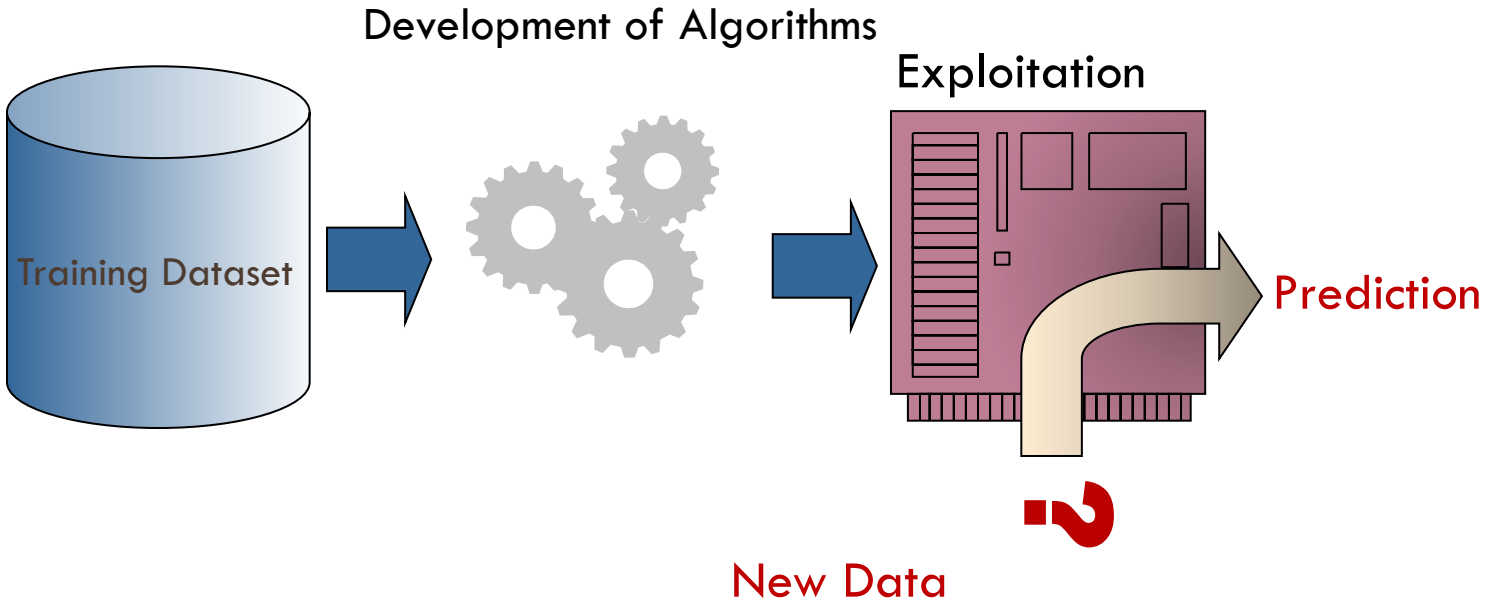
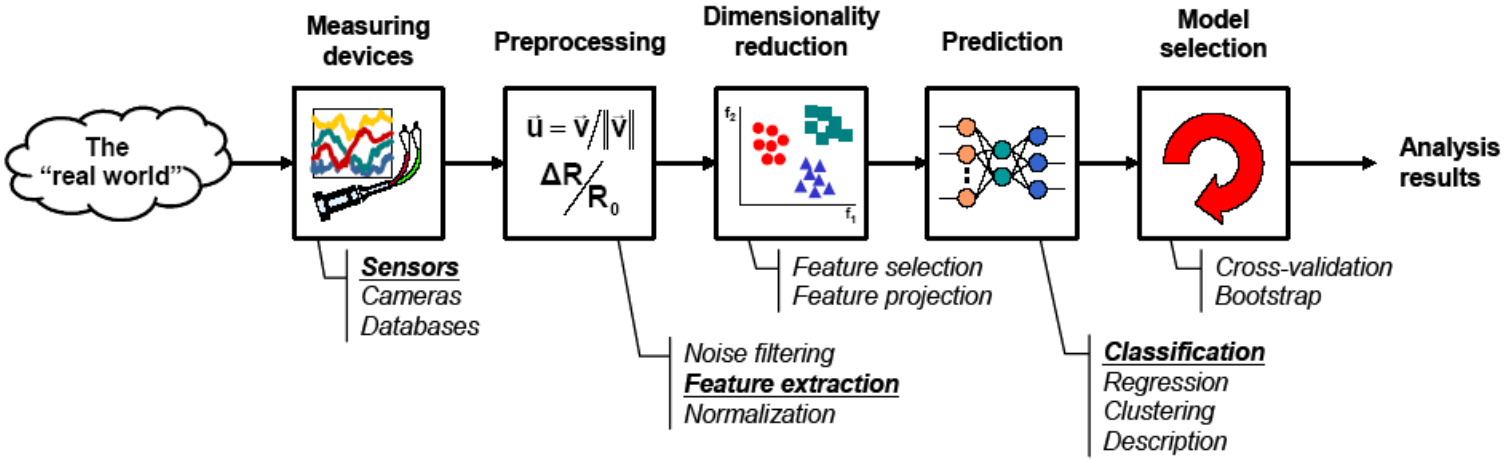
Preprocesamiento de Datos

Design Cycle



Design Cycle

- A pattern classification system contains:
  - Acquisition Sensor ➔ Raw database
  - Preprocessing Mechanisms
  - Mechanism of dimensionality reduction
  - Learning Algorithms
  - Mechanisms for validation



Design Cycle

Introduction

- Objective: To construct / use algorithms that learn from data
- Data is cheap and abundant:
  - data warehouses
  - data marts
- Knowledge is expensive and scarce.

Volumen

Velocity

Variety

Terminology: Attributes and patterns

Attribute (or variable or characteristic or descriptor), it is any distinctive aspect, quality or characteristic.

- nominal (i.e, color: white, red, yellow, green, blue, …),
- numeric (i.e., height, measured in meters).

Patterns (or Cases or Instances)

Collection (possibly ordered and structured) of descriptors (features) that represent an object.

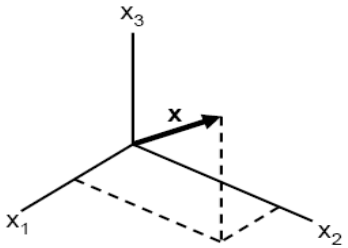
Important: patterns that describe objects of the same class have similar characteristics.

Terminology: Attributes and patterns

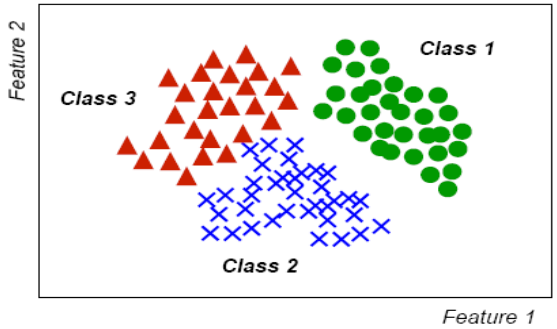
- Each pattern is represented by a set of attributes ➔ a column vector of d dimensions called attribute vector
- d-dimensional space defined by this vector is the attribute space
- The patterns are represented as points in attribute space

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \\ x_d \end{bmatrix}$$

Vector of attribute



Attribute Space

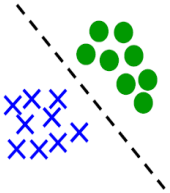


Terminology: Attributes and patterns

What is a "good" vector of attributes?

The quality of a vector of attributes is related to its ability to discriminate examples from different classes:

- The attributes of instances of the same class should have similar values
- The attributes of instances from different classes should have different values

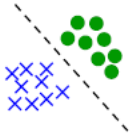


"Good" features

"Bad" features

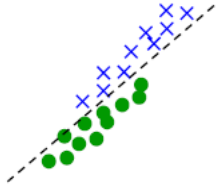
Terminology: Attributes and patterns

Additional properties related to attributes :

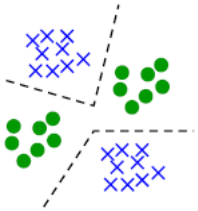


Lineal Separability

Non lineal Separability



Highly
Correlated attributes



Multi-modal

Data Preprocessing

- How can data be prepared for the next text mining /machine learning process?
  - **Data cleaning**, is the process oriented to eliminate data with noise or incorrect.
  - **Data integration**, tries to integrate different data sources in a coherent and homogeneous warehouse such as a data warehouse or a data cube.
  - **Data transformation**, or transformation of the data as, for example, a normalization.
  - **Data reduction**, it is aimed at reducing size of the data by aggregation and / or elimination of redundant features.

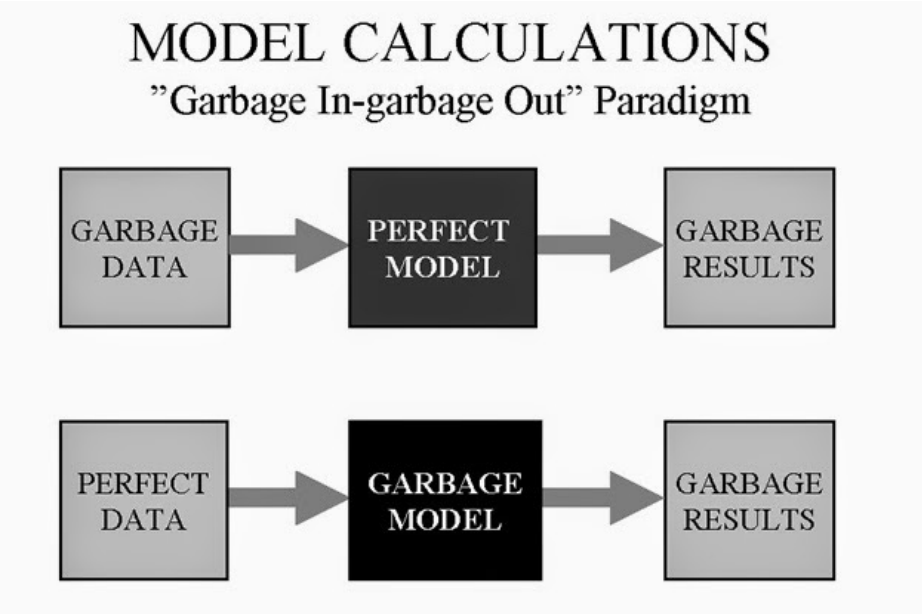
Data cleaning

- Real-world data is often presented in an incomplete, inconsistent and noisy way. We will learn to clean up the data by:
  - eliminating missing data,
  - softening the effect of noise,
  - eliminating data out of range and
  - correcting inconsistencies.
- The representation and quality of data is first and foremost before running any analysis.
- Data preparation and filtering steps can be considerable time-consuming.

Data Preprocessing

- Knowledge Discovery during the training phase is more difficult when:
  - there is much irrelevant and redundant information
  - noisy
  - unreliable data
- Data gathering methods are often loosely controlled
  - Resulting in out-of-range values (e.g., Income: -100),
  - Impossible data combinations (e.g., Gender: Male, Pregnant: Yes),
  - Missing values
- The product of data pre-processing is the final dataset.

“Garbage In – Garbage Out”

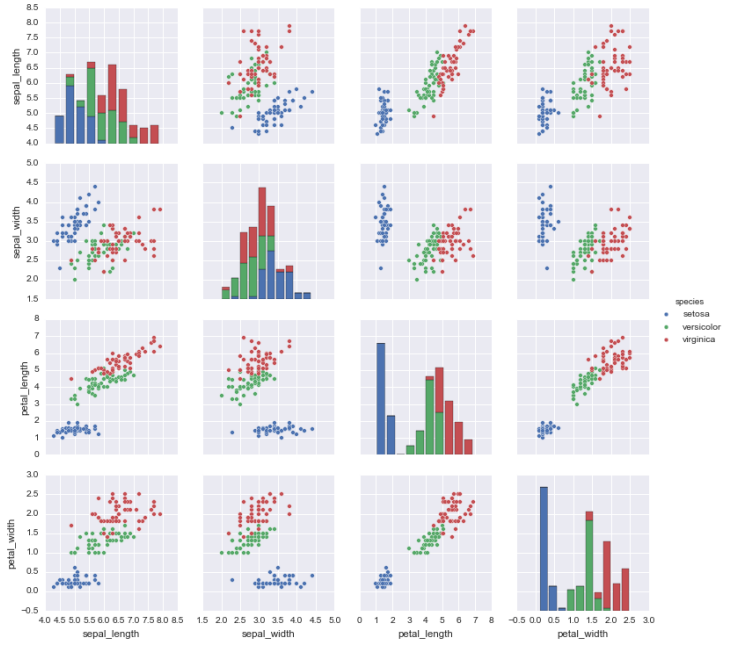


Garbage is an unwanted or undesired material or substance

Data Cleaning or Data Cleansing

Data auditing: The data is audited with the use of statistical and database methods to detect anomalies and contradictions.

- Incomplete
  - lacking attribute values
  - certain attributes of interest
  - containing only aggregate data
- Noisy
  - Containing errors
- Outliers, values which deviate from the expected
- Inconsistent



Data Cleaning – Incomplete Data

- Attributes of interest may not always be available
- Data may not be included simply because it was not considered important at the time of entry
- Relevant data may not be recorded due to a misunderstanding
- Incomplete Data due to equipment malfunctions

Data Cleaning – Incomplete Data (II)

- What to do with missing values?
  - Ignore the tuple (pattern, instance, case, …)
    - This is usually done when the class label is missing (Classification Problem)
    - In general, it´s not very effective, unless the tuple contains several attributes with missing values.
  - Fill in the missing value manually
  - Use a global constant to fill in the missing value
  - Use the attribute mean to fill in the missing value
  - Use the attribute mean for all samples belonging to the same class as the given tuple
  - Use the most probable value to fill in the missing value. Inference-based tools: regressión, bayesian formalism or decision tree induction, KNN Imputation

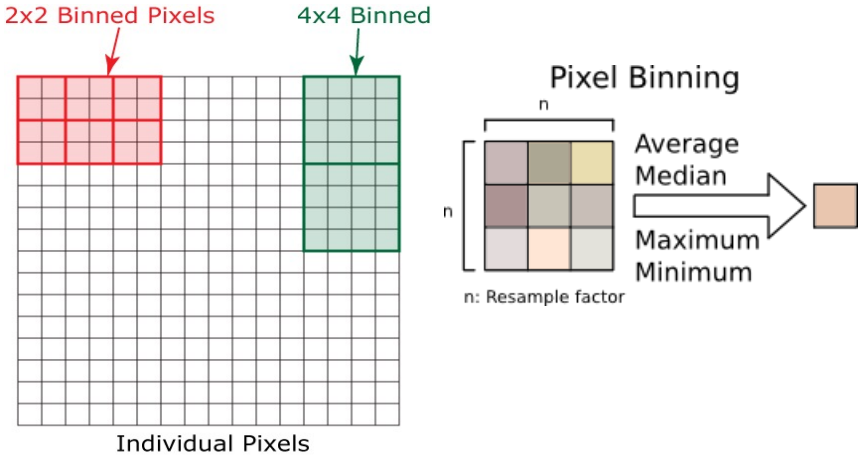
BIAS the data:
filled-in value may not be correct

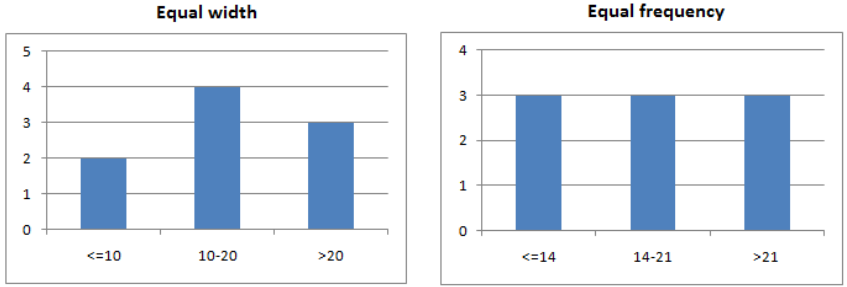
Data Cleaning – Noisy

- Noise is a random error or variance in a measured variable

$$\hat{X}=X+\varepsilon$$

- Smoothing noisy data:
  - Binning methods are a way to group a number of more or less continuous values into a smaller number of "bins"





Data={0, 4, 12, 16, 16, 18, 24, 26, 28}

Data Cleaning – Outlier Detection

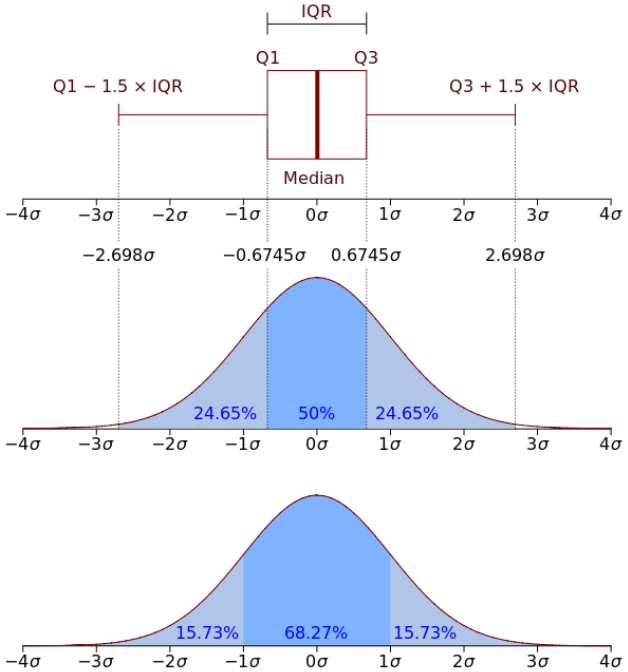
- Outliers
  - **CASE I**: the outlying value should be

deleted from the analysis (or corrected if possible)

An outlier may indicate bad data

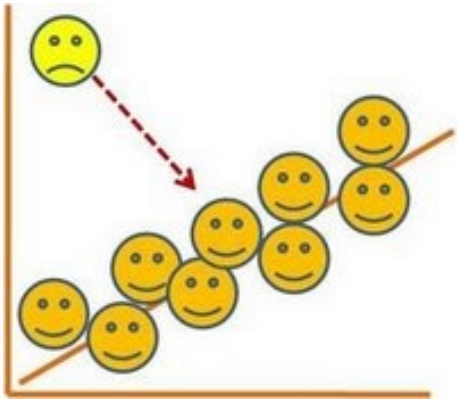
- the data may have been coded incorrectly
- an experiment may not have been run correctly

- **CASE II**:
  - outliers may be due to random variation
  - may indicate something scientifically interesting



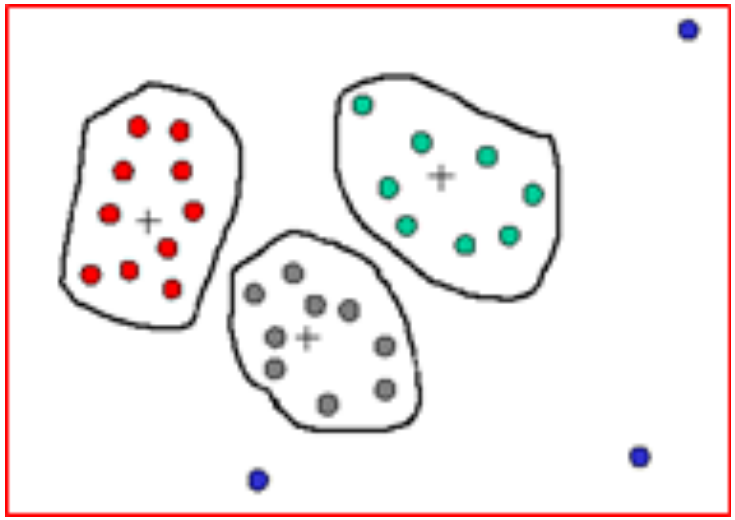
Data Cleaning – Outlier Detection

- Outliers should be investigated carefully.
- Often they contain valuable information about the process under investigation or the data gathering and recording process.
- Before considering the possible elimination of these points from the data, one should try to understand why they appeared and whether it is likely similar values will continue to appear.
- Of course, outliers **are often bad data points**.



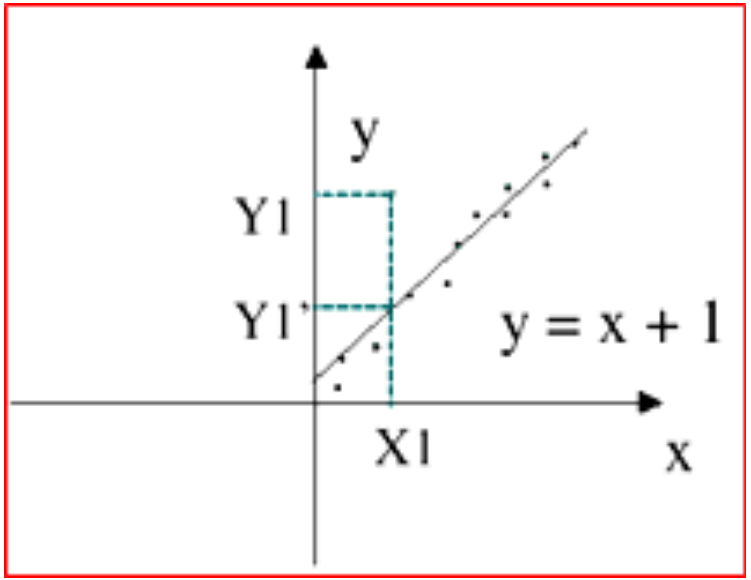
Data Cleaning – Outlier Detection (II)

- Clustering: Similar values are organized in groups or clusters. Values falling outside of the cluster maybe considered as outliers and may be candidate for elimination



Data Cleaning – Outlier Detection (III)

- Regression: Fit Data to a Function. The new values given by the function are used instead of the original values.

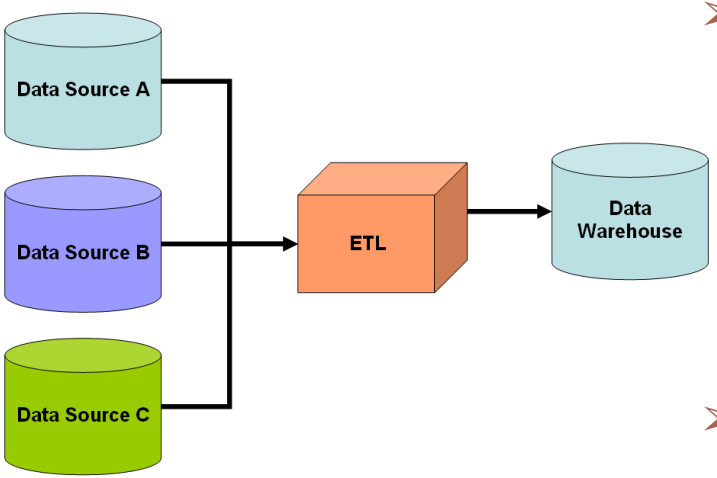


Data Cleaning – Inconsistencies

- Correct inconsistent data: use domain knowledge or expert decision.
- Inconsistent: containing discrepancies in codes or names, e.g.
  - Age = “42”, Date Birth = “03/07/2010”
  - Was rating “1, 2 or 3”, now rating “A, B or C”
- Inconsistent data are handled by:
  - Manual Corrections: tedious and expensive
  - Develop routines to detect insconsistencies

Data Integration

- Data integration involves combining data residing in different sources and providing users with a unified view of these data.



- Detecting and resolving data value conflicts:
  - Different scales, metric (feet or meters).
  - Entity identification problem, some attributes representing a given concept may have different names in different databases.

  Patients: “Bill”, “William”, “B.”

- Removing duplicates and redundant data

Data Transformation

- Normalization, where the attribute data are scaled so as to fall within a small specified range, such as [-1.0, 1.0], or [0, 1.0]
  - Min-Max normalization, linear transformation of the original data based on the minimum and maximum values of an attribute

$$x' = nuevo_{min_A} + \left(nuevo_{max_A} - nuevo_{min_A}\right) \frac{x - min_A}{max_A - min_A}$$

  - Z-Score normalization, adjusts the data from the initial distribution to a normal
    Also called "standarization"

$$x' = \frac{x - \mu_A}{\sigma_A}$$

Data Transformation

- Normalization by decimal scale, normalizes by moving the decimal points of the values of attribute A

$$x' = \frac{x}{10^j}$$

where j is the smallest integer such that $\max(|x'|)<1$

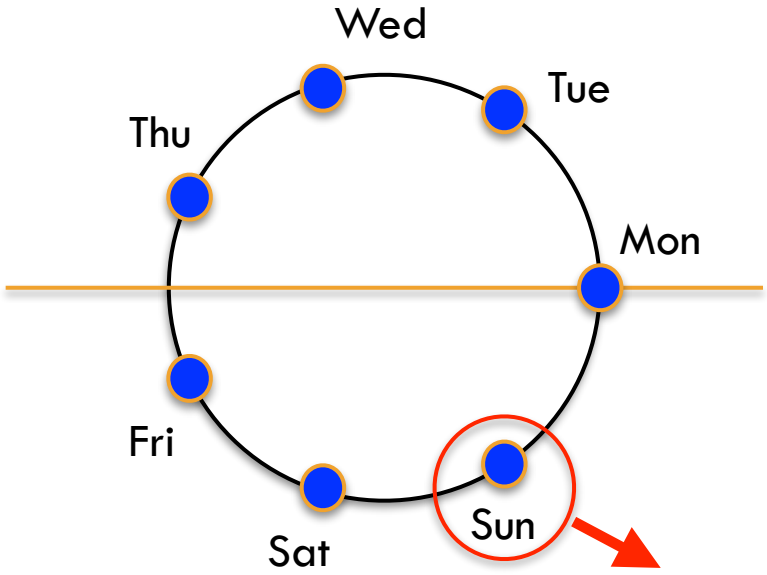
Data Transformation

- Dummy Coding: categorical predictor variables cannot be entered directly into a regression model and be meaningfully interpreted
- Dichotomous variables from categorical variables are created ➔ **dummy coding.**

	Dept	**FamilyS**	**Biology**
Family Studies	1	1	0
Biology	2	0	1
Business	3	0	0

Data Transformation

- Cyclic categories
  - Example: Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday
    - Option 1: **dummy coding** (7 dummy variables, one can be removed)
    - Option 2: cyclic coding:



N categories -> N angles

angle i = $i \cdot (360°/N)$

Mon: angle = 0°
Wed: angle = $2 \cdot (360°/7)$

2 variables x, y

x(i) = cos(angle i), y(i) = sin(angle i)

$x = \cos(6 \cdot 360° / 7), y = \sin(6 \cdot 360° / 7)$

Data Transformation

- Cyclic categories
  - Example: Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday
    - Option 1: **dummy coding** (7 dummy variables, one can be removed)
    - Option 2: cyclic coding:

Warning: usually numerical libraries such as bumpy work with angles measured in radians, not degrees

Thus we have to change degrees to radians:

angle in radians = angle in degrees · pi / 180

Data Transformation: Synthetic Variables

- New attributes, for what and how?
  - Add / replace attributes makes it easier for algorithms to analyze the dataset
  - If we can combine attributes through some interesting expression, we get that the algorithm does not have to discover that expression
    - We can create an attribute area from height and width
    - Knowledge from experts is required

NPI = [0.2 x S] + N + G, where

S =Size, N=#nodes (0 =1, 1-3 = 2, >3 = 3), G=Grade (1, 2 o 3)

False Predictor

- **Be careful not to include False Predictors. Find and eliminate them.**

A false predictor is a variable that is strongly correlated with the output class, but that is not available in a realistic prediction scenario.

- An expert in a modeling domain can spot when false predictor is buried among the input variables, because the model will be performing better than could be expected given the uncertain nature of the task.
- If the results are too good to be true, you probably have found false predictors.