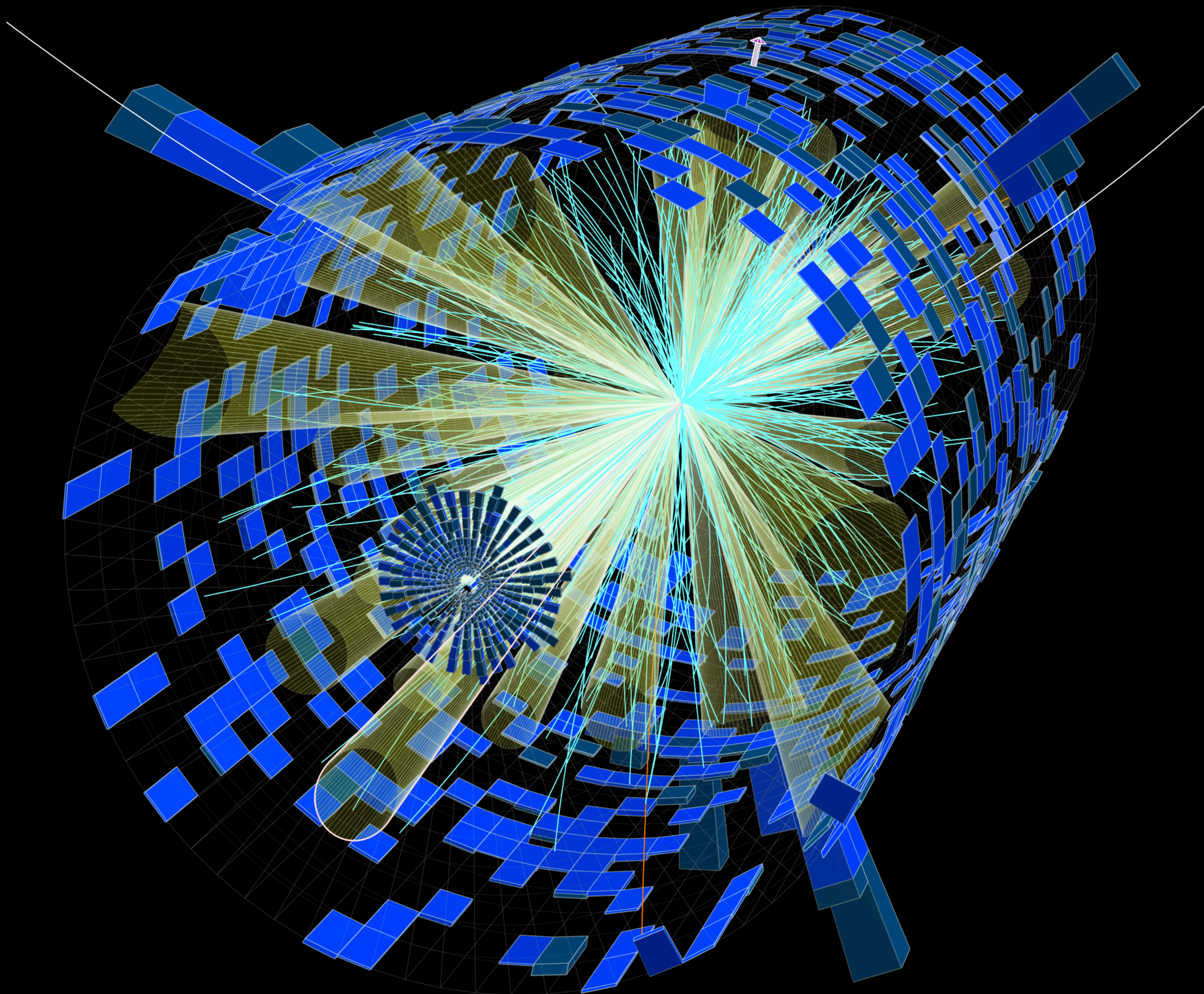
NYU Center for Data Science

Center for Cosmology and Particle Physics

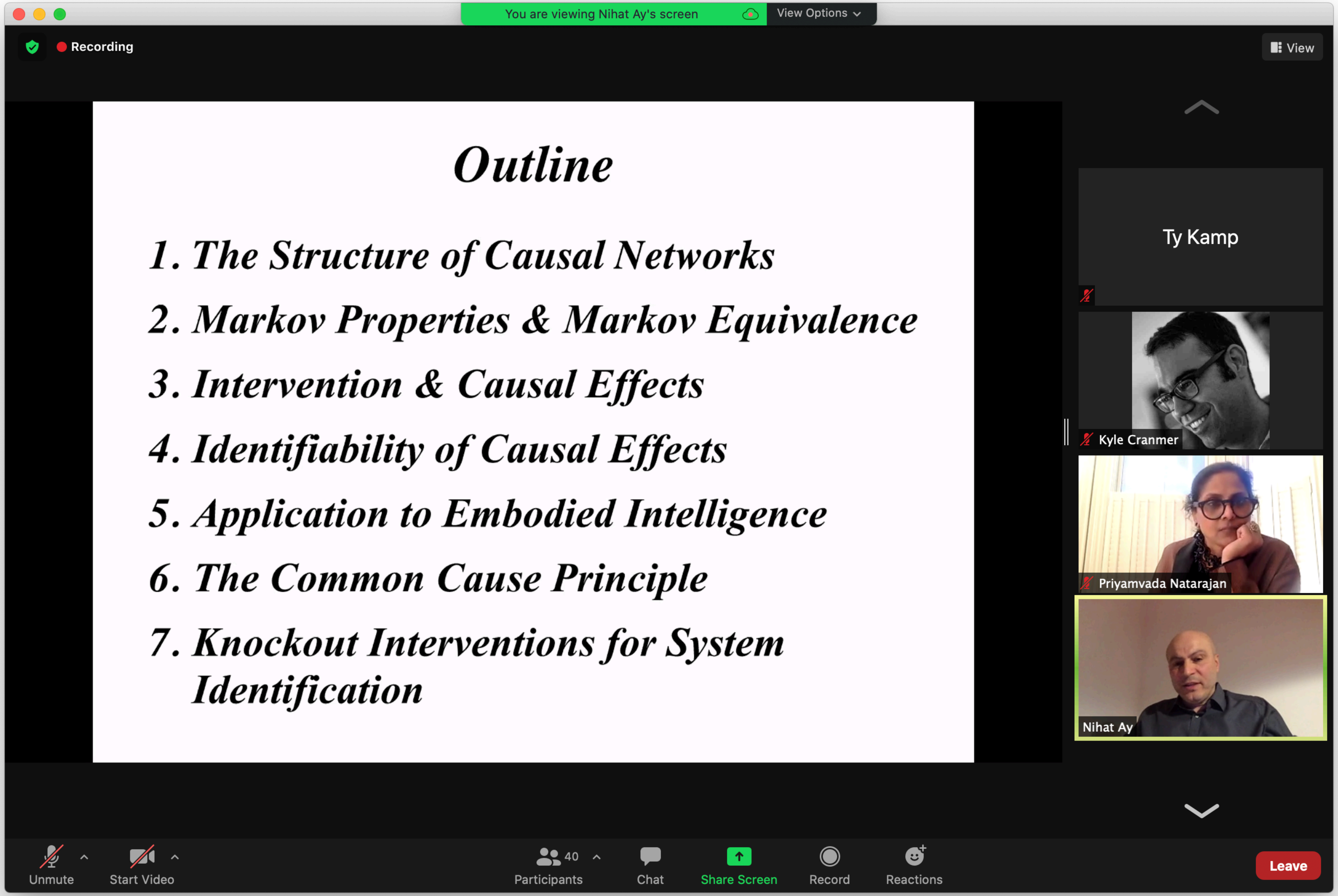
A CONVERSATION WITH

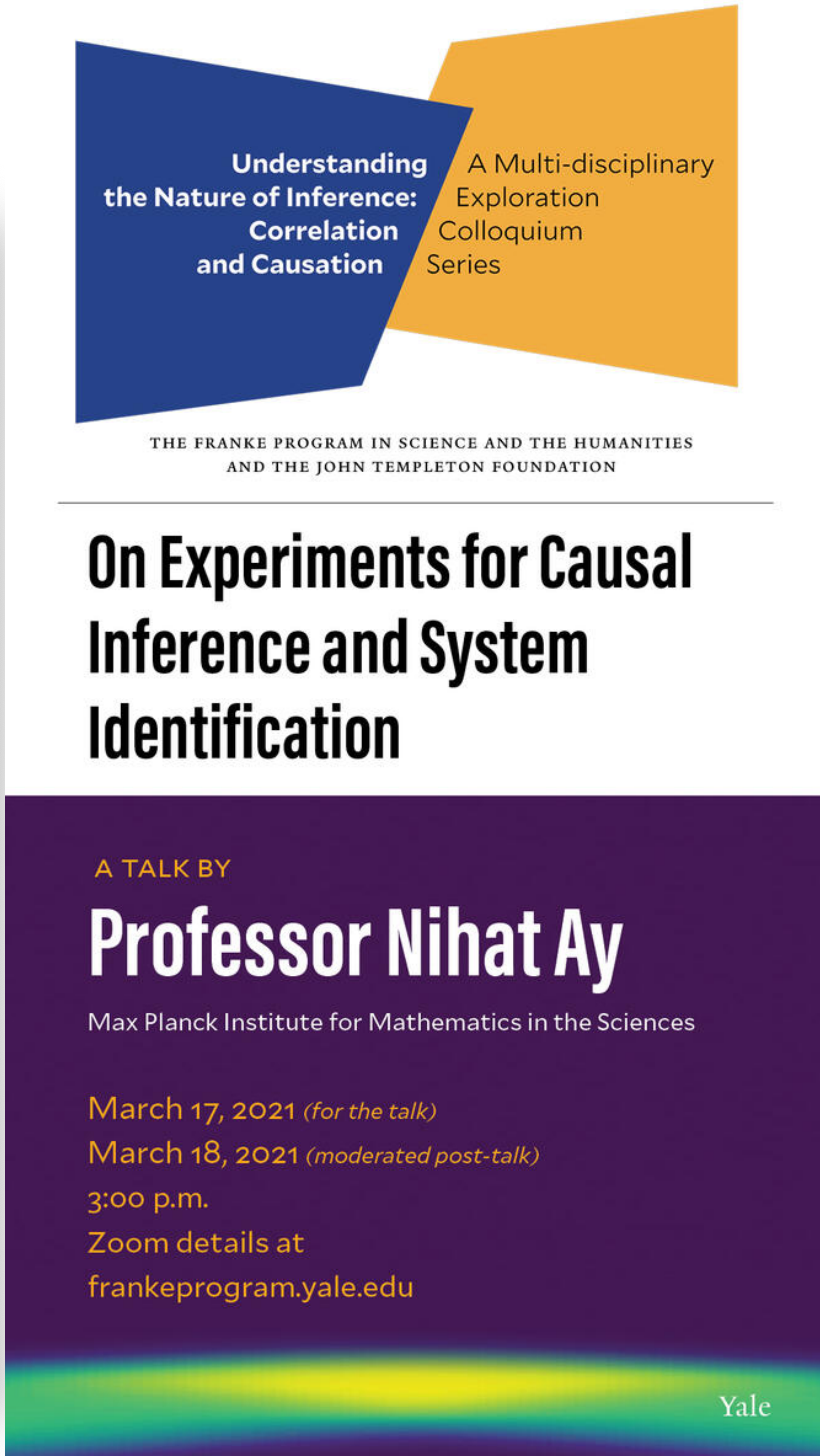
NIHAT AY

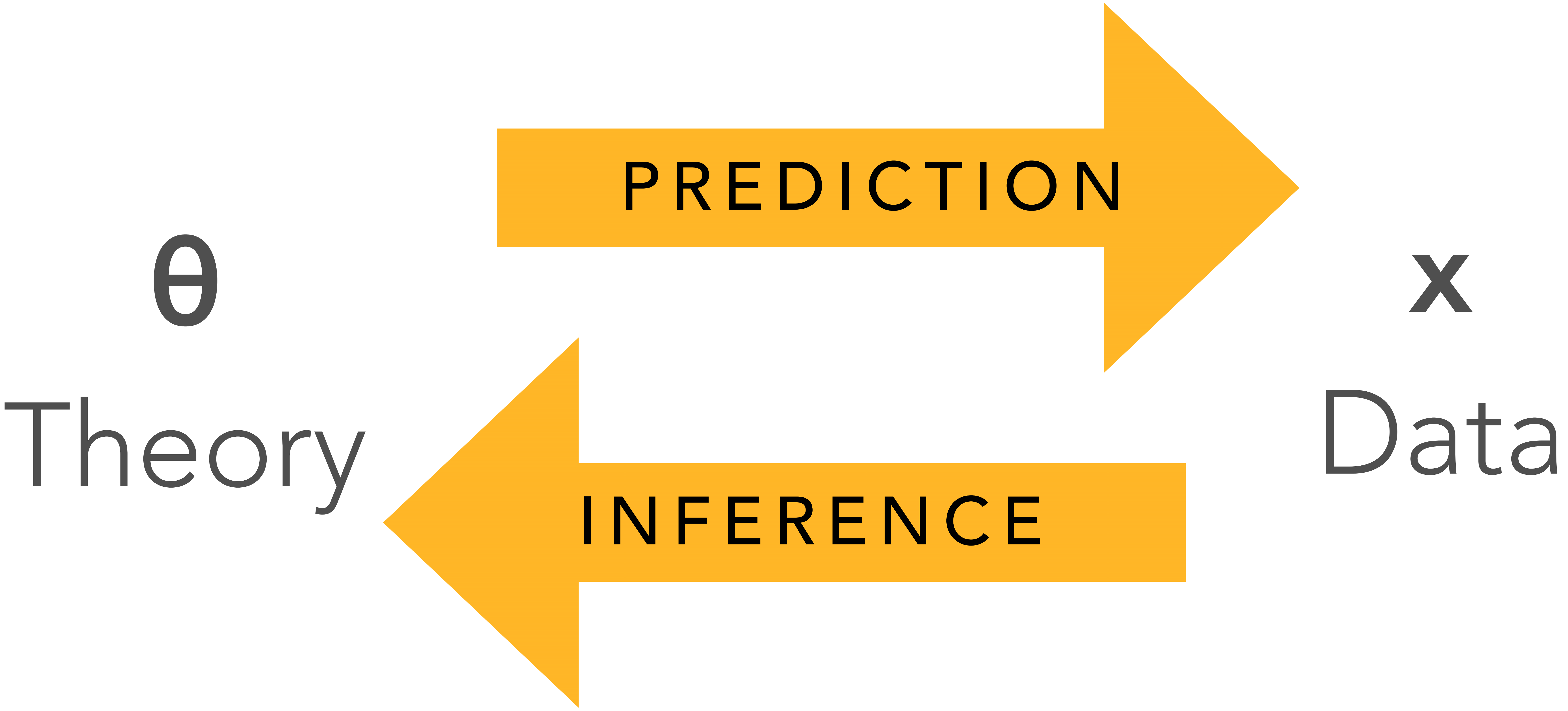


@KyleCranmer
New York University
Department of Physics
Center for Data Science
CILVR Lab

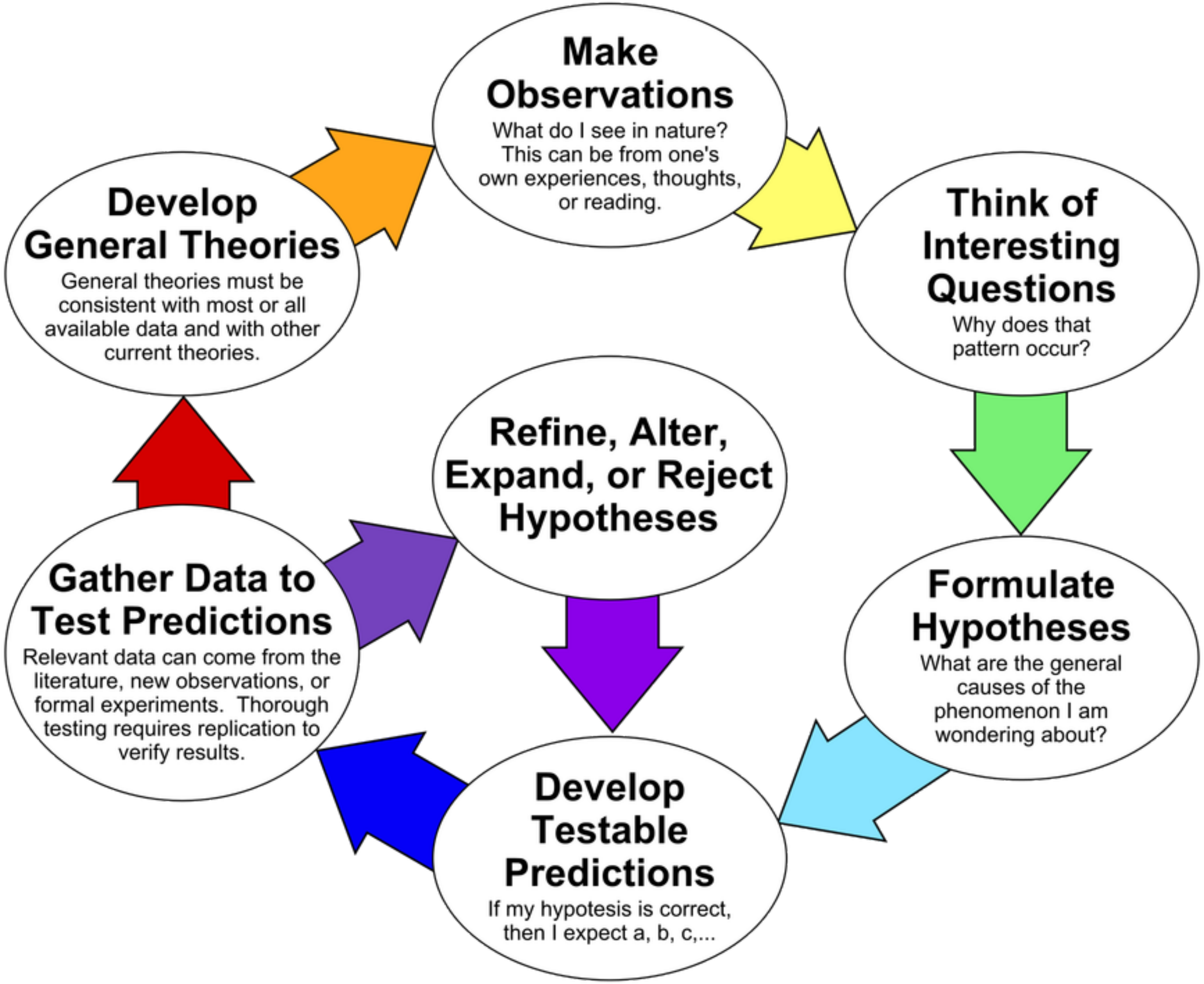
Yesterday's talk





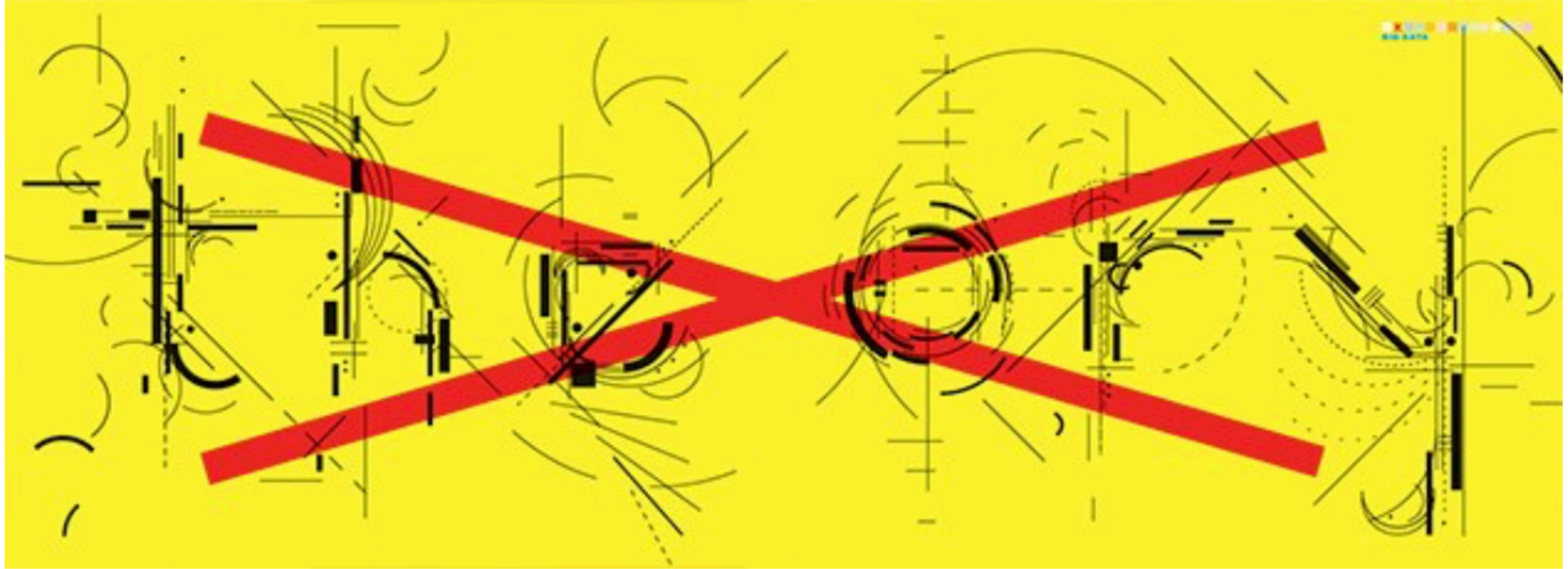

PREDICTION
INFERENCE
θ
Theory
x
Data

The Scientific Method as an Ongoing Process



CHRIS ANDERSON SCIENCE 06.23.08 12:00 PM

THE END OF THEORY: THE DATA DELUGE MAKES THE SCIENTIFIC METHOD OBSOLETE



All models are wrong, and increasingly you can succeed without them.

— Peter Norvig

The ladder of Causation

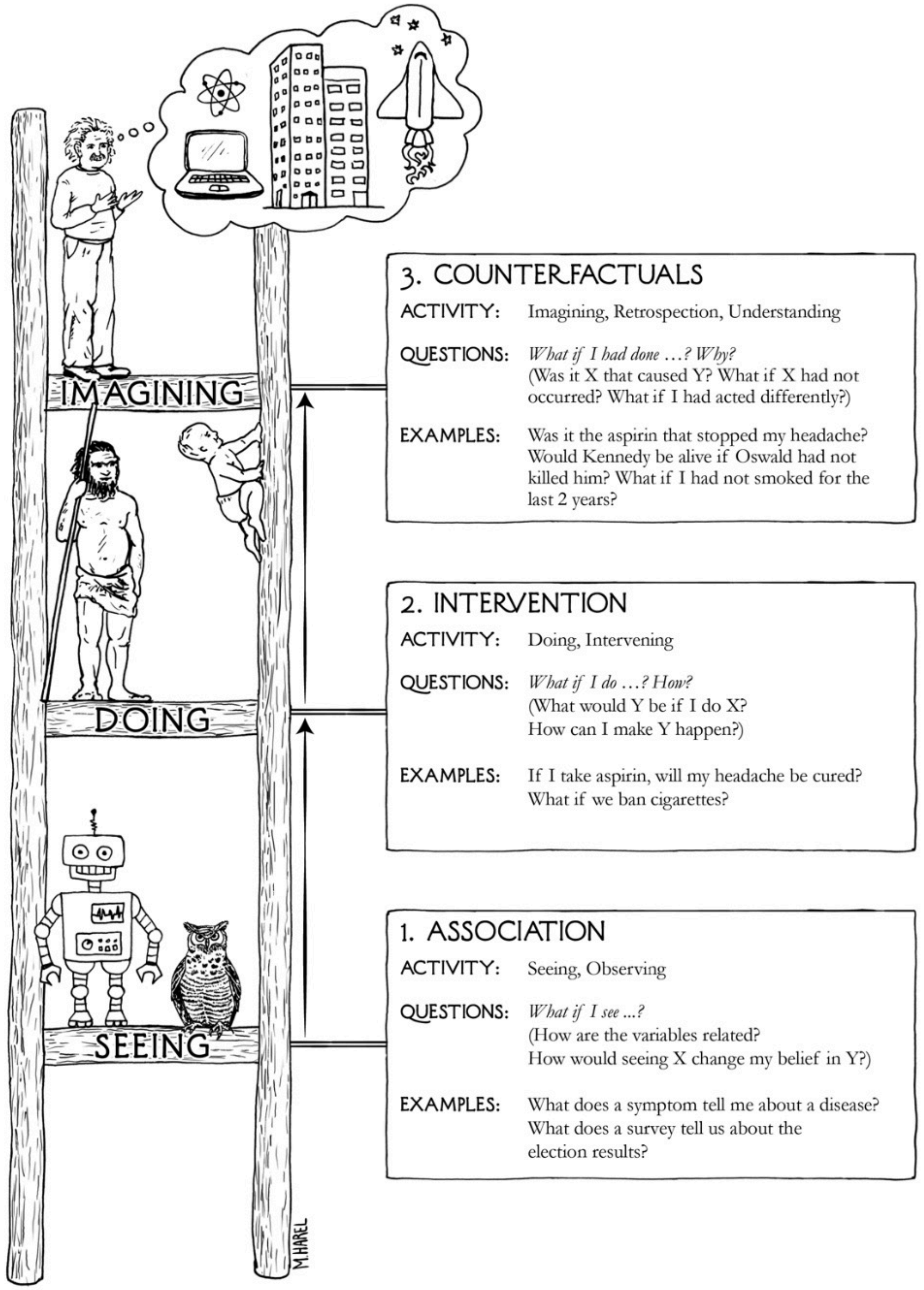


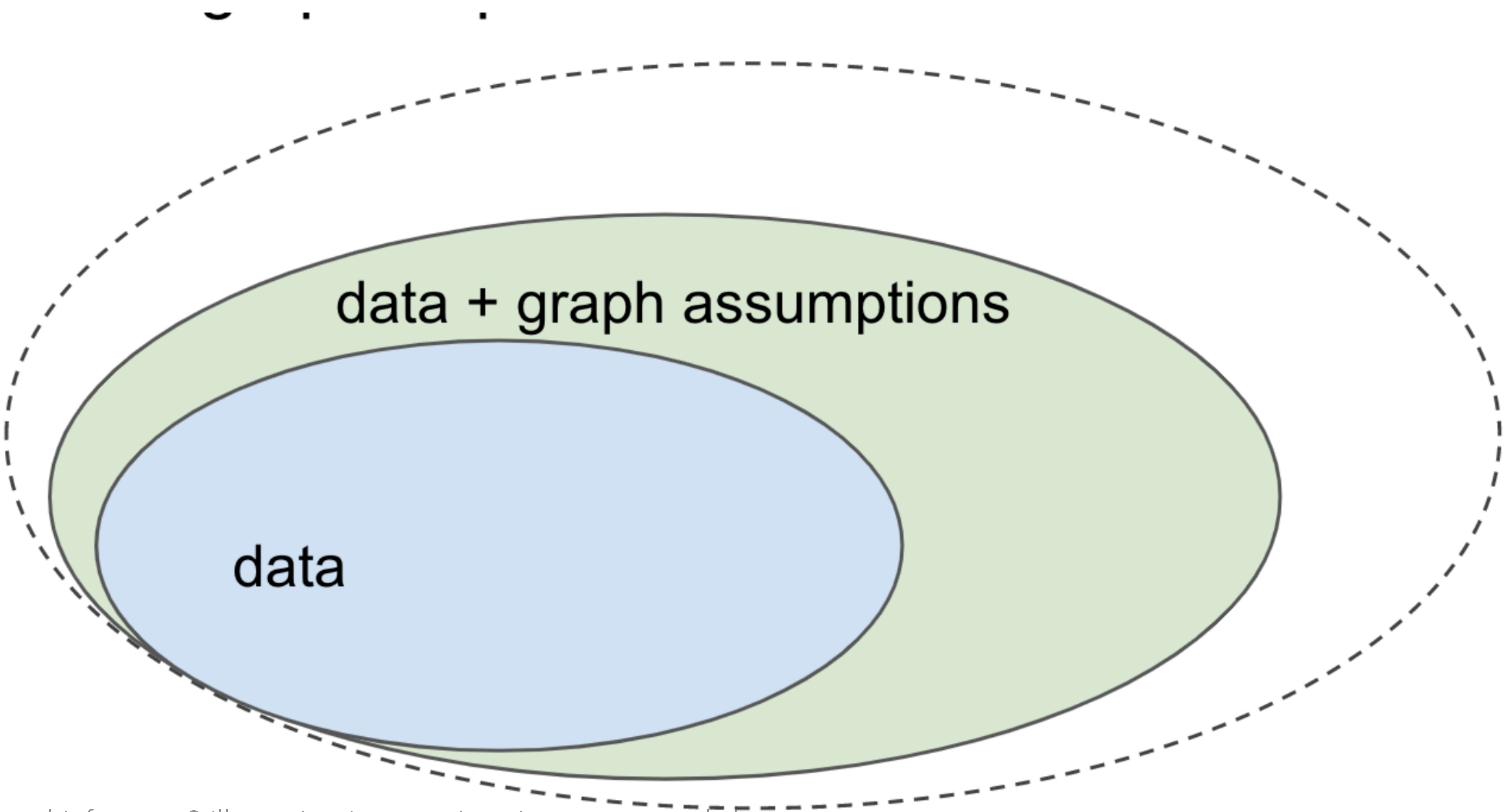
FIGURE 1.2. The Ladder of Causation, with representative organisms at each level. Most animals, as well as present-day learning machines, are on the first rung, learning from association. Tool users, such as early humans, are on the second rung if they act by planning and not merely by imitation. We can also use experiments to learn the effects of interventions, and presumably this is how babies acquire much of their causal knowledge. Counterfactual learners, on the top rung, can imagine worlds that do not exist and infer reasons for observed phenomena. (*Source:* Drawing by Maayan Harel.)

Ferenc Huszár

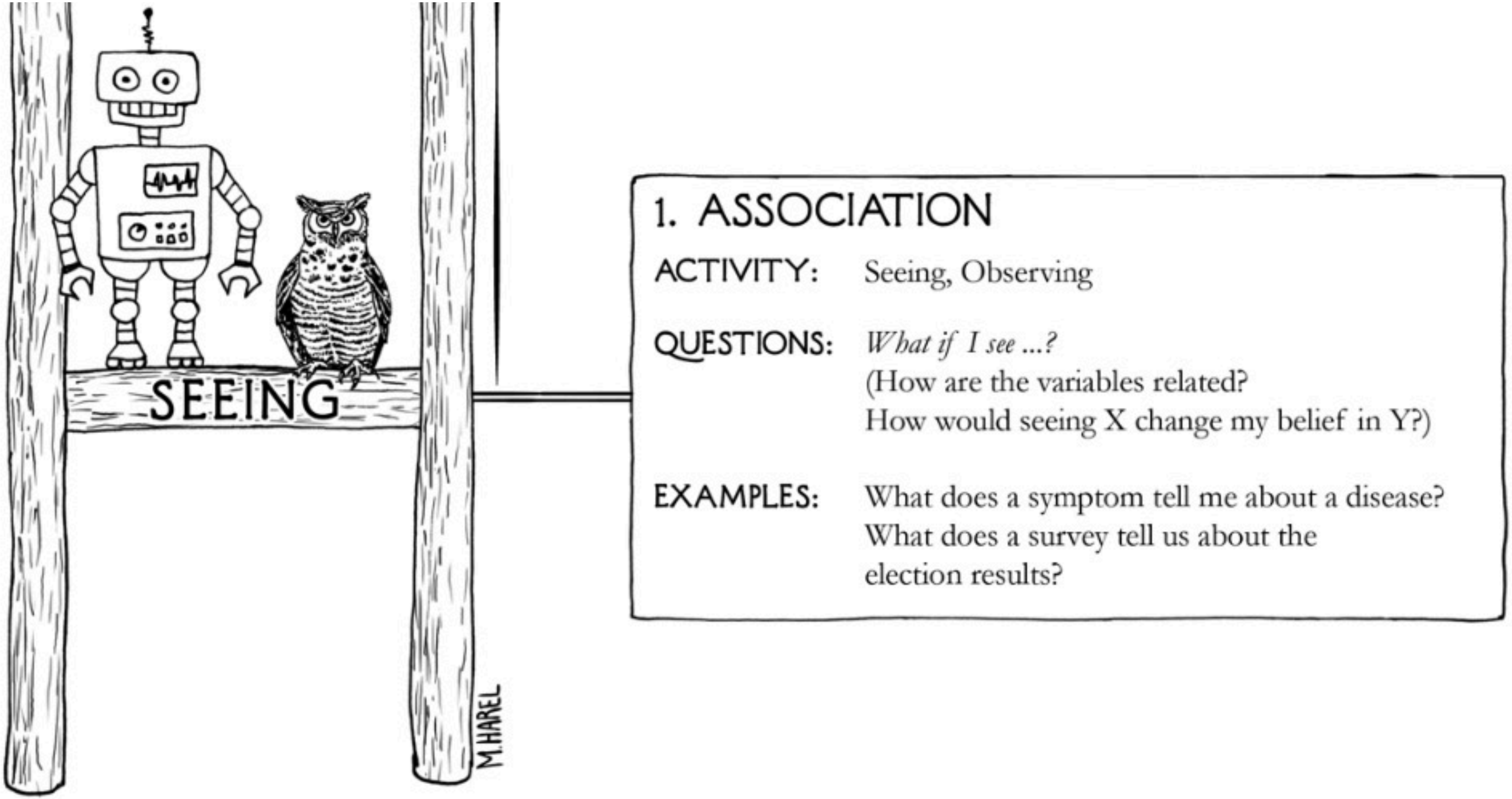
inFERENCe

The moral~~e~~ of the story

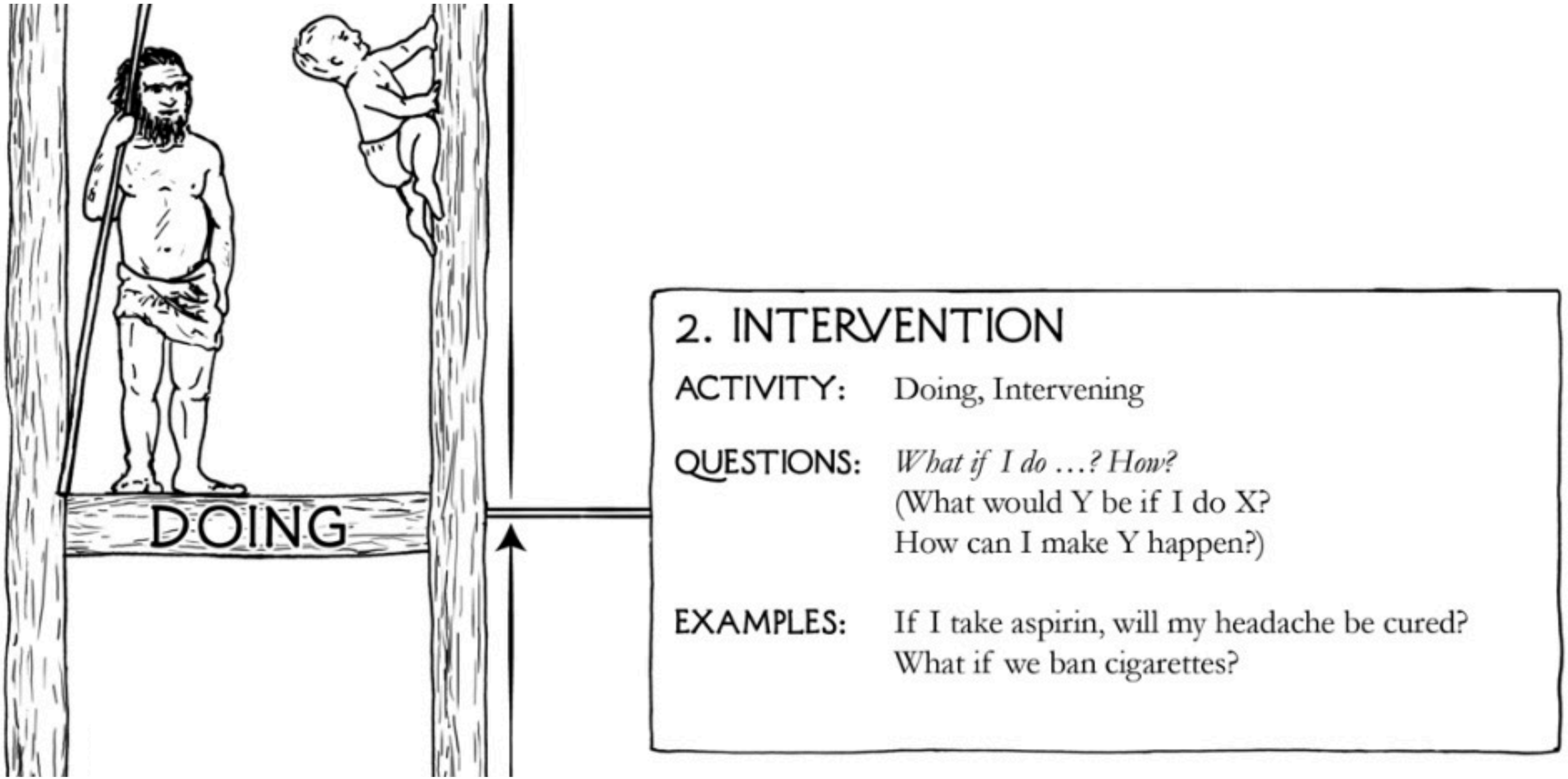
The moral~~e~~ of this story is summed up in the following picture:



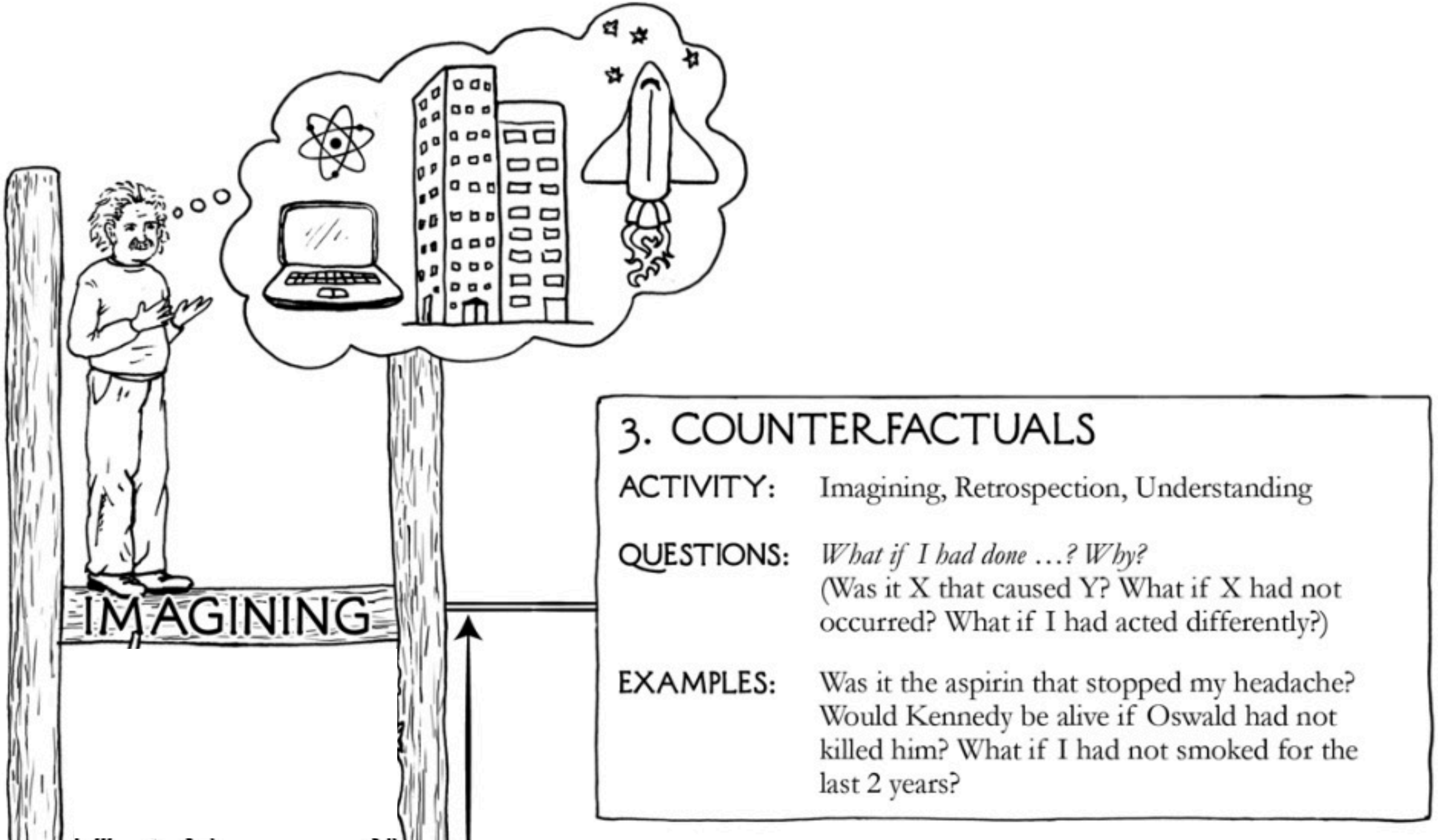
Association

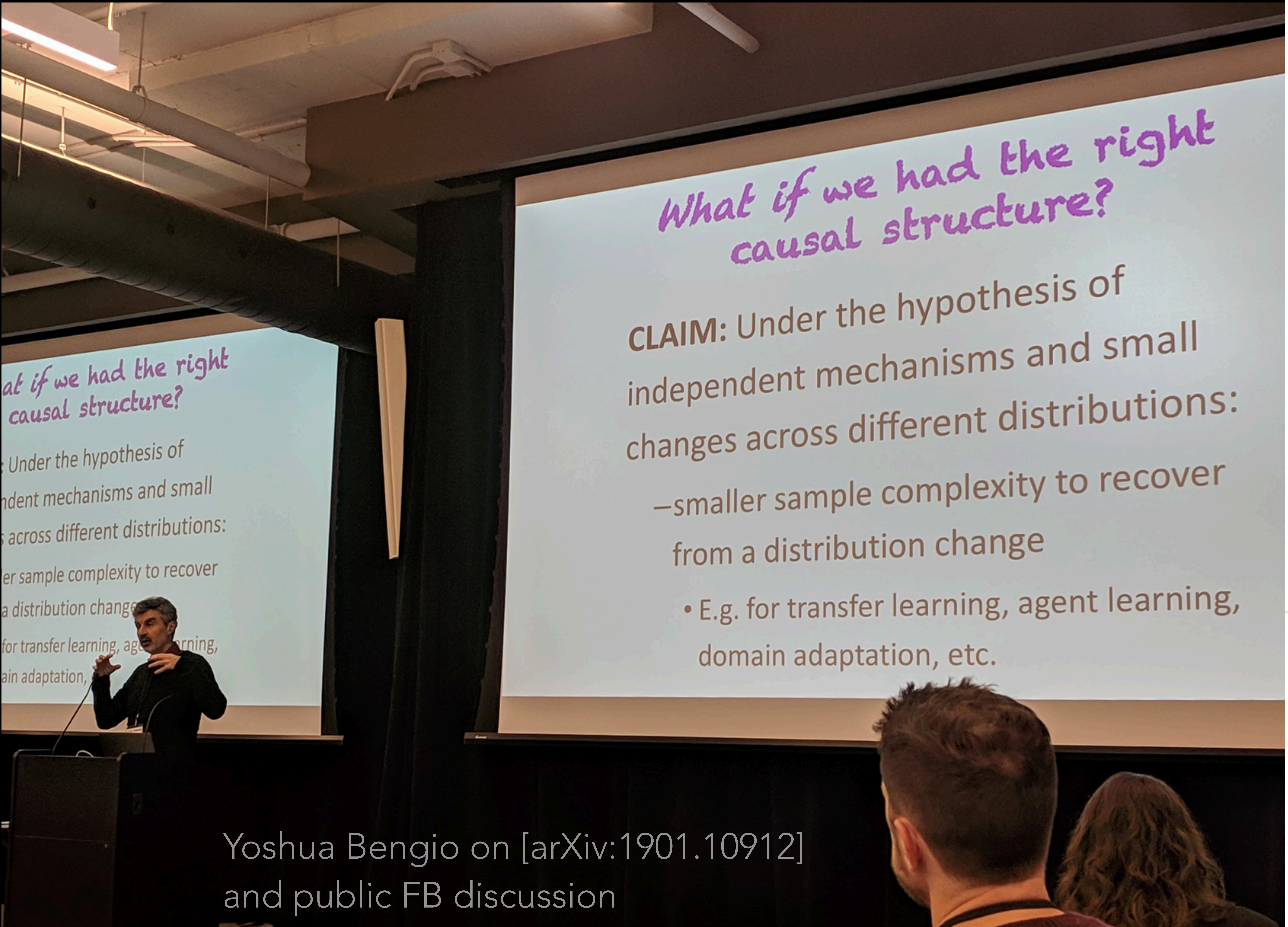


Intervention



Counterfactuals





Yoshua Bengio on [arXiv:1901.10912] and public FB discussion

Max Welling Isn't this what Bernhard Schoelkopf has been saying for a while?

Like · Reply · 6w

Yann LeCun ...and Leon Bottou ?

Like · Reply · 6w

Leon Bottou Yoshua's paper says: if you observe a distribution change that comes from a causal effect, then you'll adapt faster if your generative model matches the causal model.

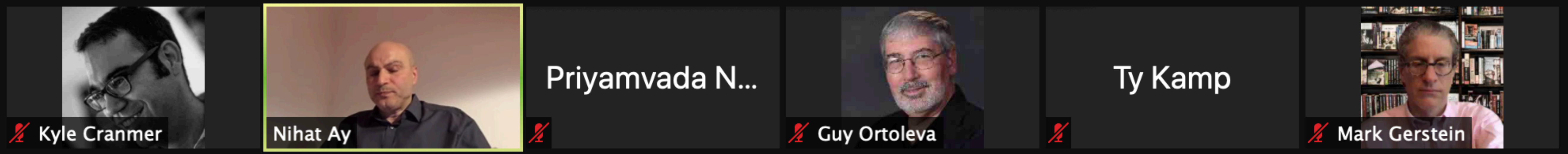
Another way of seeing it is : the right causal graph suggests a particular factorization of the joint distribution (a directed bayesian network). A causal intervention means that you only change one of these factors (or a few factors) while leaving the other ones unchanged. Therefore if your generative model is the right causal model, meaning that it factorizes the joint in the same way, it will be easy to adapt it to the change because only a few parameters need changing (those associated with the factors that actually changed).

Said like this, it feels pretty trivial. Yoshua proposes to use this to infer the right causal model from a plurality of observed distributions.

Dan Roy Max Welling yes. He's been arguing for generative models with causal structure for years as the way to extract information for rich environments. So not this

Max Welling Dan Roy I am, and I think most of us, are keenly aware that Josh has been the big proponent of this view. And I think most people agree with him on this view. Integrating this view with deep learning for more narrowly defined tasks seems to me an interesting intellectual pursuit though. I think that's what's happening here but I was not at the talk 😉

You are viewing Nihat Ay's screen



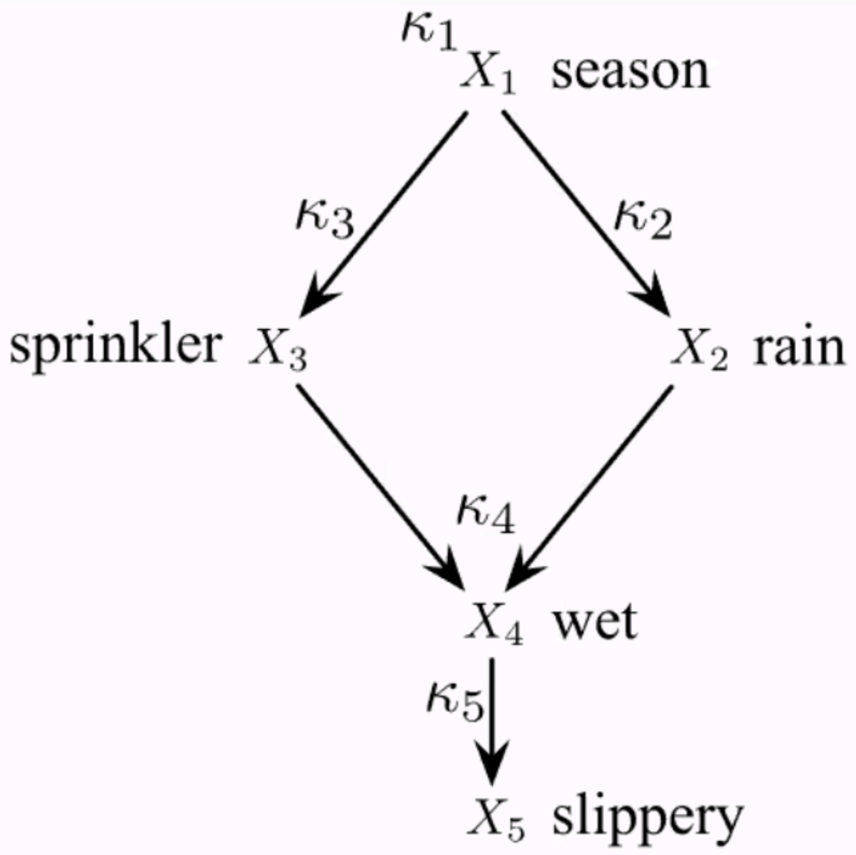
Recording

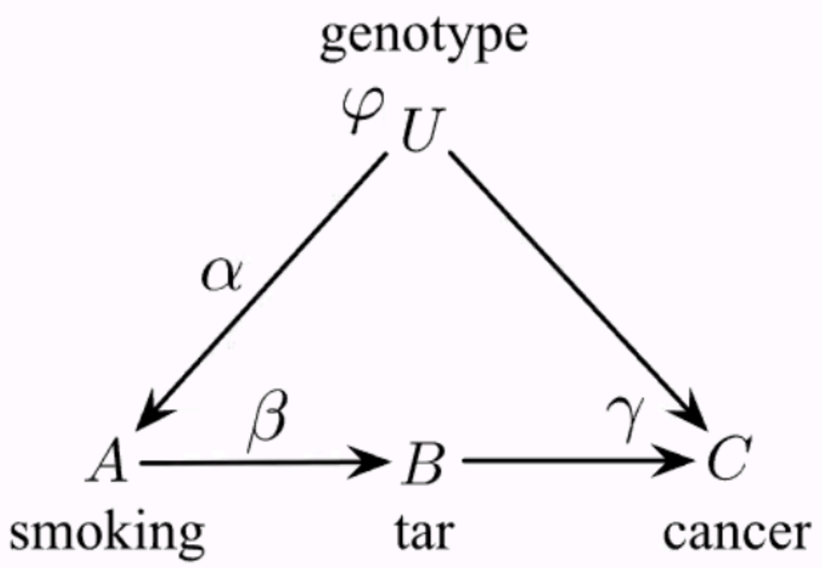
Causally interpreted Bayesian networks

Definition: A DAG $G = (V, E)$ together with a family $\kappa_v : \mathbb{X}_{pa(v)} \times \mathbb{X}_v \to [0, 1]$, $v \in V$, of Markov kernels is called *Bayesian network*. The interpretation of the Markov kernels κ_v as mechanisms of the nodes implies a causal nature of the Bayesian network.

Consider two distinct nodes $v, w \in V$.

- If $v \to w$ we call v a (*pot.*) *direct cause* of w and w a (*pot.*) *direct effect* of v.
- If $v \rightsquigarrow w$ we call v a (*pot.*) *cause* of w and w an (*pot.*) *effect* of v.

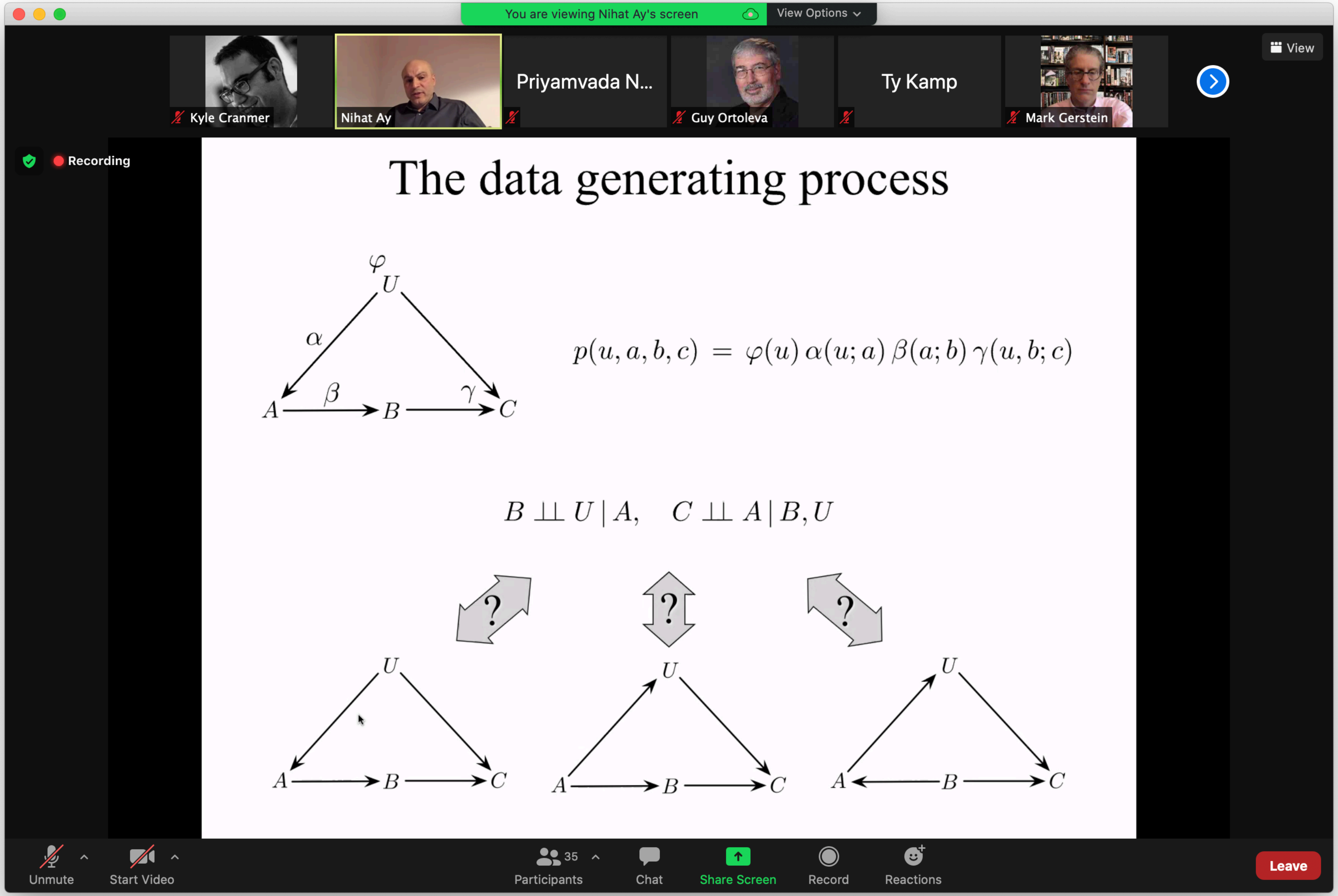




• J. Pearl. ***Causality: Models, Reasoning, and Inference***. Cambridge University Press 2000, 2009.

Unmute Start Video 38 Participants Chat Share Screen Record Reactions Leave

Markov Equivalence



A toy example

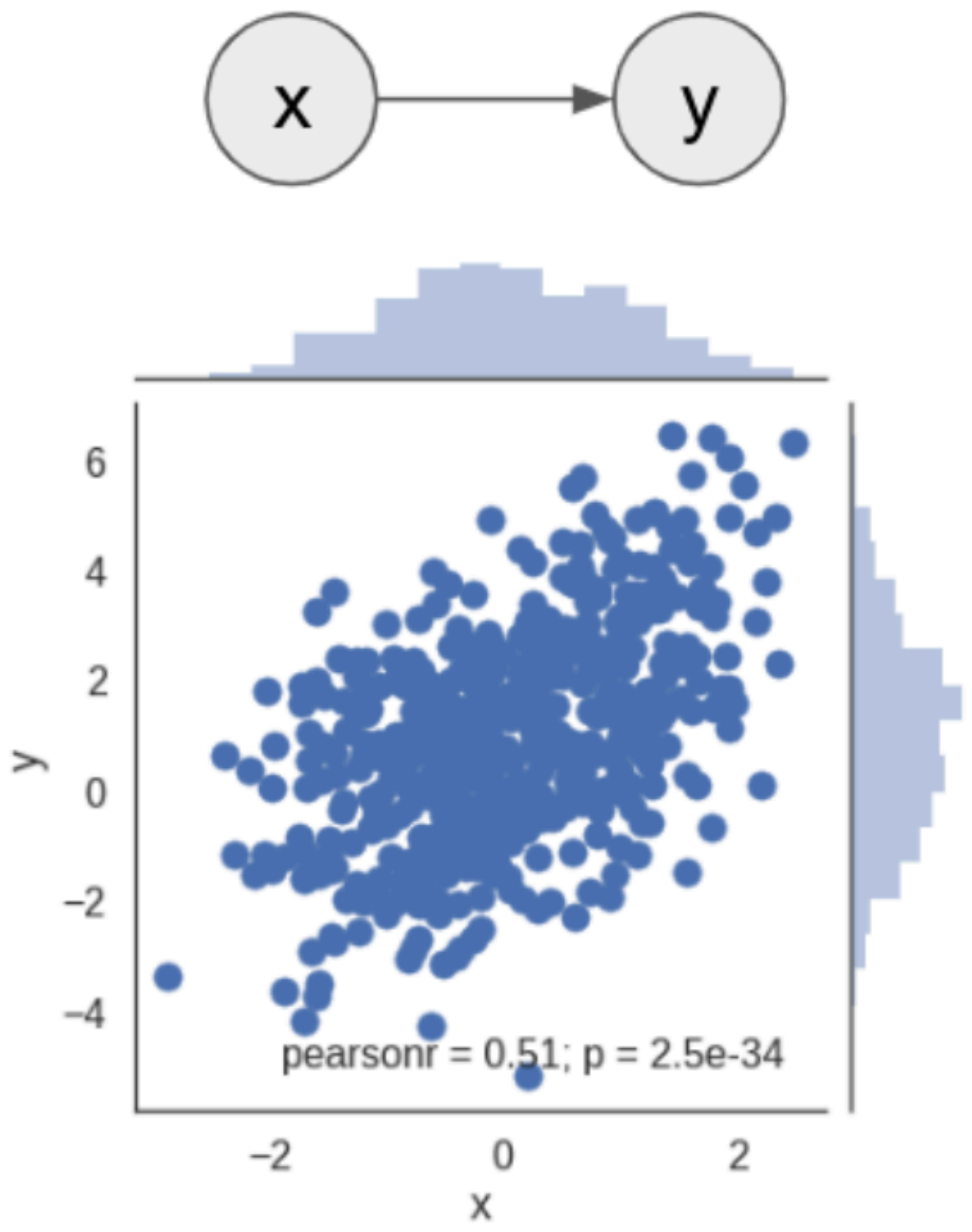
Ferenc Huszár

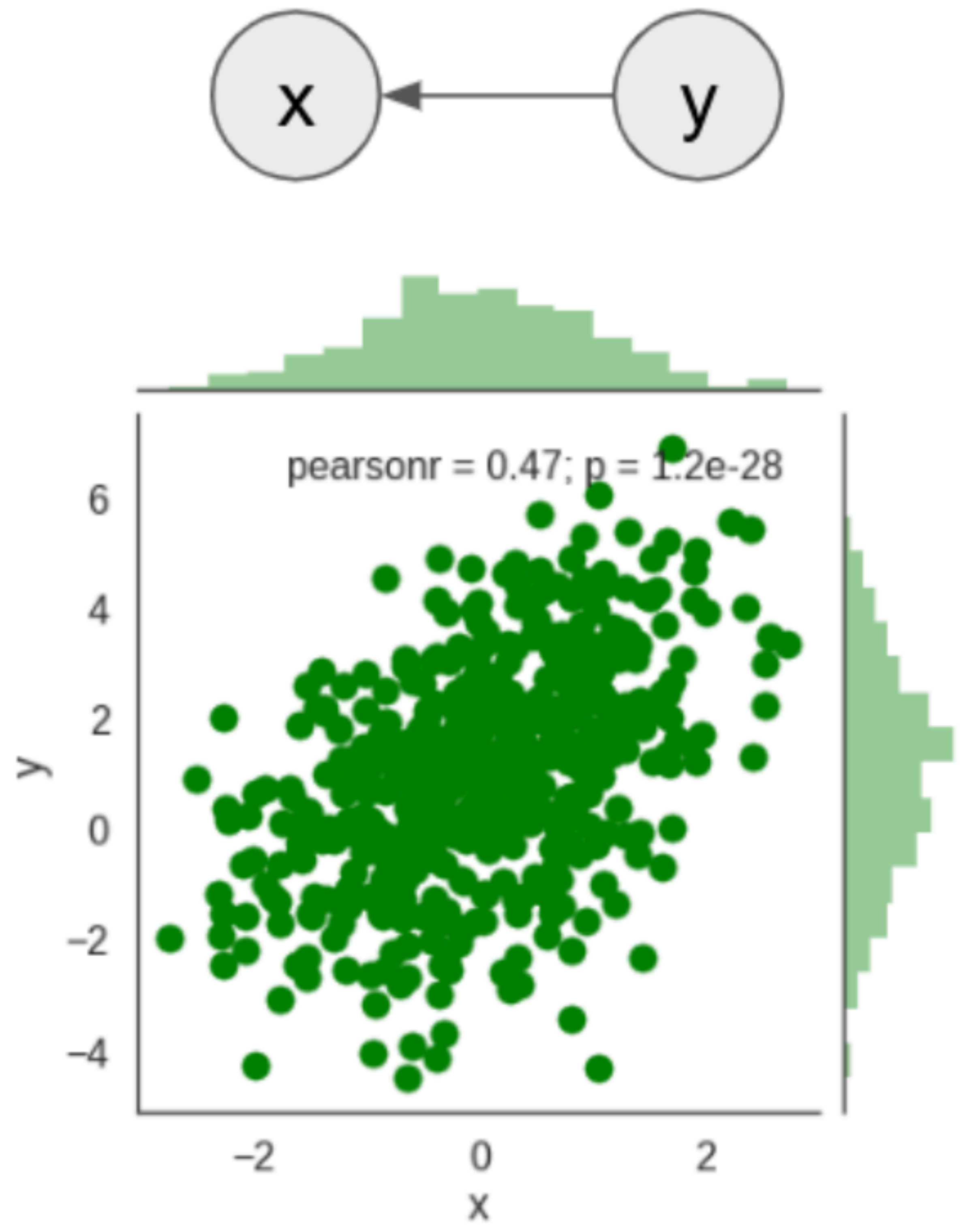
inFERENCe

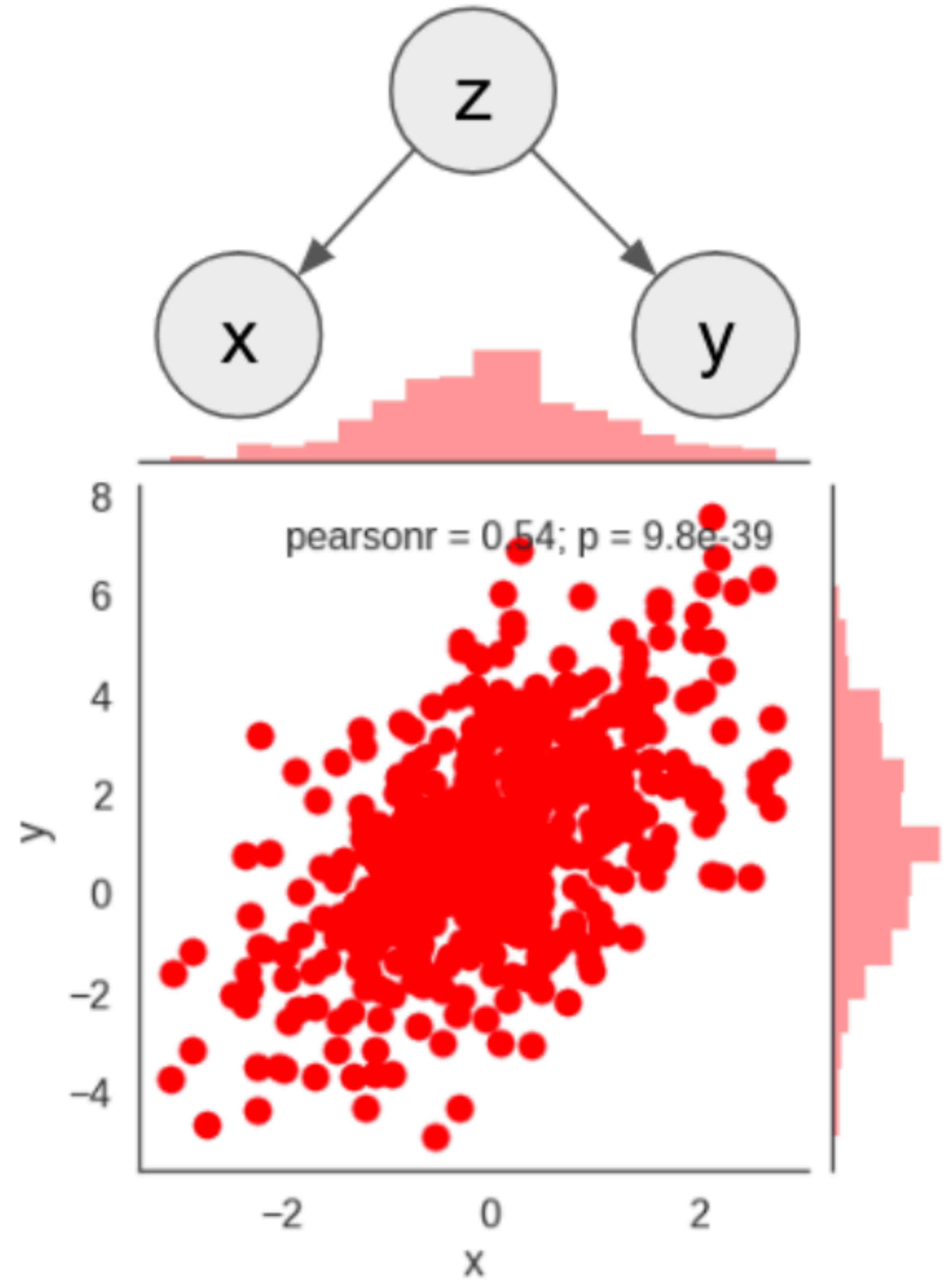
```
x = randn()
y = x + 1 + sqrt(3)*randn()
```

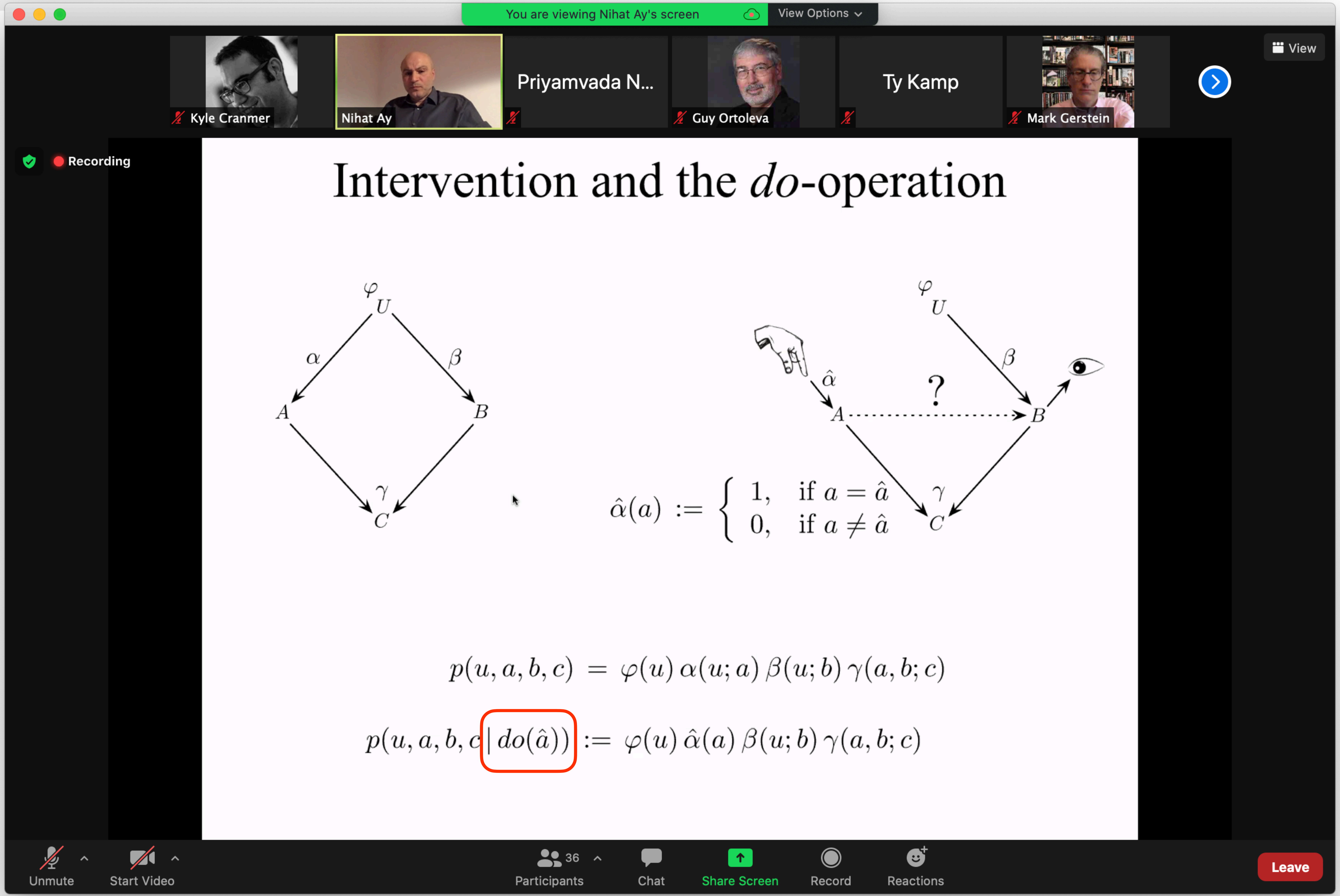
```
y = 1 + 2*randn()
x = (y-1)/4 + sqrt(3)*randn()/2
```

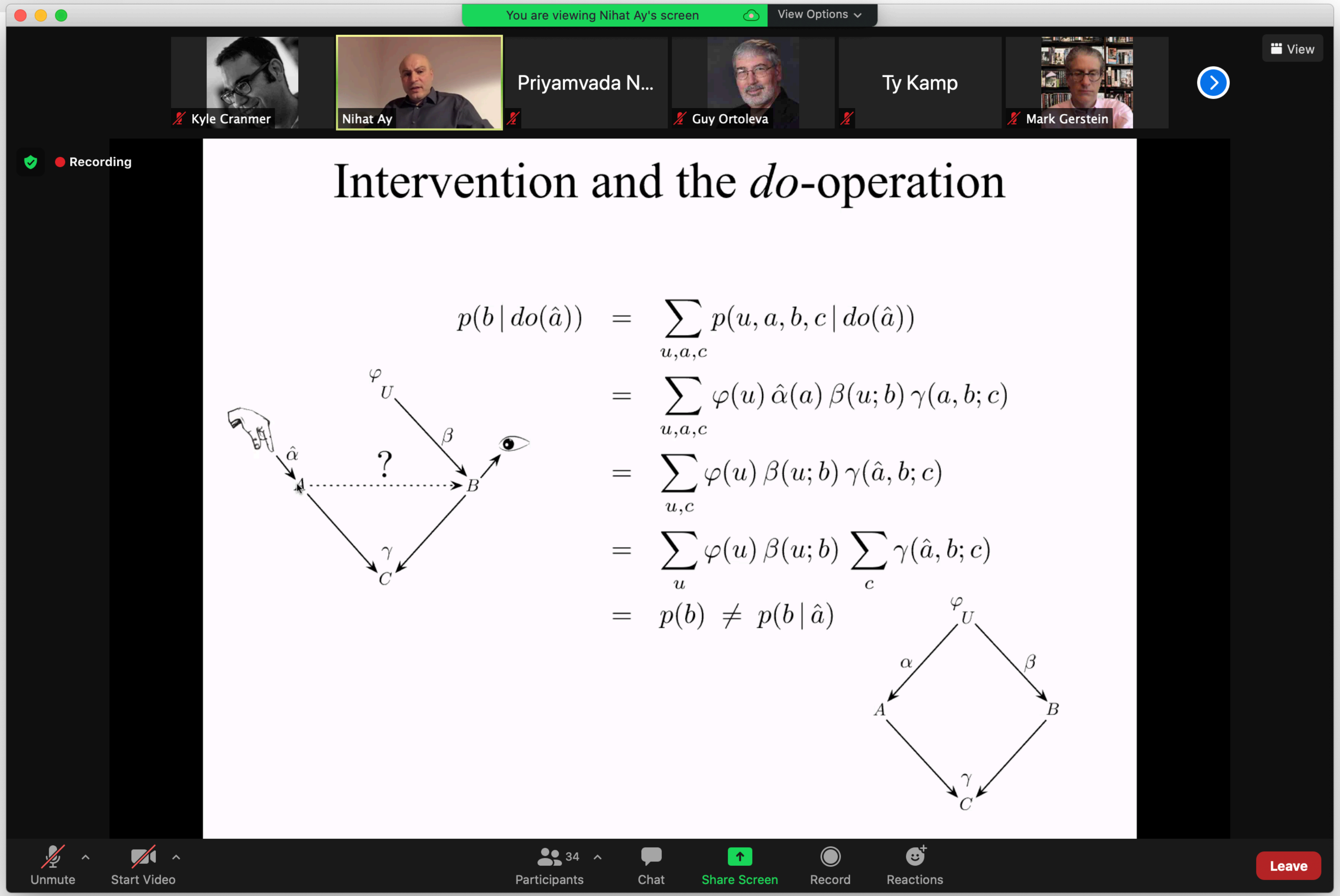
```
z = randn()
y = z + 1 + sqrt(3)*randn()
x = z
```

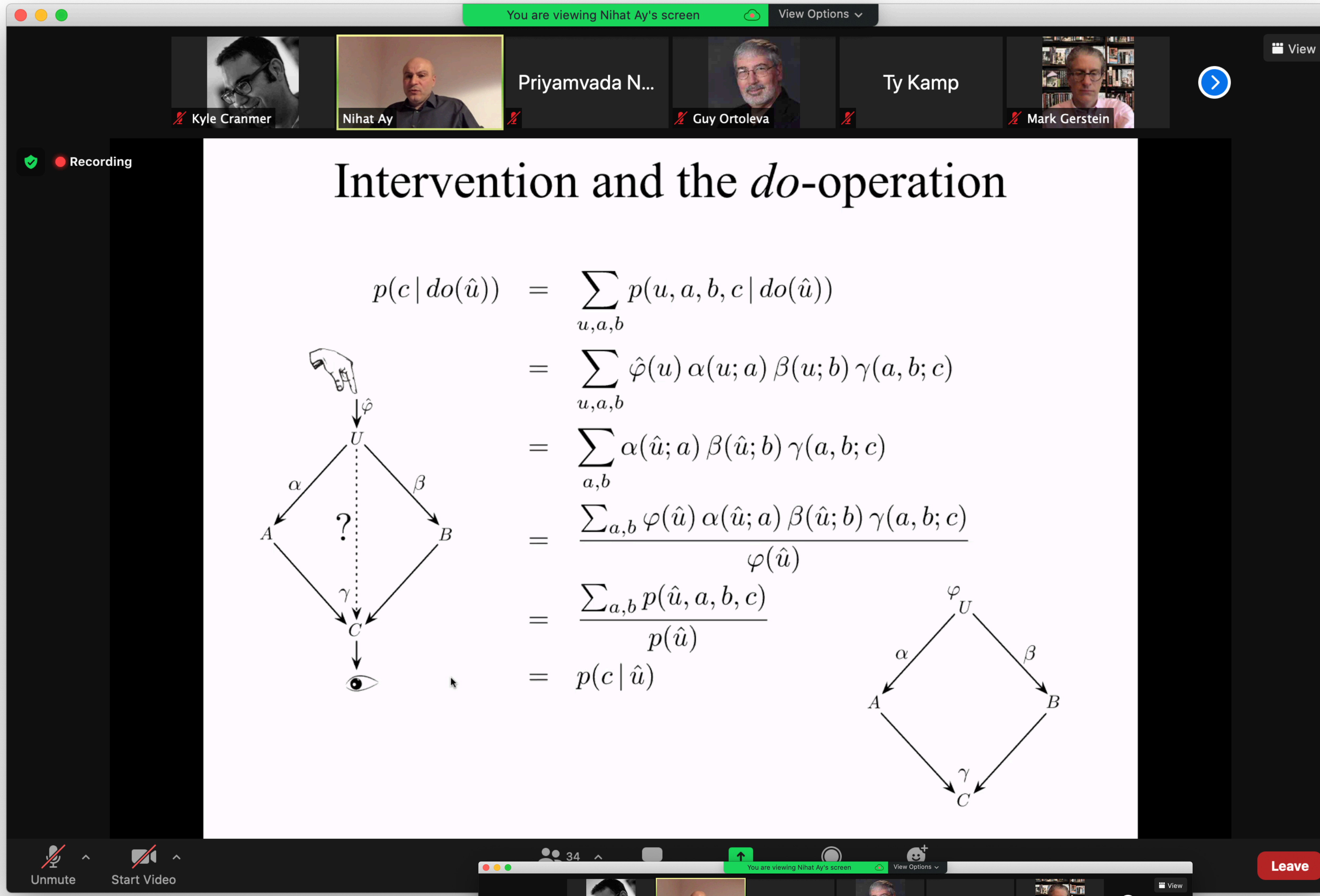


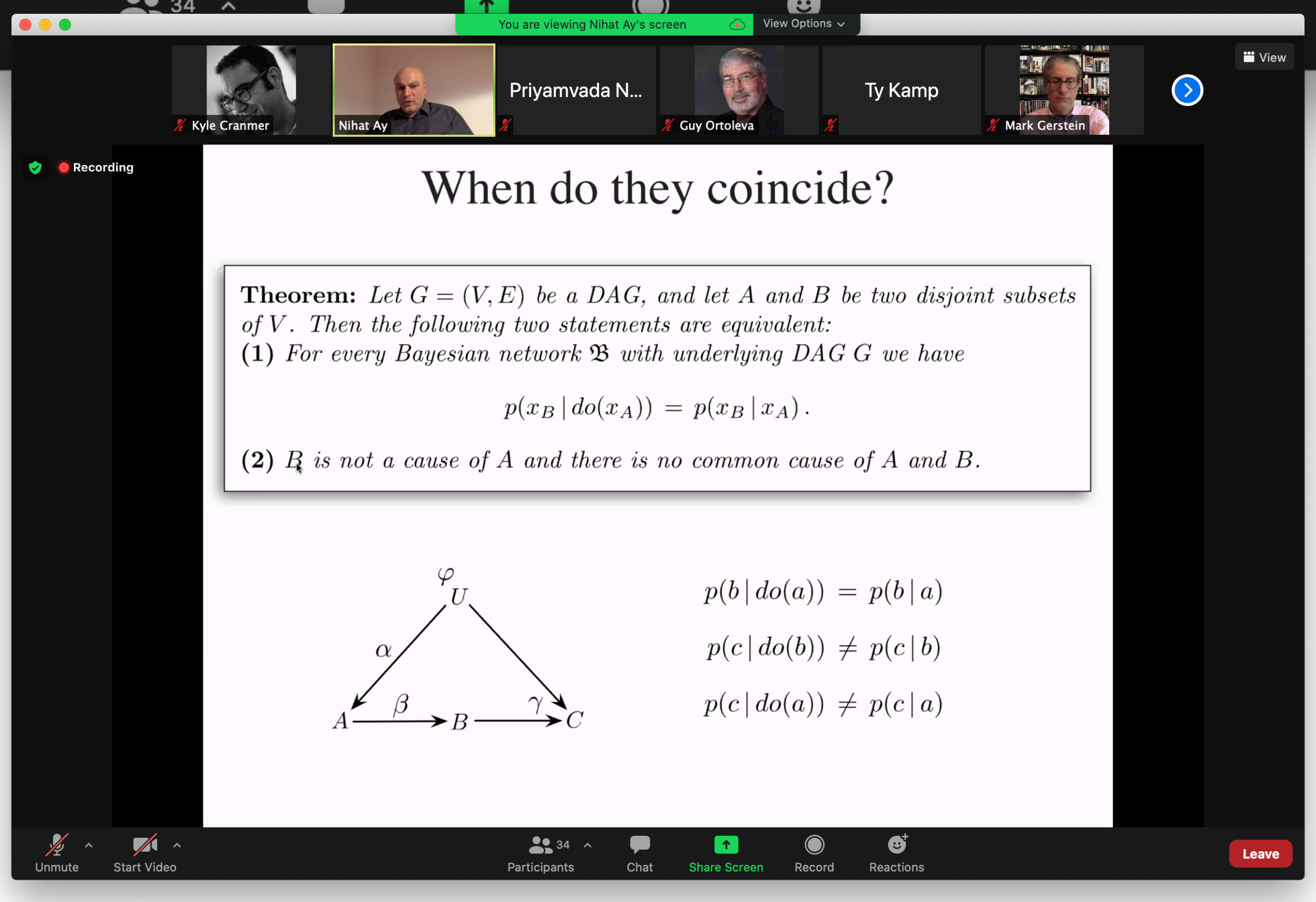





You are viewing Nihat Ay's screen
View Options
Kyle Cranmer
Nihat Ay
Priyamvada N...
Guy Ortoleva
Ty Kamp
Mark Gerstein
View
Recording
Intervention and the *do*-operation
φ U α β A B γ C
φ U $\hat{\alpha}$? β A B γ C
$\hat{\alpha}(a) := \begin{cases} 1, & \text{if } a = \hat{a} \\ 0, & \text{if } a \neq \hat{a} \end{cases}$
$p(u,a,b,c) = \varphi(u)\,\alpha(u;a)\,\beta(u;b)\,\gamma(a,b;c)$
$p(u,a,b,c \mid do(\hat{a})) := \varphi(u)\,\hat{\alpha}(a)\,\beta(u;b)\,\gamma(a,b;c)$
Unmute
Start Video
36
Participants
Chat
Share Screen
Record
Reactions
Leave


You are viewing Nihat Ay's screen
View Options
Kyle Cranmer
Nihat Ay
Priyamvada N...
Guy Ortoleva
Ty Kamp
Mark Gerstein
View
Recording
Intervention and the *do*-operation
$p(b \,|\, do(\hat{a})) = \sum_{u,a,c} p(u,a,b,c \,|\, do(\hat{a}))$
$= \sum_{u,a,c} \varphi(u)\, \hat{\alpha}(a)\, \beta(u;b)\, \gamma(a,b;c)$
$= \sum_{u,c} \varphi(u)\, \beta(u;b)\, \gamma(\hat{a},b;c)$
$= \sum_{u} \varphi(u)\, \beta(u;b) \sum_{c} \gamma(\hat{a},b;c)$
$= p(b) \neq p(b \,|\, \hat{a})$
Unmute
Start Video
Participants
Chat
Share Screen
Record
Reactions
Leave

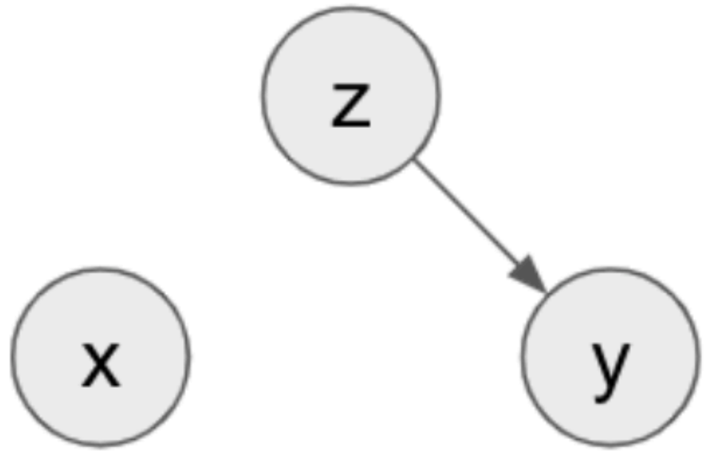

You are viewing Nihat Ay's screen
View Options
Kyle Cranmer
Nihat Ay
Priyamvada N...
Guy Ortoleva
Ty Kamp
Mark Gerstein
View
Recording
Intervention and the *do*-operation
$p(c \,|\, do(\hat{u})) = \sum_{u,a,b} p(u,a,b,c \,|\, do(\hat{u}))$
$= \sum_{u,a,b} \hat{\varphi}(u)\, \alpha(u;a)\, \beta(u;b)\, \gamma(a,b;c)$
$= \sum_{a,b} \alpha(\hat{u};a)\, \beta(\hat{u};b)\, \gamma(a,b;c)$
$= \frac{\sum_{a,b} \varphi(\hat{u})\, \alpha(\hat{u};a)\, \beta(\hat{u};b)\, \gamma(a,b;c)}{\varphi(\hat{u})}$
$= \frac{\sum_{a,b} p(\hat{u},a,b,c)}{p(\hat{u})}$
$= p(c \,|\, \hat{u})$
Unmute
Start Video
Leave


Kyle Cranmer
Nihat Ay
Priyamvada N...
Guy Ortoleva
Ty Kamp
Mark Gerstein
Recording
When do they coincide?
Theorem: *Let $G = (V,E)$ be a DAG, and let A and B be two disjoint subsets of V. Then the following two statements are equivalent:*
(1) *For every Bayesian network $\mathfrak{B}$ with underlying DAG G we have*
$p(x_B \,|\, do(x_A)) = p(x_B \,|\, x_A)$.
(2) *B is not a cause of A and there is no common cause of A and B.*
$p(b \,|\, do(a)) = p(b \,|\, a)$
$p(c \,|\, do(b)) \neq p(c \,|\, b)$
$p(c \,|\, do(a)) \neq p(c \,|\, a)$
Leave

A toy example

Ferenc Huszár

inFERENCe



$$P(y|do(X)) = p(y|x)$$

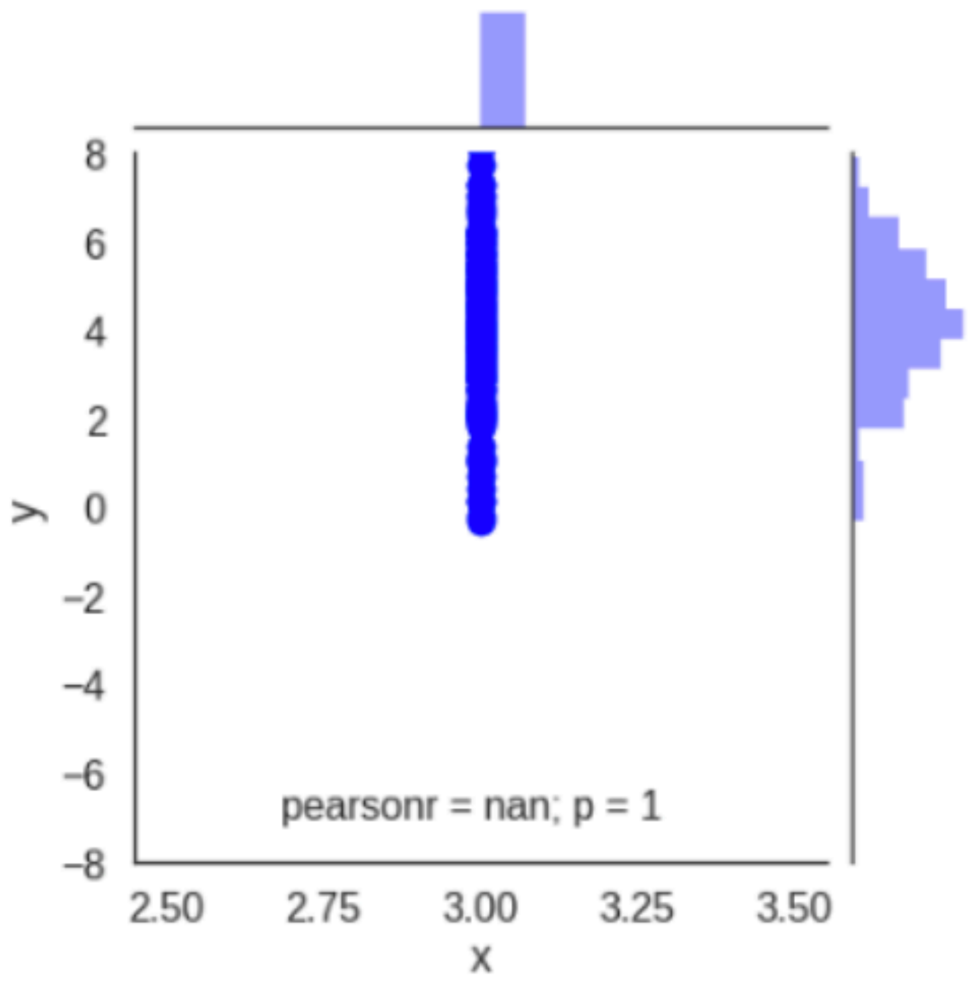
$$P(y|do(X)) = p(y)$$

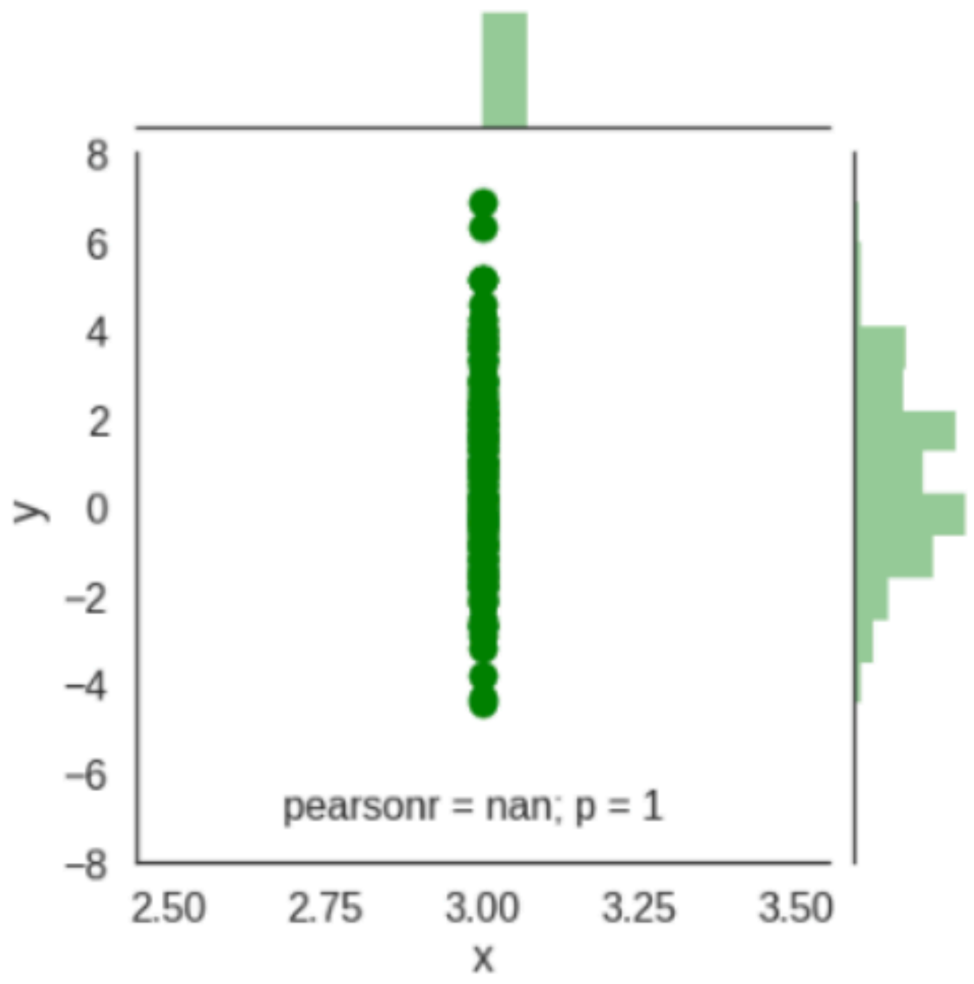
$$P(y|do(X)) = p(y)$$

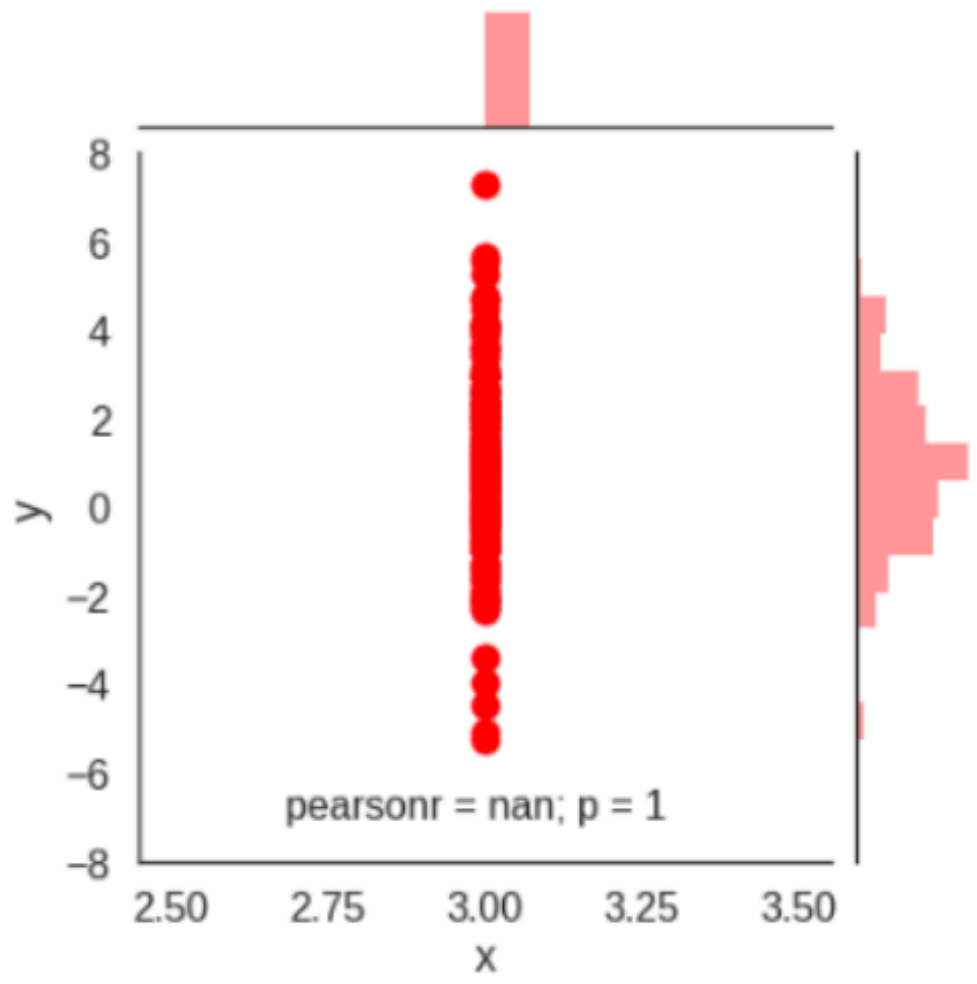
x = randn()
x = 3
y = x + 1 + sqrt(3)*randn()
x = 3

y = 1 + 2*randn()
x = 3
x = (y-1)/4 + sqrt(3)*randn()/2
x = 3

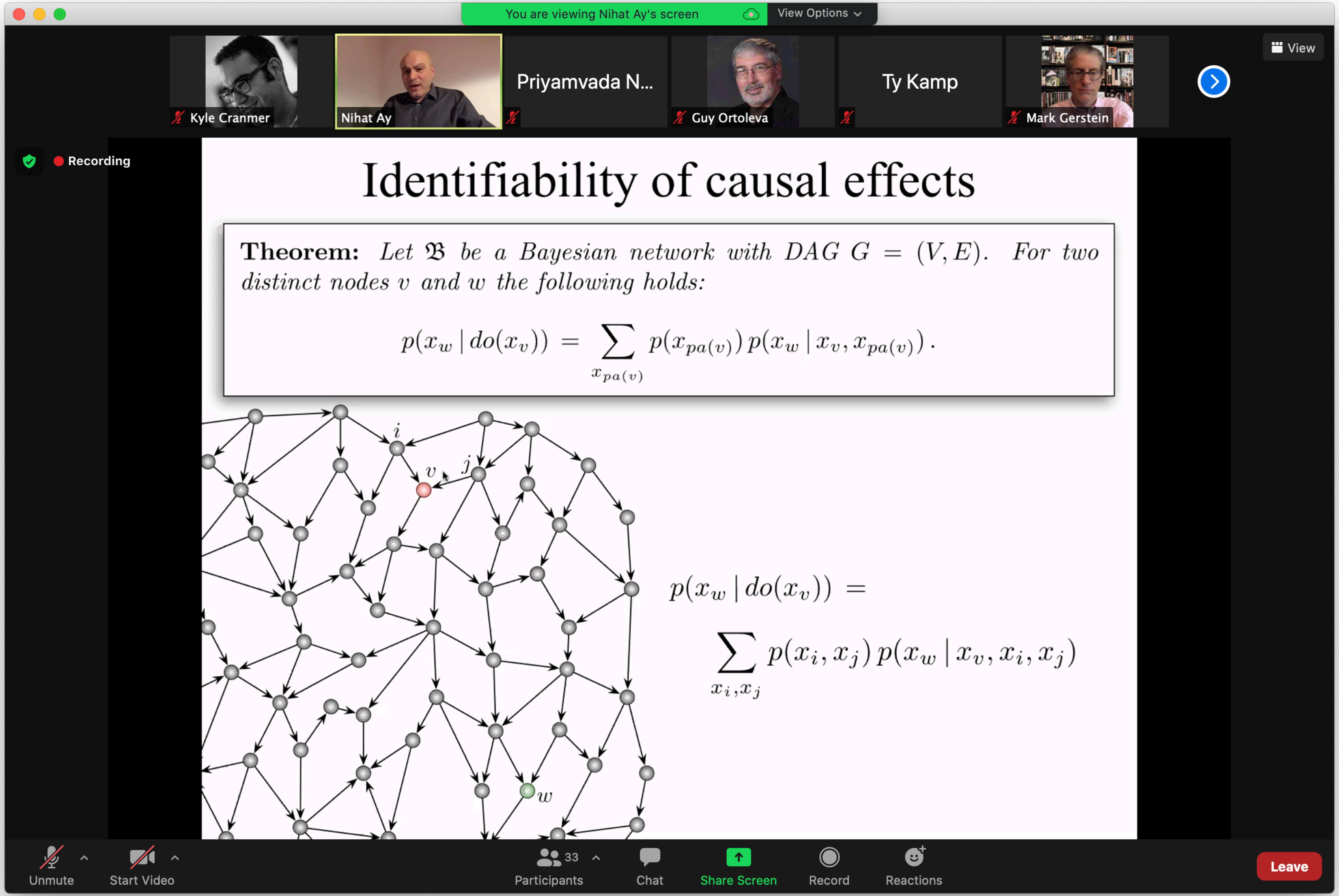
z = randn()
x = 3
x = z
x = 3
y = z + 1 + sqrt(3)*randn()
x = 3



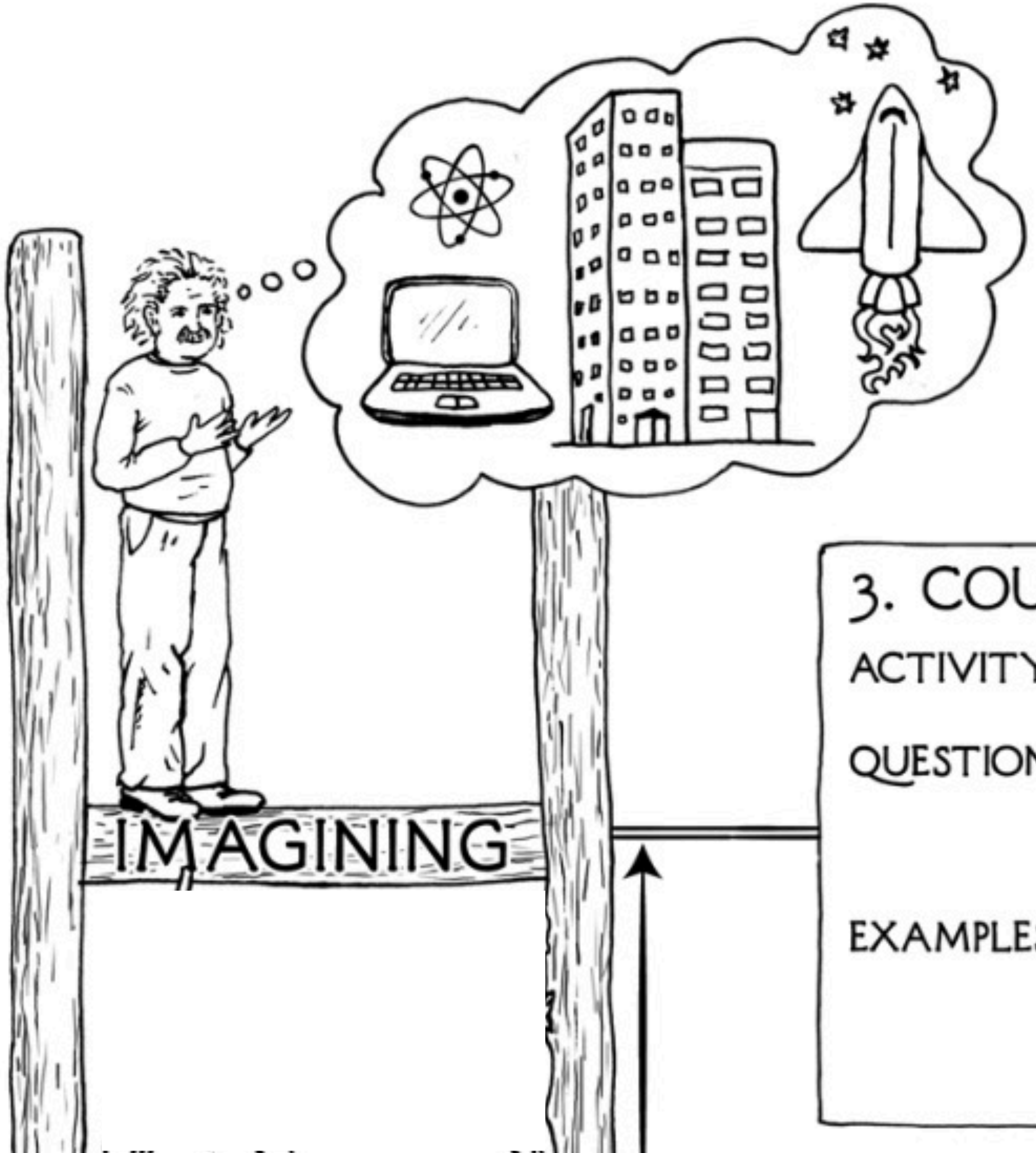




https://www.inference.vc/causal-inference-2-illustrating-interventions-in-a-toy-example/


You are viewing Nihat Ay's screen
View Options
View
Kyle Cranmer
Nihat Ay
Priyamvada N...
Guy Ortoleva
Ty Kamp
Mark Gerstein
Recording
Identifiability of causal effects
Theorem: Let $\mathfrak{B}$ be a Bayesian network with DAG $G = (V, E)$. For two distinct nodes v and w the following holds:
$p(x_w \mid do(x_v)) = \sum_{x_{pa(v)}} p(x_{pa(v)}) \, p(x_w \mid x_v, x_{pa(v)})$.
i
v
j
w
$p(x_w \mid do(x_v)) = \sum_{x_i, x_j} p(x_i, x_j) \, p(x_w \mid x_v, x_i, x_j)$
Unmute
Start Video
33
Participants
Chat
Share Screen
Record
Reactions
Leave

Counterfactuals



3. COUNTERFACTUALS

ACTIVITY: Imagining, Retrospection, Understanding

QUESTIONS: *What if I had done …? Why?*
(Was it X that caused Y? What if X had not occurred? What if I had acted differently?)

EXAMPLES: Was it the aspirin that stopped my headache? Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?

Example 1: David Blei's election example

This is an example David brought up during the **Causality Panel** and I referred back to this in my talk. I'm including it here for the benefit of those who attended my MLSS talk:

Given that Hilary Clinton did not win the 2016 presidential election, and given that she did not visit Michigan 3 days before the election, and given everything else we know about the circumstances of the election, what can we say about the probability of Hilary Clinton winning the election, had she visited Michigan 3 days before the election?

Let's try to unpack this. We are are interested in the probability that:

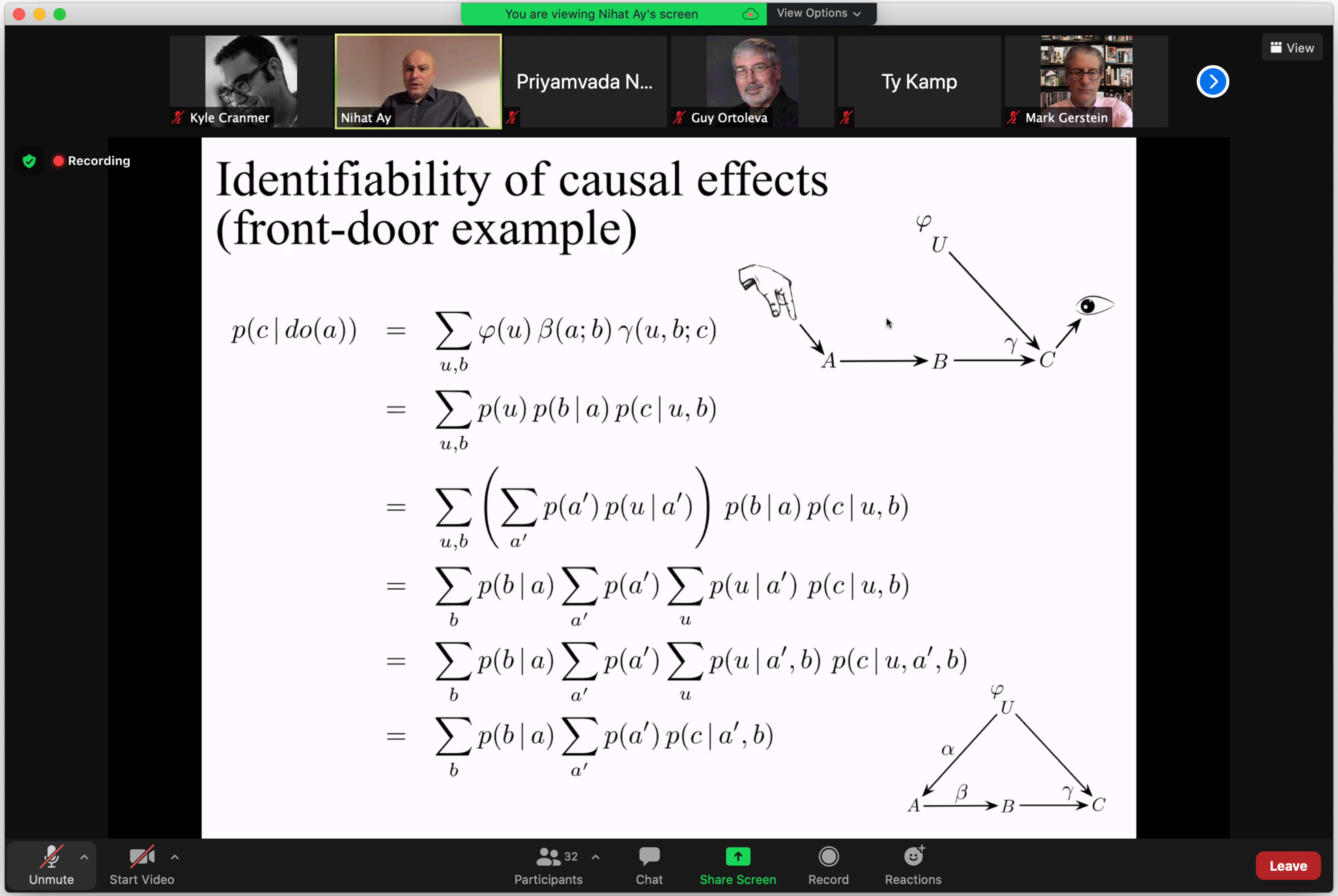
- she *hypothetically* wins the election

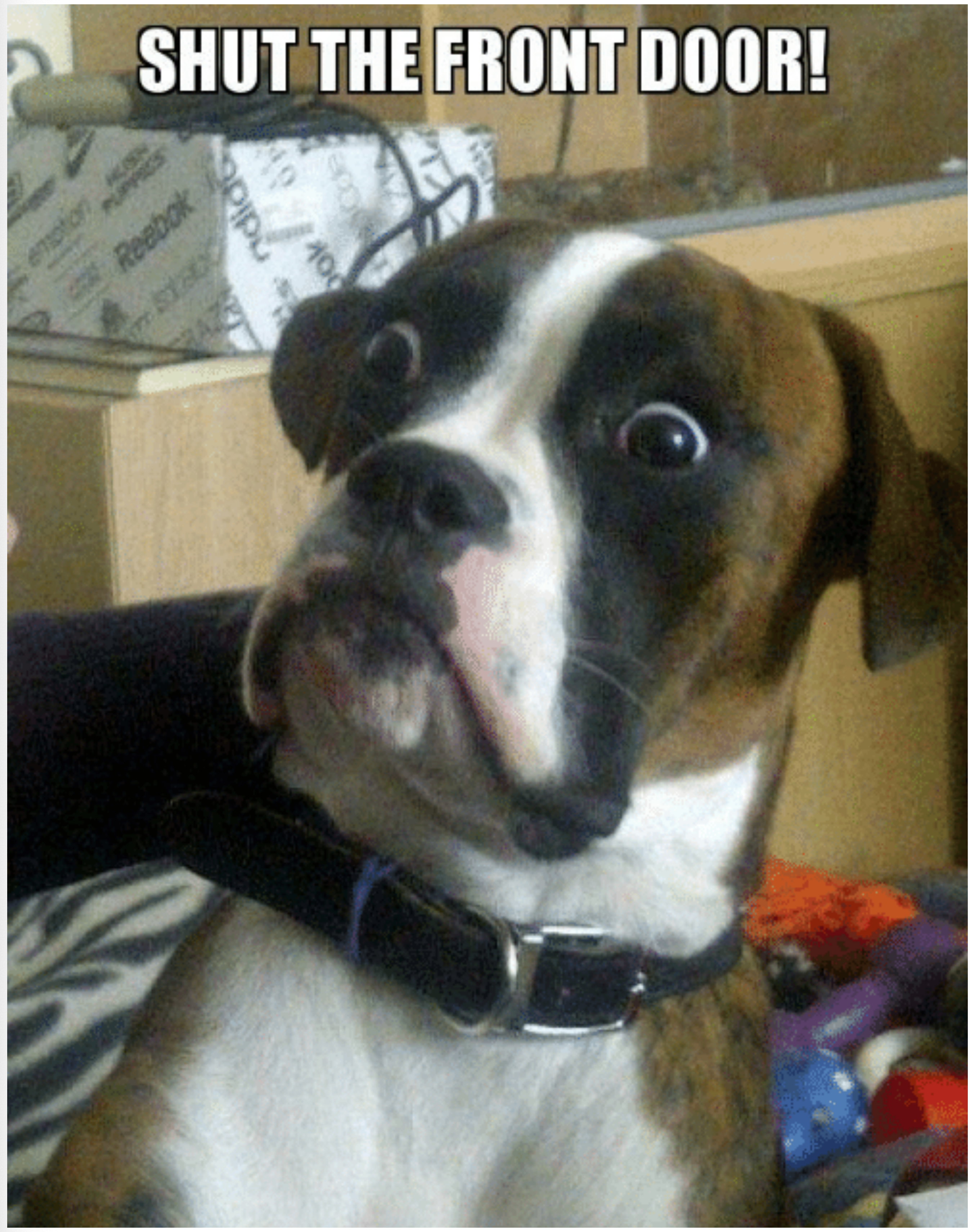
conditionied on four sets of things:

- she lost the election
- she did not visit Michigan
- any other relevant an observable facts
- she *hypothetically* visits Michigan

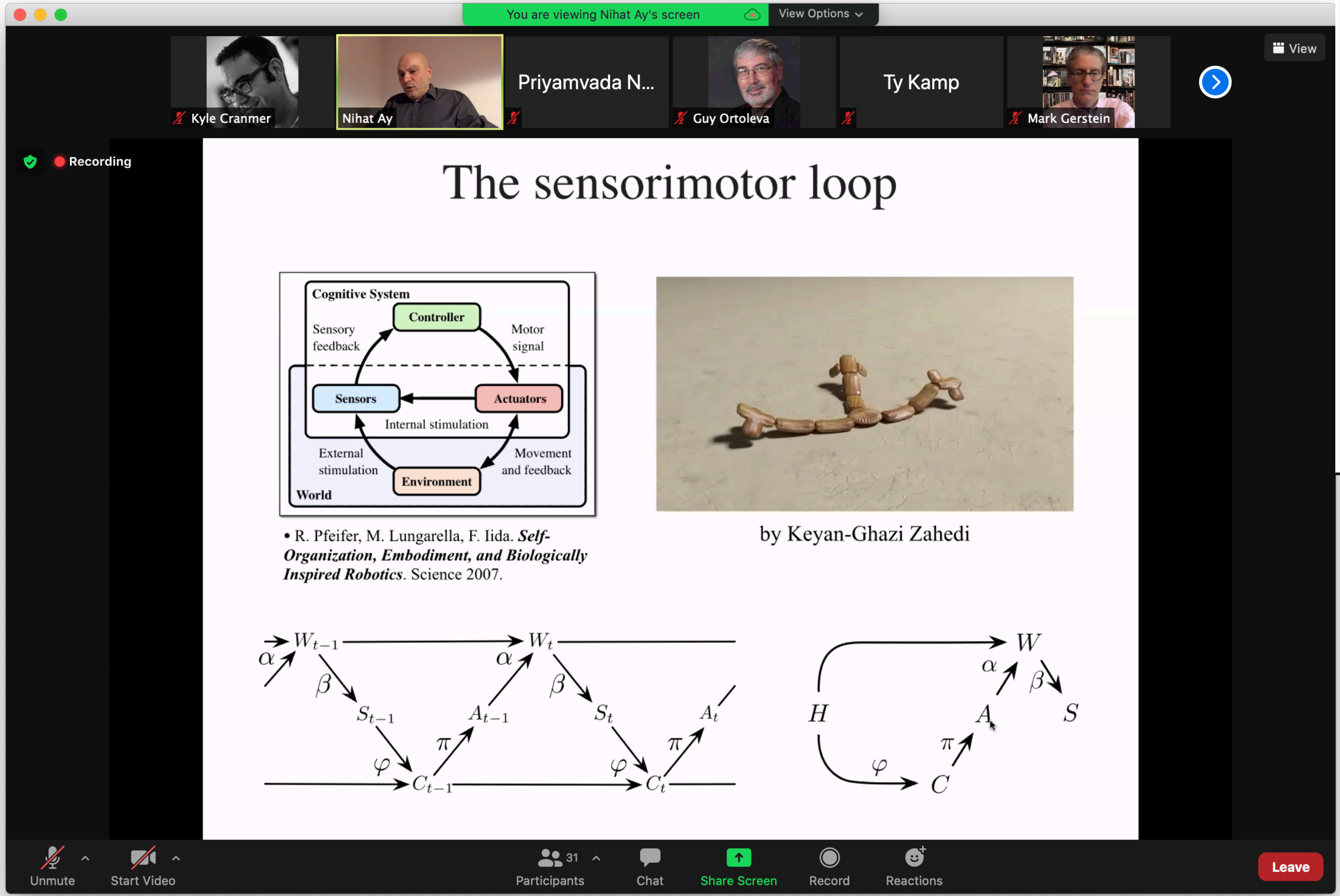
It's a weird beast: you're simultaneously conditioning on her visiting Michigan and not visiting Michigan. And you're interested in the probability of her winning the election given that she did not. WHAT?

Why would quantifying this probability be useful? Mainly for credit assignment. We want to know why she lost the election, and to what degree the loss can be attributed to her failure to visit Michigan three days before the election. Quantifying this is useful, it can help political advisors make better decisions next time.

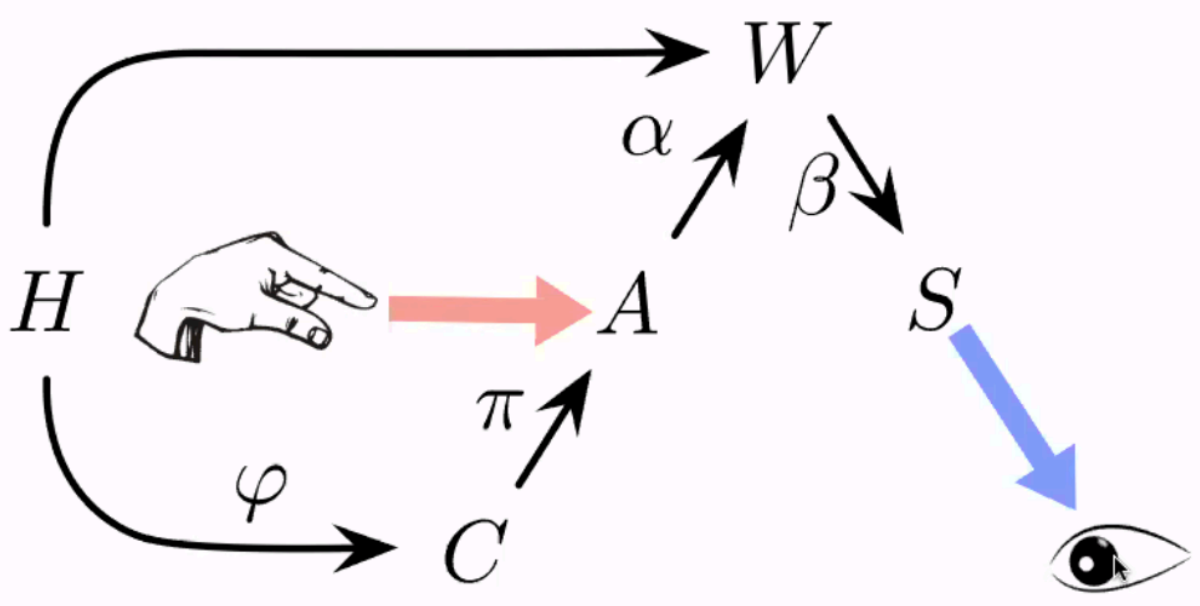

You are viewing Nihat Ay's screen
View Options
Kyle Cranmer
Nihat Ay
Priyamvada N...
Guy Ortoleva
Ty Kamp
Mark Gerstein
View
Recording
Identifiability of causal effects
(front-door example)
$p(c\,|\,do(a)) = \sum_{u,b} \varphi(u)\,\beta(a;b)\,\gamma(u,b;c)$
$= \sum_{u,b} p(u)\,p(b\,|\,a)\,p(c\,|\,u,b)$
$= \sum_{u,b} \left(\sum_{a'} p(a')\,p(u\,|\,a') \right) p(b\,|\,a)\,p(c\,|\,u,b)$
$= \sum_{b} p(b\,|\,a) \sum_{a'} p(a') \sum_{u} p(u\,|\,a')\; p(c\,|\,u,b)$
$= \sum_{b} p(b\,|\,a) \sum_{a'} p(a') \sum_{u} p(u\,|\,a',b)\; p(c\,|\,u,a',b)$
$= \sum_{b} p(b\,|\,a) \sum_{a'} p(a')\,p(c\,|\,a',b)$
Unmute
Start Video
32
Participants
Chat
Share Screen
Record
Reactions
Leave


SHUT THE FRONT DOOR!

In robotics



Causal effects in the sensorimotor loop



$$p(s \,|\, do(a)) \;=\; \sum_c p(s \,|\, c, a)\, p(c)$$

- N. Ay, K. Zahedi. ***On the Causal Structure of the Sensorimotor Loop***. GSO 2014.
- N. Ay, K. Zahedi. ***An Information-Theoretic Approach to Prediction and Deliberative Decision Making of Embodied Systems***. Proc. of ICCN 2011. Advances in Cognitive Neurodynamics, Springer 2012.

In Reinforcement Learning

Causally Correct Partial Models for Reinforcement Learning

Danilo J. Rezende [*1] Ivo Danihelka [*1 2] George Papamakarios [1] Nan Rosemary Ke [3] Ray Jiang [1]
Theophane Weber [1] Karol Gregor [1] Hamza Merzic [1] Fabio Viola [1] Jane Wang [1] Jovana Mitrovic [1]
Frederic Besse [1] Ioannis Antonoglou [1 2] Lars Buesing [1]

Causally Correct Partial Models

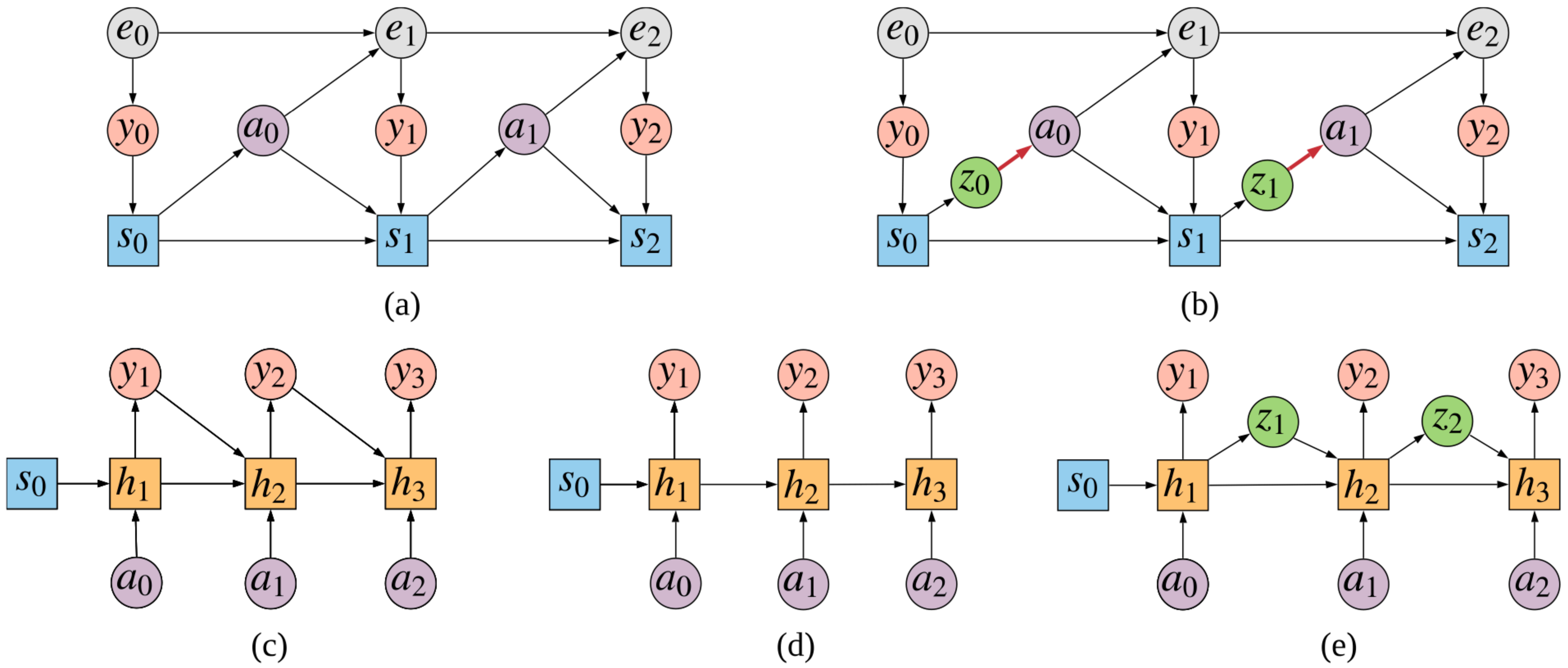


Figure 3. Graphical representations of the environment, the agent, and the various models. Circles are stochastic nodes, rectangles are deterministic nodes. (a) Agent interacting with the environment, generating a trajectory $\{y_t, a_t\}_{t=0}^{T}$. These trajectories are the training data for the models. (b) Same as (a) but also including the backdoor z_t in the generated trajectory. The red arrows indicate the locations of the interventions. (c) Standard autoregressive generative model of observations. The model predicts the observation y_t which it then feeds into h_{t+1}. (d) Example of a Non-Causal Partial Model (NCPM) that predicts the observation y_t without feeding it into h_{t+1}. (e) Proposed Causal Partial Model (CPM), with a backdoor z_t for the actions.