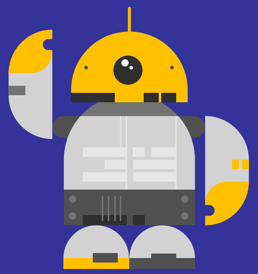


Database Systems

Lecture 4 – Machine Learning for Data Analytics

Dr. T. Akilan

takilan@lakeheadu.ca

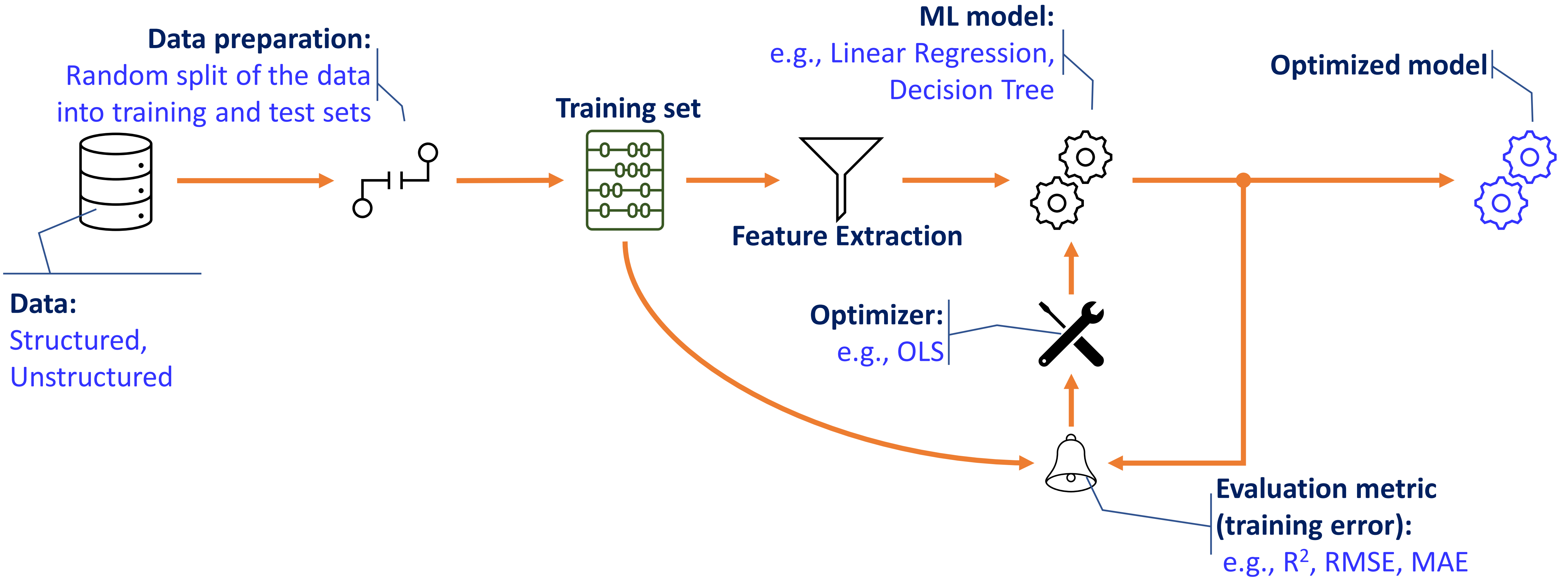
Welcome back

- ❑ Project Stage 2
- ❑ Assignment 1
- ❑ Pop quiz

This Session

- General ML Modeling Phases
  - Overfitting
  - K-fold cross validation (CV)
- Regression
  - Linear regression
    - Ordinary least square
- Classification
  - Logistic regression
    - Maximum likelihood estimation
- Evaluation metrics
  - ROC
  - AUC

General ML Model Training Phases



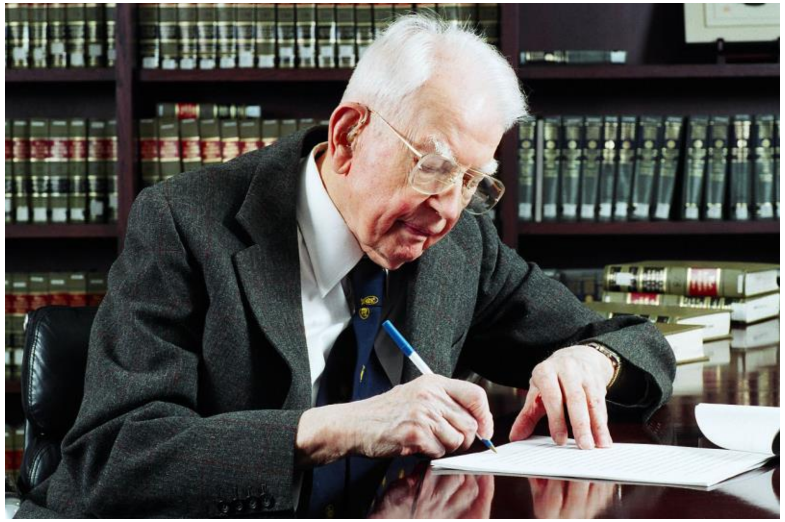
General ML Model Training Phases Cont.

- Supervised model training can be broken down into three steps:

1. **Representation**
   - ➢ Model is exposed to data (generally, large data)
   - ➢ Model transforms the input data into the desired results
   - ➢ Model learns the relationship between the raw data and which data points are strong predictors for the desired outcome
2. **Evaluation**
   - ➢ Score the models based on the predictions.
   - ➢ Make sure the **model is not overfit** to training data
3. **Optimization**
   - ➢ The model must be optimized to perform on more diverse sets of input data
   - ➢ Select the most generalized model.

Model Overfitting and Its Avoidance

- “If you torture the data long enough, it will confess to anything.”



Ronald Coase - Nobel Prize in Economics, 1991
(Image from https://en.wikipedia.org/)

Model Overfitting and Its Avoidance Cont.

- **Problem:** Model using pure **memorization**
  – possibility of a **most extreme overfitting** procedure

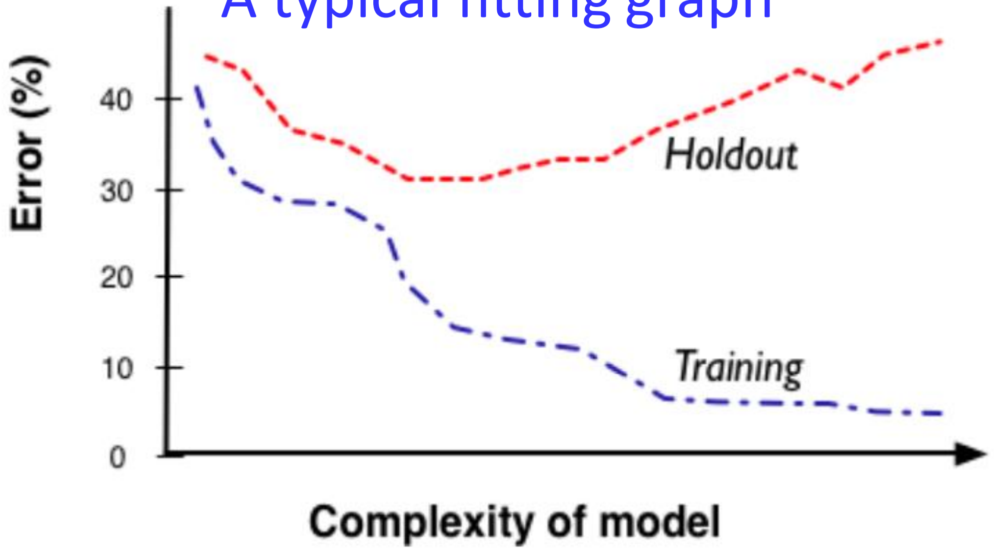
- **Reality:** All ML models have the tendency to overfit to certain degree—some more than others

- **Solution:**
  - There is no single choice or rule of thumb that can eliminate overfitting.
  - The best strategy is to recognize overfitting and manage complexity in a principled way

Model Overfitting and Its Avoidance Cont.

- **How to recognize it.**
  - Holdout Data and Fitting Graphs
  - Hold out data –
    - ✓ The value of the target variable is known
    - ✓ It will not be used to build the model (hidden true values)
    - ✓ The model will predict the outcomes for all the data points in the holdout set
    - ✓ Generalization performance is computed by comparing the predicted values with the hidden true values

A typical fitting graph

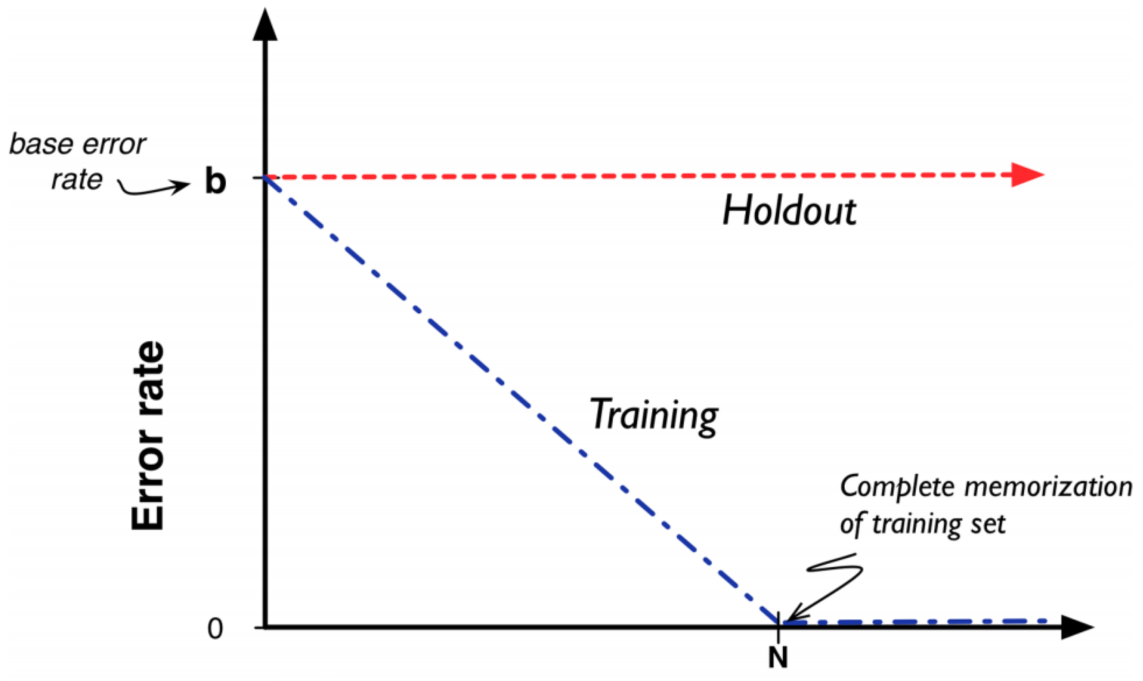


Observation:

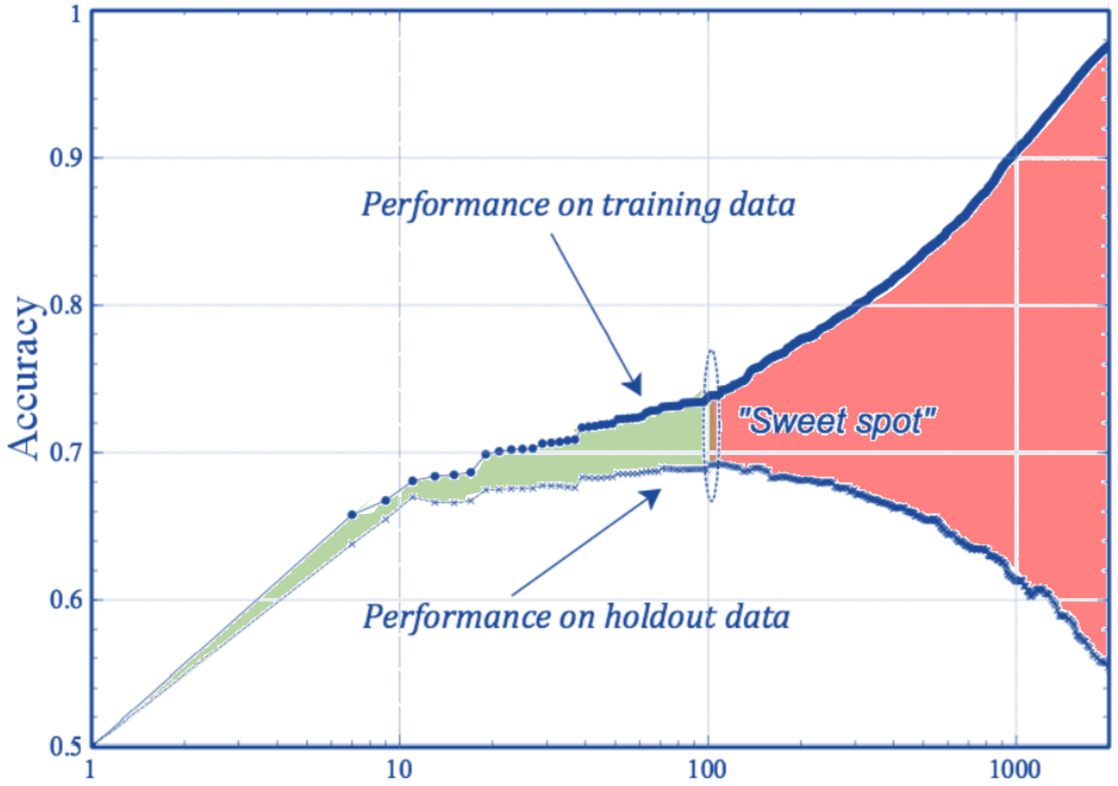
- As the models get too complex, they look very accurate on the training data.
- However, they are **overfitting**, as the training accuracy diverges from the holdout (generalization) accuracy.

Recognizing Model Overfitting

- **How to recognize it.**

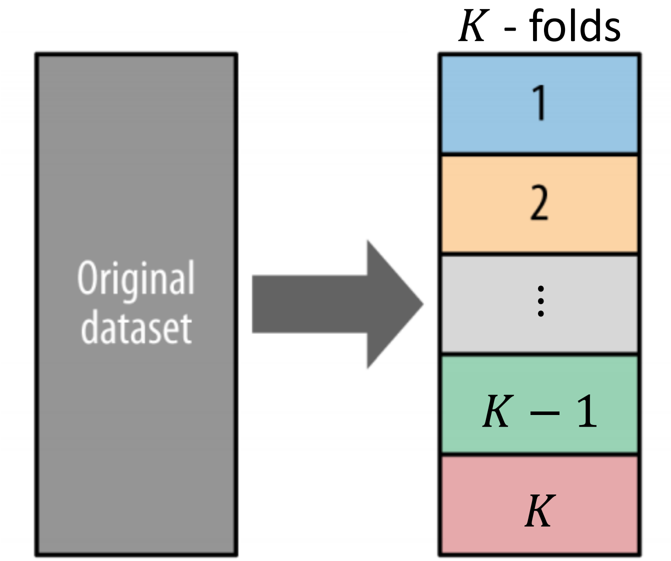


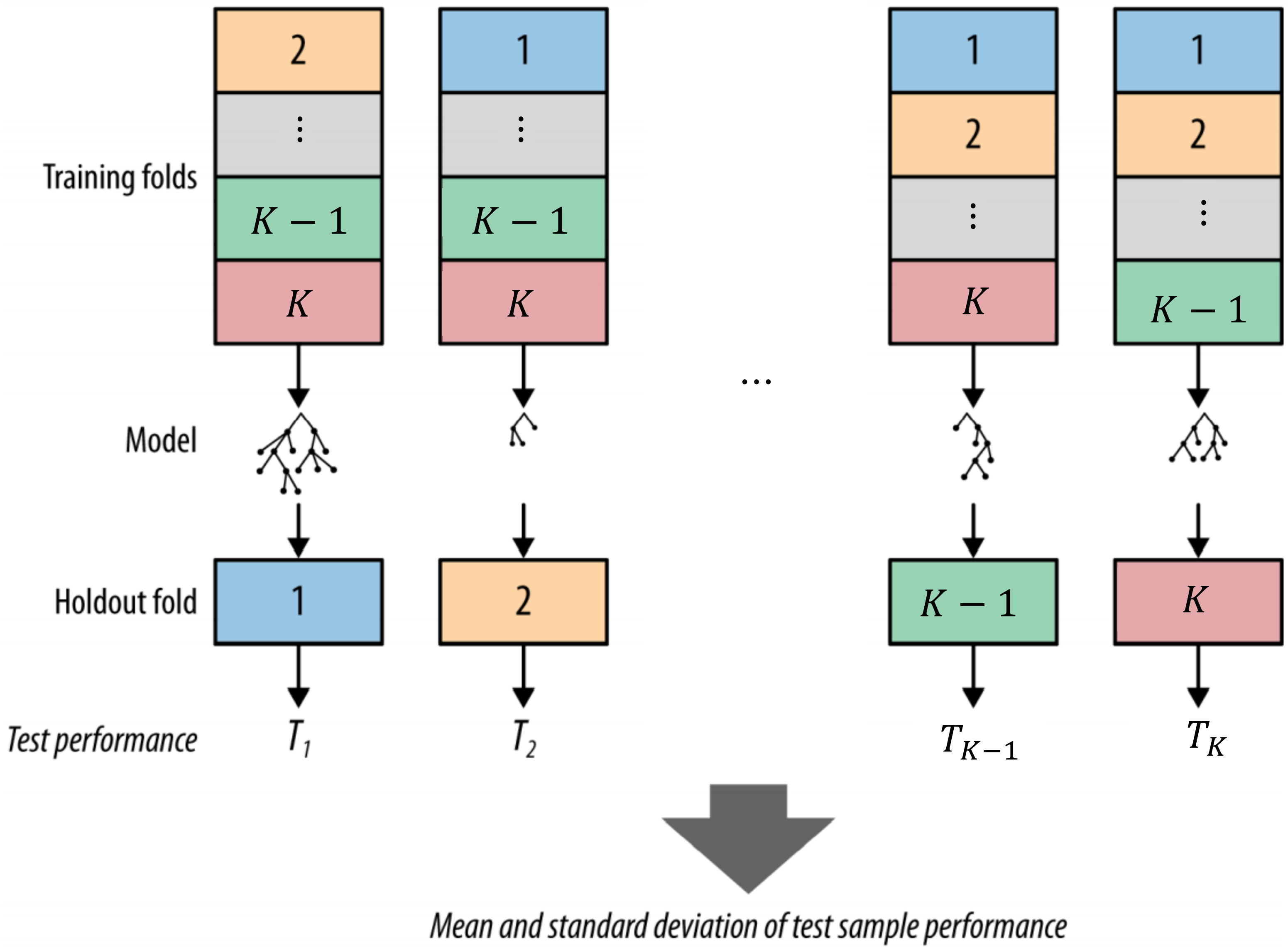
Number of Training Samples



Model's Complexity (for e.g., wrt number of nodes)

Recognizing Model Overfitting– Cross Validation





- Splitting a labeled dataset into k partitions called folds
- Each iteration, $(k-1)/k$ portion of the data used for training and $1/k$ used for testing

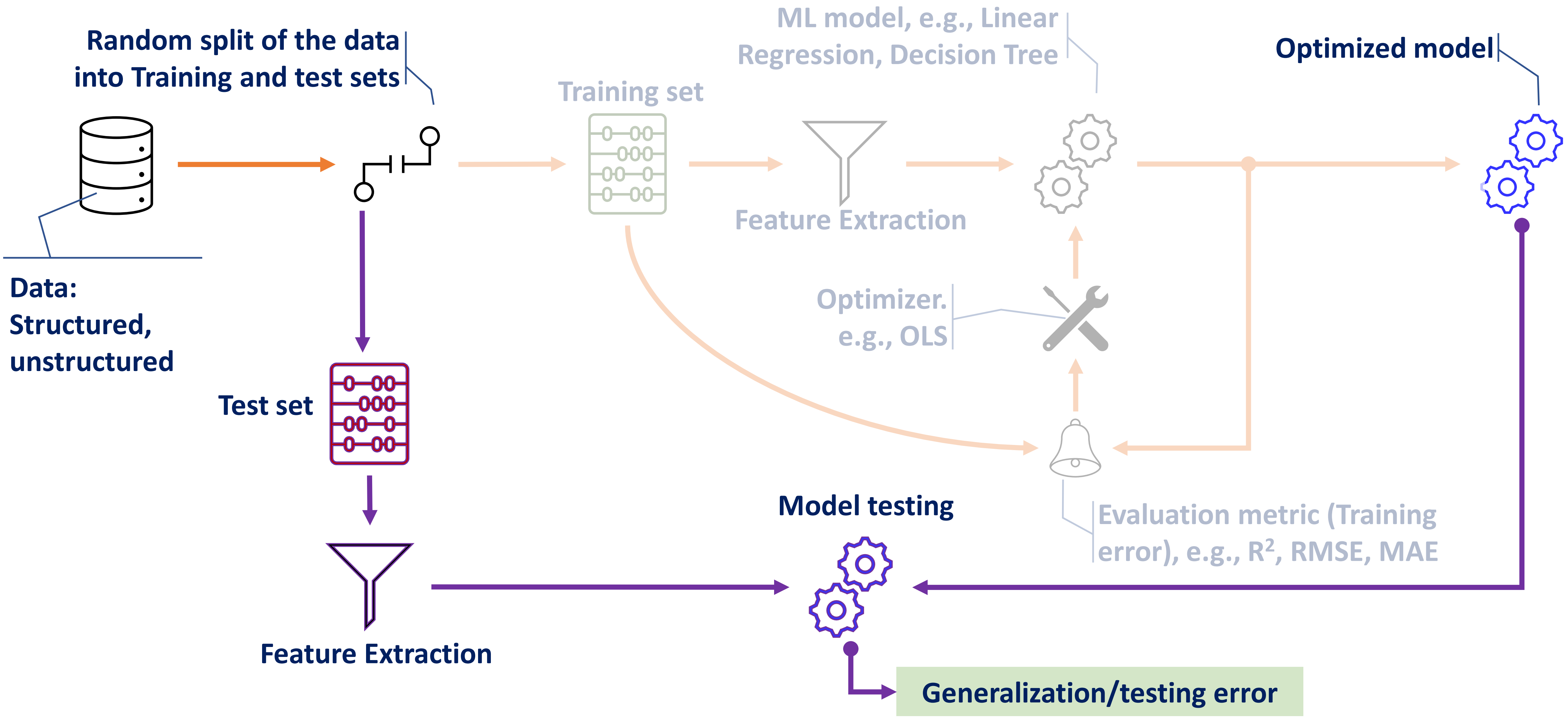
Model Overfitting and Its Avoidance

- The general approaches:
  - K-fold Cross validation

  - Attribute regularization
    - Key feature selection and removal of irrelevant features
    - Data augmentation

  - Model regularization
    - Network pruning (reduce network size when it has become too large)
    - Weight regularization
    - Adding dropout layers
    - Employing explicit complexity penalties into the objective function

General ML Model Testing Phases



General ML Models

Sample Real-world Problems	Type	Applicable ML Models
Need to determine the **relationships between (numerical) data points**	**Regression**	**Linear Regression, Logistic Regression**
Need to assign (known) **labels to (un**known) **objects**	**Classification**	**Naïve Bayes, Decision Trees**
Need to **group items** by **similarity** or **commonalities** in the attributes	**Clustering**	**K-means clustering**
Need to discover **relationships between actions** or items	**Association**	**Apriori**
Need to analyze **text** reviews or **corpus**	**NLP**	**BoW, TF-IDF, RNN**

Linear Regression

Regression

- The focus is on the **relationship between outcome(s)** and **its input variable(s)**.
  - Not only for outcome prediction, but also reasoning how **changes** in **individual drivers affect** the outcome.
  - Multiple input variables → Multivariate Regression.

- The outcome can be **continuous** or **discrete**.
  - **Discrete** - predicting the probability that the outcome will occur.
  - **Continuous** – estimating the real value of output value(s).

- **Example**: Regression analysis is useful for answering the following kinds of questions:

What is a person's expected income?

Will an applicant default on a loan?

Linear Regression

- Used to **estimate** a **continuous** value as a **linear** (additive) **function** of other variables

  - `Income = f(years of education, age, gender)`
  - `House = f(median home price in neighborhood, square footage, # of bedrooms/bathrooms)`
  - `Treatment effect = f(duration of radiation, frequency of radiation, patient attributes)`

- Input: variables can be continuous or discrete.
- Output:
  - A **linear expression** for predicting outcome as a function of drivers.
  - A **set of coefficients** that indicate the relative impact of each driver.

Linear Regression Cont.

- The preferred method for almost any problem, where we are **predicting a continuous outcome**

Try this first; if it fails, then try something more **complicated** models.

- **Example:** linear relationship between the **accidents in a state and the population of the state**.

- The fitted model (i.e., best fit line) represents the relation:
  - $y = \beta_0 + \beta_1 x$
    $= 142.7120 + 0.0001256x.$

Image credit: https://www.mathworks.com/

Linear Regression: Model Description

- **Assumption:** a **linear relationship** between the **input** variables and the **outcome** variable.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1} + \varepsilon$$

Dataset

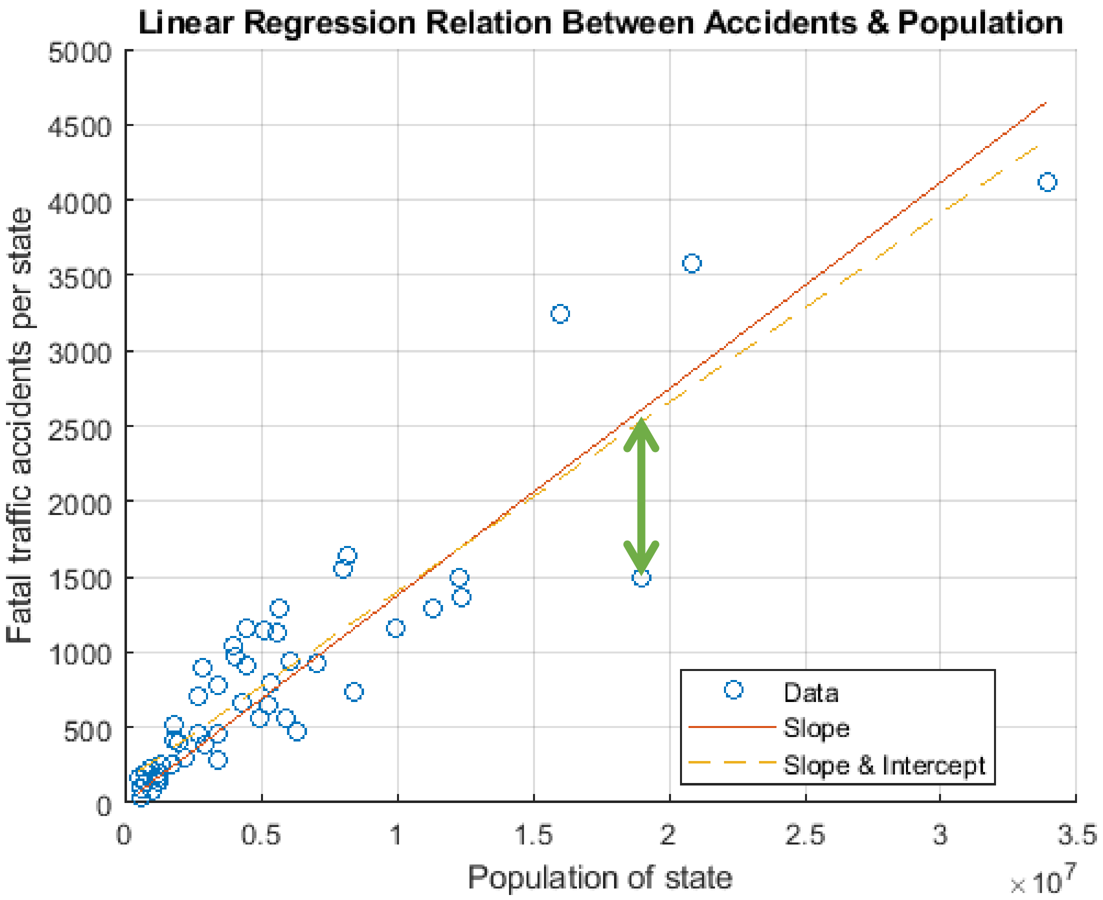
	F_1	F_2	...	F_p
S_1				
S_2				
$\vdots$				
S_N				

y - outcome variable (dependent variable).
x_j - input (independent) variables, for $j = 1, 2, \ldots, p - 1$,
(given that the dataset has P number of features)
β_0 - value of y when each x_i equals zero.
β_j - change in y based on a unit change in x_i, for $j = 1, 2, \ldots, p - 1$.
ε - a random error term that represents the difference in the model and a particular observed value for y.

- The estimates for these **unknown parameters** are chosen so that, on average, the model provides a reasonable estimate of y based on the input X.
- i.e., the **fitted model** should **minimize** the overall **error between** the **linear model** and the **actual observations**.
- **How do we find the unknown parameters?**
  - ➢ **Ordinary Least Squares** (OLS) is a common technique to estimate the parameters.

Linear Regression: Model Description Cont.

- **Ordinary Least Squares (OLS)**



- The vertical lines represent the distance between each observed y value and the linear regressor line.

- The n individual distances to be squared and then summed.
- $\sum_{i=1}^{n}\left[y_i - (\beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1})\right]^2$
- $S(\beta) = \sum_{i=1}^{n}\left|y_i - \sum_{j=1}^{p-1} x_{i,j}\beta_j\right|^2$

- **Objective:** find the optimal values for the parameters $\beta_j' s$ that **minimize** the sum of **squared errors** (SSR) or **residual** sum of squares (RSS)).