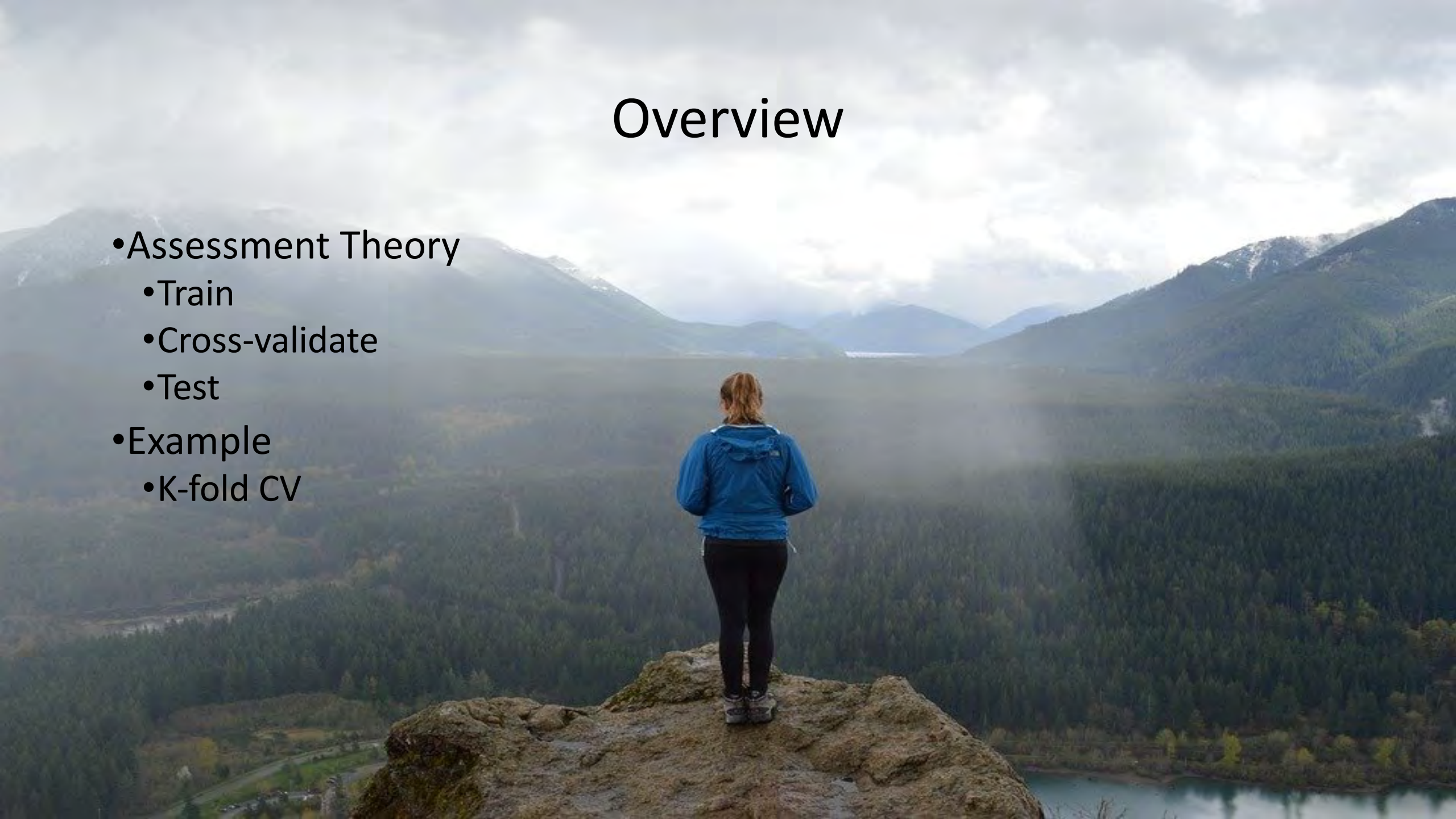
HPE DSI 311
Introduction to Machine Learning

Summer 2021

Instructor: Ioannis Konstantinidis

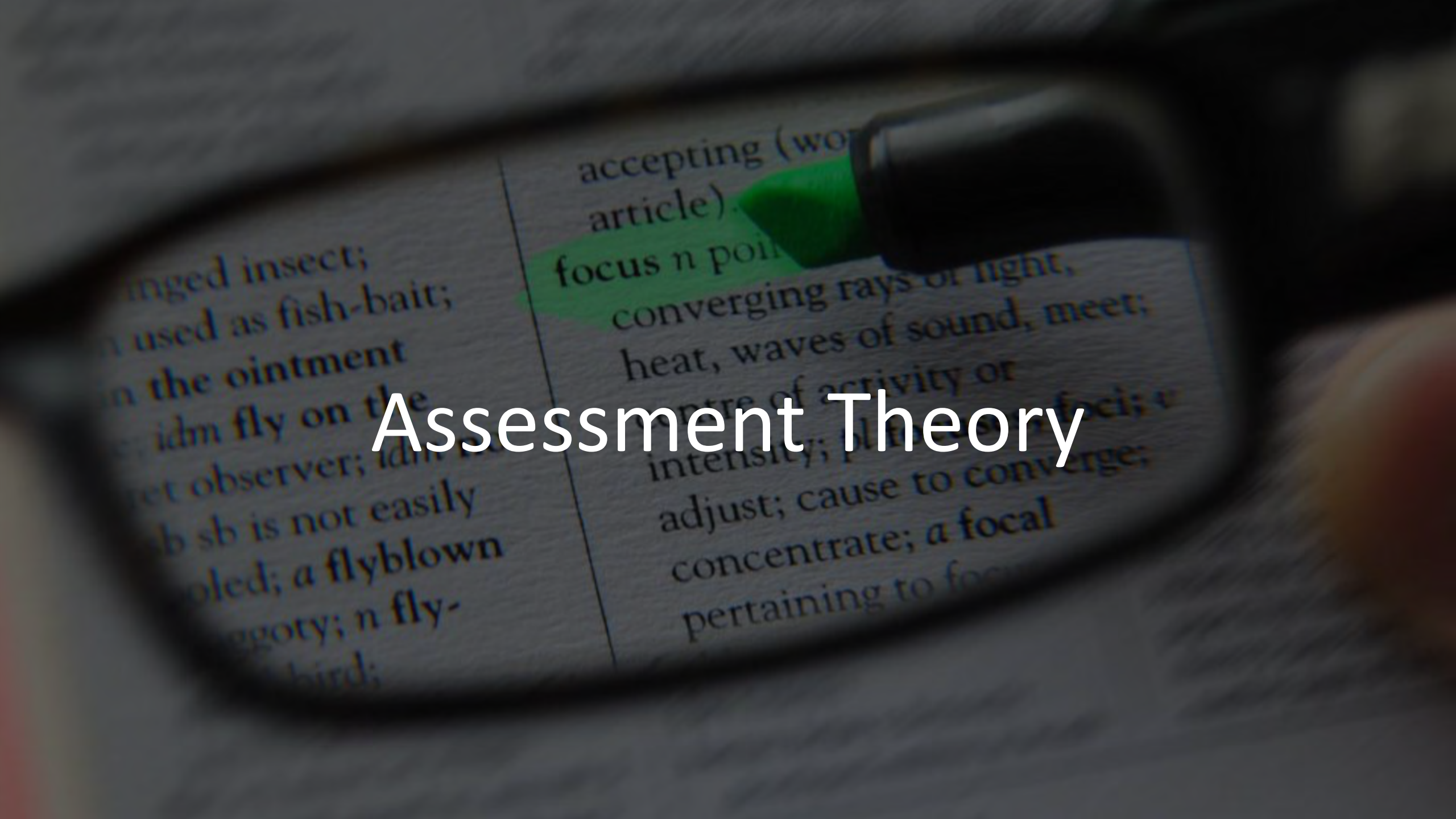
UNIVERSITY of HOUSTON
DIVISION OF RESEARCH
HEWLETT PACKARD ENTERPRISE DATA SCIENCE INSTITUTE

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)

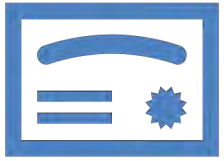
Assessment is conducted during the *learning process* in order to modify teaching and learning activities to *improve the attainment* of students

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

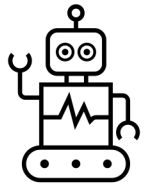


Summative assessment goal: to monitor learning ***outcomes***

- often for purposes of external accountability

Machine Learning (ML)

~~Students~~

Software models 

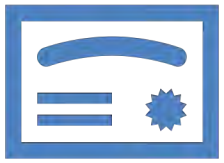
Assessment Theory (for ML)

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

model

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

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 be this tall to ride this ride!"

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 be this tall to ride this ride!”

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.

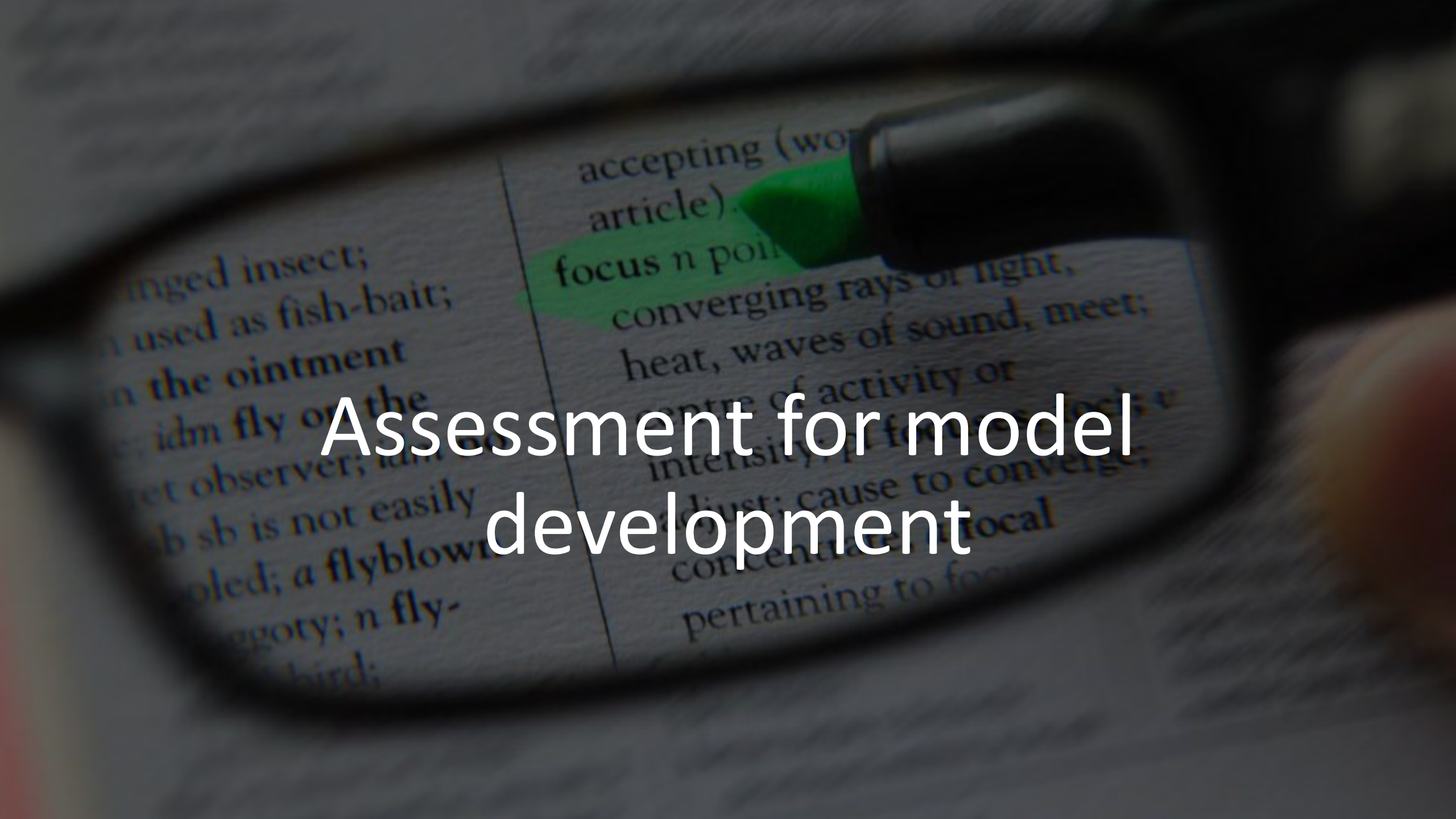
- The test assesses how closely the performance matches specific criteria, not how the student compares to others
- “You must be this tall to ride this ride!”

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 tweak model parameters to achieve the best fit possible?

- 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

Implemented in scikit-learn by GridSearchCV() or similar

Quick aside: (hyper)parameters

- Are they a special kind of parameter?
- What is the difference?

Quick aside: (hyper)parameters

Model parameters are computed to optimize an objective function (e.g., weights, coefficients)

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 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 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

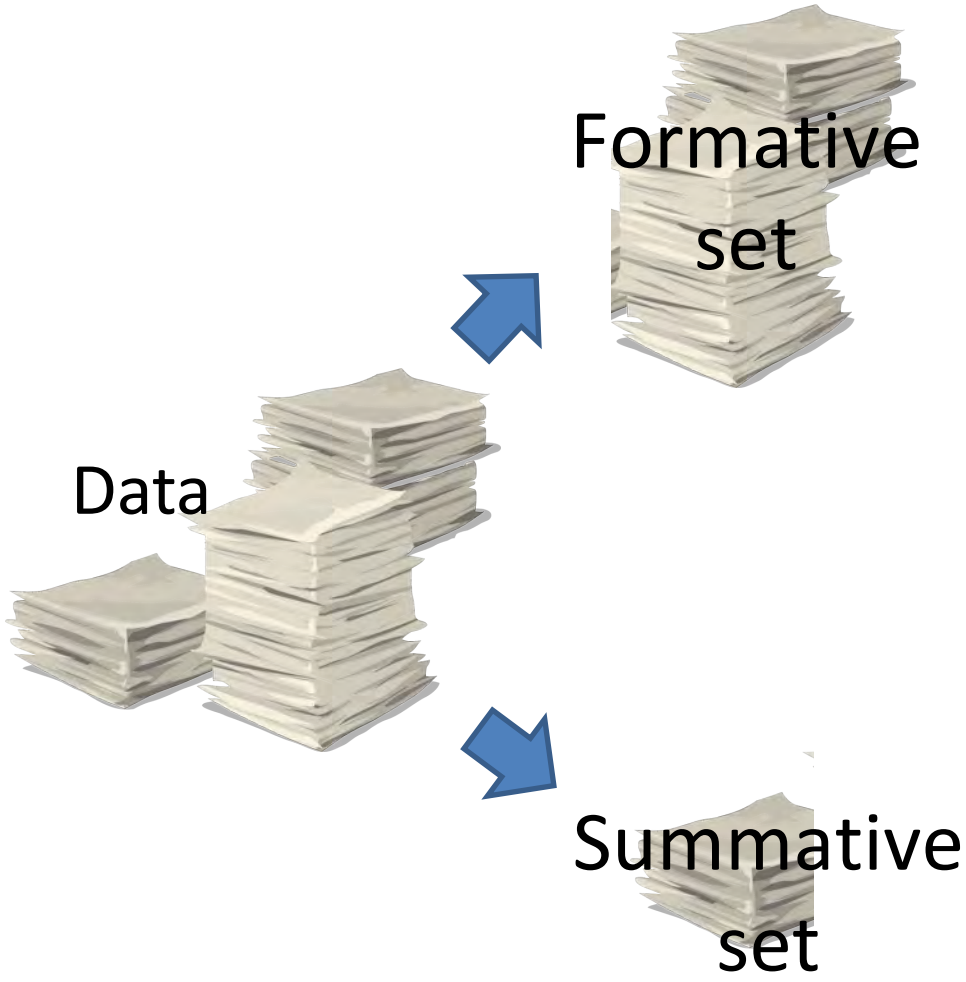

Getting the most
out of your data

Your data is the Question Bank

Your data is the Question Bank

Don’t let your model cheat!

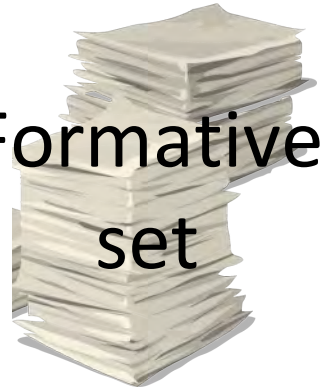
Split your data



Model Development

Model Testing

Split your data



Formative set

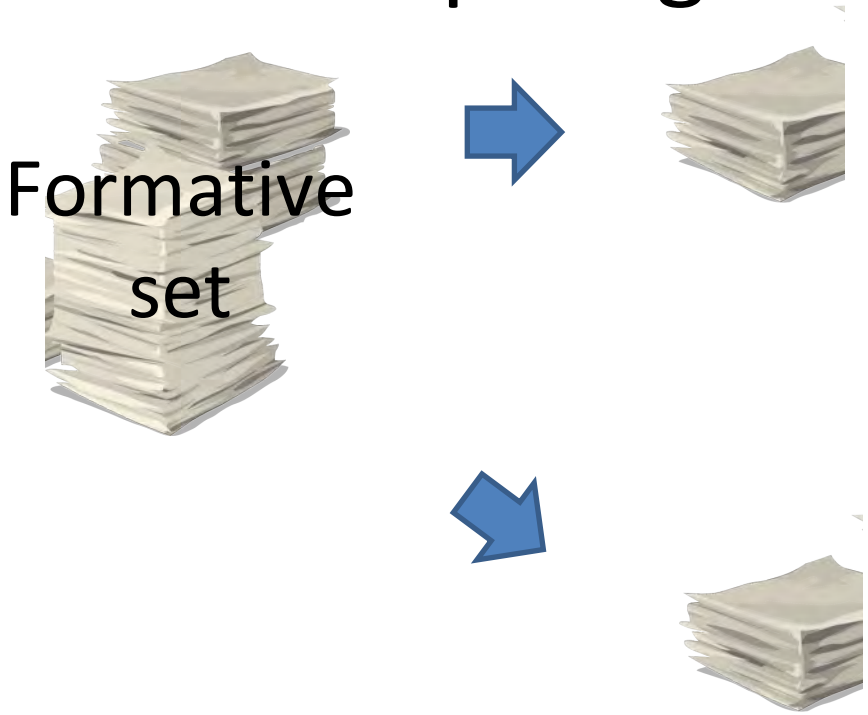
Data used only to develop the model

NO PEEKING!

Summative set

Data used only to test performance of a fully-specified model

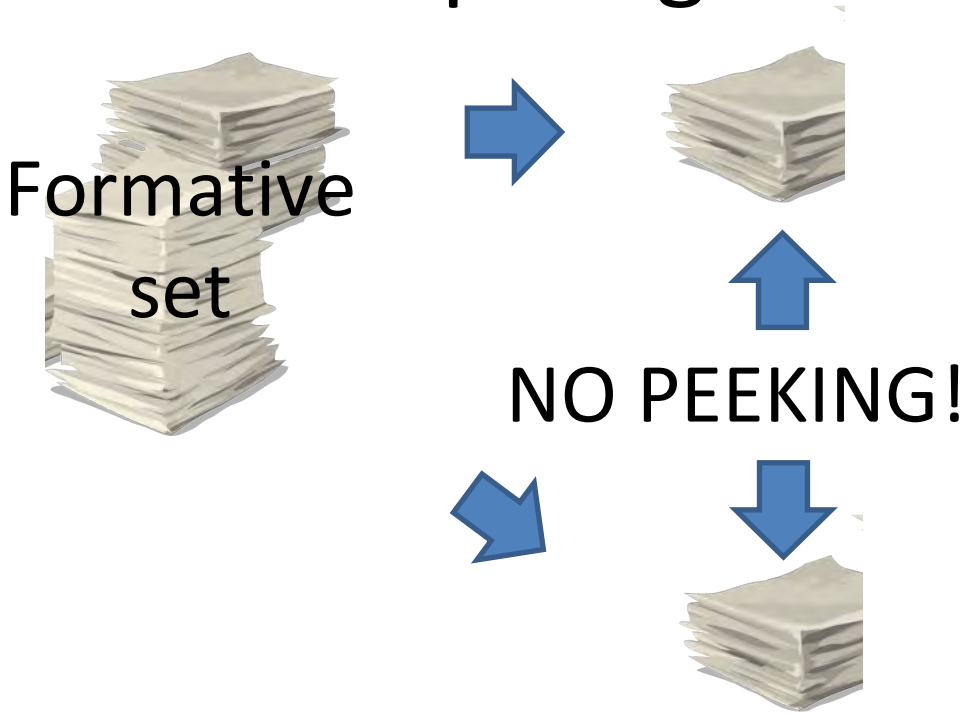
Split again



Training set: used to compute parameters

Validation set: used to evaluate capacity

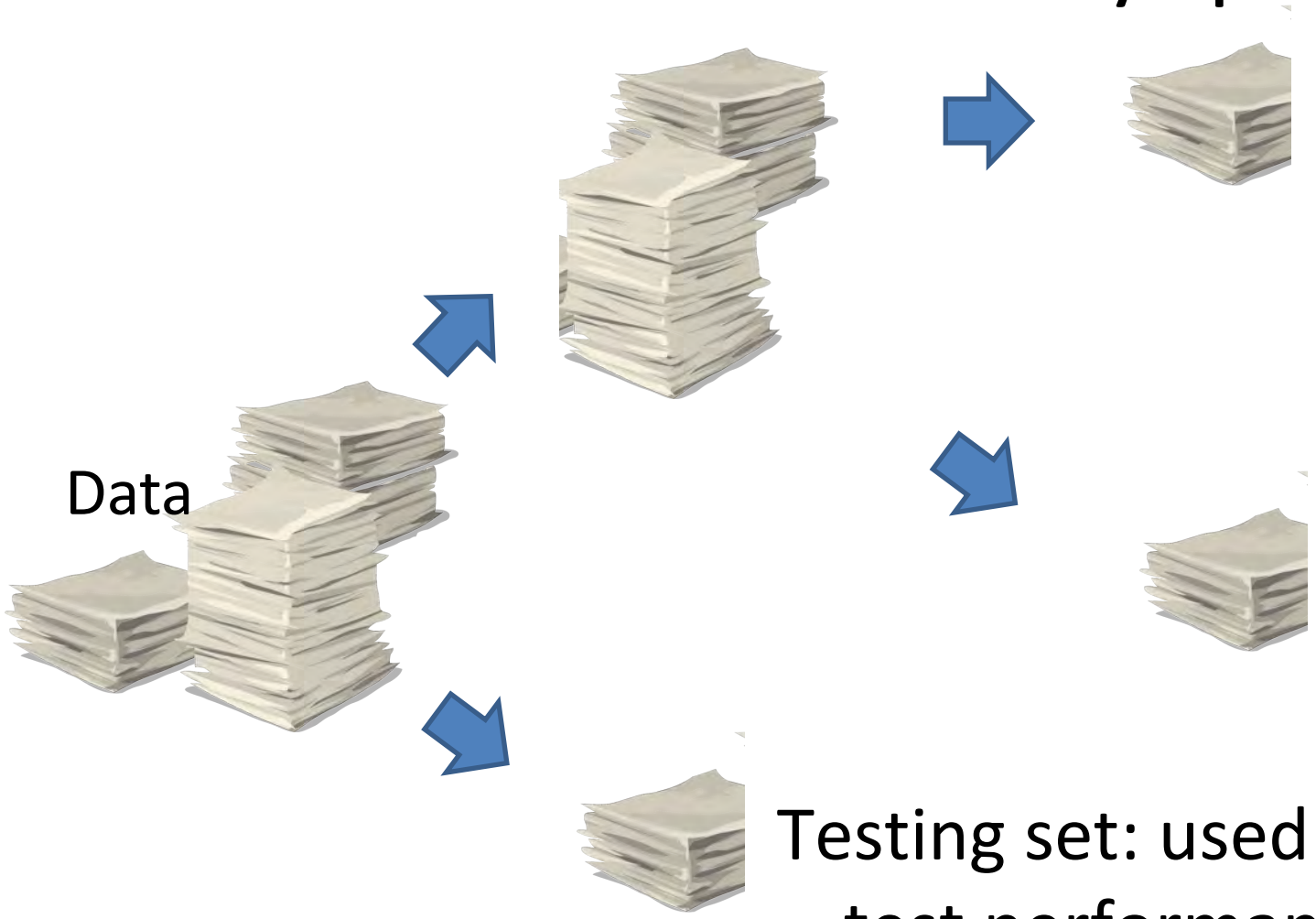
Split again



Training set: used to compute parameters

Validation set: used to evaluate capacity

Too many splits!



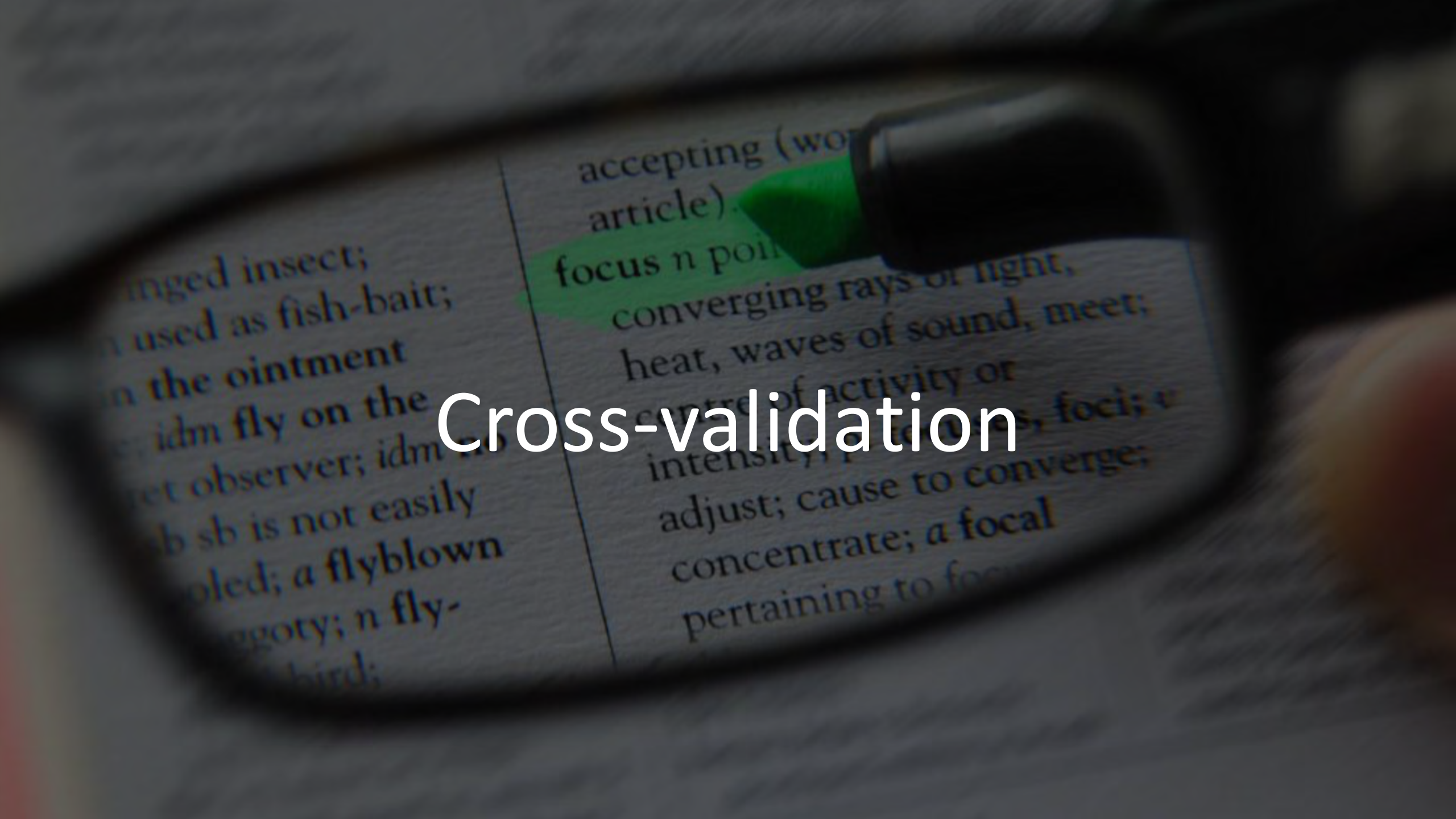
Training set: used to compute parameters

Validation set: used to evaluate capacity

Testing set: used to test performance

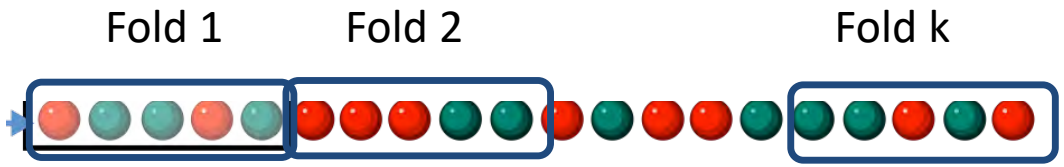
Training set loses power

vs.


Cross-validation

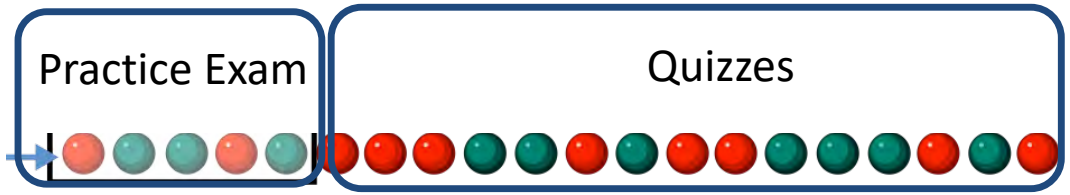
K-fold Cross-validation

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



K-fold Cross-validation

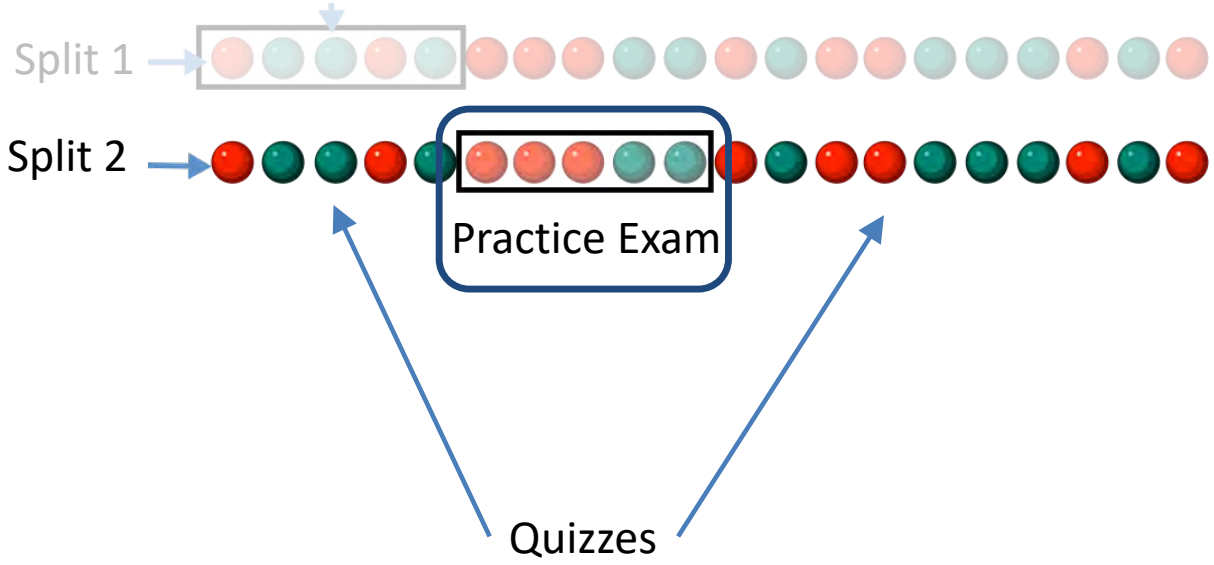
- Randomly partition the formative data into *k mutually exclusive* folds, each approximately equal size



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

K-fold Cross-validation

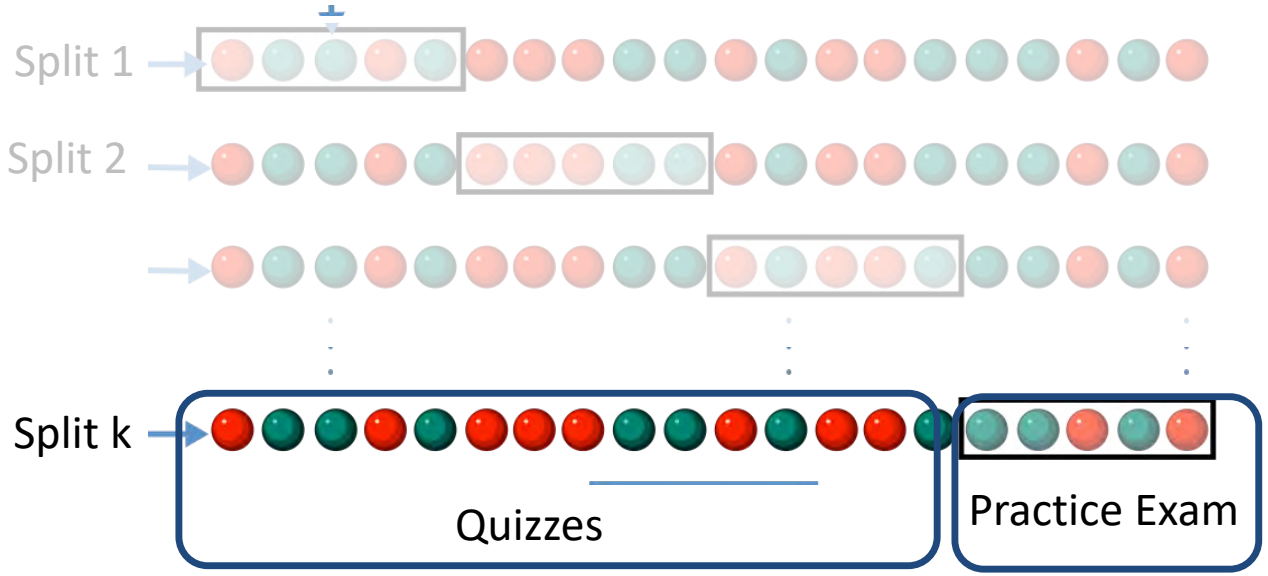
- Randomly partition the formative data into *k mutually exclusive* folds, each approximately equal size



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

K-fold Cross-validation

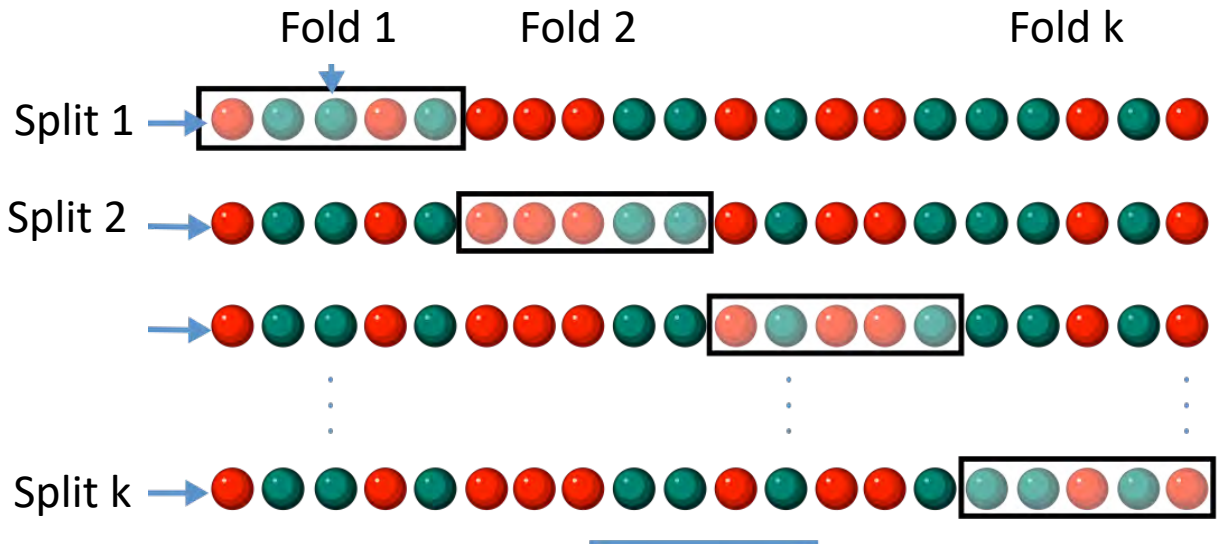
- Randomly partition the formative data into *k mutually exclusive* folds, each approximately equal size



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

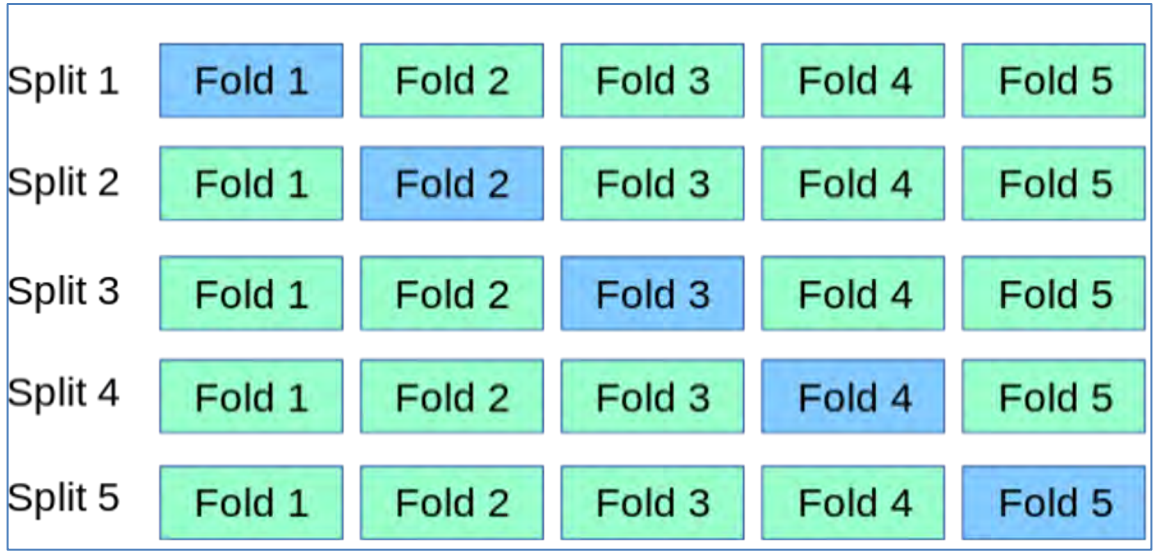
K-fold Cross-validation

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



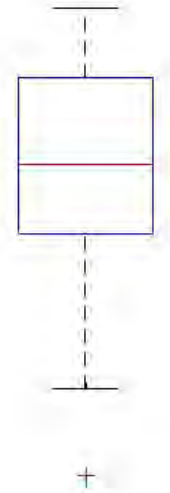
K-fold Cross-validation

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



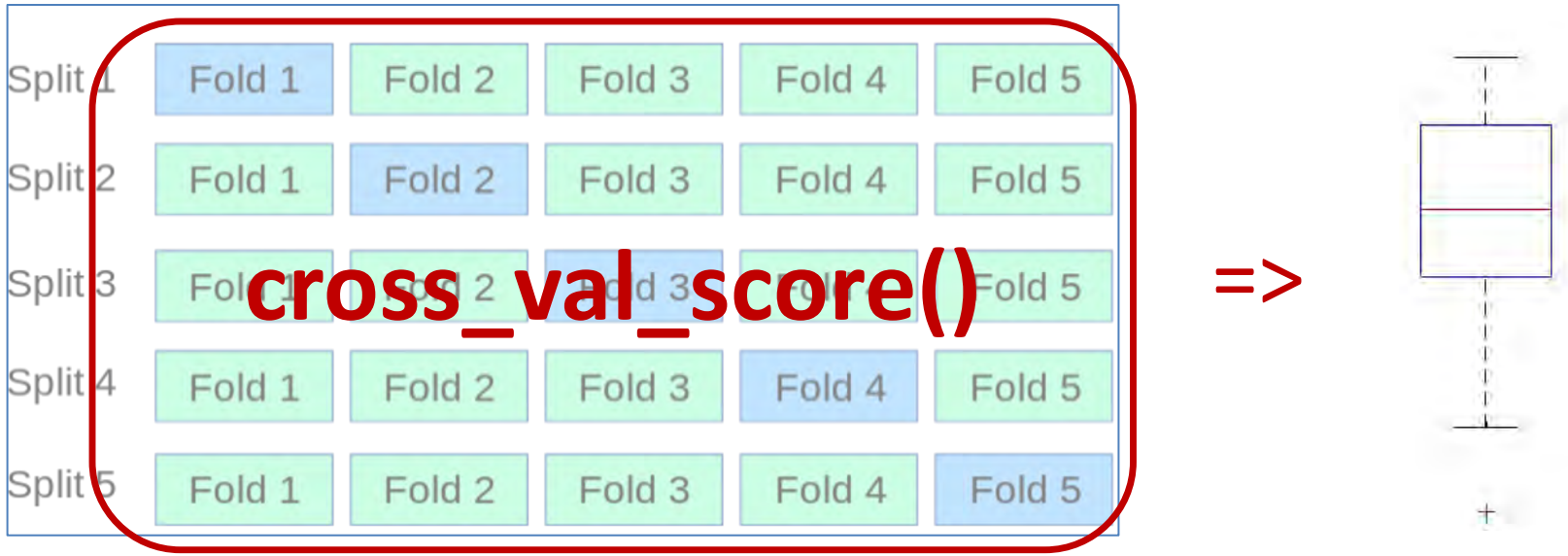
=>

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



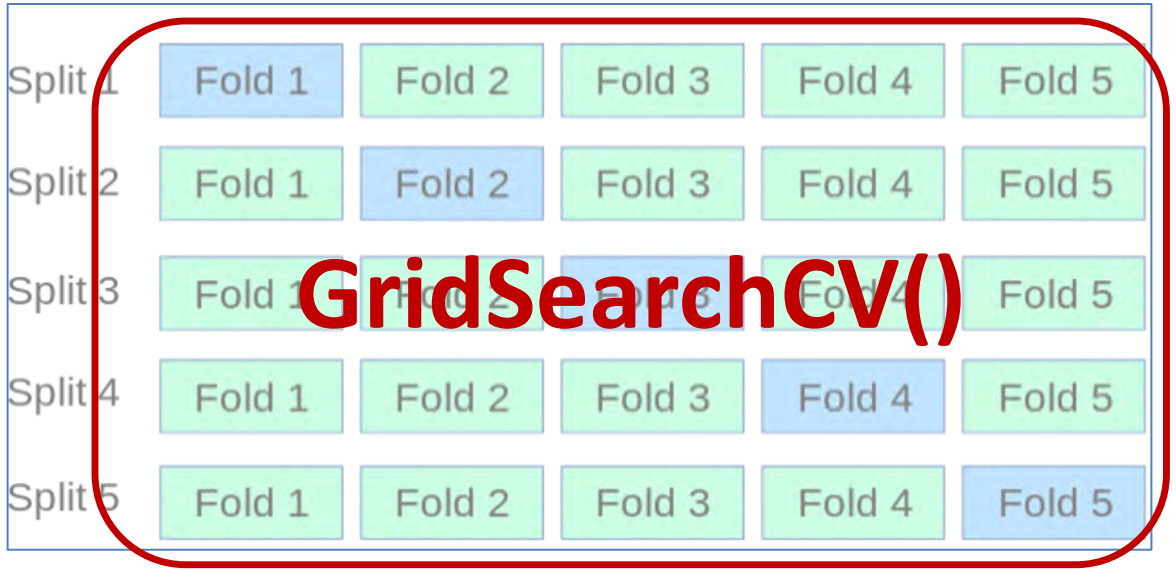
K-fold Cross-validation

- 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



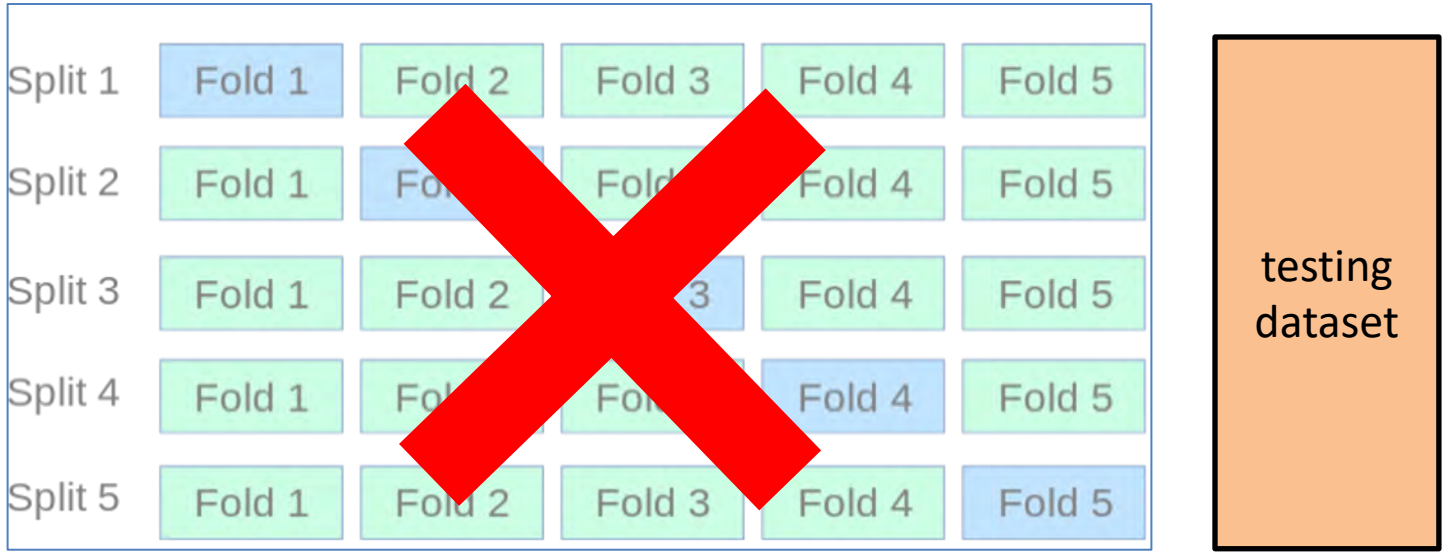
K-fold Cross-validation

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



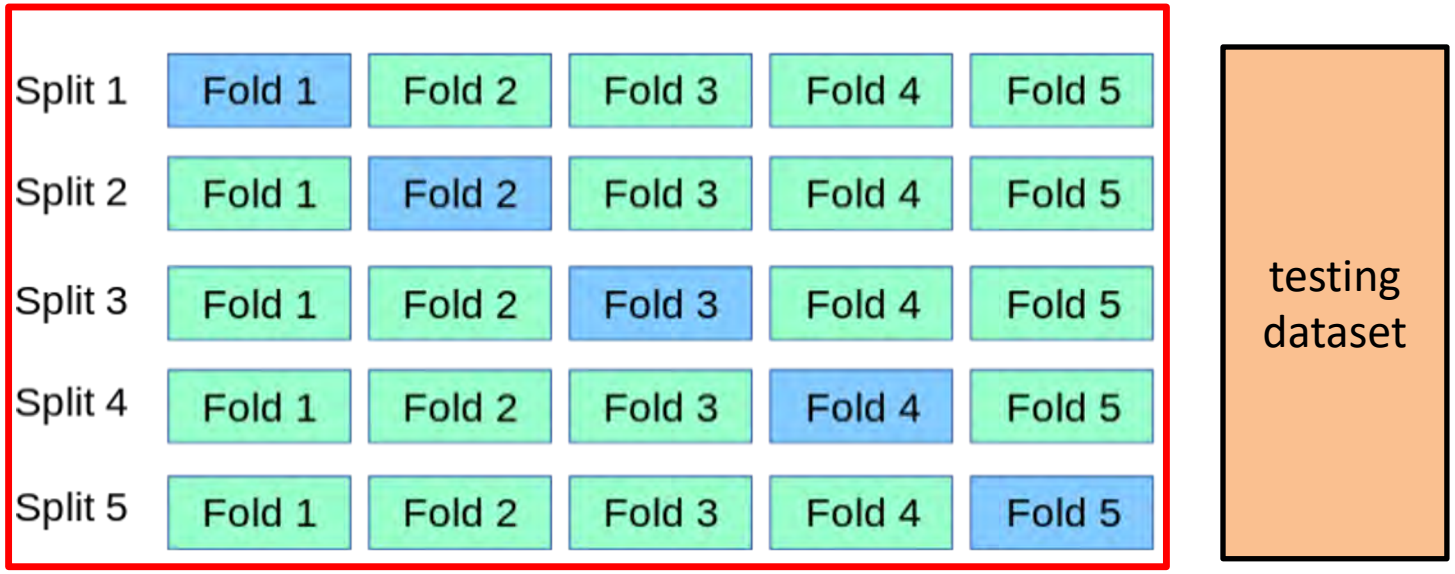
K-fold Cross-validation

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



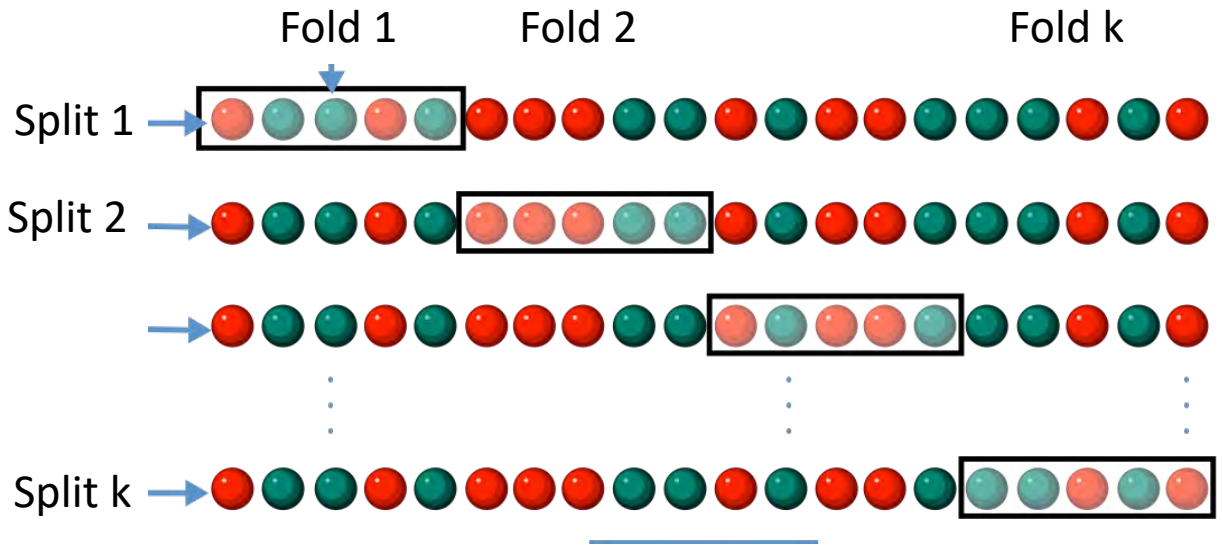
K-fold Cross-validation

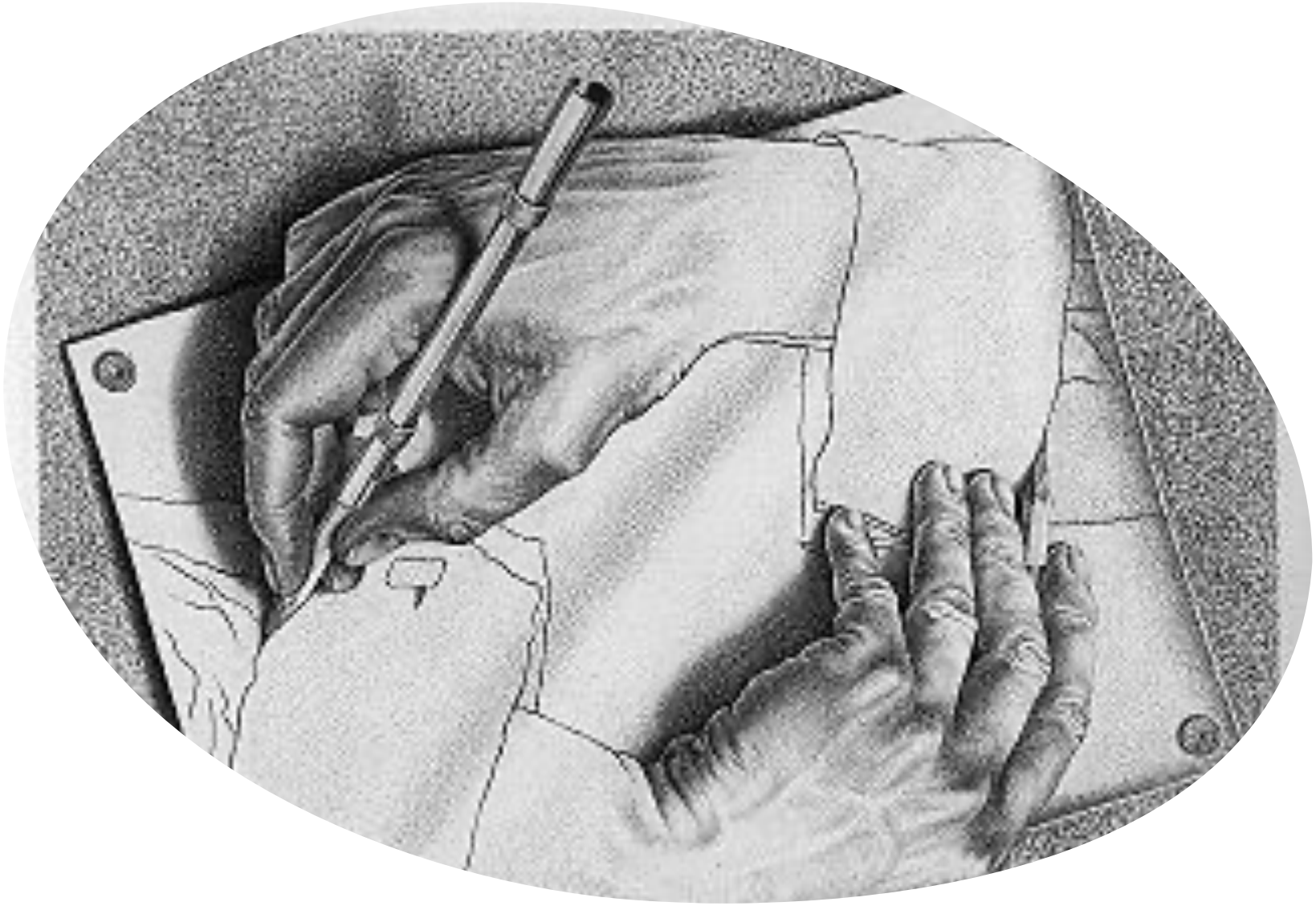
- Not a substitute for summative assessment
- Retrain using ALL of the training (development) set
- Test using the separate **summative** dataset



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?

Do you measure what matters ?

Are you measuring performance directly, or indirectly?

- Are the tests asking the question that matters, or
- a proxy question?

Is there ground truth or not?

- Recommender systems, vs.
- Martian soil classification

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 Tuesday (June 22), 11:59 pm (Central)

Your assignment is to create a Jupyter notebook that demonstrates how to do the following (use methods discussed in the class materials shared so far):

1. Load the dataset in the file named BDOShoham.csv and produce at least one table and one graph that summarize the dataset statistics; (4 points)
2. Set up a classification problem: predicting the FlowPattern value based on the values of the variables named Vsl, Vsg, and Ang. Train at least two models (e.g., k-NN, logistic regression) to solve this classification problem; (4 points)
3. Evaluate each model's performance using cross-validation on the training set you created; report on at least two different scoring methods (e.g., confusion matrix, weighted precision, macro recall, f1 score); (4 points)
4. Modify at least two hyperparameters (e.g., n_neighbors, weights, metric, penalty) and describe the improvement/degradation of a model's performance compared to its default settings; (4 points)
5. Test the performance of the best model+hyperparameters combination using the test set you created. Discuss your overall results (4 points)