Hewlett Packard Enterprise
Data Science Institute
UNIVERSITY OF HOUSTON

HPE DSI 311
Introduction to Machine Learning

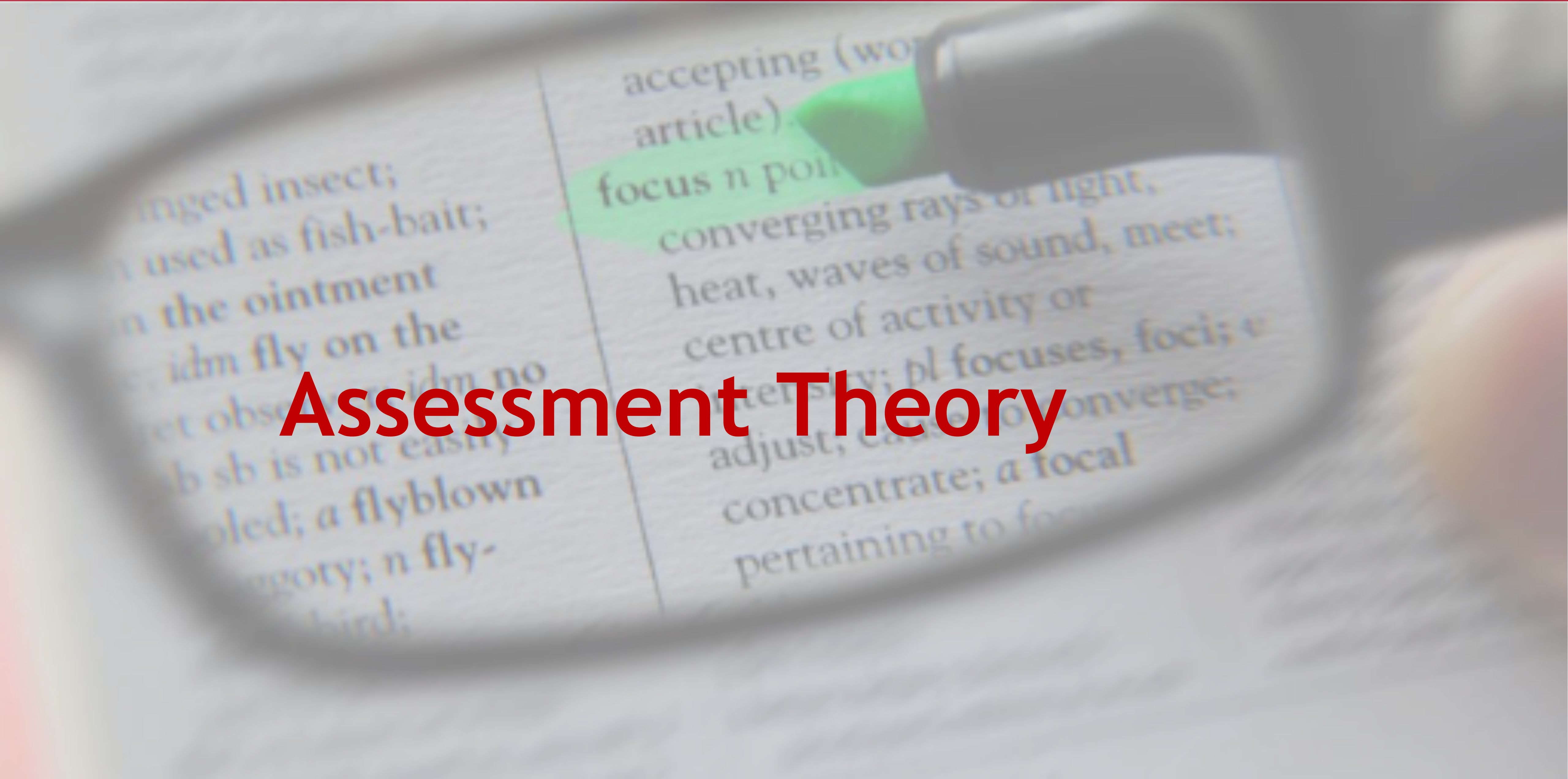
Spring 2023

Instructor: Ioannis Konstantinidis

Overview

- Assessment Theory
  - Train
  - Cross-validate
  - Test
- Example
  - K-fold CV

How do we know what the “machine” “learned”?



Assessment Theory

Assessment Theory (for humans)

Assessment is conducted during the *learning process* in order to modify teaching and learning activities to *improve the attainment* of students

Assessment Theory (for humans)

Assessment is conducted during the *learning process* in order to modify teaching and learning activities to *improve the attainment* of students

Formative assessment goal: to monitor student learning to provide ongoing ***feedback***

- identify their strengths and weaknesses
- target areas that need work

Summative assessment goal: to monitor learning ***outcomes***

- often for purposes of external accountability

Machine Learning (ML)

~~Students~~

Software models 🤖

Assessment Theory (for ML)

Assessment is conducted during the *learning process* in order to modify teaching and learning activities to *improve the attainment* of ~~students~~ model

Formative assessment goal: to monitor ~~student~~ model learning to provide ongoing ***feedback***

- identify their strengths and weaknesses
- target areas that need work

Summative assessment goal: to monitor learning ***outcomes***

- often for purposes of external accountability

Testing types

Criterion

vs.

Norm-Referenced

Criterion- vs. Norm-Referenced Tests

Criterion-referenced assessments measure individual performance: how well a student has mastered a specific learning objective.

- The test assesses how closely the performance matches specific criteria, not how the student compares to others
- Can you think of examples?

Criterion- vs. Norm-Referenced Tests

Criterion-referenced assessments measure individual performance: how well a student has mastered a specific learning objective.

- The test assesses how closely the performance matches specific criteria, not how the student compares to others
- “You must run the 100m in under 10.05 sec to qualify for the Olympics”

Criterion- vs. Norm-Referenced Tests

Criterion-referenced assessments measure individual performance: how well a student has mastered a specific learning objective.

- The test assesses how closely the performance matches specific criteria, not how the student compares to others
- "You must run the 100m in under 10.05 sec to qualify for the Olympics"

Norm-referenced assessments compare individual performance to a reference group: the overall acquisition of skills and knowledge relative to peers.

- The test usually covers a broad range of content, but what is tested is often mismatched to what is taught
- Can you think of examples?

Criterion- vs. Norm-Referenced Tests

Criterion-referenced assessments measure individual performance: how well a student has mastered a specific learning objective.

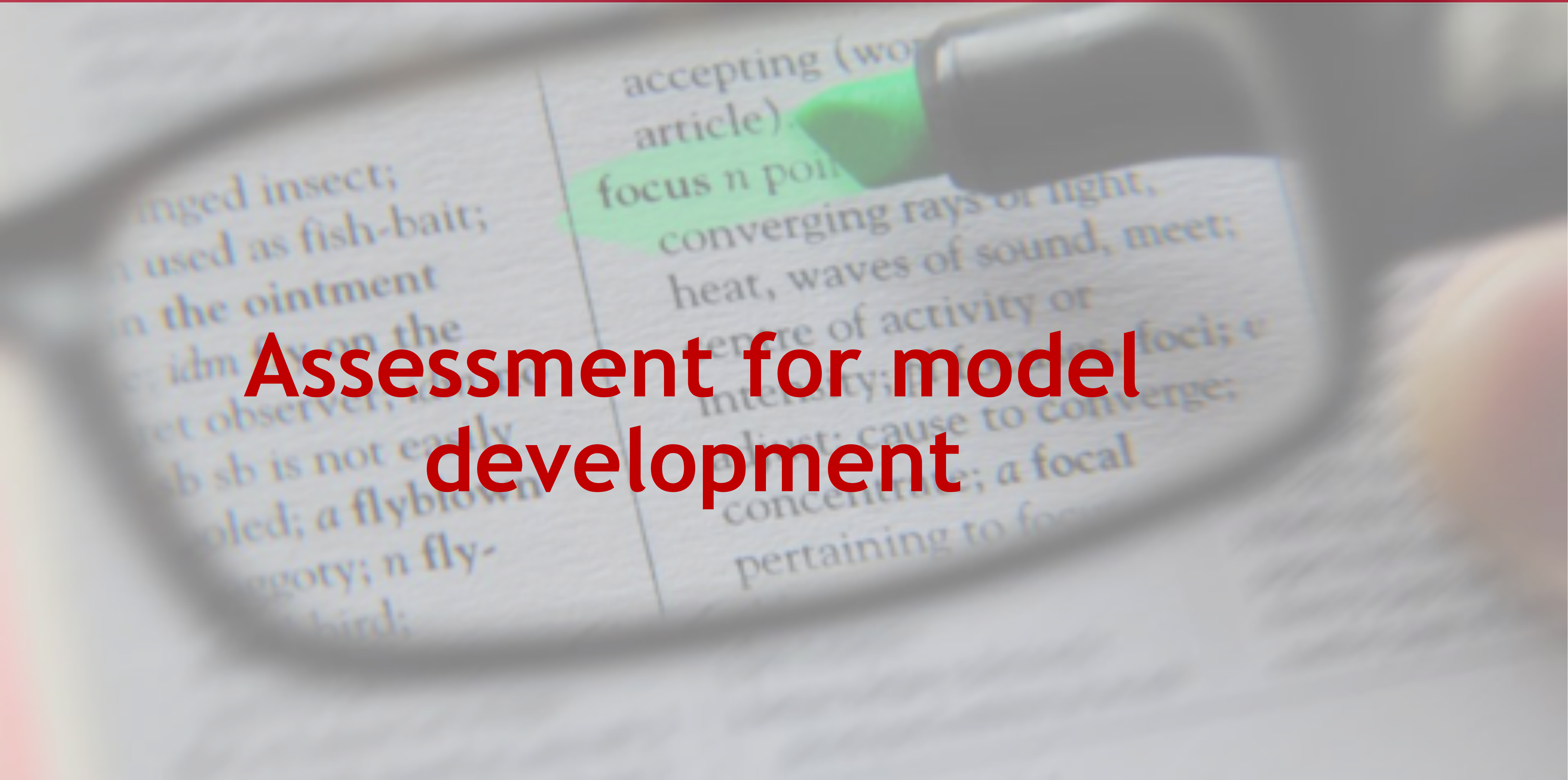
- The test assesses how closely the performance matches specific criteria, not how the student compares to others
- “You must run the 100m in under 10.05 sec to qualify for the Olympics”

Norm-referenced assessments compare individual performance to a reference group: the overall acquisition of skills and knowledge relative to peers.

- The test usually covers a broad range of content, but what is tested is often mismatched to what is taught
- “Grading on a curve” or percentile rank (e.g., SAT, GRE, IQ)

Assessment Theory (quick ref)

	Formative Assessment	**Summative Assessment**
When	During a learning activity	At the end of a learning activity
Goal	To improve learning	To make a decision
Feedback	Return to material	Final judgement
Frame of Reference	Always criterion	Sometimes criterion; Sometimes normative



Assessment for model development

Example: Train, Validate, Test

Quizzes are used to **train** students as they learn the material for the standardized test. [Formative + criterion]

Example: Train, Validate, Test

Quizzes are used to **train** students as they learn the material for the standardized test. [Formative + criterion]

Practice exams are used to **validate** how well the students learned the material, and to **evaluate** how students will perform on the standardized test. Each practice exam includes a different set of questions that were not used in the quizzes. [Summative + criterion]

Example: Train, Validate, Test

Quizzes are used to **train** students as they learn the material for the standardized test. [Formative + criterion]

Practice exams are used to **validate** how well the students learned the material, and to **evaluate** how students will perform on the standardized test. Each practice exam includes a different set of questions that were not used in the quizzes. [Summative + criterion]

The *standardized test* is used to **test** how well the students learned the material and **rank** students based on their scores. The standardized test includes one common set of questions for all students, different from all the questions used before. [Summative + norm]

Fit: quizzes

- How should we find the internal model parameters that achieve the best fit?

- Fix an *objective* function

- Keep modifying parameters until there is no room for improvement

- Implemented in scikit-learn as the fit() method

Evaluation: Practice Exams

- How well will the trained model do?
- Fix a *scoring* function
- Evaluate model **capability** for standardized test score
- Implemented in scikit-learn as the cross_val_score() method or similar

Selection: Standardized Test

- Which model **does** best?
- Use the separate testing data
- Pick the model with the best score

Quick aside: (hyper)parameters

Internal model parameters are computed to optimize an objective function (e.g., coefficients in LR)

Many times the objective function is actually a family of functions indexed by a variable, e.g.,

- λ for Ridge or LASSO regression

Other models may lack an objective function, but still rely on fixing the value of a variable, e.g.,

- k (# of neighbors) in kNN classification

This callable variable is called a hyperparameter

Quick aside: (hyper)parameters

It is best to think of two different hyperparameters as specifying the same model for purposes of understanding the theory,

BUT

they specify different, separate models for purposes of evaluation.

E.g.,

- `KNeighborsClassifier(n_neighbors=5)` and
- `KNeighborsClassifier(n_neighbors=10)`

are two separate models, just like

- `KNeighborsClassifier()` and
- `LogisticRegression()`

are two different models

Model tuning

Is the process of selecting which

- Hyperparameter choice, aka
- Objective function choice, aka
- Model choice

produces the best result

Model Development and Testing (quick ref)

	Fit	Evaluate	Select
Optimized Measure	Objective Function	Scoring Function	Scoring Function
Goal	Compute Model Parameters (weights)	Evaluate Model Capacity (scores)	Chose Model Hyperparameters / Type
Method	Guided Search (gradient descent)	Cross-validation	Comparison (list)
Data Set	Training Data	Training Data	Testing Data

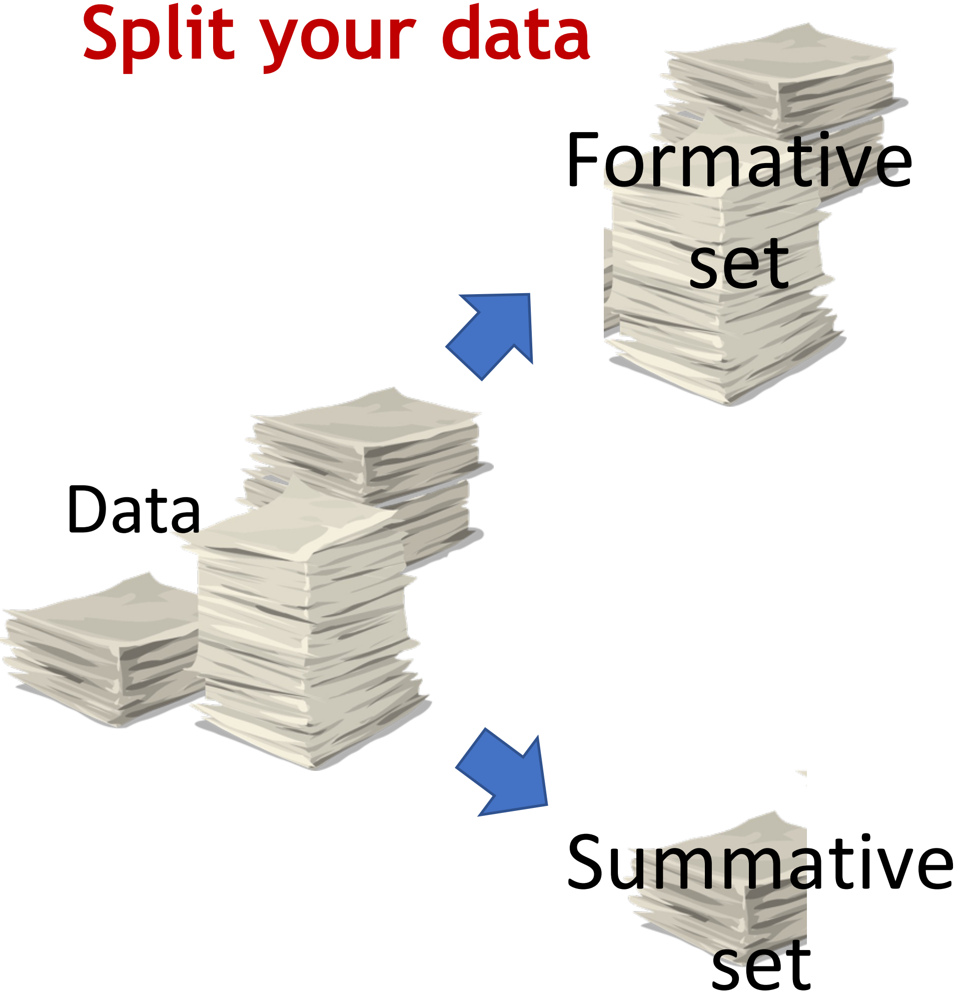
Getting the most out of your data

Your data is the Question Bank

Your data is the Question Bank

Don’t let your model cheat!

Split your data



Data

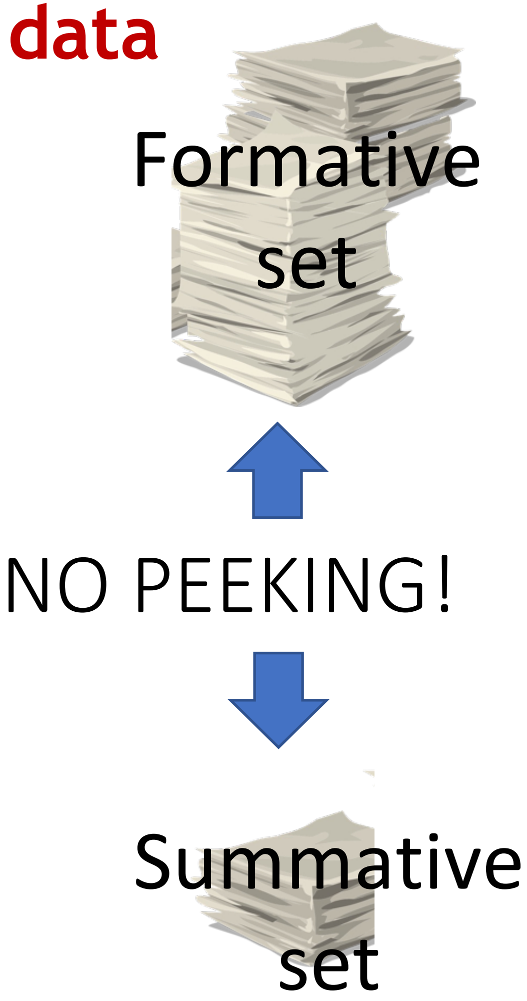
Formative set

Summative set

Model Development

Model Testing

Split your data



Formative set

NO PEEKING!

Summative set

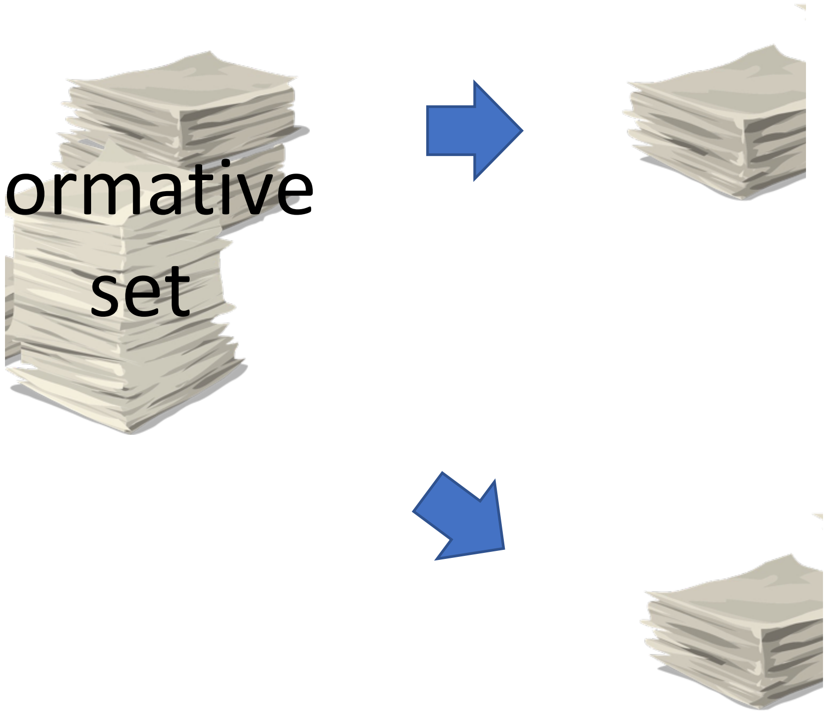
Data used only to develop the model

Data used only to test performance of a fully-specified model https://reproducible.cs.princeton.edu/

Split your data

Field	Paper	Year	# papers reviewed	# papers w/pitfalls	Pitfalls
Medicine	Bouwmeester et al.	2012	71	27	No train-test split
Neuroimaging	Whelan et al.	2014	—	14	No train-test split; Feature selection on train and test set
Autism Diagnostics	Bone et al.	2015	—	3	Duplicates across train-test split; Sampling bias
Nutrition research	Ivanescu et al.	2016	—	4	No train-test split
Satelitte imaging	Nalepa et al.	2019	17	17	Non-independence between train and test sets
Tractography	Poulin et al.	2019	4	2	No train-test split
Brain-computer interfaces	Nakanishi et al.	2020	—	1	No train-test split
Histopathology	Oner et al.	2020	—	1	Non independence between train and test sets
Computer security	Arp et al.	2020	30	30	No train-test split; Pre-processing on train and test sets together; Illegitimate features; others
Neuropsychiatry	Poldrack et al.	2020	100	53	No train-test split; pre-processing on train and test sets together
Medicine	Vandewiele et al.	2021	24	21	Feature selection on train-test sets; Non-independence between train and test sets; Sampling bias
Radiology	Roberts et al.	2021	62	62	No train-test split; duplicates in train and test sets; sampling bias

Split again

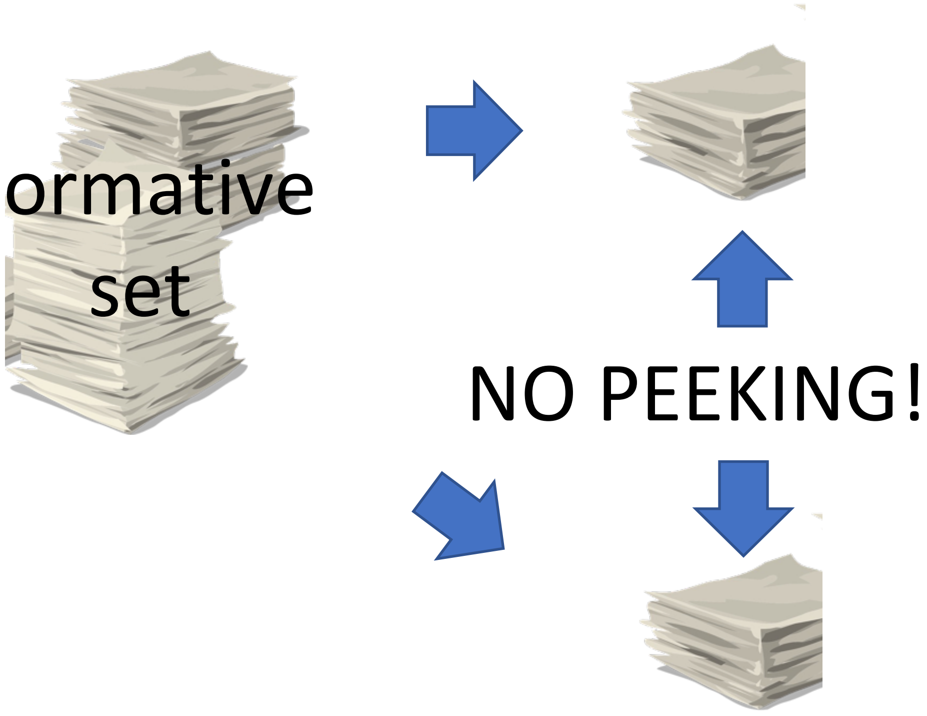


Formative set

Training set: used to compute parameters

Validation set: used to evaluate capacity

Split again



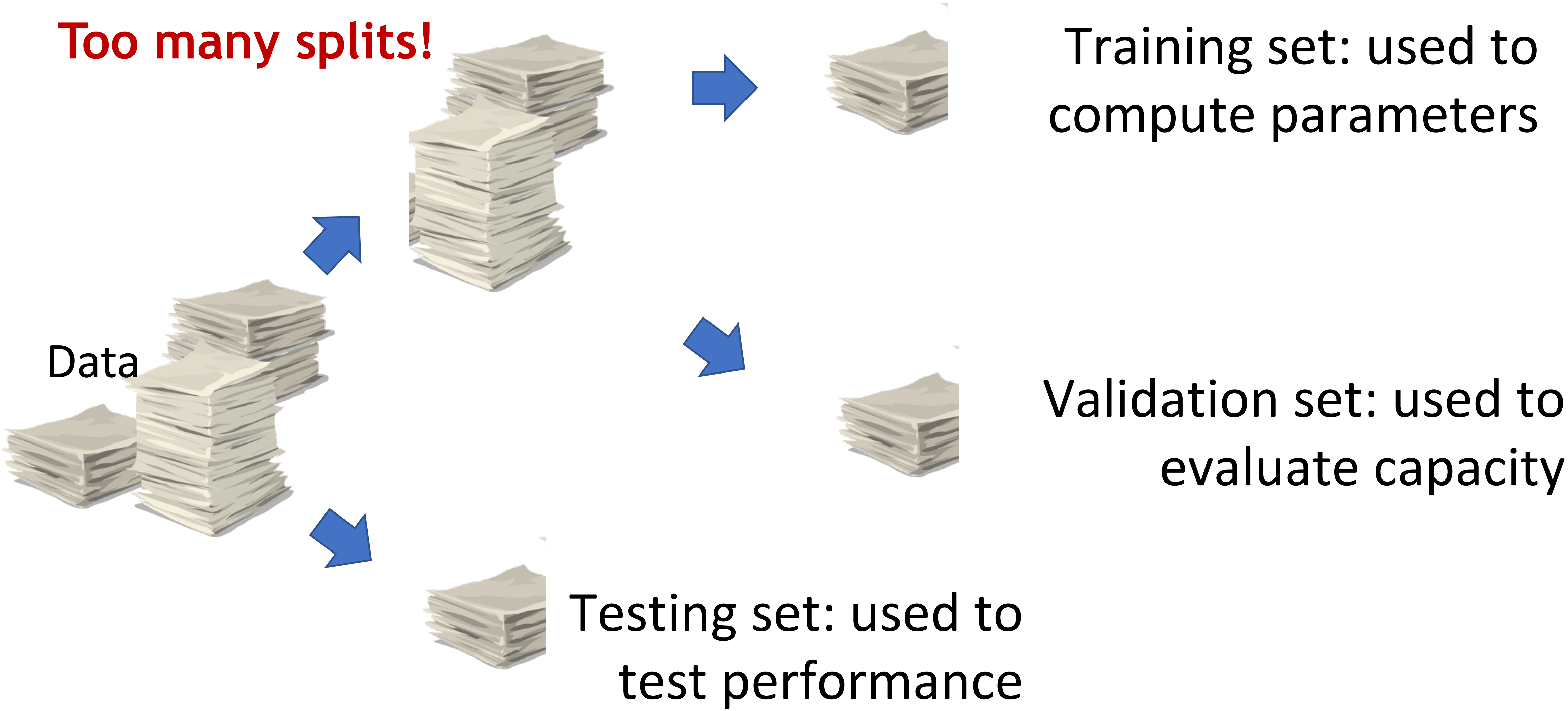
Formative set

NO PEEKING!

Training set: used to compute parameters

Validation set: used to evaluate capacity

Too many splits!



Data

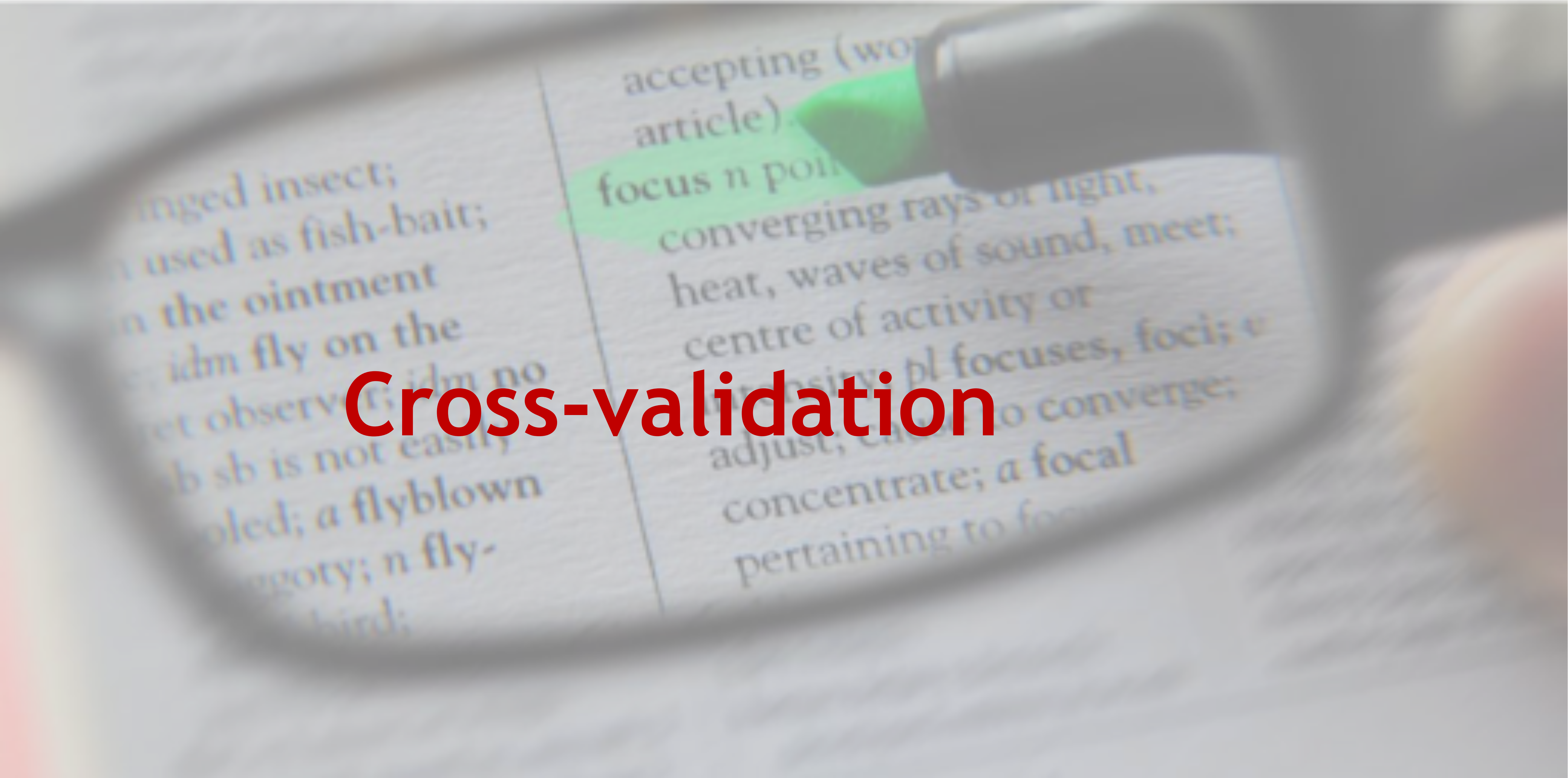
Training set: used to compute parameters

Validation set: used to evaluate capacity

Testing set: used to test performance

Training set loses power

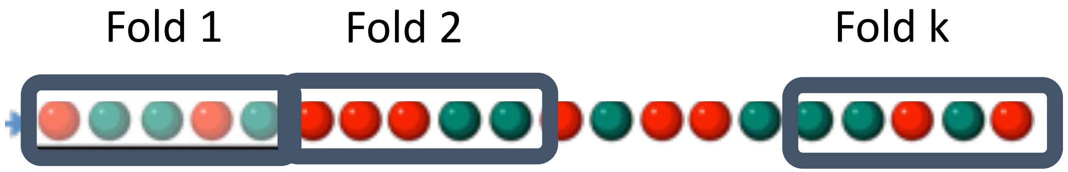
vs.



Cross-validation

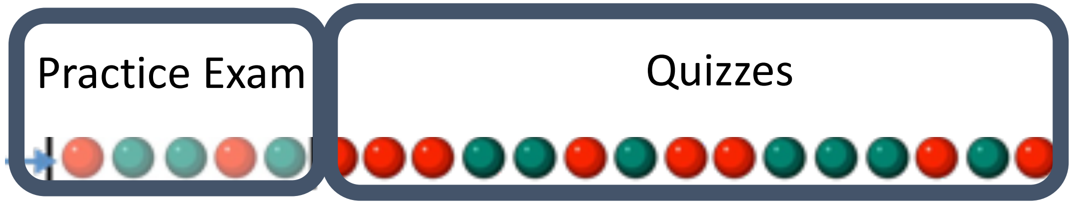
K-fold Cross-validation

- Randomly partition the formative data into *k mutually exclusive* folds, each approximately equal size



K-fold Cross-validation

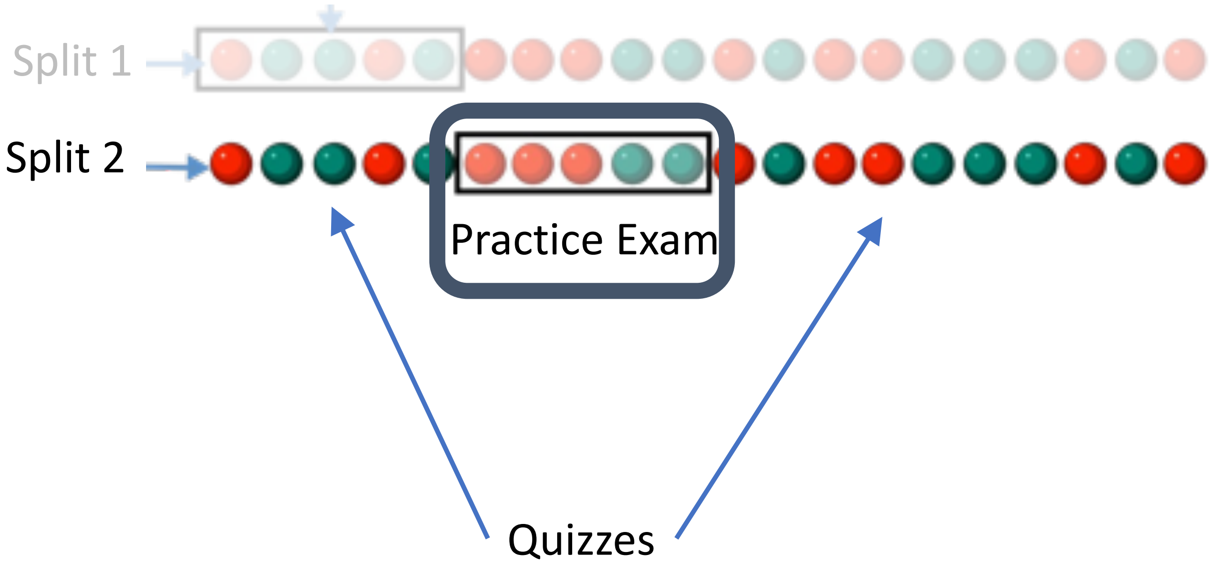
- Randomly partition the formative data into *k mutually exclusive* folds, each approximately equal size



Use one fold as an evaluation set and all others as a training set

K-fold Cross-validation

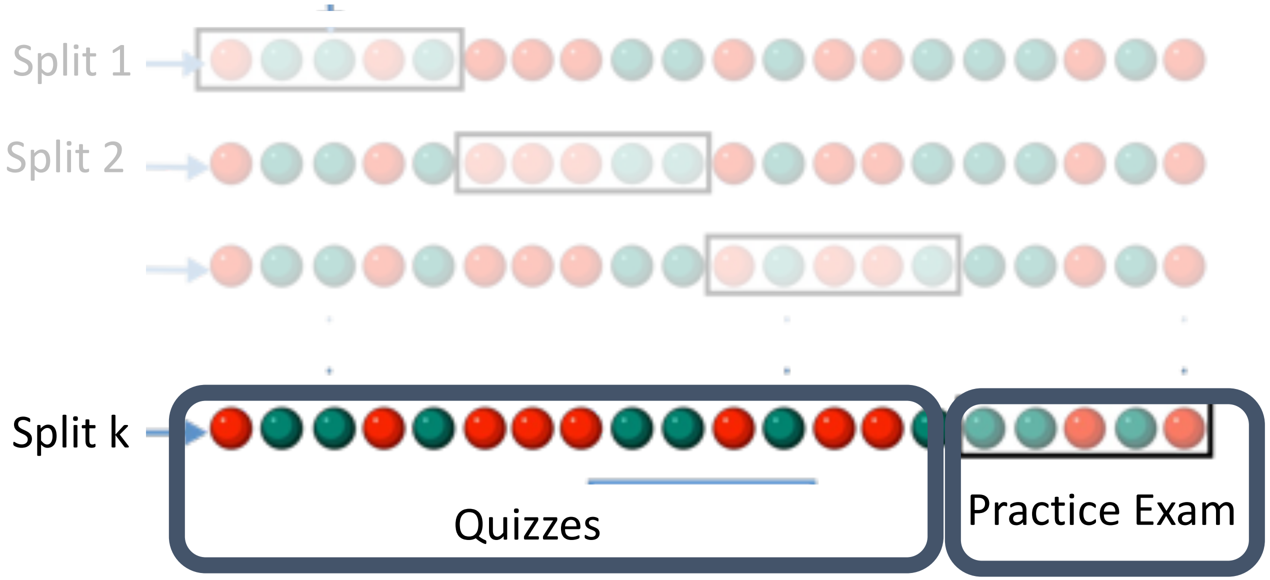
- Randomly partition the formative data into *k mutually exclusive* folds, each approximately equal size



Repeat using **another** fold as an evaluation set and all others as a training set

K-fold Cross-validation

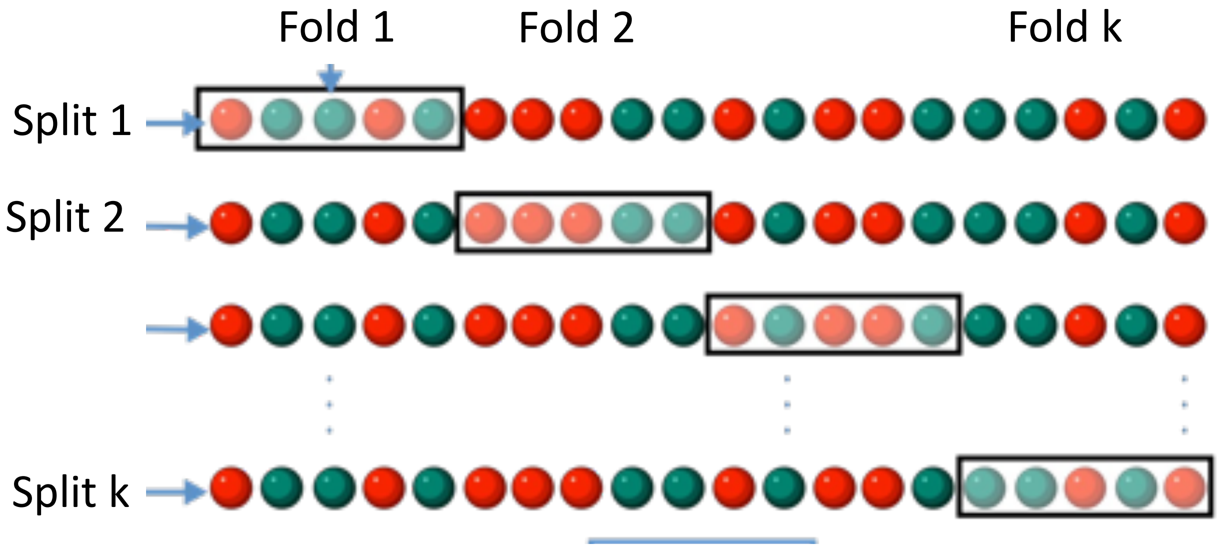
- Randomly partition the formative data into *k mutually exclusive* folds, each approximately equal size



Iterate using one fold as an evaluation set and all others as a training set

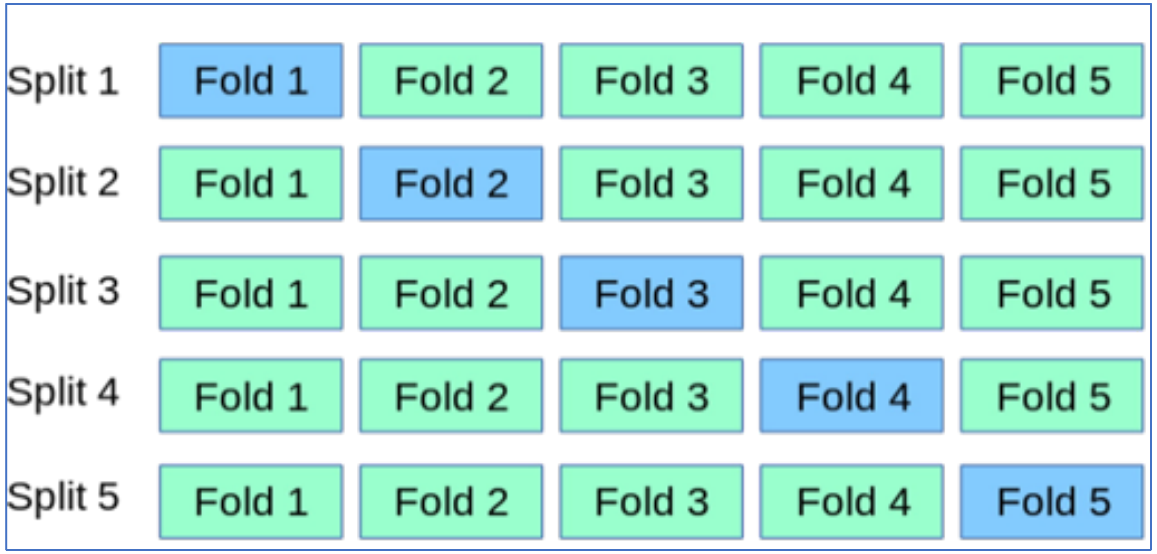
K-fold Cross-validation

- All of the **formative** data contribute to both training and evaluation, with no contamination



K-fold Cross-validation

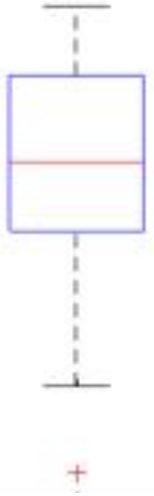
- Allows the computation of summary statistics for score centrality and dispersion (spread)



Split 1	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 2	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 3	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 4	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 5	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5

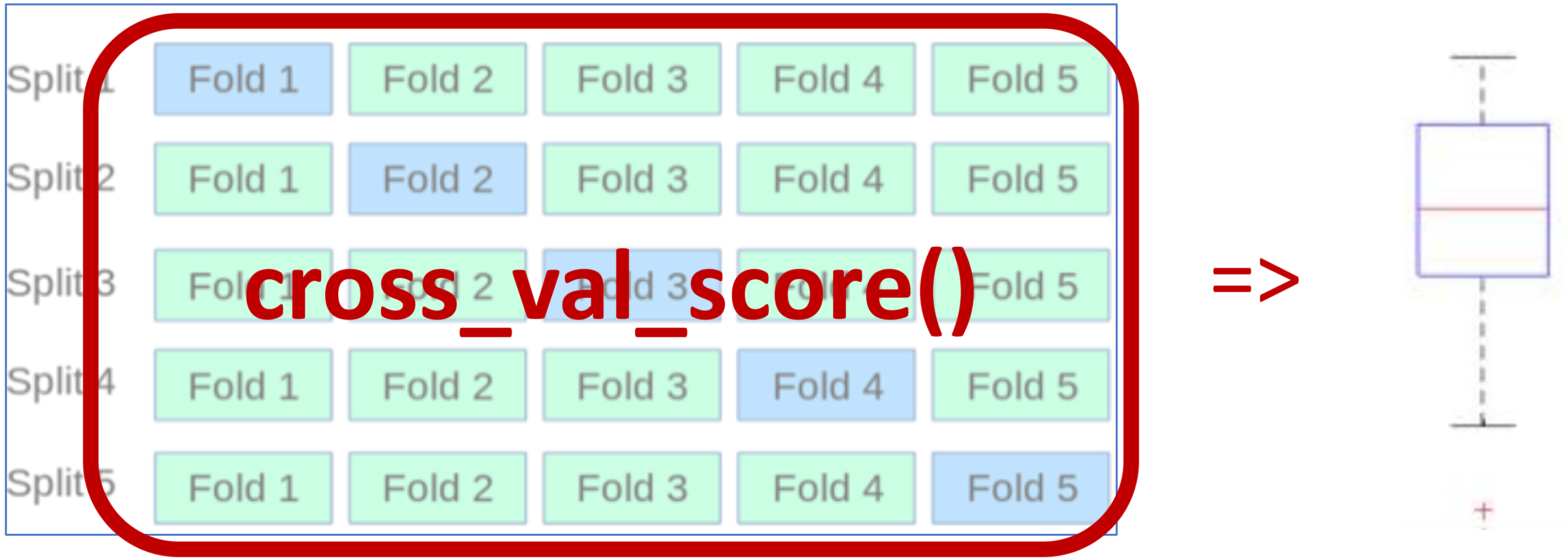
=>

Box-and-whisker plot
of score distribution
over all splits (exams)



K-fold Cross-validation

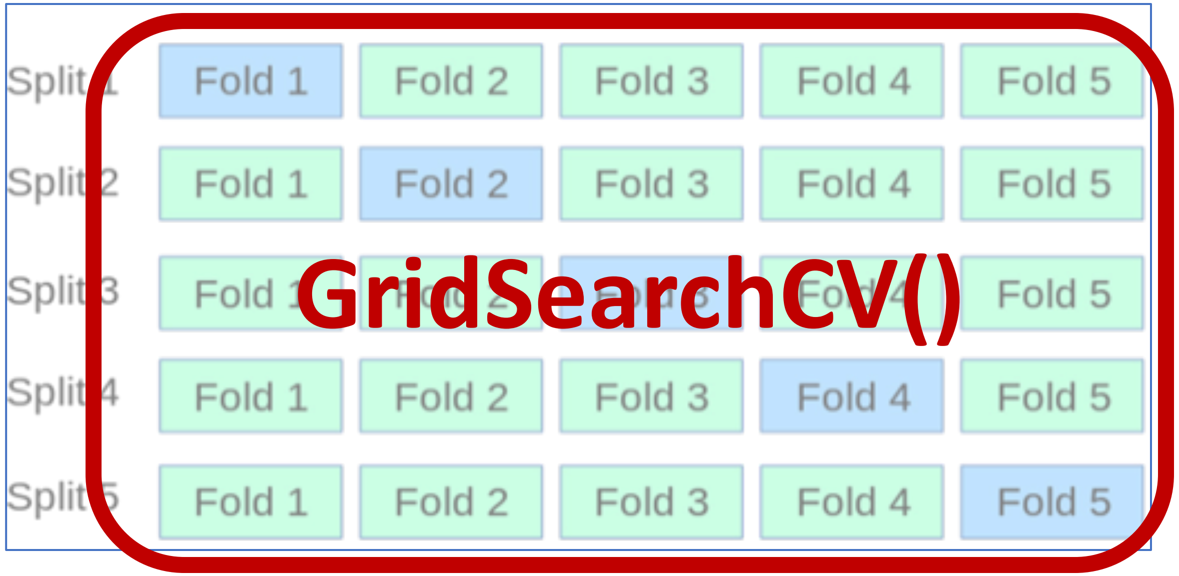
- Allows the computation of summary statistics for score centrality and dispersion (spread)
- No need to hand-code iteration loops; scikit-learn has a helper function



cross_val_score() =>

K-fold Cross-validation

- Also allows the selection of hyperparameters
- Scikit-learn has a function for that as well

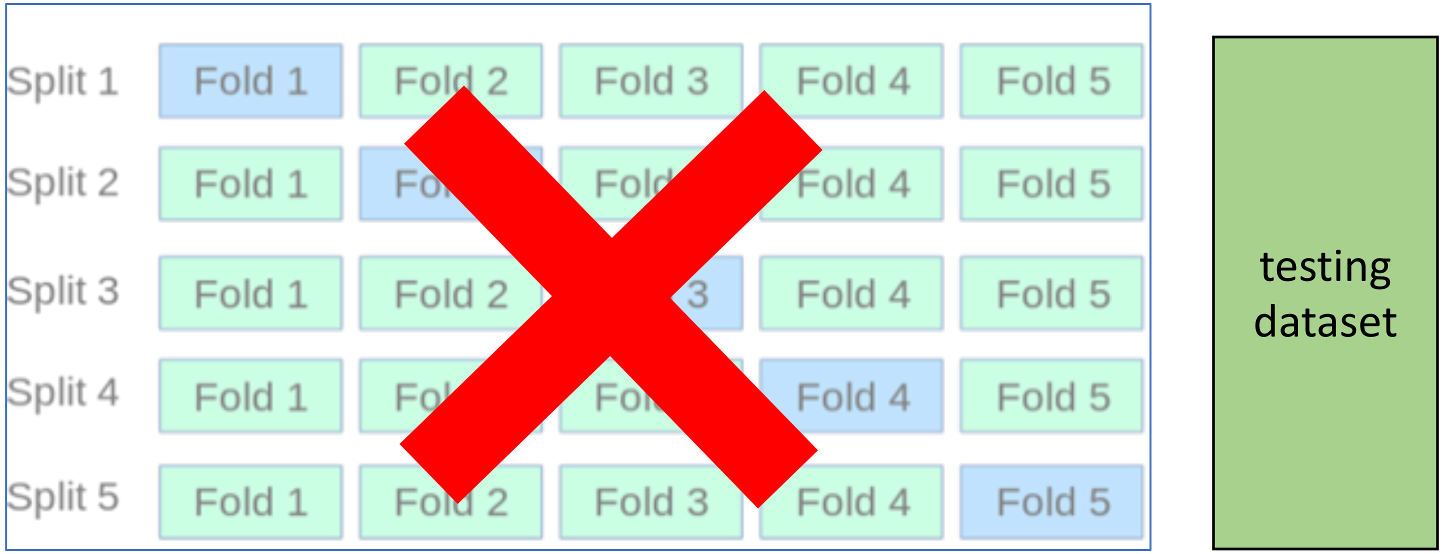


Split 1 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Split 2 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Split 3 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Split 4 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Split 5 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5

GridSearchCV()

K-fold Cross-validation

- Not a substitute for summative assessment
- Test using the separate **summative** dataset

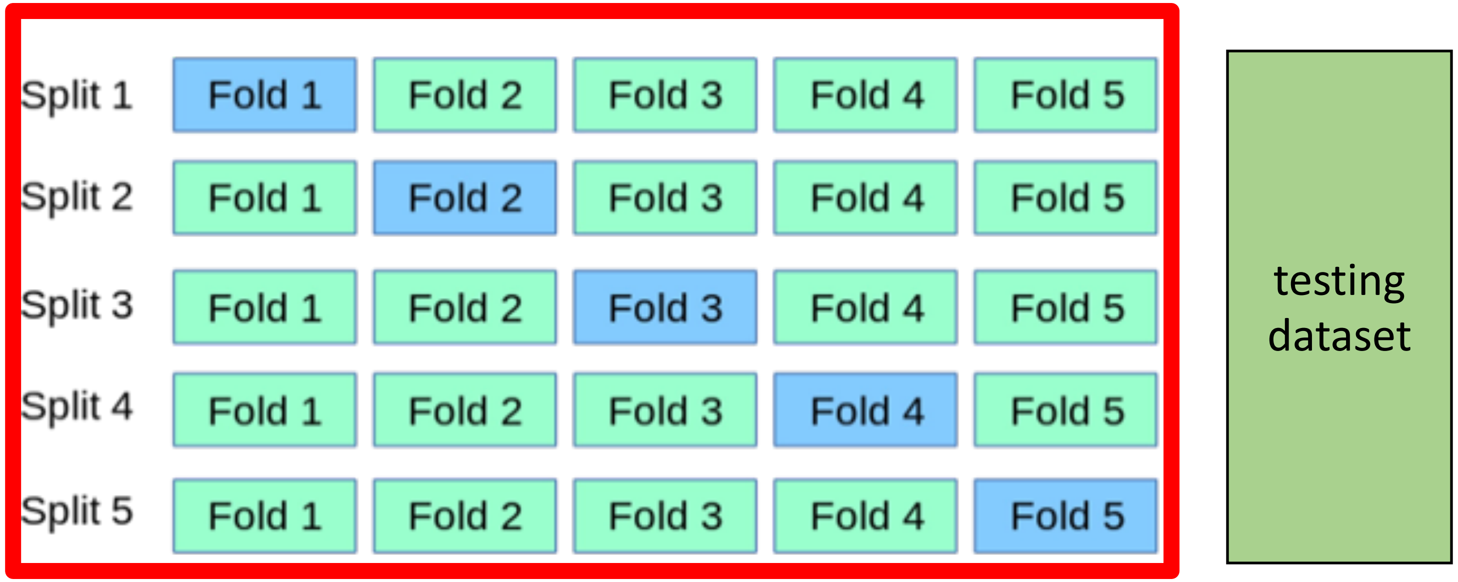


Split 1	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 2	Fold 1	Fo	Fold	Fold 4	Fold 5
Split 3	Fold 1	Fold 2	3	Fold 4	Fold 5
Split 4	Fold 1	Fo	Fo	Fold 4	Fold 5
Split 5	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5

testing dataset

K-fold Cross-validation

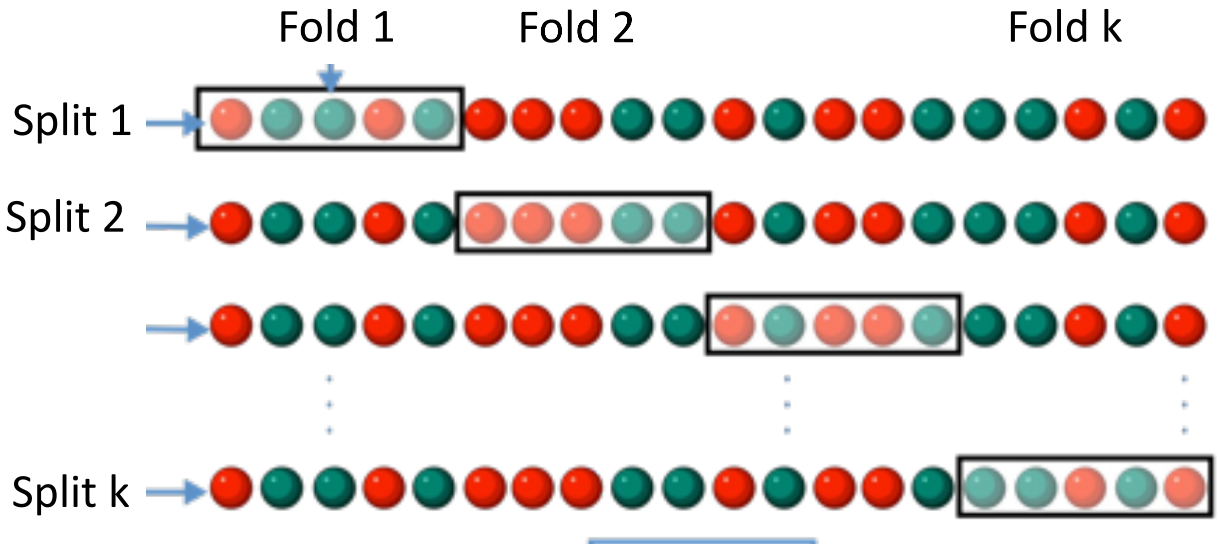
- Not a substitute for summative assessment
- Retrain using ALL of the training (development) set
- Test using the separate **summative** dataset



Split 1	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 2	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 3	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 4	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 5	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5

Stratified Cross-Validation

Folds are stratified so that class distribution in each fold is approximately the same as in the initial data



Hands-on Example:

k-fold cross validation

How to design good assessments?

Other Criteria for Performance Evaluation

Speed

- How fast can it predict
- How long does it take to train

Storage

- How much memory is needed for the model
- How much compression can be applied to the data

Scalability

- How modular is the implementation
- How large is the support community

Predictive capability

Homework Assignment #1
Due Wednesday (February 8), 11:59 pm (Central)