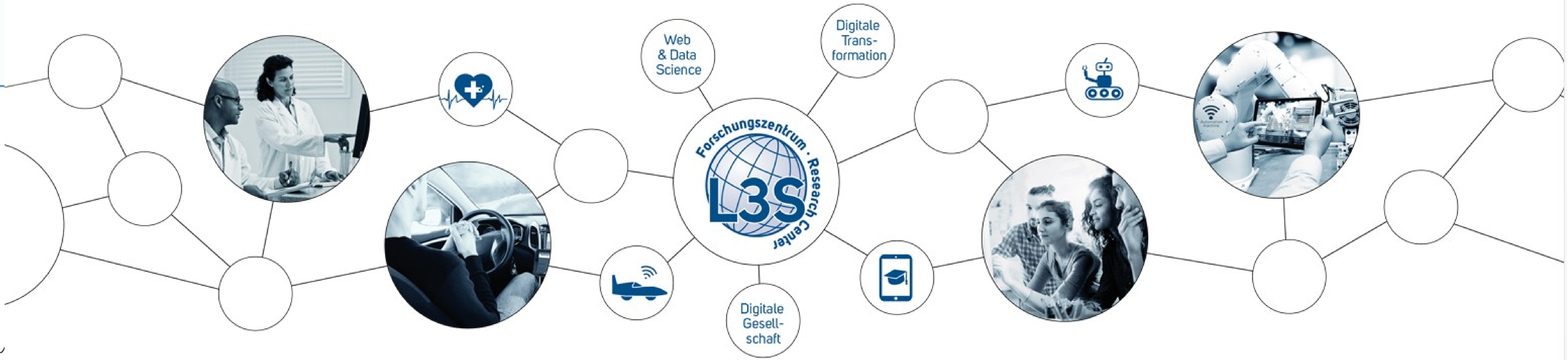
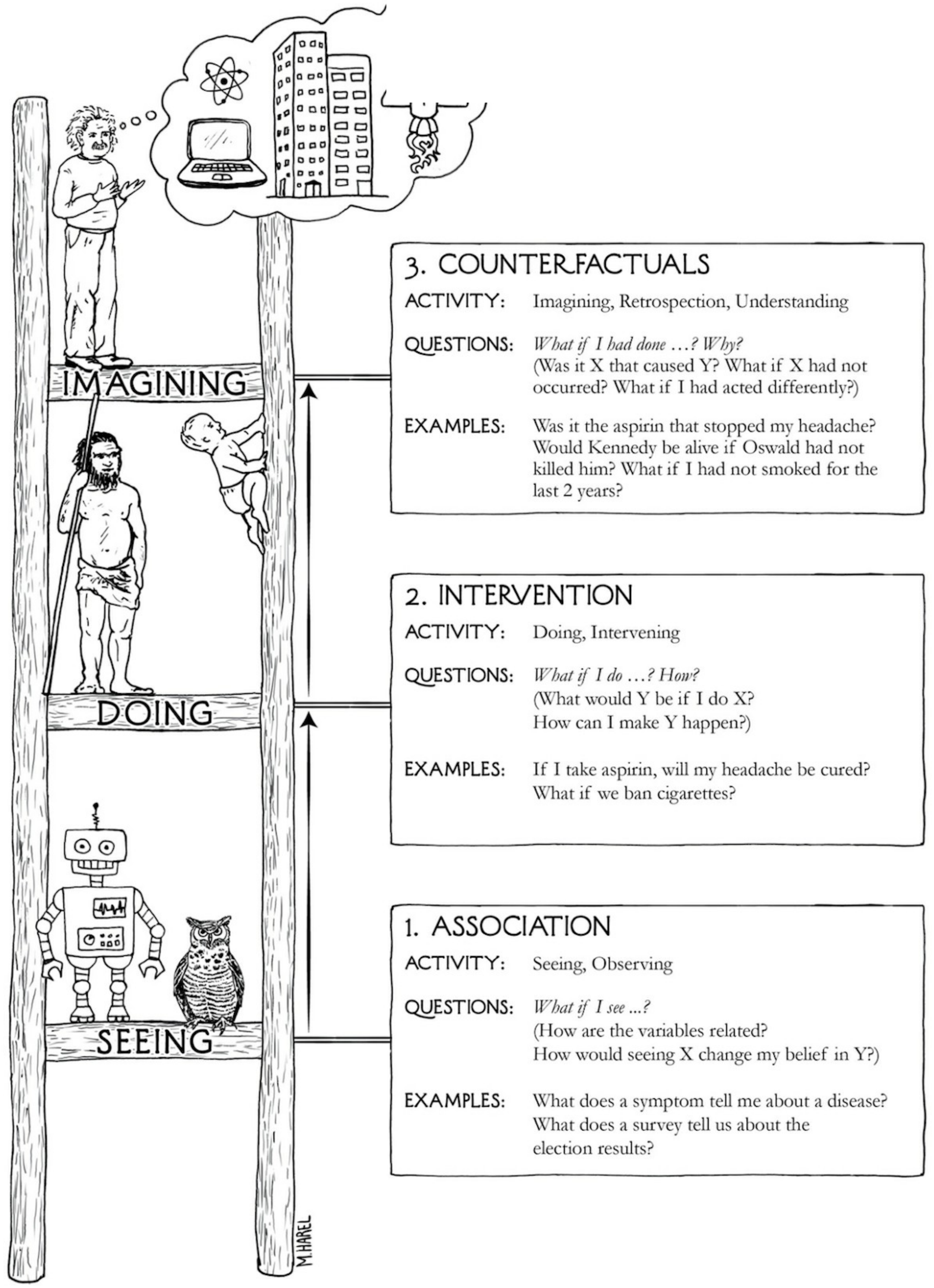




On the Role of Causality for Trustworthy Artificial Intelligence

Wolfgang Nejdl

L3S Research Center &
Leibniz Universität Hannover
Germany

Leibniz AILab

Storks Deliver Babies (p = 0.008) – Matthews, TS Vol 22/2, 2000

Country	Area (km^2)	Storks (pairs)	Humans (10^6)	Birth rate (10^3/yr)
Albania	28,750	100	3.2	83
Austria	83,860	300	7.6	87
Belgium	30,520	1	9.9	118
Bulgaria	111,000	5000	9.0	117
Denmark	43,100	9	5.1	59
France	544,000	140	56	774
Germany	357,000	3300	78	901
Greece	132,000	2500	10	106
Holland	41,900	4	15	188
Hungary	93,000	5000	11	124
Italy	301,280	5	57	551
Poland	312,680	30,000	38	610
Portugal	92,390	1500	10	120
Romania	237,500	5000	23	367
Spain	504,750	8000	39	439
Switzerland	41,290	150	6.7	82
Turkey	779,450	25,000	56	1576

Table 1. Geographic, human and stork data for 17 European countries

sample size. In our case, $n = 17$ so that $t = 3.06$, which for $(n - 2) = 15$ degrees of freedom leads to a *p*-value of 0.008.

◆ ANALYSIS ◆

What are we to make of this result, which points to a highly statistically significant degree of correlation between stork populations and birth rates? The correlation coefficient is not particularly high, but according to its *p*-value, there is only a 1 in 125 chance of obtaining at least as impressive a value *assuming* the null hypothesis of no correlation were true. Yet as with any *p*-value (and contrary to what unwary users of them believe),

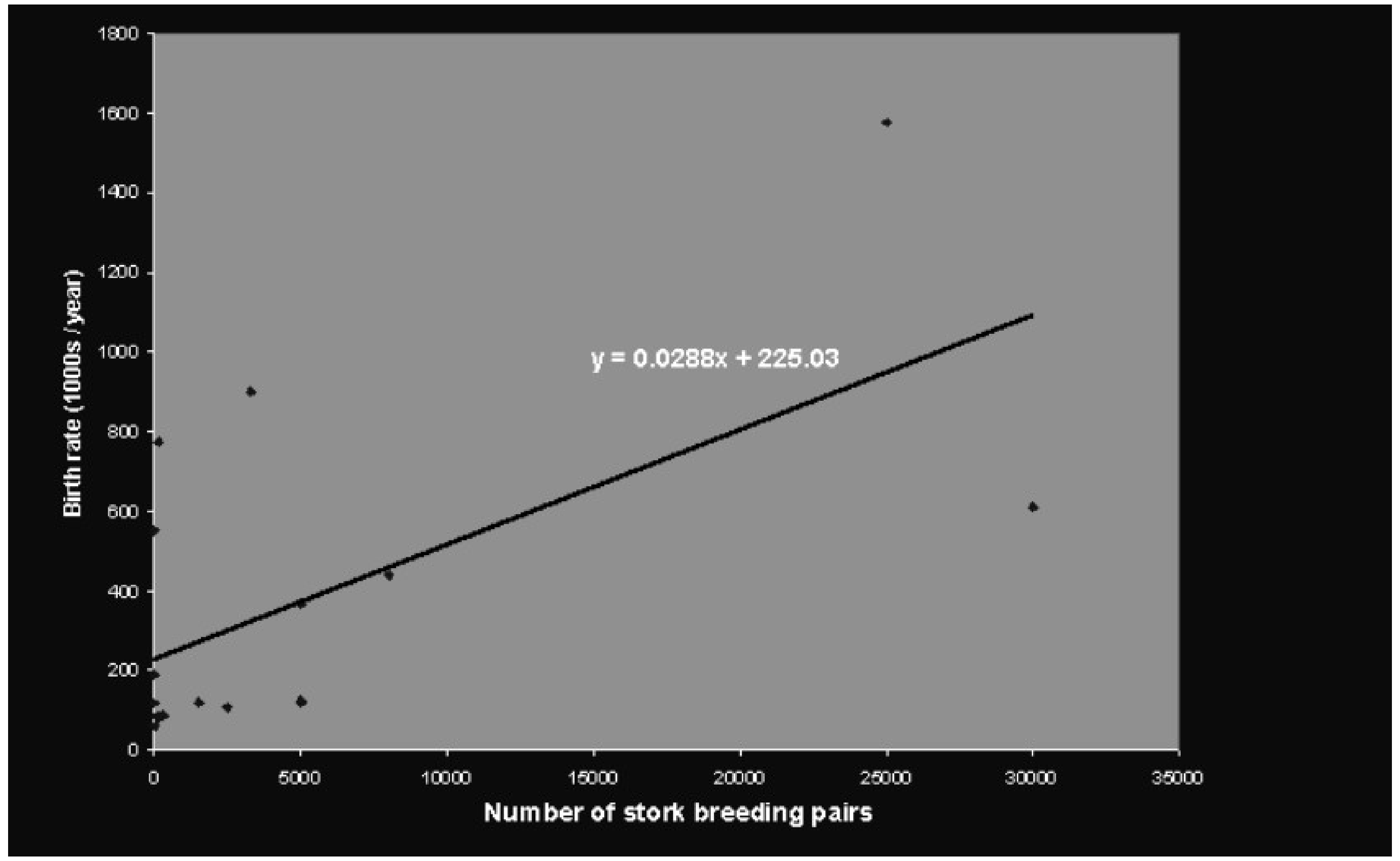
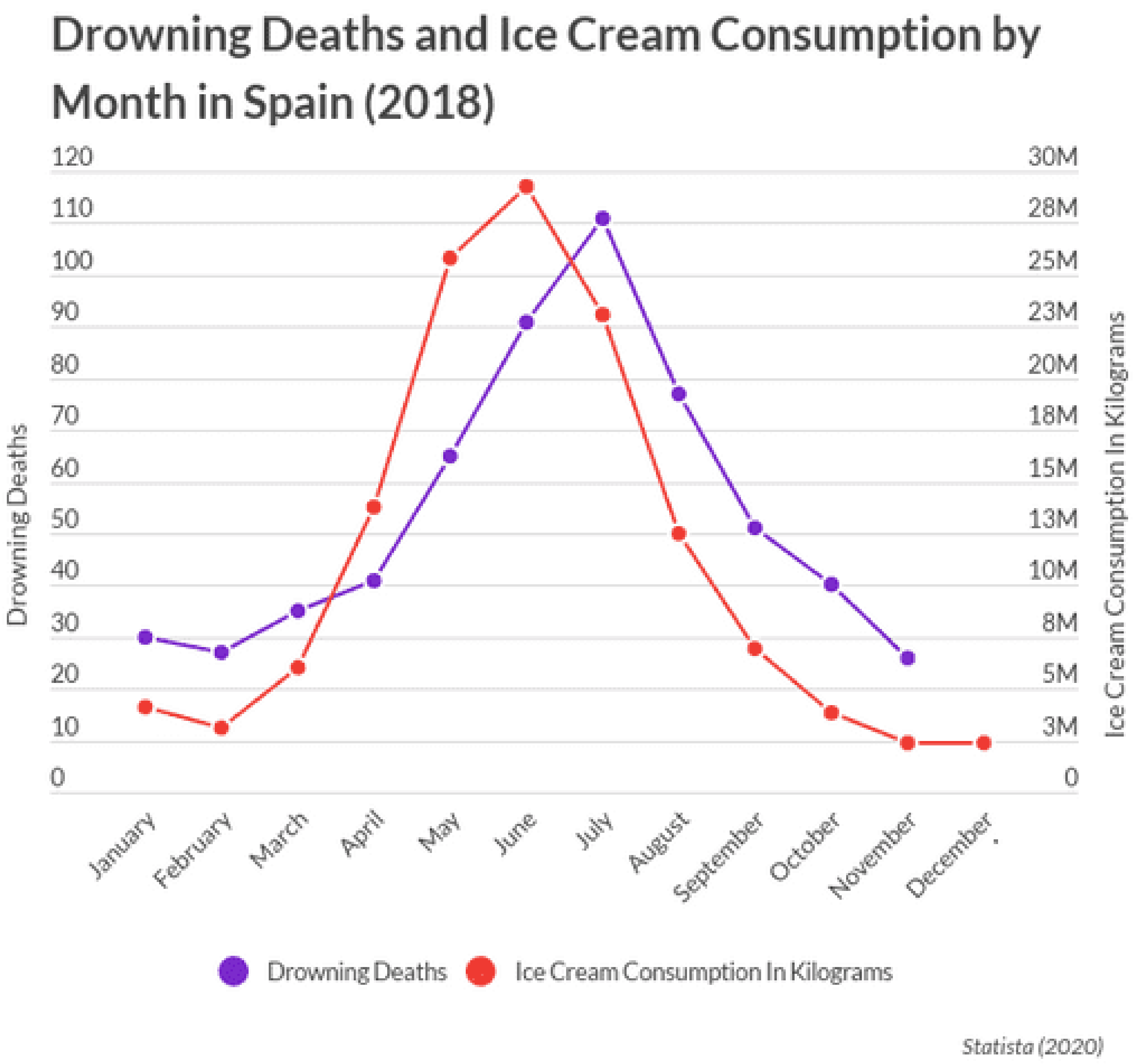
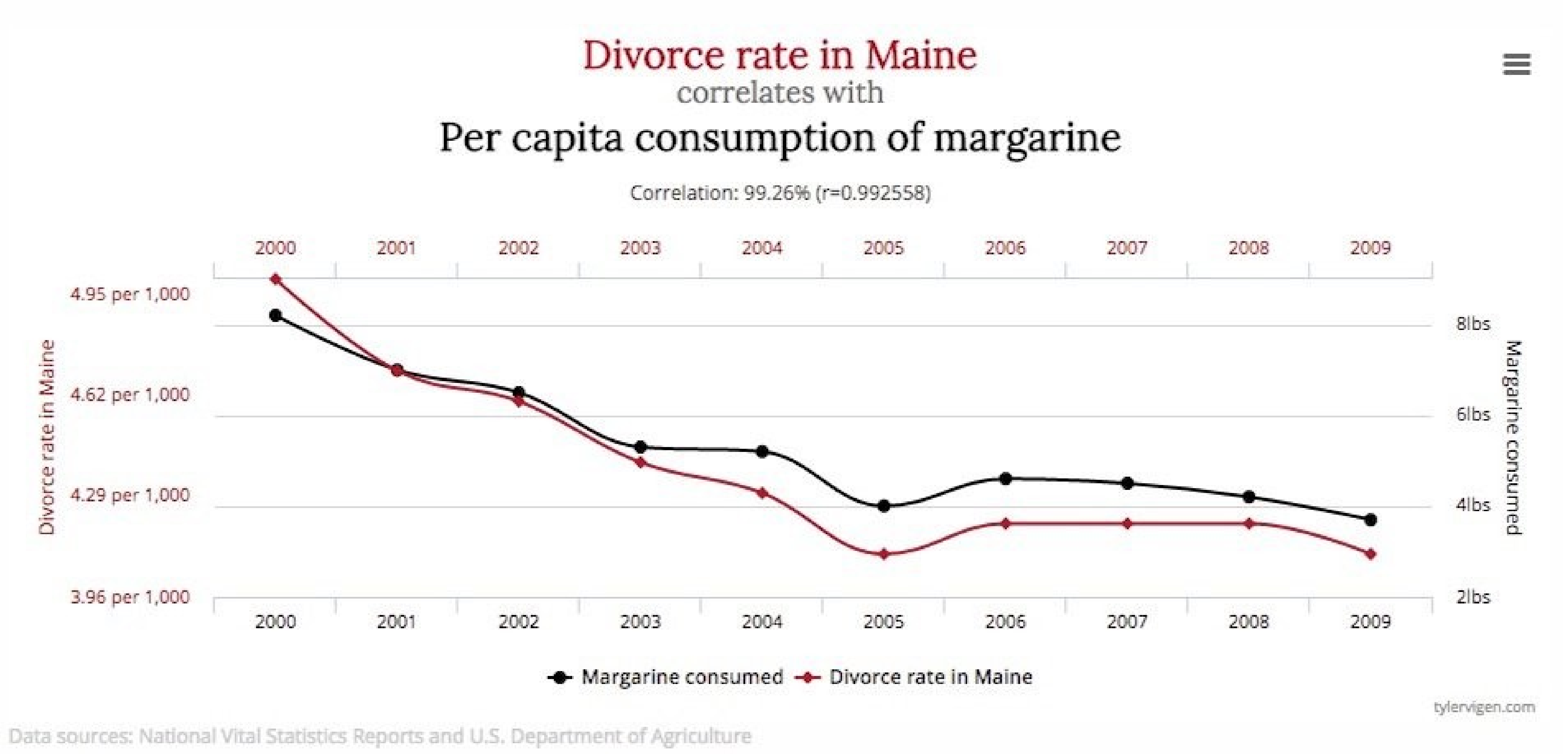


Fig 1. How the number of human births varies with stork populations in 17 European countries.

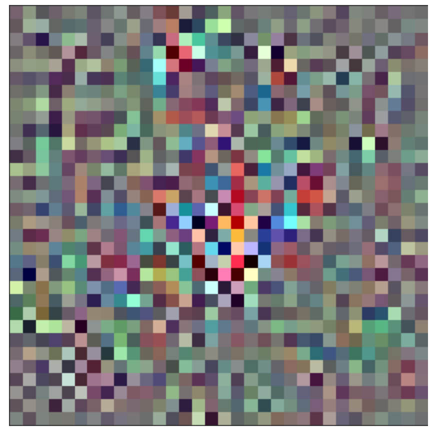

Drowning Deaths and Ice Cream Consumption by Month in Spain (2018)
Drowning Deaths
Ice Cream Consumption In Kilograms
120
110
100
90
80
70
60
50
40
30
20
10
0
30M
28M
25M
23M
20M
18M
15M
13M
10M
8M
5M
3M
0
January
February
March
April
May
June
July
August
September
October
November
December
Drowning Deaths
Ice Cream Consumption In Kilograms
Statista (2020)


Divorce rate in Maine
correlates with
Per capita consumption of margarine
Correlation: 99.26% (r=0.992558)
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
4.95 per 1,000
4.62 per 1,000
4.29 per 1,000
3.96 per 1,000
Divorce rate in Maine
8lbs
6lbs
4lbs
2lbs
Margarine consumed
Margarine consumed
Divorce rate in Maine
tylervigen.com
Data sources: National Vital Statistics Reports and U.S. Department of Agriculture

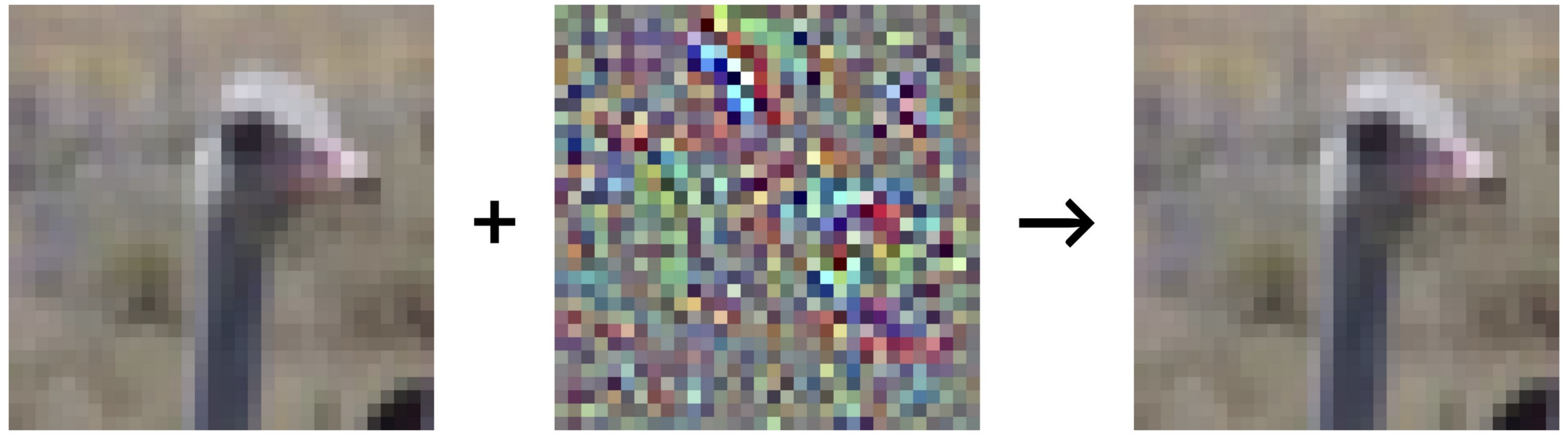
Neural networks find the „best“ features (Madry 2019)

- *Adversarial examples are not bugs, they are features.* Ilyas et al, NIPS 2019.
- https://gradientscience.org/adv/

- A tale about the planet ERM, inhabited by an alien race known as Nets.
- Each individual’s place in the social hierarchy is determined by their ability to classify bizarre 32-by-32 pixel images (meaningless to the Nets) into ten completely arbitrary categories.
- These images are drawn from a top-secret dataset, See-Far—outside of looking at those curious pixelated images, the Nets live their lives totally blind.



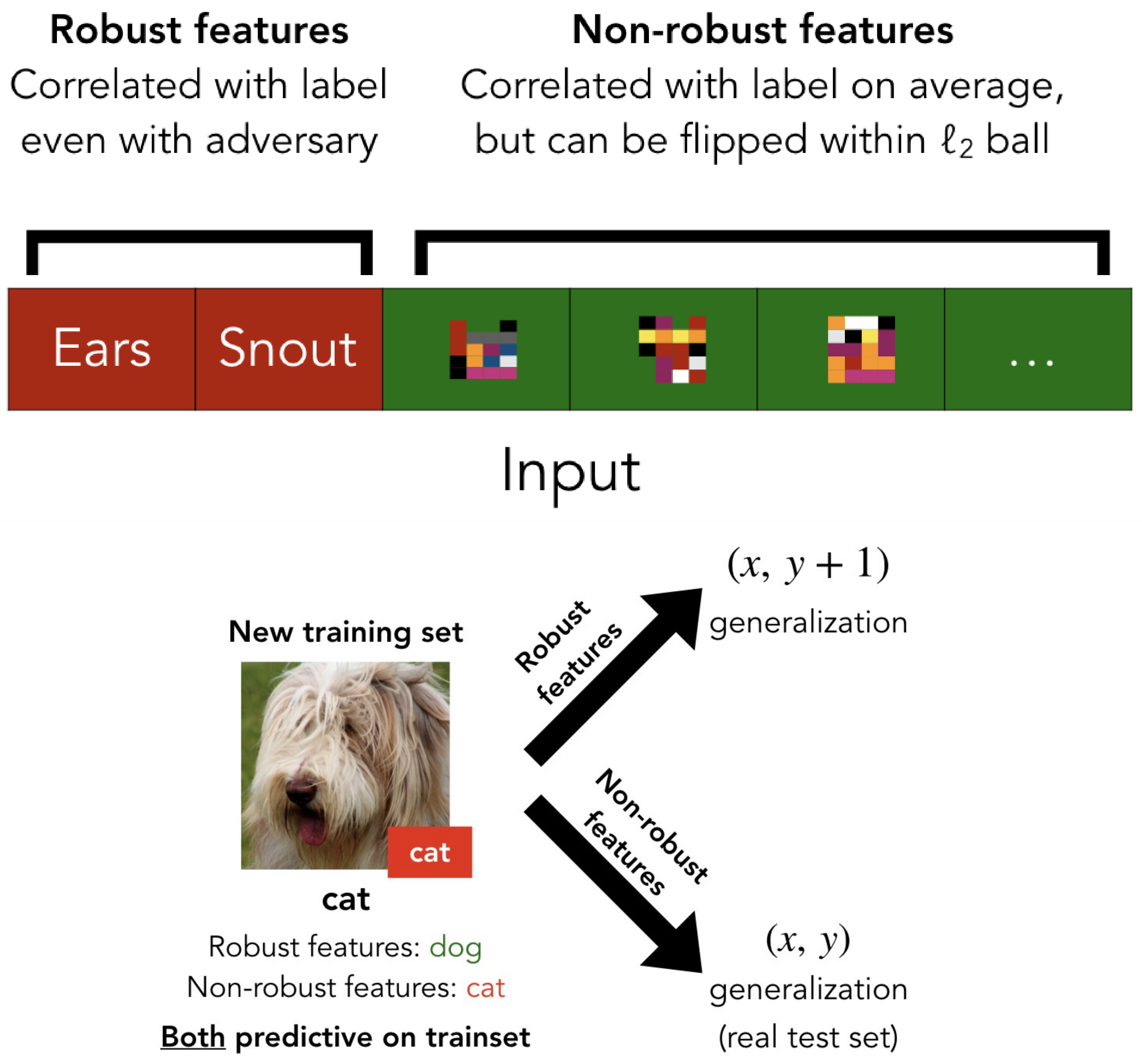
A TOOGIT, highly indicative of a "1" image. Nets are extremely sensitive to TOOGITs.



On the left is a “2”, in the middle there is a GAB pattern, which is known to indicate “4”—unsurprisingly, adding a GAB to the image on the left results in a new image, *which looks exactly like an image corresponding to the “4” category*.

Neural networks find the „best“ features (Madry 2019)

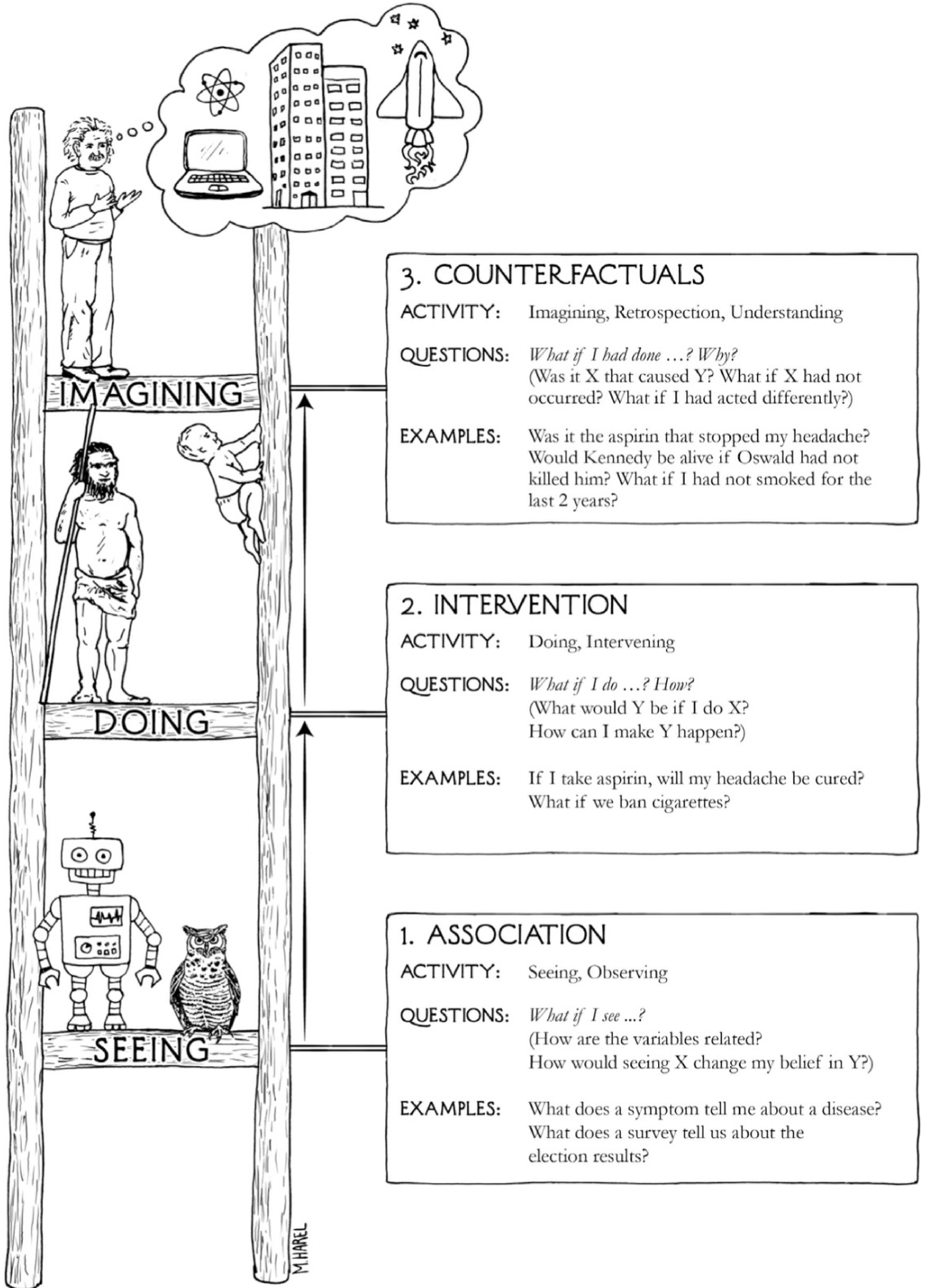
- Every training set includes „robust features“ (usually used by humans) and „non-robust features“ (which are brittle and can be disturbed easily)
- Adversarial training tries to disturb these non-robust features to make them useless as discriminators
- Interpretability and causality considerations have to be included already in the training phase
- post-hoc explanation of standard models (which might use these non-robust features) is less useful (as we cannot explain these non-robust features to a human)



Causality: The science of cause and effect

Pearl’s Ladder of Causality

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?

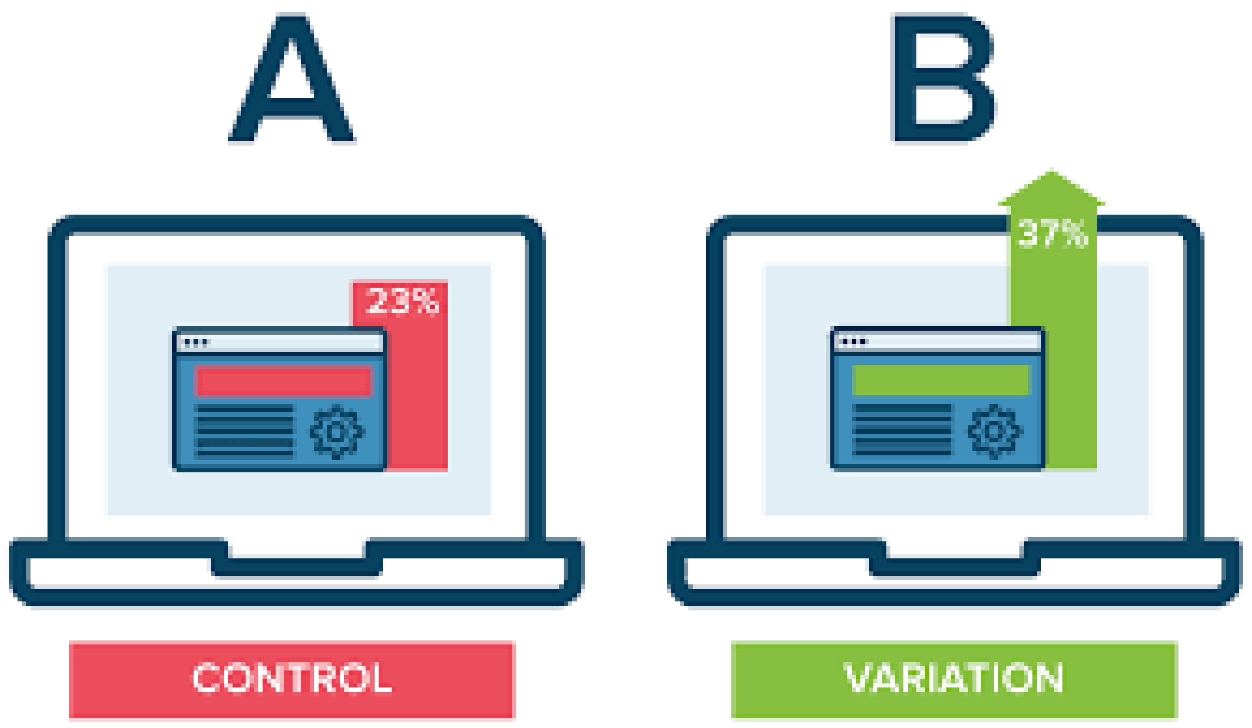


Pearl, Mackenzie: The Book of Why, Basic Books, 2020.

Primer on Causality

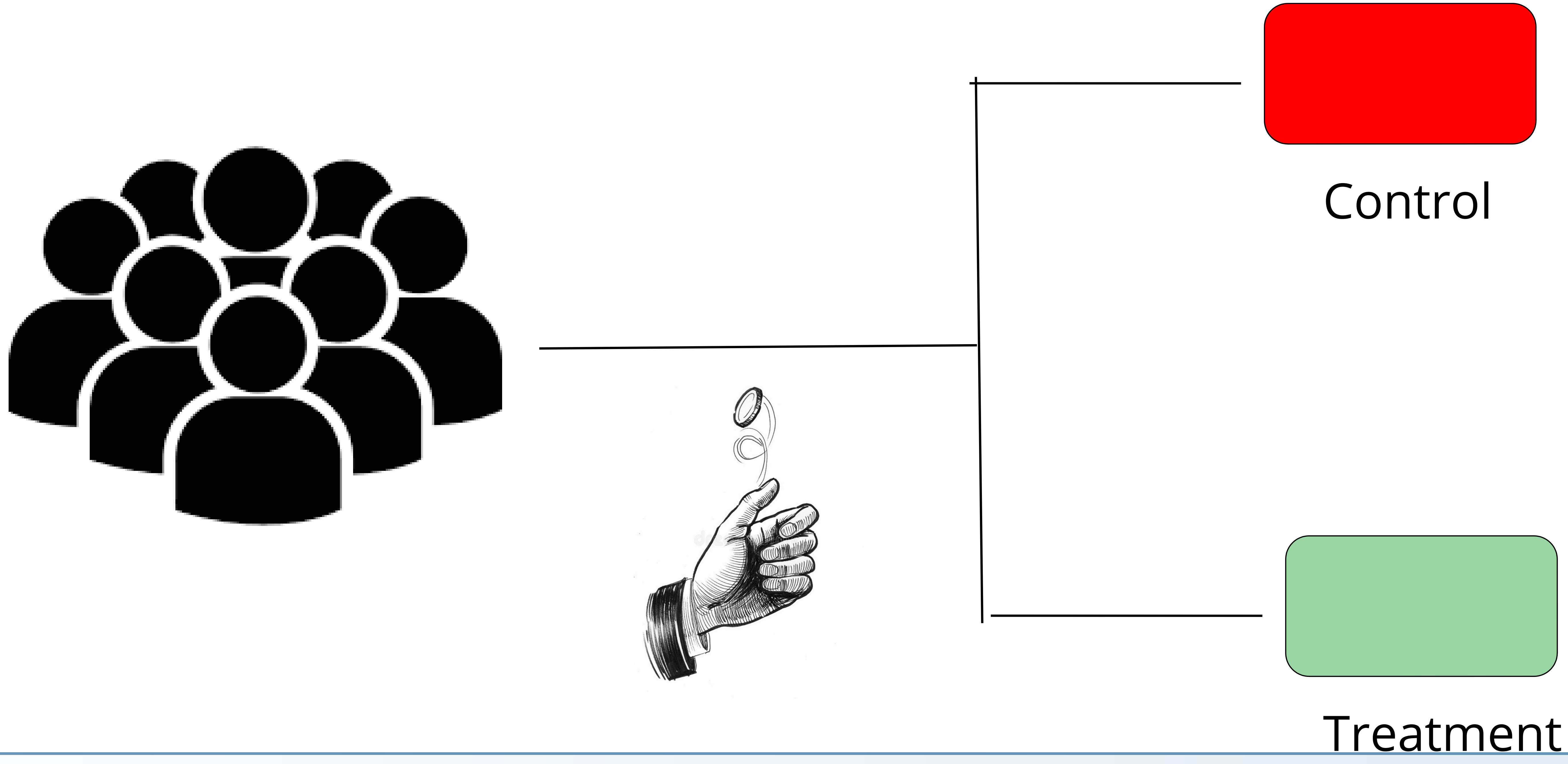
What is causality?

- Science of cause and effect – we can experiment …
  - Random controlled trials in medicine (Is the drug effective?),
    Web A/B testing (Will changing the interface or algorithm lead to more clicks?)



Randomized Control Trials

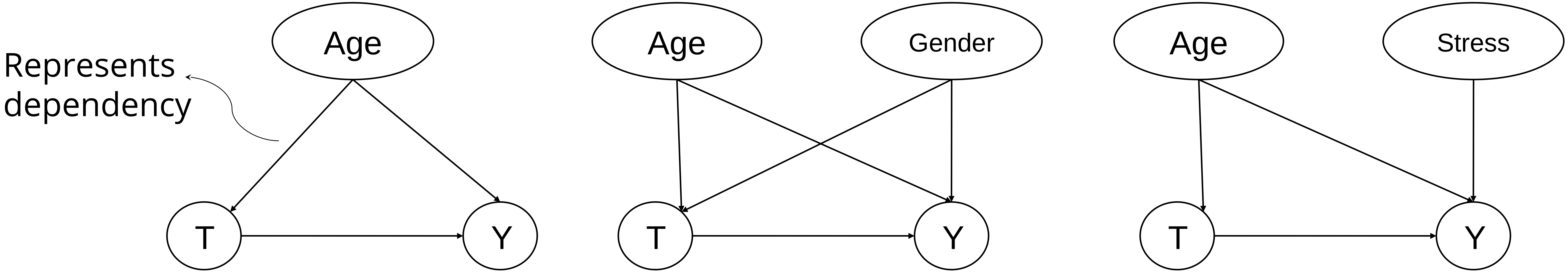
- We want to understand if X causes Y (e.g., whether changing the appearance of the website (X) increases number of clicks (Y))



Experimental vs. Observational Data

- Performing experiments is often not possible
- Only observed data is available
    - Experiments might not have been performed perfectly
    - Selection bias when deciding control/treatment individuals
    - Much easier to collect data on the Social Web than to do experiments
- How can we measure causal effect with observed data?
- Two models
    - Potential outcome framework
    - Graphical and Structural causal models

Graphical Models

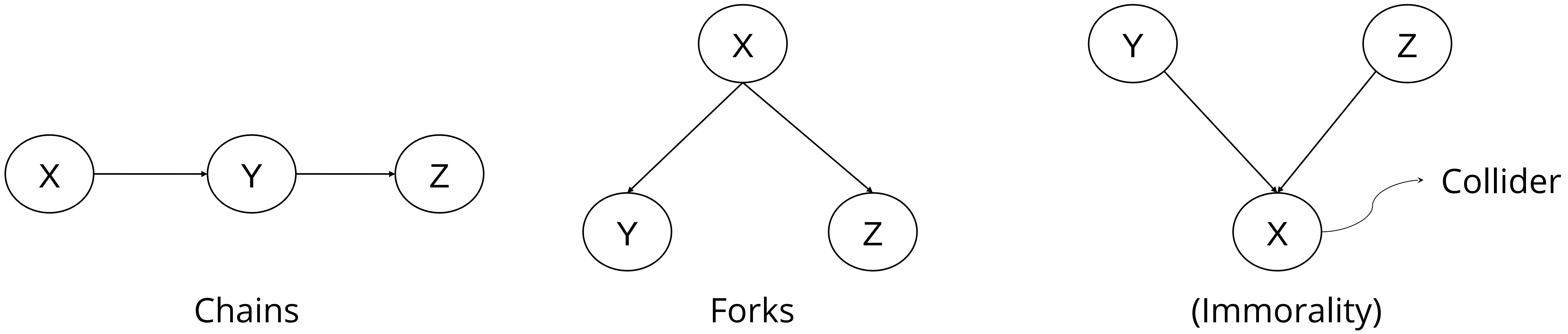


X = {Age}

X = {Age, Gender}

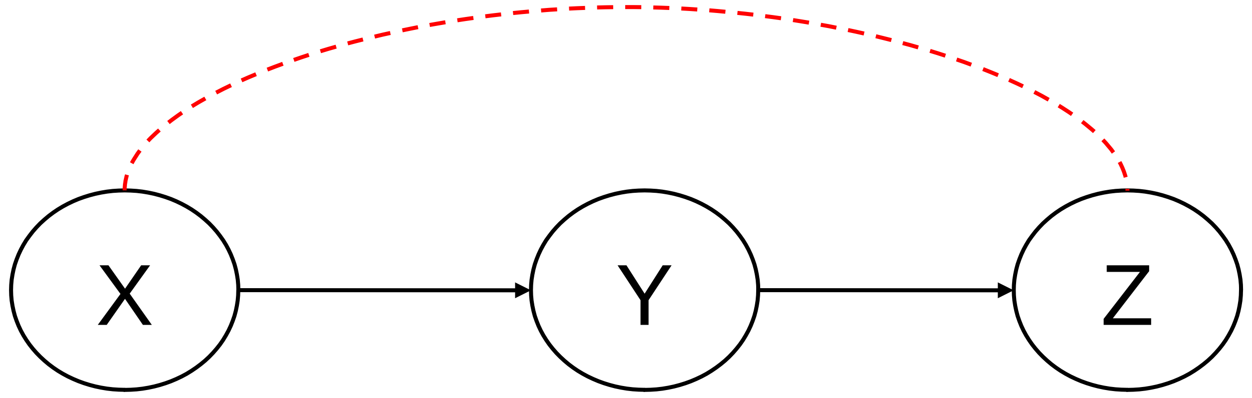
X = {Age}

Graphical Models

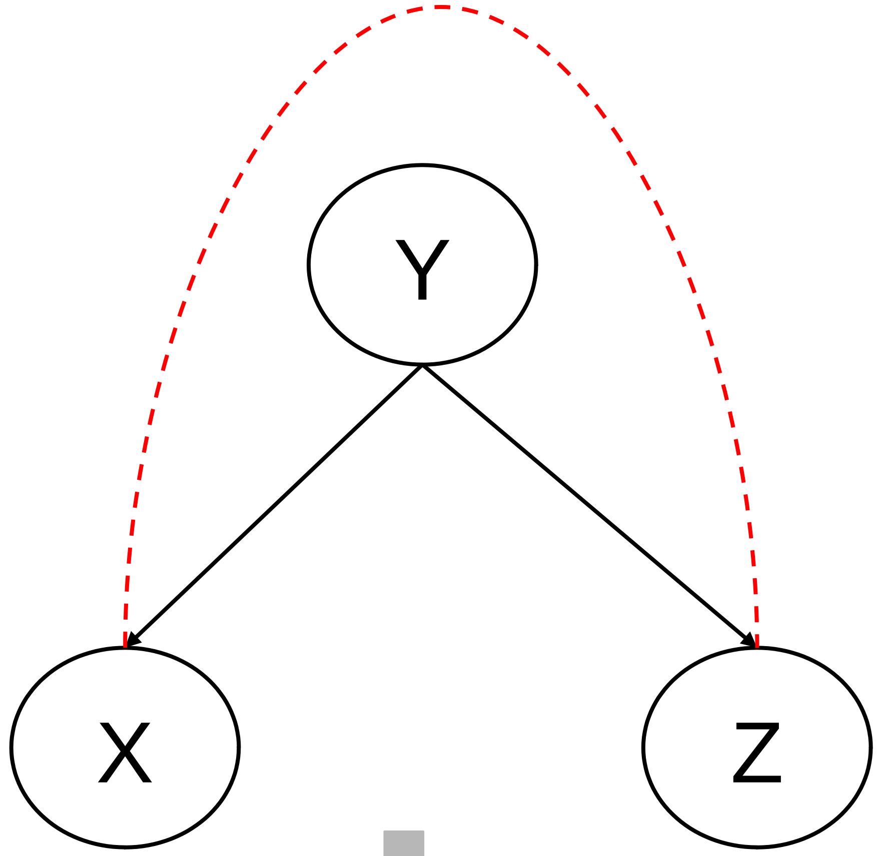


Graphical Models: (In)dependence

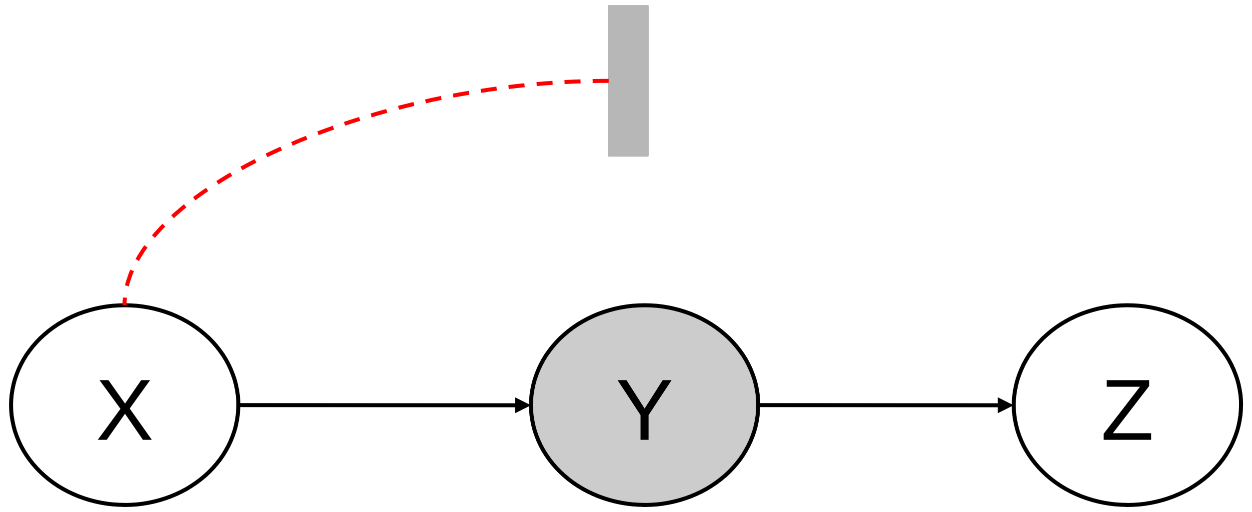
Association

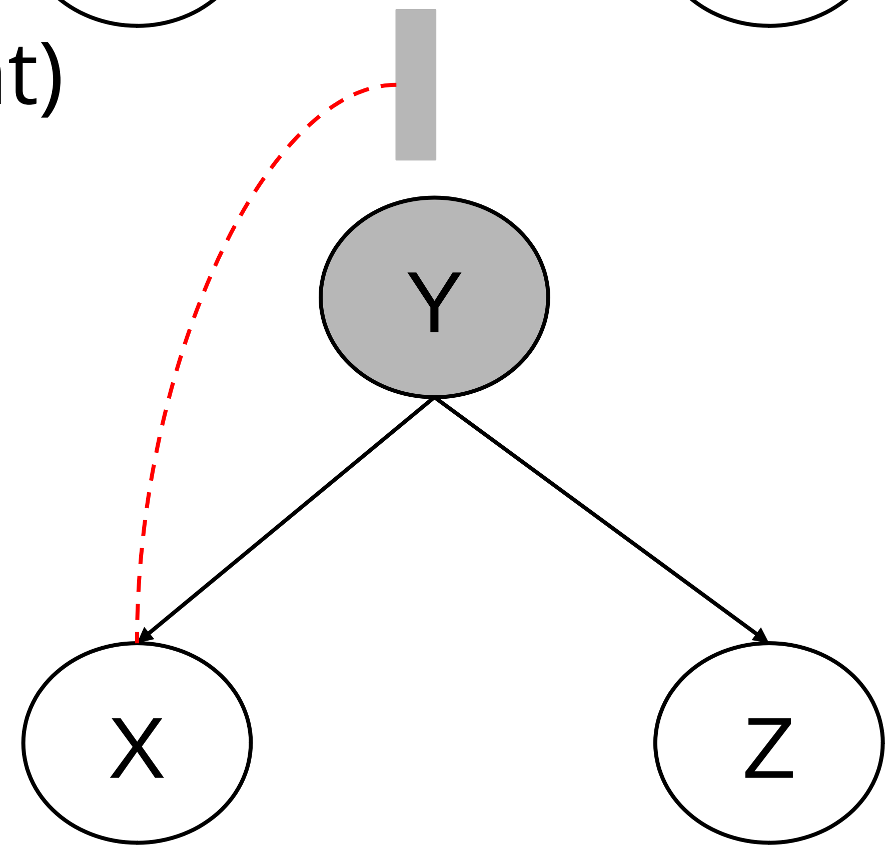


Association



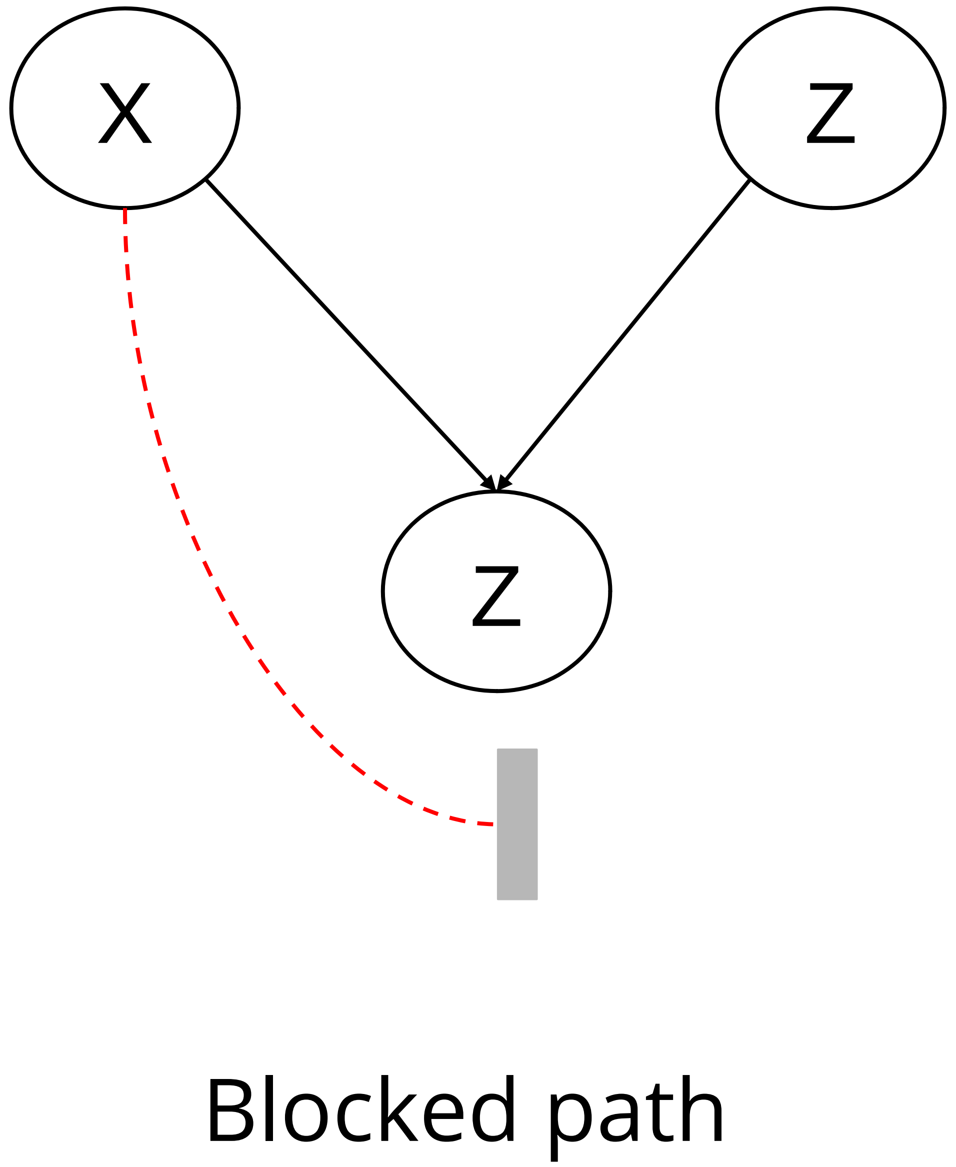
Unblocked path (X and Z are dependent)

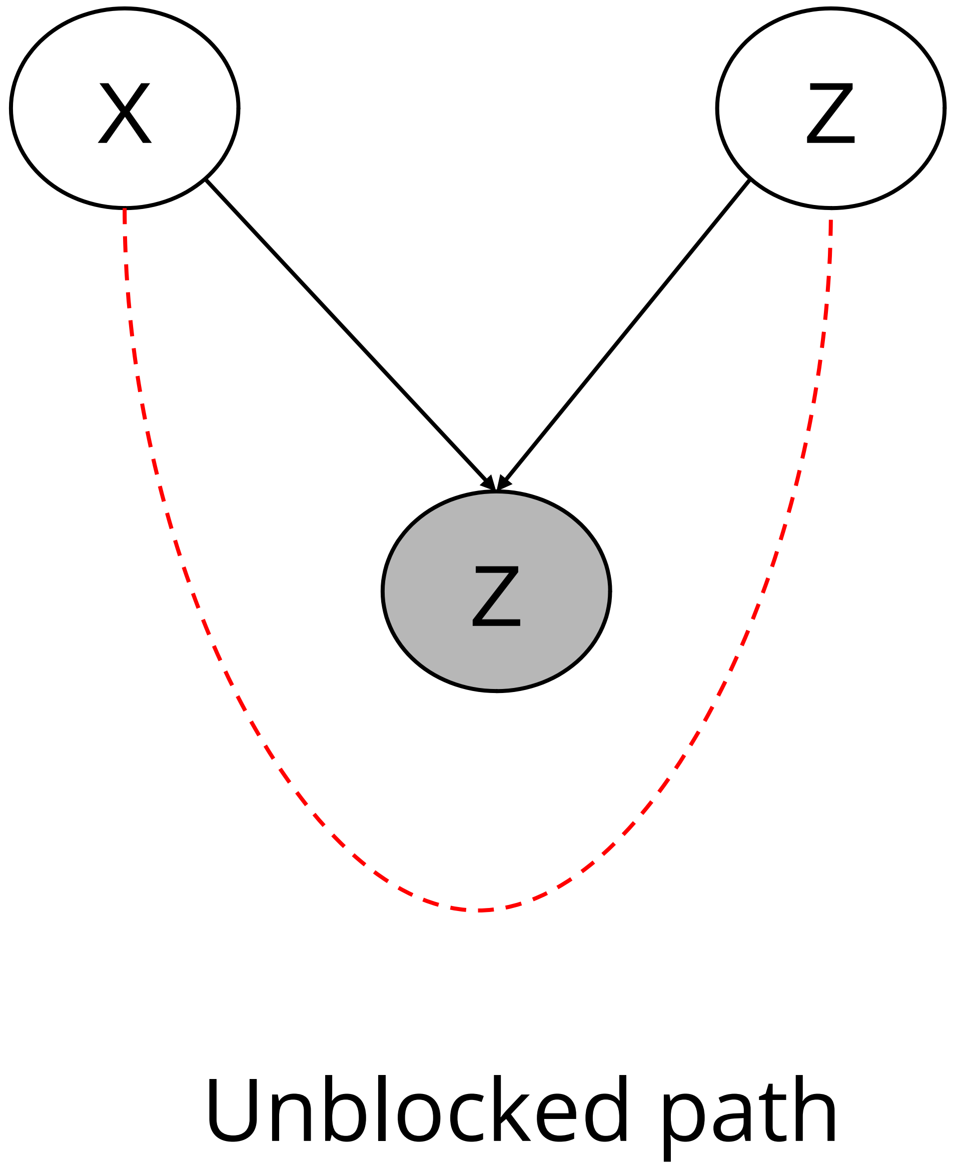




Blocked path (X and Z are independent)

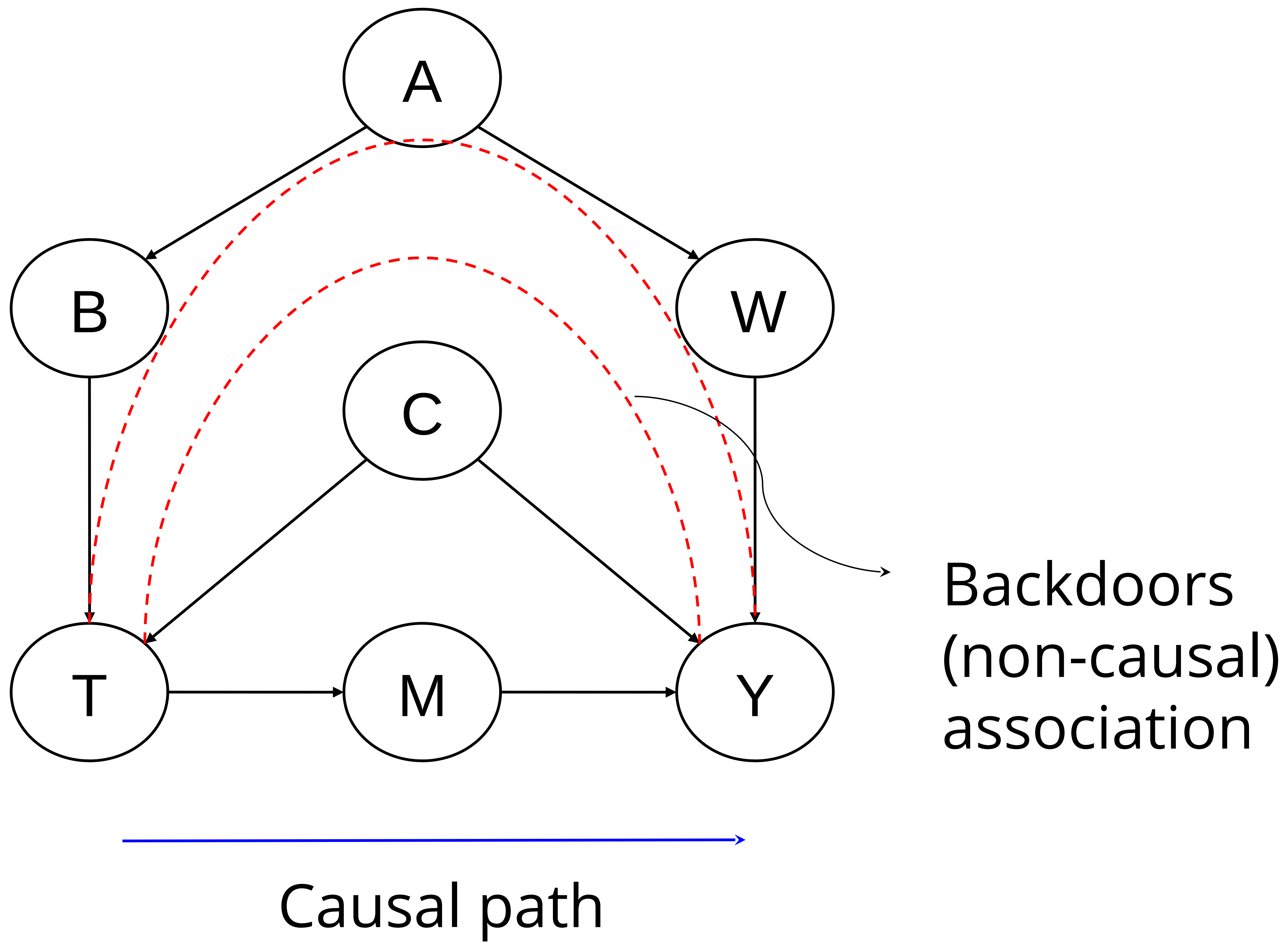
Graphical Models: (In)dependence

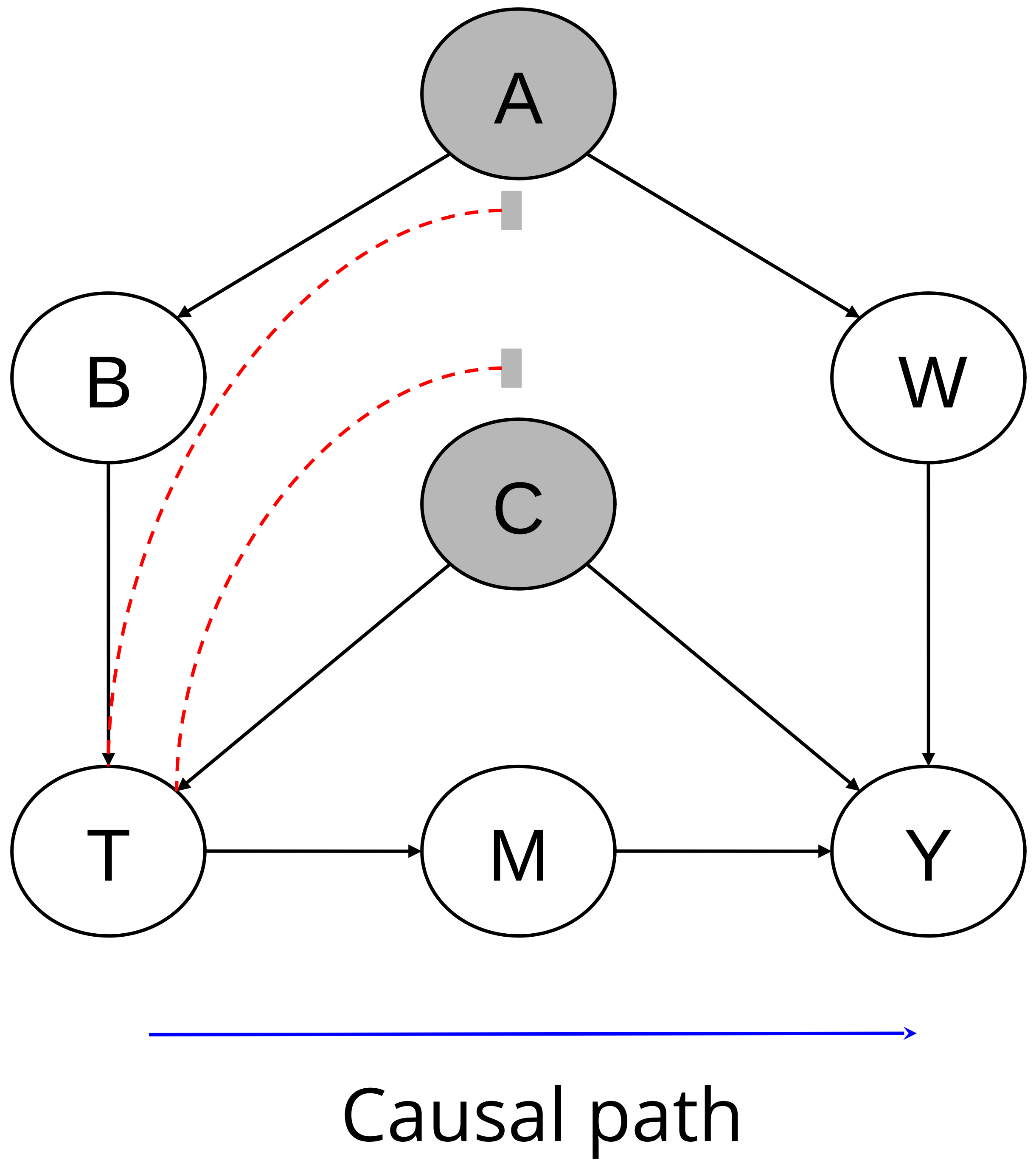




- Conditioning on the collider or any of its descendants unblocks the path
- D-separation in probabilistic graphical models

Graphical Models: Backdoor adjustments





Structural Causal Models

Structural equation for A as a cause of B

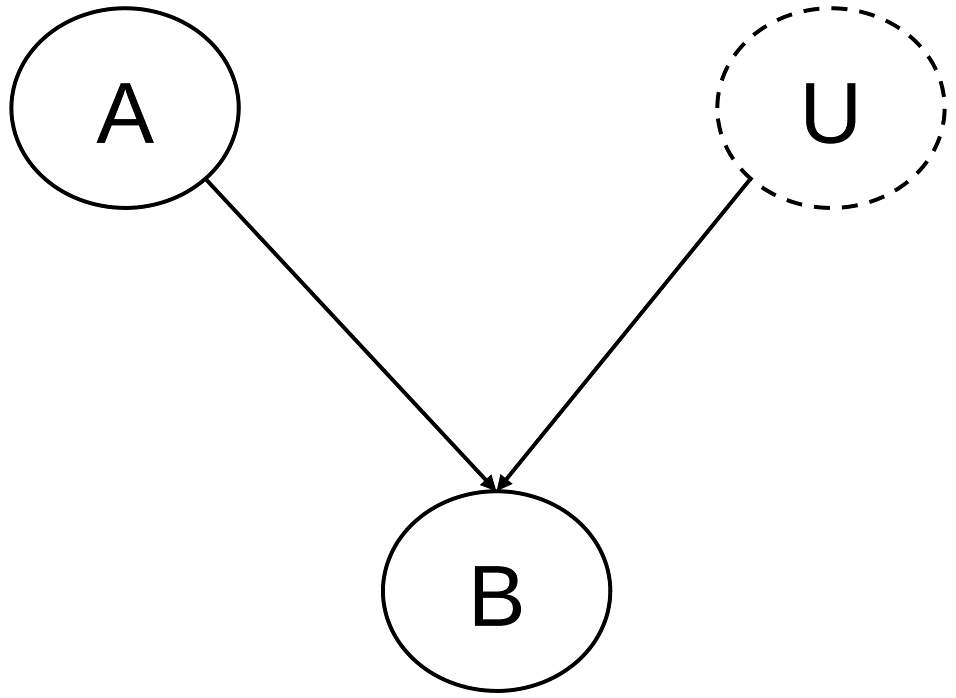
$$B := f(A)$$

Equality does not convey any causal information

Unobserved characteristics: Incorporates stochasticity

$$B := f(A, U)$$

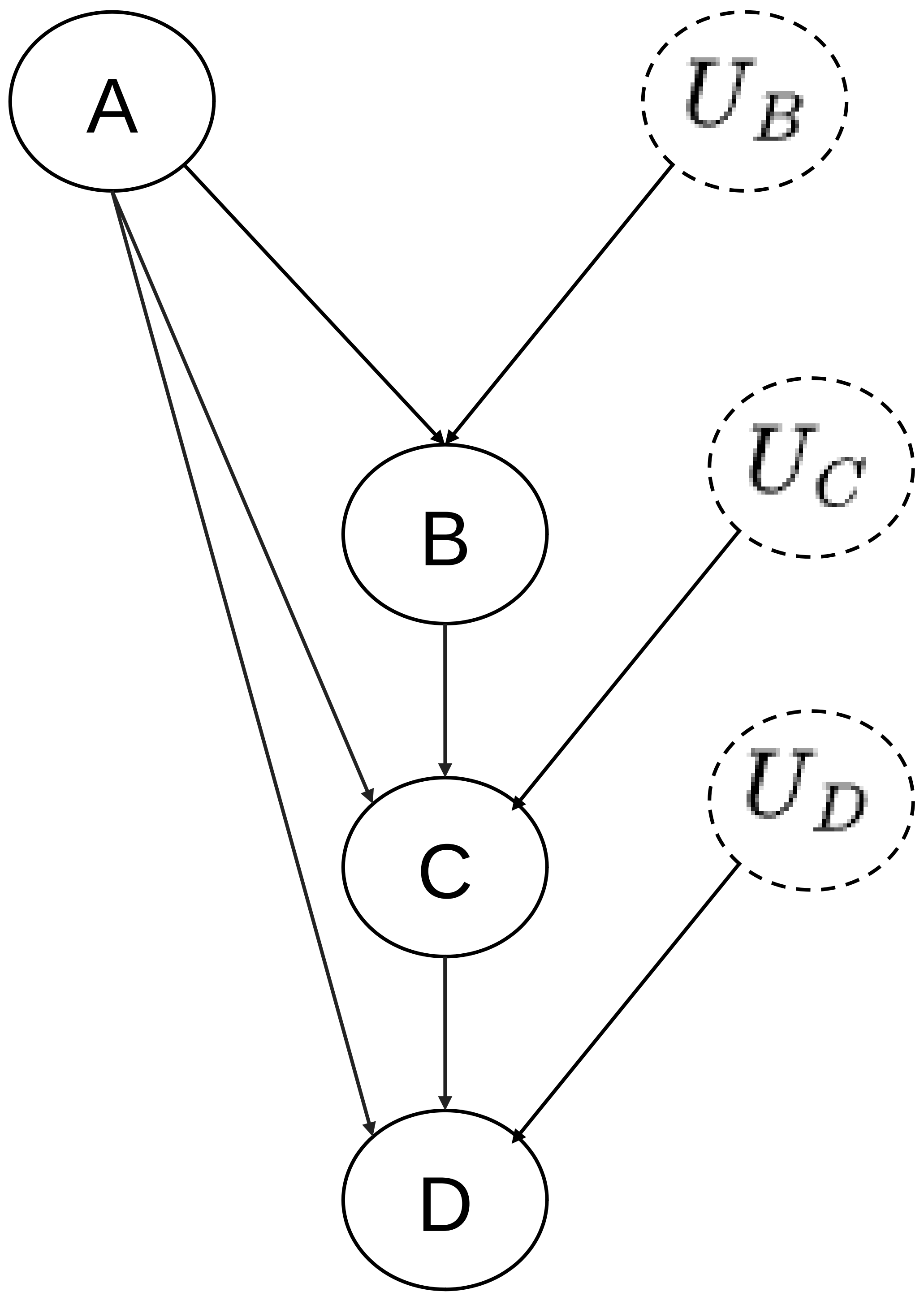
Causal mechanism: $X_i := f(\underbrace{A, B, \ldots}_{\text{Parents of } X_i})$



Structural Causal Models

$$
M: \begin{aligned} B &:= f_B(A, U_B) \\ C &:= f_C(A, B, U_C) \\ D &:= f_D(A, C, U_D) \end{aligned}
$$

- Set of endogenous variables
- Set of exogenous variables
- A set of functions, each to generate a endogenous variable from other variables



Causality and Trustworthy AI

We follow the dimensions described by the TAILOR project

- Interpretability: Providing meaningful explanations to users
- Fairness: Developing debiased and non-discriminating AI systems
- Robustness: Decreasing sensitivity towards input changes
- Privacy: Defending against privacy-evasive attacks
- Safety and Accountability: Auditing AI systems

For all references and details see:

Niloy Ganguly, Dren Fazlija, Maryam Badar, Marco Fisichella, Sandipan Sikdar, Johanna Schrader, Jonas Wallat, Koustav Rudra, Manolis Koubarakis, Gourab K. Patro, Wadhah Zai El Amri, Wolfgang Nejdl: **A Review of the Role of Causality in Developing Trustworthy AI Systems.** Feb 2023, https://arxiv.org/abs/2302.06975

Interpretability and Causality

Why do we need causal explanations?

Interpretability is often sacrificed for generalizability

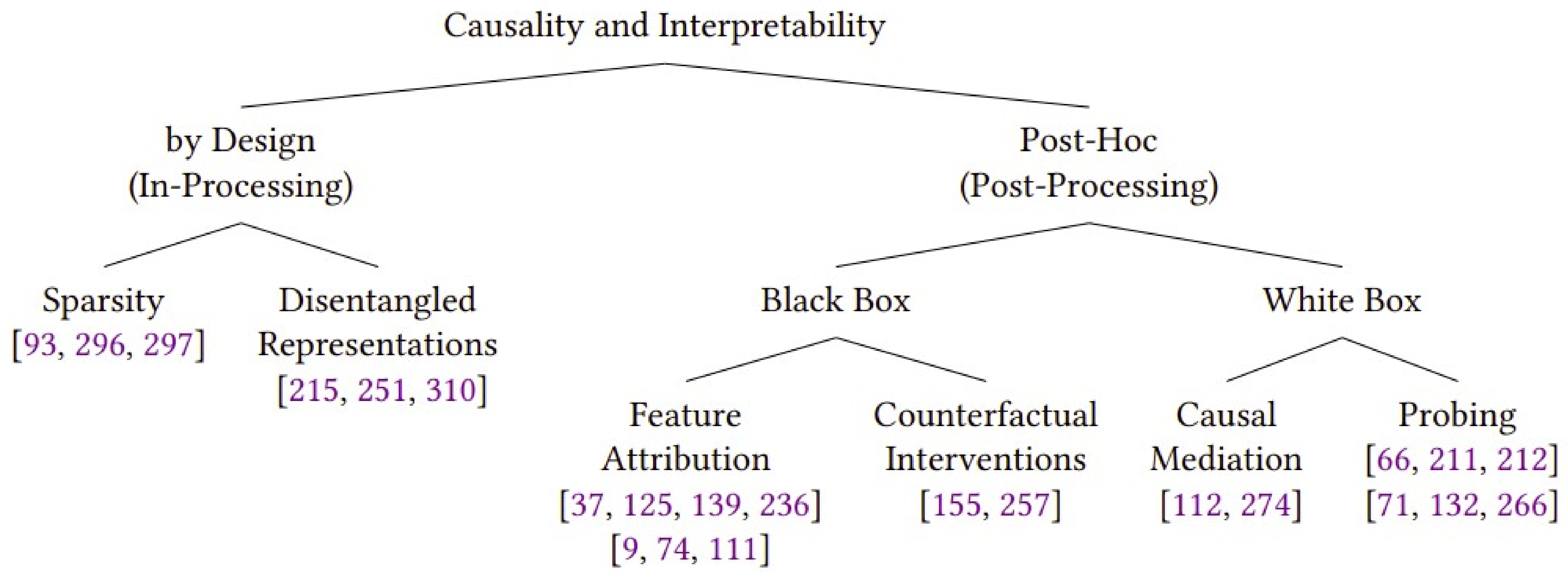
High-stake scenarios like medicine will need (and legally require) interpretability

Causal explanations can ensure that the true reasons for a prediction are communicated

Causality has been used to increase interpretability

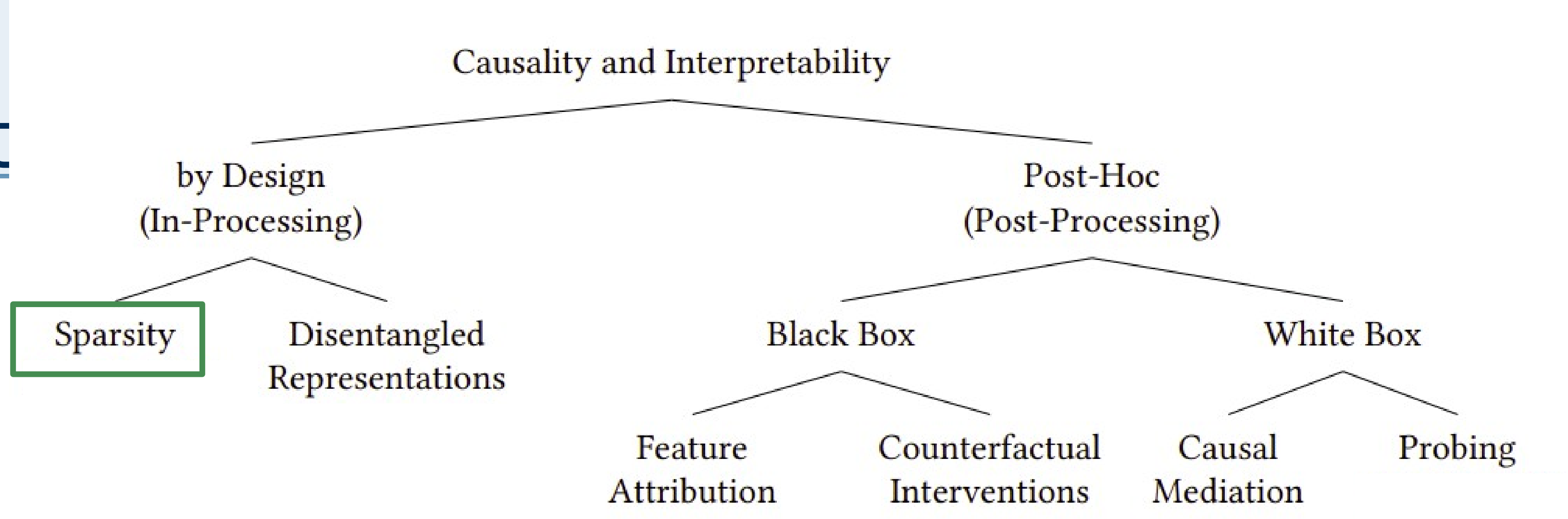
- Mainly for classifications tasks in computer vision and NLP

Interpretability



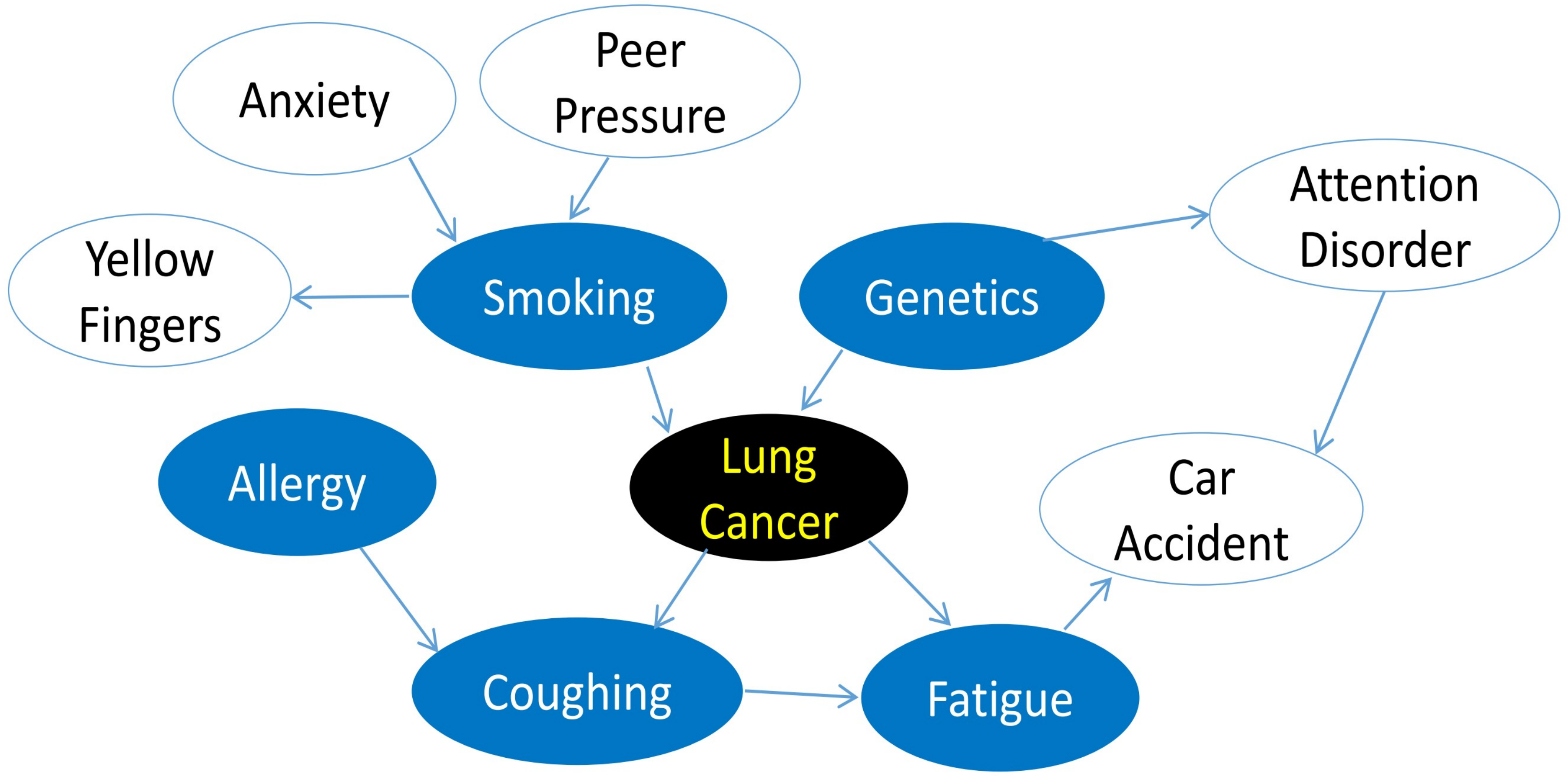
Like traditional interpretability: some models are interpretable by their model design and some methods provide post-hoc explanations for non-interpretable models

Causality-based Feature Selection: Methods and Evaluations. [Yu2020]

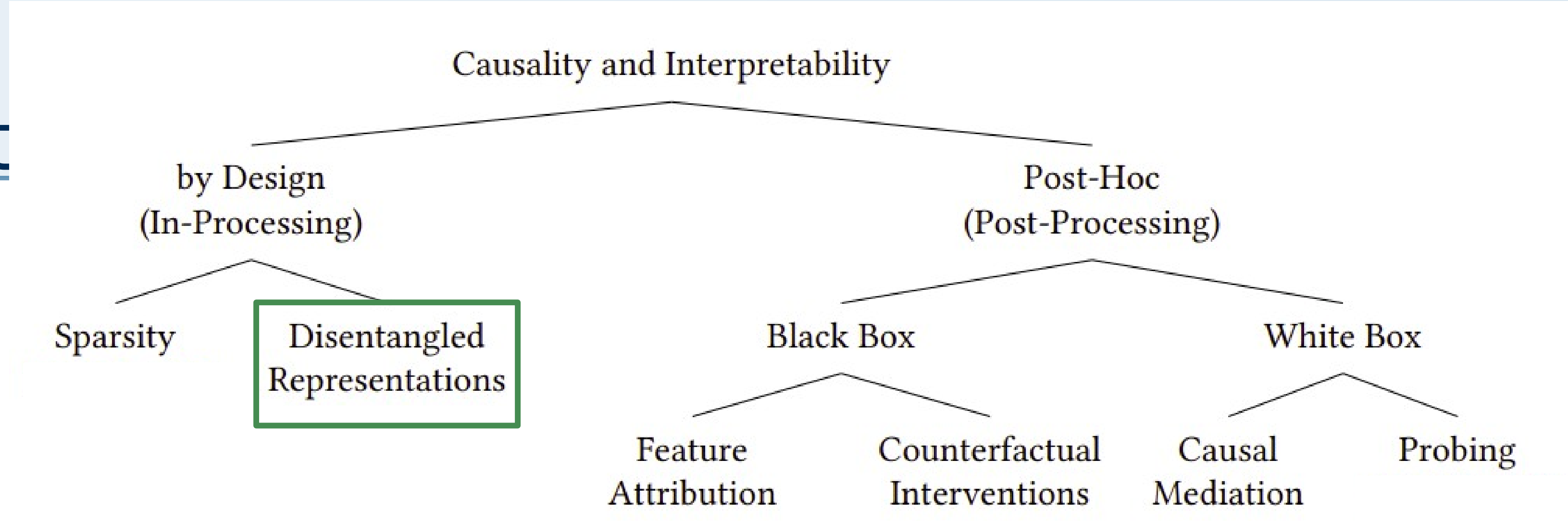


Goal: Categorize the data's attributes into relevant / irrelevant features

Solution: under the faithfulness assumption the Markov boundary of a variable in a Bayesian Network describes the variable's local causal relationships

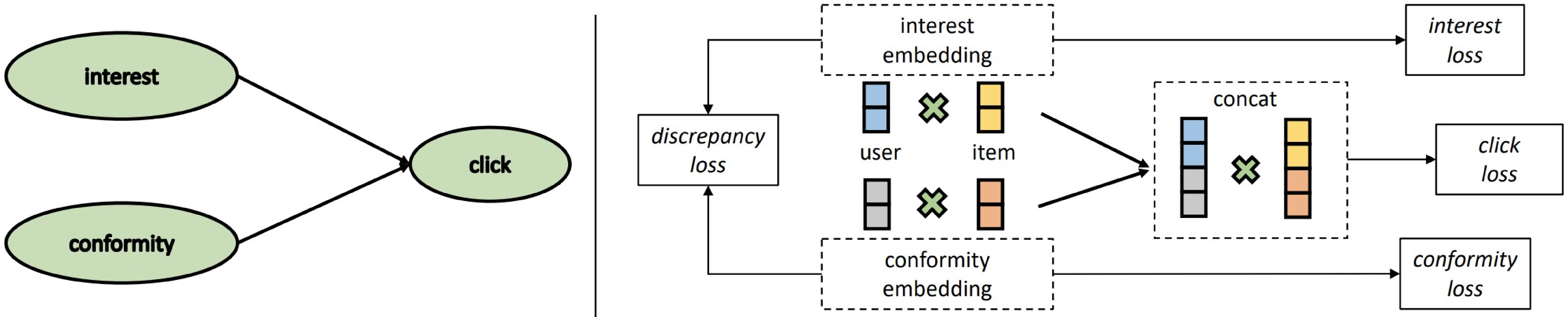


Disentangling User Interest and Conformity for Recommendation with Causal Embedding. [Zheng2021]



Goal: Disentangled latent representations that represent human-understandable concepts

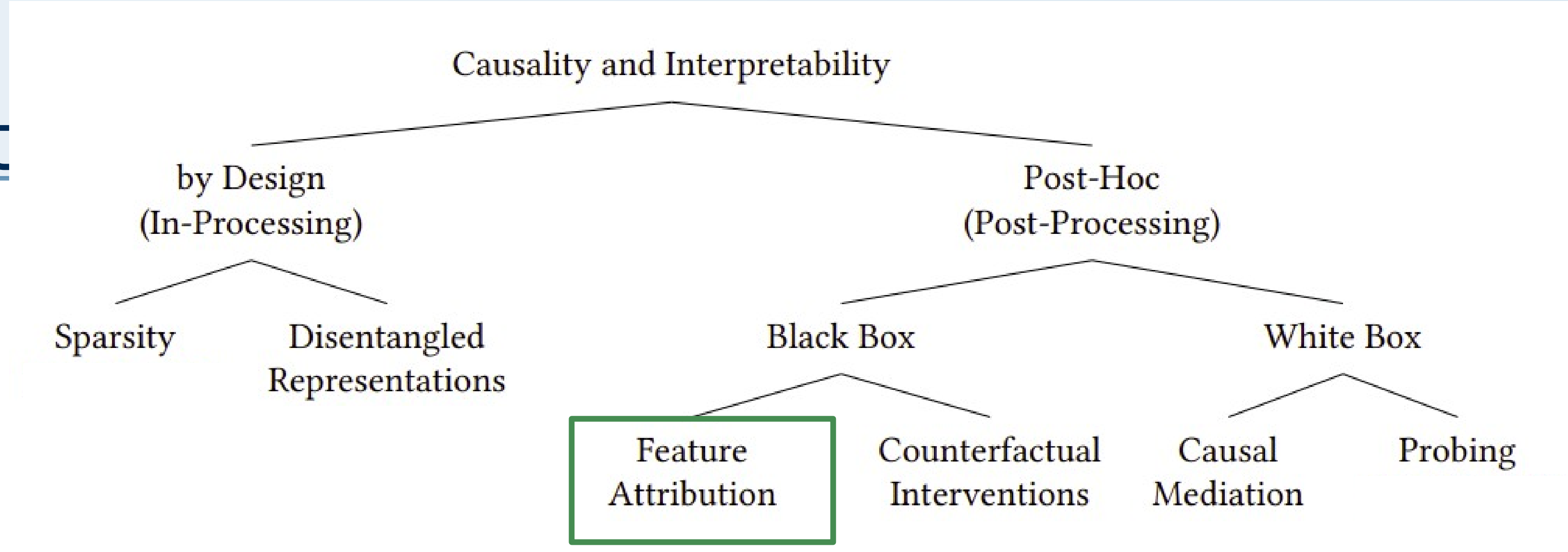
Solution: Disentangle model representations using model architecture & cause-specific training data



(a) Causal Graph

(b) Causal Embedding

Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention. [Kim2017]

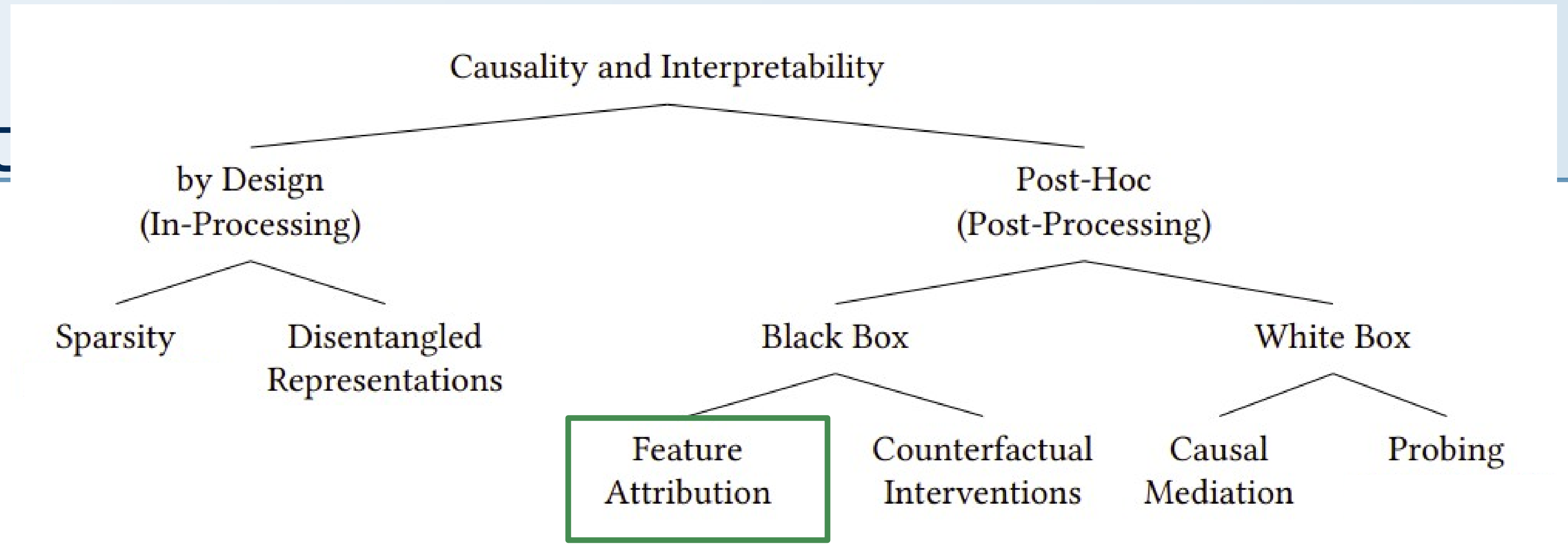


Goal: detect regions that causally influence the prediction

Solution: causal filtering by masking potential influential factors to distinguish true from spurious influences

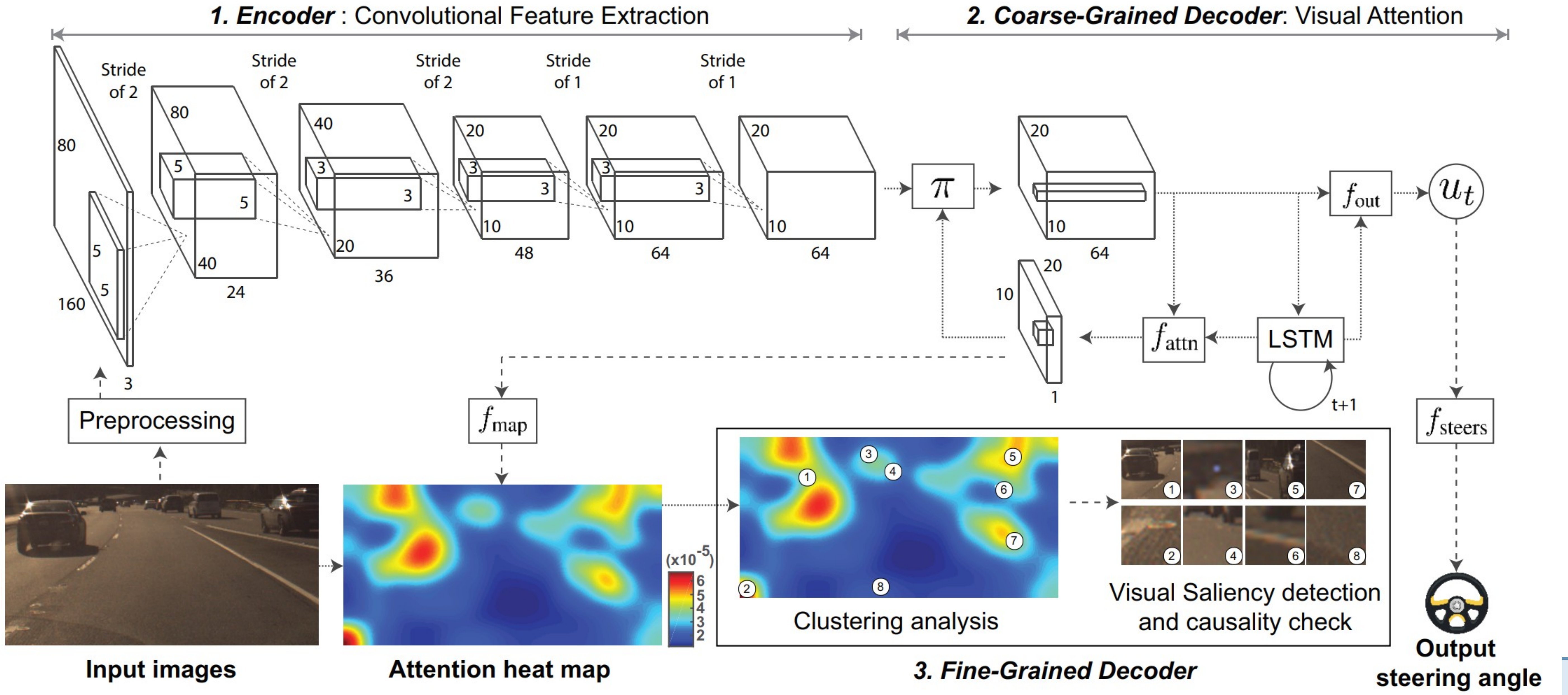
Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention.
[Kim2017]

Trust

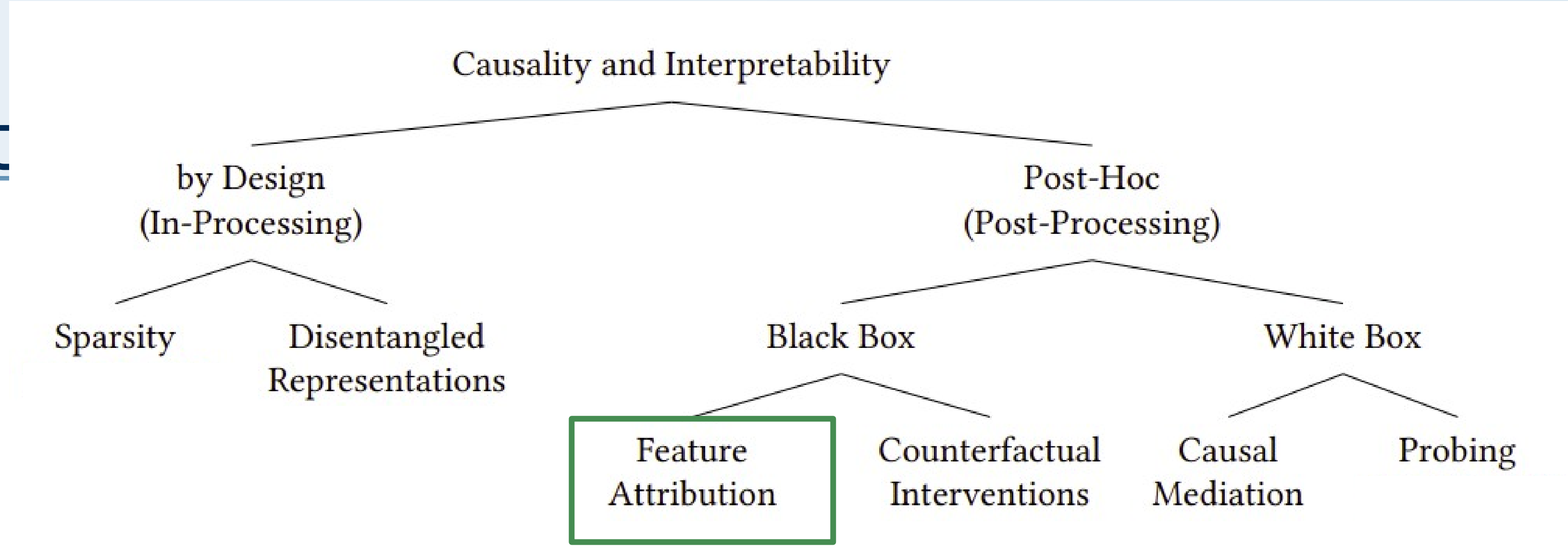


Goal: detect regions that causally influence the prediction

Solution: causal filtering by masking potential influential factors to distinguish true from spurious influences

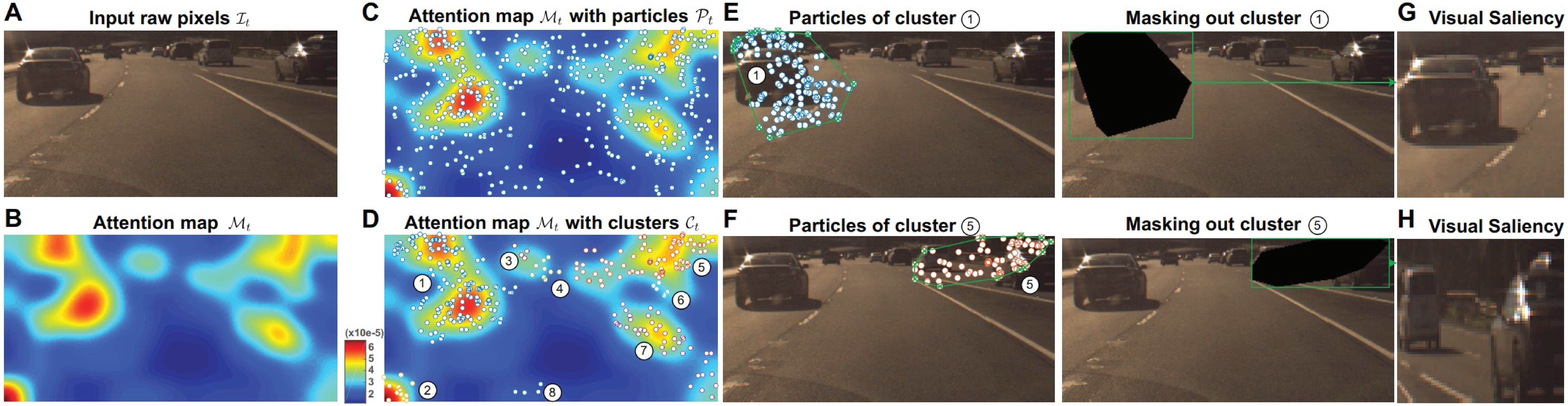


Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention. [Kim2017]

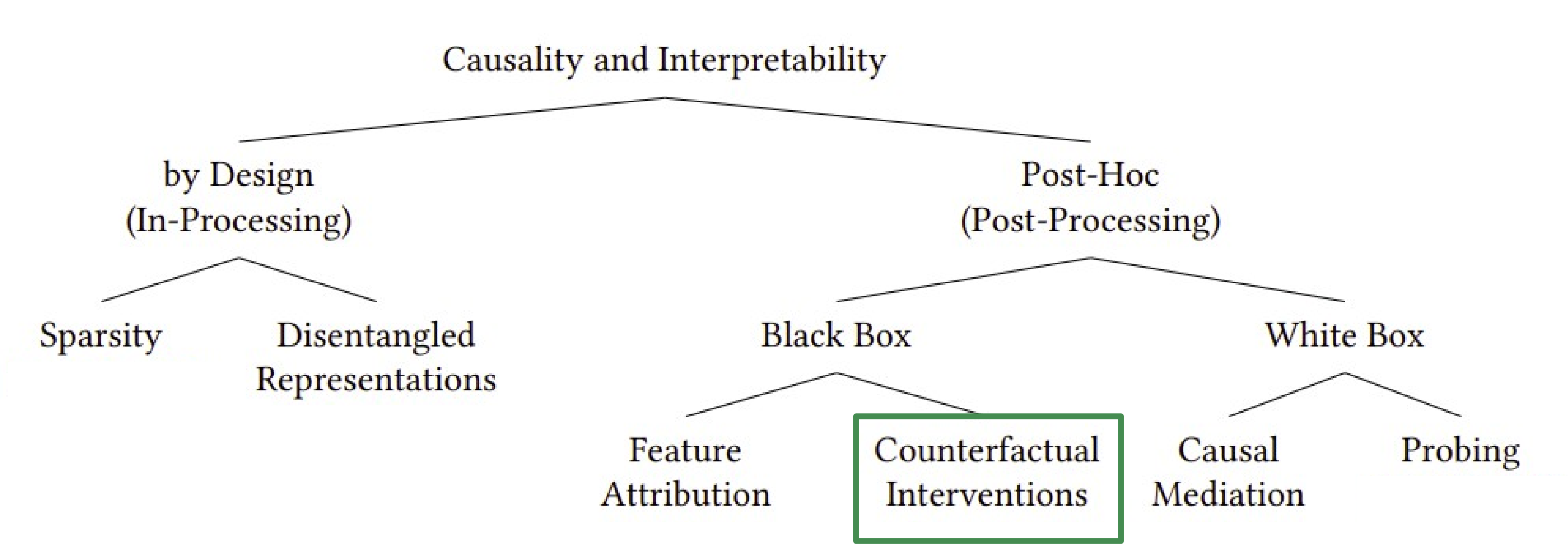


Goal: detect regions that causally influence the prediction

Solution: causal filtering by masking potential influential factors to distinguish true from spurious influences

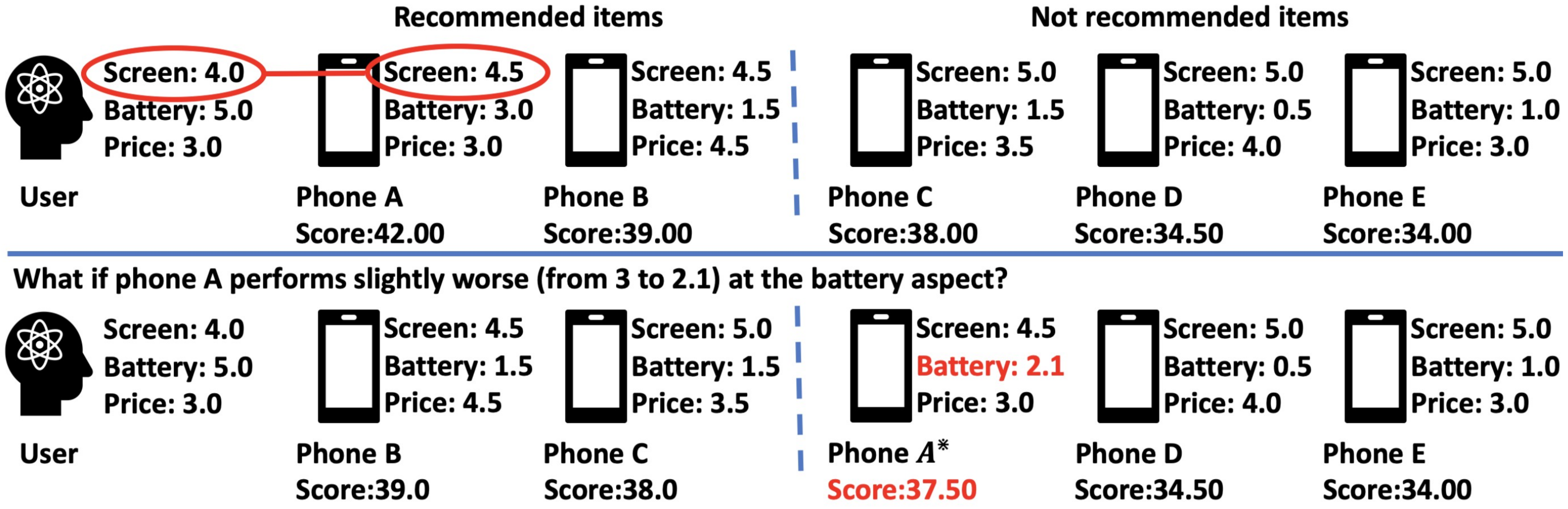


Counterfactual Explainable Recommendation.
[Tan2021]



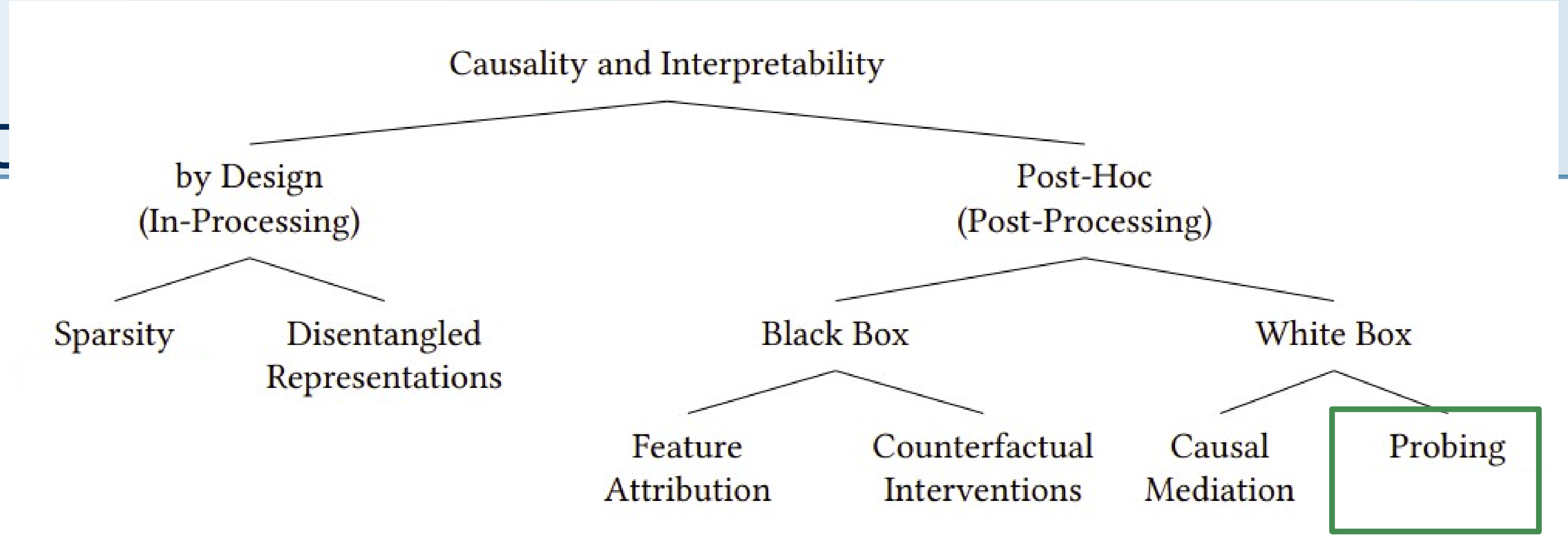
Goal: Detect attributes that could reverse an observed recommendation

Solution: Optimize for minimal changes that reverse the recommendation



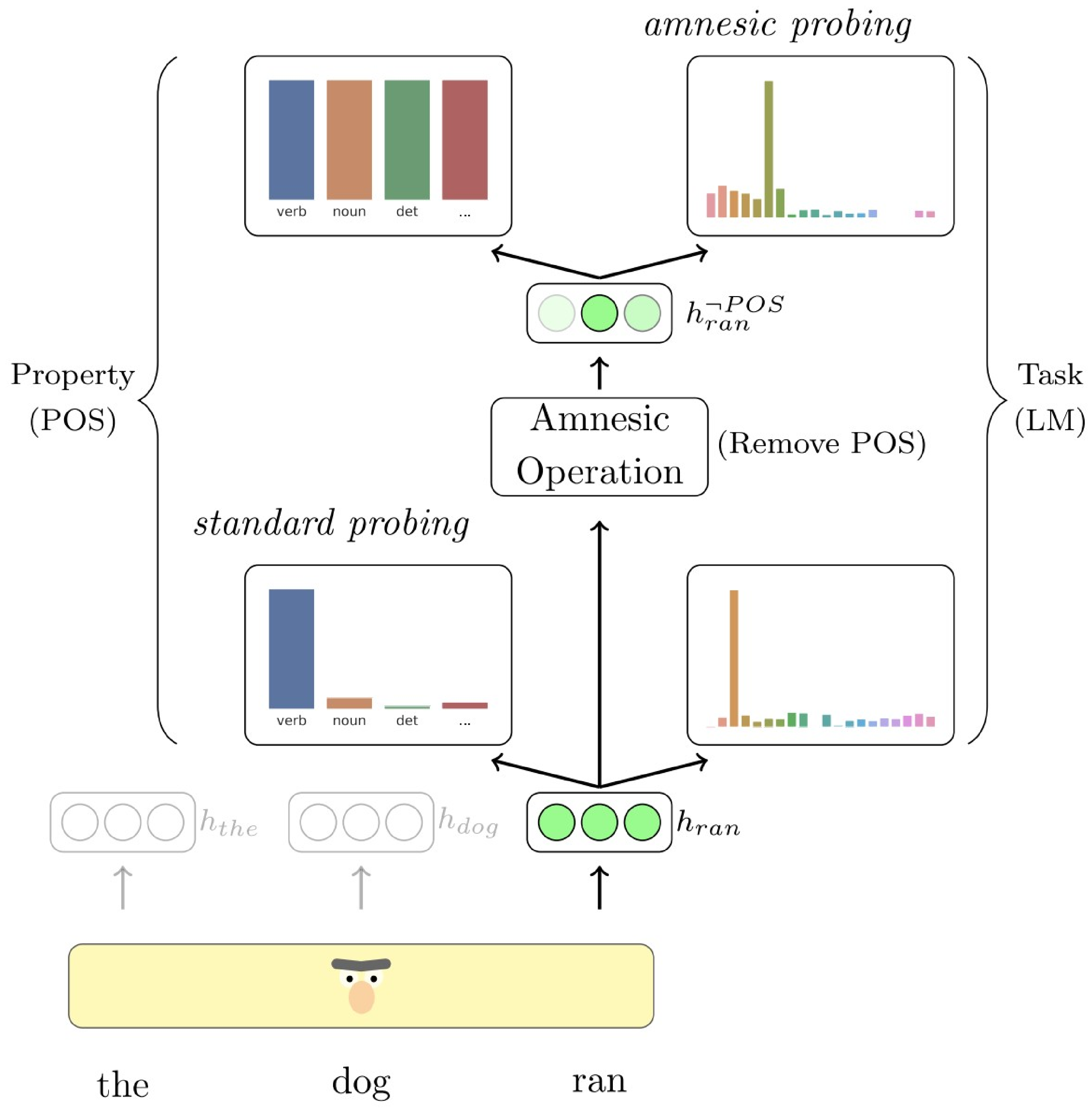

```
If the item had been slightly worse on [aspect(s)],
then it will not be recommended.
```

Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. [Elazar2020]



Goal: Investigate the effect of certain concepts (e.g., gender information/POS) on downstream tasks.

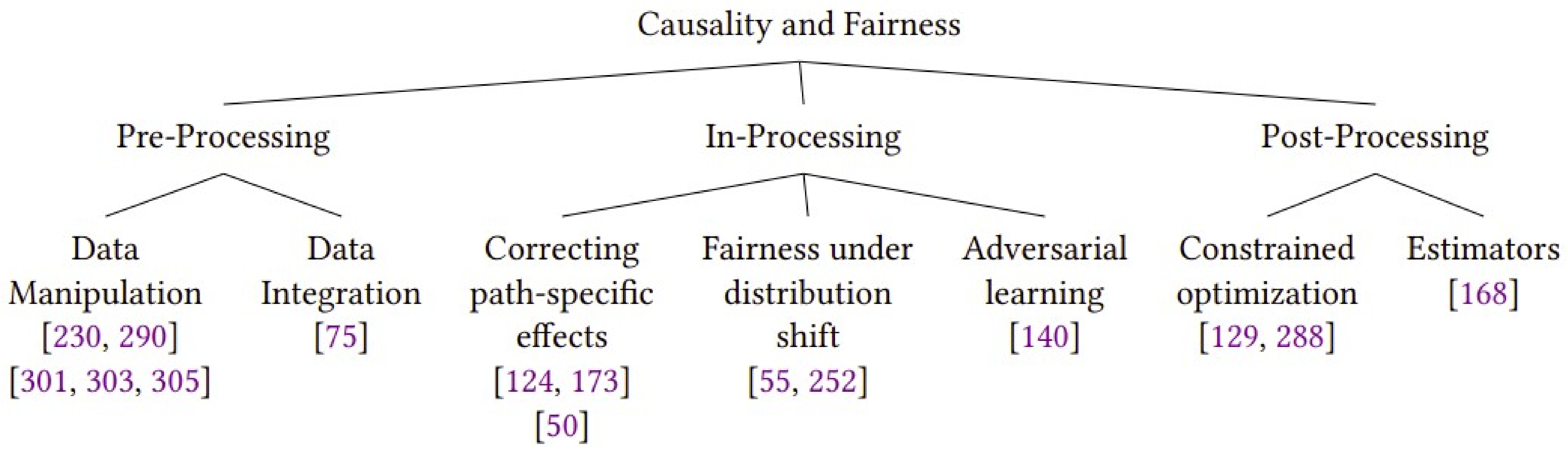
Solution: Remove information from embeddings and measure downstream task performance.



Fairness

Causality is used to quantify and describe fairness

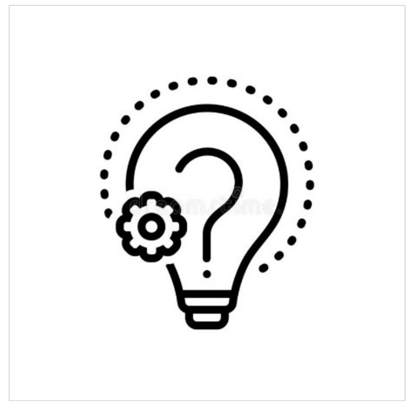
+ This requires the definition of new fairness measures for both *individual* and *group* fairness
+ Effects of sensitive attributes can be causally quantified



Discrimination is Causal in nature...

Discrimination can be causal in nature, meaning that it is often the result of systemic biases that are deeply ingrained in social, economic, and political structures.

Automatic Decision making system



Reason of Discrimination?

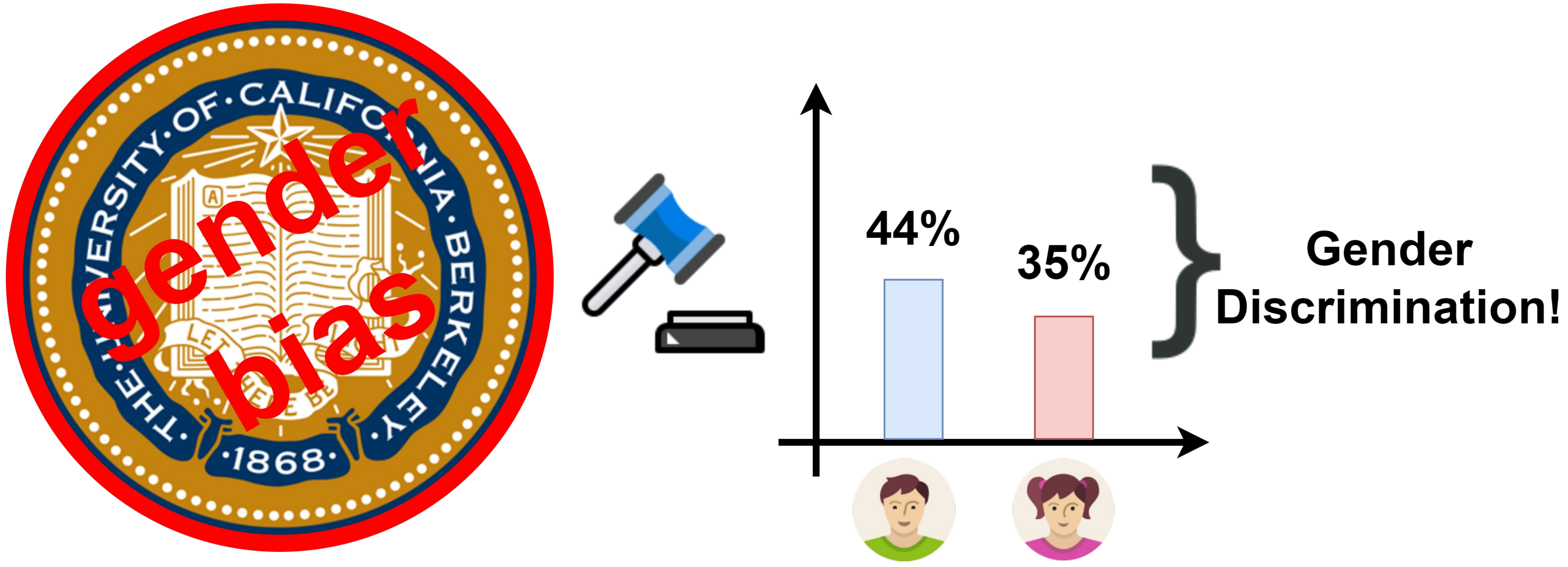
Historical data

By identifying the causal factors that contribute to discrimination, we can develop interventions and policies that address the root causes of the problem, rather than simply treating the symptoms.

UC-Berkley—Simpson’s paradox

The University of California, Berkeley in the 1970s feared a suspected gender bias in the outcomes of its graduate school admissions.

UC-Berkley—Simpson’s paradox



UC-Berkley—Simpson’s paradox

UC-Berkley—Simpson’s paradox

- Some departments had a higher proportion of male applicants, which made it more difficult for women to be admitted overall.
- Women tended to apply to departments that admitted a smaller percentage of applicants overall.
- Once the data was properly analyzed, it was found that there was no evidence of discrimination against women in the admission process at UC Berkeley.

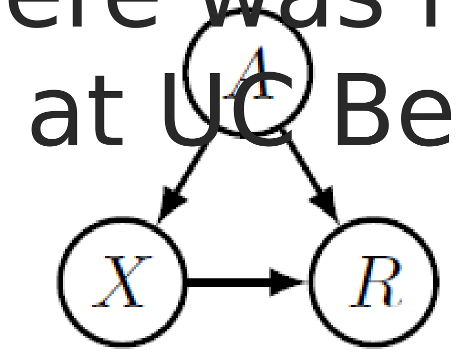
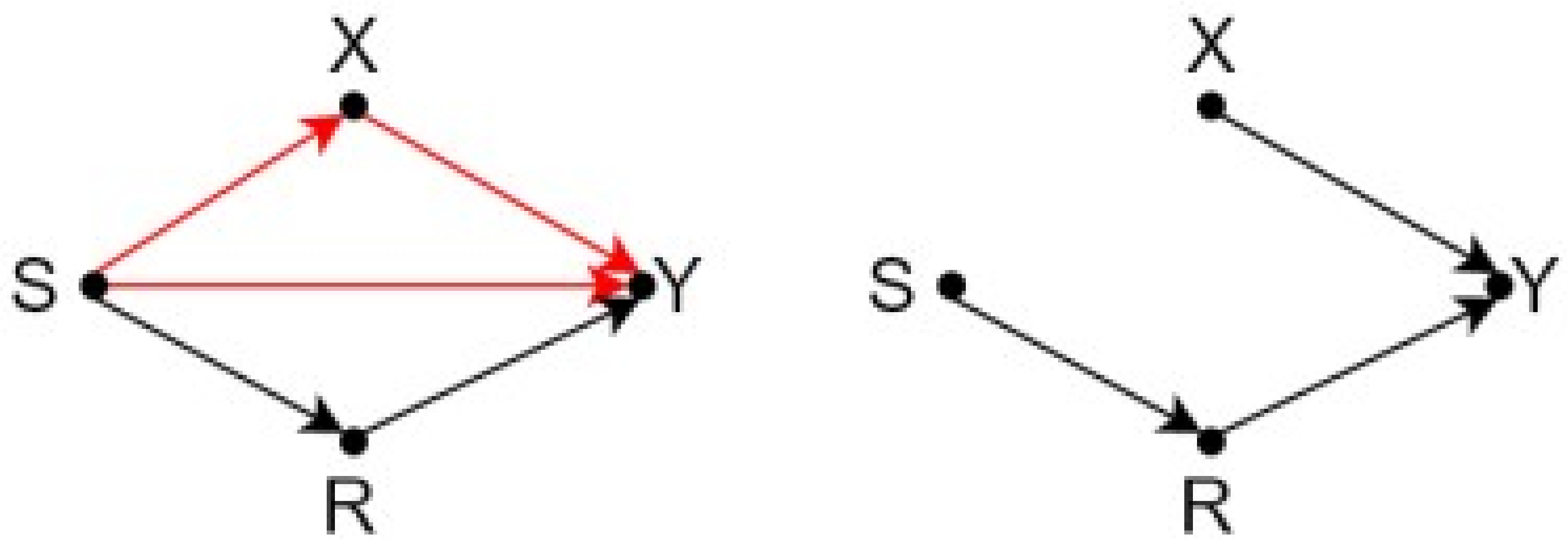


Figure 1: The admission decision R does not only directly depend on gender A, but also on department choice X, which in turn is also affected by gender A.

IF: No unresolved discrimination [Kilbertus2017]

It is a group fairness notion that focuses on the direct and indirect causal influence of sensitive attributes on the decision. It is satisfied when there is no direct path between the sensitive attributes and the outcome, except through a resolving/ admissible variable



[Kilbertus2017]: Kilbertus, N., Rojas Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. Advances in neural information processing systems, 30.

Robustness and Privacy

Robustness: Decreasing sensitivity towards input changes

(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98

(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97

(C) No Person: 0.97, **Mammal: 0.96**, Water: 0.94, Beach: 0.94, Two: 0.94

Privacy: Defending against privacy-evasive attacks

Causal solutions for both areas overlap significantly

- Robustness: Methods for centralized learning setting
- Privacy: Similar methods for decentralized/federated learning setting

Beery, Sara, Grant Van Horn, and Pietro Perona. "Recognition in terra incognita." *Proceedings of the European conference on computer vision (ECCV)*. 2018.

Statistical Machine Learning

We assume that our data is independent and identically distributed (IID)

Allows one to infer the performance of models solely through training data

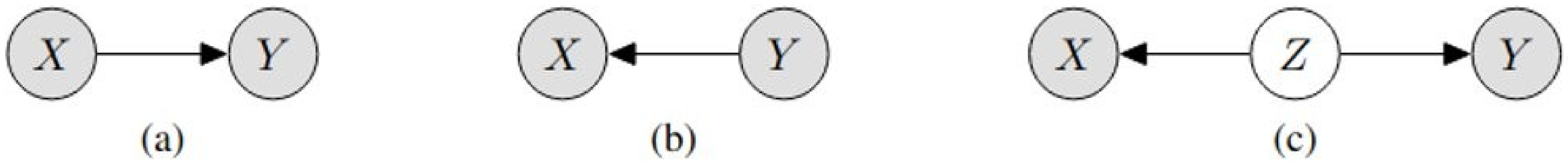
- Empirical Risk Minimization

Very unlikely that training data covers all statistical properties of real-world inference data

Susceptible to distributional shifts caused by unseen data

Enhancing AI with Causality

- No definite solution for distributional shifts
- Statistical ML models are not inclined to properly understand causal relationship
  - Simply fall back on observable correlation that works best for the training data
- Causal encodings allow us to constraint this behavior
- Achievable with pre-, in- and post-processing methods



[Schölkopf22]: Schölkopf, Bernhard, and von Kügelgen, Julius. "From Statistical to Causal Learning." *arXiv*, 2022

Robustness

The performance of models can greatly vary when facing distributional shifts

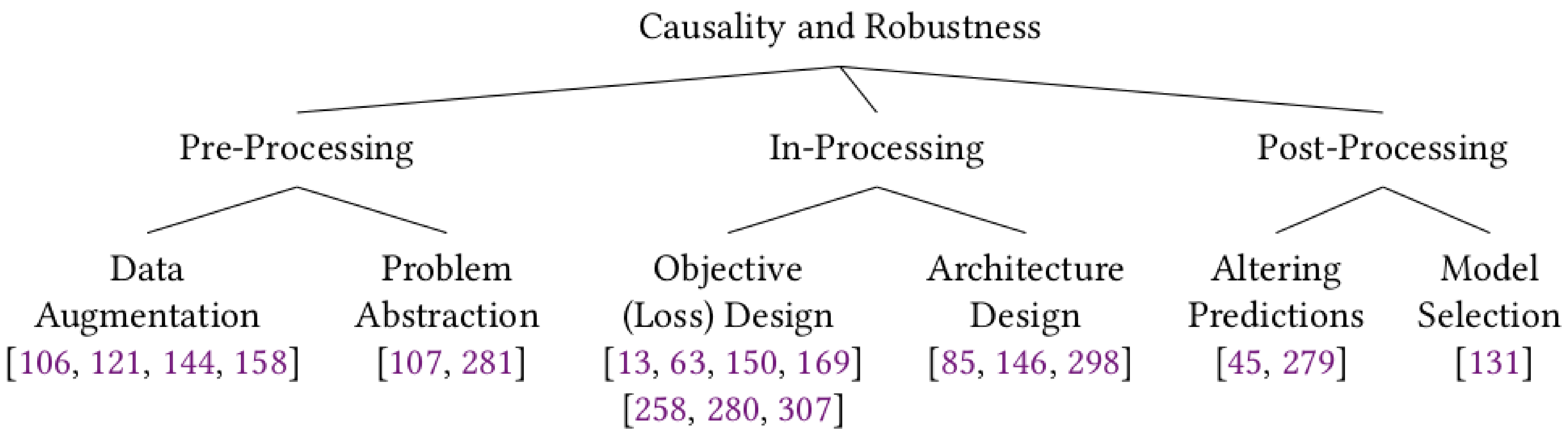
Within this survey, we differentiate between two types of shifts

- Naturally occurring shifts caused by out-of-distributional (OOD) data
- Artificially crafted shifts caused by adversarial examples (AEs)

\+ Primarily focused on increasing robustness in OOD-setting

\+ Causal solutions for a diverse set of problems and data domains

- Computer vision, recommendation, NLP, reinforcement learning and self-supervised learning



Invariant Risk Minimization
In-Processing Method for Robustness [Arjovsky2019]

- Feature invariance relates to its causal importance
  - E.g., image background can greatly vary across data points
  - Therefore, it is not important for predicting the label
- Allows one to develop causal models without causal encodings
- Idea: Promote consistent behavior across different environments
- Successful at increasing robustness of image classifiers in the OOD setting

(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98

(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97

(C) No Person: 0.97, **Mammal: 0.96**, Water: 0.94, Beach: 0.94, Two: 0.94

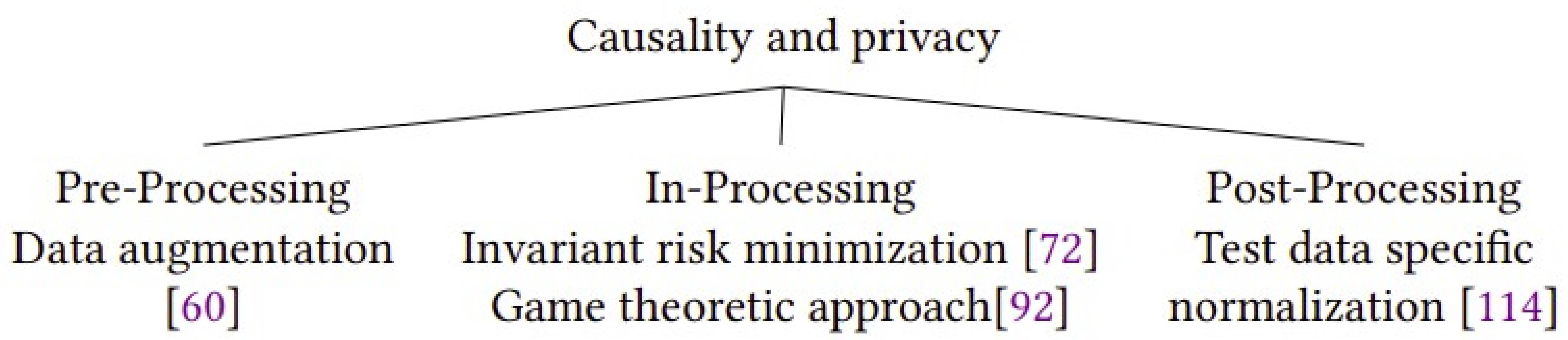
[Beery2018] Beery, Sara, Grant Van Horn, and Pietro Perona. "Recognition in terra incognita." *Proceedings of the European conference on computer vision (ECCV)*. 2018.

Privacy

Main learning paradigm of this section: Federated Learning (FL)

- o Important for domains like healthcare or autonomous driving

+ Core idea: Increase generalization ability of FL models
  - o weak generalizability problematic for *membership inference attacks*

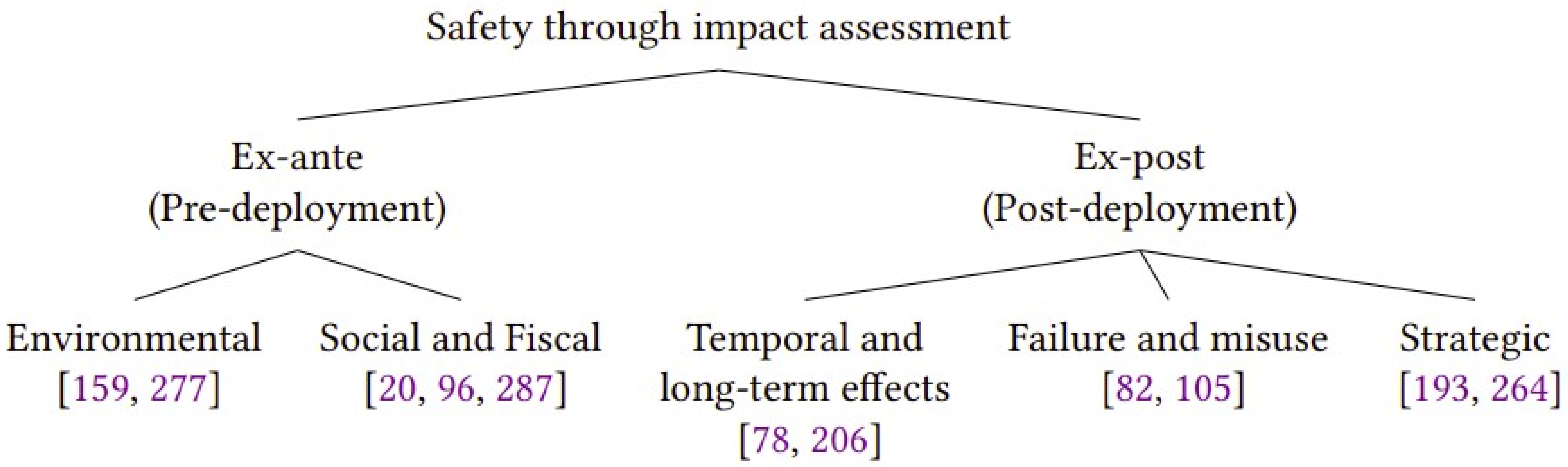


Safety & Accountability (Auditing)

Causality used purely to *assess* the impact of AI systems

- Publications from previous sections enhanced AI systems

+ Safety: estimating negative effects of deploying AI systems
+ Accountability: identifying causes of negative effects
+ Idea of impact can vary depending context, scale and domain



How can AI and humans interact successfully?
„AI must be transparent, explainable, robust and human-centered.“

from *What's so Funny about Science?* by Sidney Harris (1977)

Some L3S projects focusing on this theme:

- European GK "NoBIAS“
- Nds. GK "Responsible AI“
- Project "BIAS" of the VW Foundation
- ZDIN Future Lab Society and Work
- European Big Data Infrastructure SoBigData I + II + III
- CRC Constructing Explainability
- ERC Human-Centered AutoML
- International Leibniz Future Laboratory for Artificial Intelligence
- CAIMed: AI and Causal Methods for Medicine

Causality and Trustworthy AI

Along the dimensions of

- Interpretability: Providing meaningful explanations to users
- Fairness: Developing debiased and non-discriminating AI systems
- Robustness: Decreasing sensitivity towards input changes
- Privacy: Defending against privacy-evasive attacks
- Safety and Accountability: Auditing AI systems

For all references and details see:

Niloy Ganguly, Dren Fazlija, Maryam Badar, Marco Fisichella, Sandipan Sikdar, Johanna Schrader, Jonas Wallat, Koustav Rudra, Manolis Koubarakis, Gourab K. Patro, Wadhah Zai El Amri, Wolfgang Nejdl: **A Review of the Role of Causality in Developing Trustworthy AI Systems.** Feb 2023, https://arxiv.org/abs/2302.06975