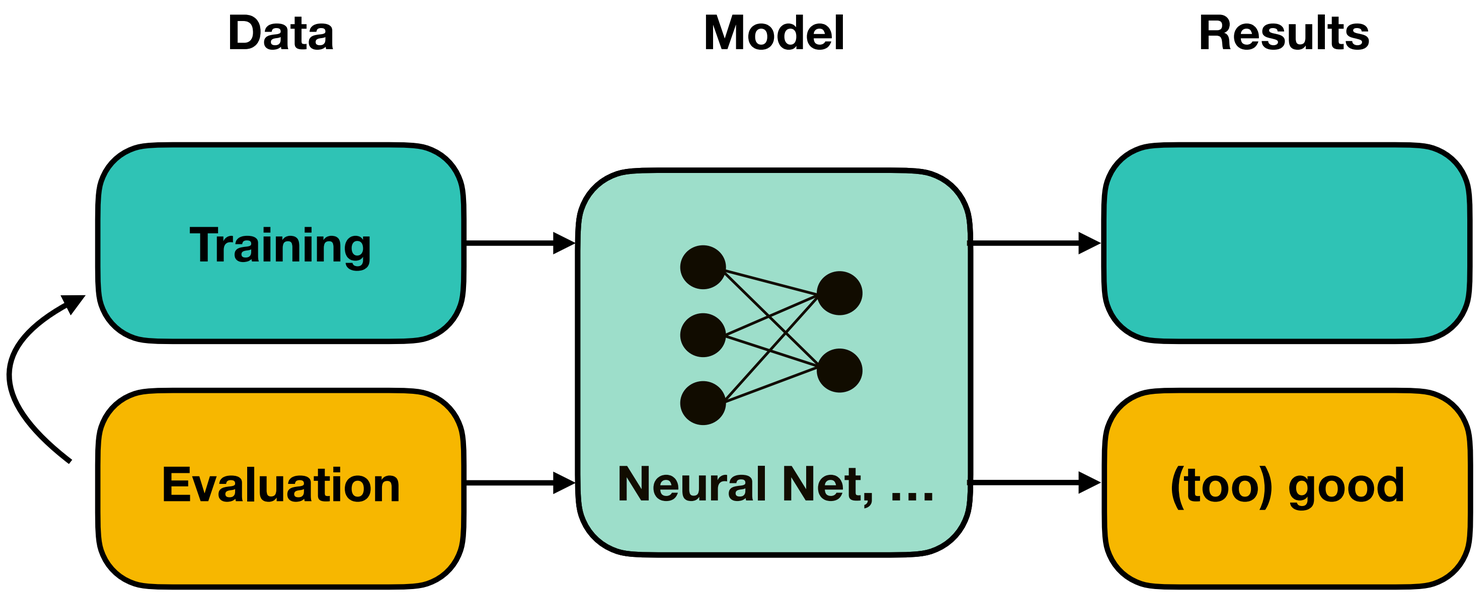
Data Leakage in ML-based Projects

Kapoor & Narayanan (2022): “Leakage and Reproducibility Crisis in ML-based Science”

Dorothea Sommer | GöHPCoffee | 31.08.2022

Data Leakage

Using evaluation information during training



- Kapoor & Narayanan (2022) meta-review: 329 papers identified **across many domains** (medicine, social science, …)
- Leads to **overoptimistic estimate** of the employed model

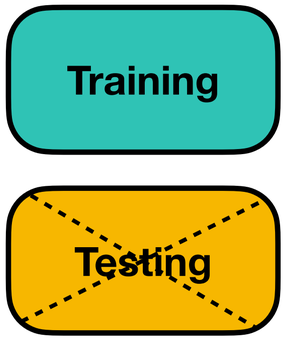
Empirical Results

Field	Paper	Number of papers reviewed	Number of papers with pitfalls	**[L1.1] No test set**	**[L1.2] Pre-proc. on train-test**	**[L1.3] Feature sel. on train-test**	**[L1.4] Duplicates**	**[L2] Illegitimate features**	**[L3.1] Temporal leakage**	**[L3.2] Non-ind. b/w train-test**	**[L3.3] Sampling bias**	Comput. reproducibility issues	Data quality issues	Metric choice issues	Standard dataset used?
Medicine	Bouwmeester et al. (2012)	71	27	○									○		
Neuroimaging	Whelan & Garavan (2014)	–	14	○		○									
Autism Diagnostics	Bone et al. (2015)	–	3				○			○			○	○	○
Bioinformatics	Blagus & Lusa (2015)	–	6		○										
Nutrition Research	Ivanescu et al. (2016)	–	4	○									○	○	
Software Eng.	Tu et al. (2018)	58	11					○			○		○		○
Toxicology	Alves et al. (2019)	–	1				○				○		○		
Satellite Imaging	Nalepa et al. (2019)	17	17						○				○		○
Tractography	Poulin et al. (2019)	4	2	○							○		○	○	○
Clinical Epidem.	Christodoulou et al. (2019)	71	48			○							○		
Brain-computer Int.	Nakanishi et al. (2020)	–	1	○											○
Histopathology	Oner et al. (2020)	–	1						○						
Neuropsychiatry	Poldrack et al. (2020)	100	53	○	○								○	○	
Medicine	Vandewiele et al. (2021)	24	21			○			○	○	○		○		○
Radiology	Roberts et al. (2021)	62	62	○			○			○	○				○
IT Operations	Lyu et al. (2021)	9	3					○							○
Medicine	Filho et al. (2021)	–	1				○								
Neuropsychiatry	Shim et al. (2021)	–	1			○					○				
Genomics	Barnett et al. (2022)	41	23			○							○		
Computer Security	Arp et al. (2022)	30	30	○	○	○		○		○	○		○	○	

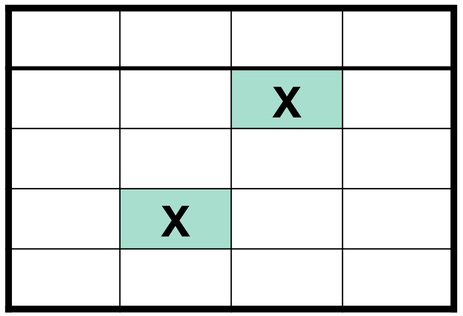
L1 No Clean Separation

Between the training and the test set

L1.1 No test set

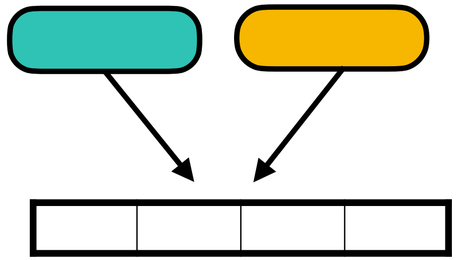


L1.2 Pre-processing on training and test set

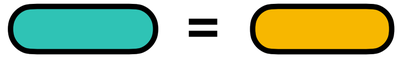


Think: Imputation (replacing missing values with what)?

L1.3 Feature Selection on both



L1.4 Duplicates



Think: How can we ensure no duplicates?

My recommendation: Use **fdupe -r .** for checking the content (not filenames!).

L2 Model uses illegimate features

Examples

- Feature as **proxy for the outcome variable**
  🔎💉 **feature**: use of anti-hypertensive drug, **prediction:** hypertension

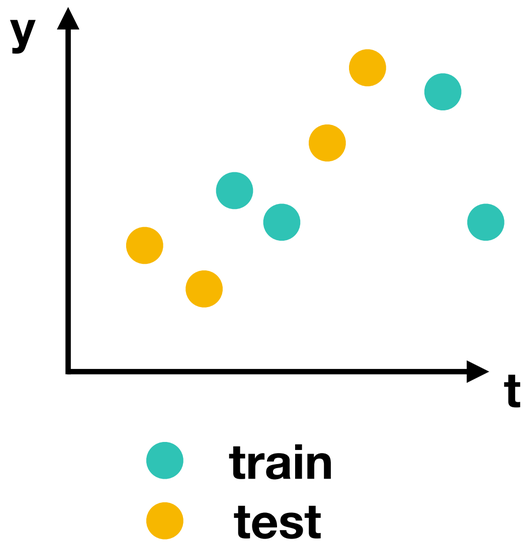
- 👾 Sometimes it can be **hidden** (own experience from replicating paper)
  natural language processing:
  - cluster words of a tweet corpus into descriptive 200 words
  - **features:** linguistic features (e.g, emoji usage, whether each word from 200 is present)
  - **prediction**: classify socioeconomic status of users (ground truth: job in profile)

Why is the feature in the model legitimate? Requires domain knowledge!

L3 Test set not properly drawn

From the distribution of scientific interest

L3.1 Temporal leakage



Think: Why is this a problem?

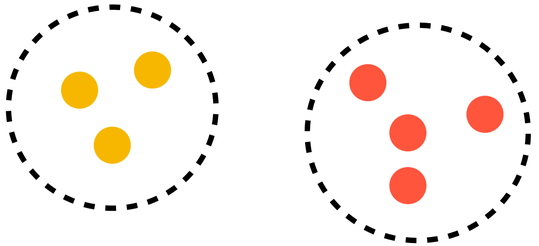
L3.2 Nonindependence between test and train samples

block cross validation (Roberts, 2017)

Dependence structure	Parametric solution	Blocking	Blocking illustration
Spatial	Spatial models (e.g.CAR, INLA, GWR)	Spatial	
Temporal	Time-series models (e.g.ARIMA)	Temporal	
Grouping	Mixed effect models (e.g. GLMM)	Group	
Hierarchical / Phylogenetic	Phylogenetic models (e.g. PGLS)	Hierarchical	

Figure 1. Examples of dependence structures, parametric solutions to parameter estimation, and the associated blocking approaches for cross-validation to increase reliability of prediction error estimates.

L3.3 Sampling bias in test distribution



spatial bias
sampling from one location, making claims about another

selection bias
ignoring borderline cases in autism diagnostic, so overoptimistic results

Solution Ideas: Model Sheet 💡

Answer questions to prevent data leakage

L1 Clean train test separation.

Argue why test set does not interact with training set. | Duplicates.

L2 Check legitimacy for each feature.

Argue why *each* feature is legitimate. | Makes you think why you assume relation.

L3 Test set is drawn from distribution of scientific interest.

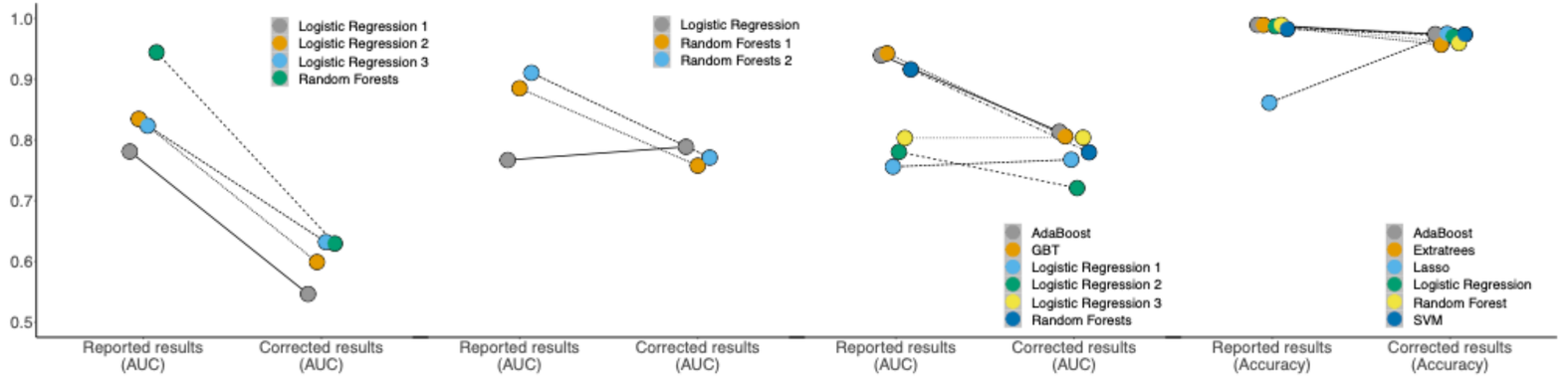
Is the distribution of scientific interest the same on which the model is tested on?

Take-away

- Use **check-list** to ensure that your data-processing goes right
  - **Model info sheet** for detecting and preventing data leakage: https://reproducible.cs.princeton.edu/model-info-sheet-template.docx (Kapoor & Narayanan, 2022)
  - **Model card** for clarifying details of training and usage contexts: https://arxiv.org/pdf/1810.03993.pdf (Mitchell, 2019)
- Thoughtfully **inspect your data** (Andrej Karpathy: spend *hours* inspecting).

Empirical Results

Corrected ML results on civil war prediction



Paper	Muchlinski et al.	Colaresi and Mahmood	Wang	Kaufman et al.
Claim	Random Forests model drastically outperforms Logistic regression models	Random Forests models drastically outperform Logistic regression model	Adaboost and Gradient Boosted Trees (GBT) drastically outperform other models	Adaboost outperforms other models
Error	**[L1.2] Pre-proc. on train-test** (Incorrect imputation)	**[L1.2] Pre-proc. on train-test** (Incorrect reuse of an imputed dataset)	**[L1.2] Pre-proc. on train-test**. (Incorrect reuse of an imputed dataset) **[L3.1] Temporal leakage** (*k*-fold cross validation with temporal data)	**[L2] Illegitimate features** (Data leakage due to proxy variables) **[L3.1] Temporal leakage** (*k*-fold cross validation with temporal data)
Impact	Random Forests perform no better than Logistic Regression	Random Forests perform no better than Logistic Regression	Difference in AUC between Adaboost and Logistic Regression drops from 0.14 to 0.01	Adaboost no longer outperforms Logistic Regression. None of the models outperform a baseline model that predicts the outcome of the previous year
Discussion	Impact of the incorrect imputation is severe since 95% of the out-of-sample dataset is missing and is filled in using the incorrect imputation method	Re-use the dataset provided by Muchlinski et al., which uses an incorrect imputation method	Re-use the dataset provided by Muchlinski et al., which uses an incorrect imputation method	Use several proxy variables for the outcome as predictors (e.g., *colwars, cowwars, sdwars*, all proxies for civil war), leading to near perfect accuracy

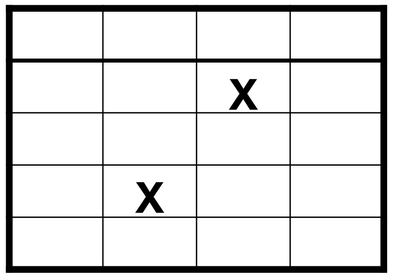
Other Issues

That are not data leakage

Computational Reproducibility

Available code?
Available data?

Data Quality



How are missing values addressed?

Metric Choice

Accuracy?

Does performance metric capture scientific problem of interest?

Use of standard data sets

1. ?	1. ?
2. ?	2. ?
3. ?	3. ?

No standard modeling and evaluation procedures.