Assessment 3

High-Level Topics overview

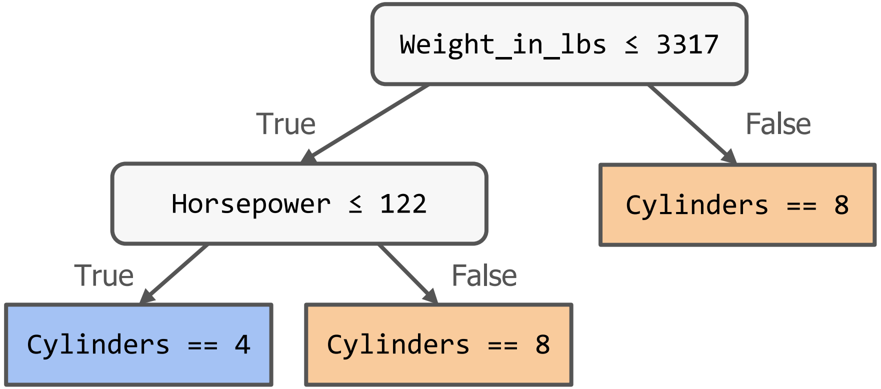
Unsupervised vs Supervised learning

- **Unsupervised learning** is when our data consists of examples (rows) and features (columns). It is the broad task of describing how our data is organized.
- **Supervised learning** is when our data consists of examples and features, as well as outcomes (labels) for each example. Broad tasks are classification and regression.

Focus of Assessment 3

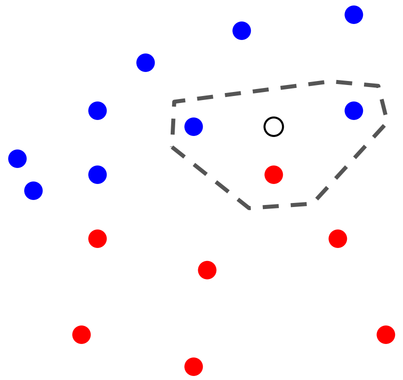
Recap: Decision Trees

- Labels must be categorical (this is a classification task)
- Features can be categorical or numerical
- A decision tree is fit to the data. We can specify depth to control complexity of the tree.



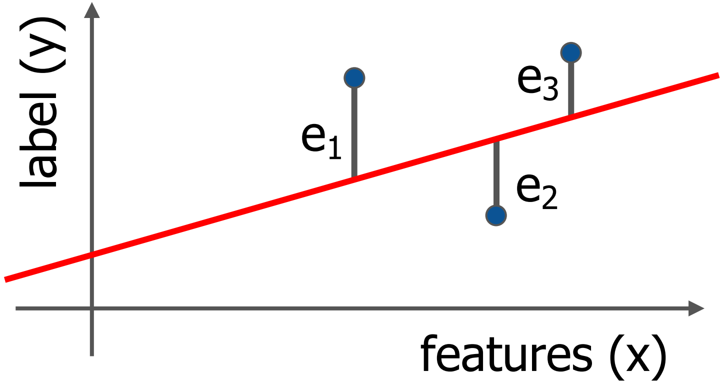
Recap: K-Nearest-Neighbors

- Labels must be categorical (this is a classification task)
- Features must be numerical (continuous values)
- A new point is classified based on a majority vote among the K nearest neighbors to the point.



Recap: Linear Regression

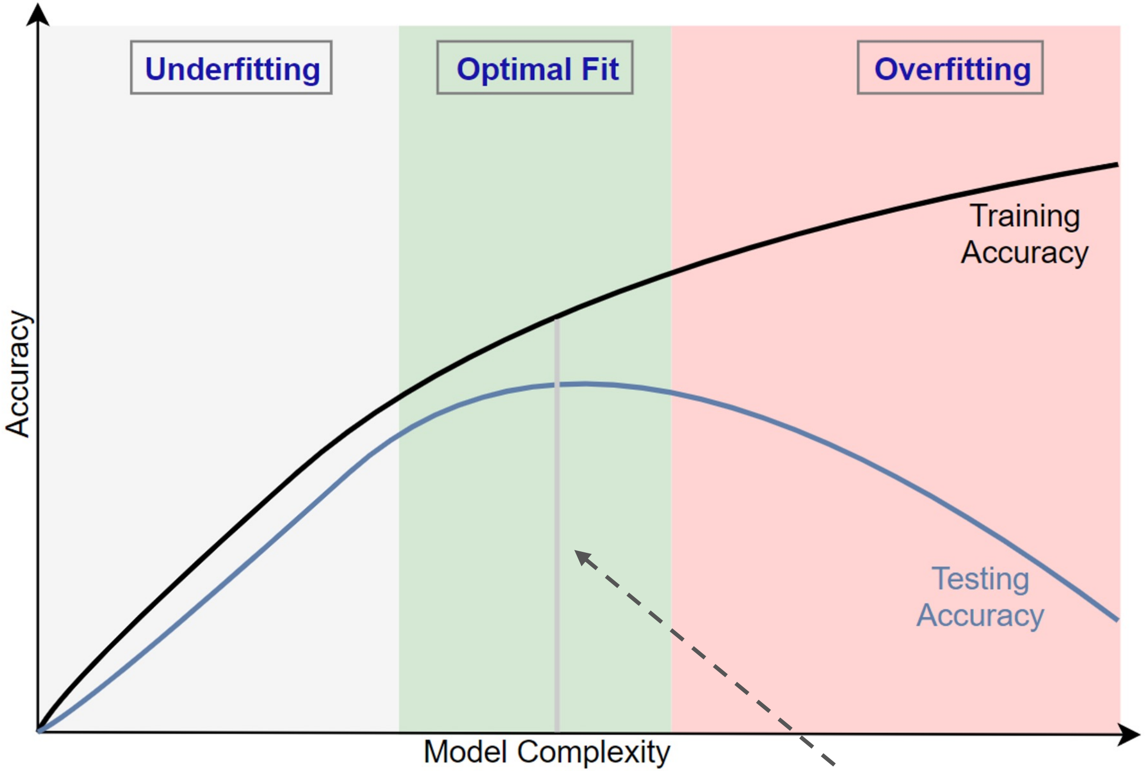
- Labels must be numerical (this is a regression task)
- Features must be numerical
- A line (or plane) of best fit is drawn through the data to minimize the sum of squared residuals (RSS).



Recap: Flavors of Regression

- Linear regression (with a single or multiple features)
- Polynomial regression
  - Nonlinear model (with respect to the original feature(s))
  - Still linear regression (with respect to the polynomial features!)
- Autoregression
  - Attempts to predict the future using past measurement
  - Still linear regression (with respect to the “lagged” features!)

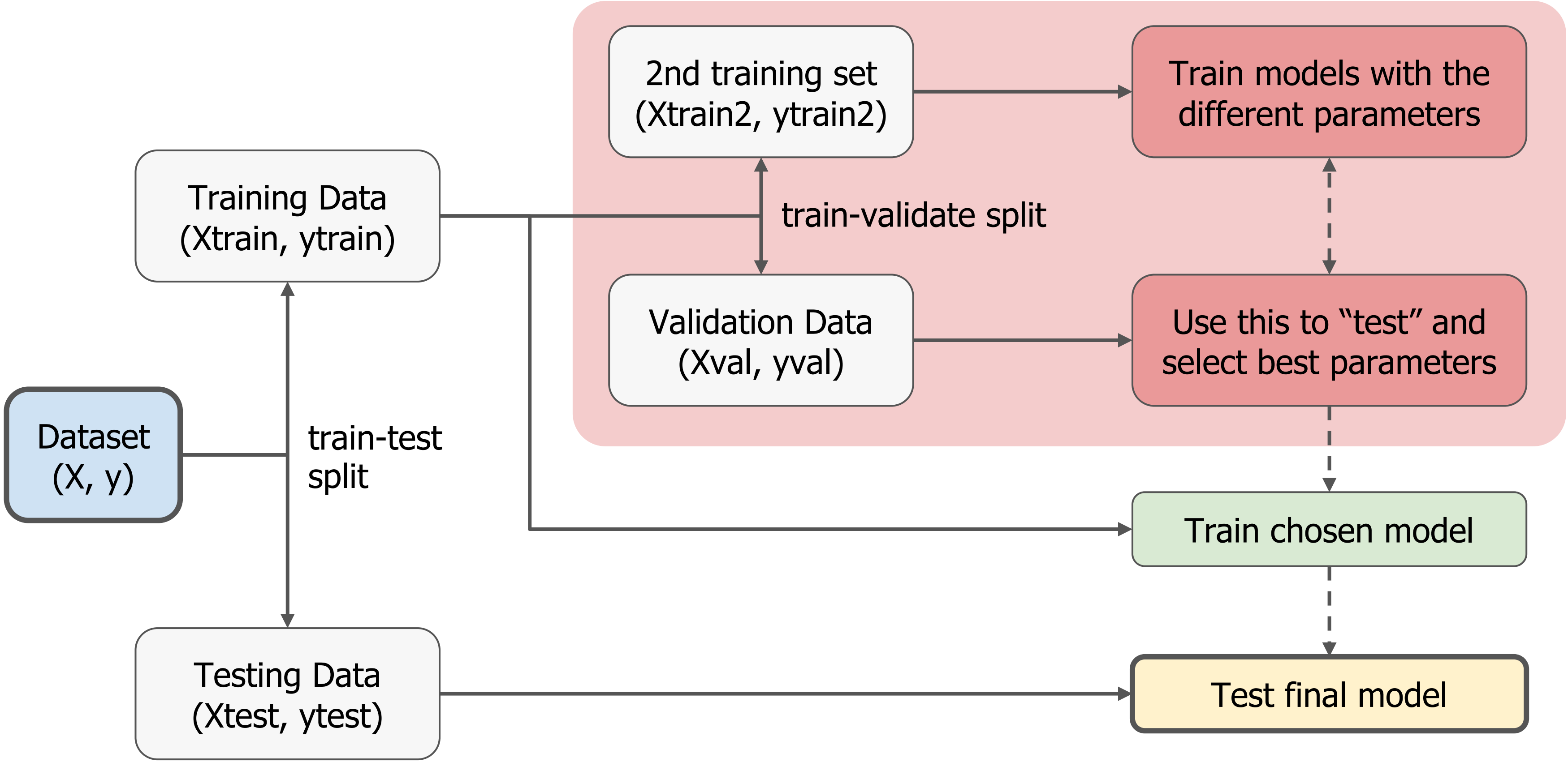
Accuracy-Complexity trade-off



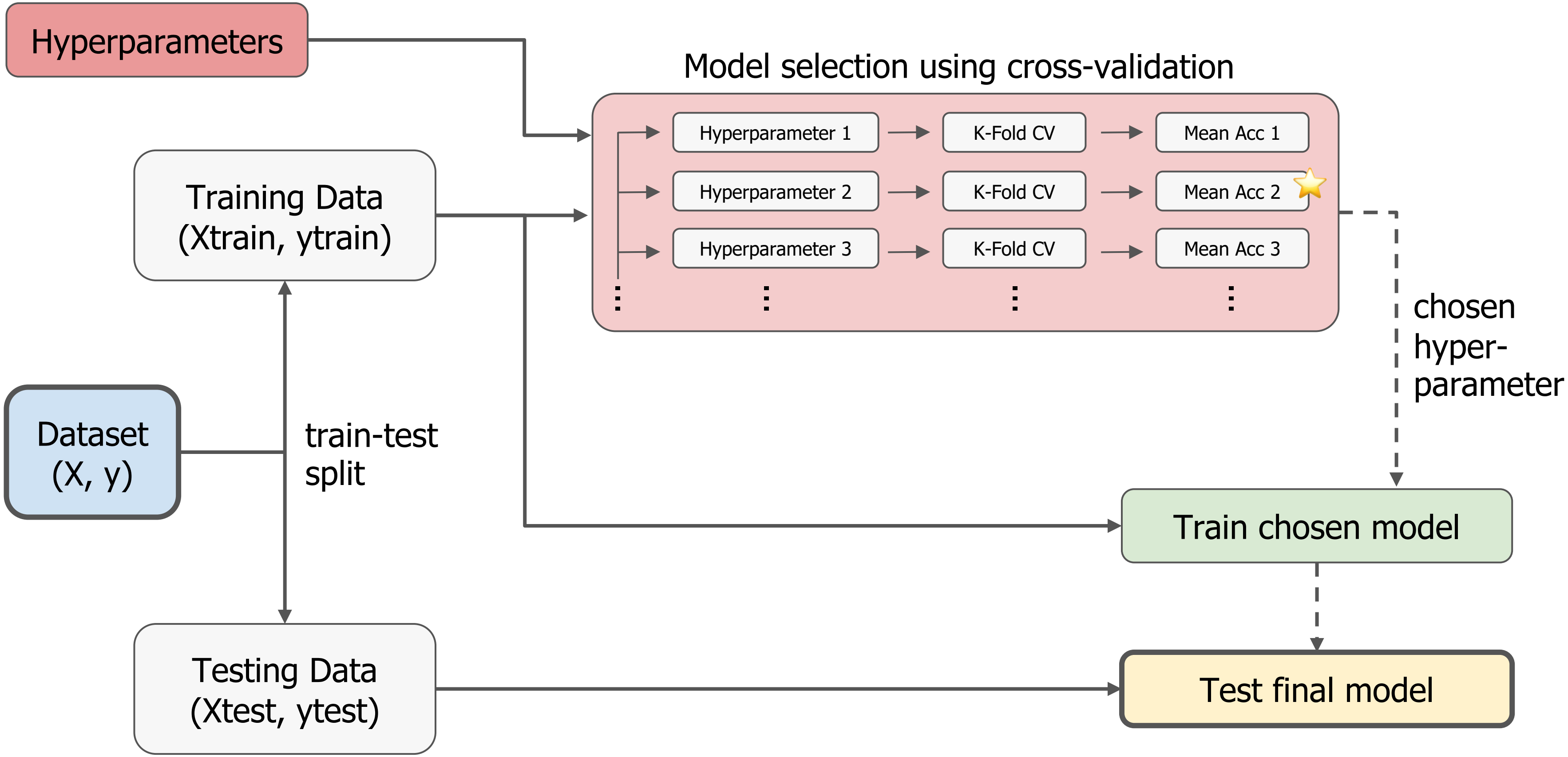
Model Complexity in our example is `max_depth` (more depth is more complex)

Ideal model complexity (best test accuracy)

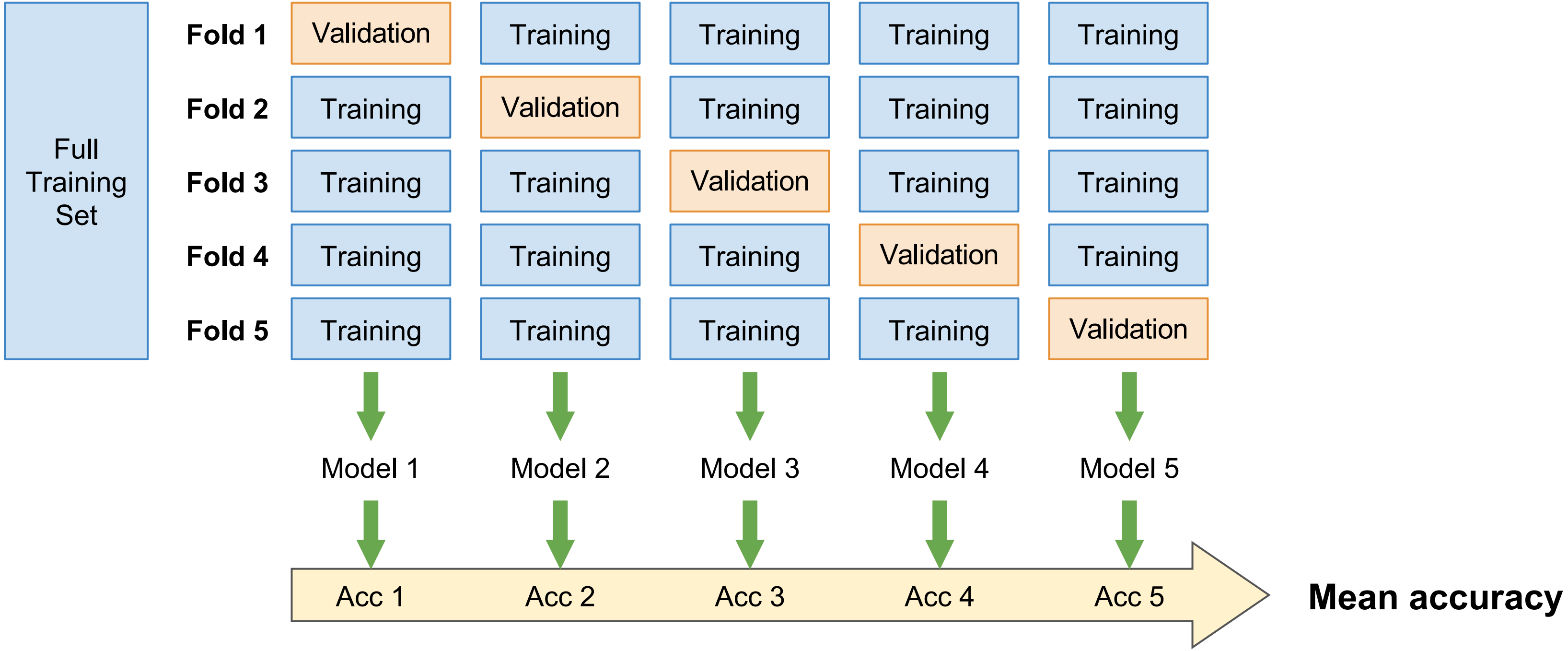
Simple Validation



Cross-validation



K-Fold cross-validation



Full Training Set

Fold 1	Validation	Training	Training	Training	Training
Fold 2	Training	Validation	Training	Training	Training
Fold 3	Training	Training	Validation	Training	Training
Fold 4	Training	Training	Training	Validation	Training
Fold 5	Training	Training	Training	Training	Validation

Mean accuracy

Other topics

- Time Series
  - Data collected at time instances
  - Needs special treatment since the order matters
- Numpy
  - Various useful methods for manipulation of arrays
- Bias in Data
  - Very broad and complex topic.
  - Examples: Selection, Survivorship, Aggregation, Algorithmic