Ethics of Data Science – Part II

Communicating uncertainty

Quantifying uncertainty is only half the job

- **Identify** sources of uncertainty
- **Robustify** model development process against uncertainty
- **Quantify** all residual sources of uncertainty
- **Edify** stakeholders about degree of confidence and uncertainty

Communicating uncertainty is its own art and science

Decreasing precision

i. A full explicit probability distribution
ii. A summary of a distribution
iii. A rounded number, range or an order-of-magnitude assessment
iv. A predefined categorisation of uncertainty
v. A qualifying verbal statement
vi. A list of possibilities or scenarios
vii. Informally mentioning the existence of uncertainty
viii. No mention of uncertainty
ix. Explicit denial that uncertainty exists

Van der Bles, Anne Marthe, et al. "Communicating uncertainty about facts, numbers and science." *Royal Society open science* 6.5 (2019): 181870.

Communicating uncertainty is its own art and science

- Direct uncertainty is that which can be quantified by the model itself
- Indirect uncertainty is doubt that arises from the quality of the data, model assumptions, or other factors that cannot be quantified
- Keeping those distinct can aid communication, though some organisations prefer to combine both into a single statement for some audiences (e.g., “possibly carcinogenic”, by IARC).

relative effect (95% CI)	no. participants (studies)	quality of the evidence (GRADE)
HR 0.94 (0.83 – 1.06)	3693 (3 RCTs)	⊕⊕⊕⊕ HIGH

Zhu J, Li R, Tiselius E, Roudi R, Teghararian O, Suo C, Song H. 2017Immunotherapy (excluding checkpoint inhibitors) for stage I to III non-small cell lung cancer treated with surgery or radiotherapy with curative intent. *Cochrane Database Syst. Rev.* **12**, Cd011300.

Communicating uncertainty is its own art and science

Quality level	Current definition	Previous definition
High	We are very confident that the true effect lies close to that of the estimate of the effect	Further research is very unlikely to change our confidence in the estimate of effect
Moderate	We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different	Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate
Low	Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect	Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate
Very low	We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect	Any estimate of effect is very uncertain

Balshem, Howard, et al. "GRADE guidelines: 3. Rating the quality of evidence." *Journal of clinical epidemiology* 64.4 (2011): 401-406.

Communicating uncertainty is its own art and science

What do you think “very likely” means in the quote on the right?

“Most of the observed increase in global average temperatures since the mid-20th century is very likely due to the observed increase in anthropogenic greenhouse gas concentrations”

- Spiegelhalter DJ, Pearson M, Short I. 2011Visualizing uncertainty about the future. *Science* **333**, 1393-1400.
- Intergovernmental Panel on Climate Change Working Group 1, *IPCC Fourth Assessment Report: The Physical Science Basis* (IPCC, 2007; www.ipcc.ch/ipccreports/ar4-wg1.htm).

Communicating uncertainty is its own art and science

What do you think “very likely” means in the quote on the right?

90%-99%

“Most of the observed increase in global average temperatures since the mid-20th century is very likely due to the observed increase in anthropogenic greenhouse gas concentrations”

- Spiegelhalter DJ, Pearson M, Short I. 2011Visualizing uncertainty about the future. *Science* **333**, 1393-1400.
- Intergovernmental Panel on Climate Change Working Group 1, *IPCC Fourth Assessment Report: The Physical Science Basis* (IPCC, 2007; www.ipcc.ch/ipccreports/ar4-wg1.htm).

Communicating uncertainty is its own art and science

What % of respondents to a U.S. survey that asked the question on the right do you think got it wrong?

"Which of the following numbers represents the biggest risk of getting a disease? 1 in 100, 1 in 1000, or 1 in 10?"

- Spiegelhalter DJ, Pearson M, Short I. 2011Visualizing uncertainty about the future. *Science* **333**, 1393-1400.
- Intergovernmental Panel on Climate Change Working Group 1, *IPCC Fourth Assessment Report: The Physical Science Basis* (IPCC, 2007; www.ipcc.ch/ipccreports/ar4-wg1.htm).

Communicating uncertainty is its own art and science

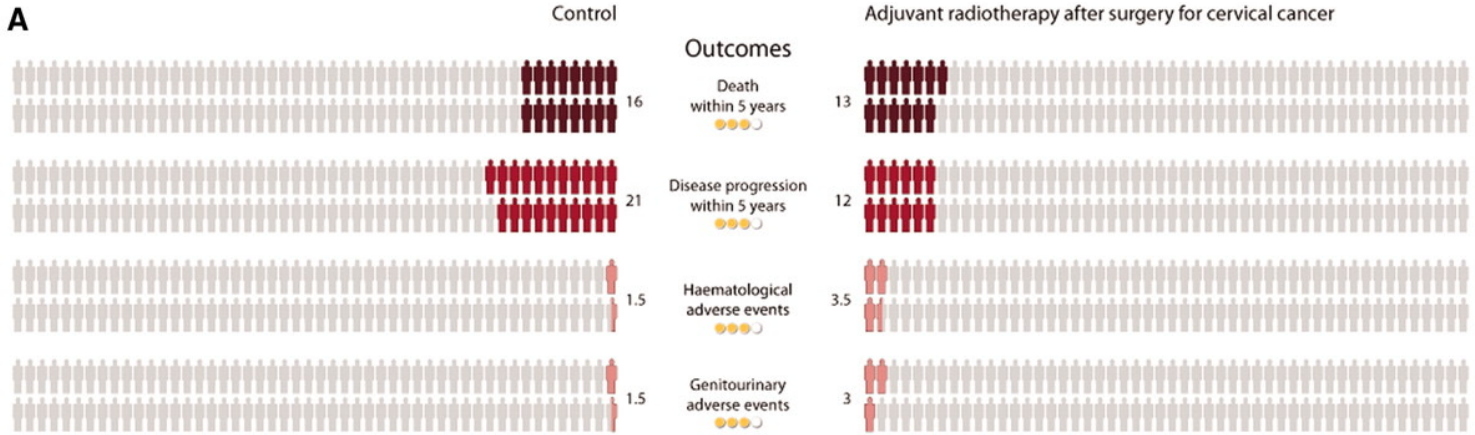
What % of respondents to a U.S. survey that asked the question on the right do you think got it wrong?

25%

"Which of the following numbers represents the biggest risk of getting a disease? 1 in 100, 1 in 1000, or 1 in 10?"

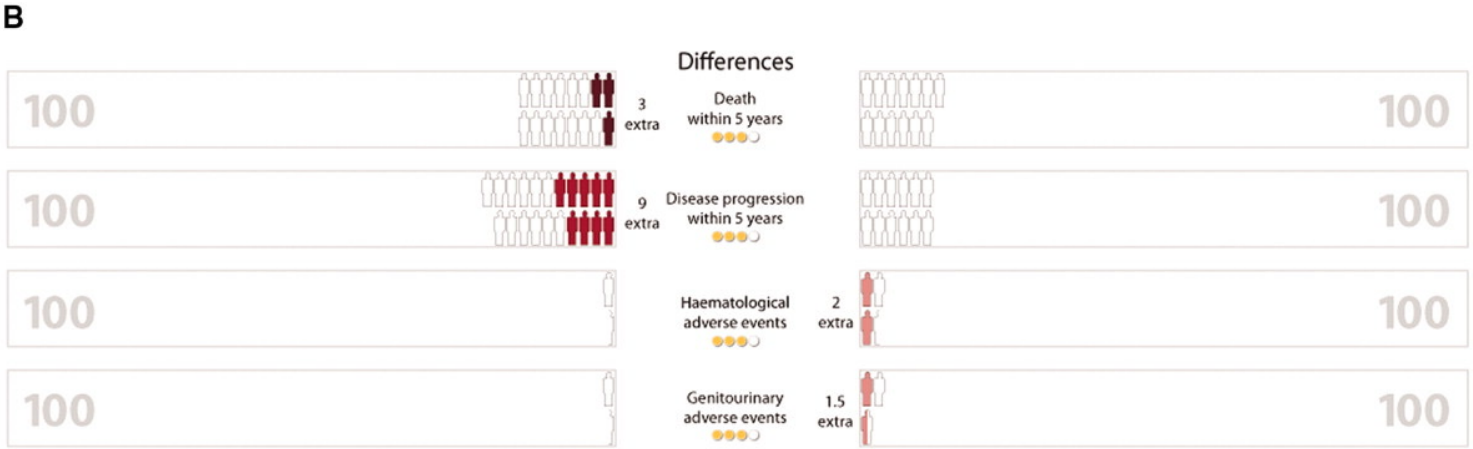
- Spiegelhalter DJ, Pearson M, Short I. 2011Visualizing uncertainty about the future. *Science* **333**, 1393-1400.
- Intergovernmental Panel on Climate Change Working Group 1, *IPCC Fourth Assessment Report: The Physical Science Basis* (IPCC, 2007; www.ipcc.ch/ipccreports/ar4-wg1.htm).

Communicating uncertainty is its own art and science



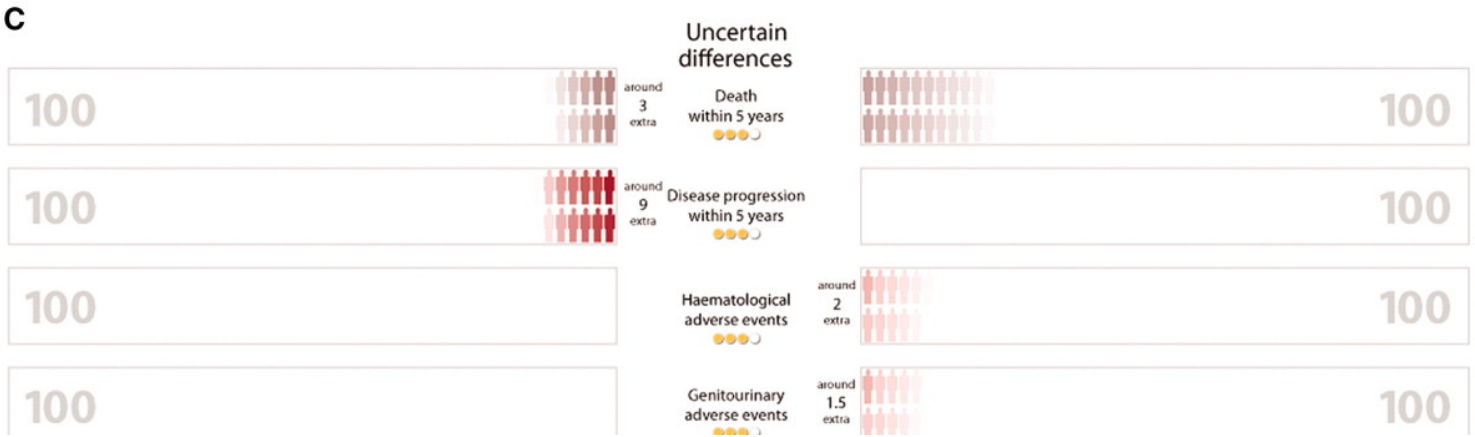
- Spiegelhalter DJ, Pearson M, Short I. 2011Visualizing uncertainty about the future. *Science* **333**, 1393-1400.
- Intergovernmental Panel on Climate Change Working Group 1, *IPCC Fourth Assessment Report: The Physical Science Basis* (IPCC, 2007; www.ipcc.ch/ipccreports/ar4-wg1.htm).

Communicating uncertainty is its own art and science



- Spiegelhalter DJ, Pearson M, Short I. 2011Visualizing uncertainty about the future. *Science* **333**, 1393-1400.
- Intergovernmental Panel on Climate Change Working Group 1, *IPCC Fourth Assessment Report: The Physical Science Basis* (IPCC, 2007; www.ipcc.ch/ipccreports/ar4-wg1.htm).

Communicating uncertainty is its own art and science



- Spiegelhalter DJ, Pearson M, Short I. 2011Visualizing uncertainty about the future. *Science* **333**, 1393-1400.
- Intergovernmental Panel on Climate Change Working Group 1, *IPCC Fourth Assessment Report: The Physical Science Basis* (IPCC, 2007; www.ipcc.ch/ipccreports/ar4-wg1.htm).

Avoiding overfitting is easier said than done

- In any single model development cycle, guidelines for avoiding overfitting are clear:
  - Use cross-validation for model selection and hold-out for performance evaluation
  - Use shrinkage/regularization (L1/L2, complexity penalties, dropout, early stopping)
  - Use feature / stability selection where appropriate
- However, across an agile lifecycle of a model, avoiding overfitting is harder to enforce:
  - Often evaluation benchmarks are fixed over time (so consulted multiple times)
  - Heavy reliance on cross-validation for optimization can result in overfitting the training data
  - Reluctance to publish negative results can over time introduce reproducibility challenges
- In classical statistics, multiple hypothesis testing shares similar challenges, and corrections heavily penalize the confidence of any one finding as a function of the number of “questions asked”.

Avoiding overfitting is easier said than done

- In one of the most heavily cited papers of all time, it was shown that the probability of a finding being successfully reproduced is a function of multiple factors, including:
  - “when there is a greater number and lesser preselection of tested relationships;"
  - “where there is greater flexibility in designs, definitions, outcomes, and analytical modes;”
- These are both considered advantages of the modern, agile approach to data science, rather than risk factors.

PLOS MEDICINE

OPEN ACCESS

ESSAY

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • https://doi.org/10.1371/journal.pmed.0020124

Avoiding overfitting is easier said than done

- Bonferroni and other multiple testing corrections assume that the hypotheses to be tested are selected **in advance**
- This assumption is violated by any form of data-driven screening (e.g., by filtering on absolute correlation between a feature and the response) or variable selection.
- It is also violated in sequential model development, i.e., when we try one thing, then another, then another. The number of "multiple tests" effectively executed there is massive.
- For generalization performance, the cost being data-driven is the need for abundant "fresh" data, whereas hypothesis-driven flows have smaller chance of overfitting and false discovery, but also less potential to uncover subtle relationships.

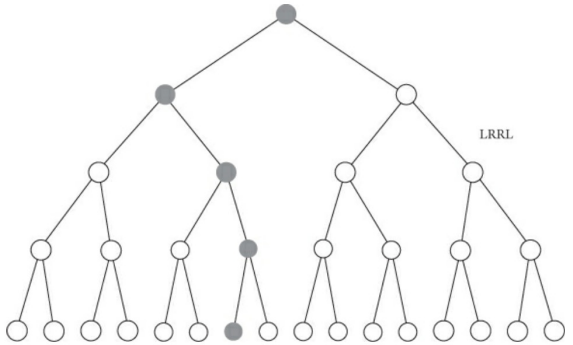


Fig. 26. Tree illustrating the dangers of adaptive data analysis and *p*-hacking. Each level of the tree corresponds to a feature that could be correlated (left) or anti-correlated (right) with the label. The gray path (LRRL) represents the outcomes of the correlation tests. Each leaf corresponds to a classifier that results from a sequence of correlation tests.

Figure 26 from Kearns, Michael, and Aaron Roth. *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press, 2019.

Summary

- Being honest about the reliability of your results is an integral part of trustworthy DS/ML
- This requires expert use of technical tools such as uncertainty quantification; a well-rounded assessment of the quality of evidence at hand (incl. within specialized frameworks in certain cases); appropriate use of language and visualisations to communicate uncertainty clearly
- Although these fields have a wealth of advice and tools for use by the data scientist, they require a culture that is accepting of and patient with negative results, and honesty about “lack of evidence”.
- Nevertheless, use of tools, frameworks and visualisations in a consistent manner is more likely to create trust over time in the role of uncertainty in decision making, and minimize confusion.
- Overfitting is particularly pernicious and does not always happen consciously. Pre-registering the hypotheses under examination or very disciplined hold-out evaluation are trustworthy techniques.

Be brave, systematic, appropriately visual, and conscious of your process.