A Computational Intelligence Perspective of Modelling and Data Analysis (Knowledge Discovery)

Varun Kumar Ojha

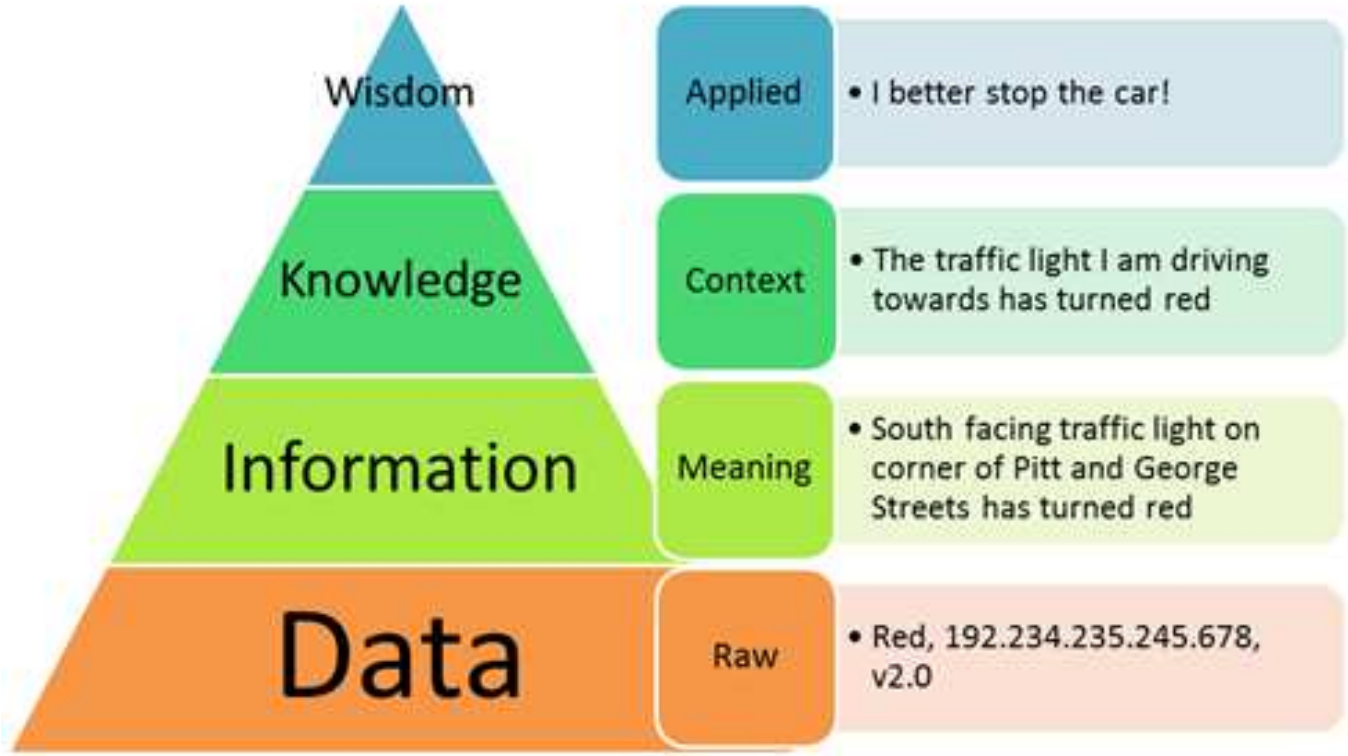
IT4Innovations, VŠB Technical University of Ostrava,

The Czech Republic.

Outline

1. Importance of Data Processing and Knowledge Discovery
2. Identifying Application
3. Generating Good Quality Data (Data Quality Assurance)
4. Cleaning and Preprocessing Data
5. Identifying Computational Methods
6. Discovering Meaning (Interpreting Discovered Knowledge)

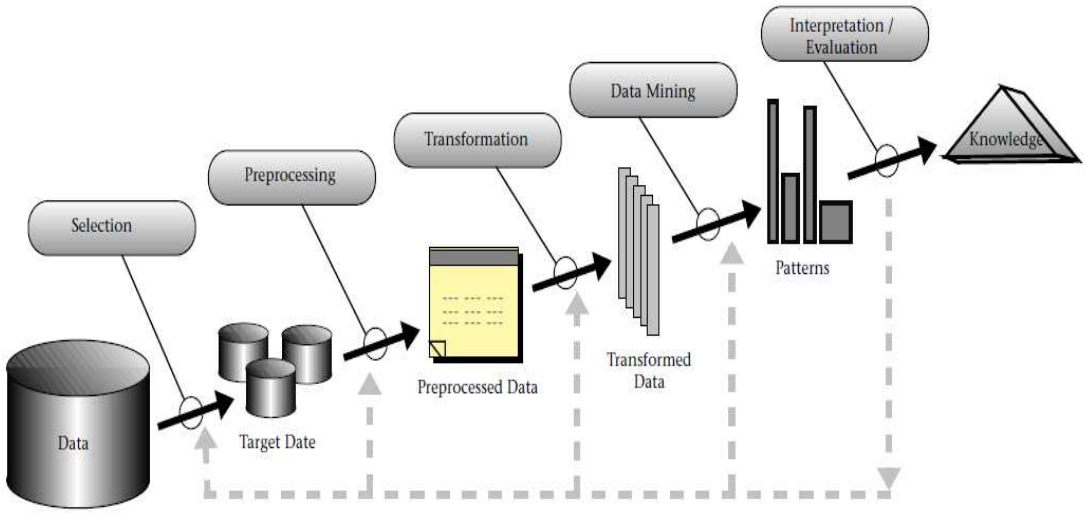
Knowledge Discovery



Source: http://www.allthingy.com/data-information-knowledge-wisdom/

Data analysis (Knowledge Discovery) Steps

1. Identifying real world application and goals
2. Generating Data
3. Cleaning Data
4. Data Reduction
5. Identifying/matching Goal (Step 1) (CI method to use)
6. Selecting Algorithms
7. Searching data pattern of interest
8. Interpreting discovered patterns
9. Using or comparing discovered meaning (interpreted knowledge) in real world. (Step 1)



Source: Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, From Data Mining to Knowledge Discovery in Databases, AI Magazine Volume 17 Number 3 (1996)

Data

- Be able to define variables.
  - Continuous variables:
    - Always numeric
    - Can be any number, positive or negative
  - Categorical variables:
    - Information that can be sorted into categories
    - Types of categorical variables – ordinal, nominal and dichotomous (binary)
    - Ordinal variable—a categorical variable with some intrinsic order or numeric value
      - Rating (excellent, good, fair, poor) or Score (0 – 5)
    - Nominal variable – a categorical variable without an intrinsic order e.g. : Sex (male, female)
    - Dichotomous (or binary) variables
    - Yes or No (0 /1)
- Be able to identify independent variables and dependent variables

What is “Dirty Data”?

The data which is

inaccurate, imprecise, unreliable, incomplete, inconsistent, irrelevant, invalid, ambiguous, redundant, and forged.

Good Quality Data Should be

#	Measure	Description
1	Accuracy	The degree of correspondence between data and the real world, i.e., Each piece of data stored is related to a real world datum in a precise way. (Fundamentally controlled by the quality of the input)
2	Precision (or Numerical Precision)	The data have sufficient detail, in this case the “accuracy” of the data refers to the fineness of measurement units. (Data is stored with the precision required to characterize it.)
3	Reliability	The data are measured and collected consistently, definitions and methodologies are the same over time. The data stored is trustable, i.e., it can be taken as true information.
4	Completeness	Every fact of the real world is represented. It is possible to consider two different aspects of completeness: first, certain values may not be present at the time; second, certain attributes cannot be stored.
5	Consistency	The data agrees with itself (There is no contradiction between the data stored).

Good Quality Data Cont....

#	Measure	Description
6	Relevant	Every piece of information stored is important in order to get a representation of the real world. (The stored information is applicable for the application problem.)
7	Validity	Valid data are considered accurate: They measure what they are intended to measure.
8	Unambiguous and Uniqueness	Each piece of data has a unique meaning. The collected data should have unique record.
9	Integrity and Correctness	The data are protected from deliberate bias or manipulation for political or personal reasons. Every set of data stored represents a real world situation
10	Amount of Data or Conscience	The number of facts stored. (Real world is represented with the minimum information required.)

Source: Mónica Bobrowski, Martina Marré, Daniel Yankelevich, Measuring Data Quality, Report no.: 99-002, Pabellón 1 - Planta Baja - Ciudad Universitaria

Questions in Data Quality Measurement

- Where data quality problem occurs
  - Data and information is not static, it flows in a data collection and usage process.
  - Data quality problem occurs during: Data gathering, Data delivery, Data storage, Data integration, Data retrieval, Data mining/analysis
- Which quality measure parameter to chose.
  - (Not all the dimensions are relevant in every situation)
- How to define the quality measure.
  - (Each dimension has several aspects that characterize it. Not every aspect is important in every situation.)

Outline

1. Importance of Data Processing and Knowledge Discovery
2. Identifying Application
3. Generating Good Quality Data (Data Quality Assurance)
4. Cleaning and Preprocessing Data
5. Identifying Computational Methods
6. Discovering Meaning (Interpreting and Discovering Knowledge)

Major Tasks in Data Preprocessing

- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration
  - Integration of multiple databases, data cubes, or files
- Data transformation
  - Normalization and aggregation
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results
- Data discretization
  - Part of data reduction but with particular importance, especially for numerical data

Source: http://www.mimuw.edu.pl/~son/datamining/DM/4-preprocess.pdf

Data Cleaning Tasks

- Data acquisition and metadata
- Fill in missing values
- Unified date format
- Converting nominal to numeric
- Identify outliers and smooth out noisy data
- Correct inconsistent data
- Convert data to a standard format (e.g. ARFF or csv)

Arranging Data (Appropriate Format)

Samples	Features (Variables)				
	Input features (Independent Variables)				Output feature (dependent variable)
	Input F1	Input F2	Input F3	Input F4	Output F1
Sample 1					
Sample 2					
Sample 3					
Sample 4					
Sample 5					
:					
Sample N					

Data Integration

- Redundant data occur often when integration of multiple databases
- The same attribute may have different names in different databases
- One attribute may be a “derived” attribute in another table, e.g., annual revenue
- Redundant data may be able to be detected by correlational analysis
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Data Transformation

- Smoothing: remove noise from data
- Aggregation: summarization, data cube construction
- Generalization: concept hierarchy climbing
- Normalization: scaled to fall within a small, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling
- Attribute/feature construction
- New attributes constructed from the given ones

Data reduction

- Feature Reduction (Feature Selection or Feature Extraction)
- Feature Selection
  - E.g. Correlation-based feature selection, Stochastic feature Selection.
- Feature Extraction
  - E.g. Principle Component Analysis. Feature Analysis

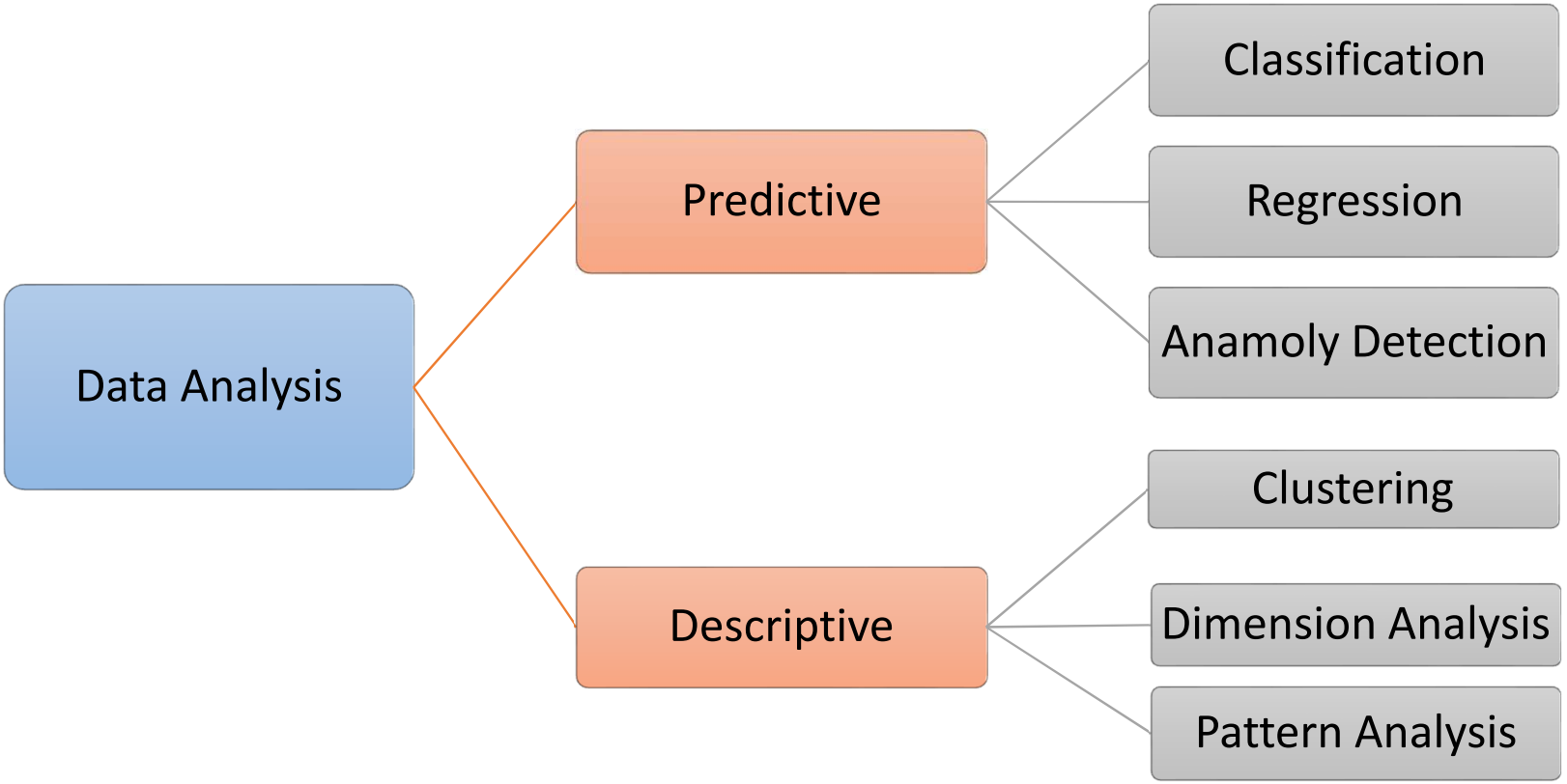
Data discretization

- Three types of attributes:
  - Nominal — values from an unordered set
  - Ordinal — values from an ordered set
  - Continuous — real numbers
- Discretization:
  - divide the range of a continuous attribute into intervals
  - Some classification algorithms only accept categorical attributes.
  - Reduce data size by discretization

Outline

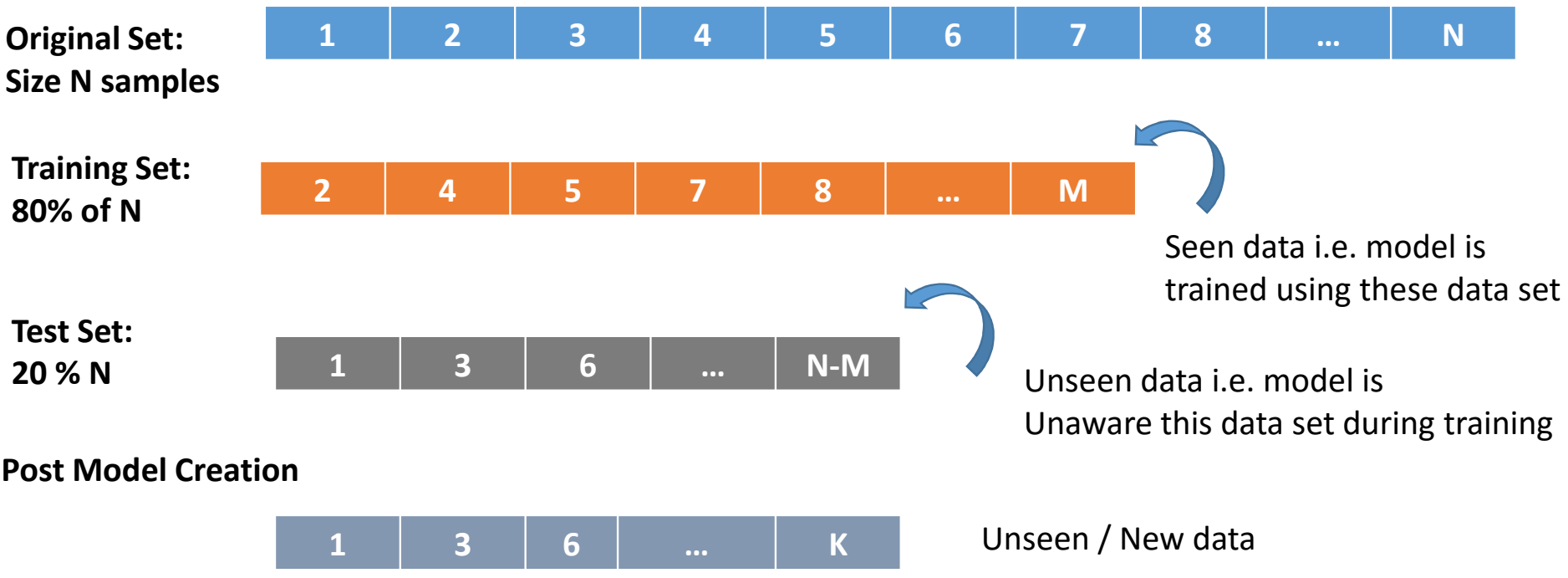
1. Importance of Data Processing
2. Identifying Application
3. Generating Good Quality Data (Data Quality Assurance)
4. Cleaning and Preprocessing Data
5. Identifying Computational Methods
6. Discovering Meaning (Interpreting and Discovering Knowledge)

Identifying Computational Methods and Goals



Training Set and Test Set

During Model Creation:



Post Model Creation

Computational Intelligent Methods

Example of Regression Problem

- **Problem** Daily Electricity Energy Consumption (Average price)
- **Input** Hydroelectric real [27881.8, 206035.0]
- **Input** Nuclear real [114760.0, 187105.0]
- **Input** Coal real [33537.0, 234833.0]
- **Input** Fuel real [0.0, 67986.5]
- **Input** Gas real [0.0, 84452.2]
- **Input** Special real [5307.0, 16357.0]

X

- **Output** Consume real **[0.765853, 5.11875]**

Y

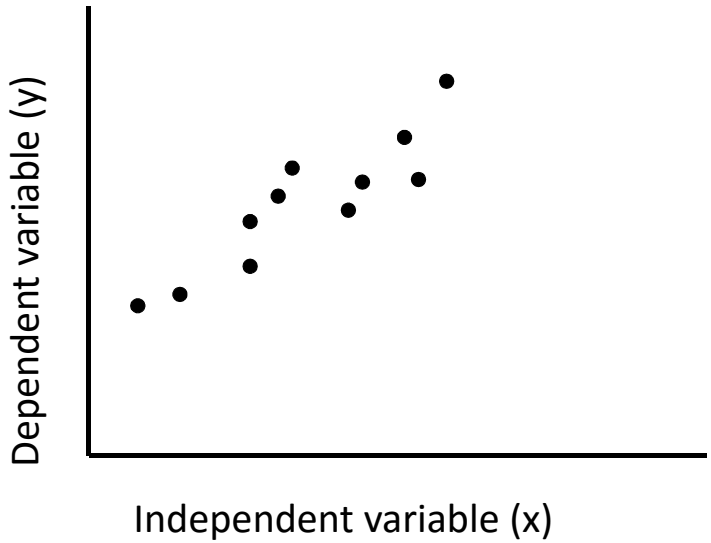
- **365 Examples**





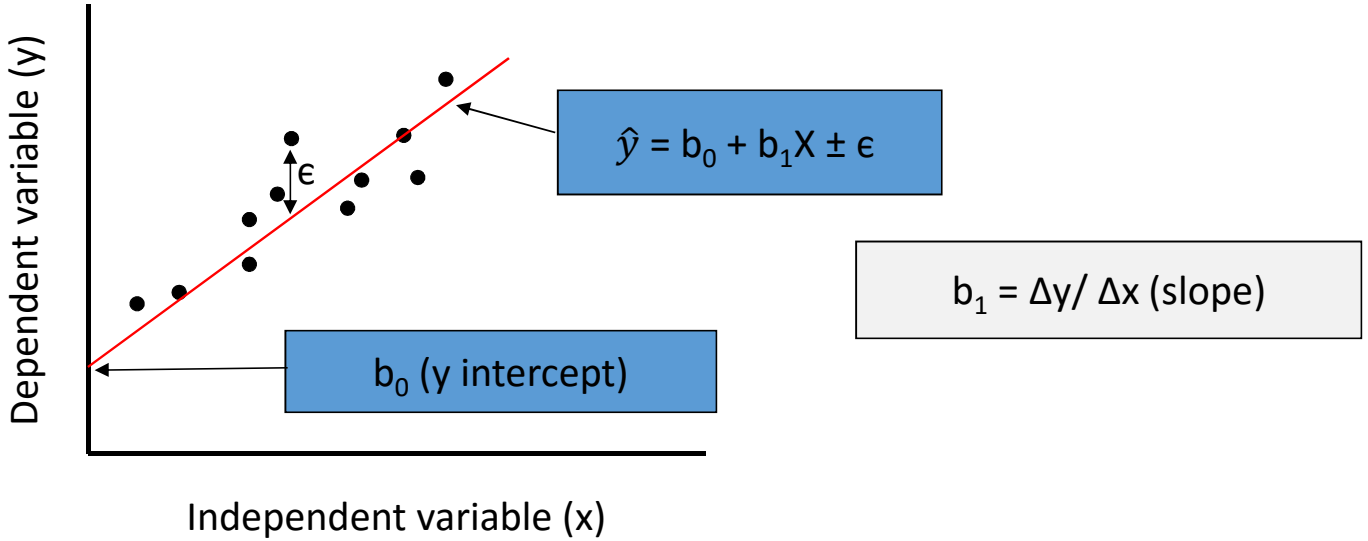

Outputs are continues numbers between 0.7658 to 5.11875

Source: http://archive.ics.uci.edu/ml/datasets.html



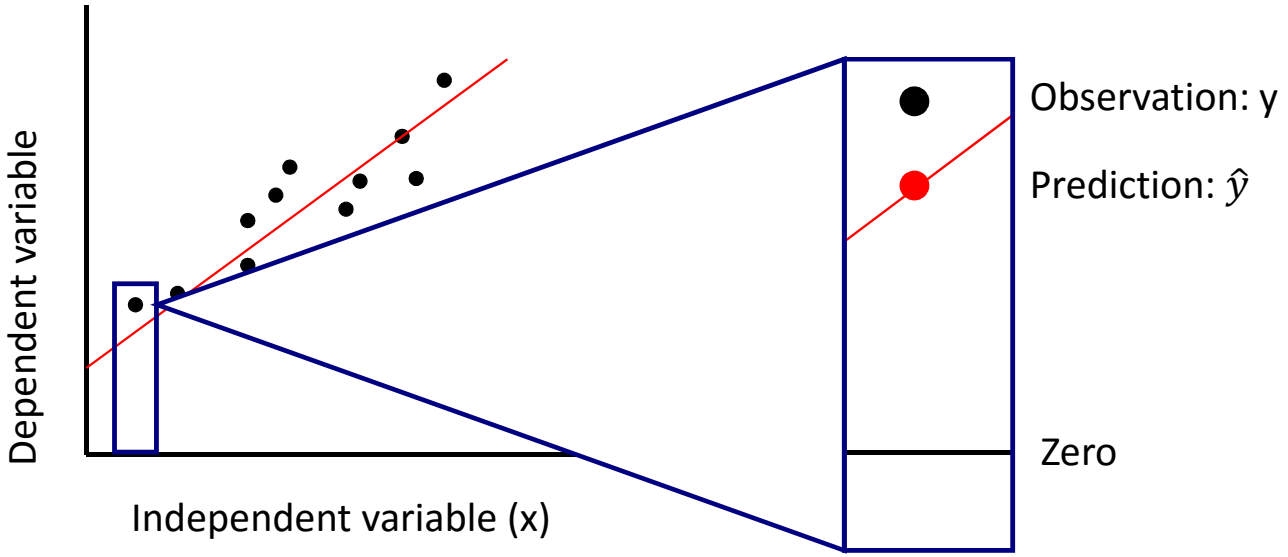
Regression is the attempt to explain the variation in a dependent variable using the variation in independent variables.

If the independent variable(x) sufficiently explain the variation in the dependent variable (y), the model can be used for prediction.



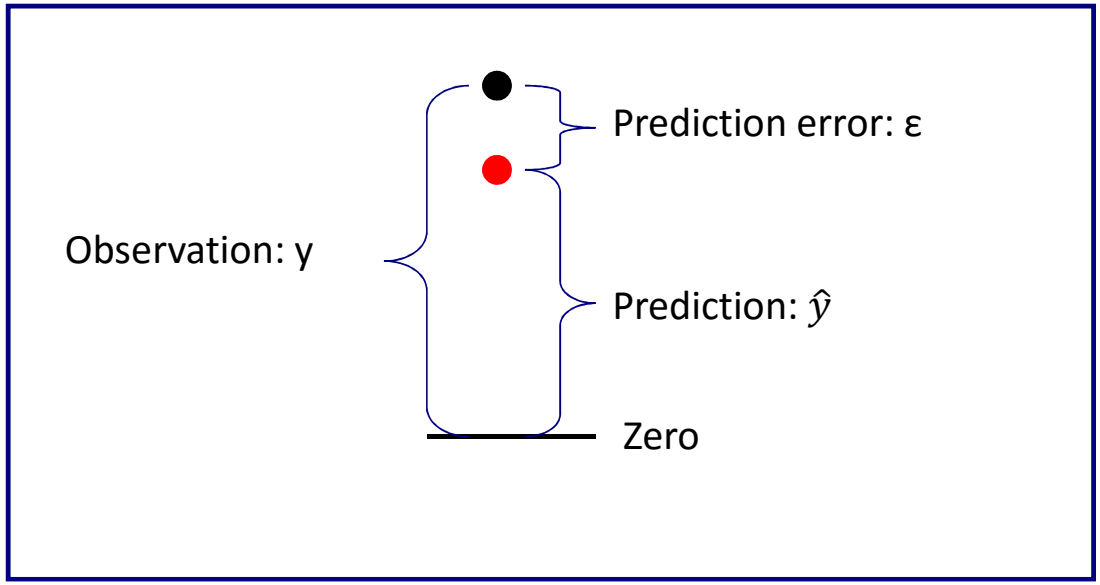
The output of a regression is a function that predicts the dependent variable based upon values of the independent variables.

Simple regression fits a straight line to the data.



The function will make a prediction for each observed data point.

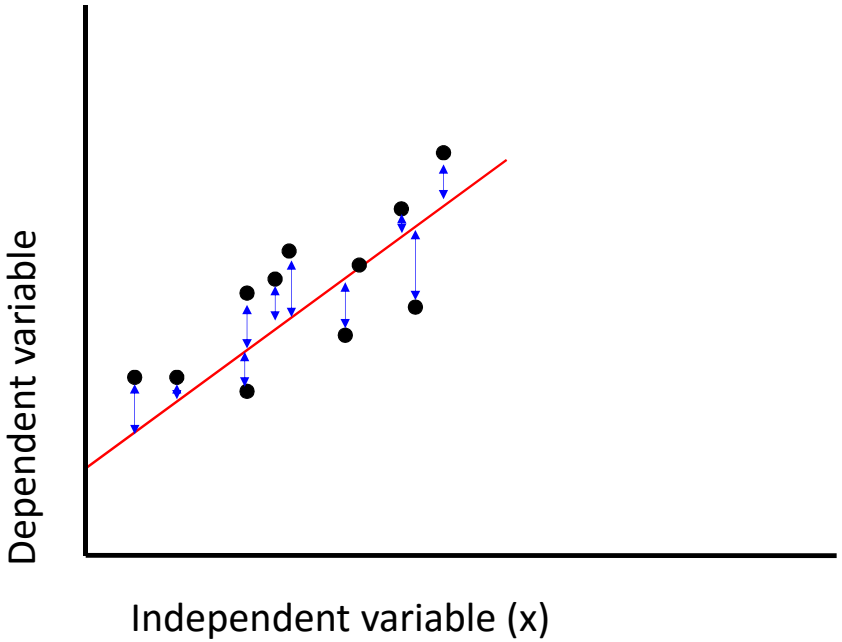
The observation is denoted by y and the prediction is denoted by $\hat{y}$.



For each observation, the variation can be described as:

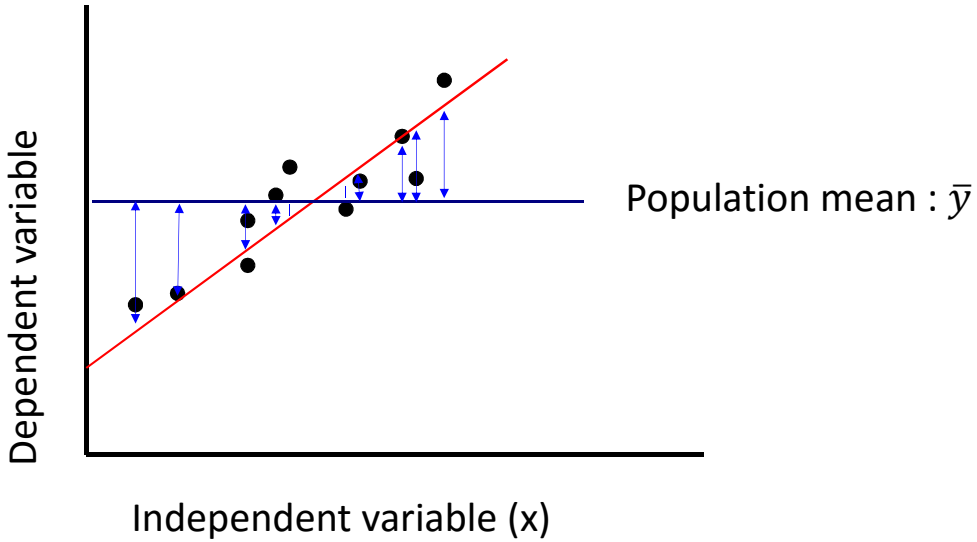
$$y = \hat{y} + \varepsilon$$

Actual = Explained + Error



A least squares regression selects the line with the lowest total sum of squared prediction errors.

This value is called the Sum of Squares of Error, or SSE.



The Sum of Squares Regression (SSR) is the sum of the squared differences between the prediction for each observation and the population mean.

Regression Goodness measure

The Total Sum of Squares (SST) is equal to SSR + SSE.

Mathematically,

SSR = $\sum (\hat{y} - \bar{y})^2$ (measure of explained variation)

SSE = $\sum (\hat{y} - y)^2$ (measure of unexplained variation)

SST = SSR + SSE = $\sum (\bar{y} - y)^2$ (measure of total variation in y)

Regression Goodness measure

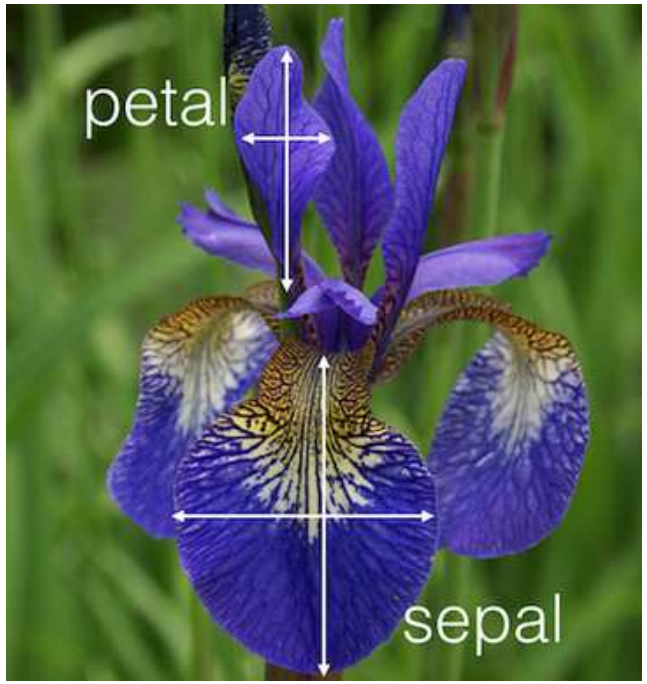
- The proportion of total variation (SST) that is explained by the regression (SSR) is known as the Coefficient of Determination, and is often referred to as $\boldsymbol{R^2}$.

$$R^2 = \frac{SSR}{SST} = \frac{SSR}{SSR + SSE}$$

- The value of $\boldsymbol{R^2}$ can range between 0 and 1, and the higher its value the more accurate the regression model is. It is often referred to as a percentage.

Example of Classification

- Problem: Iris flower identification
- Input Sepal Length real[4.3,7.9]
- Input Sepal Width real[2.0,4.4]
- Input Petal Length real[1.0,6.9]
- Input Petal Width real[0.1,2.5]
- Output {Iris-Setosa, Iris-Versicolor, Iris-Virginica}
- 150 examples

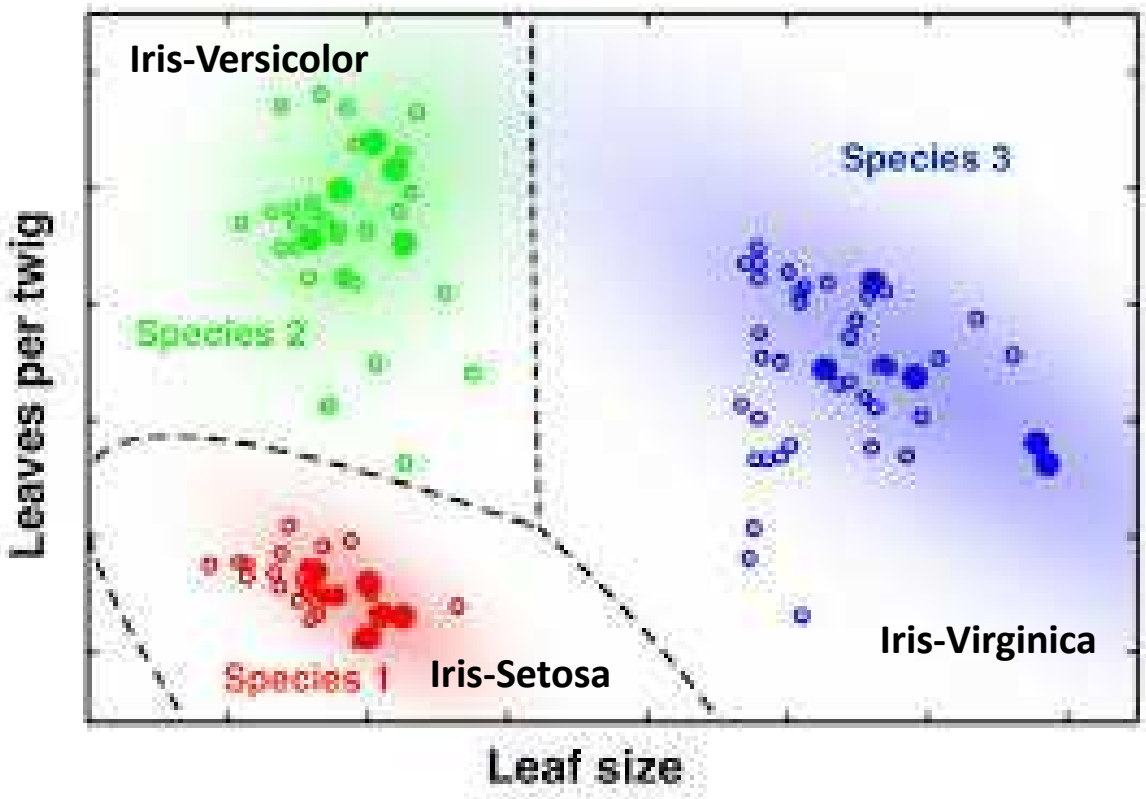




Outputs are discrete

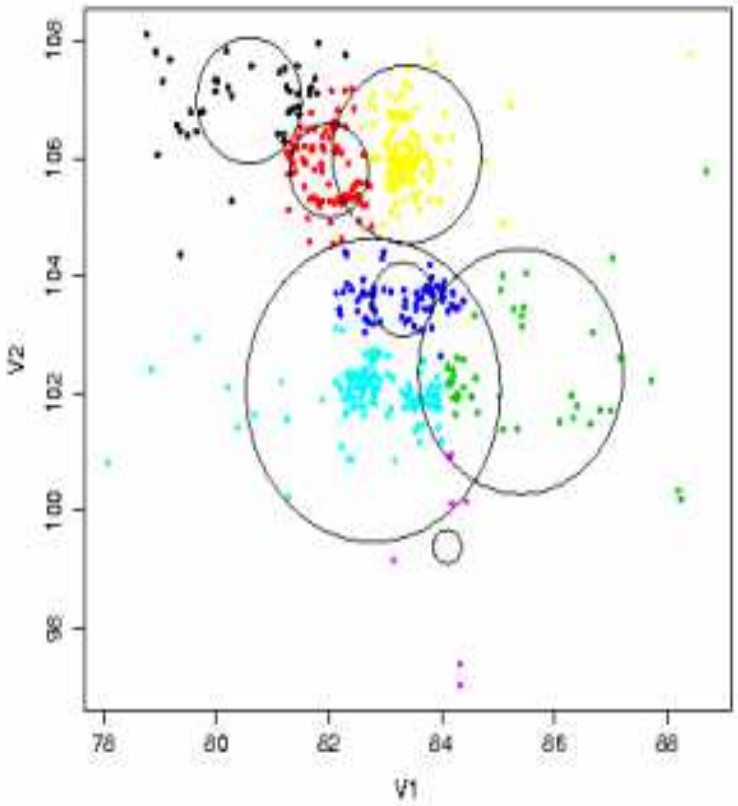
Source: http://archive.ics.uci.edu/ml/datasets.html

Classification



Clustering

- Cluster: a collection of data objects
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters
- Cluster analysis
  - Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters
- Unsupervised learning: no predefined classes.

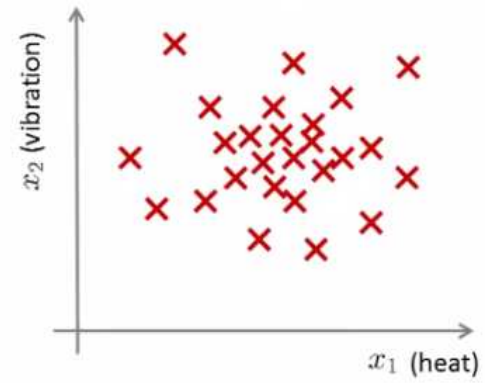


Example of Clustering

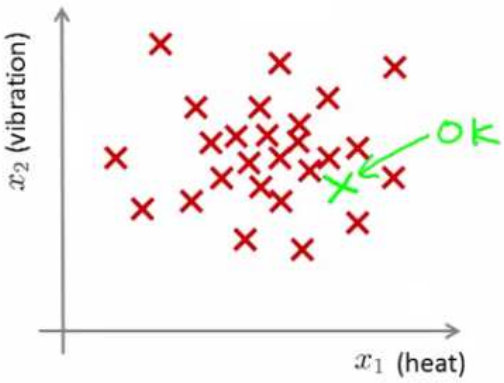
- Marketing: Finding groups of customers with similar behavior given a large database of customer data containing their properties and past buying records

- Biology: Classification of plants and animals given their features

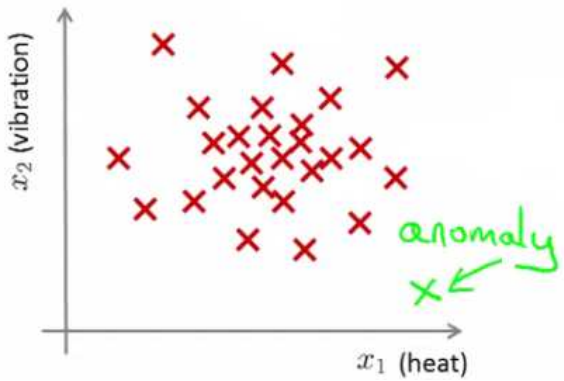
Anomaly Detection



Training Condition with mostly positive examples



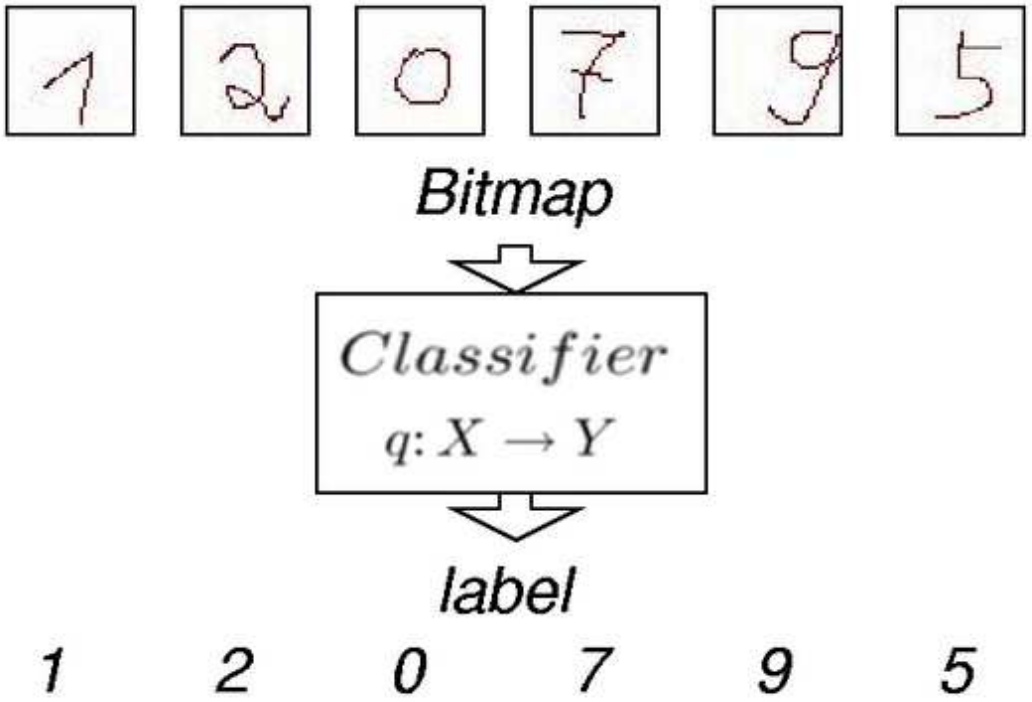
Test Condition 1



Test Condition 2

Source: http://www.holehouse.org/mlclass/15_Anomaly_Detection.html

Pattern Recognition Task



Outline

1. Importance of Data Processing
2. Identifying Application
3. Generating Good Quality Data (Data Quality Assurance)
4. Cleaning and Preprocessing Data
5. Identifying Computational Methods
6. Discovering Meaning (Interpreting and Discovering Knowledge)

Meaning to Real World

- What results says…
- Representing results in graphs and other visualization methods
- Extracting physical meaning related to real worlds problem intended.
- Comparing the results:
- What was the previous believes about the data/real world.
- What are the new knowledge obtained from the data analysis.

Thank You

Varun.kumar.ojha@vsb.cz

Q & A