From causality to compression: AI Research at Qualcomm

Johann Brehmer
Staff Engineer

Ties van Rozendaal
Senior Engineer

Jens Petersen
Senior Engineer

Qualcomm Technologies Netherlands B. V.

11:00	**Intro to Qualcomm AI Research**	Johann Brehmer
11:30	**Causality for ML and ML for causality**	Johann Brehmer
12:15	Break	
13:00	**(Adaptive) neural compression**	Ties van Rozendaal
13:45	Break	
14:00	**Perceptual quality in neural compression**	Jens Petersen

Part I
Intro to Qualcomm AI Research

Qualcomm: leading mobile innovation for over 30 years

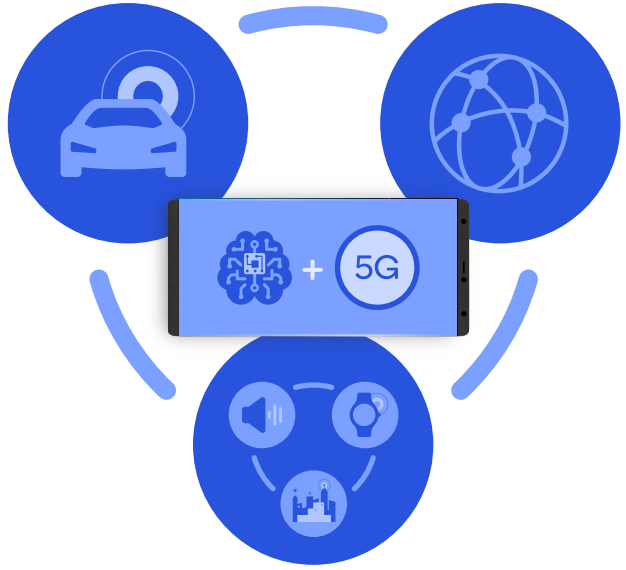
Digitized mobile communications

Analog to digital

Redefined computing

Desktop to smartphones

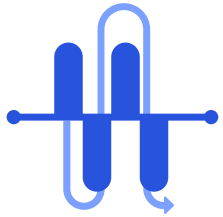
Transforming industries



Connecting virtually everything

R&D: focus on technologies 3-10 years from commercialization

Wi-Fi



Shared Spectrum Technologies

5G

Industrial IOT

Cellular - V2X

Autonomous Driving

Deep Learning

Extended Reality (XR)

Low-power Processors & Sensors

Qualcomm AI Research

- Advance AI to make its core capabilities – perception, reasoning, and action – **ubiquitous across devices**
- Make breakthroughs in **fundamental AI research** and scale them across industries
- Bring together some of the best minds in the field and push boundaries of what's possible

The challenge of AI workloads

- Very compute intensive
- Large, complicated neural network models
- Complex concurrencies
- Real-time
- Always-on

Power and thermal efficiency are essential for on-device AI

Constrained mobile environment

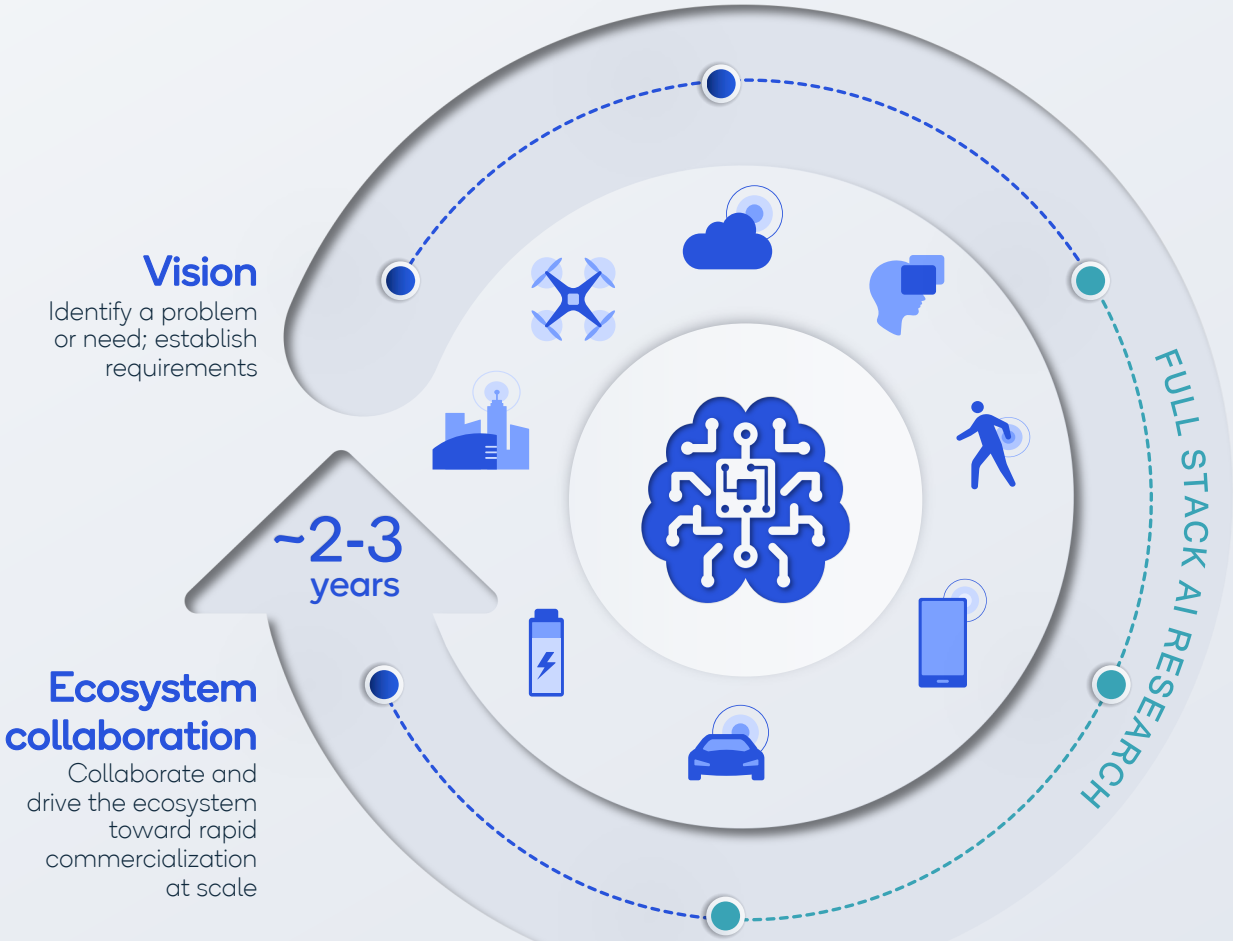
- Must be thermally efficient for sleek, ultra-light designs
- Requires long battery life for all-day use
- Storage/memory bandwidth limitations

Full stack AI research

Model, hardware, and software innovation across each layer to accelerate AI applications

Early R&D and technology inventions essential to leading the ecosystem forward

Transfer tech to commercial teams and influence future research with learnings from deployment



Invention
Invent new methods that set state-of-the-art

Model quantization & optimization
Develop tech & tools to quantize weights and modify architecture to run efficiently on hardware

Software compilation
Develop tech & tools to improve graph-level and kernel-level software compilation performance

Proof of concept
Target teams integrate models into final application for stable and intuitive demonstration

Ecosystem collaboration
Collaborate and drive the ecosystem toward rapid commercialization at scale

Vision
Identify a problem or need; establish requirements

Qualcomm AI Research in Amsterdam

Image source: Hanno Ackermann

Topics we work on in Amsterdam

Causality, interactive learning, and robotics

- Affordance / skill learning
- Imitation learning
- Causal representation learning
- Geometric structure

Combinatorial optimization

- Combinatorial Bayesian optimization
- Chip design

Data compression

- Generative models
- End-to-end video compression
- Audio compression

Wireless AI

- Wireless channel estimation
- Learning physics of RF propagation
- Error propagation with GNNs

Federated learning

- Massively distributed computing
- Differential privacy

Model efficiency

- NN compression
- Quantization
- Conditional compute

Perception

- Efficient video processing
- Detection, segmentation, scene understanding
- Wireless RF sensing

XR

- VR & AR on embedded devices
- 3D reconstruction and scene understanding

more fundamental

more applied

Example papers

Published as a conference paper at ICLR 2018

SPHERICAL CNNS

Taco S. Cohen*
University of Amsterdam

Mario Geiger*
EPFL

Jonas Köhler*
University of Amsterdam

Max Welling
University of Amsterdam & CIFAR

ABSTRACT

Convolutional Neural Networks (CNNs) have become the method of choice for learning problems involving 2D planar images. However, a number of problems of recent interest have created a demand for models that can analyze spherical images. Examples include omnidirectional vision for drones, robots, and autonomous cars, molecular regression problems, and global weather and climate modelling. A naive application of convolutional networks to a planar projection of the spherical signal is destined to fail, because the space-varying distortions introduced by such a projection will make translational weight sharing ineffective.

In this paper we introduce the building blocks for constructing spherical CNNs. We propose a definition for the spherical cross-correlation that is both expressive and rotation-equivariant. The spherical correlation satisfies a generalized Fourier theorem, which allows us to compute it efficiently using a generalized (non-commutative) Fast Fourier Transform (FFT) algorithm. We demonstrate the

Best paper award
@ ICLR 2018

Example papers

A White Paper on Neural Network Quantization

Markus Nagel*
Qualcomm AI Research†
markusn@qti.qualcomm.com

Marios Fournarakis*
Qualcomm AI Research†
mfournar@qti.qualcomm.com

Rana Ali Amjad
Qualcomm AI Research†
ramjad@qti.qualcomm.com

Yelysei Bondarenko
Qualcomm AI Research†
ybodaren@qti.qualcomm.com

Mart van Baalen
Qualcomm AI Research†
mart@qti.qualcomm.com

Tijmen Blankevoort
Qualcomm AI Research†
tijmen@qti.qualcomm.com

Abstract

While neural networks have advanced the frontiers in many applications, they often come at a high computational cost. Reducing the power and latency of neural network inference is key if we want to integrate modern networks into edge devices with strict power and compute requirements. Neural network quantization is one of the most effective ways of achieving these savings but the additional noise it induces can lead to accuracy degradation.

Example papers

FrameExit: Conditional Early Exiting for Efficient Video Recognition

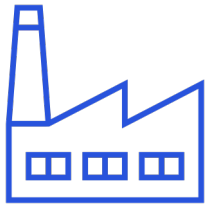
Amir Ghodrati* Babak Ehteshami Bejnordi* Amirhossein Habibian
Qualcomm AI Research†
{ghodrati, behtesha, ahabibia}@qti.qualcomm.com

Abstract

In this paper, we propose a conditional early exiting framework for efficient video recognition. While existing works focus on selecting a subset of salient frames to reduce the computation costs, we propose to use a simple sampling strategy combined with conditional early exiting to enable efficient recognition. Our model automatically learns to process fewer frames for simpler videos and more frames for complex ones. To achieve this, we employ a cascade of gating modules to automatically determine the earliest point in processing where an inference is sufficiently reliable. We generate on-the-fly supervision signals to the

Example demo

NeurIPS 2022: 3 / 8 of official expos were Qualcomm demos!



Research at industry labs

Choose topics based on business needs / product relevance and curiosity

Work by thinking, math-ing, implementing, experimenting, writing, presenting

Collaborate in flat hierarchies

Strive for papers, demos, or impact on products

Enjoy great working conditions

Research in academia

Choose topics based on curiosity (and funding opportunities)

Same

Collaborate mostly in professor-postdoc-student hierarchies

Strive for papers, theses, funding, or jobs

You tell me :)

(NB: this is my personal opinion from limited experience)

Qualcomm

Products Support Company

Research / Artificial Intelligence / Artificial Intelligence R...

Qualcomm AI Research

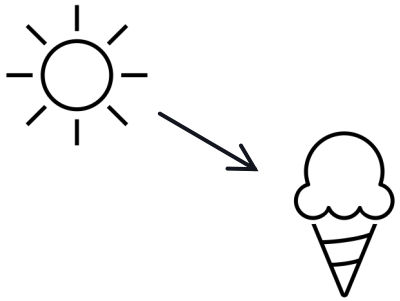
At Qualcomm AI Research, we are advancing AI to make its core capabilities – perception, reasoning, and action – ubiquitous across devices. Our mission is to make breakthroughs in fundamental AI research and scale them across industries. By bringing together some of the best minds in the field, we're pushing the boundaries of what's possible and shaping the future of AI.

About AI Research Areas AI Papers Open Source Demo Videos Webinar Videos Blog Posts Presentation Downloads Newsletter

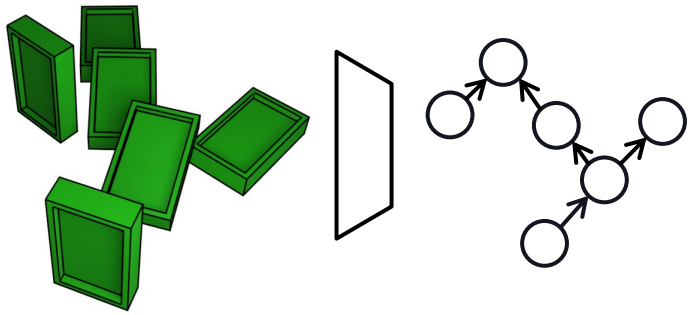
We're hiring

www.qualcomm.com/research/artificial-intelligence/ai-research

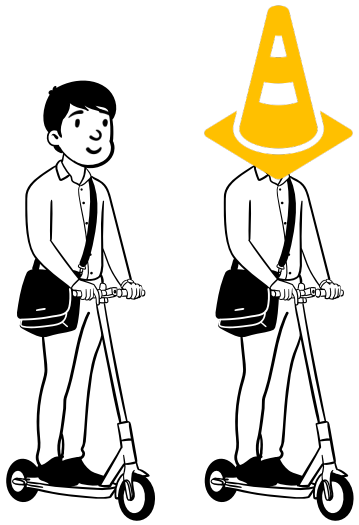
Part II
Causality for ML and ML for causality



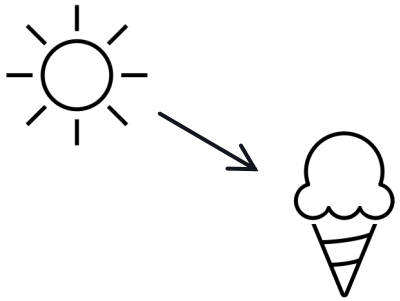
Intro:
Why causality?



ML for causality:
Learning causal variables and causal structure from non-iid data



Causality for ML:
Deconfounding imitation learning

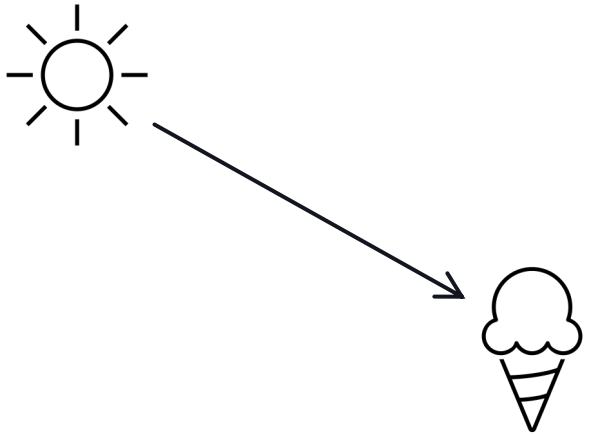


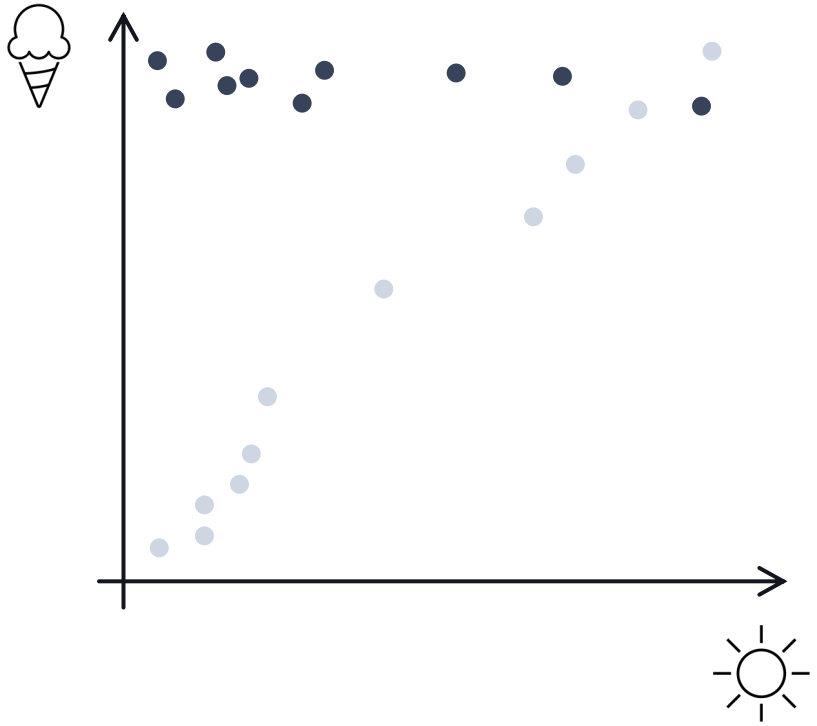
Intro:
Why causality?

ML for causality:
Learning causal variables and causal structure from non-iid data

Causality for ML:
Deconfounding imitation learning

Causality

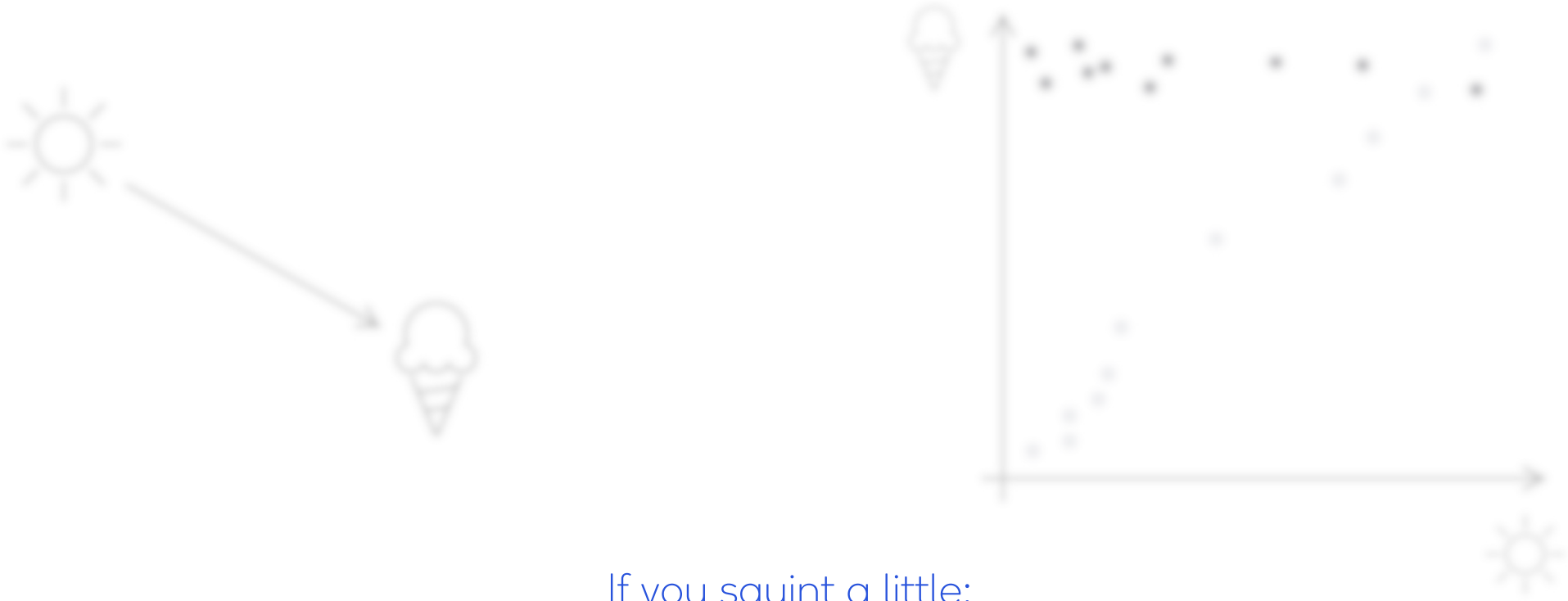




Semantically, causal models label relations between random variables as **cause-effect relations**

Functionally, causal models describe **probability distributions and how they change** under changing conditions

Causality

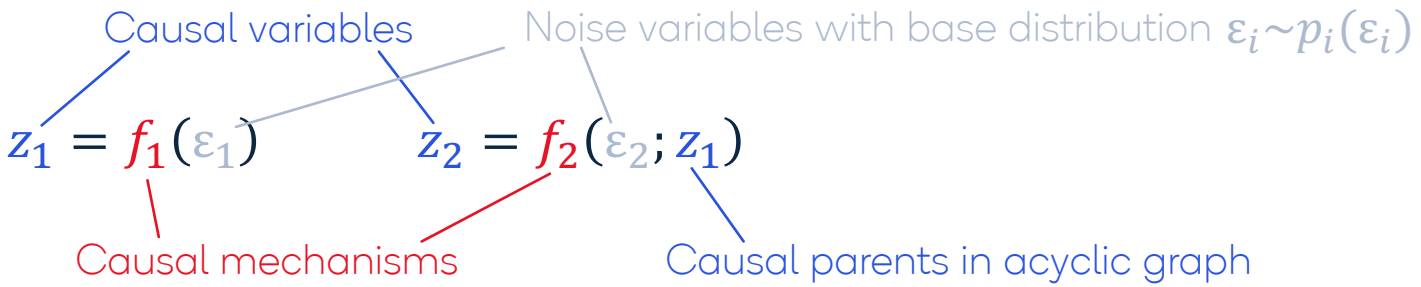


If you squint a little:

Causality is a language to reason about certain kind of **changes**

Formalism: **Structural causal models** (SCMs)

- **SCM:**



- **Solution:**

$$z = s(\varepsilon) \;\Rightarrow\; z \sim p_z(z)$$

Solution function (= successively applying causal mechanisms)

Observational distribution

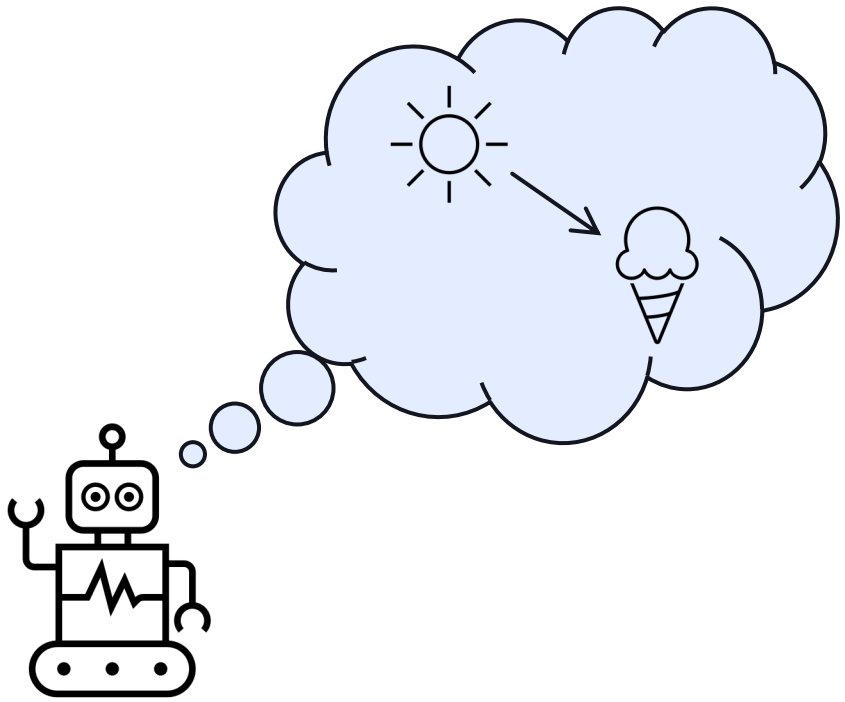
- **Interventions:**

$$f_i(\varepsilon_i; z_{\text{parents}}) \rightarrow \tilde{f}_i(\varepsilon_i) \qquad \Rightarrow \; z \sim \tilde{p}_z^i(z)$$

New mechanism (perfect intervention: no parents)

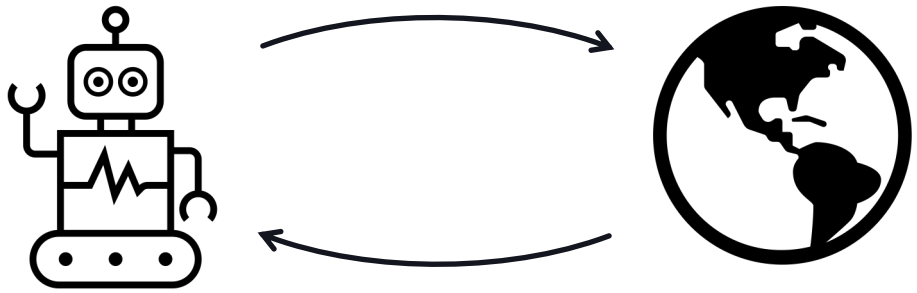
Interventional distribution

ML for causality:
Learn causal models of the world



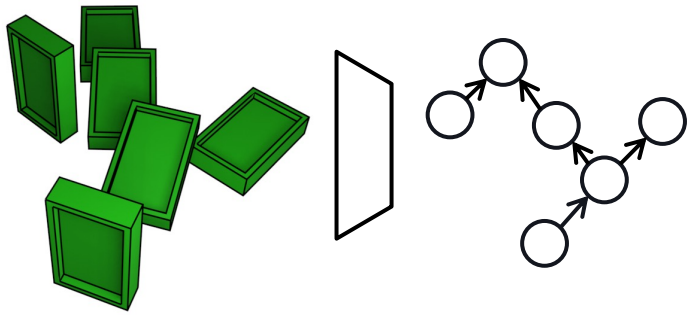
- **Causal inference / discovery**
  - given dataset of causal variables, learn causal graph and causal mechanisms
  - applications: healthcare, economics, ...
- **Causal representation learning**
  - given dataset of low-level variables (pixels), learn encoder to causal variables, causal graph, and causal mechanisms
  - applications: robotics, autonomous driving, ...
- **Causal models of the world may be useful...**
  - out of scientific interest
  - for planning
  - to increase robustness to domain shift

Causality for ML:
Take causal structure of the learning process into account



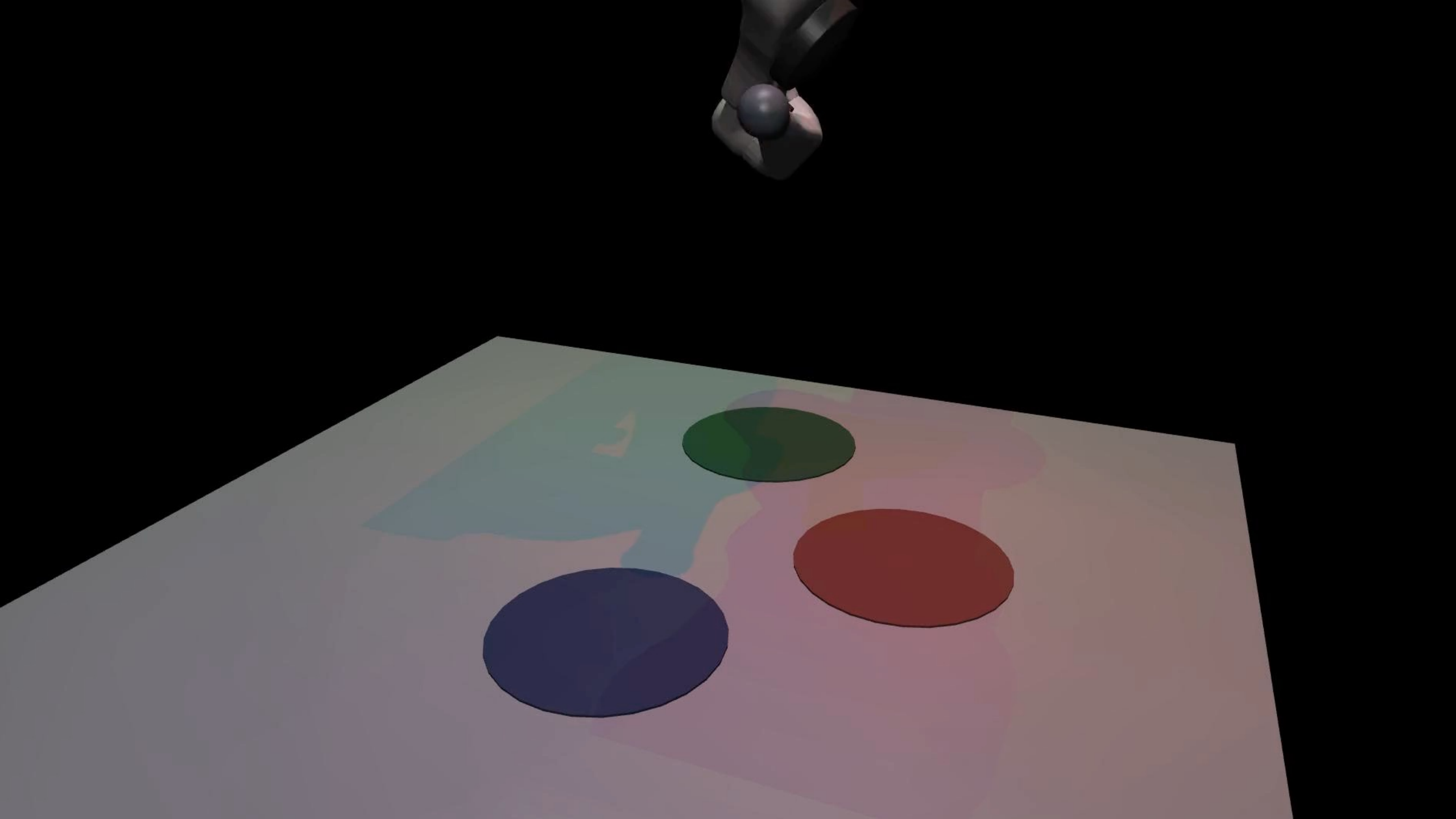
- **Imitation learning and sequence modelling**
  - given expert trajectories,
    learn policy that imitates expert's behaviour
  - need to be robust to confounding:
    expert may have access to different information
  - applications in robotics, autonomous driving, ...
- **Causality-aware learning algorithms are useful...**
  - when naïve learning algorithms make fundamental mistakes and fail even with infinite data

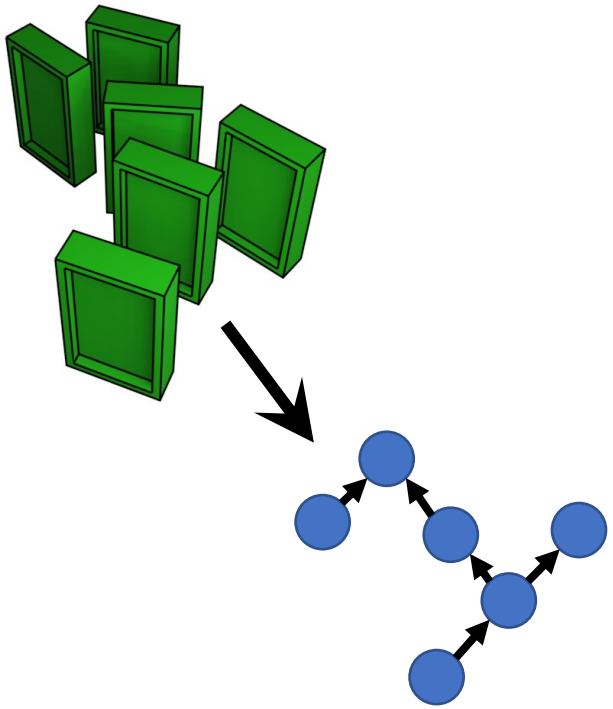
Intro:
Why causality?



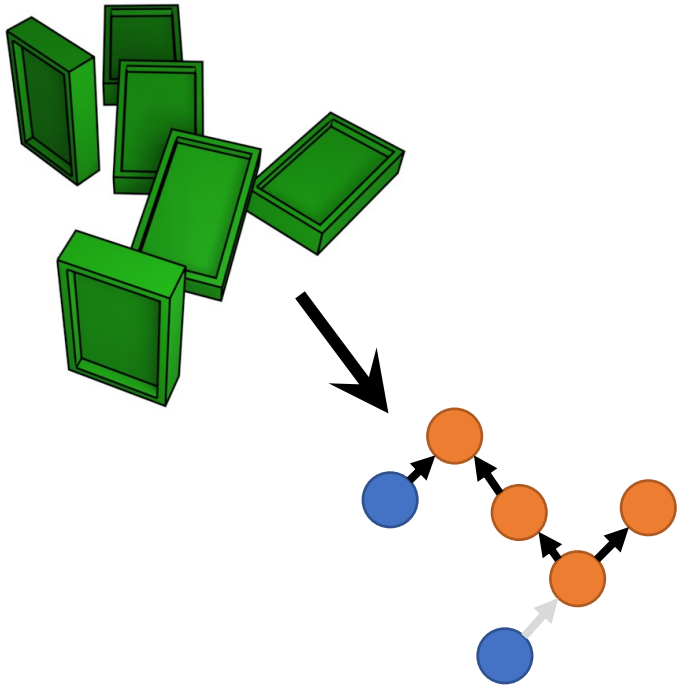
ML for causality:
Learning causal variables and causal structure from non-iid data

Causality for ML:
Deconfounding imitation learning

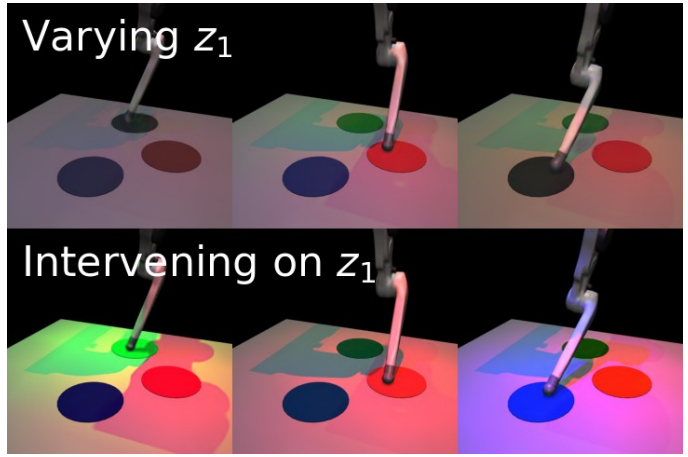




Can we **learn causal variables & causal structure from pixels**, without labels?

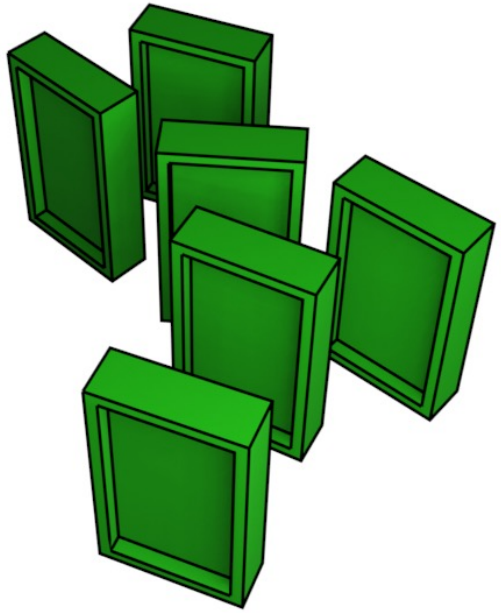


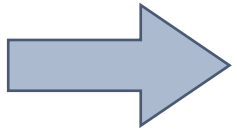
We prove: this is possible with **weak supervision**, when observing effects of interventions

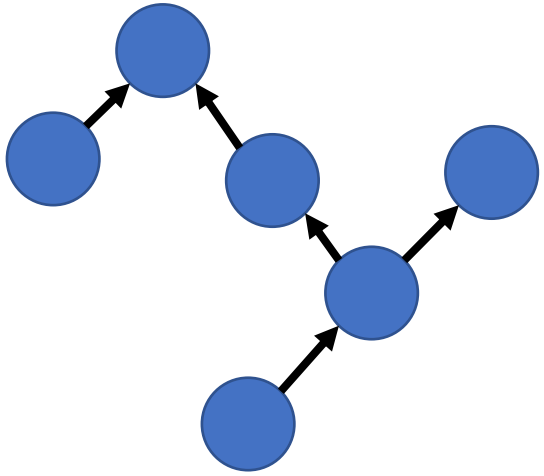


In practice, **implicit latent causal models** can identify the causal structure in image datasets

Causal representation learning



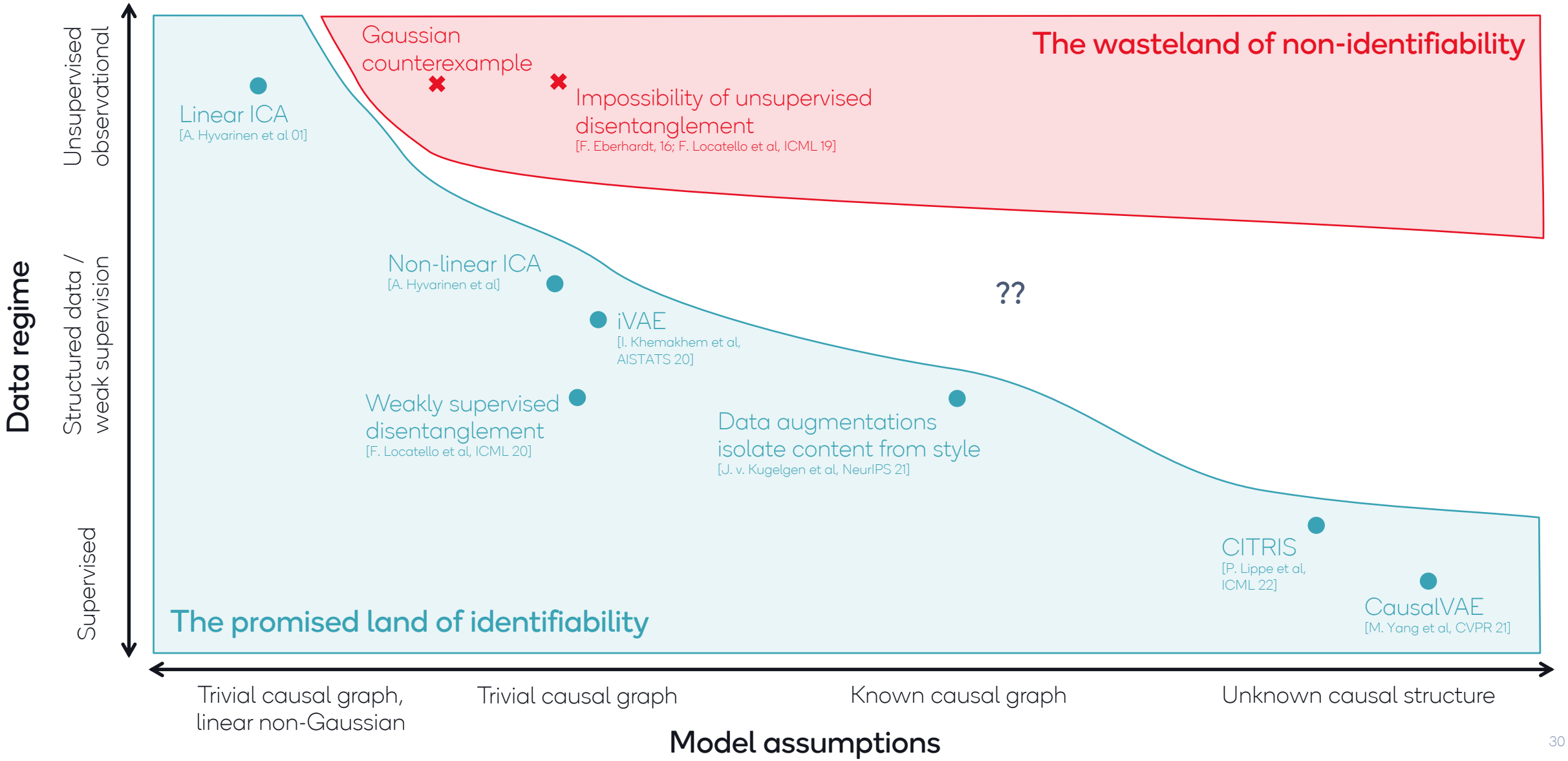




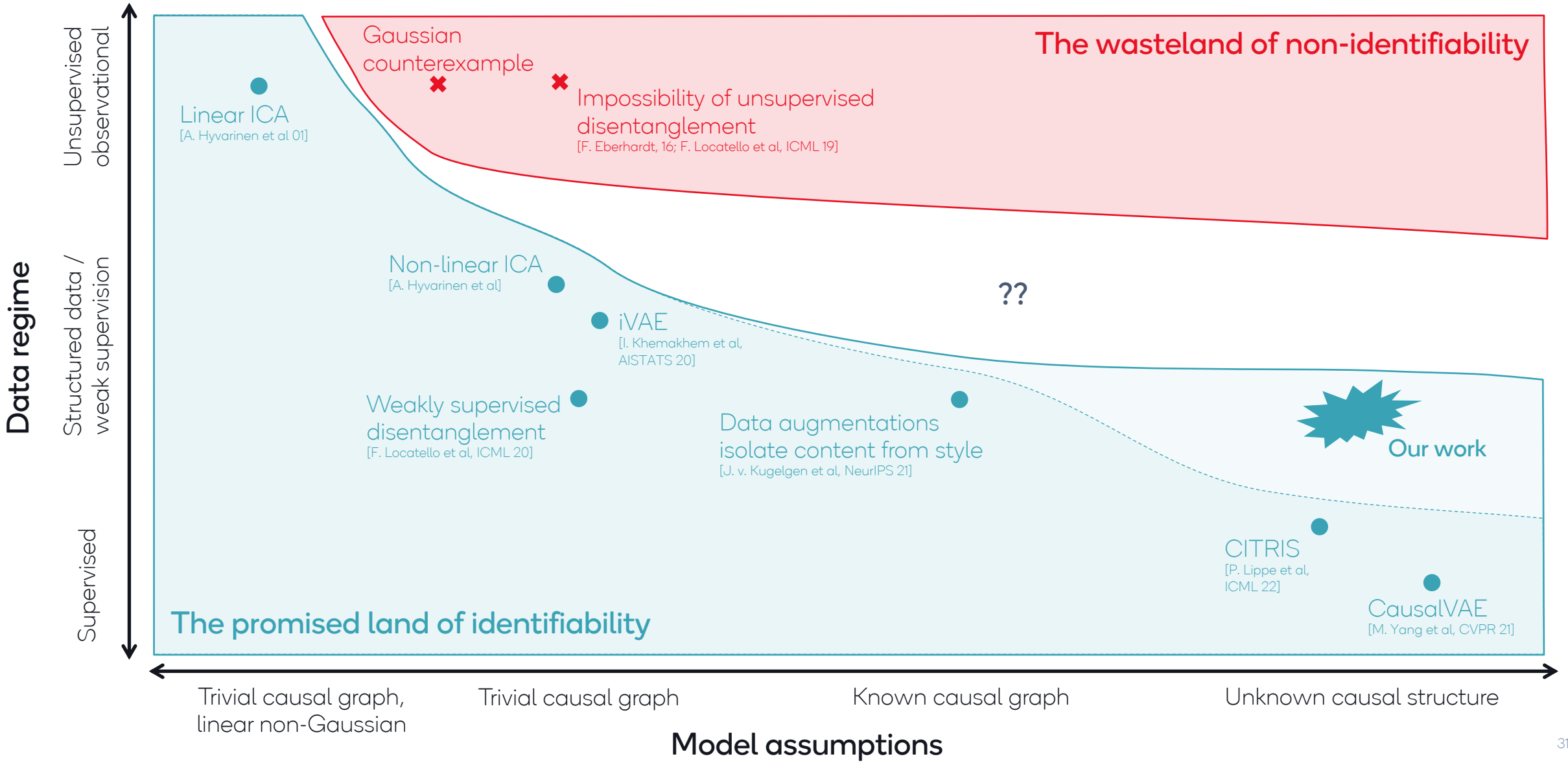
Given: **low-level, unstructured data representation**
(e.g. pixels)

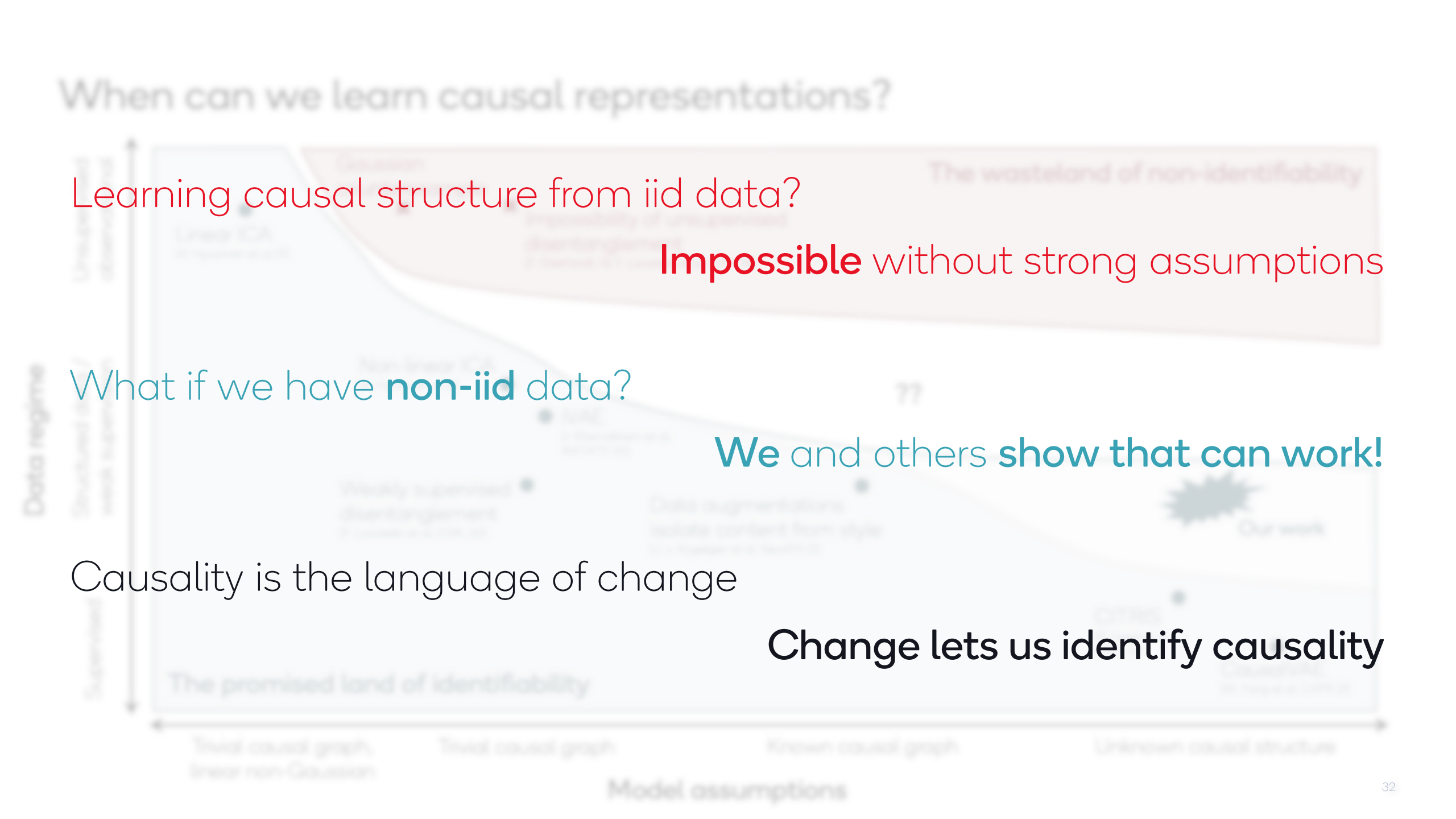
Goal: learn encoder to
high-level variables
(e.g. object positions, states, ...)
and their relations / causal structure

When can we learn causal representations?

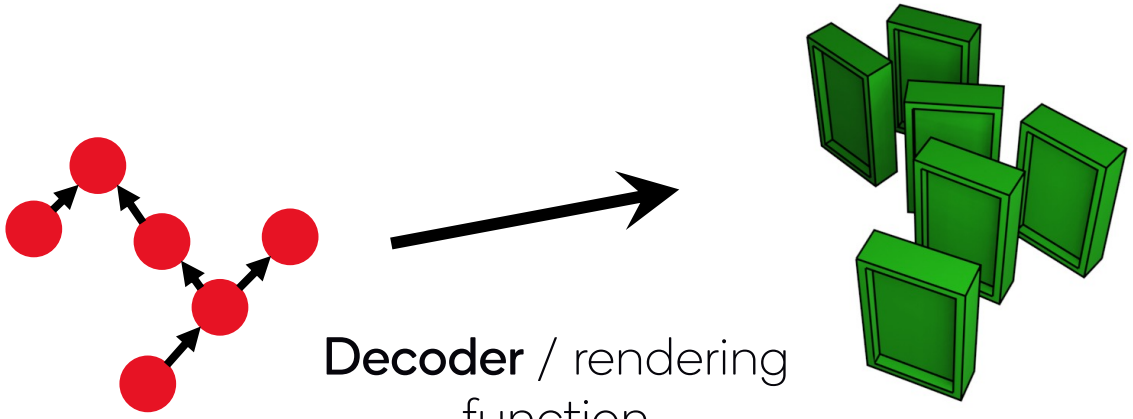


When can we learn causal representations?




When can we learn causal representations?
Learning causal structure from iid data?
Impossible without strong assumptions
What if we have non-iid data?
We and others show that can work!
Causality is the language of change
Change lets us identify causality
Data regime
Model assumptions

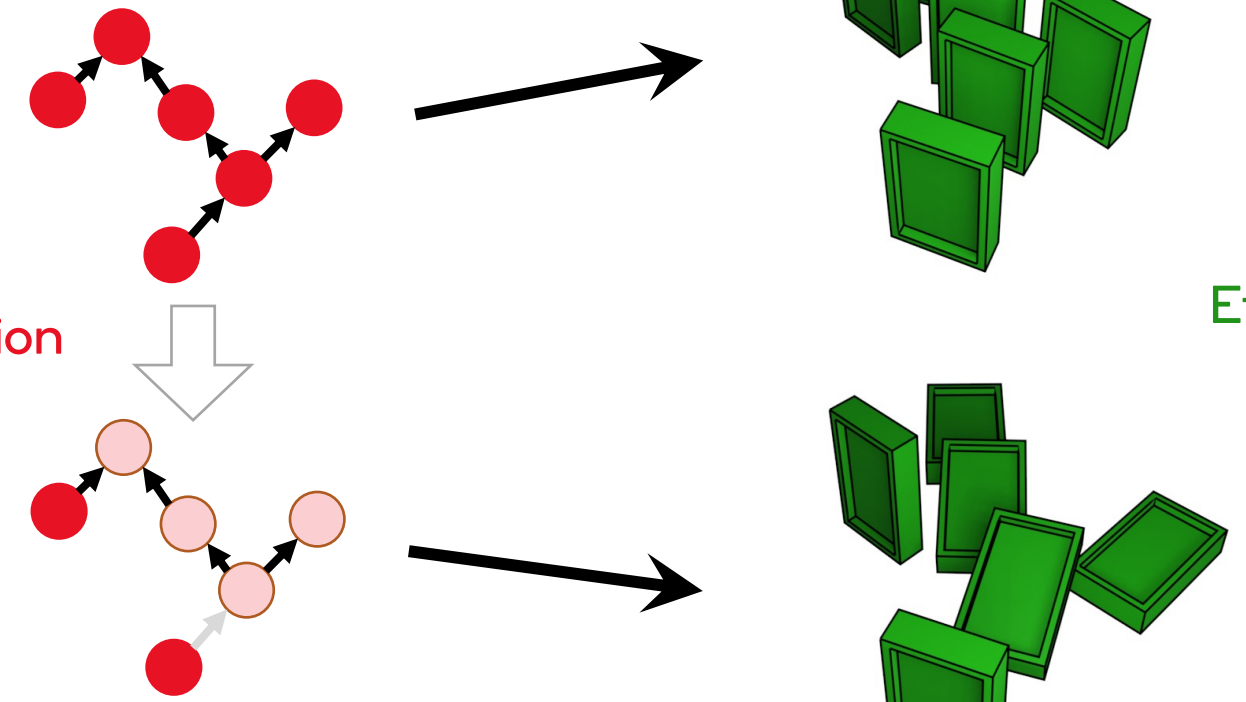
Theory: **Latent causal model**



High-level variables with a structural causal model between them

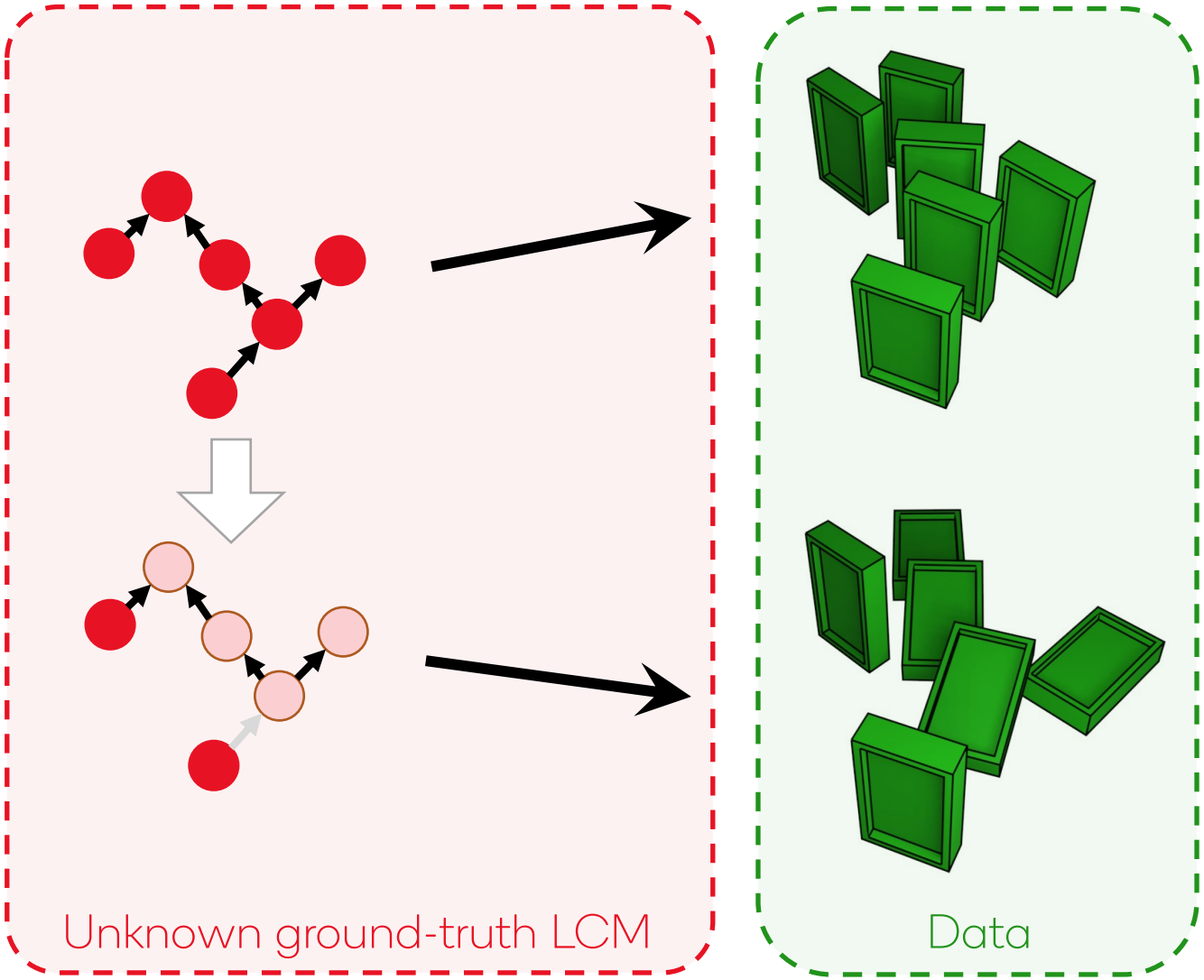
Low-level data (pixels)

Interventions



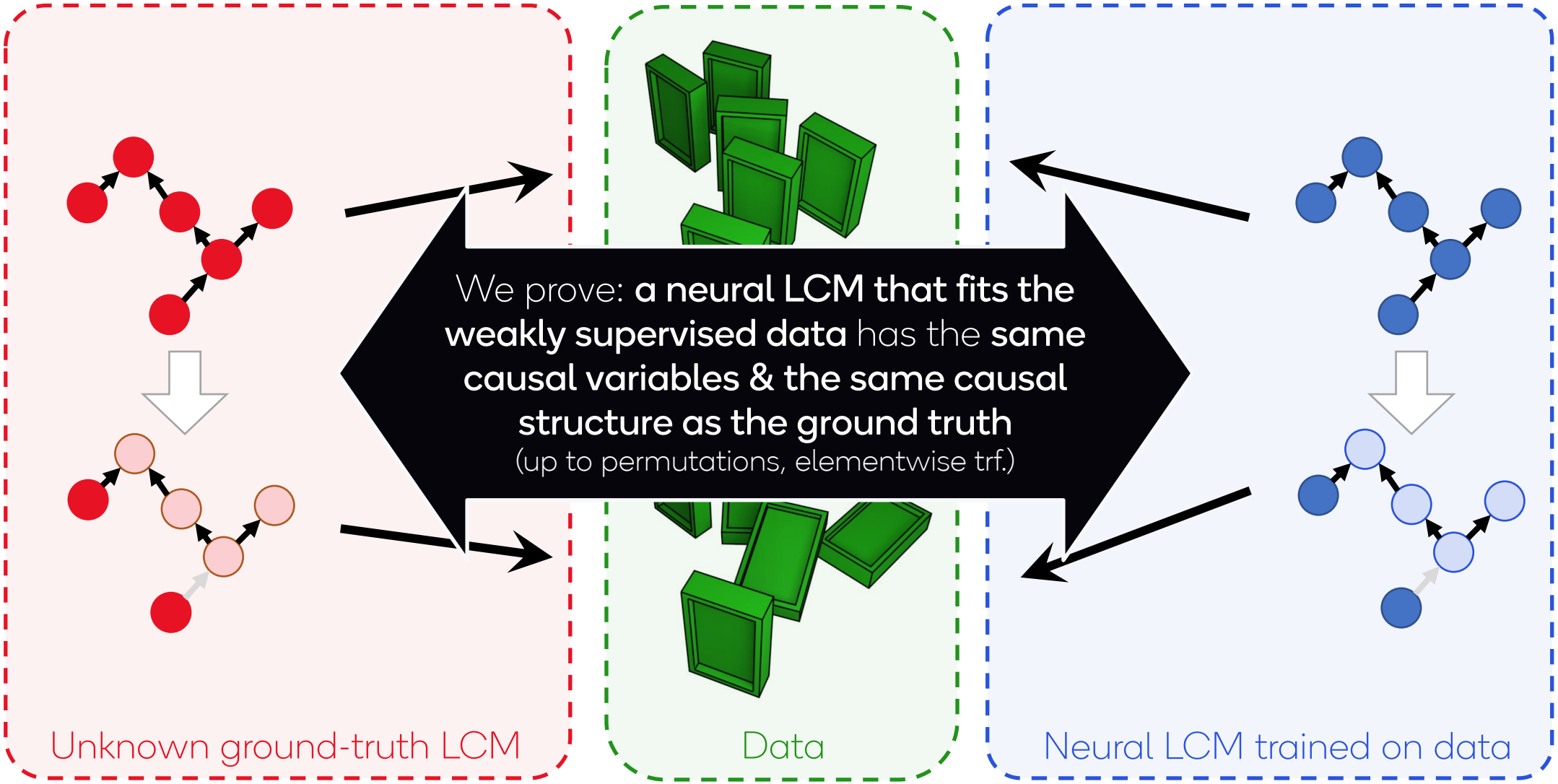
Effect of intervention
in data space

Weakly supervised data setting



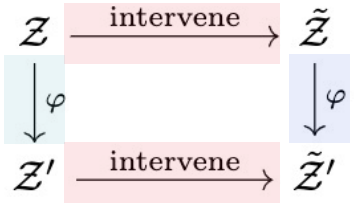
- We assume access to **data pairs of the system before and after interventions**
  - Equivalent to counterfactuals
  - Causal abstraction of time-series data
- Otherwise, **no labels**
  - Only pixel-level data is observed
  - Intervention targets are unknown

Identifiability theorem

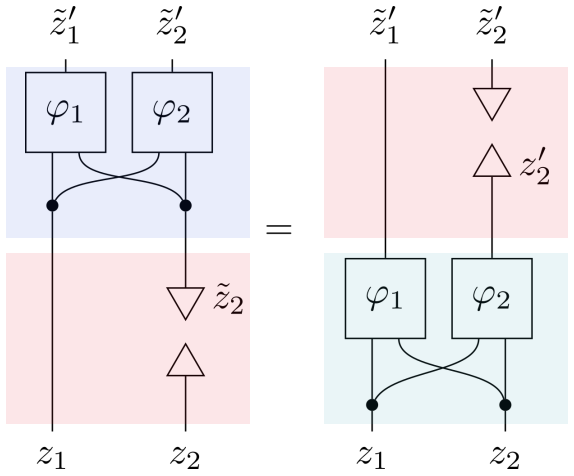


Proof sketch

1. Consider two LCMs with causal variables z and z', both matching the data.
   Define $\varphi : z \to z'$.
2. Interventions commute with φ:



3. We assume perfect interventions. Then then $\tilde{z}'_i$ is independent of z_j . For 2 variables:

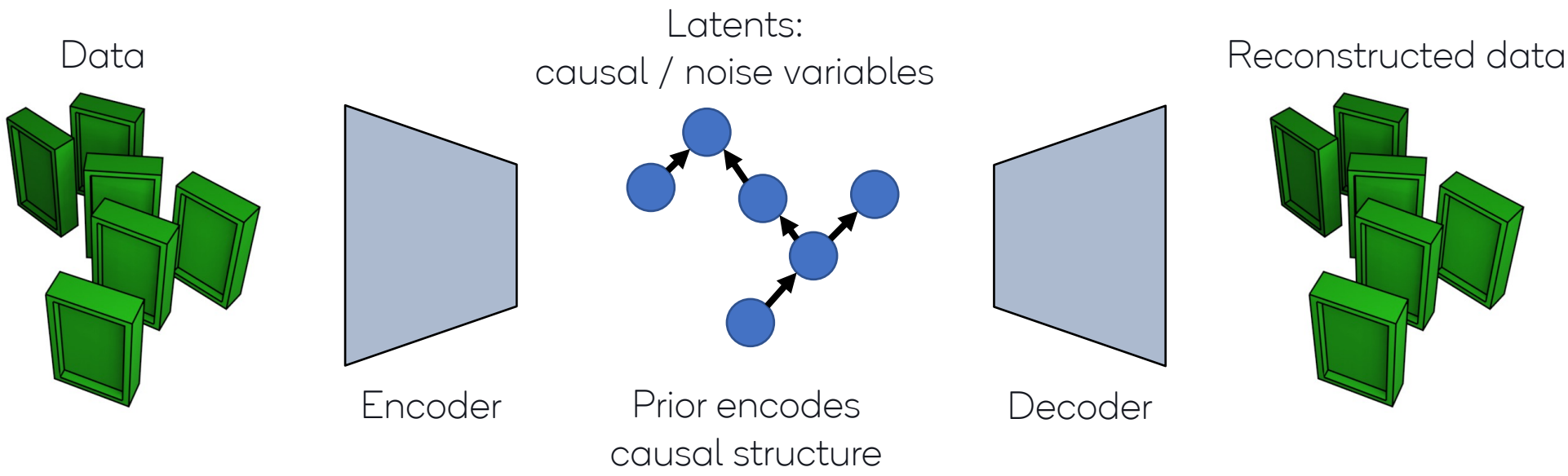


4. We assume $\mathbb{R}$-valued variables. Statistical independence then implies functional independence. Thus, $\varphi_i(z_i, z_j)$ must be constant in z_j.
5. Since this holds for any i, φ must be a permutation plus elementwise transformations.
6. Finally, we can show that the causal graphs and intervention targets in the two models are consistent with this transformation.
7. Thus the two models are isomorphic.

No theorem without assumptions

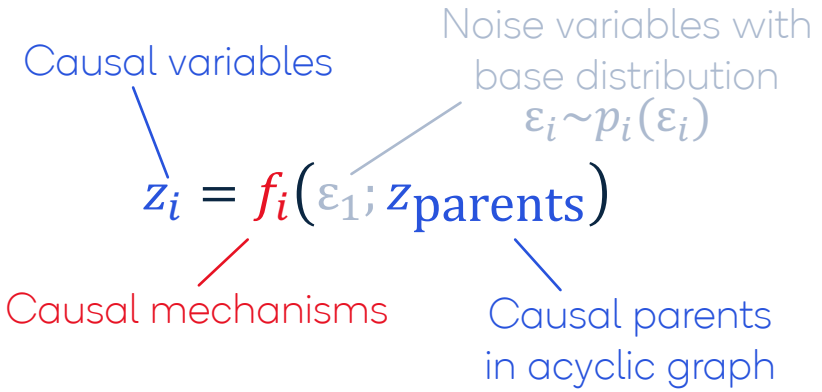
Assumption	Possible relaxation
Weakly supervised data is available	Maybe (first results)
Causal variables are $\mathbb{R}$-valued	Maybe (some ideas)
Causal mechanisms are diffeomorphic	Difficult
No hidden confounders	Difficult
Decoder is deterministic	Plausible (as in iVAE)
Interventions are perfect (Post-intervention values of intervention targets are independent of pre-intervention state)	Difficult (counterexamples)
Interventions are complete (The dataset contains interventions on any single causal variable)	Relaxation to n-target interventions plausible (incomplete interventions → partial identifiability)

Operationalizing latent causal models

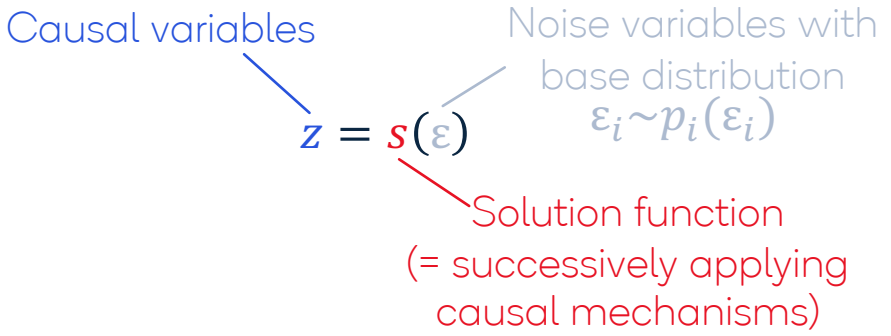


The devil in the details

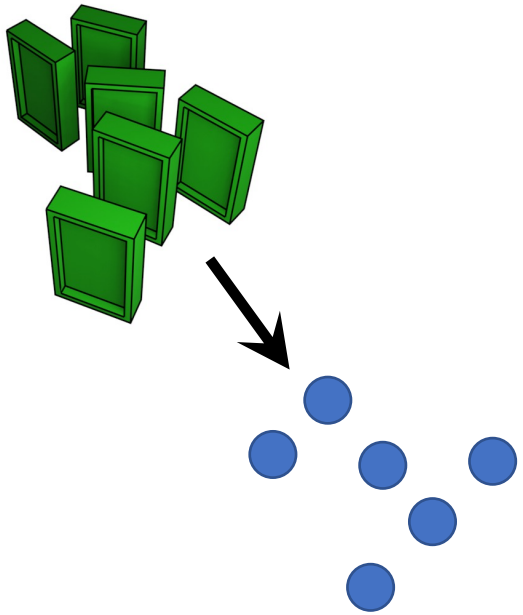
- Key question: **how do you represent the causal structure in the prior?**
- **Explicitly** through learnable causal graph and mechanisms?
  - **Does not work well empirically**
  - Learning a graph parameterization is prone to local minima



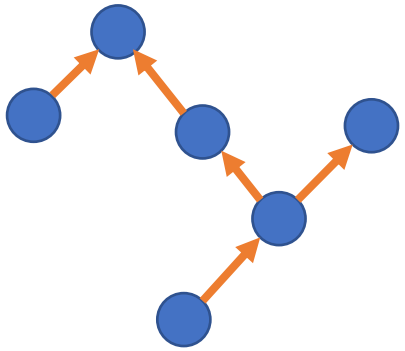
- **Implicitly** through learnable solution function!
  - **Much more robust**
  - Lesson: Don't learn graphs if you don't have to



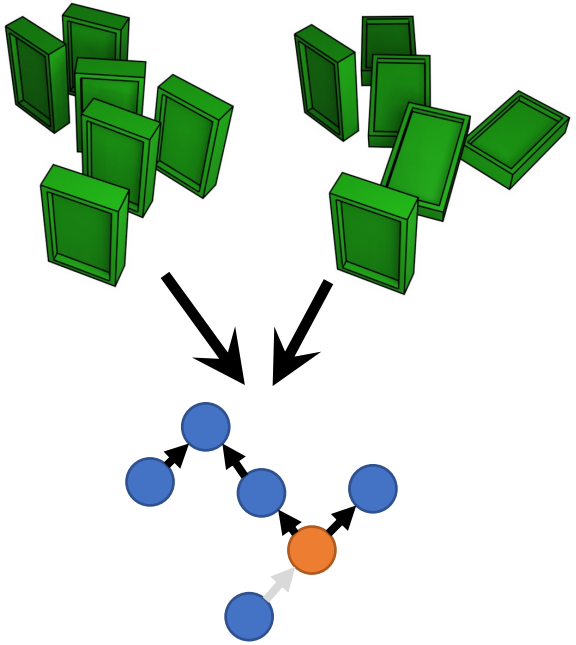
What can you do with ILCMs?



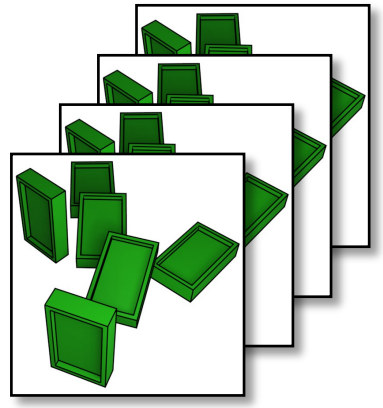
Map pixels to causal variables



Find the causal graph

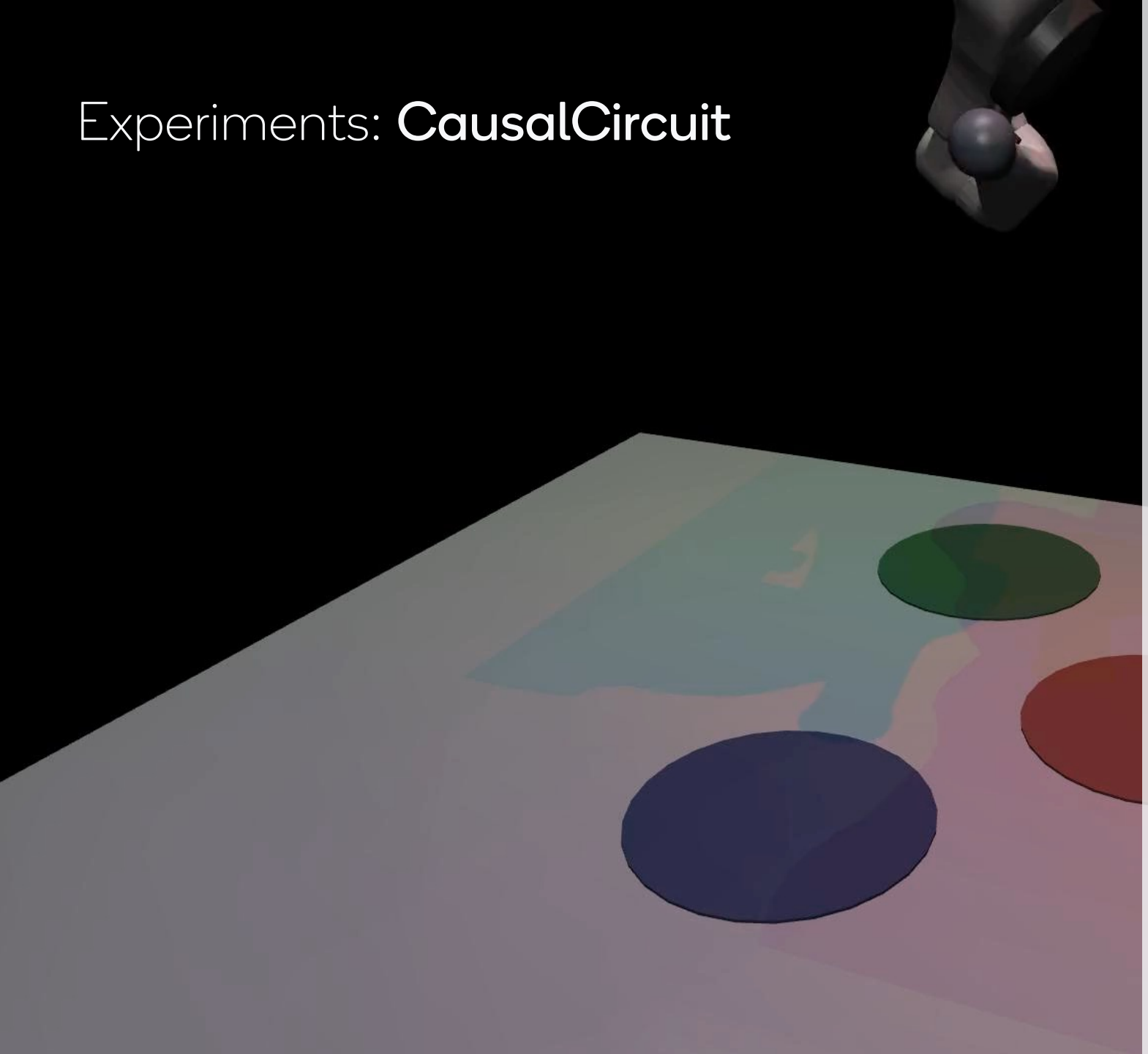


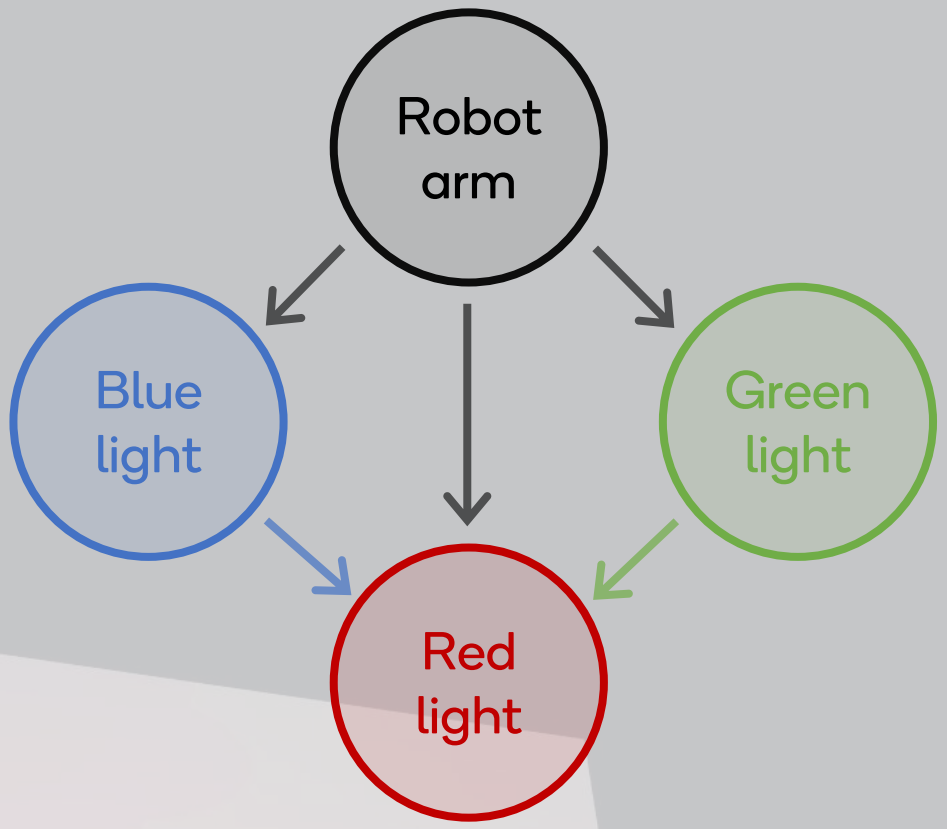
Infer interventions
from data pairs



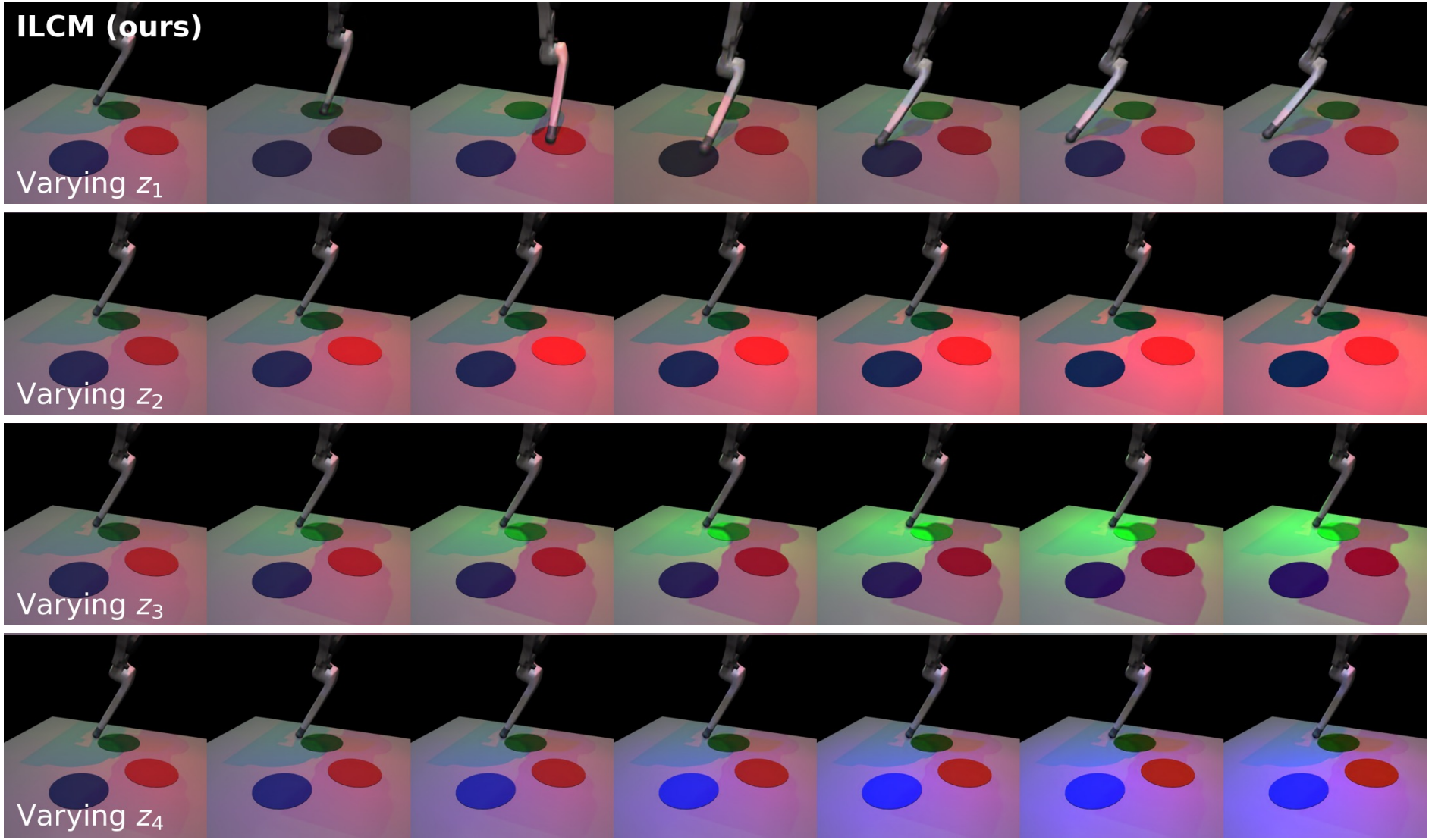
Generate observational, interventional, and counterfactual data

Experiments: **CausalCircuit**

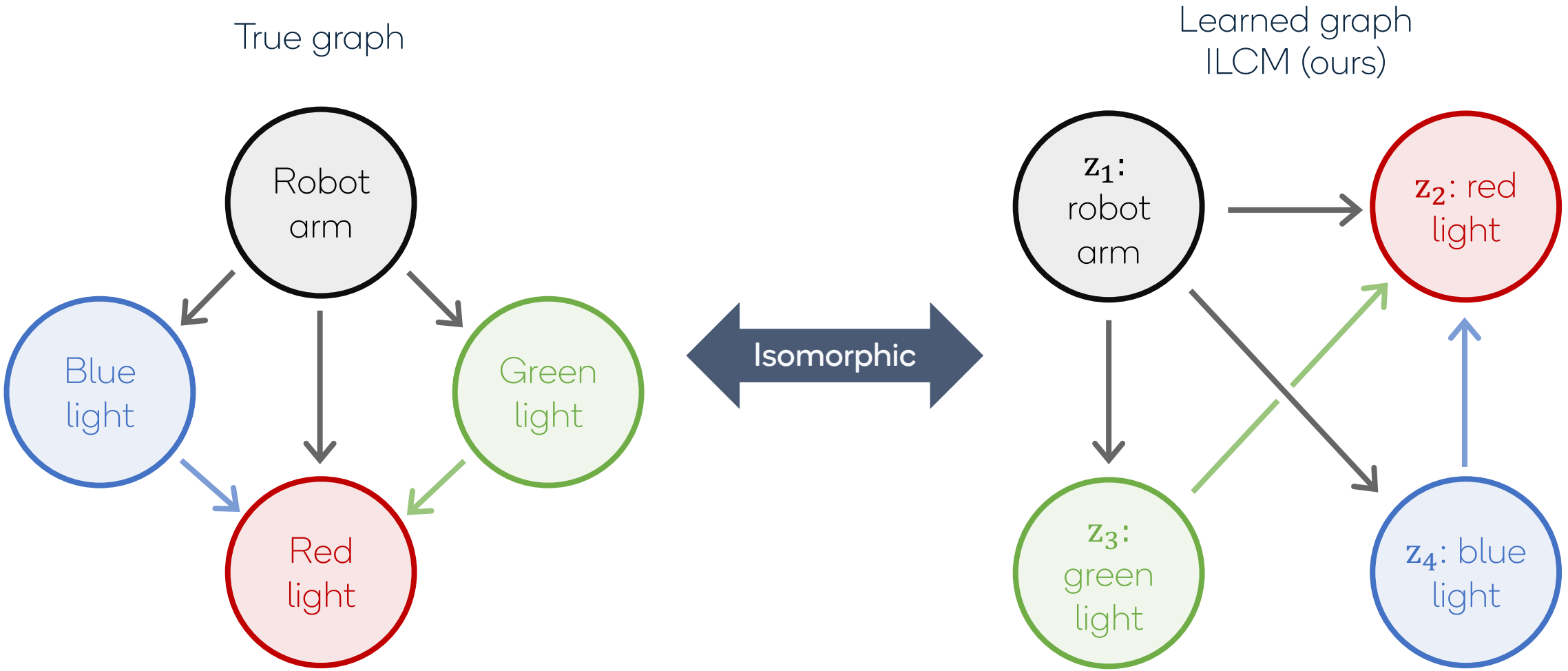




LCMs **disentangle** the causal variables

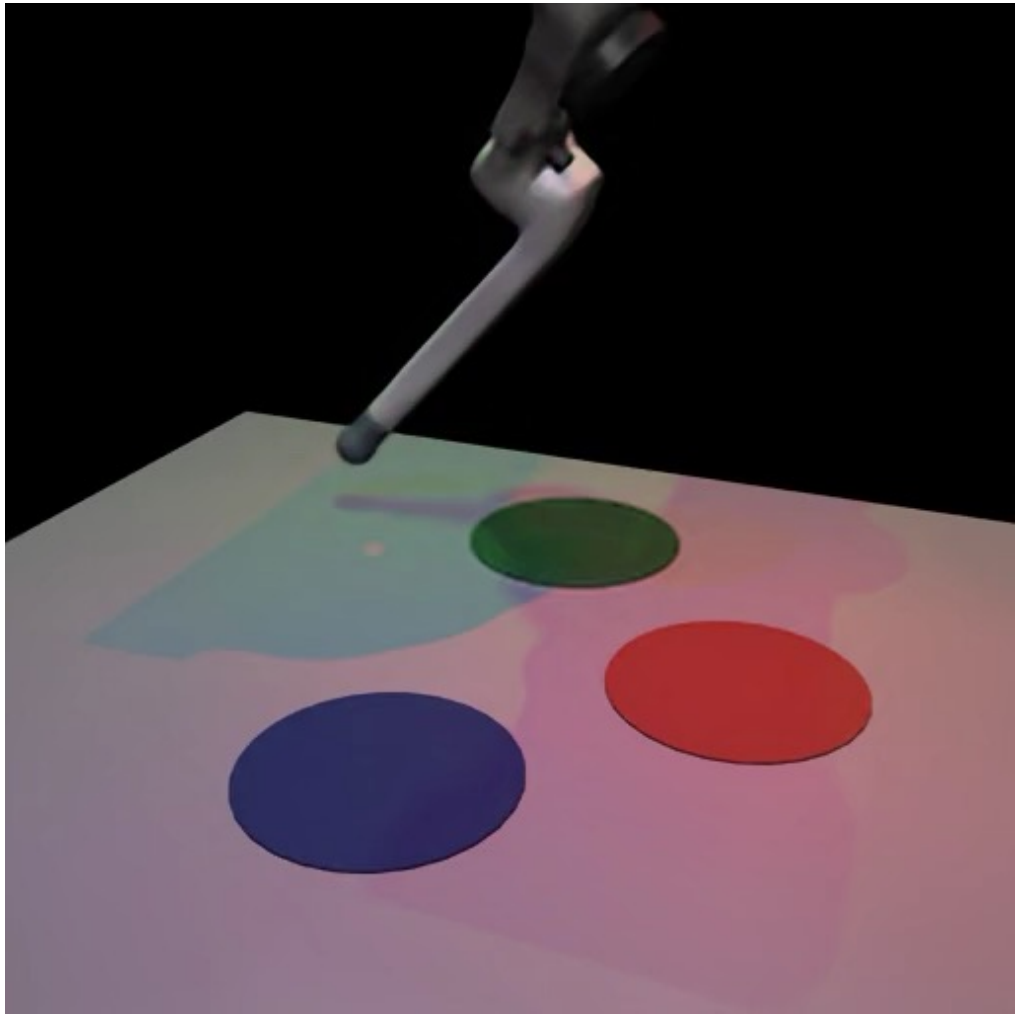


LCMs learn the **correct graph**



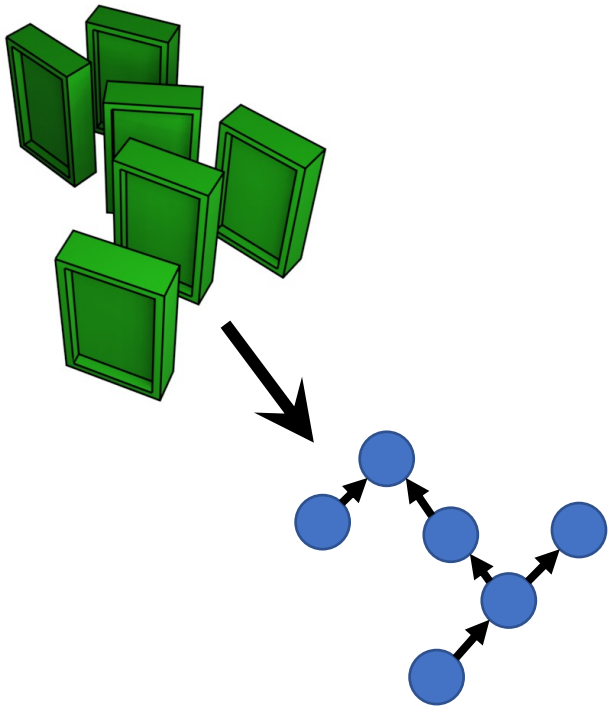
LCMs let us **reason causally**

LCM samples, **intervening** on a single latent (including causal effects)

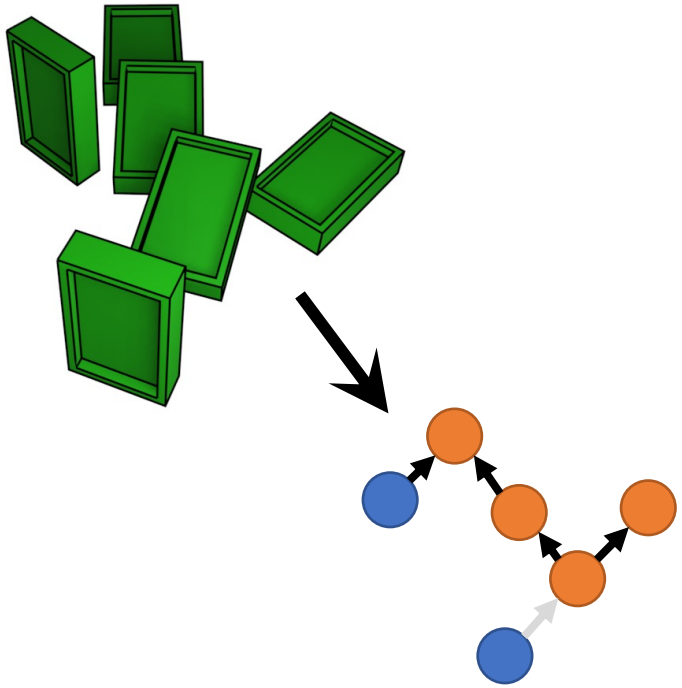


A long way to go

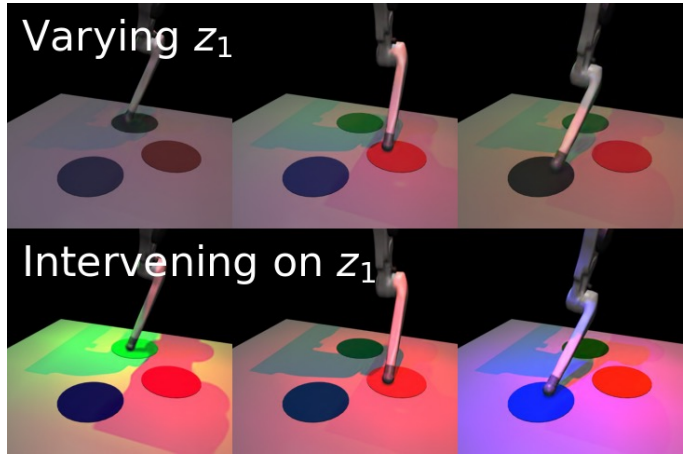
Where we are	Where we need to get
Identifiability theorems	**Demonstrate usefulness** on downstream tasks
Pre- & post-intervention data	**Realistic data regimes:** observational & interventional data, video data, ...
God-given interventions	**Learning intervention policies**
Fixed causal variables	**Variable scene composition**
Strict **DAG-based causality**	**Weaker relational structures**
Toy experiments (up to O(10) variables)	**Realistic experiments**



Can we **learn causal variables & causal structure from pixels**, without labels?



We prove: this is possible with **weak supervision**, when observing effects of interventions



In practice, **implicit latent causal models** can identify the causal structure in image datasets

Weakly supervised causal representation learning

JB*, Pim de Haan*, Phillip Lippe, Taco Cohen

*equal contribution

NeurIPS 2022

arXiv:2203.16437

Pim de Haan

Phillip Lippe

Taco Cohen

Towards causal representation learning

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer,
Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, Yoshua Bengio
IEEE 2021, arXiv:2102.11107

Weakly-supervised disentanglement without compromises

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf,
Olivier Bachem, Michael Tschannen
ICML 2020, arXiv:2002.02886

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

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel,
Bernhard Schölkopf, Michel Besserve, Francesco Locatello
NeurIPS 2021, arXiv:2106.04619

CITRIS: Causal identifiability from temporal intervened sequences

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco
Cohen, Efstratios Gavves
ICML 2022, arXiv:2202.03169

Interventional causal representation learning

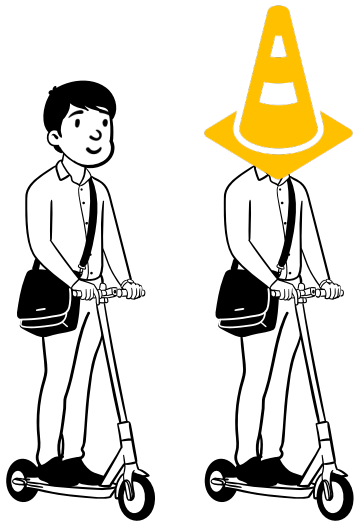
Kartik Ahuja, Divyat Mahajan, Yixin Wang, Yoshua Bengio
arXiv:2209.11924

Causal triplet: an open challenge for intervention-centric causal representation learning

Yuejiang Liu, Alexandre Alahi, Chris Russell, Max Horn, Dominik
Zietlow, Bernhard Schölkopf, Francesco Locatello
CLeaR 2023, arXiv:2301.05169

Intro:
Why causality?

ML for causality:
Learning causal variables and causal structure from non-iid data



Causality for ML:
Deconfounding imitation learning

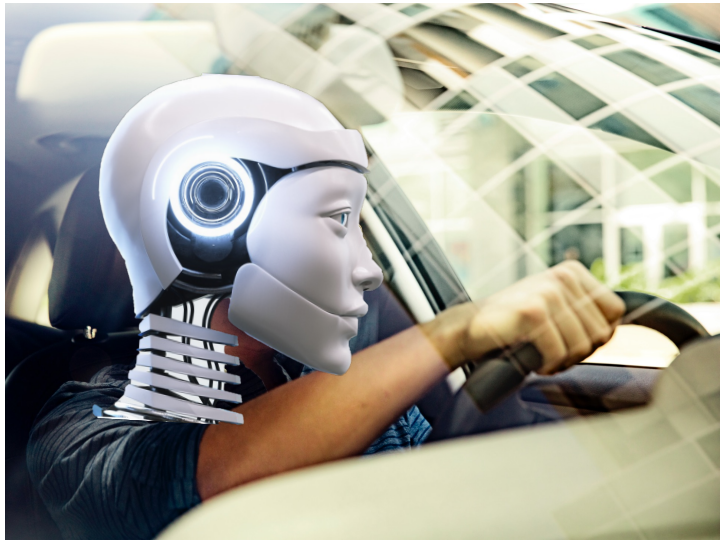
Imitation learning

Human expert

Record observations, actions

Data

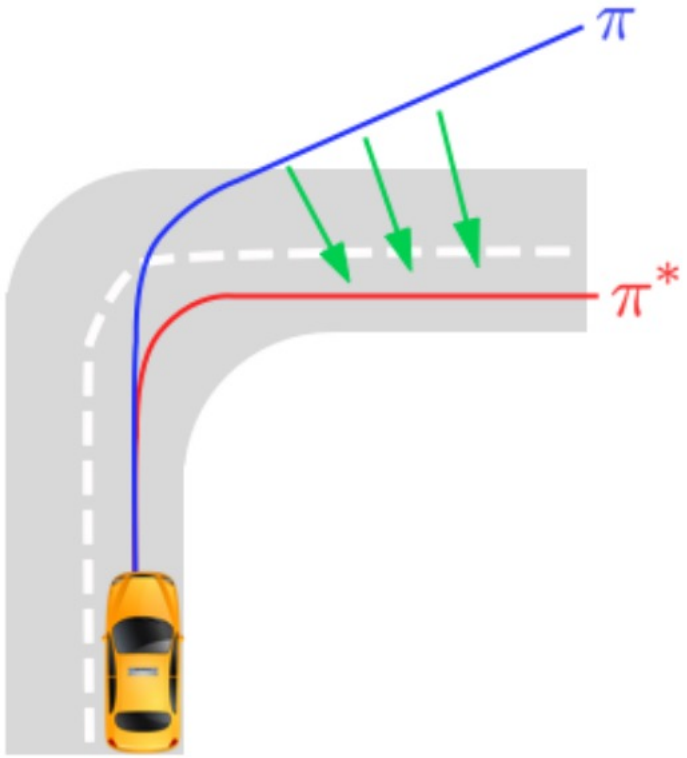
Train policy (e.g. with Behavioral Cloning)



Autonomous vehicle

Can you spot the problem?

“Easy” problem: support of training data



- **Training distribution does not have complete support**
  - Expert only performs “reasonable” actions
  - Data does not contain states only reached through unreasonable actions
- **If imitating agent deviates from support, behaviour becomes undefined**
  - Small mistakes accumulate
- **”Harmless” problem: goes away with infinite data**
  - ...when expert is perfectly trained
  - Can also fix with noise injection or expert queries (Dagger)

Image source: Attia et al, “Global overview of Imitation Learning”, 1801.06503

Hard problem: confounding

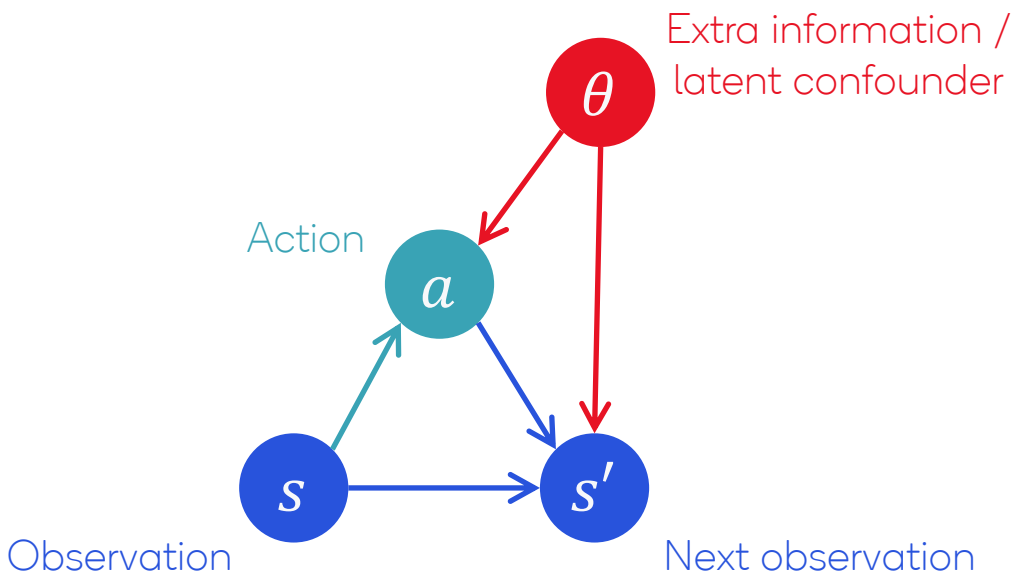
- **Expert has access to more information than imitator**
  - Human driver listens to weather forecast, or hears approaching sirens …
- **Extra information influences decisions**
  - Human expert will drive slower depending on extra information
- **Imitating agent won't be able to imitate properly**
  - Autonomous vehicle cannot sensibly choose whether it should go slow or fast
- **Extreme case: causal delusions**
  - Imitating agent can take own past actions as evidence to base next actions on
  - Autonomous vehicle may first randomly decide to drive fast, then continue driving fast because expert speed is also always consistent

Expert

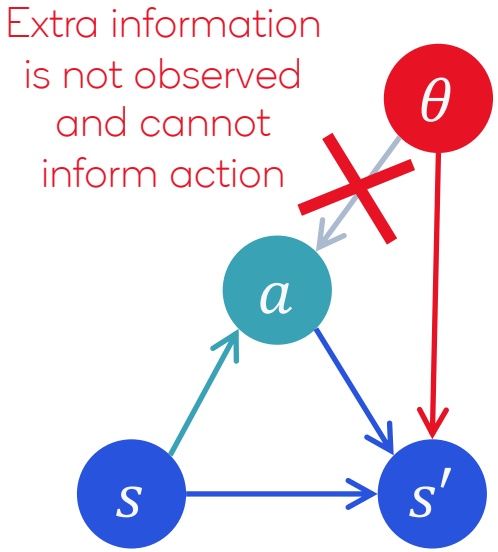
Imitator

Causal perspective

Expert interacting with the environment:



Imitator interacting with the environment:



Causal perspective: The expert and imitator settings differ by an intervention

Naïve behavioral cloning

An imitator trained with behavioral cloning learns the **conditional policy**

$$\pi_{cond}(a_t \mid s_1, a_1, \ldots, s_t) = \mathbb{E}_{\theta \sim p_{cond}(\theta|\tau)}\, \pi_{exp}(a_t|s_t, \theta)$$

Averaging over latent belief with Bayes' rule:

$$p_{cond}(\theta|\tau) \propto p(\theta) \prod_t p(s_{t+1}|s_t, a_t, \theta)\pi_{exp}(a_t|s_t, \theta)$$

This uses both **past environment transitions** and **past actions** as evidence for the latent

But: Using past actions as evidence is **wrong** when past actions were chosen by agent!

The conditional policy can thus **fail to imitate the expert, even with infinite data!**

Interventional policy

Instead, to get correct imitator behavior we need to learn the **interventional policy**

$$\pi_{int}(a_t \mid s_1, a_1, \ldots, s_t) = \mathbb{E}_{\theta \sim p_{int}(\theta|\tau)}\, \pi_{exp}(a_t|s_t, \theta)$$

Still averaging over latent belief, but now as

$$p_{int}(\theta|\tau) \propto p(\theta) \prod_t p(s_{t+1}|s_t, a_t, \theta) \cancel{\pi_{exp}(a_t|s_t, \theta)}$$

Only **past environment transitions** are used as evidence for the latent

We show (not in this talk) that this **interventional policy will converge to the expert's policy** (under some assumptions)

In theory, when can we learn the interventional policy?

Tier 1: Identifiability from demonstrations only

$\tau \sim \pi_{exp}$ 

Tier 2: Identifiability from demonstrations and simulator

$\tau \sim \pi_{exp}$ 

$\tau \sim \pi_{rnd}$ 

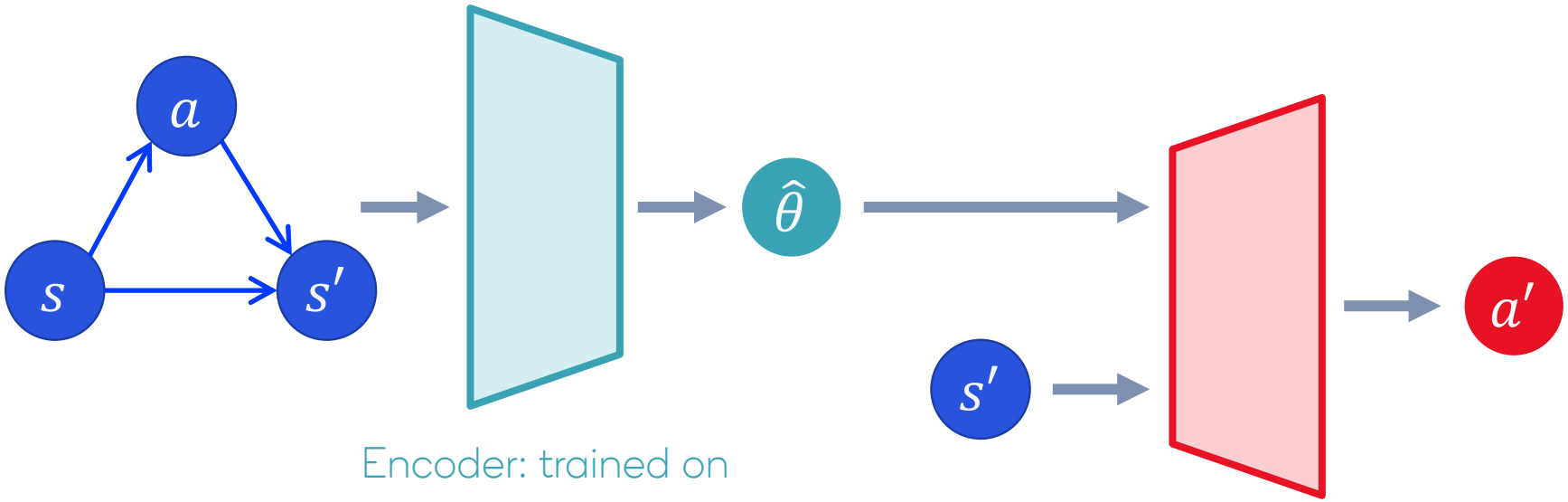
Tier 3: Identifiability from expert queries

$\tau \sim \pi_{exp}$ 

π_{exp}

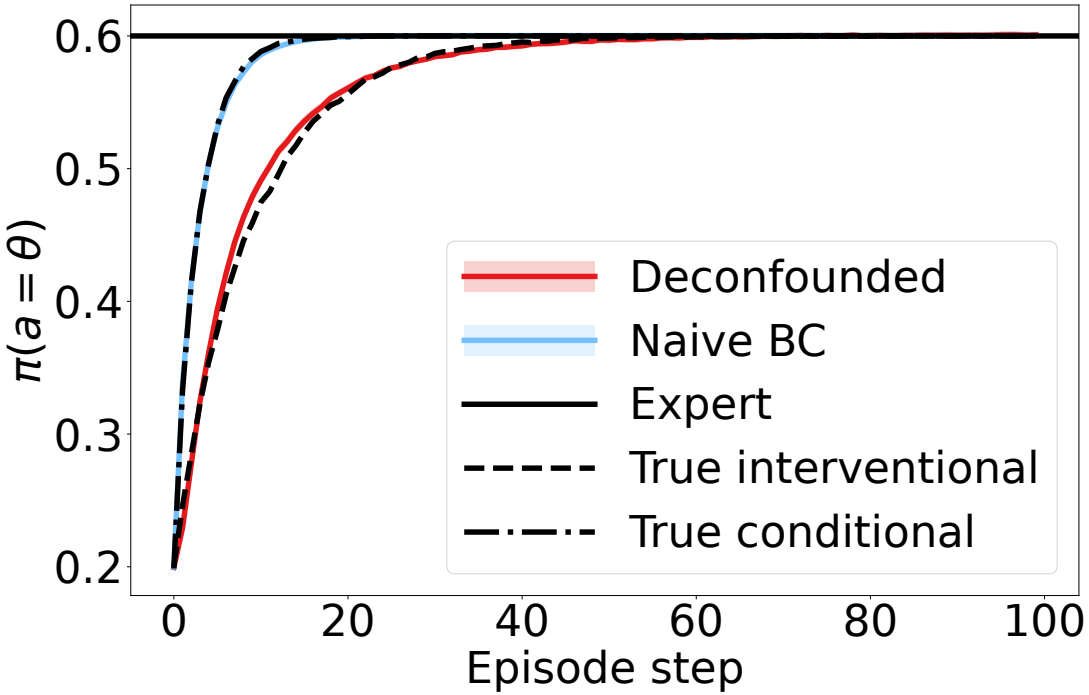
Practical algorithm

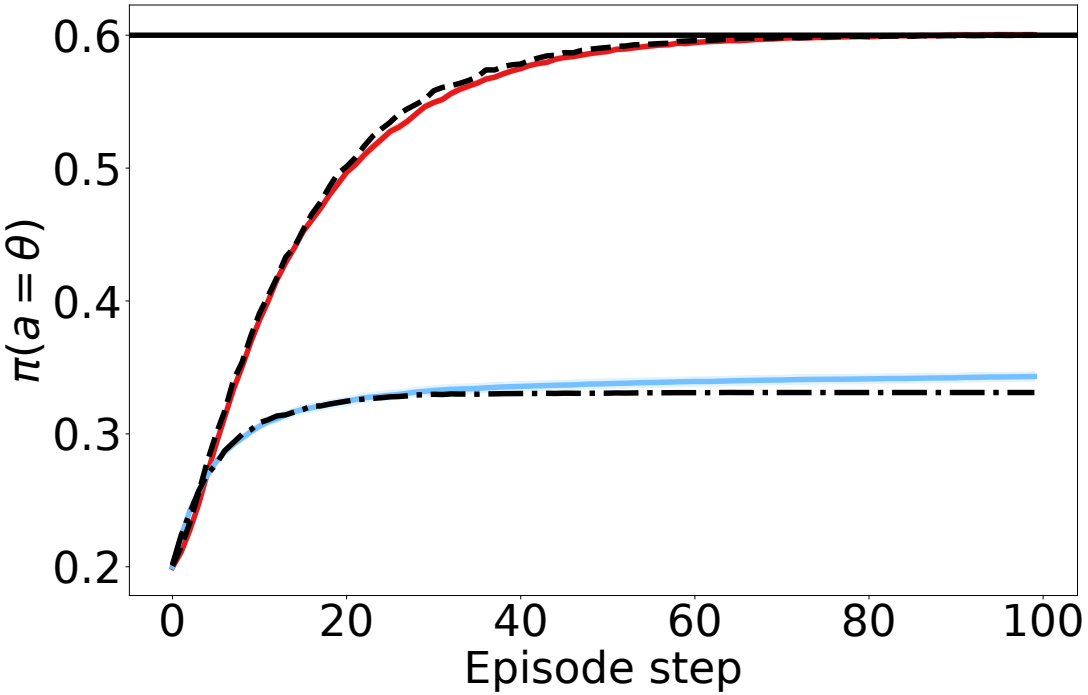
- Learn a **belief over latents** and a **latent-conditional policy**



- Test time: alternate between **updating belief with encoder** and **acting under that belief with policy**

Multi-armed bandit (expert knows most likely arm)

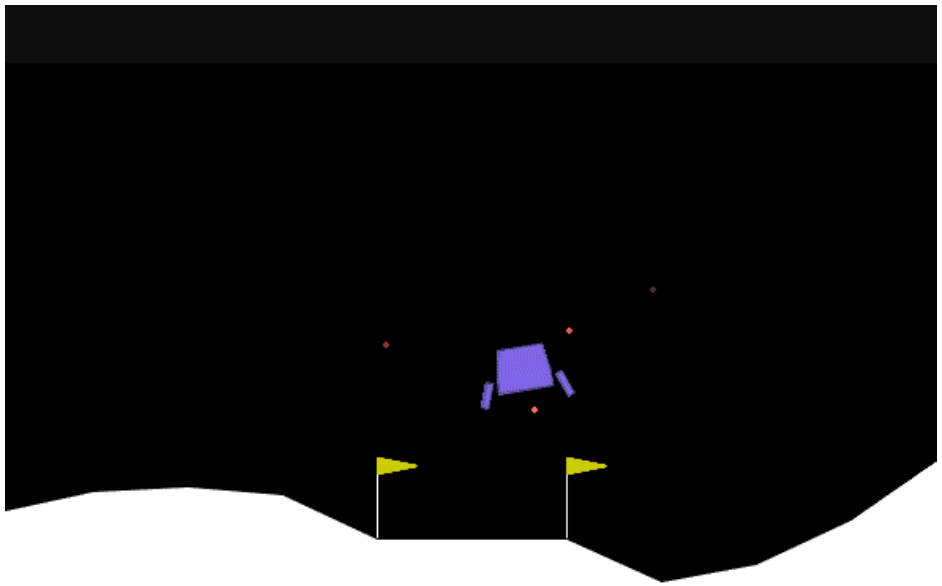


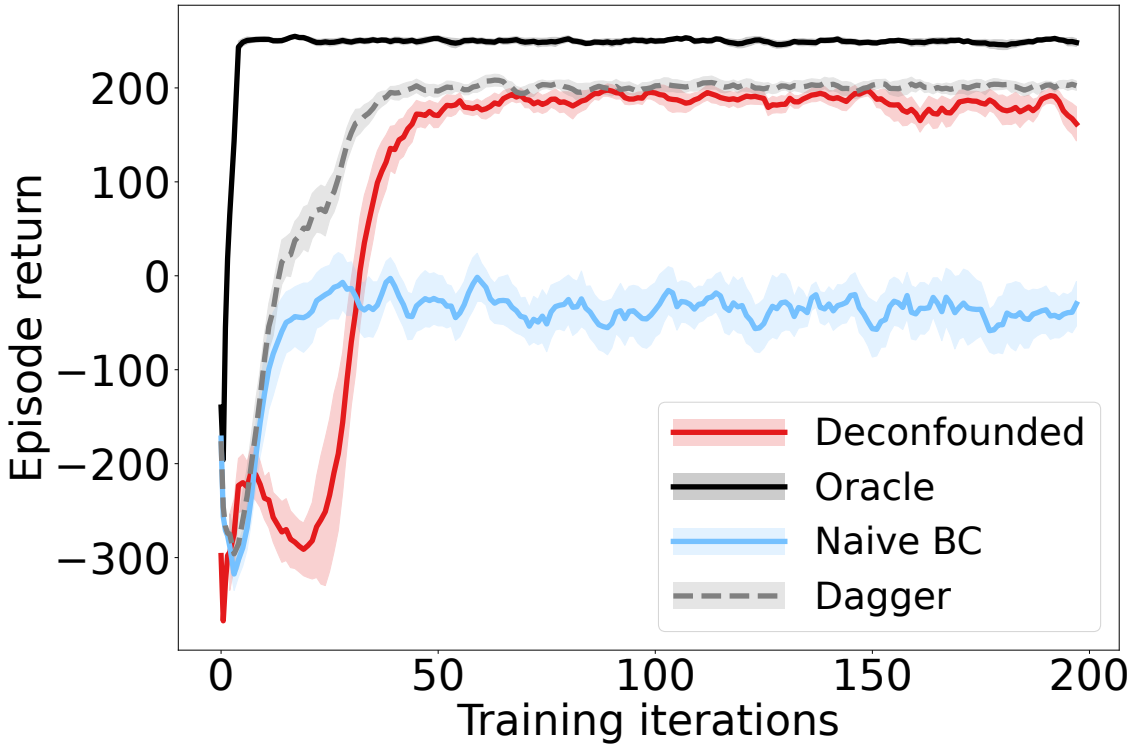


On training data collected by expert:
both naïve BC and our deconfounded algorithm fit data well

Imitator interacting with environment:
naïve BC suffers from confounding,
our deconfounded algorithm is able to infer latents and take them into account

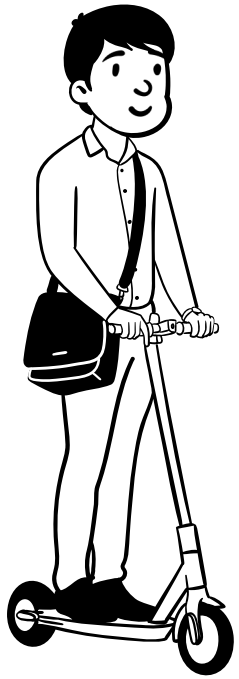
Confounded LunarLander





Environment: LunarLander with random key bindings (known to the expert, but not the imitator)

Naïve BC suffers from confounding, our deconfounded algorithm is able to infer key mapping and take them into account

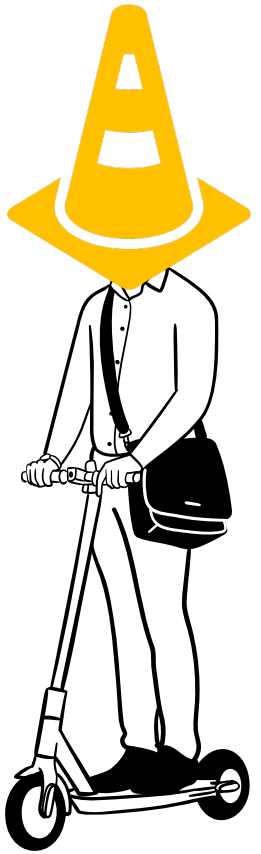


Imitation learning fails in the presence of **information asymmetries**

A **causal perspective** helps to identify issue and solution: need to learn an **interventional, rather than conditional, belief** over latents

We analyze theoretically when the interventional policy can be learned...

...and introduce a **practical algorithm** that continuously updates a belief over latents and acts under it, solving confounding problem



Deconfounded imitation learning

Risto Vuorio, JB, Hanno Ackermann, Daniel Dijkman, Taco Cohen, Pim de Haan

Under review

arXiv: 2211.02667

Risto Vuorio

Pim de Haan

Hanno Ackermann

Daniel Dijkman

Taco Cohen

Shaking the foundations: delusions in sequence models for interaction and control

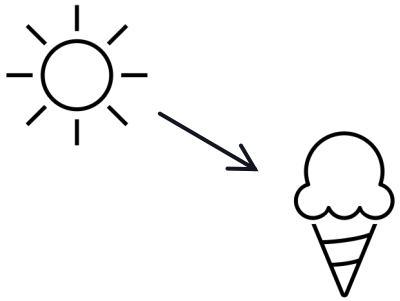
Pedro Ortega, Markus Kunesch, Grégoire Delétang, Tim Genewein, Jordi Grau-Moya, Joel Veness, Jonas Buchli, Jonas Degrave, Bilal Piot, Julien Perolat, Tom Everitt, Corentin Tallec, Emilio Parisotto, Tom Erez, Yutian Chen, Scott Reed, Marcus Hutter, Nando de Freitas, Shane Legg

arXiv:2110.10819

Sequence Model Imitation Learning with Unobserved Contexts

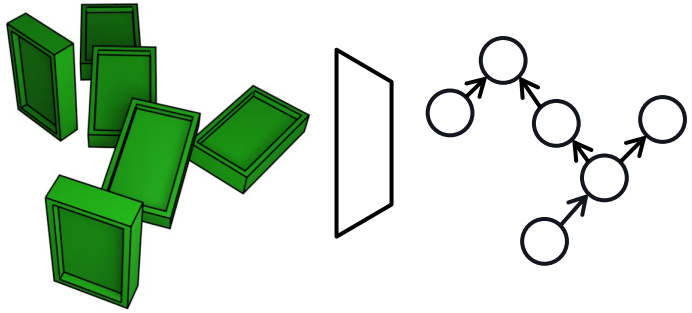
Gokul Swamy, Sanjiban Choudhury, Andrew Bagnell, Zhiwei Steven Wu

NeurIPS 2022, arXiv:2208.02225



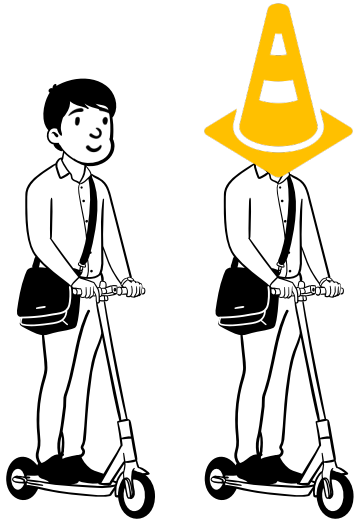
Intro:

Causality is a framework for reasoning about changes



ML for causality:

From non-iid data, we can learn the causal structure of the world



Causality for ML:

With a causal approach,
we can train agents to imitate experts even when information is asymmetric

11:00	**Intro to Qualcomm AI Research**	Johann Brehmer
11:30	**Causality for ML and ML for causality**	Johann Brehmer
12:15	Break	
13:00	**(Adaptive) neural compression**	Ties van Rozendaal
13:45	Break	
14:00	**Perceptual quality in neural compression**	Jens Petersen

Thank you

Qualcomm

Follow us on:

For more information, visit us at:

qualcomm.com & qualcomm.com/blog