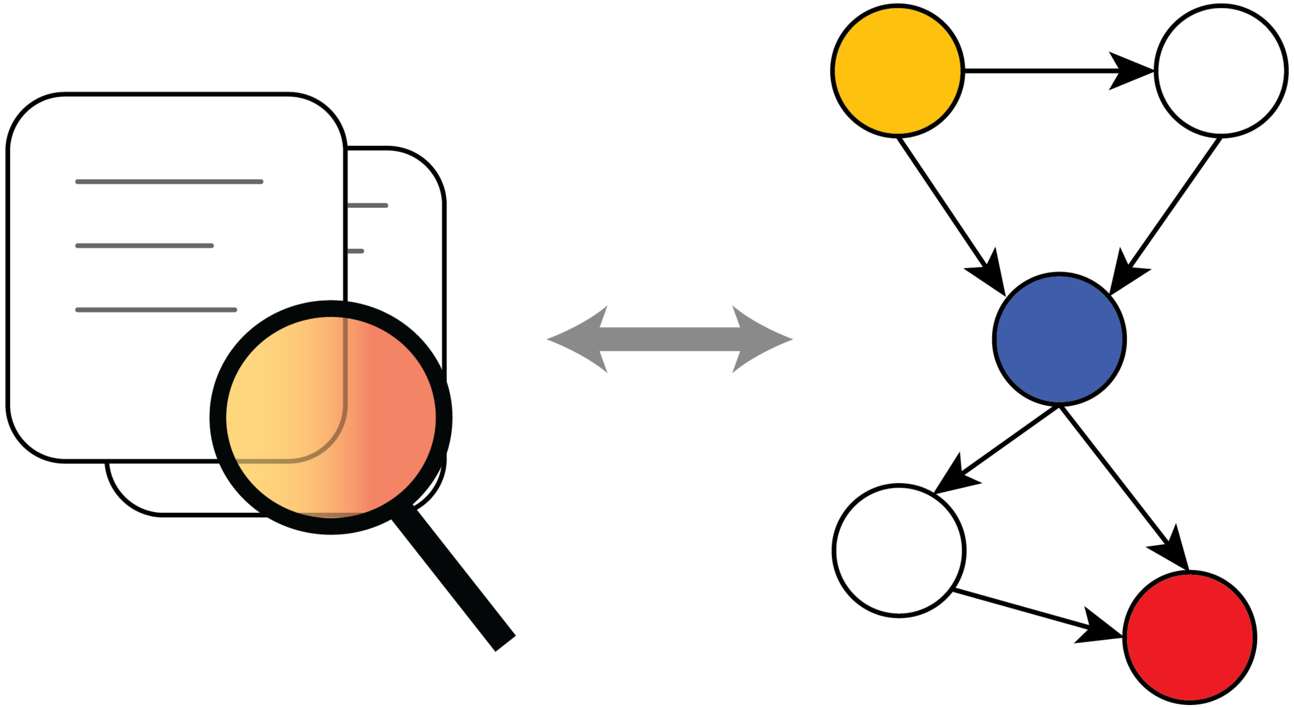
SIG Causal Data Science



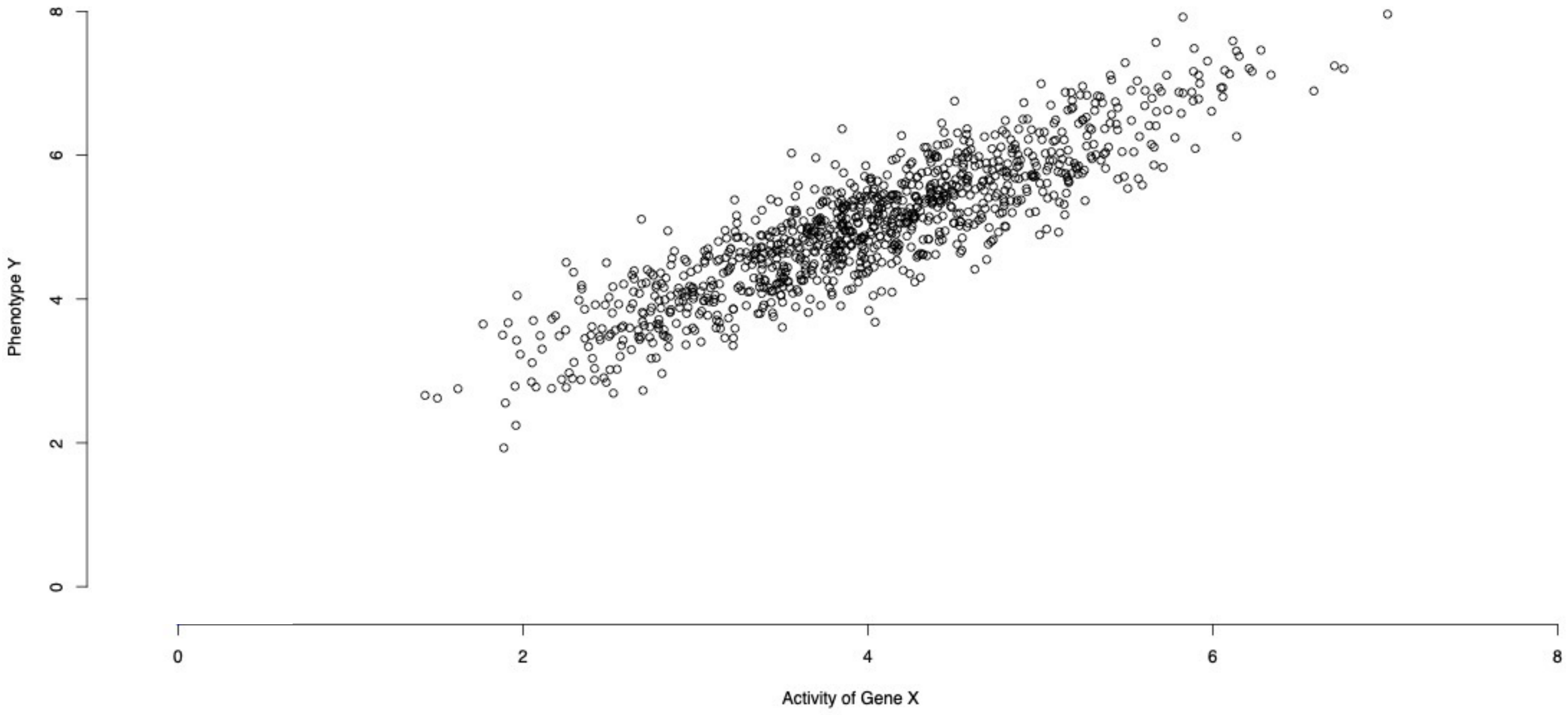
Causal Inference <-> Data Science

Statistical Modeling / Machine Learning / Data Science provide us with a variety of incredibly useful tools for performing ***certain types*** of tasks with data

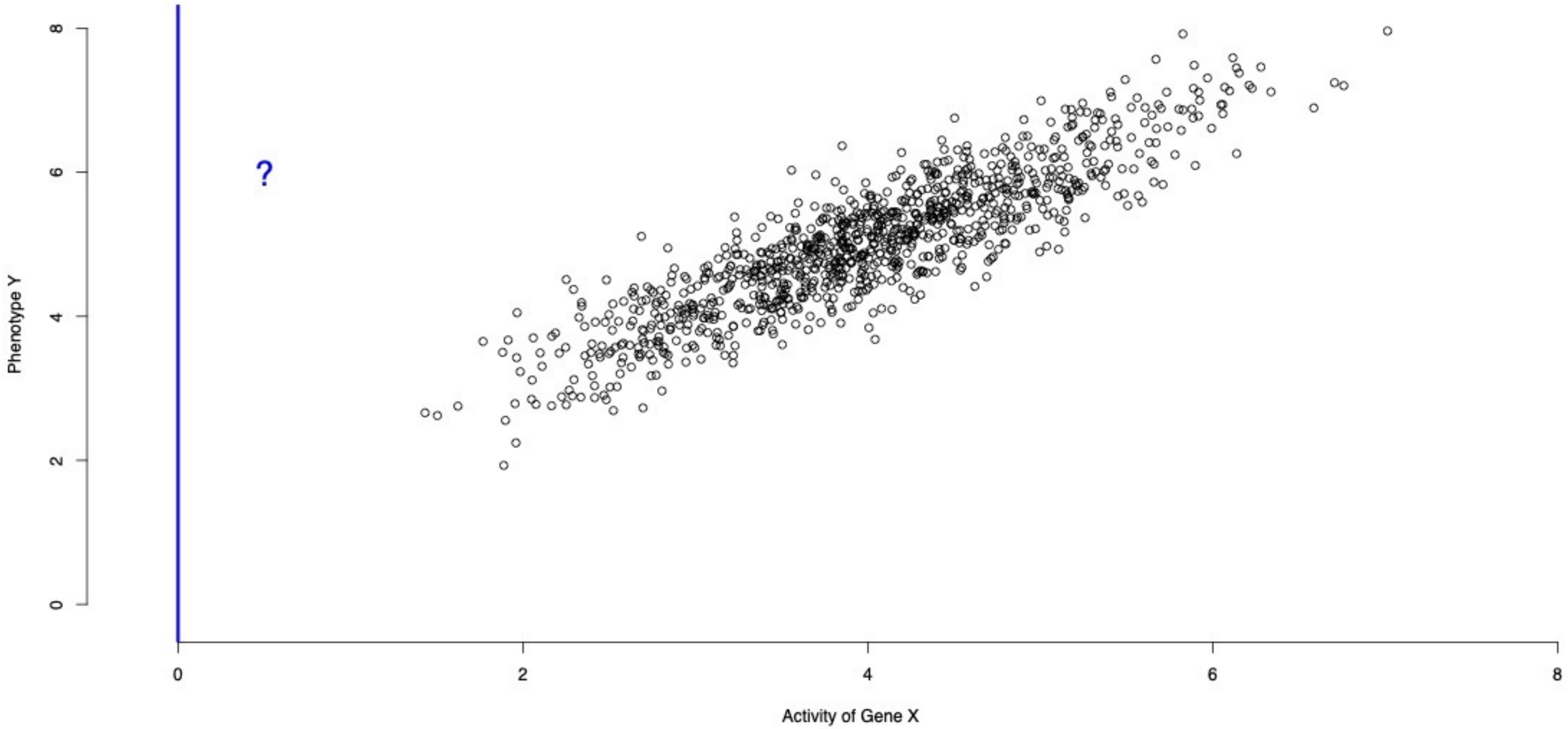
- Description / Feature Extraction / Classification
- Models of co-occurrence patterns
- Prediction
  - ….of a new data point drawn from the same population distribution under the same circumstances as the training set

However, these tools are often not sufficient to tell us how we might **interact** with or **make decisions** which impact the real world

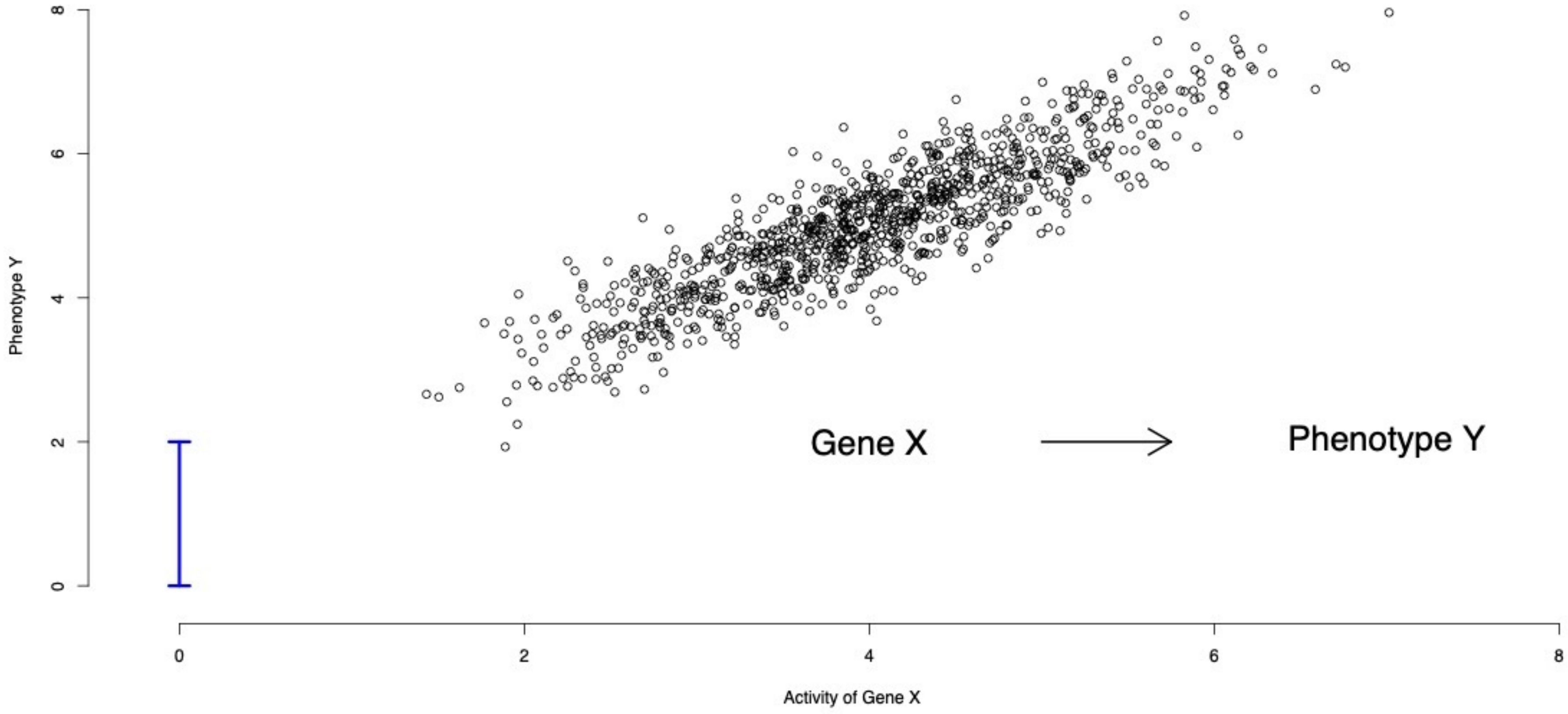
Causal Inference <-> Data Science



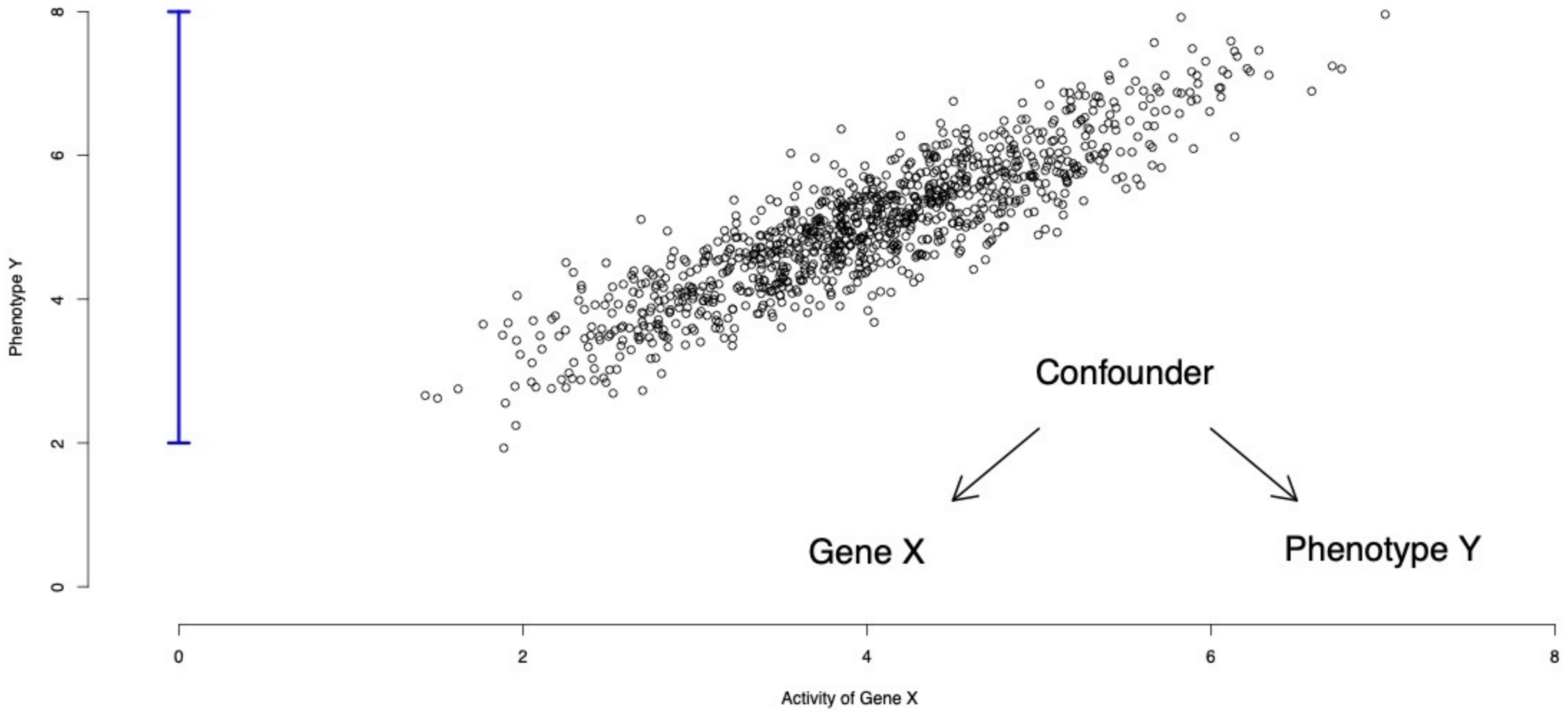
Causal Inference <-> Data Science



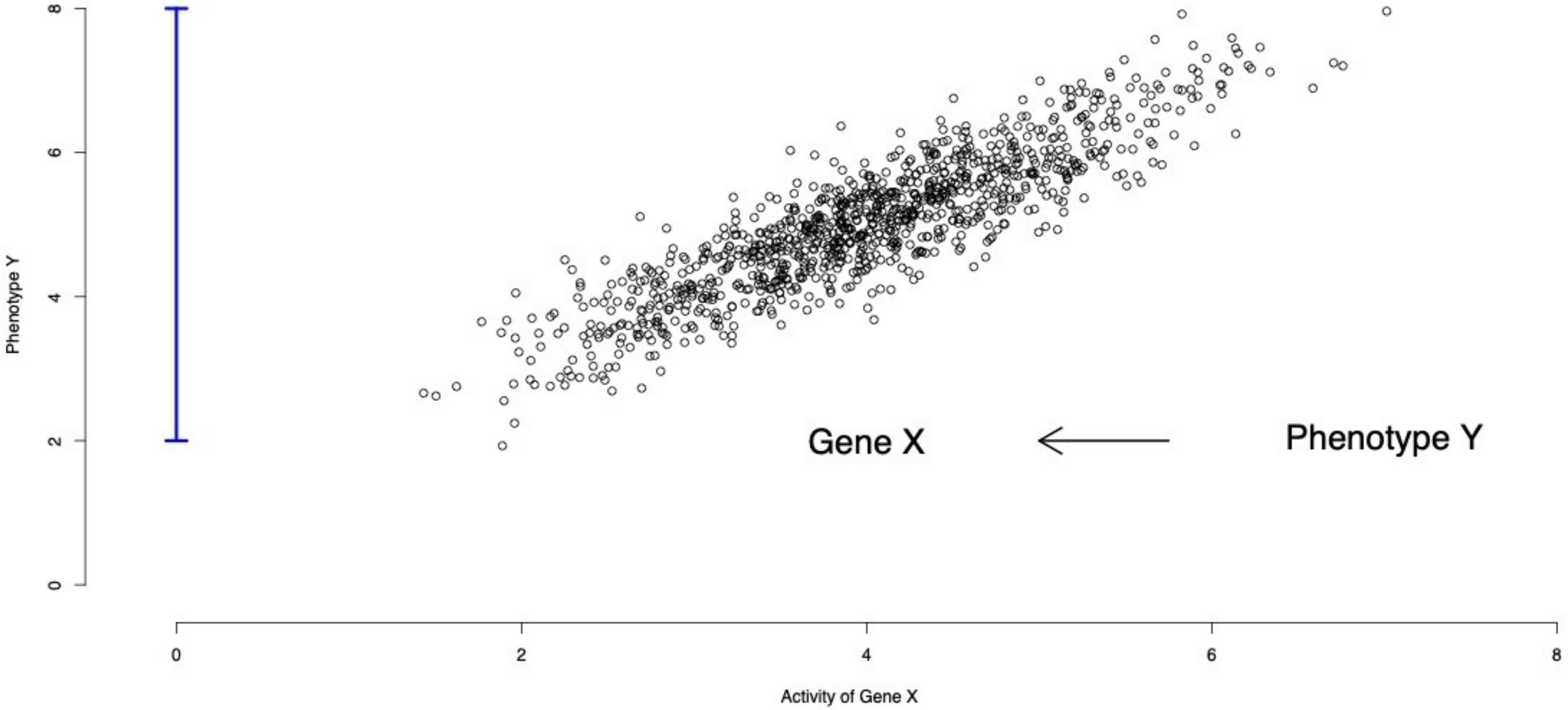
Causal Inference <-> Data Science



Causal Inference <-> Data Science



Causal Inference <-> Data Science



Causal Tasks

Causal Inference

- Estimating the effects of (real or hypothetical) interventions from data
- Using causal knowledge / causal models to do inference in combination with statistical modelling techniques
- Counterfactual Prediction vs “Factual” Prediction

Causal Discovery

- Learning the causal model itself from data
- Structure Learning, Structure Recovery, Causal Learning

Aim of the SIG

Bring together researchers who develop and apply methods or approaches that aim to answer **causal research questions**

- Does exposure to a particular factor cause disease onset?
- Was the introduction of a government policy successful in achieving a particular aim or not?
- How should we intervene in a system to achieve some outcome, and what effect can we expect that intervention to have?
- How can we design learning algorithms that yield insights into causal effects and interventions?
- How can we best leverage large-scale data sources for causal insight?
- Can prediction models be used to make decisions about optimal treatments?

We aim to be a **broad church**; different perspectives, different backgrounds, different kinds of problems

Members

Biomedical Sciences

Julius Center

Princes Maxima Centrum

Pharmaceutical Sciences

Epidemiology and Health Economics

Social and Behavioural Sciences

Methods & Statistics

Sociology

Law, Economics & Governance

REBO

Utrecht School of Economics (USE)

Science Faculty

Information and Computing Sciences

Causal Data Science Co-ordination team

Oisín Ryan (o.ryan@umcu.nl)

- Real World Evidence Team, Data Science and Biostatistics, Julius Center, UMCU
- Causal inference and discovery with large-scale administrative data sources

Thijs van Ommen (m.vanommen1@uu.nl)

- Assistant Professor, Information and Computing Sciences, UU
- Statistical and algorithmic aspects of causal discovery

Wouter van Amsterdam (W.A.C.vanAmsterdam-3@umcutrecht.nl)

- Data Science and Biostatistics, Julius Center, UMCU
- Prediction models and causal reasoning for clinical decision making

Upcoming Events

Meeting: Causal Inference and Machine Learning

Thursday May 16, 15.00 – 17.00. Location TBA (Uithof)

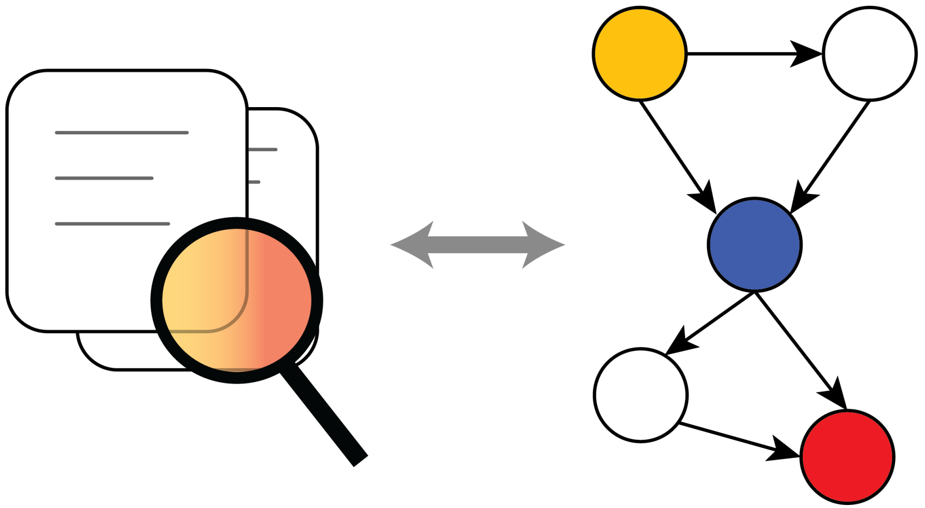
1. Causality and prediction: developing and validating models for decision making (Wouter van Amsterdam)
2. Causal discovery in the presence of unobserved confounding (Thijs van Ommen)

With plenty of time for discussion, questions, and a small borrell

Introduction to Causal Inference & Causal Data Science

Summer School **August 5th – 9th**

- **Potential Outcomes** and **Directed Acyclic Graphs**
- **Emulate a target trial** with observational data
- Causal approaches to **prediction modelling** and **structure learning**
- **Adjusting for confounders**, simple to advanced methods
- **Advanced topics:** Complex Longitudinal Settings, and causal policy evaluation
- **Hands-on sessions** every day with exercises in R



https://utrechtsummerschool.nl/courses/healthcare/introduction-to-causal-inference-and-causal-data-science

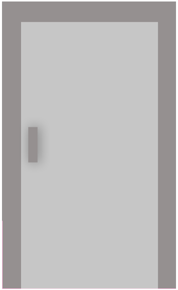
Robust Causal Domain Adaptation in a Simple Diagnostic Setting

Thijs van Ommen

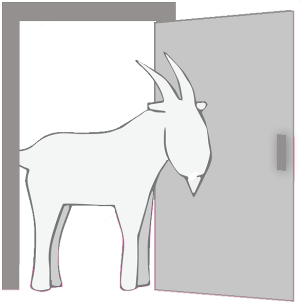
Utrecht University

ADS SIGs event, April 18, 2024

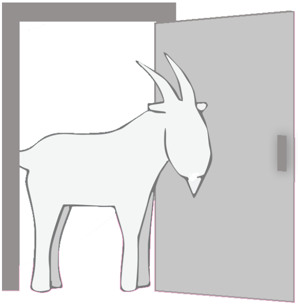
Background

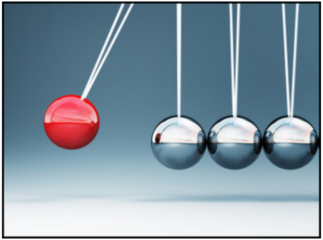


Background

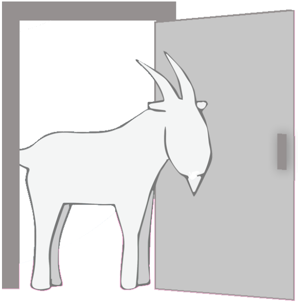


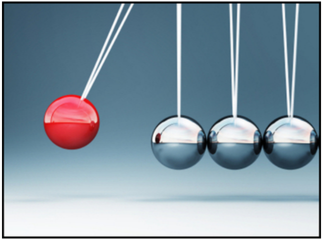
Background





Background





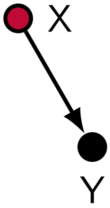
This work

Motivating example

- X: lung cancer — to be diagnosed

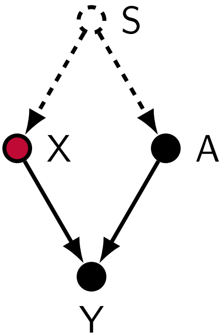
Motivating example

- X: lung cancer — to be diagnosed
- Y: chest pain



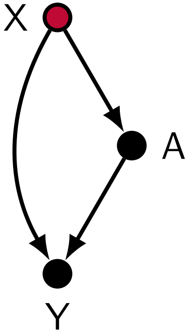
Motivating example

- X: lung cancer — to be diagnosed
- Y: chest pain
- S: smoking (unobserved variable)
- A: aspirin — may be prescribed to smokers due to their risk of heart disease



Motivating example

- X: lung cancer — to be diagnosed
- Y: chest pain
- S: smoking (unobserved variable)
- A: aspirin — may be prescribed to smokers due to their risk of heart disease



Motivating example

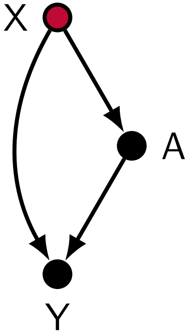
Two domains, e.g. hospitals:

- source domain ($C = 0$) where we observe data
- target domain ($C = 1$) where we want to make decisions

Same causal graph, different distributions:

source:

$$P(X \mid C = 0)$$
$$P(A \mid X, C = 0)$$
$$P(Y \mid X, A, C = 0)$$



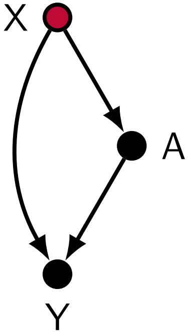
Motivating example

Two domains, e.g. hospitals:

- source domain ($C = 0$) where we observe data
- target domain ($C = 1$) where we want to make decisions

Same causal graph, different distributions:

$$
\begin{array}{rcl}
& \text{source:} & \text{target:} \\
P(X \mid C = 0) & = & P(X \mid C = 1) \\
P(A \mid X, C = 0) & & P(A \mid X, C = 1)\textbf{?} \\
P(Y \mid X, A, C = 0) & = & P(Y \mid X, A, C = 1)
\end{array}
$$



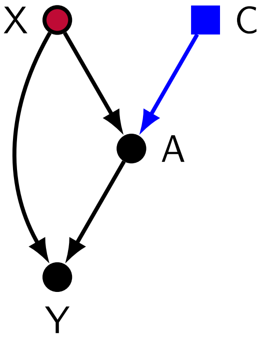
Motivating example

Two domains, e.g. hospitals:

- source domain ($C = 0$) where we observe data
- target domain ($C = 1$) where we want to make decisions

Same causal graph, different distributions:

$$\begin{array}{rcl} & \text{source:} & \text{target:} \\ P(X \mid C = 0) & = & P(X \mid C = 1) \\ P(A \mid X, C = 0) & & P(A \mid X, C = 1)\textbf{?} \\ P(Y \mid X, A, C = 0) & = & P(Y \mid X, A, C = 1) \end{array}$$



Robust approach

Let $\mathcal{P}$ be the set of all joint distributions for the target domain consistent with what we know from the source domain

- We want to take decisions that are good regardless of what $P \in \mathcal{P}$ is realized

Robust approach

Let $\mathcal{P}$ be the set of all joint distributions for the target domain consistent with what we know from the source domain

- We want to take decisions that are good regardless of what $P \in \mathcal{P}$ is realized
- Model as zero-sum game against adversary who chooses $P \in \mathcal{P}$

Robust approach

Let $\mathcal{P}$ be the set of all joint distributions for the target domain consistent with what we know from the source domain

- We want to take decisions that are good regardless of what $P \in \mathcal{P}$ is realized
- Model as zero-sum game against adversary who chooses $P \in \mathcal{P}$
- For that, we need to fix a loss function, e.g. Brier or logarithmic loss:

$$L_{\text{Brier}}(x, Q) = \sum_{x' \in \mathcal{X}} \left(\mathbf{1}_{x'=x} - Q(x')\right)^2;$$
$$L_{\log}(x, Q) = -\log Q(x).$$

(Both are strictly proper scoring rules: they are uniquely minimized when Q equals the true distribution of X)

Solution

We showed [TvO, ISIPTA 2019]

- that optimal strategies exist for both players, for these and many more loss functions;
- how to find them, analytically or numerically.

Solution

We showed [TvO, ISIPTA 2019]

- that optimal strategies exist for both players, for these and many more loss functions;
- how to find them, analytically or numerically.

In a numerical example with all variables binary, we found the optimal strategies:

- for Brier loss, and
- for logarithmic loss

The two solutions (and thus the resulting decisions) are different, even though both loss functions are strictly proper scoring rules

Conclusion

Conclusions:

- Causality helps us think about data science problems, even if they're not obviously about causality
- We can't decouple the (probabilistic) prediction from the decision making that follows

The end

Thank you!