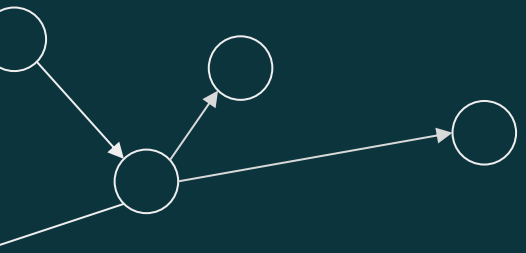
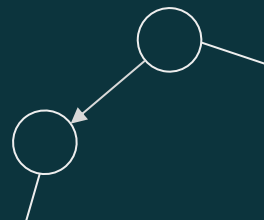


TECHNISCHE
UNIVERSITÄT
DARMSTADT

Causal Explanations of Structural Causal Models

Matej Zečević
@ CIIG

12th December 2022



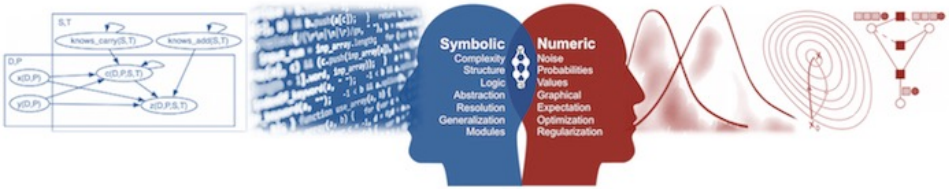
Short “Who am I?”

Prof. Dr. Kristian Kersting, FEurAI, FELLIS

Computer Science Department and Centre for Cognitive Science, TU Darmstadt
Altes Hauptgebäude, Room 074, Hochschulstrasse 1, 64289 Darmstadt, Germany
+49-6151-16-24411 kersting (at) cs (dot) tu-darmstadt (dot) de

Devendra Singh Dhami

Machine Learning Group, Computer Science Department, TU Darmstadt.
Hochschulstrasse 1, Room S1|66, 64289 Darmstadt, Germany
+49 1523 79 19263 devendra (dot) dhami (at) cs (dot) tu-darmstadt (dot) de



TECHNISCHE UNIVERSITÄT DARMSTADT

CAUSE

Matej Zečević

Machine Learning Group, Computer Science Department, TU Darmstadt.
Hochschulstrasse 1, Room S1|03 066, 64289 Darmstadt, Germany.
matej (dot) zecevic (at) cs (dot) tu-darmstadt (dot) de

Fachbereich Informatik

Mission. Contributing towards System 2 AI by unifying Causality with Machine Learning.
"As X-rays are to the surgeon, graphs are for causation.", if you know the name of the quotee by heart, then please contact me.

Personal Website.
matej-zecevic.de

Timeline.

2020 - now: Ph.D. candidate with Prof. Kersting @ CS Department, TU Darmstadt, Germany.
2019 - 2020: M.Sc. Computer Science Thesis under Prof. Schölkopf @ Max Planck Institute for Intelligent Systems, Tübingen, Germany.
2018 - 2019: Research Intern with Dr. Vinogradska @ Bosch Center for Artificial Intelligence, Renningen, Germany.
2017 - 2018: B.Sc. Computer Science Thesis under Prof. Helmstaedter @ Max Planck Institute for Brain Research, Frankfurt, Germany.
2014 - 2020: M.Sc. & B.Sc. Computer Science @ CS Department, TU Darmstadt, Germany.

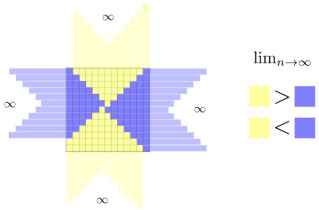
Social Media 🤣

Twitter
@matej_zecevic

Open-access articles on my Blog
(pretty much on anything but mostly on nerdy topics)
matej-zecevic.de

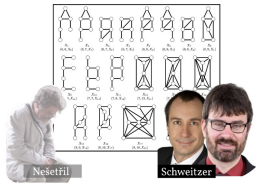
The Infinite is Useful

Talking about the foundation of mathematics usually involves at some initial point the discussion on set theory, which one arguably considers to be a “theory...



Proving Nešetřil's Conjecture: Minimal Asymmetric Graphs

In 1988 at a seminar in Oberwolfach (located in Germany and by many mathematicians considered as a sort of “mecca for mathematicians” [1]) the czech...



Sports Excellence: The Greatest Shohei Ono

Judo (柔道, japanese for “gentle way”) is a martial art centered around throwing techniques for close-quarters combat (opposed to for instance Karate, another japanese martial...

TECHNISCHE UNIVERSITÄT DARMSTADT

For Your Interest

Opportunities for Consuming and
Engaging in Causality

Weekly Meeting Community

Join the community
via
discuss.causality.link

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Slack, Zoom, Mailing List, Announcements

Next Up on 07.December '22: Luigi Gresele, Ph.D. Candidate @ MPI:IS with Causal Inference Through the Structural Causal Marginal Problem @ ICML 2022

Join the Discussion!

We get together with the authors of papers in Causality x AI/ML to discuss their papers in a lively group discussion. We meet weekly on Wednesday at 15:30 UTC / 16:30 CEST/CET in summer/winter / 10:30 EST / 23:30 JST.

Join the **Session** (Zoom)

Join the **Community** (Slack)

Join the **Mailing List** (G-Groups)

TECHNISCHE UNIVERSITÄT DARMSTADT

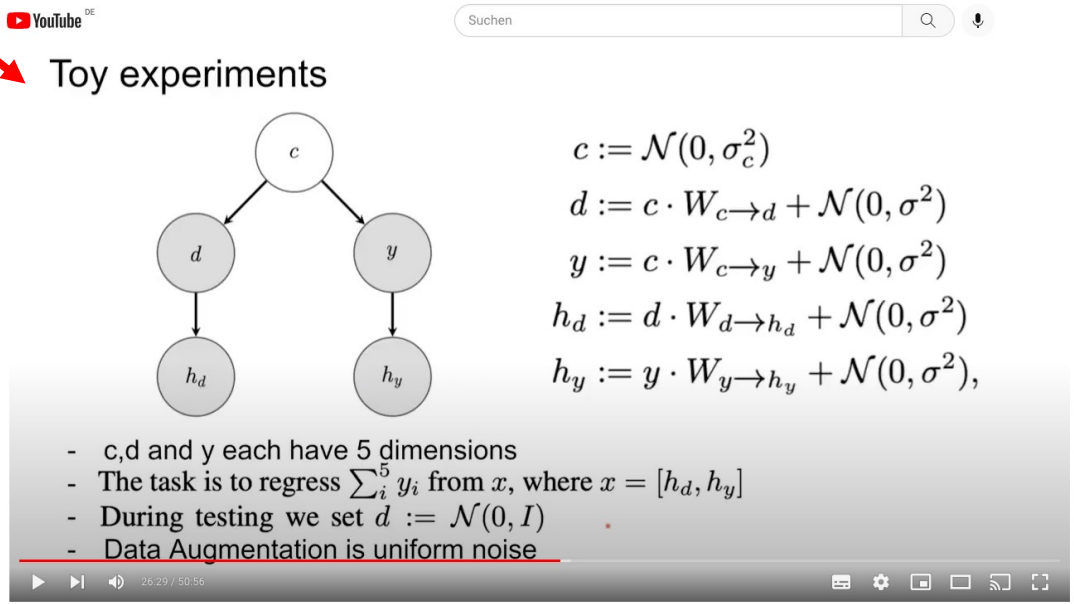
21 Sessions Completed and All Recorded

Past Sessions (see #recordings):

- Session 30.11.2022 | **Selecting Data Augmentation for Simulating Interventions** | Discussant: Maximilian Ilse
- Session 23.11.2022 | **On Disentangled Representations Learned from Correlated Data** | Discussant: Frederik Träuble
- Session 16.11.2022 | **Causal Curiosity: RL Agents Discovering Self-supervised Experiments for Causal Repr. Learning** | Discussant: Sumedh Sontakke
- Session 09.11.2022 | **Causal Machine Learning: A Survey and Open Problems** | Discussants: Jean Kaddour, Aengus Lynch
- Session 02.11.2022 | **A Critical Look at the Consistency of Causal Estimation with Deep Latent Variable Models** | Discussant: Severi Rissanen
- Session 26.10.2022 | **Nonlinear Invariant Risk Minimization: A Causal Approach** | Discussant: Chaochao Lu
- Session 19.10.2022 | **CausalVAE: Disentangled Representation Learning via Neural Structural Causal Models** | Discussant: Mengyue Yang
- Session 12.10.2022 | **Weakly Supervised Causal Representation Learning** | Discussant: Johann Brehmer
- Session 05.10.2022 | **Towards Causal Representation Learning** | Discussant: Anirudh Goyal
- Session 21.09.2022 | **Selection Collider Bias in Large Language Models** | Discussant: Emily McMilin
- Session 14.09.2022 | **The Causal-Neural Connection: Expressiveness, Learnability, and Inference** | Discussants: Kai-Zhan Lee, Kevin Xia
- Session 07.09.2022 | **Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style** | Discussant: Julius von Kügelgen
- Session 31.08.2022 | **Interventions, Where and How? Experimental Design for Causal Models at Scale** | Discussants: Panagiotis Tigas and Yashas Annadani
- Session 24.08.2022 | **Beware of the Simulated DAG! Causal Discovery Benchmarks May Be Easy To Game** | Discussant: Alexander Reisach
- Session 17.08.2022 | **Effect Identification in Cluster Causal Diagrams** | Discussants: Tara Anand and Adèle Ribeiro
- Session 10.08.2022 | **Can Foundation Models Talk Causality?** | Discussant: Moritz Willig
- Session 03.08.2022 | **Causal Conceptions of Fairness and their Consequences** | Discussants: Hamed Nilforoshan and Johann Gaebler
- Session 28.07.2022 | **Bayesian Causal Discovery under Unknown Interventions** | Discussant: Alexander Hägele
- Session 20.07.2022 | **Counterfactual Fairness** | Discussant: Toon Vanderschueren

All 21 Recorded, Re-watch!

Session 30.11.2022 | **Selecting Data Augmentation for Simulating Interventions** | Discussant: Maximilian Ilse



TECHNISCHE UNIVERSITÄT DARMSTADT

neuro (Causal ∧ Symbolic) AI

Workshop at the 36th Conference on Neural Information Processing Systems (NeurIPS)

December 2022

Visit **ncsi.cause-lab.net**

Workshop

Workshop on neuro Causal and Symbolic AI (nCSI)

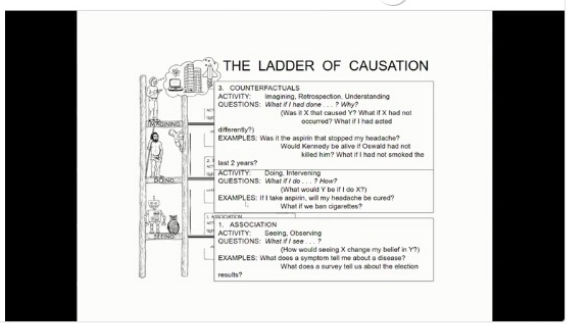
Matej Zečević · Devendra Dhami · Christina Winkler · Thomas Kipf · Robert Peharz · Petar Veličković

Virtual

Abstract Workshop Website

[Contact: ncsi-workshop@googlegroups.com]

Time	Speaker	Title
09:00	Matej Zečević (B/O Organizers)	Welcome & Opening Remarks
09:10	Judea Pearl	Opening Keynote for nCSI
9:40	Emanuele Marconato et al.	GlanceNets: Interpretable, Leak-proof Concept-based Models
9:50	Steven Piantadosi & Felix Hill	Meaning without reference in large language models
10:00		Poster Session 1 (virtual) + Break
11:30	Mausam	Neural Models with Symbolic Representations for Perceptuo-Reasoning Tasks
12:00	Dhanya Sridhar	Causal Inference from Text
12:30		Break
13:30	Yan Zhang et al.	Unlocking Slot Attention by Changing Optimal Transport Costs
13:40	Kartik Ahuja et al.	Interventional Causal Representation Learning
13:50	Jovana Mitrović	Representation Learning and Causality
14:20		Poster Session 2 (virtual) + Break
15:30	Tobias Gerstenberg	A Counterfactual Simulation Model of Causal Judgment
16:00	Guy Van den Broeck	AI can learn from data. But can it learn to reason?
16:30		Break
17:00	Tobias Gerstenberg, Sriraam Natarajan, Mausam, Guy Van den Broeck, Devendra S. Dhami (B/O Organizers)	Panel Discussion: "Heading for a Unifying View on nCSI"
17:50	Matej Zečević (B/O Organizers)	Closing Remarks



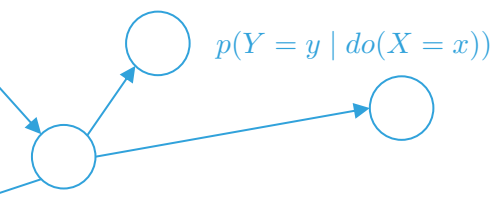
Chat **Full**

By joining this chat you are acknowledging that you have read and will abide by the NeurIPS Code of Conduct as shown on the NeurIPS website. Attendee...

great talk guy!

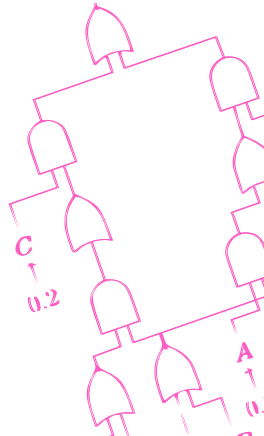
1

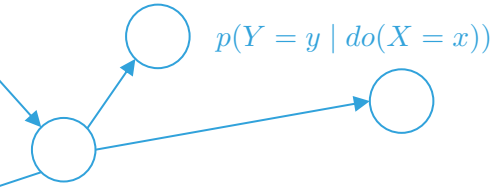
aleksander-molak-he-him 8:36 PM



The Aim of nCSI

Our aim is to bring together researchers interested in the integration of research areas in artificial intelligence including general machine and deep learning, symbolic and object-centric methods, and logic with rigorous formalizations of causality with the goal of developing next-generation AI systems.





Invited Speakers

Judea Pearl

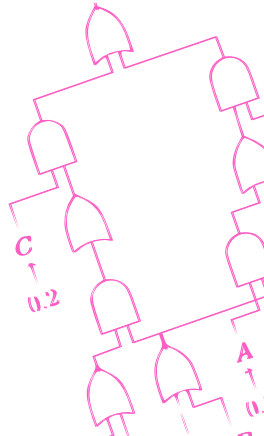
Opening Keynote for nCSI

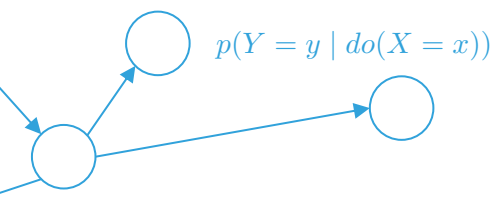
Mausam

“Neural Models with Symbolic Representations for Perceptuo-Reasoning Tasks”

Dhanya Sridhar

“Causal Inference from Text”





Invited Speakers

Jovana Mitrović

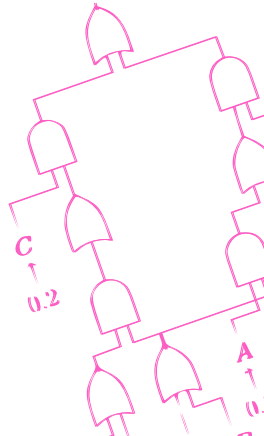
“Representation Learning and Causality”

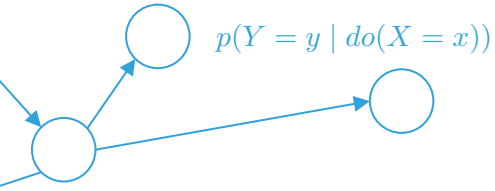
Tobias Gerstenberg

“A Counterfactual Simulation Model of Causal Judgment”

Guy Van den Broeck

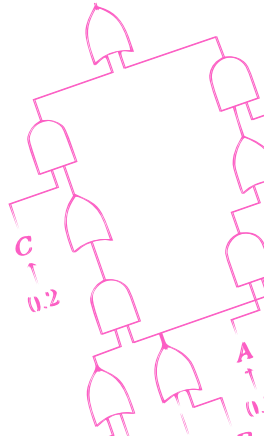
“AI can learn from data. But can it learn to reason?”

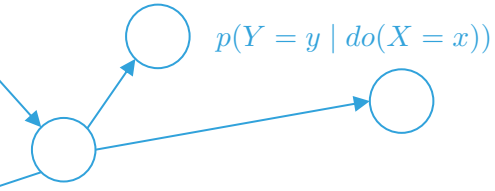




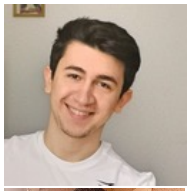
Submissions & Reviews

- 19 out of 24 submissions accepted (79%, 4 / 19 orals)
- 33 Reviewers, 7 Area Chairs, each paper 2+1 reviews

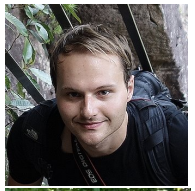




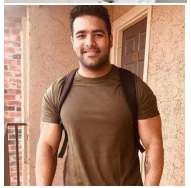
Organizers



Matej **Zečević**



Thomas **Kipf**



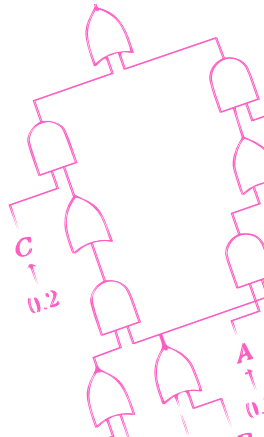
Devendra S. **Dhami**

Robert **Peharz**

Christina **Winkler**

Petar **Veličković**

Advisor: Kristian **Kersting**



Back to Today's Topic

Paper Currently Under Review

"Causal Explanations of SCM"

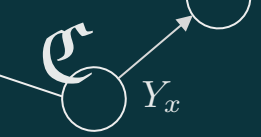
Causal Explanations of SCM

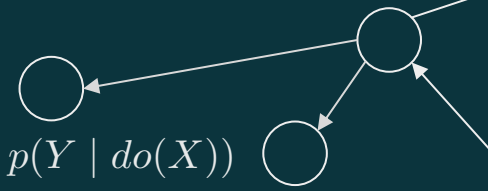
Matej Zečević, Devendra Singh Dhami,
Constantin Rothkopf, Kristian Kersting

TECHNISCHE
UNIVERSITÄT
DARMSTADT

TL;DR

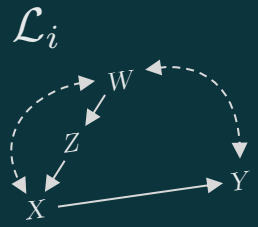
As a step towards causal XIL, we propose a solution to the lack of truly causal explanations from existing methods.





I | Why?

(Why Causality?)



TECHNISCHE
UNIVERSITÄT
DARMSTADT

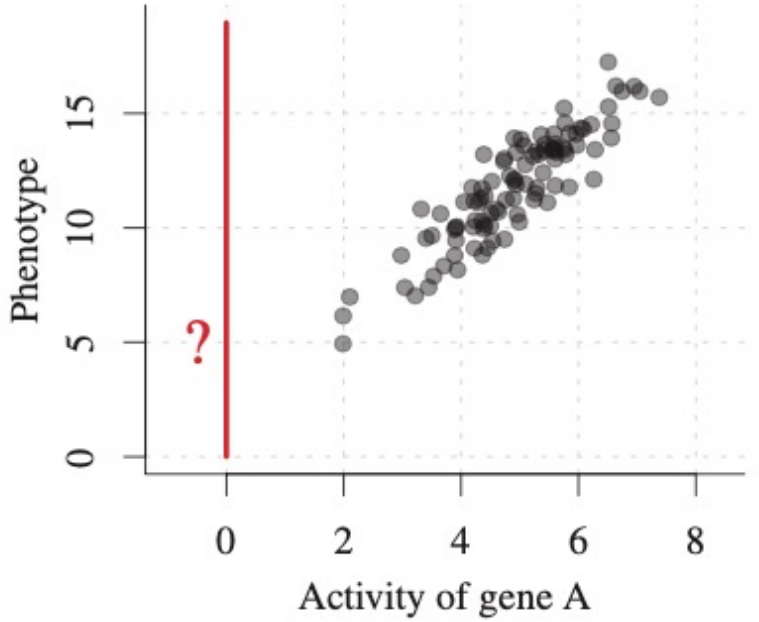


Figure from “Elements of Causal Inference” (2017) by
Jonas Peters, Dominik Janzing and Bernhard Schölkopf

The Full Picture

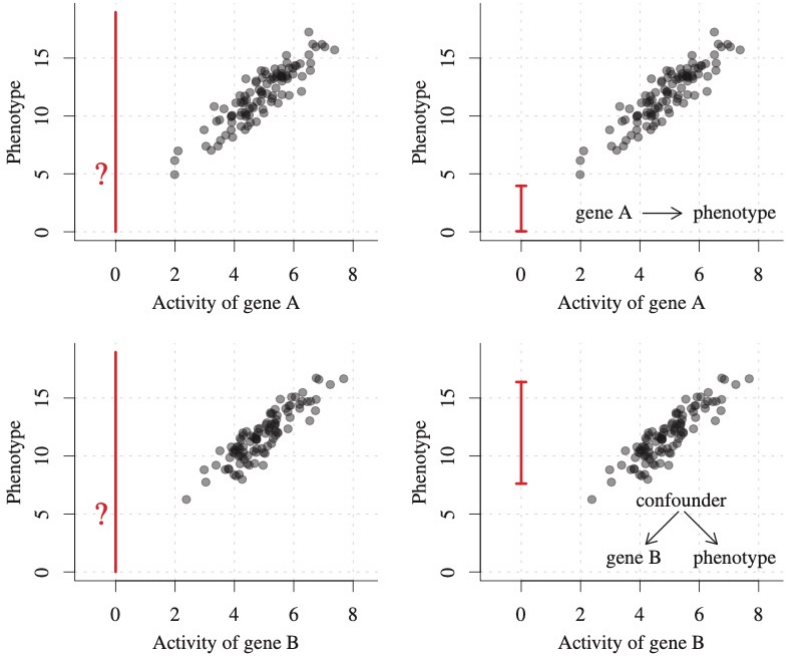
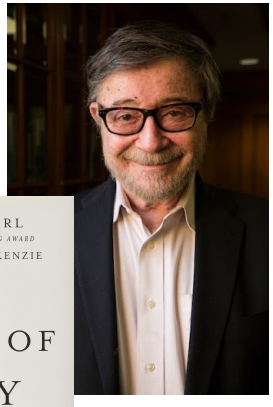
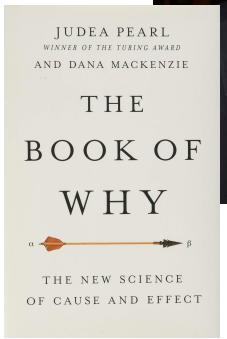


Figure from "Elements of Causal Inference" (2017) by Jonas Peters, Dominik Janzing and Bernhard Schölkopf

Why?





"To Build Truly Intelligent Machines, Teach Them Cause and Effect"

"All the impressive achievements of deep learning amount to just curve fitting"

Judea Pearl in "The Book of Why" and in an interview with quantamagazine in 2018

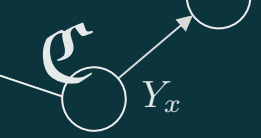
"As X-rays are to the surgeon, graphs are for causation."

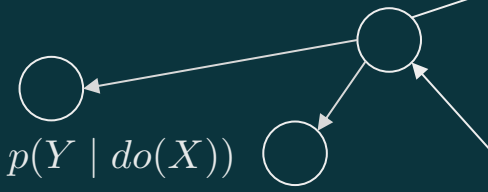
-Judea Pearl in Causality (2009)

"As graphs are to the causality, causal nets are for AI."

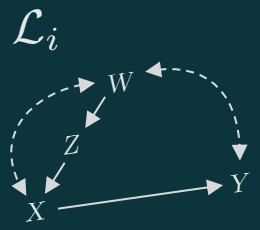
"As X-rays are to the surgeon, graphs are for causation."

-Judea Pearl in Causality (2009)





2 Background: Causality & XAI



TECHNISCHE UNIVERSITÄT DARMSTADT

Causal Inference 101

- **Structural Causal Model** (SCM) $\mathfrak{C} = (\mathbf{S}, P(\mathbf{U}))$

  describe the mechanistic relations of variables $\mathbf{V}$

  $$V_i := f_i(pa(V_i), U_i), \quad \text{where } i = 1, \ldots, d$$

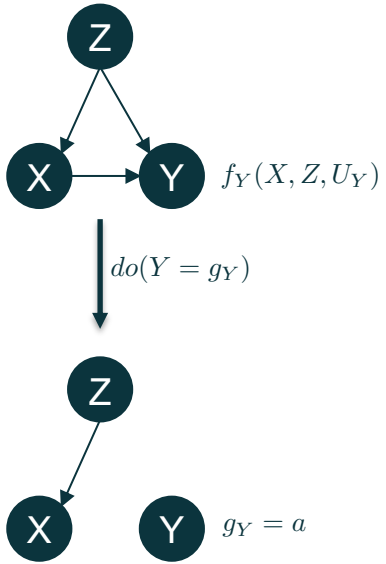
  where $P(\mathbf{U})$ factorizes the exogenous variables. (Markovian-SCM)

- An **intervention** actively replaces the origin mechanism $f_{\mathbf{W}}$ with a new mechanism, denoted by $do(\mathbf{W} = g_{\mathbf{w}})$.

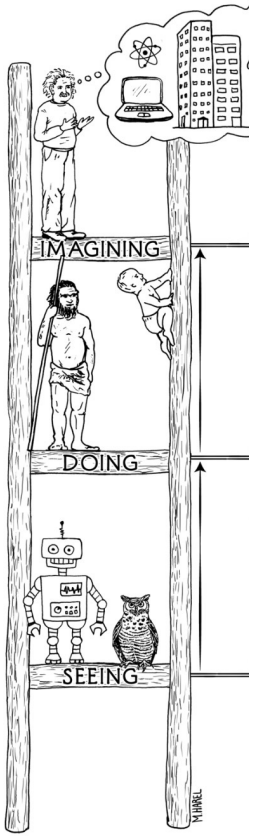
- An important view on **valuation** is given by:

  $$p^{\mathfrak{C}}(\mathbf{w} \mid do(\mathbf{z})) = \sum_{\{\mathbf{u} | \mathbf{W}_{\mathbf{z}}(\mathbf{u}) = \mathbf{w}\}} p(\mathbf{u})$$

  where $\mathbf{W}_{\mathbf{z}} : \mathbf{U} \mapsto \mathbf{W}$.



Why?



3. COUNTERFACTUALS

ACTIVITY: Imagining, Retrospection, Understanding

QUESTIONS: *What if I had done …? Why?*
(Was it X that caused Y? What if X had not occurred? What if I had acted differently?)

EXAMPLES: Was it the aspirin that stopped my headache? Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?

2. INTERVENTION

ACTIVITY: Doing, Intervening

QUESTIONS: *What if I do …? How?*
(What would Y be if I do X?
How can I make Y happen?)

EXAMPLES: If I take aspirin, will my headache be cured? What if we ban cigarettes?

1. ASSOCIATION

ACTIVITY: Seeing, Observing

QUESTIONS: *What if I see ...?*
(How are the variables related?
How would seeing X change my belief in Y?)

EXAMPLES: What does a symptom tell me about a disease? What does a survey tell us about the election results?

Pearl Causal Hierarchy (PCH), “The Book of Why” (2018)

On Pearl’s Hierarchy and the Foundations of Causal Inference

Elias Bareinboim, Juan Correa, Duligur Ibeling, Thomas Icard

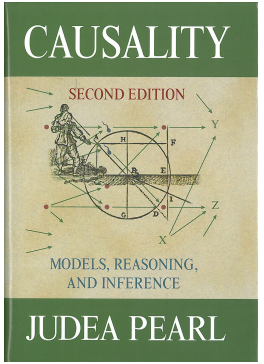
causalai.net/r60

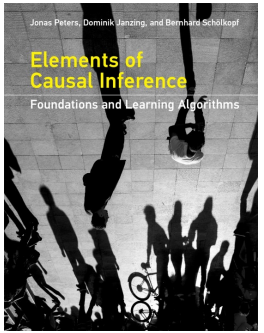
Theorem 1. *[Causal Hierarchy Theorem (CHT), formal version] With respect to the Lebesgue measure over (a suitable encoding of $\mathcal{L}_3$-equivalence classes of) SCMs, the subset in which any* PCH *collapse occurs is measure zero.* ■

TECHNISCHE UNIVERSITÄT DARMSTADT

Pointers to Causal Inference References

- Judea Pearl, “**Causality**”, Cambridge University Press, 2009.
- Peters et al., “**Elements of Causal Inference**”, MIT Press, 2017.
- Elias Bareinboim Lecture “**Causal Data Science**”, 2019.
  `https://www.youtube.com/watch?v=dUsokjG4DHc`
- Brady Neal’s Free Online Course ”**Introduction to Causal Inference**”, 2020.
  `https://www.bradyneal.com/causal-inference-course`
- Jonas Peters Lecture Series “**Causality**”, 2017.
  `https://www.youtube.com/watch?v=zvrcyqcN9Wo`

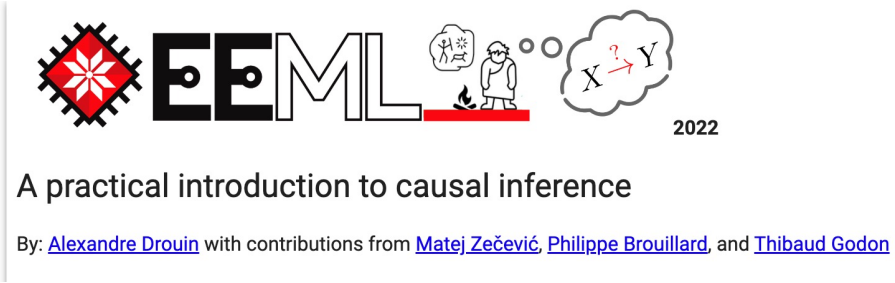




TECHNISCHE
UNIVERSITÄT
DARMSTADT

Code Tutorials, Hands-on!

- Alexandre Drouin's Code Tutorial for EEML 2022



https://colab.research.google.com/drive/13Bsvvl5l3uR1hbVdpMAFR13gdjwoJ6if?usp=sharing

- My Code Tutorial for SMLW 2022

https://github.com/zecevic-matej/SMLW-Causality-Tutorial/blob/main/Tutorial-on-Causality.ipynb

eXplainable Artificial Intelligence (XAI)

Explainable and Interpretable AI/ML. A great body of work within deep learning has provided visual means for explanations of how a neural model came up with its decision i.e., importance estimates for a model's prediction are being mapped back to the original input space e.g. raw pixels in the arguably standard use-case of computer vision (Selvaraju et al., 2017; Schulz et al., 2020). Formally defined in (Sundararajan et al.), we simply have that $A_F(\mathbf{x}) = (a_1, \ldots, a_n) \in \mathbb{R}^n$ is an attribution of predictive model F when a_i is the contribution of x_i for prediction $F(\mathbf{x})$ (with $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$). Recently, Stammer et al. (2021) argued that such explanations are insufficient for any task that requires symbolic-level knowledge while comparing the existing state of explanations to "children that are only able to point fingers but lack articulation". (Stammer et al., 2021) therefore proposed a neuro-symbolic explanation scheme to revise and ultimately circumvent "Clever Hans" like behavior from learned models in a XIL user-model loop (Teso & Kersting, 2019). On the causal end, (Schwab & Karlen, 2019) proposed a model-agnostic approach that can generate explanations following the idea of Granger causality (which is very different from Pearlian causality as it captures "temporal relatedness" which holds in their setting as input precedes output). Specifically, they train a surrogate model to capture to what degree certain inputs cause outputs in the model to be explained. They achieve this by simply comparing the prediction loss of the model for the original input $\mathcal{L}(y, \hat{y}_X)$ (where X is the input) with the alternate prediction loss when a certain feature i is being removed $\mathcal{L}(y, \hat{y}_{X \setminus \{i\}})$. On the Pearlian side of explanations, the arguably closest works on explainable AI/ML can be found in research around fairness (Kusner et al., 2017; Plecko & Bareinboim, 2022). For instance, Karimi et al. (2020) investigated how to best find a counterfactual that flips a decision of interest e.g. an applicant for a credit is rejected and the question is now which counterfactual setting (changes to the applicant) would have resulted in a credit approval. From a purely causal viewpoint, our work might be compared to the definitions of Halpern (2016) for "actual causation."

eXplainable Interactive Learning (XIL)

Algorithm 1 CAIPI takes as input a set of labelled examples $\mathcal{L}$, a set of unlabelled instances $\mathcal{U}$, and iteration budget T.

$f \leftarrow \textsc{Fit}(\mathcal{L})$
repeat
  $x \leftarrow \textsc{SelectQuery}(f, \mathcal{U})$
  $\hat{y} \leftarrow f(x)$
  $\hat{z} \leftarrow \textsc{Explain}(f, x, \hat{y})$
  Present x, $\hat{y}$, and $\hat{z}$ to the user
  Obtain y and explanation correction $\mathcal{C}$
  $\{(\bar{x}_i, \bar{y}_i)\}_{i=1}^{c} \leftarrow \textsc{ToCounterExamples}(\mathcal{C})$
  $\mathcal{L} \leftarrow \mathcal{L} \cup \{(x, y)\} \cup \{(\bar{x}_i, \bar{y}_i)\}_{i=1}^{c}$
  $\mathcal{U} \leftarrow \mathcal{U} \setminus (\{x\} \cup \{\bar{x}_i\}_{i=1}^{c})$
  $f \leftarrow \textsc{Fit}(\mathcal{L})$
until budget T is exhausted or f is good enough
return f

During interactions between the system and the user, three cases can occur: **(1) Right for the right reasons:** The prediction and the explanation are both correct. No feedback is requested. **(2) Wrong for the wrong reasons:** The prediction is wrong. As in active learning, we ask the user to provide the correct label. The explanation is also necessarily wrong, but we currently do not require the user to act on it. **(3) Right for the wrong reasons:** The prediction is correct but the explanation is wrong. We ask the user to provide an explanation *correction* $\mathcal{C}$.

“Clever Hans”

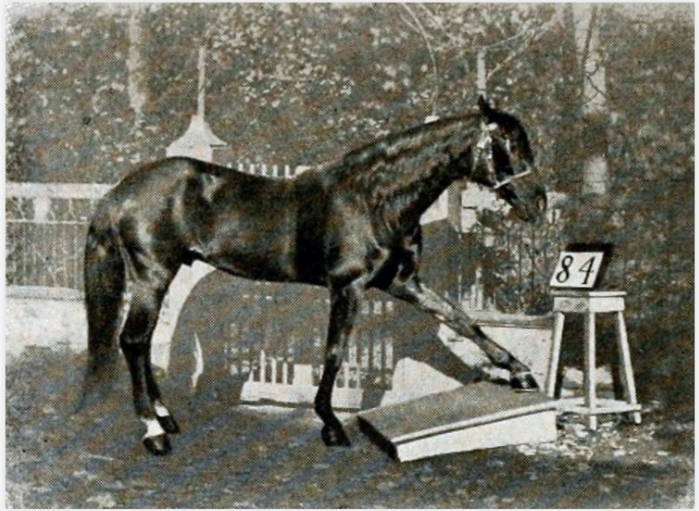


Figure 2: **The clever horse named Hans.** The image shows Hans in 1909 solving an arithmetic question. Hans is an Orlow-Traber horse, a race from Russia for which breeding began end-19th century.

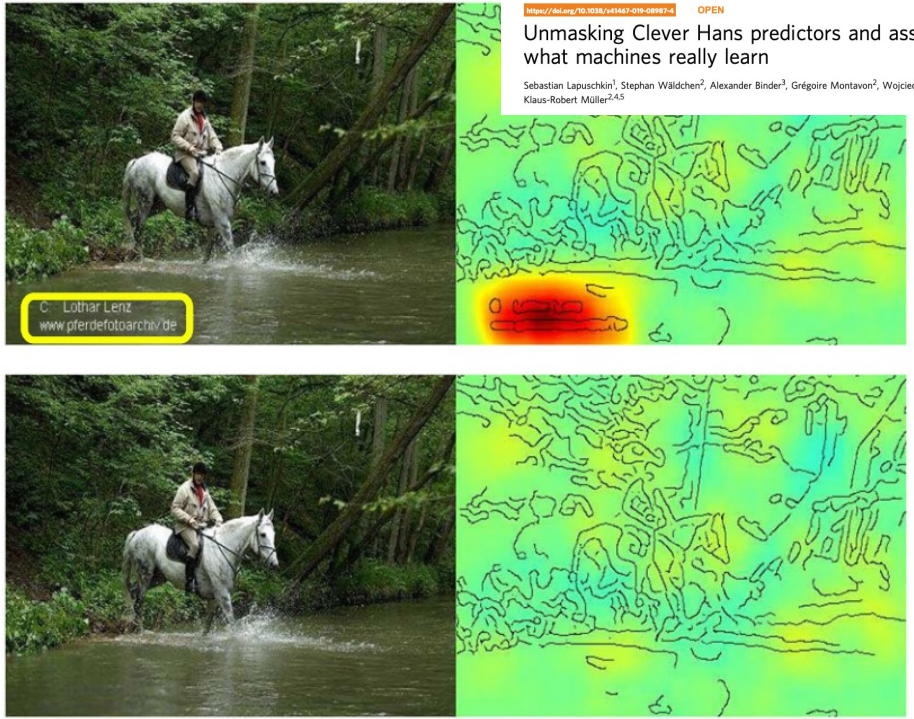
nature COMMUNICATIONS

ARTICLE

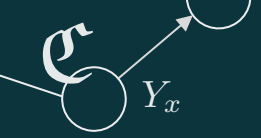
https://doi.org/10.1038/s41467-019-08987-4 OPEN

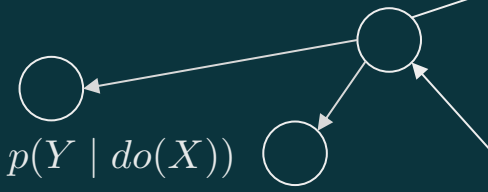
Unmasking Clever Hans predictors and assessing what machines really learn

Sebastian Lapuschkin[1], Stephan Wäldchen[2], Alexander Binder[3], Grégoire Montavon[2], Wojciech Samek[1] & Klaus-Robert Müller[2,4,5]

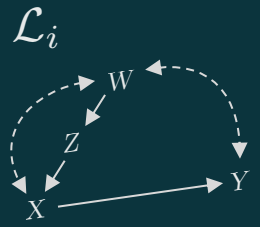


TECHNISCHE UNIVERSITÄT DARMSTADT





3 Motivation & What's New



TECHNISCHE UNIVERSITÄT DARMSTADT

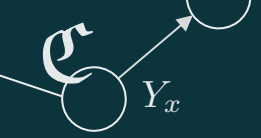
Motivation

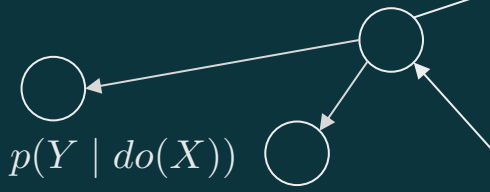
- In **explanatory interactive learning** (XIL) the user queries the learner, then the learner explains its answer to the user and finally the loop repeats.
- XIL is attractive for two reasons, (**1**) the learner becomes better and (**2**) the user's trust increases.
- For both reasons to hold, the learner's *explanations must be useful* to the user and the user must be allowed to ask useful questions. Ideally, both questions and explanations should be grounded in a *causal model* since they avoid spurious fallacies.

==> Causal XIL

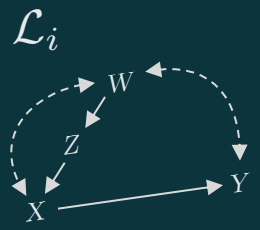
Contributions

- We show that existing explanation methods are **not** guaranteed to be causal even when provided with a Structural Causal Model (SCM)
- We derive from first principles an explanation method that makes full use of a given SCM, which we refer to as **SCE (E standing for Explanation)**
- We conduct several experiments including a user study with 22 participants to investigate the virtue of SCE as causal explanations of SCMs.



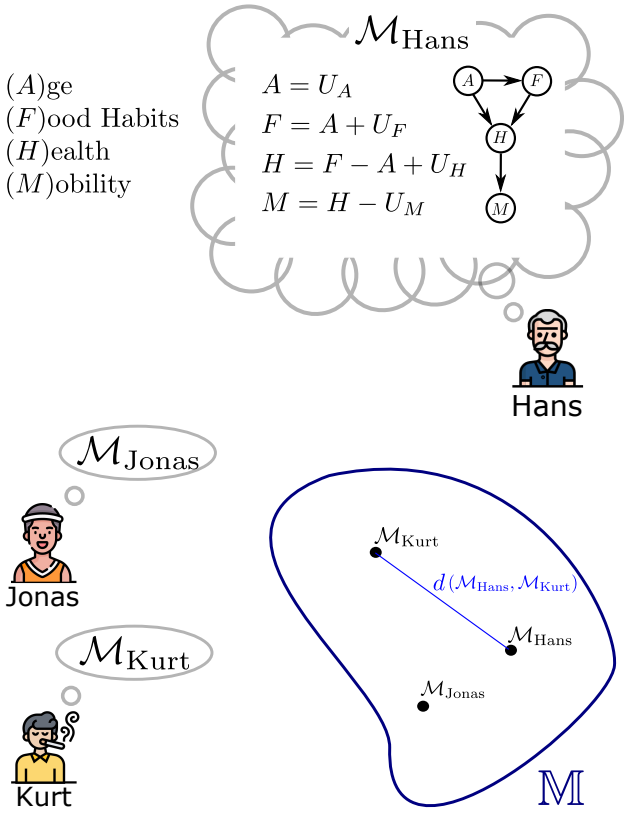


4 The “Causal Hans” Example

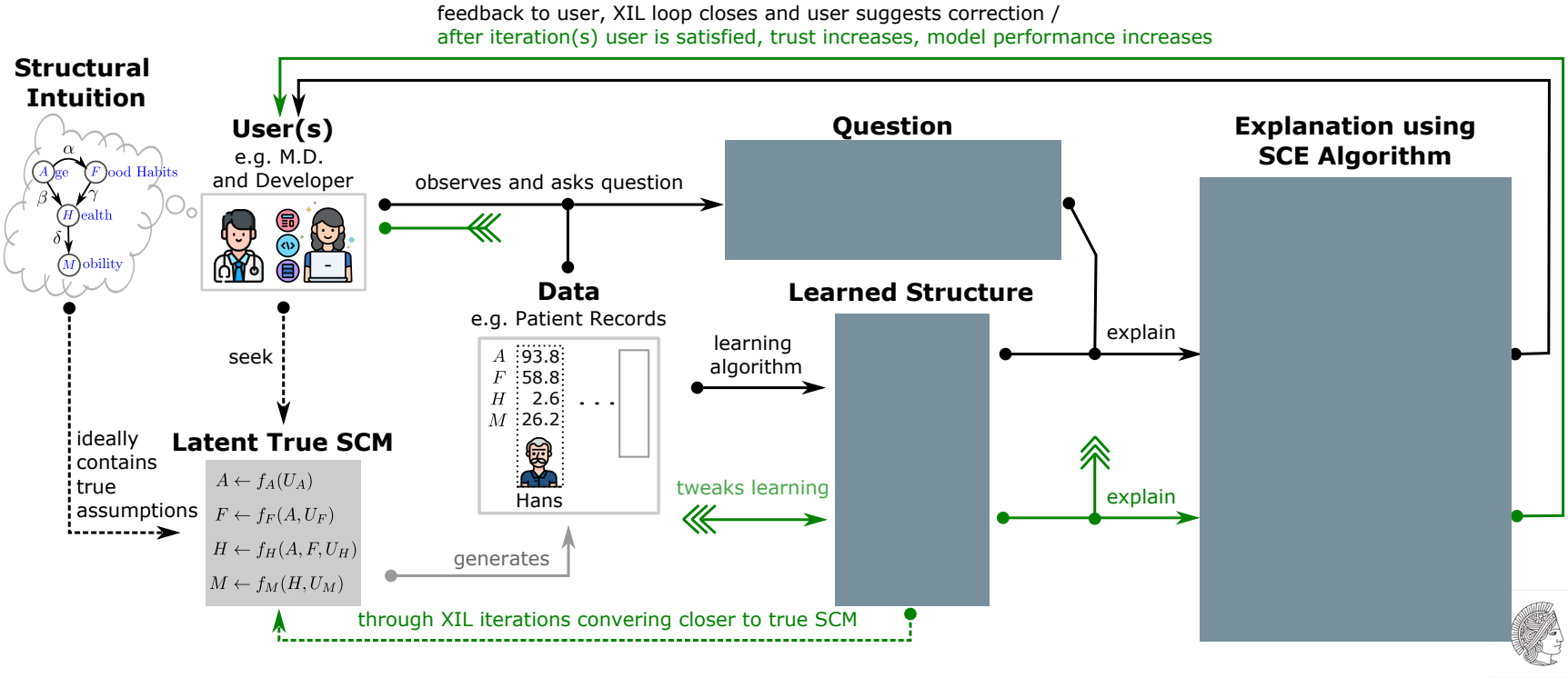


TECHNISCHE UNIVERSITÄT DARMSTADT

Philosophical Point: Human Thoughts in Terms of SCM



The Causal XIL Loop (Before Our Derivation)



Asking a Question about Individuals

We will use the Causal Hans example that we just introduced as our workbench to derive a computable algorithm for truly causal explanations. For the sake of simplicity we assume that the structural intuition that the user/expert can capture about the problem using their mental model can be described as an SCM[2]. In the following, consider an SCM that generates patients's medical records described by numerical representations for age, nutrition, overall health and mobility respectively ($\mathbf{V} = \{A, F, H, M\}$). Next, let's consider some samples from said SCM. E.g. we might observe the data set containing the individual named Hans $\mathrm{H} = (a_H, f_H, h_H, m_H) = (93.8, 58.8, 2.6, 26.2)$ where each value could be associated with a discrete label e.g. $a_H = 93.8$ would be 93 years old, whereas $m_H = 26.2$ refers to a rather immobile person. The latter label is actually implicitly the assessment $m_H < \mu^M$ where $\mu_M = 35.6$ is the average mobility value for the population, that is, we observe Hans to be a rather immobile person *when compared (or relative) to the whole population*. With this we are in the position to pose a question like

Q1: *"Why is Hans's Mobility bad?"*

where the word "bad" refers to "bad relative to the population."

Formalizing the Question Type

Definition 1 (Why Question). *Let $x \in Val(X)$ be an instance of $X \in \mathbf{V}$ of SCM $\mathcal{M}$. Further, let μ^X be the empirical mean for a set of samples ($\mu^X := \frac{1}{n}\sum_i^n x_i$) and let $R \in \{<, >\}$ be a binary ordering relation. We call $Q_X := R(x, \mu^X)$ a (single) why question if Q_X is true.*

Checking back with the definition, we see that **Q1** defines a valid question for the Causal Hans example since $Q_M := m_H < \mu^M = 26.2 < 35.6$ evaluates to true.

What an SCM Tells Us About Hans

Next, we will discuss the structural intuition of the user (e.g. the M.D. in Fig.1). Generally, the true SCM $\mathcal{M}^*$ is latent but we can realistically expect to have access to partial knowledge (or estimate) of $\mathcal{M}^*$. Say, the user has intuition for an SCM $\mathcal{M}$ that contains the relations $A \xrightarrow{\alpha} F$, $A \xrightarrow{\beta} H$, $F \xrightarrow{\gamma} H$, $H \xrightarrow{\delta} M$ where $\alpha, \beta, \gamma, \delta$ denotes the respective causal effects[3] Further, $\alpha, \gamma, \delta > 0$ while $\beta < 0$ meaning that for instance an aging (higher value for A) will have a negative, decreasing effect on health (smaller value for H), and also $\beta > \gamma$ meaning the causal effect of aging on health is greater in absolute terms than the one of food habits onto health. Now when we intend on answering **Q1** it seems reasonable to start with the queried variable first, mobility in this case. We observe that M is an effect of H with $\gamma > 0$ meaning that since Hans has also below average health (and not just below average mobility) and lower health translates to lower mobility that m_H is inline with h_H. Traversing the chain further to the causes of H, which are A, F in $\mathcal{M}$ we observe two different scenarios. Since A is above average as Hans is an elderly person and $\beta < 0$ we can conclude that a_H is definitely an explanation for h_H whereas F with $\gamma > 0$ is actually a countering factor since Hans has a good diet beneficial to health.

Formulating an Explanation

Explanation 1 (for **Q1**). *"Hans's Mobility is bad because of his bad Health which is mostly due to his high Age although his Food Habits are good."*

Explanation 1 is a truly causal answer to the observation about Hans's mobility deficiency based on SCM $\mathcal{M}$. It captures both the existence and the "strength" of a causal relation. In the following we will capture and formalize our intuition that allowed us to derive Exp.1. This will allows us to move towards computing such causal explanations automatically.

Generalizing the Key Ideas in Logic

We mainly used four ideas or pieces of knowledge in our argument above: (I) that there is a relative notion in the why question Q_M like "why ... bad?" that implicitly compares an individual (here, Hans) to the remaining population, (II) note that by definition there can only exist a causal effect from some variable to another *if and only if* one is the argument of the other in a structural equation of $\mathcal{M}$, (III) the causal effect $\alpha_{X \to Y}$ allows us to assert whether the observed values for (x, y) are "surprising" or not (e.g. it was not surprising that $m_H < \mu^M$ after observing $h_H > 0$ and knowing that $\gamma > 0$ since decreasing health means decreasing mobility in general and Hans is old), and (IV) that some causal effects are more important or influental than others (e.g. age versus food habits w.r.t. health). We can neatly collect all information from (I-III) in a single tuple which we call causal scenario.

Definition 2. *The tuple* $C_{XY} := (\alpha_{X \to Y}, x, y, \mu^X, \mu^Y)$ *is called causal scenario.*

The (IV) point we can capture separately as will be shown below. Now, we finally express our build up intuition and understanding into rules expressed in first-order logic that will then allow us to compute causal explanations like Exp.1 automatically.

Definition 3 (Explanation Rules)**.** *Let* C_{XY} *denote a causal scenario, let* $s(x) \in \{-1, 1\}$ *be the sign of a scalar, let* $R_i \in \{<, >\}$ *be a binary ordering relation and let* $\mathcal{Z}_X = \{|\alpha_{Z \to X}| : Z \in \mathrm{Pa}_X\}$ *be the set of absolute parental causal effects onto* X*. We define FOL-based rule functions as*

(ER1) If $R_1 \neq R_2$*, then:* $R_1(s(\alpha_{X \to Y}), 0) \wedge (R_2(y, \mu^Y) \vee R_1(x, \mu^X))$,

(ER2) If $R_1 \neq R_2$*, then:* $R_1(s(\alpha_{X \to Y}), 0) \wedge R_1(y, \mu^Y) \wedge R_2(x, \mu^X)$*, and*

(ER3) If $|\mathcal{Z}_X| > 1$*, then* $Y \iff \arg\max_{Z \in \mathcal{Z}_X} Z$

indicating for each rule $ERi(\cdot) \in \{-1, 0, 1\}$ *how the causal relation* $X \to Y$ *satisfies that rule.*

Pronouncing the Rules + Inspiration

*ER*1	Excitation	"Y because of X [being low/high]"
*ER*2	Inhibition	"Y although X [is low/high]"
*ER*3	Preference	"mostly" + *ER*1 or *ER*2 pronunciation

Table 1: **Pronunciation Scheme.** Right shows the natural language reading of a rule's activation.

Deriving an Algorithm
Structural Causal Explanations (SCE)

Definition 4 (SCE). *Like before let $Q_X, \mathcal{M}$ be a valid why-question and some proxy SCM. Further, let $\boldsymbol{D} \in \mathbb{R}^{n \times |\mathbf{V}|}$ denote our data set. We define a recursion*

$$\mathbf{E}(Q_X, \mathcal{M}, \boldsymbol{D}) = (\bigoplus_{Z \in \mathrm{Pa}(X)} ER(Z \to X), \bigoplus_{Z \in \mathrm{Pa}(X)} \mathbf{E}(Q_Z, \mathcal{M}, \boldsymbol{D})) \tag{1}$$

where $\bigoplus_{i=1}^{n} v_i = (v_1, \ldots, v_n)$ denotes concatenation and ER checks each rule ERi (Def.3), and the recursion's base case is being evaluated at the roots of the causal path to X, that is, for some $Z \in \mathbf{V}$ with a path $Z \to \cdots \to X$ we have

$$\mathbf{E}(Q_Z, \mathcal{M}, \boldsymbol{D}) = \emptyset. \tag{2}$$

We call $\mathbf{E}(Q_X, \mathcal{M}, \boldsymbol{D})$ Structural Causal Explanation of $\mathcal{M}$.

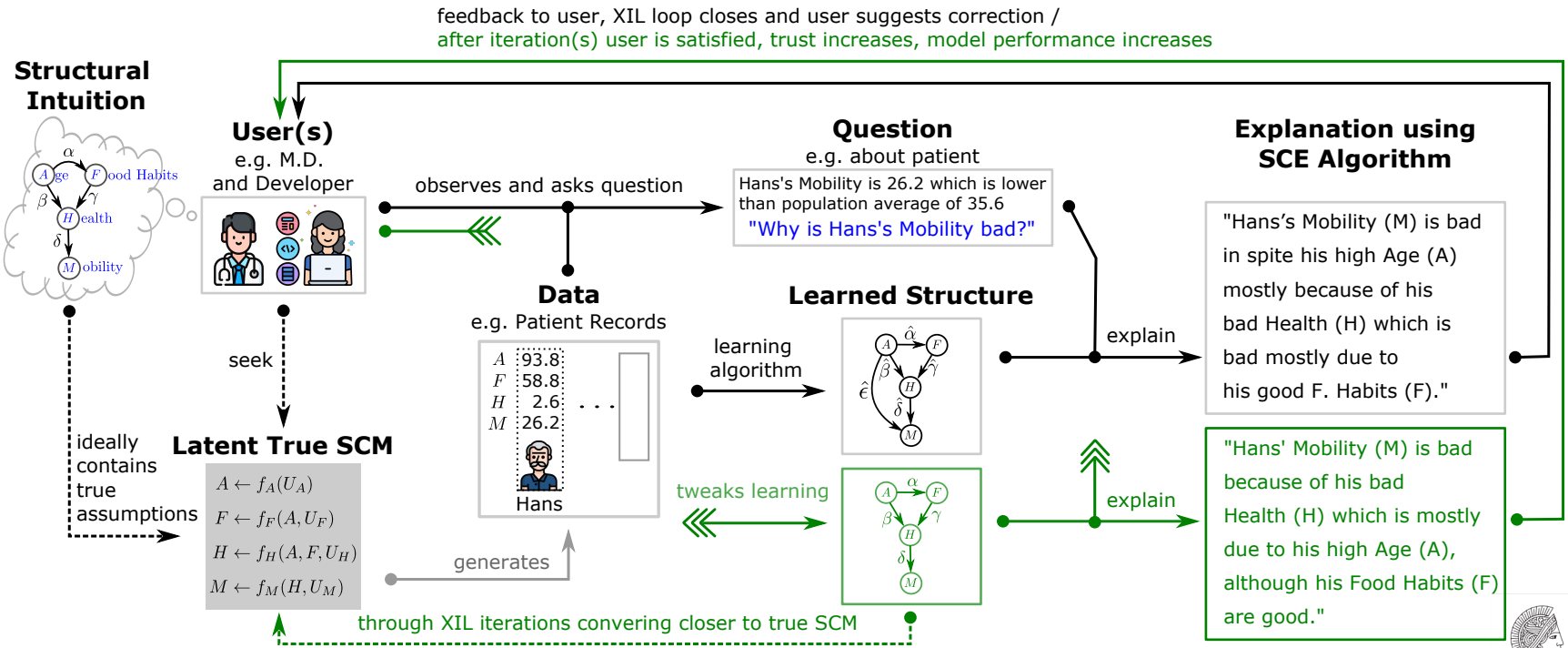
Our “Causal Hans” Example
Explanation Automated through SCE Algorithm

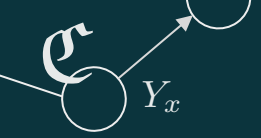
Causal Hans Example revisited (using SCE algorithm). To return one last time, we clearly see that for Q_M (corresponding to **Q1**) we can compute using Eq.1

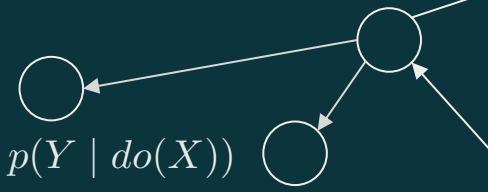
$$\begin{aligned}\mathbf{E}(Q_M, \mathcal{M}, \boldsymbol{D}) &= ((\boldsymbol{ER1} = -1), \bigoplus_{Z \in \{A,F\}} \mathbf{E}(Q_H, \mathcal{M}, \boldsymbol{D})) \\ &= (\ldots, (((\boldsymbol{ER1} = 1, \boldsymbol{ER3} = 1), \mathbf{E}(Q_A, \mathcal{M}, \boldsymbol{D})), ((\boldsymbol{ER2} = 1), \mathbf{E}(Q_F, \mathcal{M}, \boldsymbol{D})))), \\ &= (\ldots, ((\ldots, \emptyset), (\ldots, \emptyset))).\end{aligned}$$

So the recursion result is $H \to M : (\boldsymbol{ER1} = -1, \boldsymbol{ER2} = 0, \boldsymbol{ER3} = 0), A \to H : (\boldsymbol{ER1} = 1, \boldsymbol{ER2} = 0, \boldsymbol{ER3} = 1), F \to H : (\boldsymbol{ER1} = 0, \boldsymbol{ER2} = 1, \boldsymbol{ER3} = 0)$. This result *uniquely* identifies the human understandable pronunciation of our causal explanation in Exp.1. We provide a detailed explanation on the pronunciation scheme and also intuitive namings for the rules ER_i in the appendix. It is worthwhile noting that the natural language choice of words to express the interpretation is not implied by the form of the SCE e.g., while Hans’s mobility is said to be “bad”, a car’s remaining fuel is rather considered to be “low”.

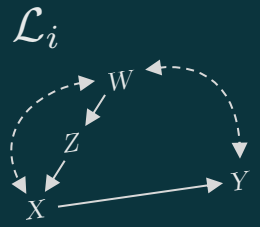
The Causal XIL Loop for SCE







5 Checking SCE, Theory & Empirics



TECHNISCHE UNIVERSITÄT DARMSTADT

Noteworthy Properties of SCE

Proposition 1. *For any causal scenario the rules ER1 and ER2 will be mutually exclusive.*

Proof. First, we code the binary ordering relations $<, >$ to represent 0 and 1 respectively. Second, we observe that ER$i \in \{<, >\}, i \in \{1, 2\}$ always involves the triplet $T = (R(s(\alpha_{X\to Y}), 0), R(y, \mu^Y), R(x, \mu^X))$. Third, let $\mathbb{T} := \{0, 1\}^3$ be the set of all such triples as their code words, so $T \in \mathbb{T}$. Looking at the total number of possible scenarios $|\mathbb{T}| = 2^3 = 8$, we easily see that ER1 covers codewords $\{010, 011, 100, 101, 000, 111\}$ and ER2 covers the codewords $\{001, 110\}$, and together they cover all codewords ER1 $\cup$ ER2 $= \mathbb{T}$. Since any single scenario C_{XY} is uniquely mapped to a codeword, it will either trigger ER1 or ER2 but never both. □

Proposition 2. *The SCE recursion always terminates.*

Proof. The recursion's base case is reached when a root node is reached i.e., a node i with $\mathrm{Pa}_i = \emptyset$. An SCM implies a finite DAG, so root nodes are reached eventually. □

Theorem 1. *The output of any causal structure learning algorithm can be used to compute SCE.*

Proof. The proof for this theorem is surprisingly simple in that the SCM $\mathcal{M}$ used in the SCE recursion is only required to provide some kind of numerical value $\alpha_{i\to j}$ for the relation of any variable pair (i, j), that is, a matrix $\mathrm{A} \in \mathbb{R}^{|\mathbf{V}|\times|\mathbf{V}|}$ which represents a linear SCM or a SCM where each $\alpha_{i\to j}$ represents a causal effect description. If the matrix A is an adjacency matrix living in $[0, 1]^{|\mathbf{V}|\times|\mathbf{V}|}$, then we simply have no information about ER3 since all causal effects are assumed to be the same. Since any causal structure learning algoirthm will produce a causal graph represented by a matrix, we have that we can compute SCE. □

Failure of Previous Methods (e.g. CXPlain) and Comparison to SCE

Query / Question

"Why is Hans's Mobility bad?"

Hans

Age,	Food Habits,	Health,	Mobility
93.8,	58.8,	2.6,	26.2

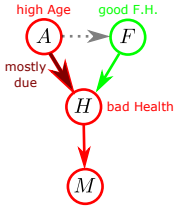
Average Mobility in Population 35.6

Causal Explanation with SCE Algorithm

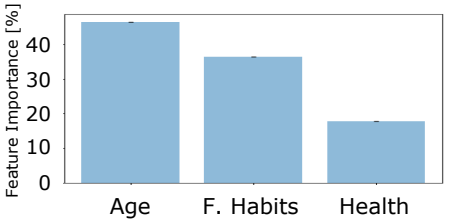
Answer:

"Hans' Mobility (M) is bad because of his *bad* Health (H) which is *mostly* due to his *high* Age (A), *although* his Food Habits (F) are *good*."

Graphical Interpretation:

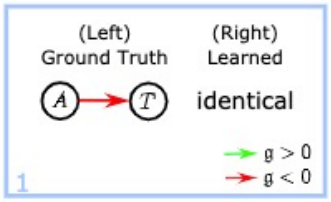


"Causal" Explanation with CXPlain

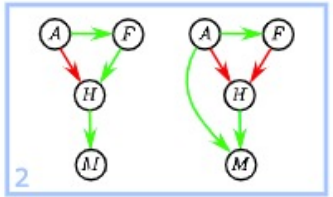


What we observe is a distribution over importance scores where all factors are being deemed relevant and "causal" to the output (which in this case is the mobility of Hans). Also the highest attribution is given to age, then food habits and finally the lowest to health. This single observation makes apparent two important shortcomings: **(I)** from the output we do not know which is a direct (H) and which are indirect (A, F) causes while the ordering of presentation in SCE clearly distincts the former from the latter, and **(II)** we have no information on the causal effect, that is, we cannot tell in which way a variable with high attribution will affect the predicted variable, for example food habits received a high importance score like age but age will have a detrimental effect on mobility whereas food habits will have a beneficial effect—again, SCE fixes this by discriminating the positive and negative cases. We also ran CXPlain for more queries and observed two further fallacies that we discuss in the Appendix. One the flip side, like SCE, the CXPlain attribution was able to identify the that the effect of aging is stronger that that of dieting.

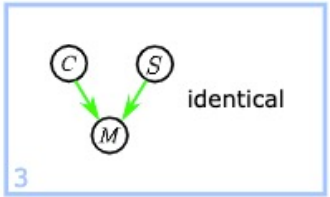
Quality of Learned Explanations



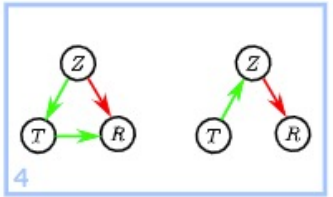
(A) Altitude, (T) Temperature



(A) Age, (F) Nutrition, (H) Health, (M) Mobility



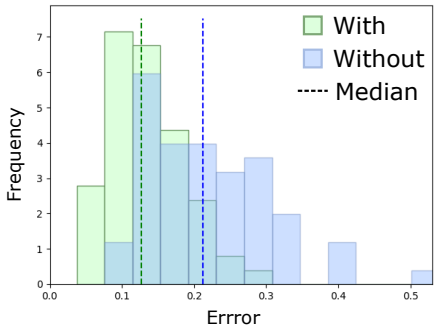
(C) Tank Size, (S) Driving Style, (M) Fuel Savings

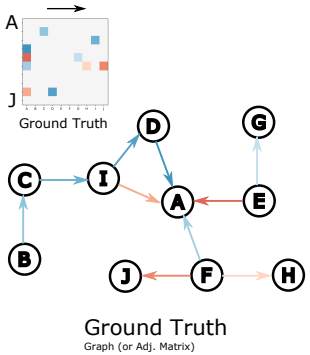


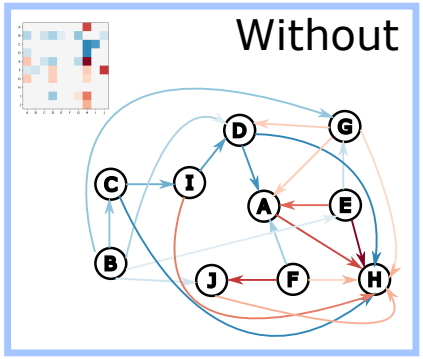
(T) Treatment, (Z) Pre-Condition, (R) Recovery

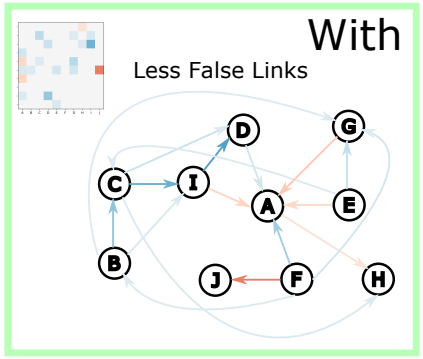
1	*"Why is the Temperature at the Matterhorn low?"*	(*Question*)
	"The Temperature at the Matterhorn is low because of the high Altitude."	(Ground Truth)
	"The Temperature at the Matterhorn is low because of the high Altitude."	**(Learned)**
2	*"Why is Hans's Mobility bad?"*	
	"Hans's Mobility is bad because of his bad Health which is mostly due to his high Age, although his Food Habits are good."	
	"Hans's Mobility, in spite his high Age, is bad mostly because of his bad Health which is bad mostly due to his good Food Habits."	
3	*"Why is your personal car's left Mileage low?"*	
	"Your left Mileage is low because of your small Car and your bad Driving Style."	
	"Your left Mileage is low because of your small Car and your bad Driving Style."	
4	*"Why did Kurt not Recover?"*	
	"Kurt did not Recover because of his bad Pre-condition, although he got Treatment."	
	"Kurt did not Recover because of his bad Pre-Condition, which were bad although he got Treatment."	

Using Explanations for Regularization in Learning the Right Graph







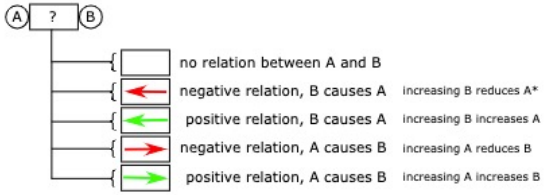


TECHNISCHE UNIVERSITÄT DARMSTADT

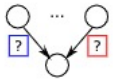
User Study with 22 Participants, Our Survey Document:

Questions

Is there a causal relation?
If yes, then is it positive or negative?
If yes, then in which direction?



If there are multiple relations, what is the order of strength?



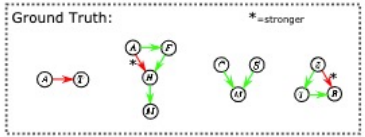
e.g. 1 most important cause
2 2nd most important cause

*dependent on context thereby the word "increasing" might change to "improving" and analogue for the negative

To avoid bias in drawing relations,
we don't provide any hints on a graph structure and
we randomize the sorting of the variables.

To provide more clarity we depict the names of the concepts
with additional illustrations.

The participants are asked to perform induction based
on personal data/experience i.e., they only see the orange and blue boxes.



Example 1

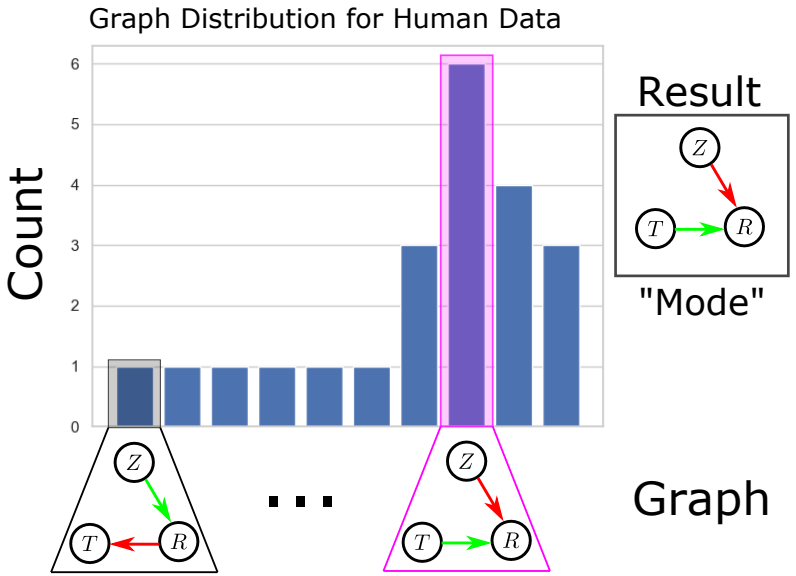
Example 2

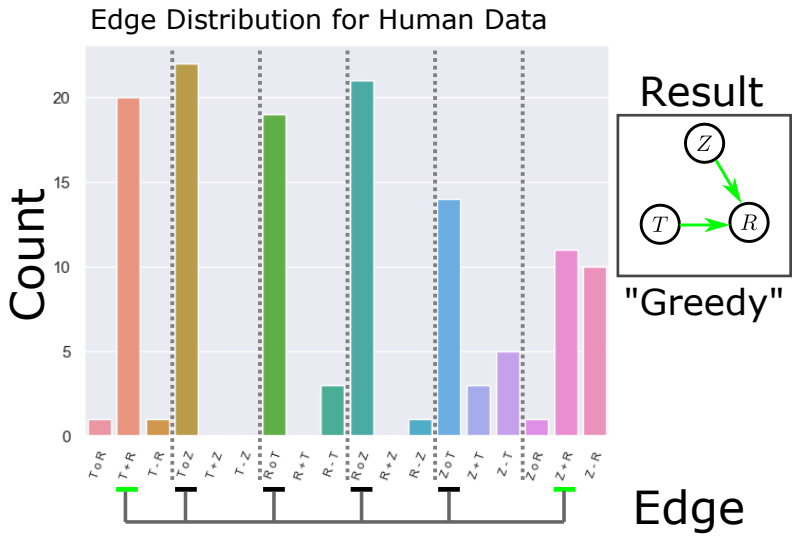
Example 3

Example 4

TECHNISCHE UNIVERSITÄT DARMSTADT

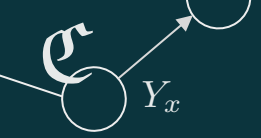
Aggregating the Human Data

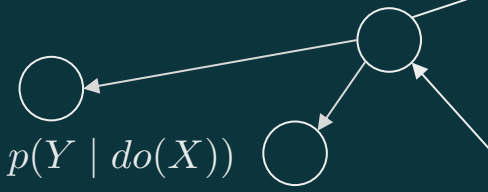




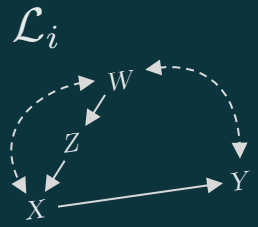
Comparing Human vs Machine SCEs

Humans: *"Hans's Mobility is bad because of his bad Health which is mostly due to his high Age, although his Food Habits are good."*
Machines: *"Hans's Mobility, in spite his high Age, is bad mostly because of his bad Health which is bad mostly due to his good Food Habits."*





Z | Recent Works of Ours



TECHNISCHE UNIVERSITÄT DARMSTADT

Relating Graph Neural Networks to Structural Causal Models

Matej Zečević, Devendra Singh Dhami, Petar Veličković, Kristian Kersting

got a lot of attention, Judea tweeted, whole community went nuts..

Finding Structure and Causality in Linear Programs

Matej Zečević, Florian Peter Busch, Devendra Singh Dhami, Kristian Kersting

published at ICLR OSC WS

Interventional Sum-Product Networks: Causal Inference with Tractable Probabilistic Models

Matej Zečević, Devendra Singh Dhami, Athresh Karanam, Sriraam Natarajan, Kristian Kersting

published in NeurIPS 2021

Can Foundation Models Talk Causality?

Moritz Willig, Matej Zečević, Devendra Singh Dhami, Kristian Kersting

got a little attention, Judea also tweeted
published at UAI CRL WS

XAI Establishes a Common Ground Between Machine Learning and Causality

Matej Zečević, Devendra Singh Dhami, Constantin A. Rothkopf, Kristian Kersting

On the Tractability of Neural Causal Inference

Matej Zečević, Devendra Singh Dhami, Kristian Kersting

Can Linear Programs Have Adversarial Examples? A Causal Perspective

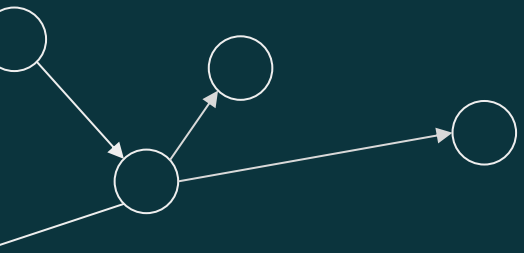
Matej Zečević, Devendra Singh Dhami, Kristian Kersting

The Causal Loss: Driving Correlation to Imply Causation

Moritz Willig, Matej Zečević, Devendra Singh Dhami, Kristian Kersting

Towards a Solution to Bongard Problems: A Causal Approach

Salahedine Youssef, Matej Zečević, Devendra Singh Dhami, Kristian Kersting



TECHNISCHE UNIVERSITÄT DARMSTADT

Thank you!

Questions?

matej.zecevic@tu-darmstadt.de | https://www.matej-zecevic.de

Further, my gratitude and thanks go out to Kristian Kersting, Devendra Dhami, all our collaborators and AIML@TU Darmstadt