Slides, codes and other resources

- Connect with us at biodatascienceandeducation@e.ntu.edu.sg or wilsongoh@ntu.edu.sg
- Check out our website at https://gohwils.github.io/biodatascience/
- Codes for generating the outputs is conducted using R

https://qrgo.page.link/GeJVk

Two missed issues in data analysis

Wilson Wen Bin Goh

iSLS9 2021

Singapore Phenome Center (LKC)

Metabonomics Workshop 2021

Masters of Biomedical Data Science (second intake)

- **First** biomedical data science graduate programme in Asia-Pacific
- **Highly selective** (only 12 places for AI track. 12 for Bioinformatics track.)
- Course run with **McKinsey and Co.**
- **DeepTech immersion scheme** -> Get paid and work in Startups or large companies
- Unique Biomedical Science immersion scheme -> learn directly from **NTU** professors
- Train with practicing data scientists in industry and academia! -> from **UpLevel**, **Quantum Black** and Bioinformatics Institute Singapore, **ASTAR**
- **Contact**: msc_biodatascience@ntu.edu.sg
- >300 applicants this round (including candidates from Columbia, Berkeley, Imperial College, Zhejiang, Fudan, Shanghai JiaoTong and Peking U.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

McKinsey & Company

Agency for Science, Technology and Research SINGAPORE

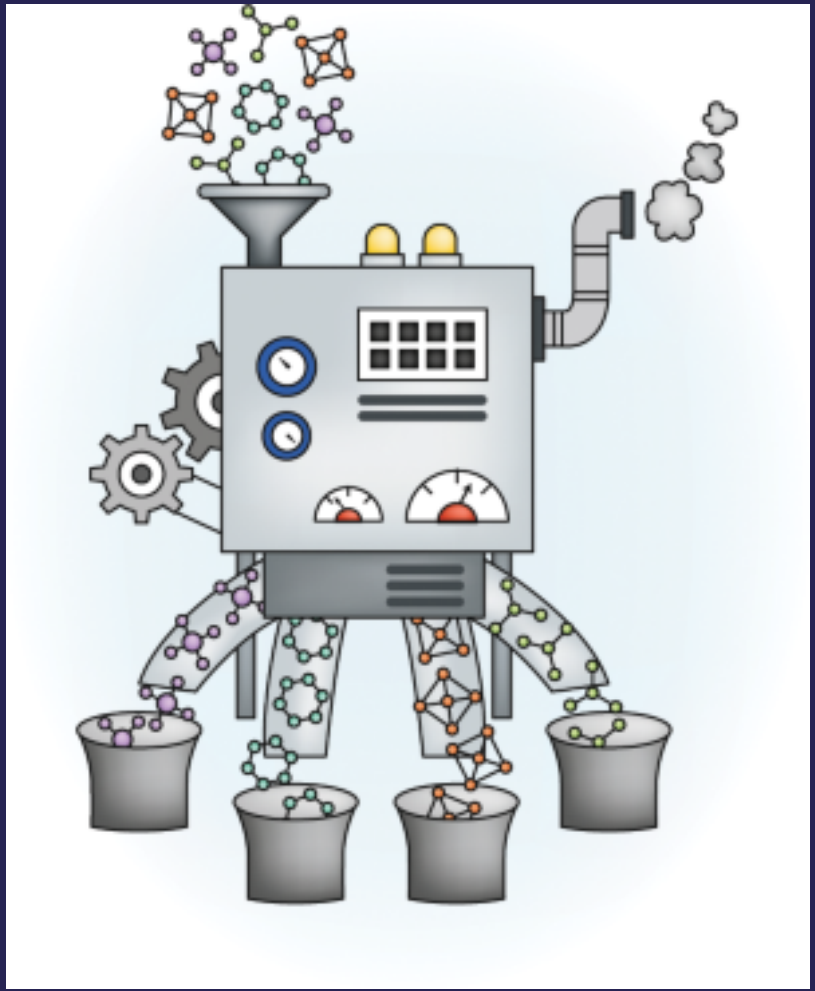
SGINNOVATE

http://sbs.ntu.edu.sg/Programmes/GraduateByCourseWork

Sophisticated tools are needed to process metabolomics data.

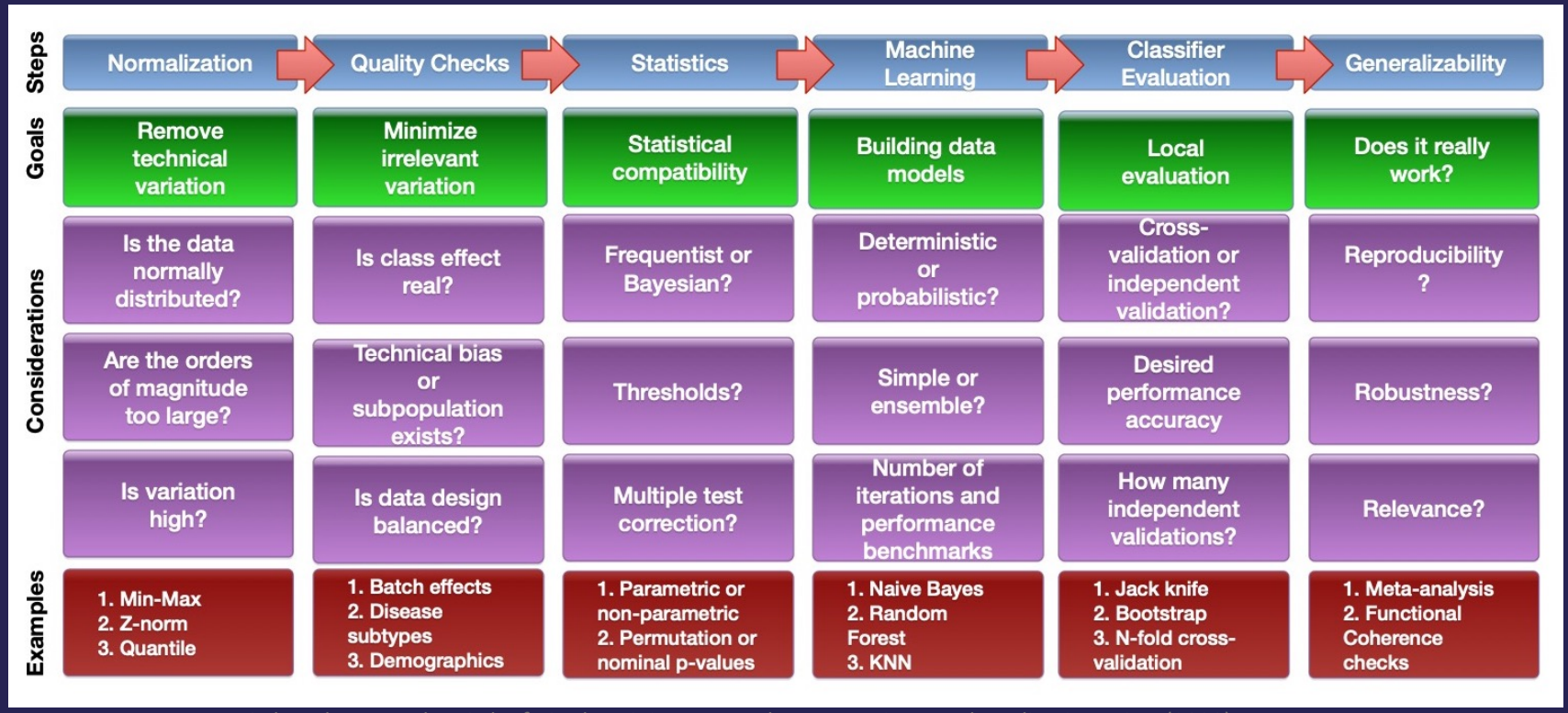
Credit: Kimberly Caesar/Springer Nature

https://www.nature.com/articles/s41592-019-0710-6



Data analysis is complex

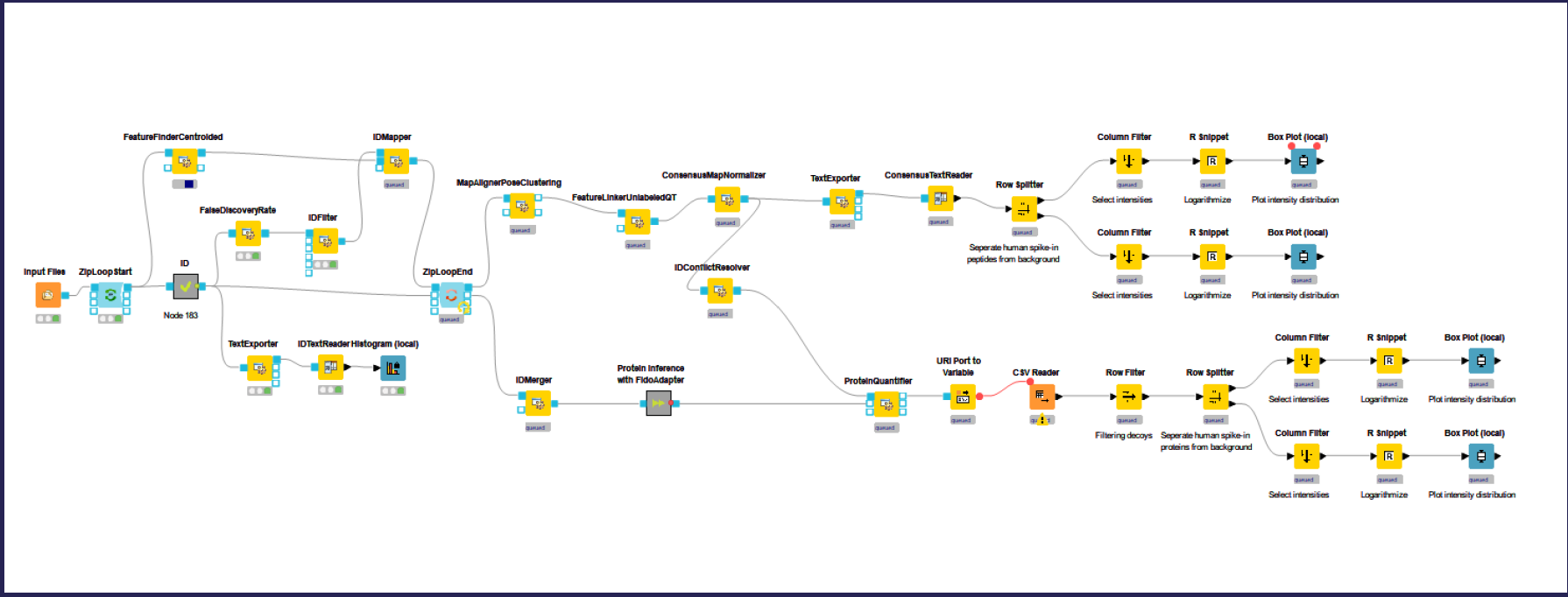
- There are dependencies and assumptions at every step
- Errors generated upstream can be propagated and obfuscated downstream



Goh and Wong, *The Birth of Bio-data Science: Trends, Expectations, and Applications. GPB* (2020)

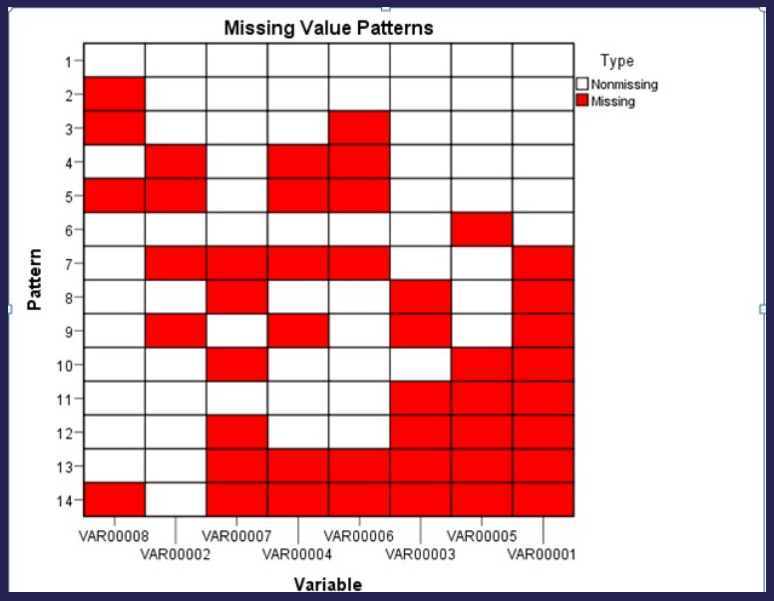
Data analysis is complex

- An MS proteomics pipeline built using OpenMS and KNIME by my PhD student Weijia
- Each node is a data processing step, requiring different parameters and settings
- How do you know if your settings are suitable?



2 missed or unknown themes in data analysis

SENSIBLE PREPROCESSING



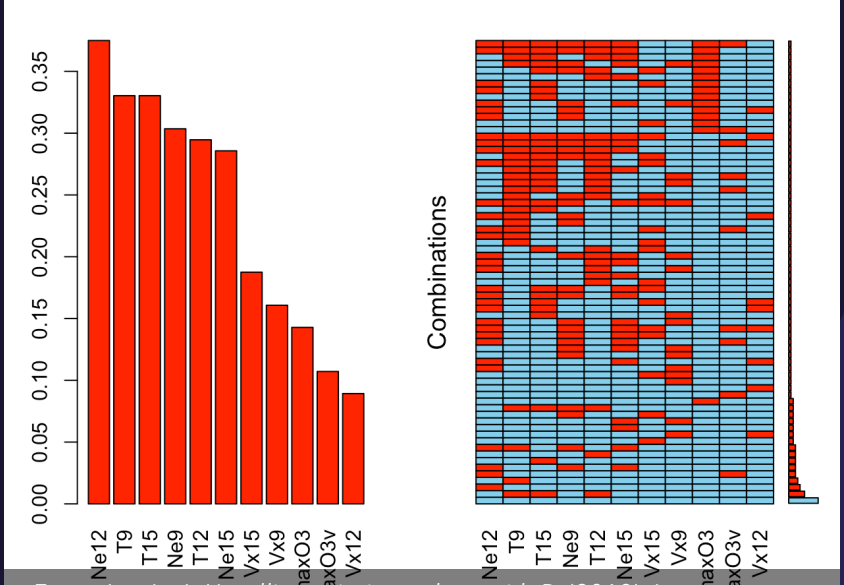
DOPPELGANGER EFFECT

Missing Value Imputation

In the presence of non-negligible co-variates

Joint work with Priscila Yun Qian Sun

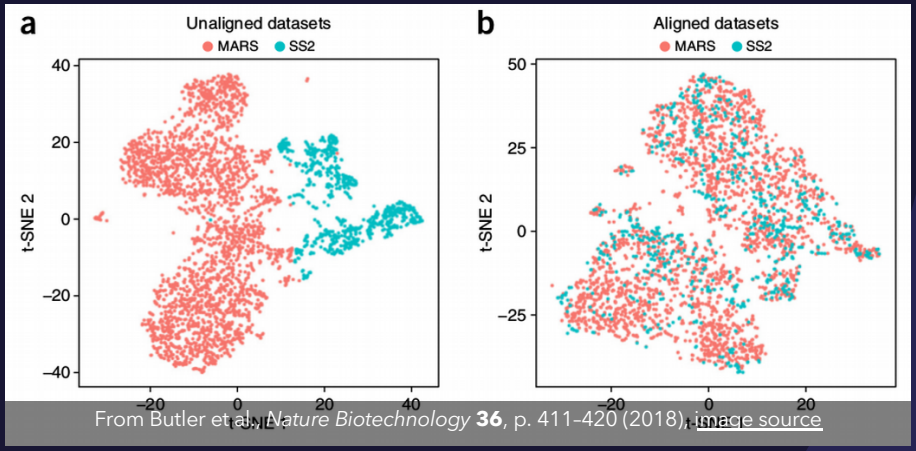
What is Missing Value Imputation (MVI)?



From Jossie J, *Handling missing values with R.* (2018), image source

- MVI is the task of estimating and inserting the value of a missing data value
- Missing values abound in real world data
- Some simple MVI approaches include 0 imputation, or variable-wise global mean imputation
- We will focus on variable-wise global mean imputation

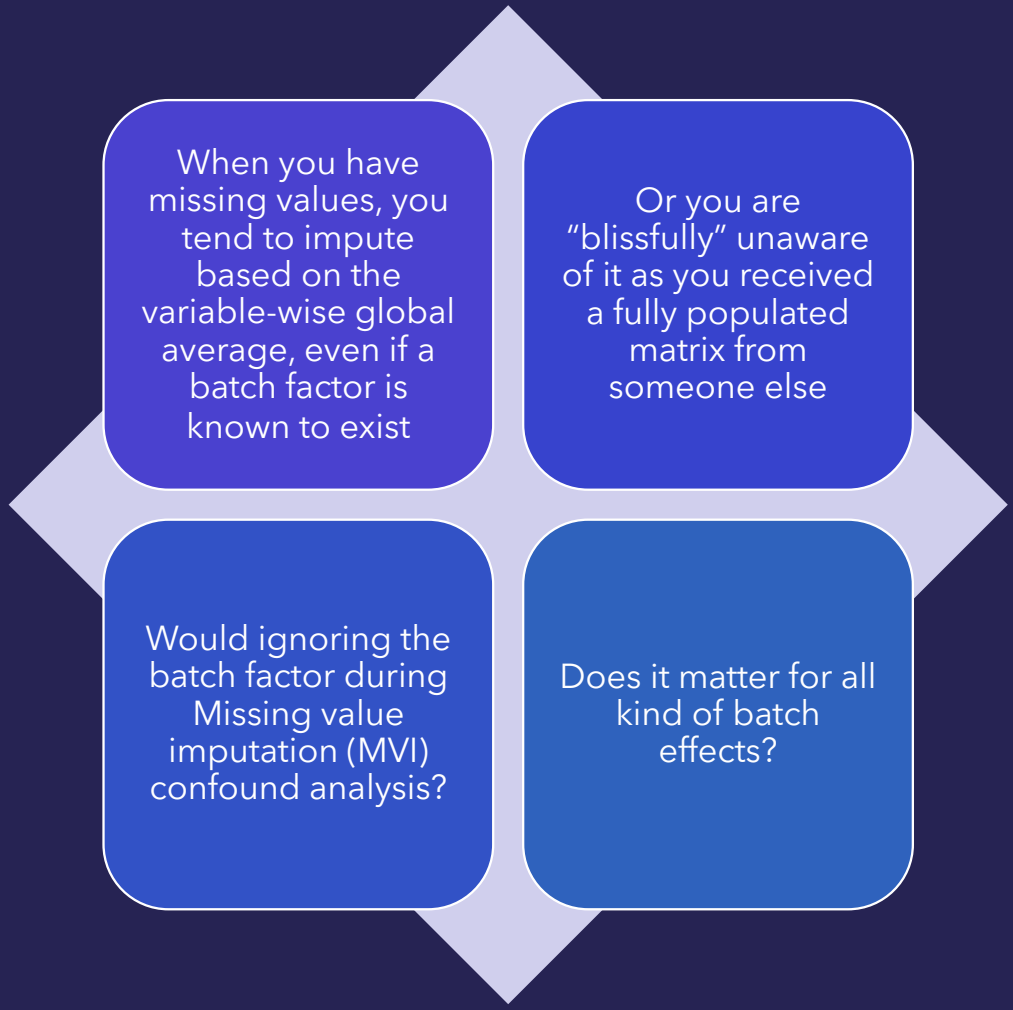
What are batch effects?



From Butler et al., *Nature Biotechnology* **36**, p. 411–420 (2018), image source

- Batch effects are technical sources of variation
- Can be due to machine, reagent, experimenter
- Can generate false positives and false negatives
- Exact nature is likely heterogeneous and complex
- Can be estimated and removed via batch effect correction algorithms (BECA)

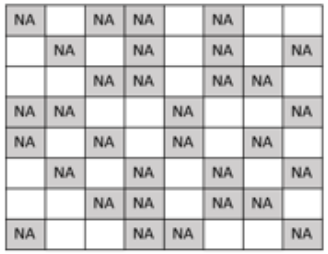
What happens when MVI is done but batch effects exist?



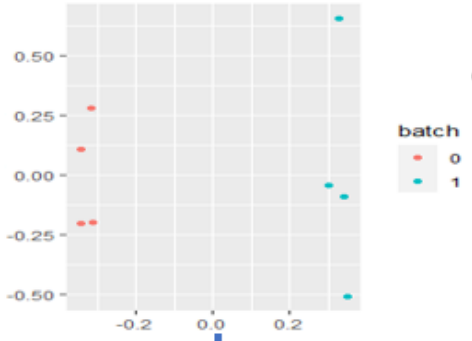
Problems in high-dimensional biological data analyses

Missing Value (MV)

- Information points being present in some samples but not others



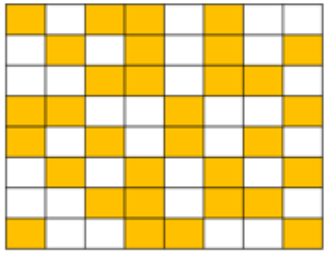
Batch Effect (BE)



- Technical sources of biases that may confound the true signal-of-interest

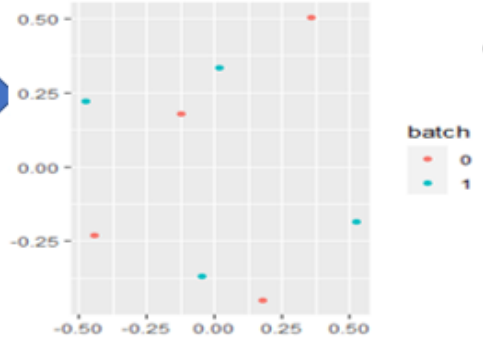
Missing Value Imputation (MVI)

- Early Pre-processing
- Usually ignore batch covariate



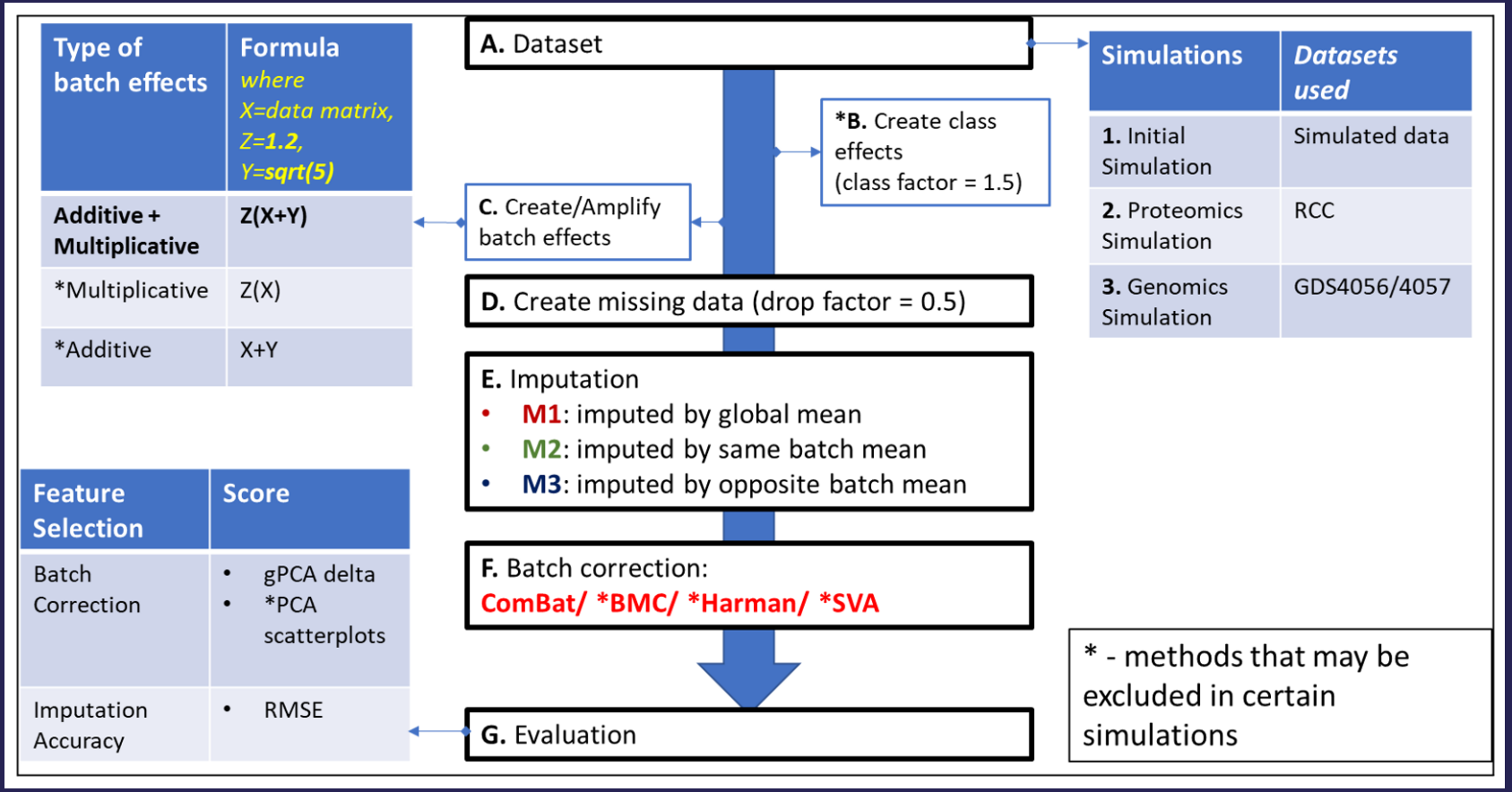
Impact?

Batch Effect Correction Algorithm (BECA)



- Late Pre-processing

Simulation design (at a glance)



Please open the folder “Missing Value Imputation”

1) Test Simulations

- 20x20 data matrix simulation on 4 BECAs
- Repeated 10 times
- Takes about 10 mins to run
- OUTPUT: RMSE & GPCA DELTA BOXPLOTS

2) 4056 and 4057 ER+ Breast cancer dataset

- 32 samples in batch 4056
- 32 samples in batch 4057
- Only one true sample class: ER+

3) Genomic simulations using ER+ Breast cancer dataset

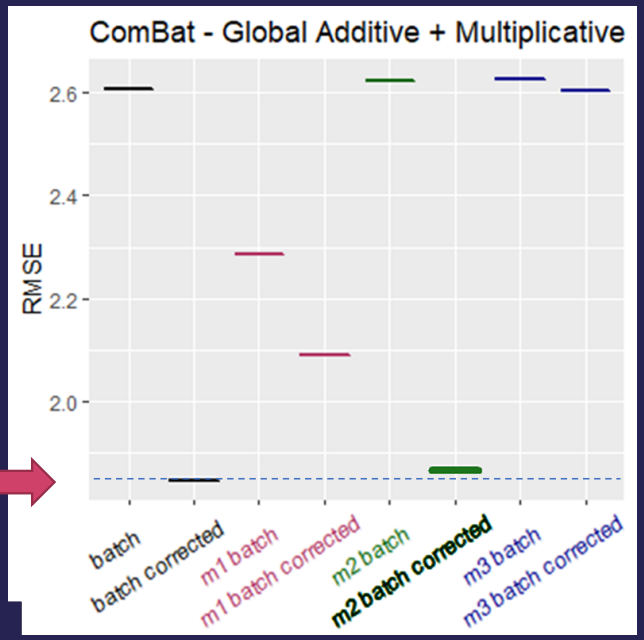
- Make sure you change your working directory to where the 4056_4057_ER+ file (2) is kept
- Simulate batch effects only
- OUTPUT: RMSE & GPCA DELTA BOXPLOTS

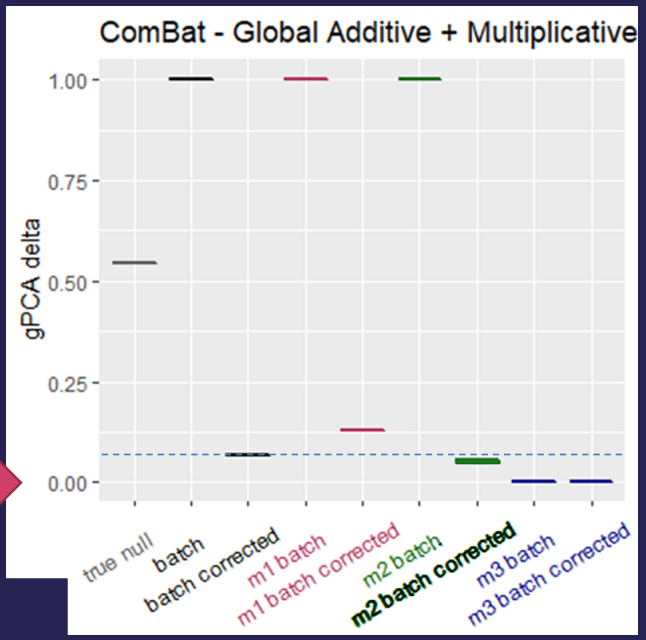
4) PCA scatterplot

- Need run genomics_sim (3) first
- Save data in workspace
- To visualise PCA scatterplots, sample boxplots and IQR plots for genomics data.

M2 works (beautifully) on genomic data

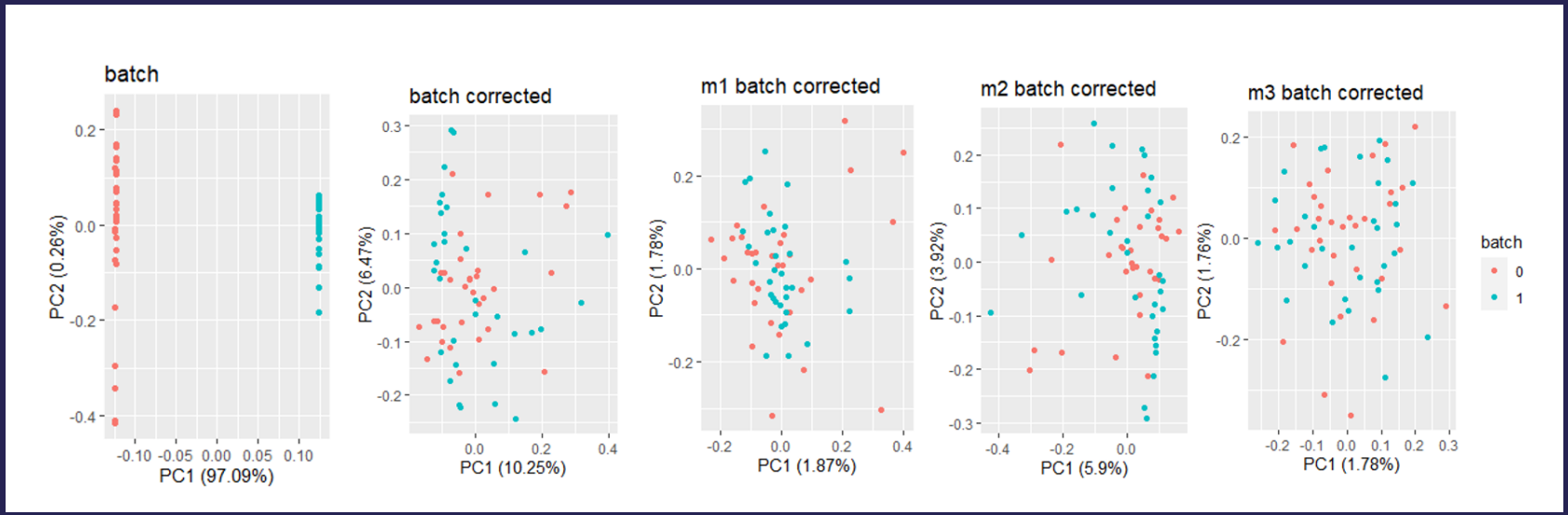
- M2 has the lowest RMSE
- M2 also has a pretty good gPCA delta (a relative estimation of batch proportion in data)





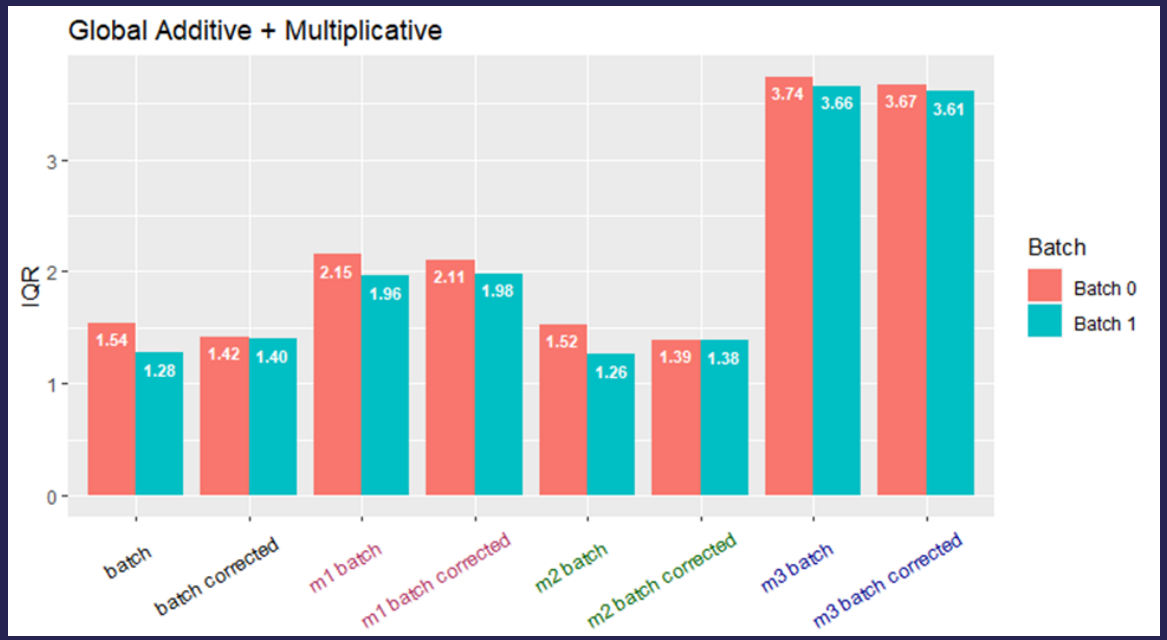
M2 works (beautifully) on genomic data

- M2 has comparable gPCA to M1
- In PCA, it also does not appear impressive…but…



M2 works (beautifully) on genomic data

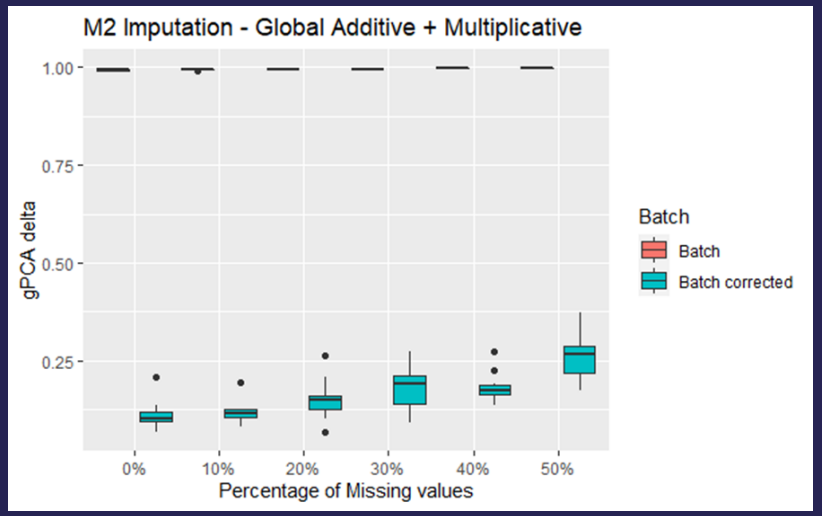
- …M1 and M3 do result in increased noise in the data even if the batch effects appear to be "mitigated" (y-axis: sample value distribution)
- Similar findings also for proteomics data (not shown)

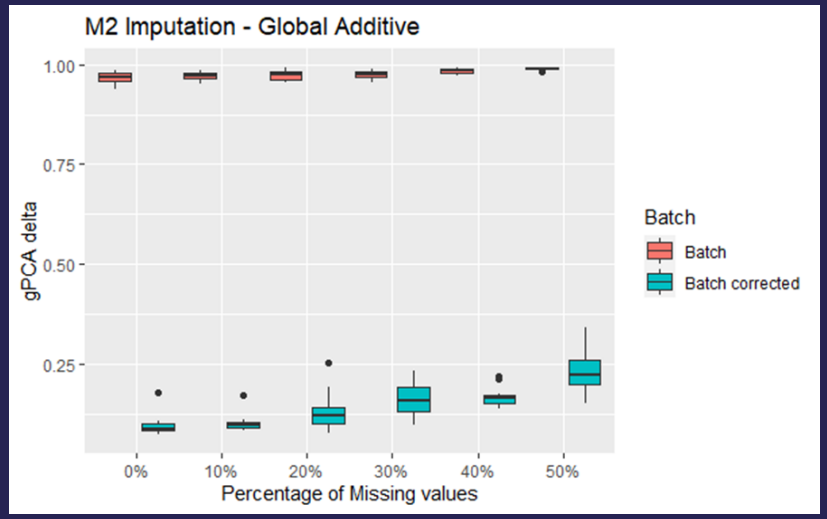


Batch corrected | M1 | M2 | M3

M2 missingness and gPCA deltas

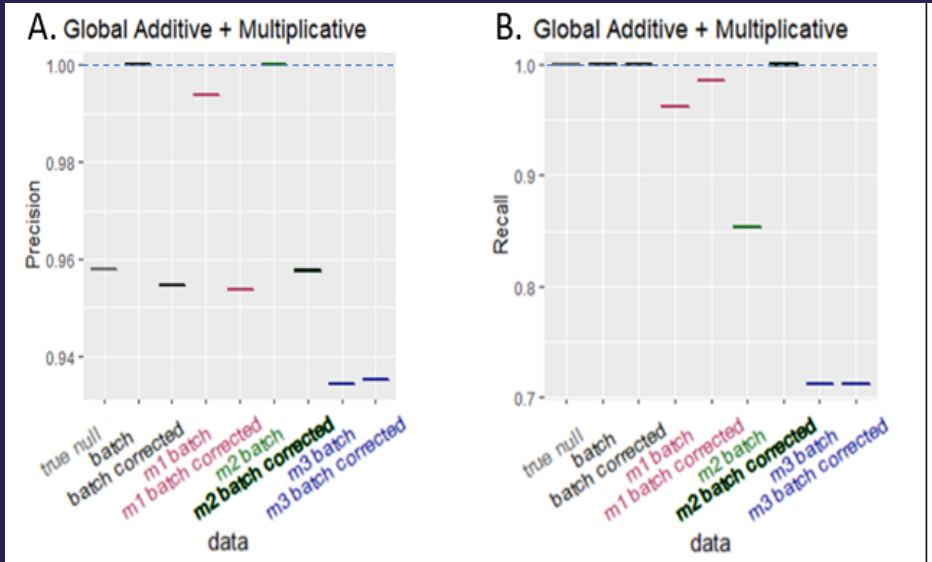
- There is positive correlation between the percentage of missing values and gPCA delta post cleaning



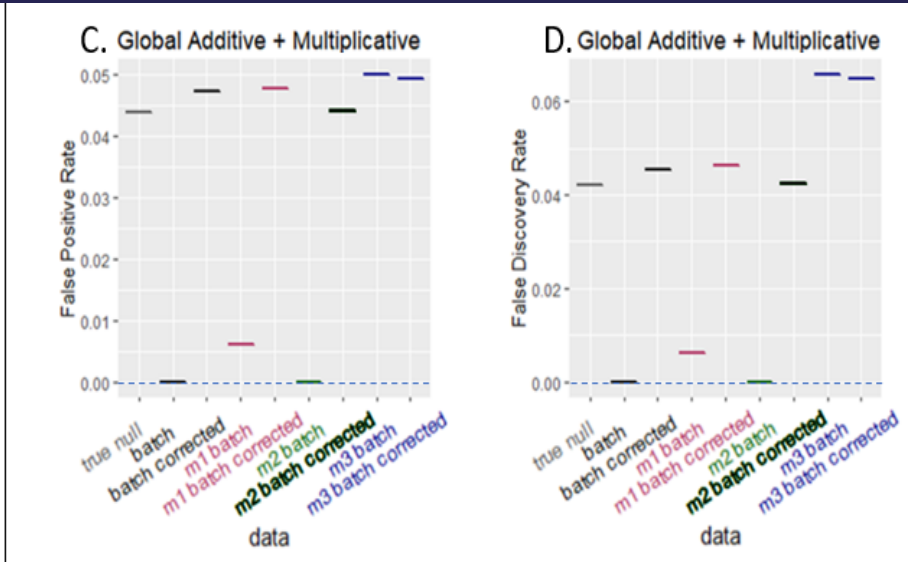


M2 dominates in feature selection

- M2 has highest precision and recall
- M2 also has the lowest false positive rates and false discovery rates
- It should be noted in general MVIs do tend to drive up errors as a whole



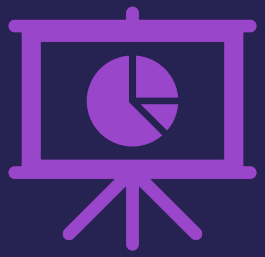
Precision & recall: the higher the better



FPR & FDR: the lower the better

Key Takeaway

If you know a batch factor (or any non-negligible co-variate e.g. age, gender, etc) exists in your data and you want to impute missing values…



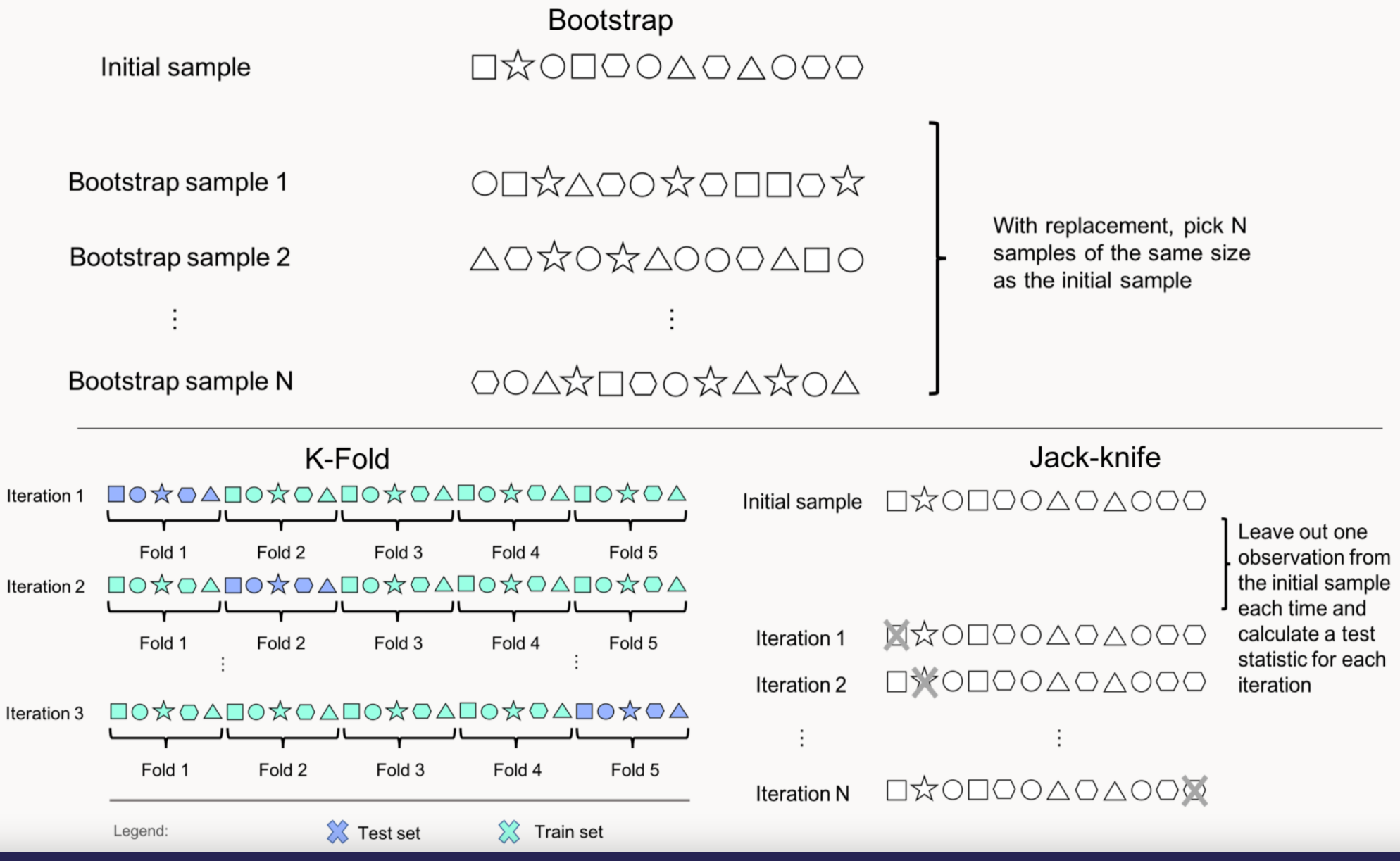
…Do not impute based on global mean or make a general assumption.

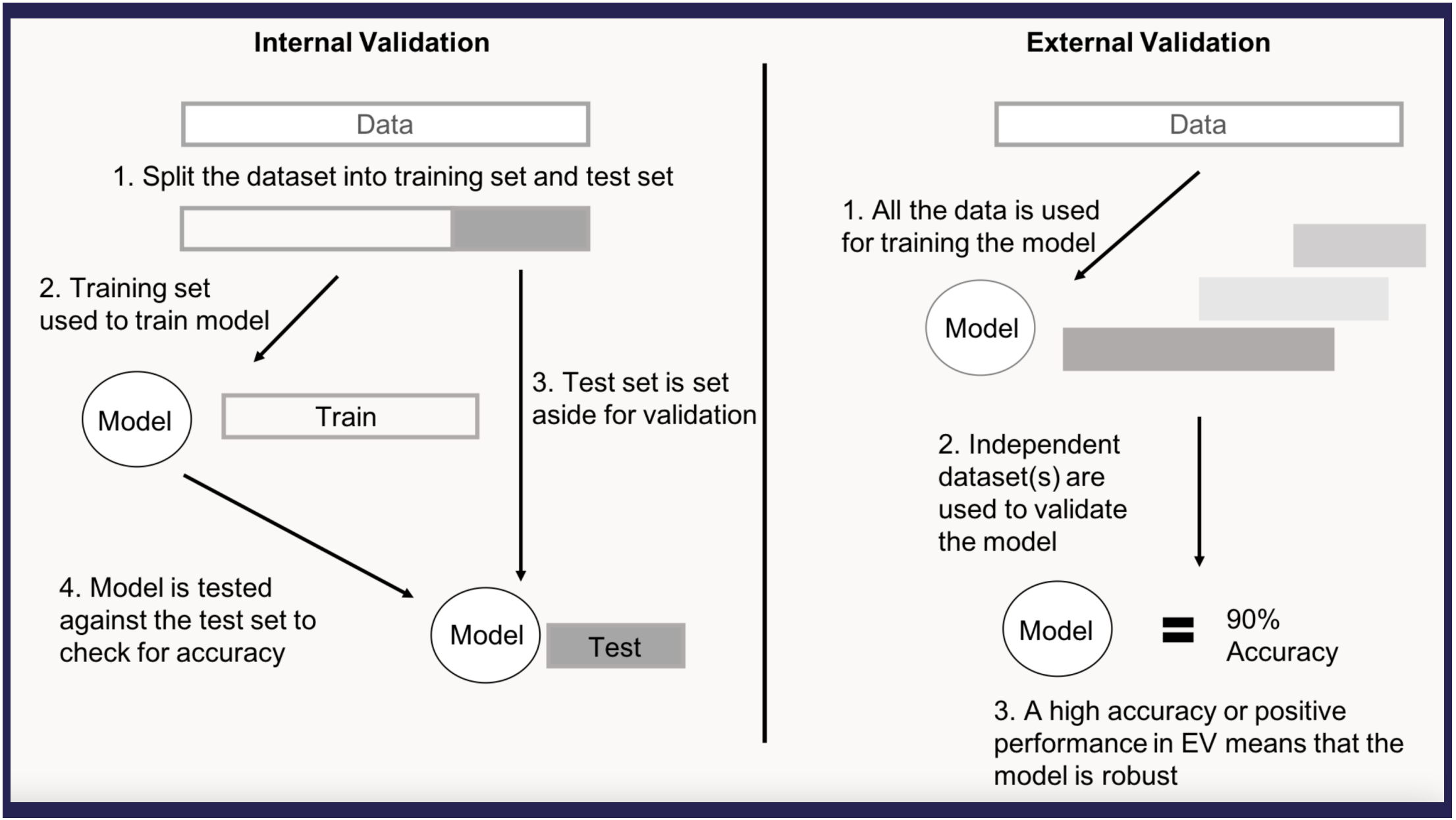
Make sure you impute based on same batch samples only...

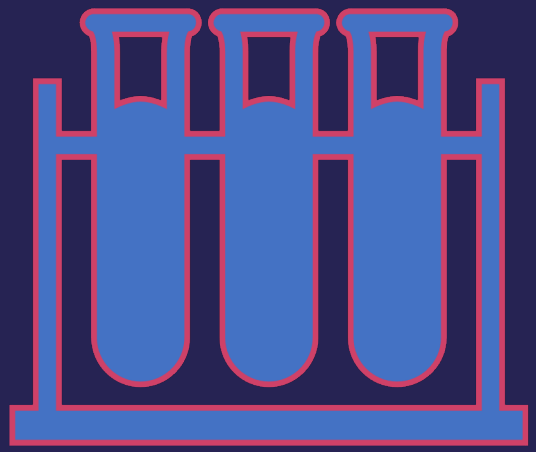
Doppelganger Effect

Beware of good performance in your machine learning experiment

Joint work with Wang Li Rong

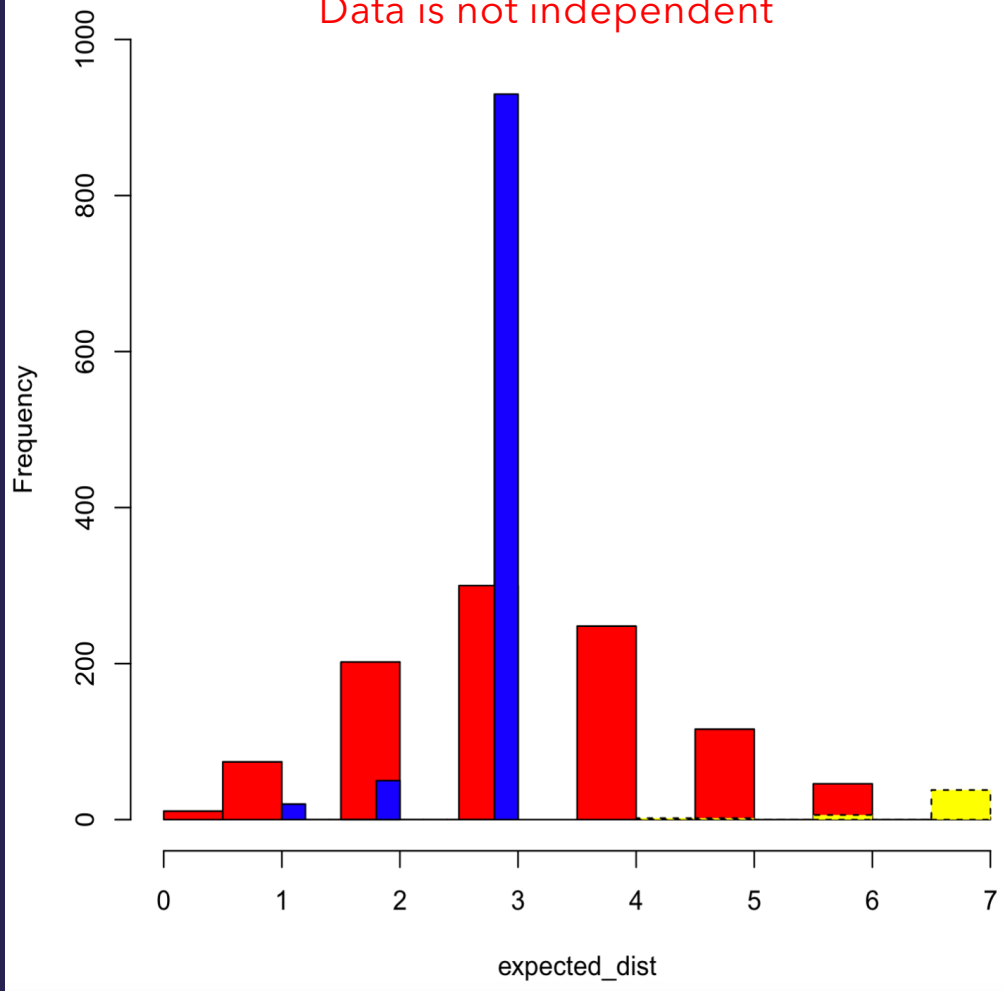

Bootstrap
Initial sample
Bootstrap sample 1
Bootstrap sample 2
Bootstrap sample N
With replacement, pick N samples of the same size as the initial sample
K-Fold
Iteration 1
Iteration 2
Iteration 3
Fold 1
Fold 2
Fold 3
Fold 4
Fold 5
Legend:
Test set
Train set
Jack-knife
Initial sample
Iteration 1
Iteration 2
Iteration N
Leave out one observation from the initial sample each time and calculate a test statistic for each iteration


Internal Validation
Data
1. Split the dataset into training set and test set
2. Training set used to train model
Model
Train
3. Test set is set aside for validation
4. Model is tested against the test set to check for accuracy
Model
Test
External Validation
Data
1. All the data is used for training the model
Model
2. Independent dataset(s) are used to validate the model
Model
90% Accuracy
3. A high accuracy or positive performance in EV means that the model is robust

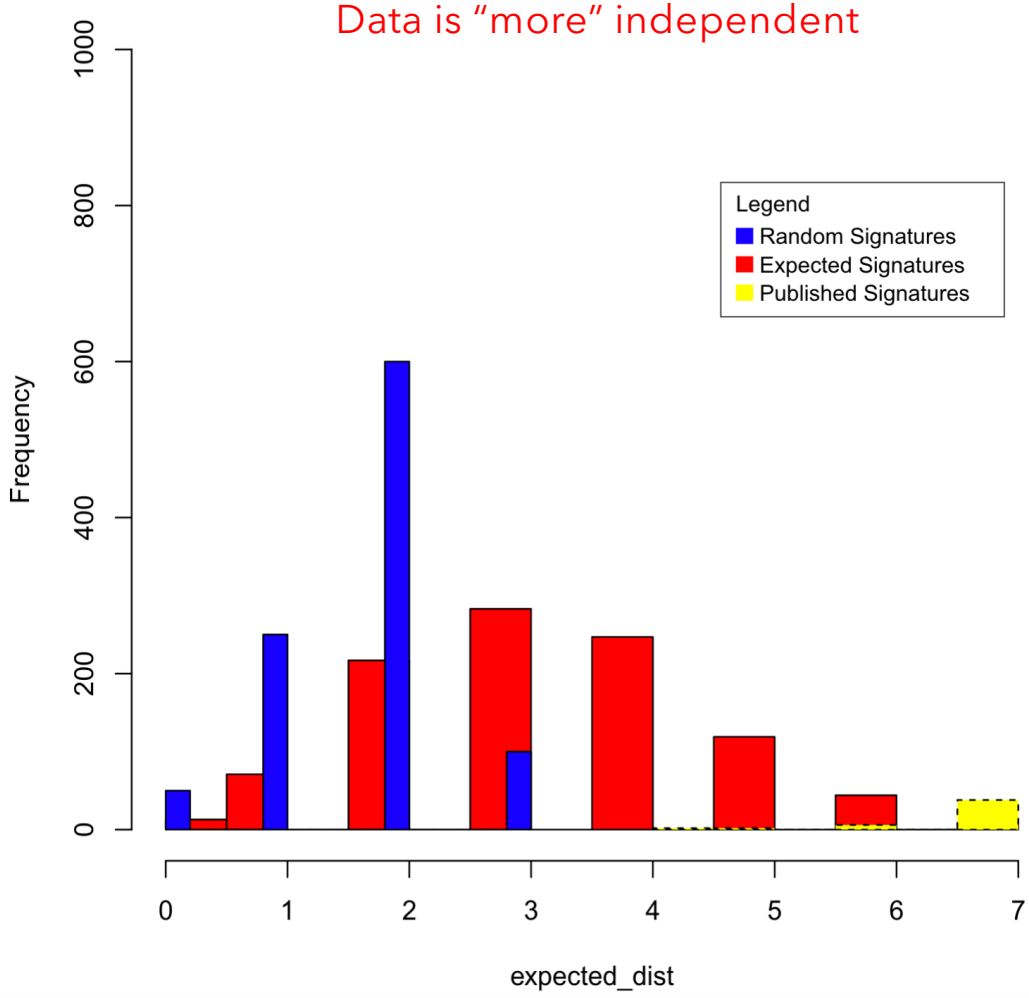


These techniques rely on assumption of independence between samples and data. Unfortunately, data is not as independent as you may hope for

Data is not independent

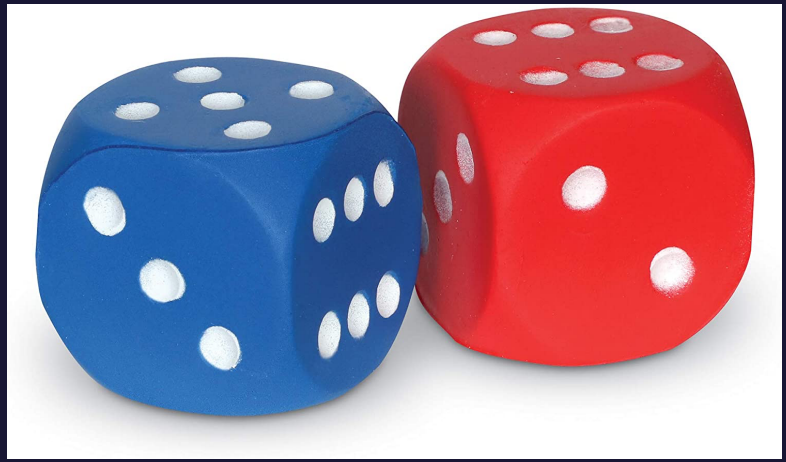


Data is “more” independent



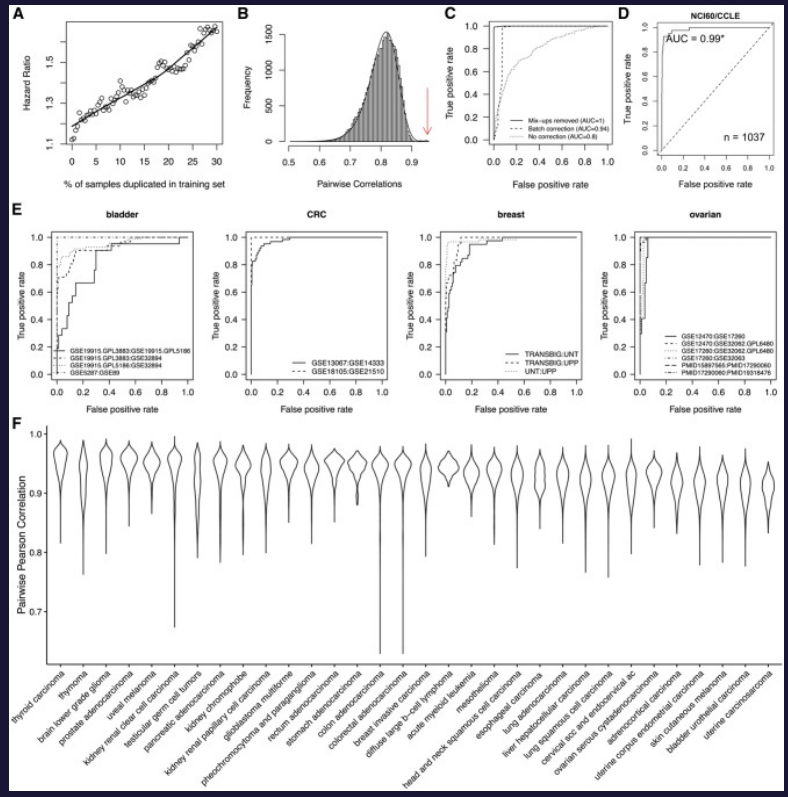
The Doppelganger effect occurs when data is not independent

- When these non-independent datasets are separated into training and test sets in ML (under typical assumption of independence) they inflate ML performance metrics
- This effect is similar to data leakage (where the same sample or replicates of the same sample end up in both training and test sets)
- But in Doppelganger effect, samples are similar due to chance



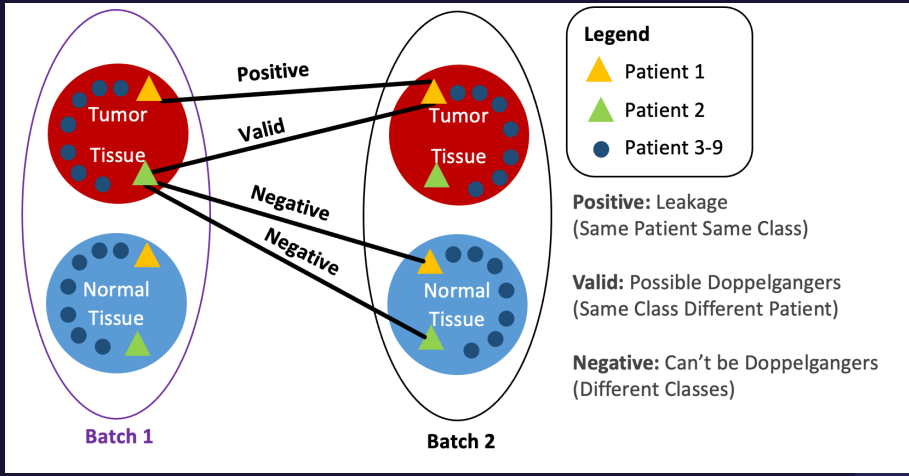
How to check for Doppelgangers?

- The Pairwise Pearson Correlation Coefficient has been proposed as one approach (Waldron, 2016)
- It works by calculating a batch-corrected correlation coefficient between samples of different data batches
- The problem(s) was that what they discovered were actually leakage due to same sample being used in different experiments.



Please open the folder “DoppelgangerCode”

An introduction to the dataset we will be using --- the Renal Cancer (RC) dataset

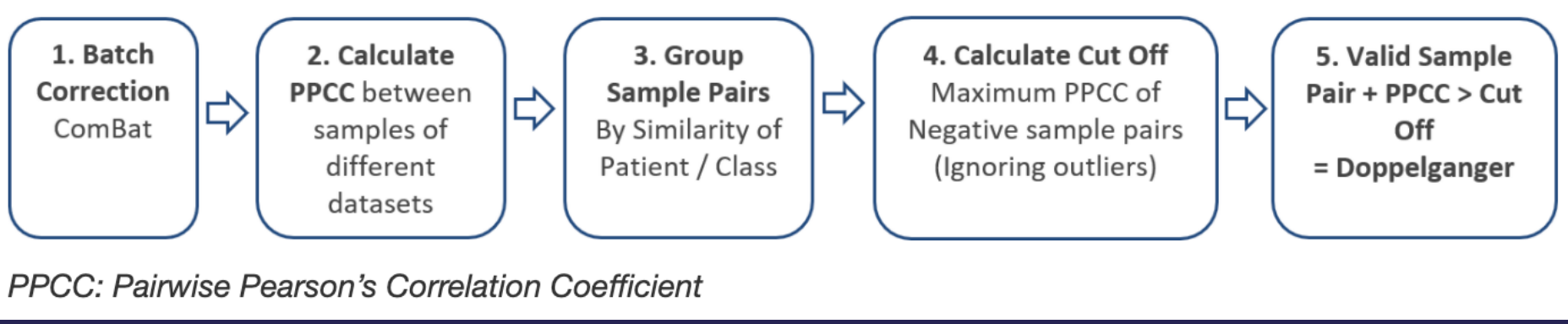


We use data where the ground truth is known a priori. And provide a proper technical definition
A doppelganger pair is when samples belonging to the same class but are not technical replicates are similar by chance

RC_Metadata

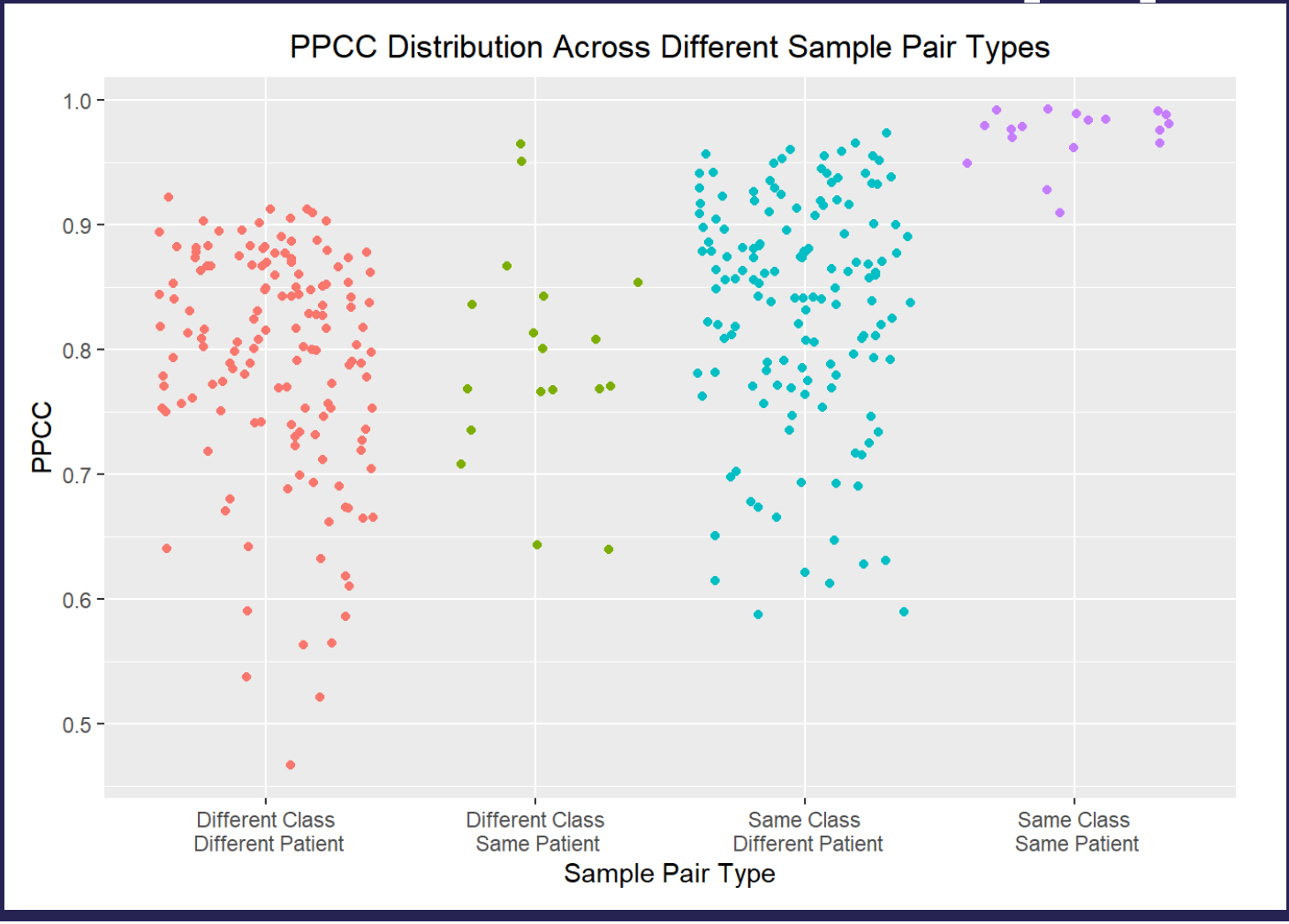
	Class	Tumour_Histological_Type	Patient_ID	Sample_ID	Replicate
normal_cc_patient_1_rep1	Normal	cc	1	1	1
normal_cc_patient_1_rep2	Normal	cc	1	19	2
tumor_cc_patient_1_rep1	Tumor	cc	1	2	1
tumor_cc_patient_1_rep2	Tumor	cc	1	20	2
normal_cc_patient_2_rep1	Normal	cc	2	3	1
normal_cc_patient_2_rep2	Normal	cc	2	21	2
tumor_cc_patient_2_rep1	Tumor	cc	2	4	1
tumor_cc_patient_2_rep2	Tumor	cc	2	22	2
normal_cc_patient_3_rep1	Normal	cc	3	5	1
normal_cc_patient_3_rep2	Normal	cc	3	23	2
tumor_cc_patient_3_rep1	Tumor	cc	3	6	1
tumor_cc_patient_3_rep2	Tumor	cc	3	24	2
normal_p_patient_4_rep1	Normal	p	4	7	1
normal_p_patient_4_rep2	Normal	p	4	25	2
tumor_p_patient_4_rep1	Tumor	p	4	8	1
tumor_p_patient_4_rep2	Tumor	p	4	26	2
normal_ch_patient_5_rep1	Normal	ch	5	9	1

Our approach at a glance

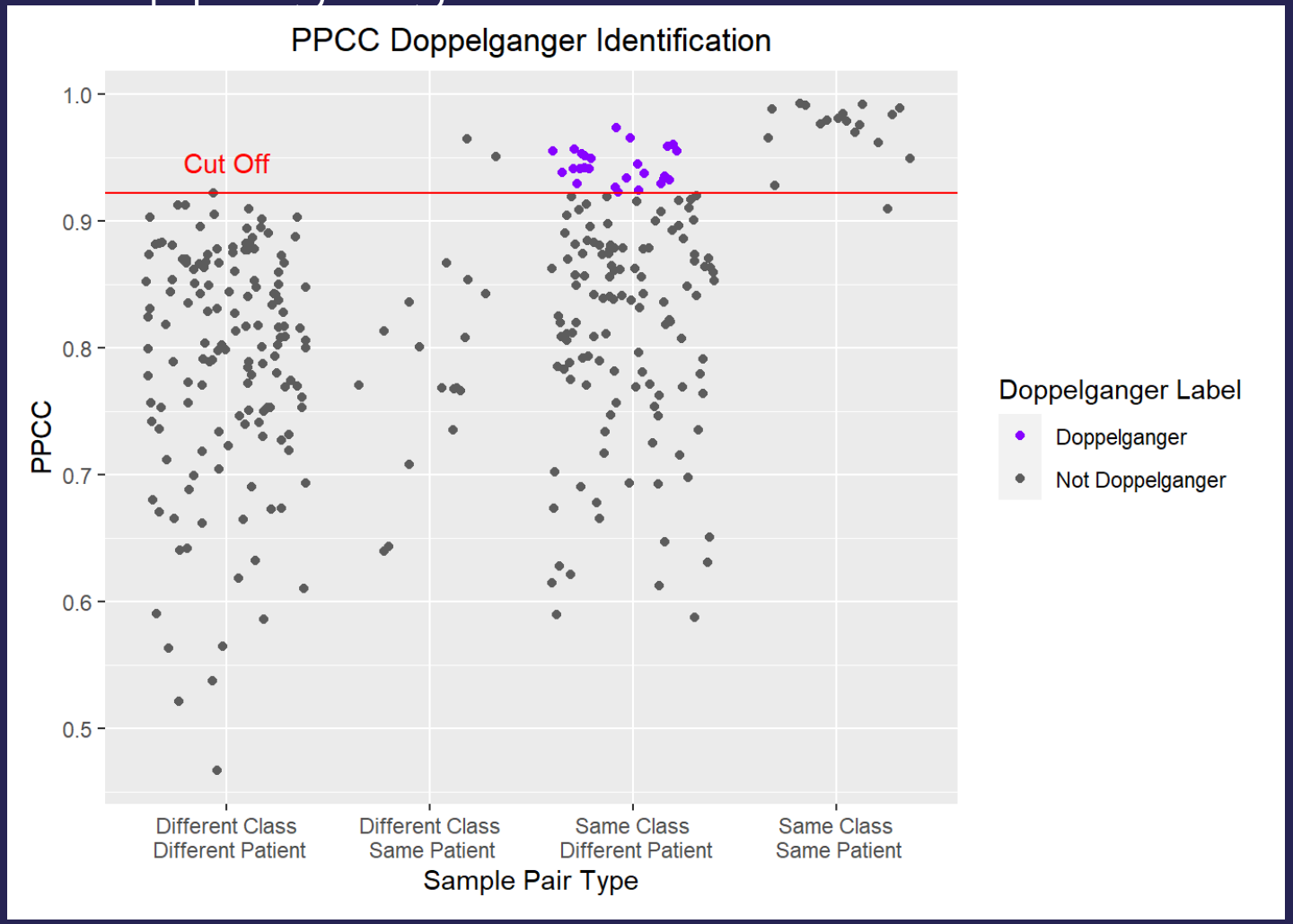


PPCC: Pairwise Pearson's Correlation Coefficient

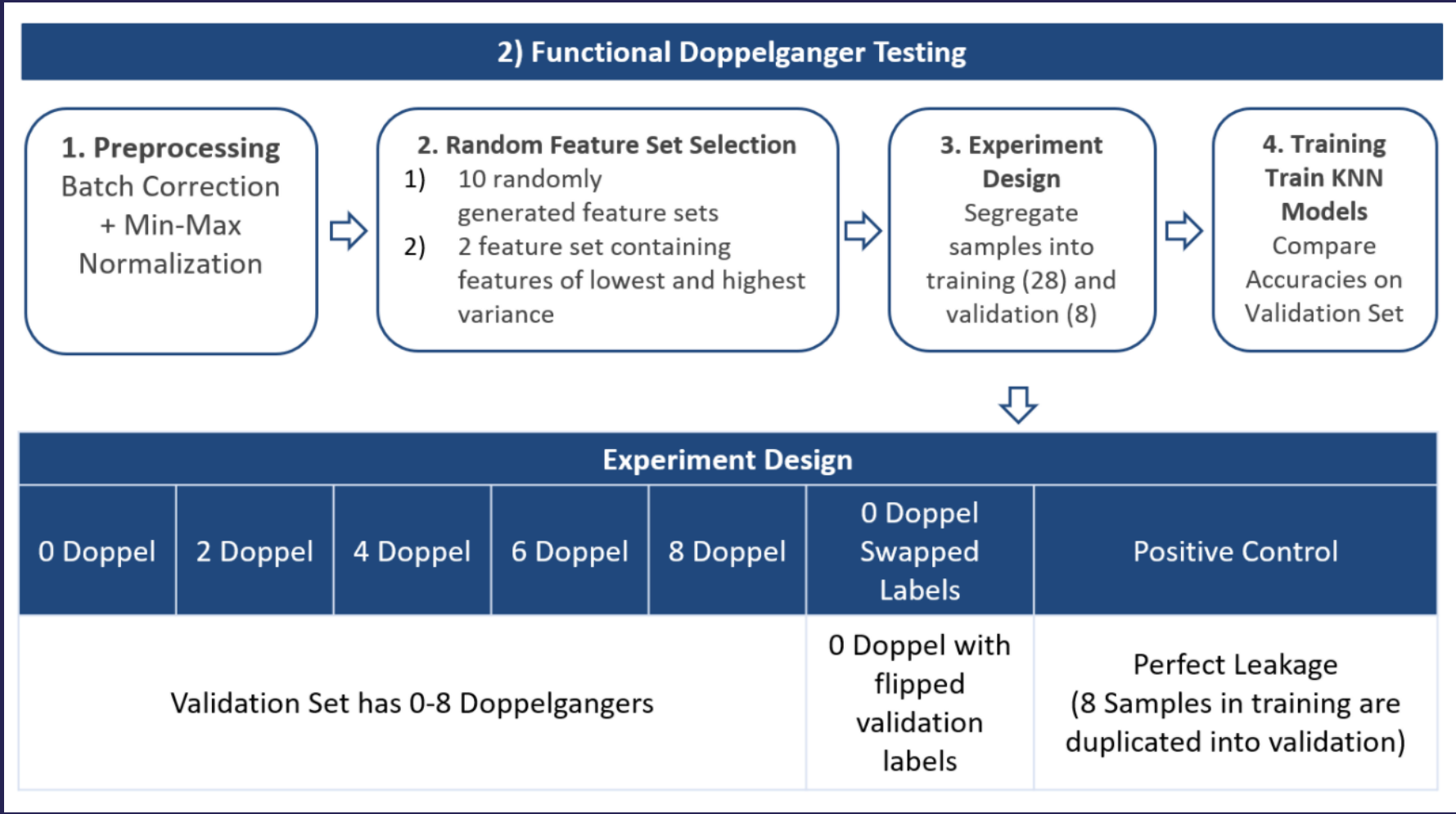
PPCC distribution across different sample pair types



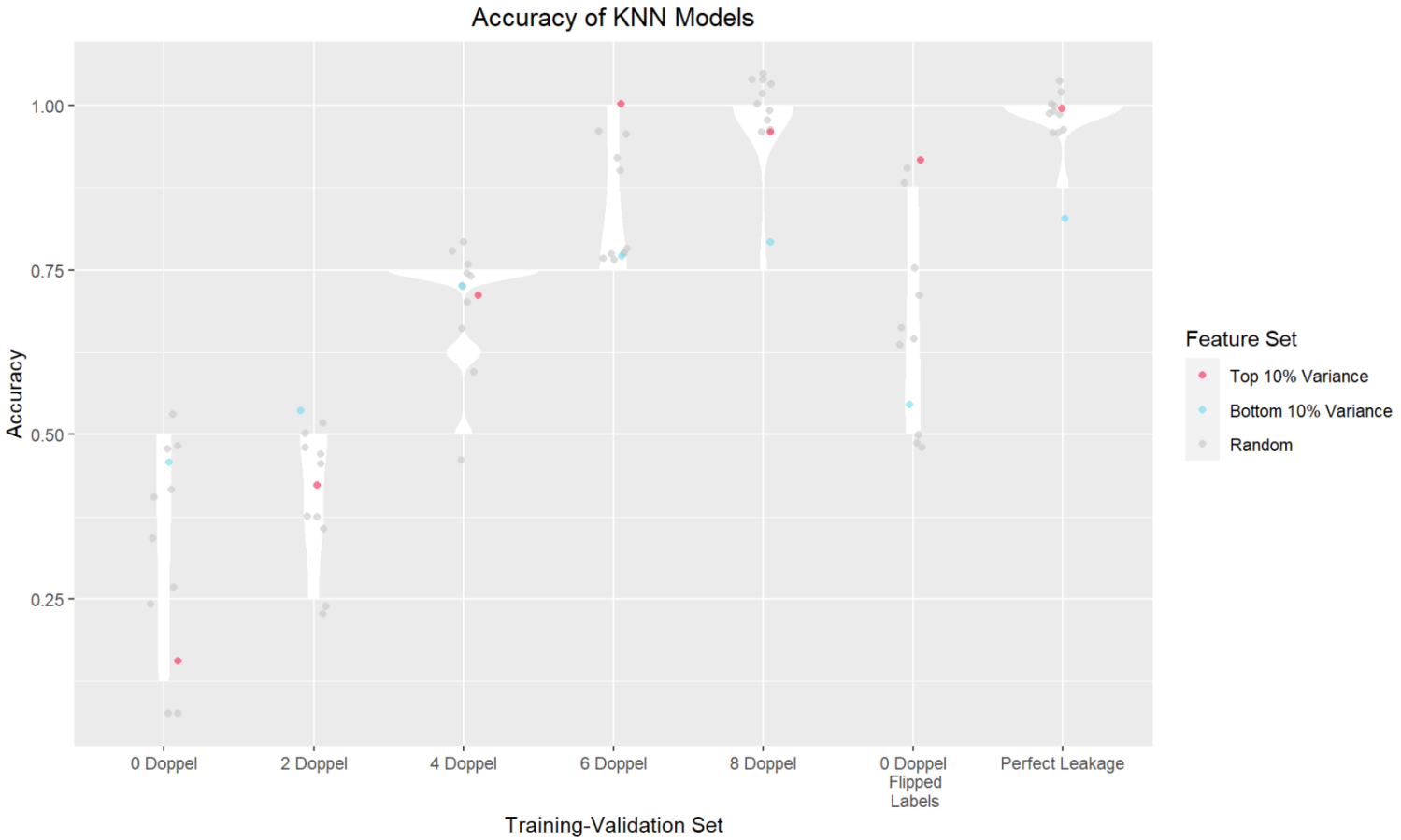
PPCC Doppelganger identification



Functional doppelganger testing in machine learning

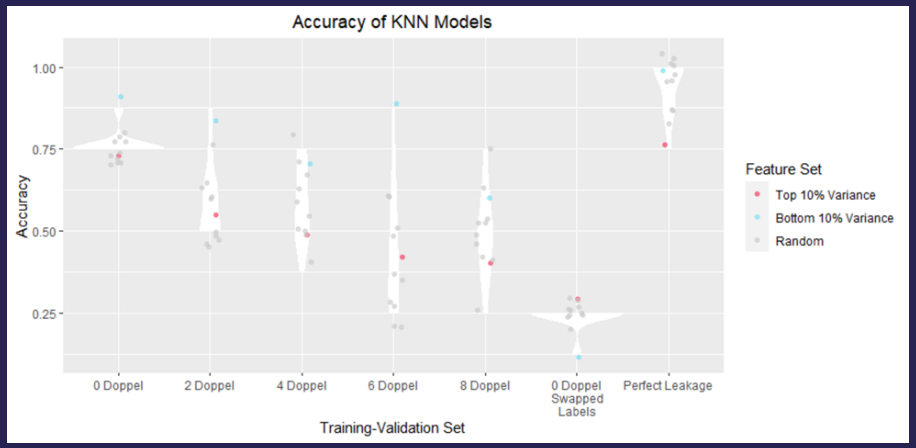


Doppelgangers act like leakage problems

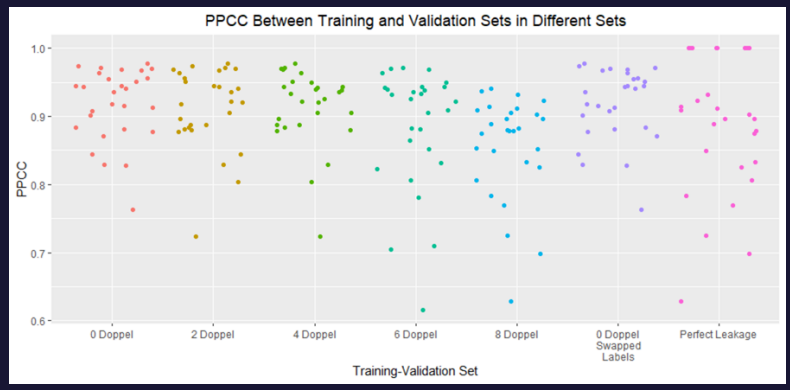


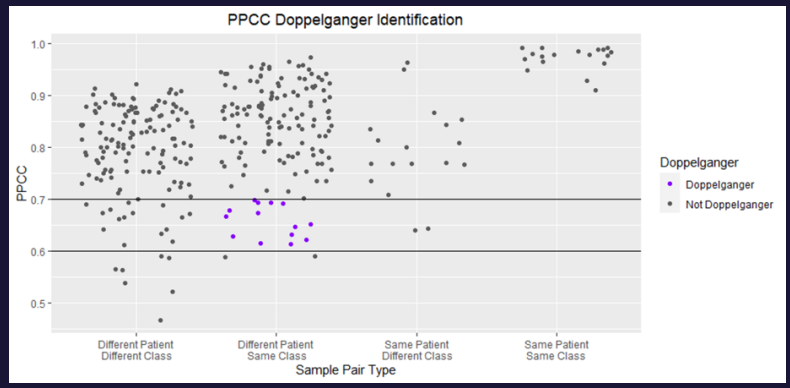
We observe a dosage dependent relationship between the number of doppelgangers and the accuracy of models on the validation set since accuracy increases as the number of doppelgangers in validation set increases.

So let us see what happens if we pick sets that are NOT doppelgangers. Does it still have a dosage effect?



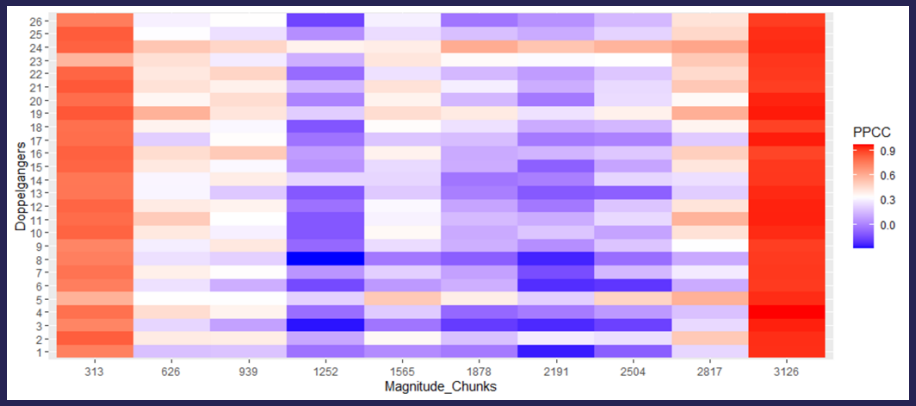
Taking weakly correlated pairs and spiking them into models does not have a clear impact on model performance





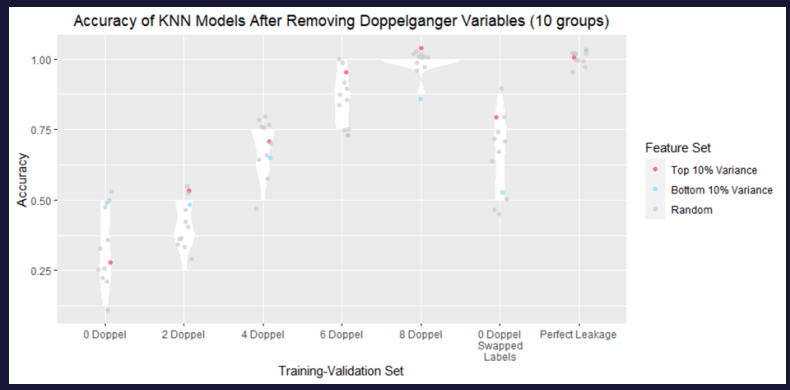
Does discretization and trimming the data help?

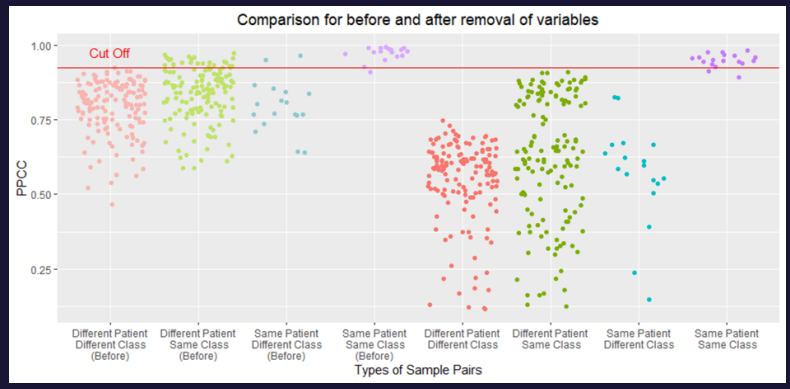
Order Variables by Magnitude (Sum) → Equal-Frequency Discretization (10 Bins) → Recalculate PPCC with Each Bin of Variables → Visualize with Heatmap



Turns out that the correlations are most extreme on the ends

Variables in bins with high PPCC (Marked in Red) = Doppelganger Variables





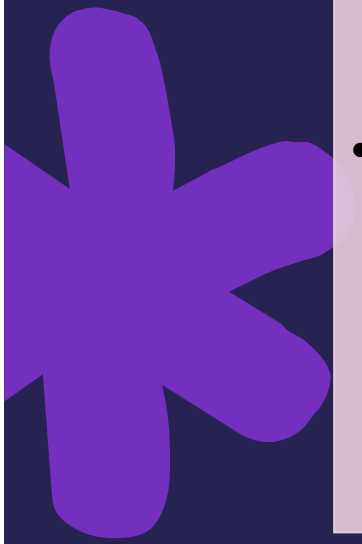
Key Takeaway

- Doppelganger effect results in an inflation of validation accuracies
- When training models with biomedical data, it is essential to check for functional doppelgangers before separating the dataset into training and validation sets
- Biological samples tends to have surprisingly high mutual correlations
- To avoid the doppelganger effect, doppelgangers should never be in both training and validation sets. They should be put either all in training or all in validation.

References (Do check out)

- Ho et al. Extensions of the external validation for checking learned model interpretability and generalizability. Patterns 2021.
- Ho et al. Avoid oversimplifications in machine learning: Going beyond the class-prediction accuracy. Patterns 2020.
- Goh and Wong. Advanced bioinformatics methods for practical applications of proteomics. Briefings in Bioinformatics 2019.
- Goh and Wong. Dealing with confounders in -omics analysis. Trends in Biotechnology 2018.

Our other project areas..



Data science (TCM)	AI in Education	Statistics	Meta-ML problems
• Predicting patient disease using TCM features • Feature engineering and database design	• Performance prediction • Intelligent coaching platforms • Co-evolution with pedagogy • Graph literacy • Bio-data science education	• Anna Karenina Principle • Relooking Neyman-Pearson approaches • Stability and reproducibility problems • Network analytics • Batch effects	• Sloppy models • No-free lunch theorem • Rashomon set problems • Doppelganger effect • Better evaluation metrics • Better feature engineering

Acknowledgements

- MOE ACRF 1 and 2
- National Research Foundation, Singapore
- Singapore Data Science Consortium
- National Medical Research Council
- My supportive colleagues in SBS, LKC and NUS
- My amazing team!
- Special thanks to my CNY-OFYP student Priscila and to my URECA student, Lirong

Questions?

Thanks for listening