Challenges of causal analysis | Feedback on DoWhy

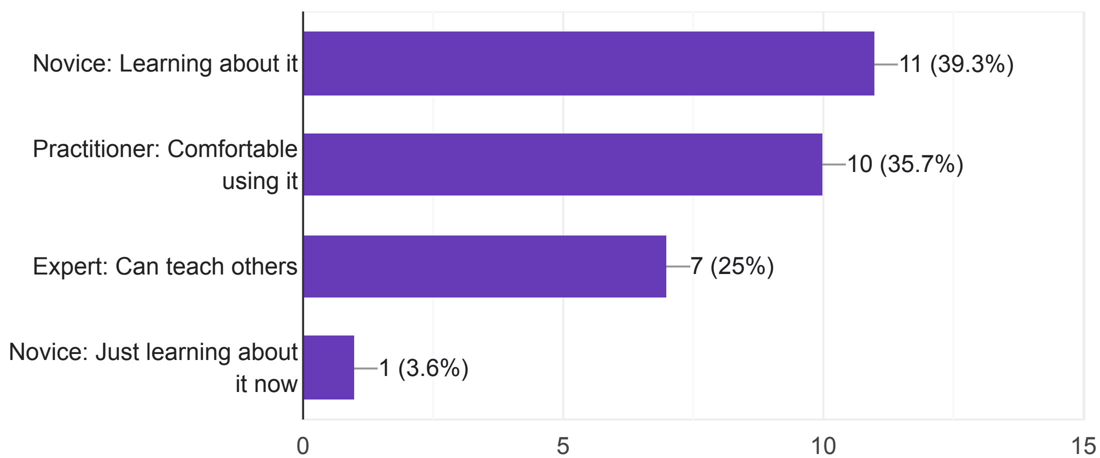
28 responses

Publish analytics

How comfortable are you with conducting a causal analysis?

Copy

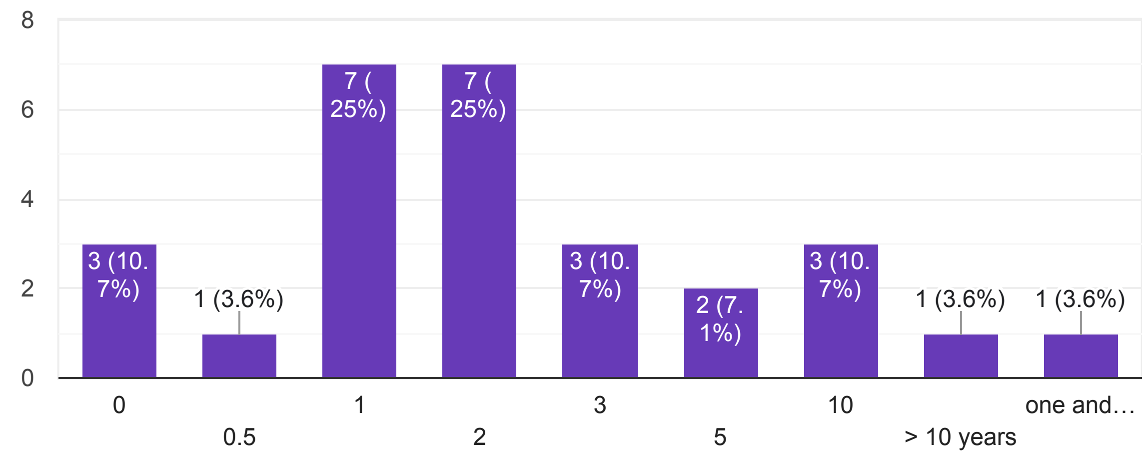
28 responses



How many years of experience with causality analysis do you have?

Copy

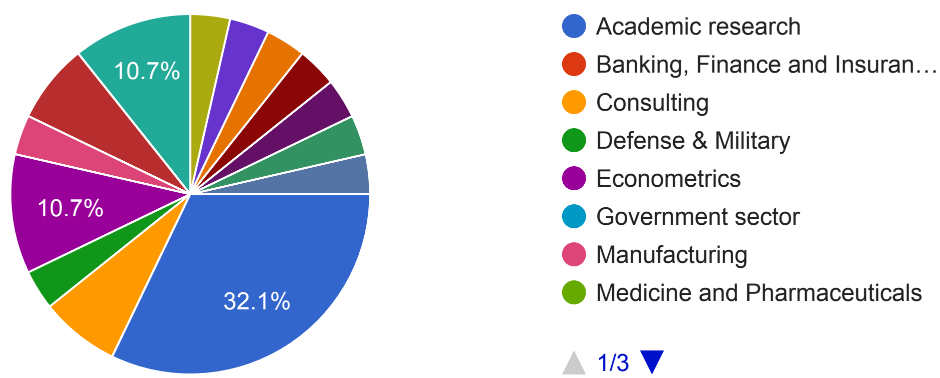
28 responses



In which context are you looking at causality analysis?

Copy

28 responses



Please describe the use-case(s) where you wish to apply or are applying causal methods.

20 responses

I wish to apply causal methods to understand issues in food supply chains.

To see economic impact

Marketing Analytics: Deriving insights and causal drivers behind time series in the context of Marketing campaigns scheduled on the Amazon Retail page

business processes, graph analysis

Many

causal contribution, root cause analysis, causal effect estimation, causal discovery, omitted variable bias estimation, policy optimization, evaluating causal models

Learning causal structure from time series and then predicting effect of an intervention.

Chemical reactions, ecology LTER

Finding relationships between different non-linear features in weather and climate science.

risk management and also areas where there is little data

Pricing, demand forecasting, public policy analysis, corporate strategy analysis

HR, Psychology

FYI: my use cases are well described by J.M. Mooij, D. Janzing, B. Scholkopf in:

https://arxiv.org/abs/1304.7920 with the addition of manipulated (controlled) variables and unobserved disturbances. Given that I am a novice user currently, I plan to contribute examples to the GitHub doWhy library in the future. Some use cases I am considering are:

1. Identify root cause of unexpected fault in energy (example: refining) or chemical production process.
2. Attribute performance "drift" due to degradation of process equipment performance (example: heat exchangers, fired heaters, catalytic reactors, filters, rotating machinery etc.) to both observed and hidden variables (non-stationary, non-linear).
3. Forecast future performance of process technology to enable optimal decisions related to maintenance timing and costs.
4. Identify equipment constraints when planning to operate in a region outside of the training dataset (never seen before in life of equipment).
5. Use counterfactual techniques, determine if delay of routine maintenance checks would result in equipment failure if the checks were deferred by xx months.
6. Include influence of management and business decisions on performance outcomes of assets (gain insight into impact of business decisions when trading off potential outcomes).

Promotion Effectiveness, Personalized marketing

COVID-19 vs Environmental factors (https://pubs.acs.org/doi/10.1021/acs.est.1c02204); fate and transport of contaminants

identify the effects of software engineering practices (or the causes of certain outcomes)

Agricultural decision making.

scientific machine learning on lithium-ion batteries for predictive maintenance

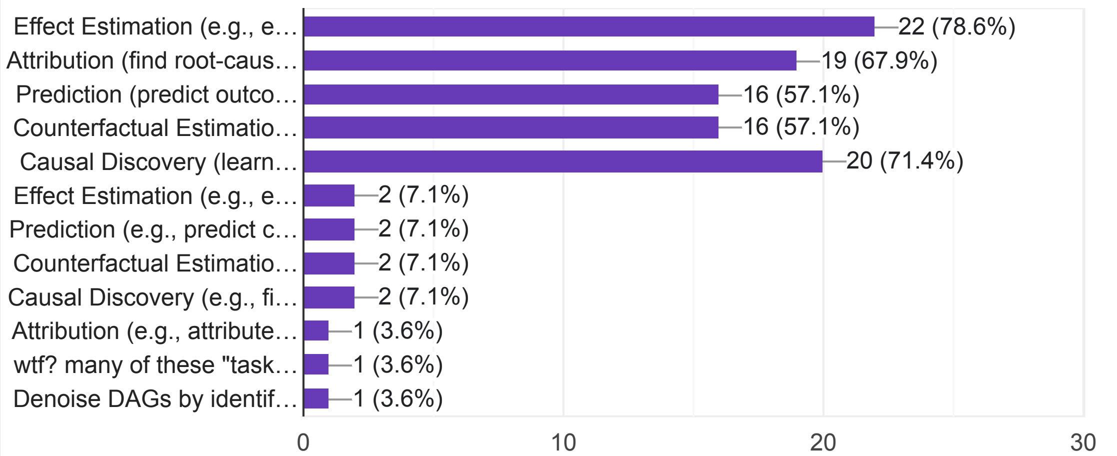
Discovering complex causal relationships from observational data

Cloud Software System Outages Analysis

Which of the following tasks are relevant to your use case?

Copy

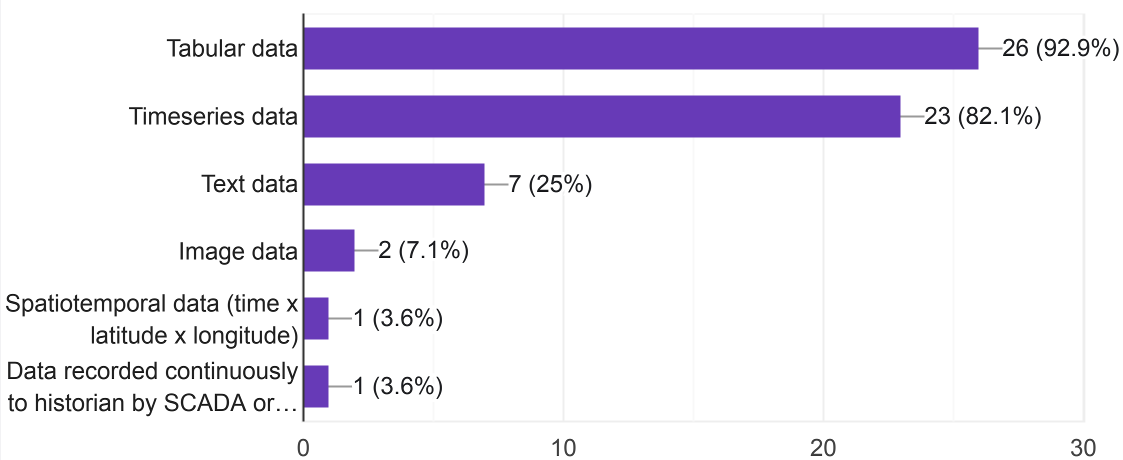
28 responses



What kind of data do you have for your use case?

Copy

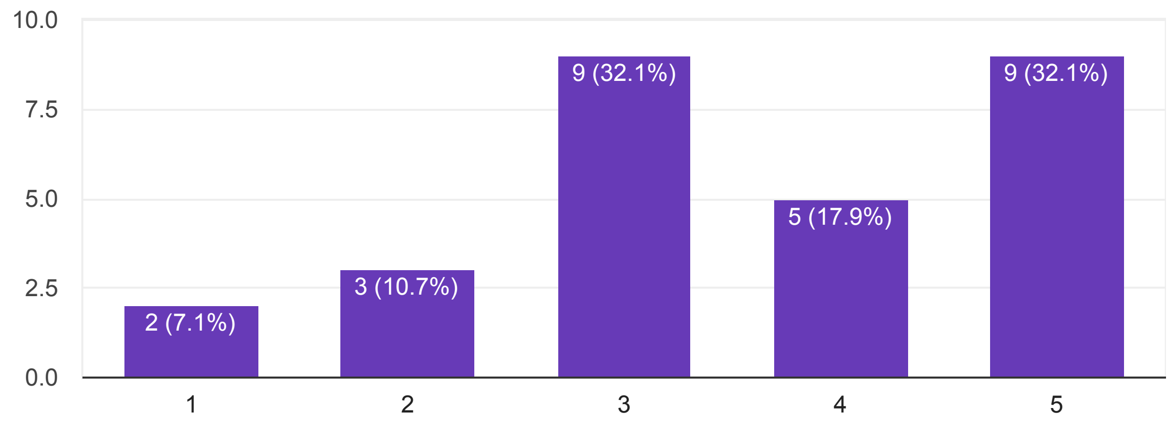
28 responses



How useful have you found DoWhy for conducting causal analysis?

Copy

28 responses



Can you elaborate?

17 responses

I have yet to actively use the library, but overall it looks great.

Not clear at all from the docs why DoWhy is useful over and above sklearn, econml, etc.

one example: graphical causal models are very usedul in assessing feature relations

User interface is a bit complicated. Limited to a few estimands that can be actually estimated. Cannot estimate effect with continuous treatment variables. I like the separation into different parts (identification, estimation, refutation). Scalability is also an issue of the package.

The documentation was very scattered and I couldn’t use functionality beyond a BD estimation procedure.

Reputation anslydis

very early stage - we actually use another causal library currently(causalnex) from quantum black

NA

The doWhy.gcm module appears to be most relevant to my use cases and the example notebooks provide sufficient starting points to apply the techniques to my datasets.

On the challenging side, interpreting the GCM module against the alternative doWhy methods has taken some effort since the relevancy of model refutation, hidden variables, Instrumental Variables, dataset sufficiency/bias in the GCM module does not appear to be as much of a concern. Also the documentation support could be improved but I see from the doWhy GitHub wiki that docs improvement is already an active discussion so will not elaborate further.

I especially liked the refutation step, which is unique for DOWHY !

Refutation testing

Refutation process

Limited experience as yet, so hard to judge: but low threshold to start, really appreciate the refutations, need support for non-binary data, showwhy still to limited and buggy.

One-stop shopping.

I love the 4-step process and the new GCM API.

N/A I am starting now using it

It is one of the only production quality libraries for learning causal graphs

What are the biggest obstacles to applying causal methods to your use case?

20 responses

I am still learning causal methods.

Create understanding of benefit of causal methods as compared with traditional methods

theoretical distance from basic applications like bayesian networks

Communication between subject matter experts and data scientists

Efficient estimation of complicated estimands. Identifying the causal structure (with and without using data).

A consistent API

Lack of theoretical background

Different methods show different results, particularly for causal discovery. How to establish the significance of a method.

I dont see any specific obstractles - but lack of skills, ease of use etc(because its not familiar to most data scientists)

NA

My comments relate nearly entirely to the GCM module since my domain typically has DAGs that are easily identifiable due to the design of the assets I work with (upstream raw materials flow to downstream finished products). Obstacles I have found so far include:

1. How to handle disturbances from hidden variables (and confounders)? If X-->Y and Z-->Y and I have data for both X and Y but not Z, then the GCM module does not appear to be

capable of attributing causes to Z through evidential learning. The datasets tend to always assume full measurement of DAG nodes. I might be mistaken on this since I only recently started using example code. If use of instrument variables are the answer then it could be made explicit in the GCM help docs.

2. Understanding dataset requirements is somewhat difficult because the approach seems very liberal when compared against the requirements of other ML and DL techniques. Normalization of inputs, bias correction, over-sampling of rare events, size of dataset, randomization/ shuffling of inputs are a couple of aspects that do not appear at all important when using DAGs and SCMs. This could be addressed more explicitly if only to resolve the discrepancy a user sees when dealing with causal modeling for first time.

3. The metrics used to assess DAG accuracy and faithfulness could be standardized or expanded. The p-value and KL divergence are used in several notebooks but utilizing graphical outputs may help the user who is used to looking at confusion matrices, F-scores, and 'Y-actual' vs 'Y-predicted' scatterplots from typical ML/DL applications. Although cross-validation does not apply, would it not be useful to hold back some training data as a validation set to display to the user as part of the workflow?

4. The theory of Causality and the techniques used in the GCM module are incredibly interesting and fascinating yet they are also intimidating to non-academic, industrial users like me. The problems handled are also quite serious (healthcare, economic development, corporate strategy etc.). It may help if some examples on lighter topics were included to engage new users and help the marketing of the subject matter. The DL community have come very far based on the classification of dogs and cats so perhaps the Causality community could find an equivalent 'fun' topic to engage in!

Validating that the Causal Structure found by the software is the correct one

Scaling models

Hard to judge which methods are (not) appropriate and when; leads to fair of wrong/unreliable conclusions

Observational data with highly diverse confounder acting across various spatial scales.

Limited set of identification methods (lack of generalized do-calculus), no easy way to generate counterfactuals using the main CausalModel API, no easy way to generate predictions from CausalModel (or lack of documentation on this functionality)

Proper setup of the problem: how to model systems' dynamics (ODEs, PDEs, ...)

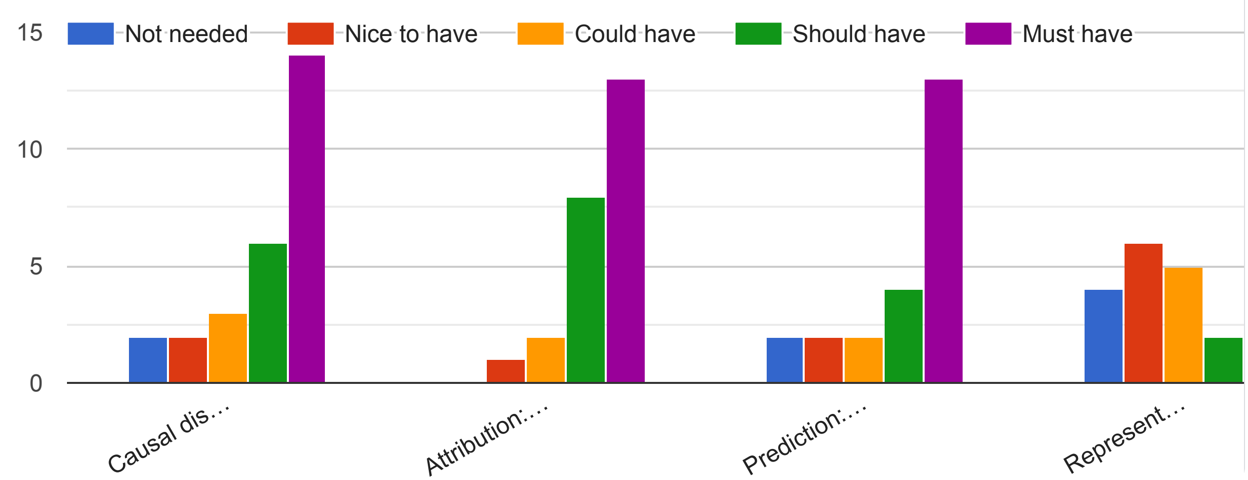
Finding usable data

Knowing what data to aggregate

Lack of easy to use tools

We are listing some of the potential next steps we are considering for DoWhy. How important are these from your perspective?

Copy



What other features and functionalities would you like to see implemented in DoWhy?

12 responses

Simulating data sets given causal graphs and then perhaps benchmarking causal discovery algorithms and estimation algorithms would be cool to see.

Alternative assements

i said yes to all the above features because i think they are all needed. nothing else to add. PS Emre is familar with my work https://www.linkedin.com/in/ajitjaokar/

NA

1. Library (or cheat sheet) of DAG relationships (structures) users should avoid so that model will learn with higher accuracy. (see J.Peters, D. Janzing, B. Scholkopf, "Elements of Causal Inference", ch 9.4 and 9.5 for examples). Also more explicit directions on how to "break" feedback loops (recycles) for users would be helpful.

2. Standardize the workflow so that users can be confident they are following good practices when developing their investigations and models.

3. DAGs naturally lend themselves to visual explanations/ presentations and I would like to see better use of graphical tools to help the user. For example a platform like MS Azure ML Designer or the MathWorks Simulink application are examples of user friendly styles to build and test DAGs. Other visualizations such as force-directed graphs are also impressive.

4. I would like to see how SCMs and FCMs handle rare events such as earthquakes, extreme storms, or failure of typically reliable equipment.

I just recently learned about a new Causal Discovery algorithm called GLOBE, which produces a fully oriented graph / DAG !!! It is based on the application of Kolmogorov complexity (that is to say: an approximation of it, by means of MDL). This is the publication: Discovering Fully Oriented Causal Networks, https://eda.mmci.uni-saarland.de/pubs/2021/globe-

mian,marx,vreeken-wapx.pdf
I am conducting experiments with the open source software of GLOBE, mainly to compare it with existing Causal Discovery software (Causal Discovery Toolbox).

Because GLOBE delivers a unique, fully oriented DAG, I think it is a very important algorithm for Causal Discovery and therefore certainly worthwhile incorporating it in DOWHY (or PYWHY for that matter) !

This would be my recommendation for implementation of GLOBE in DOWHY.

Support for Scalable models and Distributed refutation testing

Causal Discovery is not necessary since the two tasks rely on different sets of dependencies

Time series. Panel data. Predicting intervention effects.

I perceive the GCM API as very promising, would like to see more comprehensive documentation on this API.

Also a clear way to generate predictions from the CausalModel API would be great!

I don't know yet

Maybe more state of the art algorithms like NO TEARS or similar algorithms

Your email (optional)

15 responses

emolamol@gmail.com

luitpold@amazon.com

stu.frost@geminos.ai

peter.gmeiner@algobalance.com

zelimir.kurtanjek@gmail.com

manmeet.singh@utexas.edu

ajit.jaokar@conted.ox.ac.uk

acucar@gmail.com

david.cliffe@shell.com

jfhmbours@gmail.com

prakag@gmail.com

kangqiao2015@gmail.com

l.bergmans@sig.eu

agrolisboa@gmail.com

aleksander.molak@gmail.com