Symbolic, Statistical, and Causal Representations

Bernhard Schölkopf

Max Planck Institute for Intelligent Systems

AIDA

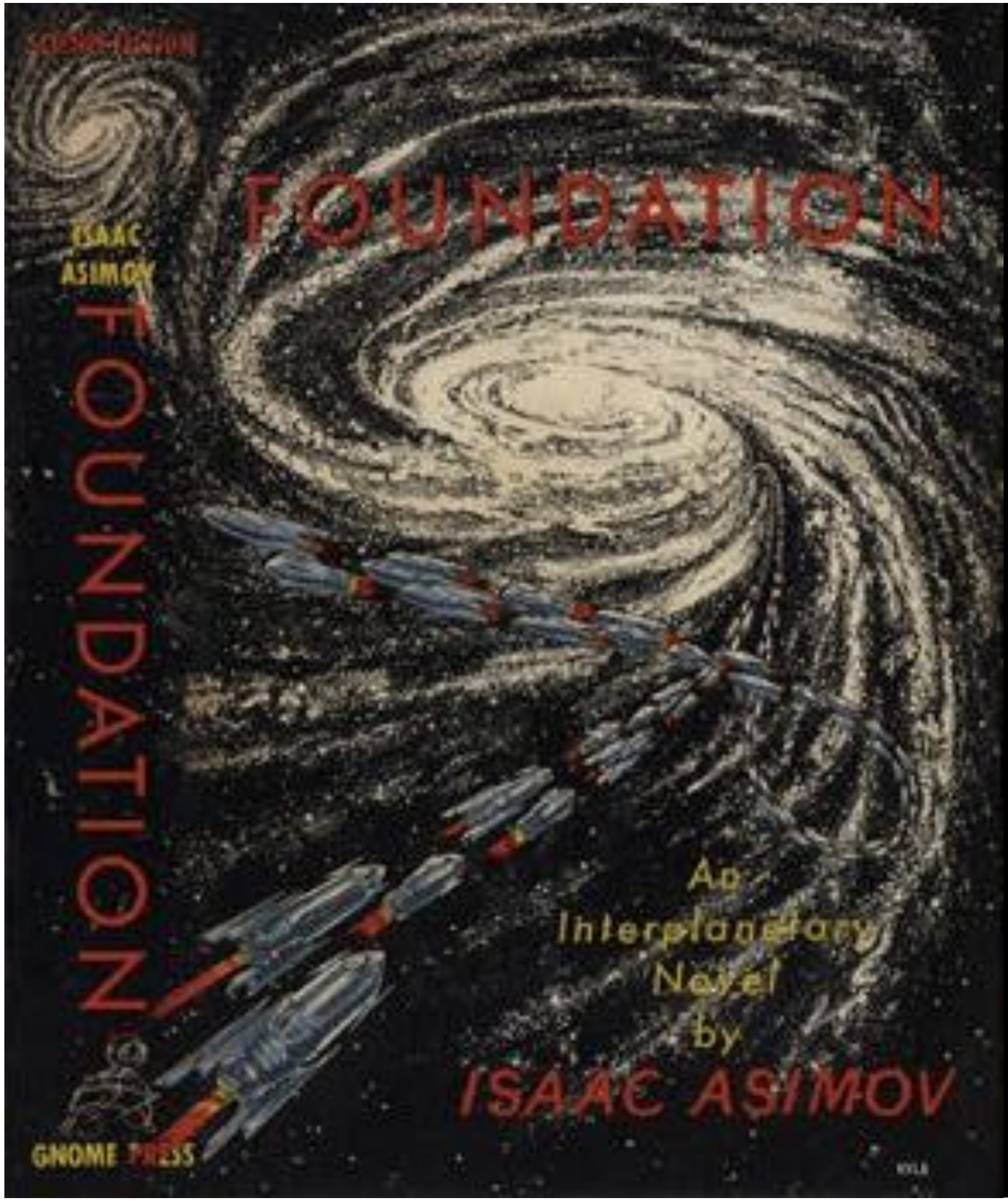
ARTIFICIAL INTELLIGENCE
DOCTORAL ACADEMY

12 13 JULY

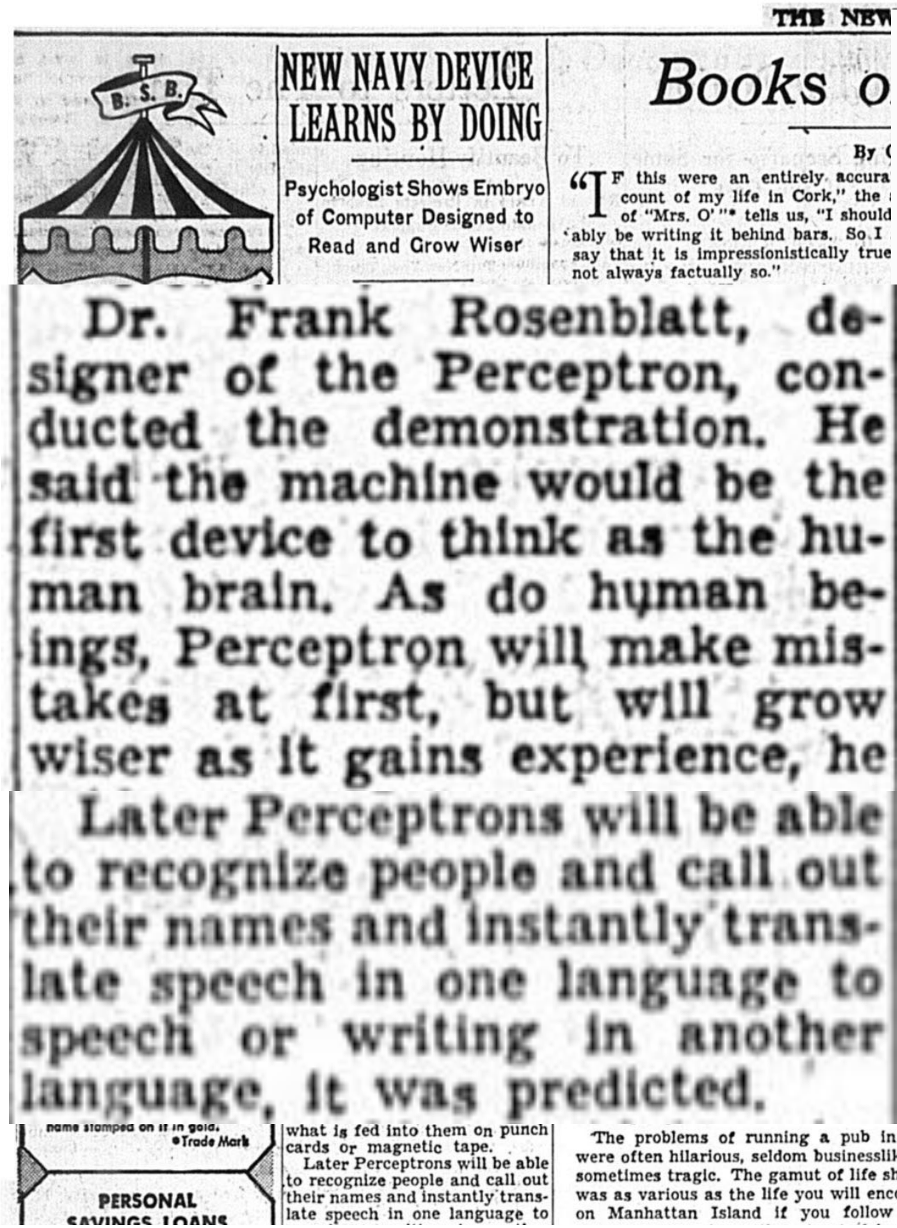
ELLIS PHD &
POSTDOC
SUMMIT '21

open to all ELLIS PhDs & Postdocs

ellis



"Hari Seldon [..] brought the science of psychohistory to its full development. [..] The individual human being is unpredictable, but the reactions of human mobs, Seldon found, could be treated statistically."



THE NEW

NEW NAVY DEVICE LEARNS BY DOING

Psychologist Shows Embryo of Computer Designed to Read and Grow Wiser

Dr. Frank Rosenblatt, designer of the Perceptron, conducted the demonstration. He said the machine would be the first device to think as the human brain. As do human beings, Perceptron will make mistakes at first, but will grow wiser as it gains experience, he

Later Perceptrons will be able to recognize people and call out their names and instantly translate speech in one language to speech or writing in another language, it was predicted.

what is fed into them on punch cards or magnetic tape.

Later Perceptrons will be able to recognize people and call out their names and instantly translate speech in one language to

Books o

By C

"IF this were an entirely accura count of my life in Cork," the of "Mrs. O'"* tells us, "I should ably be writing it behind bars. So I say that it is impressionistically true not always factually so."

ong to
buy it
eciate
signed
h this
As au-
l sim-

omers
refer-
book,
nation
ood of
s the

potted
quired
is she
napps
wine
es and
s and
terna-

The problems of running a pub in Cork were often hilarious, seldom businesslike and sometimes tragic. The gamut of life she saw was as various as the life you will encounter on Manhattan Island if you follow Park

witz with
the long
Early
she had l
stout, aft
rival call
"Wrap
three thi
seen Fox
bung, rai
that I sh
ever to b
two fairl

M

It work
livered a
place, t
around i
mention
awe, salu
say, augi
a huge h
the bar.
A rat
flowers i
past is a
ent is a
for trage
can doub
Mrs. O'.
Then t
glar. He

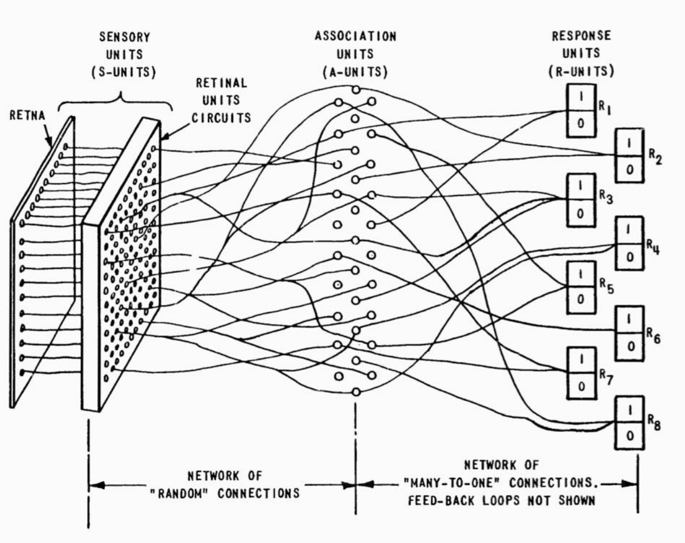
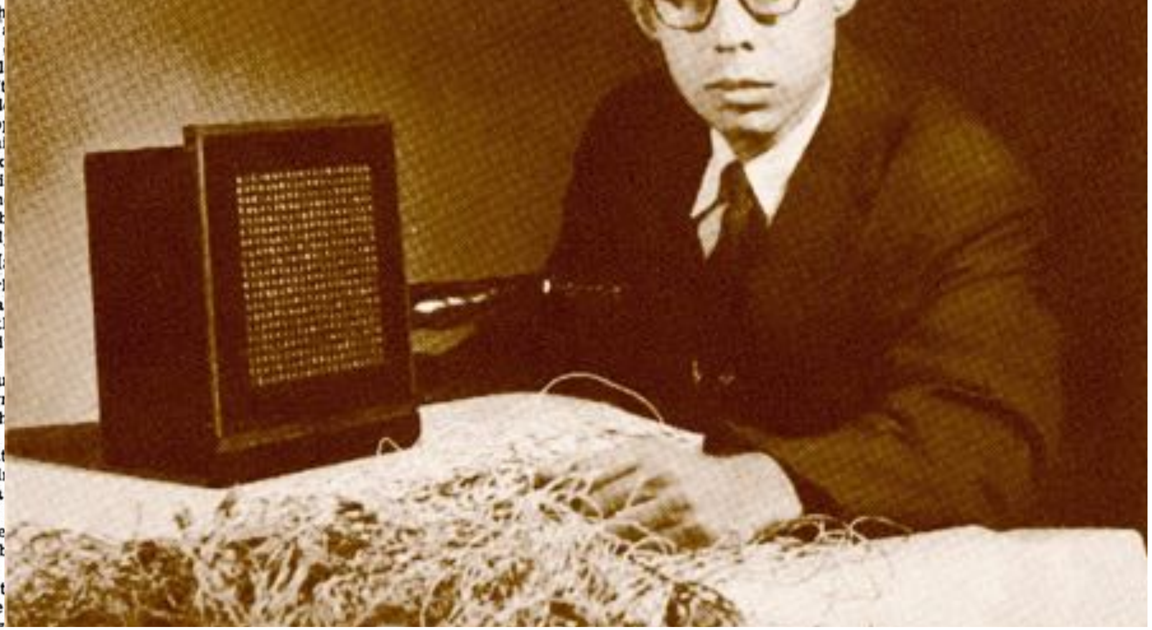

SENSORY UNITS (S-UNITS)
RETINAL UNITS CIRCUITS
ASSOCIATION UNITS (A-UNITS)
RESPONSE UNITS (R-UNITS)
RETNA
R1
R2
R3
R4
R5
R6
R7
R8
NETWORK OF "RANDOM" CONNECTIONS
NETWORK OF "MANY-TO-ONE" CONNECTIONS. FEED-BACK LOOPS NOT SHOWN

Figure I ORGANIZATION OF THE MARK I PERCEPTRON

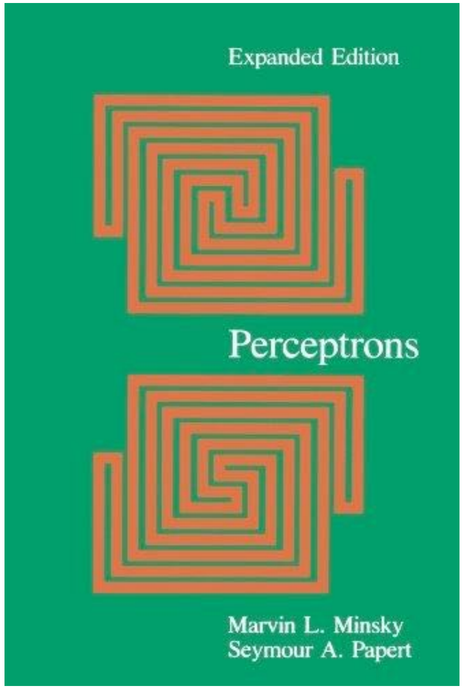


Symbolic AI

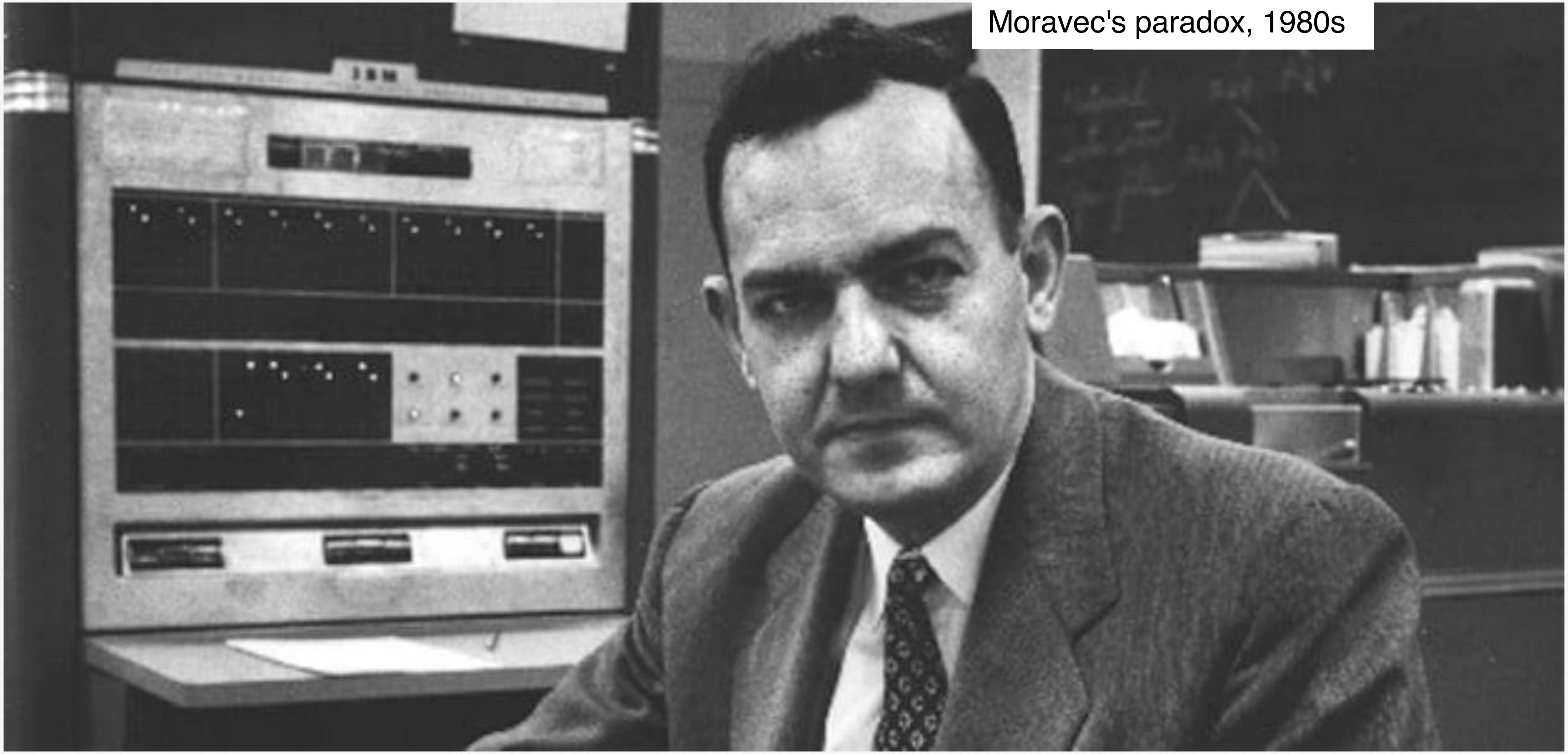
1975 Turing Award Lecture

Newell & Simon:

The Physical Symbol System Hypothesis. A physical symbol system has the necessary and sufficient means for general intelligent action.



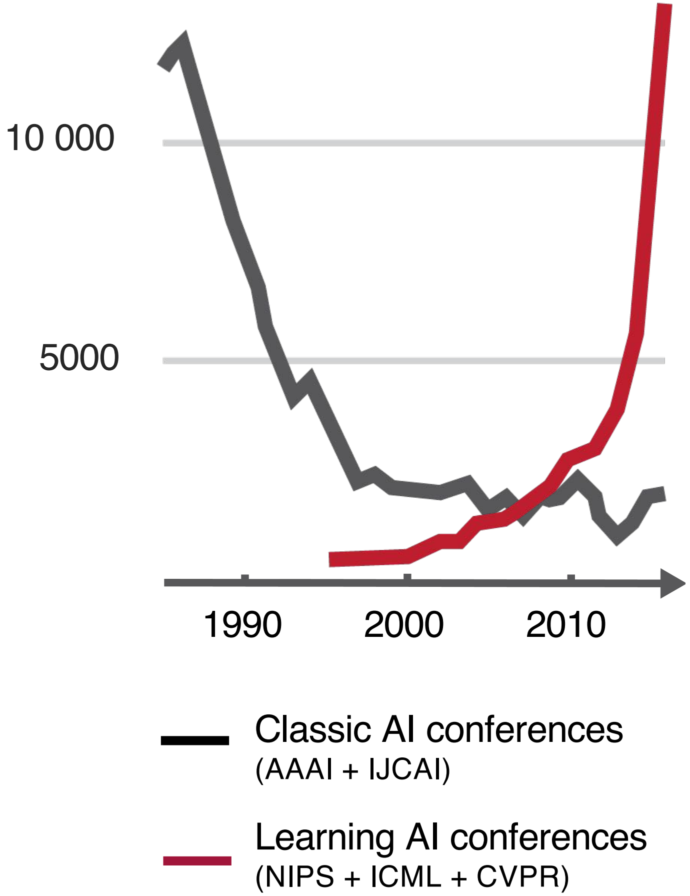
Moravec's paradox, 1980s



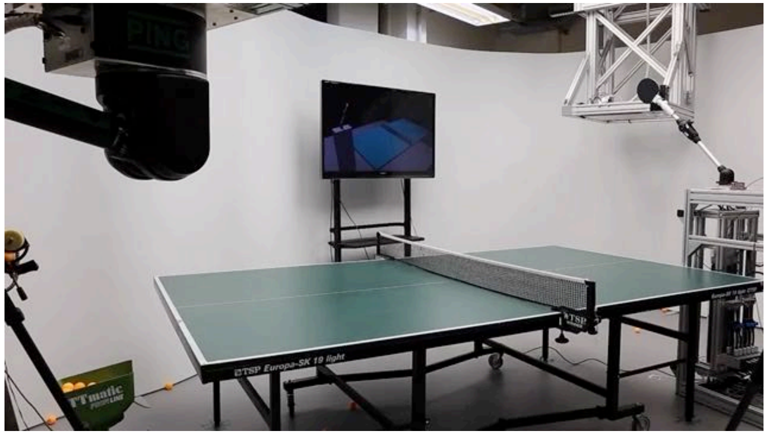
“Machines will be capable, within twenty years, of doing any work a man can do”

Classic AI

Programmed by humans



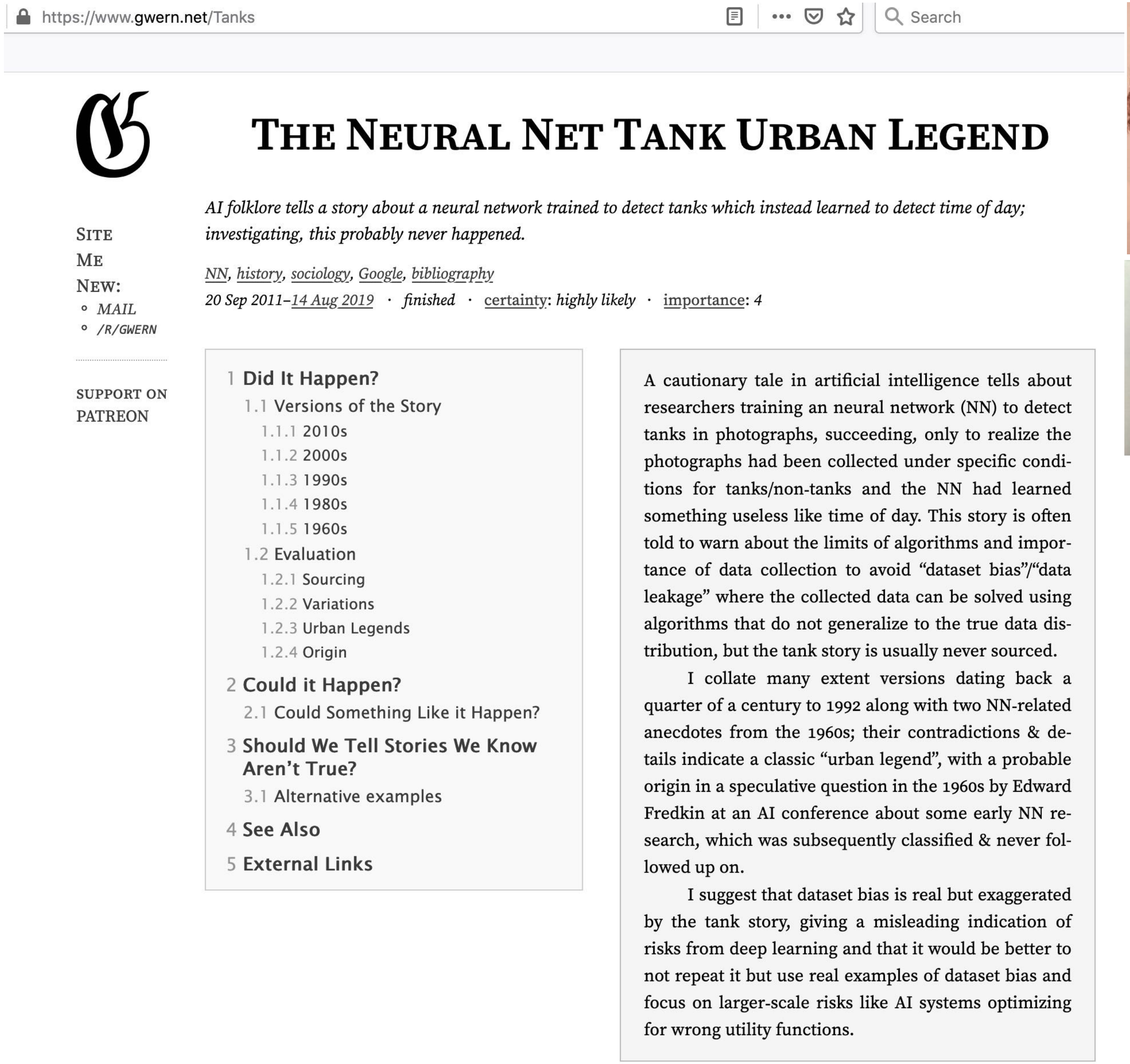
Learning AI *(Machine learning)*



Programmed by humans and learning from experience

(Video: D. Büchler)

Image credit: W. Brendel / M. Bethge



The Neural Net Tank Urban Legend

AI folklore tells a story about a neural network trained to detect tanks which instead learned to detect time of day; investigating, this probably never happened.

Site
Me
New:
- MAIL
- /R/GWERN

SUPPORT ON PATREON

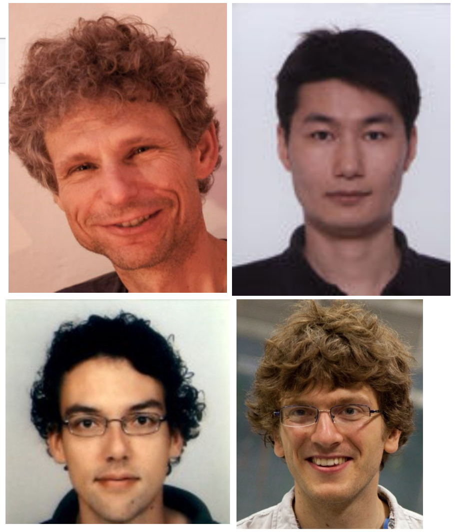
NN, history, sociology, Google, bibliography

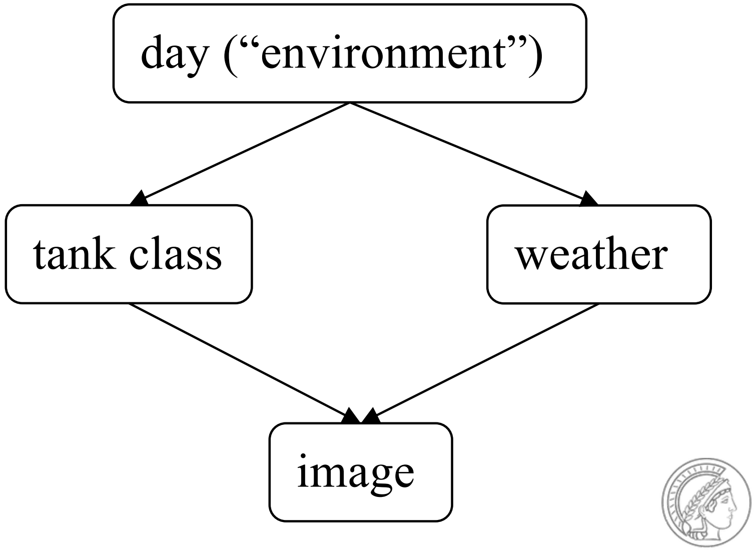
20 Sep 2011–14 Aug 2019 · finished · certainty: highly likely · importance: 4

A cautionary tale in artificial intelligence tells about researchers training an neural network (NN) to detect tanks in photographs, succeeding, only to realize the photographs had been collected under specific conditions for tanks/non-tanks and the NN had learned something useless like time of day. This story is often told to warn about the limits of algorithms and importance of data collection to avoid "dataset bias"/"data leakage" where the collected data can be solved using algorithms that do not generalize to the true data distribution, but the tank story is usually never sourced.

I collate many extent versions dating back a quarter of a century to 1992 along with two NN-related anecdotes from the 1960s; their contradictions & details indicate a classic "urban legend", with a probable origin in a speculative question in the 1960s by Edward Fredkin at an AI conference about some early NN research, which was subsequently classified & never followed up on.

I suggest that dataset bias is real but exaggerated by the tank story, giving a misleading indication of risks from deep learning and that it would be better to not repeat it but use real examples of dataset bias and focus on larger-scale risks like AI systems optimizing for wrong utility functions.





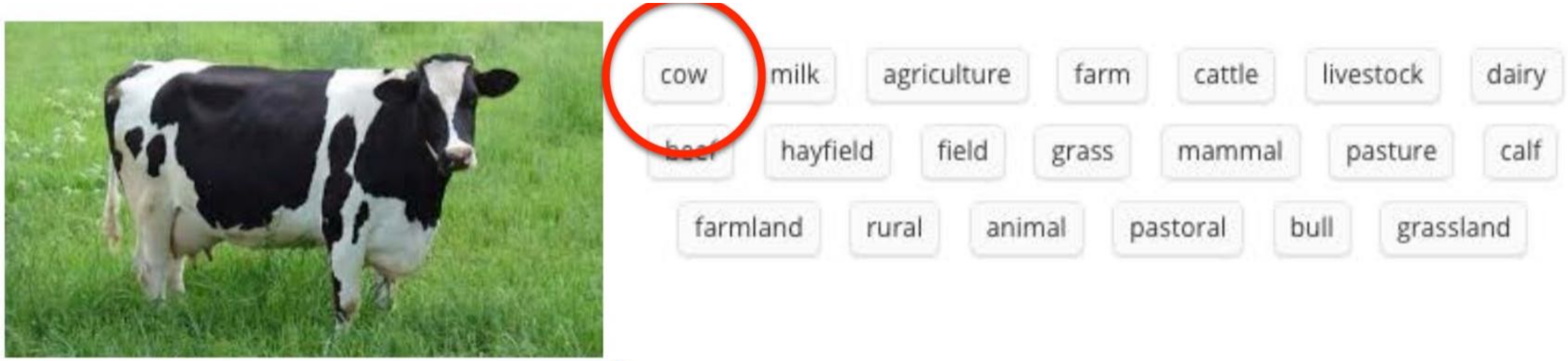
day ("environment")

tank class

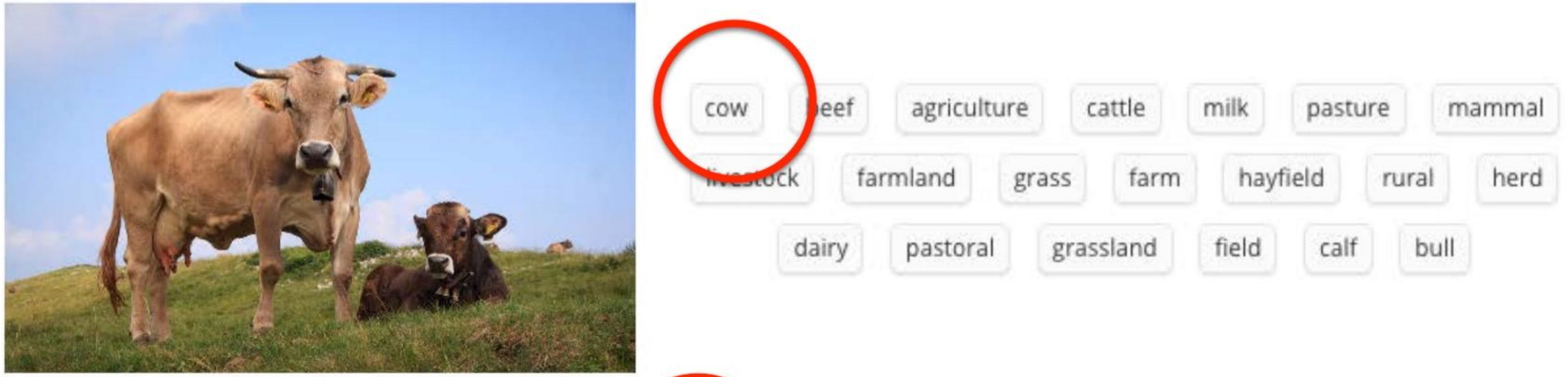
weather

image

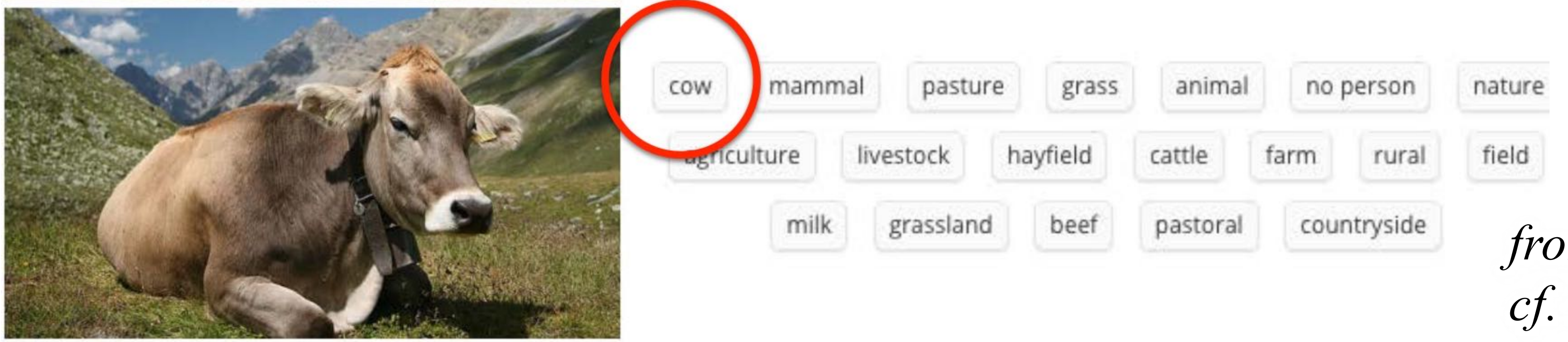
Human-level object recognition?



cow milk agriculture farm cattle livestock dairy
beef hayfield field grass mammal pasture calf
farmland rural animal pastoral bull grassland



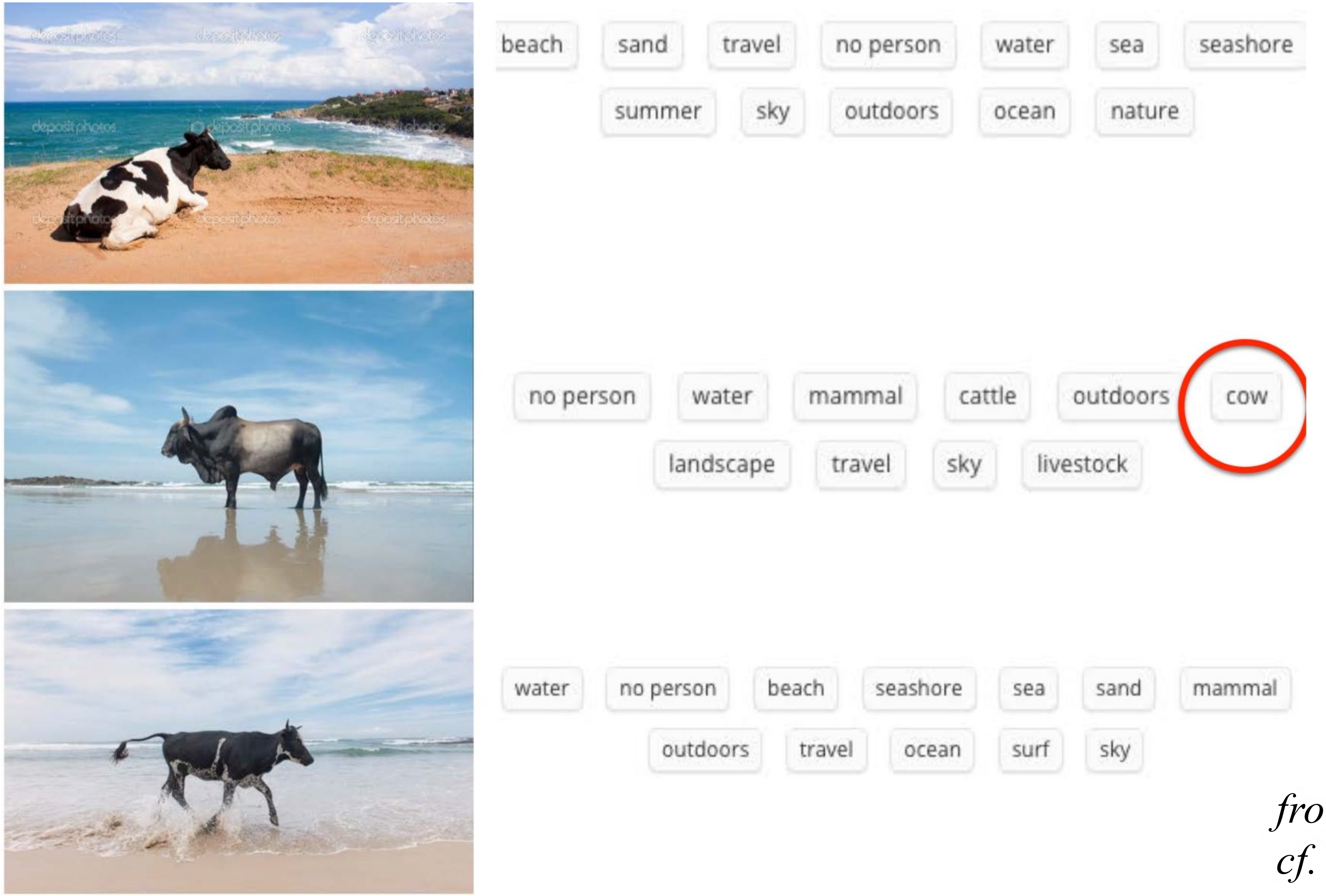
cow beef agriculture cattle milk pasture mammal
livestock farmland grass farm hayfield rural herd
dairy pastoral grassland field calf bull



cow mammal pasture grass animal no person nature
agriculture livestock hayfield cattle farm rural field
milk grassland beef pastoral countryside

from Perona, 2017;
cf. Lopez-Paz et al., 2016

Machine learning uses correlations rather than causality



beach sand travel no person water sea seashore
summer sky outdoors ocean nature

no person water mammal cattle outdoors cow
landscape travel sky livestock

water no person beach seashore sea sand mammal
outdoors travel ocean surf sky

from Perona, 2017;
cf. Lopez-Paz et al., 2016

Adversarial Vulnerability

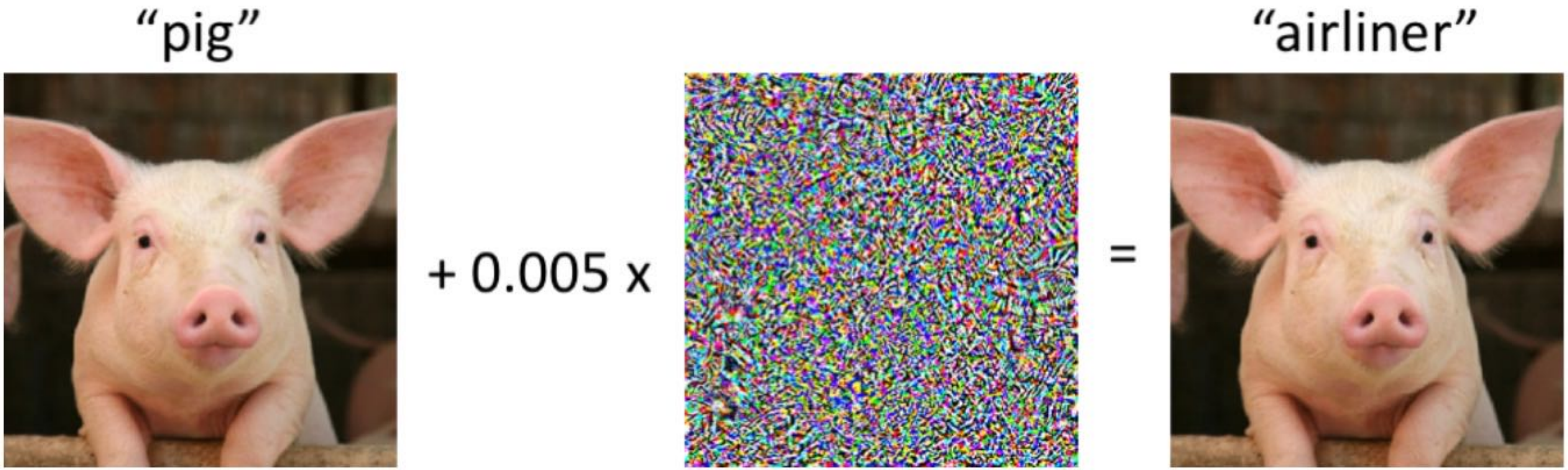


Image credit: http://people.csail.mit.edu/madry/lab/blog/adversarial/2018/07/06/adversarial_intro/

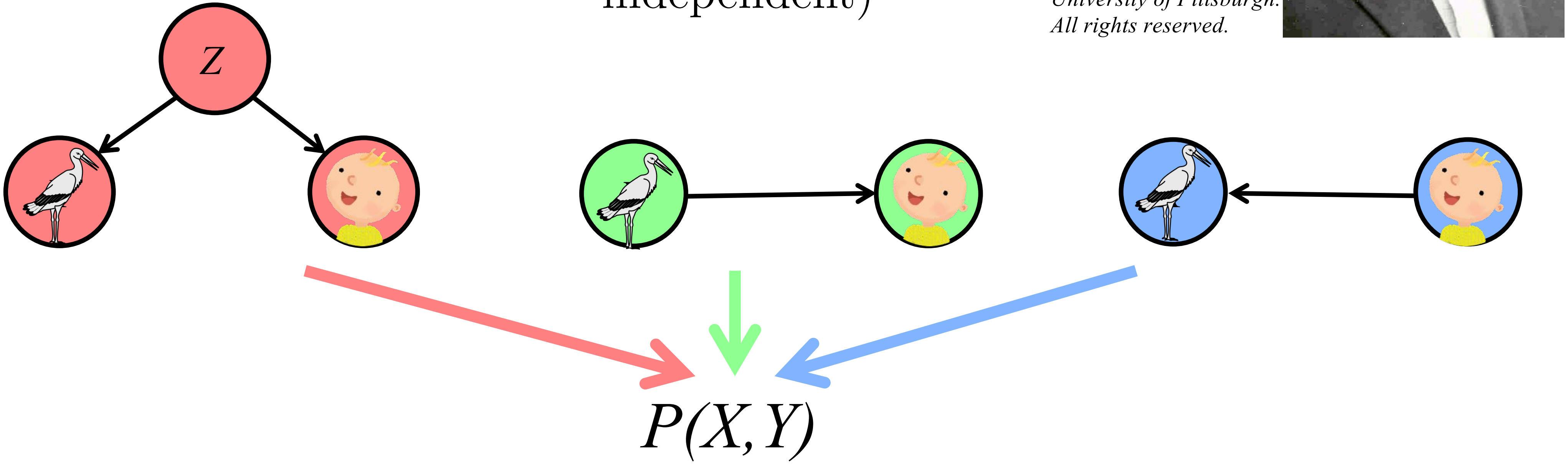
C. Szegedy et al. Intriguing properties of neural networks. *arXiv:1312.6199, 2013*

Reichenbach’s Common Cause Principle

(i) if X and Y are dependent, then there exists Z *causally* influencing both;

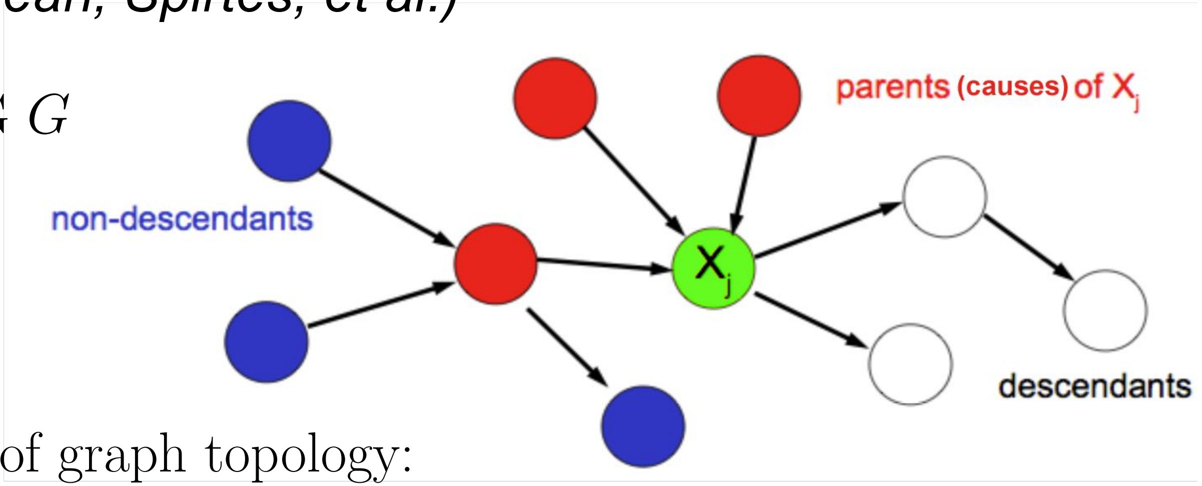
(ii) Z screens X and Y from each other (given Z, X und Y become independent)

by permission of the University of Pittsburgh. All rights reserved.



$\sum_z p(x|z)p(y|z)p(z)$ $p(x)p(y|x)$ $p(x|y)p(y)$

Structural causal models *(Pearl, Spirtes, et al.)*



- Set of observables $X_1, \ldots, X_n$ on a DAG G
- arrows represent direct causation
- $X_i := f_i(\mathrm{PA}_i, U_i)$ with independent RVs $U_1, \ldots, U_n$.
- distribution of the U_i picks up footprint of graph topology: *observational distribution* $p(X_1, \ldots, X_n)$ satisfies:

  Conditioned on its parents, X_i is independent of its non-descendants *(causal Markov condition)*

  —assay using conditional independence testing (for $n > 2$)
- (G, p) is a "graphical model" *(Lauritzen, 1996)*, $p(X_1, \ldots, X_n) = \prod_i p(X_i|\mathrm{Parents}_i)$
- *interventions / mechanism shifts* are modelled by changing functions (mechanisms); entail *interventional distribution*

Causality in differential equations

Consider the set of differential equations

$$\frac{d\mathbf{x}}{dt} = f(\mathbf{x}),\ \mathbf{x} \in \mathbb{R}^d,$$

with initial value $\mathbf{x}(t_0) = \mathbf{x}_0$.

Picard–Lindelöf: locally, if f is Lipschitz, there exists a unique solution $\mathbf{x}(t)$

$\Longrightarrow$ the immediate future of $\mathbf{x}$ is implied by its past

Using dt and $d\mathbf{x} = \mathbf{x}(t + dt) - \mathbf{x}(t)$:

$$\mathbf{x}(t + dt) = \mathbf{x}(t) + dt \cdot f(\mathbf{x}(t)).$$

This tells us which entries of $\mathbf{x}(t)$ cause the future of others $\mathbf{x}(t + dt)$, i.e., the causal structure.

https://arxiv.org/abs/1911.10500

Linking ODEs and SCMs

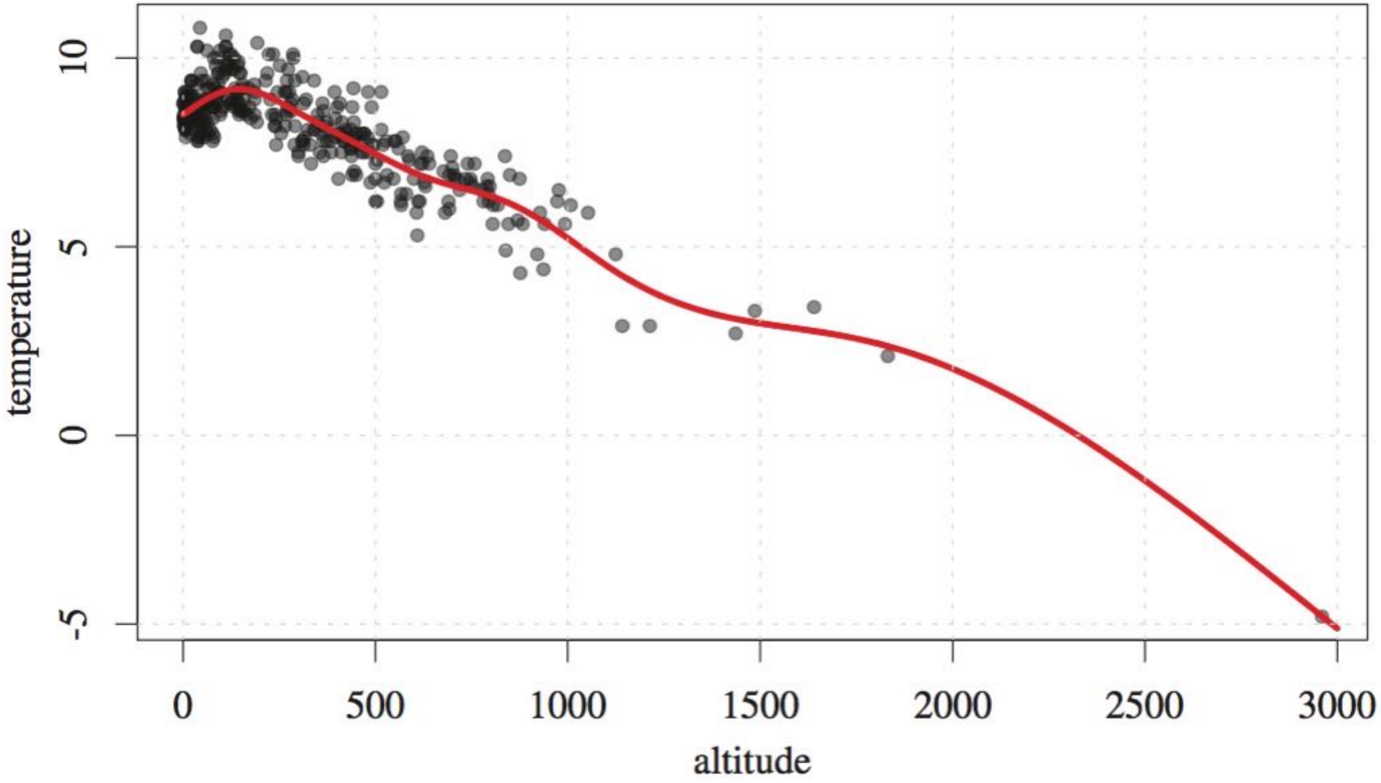
There is no reason why simple SCMs should be derivable in general.

Derive SCMs describing the interventional behavior of a coupled ODE system in equilibrium state and perturbed in an "adiabatic" way (Mooij et al., UAI 2013)

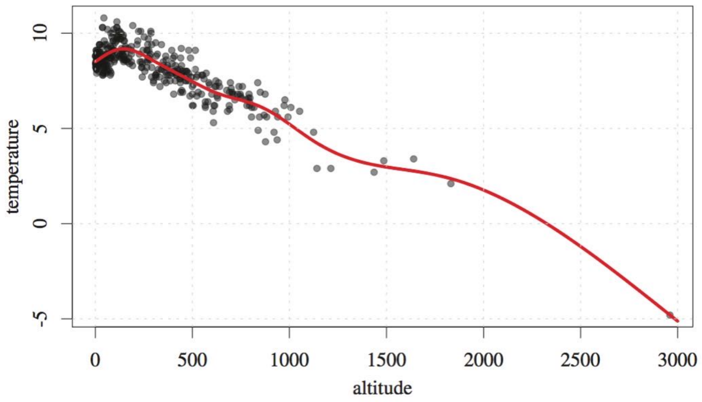
Generalization to oscillatory systems (Rubenstein et al., UAI 2018)

Subject to appropriate conditions, structural models can arise from coarse-graining of microscopic models, including microscopic structural equation models (Chalupka et al. 2015; Rubenstein et al., UAI 2017) or temporally aggregated time series (Gong et al., UAI 2017).

What is cause and what is effect?



$$p(a,t) = p(a|t)\ p(t) \quad T \rightarrow A$$
$$= p(t|a)\ p(a) \quad A \rightarrow T$$



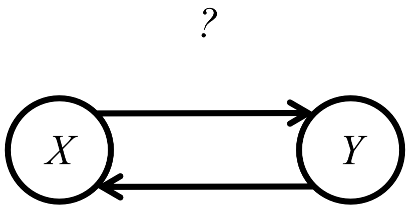
- **intervention** on a: raise the city, find that t changes
- hypothetical intervention on a: still expect that t changes, since we can think of a physical mechanism $p(t|a)$ that is **independent** of $p(a)$
- we expect that $p(t|a)$ is **invariant** across, say, different countries in a similar climate zone

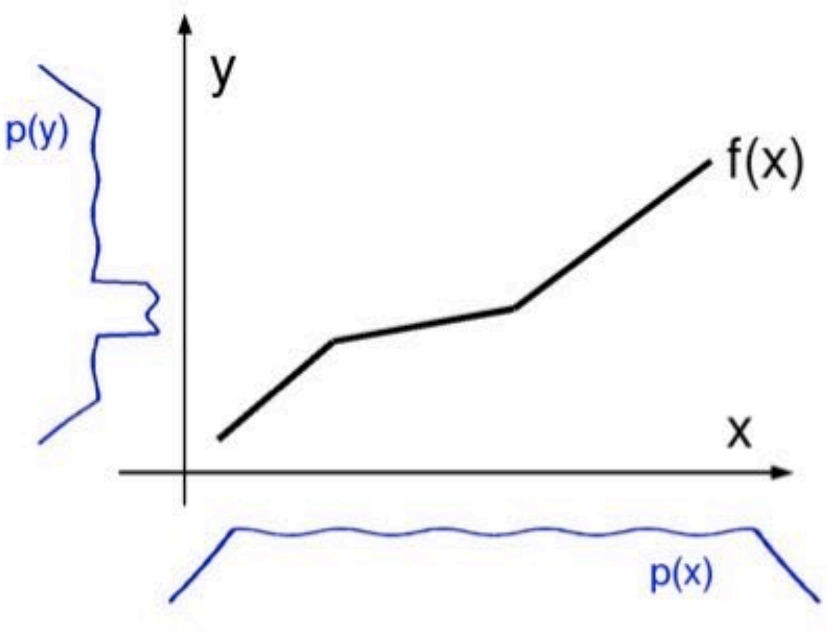
Independent Causal Mechanisms Principle (ICM):

The causal generative process is composed of autonomous modules that do not inform or influence each other.

Independence of input and mechanism

- No noise on effect variable
- Assumption: $y = f(x)$ with invertible f





Daniusis, Janzing, Mooij, Zscheischler, Steudel, Zhang, Schölkopf:
Inferring deterministic causal relations, *UAI* 2010

Causal independence implies anticausal dependence

Assume that f is a monotonically increasing bijection of $[0,1]$.

View p_x and $\log f'$ as RVs on the prob. space $[0,1]$ w. Lebesgue measure.

Postulate (independence of mechanism and input):

$$\operatorname{Cov}(\log f', p_x) = 0$$

Note: this is equivalent to

$$\int_0^1 \log f'(x) p(x) dx = \int_0^1 \log f'(x) dx,$$

since $\operatorname{Cov}(\log f', p_x) = E[\log f' \cdot p_x] - E[\log f'] E[p_x] = E[\log f' \cdot p_x] - E[\log f']$.

Proposition: If $f \neq Id$,

$$\operatorname{Cov}(\log f^{-1'}, p_y) > 0.$$

u_x, u_y uniform densities for x, y

v_x, v_y densities for x, y induced by transforming u_y, u_x via f^{-1} and f

Equivalent formulations of the postulate:

Additivity of Entropy:

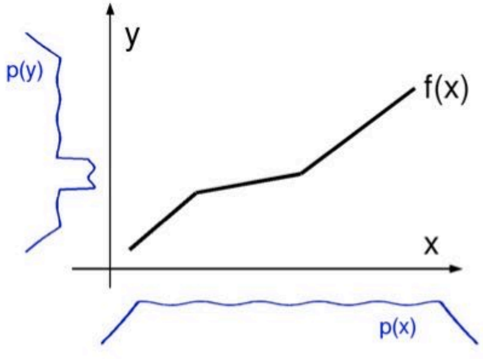
$$S(p_y) - S(p_x) = S(v_y) - S(u_x)$$

Orthogonality (information geometric):

$$D(p_x \| v_x) = D(p_x \| u_x) + D(u_x \| v_x)$$

which can be rewritten as

$$D(p_y \| u_y) = D(p_x \| u_x) + D(v_y \| u_y)$$

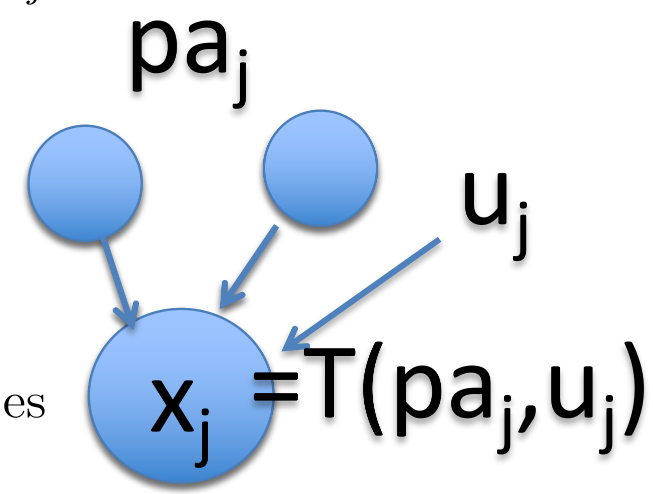


Interpretation:

irregularity of p_y = irregularity of p_x + irregularity introduced by f

Algorithmic structural causal model

- for every node x_j there exists a program u_j that computes x_j from its parents pa_j



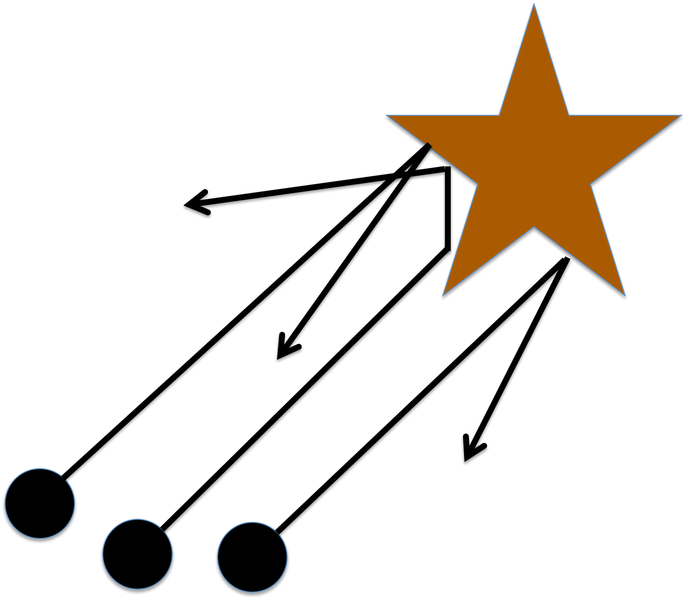
- all u_j are jointly independent
- the program u_j represents the causal mechanism that generates the effect from its causes
- u_j are the analog of the unobserved noise terms in the statistical functional model

Theorem: this model implies the causal Markov condition (replacing Shannon entropy with Kolmorogov complexity).

(Janzing & Schölkopf, IEEE Trans. Information Theory 2010)

Gedankenexperiment

Particles scattered at an object



- incoming beam: ‘cause’
- scattering at object: ‘mechanism’
- outgoing beam: ‘effect’, contains information about the object

Independence assumption

- s initial state of a physical system
- M the system dynamics applied for some fixed time

Independence Principle: s and M are algorithmically independent

$$I(s : M) \overset{+}{=} 0,$$

i.e., knowing s does not enable a shorter description of M and vice versa.

Thermodynamic Arrow of Time

Theorem [non-decrease of entropy]. Let M be a bijective map on the set of states of a system then $I(s : M) \overset{+}{=} 0$ implies

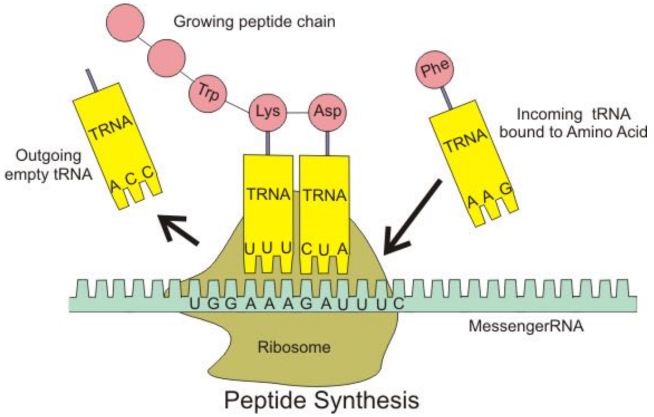
$$K(M(s)) \overset{+}{\geq} K(s)$$

Proof idea: If $M(s)$ admits a shorter description than s, knowing M admits a shorter description of s: just describe $M(s)$ and then apply M^{-1}.

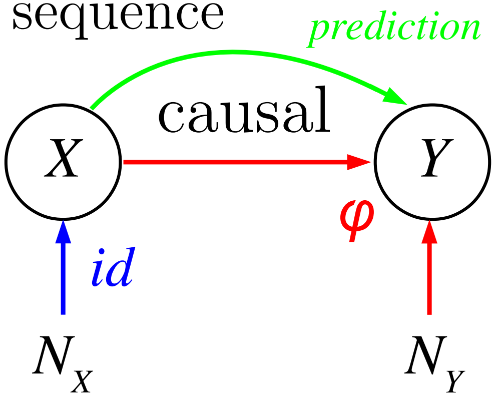
Janzing, Chaves, Schölkopf: Algorithmic independence of initial condition and dynamical law in thermodynamics and causal inference. New J. of Physics, 2016

Using cause-effect knowledge

- example 1: predict protein from mRNA sequence

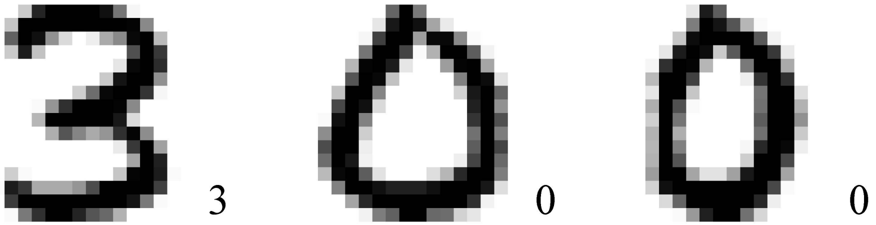


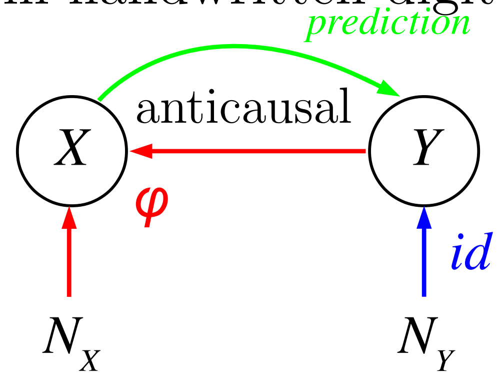
Source: http://commons.wikimedia.org/wiki/File:Peptide_syn.png



causal mechanism φ

- example 2: predict class membership from handwritten digit





Covariate Shift and Semi-Supervised Learning

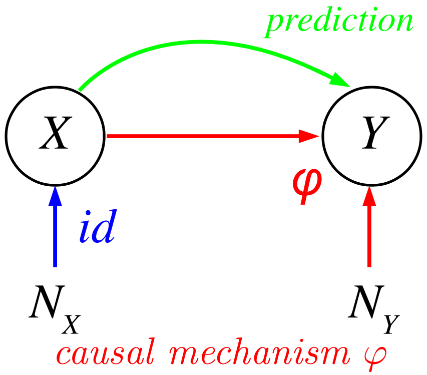
Assumption: $p(C)$ and mechanism $p(E|C)$ "independent"
Goal: learn $X \mapsto Y$, i.e., estimate (properties of) $p(Y|X)$

Semi-supervised learning: improve estimate by more data from $p(X)$
Covariate shift: $p(X)$ changes between training and test

Causal learning

$p(X)$ and $p(Y|X)$ independent

1. *semi-supervised learning impossible*
2. *$p(Y|X)$ invariant under change in $p(X)$*

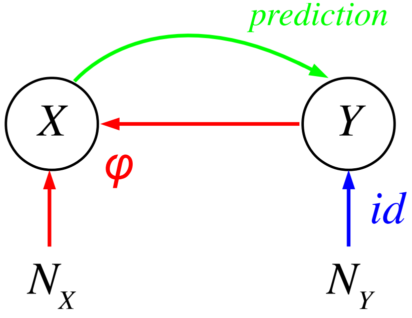


Anticausal learning

$p(Y)$ and $p(X|Y)$ independent

hence $p(X)$ and $p(Y|X)$ dependent

1. *semi-supervised learning possible*
2. *$p(Y|X)$ changes with $p(X)$*



Schölkopf, Janzing, Peters, Sgouritsa, Zhang, Mooij, 2012, cf. Storkey, 2009; Bareinboim & Pearl, 2012

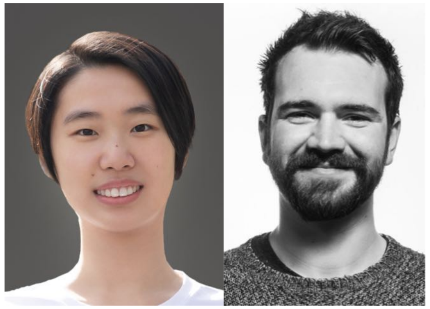
- Experimental Meta-Analysis confirms prediction

  Schölkopf et al., ICML 2012; von Kügelgen et al., UAI 2020, Jin et al., submitted

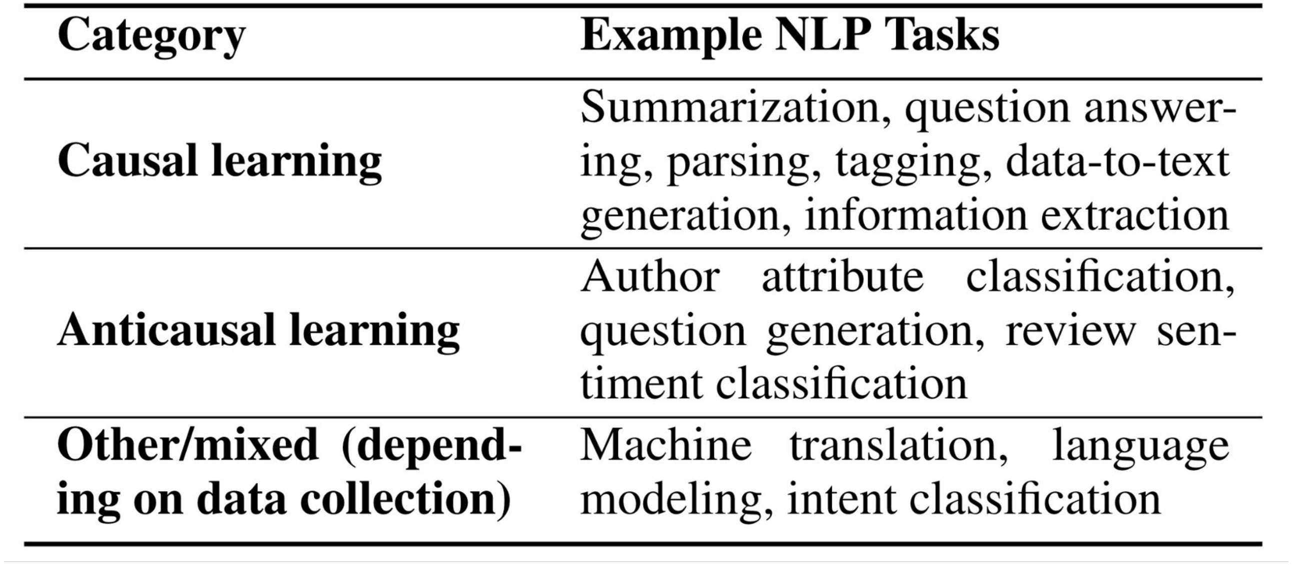
- All known SSL assumptions link $p(X)$ to $p(Y|X)$:
  - ***Cluster assumption***: points in same cluster of $p(X)$ have the same Y
  - ***Low density separation assumption***: $p(Y|X)$ should cross 0.5 in an area where $p(X)$ is small
  - ***Semi-supervised smoothness assumption***: $\mathrm{E}(Y|X)$ should be smooth where $p(X)$ is large

Independent Causal Mechanisms in NLP

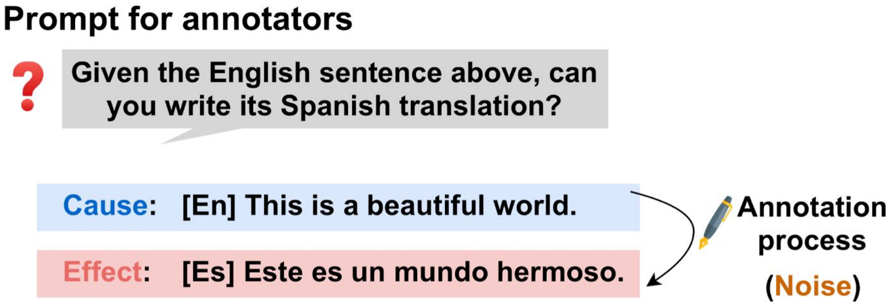
(with Zhijing Jin & Julius von Kügelgen)



Common NLP tasks:



Category	Example NLP Tasks
Causal learning	Summarization, question answering, parsing, tagging, data-to-text generation, information extraction
Anticausal learning	Author attribute classification, question generation, review sentiment classification
Other/mixed (depending on data collection)	Machine translation, language modeling, intent classification



Prompt for annotators

❓ Given the English sentence above, can you write its Spanish translation?

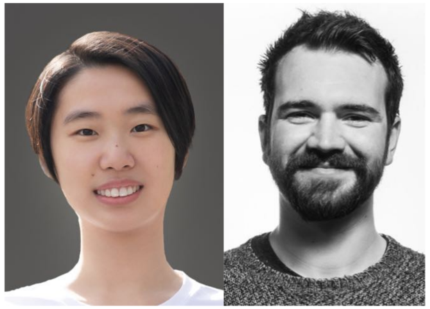
Cause: [En] This is a beautiful world.

Effect: [Es] Este es un mundo hermoso.

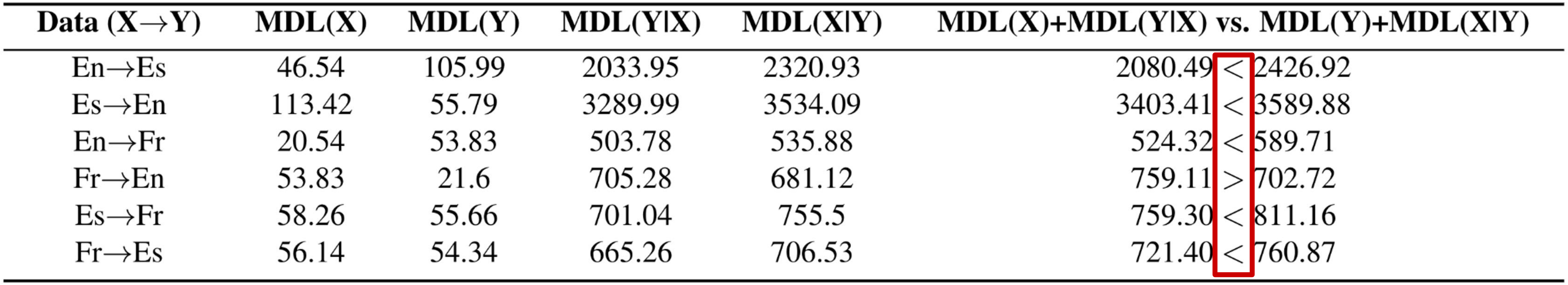
Effect = CausalMechanism (Cause, Noise)

ICM in NLP: Findings

(with Zhijing Jin & Julius von Kügelgen)



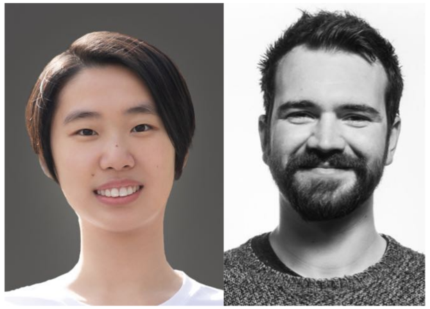
Causal direction corresponds to shorter description of machine translation data in terms of minimum description length (MDL):



Data (X→Y)	MDL(X)	MDL(Y)	MDL(Y|X)	MDL(X|Y)	MDL(X)+MDL(Y|X) vs. MDL(Y)+MDL(X|Y)
En→Es	46.54	105.99	2033.95	2320.93	2080.49 < 2426.92
Es→En	113.42	55.79	3289.99	3534.09	3403.41 < 3589.88
En→Fr	20.54	53.83	503.78	535.88	524.32 < 589.71
Fr→En	53.83	21.6	705.28	681.12	759.11 > 702.72
Es→Fr	58.26	55.66	701.04	755.5	759.30 < 811.16
Fr→Es	56.14	54.34	665.26	706.53	721.40 < 760.87

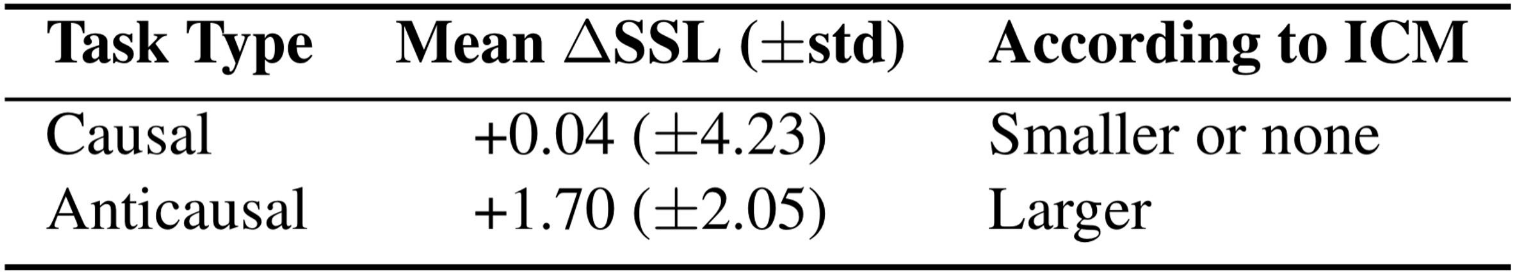
ICM in NLP: Findings

(with Zhijing Jin & Julius von Kügelgen)



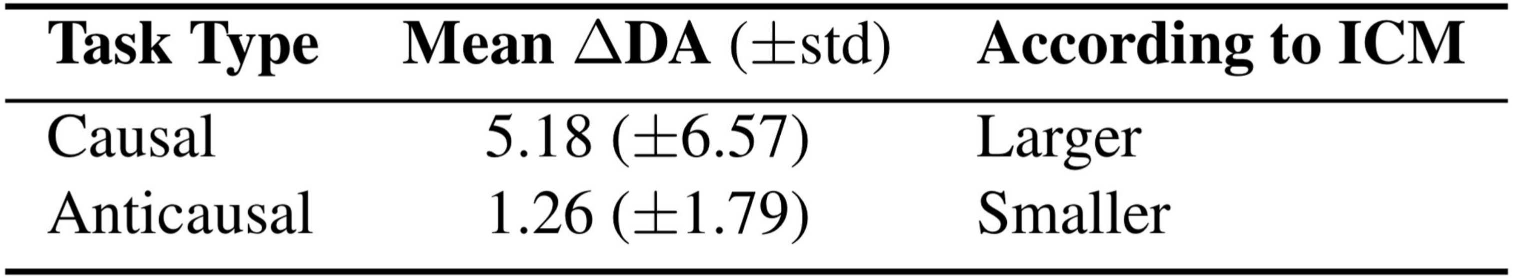
Implications of ICM for SSL and DA confirmed by NLP meta-study:

Semi-supervised learning (SSL): *anticausal* > *causal*.



Task Type	**Mean ΔSSL (±std)**	**According to ICM**
Causal	+0.04 (±4.23)	Smaller or none
Anticausal	+1.70 (±2.05)	Larger

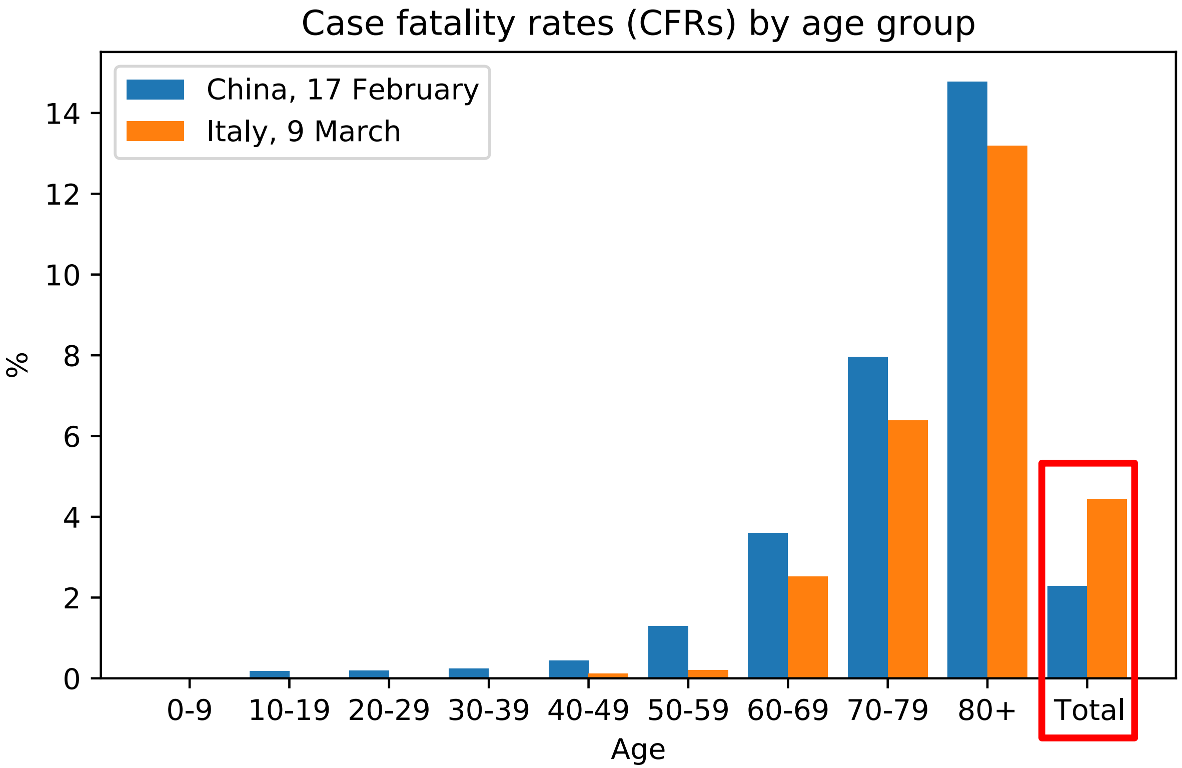
Domain adaptation (DA): *causal* > *anticausal*.



Task Type	**Mean ΔDA** (±std)	**According to ICM**
Causal	5.18 (±6.57)	Larger
Anticausal	1.26 (±1.79)	Smaller

Simpson's paradox in Covid-19 case fatality

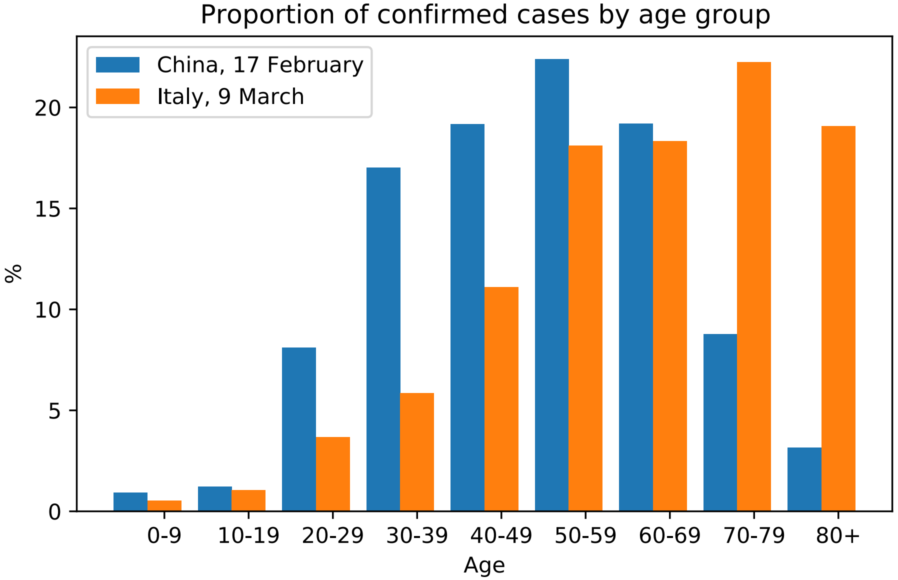
(v. Kügelgen, Gresele, https://arxiv.org/abs/2005.07180 / IEEE Trans. AI)



Case fatality rates (CFRs) in Italy are *lower* for each age group, but *higher* overall.

Simpson's paradox: opposite trends in grouped and aggregated data.

Here, it stems from a difference in case demographic:



Thanks to Elias Bareinboim

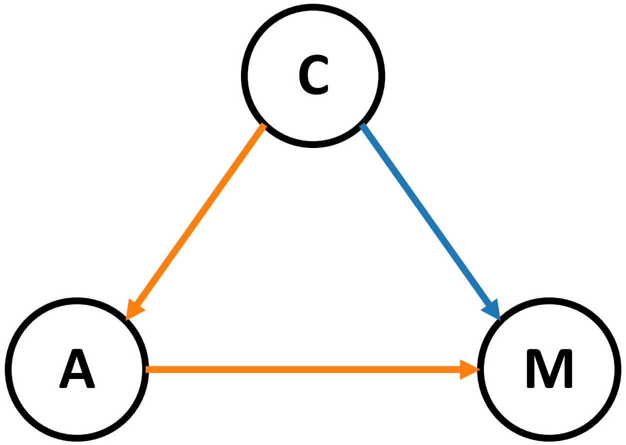
Mediation analysis

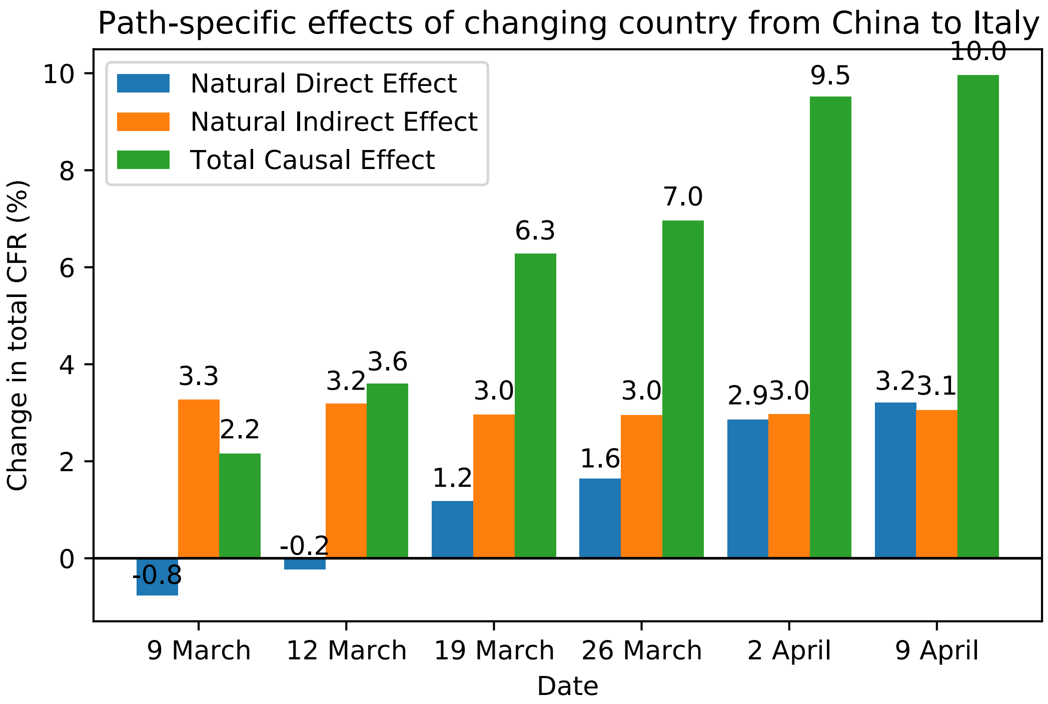
Only for linear models can **total causal effect (TCE)** be decomposed into direct effect (DE) and indirect effect (IE),

$$\text{TCE} = \text{DE} + \text{IE}$$

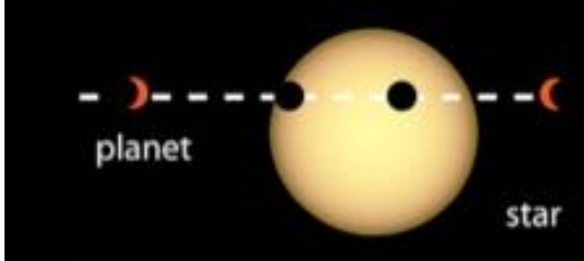
Due to interactions, DE and IE are *not uniquely defined in general*, but depend on the state of the mediator.

- **Natural Direct Effect (NDE):** case demographic kept as in China while CFRs per age group changed to those in Italy.
- **Natural Indirect Effect (NIE):** CFRs per age group kept as in China, while case demographic changed to that in Italy.

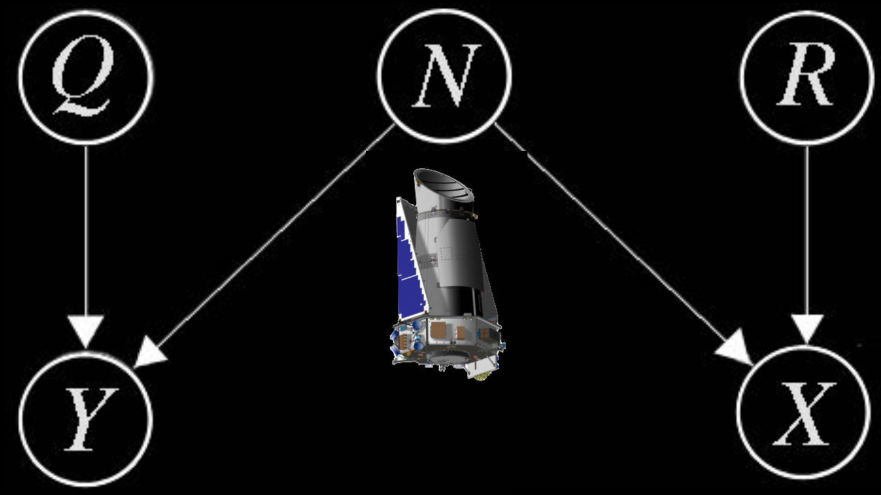


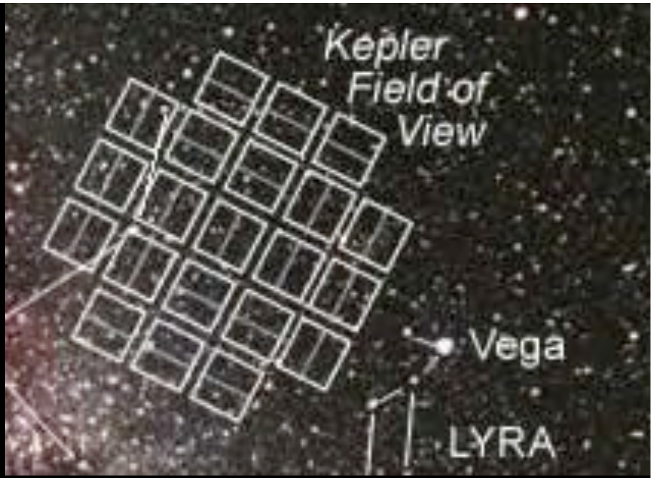

Path-specific effects of changing country from China to Italy

Causal models for exoplanet detection

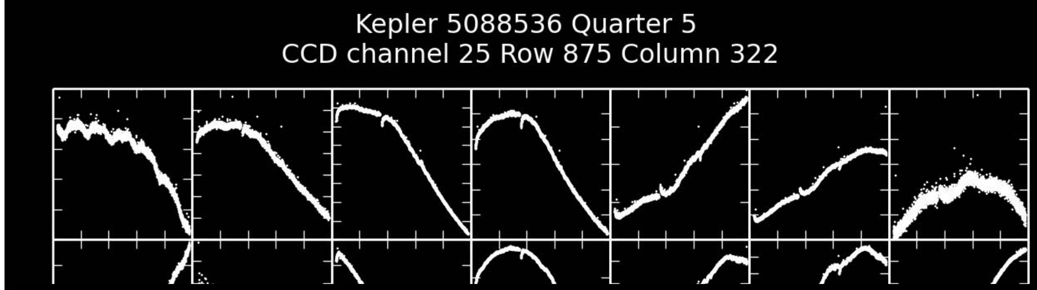


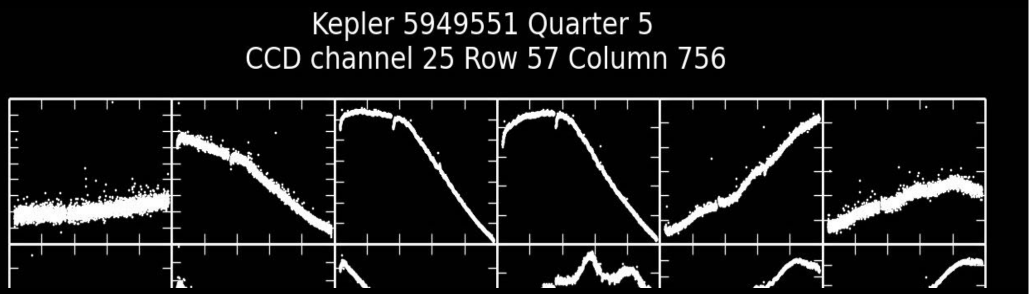
ICML 2015
Astrophysical Journal 2015
PNAS 2016





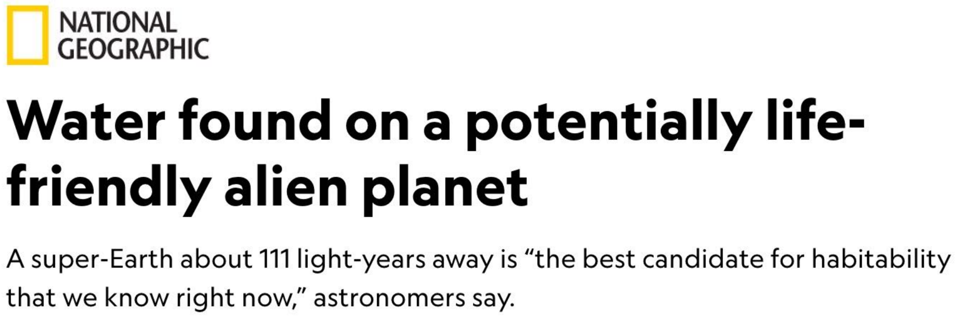
$$Q - E[Q] = Y - E[Y|X]$$





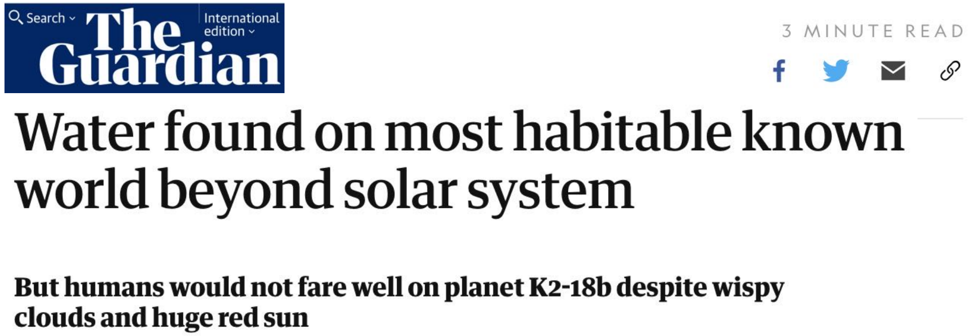
Discovered 21 new explanets. One of them received the name K2-18b.

The Milkman Problem



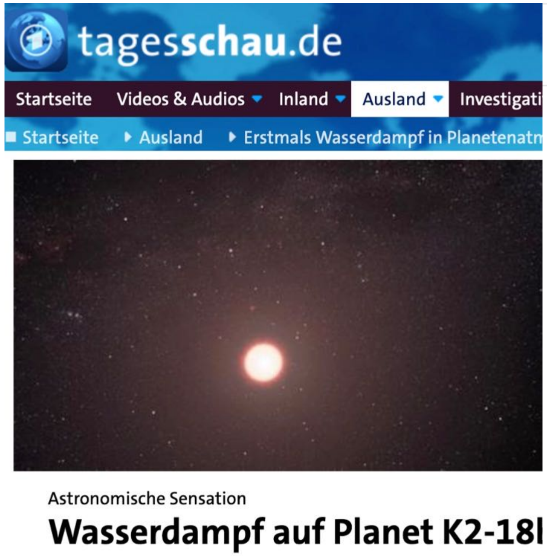
Water found on a potentially life-friendly alien planet

A super-Earth about 111 light-years away is “the best candidate for habitability that we know right now,” astronomers say.



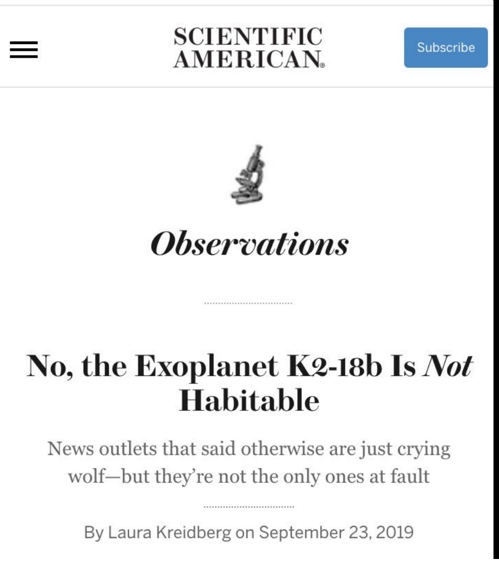
Water found on most habitable known world beyond solar system

But humans would not fare well on planet K2-18b despite wispy clouds and huge red sun



Astronomische Sensation

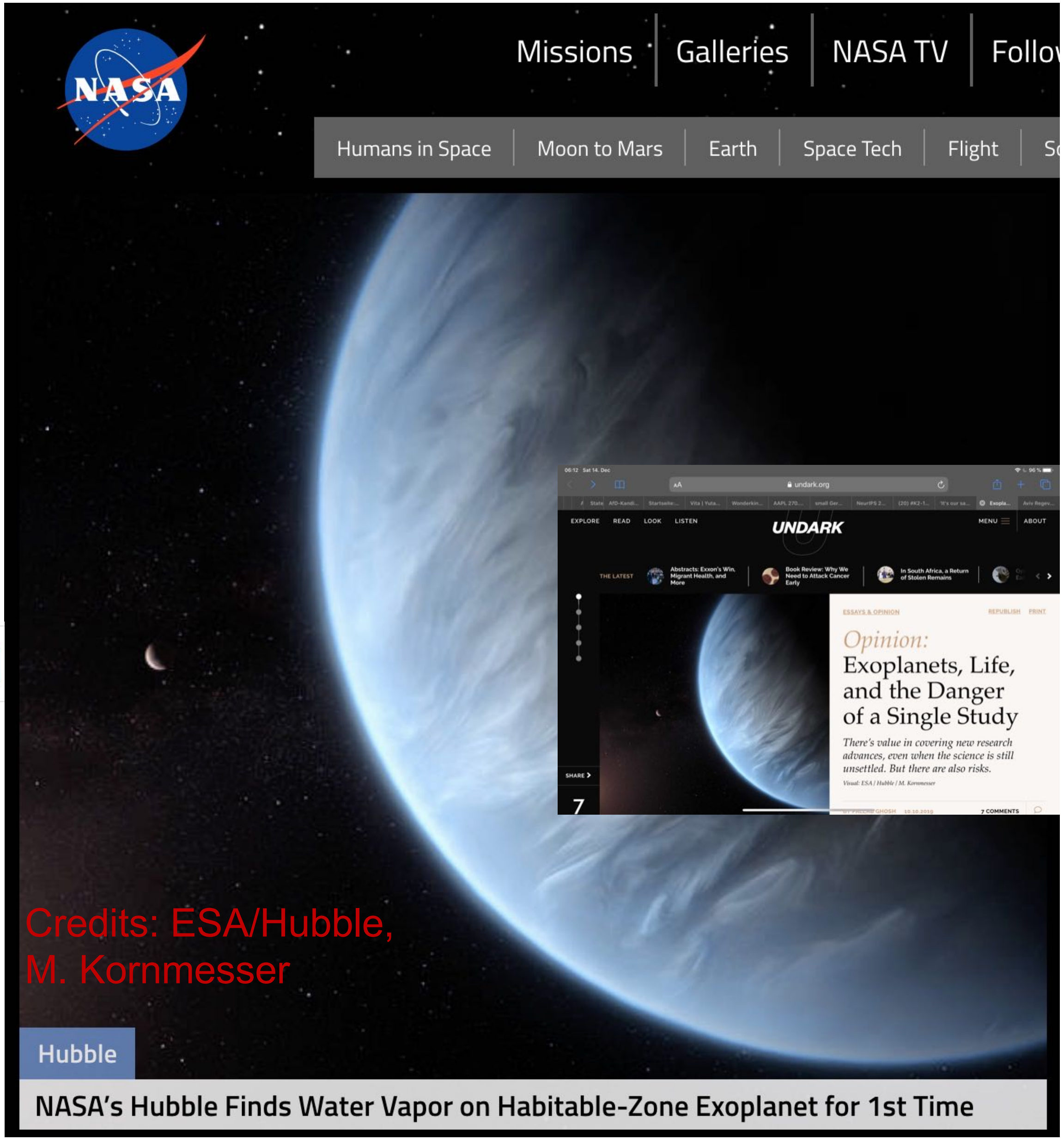
Wasserdampf auf Planet K2-18l



No, the Exoplanet K2-18b Is *Not* Habitable

News outlets that said otherwise are just crying wolf—but they’re not the only ones at fault

By Laura Kreidberg on September 23, 2019



Opinion: Exoplanets, Life, and the Danger of a Single Study

There's value in covering new research advances, even when the science is still unsettled. But there are also risks.

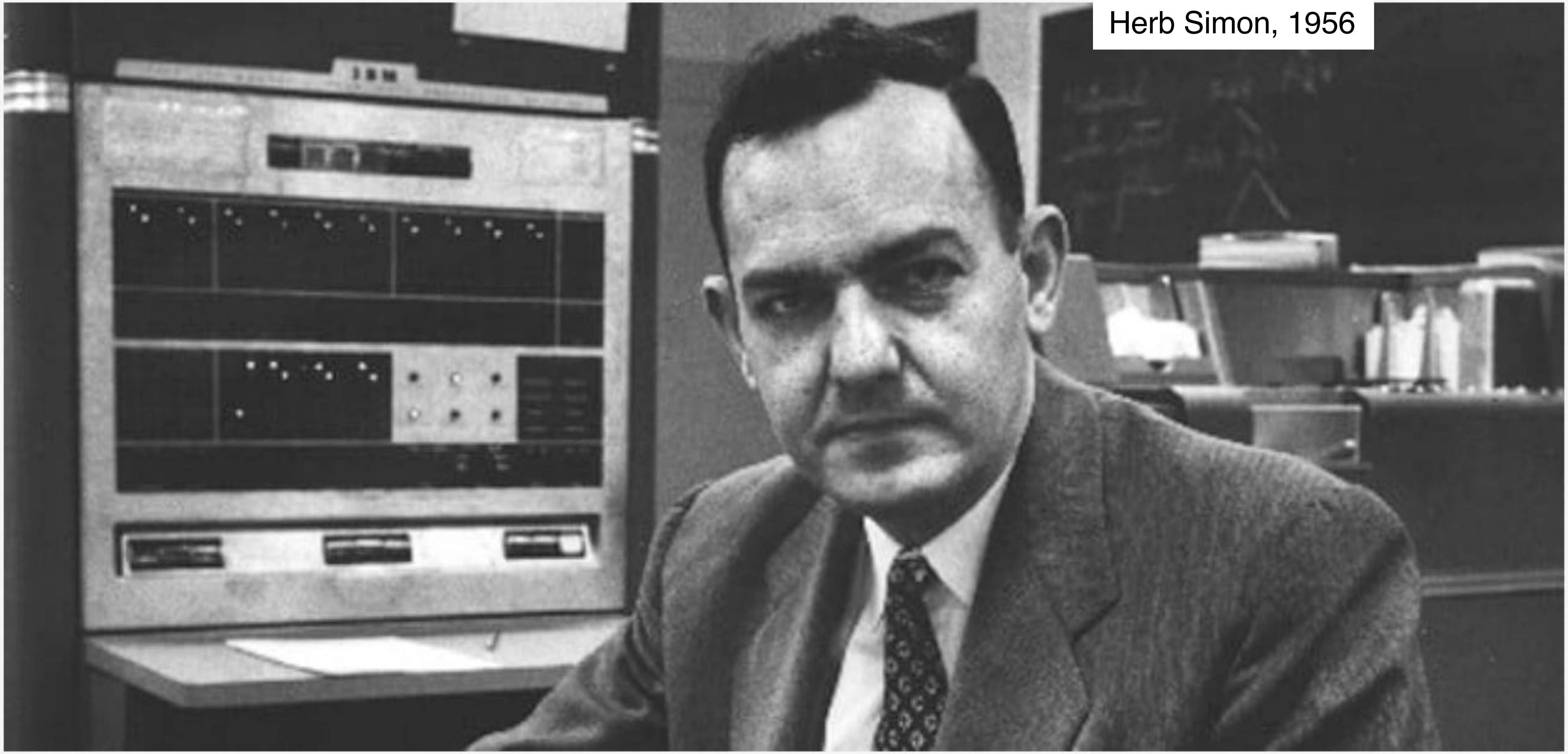
Visual: ESA / Hubble / M. Kornmesser

10.10.2019

Credits: ESA/Hubble, M. Kornmesser

Hubble

NASA’s Hubble Finds Water Vapor on Habitable-Zone Exoplanet for 1st Time



Herb Simon, 1956

“Machines will be capable, within twenty years, of doing any work a man can do”

Toward causal representation learning

Core Problem of Statistical Representations: Representation learning only includes *statistical* information — it does not capture interventions, reasoning, planning.

Core Problem of Causal Representations: SCMs are usually at the *symbolic* level — they assume the causal variables are given.

https://arxiv.org/abs/2102.11107

Independent mechanisms and the disentangled factorization

Factorization

- **independent noises** in the causal graph:

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

Independent mechanisms and the disentangled factorization

Disentangled (causal) factorization

https://arxiv.org/abs/1911.10500
https://arxiv.org/abs/2102.11107

- **independent noises** in the causal graph:

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

- **independent mechanisms**: 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**

(Janzing & Schölkopf, IEEE Trans. Inf. Th. 2010; Schölkopf et al., ICML 2012),

cf. *autonomy, (structural) invariance, separability, exogeneity, stability, modularity: (Aldrich, 1989; Pearl, 2009)*

Special case: If the graph has no edges, disentanglement reduces to statistical independence:

$$p(X_1,\dots,X_n) = \prod_{I=1}^{n} p\left(X_i\right)$$

In general, the causal factors will not be statistically independent, and independence-based methods struggle to find them (Träuble et al., ICML 2021)

Entangled factorizations

Disentangled (causal) factorization

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

Entangled (non-causal) factorizations

e.g.,

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

- cannot intervene on $p(X_i \mid X_{i+1}, \ldots, X_n)$
- changing one $p\left(X_i \mid \mathrm{PA}_i\right)$ will usually change **many** of the $p(X_i \mid X_{i+1}, \ldots, X_n)$

https://arxiv.org/abs/1911.10500

Causal viewpoint on distribution shift

Disentangled causal factorization

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

with independent mechanisms $p\left(X_i \mid \mathrm{PA}_i\right)$.

Sparse Mechanism Shift Hypothesis: small distribution changes manifest themselves sparsely in the disentangled factorization, i.e., they should usually not affect all factors simultaneously.

Here, a shift can be passive (e.g., distribution drift) or active (intervention, action).

Stated in *(Parascandolo et al., arXiv:1712.00961 (2017); Bengio et al., arXiv:1901.10912 (2019), Schölkopf, arXiv:1911:10500 (2019))*; see also *(Schölkopf et al., ICML 2012, Schölkopf, Janzing, Lopez-Paz 2016, Zhang et al., ICML 2013, Huang, Zhang et al., JMLR 2020)*

https://arxiv.org/abs/1911.10500

Causal training

Sparse mechanism shift training: require that across domain shifts or actions/interventions, only a sparse set of causal representation factors changes.

ICM training: encourage independence of mechanisms

Counterfactual training: require that interventions produce valid images (e.g., after reconstruction in an autoencoder).

For *interventions on the* U_i, this encourages statistical independence (cf. standard disentanglement).

For *interventions on the* S_i, this encourages the mechanisms f_i to be independently manipulable.

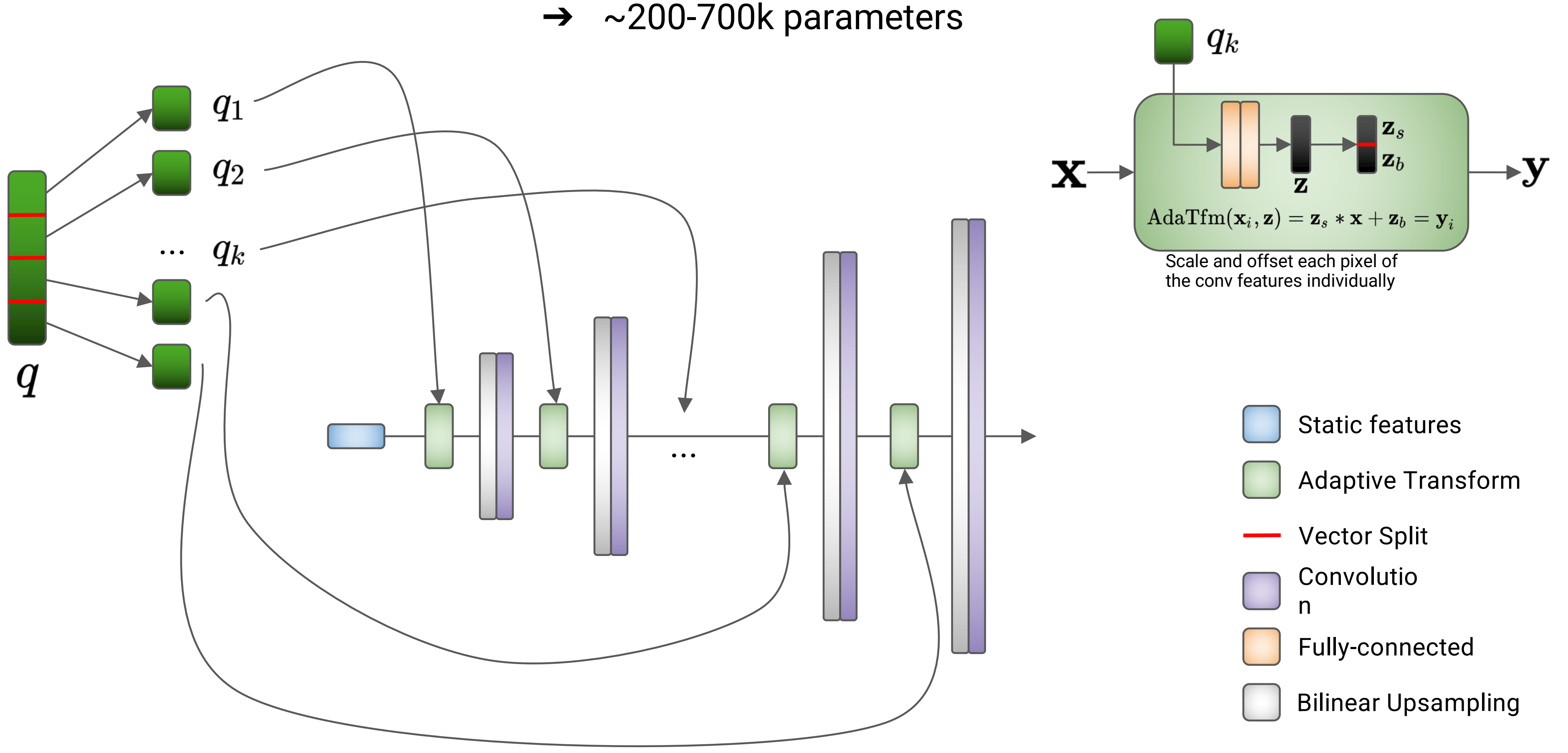
Structural training: embed SCM structure into decoder architecture and train by reconstruction error

Structural Decoders

Leeb et al.

arXiv 2006.07796

➔ ~200-700k parameters

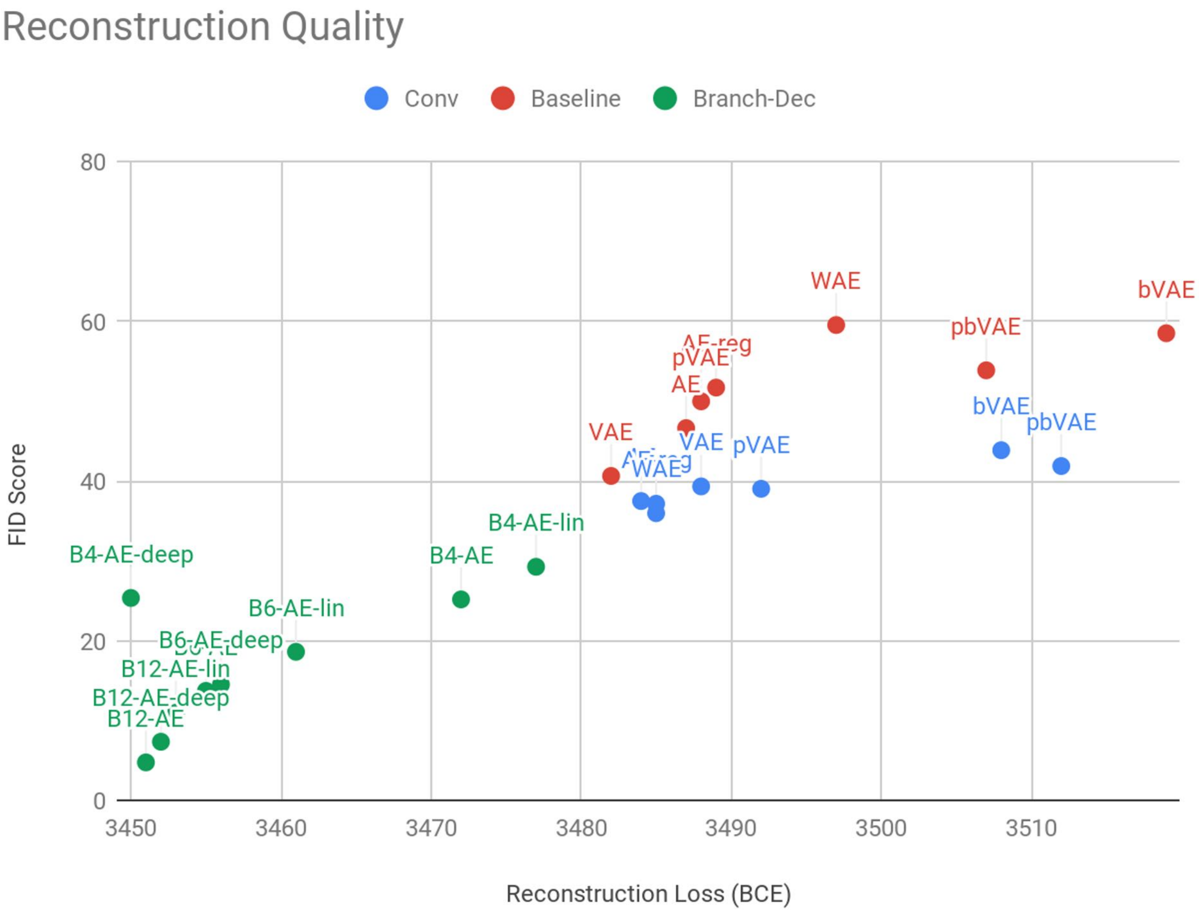


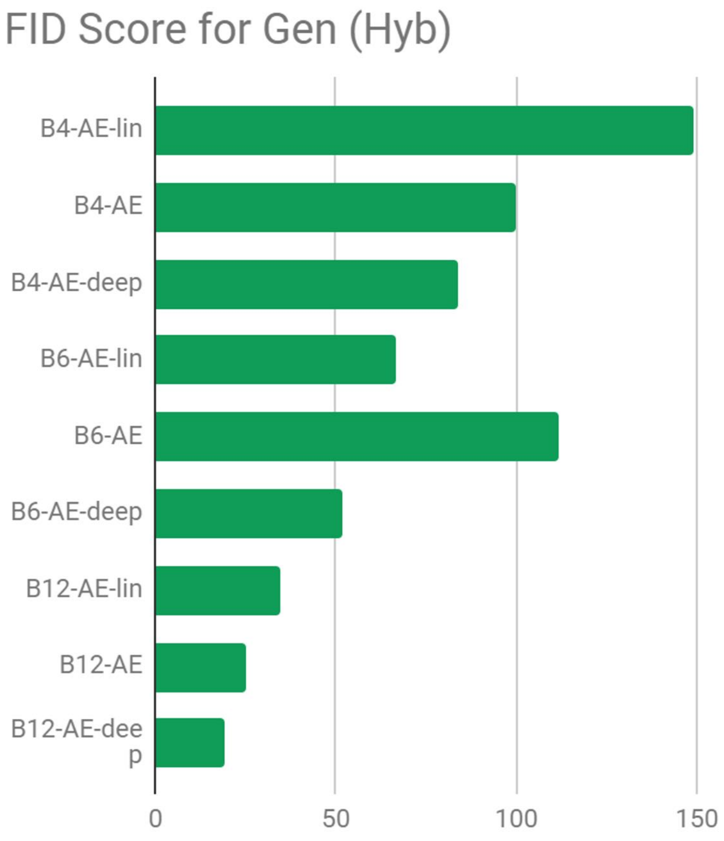
$\mathrm{AdaTfm}(\mathbf{x}_i, \mathbf{z}) = \mathbf{z}_s * \mathbf{x} + \mathbf{z}_b = \mathbf{y}_i$

Scale and offset each pixel of the conv features individually

- 🟦 Static features
- 🟩 Adaptive Transform
- 🟥 Vector Split
- 🟪 Convolutio n
- 🟧 Fully-connected
- ⬜ Bilinear Upsampling

Quantitative Results

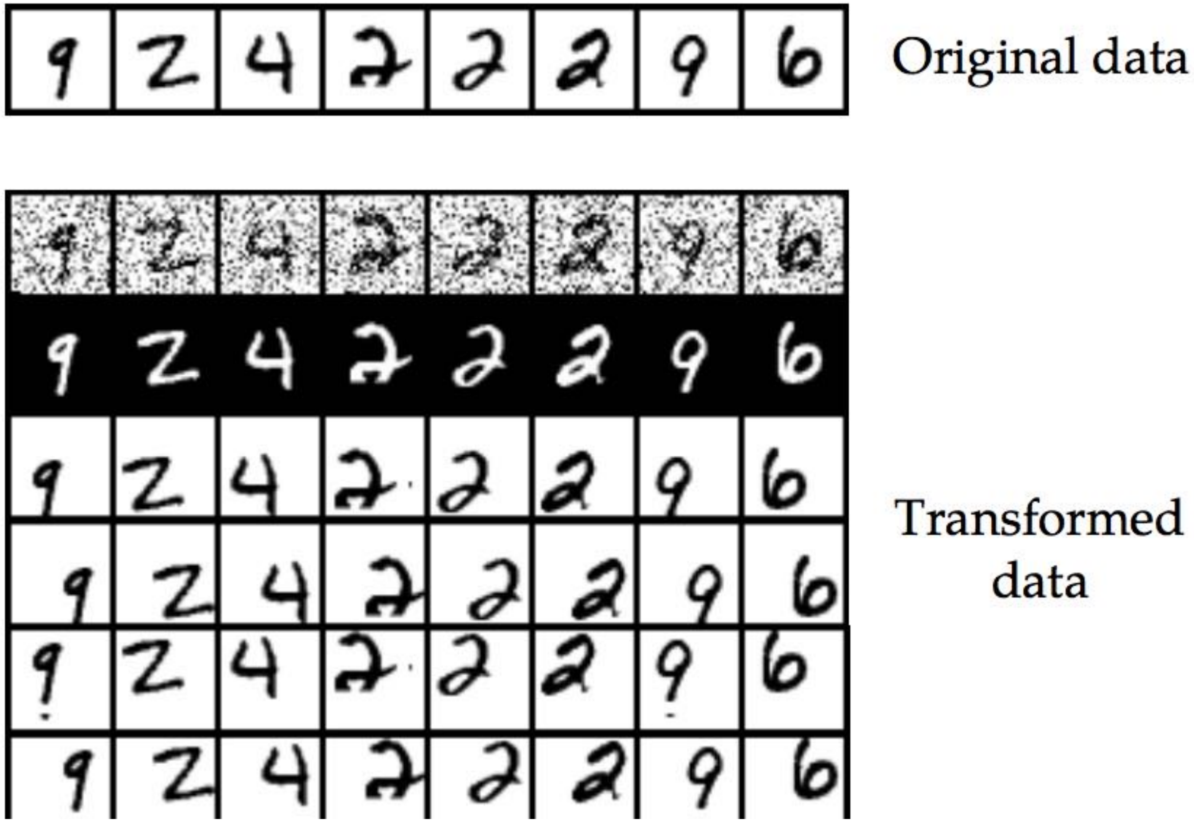

Reconstruction Quality


FID Score for Gen (Hyb)

Learning independent mechanisms

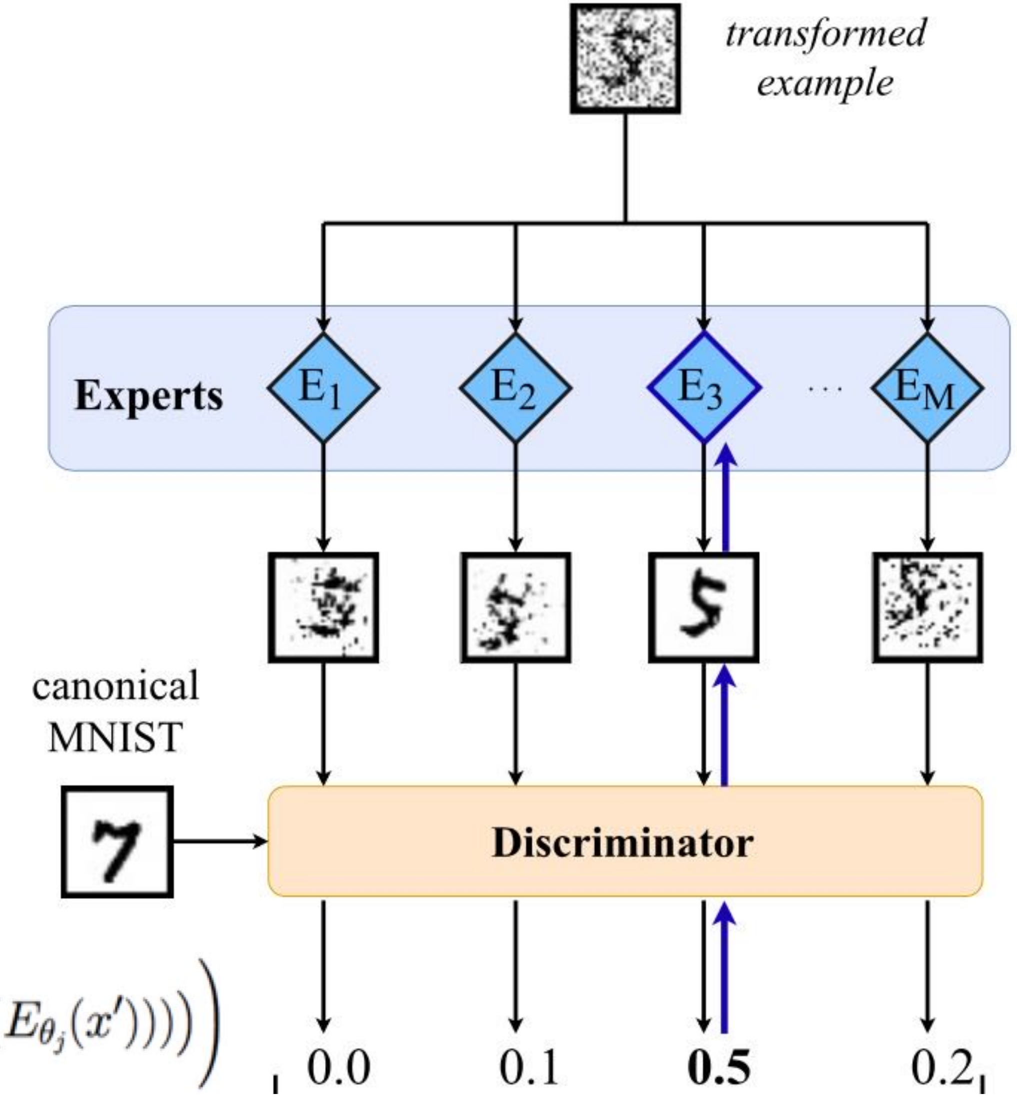
(with Parascandolo, Kilbertus, Rojas-Carulla, ICML 2018)

- 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



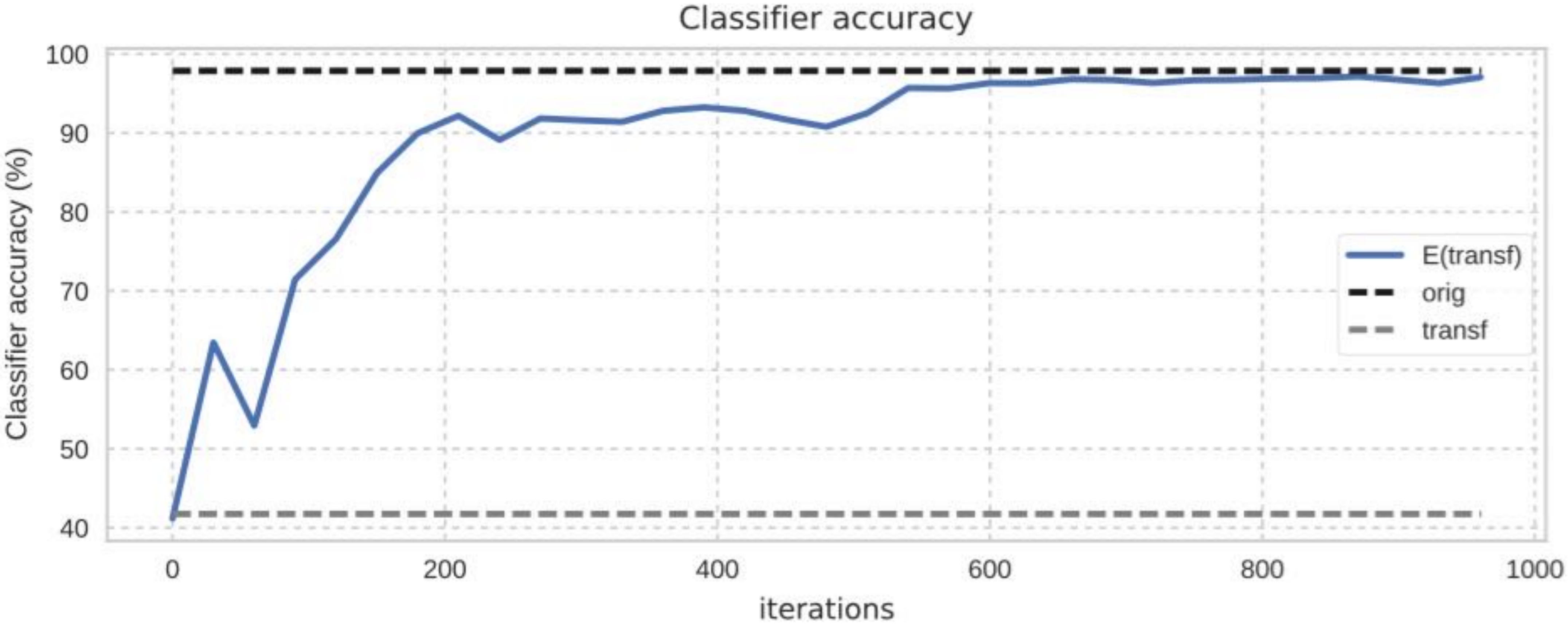
Method

- Mechanisms initialized $\approx$ 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

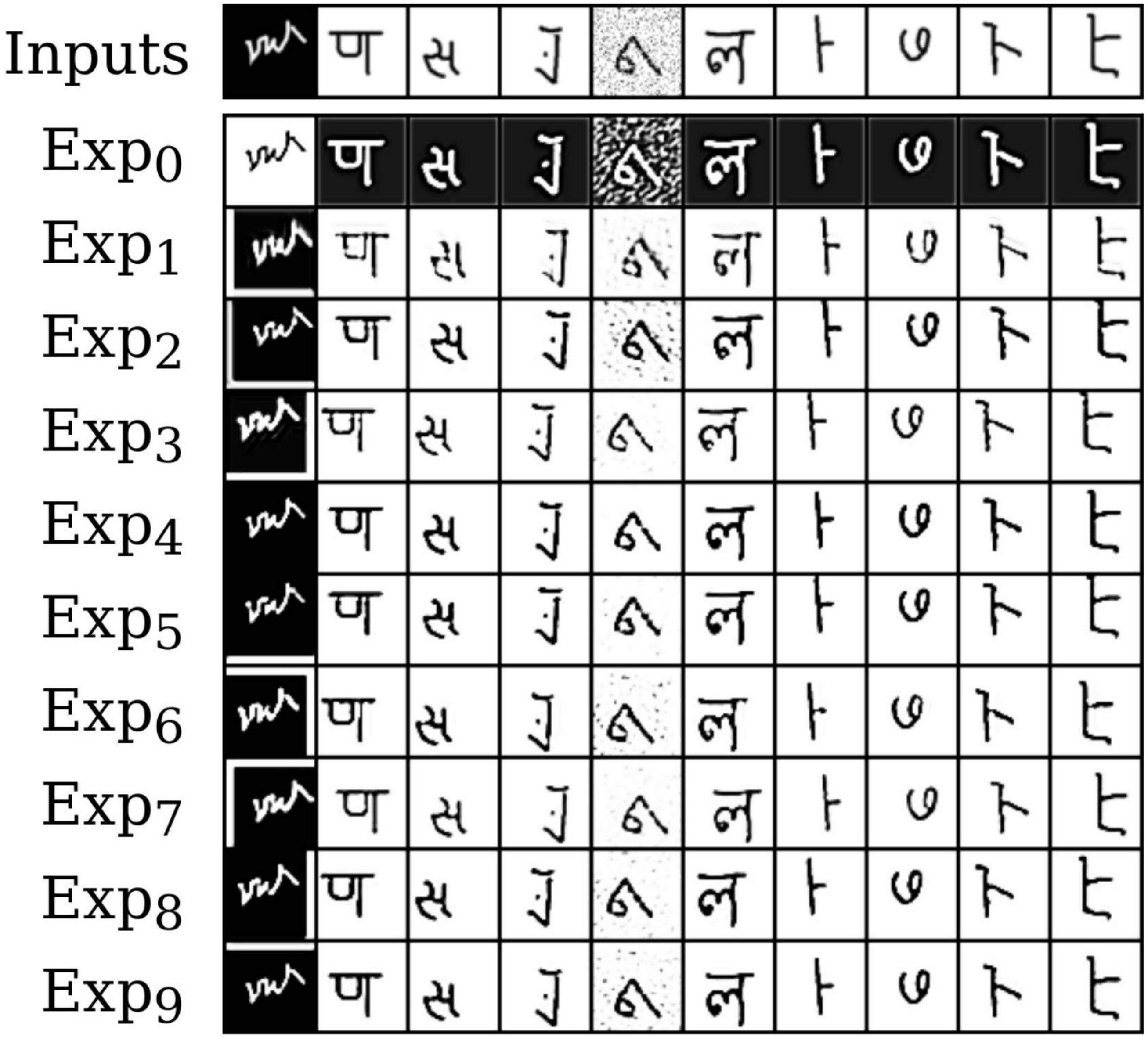


$$\max_{\theta_D}\left(\mathbb{E}_{x\sim P}\log(D_{\theta_D}(x)) + \frac{1}{N'}\sum_{j=1}^{N'}\mathbb{E}_{x'\sim Q}\left(\log(1-D_{\theta_D}(E_{\theta_j}(x')))\right)\right)$$

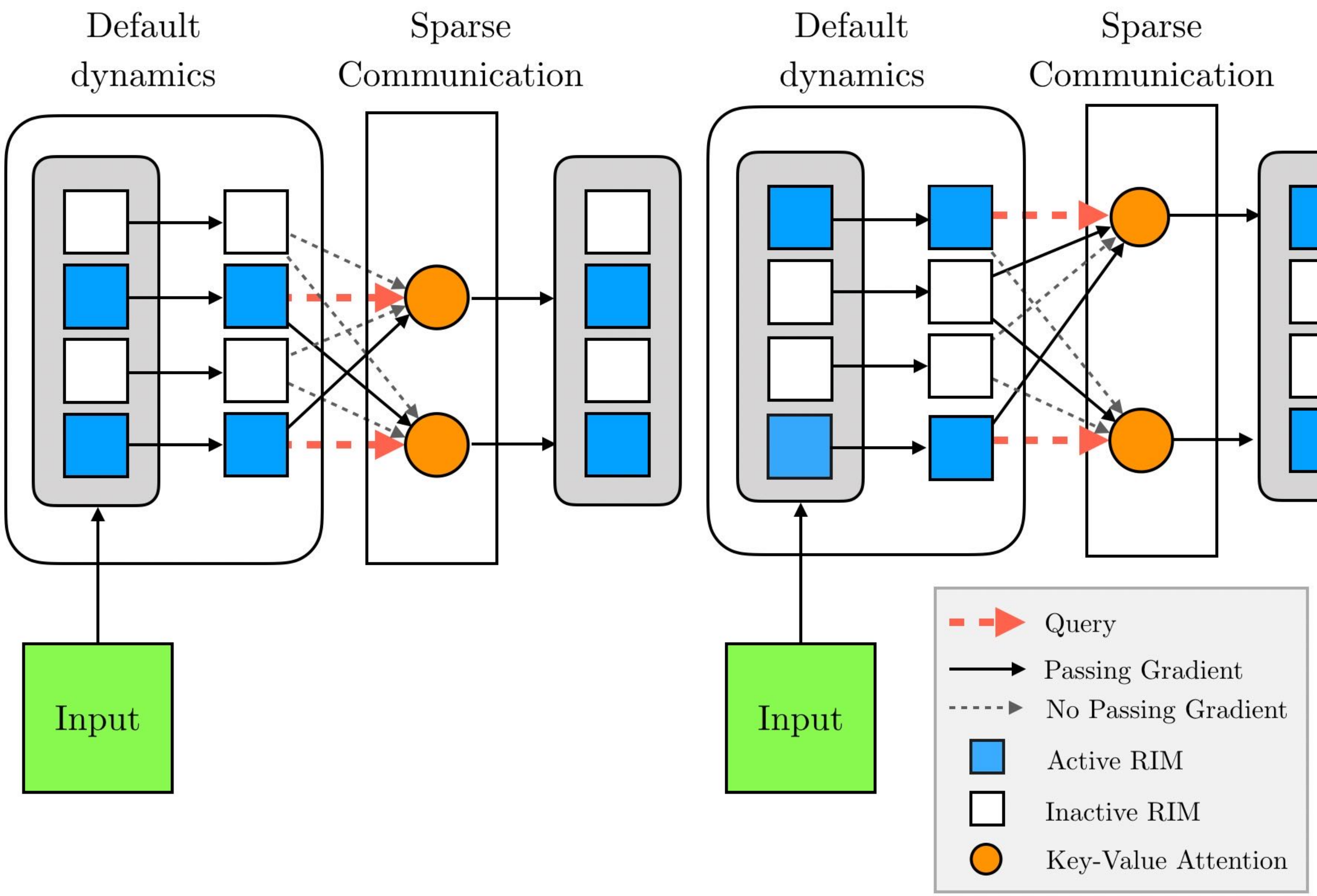
Accuracy of a CNN trained on MNIST for different test sets



Generalizing to Omniglot characters



Recurrent Independent Mechanisms

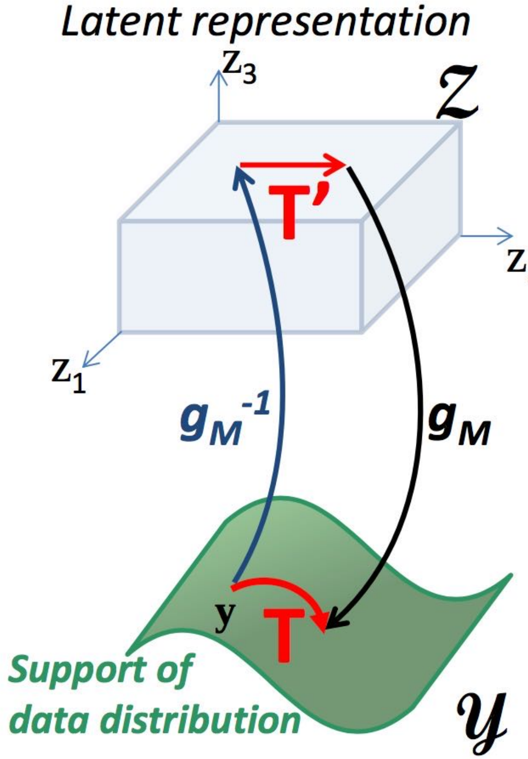


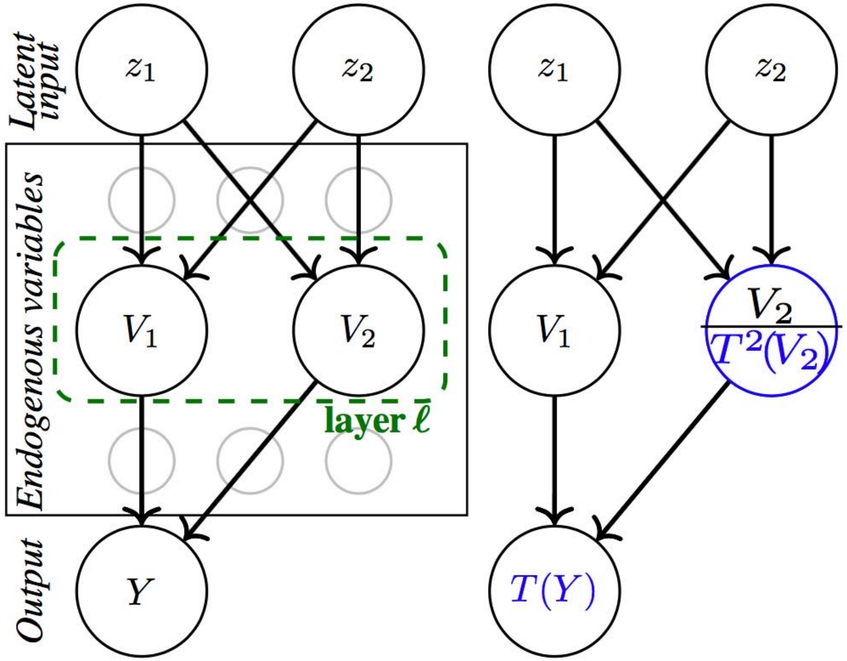
with **Anirudh Goyal**,
Alex Lamb,
Jordan Hoffmann,
Shagun Sodhani,
Sergey Levine,
Yoshua Bengio

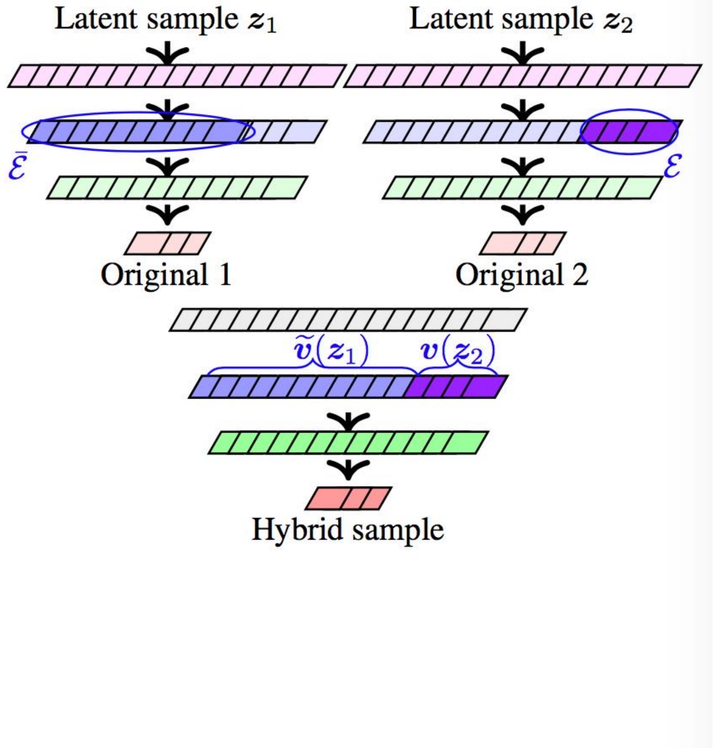
ICLR 2021

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

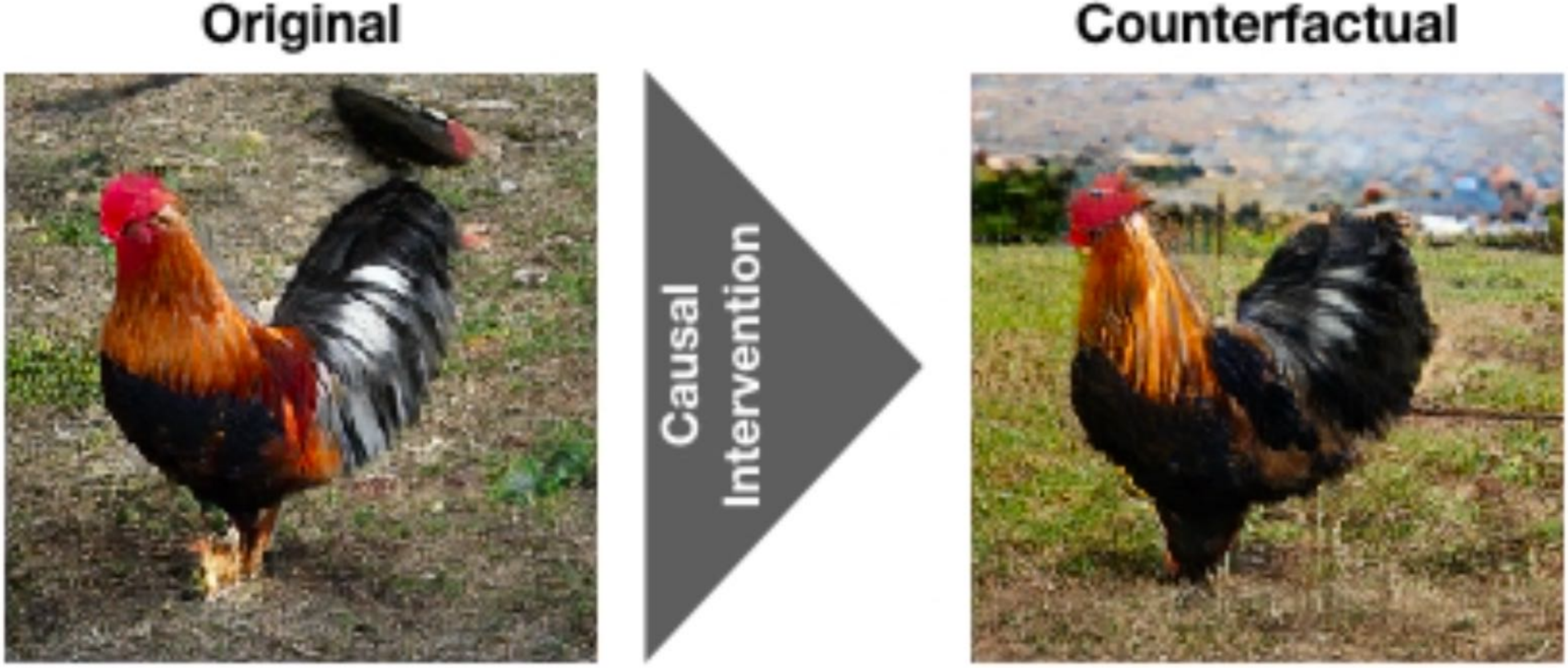
Interventional Representations *(Besserve et al., ICLR 2020)*



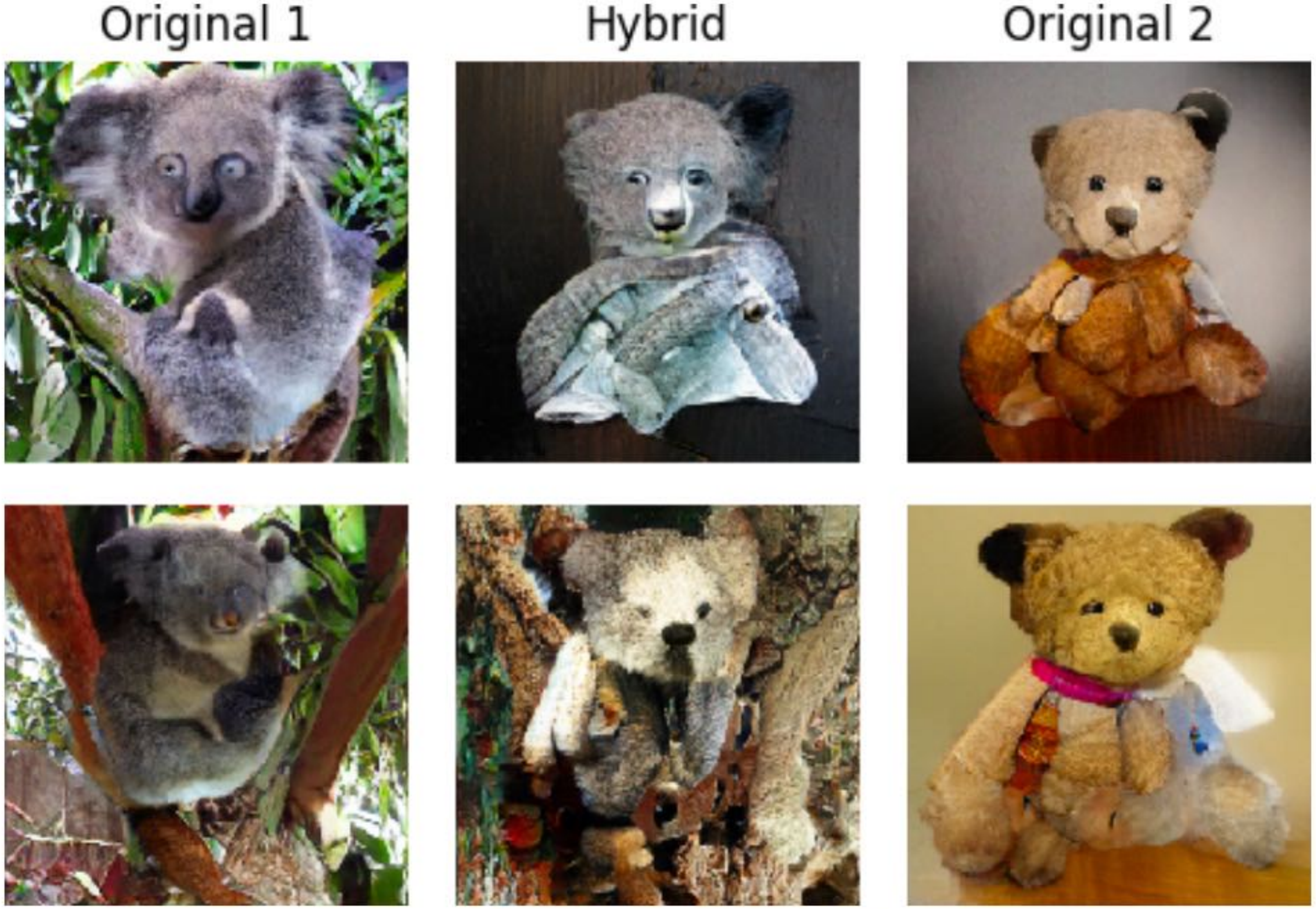




Interventional Representations *(Besserve et al., ICLR 2020)*



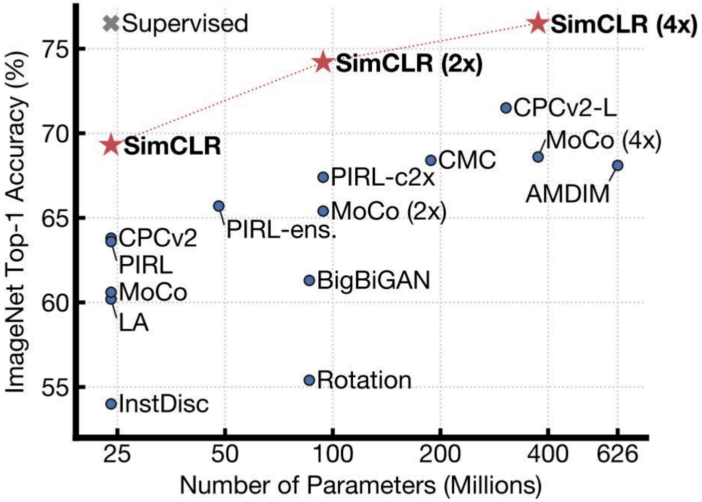
Interventional Representations *(Besserve et al., ICLR 2020)*



Self-supervised learning provably isolates content from style

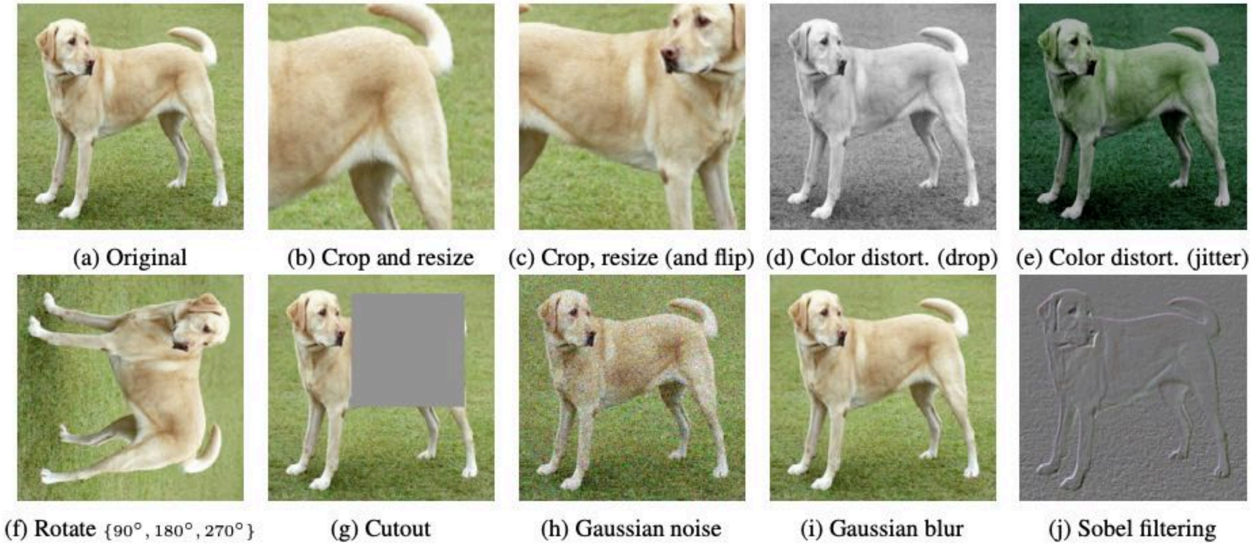
(https://arxiv.org/abs/2106.04619)

with **Julius von Kügelgen*, Yash Sharma*, Luigi Gresele*,** Wieland Brendel, Michel Besserve, Francesco Locatello



Data augmentation: very effective for self-supervised representation learning.

Useful because transformations are typically designed to leave *semantics* intact:



(a) Original (b) Crop and resize (c) Crop, resize (and flip) (d) Color distort. (drop) (e) Color distort. (jitter)

(f) Rotate {90°, 180°, 270°} (g) Cutout (h) Gaussian noise (i) Gaussian blur (j) Sobel filtering

Figures from:
SimCLR: A Simple Framework for Contrastive Learning of Visual Representations.
Chen, Kornblith, Norouzi, Hinton (ICML 2020; https://arxiv.org/abs/2002.05709)

Self-supervised learning with data augmentations provably isolates content from style

with **Julius von Kügelgen*, Yash Sharma*, Luigi Gresele***, Wieland Brendel, Michel Besserve, Francesco Locatello

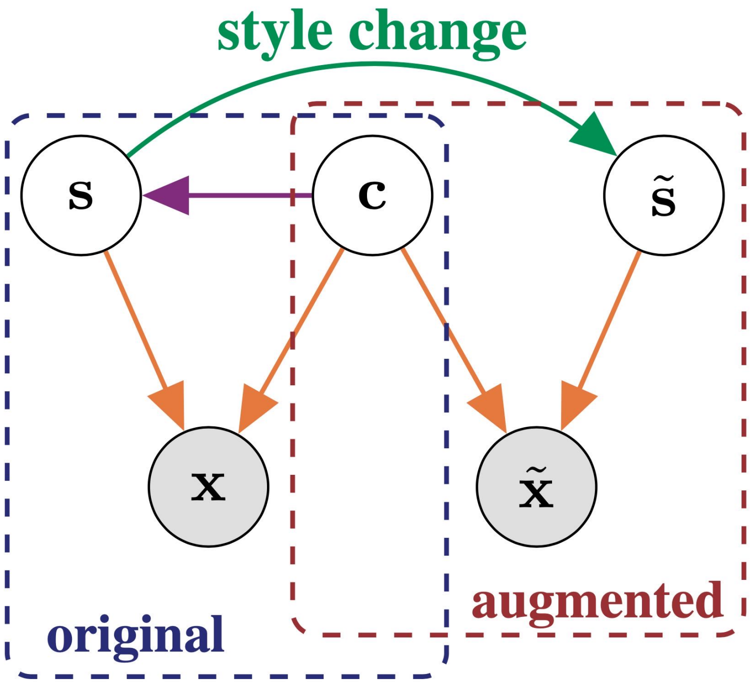
Formalise generation $\boldsymbol{x} = f(\boldsymbol{z})$ and augmentation $\tilde{\boldsymbol{x}} = f(\tilde{\boldsymbol{z}})$ processes as latent variable model with a content-style partition $\boldsymbol{z} = (\boldsymbol{c}, \boldsymbol{s})$:

- *invariant content $\boldsymbol{c}$:* always shared between pairs $(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ of views;
- *varying style $\boldsymbol{s}$:* may change across pairs $(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ of views.

Allow causal dependence of style on content (*Causal3DIdent* dataset):

augmented view $\tilde{\boldsymbol{x}}$ = counterfactual under soft style intervention on $\boldsymbol{x}$.

Theory: Can identify* invariant content partition in generative and discriminative learning with entropy maximisation (e.g., SimCLR).



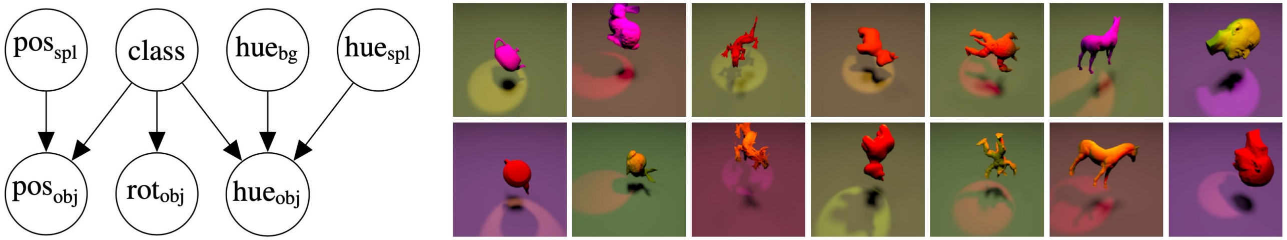


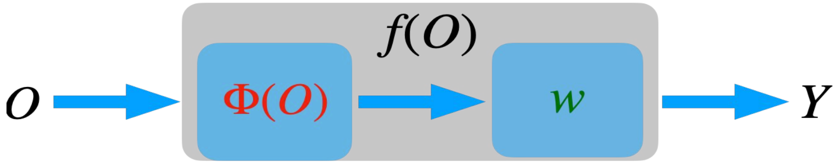
Figure 2: *(Left)* Causal graph for the *Causal3DIdent* dataset. *(Right)* Two samples from each object class.

*up to invertible transformation

Nonlinear Invariant Risk Minimization

(with Chaochao Lu, Yuhuai Wu, José Miguel Hernández-Lobato, arXiv:2102.12353)

Problem

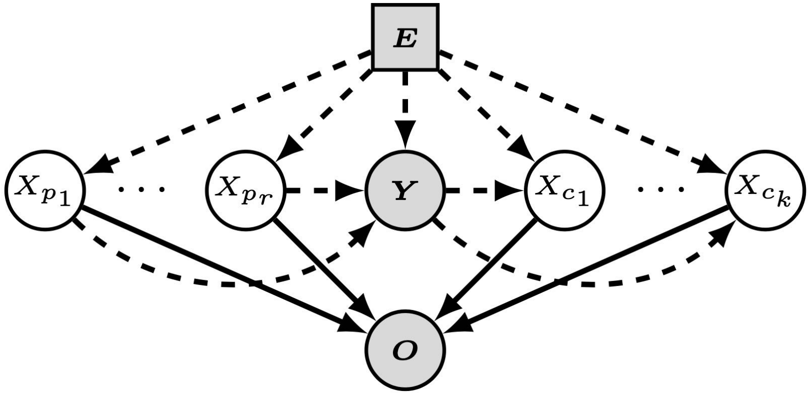


Nonlinear Data Representation

Nonlinear Invariant Classifier

Key Idea:
Data representation $\Phi(O)$ should be the direct cause of Y.

Assumption on Causal Graphs



This assumption is more general than the common Independence assumption in latent variable models.

Assumption on the Prior

$$P(X \mid Y,E) = P(X_{p_1},\ldots,X_{p_r} \mid Y,E)\prod_{i\in I_C} P(X_i \mid Y,E)$$

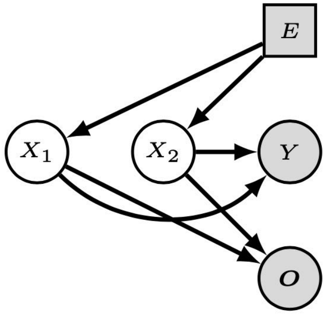
$$\Downarrow$$

$$p_{T,\lambda}(X \mid Y,E) = \frac{\mathcal{Q}(X)}{\mathcal{Z}(Y,E)}\cdot\exp\big(\langle T(X),\lambda(Y,E)\rangle\big)$$

The prior is assumed to be a general exponential family distribution leading to IDENTIFIABILITY.

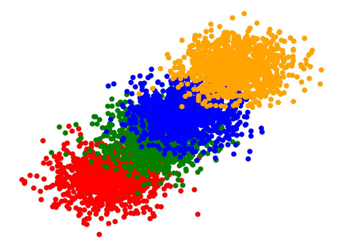
Experimental Results

(with Chaochao Lu, Yuhuai Wu, José Miguel Hernández-Lobato, arXiv:2102.12353)

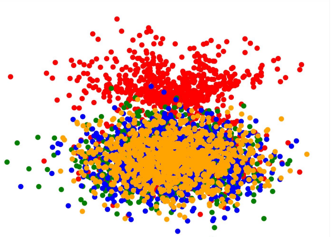


$$E \sim \mathcal{U}\{0.2, 2, 3, 5\}$$
$$X_1 \sim \mathcal{N}(X_1|E, 1)$$
$$X_2 \sim \mathcal{N}(X_2|2E, 2)$$
$$Y \sim \mathcal{N}(Y|X_1 + X_2, 1)$$
$$\boldsymbol{O} = g(X_1, X_2)$$

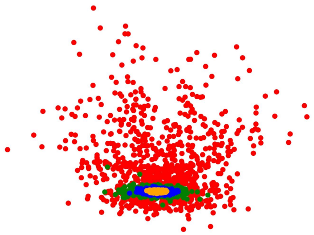
Data Generating Process



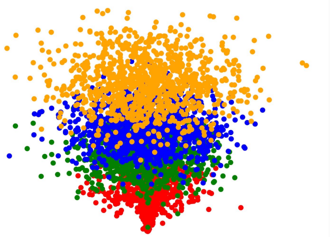
Original Data



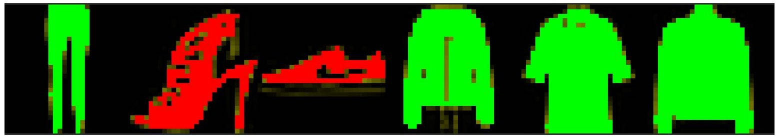
Samples from VAE
(Kingma et al. 2013)



Samples from iVAE
(Khemakhem et al. 2020)



Samples from iCaRL



Color	Red	Green
Y=0	p_e	$1 - p_e$
Y=1	$1 - p_e$	p_e

- **2 Training Envs:** $\{p_e = 0.1, p_e = 0.2\}$
- **1 Testing Env:** $\{p_e = 0.9\}$

Table 2: Colored Fashion MNIST. Comparisons in terms of accuracy (%) (mean ± std deviation).

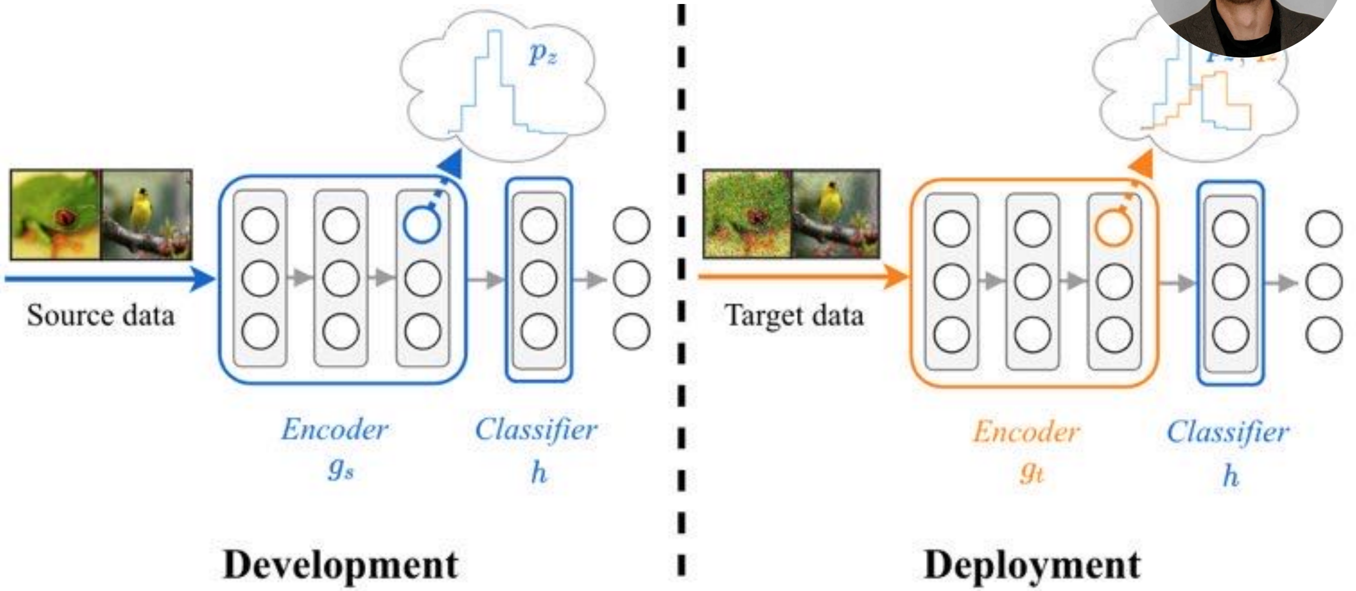
METHOD	TRAIN	TEST
ERM	83.17 ± 1.01	22.46 ± 0.68
ERM 1	81.33 ± 1.35	33.34 ± 8.85
ERM 2	84.39 ± 1.89	13.16 ± 0.82
ROBUST MIN MAX	82.81 ± 0.11	29.22 ± 8.56
F-IRM GAME	62.31 ± 2.35	69.25 ± 5.82
V-IRM GAME	68.96 ± 0.95	70.19 ± 1.47
IRM	75.01 ± 0.25	55.25 ± 12.42
iCaRL (ours)	$\mathbf{74.87 \pm 0.36}$	$\mathbf{73.56 \pm 0.75}$
ERM GRAYSCALE	74.79 ± 0.37	74.67 ± 0.48
OPTIMAL	75	75

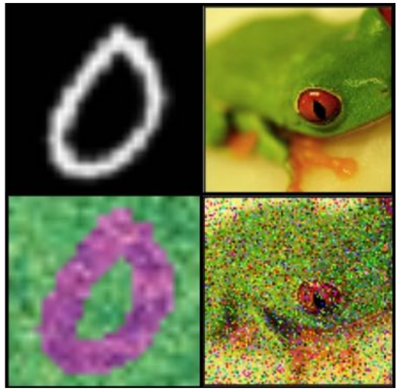
Source-Free Adaptation to Measurement Shift via Bottom-Up Feature Restoration (Cian Eastwood et al., https://arxiv.org/abs/2107.05446)

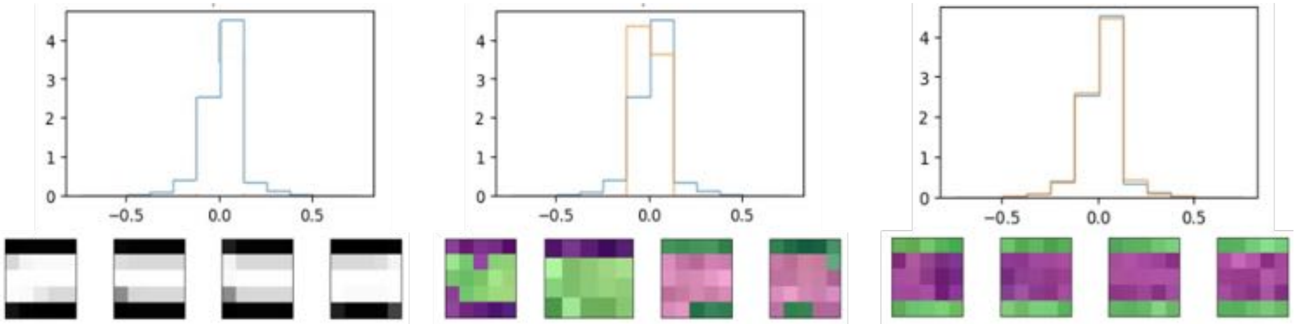
Source-free domain adaptation
-*Development:* train + equip model
-*Deployment:* adapt, no source data

Measurement shift (cf. Storkey, 2009)
-New sensor, same underlying features

Feature restoration
-**Goal:** extract same features, new env.
-**Method:** align (marginal) feature dists.

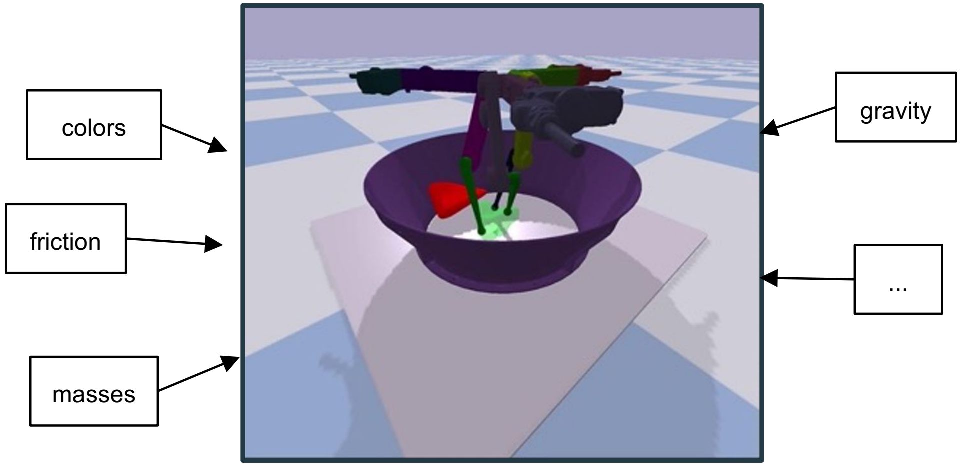






CausalWorld: A Robotic Manipulation Benchmark for Causal Structure and Transfer Learning

Ahmed and Träuble et al.,
arXiv: 2010.04296,
ICLR 2021



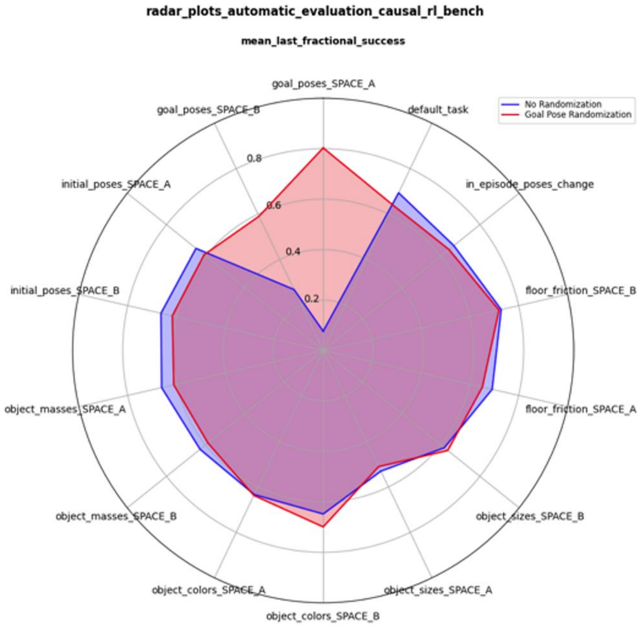
colors

friction

masses

gravity

...



Evaluate different generalization aspects by intervening on a large range of different defining variables of the hierarchical causal generative world model of the robotic environment.

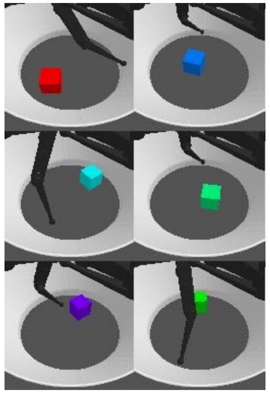
Benchmark with many challenging environments and fully documented code: https://github.com/rr-learning/CausalWorld

On the Transfer of Disentangled Representations in Realistic Settings

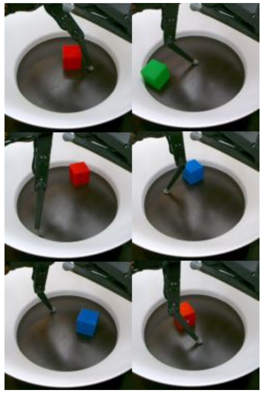
Dittadi and Träuble et al.,
arXiv: 2010.14407,
ICLR 2021

New Disentanglement Dataset

More complex and realistic, correlations between factors, occlusions, sim-to-real

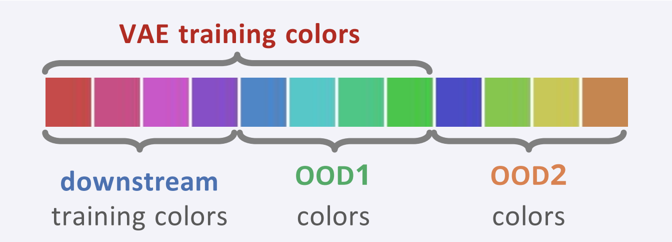


1 million simulated



1800 real (labeled)

Out-of-Distribution Generalization of Downstream Tasks



VAE training colors

downstream training colors

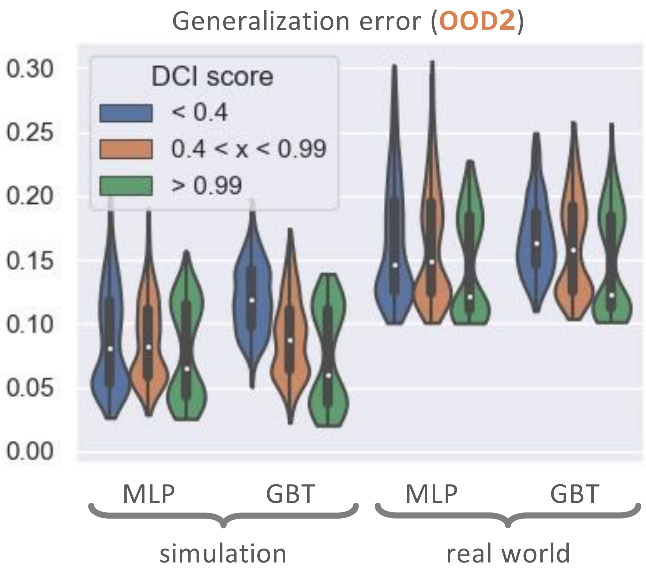
OOD1 colors

OOD2 colors

Task: **predict** value of non-OOD factors

- **Train downstream task** on **pre-trained representations**
- Test it OOD but still in the VAE's training distribution (**OOD1**)
- Test it OOD w.r.t. the VAE itself (**OOD2**)

Disentanglement has **minor role** when represent. function is OOD



Causal Curiosity: RL Agents Discovering Self-supervised Experiments for Causal Representation Learning

- Curiosity to discover causation in an environment.
- **Reward-free**
- Set of environments with interventions on causal factors
- Use Kolmogorov Complexity as reward to RL agent
- Agents producing self-supervised experiments to test out mass, size etc.

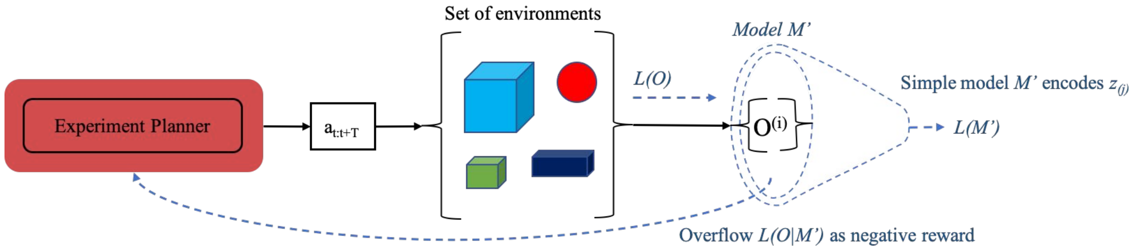


Fig 1: Experiment Discovery

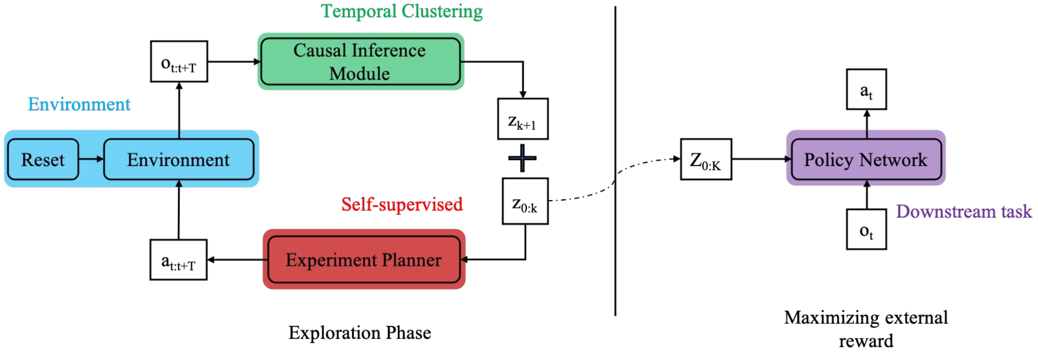
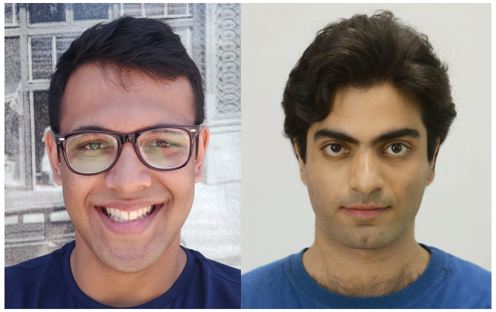
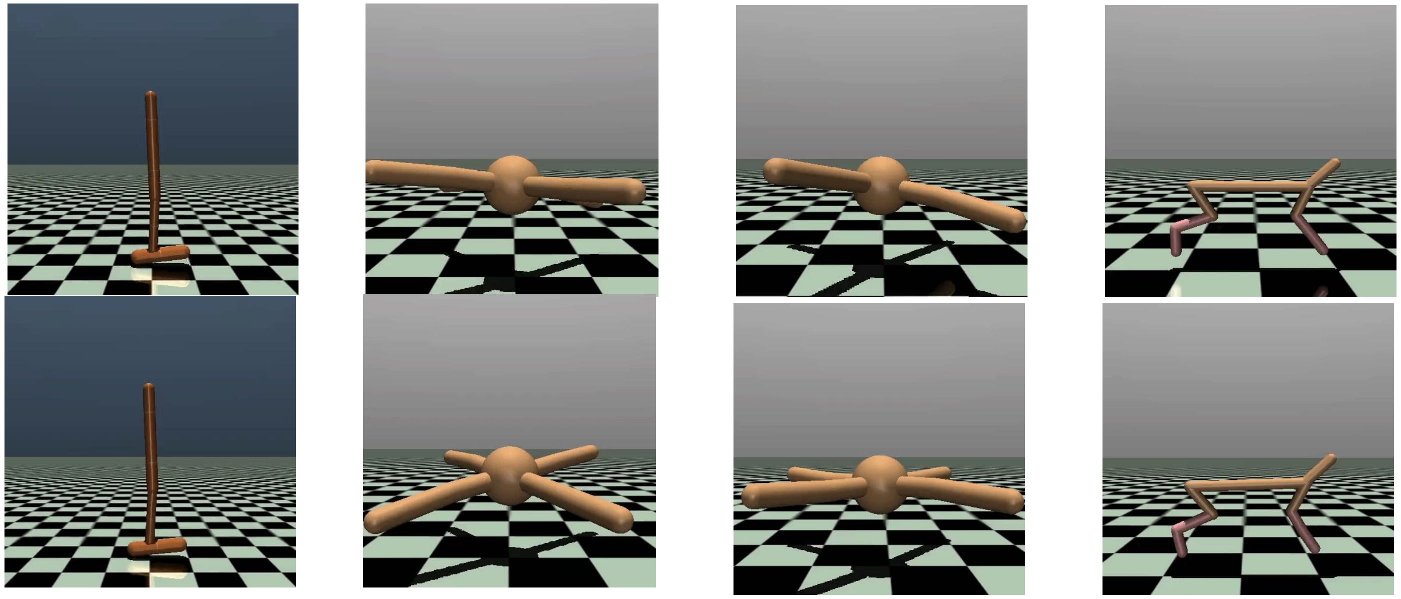


Fig 2: Performing experiments sequentially to learn causal representations. Representations used for downstream transfer.

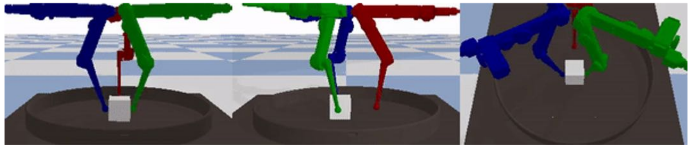


Sontakke, Sumedh A., Arash Mehrjou, Laurent Itti, and Bernhard Schölkopf. "Causal Curiosity: RL Agents Discovering Self-supervised Experiments for Causal Representation Learning." *arXiv preprint arXiv:2010.03110* (2020). To appear at *ICML 2021*

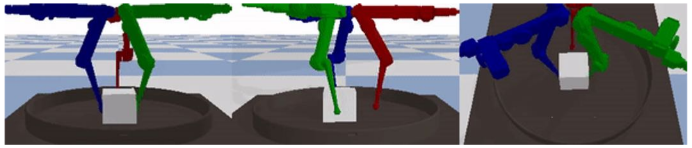
Discovered Behaviors - Mujoco

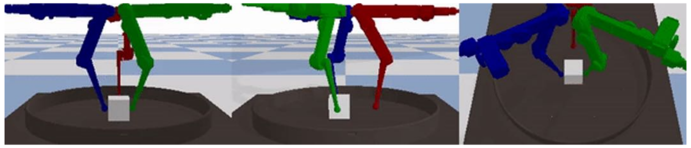


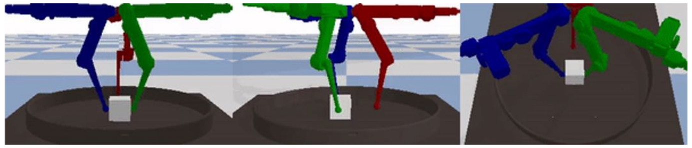
Discovered Behaviors - CausalWorld

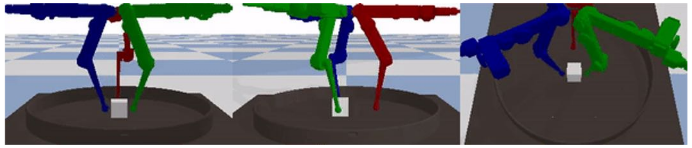


Lifting Behaviors

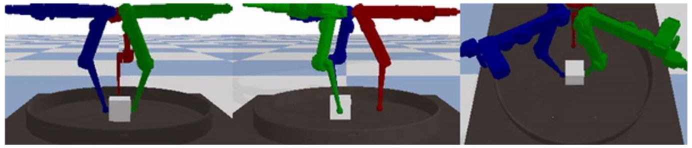


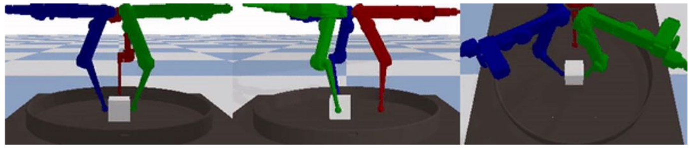




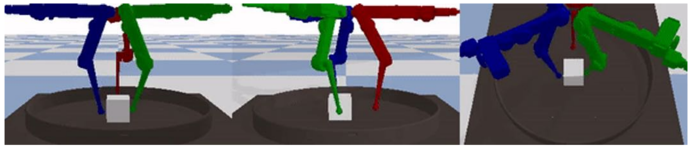


Rotate Behaviors

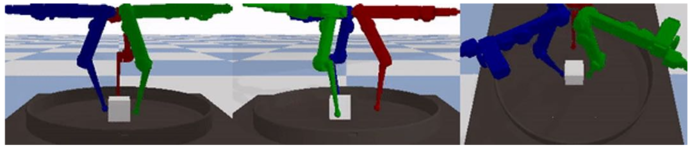




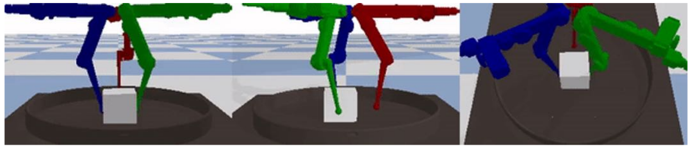
Discovered Behaviors - CausalWorld



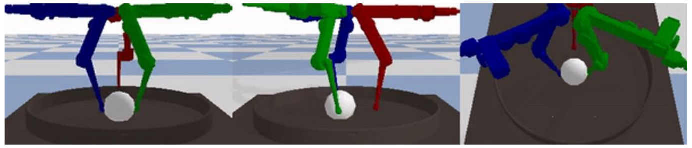
Dribble



Pushing along y



Pushing along x



Roll

Causal Influence Detection for Reinforcement Learning

(with Maximilian Seitzer and Georg Martius, arXiv:2106.03443)

Observations

- Real-world agents have limited interventional range
- Causal influence of agent on environment occurs only sparsely

Idea

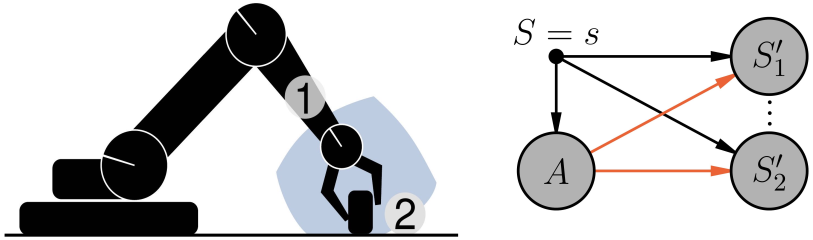
- Use causal influence to speed-up reinforcement learning

Method

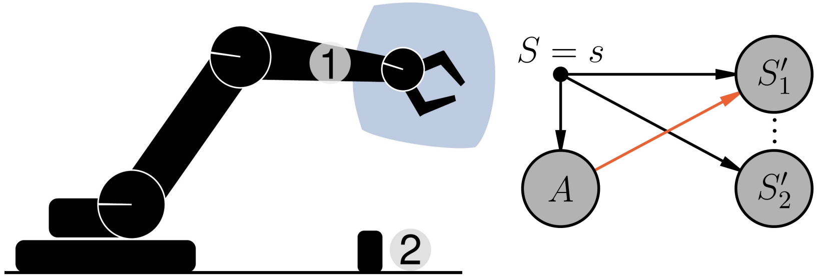
- Define measure of *causal action influence* as a conditional mutual information

$$C(s) := I(S', A \mid S = s)$$

- Estimate it from data using neural networks



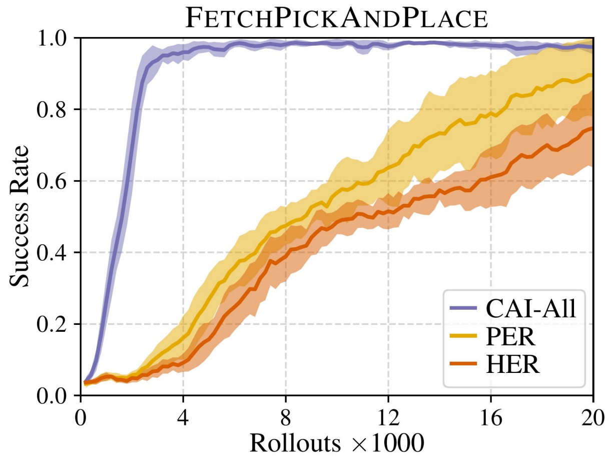
Robot can control object



Causal influence on object impossible

Causal Influence Detection for Reinforcement Learning

(with Maximilian Seitzer and Georg Martius, arXiv:2106.03443)



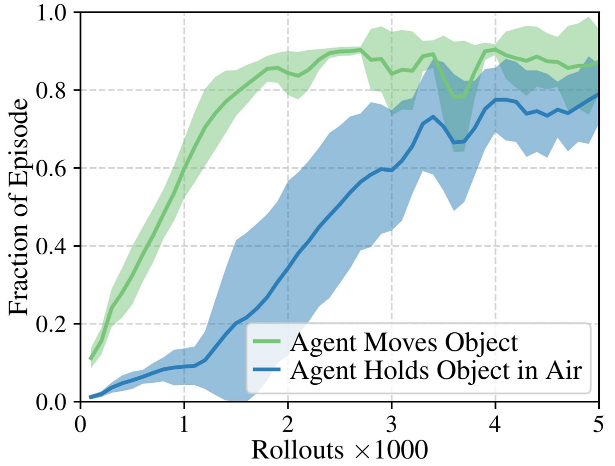
Results

- Focusing on states with causal influence (exploration and prioritization)

  ➢ Highly increased sample-efficiency on robotic manipulation tasks

- Maximizing causal influence as intrinsic motivation

  ➢ Agent quickly discovers interesting behaviors (grasping, lifting)

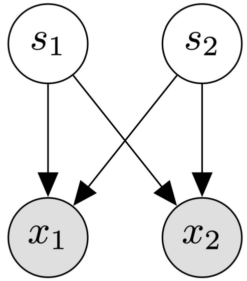


Brockmann et al. OpenAI Gym, arXiv:1606.01540

Causality for nonlinear ICA

(https://arxiv.org/abs/2106.05200)

with **Luigi Gresele*, Julius von Kügelgen***, Vincent Stimper, Michel Besserve



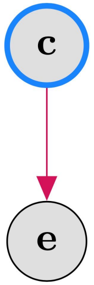
Observe: nonlinear mixtures, $\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{s})$, of independent sources $\boldsymbol{s}$

Goal: recover the unobserved sources (blind source separation)

Problem: impossible in general [Hyvärinen & Pajunen, '99]

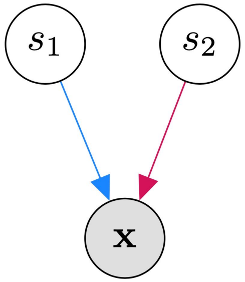
Recently: use auxiliary variables [Hyvärinen et al., '16, '17, '19]

New: interpret mixing as *causal* process & constrain $\boldsymbol{f}$ using the ICM principle



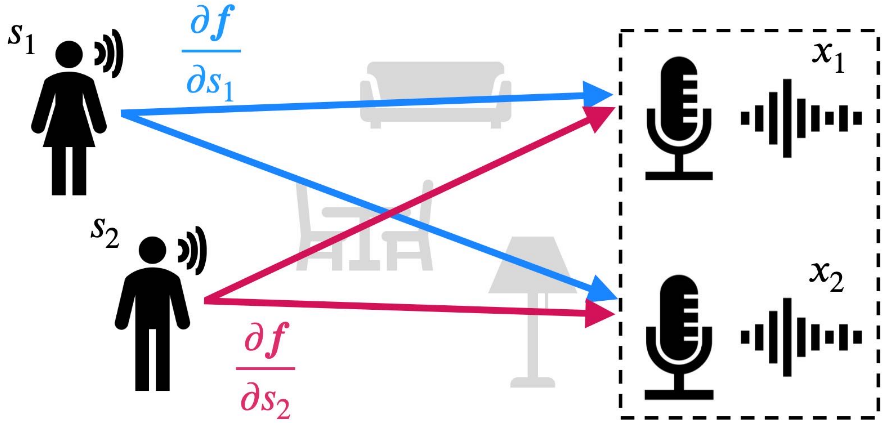
ICM usually applied to cause distribution p_c and mechanism $p_{e|c}$ (or $\boldsymbol{f}$), e.g., cause-effect discovery

But: in nonlinear ICA, cause (source distribution) is unobserved



Independent mechanism analysis (IMA):

- ICM at level of mixing function
- contributions $\frac{\partial f}{\partial s_i}$ of each source to observed distribution be "independent" (not statistical)
- speakers' positions not fine-tuned to room accoustics and microphone placement

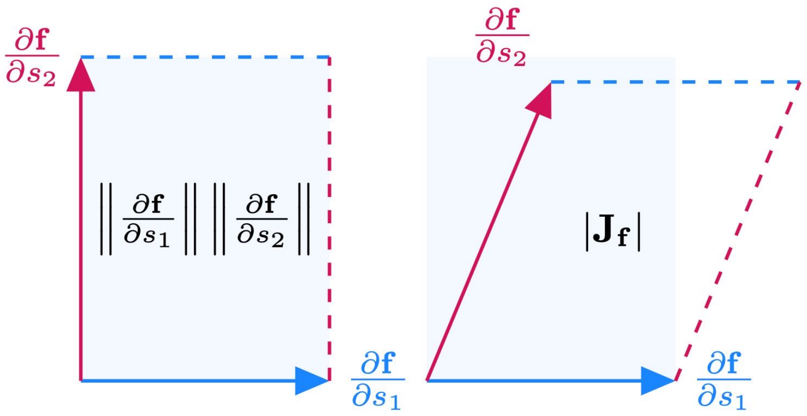

s_1
$\frac{\partial f}{\partial s_1}$
x_1
s_2
x_2
$\frac{\partial f}{\partial s_2}$

Independent mechanism analysis

with **Luigi Gresele*, Julius von Kügelgen*,** Vincent Stimper, Michel Besserve

<u>IMA Principle:</u> the influences of each source on the observed distribution are independent in the sense that:

$$\log |\mathbf{J_f}(\mathbf{s})| = \sum_{i=1}^{n} \log \left\| \frac{\partial \mathbf{f}}{\partial s_i}(\mathbf{s}) \right\|$$

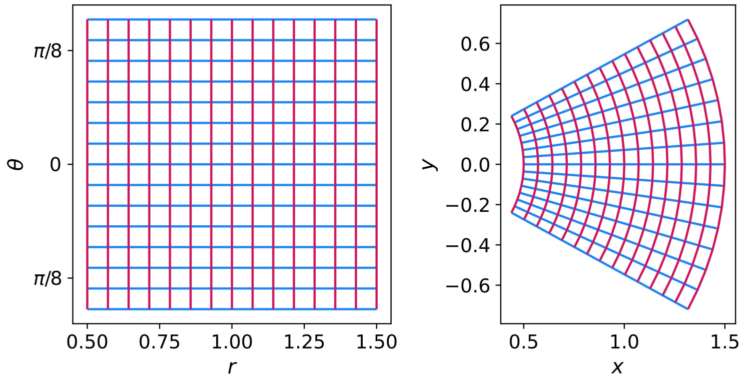

$\frac{\partial \mathbf{f}}{\partial s_2}$
$\left\| \frac{\partial \mathbf{f}}{\partial s_1} \right\| \left\| \frac{\partial \mathbf{f}}{\partial s_2} \right\|$
$\frac{\partial \mathbf{f}}{\partial s_1}$
$\frac{\partial \mathbf{f}}{\partial s_2}$
$|\mathbf{J_f}|$
$\frac{\partial \mathbf{f}}{\partial s_1}$

<u>Geometric interpretation:</u> corresponds to an *orthogonality condition* on the columns of the Jacobian.

<u>Contrast function:</u>

$$C_{IMA}(f, p_s) = \int \left(\sum_{i=1}^{n} \log \left\| \frac{\partial \boldsymbol{f}}{\partial s_i}(\boldsymbol{s}) \right\| - \log \left| \boldsymbol{J_f}(\boldsymbol{s}) \right| \right) \; p_{\boldsymbol{s}}(\boldsymbol{s}) d\boldsymbol{s}$$

- ≥ 0, with equality iff. $\boldsymbol{f}$ is an *orthogonal coordinate transformation*
- *invariant to reparametrisation* of the sources by *permutation* and *element-wise invertible nonlinearities*



Independent mechanism analysis

with **Luigi Gresele*, Julius von Kügelgen***, Vincent Stimper, Michel Besserve

Theory

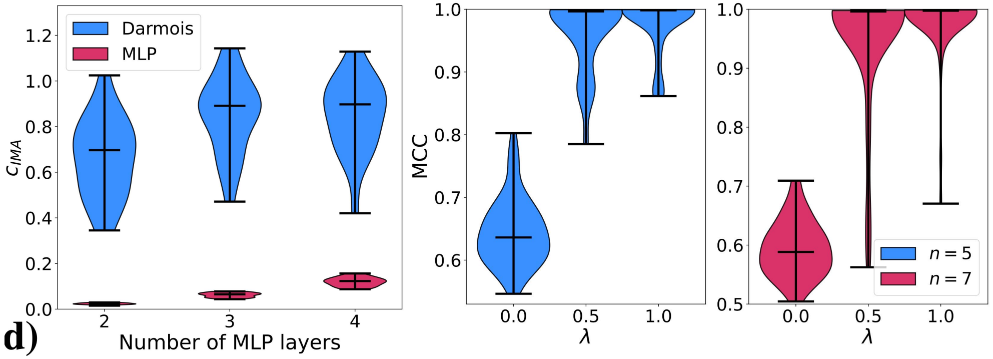
Can rule out (in the sense that C_{IMA} is larger for) well-known spurious ICA solutions:

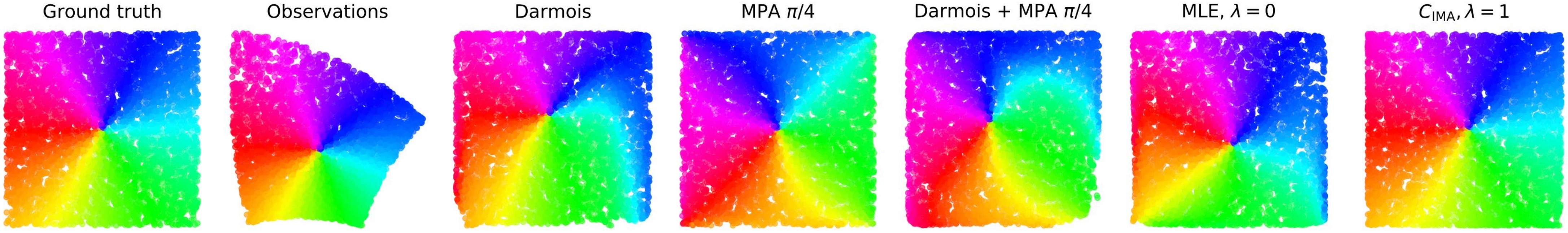
- Darmois (inverse CDF) construction
- Measure-preserving automorphisms (MPA)

Consistent with existing identifiability results for linear ICA, and conformal maps.

Experimental results

Even when assumptions are not perfectly satisfied, IMA seems useful to distinguish spurious solutions and recover the true sources





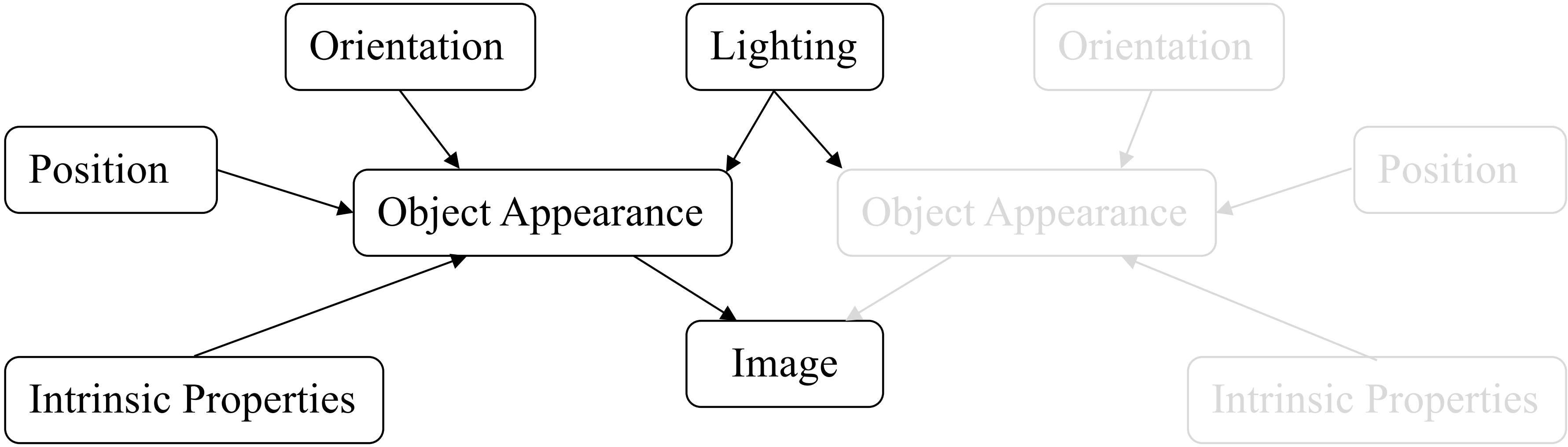
Generative models as “causal digital twins”

Disentangled (causal) factorization *https://arxiv.org/abs/1911.10500*

- **independent noises** in the causal graph:

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

- **independent mechanisms**: 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** (implies intervenability)



Presented at the ICLR 2020 workshop “Causal learning for decision making”

TOWARDS CAUSAL GENERATIVE SCENE MODELS VIA COMPETITION OF EXPERTS

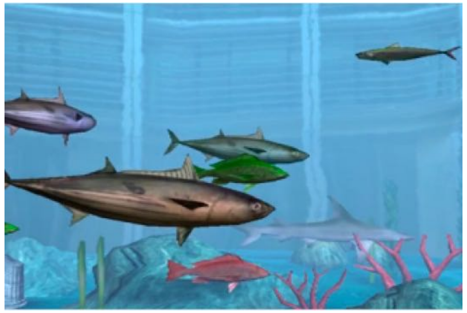
Julius von Kügelgen[*†1,2], **Ivan Ustyuzhaninov**[*†3],
Peter Gehler[‡4], **Matthias Bethge**[‡3,4], **Bernhard Schölkopf**[‡1,4]
[1]Max Planck Institute for Intelligent Systems Tübingen, Germany
[2]Department of Engineering, University of Cambridge, United Kingdom
[3]University of Tübingen, Germany
[4]Amazon Tübingen, Germany
{jvk,bs}@tuebingen.mpg.de,
{ivan.ustyuzhaninov,matthias.bethge}@bethgelab.org,
pgehler@amazon.com

ABSTRACT

Learning how to model complex scenes in a modular way with recombinable components is a pre-requisite for higher-order reasoning and acting in the physical world. However, current generative models lack the ability to capture the inherently compositional and layered nature of visual scenes. While recent work has made progress towards unsupervised learning of object-based scene representations, most models still maintain a global representation space (i.e., objects are not explicitly separated), and cannot generate scenes with novel object arrangement and depth ordering. Here, we present an alternative approach which uses an inductive bias encouraging modularity by training an ensemble of generative models (*experts*). During training, experts compete for explaining parts of a scene, and thus specialise on different object classes, with objects being identified as parts that re-occur across multiple scenes. Our model allows for controllable sampling of individual objects and recombination of experts in physically plausible ways. In contrast to other methods, depth layering and occlusion are handled correctly, moving this approach closer to a causal generative scene model. Experiments on simple toy data qualitatively demonstrate the conceptual advantages of the proposed approach.

1 INTRODUCTION

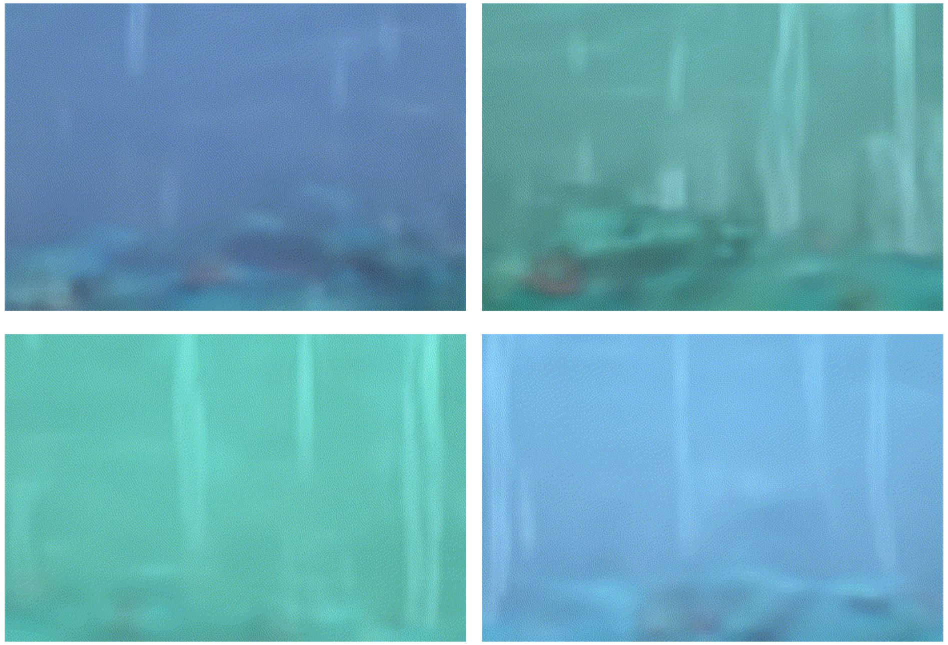
Proposed in the early days of computer vision Grenander (1976); Horn (1977), *analysis-by-synthesis* is an approach to the problem of visual scene understanding. The idea is conceptually elegant and appealing: build a system that is able to synthesize complex scenes (e.g., by rendering), and then understand analysis (inference) as the inverse of this process that decomposes new scenes into their constituent components. The main challenges in this approach are the need for generative models of objects (and their composition into scenes) and the need to perform tractable inference given new



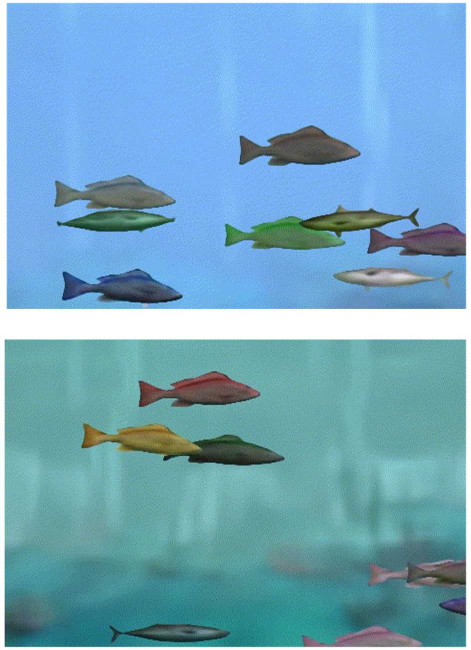
Training set

Tangemann, Schneider et al., 2021

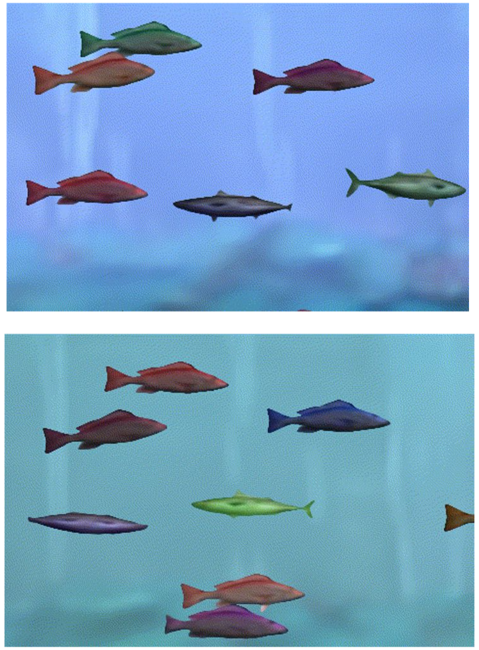
# of objects



fish identities



position



Towards causal machine learning

learn *world models (*aka *digital twins)* that are

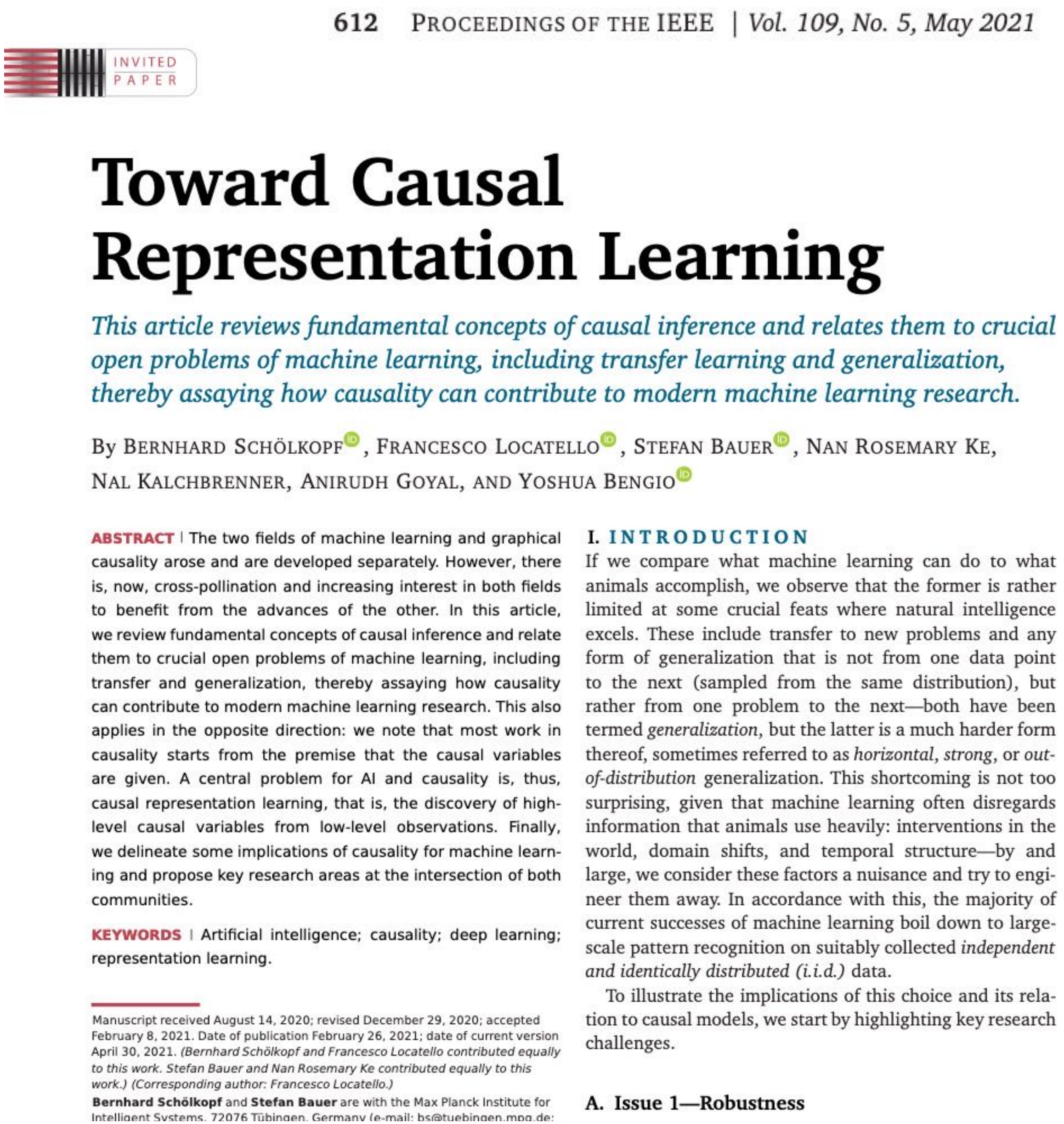
(1) data-efficient

- use data from multiple tasks in multiple environments
- use re-usable components that are robust across tasks, i.e., causal (independent) mechanisms
  - disentanglement as a causal problem
  - bias RL to search for invariance / find models where shifts are sparse

(2) interventional

- move representation learning towards interventional representations: *”thinking is acting is an imagined space”* (Konrad Lorenz) --- planning, reasoning, …

cf. Schölkopf, Janzing, Lopez-Paz 2016
ICML 2017 talk, https://vimeo.com/238274659



INVITED PAPER

Toward Causal Representation Learning

This article reviews fundamental concepts of causal inference and relates them to crucial open problems of machine learning, including transfer learning and generalization, thereby assaying how causality can contribute to modern machine learning research.

By BERNHARD SCHÖLKOPF, FRANCESCO LOCATELLO, STEFAN BAUER, NAN ROSEMARY KE,
NAL KALCHBRENNER, ANIRUDH GOYAL, AND YOSHUA BENGIO

ABSTRACT | The two fields of machine learning and graphical causality arose and are developed separately. However, there is, now, cross-pollination and increasing interest in both fields to benefit from the advances of the other. In this article, we review fundamental concepts of causal inference and relate them to crucial open problems of machine learning, including transfer and generalization, thereby assaying how causality can contribute to modern machine learning research. This also applies in the opposite direction: we note that most work in causality starts from the premise that the causal variables are given. A central problem for AI and causality is, thus, causal representation learning, that is, the discovery of high-level causal variables from low-level observations. Finally, we delineate some implications of causality for machine learning and propose key research areas at the intersection of both communities.

KEYWORDS | Artificial intelligence; causality; deep learning; representation learning.

I. INTRODUCTION

If we compare what machine learning can do to what animals accomplish, we observe that the former is rather limited at some crucial feats where natural intelligence excels. These include transfer to new problems and any form of generalization that is not from one data point to the next (sampled from the same distribution), but rather from one problem to the next—both have been termed *generalization*, but the latter is a much harder form thereof, sometimes referred to as *horizontal*, *strong*, or *out-of-distribution* generalization. This shortcoming is not too surprising, given that machine learning often disregards information that animals use heavily: interventions in the world, domain shifts, and temporal structure—by and large, we consider these factors a nuisance and try to engineer them away. In accordance with this, the majority of current successes of machine learning boil down to large-scale pattern recognition on suitably collected *independent and identically distributed (i.i.d.)* data.

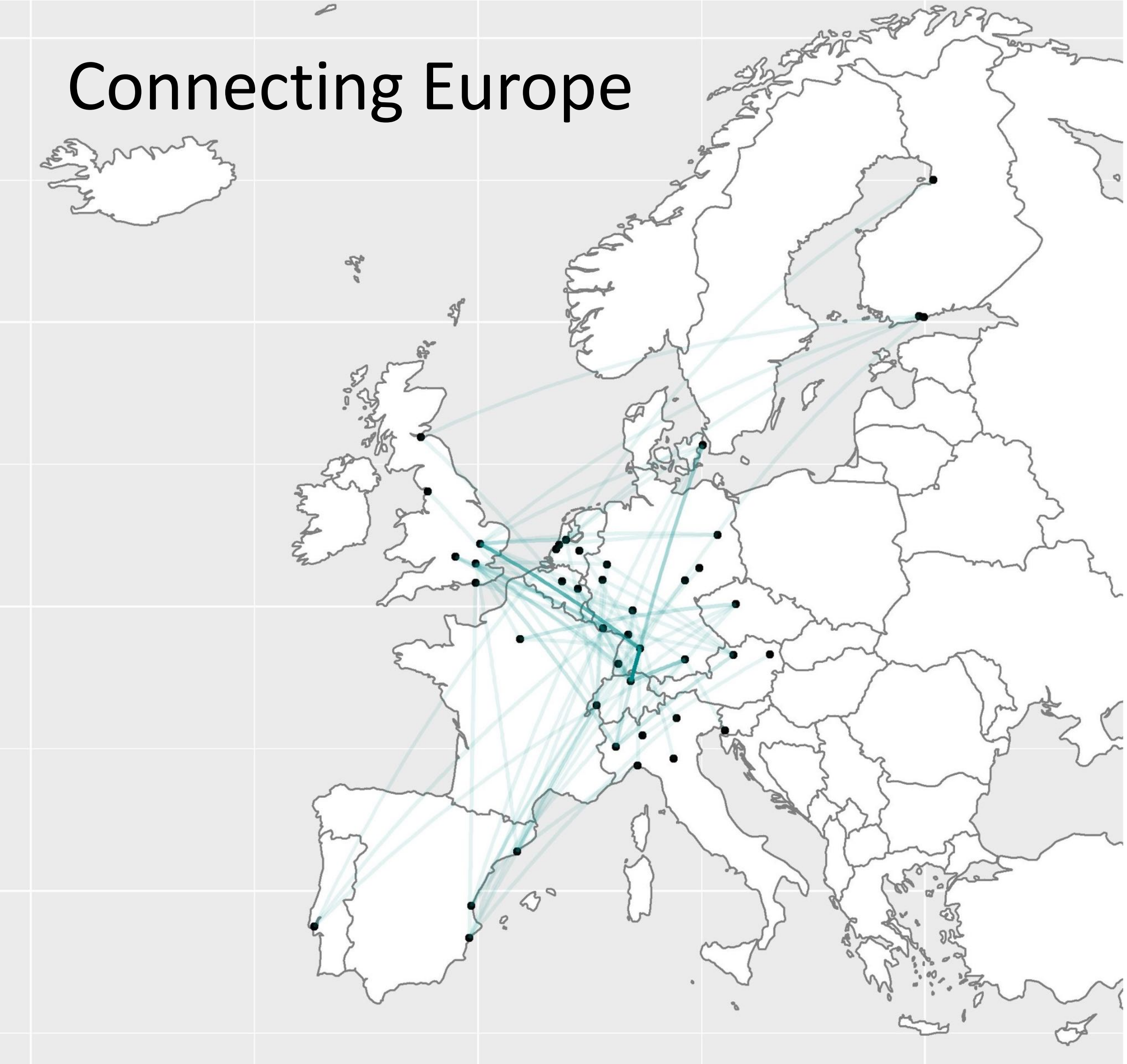
To illustrate the implications of this choice and its relation to causal models, we start by highlighting key research challenges.

A. Issue 1—Robustness

Manuscript received August 14, 2020; revised December 29, 2020; accepted February 8, 2021. Date of publication February 26, 2021; date of current version April 30, 2021. *(Bernhard Schölkopf and Francesco Locatello contributed equally to this work. Stefan Bauer and Nan Rosemary Ke contributed equally to this work.) (Corresponding author: Francesco Locatello.)*

Bernhard Schölkopf and **Stefan Bauer** are with the Max Planck Institute for Intelligent Systems, 72076 Tübingen, Germany (e-mail: bs@tuebingen.mpg.de;

Connecting Europe



> 1,300
applications

international
>70 countries

diverse
>90 nationalities

competitive
< 5% acceptance

joint supervision
>100 Fellows and
Scholars recruiting

- **120 international exchanges**
- **70 institutions** and **companies**
- **jointly advised by 116 ELLIS Fellows, Scholars and Members**

Thank You