MLSS 2020 Causal Inference II

Bernhard Schölkopf & Stefan Bauer

Thanks!

Co-organizers

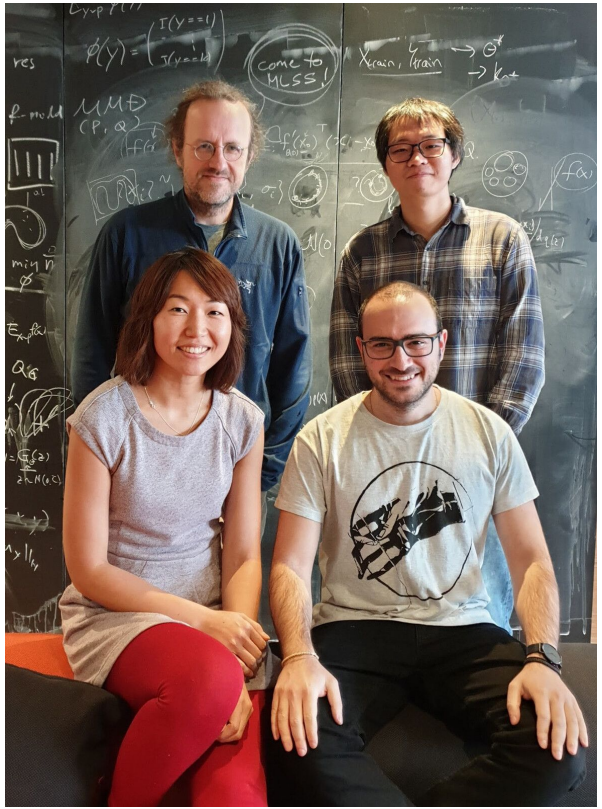
- Mijung Park
- Wittawat Jitkrittum
- Georgios Arvanitidis
- Bernhard Schölkopf

Event Management

- Barbara Kettemann
- Maria Püschel

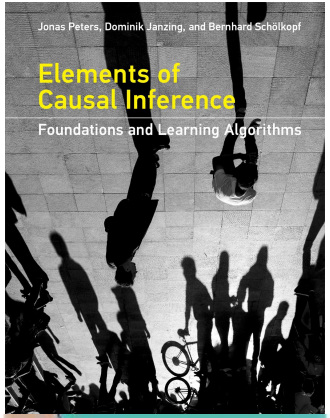
Administrative Support

- Matthias Tröndle
- Aline Dietrich



Additional Material

- Dominik Janzing & Bernhard Schölkopf: MLSS Causality Tutorial 2013 : Causality 1 https://www.youtube.com/watch?v=KsbftkwZTq4 & Causality 2 https://www.youtube.com/watch?v=Y5M6qwbidu0
- Jonas Peters: Mini Course Causality (8h) given at MIT: https://www.youtube.com/watch?v=zvrcyqcN9Wo
- Sebastian Weichwald & Dominik Janzing (4h): Causal Inference Tutorial CCN 2019 https://www.youtube.com/watch?v=WFnz_CWWFmM
- Miguel Hernan & Jamie Robbins: Causal Inference book https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/
- Jonas Peters, Dominik Janzing and Bernhard Schölkopf: Elements of Causal Inference http://web.math.ku.dk/~peters/elements.html
- Judea Pearl & Dana Mackenzie: The book of Why http://bayes.cs.ucla.edu/WHY/
- Richard Scheines: An Introduction to Causal Inference http://mlg.eng.cam.ac.uk/zoubin/SALD/Intro-Causal.pdf



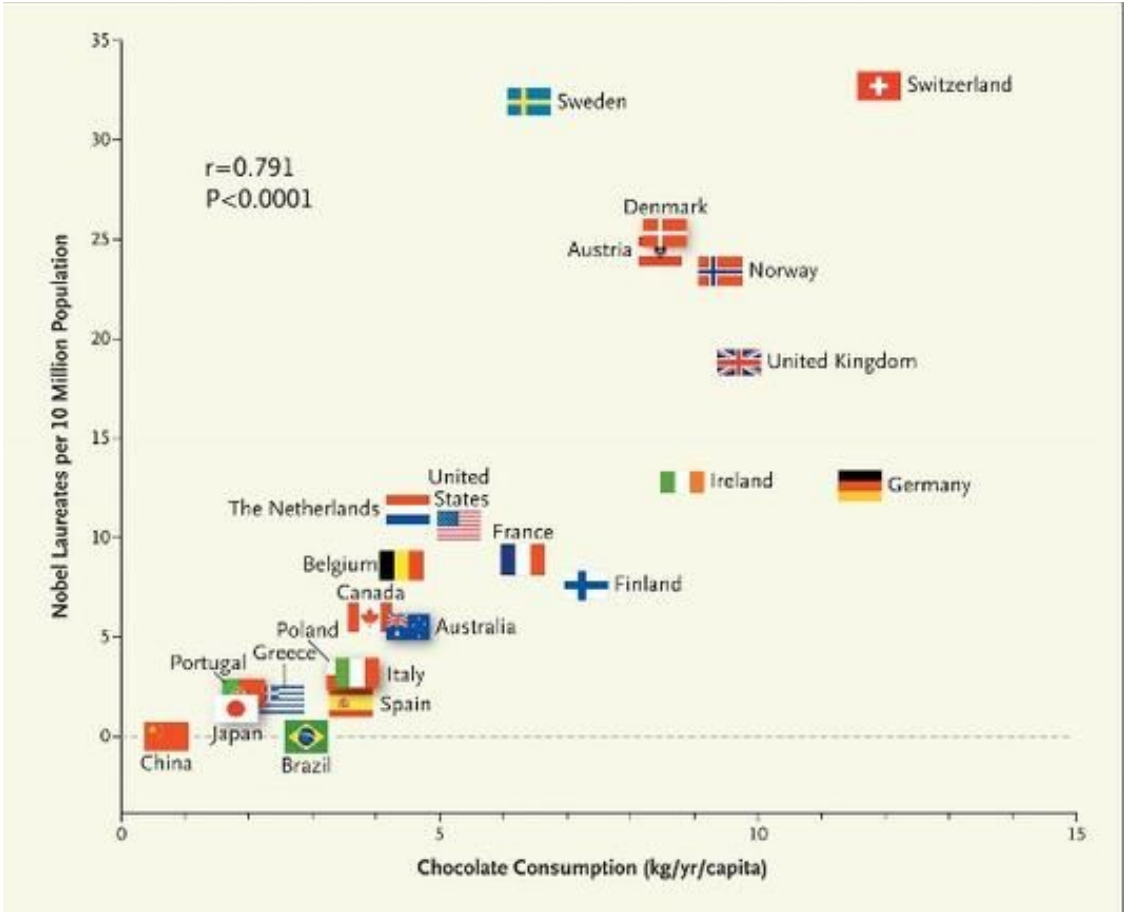


Image from: Messerli, F. H., et al. "Chocolate and Your Health." *N Engl J Med* 367.16 (2012): 1562-4.

Causal Models as Posets of Distributions.

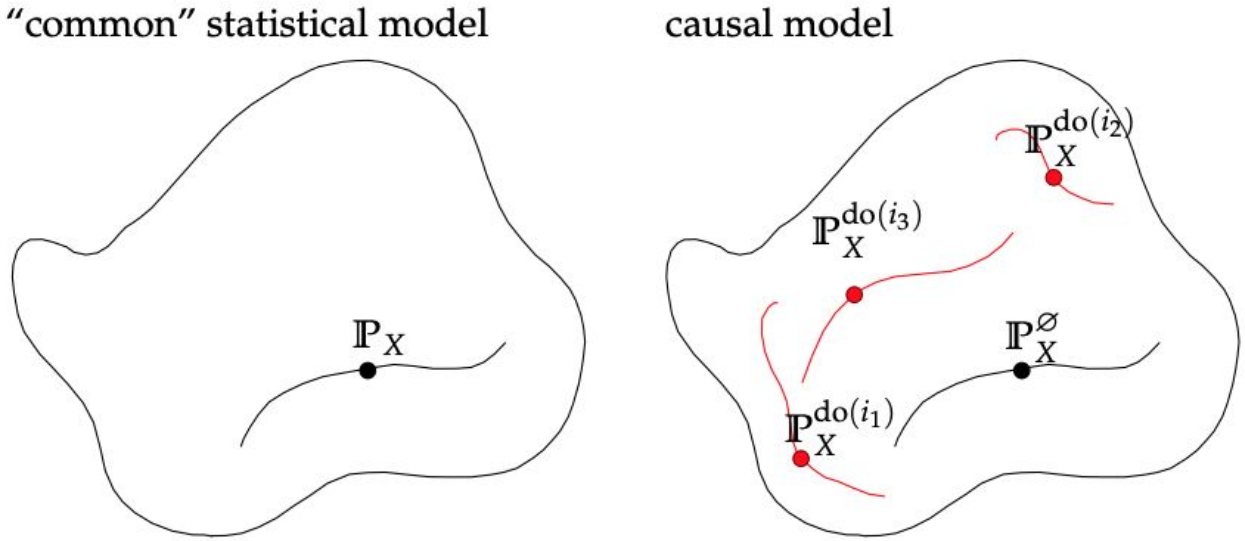
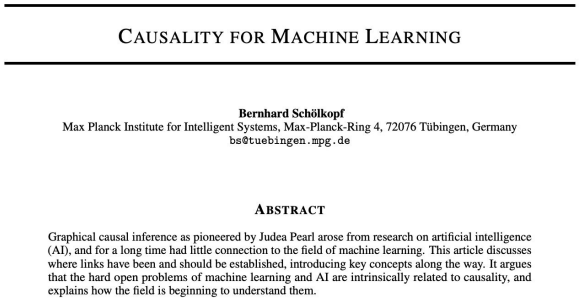


Image & motivation from: S. Weichwald, Pragmatism and variable transformations in causal modelling, PHD Thesis 2019
https://www.research-collection.ethz.ch/handle/20.500.11850/377699?locale-attribute=de

Very brief orientation

- Randomized Control Trial
- Observational Study
- Reichenbach Principle
- Causal Diagrams / Graphs as underlying assumption
  - Different results for different graphs
  - Must be made explicit what the assumptions on graph are

Today: What can we do if we do not know the graph? And what are connections with machine learning?



CAUSALITY FOR MACHINE LEARNING

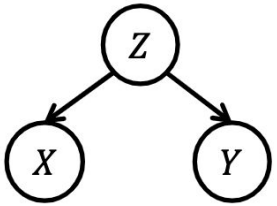
Bernhard Schölkopf
Max Planck Institute for Intelligent Systems, Max-Planck-Ring 4, 72076 Tübingen, Germany
bs@tuebingen.mpg.de

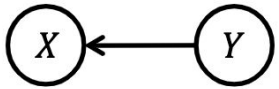
ABSTRACT

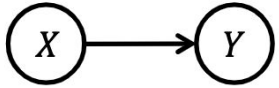
Graphical causal inference as pioneered by Judea Pearl arose from research on artificial intelligence (AI), and for a long time had little connection to the field of machine learning. This article discusses where links have been and should be established, introducing key concepts along the way. It argues that the hard open problems of machine learning and AI are intrinsically related to causality, and explains how the field is beginning to understand them.

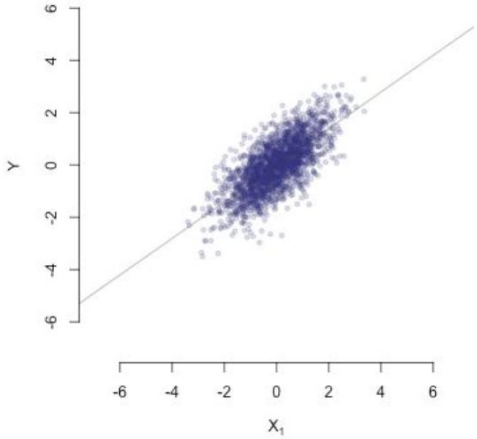
Schölkopf, Bernhard. "Causality for machine learning." *arXiv preprint arXiv:1911.10500* (2019).

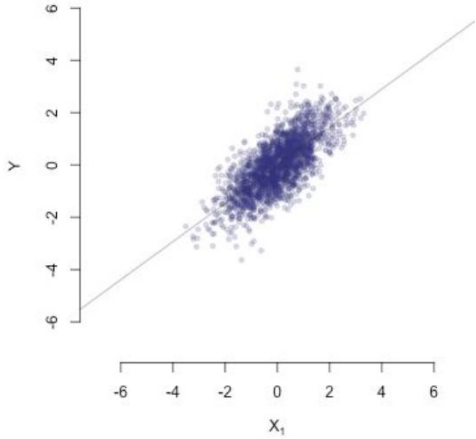
Key problem - Many SCMs generate same distribution











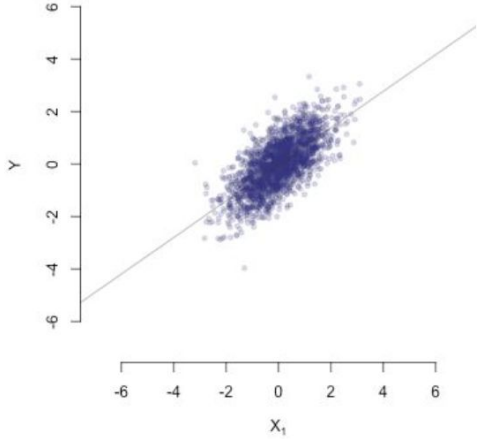
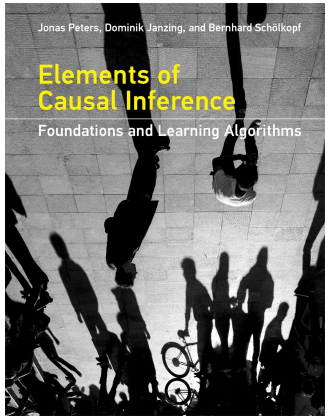


Image from: C. Hein-Deml Causality Lecture ETH 2019
https://stat.ethz.ch/lectures/ss19/causality.php#course_materials

Assumptions that enable Causal Discovery

- Faithfulness
- Independence of Mechanisms
- Additive Noise
- Linear non-Gaussian models
- …



See Chapter 2 in Elements.

Key point: None of these assumptions simply translates to more data will allow to draw causal conclusions.

Connection to modern deep learning: generalization without inductive biases seems very difficult.

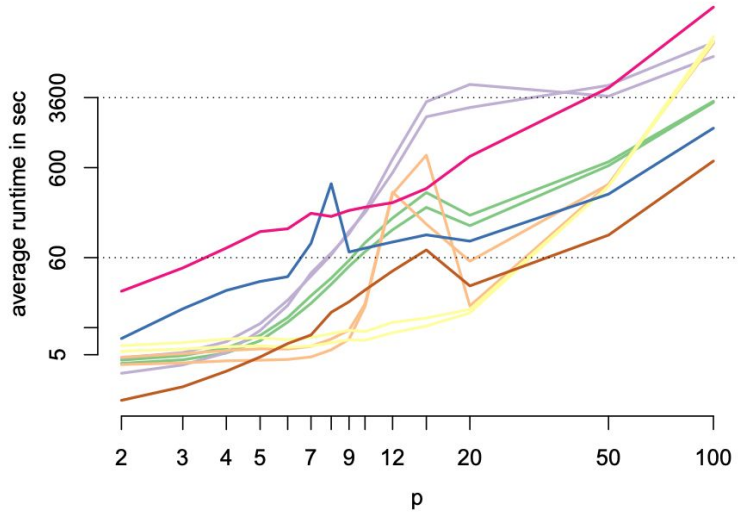
Causal Structure Learning

Main categories

- Constraint based e.g. PC
- Score based e.g. GES
- Restricted SEMs e.g. LiNGAM

Key takeaways:

- Often computational infeasible for large graphs and approximations may result in local optima
- Non-linear case seems very challenging in case of conditional independence testing.
- Require direct observations of variables.



Causal Structure Learning

Christina Heinze-Deml, Marloes H. Maathuis, and Nicolai Meinshausen
Seminar für Statistik, Department of Mathematics
ETH Zurich, Switzerland, CH-8092 Zurich
email: {heinzedeml, maathuis, meinshausen}@stat.math.ethz.ch

June 29, 2017

Christina Heinze-Deml et al.Annual Review of Statistics and Its Application, 2018. https://arxiv.org/abs/1706.09141

Identifiability of linear non-Gaussian models

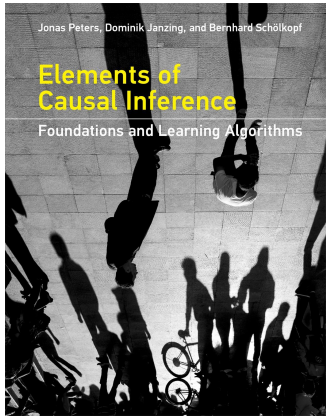
Theorem 4.2 (Identifiability of linear non-Gaussian models) *Assume that $P_{X,Y}$ admits the linear model*

$$Y = \alpha X + N_Y, \quad N_Y \perp\!\!\!\perp X, \tag{4.1}$$

with continuous random variables X, N_Y, and Y. Then there exist $\beta \in \mathbb{R}$ and a random variable N_X such that

$$X = \beta Y + N_X, \quad N_X \perp\!\!\!\perp Y, \tag{4.2}$$

if and only if N_Y and X are Gaussian.



Independent Component Analysis

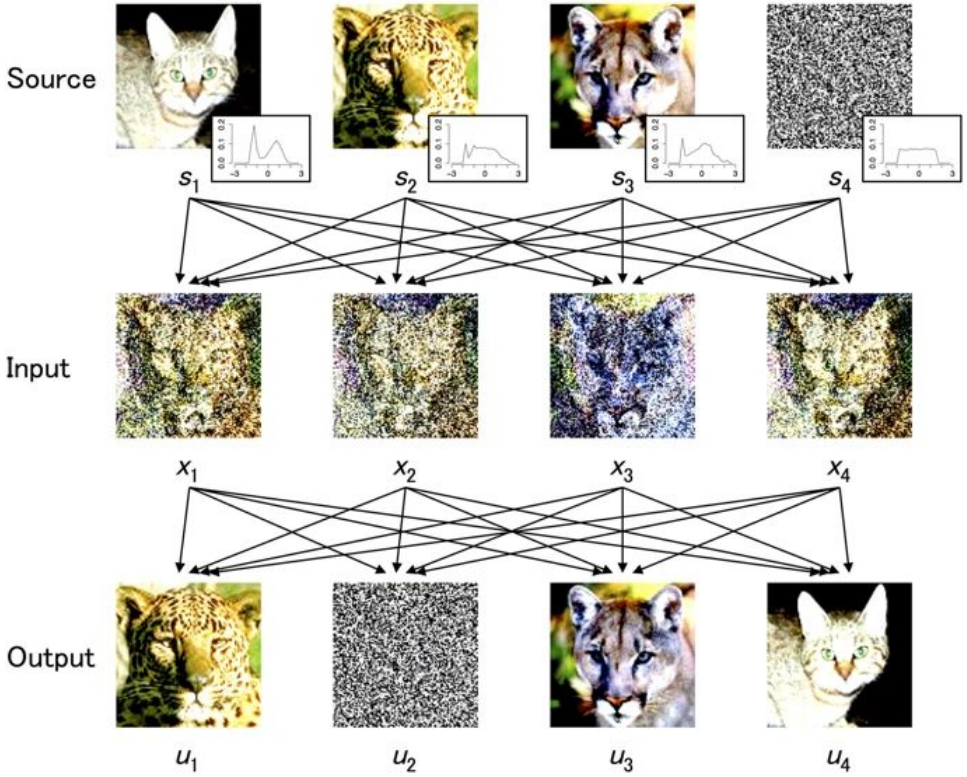


Image from: Isomura, Takuya, and Taro Toyoizumi. "A local learning rule for independent component analysis." *Scientific reports* 6 (2016): 28073.

Model: X=AS

X: observed variables
S: mutually independent, continuous latent, non-Gaussian variables (sources)
A: unobserved full rank mixing matrix

- Here: A is identifiable up to permutation, scaling and sign of the columns
- Without inductive biases, the nonlinear cases is unidentifiable: Hyvärinen, Aapo, and Petteri Pajunen. "Nonlinear independent component analysis: Existence and uniqueness results." *Neural networks* 1999.
- Khemakhem, Ilyes, et al. "Variational autoencoders and nonlinear ica: A unifying framework." *AISTATS*. 2020.

LiNGAM: Linear non-Gaussian acyclic models for causal discovery

$$X = \beta X + \varepsilon, \quad \beta \in \mathbb{R}^{p \times p}, X \in \mathbb{R}^p, \varepsilon \in \mathbb{R}^p$$

- Noise is mean-zero non-Gaussian with positive variance
- Noise components are mutually independent i.e. no hidden confounders
- Acyclicity and permutations imply that beta is strictly lower triangular

$$X = \beta X + \varepsilon$$
$$(I - \beta)X = \varepsilon$$
$$X = (I - \beta)^{-1}\varepsilon$$

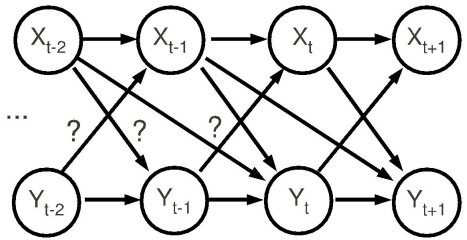
A Linear Non-Gaussian Acyclic Model for Causal Discovery

Shohei Shimizu* SHOHEIS@ISM.AC.JP
Patrik O. Hoyer PATRIK.HOYER@HELSINKI.FI
Aapo Hyvärinen AAPO.HYVARINEN@HELSINKI.FI
Antti Kerminen ANTTI.KERMINEN@HELSINKI.FI
Helsinki Institute for Information Technology, Basic Research Unit
Department of Computer Science
University of Helsinki
FIN-00014, Finland

Structure Learning: Time Series

Time series and Granger causality

Does X cause Y and/or Y cause X?



exclude instantaeous effects and common causes

- if

$$Y_{present} \not\perp\!\!\!\perp X_{past} \,|Y_{past}$$

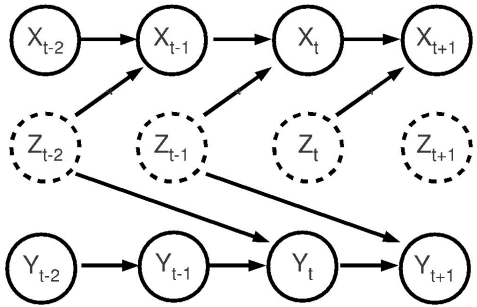
  there must be arrows from X to Y (otherwise d-separation)

- Granger (1969): the past of X helps when predicting Y_t from its past
- strength of causal influence often measured by transfer entropy

$$I(Y_{present}; X_{past} \,|Y_{past})$$

Confounded Granger

Hidden common cause Z relates X and Y



due to different time delays we have

$$Y_{present} \not\perp\!\!\!\perp X_{past} \,|Y_{past}$$

but

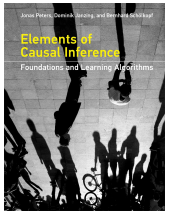
$$X_{present} \perp\!\!\!\perp Y_{past} \,|X_{past}$$

Granger infers $X \rightarrow Y$

Recent work on finding causation in confounded time series: Mastakouri et al., 2020 https://arxiv.org/abs/2005.08543

Intervention Invariance

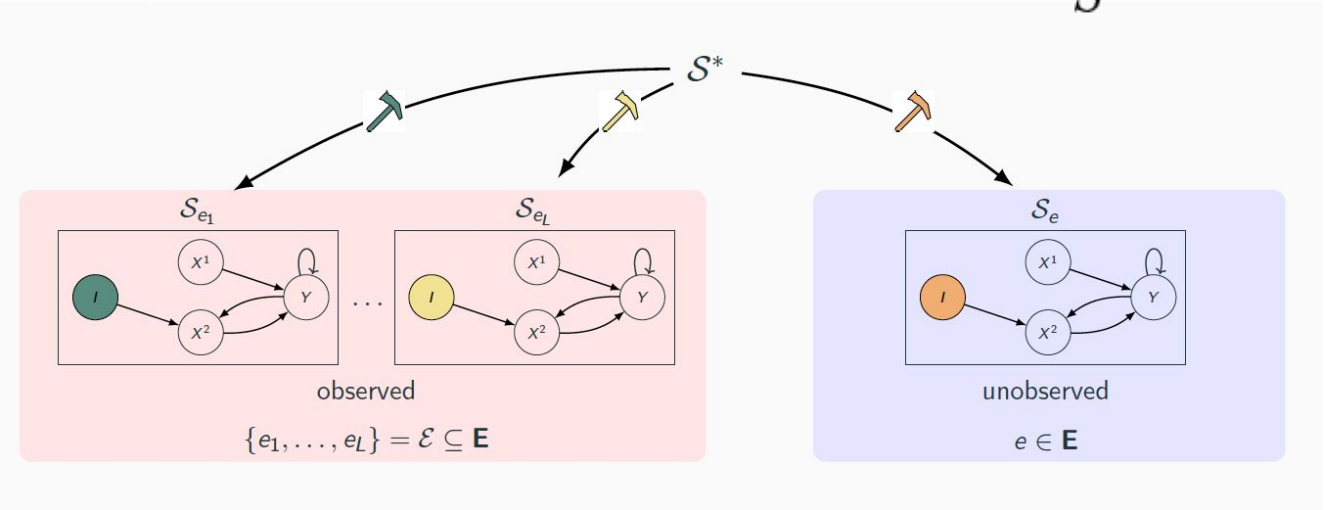
Schölkopf, Bernhard, et al. "On causal and anticausal learning." *ICML* (2012).



J. Peters, P. Bühlmann and N. Meinshausen, **Causal inference using invariant prediction: identification and confidence intervals,** arXiv:1501.01332, Journal of the Royal Statistical Society, Series B, 2016

The conditional mechanism of a target Y remains fixed across interventions.

$$Y^e = X^e \gamma^\star + \varepsilon^e, \quad \varepsilon^e \sim F_\varepsilon \text{ and } \varepsilon \perp\!\!\!\perp X^e_{S^\star}, \forall e \in E$$



SCMs for ODEs & SDEs

A deterministic causal kinetic model, over processes $(\mathbf{x}_t)_t := (x_t^1, \ldots, x_t^d)_t$ is a collection of d ODEs and initial value assignments

$$\begin{aligned}
\tfrac{d}{dt}x_t^1 &:= f^1(x_t^{\mathrm{PA}(1)}, x_t^1), \qquad & x_0^1 &:= \xi_0^1 \\
\tfrac{d}{dt}x_t^2 &:= f^2(x_t^{\mathrm{PA}(2)}, x_t^2), \qquad & x_0^2 &:= \xi_0^2 \\
&\vdots \\
\tfrac{d}{dt}x_t^d &:= f^d(x_t^{\mathrm{PA}(d)}, x_t^d), \qquad & x_0^d &:= \xi_0^d .
\end{aligned}$$

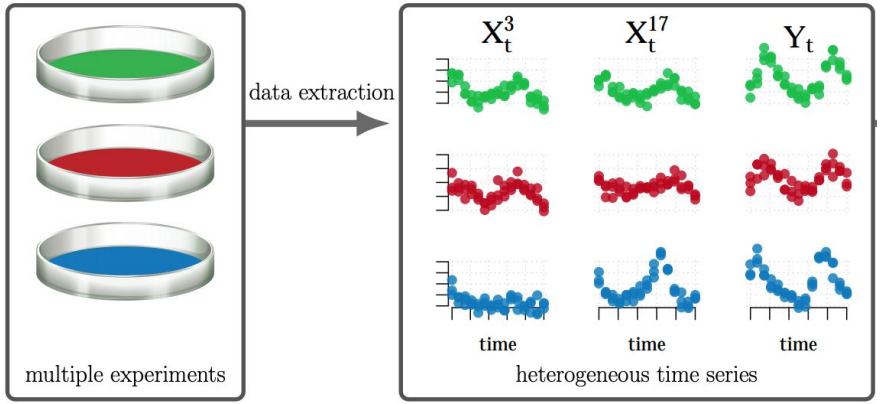
We require that this system is uniquely solvable.

J. Peters, S. Bauer, N. Pfister: Causal models for dynamical systems, to appear, https://arxiv.org/abs/2001.06208

If the network structure is unknown, can we recover it from data?

Goal: Identify system of ODEs $\dot{\mathbf{x}}_t = g_\theta(\mathbf{x}_t, t), \quad \mathbf{x}_0 = \mathbf{x}_0$

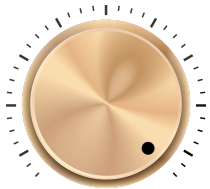
Given: Set of discrete time measurements with measurement noise $\tilde{\mathbf{X}}_t := \mathbf{x}_t + \varepsilon_t$



“Classic Approach”

$$\operatorname{argmin}_\theta \sum_{t \in \{t_1,\ldots,t_m\}} \|\tilde{\mathbf{x}}_t - \hat{\mathbf{x}}_t\|_2^2$$

$$\dot{\mathbf{X}}_t = g_\theta(\mathbf{X}_t, t), \quad \mathbf{X}_0 = \mathbf{x}_0$$

causality  prediction

Causal Approach

utilize causal structure: Invariance across Environments

Haavelmo 1944, Aldrich 1989, Pearl 2009, Peters 2015, …

$$\forall e \in \mathbf{E}: \quad \frac{d}{dt} y_t^e = g_\theta(\mathbf{x}_t^e, t) \quad \text{in } \mathcal{S}_e.$$

causality  prediction

King et al, Nature 2004, Bongard et al, PNAS 2007, Calderhead et al, NIPS 2009, Schmidt et al, Science 2009, Oates, Mukherjee, AoAS 2012, Babtie et al, PNAS 2014, Hill et al, Nature Methods 2016, Chen et al, JASA 2016, Rudy et al, Science Advances 2017, Brunten et al, Nature Comm 2017, Mikkelsen, Hansen, arXiv 1710.09308, many more...

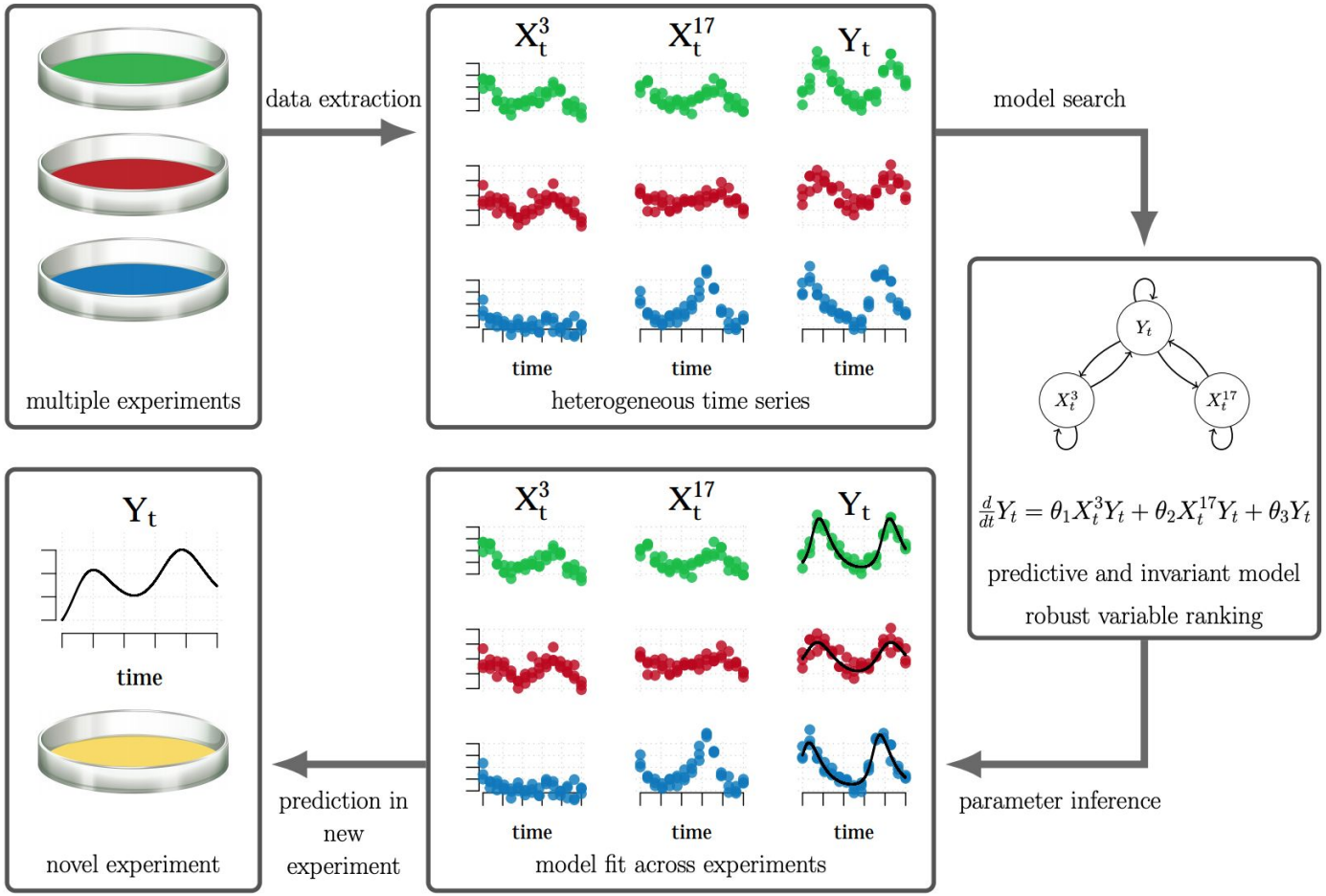
How to measure invariance of an ODE?

For a proposed ODE: $\frac{d}{dt} Y_t = \theta_1 X_t^8$?

1. For each experiment e, smooth target trajectory $\hat{y}^{(e)}$
2. Obtain fitted values for target derivatives. $\xi_{t_1}^{(e)}, \ldots, \xi_{t_m}^{(e)}$
3. Smooth target, with constraints on derivatives $\hat{y}^{(e)}$
4. Score for model ranking where

$$\sum_e \left[\mathrm{RSS}^{(e)} - \mathrm{RSS}^{(e)} \right] / \left[\mathrm{RSS}^{(e)} \right]$$

$$\mathrm{RSS}^{(e)} := \sum_\ell \left(\hat{y}_{t_\ell}^{(e)} - Y_{t_\ell}^{(e)} \right)^2$$



N. Pfister, S. Bauer, J. Peters: Learning stable and predictive structures in kinetic systems, Proceedings of the National Academy of Sciences, 2019

Application to Signalling Pathway

Biological data

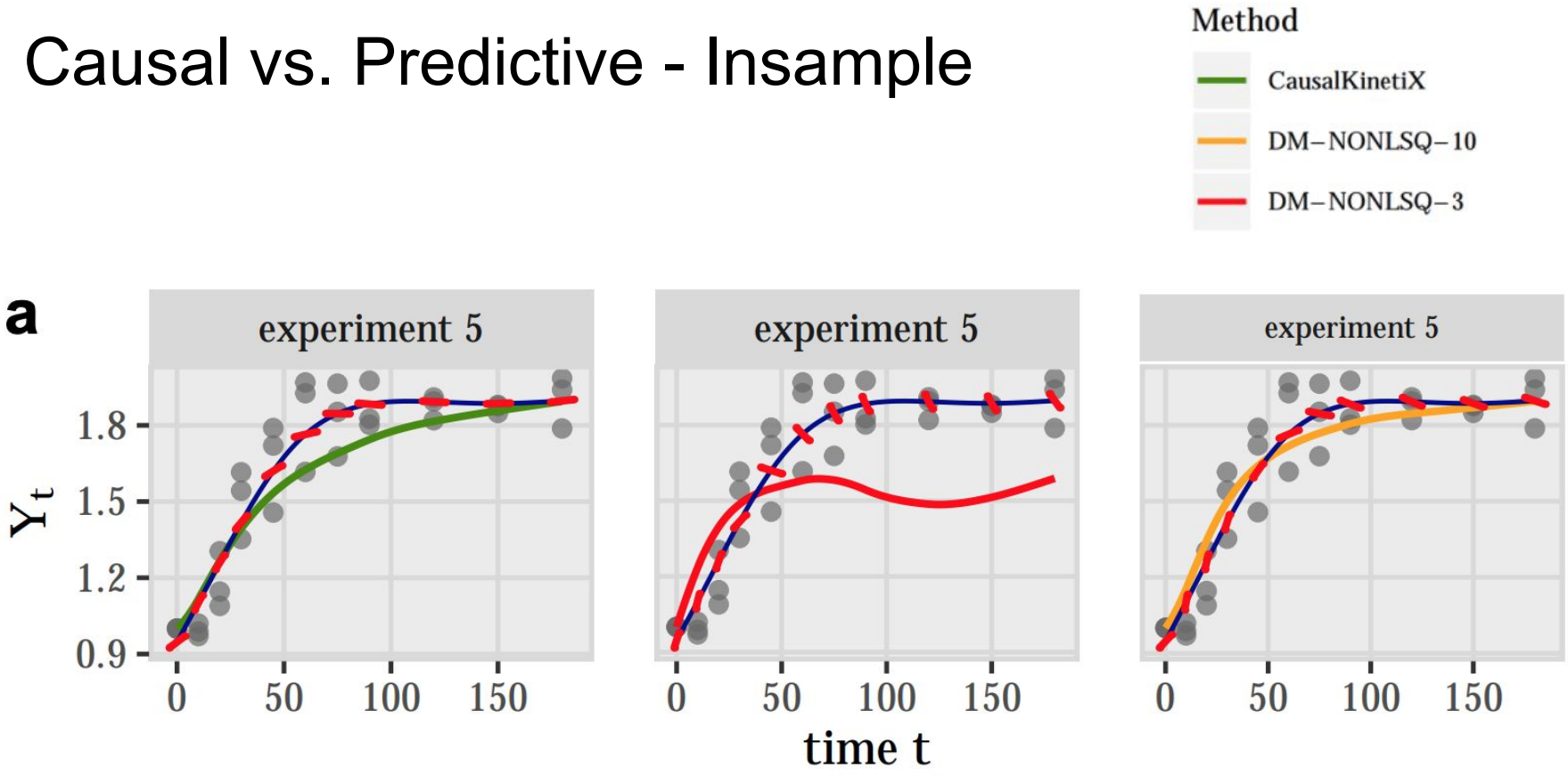
- target variable Y_t (mTOR)
- 411 metabolites $X_t^1, \dots X_t^{411}$
- observations in 5 experimental conditions
- in each experiment, each variables is measured at 11 time Points
- each measurement has 3 technical replicates

From biological knowledge we expect models of the type:

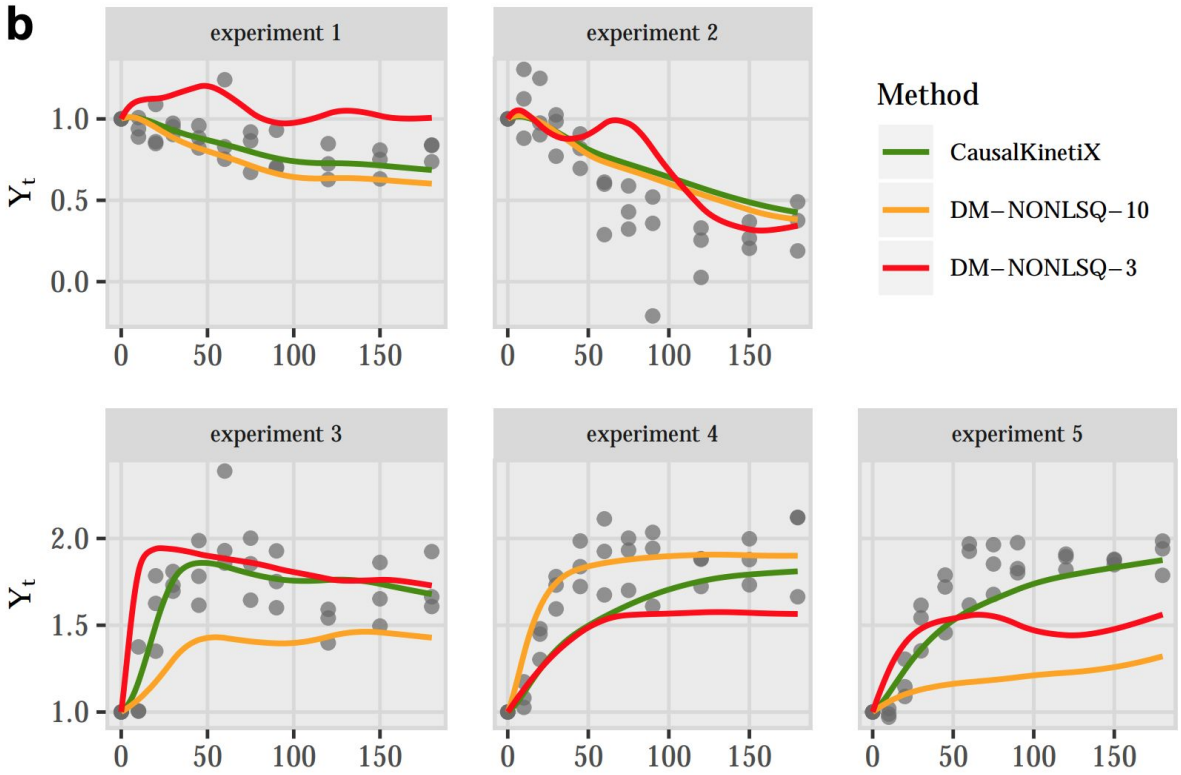
$$\frac{d}{dt}Y_t = \theta_1 X_t^j X_t^k + \theta_2 X_t^p X_t^q + \theta_3 X_t^r X_t^s$$

$$j, k, p, q, r, s \in \{1, \dots, 411\}$$

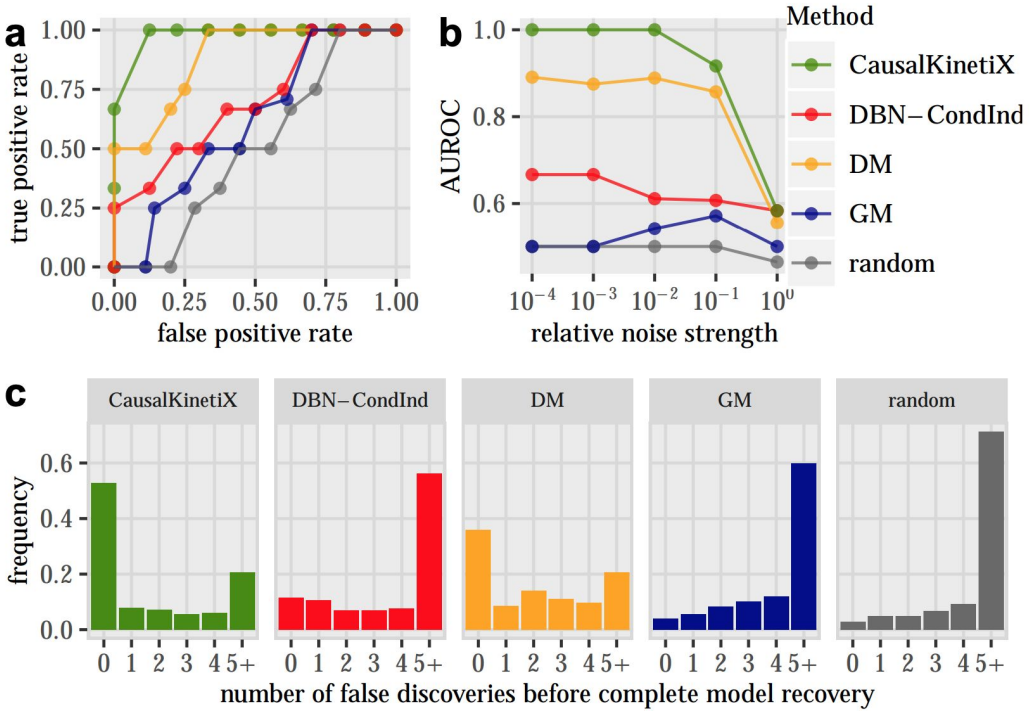
Causal vs. Predictive - Insample



Causal vs. Predictive - Out-of-Sample

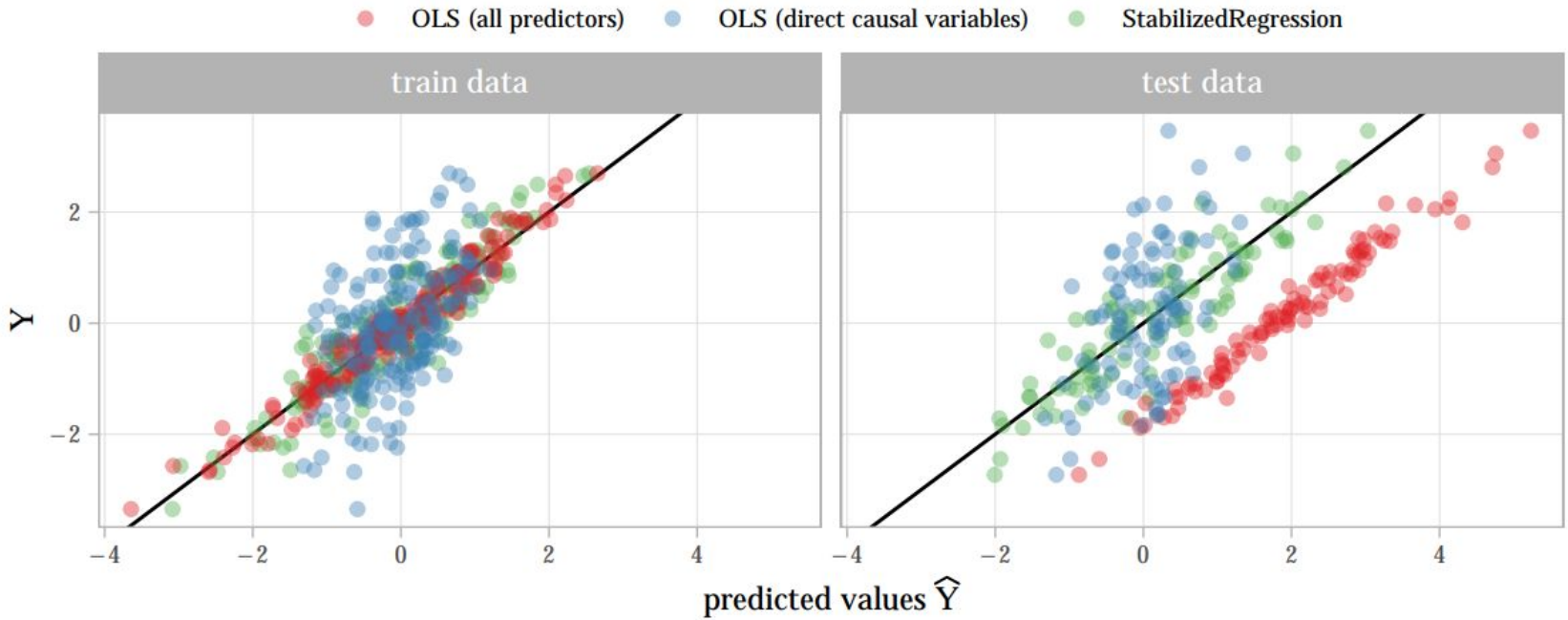


Variable Selection - Rank individual variables based on how often they appear in top ranked model.



Formalized in:
N. Pfister, E. G. William, J. Peters, R. Aebersold, P. Buhlmann: Stabilizing Variable Selection and Regression, **http://arxiv.org/abs/1911.01850**

Stabilized Regression $f_{SR}(X) = \sum_{S \subset \{1,\ldots,d\}} \omega_S f^S(X^S), \quad \sum_S \omega_S = 1$



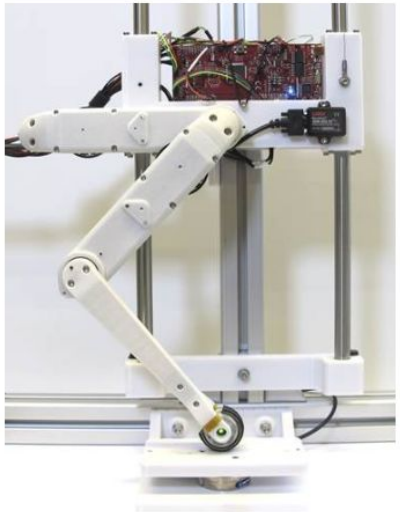
N. Pfister, E. G. William, J. Peters, R. Aebersold, P. Buhlmann: Stabilizing Variable Selection and Regression,
http://arxiv.org/abs/1911.01850

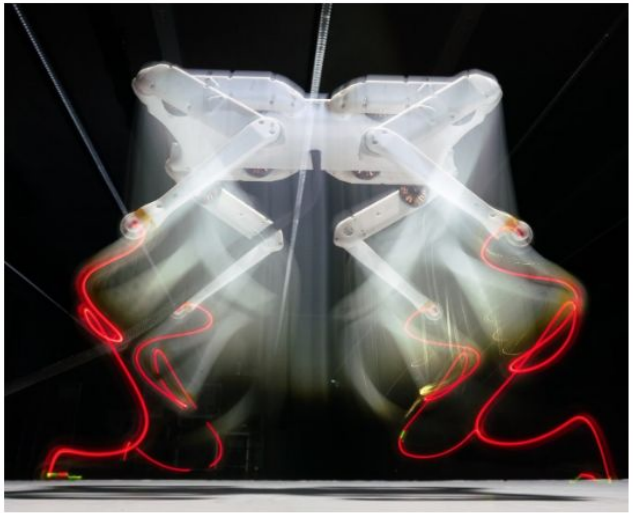
Summary I

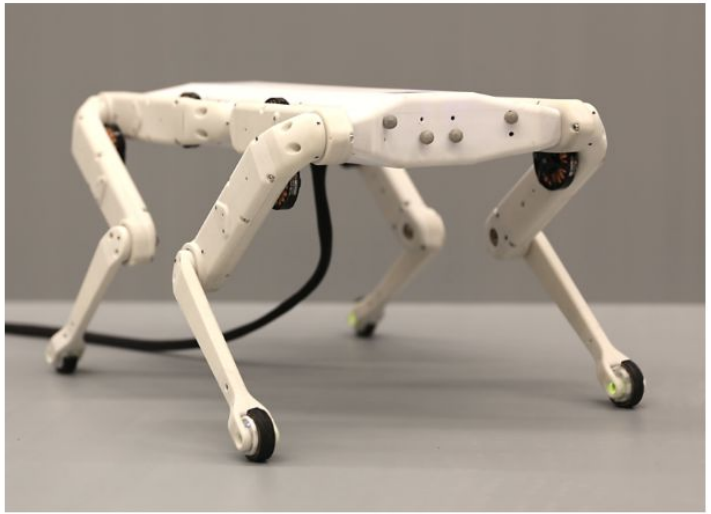
- Causal model -> Invariance but from Invariance -> Causal Model (Peters et al. Invariant Causal Prediction, JRSS B, 2016)
- Benefits of a causal approach / stability for out-of-sample, likewise in:
  - Basu, Sumanta, et al. "Iterative random forests to discover predictive and stable high-order interactions." *Proceedings of the National Academy of Sciences* 115.8 (2018): 1943-1948.
  - Ke, Nan Rosemary, et al. "Learning neural causal models from unknown interventions." *arXiv preprint arXiv:1910.01075* (2019)
- Key problem in all papers: Variables are fully observed and given, validation on real-world data is difficult / not available.
- Connection to ICA for identifying independent components.
- Data is too limited to apply deep learning approaches.

Open Dynamic Robot Initiative

An Open Torque-Controlled Modular Robot Architecture for Legged Locomotion Research





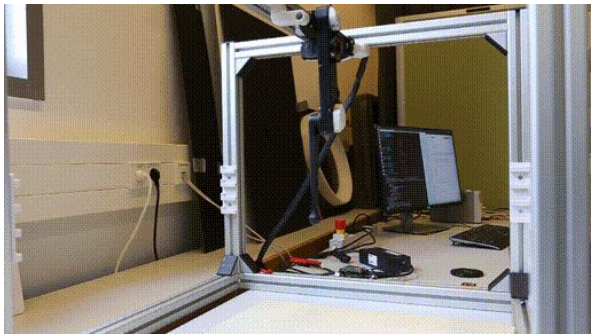


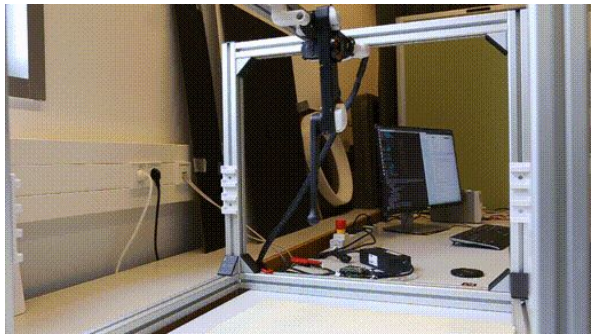
Grimminger F., et al. An Open Torque-Controlled Modular Robot Architecture for Legged Locomotion Research. IROS 2019 https://open-dynamic-robot-initiative.github.io/

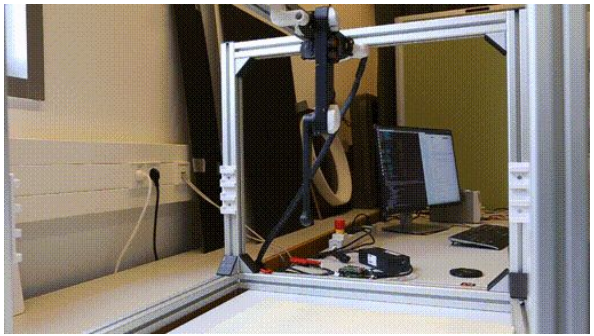
Follow-up: Transferable Dynamics Learning

Agudelo-Espana D., Zadaianchuk A., Wenk P., Garg A., Akpo J., Grimminger F., Viereck J., Naveau M., Righetti L., Martius G., Krause A., Sch□olkopf B., Bauer S., W□uthrich M., A New Robotic Dataset for Transferable Dynamics Learning, ICRA 2020

https://github.com/rr-learning/transferable_dynamics_dataset







A causal perspective on deep representation learning

Causal representation learning

Problem: SCMs are usually at the 'symbolic' level; they assume the causal variables are given.

Goal: embed an SCM into a deep learning model whose inputs and outputs are high-dimensional and unstructured.

Observation: natural connection between SCMs and deep generative models: both use the *reparametrization trick* (Kingma & Welling, 2013), i.e., they make randomness an (exogenous) input to the model rather than an intrinsic component.

Idea: realize the U_i as (latent) noise variables in a generative model.

Given (high-dimensional) $X = (X_1, \ldots, X_d)$ (think of X as an image with pixels $X_1, \ldots, X_d$), construct causal variables $S_1, \ldots, S_n$ ($n \ll d$) and mechanisms

$$S_i := f_i(\mathbf{PA}_i, U_i), \quad (i = 1, \ldots, n)$$

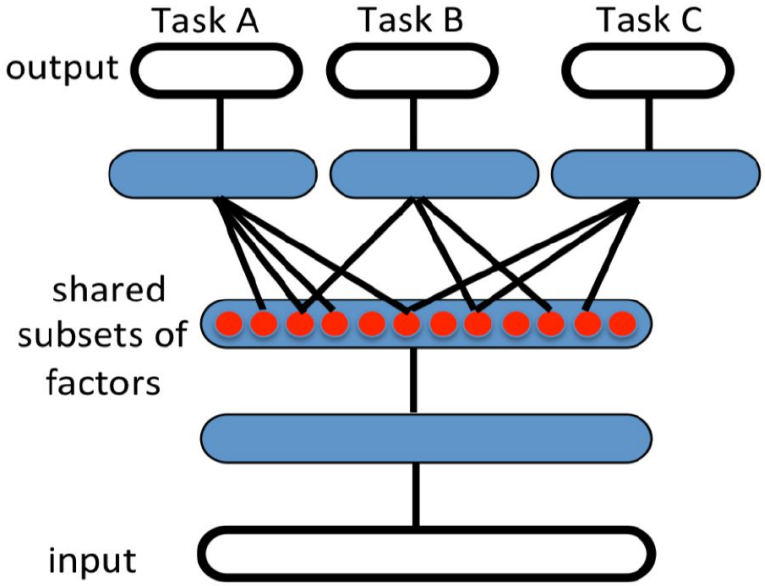
such that we get a **disentangled representation**

$$p(S_1, \ldots, S_n) = \prod_{i=1}^{n} p(S_i \mid \mathbf{PA}_i)$$

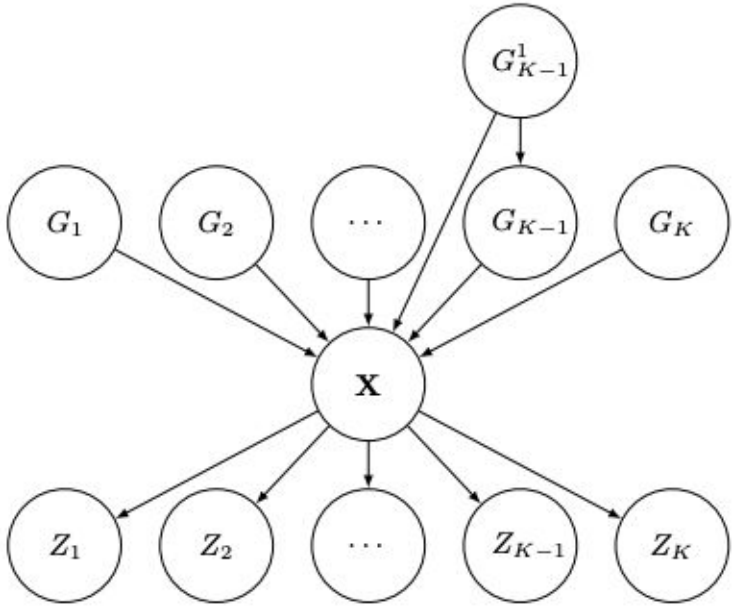
with $p(S_i \mid \mathbf{PA}_i)$ independently manipulable and largely invariant across related problems.

Representation Learning: A Review and New Perspectives

Yoshua Bengio†, Aaron Courville, and Pascal Vincent†
Department of computer science and operations research, U. Montreal
† also, Canadian Institute for Advanced Research (CIFAR)



Causal Framework



Disentangled Generative Factors:

$$\forall g_j^{\triangle} \quad p(g_i|\mathrm{do}(G_j \leftarrow g_j^{\triangle})) = p(g_i) \quad \left(\neq p(g_i|g_j^{\triangle})\right)$$

- **Disentangled (causal) factorization**

$$p(X_1, \ldots, X_n) = \prod_{I=1}^{n} p\left(X_i \mid \mathrm{PA}_i\right)$$

  according to the causal graph: **independent noises and conditionals**

- a change in a distribution always comes from a sparse change in causal conditionals / mechanisms (i.e., structural assignment/function and/or noise variable)
- changing one $p\left(X_i \mid \mathrm{PA}_i\right)$ does not change the other $p\left(X_j \mid \mathrm{PA}_j\right)$ $(j \neq i)$; they remain **invariant**
- **Entangled (non-causal) factorizations**: e.g.,

$$p(X_1, \ldots, X_n) = \prod_{I=1}^{n} p(X_i \mid X_{i+1}, \ldots, X_n).$$

  Here, changes will **not** be local.

Suter, Raphael, et al. "Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness." *ICML*. 2019.

Schölkopf, Bernhard. "Causality for machine learning." *arXiv preprint arXiv:1911.10500* 2019.

Besserve, Michel, et al. "Group invariance principles for causal generative models." *AISTATS*. 2018.

Representation learning

Observations

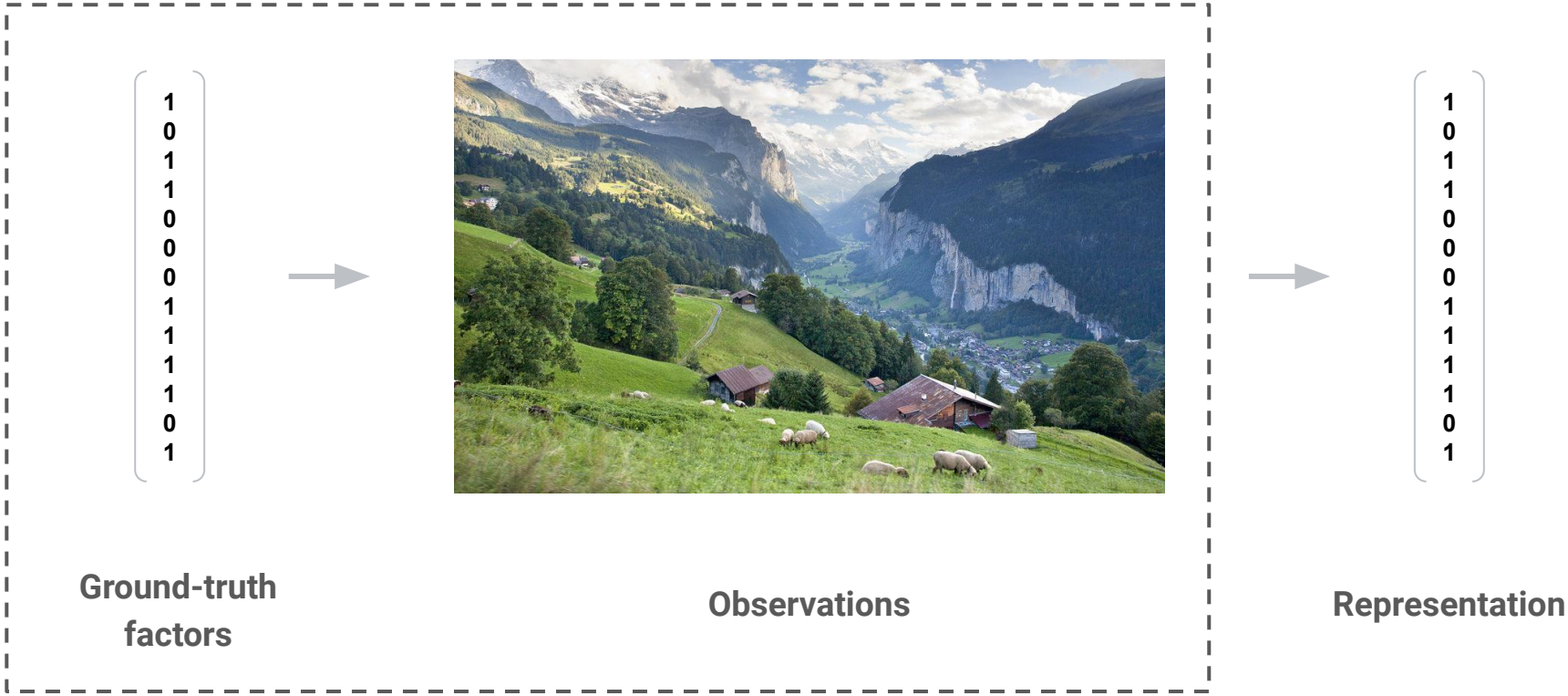
Representation learning

1
0
1
1
0
0
0
1
1
1
1
0
1

Observations

Representation

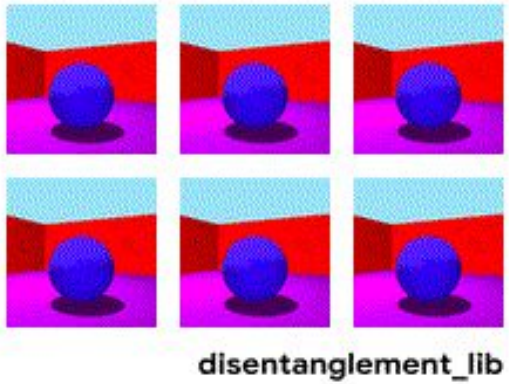
Representation learning



Disentangled representations

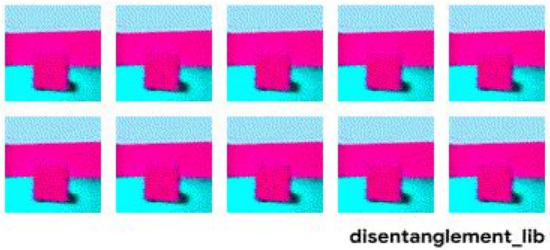
Disentanglement: Single change in factor should lead to single change in representation

What is disentanglement?

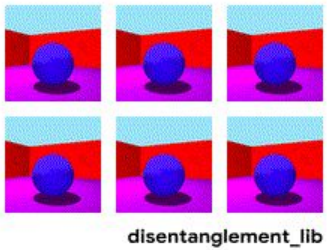


Ground-truth
factors of variation

What is disentanglement?

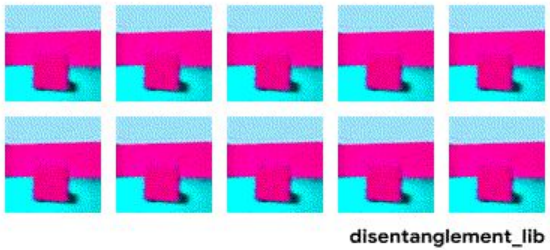


Disentangled model/representation

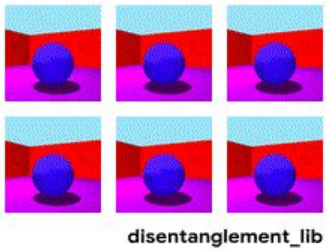


Ground-truth factors

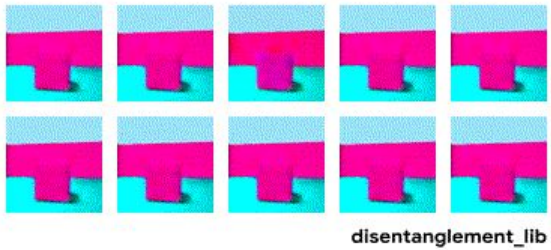
What is disentanglement?



Disentangled model/representation

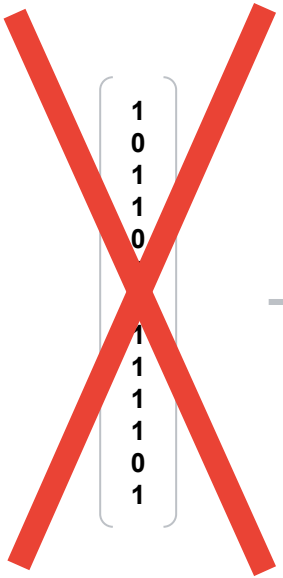


Ground-truth factors

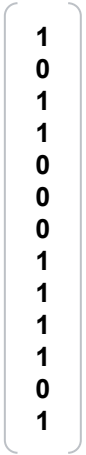


Entangled model/representation

Unsupervised Learning of Disentangled Representations

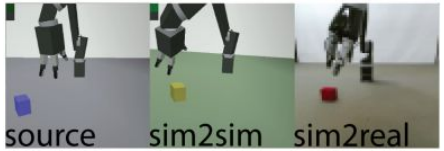


Without access to labels...

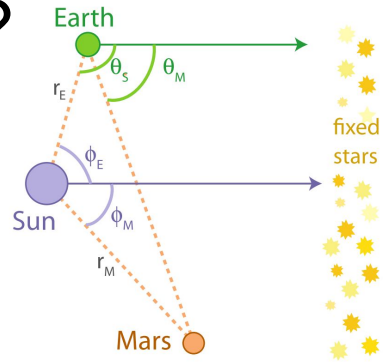


...we aim to find a disentangled representation

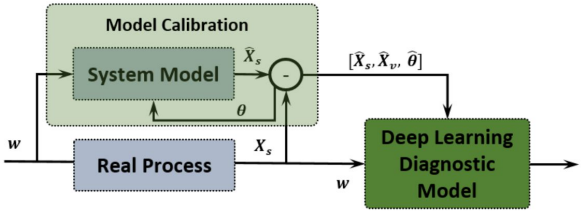
Why Disentanglement?



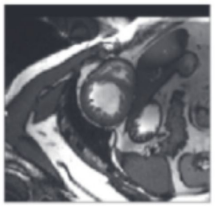
Zero-shot transfer in RL; Higgins and Pal et al., 2018



Discovering Physical Concepts; Iten et al., 2018



Fault Detection; Arias Chao et al., 2019



Cardiac Image Analysis; Chartsias et al., 2019

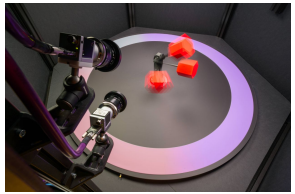

TheUpshot

ROBO RECRUITING

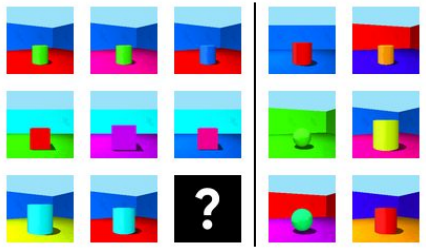
Can an Algorithm Hire Better Than a Human?

By Claire Cain Miller

Fairness; Creager et al., 2019; Locatello et al., 2019

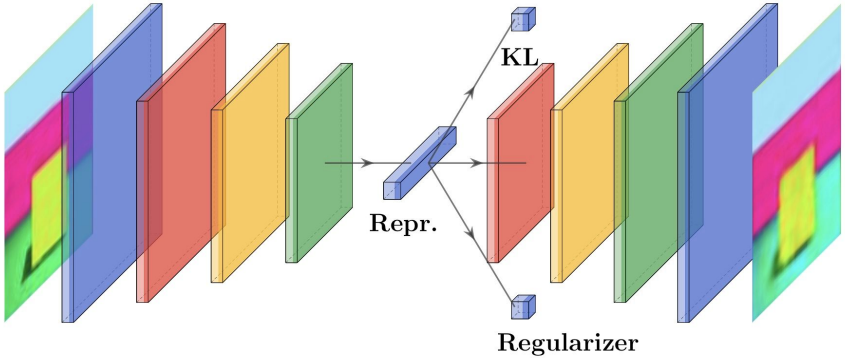


Robotics; Gondal et al., 2019



Abstract Reasoning; Van Steenkiste et al., 2019

Disentanglement methods: VAE + Regularizer



Regularization encourages factorizing representations

arbitrariness: $\hat{\boldsymbol{x}} = D(E(\boldsymbol{x})) = D(f(f^{-1}(E(\boldsymbol{x})))) = \tilde{D}(\tilde{E}(\boldsymbol{x}))$

Disentanglement $\iff$ splitting sources of variation

- Supervised: split known factors from unknowns
- Unsupervised: independence regularization, e.g.:
  - β-VAE: $D_{KL}\left(q_\phi(\boldsymbol{z}|\boldsymbol{x})\|p(\boldsymbol{z})\right)$
  - FactorVAE, β-TCVAE: $\mathrm{TC}(\boldsymbol{z}) = D_{KL}\left(q(\boldsymbol{z})\|\prod_i q(z_i)\right)$
  - DIP-VAE: factorize $q_\phi(\boldsymbol{z}) = \int q_\phi(\boldsymbol{z}|\boldsymbol{x})p(\boldsymbol{x})\,d\boldsymbol{x}$

Beta-VAE, (Higgins et al., 2017);
Fix capacity of VAE bottleneck.

Annealed-VAE, (Burgess et al., 2017);
Progressively increase capacity of VAE bottleneck.

Factor-VAE, (Kim & Mnih, 2018);
Penalize Total Correlation with adversarial training.

Beta-TCVAE, (Chen et al., 2018);
Penalize Total Correlation with Monte Carlo estimate.

DIP-VAE I and **II**, (Kumar et al., 2018)
Match moments with a disentangled prior.

Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations (ICML 2019)

Theoretical result

For arbitrary data, the unsupervised learning of disentangled representations is impossible!

Large-scale experimental study

Can we learn disentangled representations without looking at the labels?

Learning disentangled representations is challenging

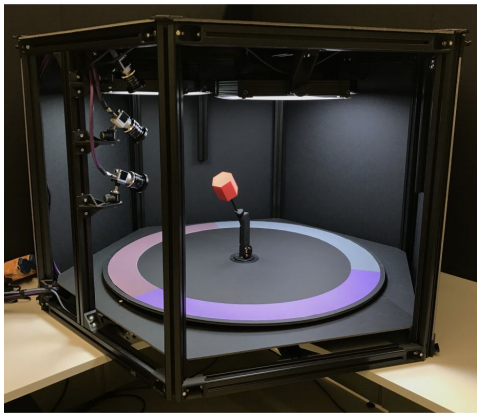
- Random seeds and hyperparameters seem to matter more than the model [1,2]
- Different methods successfully enforce properties “encouraged” by the corresponding losses [3] but
    - unsupervised methods are often NOT unsupervised but still require access to labels for model selection.
    - inductive biases play a key role.
- Official NeurIPS 2019 Challenge: What happens on a real world data?

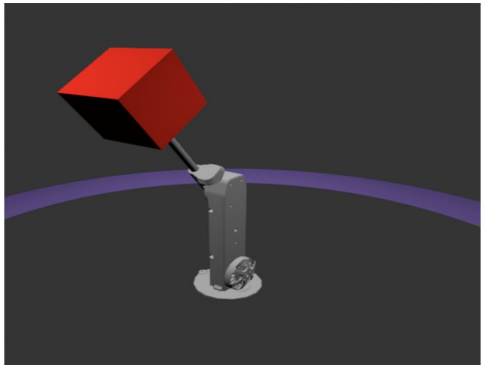
Disentanglement Challenge

Disentanglement Challenge: From Simulation to Real World

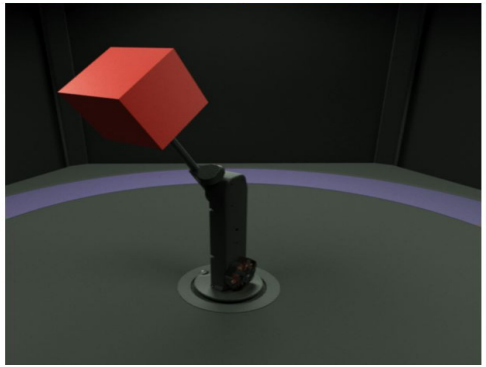
http://www.disentanglement-challenge.com/

- 2 stages
  - Stage 1: Simulated-to-real image transfer learning
  - Stage 2: Transfer to unseen objects
- Low-to-high image resolution (each over 1m images)
- The competition spanned across approx. 3 months
- Approximately 200 participants

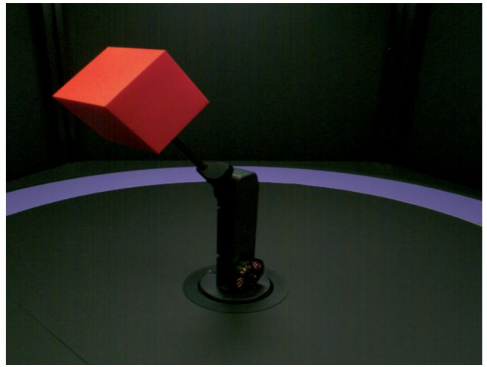




Low-quality rendering



High-quality rendering



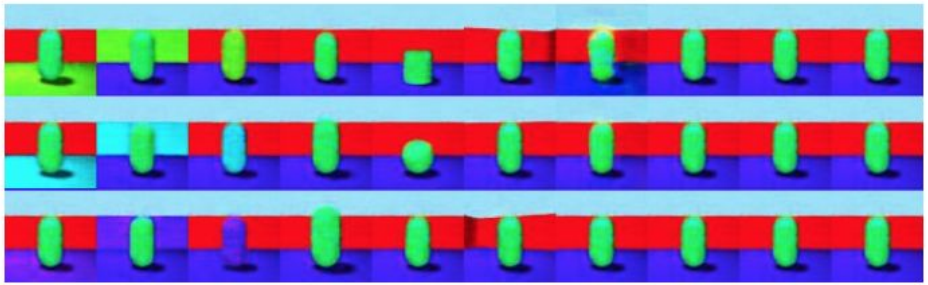
Real-life

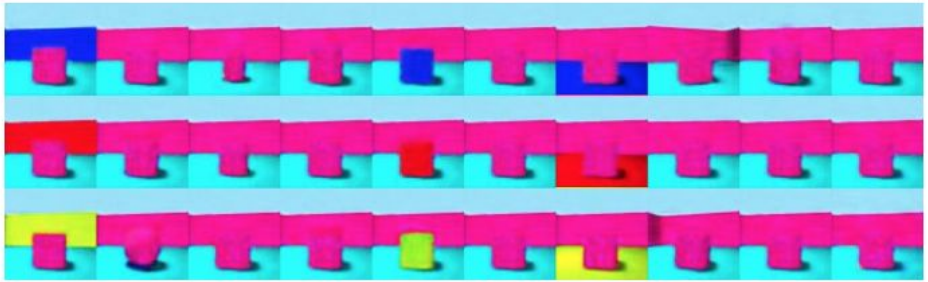
Summary and Open Questions

- Unsupervised setting seems to be impossible (in theory and practice)
  - disentanglement learning with interactions (Thomas et al., 2017)
  - when weak forms of supervision like grouping information are available (Bouchacourt et al., 2018)
  - when temporal structure is available for the learning problem.
- Transfer from simulation to real-world works surprisingly well!
- Generalization and extrapolation to new objects is insufficient.
- **Need setup where we can evaluate learned representations for concrete downstream tasks and with interventions!**

DISENTANGLING FACTORS OF VARIATION USING FEW LABELS

Francesco Locatello[1,2], Michael Tschannen[3], Stefan Bauer[2], Gunnar Rätsch[1], Bernhard Schölkopf[2], Olivier Bachem[3]
[1] Department of Computer Science, ETH Zurich
[2] Max Planck Institute for Intelligent Systems, Tübingen
[3] Google Research, Brain Team
francesco.locatello@inf.ethz.ch, bachem@google.com



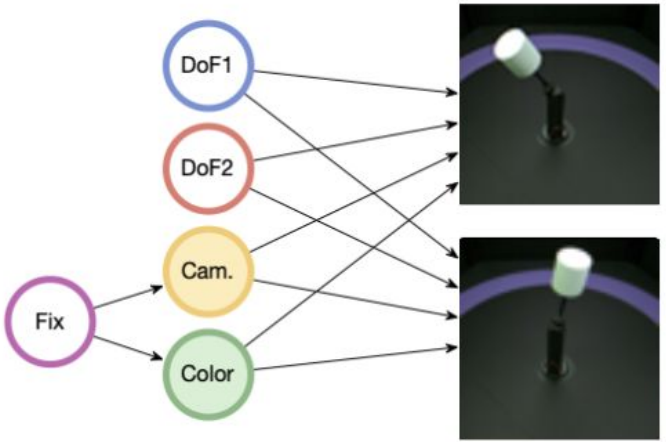


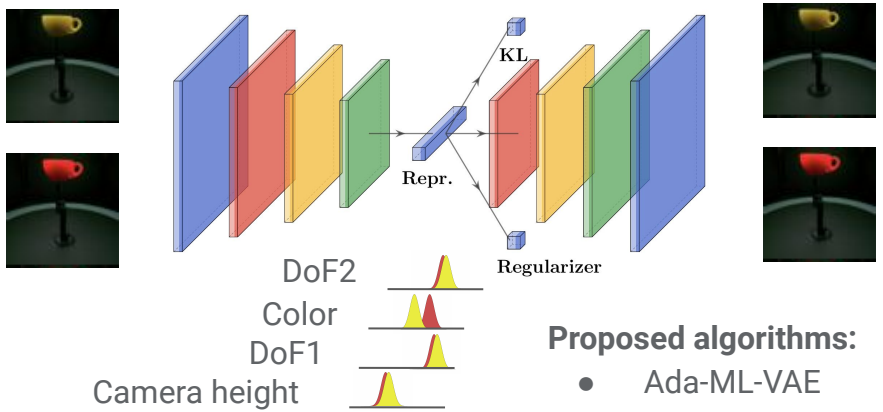
Locatello, Francesco, et al. "Disentangling factors of variation using few labels." *arXiv preprint arXiv:1905.01258* ICLR 2020. .

Weakly-Supervised Disentanglement

F. Locatello, B. Poole, G. Rätsch, B.Schölkopf, O. Bachem, M. Tschannen

arXiv:2002.02886 (ICML'20)





Proposed algorithms:

- Ada-ML-VAE
- Ada-GVAE

Theoretical result

Under some assumptions, disentangled representations can be learned and identified from weak supervision.

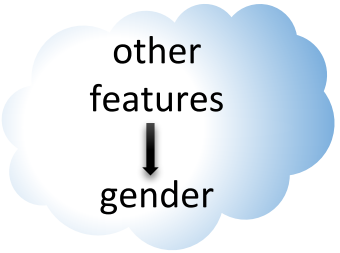
Practical aspects

Model selection based on the weakly-supervised loss, without relying on supervised disentanglement metrics

Fairness

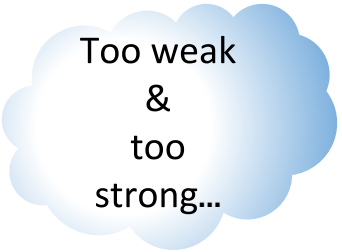
Causality and fairness

"unawareness"



"demographic parity"

$\text{decision} \perp\!\!\!\perp \text{gender}$



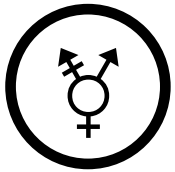
"Equalized odds/equal opportunity"

$\text{decision} \perp\!\!\!\perp \text{gender} \mid \text{suitability}$

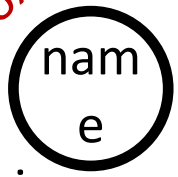
Dwork et al. 2011, Hardt et al. 2016, Berk et al. 2017, Kleinberg et al. 2016, Kamiran, Calders, …

Removing proxy discrimination

N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, D. Janzing, M. Hardt, B. Schölkopf, NIPS 2017



proxy



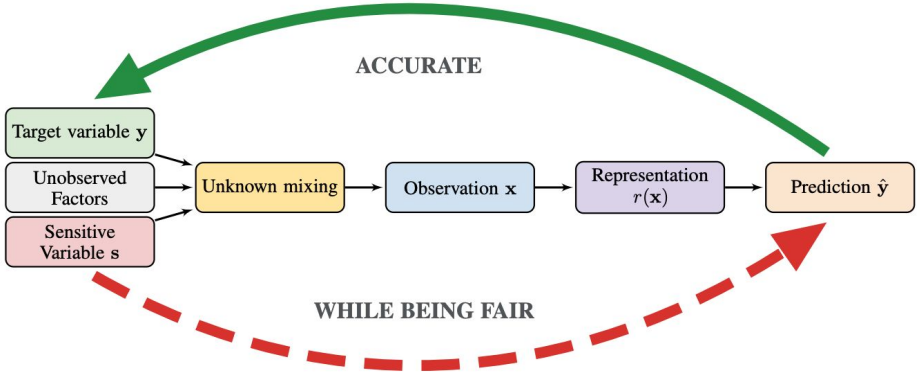
control this to cancel "unfair" paths
(if the influence of proxy on skills is additive & linear)

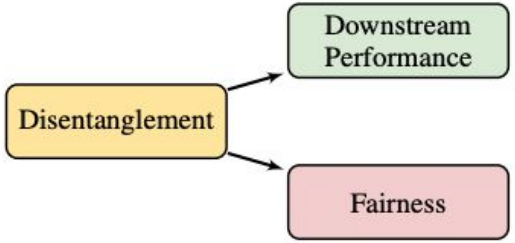
Criterion:

$$\mathbb{P}(\text{decision} \mid do(\text{proxy} = p_1)) =$$
$$\mathbb{P}(\text{decision} \mid do(\text{proxy} = p_2))$$

Pearl 2009, VanderWeele & Robinson 2014,
Bonchi et al. 2015, Qureshi et al. 2016, Kusner et al. 2017, Zhang & Wu 2017, Nabi & Shpitser 2017

Are structured representations helpful for fairness?





$$\texttt{unfairness}(\hat{\mathbf{y}}) = \frac{1}{|S|}\sum_{s} TV(p(\hat{\mathbf{y}}), p(\hat{\mathbf{y}} \mid \mathbf{s} = s))\ \forall\, y$$

Locatello, Francesco, et al. "On the fairness of disentangled representations." NeurIPS 2019.

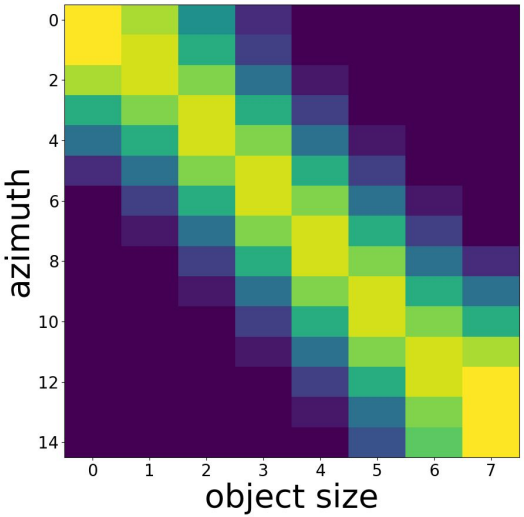
Implications of Correlations

Towards disentangled representations in real-world environments

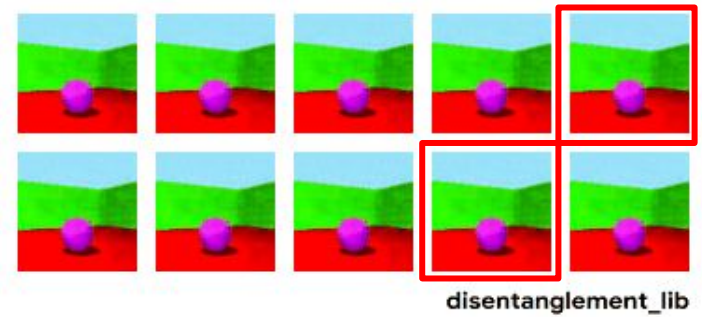
Existing disentanglement methods mostly studied within highly controlled and unrealistic settings (e.g. synthetic datasets and perfectly independent factors of variation)

What happens to disentangled representations in more realistic environments?

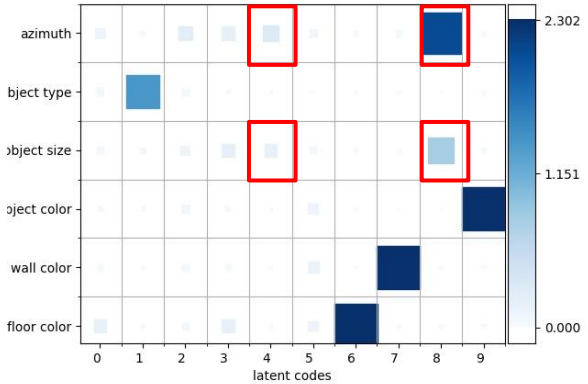
Disentangling correlated factors is nontrivial



Model with highest DCI score
and line width 20%



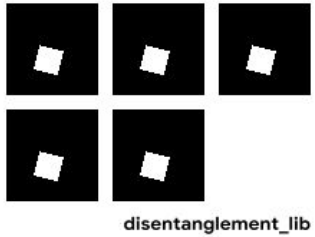
Mutual information

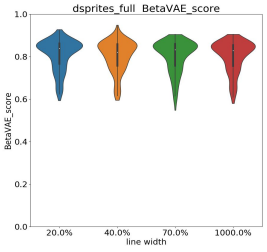


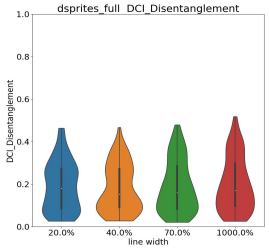
object size and azimuth has been correlated in shapes3d

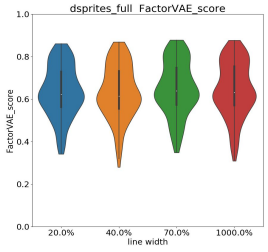
Träuble, Frederik, et al. "Is Independence all you need? On the Generalization of Representations Learned from Correlated Data." *arXiv preprint arXiv:2006.07886* (2020).

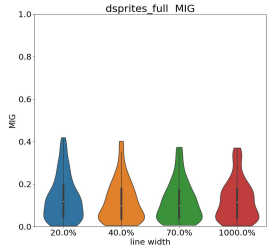
Disentanglement metrics not affected by correlated factors

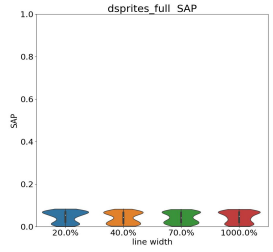


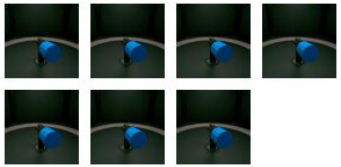


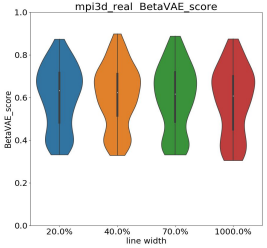


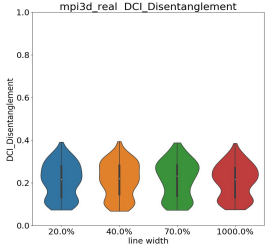


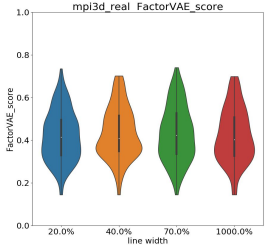


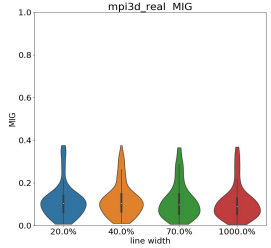


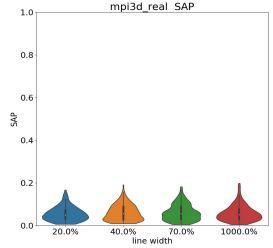


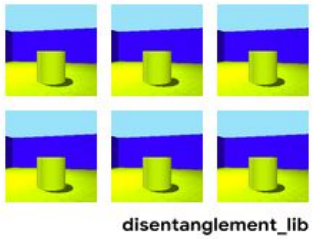


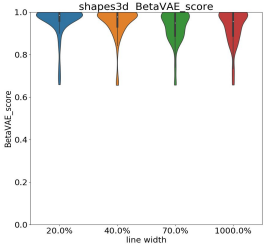


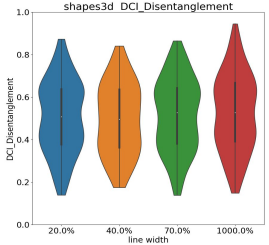


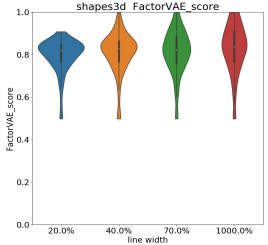


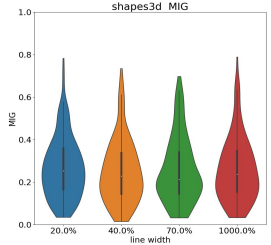


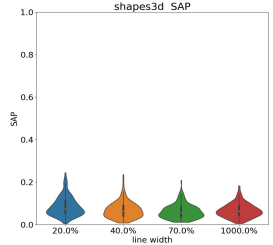




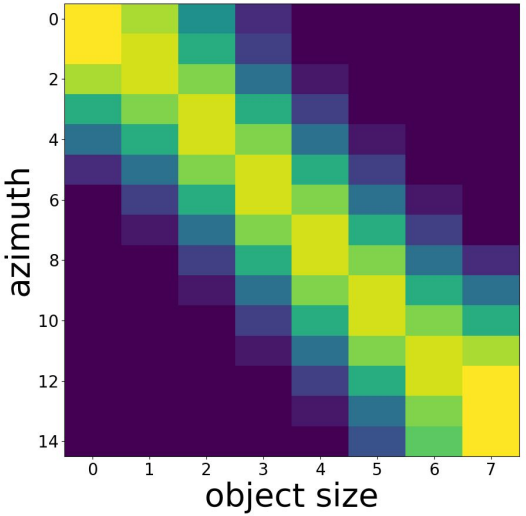




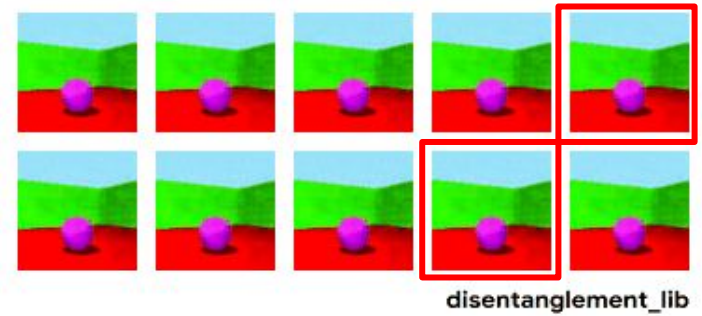




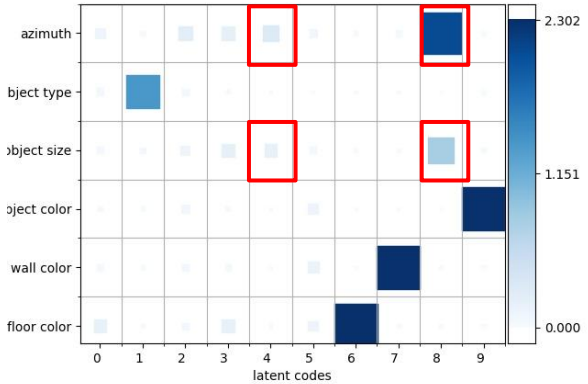
Disentangling correlated factors gets difficult for strong correlation



Model with highest DCI score and line width 20%

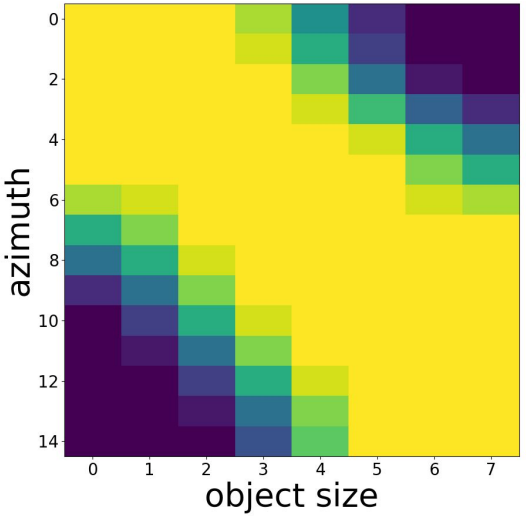


Mutual information

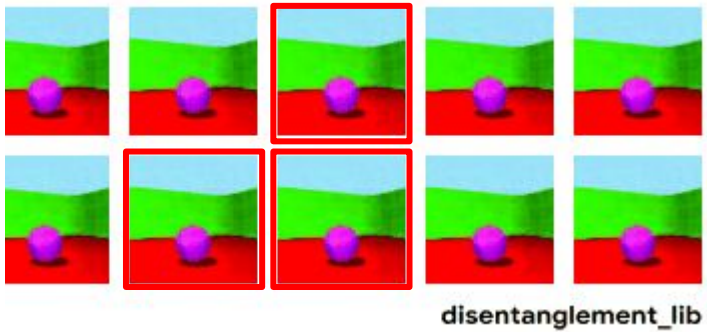


object size and azimuth has been correlated in shapes3d

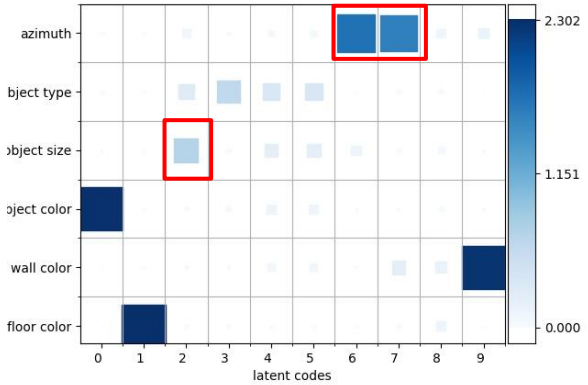
Disentangling correlated factors gets easier for weak correlation



Model with highest DCI score and line width 70%

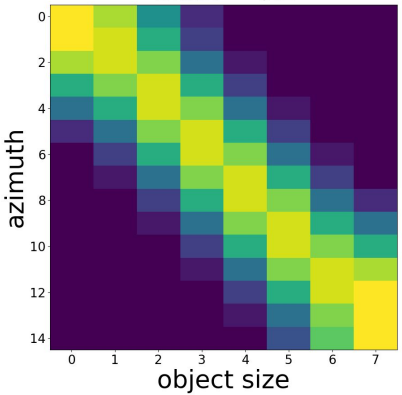


Mutual information

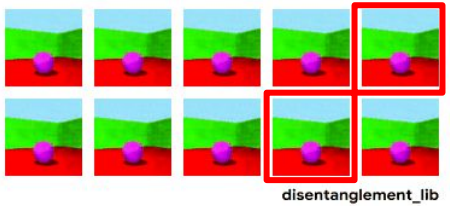


object size and azimuth has been correlated in shapes3d

What happens for the example model? Reversing entanglement with few labels!



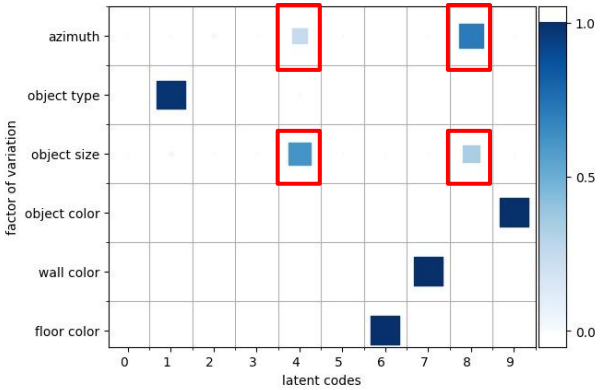
Model with highest DCI score and line width 20%



object size and azimuth has been correlated in shapes3d

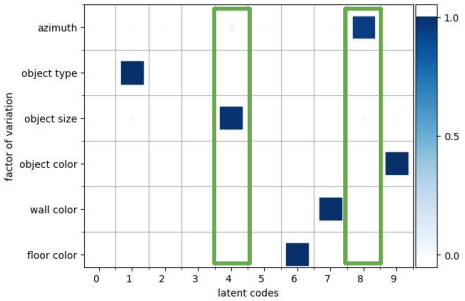
DCI: 0.87

0 labels used for fast adaptation



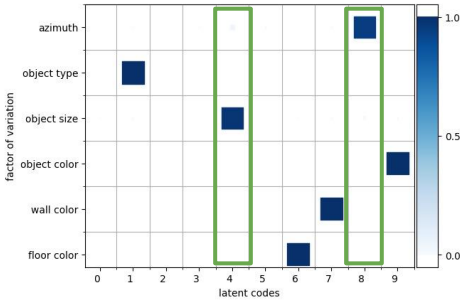
Locatello, Francesco, et al. "Disentangling factors of variation using few labels." *arXiv preprint arXiv:1905.01258* ICLR 2020

50 labels



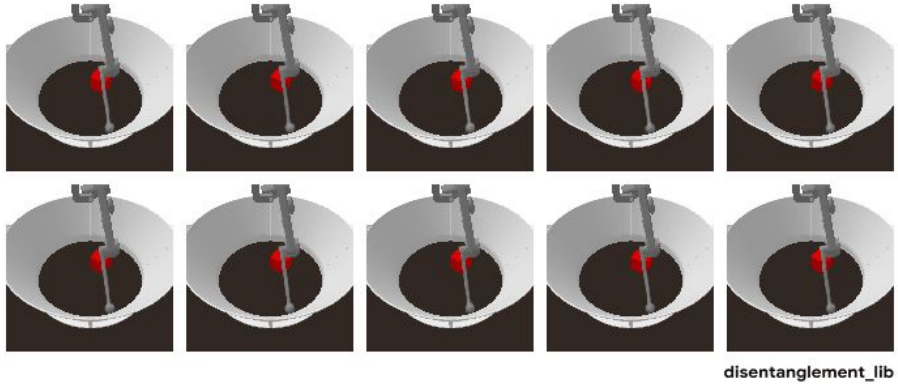
DCI: 0.97

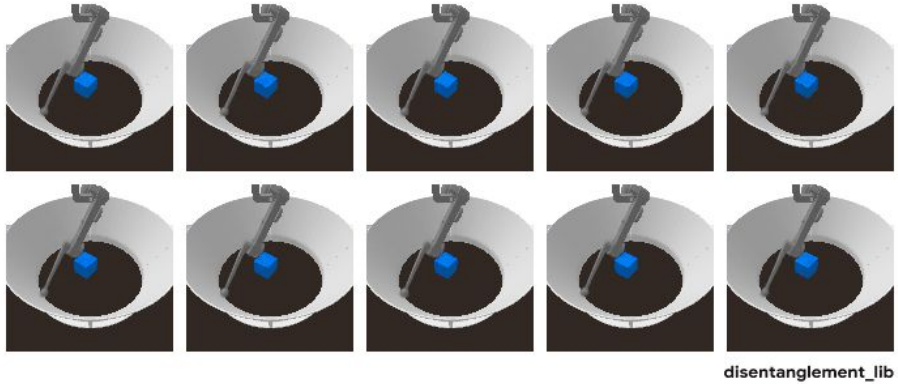
100 labels

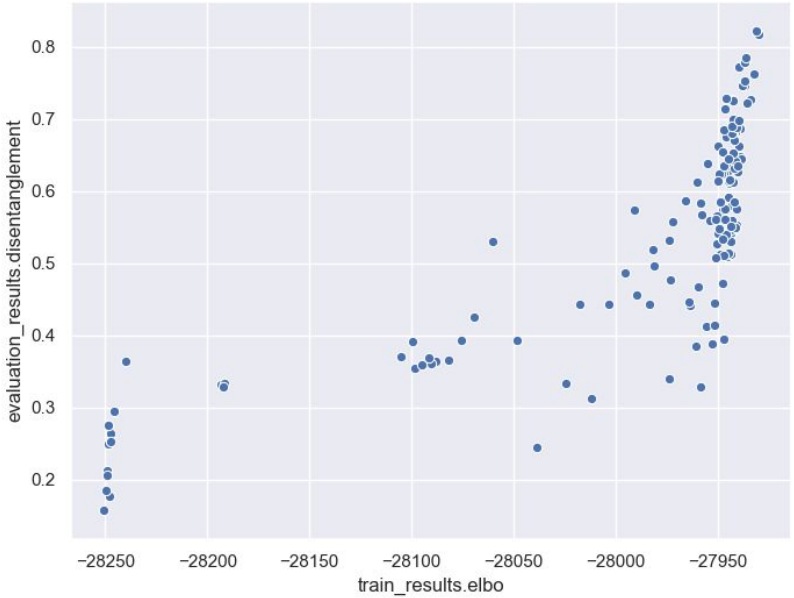


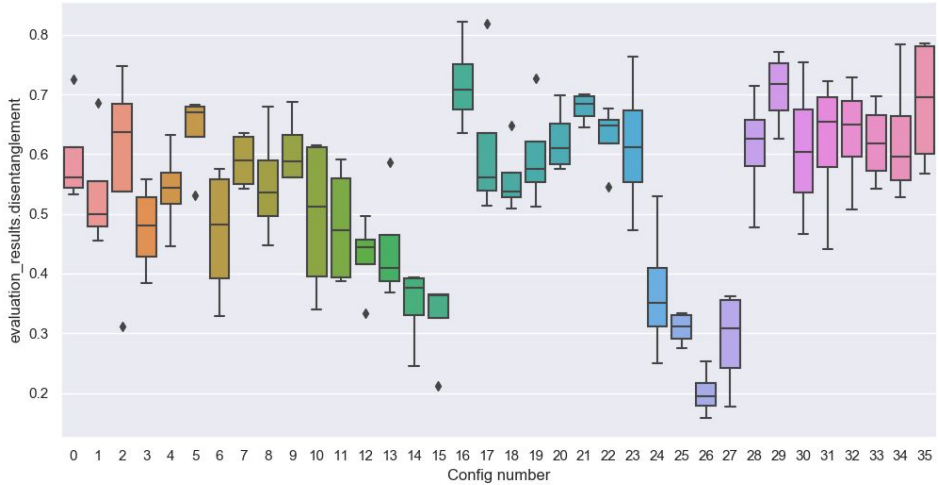
DCI: 0.97

Can we scale and transfer disentangled representations to the real-world?


disentanglement_lib


disentanglement_lib


evaluation_results.disentanglement
0.8
0.7
0.6
0.5
0.4
0.3
0.2
−28250
−28200
−28150
−28100
−28050
−28000
−27950
train_results.elbo


evaluation_results.disentanglement
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35
Config number

Structure by Architecture

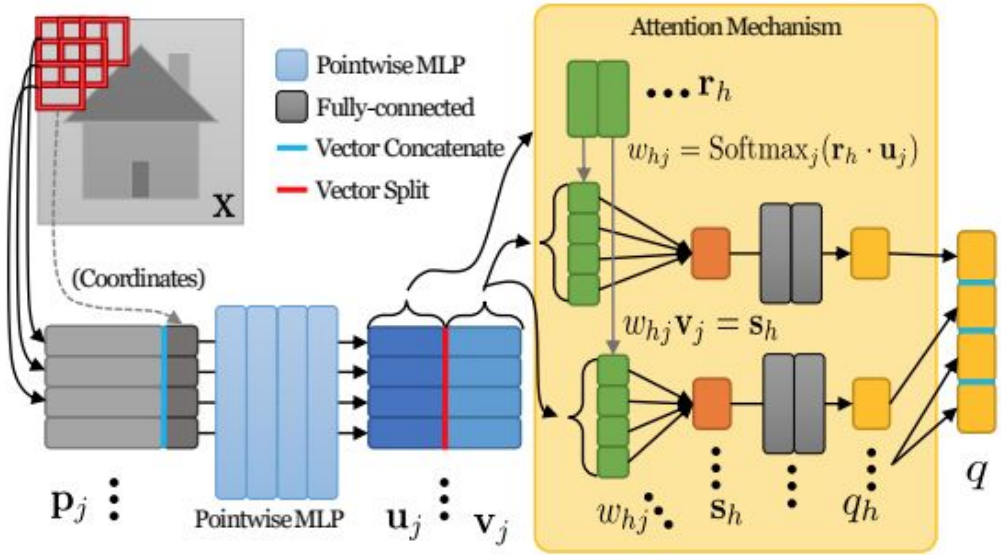
Encoding Causal Structure

Assume that the generative process that produces the observations is relatively low dimensional → **underlying (causal) mechanisms**

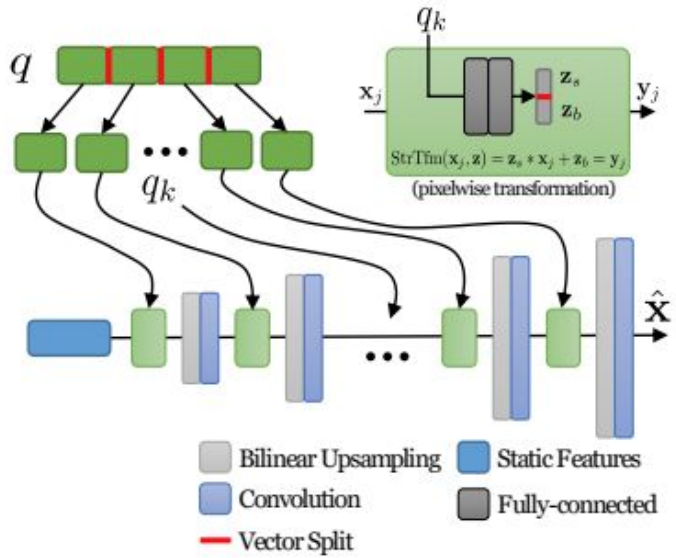
Although the true causal structure is unknown, by designing decoder to **resemble a general SCM** and training it to match the observations, it should become a useful generative model.

Structural Causal Autoencoders

Leeb, Felix, et al. "Structural Autoencoders Improve Representations for Generation and Transfer." *arXiv preprint arXiv:2006.07796* (2020).

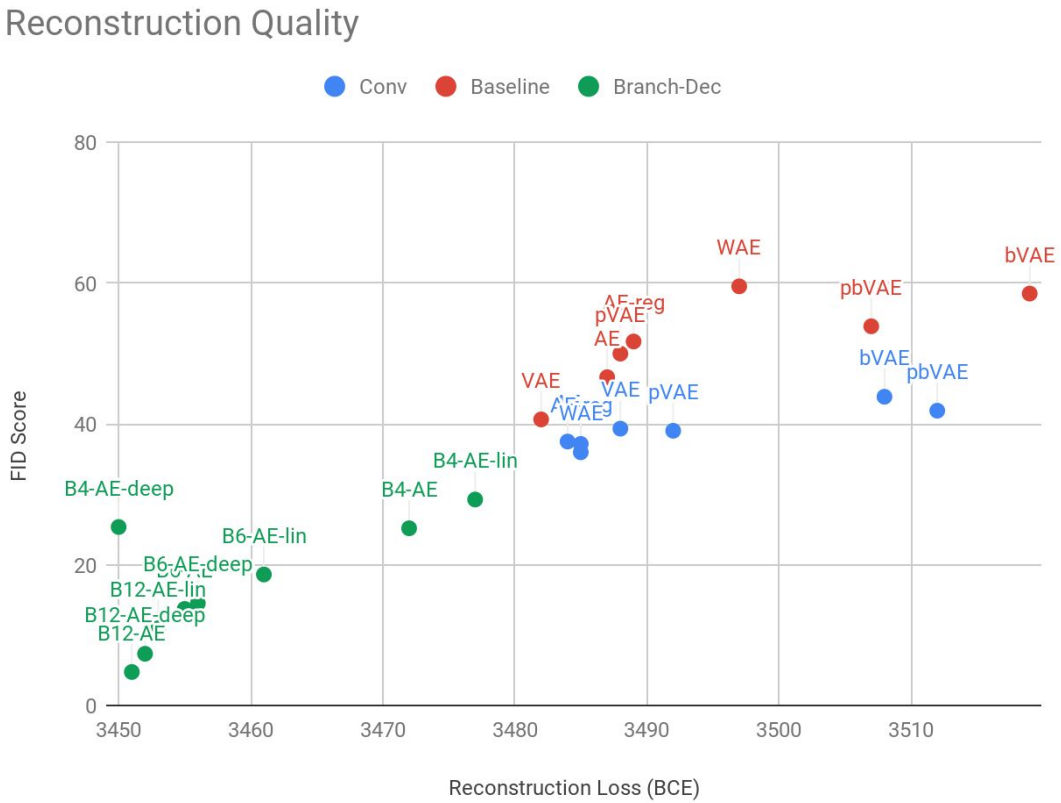


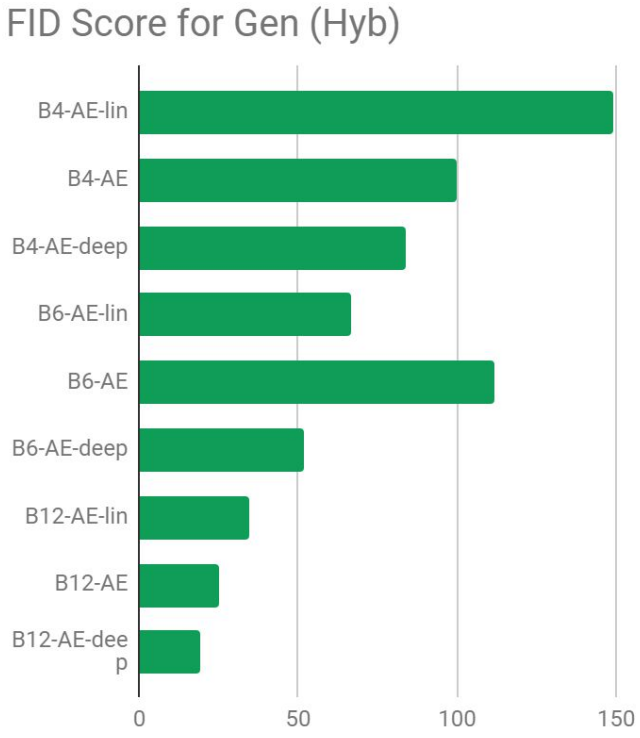
(a) Attention Encoder



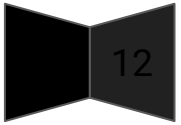
(b) Structural Decoder

Quantitative Results

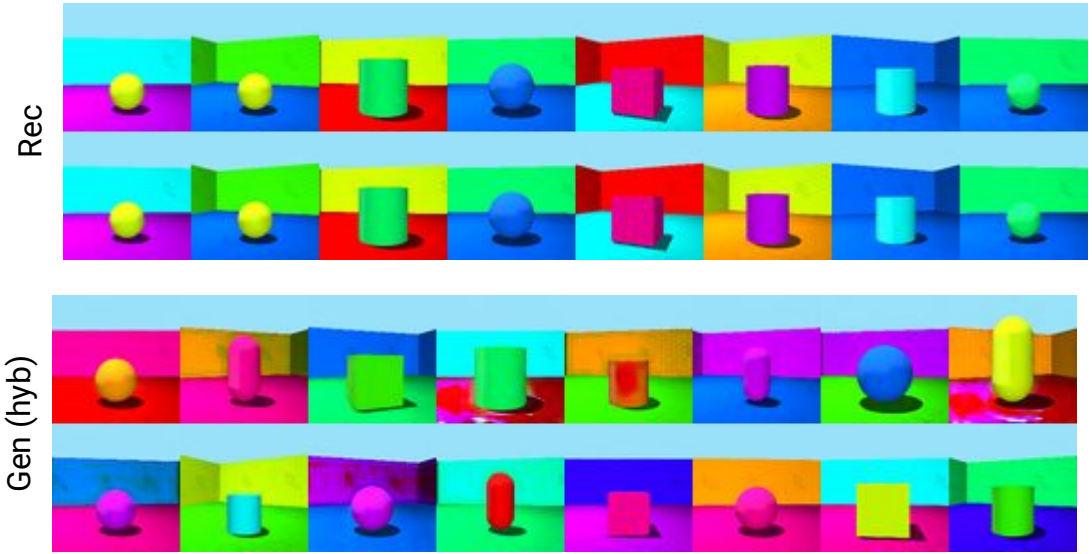


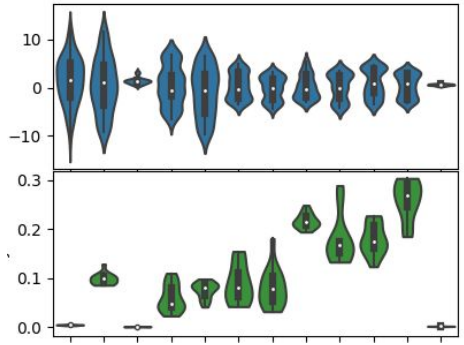


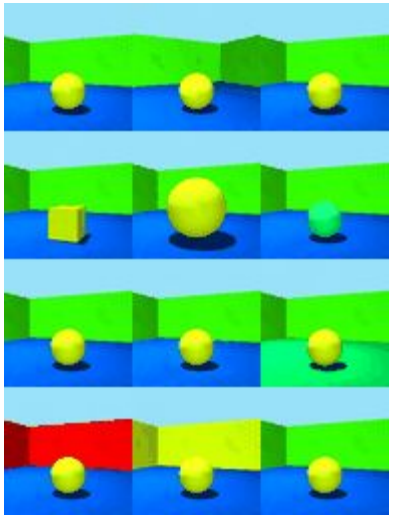
Disentanglement by Architecture



12-Branched AE







Key Insights

- By structuring the **architecture** of the encoder and decoder, the structure of the underlying mechanisms can be discovered (at least for 3D-Shapes - which is definitely a toy dataset)
- The Structural Causal Decoder learns to **distinguish and order mechanisms** from more abstract higher level features (orientation, shape, size) to the low level ones (object color, background colors)
- All evaluations are only done on simple synthetic images where we have access to the groundtruth! Difficult to evaluate on CelebA or more complex datatsets.

Outlook: Towards Causal World Models

Towards causal world models

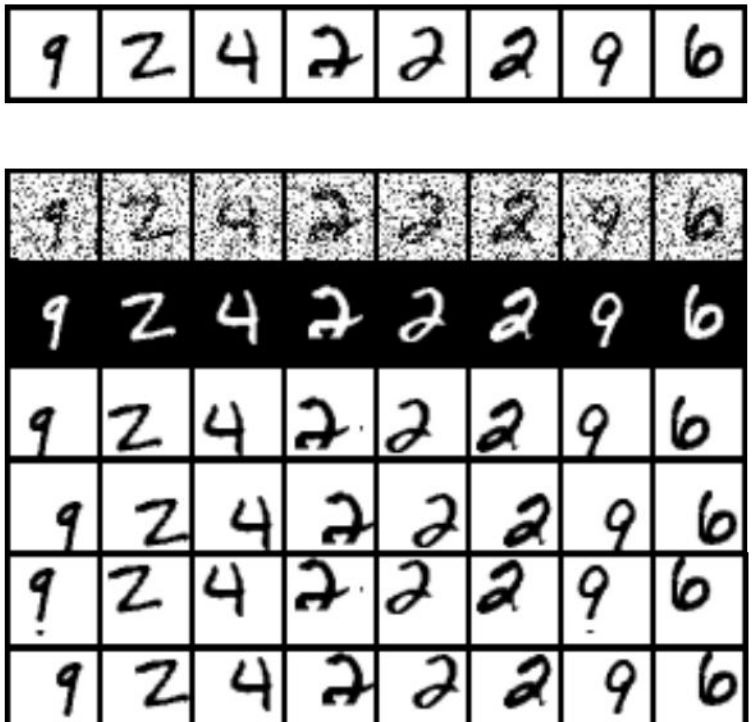
$$p(X_1, \ldots, X_n) = \prod_{I=1}^{n} p\left(X_i \mid \mathrm{PA}_i\right)$$

- learn a multi-task/environment model
  - data from multiple tasks in multiple environments
  - re-usable components that are robust across tasks, i.e., causal (independent) mechanisms
  - representation learning should move towards representations of causal world models: *"thinking is acting is an imagined space"* (Konrad Lorenz)

Learning independent mechanisms

(Parascandolo, Kilbertus, Rojas-Carulla, Schölkopf ICML 2018)

- Data drawn from $p(x)$, transformed by M mechanisms $f_1, \ldots, f_M$
- Goal: learn the independent mechanisms / factors of variation
- Method: generative model with competing mechanisms

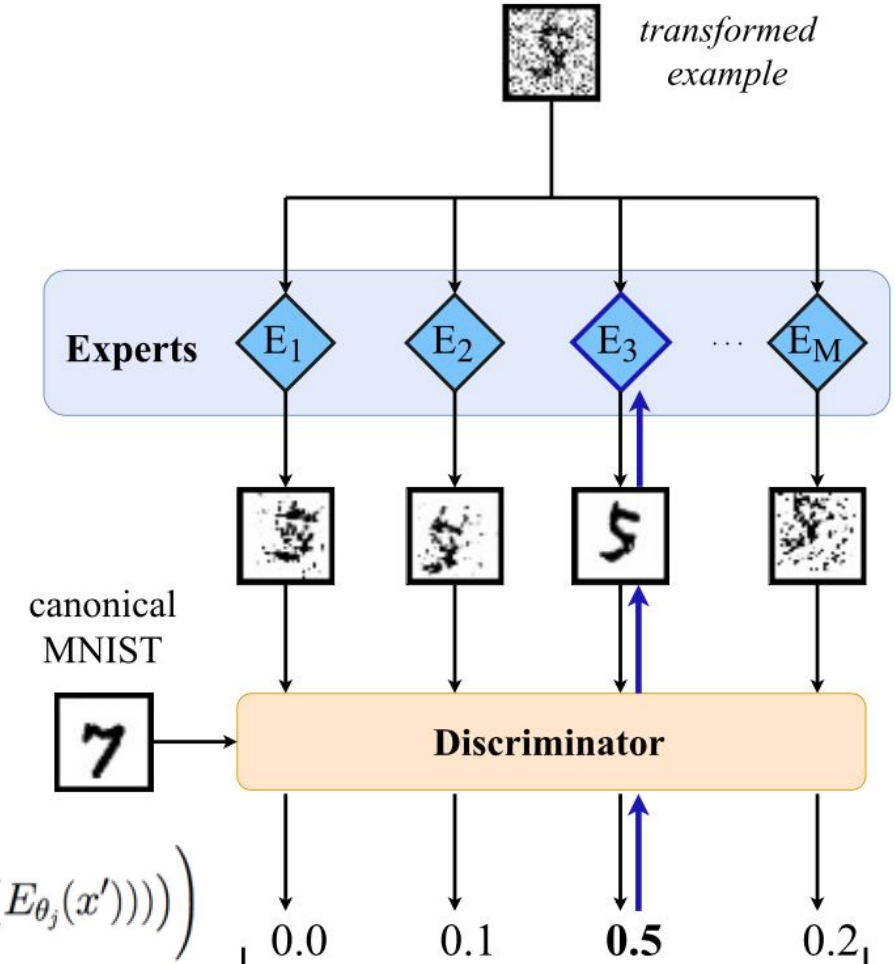


Original data

Transformed data

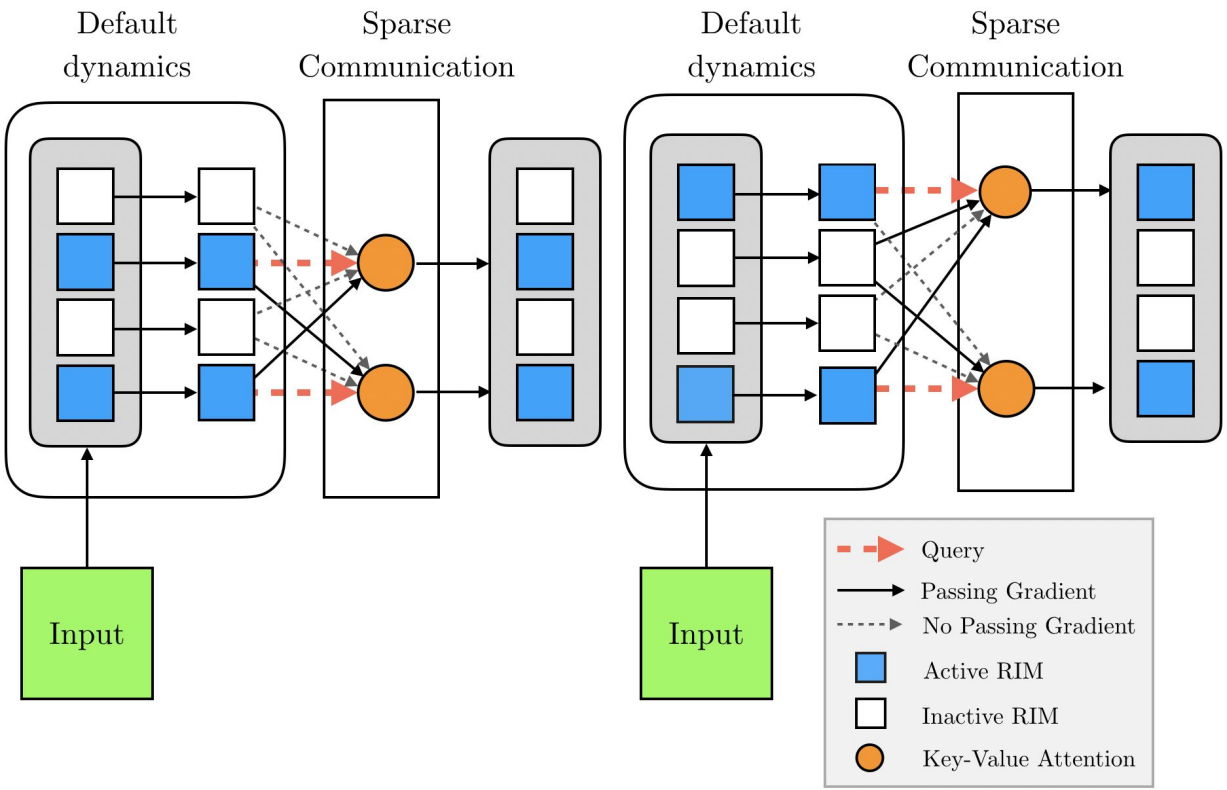
Method

- Mechanisms initialized ≈ identity
- The highest scoring mechanism against the discriminator D wins the example and is updated to increase the score
- D is trained on the original data and against the winning outputs



$$\max_{\theta_D} \left(\mathbb{E}_{x \sim P} \log(D_{\theta_D}(x)) + \frac{1}{N'} \sum_{j=1}^{N'} \mathbb{E}_{x' \sim Q} \left(\log(1 - D_{\theta_D}(E_{\theta_j}(x'))) \right) \right)$$

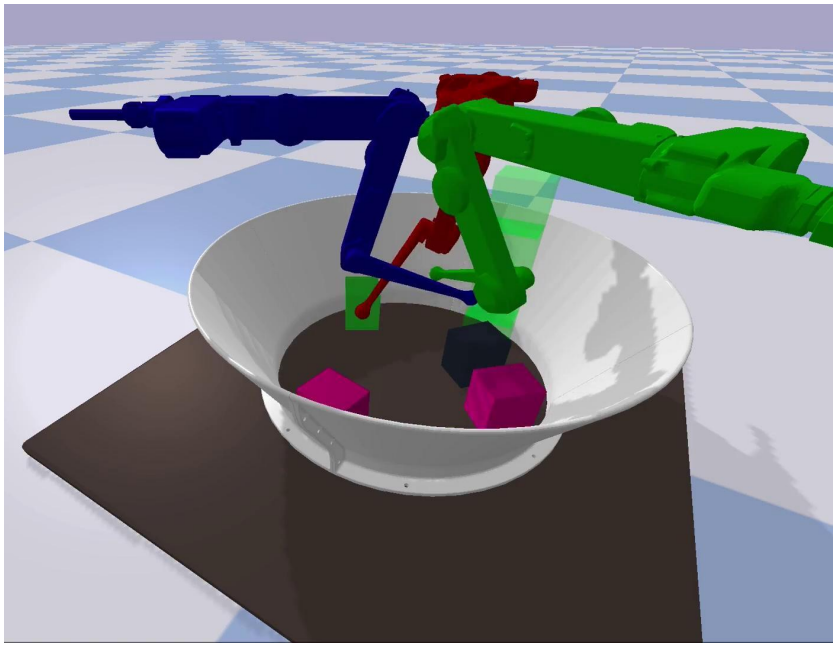
Recurrent Independent Mechanisms

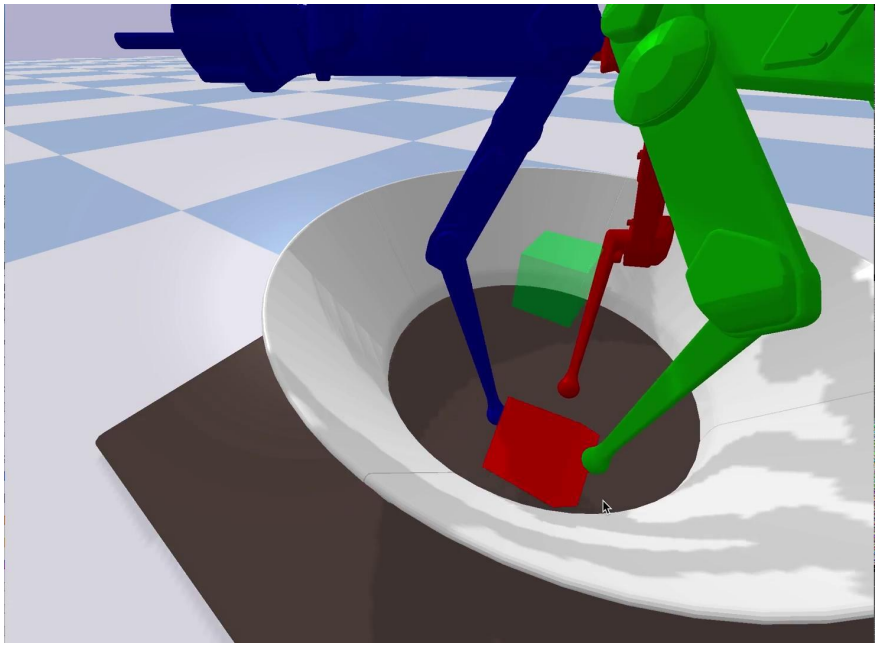


Anirudh Goyal: Modularity, Attention and Credit Assignment...computations (IAS Workshop on New Directions in Optimization Statistics and Machine Learning) https://www.youtube.com/watch?v=hwl6rgb2kQg&t=1s

A. Goyal, A. Lamb, J. Hoffmann, S. Sodhani, S. Levine, Y. Bengio, and B. Schölkopf, 2019. Recurrent independent mechanisms. arXiv:1909.10893.

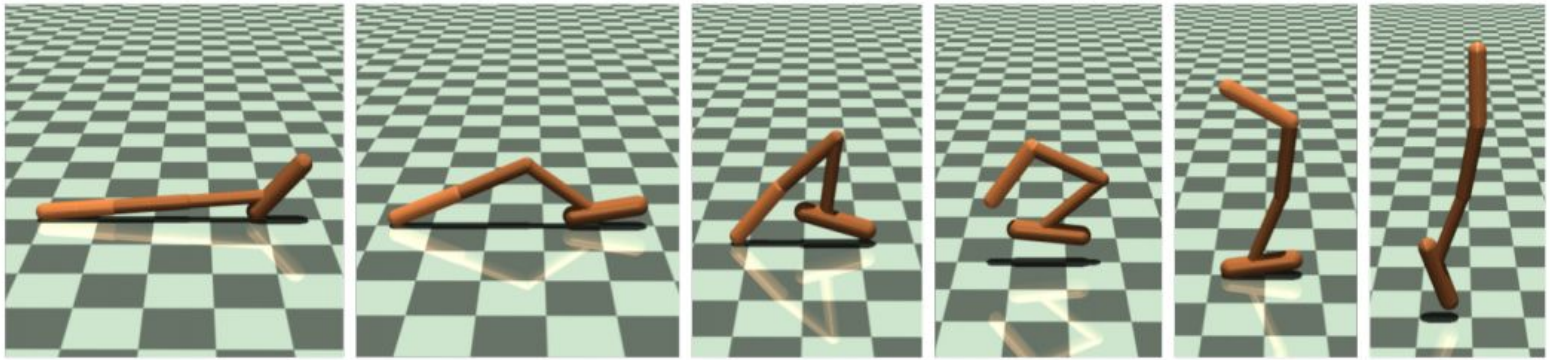
Upcoming - Using Robotic Systems as Benchmarks





Some Scepticism of Simulation Environments

- Rajeswaran, Aravind, et al. "Towards generalization and simplicity in continuous control." *Advances in Neural Information Processing Systems*. 2017. http://papers.nips.cc/paper/7233-towards-generalization-and-simplicity-in-continuous-control.pdf



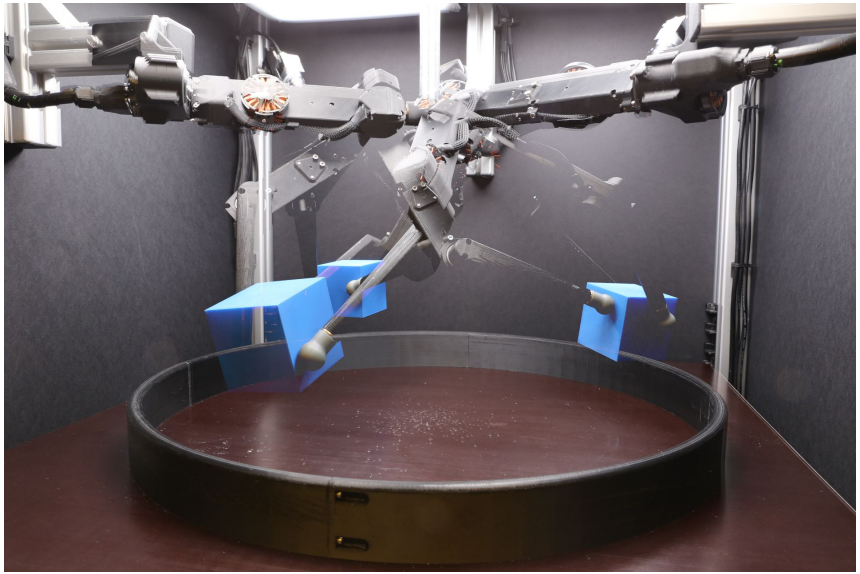
- Mania, Horia, Aurelia Guy, and Benjamin Recht. "Simple random search provides a competitive approach to reinforcement learning." *arXiv preprint arXiv:1803.07055* (2018). https://arxiv.org/pdf/1803.07055
- Fakoor, Rasool, et al. "Meta-Q-Learning." *arXiv preprint arXiv:1910.00125* (2019). https://arxiv.org/pdf/1910.00125.pdf

Hardware Design

- 3 x 3DoF-finger with shared workspace
- light panels, 3 RGB cameras
- mechanics based on [1]
  - low weight, high torque
  - 1 kHz torque control and sensing
  - robustness to impacts due to transparency of transmission

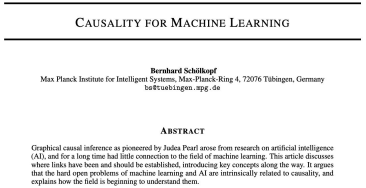
[1] Grimminger, F. et al. An Open Torque-Controlled Modular Robot Architecture for Legged Locomotion Research. *International Conference on Robotics and Automation (ICRA)*, 2020

What we have so far:



- Real-time Software Interface to code just using Python or C++
- Currently 2 platforms up two 8 in next months.
- Open-source (including software, micro-controller, boards, etc)
- Simulator

Hopefully some more updates during MLSS …



CAUSALITY FOR MACHINE LEARNING

Bernhard Schölkopf
Max Planck Institute for Intelligent Systems, Max-Planck-Ring 4, 72076 Tübingen, Germany
bs@tuebingen.mpg.de

ABSTRACT

Graphical causal inference as pioneered by Judea Pearl arose from research on artificial intelligence (AI), and for a long time had little connection to the field of machine learning. This article discusses where links have been and should be established, introducing key concepts along the way. It argues that the hard open problems of machine learning and AI are intrinsically related to causality, and explains how the field is beginning to understand them.

Summary

- Causal inference requires assumptions: Please be explicit about them!
- Key problems:
  - Scaling and computational efficiency.
  - Benchmarks and evaluations, data quantity.
  - In many applications, variables are not directly observed.
- Only at the beginning of our understanding for unstructured, non-linear inputs.
- Recent efforts especially focused on connection with deep learning. Pointers for future research:
  - Schölkopf, Bernhard. "Causality for machine learning." *arXiv preprint arXiv:1911.10500* (2019).
  - Anirudh Goyal: Modularity, Attention and Credit Assignment...computations (IAS Workshop on New Directions in Optimization Statistics and Machine Learning) https://www.youtube.com/watch?v=hwl6rgb2kQg&t=1s
  - Blaise Aguera: Social Intelligence https://slideslive.com/38922302/social-intelligence
  - Brendan Lake: Compositional generalization in minds and machines https://slideslive.com/38923478/compositional-generalization-in-minds-and-machines?ref=recommended-presentation-38922817

Advertisement - Upcoming ICML Workshop

Inductive Biases, Invariances and Generalization in RL (BIG)

International Conference on Machine Learning (ICML)

July 18, 2020

@BIGICML · #BIGICML

Contact: generalizationworkshop@gmail.com

Speakers: Yoshua Bengio, Sham Kakade, David Silver, Mengdi Wang, Martha White, Fanny Yang, Nicolas Heess, Caroline Uhler

Virtual and livestream available:
https://biases-invariances-generalization.github.io/

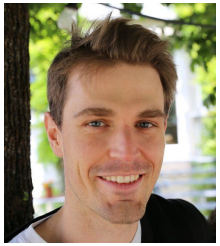
Advertisement - Open Internship Positions

Subset of results in collaboration with many amazing researchers at MPI and beyond (random order)

Diego Agudelo Espana

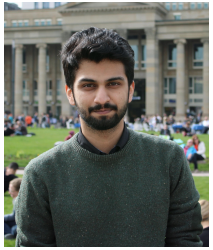
Djordje Miladinovic

Francesco Locatello



Manuel Wüthrich

Jonas Peters



Waleed Gondal

Felix Widmaier

Patrick Schwab

Joel Akpo

Niklas Pfister

Frederik Träuble

Anirudh Goyal

Andrea Dittadi

Raphael Suter

Ossama Ahmed

Rosemary Nan Ke

Felix Leeb

Thank You