Bells and Whistles of Data Science

Dr. Neha Sharma

Founder Secretary, Society for Data Science

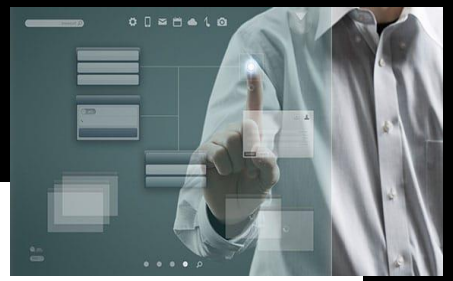
IEEE Senior Member,

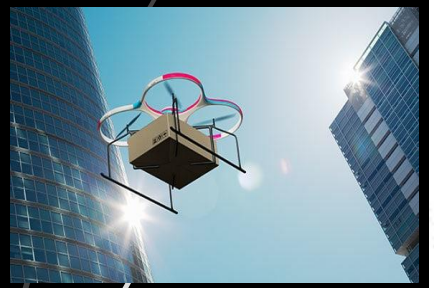
Execom Member, IEEE Pune Section

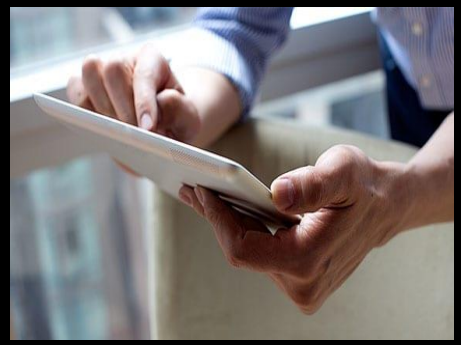
SDS SOCIETY FOR DATA SCIENCE

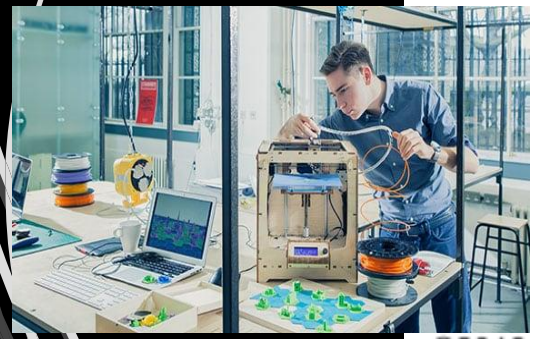
IEEE Pune Section

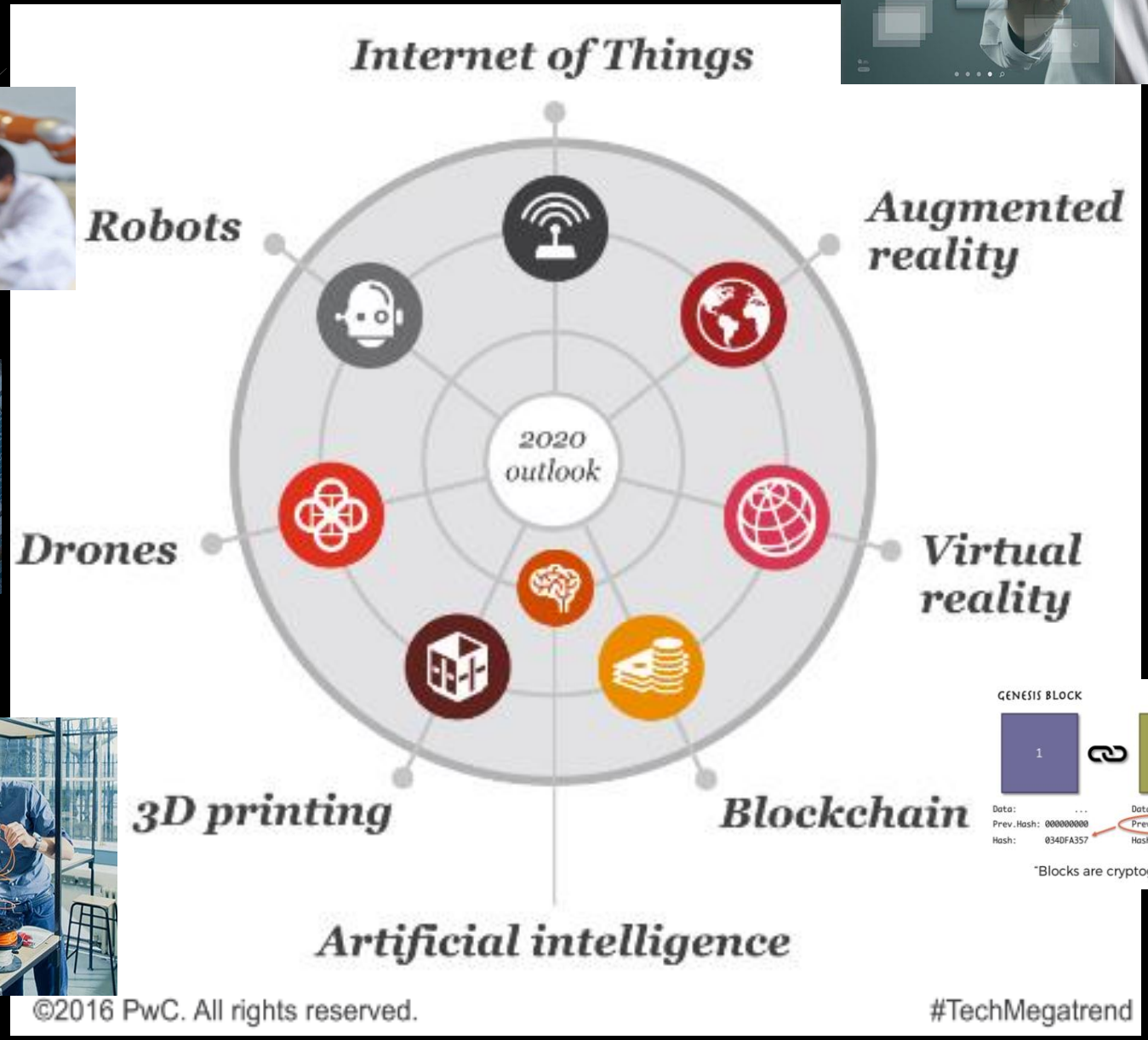
The essential eight technologies

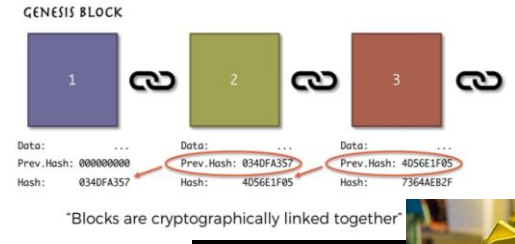










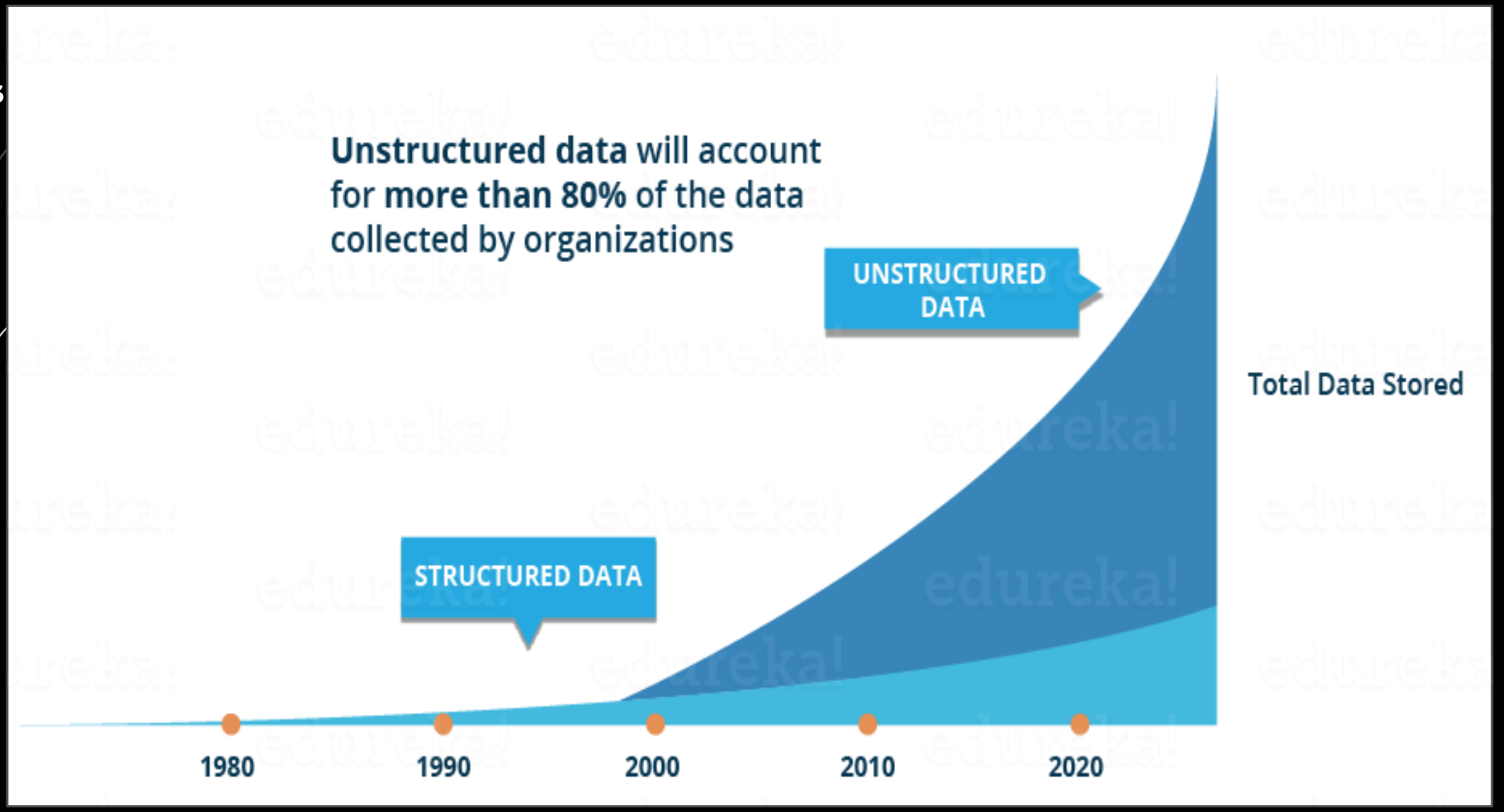


©2016 PwC. All rights reserved.

#TechMegatrend

Data in Data Science: Big Data

- Financial Logs
- Text Files
- Multimedia Forms
- Sensors
- Instruments



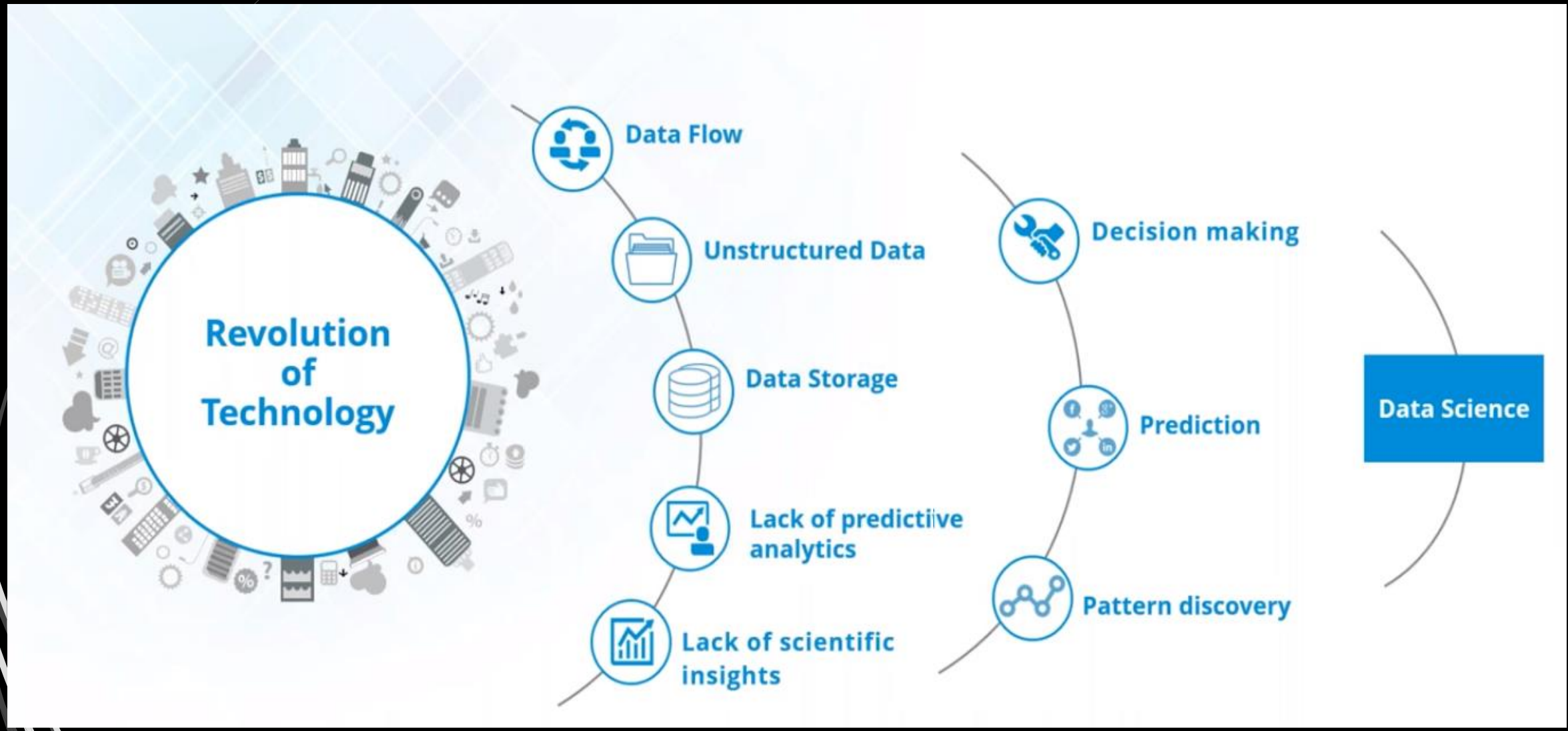
- ✓ **IBM's Definition – Big Data Characteristics**
  http://www-01.ibm.com/software/data/bigdata/

VOLUME

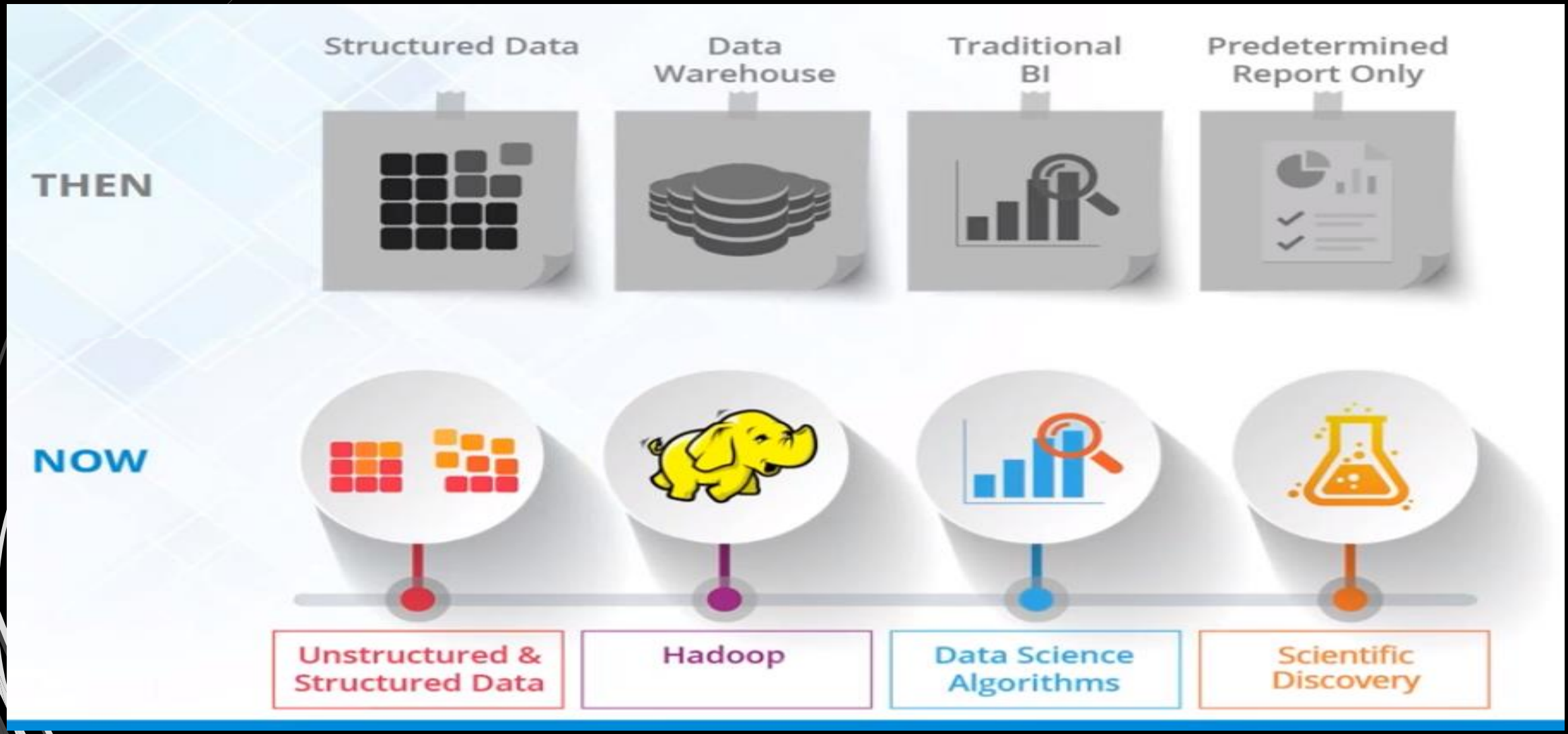
VELOCITY

VARIETY

Need For Data Science



Need For Data Science

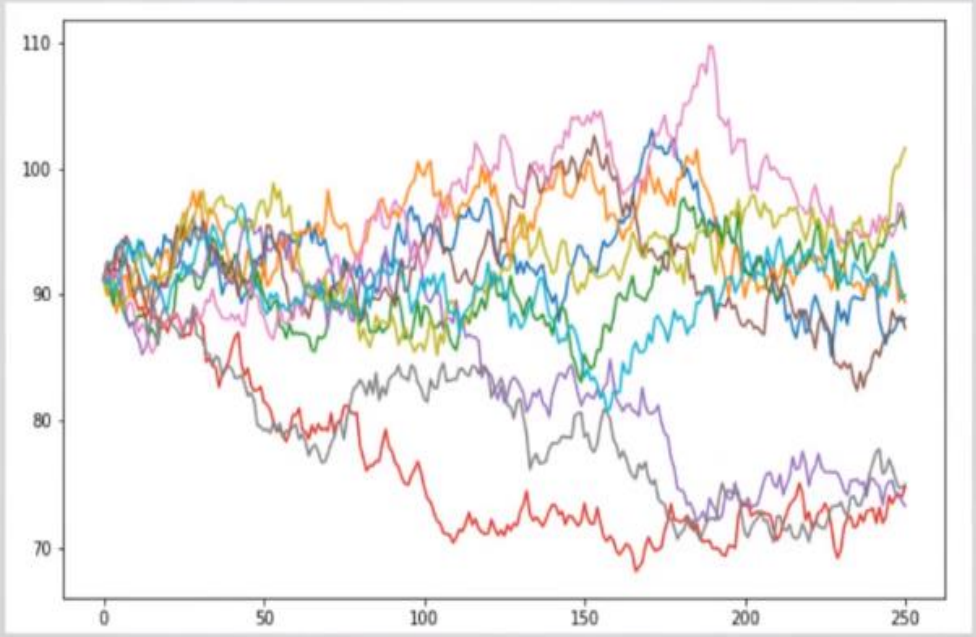


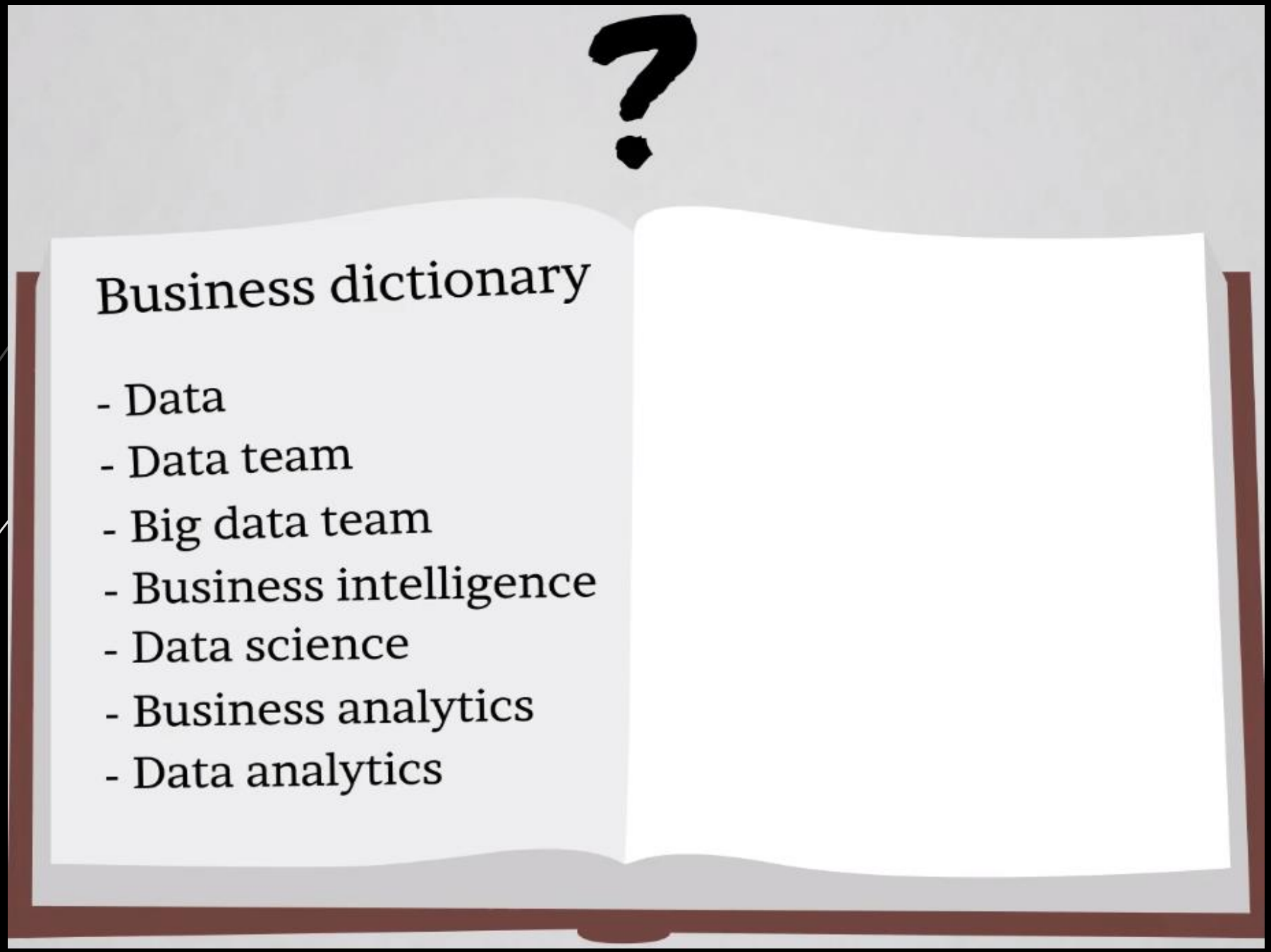
DATA SCIENCE

WHY ARE THERE SO MANY BUSINESS AND DATA SCIENCE BUZZWORDS?

Data science team

$$S_t = S_{t-1} \cdot e^{((r-\frac{1}{2}\cdot stdev^2)\cdot\delta_t + stdev\cdot\sqrt{\delta_t}\cdot Z_t)}$$



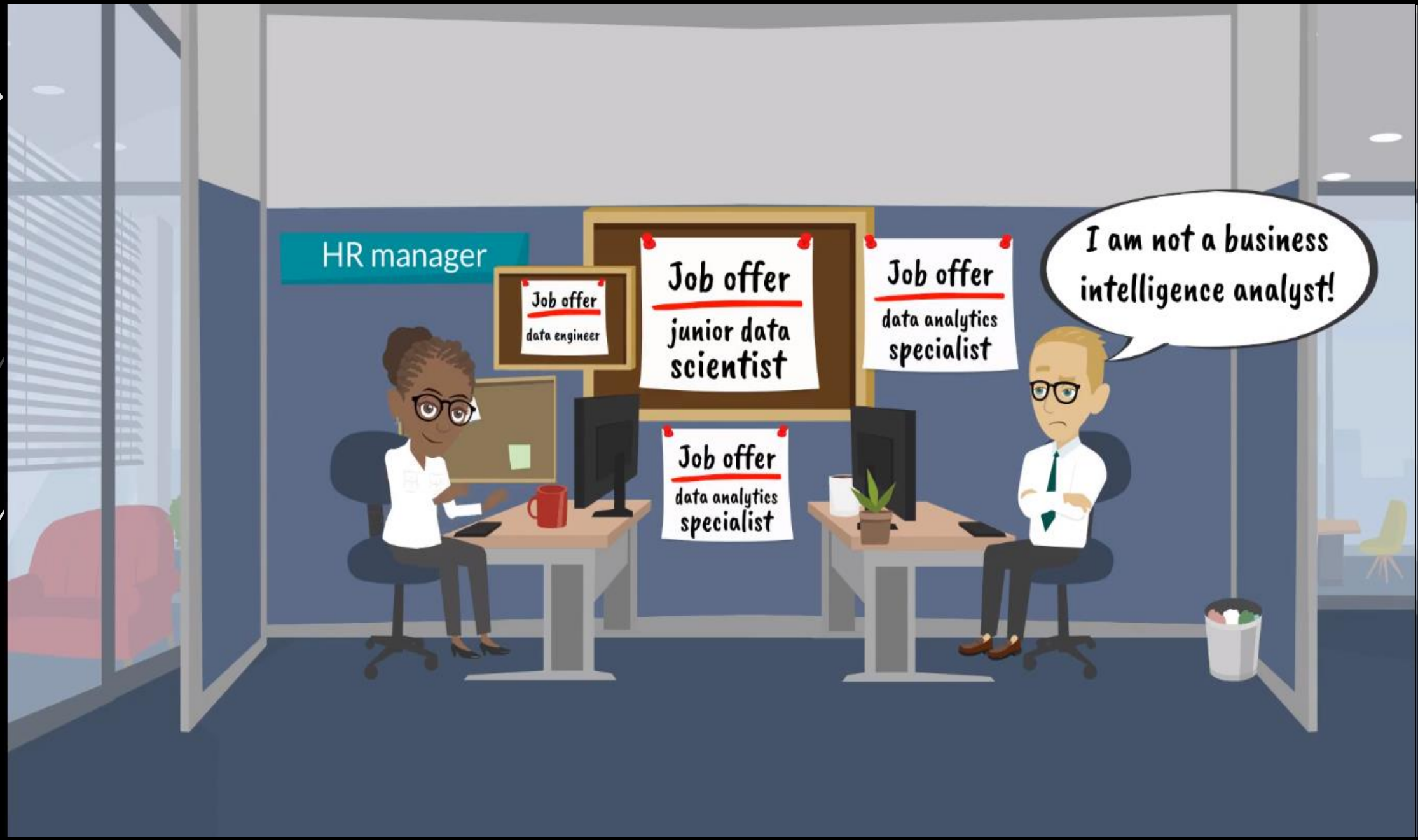

?
Business dictionary
- Data
- Data team
- Big data team
- Business intelligence
- Data science
- Business analytics
- Data analytics

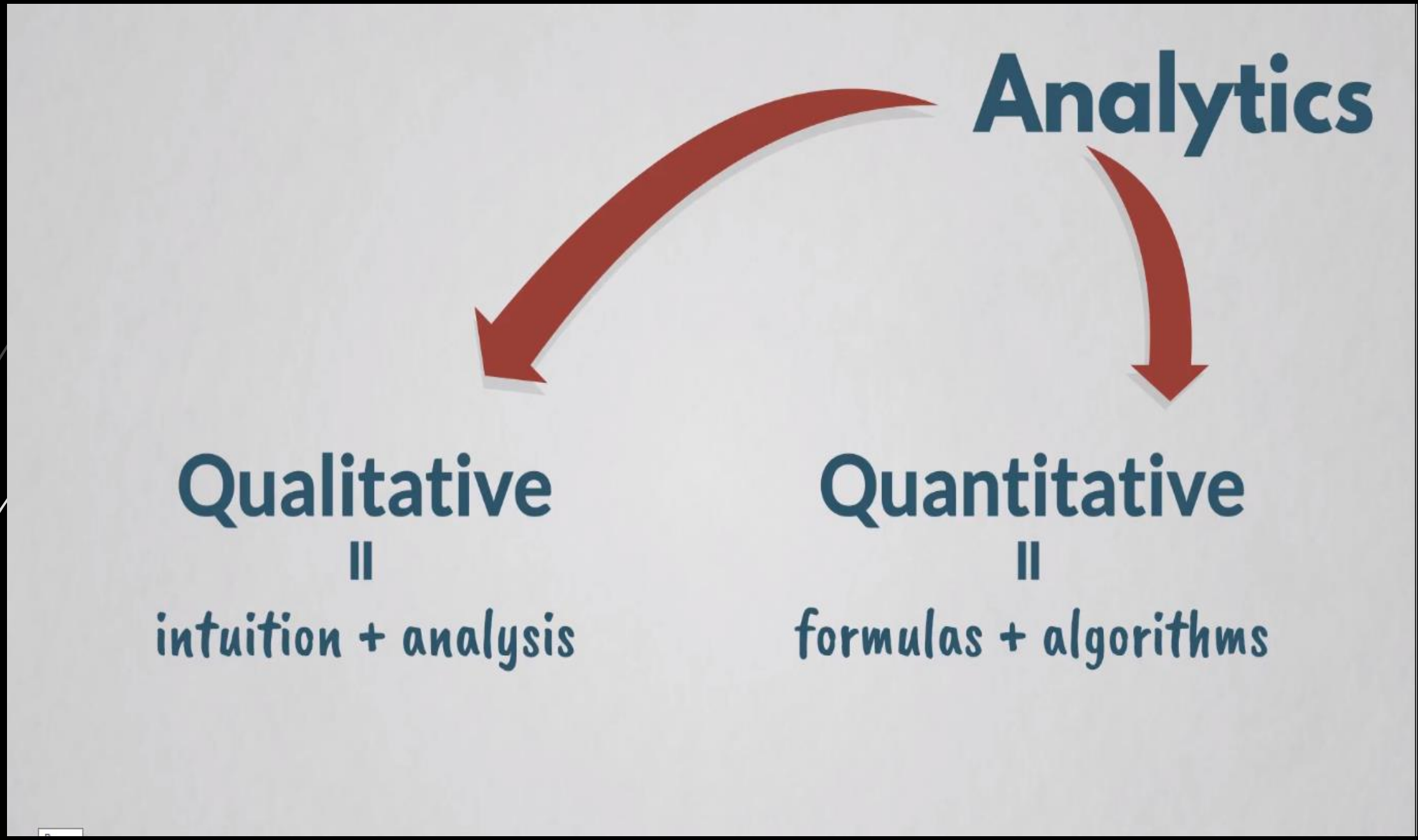
Statistics

Data mining

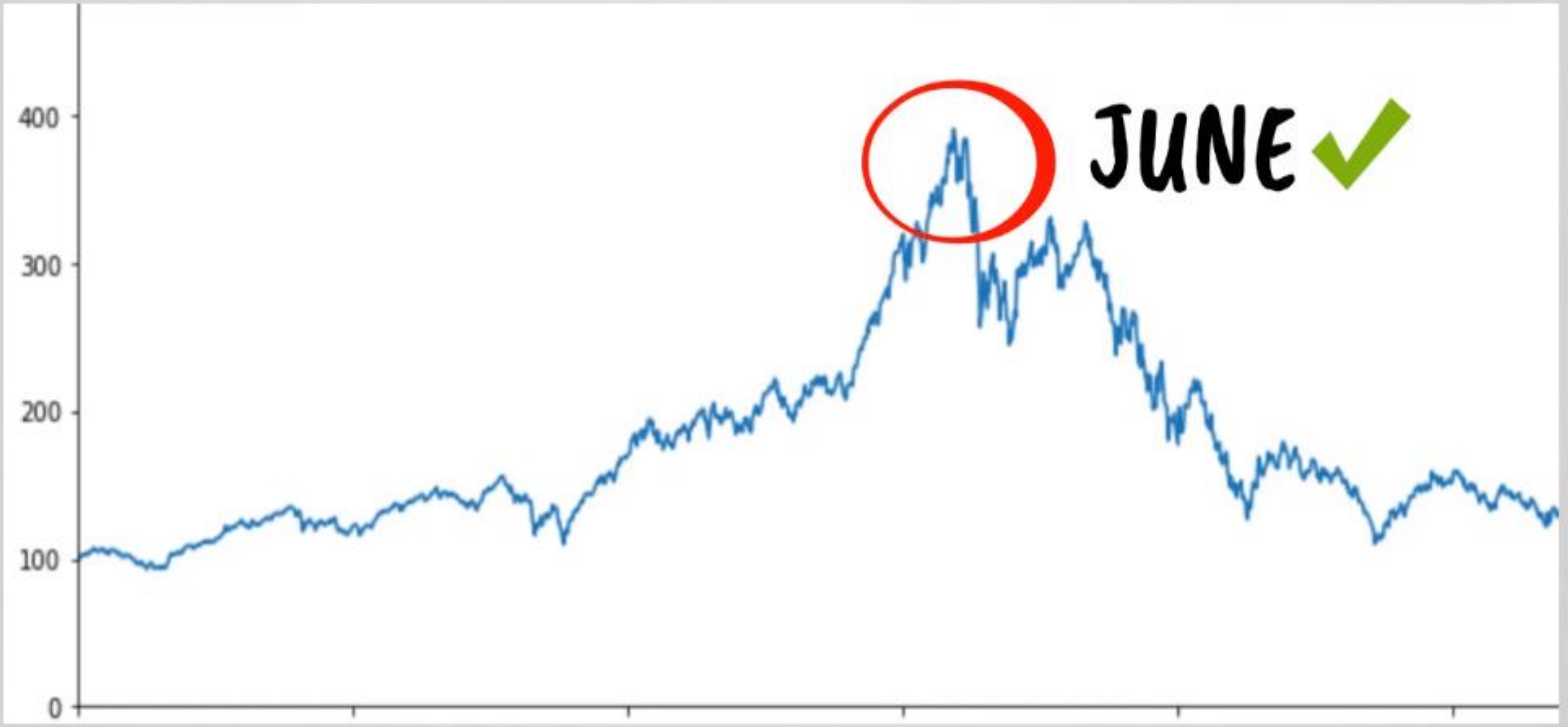
Predictive analytics

Data Science


HR manager
Job offer
data engineer
Job offer
junior data scientist
Job offer
data analytics specialist
Job offer
data analytics specialist
I am not a business intelligence analyst!


Analytics
Qualitative
=
intuition + analysis
Quantitative
=
formulas + algorithms

Quantitative analytics



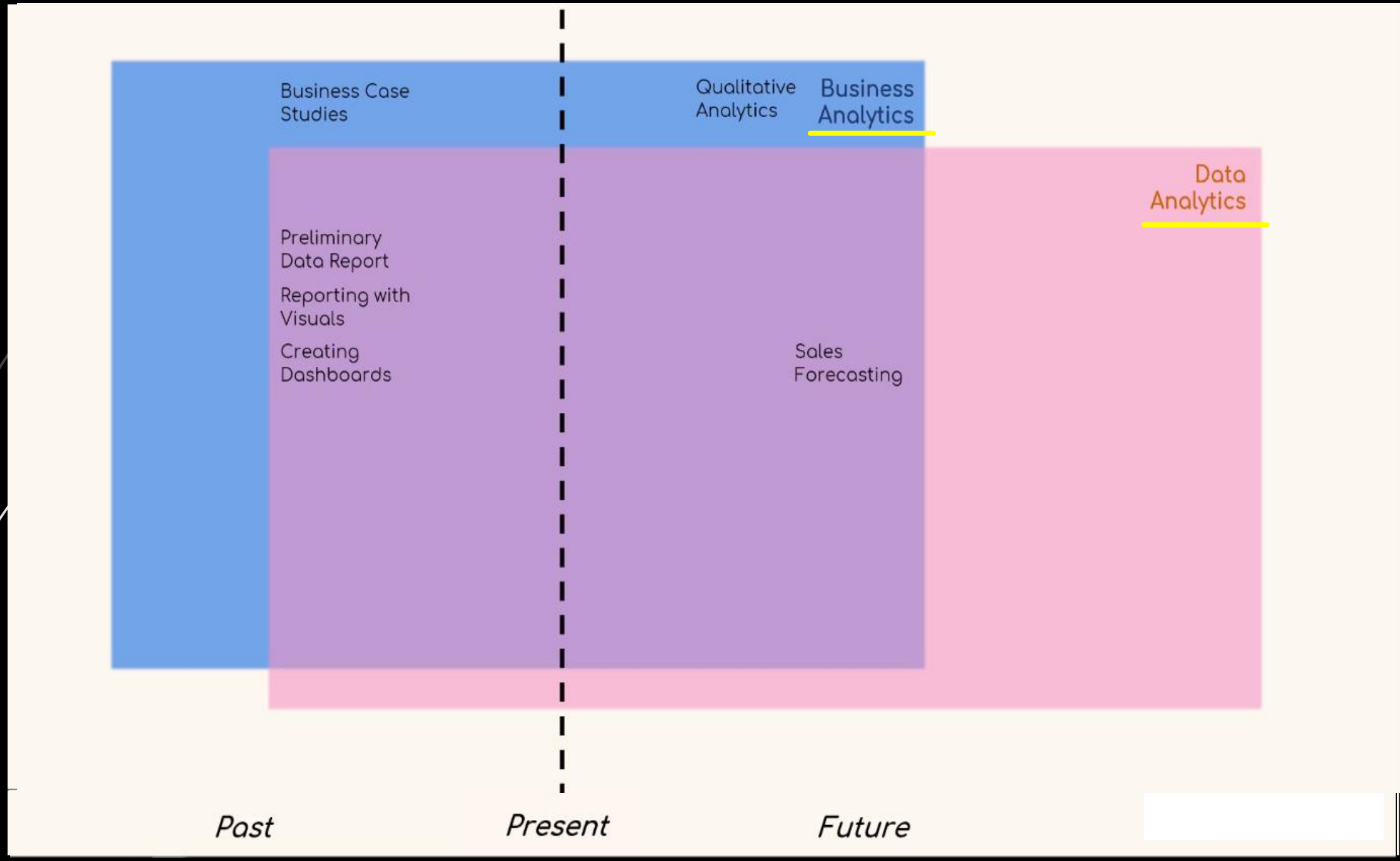
Analysis ≠ Analytics

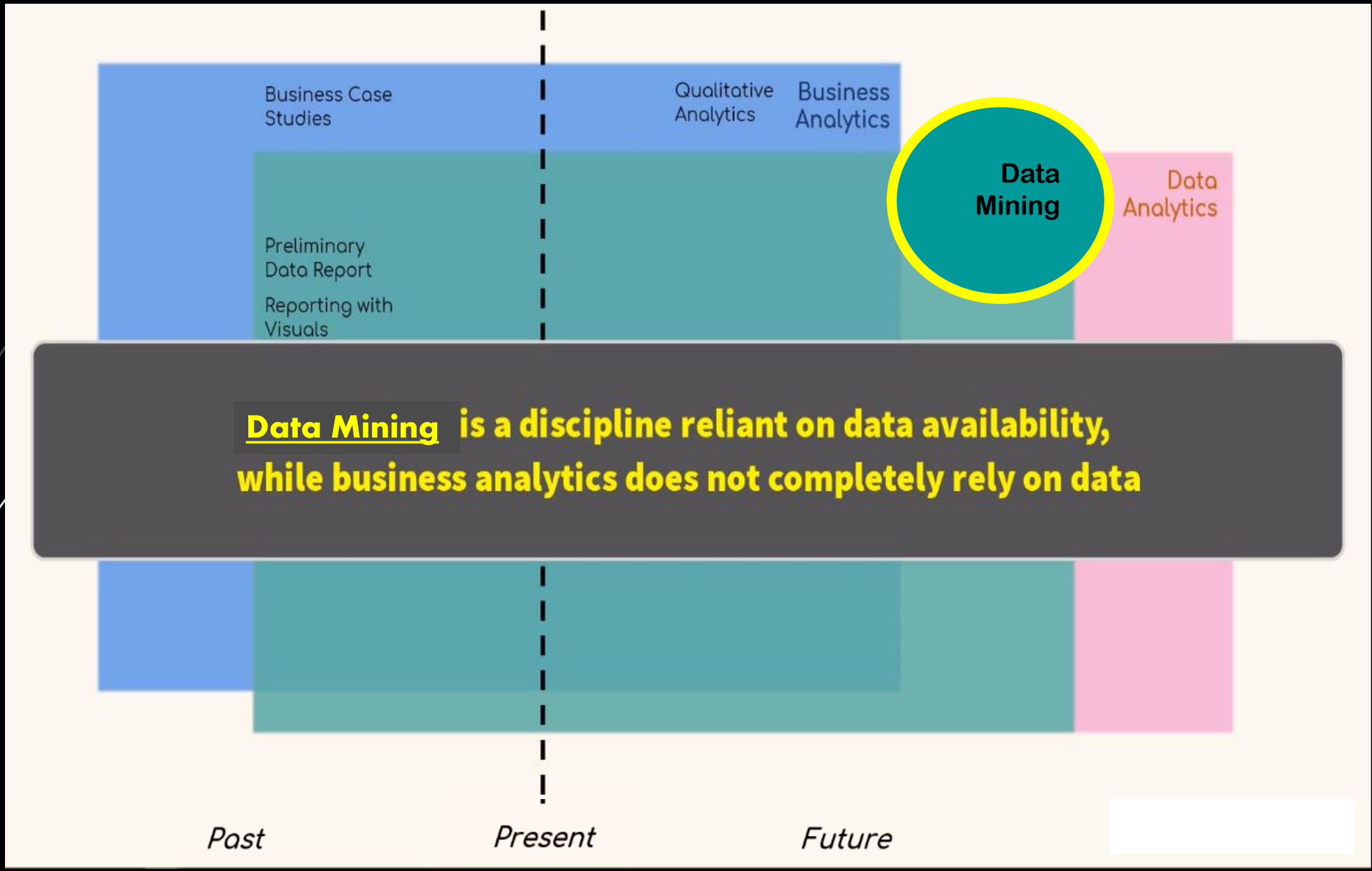
data analysis ≠ data analytics

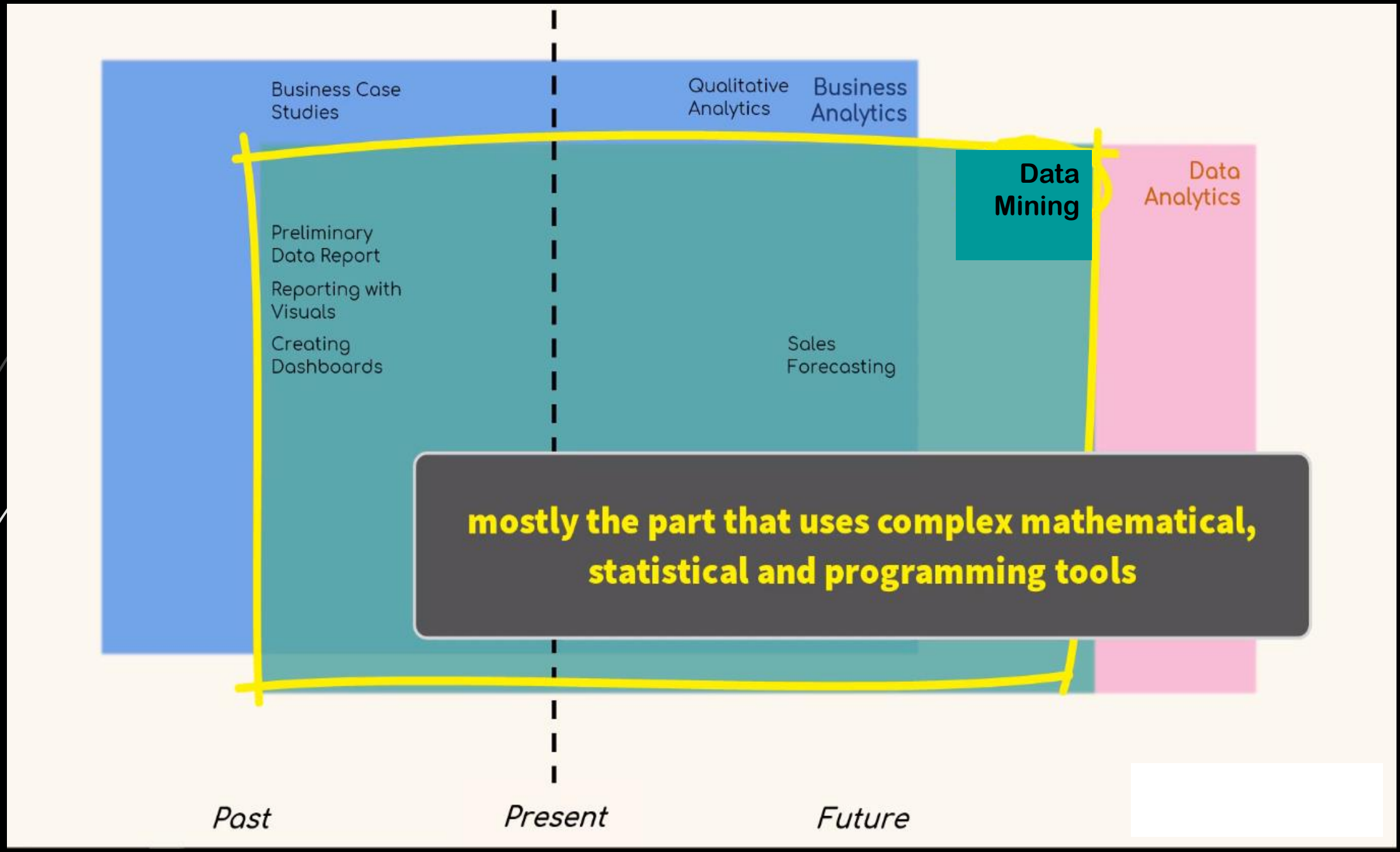
business analysis ≠ business analytics

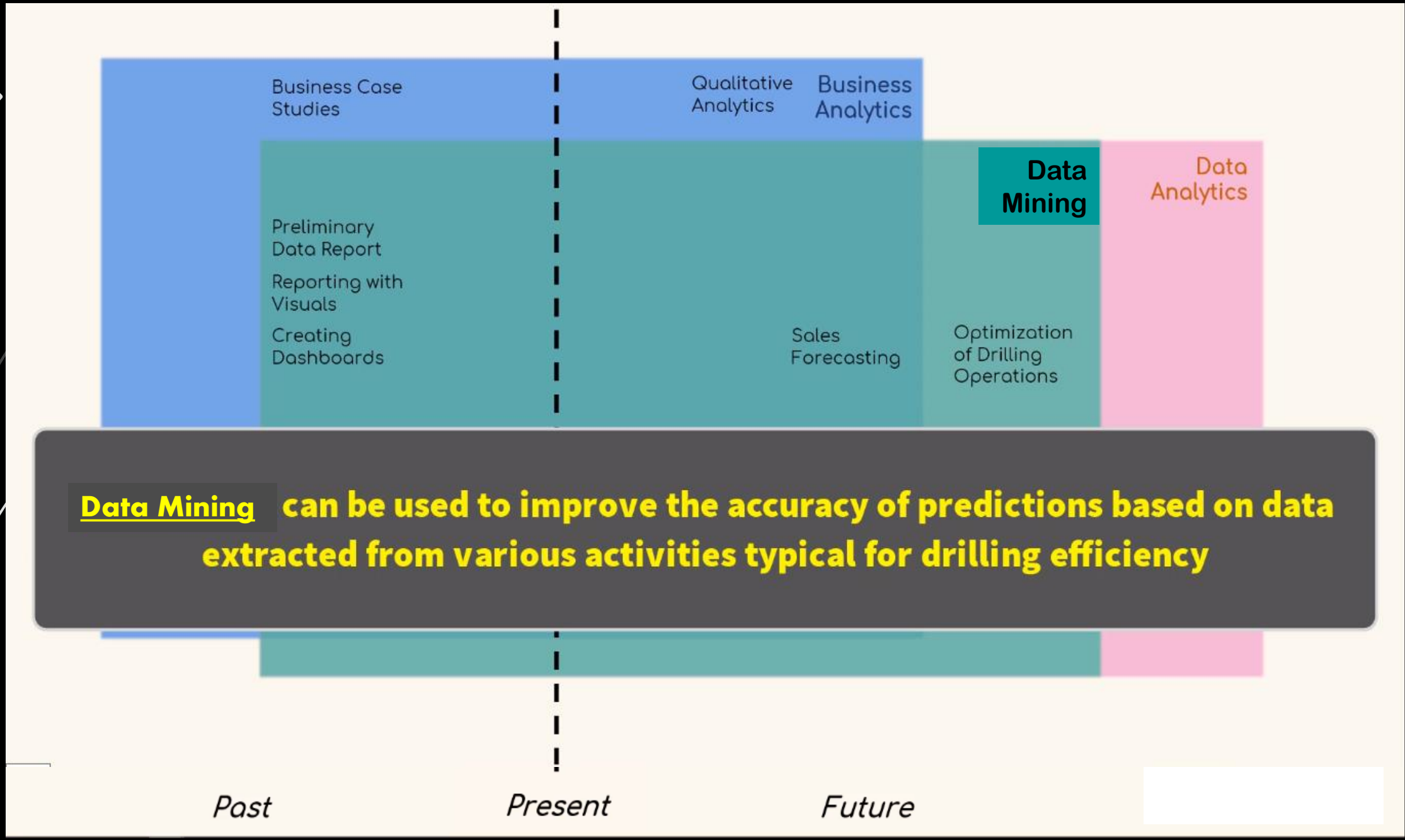
Introduction to Business Analytics, Data Analytics and Data Science

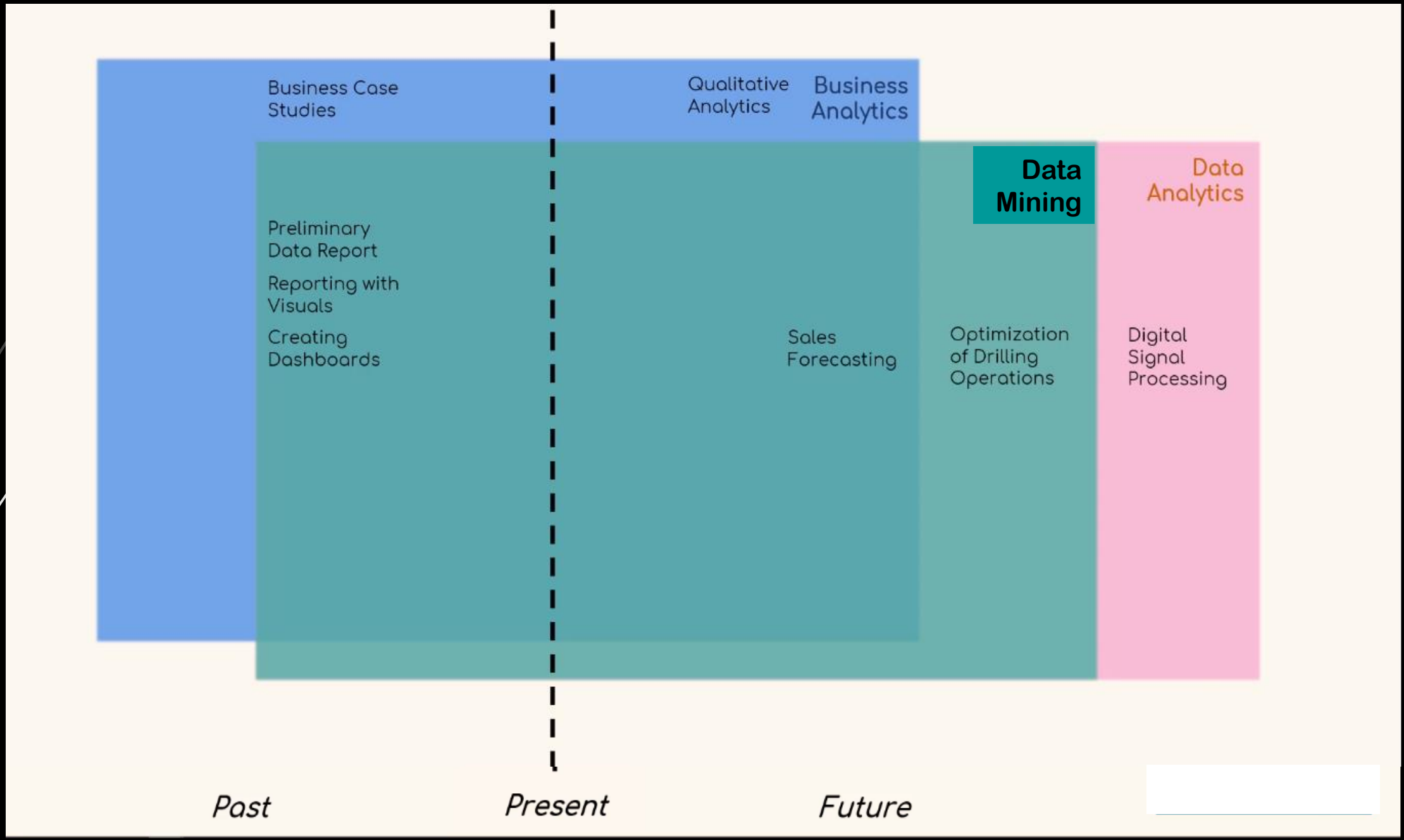

Business Case Studies
Qualitative Analytics
Business
Data
Preliminary Data Report
Reporting with Visuals
Creating Dashboards
Sales Forecasting


Business Case Studies
Qualitative Analytics
Business Analytics
Data Analytics
Preliminary Data Report
Reporting with Visuals
Creating Dashboards
Sales Forecasting
Past
Present
Future

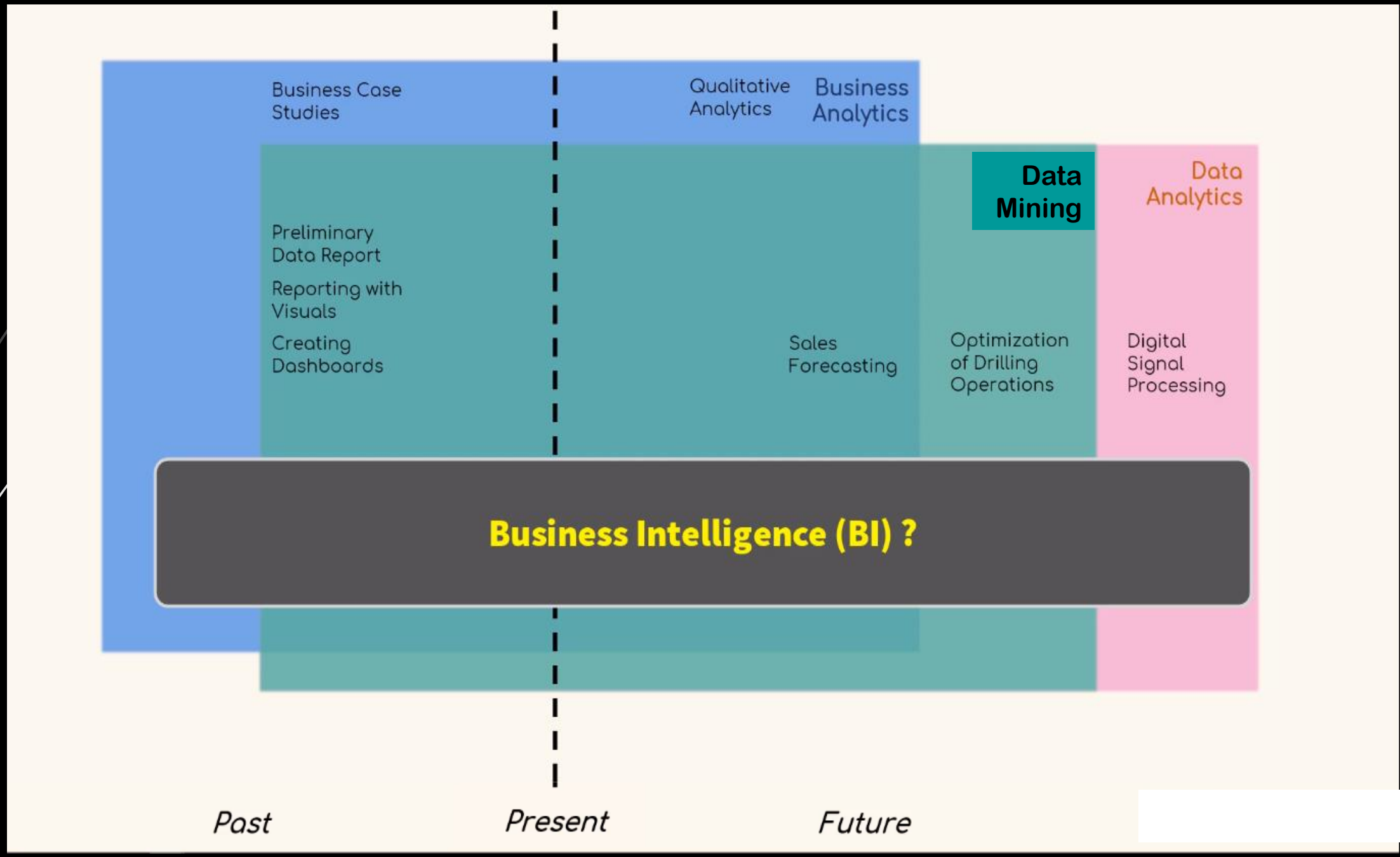

Business Case Studies
Qualitative Analytics
Business Analytics
Data Mining
Data Analytics
Preliminary Data Report
Reporting with Visuals
Data Mining is a discipline reliant on data availability, while business analytics does not completely rely on data
Past
Present
Future


Business Case Studies
Qualitative Analytics
Business Analytics
Data Mining
Data Analytics
Preliminary Data Report
Reporting with Visuals
Creating Dashboards
Sales Forecasting
mostly the part that uses complex mathematical, statistical and programming tools
Past
Present
Future


Business Case Studies
Qualitative Analytics
Business Analytics
Data Mining
Data Analytics
Preliminary Data Report
Reporting with Visuals
Creating Dashboards
Sales Forecasting
Optimization of Drilling Operations
Data Mining can be used to improve the accuracy of predictions based on data extracted from various activities typical for drilling efficiency
Past
Present
Future

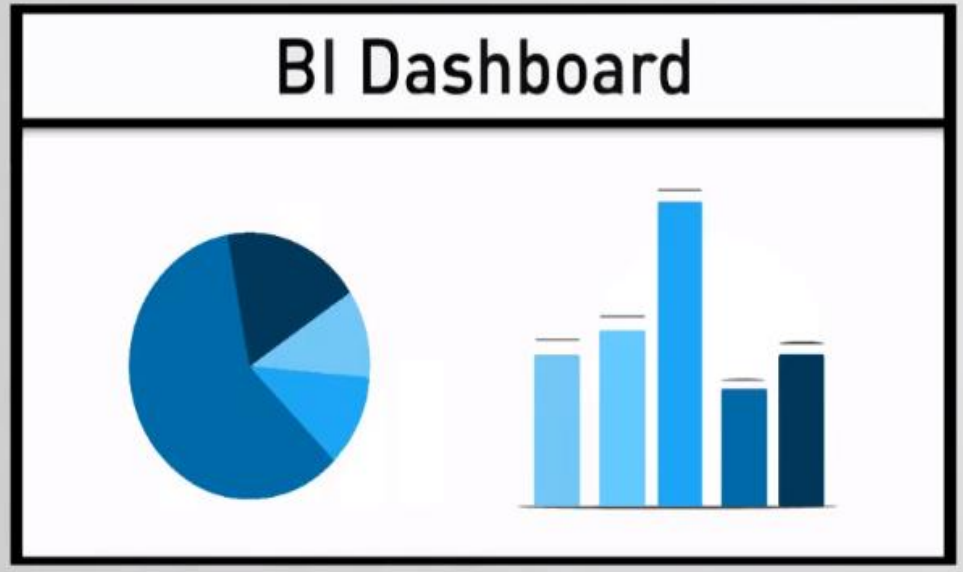

Business Case Studies
Qualitative Analytics
Business Analytics
Data Mining
Data Analytics
Preliminary Data Report
Reporting with Visuals
Creating Dashboards
Sales Forecasting
Optimization of Drilling Operations
Digital Signal Processing
Past
Present
Future

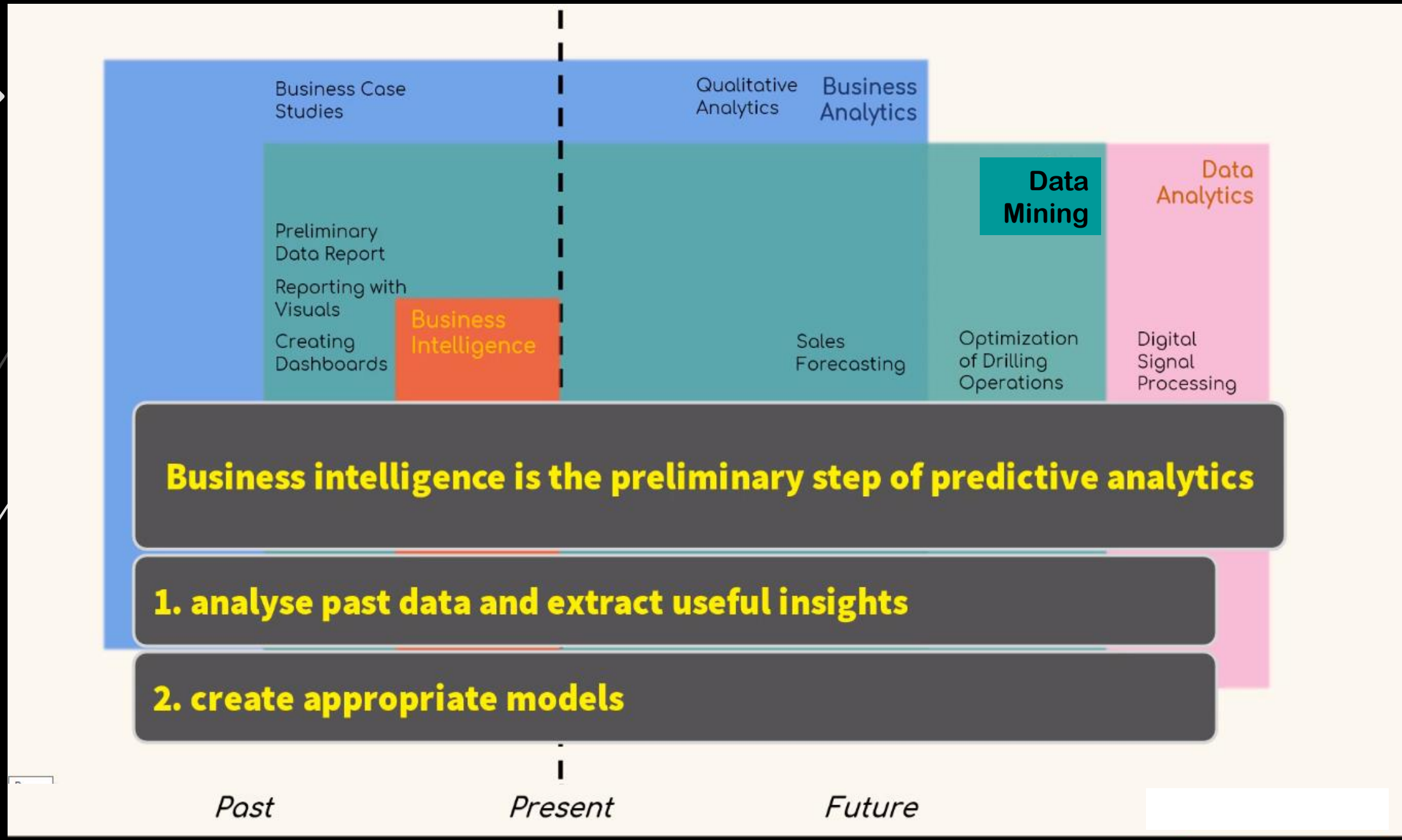
Adding Business Intelligence (BI), Machine Learning (ML) and Artificial Intelligence (AI)

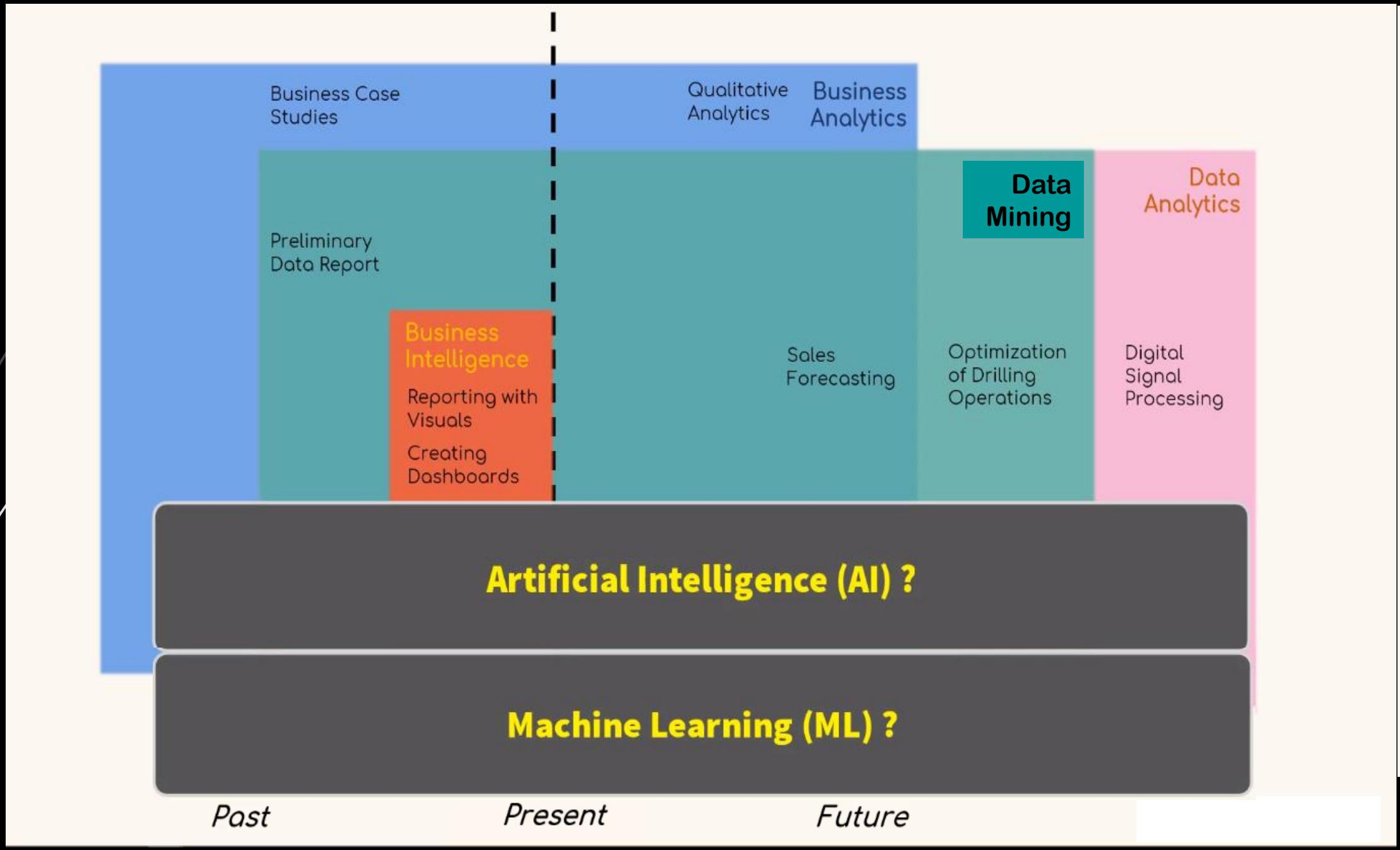

Business Case Studies
Qualitative Analytics
Business Analytics
Data Mining
Data Analytics
Preliminary Data Report
Reporting with Visuals
Creating Dashboards
Sales Forecasting
Optimization of Drilling Operations
Digital Signal Processing
Business Intelligence (BI) ?
Past
Present
Future

business intelligence (BI): the process of analysing and reporting historical business data

aims to explain past events using business data



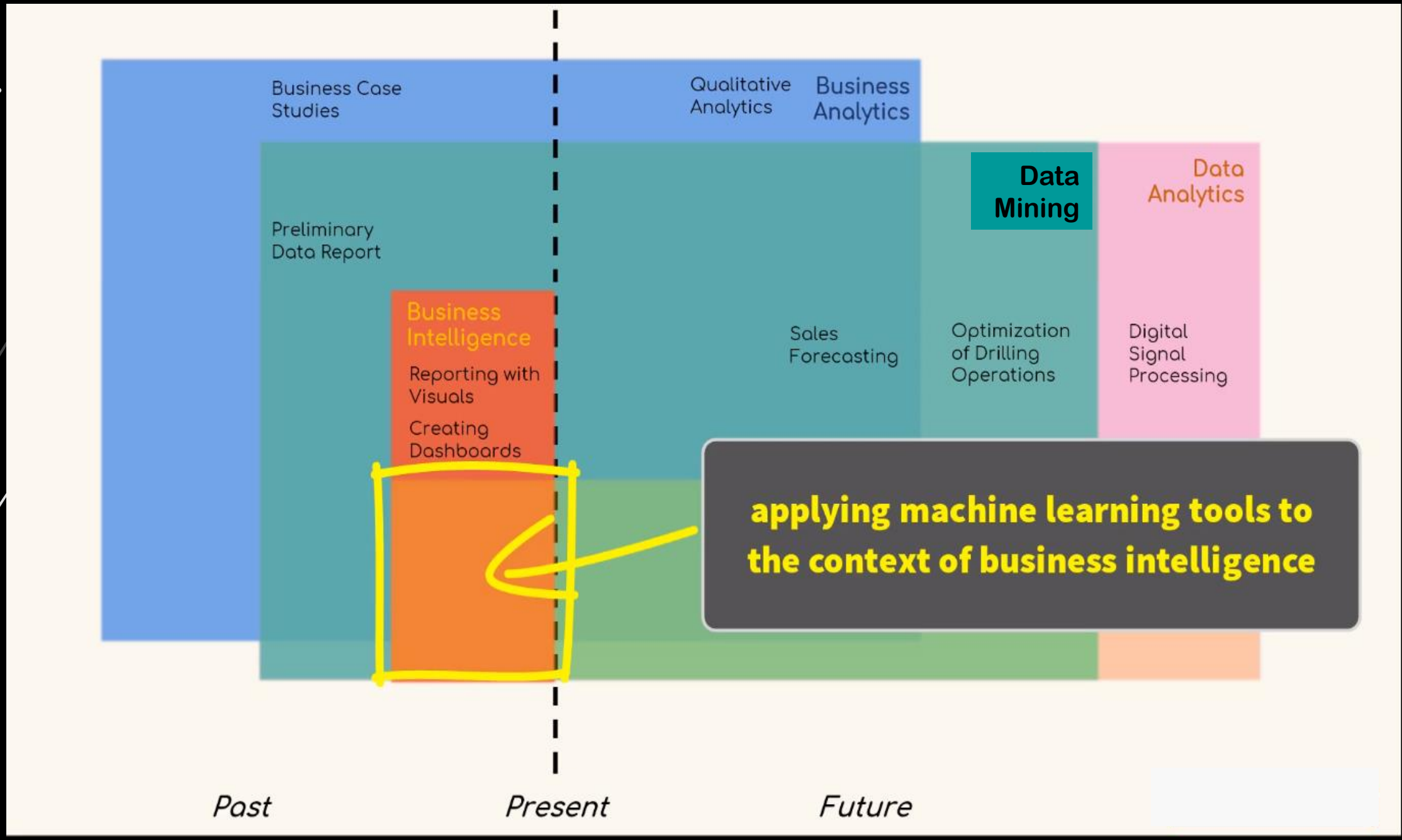

Business Case Studies
Qualitative Analytics
Business Analytics
Data Mining
Data Analytics
Preliminary Data Report
Reporting with Visuals
Creating Dashboards
Business Intelligence
Sales Forecasting
Optimization of Drilling Operations
Digital Signal Processing
Business intelligence is the preliminary step of predictive analytics
1. analyse past data and extract useful insights
2. create appropriate models
Past
Present
Future


Business Case Studies
Qualitative Analytics
Business Analytics
Data Mining
Data Analytics
Preliminary Data Report
Business Intelligence
Reporting with Visuals
Creating Dashboards
Sales Forecasting
Optimization of Drilling Operations
Digital Signal Processing
Artificial Intelligence (AI) ?
Machine Learning (ML) ?
Past
Present
Future

machine learning: The ability of machines to predict outcomes without being explicitly programmed

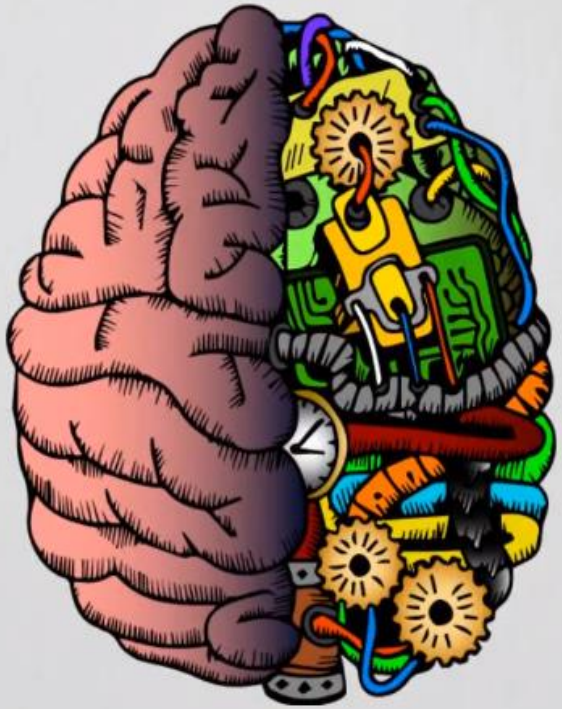
ML is about creating and implementing algorithms that let machines receive data and use this data to:

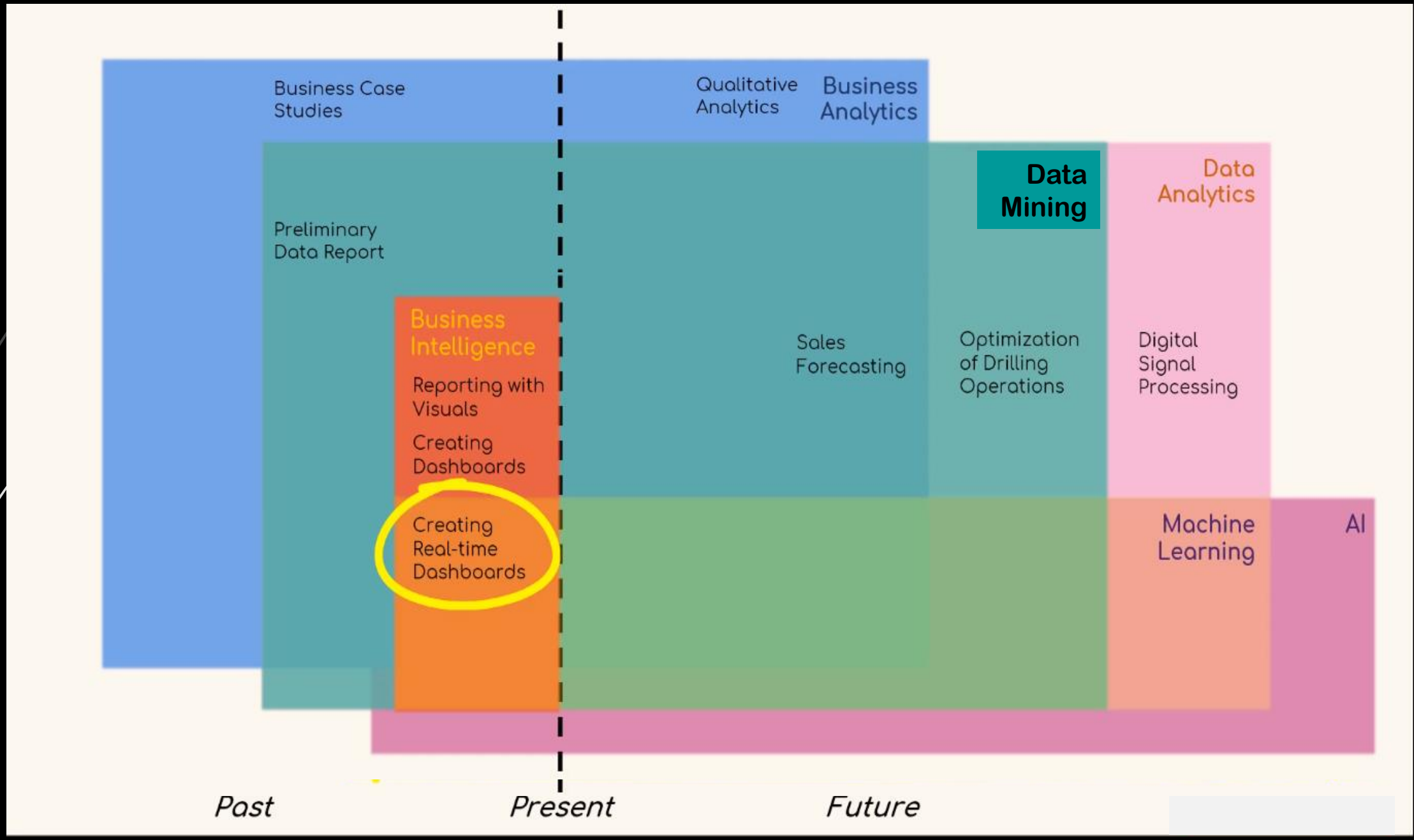
- make predictions
- analyse patterns
- give recommendations


Business Case Studies
Qualitative Analytics
Business Analytics
Data Mining
Data Analytics
Preliminary Data Report
Business Intelligence
Reporting with Visuals
Creating Dashboards
Sales Forecasting
Optimization of Drilling Operations
Digital Signal Processing
applying machine learning tools to the context of business intelligence
Past
Present
Future

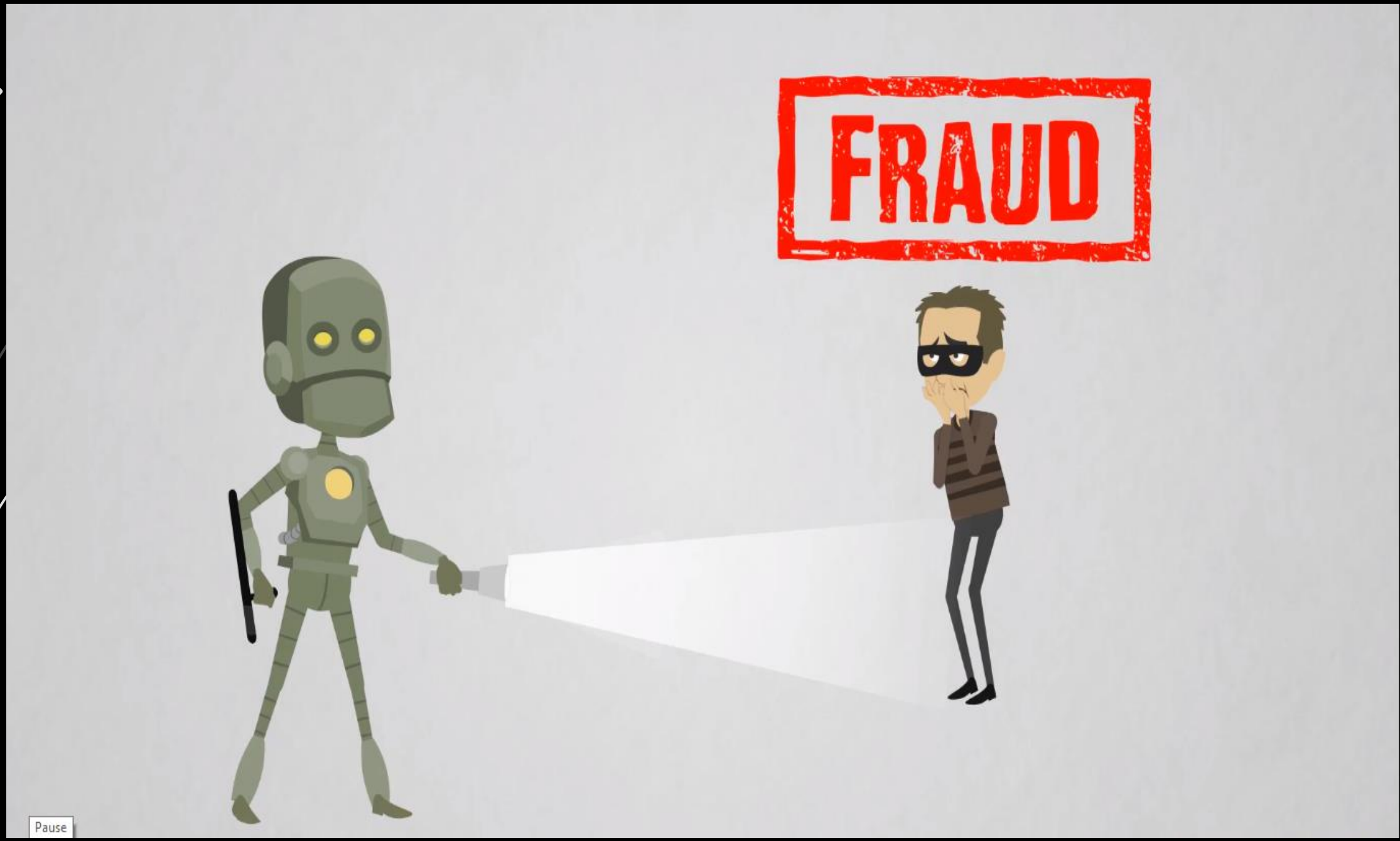
artificial intelligence: simulating human knowledge and decision making with computers

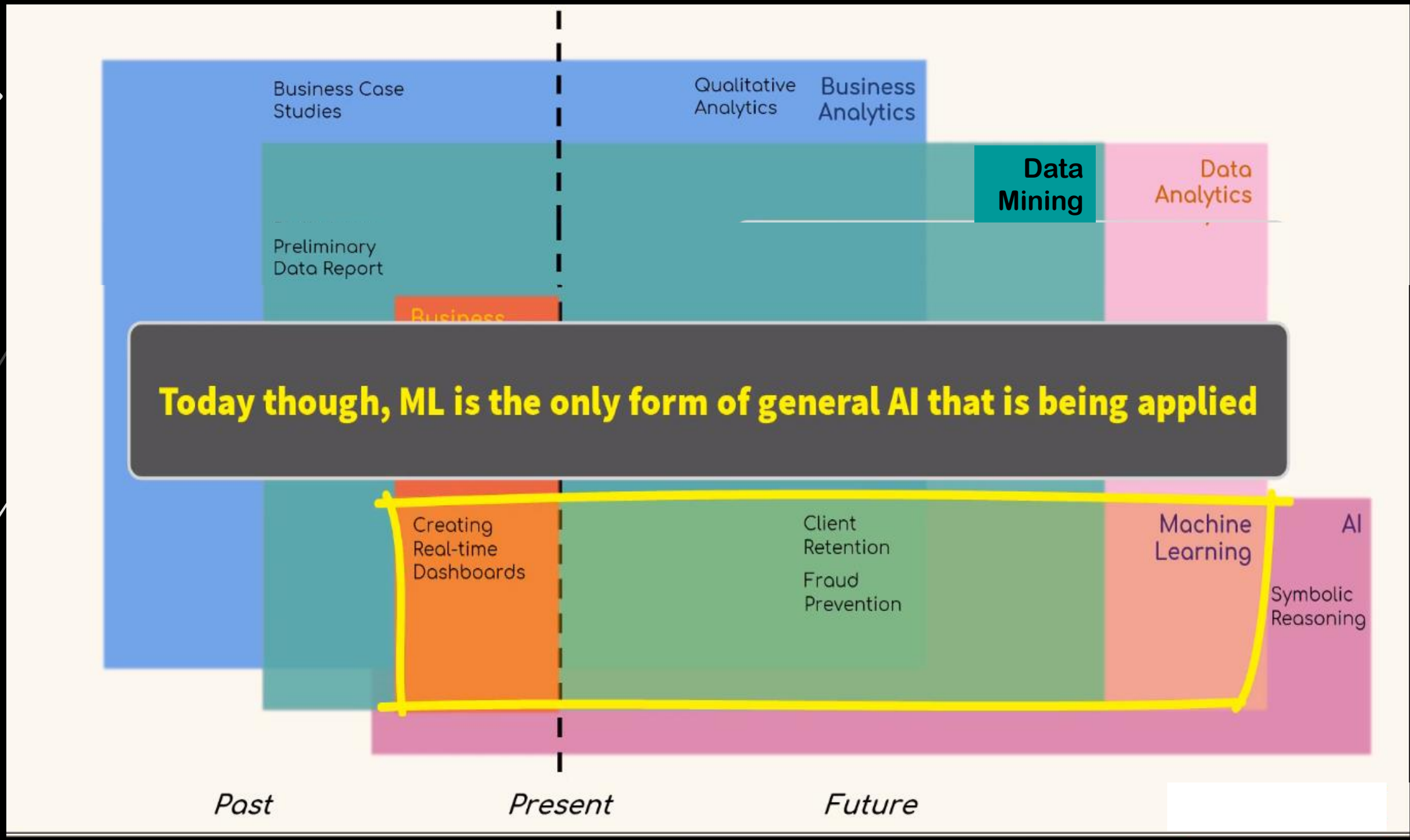
We, as humans, have only managed to reach AI through machine learning




Business Case Studies
Qualitative Analytics
Business Analytics
Data Mining
Data Analytics
Preliminary Data Report
Business Intelligence
Reporting with Visuals
Creating Dashboards
Creating Real-time Dashboards
Sales Forecasting
Optimization of Drilling Operations
Digital Signal Processing
Machine Learning
AI
Past
Present
Future

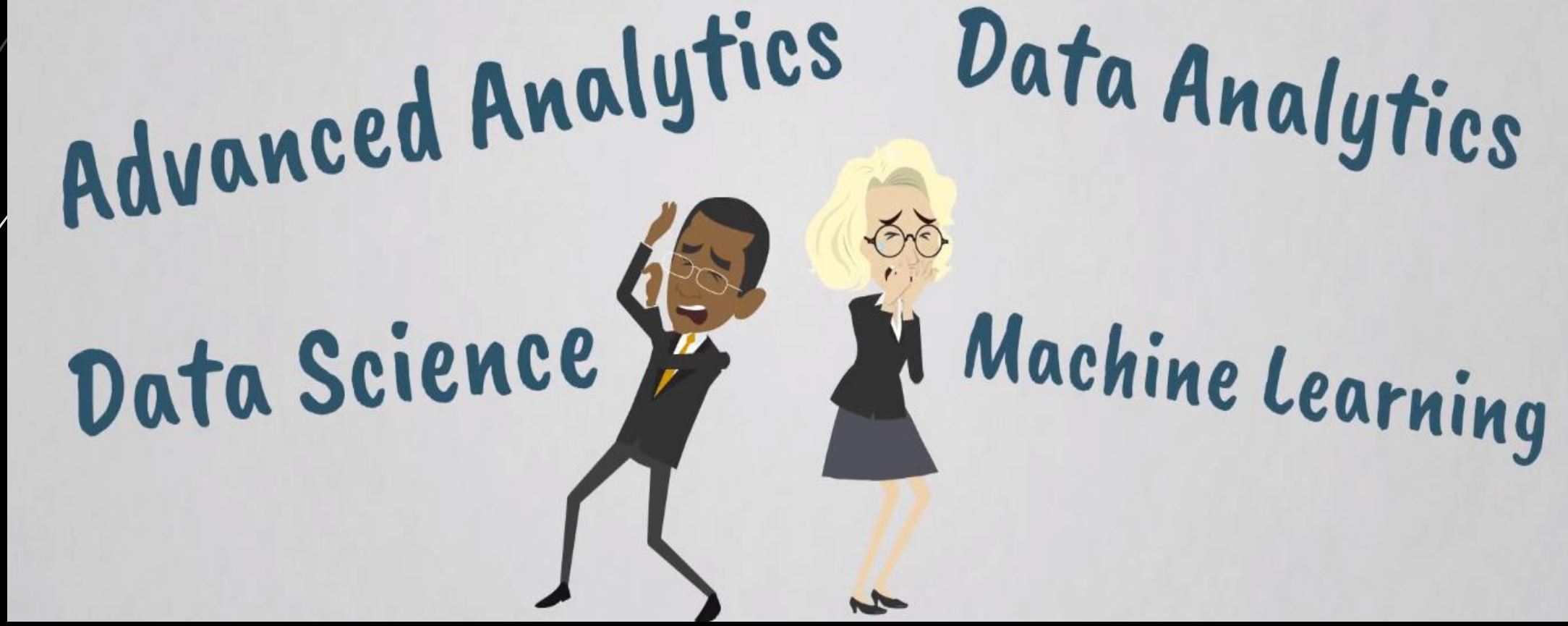
ID	Name	Age	.....
001	John	35	.....
002	Alan	22	.....
.....	.....	.....	.....

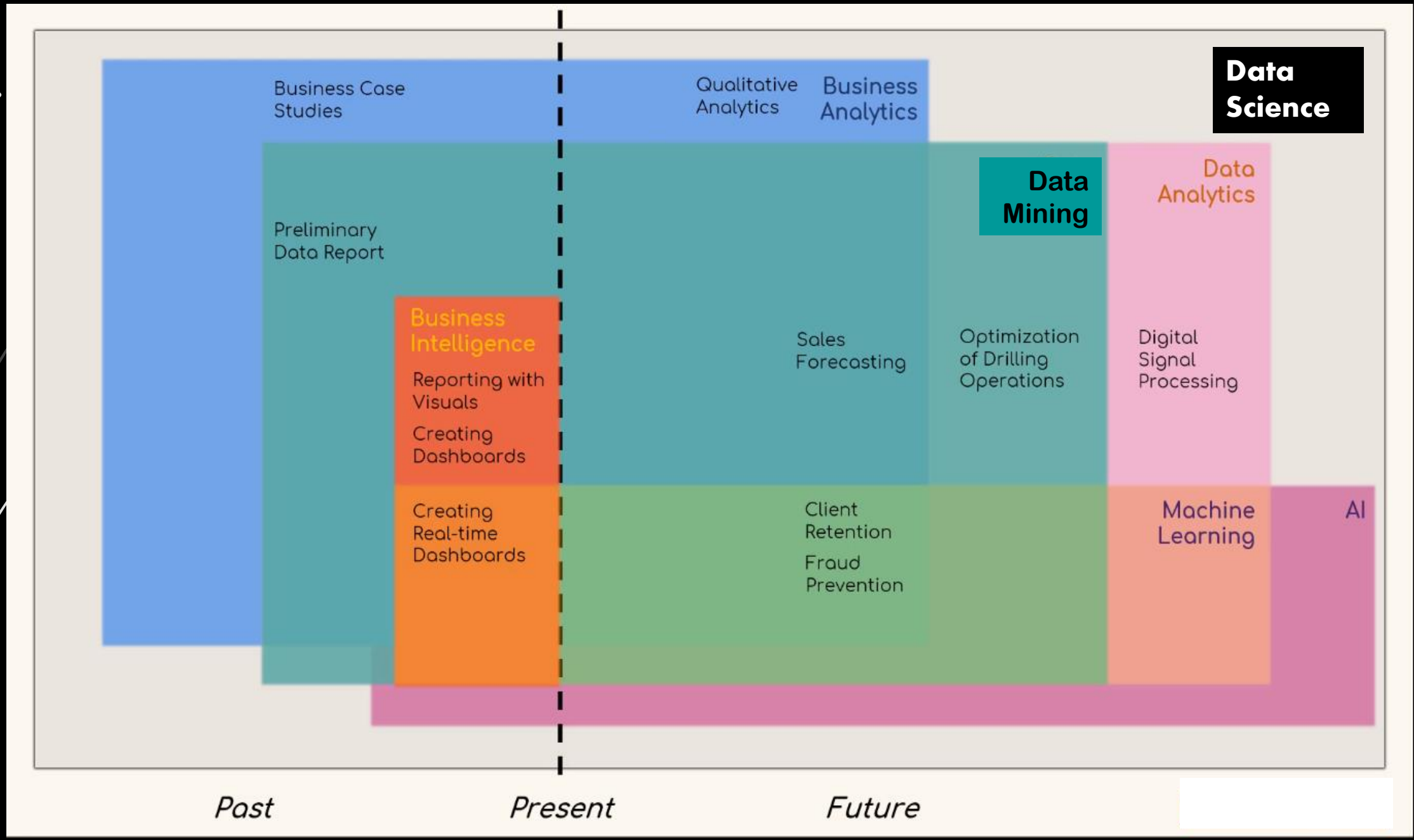

FRAUD
Pause


Business Case Studies
Qualitative Analytics
Business Analytics
Data Mining
Data Analytics
Preliminary Data Report
Today though, ML is the only form of general AI that is being applied
Creating Real-time Dashboards
Client Retention
Fraud Prevention
Machine Learning
AI
Symbolic Reasoning
Past
Present
Future

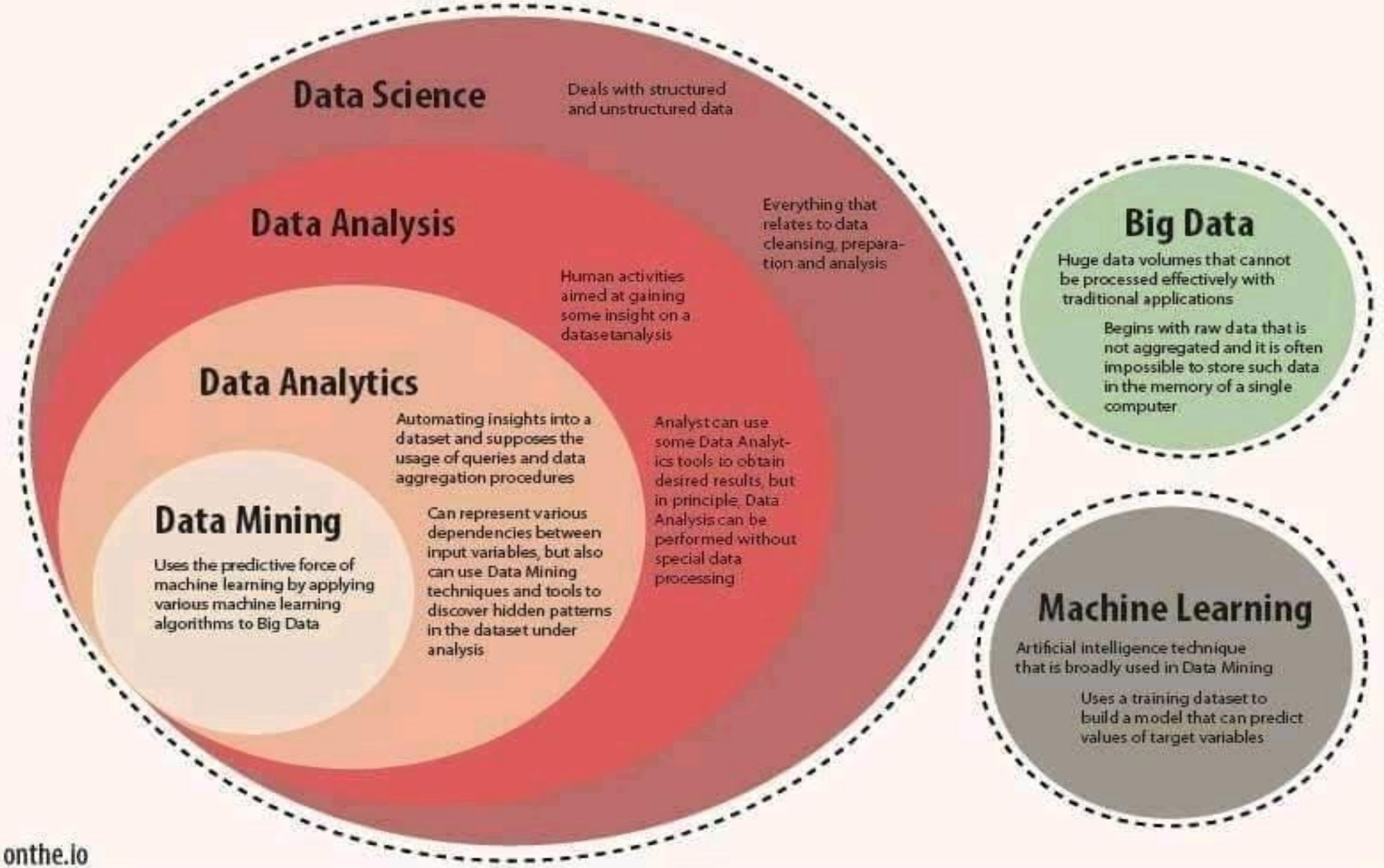
advanced analytics: a marketing term...

and Data Science




Data Science
Business Case Studies
Qualitative Analytics
Business Analytics
Data Mining
Data Analytics
Preliminary Data Report
Business Intelligence
Reporting with Visuals
Creating Dashboards
Sales Forecasting
Optimization of Drilling Operations
Digital Signal Processing
Creating Real-time Dashboards
Client Retention
Fraud Prevention
Machine Learning
AI
Past
Present
Future

What is the difference between Data Science, Data Analysis, Big Data, Data Analytics, Data Mining and Machine Learning?



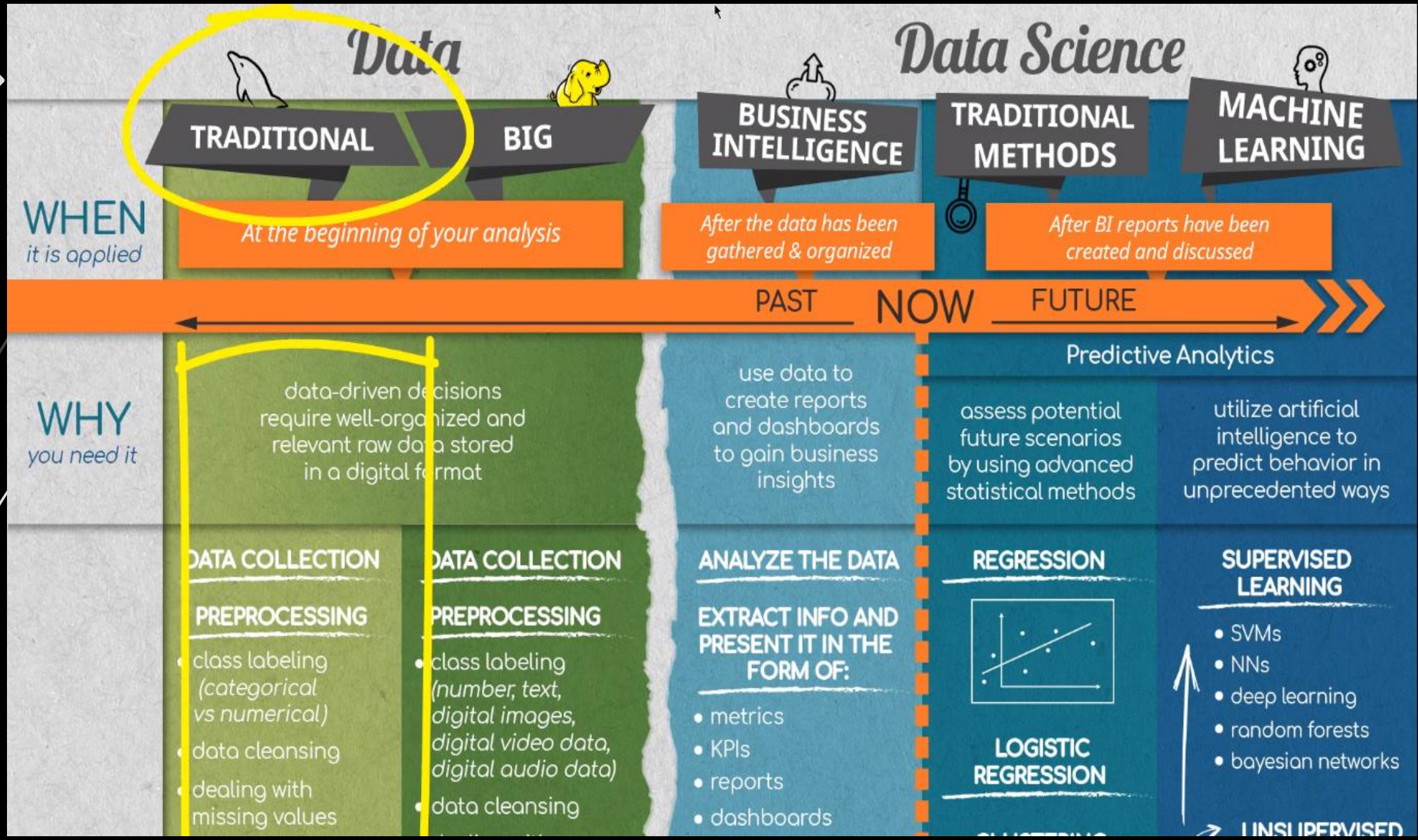
Source: onthe.io

There can be two possible scenarios where You have to use Data Science to Predict


1
?


2
?
?
ID
Name
Age
001
John
35
002
Alan
22

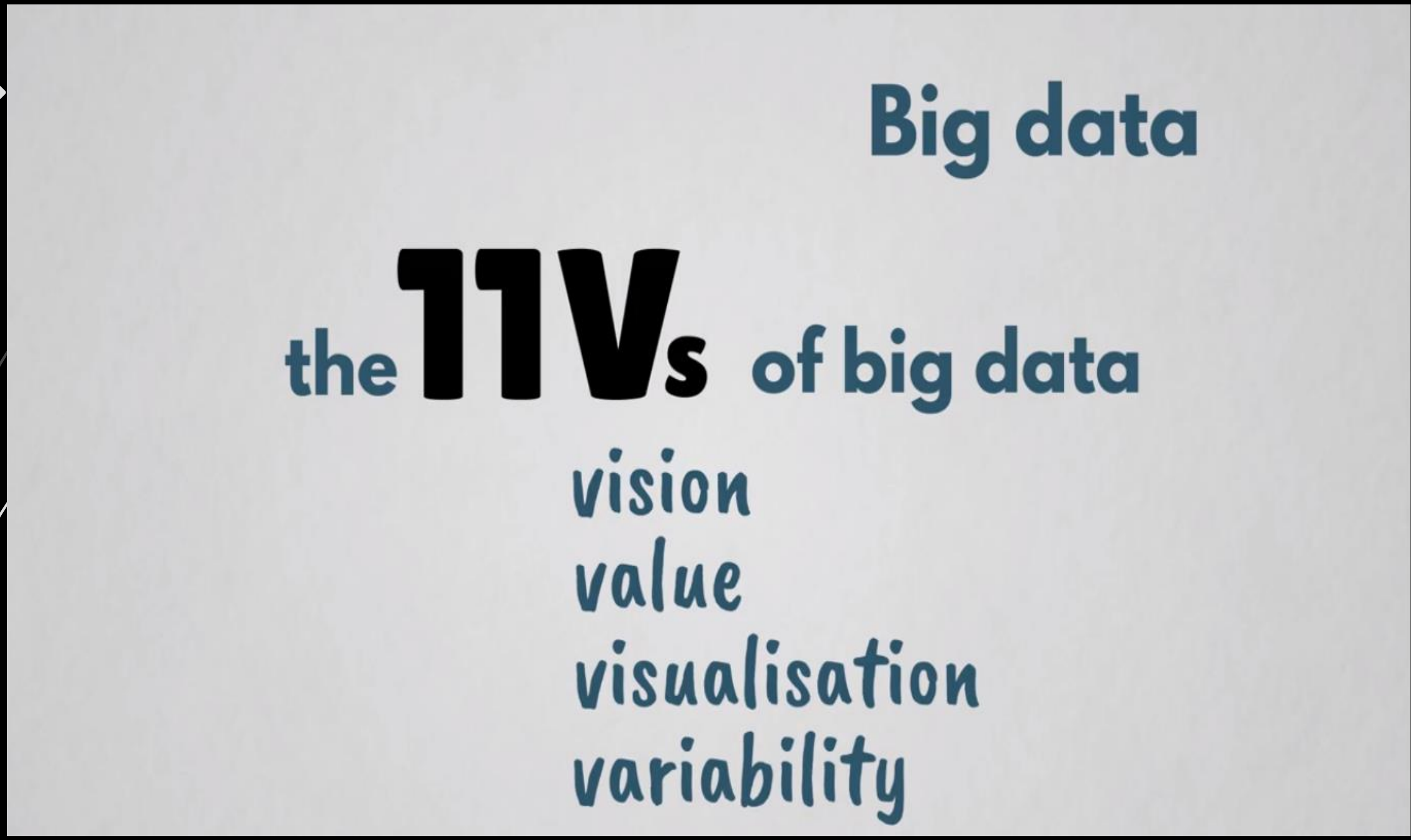
Data and Data Science


Data
Data Science
TRADITIONAL
BIG
BUSINESS INTELLIGENCE
TRADITIONAL METHODS
MACHINE LEARNING
WHEN
it is applied
At the beginning of your analysis
After the data has been gathered & organized
After BI reports have been created and discussed
PAST
NOW
FUTURE
WHY
you need it
data-driven decisions require well-organized and relevant raw data stored in a digital format
use data to create reports and dashboards to gain business insights
Predictive Analytics
assess potential future scenarios by using advanced statistical methods
utilize artificial intelligence to predict behavior in unprecedented ways
DATA COLLECTION
PREPROCESSING
class labeling (categorical vs numerical)
data cleansing
dealing with missing values
DATA COLLECTION
PREPROCESSING
class labeling (number, text, digital images, digital video data, digital audio data)
data cleansing
ANALYZE THE DATA
EXTRACT INFO AND PRESENT IT IN THE FORM OF:
metrics
KPIs
reports
dashboards
REGRESSION
LOGISTIC REGRESSION
SUPERVISED LEARNING
SVMs
NNs
deep learning
random forests
bayesian networks
UNSUPERVISED

Traditional data

- structured
-  can be managed from 1 computer

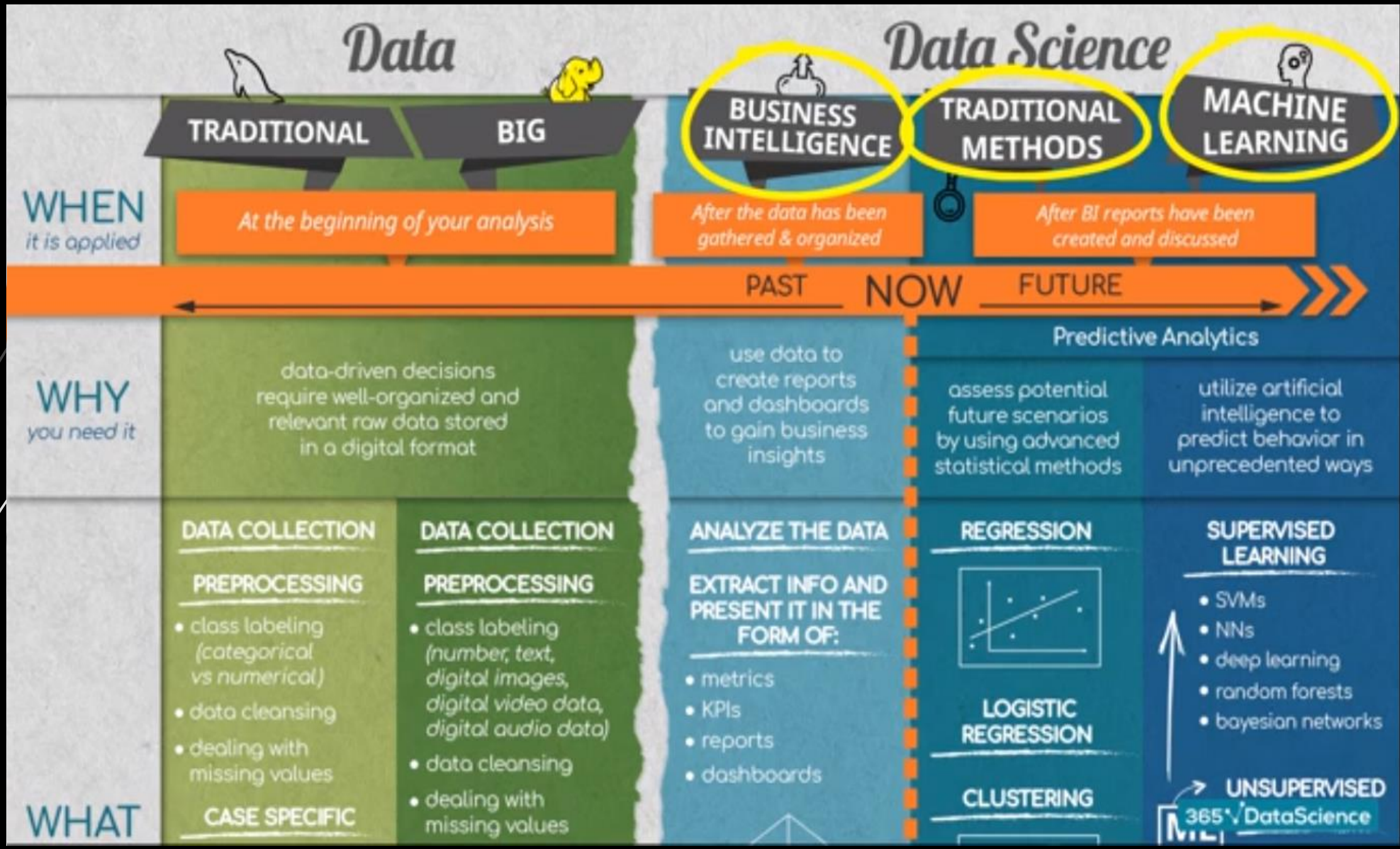
ID	Name	Age	.....
001	John	35	.....
002	Alan	22	.....
.....	.....	.....	.....

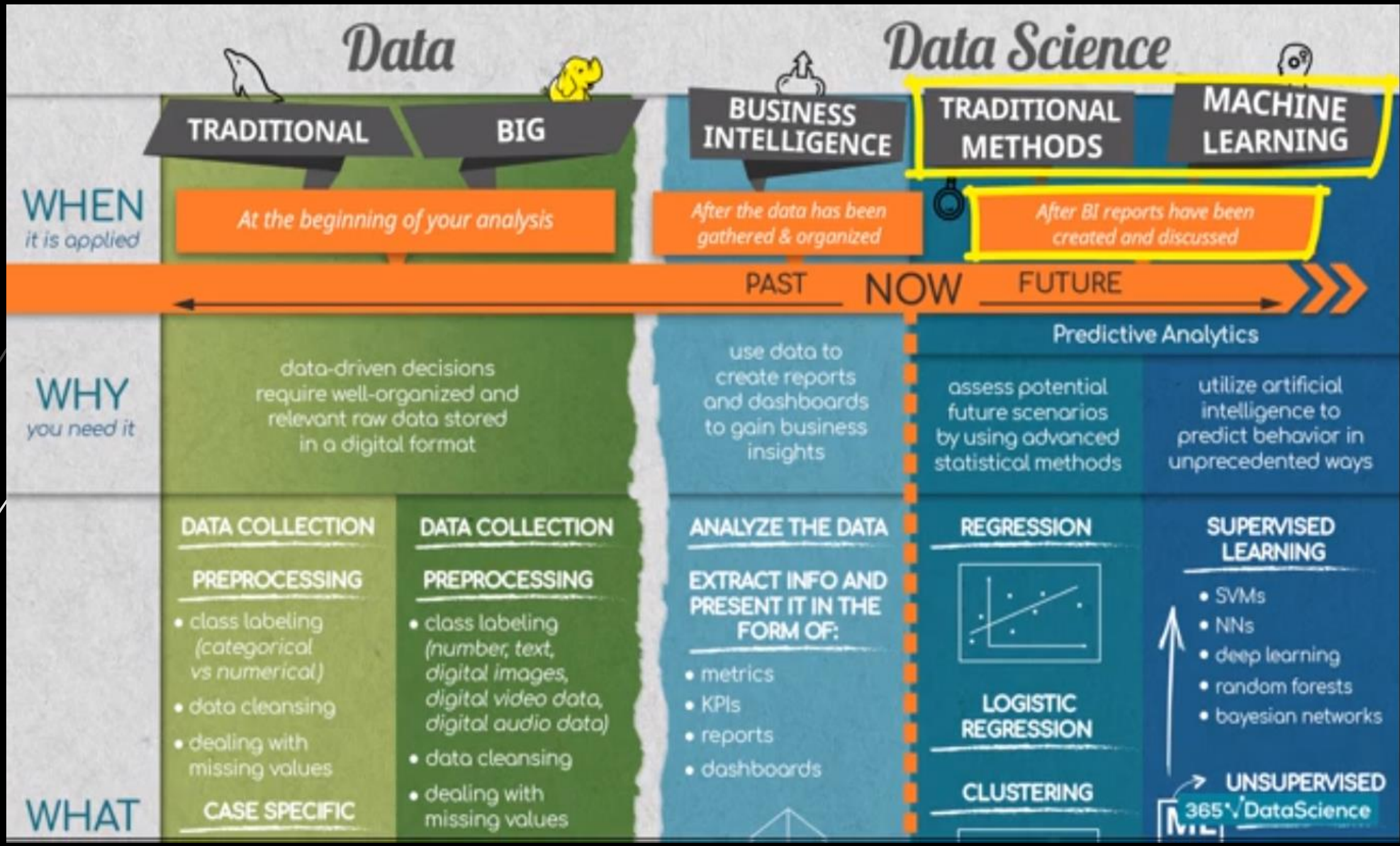

Big data
the 11Vs of big data
vision
value
visualisation
variability

Traditional data Big data

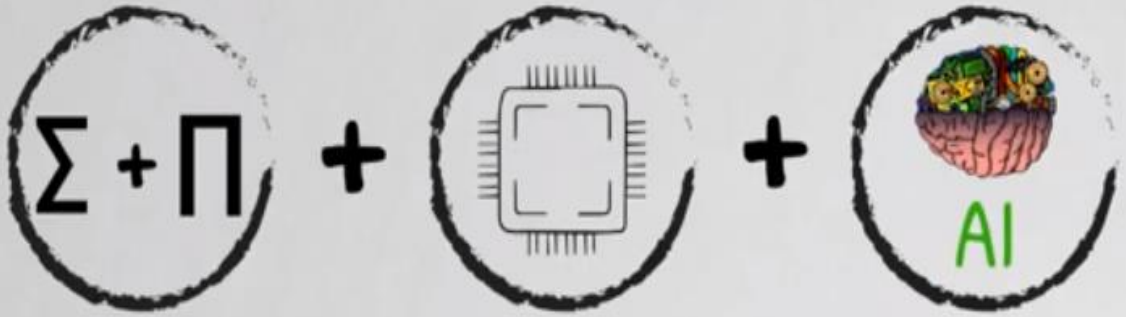
Velocity:

1 retrieved in real-time


Data
Data Science
TRADITIONAL
BIG
BUSINESS INTELLIGENCE
TRADITIONAL METHODS
MACHINE LEARNING
WHEN
it is applied
At the beginning of your analysis
After the data has been gathered & organized
After BI reports have been created and discussed
PAST
NOW
FUTURE
WHY
you need it
data-driven decisions require well-organized and relevant raw data stored in a digital format
use data to create reports and dashboards to gain business insights
Predictive Analytics
assess potential future scenarios by using advanced statistical methods
utilize artificial intelligence to predict behavior in unprecedented ways
WHAT
DATA COLLECTION
PREPROCESSING
• class labeling (categorical vs numerical)
• data cleansing
• dealing with missing values
CASE SPECIFIC
DATA COLLECTION
PREPROCESSING
• class labeling (number, text, digital images, digital video data, digital audio data)
• data cleansing
• dealing with missing values
ANALYZE THE DATA
EXTRACT INFO AND PRESENT IT IN THE FORM OF:
• metrics
• KPIs
• reports
• dashboards
REGRESSION
LOGISTIC REGRESSION
CLUSTERING
SUPERVISED LEARNING
• SVMs
• NNs
• deep learning
• random forests
• bayesian networks
UNSUPERVISED
365 DataScience

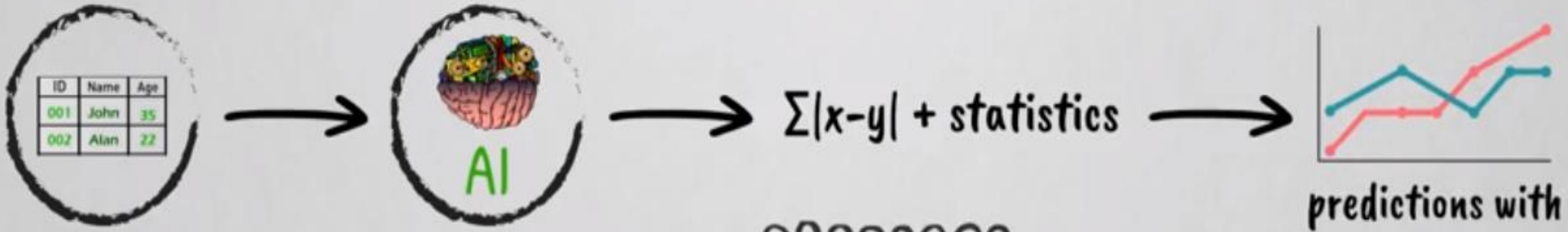

Data
Data Science
TRADITIONAL
BIG
BUSINESS INTELLIGENCE
TRADITIONAL METHODS
MACHINE LEARNING
WHEN
it is applied
At the beginning of your analysis
After the data has been gathered & organized
After BI reports have been created and discussed
PAST
NOW
FUTURE
WHY
you need it
data-driven decisions require well-organized and relevant raw data stored in a digital format
use data to create reports and dashboards to gain business insights
Predictive Analytics
assess potential future scenarios by using advanced statistical methods
utilize artificial intelligence to predict behavior in unprecedented ways
WHAT
DATA COLLECTION
PREPROCESSING
• class labeling (categorical vs numerical)
• data cleansing
• dealing with missing values
CASE SPECIFIC
DATA COLLECTION
PREPROCESSING
• class labeling (number, text, digital images, digital video data, digital audio data)
• data cleansing
• dealing with missing values
ANALYZE THE DATA
EXTRACT INFO AND PRESENT IT IN THE FORM OF:
• metrics
• KPIs
• reports
• dashboards
REGRESSION
LOGISTIC REGRESSION
CLUSTERING
SUPERVISED LEARNING
• SVMs
• NNs
• deep learning
• random forests
• bayesian networks
UNSUPERVISED
365 DataScience

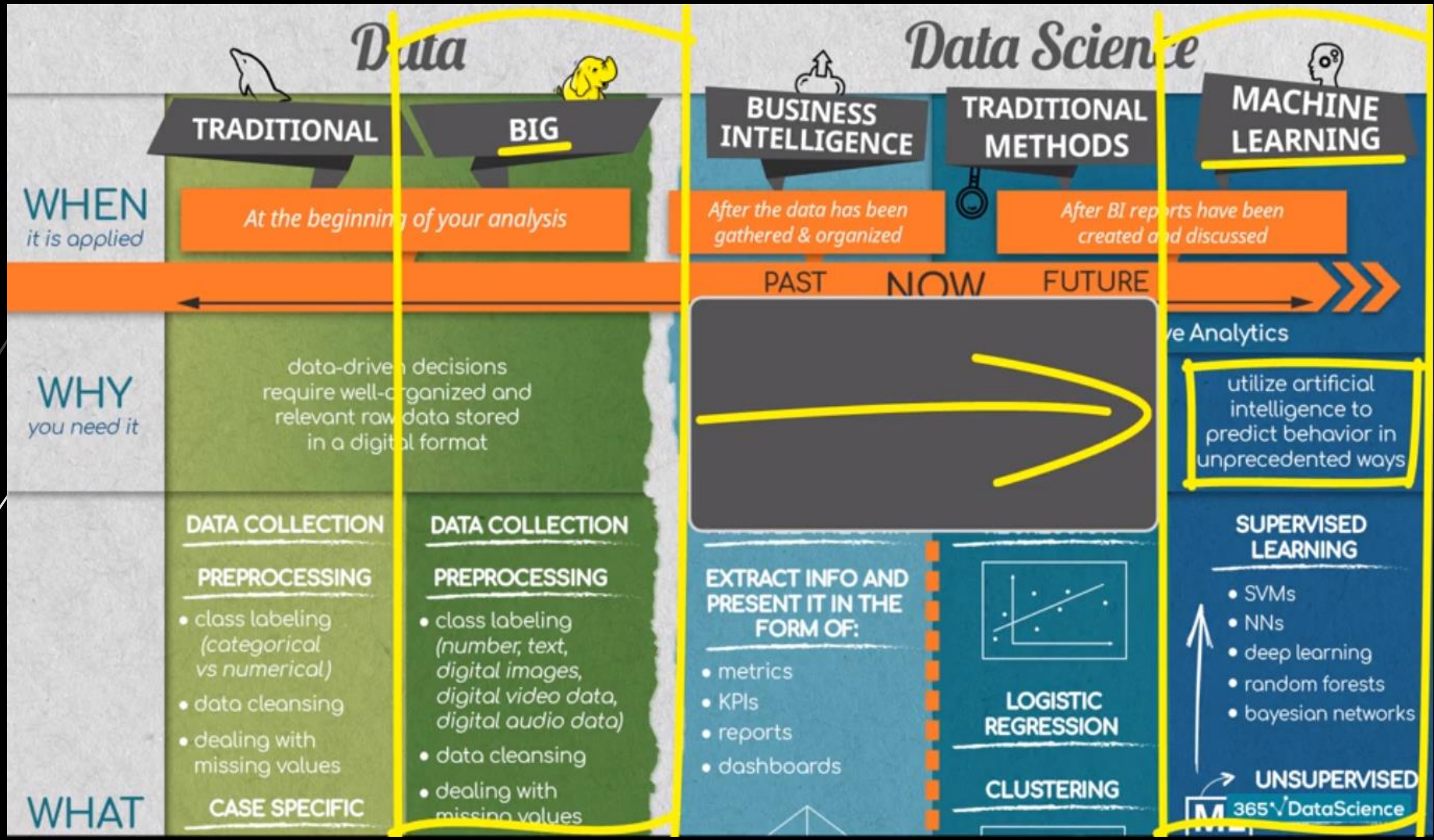
Machine learning



= predict outcomes from data without being explicitly programmed to

Algorithm:

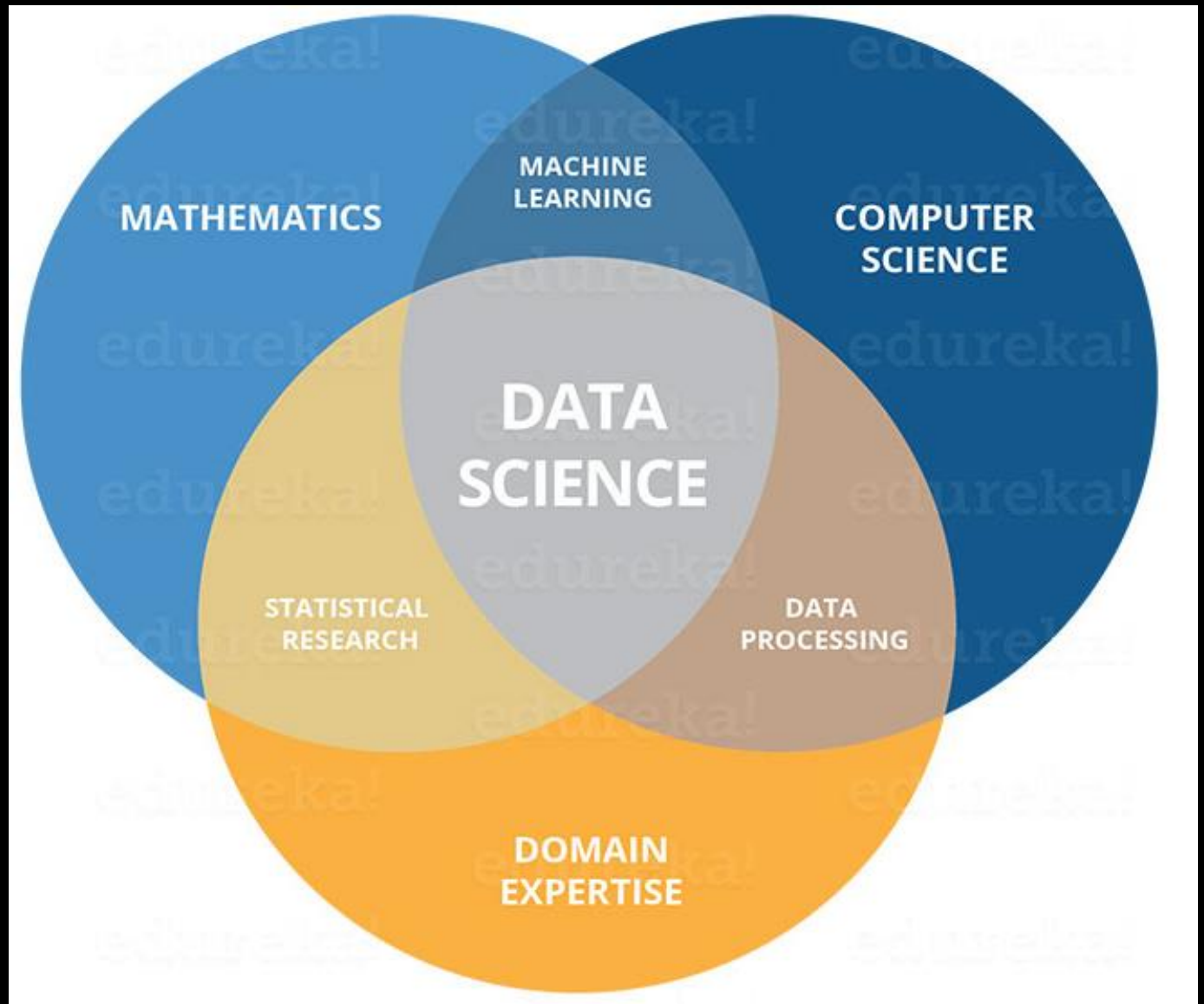



Data
Data Science
TRADITIONAL
BIG
BUSINESS INTELLIGENCE
TRADITIONAL METHODS
MACHINE LEARNING
WHEN
it is applied
At the beginning of your analysis
After the data has been gathered & organized
After BI reports have been created and discussed
PAST
NOW
FUTURE
Analytics
WHY
you need it
data-driven decisions require well-organized and relevant raw data stored in a digital format
utilize artificial intelligence to predict behavior in unprecedented ways
WHAT
DATA COLLECTION
PREPROCESSING
• class labeling (categorical vs numerical)
• data cleansing
• dealing with missing values
CASE SPECIFIC
DATA COLLECTION
PREPROCESSING
• class labeling (number, text, digital images, digital video data, digital audio data)
• data cleansing
• dealing with missing values
EXTRACT INFO AND PRESENT IT IN THE FORM OF:
• metrics
• KPIs
• reports
• dashboards
LOGISTIC REGRESSION
CLUSTERING
SUPERVISED LEARNING
• SVMs
• NNs
• deep learning
• random forests
• bayesian networks
UNSUPERVISED
365 DataScience

What is Data Science?

- Data Science is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the raw data.
- Data Science is primarily used to make decisions and predictions.

What is Data Science?



Data Analysis Vs Data Science

Features	Business Intelligence	Data Science
Data Sources	Structured (Usually SQL, often Data Warehouse)	Both Structured and Unstructured (logs, cloud data, SQL, NoSQL, text, Tweeter Feed)
Approach	Statistics and Visualization	Statistics, Machine Learning, Graph Analysis, Neuro- linguistic Programming (NLP)
Focus	Past and Present	Present and Future
Tools	Pentaho, Microsoft BI, QlikView, R	RapidMiner, BigML, Weka, R

Application of Data Science

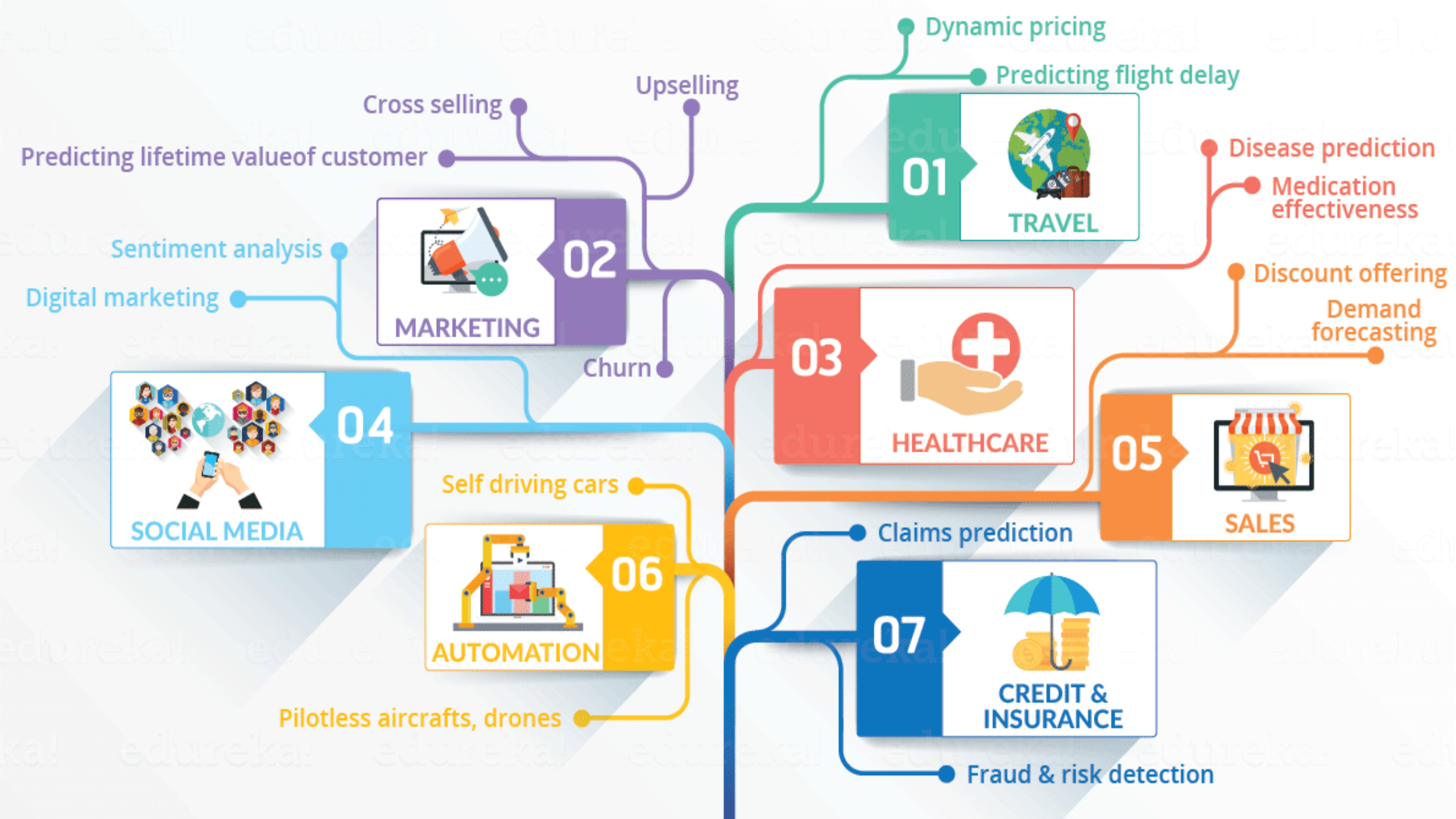
Recommend the right product to right customer to enhance business

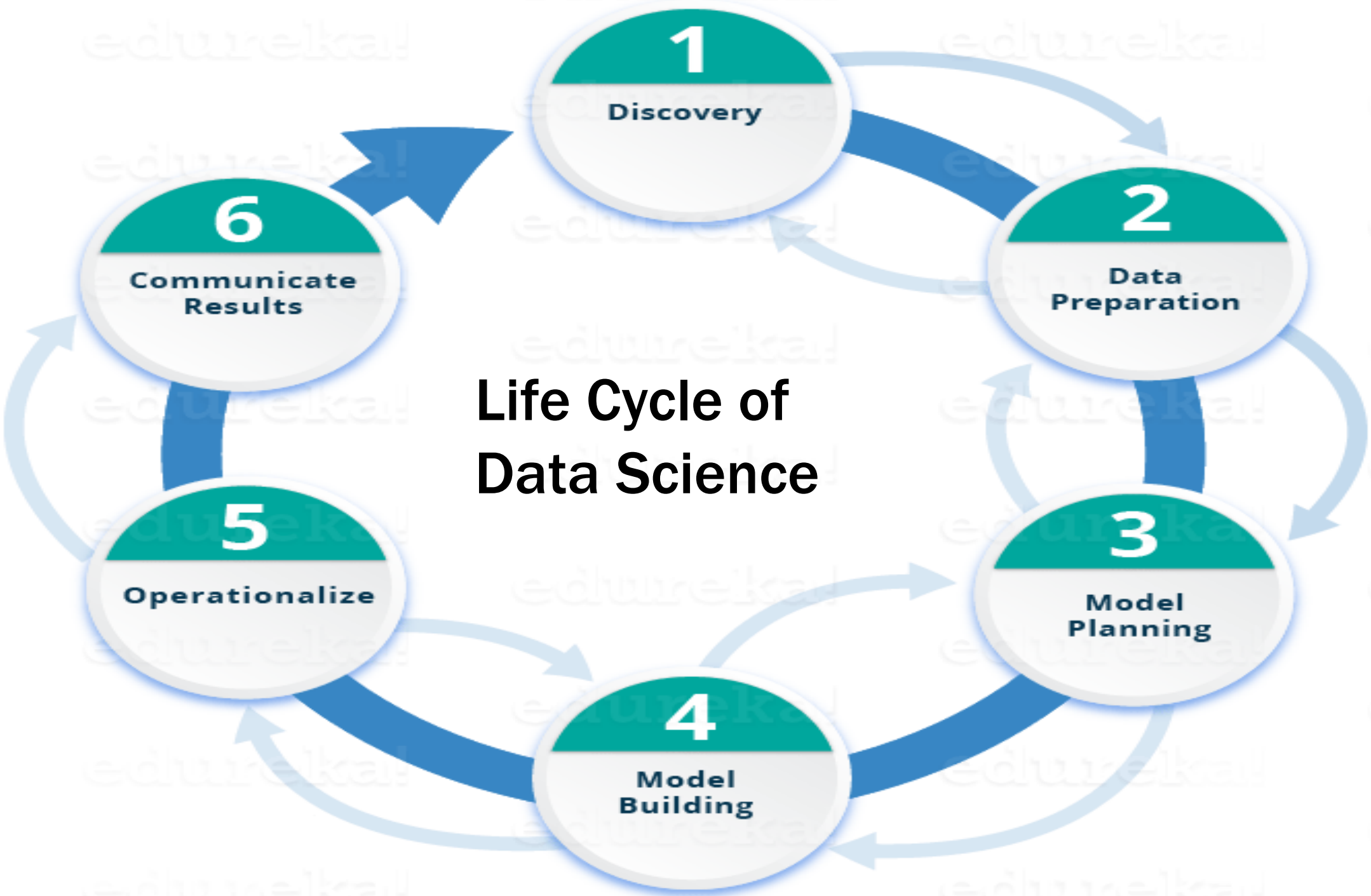
Predict the characteristics of high LTV customers and helps in customer segmentation

Build Intelligence and ability in Machines

Predict fraudulent transaction beforehand

Perform sentiment analysis to predict the outcome of elections


Dynamic pricing
Predicting flight delay
01
TRAVEL
Disease prediction
Medication effectiveness
Upselling
Cross selling
Predicting lifetime valueof customer
02
MARKETING
Sentiment analysis
Digital marketing
Churn
03
HEALTHCARE
Discount offering
Demand forecasting
04
SOCIAL MEDIA
05
SALES
Self driving cars
06
AUTOMATION
Claims prediction
07
CREDIT & INSURANCE
Pilotless aircrafts, drones
Fraud & risk detection


1
Discovery
2
Data Preparation
3
Model Planning
4
Model Building
5
Operationalize
6
Communicate Results
Life Cycle of Data Science

Phase 1—Discovery

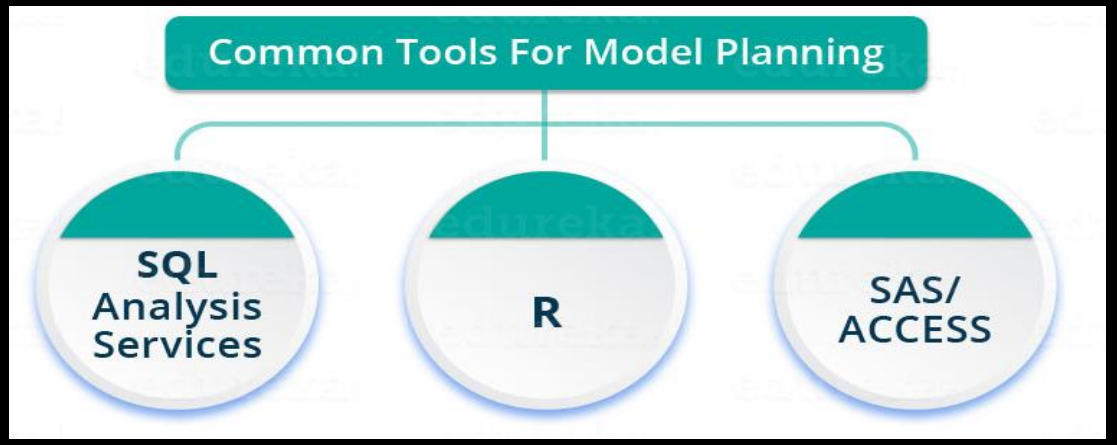
- Understand - Specifications, Requirements, Priorities and Required Budget.
- Must possess the ability to ask the right questions.
- Assess the availability of required resources present in terms of **people**, **technology**, **time** and **data** to support the project.
- Frame the business problem and formulate initial hypotheses (IH) to test.

Phase 2—Data Preparation

Preparing the analytics Sandbox → Performing ETLT → Data Conditioning → Survey & Visualize

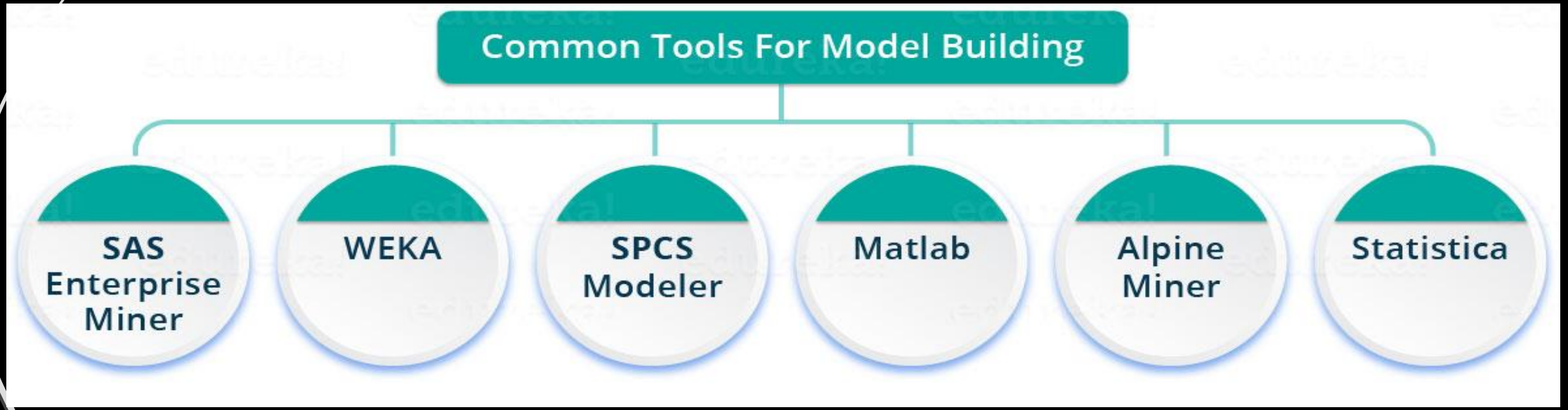
Phase 3 — Model Planning

- Determine the methods and techniques to draw the relationships between variables.
- These relationships will set the base for the algorithms which will be implemented in the next phase.
- Apply Exploratory Data Analytics (EDA) using various statistical formulas and visualization tools.



Phase 4 — Model Building

- Develop datasets for training and testing purposes.
- Consider whether the existing tools will suffice for running the models or it will need a more robust environment (like fast & parallel processing).
- Analyze various learning techniques like classification, association and clustering to build the model.



Phase 5 — Operationalize

- Deliver final reports, briefings, code and technical documents.
- Pilot project is also implemented in a real-time production environment to provide the clear picture of the performance and other related constraints on a small scale before full deployment

Phase 6 — Communicate Results

- Evaluate if planned goals have been achieved.
- So, in the last phase, identify all the key findings, communicate to the stakeholders and determine if the results of the project are a success or a failure based on the criteria developed in Phase 1.


Now, lets understand Data Science with the help of some use cases.

Case Study: Diabetes Prevention

- Step 1: Data Discovery

;npreg;glu;bp;skin;bmi;ped;age,income
1;6;148;72;35;33.6;0.627;50
2;1;85;66;29;26.6;0.351;31
3;1;89;80;23;28.1;0.167;21
4;3;78;50;32;31;0.248;26
5;2;197;70;45;30.5;0.158;53
6;5;166;72;19;25.8;0.587;51
7;0;118;84;47;45.8;0.551;31
8;1;103;30;38;43.3;0.183;33
9;3;126;88;41;39.3;0.704;27
10;9;119;80;35;29;0.263;29
11;1;97;66;15;23.2;0.487;22
12;5;109;75;26;36;0.546;60
13;3;88;58;11;24.8;0.267;22
14;10;122;78;31;27.6;0.512;45
15;4;97;60;33;24;0.966;33
16;9;102;76;37;32.9;0.665;46
17;2;90;68;42;38.2;0.503;27
18;4;111;72;47;37.1;1.39;56
19;3;180;64;25;34;0.271;26
20;7;106;92;18;39;0.235;48
21;9;171;110;24;45.4;0.721;54

Attributes:

npreg – Number of times pregnant
glucose – Plasma glucose concentration
bp – Blood pressure
skin – Triceps skinfold thickness
bmi – Body mass index
ped – Diabetes pedigree function
age – Age
income – Income

Case Study: Diabetes Prevention

- Step 2: Data Preparation

	npreg	glu	bp	skin	bmi	ped	age	income
1	6	148	72	35	33.6	0.627	50	
2	1	85	66	29	26.6	0.351	31	
3	1	89	6600	23	28.1	0.167	21	
4	3	78	50	32	31	0.248	26	
5	2	197	70	45	30.5	0.158	53	
6	5	166	72	19	25.8	0.587	51	
7	0	118	84	47	45.8	0.551	31	
8	one	103	30	38	43.3	0.183	33	
9	3	126	88	41	39.3	0.704	27	
10	9	119	80	35	29	0.263	29	
11	1	97	66	15	23.2	0.487	22	
12	5	109	75	26	36	0.546	60	
13	3	88	58	11	24.8	0.267	22	
14	10	122	78	31	27.6	0.512	45	
15	4		60	33	24	0.966	33	
16	9	102	76	37	32.9	0.665	46	
17	2	90	68	42	38.2	0.503	27	
18	4	111	72	47	37.1	1.39	56	
19	3	180	64	25	34	0.271	26	
20	7	106	92	18		0.235	48	
21	9	171	110	24	45.4	0.721	54	

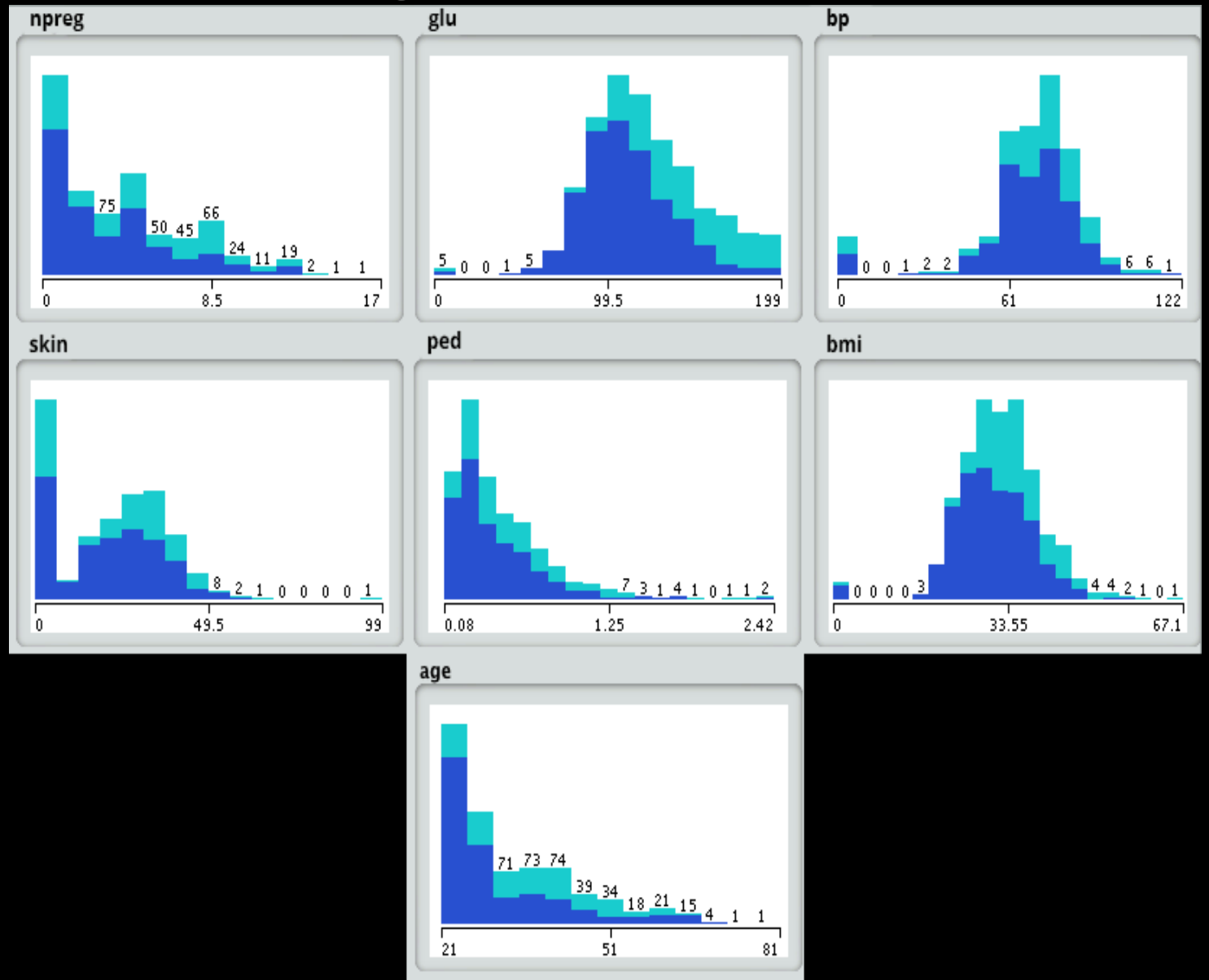
Case Study: Diabetes Prevention

Step 2: Data Preparation

	npreg	glu	bp	skin	bmi	ped	age
1	6	148	72	35	33.6	0.627	50
2	1	85	66	29	26.6	0.351	31
3	1	89	80	23	28.1	0.167	21
4	3	78	50	32	31	0.248	26
5	2	197	70	45	30.5	0.158	53
6	5	166	72	19	25.8	0.587	51
7	0	118	84	47	45.8	0.551	31
8	1	103	30	38	43.3	0.183	33
9	3	126	88	41	39.3	0.704	27
10	9	119	80	35	29	0.263	29
11	1	97	66	15	23.2	0.487	22
12	5	109	75	26	36	0.546	60
13	3	88	58	11	24.8	0.267	22
14	10	122	78	31	27.6	0.512	45
15	4	97	60	33	24	0.966	33
16	9	102	76	37	32.9	0.665	46
17	2	90	68	42	38.2	0.503	27
18	4	111	72	47	37.1	1.39	56
19	3	180	64	25	34	0.271	26
20	7	106	92	18	39	0.235	48
21	9	171	110	24	45.4	0.721	54

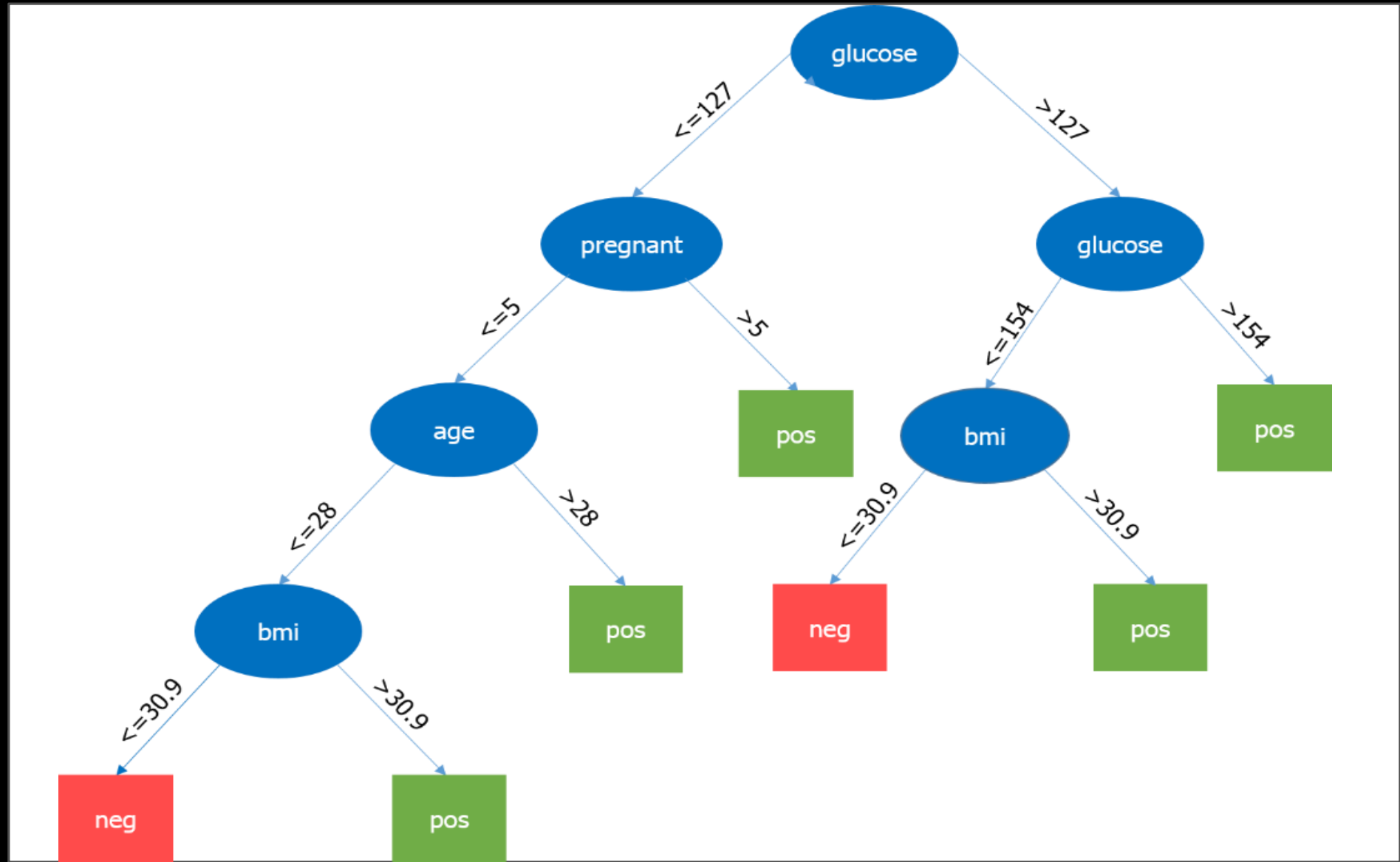
Case Study: Diabetes Prevention

- Step 3: Model Planning



Case Study: Diabetes Prevention

- Step 4: Model Building



Phase 5 — Operationalize

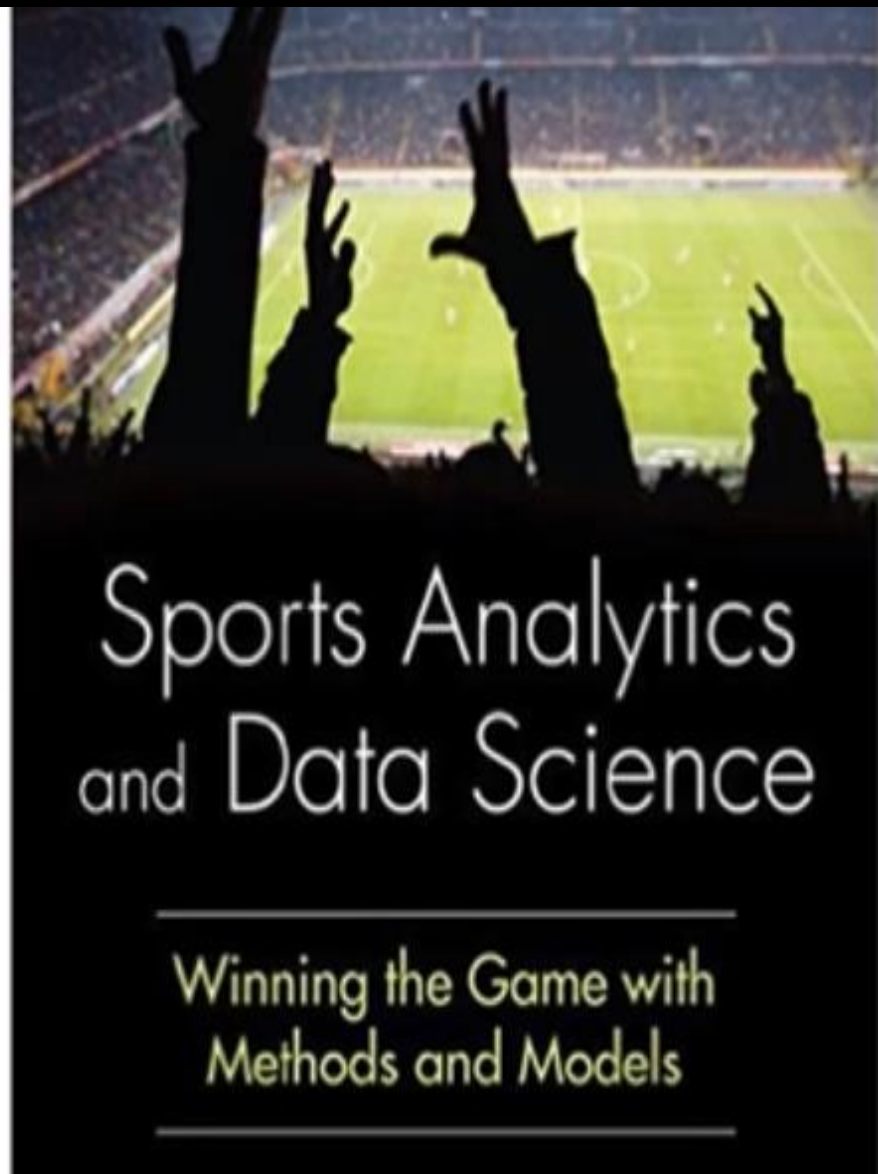
- Run a small pilot project to check if our results are appropriate. Look for performance constraints if any.
- If the results are not accurate, then we need to replan and rebuild the model.

Phase 6 — Communicate Results

- Share the output for full deployment.

Few More Use Cases

- Basketball teams are using data for tracking team strategies and outcome of matches.
- Below parameters will be used for model building.
  - Average pass time of ball.
  - Number of successful passes.
  - Speed and accuracy of successful baskets.
  - Area of court the player on average is shadowing.
- Models built on the basis of data science algorithms help in pattern discovery of player game.



Few More Use Cases

- Amazon has huge amount of consumer purchasing data.
- The data consists of consumer demographics (age, sex, location), purchasing history, past browsing history.
- Based on this data, Amazon segments its customers, draws a pattern and recommends the right product to the right customer at the right time.

Few More Use Cases

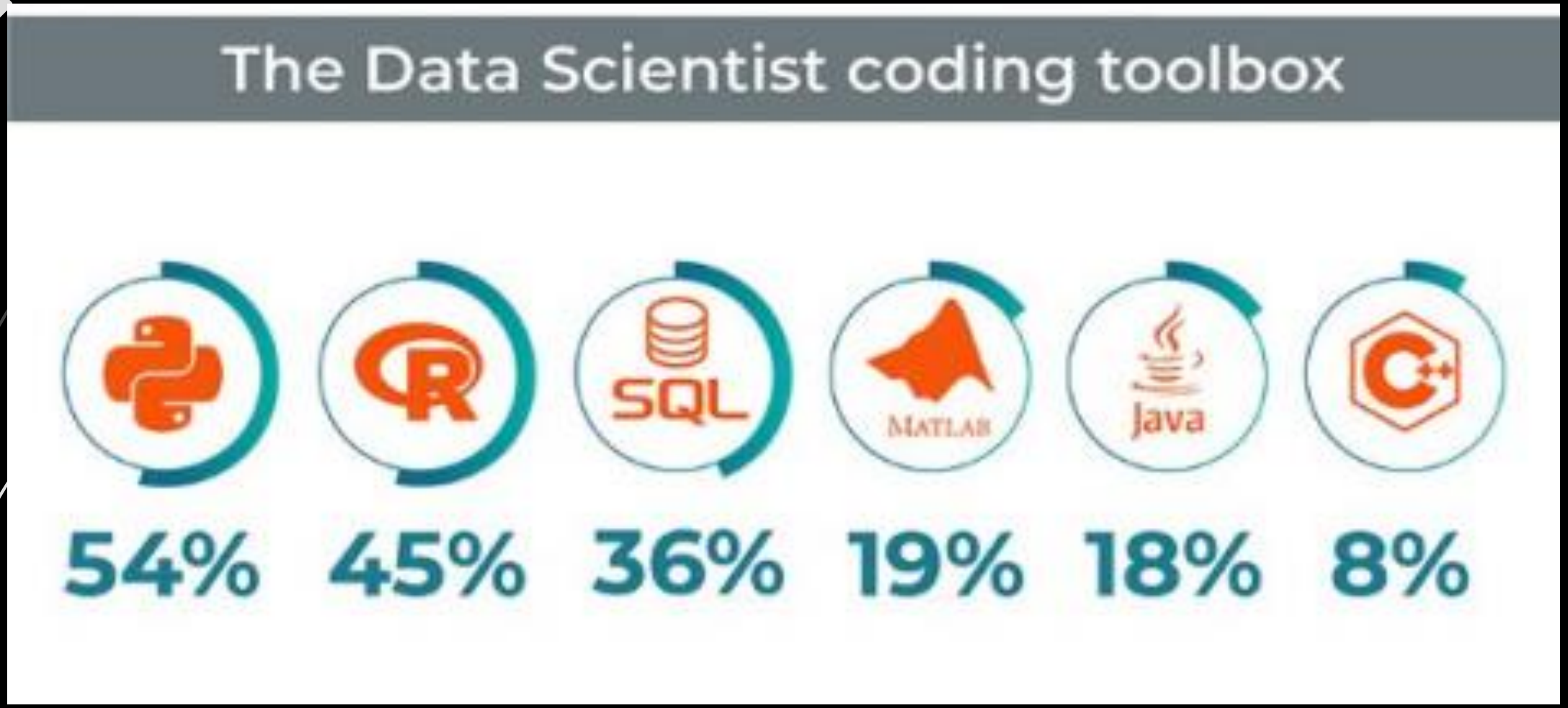
- Google self driving car is a smart, driverless car.
- It collects data from environment through sensors.
- Takes decisions like when to speed up, when to speed down, when to overtake and when to turn.

Role of Data Scientist

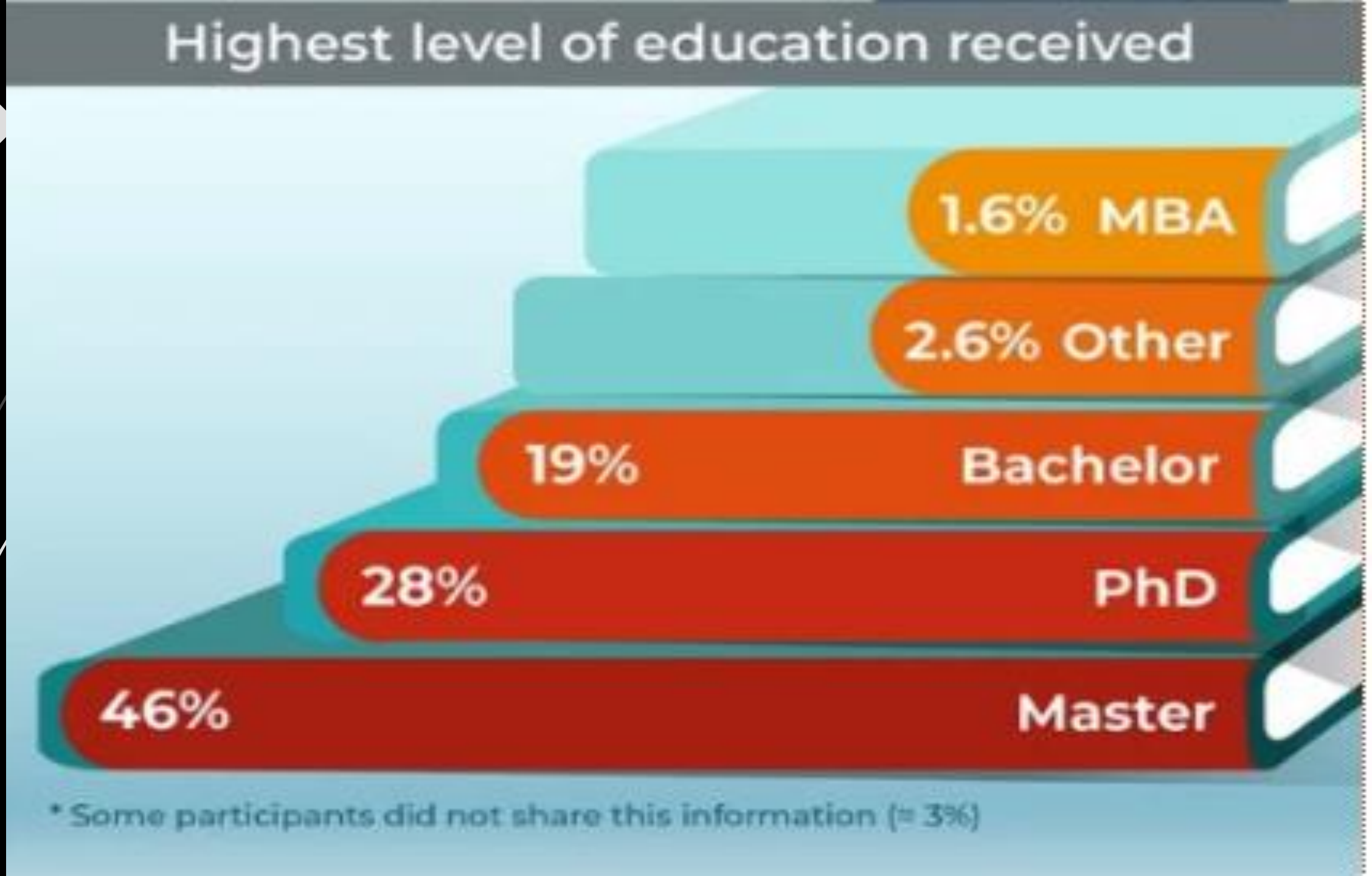
The Data Scientist will be responsible for designing and creating processes and layouts for complex, large-scale data sets used for modeling, data mining, and research purposes.

Responsibilities

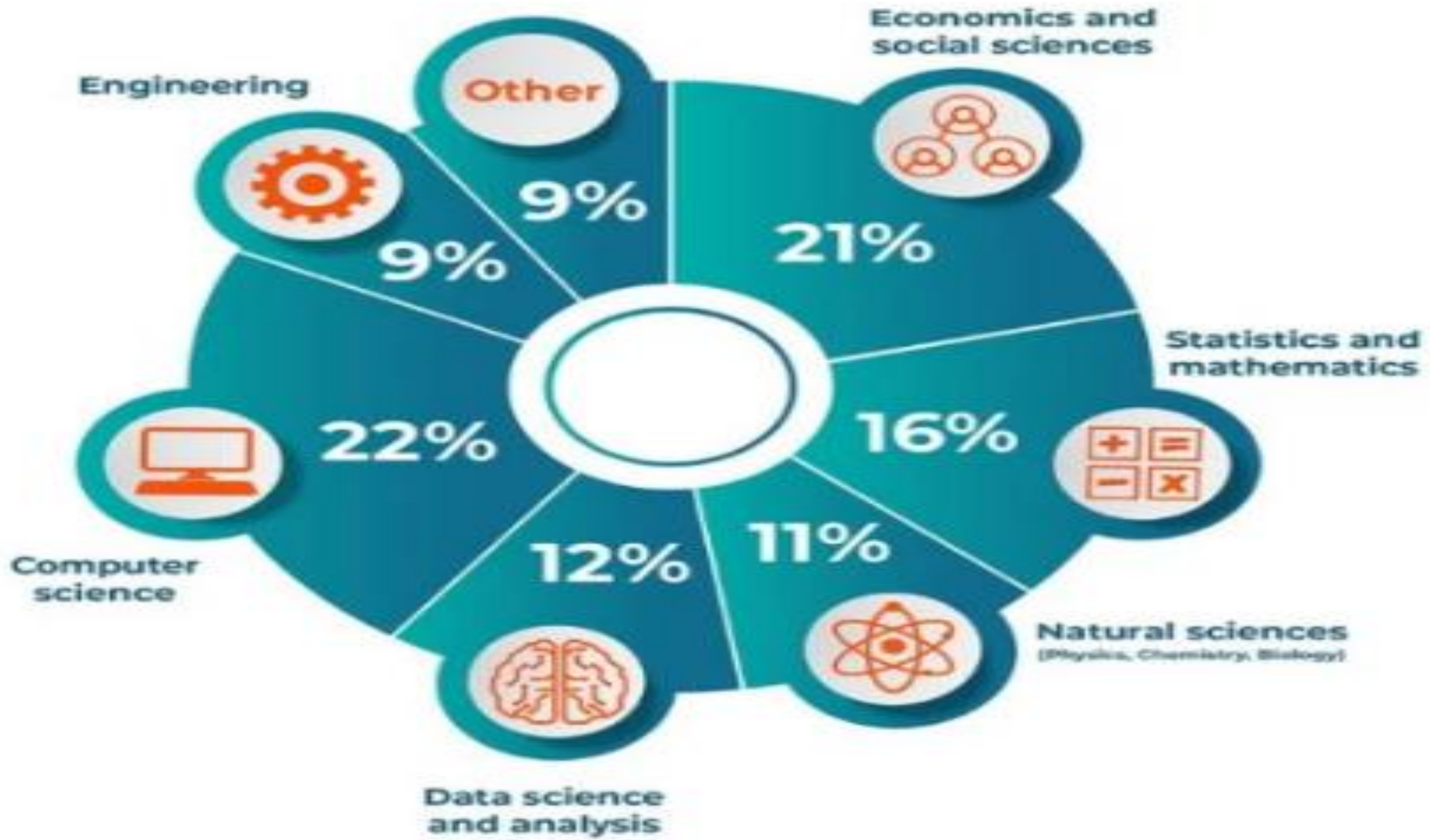
- Selecting features, building and optimizing classifiers using machine learning techniques.
- Data mining using state-of-the-art methods.
- Extending company's data with third party sources of information when needed.
- Processing, cleansing, and verifying the integrity of data for analysis.
- Building predictive models using Machine Learning algorithms.


The Data Scientist coding toolbox
SQL
MATLAB
Java
54%
45%
36%
19%
18%
8%

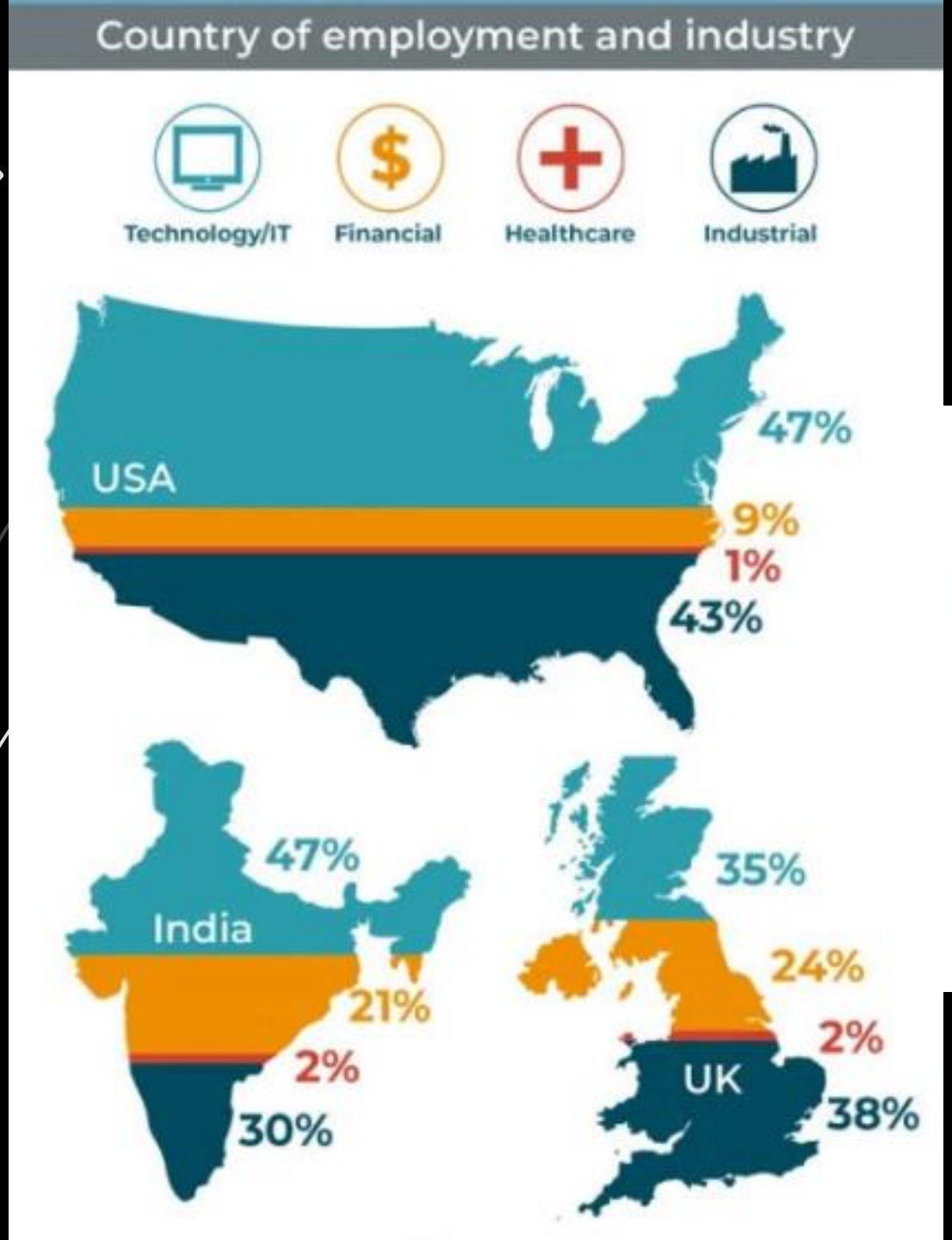
Certifications and accomplishments

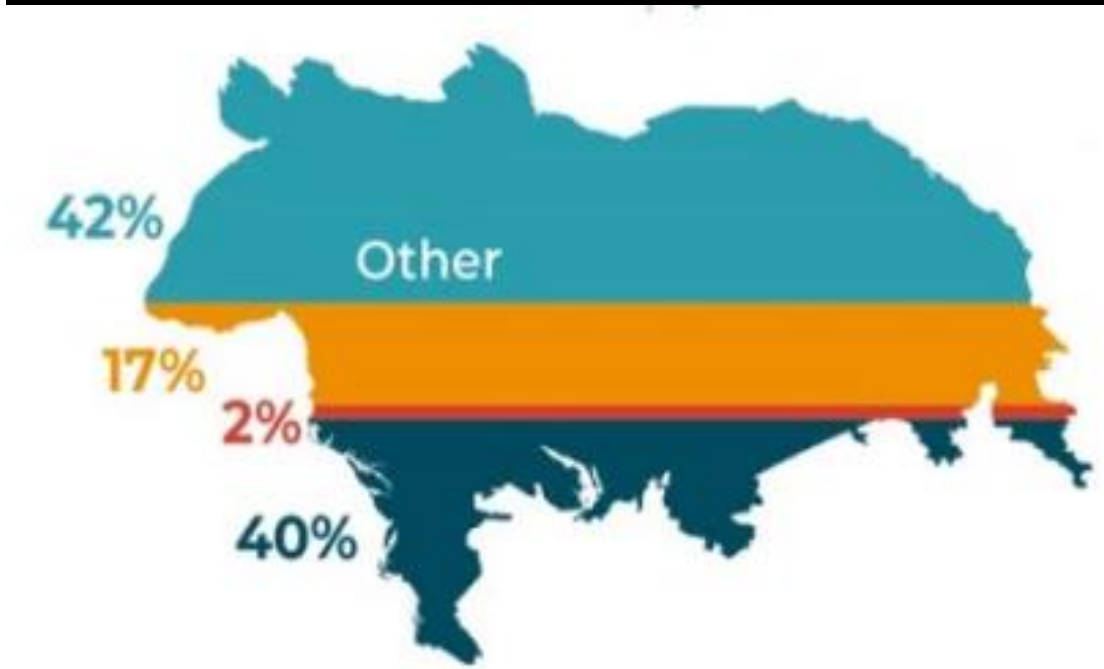

Highest level of education received
1.6% MBA
2.6% Other
19% Bachelor
28% PhD
46% Master
* Some participants did not share this information (= 3%)

Area of academic studies



Industries hiring Data Scientists


Country of employment and industry
Technology/IT
Financial
Healthcare
Industrial
USA
47%
9%
1%
43%
India
47%
21%
2%
30%
UK
35%
24%
2%
38%


Other
42%
17%
2%
40%

Think Big, Start Small, Scale Fast and Innovate in the era of disruption

THANKYOU

- Dr. Neha Sharma
- www.drnehasharma.in
- nvsharma@rediffmail.com
- 9923602490