Data Science

What is Data Science

What Will You Learn Today?

1 Need Of Data Science

2 What is Data Science

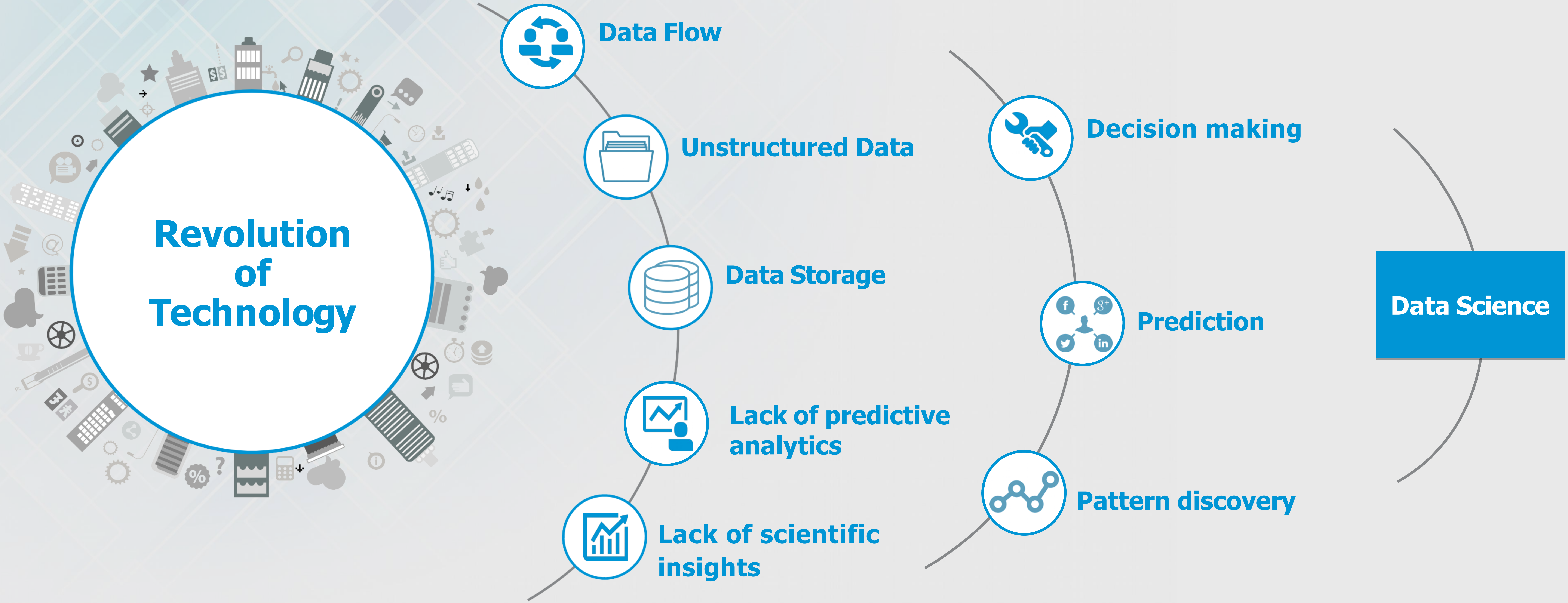
3 Use case of Data Science

4 Business Intelligence vs. Data Science

5 Tools used in Data Science

6 Lifecycle of Data Science

Need Of Data Science



Revolution of Technology

Data Flow

Unstructured Data

Data Storage

Lack of predictive analytics

Lack of scientific insights

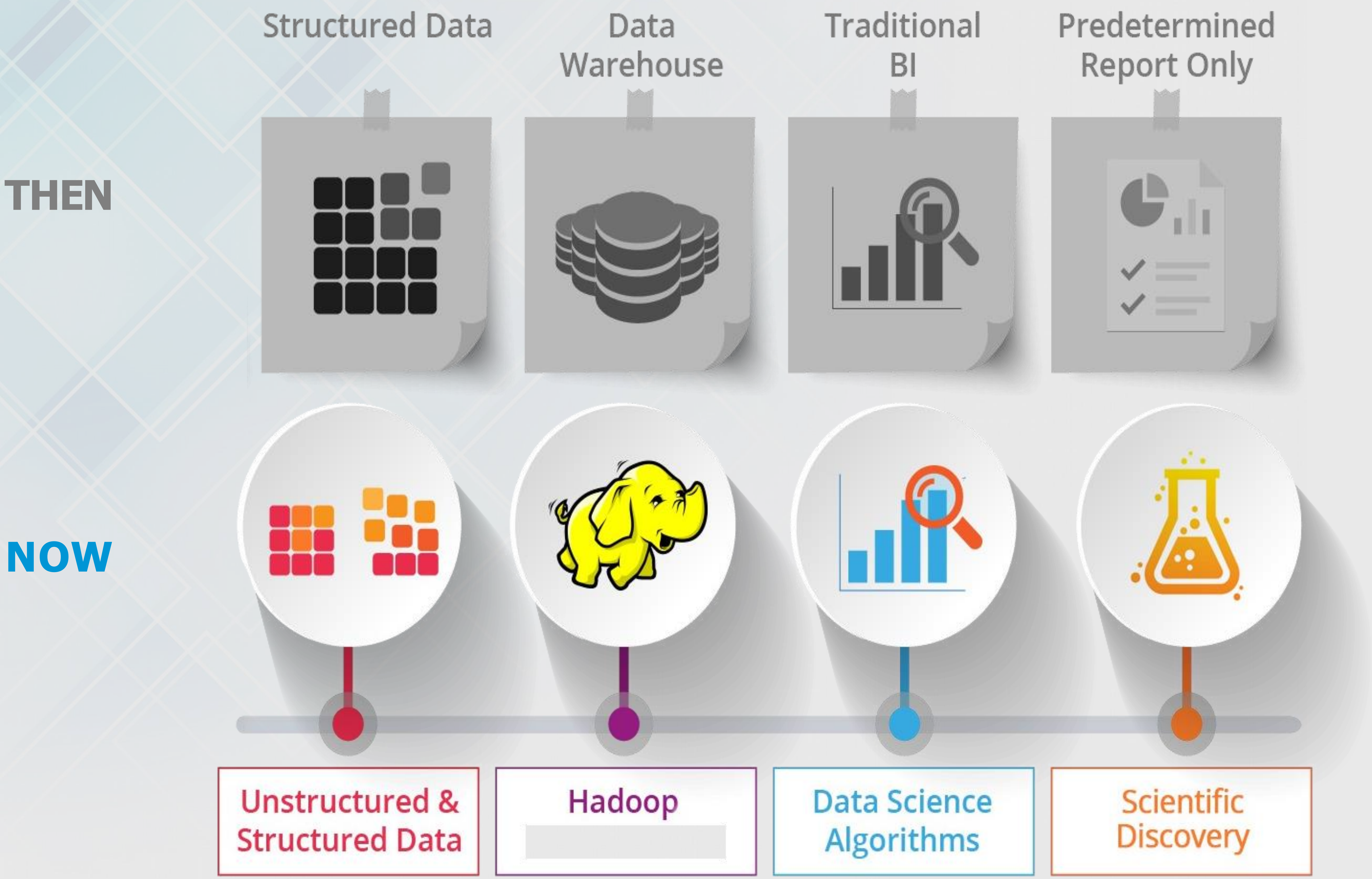
Decision making

Prediction

Pattern discovery

Data Science

Need Of Data Science



THEN

Structured Data

Data Warehouse

Traditional BI

Predetermined Report Only

NOW

Unstructured & Structured Data

Hadoop

Data Science Algorithms

Scientific Discovery

Need Of Data Science

You can use Data Science to

- ➢ Recommend the right product to the 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 transactions beforehand.
- ➢ Perform sentiment analysis to predict the outcome of elections.

What Is Data Science

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.

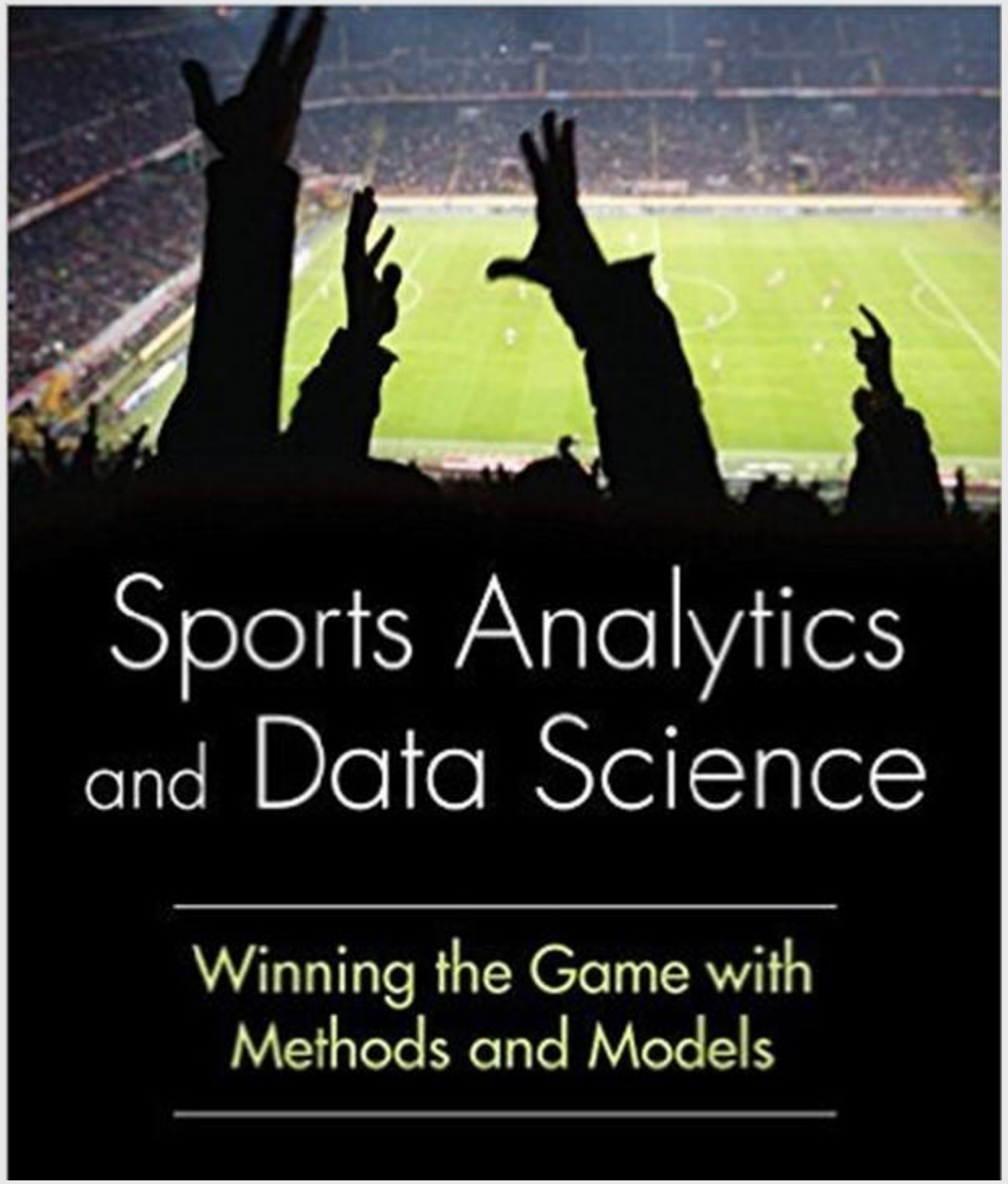
Now, lets understand Data Science with the help of some use cases.

What Is Data Science?

➢ 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.



Sports Analytics and Data Science

Winning the Game with Methods and Models

What Is Data Science?

- Amazon has huge amount of consumer purchasing data.

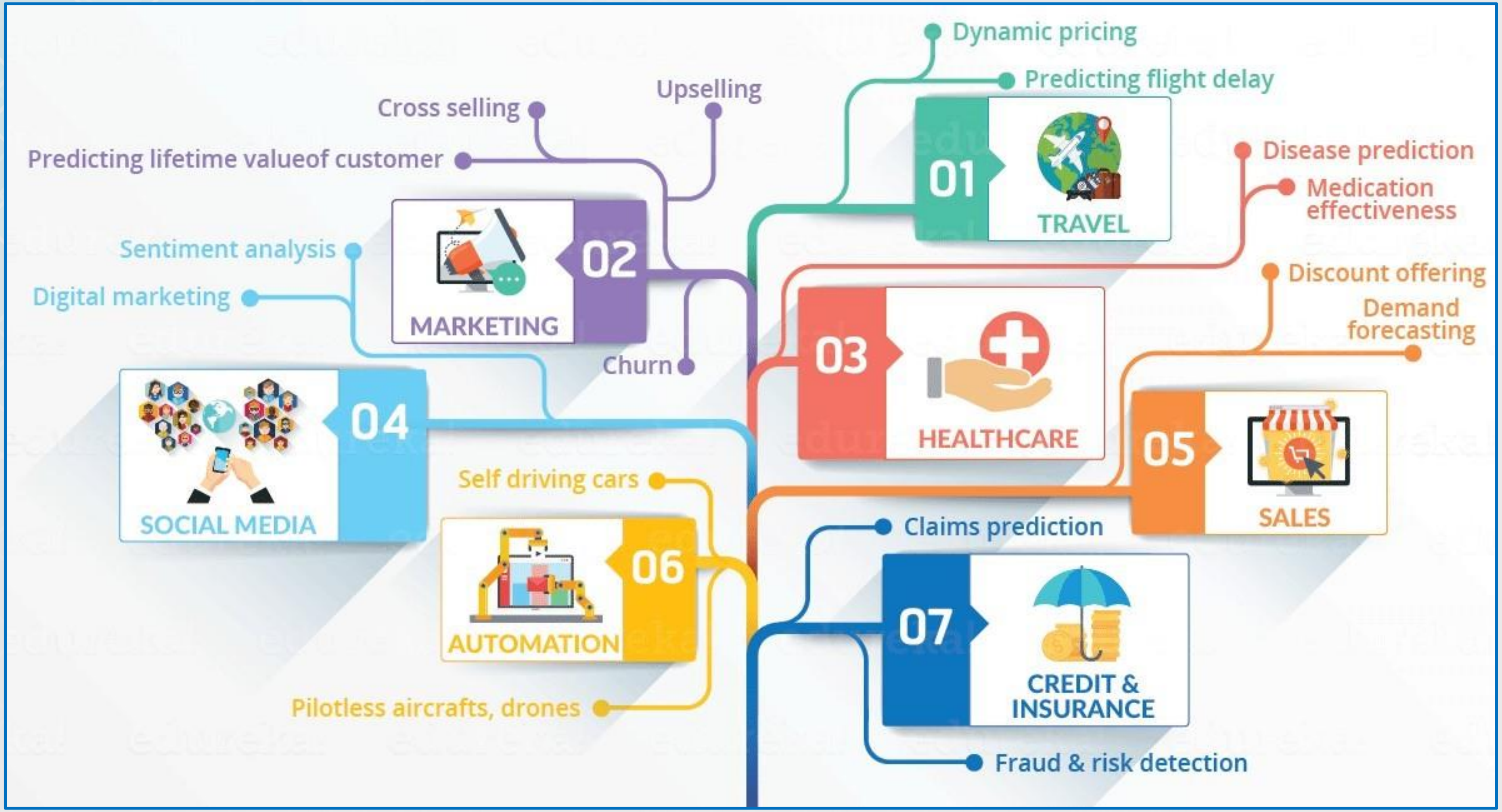
- The data consists of consumer demographics (age, sex, location), purchasing history, past browsing history.

- Based on this data, Amazon segments its customers, draws a pattern and recommends the right product to the right customer at the right time.

What Is Data Science?

- 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.

Use Cases Of Data Science



Being a Data Scientist

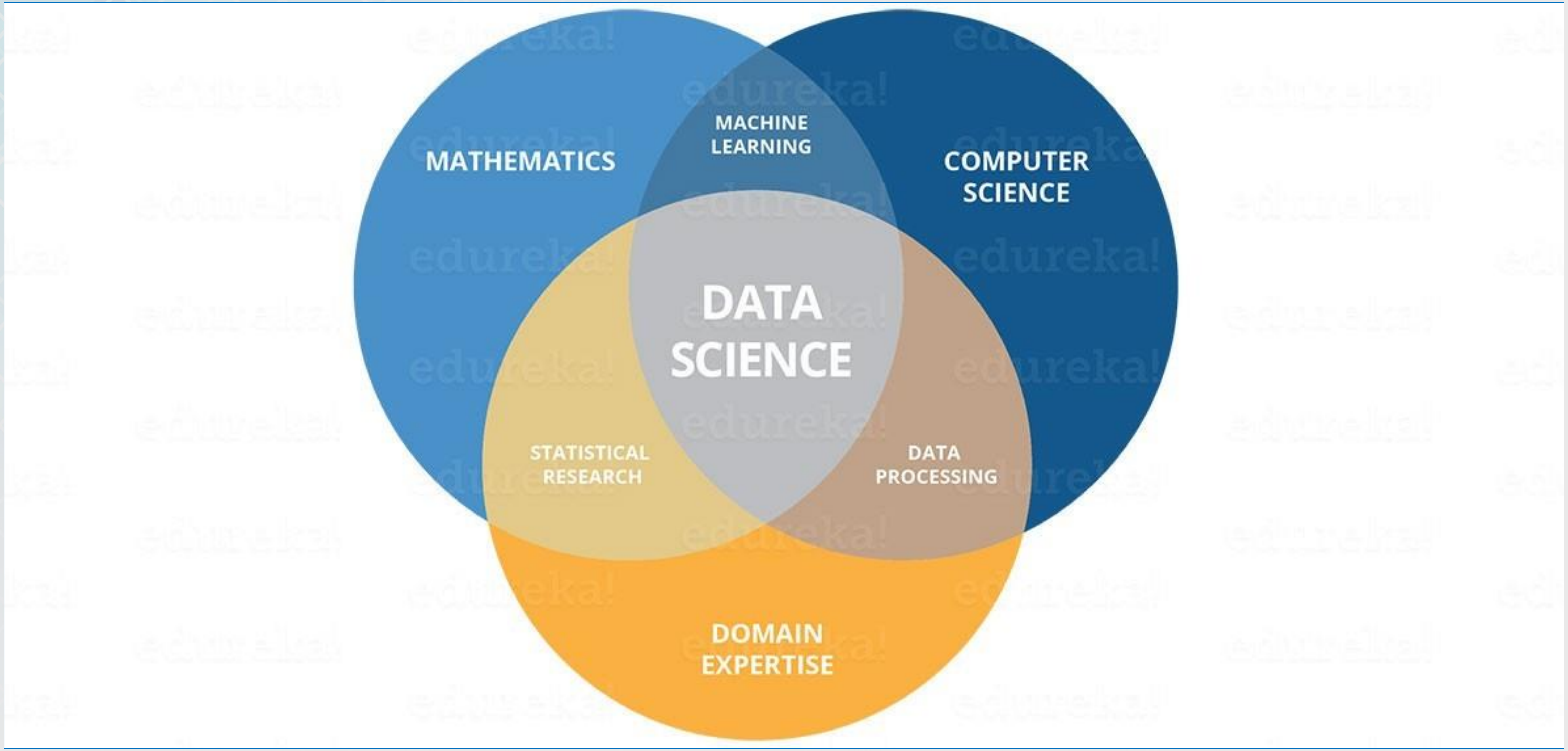
- « Data Scientist – the most sexy job of the 21st century »

  Thomas H. Davenport

- Data Scientist: A person who is better at statistics than any software engineer and better at software engineering than any statistician »»

  Josh Wills

Skills Of Data Scientist



Role Of A 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.

BI Vs. Data Science

Characteristics	Business Intelligence	Data Science
Perspective	Looking Backward	Looking Forward
Data Sources	Structured (Usually SQL, often Data Warehouse)	Both Structured and Unstructured (logs, cloud data, SQL, NoSQL, text)
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

Tools Used In Data Science

Commonly used tools by Data Scientists

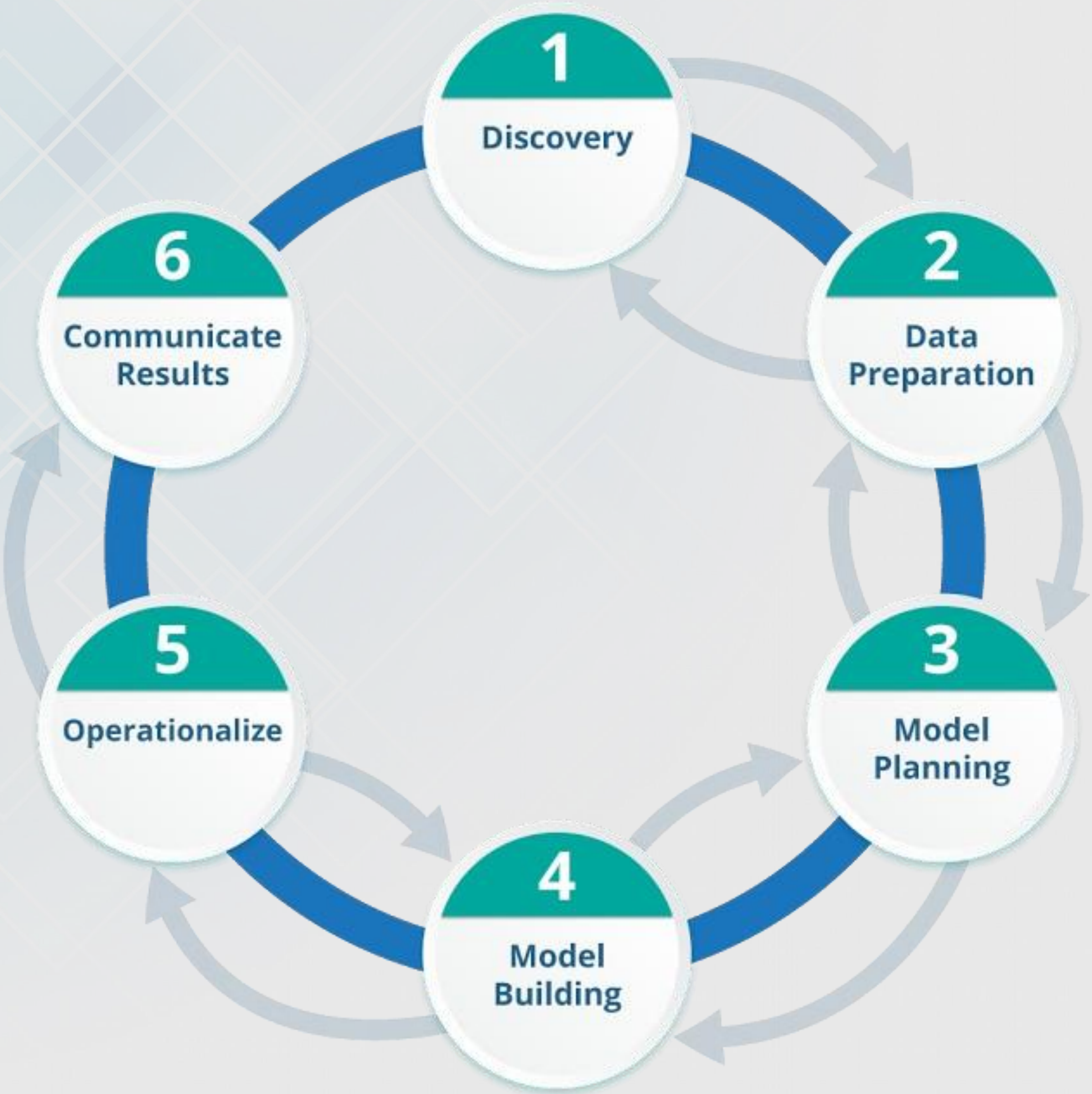
Data analysis	Data warehousing	Data visualization	Machine learning
• R	• Hadoop	• R	• Spark
• Spark	• SQL	• Tableau	• Mahout
• Python	• Hive	• Raw	• Azure ML studio
• SAS			

Lifecycle Of Data Science

What if we could predict the occurrence of diabetes and take appropriate measures beforehand to prevent it?

Definitely! Let me take you through the steps to predict the vulnerable patients.

Lifecycle Of Data Science



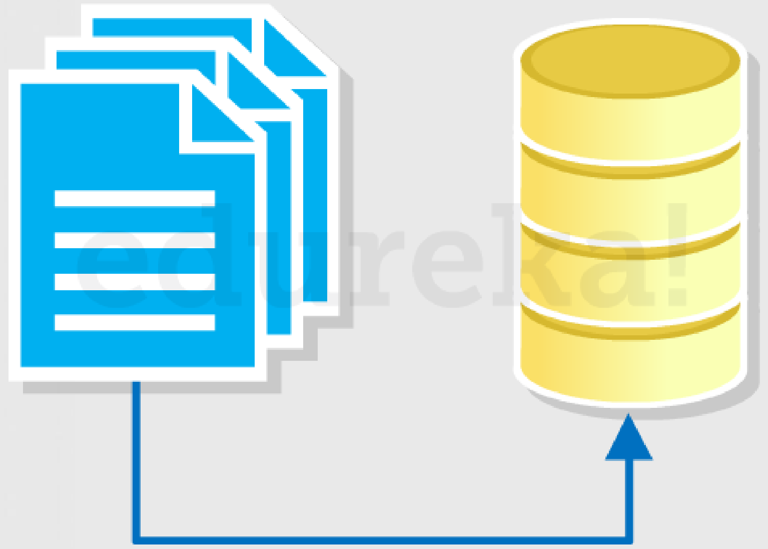
Lifecycle Of Data Science

- ● **Discovery**
- ○ **Data Preparation**
- ○ **Model Planning**
- ○ **Model Building**
- ○ **Operationalize**
- ○ **Communicate Results**

➢ Discovery involves acquiring data from all the identified internal and external sources that can help answer the business question.

➢ This data could be

- logs from webservers
- social media data
- census datasets
- data streamed from online sources via APIs



Lifecycle Of Data Science

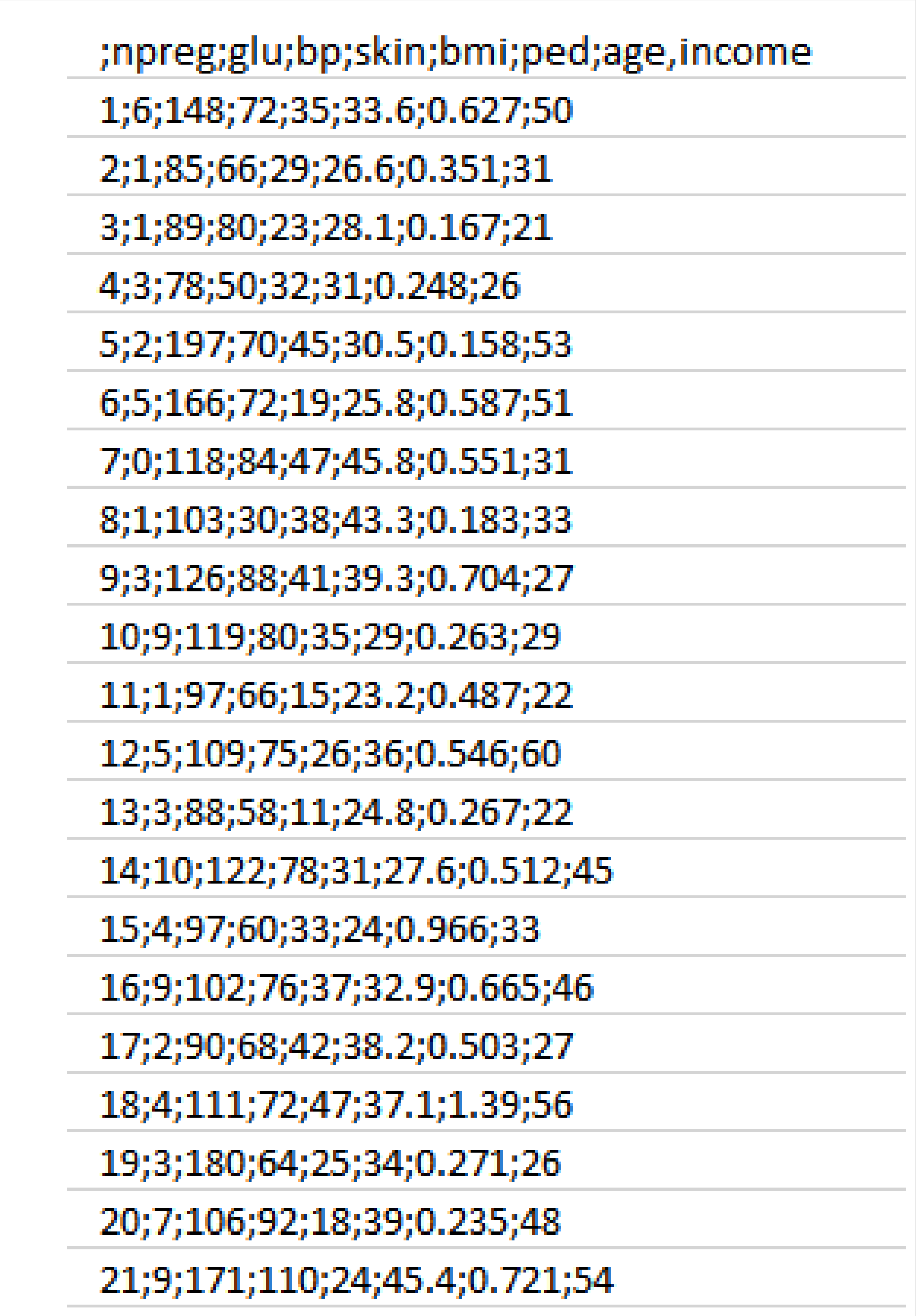
- ● **Discovery**
- ○ Data Preparation
- ○ Model Planning
- ○ Model Building
- ○ Operationalize
- ○ Communicate Results

Doctor gets this data from the medical history of the patient.

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

Income is an irrelevant attribute in the prediction of diabetes



```
;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
```

Lifecycle Of Data Science

○ Discovery

● **Data Preparation**

○ Model Planning

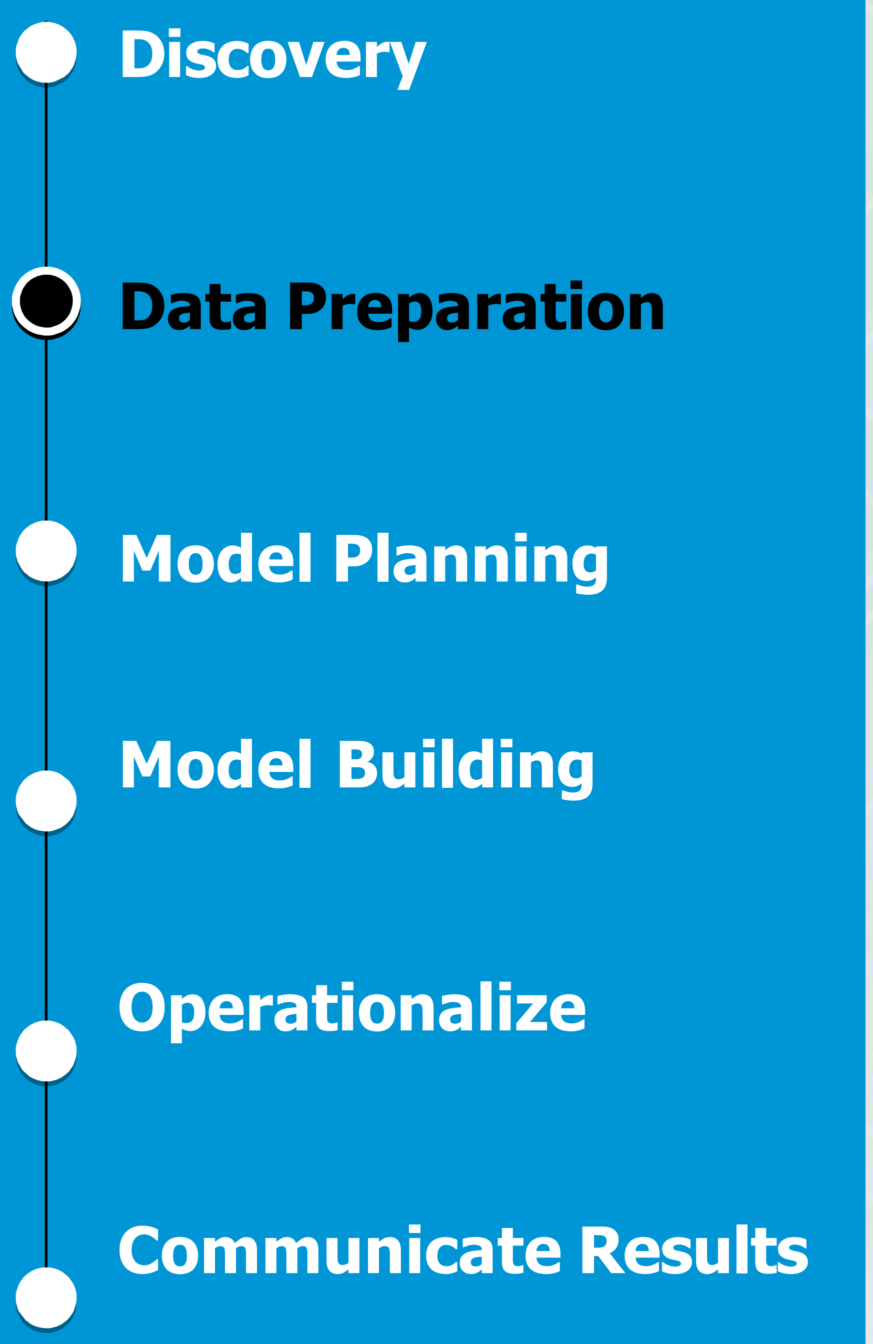
○ Model Building

○ Operationalize

○ Communicate Results

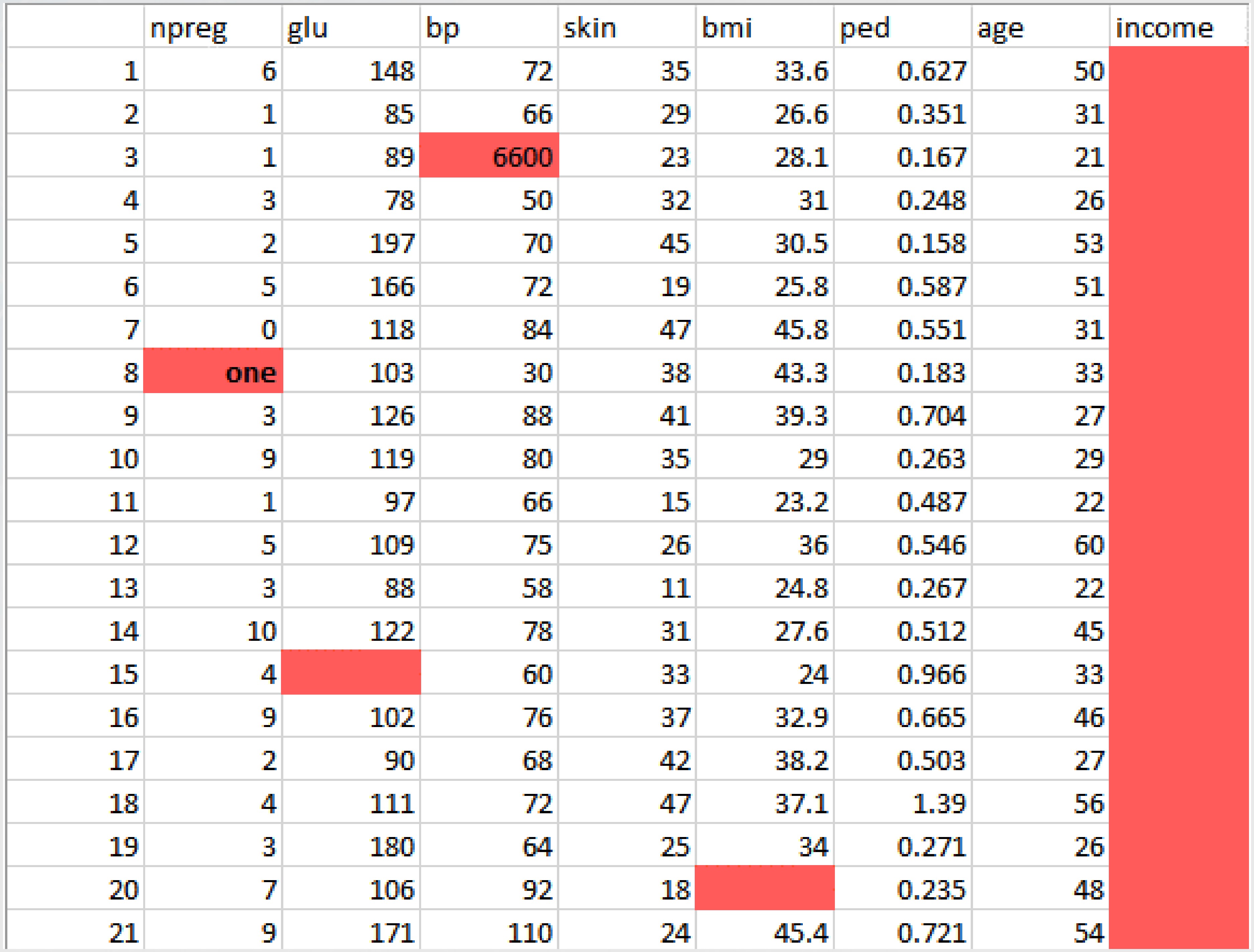
- ➢ The data can have a lot of inconsistencies like missing values, blank columns, abrupt values and incorrect data format which need to be cleaned.
- ➢ It is required to explore, preprocess and condition data prior to modeling.
- ➢ This will help you to spot the outliers and establish a relationship between the variables.

Lifecycle Of Data Science



- Discovery
- **Data Preparation**
- Model Planning
- Model Building
- Operationalize
- Communicate Results

This data has lot of anomalies and needs cleansing before further analysis can be done.

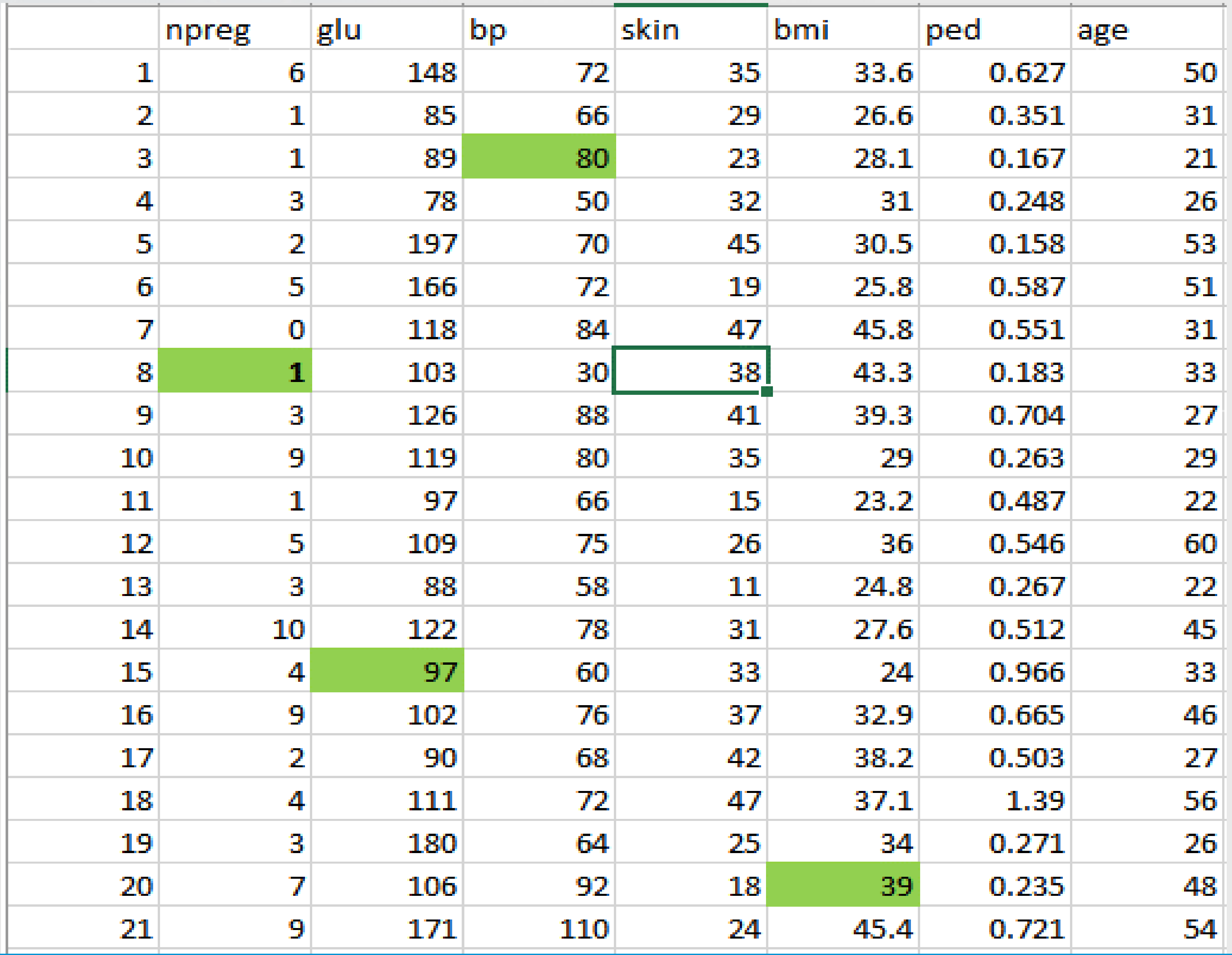


	npreg	glu	bp	skin	bmi	ped	age	income
1	6	148	72	35	33.6	0.627	50	
2	1	85	66	29	26.6	0.351	31	
3	1	89	6600	23	28.1	0.167	21	
4	3	78	50	32	31	0.248	26	
5	2	197	70	45	30.5	0.158	53	
6	5	166	72	19	25.8	0.587	51	
7	0	118	84	47	45.8	0.551	31	
8	one	103	30	38	43.3	0.183	33	
9	3	126	88	41	39.3	0.704	27	
10	9	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	

Lifecycle Of Data Science

- ○ Discovery
- ● **Data Preparation**
- ○ Model Planning
- ○ Model Building
- ○ Operationalize
- ○ Communicate Results

We clean and preprocess this data by removing the outliers, filling up the null values and normalizing the data type.



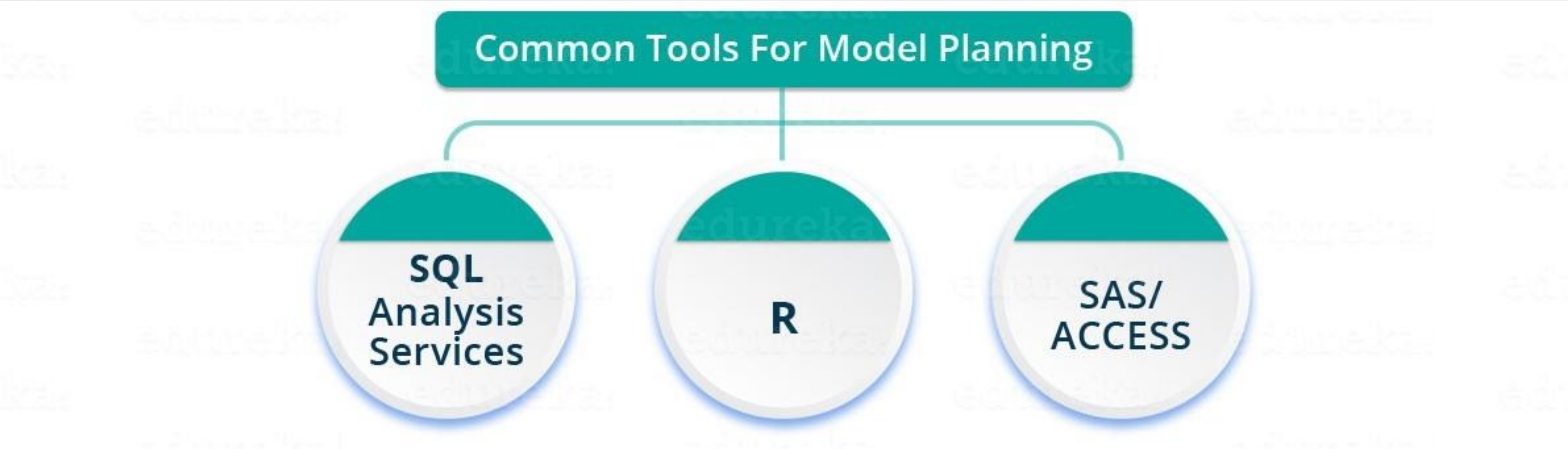
	npreg	glu	bp	skin	bmi	ped	age
1	6	148	72	35	33.6	0.627	50
2	1	85	66	29	26.6	0.351	31
3	1	89	80	23	28.1	0.167	21
4	3	78	50	32	31	0.248	26
5	2	197	70	45	30.5	0.158	53
6	5	166	72	19	25.8	0.587	51
7	0	118	84	47	45.8	0.551	31
8	1	103	30	38	43.3	0.183	33
9	3	126	88	41	39.3	0.704	27
10	9	119	80	35	29	0.263	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

Lifecycle Of Data Science

- Discovery
- Data Preparation
- Model Planning
- Model Building
- Operationalize
- Communicate Results

➢ Here, we determine the methods and techniques to draw the relationships between variable.

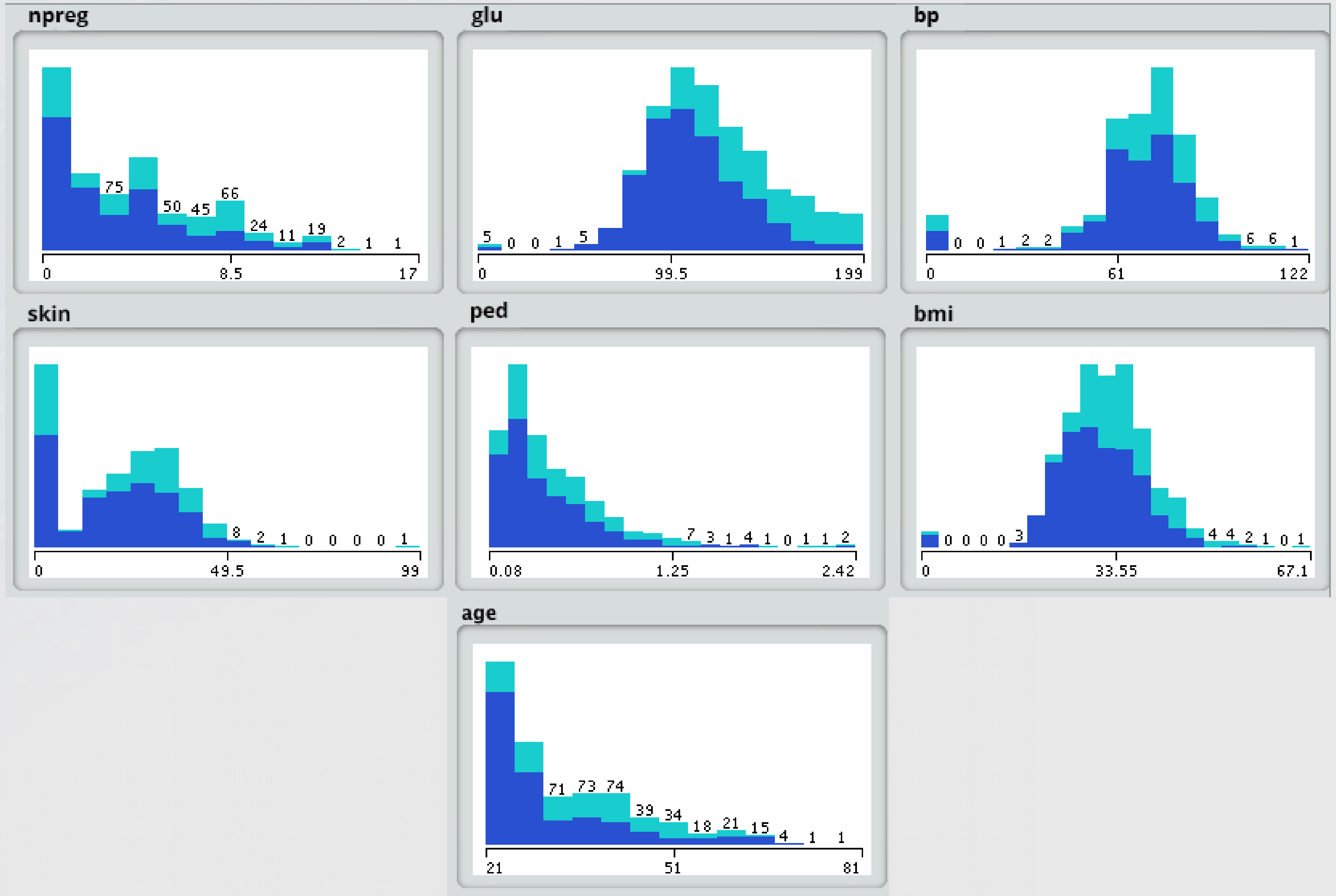
➢ Apply Exploratory Data Analytics (EDA) using various statistical formulas and visualization tools.



Lifecycle Of Data Science

- Discovery
- Data Preparation
- **Model Planning**
- Model Building
- Operationalize
- Communicate Results

Use of visualization techniques like histograms, line graphs, box plots to get a fair idea of the distribution of data.



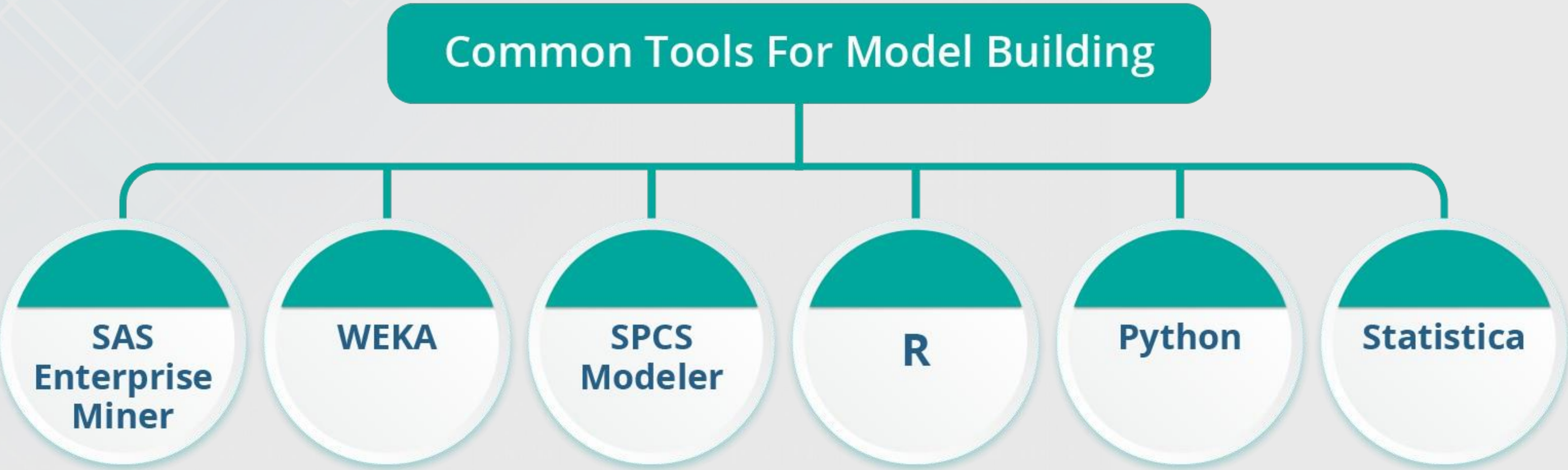
Lifecycle Of Data Science

- Discovery
- Data Preparation
- Model Planning
- **Model Building**
- Operationalize
- Communicate Results

➢ Develop datasets for training and testing purposes.

➢ Consider whether existing tools will suffice for running the models.

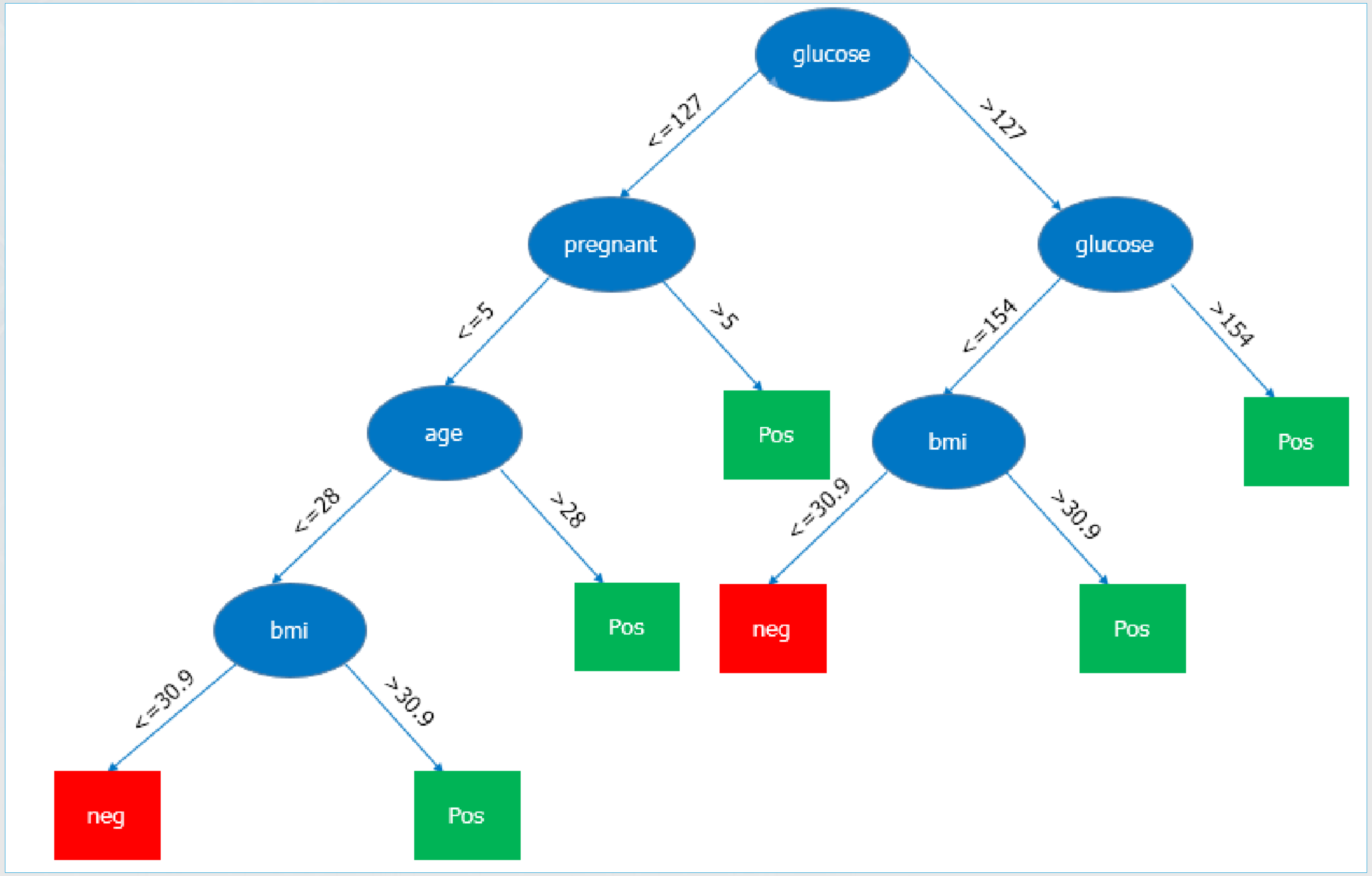
➢ Analyze various learning techniques like classification, association and clustering to build the model.



Lifecycle Of Data Science

- Discovery
- Data Preparation
- Model Planning
- **Model Building**
- Operationalize
- Communicate Results

This is a decision tree based on different attributes.



Lifecycle Of Data Science

○ **Discovery**

○ **Data Preparation**

○ **Model Planning**

○ **Model Building**

● **Operationalize**

○ **Communicate Results**

- ➢Deliver final reports, briefings, code and technical documents.
- ➢Implement pilot project in a real-time production environment.
- ➢Look for performance constraints if any.

Lifecycle Of Data Science

- Discovery
- Initialization
- Model Planning
- Model Building
- Deployment
- **Communicate Results**

➢ Identify all the key findings and communicate to the stakeholders.

➢ Explaining the model and result to medical authorities.

➢ Determine if the results of the project are a success or a failure based on the criteria developed.

Lifecycle Of Data Science

- Discovery
- Initialization
- Model Planning
- Model Building
- Deployment
- ● Communicate Results

➢ **Diabetes Positive set:**

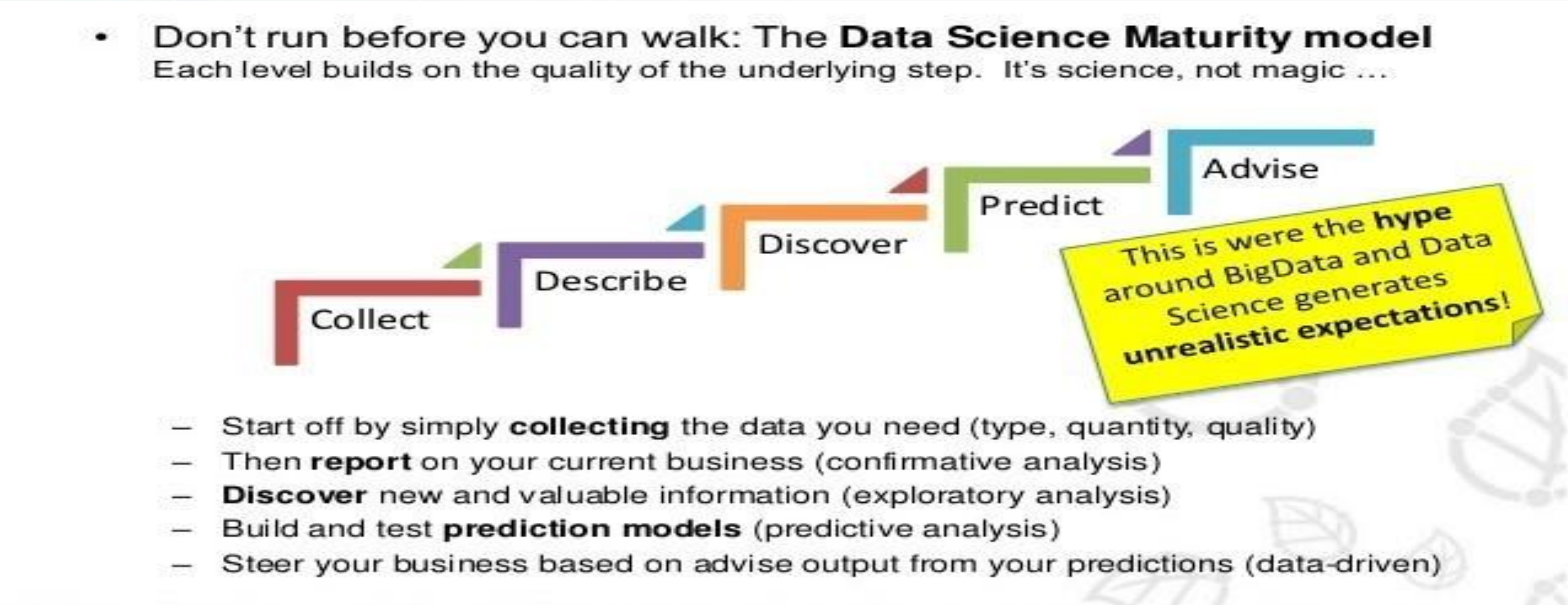
- glucose > 154
- glucose >127 & <= 154 + bmi >30.9
- glucose<=127 + pregnant >5
- glucose<=127 + pregnant <=5 + age >28
- glucose<=127 + pregnant <=5 + age <=28 +bmi > 30.9

➢ **Diabetes Negative set:**

- glucose > 154
- glucose >127 & <= 154 + bmi <=30.9
- glucose<=127 + pregnant <=5 + age <=28 +bmi <= 30.9

➢ We can use this decision tree result to know whether the patient is vulnerable to diabetes or not.

The Data Science maturity model



- Don’t run before you can walk: The **Data Science Maturity model**
  Each level builds on the quality of the underlying step. It's science, not magic …

This is were the **hype** around BigData and Data Science generates **unrealistic expectations**!

- – Start off by simply **collecting** the data you need (type, quantity, quality)
- – Then **report** on your current business (confirmative analysis)
- – **Discover** new and valuable information (exploratory analysis)
- – Build and test **prediction models** (predictive analysis)
- – Steer your business based on advise output from your predictions (data-driven)

Data Science

Thank You

All the Best…!