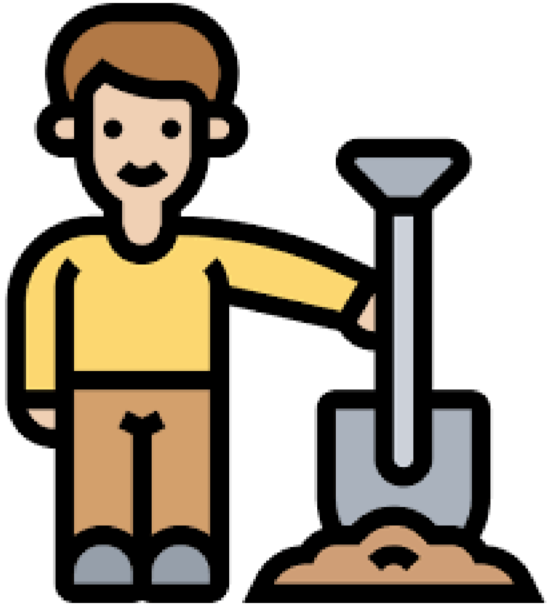
CCS2313 Data Mining
2024

جامعة البخاري الدولية
ALBUKHARY INTERNATIONAL UNIVERSITY



Chapter 1

Introduction to **Data Mining**

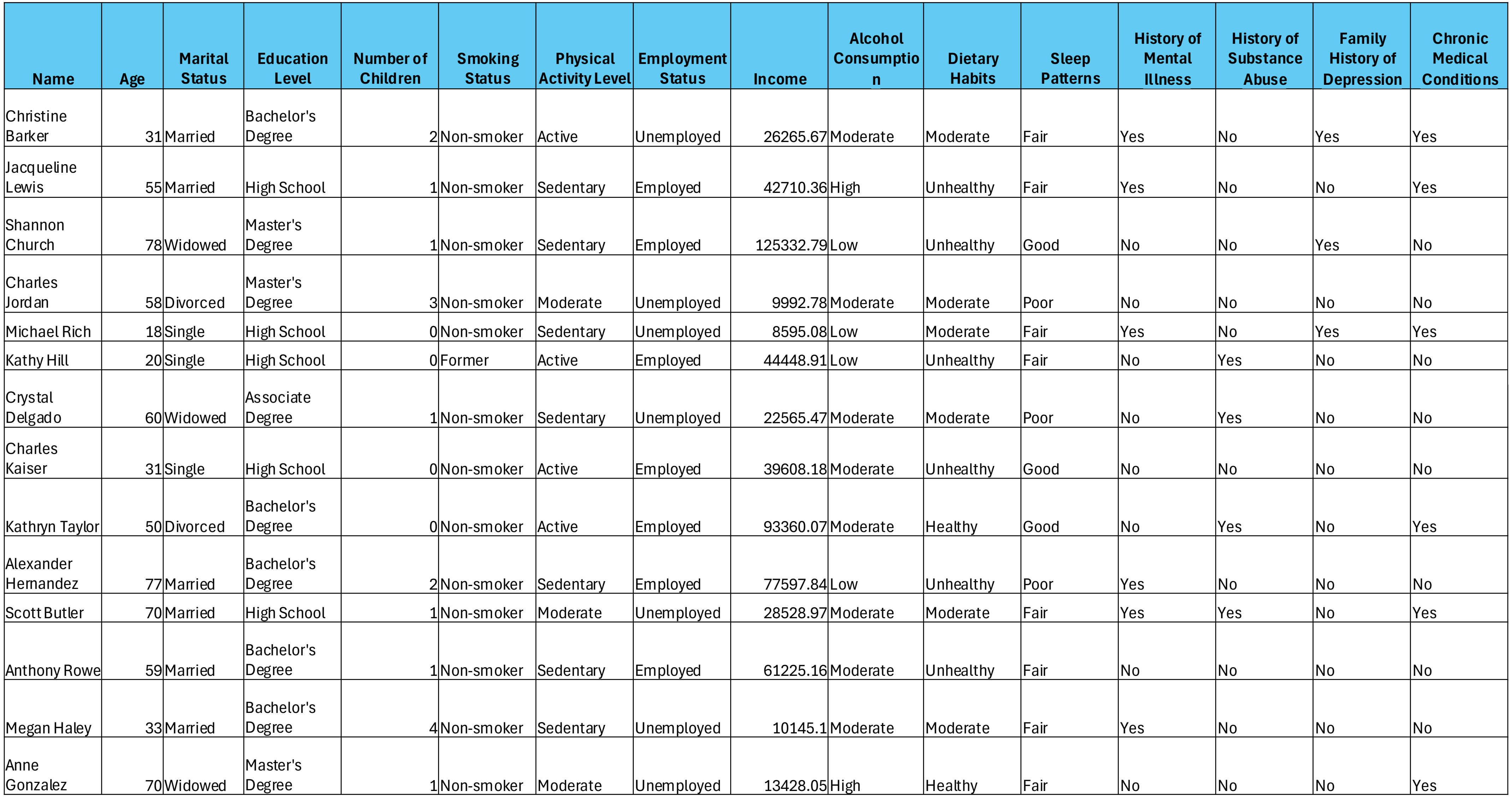
Outline

- Data mining definition
- What data mining can offer?
- Data mining task
- Why need data mining?
- The dig data
- Data mining process
- Challenges in data mining
- Tools for data mining

Data Mining at Glance

Depression Dataset

https://www.kaggle.com/datasets/anthonytherrien/depression-dataset

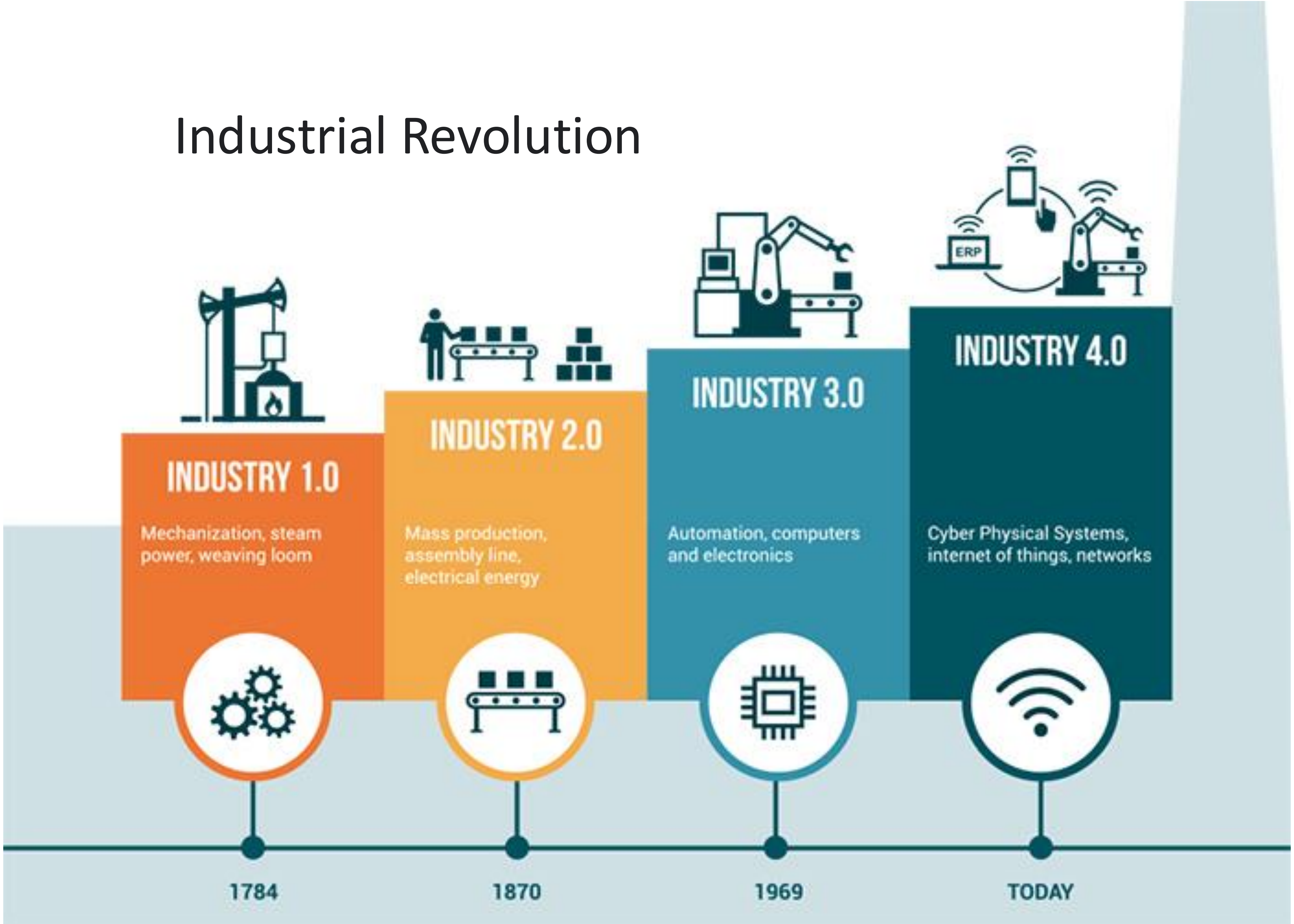


Name	Age	Marital Status	Education Level	Number of Children	Smoking Status	Physical Activity Level	Employment Status	Income	Alcohol Consumption	Dietary Habits	Sleep Patterns	History of Mental Illness	History of Substance Abuse	Family History of Depression	Chronic Medical Conditions
Christine Barker	31	Married	Bachelor's Degree	2	Non-smoker	Active	Unemployed	26265.67	Moderate	Moderate	Fair	Yes	No	Yes	Yes
Jacqueline Lewis	55	Married	High School	1	Non-smoker	Sedentary	Employed	42710.36	High	Unhealthy	Fair	Yes	No	No	Yes
Shannon Church	78	Widowed	Master's Degree	1	Non-smoker	Sedentary	Employed	125332.79	Low	Unhealthy	Good	No	No	Yes	No
Charles Jordan	58	Divorced	Master's Degree	3	Non-smoker	Moderate	Unemployed	9992.78	Moderate	Moderate	Poor	No	No	No	No
Michael Rich	18	Single	High School	0	Non-smoker	Sedentary	Unemployed	8595.08	Low	Moderate	Fair	Yes	No	Yes	Yes
Kathy Hill	20	Single	High School	0	Former	Active	Employed	44448.91	Low	Unhealthy	Fair	No	Yes	No	No
Crystal Delgado	60	Widowed	Associate Degree	1	Non-smoker	Sedentary	Unemployed	22565.47	Moderate	Moderate	Poor	No	Yes	No	No
Charles Kaiser	31	Single	High School	0	Non-smoker	Active	Employed	39608.18	Moderate	Unhealthy	Good	No	No	No	No
Kathryn Taylor	50	Divorced	Bachelor's Degree	0	Non-smoker	Active	Employed	93360.07	Moderate	Healthy	Good	No	Yes	No	Yes
Alexander Hernandez	77	Married	Bachelor's Degree	2	Non-smoker	Sedentary	Employed	77597.84	Low	Unhealthy	Poor	Yes	No	No	No
Scott Butler	70	Married	High School	1	Non-smoker	Moderate	Unemployed	28528.97	Moderate	Moderate	Fair	Yes	Yes	No	Yes
Anthony Rowe	59	Married	Bachelor's Degree	1	Non-smoker	Sedentary	Employed	61225.16	Moderate	Unhealthy	Fair	No	No	No	No
Megan Haley	33	Married	Bachelor's Degree	4	Non-smoker	Sedentary	Unemployed	10145.1	Moderate	Moderate	Fair	Yes	No	No	No
Anne Gonzalez	70	Widowed	Master's Degree	1	Non-smoker	Moderate	Unemployed	13428.05	High	Healthy	Fair	No	No	No	Yes

What can you do with this information?

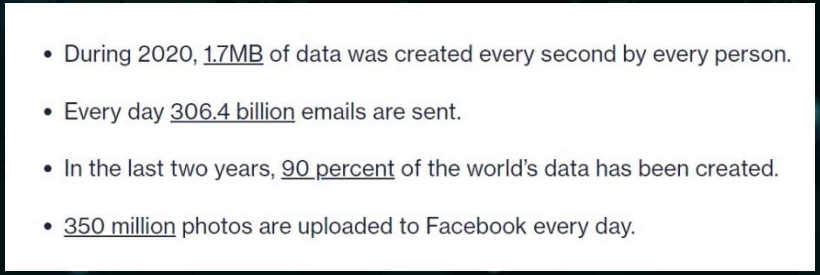
Data Mining

Industrial Revolution


INDUSTRY 1.0
Mechanization, steam power, weaving loom
INDUSTRY 2.0
Mass production, assembly line, electrical energy
INDUSTRY 3.0
Automation, computers and electronics
INDUSTRY 4.0
Cyber Physical Systems, internet of things, networks
ERP
1784
1870
1969
TODAY

Data is generated everywhere!

Data in Reality!



- During 2020, 1.7MB of data was created every second by every person.
- Every day 306.4 billion emails are sent.
- In the last two years, 90 percent of the world's data has been created.
- 350 million photos are uploaded to Facebook every day.

1.8 Trillion GB 2011

2.8 Trillion GB 2012

40 Trillion GB 2020

74 Trillion GB 2021

“information age”

production of electronic data

How much your average cell phone data/month? 2GB? 1GB?

Super Huge Databases

YouTube

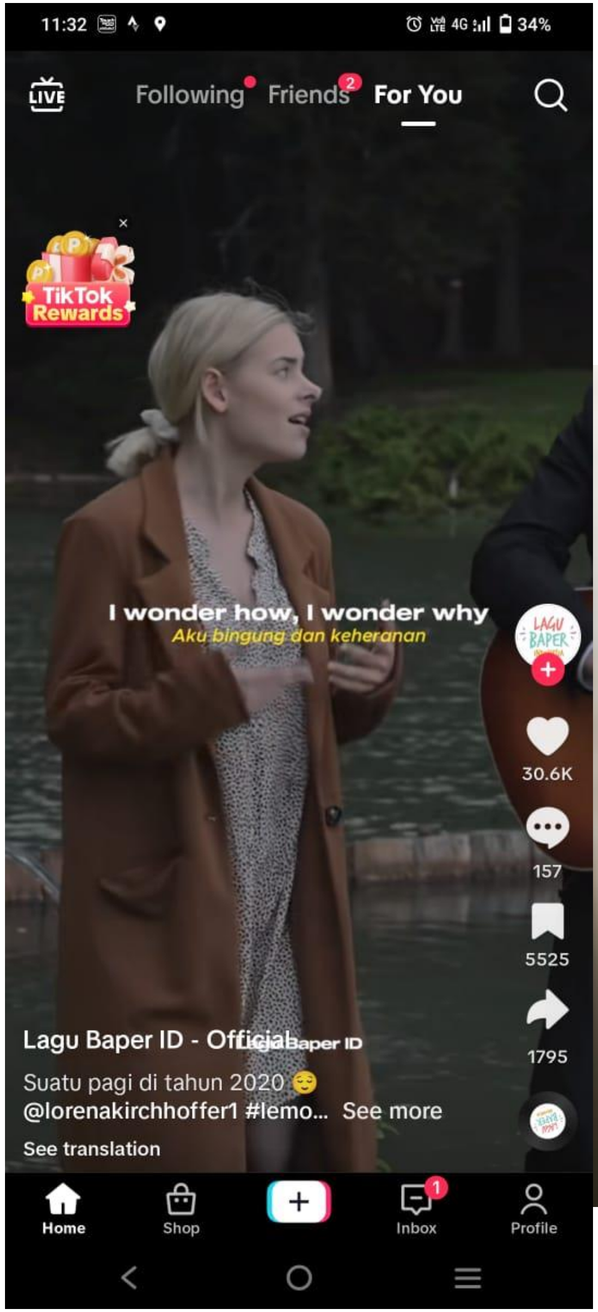
- 100 million videos watched per day
- 65,000 videos added each day
- 60% of all videos watched online
- At least 45 terabytes of videos

- 91 million searches per day
- accounts for 50% of all internet searches
- Virtual profiles of countless number of users
- In terms of internet databases, Google is king.

Google

amazon.com

- 59 million active customers
- More than 42 terabytes of data
- more than 250,000 full text books available online, allows users to comment and interact on virtually every page of the website



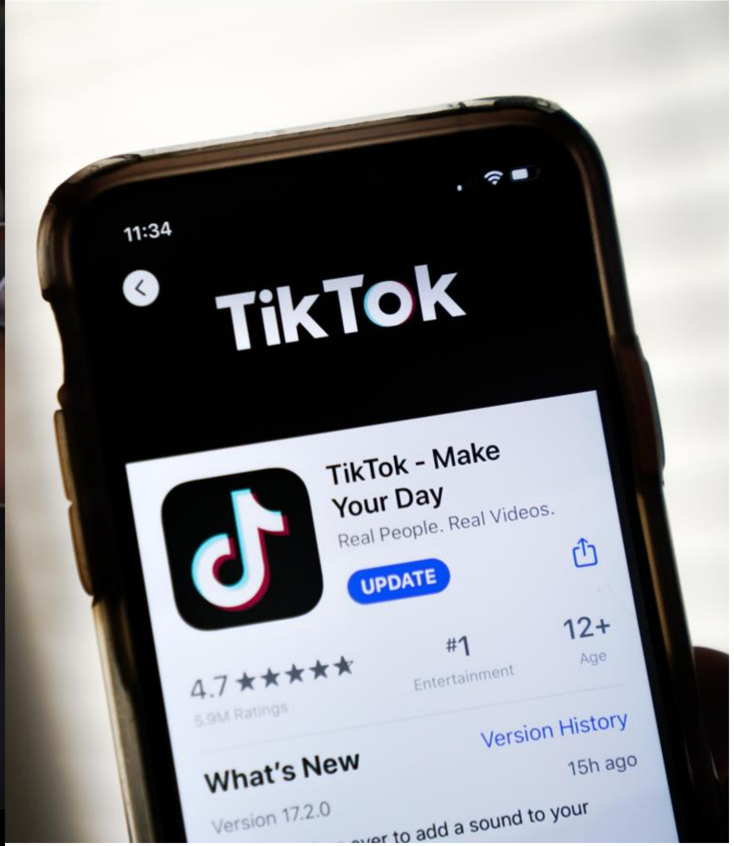
I wonder how, I wonder why

Aku bingung dan keheranan

Lagu Baper ID - Official

Suatu pagi di tahun 2020 😌
@lorenakirchhoffer1 #lemo... See more

See translation



https://www.usesignhouse.com/blog/tiktok-stats

TikTok

45.26

million 2022

2021
17,100
posted per min

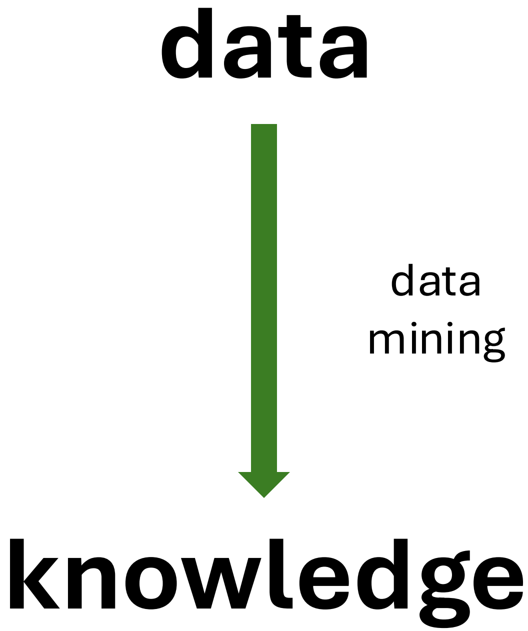
2021
16,322
posted per min

2021
272
posted per
SECOND

Videos

What is data mining?

The **process** of discovering **patterns** and extracting **valuable information** from large datasets using **statistical** & **computational techniques** .



data

data mining

knowledge

What data mining can **offer**?

from data

insight

hidden knowledge

Improve decision making

how is the **knowledge** looks?

Knowledge from data mining

1 Discovering Patterns and Trends

1. Correlations Identification
2. Uncovering hidden patterns
3. Detecting trends
4. Customer segmentation
5. Market basket analysis
6. Sentiment Analysis

2 Making Predictions

1. Predictive modelling.
2. Classification modelling
3. Fraud detection

3 Optimize Processes & Competitive Advantage

1. Supply chain optimization.
2. Market analysis
3. Risk assessment
4. Personalized marketing

Better decision making

Data Mining Task

Descriptive analytics

The most basic type of analytics, and it involves summarizing and describing data.

It is used to answer questions like "What happened?" and "What is happening now?"

Descriptive analytics includes techniques like data visualization, summarization, and clustering

Predictive analytics

Used to make predictions about future events based on past data.

It is used to answer questions like "What is likely to happen in the future?"

Predictive analytics includes techniques like regression analysis, time series analysis, and machine learning algorithms.

Prescriptive analytics

The most advanced type of analytics, and it involves using data to recommend actions or decisions.

It is used to answer questions like "What should we do?" and "What is the best course of action?"

Prescriptive analytics includes techniques like optimization algorithms and decision trees.

Data Mining Task

Descriptive Data Mining

1. **Summarization:** Creating a concise overview of the data, often using statistical measures and visualizations.
2. **Association Rule Mining:** Discovering relationships between items or events (e.g., "people who bought X also bought Y").
3. **Clustering:** Grouping similar data points together without predefined labels.
4. **Outlier Detection:** Identifying data points that significantly deviate from the norm.

Predictive Data Mining

1. **Classification:** Assigning data points to predefined categories or classes.
2. **Regression:** Predicting numerical values based on input variables.
3. **Time Series Analysis:** Analyzing data points collected over time to identify trends and patterns.

Prescriptive Data Mining

1. **Optimization:** Finding the best possible solution to a problem with given constraints.
2. **Simulation:** Creating virtual models to test different scenarios and their outcomes.
3. **Recommendation Systems:** Suggesting items or actions based on user preferences and behavior.

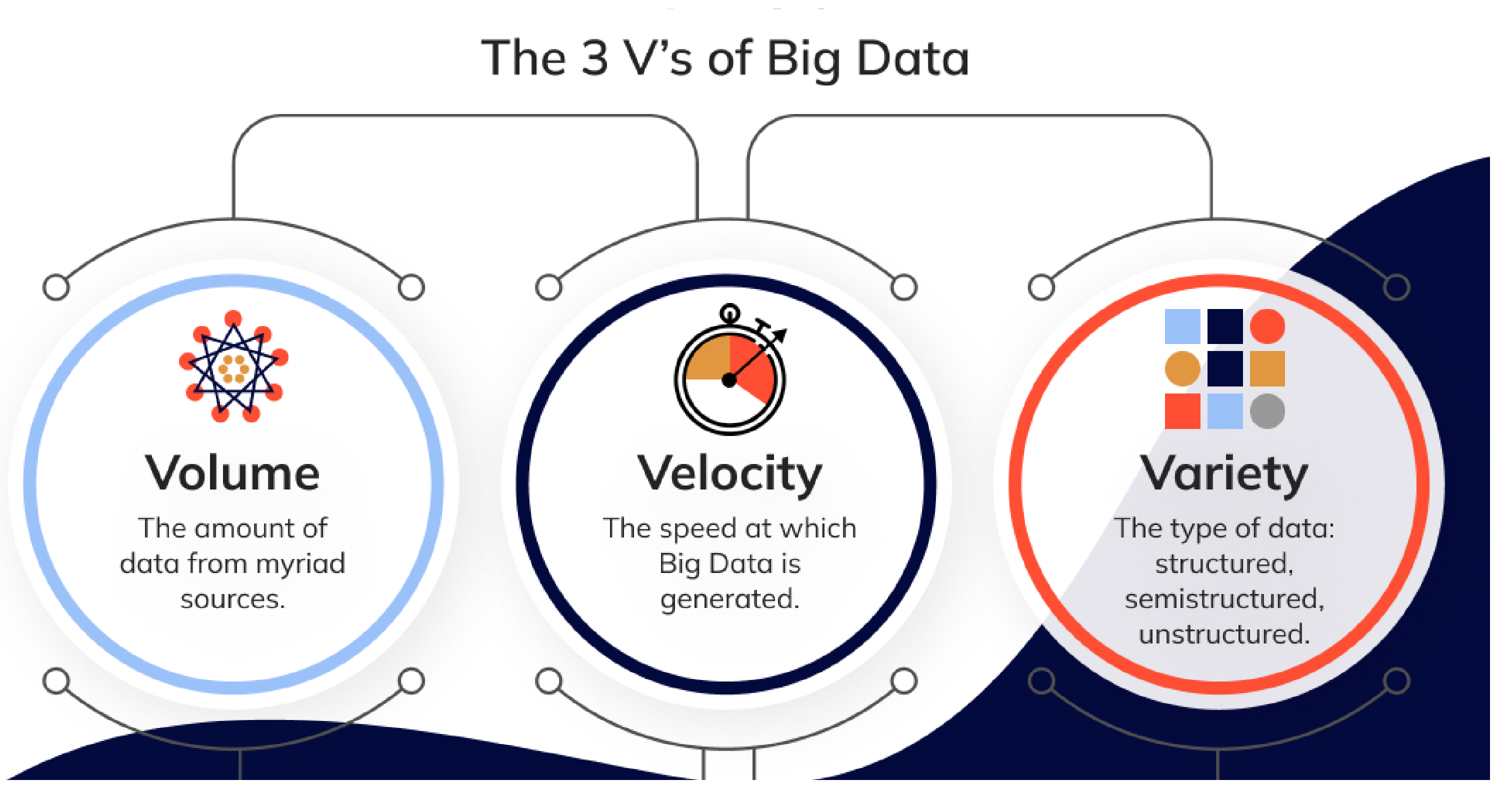
Why data mining **now?**

Big Data

phenomena

Extremely **large** that are **complex** dataset

“3V”


The 3 V’s of Big Data
Volume
The amount of data from myriad sources.
Velocity
The speed at which Big Data is generated.
Variety
The type of data: structured, semistructured, unstructured.

https://intellisoft.io/big-data-security-intelligence-what-you-need-to-know/

How big data & data mining is driving the business?

Competitive Advantage

- Informed decision making
- Customer Understanding
- Risk Management

Operational Efficiency

- Process optimization
- Cost reduction
- Fraud prevention

Innovation

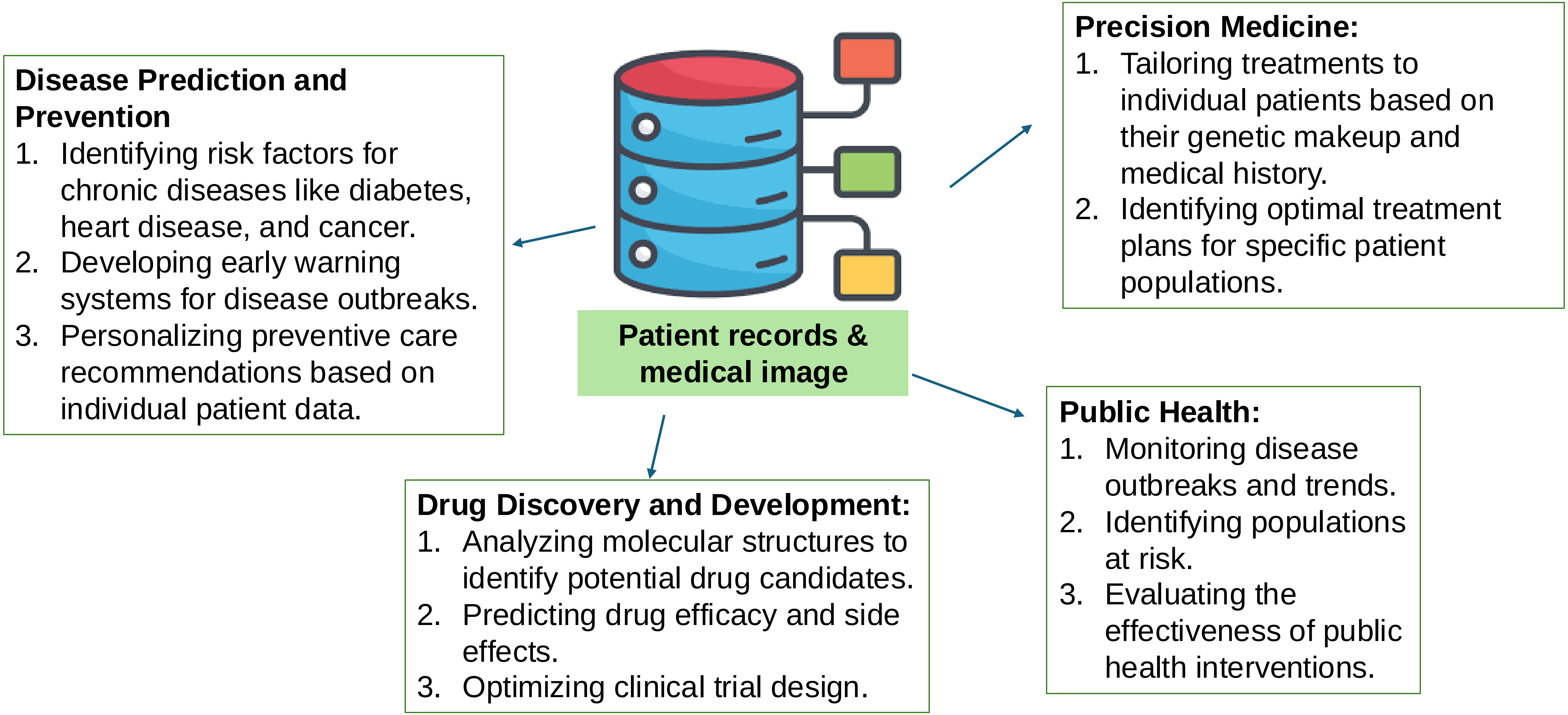
- New product development
- Personalized experience

Data Mining in Health Care

Generates vast amounts of data from **patient records** to **medical images**

It be harnessed to **improve patient outcomes**, **optimize operations**, and drive medical **research**.

Key Applications of Data Mining in Healthcare



Patient records & medical image

Disease Prediction and Prevention

1. Identifying risk factors for chronic diseases like diabetes, heart disease, and cancer.
2. Developing early warning systems for disease outbreaks.
3. Personalizing preventive care recommendations based on individual patient data.

Precision Medicine:

1. Tailoring treatments to individual patients based on their genetic makeup and medical history.
2. Identifying optimal treatment plans for specific patient populations.

Drug Discovery and Development:

1. Analyzing molecular structures to identify potential drug candidates.
2. Predicting drug efficacy and side effects.
3. Optimizing clinical trial design.

Public Health:

1. Monitoring disease outbreaks and trends.
2. Identifying populations at risk.
3. Evaluating the effectiveness of public health interventions.

Key Applications of Data Mining in **Agriculture**

Precision Agriculture

1. Yield Prediction: Analyzing historical data on weather patterns, soil conditions, and crop performance to predict yields.
2. Fertilizer Optimization: Determining optimal fertilizer application based on soil composition, crop needs, and weather conditions.

Crop Monitoring and Disease Detection

1. Image Analysis: Using drones and satellites to monitor crop health, detect pests and diseases, and assess crop growth.
2. Early Warning Systems: Developing models to predict disease outbreaks based on historical data and real-time conditions.

Supply Chain Optimization

1. Demand Forecasting: Predicting crop demand based on market trends, consumer preferences, and economic factors.
2. Inventory Management: Optimizing inventory levels to minimize waste and maximize profits.
3. Logistics Optimization: Improving transportation and distribution efficiency

Key Applications of Data Mining in **Maybank**

Fraud detection: machine learning algorithms is used to identify patterns and anomalies in customer transactions, which can help detect fraudulent activities and prevent financial losses.

Customer segmentation: Data analytics is used to segment its customers based on various factors such as income, age, and spending habits. This helps the bank personalize its services and offers to different customer groups.

Predictive analytics: Maybank uses predictive analytics to forecast customer behavior and anticipate their needs. For example, the bank can predict when a customer is likely to apply for a loan or make a large purchase, and offer targeted promotions accordingly.

Risk management: Maybank uses data analytics to assess and manage risks associated with its lending and investment activities. This helps the bank make informed decisions and minimize potential losses.

Chatbots: Maybank uses natural language processing (NLP) and machine learning to develop chatbots that can handle customer queries and requests. The chatbots are trained on historical customer interactions, which helps them provide accurate and personalized responses

Key Applications of Data Mining in **McDonald's**

36,000 Outlets Worldwide

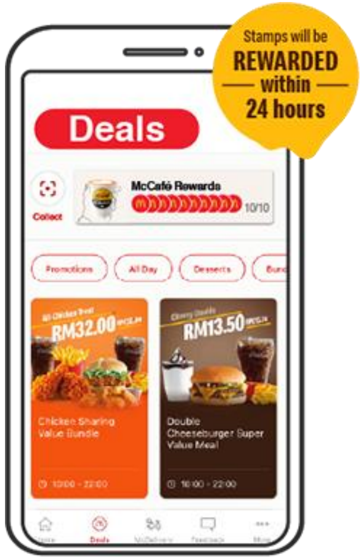
🇺🇸 13, 257 🇲🇾 282

1. Predictive Staffing: McDonald's uses data analytics to predict how many employees they need to staff their restaurants based on historical sales data, time of day, and other factors. This helps them to reduce wait times, improve customer satisfaction, and optimize labor costs.

2. Menu Optimization: McDonald's uses data analytics to analyze customer preferences and optimize their menu by adding or removing items based on popularity and profitability. This helps them to increase sales and improve customer loyalty.

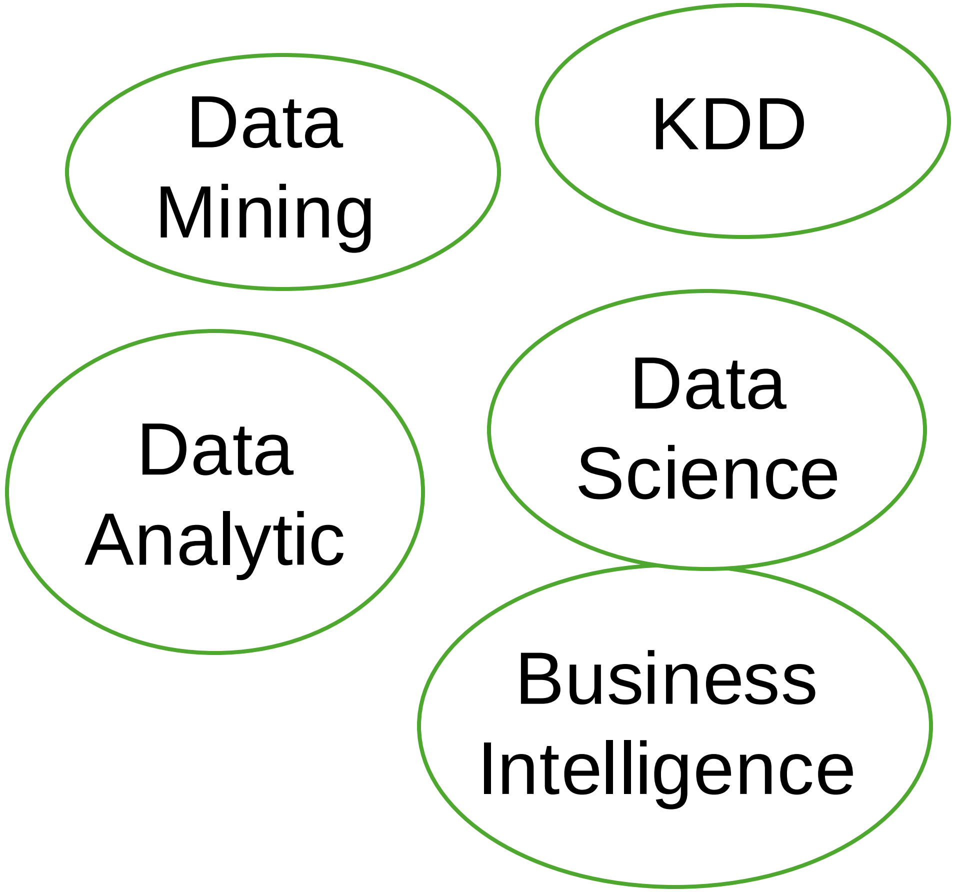
3. Real-time Sales Monitoring: McDonald's uses data analytics to monitor sales in real-time and identify issues such as low sales or supply chain disruptions. This helps them to respond quickly and prevent any negative impact on their business.

4. Customer Analytics: McDonald's uses data analytics to analyze customer data and gain insights into their behavior, preferences, and demographics. This helps them to personalize their marketing campaigns, improve their menu offerings, and enhance the overall customer experience



Stamps will be REWARDED within 24 hours

Key Concepts and Terminology



Data Mining

KDD

Data Analytic

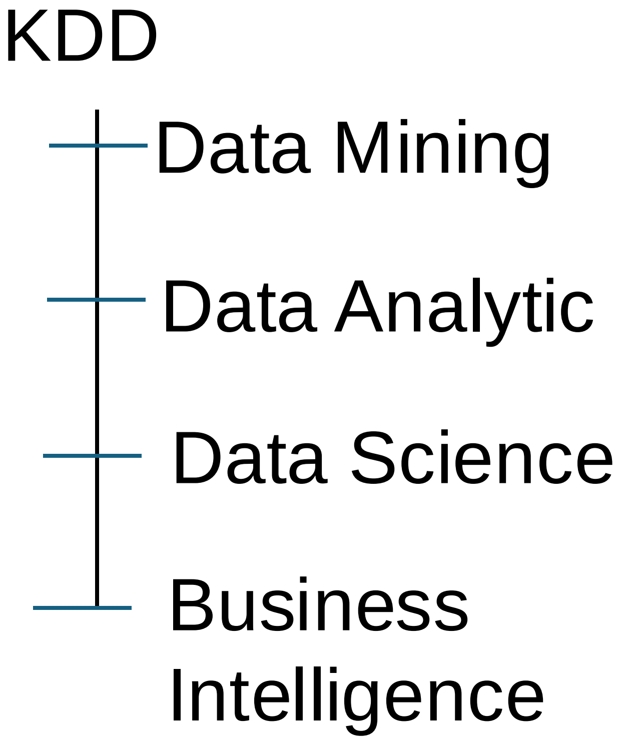
Data Science

Business Intelligence

Shared similarity

Extracting insights/knowledge from databases

Key Concepts and Terminology



KDD

Data Mining

Data Analytic

Data Science

Business Intelligence

The depth, breadth, and application of those insights are difference…

1. KDD (Knowledge Discovery in Databases): A structured framework for the **entire knowledge discovery process**

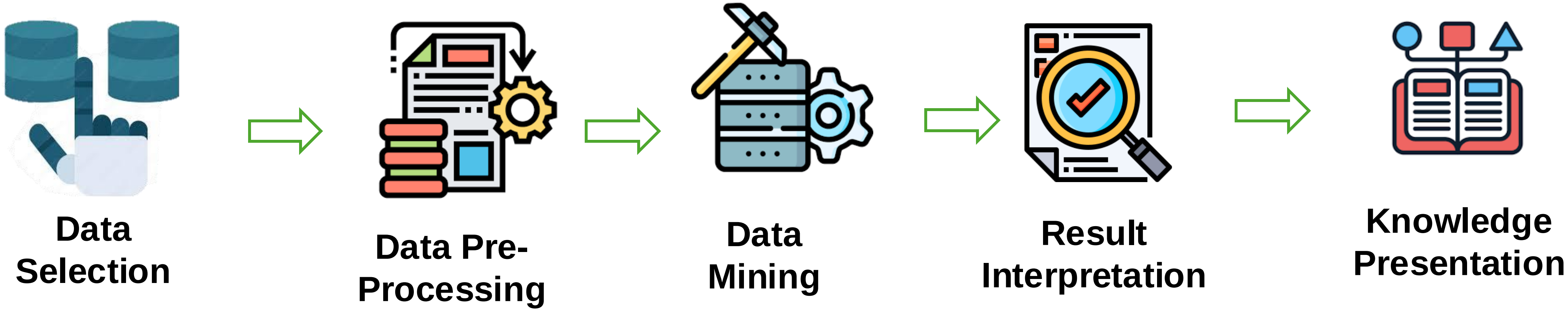
2. Data Mining: A core component of KDD, focusing on **discovering patterns** within data.

3. Data Analytics: A broader field that builds upon data mining to extract insights & uses those patterns to make informed decisions

4. Data Science: A broader field encompassing data mining, data analytics, and additional methodologies.

5. Business intelligence: translates insights into actionable information for **business decision-making**.

Data Mining Process



Data Selection

Data Pre-Processing

Data Mining

Result Interpretation

Knowledge Presentation

Data Mining Process

Organization data

survey

Data Selection

Involves: Identifying relevant data sources for the specific problem or question at hand.

Goal: To gather data that is essential for the analysis and avoid unnecessary data.

Example: For a customer churn analysis, selecting customer demographics, usage patterns, and billing information.

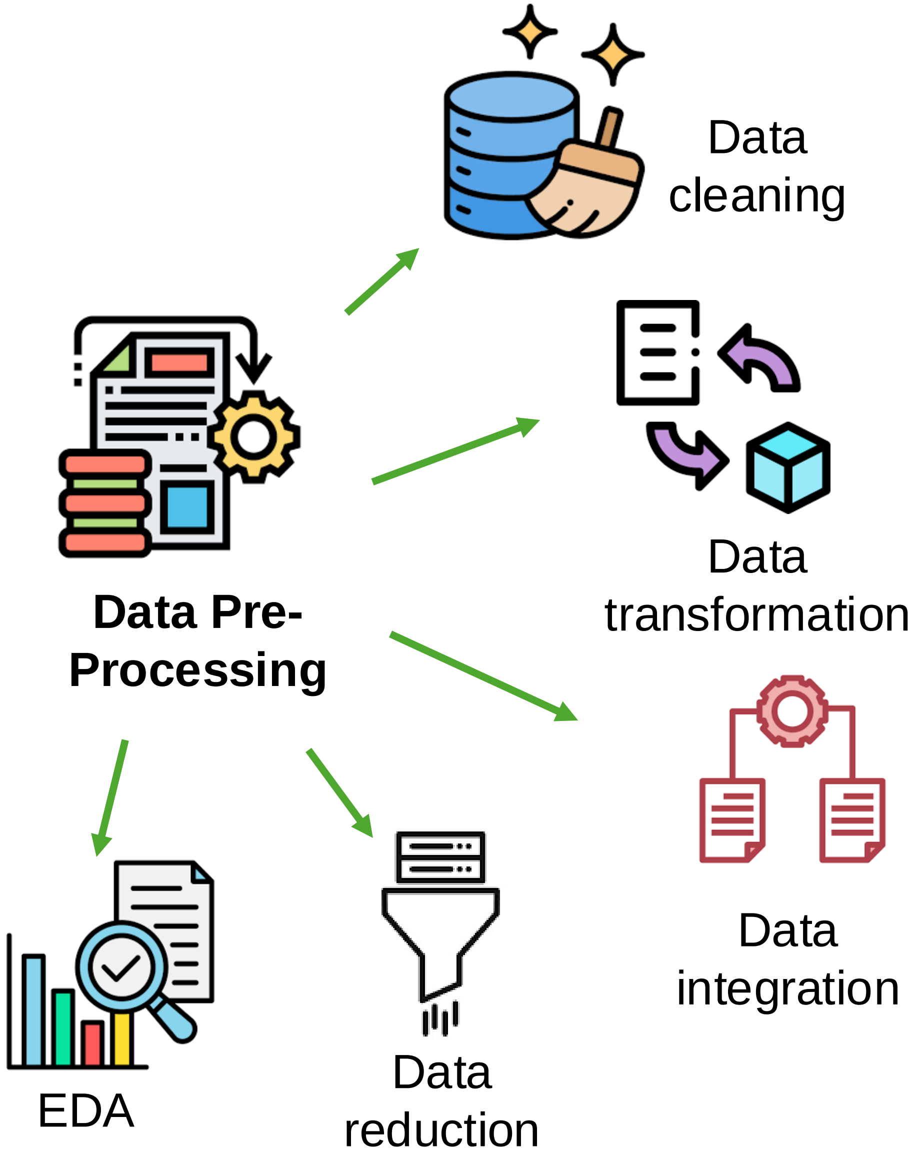
kaggle

UCI Machine Learning Repository

Public- shared database

Web data scraping

Data Mining Process



Data cleaning

Data transformation

Data Pre-Processing

Data integration

EDA

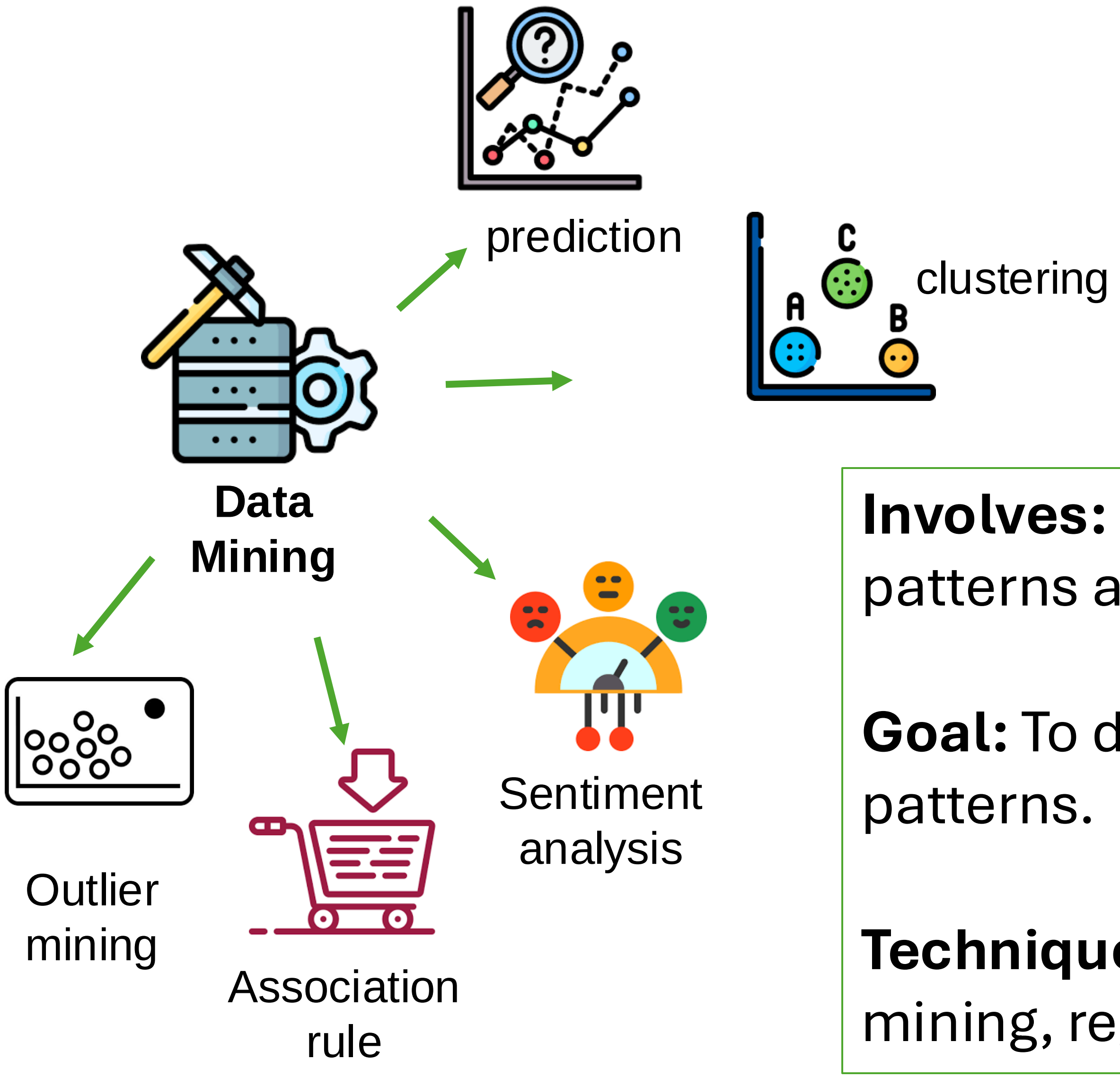
Data reduction

Involves: Cleaning, transforming, and preparing data for analysis.

Goal: To ensure data quality and consistency.

Tasks: Handling missing values, outliers, inconsistencies, and data normalization.

Data Mining Process



prediction

clustering

Data Mining

Sentiment analysis

Outlier mining

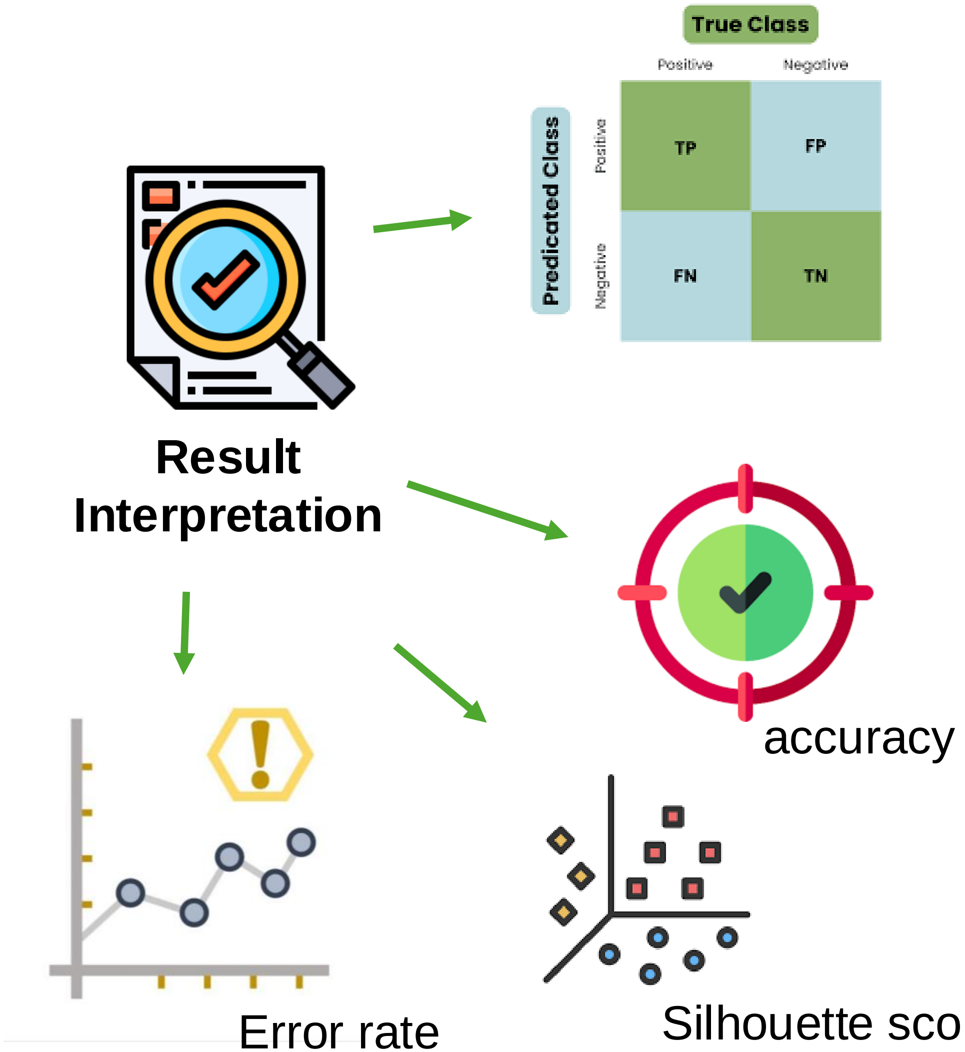
Association rule

Involves: Applying **algorithms** and **techniques** to extract patterns and knowledge from the prepared data.

Goal: To discover hidden relationships, trends, and patterns.

Techniques: Clustering, classification, association rule mining, regression, etc.

Data Mining Process



Confusion metric

		True Class	
		Positive	Negative
Predicated Class	Positive	TP	FP
	Negative	FN	TN

Result Interpretation

accuracy

Error rate

Silhouette score

Involves: Analyzing the discovered patterns and understanding their implications.

Goal: To derive meaningful insights from the data mining results.

Tasks: Identifying significant patterns, evaluating patterns against business objectives.

Data Mining Process

Knowledge Presentation

Involves: Communicating the findings to stakeholders in a clear and understandable manner.

Goal: To share insights and support decision-making.

Methods: Reports, visualizations, dashboards.

Report

prototype

dashboard

Challenges and Issues in Data Mining

Data Quality Issues

- **Noise and Inconsistency:** Data often contains errors, outliers, and inconsistencies that can affect the accuracy of results.
- **Missing Values:** Incomplete data can lead to biased results or reduced model performance.
- **Data Integration:** Combining data from multiple sources can be challenging due to inconsistencies in formats and definitions.

Challenges and Issues in Data Mining

Data Privacy and Security

- **Sensitive Information:** Data mining often involves handling sensitive personal information, raising privacy concerns.
- **Data Breaches:** The risk of data breaches and unauthorized access to sensitive information is significant.
- **Compliance:** Adhering to data protection regulations (e.g., GDPR, CCPA) can be complex.

Challenges and Issues in Data Mining

Scalability and Efficiency

- **Big Data:** Handling massive datasets requires efficient algorithms and distributed computing resources.
- **Computational Cost:** Data mining can be computationally expensive, especially for complex models and large datasets.

Challenges and Issues in Data Mining

Interpretability and Explainability

- **Black Box Models:** Some models, like deep learning, can be difficult to interpret, making it challenging to understand the reasons behind predictions.
- **Complex Patterns:** Discovering complex patterns might not always lead to easily understandable explanations.

Challenges and Issues in Data Mining

Overfitting and Underfitting

- **Overfitting:** Models that are too complex may fit the training data too well but perform poorly on new data.
- **Underfitting:** Models that are too simple may not capture the underlying patterns in the data.

Challenges and Issues in Data Mining

Ethical Considerations

- **Bias:** Data mining models can perpetuate existing biases in the data.
- **Fairness:** Ensuring that data mining algorithms treat different groups fairly is crucial.
- **Accountability:** Organizations must be accountable for the outcomes of data mining models.

Tools for Data Mining

Open-Source Tools

- **Python:** A versatile language with libraries like NumPy, Pandas, Scikit-learn, and TensorFlow for data manipulation, analysis, and machine learning.
- **R:** Specifically designed for statistical computing and graphics, with packages for data mining tasks.
- **Weka:** A Java-based platform with a graphical user interface for various data mining algorithms.
- **RapidMiner:** A visual workflow environment for data mining tasks.
- **Orange:** A Python-based data mining and machine learning toolkit with a visual interface.
- **KNIME:** An open-source data analytics platform with a workflow-based interface.

Python for data mining.

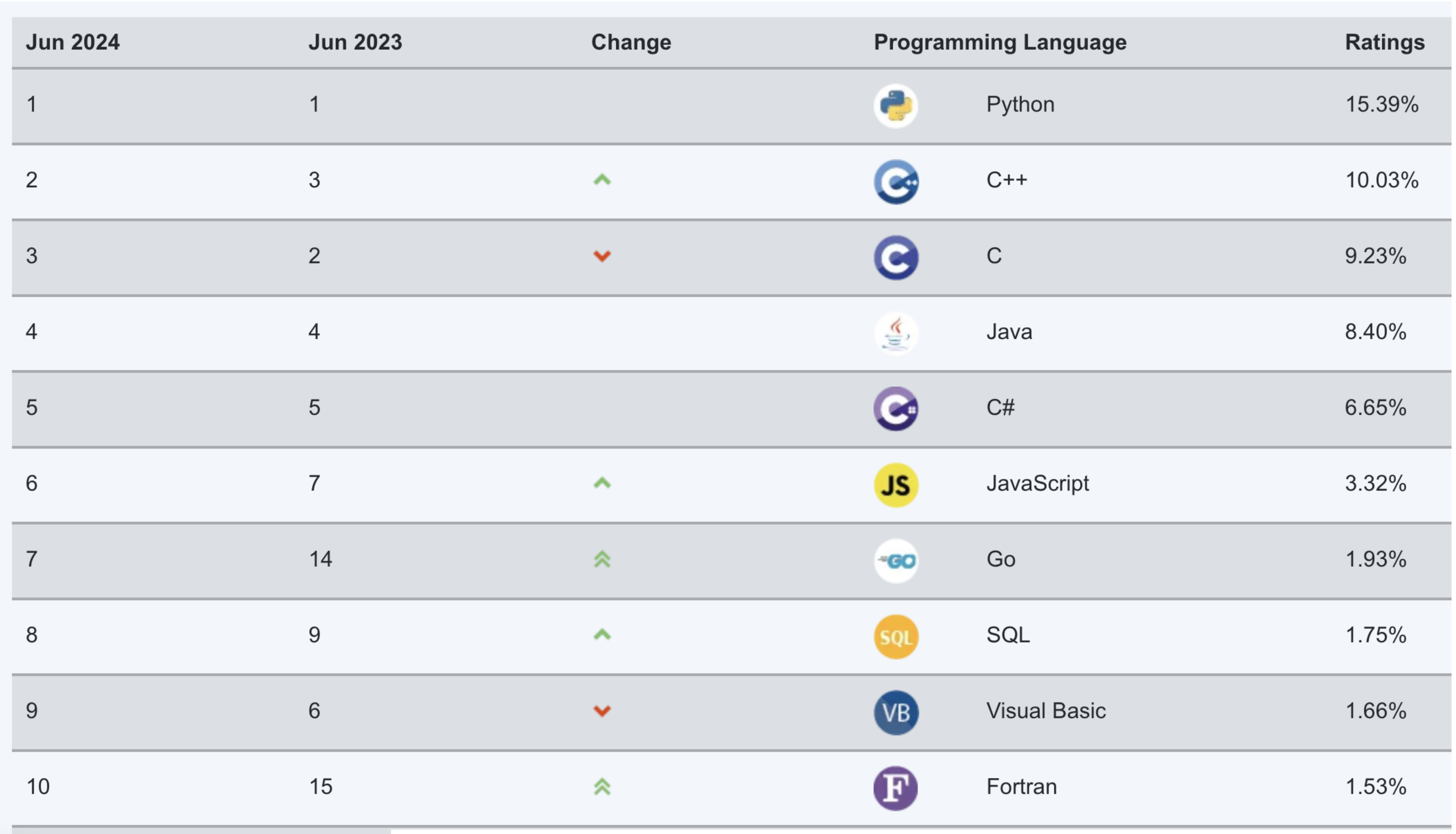
TIOBE
the software quality company

Popularity of Python

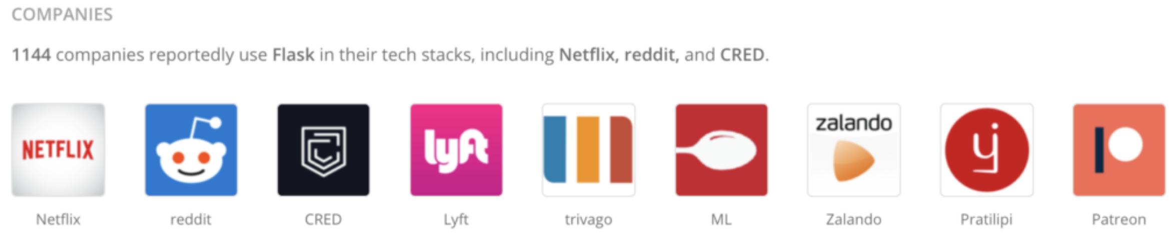
Readability
Versatility of Use
Efficient & Reliable
Active & Supportive Community
Library & Framework

Beginner

Expert

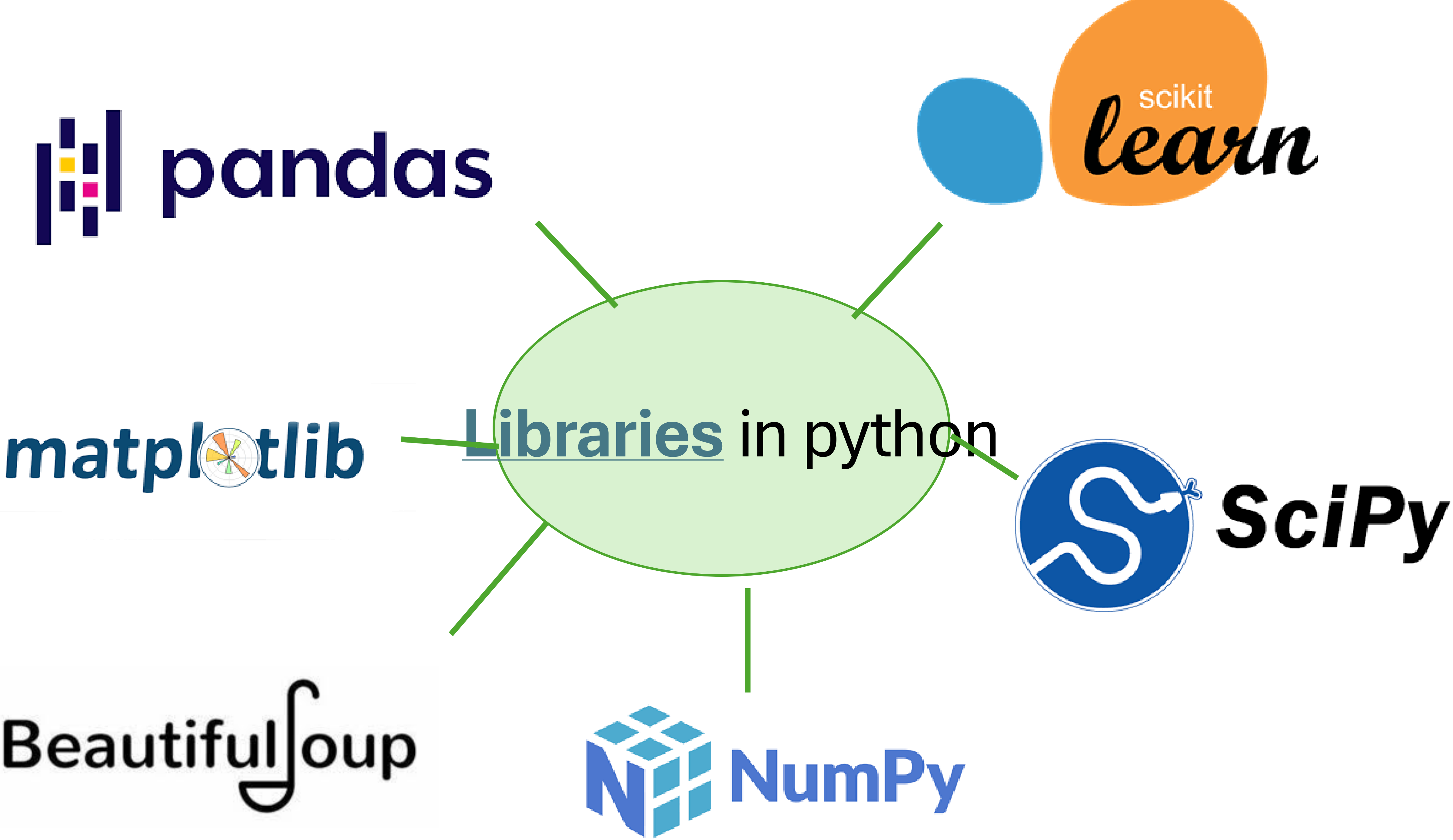


Jun 2024	Jun 2023	Change	Programming Language	Ratings
1	1		Python	15.39%
2	3	⌃	C++	10.03%
3	2	⌄	C	9.23%
4	4		Java	8.40%
5	5		C#	6.65%
6	7	⌃	JavaScript	3.32%
7	14	⌃⌃	Go	1.93%
8	9	⌃	SQL	1.75%
9	6	⌄	Visual Basic	1.66%
10	15	⌃⌃	Fortran	1.53%



COMPANIES

1144 companies reportedly use Flask in their tech stacks, including Netflix, reddit, and CRED.

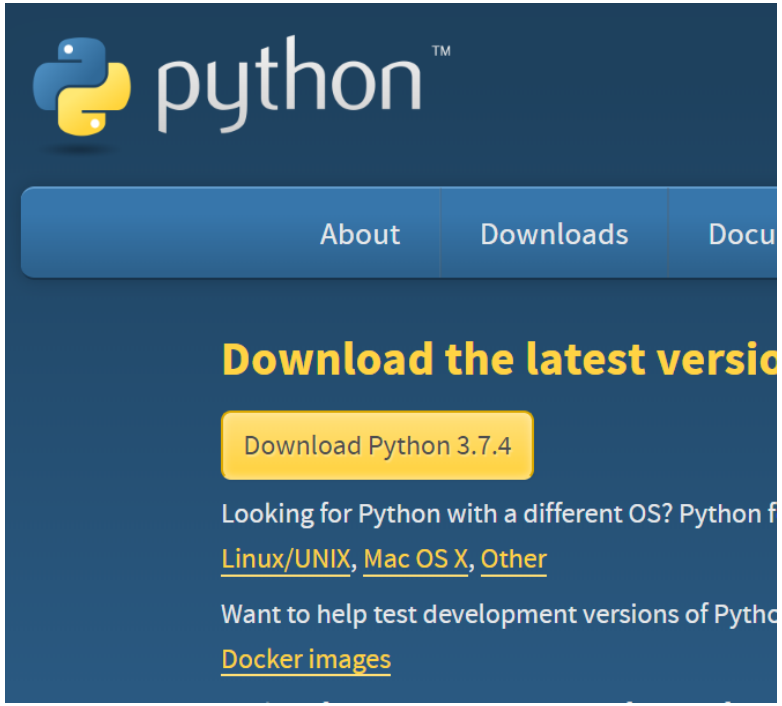


Libraries in python

Two requirements you need to prepare to code in python

First…

- Download and Install **Python** https://www.python.org/downloads/



Download the latest versio

Download Python 3.7.4

Looking for Python with a different OS? Python f

Linux/UNIX, Mac OS X, Other

Want to help test development versions of Pytho

Docker images

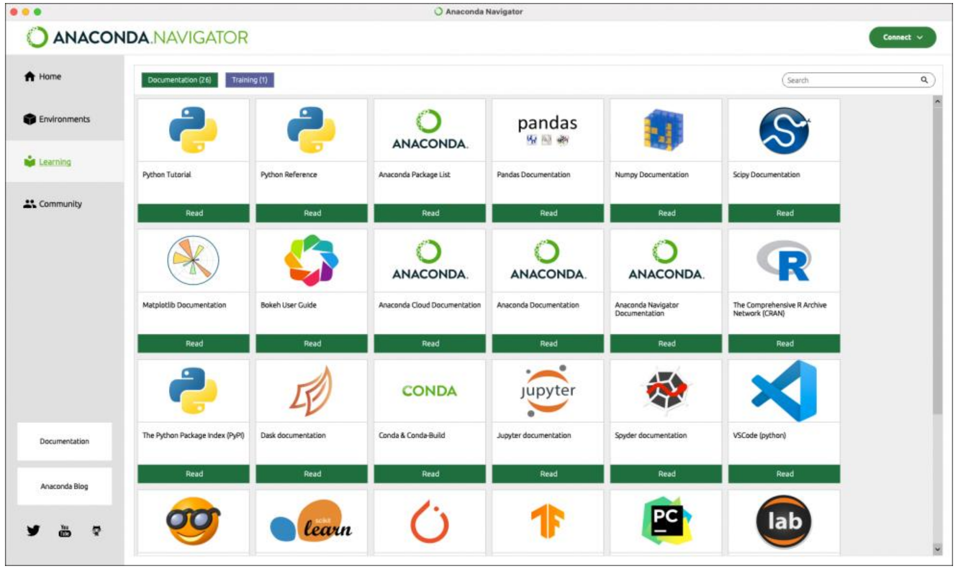
- Code Editor/ Integrated Development Environment (**IDE**)

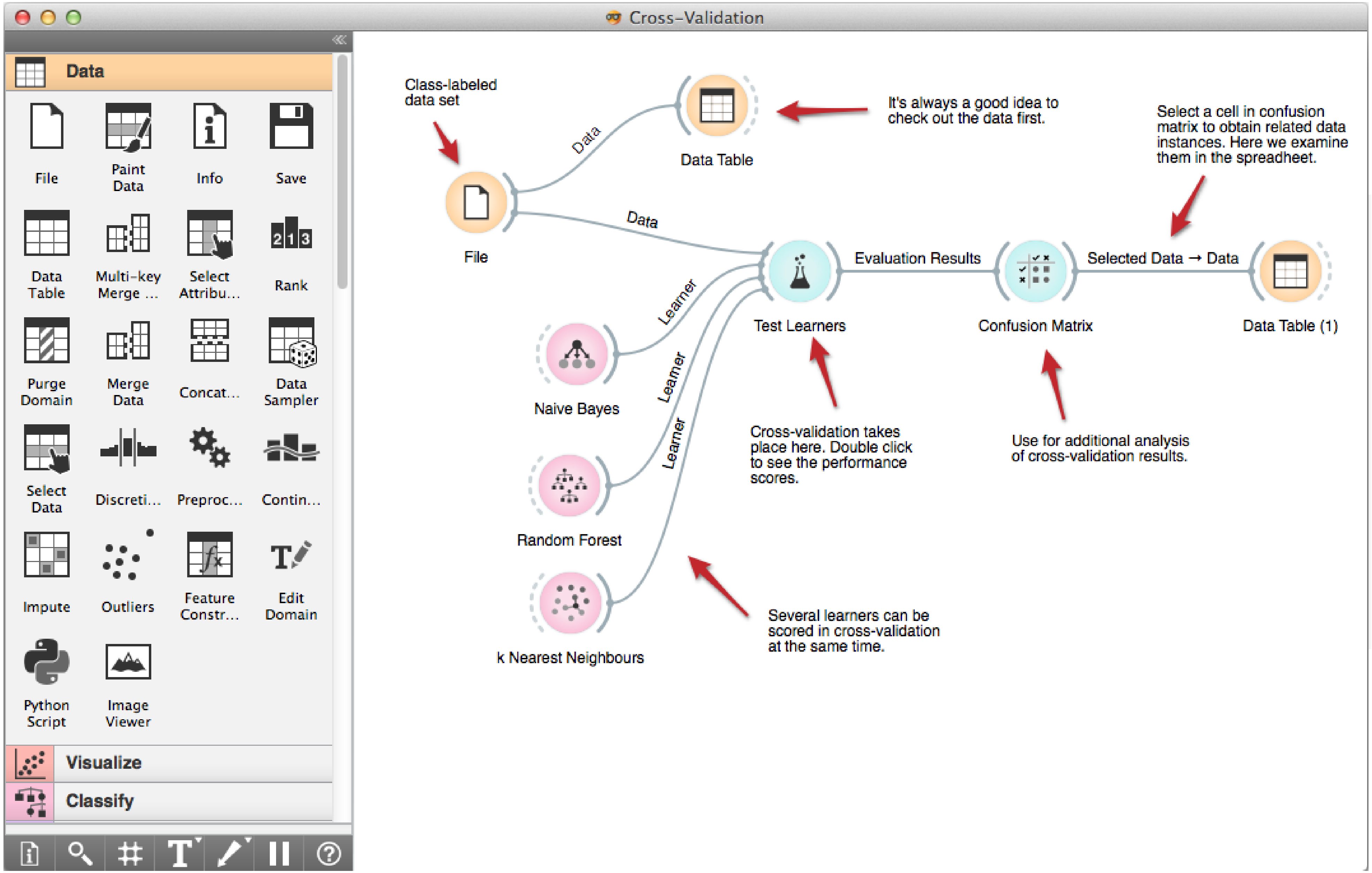
Second… Download and Install **Anaconda** https://www.anaconda.com/download/success

Jupyter Notebook

PyCharm

VS Code





Class-labeled data set

It's always a good idea to check out the data first.

Select a cell in confusion matrix to obtain related data instances. Here we examine them in the spreadsheet.

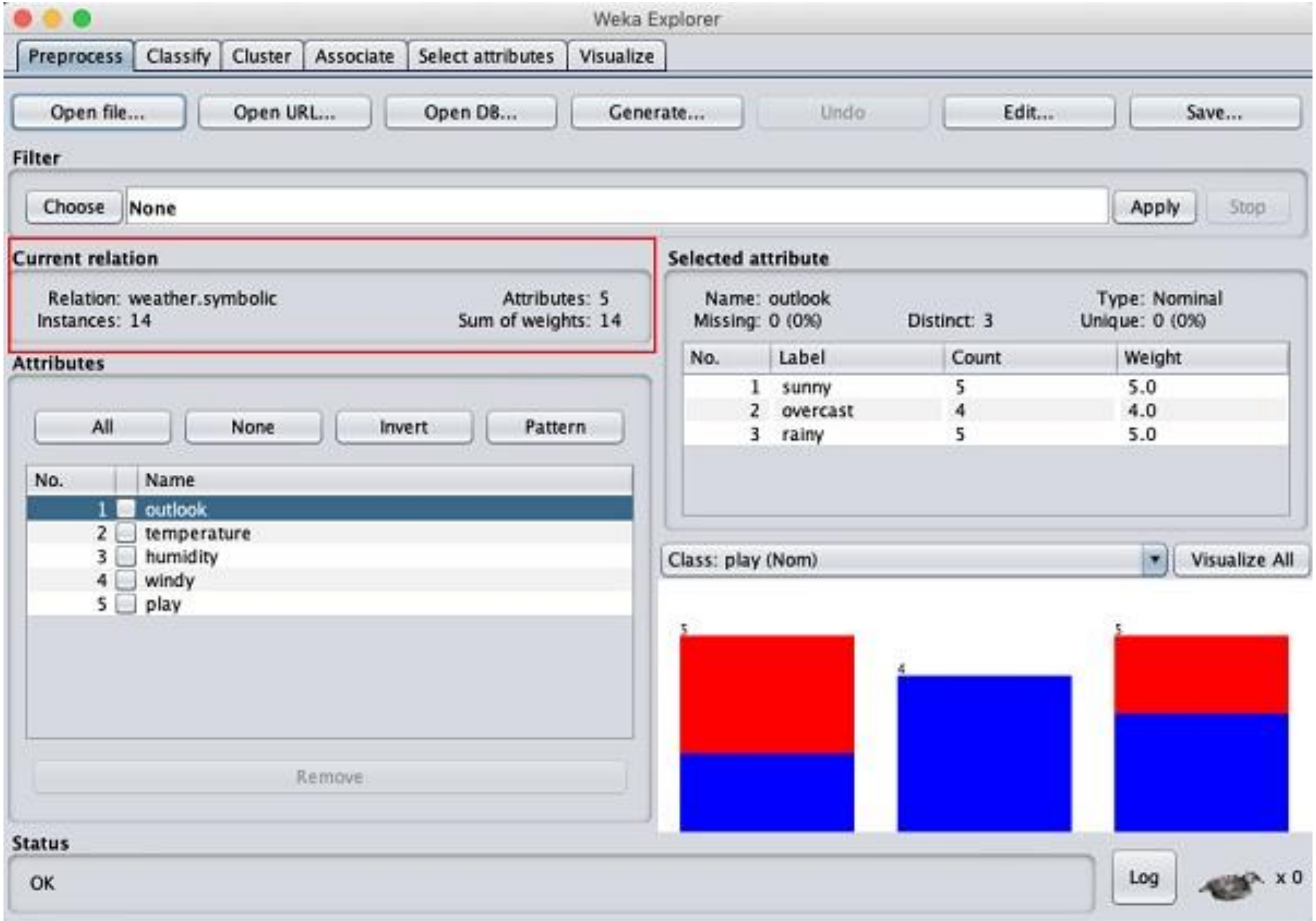
Cross-validation takes place here. Double click to see the performance scores.

Use for additional analysis of cross-validation results.

Several learners can be scored in cross-validation at the same time.

orange
DATA MINING

https://orangedatamining.com/download/



Filter

None

Current relation

Relation: weather.symbolic
Instances: 14
Attributes: 5
Sum of weights: 14

Attributes

No.		Name
1	□	outlook
2	□	temperature
3	□	humidity
4	□	windy
5	□	play

Selected attribute

Name: outlook
Missing: 0 (0%)
Distinct: 3
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	sunny	5	5.0
2	overcast	4	4.0
3	rainy	5	5.0

Class: play (Nom)

Status

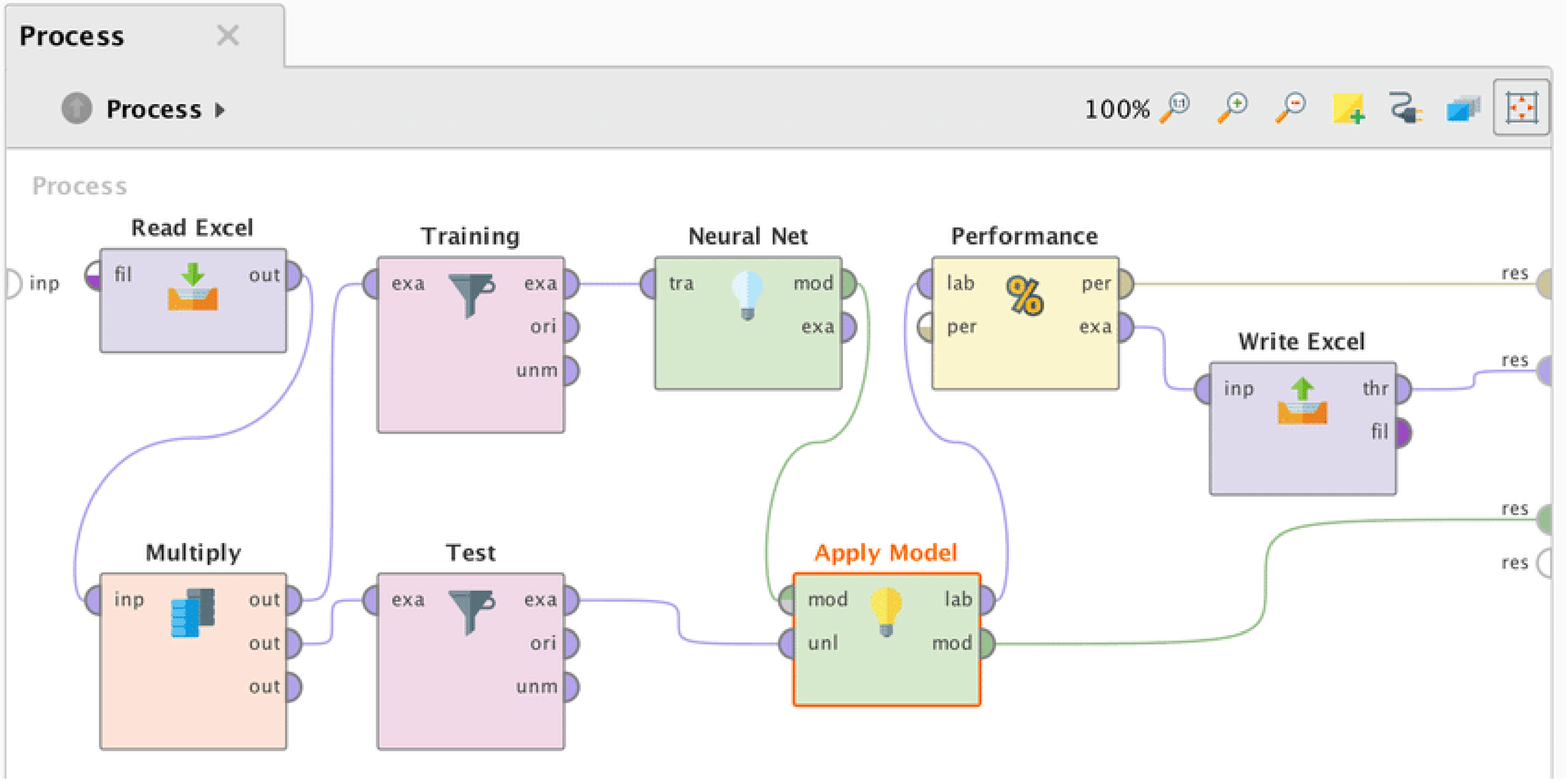
OK

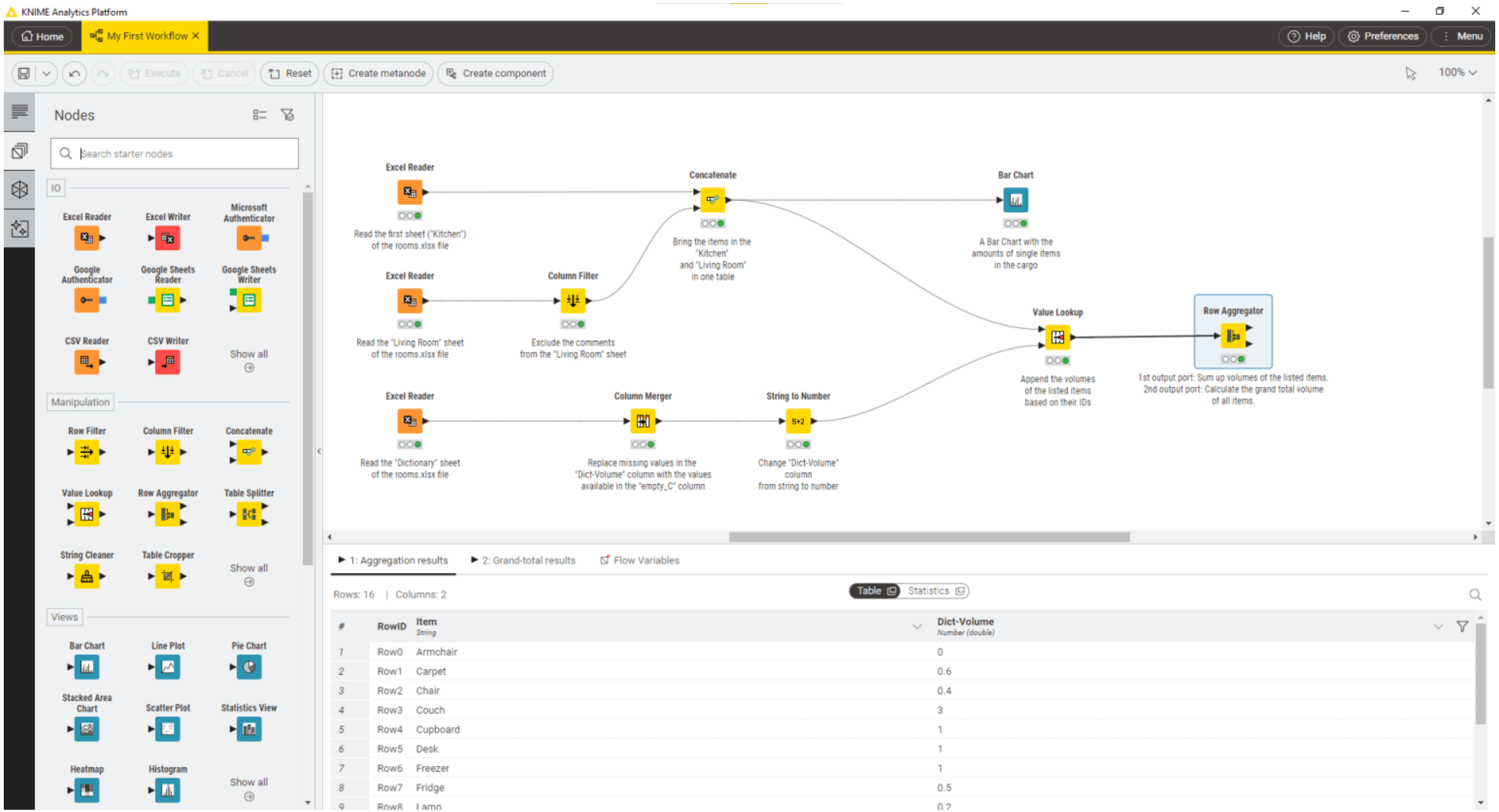
WEKA

https://waikato.github.io/weka-wiki/downloading_weka/

RAPID MINER

https://docs.rapidminer.com/9.9/studio/installation/





Read the first sheet ("Kitchen") of the rooms.xlsx file

Read the "Living Room" sheet of the rooms.xlsx file

Exclude the comments from the "Living Room" sheet

Bring the items in the "Kitchen" and "Living Room" in one table

A Bar Chart with the amounts of single items in the cargo

Read the "Dictionary" sheet of the rooms.xlsx file

Replace missing values in the "Dict-Volume" column with the values available in the "empty_C" column

Change "Dict-Volume" column from string to number

Append the volumes of the listed items based on their IDs

1st output port: Sum up volumes of the listed items.
2nd output port: Calculate the grand total volume of all items.

Rows: 16 | Columns: 2

#	RowID	Item String	Dict-Volume Number (double)
1	Row0	Armchair	0
2	Row1	Carpet	0.6
3	Row2	Chair	0.4
4	Row3	Couch	3
5	Row4	Cupboard	1
6	Row5	Desk	1
7	Row6	Freezer	1
8	Row7	Fridge	0.5
9	Row8	Lamp	0.2

KNIME

https://www.knime.com/downloads

Tools for Data Mining

Commercial Tools

- **IBM SPSS Modeler:** A comprehensive platform for data mining, predictive analytics, and machine learning.
- **SAS Enterprise Miner:** A powerful tool for advanced analytics and data mining.
- **MATLAB:** A high-performance computing language with toolboxes for data analysis and machine learning.

Tools for Data Mining

Cloud-Based Platforms

- **Google Cloud Platform (GCP):** Offers a range of data mining and machine learning services.
- **Amazon Web Services (AWS):** Provides various tools and services for data processing and analysis.
- **Microsoft Azure:** Offers a comprehensive cloud platform for data mining and machine learning.

Considerations When Choosing a Tool

The choice of data mining tool depends on factors like the size of the dataset, the complexity of the analysis, the desired level of programming involvement, and the specific algorithms required

- **Ease of use:** Consider the tool's interface and learning curve.
- **Scalability:** Ensure the tool can handle the size and complexity of your datasets.
- **Functionality:** Check if the tool supports the required data mining algorithms and techniques.
- **Cost:** Evaluate the pricing model and licensing options.
- **Community support:** Access to forums, tutorials, and documentation can be helpful.