Ethics of Data Science – Part II

Degrees of evidence

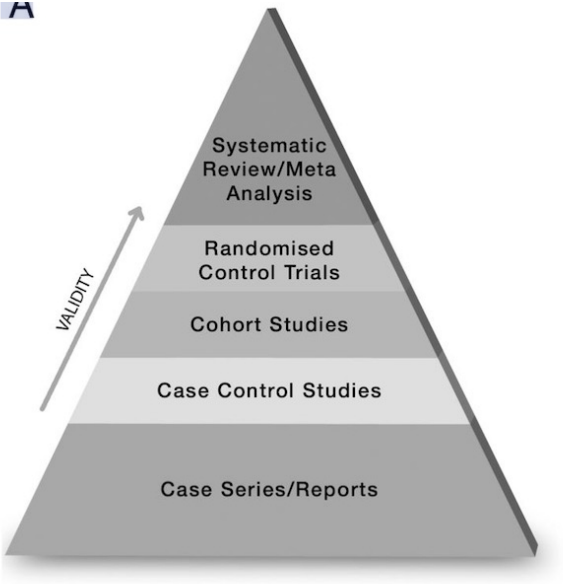
Degrees of evidence

- Generating trust in large part relies on honesty
- Honest communication of uncertainty in data science outputs is a challenging, multi-faceted exercise
- Today we will discuss:
  - General characteristics of analyses that determine the strength of the evidence
  - Specific sources of uncertainty in machine learning models
  - Uncertainty quantification
  - Ways for communicating uncertainty

1. Schünemann, Holger J., et al. "Letters, numbers, symbols and words: how to communicate grades of evidence and recommendations." *Cmaj* 169.7 (2003): 677-680.
2. Van der Bles, Anne Marthe, et al. "Communicating uncertainty about facts, numbers and science." *Royal Society open science* 6.5 (2019): 181870.
3. Abdar, Moloud, et al. "A review of uncertainty quantification in deep learning: Techniques, applications and challenges." *Information Fusion* 76 (2021): 243-297.

General characteristics determining degree of evidence

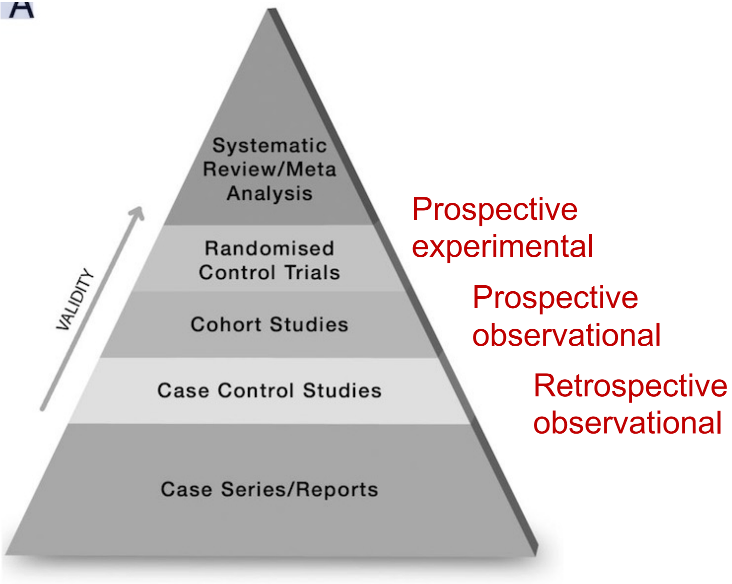
- Before we investigate internal, data-driven ways to measure and understand uncertainty, it is important to take a step back and think about the quality/reliability of the data selection process.
- To complete our journey into the life sciences, let's briefly discuss the "hierarchy of evidence":
  - Anecdotal evidence
  - Observational study with matched controls
  - Prospective observational study (cohort study)
  - RCT
  - Systematic review*



*: Some people disagree with systematic review being on the top of the pyramid

General characteristics determining degree of evidence

- Before we investigate internal, data-driven ways to measure and understand uncertainty, it is important to take a step back and think about the quality/reliability of the data selection process.
- To complete our journey into the life sciences, let's briefly discuss the "hierarchy of evidence":
  - Anecdotal evidence
  - Observational study with matched controls
  - Prospective observational study (cohort study)
  - RCT
  - Systematic review*



*: Some people disagree with systematic review being on the top of the pyramid

General characteristics determining degree of evidence

- More generally, within the predictive machine learning context, we can think about the representativeness of the evaluation data in particular:
  - Is our evaluation dataset representative of the future deployment of the algorithm?
  - How much variability in our estimated generalization performance is there?
  - Are we looking to re-train the model in the future? How sensitive is performance to that?
  - Is our performance metric the right / only thing to measure?

General characteristics determining degree of evidence

- More generally, within the predictive machine learning context, we can think about the representativeness of the evaluation data in particular:
  - Is our evaluation dataset representative of the future deployment of the algorithm?
  - How much variability in our estimated generalization performance is there?
  - Are we looking to re-train the model in the future? How sensitive is performance to that?
  - Is our performance metric the right / only thing to measure?

- Dataset shift can take many forms. In supervised learning y = f(X), for example:
  - **Covariate shift** is when the distribution of X changes but f stays constant. This can still have detrimental performance when f(.) is a flexible function as it might be required to extrapolate to regions of X it had not observed prior (e.g., very large values of X)
  - **Concept shift/drift** is when the relationship f(.) changes over time.
  - **Prior shift** is when the two classes stay the same but one becomes more prevalent over time.

Specific sources of uncertainty in machine learning

Predictive accuracy is driven by a number of sources of “error”:

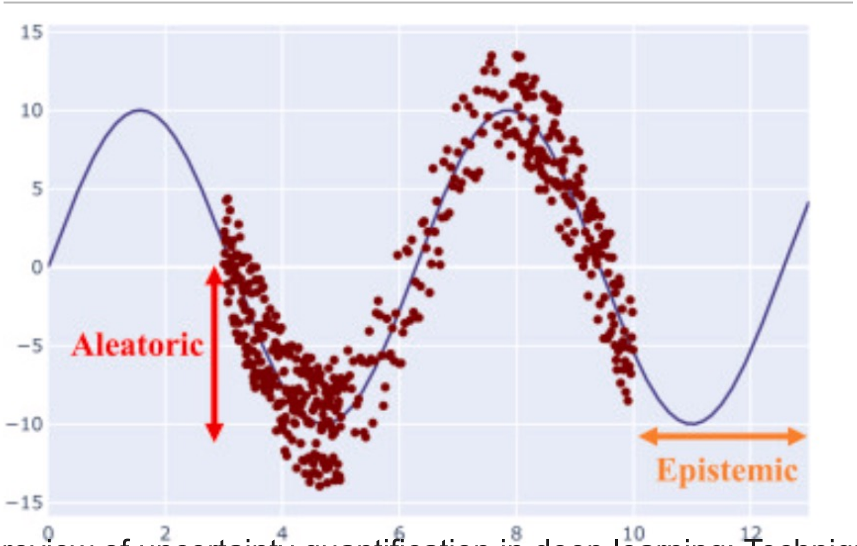
- Estimation bias (persists in larger sample sizes, generally larger for simple models)
- Estimation variance (increases with more complex models, vanishes with large samples)
- Variance within the ML algorithm (e.g., seed in random forest)
- Bias/variance in estimated accuracy (due to dataset shift and holdout sample size)
- “Irreducible” variance (e.g., natural variation in “outcome” variance)

Epistemic

Aleatoric

Some model classes (e.g., Bayesian modelling) produce an estimate of their own uncertainty – others do not, and wrapper methods need to be applied to produce an estimate.

Specific sources of uncertainty in machine learning

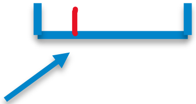


Abdar, Moloud, et al. "A review of uncertainty quantification in deep learning: Techniques, applications and challenges." *Information Fusion* 76 (2021): 243-297.

Specific sources of uncertainty in machine learning

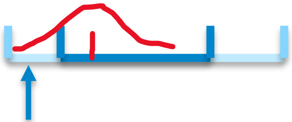
A useful distinction is that between a confidence interval and a prediction interval:

A confidence interval for an estimator will contain the true value of the estimand 95% of the time.



This could be the real value of the mean

A prediction interval extends the confidence interval to also reflect the irreducible variance.



This could be the value of a single future observation

Self-reported estimates of uncertainty rely themselves on (2nd order) estimates that are subject to noise. Bayesian models naturally report prediction intervals. Frequentist models can too, but it's hard.

Specific sources of uncertainty in machine learning

Unlike Bayesian methods, deep and non-parametric learning does not have a mechanism to produce uncertainty about the model's own predictions, neither at the level of individual parameters (posterior uncertainty on a neural network weight) nor at the level of the prediction (predictive posterior).

A number of techniques that leverage re-sampling as a way to assess the sensitivity of the prediction on the sample uncertainty have appeared that can produce approximate posteriors from DNNs, including **drop-out,** where multiple predictions from the NN are obtained by re-sampling which units to drop, e.g.,:

$$\mathbb{E}_{q(y^\star|x^\star)}[y^\star] \approx \frac{1}{T}\sum_{t=1}^{T} \hat{y}^\star(x^\star, W_1^t, \ldots, W_L^t)$$

Eq (6) from the seminal paper from Gal, Yarin, and Zoubin Ghahramani. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning." *international conference on machine learning*. PMLR, 2016.

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