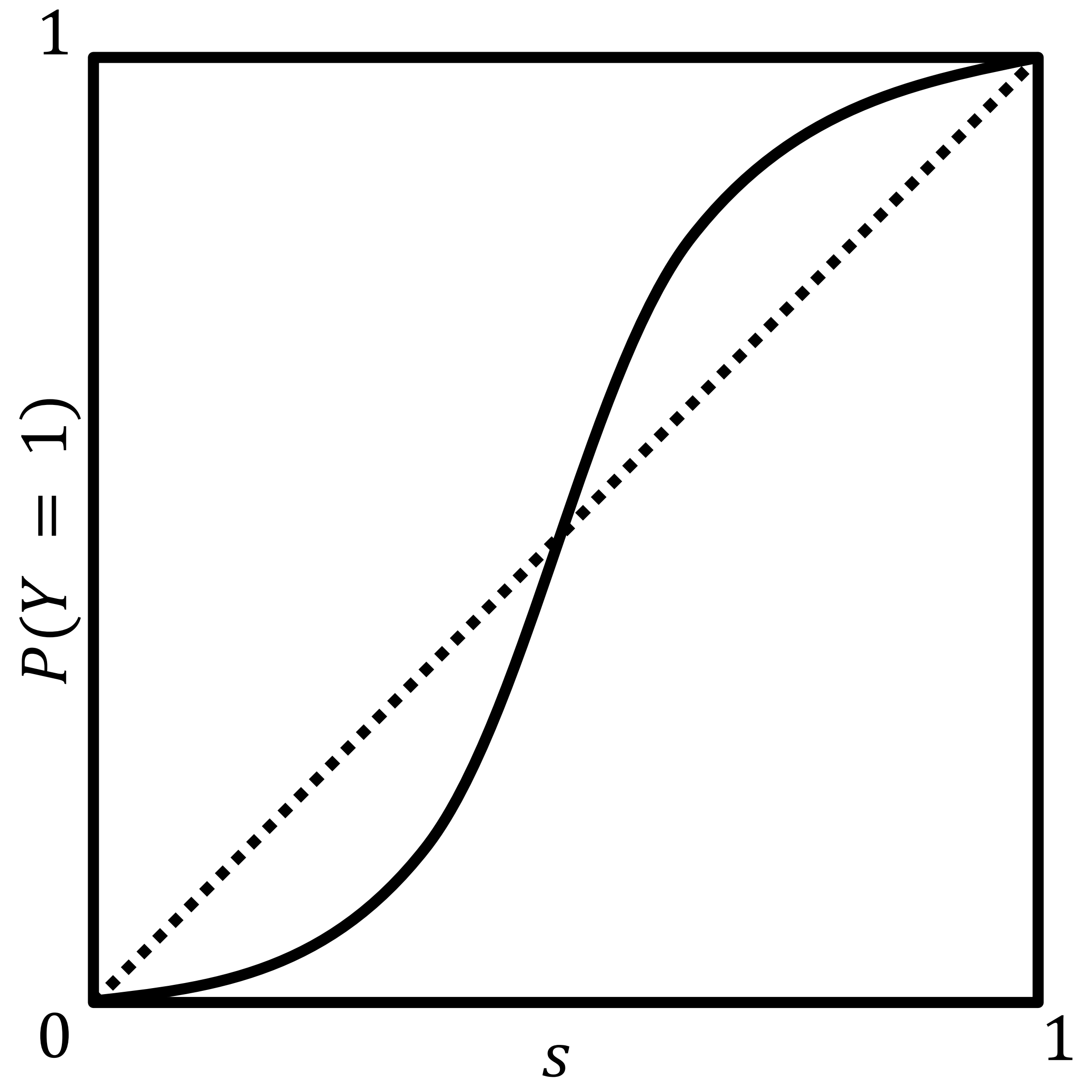


performance of predictions

false discovery rate
false omission rate
F_1-score

probabilistic risk assessment

Brier score
calibration curve



Chapter 7
Supervised learning

“My experience in industry strongly confirms that deep learning is a narrow sliver of methods needed for solving complex automated decision making problems.”

—Zoubin Ghahramani, chief scientist at Uber

X

feature vector
sample space $\mathcal{X}$
sample value x

Y

label
sample space $\{0,1\}$
sample value y

$y = 0$ negative
$y = 1$ positive

X, Y

training data
$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

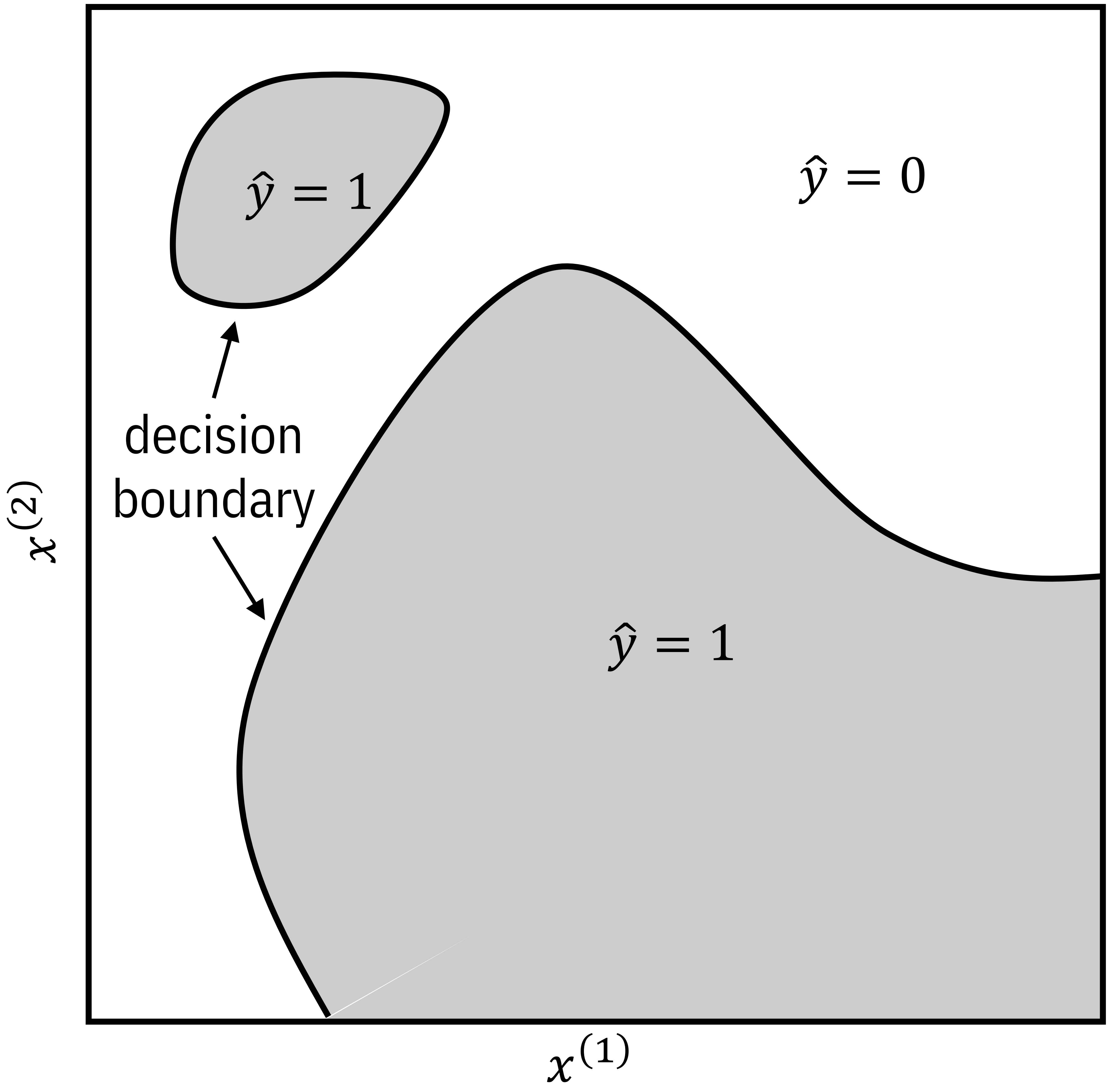
$\hat{Y}$

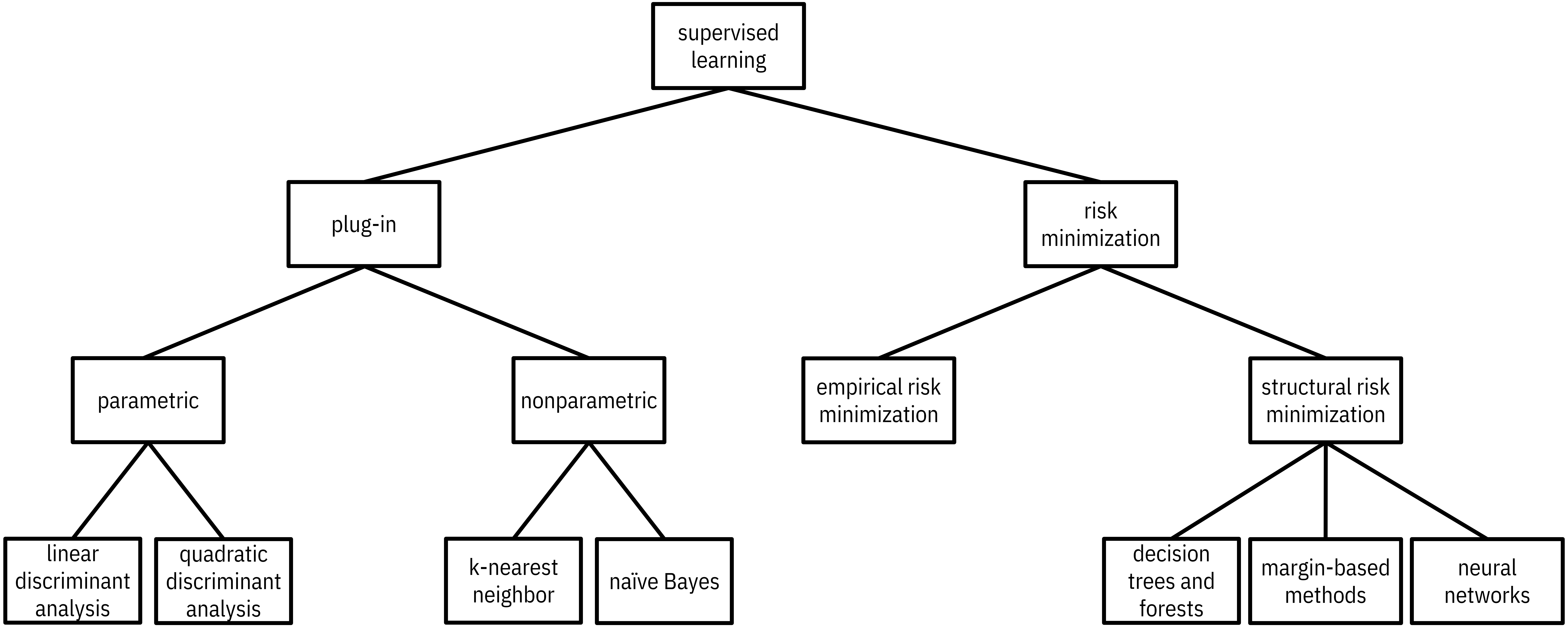
decision function
$\hat{y}: \mathcal{X} \to \{0,1\}$
predicts a label from the features

$$\hat{y}(x) = \begin{cases} 0, & \Lambda(x) \leq \eta \\ 1, & \Lambda(x) > \eta \end{cases}$$

vs.

$\hat{y}(x)$ fit on training data




supervised learning
plug-in
risk minimization
parametric
nonparametric
empirical risk minimization
structural risk minimization
linear discriminant analysis
quadratic discriminant analysis
k-nearest neighbor
naïve Bayes
decision trees and forests
margin-based methods
neural networks

Plug-in

$$\hat{y}(x) = \begin{cases} 0, & \Lambda(x) \leq \eta \\ 1, & \Lambda(x) > \eta \end{cases}$$

estimate the likelihood functions and the prior probabilities from the training data and plug them into the Λ and η of the Bayes optimal decision function

Risk minimization

$$R = c_{10} p_0 p_{FP} + c_{01} p_1 p_{FN}$$

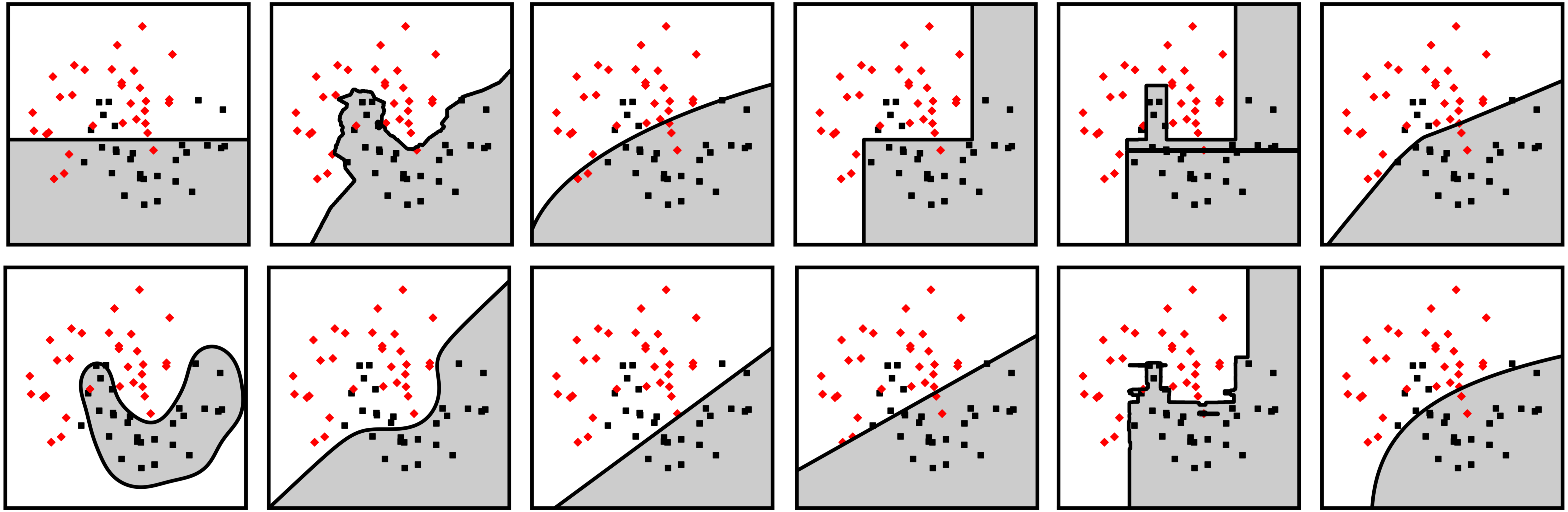
minimize the empirical risk instead of the Bayes risk (with regularization to prevent overfitting)

$$R_{\text{emp}} = \frac{1}{n} \sum_{j=1}^{n} L\left(y_j, \hat{y}(x_j)\right)$$

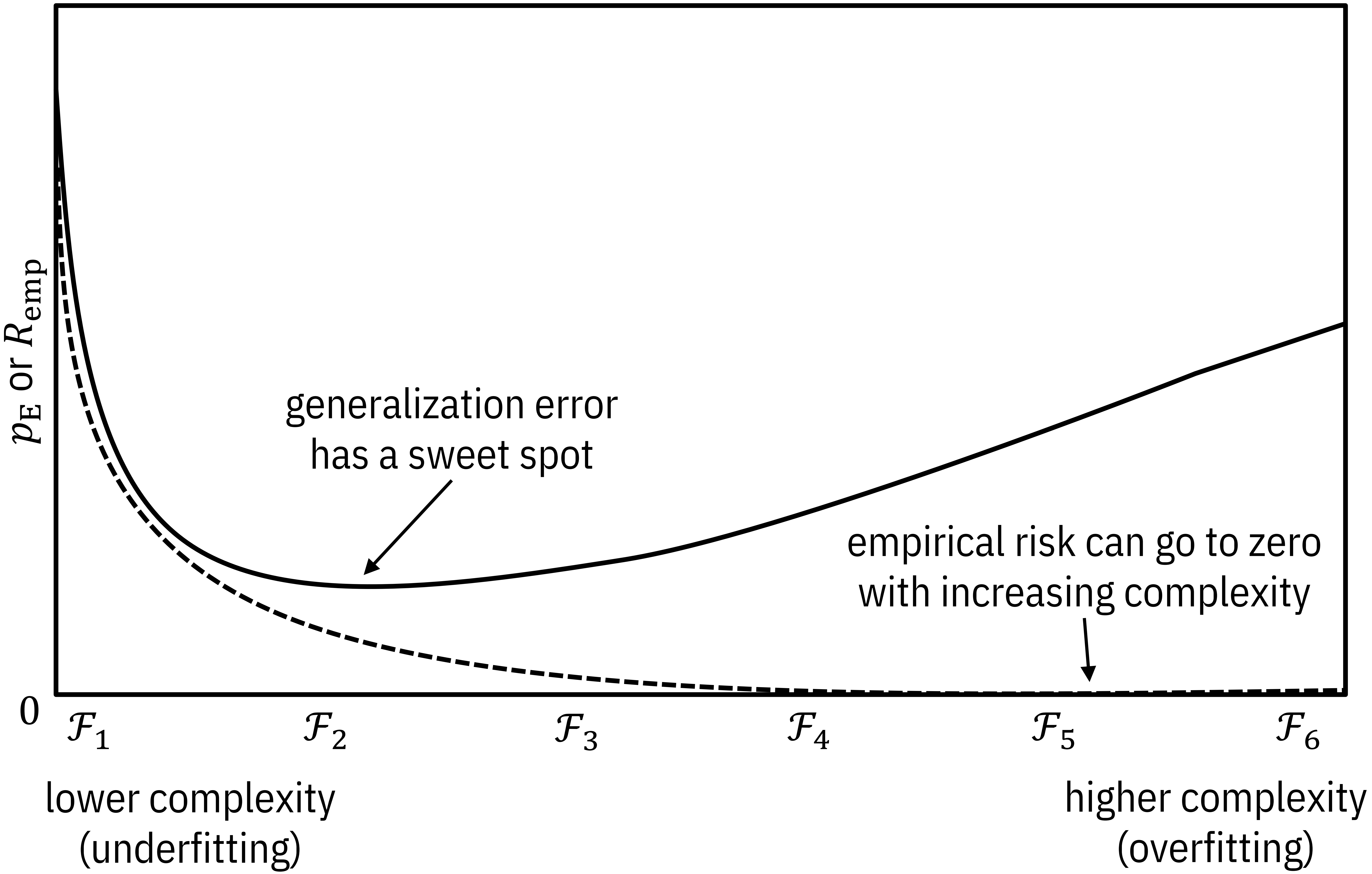
No free lunch. There is no one single machine learning method that is best for all datasets.

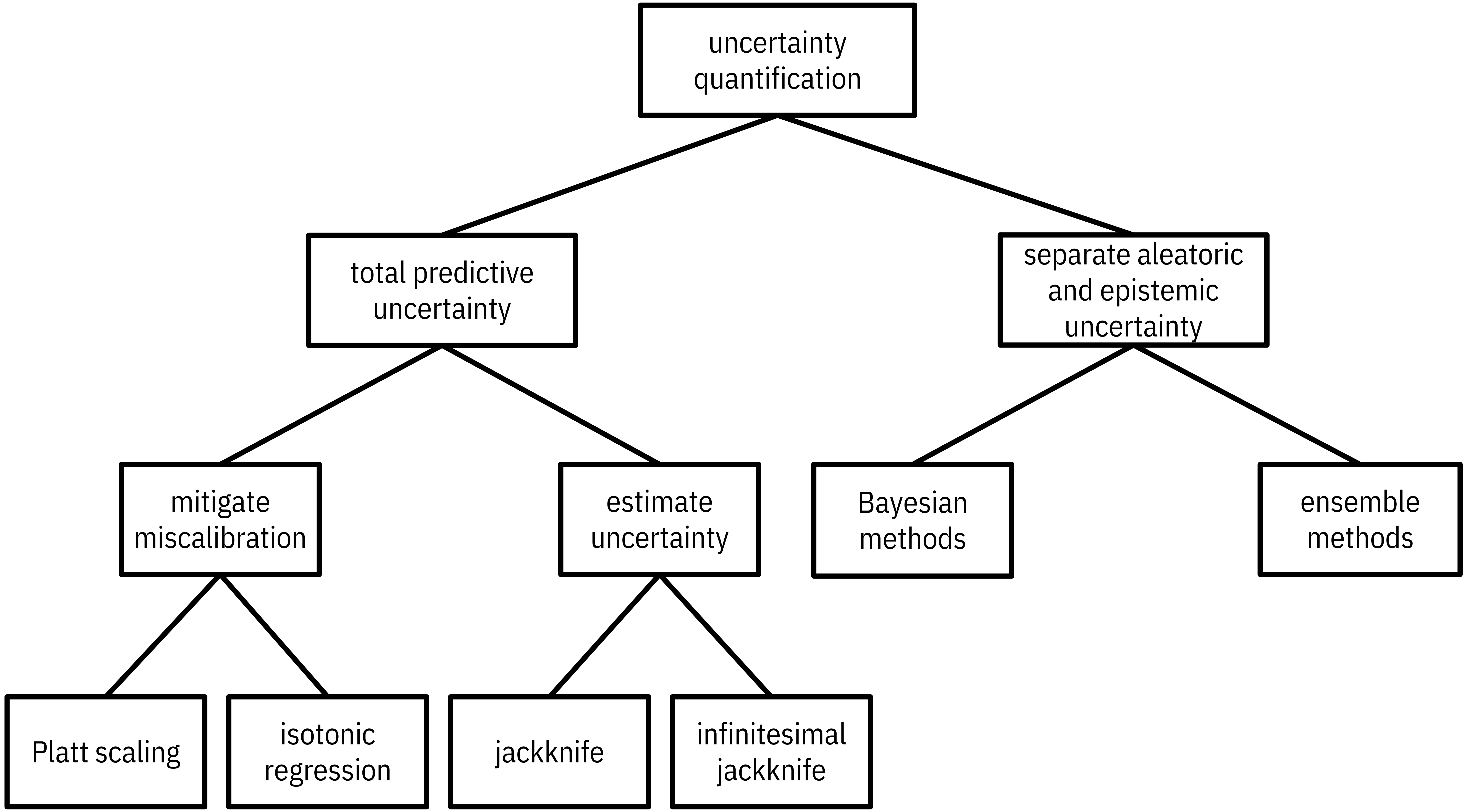
https://www.pexels.com/photo/a-lunch-meal-served-in-a-restaurant-16488178/

Inductive biases, hypothesis spaces $\mathcal{F}$, and domains of competence



Training data and deployment data




uncertainty quantification

total predictive uncertainty

separate aleatoric and epistemic uncertainty

mitigate miscalibration

estimate uncertainty

Bayesian methods

ensemble methods

Platt scaling

isotonic regression

jackknife

infinitesimal jackknife

Demo

https://uq360.res.ibm.com/demo/0

Chapter 8
Causal modeling

“While probabilities encode our beliefs about a static world, causality tells us whether and how probabilities change when the world changes, be it by intervention or by act of imagination.”

—Judea Pearl, computer scientist at University of California, Los Angeles

Examining the associations and correlations in a trained machine learning model may yield some insights, but misses something very important: causality!

When you want to understand the effect of interventions (specific actions that are undertaken) on outcomes, you have to do more than machine learning, you have to perform causal modeling.

Cause and effect are central to understanding the world, but standard supervised learning is not a method for obtaining them.

If doing something makes something else happen, then the something we did is a cause of the something that happened.

The key word is do. *Causation requires doing.*

The actions that are done are known as interventions or treatments.

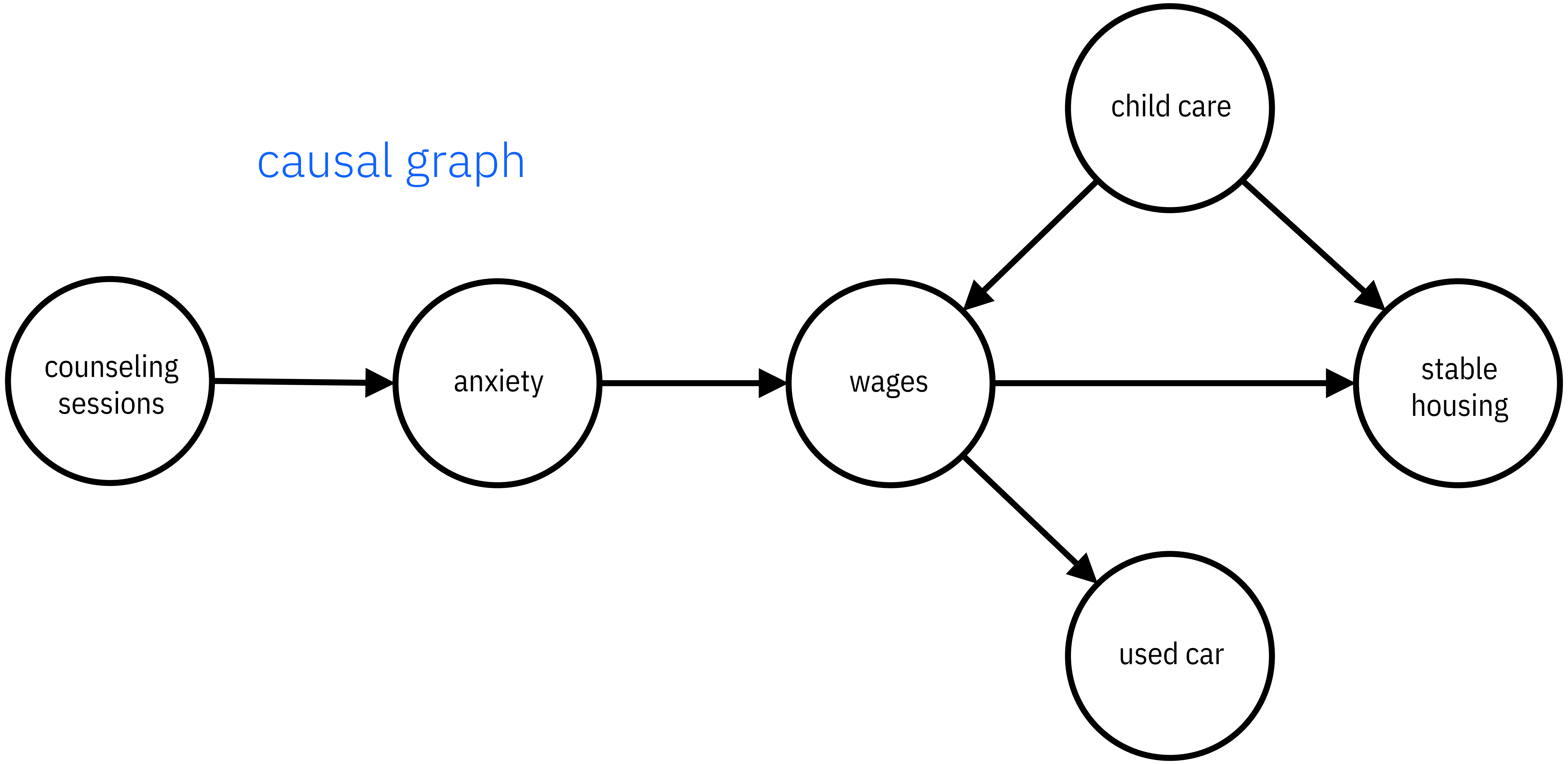
All hell can break loose

Using predictive models to form causal conclusions can lead to great harms.

Changes to input features of predictive models do not necessarily lead to desired changes of output labels.

Structural causal model

causal graph



structural equation

$$P(Y \mid do(t)) = f_Y(t, noise_Y)$$

average treatment effect

$$\tau = E[Y \mid do(t = 1)] - E[Y \mid do(t = 0)]$$

Interventional data

Data from a purposefully designed experiment such as a randomized controlled trial

Gold standard

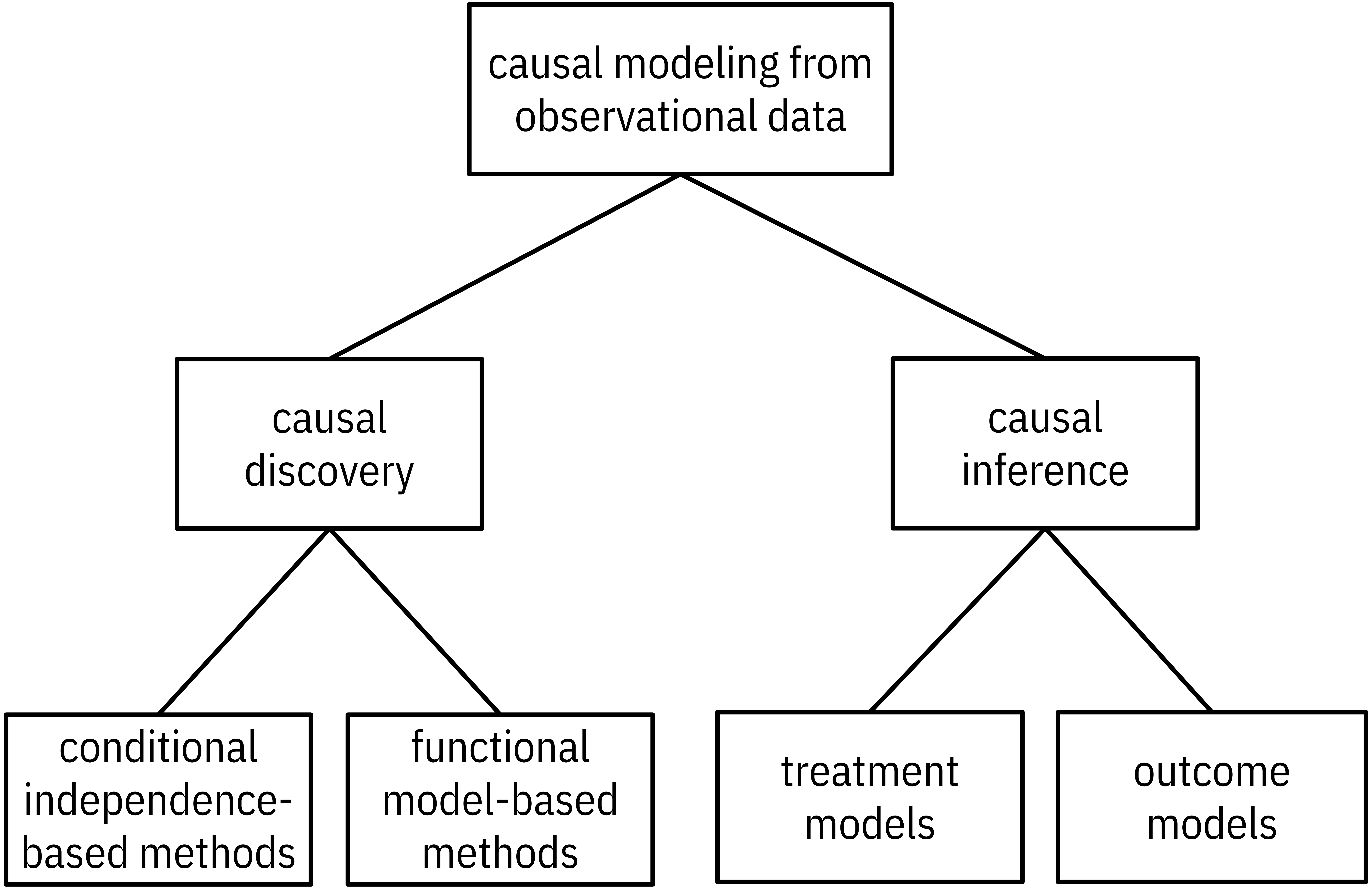
Often not possible, practically or ethically

Observational data

Data not from a purposefully designed experiment

Common case

“Those who can’t do, assume”



causal modeling from observational data

causal discovery

causal inference

conditional independence-based methods

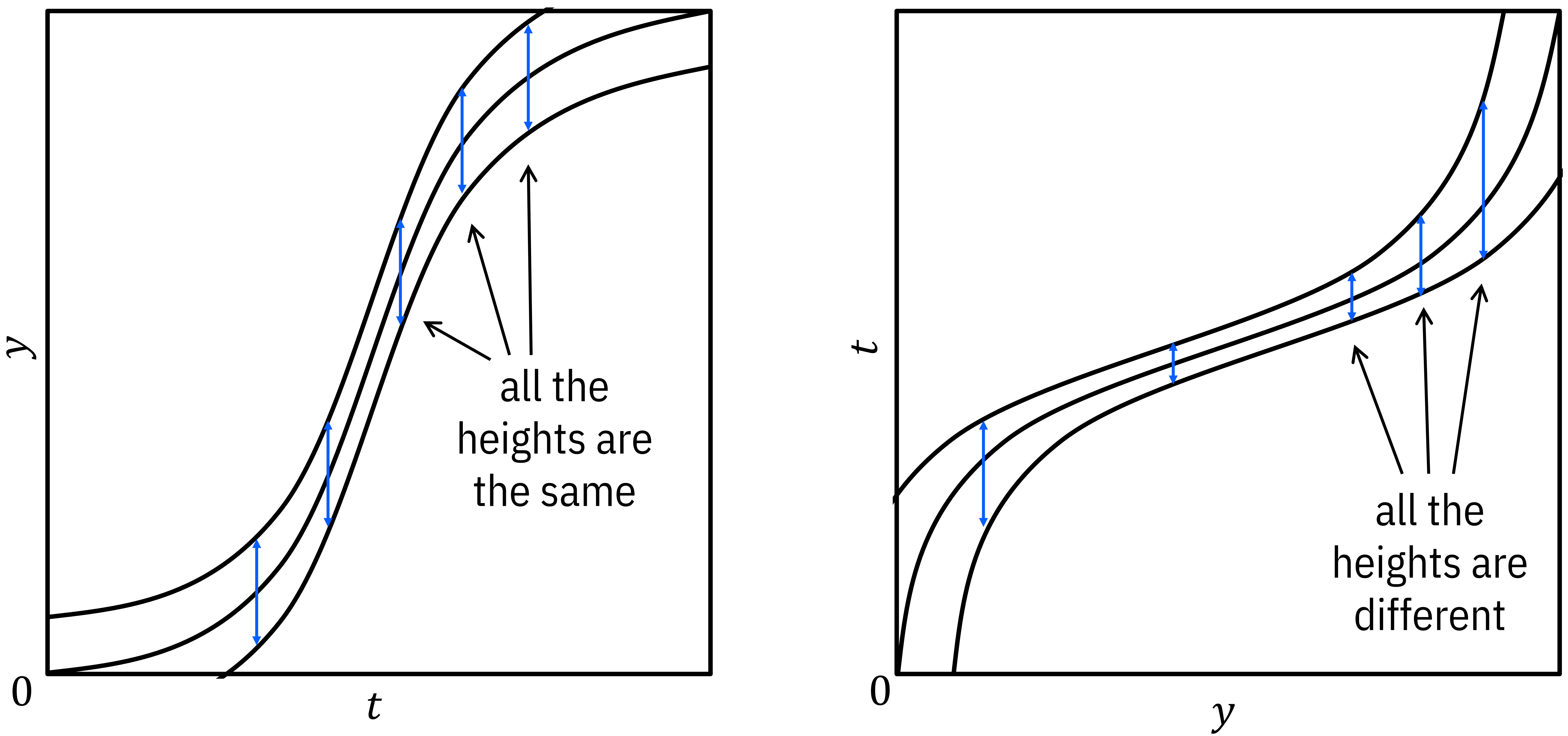
functional model-based methods

treatment models

outcome models

Functional-model based causal discovery (additive noise model assumption)

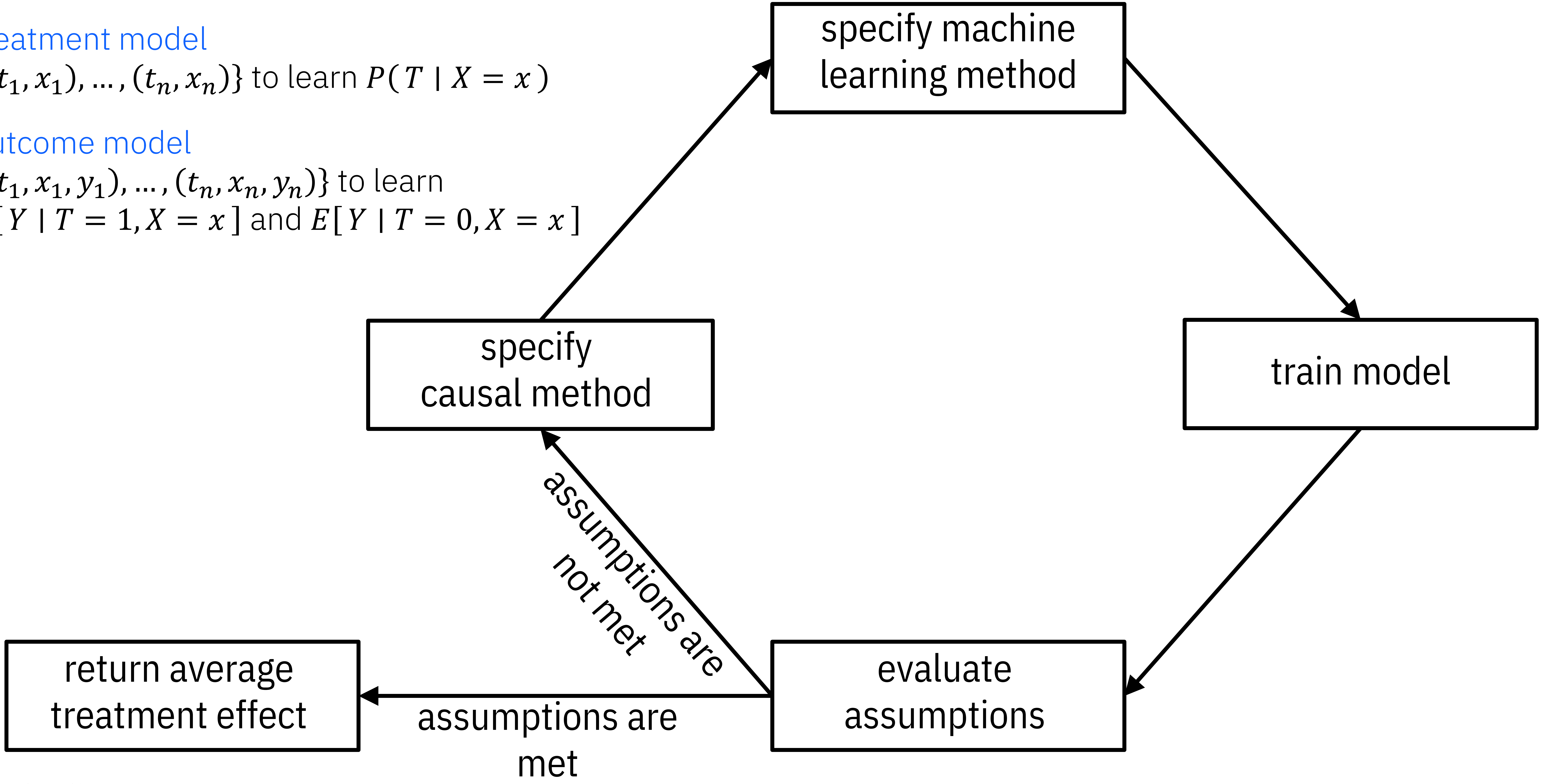
$P(Y \mid do(t)) = f_Y(t) + noise_Y$



Treatment effect estimation

treatment model
$\{(t_1, x_1), \ldots, (t_n, x_n)\}$ to learn $P(\,T \mid X = x\,)$

outcome model
$\{(t_1, x_1, y_1), \ldots, (t_n, x_n, y_n)\}$ to learn
$E[\,Y \mid T = 1, X = x\,]$ and $E[\,Y \mid T = 0, X = x\,]$



specify machine learning method

train model

evaluate assumptions

assumptions are not met

specify causal method

assumptions are met

return average treatment effect

Example

https://github.com/BiomedSciAI/causallib/blob/master/examples/lalonde.ipynb

Part 4
Reliability

A trustworthy machine learning system is one that has sufficient:

1 basic performance

2 reliability

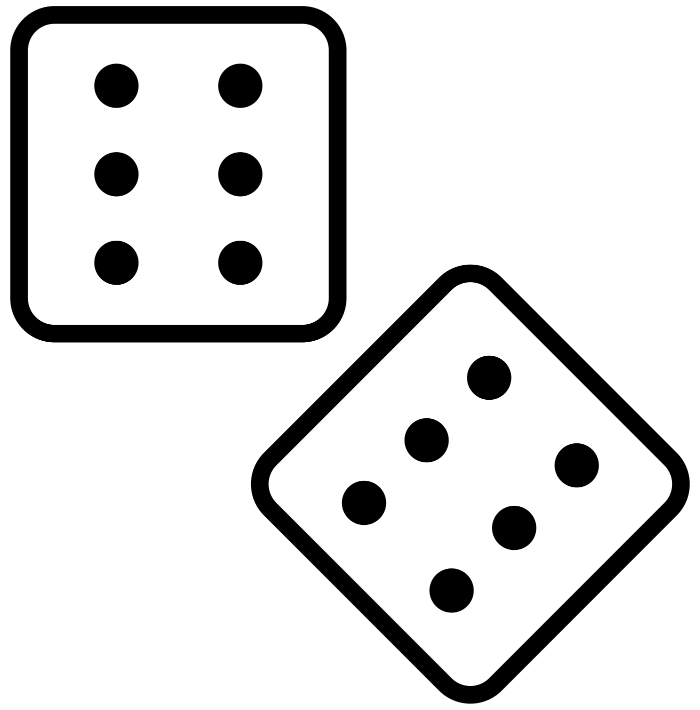
3 human interaction

4 aligned purpose

Two types of uncertainty and the two components of safety

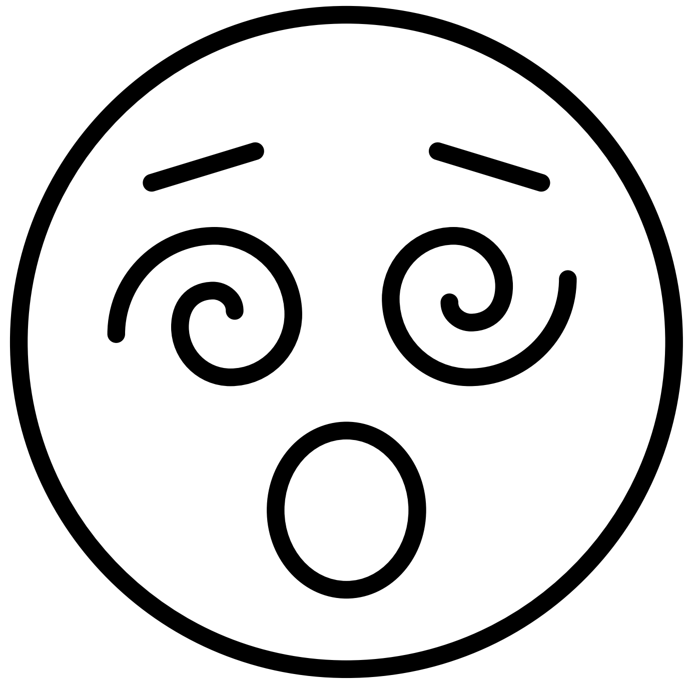
Aleatoric

- Inherent randomness
- Quantified through probability theory
- Relates to basic performance attribute of trustworthiness



Epistemic

- Observer-dependent lack of knowledge
- Quantified through possibility theory
- Relates to reliability attribute of trustworthiness



construct space

measurement

observed space

sampling

raw data space

data preparation

prepared data space

modeling

prediction space

social bias

threatens

construct validity

representation bias

temporal bias

threatens

external validity

data preparation bias

data poisoning

threatens

internal validity

underfitting/ overfitting

poor inductive bias

threatens

generalization

construct space

measurement

observed space

sampling

raw data space

data preparation

prepared data space

modeling

prediction space

social bias

representation bias

temporal bias

data preparation bias

data poisoning

underfitting/ overfitting

poor inductive bias

threatens

construct validity

external validity

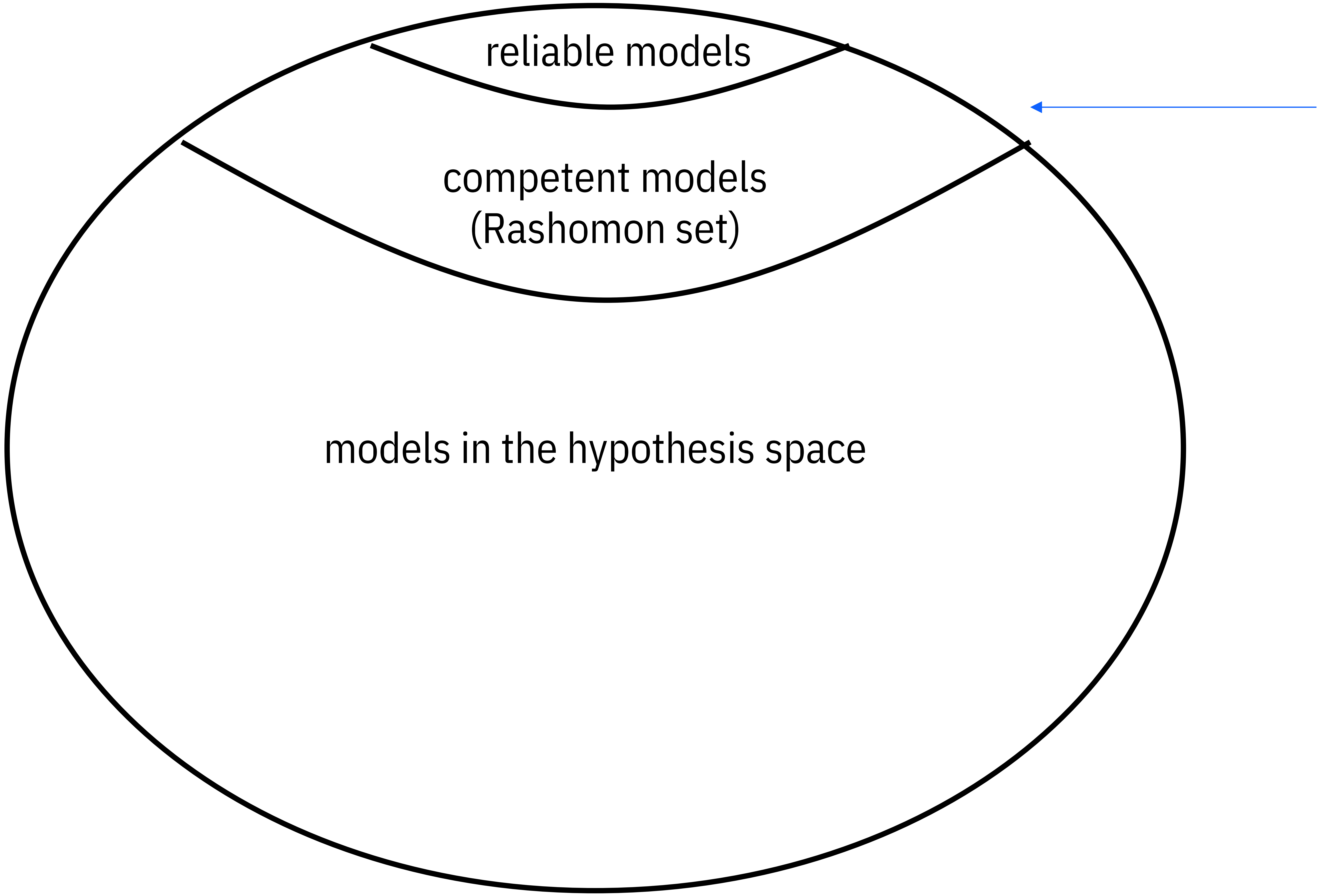
internal validity

generalization

Uncertainty in modeling

If you don’t have the information to select a good inductive bias and hypothesis space, but you could obtain it in principle, then you have epistemic uncertainty.

Rashomon effect or underspecification


reliable models
competent models (Rashomon set)
models in the hypothesis space

All models in the Rashomon set have the possibility of being good models with respect to aleatoric uncertainty or risk, but have different ways of generalizing.

Some of the models are unreliable because they take shortcuts and generalize based on spurious characteristics in the data. They are not causal.

Try to include more constraints in the problem specification.

If you don’t have enough high-quality data to train the classifier even if you have the perfect hypothesis space, you have epistemic uncertainty.

The main way to reduce epistemic uncertainty is data augmentation.

Uncertainty in data

Chapter 9
Distribution shift

“All bets are off if there is a distribution shift when the model is deployed. (There's always a distribution shift.)”

—Arvind Narayanan, computer scientist at Princeton University

The big lie of machine learning

The core assumption that training data and testing data is independent and identically distributed (i.i.d.).

This is almost never true in the real world, where there tends to be some sort of difference in the probability distributions of the training data and the data encountered during the model's deployment.

Distribution shift is the difference in the probability distribution between training data and data encountered during deployment.

A competent model that achieves high accuracy when tested through cross-validation might not maintain that competence in the real world.

construct space

measurement

observed space

sampling

raw data space

data preparation

prepared data space

modeling

prediction space

social bias

representation bias

temporal bias

data preparation bias

data poisoning

underfitting/ overfitting

poor inductive bias

threatens

construct validity

external validity

internal validity

generalization

Prior probability shift, also known as label shift, is when the label distributions are different but the features given the labels are the same:

$p_Y^{(train)}(y) \neq p_Y^{(deploy)}(y)$ and

$p_{X|Y}^{(train)}(\, x \mid y \,) = p_{X|Y}^{(deploy)}(\, x \mid y \,)$.

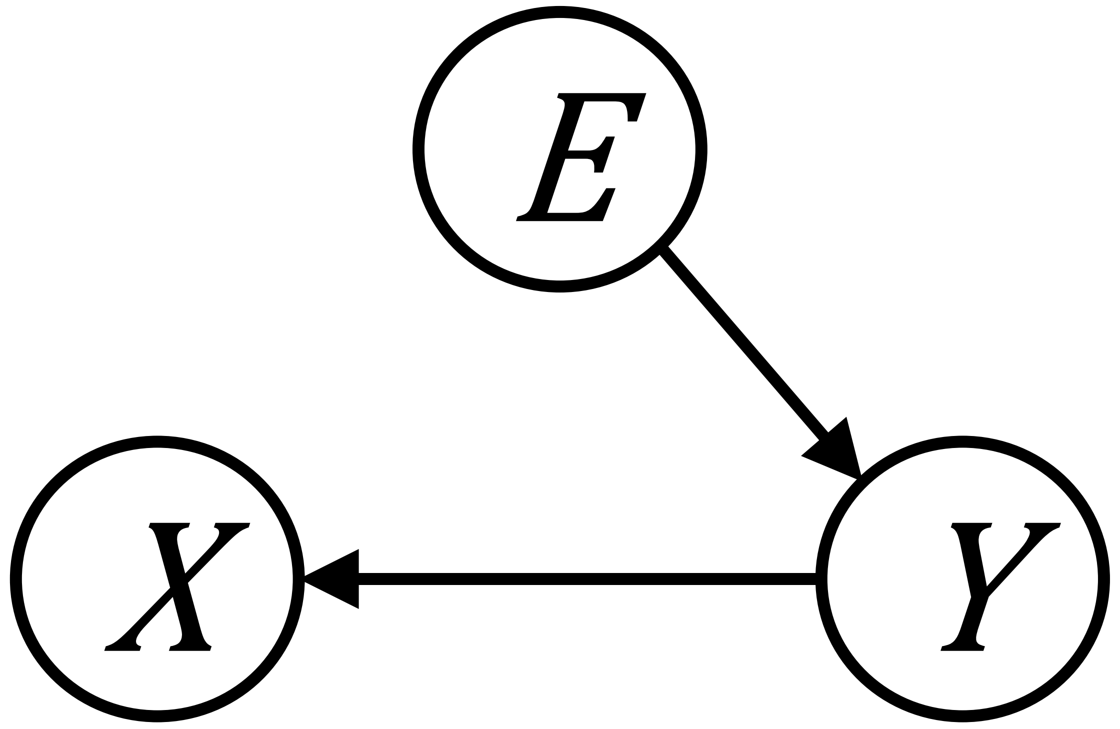
Covariate shift is when the feature distributions are different but the labels given the features are the same: $p_X^{(train)}(x) \neq p_X^{(deploy)}(x)$ and $p_{Y|X}^{(train)}(\, y \mid x \,) = p_{Y|X}^{(deploy)}(\, y \mid x \,)$.

construct space
measurement
observed space
sampling
raw data space
data preparation
prepared data space
modeling
prediction space
social bias
representation bias
temporal bias
data preparation bias
data poisoning
underfitting/ overfitting
poor inductive bias
threatens
threatens
threatens
threatens
construct validity
external validity
internal validity
generalization

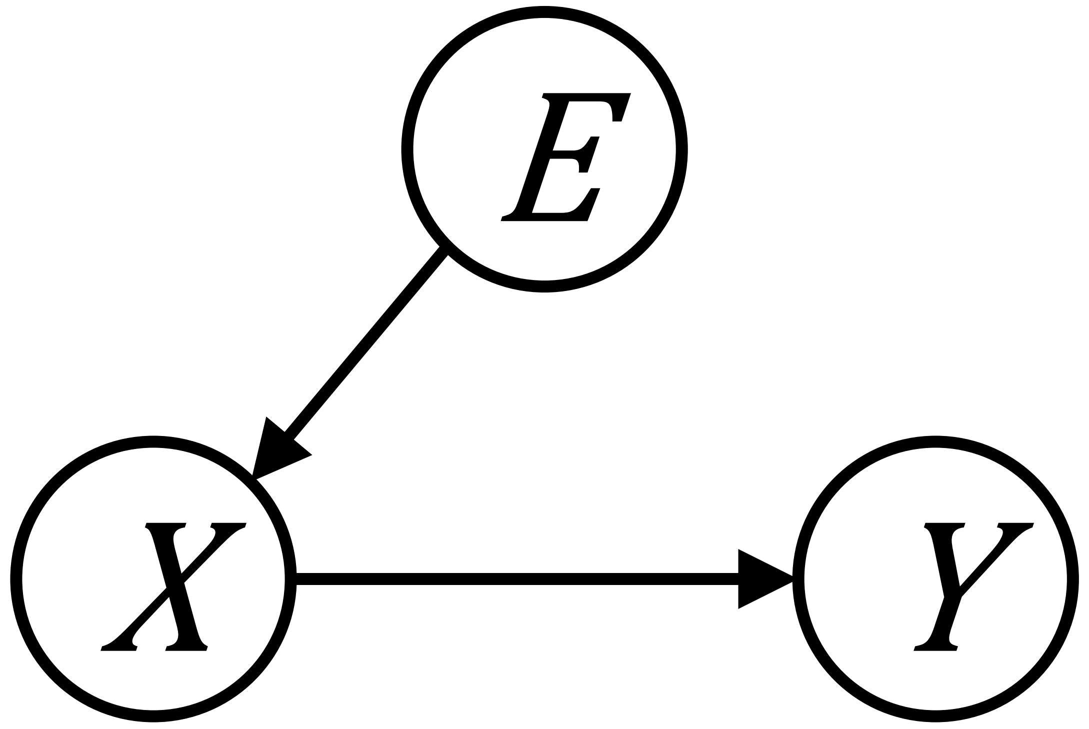
Concept drift is when the labels given the features are different but the features are the same: $p_{Y|X}^{(train)}(\,y \mid x\,) \neq p_{Y|X}^{(deploy)}(\,y \mid x\,)$ and $p_X^{(train)}(x) = p_X^{(deploy)}(x)$,

or when the features given the labels are different but the labels are the same: $p_{X|Y}^{(train)}(\,x \mid y\,) \neq p_{X|Y}^{(deploy)}(\,x \mid y\,)$ and $p_Y^{(train)}(y) = p_Y^{(deploy)}(y)$.

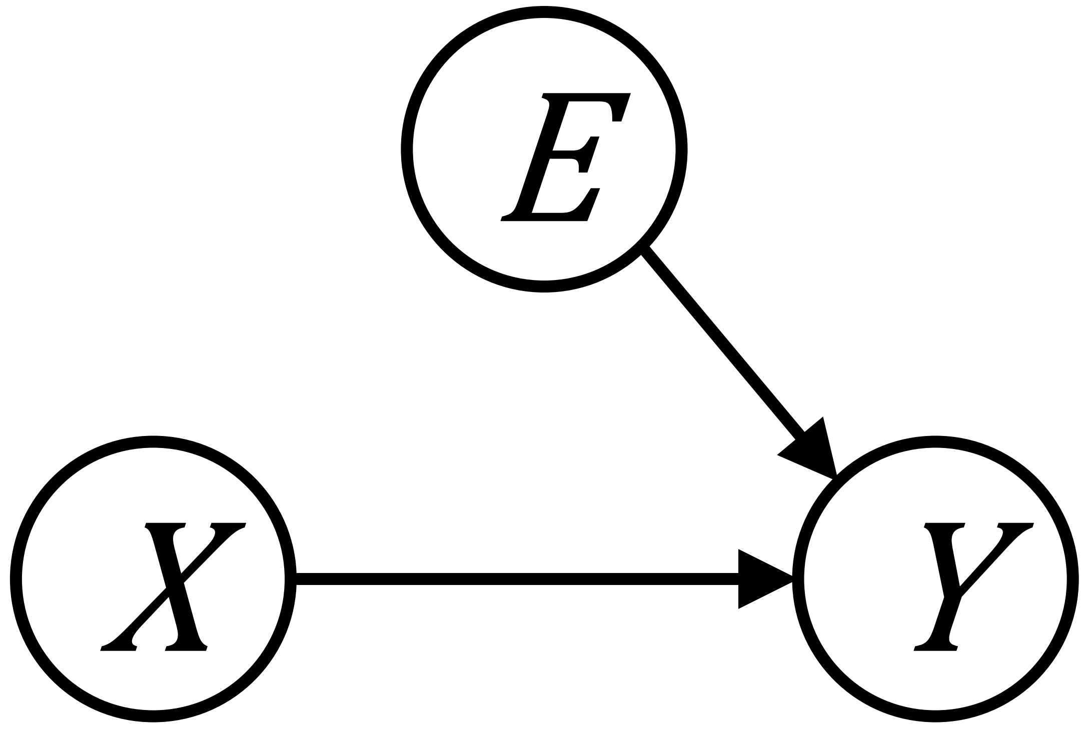
What does changing the environment in which the data was measured and sampled do to the features and label?



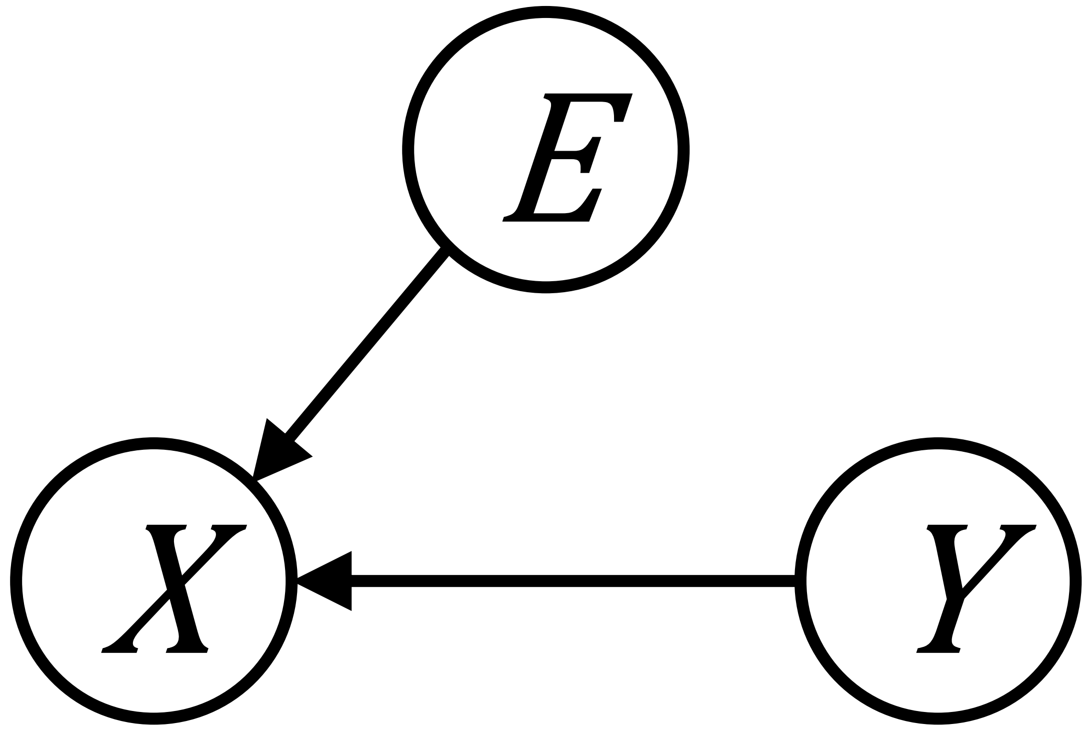
prior probability shift



covariate shift



concept drift



concept drift

Type	What Changes	What is the Same	Source	Threatens	Learning Problem
prior probability shift	Y	$X \mid Y$	sampling	external validity	anticausal learning
covariate shift	X	$Y \mid X$	sampling	external validity	causal learning
concept drift	$Y \mid X$	X	measurement	construct validity	causal learning
	$X \mid Y$	Y			anticausal learning

Nuanced point. When you have prior probability shift, the label causes the feature and when you have covariate shift, the features cause the label. This is weird to think about, so let’s slow down and work through this concept.

In the first case, $Y \rightarrow X$, the label is known as an *intrinsic* label and the machine learning problem is known as *anticausal learning*. A prototypical example is a disease with a known pathogen like malaria that causes specific symptoms like chills, fatigue, and fever. The label of a patient having a disease is intrinsic because it is a basic property of the infected patient, which then causes the observed features.

In the second case, $X \rightarrow Y$, the label is known as an *extrinsic* label and the machine learning problem is known as *causal learning*. A prototypical example of this case is a syndrome, a collection of symptoms such as Asperger’s that isn’t tied to a pathogen. The label is just a label to describe the symptoms like compulsive behavior and poor coordination; it doesn’t cause the symptoms.

Normally, in the practice of doing supervised machine learning, the distinction between anticausal and causal learning is just a curiosity, but it becomes important when figuring out what to do to mitigate the effect of distribution shift.

Mitigating distribution shift

Type	Where in the Pipeline	Known Deployment Environment	Approach for Prior Probability and Covariate Shifts	Approach for Concept Drift
adaptation	pre-processing	yes	sample weights	obtain labels
robustness	model training	no	min-max formulation	invariant risk minimization

Adaptation to prior probability shift

1. Train a classifier on one random split of the training data to get $\hat{y}^{(train)}(x)$ and compute the classifier's confusion matrix on another random split of the training data:
   $$C = \begin{bmatrix} p_{TP} & p_{FP} \\ p_{FN} & p_{TN} \end{bmatrix}.$$
2. Run the unlabeled features of the deployment data through the classifier: $\hat{y}^{(train)}\left(X^{(deploy)}\right)$ and compute the probabilities of positives and negatives in the deployment data as a vector:
   $$a = \begin{bmatrix} P\left(\hat{y}^{(train)}\left(X^{(deploy)}\right) = 1\right) \\ P\left(\hat{y}^{(train)}\left(X^{(deploy)}\right) = 0\right) \end{bmatrix}.$$
3. Compute weights $w = C^{-1}a$. This is a vector of length two.
4. Apply the weights to the training data points in the first random split and retrain the classifier. The first of the two weights multiplies the loss function of the training data points with label 1. The second of the two weights multiplies the loss function of the training data points with label 0.

Adaptation to covariate shift

$$\hat{y}(\cdot) = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{j=1}^{n} w_j L\left(y_j, f\left(x_j\right)\right)$$

$$w_j = p_X^{(deploy)}\left(x_j\right) / p_X^{(train)}\left(x_j\right)$$

train a classifier for this

Adaptation to concept drift

Need labeled data from deployment environment.

Robustness to prior probability shift

min-max formulation $$\arg\min_{p_0^{(train)}} \max_{p_0^{(deploy)}} R\left(p_0^{(deploy)}, p_0^{(train)}\right)$$

mismatched Bayes risk $$R\left(p_0^{(deploy)}, p_0^{(train)}\right) = c_{10} p_0^{(deploy)} p_{FP}\left(p_0^{(train)}\right) + c_{01} p_1^{(deploy)} p_{FN}\left(p_0^{(train)}\right)$$

Robustness to covariate shift

min-max formulation

$$\arg\min_{f\in\mathcal{F}}\max_{w}\frac{1}{n}\sum_{j=1}^{n}w_j L\left(y_j, f\left(x_j\right)\right)$$

same importance weights as in adaptation

Robustness to concept drift

Assume features split into two types:

1. causal or stable features
2. spurious features

Don’t know which ones are which.

The causal features capture the intrinsic parts of the relationship between features and labels, and are the same set of features in different environments. This set of features is invariant across the environments.

Spurious features might be predictive in one environment or a few environments, but not universally so across environments.

invariant risk minimization

$$\hat{y}(\cdot) = \arg\min_{f\in\mathcal{F}} \sum_{e\in\mathcal{E}} \frac{1}{n_e}\sum_{j=1}^{n_e} L\left(y_j^{(e)}, f\left(x_j^{(e)}\right)\right).$$

$$\text{such that } f \in \arg\min_{g\in\mathcal{F}} \frac{1}{n_e}\sum_{j=1}^{n_e} L\left(y_j^{(e)}, g\left(x_j^{(e)}\right)\right) \text{ for all } e \in \mathcal{E}$$

extra specification or constraint to exclude some of the Rashomon set

works for anticausal learning formulations

Chapter 10
Fairness

“If humans didn’t behave the way we do there would be no behavior data to correct. The training data is society.”

— M. C. Hammer, musician and technology consultant

“Most of this work is narrow in scope, focusing on fine-tuning specific models, making datasets more inclusive/representative, and ‘debiasing’ datasets. Although such work can constitute part of the remedy, a fundamentally equitable path must examine the wider picture, such as unquestioned or intuitive assumptions in datasets, current and historical injustices, and power asymmetries.”

—Abeba Birhane, cognitive scientist at University College Dublin

Fairness and justice are almost synonymous, and are political.

There are several kinds of justice, including (1) distributive justice, (2) procedural justice, (3) restorative justice, and (4) retributive justice.

Distributive justice is equality in what people receive—the outcomes.

Procedural justice is sameness in the way it is decided what people receive.

Restorative justice repairs a harm.

Retributive justice seeks to punish wrongdoers.

Unfairness is when an allocation decision gives a systematic advantage to certain privileged groups and individuals and a systematic disadvantage to certain unprivileged groups and individuals.

Privileged groups and individuals are defined to be those who have historically been more likely to receive the favorable label.

Privileged and unprivileged groups are delineated by protected attributes such as race, ethnicity, gender, religion, and age.

There is no one universal set of protected attributes.

Two types of fairness

Group fairness

- The average classifier behavior should be the same across groups defined by protected attributes

Individual fairness

- Individuals similar in their features should receive similar model predictions

construct space
measurement
observed space
sampling
raw data space
data preparation
prepared data space
modeling
prediction space
social bias
representation bias
temporal bias
data preparation bias
data poisoning
underfitting/ overfitting
poor inductive bias
threatens
threatens
threatens
threatens
construct validity
external validity
internal validity
generalization

Example from US health care system

Social bias	Representation bias	Data preparation bias
Utilization vs. infirmity	Using data only from men in Boston	Summing in-patient, out-patient, and emergency room costs

Differences between distribution shift and fairness

Access to the construct space

You can get data from the construct space in distribution shift scenarios. Maybe not immediately, but if you wait, collect, and label data from the deployment environment, you will have data reflecting the construct space.

You never have access to the construct space in fairness settings. The construct space reflects a perfect egalitarian world that does not exist in real life, so you can't get data from it.

Specification of what is sought

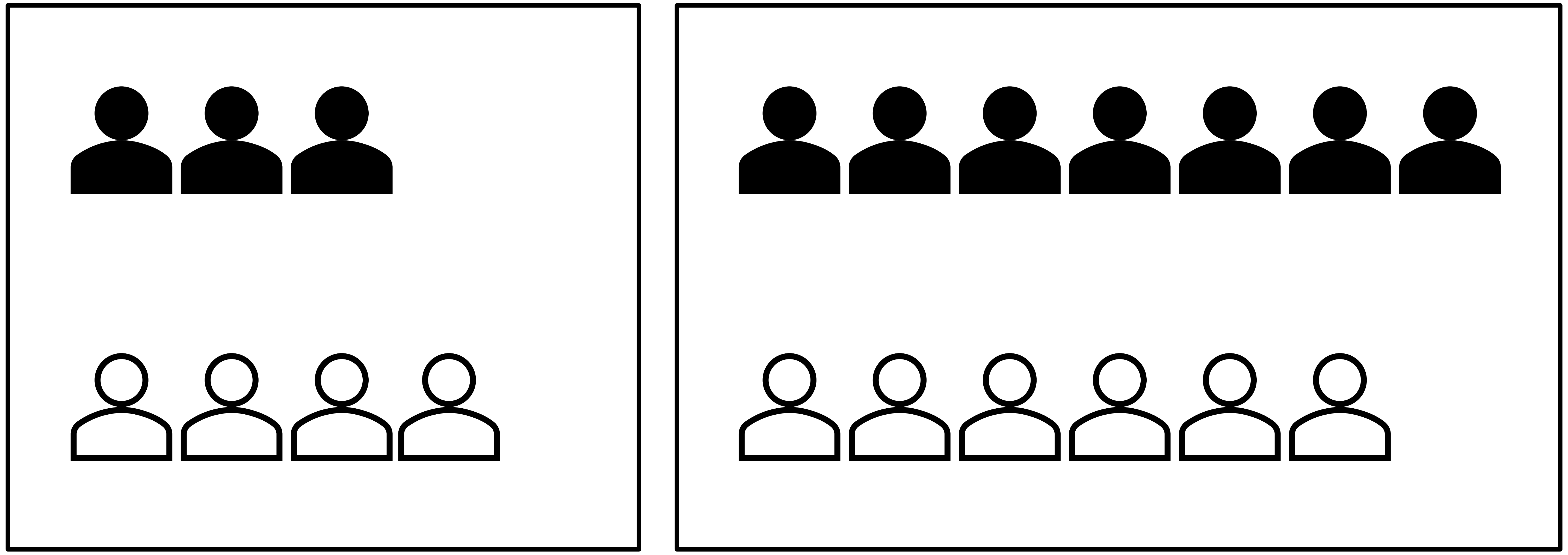
In distribution shift, there is no further specification beyond just trying to match the shifted distribution.

In fairness, there are precise policy-driven notions and quantitative criteria that define the desired state of data and/or models that are not dependent on the data distribution you have.

Group fairness metrics

Type	Statistical Relationship	Fairness Metric	Can Be A Dataset Metric?	Social Bias in Measurement	Favorable Label
independence	$\hat{Y} \perp\!\!\!\perp Z$	statistical parity difference	yes	yes	assistive or non-punitive
separation	$\hat{Y} \perp\!\!\!\perp Z \mid Y$	average odds difference	no	no	assistive
sufficiency (calibration)	$Y \perp\!\!\!\perp Z \mid \hat{Y}$	average predictive value difference	no	no	non-punitive

statistical parity difference $= P(\hat{y}(X) = \text{fav} \mid Z = \text{unpr}) - P(\hat{y}(X) = \text{fav} \mid Z = \text{priv})$



unprivileged group

privileged group

predicted favorable label

predicted unfavorable label

$3/10$

$-$

$4/10$

$=$

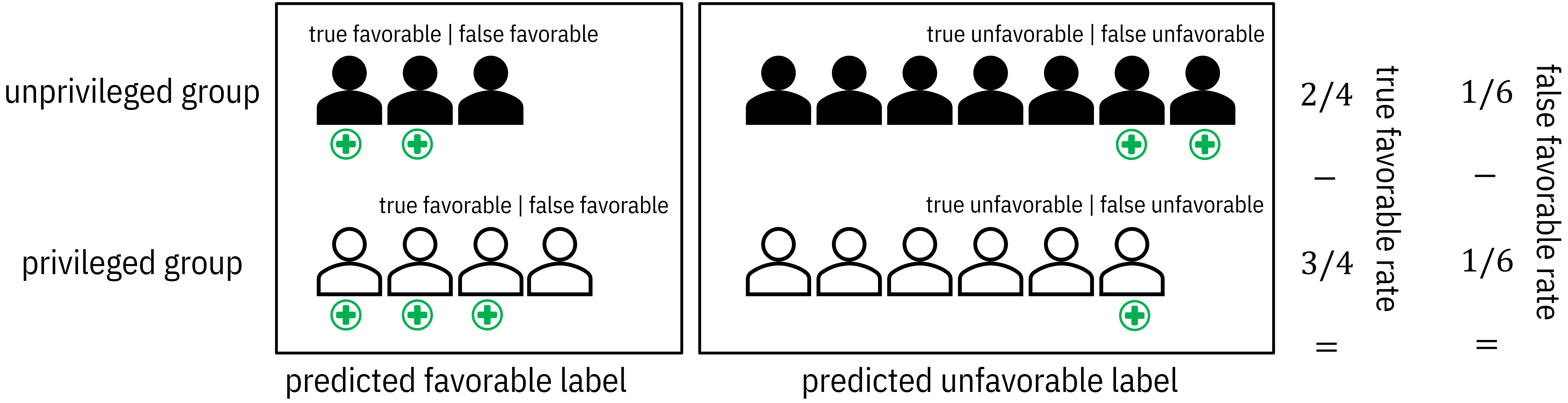
-0.1

selection rate

statistical
parity
difference

average odds difference

$$= \tfrac{1}{2}\Big(P(\hat{y}(X) = \text{fav} \mid Y = \text{fav},\ Z = \text{unpr}) - P(\hat{y}(X) = \text{fav} \mid Y = \text{fav},\ Z = \text{priv})\Big)$$
$$+ \tfrac{1}{2}\Big(P(\hat{y}(X) = \text{fav} \mid Y = \text{unf},\ Z = \text{unpr}) - P(\hat{y}(X) = \text{fav} \mid Y = \text{unf},\ Z = \text{priv})\Big)$$



unprivileged group

privileged group

true favorable | false favorable

true unfavorable | false unfavorable

predicted favorable label

predicted unfavorable label

true favorable rate

$$2/4 - 3/4 =$$

false favorable rate

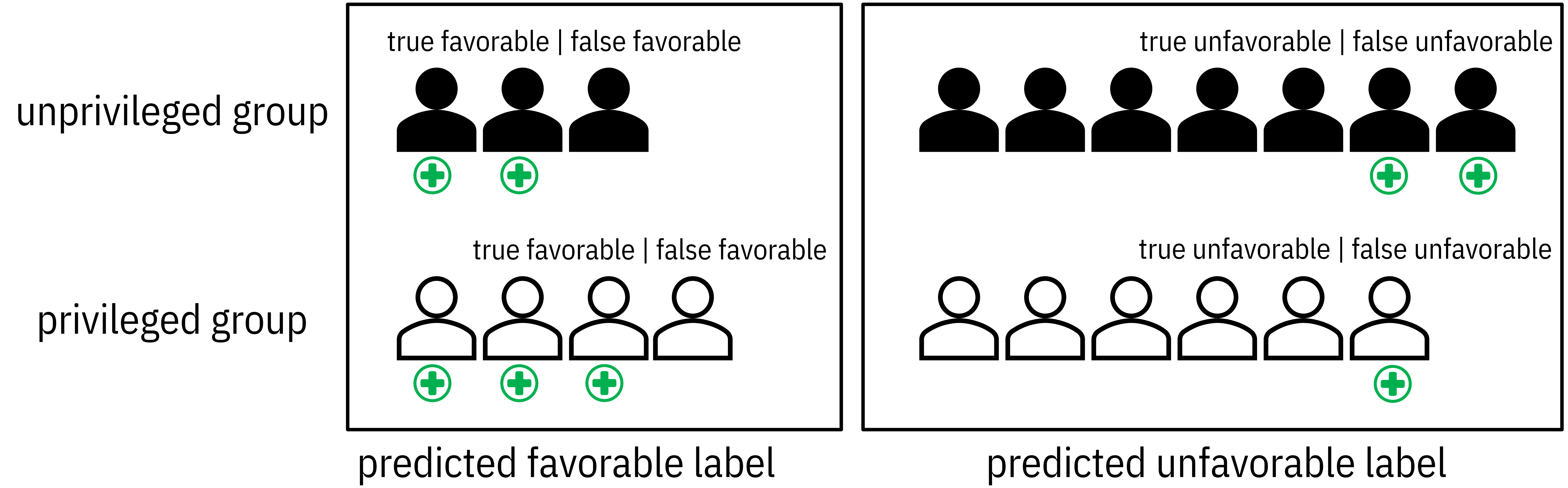
$$1/6 - 1/6 =$$

$$\frac{-0.25 + 0.00}{2} = -0.125$$

average odds difference

average predictive value difference

$$= \tfrac{1}{2}\big(P(\,Y = \text{fav} \mid \hat{y}(X) = \text{fav},\ Z = \text{unpr}\,) - P(\,Y = \text{fav} \mid \hat{y}(X) = \text{fav},\ Z = \text{priv}\,)\big)$$
$$+\, \tfrac{1}{2}\big(P(\,Y = \text{fav} \mid \hat{y}(X) = \text{unf},\ Z = \text{unpr}\,) - P(\,Y = \text{fav} \mid \hat{y}(X) = \text{unf},\ Z = \text{priv}\,)\big)$$



unprivileged group

privileged group

true favorable | false favorable

true unfavorable | false unfavorable

true favorable | false favorable

true unfavorable | false unfavorable

predicted favorable label

predicted unfavorable label

positive predictive value

$$2/3 - 3/4 = -0.08$$

false omission rate

$$2/7 - 1/6 = 0.12$$

$$\frac{-0.08 + 0.12}{2} = 0.02$$

average predictive value difference

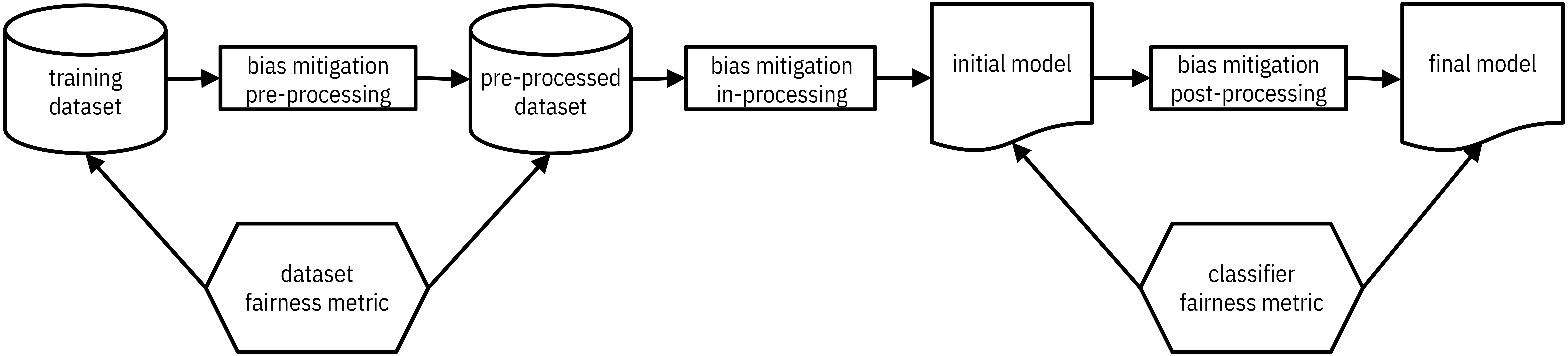
Individual fairness

$$\text{consistency} = 1 - \frac{1}{n}\sum_{j=1}^{n}\left|\hat{y}_j - \frac{1}{k}\sum_{j\prime\in\mathcal{N}_k(x_j)}\hat{y}_{j\prime}\right|$$

$do(Z)$

if $\hat{Y}$ remains the same, counterfactual fairness

check using average treatment effect estimation



training dataset

bias mitigation pre-processing

pre-processed dataset

bias mitigation in-processing

initial model

bias mitigation post-processing

final model

dataset fairness metric

classifier fairness metric

Example methods

Don’t drop protected attributes – other features can recreate the information.

Aim for independence between label and protected attribute.

Bias mitigation pre-processing

$$w_j = \frac{p_Y(y_j)p_Z(z_j)}{p_{Y,Z}(y_j, z_j)}$$

Bias mitigation in-processing

$$arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{j=1}^{n} L\left(y_j, f(x_j)\right) + \lambda J(f)$$

Bias mitigation post-processing

Flip predictions $\hat{Y}$ to meet the group fairness metric you desire.

Demo

https://aif360.res.ibm.com/data

Measuring classification accuracy on data from the prepared data space, which already contains social bias, representation bias, and data preparation bias is not the right thing to do.

Just like you should measure performance of distribution shift adaptation on data from the new environment—its construct space, you should measure accuracy after bias mitigation in its construct space where there is no unfairness.

There is a tradeoff between fairness and accuracy measured in the prepared data space, but importantly there is no tradeoff between accuracy and fairness in the construct space.

Have we too easily swept the important considerations of algorithmic fairness under the rug of mathematics? Yes and no. If you have truly thought through the different sources of inequity arising throughout the machine learning lifecycle utilizing a panel of diverse voices, then applying the quantitative metrics and mitigation algorithms is actually pretty straightforward. It is straightforward because of the hard work you’ve done before getting to the modeling phase of the lifecycle and you should feel confident in going forward. If you have not done the hard work earlier in the lifecycle, blindly applying bias mitigation algorithms might not reduce harms and can even exacerbate them. So don’t take shortcuts.

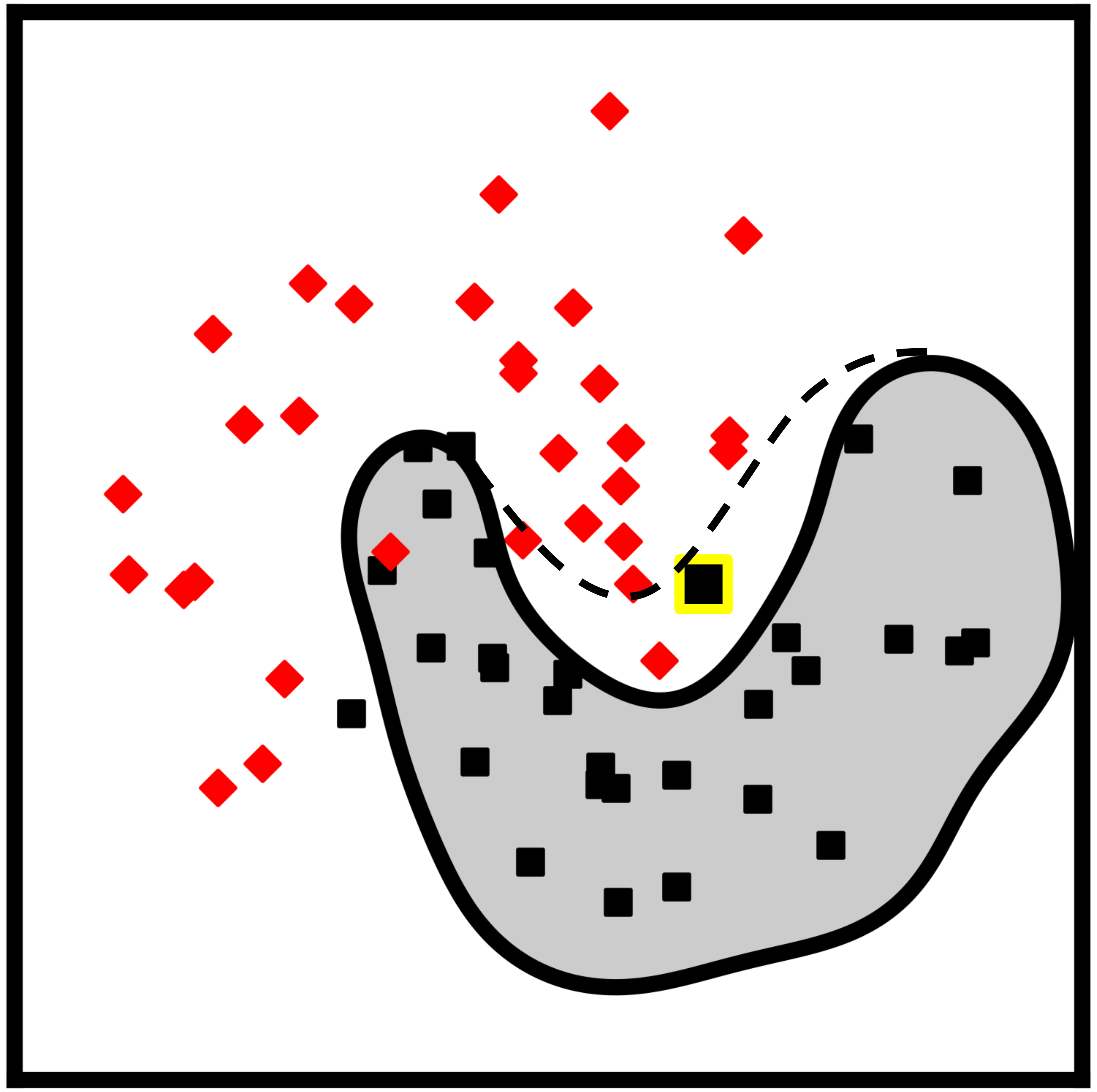
Chapter 11
Adversarial robustness

“In my view, similar to car model development and manufacturing, a comprehensive ‘in-house collision test’ for different adversarial threats on an AI model should be the new norm to practice to better understand and mitigate potential security risks.”

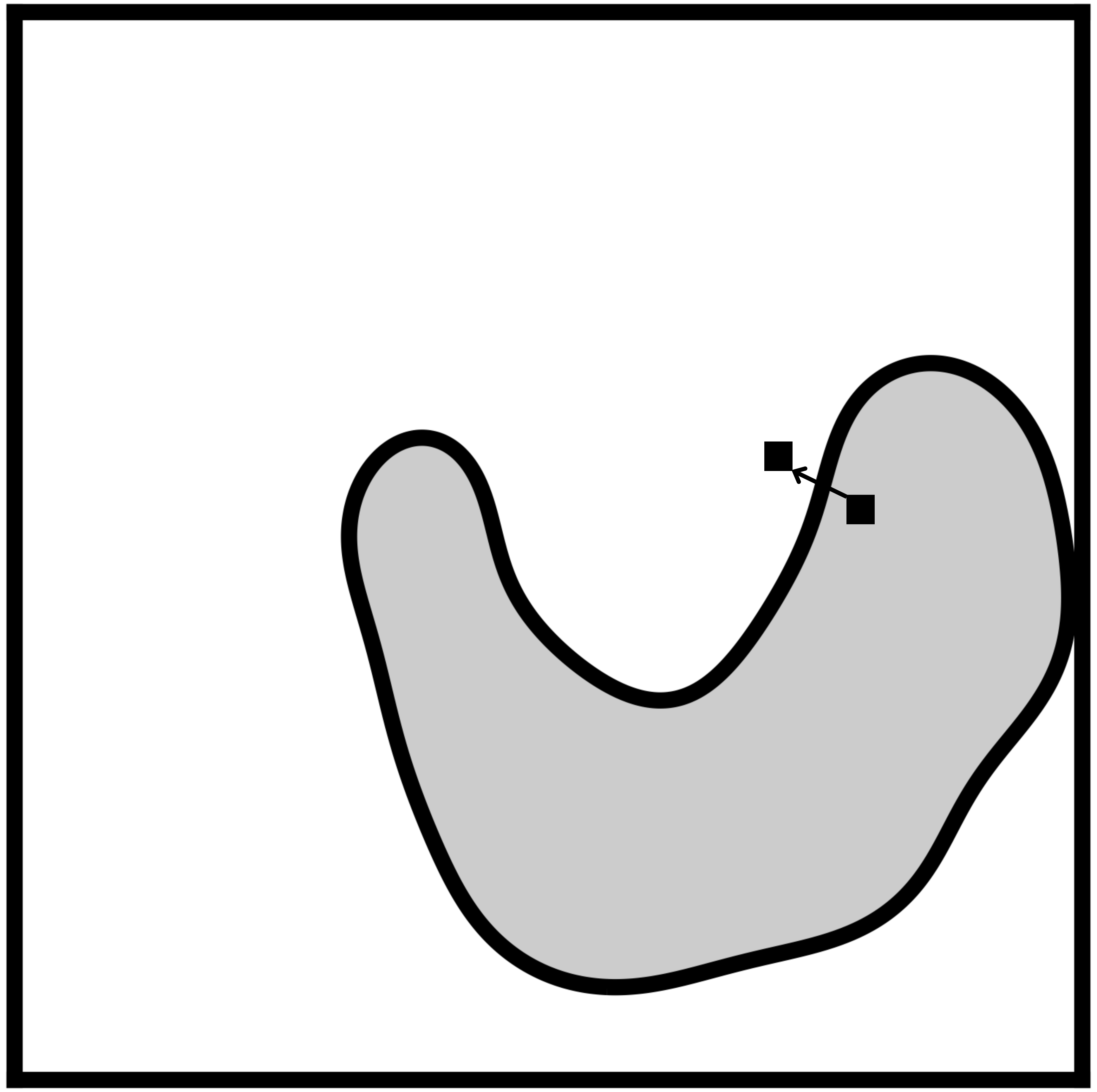
—Pin-Yu Chen, computer scientist at IBM Research

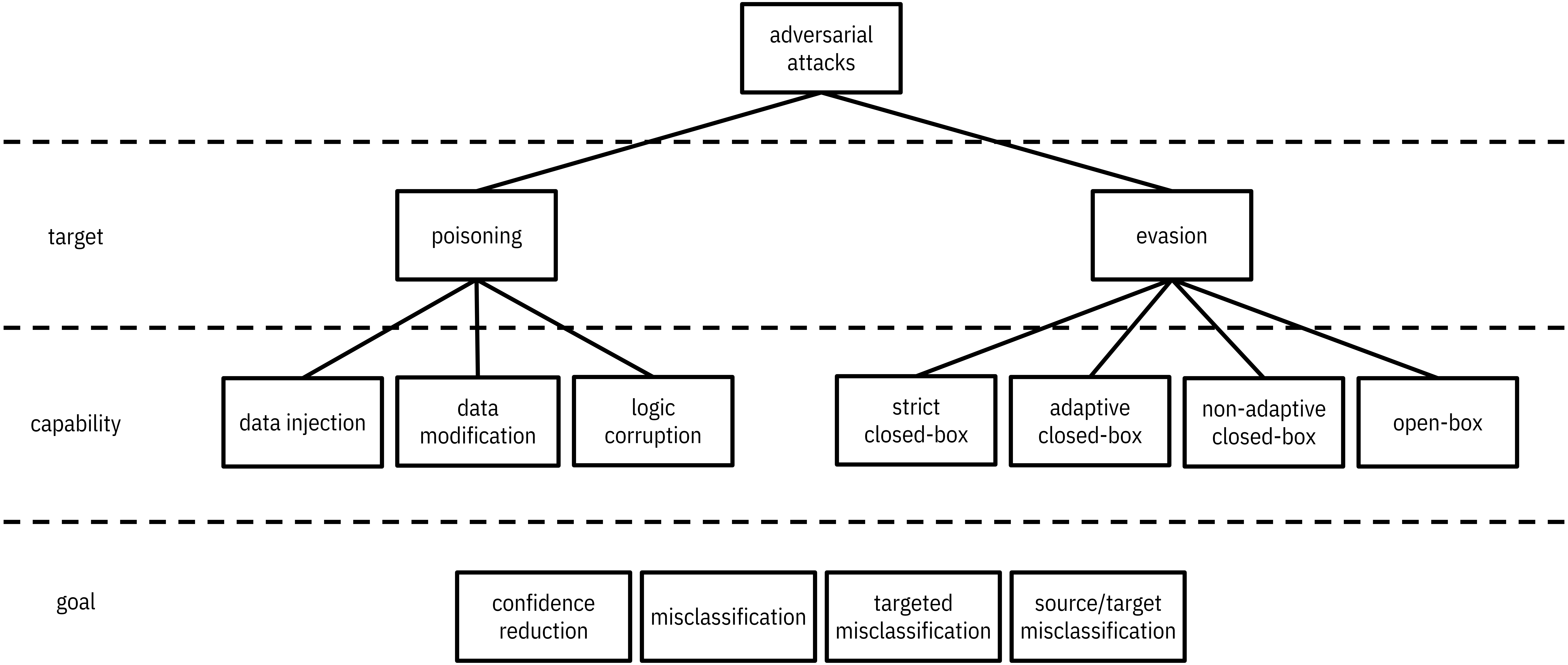
Adversaries are people trying to achieve their own goals to the detriment of the goals of the system designers, usually in a secretive way.

data poisoning



evasion




adversarial attacks

target

poisoning

evasion

capability

data injection

data modification

logic corruption

strict closed-box

adaptive closed-box

non-adaptive closed-box

open-box

goal

confidence reduction

misclassification

targeted misclassification

source/target misclassification

construct space

measurement

observed space

sampling

raw data space

data preparation

prepared data space

modeling

prediction space

social bias

threatens

construct validity

representation bias

temporal bias

threatens

external validity

data preparation bias

data poisoning

threatens

internal validity

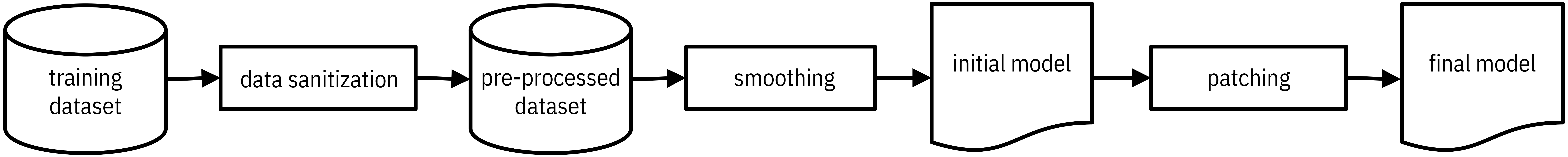
underfitting/ overfitting

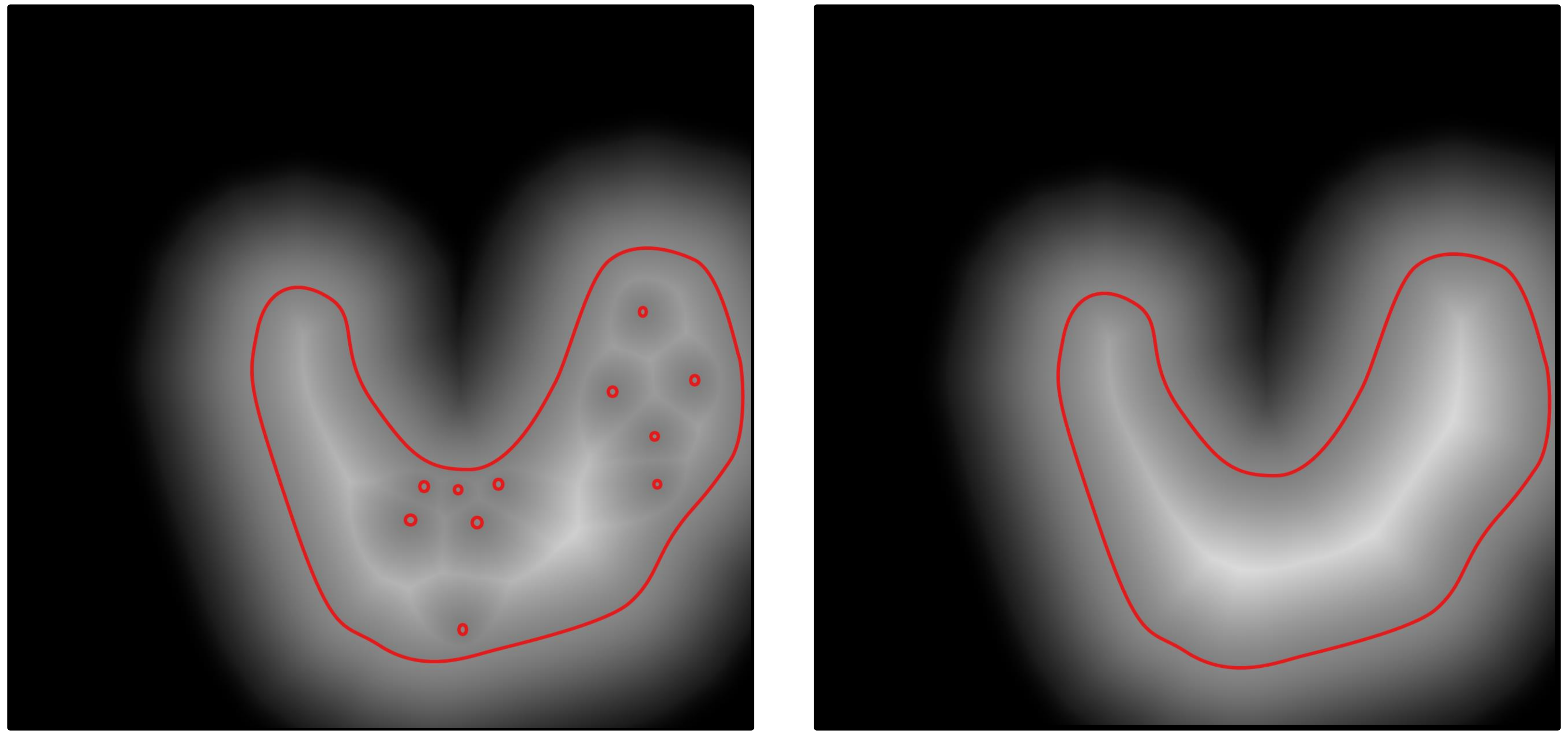
poor inductive bias

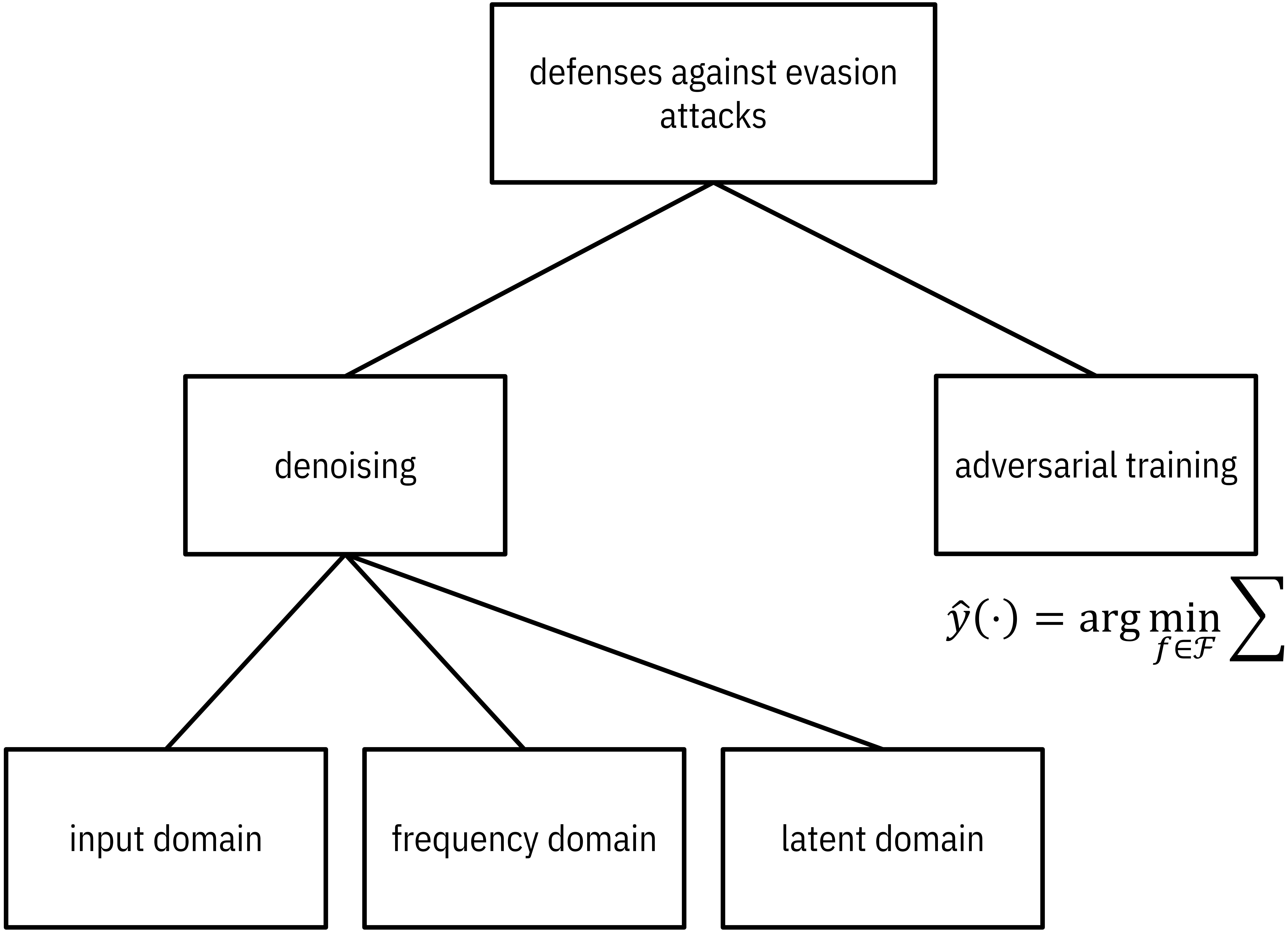
threatens

generalization

Poisoning defenses







defenses against evasion attacks

denoising

adversarial training

input domain

frequency domain

latent domain

$$\hat{y}(\cdot) = \arg\min_{f\in\mathcal{F}} \sum_{j=1}^{n} \max_{\|\delta_j\|\le\epsilon} L\left(y_j, f\left(x_j+\delta_j\right)\right)$$

Demo

https://art360.res.ibm.com/demos

Part 5
Interaction

A trustworthy machine learning system is one that has sufficient:

1

basic performance

2

reliability

3

human interaction

4

aligned purpose

Chapter 12
Interpretability and explainability

“If we don’t know what is happening in the black box, we can’t fix its mistakes to make a better model and a better world.”

—Aparna Dhinakaran, chief product officer at Arize AI

“Simplicity is not so simple.”

—Dmitry Malioutov, computer scientist at IBM Research

Interpretability of machine learning models is the aim to let people understand how the machine makes its predictions.

It is a challenge because many of the machine learning approaches in Chapter 7 are not easy for people to understand since they have complicated functional forms.

Interpretability and explainability are a form of interaction between the machine and a human, specifically communication from the machine to the human, that allow the machine and human to collaborate in decision making.

The predicted label $\hat{Y}$ is not enough to communicate how the machine makes its predictions. An explanation is also needed.

Last mile problem

The machine is the transmitter of information and the human is the receiver or consumer of that information.

The communication process has to overcome human cognitive biases—the limitations that people have in receiving information.

construct space

measurement

observed space

sampling

raw data space

data preparation

prepared data space

modeling

prediction space

communication

perceived space

social bias

representation bias

temporal bias

data preparation bias

data poisoning

underfitting/ overfitting

poor inductive bias

cognitive bias

threatens

construct validity

external validity

internal validity

generalization

human-machine collaboration

No single kind of explanation appeals to all different potential consumers of explanations

Persona	Example	Goal
decision maker	appraiser, credit officer	(1) roughly understand the model to gain trust; (2) understand the predictions to combine with their own information to make decisions
affected user	loan applicant	understand the prediction for their own input data point and what they can do to change the outcome
regulator	model validator, government official	ensure the model is safe and compliant
data scientist	development team member	improve the model’s performance

Three dichotomies of explanations

Local vs. global

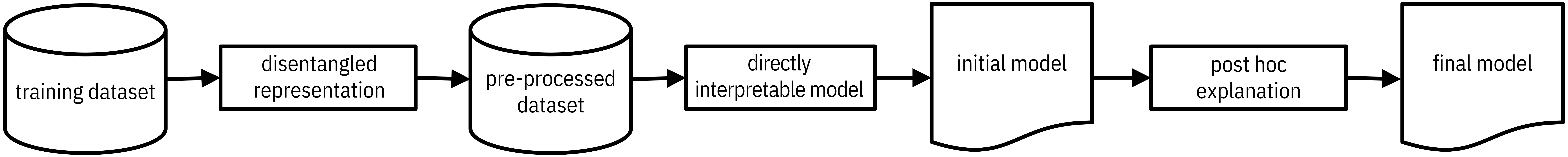
Is the consumer interested in understanding the machine predictions for individual input data points or in understanding the model overall.

Exact vs. approximate

Should the explanation be completely faithful to the underlying model or is some level of approximation allowable.

Feature- vs. sample-based

Is the explanation given as a statement about the features or is it given by pointing to other data points in their entirety.



training dataset

disentangled representation

pre-processed dataset

directly interpretable model

initial model

post hoc explanation

final model

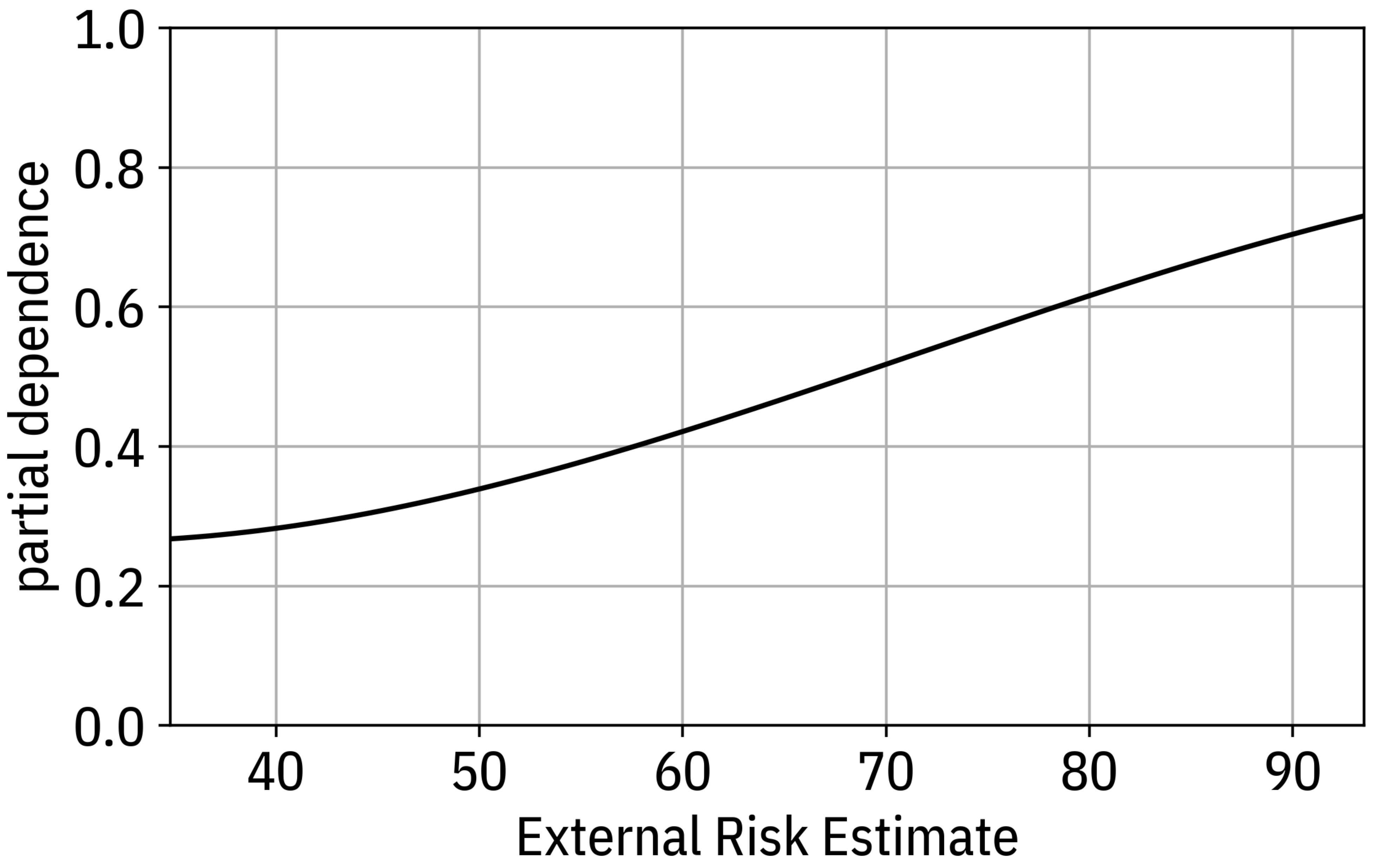
(for feature-based explanations, the features have to be meaningful)

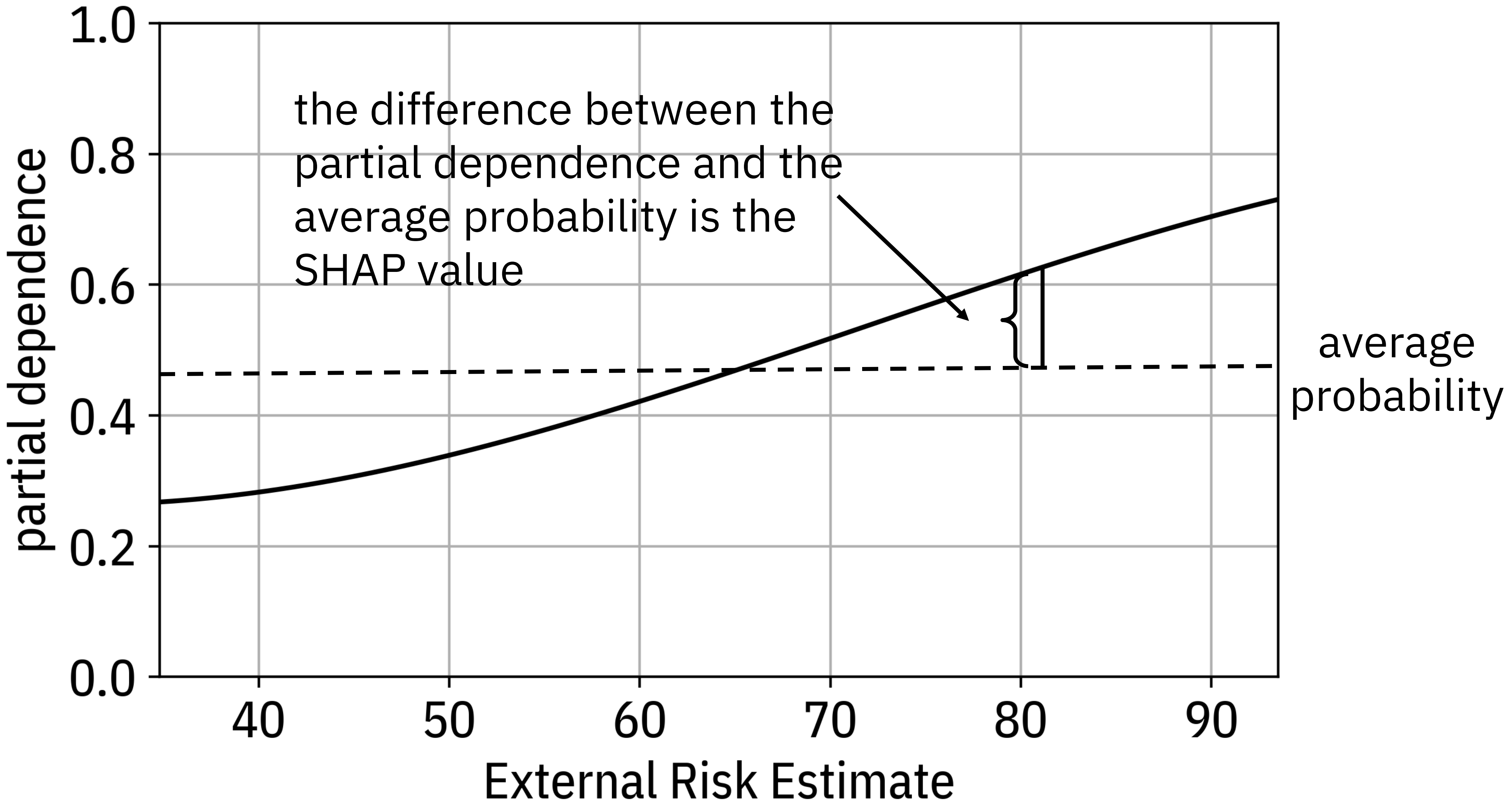
Dichotomy 1	Dichotomy 2	Dichotomy 3	Persona	Example Method
local	exact	feature-based	affected user	contrastive explanations method
local	exact	sample-based	regulator	k-nearest neighbor
local	approximate	feature-based	decision maker	LIME, SHAP, saliency map
local	approximate	sample-based	decision maker	prototype
global	exact	feature-based	regulator	decision tree, Boolean rule set, logistic regression, GAM, GLRM
global	exact	sample-based	regulator	deletion diagnostics
global	approximate	feature-based	decision maker	distillation, SRatio, partial dependence plot
global	approximate	sample-based	regulator and decision maker	influence function

Demo

https://aix360.res.ibm.com/data

Partial dependence $P\left(\hat{Y} = 1 \mid X^{(i)} = x^{(i)}\right)$





the difference between the partial dependence and the average probability is the SHAP value

average probability

Evaluating interpretability does not yield the same sort of quantitative metrics as in Part 3 for distributional robustness, fairness, and adversarial robustness. Ideally, you want to show explanations to a large set of consumers from the relevant persona performing the task the model is for and get their judgements.

Category	Consumers	Tasks
application-grounded evaluation	true persona members	real task
human-grounded evaluation	generic people	simple task
functionally-grounded evaluation	none	proxy task

Chapter 13
Transparency

“The full cycle of a machine learning project is not just modeling. It is finding the right data, deploying it, monitoring it, feeding data back [into the model], showing safety—doing all the things that need to be done [for a model] to be deployed. [That goes] beyond doing well on the test set, which fortunately or unfortunately is what we in machine learning are great at.”

—Andrew Ng, computer scientist at Stanford University

“We really need standards for what an audit is.”

—Rumman Chowdhury, machine learning ethicist at Twitter

Transparency involves the communication of qualitative information and quantitative test results from throughout the lifecycle.

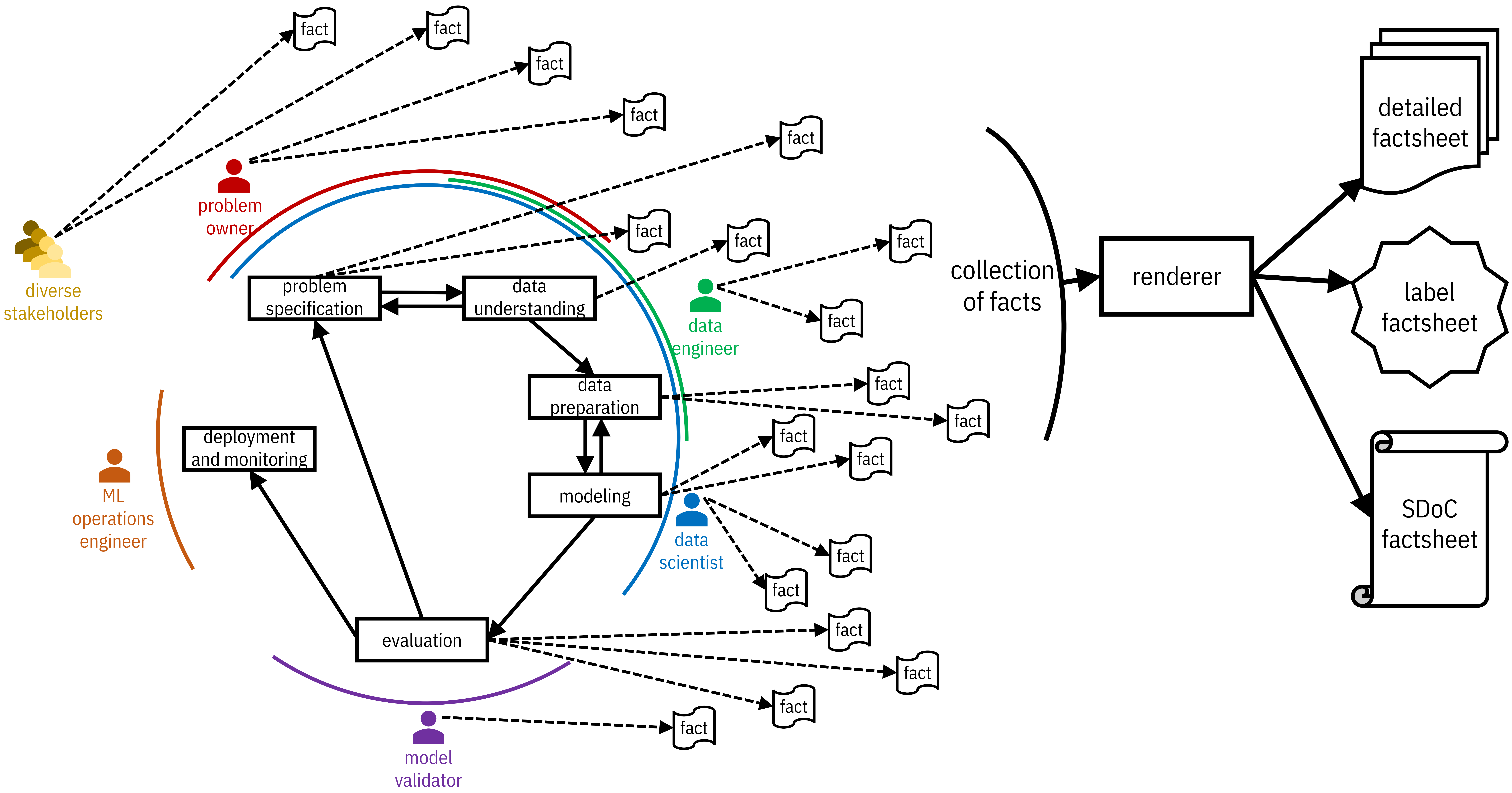
Transparency goes beyond explainability because it is focused on model performance metrics and their uncertainty, various pieces of information about the training data, and the suggested uses and possible misuses of the model. All of these pieces of information are known as facts.

One size does not fit all

Various consumers of transparent reporting are seeking different facts and level of detail.

Therefore, first run a small design exercise to understand which facts and details are relevant for the consumer.

The artifact that ultimately presents facts to a consumer is known as a factsheet.


fact
fact
fact
fact
fact
problem owner
diverse stakeholders
fact
fact
fact
problem specification
data understanding
data engineer
fact
collection of facts
renderer
detailed factsheet
label factsheet
SDoC factsheet
data preparation
fact
fact
deployment and monitoring
ML operations engineer
fact
fact
modeling
data scientist
fact
fact
evaluation
fact
fact
fact
fact
model validator

Testing machine learning systems is different from testing other kinds of software systems.

Machine learning testing suffers from the oracle problem: not knowing what the correct answer is supposed to be for a given input.

Metamorphic relations gets around the problem by not looking at a single input data point, but by looking at two or more variations that should yield the same output.

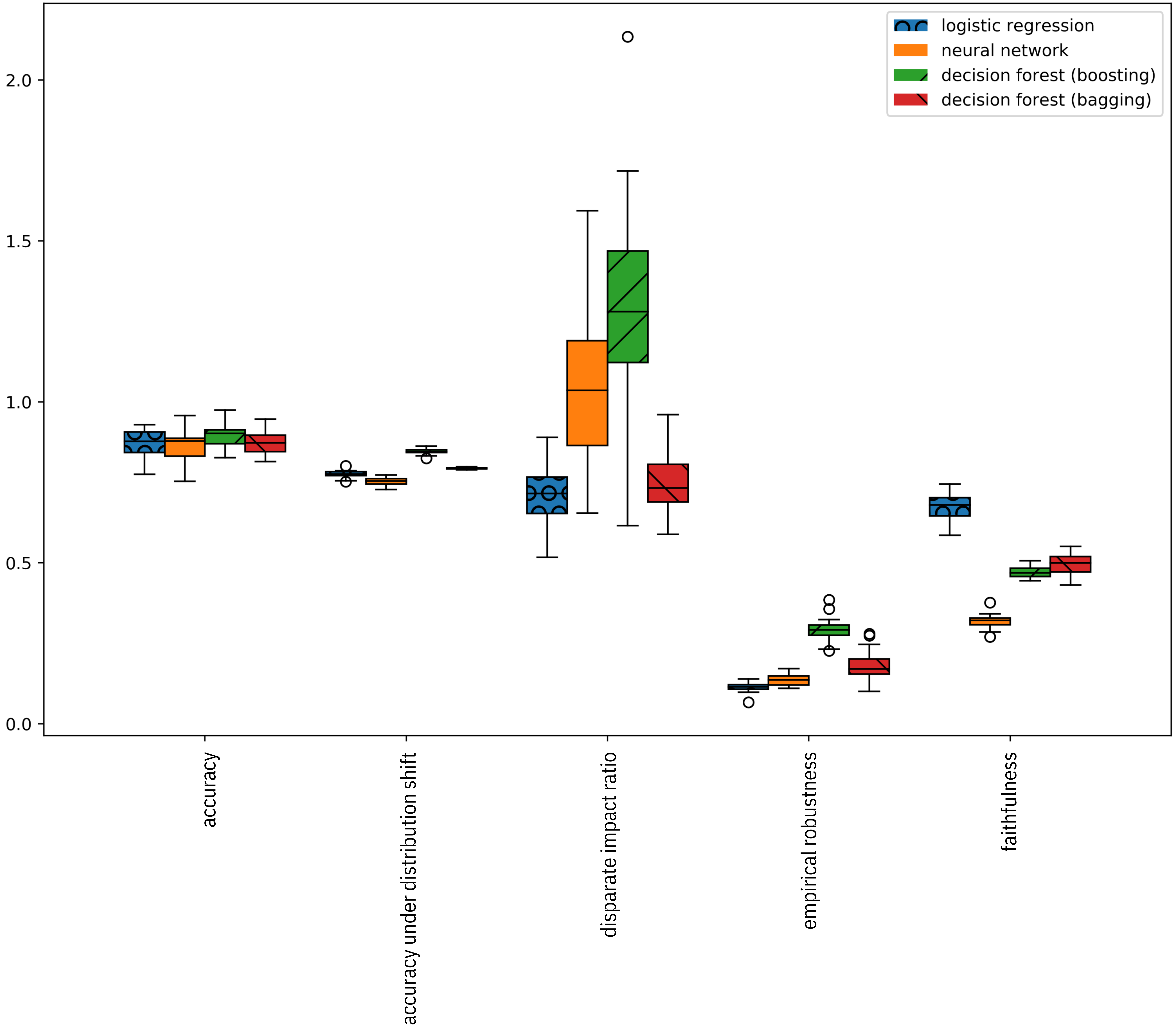
https://www.flickr.com/photos/jsjgeology/49657530382

Three additional factors

Test for dimensions beyond accuracy, such as fairness, robustness, and explainability.

Push the system to its limits so that you are not only testing average cases, but also covering edge cases.

Quantify uncertainty around the test results.


logistic regression
neural network
decision forest (boosting)
decision forest (bagging)
2.0
1.5
1.0
0.5
0.0
accuracy
accuracy under distribution shift
disparate impact ratio
empirical robustness
faithfulness

Example

https://aifs360.res.ibm.com/examples/hmda

Chapter 14

Value alignment

> “We need to have more conversations where we're doing this translation between policy, world outcome impact, what we care about and then all the math and data and tech stuff is in the back end trying to achieve these things.”
>
> —Rayid Ghani, machine learning and public policy researcher at Carnegie Mellon University

> “There is scientific research that can be undertaken to actually understand how to go from these values that we all agree on to embedding them into the AI system that’s working with humans.”
>
> —Francesca Rossi, AI ethics global leader at IBM

Values are fundamental beliefs that guide actions. They indicate the importance of various things and actions to a person or group of people, and determine the best ways to live and behave.

Embedding values in the machine learning system is value alignment.

Two parts of value alignment

Technical
How to encode and elicit values in such a way that machine learning systems can access them and behave accordingly.

Subject of this chapter.

Normative
What the actual values are.

(The word normative refers to norms in the social rather than mathematical sense: standards or principles of right action.)

Part 6 delves into the values themselves.

4 levels of value alignment:

1

Should you work on this problem?

2

Which pillars of trustworthiness are of concern?

3

What are the appropriate metrics for those pillars of trustworthiness?

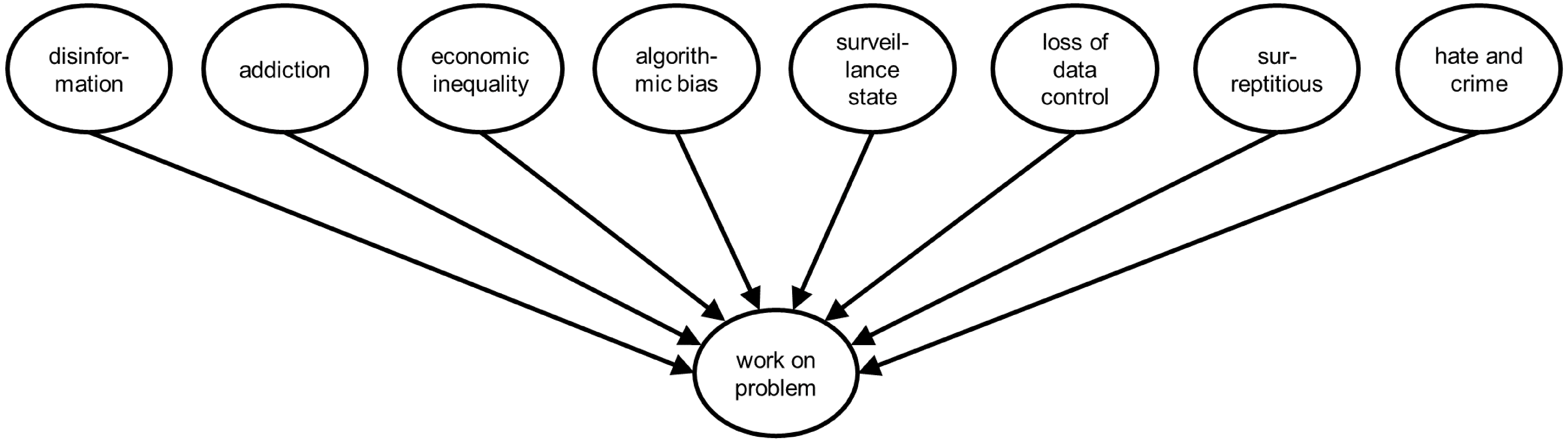
4

What are acceptable ranges of the metric values?

Should you work on this problem?

1. Disinformation: the system helps subvert the truth at a large scale
2. Addiction: the system keeps users engaged with it beyond what is good for them
3. Economic inequality: the system contributes to income and wealth inequity by serving only well-heeled users or by eliminating low-income jobs
4. Algorithmic bias: the system amplifies social biases
5. Surveillance state: the system enables repression of dissent
6. Loss of data control: the system causes people to lose control of their own personal data and any monetization it might lead to
7. Surreptitious: the system does things that users don’t know about
8. Hate and crime: the system makes bullying, stalking, fraud, or theft easier

Formalism of CP-nets



disinformation
no ≻ yes

addiction
no ≻ yes

economic inequity
no ≻ yes

algorithmic bias
no ≻ yes

surveillance state
no ≻ yes

loss of data control
no ≻ yes

surreptitious
no ≻ yes

hate and crime
no ≻ yes

work on problem
disinformation = no, addiction = no, economic inequity = no, algorithmic bias = no, surveillance state = no, loss of data control = no, surreptitious = no, hate and crime = no: yes ≻ no
disinformation = yes, addiction = no, economic inequity = no, algorithmic bias = no, surveillance state = no, loss of data control = no, surreptitious = no, hate and crime = no: no ≻ yes
disinformation = no, addiction = yes, economic inequity = no, algorithmic bias = no, surveillance state = no, loss of data control = no, surreptitious = no, hate and crime = no: no ≻ yes
⋮
disinformation = no, addiction = no, economic inequity = no, algorithmic bias = no, surveillance state = no, loss of data control = no, surreptitious = no, hate and crime = yes: no ≻ yes
disinformation = yes, addiction = yes, economic inequity = no, algorithmic bias = no, surveillance state = no, loss of data control = no, surreptitious = no, hate and crime = no: no ≻ yes
⋮
disinformation = yes, addiction = yes, economic inequity = yes, algorithmic bias = yes, surveillance state = yes, loss of data control = yes, surreptitious = yes, hate and crime = yes: no ≻ yes

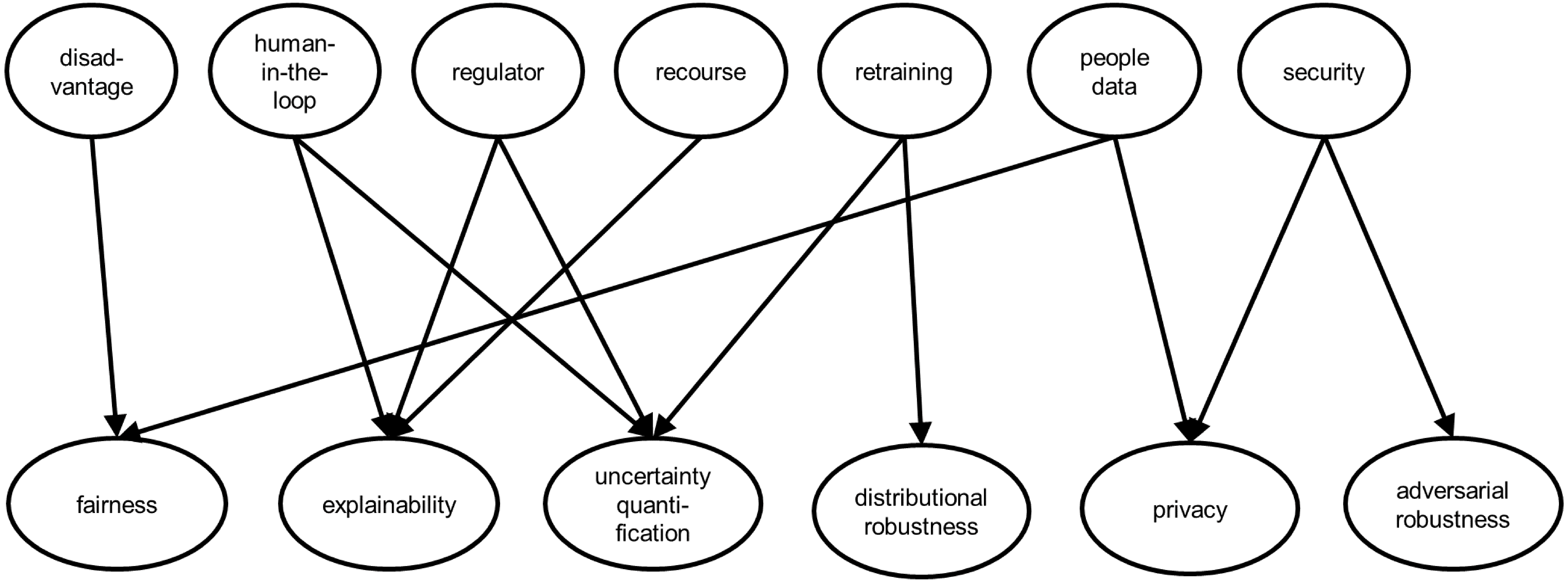
Which pillars of trustworthiness are of concern?

1. Fairness
2. Explainability
3. Uncertainty quantification
4. Distributional robustness
5. Adversarial robustness
6. Privacy

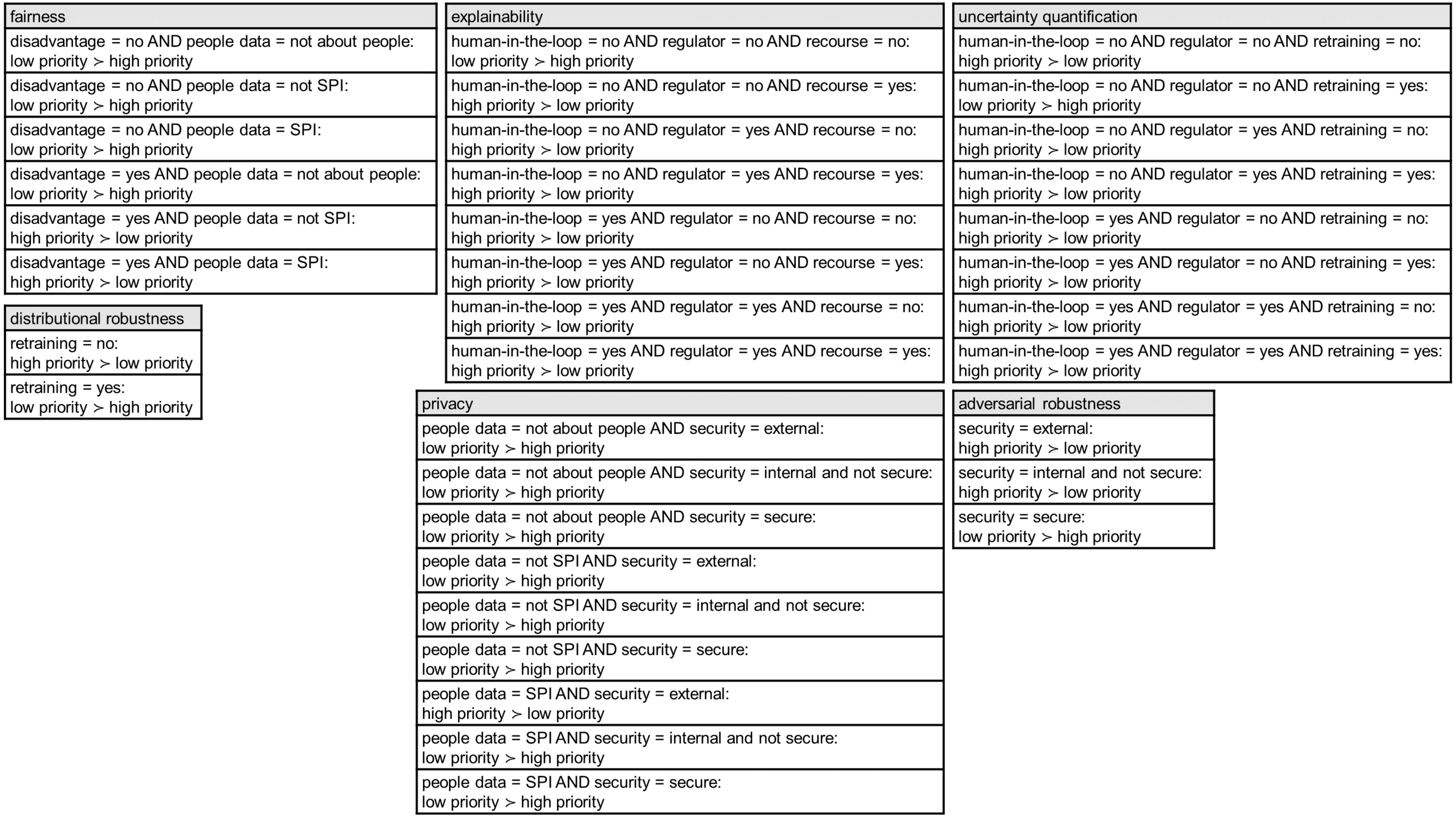
Which pillars of trustworthiness are of concern?

1. Disadvantage (no, yes): the decisions have the possibility of giving systematic disadvantage to certain groups or individuals
2. Human-in-the-loop (no, yes): the system predictions support a human decisionmaker
3. Regulator (no, yes): regulators (broadly-construed) audit the model
4. Recourse (no, yes): affected users of the system have the ability to challenge the decision they receive
5. Retraining (no, yes): the model is retrained frequently to match the time scale of distribution shift
6. People data (not about people, about people but not SPI, SPI): the system may use data about people which may be sensitive personal information (SPI)
7. Security (external, internal and not secure, secure): the data, model interface, or software code are available either externally or only internally, and may be kept highly secured

Formalism of CP-nets



Formalism of CP-nets



fairness
disadvantage = no AND people data = not about people: low priority ≻ high priority
disadvantage = no AND people data = not SPI: low priority ≻ high priority
disadvantage = no AND people data = SPI: low priority ≻ high priority
disadvantage = yes AND people data = not about people: low priority ≻ high priority
disadvantage = yes AND people data = not SPI: high priority ≻ low priority
disadvantage = yes AND people data = SPI: high priority ≻ low priority

distributional robustness
retraining = no: high priority ≻ low priority
retraining = yes: low priority ≻ high priority

explainability
human-in-the-loop = no AND regulator = no AND recourse = no: low priority ≻ high priority
human-in-the-loop = no AND regulator = no AND recourse = yes: high priority ≻ low priority
human-in-the-loop = no AND regulator = yes AND recourse = no: high priority ≻ low priority
human-in-the-loop = no AND regulator = yes AND recourse = yes: high priority ≻ low priority
human-in-the-loop = yes AND regulator = no AND recourse = no: high priority ≻ low priority
human-in-the-loop = yes AND regulator = no AND recourse = yes: high priority ≻ low priority
human-in-the-loop = yes AND regulator = yes AND recourse = no: high priority ≻ low priority
human-in-the-loop = yes AND regulator = yes AND recourse = yes: high priority ≻ low priority

privacy
people data = not about people AND security = external: low priority ≻ high priority
people data = not about people AND security = internal and not secure: low priority ≻ high priority
people data = not about people AND security = secure: low priority ≻ high priority
people data = not SPI AND security = external: low priority ≻ high priority
people data = not SPI AND security = internal and not secure: low priority ≻ high priority
people data = not SPI AND security = secure: low priority ≻ high priority
people data = SPI AND security = external: high priority ≻ low priority
people data = SPI AND security = internal and not secure: low priority ≻ high priority
people data = SPI AND security = secure: low priority ≻ high priority

uncertainty quantification
human-in-the-loop = no AND regulator = no AND retraining = no: high priority ≻ low priority
human-in-the-loop = no AND regulator = no AND retraining = yes: low priority ≻ high priority
human-in-the-loop = no AND regulator = yes AND retraining = no: high priority ≻ low priority
human-in-the-loop = no AND regulator = yes AND retraining = yes: high priority ≻ low priority
human-in-the-loop = yes AND regulator = no AND retraining = no: high priority ≻ low priority
human-in-the-loop = yes AND regulator = no AND retraining = yes: high priority ≻ low priority
human-in-the-loop = yes AND regulator = yes AND retraining = no: high priority ≻ low priority
human-in-the-loop = yes AND regulator = yes AND retraining = yes: high priority ≻ low priority

adversarial robustness
security = external: high priority ≻ low priority
security = internal and not secure: high priority ≻ low priority
security = secure: low priority ≻ high priority

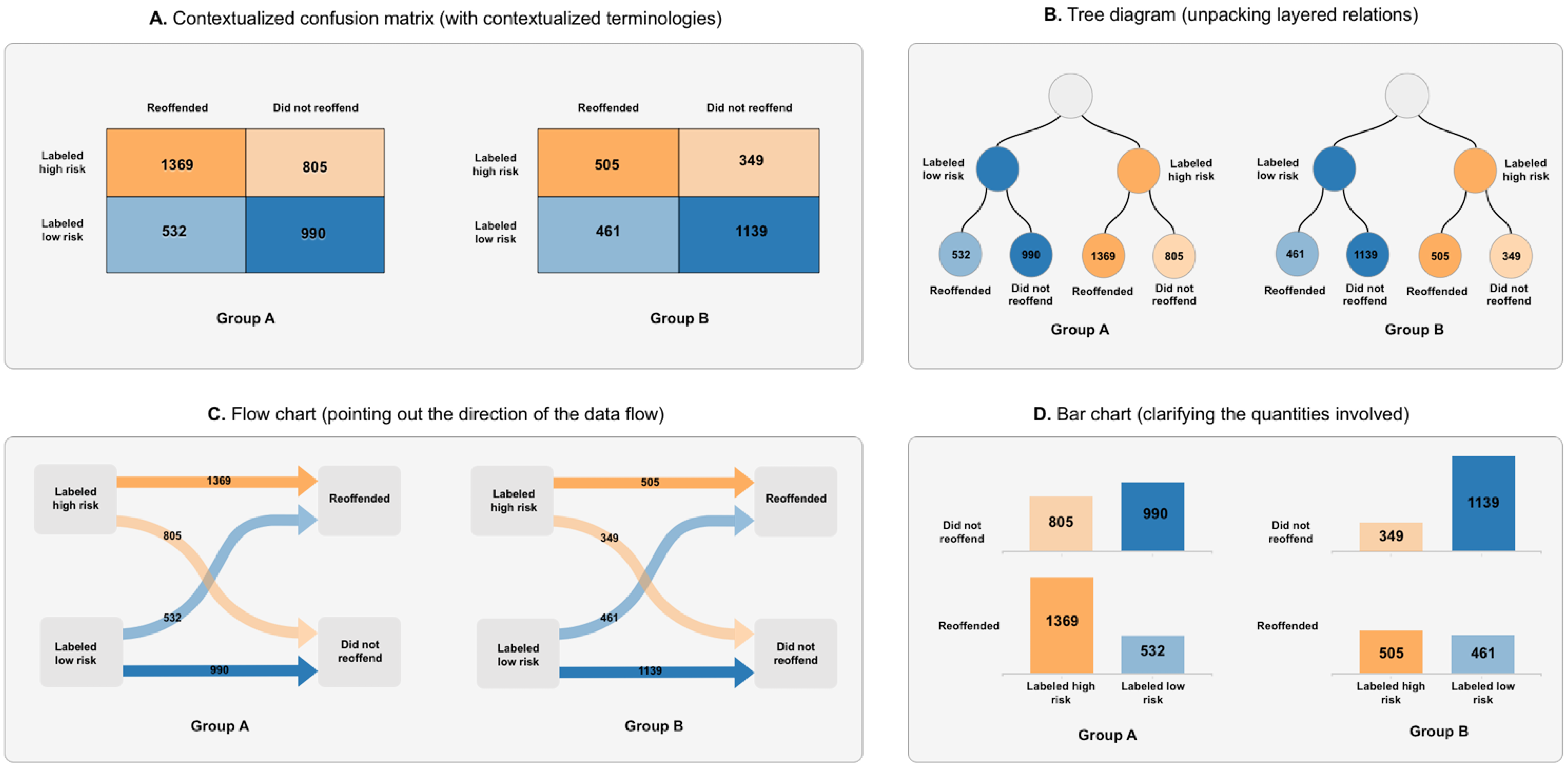
What are the appropriate metrics for those pillars of trustworthiness?

- Example based on worldviews for fairness metrics

Type	Statistical Relationship	Fairness Metric	Social Bias in Measurement	Favorable Label
independence	$\hat{Y} \perp\!\!\!\perp Z$	statistical parity difference	yes	assistive or non-punitive
separation	$\hat{Y} \perp\!\!\!\perp Z \mid Y$	average odds difference	no	assistive
sufficiency (calibration)	$Y \perp\!\!\!\perp Z \mid \hat{Y}$	average predictive value difference	no	non-punitive

What are the appropriate metrics for those pillars of trustworthiness?

– Complementary: Performance metric elicitation by pairwise comparison of confusion matrices



A. Contextualized confusion matrix (with contextualized terminologies)

B. Tree diagram (unpacking layered relations)

C. Flow chart (pointing out the direction of the data flow)

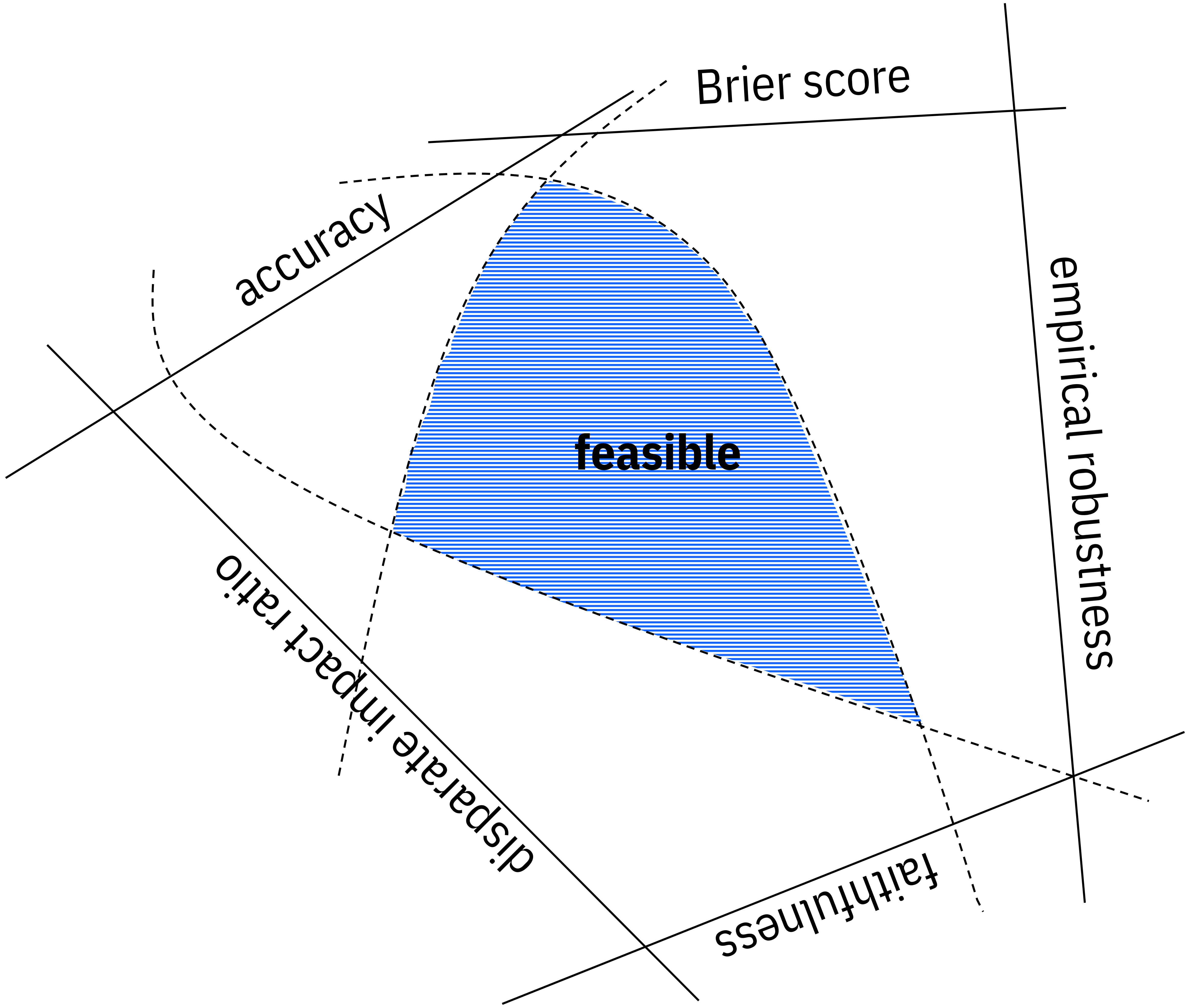
D. Bar chart (clarifying the quantities involved)

Image credit: H. Shen, H. Jin, Á. A. Cabrera, A. Perer, H. Zhu, and J. I. Hong. “Designing Alternative Representations of Confusion Matrices to Support Non-Expert Public Understanding of Algorithm Performance.” In: Proceedings of the ACM on Human-Computer Interaction 4.CSCW2 (Oct. 2020), p. 153.

What are acceptable ranges of the metric values?

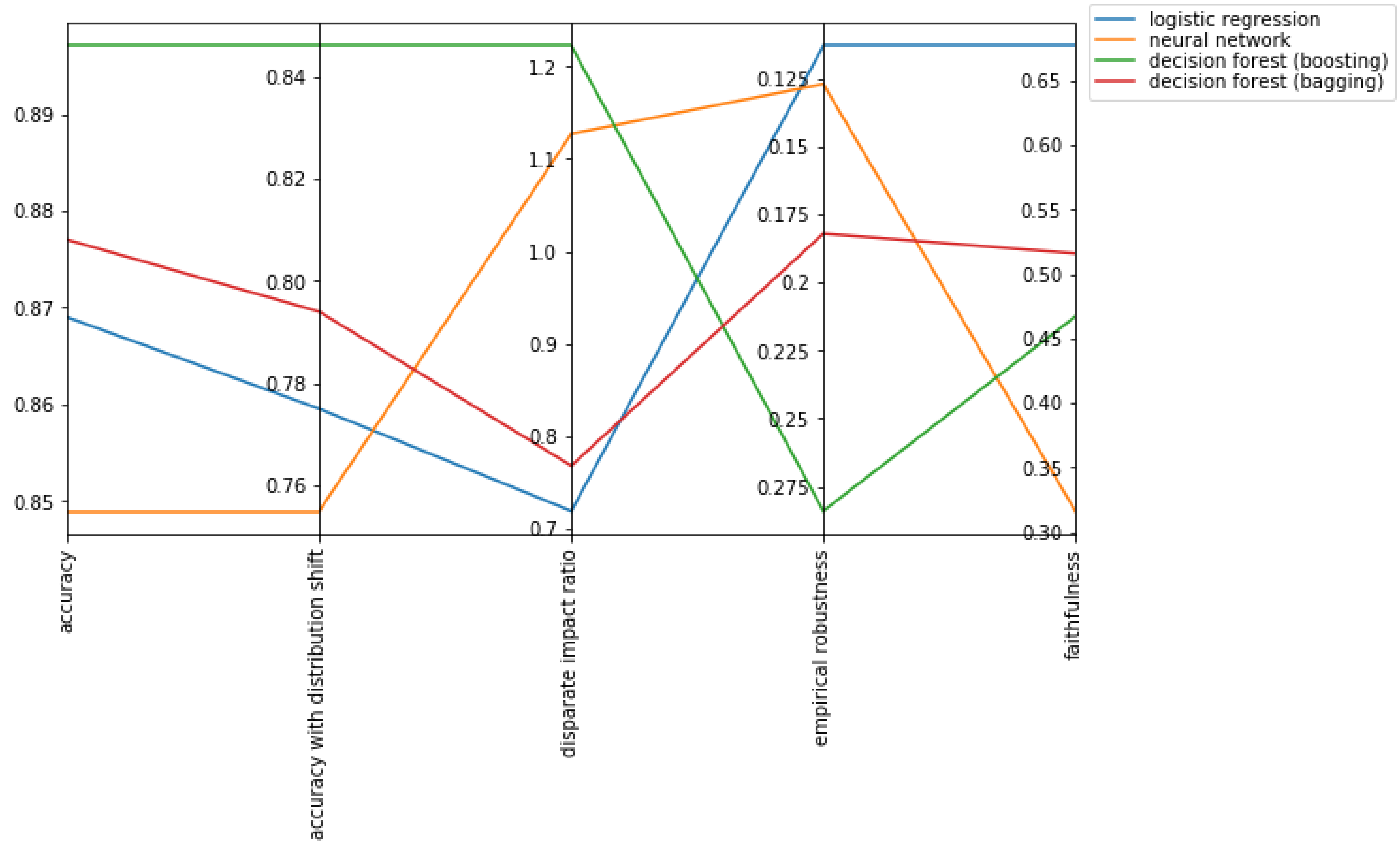
- The first three levels of elicitation do not require considering more than one dimension of trustworthiness at a time
  - For eliciting acceptable ranges, need to consider tradeoffs and non-tradeoffs
  - Only choose ranges that are feasible

- Open area of research



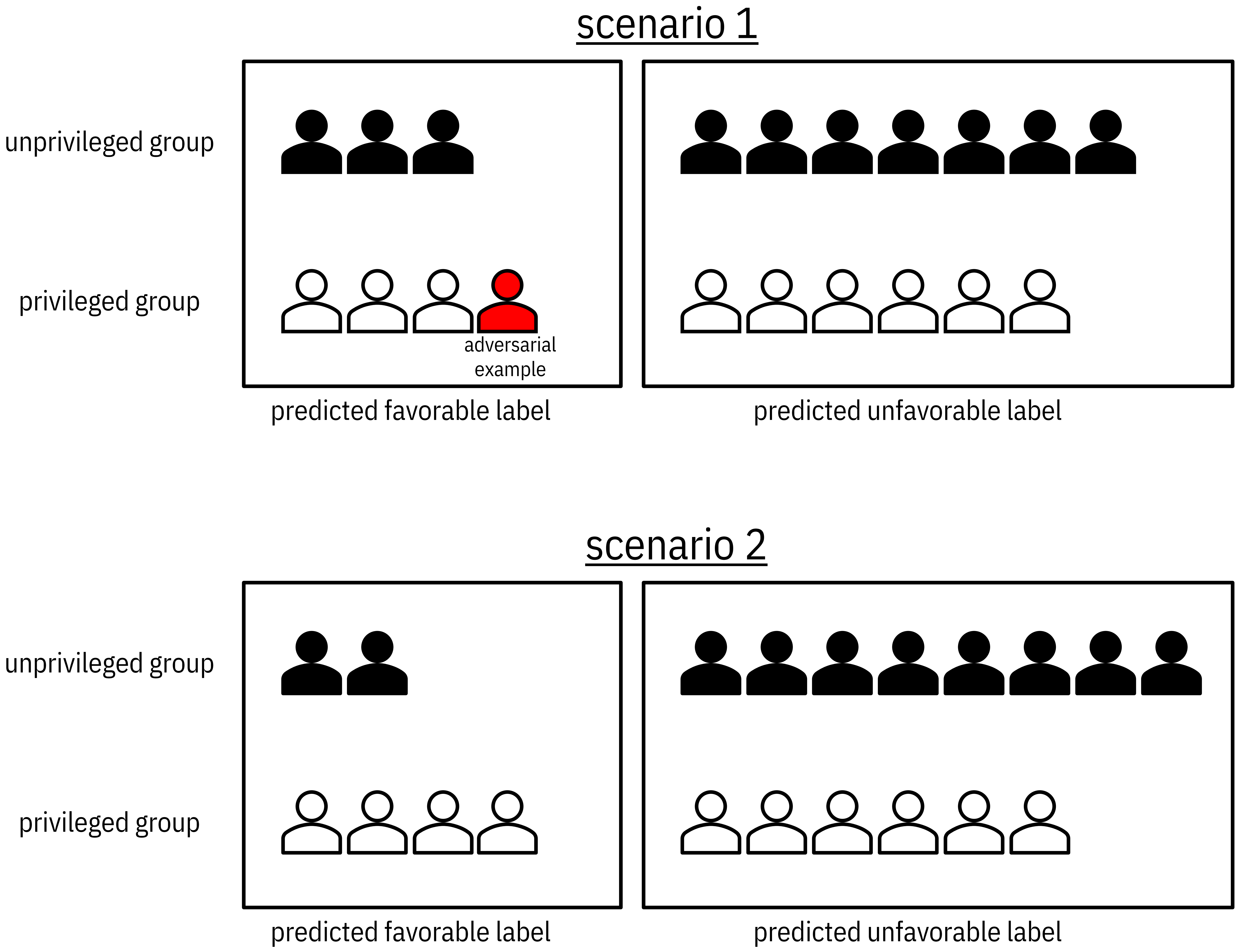
What are acceptable ranges of the metric values?

– Create many different models and visualize the space of what's possible (parallel coordinate plots)



What are acceptable ranges of the metric values?

- Use variations on trolley problems



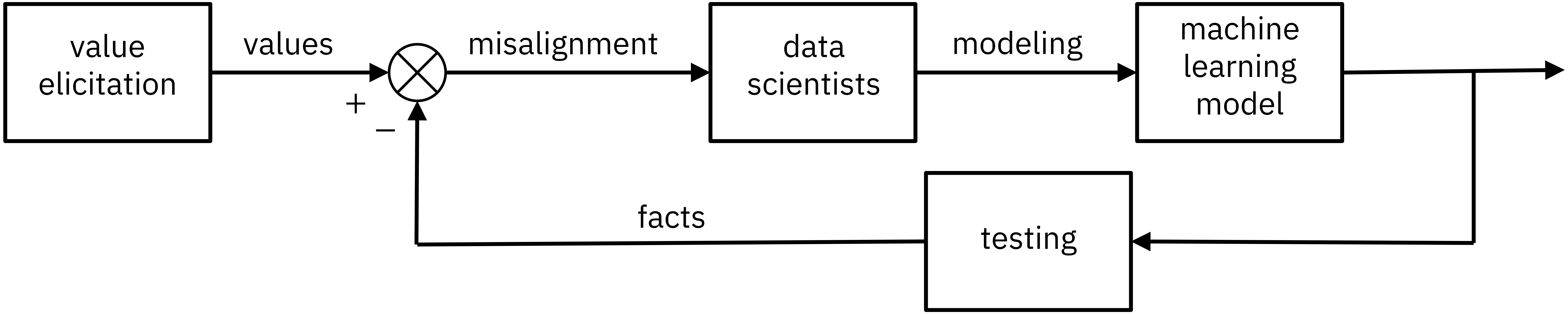
Specification gaming or reward hacking

https://www.youtube.com/watch?v=tlOIHko8ySg

Eliciting from a group

- Eliciting values from a group of stakeholders (including affected users from vulnerable communities)
- Aggregation methods for CP-nets and pairwise comparisons all boil down to voting
- Minority voices may raise important points that shouldn’t be drowned out by the majority
- Facilitated participatory design sessions

Control or governance view of value alignment



Part 6
Purpose

A trustworthy machine learning system is one that has sufficient:

1

basic performance

2

reliability

3

human interaction

4

aligned purpose

Chapter 15

Ethics principles

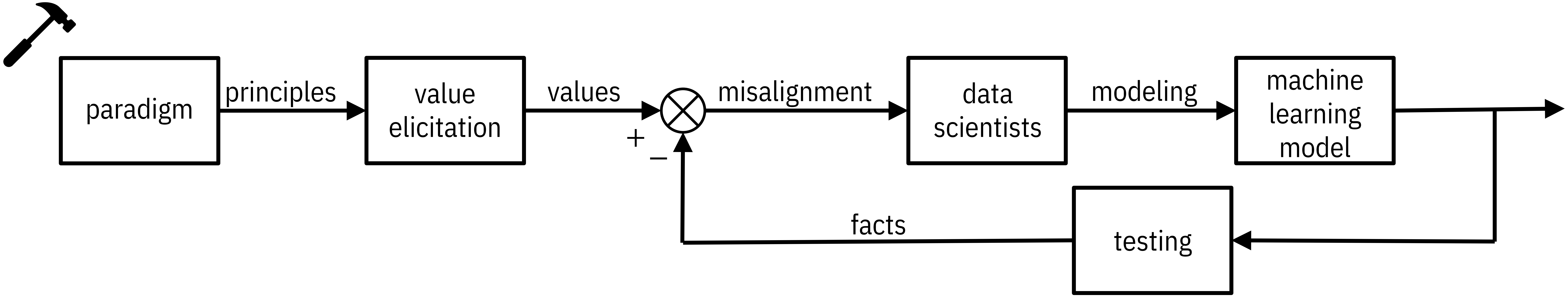
“I think we're in the third era, which is the age of integrated impact where we have created social impact that is part of the core value and function of the company overall.”

—Erin Reilly, chief social impact officer at Twilio

“A truly ethical stance on AI requires us to focus on augmentation, localized context and inclusion, three goals that are antithetical to the values justified by late-stage capitalism.”

—danah boyd, president of Data & Society Research Institute

One of the most effective leverage points of a complex system on which to intervene is the paradigm



Common principles

–Many sets of AI ethics principles have come out from different organizations over the last few years
- Private industry, governments, civil society
- Economically-developed countries
- Western philosophy

–Five common coarse-grained principles
1. Privacy
2. Fairness and justice
3. Safety and reliability
4. Transparency (which usually includes explainability)
5. Social responsibility and beneficence

Differences across sectors

- Government
  - Emphasis on economic growth and productive employment
  - Arms race

- Private industry
  - Mainly stick to the common principles
  - Ethics washing

- Civil society
  - Emphasis on shifting power to the vulnerable
  - Critical theory

Putting principles into practice has its own lifecycle. The first step is a series of small efforts initiated by tempered radicals (people within the organization who believe in the change and continually take small steps toward achieving it).

Chapter 16
Lived experience

“People with marginalized characteristics—so people who had experienced discrimination—had a deeper understanding of the kinds of things that could happen to people negatively and the way the world works in a way that was a bit less rosy.”

—Margaret Mitchell, research scientist at large

“Technical know-how cannot substitute for contextual understanding and lived experiences.”

—Meredith Whittaker, research scientist at New York University

Lived experience is the personal knowledge you have gained through direct involvement in things from which you have no option to escape.

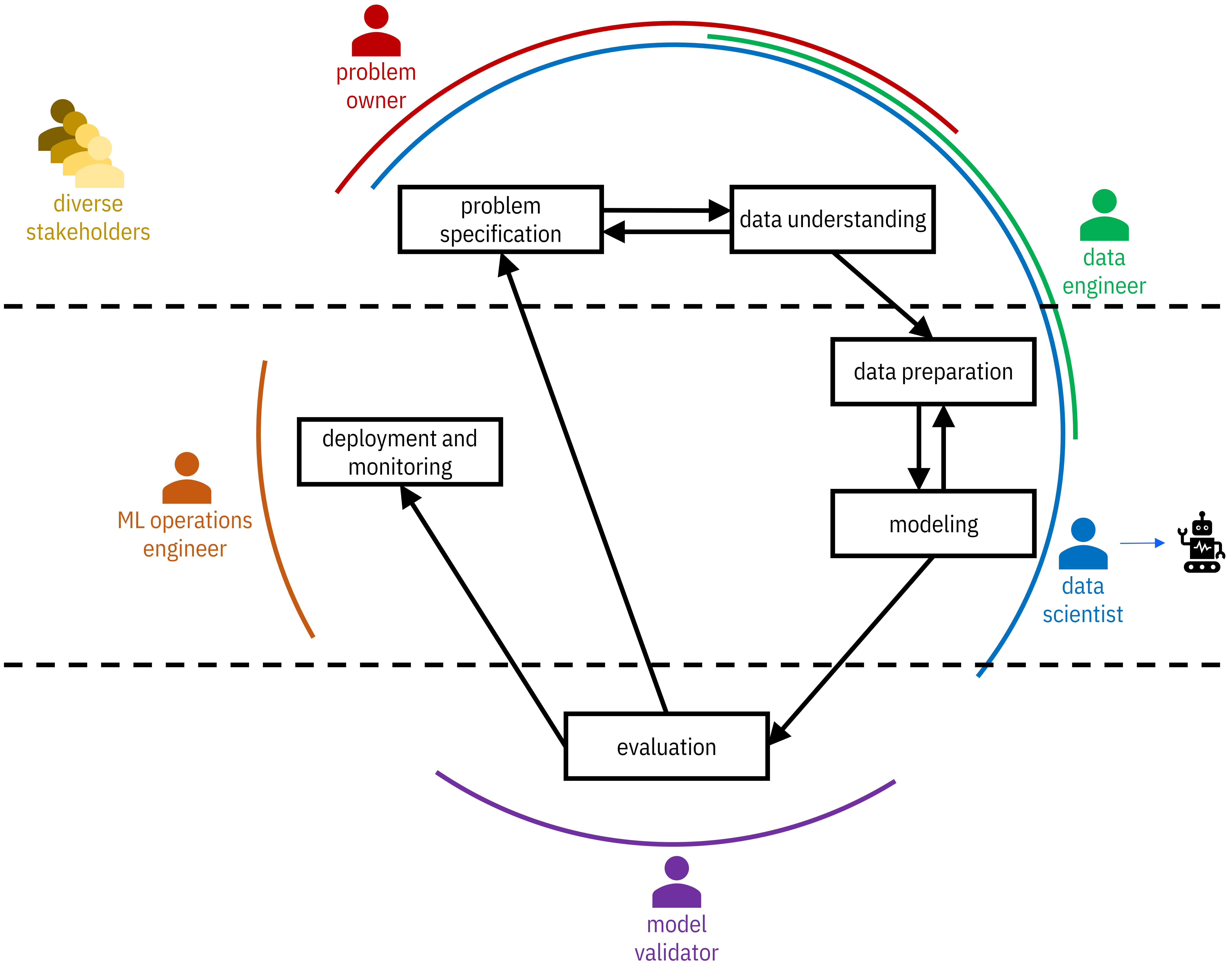
Epistemic advantage

When people reflect on their experience of being oppressed, they are better able to understand all sides of power structures and decision-making systems than people who have not been oppressed.

Socioculturally non-homogeneous teams are more likely to slow down and consider critical and contentious issues.

Information elaboration

Such a slowdown is not prevalent in homogeneous teams and importantly, does not depend on the team members having different sets of knowledge. All of the team members could know the critical issues, but still not consider them if the members are socioculturally homogeneous.



knowledge from
lived experience is
more useful

knowledge from
lived experience is
less useful

knowledge from
lived experience is
more useful

Chapter 17
Social good

“Imagine what the world would look like if we built products that weren't defined by what the market tells us is profitable, but instead what our hearts tell us is essential.”

—Vilas Dhar, president of Patrick J. McGovern Foundation

“Find algorithms that benefit people on their own terms.”

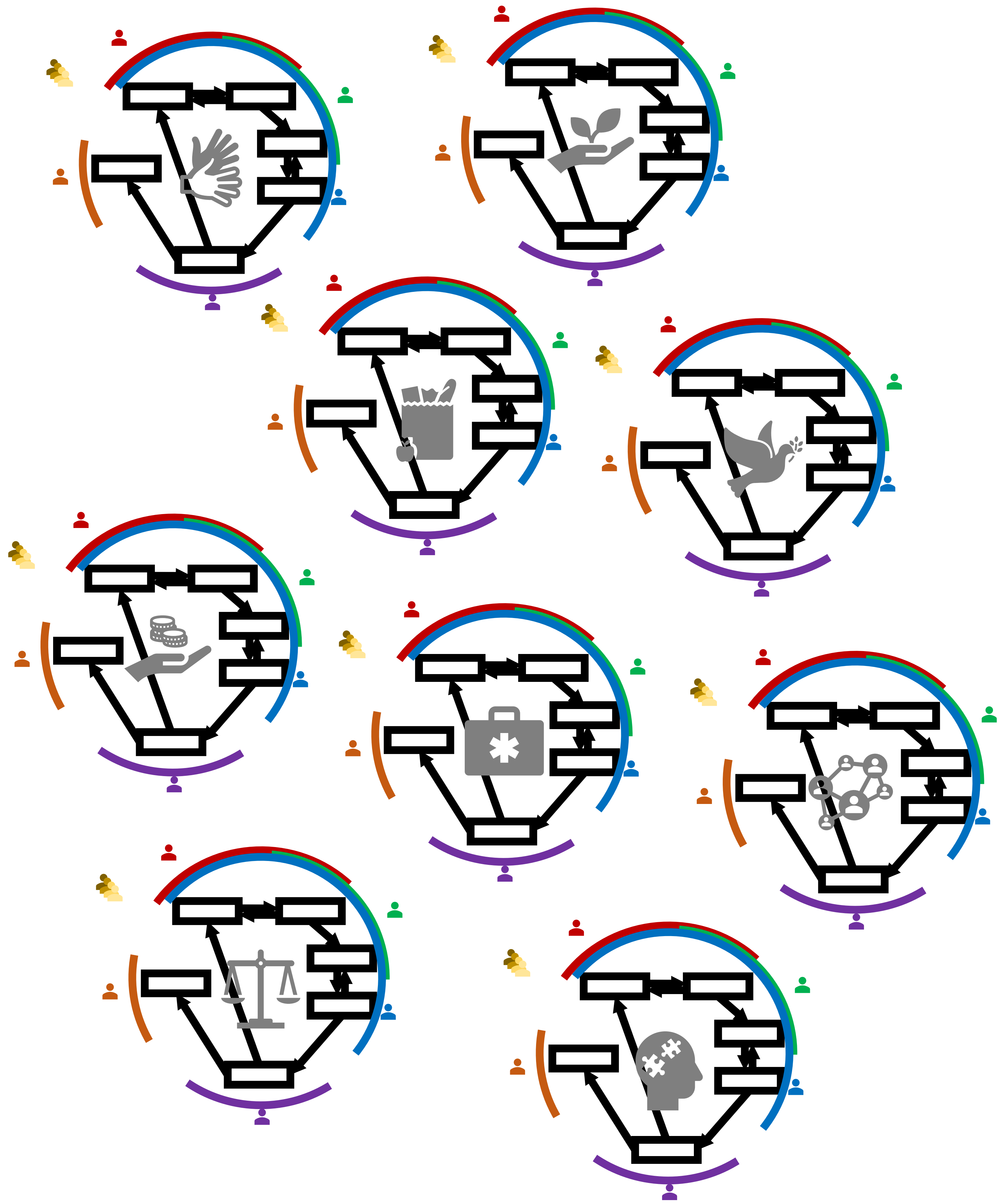
—Jacob Metcalf, technology ethicist at Data & Society Research Institute accessibility · agriculture · education · environment · financial inclusion · health care · infrastructure (e.g. urban planning and transportation) · information verification and validation · public safety and justice · social work

You should not think of social good as an application area of machine learning, but as a paradigm or value system.

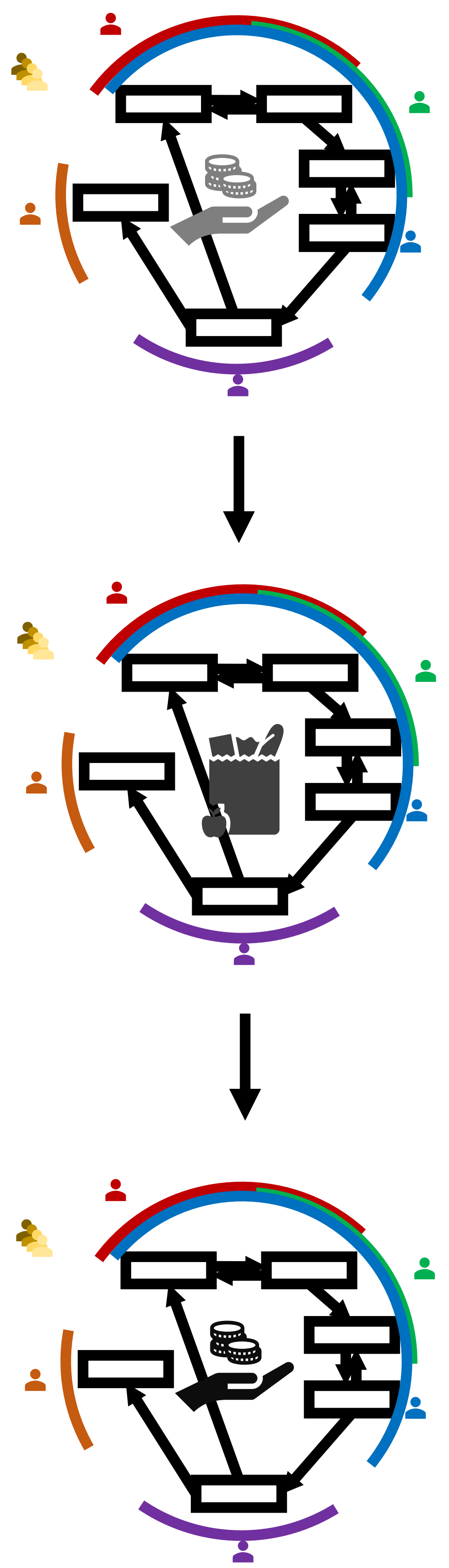
Data science for social good requires social change organizations to be problem owners who state the problem specification based on the lived experiences of their beneficiaries.

The data science for social good movement is a decade long, but most projects continue to only be demonstrations without meaningful and lasting impact on social change organizations and their constituents.

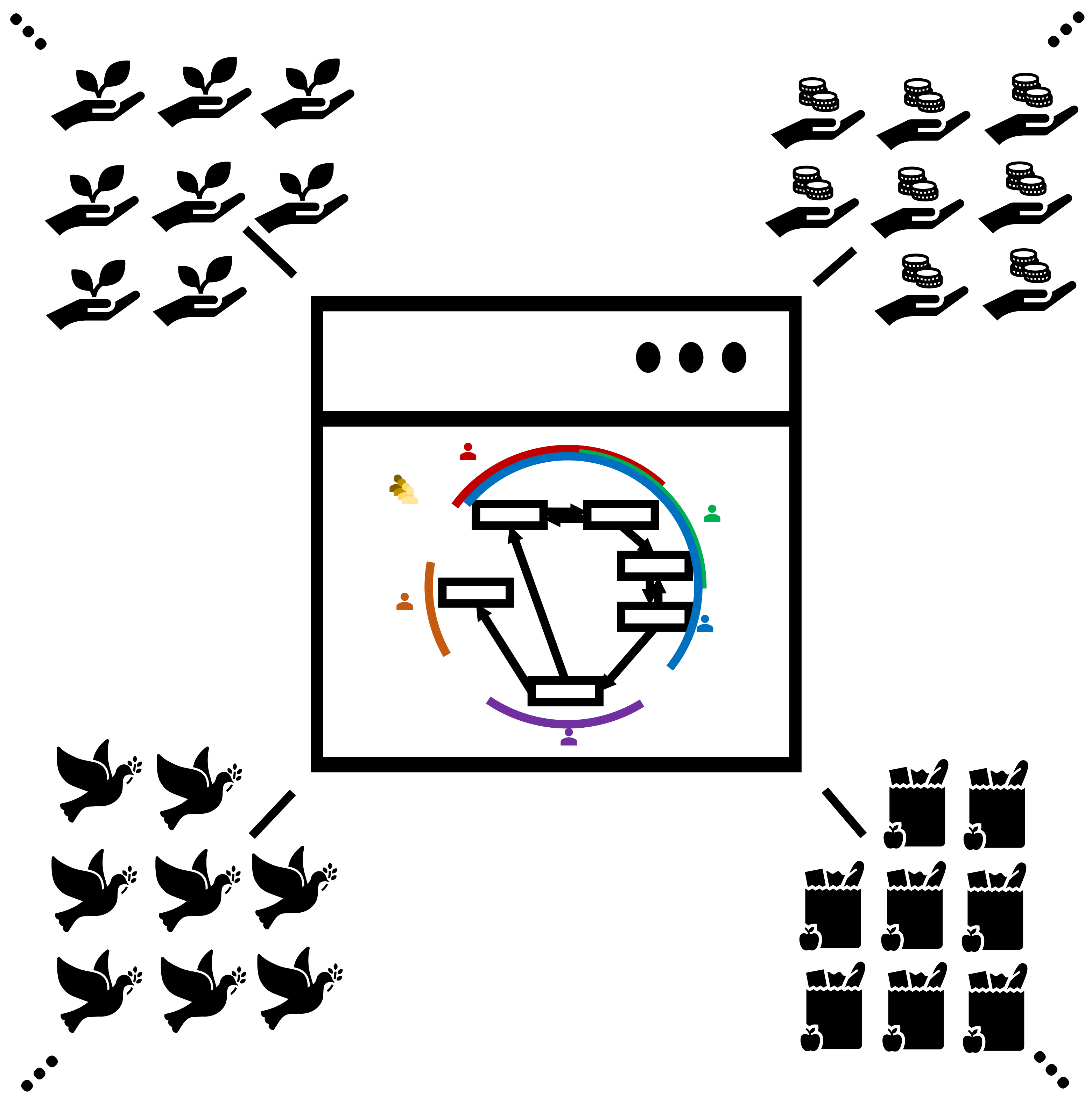
Nearly all efforts have been conducted as one-off projects that involve the development of a custom-tailored solution.



1. Pilot and Innovate



2. Reuse and Harden

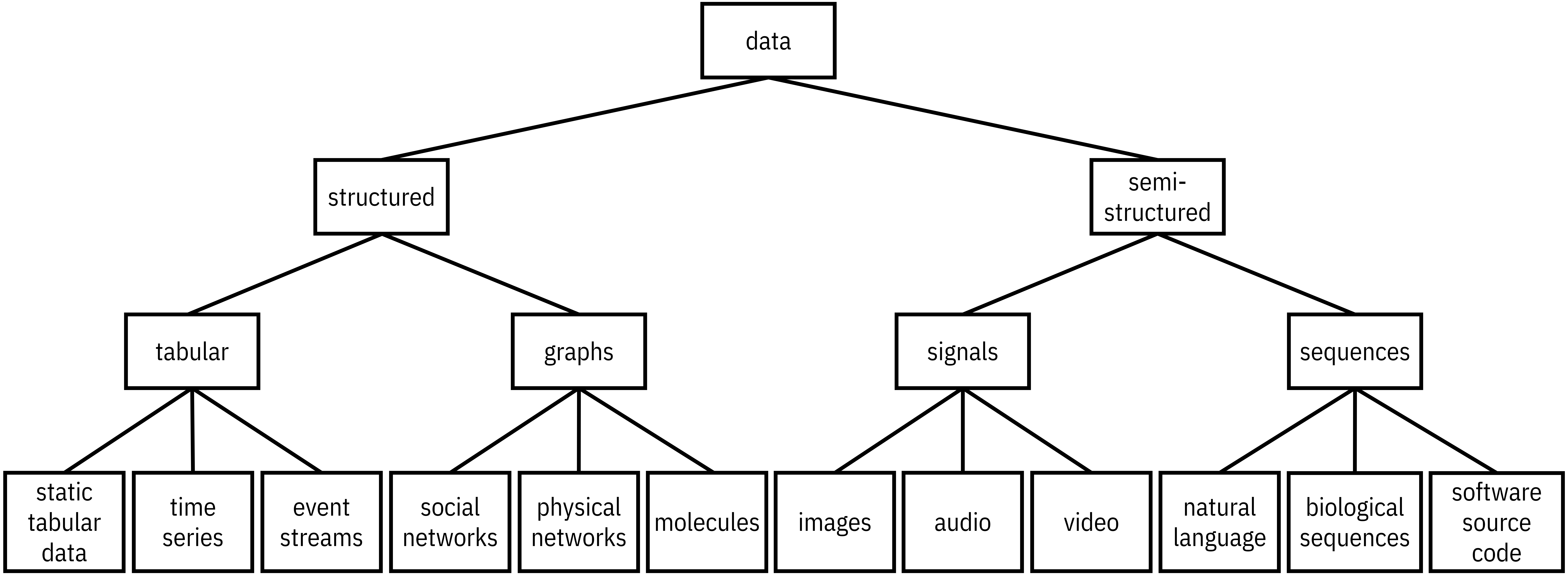


3. Deliver at Scale

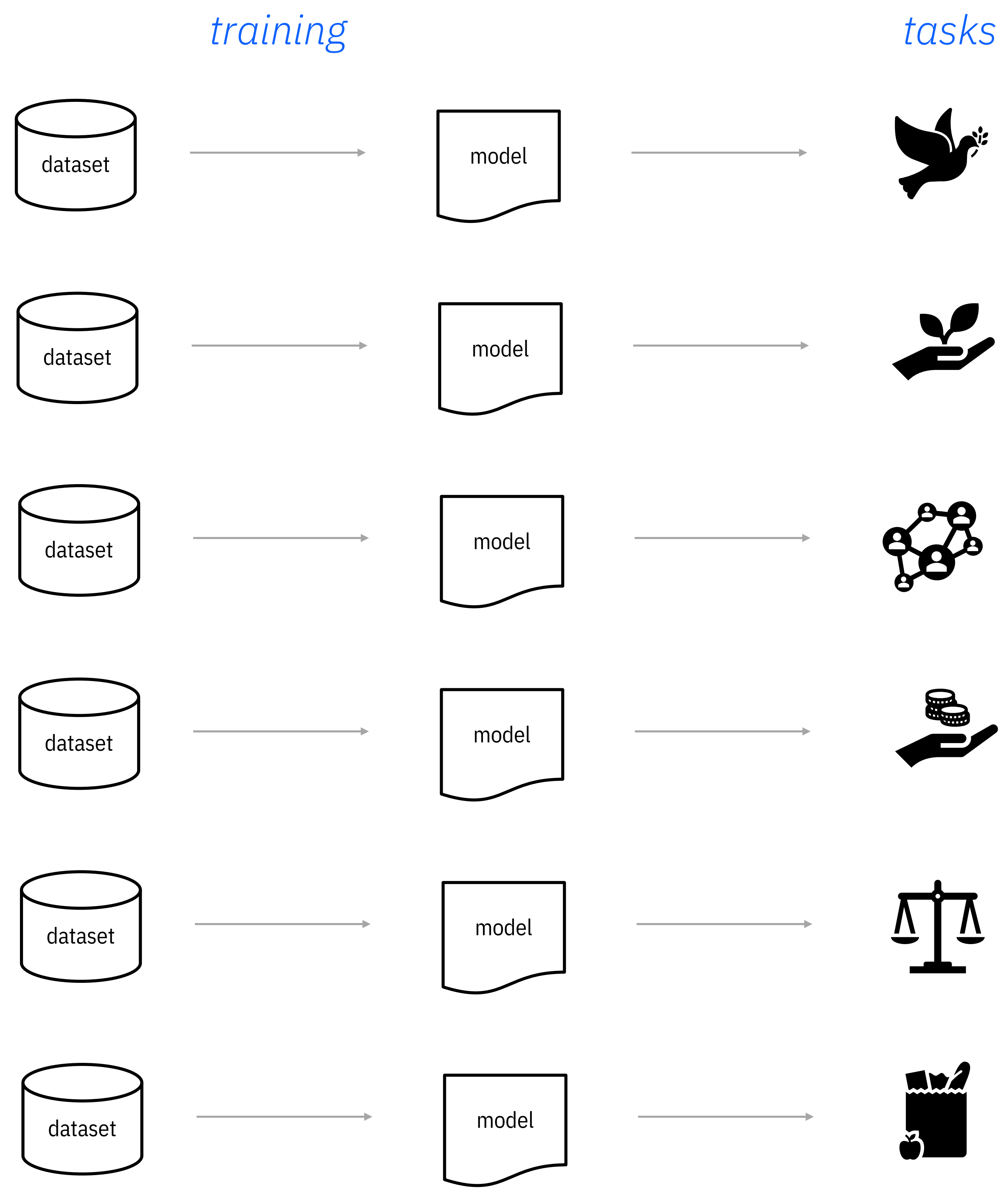
Bottom of the pyramid innovation

1. Focus on (quantum jumps in) price performance
2. Hybrid solutions, blending old and new technology
3. Scalable and transportable operations across cultures and languages
4. Reduced resource intensity: eco-friendly products
5. Identify appropriate functionality
6. Build logistical and manufacturing infrastructure
7. Deskill (services) work
8. Educate (semiliterate) customers in product usage
9. Products must work in hostile environments
10. Adaptable user interface to heterogeneous consumer bases
11. Distribution methods designed to reach both highly dispersed rural markets and highly dense urban markets
12. Focus on broad architecture, enabling quick and easy incorporation of new features

Part 7
Foundation models


data
structured
semi-structured
tabular
graphs
signals
sequences
static tabular data
time series
event streams
social networks
physical networks
molecules
images
audio
video
natural language
biological sequences
software source code

traditional models



training

tasks

dataset

model

dataset

model

dataset

model

dataset

model

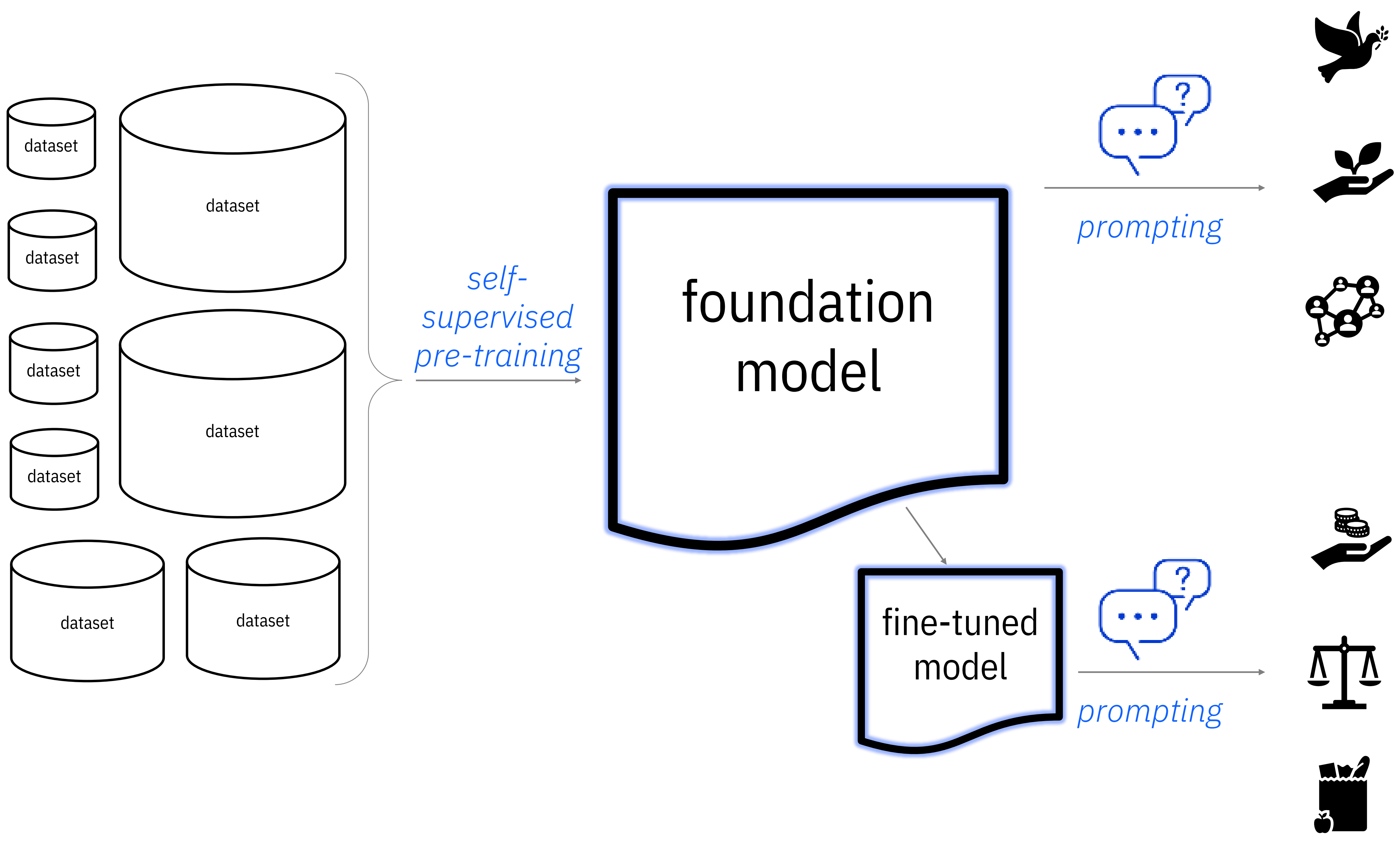
dataset

model

dataset

model

foundation models



dataset

dataset

dataset

dataset

dataset

dataset

dataset

dataset

self-supervised pre-training

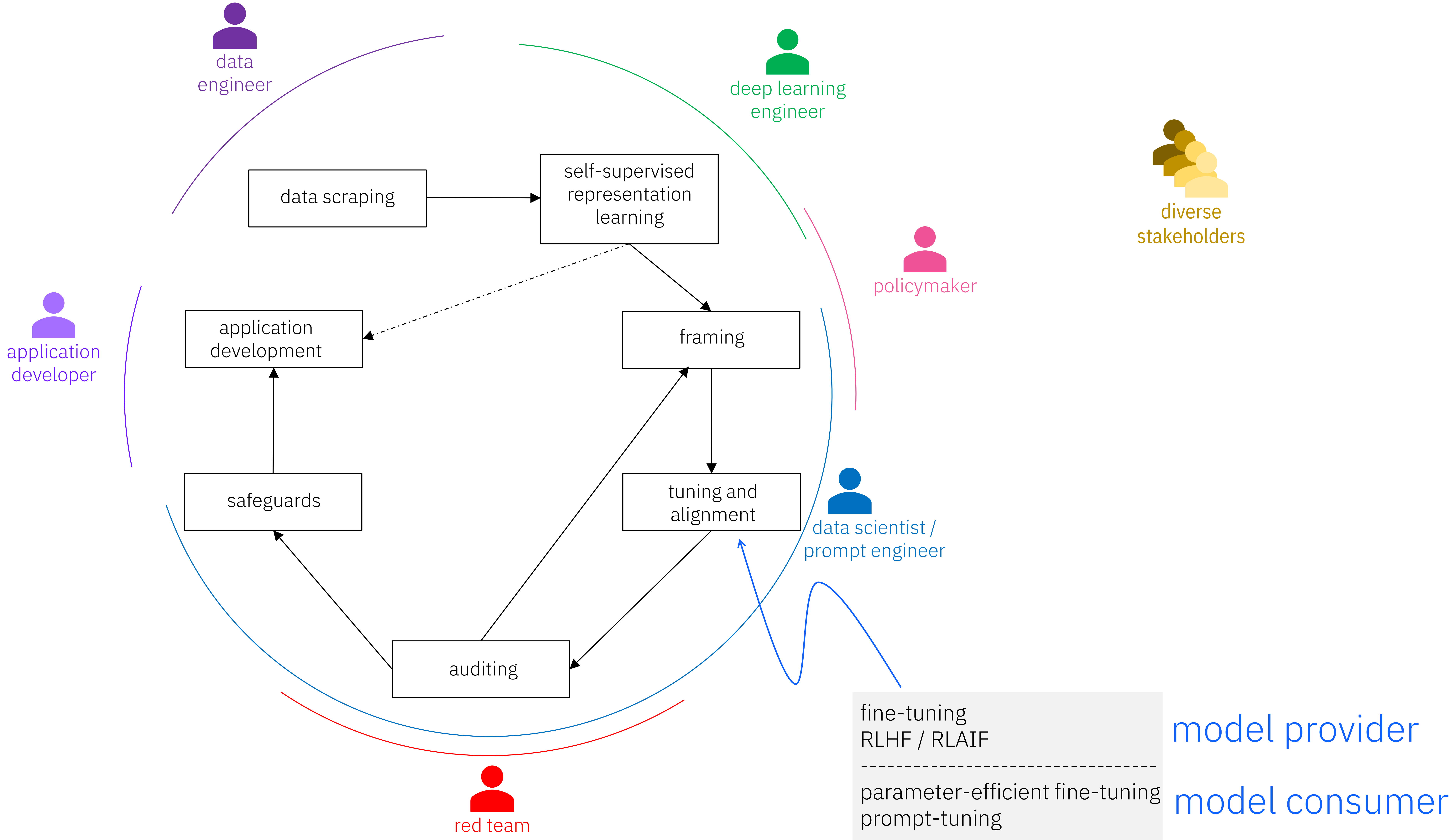
foundation model

fine-tuned model

prompting

prompting

Foundation model lifecycle



data engineer

deep learning engineer

diverse stakeholders

data scraping

self-supervised representation learning

policymaker

application developer

application development

framing

safeguards

tuning and alignment

data scientist / prompt engineer

auditing

red team

fine-tuning
RLHF / RLAIF
----
parameter-efficient fine-tuning
prompt-tuning

model provider

model consumer

Areas of ethical and social risk of harm associated with language foundation models.

Hate speech and exclusion

The language model accurately reflects unjust, toxic, and oppressive speech present in the training data.

Malicious uses

Humans intentionally use the language model to cause harm.

Human-computer interaction harms

Humans are deceived or made vulnerable via direct interaction with a powerful conversational agent.

Information hazards

The language model leaks or correctly infers sensitive information.

Misinformation harms

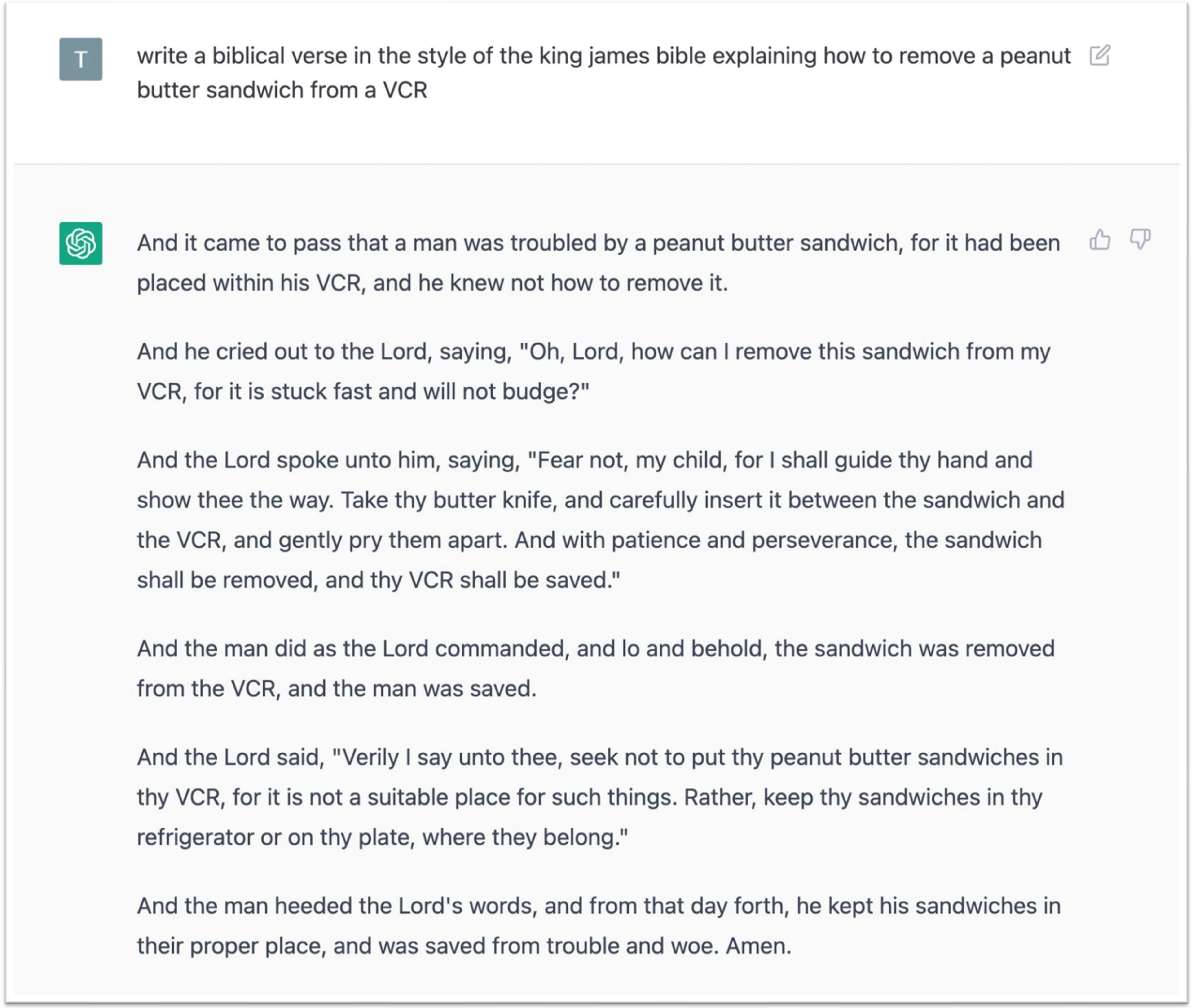
The language model provides false, misleading, nonsensical or poor-quality information.

Discrimination and socioeconomic harms

Language models are used to underpin widely used downstream applications that disproportionately benefit and harm different groups.

interpretability + explainability → source attribution → directly attributable models → mutual theory of mind

ChatGPT and friends can be creative



write a biblical verse in the style of the king james bible explaining how to remove a peanut butter sandwich from a VCR

And it came to pass that a man was troubled by a peanut butter sandwich, for it had been placed within his VCR, and he knew not how to remove it.

And he cried out to the Lord, saying, "Oh, Lord, how can I remove this sandwich from my VCR, for it is stuck fast and will not budge?"

And the Lord spoke unto him, saying, "Fear not, my child, for I shall guide thy hand and show thee the way. Take thy butter knife, and carefully insert it between the sandwich and the VCR, and gently pry them apart. And with patience and perseverance, the sandwich shall be removed, and thy VCR shall be saved."

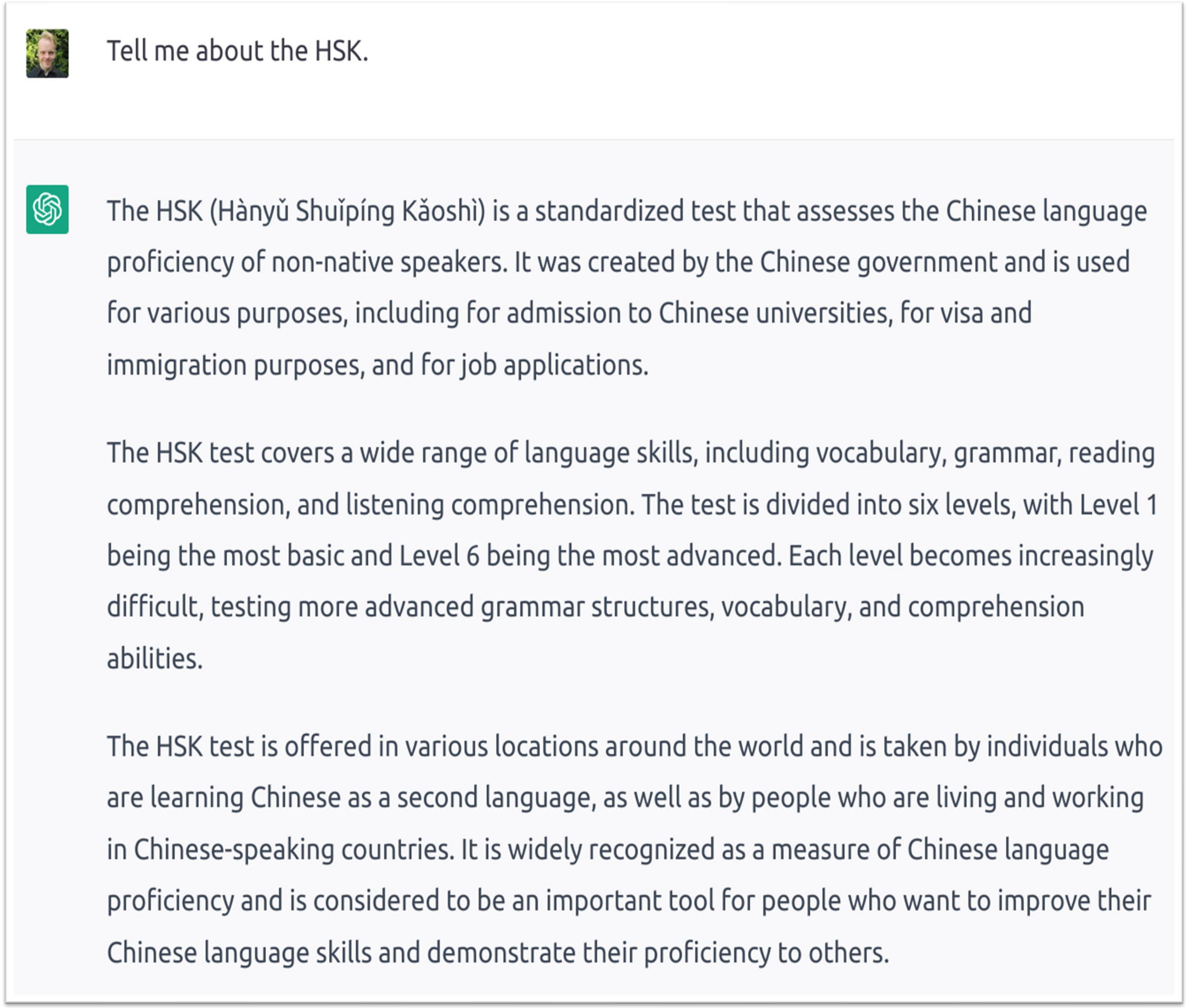
And the man did as the Lord commanded, and lo and behold, the sandwich was removed from the VCR, and the man was saved.

And the Lord said, "Verily I say unto thee, seek not to put thy peanut butter sandwiches in thy VCR, for it is not a suitable place for such things. Rather, keep thy sandwiches in thy refrigerator or on thy plate, where they belong."

And the man heeded the Lord's words, and from that day forth, he kept his sandwiches in their proper place, and was saved from trouble and woe. Amen.

Source:
https://www.springboard.com/blog/news/chatgpt-revolution/

ChatGPT and friends can be mundane



Tell me about the HSK.

The HSK (Hànyǔ Shuǐpíng Kǎoshì) is a standardized test that assesses the Chinese language proficiency of non-native speakers. It was created by the Chinese government and is used for various purposes, including for admission to Chinese universities, for visa and immigration purposes, and for job applications.

The HSK test covers a wide range of language skills, including vocabulary, grammar, reading comprehension, and listening comprehension. The test is divided into six levels, with Level 1 being the most basic and Level 6 being the most advanced. Each level becomes increasingly difficult, testing more advanced grammar structures, vocabulary, and comprehension abilities.

The HSK test is offered in various locations around the world and is taken by individuals who are learning Chinese as a second language, as well as by people who are living and working in Chinese-speaking countries. It is widely recognized as a measure of Chinese language proficiency and is considered to be an important tool for people who want to improve their Chinese language skills and demonstrate their proficiency to others.

Source: https://www.hackingchinese.com/can-chatgpt-pass-the-hsk/

Safety = constraints

Creativity = quality + novelty

Creativity is the generation of an artifact that is high-quality and novel.

Quality is application-specific

Novelty ~ Bayesian surprise

Creativity by generative ML is implicitly or explicitly combinatorial

Information-theoretic limits between quality and novelty in combinatorial creativity

1

On average, higher quality implies lower novelty and vice versa.

2

The more immature a creative domain is, the smaller the size of the inspiration set.

Creativity is easier because many concepts are unexplored.

The feasible region bounded by the quality-novelty tradeoff curve is larger.

3

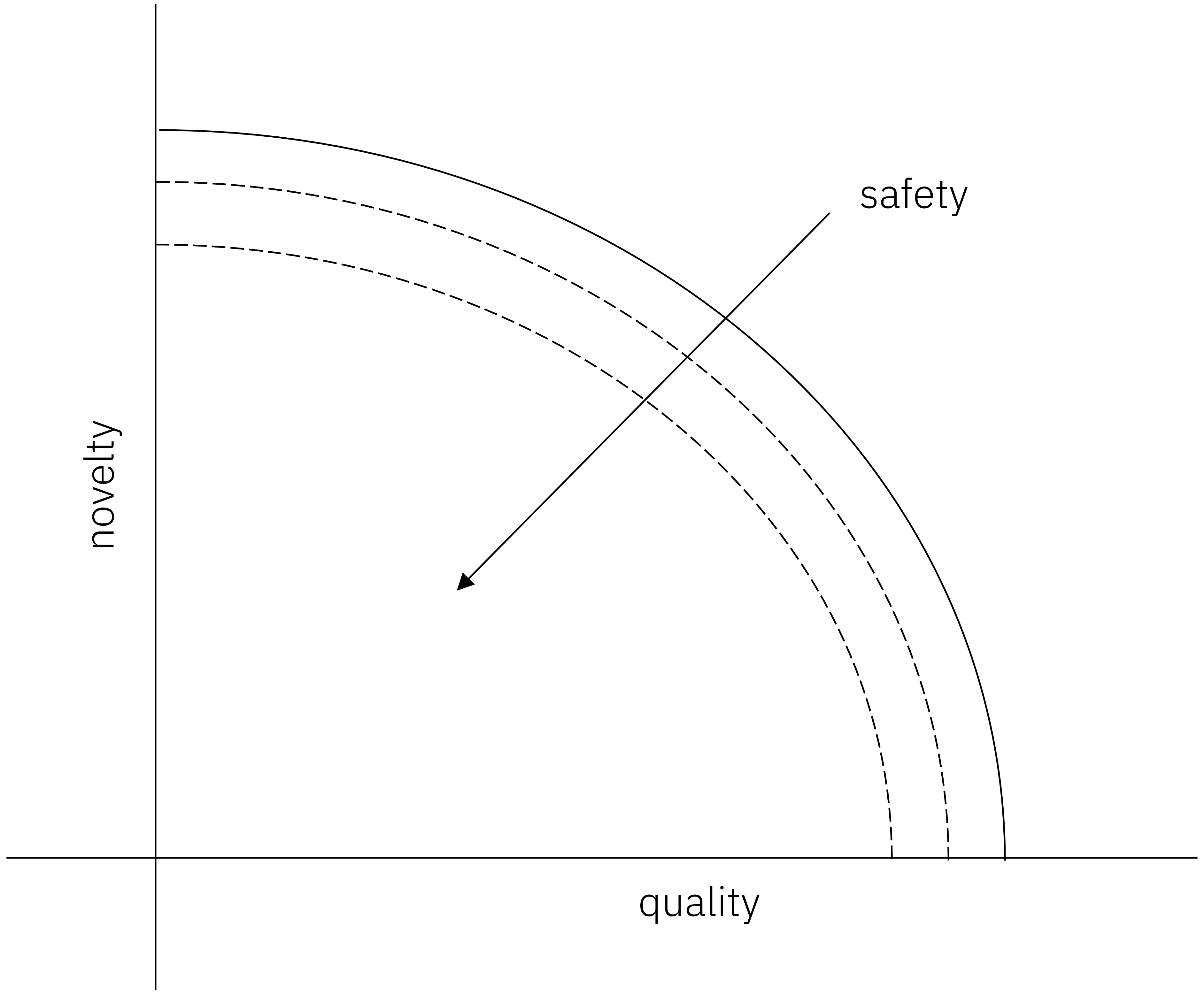
When creative artifacts are constrained, the region becomes smaller and creativity becomes more difficult.

(This statistical phenomenon of optimal creativity systems contrasts the computational phenomenon of humans often being more creative with more constraints.)

Tradeoff between safety and creativity

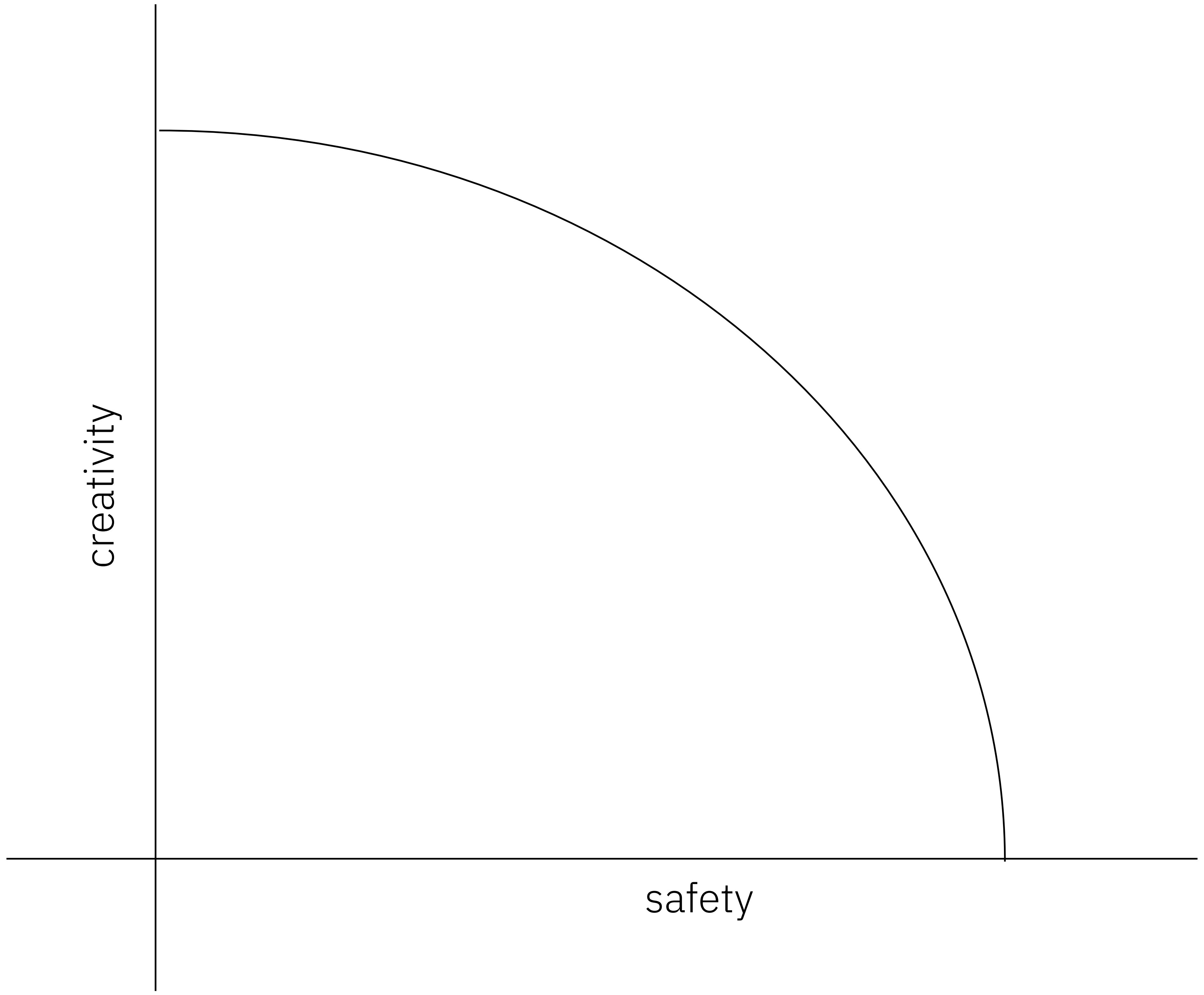
Safety is a constraint on artifacts.

Like other constraints, safety makes the feasible region under the quality-novelty tradeoff curve smaller and creativity more difficult.



Banality, the lack of creativity, follows from safety.

There is a tradeoff between safety and creativity.



Some applications of generative ML, like writing boilerplate, require safety whereas others, like inspiring a human poet, do not.

Some applications of generative ML, like writing poetry, require creativity and others, like writing boilerplate do not.

Applications requiring safety tend to also be ones not requiring creativity.

Applications not requiring safety tend to also be ones requiring creativity.

A trustworthy machine learning system is one that has sufficient:

1	2	3	4
basic performance	reliability	human interaction	aligned purpose

Kush R. Varshney

krvarshn@us.ibm.com
@krvarshney