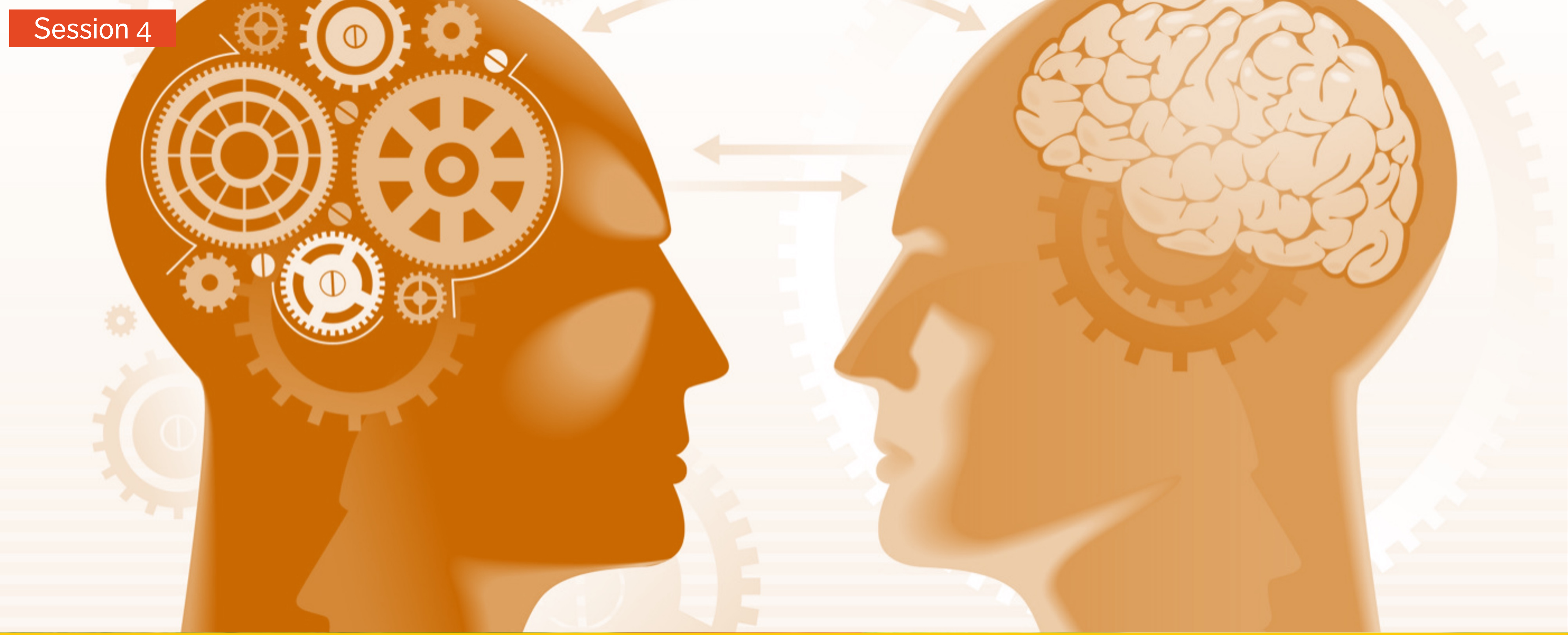


12. Miscellanea

Biases, Fallacies, and Interpretation

When consulting (or conducting) studies, beware:

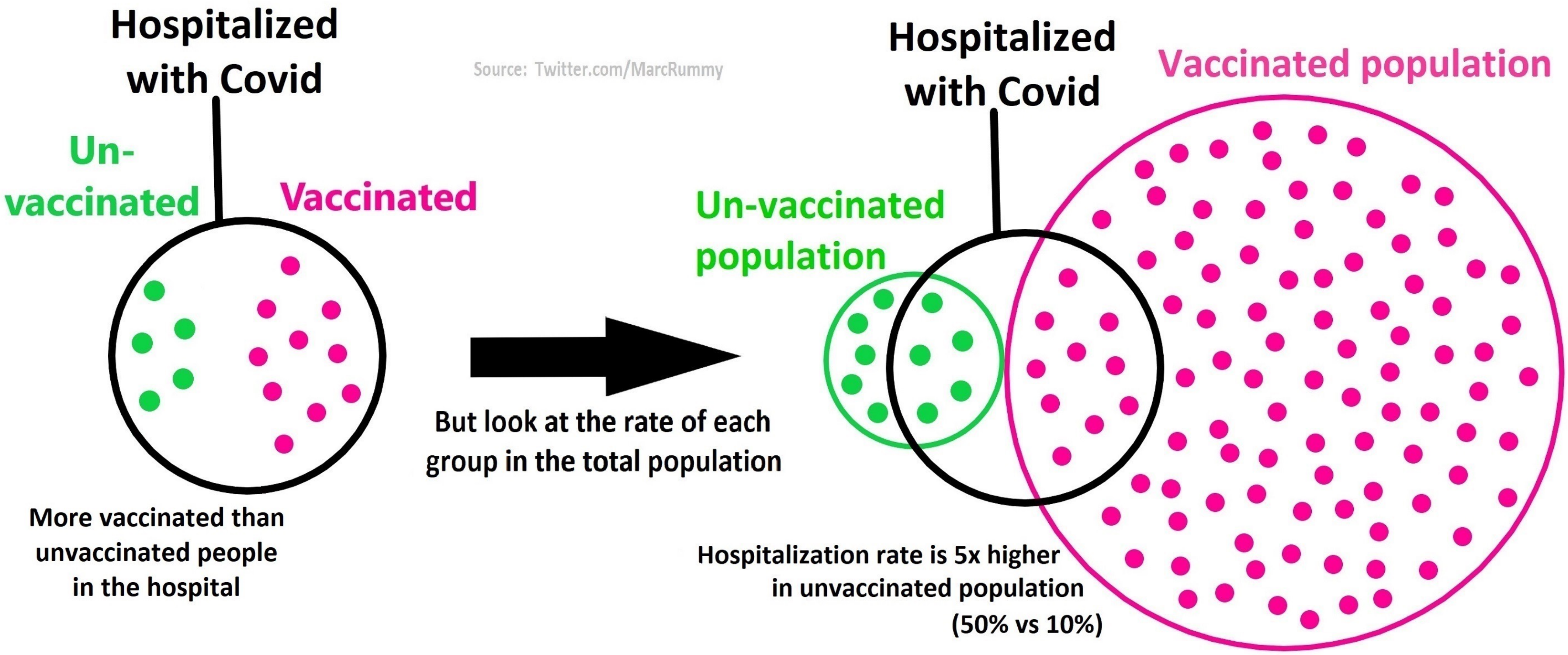
- **selection bias** (what data was included, how was it selected?)
- **omitted-variable bias** (were relevant variables ignored?)
- **detection bias** (did prior knowledge affect the results?)
- **funding bias** (who's paying for this?)
- **publication bias** (what's not being published?)
- **data-snooping bias** (trying too hard?)
- **analytical bias** (did the choice of specific method affect the results?)
- **exclusion bias** (are specific observations/units being excluded?)

But: does the presence of bias necessarily invalidate the results?

Biases, Fallacies, and Interpretation

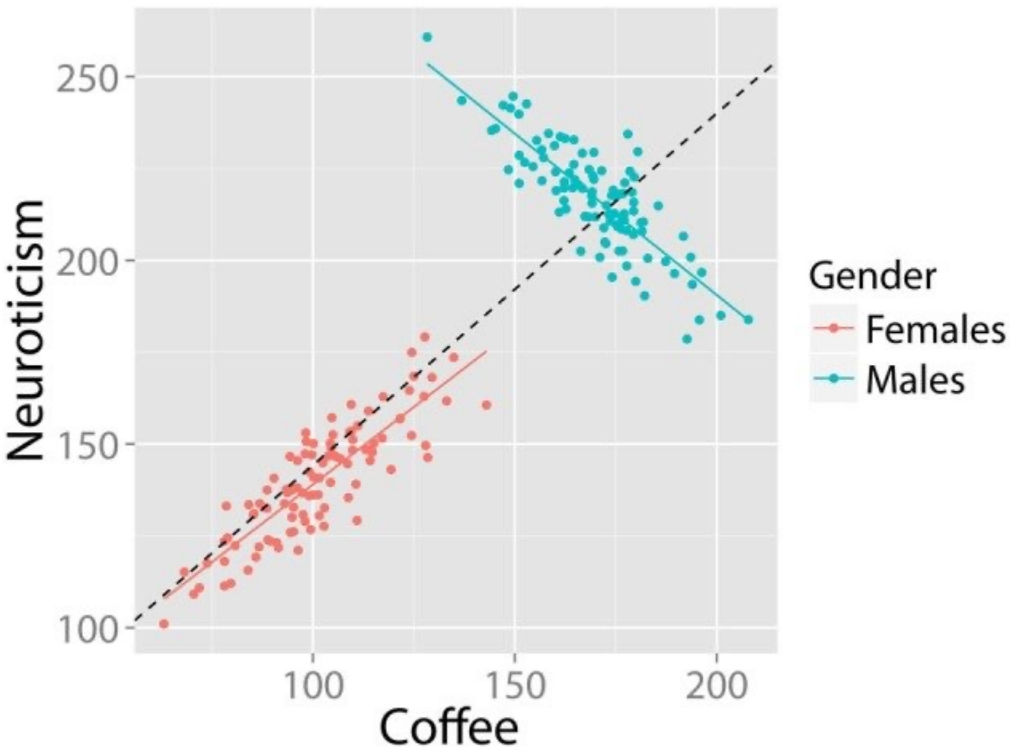
Remember:

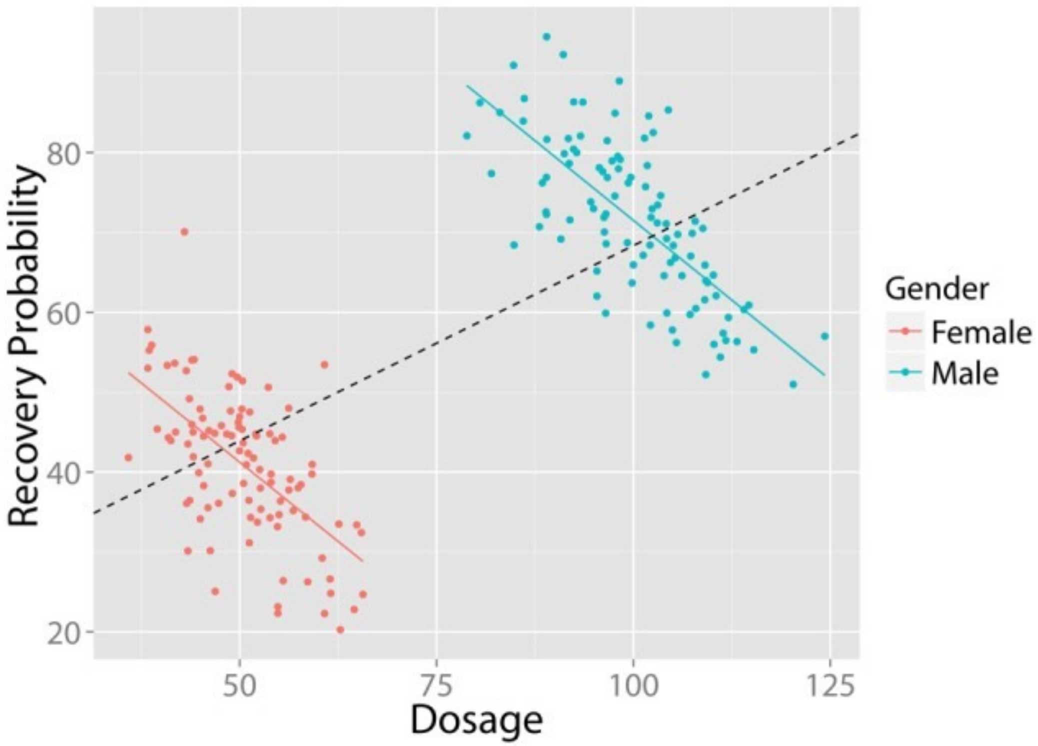
- correlation is not causation (but it is a hint!)
- extreme patterns can mislead
- stay within a study's range
- keep the base rate in mind
- counter-intuitive results are not always wrong (Simpson's Paradox, Benford’s Law, etc.)
- randomness plays a role
- there is a human component to any analytical activity
- small effects can still be (statistically) significant
- beware of sacrosanct statistics (p-value, etc.)



Note: The ratios presented are made to illustrate the concept of the base rate fallacy when the vaccination rate is high

[Kievit et al., *Simpson's Paradox in Psychological Science: A Practical Guide*]





DS/ML Myths and Mistakes

Myths:

- DS/ML is about algorithms
- DS/ML is about predictive accuracy
- DS/ML requires data warehouses and fancy infrastructure
- DS/ML requires a large quantity of data
- DS/ML requires technical experts (?)

DS/ML Myths and Mistakes

Mistakes:

- selecting the wrong problem
- getting buried under tons of data without metadata understanding
- not planning the data analysis process
- having insufficient business and domain knowledge
- using incompatible data analysis tools
- using tools that are too specific
- ignoring individual predictions/records in favour of aggregated results
- running out of time
- measuring results differently than the sponsor/stakeholders
- naïvely believing what one's told about the data

The Future of DS/ML/AI

What we didn't talk about:

- tons of classification and clustering algorithms
- recommender systems
- data streams
- bayesian data analysis
- natural language processing and text mining
- feature selection and dimension reduction (curse of dimensionality)
- data engineering
- … and much, much more!

The Future of DS/ML/AI

Future tasks:

- self-driving vehicles
- machine translation and language understanding
- detection and prevention of climate and ecosystem disturbances
- automated data science (?!)
- detection and prevention of astronomical catastrophic events
- explainable A.I.

The Future of DS/ML/AI

Future trends:

- new questions
- new tools
- new data sources
- data science as job component
- augmented/swarm intelligence

Session 4

In Conclusion

DS/ML is a team activity.

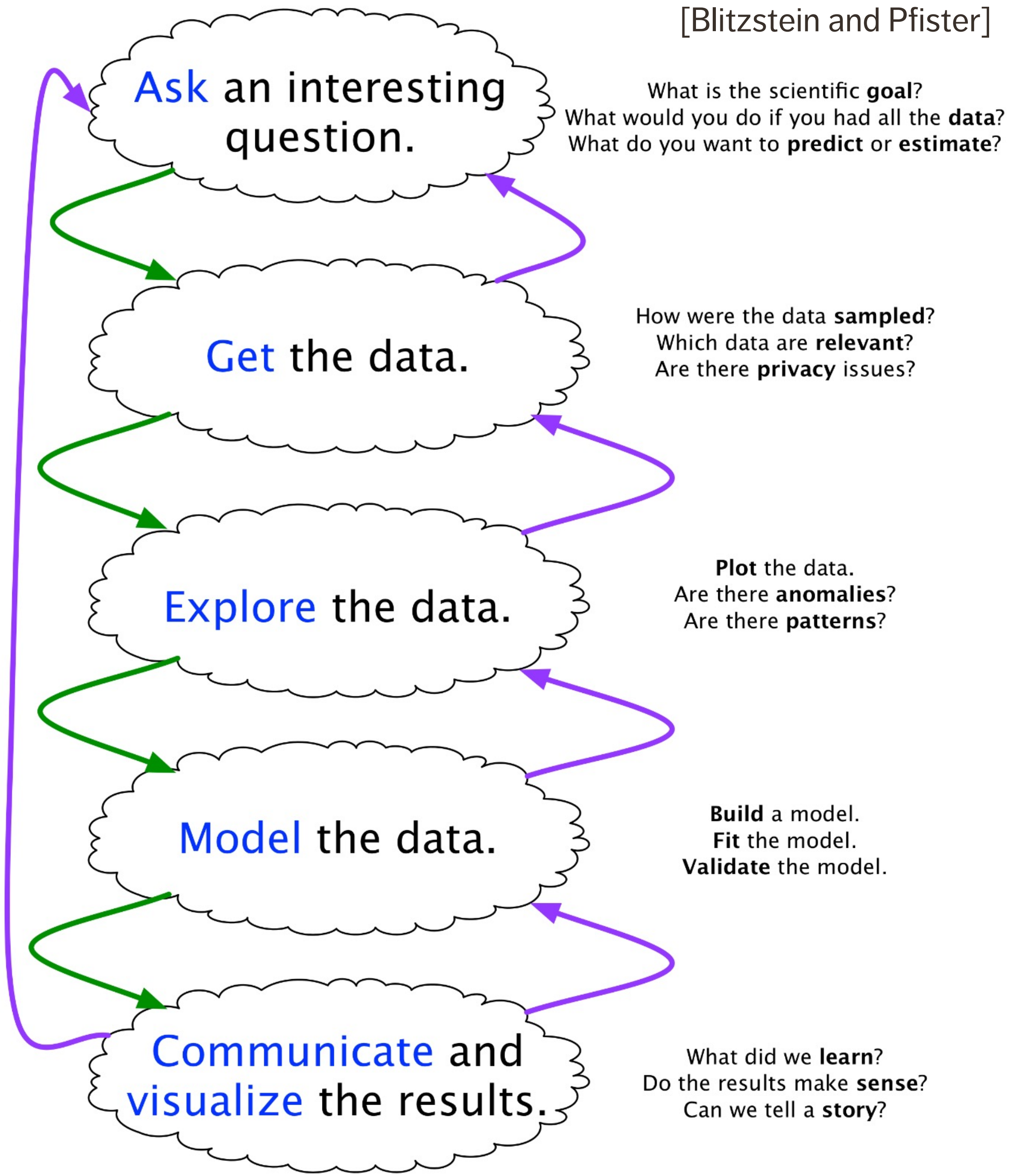
Ethical considerations are crucial.

Let the data speak.

Look for actionable insights.

Supervised vs. unsupervised.

Be ready to clean, prepare, & visualize data.



Exercises

Miscellanea

1. What is your preferred approach: “tried, tested, and true” or “disruptive data science”? What would it take for you to consider the other side of the coin?

2. True or False?
   a. The predictive performance of a supervised model is evaluated on the training set.
   b. Cross-validation can be used to reduce the risk of overfitting a predictive model.
   c. It is always better to use as many variables as possible in a model.
   d. If observations with missing values are deleted, this may lead to bias and errors.
   e. We can use a clustering algorithm to predict class membership.

Exercises

Miscellanea

2. True or False? (cont.)
    f. If all methods don't yield the same result, it is a proof that the question cannot be answered.
    g. Business and domain knowledge is only necessary when working with old data.
    h. Sponsors and clients need to know all analytical details.
    i. It's impossible to plan the data analysis process before we know what the data looks like.
    j. The available data is not always appropriate/representative of the situation we are modeling.

3. In what ways can you see DS/ML becoming a crucial part of your work? Is this development welcomed? How do you want to be involved?