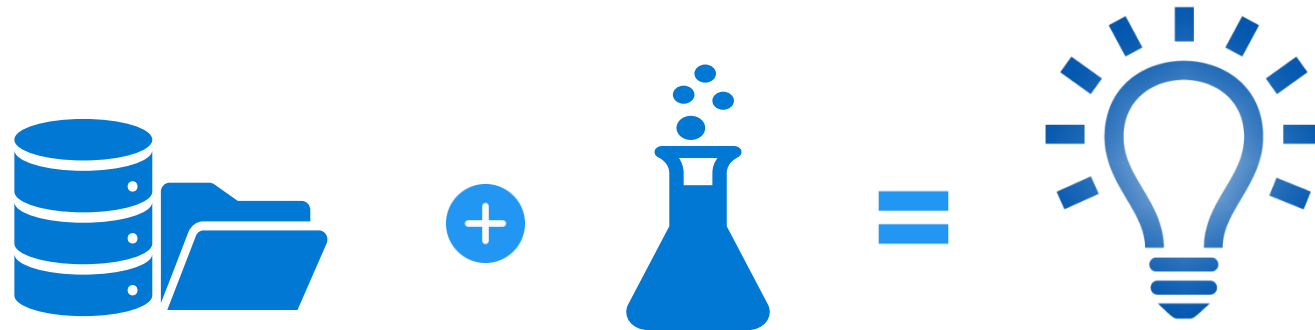

Data Science for Business Leaders



DAY 1 - FLASHBACK

- 1 Goal of Data Science in Telco
- 2 Concept of Data Science
- 3 Types of Data
- 4 Scientific Methods to get Knowledge from Data
- 5 When we need to use which Methods

Apply **Scientific Methods** to extract **Knowledge** from **Data**.

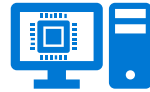


Scientific Methods



Statistics

Designed for inference about the relationships between variables



Machine Learning

Designed to make the most accurate predictions possible



Artificial Intelligence

Designed to mimic human behavior using ML and Deep Learning

DAY 2 - OUTLINE

- 1 Statistical Data Analysis
- 2 Data Analysis Use Case with Excel
- 3 Explore Machine Learning
- 4 Steps of Machine Learning
- 5 Predictive Analytics with Machine Learning

Analysis $\stackrel{?}{=}$ Analytics

Analysis



Past

Explain
How? Why?



STATS/BI

We can use different tools to explain the previous trends like, Power BI, Tableau, QlikView, MicroStrategy etc.

Analytics



Future

Explore potential future events



ML/AI

We can use different language packages and framework to implement ML/AI model.

Business Analytics



Descriptive

What has happened?



Diagnostic

Why did it happen?



Predictive

What will happen next?



Prescriptive

What should I do?

← Looking back

Looking forward →

Case Study

A **credit card company** wants to **reduce** the number of customers defaulting on their payments. They gather historical data on customer transactions, payment history, credit scores, and demographic information. By analyzing this data, they aim to create a model that can forecast the likelihood of a customer defaulting on their next payment. The focus is on developing a predictive model that can identify early warning signs of potential defaults based on patterns and trends observed in the historical data. This analysis will enable the company to proactively intervene and offer assistance to customers at risk of default, thereby minimizing financial losses and maintaining a healthier customer base.

Case Study

A **manufacturing plant** experiences a sudden and **significant drop** in its production output. The operations team gathers data on machine performance, maintenance logs, and production schedules. By analyzing this data, they aim to pinpoint the exact factors that led to the production decline. The focus is on identifying any equipment malfunctions, breakdowns, or operational bottlenecks that might have contributed to the drop in output. This analysis will help the team determine the root causes of the issue and develop strategies to address the problem promptly, ensuring the plant returns to its optimal production levels.

Case Study

An e-commerce company is analyzing its sales data from the past year. The company's data team compiles information on product purchases, order dates, customer locations, and purchase amounts. By examining this data, the company aims to uncover trends and patterns in its sales performance. The focus is on understanding which products sold well during specific periods, identifying peak buying times, and discerning whether there are any geographical preferences among customers. This analysis will guide the company in making informed decisions about inventory management, marketing strategies, and potential expansions into new markets.

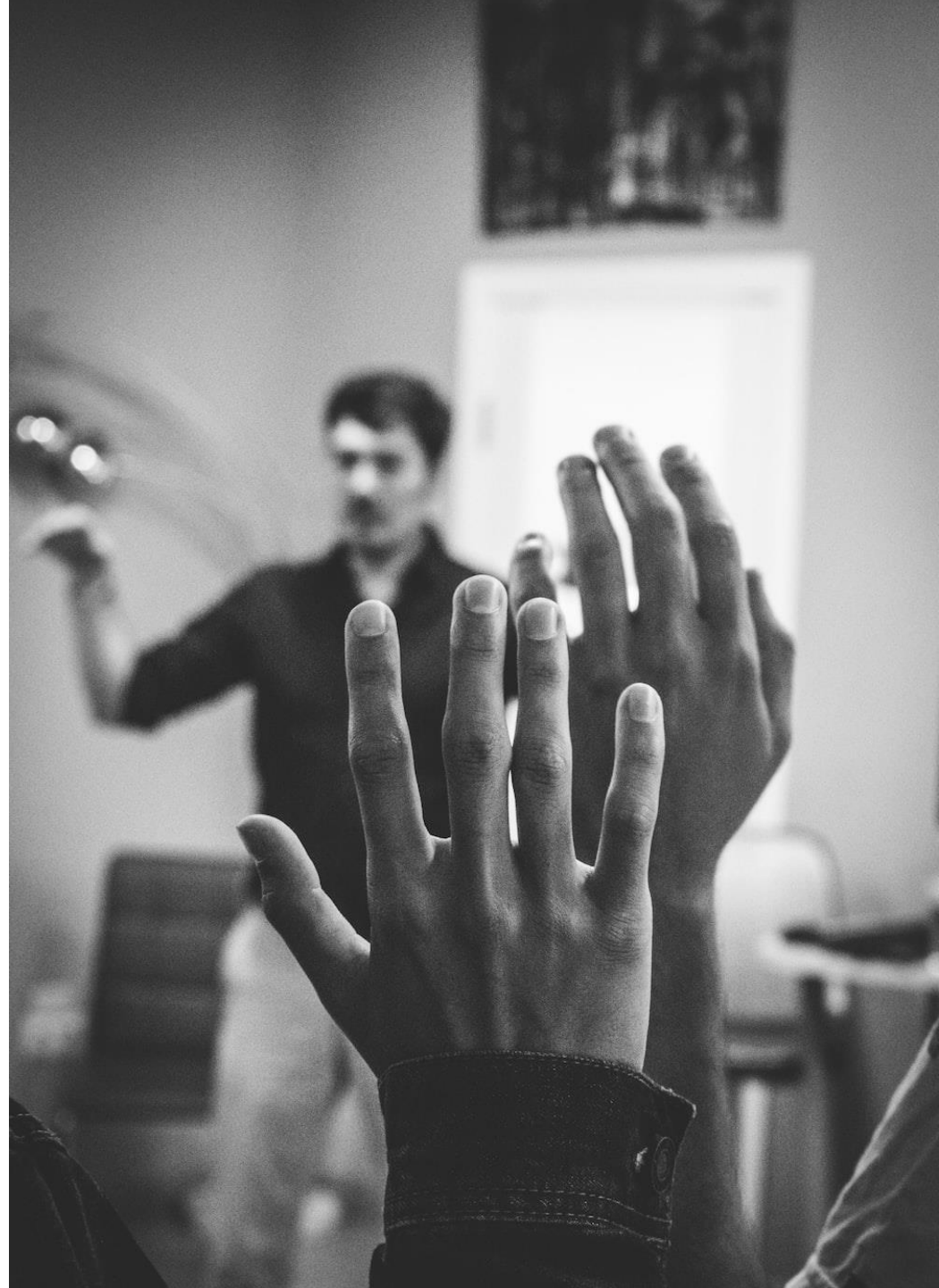
Case Study

A **healthcare provider** notices an **increase** in patient readmissions for a specific chronic condition. They gather data on patient medical histories, treatment plans, medications prescribed, and post-discharge follow-up procedures. By analyzing this data, they aim to develop a system that recommends personalized treatment plans for patients with the identified chronic condition. The focus is on creating a solution that can predict potential complications based on historical patient data and suggest optimal interventions to prevent readmissions. This analysis will guide healthcare professionals in making more informed decisions about patient care, ultimately reducing readmission rates and improving overall patient outcomes.

Data-Driven Decision Making

1

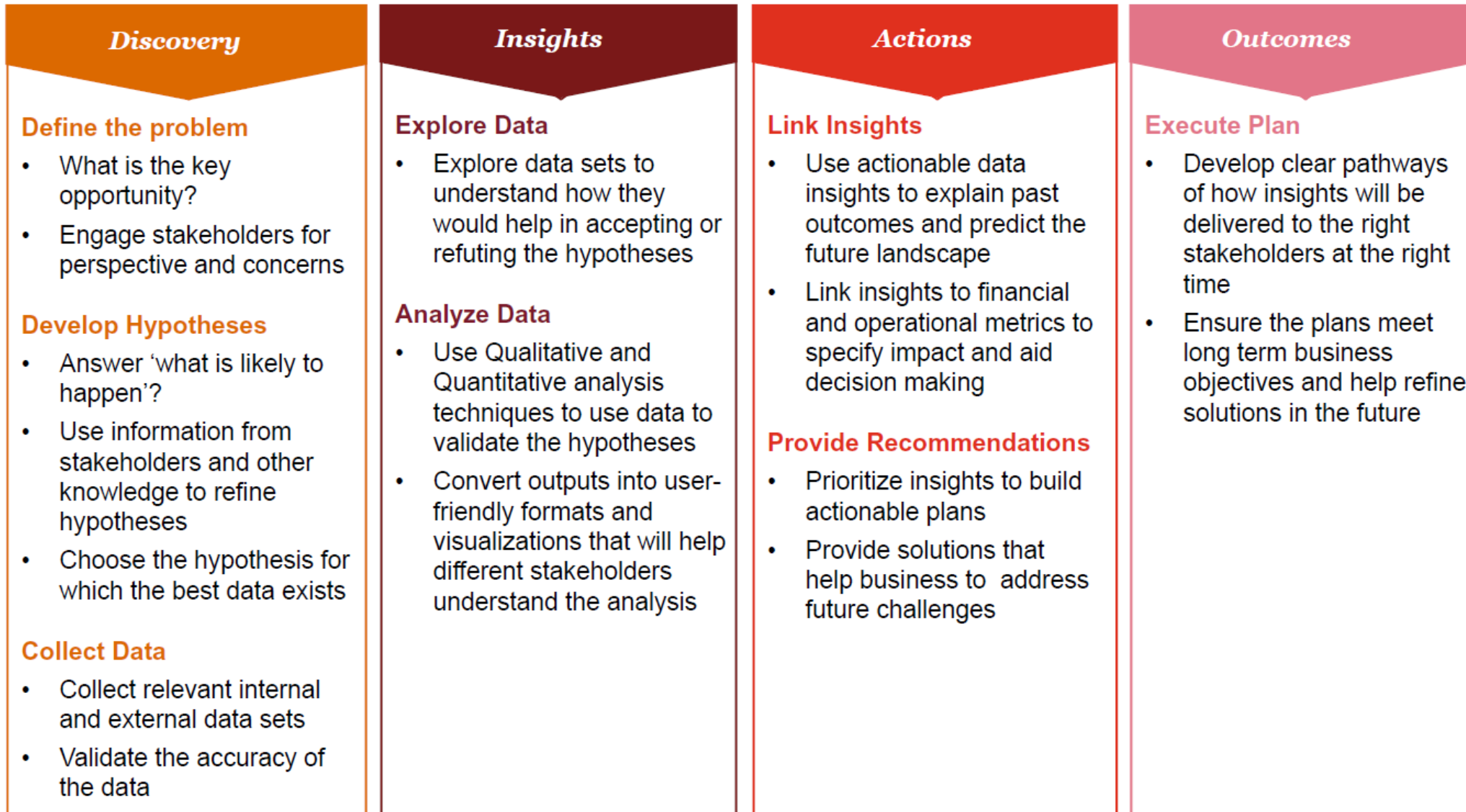
Ask **Questions** to
your Data !



Data and Analytics Framework



Putting the Framework into Action



Data Driven Decision Making - Simulation Exercise

<https://bit.ly/3dm-blink>



2

Understand your **Numbers !**



Statistics



Descriptive Statistics

Understand what your sample data looks like



Probability Distributions

If the sample data fits a probability distribution, use it as a **model** for the entire population

$$\leftarrow \mu \rightarrow$$

Confidence Intervals

If the sample doesn't fit a distribution, use the central limit theorem to make **estimates** about population parameters



Hypothesis Tests

Continue to leverage the central limit theorem to draw **conclusions** about what a population looks like based on a sample



Regression Analysis

Use additional variables to increase the accuracy of your estimates and make **predictions** based on their relationships

MAVEN PIZZA PARLOR | PROJECT BRIEF



You are a BI Consultant that has just been approached by **Maven Pizza Parlor**, a new pizza place in New Jersey that needs help with their demand planning



From: **Mary Margherita** (*Owner*)

Subject: **Daily Pizza Sales**


Hi!

We we're extract our daily pizza sales from our POS system, and we want to use this for planning, but none in the team is data savvy.

Is that something you could help us with?

We want to know how many pizza sales to expect every day, how much they typically vary, and if they fluctuate by day of the week.

Thank you!

 Pizza_Sales.xlsx

 Reply

 Forward

MAVEN PIZZA PARLOR | PROJECT BRIEF



You are a BI Consultant that has just been approached by **Maven Pizza Parlor**, a new pizza place in New Jersey that needs help with their demand planning



From: **Mary Margherita** (*Owner*)

Subject: **Daily Pizza Sales**

Hi!

We we're extract our daily pizza sales from our POS system, and we want to use this for planning, but none in the team is data savvy.

Is that something you could help us with?

We want to know how many pizza sales to expect every day, how much they typically vary, and if they fluctuate by day of the week.

Thank you!

 Pizza_Sales.xlsx

 Reply

 Forward

Key Objectives

1. Summarize the daily pizza sales by using descriptive statistics

MAVEN MEDICAL CENTER | PROJECT BRIEF



You are a Data Analyst at the **Maven Medical Center** in Springfield, MA and just received a project request from the chief gynecologist



From: **Betty Birth** (*Chief Gynecologist*)


Subject: **Need some probability figures**

Good morning!

We've had over 30% of the babies born this year weigh under 2.5kg, which is considered low. The percentage itself seems a little high to me though. Is there any way you could check what the probability of a baby weighing under 2.5kg is with the data we have?

I could also use the number of births we've had so far in the top & bottom 1% if possible.

Thank you!

 Birth_Weights.xlsx

 Reply

 Forward

MAVEN MEDICAL CENTER | PROJECT BRIEF



You are a Data Analyst at the **Maven Medical Center** in Springfield, MA and just received a project request from the chief gynecologist



From: **Betty Birth** (*Chief Gynecologist*)


Subject: **Need some probability figures**

Good morning!

We've had over 30% of the babies born this year weigh under 2.5kg, which is considered low. The percentage itself seems a little high to me though. Is there any way you could check what the probability of a baby weighing under 2.5kg is with the data we have?

I could also use the number of births we've had so far in the top & bottom 1% if possible.

Thank you!

 Birth_Weights.xlsx

 Reply

 Forward

Key Objectives

1. Check if the weights can be assumed to follow a normal distribution
2. If so, calculate the probability of a baby weighing 2.5kg or less
3. Estimate the values at the 1% and 99% cumulative probabilities
4. Count the number of births under and over those thresholds

3

Visualize your Data!



A row of wooden figures, with one red figure in the center. The background is a blurred grey.

HR Analysis: Employee Retention

Employee Turnover Analysis

Problem Statement: Management wants to understand how to reduce employee turnover.

Goal: HR wants to create an employee retention program.

Task: Analysis, hypothesis and data story on reasons for churn.

Data: ~15,000 employee records.

Questions from Management:

- What is the main cause of turnover?
- Is there something surprising in the data?
- What segment should we focus on?
- Which department has the highest turnover?
- Do we need to increase X or decrease X?
- Where should we put our pilot program

Insight Development

How to develop insights?(W.H.W)

1. What's the goals of the business?

Make money/reduce employee churn/limit recruitment cost

2. What is the metric of success or failure?

Employee retention/churn

3. What are the trends?(positive or negative)

Departments with high and low churn

4. What influences our metrics and trends?

Other metrics' affect on churn

5. How can we fix the trends?

Lowering/increasing X may lower or increase

Tools & Techniques

Tools: Excel and PowerPoint

Techniques: Pivot Table, Power Query, DAX

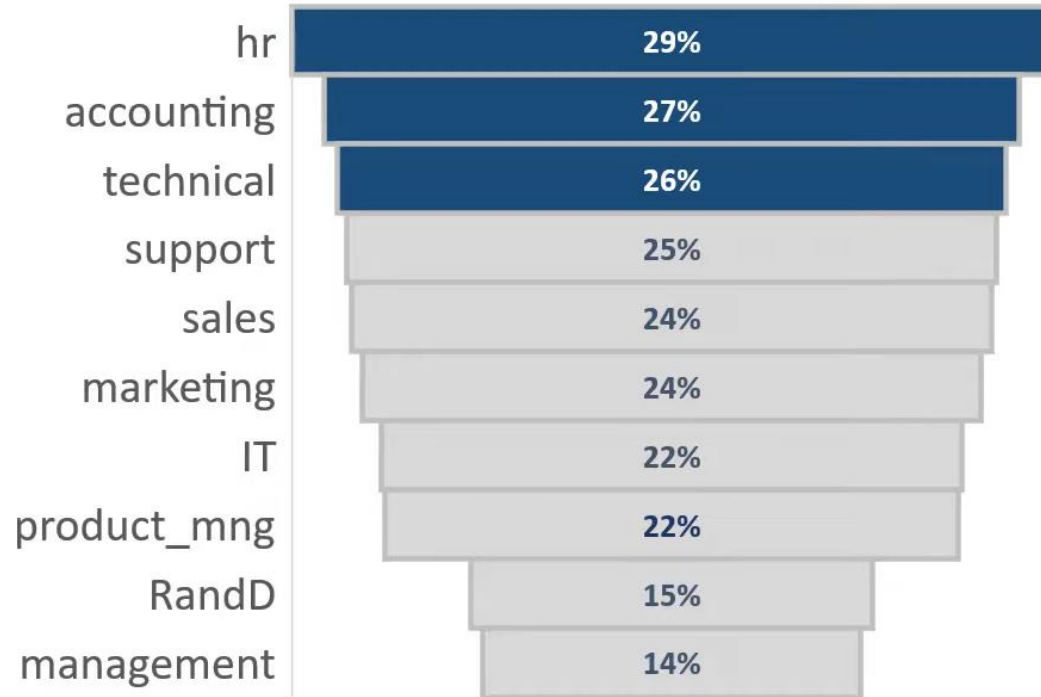
Statistics: Mean, Median, Sum, Count, Percentile

Visuals: Stacked Bar, Boxplot, Funnels, Pie Charts

Where Do We Have the Most Churn?

24%

Company Turnover



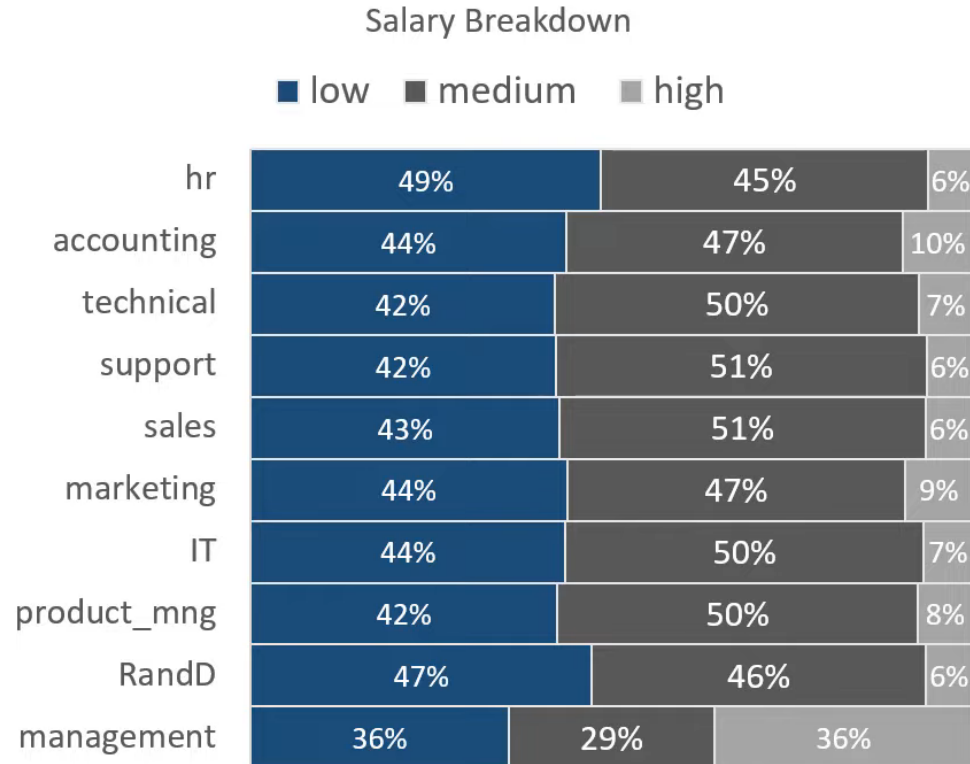
DEPARTMENT TURNOVER

These departments have the most churn. However, we need to ask what is the representation of these departments in the company and what is driving this churn?

Does Salary Affect Employee Retention?

High Churn & Low Salary

The departments with the most churn also have the most employee in the low salary range.

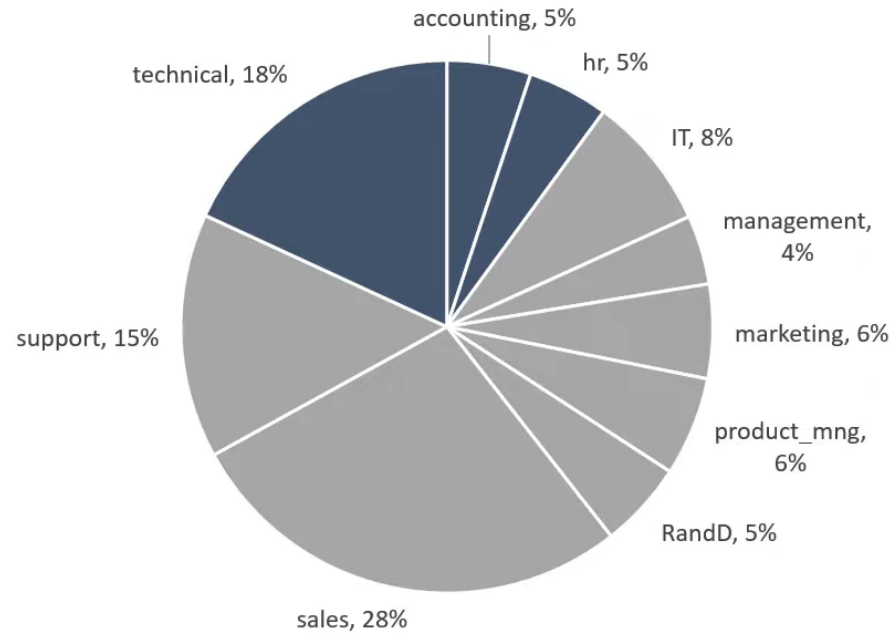


Salary

Although salary are lower for the top 3 departments with the lowest retention. Not all the categories have the lowest salaries. However, high medium and high salaries do show greater retention.

Does Salary Affect Employee Retention?

Where are most employees concentrated?

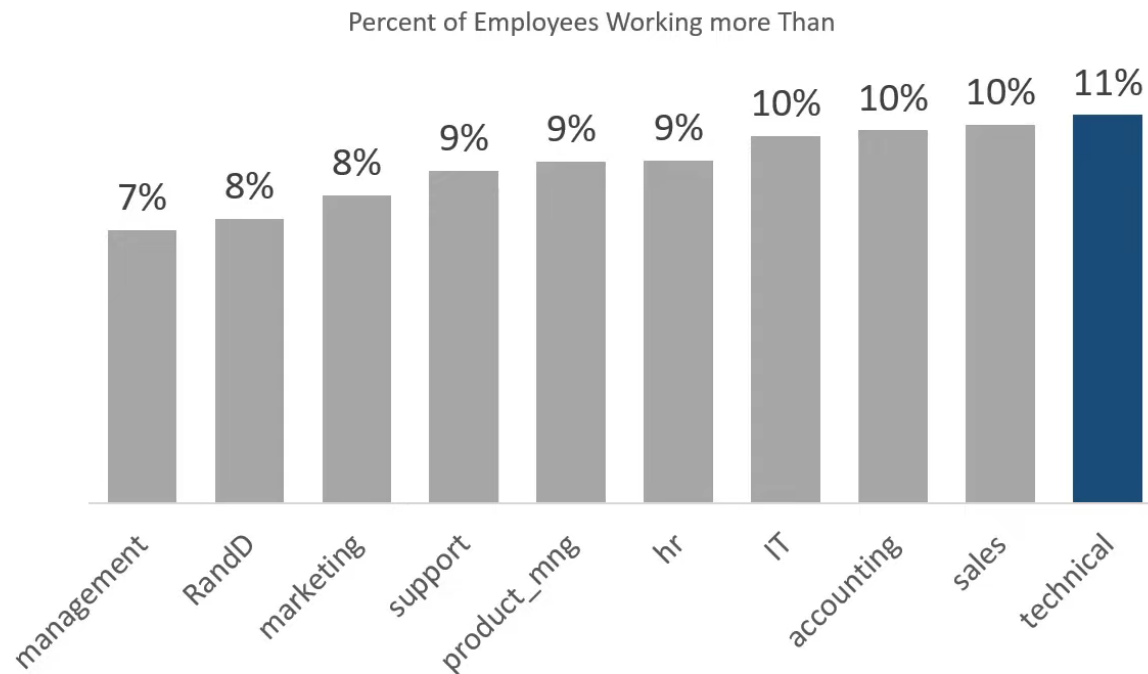


Top 3 Departments by Churn and Employees

Although these departments have the most churn there don't necessarily equal large volume of employee. However, we should evaluate these departments difficult in recruitment.

Is Working Long Hours Affected Churn ?

Who is working the longest hours ?



Top 3 Departments by Churn and Long Hours

When evaluating the long hours outliers which would be at the 90th percentile. It's easy to determine that the technical department has the highest amount of employee in this segment.

Summary & Recommendations

Summary:

The **overall churn of the companies sits at 24%**. This indicates that there may be an issue since the **industry average is between 12 and 15%**.

We have identified 3 candidates for a pilot program who have the highest churn. Out three segments, the **technical has the greatest number of employees at 18% at churn of 26% while HR(29% churn) and accounting(26 % churn) make up 5% of employee each, respectively.**

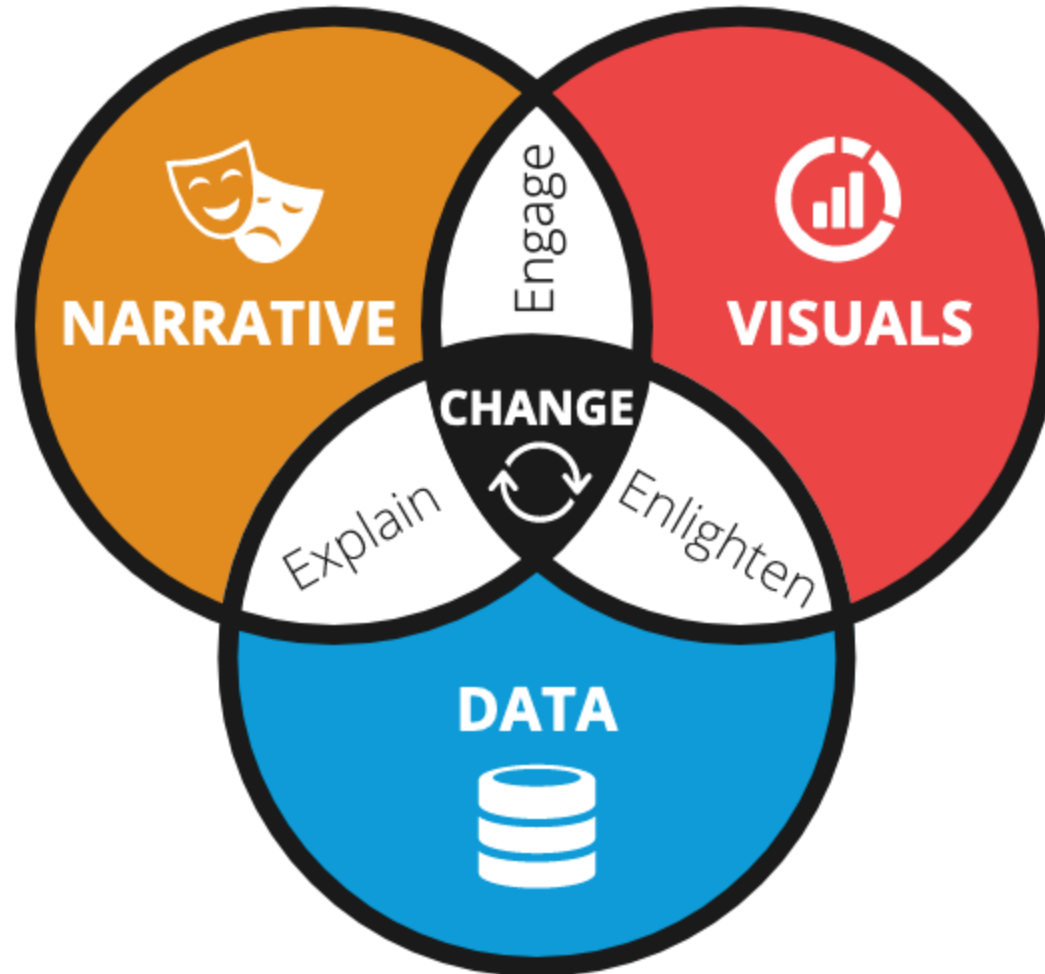
Salary and work hours may factor into the department churn with these segments having the majority of employees in the low and mid salary ranges. **Technical employees have 11% of employee working more than 267 hours month or more per month. This would be the best candidate for the pilot program.**

4

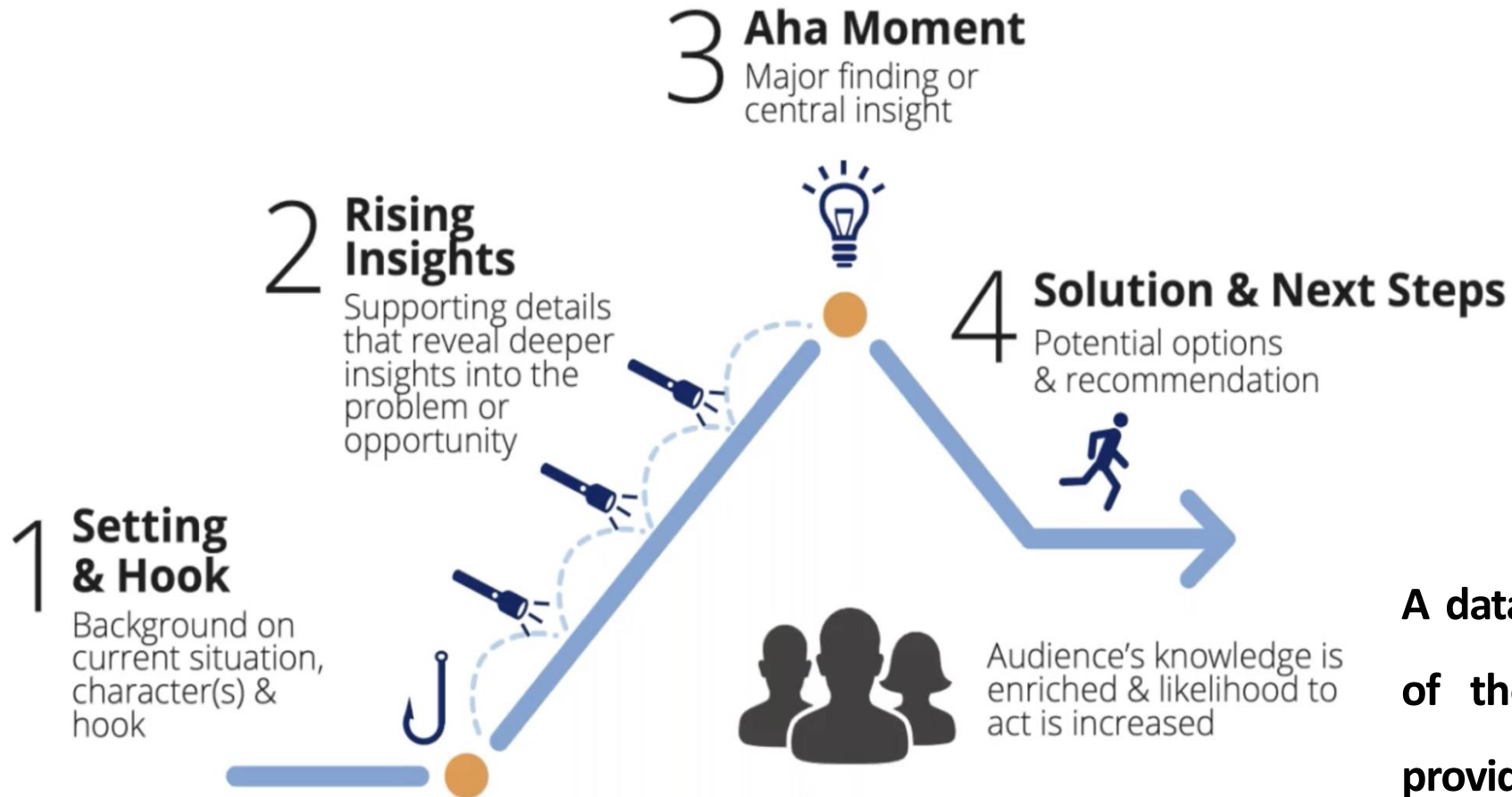
Data **Storytelling** Connect Dots!



Telling Effective Data Stories with Narrative, Data, and Visuals



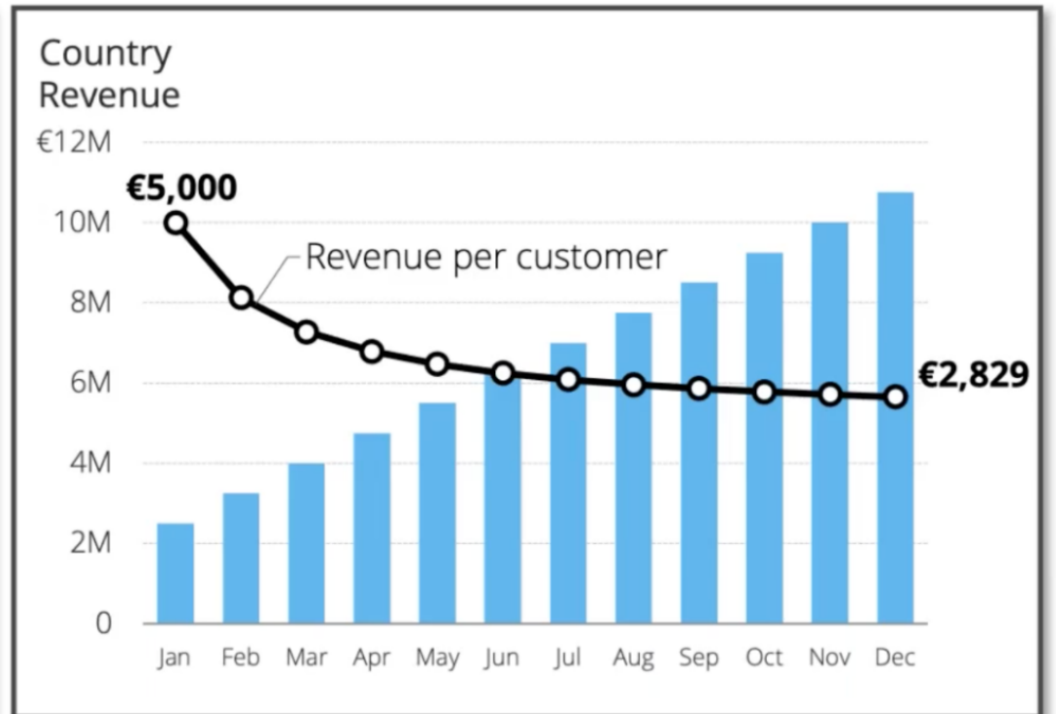
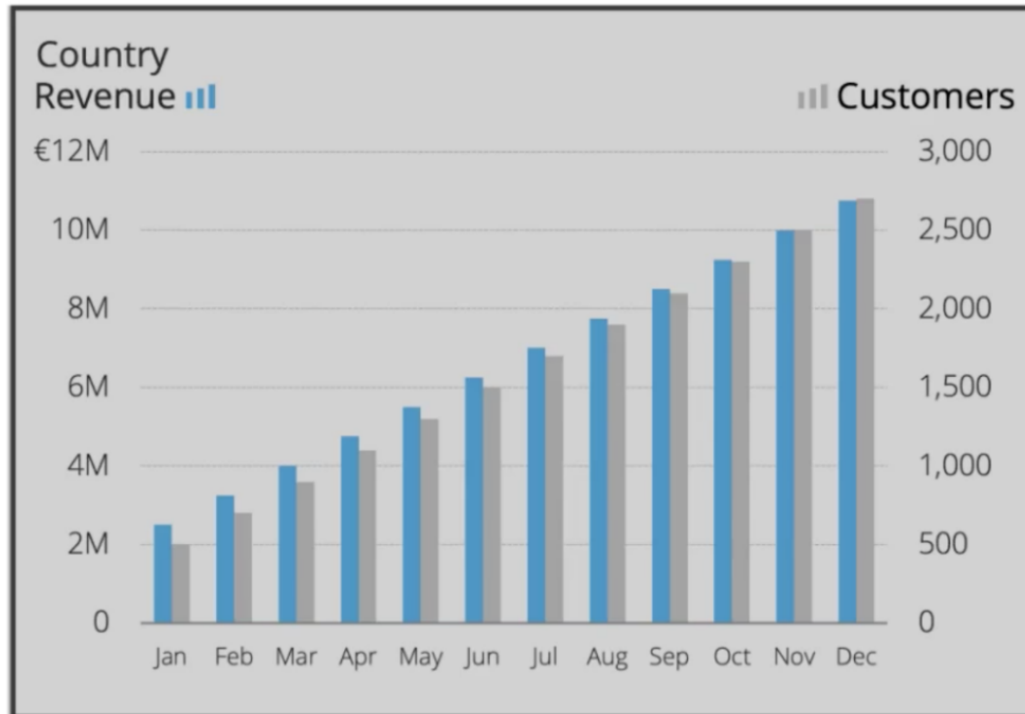
Narrative Structure



A data story **begins** by setting the scene of the current situation, **proceeds** by providing insights that **lead** up to the central insight, and **ends** with relevant recommendations.

Identify Right Data for Your Data Story

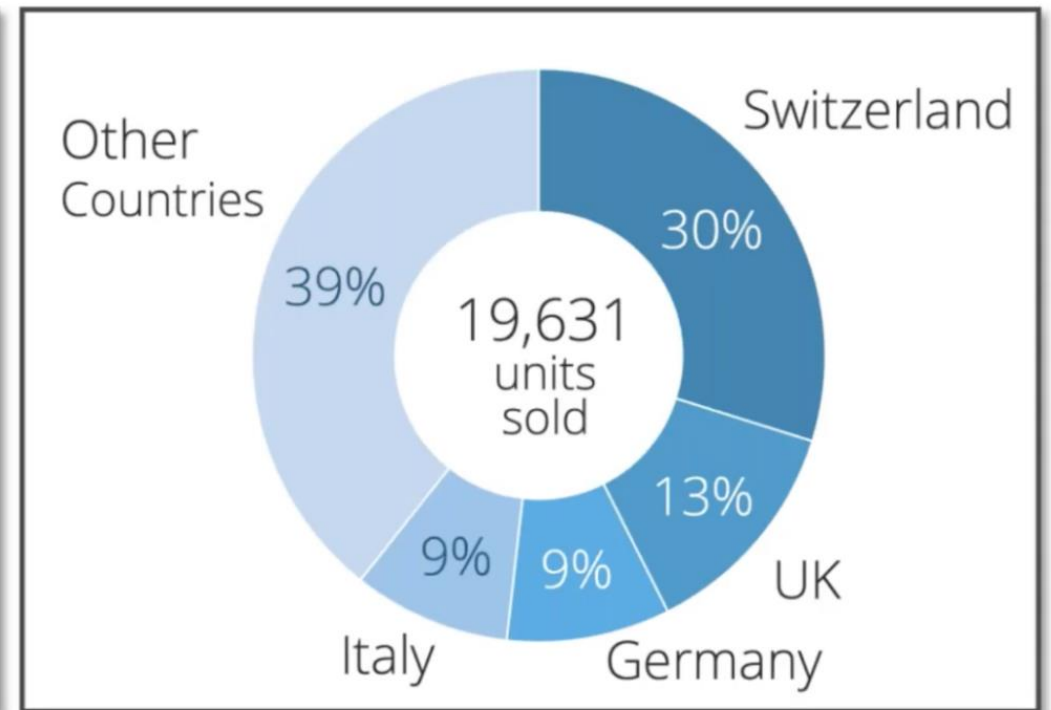
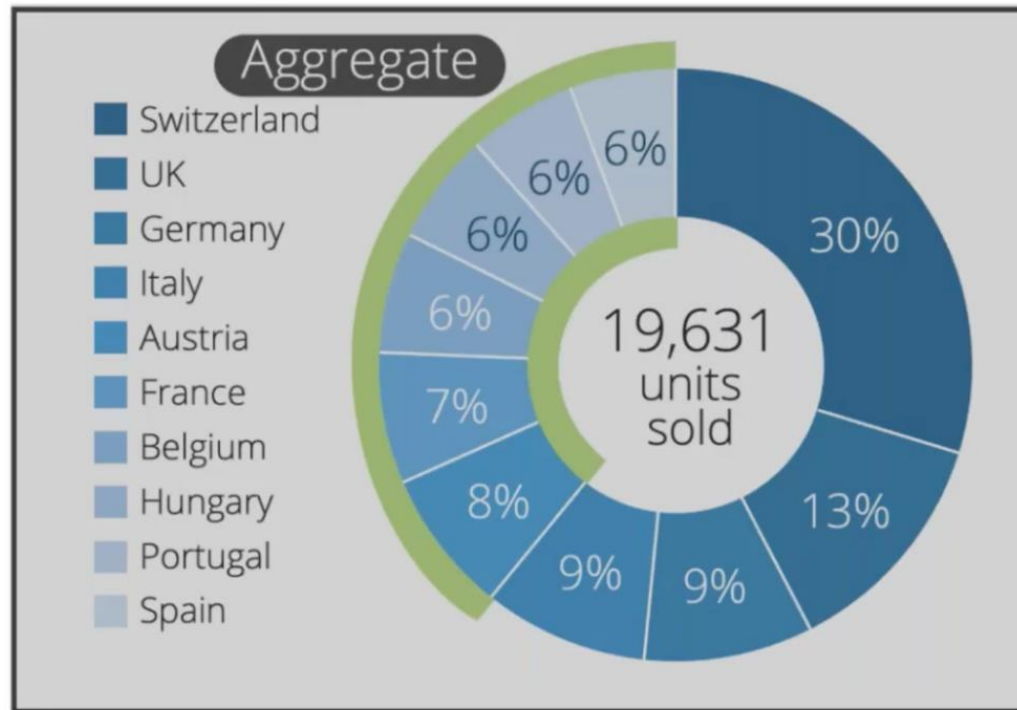
Calculated metrics may be more insightful than total values.



Explicitly demonstrating that the revenue per customer is falling (**right**) is a better choice than plotting the total revenue and customer side-by-side (**left**)

Aggregate Less Important Information

To simplify charts, you can **aggregate less critical data** to reduce the cognitive load.



The market shares of the largest markets become apparent when the smallest markets are aggregated.

5

Data **Modelling** Mingle your Data!





FACT TABLE

VS



DIMENSION TABLE

Fact Table

Order ID	Date	Product ID	Customer ID	Quantity	Price	Total Order Amount
123456	12-04-2000	1555	4564	3	1000	3000

Dimension Table

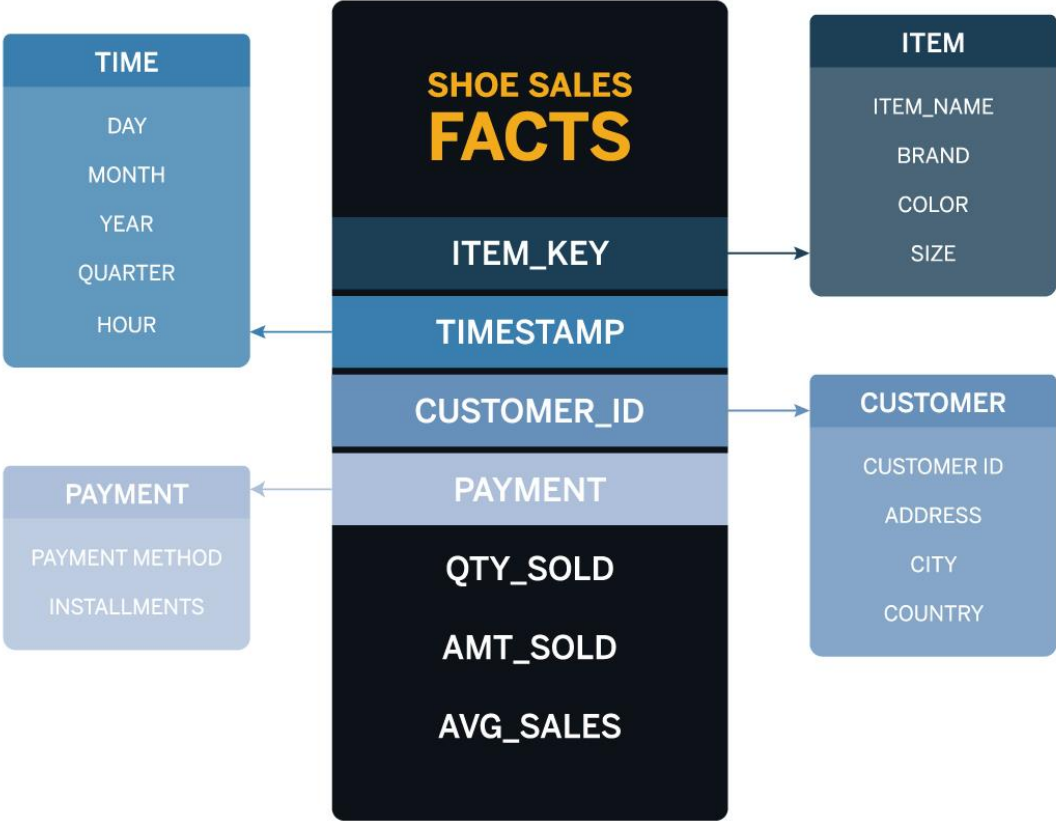
Product ID	Product Name	Category	Sub-Category	Brand	Price
1555	Chair	Furniture	Household	ABC	1000

Fact vs Dimension Table

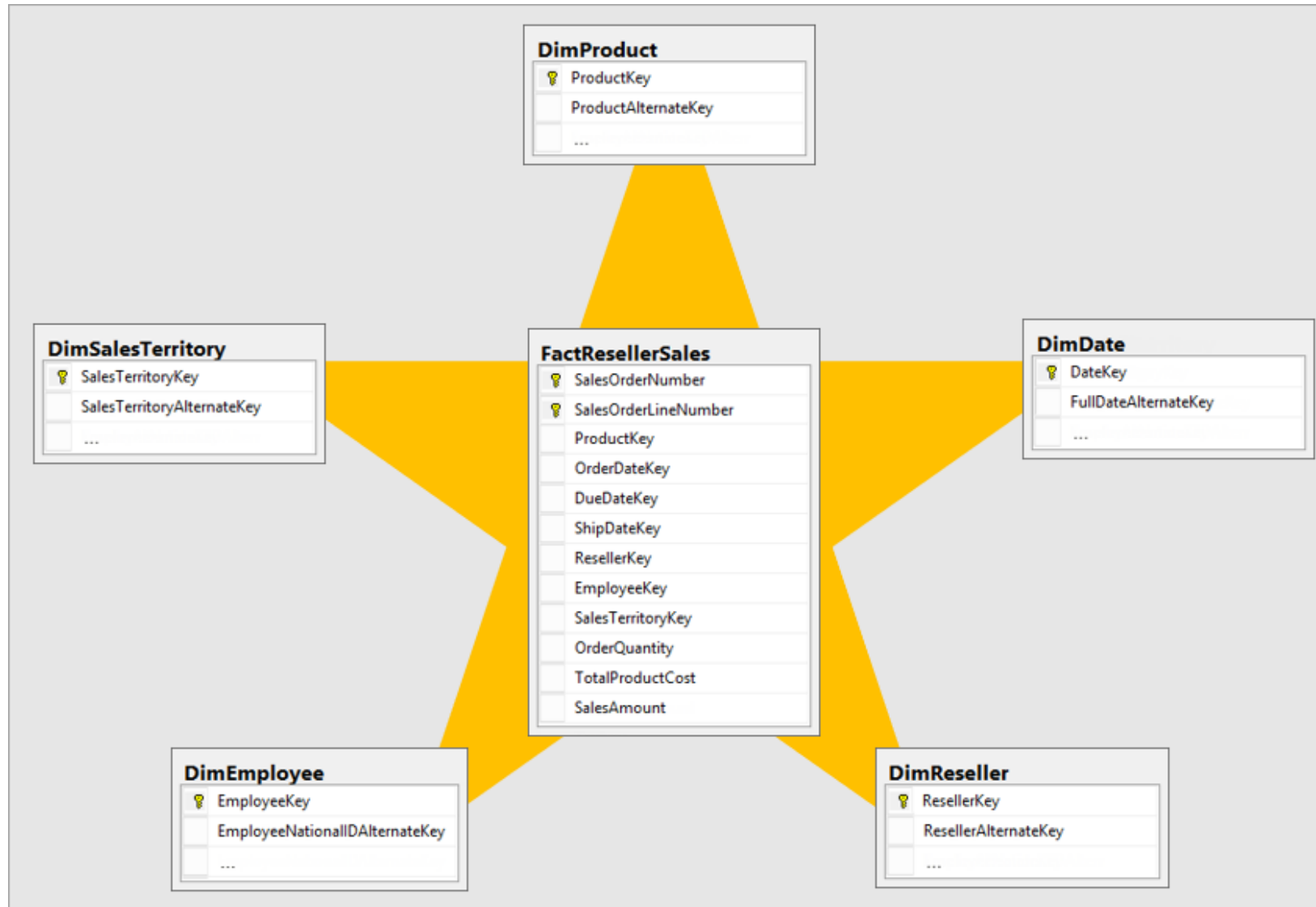
Basis	Fact Table	Dimension Table
Contents	Numeric values and transactional data.	Categorical data and descriptive attributes.
Purpose	Stores quantitative measures and metrics.	Provides descriptive attributes and context.
Size	Larger in terms of data volume.	Smaller in terms of data volume.
Aggregation	Aggregates data for analysis and reporting.	Provides context for data aggregation.
Querying	Provides data for analysis and calculations.	Used for filtering and categorization
Examples	Sales transactions and inventory levels.	Date, product, store, and customer dimensions
Rows	Many rows.	Fewer rows.

Star Schema

BEST RUN SHOES



Star Schema



Let's do that in using **Power Pivot**



Let's Play

www.kahoot.it

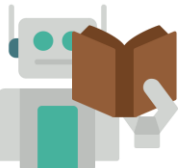
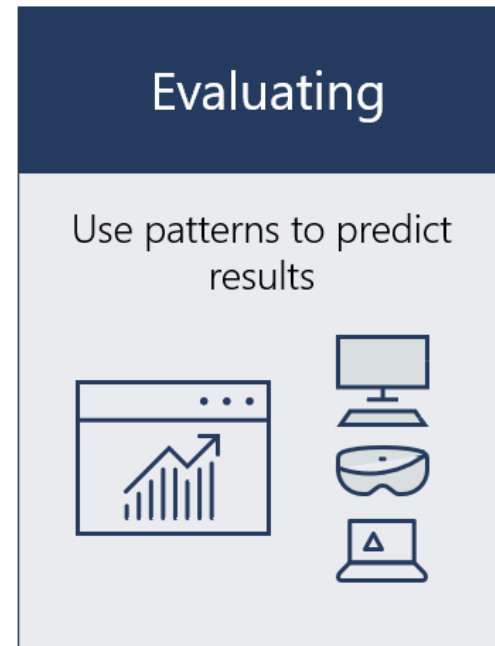
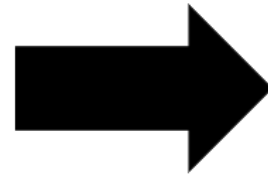
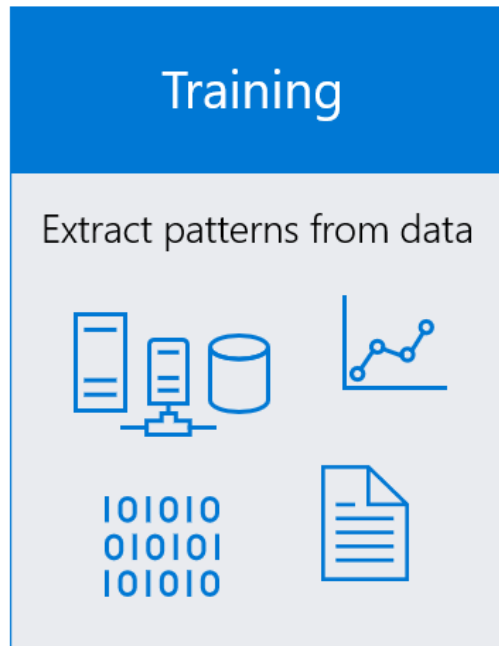
6

Predictive Analytics – with ML

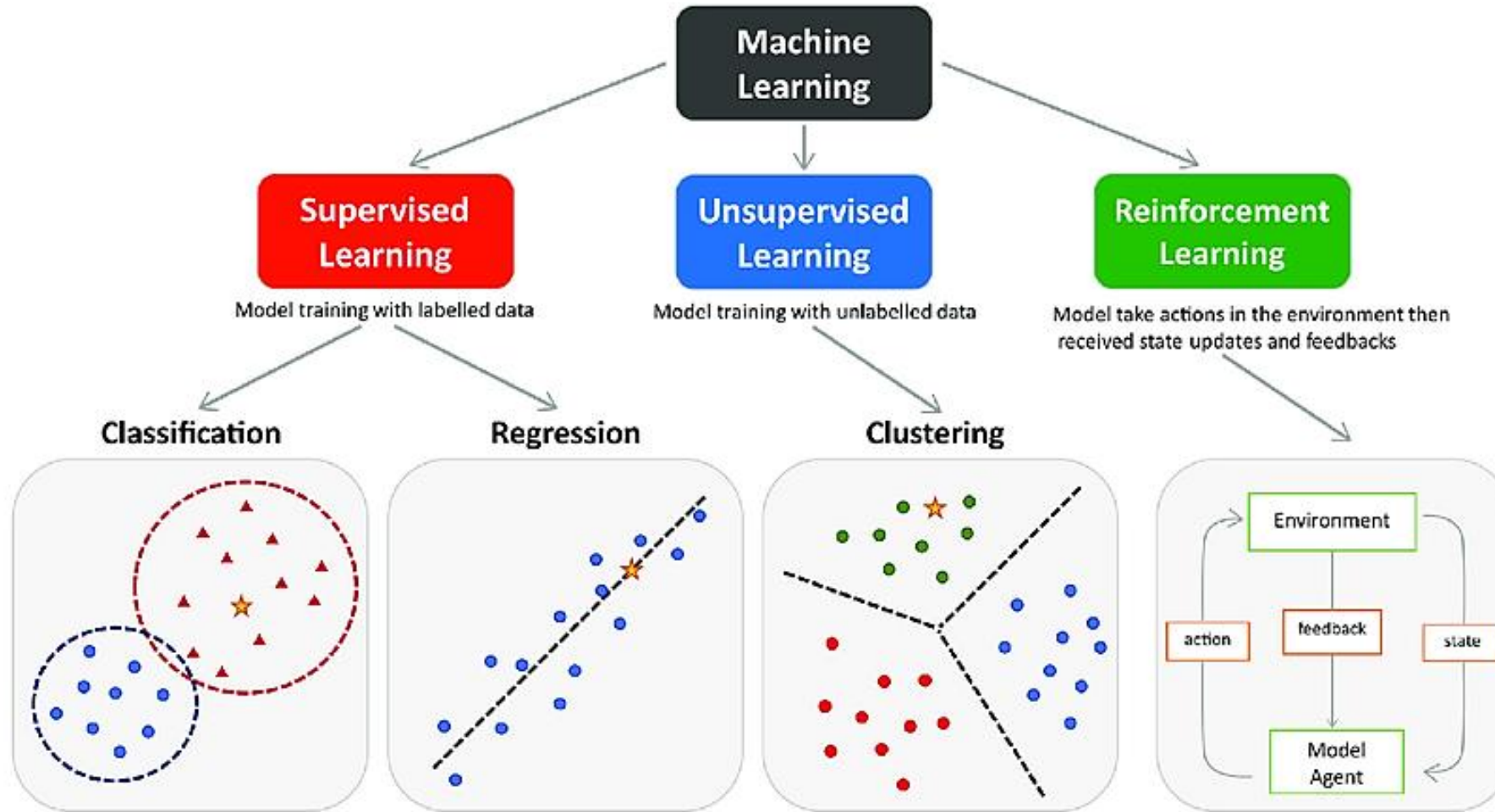


Machine Learning

Machine (computer) tries to find the pattern (self-learn) from the data.



Types of Machine Learning



Churn Prediction

Customer ID	Age	Monthly Charges	Contract Length	Data Usage	Customer Service Calls	Churn (Target)
1	30	50	12	150 GB	2	No
2	45	75	24	300 GB	1	No
3	22	35	6	50 GB	3	Yes
4	55	60	12	200 GB	2	No
5	40	85	24	250 GB	1	Yes
...

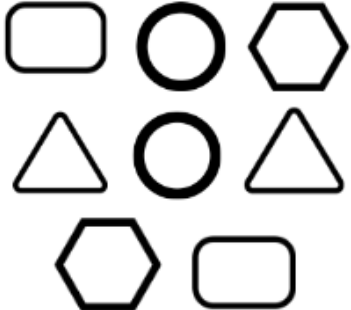
Predicting customer churn can help companies proactively address customer issues, improve service quality, and implement targeted retention strategies. Supervised machine learning algorithms can be employed to build predictive models that identify customers at risk of churning.

⚙️ Set of inputs ~ [Features] / [Independent Variables] / [X]

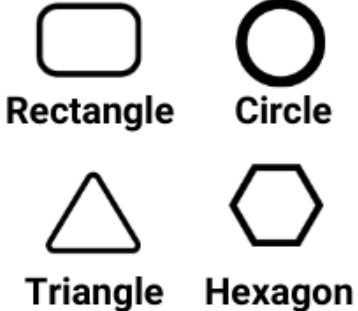
⚙️ Outputs ~ [Labels] / [Dependent Variables] / [Y]

Supervised Learning

Labeled Data



Labels



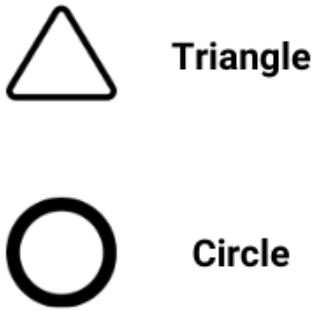
Machine



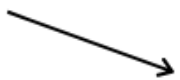
ML Model



Predictions



Test Data



Supervised Learning

Classification

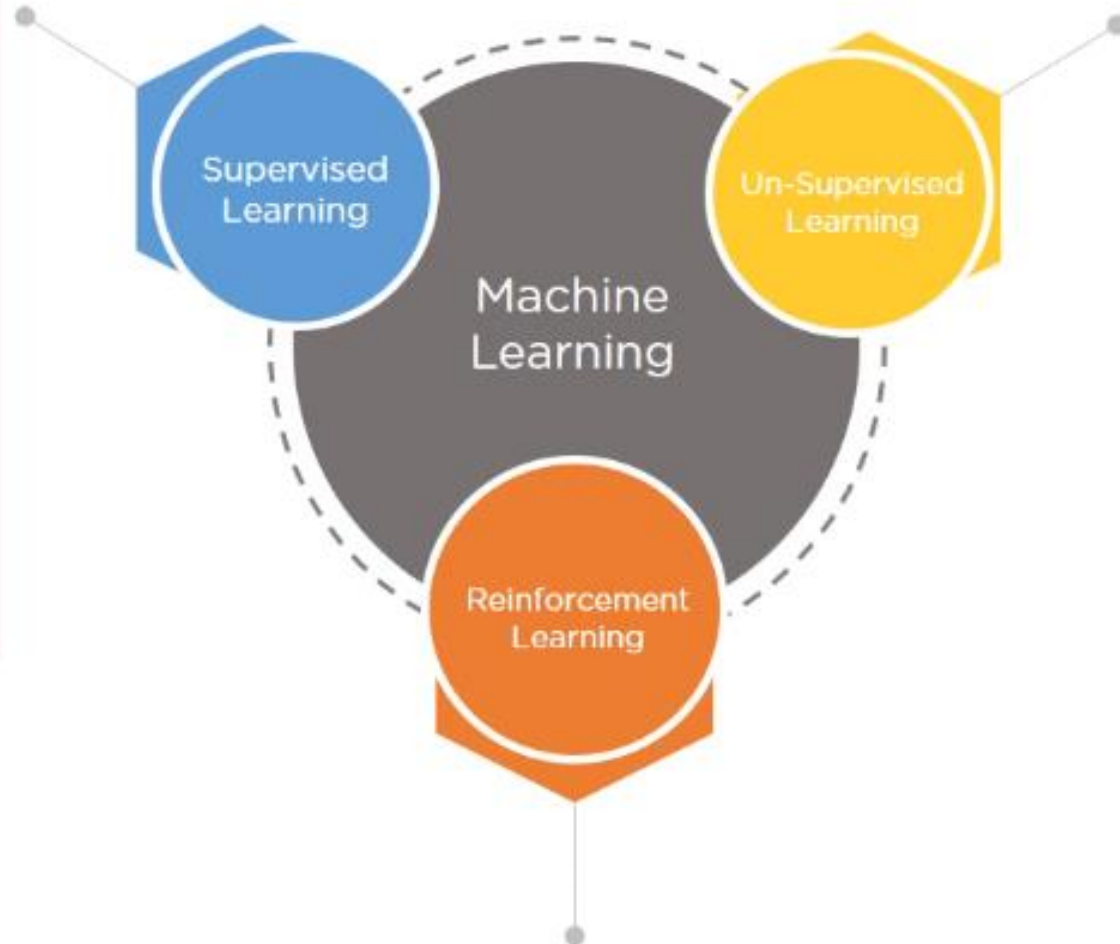
- Fraud Detection
- Email Spam Detection
- Image Classification

Categorical

Regression

- Weather Forecasting
- Risk Assessment
- Score Prediction

Numerical



Supervised Learning

User ID	Gender	Age	Salary	Purchased
15624510	Male	19	19000	0
15810944	Male	35	20000	1
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	1
15728773	Male	27	58000	1
15598044	Female	27	84000	0
15694829	Female	32	150000	1
15600575	Male	25	33000	1
15727311	Female	35	65000	0
15570769	Female	26	80000	1
15606274	Female	26	52000	0
15746139	Male	20	86000	1
15704987	Male	32	18000	0
15628972	Male	18	82000	0
15697686	Male	29	80000	0
15733883	Male	47	25000	1

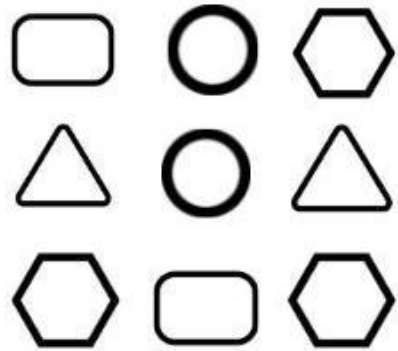
Figure A: CLASSIFICATION

Temperature	Pressure	Relative Humidity	Wind Direction	Wind Speed
10.69261758	986.882019	54.19337313	195.7150879	3.278597116
13.59184184	987.8729248	48.0648859	189.2951202	2.909167767
17.70494885	988.1119385	39.11965597	192.9273834	2.973036289
20.95430404	987.8500366	30.66273218	202.0752869	2.965289593
22.9278274	987.2833862	26.06723423	210.6589203	2.798230886
24.04233986	986.2907104	23.46918024	221.1188507	2.627005816
24.41475295	985.2338867	22.25082295	233.7911987	2.448749781
23.93361956	984.8914795	22.35178837	244.3504333	2.454271793
22.68800023	984.8461304	23.7538641	253.0864716	2.418341875
20.56425726	984.8380737	27.07867944	264.5071106	2.318677425
17.76400389	985.4262085	33.54900114	280.7827454	2.343950987
11.25680746	988.9386597	53.74139903	68.15406036	1.650191426
14.37810685	989.6819458	40.70884681	72.62069702	1.553469896
18.45114201	990.2960205	30.85038484	71.70604706	1.005017161
22.54895853	989.9562988	22.81738811	44.66042709	0.264133632
24.23155922	988.796875	19.74790765	318.3214111	0.329656571

Figure B: REGRESSION

Unsupervised Learning

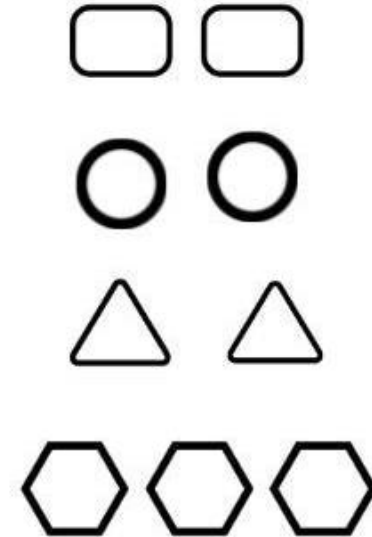
Unlabelled Data



Machine



Results



Customer Segmentation

Customer	Call Minutes	Data Usage	Text Messages	Tenure	Service Interactions	Avg. Monthly Bill
1	250	2.5	100	12	2	\$50
2	100	1.0	20	24	1	\$30
3	800	5.0	300	6	3	\$70
...

Unsupervised machine learning techniques can be applied to segment customers based on their usage patterns, preferences, and behavior. One common approach is using clustering algorithms to group customers with similar characteristics together.

Cluster 1 (High Usage):

- ❑ Customers with high call minutes, data usage, and text messages
- ❑ Relatively short tenure
- ❑ Moderate to high service interactions
- ❑ Higher average monthly bill

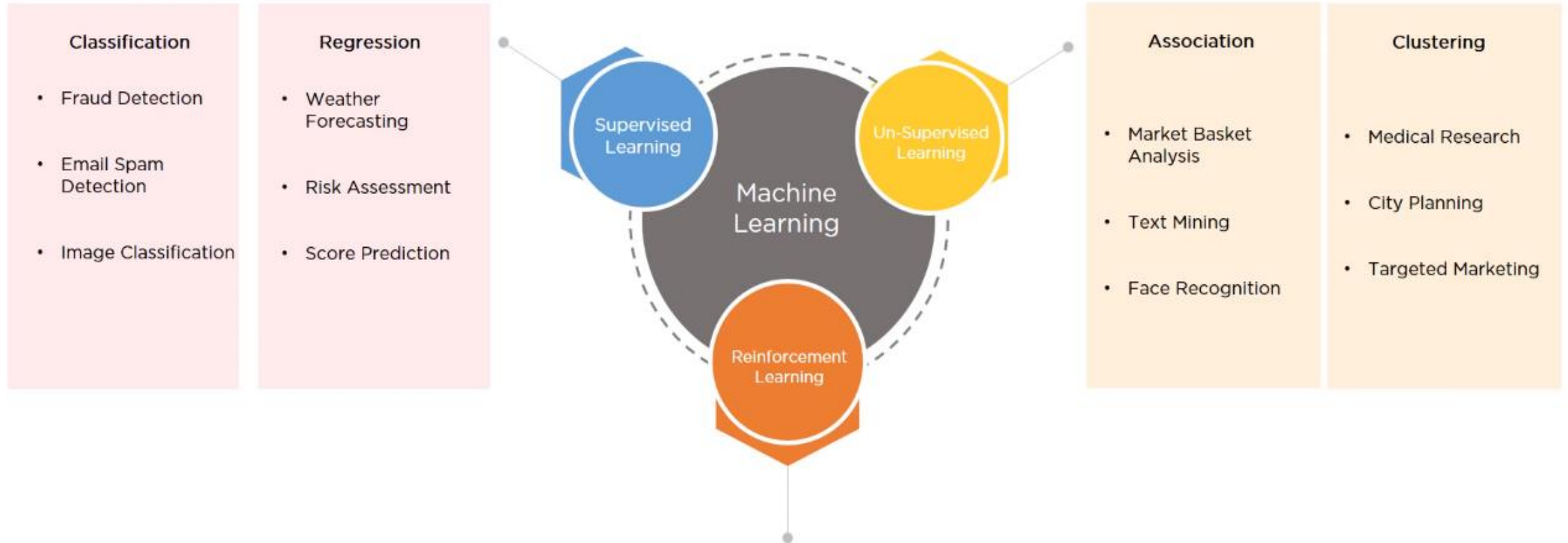
Cluster 2 (Medium Usage):

- ❑ Customers with moderate call minutes, data usage, and text messages
- ❑ Medium tenure
- ❑ Moderate service interactions
- ❑ Moderate average monthly bill

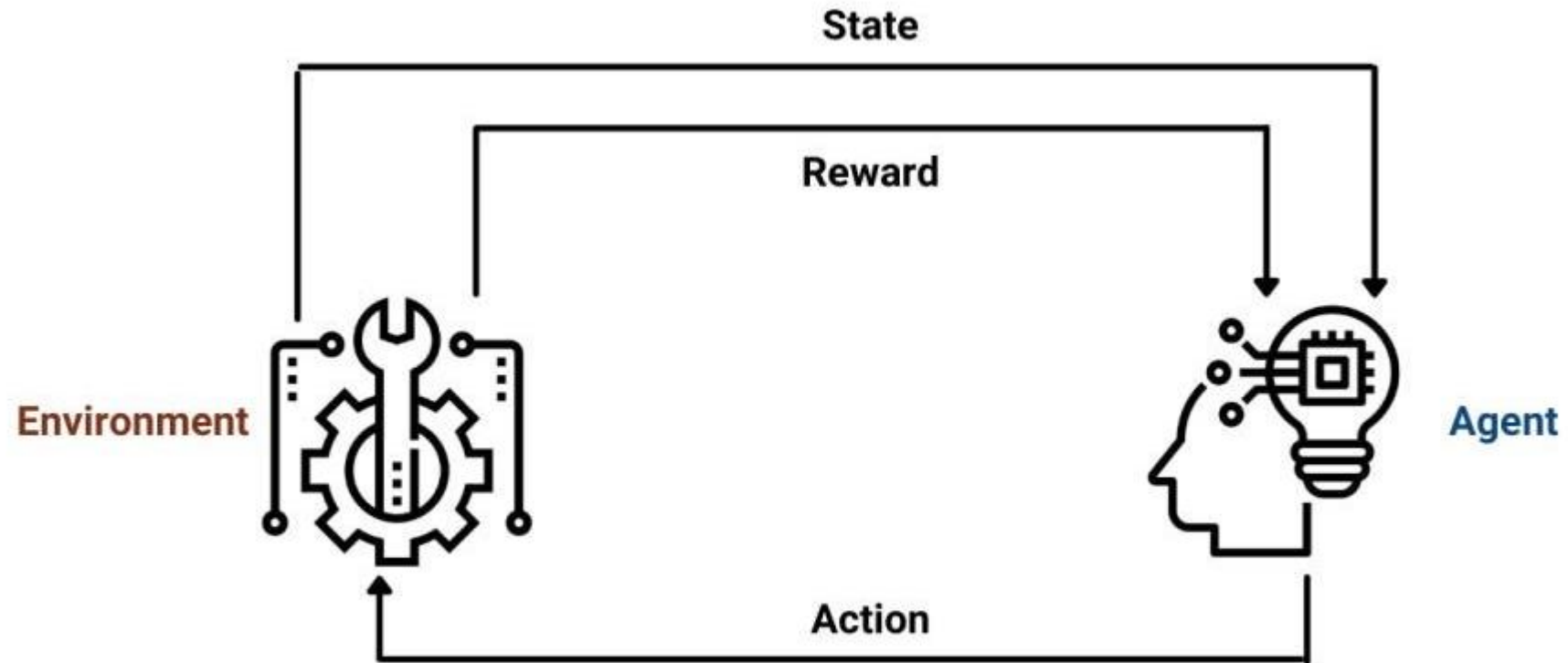
Cluster 3 (Low Usage):

- ❑ Customers with low call minutes, data usage, and text messages
- ❑ Longer tenure
- ❑ Low service interactions
- ❑ Lower average monthly bill

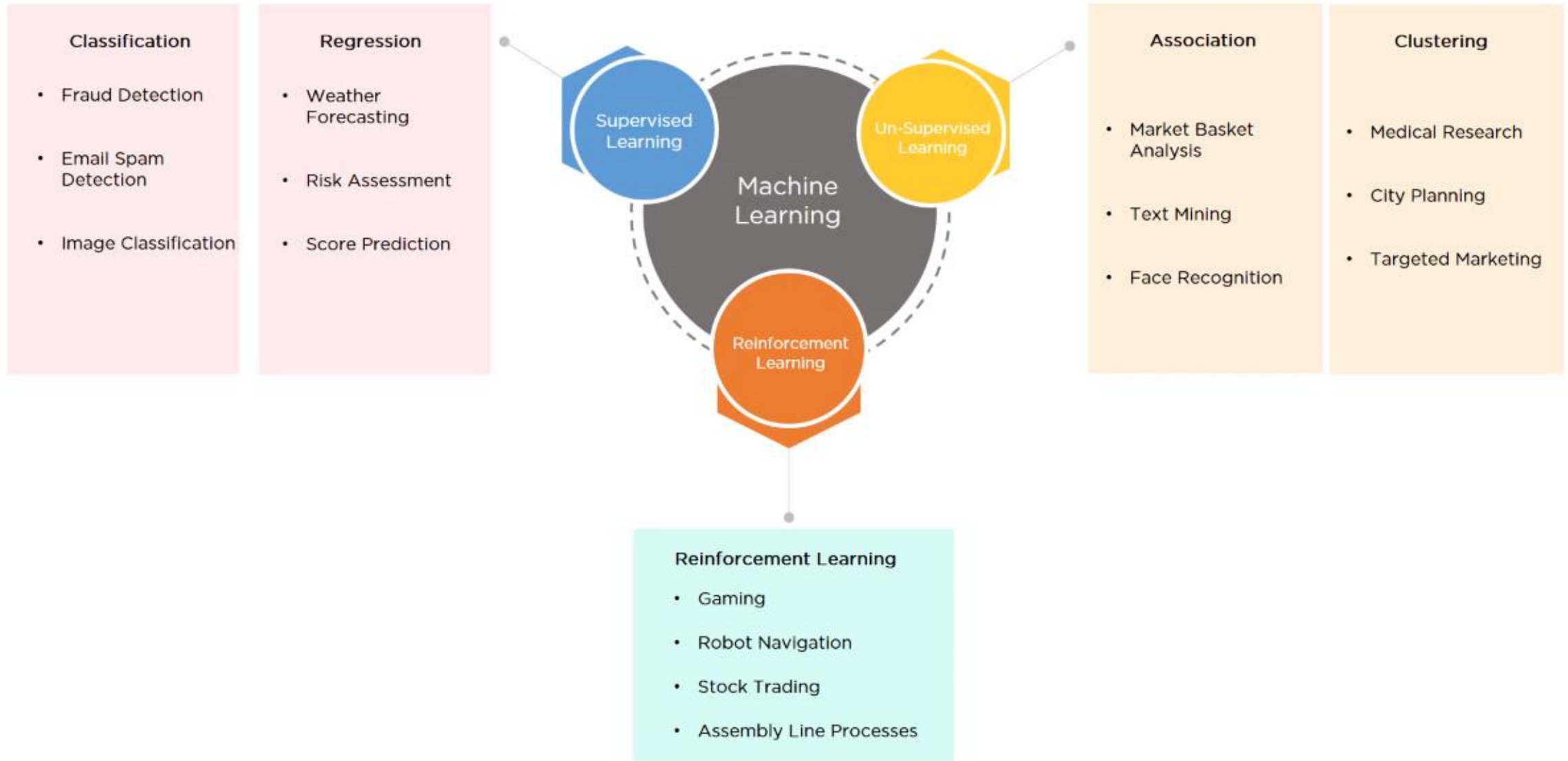
Unsupervised Learning



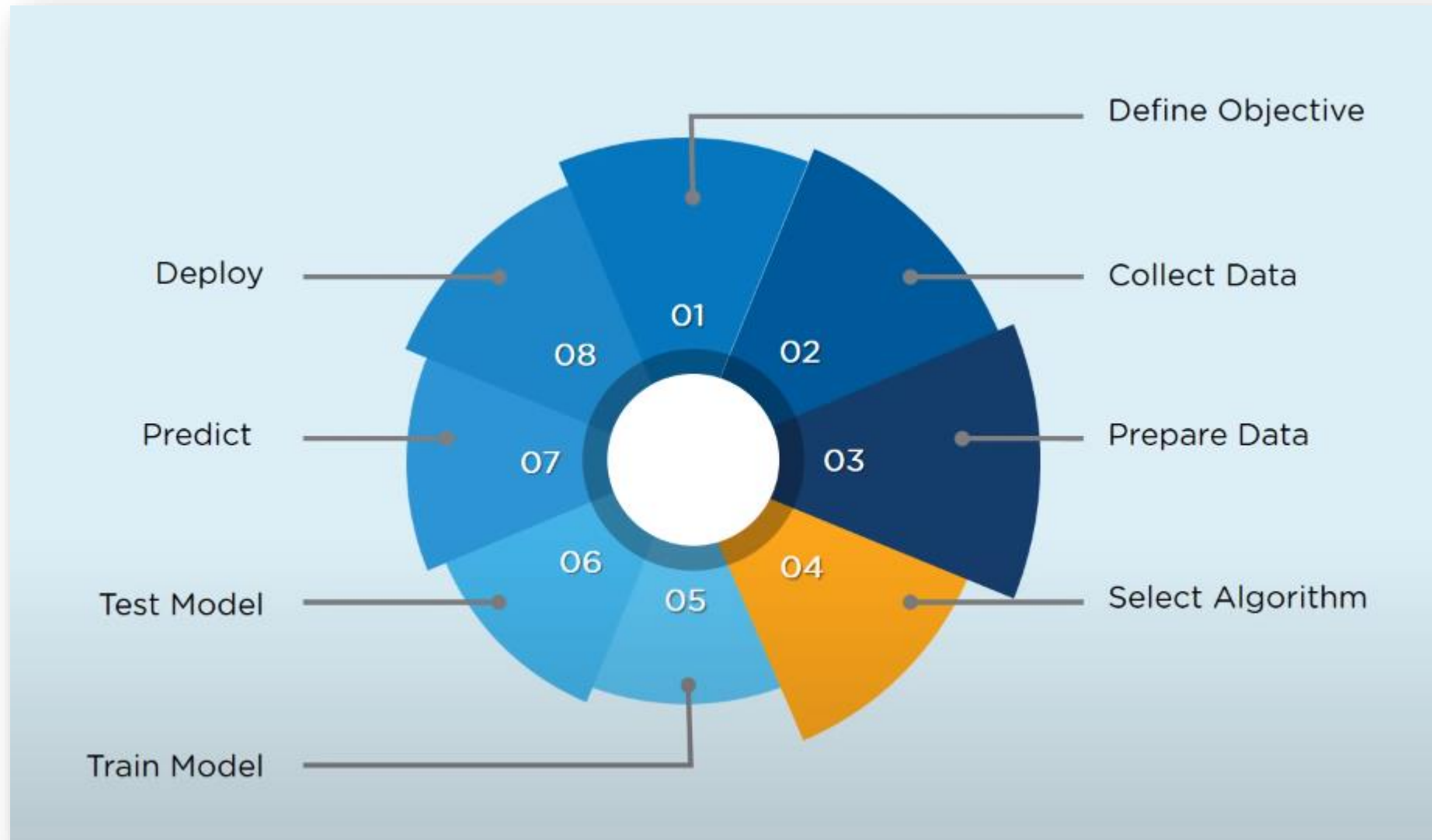
Reinforcement Learning



Reinforcement Learning



Processing Steps for Machine Learning



Performance Metrics - Regression

○ BAD MODEL

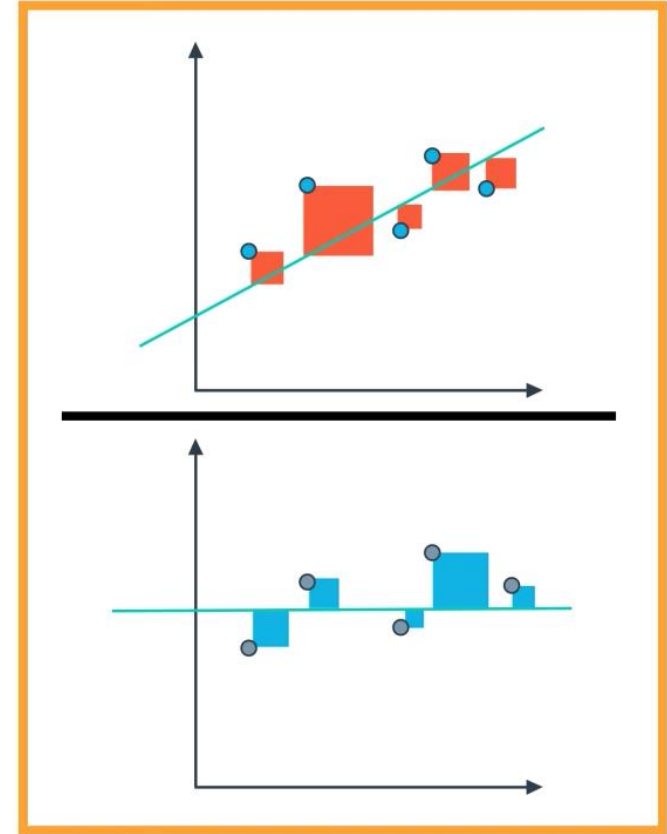
The errors should be similar.
R2 score should be close to 0.

○ GOOD MODEL

The mean squared error for the linear regression model should be a lot smaller than the mean squared error for the simple model.

R2 score should be close to 1.

$$R^2 = 1 -$$



R² (R-Square) score can be interpreted as a coefficient of determination where value propagate from **0 to 1**, where 1 is the best.

Performance Metrics - Classification

CONFUSION MATRIX

		Diagnosis prediction	
		Diagnosed sick	Diagnosed healthy
Patients	Actually sick	1,000 True positives	200 False negatives
	Actually healthy	800 False positives	8,000 True negatives

Patients 10,000

Performance Metrics - Classification

$$\text{F1 score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

F1 score can be interpreted as a measure of overall model performance from **0 to 1**, where 1 is the best.

Hands-on

Supervised Machine Learning

IDE : No Code Machine Learning with [Azure Machine Learning Studio \(Classic\)](#)
Desktop Based Data Mining : [Orange](#)

Hands-on

Step 1 : Please go to this site <https://studio.azureml.net/>

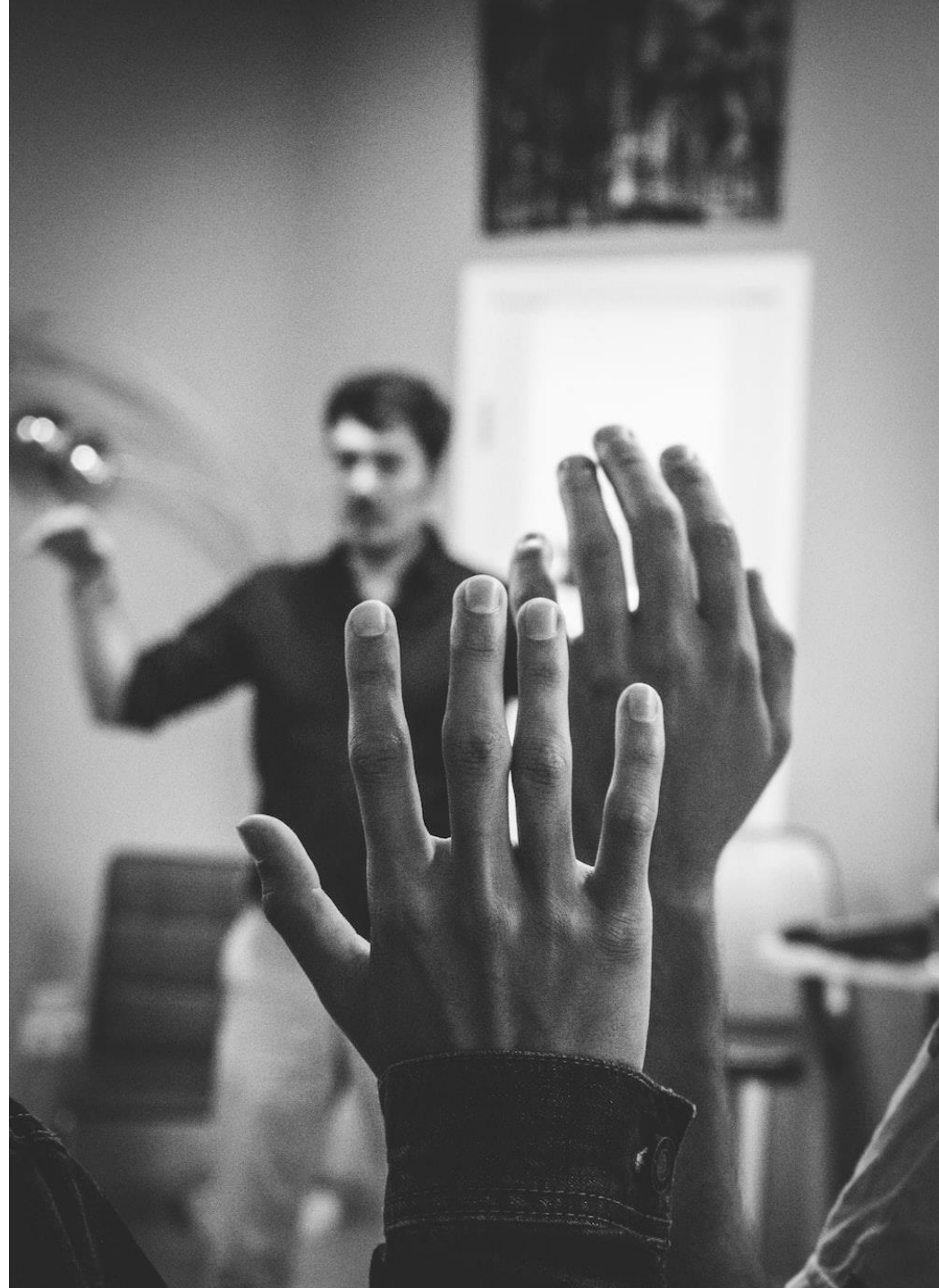
Step 2 : Use any Microsoft Account to Register and Login

Step 3 : Let's do some Prediction

Please go to this page for all resources : <http://arif.works/blink>

1

Ask **Questions** to
your Data !



2

Understand your Numbers !



3

Visualize your
Data!



4

Data **Storytelling** Connect Dots!



5

Data **Modelling** Mingle your Data!



6

Predictive Analytics – with ML

