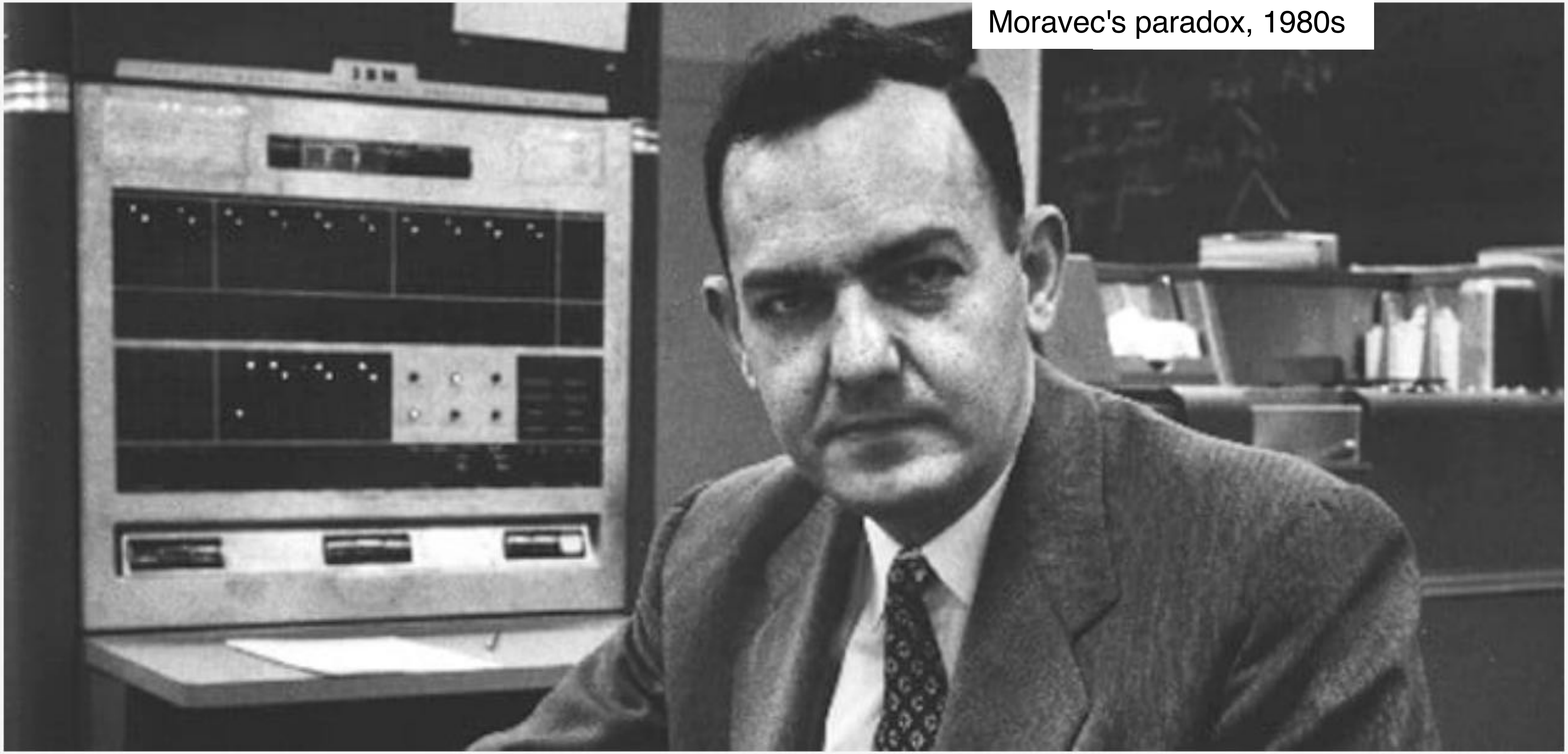
Causality

Bernhard Schölkopf

Max Planck Institute for Intelligent Systems & ETH Zürich

Max Planck ETH Center for Learning Systems

Moravec's paradox, 1980s



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

https://www.gwern.net/Tanks

Search

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

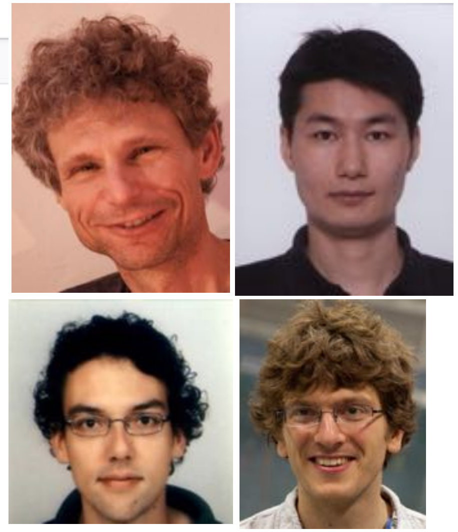
NN, history, sociology, Google, bibliography

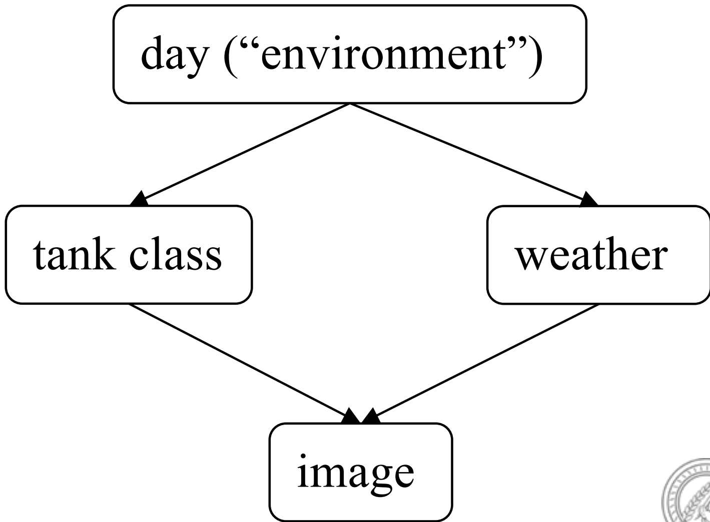
20 Sep 2011–14 Aug 2019 · *finished* · certainty: *highly likely* · importance: 4

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.

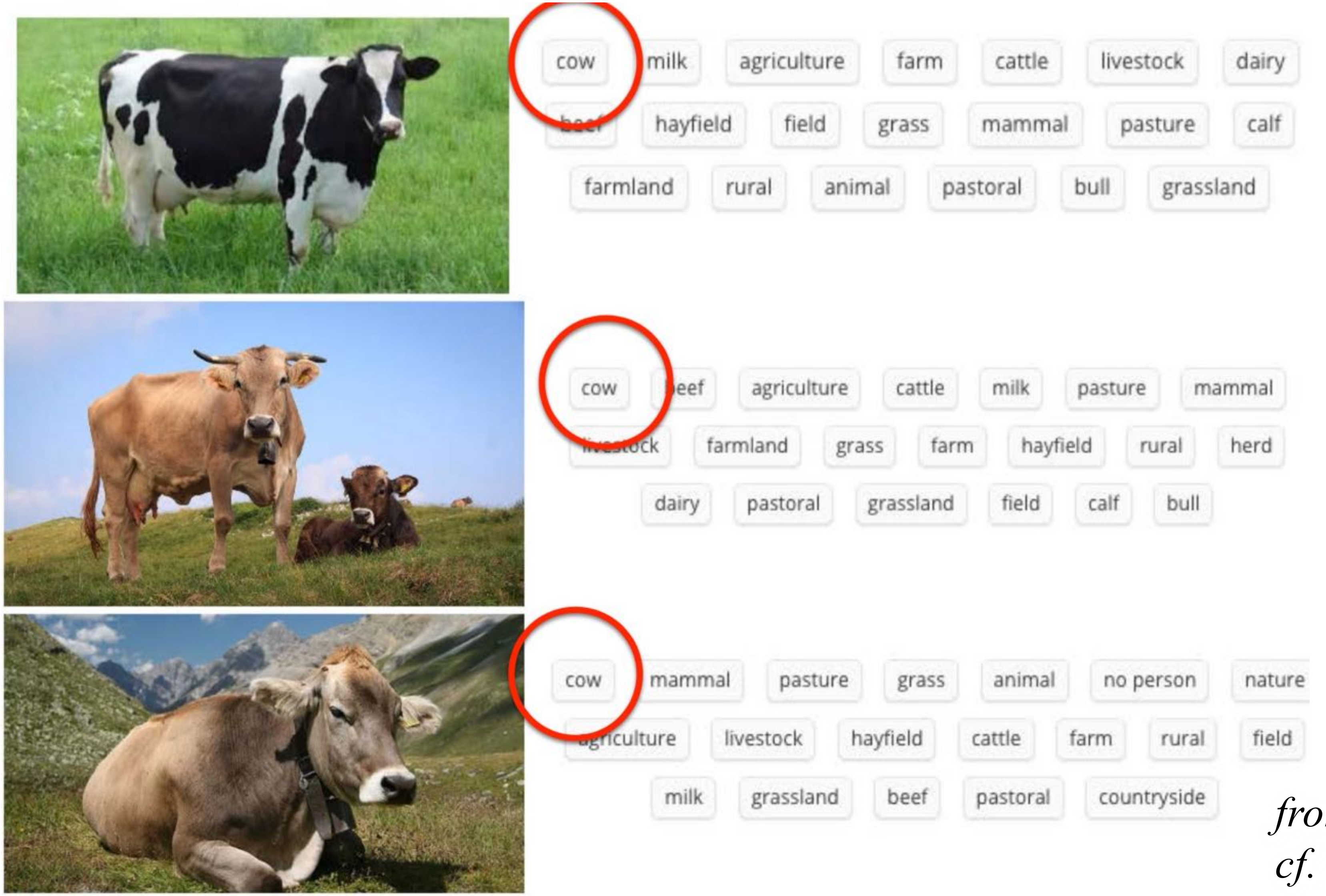
I collate many extent versions dating back a quarter of a century to 1992 along with two NN-related anecdotes from the 1960s; their contradictions & details indicate a classic "urban legend", with a probable origin in a speculative question in the 1960s by Edward Fredkin at an AI conference about some 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.



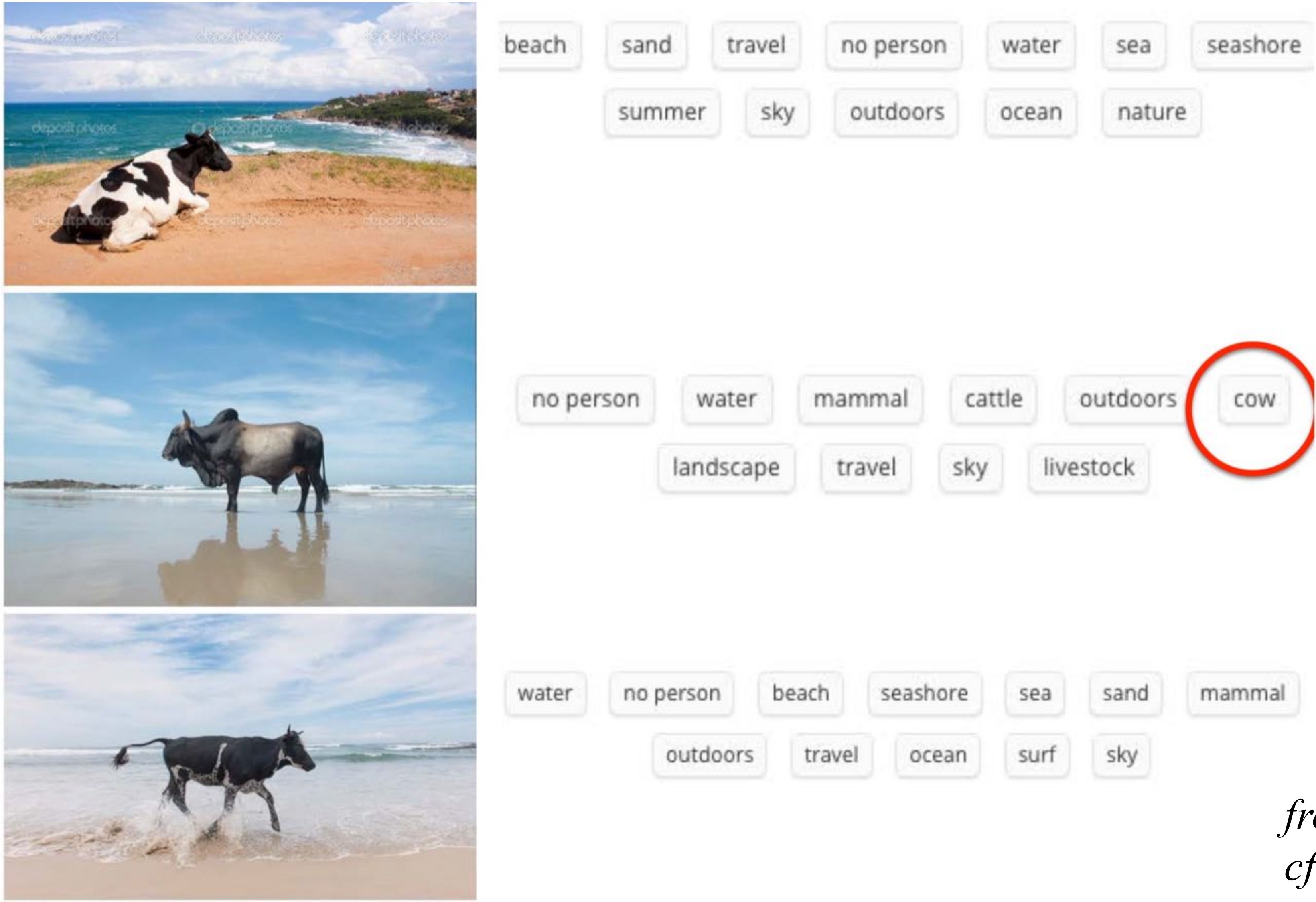


Human-level object recognition?



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

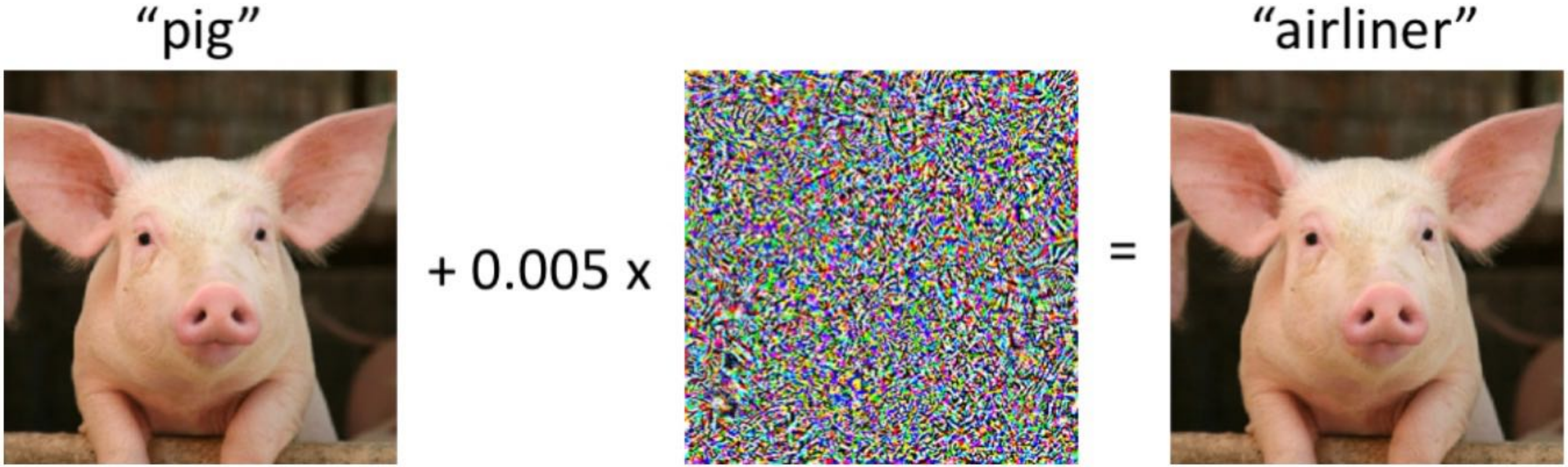


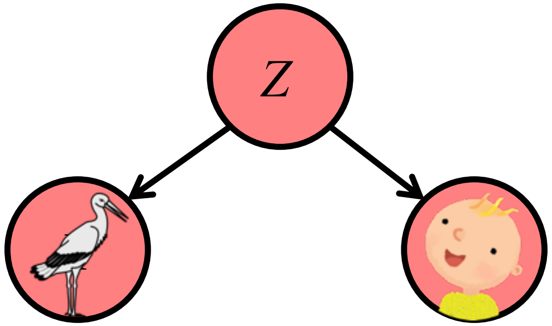
Image credit: http://people.csail.mit.edu/madry/lab/blog/adversarial/2018/07/06/adversarial_intro/

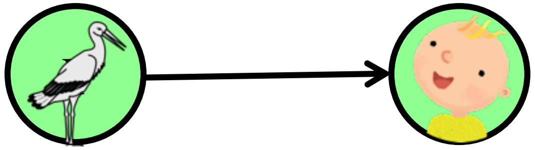
C. Szegedy et al. Intriguing properties of neural networks. *arXiv:1312.6199, 2013*

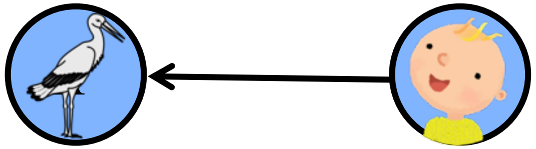
Reichenbach's Common Cause Principle

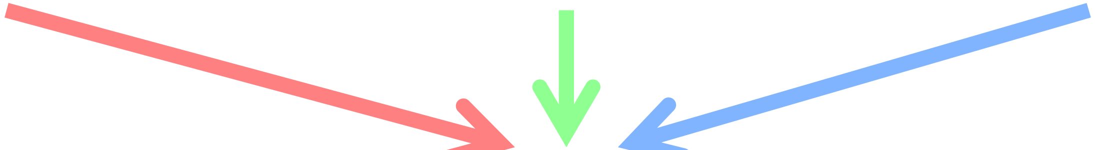
(i) if X and Y are dependent, then there exists Z *causally* influencing both;

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







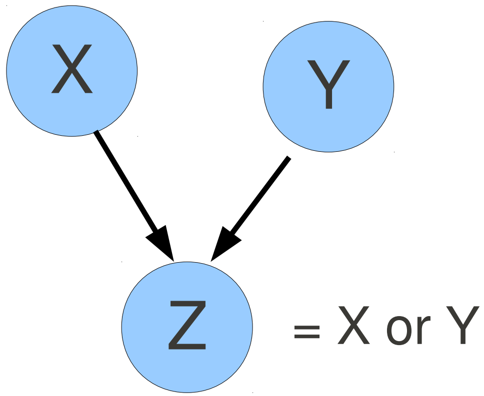


$P(X,Y)$

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

Correlation by conditioning on common effects

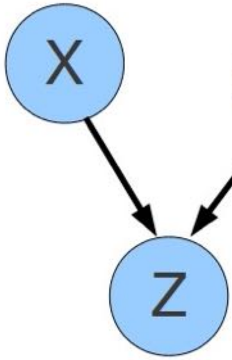
Berkson's paradox (1946)
Example: X, Y, Z binary

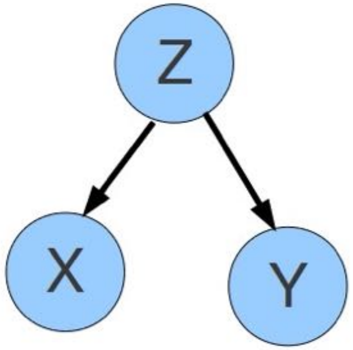


$$X \perp\!\!\!\perp Y \quad \text{but} \quad X \not\perp\!\!\!\perp Y \,|\, Z$$

- assumption 1: there is no correlation between being a good speaker (X) and being a good scientist (Y)
- assumption 2: to be successful, you need to be either a good speaker or a good scientist (or both)
- among the successful scientists, there is a *negative* correlation between being a good speaker and being a good scientist

Asymmetry under inverting arrows

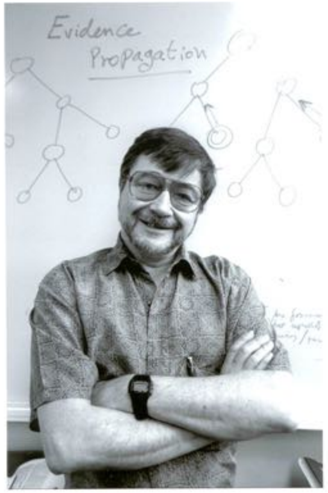




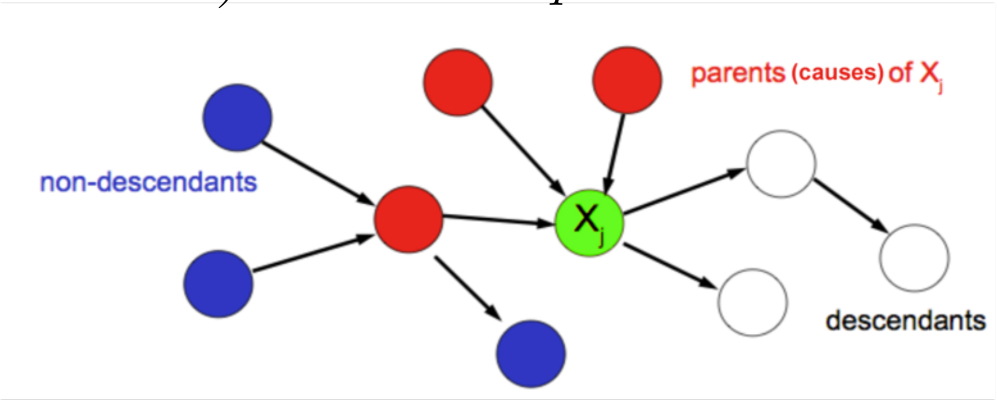
$X \perp\!\!\!\perp Y$ $\qquad$ $X \not\!\perp\!\!\!\perp Y$

$X \not\!\perp\!\!\!\perp Y \,|\, Z$ $\qquad$ $X \perp\!\!\!\perp Y \,|\, Z$

Definition of a Structural Causal Model *(Pearl et al.)*



- directed acyclic graph G with vertices $X_1, \ldots, X_n$
  (following arrows does not lead to loops)
- Semantics: vertices = observables, arrows = direct causation
- $X_i := f_i(\mathrm{PA}_i, U_i)$, with independent RVs $U_1, \ldots, U_n$ that possess a joint density
- U_i stands for "unexplained" (alternatively "noise" or "exogenous variable")
- this is also called a *(nonlinear) structural equation model*



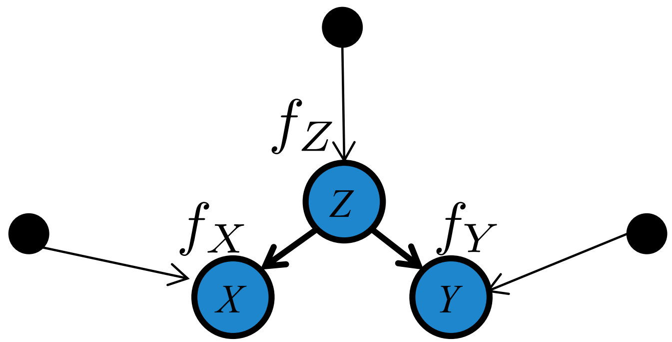
Reichenbach's Principle and causal sufficiency

- Independence of noises is a form of "causal sufficiency:" if the noises were dependent, then Reichenbach's principle would tell us the causal graph is incomplete

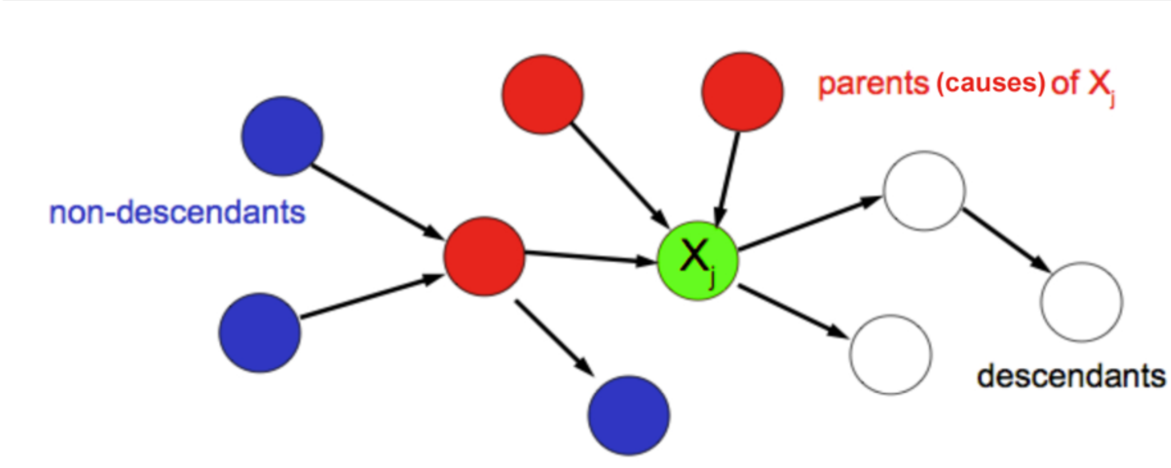
- The SCM model satisfies Reichenbach's principle:

  1. functions of independent variables are independent, hence dependence can only arise in two vertices that depend (partly) on the same noise term(s).

  2. if we condition on these noise terms, the variables become independent

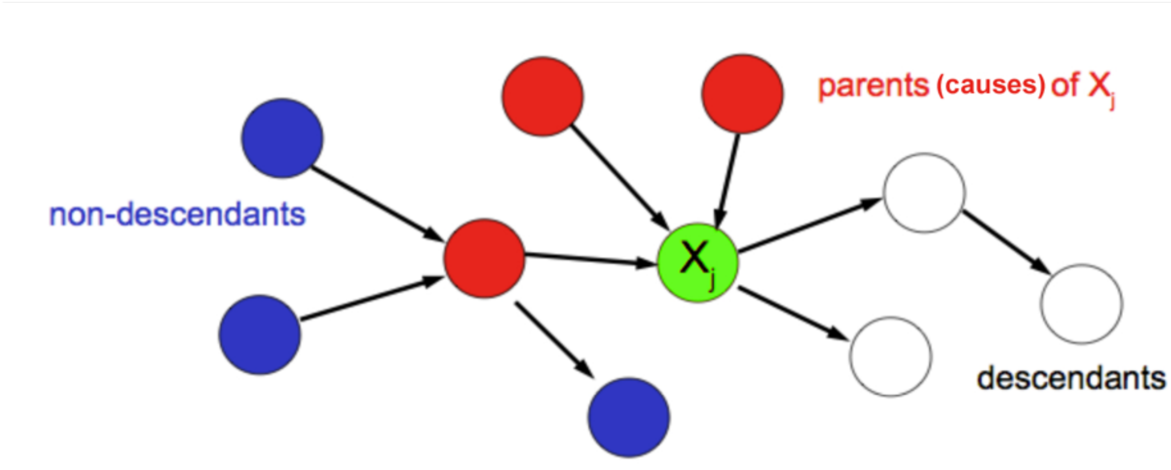


Entailed distribution



- $X_i := f_i(\mathrm{PA}_i, U_i)$, with independent $U_1, \ldots, U_n$.
- Recursively substitute the parent equations to get $X_i = g_i(U_1, \ldots, U_n)$, with independent $U_1, \ldots, U_n$.
- Each X_i is thus a RV and we get a joint distribution of $X_1, \ldots, X_n$, called the *observational distribution.*
- The distribution and the DAG form a *directed graphical model* and any directed graphical model can be written as a functional causal model.

Entailed distribution



- A structural causal model entails a joint distribution $p(X_1, \ldots, X_n)$.

  Questions:

  (1) What can we say about it?

  (2) Can we recover G from p?

Markov conditions *(Lauritzen 1996, Pearl 2000)*

Theorem: the following are equivalent:

– Existence of a structural causal model

– Local Causal Markov condition: X_i *statistically independent* of non-descendants, given parents (i.e.: every information exchange with its non-descendants involves its parents)

– Global Causal Markov condition: "d-separation" (characterizes the set of independences implied by local Markov condition — see below)

– Factorization $p(X_1, \ldots, X_n) = \prod_i p\left(X_i \mid \mathrm{PA}_i\right)$

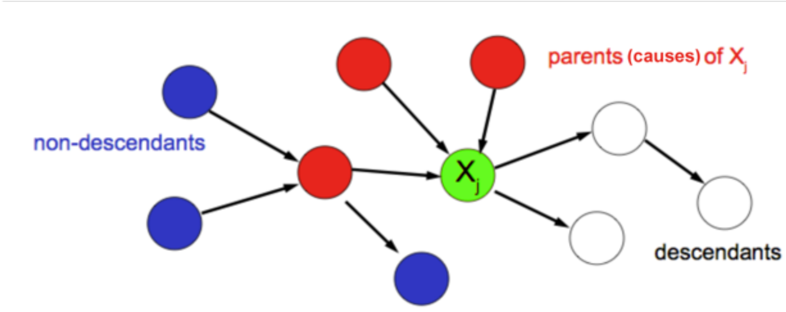
(subject to technical conditions)

$p\left(X_i \mid \mathrm{PA}_i\right)$ is called a *causal conditional* or *causal Markov kernel*.
It corresponds to the structural "equation" $X_i := f_i(\mathrm{PA}_i, U_i)$.

Not every conditional is causal — only those that condition on the parents in our DAG.

Graphical Causal Inference *(Spirtes, Glymour, Scheines, Pearl, ...)*

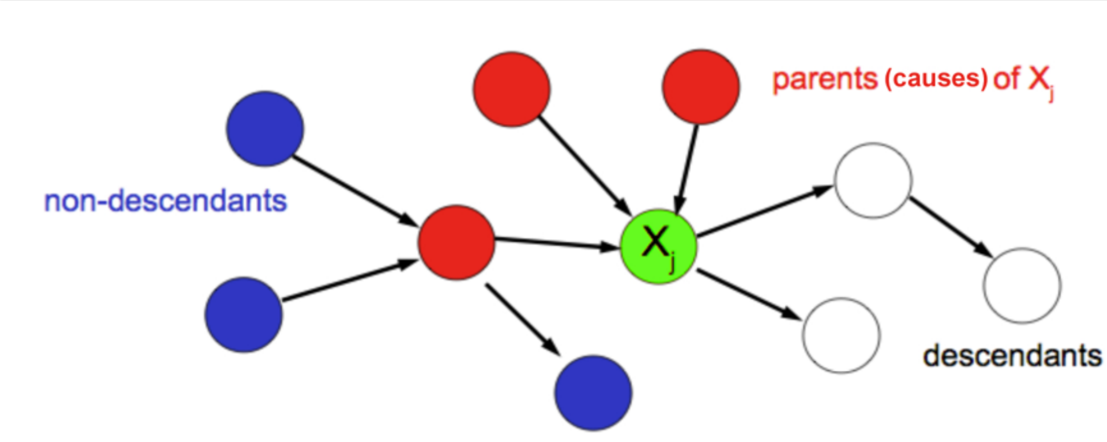
Question: How can we recover G from a single p (e.g., from the observational distribution)?
Answer: by conditional independence testing, infer a class containing the correct G
(i.e., track how the noise information spreads).



Problems:

- Markov condition states $(X \perp\!\!\!\perp Y \,|Z)_G \quad \Rightarrow \quad (X \perp\!\!\!\perp Y \,|Z)_p$, but
  we need "faithfulness" $\quad (X \perp\!\!\!\perp Y \,|Z)_G \quad \Leftarrow \quad (X \perp\!\!\!\perp Y \,|Z)_p$ *(Sprites, Glamour, Scheines 2001)*
  Hard to justify for finite data *(Uhler, Raskutti, Bühlmann, Yu, 2013)*.
- if the f_i are complex, then conditional independence testing based on finite samples becomes arbitrarily hard
- for **two variables** only, there are no conditional independences

Interventions and shifts



- **Definition.** Replacing $X_i := f_i(\mathrm{PA}_i, U_i)$ with another assignment (e.g., $X_i := const.$) is called an *intervention* on X_i.
- The entailed distribution is called the *interventional distribution.*
- This contains as special cases: domain shift distribution and covariate shift distribution (see below).
- A general intervention corresponds to changing some *causal conditionals* $p(X_i|\mathrm{PA}_i)$

Pearl's do-calculus

- Motivation: goal of causality is to infer the effect of interventions

- distribution of Y given that X is set to x: $p(Y|do\, X = x)$ or $p(Y|do\, x)$
- don't confuse it with $p(Y|x)$
- can be computed from p and G

Difference between seeing and doing

$p(y|x)$

Probability a participant of this course can get a NeurIPS paper accepted

$p(y \mid \mathrm{do}\ x)$

Probability that anyone can get a NeurIPS paper accepted after being made to participate in this course

Computing $p(X_1, \ldots, X_n | do\, x_i)$

from $p(X_1, \ldots, X_n)$ and G

- Start with causal factorization

$$p(X_1, \ldots, X_n) = \prod_{j=1}^{n} p(X_j | PA_j)$$

- Replace $p(X_i | PA_i)$ with $\delta_{X_i x_i}$

$$p(X_1, \ldots, X_n | do\, x_i) := \prod_{j \neq i} p(X_j | PA_j) \delta_{X_i x_i}$$

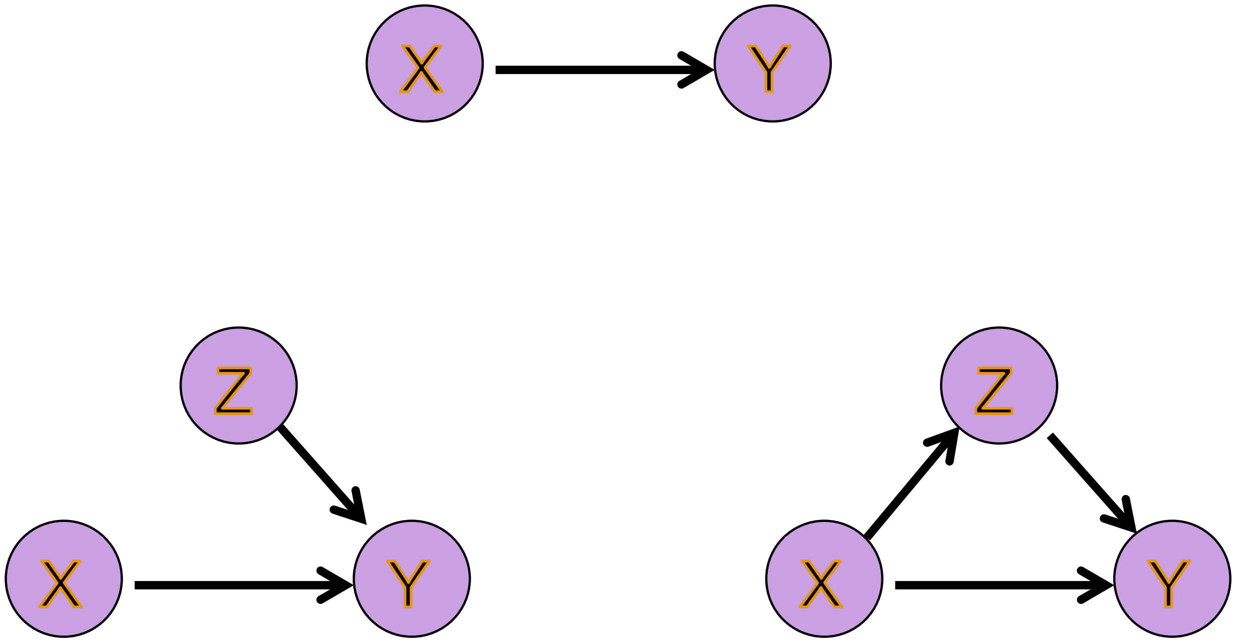
Computing $p(X_k|do\, x_i)$

Sum over x_i to get

$$p(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n|do\, x_i) = \prod_{j \neq I} p(X_j|PA_j(x_i)).$$

- i.e.: for $j \neq i$, drop $p(X_i|PA_i)$ and substitute x_i for X_i
- obtain $p(X_k|do\, x_i)$ by marginalisation

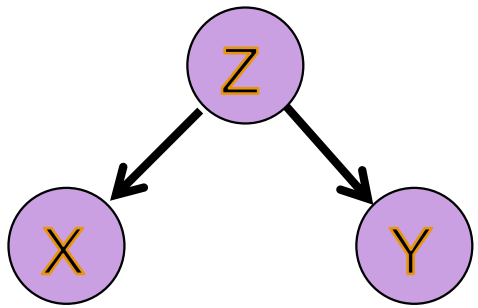
Examples for $p(.|do\,x) = p(.|x)$



Examples for $p(.|do\,x) \neq p(.|x)$

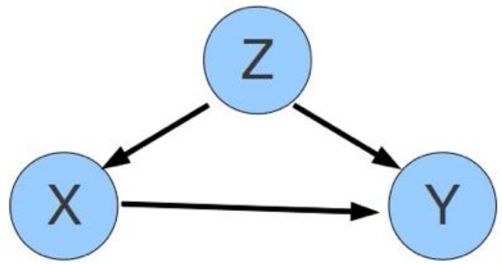
- $p(Y|do\,x) = P(Y) \neq P(Y|x)$

- $p(Y|do\,x) = P(Y) \neq P(Y|x)$



Controlling for confounding / adjustment formula

Y depends on X due to $X \to Y$ *and* the confounder Z



- Causal factorization

$$p(X, Y, Z) = p(Z)\, p(X|Z)\, p(Y|X, Z)$$

- Replace $p(X|Z)$ with δ_{Xx} and integrate out X:

$$\begin{aligned} p(X, Y, Z|do\ x) &= p(Z)\, \delta_{Xx}\, p(Y|X, Z) \\ p(Y, Z|do\ x) &= p(Z)\, p(Y|x, Z) \end{aligned}$$

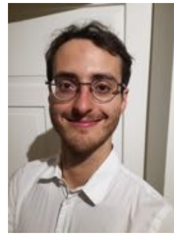
- marginalize over Z to get the "adjustment formula"

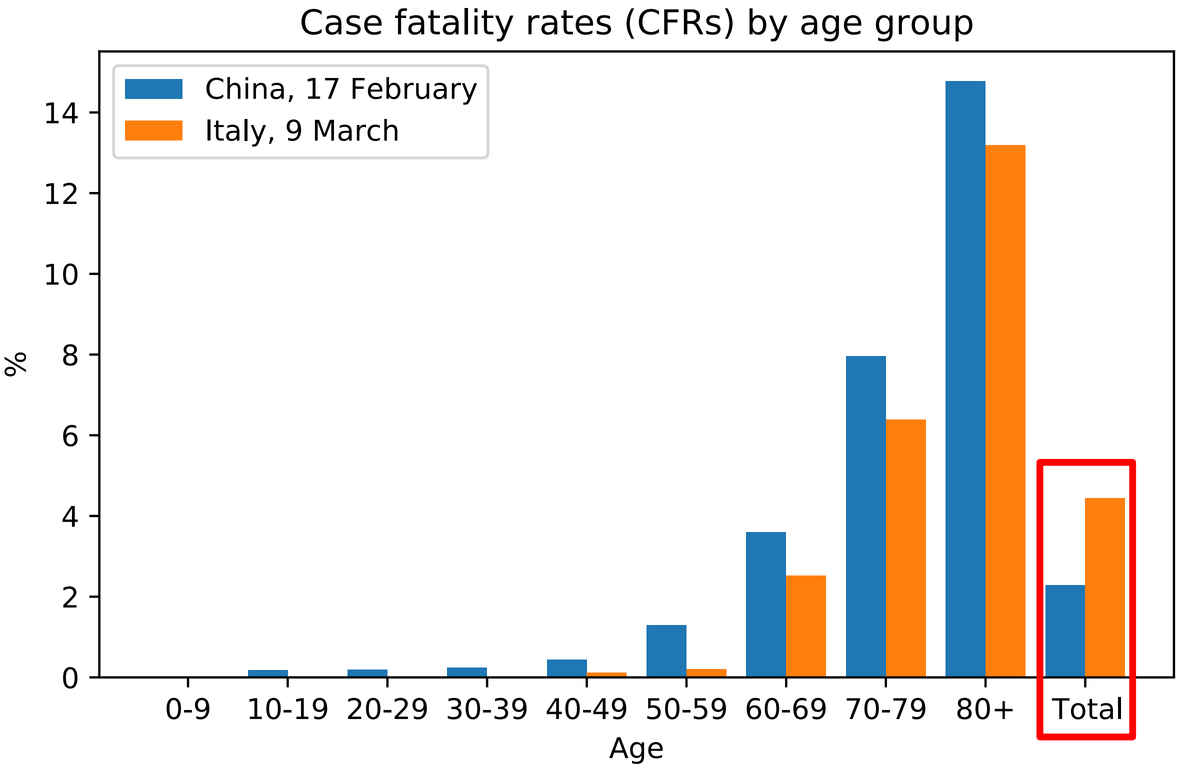
$$p(Y|do\ x) = \sum_z p(z)\, p(Y|x, z)$$

This is different from $p(Y|x)$ (Simpson's paradox).

Simpson's paradox in Covid-19 case fatality rates

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

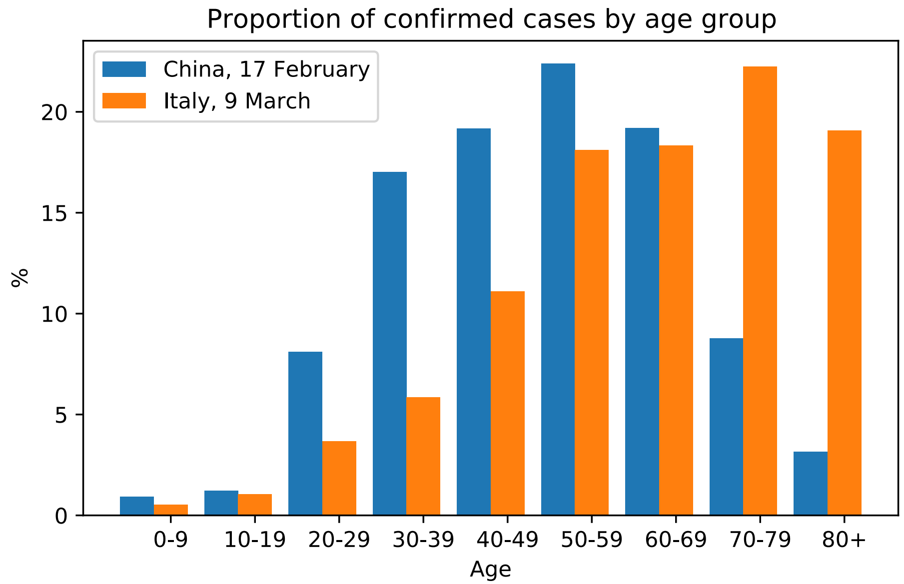




Case fatality rates (CFRs) in Italy are *lower* for each age group, but *higher* overall.

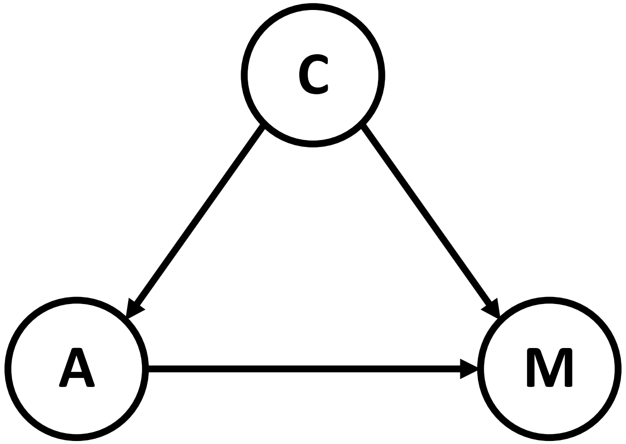
Simpson's paradox: opposite trends in grouped and aggregated data.

Here, it stems from a difference in case demographic:



Thanks to Elias Bareinboim

Coarse-grained causal graph



Data generating process:

- Randomly pick a **country C**
- Given **C**, sample a *positively-tested* patient with **age group A**
- Given **C** and **A**, sample **medical outcome, or mortality, M** (deceased at time of reporting?)

Assumptions & meaning of directed arrows:

- **C → A**: general population demographic, inter-generational mixing, age-specific social-distancing, …
- **A → M**: age-related health condition & other comorbidities.
- **C → M**: number of ventilators & ICU beds, …

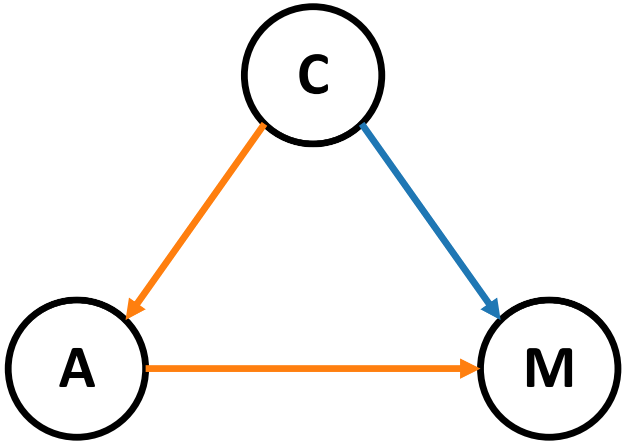
Mediation analysis

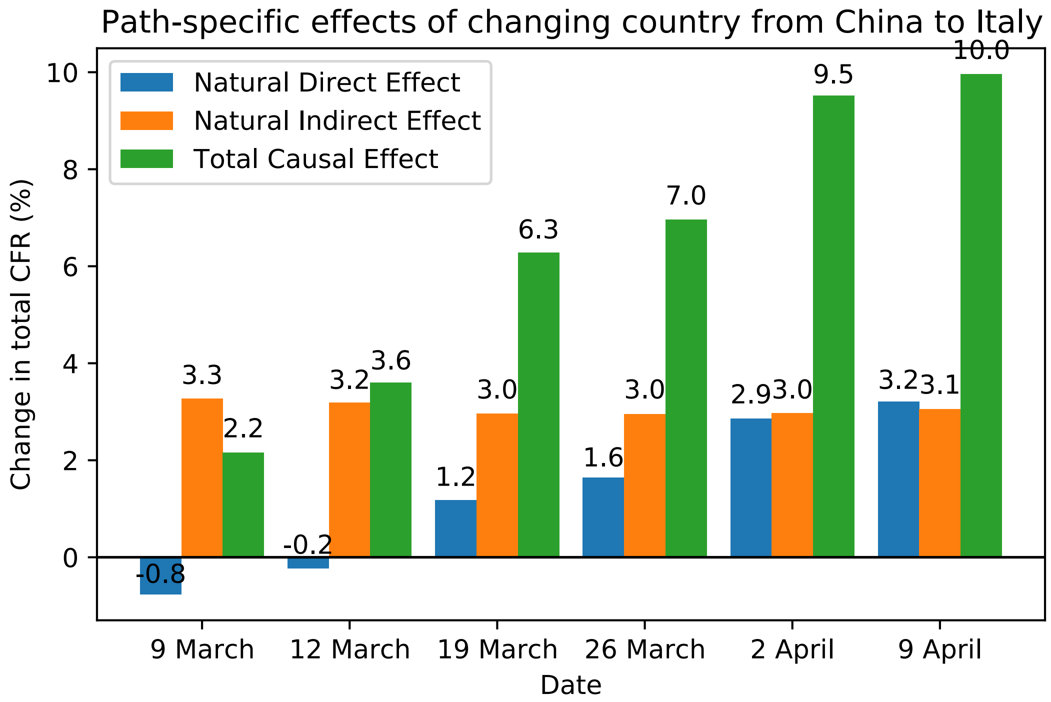
Only for linear models can **total causal effect (TCE)** be decomposed into direct effect (DE) and indirect effect (IE),

$$\mathbf{TCE} = \mathrm{DE} + \mathrm{IE}$$

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.





Does it make sense to talk about causality without mentioning time?

Does it make sense to talk about statistics without mentioning time?

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

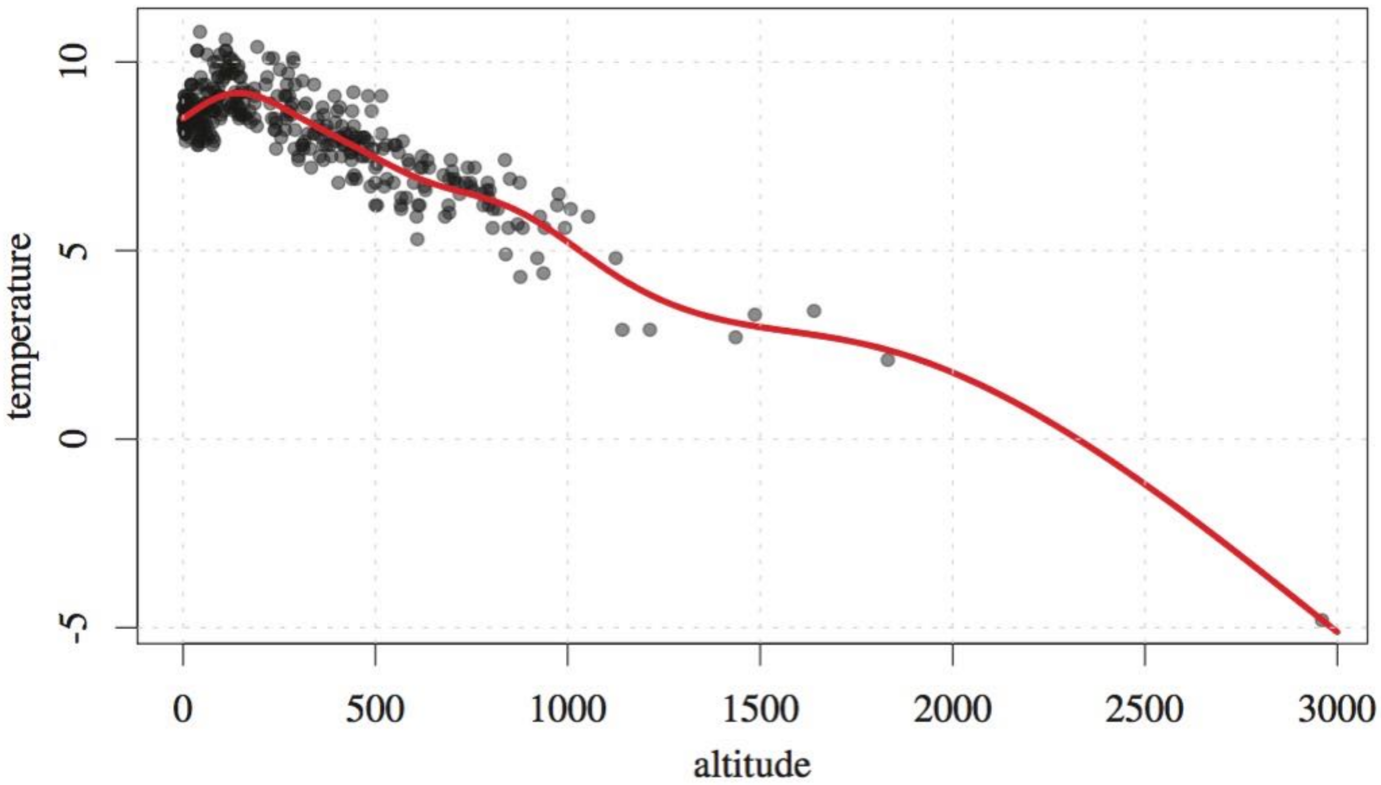
Using dt and $d\mathbf{x} = \mathbf{x}(t + dt) - \mathbf{x}(t)$:

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

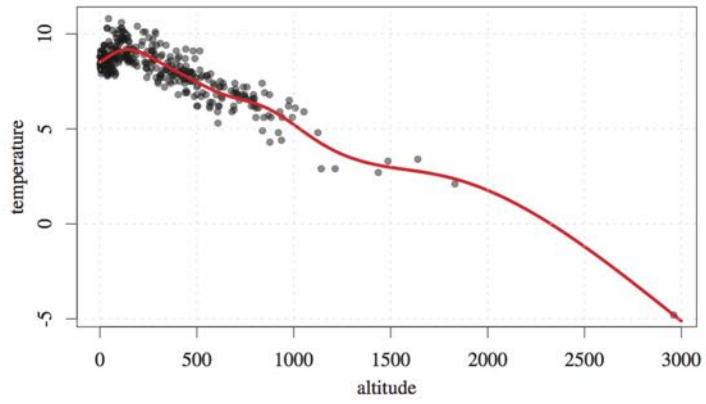
This tells us which entries of $\mathbf{x}(t)$ cause the future of others $\mathbf{x}(t + dt)$, i.e., the causal structure.

https://arxiv.org/abs/1911.10500

What is cause and what is effect?



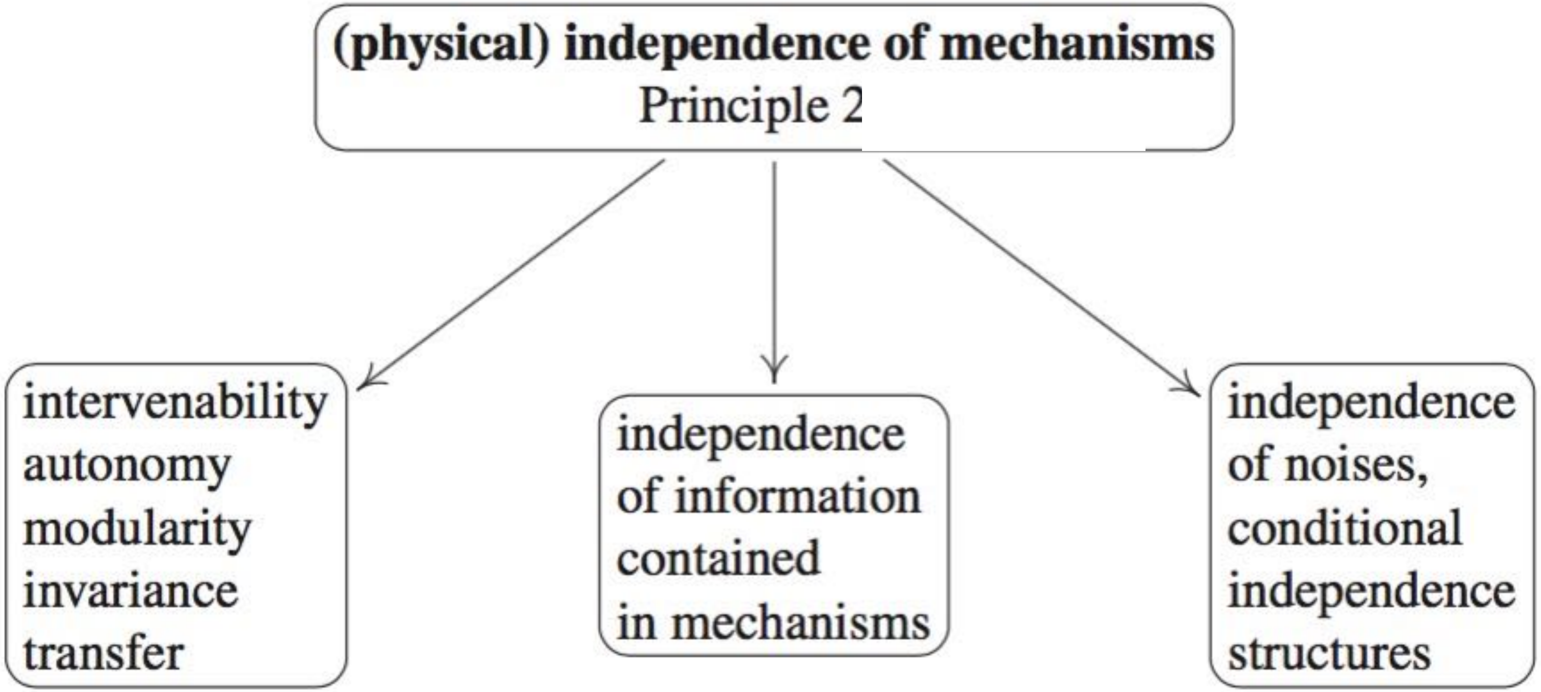
$$p(a,t) = p(a|t)\ p(t) \quad T \to A$$
$$= p(t|a)\ p(a) \quad A \to T$$



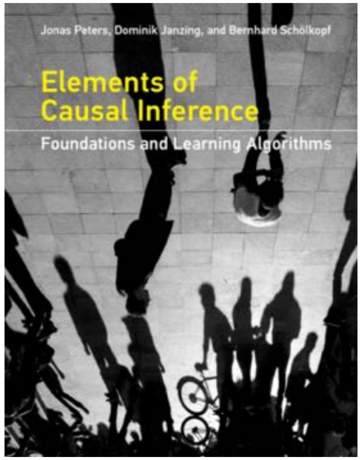
- **intervention** on a: raise the city, find that t changes
- hypothetical intervention on a: still expect that t changes, since we can think of a physical mechanism $p(t|a)$ that is **independent** of $p(a)$
- we expect that $p(t|a)$ is **invariant** across, say, different countries in a similar climate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

Independent Causal Mechanisms Principle (ICM):

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

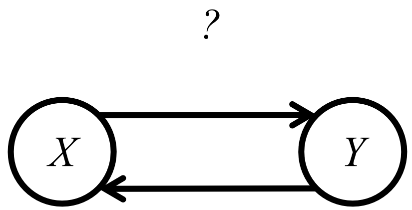


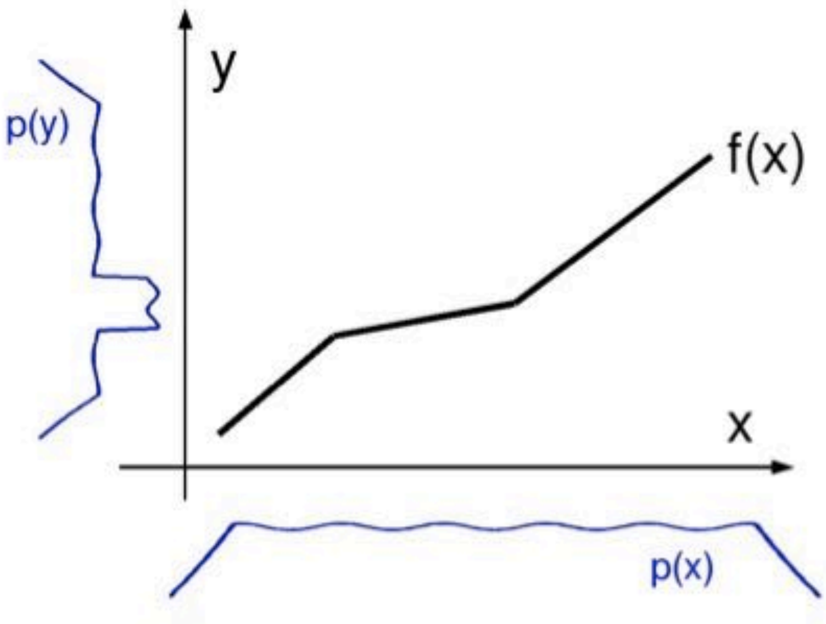
Peters, Janzing, Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms.* MIT Press, 2017
http://www.math.ku.dk/~peters/jonas_files/bookDRAFT11-online-2017-06-28.pdf



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):

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

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:

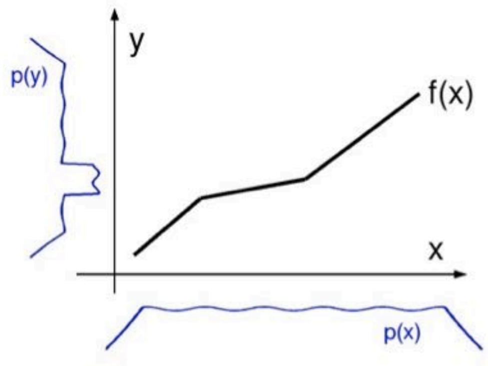
$$S(p_y) - S(p_x) = S(v_y) - S(u_x)$$

Orthogonality (information geometric):

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

which can be rewritten as

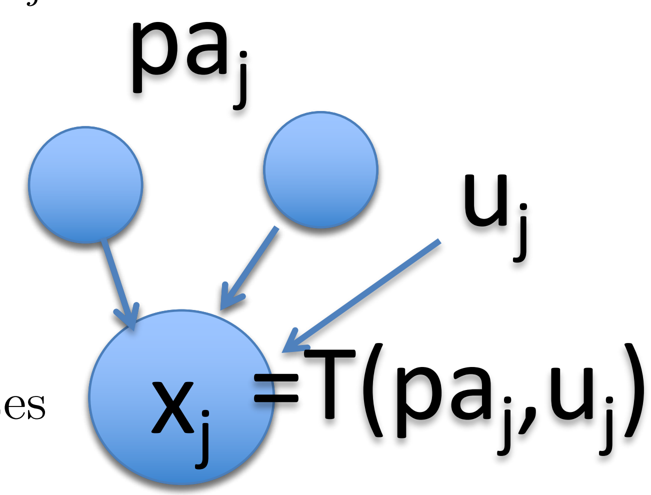
$$D(p_y \| u_y) = D(p_x \| u_x) + D(v_y \| u_y)$$



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



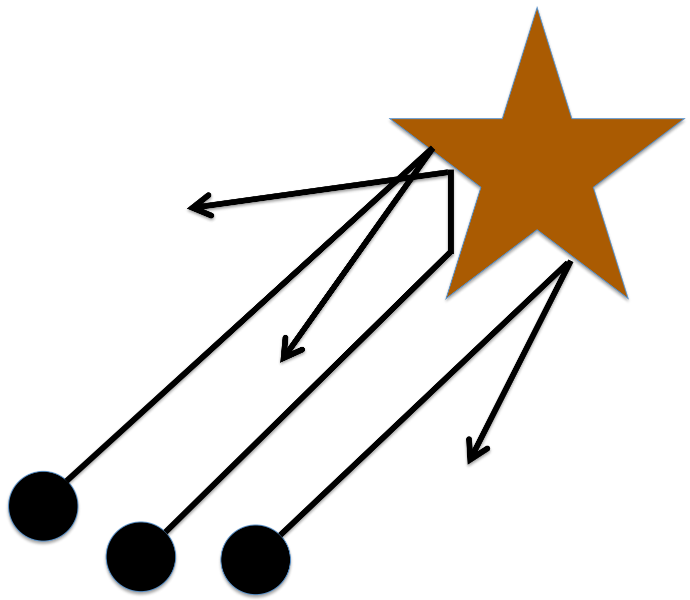
- all u_j are jointly independent
- the program u_j represents the causal mechanism that generates the effect from its causes
- u_j are the analog of the unobserved noise terms in the statistical functional model

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

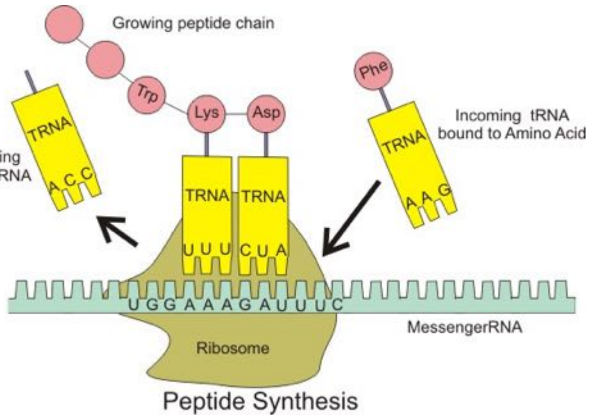
$$K(M(s)) \stackrel{+}{\geq} K(s)$$

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}.

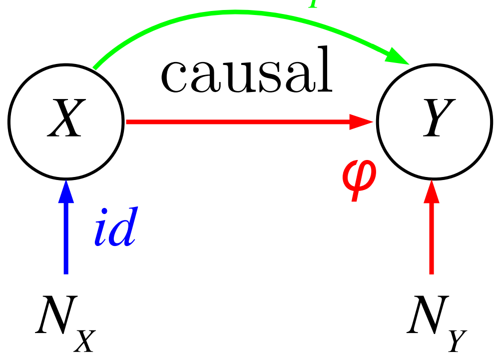
Janzing, Chaves, Schölkopf: Algorithmic independence of initial condition and dynamical law in thermodynamics and causal inference. New J. of Physics, 2016

Using cause-effect knowledge

- example 1: predict protein from mRNA sequence

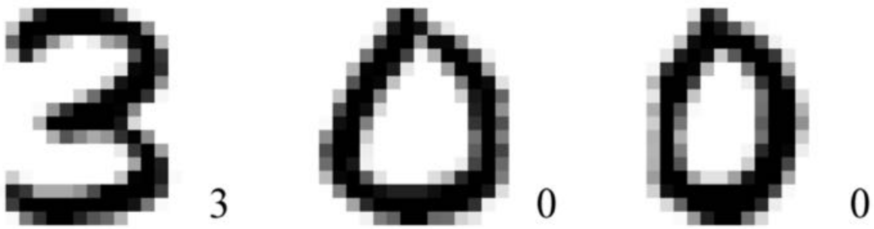


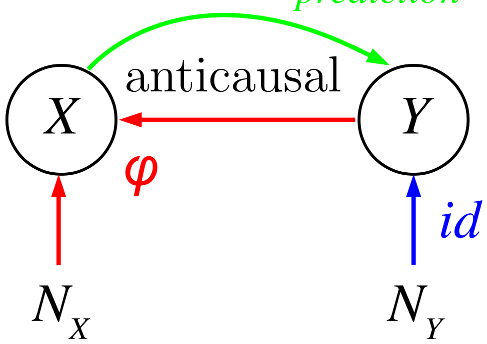
Source: http://commons.wikimedia.org/wiki/File:Peptide_syn.png



causal mechanism φ

- example 2: predict class membership from handwritten digit





Covariate Shift and Semi-Supervised Learning

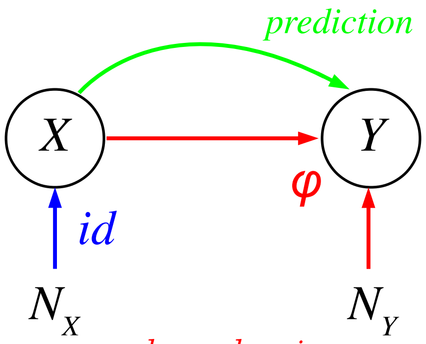
Assumption: $p(C)$ and mechanism $p(E|C)$ "independent"
Goal: learn $X \mapsto Y$, i.e., estimate (properties of) $p(Y|X)$

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 $p(X)$*

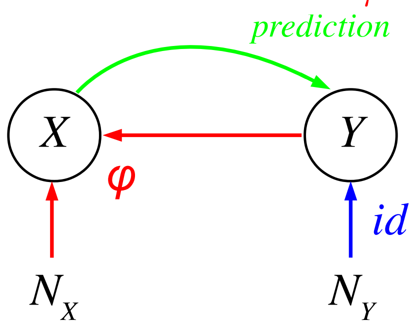


Anticausal learning

$p(Y)$ 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

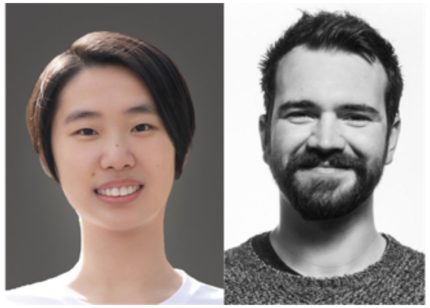
- Experimental Meta-Analysis confirms prediction

  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)

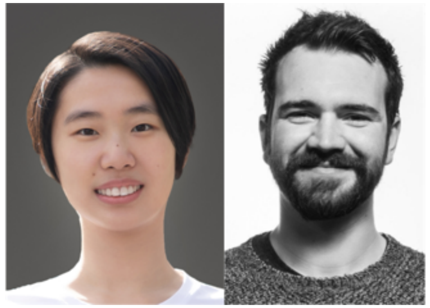
Effect = CausalMechanism (Cause, Noise)

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

ICM in NLP: Findings

(with Zhijing Jin & Julius von Kügelgen)

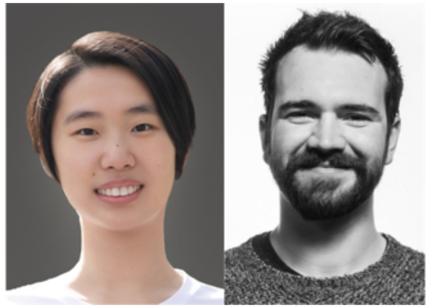


Causal direction corresponds to shorter description of machine translation data in terms of minimum description length (MDL):

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:

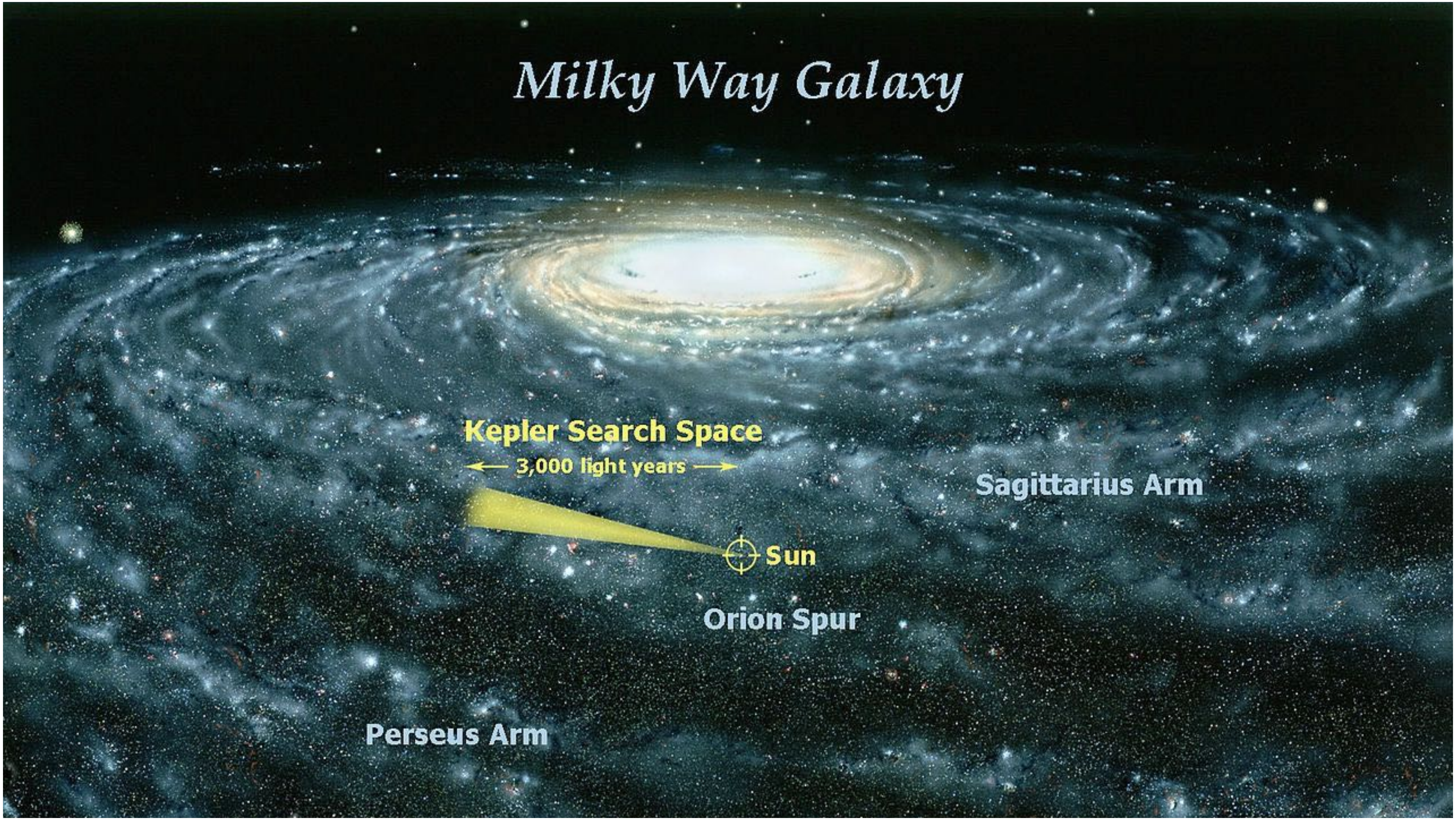
Semi-supervised learning (SSL): *anticausal* > *causal*.

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

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

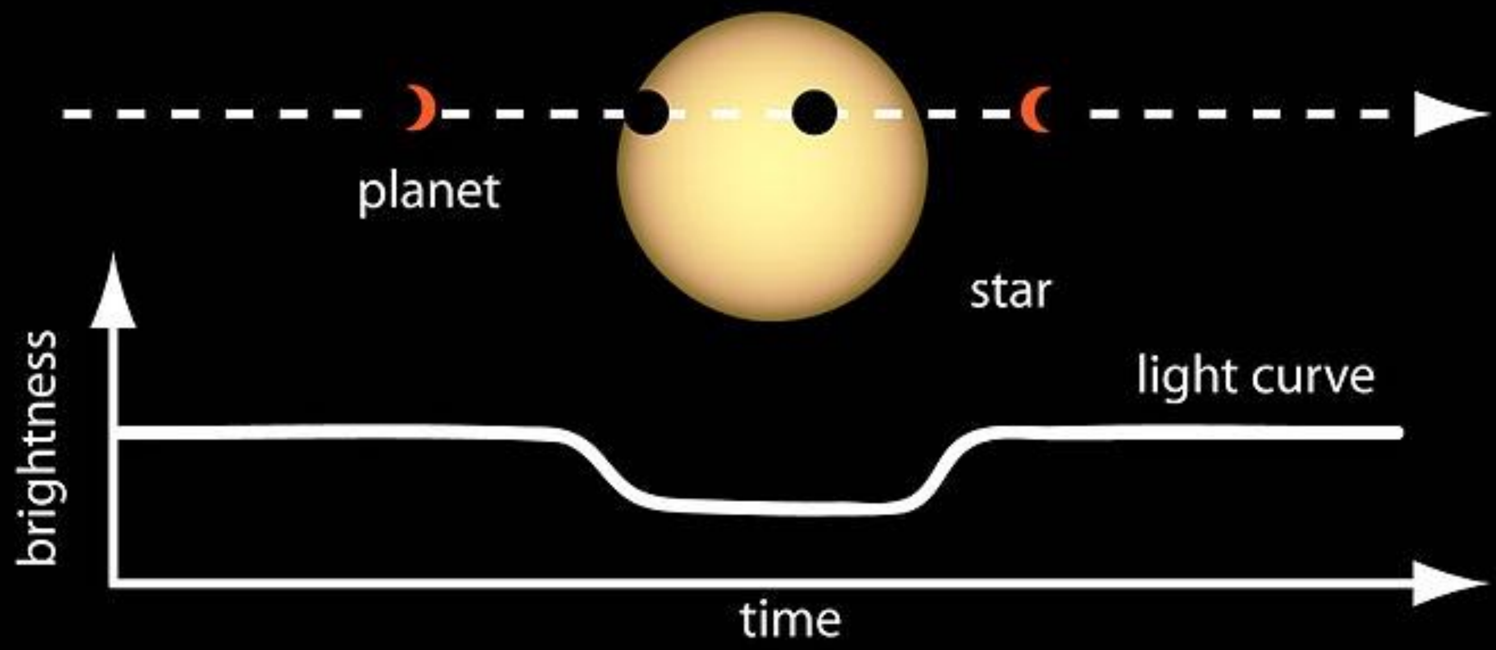
Task Type	Mean ΔDA ($\pm$std)	According to ICM
Causal	5.18 ($\pm$6.57)	Larger
Anticausal	1.26 ($\pm$1.79)	Smaller

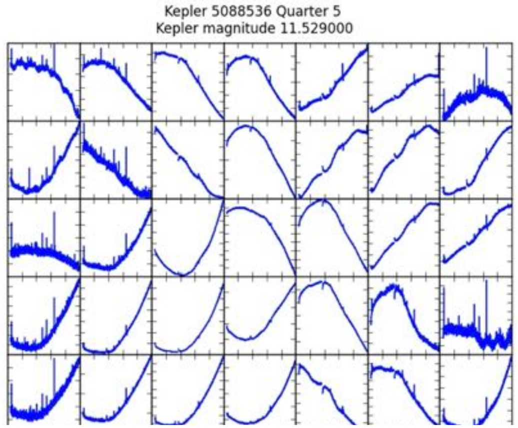
Causal Modeling for Confounder Removal in Exoplanet Detection

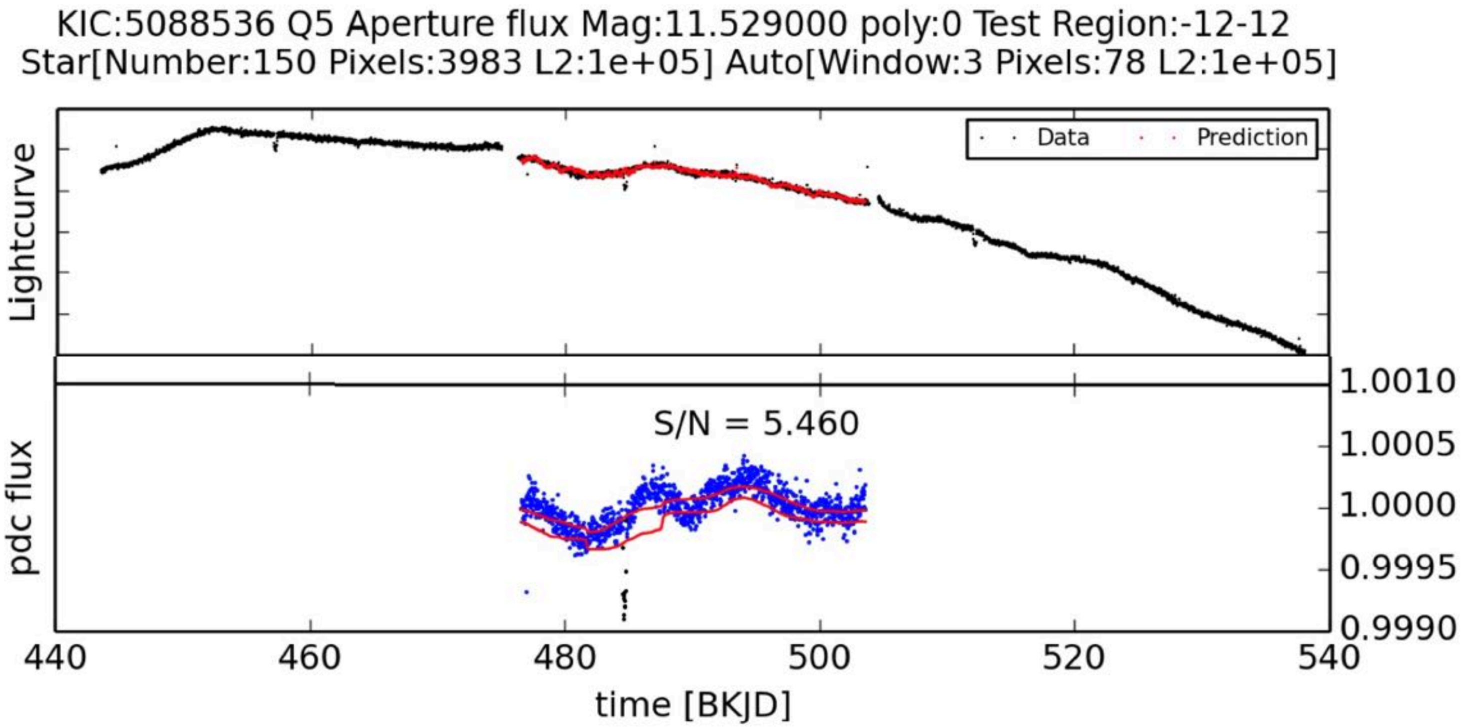


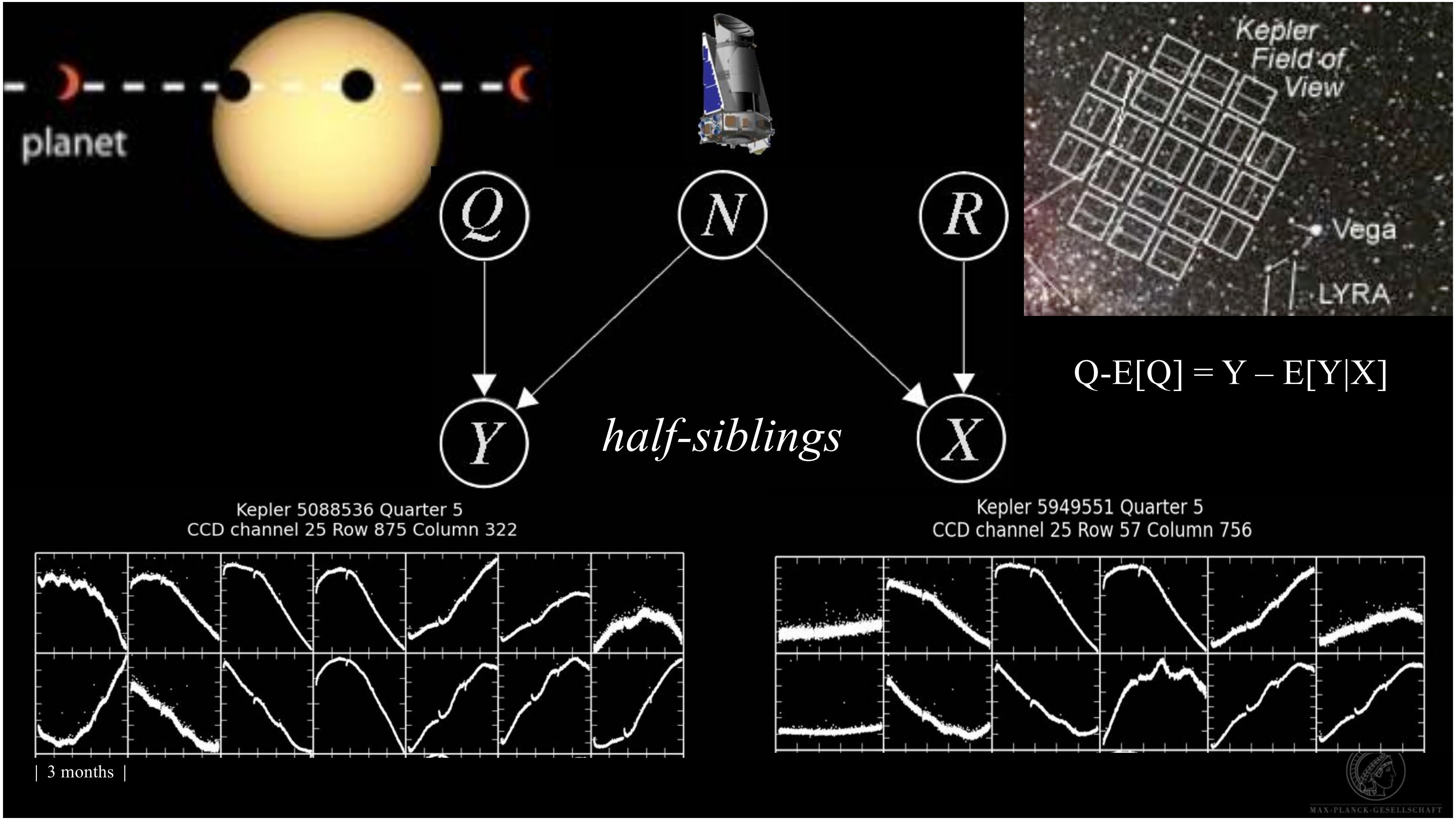
Milky Way Galaxy
Kepler Search Space
3,000 light years
Sagittarius Arm
Sun
Orion Spur
Perseus Arm

Exoplanet Transits

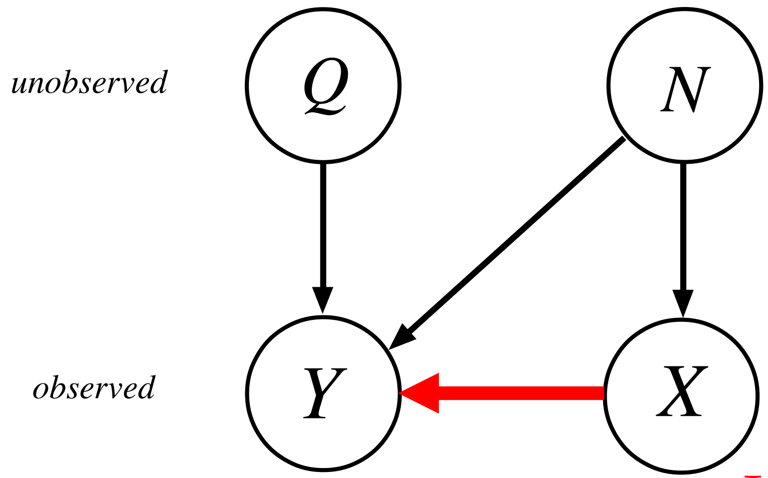



Kepler 5088536 Quarter 5
Kepler magnitude 11.529000


KIC:5088536 Q5 Aperture flux Mag:11.529000 poly:0 Test Region:-12-12
Star[Number:150 Pixels:3983 L2:1e+05] Auto[Window:3 Pixels:78 L2:1e+05]
Data
Prediction
Lightcurve
S/N = 5.460
pdc flux
1.0010
1.0005
1.0000
0.9995
0.9990
440
460
480
500
520
540
time [BKJD]


planet
Kepler Field of View
Vega
LYRA
Q
N
R
Y
X
half-siblings
Q-E[Q] = Y – E[Y|X]
Kepler 5088536 Quarter 5
CCD channel 25 Row 875 Column 322
Kepler 5949551 Quarter 5
CCD channel 25 Row 57 Column 756
| 3 months |
MAX-PLANCK-GESELLSCHAFT

Half-Sibling Regression



Idea: remove $E[Y|X]$ from Y to reconstruct Q.

$X \perp\!\!\!\perp Q$

X and Y share information (only) through N

If we try to predict Y from X, we only pick up the part due to N

with David Hogg, Dan Foreman-Mackey, Dun Wang, Dominik Janzing, Jonas Peters, Carl-Johann Simon-Gabriel *(ICML 2015)*

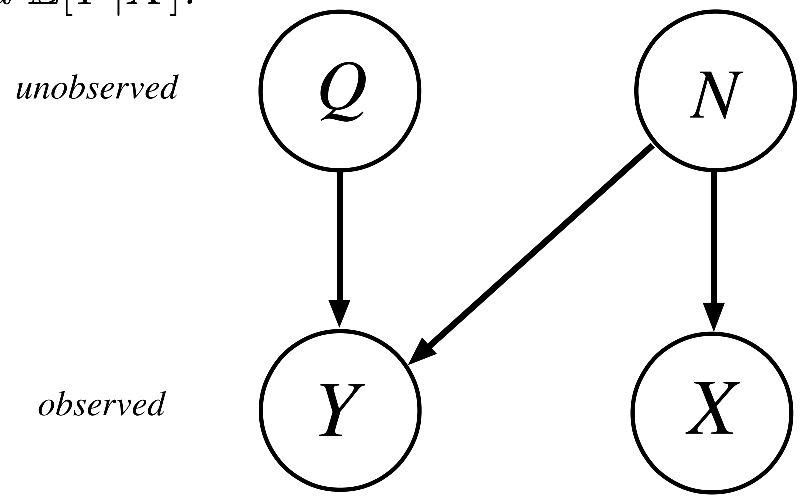
Proposition. Q, N, Y, X random variables, $X \perp\!\!\!\perp Q$, and f measurable.
Suppose

- $Y = Q + f(N)$ *(additive noise model)*

- $f(N) = \psi(X)$ for some ψ *(complete information).*

Then $\hat{Q} := Y - \mathbb{E}[Y|X] = Q - \mathbb{E}[Q]$.

Q can be reconstructed, up to a constant offset, from Y and $\mathbb{E}[Y|X]$.

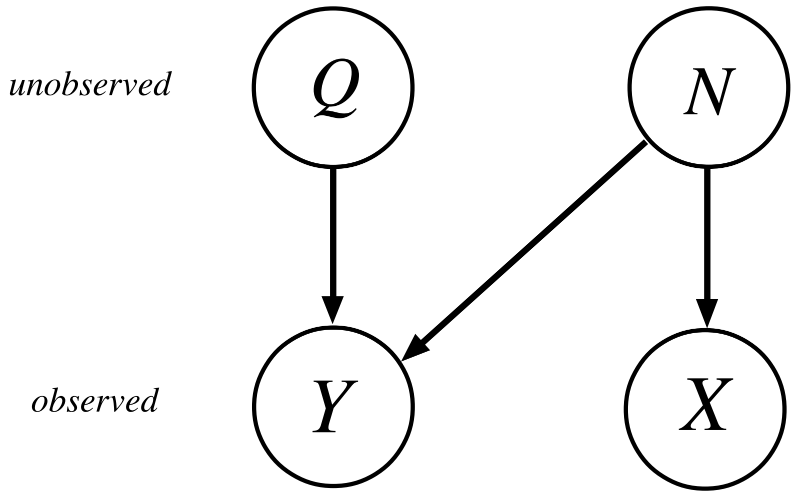


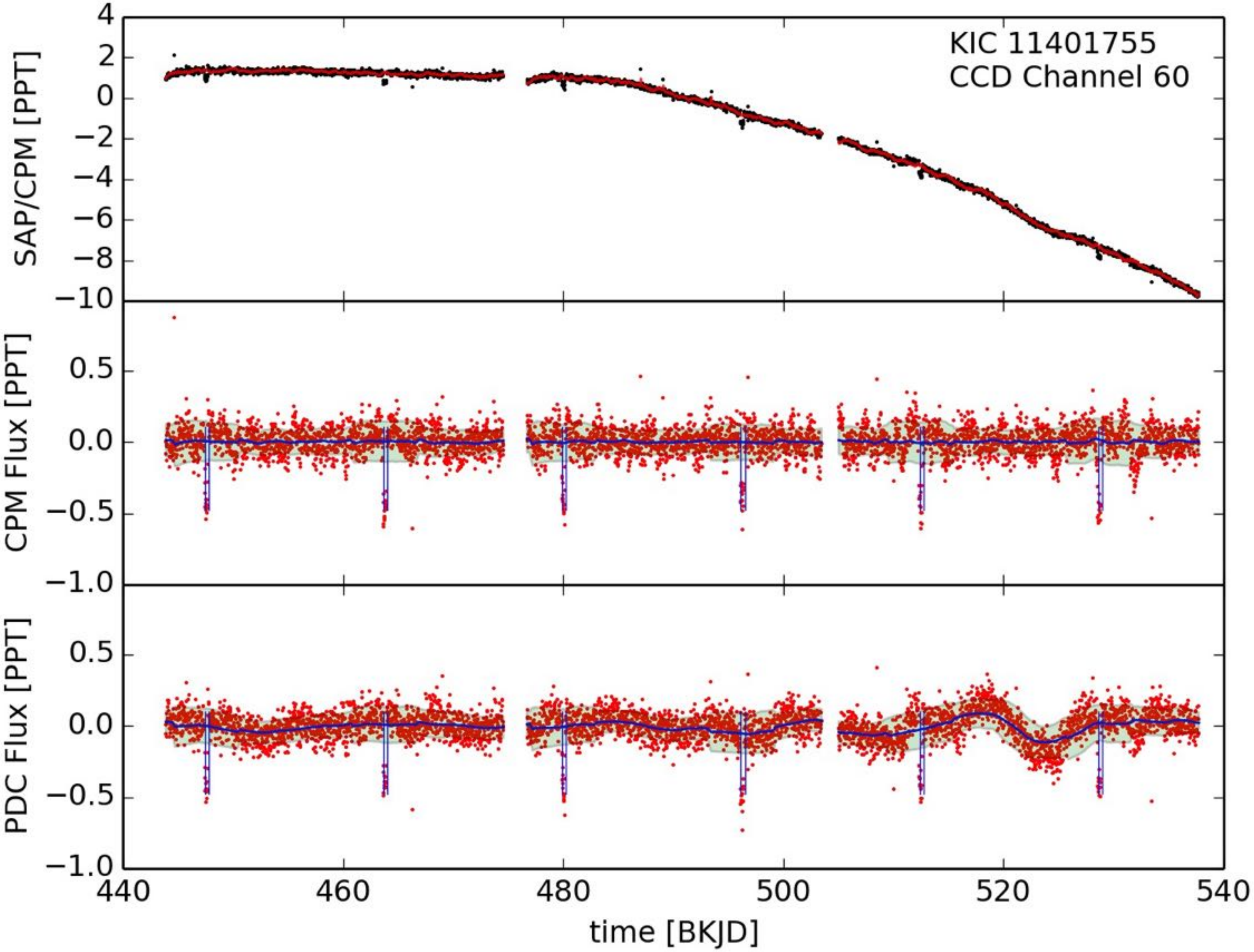
Proposition. Q, N, Y, X random variables, $X \perp\!\!\!\perp Q$, and f measurable.
Suppose

- $Y = Q + f(N)$ *(additive noise model)*

Then $E[(\hat{Q} - (Q - E[Q]))^2] = E[\mathrm{Var}[f(N)|X]]$.

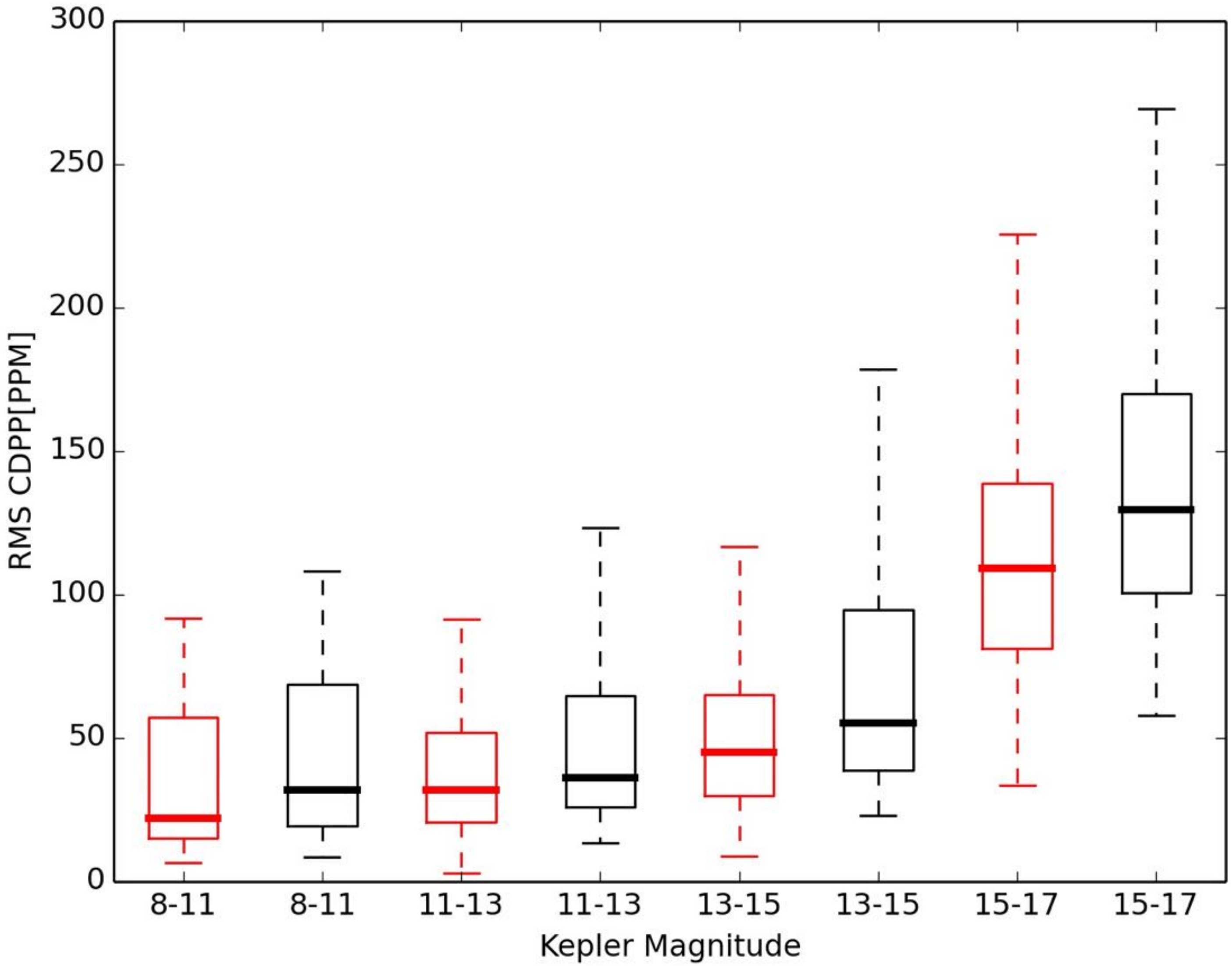
If $f(N)$ can (in principle) be predicted well from X, then Q can be reconstructed well.





Bernhard Schölkopf

MAX-PLANCK-GESELLSCHAFT

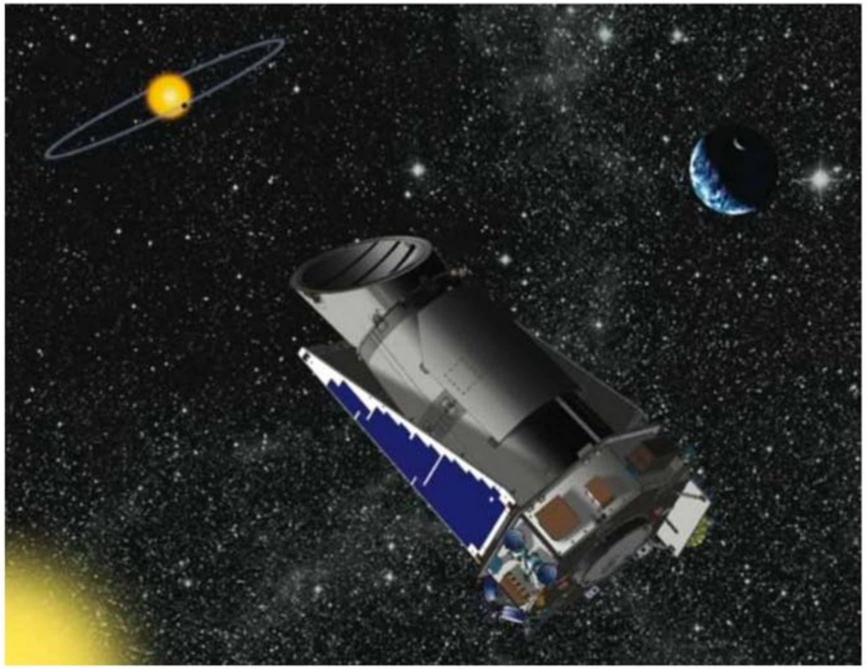


Bernhard Schölkopf

MAX-PLANCK-GESELLSCHAFT

Planet-Hunting Kepler Spacecraft Suffers Major Failure, NASA Says

By Mike Wall May 15, 2013 Science & Astronomy

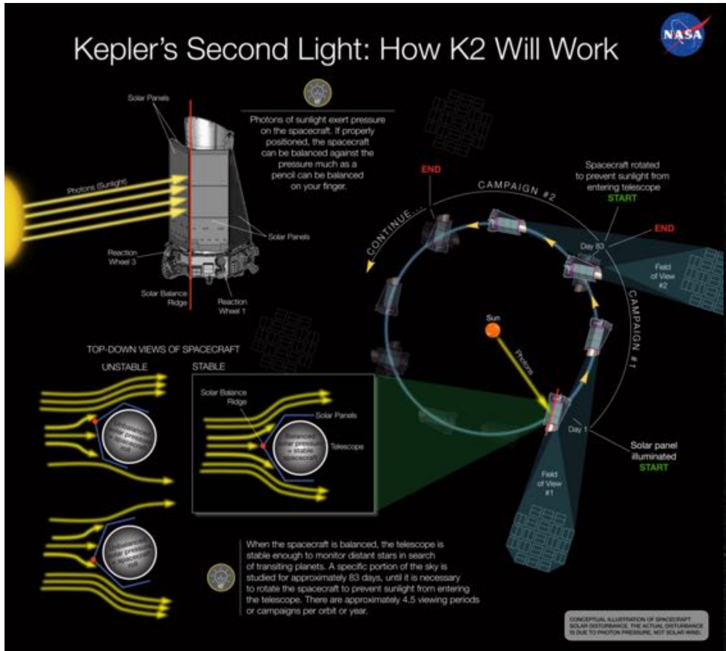


An artist's interpretation of the Kepler observatory in space. (Image: © NASA.)

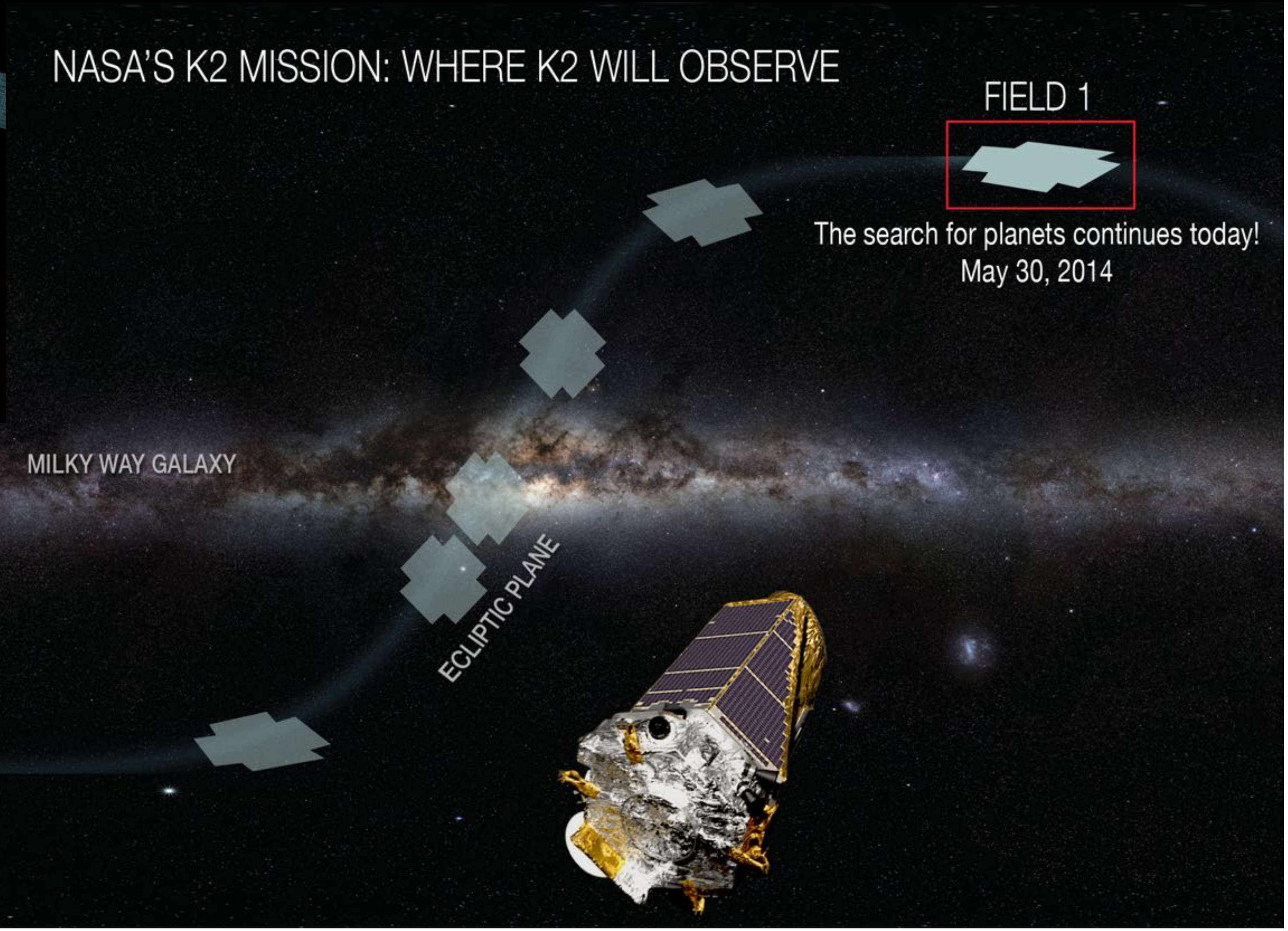
This story was updated at 5:20 p.m. EDT.

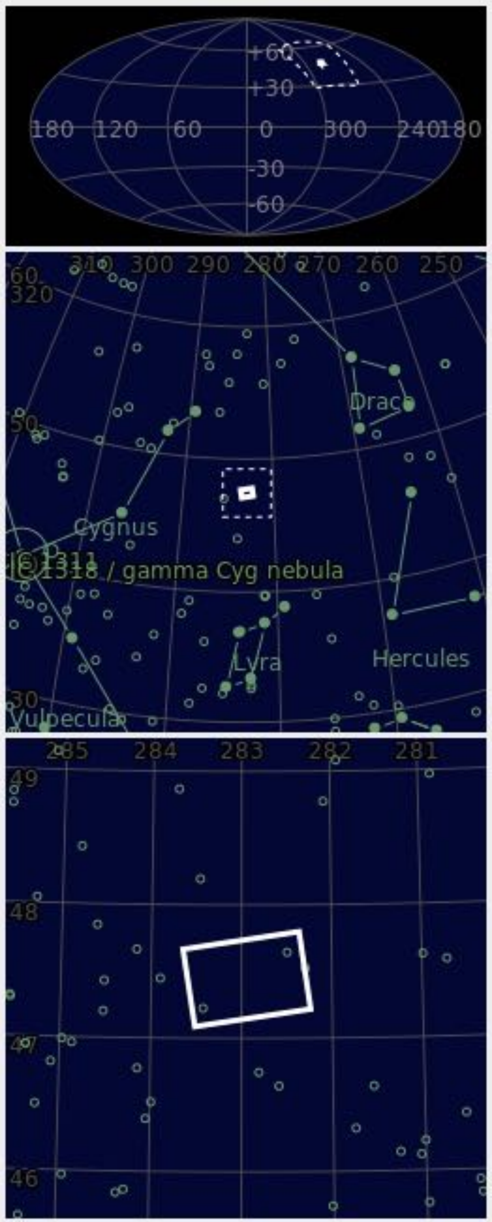
The planet-hunting days of NASA's prolific Kepler space telescope, which has discovered more than 2,700 potential alien worlds to date, may be over.

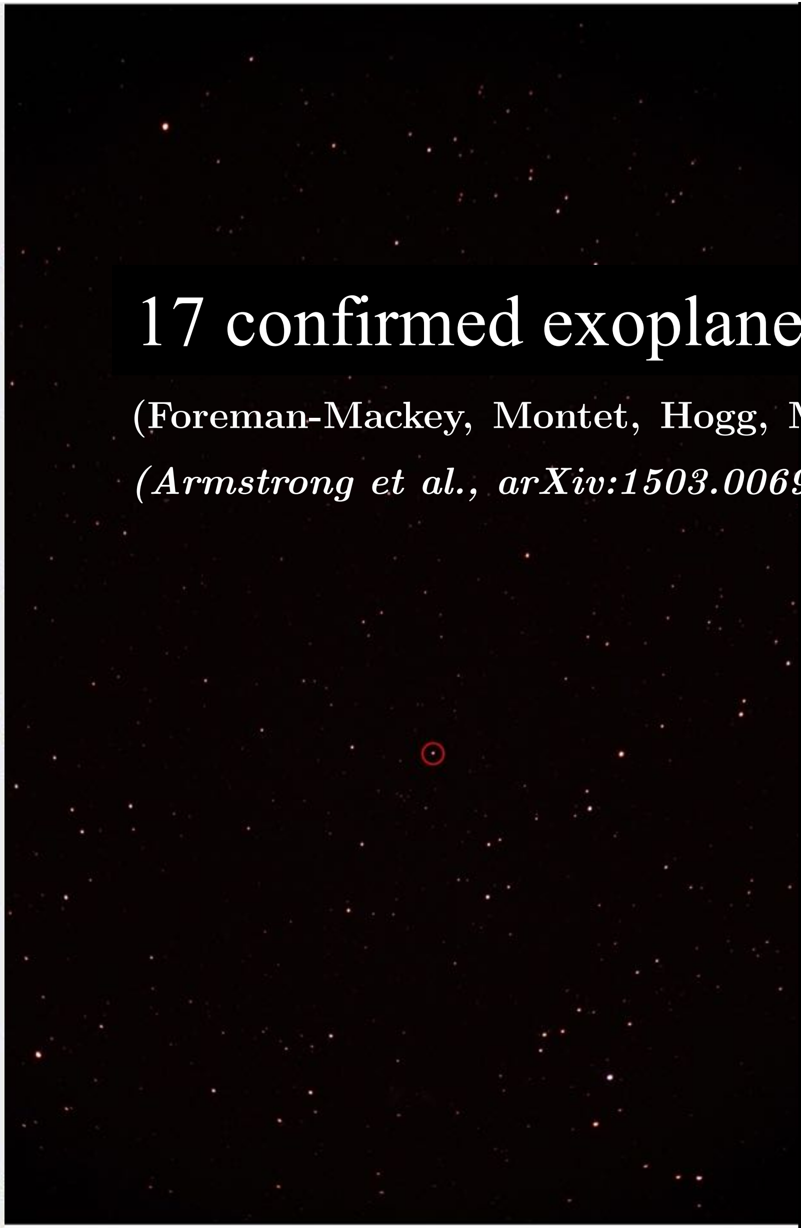
The second of Kepler's four reaction wheels — devices that allow the observatory to maintain its position in space — has failed, NASA officials announced Wednesday (May 15).



Credits: NASA Ames/W Stenzel







17 confirmed exoplanets

(Foreman-Mackey, Montet, Hogg, Morton, Wang, Schölkopf, arXiv:1502.04715)*:*
(Armstrong et al., arXiv:1503.00692; Montet et al., arXiv:1503.07866).

The Astrophysical Journal, 806:215 (13pp), 2015 June 20 doi:10.1088/0004-637X/806/2/215

A SYSTEMATIC SEARCH FOR TRANSITING PLANETS IN THE *K2* DATA

Daniel Foreman-Mackey[1], Benjamin T. Montet[2,3], David W. Hogg[1,4,5],
Timothy D. Morton[6], Dun Wang[1], and Bernhard Schölkopf[7]
[1] Center for Cosmology and Particle Physics, Department of Physics, New York University,
4 Washington Place, New York, NY 10003, USA; danfm@nyu.edu
[2] Cahill Center for Astronomy and Astrophysics, California Institute of Technology, Pasadena, CA 91125, USA
[3] Harvard-Smithsonian Center for Astrophysics, Cambridge, MA 02138, USA
[4] Max-Planck-Institut für Astronomie, Königstuhl 17, D-69117, Heidelberg, Germany
[5] Center for Data Science, New York University, 726 Broadway, 7th Floor, New York, NY 10003, USA
[6] Department of Astrophysics, Princeton University, Princeton, NJ 08544, USA
[7] Max Planck Institute for Intelligent Systems, Spemannstrasse 38, D-72076, Tübingen, Germany
Received 2015 February 16; accepted 2015 May 12; published 2015 June 18

ABSTRACT

Photometry of stars from the *K2* extension of NASA's *Kepler* mission is afflicted by systematic effects caused by small (few-pixel) drifts in the telescope pointing and other spacecraft issues. We present a method for searching *K2* light curves for evidence of exoplanets by simultaneously fitting for these systematics and the transit signals of interest. This method is more computationally expensive than standard search algorithms but we demonstrate that it can be efficiently implemented and used to discover transit signals. We apply this method to the full Campaign 1 data set and report a list of 36 planet candidates transiting 31 stars, along with an analysis of the pipeline performance and detection efficiency based on artificial signal injections and recoveries. For all planet candidates, we present posterior distributions on the properties of each system based strictly on the transit observables.

Key words: catalogs – methods: data analysis – methods: statistical – planetary systems – stars: statistics

1. INTRODUCTION

The *Kepler* Mission was incredibly successful at finding transiting exoplanets in the light curves of stars. The Mission

a few percent of the data are actually stored and downloaded to Earth, there is not enough information in the data to derive or infer a complete or accurate flat-field map. Therefore, work on

Table 2
The Catalog of Planet Candidates and their Observable Properties

EPIC	*Kepler* mag	R.A. (J2000)	Decl. (J2000)	P (days)	t_0 [BJD-2456808]	$R_P/R_\star$
201208431	14.41	174.745640	−3.905585	$10.0040^{+0.0018}_{-0.0016}$	$7.5216^{+0.0098}_{-0.0090}$	$0.0349^{+0.0034}_{-0.0026}$
201257461	11.51	178.161109	−3.094936	$50.2677^{+0.0083}_{-0.0074}$	$20.3735^{+0.0147}_{-0.0098}$	$0.0334^{+0.0054}_{-0.0017}$
201295312	12.13	174.011630	−2.520881	$5.6562^{+0.0007}_{-0.0007}$	$3.7228^{+0.0086}_{-0.0091}$	$0.0175^{+0.0020}_{-0.0009}$
201338508	14.36	169.303502	−1.877976	$10.9328^{+0.0022}_{-0.0021}$	$6.5967^{+0.0088}_{-0.0081}$	$0.0339^{+0.0025}_{-0.0030}$
201338508	14.36	169.303502	−1.877976	$5.7350^{+0.0006}_{-0.0006}$	$0.8626^{+0.0054}_{-0.0055}$	$0.0331^{+0.0025}_{-0.0023}$
201367065	11.57	172.334949	−1.454787	$10.0542^{+0.0004}_{-0.0004}$	$5.4186^{+0.0018}_{-0.0018}$	$0.0354^{+0.0022}_{-0.0011}$
201367065	11.57	172.334949	−1.454787	$24.6470^{+0.0014}_{-0.0016}$	$4.2769^{+0.0030}_{-0.0029}$	$0.0272^{+0.0016}_{-0.0013}$
201384232	12.51	178.192260	−1.198477	$30.9375^{+0.0029}_{-0.0052}$	$19.5035^{+0.0053}_{-0.0039}$	$0.0260^{+0.0011}_{-0.0011}$
201393098	13.05	167.093771	−1.065755	$28.6793^{+0.0105}_{-0.0116}$	$16.6212^{+0.0305}_{-0.0177}$	$0.0231^{+0.0028}_{-0.0020}$
201403446	11.99	174.266344	−0.907261	$19.1535^{+0.0050}_{-0.0050}$	$7.3437^{+0.0116}_{-0.0143}$	$0.0154^{+0.0014}_{-0.0013}$
201445392	14.38	169.793665	−0.284375	$10.3527^{+0.0011}_{-0.0011}$	$5.6110^{+0.0047}_{-0.0051}$	$0.0349^{+0.0045}_{-0.0025}$
201445392	14.38	169.793665	−0.284375	$5.0644^{+0.0006}_{-0.0006}$	$5.0690^{+0.0059}_{-0.0064}$	$0.0274^{+0.0025}_{-0.0020}$
201465501	14.96	176.264468	0.005301	$18.4488^{+0.0015}_{-0.0015}$	$14.6719^{+0.0035}_{-0.0032}$	$0.0531^{+0.0061}_{-0.0039}$
201505350	12.81	174.960319	0.603575	$11.9069^{+0.0005}_{-0.0004}$	$9.2764^{+0.0013}_{-0.0015}$	$0.0446^{+0.0009}_{-0.0006}$
201505350	12.81	174.960319	0.603575	$7.9193^{+0.0001}_{-0.0001}$	$5.3840^{+0.0006}_{-0.0008}$	$0.0747^{+0.0016}_{-0.0013}$
201546283	12.43	171.515165	1.230738	$6.7713^{+0.0001}_{-0.0001}$	$4.8453^{+0.0012}_{-0.0011}$	$0.0481^{+0.0020}_{-0.0012}$
201549860	13.92	170.103081	1.285956	$5.6083^{+0.0005}_{-0.0006}$	$4.1195^{+0.0045}_{-0.0047}$	$0.0283^{+0.0041}_{-0.0023}$
201555883	15.06	176.075940	1.375947	$5.7966^{+0.0002}_{-0.0002}$	$5.3173^{+0.0027}_{-0.0050}$	$0.0604^{+0.0068}_{-0.0032}$
201565013	16.91	176.992193	1.510249	$8.6381^{+0.0003}_{-0.0002}$	$3.4283^{+0.0016}_{-0.0015}$	$0.1538^{+0.0355}_{-0.0243}$
201569483	11.77	167.171299	1.577513	$5.7969^{+0.0000}_{-0.0000}$	$5.3130^{+0.0002}_{-0.0003}$	$0.3587^{+0.0379}_{-0.0334}$
201577035	12.30	172.121957	1.690636	$19.3062^{+0.0013}_{-0.0013}$	$11.5790^{+0.0025}_{-0.0027}$	$0.0380^{+0.0023}_{-0.0012}$
201596316	13.15	169.042002	1.986840	$39.8415^{+0.0136}_{-0.0155}$	$21.8572^{+0.0120}_{-0.0101}$	$0.0267^{+0.0034}_{-0.0022}$
201613023	12.14	173.192036	2.244884	$8.2818^{+0.0006}_{-0.0007}$	$7.3752^{+0.0055}_{-0.0052}$	$0.0205^{+0.0012}_{-0.0008}$
201617985	14.11	179.491659	2.321476	$7.2823^{+0.0007}_{-0.0008}$	$4.6337^{+0.0050}_{-0.0050}$	$0.0333^{+0.0072}_{-0.0032}$
201629650	12.73	170.155528	2.502696	$40.0492^{+0.0186}_{-0.0259}$	$4.5363^{+0.0202}_{-0.0172}$	$0.0241^{+0.0025}_{-0.0020}$
201635569	15.55	178.057026	2.594245	$8.3681^{+0.0002}_{-0.0002}$	$3.4514^{+0.0015}_{-0.0014}$	$0.0991^{+0.0120}_{-0.0078}$
201649426	13.22	177.234262	2.807619	$27.7704^{+0.0001}_{-0.0001}$	$13.3476^{+0.0001}_{-0.0002}$	$0.4365^{+0.0777}_{-0.0583}$
201702477	14.43	175.240794	3.681584	$40.7365^{+0.0026}_{-0.0025}$	$3.5451^{+0.0026}_{-0.0025}$	$0.0808^{+0.0043}_{-0.0114}$
201736247	14.40	178.110797	4.254747	$11.8106^{+0.0016}_{-0.0019}$	$3.8483^{+0.0093}_{-0.0071}$	$0.0347^{+0.0030}_{-0.0024}$
201754305	14.30	175.097258	4.557340	$19.0726^{+0.0048}_{-0.0049}$	$1.4893^{+0.0128}_{-0.0133}$	$0.0297^{+0.0042}_{-0.0030}$
201754305	14.30	175.097258	4.557340	$7.6202^{+0.0012}_{-0.0011}$	$3.6813^{+0.0061}_{-0.0057}$	$0.0281^{+0.0034}_{-0.0026}$
201779067	11.12	168.542699	4.988131	$27.2429^{+0.0001}_{-0.0001}$	$12.2599^{+0.0002}_{-0.0003}$	$0.2535^{+0.0369}_{-0.0259}$
201828749	11.56	175.654342	5.894323	$33.5093^{+0.0023}_{-0.0018}$	$5.1554^{+0.0037}_{-0.0032}$	$0.0267^{+0.0021}_{-0.0020}$
201855371	13.00	178.329775	6.412261	$17.9715^{+0.0015}_{-0.0017}$	$9.9412^{+0.0033}_{-0.0038}$	$0.0311^{+0.0030}_{-0.0017}$
201912552	12.47	172.560460	7.588391	$32.9410^{+0.0039}_{-0.0032}$	$28.1834^{+0.0057}_{-0.0105}$	$0.0513^{+0.0035}_{-0.0056}$
201929294	12.97	174.656969	7.959611	$5.0084^{+0.0001}_{-0.0001}$	$4.5703^{+0.0022}_{-0.0012}$	$0.1163^{+0.0011}_{-0.0014}$

THE ASTROPHYSICAL JOURNAL, 809:25 (15pp), 2015 August 10 doi:10.1088/0004-637X/809/1/25

STELLAR AND PLANETARY PROPERTIES OF *K2* CAMPAIGN 1 CANDIDATES AND VALIDATION OF 17 PLANETS, INCLUDING A PLANET RECEIVING EARTH-LIKE INSOLATION

BENJAMIN T. MONTET[1,2], TIMOTHY D. MORTON[3], DANIEL FOREMAN-MACKEY[4,5], JOHN ASHER JOHNSON[2], DAVID W. HOGG[4,5,6], BRENDAN P. BOWLER[1,8], DAVID W. LATHAM[2], ALLYSON BIERYLA[2], AND ANDREW W. MANN[7,9]
[1] Cahill Center for Astronomy and Astrophysics, California Institute of Technology, Pasadena, CA 91125, USA; btm@astro.caltech.edu
[2] Harvard-Smithsonian Center for Astrophysics, Cambridge, MA 02138, USA
[3] Department of Astrophysics, Princeton University, Princeton, NJ 08544, USA
[4] Center for Cosmology and Particle Physics, Department of Physics, New York University, 4 Washington Place, New York, NY 10003, USA
[5] Center for Data Science, New York University, 726 Broadway, 7th Floor, New York, NY 10003, USA
[6] Max-Planck-Institut für Astronomie, Königstuhl 17, D-69117 Heidelberg, Germany
[7] Department of Astronomy, The University of Texas at Austin, Austin, TX 78712, USA
Received 2015 March 27; accepted 2015 July 2; published 2015 August 5

ABSTRACT

The extended *Kepler* mission, *K2*, is now providing photometry of new fields every three months in a search for transiting planets. In a recent study, Foreman-Mackey and collaborators presented a list of 36 planet candidates orbiting 31 stars in *K2* Campaign 1. In this contribution, we present stellar and planetary properties for all systems. We combine ground-based seeing-limited survey data and adaptive optics imaging with an automated transit analysis scheme to validate 21 candidates as planets, 17 for the first time, and identify 6 candidates as likely false positives. Of particular interest is K2-18 (EPIC 201912552), a bright ($K = 8.9$) M2.8 dwarf hosting a 2.23 ± 0.25 $R_\oplus$ planet with $T_{\rm eq} = 272 \pm 15$ K and an orbital period of 33 days. We also present two new open-source software packages which enable this analysis. The first, `isochrones`, is a flexible tool for fitting theoretical stellar models to observational data to determine stellar properties using a nested sampling scheme to capture the multimodal nature of the posterior distributions of the physical parameters of stars that may plausibly be evolved. The second is `vespa`, a new general-purpose procedure to calculate false positive probabilities and statistically validate transiting exoplanets.

Key words: catalogs – planetary systems – planets and satellites: detection – stars: fundamental parameters

Table 3
Planet Properties for All Objects of Interest

Candidate	Period (days)	Epoch (BJD-2456808)	Radius ($R_\oplus$)	$a/R_\star$	a (AU)	T_{eq} (K)	Disposition
201208431.01/K2-4b	10.00329 ± 0.00159	7.5212 ± 0.0080	2.37 ± 0.40	27.79 ± 0.72	0.0777 ± 0.0012	563 ± 11	Planet
201257461.01	50.27762 ± 0.00785	20.3735 ± 0.0397	209.52 ± 99.23	6.19 ± 0.52	0.3049 ± 0.0030	1466 ± 52	FP
201295312.01	5.65706 ± 0.00079	3.7187 ± 0.0082	2.16 ± 0.57	12.94 ± 4.07	0.0633 ± 0.0019	1211 ± 154	Candidate
201338508.01/K2-5c	10.93406 ± 0.00205	6.5947 ± 0.0080	1.92 ± 0.20	32.27 ± 0.71	0.0783 ± 0.0007	511 ± 9	Planet
201338508.02/K2-5b	5.73491 ± 0.00061	0.8640 ± 0.0063	1.92 ± 0.23	20.99 ± 0.46	0.0509 ± 0.0004	634 ± 12	Planet
201367065.01/K2-3b	10.05448 ± 0.00033	5.4177 ± 0.0015	1.98 ± 0.10	30.72 ± 0.75	0.0740 ± 0.0009	504 ± 9	Planet
201367065.02/K2-3c	24.64745 ± 0.00152	4.2759 ± 0.0030	1.56 ± 0.10	55.85 ± 1.36	0.1345 ± 0.0016	374 ± 7	Planet
201384232.01/K2-6b	30.94191 ± 0.00467	19.5014 ± 0.0090	2.50 ± 0.88	50.27 ± 24.56	0.1898 ± 0.0056	615 ± 105	Planet
201393098.01/K2-7b	28.67992 ± 0.00947	16.6155 ± 0.0149	2.67 ± 0.56	40.29 ± 8.19	0.1814 ± 0.0043	651 ± 61	Planet
201403446.01	19.15344 ± 0.00607	7.3412 ± 0.0152	2.04 ± 0.46	27.05 ± 5.87	0.1408 ± 0.0040	889 ± 88	Candidate
201445392.01/K2-8b	10.35176 ± 0.00133	5.6119 ± 0.0053	2.97 ± 0.51	24.94 ± 0.79	0.0856 ± 0.0012	691 ± 14	Planet
201445392.02	5.06468 ± 0.00063	5.0663 ± 0.0071	2.31 ± 0.33	15.49 ± 0.49	0.0531 ± 0.0008	877 ± 17	Candidate
201465501.01/K2-9b	18.44883 ± 0.00137	14.6723 ± 0.0030	1.60 ± 0.42	74.76 ± 6.66	0.0848 ± 0.0050	284 ± 14	Planet
201505350.01/K2-19c	11.90691 ± 0.00037	9.2764 ± 0.0018	4.31 ± 0.49	24.09 ± 2.48	0.0965 ± 0.0017	797 ± 42	Planet
201505350.02/K2-19b	7.91943 ± 0.00007	5.3836 ± 0.0005	7.11 ± 0.81	18.35 ± 1.89	0.0735 ± 0.0013	913 ± 48	Planet
201546283.01	6.77131 ± 0.00012	4.8440 ± 0.0022	5.77 ± 3.24	17.56 ± 9.24	0.0668 ± 0.0029	991 ± 239	Candidate
201549860.01	5.60840 ± 0.00055	4.1181 ± 0.0047	2.20 ± 0.40	17.42 ± 0.46	0.0555 ± 0.0008	766 ± 14	Candidate
201555883.01	⋯	⋯	⋯	⋯	⋯	⋯	FP[b]
201565013.01	8.63810 ± 0.00024	3.4284 ± 0.0016	15.99 ± 9.19	28.07 ± 2.68	0.0669 ± 0.0031	536 ± 37	Candidate
201569483.01	5.79687 ± 0.00000	5.3135 ± 0.0004	27.81 ± 3.56	15.68 ± 1.91	0.0589 ± 0.0015	930 ± 51	FP
201577035.01/K2-10b	19.30691 ± 0.00127	11.5768 ± 0.0033	3.92 ± 0.69	32.74 ± 5.15	0.1374 ± 0.0025	703 ± 55	Planet
201596316.01/K2-11b	39.93767 ± 0.23229	21.8290 ± 0.1156	7.55 ± 9.33	45.08 ± 58.53	0.2257 ± 0.0143	734 ± 253	Planet
201613023.01/K2-12b	8.28212 ± 0.00060	7.3734 ± 0.0054	2.33 ± 0.58	17.47 ± 5.05	0.0802 ± 0.0021	1003 ± 121	Planet
201617985.01	7.28161 ± 0.00078	4.6366 ± 0.0047	1.78 ± 0.43	26.04 ± 1.16	0.0586 ± 0.0012	518 ± 16	Candidate
201629650.01/K2-13b	39.91488 ± 0.32477	4.5250 ± 0.0146	1.89 ± 0.95	79.69 ± 63.37	0.2114 ± 0.0061	511 ± 126	Planet
201635569.01/K2-14b	8.36802 ± 0.00019	3.4513 ± 0.0013	4.81 ± 0.42	30.16 ± 0.69	0.0627 ± 0.0006	488 ± 8	Planet
201649426.01	27.77045 ± 0.00008	13.3482 ± 0.0012	32.79 ± 9.01	59.26 ± 13.58	0.1517 ± 0.0097	441 ± 42	FP
201702477.01	40.73620 ± 0.00266	3.5455 ± 0.0025	7.28 ± 1.10	56.98 ± 7.61	0.2205 ± 0.0053	529 ± 36	Candidate
201736247.01/K2-15b	11.81040 ± 0.00204	3.8509 ± 0.0076	2.48 ± 0.30	28.84 ± 1.98	0.0910 ± 0.0018	676 ± 26	Planet
201754305.01/K2-16c	19.07536 ± 0.00490	1.4854 ± 0.0119	2.14 ± 0.41	41.43 ± 1.34	0.1220 ± 0.0021	523 ± 12	Planet
201754305.02/K2-16b	7.62067 ± 0.00095	3.6802 ± 0.0054	2.13 ± 0.37	22.47 ± 0.73	0.0662 ± 0.0011	710 ± 16	Planet
201779067.01	27.24273 ± 0.00012	12.2601 ± 0.0003	31.73 ± 5.25	38.25 ± 3.72	0.1718 ± 0.0022	707 ± 34	FP
201828749.01	33.51569 ± 0.00232	5.1504 ± 0.0034	3.83 ± 3.25	67.09 ± 67.64	0.1875 ± 0.0090	613 ± 239	Candidate
201855371.01/K2-17b	17.96753 ± 0.00152	9.9462 ± 0.0035	2.23 ± 0.20	39.38 ± 0.85	0.1190 ± 0.0020	487 ± 10	Planet
201912552.01/K2-18b[a]	32.94488 ± 0.00281	28.1849 ± 0.0027	2.24 ± 0.23	83.83 ± 9.03	0.1491 ± 0.0055	272 ± 15	Planet
201929294.01	⋯	⋯	⋯	⋯	⋯	⋯	FP[b]

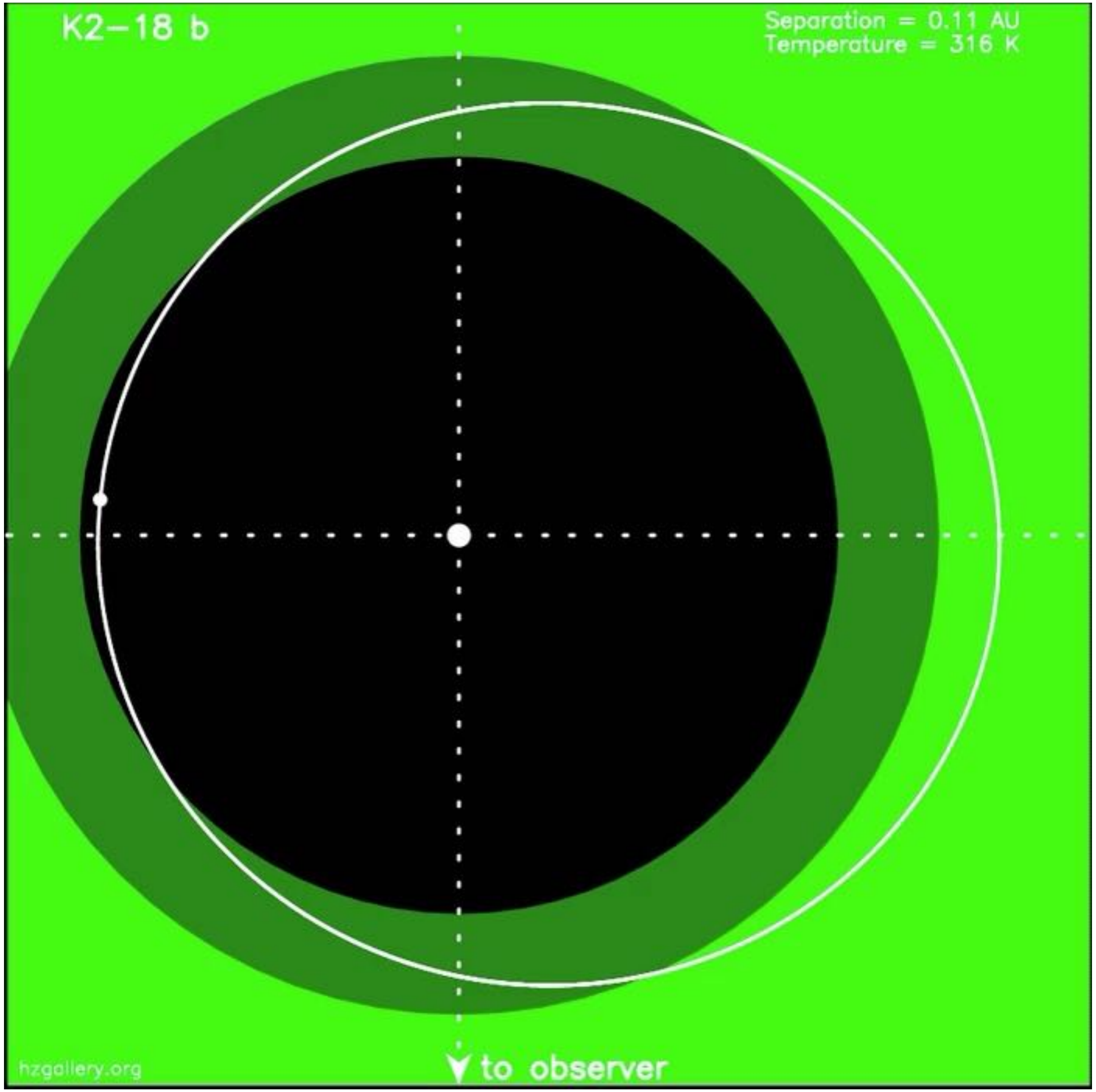
Habitable Zone Gallery

Home Plots Table Gallery Movies About Links

This site is dedicated to tracking the orbits of exoplanets in relation to their Habitable Zones.

Planets: 3706 **Systems:** 2806
Planets with orbits entirely within the Habitable Zone: 129 [?]
Updated: 2019 08 29 14:39:34 PDT

"The Earth is the only world known so far to harbor life. There is nowhere else, at least in the near future, to which our species could migrate. Visit, yes. Settle, not yet. Like it or not, for the moment the Earth is where we make our stand." - Carl Sagan


K2−18 b
Separation = 0.11 AU
Temperature = 316 K
hzgallery.org
to observer

NATIONAL GEOGRAPHIC

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.

The Guardian

3 MINUTE READ

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

tagesschau.de

Startseite | Videos & Audios | Inland | Ausland | Investigati

Startseite ▸ Ausland ▸ Erstmals Wasserdampf in Planetenatm

Astronomische Sensation

Wasserdampf auf Planet K2-18l

SCIENTIFIC AMERICAN

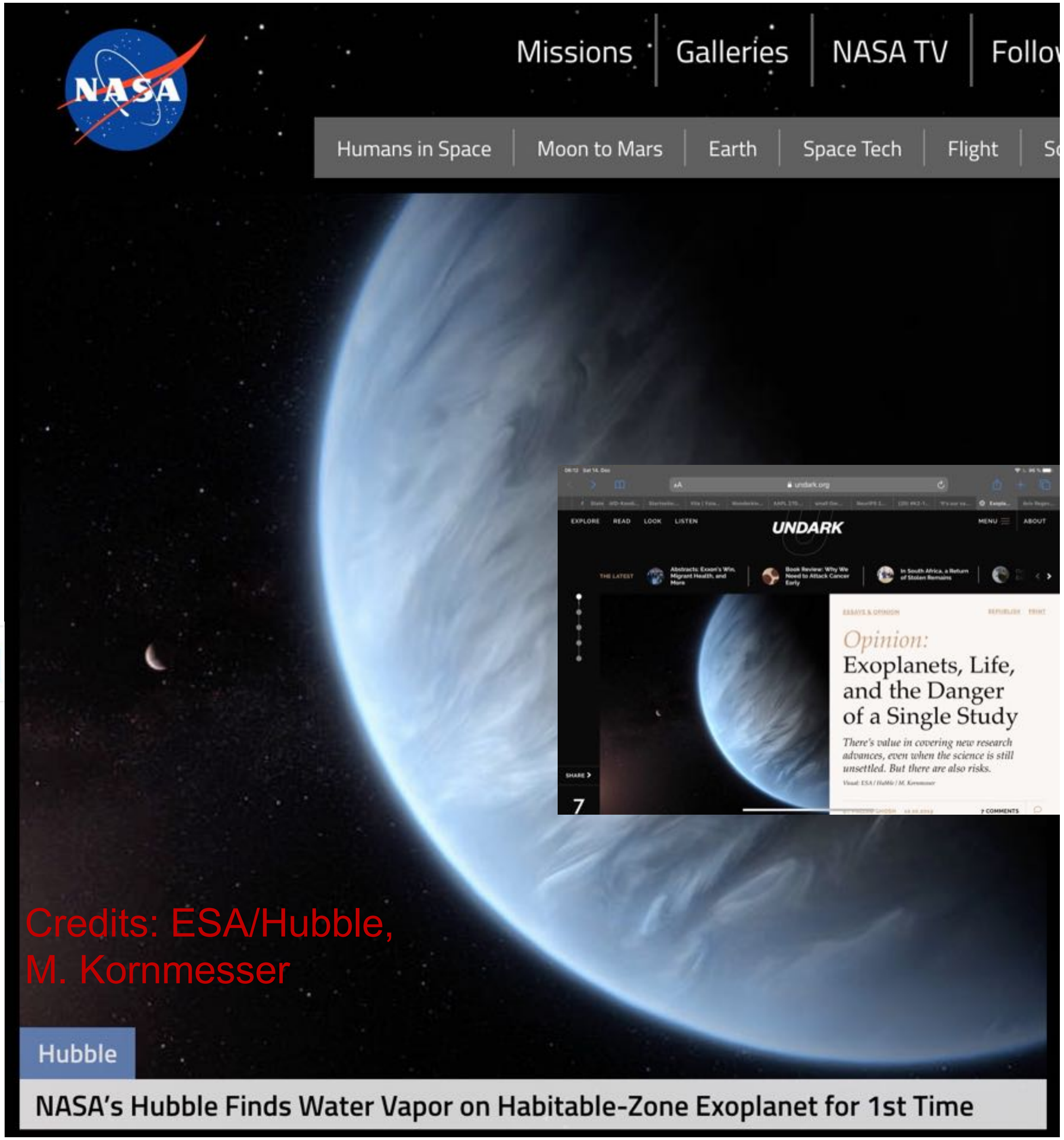
Subscribe

Observations

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



Draft version September 11, 2019
Typeset using LaTeX **twocolumn** style in AASTeX62

Water Vapor on the Habitable-Zone Exoplanet K2-18b

Björn Benneke,[1] Ian Wong,[2,3] Caroline Piaulet,[1] Heather A. Knutson,[4] Ian J.M. Crossfield,[5] Joshua Lothringer,[6] Caroline V. Morley,[7] Peter Gao,[8,3] Thomas P. Greene,[9] Courtney Dressing,[8] Diana Dragomir,[5,10] Andrew W. Howard,[11] Peter R. McCullough,[6] Eliza M.-R. Kempton,[12,13] Jonathan J. Fortney,[14] and Jonathan Fraine[15]

[1] *Institute for Research on Exoplanets and Department of Physics, Université de Montréal, Montreal, QC, Canada*
[2] *Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA, 02139, USA*
[3] *51 Pegasi b Fellow*
[4] *Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, CA 91125, USA*
[5] *Department of Physics and Kavli Institute of Astronomy, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA, 02139, USA*
[6] *Department of Physics and Astronomy, Johns Hopkins University, Baltimore, MD 21218, USA*
[7] *Department of Astronomy, University of Texas, Austin, TX 78712, USA*
[8] *Department of Astronomy, University of California - Berkeley , Berkeley, CA, 94720, USA*
[9] *NASA Ames Research Center, Moffett Field, CA, 94035, USA*
[10] *NASA Hubble Fellow*
[11] *Department of Astronomy, California Institute of Technology, Pasadena, CA 91125, USA*
[12] *Department of Astronomy, University of Maryland, College Park, MD 20742, USA*
[13] *Department of Physics, Grinnell College, 1116 8th Avenue, Grinnell, IA 50112, USA*
[14] *Department of Astronomy, University of California, Santa Cruz, CA 95064, USA*
[15] *Center for Extrasolar Planetary Systems, Space Science Institute, Boulder, CO 80301, USA*

ABSTRACT

Ever since the discovery of the first exoplanet, astronomers have made steady progress towards finding and probing planets in the habitable zone of their host stars, where the conditions could be right for liquid water to form and life to sprawl. Results from the Kepler mission indicate that the occurrence rate of habitable-zone Earths and super-Earths may be as high as 5–20%. Despite this abundance, probing the conditions and atmospheric properties on any of these habitable-zone planets is extremely difficult and has remained elusive to date. Here, we report the detection of water vapor and the likely presence of liquid water clouds in the atmosphere of the 8.6 $M_\oplus$ habitable-zone planet K2-18b. With a 33 day orbit around a cool M3 dwarf, K2-18b receives virtually the same amount of total radiation from its host star (1441 ± 80 W/m^2) as the Earth receives from the Sun (1370 W/m^2), making it a good candidate to host liquid water clouds. In this study we observed eight transits using HST/WFC3 in order to achieve the necessary sensitivity to detect water vapor. While the thick gaseous envelope of K2-18b means that it is not a true Earth analogue, our observations demonstrate that low-mass habitable-zone planets with the right conditions for liquid water are accessible with state-of-the-art telescopes.

Keywords: planets and satellites: individual (K2-18b) – planets and satellites: atmospheres

1. INTRODUCTION

The recent discovery of the transiting 8.63 ± 1.35 $M_\oplus$ exoplanet K2-18b in the habitable zone of a bright, nearby M3-dwarf provides us with an opportunity to carry out the spectroscopic study of the atmosphere of a habitable-zone planet outside our solar system (Montet et al. 2015, Benneke et al. 2017, Cloutier et al. 2019). K2-18b is an intriguing planet because its equilibrium temperature (265 ± 5K at an albedo of $A = 0.3$) is potentially very close to that of the Earth (257 K). The planet's predicted temperature provides the right con-

Corresponding author: Björn Benneke
bbenneke@astro.umontreal.ca

nature astronomy
LETTERS
https://doi.org/10.1038/s41550-019-0878-9

Water vapour in the atmosphere of the habitable-zone eight-Earth-mass planet K2-18 b

Angelos Tsiaras*, Ingo P. Waldmann*, Giovanna Tinetti, Jonathan Tennyson and Sergey N. Yurchenko

In the past decade, observations from space and the ground have found water to be the most abundant molecular species, after hydrogen, in the atmospheres of hot, gaseous extrasolar planets[1–5]. Being the main molecular carrier of oxygen, water is a tracer of the origin and the evolution mechanisms of planets. For temperate, terrestrial planets, the presence of water is of great importance as an indicator of habitable conditions. Being small and relatively cold, these planets and their atmospheres are the most challenging to observe, and therefore no atmospheric spectral signatures have so far been detected[6]. Super-Earths—planets lighter than ten Earth masses—around later-type stars may provide our first opportunity to study spectroscopically the characteristics of such planets, as they are best suited for transit observations. Here, we report the detection of a spectroscopic signature of water in the atmosphere of K2-18 b—a planet of eight Earth masses in the habitable zone of an M dwarf[7]—with high statistical confidence (Atmospheric Detectability Index[5] = 5.0, ~3.6σ (refs. [8,9])). In addition, the derived mean molecular weight suggests an atmosphere still containing some hydrogen. The observations were recorded with the Hubble Space Telescope/Wide Field Camera 3 and analysed with our dedicated, publicly available, algorithms[5,9]. Although the suitability of M dwarfs to host habitable worlds is still under discussion[10–13], K2-18 b offers an unprecedented opportunity to gain insight into the composition and climate of habitable-zone planets.

Atmospheric characterization of super-Earths is currently within reach of the Wide Field Camera 3 (WFC3) on board the Hubble Space Telescope (HST), combined with the recently implemented spatial scanning observational strategy[14]. The spectra of three hot transiting planets with radii less than 3.0 Earth radii ($R_\oplus$) have been published so far: Gliese 1214 b[15], HD 97658 b[16] and 55 Cancri e[17]. The first two do not show any evident transit depth modulation with wavelength, suggesting an atmosphere covered by thick clouds or made of molecular species heavier than hydrogen, while only the spectrum of 55 Cancri e has revealed a light-weight atmosphere, suggesting hydrogen–helium (H_2–He) still being present. In addition, transit observations of six temperate Earth-size planets around the ultra-cool dwarf TRAPPIST-1—planets b, c, d, e, f[6] and g[18]—have not shown any molecular signatures and have excluded the presence of cloud-free H_2–He atmospheres around them.

K2-18 b was discovered in 2015 by the Kepler spacecraft[7] and is orbiting around an M2.5 (metallicity [Fe/H] = 0.123 ± 0.157 dex (units of decimal exponent), effective temperature $T_{\mathrm{eff}} = 3{,}457 \pm 39$ K, stellar mass $M_\star = 0.359 \pm 0.047$ solar masses ($M_\odot$), stellar radius $R_\star = 0.411 \pm 0.038$ solar radii ($R_\odot$))[19] dwarf star, 34 pc away from the Earth. The star–planet distance of 0.1429 AU (ref. [19]) suggests a planet within the star's habitable zone (~0.12–0.25 AU) (ref. [20]), with effective temperature between 200 K and 320 K, depending on the albedo and the emissivity of its surface and/or its atmosphere. This crude estimate accounts for neither possible tidal energy sources[21] nor atmospheric heat redistribution[11,13], which might be relevant for this planet. Measurements of the mass and the radius of K2-18 b (planetary mass $M_{\mathrm{p}} = 7.96 \pm 1.91$ Earth masses ($M_\oplus$) (ref. [22]), planetary radius $R_{\mathrm{p}} = 2.279 \pm 0.0026\,R_\oplus$ (ref. [19])) yield a bulk density of 3.3 ± 1.2 g cm^{-1} (ref. [22]), suggesting either a silicate planet with an extended atmosphere or an interior composition with a water (H_2O) mass fraction lower than 50% (refs. [22–24]).

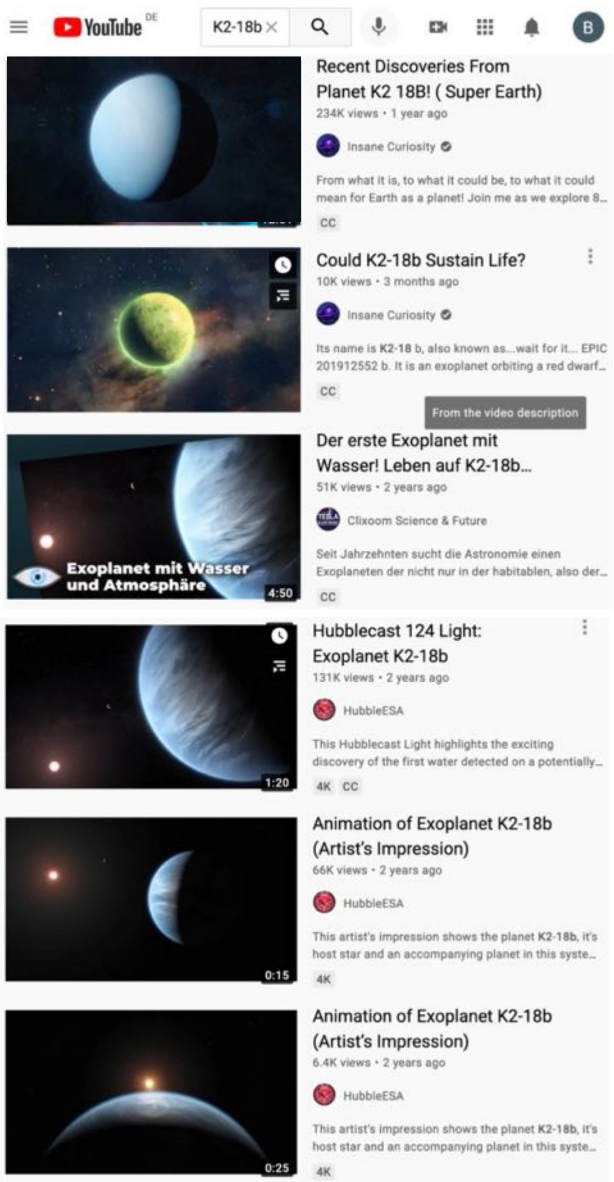
We analyse here eight transits of K2-18 b, obtained with the WFC3 camera on board the HST. We used our publicly available tools, specialized for HST/WFC3 data[5,9], to perform the end-to-end analysis from the raw data to the atmospheric parameters. The techniques used here have been validated by the analysis of the largest catalogue of exoplanetary spectra from WFC3[5]. Details can be found in Methods, and links to the data and the codes used can be found in 'Data availability' and 'Code availability', respectively. Along with the data, we provide descriptions of the data structures and instructions on how to reproduce the results presented here. Our analysis resulted in the detection of an atmosphere around K2-18 b with an Atmospheric Detectability Index[5] (ADI; a positively defined logarithmic Bayes factor) of 5.0, or approximately 3.6σ confidence[8,9], making K2-18 b the first habitable-zone planet in the super-Earth mass regime (1–10 $M_\oplus$) with an observed atmosphere around it.

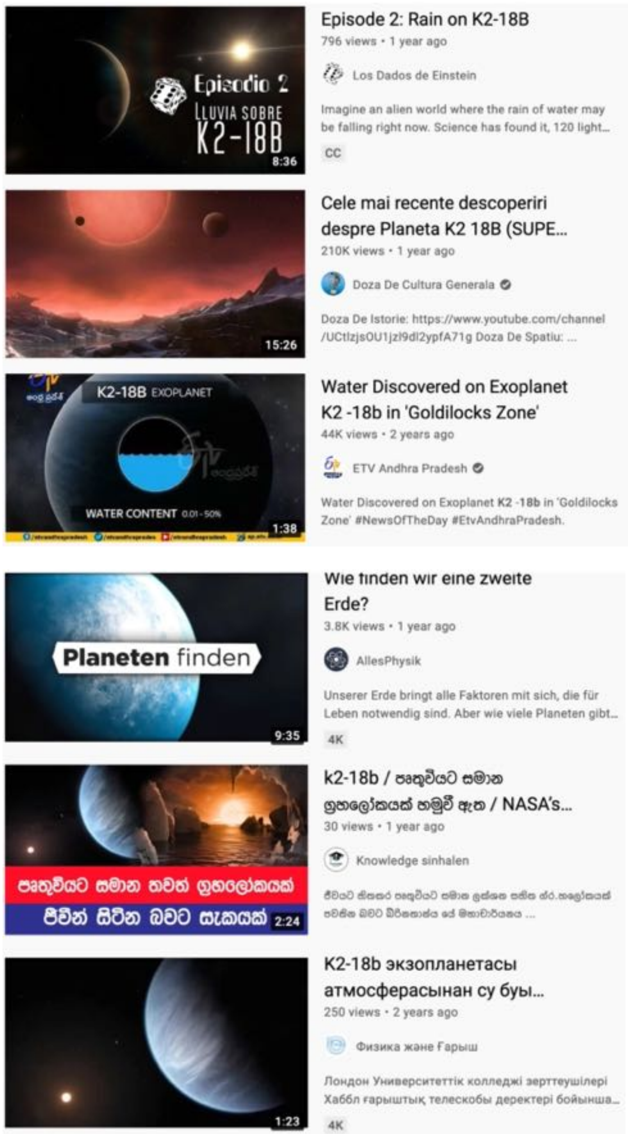
More specifically, nine transits of K2-18 b were observed as part of the HST proposals 13665 and 14682 (principal investigator: Björn Benneke), and the data are available through the Mikulski Archive for Space Telescopes (MAST; see 'Data availability'). Each transit was observed during five HST orbits, with the G141 infrared grism (1.1–1.7 μm), and each exposure was the result of 16 up-the-ramp samples in the spatial scanning mode. The ninth transit observation suffered from pointing instabilities, and we therefore decided not to include it in this analysis. We extracted the white and the spectral light curves from the reduced images, following our dedicated methodology[5,17,25], which has been integrated into an automated, self-consistent and user-friendly Python package named Iraclis (see 'Code availability'). No systematic variations of the white light curve, $R_{\mathrm{p}}/R_\star$, appeared between the eight different observations. This level of stability among the extracted broadband transit depths is not always guaranteed, as consistency problems among different observations emerged in previous analyses[5,16].

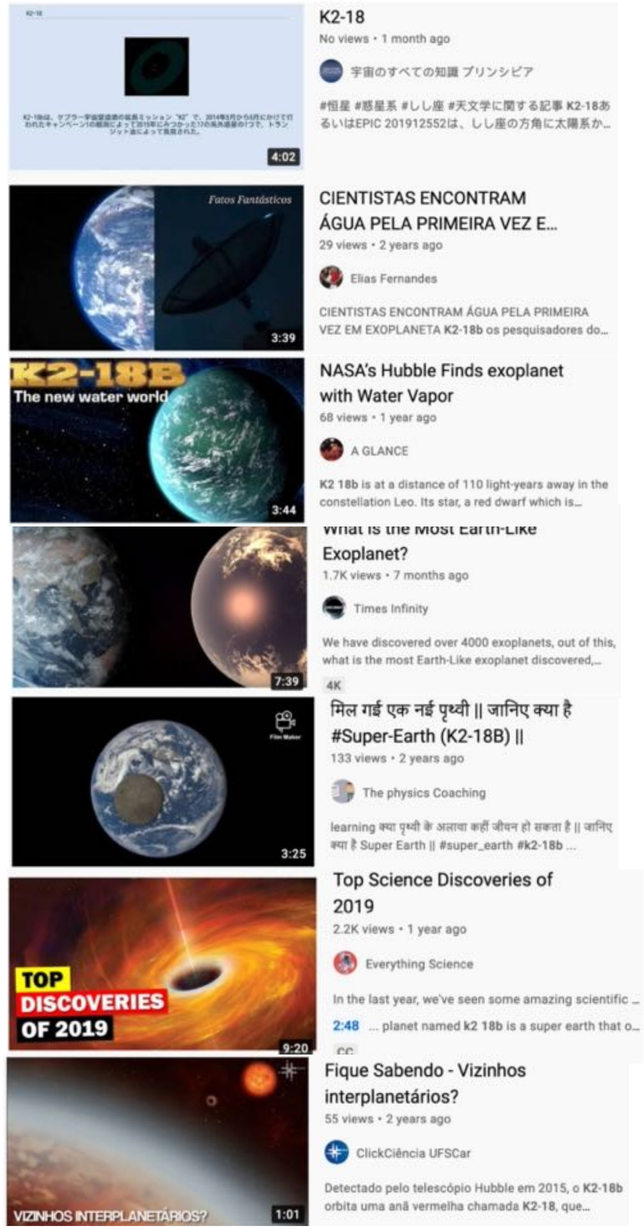
In our analysis, we found that the measured mid-transit times were not consistent with the expected ephemeris[19]. We used these results to refine the ephemeris of K2-18 b to be $P = 32.94007 \pm 0.00003$ days and $T_0 = 2457363.2109 \pm 0.0004\,\mathrm{BJD_{TDB}}$ (ref. [26]), where P

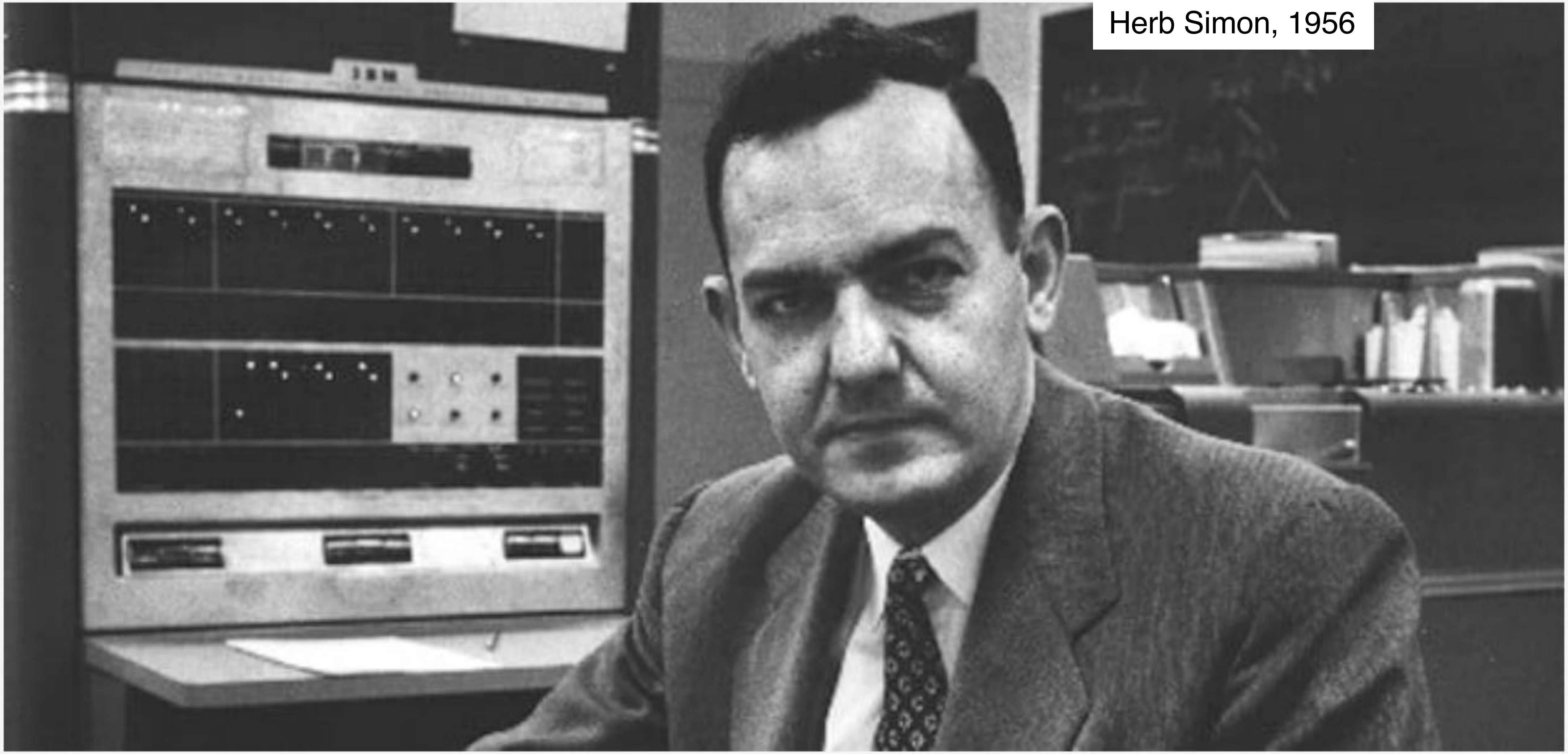
Department of Physics & Astronomy, University College London, London, UK. *e-mail: atsiaras@star.ucl.ac.uk; ingo@star.ucl.ac.uk

http://people.tue.mpg.de/bs/K2-18b.html


YouTube DE
K2-18b
Recent Discoveries From Planet K2 18B! (Super Earth)
234K views • 1 year ago
Insane Curiosity
From what it is, to what it could be, to what it could mean for Earth as a planet! Join me as we explore 8...
CC
Could K2-18b Sustain Life?
10K views • 3 months ago
Insane Curiosity
Its name is K2-18 b, also known as...wait for it... EPIC 201912552 b. It is an exoplanet orbiting a red dwarf...
CC
From the video description
Der erste Exoplanet mit Wasser! Leben auf K2-18b...
51K views • 2 years ago
Clixoom Science & Future
Seit Jahrzehnten sucht die Astronomie einen Exoplaneten der nicht nur in der habitablen, also der...
CC
Exoplanet mit Wasser und Atmosphäre
4:50
Hubblecast 124 Light: Exoplanet K2-18b
131K views • 2 years ago
HubbleESA
This Hubblecast Light highlights the exciting discovery of the first water detected on a potentially...
4K CC
1:20
Animation of Exoplanet K2-18b (Artist's Impression)
66K views • 2 years ago
HubbleESA
This artist's impression shows the planet K2-18b, it's host star and an accompanying planet in this syste...
4K
0:15
Animation of Exoplanet K2-18b (Artist's Impression)
6.4K views • 2 years ago
HubbleESA
This artist's impression shows the planet K2-18b, it's host star and an accompanying planet in this syste...
4K
0:25


Episode 2: Rain on K2-18B
796 views • 1 year ago
Los Dados de Einstein
Imagine an alien world where the rain of water may be falling right now. Science has found it, 120 light...
CC
Episodio 2
LLUVIA SOBRE K2-18B
8:36
Cele mai recente descoperiri despre Planeta K2 18B (SUPE...
210K views • 1 year ago
Doza De Cultura Generala
Doza De Istorie: https://www.youtube.com/channel /UCtlzjsOU1jzl9dl2ypfA71g Doza De Spatiu: ...
15:26
Water Discovered on Exoplanet K2 -18b in 'Goldilocks Zone'
44K views • 2 years ago
ETV Andhra Pradesh
Water Discovered on Exoplanet K2 -18b in 'Goldilocks Zone' #NewsOfTheDay #EtvAndhraPradesh.
K2-18B EXOPLANET
WATER CONTENT 0.01-50%
1:38
Wie finden wir eine zweite Erde?
3.8K views • 1 year ago
AllesPhysik
Unserer Erde bringt alle Faktoren mit sich, die für Leben notwendig sind. Aber wie viele Planeten gibt...
4K
Planeten finden
9:35
k2-18b / පෘතුවියට සමාන ග්‍රහලෝකයක් හමුවී ඇත / NASA's...
30 views • 1 year ago
Knowledge sinhalen
2:24
K2-18b экзопланетасы атмосферасынан су буы...
250 views • 2 years ago
Физика және Ғарыш
Лондон Университеттік колледжі зерттеушілері Хаббл ғарыштық телескобы деректері бойынша...
4K
1:23


K2-18
No views • 1 month ago
宇宙のすべての知識 プリンシピア
4:02
CIENTISTAS ENCONTRAM ÁGUA PELA PRIMEIRA VEZ E...
29 views • 2 years ago
Elias Fernandes
CIENTISTAS ENCONTRAM ÁGUA PELA PRIMEIRA VEZ EM EXOPLANETA K2-18b os pesquisadores do...
Fatos Fantásticos
3:39
NASA's Hubble Finds exoplanet with Water Vapor
68 views • 1 year ago
A GLANCE
K2 18b is at a distance of 110 light-years away in the constellation Leo. Its star, a red dwarf which is...
K2-18B
The new water world
3:44
What is the Most Earth-Like Exoplanet?
1.7K views • 7 months ago
Times Infinity
We have discovered over 4000 exoplanets, out of this, what is the most Earth-Like exoplanet discovered,...
4K
7:39
मिल गई एक नई पृथ्वी || जानिए क्या है #Super-Earth (K2-18B) ||
133 views • 2 years ago
The physics Coaching
3:25
Top Science Discoveries of 2019
2.2K views • 1 year ago
Everything Science
In the last year, we've seen some amazing scientific ...
2:48 ... planet named k2 18b is a super earth that o...
TOP DISCOVERIES OF 2019
9:20
Fique Sabendo - Vizinhos interplanetários?
55 views • 2 years ago
ClickCiência UFSCar
Detectado pelo telescópio Hubble em 2015, o K2-18b orbita uma anã vermelha chamada K2-18, que...
VIZINHOS INTERPLANETÁRIOS?
1:01



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:

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

- **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 $p\left(X_i \mid \mathrm{PA}_i\right)$.

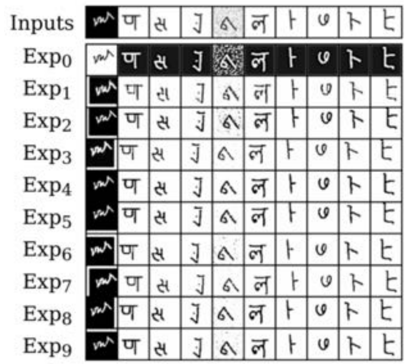
> **Sparse Mechanism Shift Hypothesis**: small distribution changes manifest themselves sparsely in the disentangled factorization, i.e., they should usually not affect all factors simultaneously.

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)*

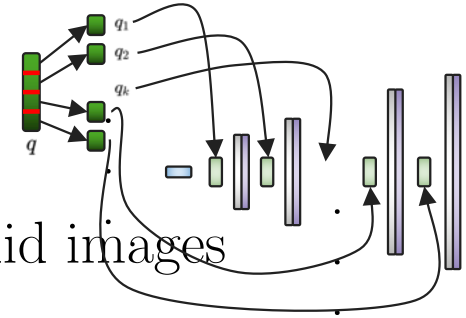
https://arxiv.org/abs/1911.10500

Causal training

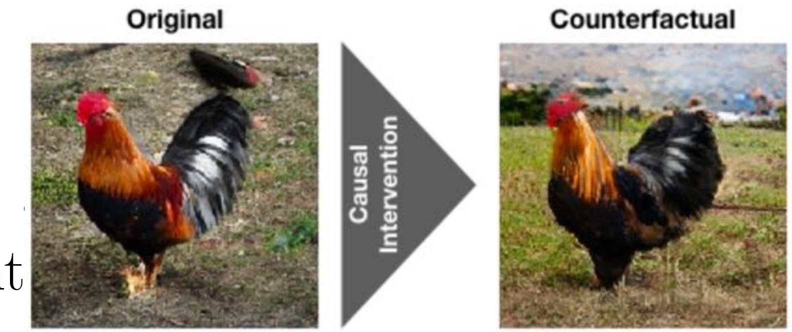


ICM training: encourage independence of mechanisms

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



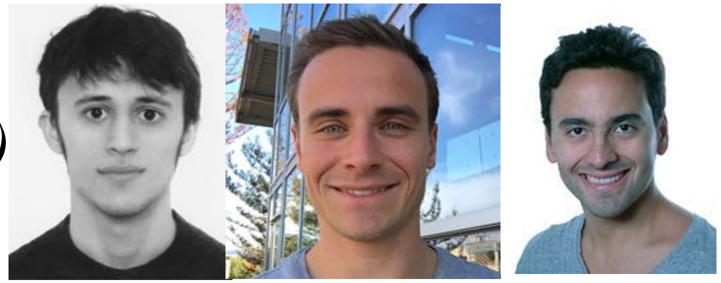
Counterfactual training: require that interventions produce valid images (e.g., after reconstruction in an autoencoder).



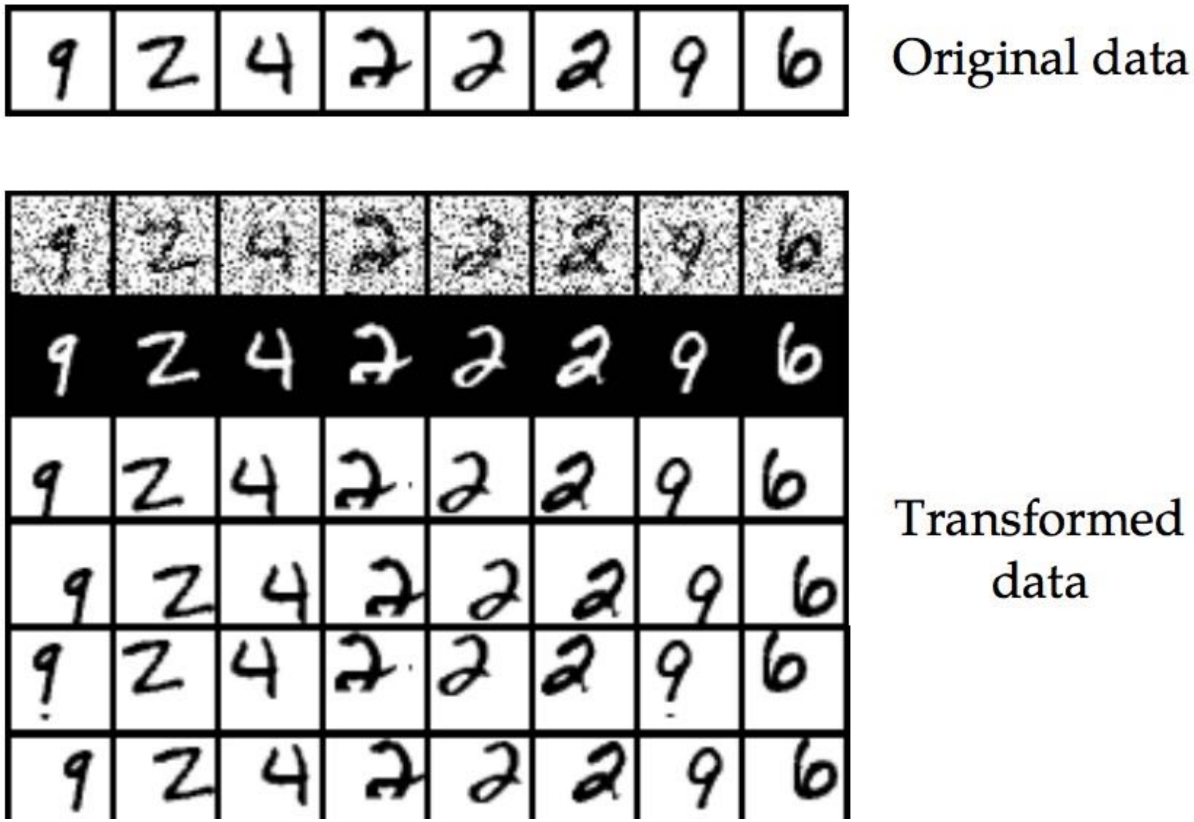
Sparse mechanism shift training: require that tions/interventions, only a sparse set of causal represent

Learning independent mechanisms

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

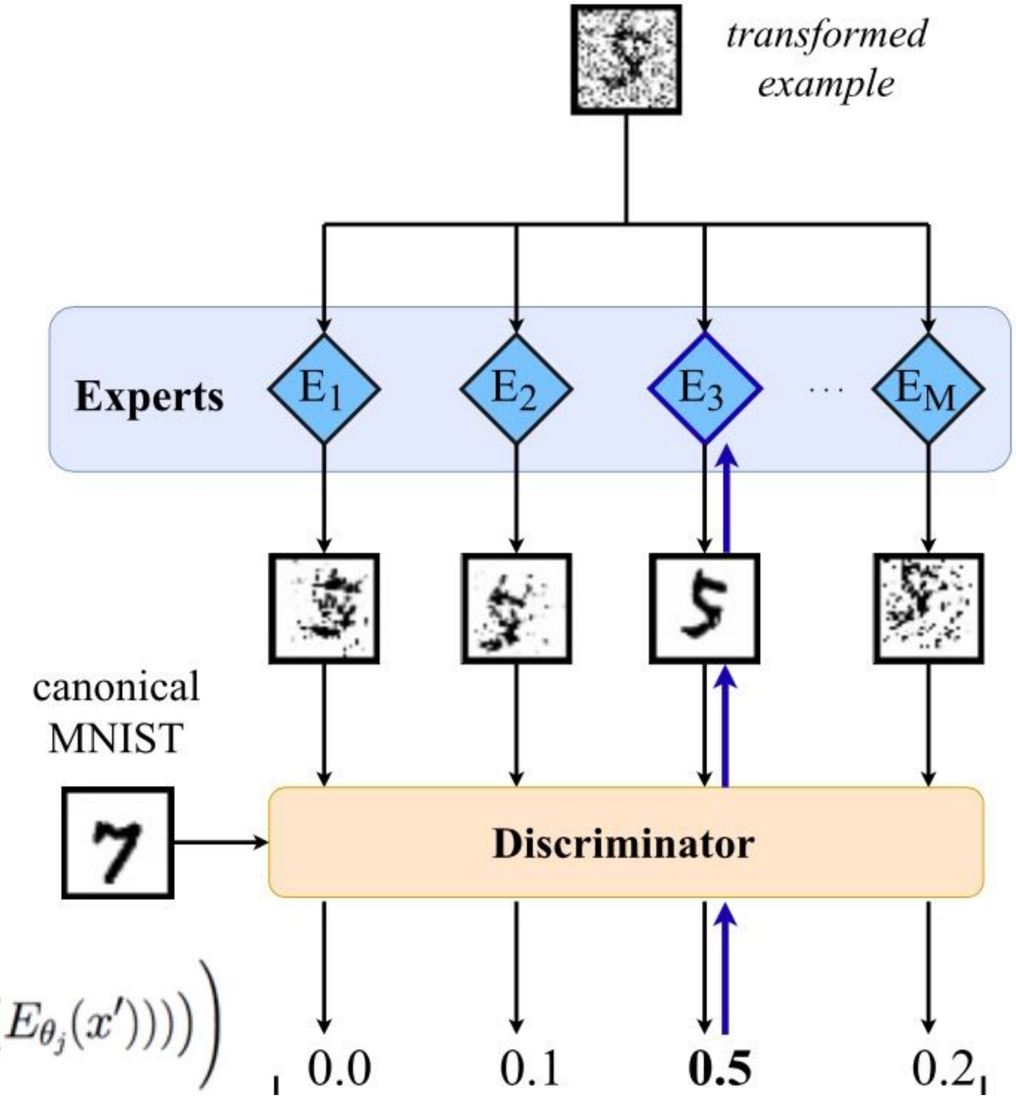


- Data drawn from $p(x)$, transformed by M mechanisms $f_1, \ldots, f_M$
- Goal: learn the independent mechanisms / factors of variation
- Method: generative model with competing mechanisms



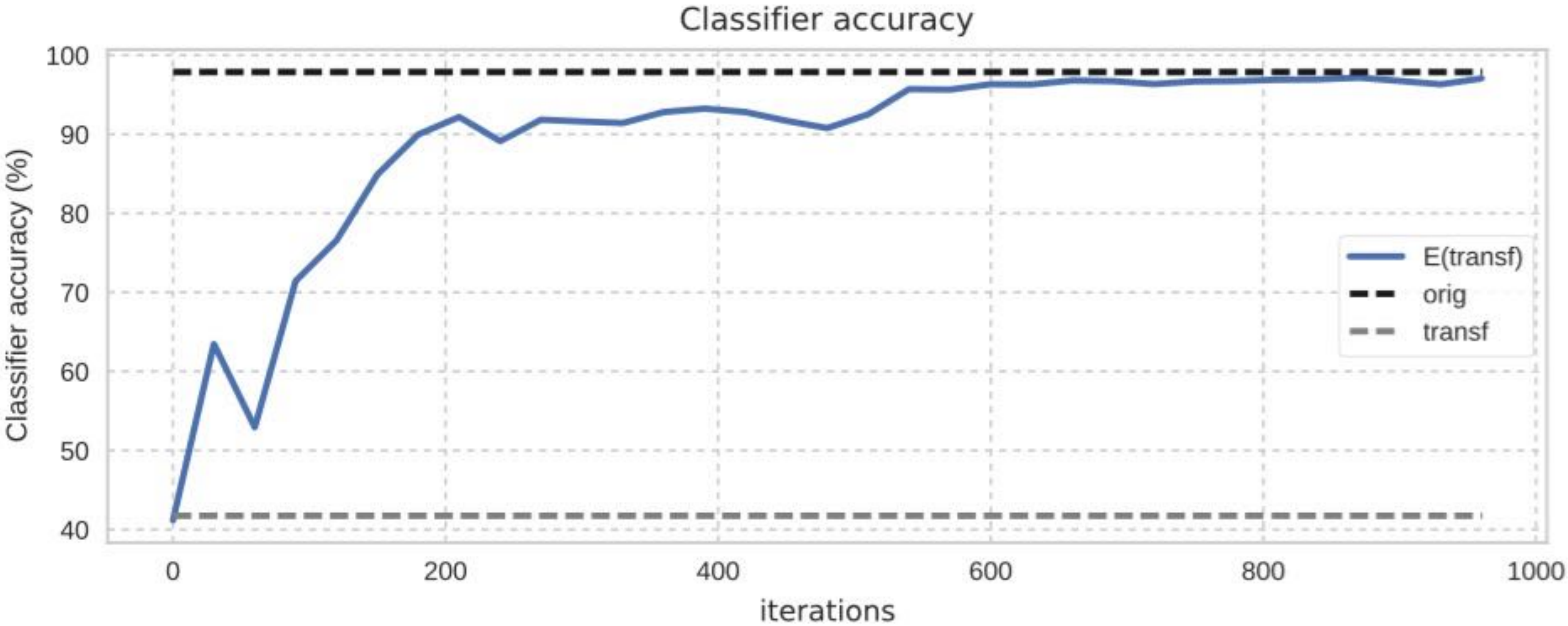
Method

- 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

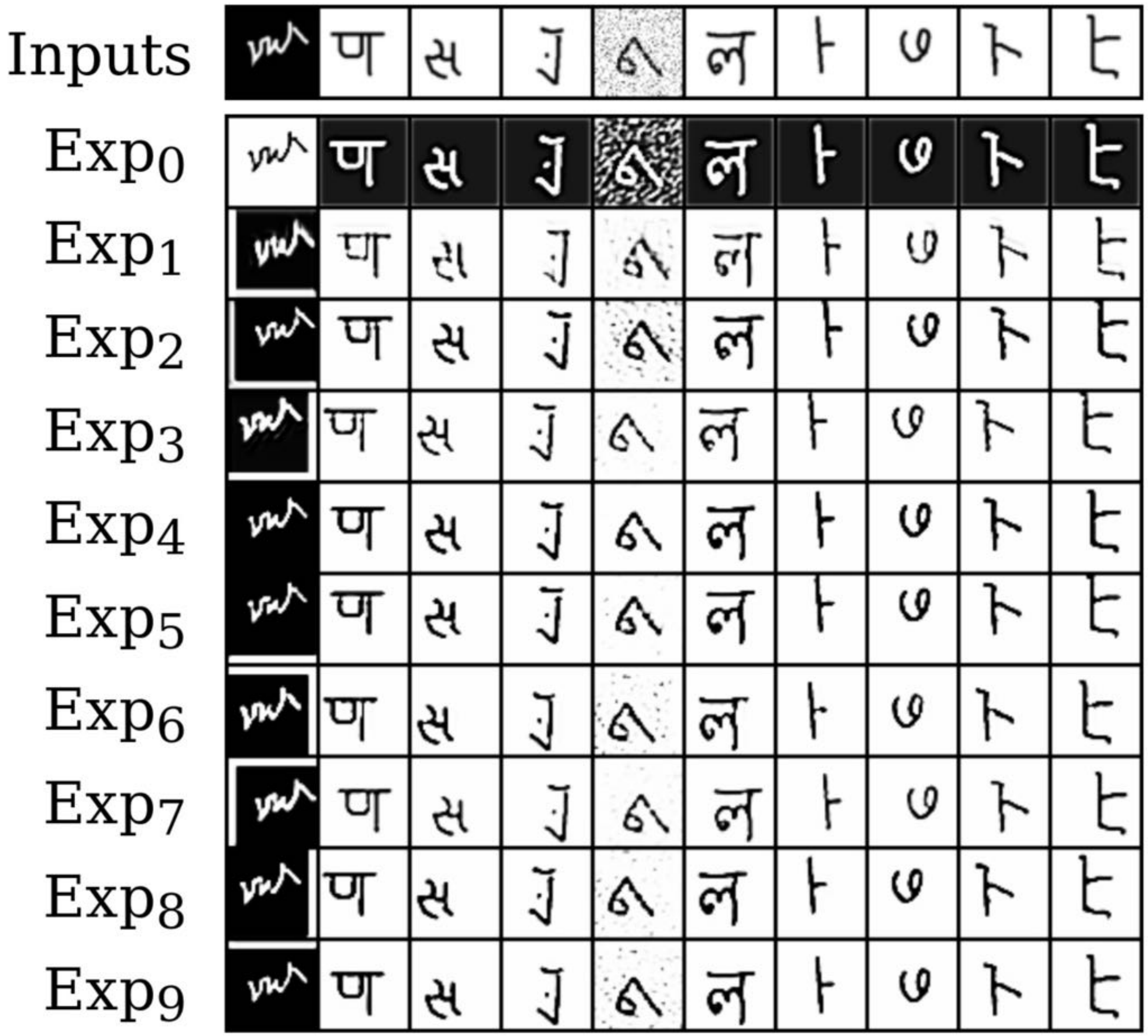


$$\max_{\theta_D} \left(\mathbb{E}_{x\sim P} \log(D_{\theta_D}(x)) + \frac{1}{N'} \sum_{j=1}^{N'} \mathbb{E}_{x'\sim Q} \left(\log(1 - D_{\theta_D}(E_{\theta_j}(x'))) \right) \right)$$

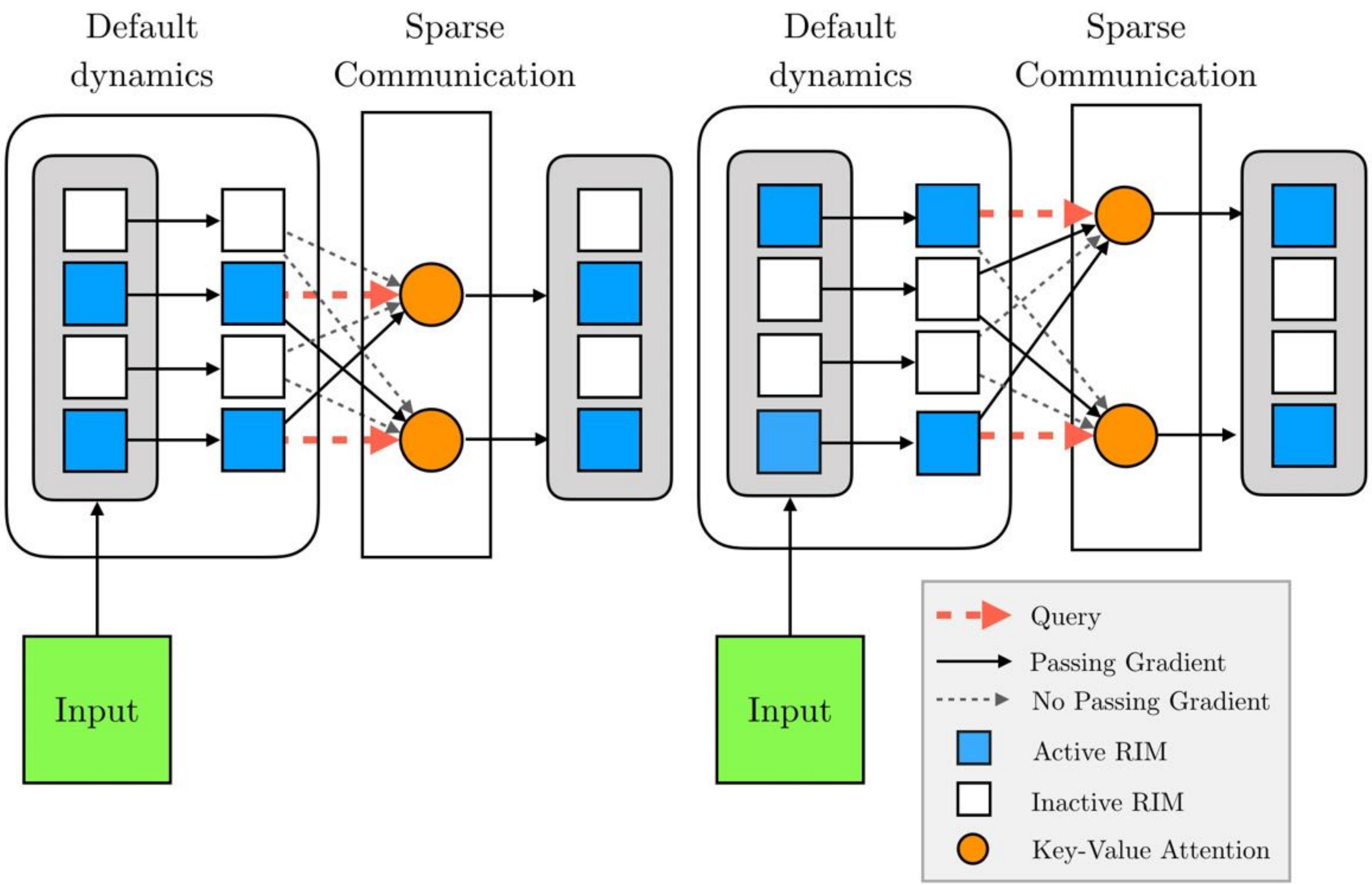
Accuracy of a CNN trained on MNIST for different test sets



Generalizing to Omniglot characters



Recurrent Independent Mechanisms



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

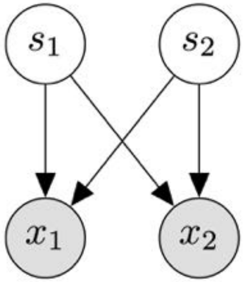
ICLR 2021

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

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)

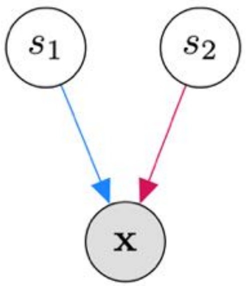
Problem: impossible in general [Hyvärinen & Pajunen, '99]

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

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

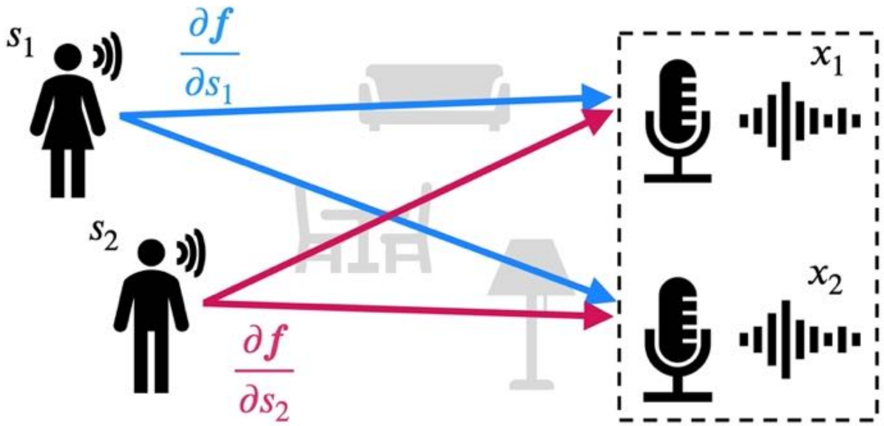
ICM usually applied to cause distribution p_c and mechanism $p_{e|c}$ (or $\boldsymbol{f}$), e.g., cause-effect discovery

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

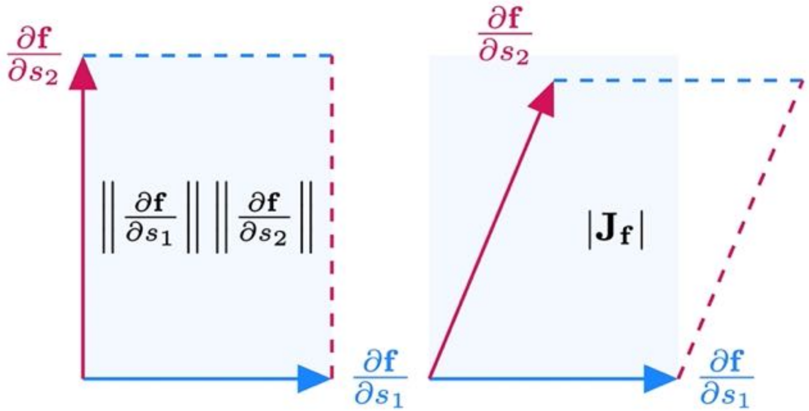


Independent mechanism analysis

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:

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

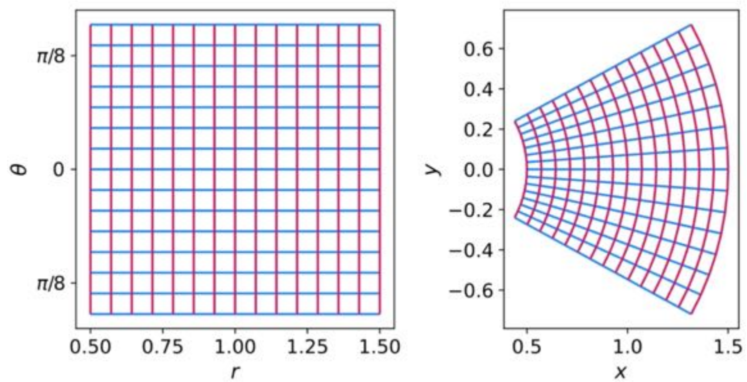


<u>Geometric interpretation:</u> corresponds to an *orthogonality condition* on the columns of the Jacobian.

<u>Contrast function:</u>

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

- ≥ 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

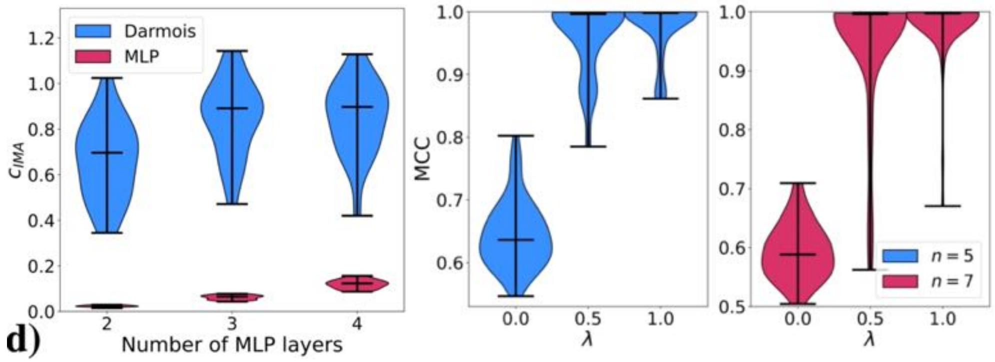
Can rule out (in the sense that C_{IMA} is larger for) well-known spurious ICA solutions:

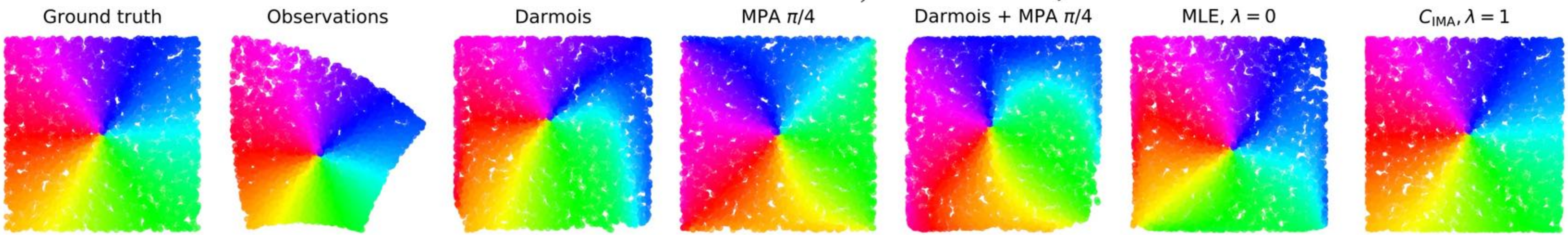
- Darmois (inverse CDF) construction
- Measure-preserving automorphisms (MPA)

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



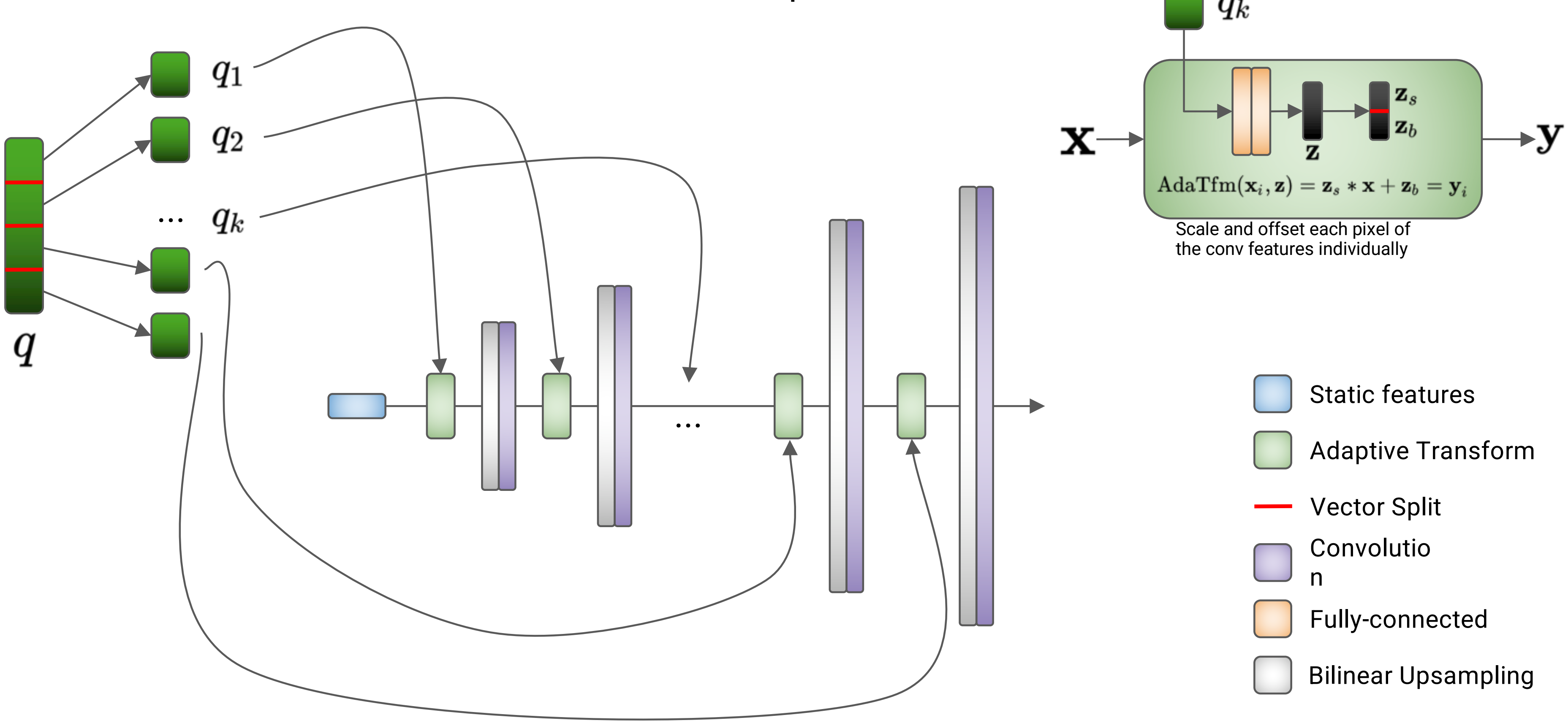


Structural Decoders

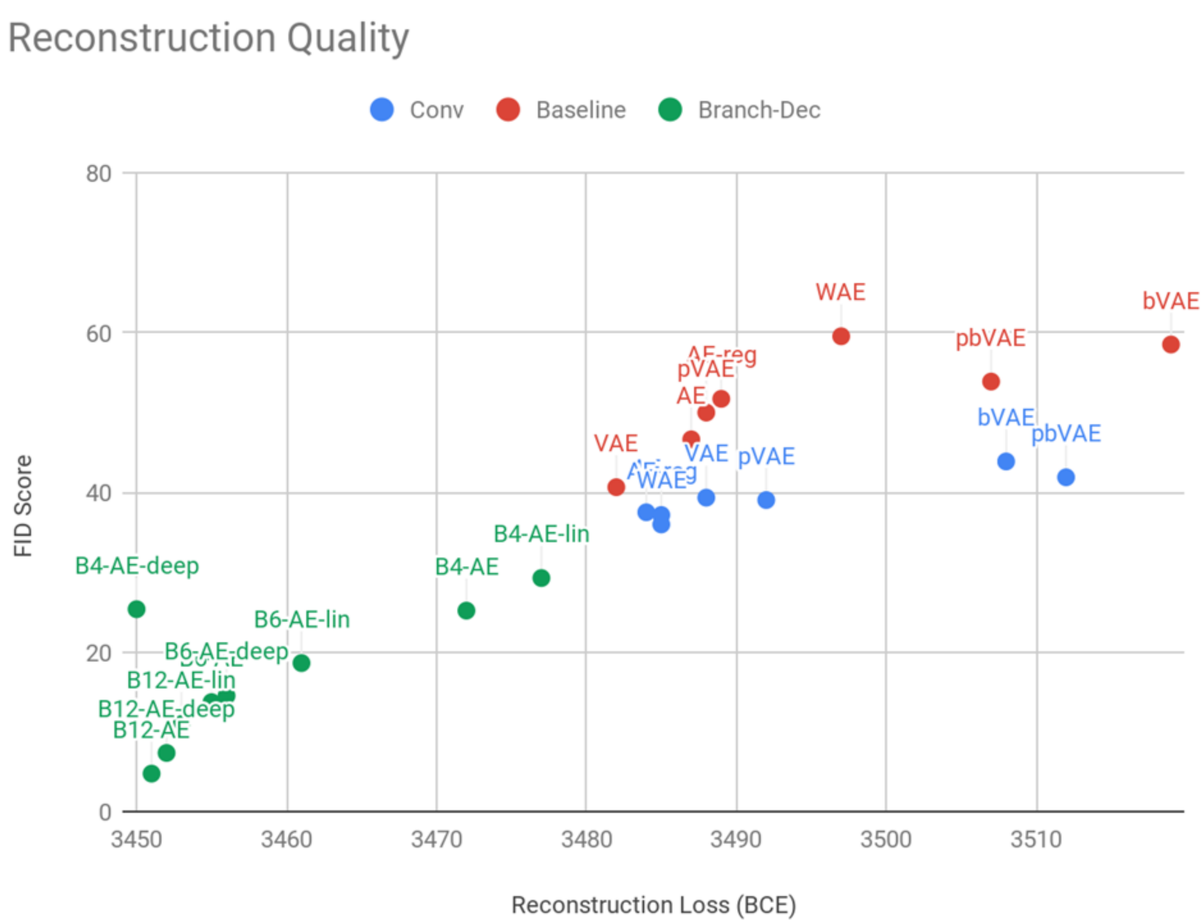
Leeb et al.

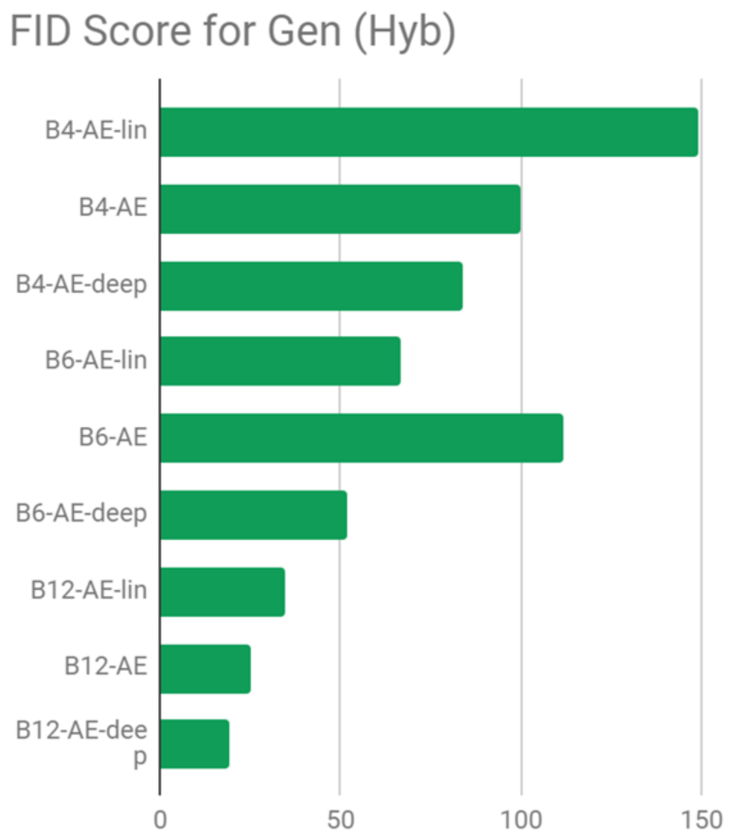
arXiv 2006.07796

➔ ~200-700k parameters

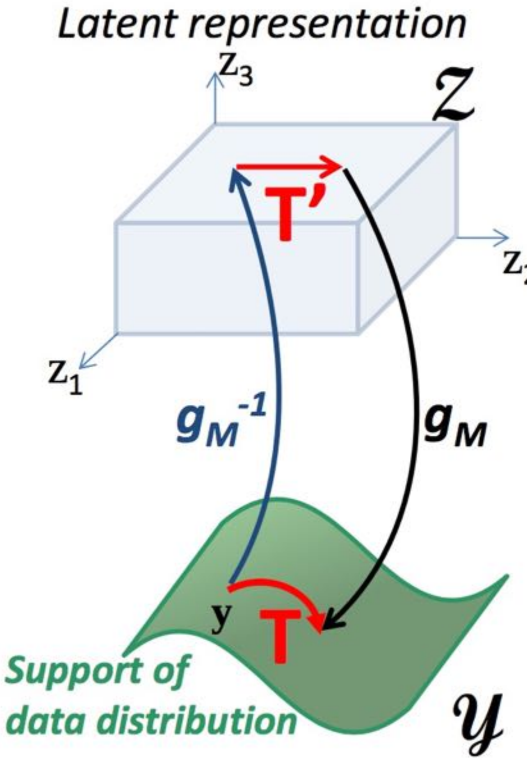


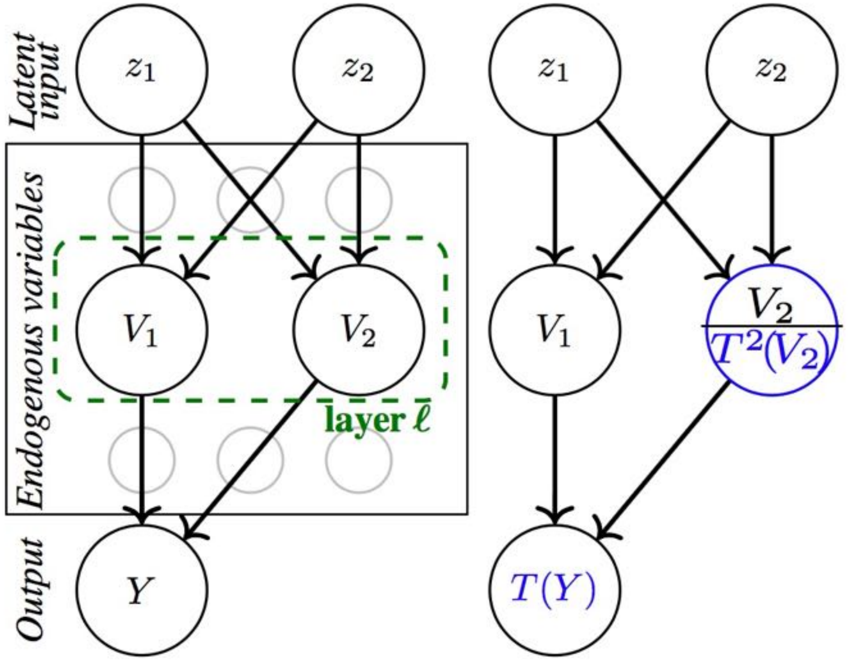
Quantitative Results

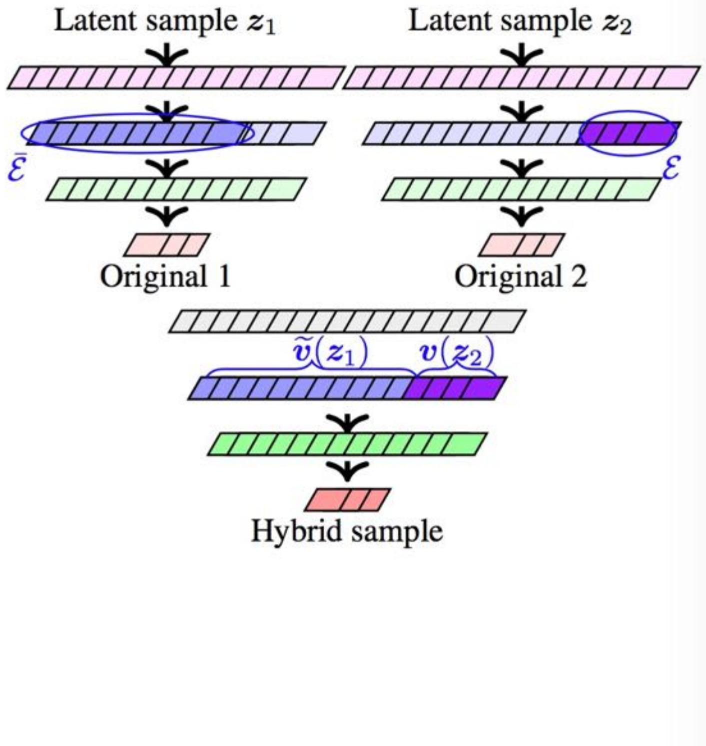




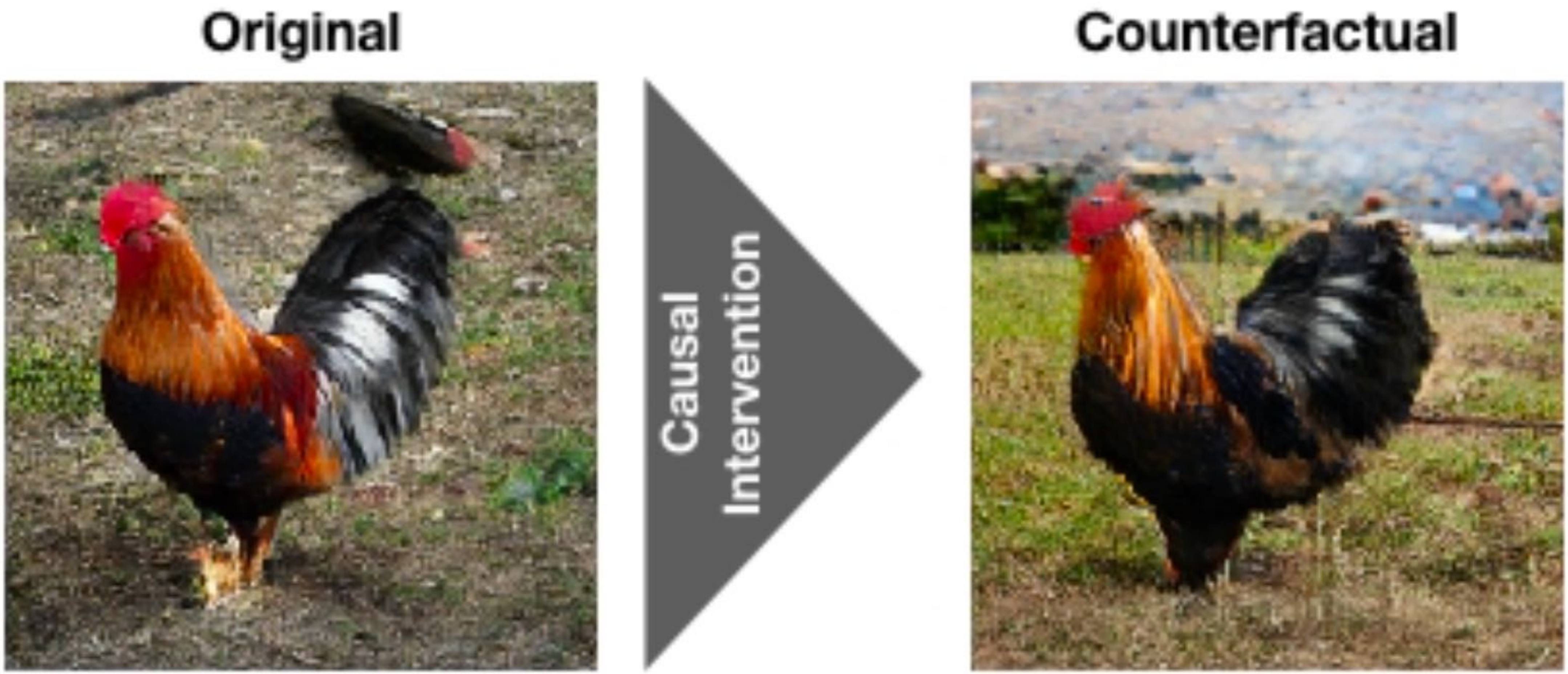
Interventional Representations *(Besserve et al., ICLR 2020)*







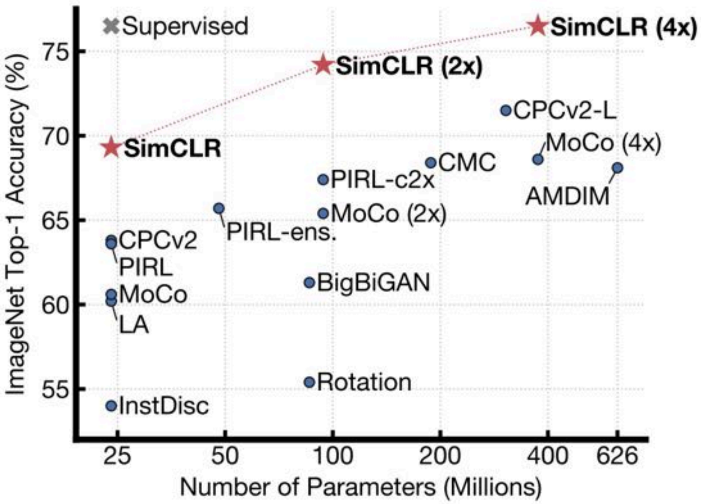
Interventional Representations *(Besserve et al., ICLR 2020)*



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

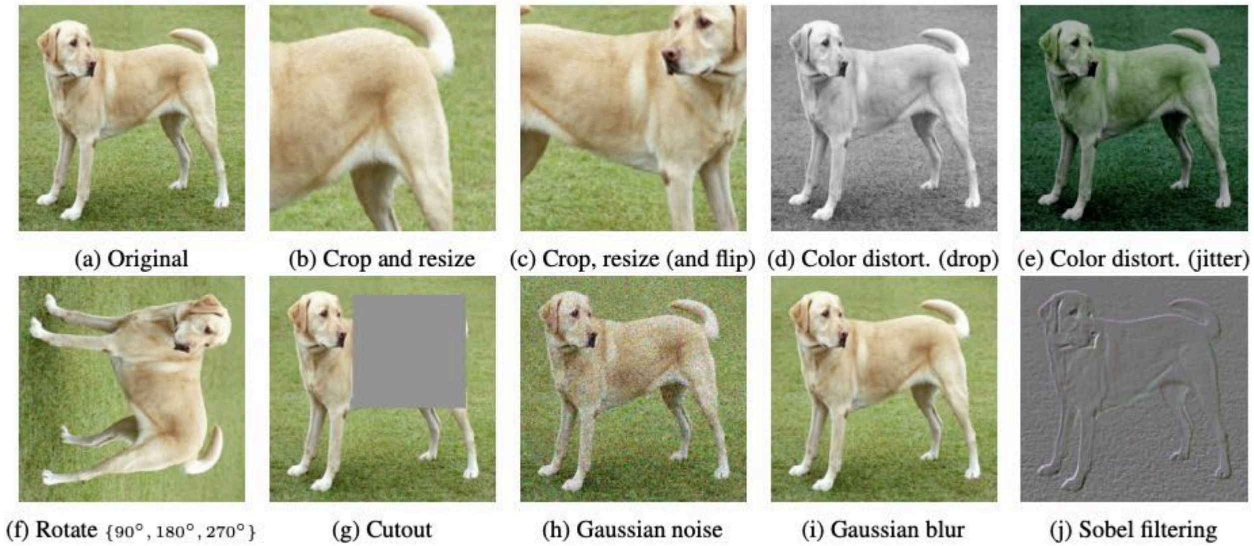
Self-supervised learning provably isolates content from style
(https://arxiv.org/abs/2106.04619)

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



Self-supervised learning using contrastive training learn a representation which is insensitive to augmentation but sensitive to changing the example (NCE).

Can think of both as interventions.



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 provably isolates content from style

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

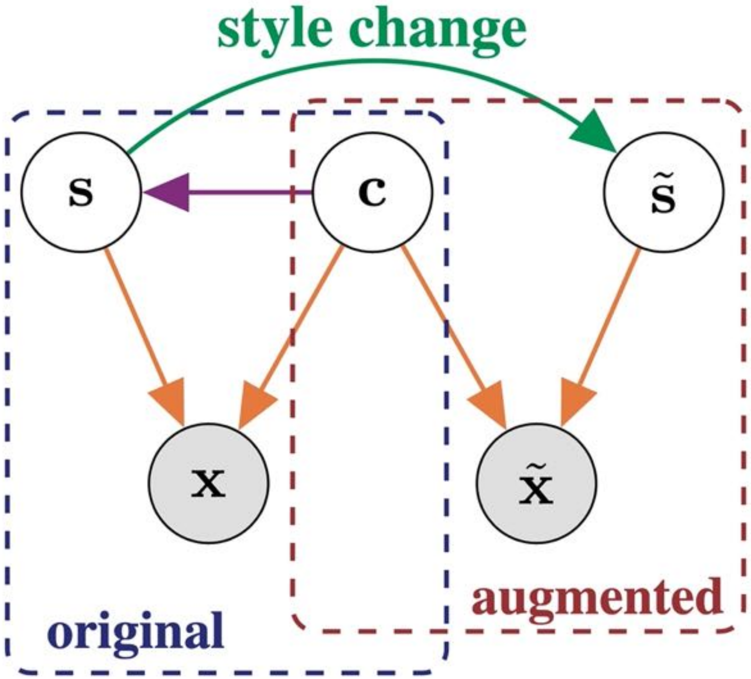
Formalise generation $\boldsymbol{x} = f(\boldsymbol{z})$ and augmentation $\tilde{\boldsymbol{x}} = f(\tilde{\boldsymbol{z}})$ processes as latent variable model with unknown content-style partition $\boldsymbol{z} = (\boldsymbol{c}, \boldsymbol{s})$, interpreting style change as an intervention.

- *invariant content $\boldsymbol{c}$:* 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).

Given data $(\boldsymbol{x}, \widetilde{\boldsymbol{x}})$ (nonlinear mixtures of content and style):

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



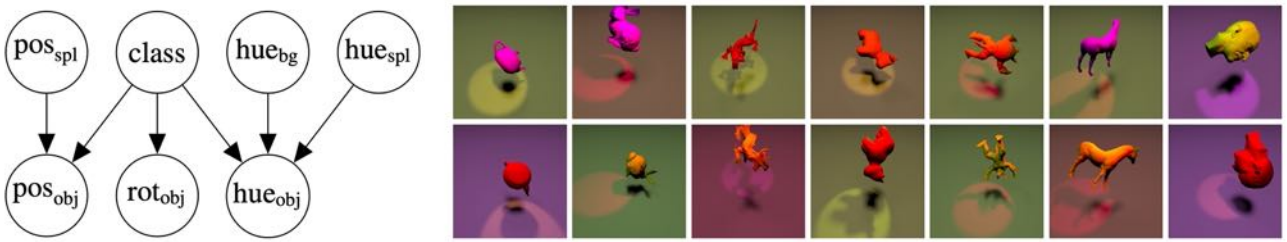


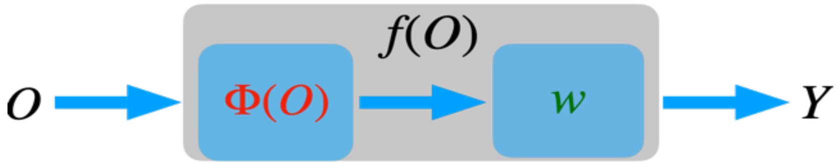
Figure 2: *(Left)* Causal graph for the *Causal3DIdent* dataset. *(Right)* Two samples from each object class.

*up to invertible transformation

Nonlinear Invariant Risk Minimization

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

Problem

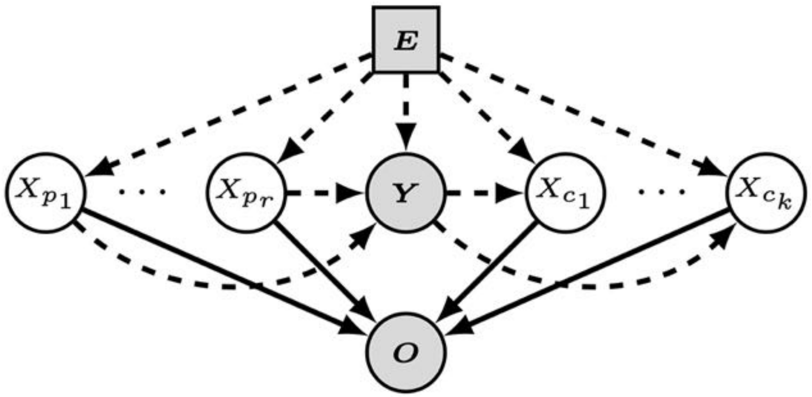


Nonlinear
Data Representation

Nonlinear
Invariant Classifier

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

Assumption on Causal Graphs



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

Assumption on the Prior

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

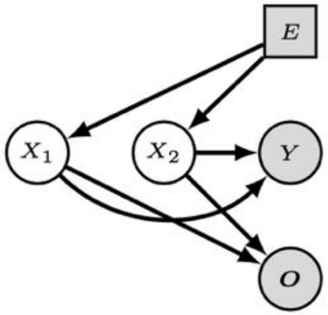
$$\Downarrow$$

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

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

Experimental Results

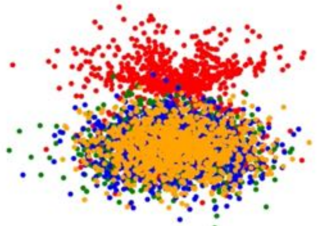
(with Chaochao Lu, Yuhuai Wu, José Miguel Hernández-Lobato, arXiv:2102.12353)



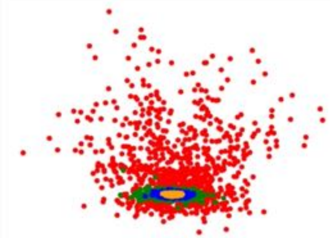
$$E \sim \mathcal{U}\{0.2, 2, 3, 5\}$$
$$X_1 \sim \mathcal{N}(X_1|E, 1)$$
$$X_2 \sim \mathcal{N}(X_2|2E, 2)$$
$$Y \sim \mathcal{N}(Y|X_1 + X_2, 1)$$
$$\boldsymbol{O} = g(X_1, X_2)$$

Data Generating Process

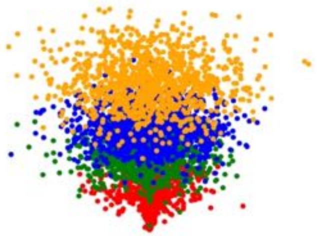
Original Data



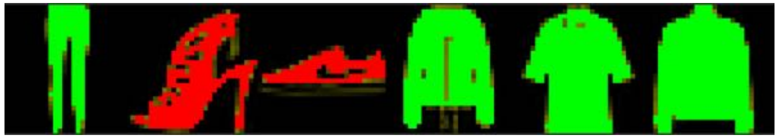
Samples from VAE
(Kingma et al. 2013)



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).

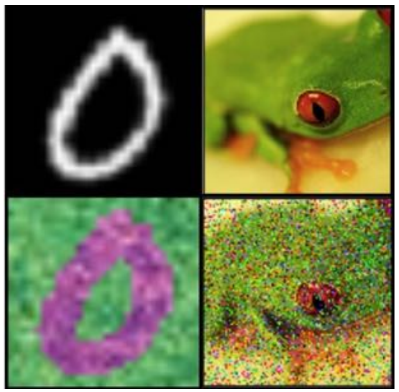
METHOD	TRAIN	TEST
ERM	83.17 ± 1.01	22.46 ± 0.68
ERM 1	81.33 ± 1.35	33.34 ± 8.85
ERM 2	84.39 ± 1.89	13.16 ± 0.82
ROBUST MIN MAX	82.81 ± 0.11	29.22 ± 8.56
F-IRM GAME	62.31 ± 2.35	69.25 ± 5.82
V-IRM GAME	68.96 ± 0.95	70.19 ± 1.47
IRM	75.01 ± 0.25	55.25 ± 12.42
iCaRL (ours)	$\mathbf{74.87 \pm 0.36}$	$\mathbf{73.56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

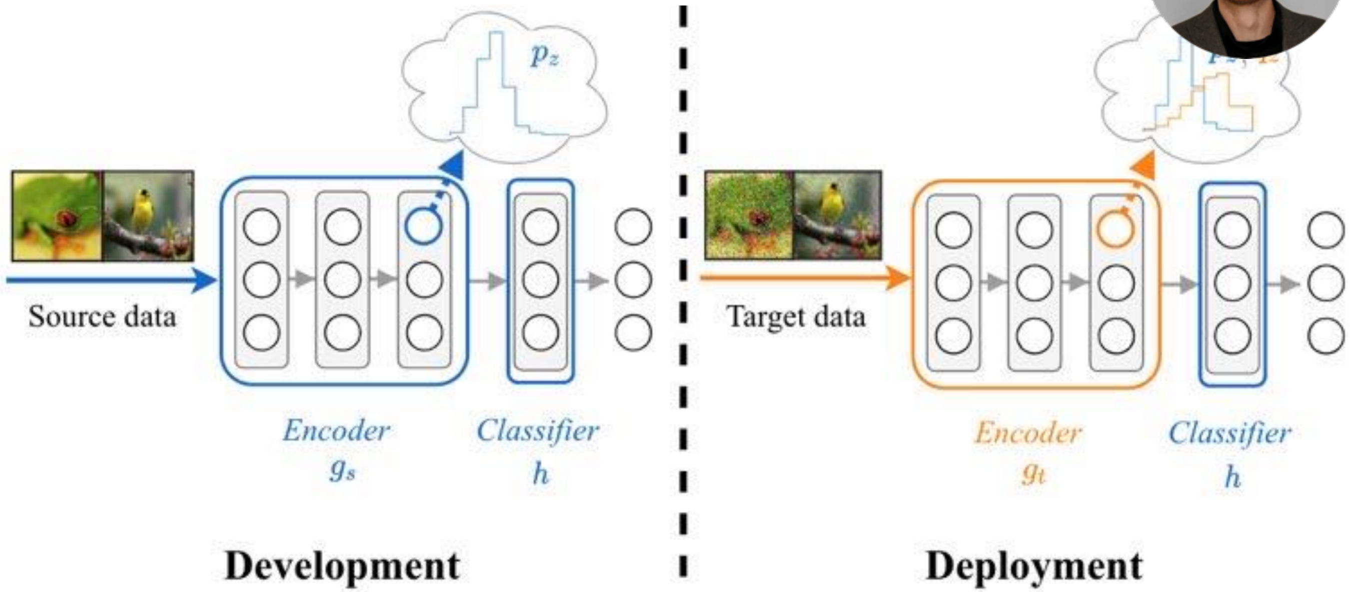
Source-Free Adaptation to Measurement Shift via Bottom-Up Feature Restoration (Cian Eastwood et al., https://arxiv.org/abs/2107.05446)

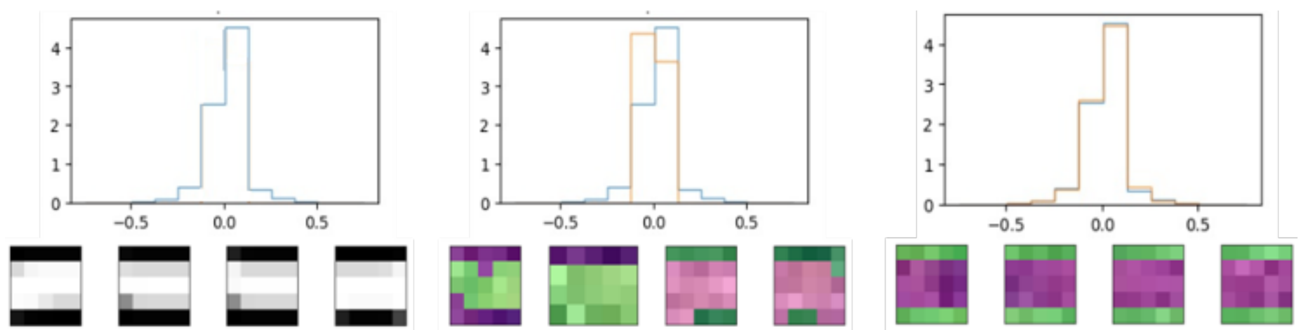
<u>*Source-free domain adaptation*</u>
-*Development:* train + equip model
-*Deployment:* adapt, no source data

<u>*Measurement shift (cf. Storkey, 2009)*</u>
-New sensor, same underlying features

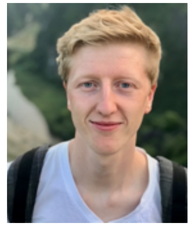
<u>*Feature restoration*</u>
-**Goal:** extract same features, new env.
-**Method:** align (marginal) feature dists.



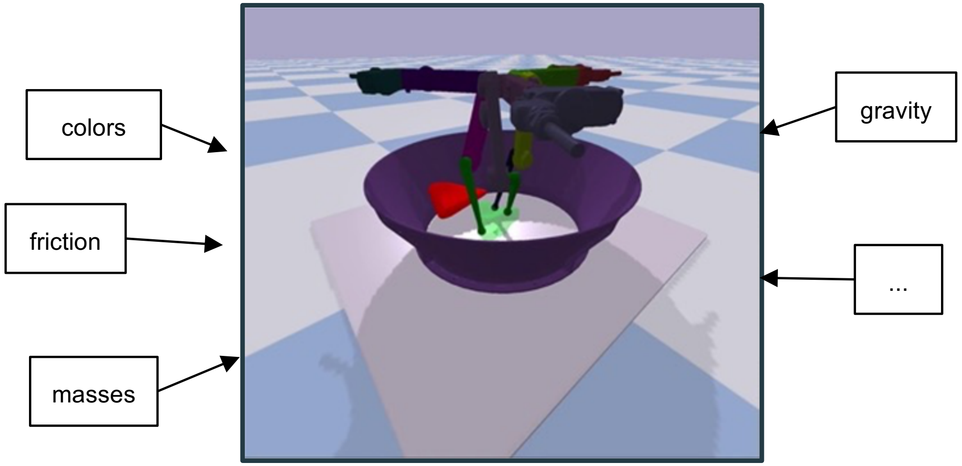


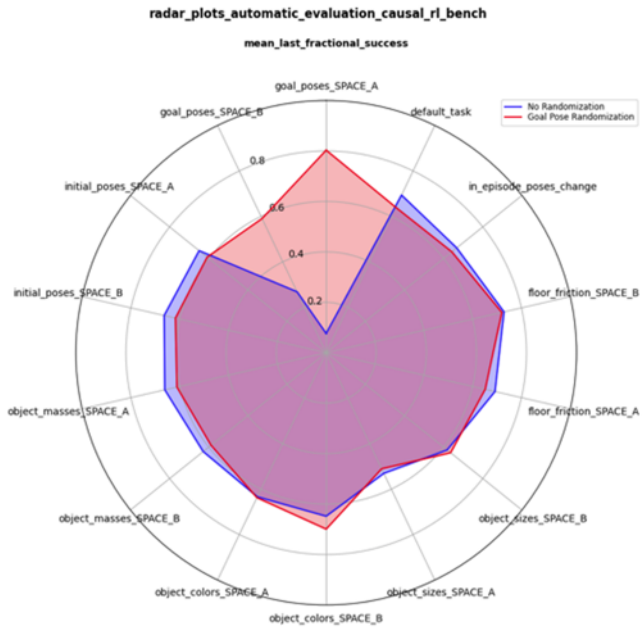


CausalWorld: A Robotic Manipulation Benchmark for Causal Structure and Transfer Learning



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

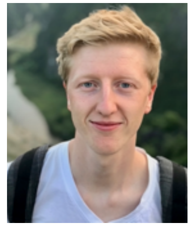




Evaluate different generalization aspects by intervening on a large range of different defining variables of the hierarchical causal generative world model of the robotic environment.

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

On the Transfer of Disentangled Representations in Realistic Settings

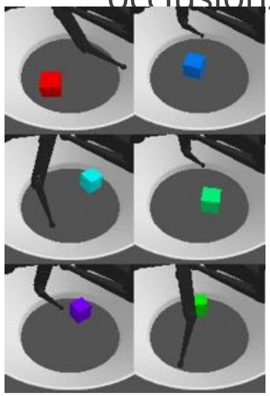



Dittadi and Träuble et al.,

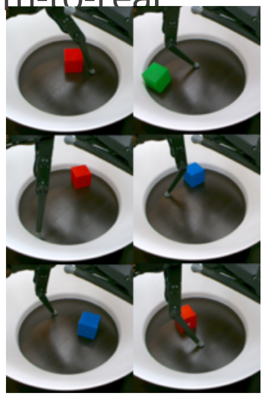
arXiv: 2010.14407,
ICLR 2021

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

OOD 1 color

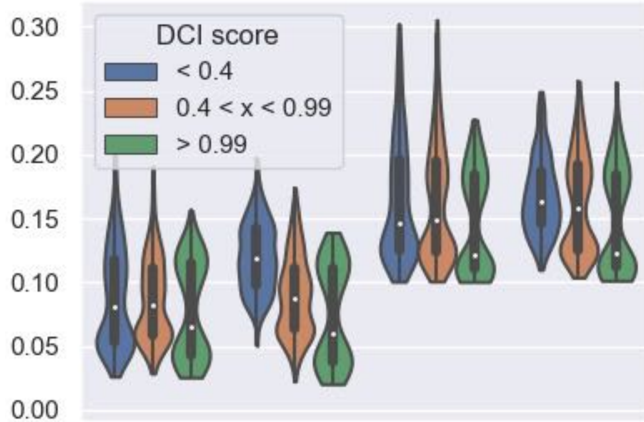
OOD 2 color

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

Generalization error (**OOD2**)



MLP GBT MLP GBT

simulation real world

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.

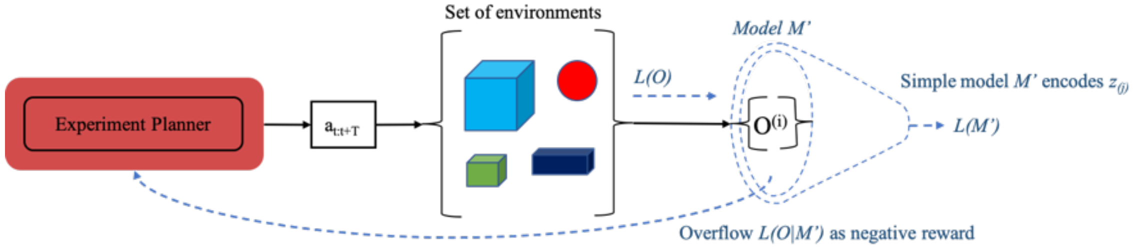


Fig 1: Experiment Discovery

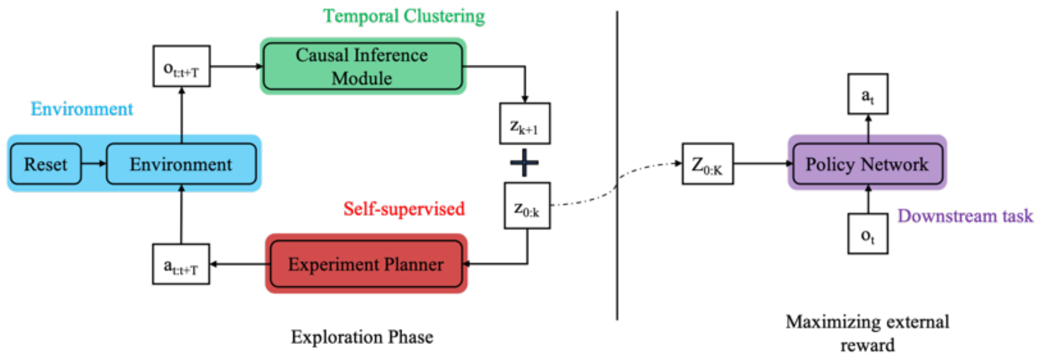
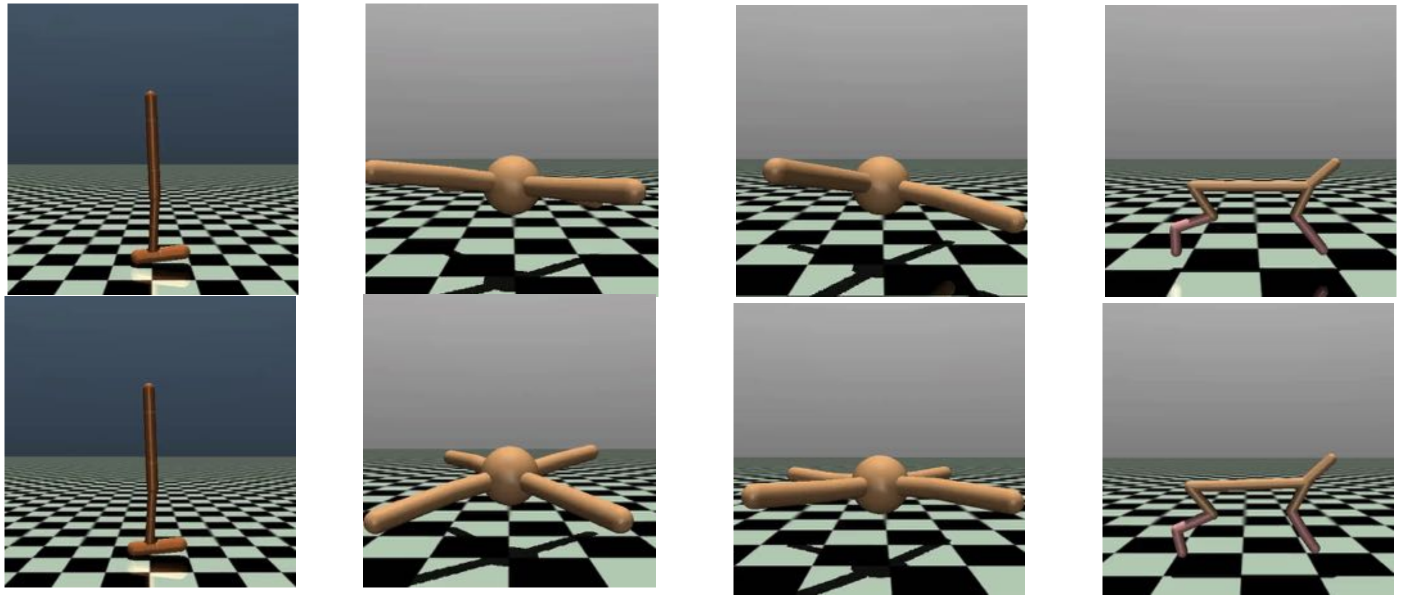


Fig 2: Performing experiments sequentially to learn causal representations. Representations used for downstream transfer.


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

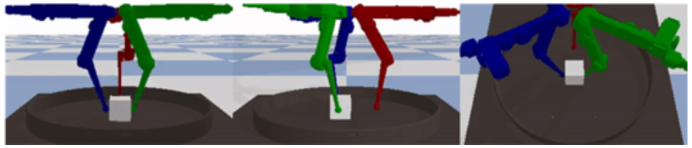


Heavy

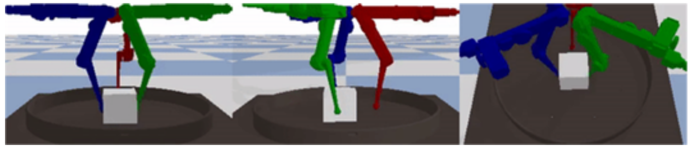
Light

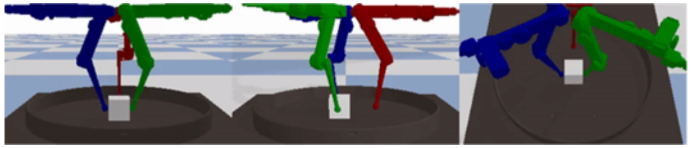
Hop Pirouette Leap Pushup

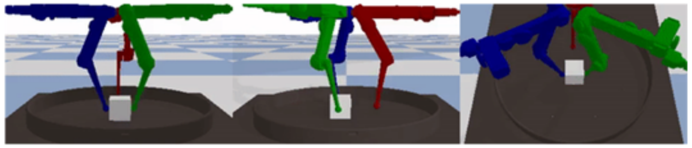
Discovered Behaviors - CausalWorld

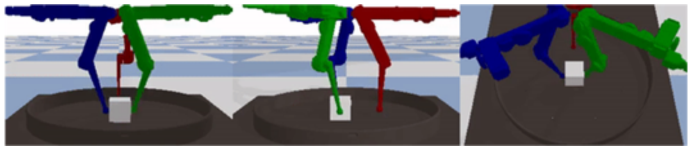


Lifting Behaviors

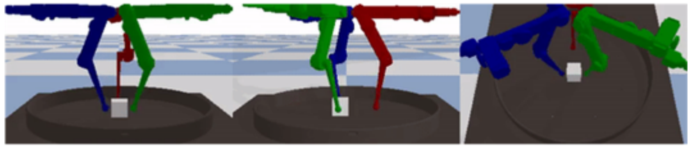


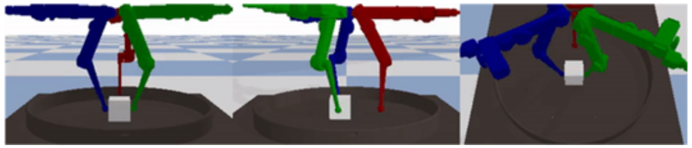




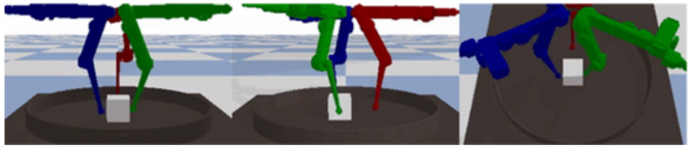


Rotate Behaviors

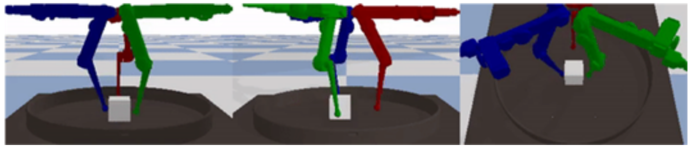




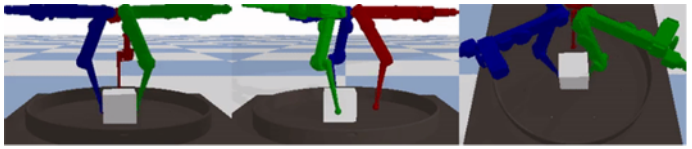
Discovered Behaviors - CausalWorld



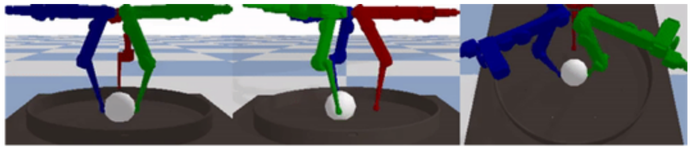
Dribble



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

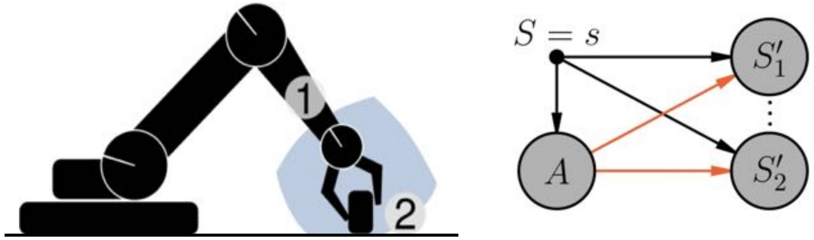
- Use causal influence to speed-up reinforcement learning

Method

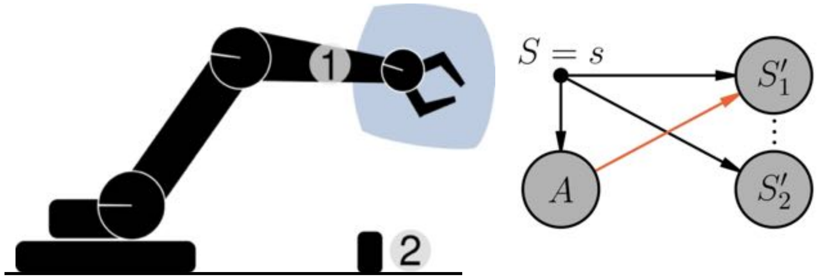
- Define measure of *causal action influence* as a conditional mutual information

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

- Estimate it from data using neural networks



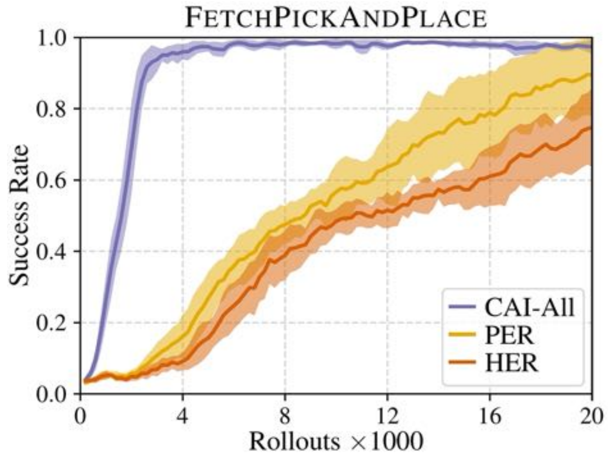
Robot can control object



Causal influence on object impossible

Causal Influence Detection for Reinforcement Learning

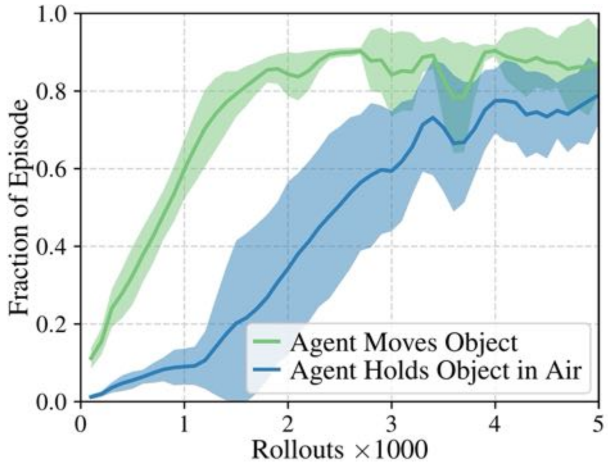
(with Maximilian Seitzer and Georg Martius, arXiv:2106.03443)



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

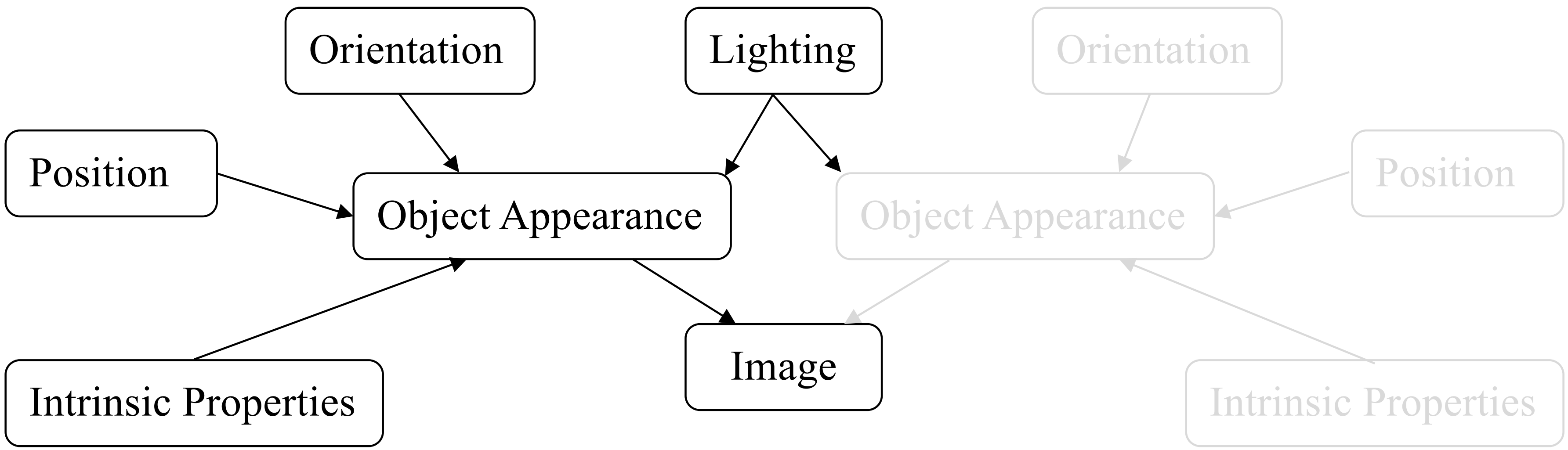
Generative scene models as causal models

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

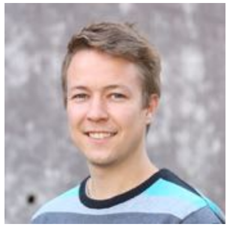
ABSTRACT

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

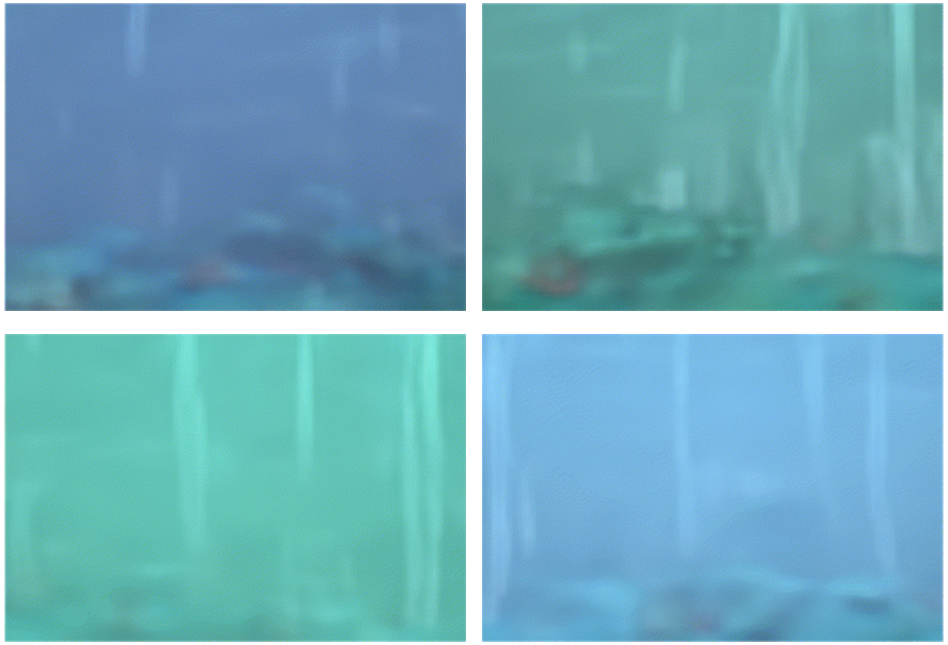
Proposed in the early days of computer vision Grenander (1976); Horn (1977), *analysis-by-synthesis* is an approach to the problem of visual scene understanding. The idea is conceptually elegant and appealing: build a system that is able to synthesize complex scenes (e.g., by rendering), and then understand 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

Tangemann, Schneider et al., 2021

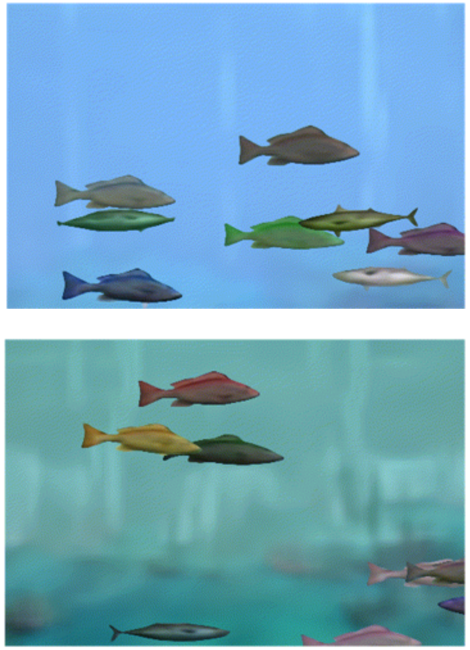
  

Training set

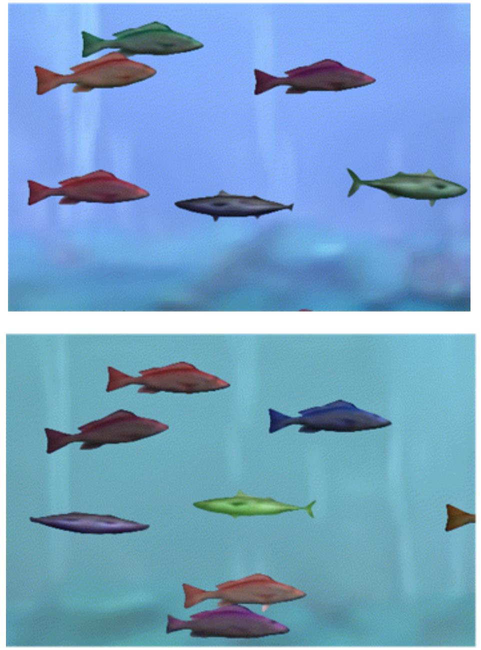
# of objects



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, …

ellis
European Laboratory for Learning and Intelligent Systems

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

MAX-PLANCK-GESELLSCHAFT

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

MAX-PLANCK-GESELLSCHAFT

A Modeling Taxonomy

Task	statistical model	causal model	differential equation model	animal
Predict in i.i.d. setting, pattern recognition, *"generalization"*	y	y	y	y
Predict under shift & intervention, *"horizontal generalization"*	n	y	y	y
Think/Reason, *"act in an imagined space"*	n	?	?	y
Learn from data	y	?	n	y
Provide physical insight, understand predictions	n	?	y/?	n

Thank You