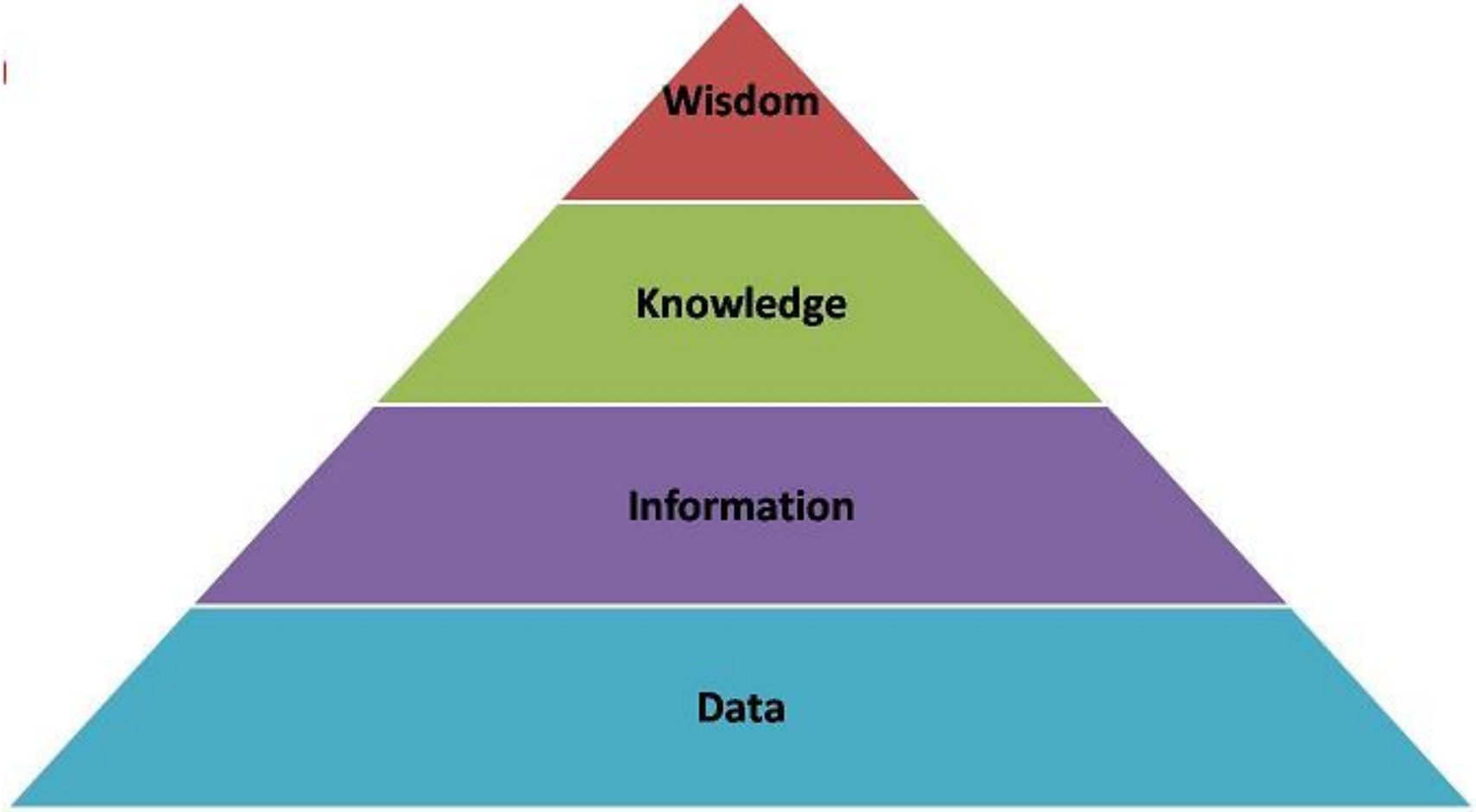
DATA-MINING

DBMS

Collecting Data
Storing Data
Managing Data

Data is collected and storing from various sources TB/hours
In today scenario Data is enormous it means we have large data but little Information .

Rich Data But Poor Information

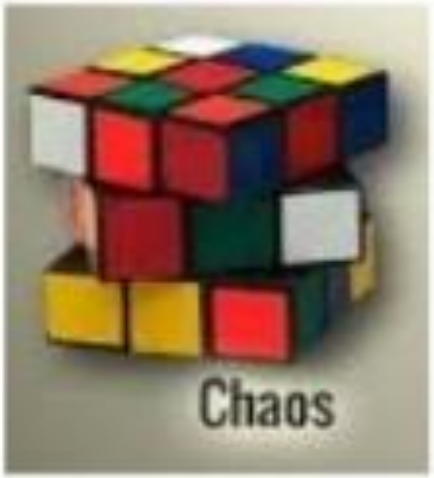


From Data to Information to Knowledge

Data Mining

There is a huge amount of data available in the Information Industry. This data is of no use until it is converted into useful information. It is necessary to analyze this huge amount of data and extract useful information from it.

Data Mining is defined as the procedure of extracting information from huge sets of data. In other words, we can say that data mining is mining knowledge from data.



DATA MINING ⟹

The Process of Discovering interesting and useful pattern and relationship is large volumes of data

Extraction of information is not the only process we need to perform; data mining also involves other processes such as Data Cleaning, Data Integration, Data Transformation, Data Mining, Pattern Evaluation and Data Presentation.

Data Mining Applications

Market Analysis
Fraud Detection
Customer Retention
Production Control
Science Exploration

Data mining is also known as Knowledge Discovery in Database (KDD).

Service providers

E-commerce

Crime agencies

KDD

- **Data Cleaning**
- **Data Integration**
- **Data Selection**
- **Data Transformation**
- **Data Mining**
- **Pattern Evaluation**
- **Knowledge Representation**

Data Cleaning

Data Cleaning is refer to remove noise , Missing Data. Generally Data Mining is used to Remove error from Data to improve Data Quality

Data Integration helps accuracy and speed

Data Transformation convert data in appropriate from for mining.

Classification In Data Mining

Predicting Certain Outcome Based on given Data

Classification is mathematical and symmetrical based on

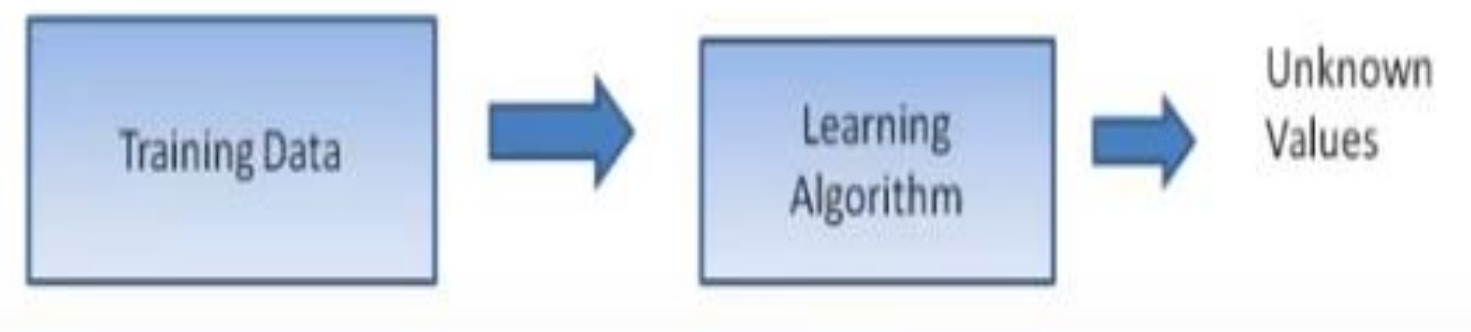
1. Decision Tree
2. Linear Programming
3. Neural Network
4. statistics

Classification Algorithms

- Decision Tree
- Rule Based Induction
- Neural Network
- Bayesian Network
- Genetic Algorithms

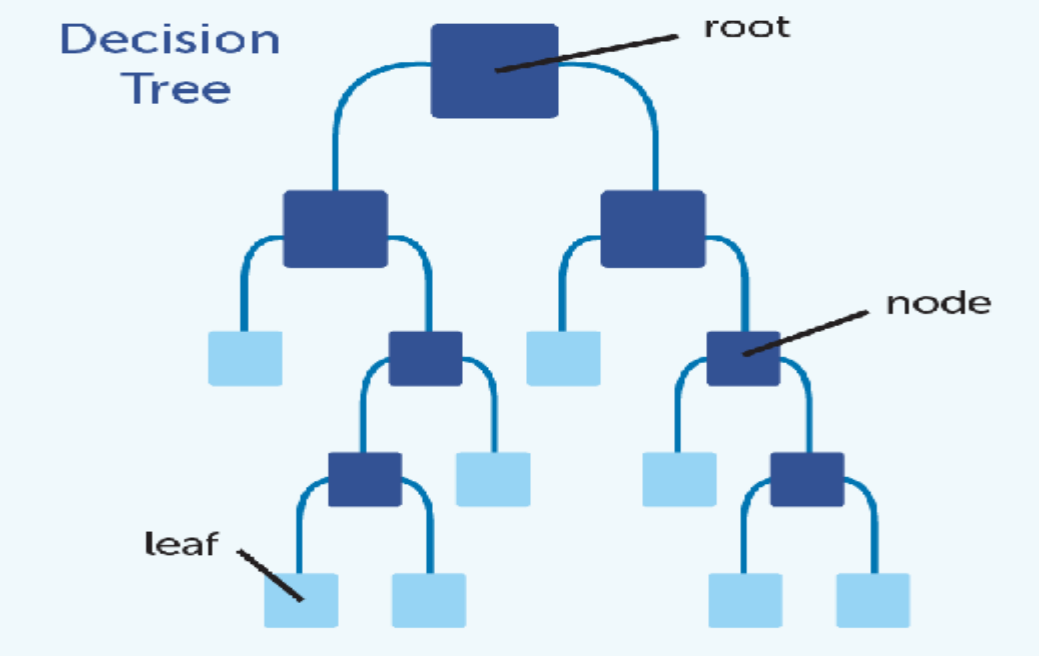
Classification Process

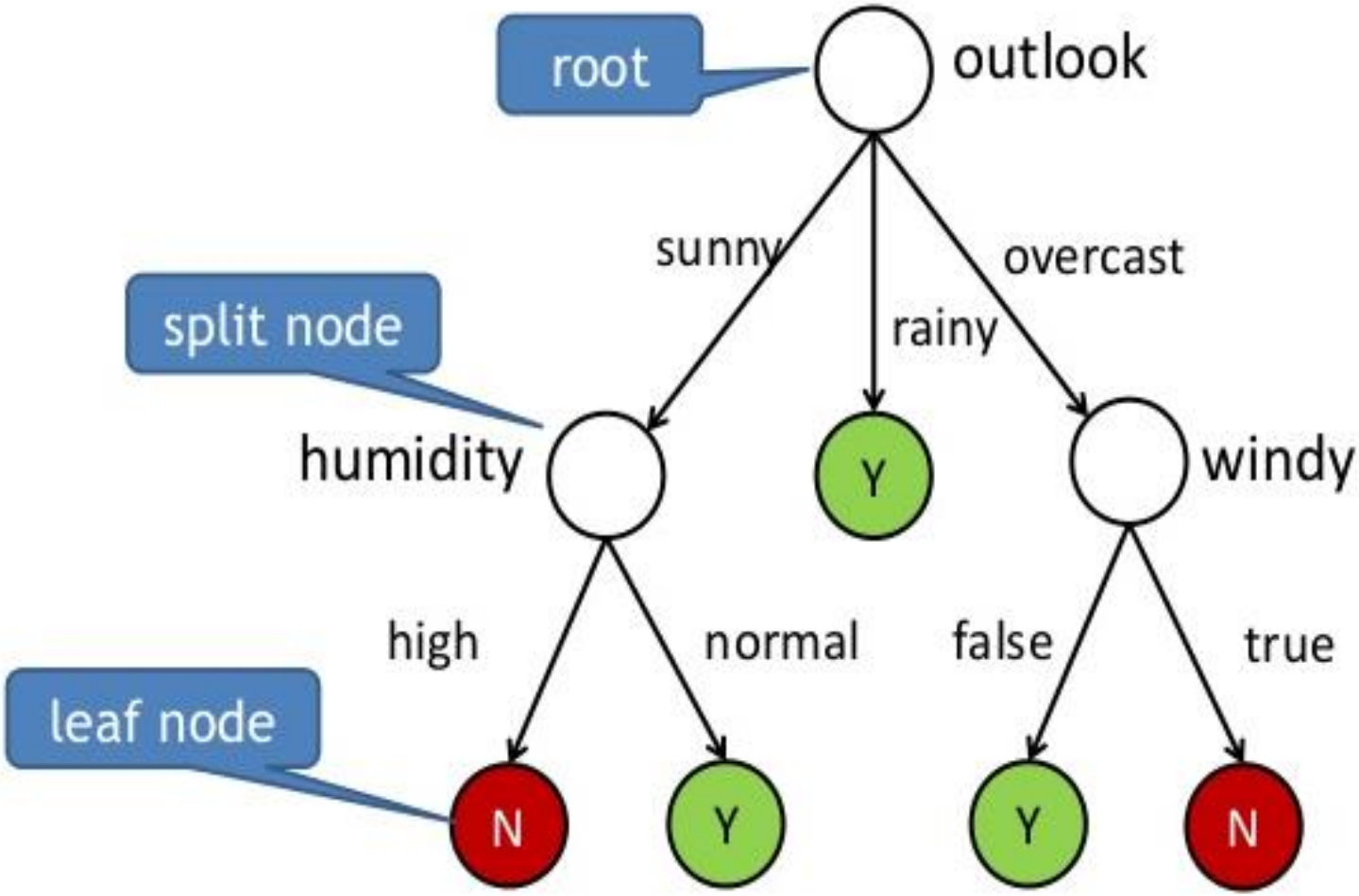
- Learning Phase
- Classification Phase



Decision Tree

A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node.




root
outlook
sunny
overcast
split node
rainy
humidity
Y
windy
high
normal
false
true
leaf node
N
Y
Y
N

Predicting Model

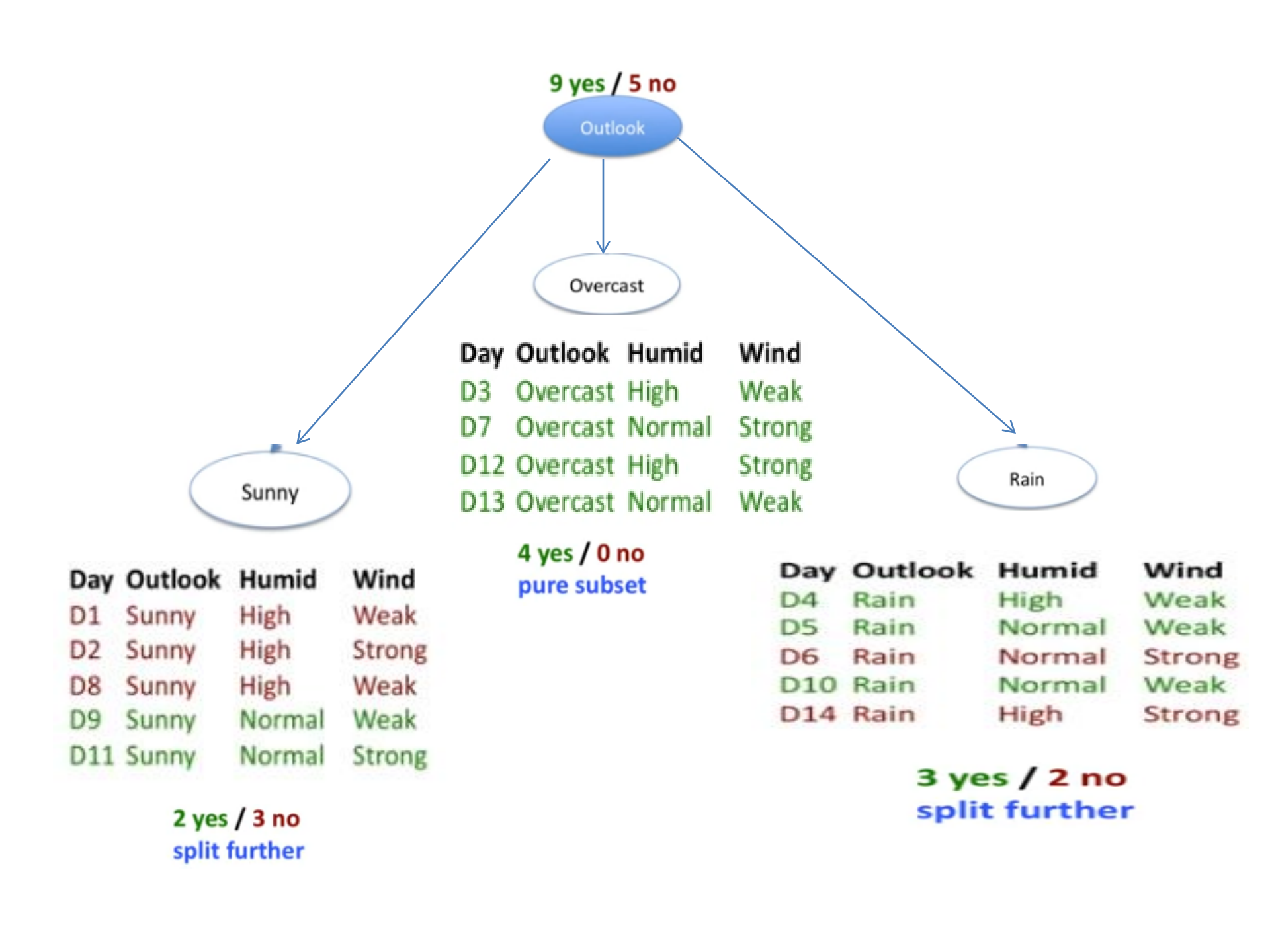
Training Data : Observe Data

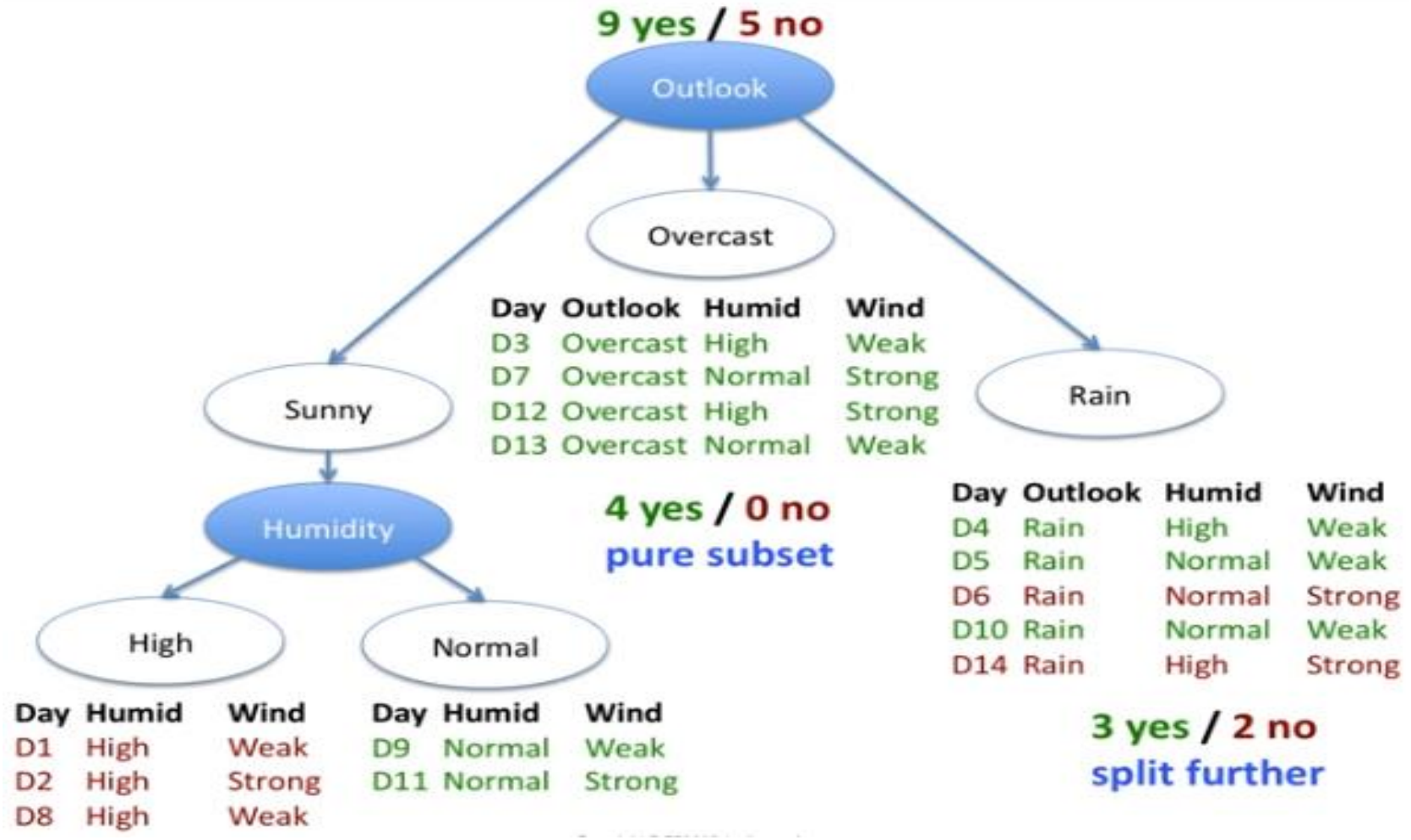
Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

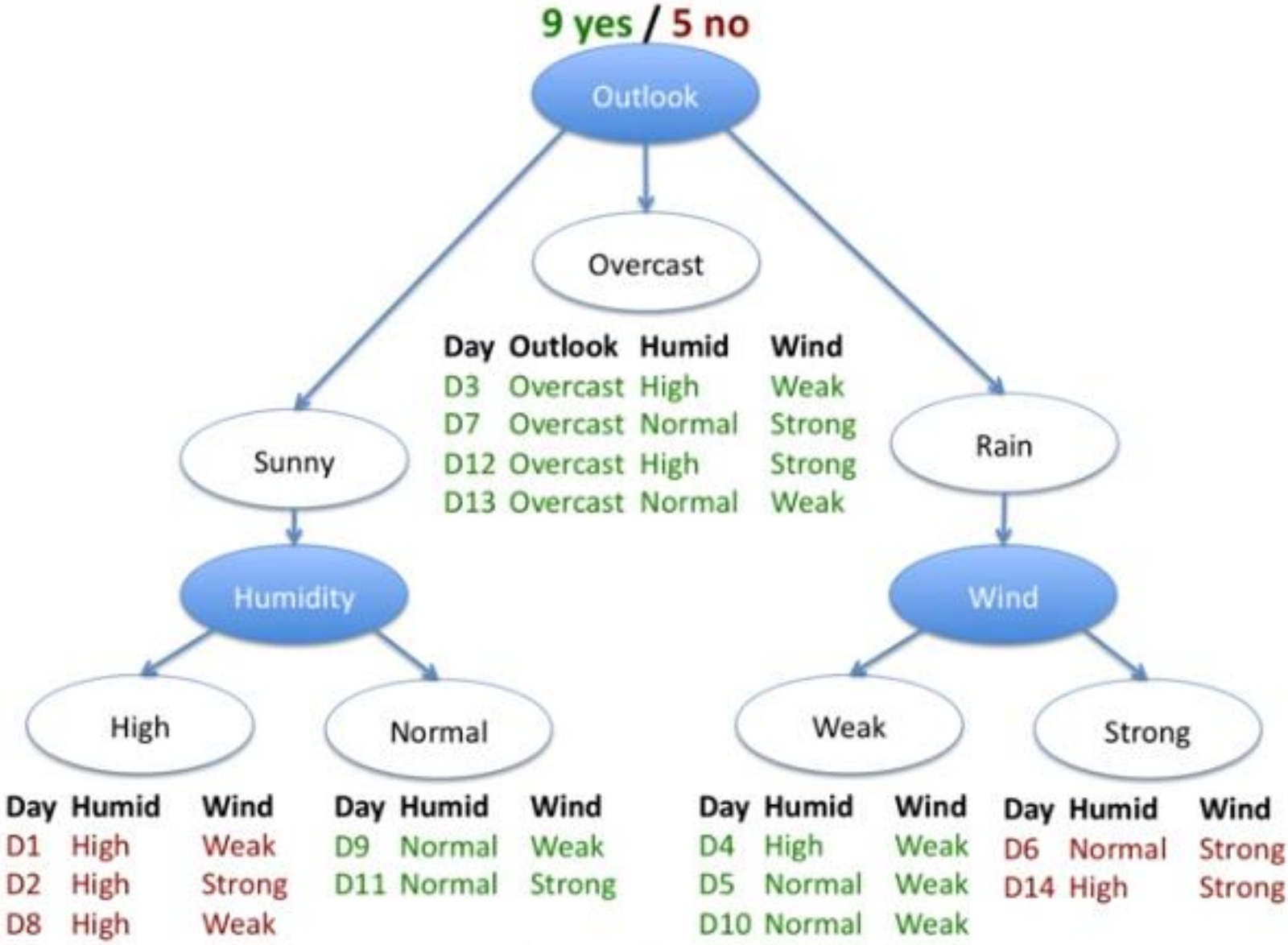
9 YES/5 NO

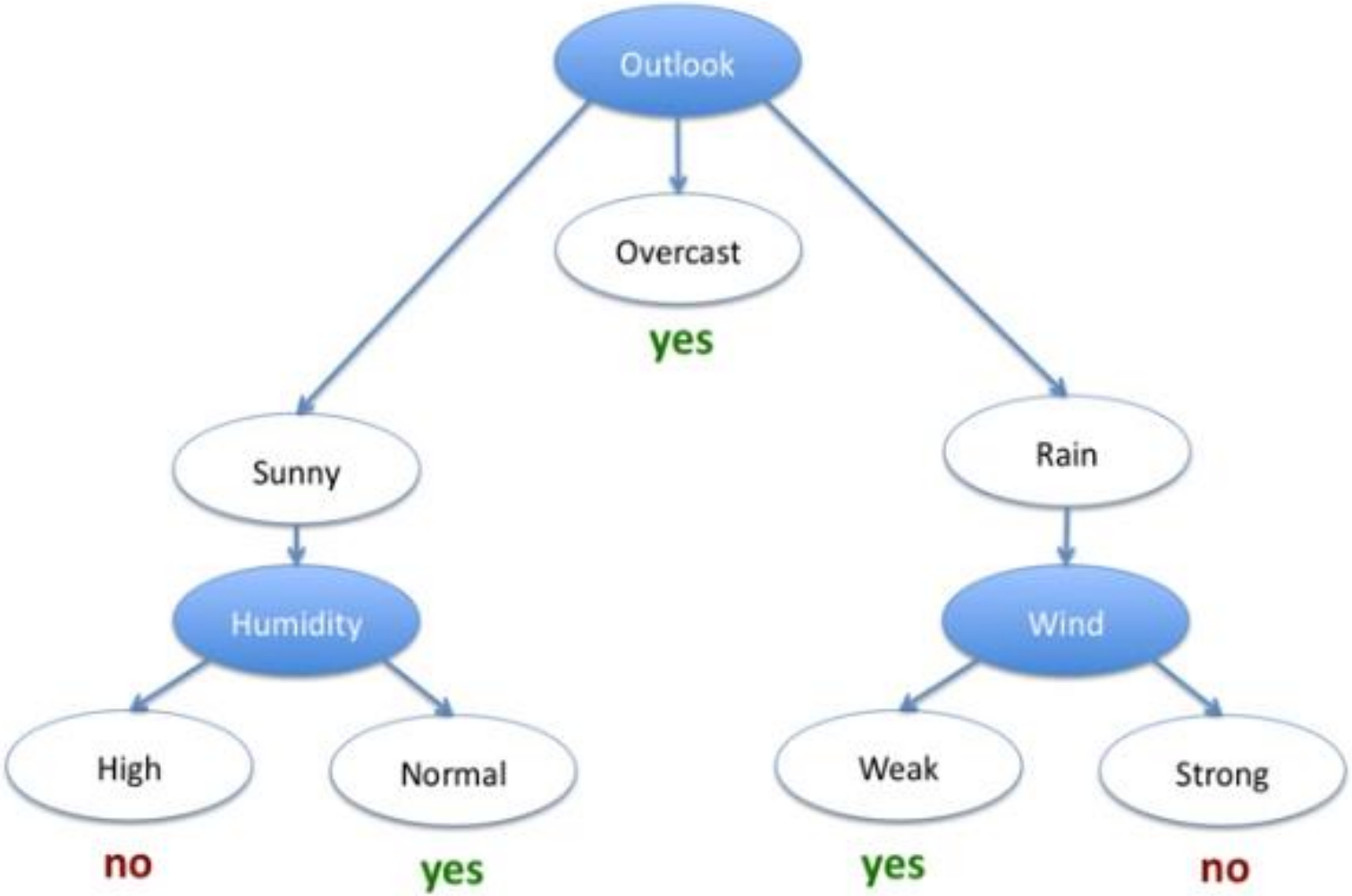
New data:

D15	Rain	High	Weak	?

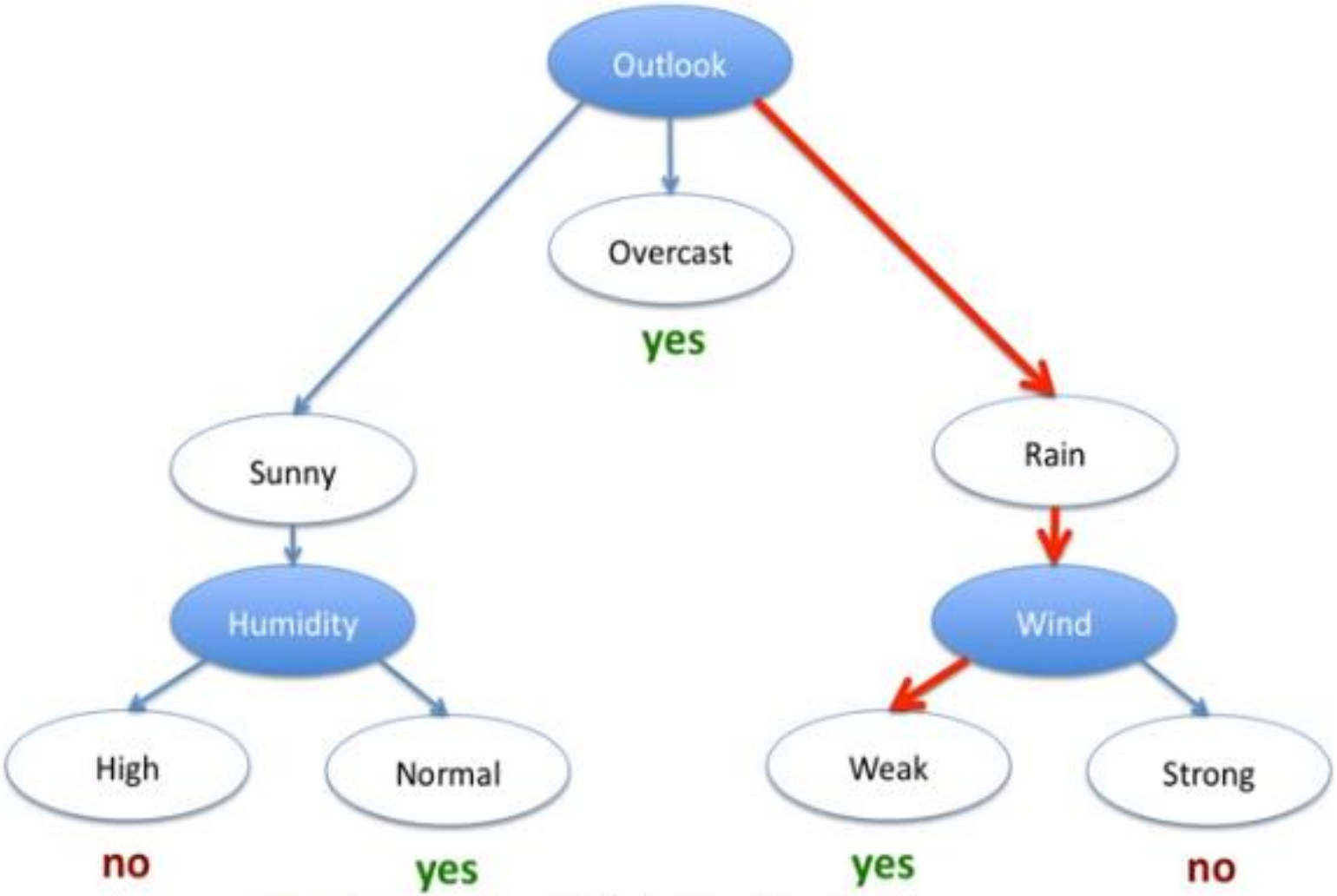

9 yes / 5 no
Outlook
Overcast
Sunny
Rain
Day Outlook Humid Wind
D3 Overcast High Weak
D7 Overcast Normal Strong
D12 Overcast High Strong
D13 Overcast Normal Weak
4 yes / 0 no
pure subset
Day Outlook Humid Wind
D1 Sunny High Weak
D2 Sunny High Strong
D8 Sunny High Weak
D9 Sunny Normal Weak
D11 Sunny Normal Strong
2 yes / 3 no
split further
Day Outlook Humid Wind
D4 Rain High Weak
D5 Rain Normal Weak
D6 Rain Normal Strong
D10 Rain Normal Weak
D14 Rain High Strong
3 yes / 2 no
split further


9 yes / 5 no
Outlook
Overcast
Sunny
Rain
Humidity
High
Normal
Day Outlook Humid Wind
D3 Overcast High Weak
D7 Overcast Normal Strong
D12 Overcast High Strong
D13 Overcast Normal Weak
4 yes / 0 no
pure subset
Day Outlook Humid Wind
D4 Rain High Weak
D5 Rain Normal Weak
D6 Rain Normal Strong
D10 Rain Normal Weak
D14 Rain High Strong
3 yes / 2 no
split further
Day Humid Wind
D1 High Weak
D2 High Strong
D8 High Weak
Day Humid Wind
D9 Normal Weak
D11 Normal Strong


9 yes / 5 no
Outlook
Overcast
Sunny
Rain
Humidity
Wind
High
Normal
Weak
Strong
Day Outlook Humid Wind
D3 Overcast High Weak
D7 Overcast Normal Strong
D12 Overcast High Strong
D13 Overcast Normal Weak
Day Humid Wind
D1 High Weak
D2 High Strong
D8 High Weak
Day Humid Wind
D9 Normal Weak
D11 Normal Strong
Day Humid Wind
D4 High Weak
D5 Normal Weak
D10 Normal Weak
Day Humid Wind
D6 Normal Strong
D14 High Strong

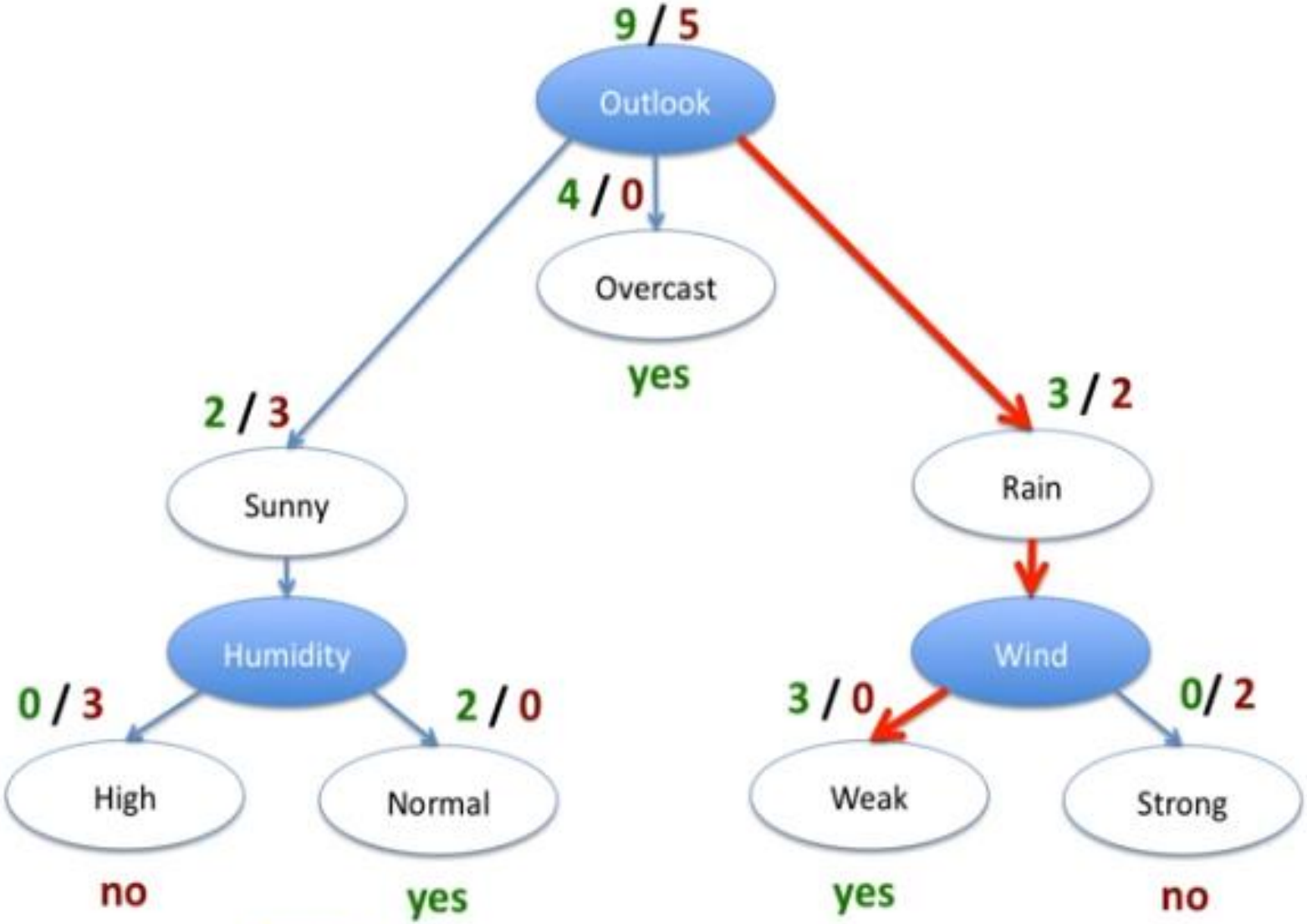


Day	Outlook	Humid	Wind
D15	Rain	High	Weak



New data:

Day	Outlook	Humid	Wind	
D15	Rain	High	Weak	→ Yes



New data:

Day	Outlook	Humid	Wind	
D15	Rain	High	Weak	→ Yes

Which Attribute to Split On

- pure set (4 yes / 0 no) => completely certain (100%)
- impure (3 yes / 3 no) => completely uncertain (50%)

- must be symmetric: 4 yes / 0 no as pure as 0 yes / 4 no

Entropy

Measure of purity of subset

- Entropy: $H(S) = - p_{(+)} \log_2 p_{(+)} - p_{(-)} \log_2 p_{(-)}$ bits
  - S ... subset of training examples
  - $p_{(+)}$ / $p_{(-)}$ ... % of positive / negative examples in S

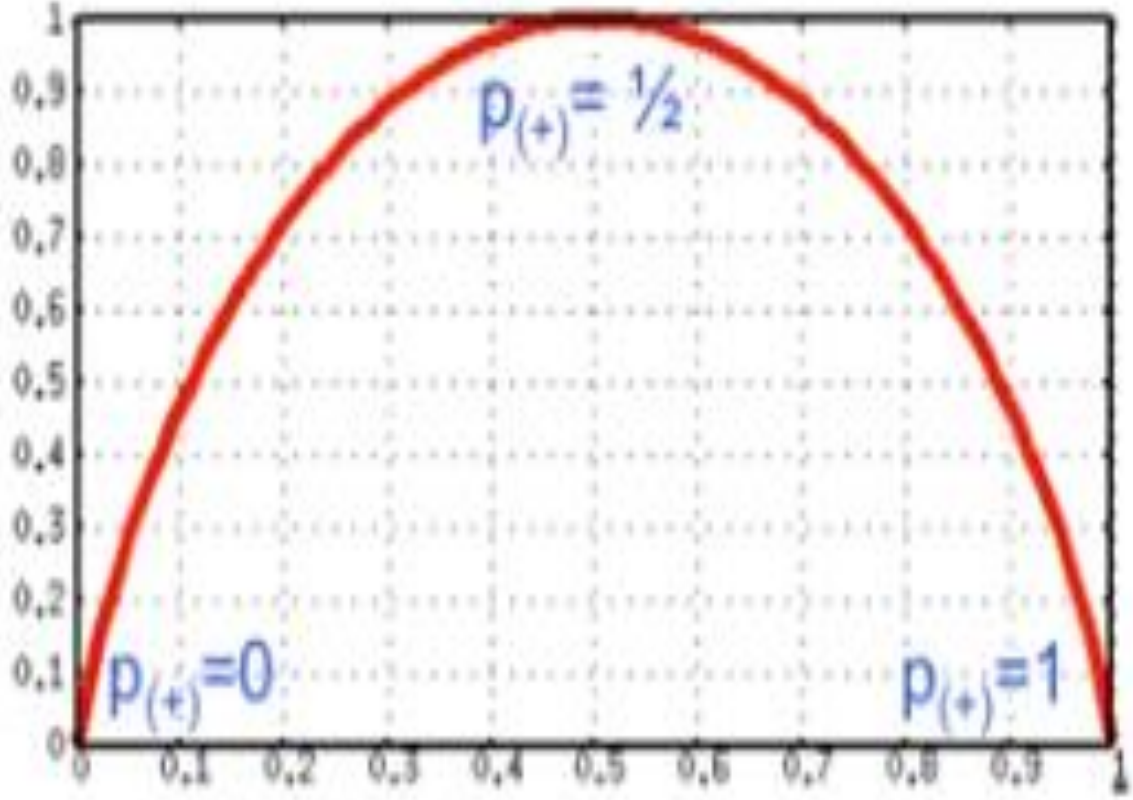
How many bits required to represent yes or no

- impure (3 yes / 3 no):

$$H(S) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1 \text{ bits}$$

- pure set (4 yes / 0 no):

$$H(S) = -\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4} = 0 \text{ bits}$$


$p_{(+)} = \frac{1}{2}$
$p_{(+)}=0$
$p_{(+)}=1$
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Information Gain

- Average of all Entropy

Gain (S, Wind)

$= H(S) - {}^{8}/_{14}\, H(S_{weak}) - {}^{6}/_{14}\, H(S_{weak})$

$= 0.94 - {}^{8}/_{14} * 0.81 - {}^{6}/_{14} * 1.0$

$= 0.049$

Data Mining Association Rule

Discovering Relationship between large Data set
This is if and then rule based
Analyzes customer behaviour

Bread => butter.

buys{ onions,potatoes} => buys {tomatoes}

Part of Association Rule

Bread => butter[20 %,45%].

Bread : Ancedent
Butter : Consequent
20%: Support
45%: Confidence

A => B

- Support denotes probability that contains both A & B.
- Confidence denotes probability that a transaction containing A also contains B.

  - Consider, in a Super Market

  Total transactions: 100.

  Bread : 20

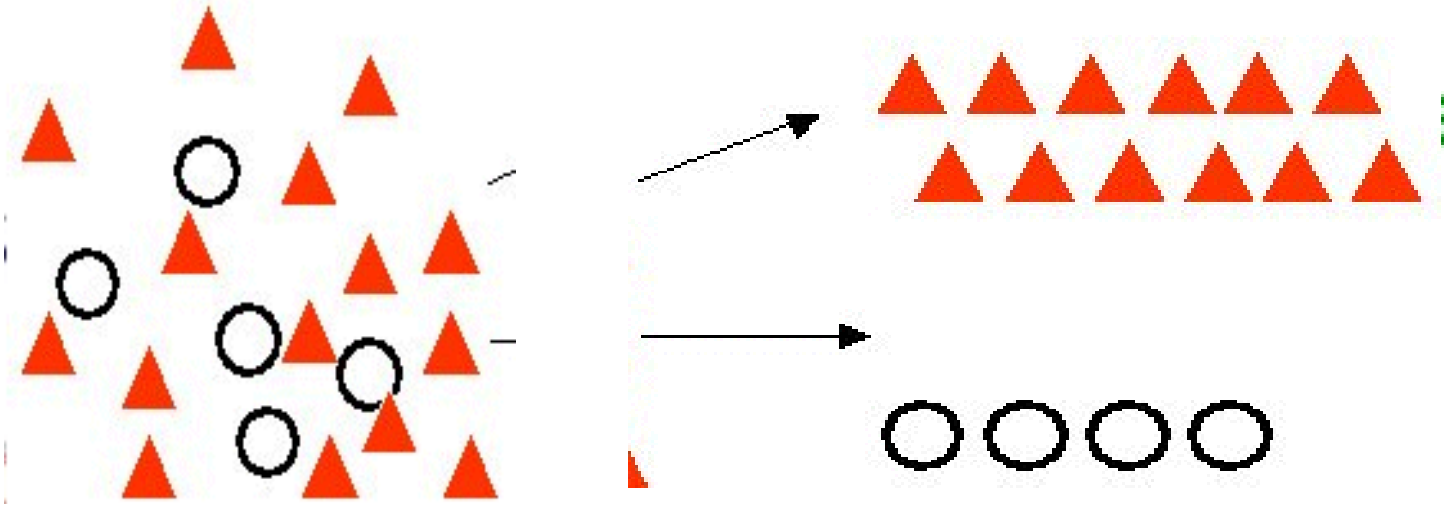
  So, 20/100 * 100 = 20 % which is support.

  In 20 transaction, butter: 9 transactions

  So, 9/20 * 100 = 45% which is confidence

Clustering

Partitioning a data into subclasses
Partitioning a data into subclasses is called cluster.
Grouping similar objects
Partitioning the data based on similarity.



Data Warehouse

Data Warehouse is relation Database which is Developed with a aim for query and analysis rather than for transaction processing.

A data warehouses provides us generalized and consolidated data in multidimensional view. Along with generalized and consolidated view of data, a data warehouses also provides us Online Analytical Processing (OLAP)

Data mining functions such as association, clustering, classification, prediction can be integrated with OLAP operations to enhance the interactive mining of knowledge at multiple level of abstraction.

Data Warehouse

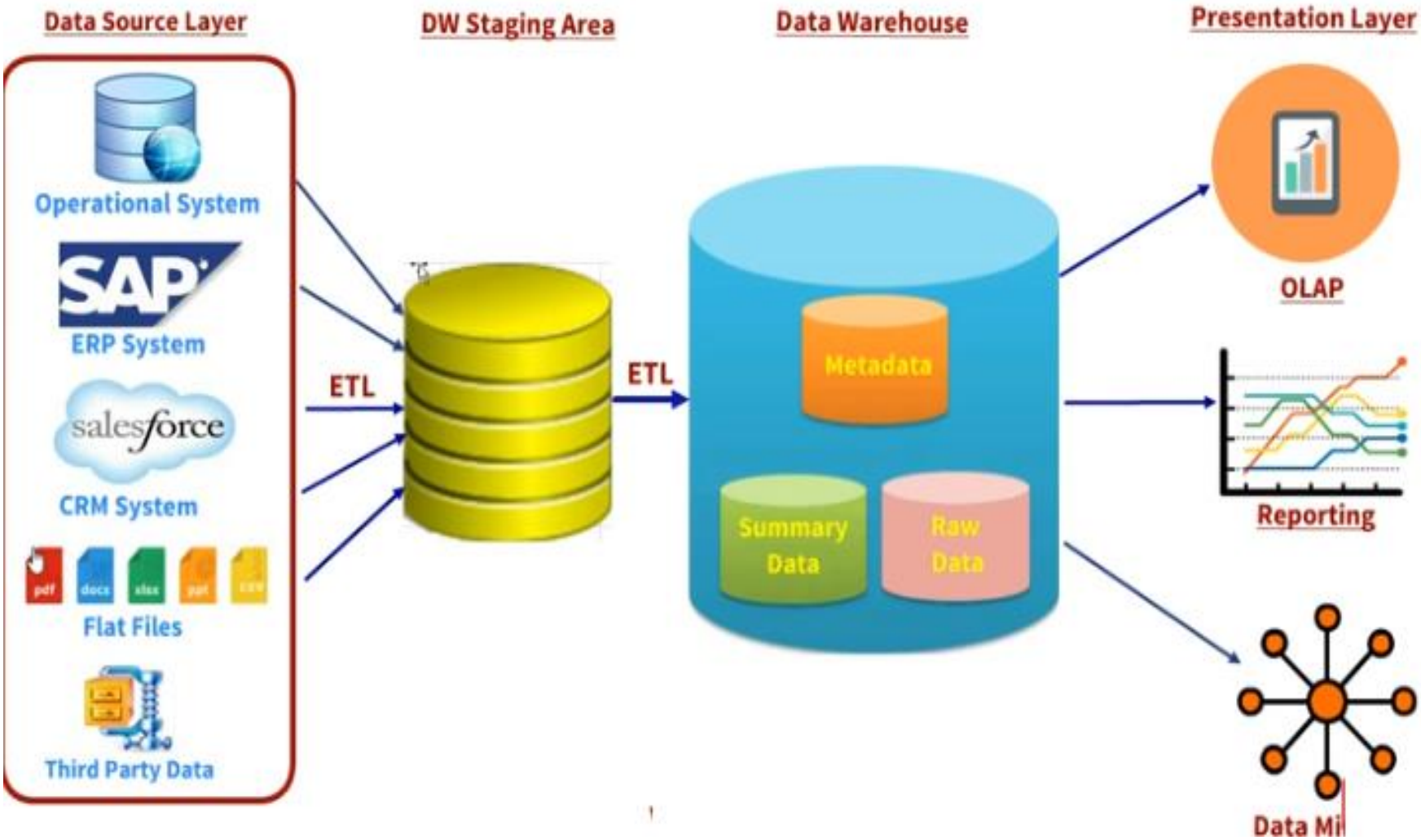
- **Subject Oriented**

  It Focuses on a subject rather than ongoing operation.
- **Integrated**

  Integrates data from multiple data Sources
- **Time Variant**
- **Non Volatile**

  Data should not change once in the warehouse.

  Previous data is not erased when new data is added to DWH


Data Source Layer
DW Staging Area
Data Warehouse
Presentation Layer
Operational System
ERP System
CRM System
Flat Files
Third Party Data
ETL
ETL
Metadata
Summary Data
Raw Data
OLAP
Reporting
Data Mi

ETL

Extract Transform & Load

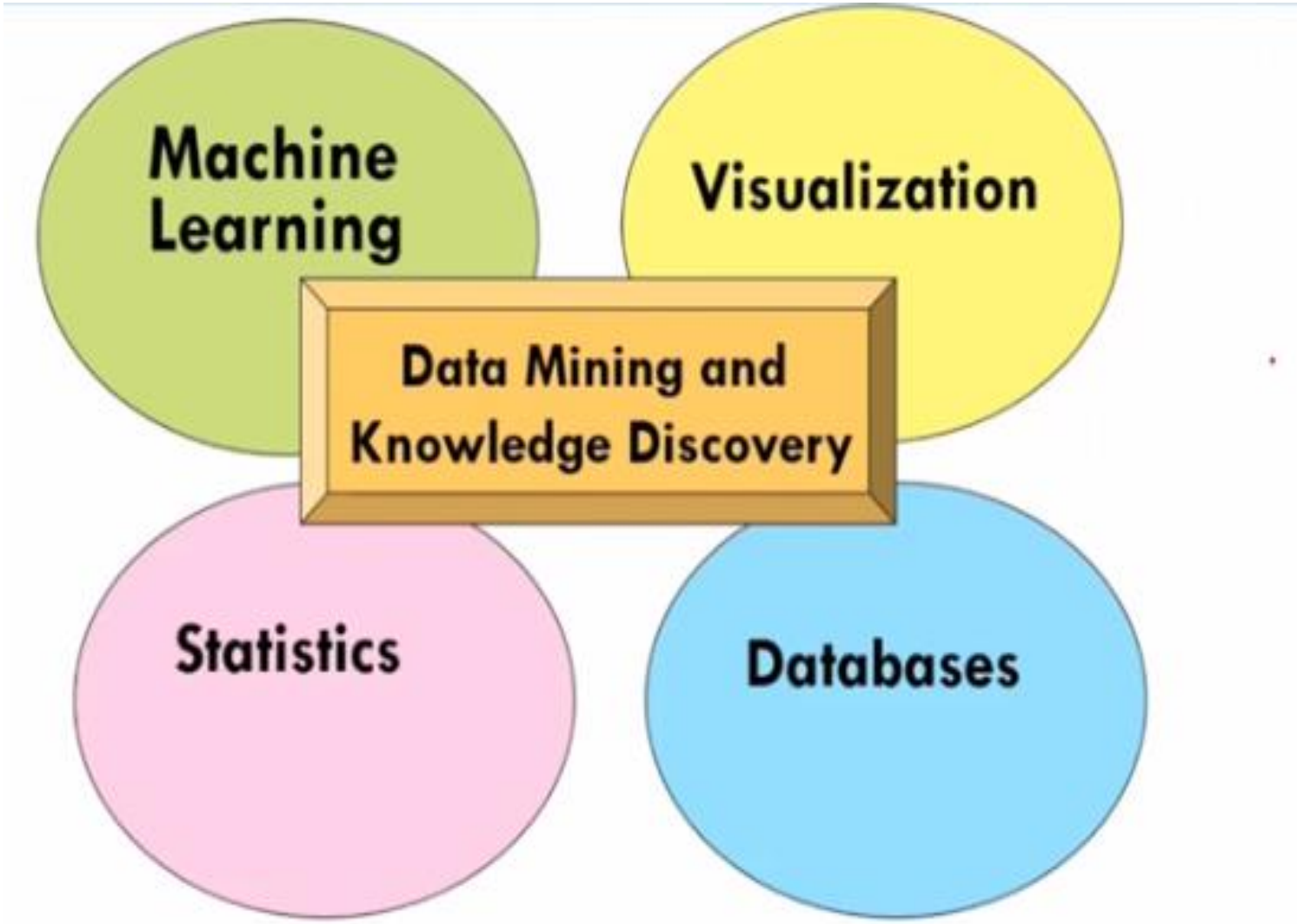
Full Extraction
Partial Extraction :with Update Notification
Partial Extraction : without Update Notification

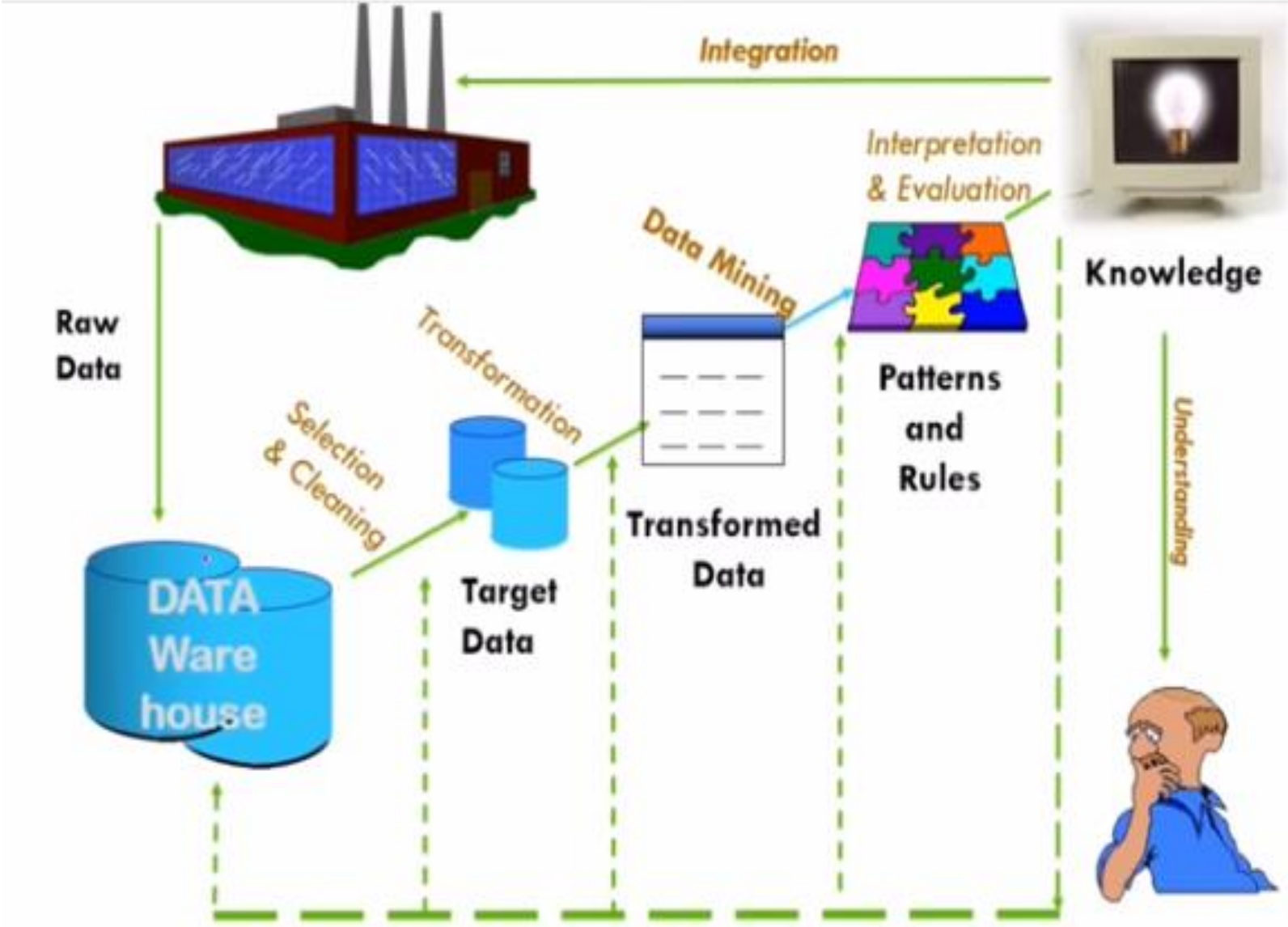
Data Extracted into a staging server is a raw data and can not be used it is .it needs to be Cleansed,Mapped and Trasformed.
Basic Transformation:
Selection
Matching
Data Cleansing or Enrichment
Consolidations or Summarization

ETL Tools

Enterprise Softwares

Open Source/Community Softwares


Machine Learning
Visualization
Data Mining and Knowledge Discovery
Statistics
Databases


Integration
Interpretation
& Evaluation
Knowledge
Data Mining
Patterns
and
Rules
Raw
Data
Transformation
Selection
& Cleaning
Transformed
Data
Target
Data
DATA
Ware
house
Understanding

Data Mining Technique

- Statistics
- Classification
- Machine Learning
- Pattern Recognition
- Regression
- Clustering
- Association Rule

What is Data?

Data is all around us. But what exactly is it? Data is a value assigned to a thing

What can we say about these?

They are golf balls, right?
So one of the first data points we have is that they are used for golf. Golf is a category of sport.
We have the colour: “white”, the condition “used”. They all have a size, there is a certain number of them and they probably have some monetary value, and so on.

Data, when collected and structured suddenly becomes a lot more useful. Let’s do this in the table below

Colour	White
Category	Sport – Golf
Condition	Used
Diameter	43mm
Price (per ball)	$0.5 (AUD)

Types of data

Qualitative data is everything that refers to the quality of something: A description of colours, texture and feel of an object , a description of experiences, and interview are all qualitative data.

Quantitative data is data that refers to a number. E.g. the number of golf balls, the size, the price, a score on a test etc.

Categorical data puts the item you are describing into a category: In our example the condition “used” would be categorical (with categories such as “new”, “used” ,”broken” etc.)

Unstructured vs. Structured data

Data for Humans

A plain sentence – “we have 5 white used golf balls with a diameter of 43mm at 50 cents each” – might be easy to understand for a human, but for a computer this is hard to understand. The above sentence is what we call unstructured data.

Data for Computers

Computers are inherently different from humans. It can be exceptionally hard to make computers extract information from certain sources. Some tasks that humans find easy are still difficult to automate with computers. For example, interpreting text that is presented as an image is still a challenge for a computer.