

Machine Learning Concepts

1.2

Quiz

Course assignments and exam

Grading:

- 45% Weekly Assignments (homework exercises). These will be submitted using Moodle.
- 45% Final Exam
- 0% Competition results
- 10% Quizzes at the beginning of each lecture, except the first lectures of each course.
Missing a lecture or being late will result in 0% for the actual quiz score.

Weekly assignment acceptance policy and achievable grades:

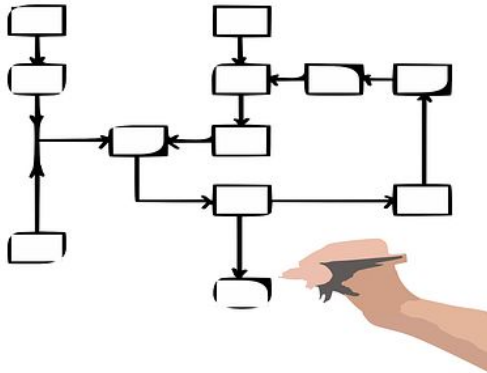
- 100% until due date
- 50% within 24 hours past due date
- 0% after that.

```
install.packages ( ' caret ' )  
install.packages ( ' party ' )
```

Topics from last week

- Process
- Visualization
- Supervised vs. Unsupervised
- Definitions of Machine Learning, why to study
- “AI is different”
- Tools, languages
- Reproducible research
- Exploratory data analysis

Topics from last week



Data Mining Process



Exploratory Data Analysis



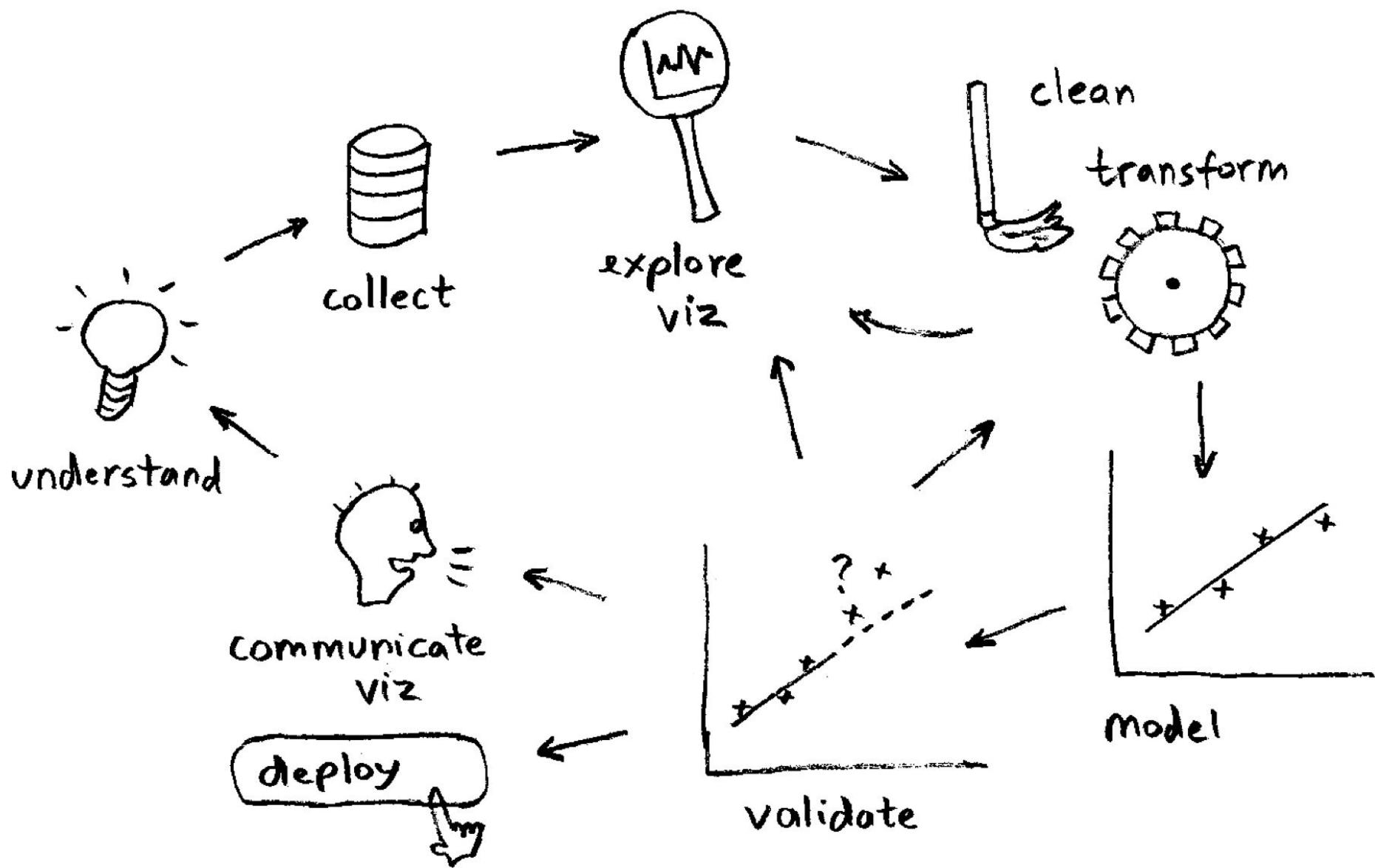
Data Visualization

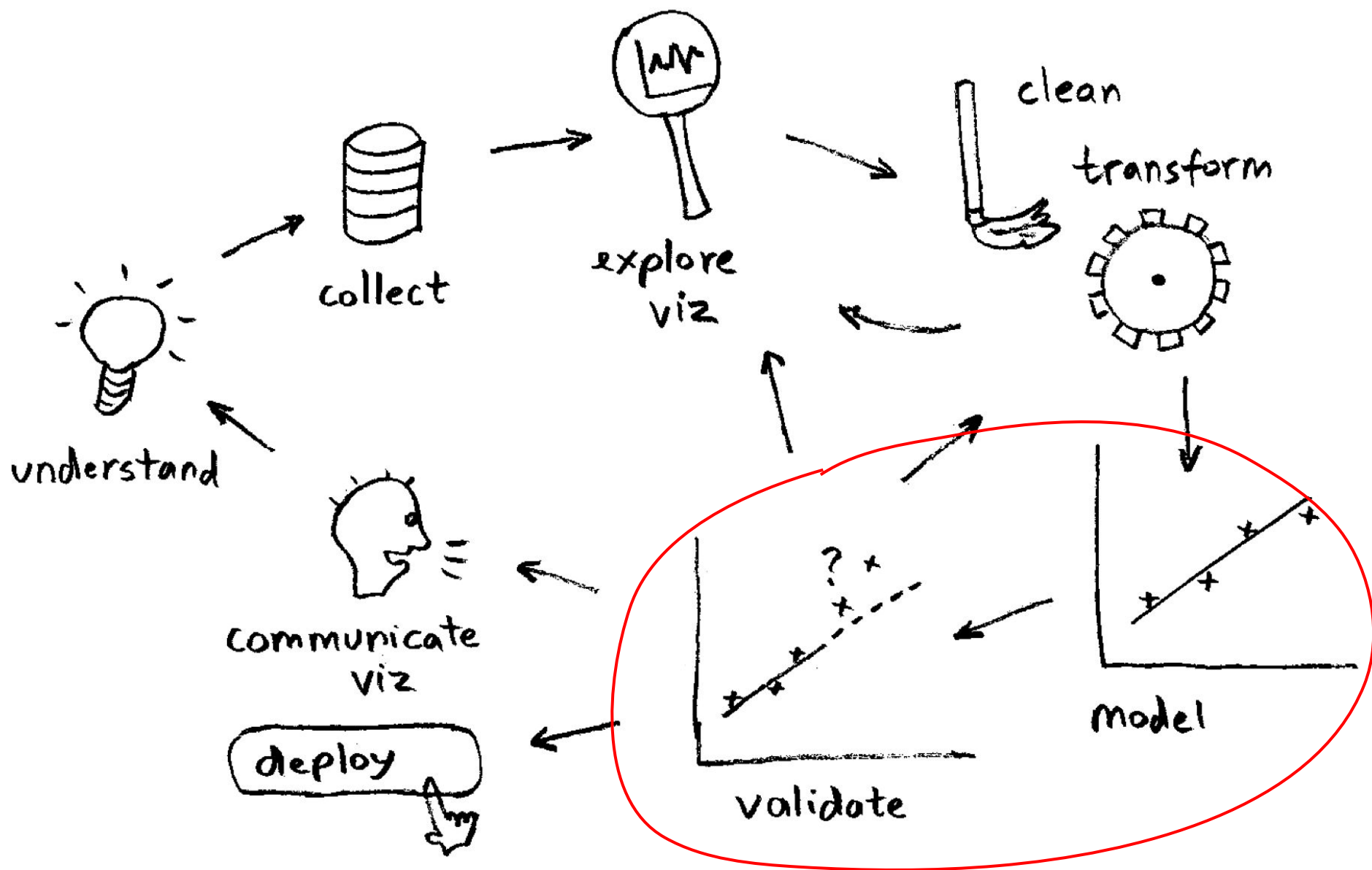


Data Science Tools



Reproducible Research





Build a Model by Hand

Bike Sharing Demand

Data Fields

datetime - hourly date + timestamp

season - 1 = Q1, 2 = Q2, 3 = Q3, 4 = Q4

holiday - whether the day is considered a holiday

workingday - whether the day is neither a weekend nor holiday

weather

1. Clear, Few clouds, Partly cloudy, Partly cloudy
2. Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
3. Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
4. Heavy Rain + Ice Pellets + Thunderstorm + Mist, Snow + Fog

temp - temperature in Celsius

atemp - "feels like" temperature in Celsius

humidity - relative humidity

windspeed - wind speed

casual - number of non-registered user rentals initiated

registered - number of registered user rentals initiated

count - number of total rentals

https://github.com/pappzoltan

 pappzoltan / machine-learning-course

forked from szilard/teach-ML-CEU-master-bizanalytics

<> Code

 Pull requests

 Actions


 Projects

 Wiki

 Security

 Insights

 Settings

 master ▾

 4 branches

 2 tags

Go to file

Add file ▾

 Code ▾

- Lectures
 - a. Slides
 - b. Code samples
 - c. Data
- Labs
 - a. Code samples
 - b. Data

Coding Exercise: Bike Sharing Demand

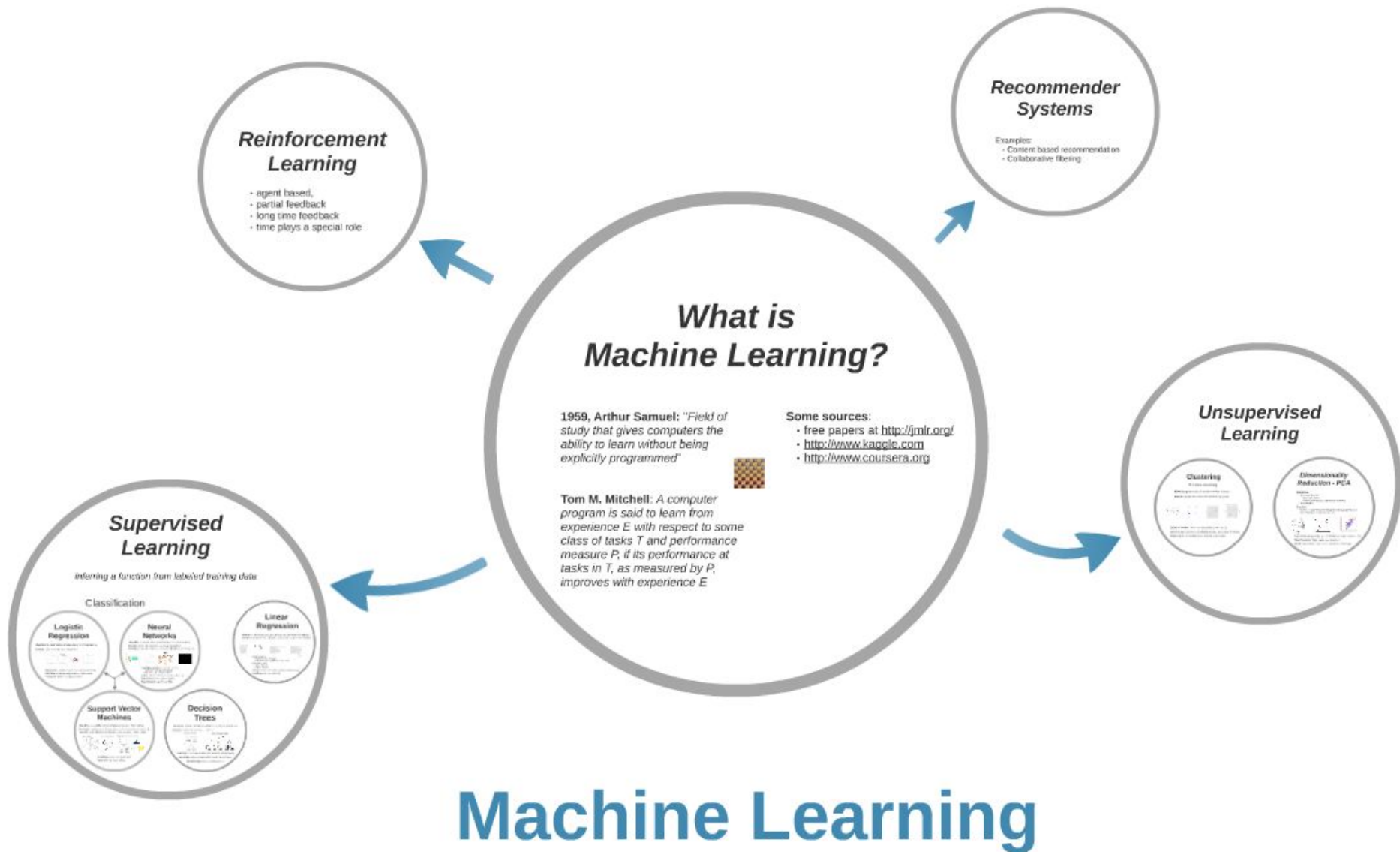
Exercise: build a model by hand, using the Bike Sharing Demand data.

Start with: `ml.1.1.2/lect/prediction_by_hand.R`



Conclusions

- We started with a **Benchmark Model**, and used it as a reference to the error level of a naive approach.
- We split the data into **Train Set** and **Test Set** for accurate error calculation
- **Train** and **Test Error** do not change together
- Increasing **Model Complexity** can reduce **Train Error** but can harm **Test Error**. This is called **Overfitting**.





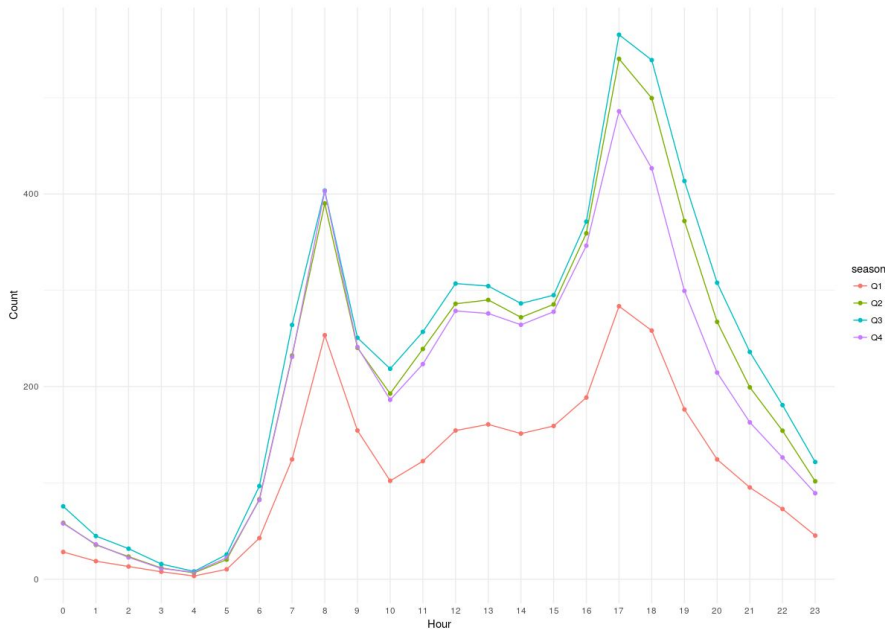
Supervised Learning?

Learning by Example

Introduction to supervised learning

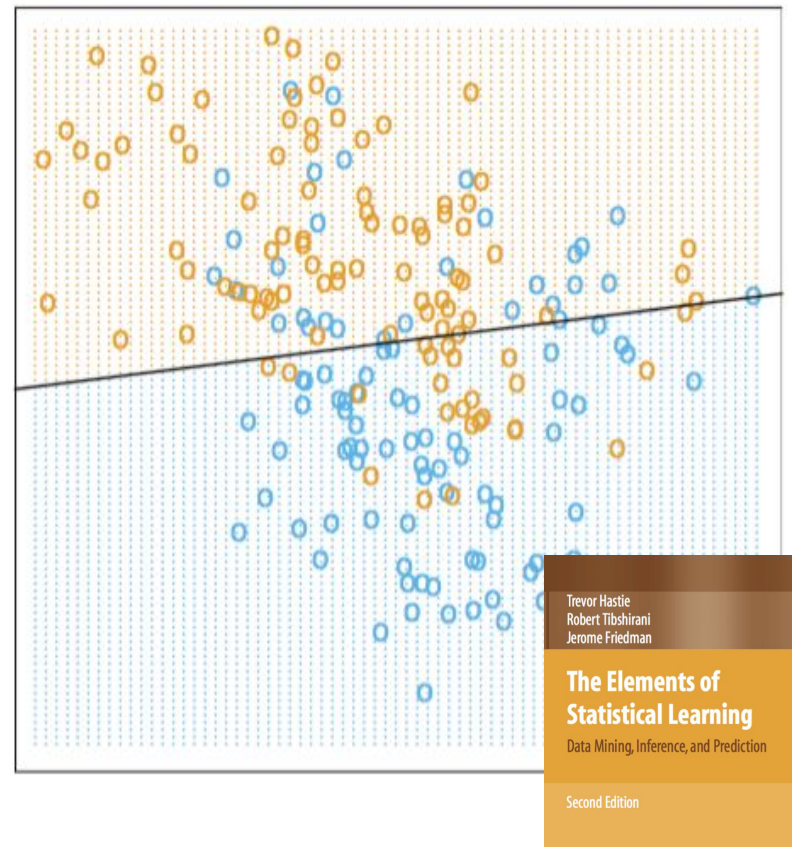
Regression

People rent bikes more in Fall, and much less in Spring.



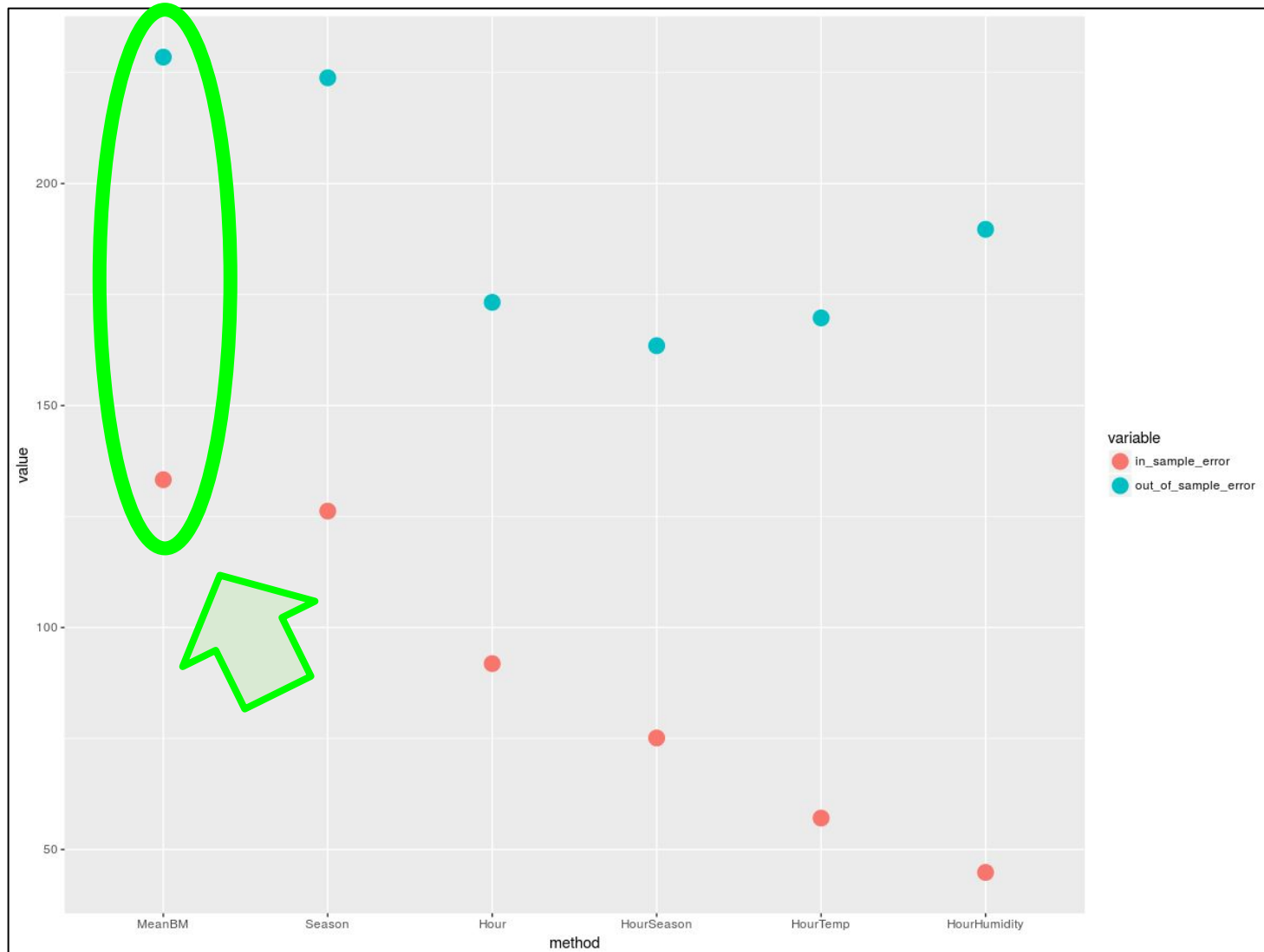
Classification

Linear Regression of 0/1 Response



Benchmark Selection

Benchmark Selection



A good Benchmark Model

Previous, available results

Results available:

- In the literature
- At your company
- etc.



A simple model

It should be:

- Simple
- Easy to understand
- Easy to calculate
- Intuitive

General guidelines:

- Regression: mean of test values for
- Classification: the most typical label

Example:

Prime number test, using input x :

- *"Is x a prime number?" - "NO!"*

Is correct for about 87% of the integers until 8000

Training, Validation and Test

Validation

Training dataset

A training dataset is a dataset of examples used for learning, that is to fit the parameters (e.g., weights) of a regressor or a classifier.

Validation dataset

A validation dataset is a set of examples used to tune the hyperparameters of a regressor or classifier. As well as the testing set, it should follow the same probability distribution as the training dataset.

Test dataset

A test dataset is a dataset that is independent of the training dataset, but that follows the same probability distribution as the training dataset.

The test dataset is used to obtain the performance characteristics such as accuracy, sensitivity, specificity, F-measure,

Cross vs. Causal Validation

Regular Machine Learning



Time Series Prediction



Cross-validation

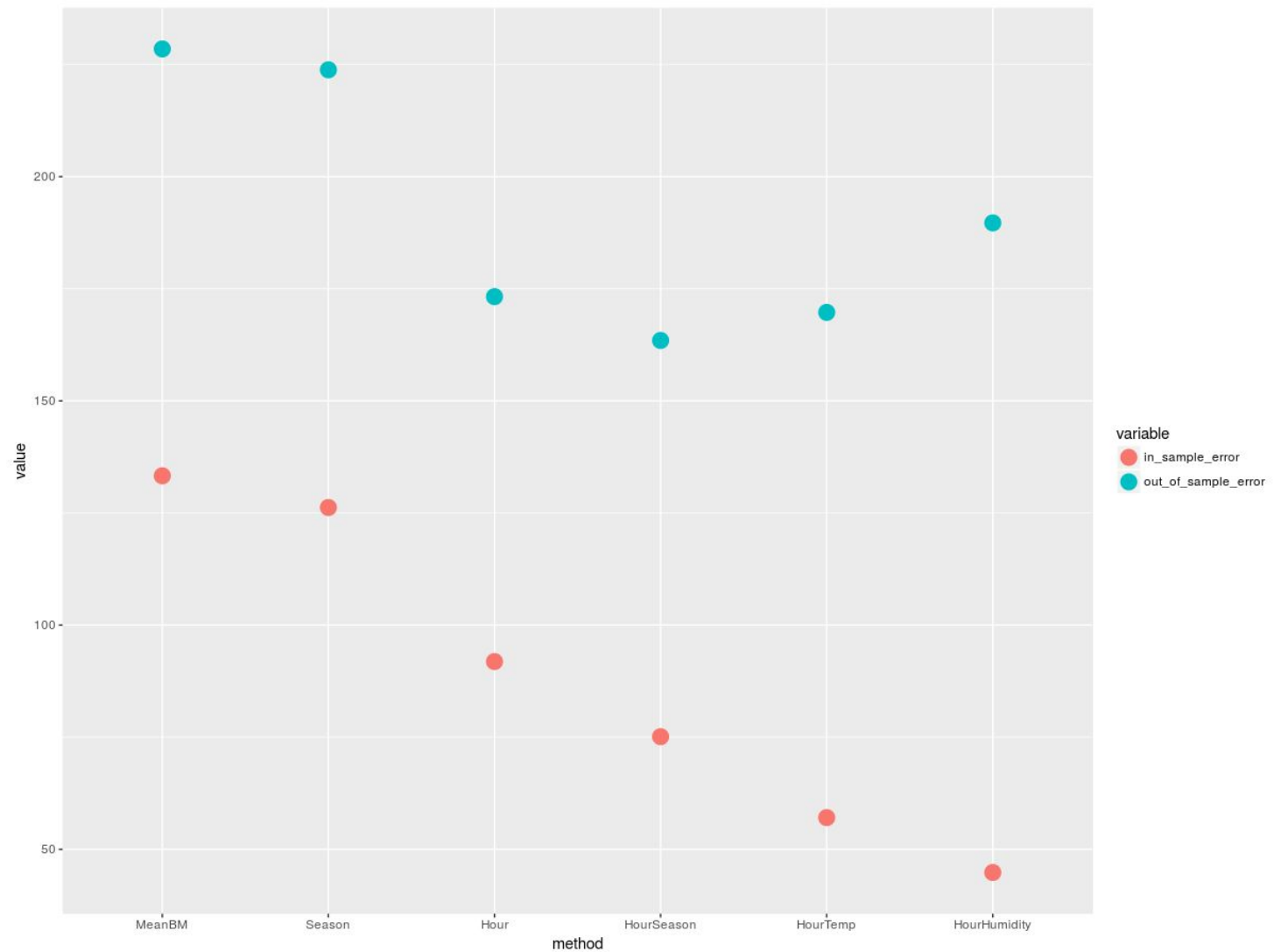
- Exhaustive cross-validation
 - Leave-p-out cross-validation: has excessive computation time due to the combinatorial explosion
 - Leave-one-out cross-validation: efficient and works well with little data
- Non-exhaustive cross-validation
 - K-fold cross-validation: randomly partitioned into k equal sized subsamples
 - Holdout method: simple validation, randomly partitioning data into train and validation set
 - Repeated random sub-sampling validation (Monte Carlo cross-validation)

Practical advice: small data for training

- Exhaustive cross-validation
 - **Leave-p-out cross-validation:** has excessive computation time due to the combinatorial explosion
 - **Leave-one-out cross-validation:** efficient and works well with little data
- Non-exhaustive cross-validation
 - **K-fold cross-validation:** randomly partitioned into k equal sized subsamples
 - **Holdout method:** simple validation, randomly partitioning data into train and validation set
 - **Repeated random sub-sampling validation (Monte Carlo cross-validation)**

Model Selection

Train Error and Test error



Flexibility / Power ↔ Interpretability



Model Selection Practice

There is usually a **inverse relationship** between model **flexibility/power** and **interpretability**. In the best case, we would like a parsimonious and interpretable model that has excellent performance. Unfortunately, that is not usually realistic.

One strategy:

1. start with the most powerful black-box type models
2. get a sense of the best possible performance
3. then fit more simplistic, understandable models
4. evaluate the performance cost of using a simpler model

Model Predictive Power

Regression

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2.$$

Mean Squared Error

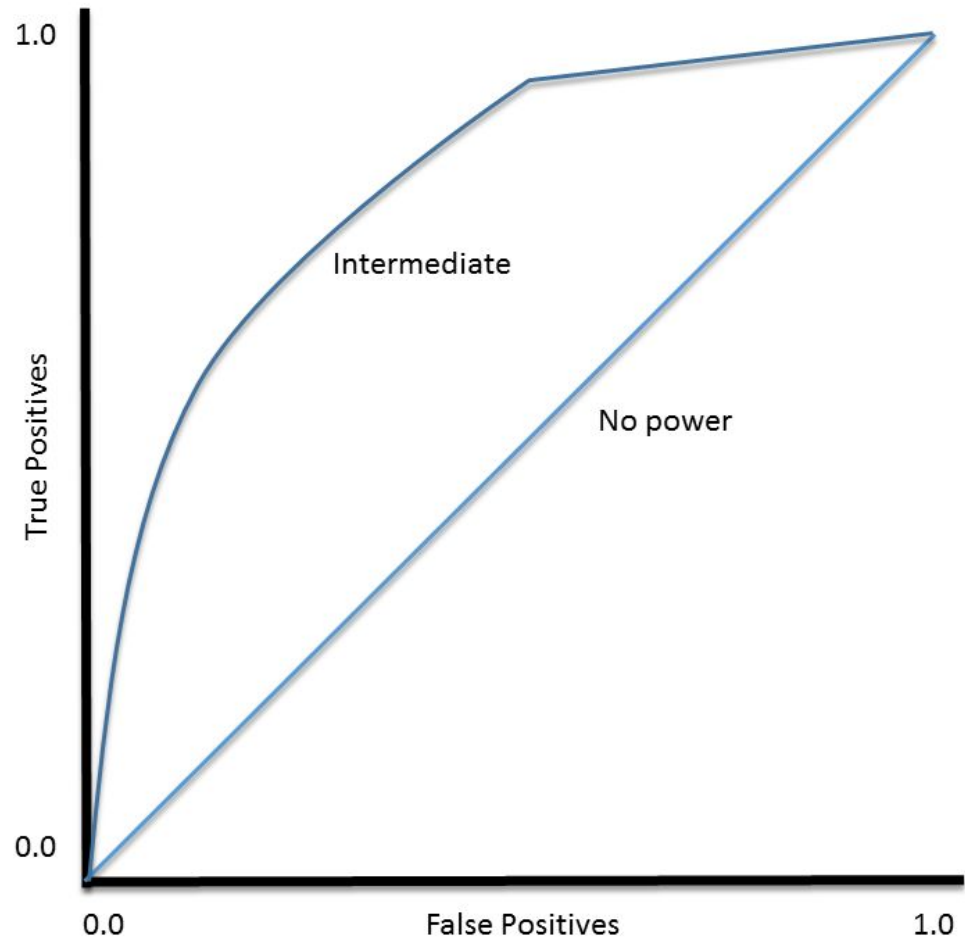
Classification

		Predicted class	
		<i>P</i>	<i>N</i>
Actual Class	<i>P</i>	True Positives (TP)	False Negatives (FN)
	<i>N</i>	False Positives (FP)	True Negatives (TN)

Confusion Matrix

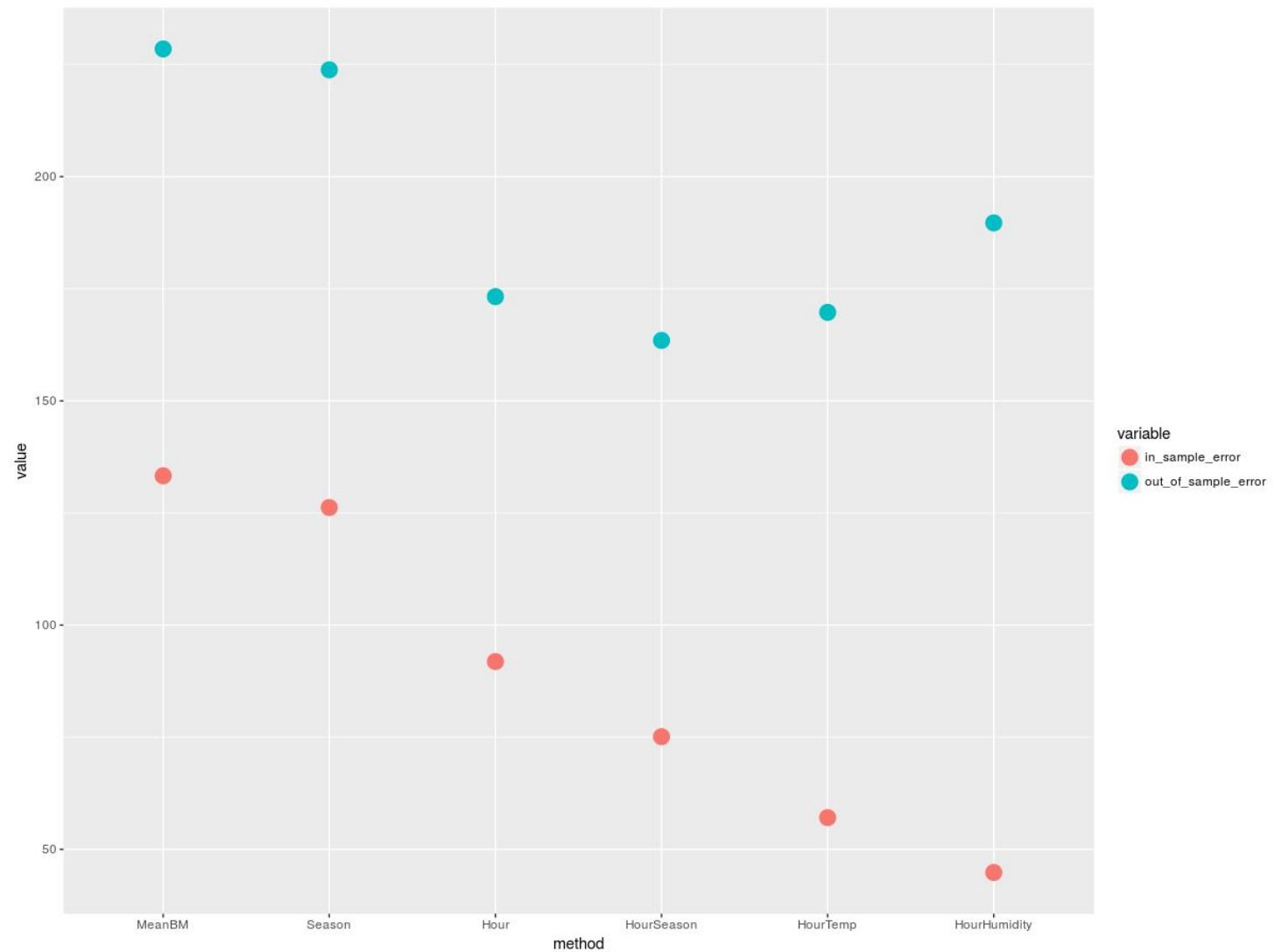
Classification ROC / AUC

		Predicted class	
		<i>P</i>	<i>N</i>
Actual Class	<i>P</i>	True Positives (TP)	False Negatives (FN)
	<i>N</i>	False Positives (FP)	True Negatives (TN)



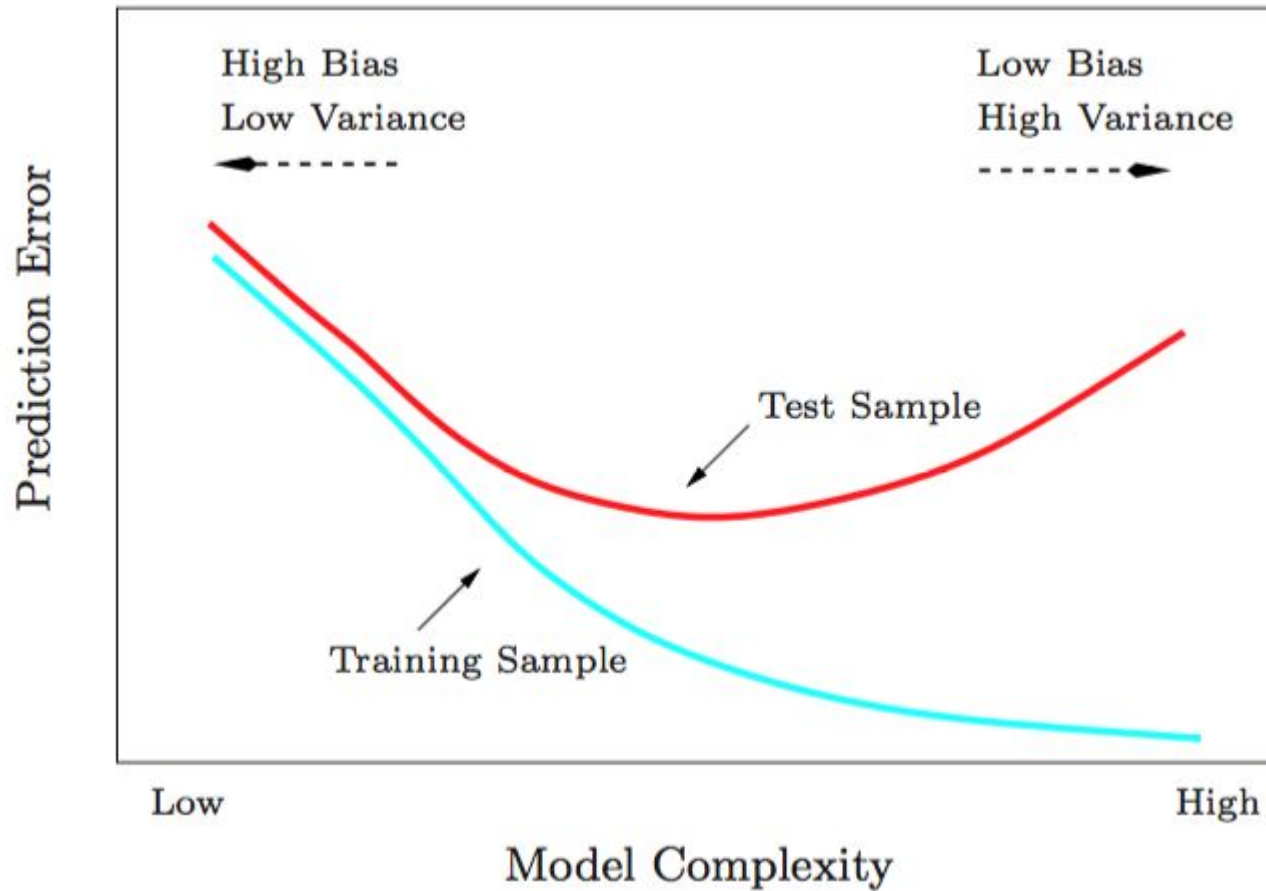
Receiver Operating Characteristic

Train and Test Error

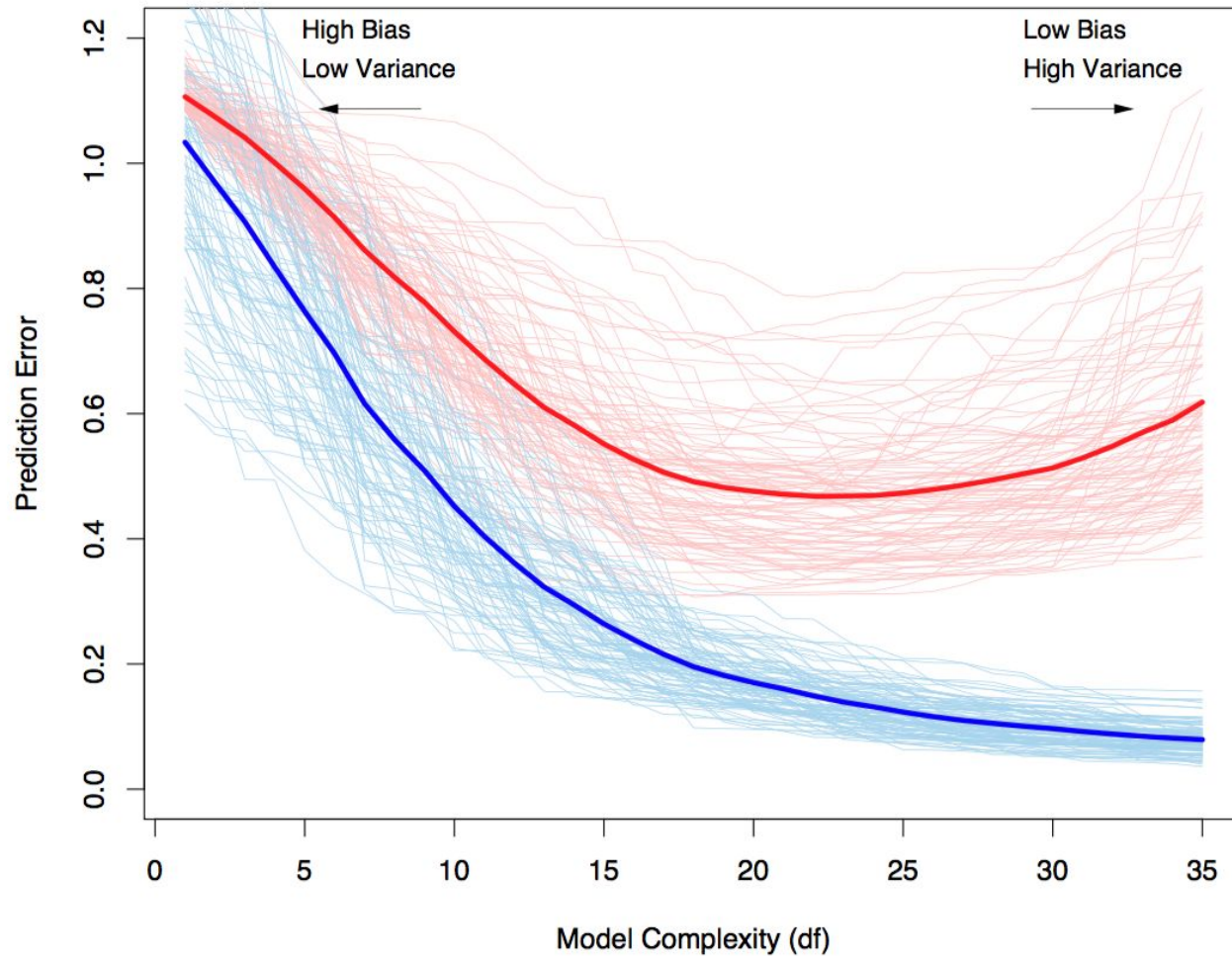


Bias vs. Variance

Bias vs. Variance



Bias vs. Variance



Bias vs. Variance

What is bias?

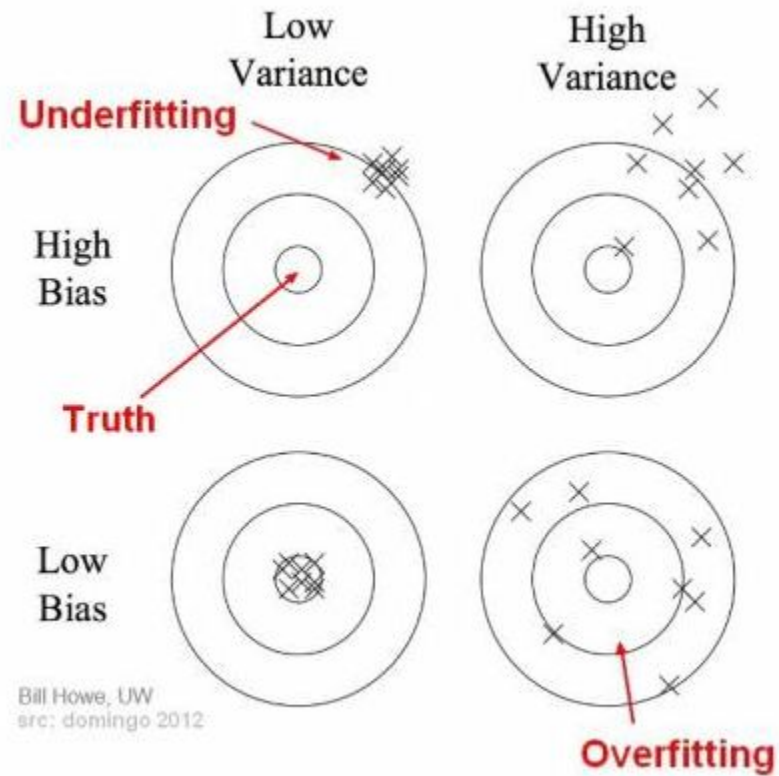
Bias is the difference between the average prediction of our model and the correct value which we are trying to predict. Model with high bias pays very little attention to the training data and oversimplifies the model. It always leads to high error on training and test data.

What is variance?

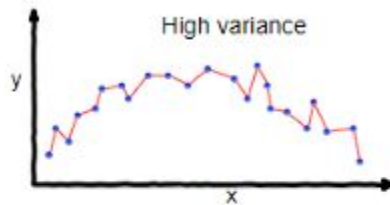
Variance is the variability of model prediction for a given data point or a value which tells us spread of our data. Model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn't seen before. As a result, such models perform very well on training data but has high error rates on test data.

Source: <https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229>

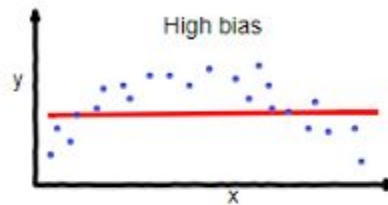
Bias vs. Variance



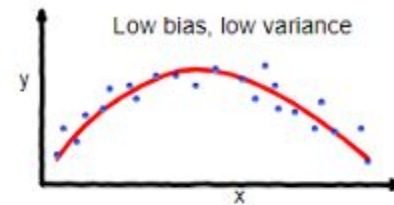
Bias vs. Variance



overfitting



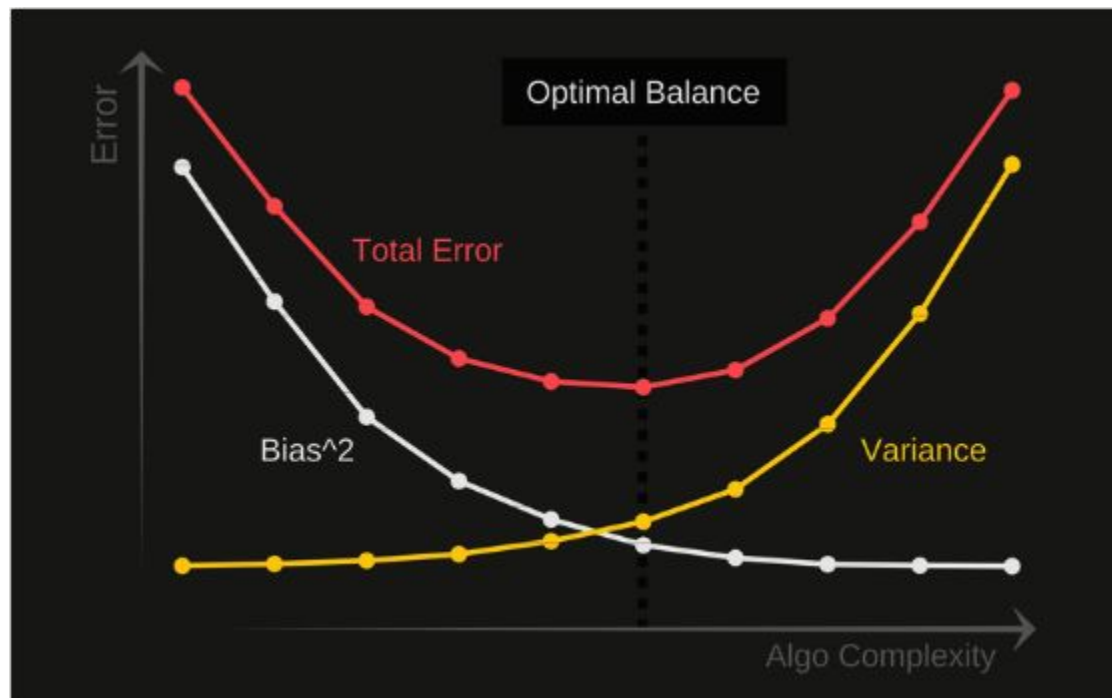
underfitting



Good balance

Bias vs. Variance

$$\text{Total Error} = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$



Linear Model

Linear Regression Model - output

Call:

```
lm(formula = count ~ season + weekday, data = bikeTrain)
```

Residuals:

Min	1Q	Median	3Q	Max
-184.13	-97.84	-28.24	62.80	507.68

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	76.6538	5.6346	13.604	<2e-16 ***
season	27.1605	1.5824	17.164	<2e-16 ***
weekday	-0.1629	0.8768	-0.186	0.853

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 129.9 on 5419 degrees of freedom

Multiple R-squared: 0.05159, Adjusted R-squared: 0.05124

F-statistic: 147.4 on 2 and 5419 DF, p-value: < 2.2e-16

R^2 : explained variance

$$R^2 = 1 - \frac{\sum (Y_{actual} - Y_{predicted})^2}{\sum (Y_{actual} - Y_{mean})^2}$$

$$R^2_{adjusted} = 1 - \frac{(1 - R^2) * (N - 1)}{N - p - 1}$$

p = Number of predictors

N = total sample size

Linear Regression Model: assumptions

Assumptions:

- Linear Separability
- Normal Distribution of errors
- etc.

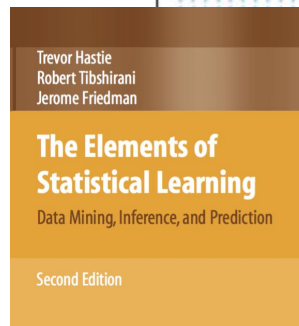
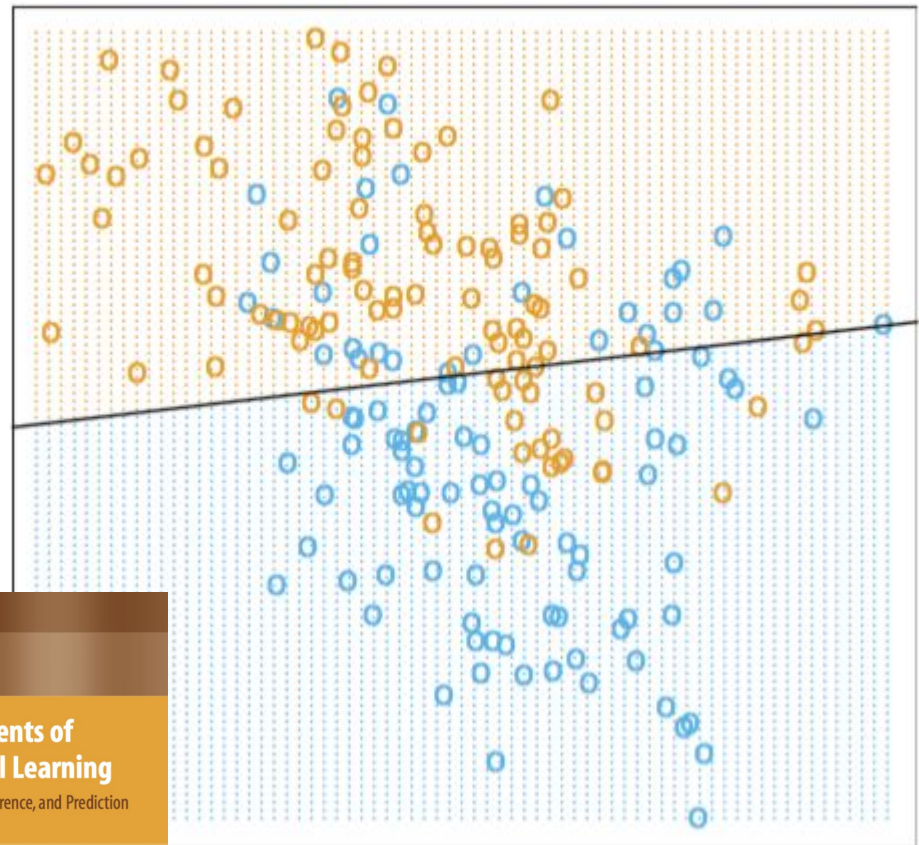
Workaround:

- Feature engineering
- Feature transformation:
log, exp, pow
- Poly

Advantage:

- Easy and quick to fit
- Easy to understand

Linear Regression of 0/1 Response



source:

Coding Exercise: Improve lm results

Exercise 1: build a linear model for the Bike Sharing Demand data.

Start with: `m1.1.2/lect/0-model_evaluation/model_evaluation.R`

Compare results to manual model.



Nearest Neighbour

Nearest Neighbour: How does it work?

Steps:

- Define a distance on feature space
- For a test data:
 - Find k closest train examples
 - Take the average for regression
 - Majority for classification
- Make this value the prediction

Distance?

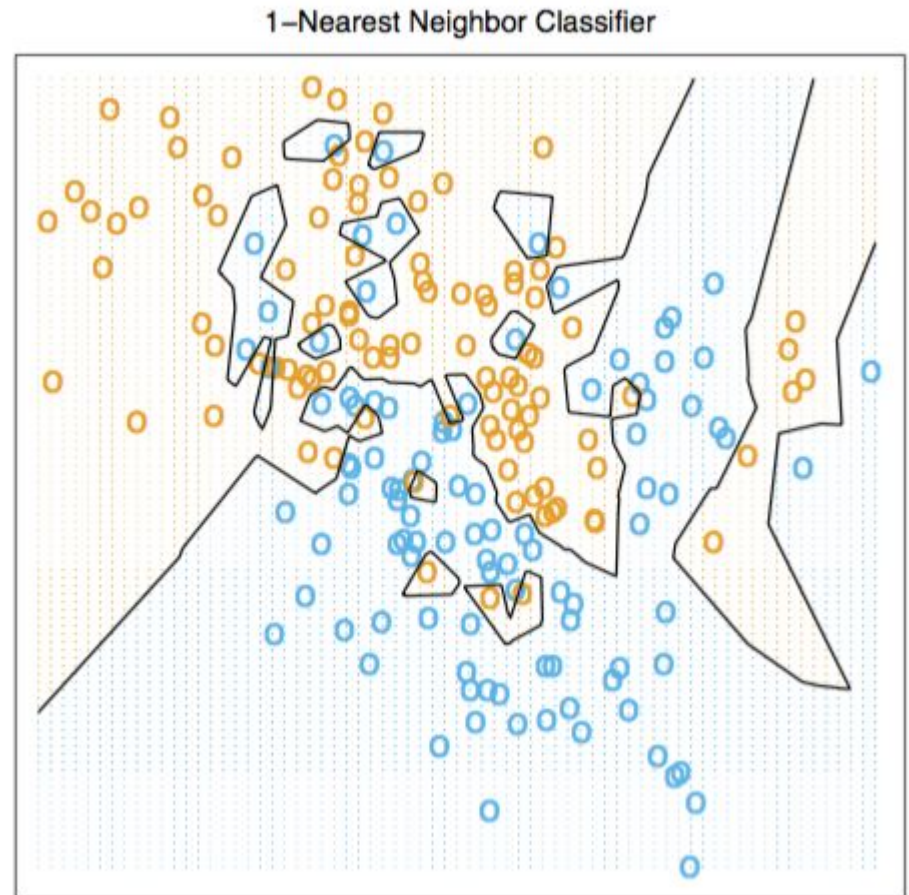
Euclidean distance is a good candidate.

Advantage?

For uniform features it is intuitive with ability for complexity.

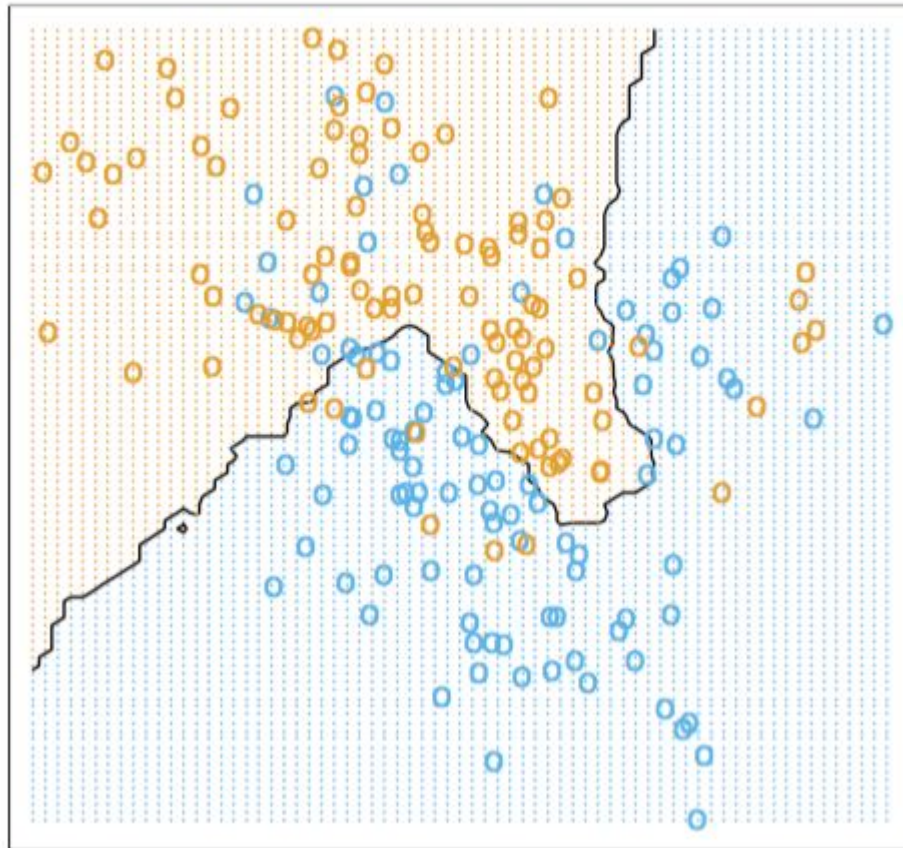
Difficulties?

Equal weight for all features?

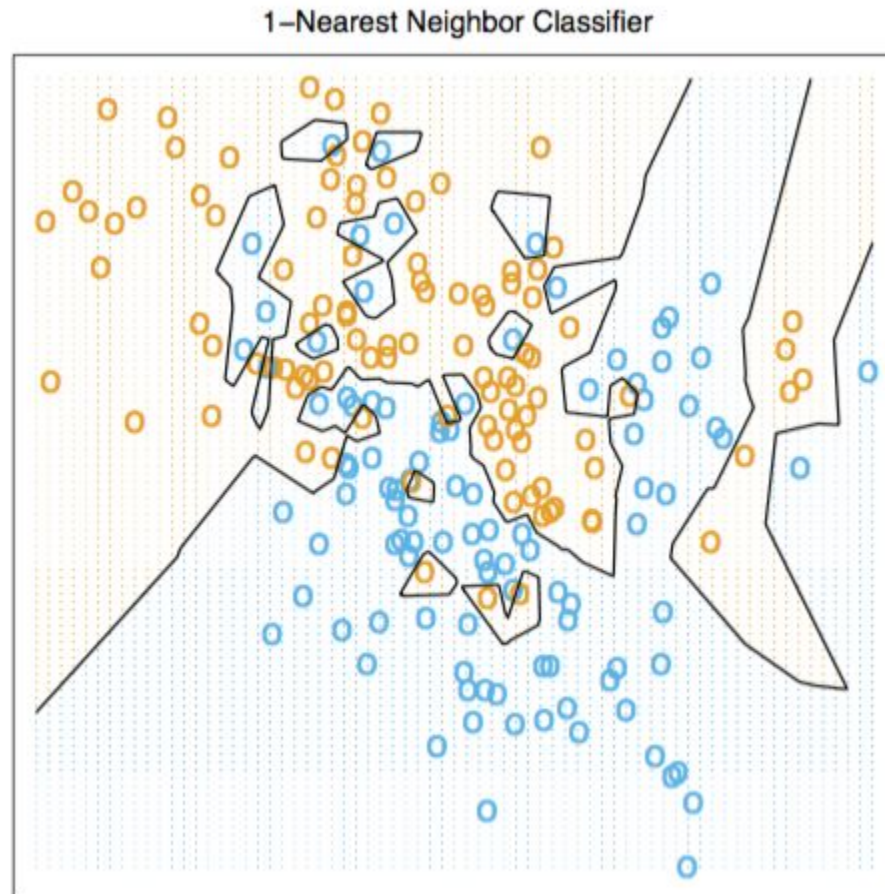


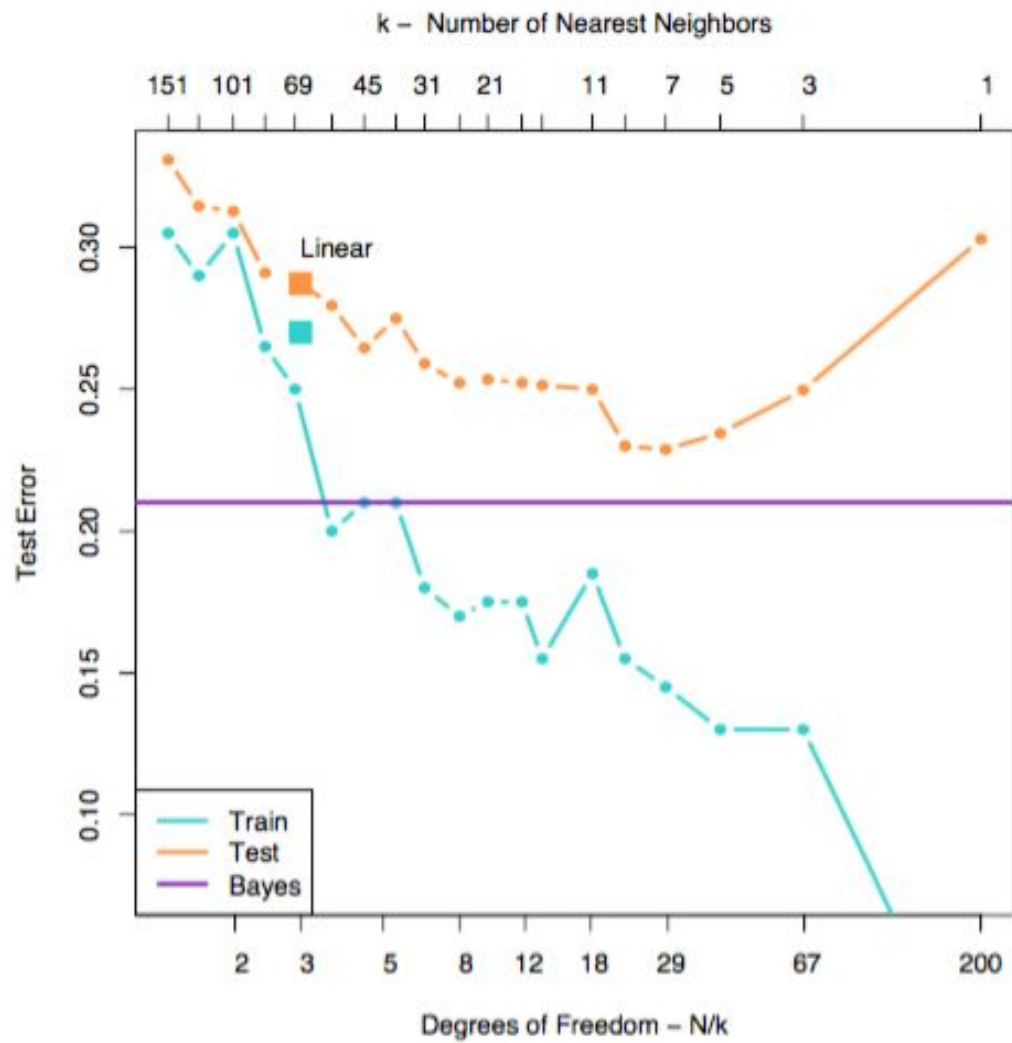
Nearest Neighbour: a high K value

15-Nearest Neighbor Classifier

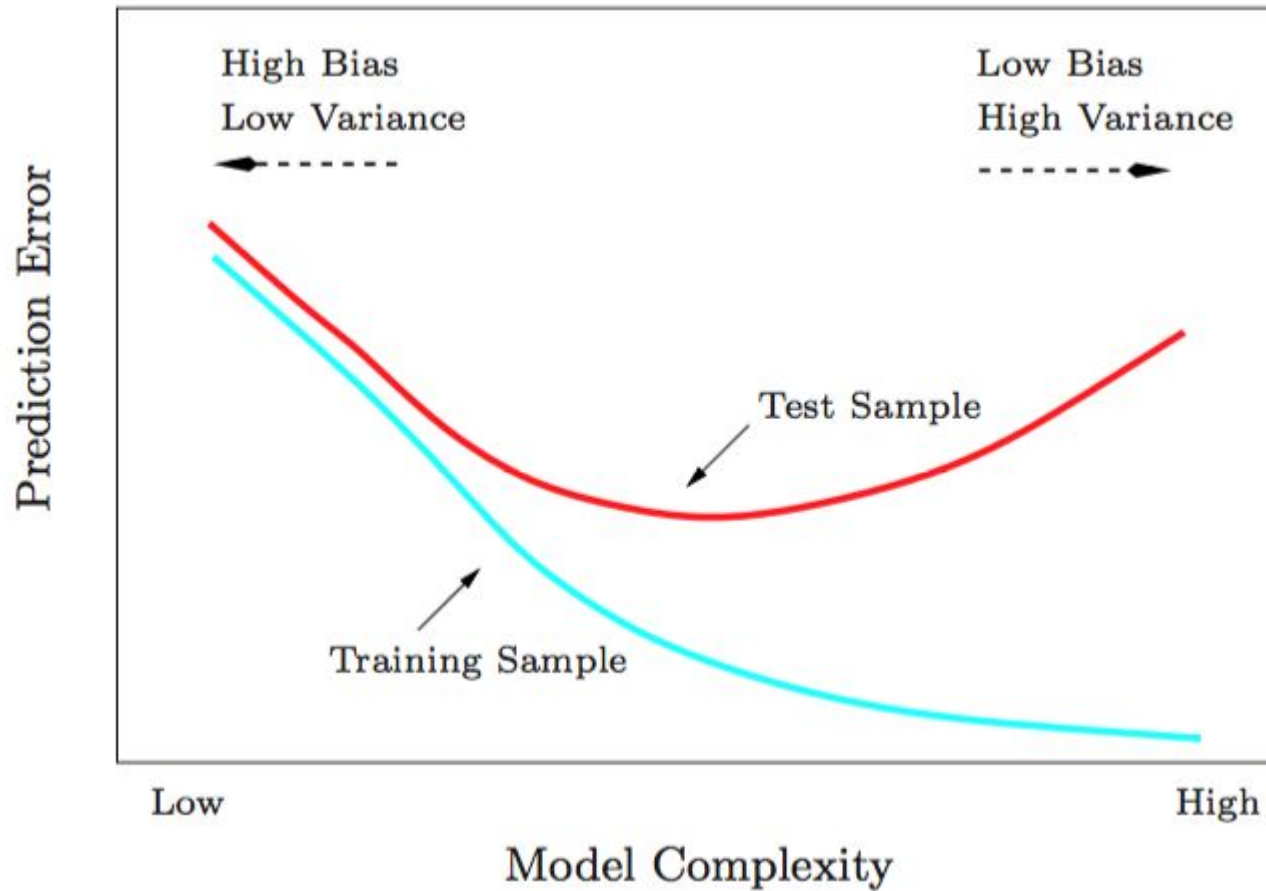


Nearest Neighbour: a low K value



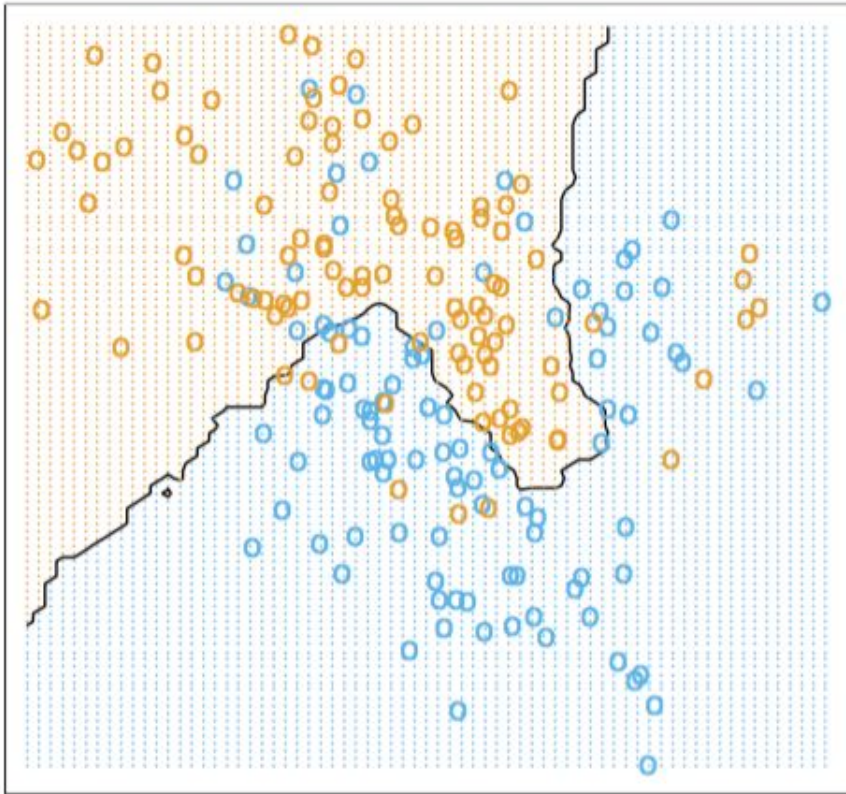


Bias vs. Variance

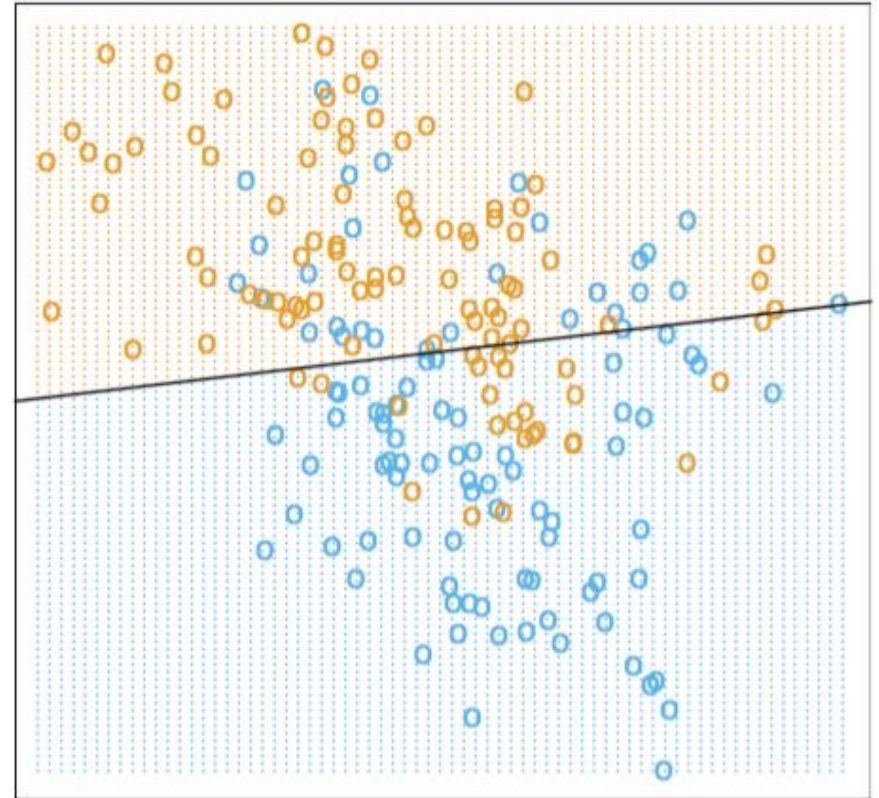


Linear Model vs. Nearest Neighbour

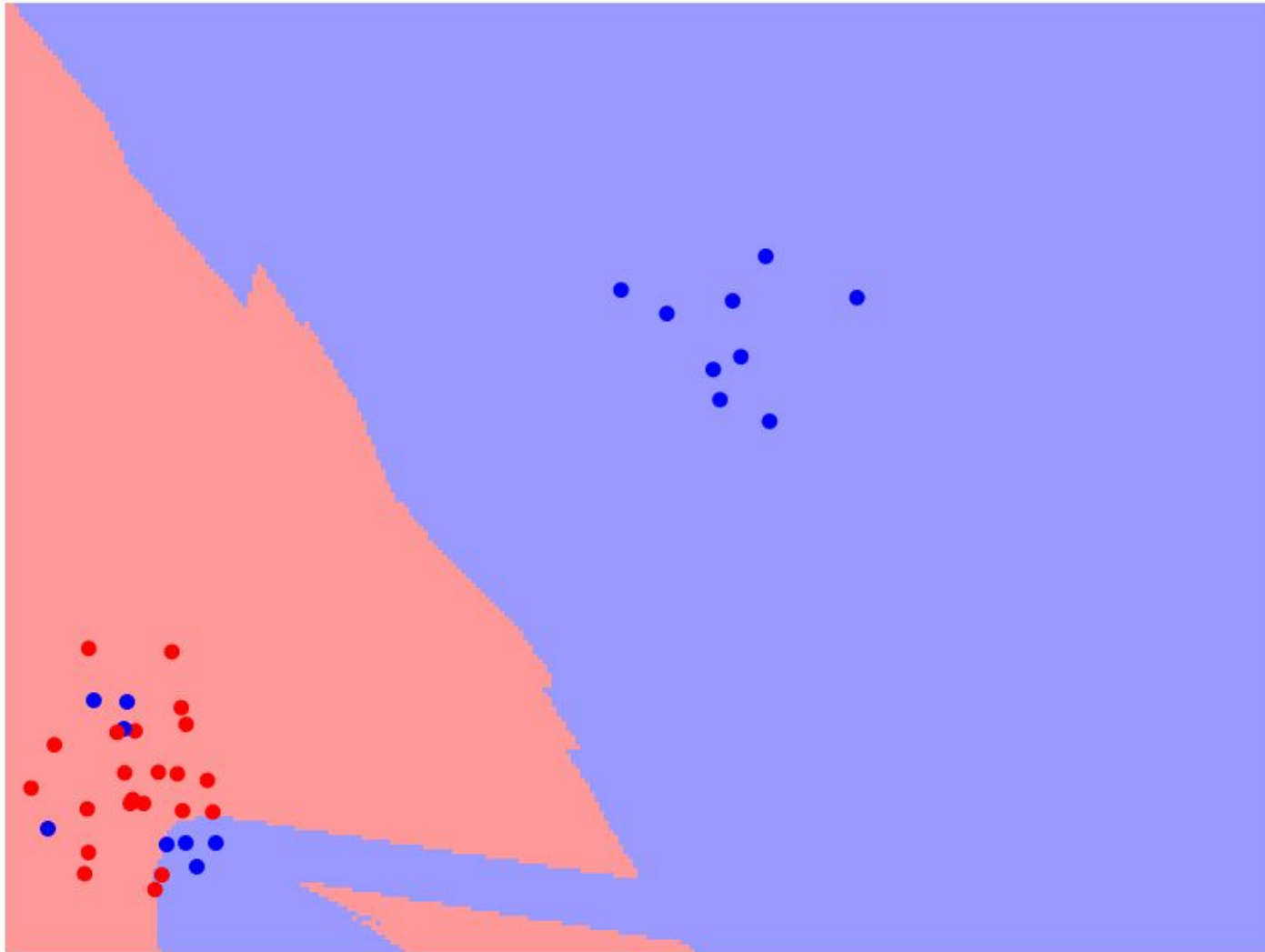
15-Nearest Neighbor Classifier



Linear Regression of 0/1 Response



Interactive Example



<http://vision.stanford.edu/teaching/cs231n-demos/knn/>

Coding Exercise: Improve NN results

Exercise 2: build a NN model for the Bike Sharing Demand data.

Start with: `ml.1.1.2/lect/0-model_evaluation/model_evaluation.R`

Compare results to manual model.



Coding Exercise: Improve Nearest Neighbour results

Exercise 2: build a NN model for the Bike Sharing Demand data.

Start with: `ml.1.3/lect/1_regression_tools.R`

Which k is the right value? How does CV help to find it?



Summary

We have reviewed some general concepts in Supervised Learning:

- Measuring **prediction power** of regression and classification models
- Separating **Training** and **Test Error**
- Separating data into **Training**, **Validation** and **Test** sets
- A number of **Cross-validation Techniques**
- Model complexity consequences: **Bias vs. Variance**
- **Model selection** criterias
- **Benchmark** Model selection

We reviewed a number of Machine Learning Models and evaluated their parameters to control complexity

- **Manual Models** to familiarize with the concept of learning
- **Linear Model:** assumptions, advantage, feature transformations
- **Nearest Neighbour:** algorithm, advantages, difficulties, complexity

