A short guide to survival in statistics

In an AI world

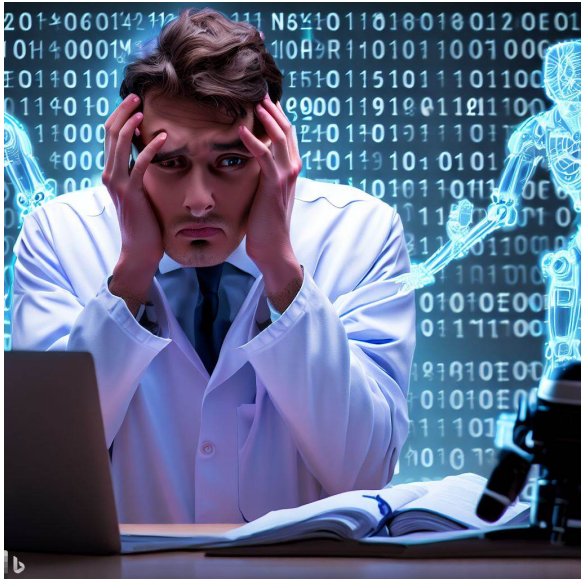
Piotr Tymoszuk, PhD

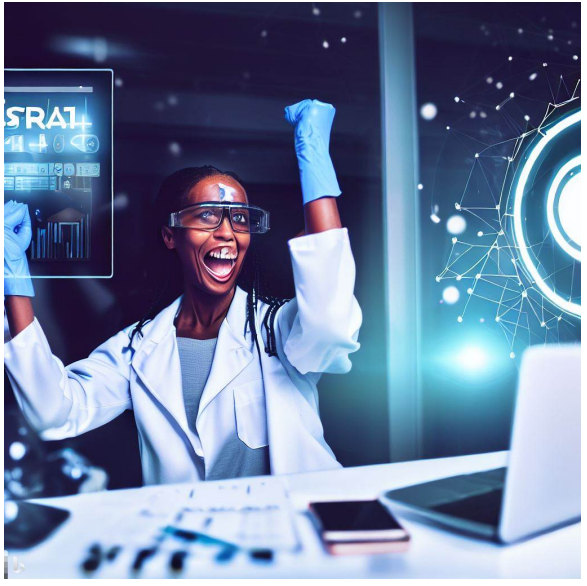
Freelance Data Scientist,
daas.tirol

Supported by GPT-4

The general aims

- Make you understand basic concepts of modeling and machine learning in clinical research
- Turn your researcher, reviewer and clinician experience with machine learning from a struggle to a great adventure





What you are NOT going to hear about

✗ Mean, standard deviation, median, quantiles

✗ Statistical tests, ANOVA, correlations, linear regression (well almost..)

✗ Odds and hazard ratios, confidence intervals

✗ P values

What you will DO hear about

✓The principle and considerations of modeling

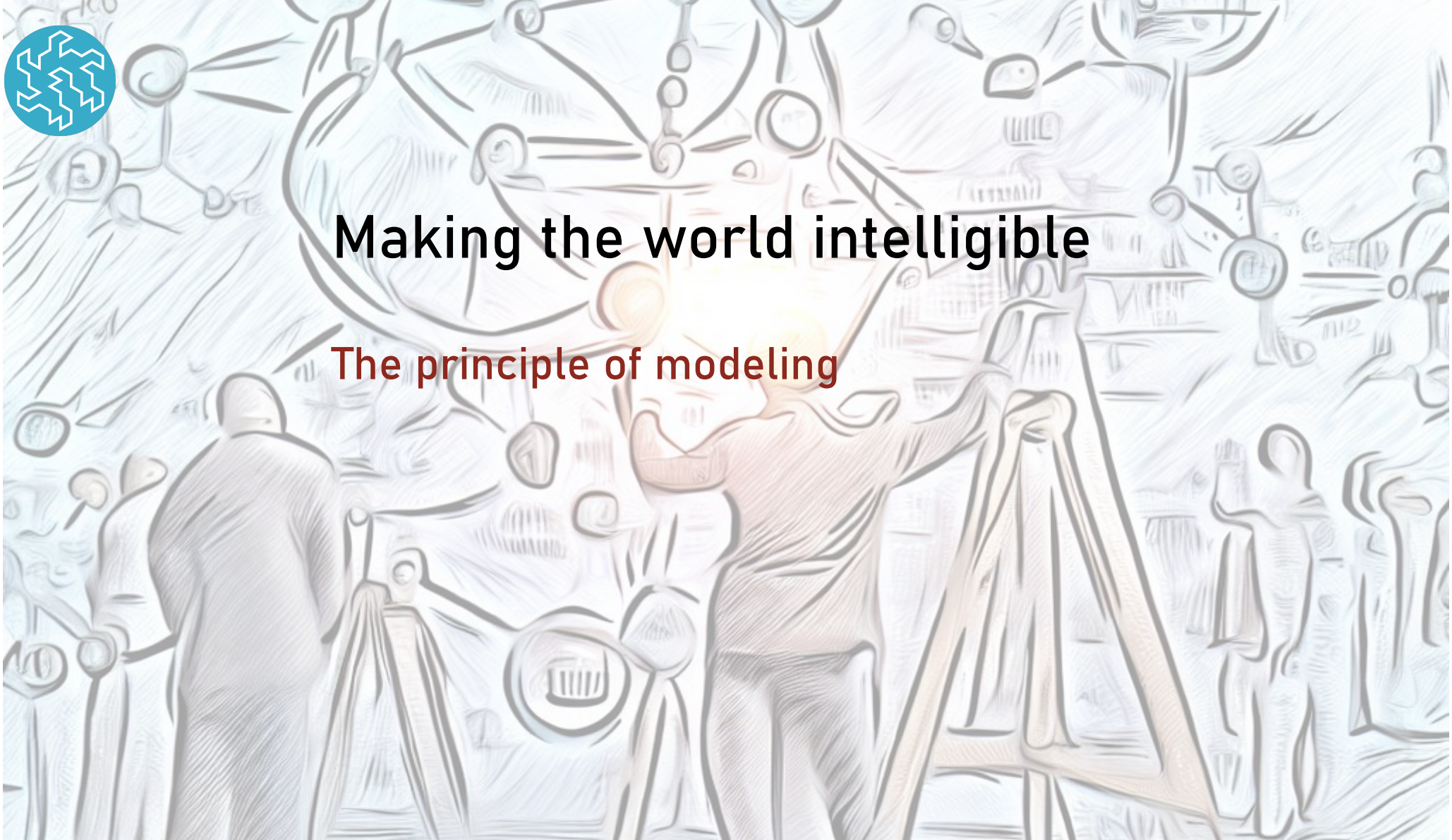
✓How to validate a model (why the hell do that?!)

✓The most common machine learning algorithms:

- Linear models
- Tree models: random forests and gradient boosted machines
- Neuronal networks
- Support vector machines (SVM)

✓How to evaluate performance and meaningfulness of a model?

✓Is machine learning a black box? How to interpret an AI model?



Making the world intelligible

The principle of modeling

The principle of modeling: analysis goals

- Every statistical analysis begins with a curious or perplexing question
- This question must be a scientifically sound one, i. e. we must have analytic or experimental tool to challenge it:

> *In so far as a scientific statement speaks about reality, it must be falsifiable; and in so far as it is not falsifiable, it does not speak about reality*
>
> *Karl Popper*

- In statistics, such scientific question is called hypothesis. Its negation is called null hypothesis.

The principle of modeling: scientifically unsound questions

- What was before the Big Bang?
- Who killed J.F. Kennedy?
- Basically all conspiracies!
- Questions considering religion, theology and metaphysics
- Every question, which cannot be answered with yes or no by results of an experiment or a human study!

The principle of modeling: the clinician's questions

- We see quite a lot of structural lung lesions in CT in our patients. This lesions are of different types and severity.
- Are such lesions clinically/functionally significant?
- Which lesions? Do they need to be more frequent or severe to have an effect on lung function?
- Is there a group of patients, which is going to have lung function problems following COVID-19?

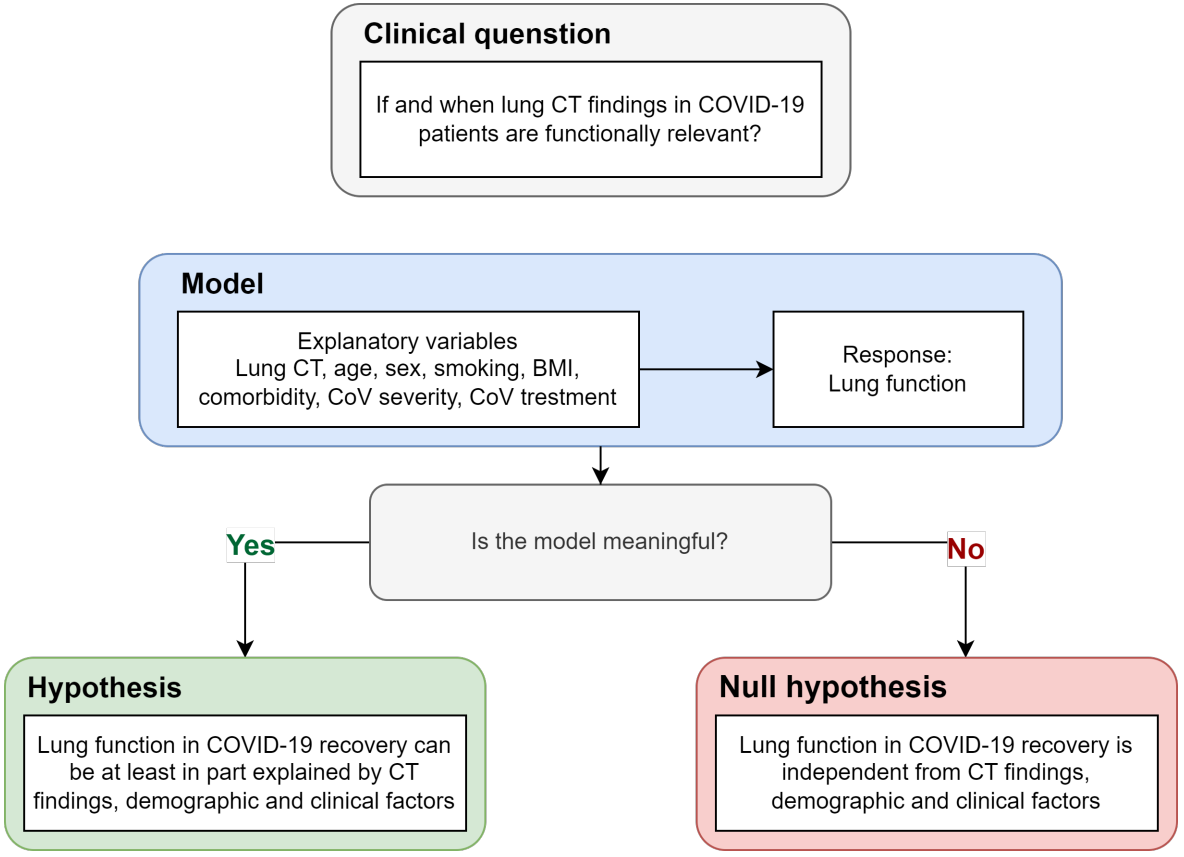
The principle of modeling: the data scientist's questions

- Hypothesis: we can explain lung function problems in COVID-19 convalescents by presence, type and severity of radiological findings, along with demographic, medical history and acute disease data

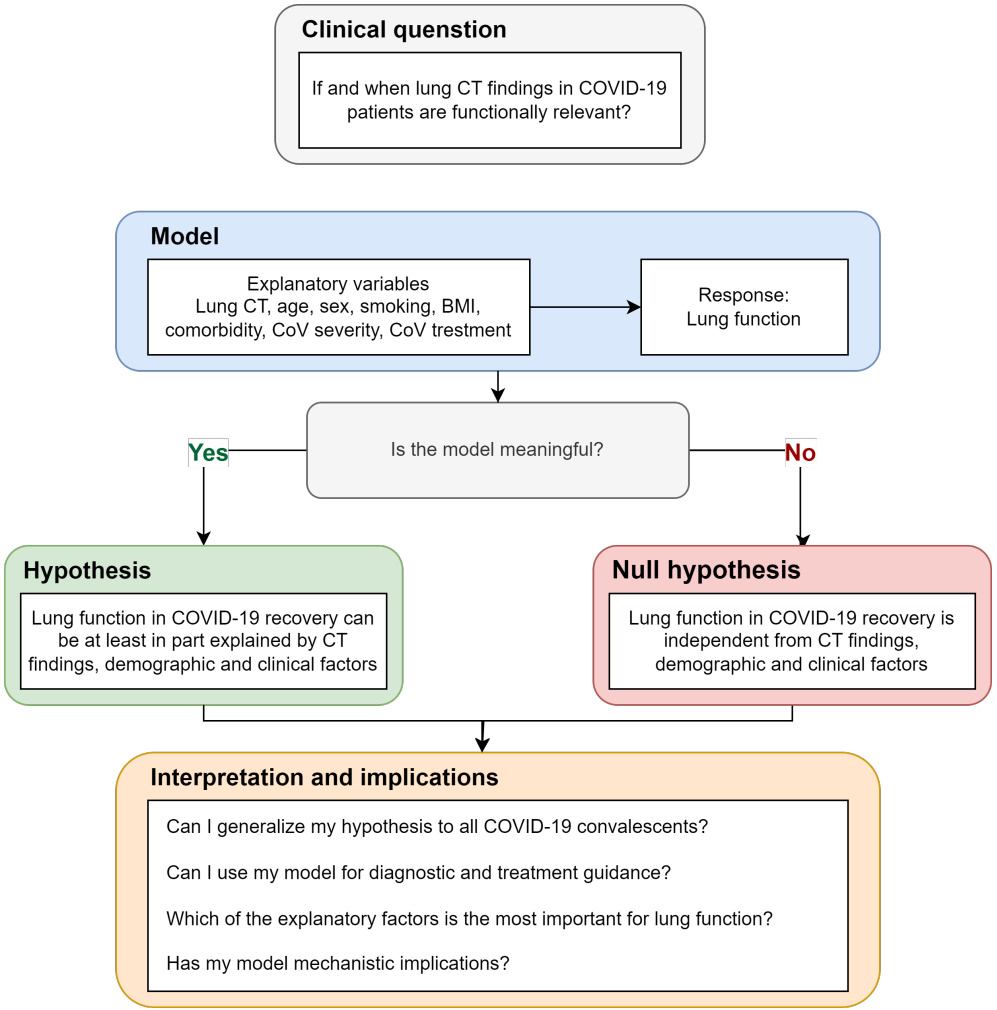
- Null hypothesis: lung function problems are independent from the variables listed above

- The tool: multi-parameter modeling. If hypothesis is true, I can build a meaningful model. Else if null hypothesis is true, my model wont be better than a random guess (dummy model)

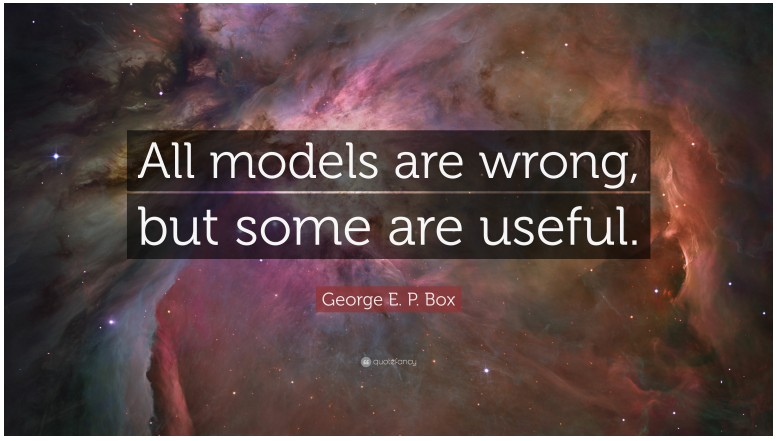
The principle of modeling



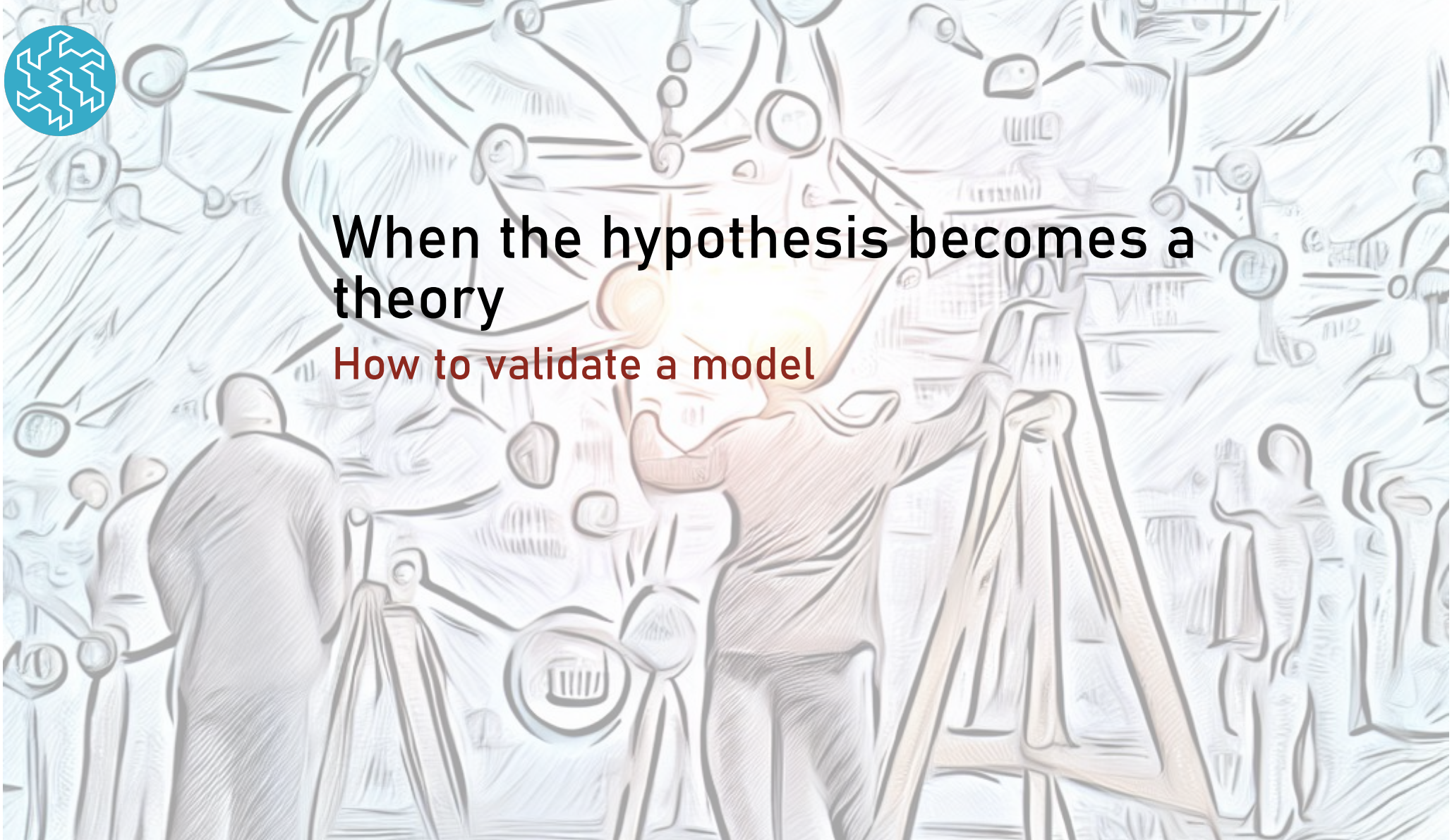
The principle of modeling



The principle of modeling: usefulness and generalization



- Well, with my study cohort dataset, I found a nice model of lung function – it's feasible that lung function is associated with CT findings, age, sex, CoV course and so on…
- Does my model describe only my cohort? Can I generalize it for all patients?
- The answer: validate it!



When the hypothesis becomes a theory

How to validate a model

How to validate a model

- ✗ A negative example (doi:10.3390/ijerph20176678). N = 189, hospitalized COVID-19 patients. The modeling response: persistent symptoms (long COVID)
- ✗ The tool: multi-variable logistic regression
- ✗ What's wrong with the analysis?

Table 4. Factors independently associated with persistent clinical symptom(s) at first and second follow-up visits in included patients' results of the multivariate logistic regression.

	First Visit		Second Visit	
Variables	**Adjusted OR (95% CI)**	***p***	**Adjusted OR (95% CI)**	***p***
Personal history				
BMI > 30 kg/m^2	3.52 (1.25–9.91)	0.02	3.80 (0.97–14.91)	0.06
Rheumatic diseases	-	-	5.25 (0.75–36.84)	0.09
Factors related to acute COVID-19				
Ageusia	-	-	0.17 (0.03–0.92)	0.04
Anosmia	-	-	13.34 (2.07–86.21)	0.006
AST according to quartiles (ref < 31)		0.003		0.03
31–42	8.68 (2.41–31.28)	0.001	10.27 (1.88–56.29)	0.007
>42	3.69 (1.32–10.31)	0.01	2.52 (0.60–10.68)	0.21
Delay (days) between onset of clinical symptoms and hospital admission	-	-	1.14 (1.01–1.29)	0.04
Pain	-	-	4.31 (1.23–15.05)	0.02
ICU stay	-	-	5.43 (1.39–21.25)	0.02

AST, aspartate transaminase; BMI, body mass index; CI, confidence interval; ICU, intensive care unit; OR: odds ratio.

How to validate a model

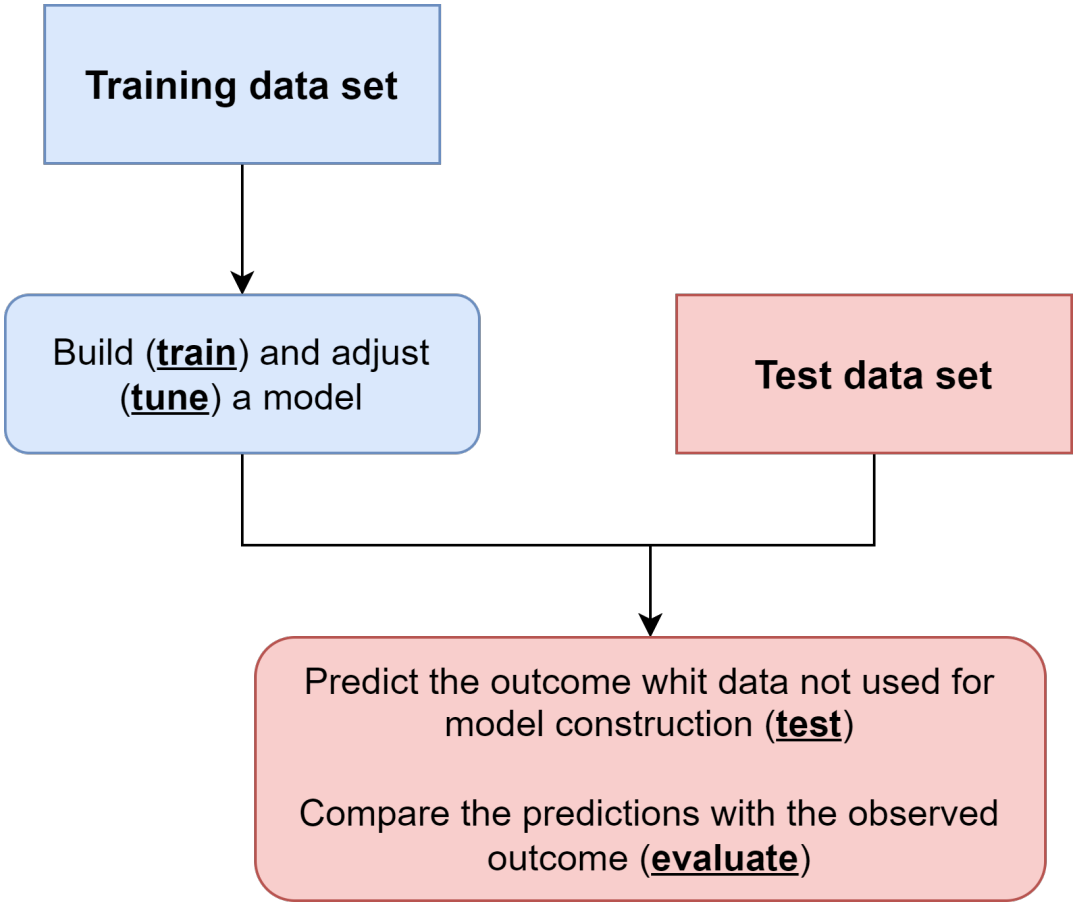
✗ How can you be sure that your model is useful, i.e. it describes a meaningful characteristic of long COVID?

✗ Could it be that your model reflects a peculiarity of your cohort?

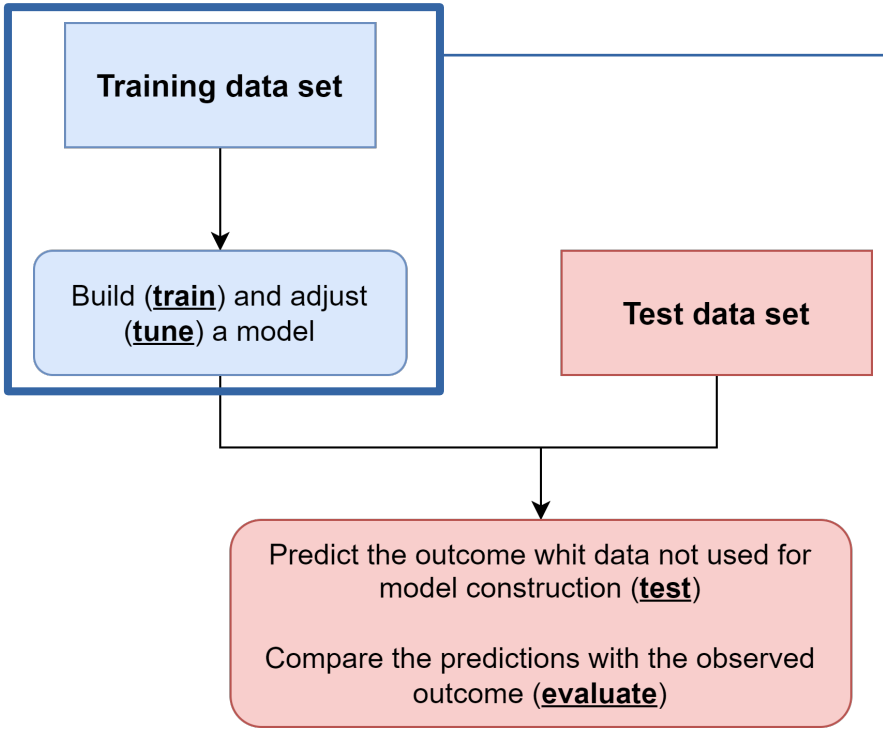
✗ Or it depends on few bizarre observations in your data set?

The answer: validate it!

How to validate a model: machine learning



How to validate a model: machine learning

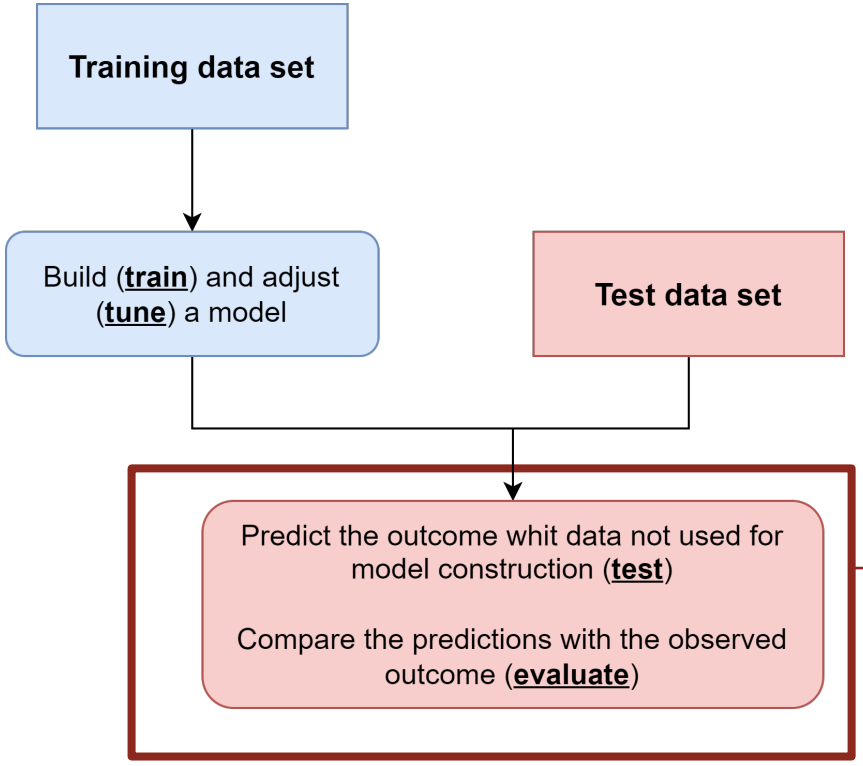


✗ My model fails during training

✗ Is my hypothesis wrong?

✗ Is there a technical problem with the training dataset or with the model?

How to validate a model: machine learning



✗ My model fails during testing/evaluation

✗ My hypothesis applies only to the training dataset, the model is useless

✗ Is my training data set not representative (bias!), noisy (poor data quality) or has many outliers?

✗ Is there a technical problem with the test data set?

How to validate a model: internal validation

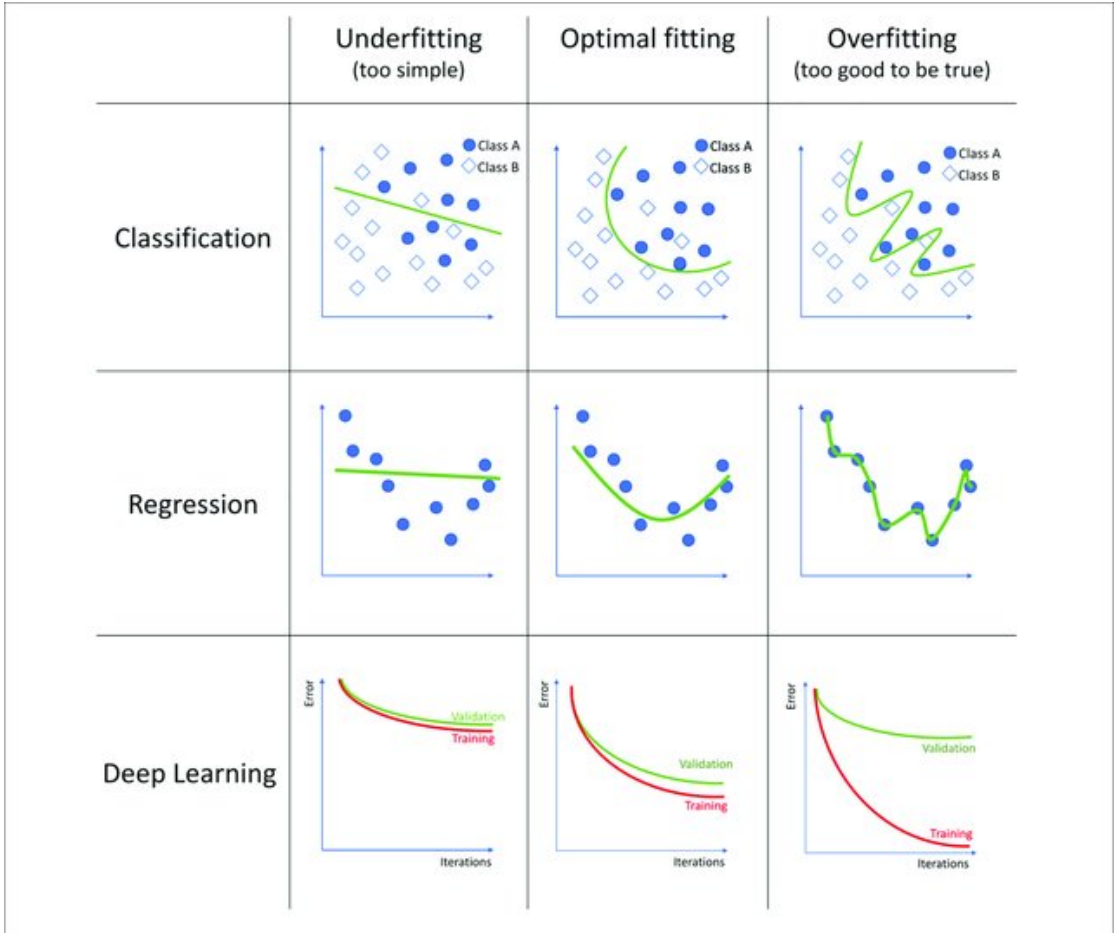
Sounds great, but I have just the one small patient cohort! Can I validate it anyway?

How to validate a model: internal validation

Sounds great, but I have just the one small patient cohort! Can I validate it anyway?

Yes, you should validate it internally anyway to make sure your model does not depend on atypical observation and does not overfit

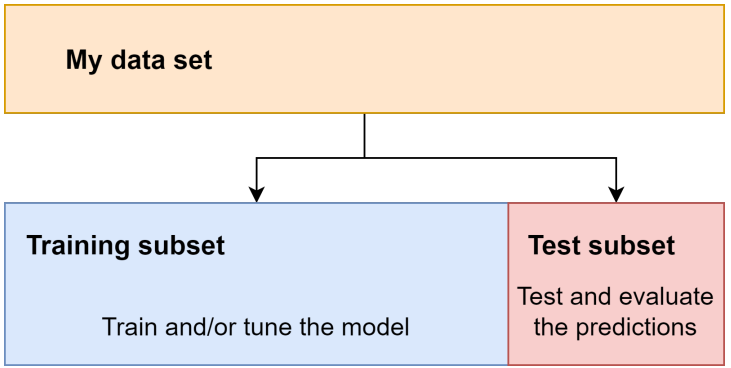
How to validate a model: internal validation



- your model has an excellent concordance with the training data
- but can you reproduce it during validation in the test data set?
- If there's an overfitting, your model is likely to fail in the test data set or internal validation

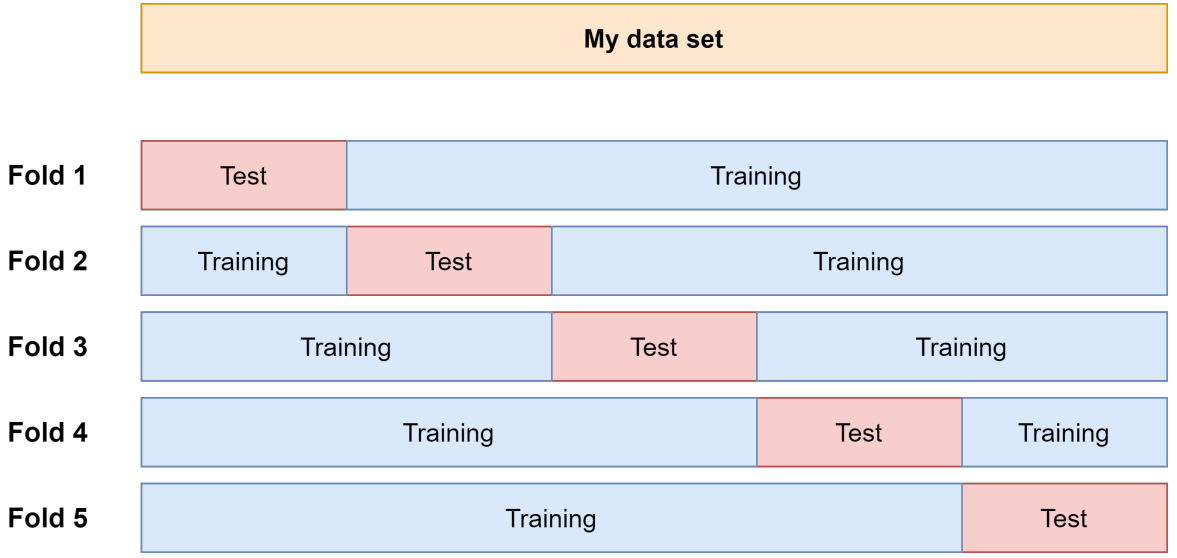
Solanes et al. 10.3389/fpsyt.2022.826111

How to validate a model: internal validation



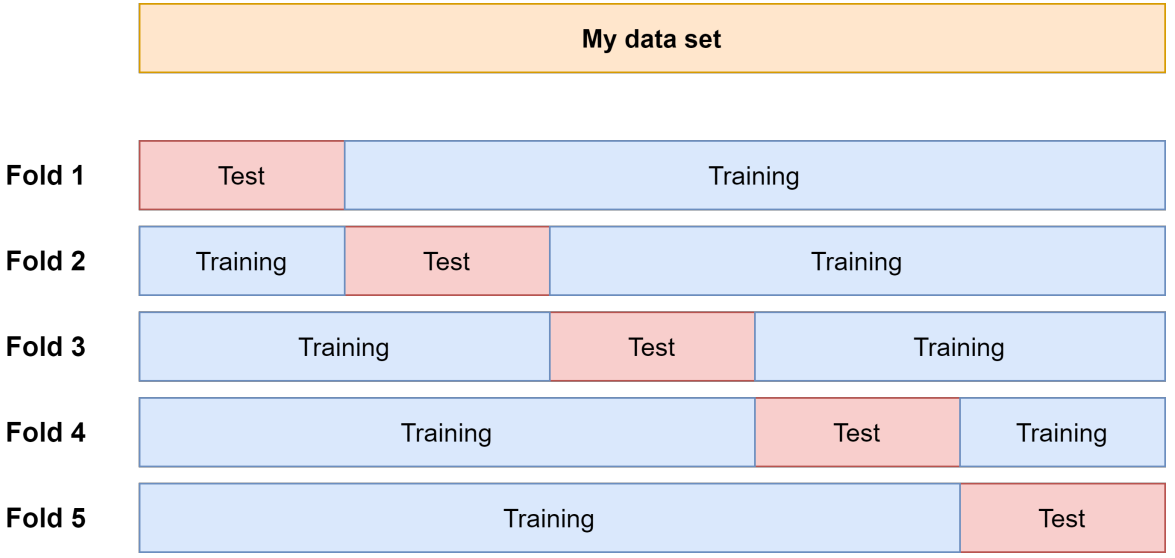
- Internal validation requires splitting your data set into a training (50 – 90%) and test subset (10 – 50%)
- The training – test split should be random and the test subset big enough to evaluate the modeling outcome
- The training – test split can be done once (hold-out) or multiple times (repeated hold-out)
- There are also alternative approaches like bootstrap or sub-setting or jackknife

How to validate a model: internal validation



- K-fold cross-validation (CV) is the ultimate form of internal validation
- In CV, each of the observation is used at least once for testing/validation of the model
- Predictions in the test subsets of CV are called of out-of-fold (OOF) predictions

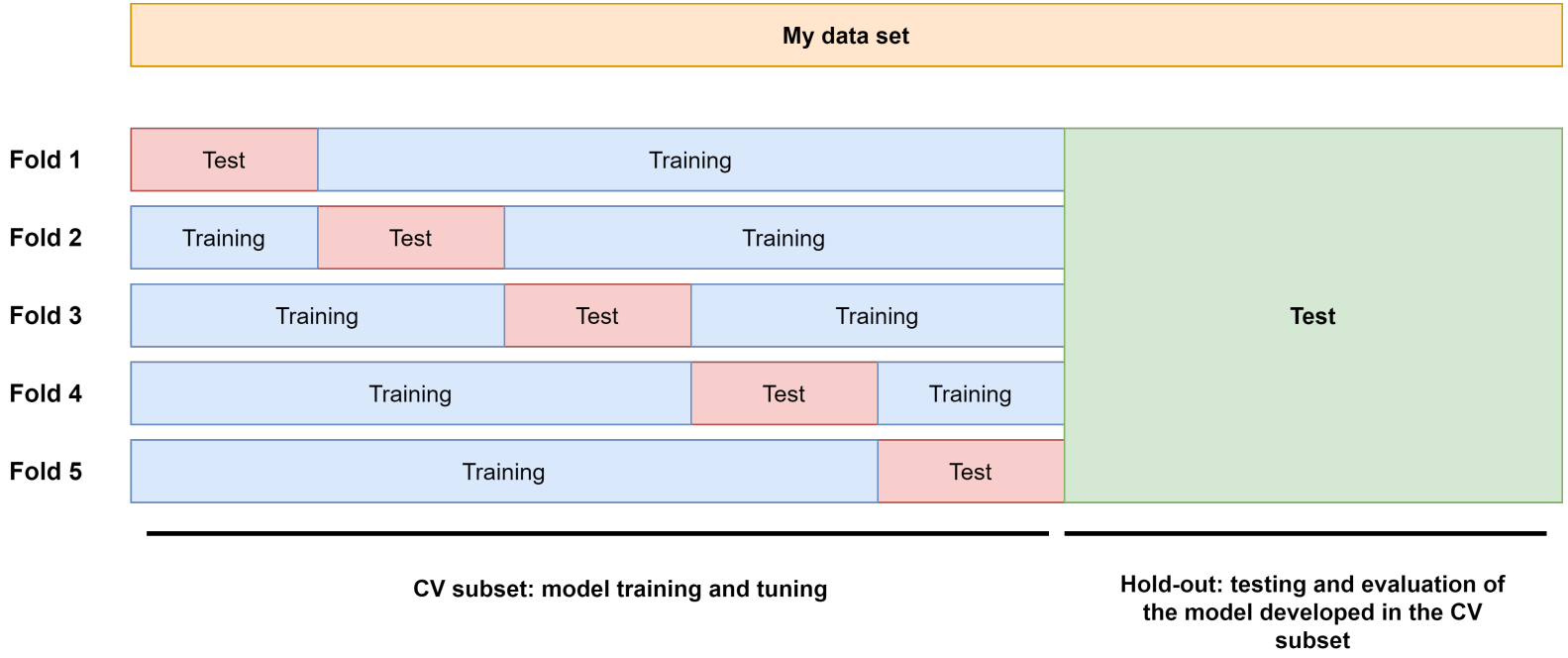
How to validate a model: internal validation



- K number of fold defines how much of the dataset will be used for testing (i.e. 10-fold CV: 10%, 5-fold CV: 20% of the entire dataset)

- CV can be done once or repeated multiple times (n-repeats k-fold CV)

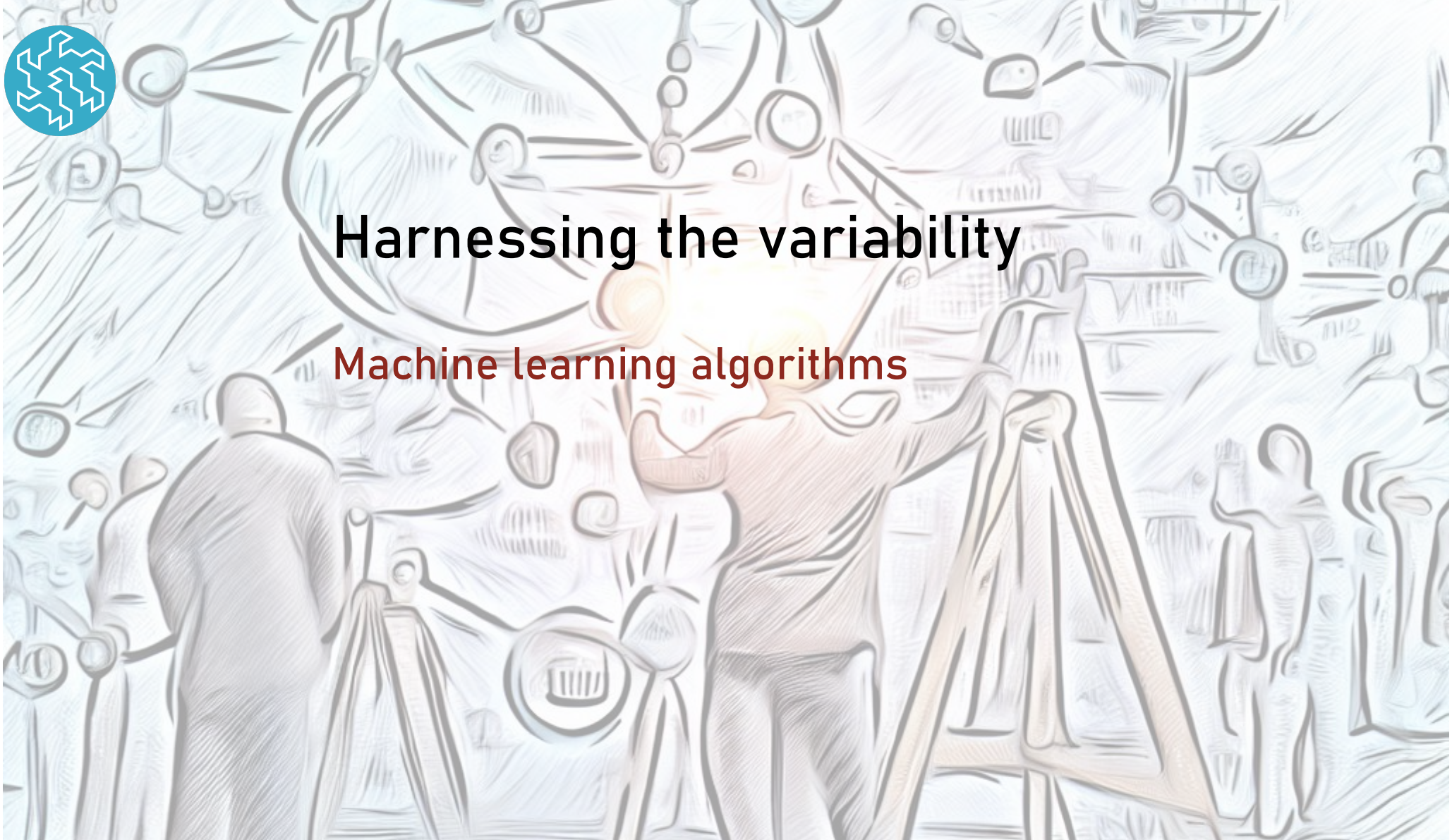
How to validate a model: internal validation



- The best practice of internal validation is to combine the CV and hold-out techniques. Why?

How to validate a model: pitfalls

- Matched data, e. g. by participant. Avoid placing observation from one patient both in the training and test subsets.
- Time series. Avoid training the model with future data and test it with past data.
- Check frequency of the outcome (e. g. lung function findings) in the test and training subsets. Why?
- Check frequency of rare explanatory features!



Harnessing the variability

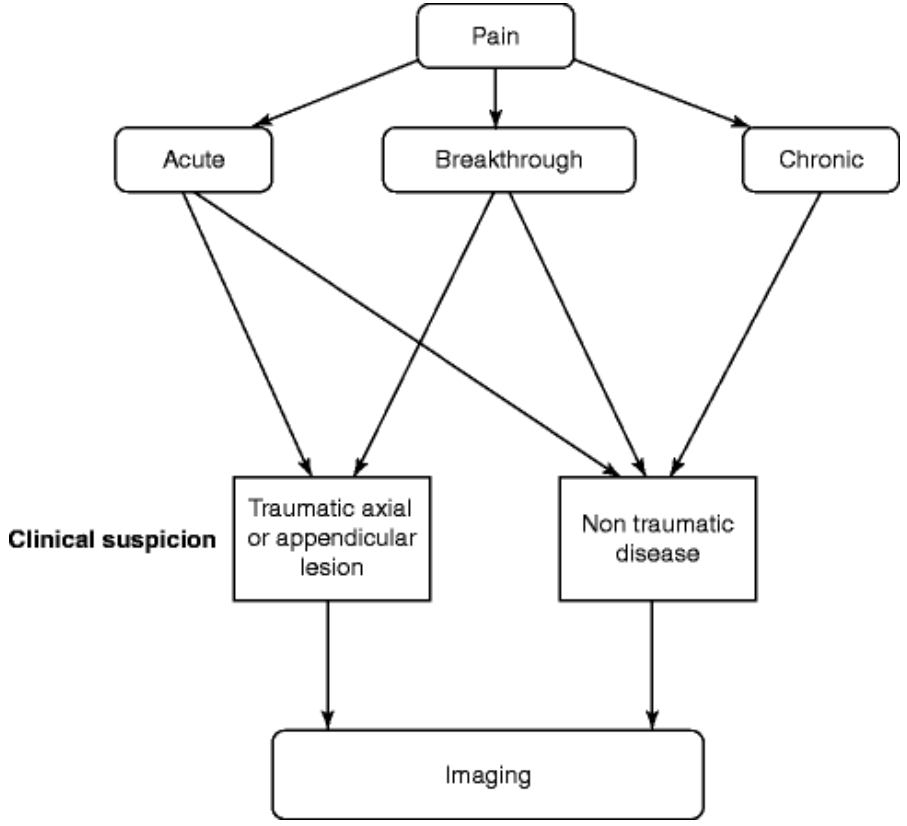
Machine learning algorithms

Machine learning algorithms

- Have you ever validated a model, i. e. played with machine learning?
- Do you know at least one machine learning algorithm?
- You will be surprised by your answers!

Machine learning algorithms

- Linear and logistic regression are common and extremely useful machine learning algorithms!
- A logistic regression model is the simplest form of a neuronal network!
- Basically every medical treatment guideline is a tree model!

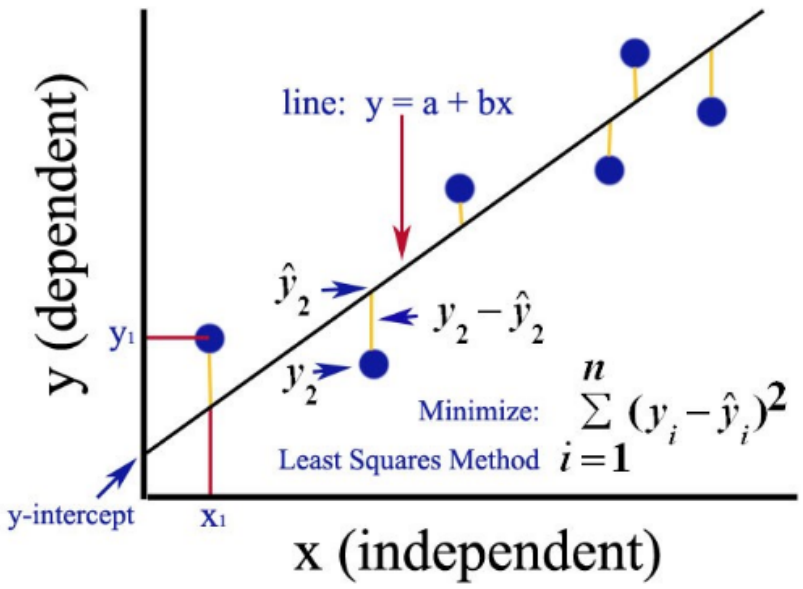


https://radiologykey.com/diagnostic-algorithm-in-the-elderly/

Machine learning algorithms: types

- If my modeling response is a numeric one: regression. Examples: size of a lesion, diffusion capacity of the lung.
- Regression models return predicted values in the same form as the modeling response
- If my modeling response is a categorical one: classification. Examples: presence of long COVID, quality classes of an image, Likert scale of quality of life
- Classification models return probability for observation belonging to classes, i.e. patient 1 has 65% probability of having cancer
- Deciding whether your model should tackle regression of classification problems can be extremely tricky!

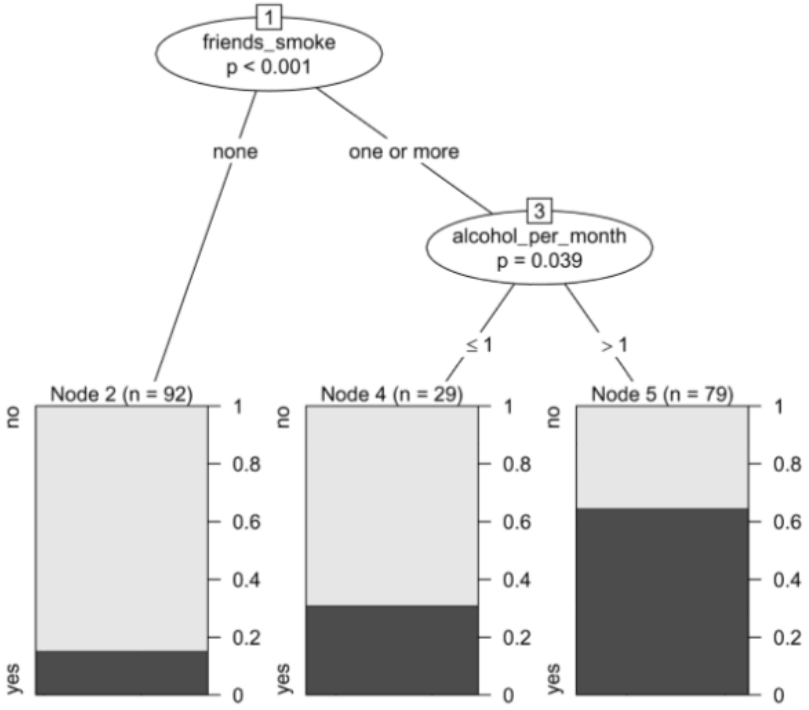
Machine learning algorithms: linear models



https://medium.com/analytics-vidhya/ordinary-least-square-ols-method-for-linear-regression-ef8ca10aadfc

- A straight linear trend or a related curve (e.g. polynomial, log, logit) is fitted to the data
- The optimal fit criterion (loss function) is usually the least square
- A robust, simple algorithm, easy to interpret (why?)
- Has specific assumptions for the response and explanatory variables!
- High risk of overfitting

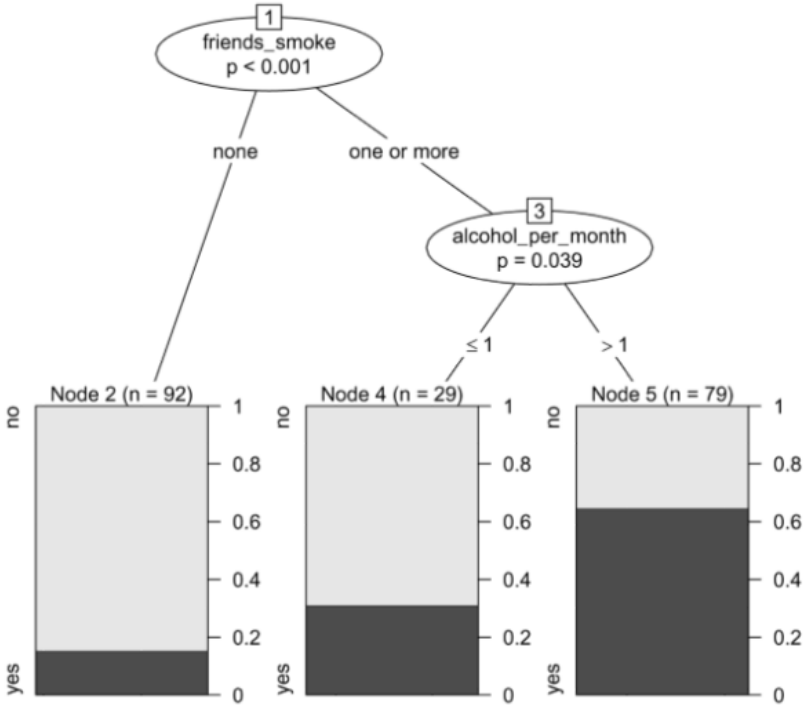
Machine learning algorithms: tree models



Strobl C, Malley J, Tutz G. doi:10.1037/A0016973

- Observations are partitioned – assigned to the tree nodes – based on differences in explanatory variables
- New nodes are added until no further improvement in differences between the terminal nodes (leaves), or the process is stopped by a user-defined criterion (e.g. minimal number of observations per node, maximal node number)
- There are multiple possible splitting rules for the tree nodes, e.g. based on information entropy, difference in variance or Gini index, statistical tests

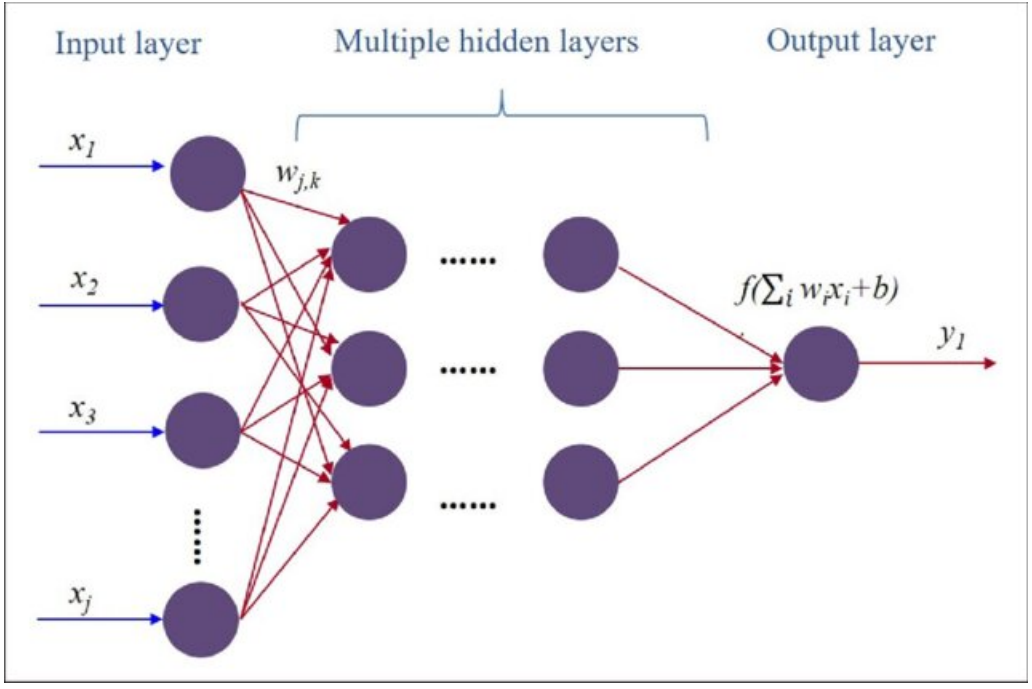
Machine learning algorithms: tree models



Strobl C, Malley J, Tutz G. doi:10.1037/A0016973

- Trees are extremely prone to overfitting
- Algorithms employing single trees are not common anymore in modern machine learning
- The most powerful tree algorithms, random forest and gradient boosted machines (GBM) use multiple simple trees (up to several thousands)
- Such ensembles of trees can fit virtually all types of data and are moderately resistant to overfitting (why?)

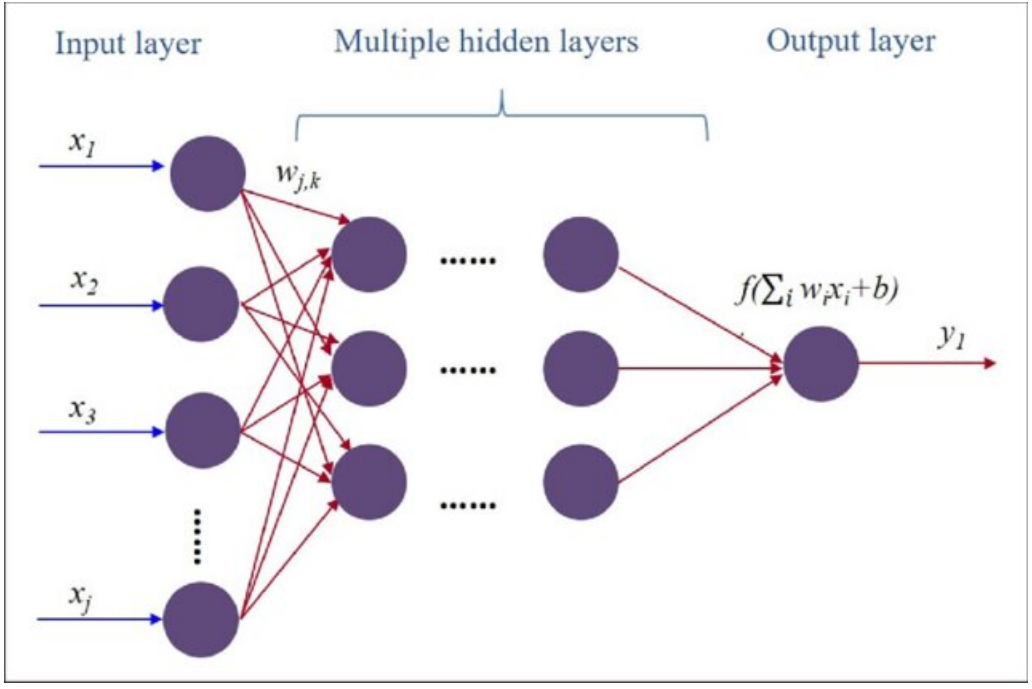
Machine learning algorithms: neuronal networks



Wong ZSY et al. doi:10.1177/1460458219889798

- Explanatory variables are assigned weights (w) by the input layer neurons based on some mathematical function
- The weighted inputs are subsequently integrated by a hidden layer of neurons (i. e. assigned new weights via an activation function) and passed to another higher layer or the output layer which provides modeling results
- This process may be repeated multiple times by adjusting weights to improve accuracy at the output

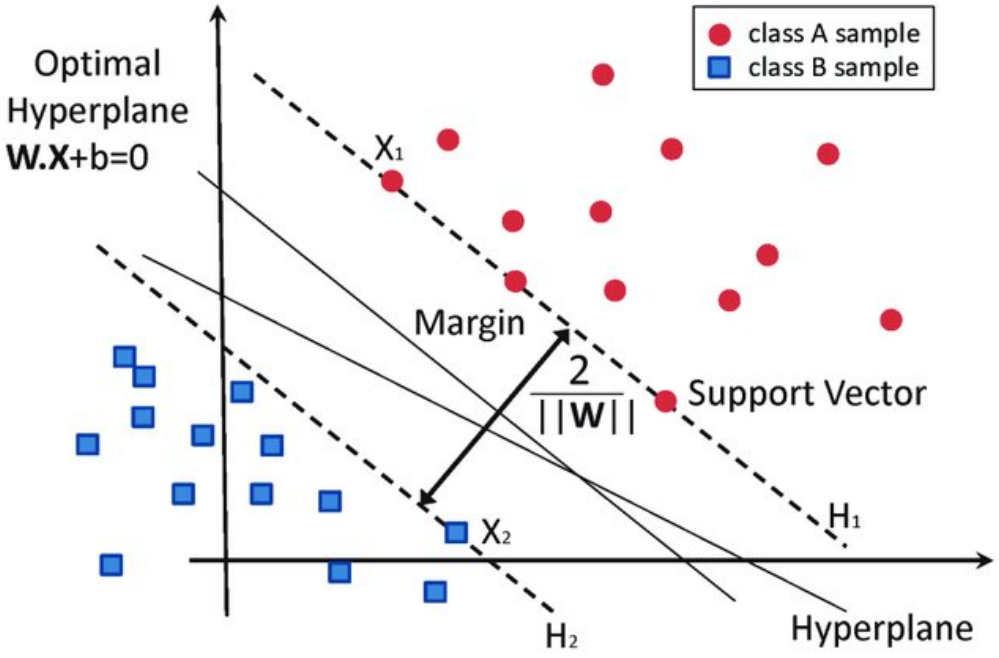
Machine learning algorithms: neuronal networks



Wong ZSY et al. doi:10.1177/1460458219889798

- Neuronal networks can handle all types of data and have usually no assumptions for the response and explanatory data
- The models are prone to overfitting and virtually not interpretable by a human!

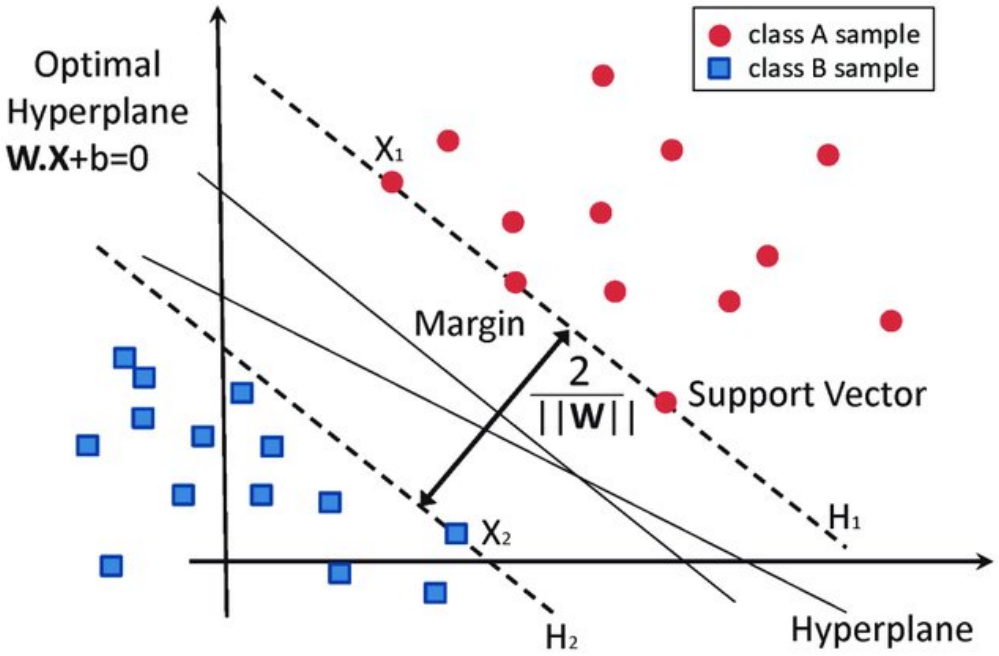
Machine learning algorithms: support vector machines



García-Gonzalo E, doi:10.3390/MA9070531

- In its canonical form, support vector machines (SVM) try to find an optimal division line (in n-dimensions: hyperplane) between classes of the outcome
- Definition of the hyperplane requires only few selected 'marginal' points of the outcome classes: so called support vectors
- Using of only few support vectors instead of all observations has multiple advantages – can you imagine which ones?

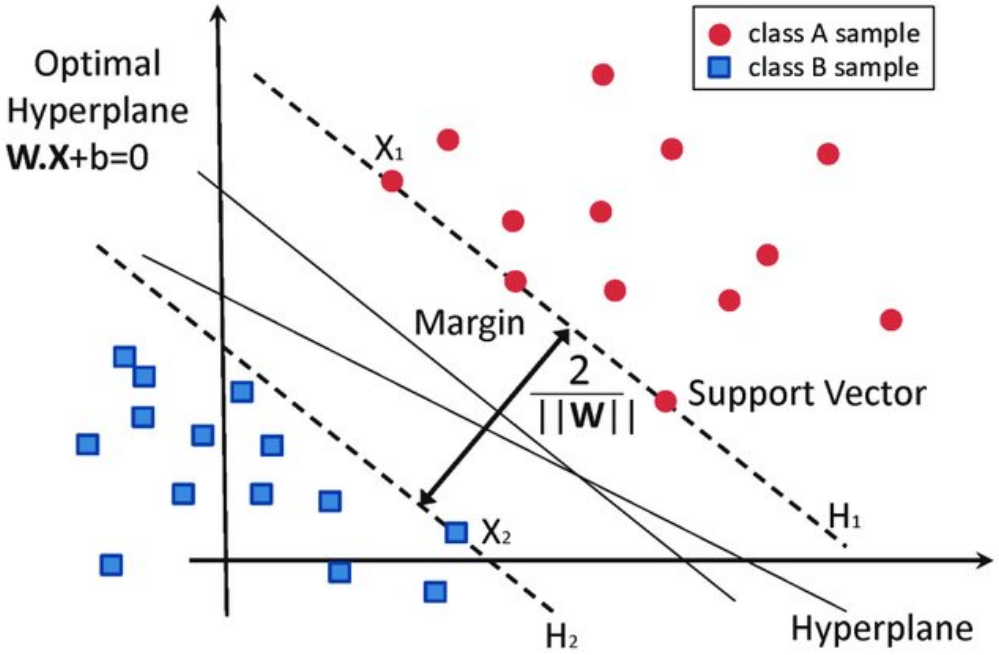
Machine learning algorithms: support vector machines



García-Gonzalo E, doi:10.3390/MA9070531

- Still there are observations that fall between the margins defined by support vectors or are even on the bad side of the hyperplane
- We are applying a penalty to such observations - so called cost (C) – that defines how much such problematic observations contribute to the model error
- Low C: wide margin between the support vectors but possibly low accuracy
- High C: narrow margin, high accuracy but a risk of overfitting

Machine learning algorithms: support vector machines



García-Gonzalo E, doi:10.3390/MA9070531

- SVM are fast and powerful learners
- Moderate risk of overfitting
- May handle multiple types of data and challenging cases of classification with so called kernels

Better than randomness?

Evaluating a model

Evaluating a model: accuracy

$$Accuracy = \frac{1}{N} \times \sum_{i=0}^{N} (prediction_i = observed_i)$$

- Overall accuracy is simply the ratio of correct predictions by the model to all predictions
- Accuracy can range from 0 (bad model) to 1 (excellent model)
- What's problematic about it? What if observations e.g. with insufficient lung function are rare?

Evaluating a model: accuracy

- An example: a model predicting Christmas given a date

https://xkcd.com/

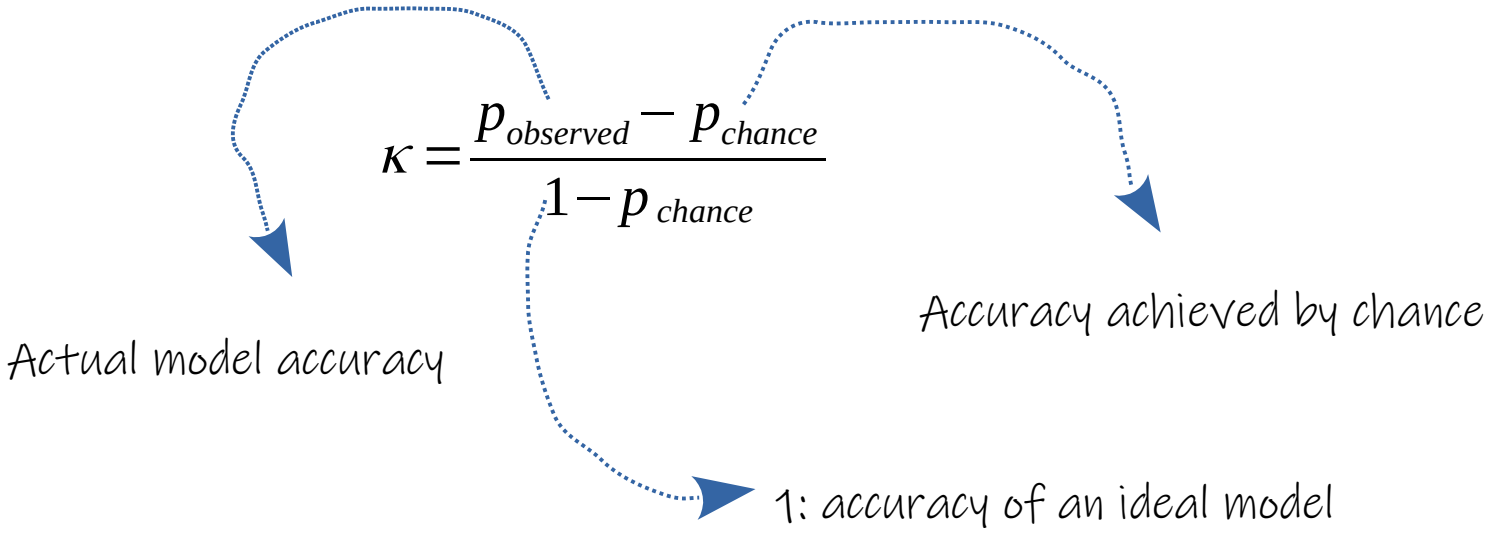
https://medium.com/eliiza-ai/modelling-rare-events-c169cb081d8b

- "no Christmas" for any day of the year: accurate but useless!

Evaluating a model: Cohen's kappa

- For rare or extremely frequent events its easy to build meaningless models predicting the outcome simply by chance

- To measure how my model outperforms such dummy models, the inter-rater reliability κ statistic was introduced



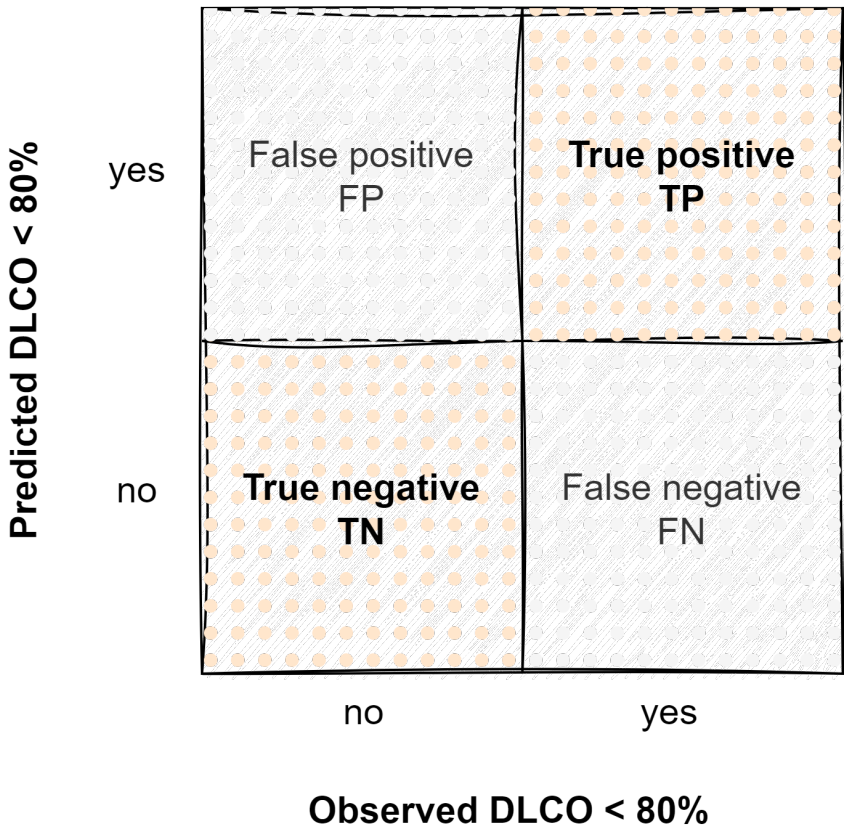
Evaluating a model: Cohen's kappa

$$\kappa = \frac{p_{observed} - p_{chance}}{1 - p_{chance}}$$

- Dummy Christmas model: κ = 0 but accuracy = 0.9973
- Every week there's one Christmas day: κ = 0.03 and accuracy = 0.86
- Correct Christmas model: κ = 1.0 and accuracy = 1.0

Evaluating a model: sensitivity and specificity

Confusion matrix



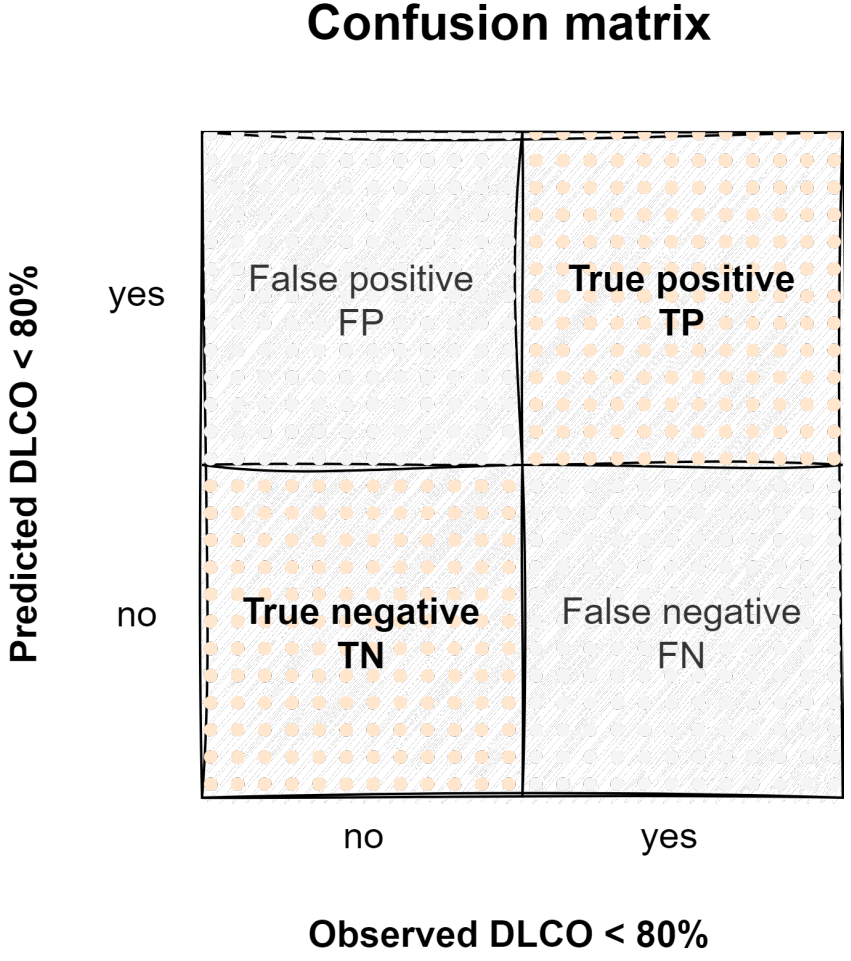
$$Se = \frac{TP}{TP + FN}$$

$$Sp = \frac{TN}{TN + FP}$$

$$J = Se + Sp - 1$$

Possible problems?

Evaluating a model: sensitivity and specificity



- Dummy Christmas model:
  - Se = 0
  - Sp ≈ 1
  - J = 0
- Every week there's one Christmas day:
  - Se ≈ 1
  - Sp = 0.8699
  - J = 0.8699
- Correct Christmas model:
  - Se = 1
  - Sp = 1
  - J = 1

Evaluating a model: Brier score

- So far, we had a look at statistics which works with the class assignment
- An example: patient with 51% chance of insufficient lung function returned by a model is classified as insufficient lung function
- Which prediction would you trust more? A model predicting 51% or a model predicting 80% chance of lung function impairment?

Evaluating a model: Brier score

- **For an accurate model, higher prediction probabilities should result in higher chance of correct predictions. Brier score follows this logic!**

$$BS = \frac{1}{N} \times \sum_{i=0}^{N} (p_i - outcome_i)^2$$

Model-predicted probability of impaired lung function, e.g. 0.9, for the given observation

Outcome:
0: the observation has normal lung function
1: the observation has abnormal lung function

Evaluating a model: Brier score

$$BS = \frac{1}{N} \times \sum_{i=0}^{N} (p_i - outcome_i)^2$$

Model-predicted probability of impaired lung function, e.g. 0.9, for the given observation

Outcome:
0: the observation has normal lung function
1: the observation has abnormal lung function

- **The model predicts insufficient lung function with 90% probability for the patient X but they have a normal value → high Brier score (square term = 0.81)**
- **The model predicts insufficient lung function with 51% probability but the patient X has normal values → a bit better Brier score (square term = 0.26)**

Evaluating a model: Brier score

$$BS = \frac{1}{N} \times \sum_{i=0}^{N} (p_i - outcome_i)^2$$

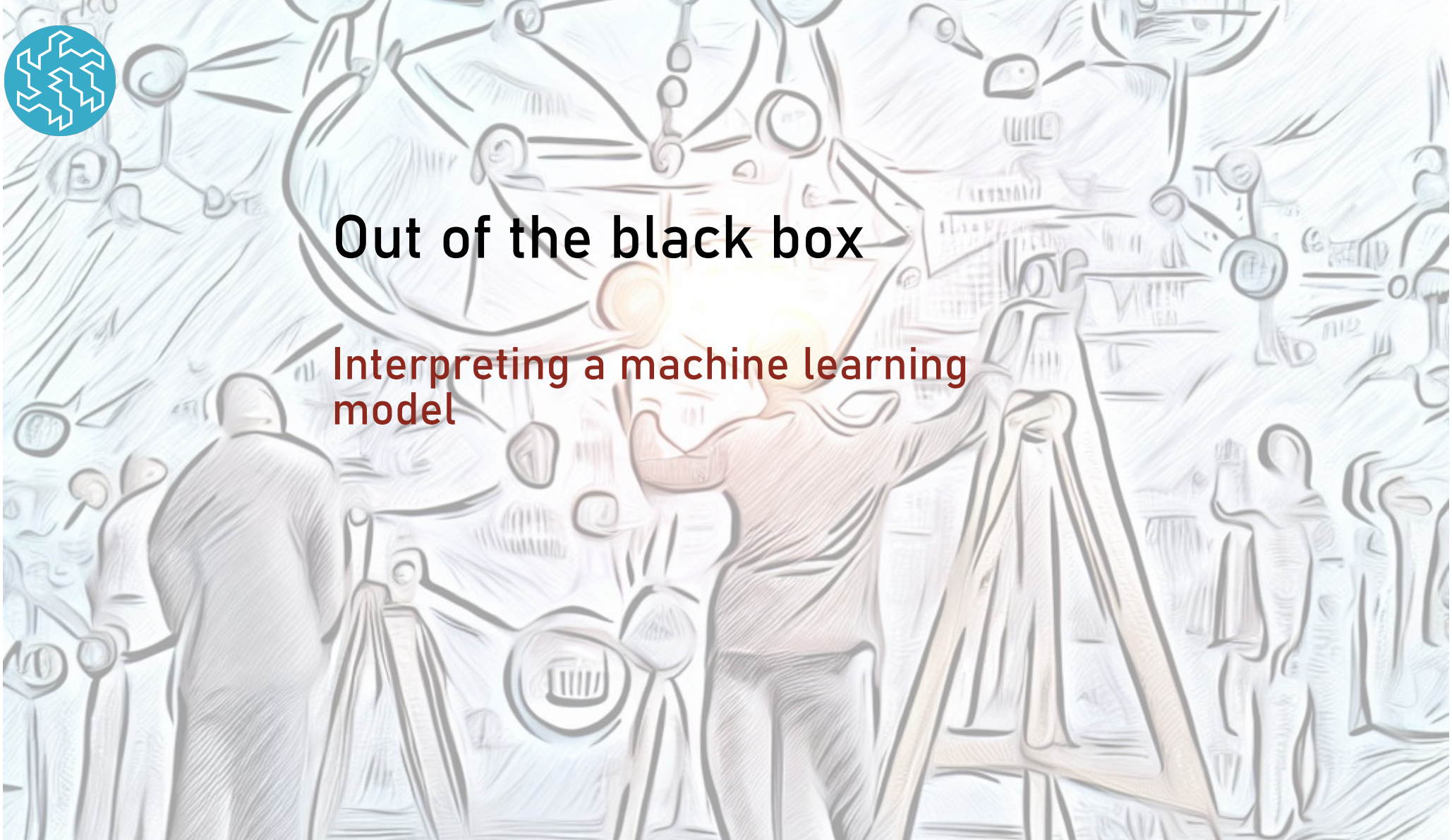
Model-predicted probability of impaired lung function, e.g. 0.9, for the given observation

Outcome:
0: the observation has normal lung function
1: the observation has abnormal lung function

- **The model predicts insufficient lung function with 90% probability for the patient X with insufficiency → excellent Brier score (square term = 0.01)**
- **The model predicts insufficient lung function with 51% probability for the patient X with insufficiency → worse Brier score (square term = 0.26)**

Evaluating a model: the CovILD data set

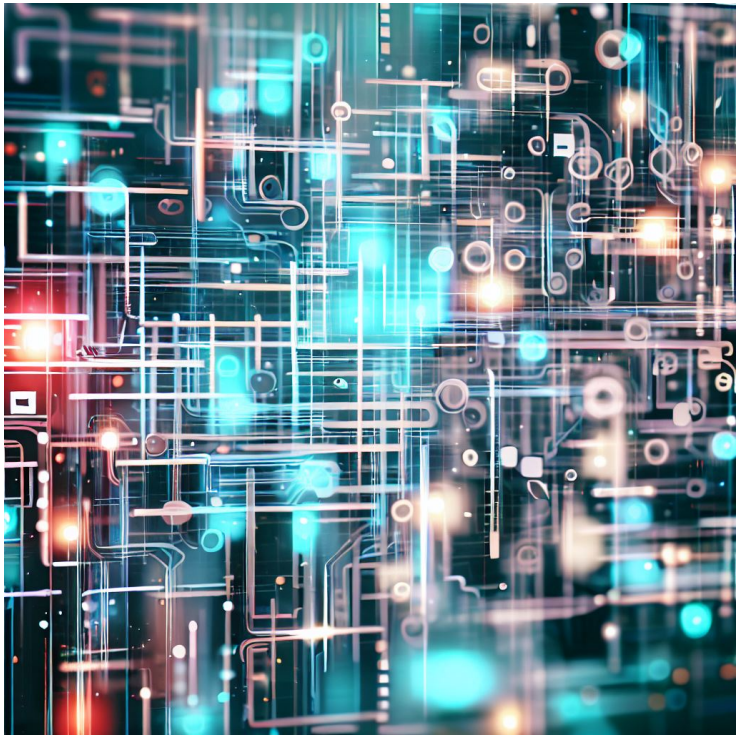
- We know now the basic statistics used for evaluation of a binary classification model:
  - accuracy and κ: how good my model's predictions correspond with the reality
  - sensitivity, specificity and J: how good my model detects the outcome of interest (e.g. DLCO < 80%, FEV1 < 80%)
  - Brier score: how the predicted event probability corresponds with the true probability of an outcome



Out of the black box

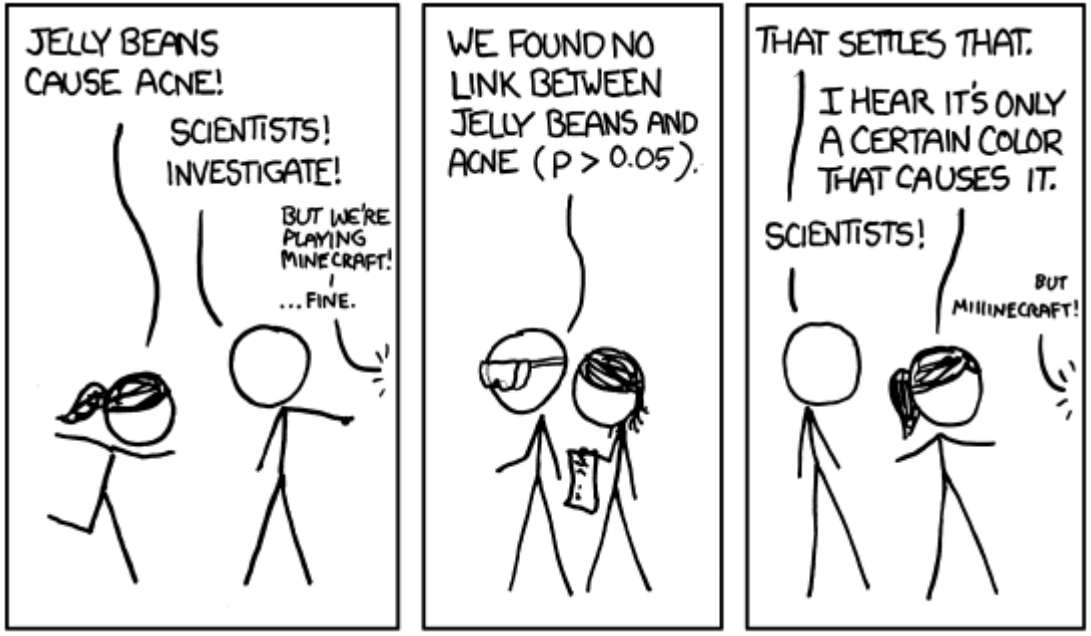
Interpreting a machine learning model

Interpreting a model



- Many people regard machine learning models as ‘black boxes’
- It does not refer e. g. to linear models or simple trees (why?)
- But definitively applies to random forests, neuronal networks or Chat GPT
- In such cases, we need robust tools to check how much an explanatory variable contributes to the model fit
- In AI slang, such variable contribution to the model fit is called variable importance

Interpreting a model



https://xkcd.com/882/

- Does it remind you of a study looking for disease markers?

Interpreting a model: SHAP

- In a multi-parameter model, single variables interact with each other in hardly predictable ways

- Often, it makes no sense to look at them one-by-one (univariable analysis) and hope for significant effects with $p < 0.05$ – hopefully after multiple-testing adjustment!

- Examples of such (too) simplistic approach:

  - Lionel Messi is a significant predictor of good scoring in the Champions League

  - Green jelly beans are significantly associated with acne

Interpreting a model: SHAP

- Instead it's more sound to check how the model performs if we change or remove a variable or few variables
- Examples of such more sophisticated approach:
  - Performance of a team in the Champions league with and without Lionel Messi
  - Acne in consumers of jelly bean mixes with and without green ones
- This is the logic of SHAP (Shapley Additive Explanations)

Interpreting a model: SHAP

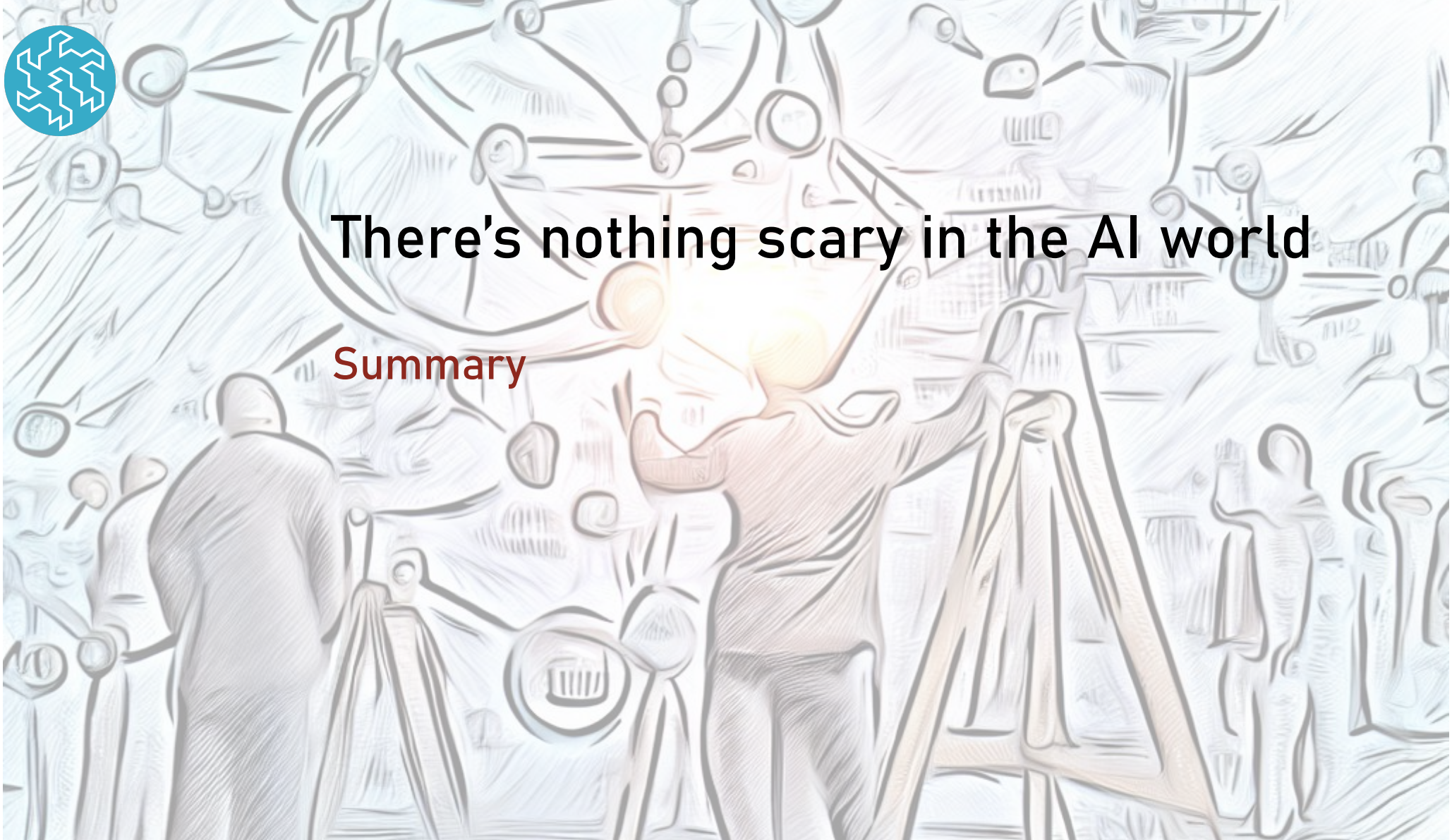
$$\varphi_i(v) = \frac{1}{n} \sum_{S \subseteq N \setminus \{i\}} \binom{n-1}{|S|}^{-1} (v(S \cup \{i\}) - v(S))$$

Importance of the variable i, e.g. importance of Messi for the team performance

Scaling coefficient: roughly matches without Messi to all matches

Performance of the model with all variable sets containing the variable i: e.g. team performance in all matches with Messi

Performance of the model with all variable sets without the variable i: e.g. team performance in all matches without Messi



There's nothing scary in the AI world

Summary

Summary: take home messages

- Building a model is not a purely analytic process: it takes a question, analytically sound hypothesis, validation, evaluation and interpretation of the results
- Validation of a model is the key. Internal validation is obligatory. External validation is paramount
- Always challenge your model (or models of others as a reviewer) and request a thorough evaluation in the training and validation data sets
- Accuracy, sensitivity and specificity alone are not sufficient to evaluate model performance

There's nothing scary in the AI world

Thanks for your attention!!!