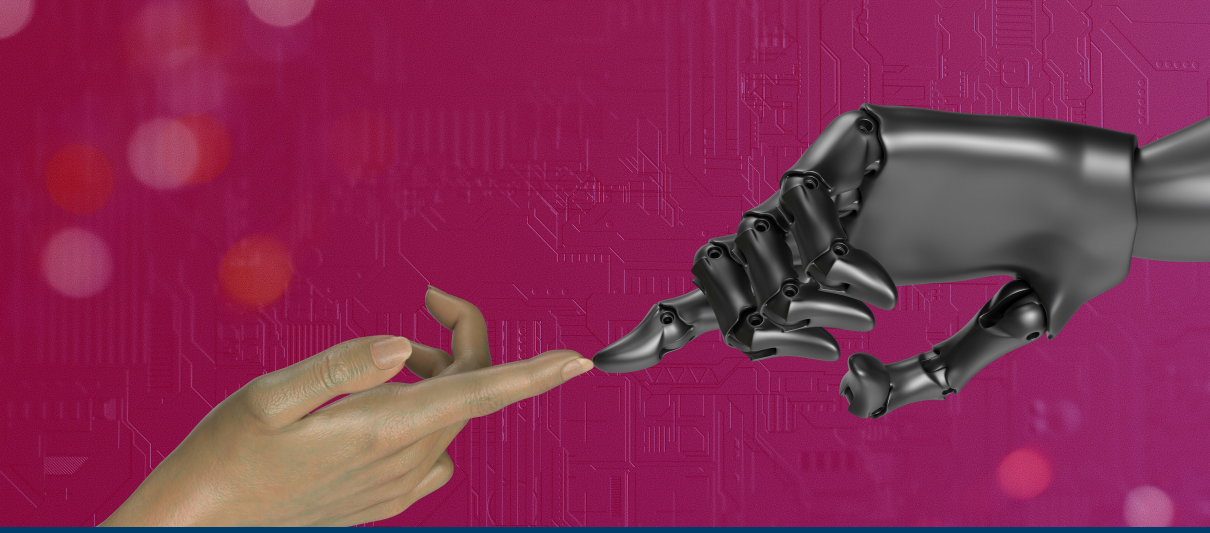


Hands-On Machine Learning

Dr. Derek Bridge | 2026/2027

UCC Coláiste na hOllscoile Corcaigh
University College Cork, Ireland

School of Computer Science and Information Technology
Scoil na Ríomheolaíochta agus na Teicneolaíochta Faisnéise

21. Bias in Machine Learning

In Machine Learning, the word "bias" has numerous meanings. Most describe ways in which a model's predictions may be skewed in some way — especially where the skew gives rise to unfair treatments.

But we start with a couple of very different definitions — ones that are not really about fairness or unfairness — just in case you encounter them in future study. These definitions are often presented mathematically — we look at informal definitions.

Only then do we look at the kind of bias that can lead to unfairness. This is a huge topic — and an active area of research. We only skim the surface in this lecture.

We use a synthetic dataset, with deliberate bias. It comes from The Centre for Data Ethics and Innovation.

Inductive Bias

- The **inductive bias** of a learning algorithm is the set of assumptions that underpin the model that it learns.
- The inductive bias restricts the kind of hypotheses (candidate models) that the learner chooses between.
- E.g. Linear Regression only chooses between linear models.
- E.g. Lasso Regression has a bias towards sparsity.
- E.g. Decision trees learn step-functions.

No Free Lunch Theorems

"if an algorithm performs well on a certain class of problems then it necessarily pays for that with degraded performance on the set of all remaining problems."

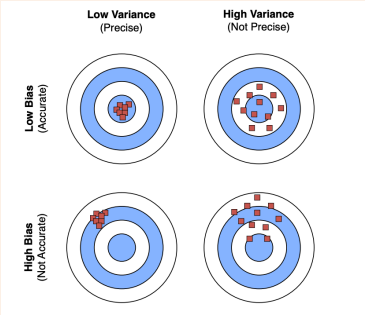
— D.H. Wolpert & W.G. Macready: *No Free Lunch Theorems for Optimization*". IEEE Transactions on Evolutionary Computation, vol.1, pp.67–82, 1997.

Errors: Bias and Variance

$$\hat{y} = y + \textit{bias error} + \textit{variance error} + \textit{irreducible error}$$

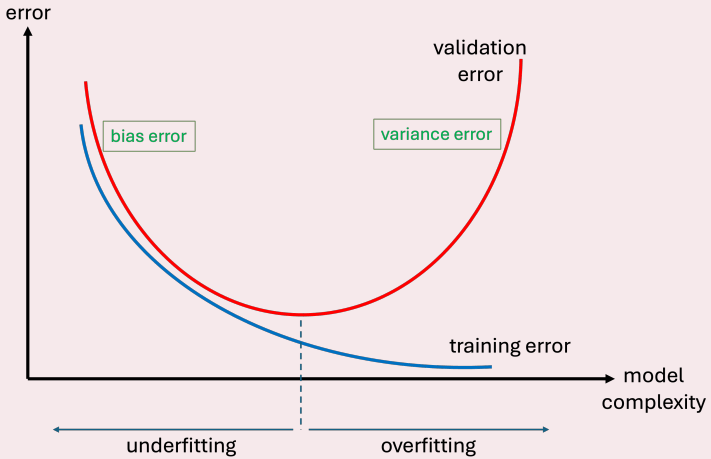
- **Bias error** is caused by choosing a learning algorithm that makes the wrong assumptions. Most commonly, the cause is a model of insufficient complexity. Hence, this is related to underfitting.
- **Variance error** is caused by sensitivity to small variations in the training set. Most commonly, the cause is a model of excessive complexity. Hence, this is related to overfitting.
- **Irreducible error** is error we cannot reduce by choosing a different model. It is caused, e.g., by randomness in the problem.

- Imagine training a model on different subsamples of the training set.
- Bias is about the magnitude of the errors.
- Variance is about the spread of the errors.



The Bias-Variance Trade-Off

- Typically, by increasing complexity, we trade-off bias for variance.



Decision-Making Bias

- In decision-making, **bias** means that disproportionate weight is given to a certain factor: based on this factor, certain outcomes become more likely than others.
- If, according to our principles, we deem this factor to be irrelevant to our decision-making task, then the decision-making is **unfair**.
- In particular, if the factor is what we deem to be a **sensitive feature**, such as race, ethnicity, age, sex, sexual orientation, physical and mental traits, ideology, interests, location, language, income, etc., then we talk of **discrimination**.
- Systematic, unfair discrimination is **prejudice**.

Why Care?

- The systems we build and deploy should reflect our values.
- There are laws against discrimination.

Discuss

"The features that we have used in our model exclude race. Therefore, our model cannot be racist."

Discuss

"The way to overcome sampling bias is to get a larger dataset."

Bias in ML

- Individual people, groups of people and whole societies exhibit bias.
- This bias will be reflected in our datasets.
- But, when developing ML models, we may amplify existing bias — or even introduce new sources of bias.
- Bias in ML may be introduced at any stage: data collection, data labelling, data cleaning, data preprocessing, learning itself, the evaluation of results in both model selection and error estimation, the interpretation of the results in deciding whether to deploy, and the on-going operation of any deployed system.

Exercise

Describe how, in each of the following stages of an ML project, we may bring bias into our model.

- Data collection;
- Data labelling;
- Data cleaning;
- Data preprocessing.

Amazon Case Study

Quoting at length from Amazon scraps secret AI recruiting tool that showed bias against women, by Jeffrey Dastin, October 2018, Reuters News Agency

- Amazon had been building computer programs since 2014 to review job applicants' resumes with the aim of mechanizing the search for top talent...
- But by 2015, the company realized its new system was not rating candidates for software developer jobs and other technical posts in a gender-neutral way.
- That is because Amazon's computer models were trained to vet applicants by observing patterns in resumes submitted to the company over a 10-year period. Most came from men, a reflection of male dominance across the tech industry.
- It penalized resumes that included the word "women" as in "women's chess club captain"...
- Amazon edited the programs to make them neutral to these particular terms. But that was no guarantee that the machines would not devise other ways of sorting candidates that could prove discriminatory...
- The group created 500 computer models focused on specific job functions and locations. They taught each to recognise some 50,000 terms that showed up on past candidates' resumes. The algorithms learned to assign little significance to skills that were common across IT applicants, such as the ability to write various computer codes,... Instead, the technology favoured candidates who described themselves using verbs more commonly found on male engineers' resumes, such as "executed" and "captured"...
- Amazon's recruiters looked at the recommendations generated by the tool when searching for new hires, but never relied solely on those rankings...
- The Seattle company ultimately disbanded the team by the start of last year because executives lost hope for the project...
- Amazon declined to comment on the technology's challenges, but said the tool "was never used by Amazon recruiters to evaluate candidates."
- ...the company managed to salvage some of what it learned from its failed AI experiment. It now uses a "much-watered down version" of the recruiting engine to help with some rudimentary chores, including culling duplicate candidate profiles from databases...
- ...a new team in Edinburgh has been formed to give automated employment screening another try, this time with a focus on diversity.

Exercise

The project was a failure. Identify the failings.

Bias Detection

- Aggregate evaluation, e.g. MAE or accuracy for the whole validation/ test set, will hide bias.
- We must additionally calculate metrics for different subgroups and combinations of subgroups and look for parity.
- But you will need to decide in advance what you regard as an equitable outcome.
- E.g. if your model has similar accuracy across subgroups, is that fairness?

COMPAS Case Study

Paraphrasing Wikipedia and the article Machine Bias, published by ProPublica.

- COMPAS is a decision-support tool, developed by Northpointe (now Equivant).
- U.S. courts use it to predict the likelihood that a defendant will re-offend (recidivism).
- It is used to guide pre-trial release, sentencing and parole.
- It uses 137 features extracted from the defendant's criminal record and from answers to questions (e.g. "Was one of your parents ever sent to jail or prison?"). Race is not one of the features.
- The predicted probability of re-offending is calculated by a weighted formula — the weights are proprietary.
- Overall accuracy, based on a 2-year follow-up period, is about 61% — a bit better than a coin toss and about the same as a random member of the public.

- The argument in favour of COMPAS is that it is consistent in its decision-making — e.g. it does not exhibit the hungry-judge phenomenon.
- One general argument against is that, being proprietary, affected parties are not able to examine its operation in general (interpretability) or in a particular case (explainability).

COMPAS Case Study

- Investigative journalists from non-profit ProPublica obtained data about COMPAS predictions.
- From their analysis, they claim "blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend" (higher false positives for African Americans) and "[whites] are much more likely than blacks to be labeled lower-risk but go on to commit other crimes" (higher false negatives for whites).

	Caucasian	African American
False Positive Rate	23%	45%
False Negative Rate	48%	28%

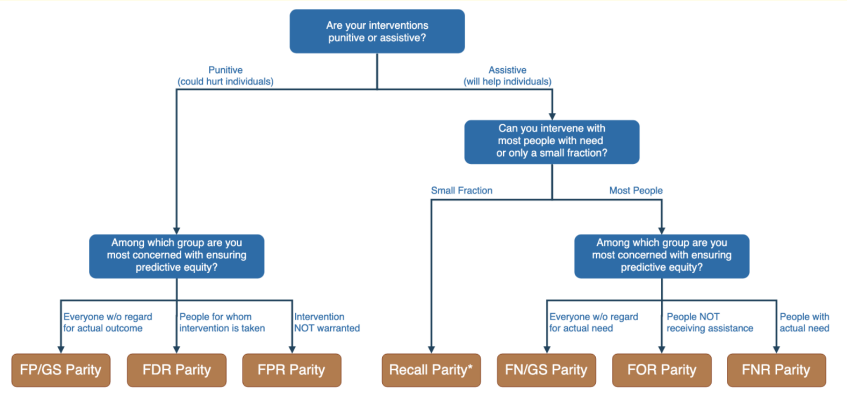
- Their analysis has been disputed in lots of ways.
- Northpointe's defence is that the Precision is similar for the two groups — it is equally accurate.

	Caucasian	African American
Precision	59%	63%

- Who is right?
- It can be shown that if prevalence is unequal across groups and precision is equal across groups, then either FPR or FNR can be equal across groups but not both.

The Fairness Tree

The **Fairness Tree** is not meant to be used slavishly. It is intended to guide the conversation about what metrics are appropriate to your project.



GS = group size

Bias Mitigation

Fix the Data

- Gather more examples.
- Augment with synthetic examples.
- Upsample or downsample.
- Stratify additionally on the sensitive feature.

Fix the Learning Algorithm

- Use weighted examples.
- Use a loss function that includes your definition of fairness.

Fix the Evaluation

- Use your chosen fairness metric in model selection and error estimation.

Fix the World!

Ummm...

Detecting and Mitigating Bias in ML

- Some bias may be inevitable — but that doesn't mean we should do nothing.
- Technical 'fixes' are not enough.

Culture

Organizational culture is key.

Diversity

Diversity in ML teams is key.

Measurement

Identification and measurement of possible harms is key.

Process

The right process, from end-to-end, is key.

Image credits

Taken from SebastianRaschka's STAT451: Introduction to Machine Learning Lecture Notes

Taken from Kit T. Rodolfa, Pedro Saleiro & Rayid Ghani: *Bias and Fairness*, Chapter 11 in Ian Foster, Rayid Ghani, Ron S. Jarmin, Frauke Kreuter & Julia Lane (eds.), *Big Data and Social Science (2nd edn.)*