# Machine Learning and Artificial Intelligence: Part 1

## *An introduction to core concepts*

*George Bezerra*
*Director of Data Science, TripAdvisor*

SFI Complex Systems Summer School 2018

# What is machine learning?

- The science of building computer models that:
  - Learn from data how to perform a task
  - Self-tune their parameters to optimize performance
  - Generalize behavior to new/unseen data

- The primary goal of ML is to provide solutions to practical real-world problems:
  - Inspiration from biology is welcome but not required
  - Explaining nature is a plus but not a must

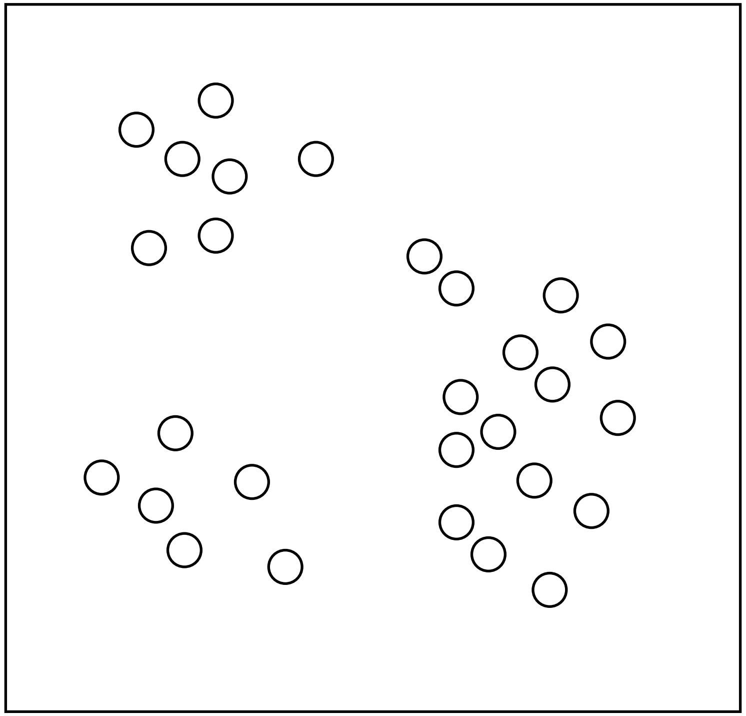
# Types of learning

- Unsupervised learning:
  - Training data is unlabeled. The model searches for a compact representation of the inherent structure in the data.
- Supervised learning:
  - Training data is labeled. The goal is to create a compact mapping between the input features and the target.
- Reinforcement learning:
  - Interactive learning where training data (labeled and unlabeled) is obtained by interacting with an environment.

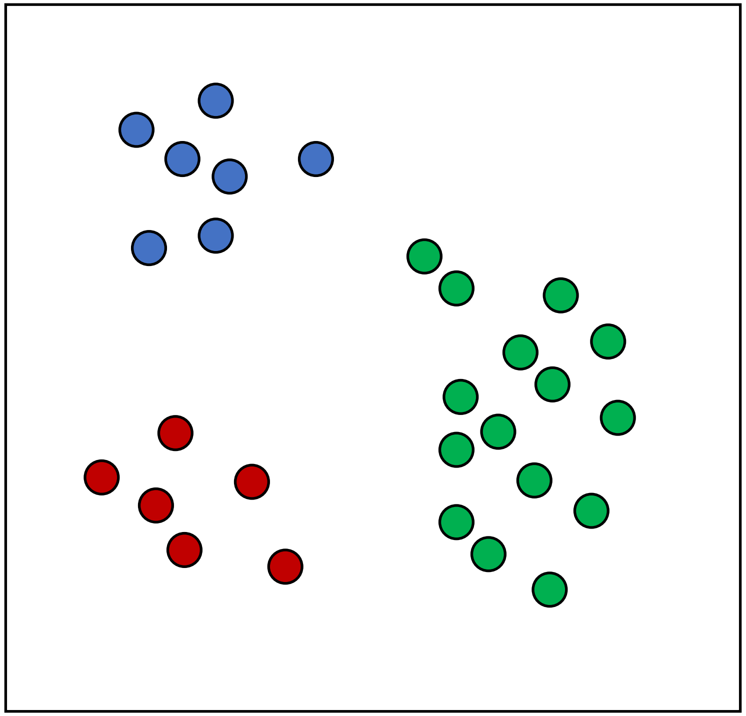
# Types of machine learning problems

- Clustering
  - Finding inherent groups in the data (unsupervised)
- Classification
  - Predict the (discrete) class of each data point (supervised)
- Regression
  - Predict a continuous, real-valued variable (supervised)
- Dimensionality reduction
  - Represent the data using a reduced number of variables (unsupervised)

# Clustering – Example 1

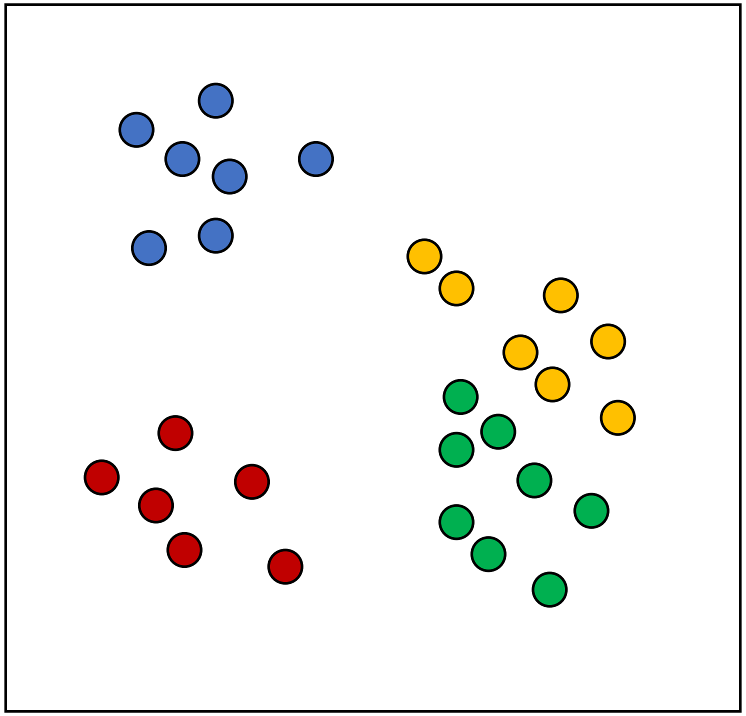


# Clustering – Example 1



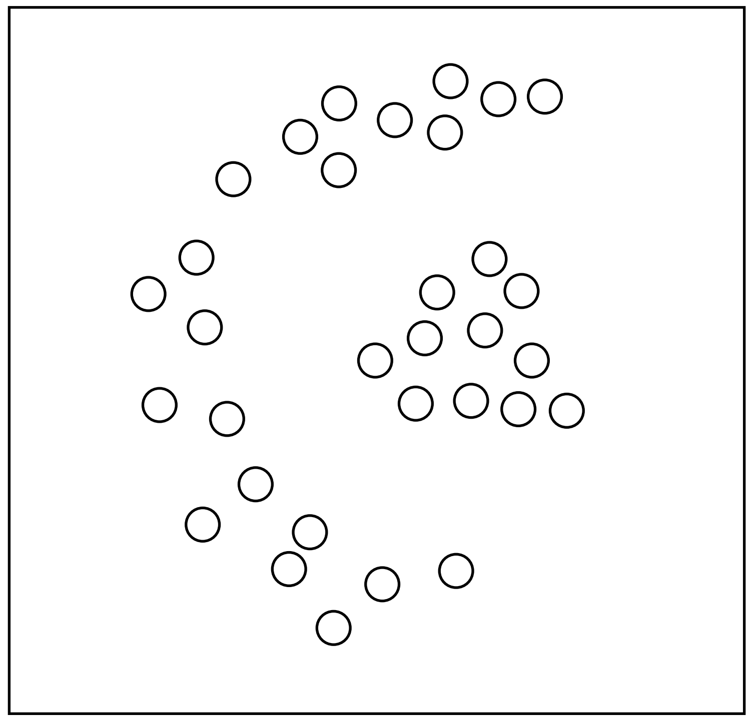
3 clusters

# Clustering – Example 1



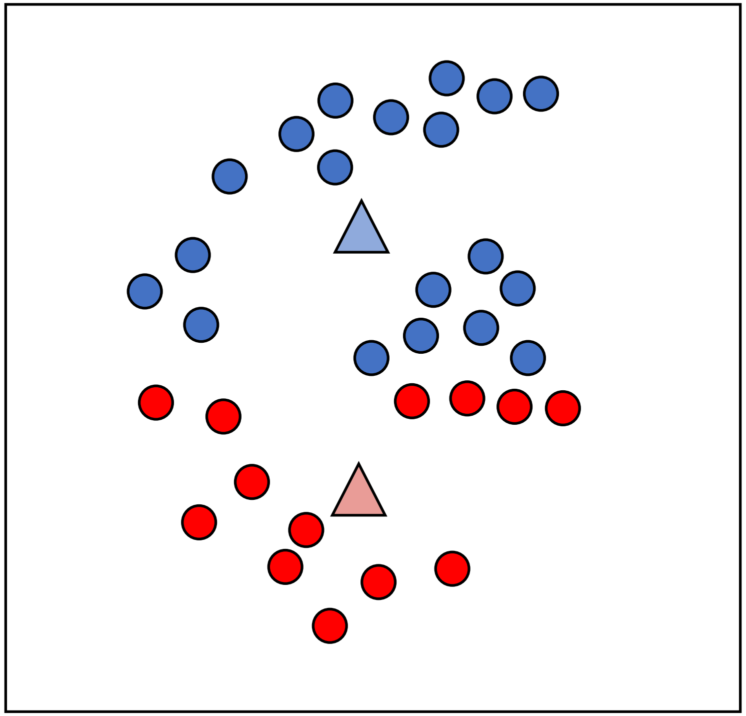
4 clusters

# Clustering – Example 2



# Clustering – Example 2

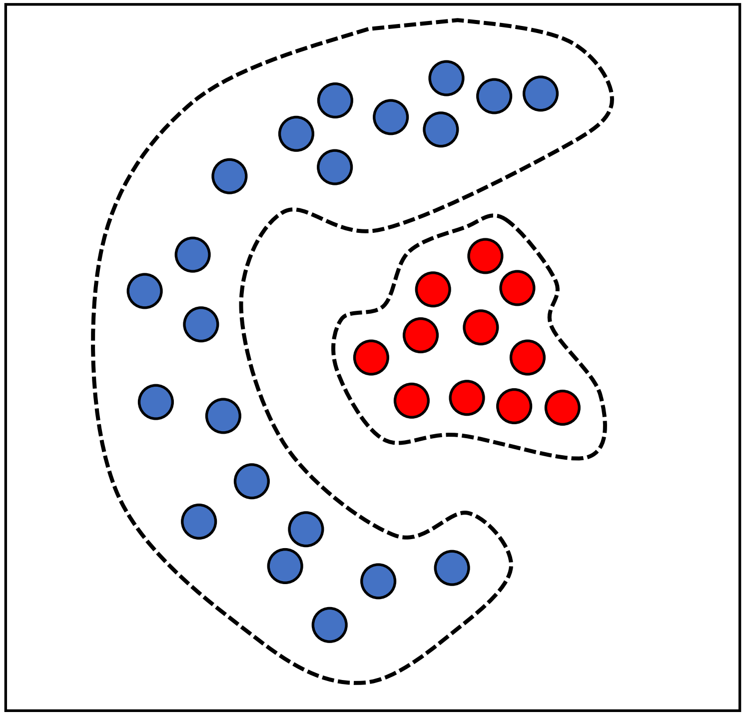
Centroid-based



2 clusters

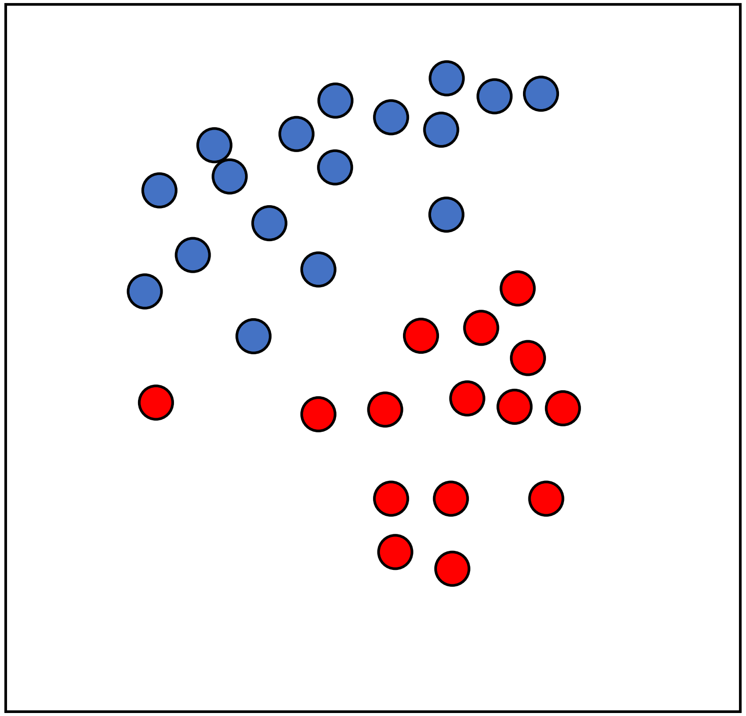
# Clustering – Example 2

Density-based



2 clusters

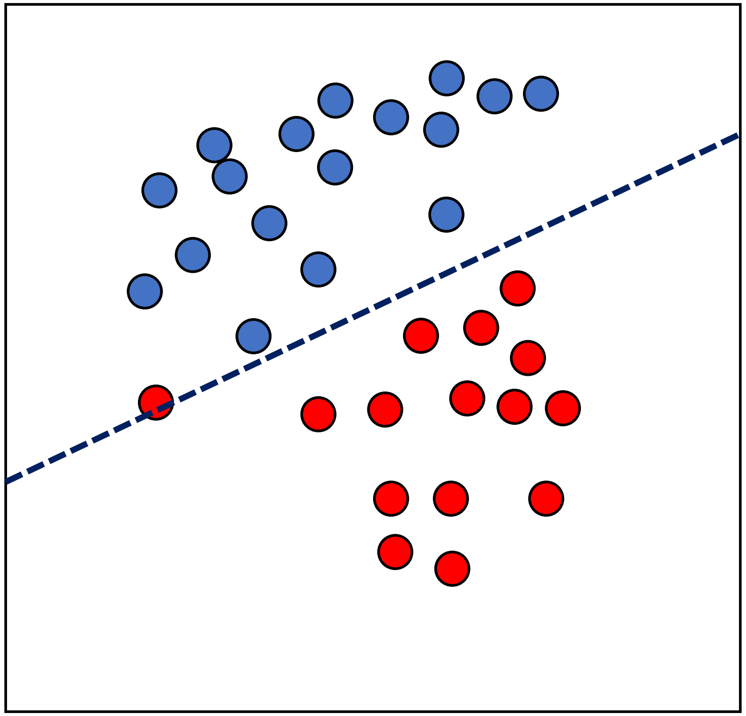
# Classification – Example 1



2 classes

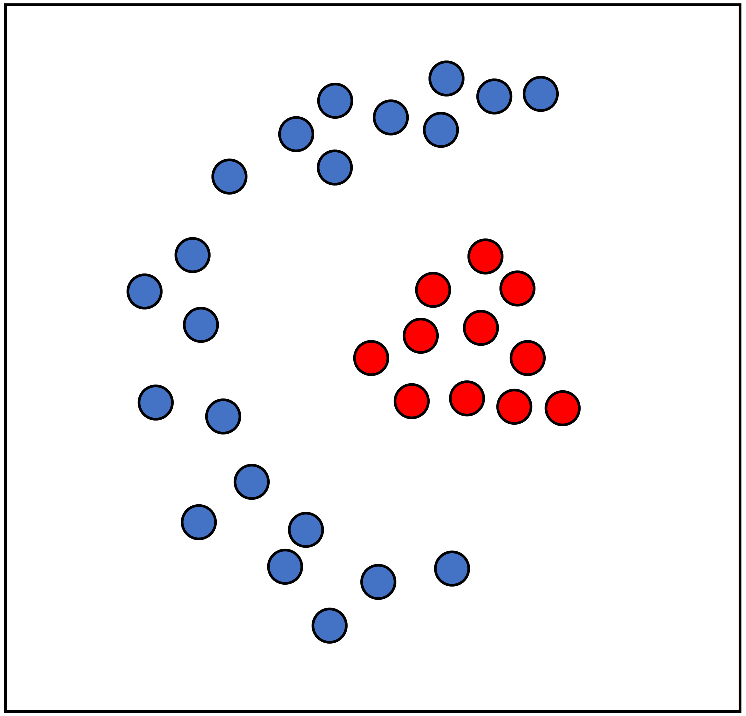
# Classification – Example 1

Linear classification



2 classes

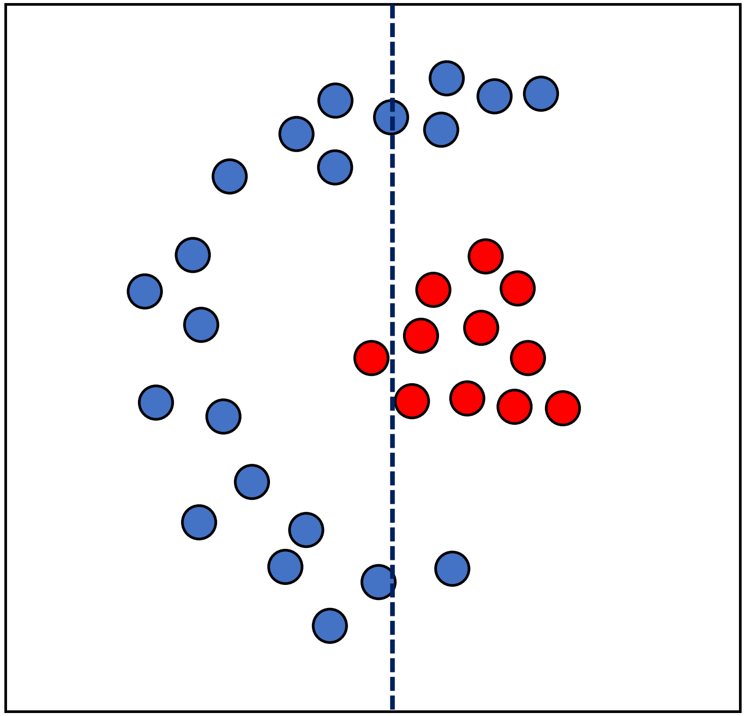
# Classification – Example 2



2 classes

# Classification – Example 2

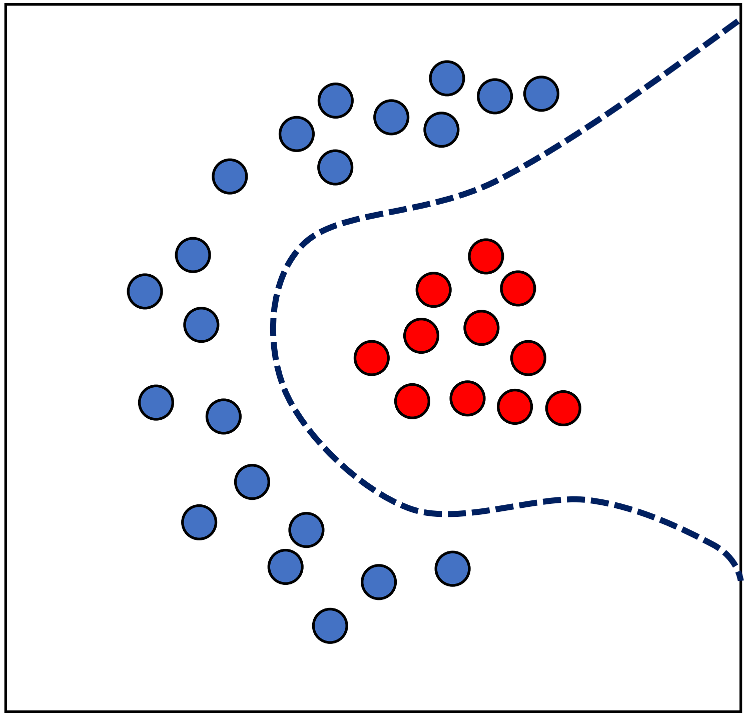
Linear classification



2 classes

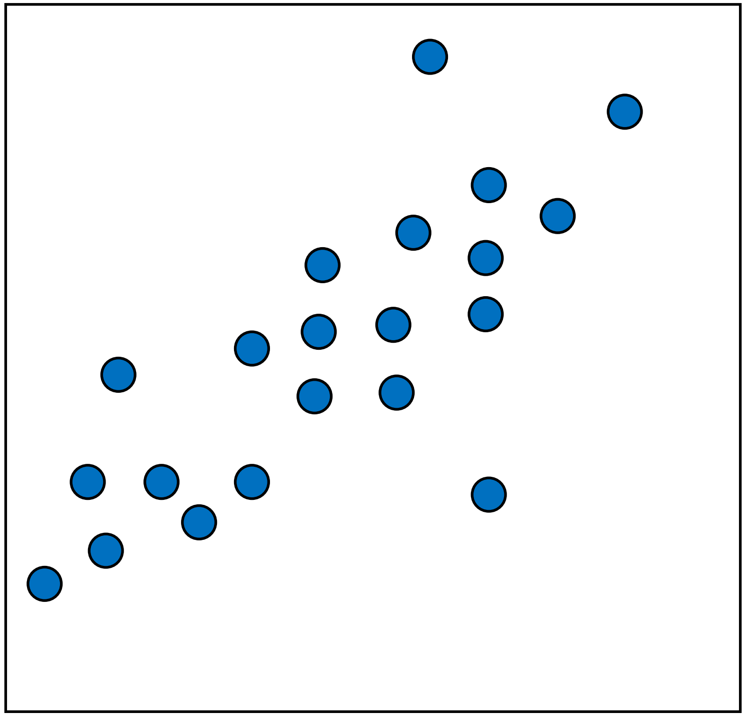
# Classification – Example 2

Non-linear classification



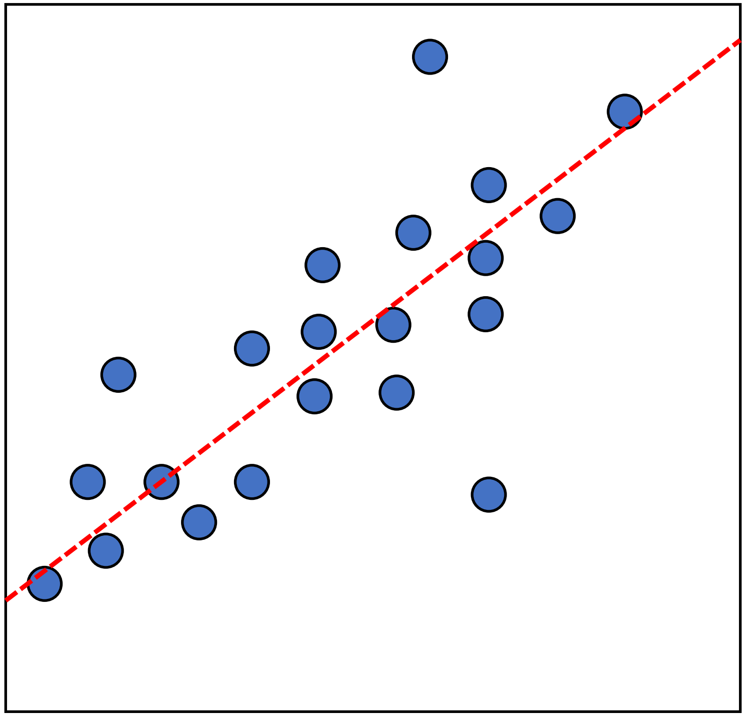
2 classes

# Regression – Example 1

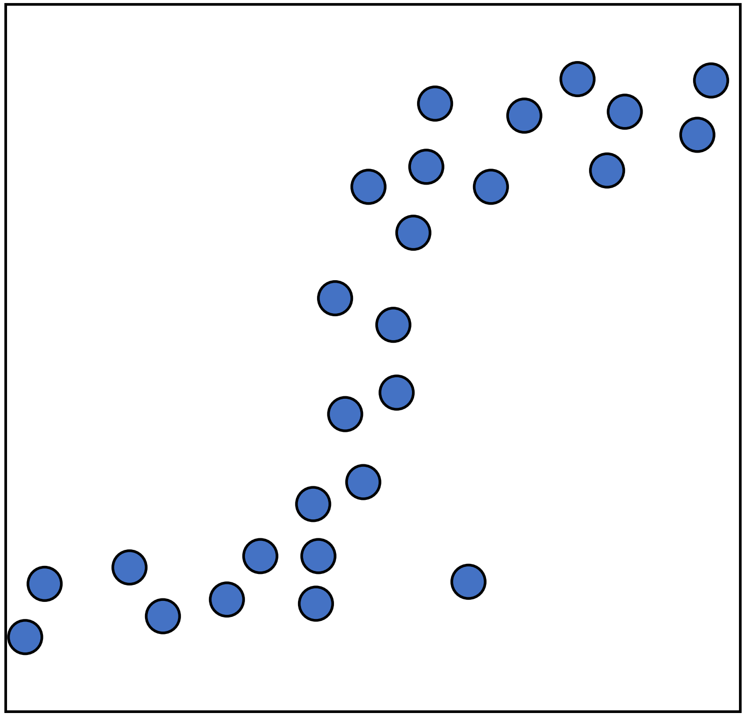


# Regression – Example 1

Linear regression

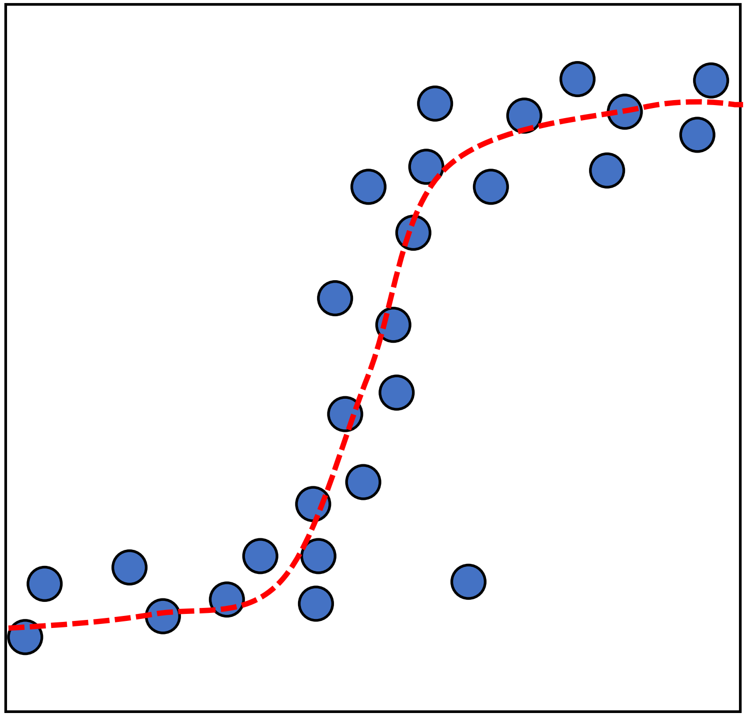


# Regression – Example 2

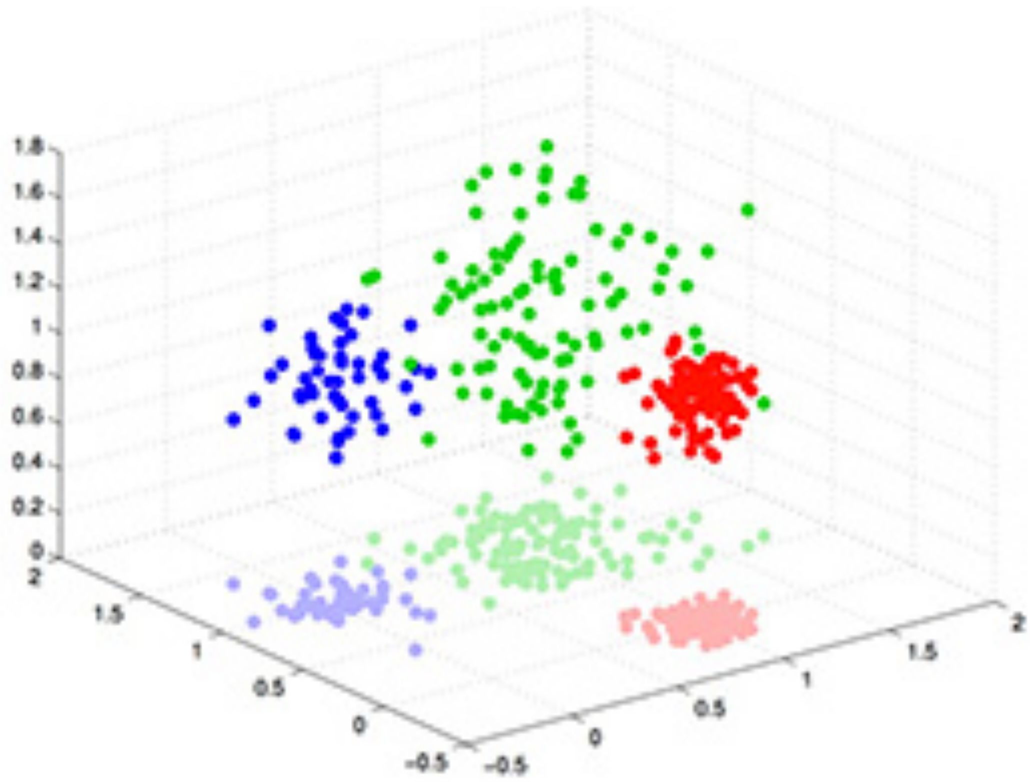


# Regression – Example 2

Non-linear regression

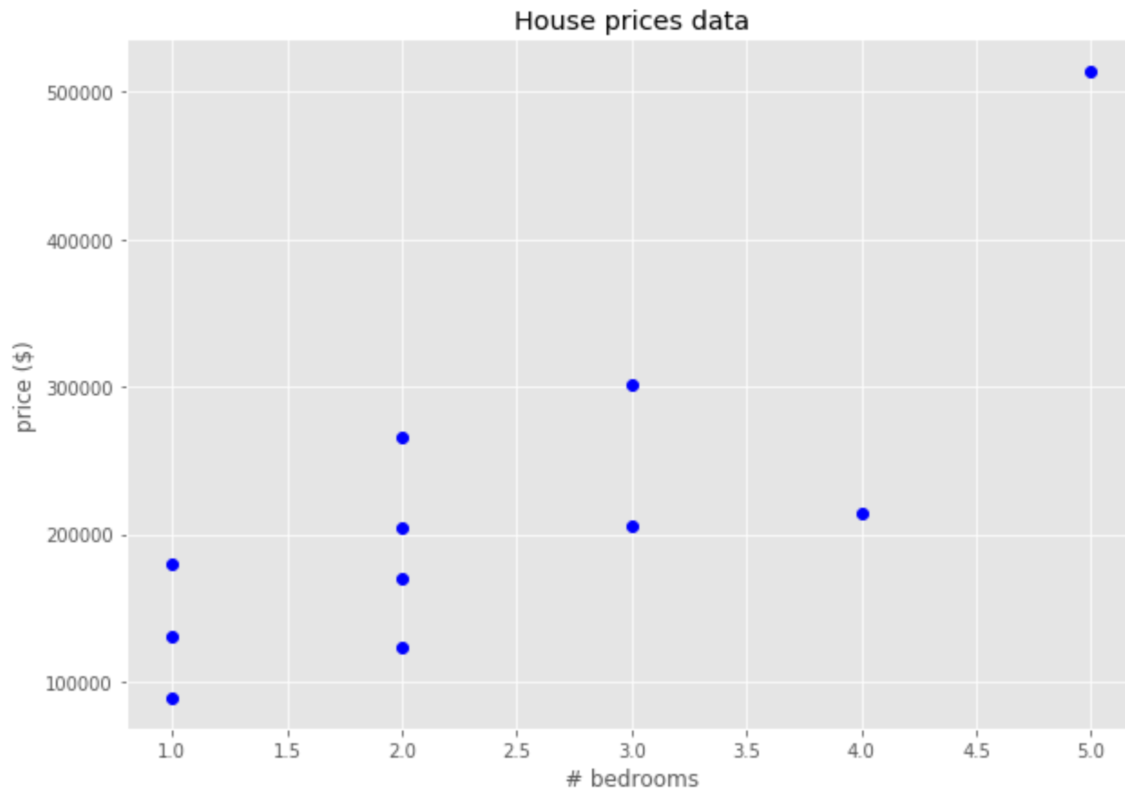


# Dimensionality reduction



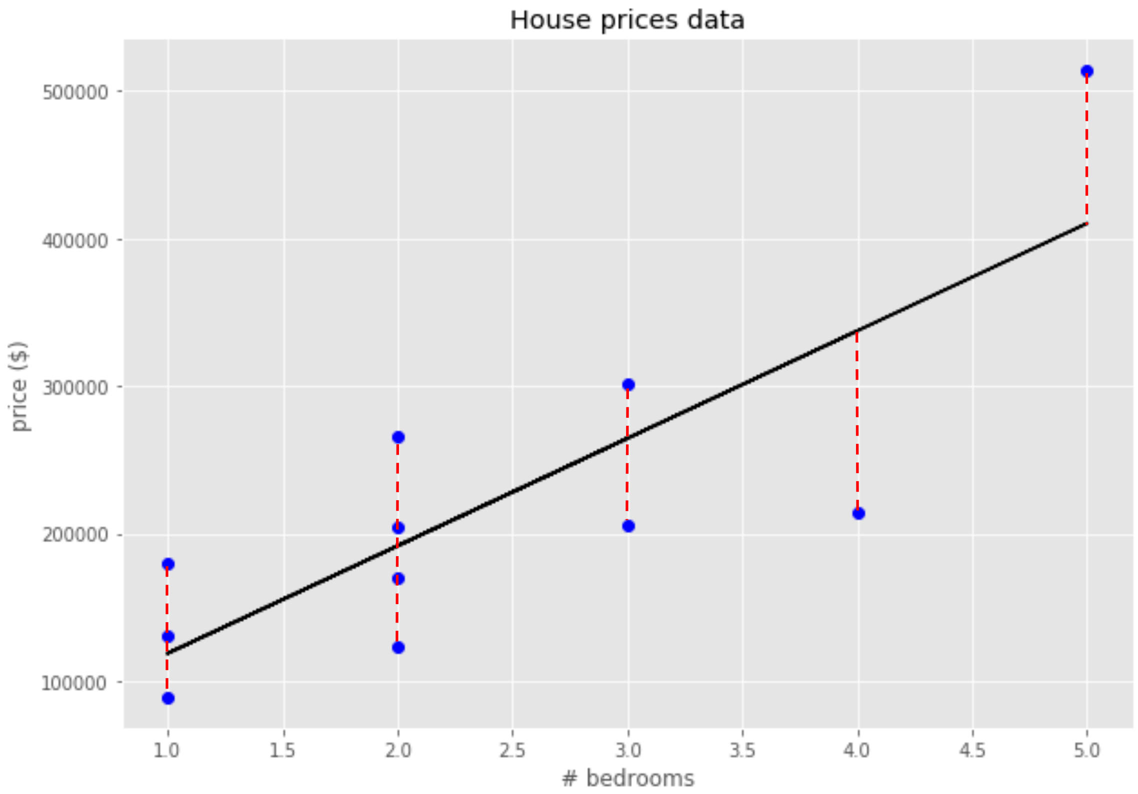
# A simple model: Linear regression

| # bdrs ($x$) | price ($\hat{y}$) |
|---|---|
| 1 | 130k |
| 2 | 122k |
| 1 | 89k |
| 3 | 301k |
| 2 | 204k |
| 5 | 514k |
| 2 | 169k |
| 1 | 180k |
| 4 | 213k |
| 2 | 266k |
| 3 | 205k |



# A simple model: Linear regression

| # bdrs ($x$) | price ($\hat{y}$) |
|---|---|
| 1 | 130k |
| 2 | 122k |
| 1 | 89k |
| 3 | 301k |
| 2 | 204k |
| 5 | 514k |
| 2 | 169k |
| 1 | 180k |
| 4 | 213k |
| 2 | 266k |
| 3 | 205k |




$$y_i = a + bx_i$$

$$Minimize: \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$

# A simple model: Linear regression

Feature matrix:

| | | | | | | |
|---|---|---|---|---|---|---|
| $x_{11}$ | $x_{12}$ | . | . | . | | $x_{1M}$ |
| $x_{21}$ | | | | | | |
| | | | | | | |
| . | | | | | | |
| . | | | | | | |
| . | | | | | | |
| | | | | | | |
| $x_{N1}$ | | . | . | . | | $x_{NM}$ |

Target:

| |
|---|
| $y_1$ |
| $y_2$ |
| . |
| . |
| . |
| . |
| . |
| $y_N$ |

M = Number of dimensions (features)
N = Number of samples

Model Equation:

$$y_i = \sum_{j=0}^{M} w_j x_{ij}$$

Loss function (mean squared error):

$$Loss = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$

# Logistic regression (classification)

Log-odds transformation:

$$log\left(\frac{p}{1-p}\right) = \sum_{j=0}^{M} w_j x_j$$

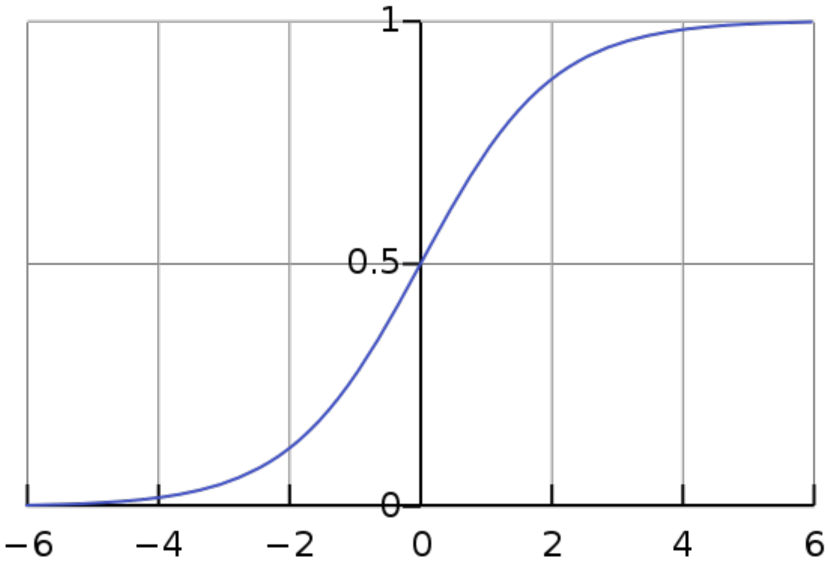
# Logistic regression (classification)

Log-odds transformation:

$$log\left(\frac{p}{1-p}\right) = \sum_{j=0}^{M} w_j x_j$$

Logistic sigmoid

$$p = \frac{1}{1 + e^{-\left(\sum_{j=0}^{M} w_j x_j\right)}}$$



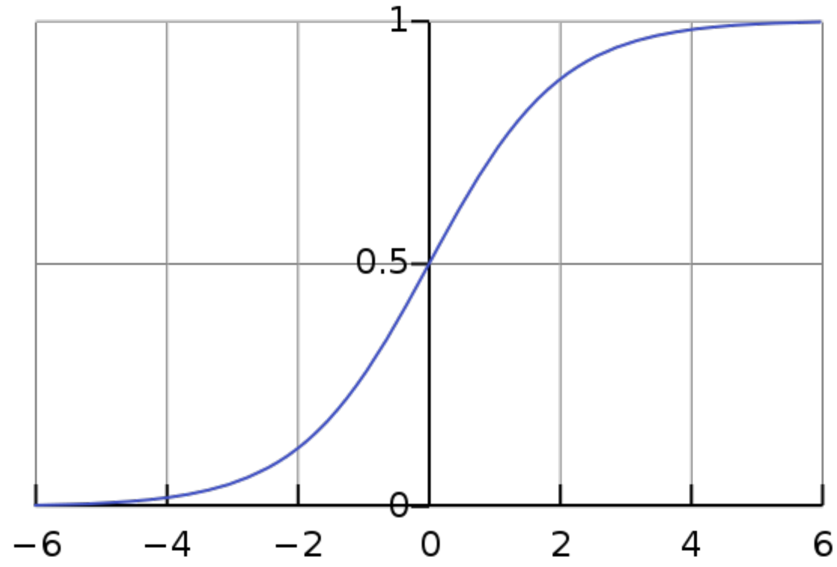
# Logistic regression (classification)

Log-odds transformation:

$$log\left(\frac{p}{1-p}\right) = \sum_{j=0}^{M} w_j x_j$$

Logistic sigmoid

$$p = \frac{1}{1 + e^{-\left(\sum_{j=0}^{M} w_j x_j\right)}}$$

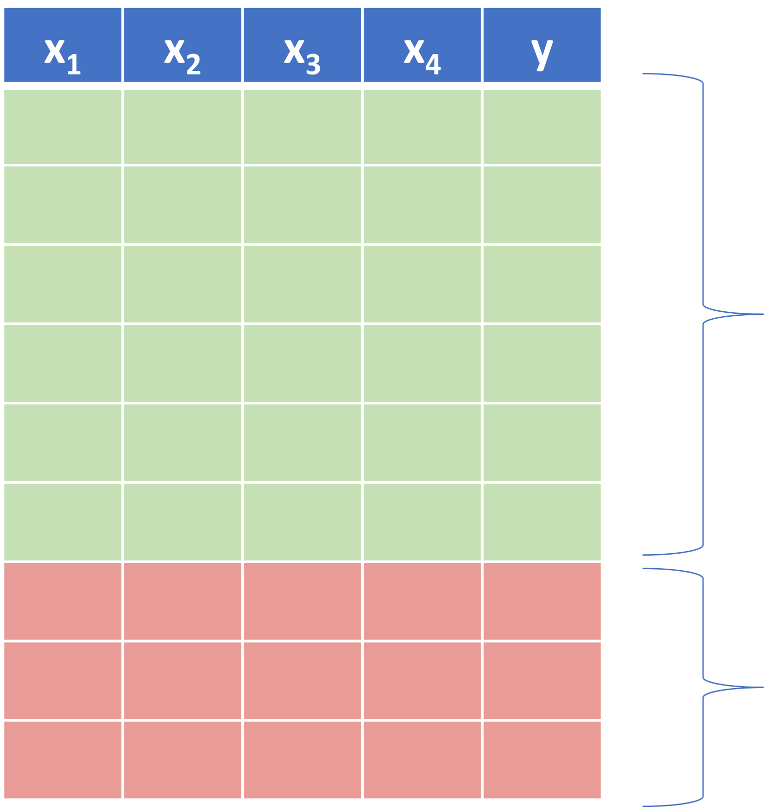



Binary cross-entropy (Log-loss):

$$Loss = -\sum_{i}^{N} \hat{y} \cdot \log p + (1-\hat{y}) \log(1-p)$$

# Model validation

- Splitting between training and test data

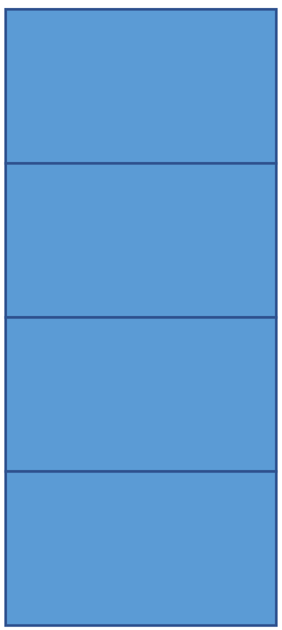



**Training data:** Used to fit the model

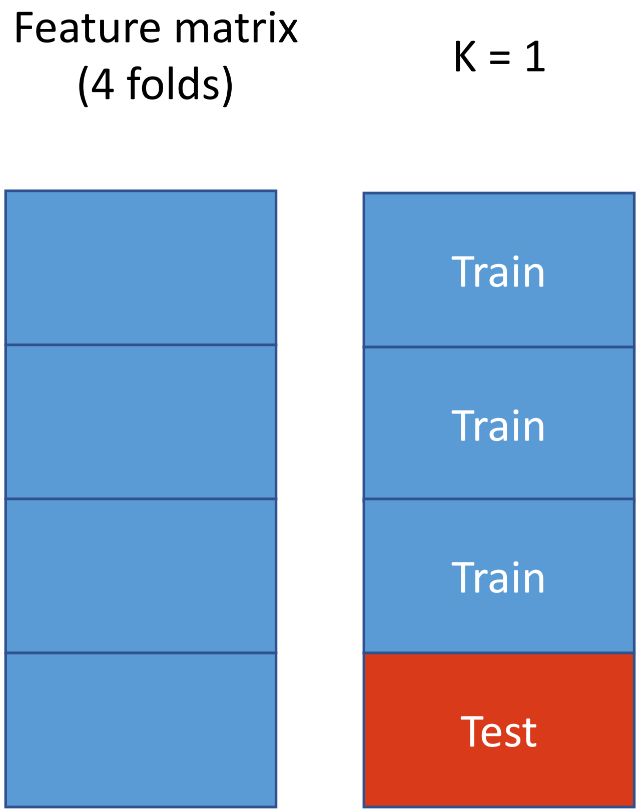
**Test data:** Used evaluate the model’s generalization performance

# K-fold cross-validation

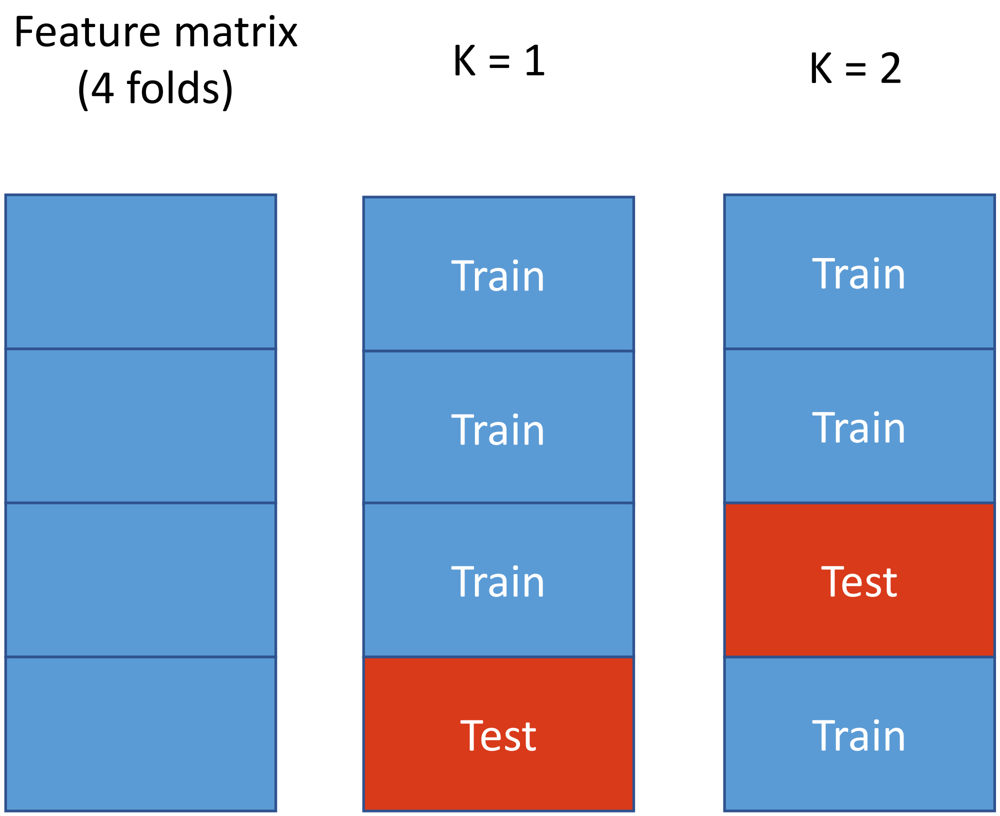
Feature matrix
(4 folds)



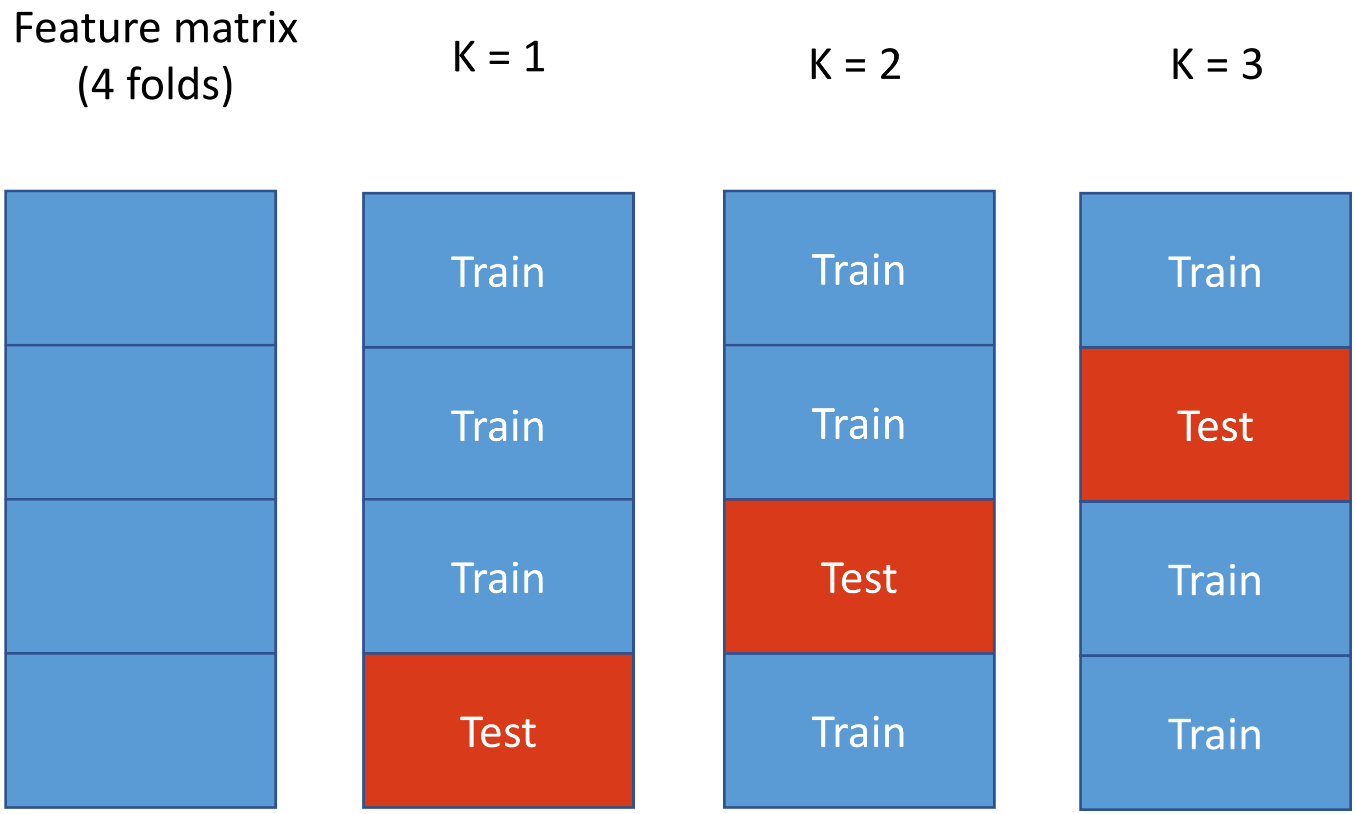
# K-fold cross-validation



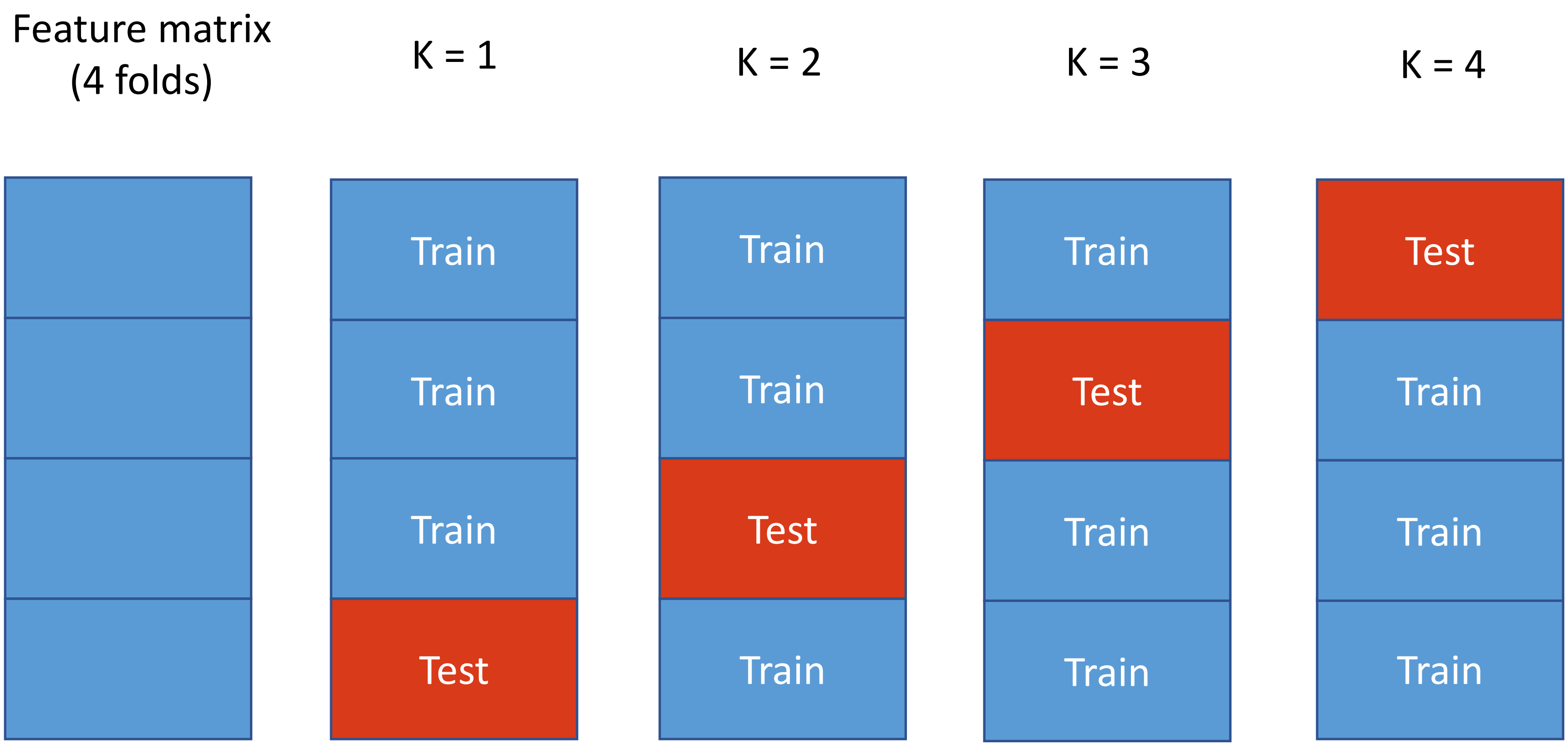
# K-fold cross-validation



# K-fold cross-validation

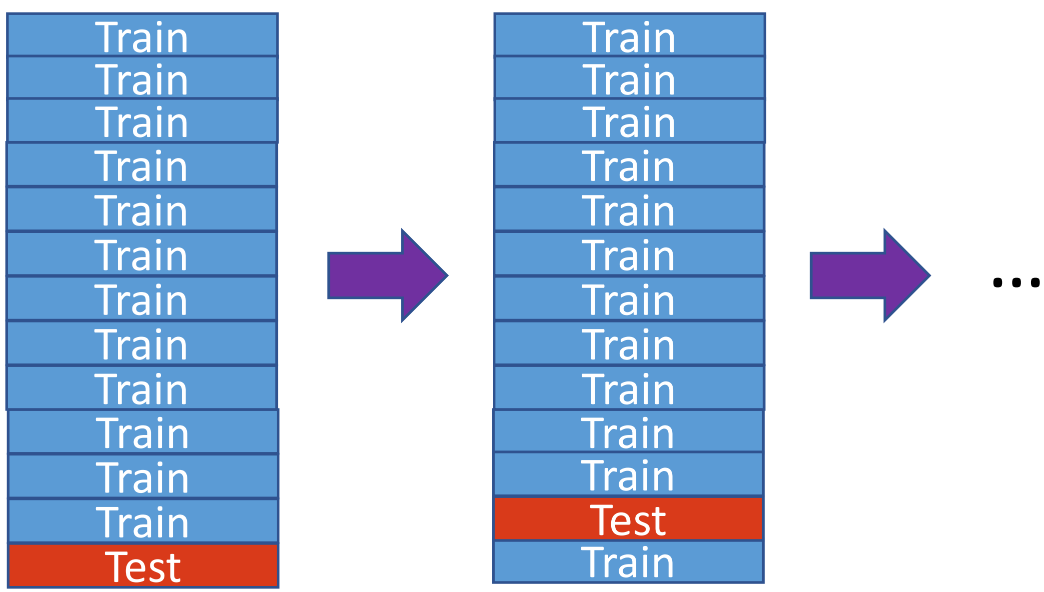


# K-fold cross-validation



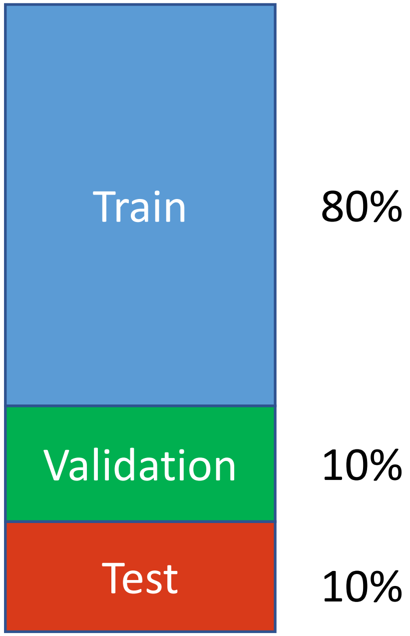
# Leave-one-out cross-validation

- Useful when very little data is available
- K = N: Use a single data point for testing; train on the remaining samples in the data set

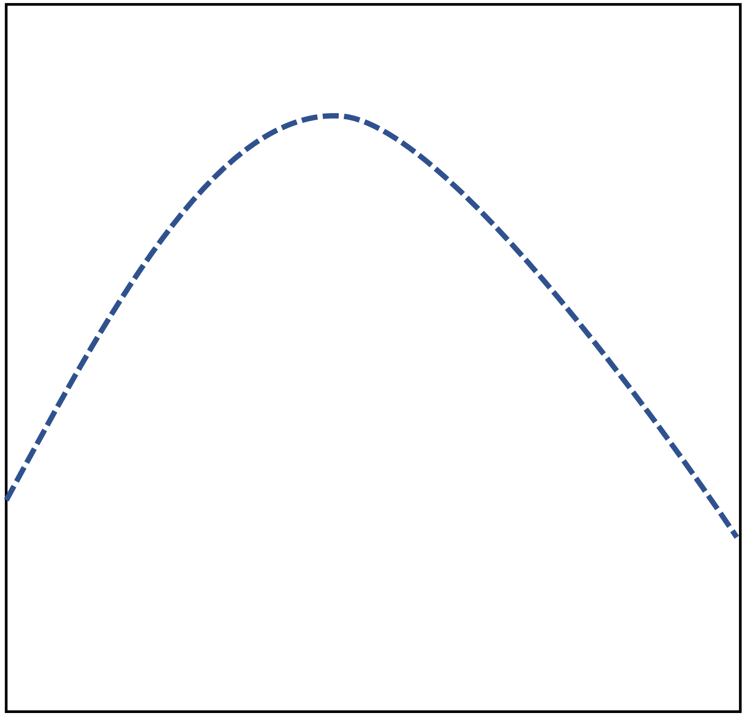


# Train-validation-test split

- Sometimes your model takes too long to train or there is plenty of data available.
- Use a fixed split into train, validation, and test sets.

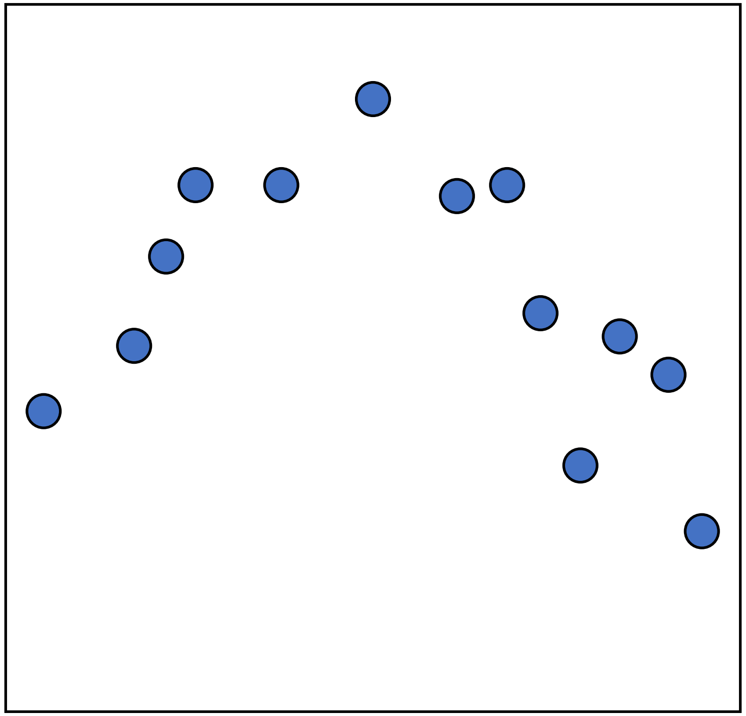


# Model selection



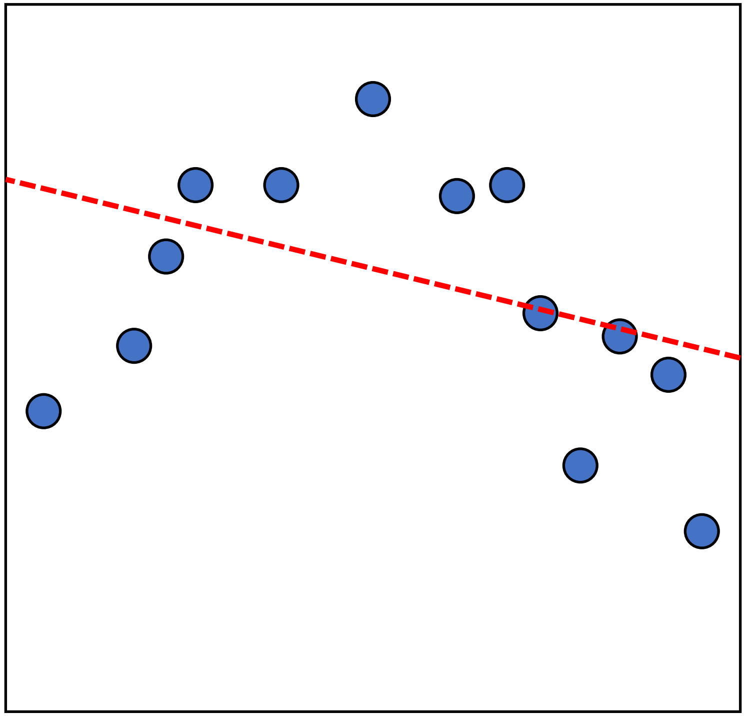
True function: $y = f(x)$

# Model selection



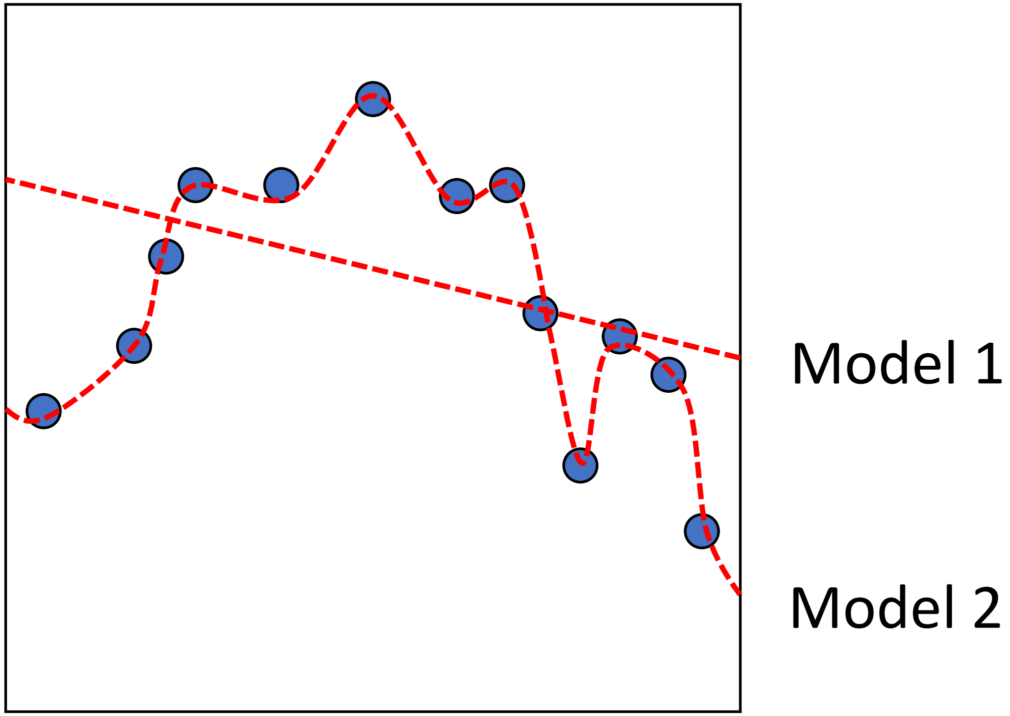
Sample of *f(x)* with Gaussian noise

# Model selection




Linear model approximation of *f(x)*

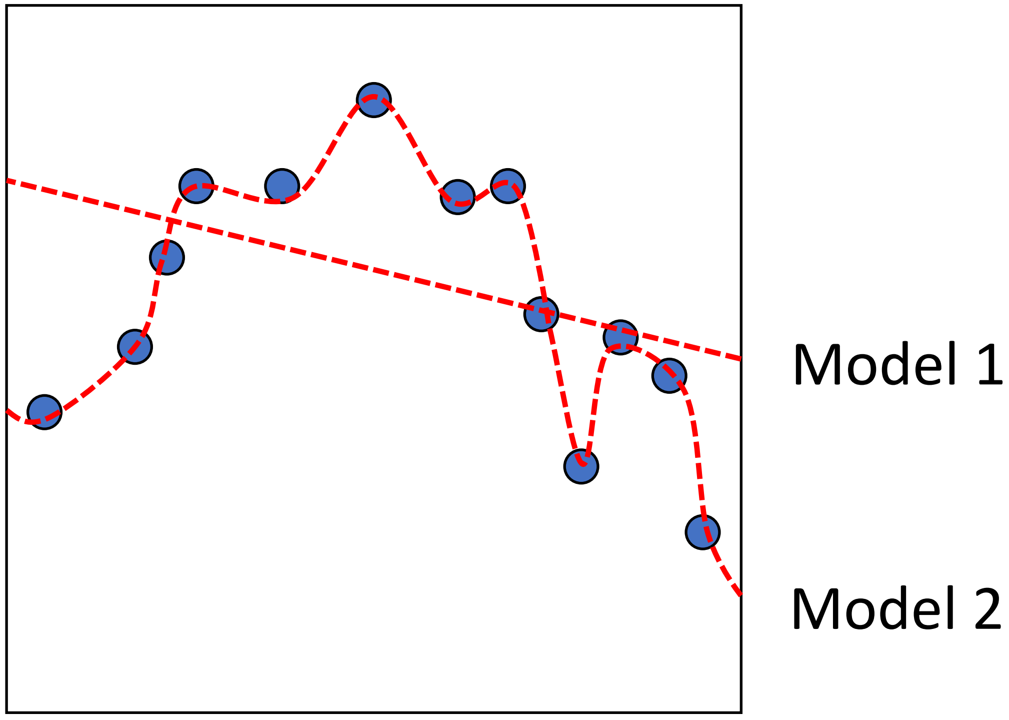
# Model selection



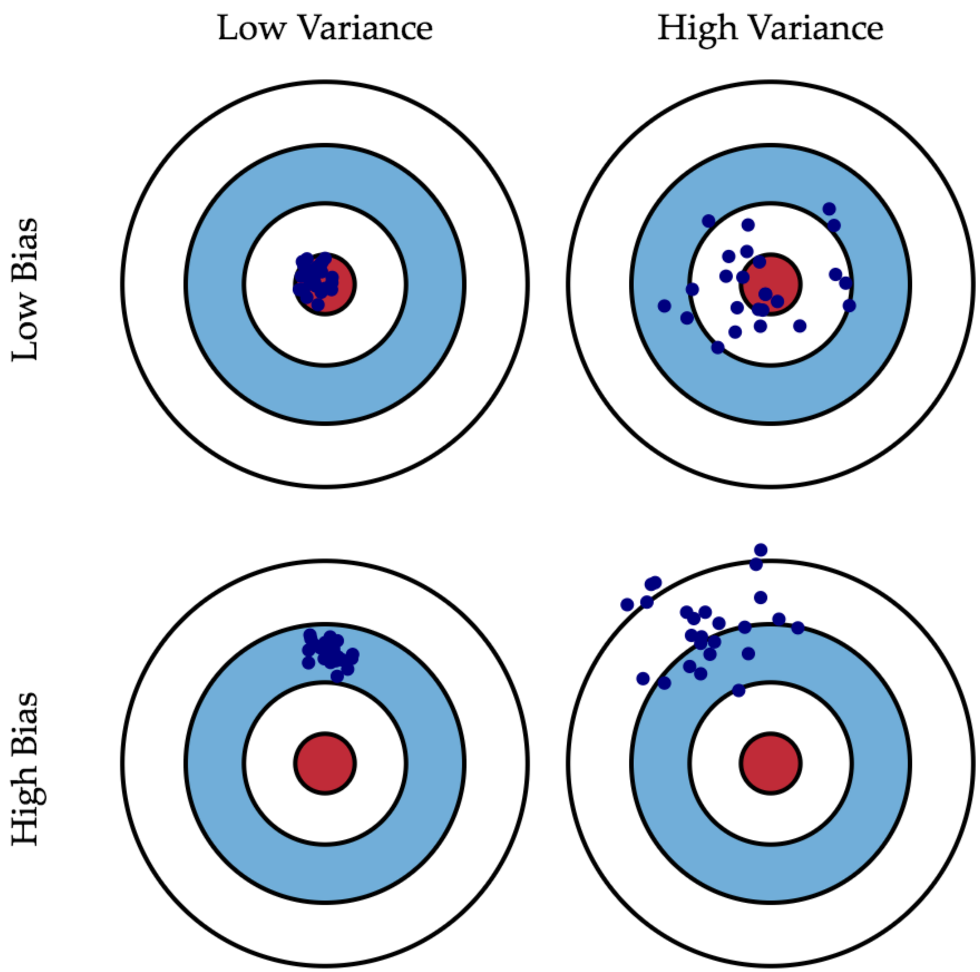

Non-linear model approximation of *f(x)*

# Model selection

- Model 1 is overly simplistic. It **under-fits** the data
- Model 2 is overly complex. It **over-fits** the data.
- Finding the right model is more than just reducing the error.



# The bias-variance tradeoff

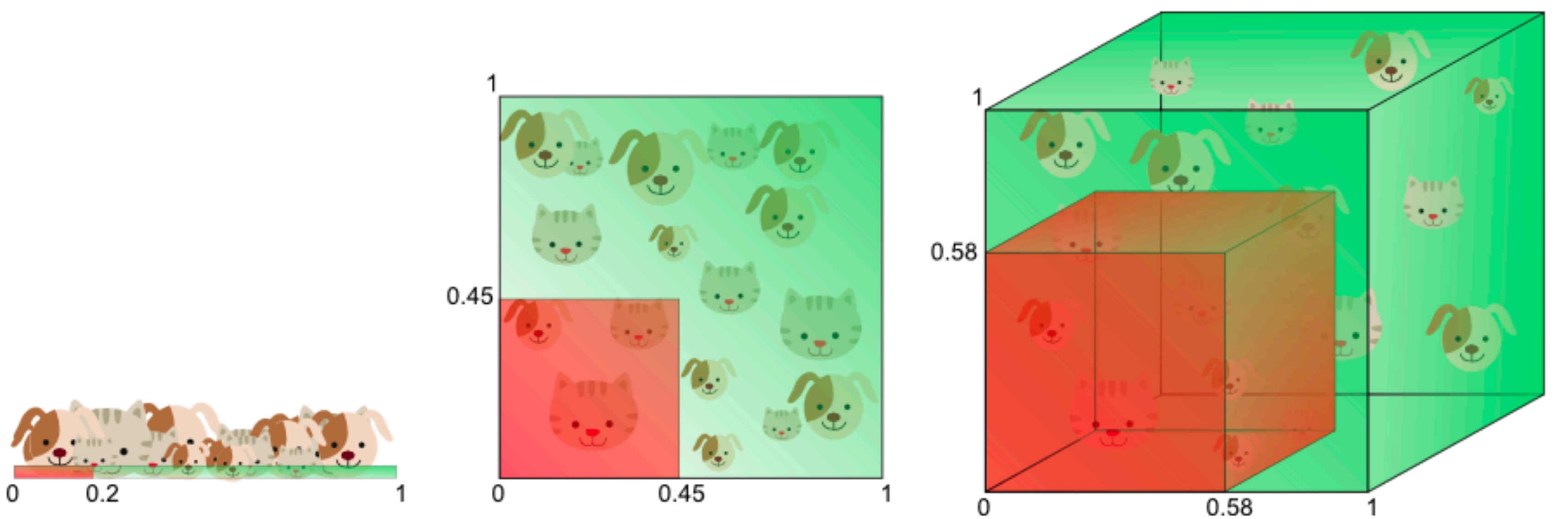



- ***Bias*** is how much the average model prediction differs from the desired value (*mean test error*).
- ***Variance*** is how much the model predictions change with each training data (*variance of test error*).

*(Image from "Understanding the Bias-Variance Tradeoff", by Scott Fortmann-Roe.)*

# The curse of dimensionality

- In order to maintain the same density, the number of data samples must grow exponentially with the number of dimensions in the feature space.
- When data is sparse, models tend to overfit.




*(Image from http://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/)*

# Regularization

- Penalty to model complexity that reduces the model's flexibility and improves generalization.

- Complexity = Tunable variables in the model.

# L1 and L2 norms

- L2 norm: $$\|\boldsymbol{w}\|_2 = \sum_{j}^{M} {w_j}^2$$
    - Gaussian prior on the weight distribution
- L1 norm: $$\|\boldsymbol{w}\|_1 = \sum_{j}^{M} |w_j|$$
    - Laplace prior on the weight distribution
    - Generates sparse solutions
    - Useful in supervised feature selection

# Regularized linear regression

- Ridge regression (L2):

$$Loss = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 + \gamma\sum_{j}^{M}{w_j}^2$$

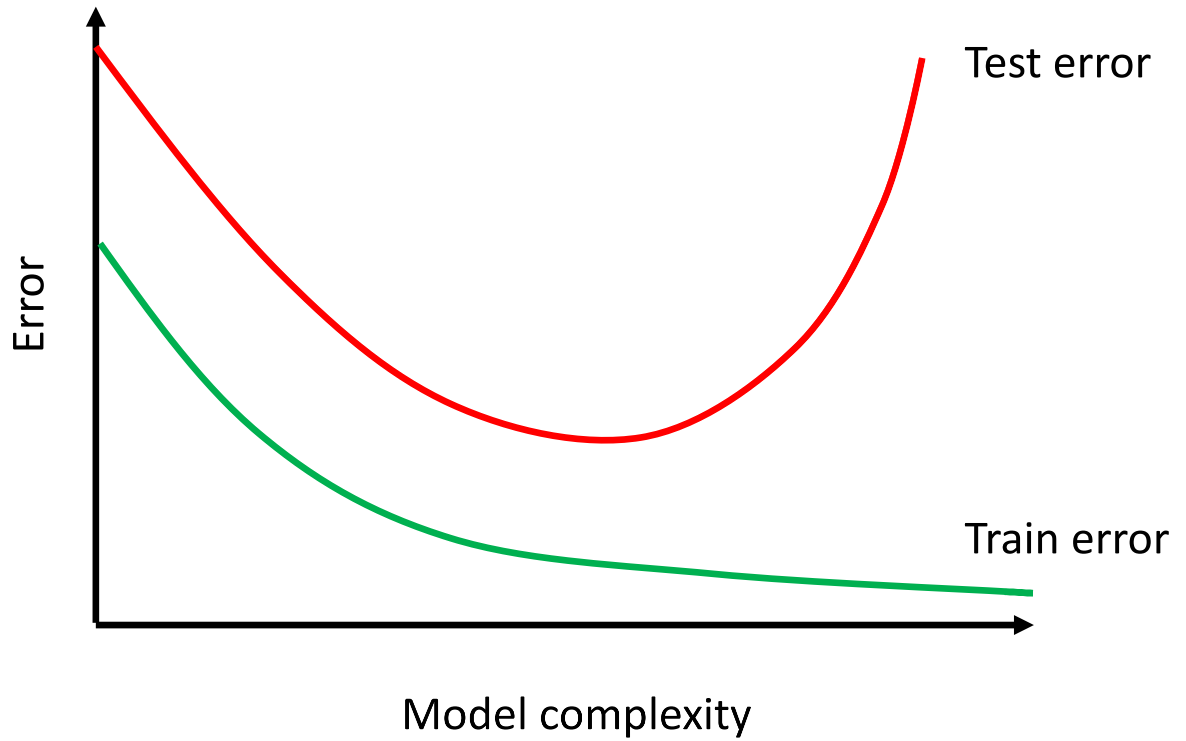
- Lasso regression (L1):

$$Loss = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 + \lambda\sum_{j}^{M}|w_j|$$

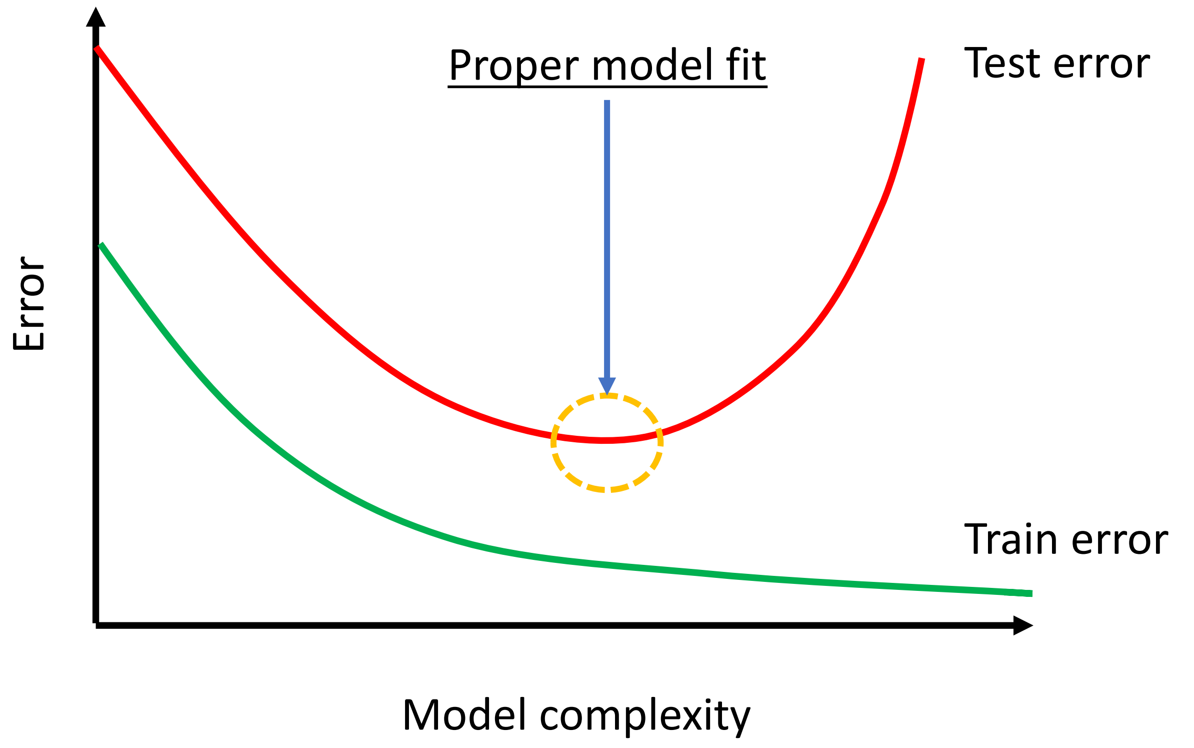
- Elastic-net (L1 + L2):

$$Loss = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 + \lambda\sum_{j}^{M}|w_j| + \gamma\sum_{j}^{M}{w_j}^2$$

# Hyper-parameter tuning



# Hyper-parameter tuning



# Model complexity

Statistical distributions

Hypothesis testing

Linear regression

Logistic regression

Decision trees

K-means clustering

Random forests

Gradient boosting (XGBoost)

Neural networks (deep learning)

Performance

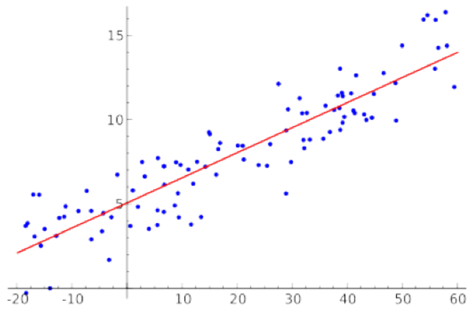
Explainability

# House prices in the real world

Features:

- # bedrooms
- # bathrooms
- Square footage
- Amenities (HDTV, internet, swimming pool, parking, etc.)
- Construction date
- Location

Model

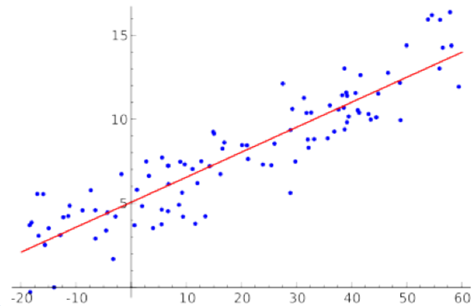



**Accuracy = 75%**

# House prices in the real world

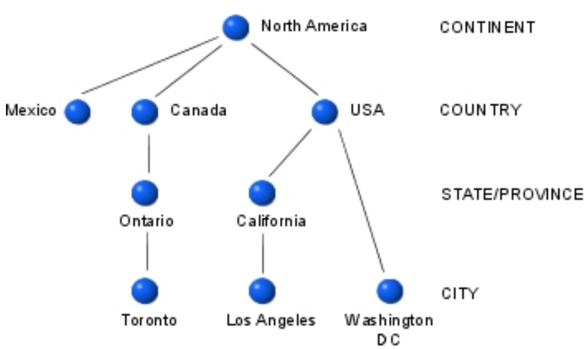
Features:
- # bedrooms
- # bathrooms
- Square footage
- Amenities (HDTV, internet, swimming pool, parking, etc.)
- Construction date
- Location

Model




**Accuracy = 82%**

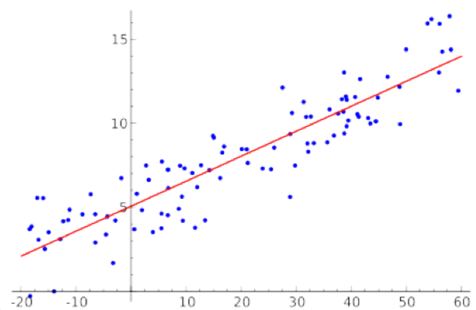
Location hierarchy



# House prices in the real world

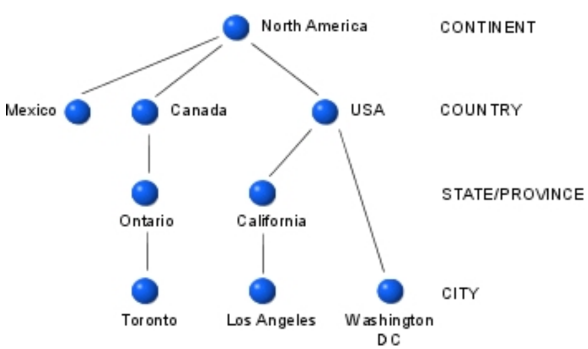
**Features:**
- # bedrooms
- # bathrooms
- Square footage
- Amenities (HDTV, internet, swimming pool, parking, etc.)
- Construction date
- Location

Model

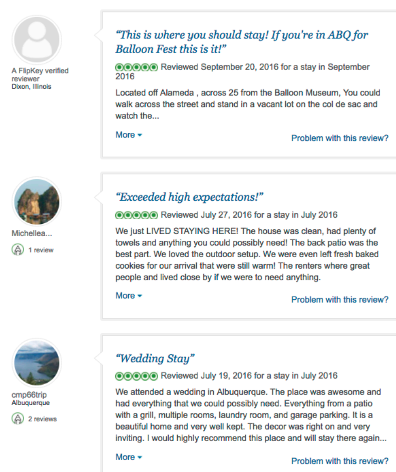



**Accuracy = 85%**

Location hierarchy




Text data



A FlipKey verified reviewer
Dixon, Illinois

*"This is where you should stay! If you're in ABQ for Balloon Fest this is it!"*

Reviewed September 20, 2016 for a stay in September 2016

Located off Alameda , across 25 from the Balloon Museum, You could walk across the street and stand in a vacant lot on the col de sac and watch the...

More

Problem with this review?

Michellea...
1 review

*"Exceeded high expectations!"*

Reviewed July 27, 2016 for a stay in July 2016

We just LIVED STAYING HERE! The house was clean, had plenty of towels and anything you could possibly need! The back patio was the best part. We loved the outdoor setup. We were even left fresh baked cookies for our arrival that were still warm! The renters where great people and lived close by if we were to need anything.

More

Problem with this review?

cmp66trip
Albuquerque
2 reviews

*"Wedding Stay"*

Reviewed July 19, 2016 for a stay in July 2016

We attended a wedding in Albuquerque. The place was awesome and had everything that we could possibly need. Everything from a patio with a grill, multiple rooms, laundry room, and garage parking. It is a beautiful home and very well kept. The decor was right on and very inviting. I would highly recommend this place and will stay there again...

More

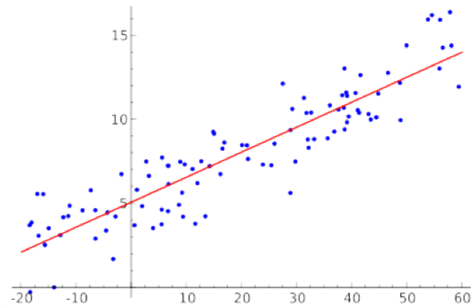
Problem with this review?

# House prices in the real world

Features:

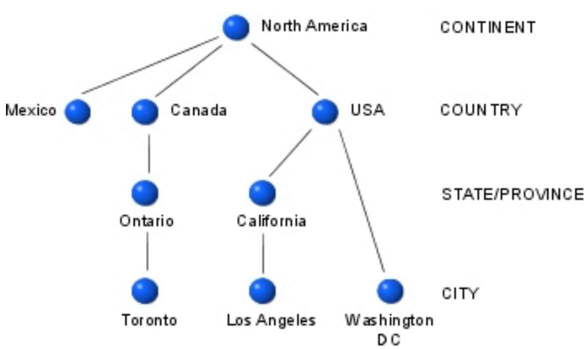
- # bedrooms
- # bathrooms
- Square footage
- Amenities (HDTV, internet, swimming pool, parking, etc.)
- Construction date
- Location

Model

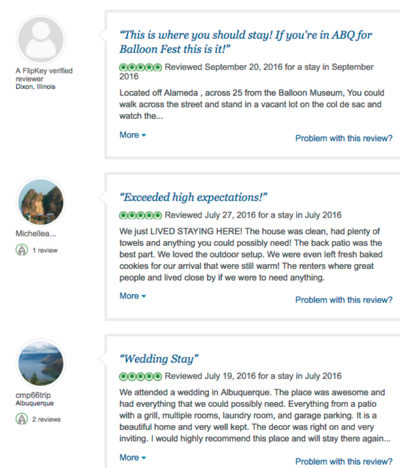



Accuracy = 87%

Location hierarchy




Text data



A FlipKey verified reviewer
Dixon, Illinois

"This is where you should stay! If you're in ABQ for Balloon Fest this is it!"

Reviewed September 20, 2016 for a stay in September 2016

Located off Alameda , across 25 from the Balloon Museum, You could walk across the street and stand in a vacant lot on the col de sac and watch the...

More ▾ Problem with this review?

Michellea...
1 review

"Exceeded high expectations!"

Reviewed July 27, 2016 for a stay in July 2016

We just LIVED STAYING HERE! The house was clean, had plenty of towels and anything you could possibly need! The back patio was the best part. We loved the outdoor setup. We were even left fresh baked cookies for our arrival that were still warm! The renters where great people and lived close by if we were to need anything.

More ▾ Problem with this review?

cmp66trip
Albuquerque
2 reviews

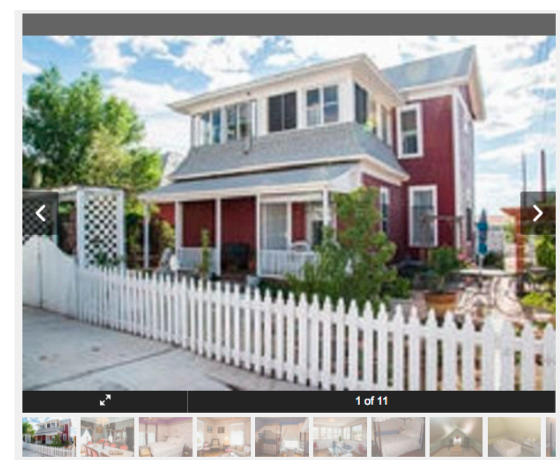
"Wedding Stay"

Reviewed July 19, 2016 for a stay in July 2016

We attended a wedding in Albuquerque. The place was awesome and had everything that we could possibly need. Everything from a patio with a grill, multiple rooms, laundry room, and garage parking. It is a beautiful home and very well kept. The decor was right on and very inviting. I would highly recommend this place and will stay there again...

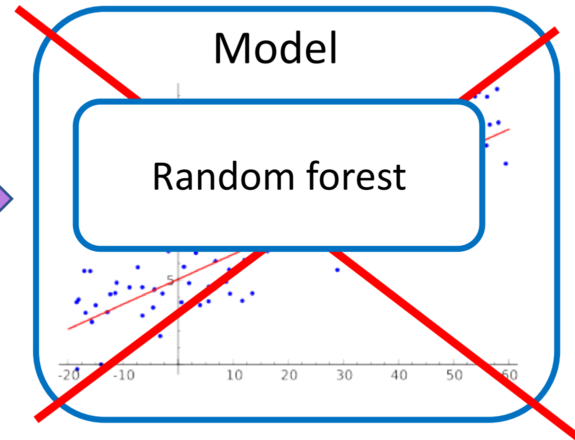
More ▾ Problem with this review?

Photos



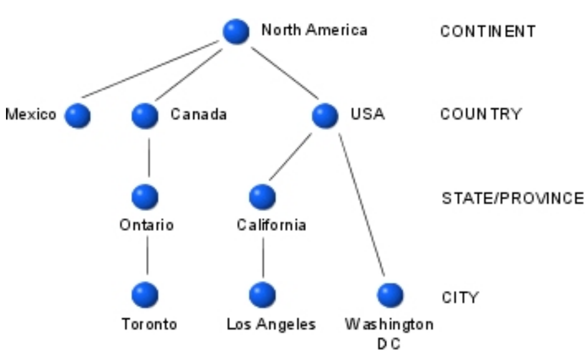
# House prices in the real world

**Features:**
- # bedrooms
- # bathrooms
- Square footage
- Amenities (HDTV, internet, swimming pool, parking, etc.)
- Construction date
- Location

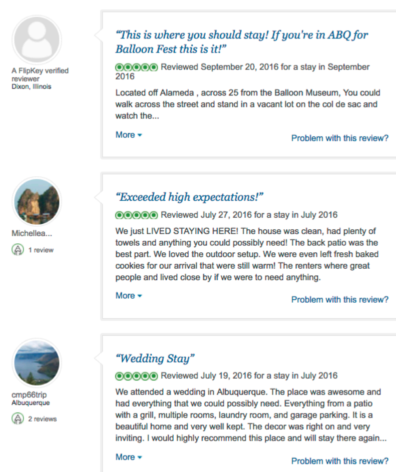



**Accuracy = 90%**

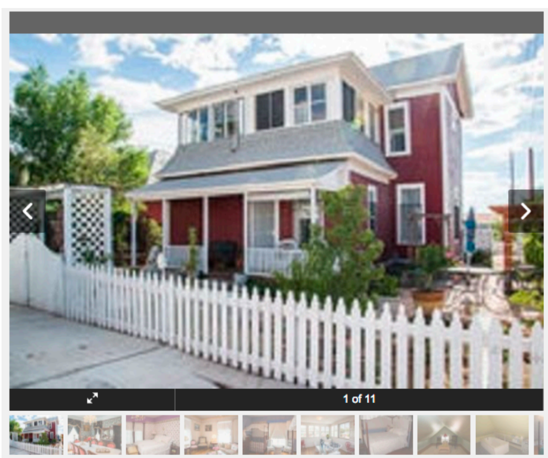
Location hierarchy




Text data

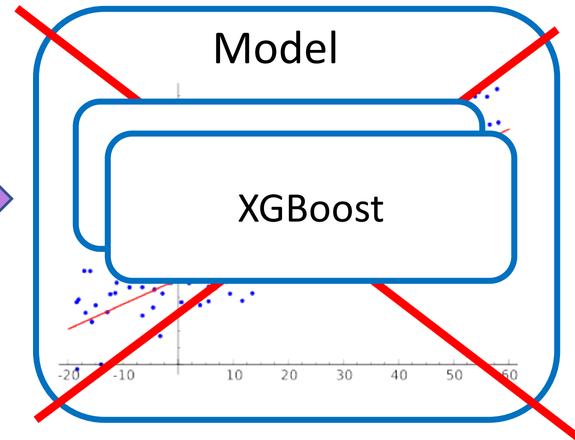



Photos



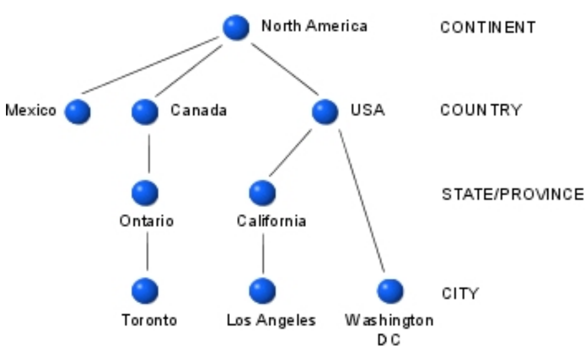
# House prices in the real world

Features:
- # bedrooms
- # bathrooms
- Square footage
- Amenities (HDTV, internet, swimming pool, parking, etc.)
- Construction date
- Location

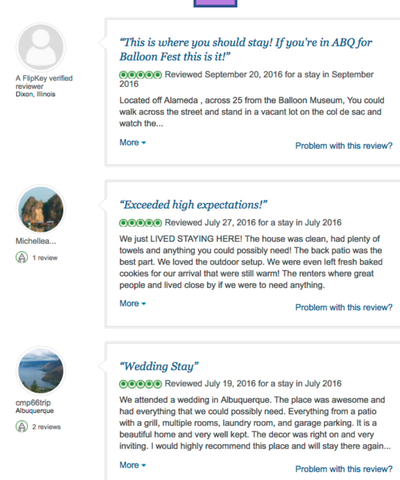



**Accuracy = 93%**

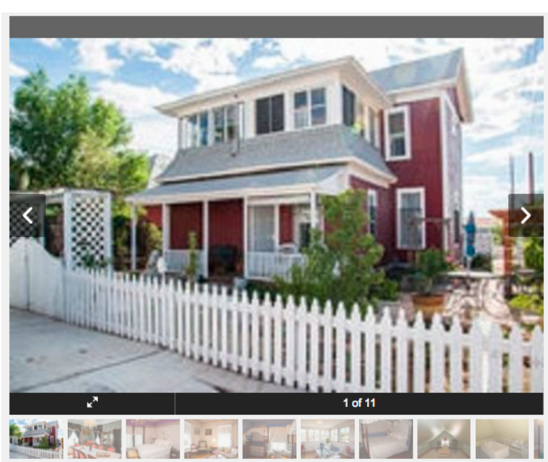
Location hierarchy




Text data

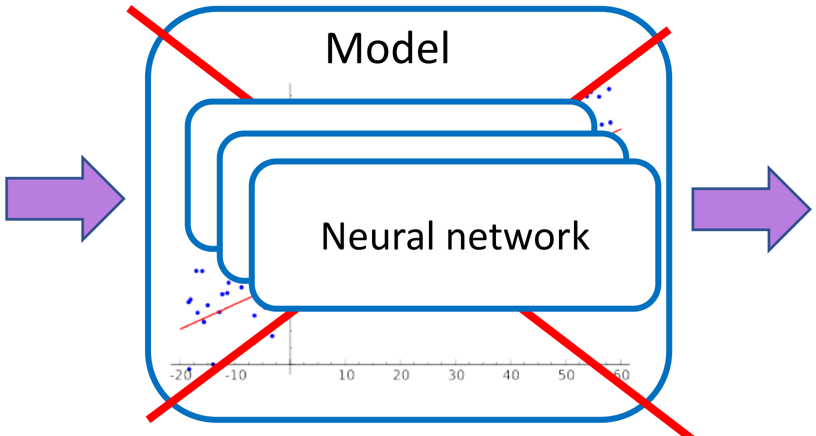



Photos



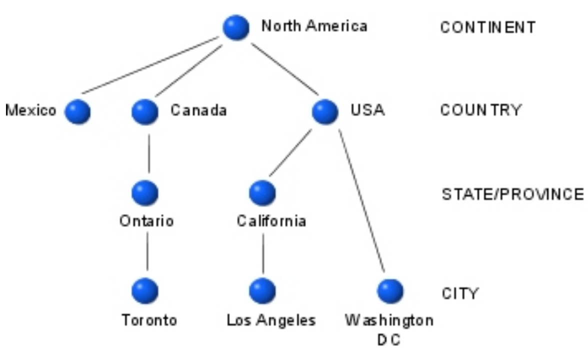
# House prices in the real world

Features:
- # bedrooms
- # bathrooms
- Square footage
- Amenities (HDTV, internet, swimming pool, parking, etc.)
- Construction date
- Location

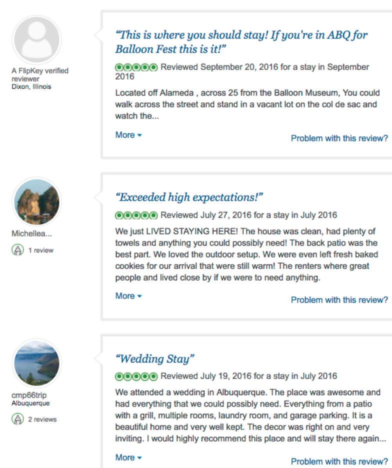



**Accuracy = 94%**

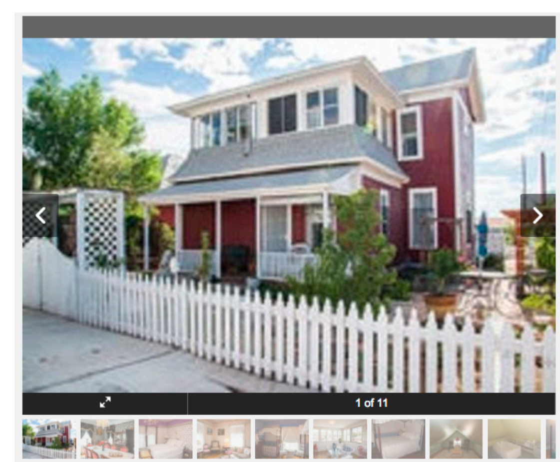
Location hierarchy




Text data



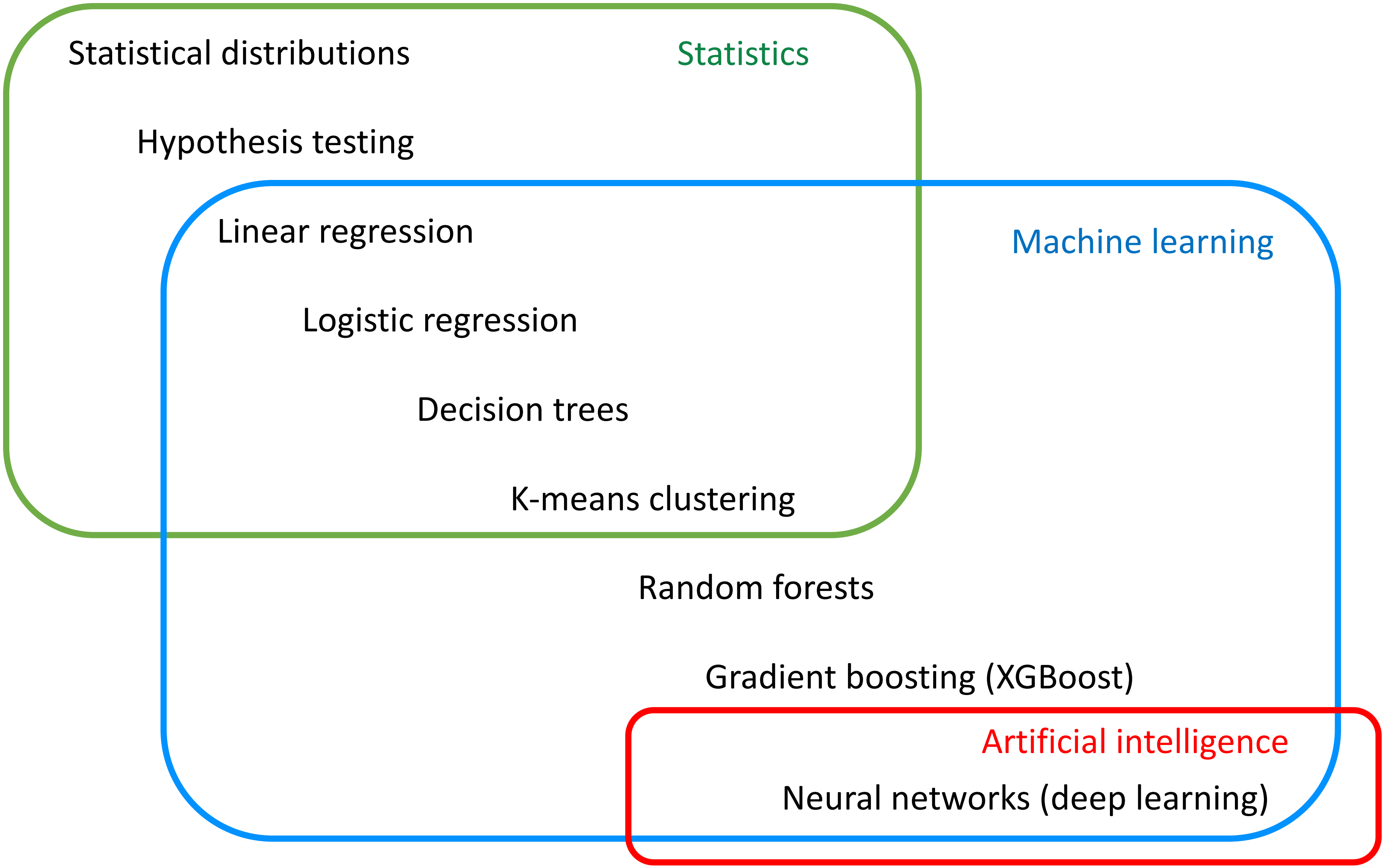

Photos



# Recap (Part 1)

- Most problems in machine learning can be reduced to only 4 types: Clustering, regression, classification, and dimensionality reduction.
- The goal of a model is **not** to reduce error/increase accuracy but to <u>generalize to unseen data</u>.
- Model complexity increases flexibility, but requires more training data and reduces interpretability.
- Regularization techniques constrain model complexity and improve generalization.
- Model validation techniques can be used to properly tune hyperparameters and achieve the right bias-variance trade-off.

# Statistics x ML x AI



# What is machine learning?

- The science of building computer models that:
  - Learn from data how to perform a task
  - Self-tune their parameters to optimize performance
  - Generalize behavior to new/unseen data