Data Analytics in Cyber Security
CT115-3-M (Version E)

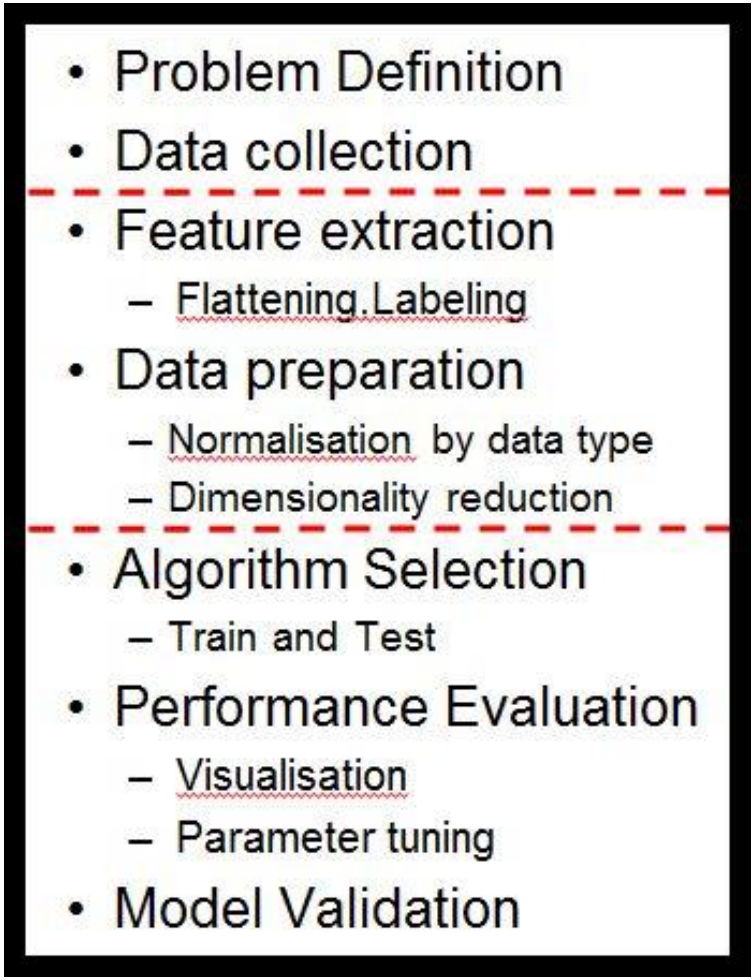
Model Selection

TOPIC LEARNING OUTCOMES

At the end of this topic, you should be able to:

1. Understand the essential steps in model validation.
2. Understand different model comparison and selection techniques.

Machine Learning Pipeline



- Problem Definition
- Data collection
- Feature extraction
  - Flattening.Labeling
- Data preparation
  - Normalisation by data type
  - Dimensionality reduction
- Algorithm Selection
  - Train and Test
- Performance Evaluation
  - Visualisation
  - Parameter tuning
- Model Validation

Iteration:
Model Validation

- The model validation phase should lead to a new round of performance evaluation, and the set of candidate algorithms will gradually be reduced until one algorithm (or one ensemble of algorithms) is selected.

Contents & Structure

- Model selection
- Comparing models
- Case study

The Final Model

- The final model is the one prepared on the entire training dataset, once we have chosen an algorithm and configuration.
- In statistics, more samples give us higher confidence that our model is representative of the population.
- One strategy to gather a population of performance measures is to use several runs of k-fold cross validation.
- Or, build an ensemble of models, each trained with a different random number seed.
  - Using a simple voting ensemble, each model makes a prediction and the mean of all predictions is reported as the final prediction.

random_state

- Most classifiers make use of randomness during the process of constructing a model from the training data
- This has the effect of fitting a different model each time same algorithm is run on the same data.
- In turn, the slightly different models have different performance when evaluated on the same test dataset.
- The proper name for this difference or random behavior within a range is *stochastic*.

➢ **Expect the performance to be a range and not a single value.**

➢ **Expect there to be a range of models to choose from and not a single model.**

random_state

- Random numbers are generated in software using a pseudo random number generator. It’s a simple math function that generates a sequence of numbers that are random enough for most applications.
- This math function is deterministic. If it uses the same starting point called a seed number, it will give the same sequence of random numbers.
- We can get reproducible results by fixing the random number generator’s seed before each model we construct.
- We do this by setting the *random_state* hyperparameter in the call to the classifier.

Report the Uncertainty

Then report summary statistics on this population

- Report the mean and standard deviation of performance, and the highest and lowest performance observed (the range)
  - Create a figure with a “box and whisker” plot for each algorithm’s sample of results.
  - The box shows the middle 50 percent of the data, the orange line in the middle of each box shows the median of the sample, and the green triangle in each box shows the mean of the sample.

Comparing Models

- Differences in performance measures might easily turn out to be merely by chance, not because one model systematically predicts ***better*** than the other.
- Statistical significance tests can determine if the difference between one population of result measures is significantly different from a second population of results.
  - many methods exist that apply statistics to the selection of Machine Learning models.
- The most robust way to do such comparisons is called *paired designs*, which compare the performance of two models (or algorithms) on the same data.

Statistical Hypothesis Testing

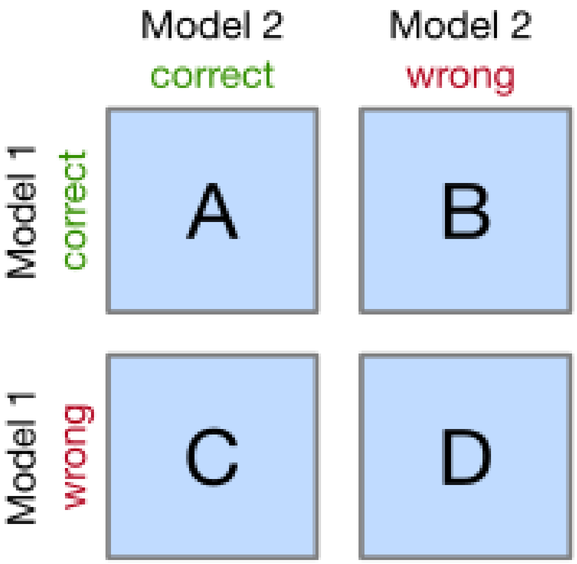
- A statistical hypothesis test quantifies how plausible it is to say two data samples have not been drawn from the same distribution.

- That describes the *null hypothesis*. We can test this null hypothesis by applying some statistical calculations.

- If the test result infers insufficient evidence to reject the null hypothesis, then any observed difference in the model scores is a happened by chance.

- If the test result infers sufficient evidence to reject the null hypothesis, then any observed difference in model scores is real.

McNemar's Test



	Model 2 correct	Model 2 wrong
Model 1 correct	A	B
Model 1 wrong	C	D

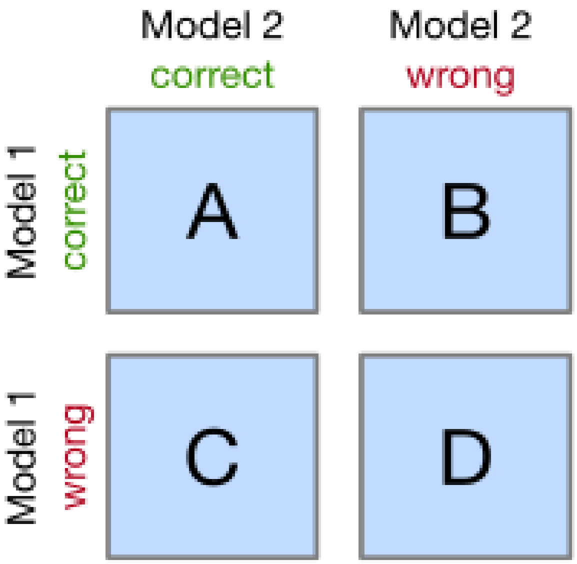
McNemar's test compares the predictions of two models based on a version of a 2x2 confusion matrix

Cells B and C (the off-diagonal entries) tell us how the models differ

Null hypothesis: the performance of the two models is the same

1. Set a significance threshold, normally $\alpha = 0.05$
2. Compute the p-value, the probability of observing the given empirical (or a larger) chi-squared value
3. If the p-value is lower than our chosen significance level, we can reject the null hypothesis that the two model's performances are equal

McNemar's Test



	Model 2 correct	Model 2 wrong
Model 1 correct	A	B
Model 1 wrong	C	D

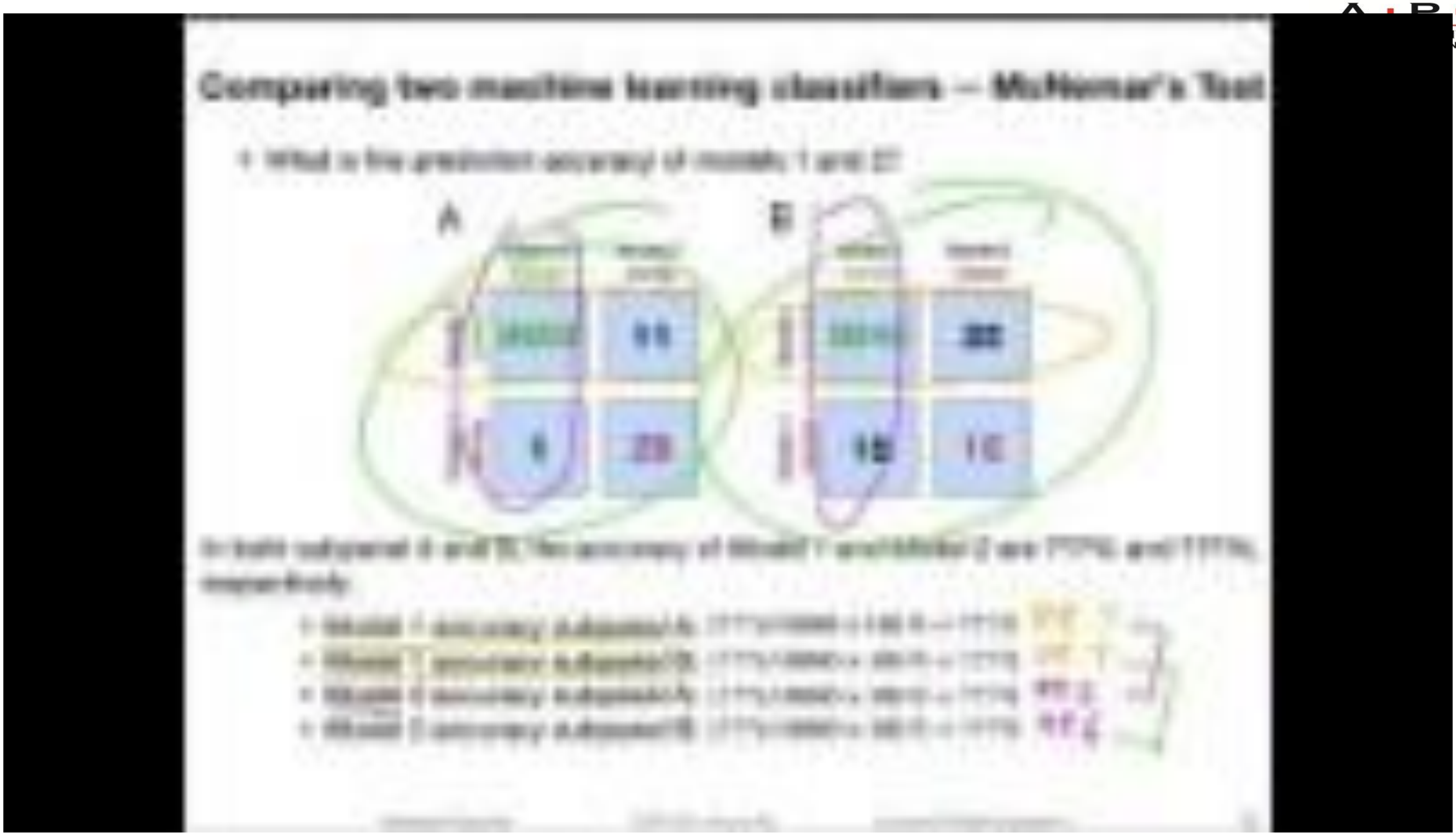
McNemar's test compares the predictions of two models based on a version of a 2x2 confusion matrix

Cells B and C (the off-diagonal entries) tell us how the models differ

Null hypothesis: the performance of the two models is the same

McNemar's test is unique because it uses the actual predictions from the models – these other tests use a metric derived from the confusion matrix, like accuracy, precision, recall, F1, roc_auc, … etc.

McNemar's Test



Comparing two machine learning classifiers – McNemar's Test

- What is the prediction accuracy of models 1 and 2?

A

	Model 2 correct	Model 2 wrong
Model 1 correct	9959	11
Model 1 wrong	1	29

B

	Model 2 correct	Model 2 wrong
Model 1 correct	9945	25
Model 1 wrong	15	15

In both subpanel A and B, the accuracy of Model 1 and Model 2 are [illegible] and [illegible], respectively.

- Model 1 accuracy subpanel A: [illegible]
- Model 1 accuracy subpanel B: [illegible]
- Model 2 accuracy subpanel A: [illegible]
- Model 2 accuracy subpanel B: [illegible]

https://youtu.be/nzznkiW8ulk?t=143

5x2cv paired (t or f) test

- **5x2cv paired t-test:** Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. Neural computation, 10(7), 1895-1923
- Low false positive rate, slightly more powerful than McNemar
- Recommended if computational efficiency (runtime) is not an issue (10 times more computations than McNemar)
- https://rasbt.github.io/mlxtend/user_guide/evaluate/paired_ttest_5x2cv/

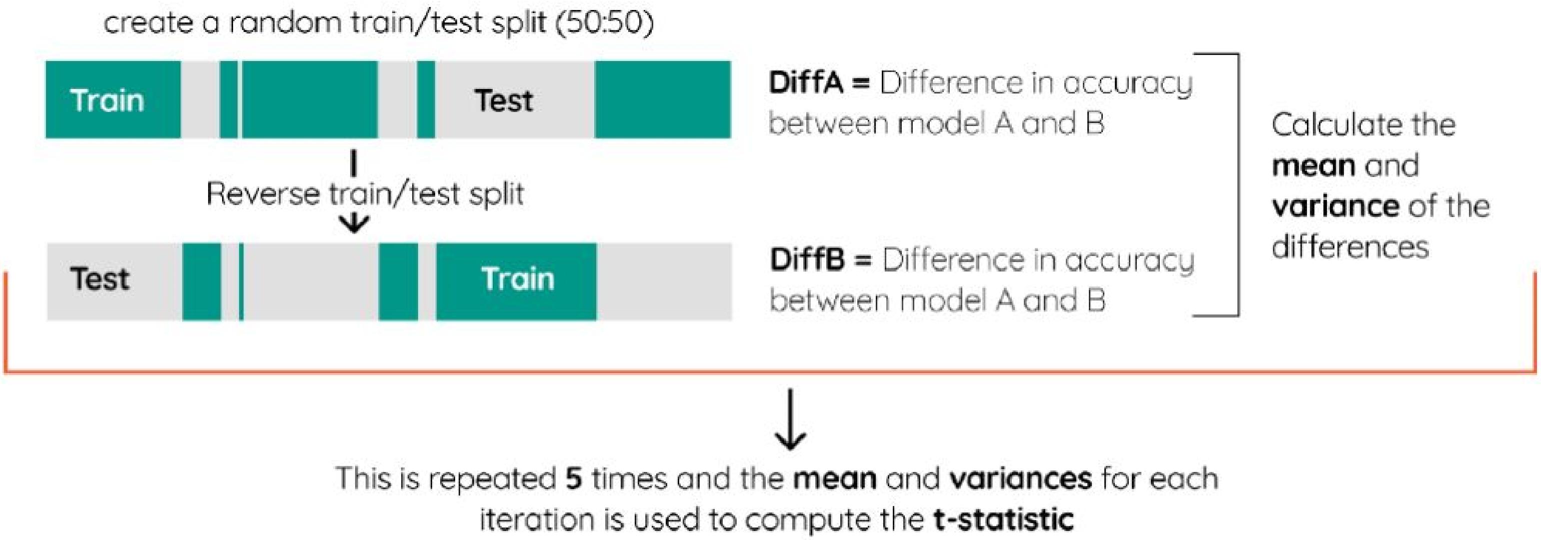
- **Combined 5×2cv f-test:** Alpaydin, Ethem (1999). "Combined 5×2cv f-test for comparing supervised classification learning algorithms." Neural computation 11(8), 1885-1892
- More robust than **5x2cv paired t-test**
- https://rasbt.github.io/mlxtend/user_guide/evaluate/combined_ftest_5x2cv/

The difference is the distribution used to calculate the p-value:

t-distribution
or
f-distribution

(McNiemar uses chi-squared distribution)

5x2CV paired (t or f) test procedure


create a random train/test split (50:50)
Train
Test
Reverse train/test split
Test
Train
DiffA = Difference in accuracy between model A and B
DiffB = Difference in accuracy between model A and B
Calculate the **mean** and **variance** of the differences
This is repeated **5** times and the **mean** and **variances** for each iteration is used to compute the **t-statistic**

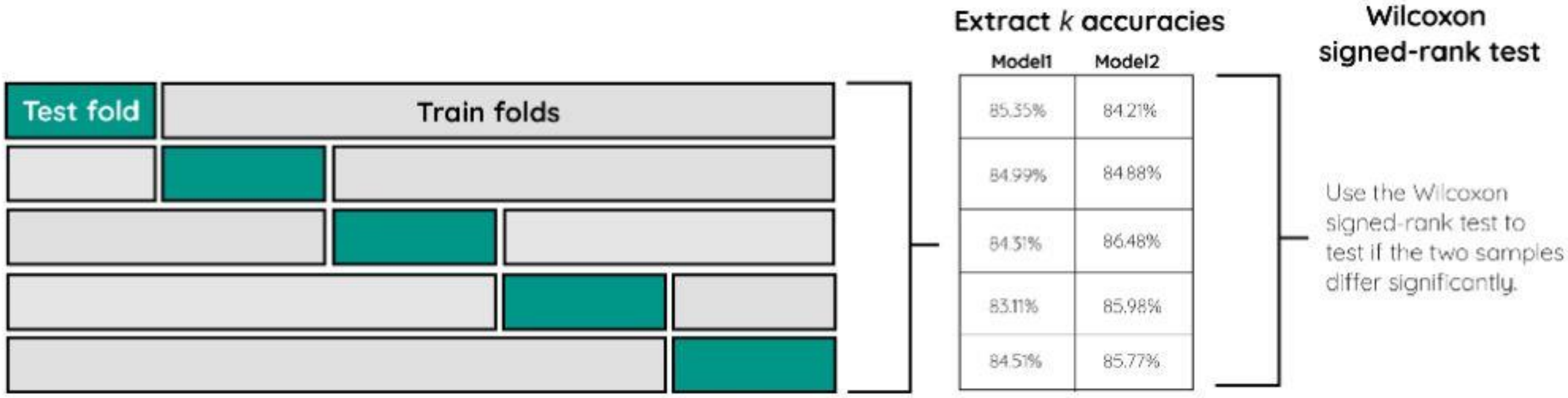
Let’s say we have two classifiers, A and B. We randomly split the data in 50% training and 50% test. Then, we train each model on the training data and compute the difference in accuracy between the models from the test set, called DiffA. Then, the training and test splits are reversed and the difference is calculated again in DiffB. This is repeated five times, after which the mean and variance of the differences is computed.

Wilcoxon Signed-Rank Test

- The Wilcoxon signed-rank test is the non-parametric version of the paired Student’s t-test. It can be used when the sample size is small and the data does not follow a normal distribution.
- We can apply this significance test for comparing two Machine Learning models.

1. Create a set of accuracy scores for each model, using k-fold cross validation or several runs with different random number seeds.
    - This gives us two samples, one for each model.
2. Then, we use the Wilcoxon signed-rank test to see if the two accuracy score samples differ significantly from each other.
    - If they do, then one is more accurate than the other.

Wilcoxon Signed-Rank Test



Extract *k* accuracies

Model1	Model2
85.35%	84.21%
84.99%	84.88%
84.31%	86.48%
83.11%	85.98%
84.51%	85.77%

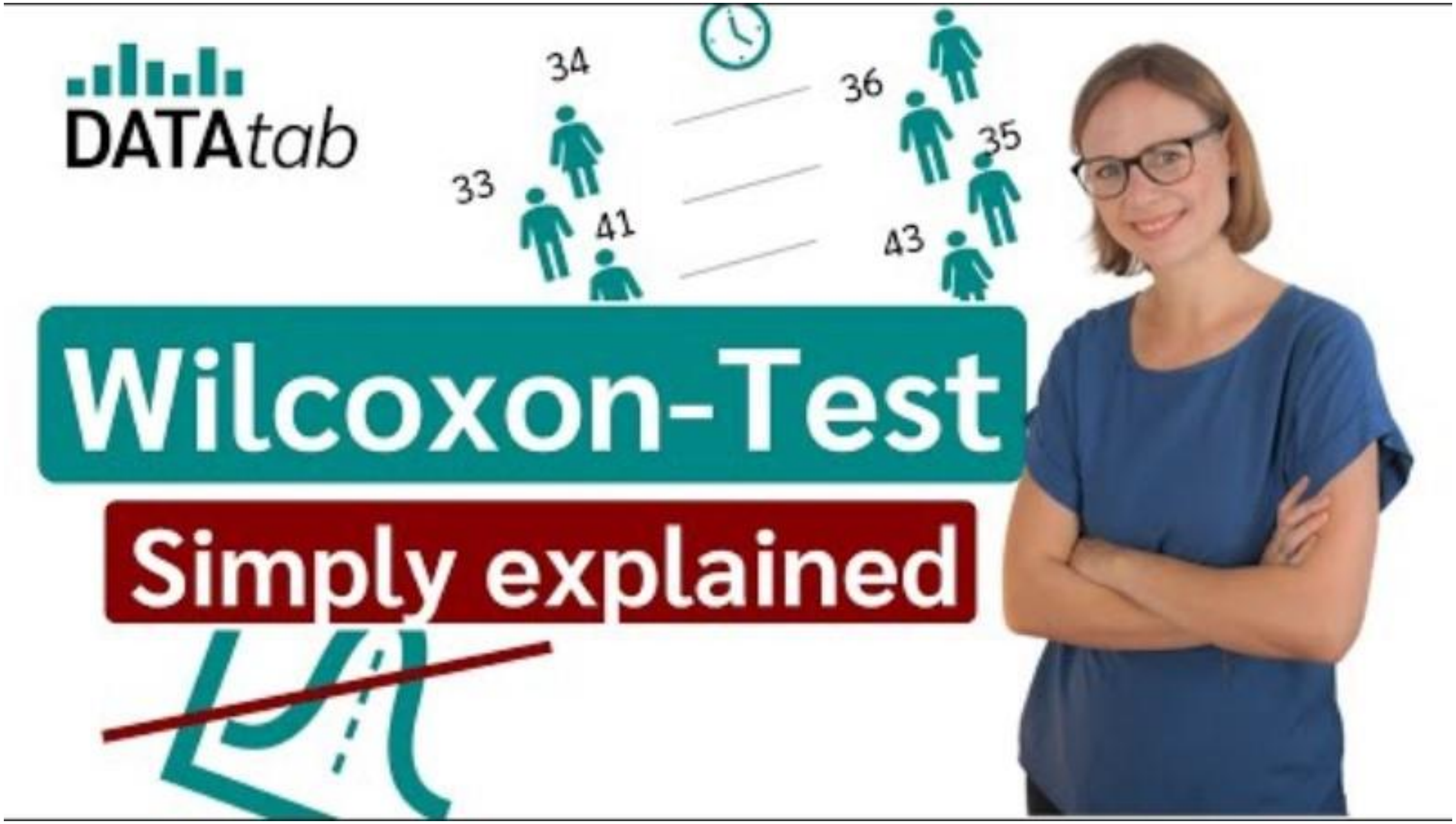
Wilcoxon signed-rank test

Use the Wilcoxon signed-rank test to test if the two samples differ significantly.

Wilcoxon signed-rank test procedure for comparing two Machine Learning models

- The result will be a p-value. If that value is lower than 0.05 we can reject the null hypothesis that there are no significant differences between the models.
- ***NOTE: It is important to <u>keep the same folds between the models</u> to <u>make sure the samples are drawn from the same population</u>. This is achieved by simply using the same random_state value.***

Wilcoxon Signed-Rank Test



https://youtu.be/NZsL2eDQiDQ

Review Questions

- What are the essential steps in model validation?
- What are the different model comparison and selection techniques?

Summary / Recap of Main Points

1. Understand the essential steps in model validation.
2. Understand different model comparison and selection techniques.