ISSUES AND CHALLENGES

ADVANCED DATA SCIENCE TRAINING

OUTLINE

1. Bad Data
2. Overfitting
3. Big Data
4. Appropriateness and Transferability
5. Biases, Fallacy, Interpretation
6. Myths and Mistakes
7. The Future of D.S./A.I./M.L.
8. In Conclusion

LEARNING OBJECTIVES

Be able to describe, at a high level, some of the common issues and challenges associated with data analysis.

Understand the value of an approximate model.

Be familiar with the V5 description of big data.

Appreciate appropriate uses of data science results.

Awareness of some common types of bias in data science.

Awareness of some classic myths and mistakes in data science.

BAD DATA

ISSUES AND CHALLENGES

"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 quite care for data, and it sounds like I'm not alone."

(Q.E. McCallum, *Bad Data Handbook*)

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
- it's 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

DISCUSSION

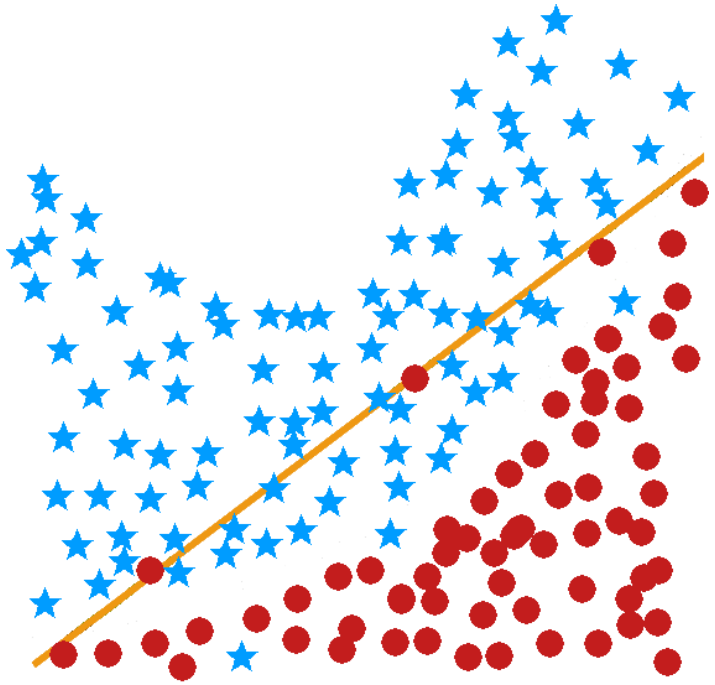
As the saying goes, “garbage in, garbage out”. What are the business and public policy consequences of making decisions based on bad data?

OVERFITTING

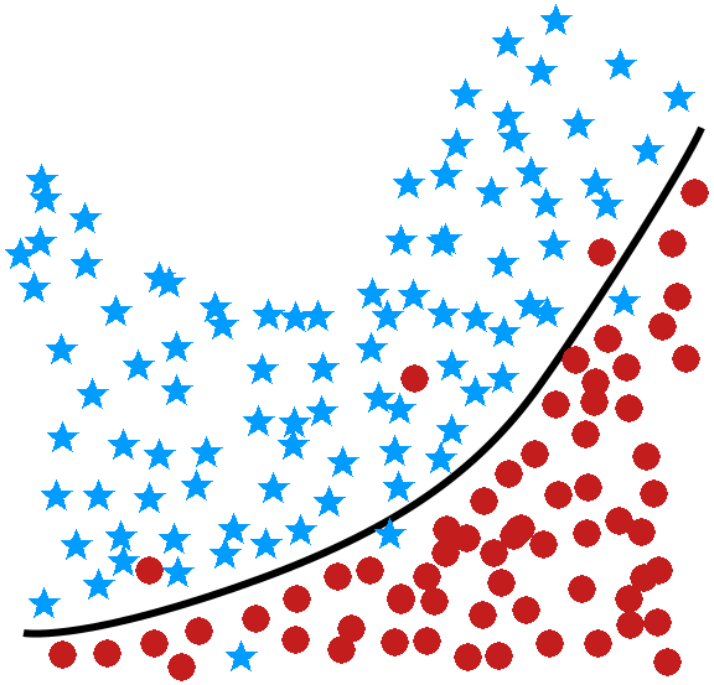
ISSUES AND CHALLENGES

(AMAR GONDALIYA, PINGAX)

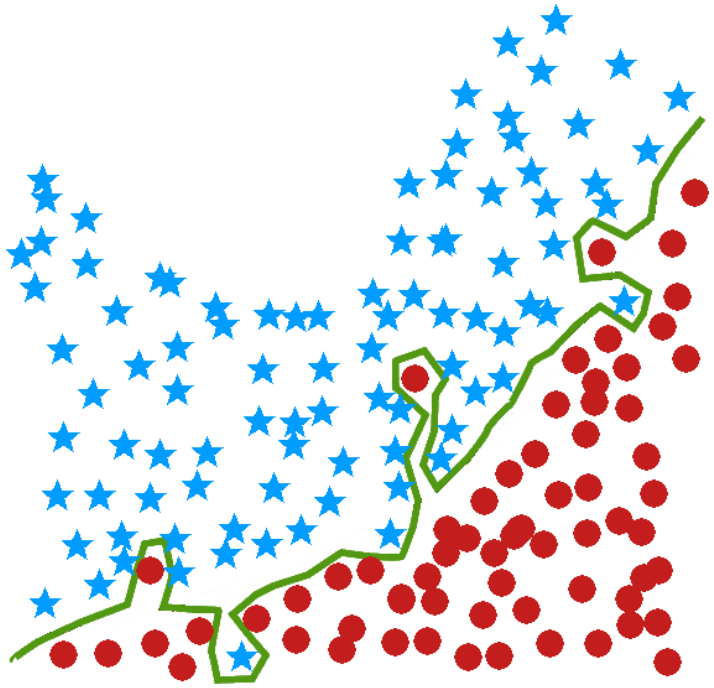
Goldilocks and the Three Models



underfit



just right



overfit

FUNDAMENTALS

The hope is for rules or models generated by any technique on a **training set** to be generalizable to **new data** (or **validation/ testing sets**).

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

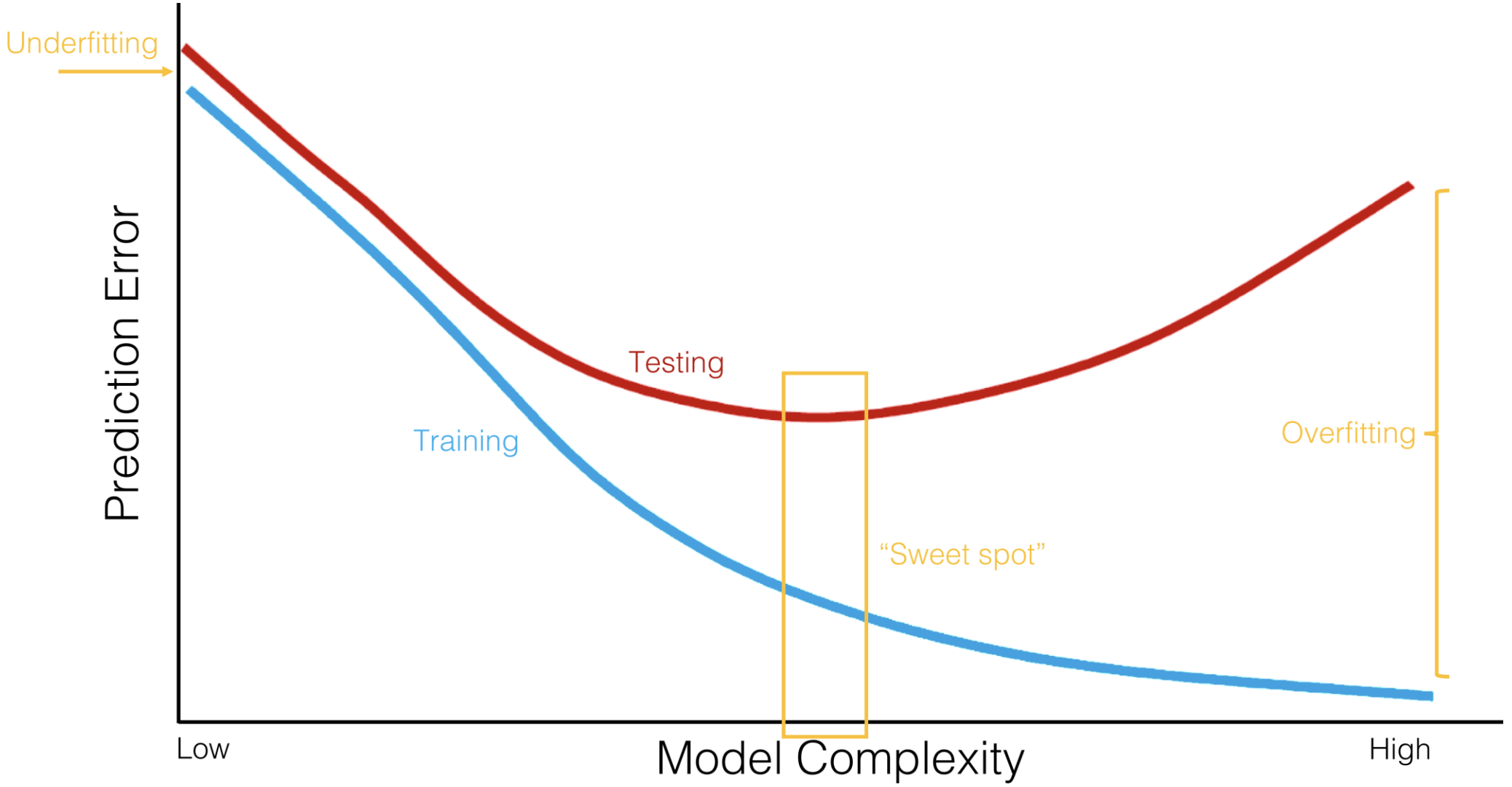
Rule I: based on a survey of 400 Germans, we infer that 43.75% of the world's population has black hair, 37.5% have brown hair, 9% have blond hair, 0.25% have red hair, and 9.5% grey hair.

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.

DISCUSSION

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


Underfitting
Prediction Error
Testing
Training
Overfitting
"Sweet spot"
Low
Model Complexity
High

OVERFITTING

ALWAYS evaluate models on unseen (testing) data.

POSSIBLE SOLUTIONS

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

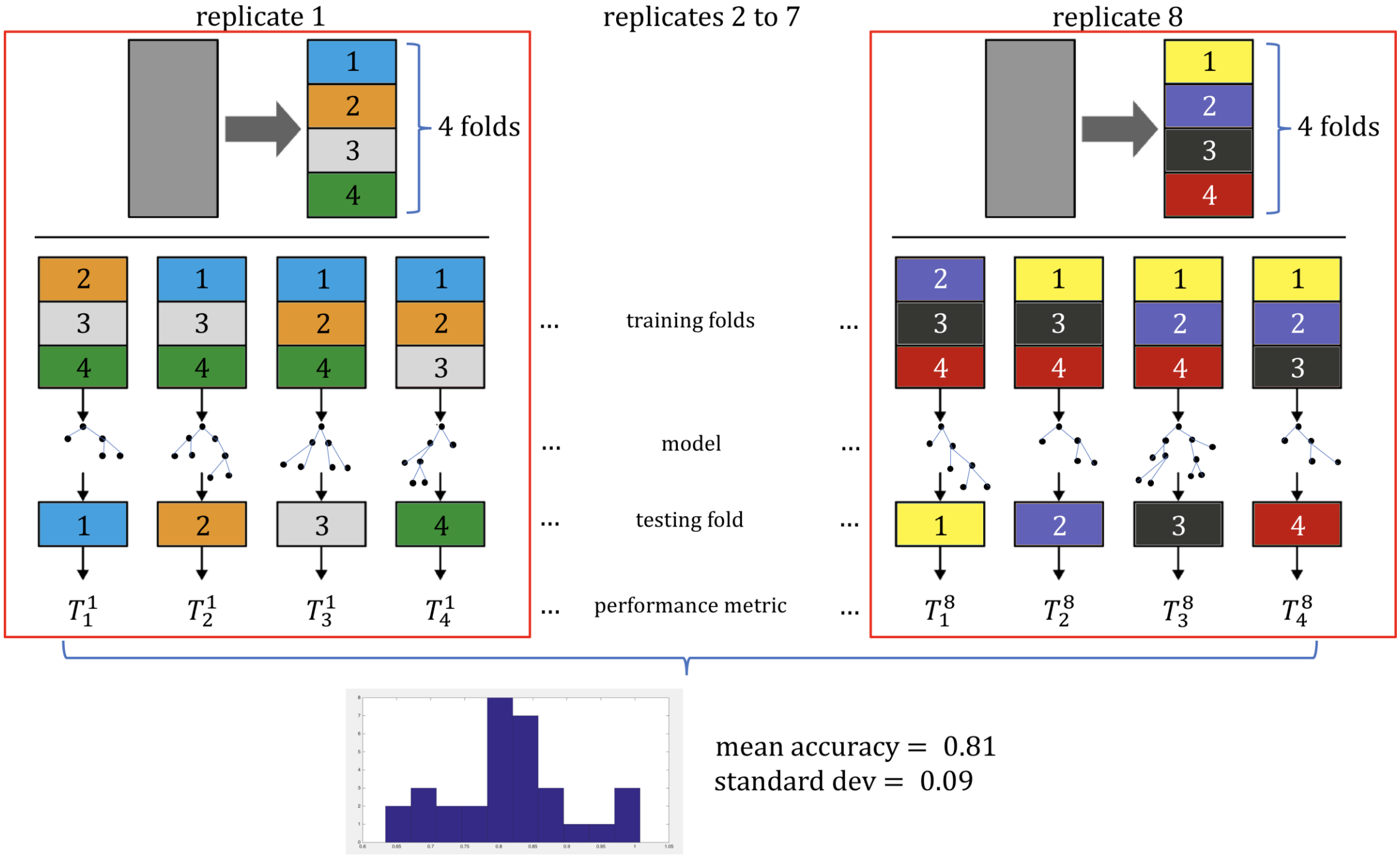
Average-sized datasets (less than a few thousand observations)

- 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

Note: decision boundaries depend on computing power and number of tasks/workflows.



mean accuracy = 0.81
standard dev = 0.09

BIG DATA

ISSUES AND CHALLENGES

"Data, big or small, is only as useful as the questions you ask of it."

(Milo Jones and Philippe Silberzahn, Forbes Magazine)

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

THE 5-V PARADIGM

Volume: large amounts of data

Velocity: speed at which data is created, accessed, processed

Variety: different types of available data, can't all be saved in relational databases (tables, pictures,...)

Veracity: quality and accuracy of big data is harder to control

Value: turn the data into something useful

Variability
Visualization

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**.

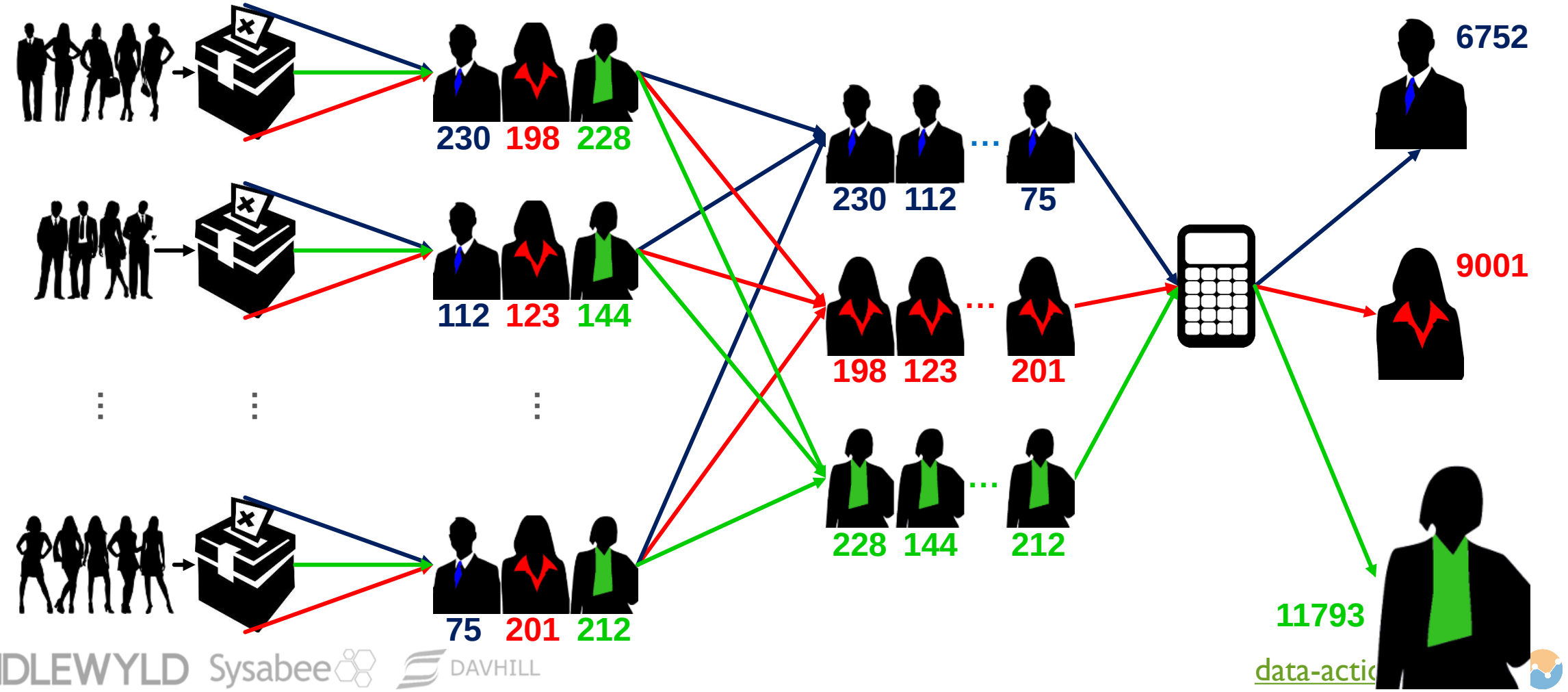
DISTRIBUTED COMPUTING

Splitting the computations among multiple CPU cores/CPUs can divide the computation time by a factor of 4, or 32, or 1000. 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: BOARD ELECTION



ANALOGY: PIZZERIA

The gains from parallelism 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).

APPROPRIATENESS AND TRANSFERABILITY

ISSUES AND CHALLENGES

"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)

APPROPRIATENESS AND TRANSFERABILITY

Data science models will be used heavily in the coming years (it has already started).

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

Data Science methods are **not** appropriate if:

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

APPROPRIATENESS AND TRANSFERABILITY

Data Science methods are **not** appropriate if (continued):

- 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 will attempt to predict class membership using an unsupervised learning algorithm

  Example: clustering loan default data might lead to a cluster contains many defaulters. If new instances get added to this cluster, should they be viewed as loan defaulters?

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 is impossible.

- **Example:** can we use a model that predicts mortgage defaulters to also predict car loan defaulters?

DISCUSSION

Is there truly no link between mortgage defaults and car loan defaults?

BIASES, FALLACIES, AND INTERPRETATION

ISSUES AND CHALLENGES

"If two poll numbers differ by less than the margin of error, it's not a news story. Scientific facts are not determined by public opinion polls. A poll taken of your viewers/internet users is not a scientific poll.
What if all polls included the option "Don't care"?"

(Jorge Chan, Piled Higher and Deeper)

BIASES, FALLACIES, AND INTERPRETATION

When consulting (or conducting) studies, you should try to determine how the following biases could have come into play:

- **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?)

BIASES, FALLACIES, AND INTERPRETATION

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.

Odd results sometimes happen (Simpson's Paradox).

BIASES, FALLACIES, AND INTERPRETATION

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.).

DISCUSSION

Does the presence of bias necessarily invalidate the results?

MYTHS AND MISTAKES

ISSUES AND CHALLENGES

"Nothing is always absolutely so."

(Sturgeon's First Law)

"95% of everything is crud."

(Sturgeon's Maxim)

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

DATA SCIENCE MYTHS & MISTAKES

Myth #1 – Data science (DS) is about algorithms.

Myth #2 – DS is about predictive accuracy.

Myth #3 – DS requires a data warehouse.

Myth #4 – DS requires a large quantity of data.

Myth #5 – DS requires technical experts.

DATA SCIENCE MYTHS & MISTAKES

Mistake #1 – Selecting the wrong problem.

Mistake #2 – Getting buried under tons of data without metadata understanding.

Mistake #3 – Not planning the data analysis process.

Mistake #4 – Insufficient business and domain knowledge.

Mistake #5 – Using incompatible data analysis tools.

DATA SCIENCE MYTHS & MISTAKES

Mistake #6 – Using tools that are too specific.

Mistake #7 – Ignoring individual predictions/records in favour of aggregated results.

Mistake #8 – Running out of time.

Mistake #9 – Measuring results differently than the sponsor.

Mistake #10 – Naïvely believing what one's told about the data.

DISCUSSION

Data science is about asking the right questions and accepting imaginative solutions.

In the battle between “tried, tested, and true” and “disruptive data science”, with whom do you side?

EXERCISE – TRUE / FALSE QUESTIONS

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

EXERCISE – TRUE / FALSE QUESTIONS

6. If all methods don't yield the same result, it is a proof that the question cannot be answered.
7. Business and domain knowledge is only necessary when working with old data.
8. Sponsors and clients need to be told all analytical details.
9. It's impossible to plan the data analysis process before we know what the data looks like.
10. The available data is not always appropriate/representative of the situation we are modeling.

THE FUTURE OF D.S./A.I./M.L.

ISSUES AND CHALLENGES

WHAT WE DIDN'T TALK ABOUT

Tons of other classification and clustering algorithms

Recommender systems

Data streams

Bayesian data analysis

Natural language processing (in depth)

Feature selection and dimension reduction (curse of dimensionality)

Data engineering

... and many more!

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.

FUTURE TRENDS

New questions

New tools

New data sources

Data science as job component

Augmented/swarm intelligence

IN CONCLUSION

ISSUES AND CHALLENGES

[Joe Blitzstein and Hanspeter Pfister, http://cs109.org]

Data science is a team activity, with subject matter experts.

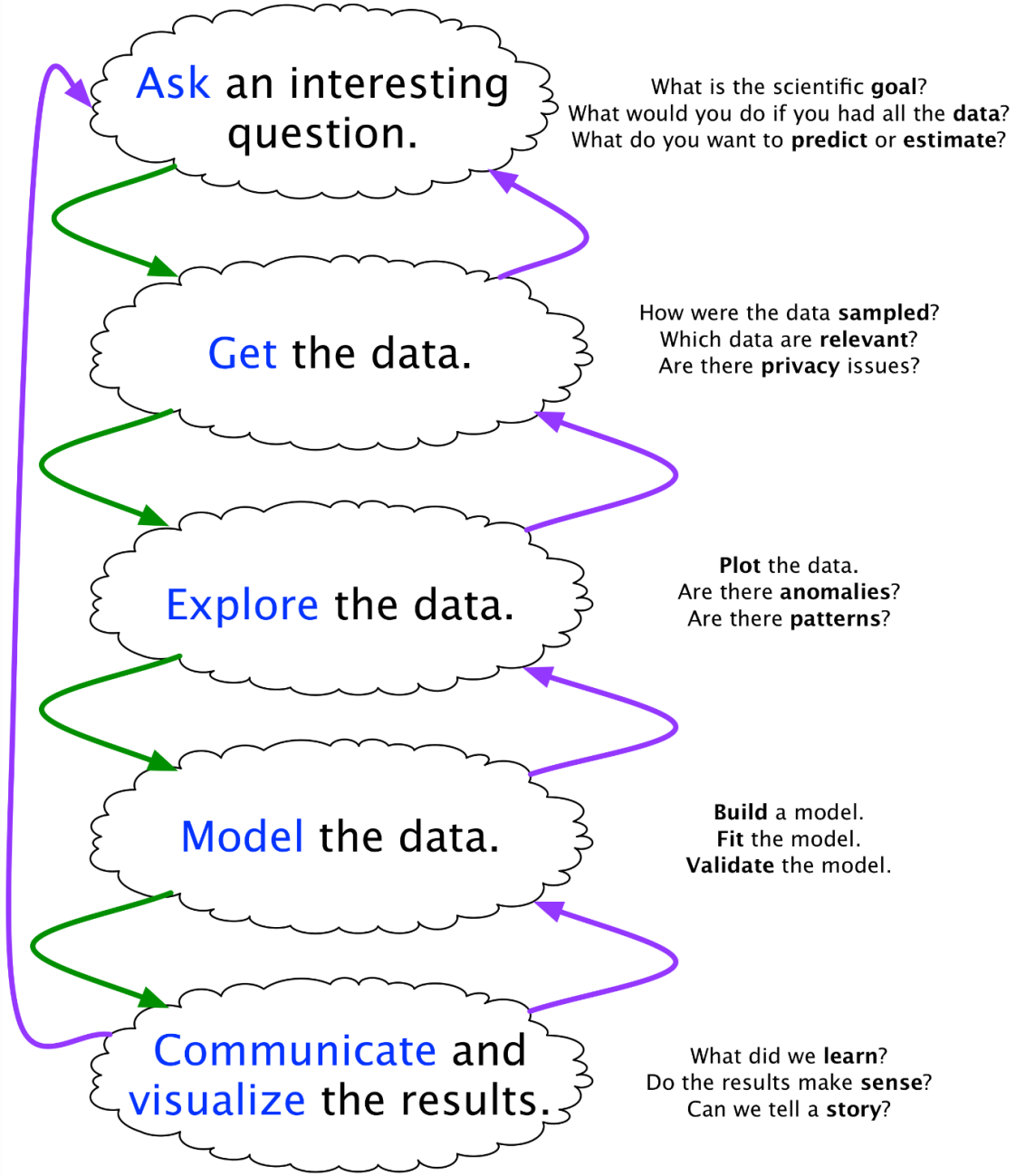
Ethical considerations are paramount, and need not be in conflict with profitability.

Let the data speak.

Look for actionable insights!

Supervised vs. unsupervised.

Large proportion of analysis time spent on data preparation.



REFERENCES

ISSUES AND CHALLENGES

REFERENCES

Aggarwal, C.C. (ed.) [2015], *Data Classification: Algorithms and Applications*, CRC Press.

Aggarwal, C.C., Reddy, C.K. (eds.) [2014], *Data Custering: Algorithms and Applications*, CRC Press.

Torgo, L. [2017], *Data Mining with R: Learning with Case Studies* (2nd ed.), CRC Press

McCallum, Q.E. [2013], Bad Data Handbook, O'Reilly.

Maheshwari, A.K. [2015], Business Intelligence and Data Mining, Business Expert Press.

Provost, F., Fawcett, T. [2013], Data Science for Business, O'Reilly.

Frank, E., Witten, I.H. [2005], Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed., Elsevier.

REFERENCES

https://hbr.org/2013/07/how-google-flu-trends-is-getting-to-the-bottom

9 types of research bias and how to avoid them

Wikipedia entry for Bias: https://en.wikipedia.org/wiki/Bias

Wikipedia entry for Selection Bias: https://en.wikipedia.org/wiki/Selection_bias

Assessing Risk of Bias in Included Studies, Cochrane Methods

Data Snooping Bias, Quantshare

Wikipedia entry for Statistical Biases: https://en.wikipedia.org/wiki/Bias_(statistics)

Wikipedia entry for Benford's Law: https://en.wikipedia.org/wiki/Benford%27s_law

REFERENCES

Silver, N. [2012], The Signal and the Noise: Why So Many Predictions Fail – But Some Don't, Penguin Press, New York

Lewis, M. [2003], Moneyball: The Art of Winning an Unfair Game, Norton, New York

Uri Simonshon, http://opim.wharton.upenn.edu/~uws/

https://en.wikipedia.org/wiki/Data_analysis_techniques_for_fraud_detection

Flaherty, D., "The Vaccine-Autism Connection: A Public Health Crisis Caused by Unethical Medical Practices and Fraudulent Science," Ann Pharmacother, Oct2011 v45 n10 1302-04

Reinhart, A., Statistics Done Wrong

REFERENCES

https://www.datacamp.com/community/blog/data-science-past-present-future

Kargupta, H., Han, J., Yu, P.S., Motwani, R., Kumar, V. (eds) [2009], *Next Generation of Data Mining*, CRC/Chapman & Hall.