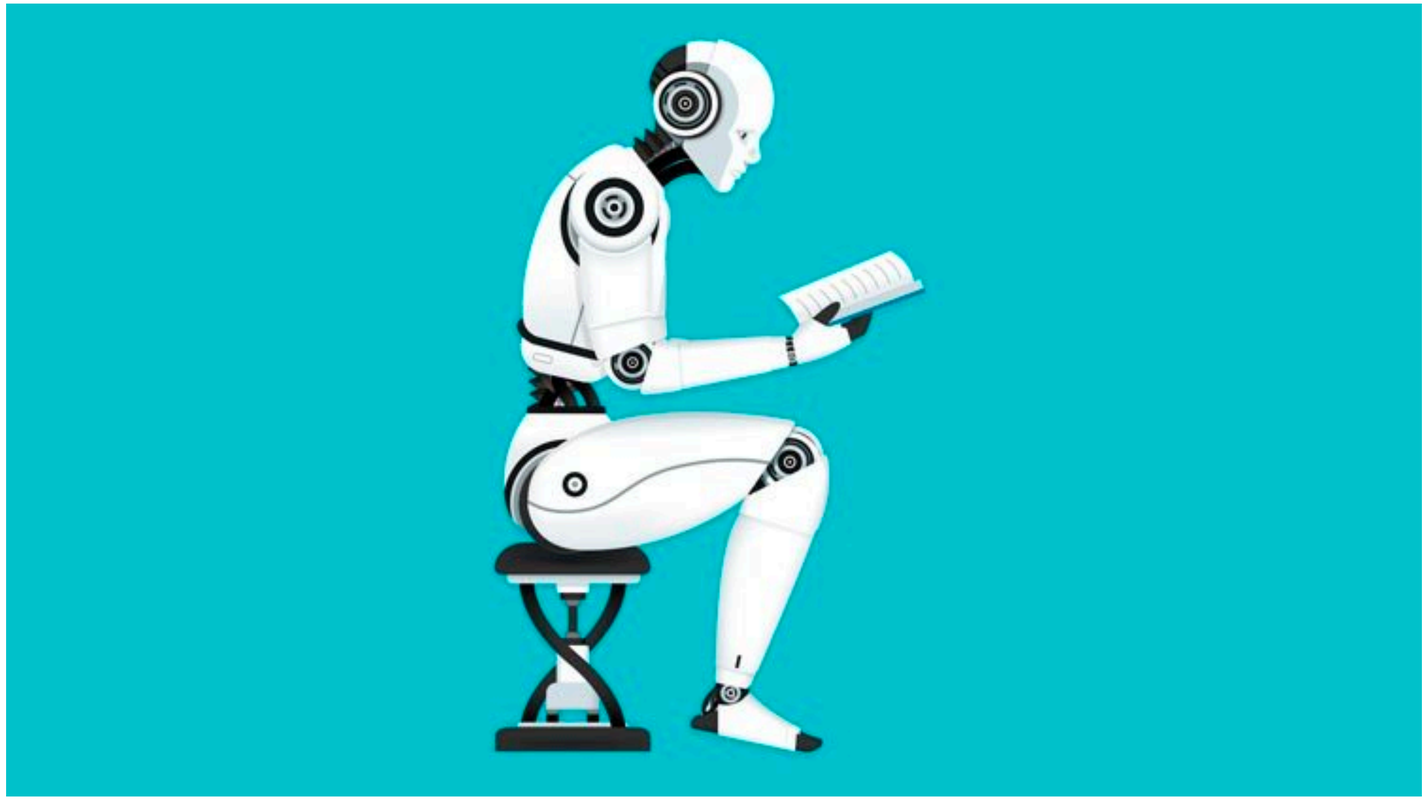
UNIVERSITAT DE BARCELONA



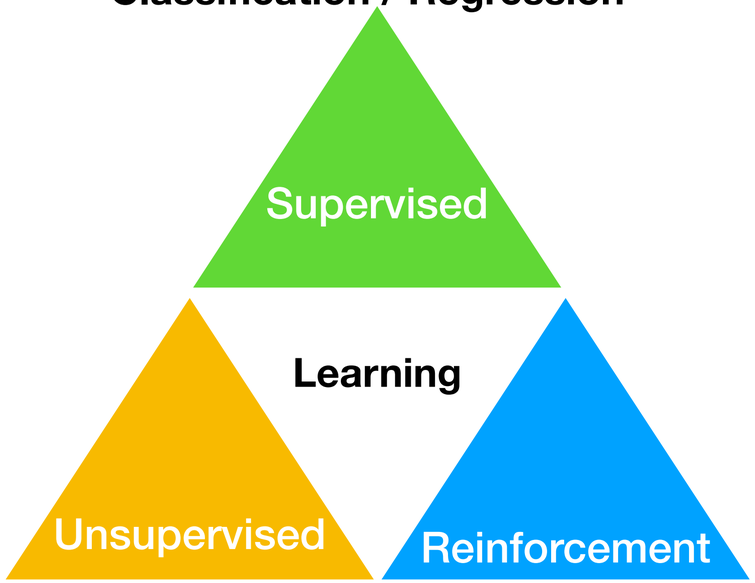
A typical Machine Learning Project

Machine Learning | Enginyeria Informàtica

Santi Seguí | 2022-2023

Type of Machine Learning

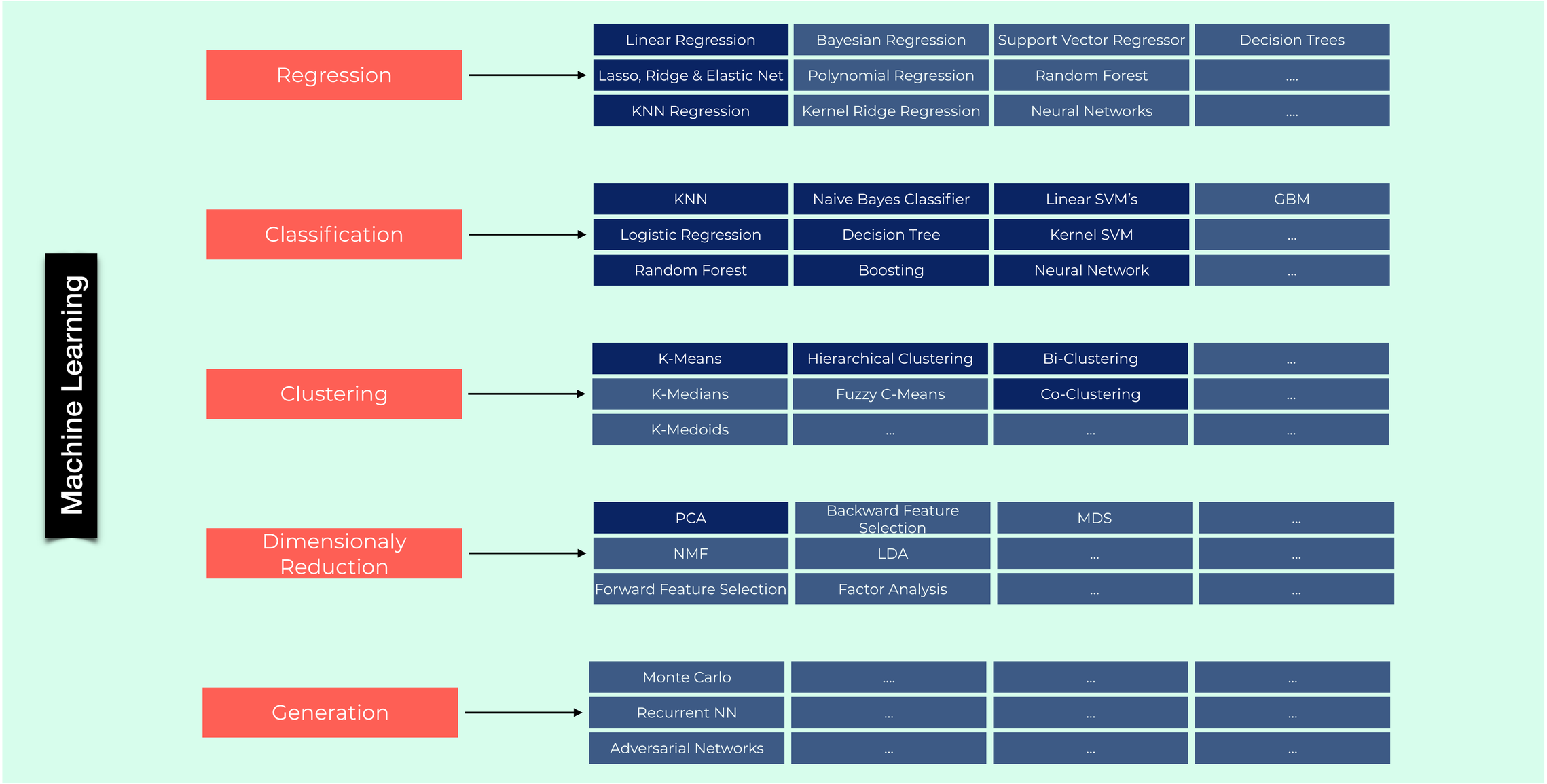
Labeled Data
Direct Feedback
Predict outcome/future
Classification / Regression



No labels
No feedback
“find hidden structure”
Clustering

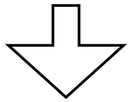
Decision process
Reward system
Learn series of actions
Learn from mistakes

Classification	**To which data category does this data point belongs?**	**Medical diagnosis**: Does this tissue show signs of diseases? **Banking**: Is this transaction fraudulent? **Computer Vision**: What type of object is in this picture? Is it a person? Is it a building?
Regression	**Given this input from a dataset, what is the likely value of a particular quantity?**	**Finance:** What is the value of this stock going to be tomorrow? **Housing:** What would be the price of this house if it was sold today? **Food Quality**: When should I pick this strawberry?
Clustering	**Which data point are similar to each other?**	**E-Commerce:** which customers are exhibiting similar behavior to each other, how do they group together? **Video Streaming**: what are the different types of video genres in our catalogue, and which ones are in the same genre?
Dimensionality reduction	**What are the most significant features of this data and how can be summarized?**	**E-Commerce:** What combinations of features allows us to summarize the behavior of our customers? **Molecular biology:** How can scientists summarize the behavior of all 20.000 human gens in a particular diseases?
Semi-supervised learning	**How can be labelled and unlabelled data be combined?**	**Computer Vision:** How can an object detection be developed, with only a small training data set? **Drug Discovery:** Which of the million possible drugs could be effective agains disease, we have so far only tested a few?
Reinforcement learning	**What actions will most effectively achieve a desired endpoint?**	**Robots:** How can a robot move through its environment? **Games:** Which moves were more important in helping the computer win a particular game?

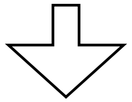

Machine Learning
Regression
Linear Regression
Bayesian Regression
Support Vector Regressor
Decision Trees
Lasso, Ridge & Elastic Net
Polynomial Regression
Random Forest
....
KNN Regression
Kernel Ridge Regression
Neural Networks
....
Classification
KNN
Naive Bayes Classifier
Linear SVM's
GBM
Logistic Regression
Decision Tree
Kernel SVM
...
Random Forest
Boosting
Neural Network
...
Clustering
K-Means
Hierarchical Clustering
Bi-Clustering
...
K-Medians
Fuzzy C-Means
Co-Clustering
...
K-Medoids
...
...
...
Dimensionaly Reduction
PCA
Backward Feature Selection
MDS
...
NMF
LDA
...
...
Forward Feature Selection
Factor Analysis
...
...
Generation
Monte Carlo
....
...
...
Recurrent NN
...
...
...
Adversarial Networks
...
...
...

No Free Lunch Theorem

Our model is a simplification of reality



Simplification is based on assumptions (bias)



Assumptions fail in certain situations

Roughly speaking: "**There is not a model that works best for all possible situations**."

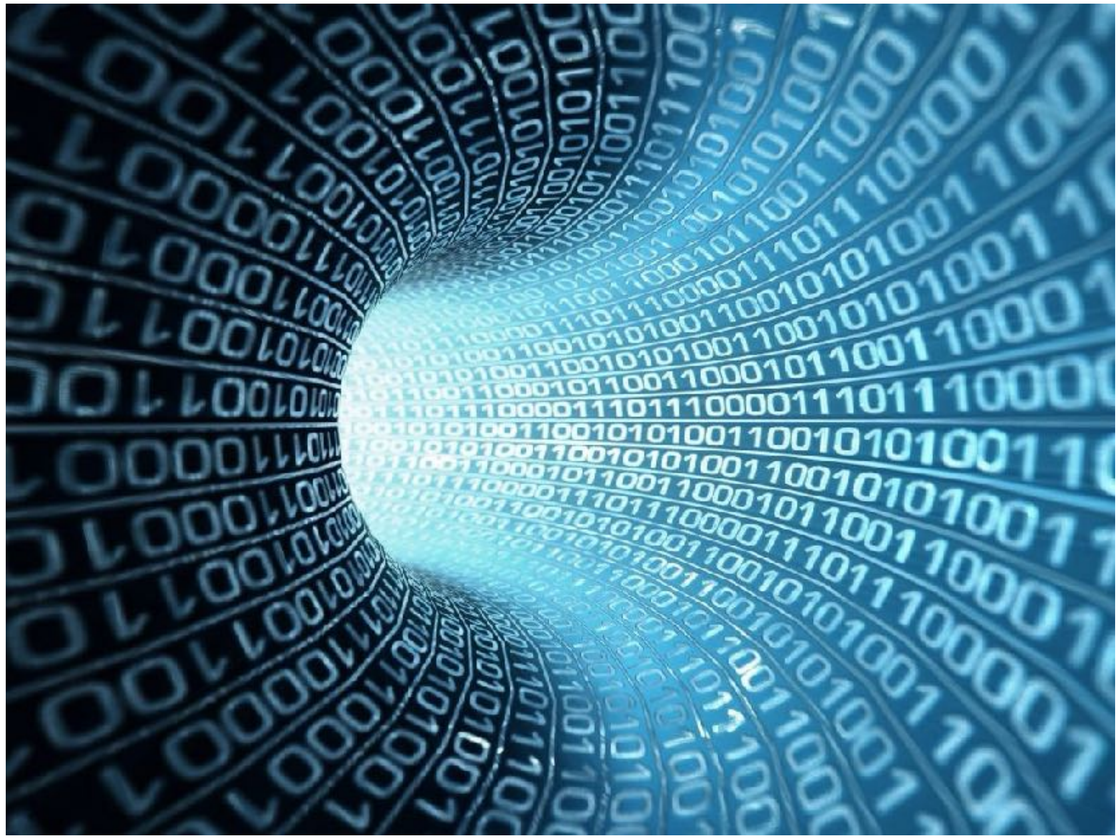
- **Data** are predicted
  - on the basis of a set of **features** (e.g. clinical measurements);
  - from a set of (observed) **training data** on these features;
  - for a set of **objects/instances** (e.g. people).

- **Inputs** for the problems are also called predictor or independent variables
- **Outputs** are also called responses or dependent variables

- The predictor model is also called estimator

Data?

- What is data? Which types of data exist?



Data

- Numerical Values (Age, Salary, Blood pressure,…)
- Categories (blood group, city,..)
- Images
- Videos
- Text
- Genes
- ….

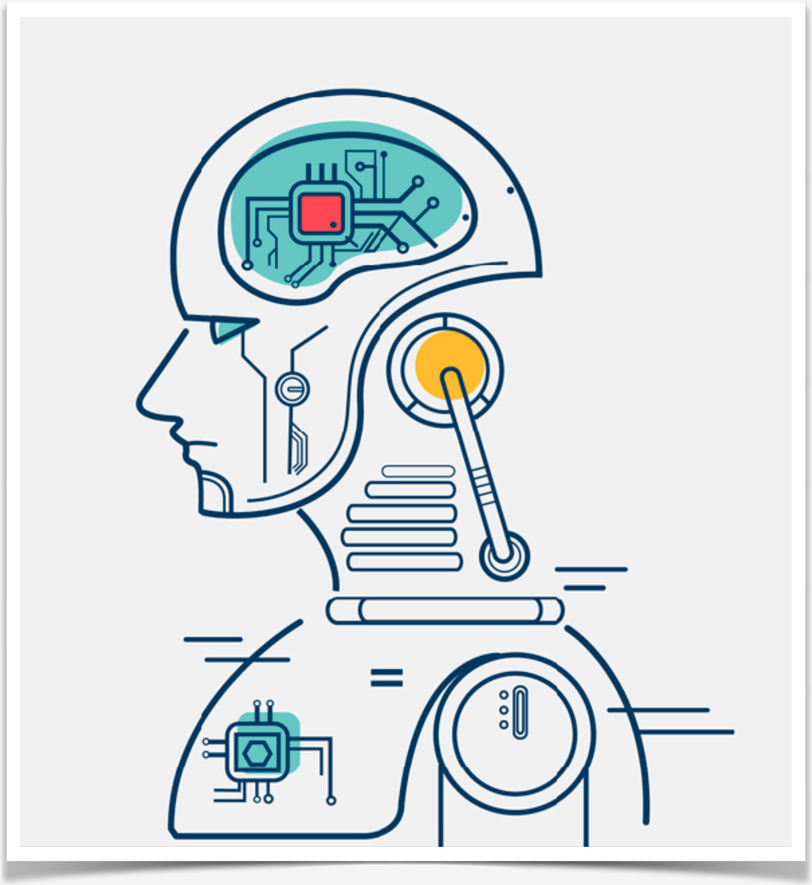
https://github.com/ssegui/ml_ub/blob/master/notebooks/Session1.ipynb

First Project:

https://www.kaggle.com/t/b629fe86ac2d4b4eb917f079db9ddc18

Score:

1 point + 0.25 for the winner



The typical Machine Learning Project

- **The main step of any machine learning project can be resumed as:**
  1. Look at the big picture
  2. Get the data
  3. Discover and visualize the data to gain insights
  4. Prepare the data for Machine Learning algorithms
  5. Select a model and train it
  6. Fine-tune your model
  7. Present your solution
  8. Launch, monitor and maintain your system

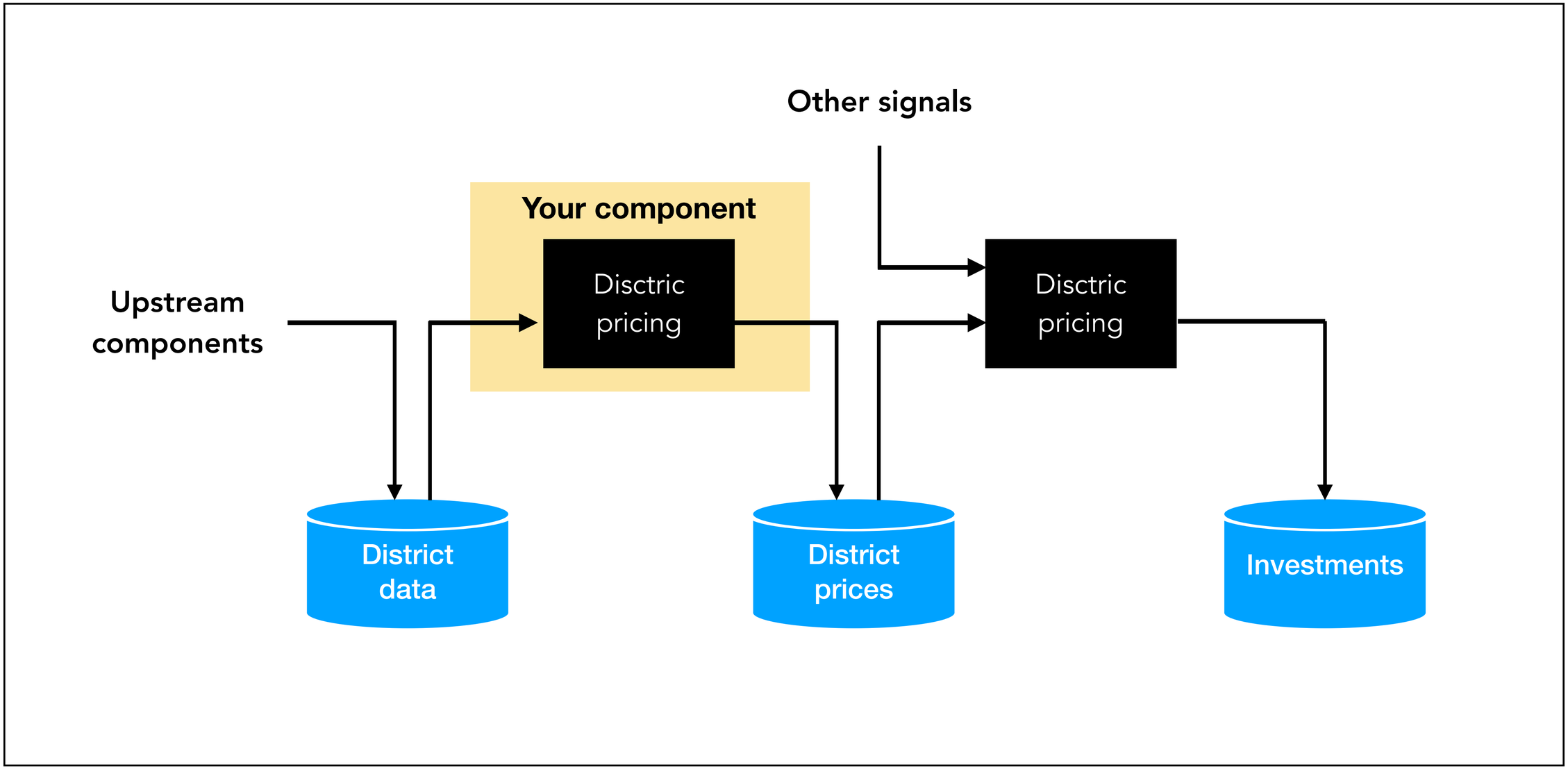
Look and the big Picture

Housing price prediction

A model that uses California census to predict the median housing price in any district

The typical Machine Learning Project

- **Look at the Big Picture**
    1. Frame the Problem
        1. First question to ask is what the exactly the “business” objective is.
        2. Building a model sometimes is not the end goal.
            1. How does the “company” expect to use and benefit from this model?
        3. Knowing the objective will determine how to frame the problem:
            1. Which algorithms to select;
            2. Which performance measure you will use to frame the problem;
            3. How much effort (or accuracy) you will spend with it
        4. Is there a current method. If not
    2. Select the Performance Measure
    3. Check the Assumptions



A Machine Learning pipeline for real state investments

Select the Performance Measure

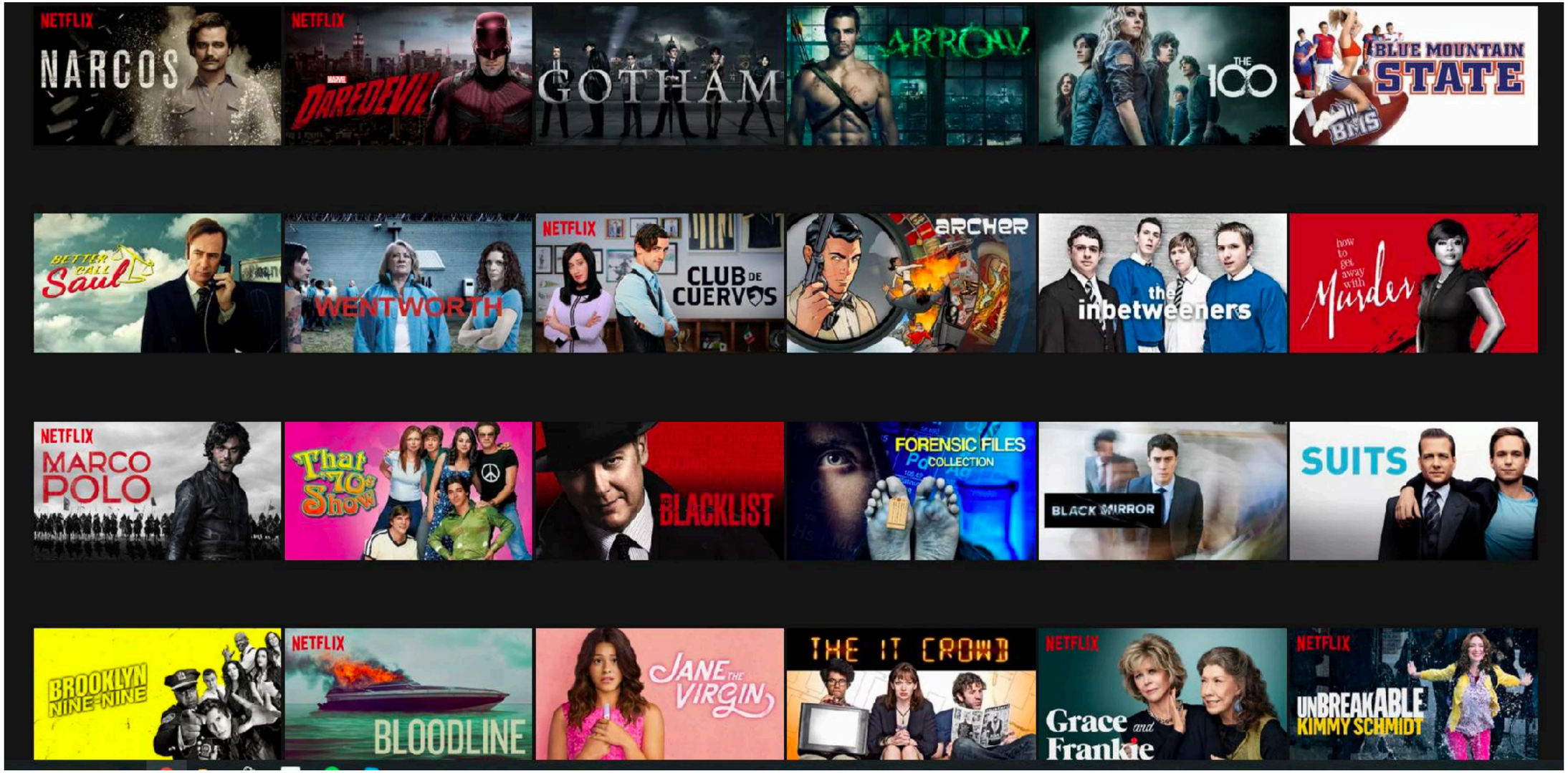
- A typical performance measure for regression problems:
  - Root Mean Square Error (RMSE):

$$RMSE(X, h) = \sqrt{1/m \sum_{i=1}^{m} (h(\boldsymbol{x}^{(i)}) - y^{(i)})^2}$$

  - Mean Absolute Error

$$MAE(\boldsymbol{X}, h) = \frac{1}{m} \sum_{i=1}^{m} | h(\boldsymbol{x}^{(1)}) - y^{(i)} |$$

Look at the Big Picture of this Problem



Get the data

- **Data** are predicted
  - on the basis of a set of **features** (e.g. clinical measurements);
  - from a set of (observed) **training data** on these features;
  - for a set of **objects/instances** (e.g. people).

- **Inputs** for the problems are also called predictor or independent variables
- **Outputs** are also called responses or dependent variables

- The predictor model is also called estimator

Get the Data

- Data can be available from different sources and formats:
  - txt, csv, xls, xlxs, structured databases, unstructured databases,…

Create your a Test set

Train Data

Test Data

Used to **create/train** your model

Used to **evaluate** your model performance

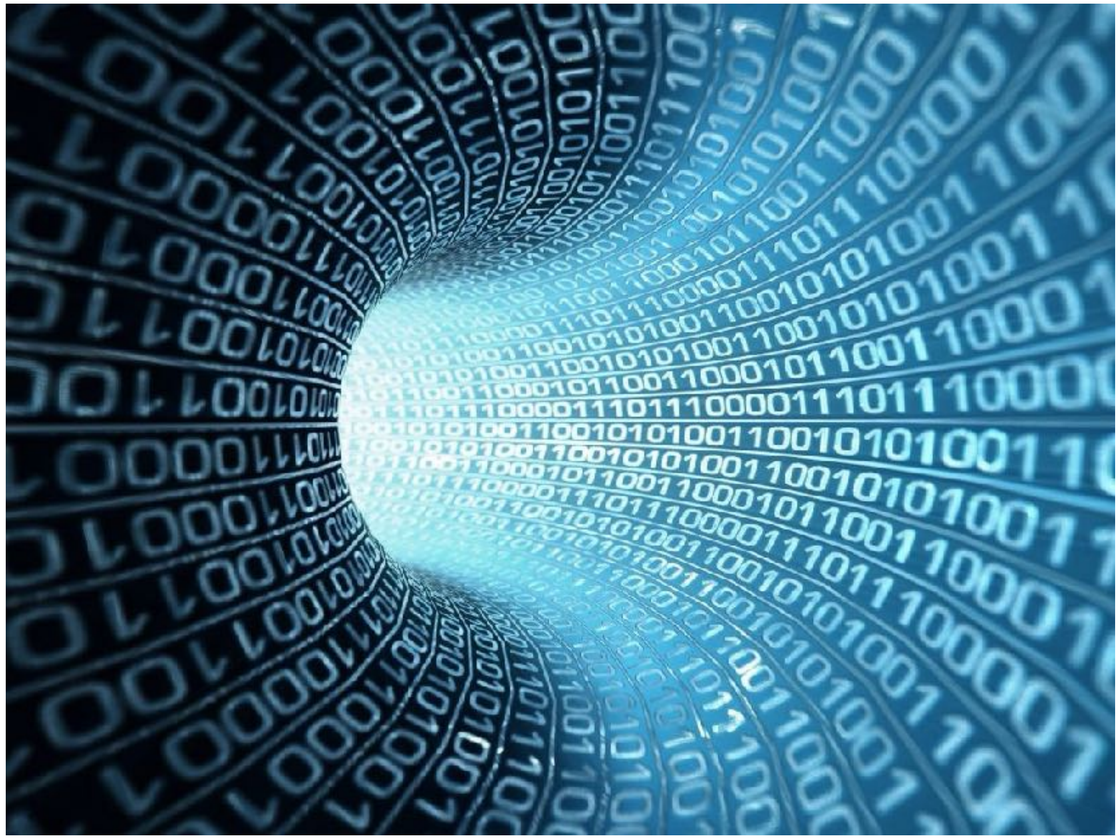
Any decision can be taken using this data.
Must be hidden until the model is ready

and hide it!!!!

Discover and visualize the data to gain insights

Data?

- What is data? Which types of data exist?

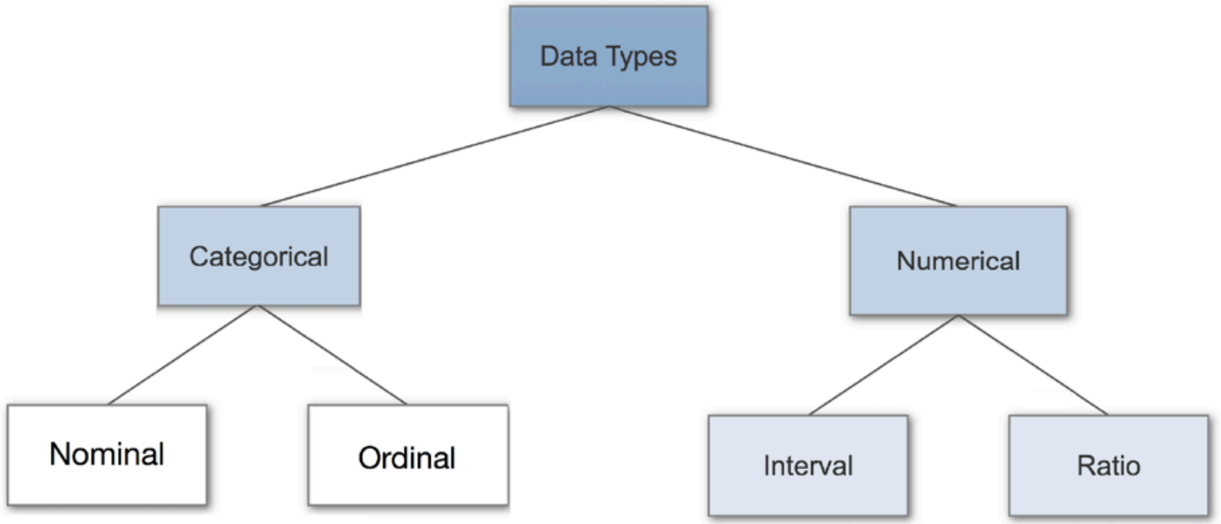


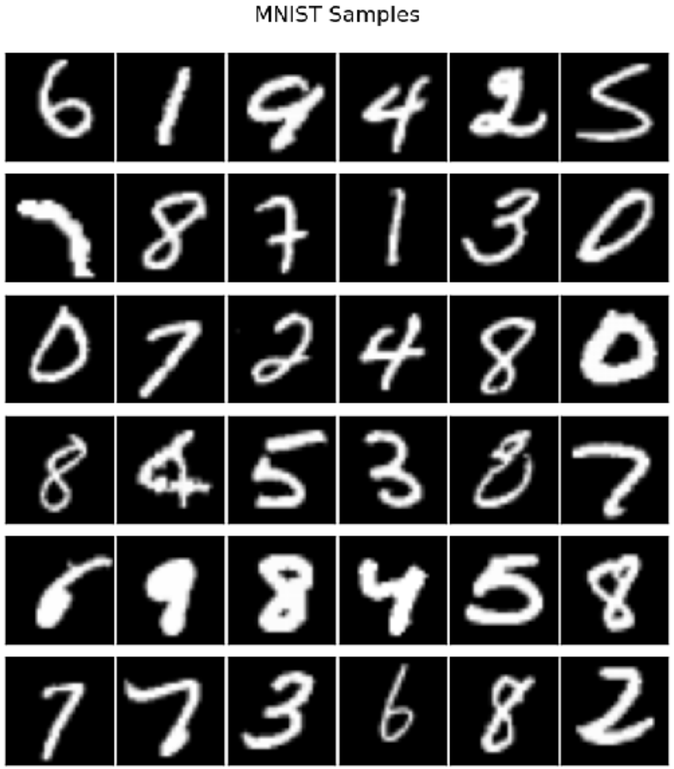
Data

- Numerical Values (Age, Salary, Blood pressure,…)
- Categories (blood group, city,..)
- Images
- Videos
- Text
- Genes
- ….

Data

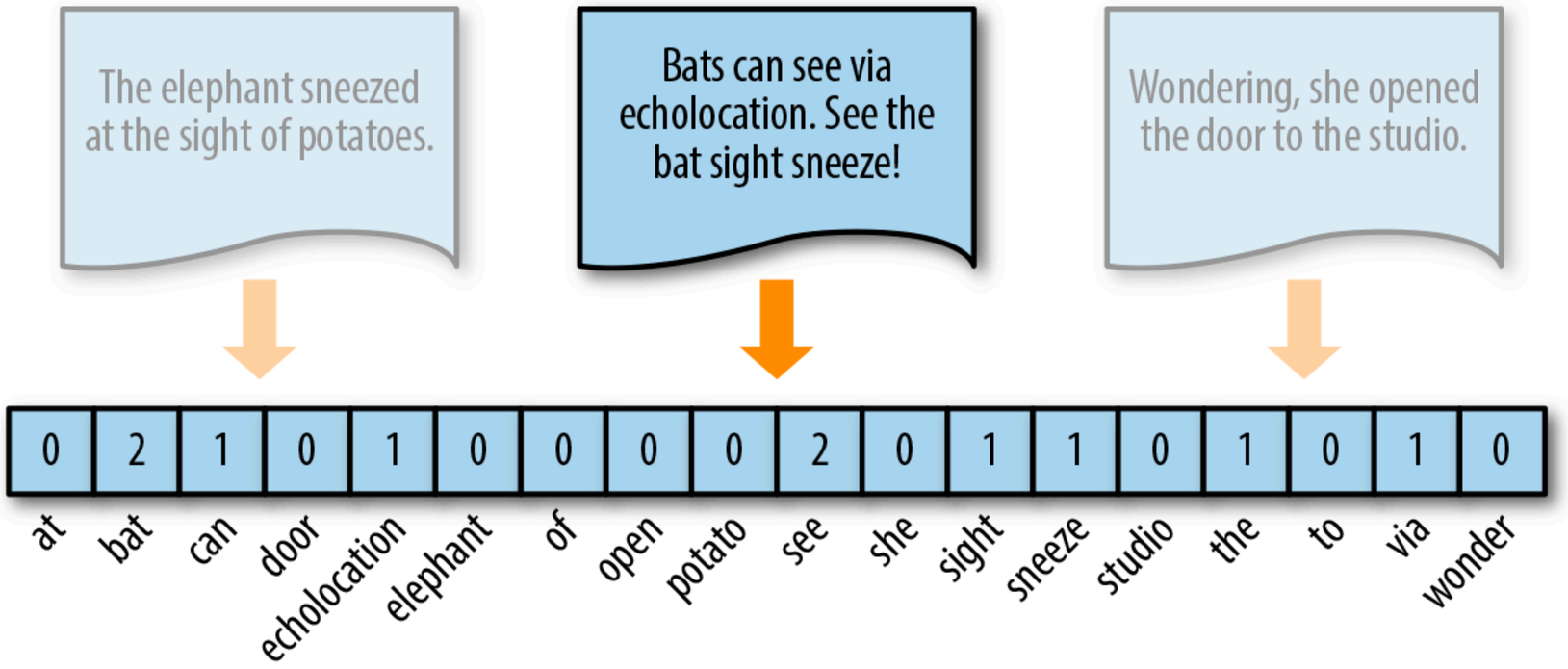
- What is data? Which types of data exist?
  - **Quantitative** (numerical): e.g. price, age,..
  - **Categorical** (discrete, often binary): Cancer/no cancer





$x \in \chi \subset \{0, 1, ..., 255\}^{400}$ is a vector of pixel intensities in a $20x20$ images.

Text representation: Bag of Words



$x_i = (x_1^i, x_2^i, \cdots, x_d^i)$ **where** x_i^α **is the number of occurrences of the** α^{th} **word in a dictionary document** i

Statistic Concepts (a review)

Review: Some statistic concepts

- **Mean** (or arithmetic mean):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_1 + \ldots + x_n}{n}$$

  sometimes is called μ

- **Median**:

If n is odd then $M_e = x^*_{((n+1)/2)}$

If n is even then $M_e = \frac{x^*_{n/2} + x^*_{(n/2+1)}}{2}$

where x^* is the sorted version of x.

Review: Some statistic concepts

- **Variance** (Var(X)): measures how far a set of numbers are spread out

$$Var(X) = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$std(X) = \sigma = \sqrt{Var(X)}$$

Review: Some statistic concepts

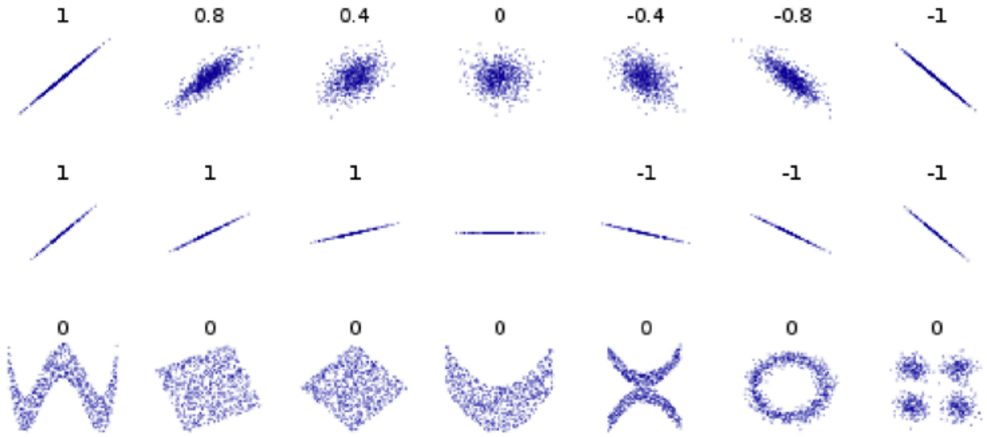
- **Covariance** (Cov(X,Y)) : measures how much two random variables change together:

$$Cov(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Review: Some statistic concepts

- **Correlation** (p): is a measure of the linear correlation between two variables X and Y, giving a value between +1 and -1 inclusive, where 1 is total positive correlation, 0 is no correlation, and -1 is total negative correlation.

$$\rho(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$

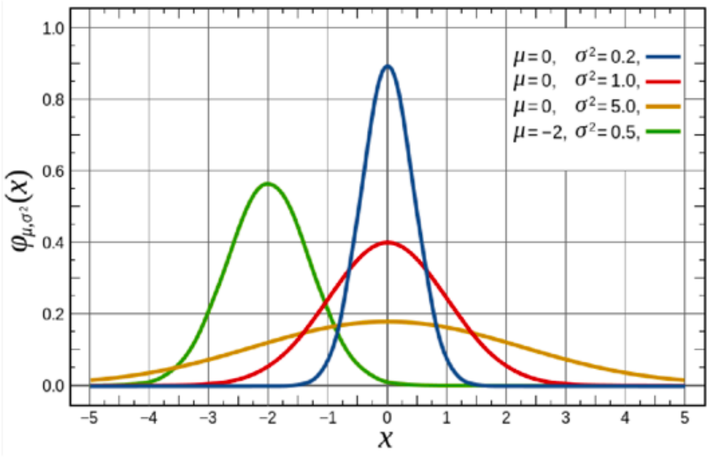


Review: Some statistic concepts

- **Normal** (or **gaussian**) **distribution** ($\mathcal{N}(\mu, \sigma^2)$)

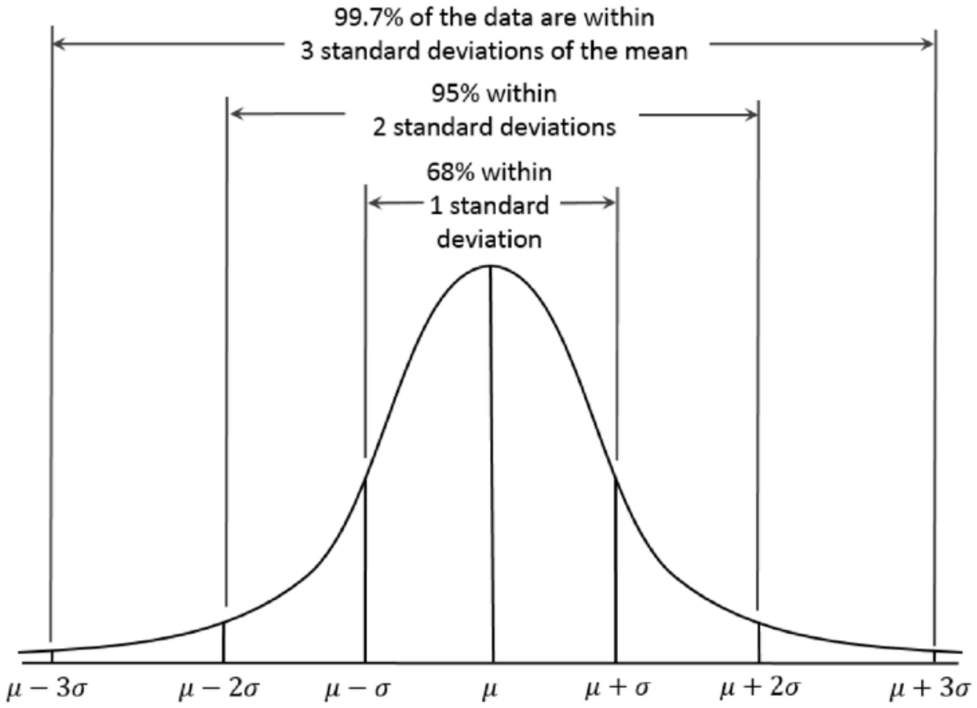
- The probability function of a normal distribution is:

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$



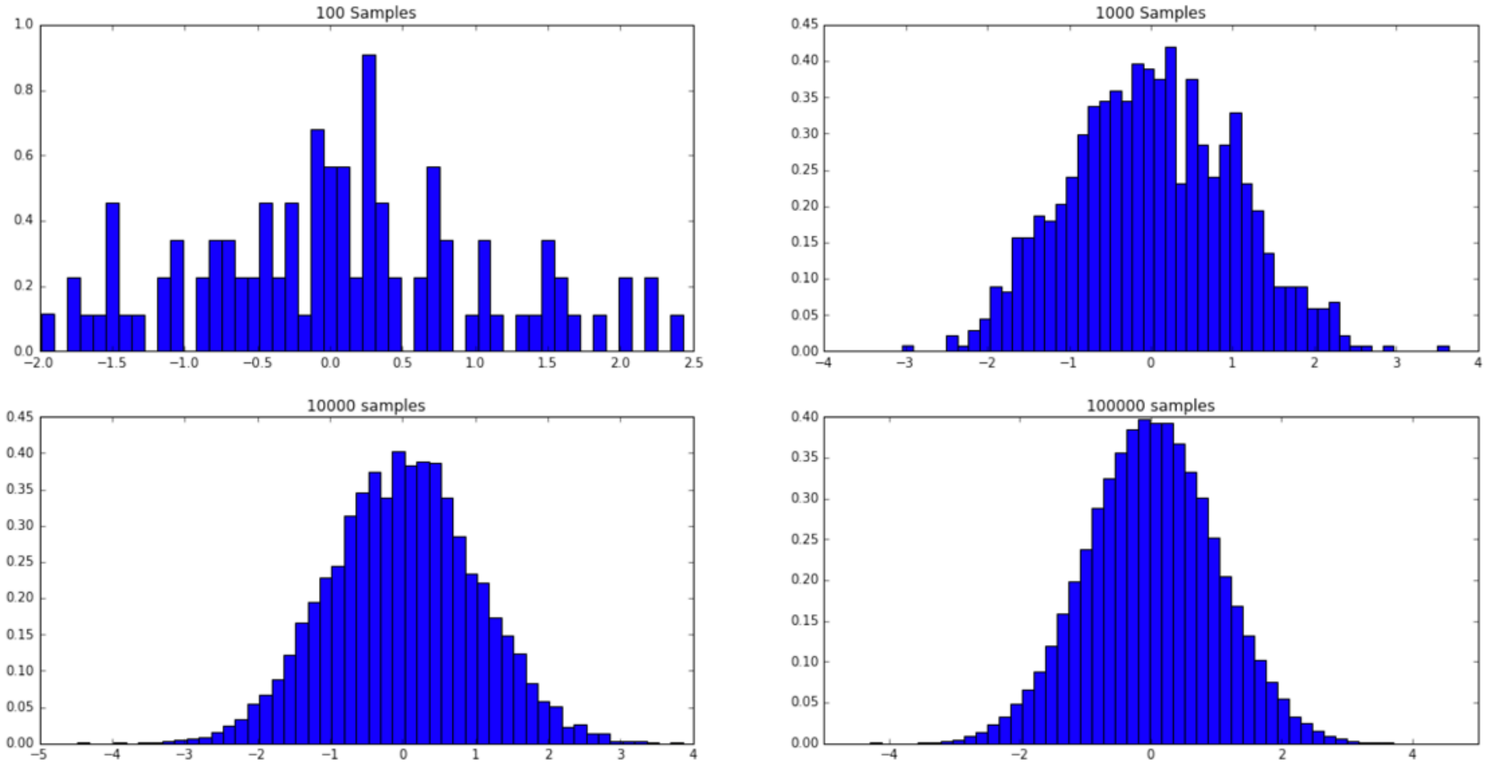
Review: Some statistic concepts

- Normal (or gaussian) distribution ($\mathcal{N}(\mu, \sigma^2)$)

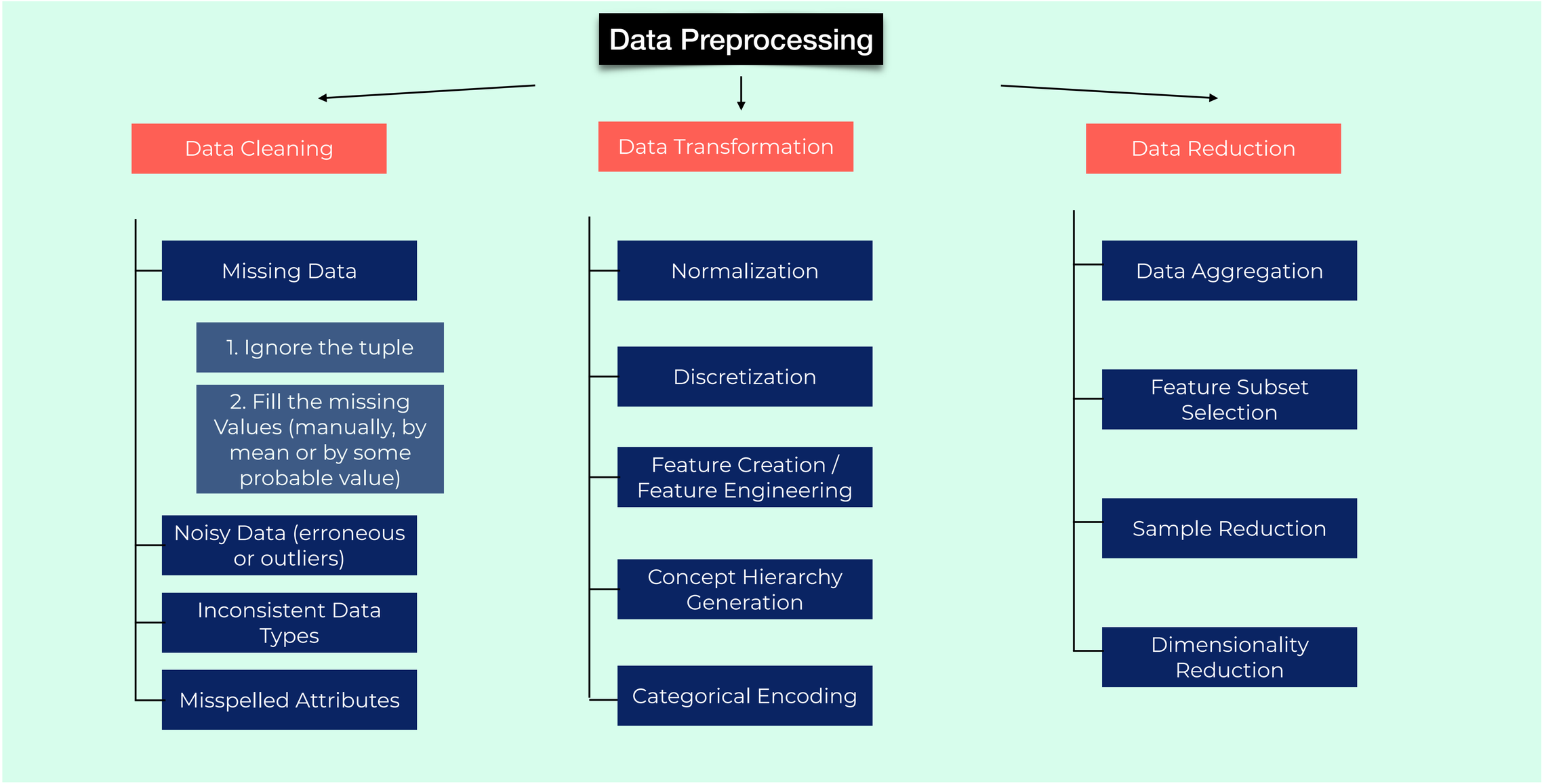


Review: Some statistic concepts

- Normal (or gaussian) distribution ($\mathcal{N}(\mu, \sigma^2)$)



Prepare the data for Machine Learning algorithms

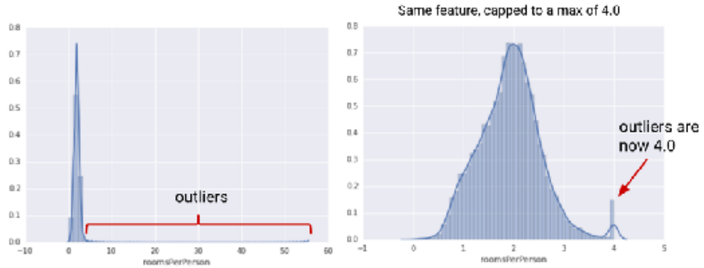

Data Preprocessing
Data Cleaning
Data Transformation
Data Reduction
Missing Data
1. Ignore the tuple
2. Fill the missing Values (manually, by mean or by some probable value)
Noisy Data (erroneous or outliers)
Inconsistent Data Types
Misspelled Attributes
Normalization
Discretization
Feature Creation / Feature Engineering
Concept Hierarchy Generation
Categorical Encoding
Data Aggregation
Feature Subset Selection
Sample Reduction
Dimensionality Reduction

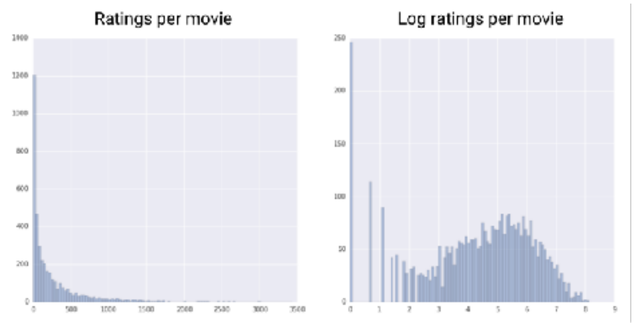
Data Transformation

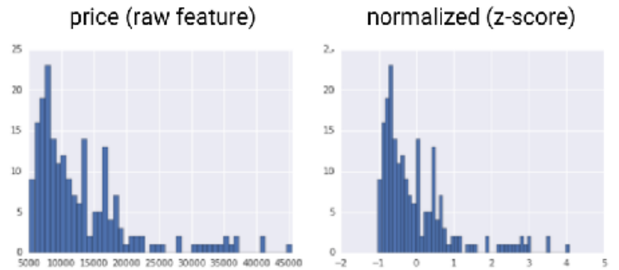
- In order to handle noise in the data we can transform it by **normalization** or **discretization**.
  - Normalization/Discretization can be different from data type to data type.

Normalization

- **Goal**: The goal of normalization is to transform features to be on a similar scale. This improves the performance and training stability of the model. Common techniques:
  - **Scaling to a range**. *Min-Max* normalization (for example into [0,1] or [-1,+1] range):
    - $x' = \frac{x - min}{max - min}$
  - **Clipping**. If your data set contains extreme outliers, you might try feature clipping, which caps all feature values above (or below) a certain value to fixed value. For example, you could clip all temperature values above 40 to be exactly 40.
  - **Log scaling**. Log scaling computes the log of your values to compress a wide range to a narrow range. It is used when the feature conforms to the power law. Log scaling changes the feature distribution, helping to improve linear model performance.
  - **Z-score normalization** (also know as standardization). It's useful when there are a few outliers, but not so extreme that you need clipping.
    - $x' = \frac{x - mean}{std}$

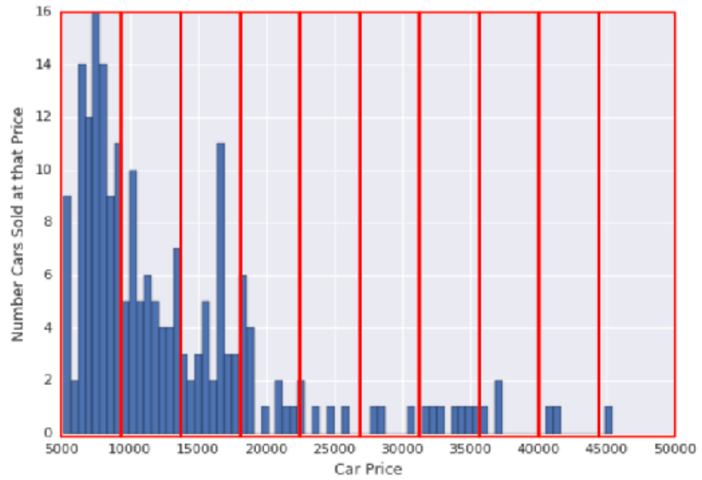


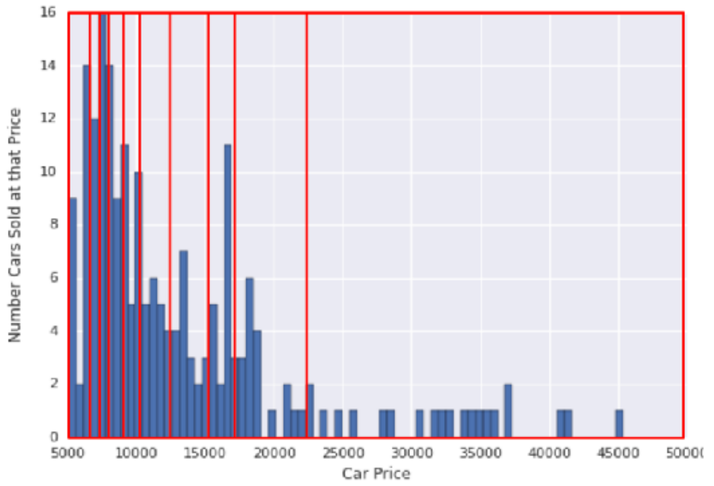




Data Discretization

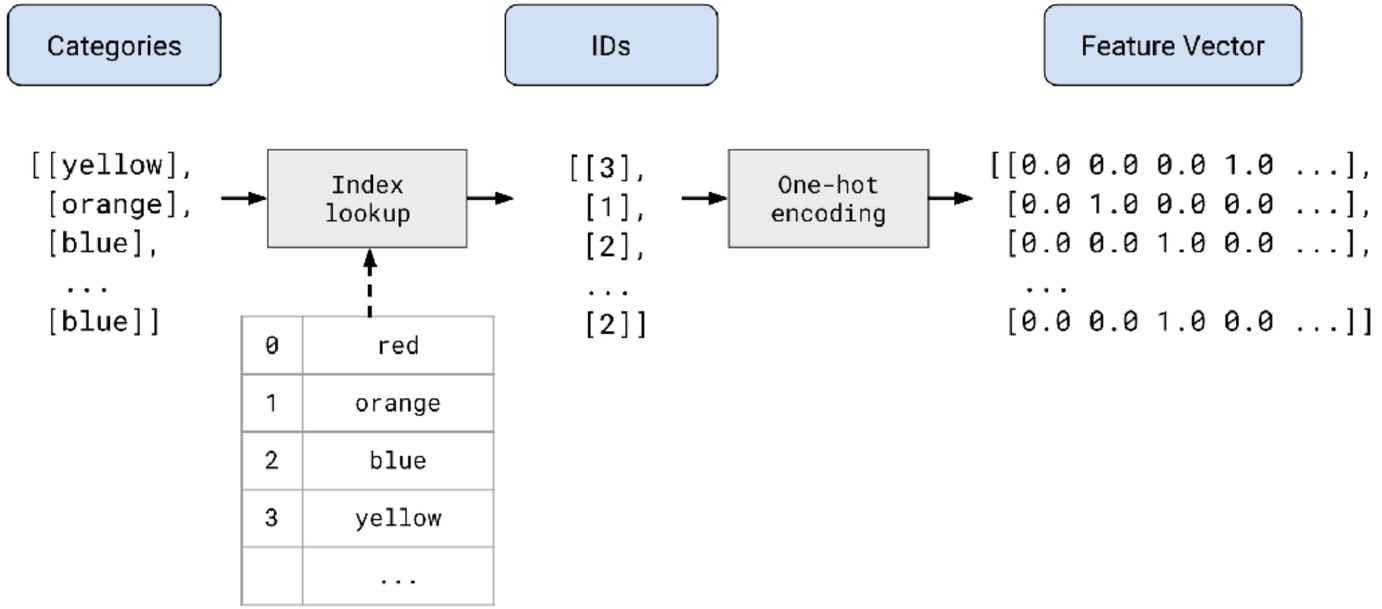
- Bucketing: Transforming numeric (usually continuous) data to categorical data.
  - **Buckets with equally spaced boundaries:** Some buckets could contain many points, while others could have few or none.
  - **Buckets with quantile boundaries**: each bucket has the same number of points. The boundaries are not fixed and could encompass a narrow or wide span of values.





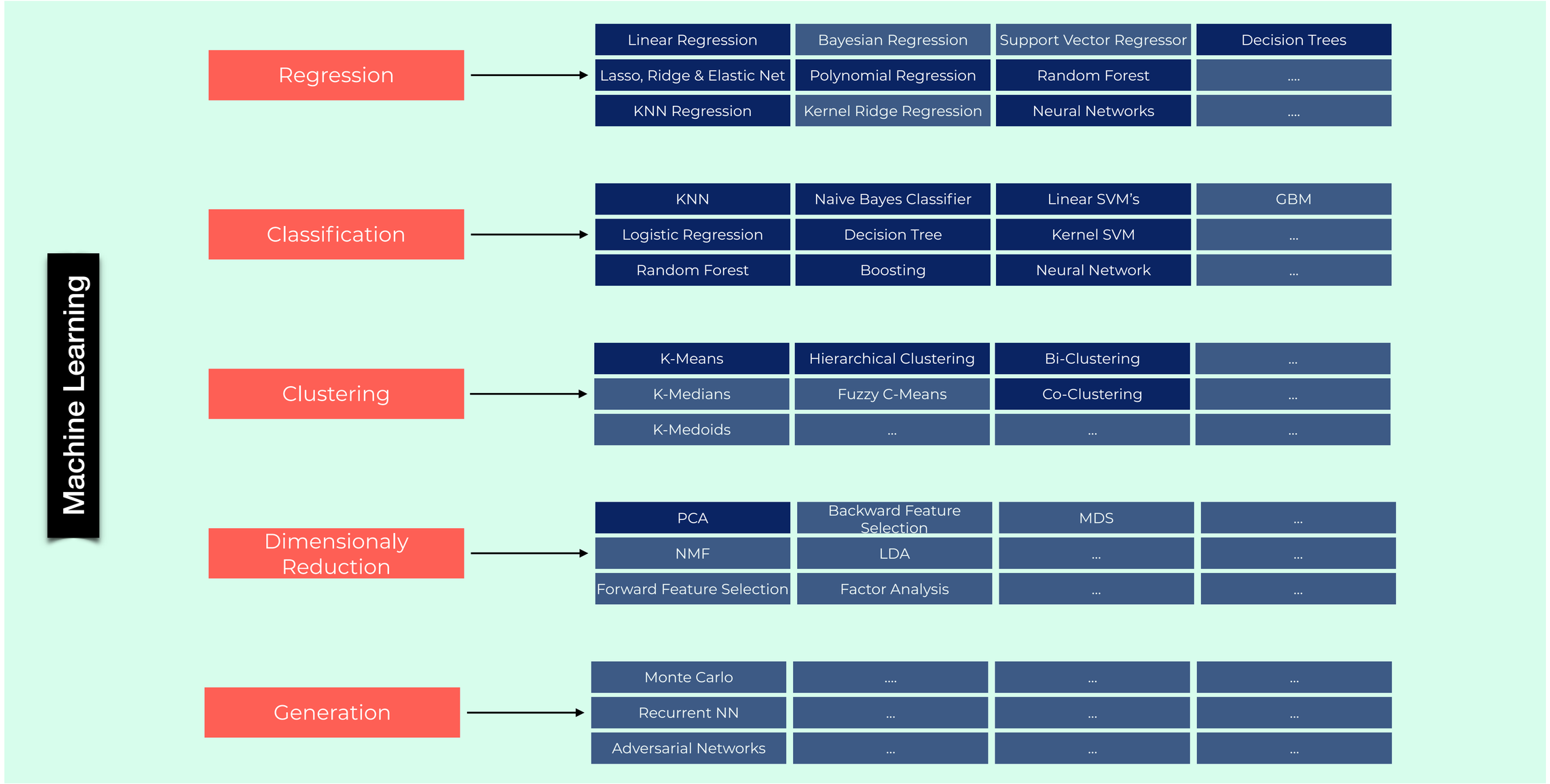
Categorical Encoding

- Ordinal Encoding
- One-Hot-Encoding
- Embeddings
  - *We will see this later in the course*

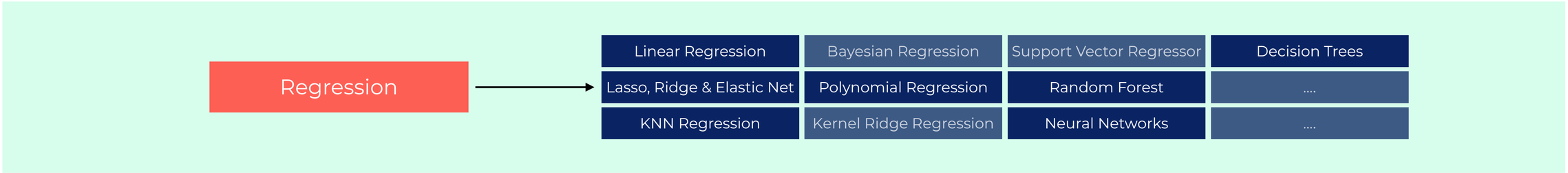


Select a model and train it

Classification	**To which data category does this data point belongs?**	**Medical diagnosis**: Does this tissue show signs of diseases? **Banking**: Is this transaction fraudulent? **Computer Vision**: What type of object is in this picture? Is it a person? Is it a building?
Regression	**Given this input from a dataset, what is the likely value of a particular quantity?**	**Finance:** What is the value of this stock going to be tomorrow? **Housing:** What would be the price of this house if it was sold today? **Food Quality**: When should I pick this strawberry?
Clustering	**Which data point are similar to each other?**	**E-Commerce:** which customers are exhibiting similar behavior to each other, how do they group together? **Video Streaming**: what are the different types of video genres in our catalogue, and which ones are in the same genre?
Dimensionality reduction	**What are the most significant features of this data and how can be summarized?**	**E-Commerce:** What combinations of features allows us to summarize the behavior of our customers? **Molecular biology:** How can scientists summarize the behavior of all 20.000 human gens in a particular diseases?
Semi-supervised learning	**How can be labelled and unlabelled data be combined?**	**Computer Vision:** How can an object detection be developed, with only a small training data set? **Drug Discovery:** Which of the million possible drugs could be effective agains disease, we have so far only tested a few?
Reinforcement learning	**What actions will most effectively achieve a desired endpoint?**	**Robots:** How can a robot move through its environment? **Games:** Which moves were more important in helping the computer win a particular game?

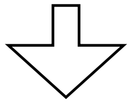

Machine Learning
Regression
Linear Regression
Bayesian Regression
Support Vector Regressor
Decision Trees
Lasso, Ridge & Elastic Net
Polynomial Regression
Random Forest
....
KNN Regression
Kernel Ridge Regression
Neural Networks
....
Classification
KNN
Naive Bayes Classifier
Linear SVM's
GBM
Logistic Regression
Decision Tree
Kernel SVM
...
Random Forest
Boosting
Neural Network
...
Clustering
K-Means
Hierarchical Clustering
Bi-Clustering
...
K-Medians
Fuzzy C-Means
Co-Clustering
...
K-Medoids
...
...
...
Dimensionaly Reduction
PCA
Backward Feature Selection
MDS
...
NMF
LDA
...
...
Forward Feature Selection
Factor Analysis
...
...
Generation
Monte Carlo
....
...
...
Recurrent NN
...
...
...
Adversarial Networks
...
...
...

Regression

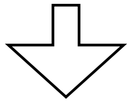

Regression
Linear Regression
Bayesian Regression
Support Vector Regressor
Decision Trees
Lasso, Ridge & Elastic Net
Polynomial Regression
Random Forest
....
KNN Regression
Kernel Ridge Regression
Neural Networks
....

No Free Lunch Theorem

Our model is a simplification of reality



Simplification is based on assumptions (bias)

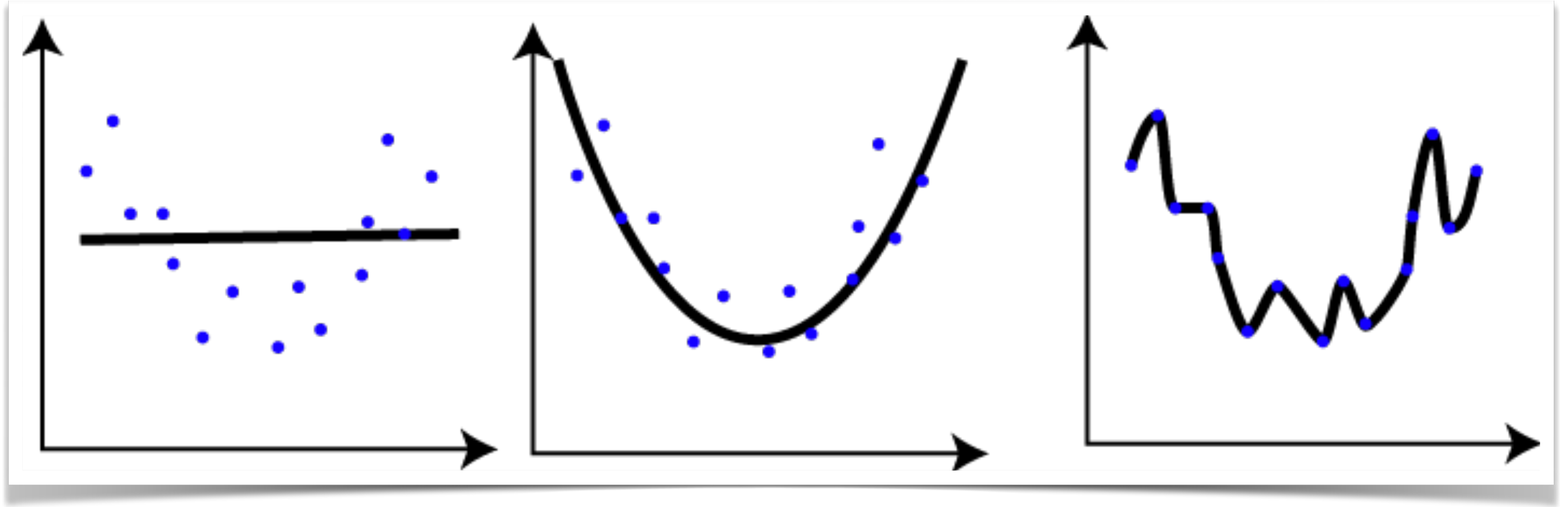


Assumptions fail in certain situations

Roughly speaking: "**There is not a model that works best for all possible situations**."

Generalization

- Undeffitting and Overffiting



Overfitting

Error vs. Sample Size

Error vs. feature set size

Fine-tune your model

Fine-tune your model

- Models can have several parameters.
- Fine-tune your model consists of **finding** the **best parameter** setup for your(s) model and evaluate which is the best.
- The chosed one should be the one to put in production.

How do you know which is the best model?

Train Data

Test Data

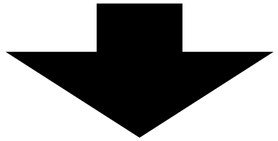
Used to **create/train** your model

Used to **evalute** your model perfomance

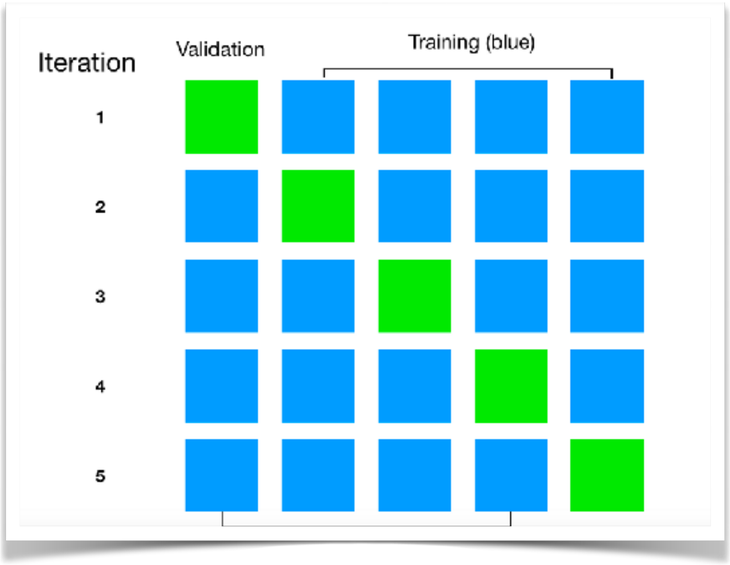
and hide it!!!!

Don’t forget!
You can not use it!

Train Data



OR



train - validation split

cross-validation

Present your solution

Present your solution

- At this stage you should get your **choosed model (only that one)** and **evaluate on the test data**.

- The results on the test data should be similar to the ones obtained using the validation data. If does not look similar it is possible that you have in your pipeline/methodology.

- The results should be presented to your boss/client

Launch, monitor
and mantain your system