How to avoid machine learning pitfalls

a guide for academic researchers

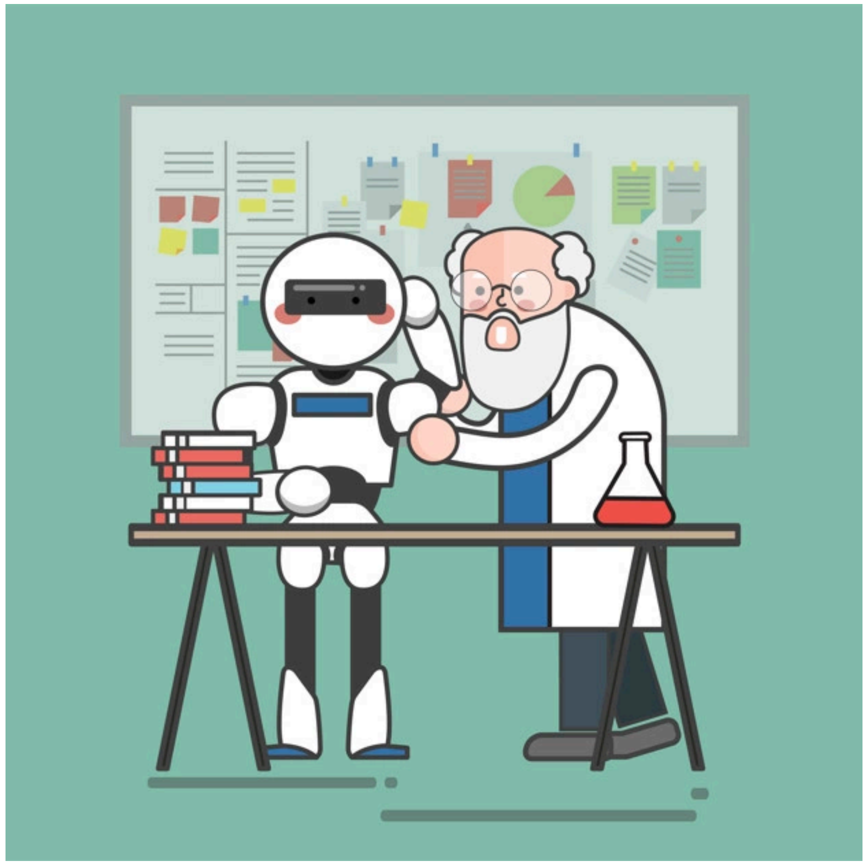
Michael A. Lones

Sermkiat Lolak 🔮

- Help newcomers avoid some of the mistakes

- ML within an academic research context

- Informally, in a **Dos and Don’ts** style.

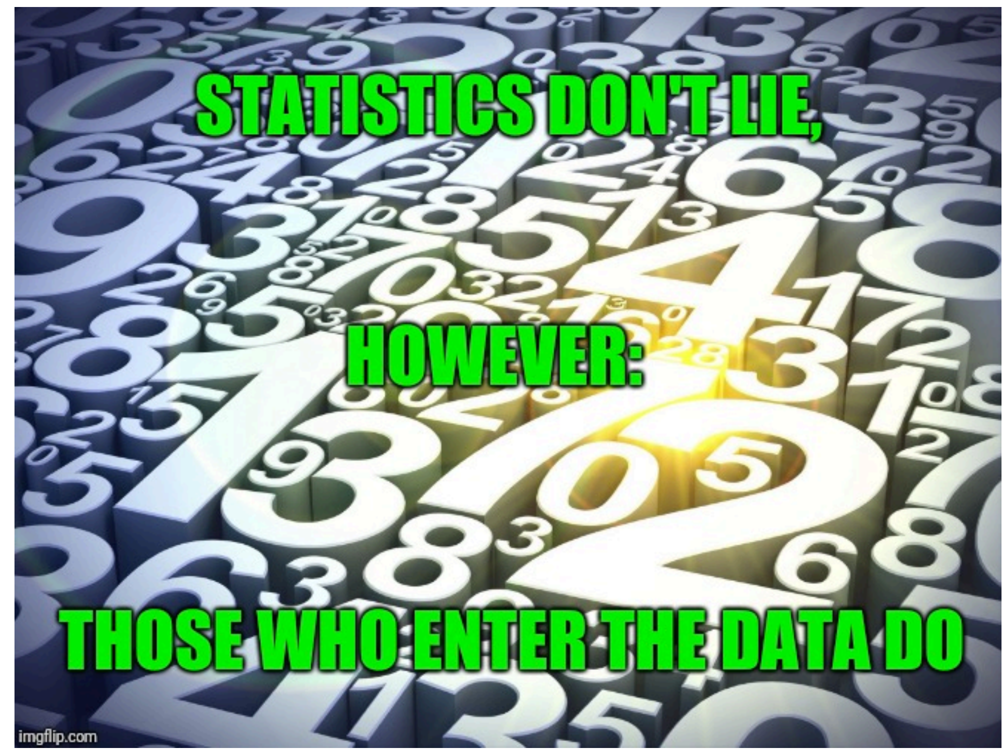


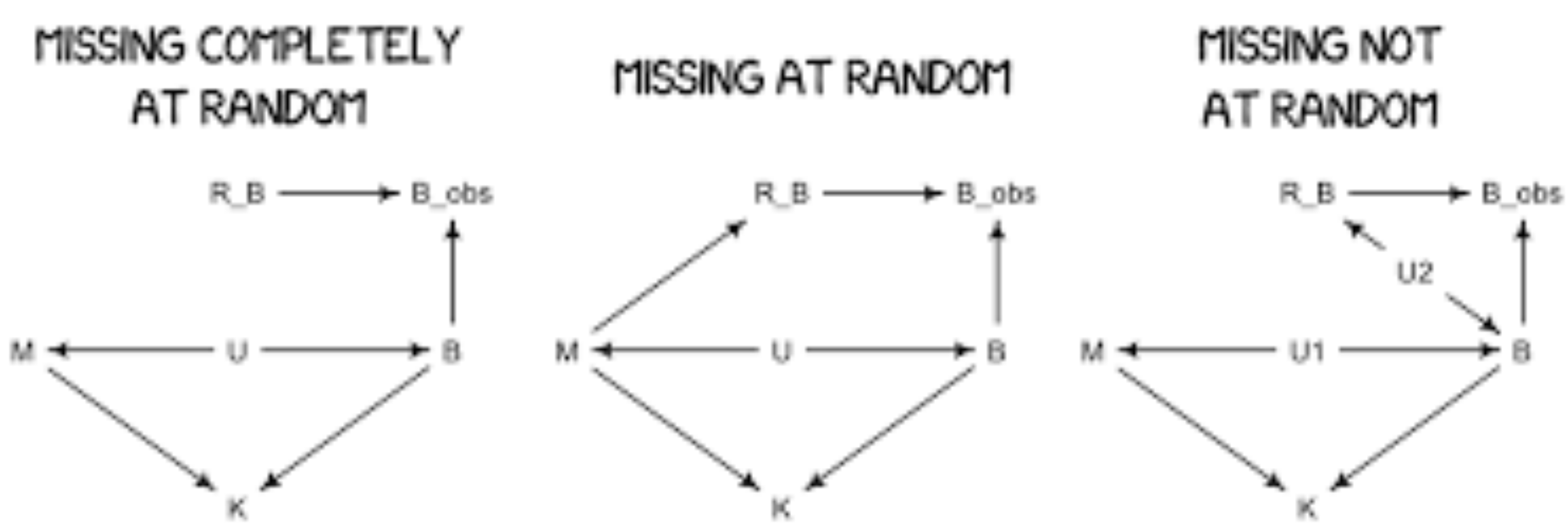
Before you start to build models

- Do take the time to understand your data
- Don’t look at all you data
- Do make sure you have enough data
- Do talk to domain experts
- Do survey the literature
- Do think about how your model will be deployed

Do take the time to understand your data

- Public dataset? Published?
- **garbage** in **garbage** out
- Exploratory data analysis
- Look for missing or inconsistent records





Don’t look at all you data

- OK to spot patterns and make insights (training set)

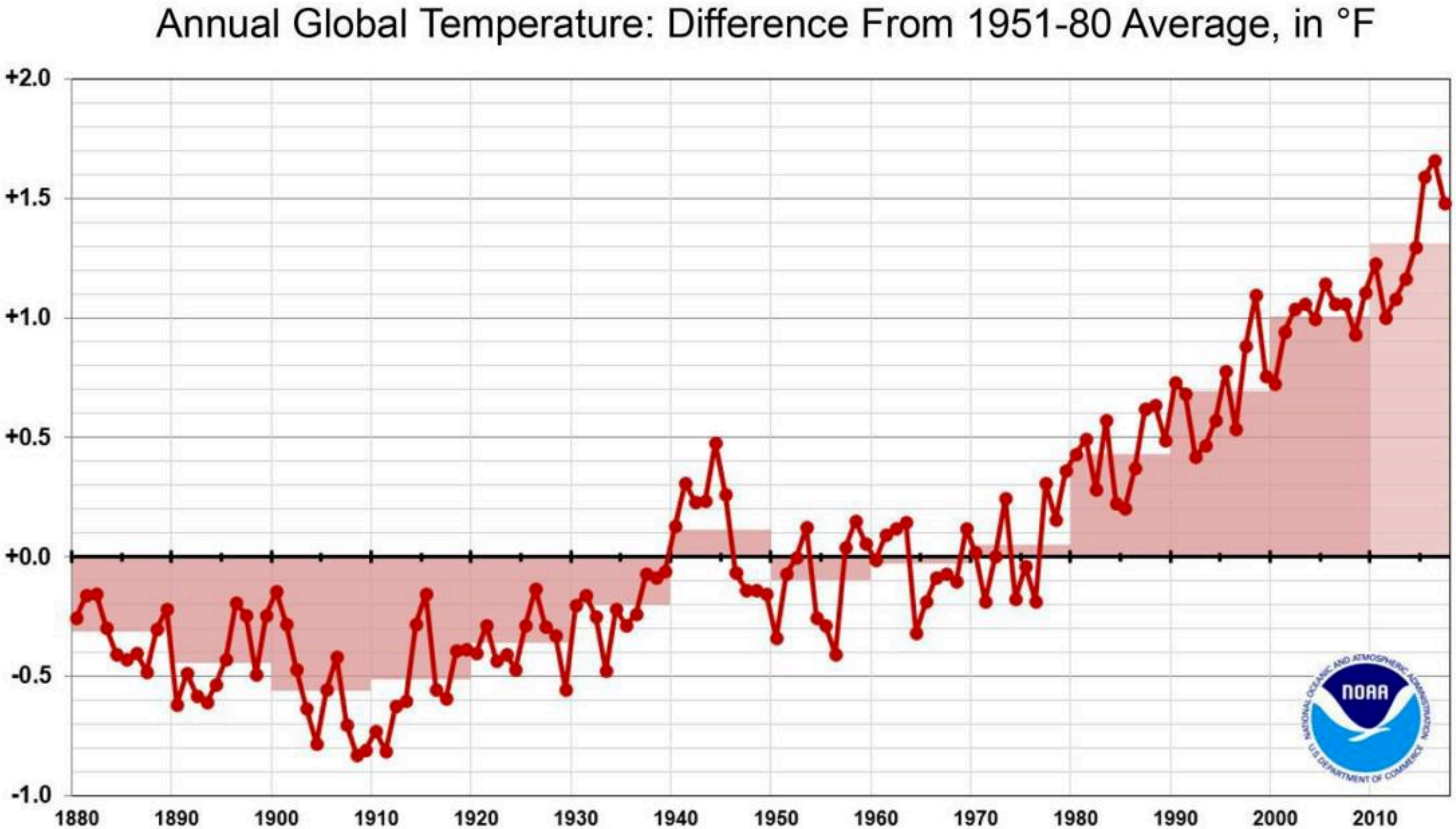
- Made assumption **only** in training set

- avoid looking closely at any test data

- Else, limit the generality of model in an untestable way.

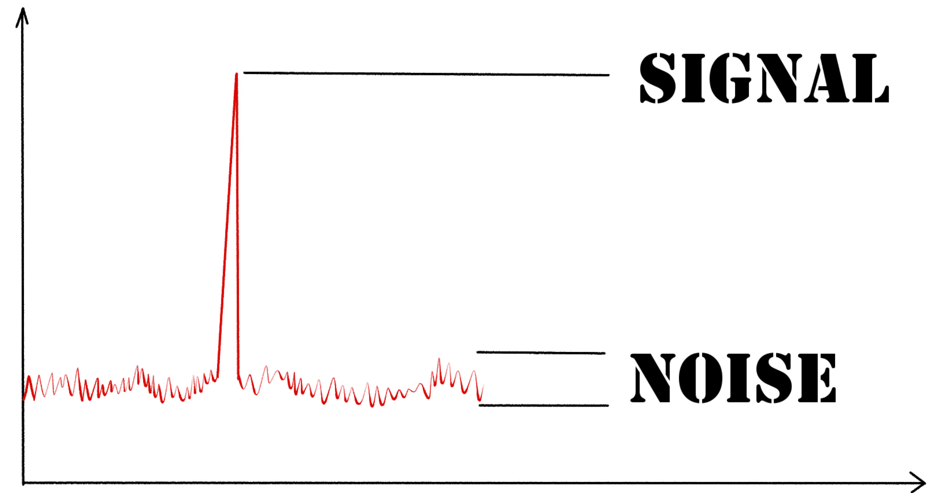
Global Temperature Time Series

NOAA GlobalTemp

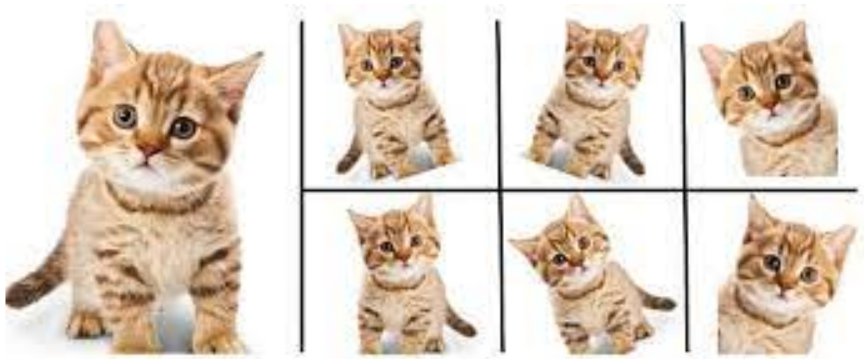


Do make sure you have enough data

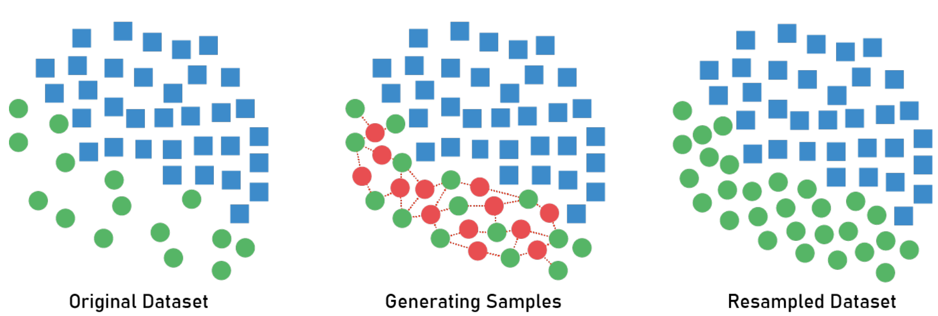
- signal to noise ratio in the data set



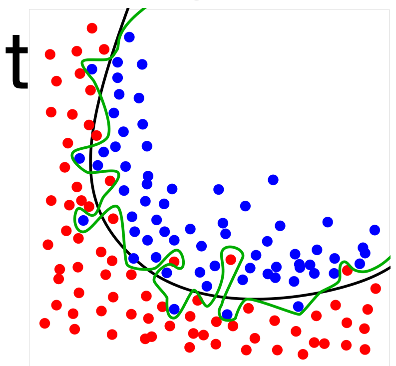
- Use Cross validation
- Data augmentation

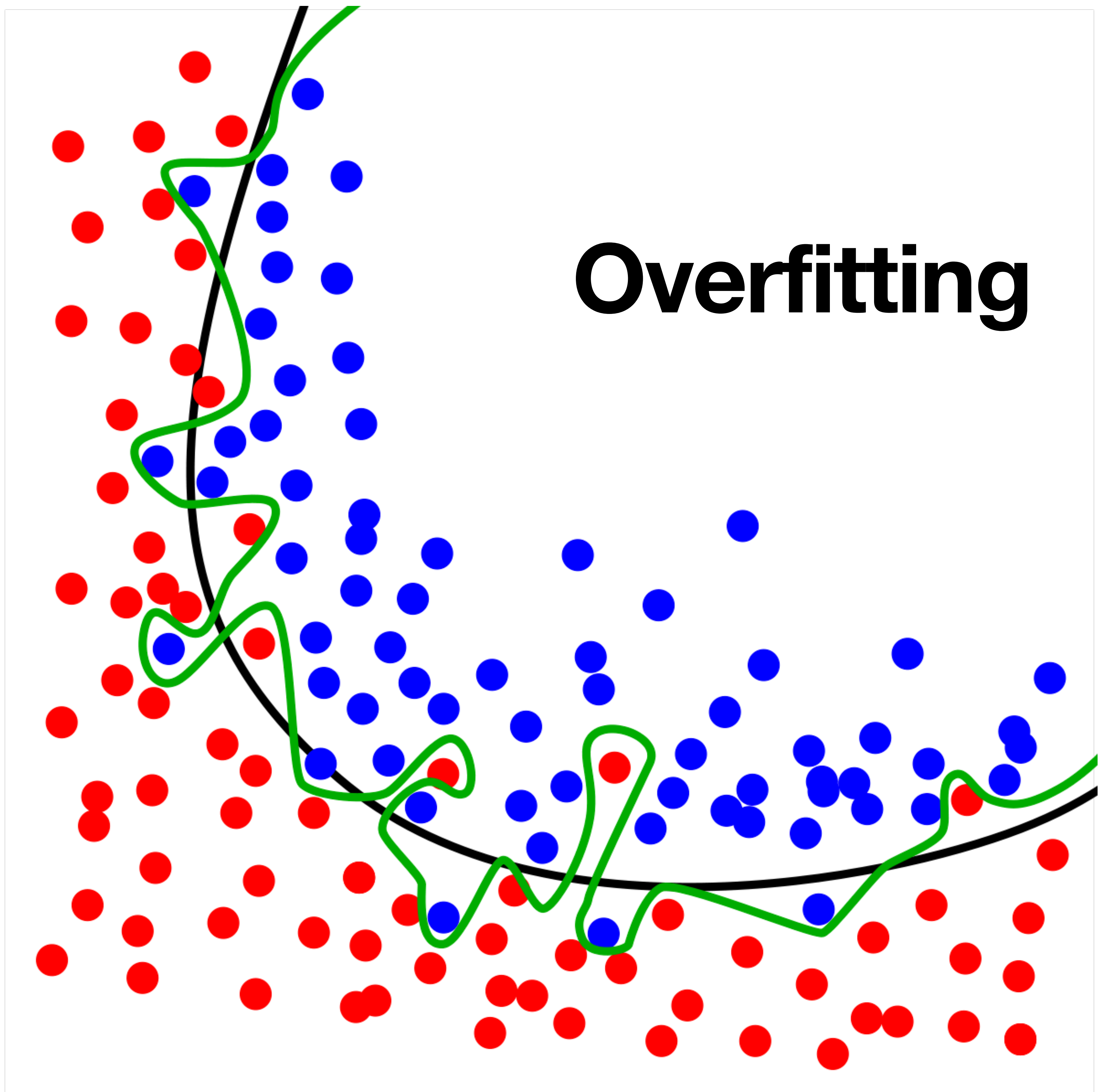


Synthetic Minority Oversampling Technique



- Augment minor class in Imbalanced dataset
- Limit model complexity —> prevent overfit




Overfitting

Do talk to domain experts

- choose the most appropriate feature set and ML model to use

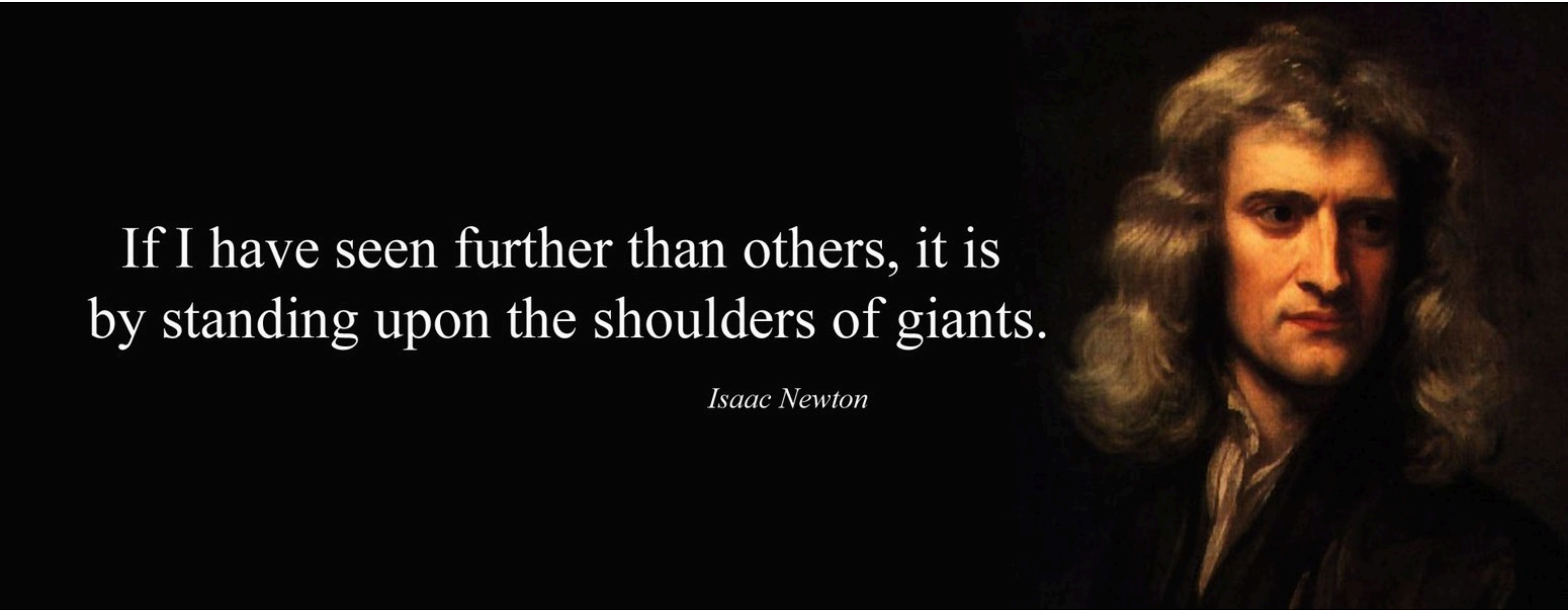
- publish to the most appropriate audience

- help you to understand the data

- Example : Opaque model where it need transparent

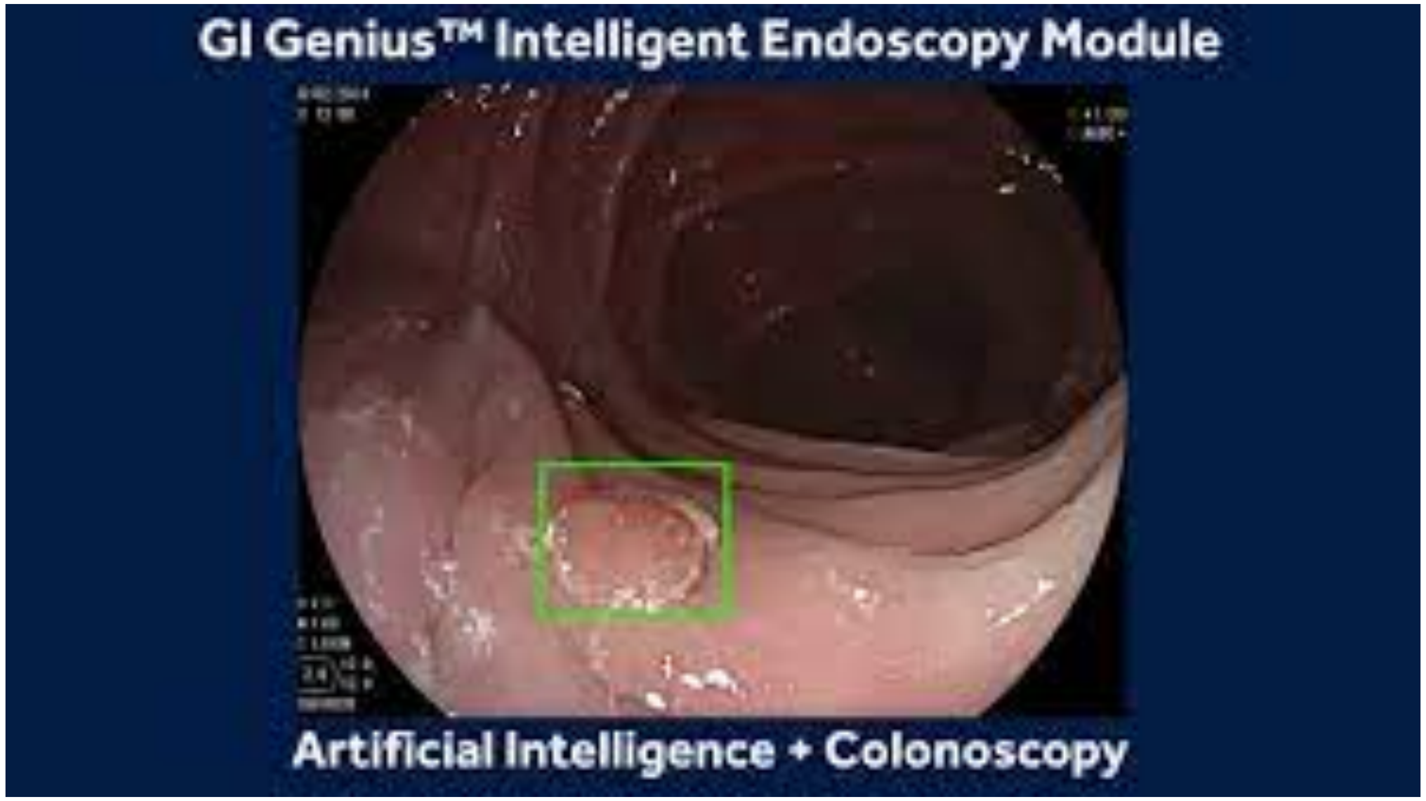
Do survey the literature

- Other people having worked on the same problem isn't a bad thing
- most likely left plenty of avenues of investigation still open



Do think about how your model will be deployed

- Why do you want to build an ML model?
- paper vs real-world
- resource-limited environment , milliseconds response?
- ML Ops



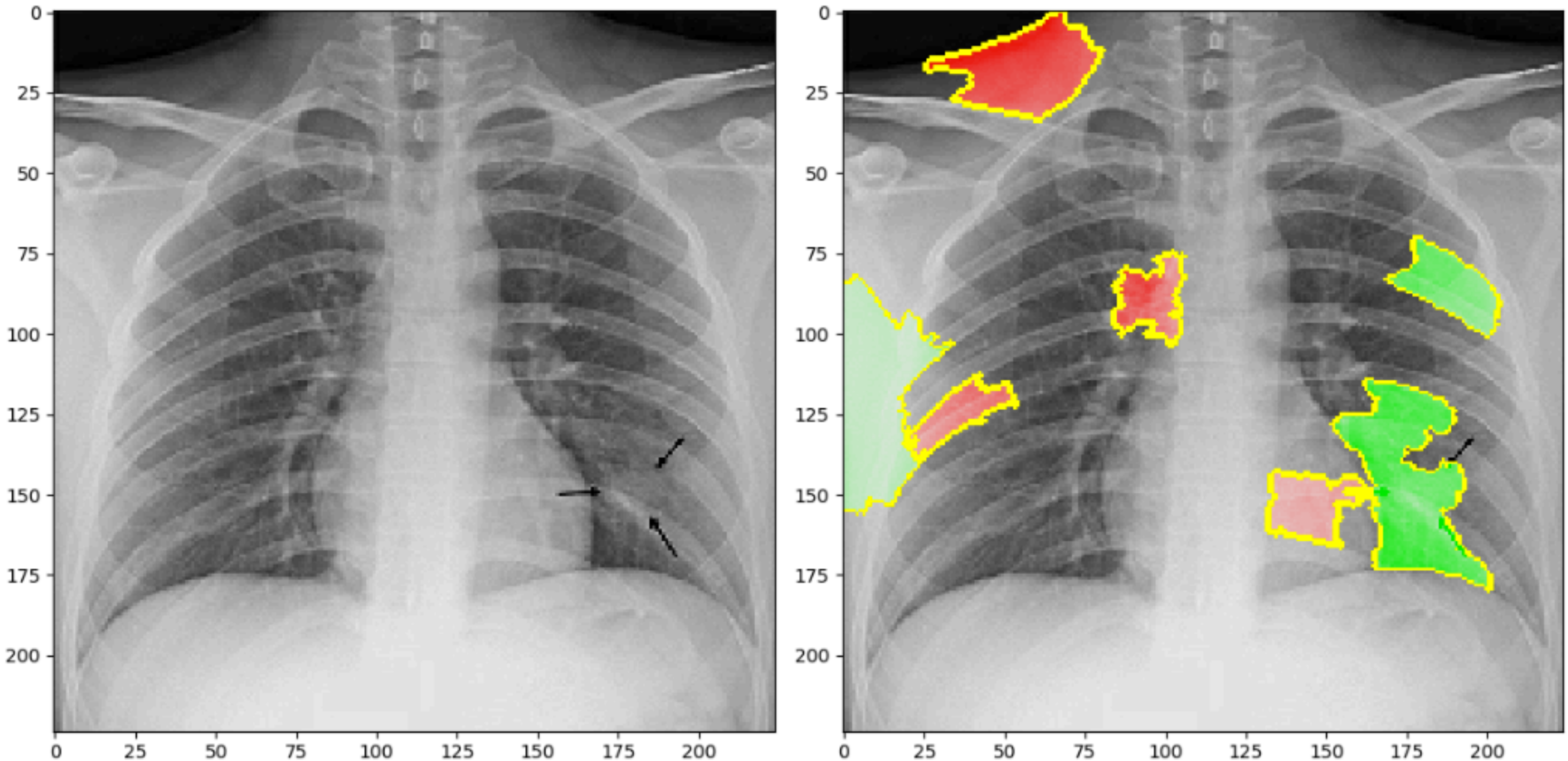
How to reliably build models

- Don’t allow test data to leak into the training process
- Do try out a range of different models
- Don’t use inappropriate models
- Do optimise your model’s hyperparameters
- Do be careful where you optimise hyperparameters and select features

Don’t allow test data to leak into the training process

- common reason : published ML models **FAIL**
  - ❌ whole data set variable scaling
  - ❌ feature selection before partitioning the data
  - ❌ using the same test data to evaluate the generality of multiple models —> **over-fit the test set**
  - ✅ use independent test set once to measure the generality of a single model at the end of the project

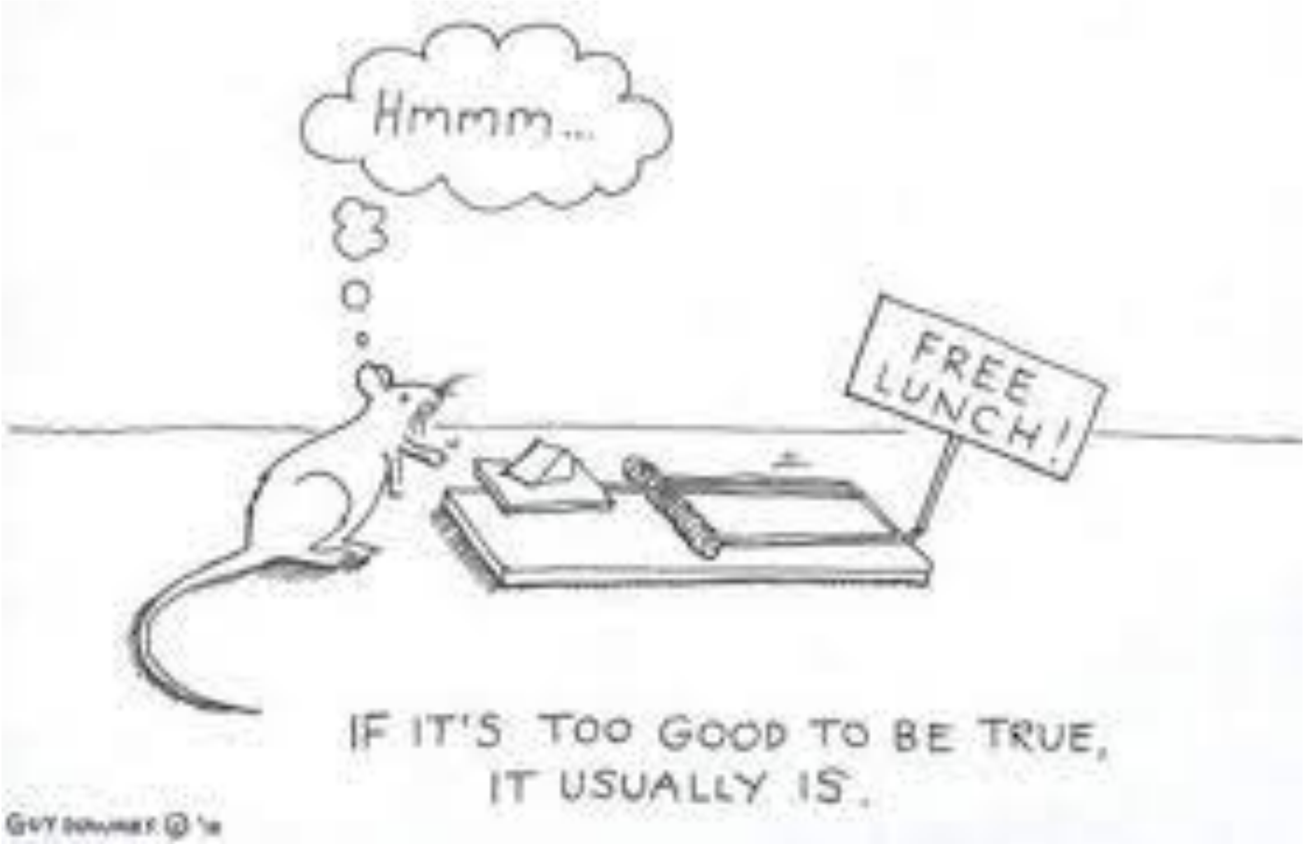
Ground Truth Class: 1 (COVID-19)
Predicted Class: 1 (COVID-19)
Prediction probabilities: ['0.01', '0.99']



Do try out a range of different models

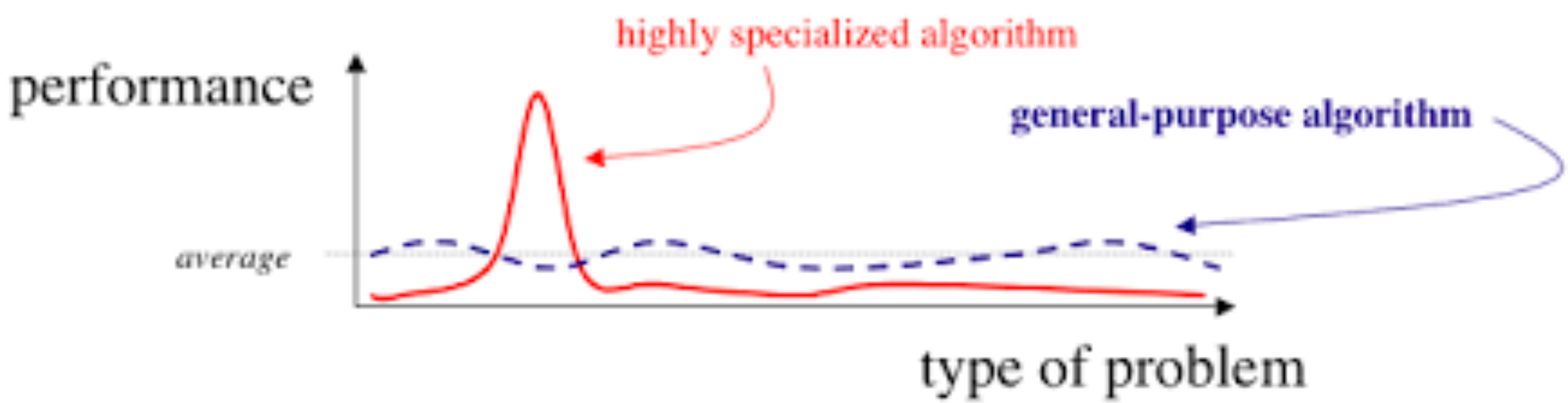
- **No Free Lunch theorem** : no single ML approach is best for every possible problem

- Find the ML model that works well for particular problem.



No-free-lunch theorem in ML

	Algorithm 1	Algorithm 2	Algorithm 3	Algorithm 4	Algorithm 5	Algorithm 6	Algorithm 7	Algorithm 8	...
Problem 1	78.90158096	38.18696053	83.9788141	3.128185533	93.71767489	3.612131384	38.02555482	46.02033283	...
Problem 2	63.63661246	51.21726878	6.915100117	92.46504485	20.63056606	90.15194724	6.628150576	88.92628997	...
Problem 3	5.467817525	78.82129795	19.01963224	16.18471759	59.57316925	26.61430506	41.45446652	62.38540108	...
Problem 4	40.96337067	55.59045049	25.47959077	77.75563723	90.98183523	42.23275523	92.4381591	80.17316672	...
Problem 5	17.32640301	80.17604054	48.01380213	9.378352179	13.25844413	66.24497877	17.39991202	46.86218446	...
Problem 6	2.90117365	14.18732284	88.12091607	28.32526953	88.17950692	43.16349405	78.48956349	76.09121009	...
Problem 7	74.22339559	71.35440724	46.26625983	69.9710712	66.9510279	68.97533166	14.29350951	56.8139594	...
Problem 8	69.06790479	89.53420767	17.7105817	71.3419208	48.8622438	3.348772613	70.81053152	3.855765825	...
Problem 9	19.94675498	3.137513385	10.68373549	4.011603637	49.49135388	37.92530089	99.49914362	54.10622766	...
Problem 10	7.510870987	58.55534993	57.60647147	80.17271882	80.41639739	25.77488384	55.59960103	94.67596268	...
Problem 11	98.30840803	40.16271408	15.063453	80.71102508	67.38435353	2.092705478	54.93369837	34.34560747	...
Problem 12	56.35291015	99.47783881	73.23060569	79.11112105	58.89165367	51.21548188	72.3854659	54.63516655	...
Problem 13	42.95441914	5.055088383	20.45995021	60.02150262	2.129162205	0.03549031414	90.26590811	1.821852475	...
Problem 14	44.26664262	55.68963431	33.72502344	56.30721179	88.24480947	42.89040502	29.76489645	6.234549423	...
Problem 15	91.00330356	24.51201295	90.63002494	53.41813975	93.87696033	28.00711639	23.69333881	40.15298867	...
...	...	...	...	...	...	...	...	...	...
Average	100	100	100	100	100	100	100	100	



Don’t use inappropriate models

- modern ML libraries : easy to apply inappropriate models
- ❌ put numeric features to models that expect categorical features
- ❌ put time series data to model expecting i.i.d
- ❌ unnecessarily complex
- reporting results from

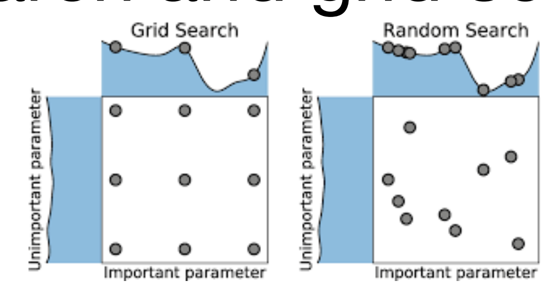
inappropriate models :

piss reviewers

Do optimise your model's hyperparameters

- significantly effect the performance : no one-size-fits-all.

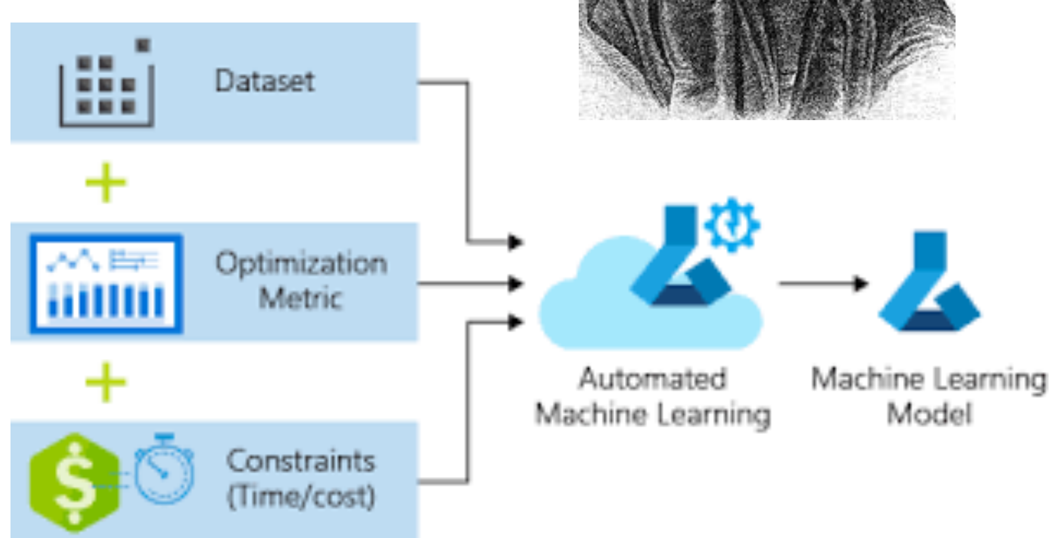
- **hyperparameter optimisation** :random search and grid search, but might not scale well



- **Bayesian Optimisation**

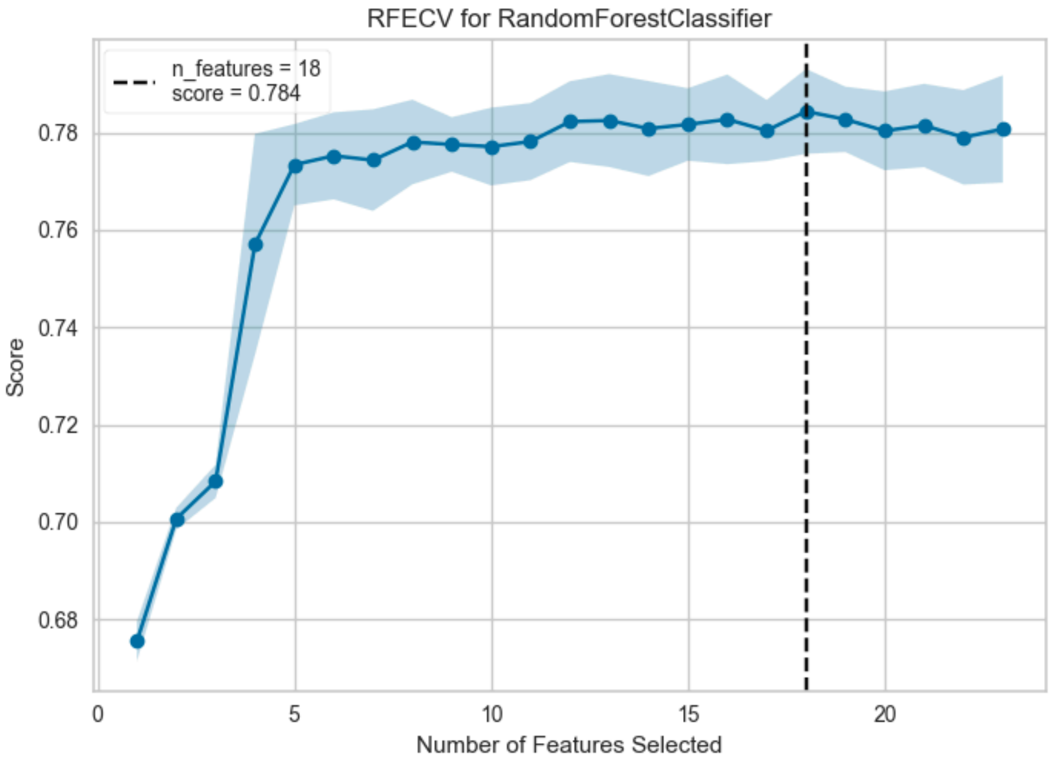
Me again

- AutoML

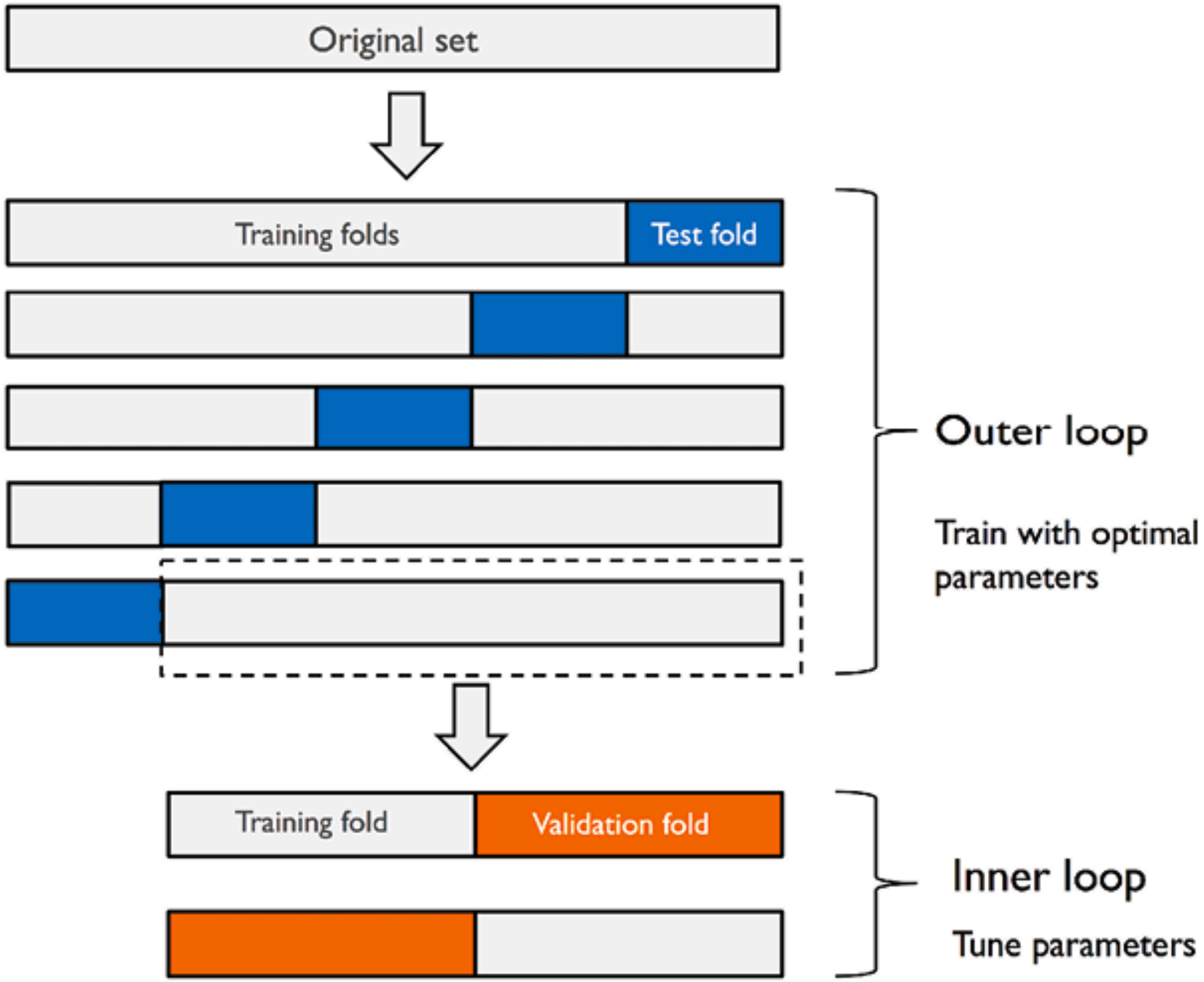


Do be careful where you optimise hyperparameters and select features

- Hyperparameter optimisation and feature selection : Do treat them as part of **model training**
- ❌ common error : feature selection on whole data —> information leaking
- nested cross-validation (double cross-validation) eg. RFECV , GridsearchCV



Nested CV / Double CV



How to robustly evaluate models

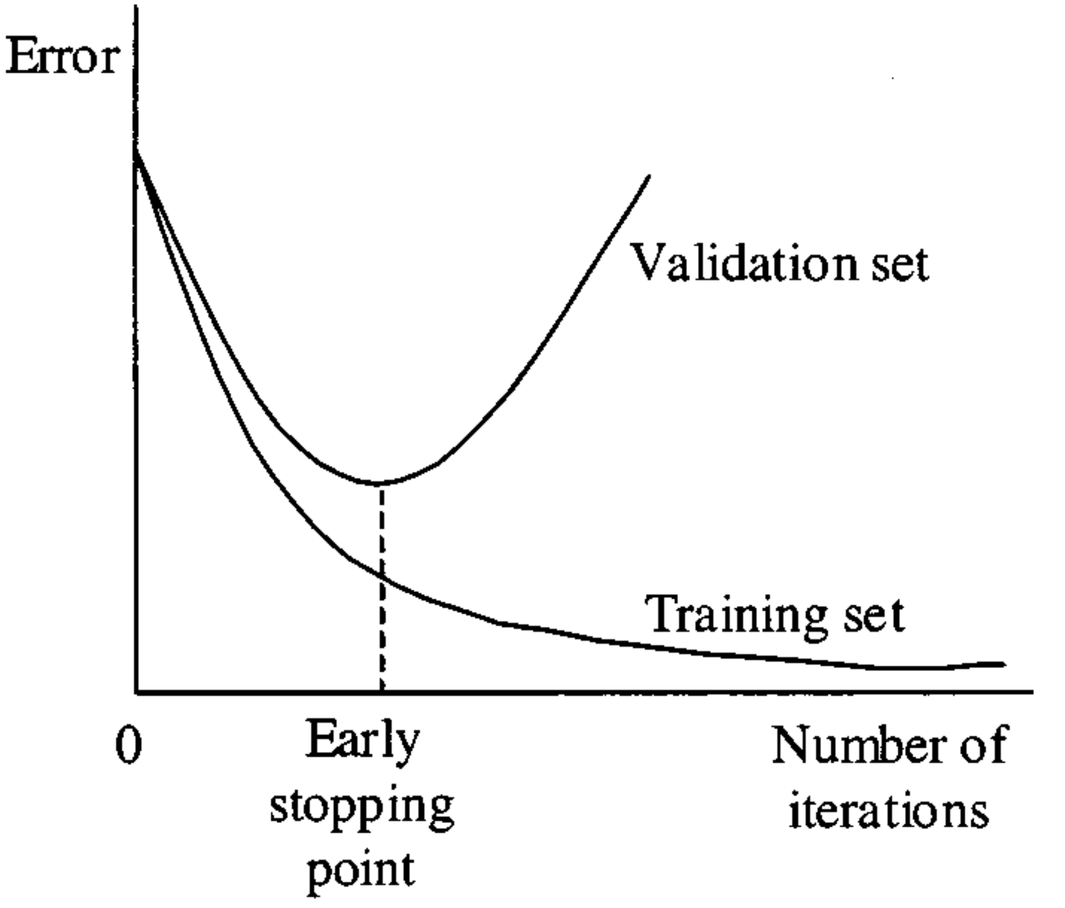
- Do use an appropriate test set
- Do use a validation set
- Do evaluate a model multiple times
- Do save some data to evaluate your final model instance
- Don’t use accuracy with imbalanced data sets

Do use an appropriate test set

- always use a test set to measure the generality of an ML model
- Appropriate test set : not overlap training set + wider population

Do use a validation set

- train multiple models : using knowledge gained about each model's performance to guide the configuration of the next.
- **not** to use the test set within this process
- Early stopping , prevent overfit



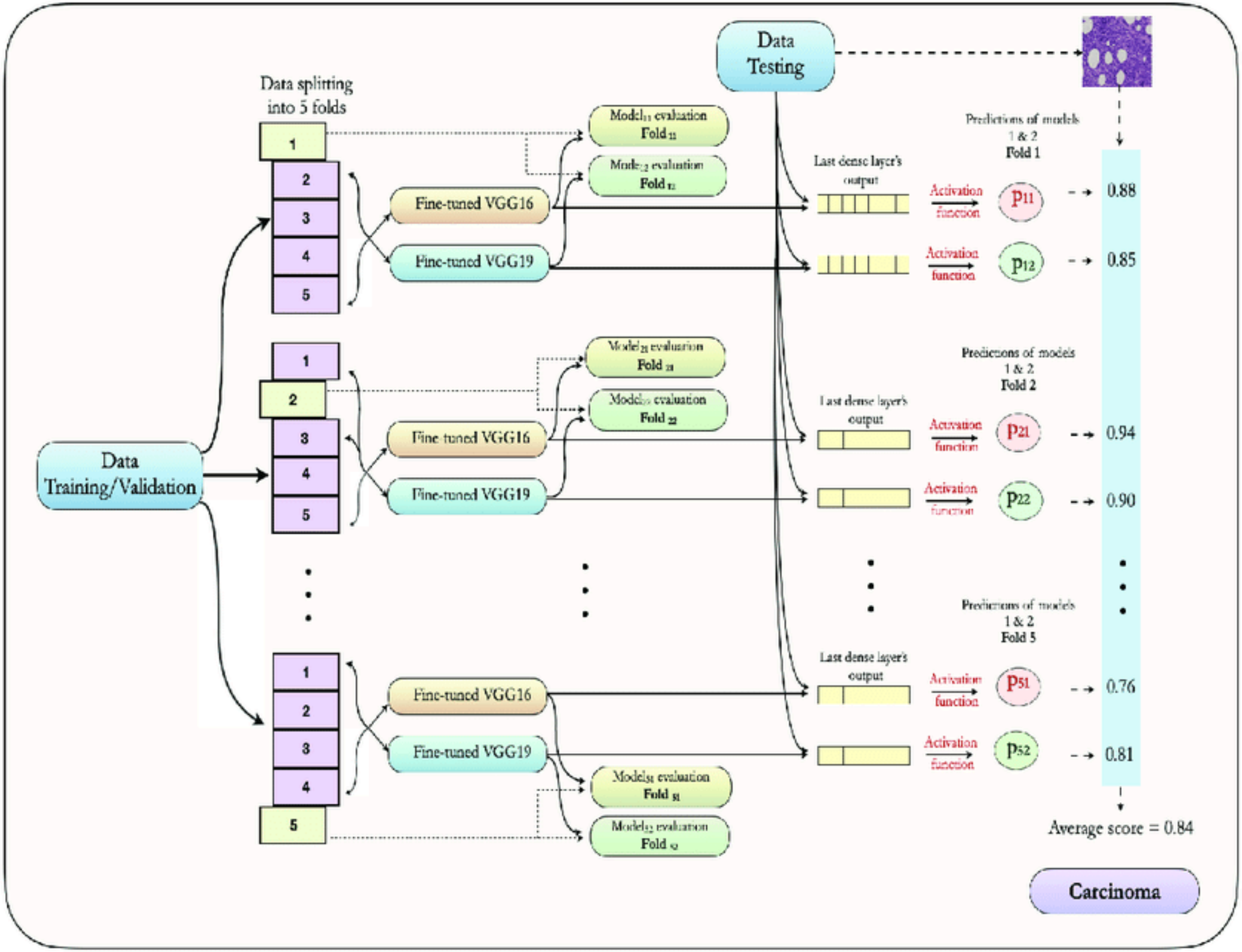
Do evaluate a model multiple times

- **Crossvalidation (CV)**
- Repeated CV: CV process is repeated multiple times with different partitionings of the data
- **Stratification** if imbalanced data

Do save some data to evaluate your final model instance

- 10-folds CV : 10 models
- Which one to report? Which one to use? The best one?

- Better to have untouched test set


Data Testing
Data splitting into 5 folds
Data Training/Validation
Fine-tuned VGG16
Fine-tuned VGG19
Model11 evaluation Fold 11
Model12 evaluation Fold 12
Model21 evaluation Fold 21
Model22 evaluation Fold 22
Model51 evaluation Fold 51
Model52 evaluation Fold 52
Last dense layer's output
Activation function
Predictions of models 1 & 2 Fold 1
Predictions of models 1 & 2 Fold 2
Predictions of models 1 & 2 Fold 5
P11
P12
P21
P22
P51
P52
0.88
0.85
0.94
0.90
0.76
0.81
Average score = 0.84
Carcinoma

Don’t use accuracy with imbalanced data sets

How to compare models fairly

- Don’t assume a bigger number means a better model
- Do use statistical tests when comparing models
- Do correct for multiple comparisons
- Don’t always believe results from community benchmarks
- Do consider combinations of models

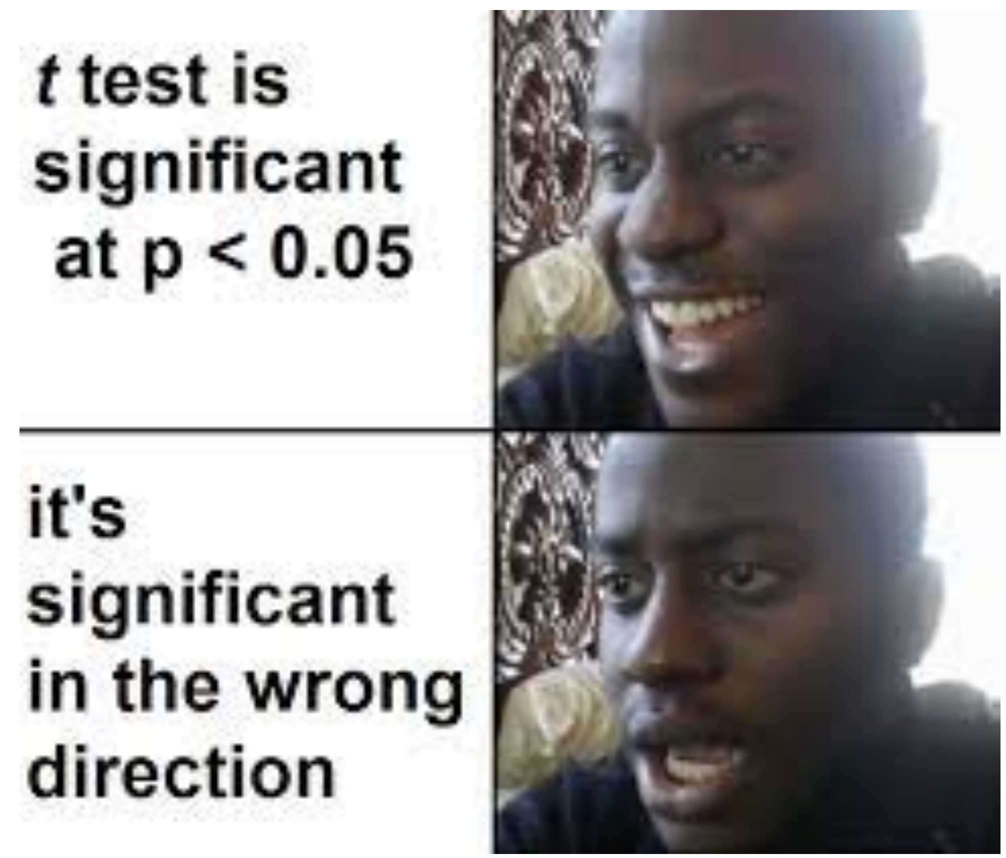
Don’t assume a bigger number means a better model

- 94% 🆚 95% ?
- Different partition of same dataset 🆚 different dataset
- Vanilla 🆚 optimised model
- ✅freshly implement the models
- ✅optimise each one to the same degree,
- ✅carry out multiple evaluations
- ✅use statistical test

Do use statistical tests when comparing models

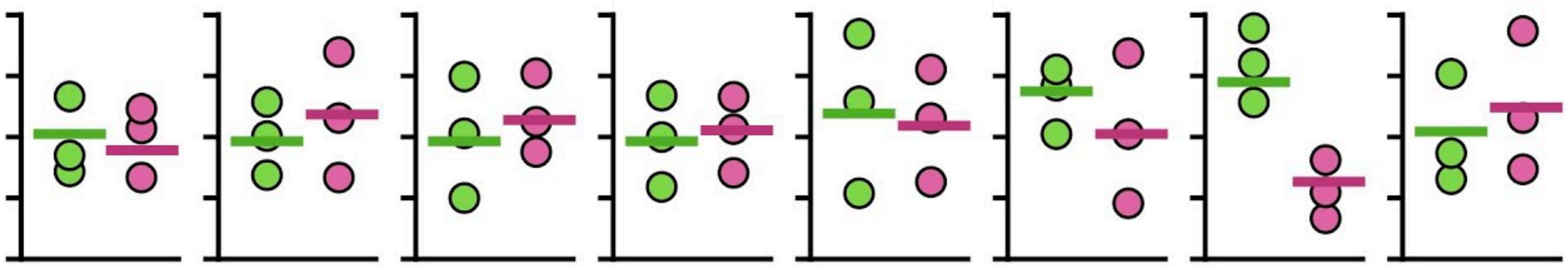
- Compare same type of model
  - **McNemar's test** : two classifiers—> comparing the classifiers' output labels for each sample in the test set

- Compare two different model
  - **Student's T test** :only normally distributed, which is often not the case.
  - **Mann-Whitney's U test** : does not assume that the distributions are normal.



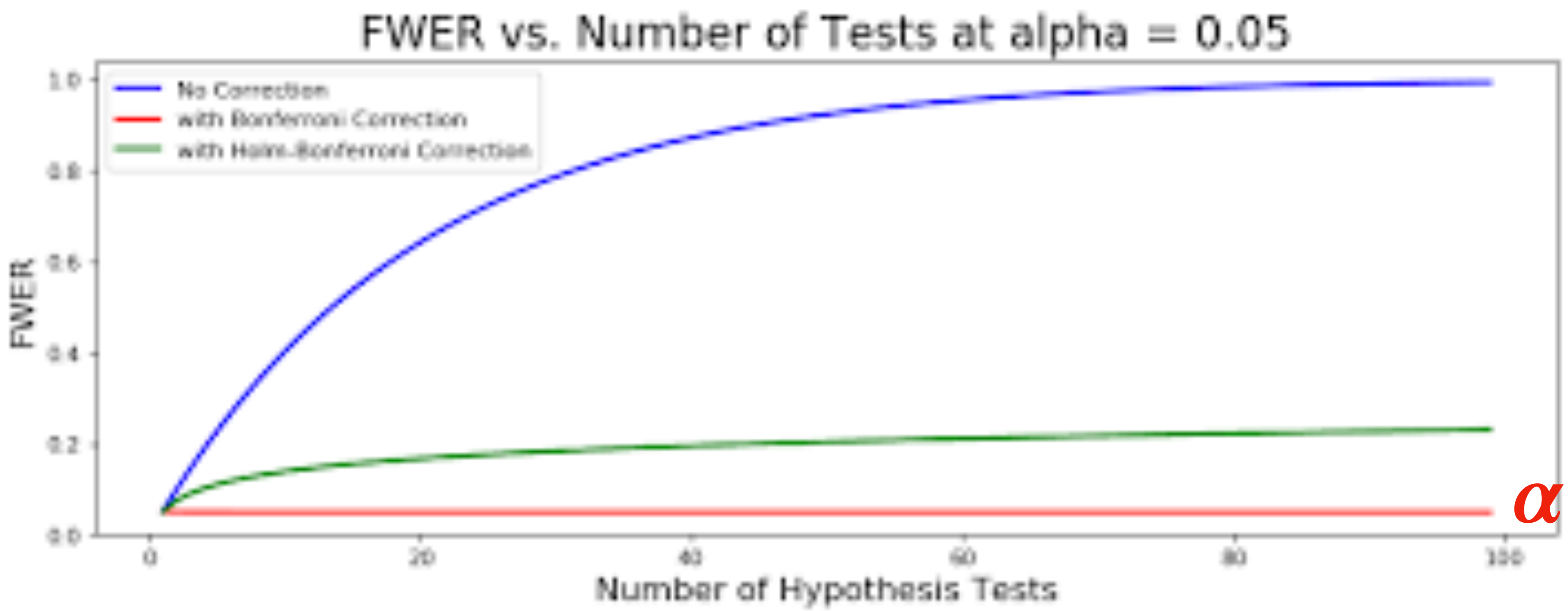
Do correct for multiple comparisons

- **Multiplicity effect** : compare multiple times of pairs : incremental chance of wrongly significant : False-positive
- **data dredging** or **p-hacking**



- **Bonferroni** correction,
  - lowers the significance threshold based on the number of tests that are being carried out

The probability of finding a significant result $= 1 - (1 - \overset{\alpha}{0.05})^{7}$
$= 0.30$



α / n

The original p value

$$\textit{Bonferroni-corrected p value} = \frac{\alpha}{n}$$

The number of tests performed

Don’t always believe results from community benchmarks

- Using benchmark dataset in certain problem
- Restricted (same) test set for everyone?
- comparing lots of models on the same test set: over-fit the test set
- **careful** : don’t assume that a small increase in performance is significant.

Diabetes Data Set

Download: Data Folder, Data Set Description

Abstract: This diabetes dataset is from AIM '94

Data Set Characteristics:	Multivariate, Time-Series	**Number of Instances:**	N/A	**Area:**	Life
Attribute Characteristics:	Categorical, Integer	**Number of Attributes:**	20	**Date Donated**	N/A
Associated Tasks:	N/A	**Missing Values?**	N/A	**Number of Web Hits:**	591748

Papers That Cite This Data Set[1]**:**

Rexa.info
Research • People • Connections

Prem Melville and Raymond J. Mooney. Diverse ensembles for active learning. ICML. 2004. [View Context].

Jeroen Eggermont and Joost N. Kok and Walter A. Kosters. Genetic Programming for data classification: partitioning the sea

Zhi-Hua Zhou and Yuan Jiang. NeC4.5: Neural Ensemble Based C4.5. IEEE Trans. Knowl. Data Eng, 16. 2004. [View Conte

Zhihua Zhang and James T. Kwok and Dit-Yan Yeung. Parametric Distance Metric Learning with Label Information. IJCAI. 2

Michael L. Raymer and Travis E. Doom and Leslie A. Kuhn and William F. Punch. Knowledge discovery in medical and biolo
Transactions on Systems, Man, and Cybernetics, Part B, 33. 2003. [View Context].

Eibe Frank and Mark Hall. Visualizing Class Probability Estimators. PKDD. 2003. [View Context].

Krzysztof Krawiec. Genetic Programming-based Construction of Features for Machine Learning and Knowledge Discovery T
[View Context].

Ilya Blayvas and Ron Kimmel. Multiresolution Approximation for Classification. CS Dept. Technion. 2002. [View Context].

Peter Sykacek and Stephen J. Roberts. Adaptive Classification by Variational Kalman Filtering. NIPS. 2002. [View Context].

Kristin P. Bennett and Ayhan Demiriz and Richard Maclin. Exploiting unlabeled data in ensemble methods. KDD. 2002. [View

Marina Skurichina and Ludmila Kuncheva and Robert P W Duin. Bagging and Boosting for the Nearest Mean Classifier: Effe
2002. [View Context].

Jochen Garcke and Michael Griebel and Michael Thess. Data Mining with Sparse Grids. Computing, 67. 2001. [View Conte

Peter L. Hammer and Alexander Kogan and Bruno Simeone and Sandor Szedm'ak. R u t c o r Research R e p o r t. Rutgers

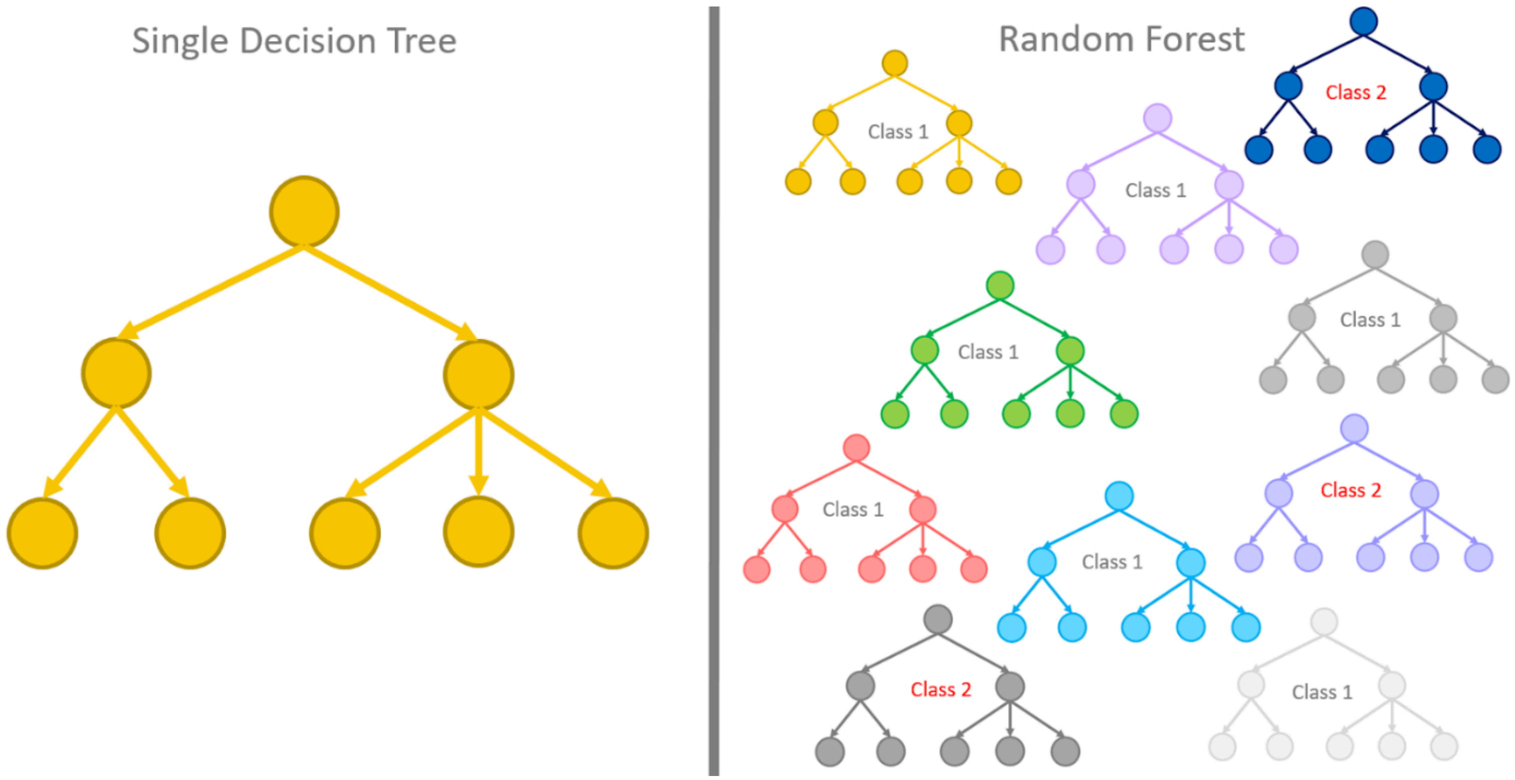
Robert Burbidge and Matthew Trotter and Bernard F. Buxton and Sean B. Holden. STAR - Sparsity through Automated Reje

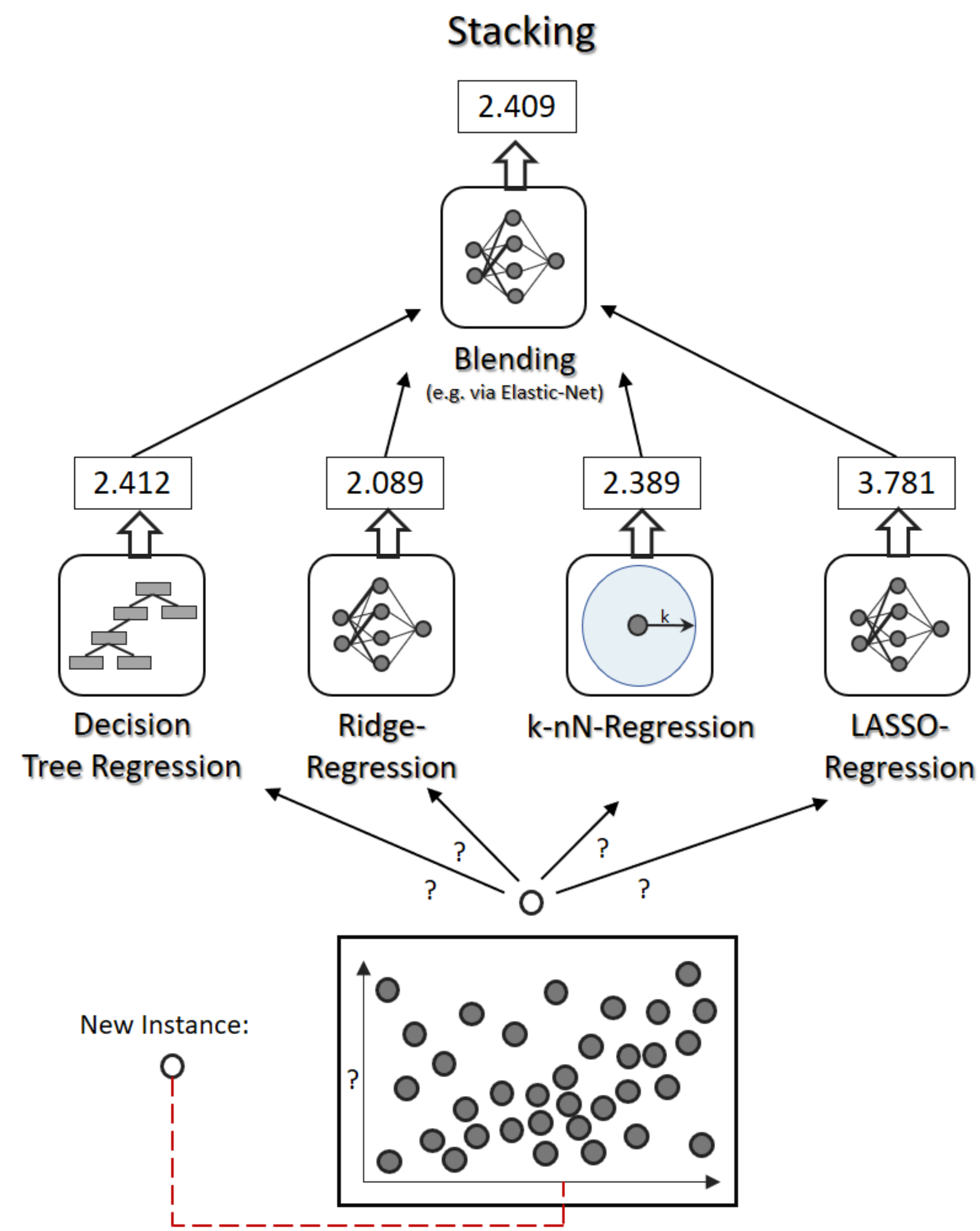
Mark A. Hall. Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. ICML. 2000. [View Co

Endre Boros and Peter Hammer and Toshihide Ibaraki and Alexander Kogan and Eddy Mayoraz and Ilya B. Muchnik. An Im

Source:

Do consider combinations of models : Ensembles



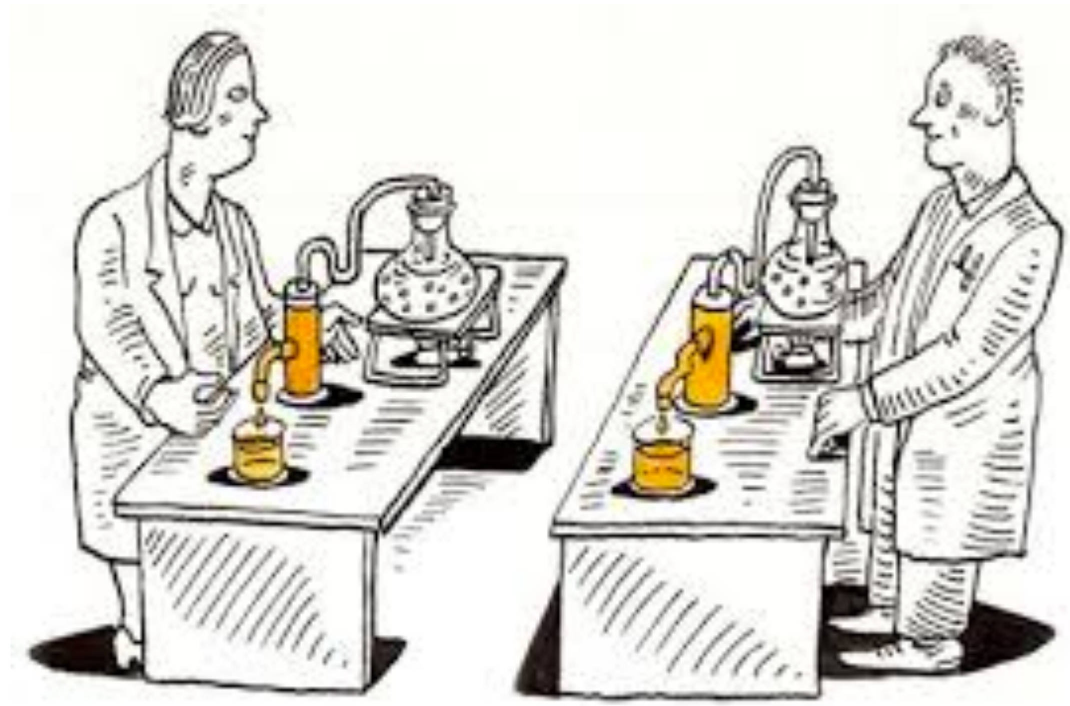


How to report your results

- Do be transparent
- Do report performance in multiple ways
- Don’t generalise beyond the data
- Do be careful when reporting statistical significance
- Do look at your models

Do be transparent

- Share model , script
- encourages to be more careful, document experiments well, and write clean code
- **reproducibility** : prominence in the ML community



Yoshua Bengio
Yesterday at 09:06 ·

Terhoch Solange
30 August at 16:40 ·

There was a farmer growing an excellent quality corn. Every year he won the award for best grown corn. One year, a journalist interviewed him and learned something interesting about how he cultivated it.
The journalist found that the farmer shared his seed corn with his neighbors. ″ How can you afford to share your best seed corn with your neighbors as they produce corn competing with yours every year?" says the journalist.
″ Why sir ", said the farmer, ″ You didn't know that? Wind picks up pollen from corn ripening and swirls it from field to field. If my neighbors grow lower quality corn, cross pollination will gradually deteriorate the quality of my corn. If I want to grow good corn, I have to help my neighbors grow good corn ".
Same goes for our lives... Those who want to live usefully, must contribute to enriching the lives of others, because the value of a life is measured by the lives it touches.
And those who choose to be happy must help others find happiness, because everyone's well-being is linked to the well-being of all...

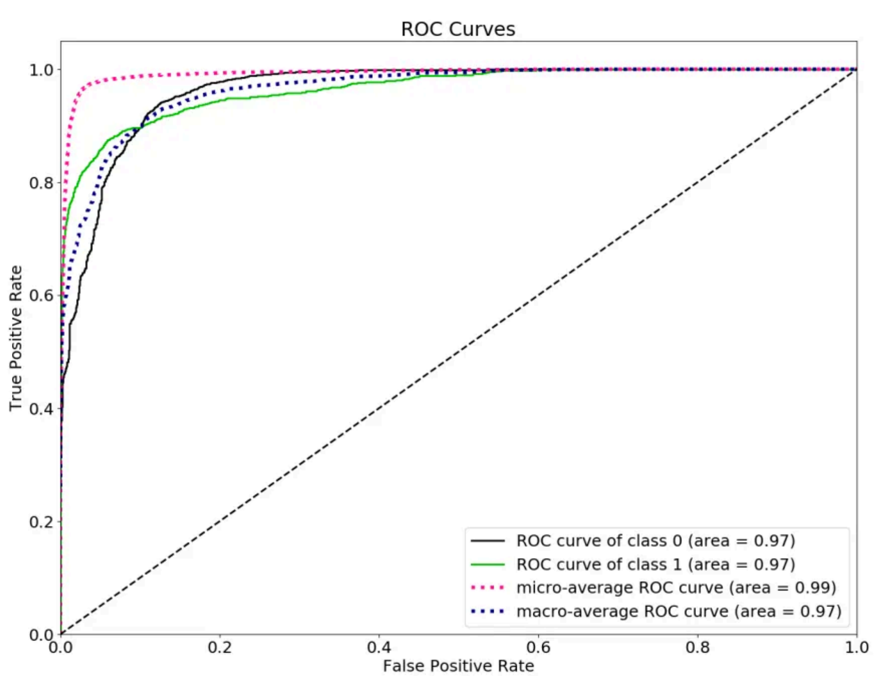
Do report performance in multiple ways

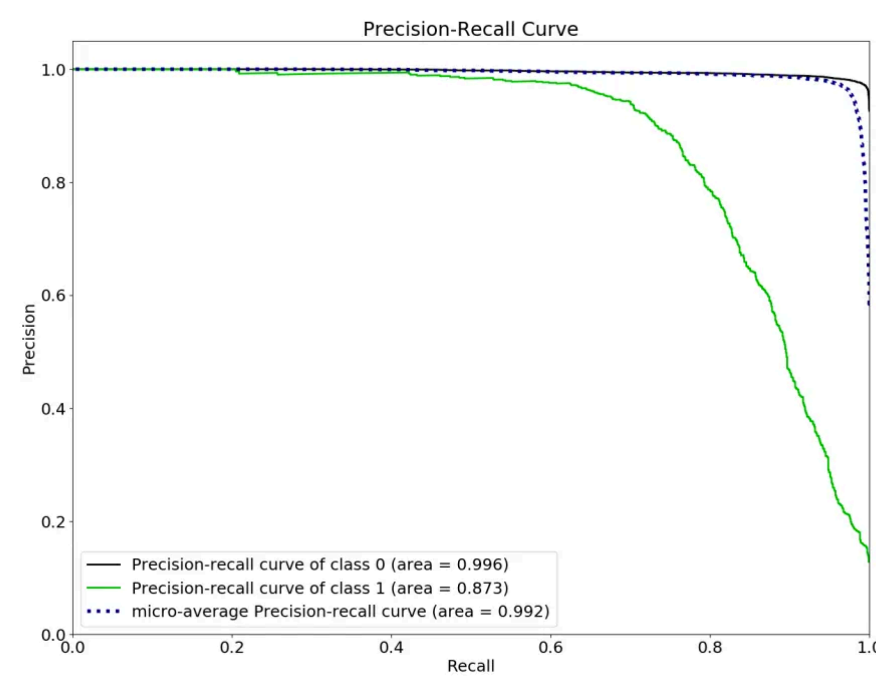
- Evaluate in multiple datasets
- Multiple metrics

$$ACC = \frac{tp + tn}{tp + fp + tn + fn}$$

$$F_{beta} = (1+\beta^2)\frac{precision * recall}{\beta^2 * precision + recall}$$

		Disease ⊕	Disease ⊖	Predictive Value	
Test	⊕	**A** True Positive (TP)	**B** False Positive (FP)	Positive Predictive Value (PPV) $\frac{TP}{TP+FP} = \frac{A}{A+B}$	Total Positive Results (A + B)
	⊖	**C** False Negative (FN)	**D** True Negative (TN)	Negative Predictive Value (NPV) $\frac{TN}{FN+TN} = \frac{D}{C+D}$	Total Negative Results (C + D)
Sensitivity & Specificity		Sensitivity $\frac{TP}{TP+FN} = \frac{A}{A+C}$	Specificity $\frac{TN}{FP+TN} = \frac{D}{B+D}$		
		All diseased patients (A + C)	All non-diseased patients (B + D)		





Don’t generalise beyond the data

- Sampling bias : Dataset not reflect the real world

Google medical researchers humbled when AI screening tool falls short in real-life testing

Devin Coldewey @techcrunch / 4:03 AM GMT+7 • April 28, 2020 Comment

- Quality of dataset : image from studio vs r-ma ’s mobile

Do be careful when reporting statistical significance

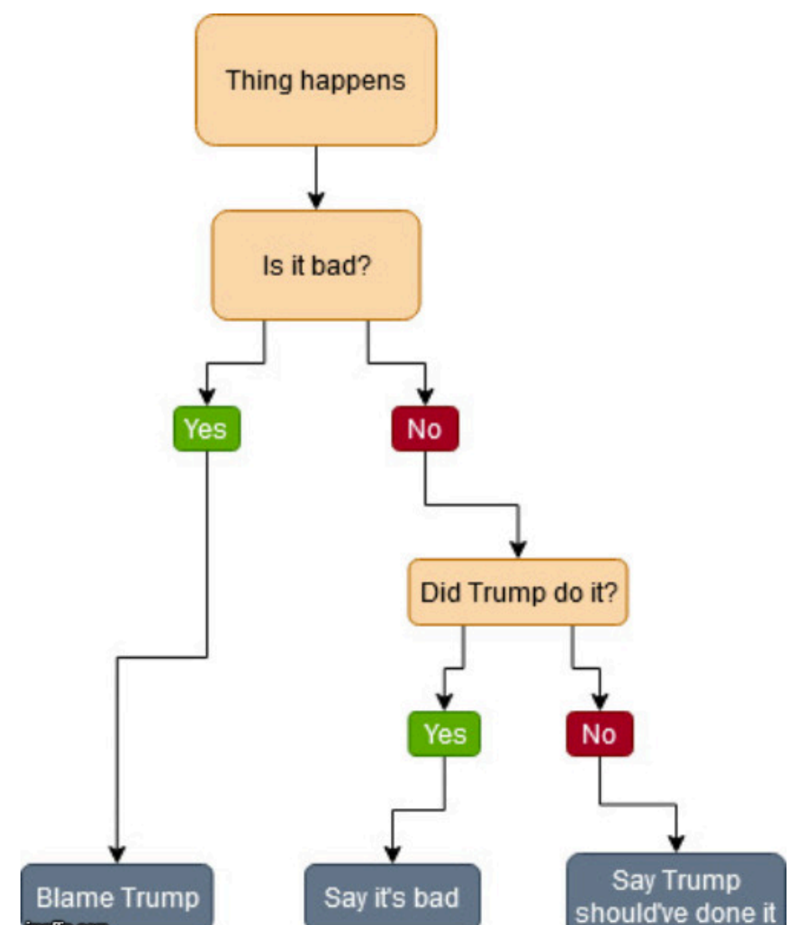
- 95% CI threshold : **1:20 false positive**
- Large samples : sig. differences, even when the actual difference in performance is miniscule
- statisticians are increasingly arguing : better not to use thresholds, just report p-values and leave it to the reader to interpret
- **Effect size** : Cohen's d statistic , Kolmogorov-Smirnov

$$d = \frac{M_E - M_C}{\text{Sample } SD \text{ pooled}} \boxed{\times \left(\frac{N-3}{N-2.25}\right) \times \sqrt{\frac{N-2}{N}}}$$

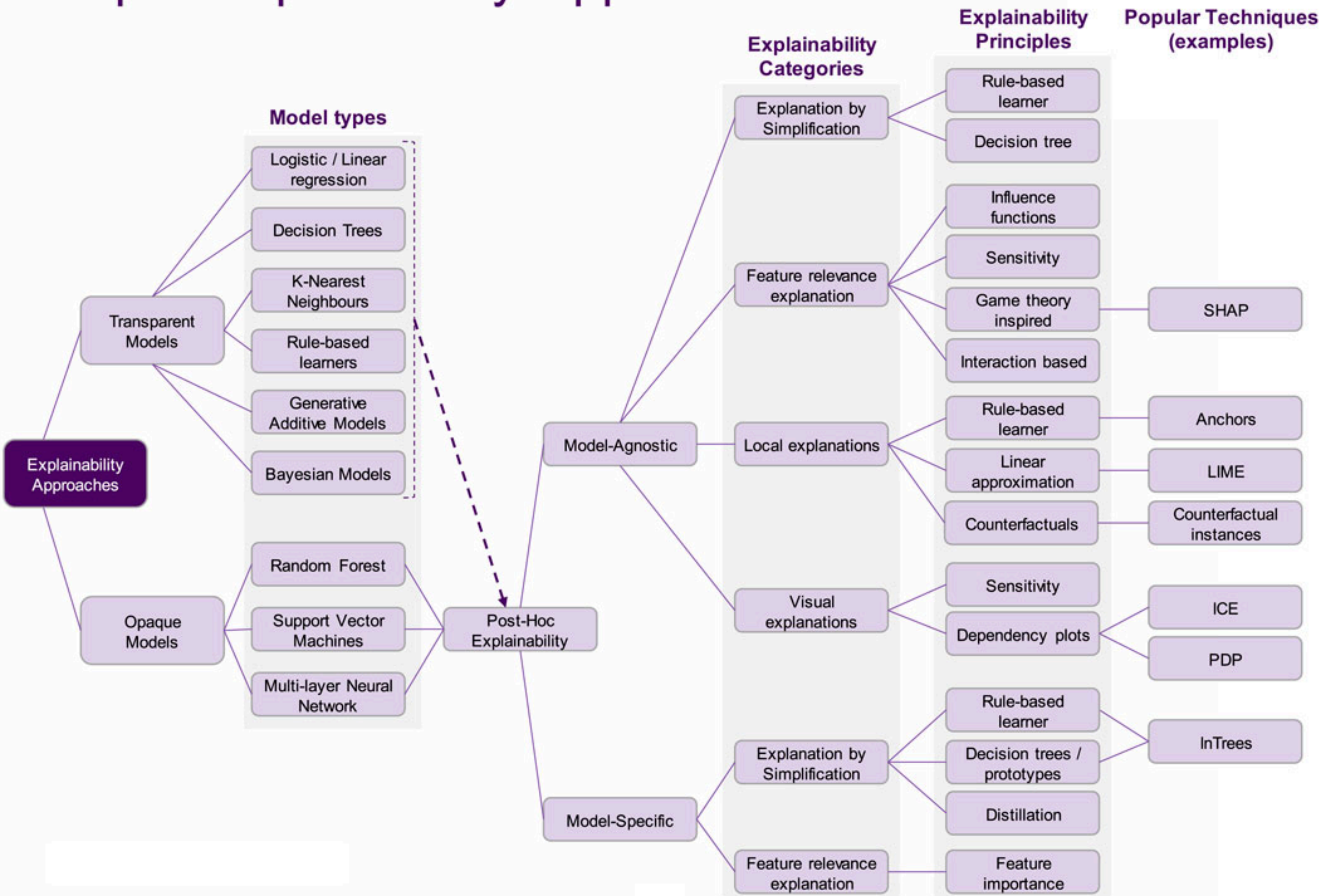
correction factor for
small samples <50

Do look at your models

- **aim of research** : not to get a slightly higher accuracy than everyone else.
- **Generate knowledge** / understanding + share with the research community
- Look inside models + try to understand
  - Decision trees : provide visualisations
  - XAI Explainable AI techniques for complex model



Map of Explainability Approaches



Q & A