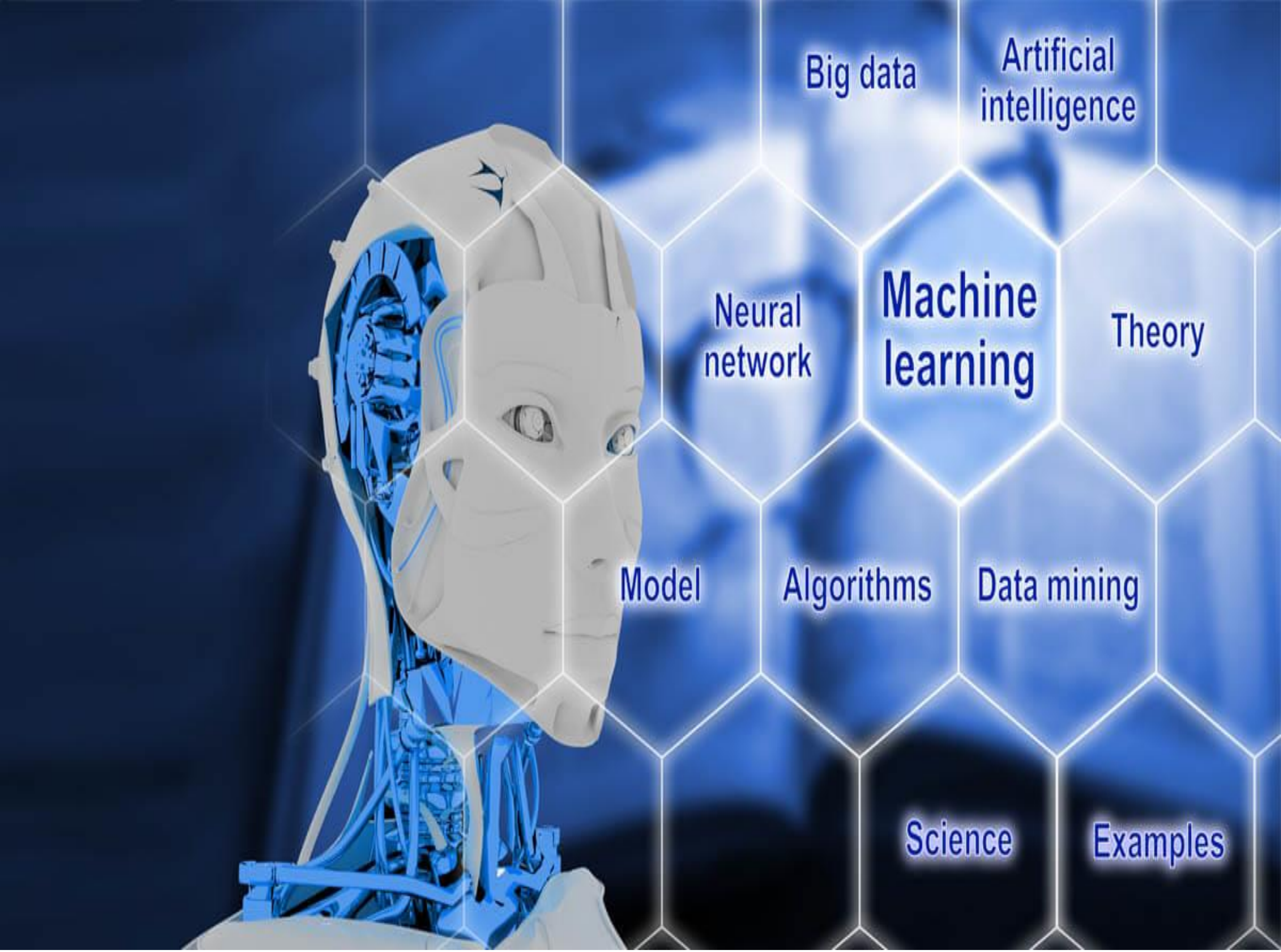
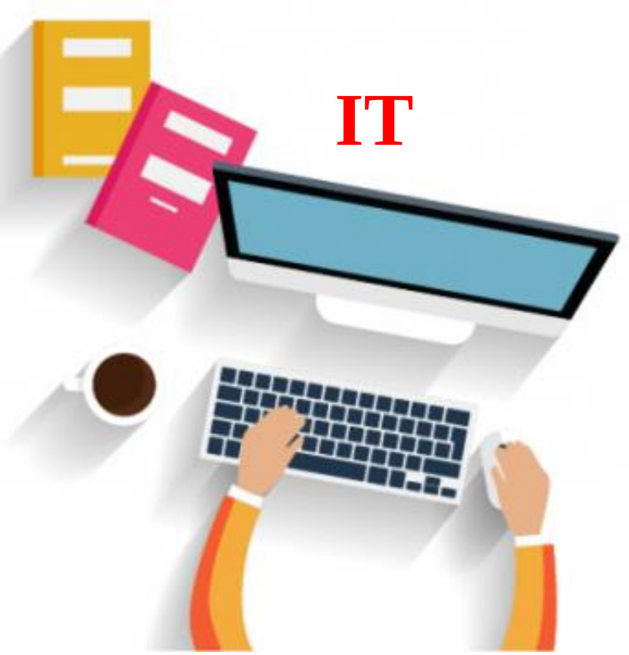

Big data
Artificial intelligence
Neural network
Machine learning
Theory
Model
Algorithms
Data mining
Science
Examples

Data-Sciences

No matter which industry you work in, there is no doubt that data affects your life and work.



IT

the FOOD COMPANY

FASHION

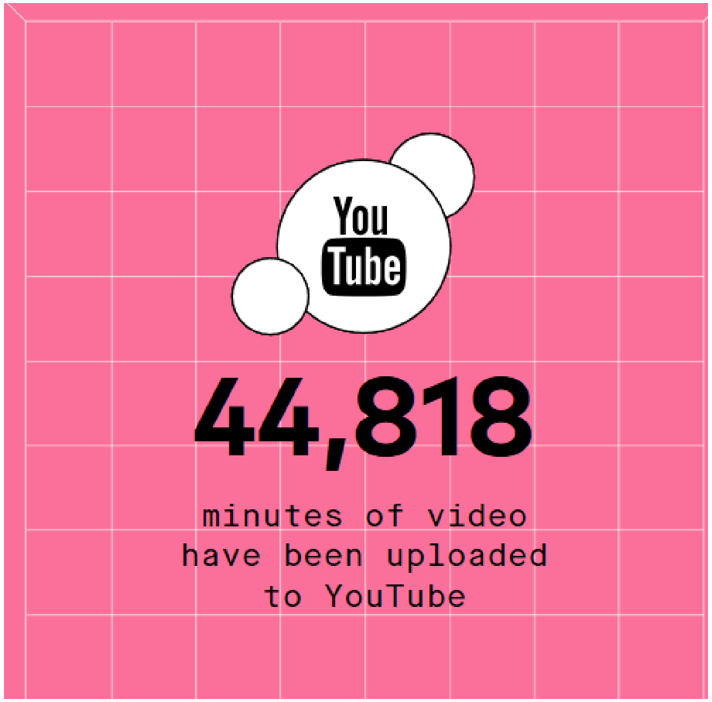
Finance

AGRICULTURE

Automobile Company

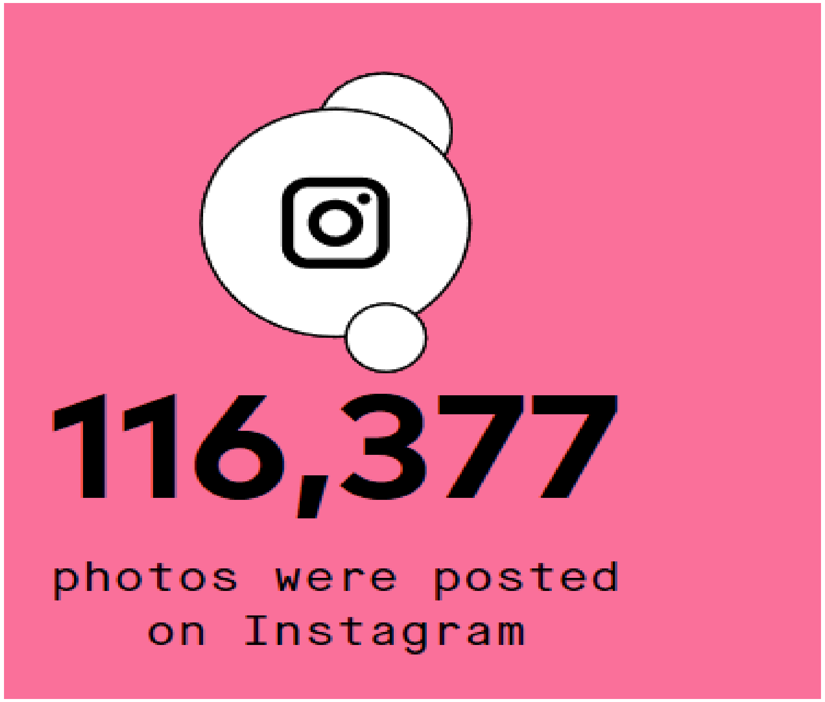
Data

It's gotten so easy to write data, and so cheap to store it, that sometimes companies don't even know what value they can get from that data.



44,818

minutes of video have been uploaded to YouTube

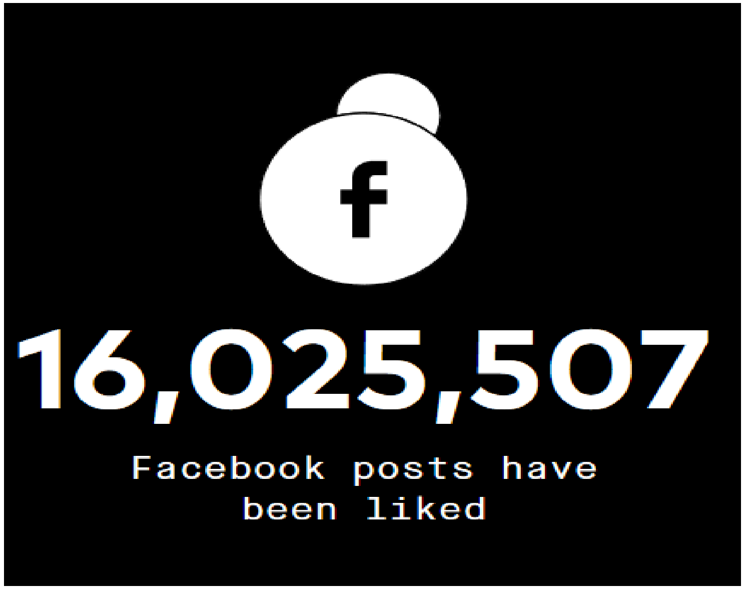


116,377

photos were posted on Instagram

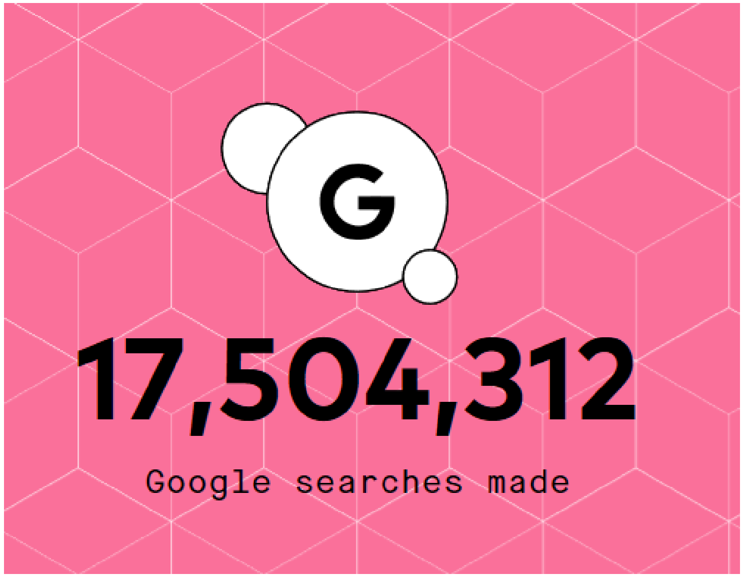
1,132,894

tweets have been tweeted



16,025,507

Facebook posts have been liked



17,504,312

Google searches made

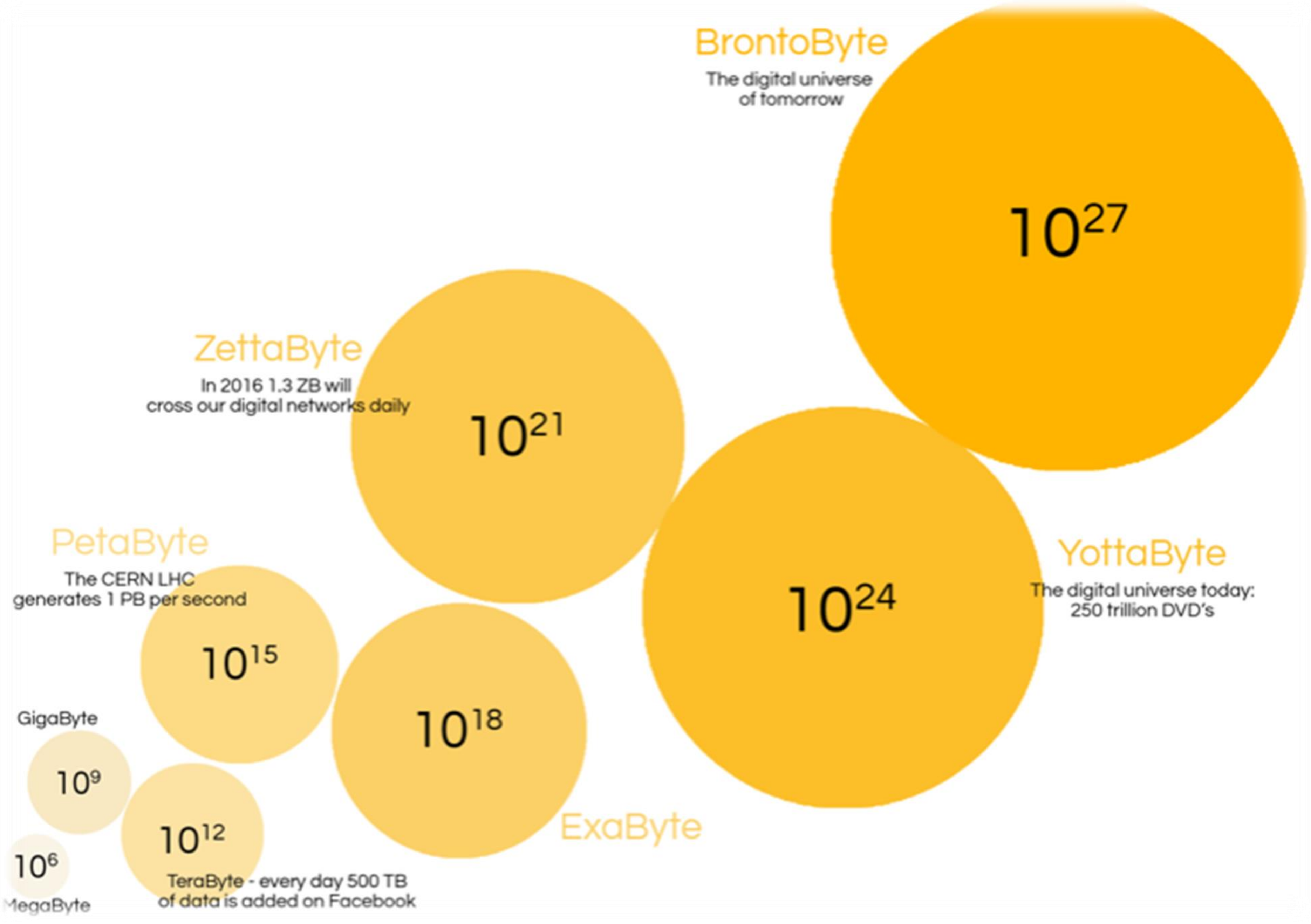
73,862,481

text messages sent

Rich Data But Poor Information

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.

At the moment ,every day 1EB of data is created on the internet.
that is the equivalent of 250 million DVD's.



10^{6}

MegaByte

10^{9}

GigaByte

10^{12}

TeraByte - every day 500 TB
of data is added on Facebook

10^{15}

PetaByte

The CERN LHC
generates 1 PB per second

10^{18}

ExaByte

10^{21}

ZettaByte

In 2016 1.3 ZB will
cross our digital networks daily

10^{24}

YottaByte

The digital universe today:
250 trillion DVD's

10^{27}

BrontoByte

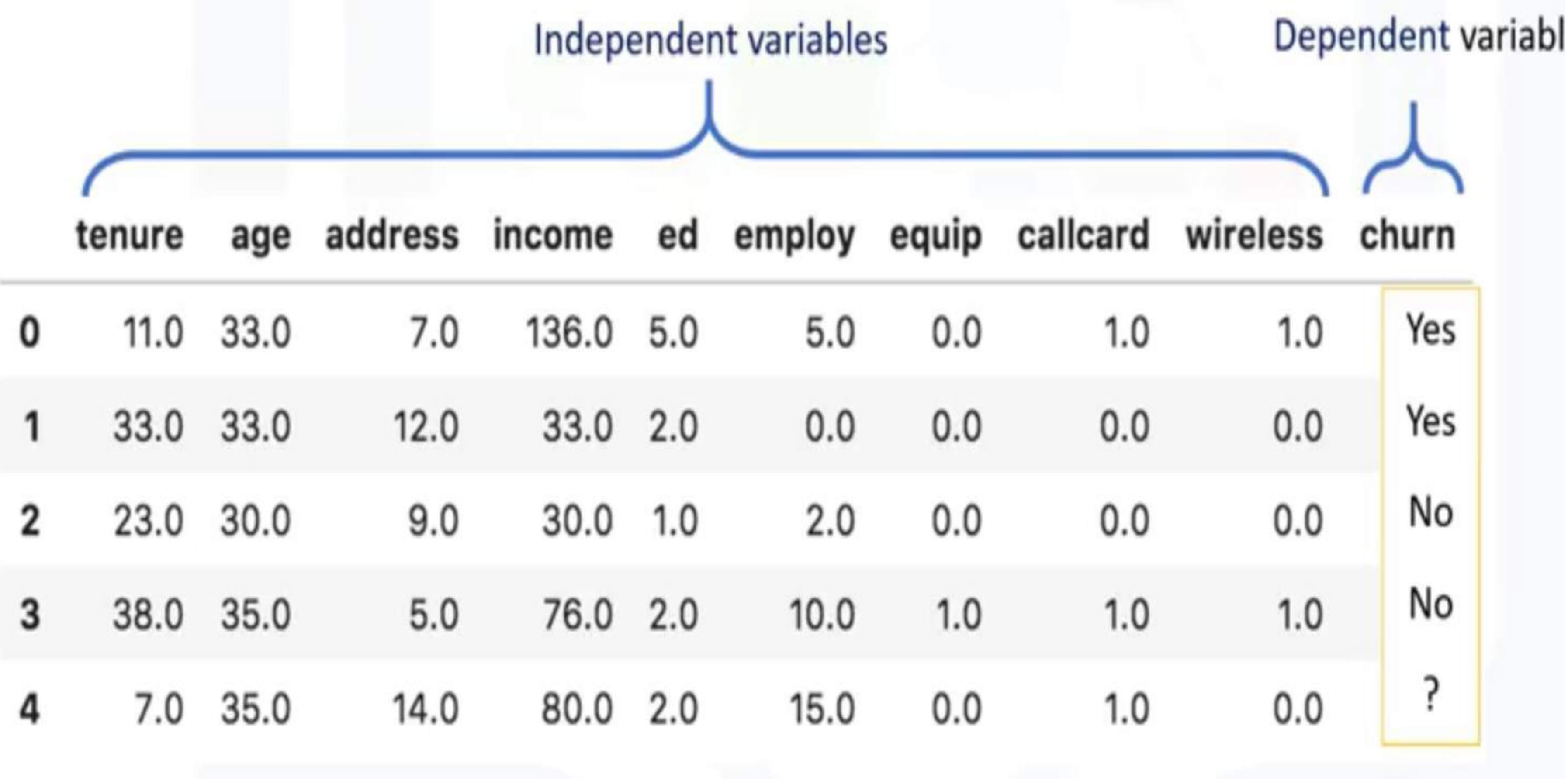
The digital universe
of tomorrow

Mobile

Social Media

IOT/Sensors

Data-set

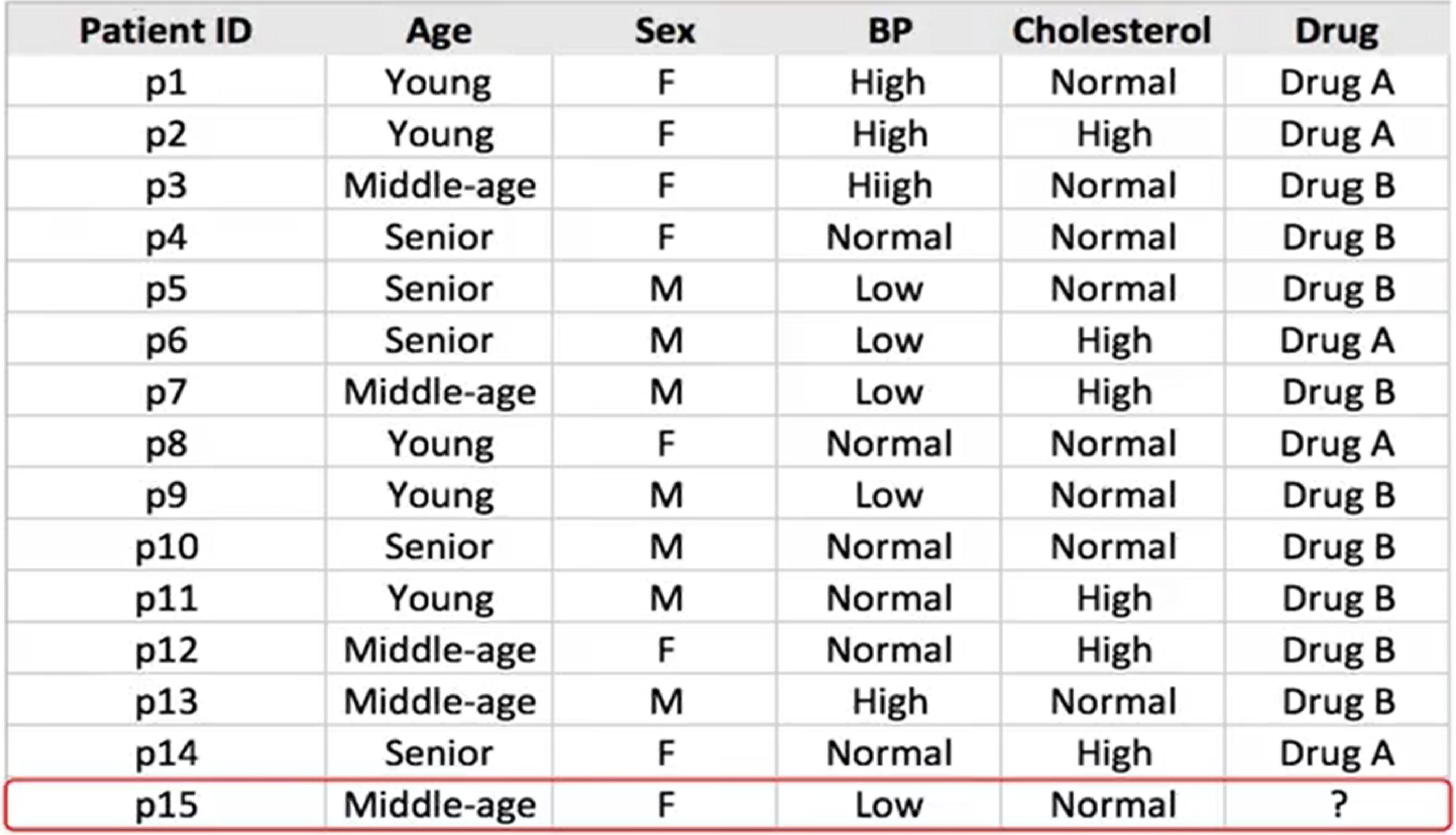


Independent variables

Dependent variabl

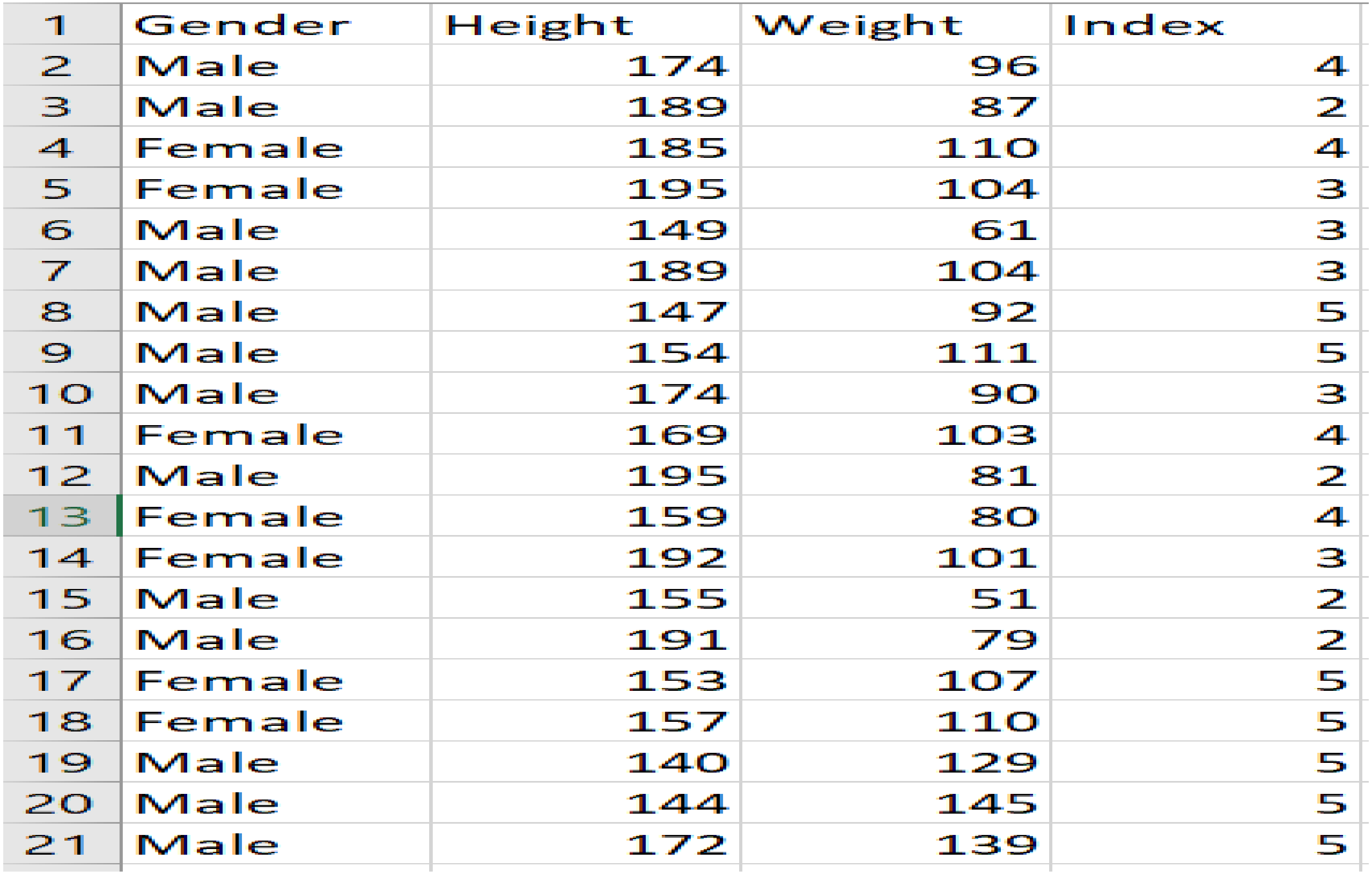
	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	Yes
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	Yes
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	No
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	No
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	?

Data-set



Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	Hiigh	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A
p15	Middle-age	F	Low	Normal	?

Data-set



1	Gender	Height	Weight	Index
2	Male	174	96	4
3	Male	189	87	2
4	Female	185	110	4
5	Female	195	104	3
6	Male	149	61	3
7	Male	189	104	3
8	Male	147	92	5
9	Male	154	111	5
10	Male	174	90	3
11	Female	169	103	4
12	Male	195	81	2
13	Female	159	80	4
14	Female	192	101	3
15	Male	155	51	2
16	Male	191	79	2
17	Female	153	107	5
18	Female	157	110	5
19	Male	140	129	5
20	Male	144	145	5
21	Male	172	139	5

Data-Sciences

Using Data to Make Decision

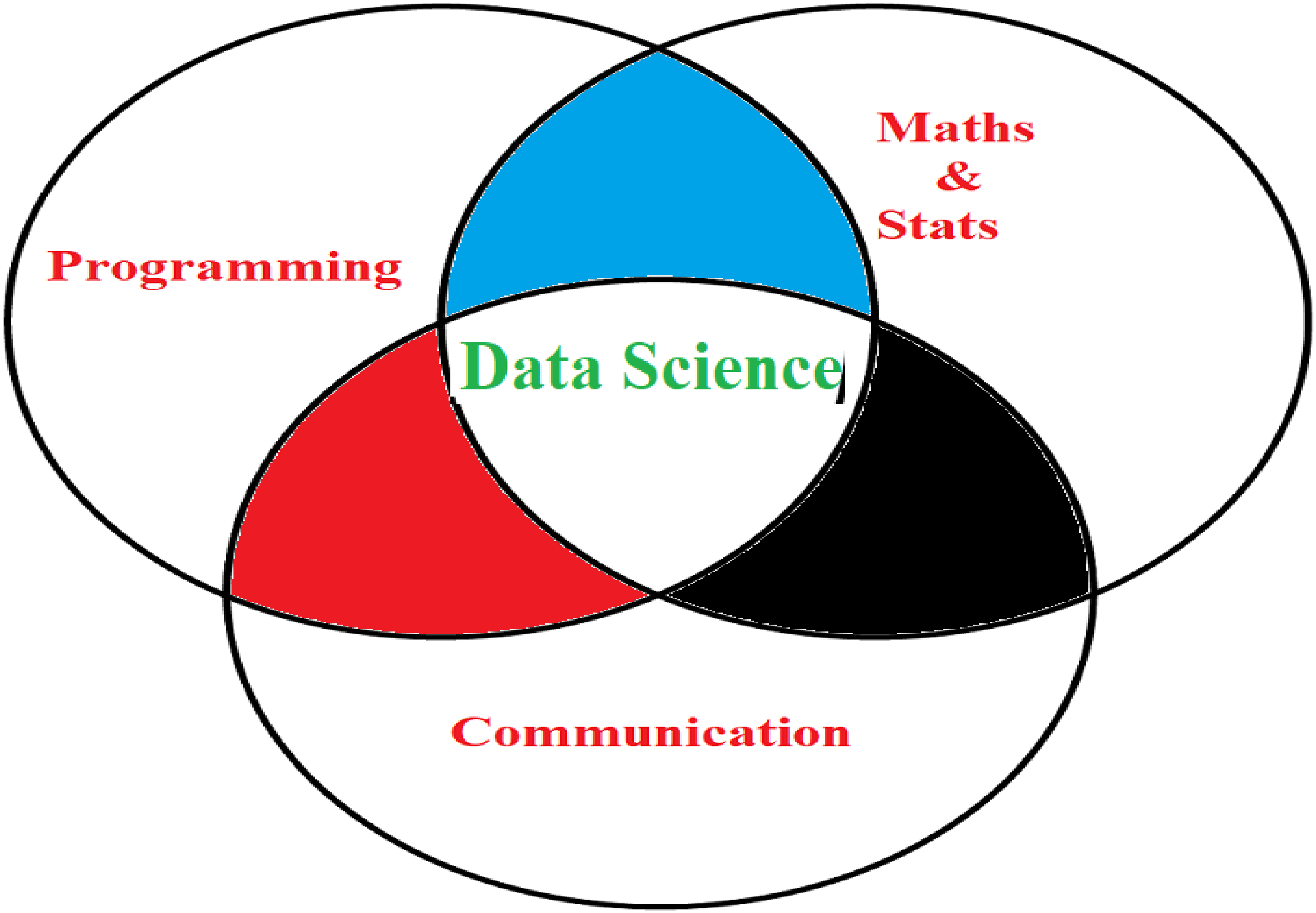
Data science is the process of capturing customer data, processing it, communicating and analyzing it, and then maintaining it.

Data science is the process of analyzing data which involves applying **Machine learning** models, statistical models to derive insights and value from data.

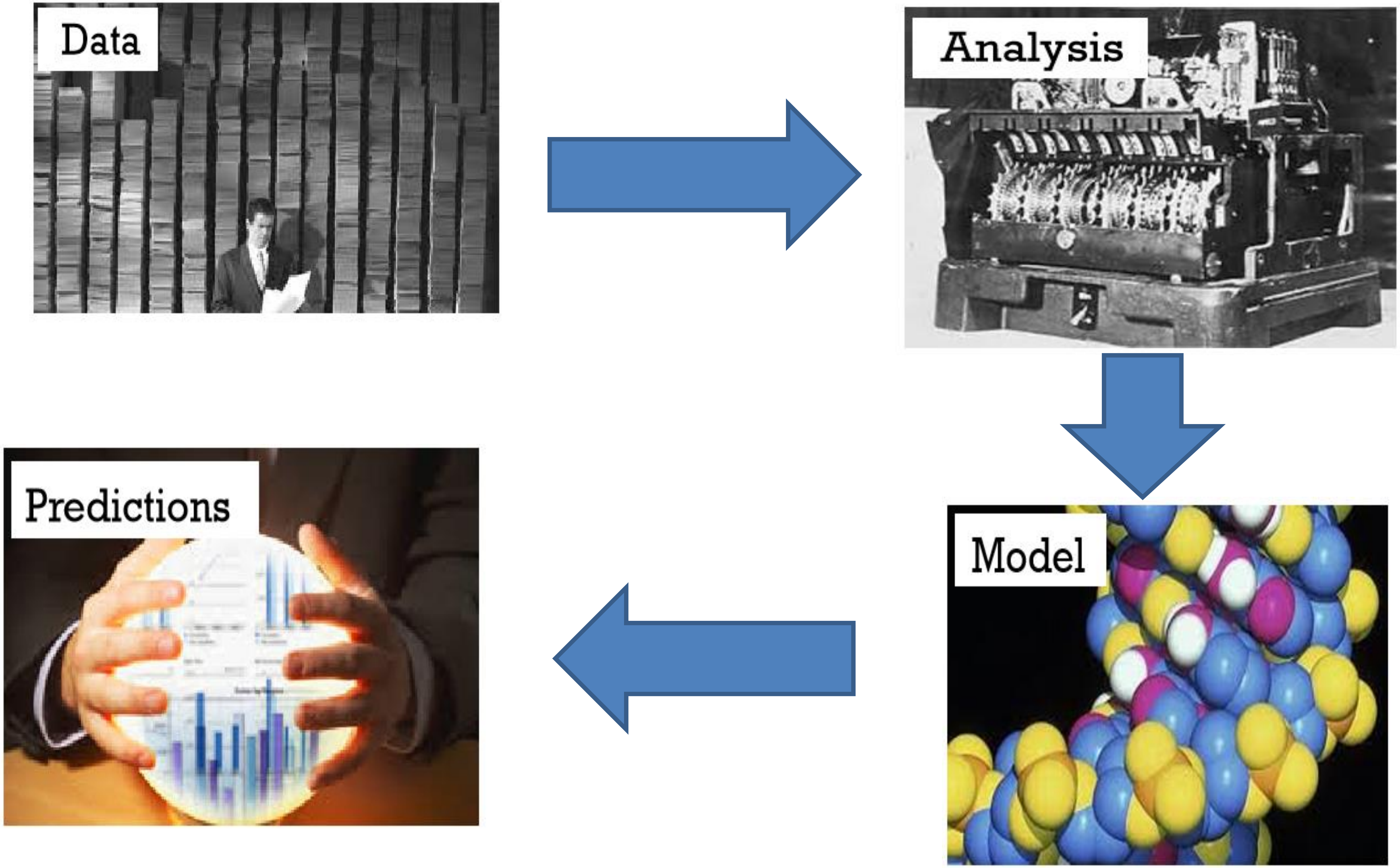
Data Science is the extraction of knowledge from data

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

Data-Science (Data-Analytics)



EXTRACTING INFORMATION FROM DATA



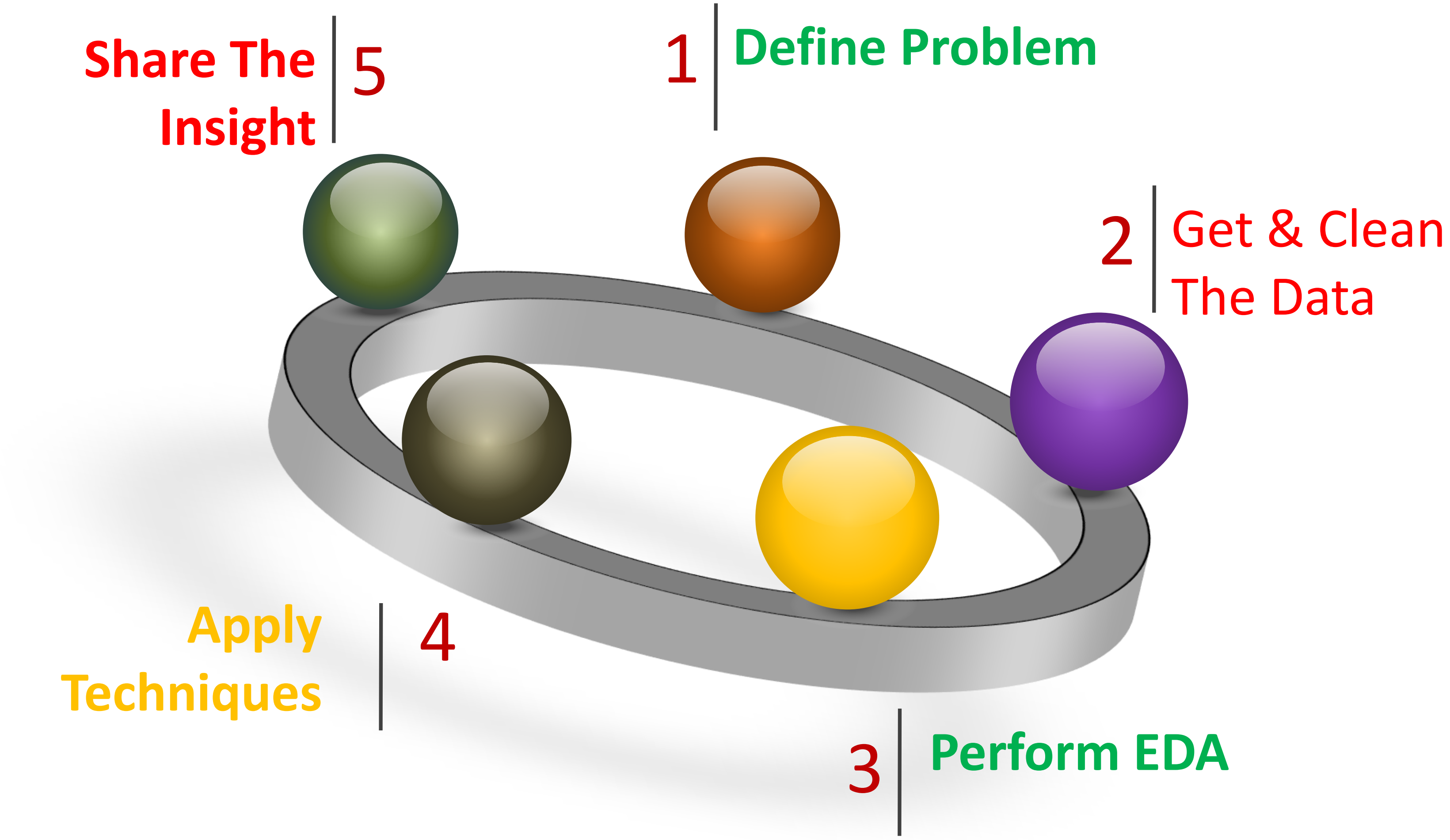
Data

Analysis

Model

Predictions

Data-Science Workflow



1 Define Problem

2 Get & Clean The Data

3 Perform EDA

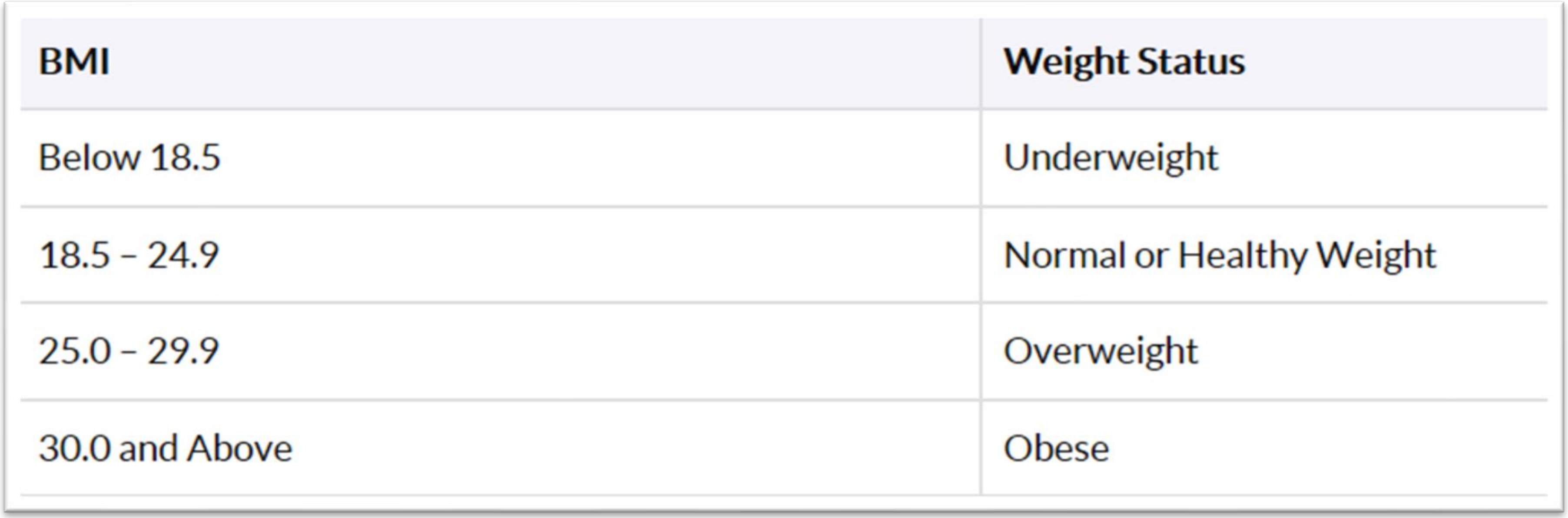
4 Apply Techniques

5 Share The Insight

Step-1 Define The Problem

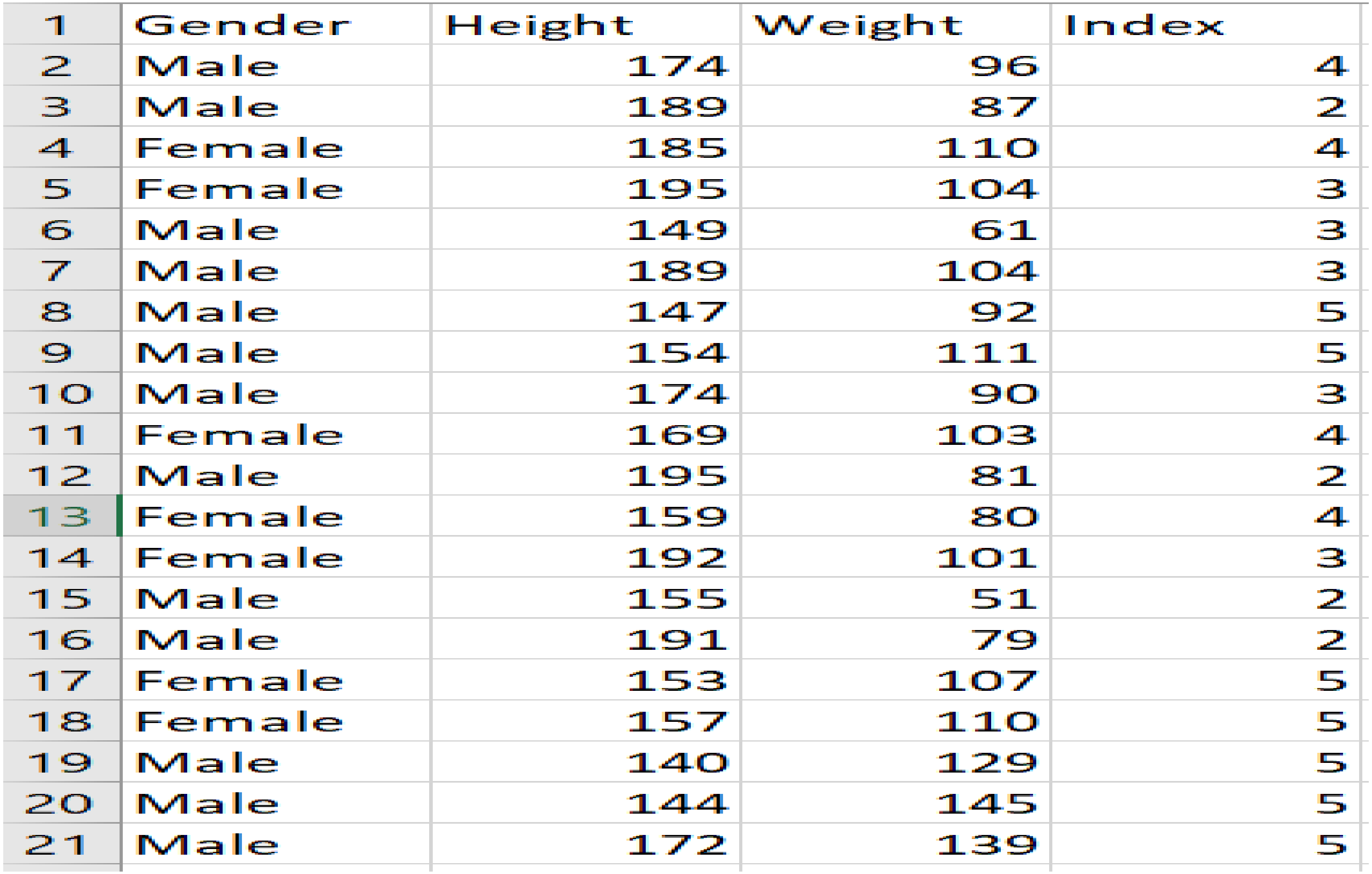
My weight is 68 kg, Height ,165 cm ,
am I normal?

The standard weight status categories associated with BMI ranges for adults are



BMI	Weight Status
Below 18.5	Underweight
18.5 – 24.9	Normal or Healthy Weight
25.0 – 29.9	Overweight
30.0 and Above	Obese

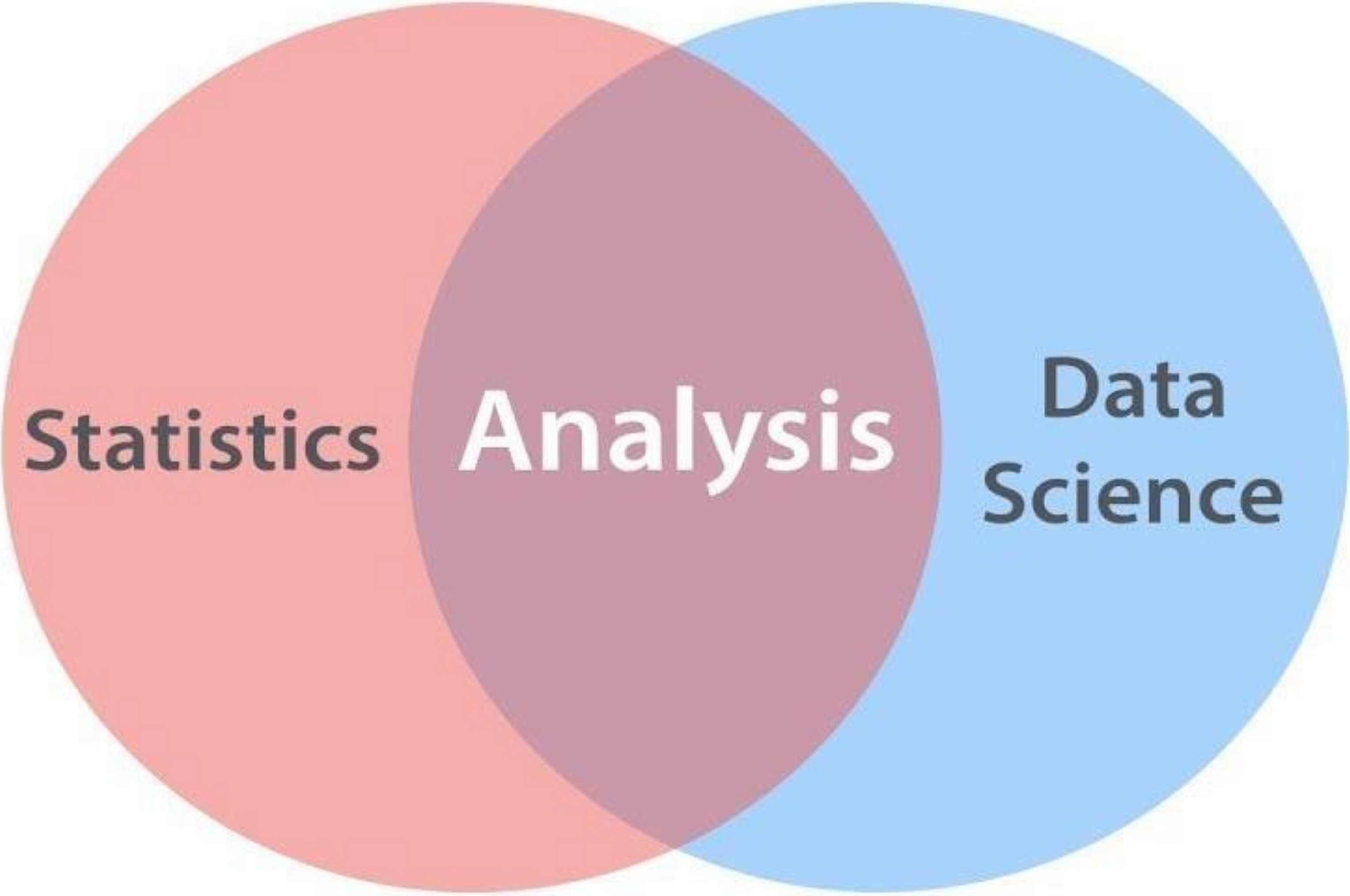
Step 2:Data –Collection



1	Gender	Height	Weight	Index
2	Male	174	96	4
3	Male	189	87	2
4	Female	185	110	4
5	Female	195	104	3
6	Male	149	61	3
7	Male	189	104	3
8	Male	147	92	5
9	Male	154	111	5
10	Male	174	90	3
11	Female	169	103	4
12	Male	195	81	2
13	Female	159	80	4
14	Female	192	101	3
15	Male	155	51	2
16	Male	191	79	2
17	Female	153	107	5
18	Female	157	110	5
19	Male	140	129	5
20	Male	144	145	5
21	Male	172	139	5

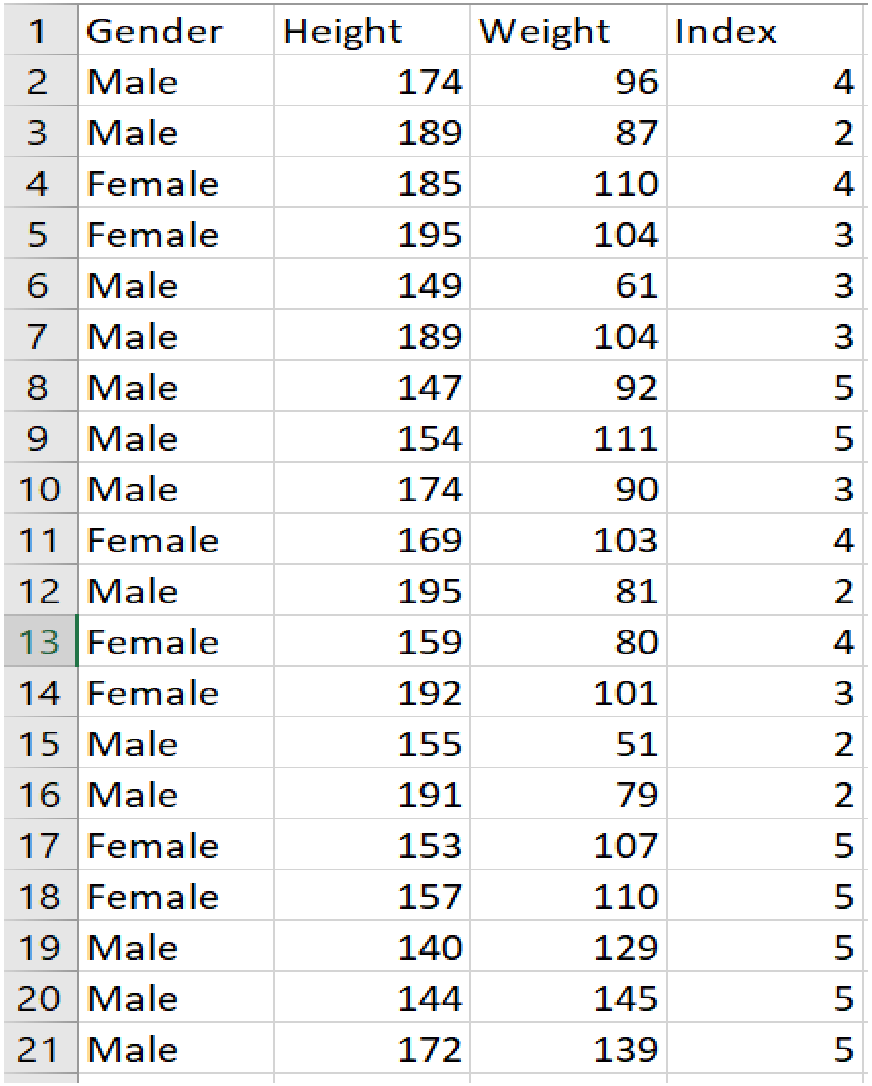
Step 3:Exploratory data analysis (**EDA**)

In **statistics**, exploratory data analysis (**EDA**) is an approach analyzing data sets to summarize their main characteristics, often with visual methods


Statistics
Analysis
Data
Science

When looking at a new dataset, whether it is familiar to you or not, it is important to use the following questions as guidelines for your preliminary analysis .

1. Is the data organized or not?
2. What does each row represent?
3. What does each column represent?
4. Are there any missing data points?
5. Do we need to perform any transformations on the columns?



1	Gender	Height	Weight	Index
2	Male	174	96	4
3	Male	189	87	2
4	Female	185	110	4
5	Female	195	104	3
6	Male	149	61	3
7	Male	189	104	3
8	Male	147	92	5
9	Male	154	111	5
10	Male	174	90	3
11	Female	169	103	4
12	Male	195	81	2
13	Female	159	80	4
14	Female	192	101	3
15	Male	155	51	2
16	Male	191	79	2
17	Female	153	107	5
18	Female	157	110	5
19	Male	140	129	5
20	Male	144	145	5
21	Male	172	139	5

Gender : Male / Female

Height : Number (cm)

Weight : Number (Kg)

Index :

0 - Extremely Weak

1 - Weak

2 - Normal

3 - Overweight

4 - Obesity

5 - Extreme Obesity

Step 4:Apply the Techniques

Here the Machine Learning Technique come into Picture to solve the problem.

Step 5:Share the Insights

How to solve a problem in Data Science

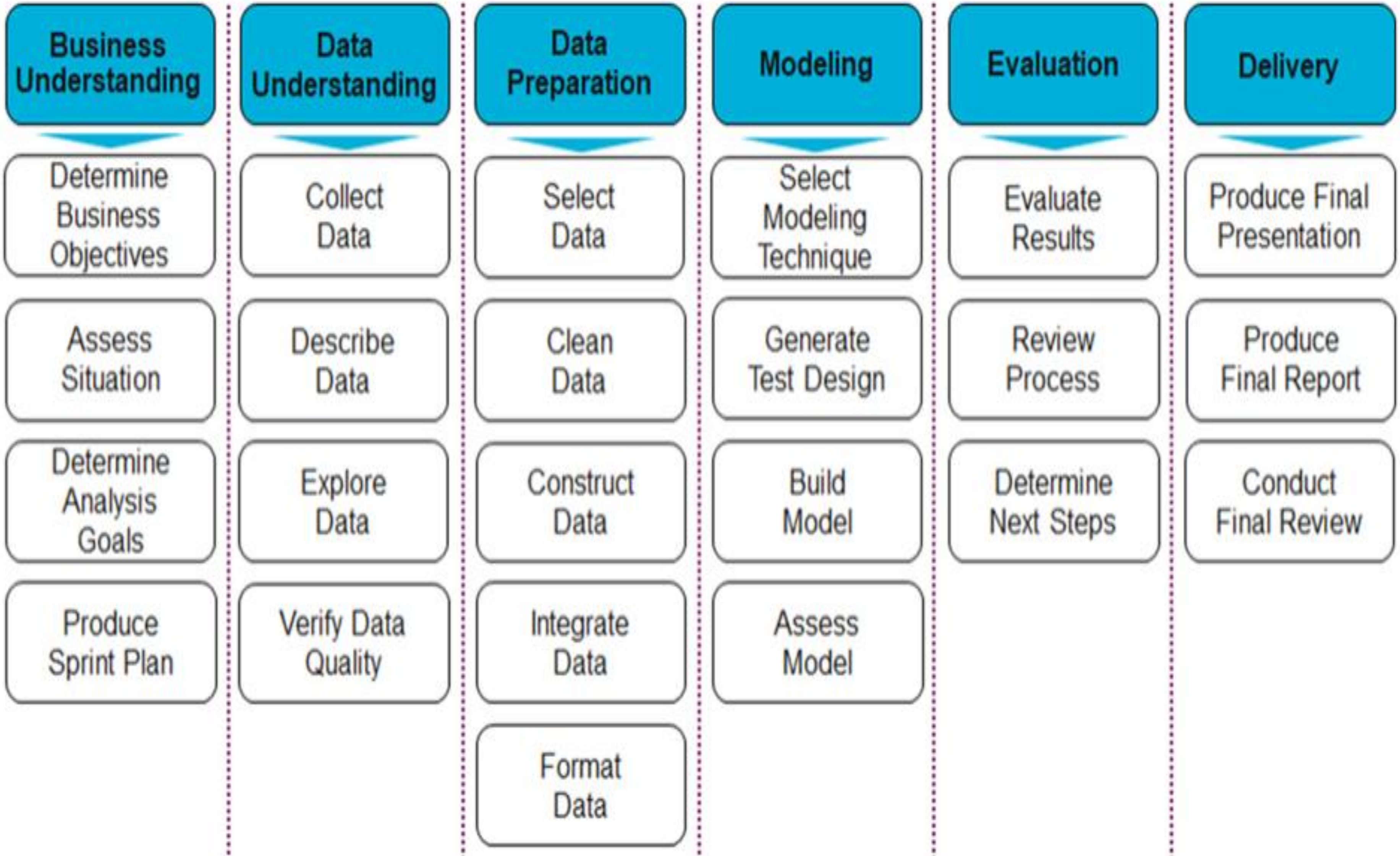
Problems in Data Science are solved using Algorithms. But, the biggest thing to judge is which algorithm to use and when to use it?

Machine Learning Algorithms

How to solve Machine Learning problems?

CRISP-DM: Data mining methodology to investigate (Big) Data
CRoss-Industry Standard Process for Data Mining

1. **Business Understanding**
2. **Data Understanding**
3. **Data preparation**
4. **Modeling**
5. **Evaluation**
6. **Deployment**
7. **Start again in Iterative process**



Business Understanding
- Determine Business Objectives
- Assess Situation
- Determine Analysis Goals
- Produce Sprint Plan

Data Understanding
- Collect Data
- Describe Data
- Explore Data
- Verify Data Quality

Data Preparation
- Select Data
- Clean Data
- Construct Data
- Integrate Data
- Format Data

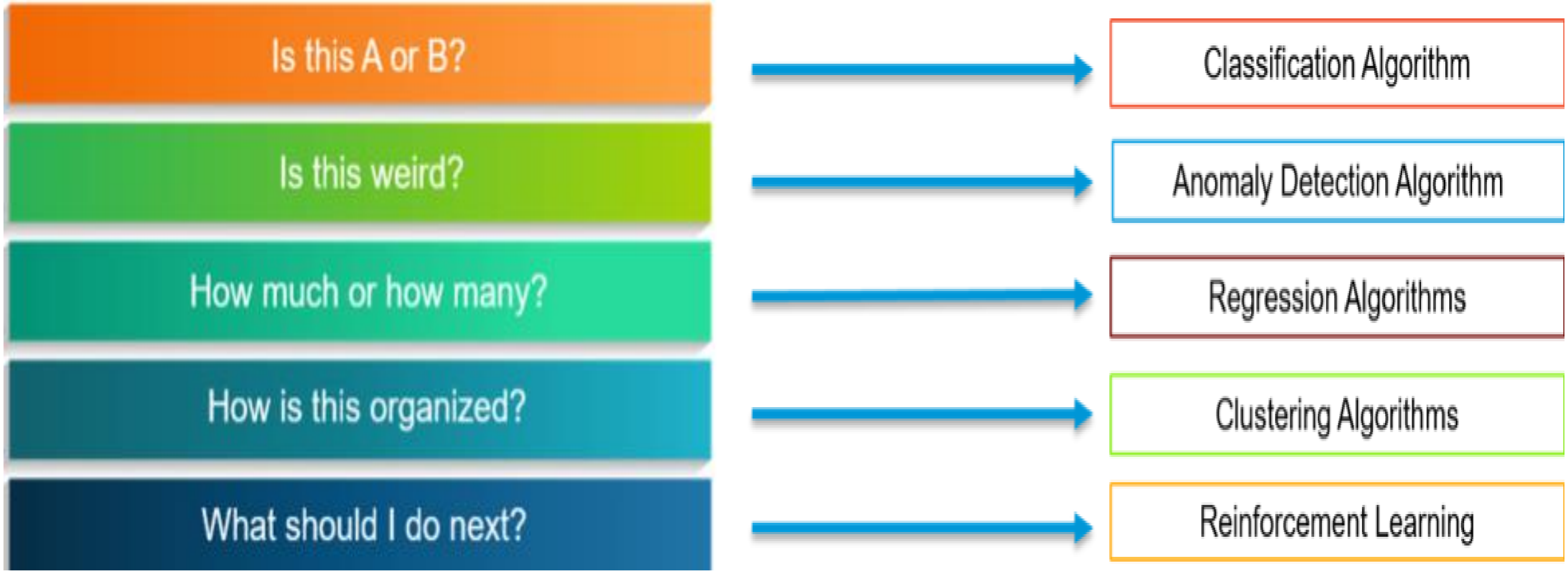
Modeling
- Select Modeling Technique
- Generate Test Design
- Build Model
- Assess Model

Evaluation
- Evaluate Results
- Review Process
- Determine Next Steps

Delivery
- Produce Final Presentation
- Produce Final Report
- Conduct Final Review

Type of Problem

1. Is this fruit Sweet or Sour ?
2. Is this Weird?
3. How much or many ?
4. how is this organized?
5. What should i do next ?



Is this A or B? → Classification Algorithm

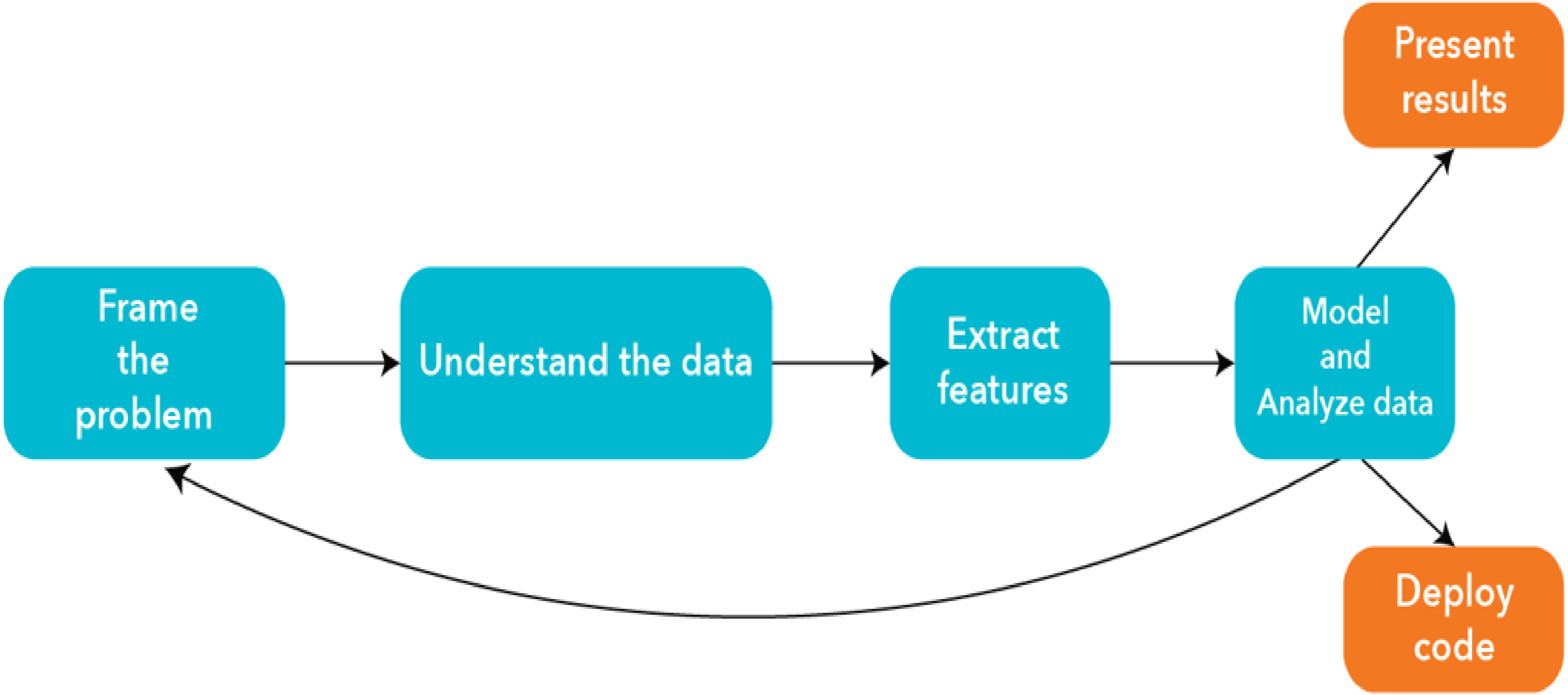
Is this weird? → Anomaly Detection Algorithm

How much or how many? → Regression Algorithms

How is this organized? → Clustering Algorithms

What should I do next? → Reinforcement Learning

Data Science Road Map



Frame the problem

Understand the data

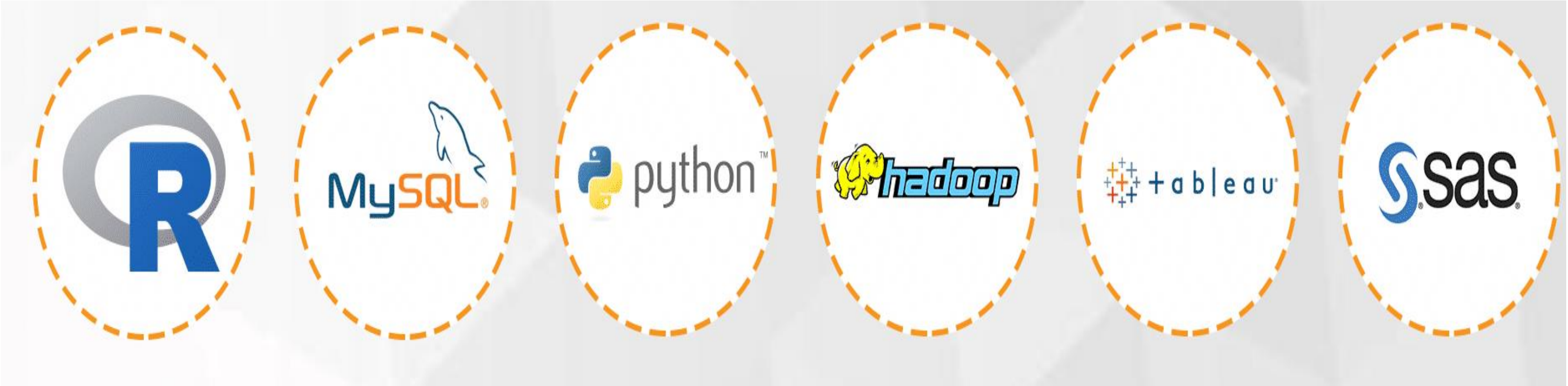
Extract features

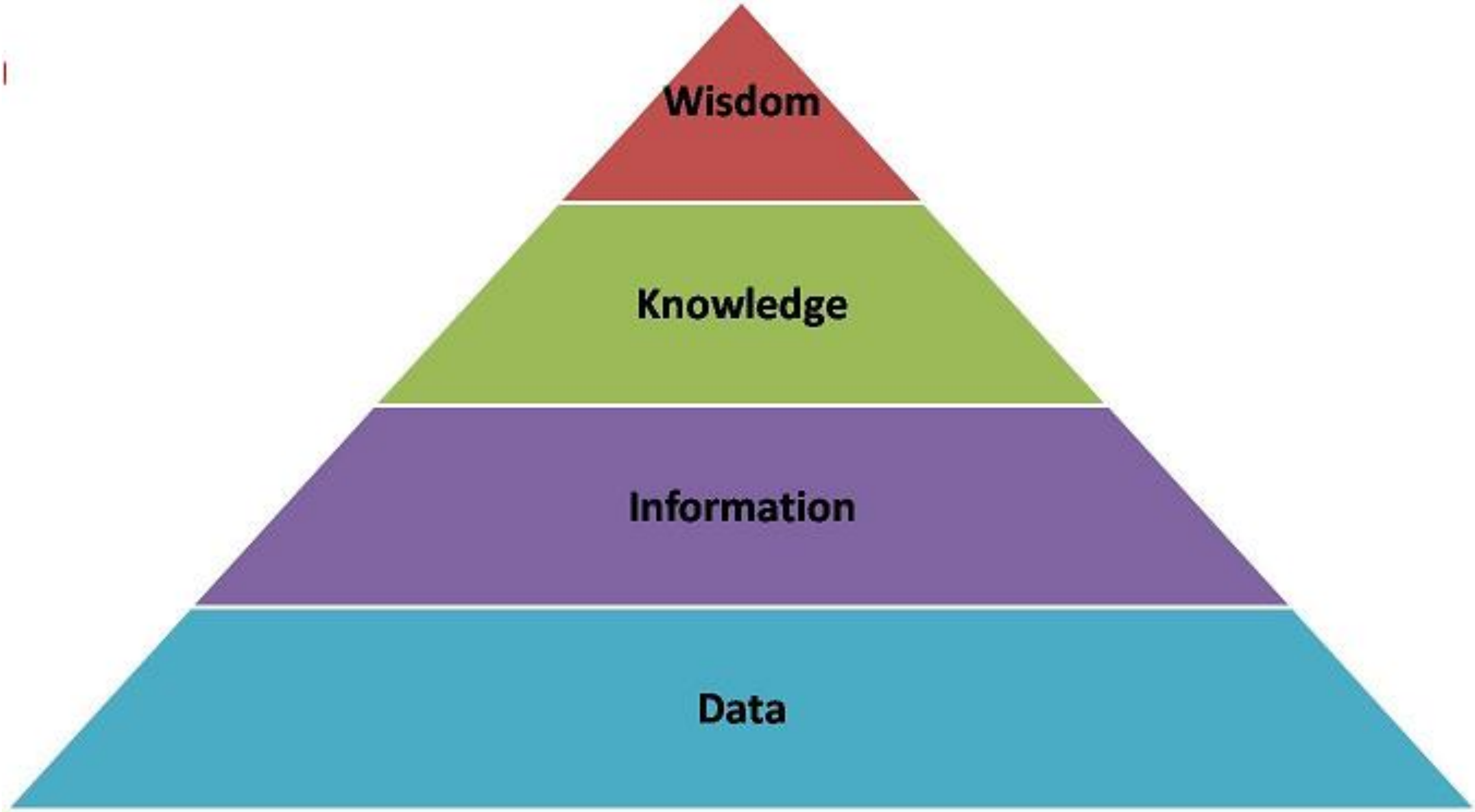
Model and Analyze data

Present results

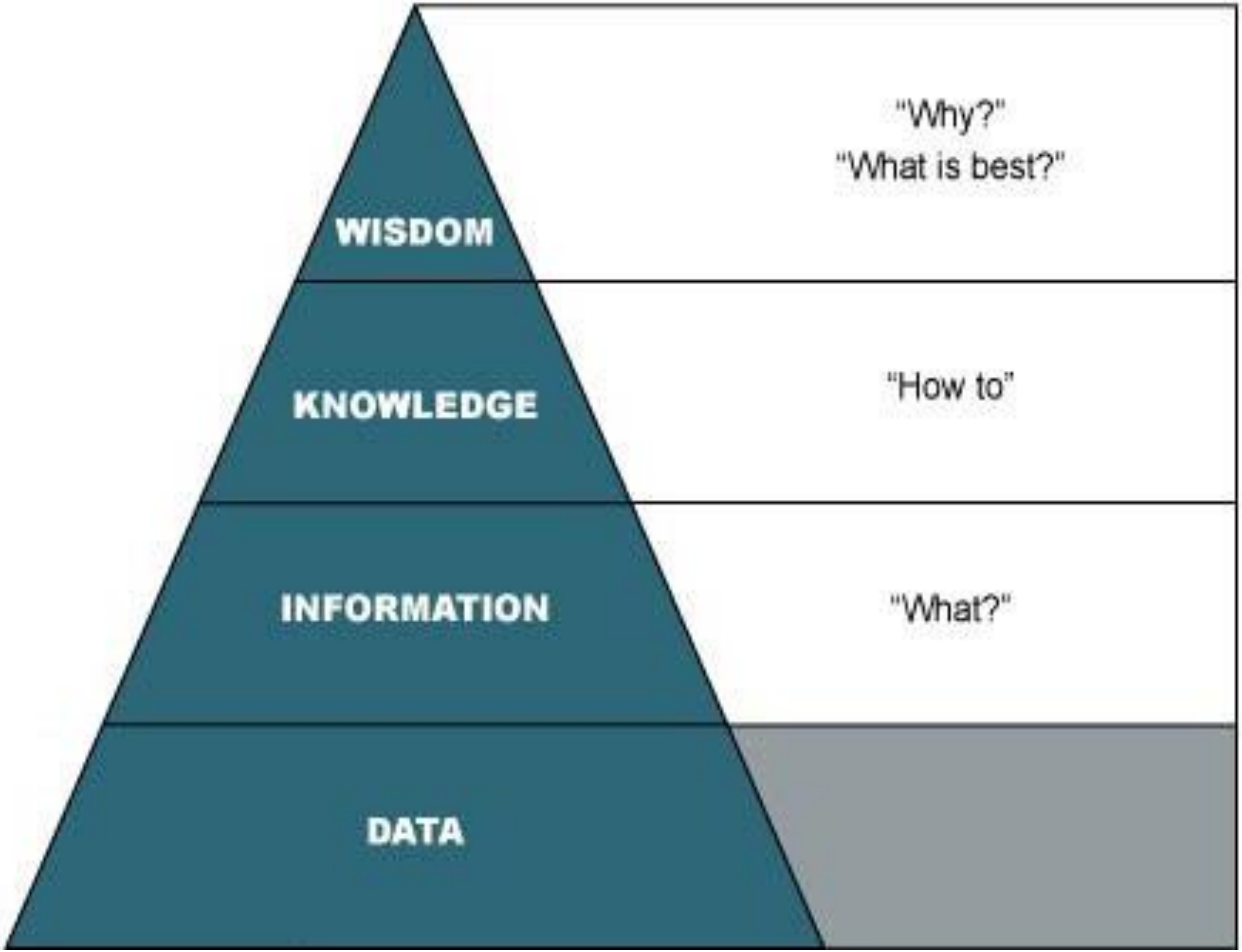
Deploy code

DATA SCIENCE
ANALYSIS
STRUCTURE
ALGORITHM
001011
101110
PROCESS
PROGRAMMING
SOLVING
KNOWLEDGE


R
MySQL
python
hadoop
tableau
sas



From Data to Information to Knowledge


WISDOM
"Why?"
"What is best?"
KNOWLEDGE
"How to"
INFORMATION
"What?"
DATA

Data

- Structured data
- Unstructured data
- Qualitative data
- Quantitative data
- Discrete data
- Continuous data
- Nominal
- Ordinal
- Interval

Tool

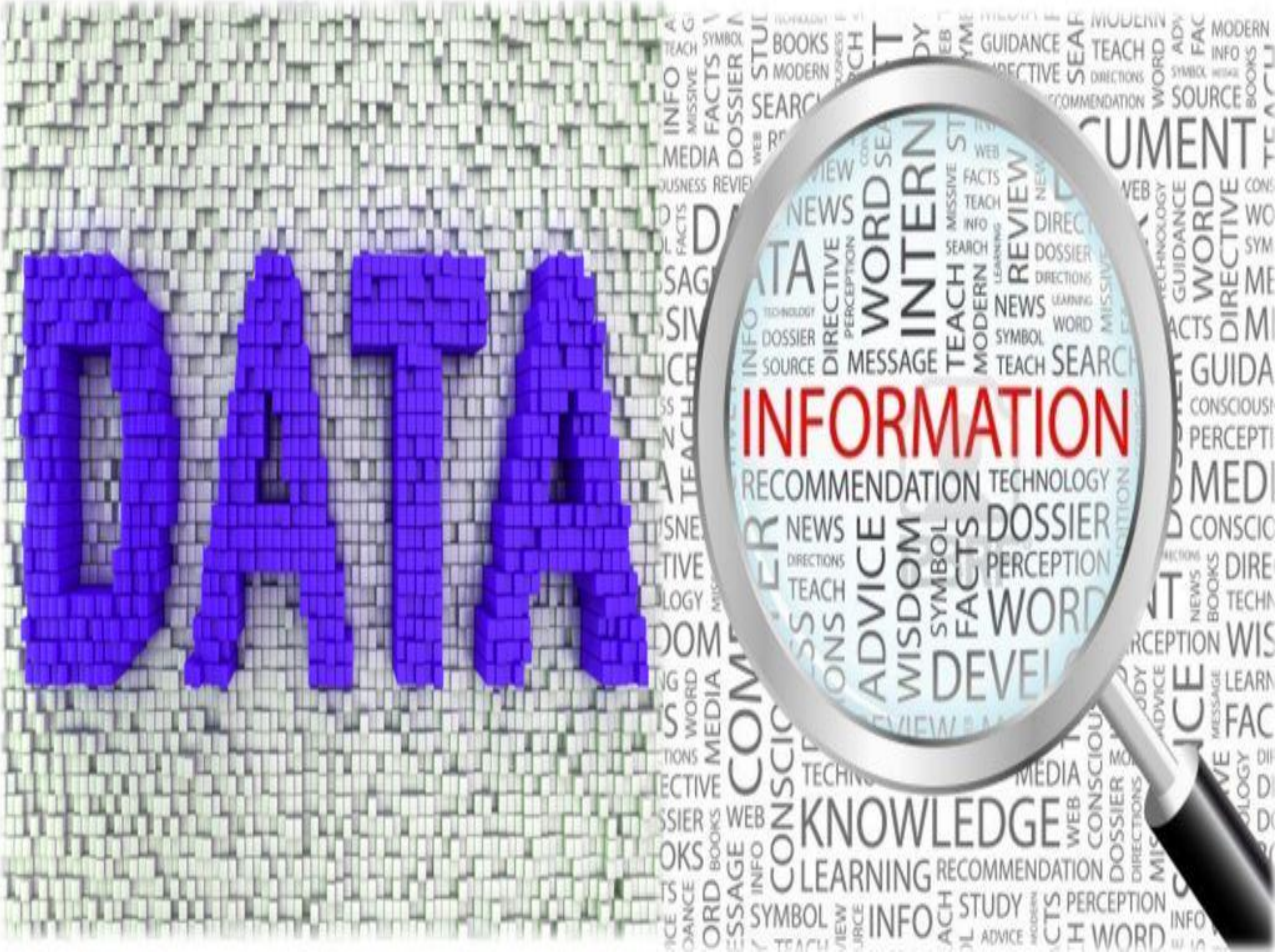
- Pandas
- Numpy
- Matplotlib
- Seaborn
- Sklearn
- R
- Tensorflow
- keras
- NLTK
- Request
- Beautiful Soup
- Pickle

Math

- Probability Theory and Statistics
- Bayes' Theorem
- Random Variables
- Variance and Expectation
- Conditional and Joint Distributions
- Standard Distribution
- Calculus

Algorithms

- Supervised
- Unsupervised
- Semi-Supervised
- Reinforcement


DATA
INFORMATION
RECOMMENDATION
TECHNOLOGY
DOSSIER
PERCEPTION
WORD
NEWS
ADVICE
WISDOM
SYMBOL
FACTS
MESSAGE
SEARCH
TEACH
DIRECTIVE
SOURCE
KNOWLEDGE
LEARNING
MEDIA
STUDY
INFO
BOOKS
GUIDANCE
MODERN
CONSCIOUSNESS

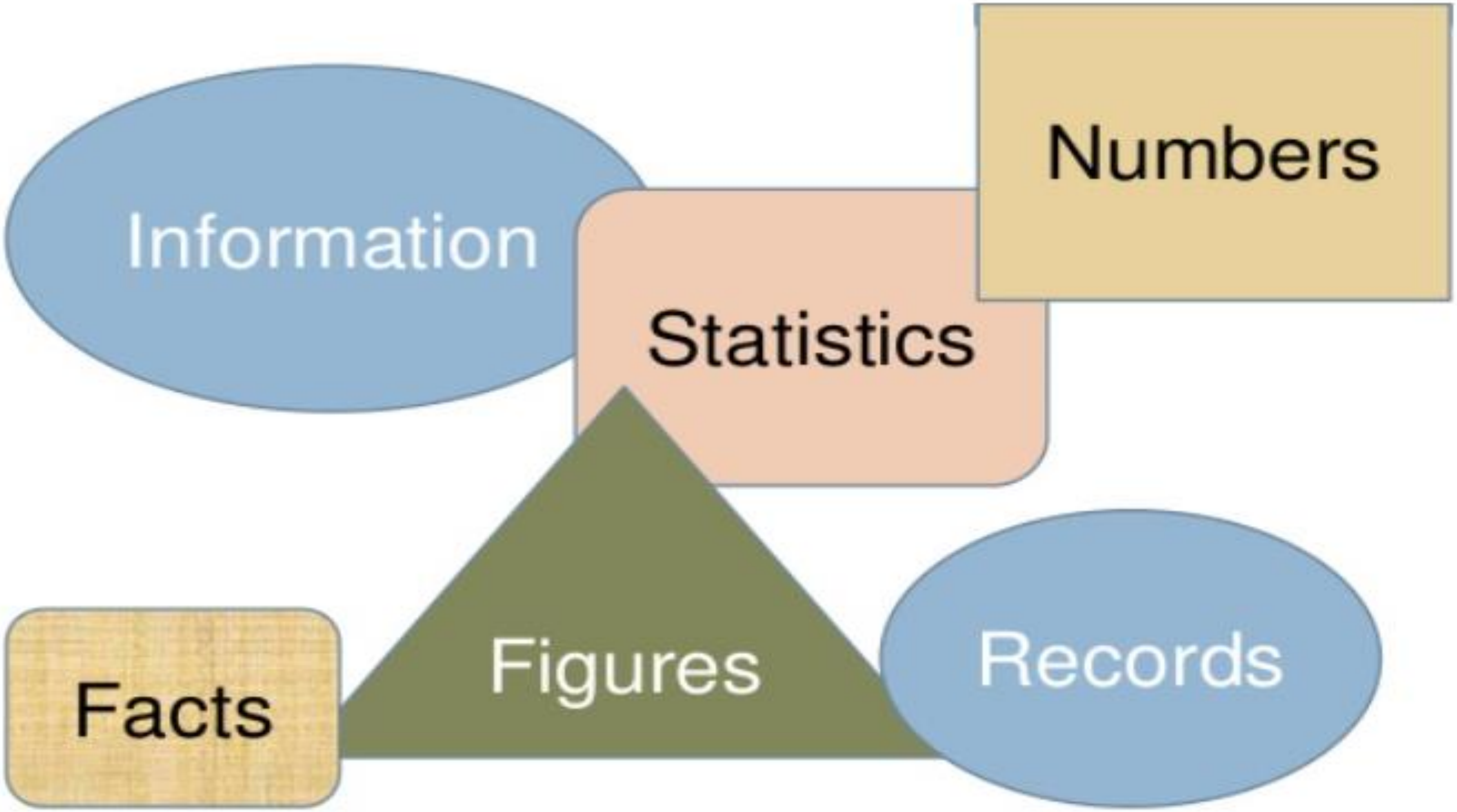
DATA

Data is a collection of figures and facts, and is raw, unprocessed, and unorganized.

The Latin root of the word “data” means **“something given**”, which is a good way to look at it.

Data is a value assigned to a thing.

DATA



Information

Numbers

Statistics

Figures

Records

Facts

Unstructured vs. Structured data

Structured data: data stored in rows and columns, mostly numerical, where the meaning of each data item is defined. This type of data constitutes about 10% of the today's total data and is accessible through database management systems.

Unstructured data: data of different forms like e.g. text, image, video, document, etc. It can also be in the form of customer complaints, contracts, or internal emails. This type of data accounts for about 90% of the data created in this century

Qualitative and Quantitative data

Qualitative data is everything that refers to the quality of something: A description of colors, 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.

Discrete data and Continuous data

Discrete data is numerical data that has gaps in it: e.g. the count of golf balls. There can only be whole numbers of golf ball (there is no such thing as 0.3 golf balls).

Continuous data is numerical data with a continuous range: e.g. size of the golfballs can be any value (e.q. 10.53mm or 10.54mm but also 10.536mm).

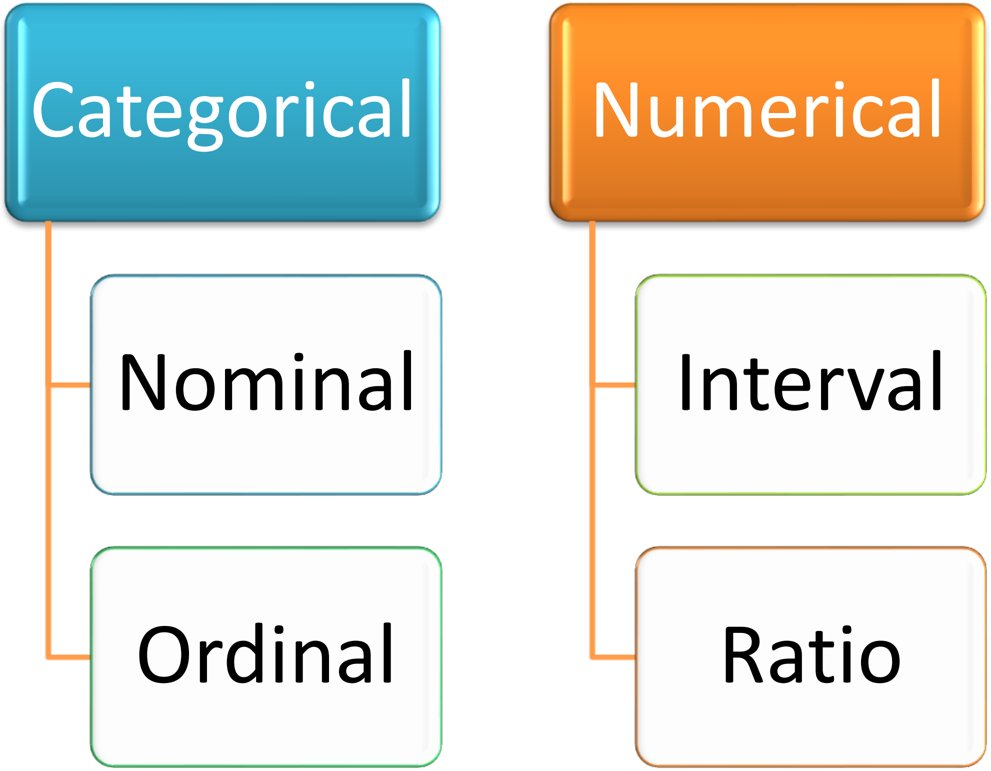
In continuous data, all values are possible with no gaps in between.

Data and Information

- Facts , statistics used for reference or analysis.
- Data must be interpreted , by a human or machine to drive meaning.
- Data is meaningless.

- Data that has been processed within a context to give it meaning.
- Information is data that has been processed
- Information is interpreted data
- information is meaningful

Types of Data



Categorical

Nominal

Ordinal

Numerical

Interval

Ratio

Categorical

Qualitative data are often termed categorical data.

Nominal (Unordered list)
A variable that has two or more categories, without any implied ordering.

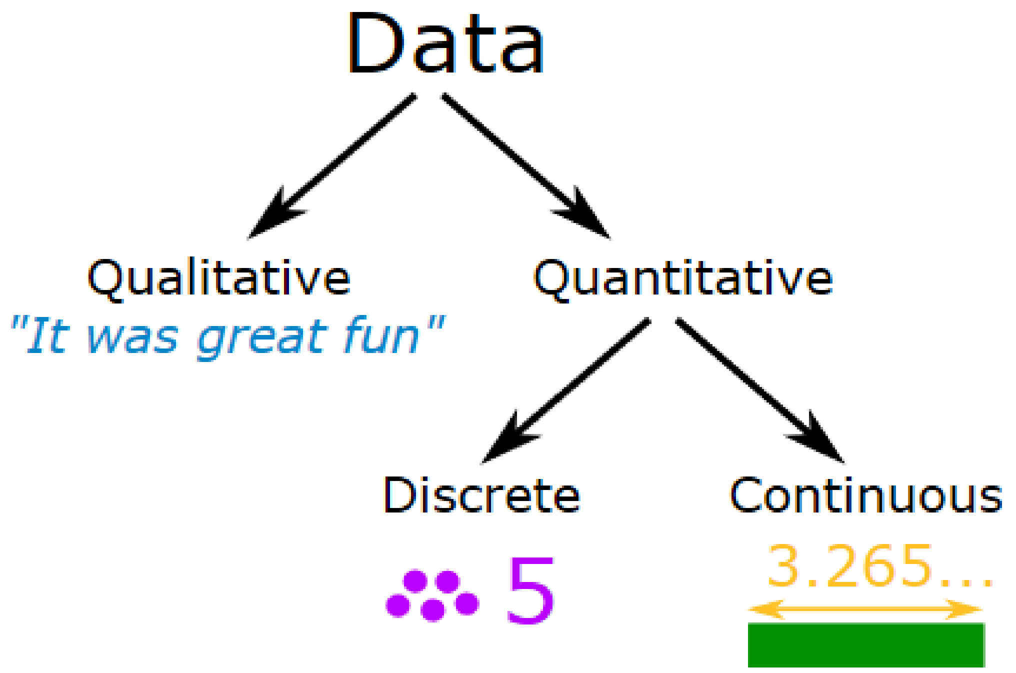
Ordinal Variable (Ordered list)
A variable that has two or more categories, with clear ordering.

- Gender - Male, Female
- Marital Status - Unmarried, Married, Divorcee
- State - New Delhi, Haryana, U.P

- Scale - Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree
- Rating - Very low, Low, Medium, Great, Very great

Data can be qualitative or quantitative

Qualitative data is descriptive information (it *describes* something)
Quantitative data is numerical information (numbers)



Qualitative

He is brown and black
He has long hair
He has lots of energy

Quantitative

He has 4 legs
He has 2 brothers
He weighs 25.5 kg
He is 565 mm tall

When To Use What In Descriptive Statistics To Measure Central Tendency?

- *For Nominal:* ***Mode***
- *For Ordinal:* ***Median***
- *For Interval/Ratio (not skewed): Mean*
- *For Interval/Ratio (skewed):* ***Median***