Towards Causal NLP

Bernhard Schölkopf

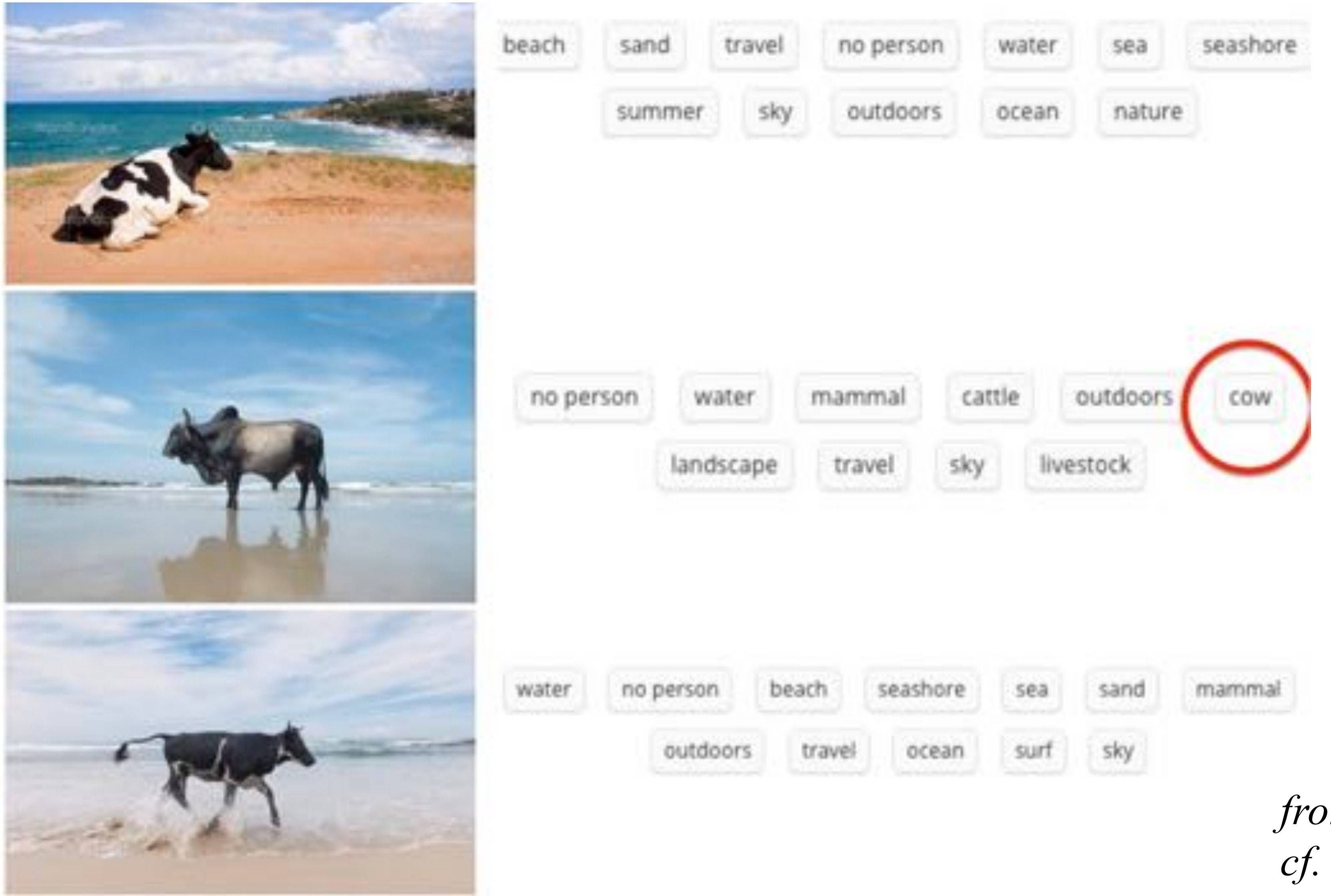
Max Planck Institute for Intelligent Systems

Max Planck ETH Center for Learning Systems

Human-level object recognition?

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

Machine learning uses correlations rather than causality



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

Adversarial Vulnerability

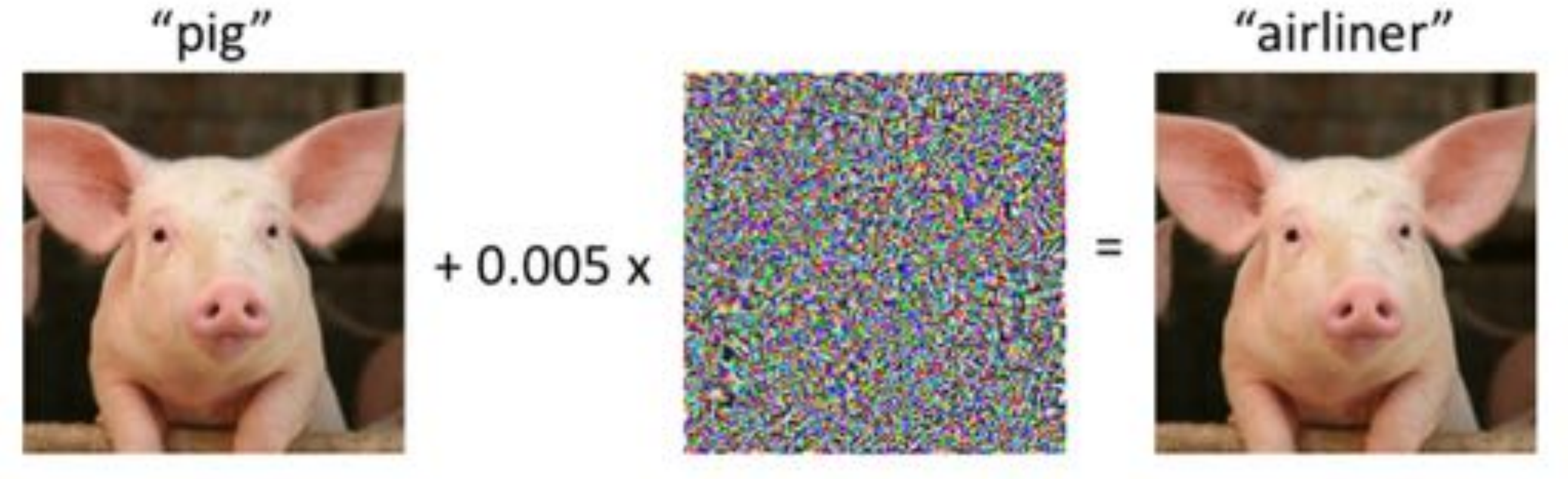


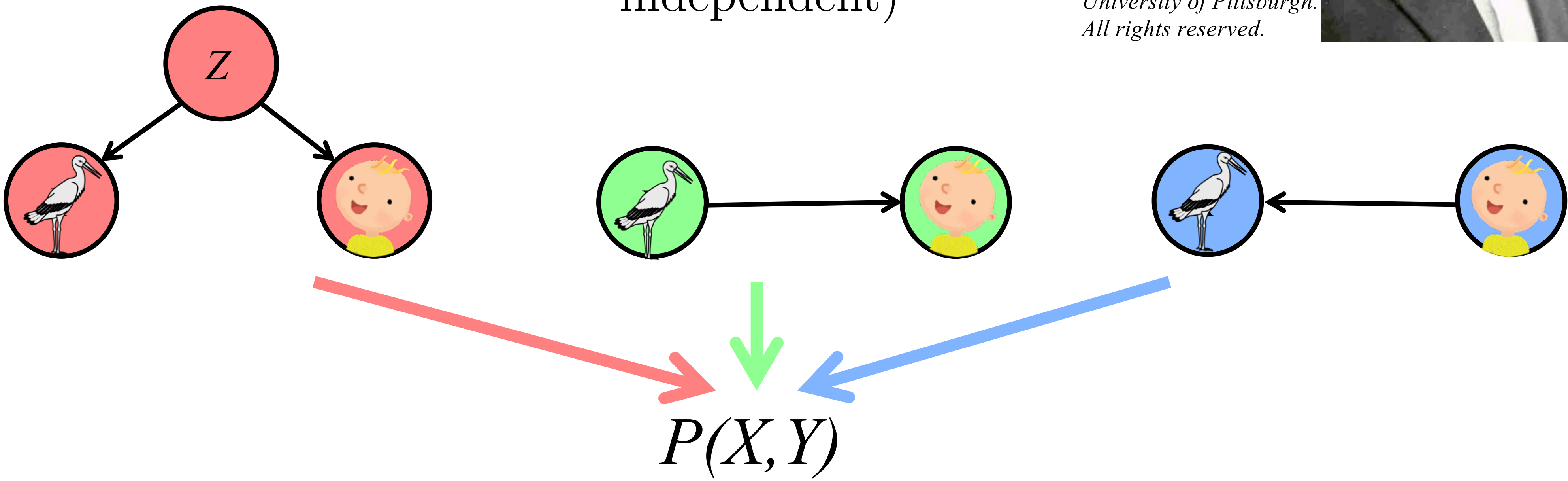
Image credit: http://people.csail.mit.edu/madry/lab/blog/adversarial/2018/07/06/adversarial_intro/

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

Reichenbach's Common Cause Principle

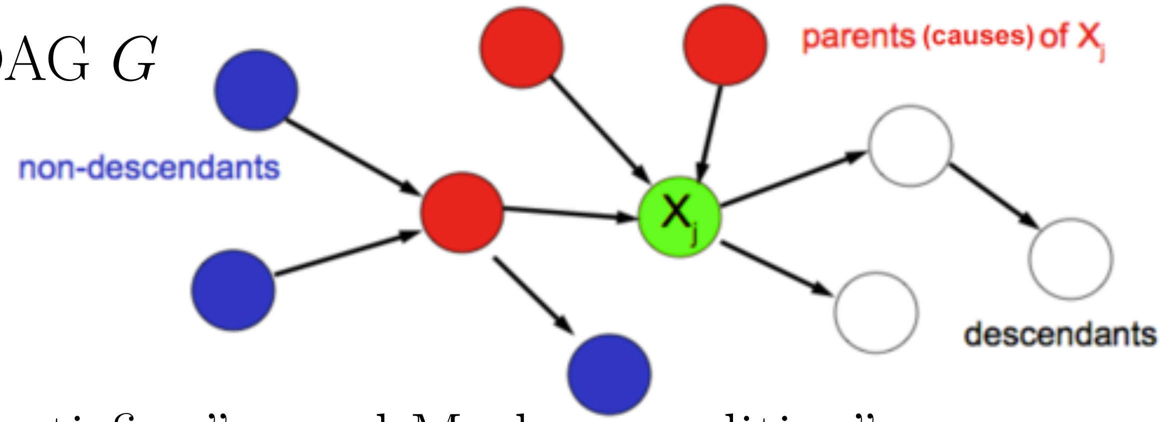
(i) if X and Y are dependent, then there exists Z *causally* influencing both;

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



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

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 $U_1, \ldots, U_n$.
- *entailed distribution* $p(X_1, \ldots, X_n)$ satisfies "causal Markov condition"

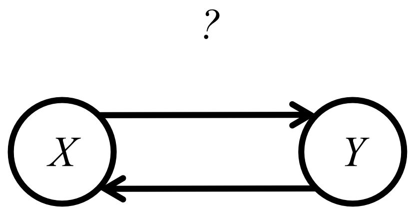
- (G, p) is a "graphical model" *(Lauritzen, 1996)*
- Causal factorization $p(X_1, \ldots, X_n) = \prod_i p(X_i|\mathrm{Parents}_i)$
- the $p(X_i|\mathrm{Parents}_i)$ are "independent" from each other.
  Can *intervene* on some and the other terms remain *invariant*

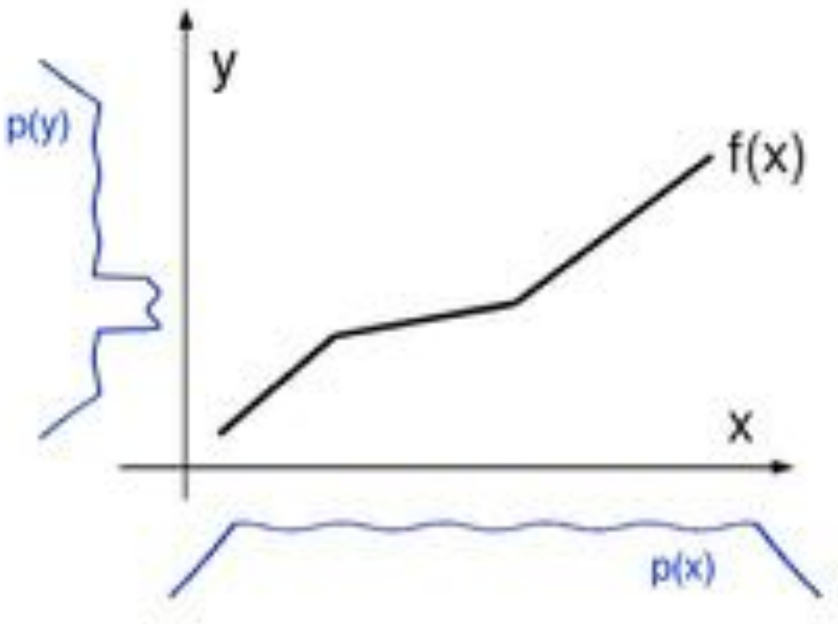
Independent Causal Mechanisms Principle (ICM):

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

MAX-PLANCK-GESELLSCHAFT

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):

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

Note: this is equivalent to

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

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

Proposition: If $f \neq Id$,

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

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:

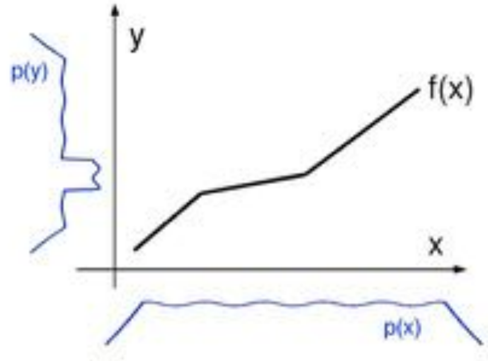
$$S(p_y) - S(p_x) = S(v_y) - S(u_x)$$

Orthogonality (information geometric):

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

which can be rewritten as

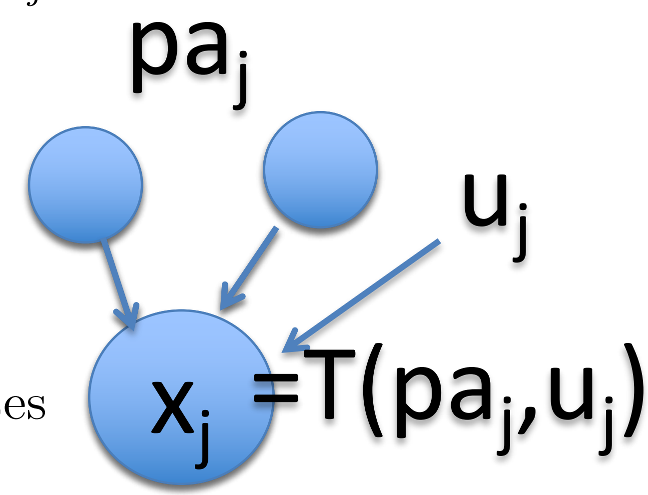
$$D(p_y \| u_y) = D(p_x \| u_x) + D(v_y \| u_y)$$



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



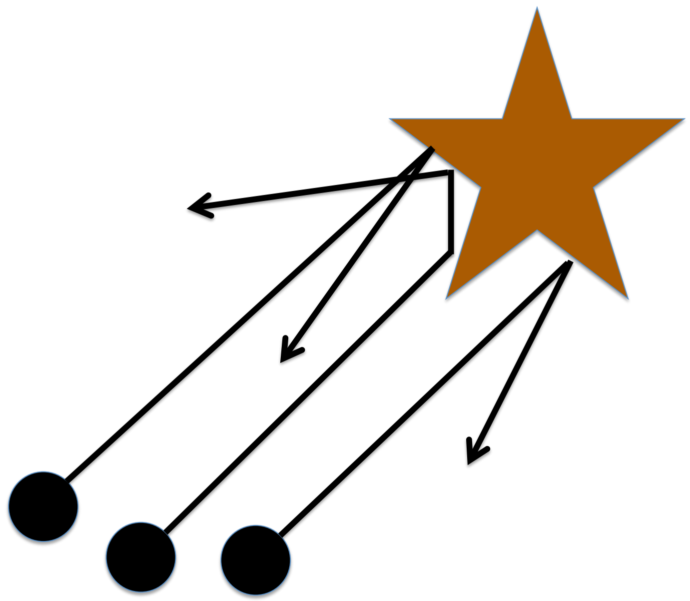
- all u_j are jointly independent
- the program u_j represents the causal mechanism that generates the effect from its causes
- u_j are the analog of the unobserved noise terms in the statistical functional model

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

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

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 $I(s:M) \stackrel{+}{=} 0$ implies

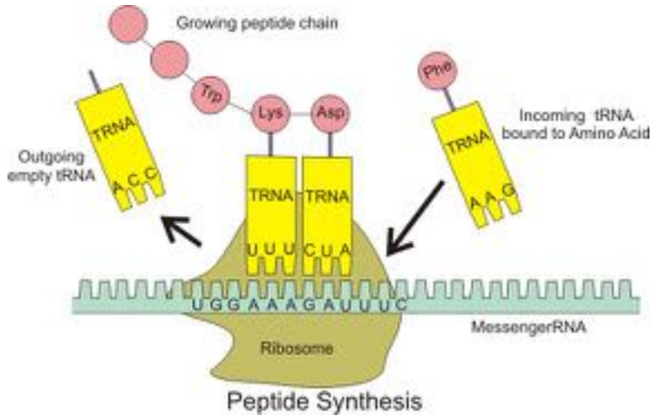
$$K(M(s)) \stackrel{+}{\geq} K(s)$$

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}.

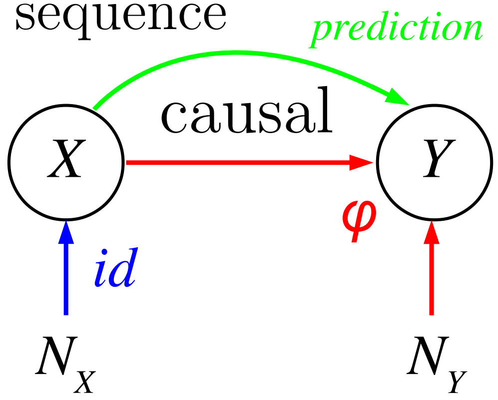
Janzing, Chaves, Schölkopf: Algorithmic independence of initial condition and dynamical law in thermodynamics and causal inference. New J. of Physics, 2016

Using cause-effect knowledge

- example 1: predict protein from mRNA sequence

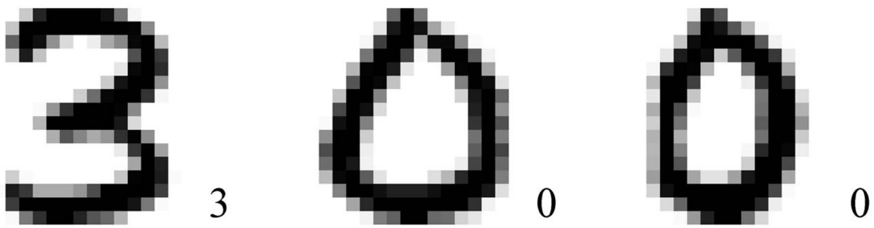


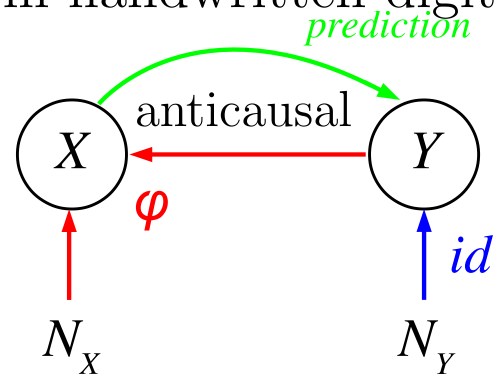
Source: http://commons.wikimedia.org/wiki/File:Peptide_syn.png



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

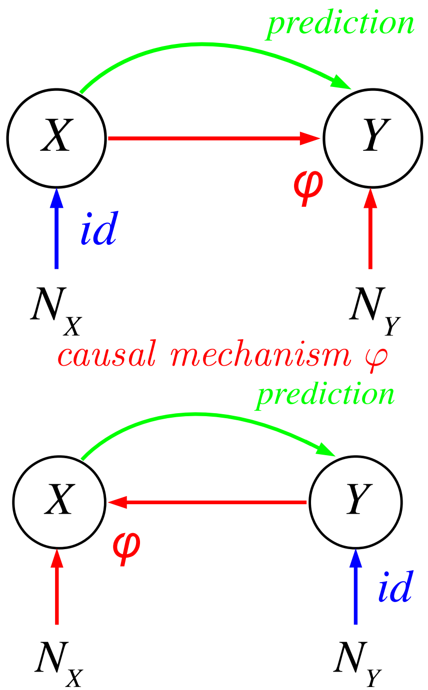
1. *semi-supervised learning impossible*
2. *$p(Y|X)$ invariant under change in $p(X)$*

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)


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.

Annotation process (Noise)

Effect = CausalMechanism (Cause, Noise)

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

Causal Direction of Data Collection Matters: Implications of **Causal and Anticausal Learning for NLP**

EMNLP 2021 (Oral)

arxiv.org/abs/2110.03618







Zhijing Jin* Julius von Kügelgen* Jingwei Ni Tejas Vaidhya Ayush Kaushal Mrinmaya Sachan Bernhard Schoelkopf

Bernhard Schölkopf

MAX-PLANCK-GESELLSCHAFT

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:

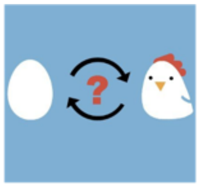
Semi-supervised learning (SSL): *anticausal* > *causal*.

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

Connecting Causal Inference and NLP



Causality

Focus of Our Lab

(Jin et al., 2021a) EMNLP oral

- **Independent Causal Mechanism (ICM)**
- **Causal representation learning**
  Ongoing: debias LM using causality
- **Robustness, domain adaptation, …**
- **Logical fallacy detection**

(Jin et al., 2021b)

- **Causal discovery for text data**
- **Causal effect estimation for text data**

1. Causality on COVID policies (Jin et al., 2021c) EMNLP Findings
2. Causality on changes of slang (Keidar et al., 2021)
3. [More ongoing work]

Symbolic AI

Classic AI:
Symbols provided a priori
Rules provided a priori

Machine learning:
Representations learnt from data
Only include *statistical* information.

Causal Modeling:
Structural causal models assume the causal variables are given.

https://arxiv.org/abs/2102.11107, Proceedings of the IEEE 2021

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:

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

- **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**

(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:

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

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

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

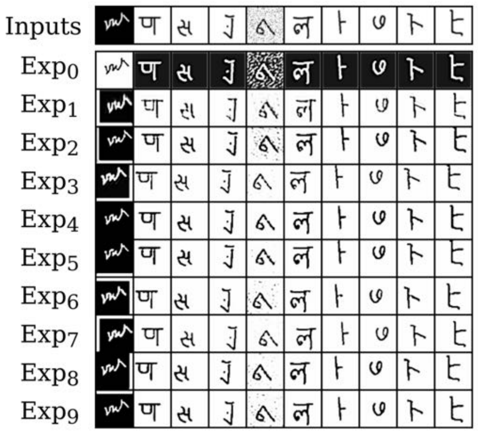
with independent mechanisms $p\left(X_i \mid \mathrm{PA}_i\right)$.

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).

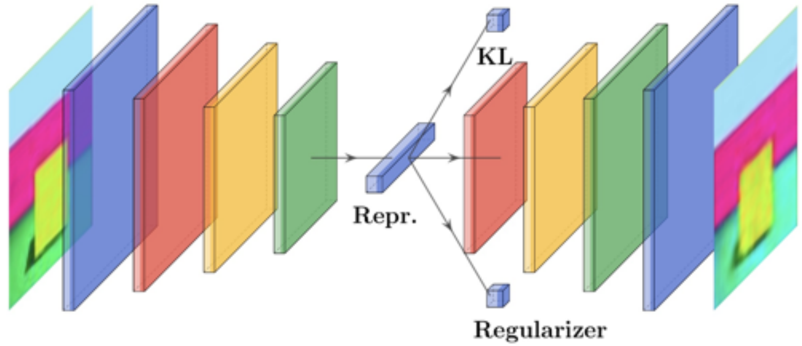
Stated in *(Parascandolo et al., arXiv:1712.00961 (2017); Bengio et al., arXiv:1901.10912 (2019), Schölkopf, arXiv:1911:10500 (2019))*; see also *(Schölkopf et al., ICML 2012, Schölkopf, Janzing, Lopez-Paz 2016, Zhang et al., ICML 2013, Huang, Zhang et al., JMLR 2020)*

https://arxiv.org/abs/1911.10500



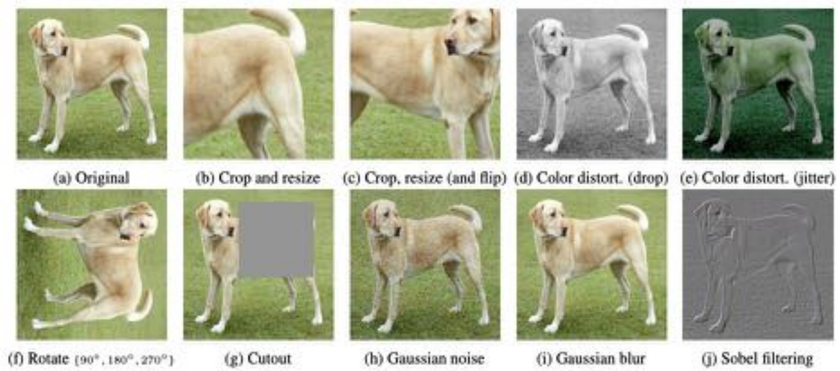
Learning independent mechanisms
(Parascandolo et al., ICML 2018)

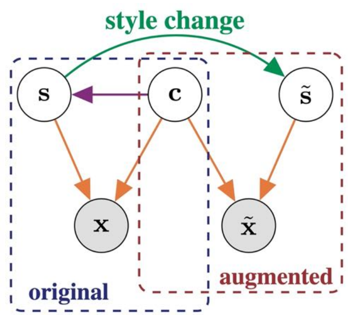
Real Robot Challenge
(Bauer et al., 2020)



Hardness of disentanglement
(Locatello et al., ICML 2019, best paper prize)

Interventional Representations
(Besserve et al., ICLR 2020)





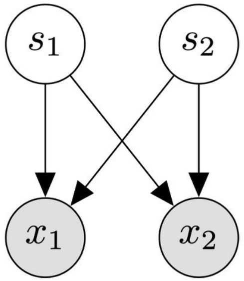
Self-supervised learning provably isolates content from style *(von Kügelgen et al., NeurIPS 2021)*

MAX-PLANCK-GESELLSCHAFT

Causality for nonlinear ICA

(https://arxiv.org/abs/2106.05200, NeurIPS 2021)

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

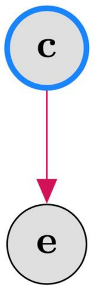
Observe: nonlinear mixtures, $\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{s})$, of independent sources $\boldsymbol{s}$

Goal: recover the unobserved sources (blind source separation)

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

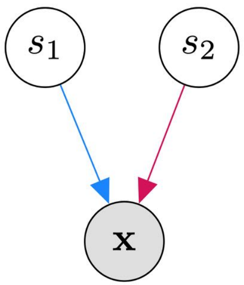
Recently: use auxiliary variables [Hyvärinen et al., '16, '17, '19]

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



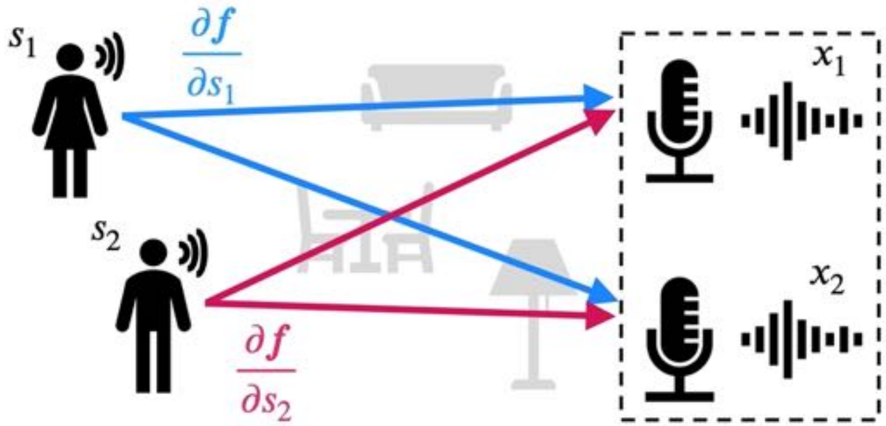
ICM usually applied to cause distribution p_c and mechanism $p_{e|c}$ (or $\boldsymbol{f}$), e.g., cause-effect discovery

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



Towards causal machine learning

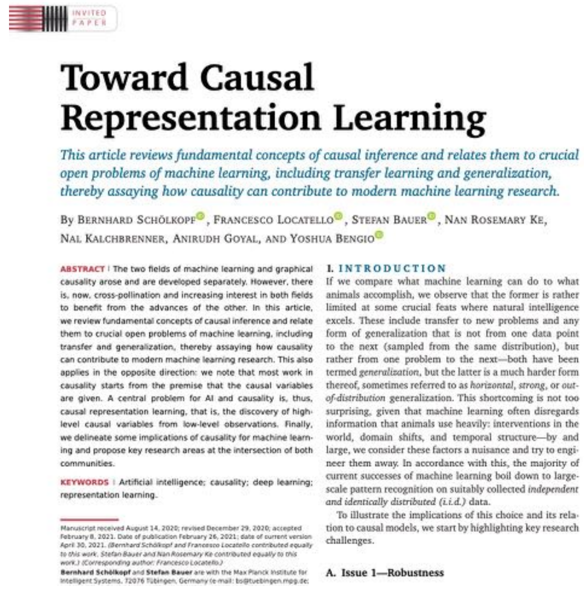
learn *world models* that are

(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, …


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.

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;

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

European Laboratory for Learning and Intelligent Systems

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

MAX-PLANCK-GESELLSCHAFT