Intro to Machine Learning

By: Farisology

Intro

What is Artificial Intelligence?

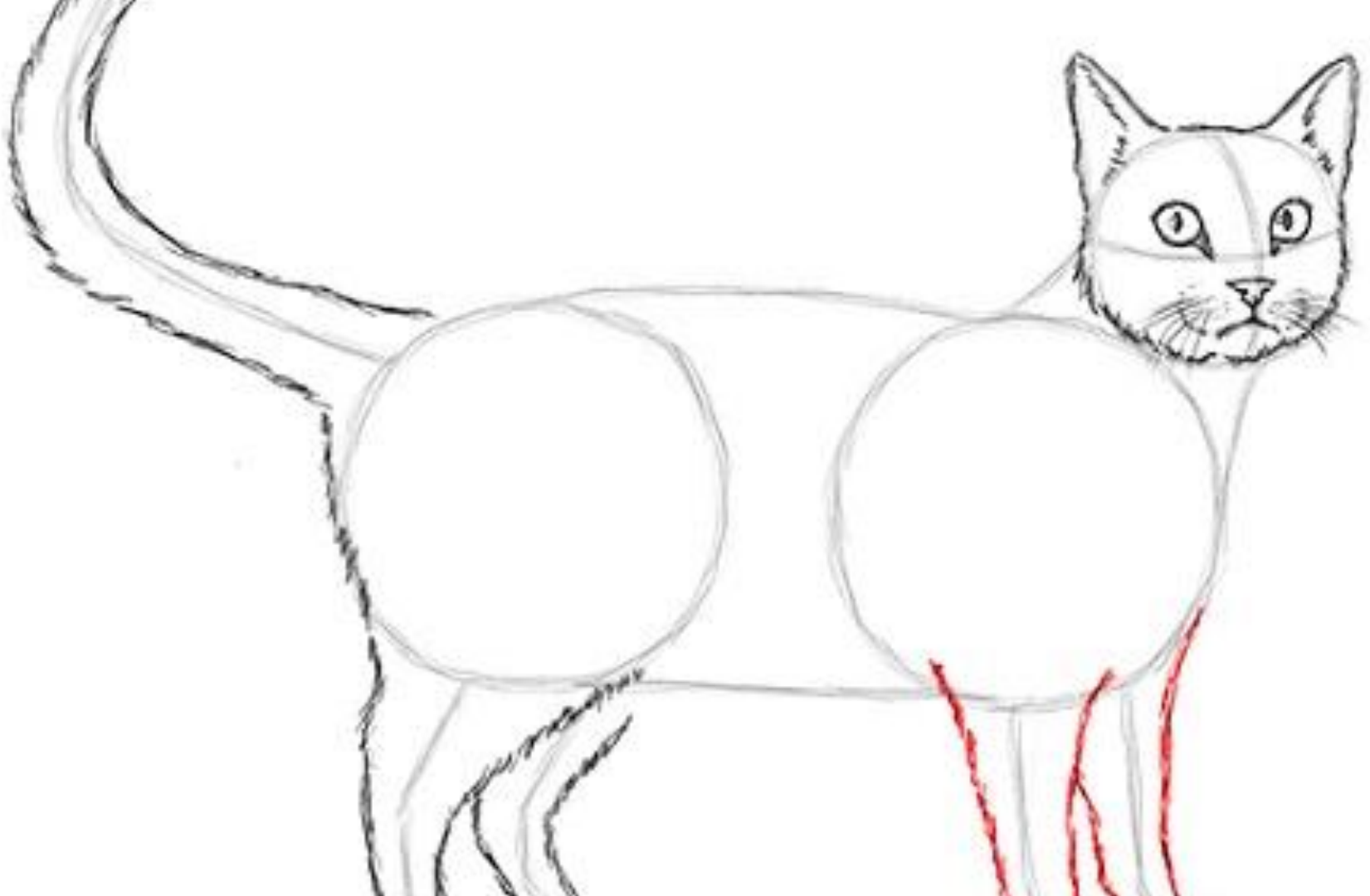
How does a machine learn to recognize objects?

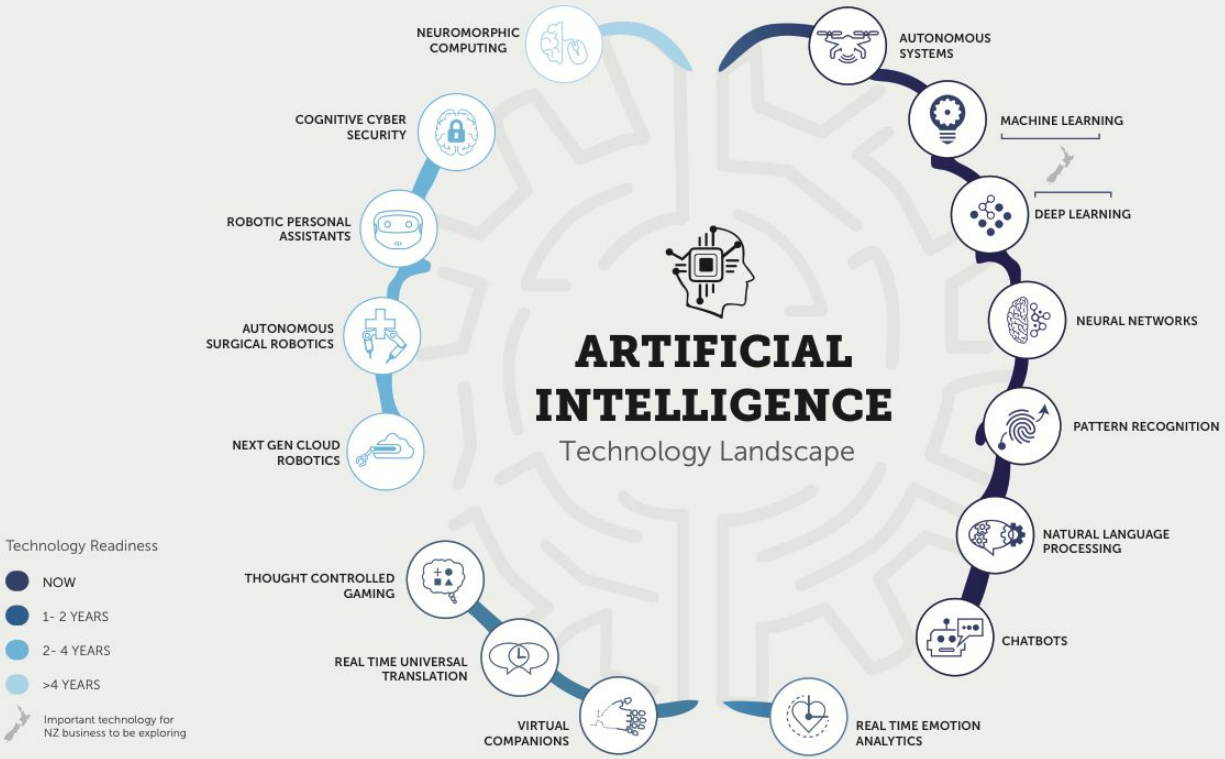
How do we humans learn?

How did you learn to recognize blue color?

How did you recognize cars/dogs/cats?

How does a machine learn to recognize objects?





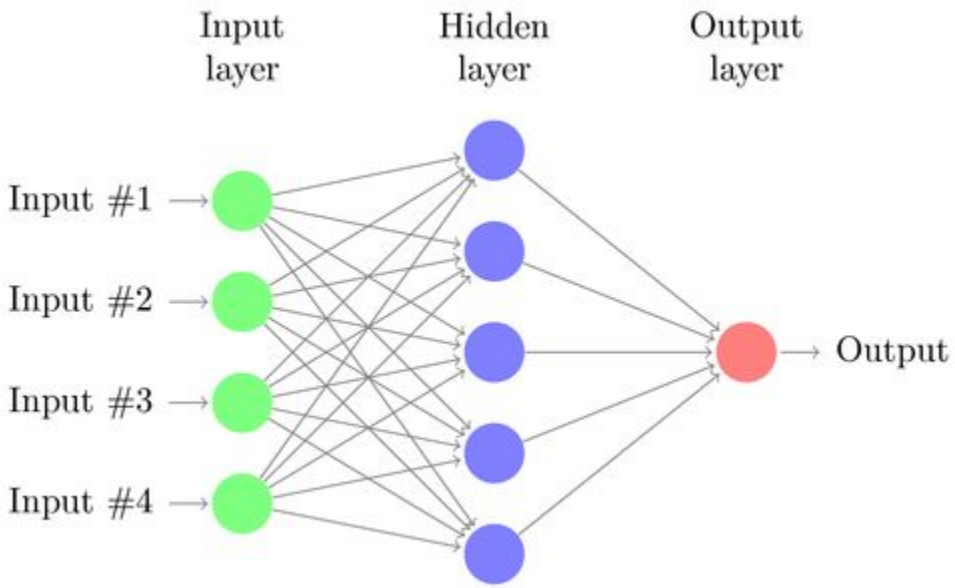
Sources:
Frost & Sullivan "Artificial Intelligence- R&D and Applications Road Map" (Dec 2016), Harvard Business Review- The competitive landscape for Machine Intelligence (Nov 2016), Shivon Zilis and James Chan "The State of Machine Intelligence, 2016" (2016), Stanford University. "Artificial Intelligence and Life in 2030" (2016), https://en.wikipedia.org/wiki/Artificial_intelligence (2017)

CallaghanInnovation

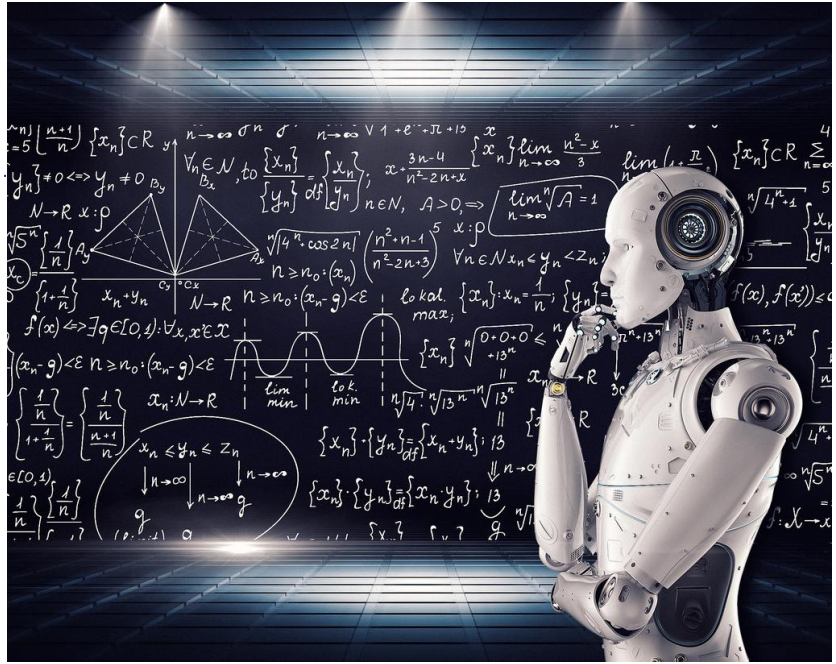
Machine defeating human in Alphago

Self driving cars

Artificial Intelligence



If-else, search theory, reasoning, fuzzy logic,



Preface

Agenda

Introduction:

- **Definition of Machine Learning.**
- **The concept of learning.**
- **Types of learning.**
- **Using data to make informed decisions.**
- **Machine learning workflow: from data to deployment.**

Real world data:

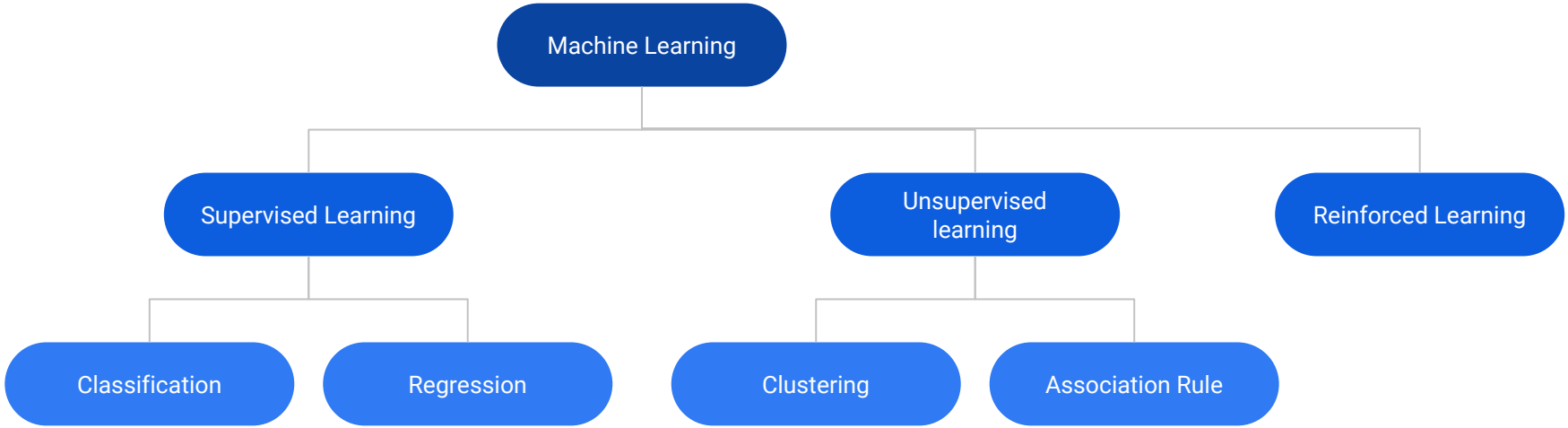
- **Data Collection.**
- **Pre-processing for modelling.**
- **Using data visualization.**

Machine Learning Definition

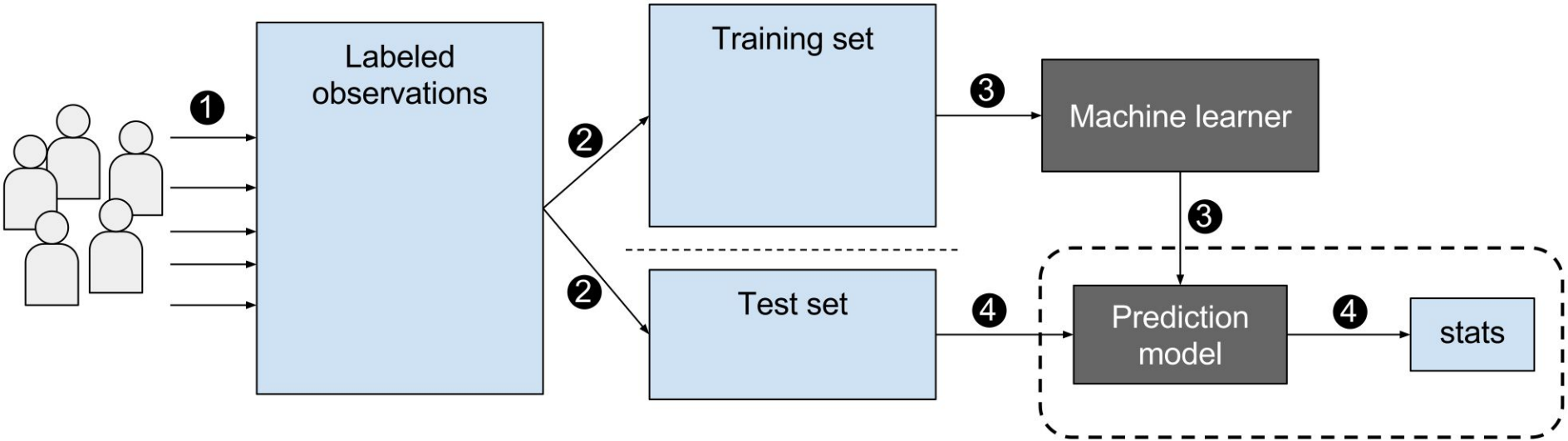
Arthur 1959: The subfield of computer science that gives computers the ability to learn without being explicitly programmed.

Mitchel 1997: A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E.

Types of Learning



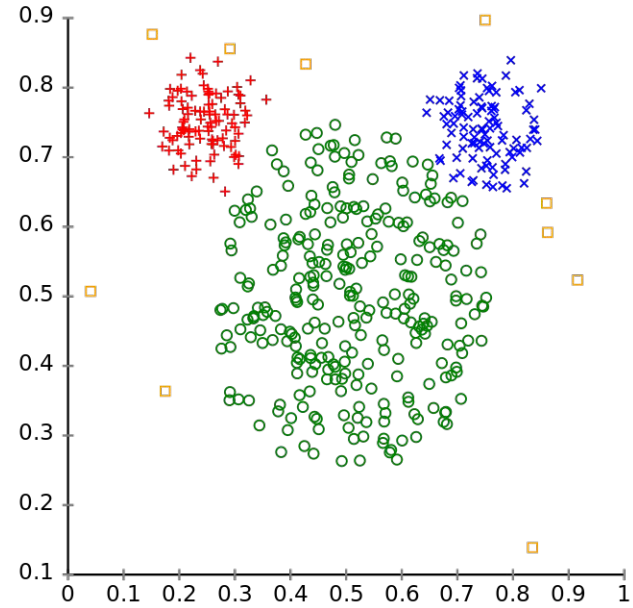
Supervised Learning



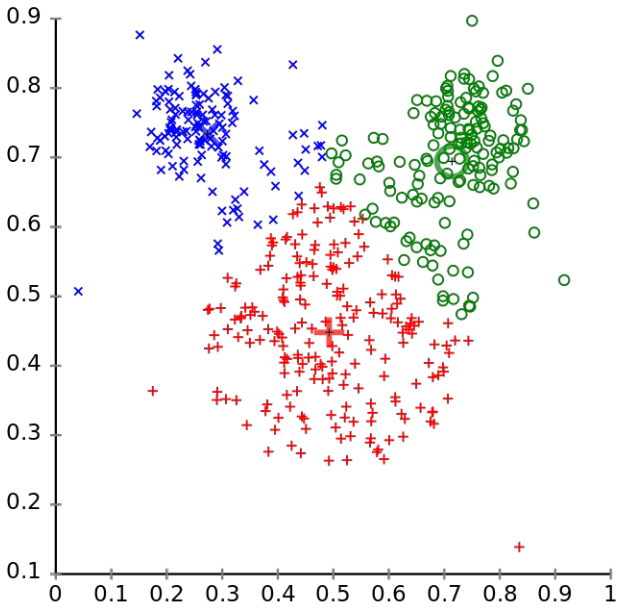
Unsupervised Learning

Different cluster analysis results on "mouse" data set:

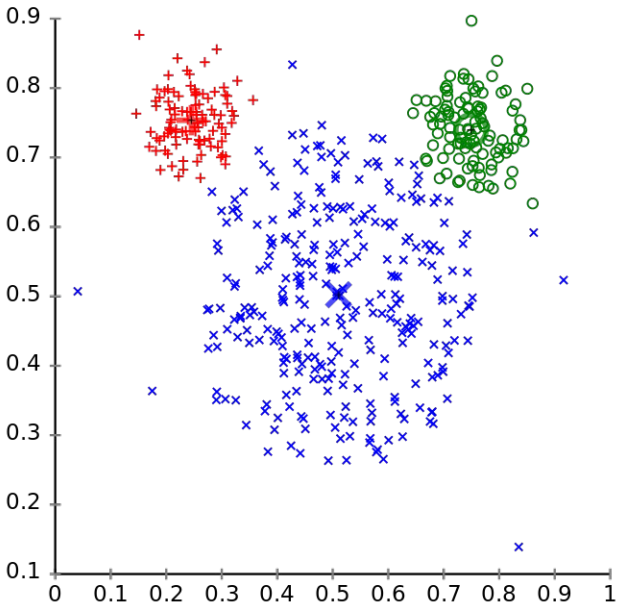
Original Data



k-Means Clustering



EM Clustering



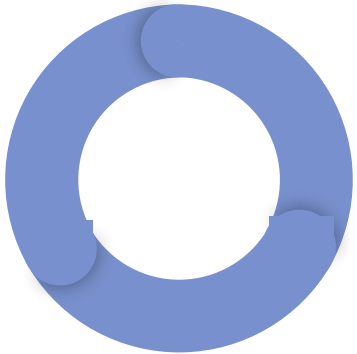
The difference between ML and Sequential logic

Machine learning Paradigm

Input

Output

Training



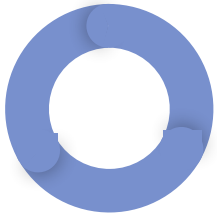
Input: the training data, features the represent an entity in our real world.

Output: the target that we want predict/detect. Basically we want to train the machine to figure out this part on it own.

Based on any model the training is the process of autonomously recognize the patterns and the relationship that connects the input to the output. Mathematically, figuring out the coefficients of an equation.

This model is a trained algorithm to perform certain type of tasks in its own without being explicitly programmed.

New input



Output

New data coming from our business (user profiles)

The trained model will predict the type of the input value.

Classification of the new user profiles.

Supervised Learning

Classification

Predicting Categories

Types of an entity

- Bad - Good - Medium
- Sick - not sick
- Hot dog- not hot dog
- Sad - happy - surprised - angry

Regression

Predicting values

Continuous values

- Sales
- Coordinates
- Time
- Age
- Power
- pressure


ABT
AKU BARU TAHU
imgflip.com

Drive Impact
Create Value

XX% sales increase
XX% ROI growth

Applications of Machine Learning

FinTech

- Credit Risk Scoring
- Fraud Detection and Prevention
- Marketing, Customer Retention, and Loyalty Programs
- Asset management

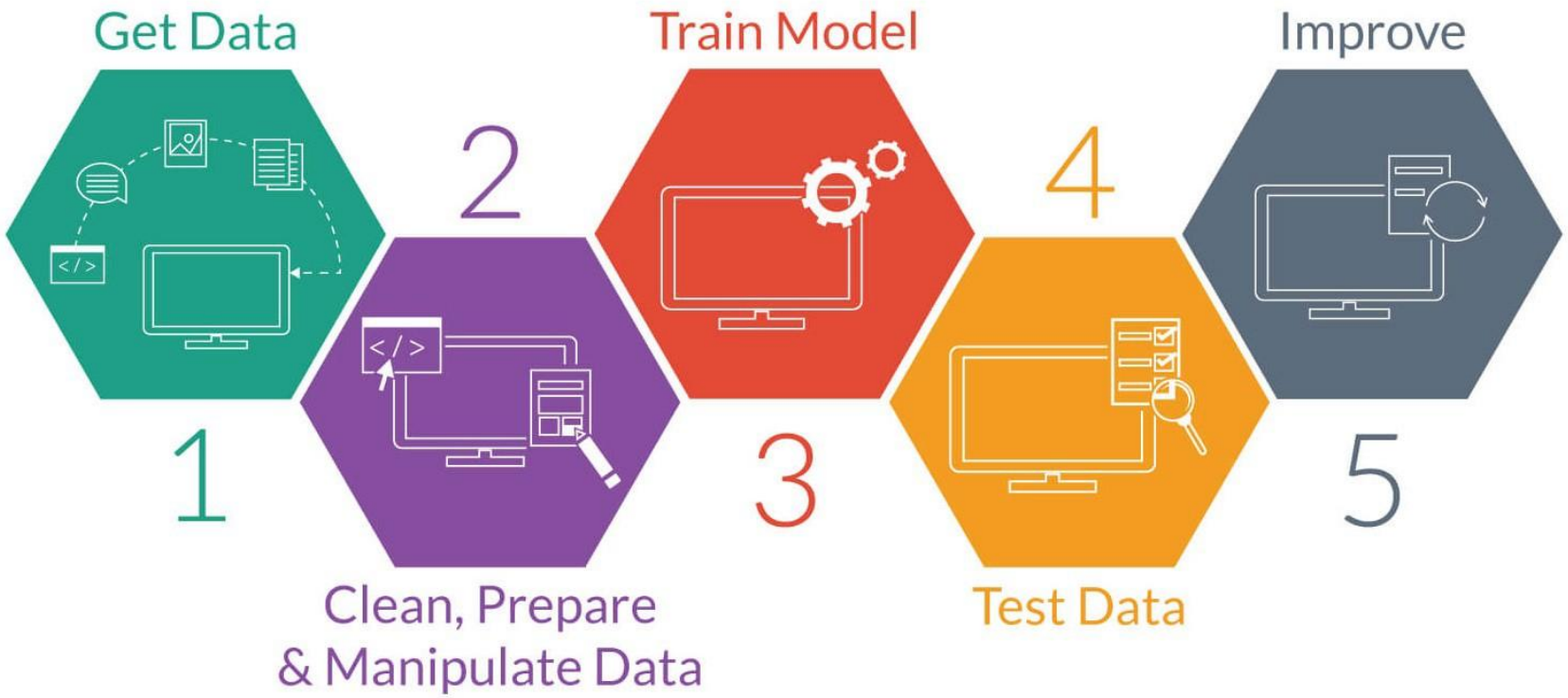
Health Care

- Improving diagnostic accuracy and efficiency
- Drug discovery
- Using Wearable devices to monitor and prevent health problems
- Optimizing clinic performance through actionable insights

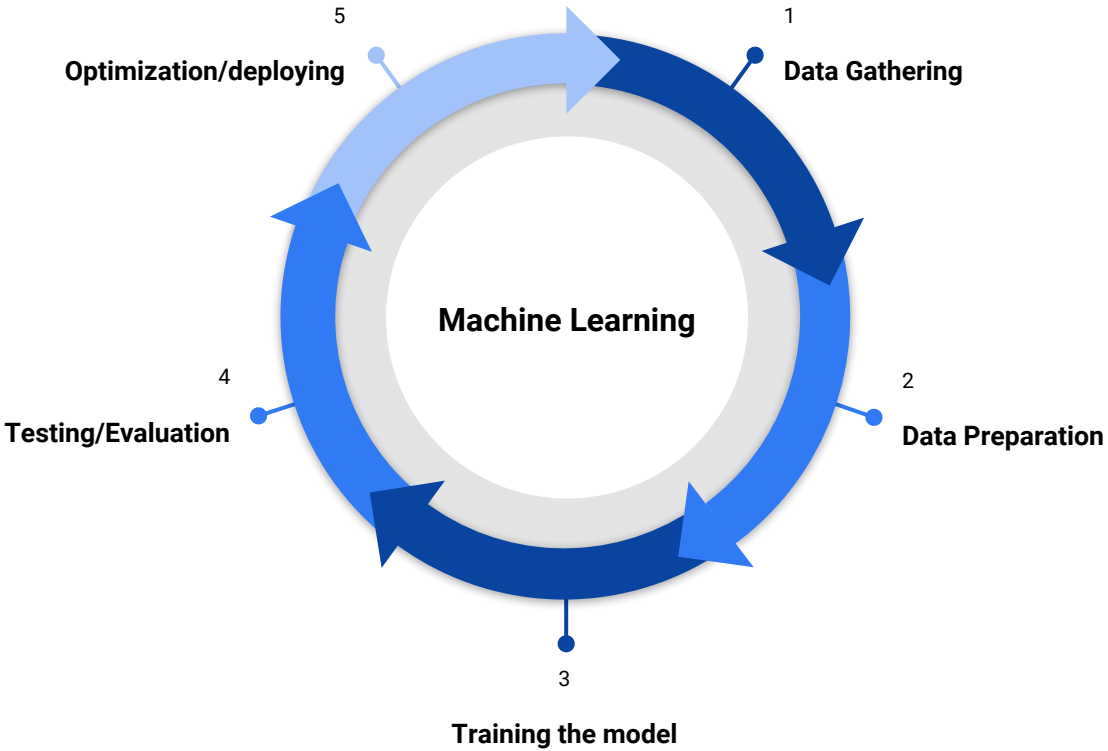
Business

- Intelligent Marketing
- Customer Segmentation
- Enhanced Decision Making
- Churn prediction
- Predicting Customer lifetime value

Machine Learning Workflow

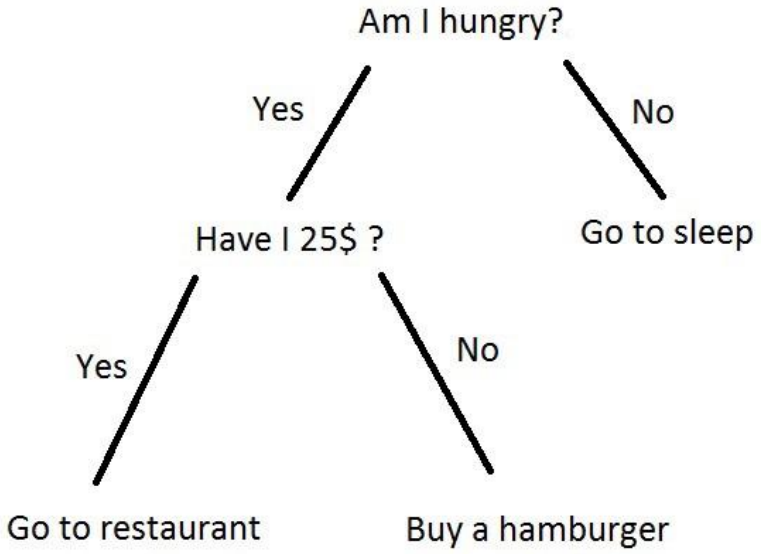


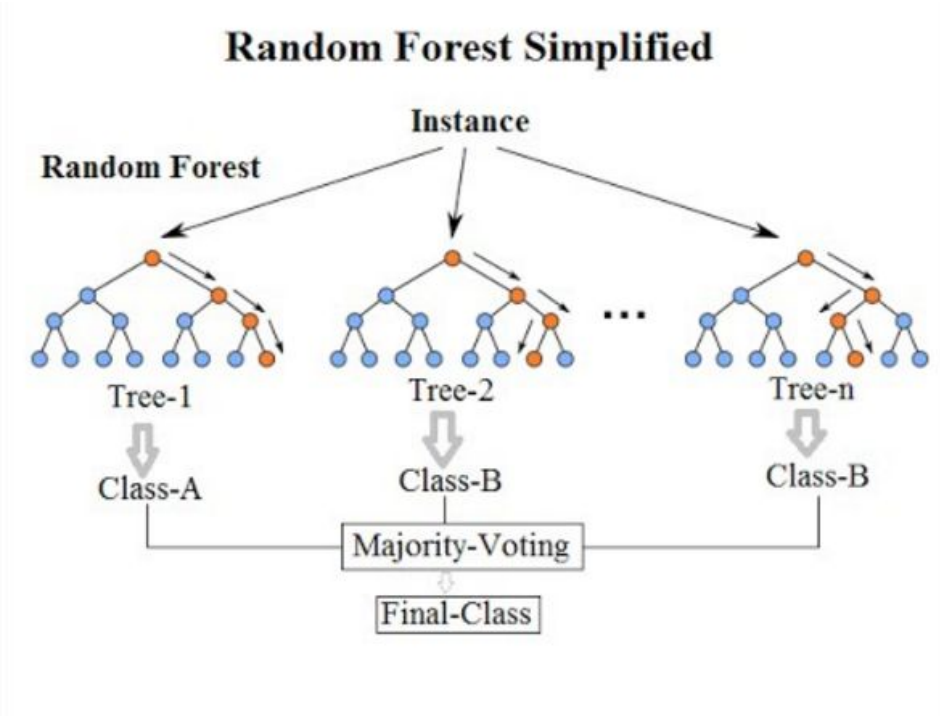
Machine Learning workflow



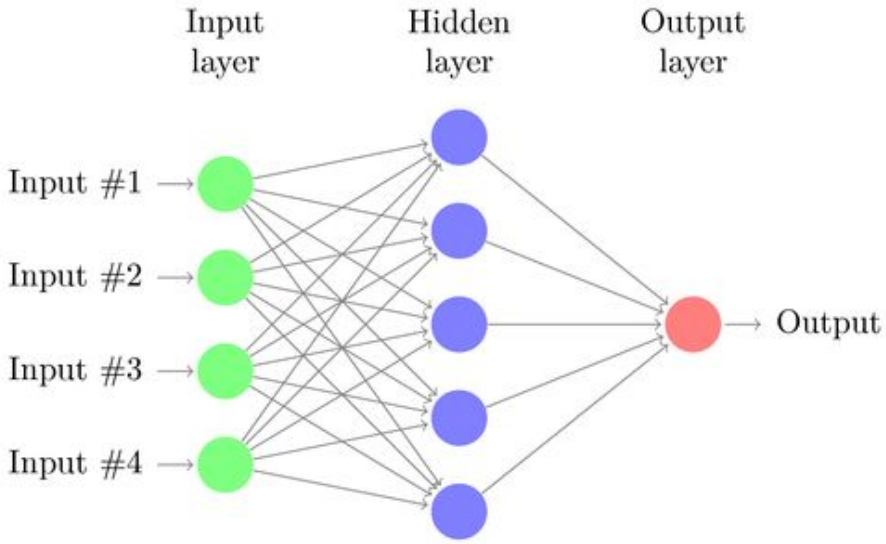
ML Algorithms

Tree Based Algorithms

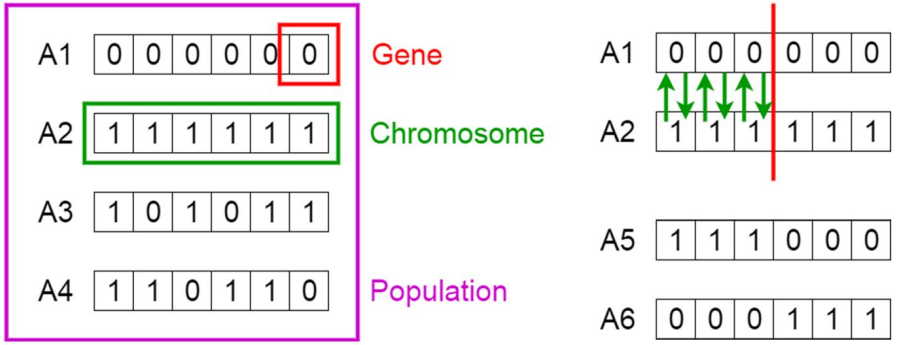




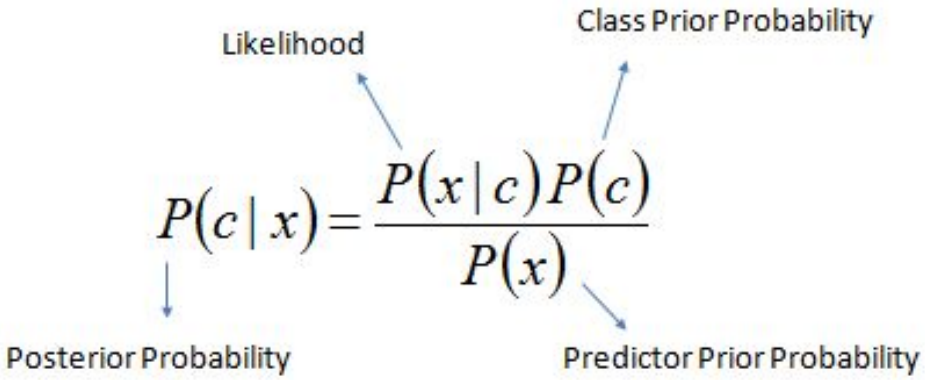
Bio-inspired Algorithms



Genetic Algorithms



Probabilistic



$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$


ALGORITHMS EVERYWHERE
imgflip.com

Data

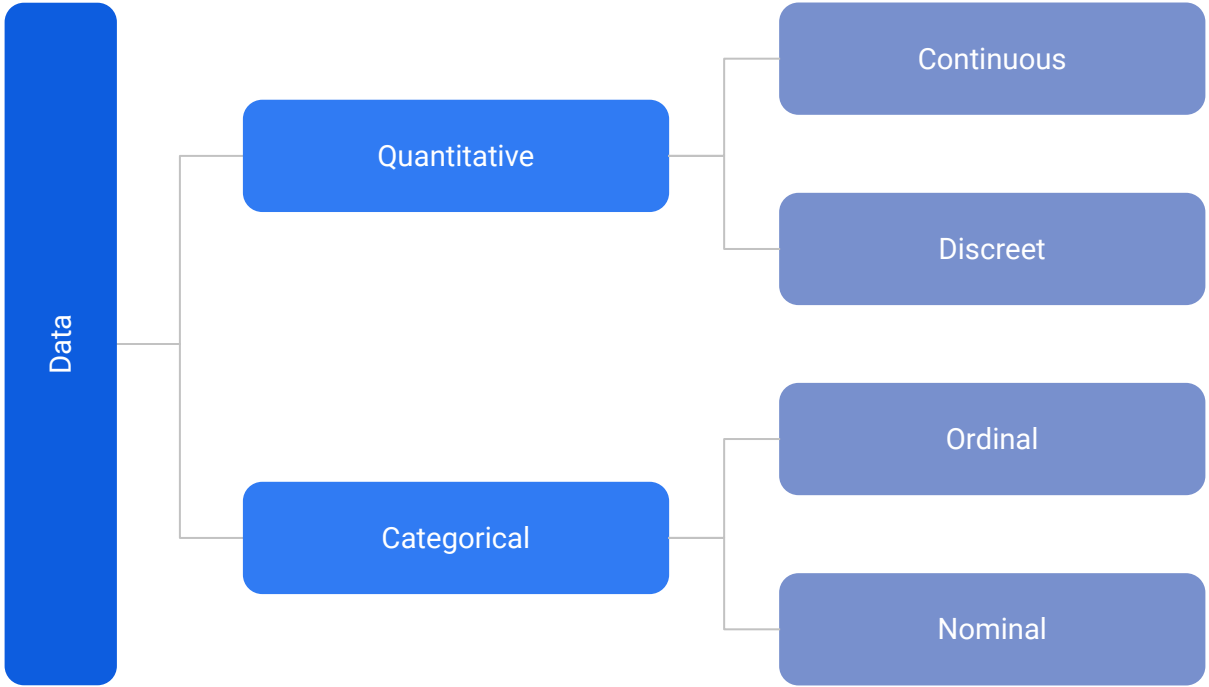
“It is a capital mistake to theorize before one has data.”

— Sherlock Holmes

What is Data?

Information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer:

Types of Data



Quantitative Data

Continuous:

Age

Weight

Time

We can divide it to much smaller units.

Discreet:

The number of people in the room.

How many cars you have.

We cannot say there is 10.5 people in the room, or I have 2.9 cars.

Categorical Data

Ordinal:

- Good/medium/bad
- V. High/high/low/V.Low
- Positive/Neutral/Negative

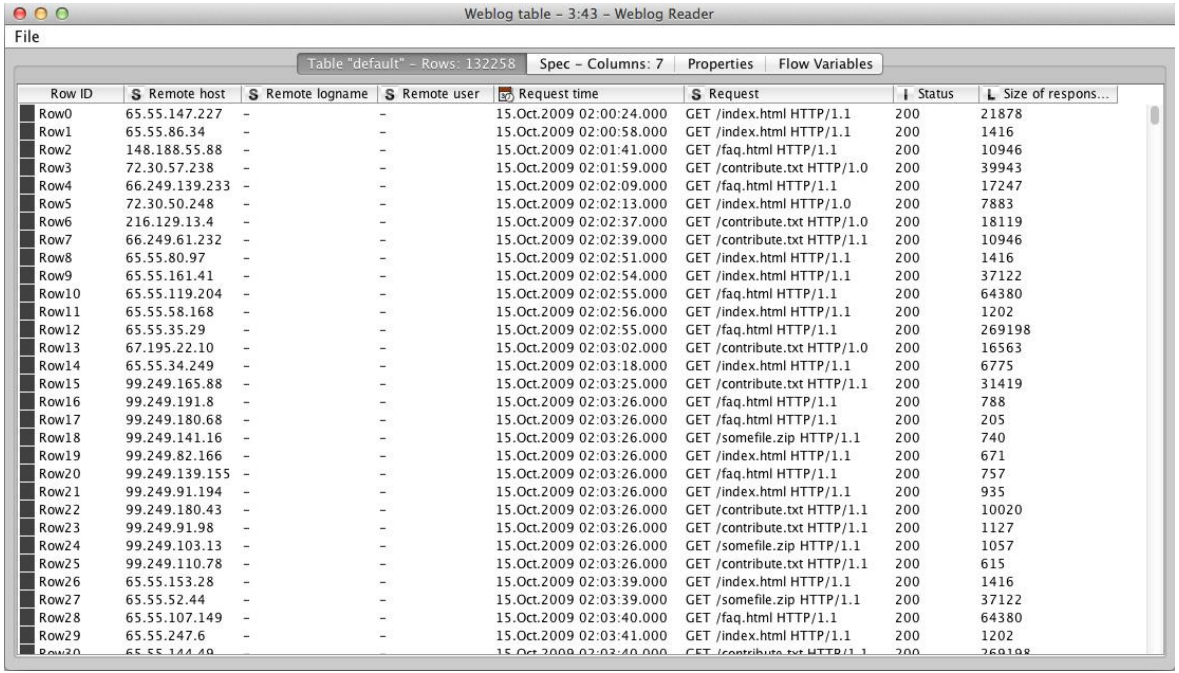
Types that has a hierarchy

Nominal:

- Persian/Siamese/British Shorthair
- Husky/Pug/Bulldog/Chihuahua
- Kampung/D24/Musang King/Udang Merah

Types that has no natural order

Raw Data



Row ID	Remote host	Remote logname	Remote user	Request time	Request	Status	Size of respons...
Row0	65.55.147.227	-	-	15.Oct.2009 02:00:24.000	GET /index.html HTTP/1.1	200	21878
Row1	65.55.86.34	-	-	15.Oct.2009 02:00:58.000	GET /index.html HTTP/1.1	200	1416
Row2	148.188.55.88	-	-	15.Oct.2009 02:01:41.000	GET /faq.html HTTP/1.1	200	10946
Row3	72.30.57.238	-	-	15.Oct.2009 02:01:59.000	GET /contribute.txt HTTP/1.0	200	39943
Row4	66.249.139.233	-	-	15.Oct.2009 02:02:09.000	GET /faq.html HTTP/1.1	200	17247
Row5	72.30.50.248	-	-	15.Oct.2009 02:02:13.000	GET /index.html HTTP/1.0	200	7883
Row6	216.129.13.4	-	-	15.Oct.2009 02:02:37.000	GET /contribute.txt HTTP/1.0	200	18119
Row7	66.249.61.232	-	-	15.Oct.2009 02:02:39.000	GET /contribute.txt HTTP/1.1	200	10946
Row8	65.55.80.97	-	-	15.Oct.2009 02:02:51.000	GET /index.html HTTP/1.1	200	1416
Row9	65.55.161.41	-	-	15.Oct.2009 02:02:54.000	GET /index.html HTTP/1.1	200	37122
Row10	65.55.119.204	-	-	15.Oct.2009 02:02:55.000	GET /faq.html HTTP/1.1	200	64380
Row11	65.55.58.168	-	-	15.Oct.2009 02:02:56.000	GET /index.html HTTP/1.1	200	1202
Row12	65.55.35.29	-	-	15.Oct.2009 02:02:55.000	GET /faq.html HTTP/1.1	200	269198
Row13	67.195.22.10	-	-	15.Oct.2009 02:03:02.000	GET /contribute.txt HTTP/1.0	200	16563
Row14	65.55.34.249	-	-	15.Oct.2009 02:03:18.000	GET /index.html HTTP/1.1	200	6775
Row15	99.249.165.88	-	-	15.Oct.2009 02:03:25.000	GET /contribute.txt HTTP/1.1	200	31419
Row16	99.249.191.8	-	-	15.Oct.2009 02:03:26.000	GET /faq.html HTTP/1.1	200	788
Row17	99.249.180.68	-	-	15.Oct.2009 02:03:26.000	GET /faq.html HTTP/1.1	200	205
Row18	99.249.141.16	-	-	15.Oct.2009 02:03:26.000	GET /somefile.zip HTTP/1.1	200	740
Row19	99.249.82.166	-	-	15.Oct.2009 02:03:26.000	GET /index.html HTTP/1.1	200	671
Row20	99.249.139.155	-	-	15.Oct.2009 02:03:26.000	GET /faq.html HTTP/1.1	200	757
Row21	99.249.91.194	-	-	15.Oct.2009 02:03:26.000	GET /index.html HTTP/1.1	200	935
Row22	99.249.180.43	-	-	15.Oct.2009 02:03:26.000	GET /contribute.txt HTTP/1.1	200	10020
Row23	99.249.91.98	-	-	15.Oct.2009 02:03:26.000	GET /contribute.txt HTTP/1.1	200	1127
Row24	99.249.103.13	-	-	15.Oct.2009 02:03:26.000	GET /somefile.zip HTTP/1.1	200	1057
Row25	99.249.110.78	-	-	15.Oct.2009 02:03:26.000	GET /contribute.txt HTTP/1.1	200	615
Row26	65.55.153.28	-	-	15.Oct.2009 02:03:39.000	GET /index.html HTTP/1.1	200	1416
Row27	65.55.52.44	-	-	15.Oct.2009 02:03:39.000	GET /somefile.zip HTTP/1.1	200	37122
Row28	65.55.107.149	-	-	15.Oct.2009 02:03:40.000	GET /faq.html HTTP/1.1	200	64380
Row29	65.55.247.6	-	-	15.Oct.2009 02:03:41.000	GET /index.html HTTP/1.1	200	1202
Row30	65.55.144.49	-	-	15.Oct.2009 02:03:40.000	GET /contribute.txt HTTP/1.1	200	269198

Cooked Data

Date	Successful Sales	No. of Customers	Value	Rejected Sales
1/7/2018	4543	254	54123	20
2/7/2018	10432	341	23432	33
3/7/2018	2234	543	65321	43
4/7/2018	4123	432	394532	11
5/7/2018	5321	123	20987	23
6/7/2018	543	123	55754	22

Summary of events.

Performance of business

Real World Data Facts

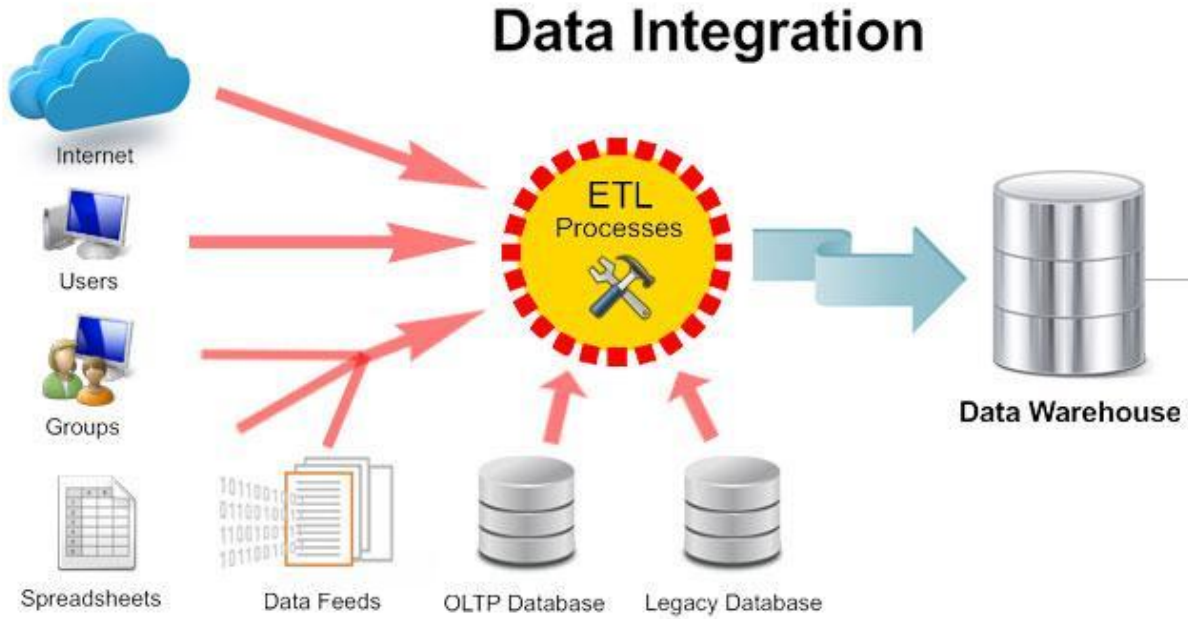
Real world data are generally

- Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
- Noisy: containing errors or outliers
- Inconsistent: containing discrepancies in codes or names.

Tasks in data preprocessing

- **Data integration:** using multiple databases, data cubes, or files.
- **Data cleaning:** fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.
- **Data transformation:** normalization and aggregation.
- **Data reduction:** reducing the volume but producing the same or similar analytical results.
- **Data discretization:** part of data reduction, replacing numerical attributes with nominal ones.

Data Integration



Data Cleaning

1. Handling missing values
2. Identify outliers and smooth out noise.
3. Eliminate inconsistency.

Fill-in with statistical values. (mean/median...etc).

Binning: sort out all values and categorise into 4 bins.

Use domain knowledge to detect inconsistency.

Data Transformation

1. Normalization.
2. Aggregation.
3. Generalization.
4. Attribute Construction.

Scaling values to fall within a specific range. Or scaling by recentering values using the statistical mean and std.

Moving up the concept hierarchy on numeric/nominal attributes.

Re-constructing an attribute based on An old attribute.

Aggregation / Generalization

COUNTRY	STATE	PRODUCT	…
US	CA	A	
US	CA	B	
…			
US	IL	A	
US	IL	C	
US	IL	D	
…			
US	TX	A	
…			
US	CO	D	
US	CO	F	
US	CO	H	
…			
US	NY	A	
US	NY	A	
US	NY	G	
…			

```
SELECT …
 county_id,
 state_id,
 APPROX_COUNT_DISTINCT(product)
FROM sales
GROUP BY county_id,
         state_id
```

COUNTRY	STATE	APPROX. CNT DIST PRODUCT
US	CA	2
US	IL	3
US	TX	1
US	CO	3
US	NY	2

?

COUNTRY	CNT DIST PRODUCT
US	84

```
SELECT …
 county_id,
 APPROX_COUNT_DISTINCT(product)
FROM sales_state_county
GROUP BY county_id
```

Data Reduction

Data reduction is the transformation of numerical or alphabetical digital information derived empirically or experimentally into a corrected, ordered, and simplified form. It can be performed in several ways:

- Reduce the number of attributes (features).
- Reduce the number of attribute values.
- Statistical sampling.

Principal Component Analysis (PCA)

Data Cleaning

1. Handling missing values
2. Identify outliers and smooth out noise.
3. Eliminate inconsistency.

Fill-in with statistical values. (mean/median...etc).

Binning: sort out all values and categorise into 4 bins.

Use domain knowledge to detect inconsistency.

Descriptive Statistics

We use statistics to describe data.

Categorical data is analyzed and described in the aspect of counts and distribution. Measuring the number of people falling into each category.

Quantitative data is measured and analyzed using four aspects:

1. Measures of Center
2. Measures of Spread
3. The Shape of the data.
4. Outliers

Measure of Center

Mean
Median
Mode

Mean is the average value or the expected value in mathematics. Mean is calculated by summing up the values and divide them by the number of the available values.

Median is exactly the center value or the value that is located in the middle of an ordered set of values. What about median of even values?

Mode is the most frequent value in the set.

Measure of Spread

Range
Interquartile Range
Standard Deviation
Variance

Range is simply the difference between the min and max value.

Five Number summary:

1. **Minimum:** The smallest number in the dataset.
2. Q_1: The value such that 25% of the data fall below.
3. Q_2: The value such that 50% of the data fall below.
4. Q_3: The value such that 75% of the data fall below.
5. **Maximum:** The largest value in the dataset.

The interquartile range is calculated as the difference between **Q3** and **Q1**.

Data Visualization

Data visualization is a general term that describes any effort to help people understand the significance of data by placing it in a visual context. Patterns, trends and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization software.

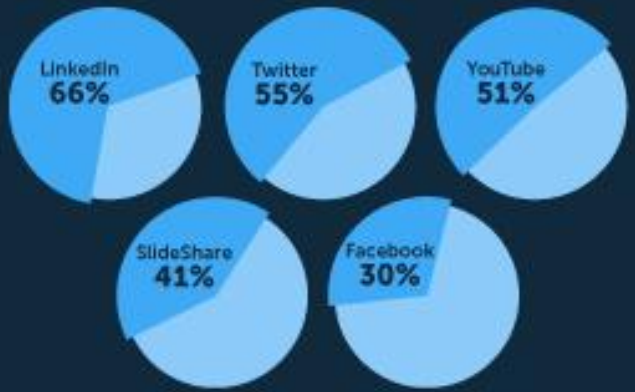
Traditional Vs. Visualization

TRADITIONAL STATISTICS

66% of B2B marketers rank LinkedIn as the most effective social media platform for their business. Other effective platforms include Twitter (55%), followed by YouTube (51%), SlideShare (41%) and Facebook with the smallest percent of marketer's selecting the platform for the most effective platform. (30%)

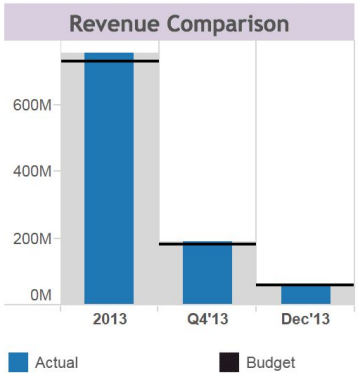
VISUALIZATION STATISTICS

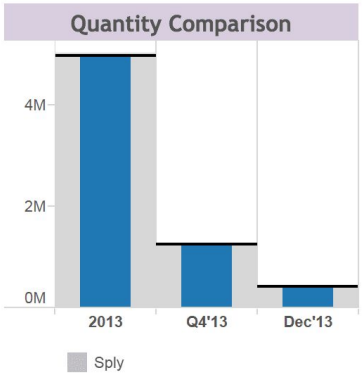
B2B marketer's first choice for most effective social media platforms.

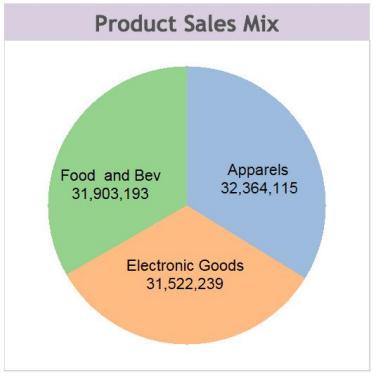


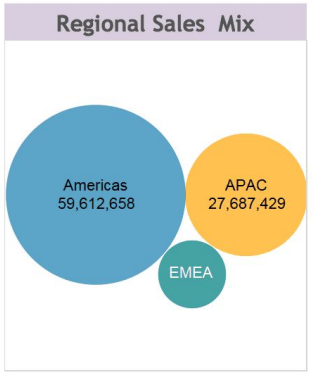
AUNICORE 2016 | SOURCE: FORBES

Sales KPI Performance - Sales Summary



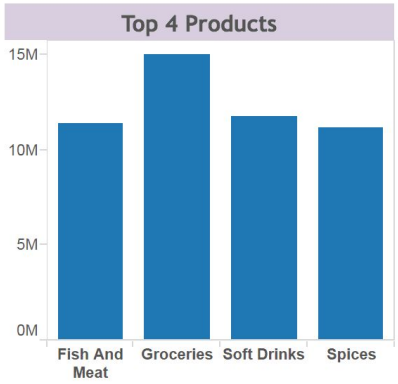


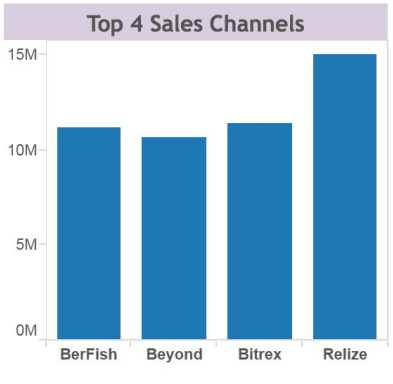


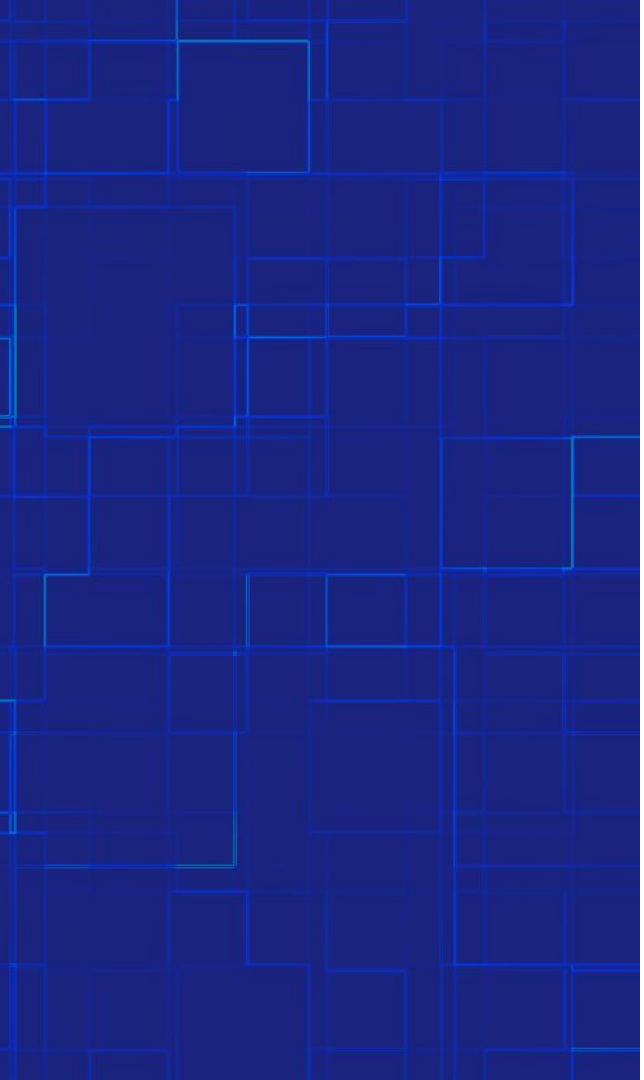


Region Select

- All
- Americas
- APAC
- EMEA







Ask me questions?