Session 4

Issues and Challenges

INTRODUCTION TO MACHINE LEARNING

We all *say* we like data, but we don’t. We like getting insight out of data. That’s not quite the same as liking data itself. In fact, I dare say that I don’t care for data, and it sounds like I’m not alone. [Q.E. McCallum, *Bad Data Handbook*]

Data, big or small, is only as useful as the questions you ask of it. [M. Jones, P. Silberzahn]

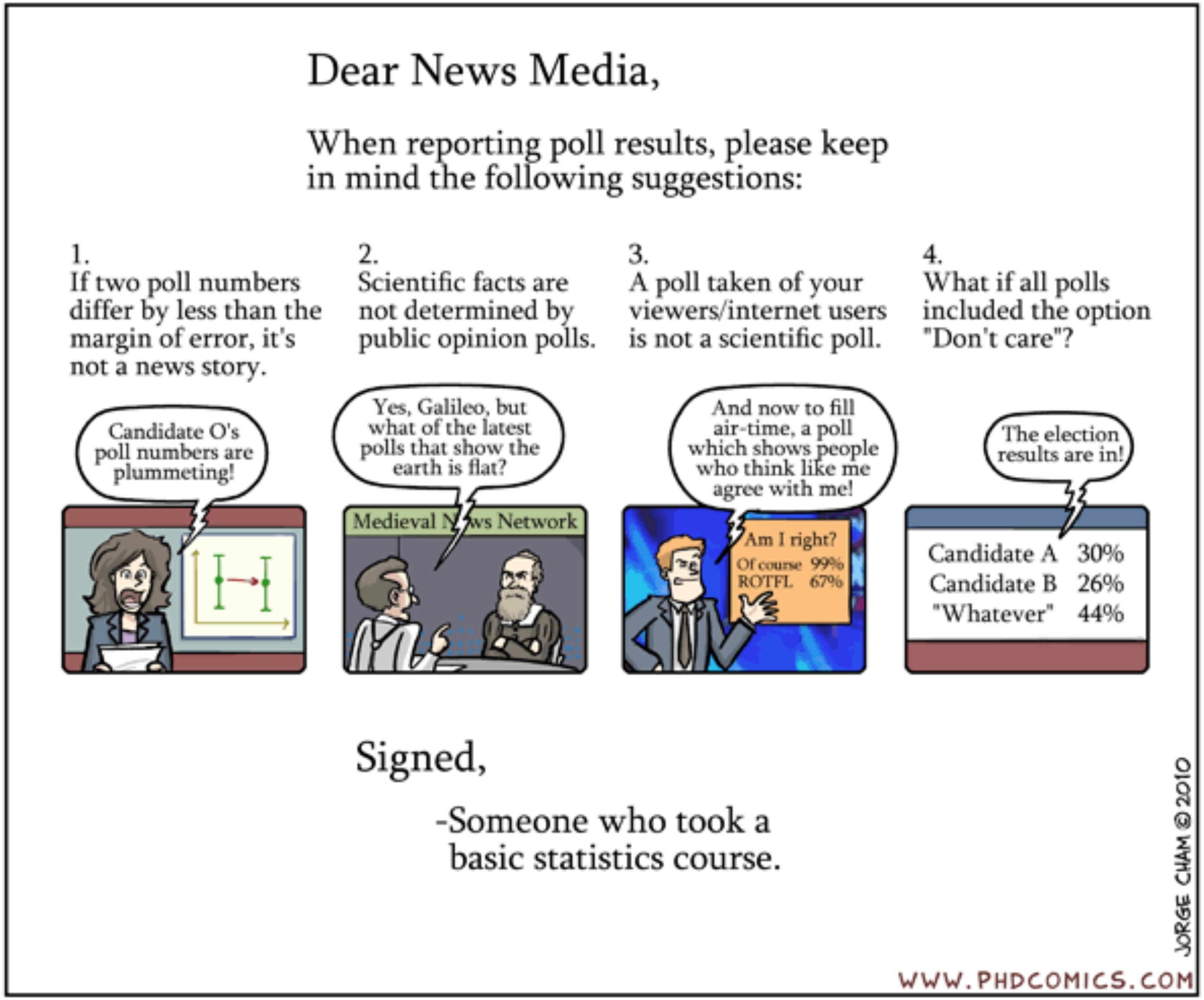
Nothing is always absolutely so. [Sturgeon’s First Law]

95% of everything is crud. [Sturgeon’s Maxim]

It can be tempting to use data as a crutch in decision-making: “The data says so!” But **sometimes the data lets us down** and that exciting correlation you found is just a by-product of a messy, biased sample. [...] Smart skeptics can help step back, reflect, and ask if **what the data is saying actually fits** with what you know and expect about the world.

[Nicholas Diakopoulos, Harvard Business Review]

Session 4



Dear News Media,

When reporting poll results, please keep in mind the following suggestions:

1. If two poll numbers differ by less than the margin of error, it's not a news story.

2. Scientific facts are not determined by public opinion polls.

3. A poll taken of your viewers/internet users is not a scientific poll.

4. What if all polls included the option "Don't care"?

Candidate O's poll numbers are plummeting!
Yes, Galileo, but what of the latest polls that show the earth is flat?
Medieval News Network
And now to fill air-time, a poll which shows people who think like me agree with me!
Am I right?
Of course 99%
ROTFL 67%
The election results are in!
Candidate A 30%
Candidate B 26%
"Whatever" 44%

Signed,

-Someone who took a basic statistics course.

JORGE CHAM © 2010

WWW.PHDCOMICS.COM

Session 4

USE THE CRS DATABASE TO SIZE THE MARKET.
THAT DATA IS WRONG.
scottadams@aol.com
www.dilbert.com
THEN USE THE SIBS DATABASE.
THAT DATA IS ALSO WRONG.
5·7·08 ©2008 Scott Adams, Inc./Dist. by UFS, Inc.
CAN YOU AVERAGE THEM?
SURE. I CAN MULTIPLY THEM TOO.

10. Bad Data and Big Data

Bad Data

Does the dataset pass the **smell test**?

- invalid entries, anomalous observations, etc.

Data formatted for human consumption, not machine readability

Difficulties with **text processing**

- encoding
- application-specific characters

Bad Data

Collecting data **online**

- legality of obtaining data
- storing offline versions

Detecting **lies** and **mistakes**

- reporting errors (lies or mistakes)
- use of polarizing language

Data and reality

- bad data
- bad reality?

Bad Data

Sources of **bias** and **errors**

- imputation bias
- top/bottom coding (replacing extreme values with average values)
- proxy reporting (head of household for household)

Seeking **perfection**

- academic data
- professional data
- government data
- service data

Bad Data

Data science **pitfalls**

- analysis without understanding
- using only one tool (by choice or by fiat)
- analysis for the sake of analysis
- unrealistic expectations of data science
- on a need-to-know basis … and you don’t need to know

Databases vs. files vs. cloud computing

- the cloud will solve all of our problems!

Bad Data

When is **close enough, good enough**?

- completeness
- coherence
- correctness
- accountability

Big Data – A Word of Warning

Big Data is no crystal ball

- “Past performance does not guarantee future results”

Big Data can't dictate personal or organizational values

- The right value answer may be the wrong data science answer
- Data-based conclusions do not live in a vacuum: context matters
- Blind obedience to data-driven results is just as dangerous as rejection based on gut-reaction

Big Data can't solve every problem

- “When all you have is a hammer, everything looks like a nail”

Big Data vs. Small Data

What is the main difference?

- the datasets are **LARGE**
- issues: collection, capture, access, storage, analysis, visualization

Where does the data come from?

- technology advances are lifting the limits on data processing speeds
- information-sensing, mobile devices, cameras and wireless networks

What are the challenges?

- most techniques were built for very small dataset
- direct approach will leave the best analyst waiting years for results

6. **variability**
7. **visualization**

The 5V$_{(7V?)}$ Paradigm

1. **volume:** large amounts of data
2. **velocity:** speed at which data is created, accessed, processed
3. **variety:** different types of available data, can’t all be saved in relational databases (tables, pictures,...)
4. **veracity:** quality and accuracy of big data is harder to control
5. **value:** turn the data into something useful

The Big Data Problem

Many computations happen **instantly**, others take a **significant** amount of time.

Crunching very large datasets is a perfect example. Analysis in *R* or *Python* with steadily increasing datasets leads to computer lags. Eventually, the time required becomes **impractically long**.

Optimizing code and using a faster CPU can only provide so much relief.

That is the **Big Data problem**.

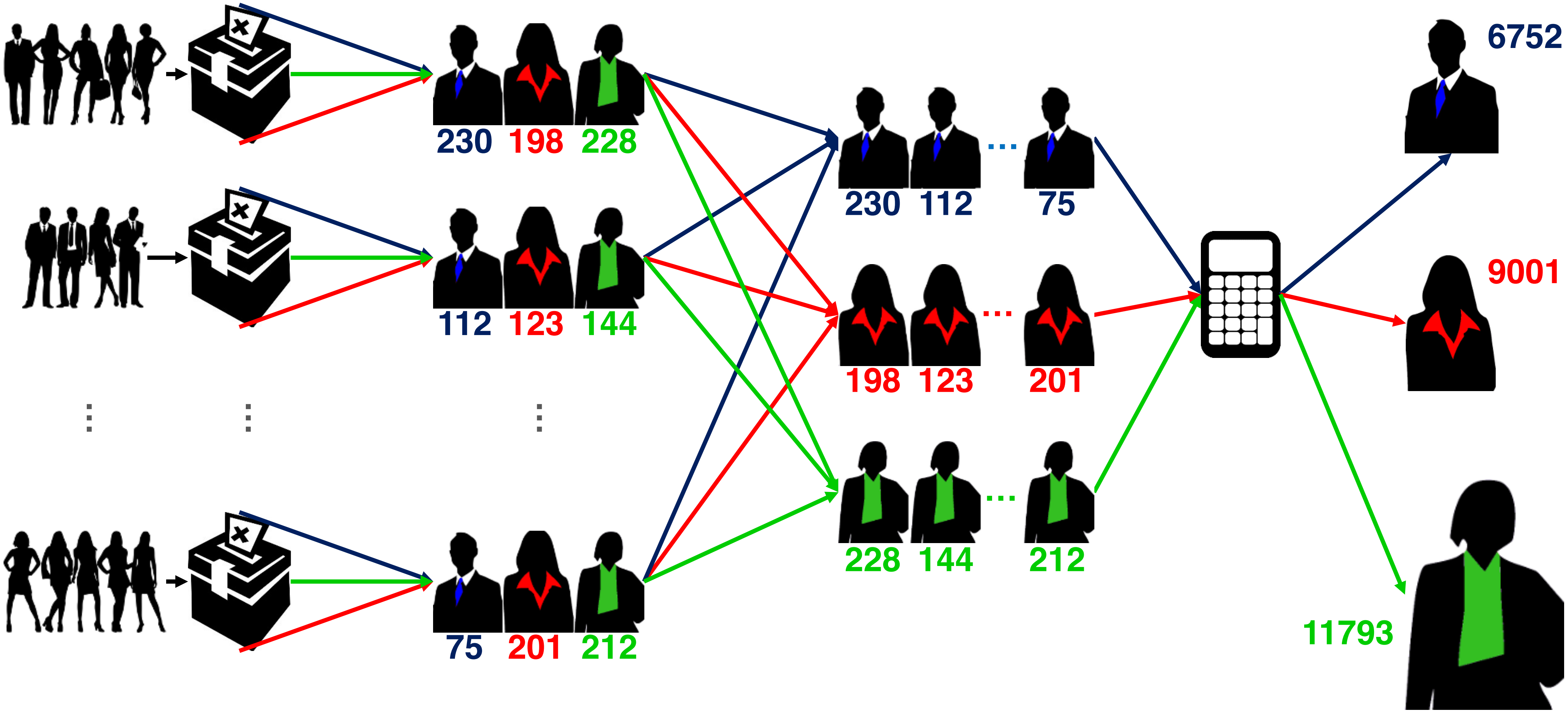
Distributed Computing

Splitting the computations among multiple CPU cores/CPUs can divide the computation time by a factor of 4, or 32, or 1000, or … This allows algorithms to run on big data to keep analytics, smart services, and recommendations updated **daily**, **hourly**, in **real time**.

Election analogy to parallelization:

- counting votes at different polling stations in a riding
- each station simultaneously counts its own votes and reports their total
- the totals of all polling stations are aggregated at Elections HQ
- one person counting all the ballots would eventually get the same result, but it would take *too long* to get the result.

Analogy: Elections



Analogy: Pizzeria

Parallelism gains depend on whether serial algorithms can be **adapted** to make use of **parallel hardware**.

Pizzeria analogy for limitations of parallelization/bottleneck:

- multiple cooks can prepare toppings in parallel
- but baking the crust can't be parallelized
- doubling oven space will increase the number of pizzas that can be made simultaneously but won't substantially speed up any one pizza
- sometimes bottlenecks prevent any gains from parallelism: people line up on both sides of a table to get some soup but there's only one ladle

Good News

Most practical computational tasks can be and are parallelized.

Modern data scientists use frameworks where distributed computing are already implemented (Apache Spark implements *MapReduce*, for instance).

Take some time to think about this potential issue **before** the start of the data collection/data analysis process – it will save headaches in the long run.

Suggested Reading

Bad Data and Big Data

J. Leskovec, A. Rajamaran, J. D. Ullman, *Mining of Massive Datasets*. Cambridge Press, 2014.

Data Understanding, Data Analysis, Data Science
Volume 3: Spotlight on Machine Learning

19. Introduction to Machine Learning
    - 19.6 Issues and Challenges
        - Bad Data

Exercises

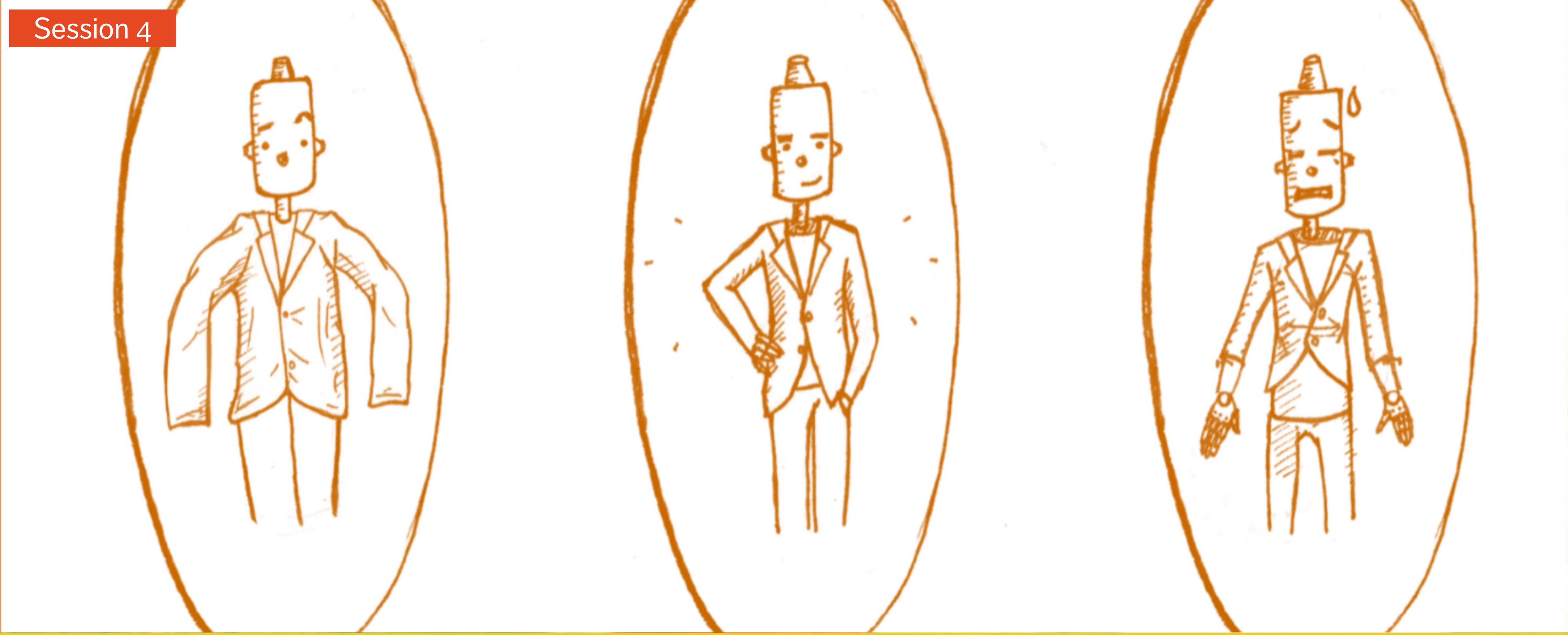
Bad Data and Big Data

1. As the saying goes, “garbage in, garbage out”. What are the analytical, business, and public policy consequences of making decisions based on bad data?

2. Whether a dataset is considered small or “big” depends not only on the dataset, but also on the available tools.

   Generate increasingly larger random datasets (3 variables + 1 class) to cluster with `kmeans()` and classify with `rpart()`. Keep track of the runtime. How does the runtime vary with the number of observations? At what size do you predict that the algorithms will be too slow and cumbersome for your needs?

Session 4



11. Underfitting and Overfitting/Transferability

Fundamentals

Rules or models generated by any technique on a **training set** have to be generalizable to **new data** (or **validation/ testing sets**) to be useful.

Problems arise when knowledge that is gained from **supervised learning** does not generalize properly to the data.

Unsupervised learning can also be affected.

Ironically, this may occur if the rules or models fit the training set **too well** – the results are **too specific to the training set**.

Example of Rules

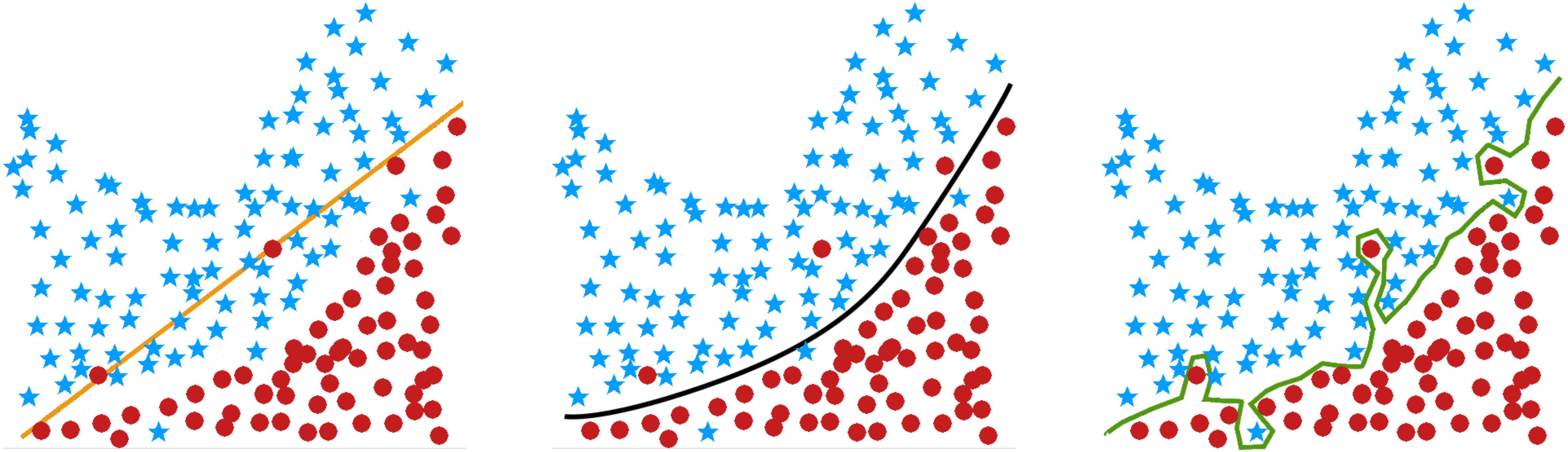
Rule I: based on a survey of 400 Germans, we infer that 43.75% of the world's population has black hair, 37.5% have brown hair, 9% have blond hair, 0.25% have red hair, and 9.5% grey hair.

Rule II: humans' hair colour is either black, brown, blond, red, or grey.

Rule III: approx. 40% of humans have black hair, 40% have brown hair, 5% blond, 2% red and 13% grey.

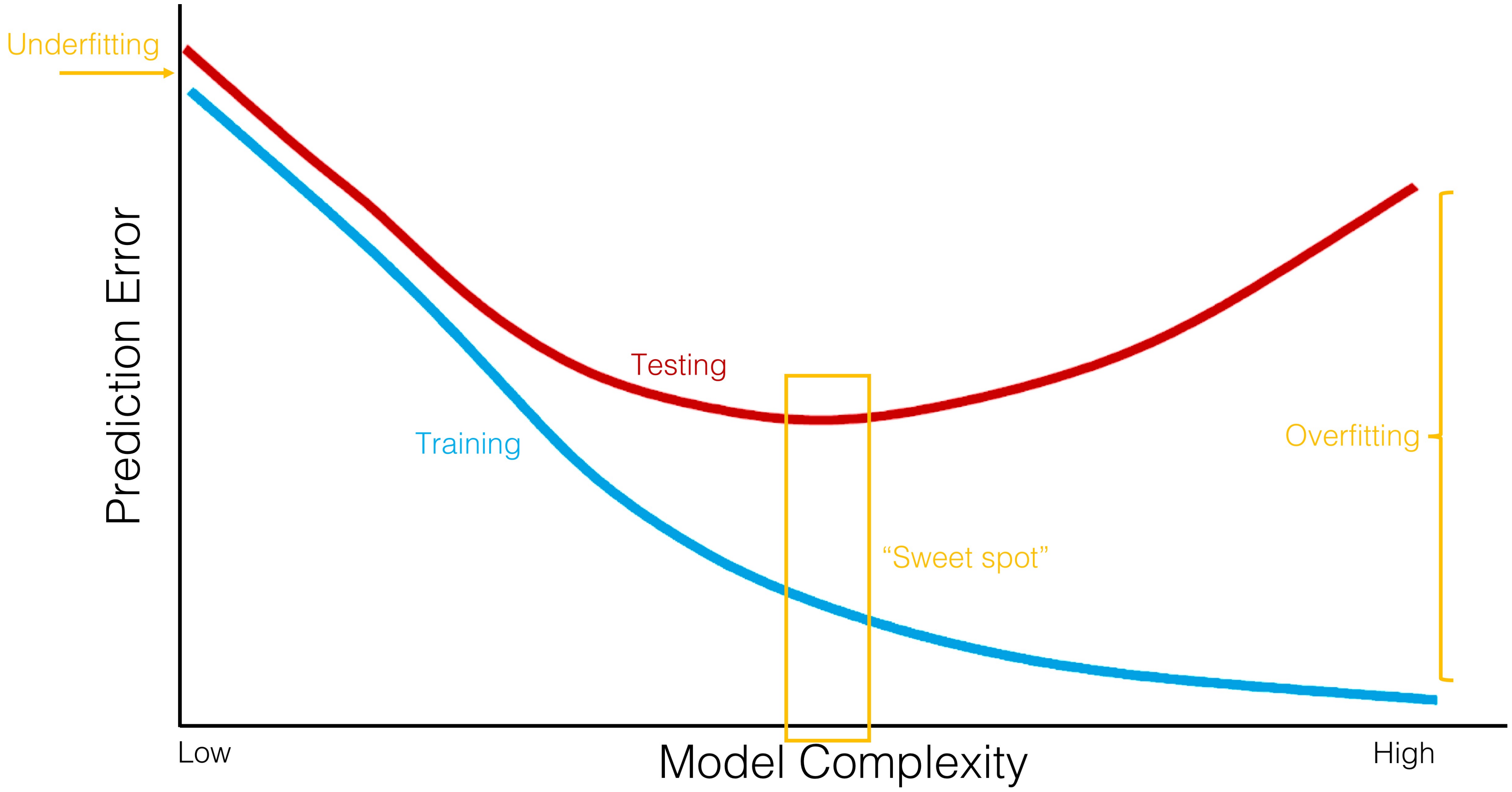
Which of the 3 rules is most useful? The most vague? Which is overly specific?

Goldilocks and the Three Models



Look into
double descent,
however.

Bias-Variance Trade-Off



ALWAYS evaluate models on unseen (testing) data!

Bias-Variance Trade-Off

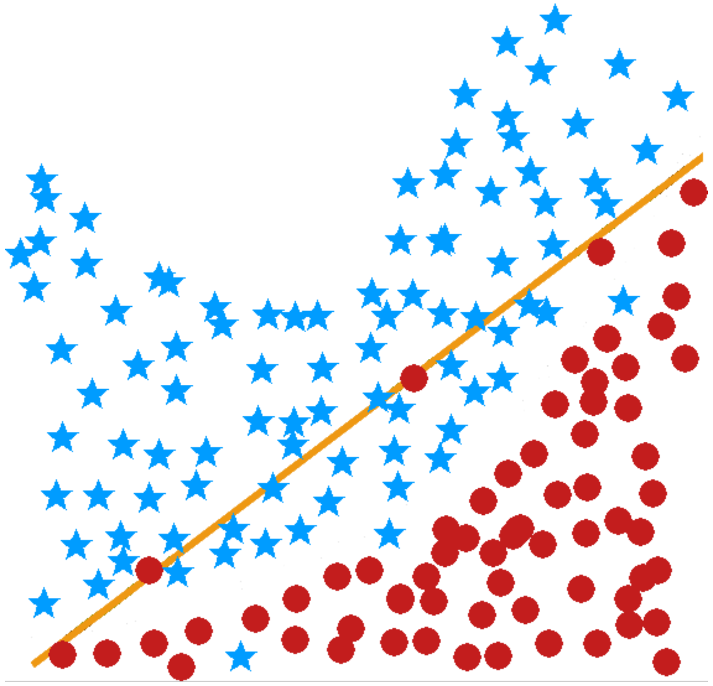
We **build** a model on **historical data** and **evaluate** its performance on **new data**.

Let $\mathrm{Error}_{\mathrm{Test}}$ be the model's performance on test data:

$$\mathrm{Error}_{\mathrm{Test}} = \mathrm{Bias}^2_{\mathrm{Model}} + \mathrm{Variance}_{\mathrm{Model}}$$

The **bias** measures the model's prediction **accuracy**; the **variance**, its **sensitivity** to (small) changes in the data.

Goldilocks and the Three Models



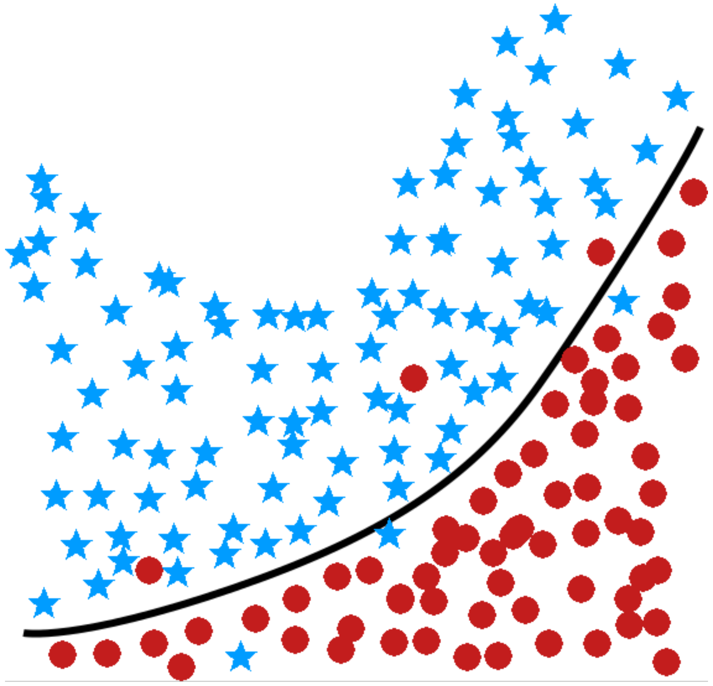
underfit

bias = high

variance = low

error = **high**

predictions are not very accurate

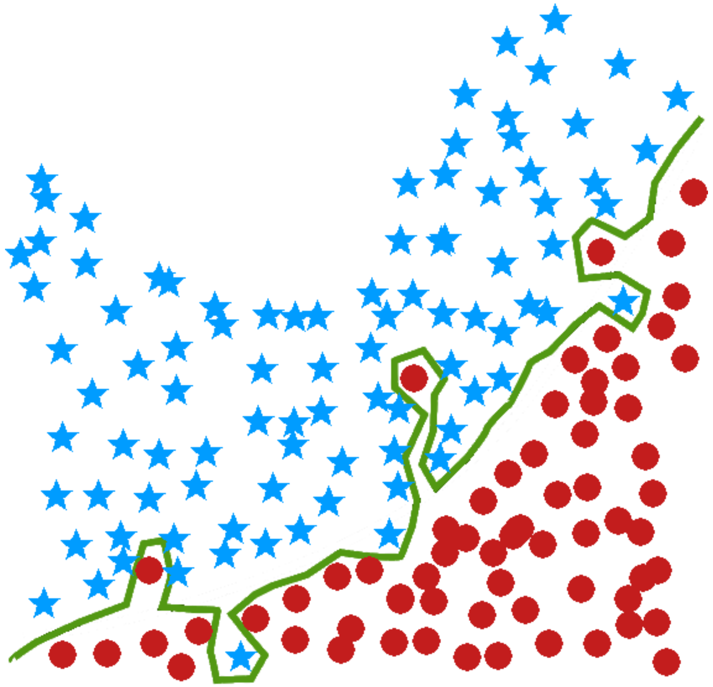


just right

bias = medium

variance = medium

error = **medium**



overfit

bias = low

variance = high

error = **high**

model is too specific to the data

Possible Solutions

Underfitting can be overcome by considering models that are more complex.

Overfitting can be overcome in several ways:

- **using multiple training sets**
  overlap is allowed (or not: see cross-validation)
- **using larger training sets**
  70% - 30% split is suggested
- **optimizing the data instead of the model**
  models are only as good as the data

Recommended Procedures

Small datasets (less than a few hundred observations)

- use 100-200 repetitions of a **bootstrap** procedure

Average-sized datasets (less than a few thousand observations)

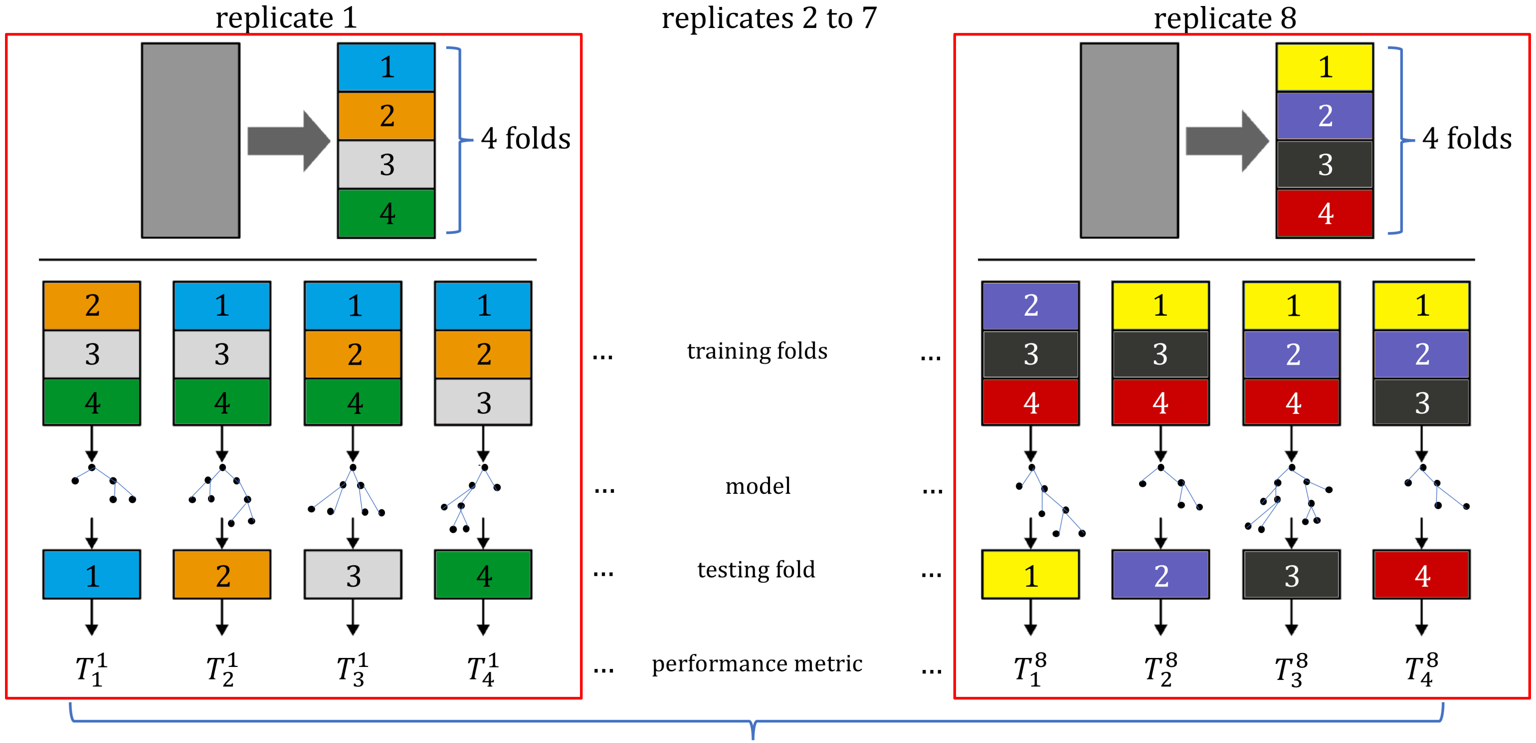
- use a few repetitions of 10-fold **cross-validation** on the training set (see next slide)

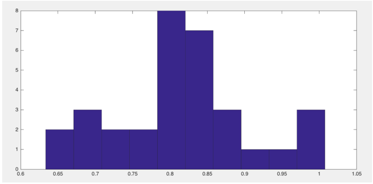
Large datasets

- use a few repetitions of **holdout** (70%-30%) split

The **decision boundaries** depend on computing power and the number of tasks in the workflows.

Cross-Validation





mean accuracy = 0.81
standard dev = 0.09

Appropriateness and Transferability

Data science and machine learning models will continue to be used heavily in the coming years.

We have discussed pros and cons of some of the applications on ethical and other non-technical grounds, but there are also **technical challenges**.

DS/ML methods are **not appropriate**:

- If you must absolutely use an existing (**legacy**) datasets instead of an **ideal** dataset ("it's the best data we have!")

Appropriateness and Transferability

DS/ML methods are **not appropriate** (cont.):

- if the dataset has attributes that usefully predict a value of interest, but which are **not available** when a prediction is required

  Example: the total time spent on a website may be predictive of a visitor's future purchases, but the prediction must be made before the total time spent on the website is known.

- if you attempt to predict **class membership** using an **unsupervised** learning algorithm

  Example: clustering loan default data might lead to a cluster that contains multiple defaulters. If new instances get added to this cluster, should they automatically be viewed as loan defaulters? (no)

Non-Transferable Assumptions

Every model makes certain assumptions about what is and is not **relevant** to its workings, but there is a tendency to only gather data which is **assumed** to be relevant to a particular situation.

If data is used in other contexts, or to make predictions depending on attributes without data, validating the results may prove impossible.

- **Example:** can we use a model that predicts mortgage defaulters to also predict car loan defaulters? A car is not a house: they play different roles in an individual's life; the values are of different magnitudes; and so on...
- That being said, is there truly no link between mortgage defaults and car loan defaults?

Session 4

Suggested Reading

Underfitting and Overfitting/ Transferability

Data Understanding, Data Analysis, Data Science
Volume 3: Spotlight on Machine Learning

19. Introduction to Machine Learning

19.6 Issues and Challenges

- Overfitting/Underfitting
- Transferability

20. Regression and Value Estimation

20.1 Statistical Learning

- Model Evaluation
- Bias-Variance Trade-Off

20.3 Resampling Methods

- Cross-Validation

20.4 Model Selection

- Optimal Models

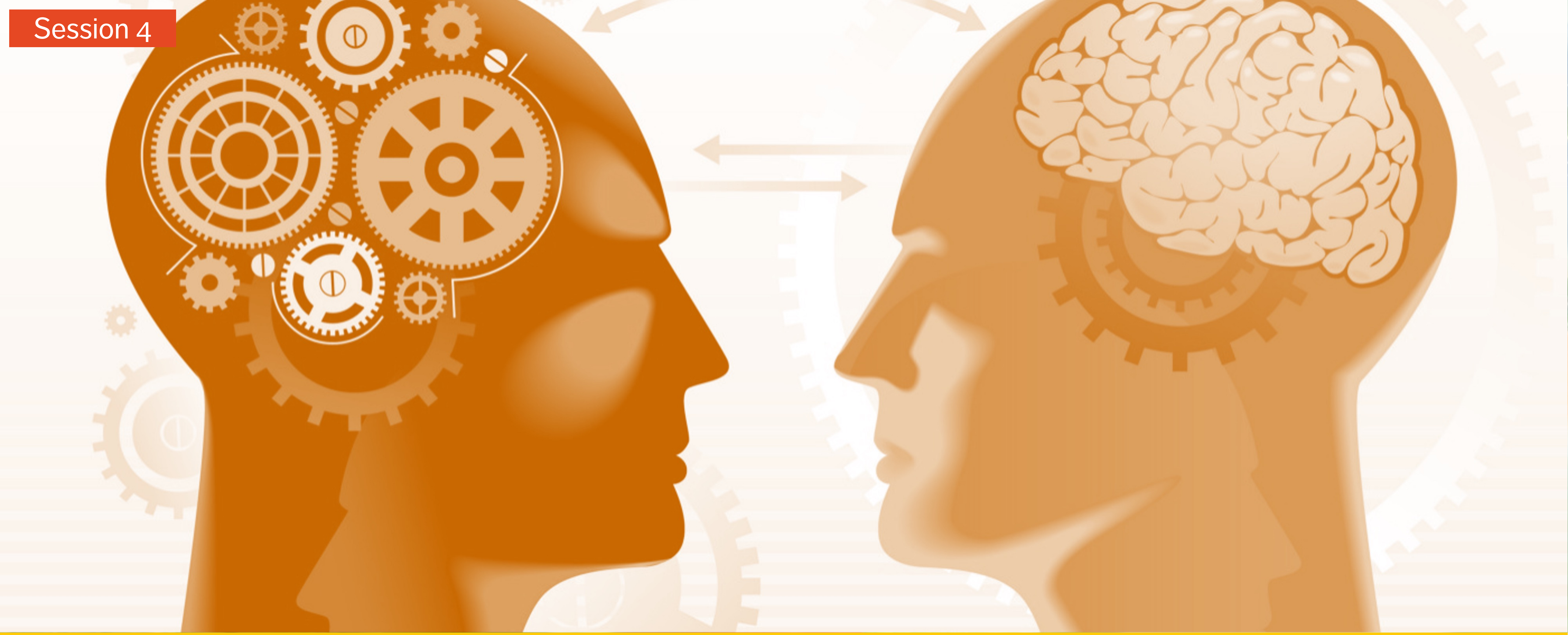
Exercises

Underfitting and Overfitting/ Transferability

1. This exercise illustrates overfitting/underfitting.
   a. Randomly generate $n = 150$ values in $[0,10]$ for the predictor x.
   b. Randomly generate $n = 150$ response values according to $y = 10 + x - 2x^2 + 17x^3 + \varepsilon$, where ε is a random error term of your choice.
   c. Fit a linear model, a quadratic model, a cubic model, and a polynomial model of degree 10 to the data.
   d. Add 3 observations to the data as in steps a. and b. Repeat step c. Do the models change much?
   e. Which model(s) would you trust to make predictions on new data?

2. Modify the Gapminder example from Cross-Validation to select a model in the previous question.

Session 4



12. Miscellanea

Biases, Fallacies, and Interpretation

When consulting (or conducting) studies, beware:

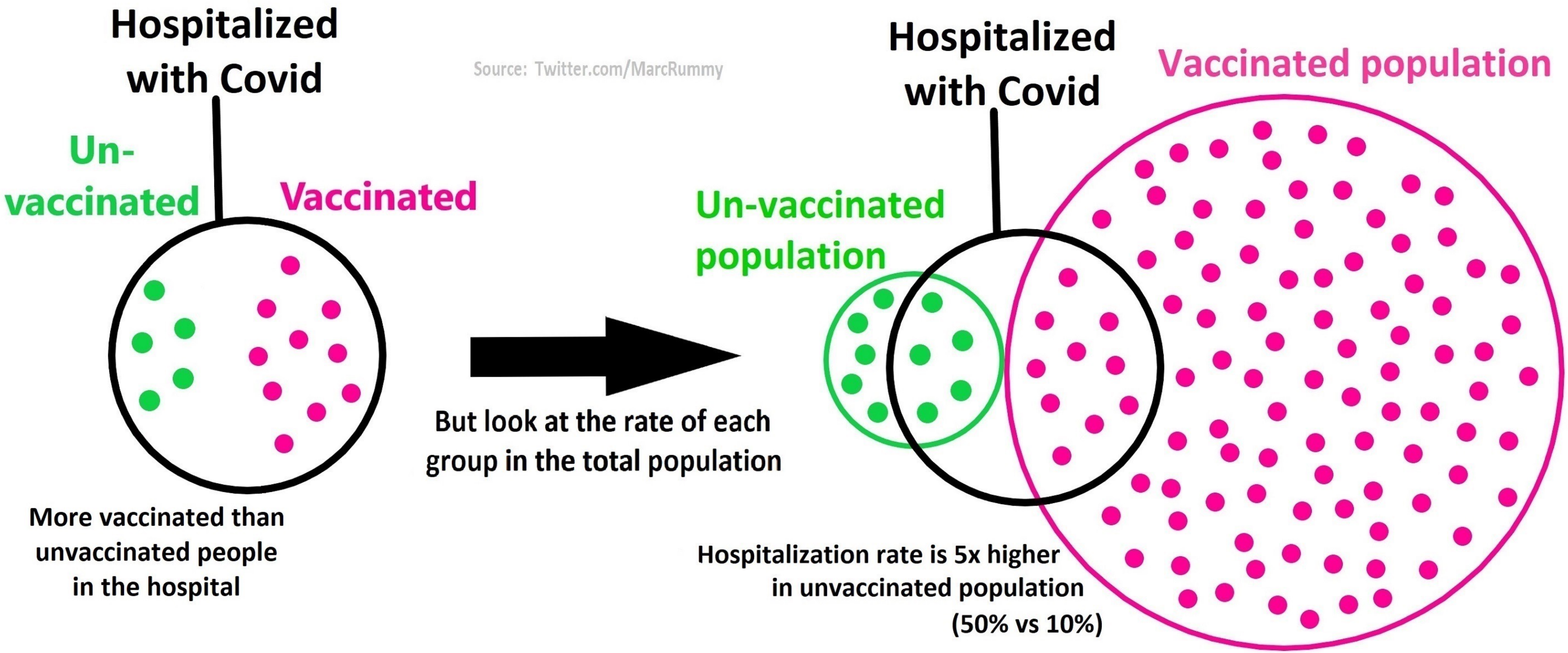
- **selection bias** (what data was included, how was it selected?)
- **omitted-variable bias** (were relevant variables ignored?)
- **detection bias** (did prior knowledge affect the results?)
- **funding bias** (who's paying for this?)
- **publication bias** (what's not being published?)
- **data-snooping 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.)



Source: Twitter.com/MarcRummy

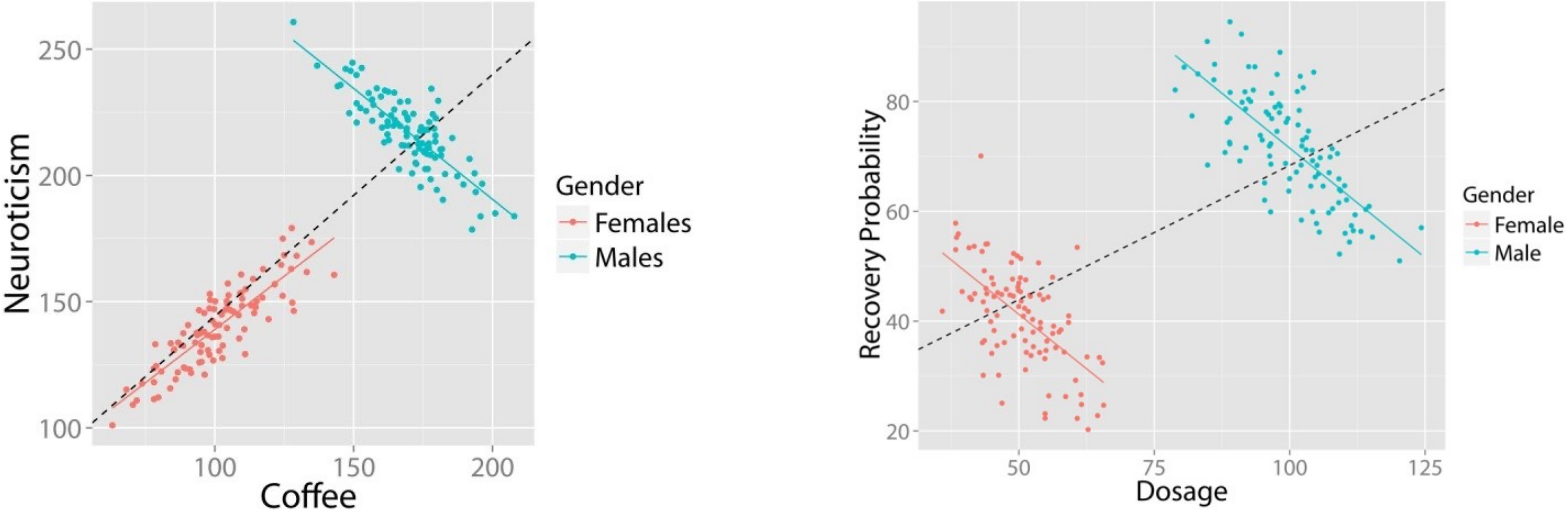
More vaccinated than unvaccinated people in the hospital

But look at the rate of each group in the total population

Hospitalization rate is 5x higher in unvaccinated population (50% vs 10%)

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*]



[A.K. Maheshwari, *Business Intelligence and Data Mining*]

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 (?)

[A.K. Maheshwari, *Business Intelligence and Data Mining*]

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.

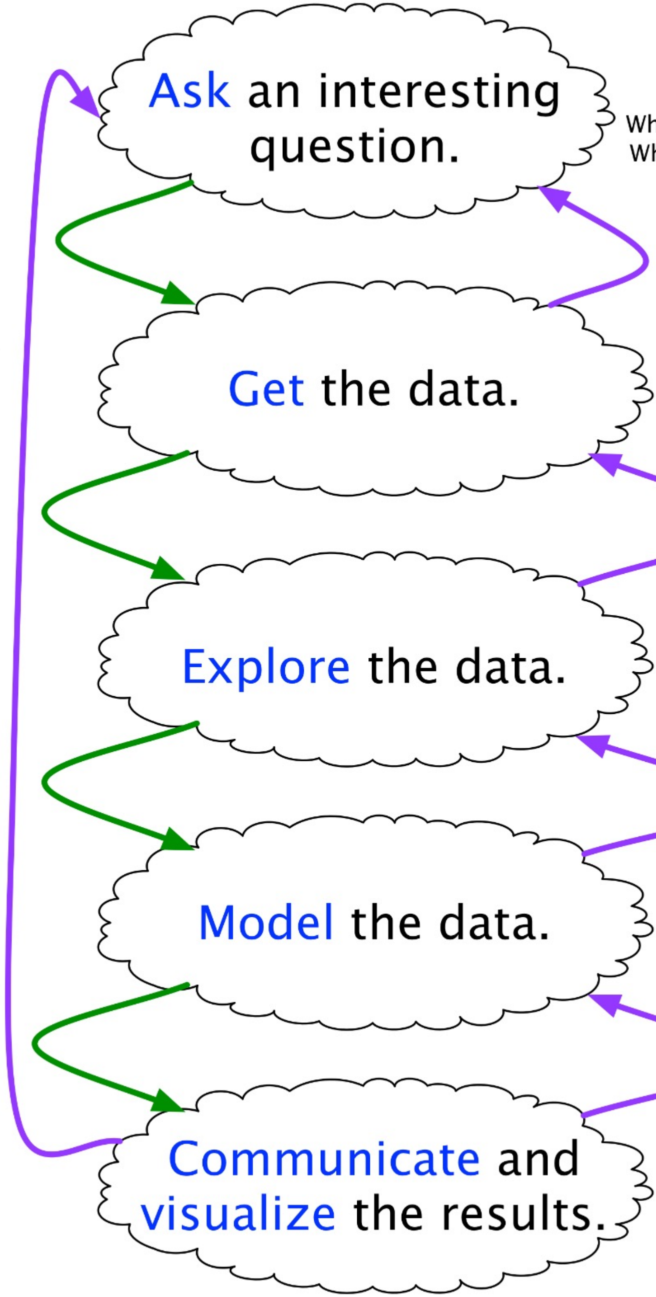
Ethical considerations are crucial.

Let the data speak.

Look for actionable insights.

Supervised vs. unsupervised.

Be ready to clean, prepare, & visualize data.



Ask an interesting question.
Get the data.
Explore the data.
Model the data.
Communicate and visualize the results.

[Blitzstein and Pfister]

What is the scientific **goal**?
What would you do if you had all the **data**?
What do you want to **predict** or **estimate**?

How were the data **sampled**?
Which data are **relevant**?
Are there **privacy** issues?

Plot the data.
Are there **anomalies**?
Are there **patterns**?

Build a model.
Fit the model.
Validate the model.

What did we **learn**?
Do the results make **sense**?
Can we tell a **story**?

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?