Acting Data-Driven - But How?

Karsten Lübke, Matthias Gehrke (FOM)

Jörg Horst (FH Bielfeld)

Sebastian Sauer (HS Ansbach)

Gero Szepannek (HS Stralsund)

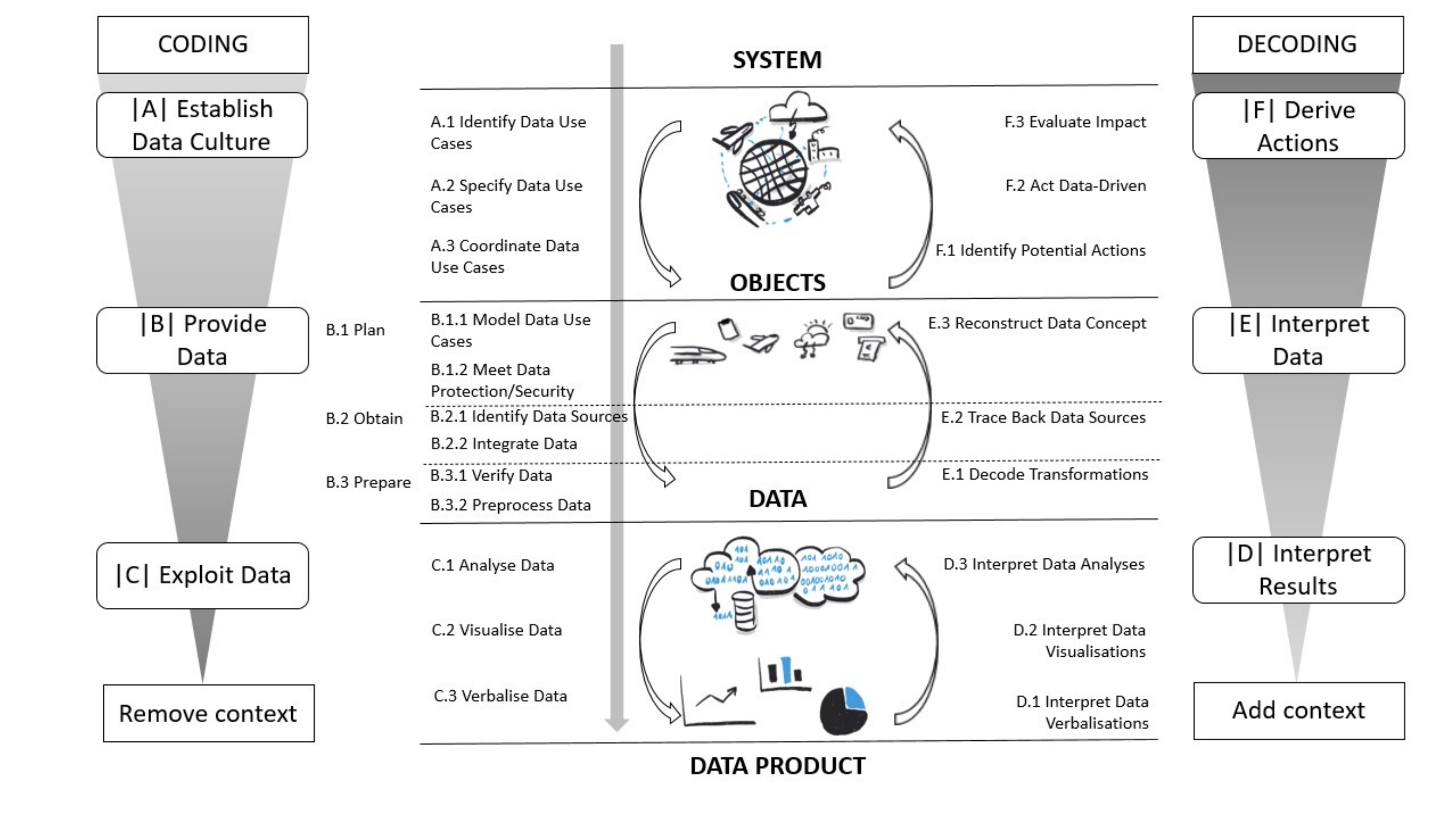
ECDA 2022

Intro

Inference from data analysis

Please take part in a (very) short survey: https://bit.ly/30sJNbm

Data Literacy



Source: Schüller (2020), cf. Data Literacy Charter

A wobbly bridge

From *A1: Data Use Case* to *F2: Act Data-Driven*:

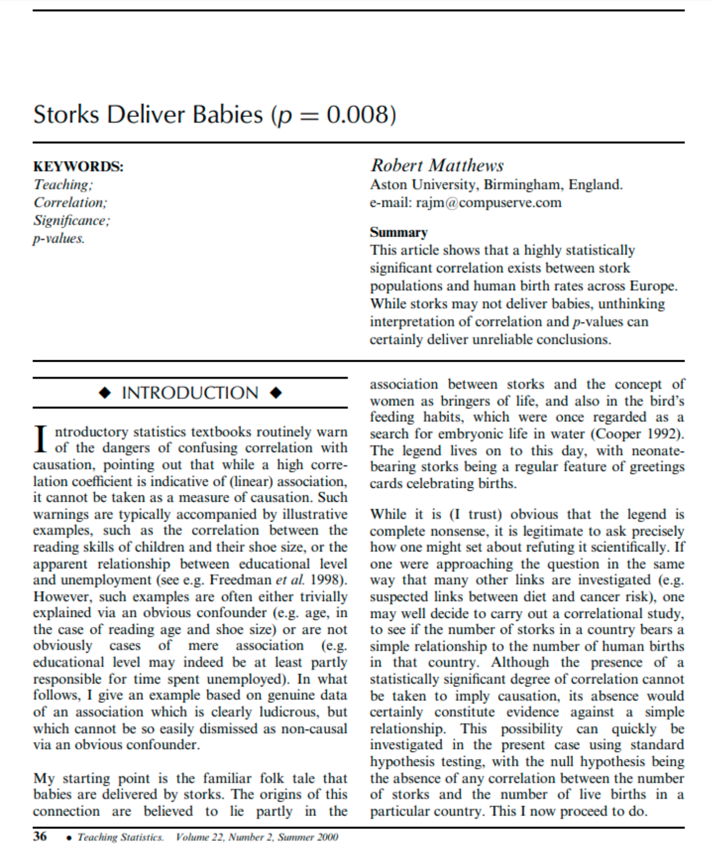
via GIPHY

Data science tasks

Hernán et al. (2019) distinguish:

- **Description**: "How can women aged 60–80 years with stroke history be partitioned in classes defined by their characteristics?"
- **Prediction**: "What is the probability of having a stroke next year for women with certain characteristics?"
- **Causal inference**: "Will starting a statin reduce, on average, the risk of stroke in women with certain characteristics?"

The challenge



Storks Deliver Babies ($p = 0.008$)

KEYWORDS:
Teaching;
Correlation;
Significance;
p-values.

Robert Matthews
Aston University, Birmingham, England.
e-mail: rajm@compuserve.com

Summary
This article shows that a highly statistically significant correlation exists between stork populations and human birth rates across Europe. While storks may not deliver babies, unthinking interpretation of correlation and *p*-values can certainly deliver unreliable conclusions.

◆ INTRODUCTION ◆

Introductory statistics textbooks routinely warn of the dangers of confusing correlation with causation, pointing out that while a high correlation coefficient is indicative of (linear) association, it cannot be taken as a measure of causation. Such warnings are typically accompanied by illustrative examples, such as the correlation between the reading skills of children and their shoe size, or the apparent relationship between educational level and unemployment (see e.g. Freedman *et al.* 1998). However, such examples are often either trivially explained via an obvious confounder (e.g. age, in the case of reading age and shoe size) or are not obviously cases of mere association (e.g. educational level may indeed be at least partly responsible for time spent unemployed). In what follows, I give an example based on genuine data of an association which is clearly ludicrous, but which cannot be so easily dismissed as non-causal via an obvious confounder.

My starting point is the familiar folk tale that babies are delivered by storks. The origins of this connection are believed to lie partly in the association between storks and the concept of women as bringers of life, and also in the bird's feeding habits, which were once regarded as a search for embryonic life in water (Cooper 1992). The legend lives on to this day, with neonate-bearing storks being a regular feature of greetings cards celebrating births.

While it is (I trust) obvious that the legend is complete nonsense, it is legitimate to ask precisely how one might set about refuting it scientifically. If one were approaching the question in the same way that many other links are investigated (e.g. suspected links between diet and cancer risk), one may well decide to carry out a correlational study, to see if the number of storks in a country bears a simple relationship to the number of human births in that country. Although the presence of a statistically significant degree of correlation cannot be taken to imply causation, its absence would certainly constitute evidence against a simple relationship. This possibility can quickly be investigated in the present case using standard hypothesis testing, with the null hypothesis being the absence of any correlation between the number of storks and the number of live births in a particular country. This I now proceed to do.

How can we be sure that no human or artificial intelligence does not start colonizing storks to increase birth rate?

Results

Back to the Survey: What is inferred?

Structual causal model for data in survey question:

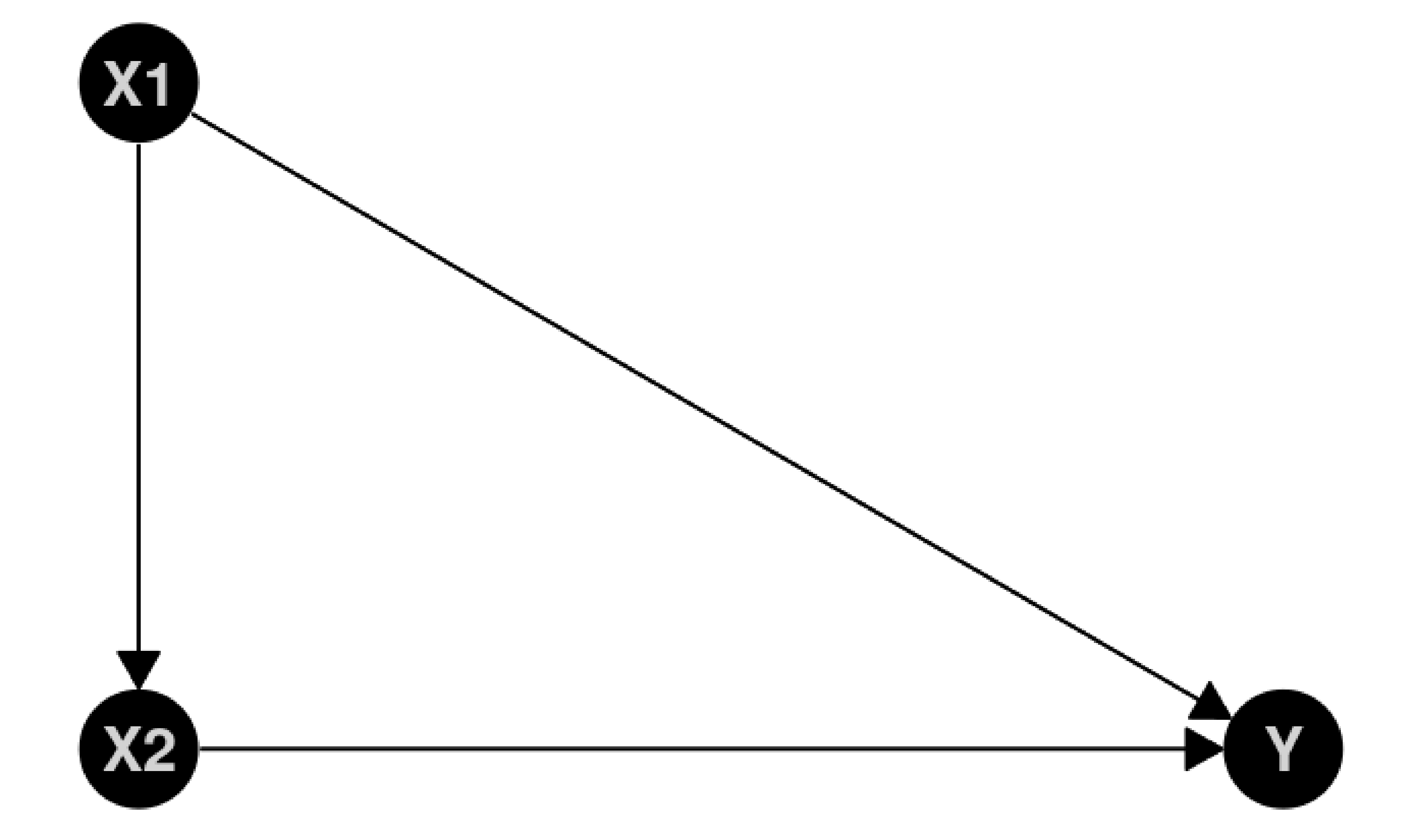
$$X_1 = U_{X_1},\ U_{X_1} \sim \mathcal{N}(0,\, 10), \quad X_2 = -2X_1 + U_{X_2},\ U_{X_2} \sim \mathcal{N}(0,\, 1),$$

$$Y = 5X_1 + X_2 + U_Y, \quad U_Y \sim \mathcal{N}(0,\, 5).$$

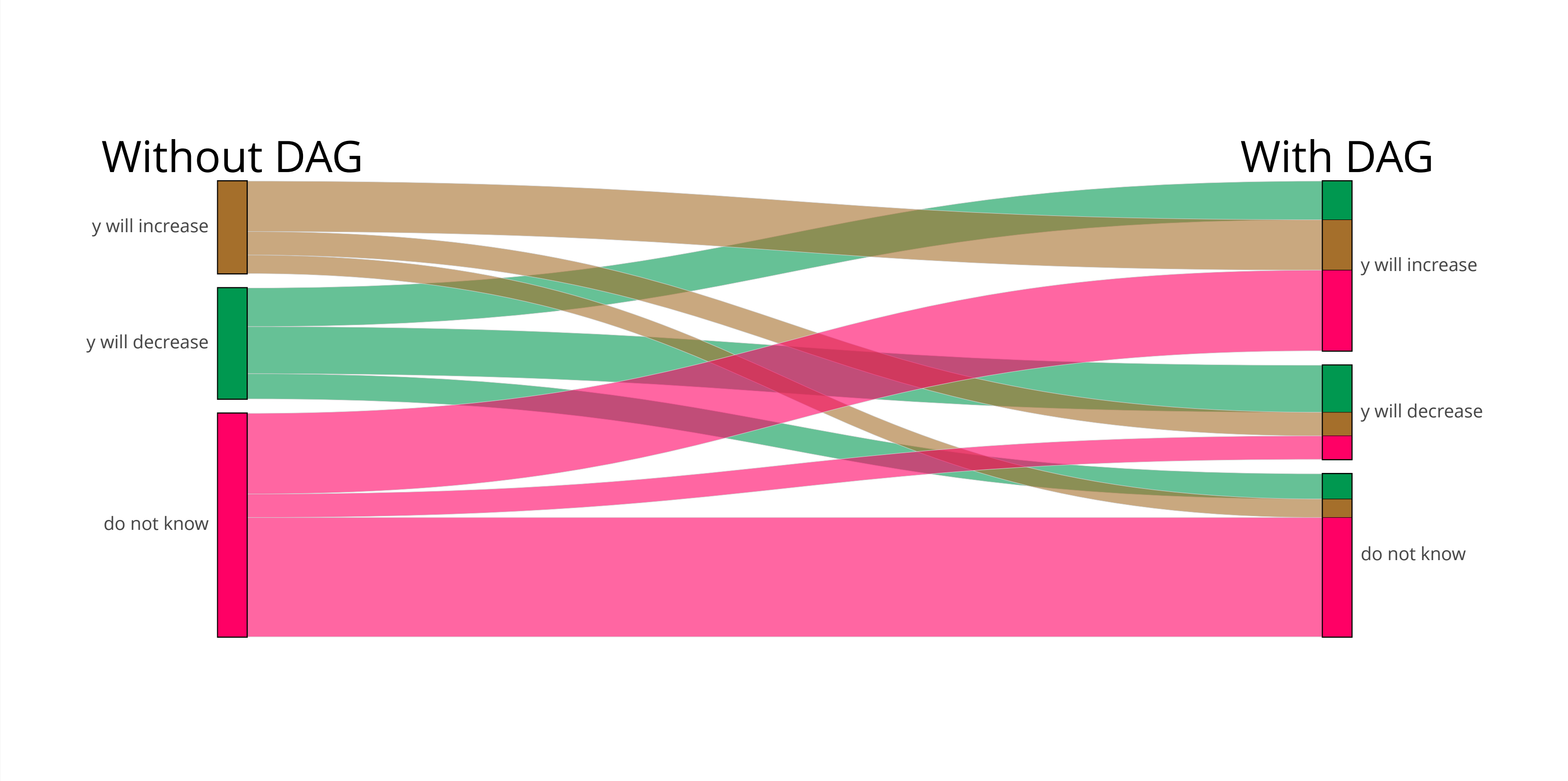
Based on linear regression result:

- $\hat{\beta}_2^{(1)} = -1.505$ (excluding x_1)
- $\hat{\beta}_2^{(2)} = 0.909$ (including x_1)

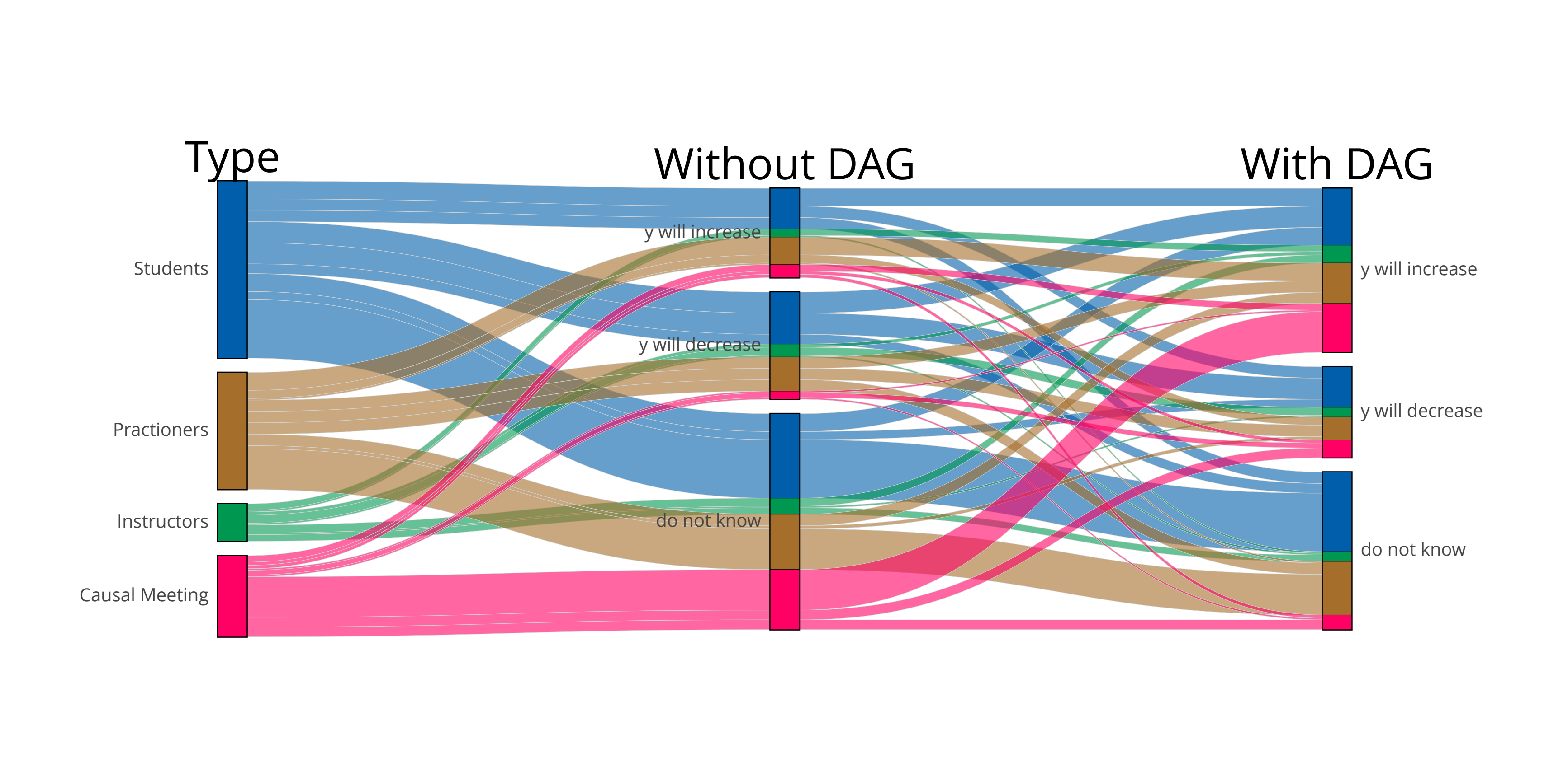
DAG



Alluvial diagram



Alluvial diagram - grouped



Numerical summary

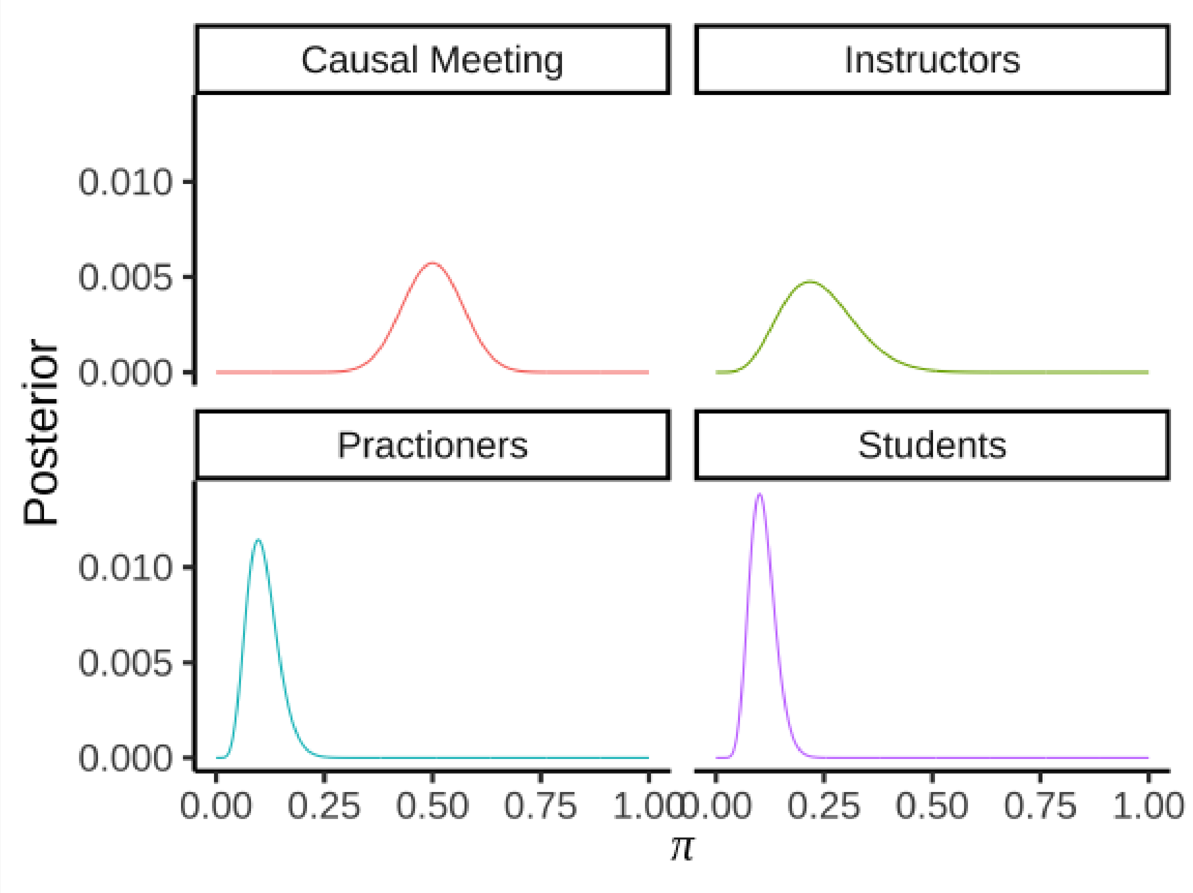
Correct on both answers:

Type	n	p.correct
Causal Meeting	50	0.500
Instructors	23	0.217
Practioners	72	0.097
Students	109	0.101

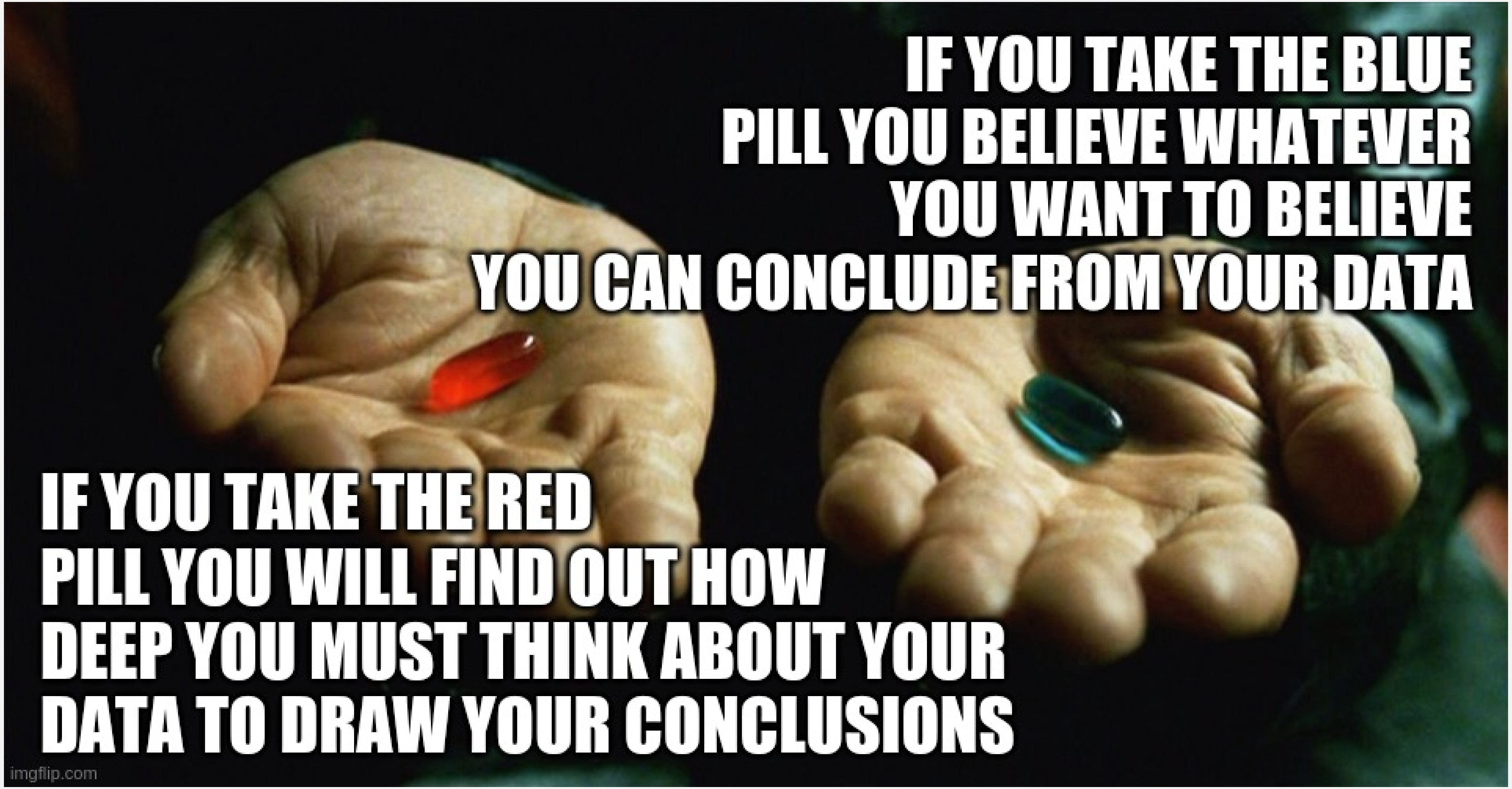
Freuqentist inference

- For the aggregated data the result is with a p-value of 5.9322569×10^{-5} statistically discernible $> 1/9$.
- With a p-value of 5.6886604×10^{-8} there are statistically discernible differences between the groups.

Baysian analysis (uniform prior)



Outro


IF YOU TAKE THE BLUE PILL YOU BELIEVE WHATEVER YOU WANT TO BELIEVE YOU CAN CONCLUDE FROM YOUR DATA
IF YOU TAKE THE RED PILL YOU WILL FIND OUT HOW DEEP YOU MUST THINK ABOUT YOUR DATA TO DRAW YOUR CONCLUSIONS
imgflip.com

If you've just woken up

Acting Data-Driven - But How?

👉 Far too many draw incorrect conlusions from data analysis. Data analysis skills are not enough to avoid drowning in the data. Integration of DAGs in data science education may be a step in that direction. More research is needed.

The wrong lesson

Danny Kaplan:

> *What I was saying* ... Data don't speak, they inform our judgment. Interpret data in the context of a whole system.
>
> *What they were hearing* ... The data will say anything you want, depending on how you cut it.

How can we provide a framework to discuss science with data with all stakeholders?

The End

❤ Thank you for your participation ❤

- ✉: karsten.luebke@fom.de