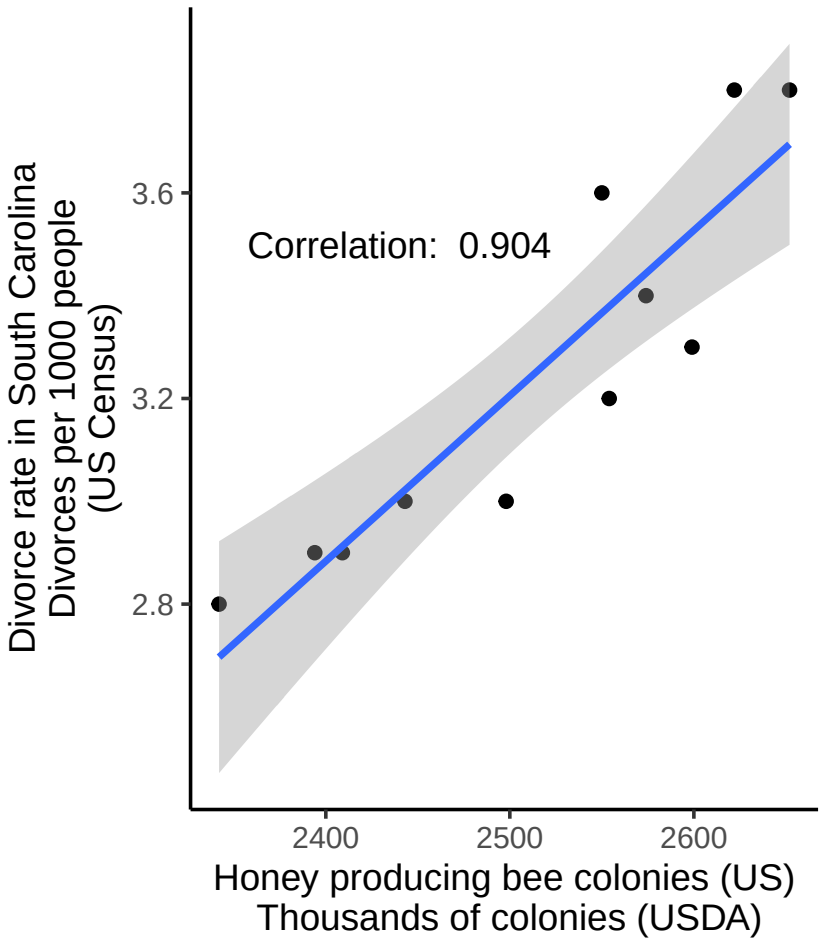
Causality, Dependency, Correlation, and Designed Experiments

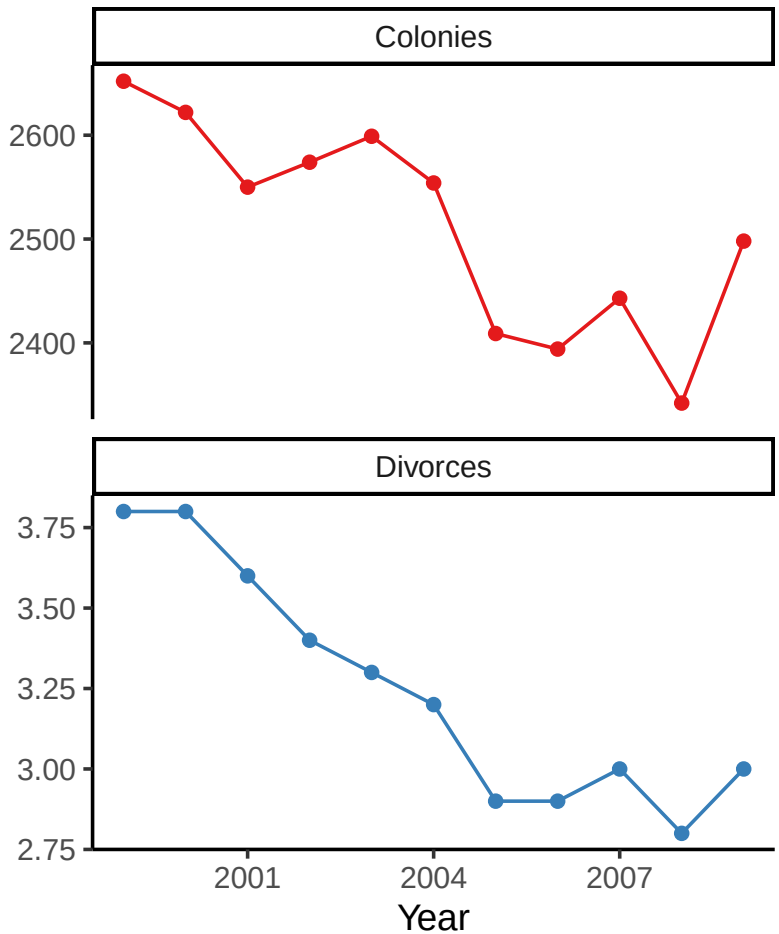
Arnaud Legrand and Jean-Marc Vincent

Scientific Methodology and Performance Evaluation

Outline

Correlation and causation





Source: *Spurious correlations*. For the good of the US society, we should try to get rid of honey bees 😉

Typical causal questions:

Causal analysis is the field of experimental design and statistics pertaining to establishing cause and effect:

- Did the fertilizer cause the crops to grow?
- Can a given sickness be prevented?
- Why is my friend depressed?

Causality can be construed from counterfactual states

What would have happened otherwise ?

Alternative Universes

For example, one could run an experiment on identical twins who were known to consistently get the same grades on their tests

- One twin is sent to study for six hours
- The other is sent to the amusement park
- If their test scores suddenly diverged by a large degree, this would be strong evidence that studying (or going to the amusement park) had a causal effect on test scores
  In this case, correlation between studying and test scores would almost certainly imply causation

This is unfortunately hard to implement

Well Designed Experiments as an Alternative

Well-designed experimental studies replace equality of individuals as in the previous example by equality of groups

Randomized controlled trial: Build two groups that are similar except for the treatment that the groups receive

- Select subjects from a single population and randomly assigning them to two or more groups
- The likelihood of the groups behaving similarly to one another (on average) rises with the number of subjects in each group.
- If (the groups are essentially equivalent except for the treatment they receive), and (a difference in the outcome for the groups is observed), then this constitutes evidence that the treatment is responsible for the outcome

Note: An observed effect could also be caused "by chance", for example as a result of random perturbations in the population but this is what statistical tests are meant for

Causation from Correlation

- Regression analysis techniques handle such queries when data is collected using designed experiments
- Otherwise, they don't. Why ?

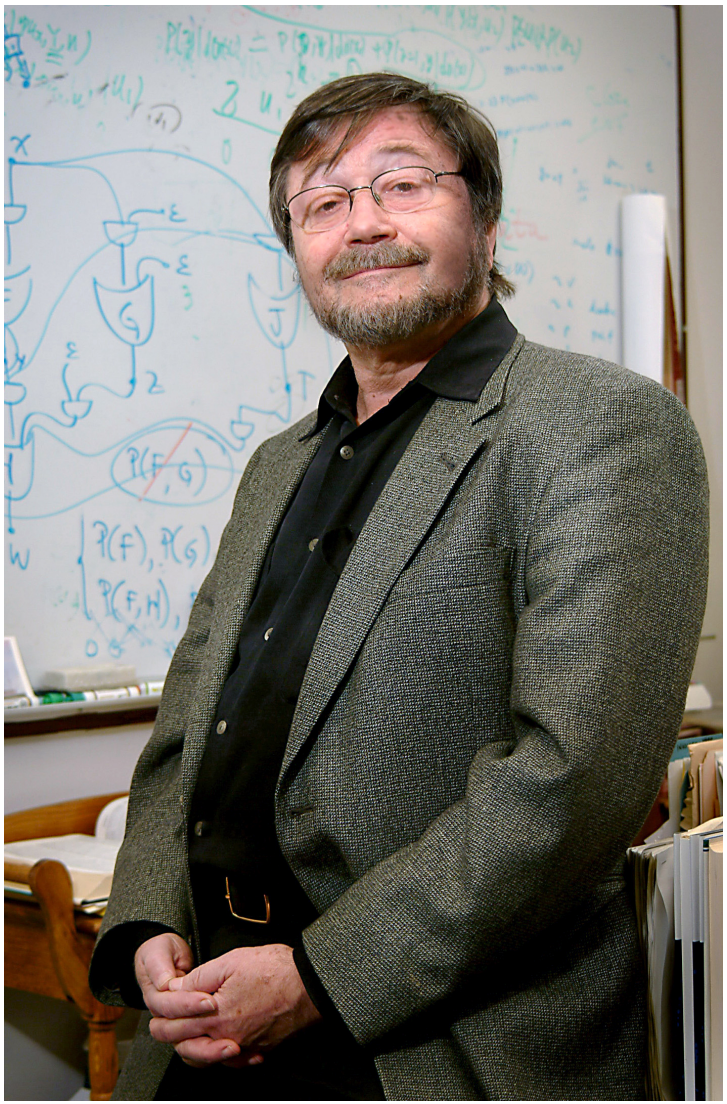
Causation from Correlation

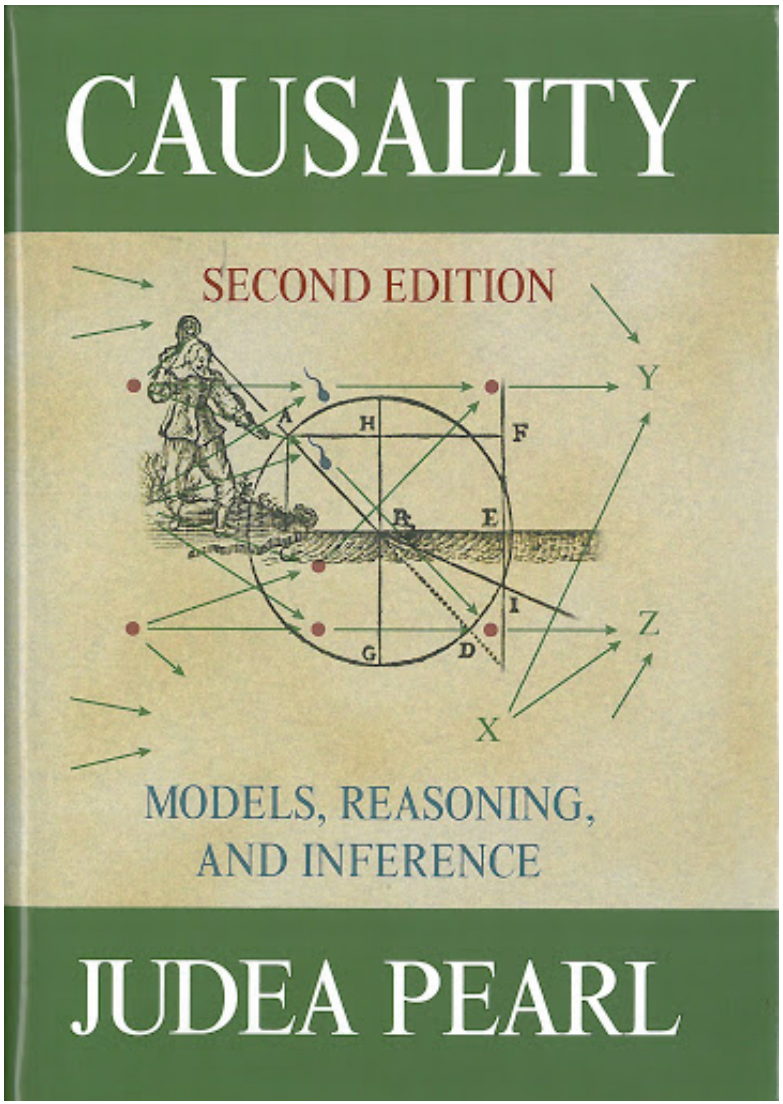
- Regression analysis techniques handle such queries when data is collected using designed experiments
- Otherwise, they don't. Why ?
  - Confounding variable
    - Hot Temperature $\rightarrow$ Rate of Ice cream consumption
    - Hot Temperature $\rightarrow$ Number of sunburns
    - Concluding "Rate of Ice cream consumption $\rightarrow$ Number of sunburns" would be silly
  - Conditioning does not work!
    - E[Life | Smoking] and E[Life | !Smoking] restrict to both categories
    - E[Life | do(!Smoking)] is about the entire population

Causation from Correlation

- Regression analysis techniques handle such queries when data is collected using designed experiments
- Otherwise, they don't. Why ?
  - Confounding variable
    - Hot Temperature $\rightarrow$ Rate of Ice cream consumption
    - Hot Temperature $\rightarrow$ Number of sunburns
    - Concluding "Rate of Ice cream consumption $\rightarrow$ Number of sunburns" would be silly
  - Conditioning does not work!
    - E[Life | Smoking] and E[Life | !Smoking] restrict to both categories
    - E[Life | do(!Smoking)] is about the entire population
- Data collected in observational studies require different techniques for causal inference
  - It requires to list possible explanations (variables and their relations)
  - This *ad hoc* model can then be used to decide whether causal relations can be inferred or not
    - The more variables and the more connected the variables, the harder
  - It can guide you toward which experiments to conduct

Judea Pearl (2011 Turing Award)





Outline

1. Causal analysis

2. Machine Learning Counterfactual Explanations

A few motivating examples (1/2)

From A Guide for Making Black Box Models Explainable

Peter applies for a loan and gets rejected by the (machine learning powered) banking software. He wonders why his application was rejected and how he might improve his chances to get a loan.

- The question of "why" can be formulated as a counterfactual
  - What is the smallest change to the features (income, number of credit cards, age, ...) that would change the prediction from rejected to approved?
- One possible answer could be: If Peter would earn 10,000 Euro more per year, he would get the loan.
- Or if Peter had fewer credit cards and had not defaulted on a loan 5 years ago, he would get the loan.

A few motivating examples (2/2)

Anna wants to rent out her apartment, but she is not sure how much to charge for it

- So she decides to train a machine learning model to predict the rent. Of course, since Anna is a data scientist, that is how she solves her problems 😉
- After entering all the details about size, location, whether pets are allowed and so on, the model tells her that she can charge 900 Euro.

A few motivating examples (2/2)

Anna wants to rent out her apartment, but she is not sure how much to charge for it

- So she decides to train a machine learning model to predict the rent. Of course, since Anna is a data scientist, that is how she solves her problems 😉
- After entering all the details about size, location, whether pets are allowed and so on, the model tells her that she can charge 900 Euro.
- She expected 1000 Euro or more, but she trusts her model and decides to play with the feature values of the apartment to see how she can improve the value of the apartment.
  - She finds out that the apartment could be rented out for over 1000 Euro, if it were 15 m2 larger. Interesting, but non-actionable knowledge, because she cannot enlarge her apartment.
  - Finally, by tweaking only the feature values under her control (built-in kitchen yes/no, pets allowed yes/no, type of floor, etc.), she finds out that if she allows pets and installs windows with better insulation, she can charge 1000 Euro.

Anna intuitively works with counterfactuals to change the outcome

Counterfactual "Explanation"

Counterfactual explanation of a prediction = the smallest change to the input values that changes the prediction to a predefined output

- Counterfactuals are human-friendly explanations, because they are contrastive to the current instance and because they are selective (focus on a small number of changes) related with Occam's razor
- But counterfactuals suffer from the *Rashomon* effect
  - *Rashomon* is a Japanese movie in which the murder of a Samurai is told by different people. Each of the stories explains the outcome equally well, but the stories contradict each other.
  - The same can also happen with counterfactuals, since there are usually multiple different counterfactual explanations.
  - Each counterfactual tells a different "story" of how a certain outcome was reached
    - One counterfactual might say to change feature A
    - The other counterfactual might say to leave A the same but change feature B, which is a "contradiction"

Conclusion

- This is still open research (best/simplest model/explanation, causality)
  - Applying Machine Learning counterfactuals to reality (like Anna) only makes sense if you have a very faithful model

- Modeling is at the heart of the scientific methodology
  - If you have (1) a sound question to answer and (2) a model of reality, you will know how to conduct your experiments and do the statistical analysis