UNIVERSITY OF COPENHAGEN

Discovering the cause: Tools for structure learning in R

Anne Helby Petersen

Github: annennenne, ahpe@sund.ku.dk

Section of Biostatistics, University of Copenhagen

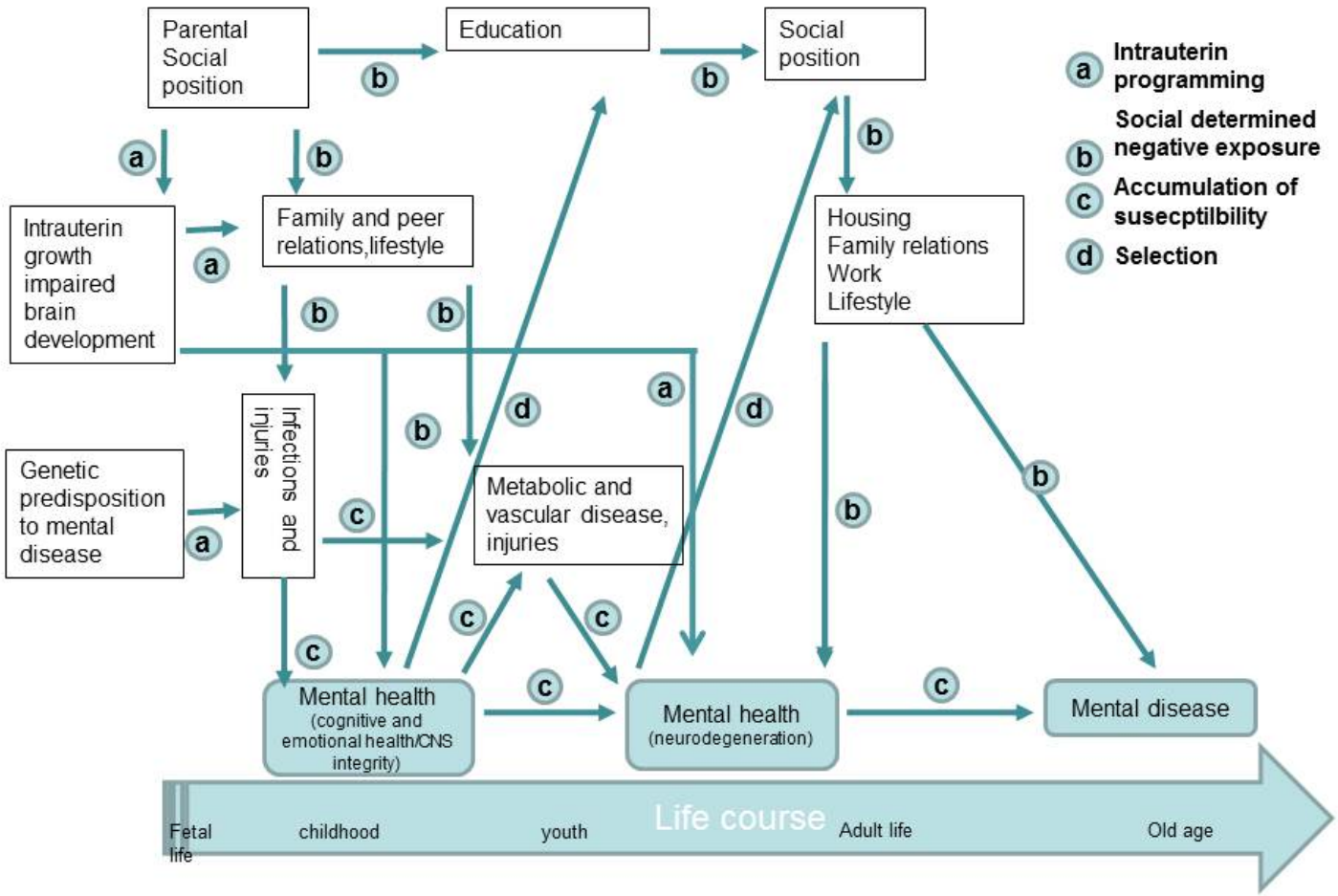
Looking for a cause

RQ: What factors influence development of dementia, depression and alcohol abuse?

Looking for a cause

RQ: What factors influence development of dementia, depression and alcohol abuse?

Fig 1. Life-course model for mental health with an indication of the mechanisms linking life expoures and mental disease



More automation, please!

P://causalDisco - master - RStudio Sou...

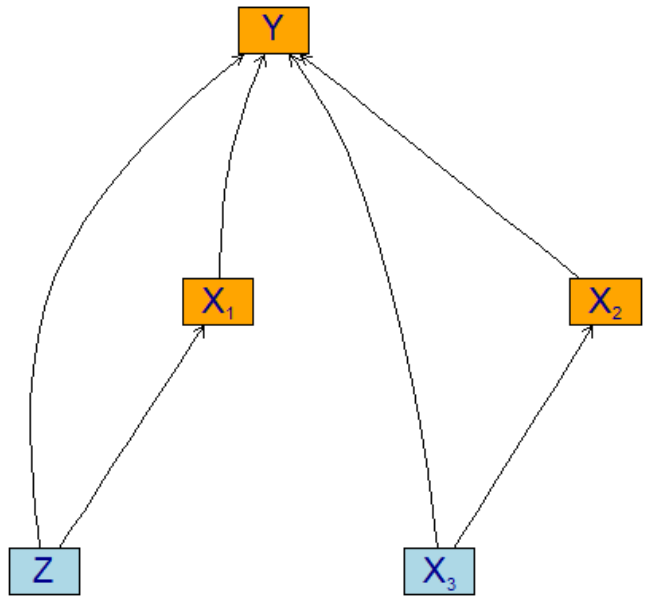
numData

Filter

	X1	X2	X3	Z	Y
1	3.391729	7.569873	6.029135	9.439524	13.454731
2	3.414703	14.188453	9.712695	9.769823	16.038376
3	3.698171	9.334827	6.896619	11.558708	13.107802
4	4.202275	10.043174	8.131201	10.070508	18.803295
5	4.168309	6.660888	5.917512	10.129288	20.587377
6	4.655413	12.207344	8.296038	11.715065	23.831699
7	4.129180	14.153822	8.673465	10.460916	22.983059
8	3.066846	10.600475	7.478397	8.734939	13.608160
9	3.062538	11.641169	9.343594	9.313147	12.973388
10	3.534678	13.879142	9.159190	9.554338	17.606833
11	5.052163	15.668988	9.916494	11.224082	31.416680
12	3.753359	11.015555	8.334251	10.359814	15.905559

Showing 1 to 12 of 1,000 entries

?



Q: Can we infer causal models from data?

More automation, please!

P://causalDisco - master - RStudio Sou...

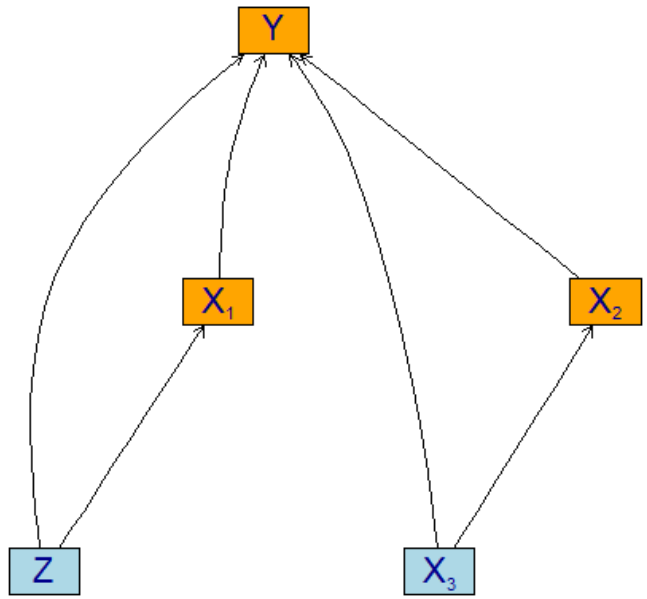
numData

Filter

	X1	X2	X3	Z	Y
1	3.391729	7.569873	6.029135	9.439524	13.454731
2	3.414703	14.188453	9.712695	9.769823	16.038376
3	3.698171	9.334827	6.896619	11.558708	13.107802
4	4.202275	10.043174	8.131201	10.070508	18.803295
5	4.168309	6.660888	5.917512	10.129288	20.587377
6	4.655413	12.207344	8.296038	11.715065	23.831699
7	4.129180	14.153822	8.673465	10.460916	22.983059
8	3.066846	10.600475	7.478397	8.734939	13.608160
9	3.062538	11.641169	9.343594	9.313147	12.973388
10	3.534678	13.879142	9.159190	9.554338	17.606833
11	5.052163	15.668988	9.916494	11.224082	31.416680
12	3.753359	11.015555	8.334251	10.359814	15.905559

Showing 1 to 12 of 1,000 entries

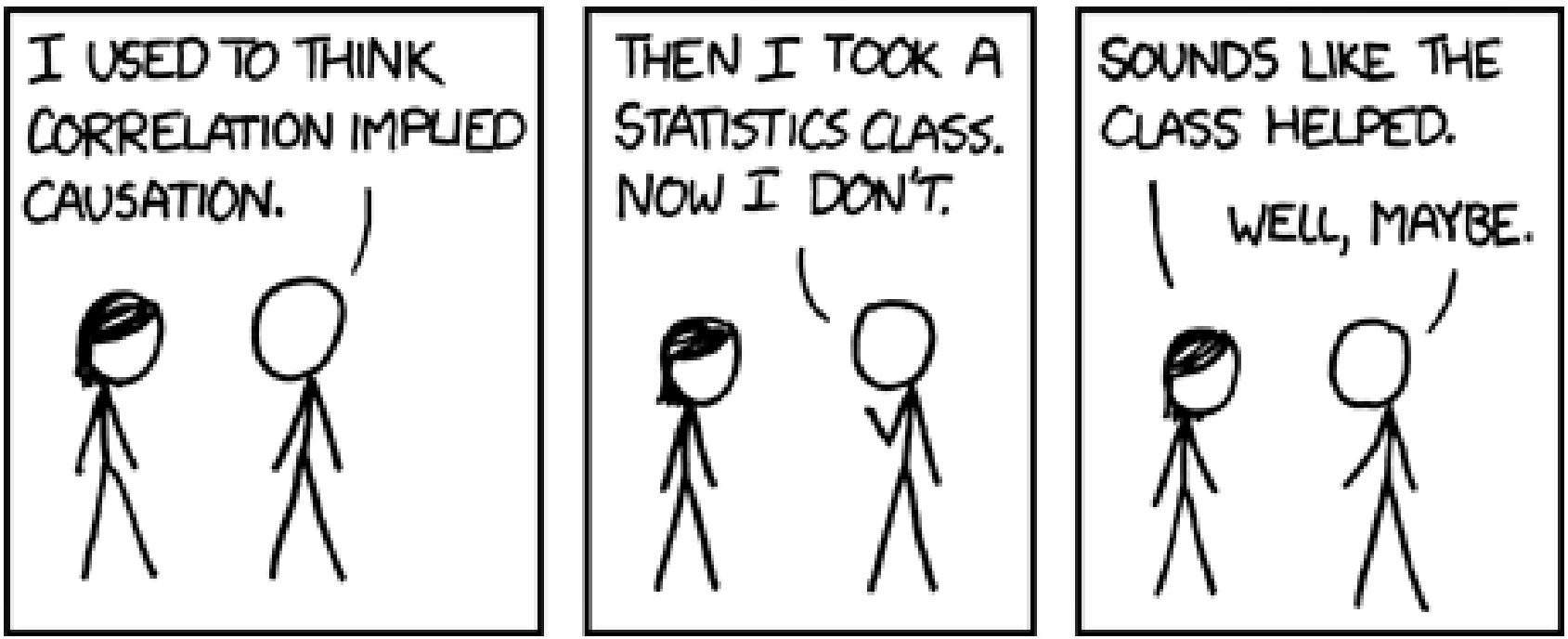
?



Q: Can we infer causal models from data?

A: Yes – sometimes!

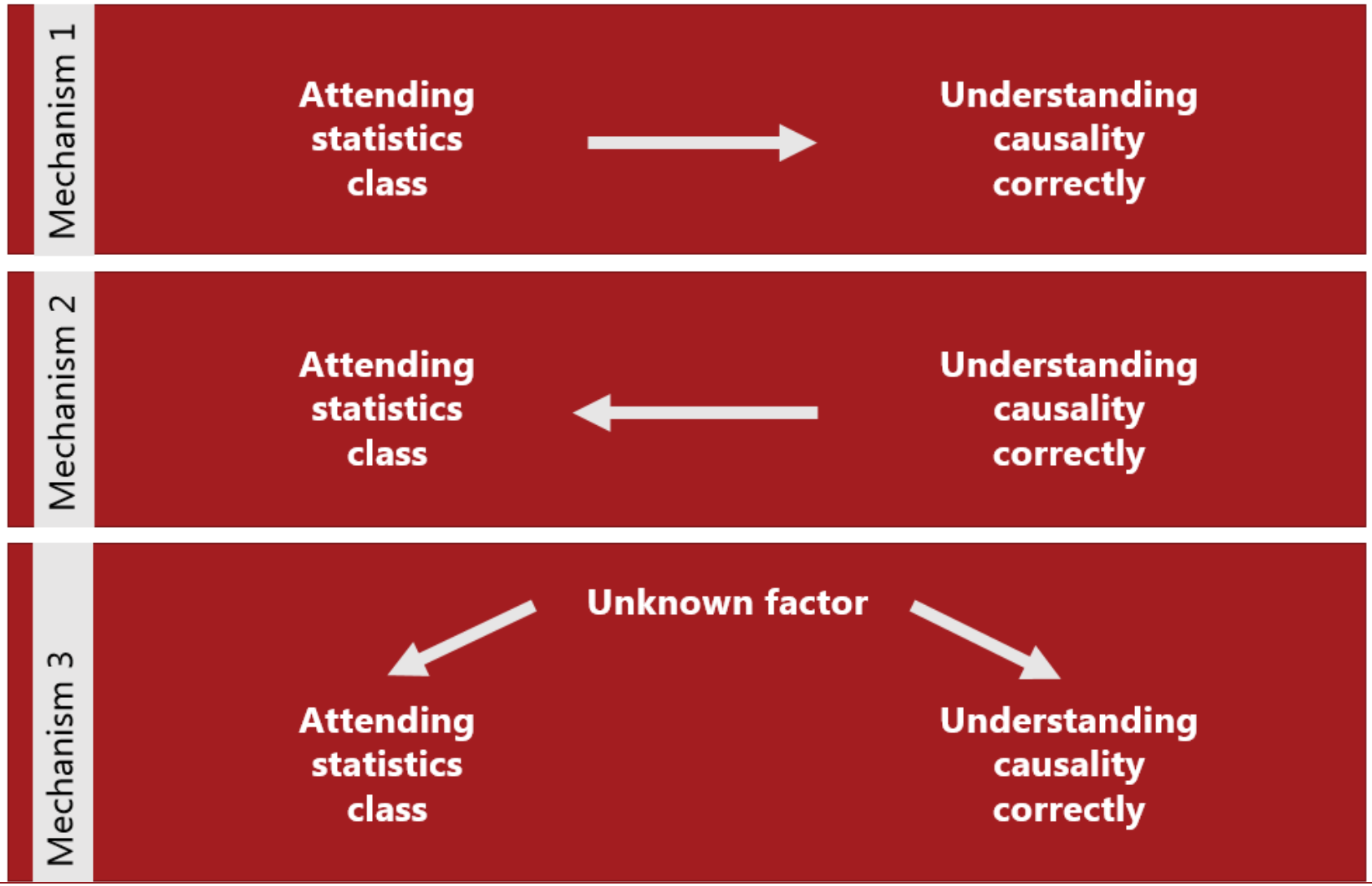
Correlation does **not** imply causation



Source: www.xkcd.com/552/

. . . but causation may imply correlation

Reichenbach's common cause principle: A correlation occurs due to one of three possible mechanisms:



Causal discovery (aka *structure learning*)

Main idea: **Causal relationships leave behind traces in data that can be used to reconstruct (parts of) the causal model.**

Note: This detective work is a matter of **data analysis**.

Causal discovery (aka *structure learning*)

Main idea: **Causal relationships leave behind traces in data that can be used to reconstruct (parts of) the causal model.**

Note: This detective work is a matter of **data analysis**.

Which `R` procedures that can be applied depends on:

- What type of data you have - numerical? Categorical? Mixed?

Causal discovery (aka *structure learning*)

Main idea: **Causal relationships leave behind traces in data that can be used to reconstruct (parts of) the causal model.**

Note: This detective work is a matter of **data analysis**.

Which `R` procedures that can be applied depends on:

- What type of data you have - numerical? Categorical? Mixed?
- What you are willing to assume about the data generating mechanism

Causal discovery (aka *structure learning*)

Main idea: **Causal relationships leave behind traces in data that can be used to reconstruct (parts of) the causal model.**

Note: This detective work is a matter of **data analysis**.

Which `R` procedures that can be applied depends on:

- What type of data you have - numerical? Categorical? Mixed?
- What you are willing to assume about the data generating mechanism
- What is feasible for your data size

Causal discovery (aka *structure learning*)

Main idea: **Causal relationships leave behind traces in data that can be used to reconstruct (parts of) the causal model.**

Note: This detective work is a matter of **data analysis**.

Which `R` procedures that can be applied depends on:

- What type of data you have - numerical? Categorical? Mixed?
- What you are willing to assume about the data generating mechanism
- What is feasible for your data size
- What is missing in your data - observations? Full variables?

Causal discovery (aka *structure learning*)

Main idea: **Causal relationships leave behind traces in data that can be used to reconstruct (parts of) the causal model.**

Note: This detective work is a matter of **data analysis**.

Which `R` procedures that can be applied depends on:

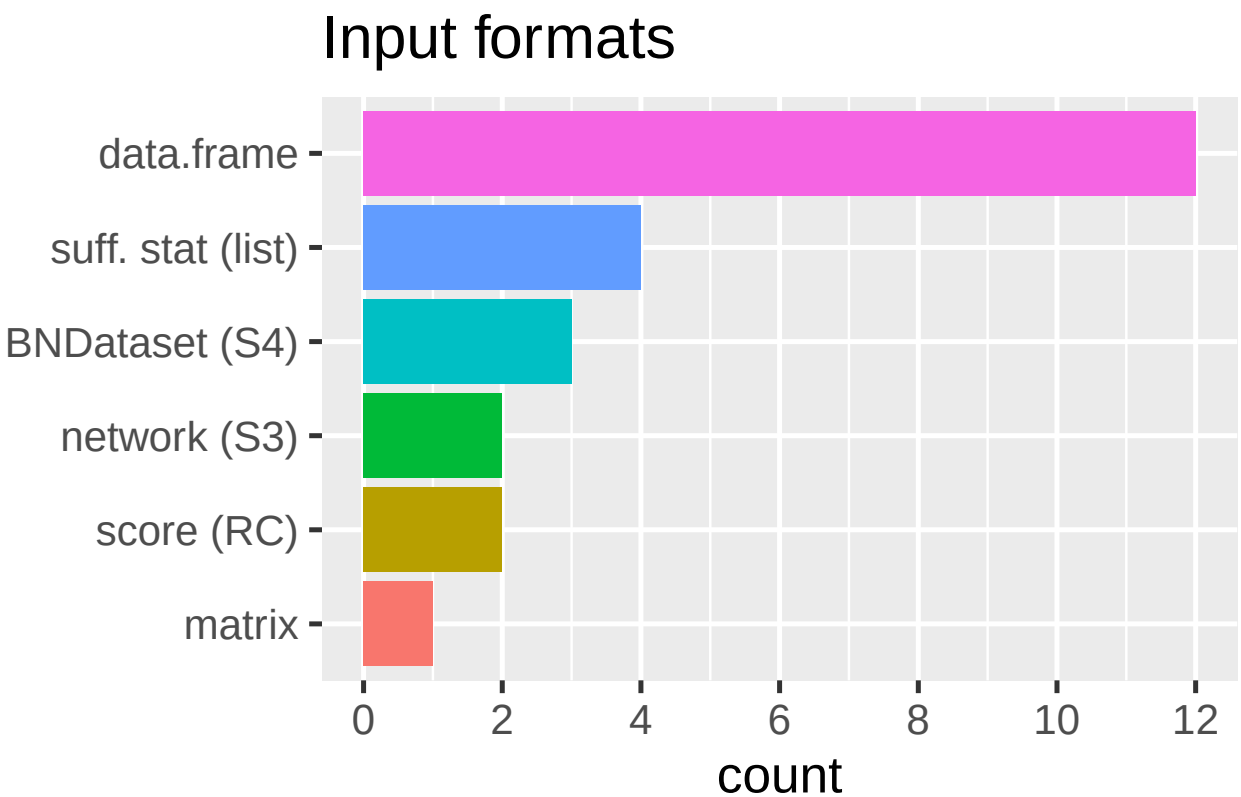
- What type of data you have - numerical? Categorical? Mixed?
- What you are willing to assume about the data generating mechanism
- What is feasible for your data size
- What is missing in your data - observations? Full variables?
- ...

Causal discovery in R

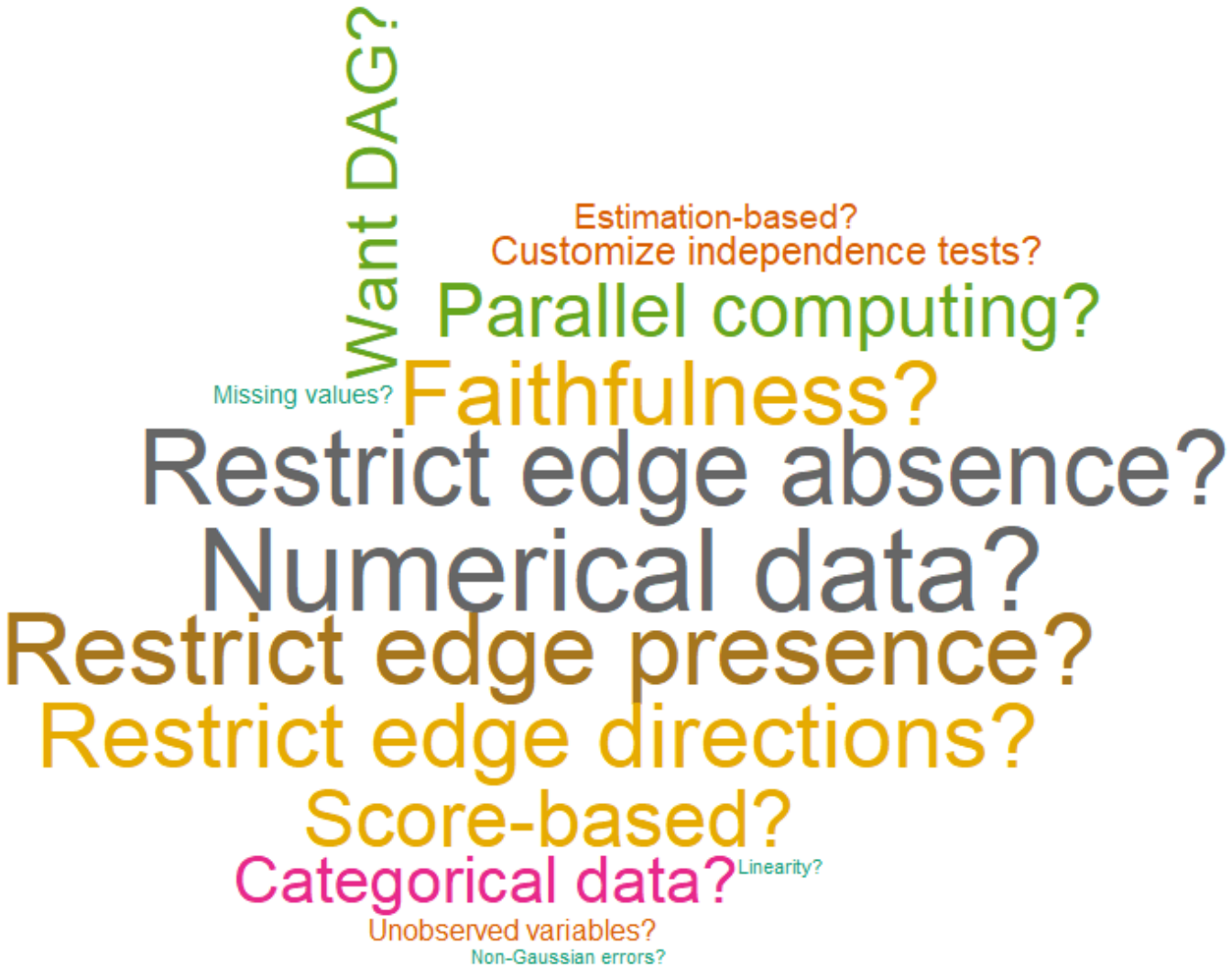
- I have looked at 24 causal discovery procedures from 6 different packages: `pcalg`, `bnstruct`, `bnlearn`, `catnet`, `stablespec`, `deal`.
- Each procedure classified according to 14 properties.
- Minimal code example and description for each procedure.

Causal discovery in R

- I have looked at 24 causal discovery procedures from 6 different packages: `pcalg`, `bnstruct`, `bnlearn`, `catnet`, `stablespec`, `deal`.
- Each procedure classified according to 14 properties.
- Minimal code example and description for each procedure.



Asking the right questions



Getting a proper overview of the answers

causalDisco 

biostatistics.dk/causaldisco

Two restrictions:

1. Only consider procedures for *observational* data
2. Only consider procedures for *acyclic* models

The causalDisco web tool

causalDisco

Choose properties

- [ ] Assume faithfulness?
- [ ] Allow for unobserved variables?
- [ ] Allow for external restrictions on edge presence?
- [ ] Allow for external restrictions on edge absence?
- [ ] Allow for custom independence tests?
- [ ] Allow for external restrictions on edge directions?
- [ ] Score-based approach?
- [x] Constraint-based approach?
- [ ] Assume linearity?
- [ ] Assume non-Gaussian errors?
- [ ] Estimation-based approach (complete identification)?
- [ ] Support missing information?
- [x] Support numerical data?
- [ ] Support categorical data?
- [ ] Return DAG (all edges directed)?
- [ ] Support parallel computing?

Available procedures:
pcalg::pc pcalg::fci pcalg::rfci pcalg::fciPlus pgalg::ges as ARGES bnlearn::gs

Procedure information | About the data | About the tool

pcalg::pc

Constraint-based learning using the PC algorithm

Package: pcalg
Function: pc
Input: sufficient statistic
Output: S4 object pcAlgo
Documentation: https://cran.r-project.org/web/packages/pcalg/vignettes/vignette2018.pdf
Article: https://www.jstatsoft.org/article/view/v047i11
Note:

- Defaults to the stable version of the algorithm (i.e. less order dependence).
- User-supplied restrictions on edge orientation is only possible after the structure has been learned, and afterwards, the algorithmic edge orientation step may be repeated.

Minimal code example:

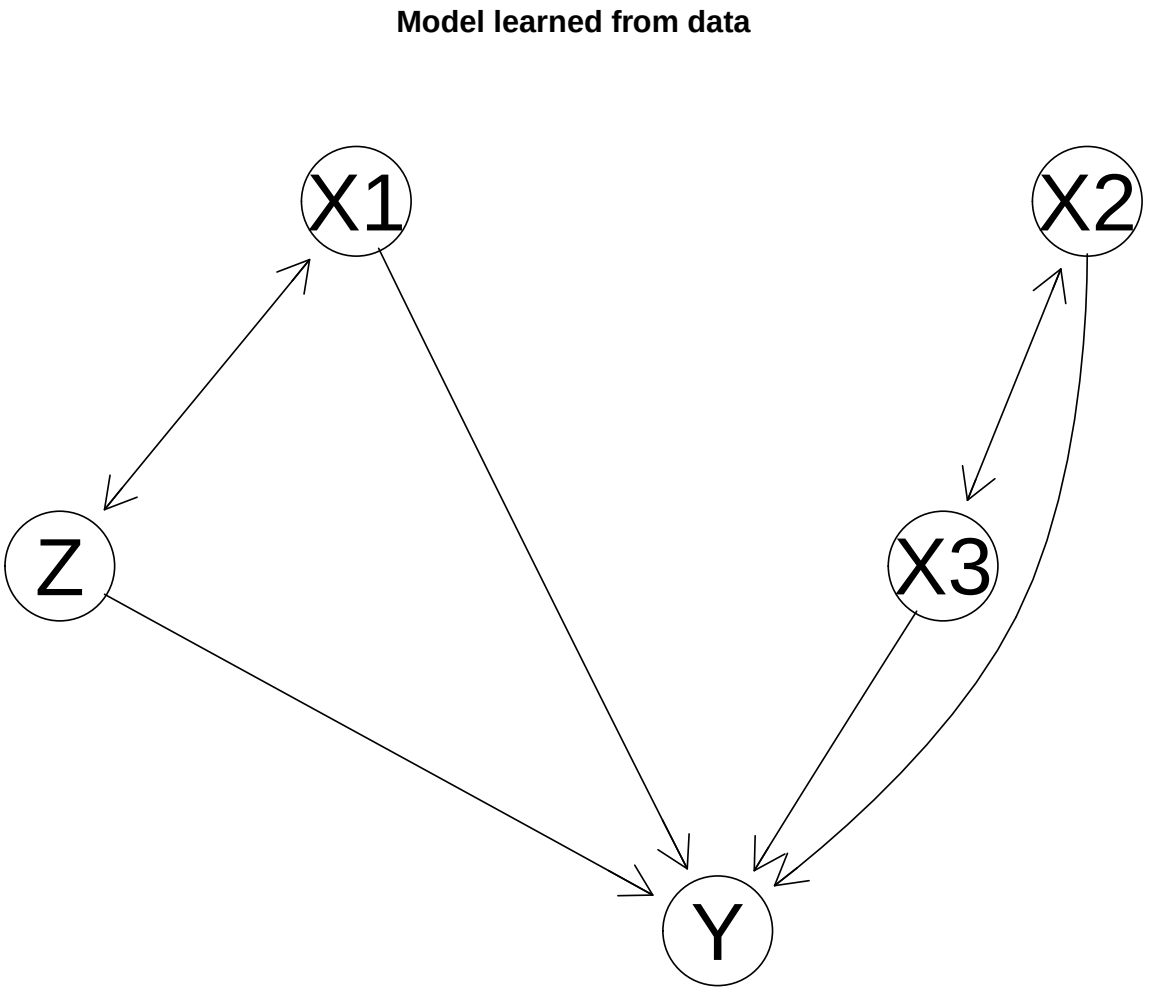
```
#Load numeric dataset numData
load(url("https://github.com/annennenne/causalDisco/raw/master/data/exampledata_numData.rda"
))

#Load package
library(pcalg)

#Prepare data for pc() call
##Note: this choice of sufficient statistics is valid for Gaussian data.
pcalg_suffstat_numData <- list(C = cor(numData), n = nrow(numData))
```

Learning the structure of numData

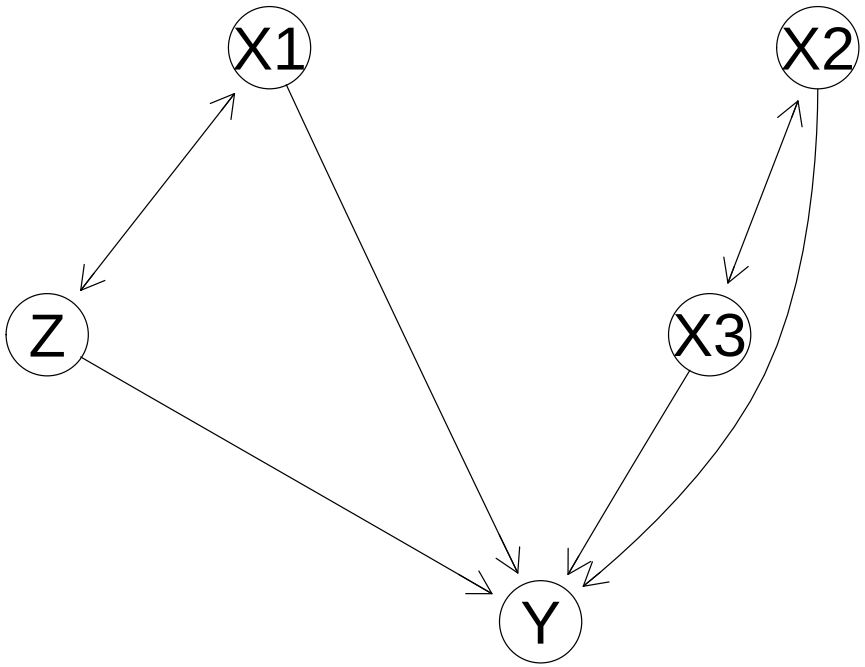
```
load(url(paste(
  "https://github.com/annennenne/causalDisco/",
  "raw/master/data/exampledata_numData.rda",
  sep = "")
  ))

library(pcalg)

pcalg_suffstat_numData <- list(C = cor(numData),
                               n = nrow(numData))

pcalg_pc_out <- pc(pcalg_suffstat_numData,
                   labels = names(numData),
                   indepTest = gaussCItest,
                   alpha = 0.01)
```

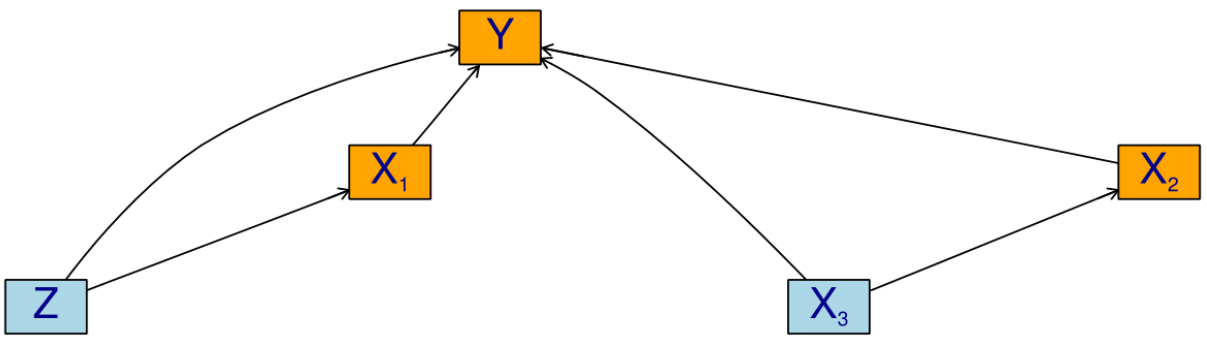
Look at the model graph

```
plot(pcalg_pc_out, main = "Model learned from data")
```



Compare with true model





Directions for future work

- **Crowdsourcing**: Make it easy for users to report – and for developers to see – what procedures are needed but not yet available
  - Currently missing procedures for: categorical data with unobserved variables, numerical data with missing information, ...
- Implement **one interface** for all available methods
  - Allow for hybrid queries combining methods from several backends
  - Allow for dynamic manipulation of assumptions

Thank you!