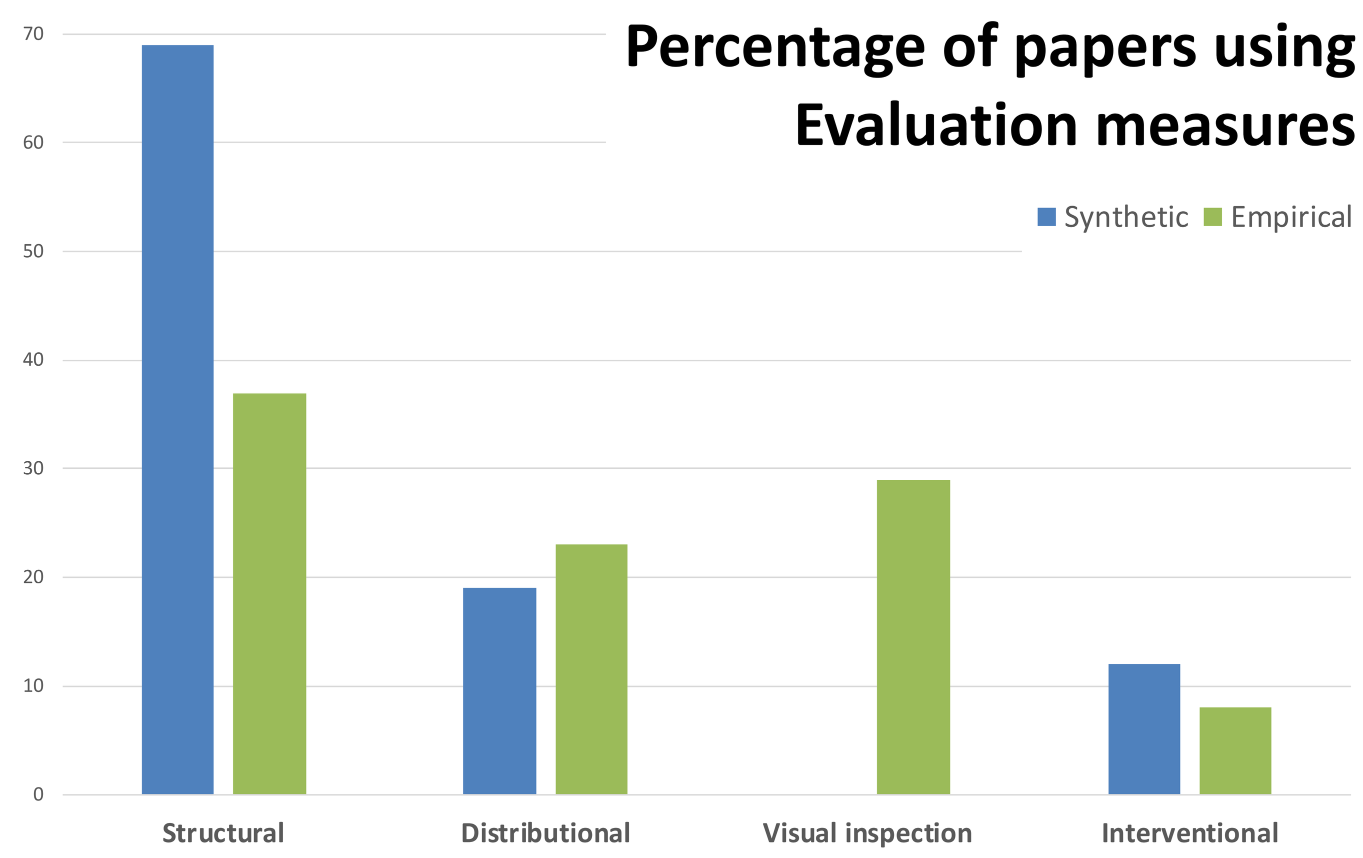
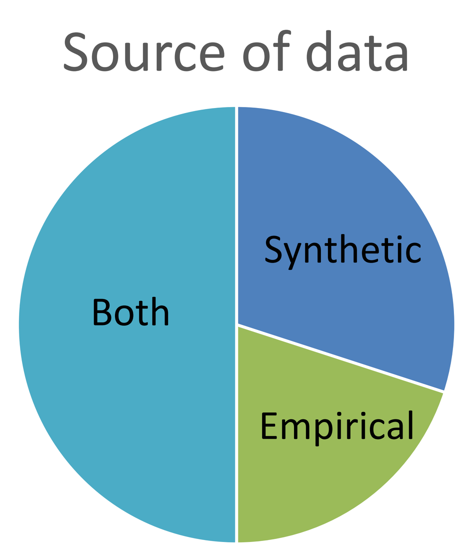
The Case for Evaluating Causal Models Using Interventional Measures and Empirical Data

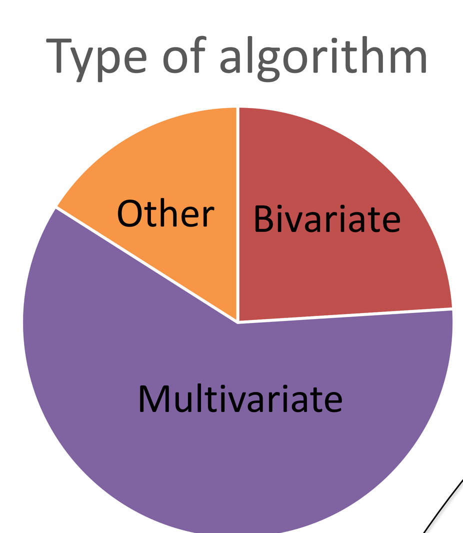
Amanda Gentzel, Dan Garant, and David Jensen
College of Information and Computer Sciences,University of Massachusetts Amherst

Q1. Don't we do this already?

- We surveyed 91 causality papers from the past 5 years of NeurIPS, AAAI, KDD, UAI, and ICML.



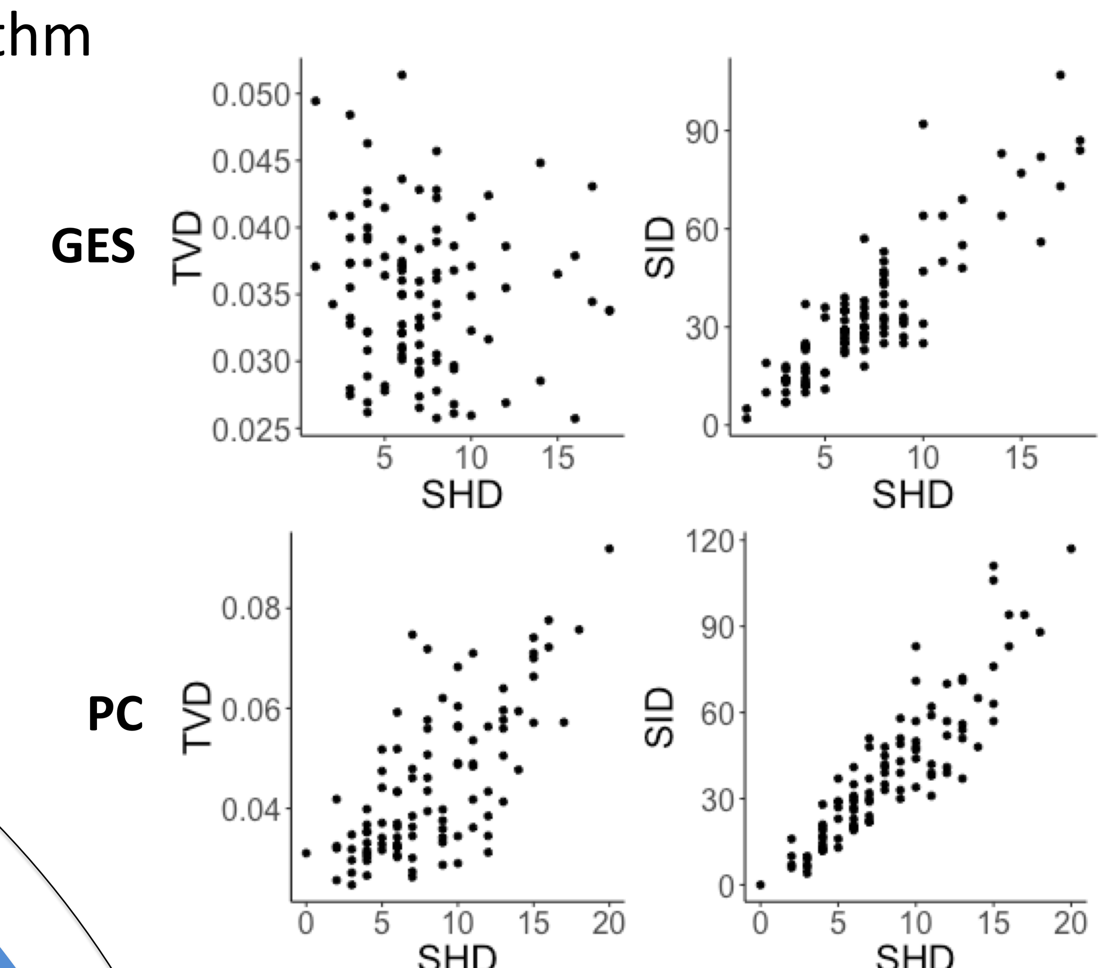




A. No, fewer than 10% of papers published in the past 5 years use a combination of empirical data and interventional measures.

Q2. Aren't structural measures enough?

- Most structural measures penalize all errors equally
- Even structural measures designed to consider interventions (ex: Structural Intervention Distance) produce similar results to other structural measures
- Interventional measures (ex: Total variation distance)
- Generated random DAGs and compare different evaluation measures, for both GES and PC
- TVD and SHD produce significantly different results, varying by algorithm



A. No, structural measures correspond poorly to measures of interventional effect, such as TVD.

Summary

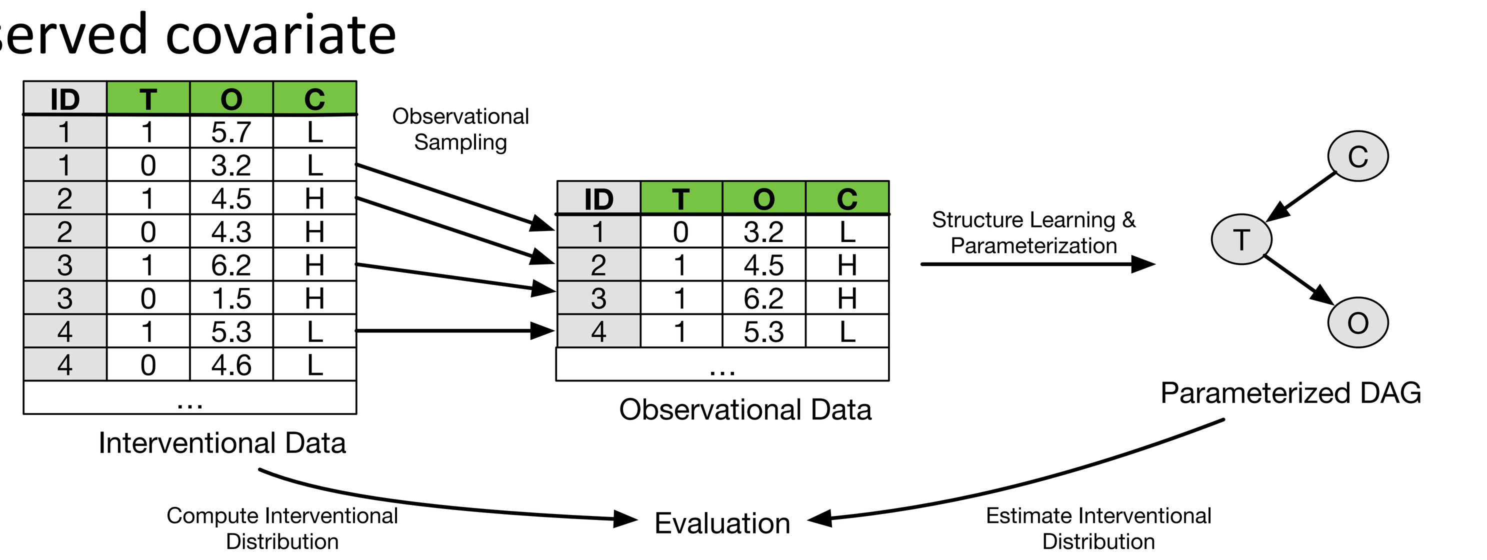
Algorithms for causal discovery are unlikely to be widely accepted unless they can be shown to accurately predict the effects of intervention when applied to empirical data.

Measures of interventional effect are necessary when trying to assess how well an algorithm learns actual causal effects

Effective methods exist to create empirical data for evaluating algorithms for causal discovery.

Q3. Is there any data that supports this type of evaluation?

- Many datasets exist with known interventional effects (DREAM, ACIC 2016 challenge, flow cytometry, cause-effect pairs challenge, etc.)
- We can collect data from computational systems
- Many advantages:
  - Empirical
  - Easily intervenable
  - Natural stochasticity
- We collected data from Postgres, the JDK, and networking infrastructure and intervened by changing system parameters
- Can create pseudo-observational data by biasing with an observed covariate



ID	T	O	C
1	1	5.7	L
1	0	3.2	L
2	1	4.5	H
2	0	4.3	H
3	1	6.2	H
3	0	1.5	H
4	1	5.3	L
4	0	4.6	L
...			

Interventional Data

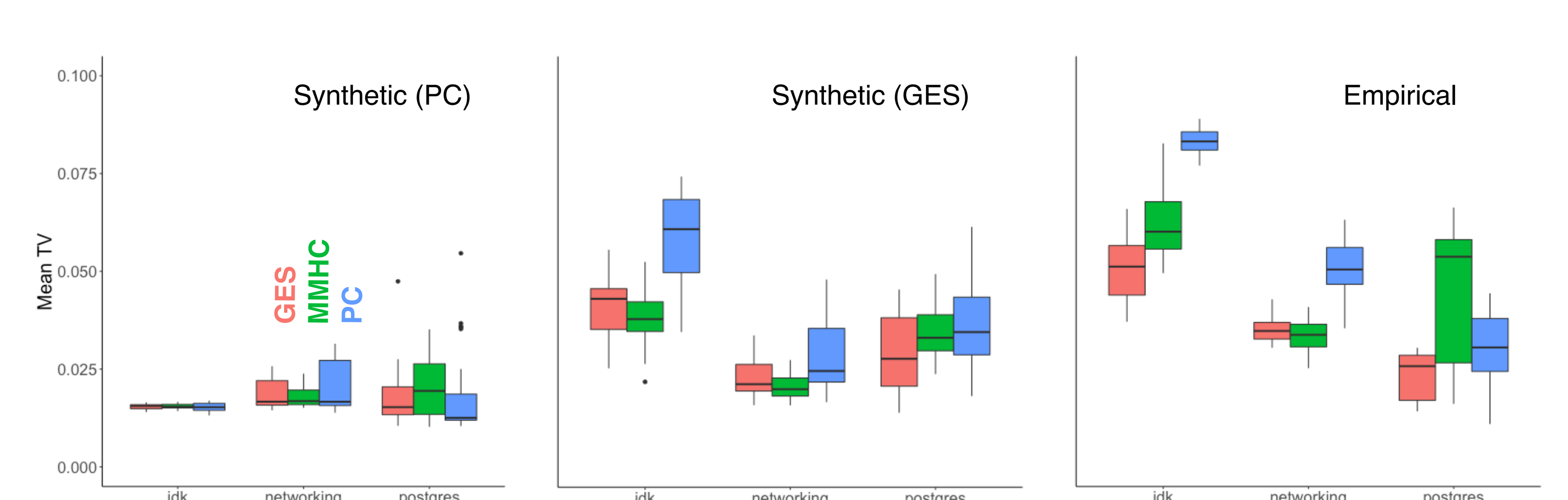
ID	T	O	C
1	0	3.2	L
2	1	4.5	H
3	1	6.2	H
4	1	5.3	L
...			

Observational Data

A. Yes, several data sets exist, including ones we have recently created from experimentation with computational systems.

Q4. Does empirical data add any value?

- Empirical data provides realistic complexity generally not present in synthetic data
- Data not generated by the researcher is less likely to contain unintentional biases and can be standardized across the community
- Provides a stronger demonstration of effectiveness
- Learned causal structure of computational systems using PC (left) and GES (center)
- Generated synthetic data based on these structures and evaluated performance of GES, MMHC, and PC
- Compare to performance of GES, MMHC, and PC on the original empirical data – significantly different relative order



A. Yes, results on empirical data sets appear to differ substantially from results on 'look-alike' synthetic data sets.