DATA MINING
DATA SCIENCE

WHAT IS THE DIFFERENCE?

DATA MINING

Data Mining is about finding the trends in a data set. And using these trends to identify future patterns. It often includes analyzing the vast amount of historical data which was previously ignored.

DATA SCIENCE

Data Science is a field of study which includes everything from Big Data Analytics, Data Mining, Predictive Modeling, Data Visualization, Mathematics, and Statistics.

#1. What is it?

Data Mining

A Technique.

Data Science

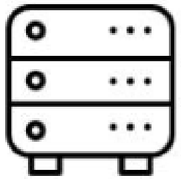
An Area

#2. Focus

Data Mining	Data Science
	
Business Process.	Scientific Study.

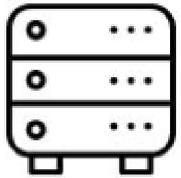
#3. Goal

Data Mining



Make
data more usable.

Data Science



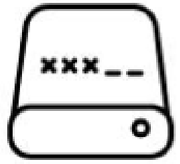
Building
Data-centric products for an
organization.

#4. Output

Data Mining	Data Science
	
Patterns.	Evaried.

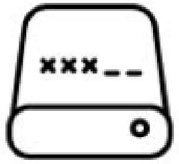
#5. Purpose

Data Mining



Finding
trends previously not known.

Data Science



Social analysis, building predictive models, unearthing unknown facts, and more,

#6. Vocational Perspective

Data Mining

Someone
with a knowledge of navigating
across data and statistical
understanding can
conduct data mining.

Data Science

A
person needs to understand
Machine Learning,
Programming, info-graphic
techniques and have domain
knowledge to become a data
scientist

#7. Extent

Data Mining

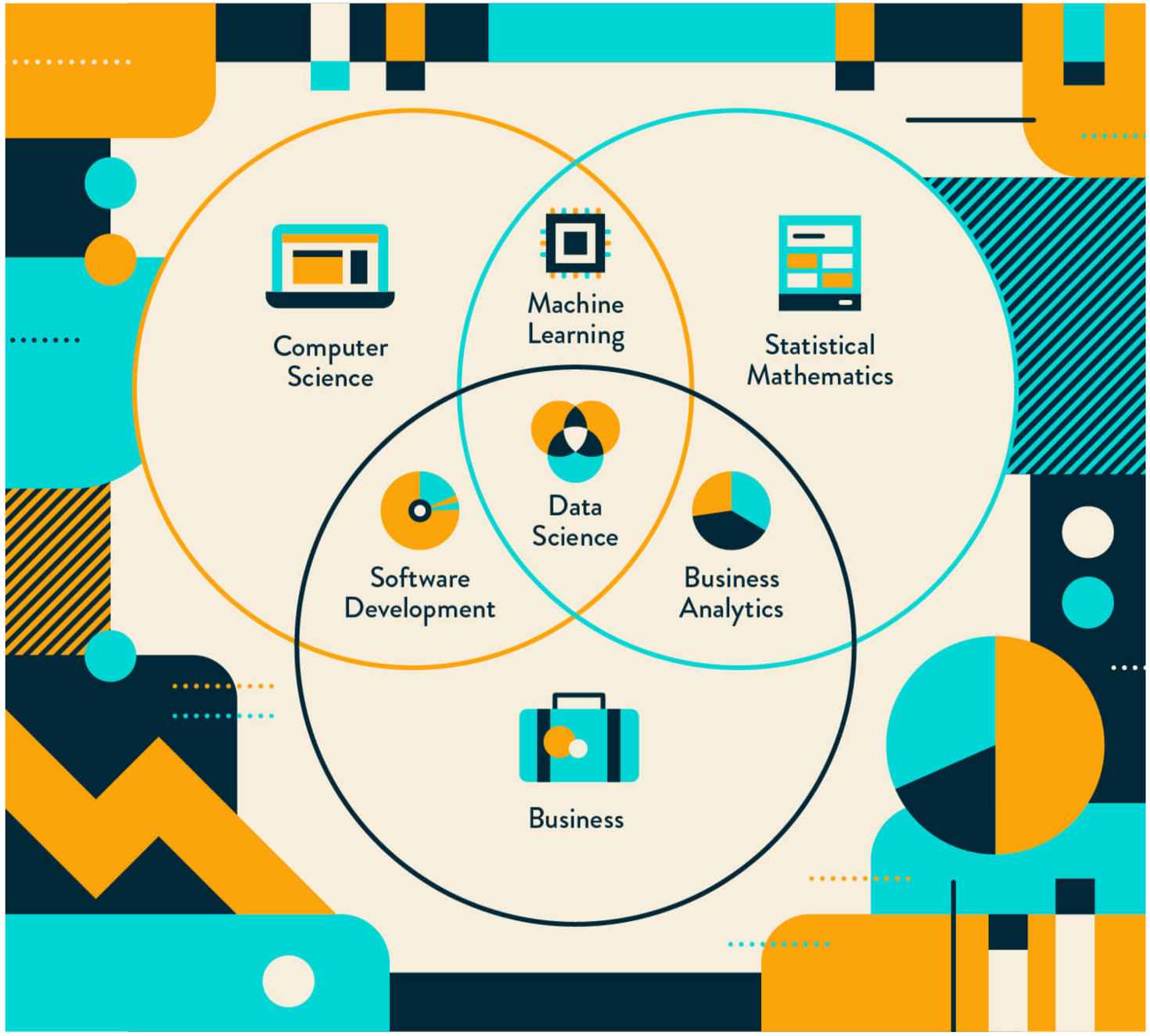
Data
mining can be a subset of Data Science as Mining activities are part of Data Science pipeline.

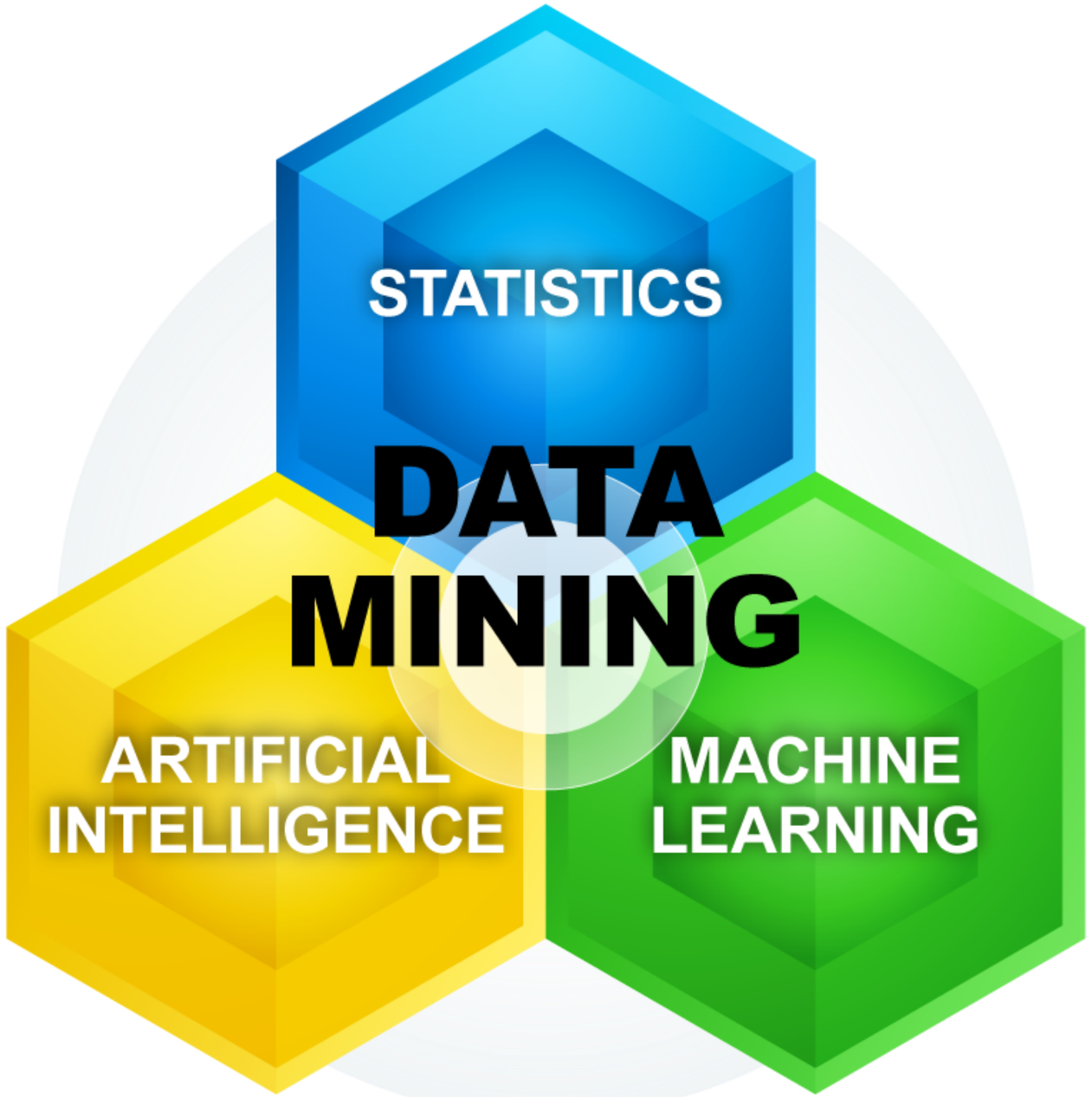
Data Science

Multidisciplinary
– Data Science consists of Data Visualizations, Computational Social Sciences, Statistics, Data Mining, Natural Language Processing, et cetera

#8. Deals with (type of data)

Data Mining	Data Science
	
Mostly structured.	All forms of data – structured, semi-structured and unstructured.

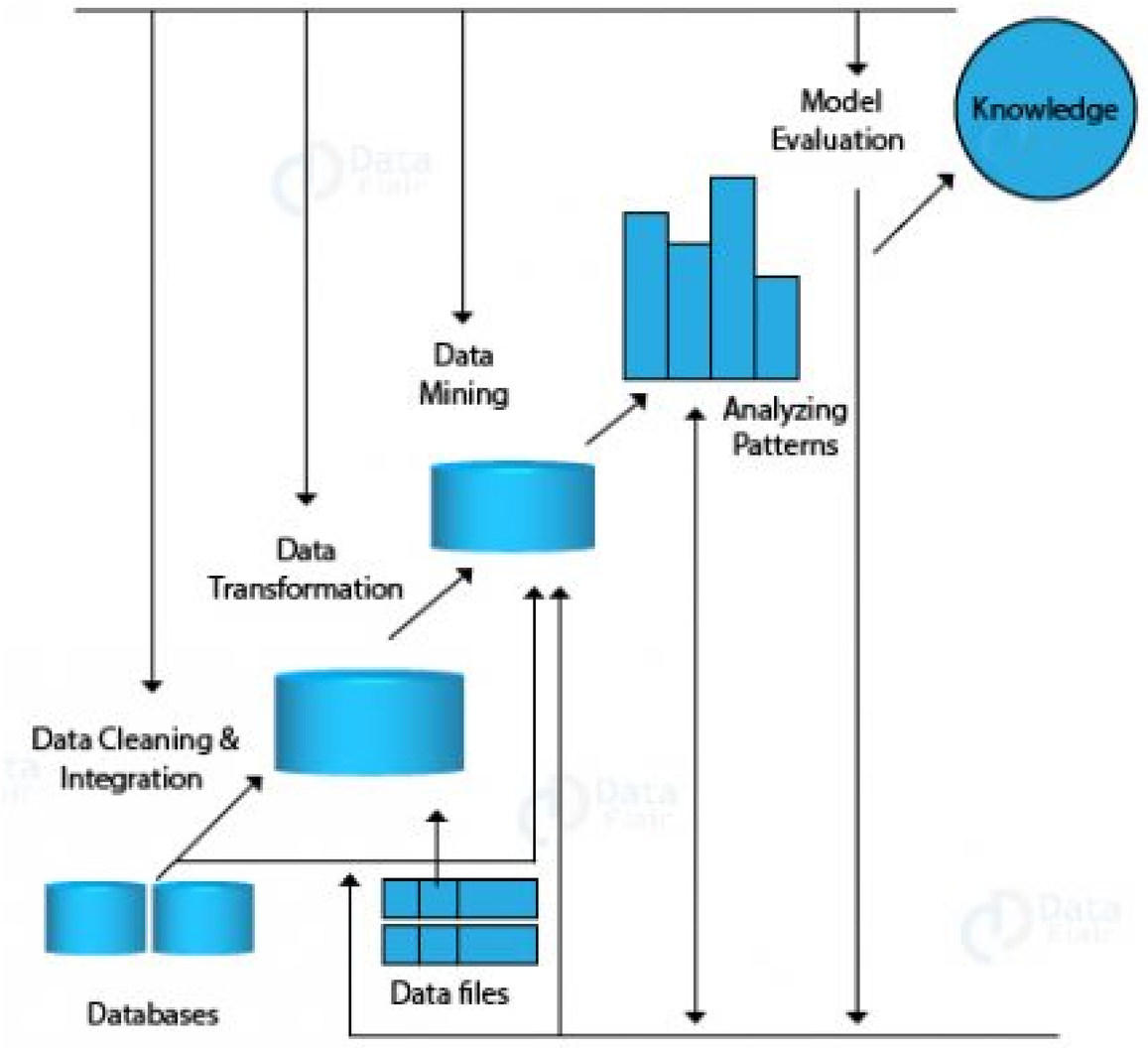

Computer Science
Machine Learning
Statistical Mathematics
Data Science
Software Development
Business Analytics
Business


STATISTICS
DATA MINING
ARTIFICIAL INTELLIGENCE
MACHINE LEARNING

KEY DIFFERENCES BETWEEN DATA SCIENCE ∇S DATA MINING

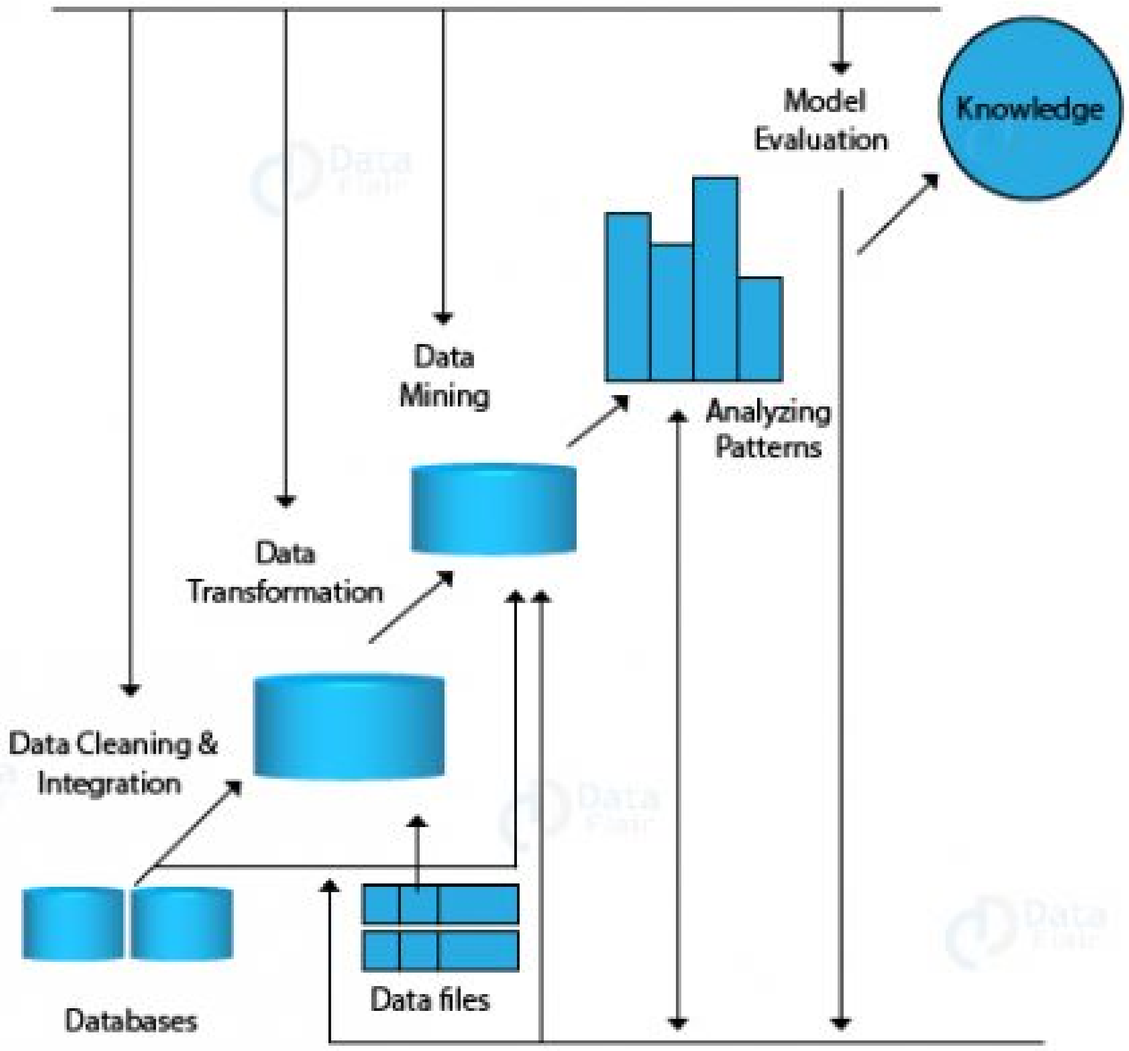
- **Data Mining** is an activity which is a part of a broader Knowledge Discovery in Databases (KDD) Process while **Data Science** is a field of study just like Applied Mathematics or Computer Science.
- Often **Data Science** is looked upon in a broad sense while **Data Mining** is considered a niche.
- Some activities under **Data Mining** such as statistical analysis, writing data flows and pattern recognition can intersect with **Data Science**. Hence, Data Mining becomes a subset of Data Science.
- Machine Learning in Data Mining is used more in pattern recognition while in Data Science it has a more general use.

DATA MINING STEPS



DATA MINING STEPS

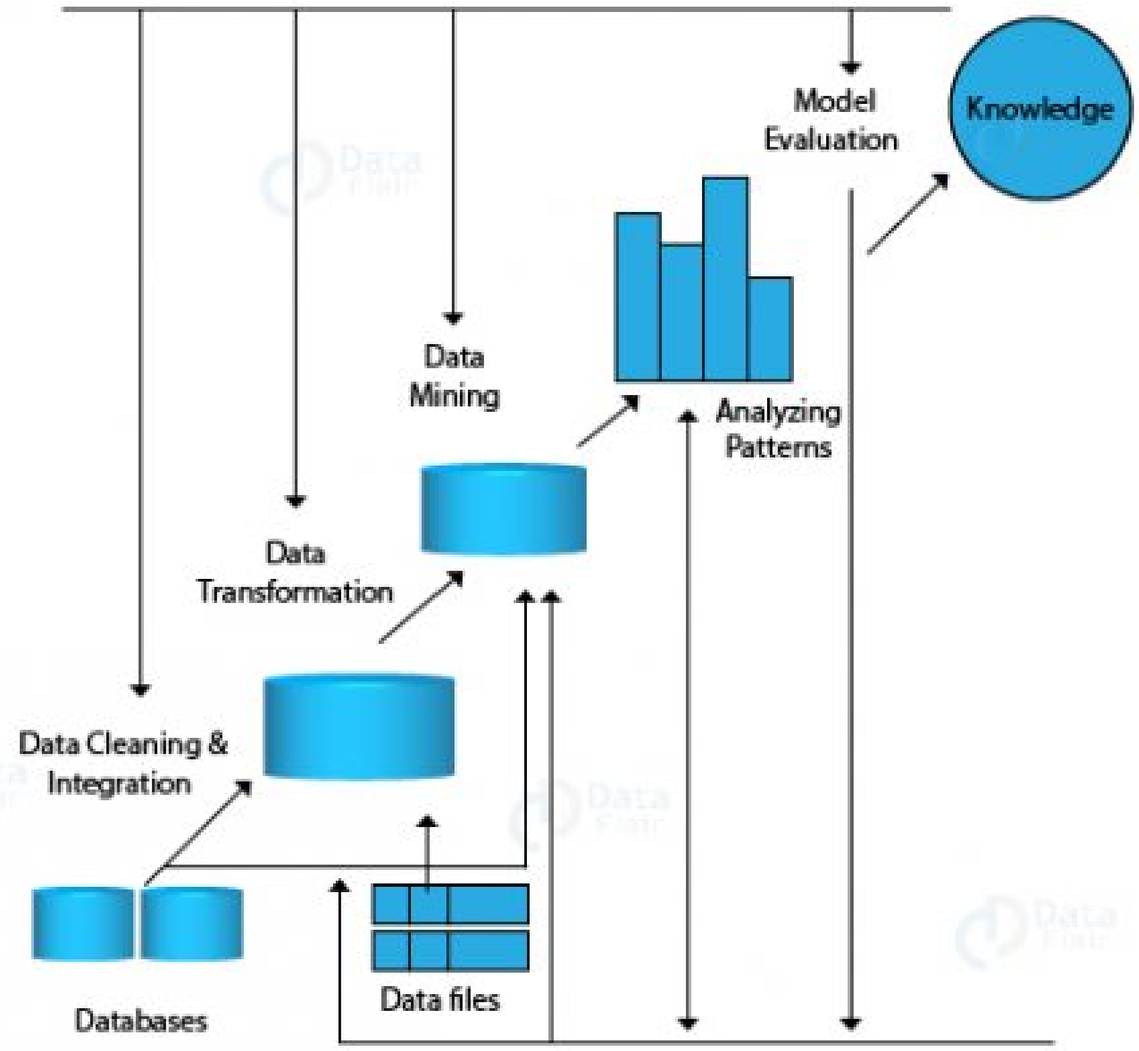
01



- Integrating data- To collect and combine data from all different sources.
- Selecting data- Not all the data gathered is useful, so in the next step, we select only the data which is useful for data mining.
- Data cleaning- The selected data may contain errors, missing values, and inconsistency that needs to be cleaned. Different techniques and tools are required in this process.

DATA MINING STEPS

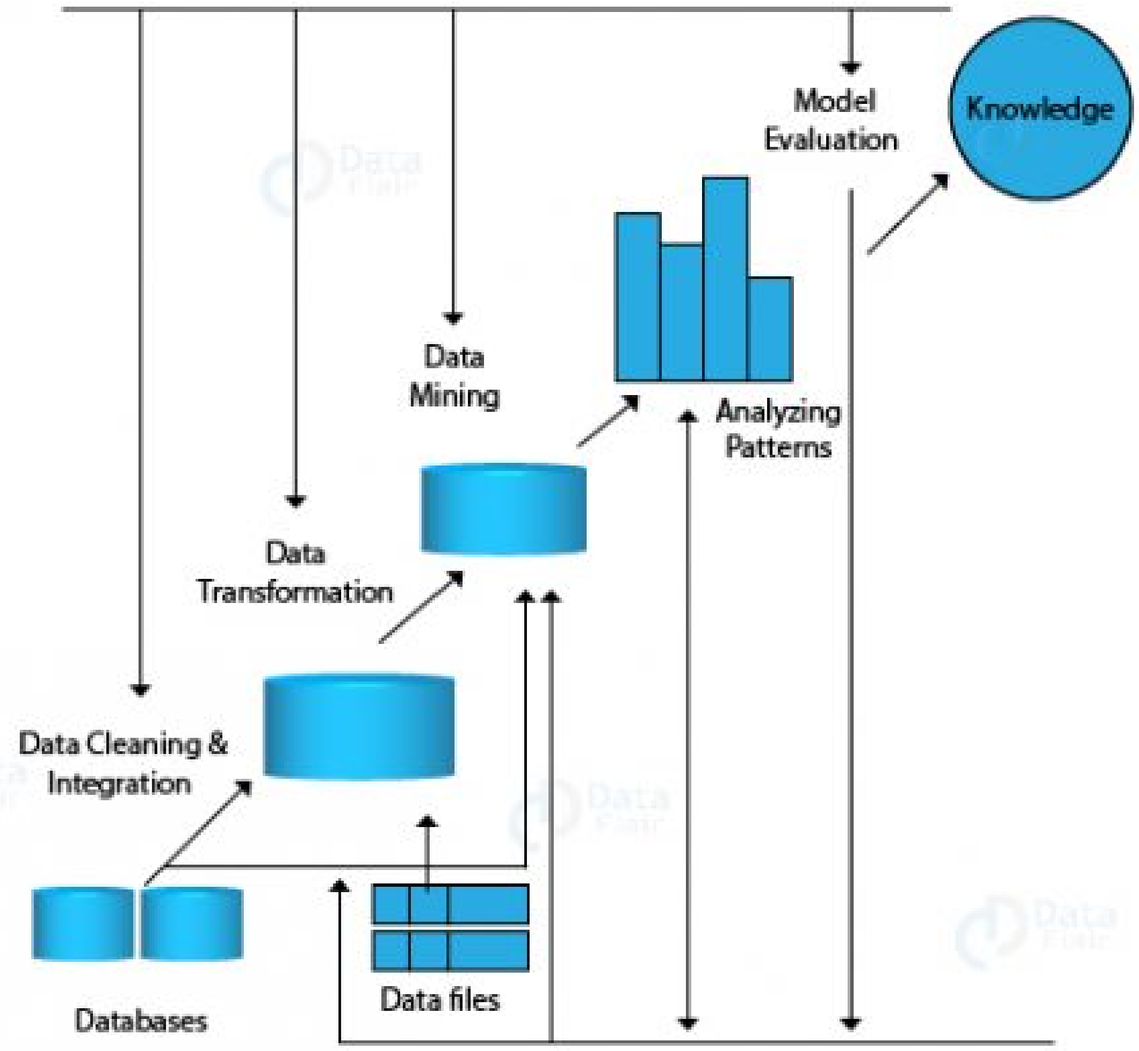
02



- Data transformation- Smoothing, aggregation, normalization, etc are some techniques used to transform data into an understandable format.

DATA MINING STEPS

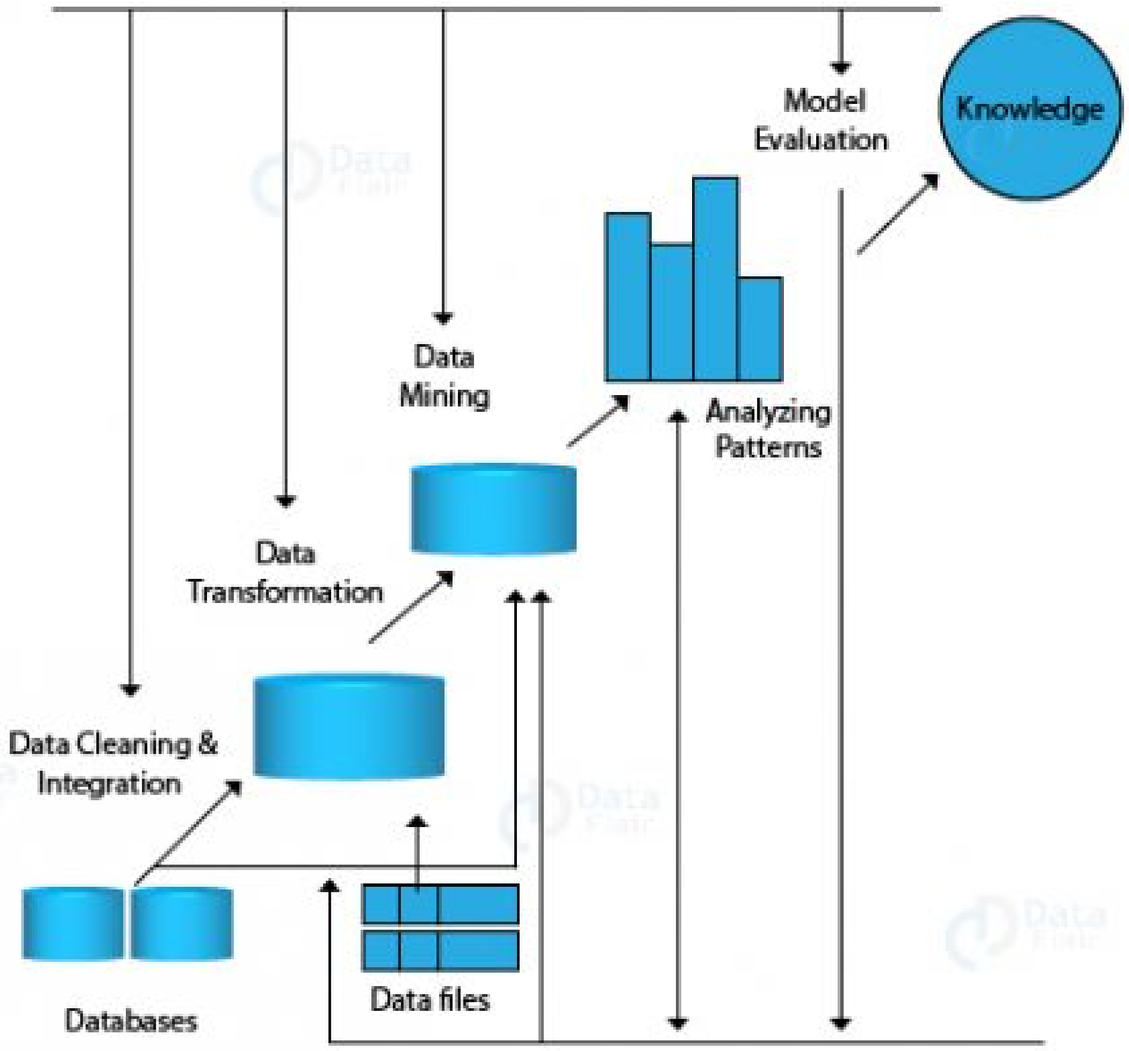
03



- Data mining- Now is the time to apply techniques like clustering and association analysis for the process of data mining and to discover interesting patterns.

DATA MINING STEPS

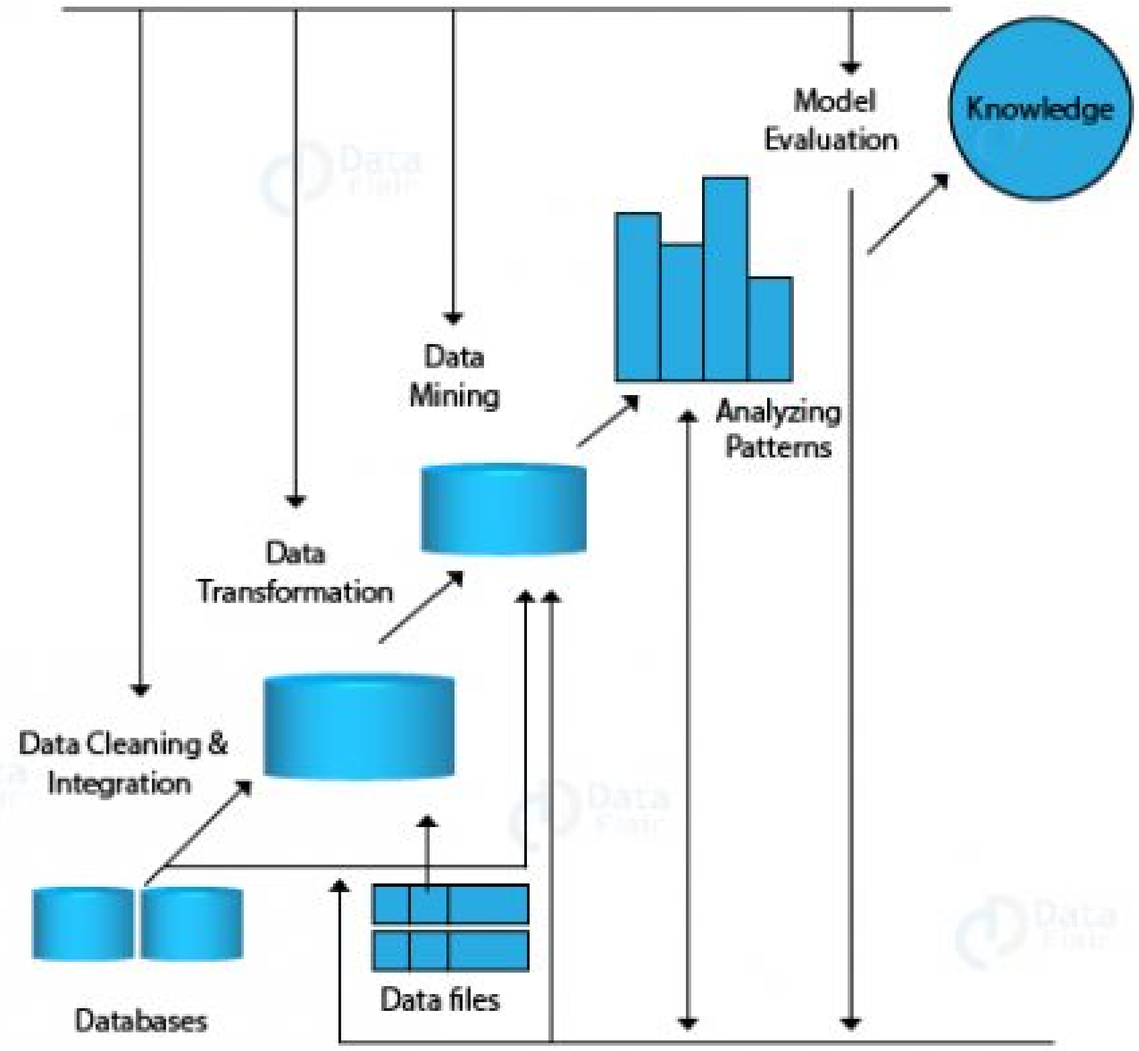
04



- Pattern evaluation- Removing redundant patterns to avoid confusion and analyzing remaining patterns is an essential step in this process.

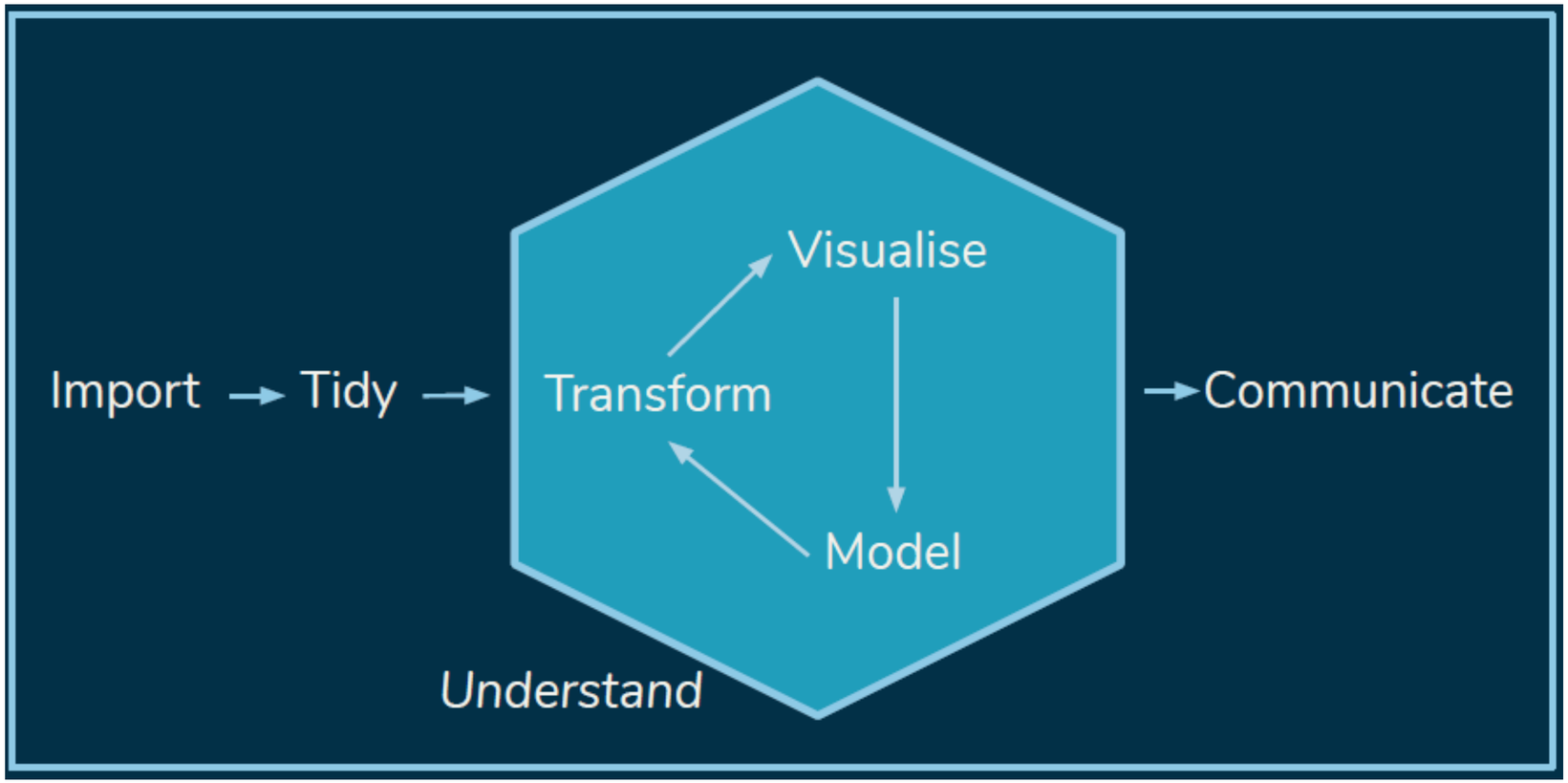
DATA MINING STEPS

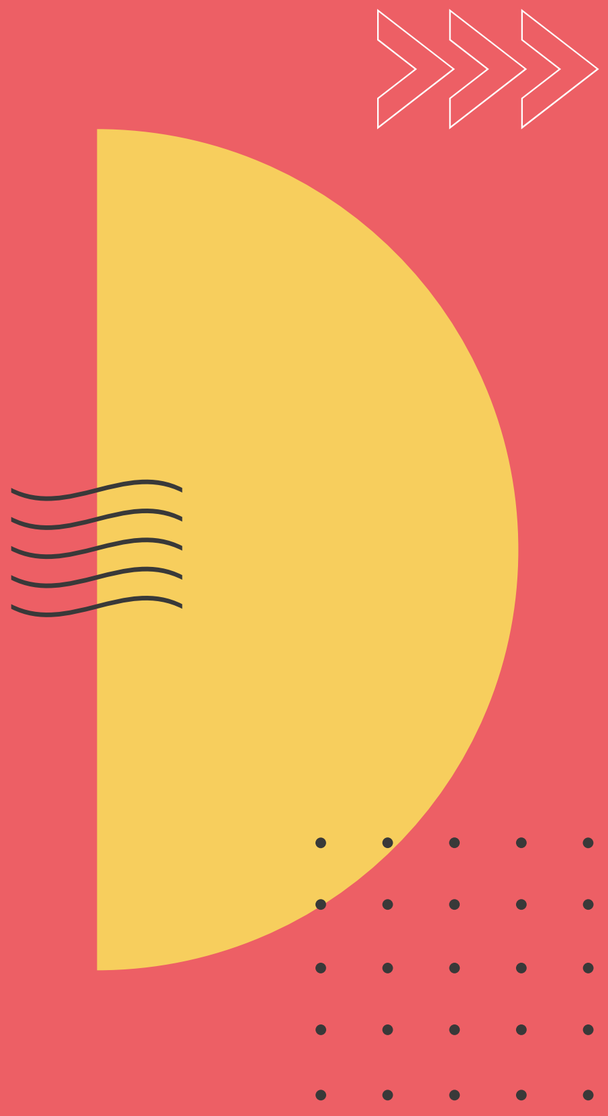
05



- Using the discovered knowledge- The final step in this process is to make proper use of the knowledge that is discovered in this process to make decisions.

DATA SCIENCE STEPS





CORE IDEAS IN DATA MINING

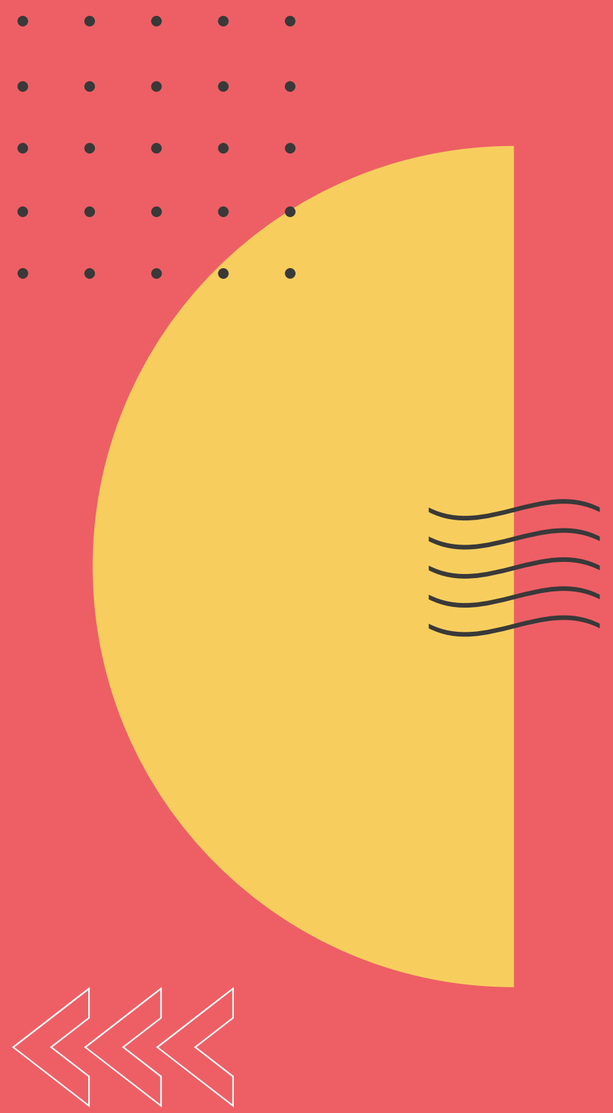
CLASSIFICATION

The most basic form on data analysis. A common task in data mining is to examine data where the classification is unknown or will occur in the future, with the goal of predicting what that classification is or will be. Similar data where the classification is known are used to develop rules, which are then applied to the data with the unknown classification.

CORE IDEAS IN DATA MINING

PREDICTION

Similar to classification, except that we are trying to predict the value of a numerical variable (e.g., amount of purchase) rather than a class (e.g., purchaser or nonpurchaser). In classification we are trying to predict a class, but the term prediction refers to the prediction of the value of a continuous variable. (Sometimes in the data mining literature, the terms estimation and regression are used to refer to the prediction of the value of a continuous variable, and prediction may be used for both continuous and categorical data.)



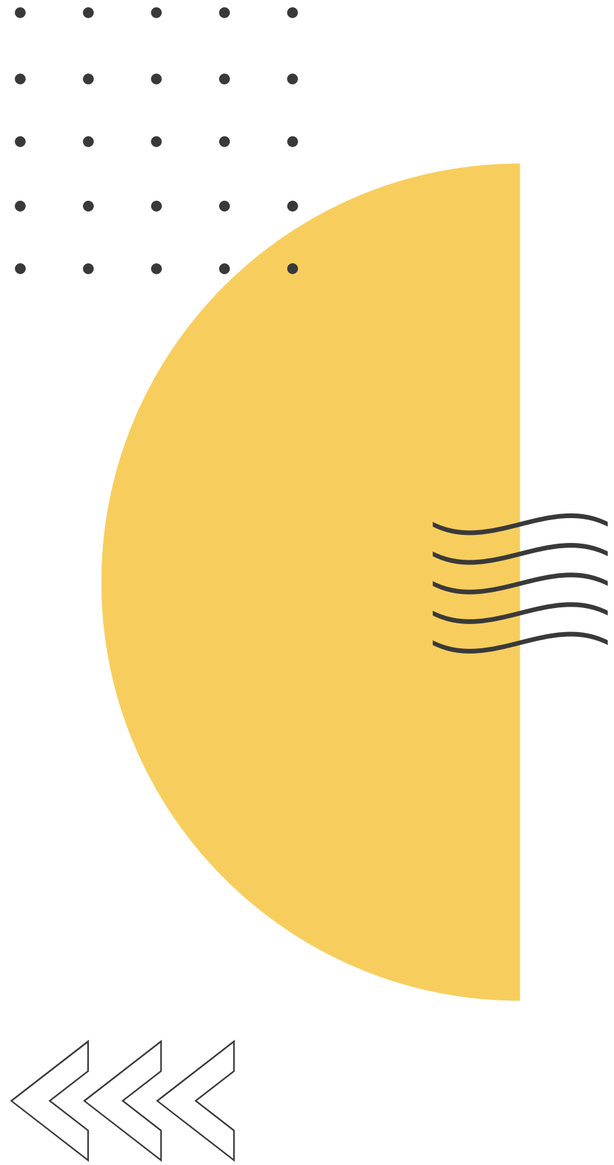
CORE IDEAS IN DATA MINING

ASSOCIATION RULES AND RECOMMENDATION SYSTEMS

Association rules, or affinity analysis, is designed to find such general associations patterns between items in large databases. The rules can then be used in a variety of ways.

EXAMPLES:

1. Large databases of customer transactions lend themselves naturally to the analysis of associations among items purchased, or "what goes with what."
    - the rules for weekly promotional offers or for bundling products.
2. Hospital database on patients' symptoms during consecutive hospitalizations can help find "which symptom is followed by what other symptom" to help predict future symptoms for returning patients.



CORE IDEAS IN DATA MINING

PREDICTIVE ANALYTICS

Classification, prediction, and to some extent, association rules and collaborative filtering constitute the analytical methods employed in predictive analytics. The term predictive analytics is sometimes used to also include data pattern identification methods such as clustering.

CORE IDEAS IN DATA MINING

DATA REDUCTION AND DIMENSION REDUCTION

The process of consolidating a large number of records (or cases) into a smaller set is termed data reduction. Reducing the number of variables is typically called dimension reduction. Dimension reduction is a common initial step before deploying data mining methods, intended to improve predictive power, manageability, and interpretability.

CORE IDEAS IN DATA MINING

SUPERVISED LEARNING

Supervised learning algorithms are those used in classification and prediction.

- Data contain outcome of interest is known.
  - also called “labeled data,” since they contain the label (outcome value) for each record or "training data"
  - the classification or prediction algorithm “learns,” or is “trained,” about the relationship between predictor variables and the outcome variable
- The validation data where the outcome is known but initially hidden, to see how well it does in comparison to other models.

CORE IDEAS IN DATA MINING

UNSUPERVISED LEARNING

Unsupervised learning algorithms are those used where there is no outcome variable to predict or classify. Hence, there is no "learning" from cases where such an outcome variable is known. Association rules, dimension reduction methods, and clustering techniques are all unsupervised learning methods.

DISKUSI:

Metode/Model/Algoritma apa saja yang telah di pelajari di mata kuliah: *Machine Learning*, Sistem Pakar, Sistem Cerdas & Pendukung Keputusan / *Artificial Intelligence*?