INDIAN SCHOOL OF BUSINESS
ISB
BHARTI INSTITUTE OF PUBLIC POLICY
MAX INSTITUTE OF HEALTHCARE MANAGEMENT
MUNJAL GLOBAL MANUFACTURING INSTITUTE
PUNJ LLOYD INSTITUTE OF
PHYSICAL INFRASTRUCTURE MANAGEMENT

Applied Business Analytics

Week 2 – Data Preliminaries for Analytics

ISB | Executive Education

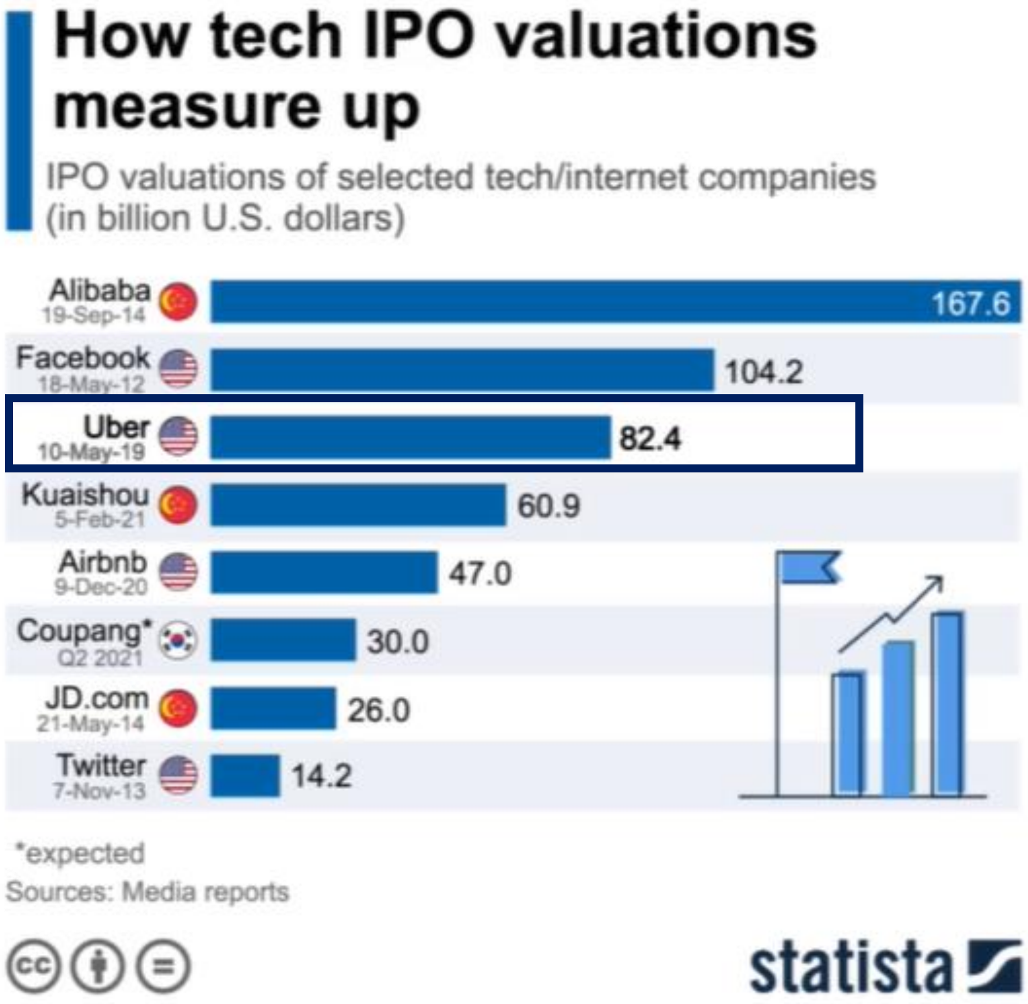
Data Preliminaries: A Motivating Example (Uber)

- March 2009 Uber was founded
- By 2014: Uber’s valuation was $40 billion
- Why is Uber valued so highly?

Valuation:

The Net Present Value (NPV) of a firm’s lifetime profits

Uber’s Current Value



How tech IPO valuations measure up

IPO valuations of selected tech/internet companies (in billion U.S. dollars)

*expected

Sources: Media reports

Source:
https://www.statista.com/chart/17965/uber-ipo-and-selected-tech-ipos-company-values/

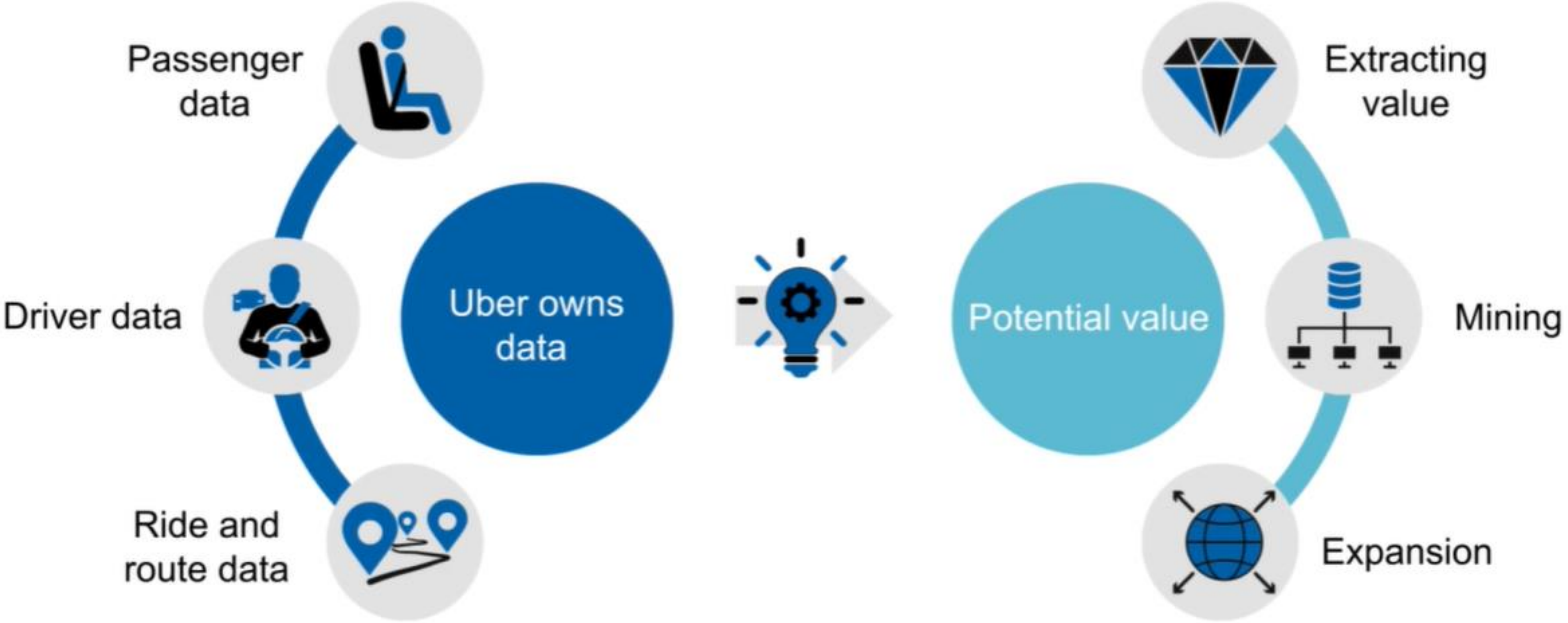
What Makes Uber So Special?

Competitive advantage

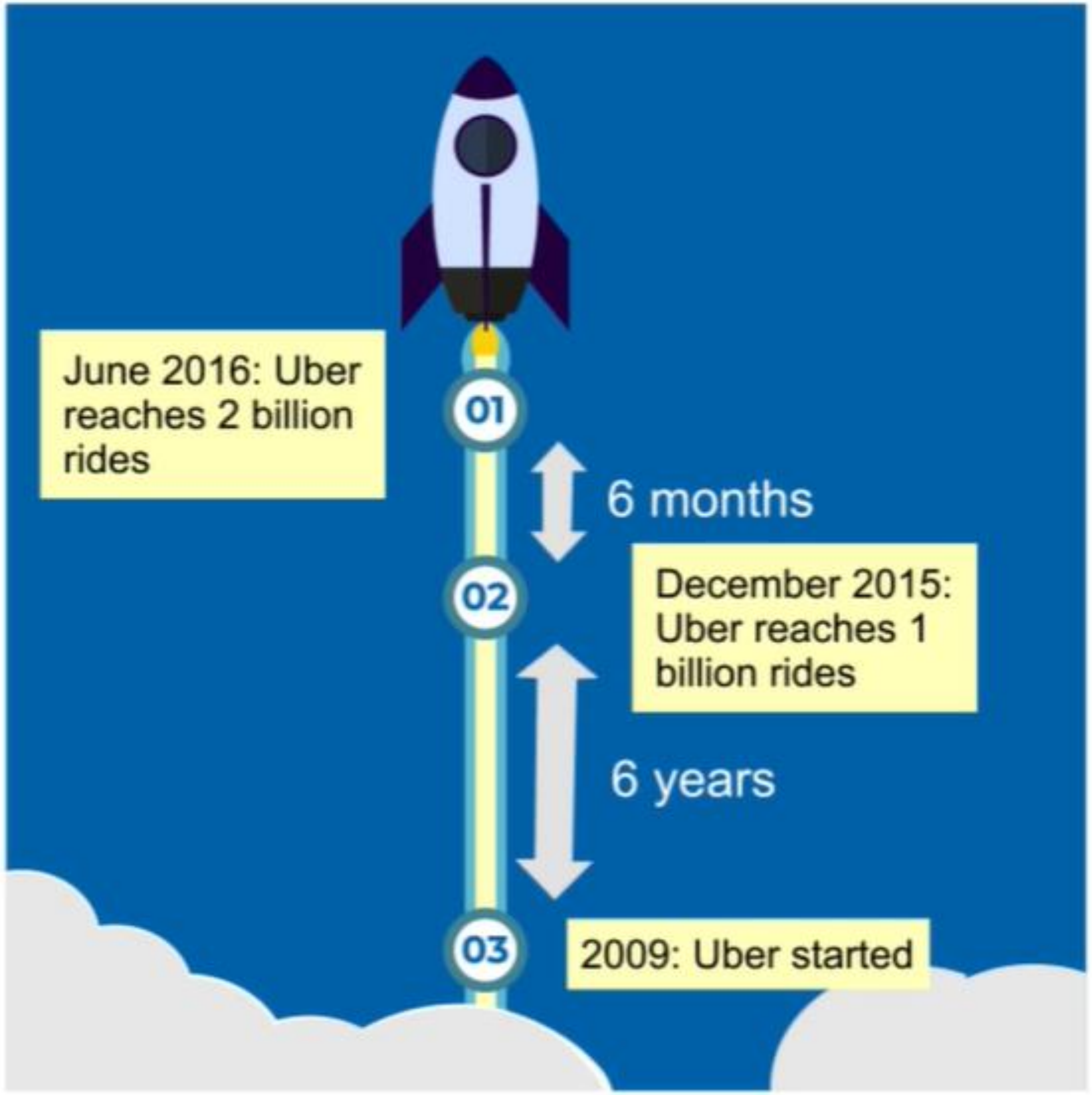
Strategic asset

Enabling platform

What Does Uber Own?



What’s the Worth of Uber’s Data?



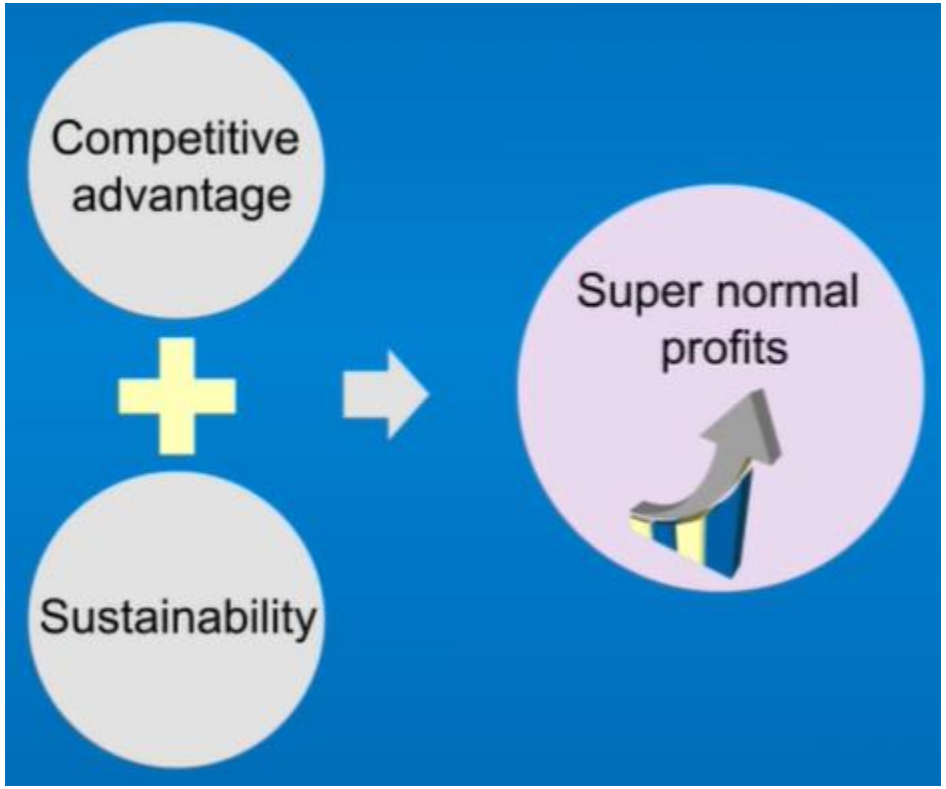
June 2016: Uber reaches 2 billion rides

December 2015: Uber reaches 1 billion rides

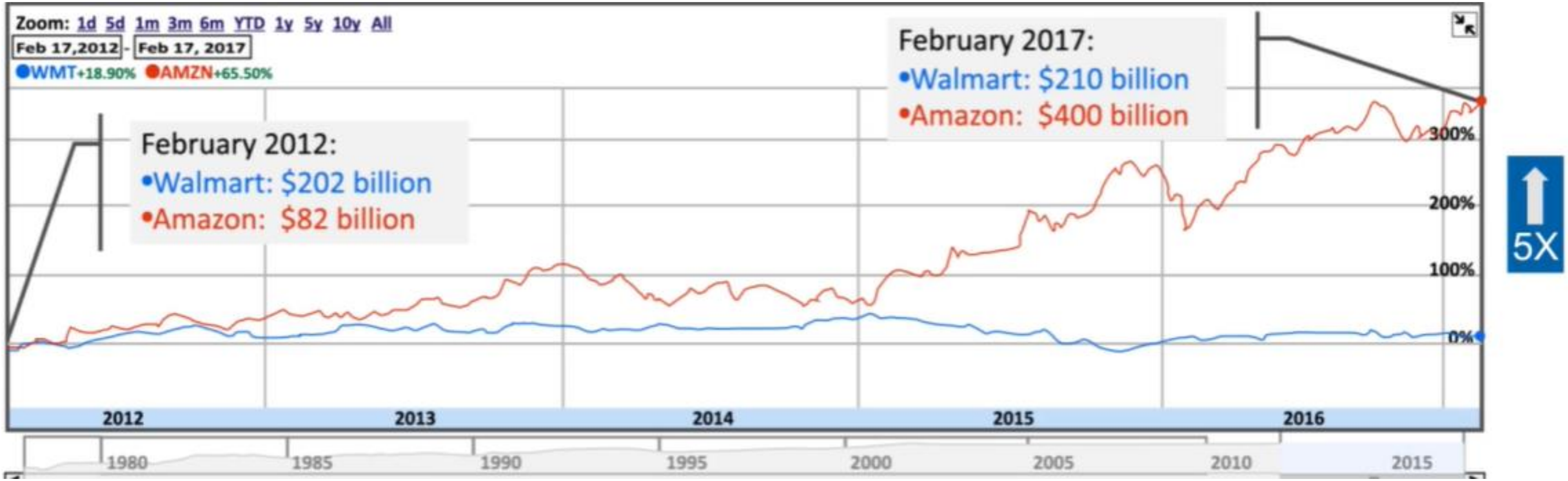
2009: Uber started

The Power of Data

Data is a valuable asset that might help build a sustainable competitive advantage



Stock Performance of Amazon vs Walmart



Zoom: 1d 5d 1m 3m 6m YTD 1y 5y 10y All

Feb 17,2012 - Feb 17, 2017

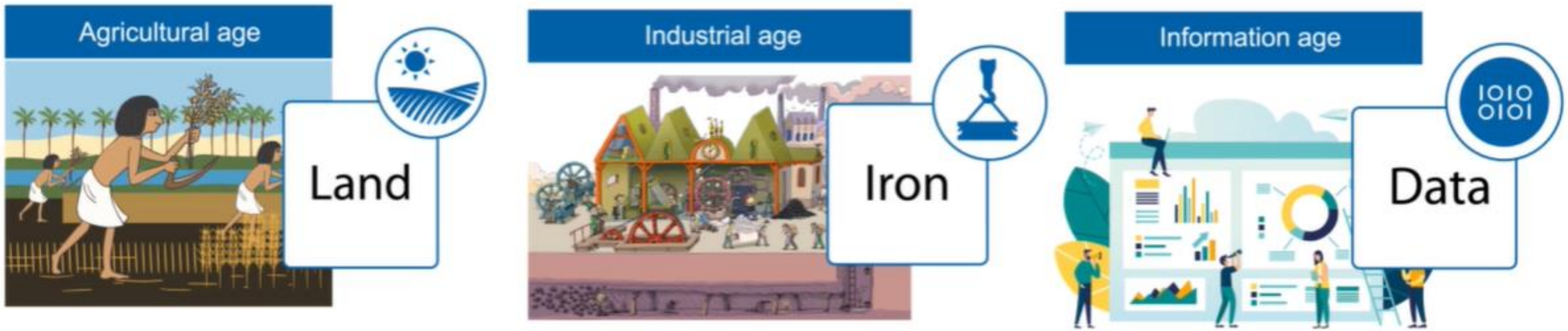
●WMT+18.90% ●AMZN+65.50%

February 2012:
- Walmart: $202 billion
- Amazon: $82 billion

February 2017:
- Walmart: $210 billion
- Amazon: $400 billion

Data Is Key in the Information Age

If land was the primary raw material of the agricultural age, iron and coal of the industrial age, then data is the primary raw material of today’s Information age.



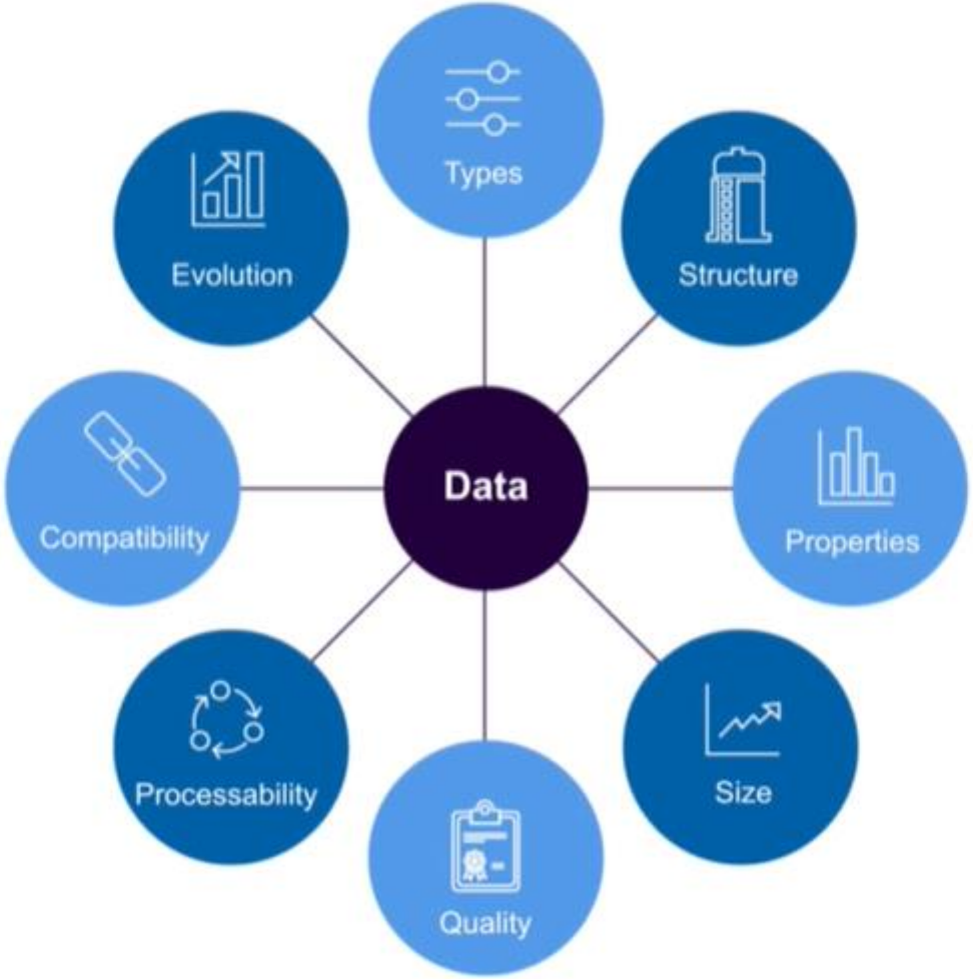
What is Data?

The word “data” originates from the Latin word “datum” meaning “given” (known or assumed as facts).

> “Data are characteristics or information, usually numerical, that are collected through observation. In a more technical sense, data are a set of values of qualitative or quantitative variables about one or more persons or objects, while a datum is a single value of a single variable.”
>
> \- Wikipedia

Key Questions about Data

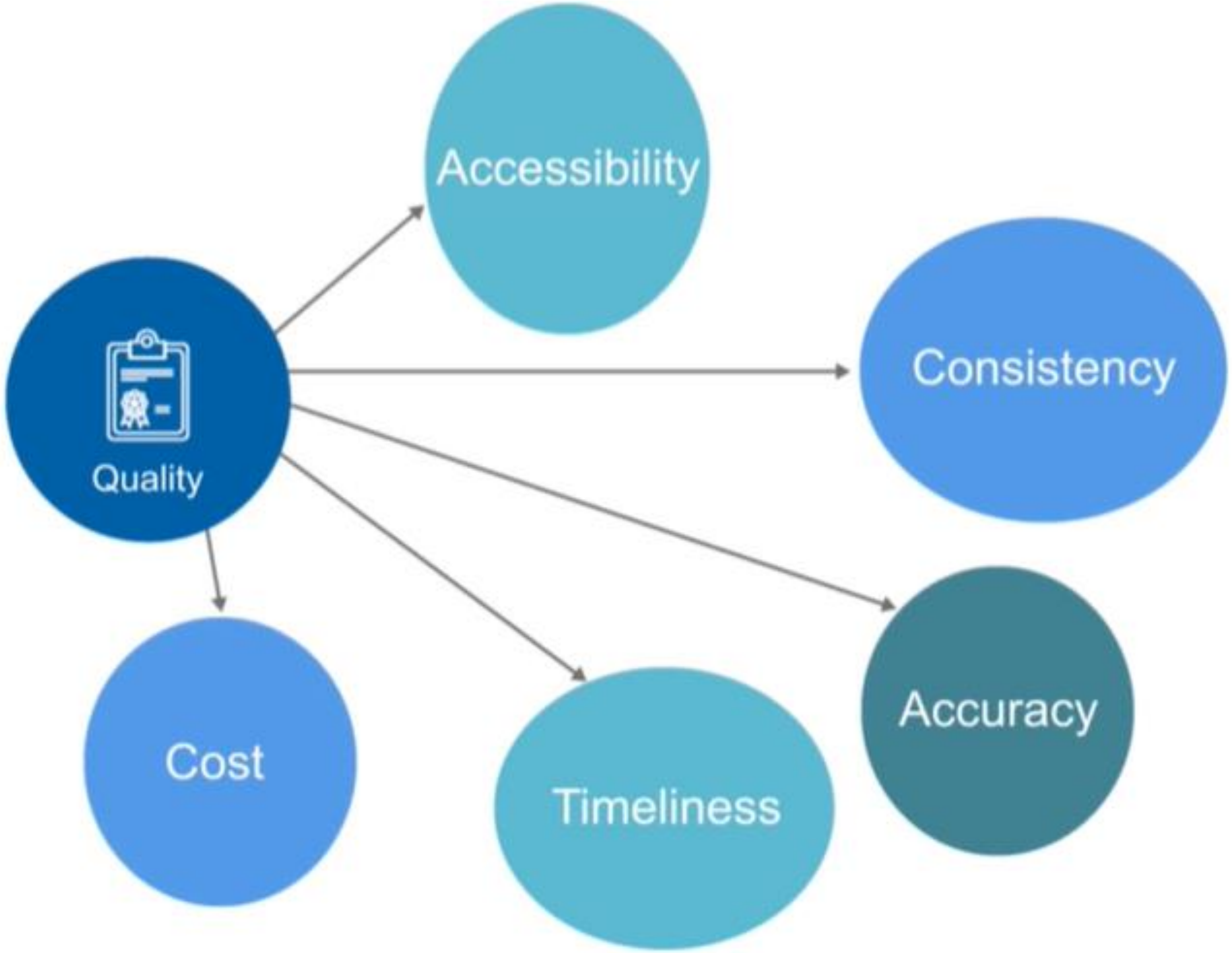
Different aspects of data



Each of these are multidimensional

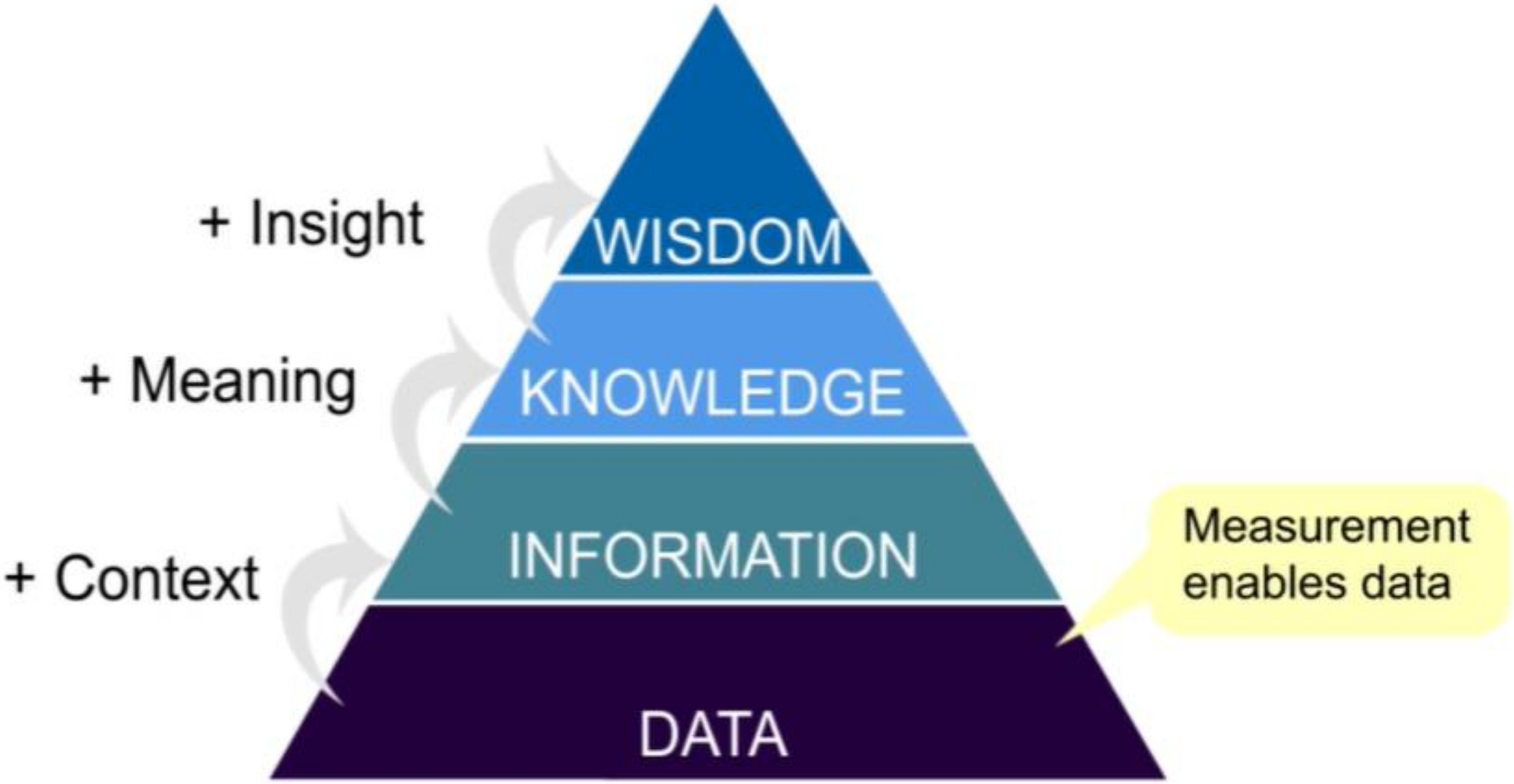
Key Questions about Data

Different aspects of data quality

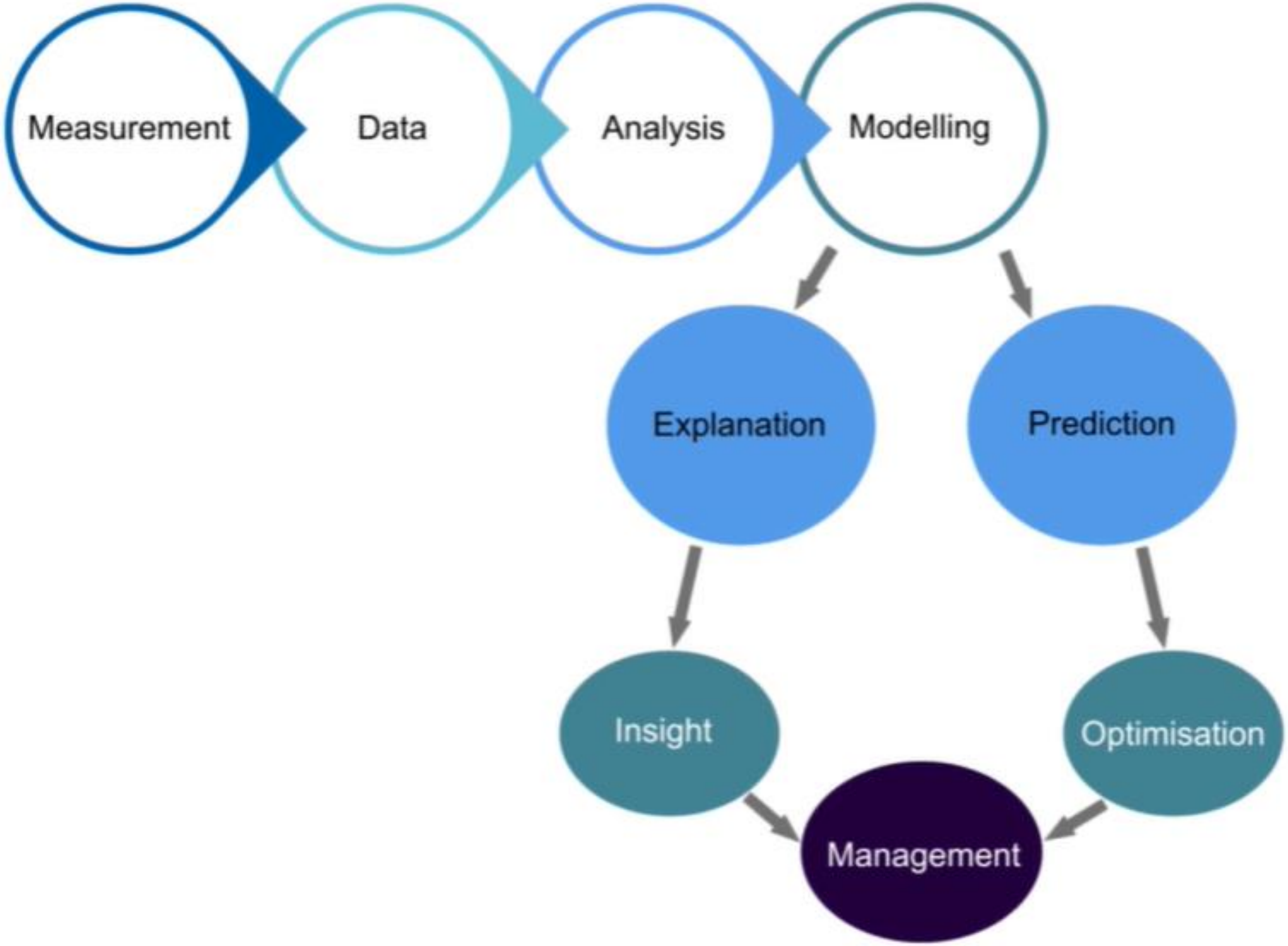


The Knowledge Hierarchy

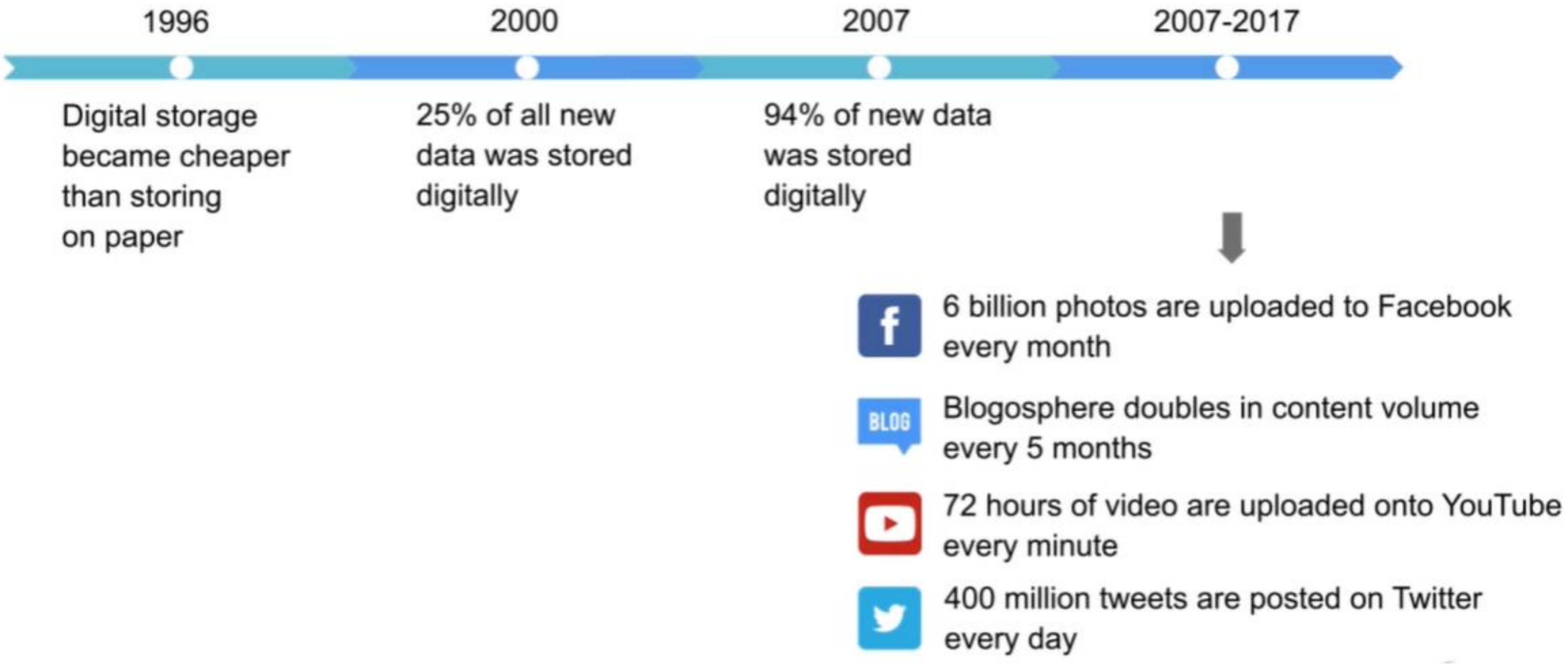
To be useful in some sense, data has to be transformed into certain higher-order entities.



Data and Measurement



Data Storage through the Years



1996

Digital storage became cheaper than storing on paper

2000

25% of all new data was stored digitally

2007

94% of new data was stored digitally

2007-2017

f 6 billion photos are uploaded to Facebook every month

BLOG Blogosphere doubles in content volume every 5 months

72 hours of video are uploaded onto YouTube every minute

400 million tweets are posted on Twitter every day

Data Storage: Conclusion

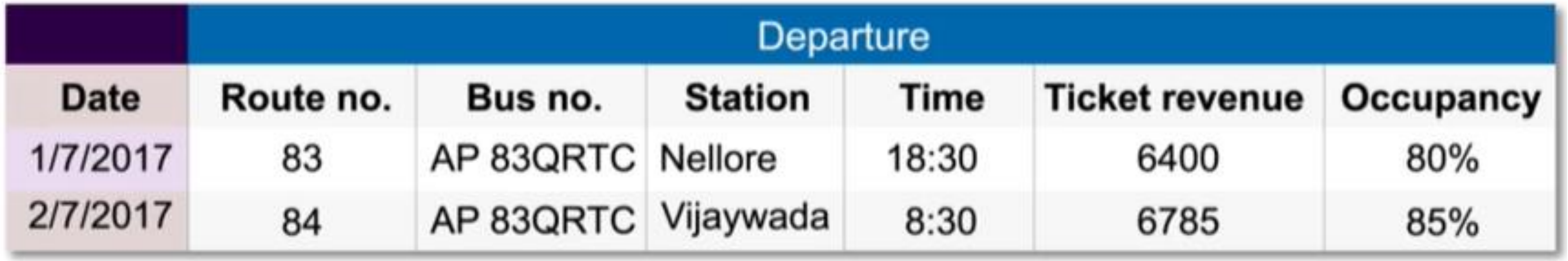
Lower digital storage costs have enabled large amounts of data to be generated and stored easily.

As a result:

- Evermore data is generated year-on-year
- Evermore of that data is native to digital means of storage, processing and transformation

Introduction to Data Dichotomies

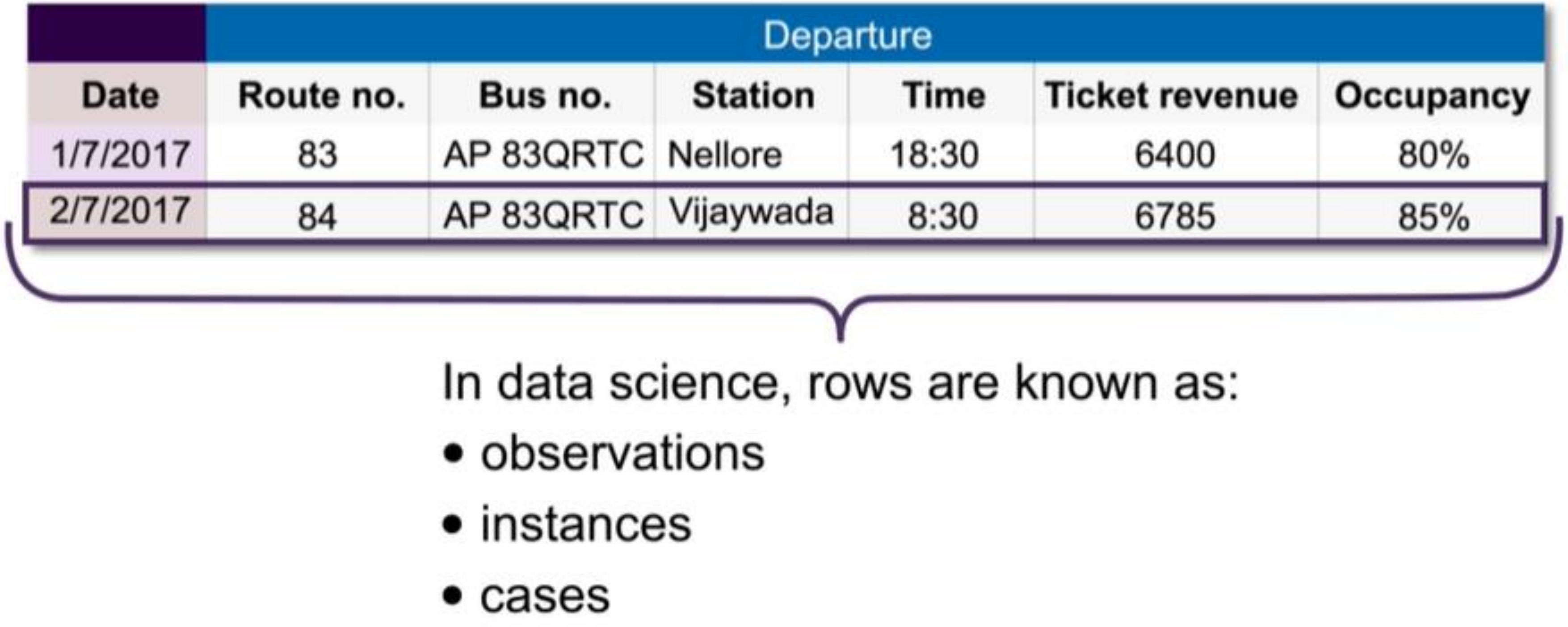
Data format: example



	Departure					
Date	Route no.	Bus no.	Station	Time	Ticket revenue	Occupancy
1/7/2017	83	AP 83QRTC	Nellore	18:30	6400	80%
2/7/2017	84	AP 83QRTC	Vijaywada	8:30	6785	85%

Data Format : Example

Data is structured when organised in rows and columns



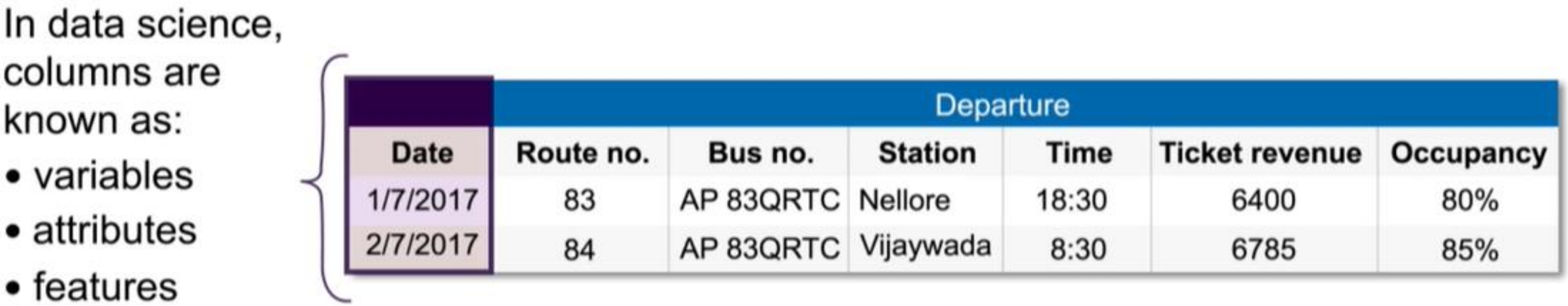
	Departure					
Date	Route no.	Bus no.	Station	Time	Ticket revenue	Occupancy
1/7/2017	83	AP 83QRTC	Nellore	18:30	6400	80%
2/7/2017	84	AP 83QRTC	Vijaywada	8:30	6785	85%

In data science, rows are known as:

- observations
- instances
- cases

Data Format : Example

Data is structured when organised in rows and columns

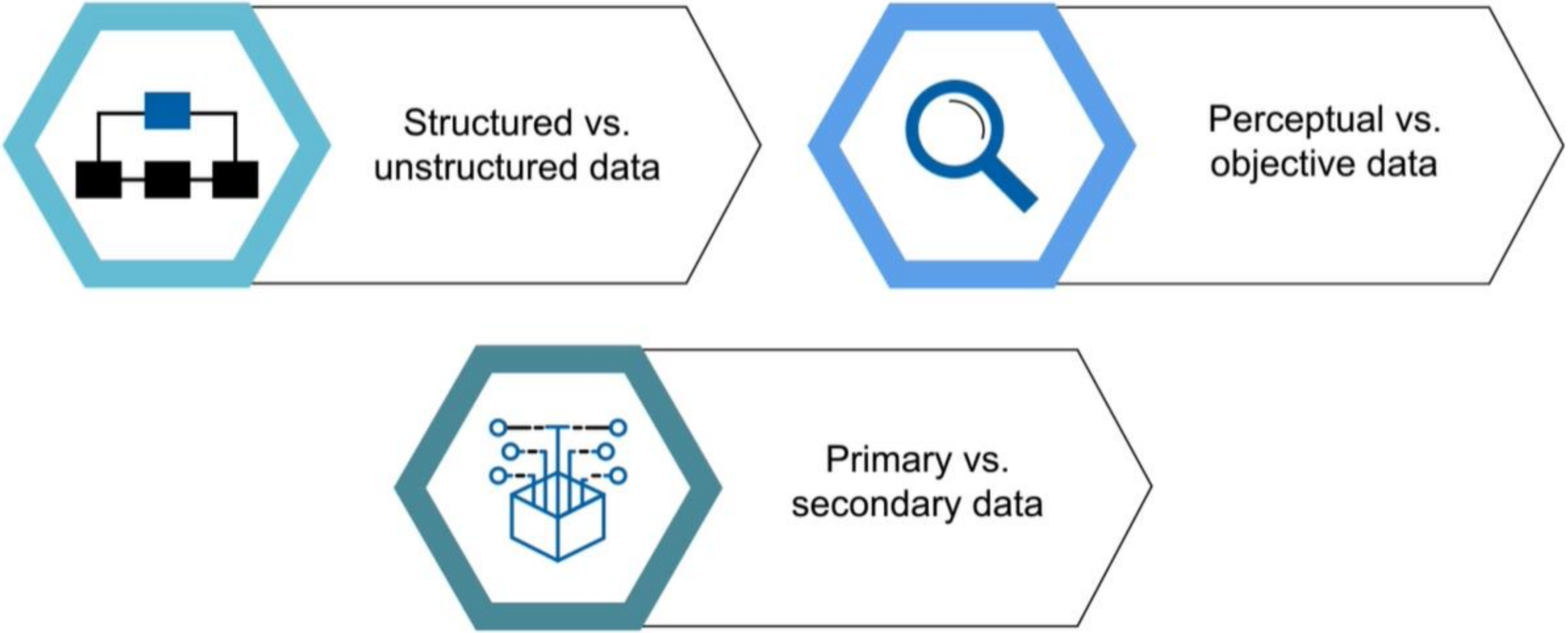


In data science, columns are known as:

- variables
- attributes
- features

	Departure					
Date	Route no.	Bus no.	Station	Time	Ticket revenue	Occupancy
1/7/2017	83	AP 83QRTC	Nellore	18:30	6400	80%
2/7/2017	84	AP 83QRTC	Vijaywada	8:30	6785	85%

Basic Data Dichotomies

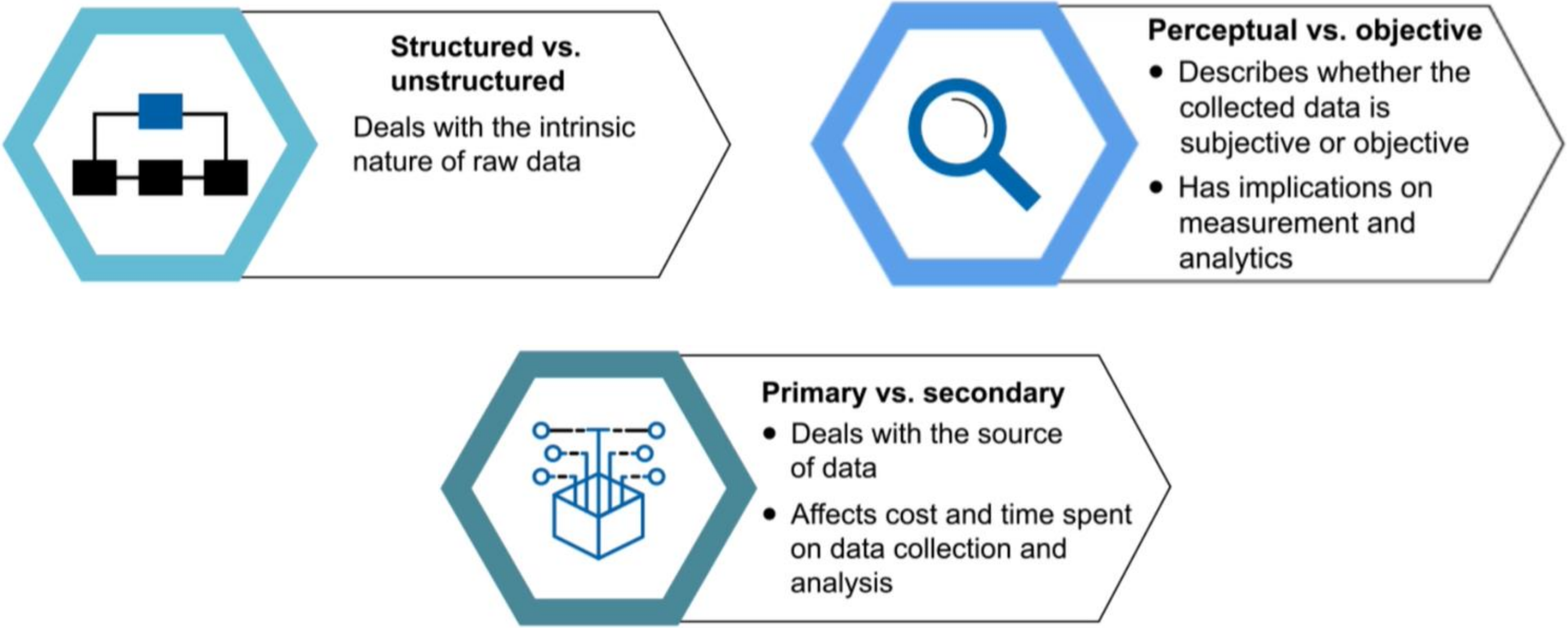


Structured vs. unstructured data

Perceptual vs. objective data

Primary vs. secondary data

Basic Data Dichotomies



Structured vs. unstructured

Deals with the intrinsic nature of raw data

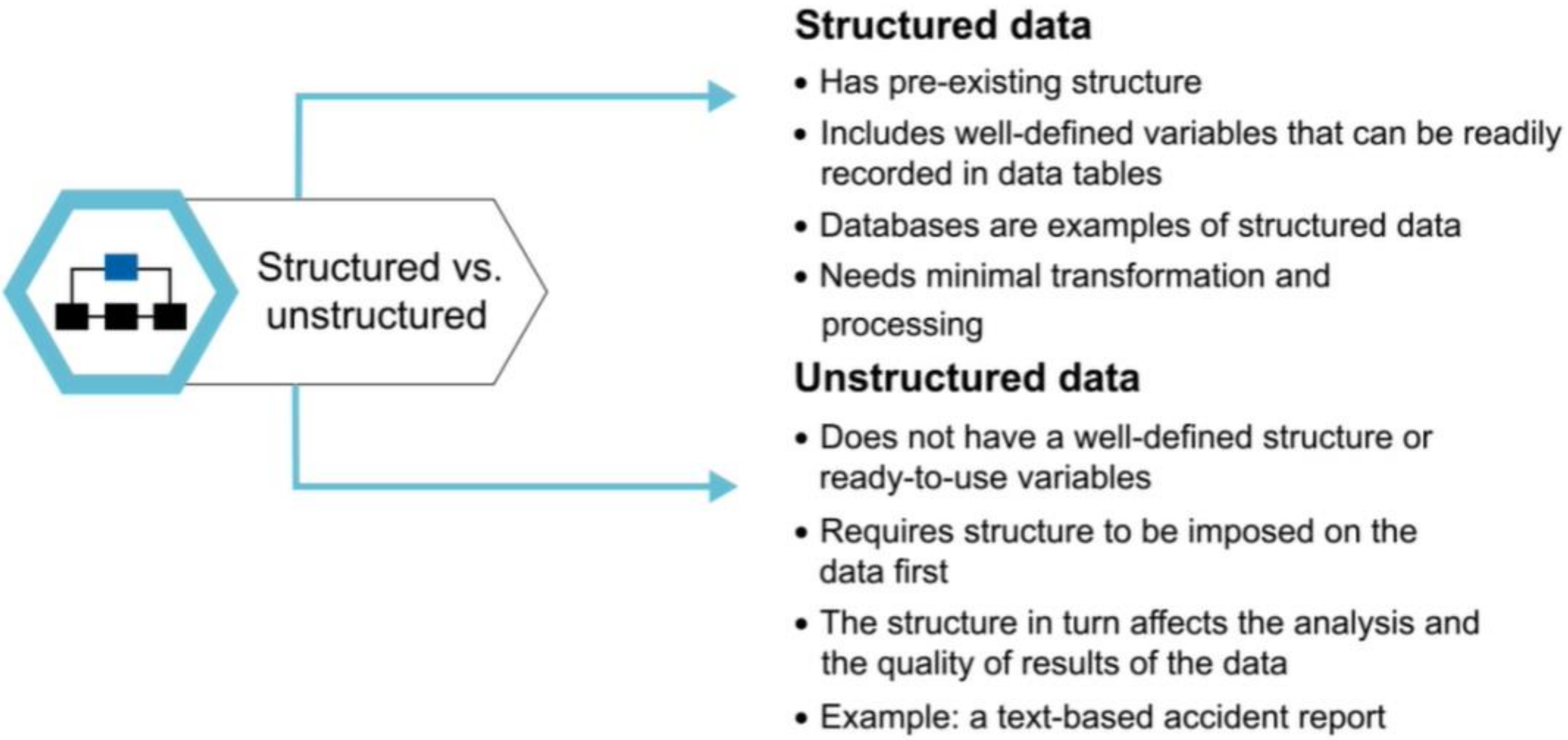
Perceptual vs. objective

- Describes whether the collected data is subjective or objective
- Has implications on measurement and analytics

Primary vs. secondary

- Deals with the source of data
- Affects cost and time spent on data collection and analysis

Structured and Unstructured Data



Structured data

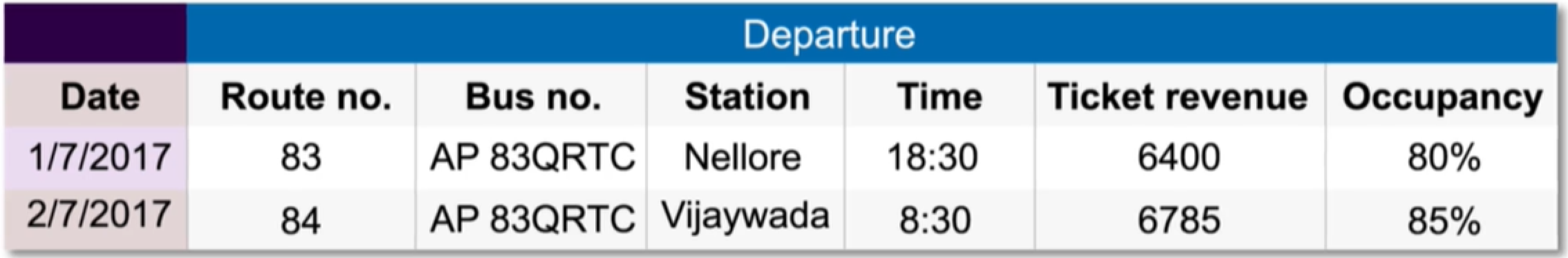
- Has pre-existing structure
- Includes well-defined variables that can be readily recorded in data tables
- Databases are examples of structured data
- Needs minimal transformation and processing

Unstructured data

- Does not have a well-defined structure or ready-to-use variables
- Requires structure to be imposed on the data first
- The structure in turn affects the analysis and the quality of results of the data
- Example: a text-based accident report

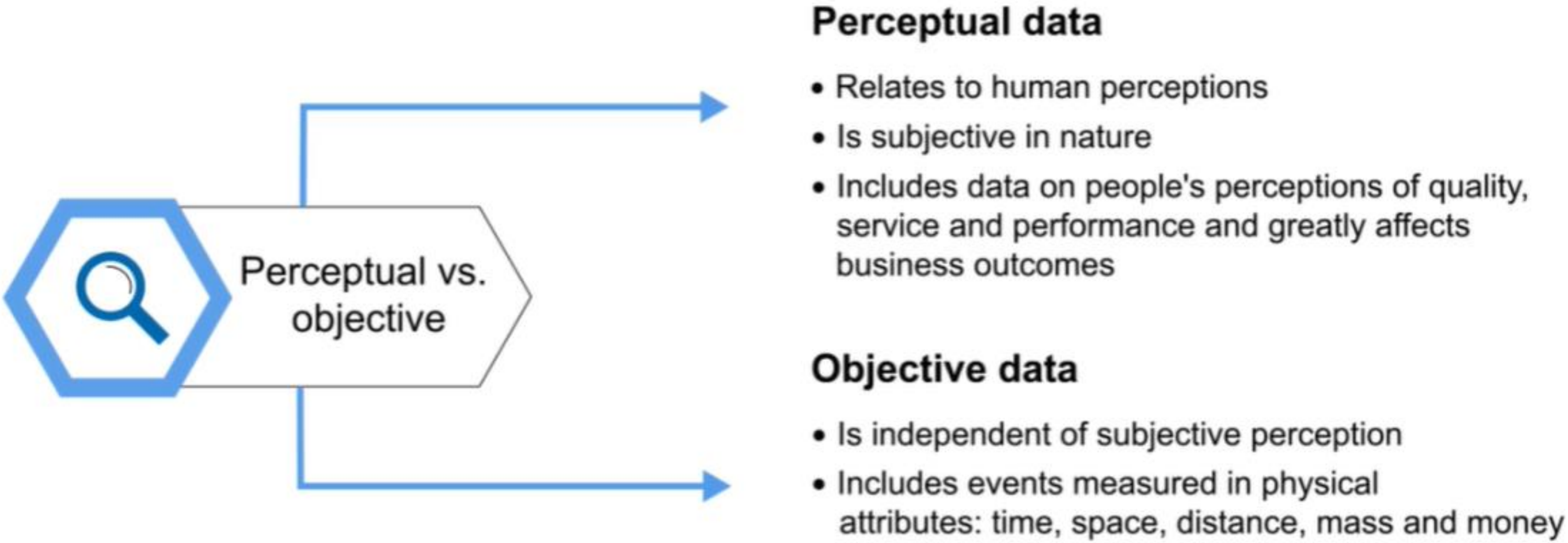
Structured and Unstructured Data

Example of structured data



	Departure					
Date	**Route no.**	**Bus no.**	**Station**	**Time**	**Ticket revenue**	**Occupancy**
1/7/2017	83	AP 83QRTC	Nellore	18:30	6400	80%
2/7/2017	84	AP 83QRTC	Vijaywada	8:30	6785	85%

Perceptual and Objective Data



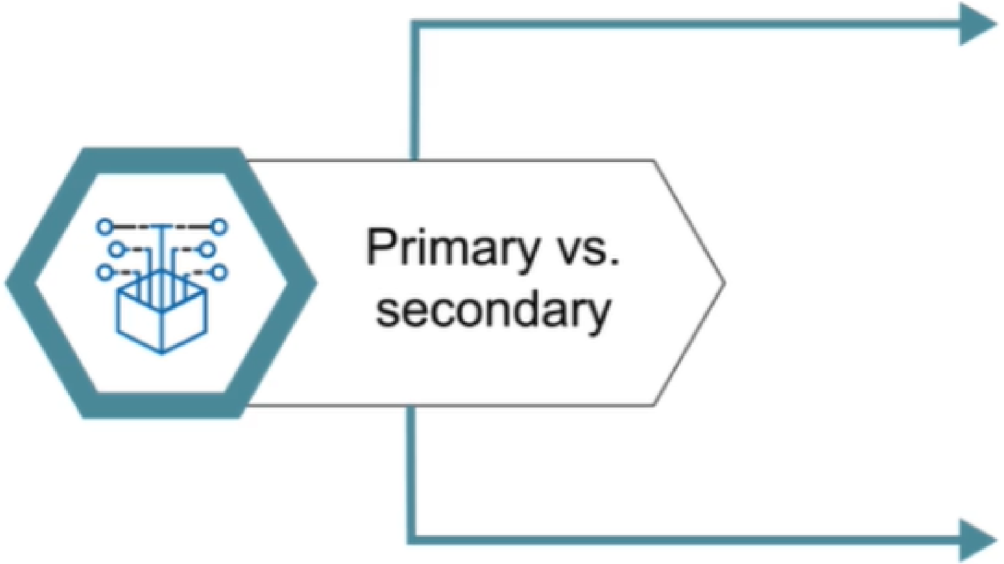
Perceptual data

- Relates to human perceptions
- Is subjective in nature
- Includes data on people's perceptions of quality, service and performance and greatly affects business outcomes

Objective data

- Is independent of subjective perception
- Includes events measured in physical attributes: time, space, distance, mass and money

Primary and Secondary Data



Data collection for research and analytics

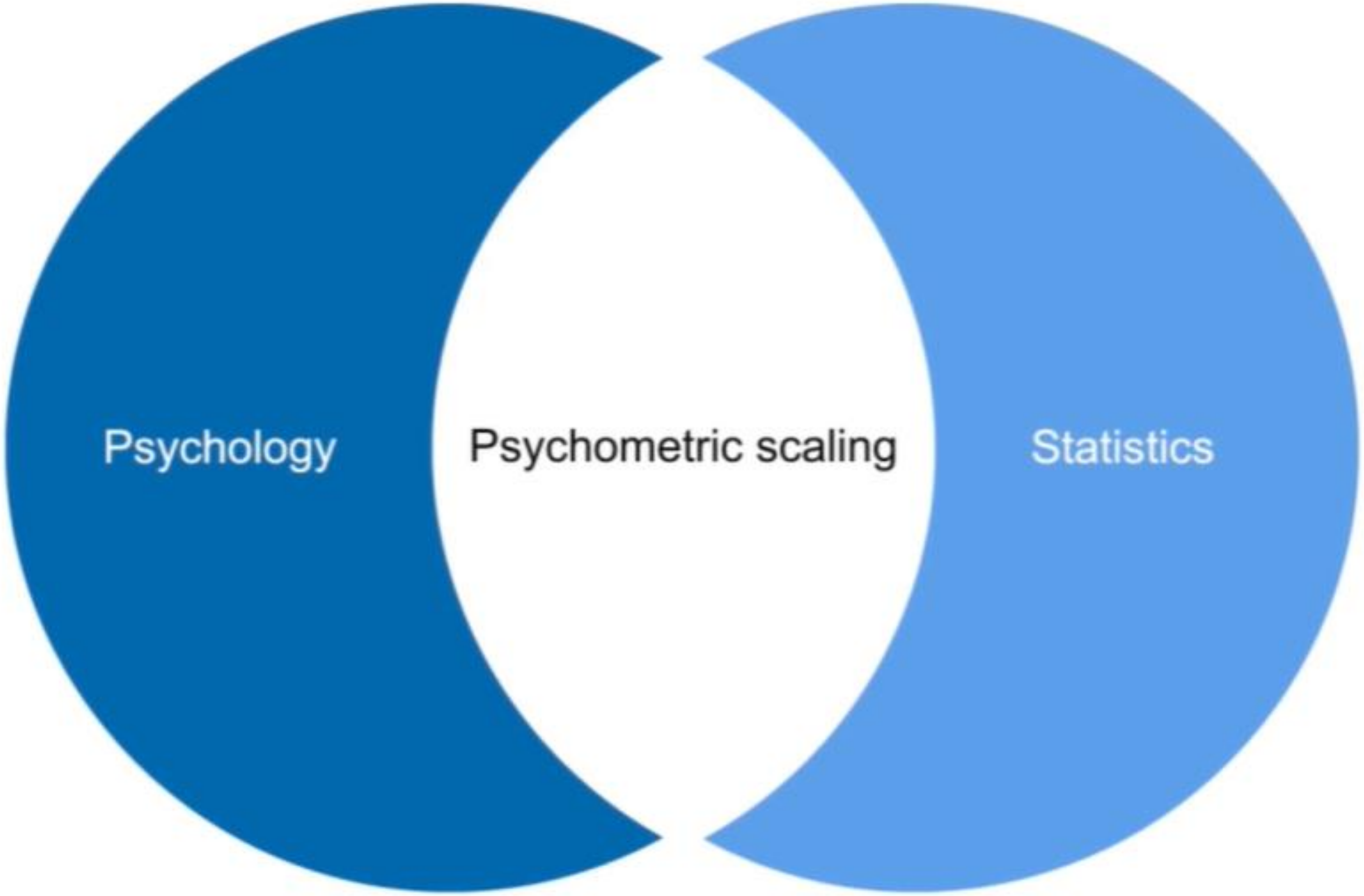
Primary data

- Data collected at source specifically for the research at hand
- Data source could include individuals, groups, organisations
- Surveys, interviews and focus groups serve as tools to collect primary data

Secondary data

- All data that are not primary
- Data collected previously for some other purpose but not for the research at hand
- Example: sales records, industry reports, interview transcripts
- APIs are an important source of secondary data

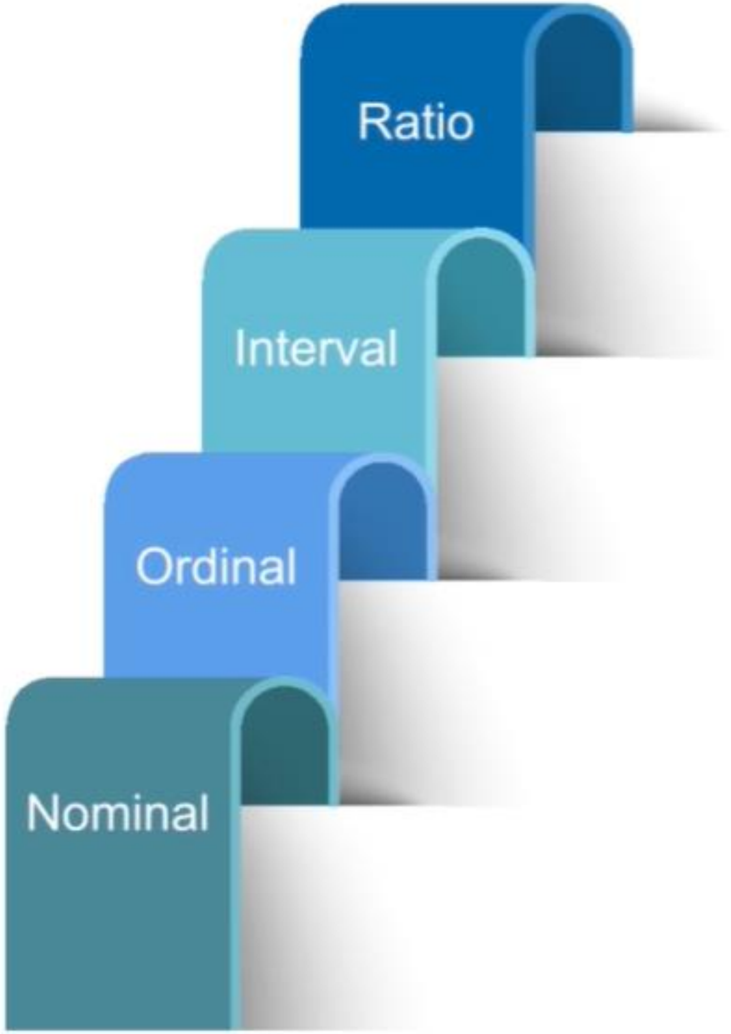
Introduction to Data Types



Psychometric scaling is the intersection of psychology and statistics

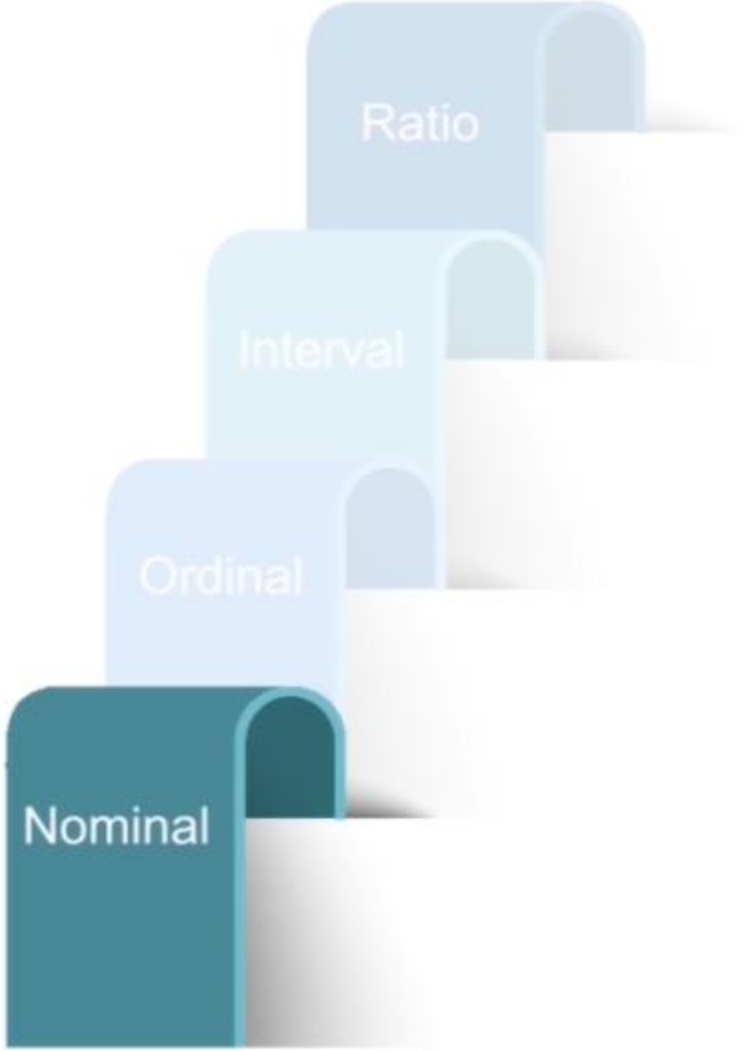
Stanley Smith Stevens' Theory of Scales

Different levels of measurement



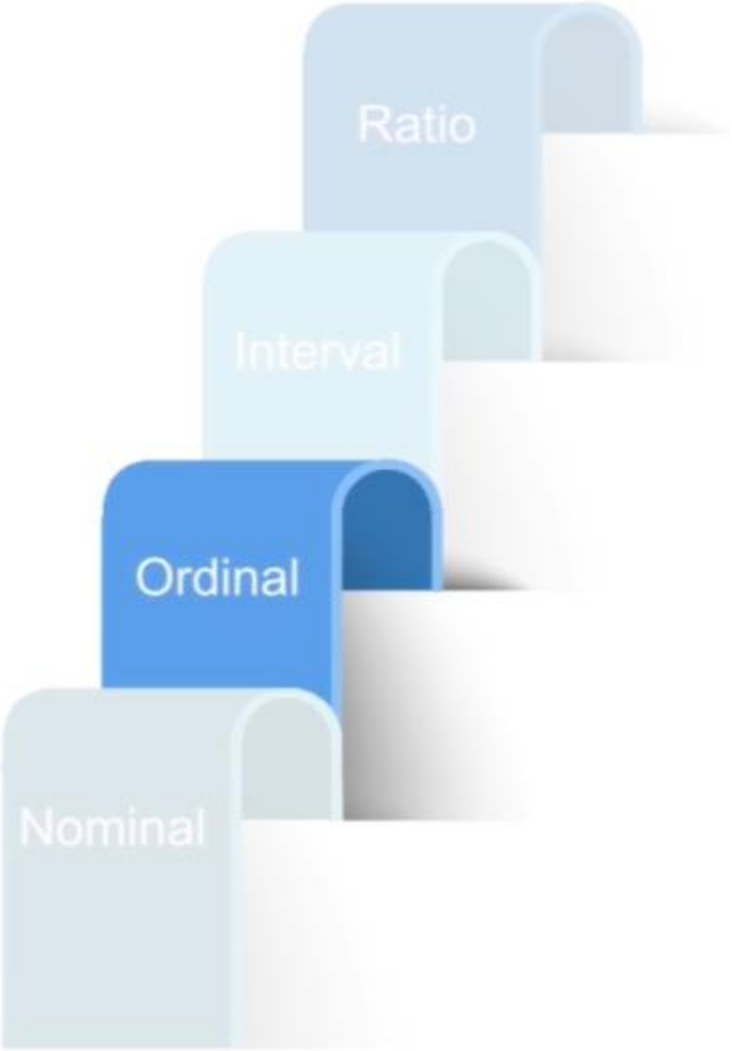
- For each type of feature, there are specific sets of permissible analytic or statistical operations
- Therefore, the scale in which data is collected matters

Data Types and Their Corresponding Primary Scales



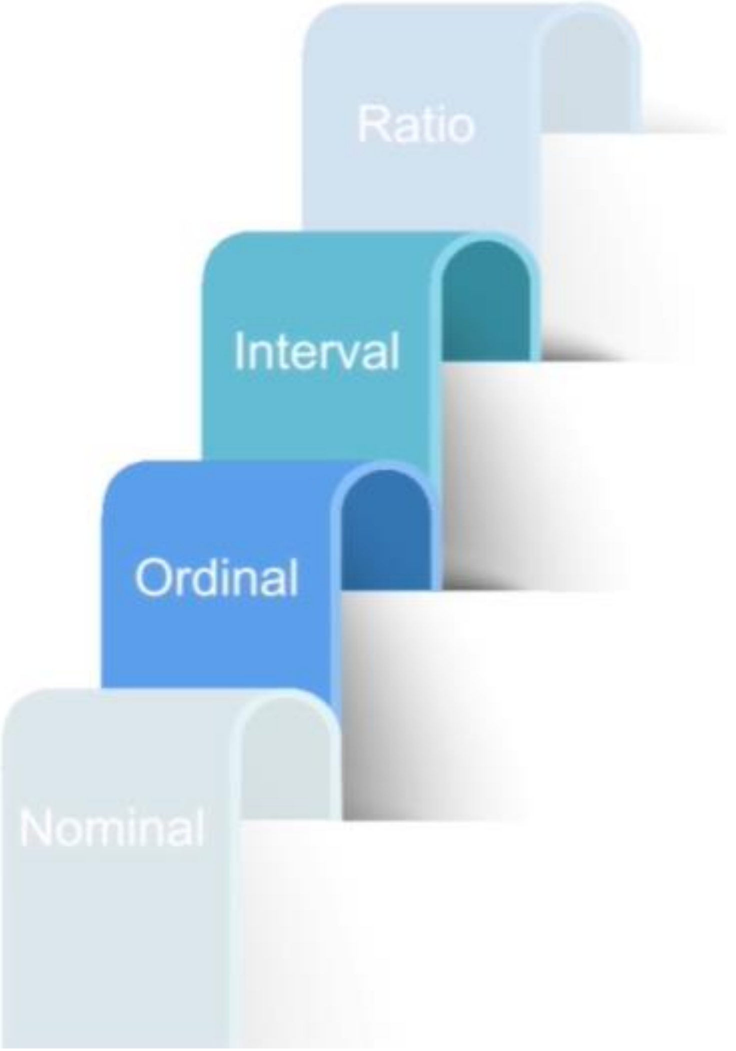
- Labels or names
- No further information can be gleaned beyond that label
- Example: A and B

Data Types and Their Corresponding Primary Scales



- Implies order
- Conveys preference information
- Conveys direction
- Example: A preferred to B, A > B,
  A is more than B and A is better than B

Data Types and Their Corresponding Primary Scales

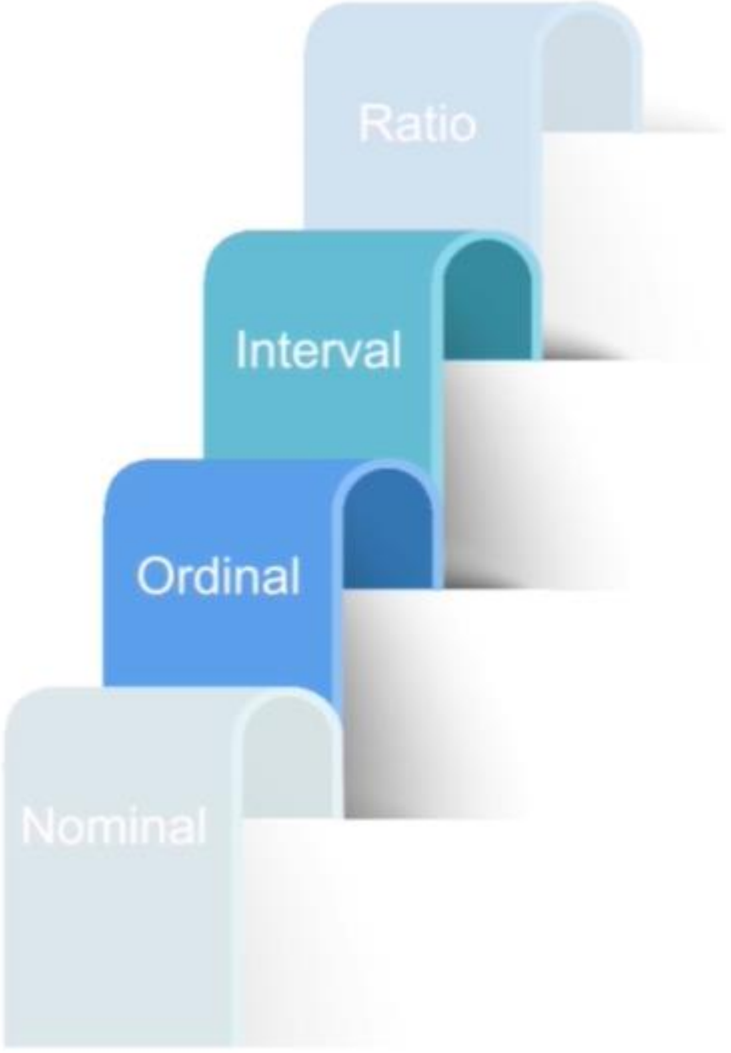


- Implies a uniform interval between any two ratings
- Conveys relative magnitude information and preference information

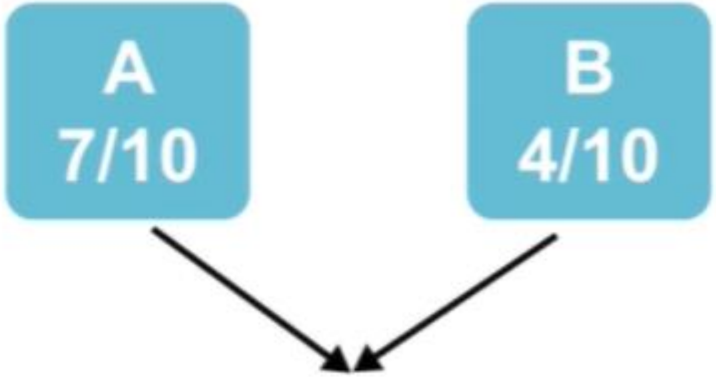
A is better than B (ordering is implied)
(also contains nominal information as it contains the names A and B)

- A is better than B – how much better?
- Conveys relative magnitude information

Data Types and Their Corresponding Primary Scales

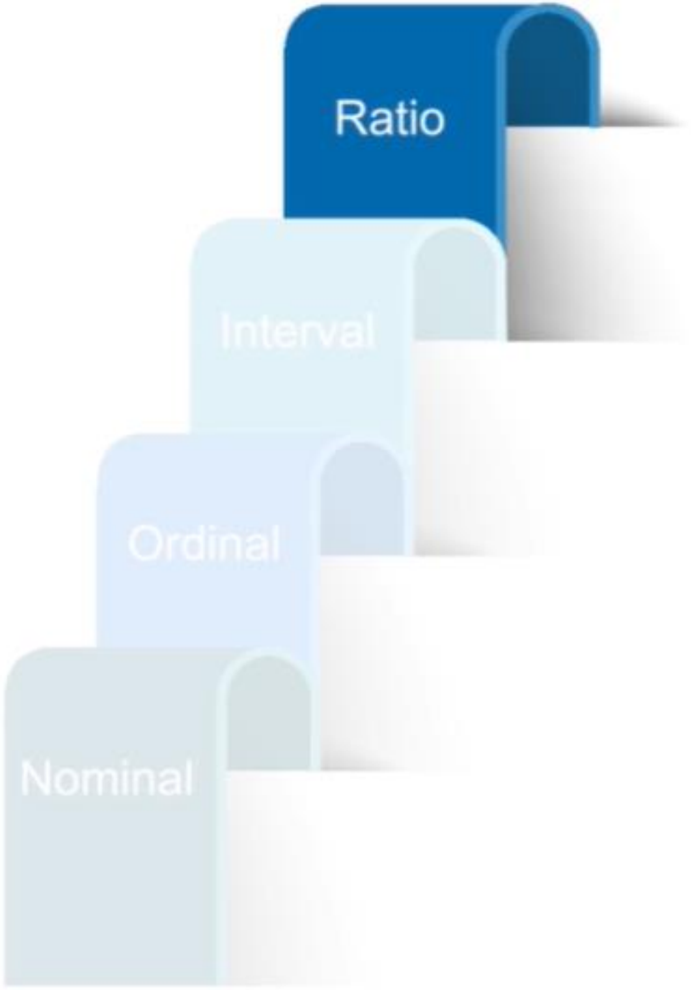


Example



- Contains nominal information
- Contains ordinal information
- Contains magnitude information

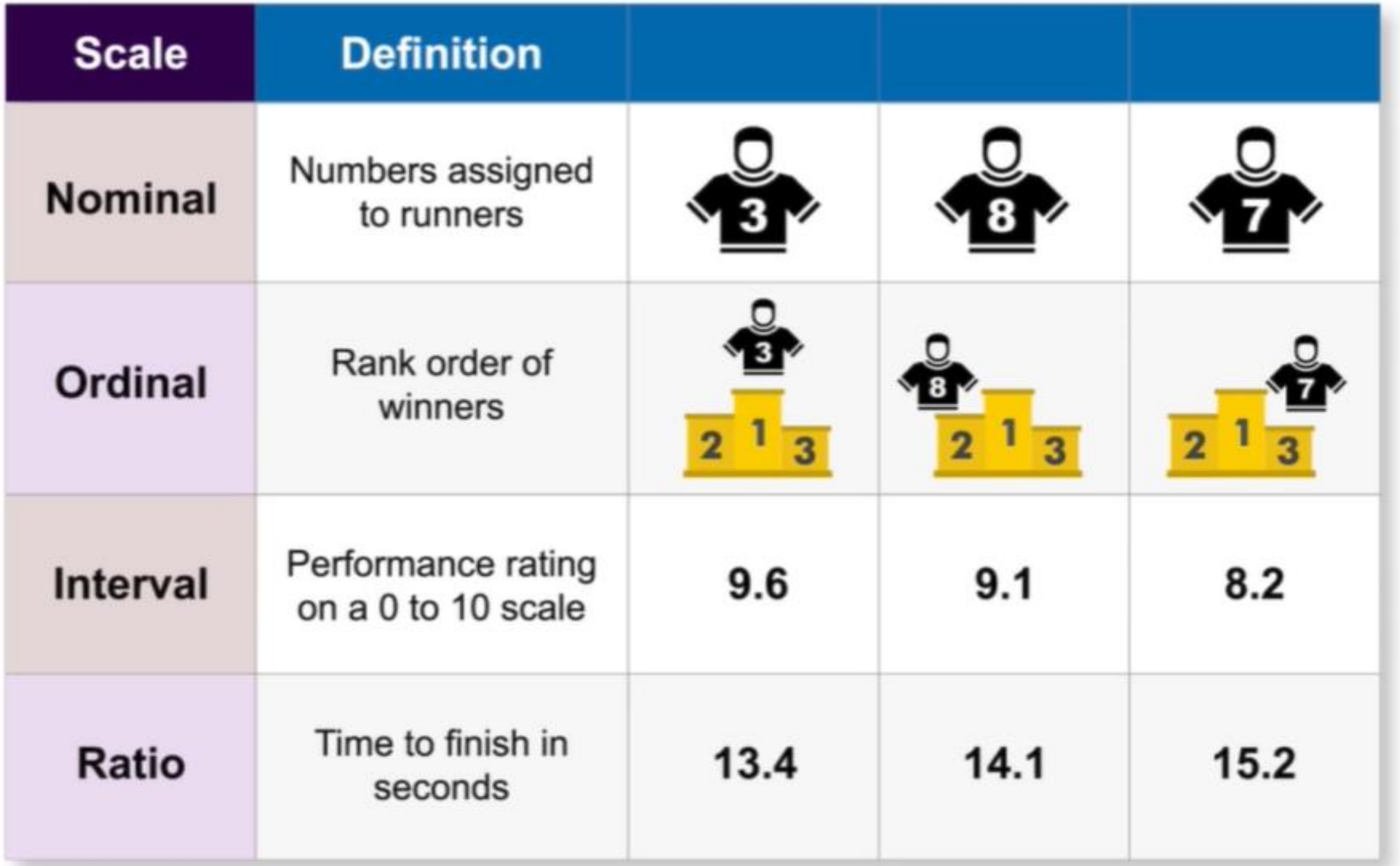
Data Types and Their Corresponding Primary Scales



- Gold standard in scales
- Highest-quality scale
- Conveys information on an absolute level
- The absolute zero is objectively defined and is independent of observer

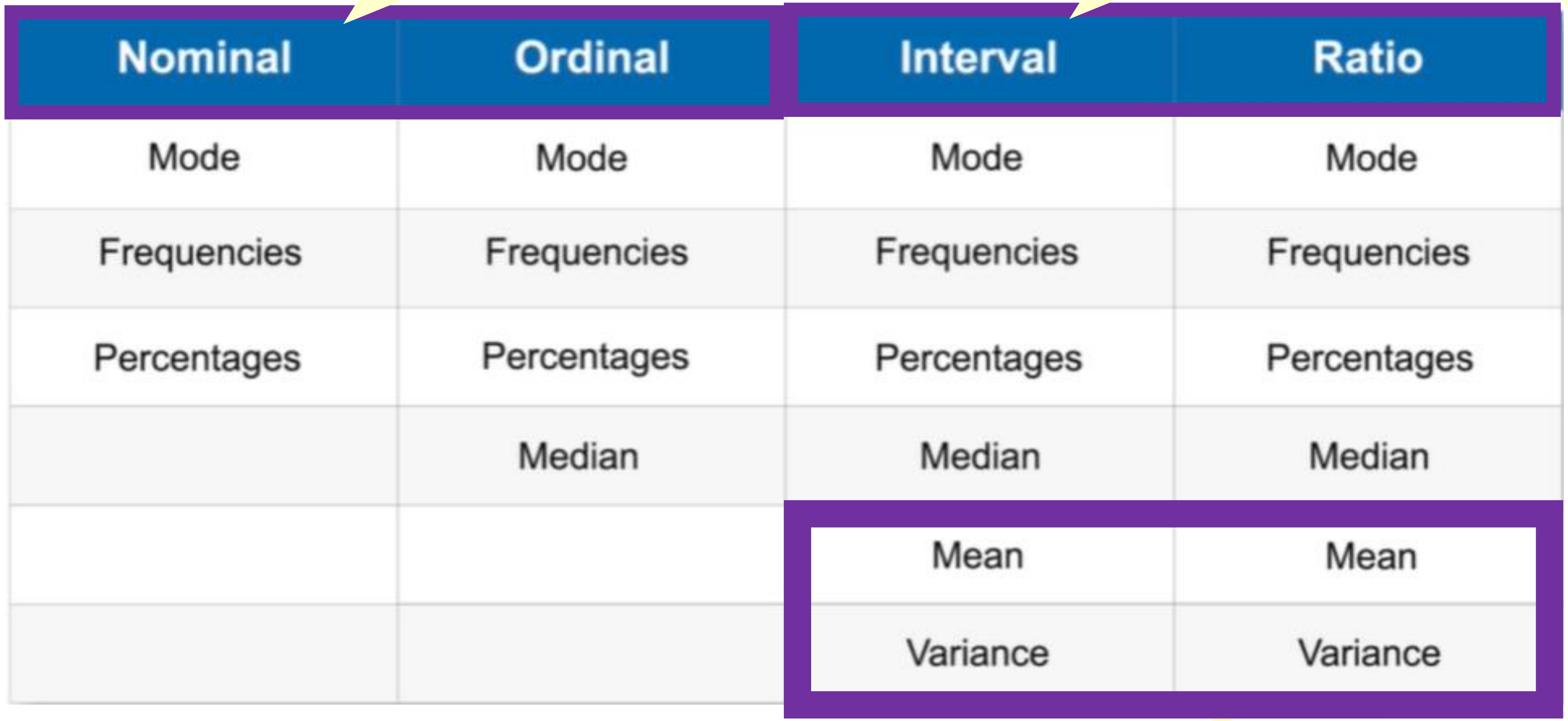
Primary Data Scales: Examples

100-meter dash sprint



Scale	Definition			
Nominal	Numbers assigned to runners	3	8	7
Ordinal	Rank order of winners	3 2 1 3	8 2 1 3	7 2 1 3
Interval	Performance rating on a 0 to 10 scale	9.6	9.1	8.2
Ratio	Time to finish in seconds	13.4	14.1	15.2

Choosing the Right Scale for Analysis



Nominal	Ordinal	Interval	Ratio
Mode	Mode	Mode	Mode
Frequencies	Frequencies	Frequencies	Frequencies
Percentages	Percentages	Percentages	Percentages
	Median	Median	Median
		Mean	Mean
		Variance	Variance

Non- metric data

Metric data

Metric scales

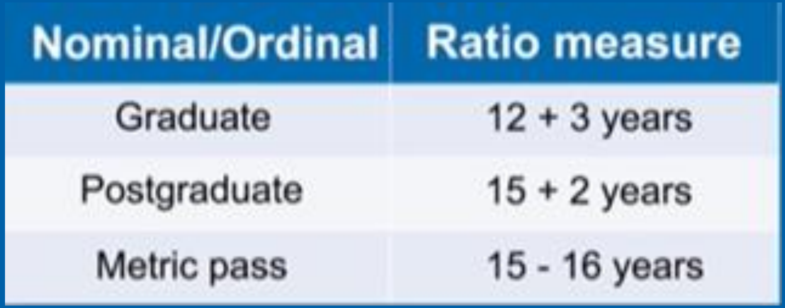
Metric space

Choosing the Right Scale for Analysis

As far as possible, collect your data using metric scales

Choosing the Right Scale for Analysis

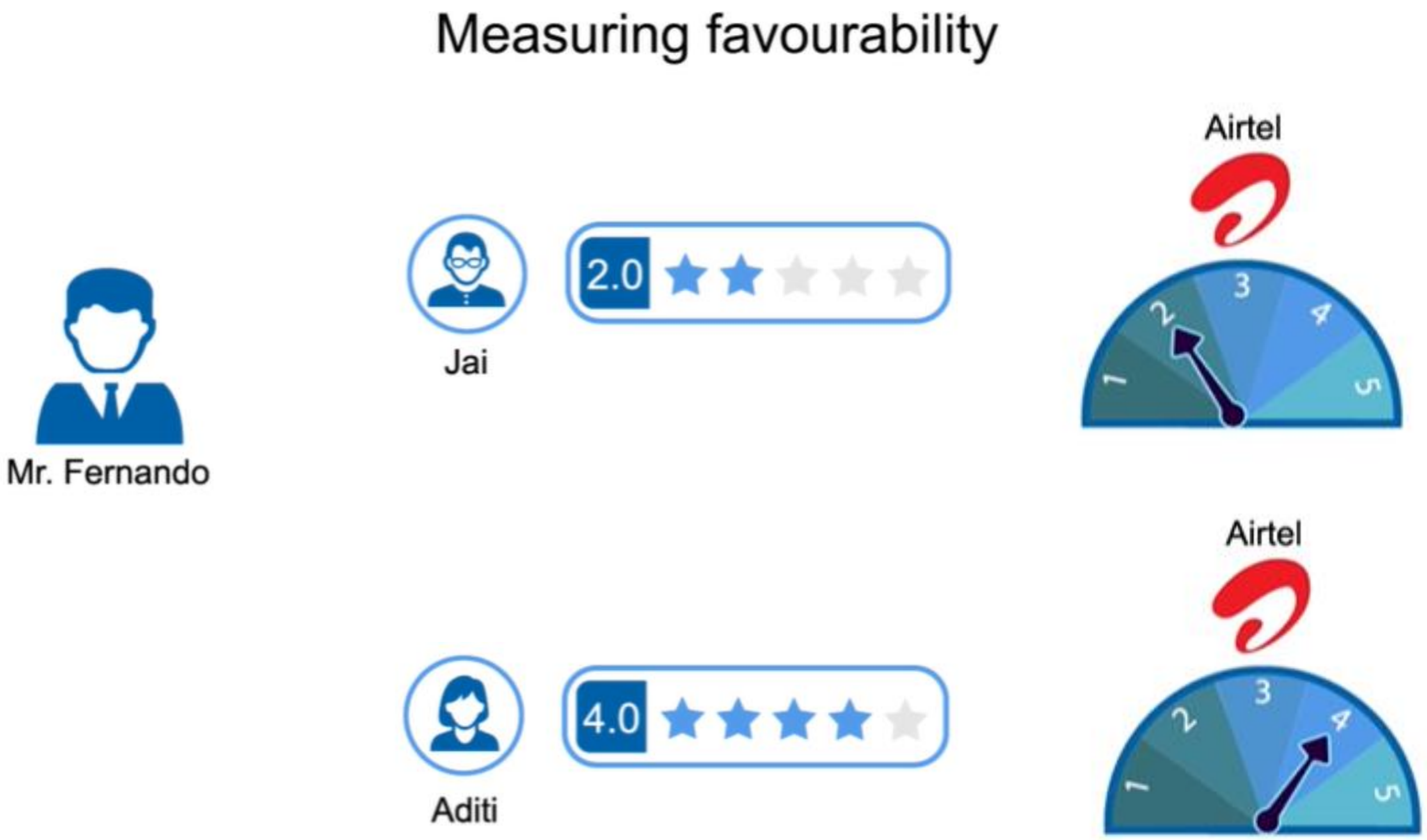
Measuring education levels in a population



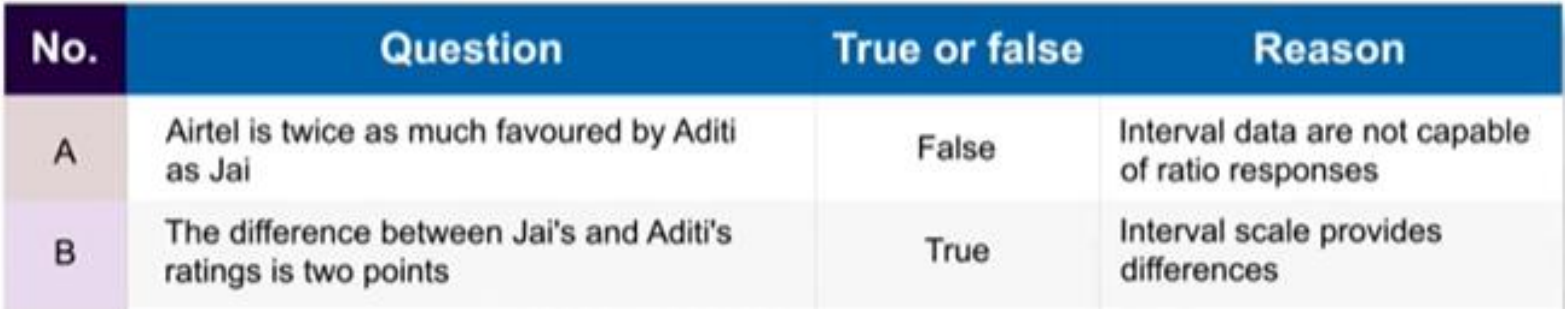
Nominal/Ordinal	Ratio measure
Graduate	12 + 3 years
Postgraduate	15 + 2 years
Metric pass	15 - 16 years

The quality of information contained in the collected data will affect the subsequent analysis

Debrief of Activity: Example-Interval Data

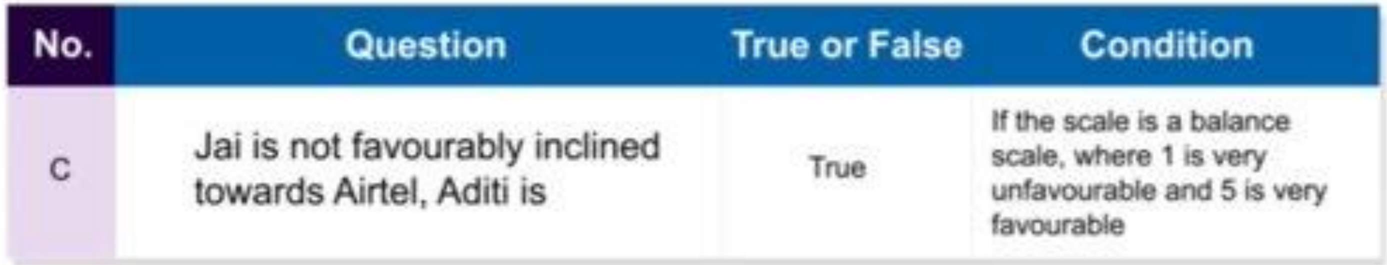


Example: Interval Data

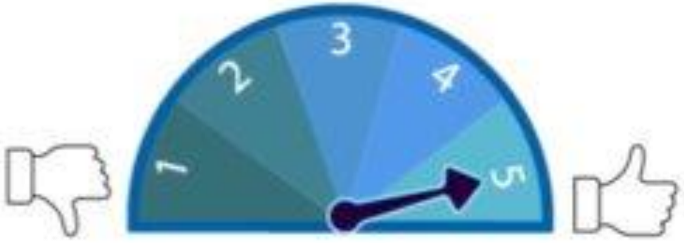


No.	Question	True or false	Reason
A	Airtel is twice as much favoured by Aditi as Jai	False	Interval data are not capable of ratio responses
B	The difference between Jai's and Aditi's ratings is two points	True	Interval scale provides differences

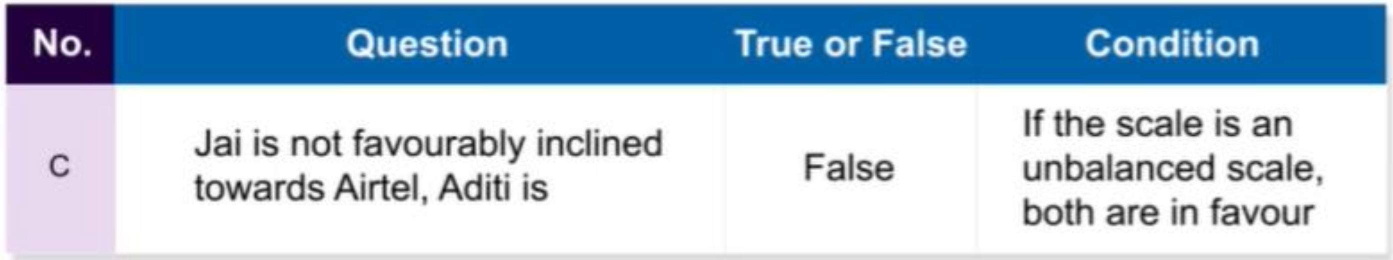
Example: Interval Data



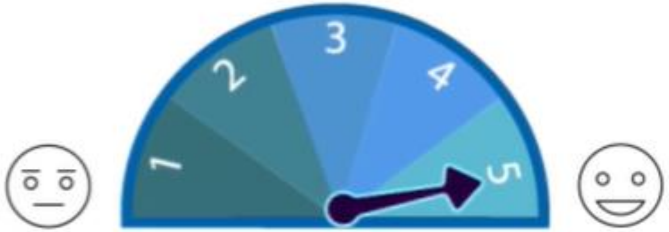
No.	Question	True or False	Condition
C	Jai is not favourably inclined towards Airtel, Aditi is	True	If the scale is a balance scale, where 1 is very unfavourable and 5 is very favourable



Balanced scale



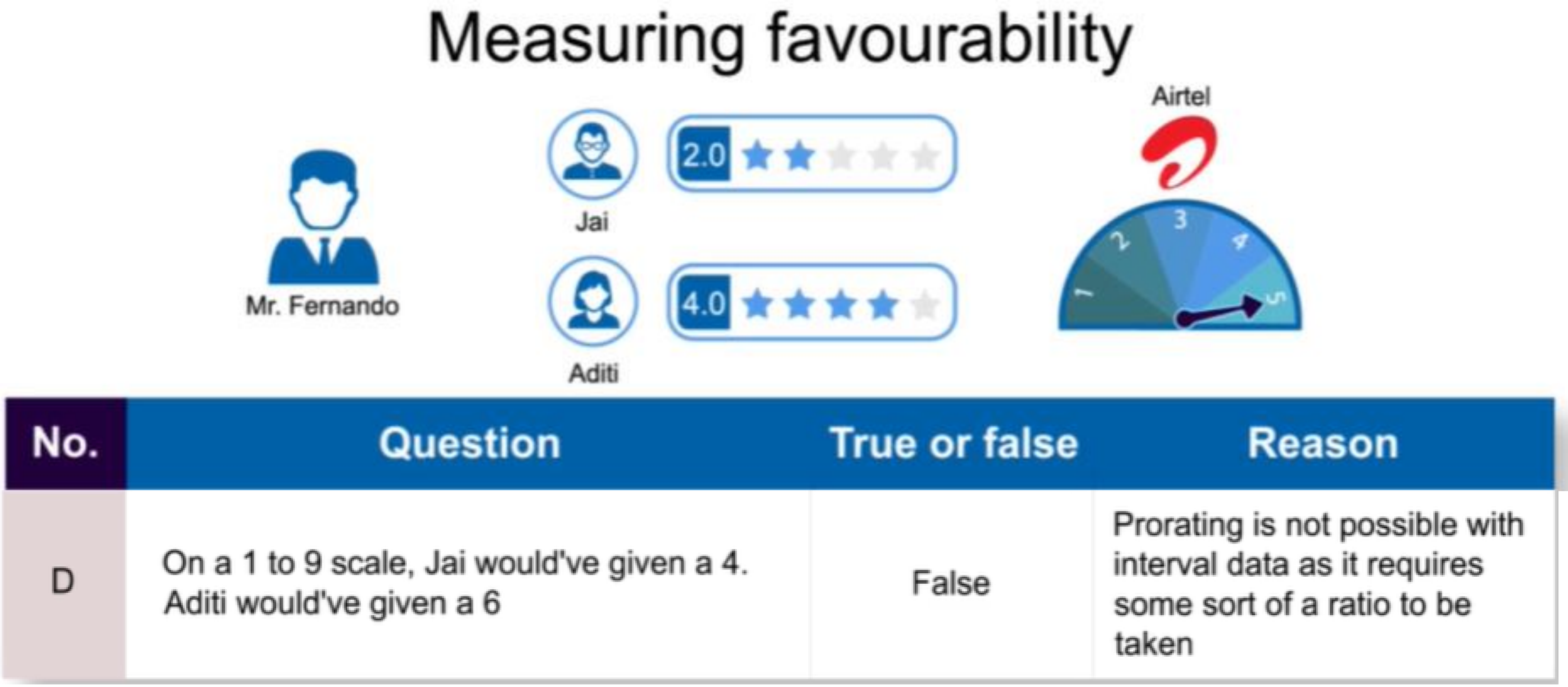
No.	Question	True or False	Condition
C	Jai is not favourably inclined towards Airtel, Aditi is	False	If the scale is an unbalanced scale, both are in favour



Unbalanced scale

Therefore, it is necessary to know scale guidance in primary perceptual data

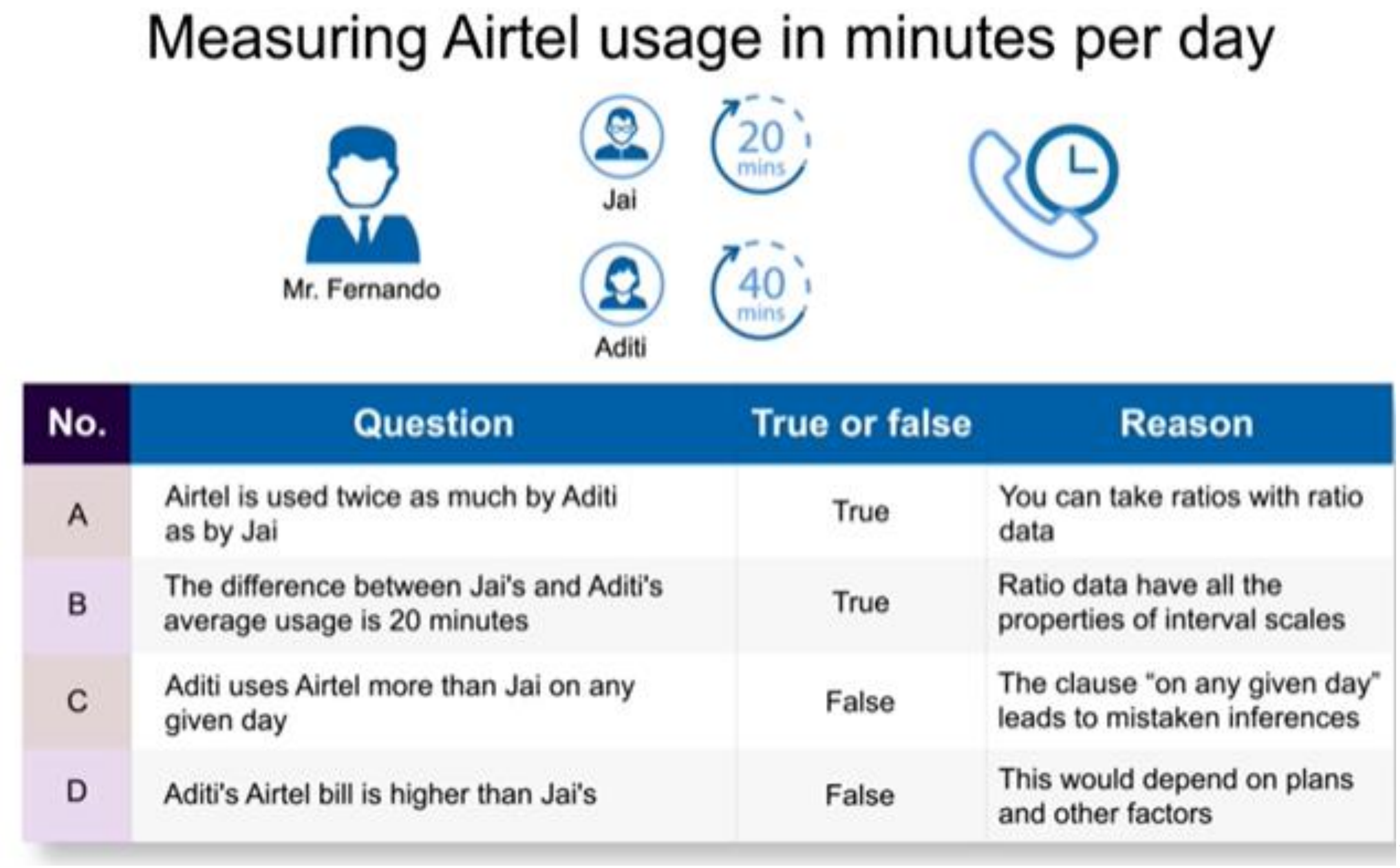
Data Preliminaries for Analytics: Summary



Measuring favourability

No.	Question	True or false	Reason
D	On a 1 to 9 scale, Jai would've given a 4. Aditi would've given a 6	False	Prorating is not possible with interval data as it requires some sort of a ratio to be taken

Example: Ratio Data



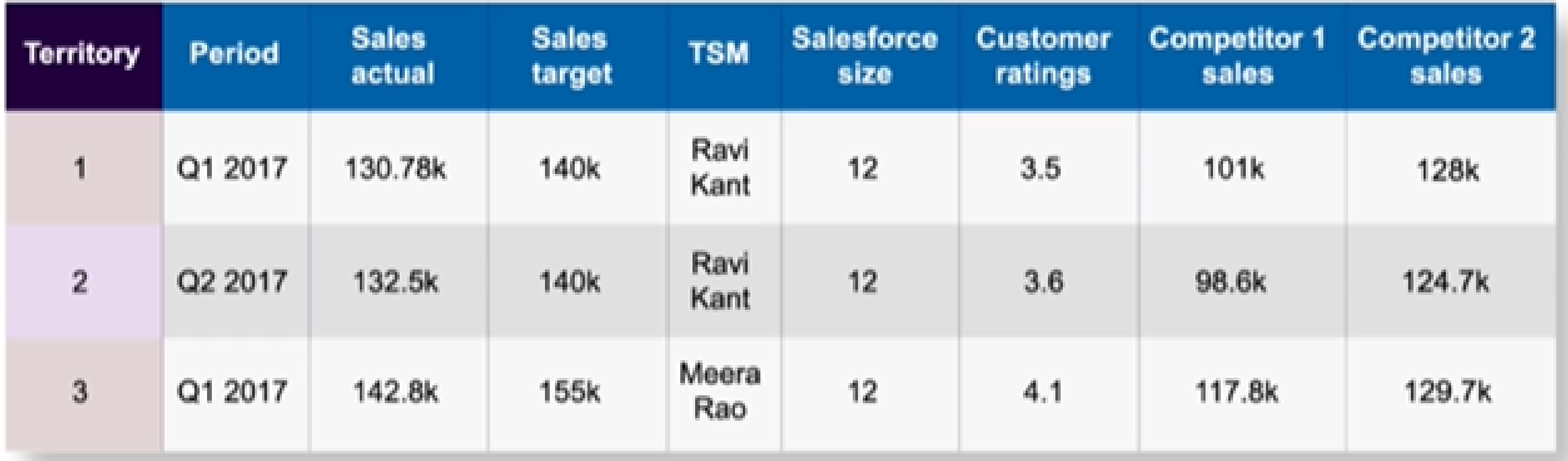
Measuring Airtel usage in minutes per day

Mr. Fernando
Jai
20 mins
Aditi
40 mins

No.	Question	True or false	Reason
A	Airtel is used twice as much by Aditi as by Jai	True	You can take ratios with ratio data
B	The difference between Jai's and Aditi's average usage is 20 minutes	True	Ratio data have all the properties of interval scales
C	Aditi uses Airtel more than Jai on any given day	False	The clause "on any given day" leads to mistaken inferences
D	Aditi's Airtel bill is higher than Jai's	False	This would depend on plans and other factors

Example: Metric vs Non-metric

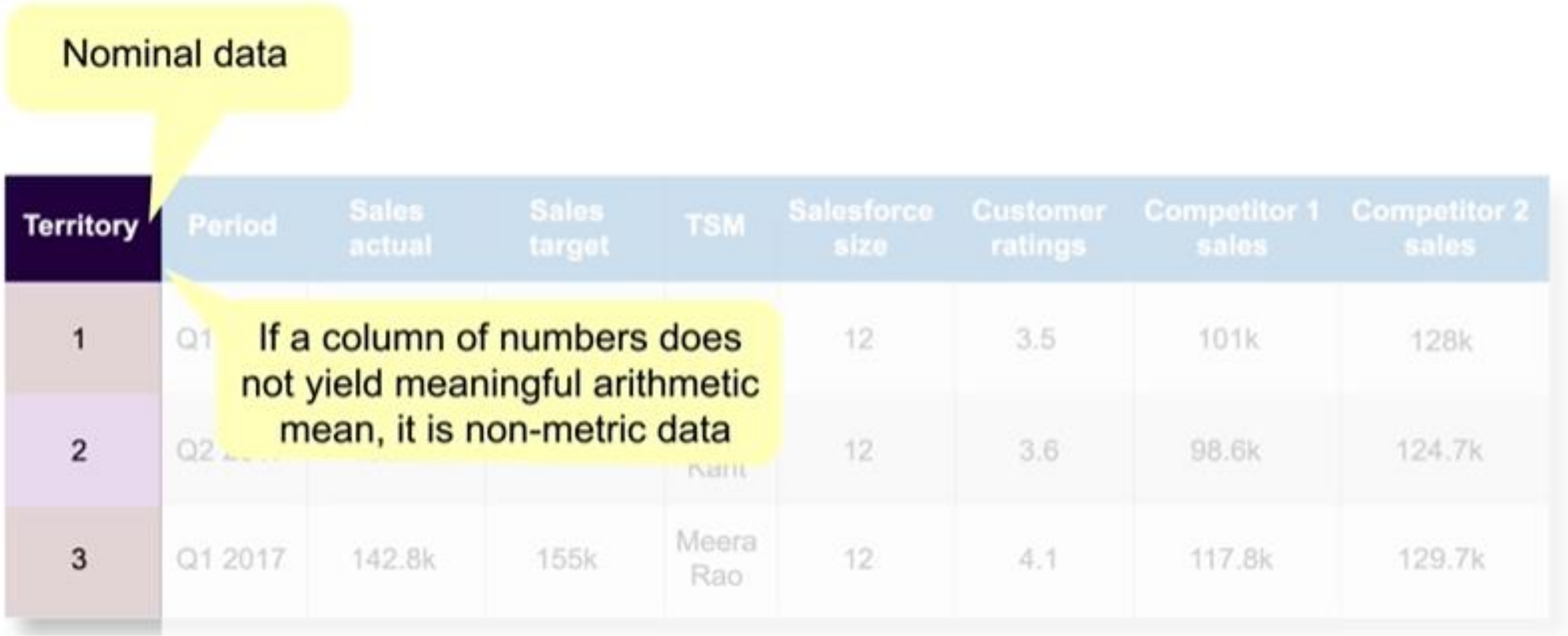
Salesforce data



Territory	Period	Sales actual	Sales target	TSM	Salesforce size	Customer ratings	Competitor 1 sales	Competitor 2 sales
1	Q1 2017	130.78k	140k	Ravi Kant	12	3.5	101k	128k
2	Q2 2017	132.5k	140k	Ravi Kant	12	3.6	98.6k	124.7k
3	Q1 2017	142.8k	155k	Meera Rao	12	4.1	117.8k	129.7k

- Which of these variables in the table indicate metric data?
- What kind within metrics? Is it interval or non? Or a ratio?
- Which of the variables are non-metric?
- What kind within non-metric?

Example: Metric vs Non-metric

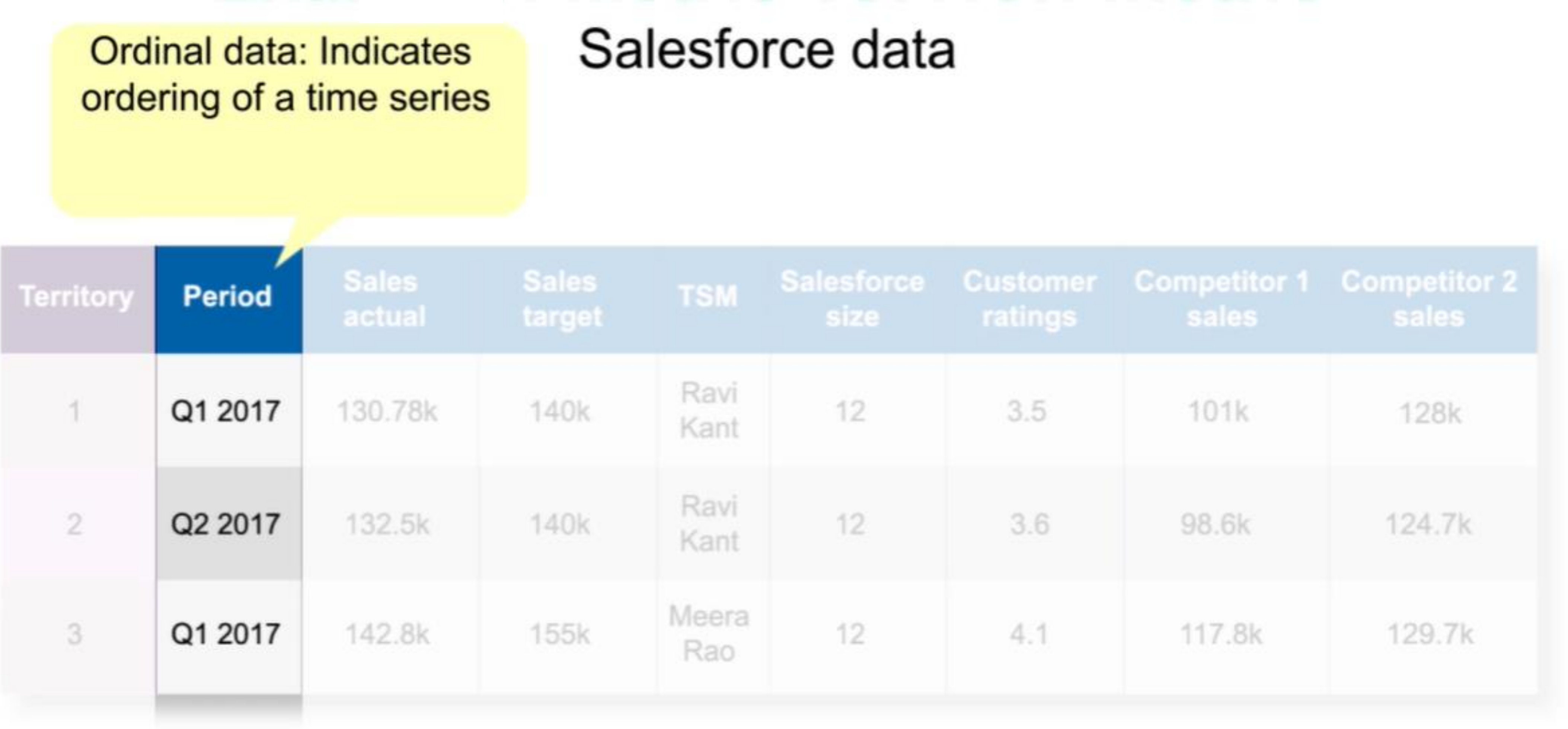


Nominal data

Territory	Period	Sales actual	Sales target	TSM	Salesforce size	Customer ratings	Competitor 1 sales	Competitor 2 sales
1	Q1 [illegible]	[illegible]	[illegible]	[illegible]	12	3.5	101k	128k
2	Q2 [illegible]	[illegible]	[illegible]	[illegible] Kant	12	3.6	98.6k	124.7k
3	Q1 2017	142.8k	155k	Meera Rao	12	4.1	117.8k	129.7k

If a column of numbers does not yield meaningful arithmetic mean, it is non-metric data

Example: Metric vs Non-metric



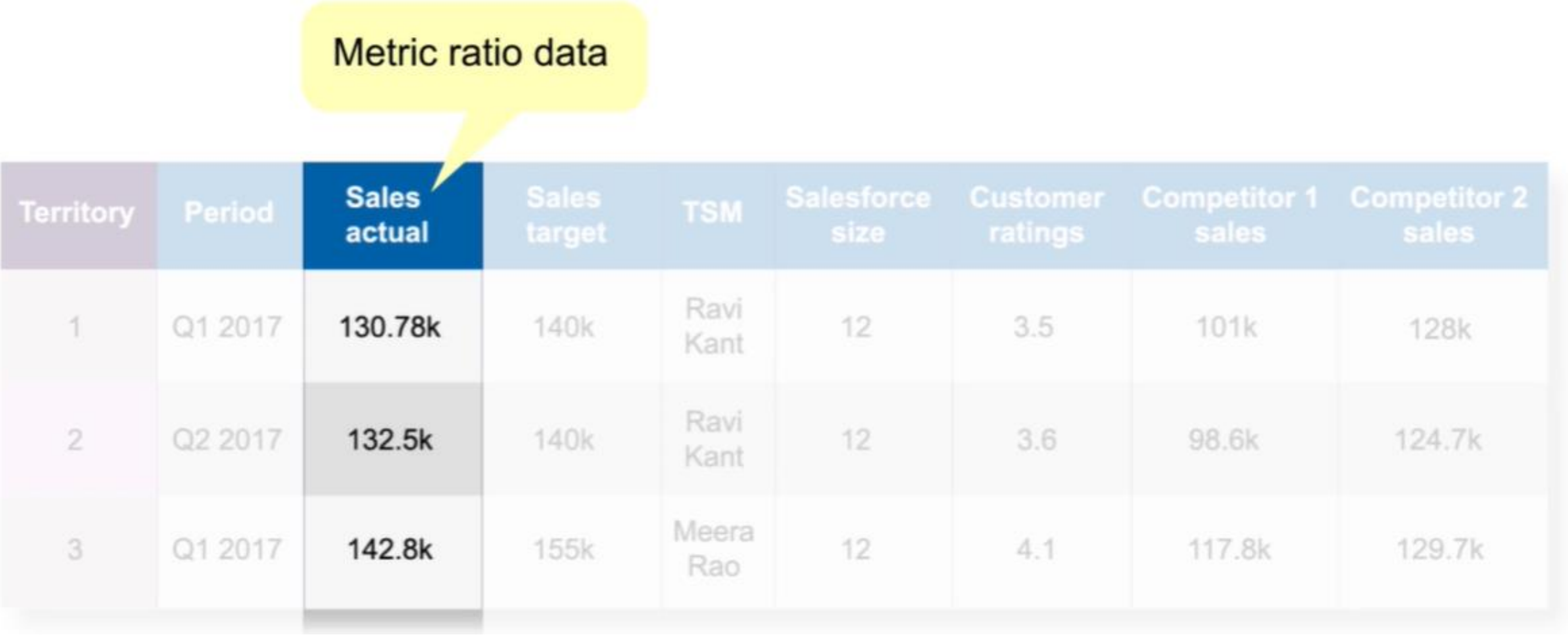
Salesforce data

Ordinal data: Indicates ordering of a time series

Territory	Period	Sales actual	Sales target	TSM	Salesforce size	Customer ratings	Competitor 1 sales	Competitor 2 sales
1	Q1 2017	130.78k	140k	Ravi Kant	12	3.5	101k	128k
2	Q2 2017	132.5k	140k	Ravi Kant	12	3.6	98.6k	124.7k
3	Q1 2017	142.8k	155k	Meera Rao	12	4.1	117.8k	129.7k

Example: Metric vs Non-metric

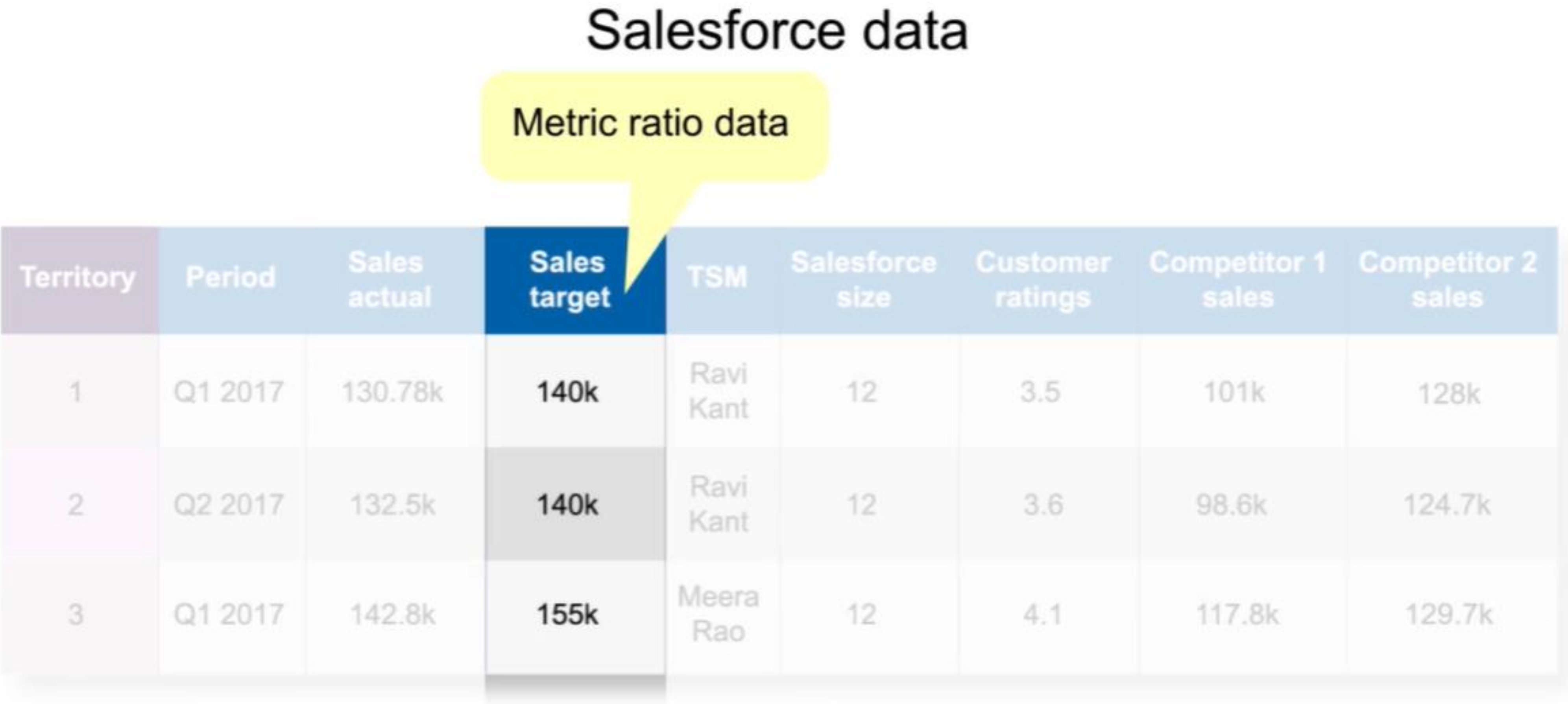
Salesforce data



Metric ratio data

Territory	Period	Sales actual	Sales target	TSM	Salesforce size	Customer ratings	Competitor 1 sales	Competitor 2 sales
1	Q1 2017	130.78k	140k	Ravi Kant	12	3.5	101k	128k
2	Q2 2017	132.5k	140k	Ravi Kant	12	3.6	98.6k	124.7k
3	Q1 2017	142.8k	155k	Meera Rao	12	4.1	117.8k	129.7k

Example: Metric vs Non-metric

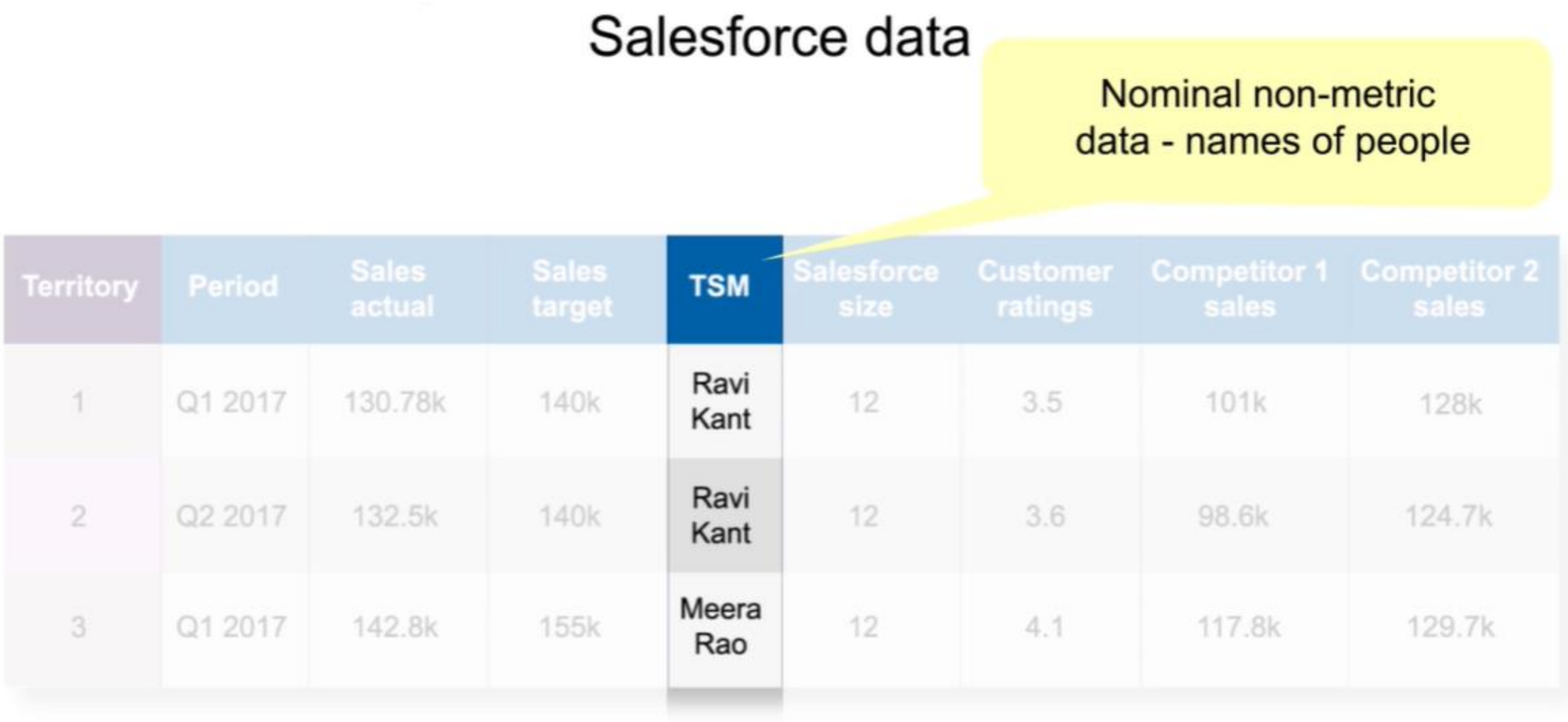


Salesforce data

Metric ratio data

Territory	Period	Sales actual	Sales target	TSM	Salesforce size	Customer ratings	Competitor 1 sales	Competitor 2 sales
1	Q1 2017	130.78k	140k	Ravi Kant	12	3.5	101k	128k
2	Q2 2017	132.5k	140k	Ravi Kant	12	3.6	98.6k	124.7k
3	Q1 2017	142.8k	155k	Meera Rao	12	4.1	117.8k	129.7k

Example: Metric vs Non-metric



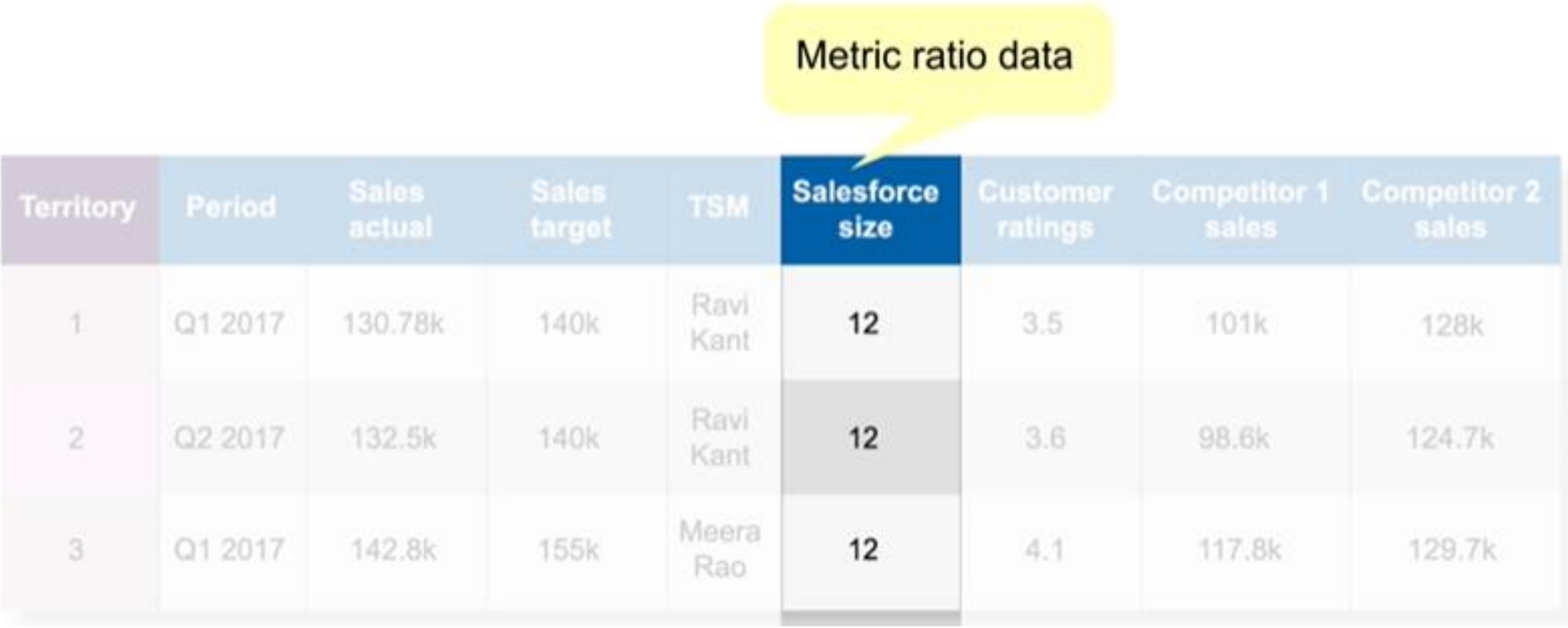
Salesforce data

Nominal non-metric data - names of people

Territory	Period	Sales actual	Sales target	TSM	Salesforce size	Customer ratings	Competitor 1 sales	Competitor 2 sales
1	Q1 2017	130.78k	140k	Ravi Kant	12	3.5	101k	128k
2	Q2 2017	132.5k	140k	Ravi Kant	12	3.6	98.6k	124.7k
3	Q1 2017	142.8k	155k	Meera Rao	12	4.1	117.8k	129.7k

Example: Metric vs Non-metric

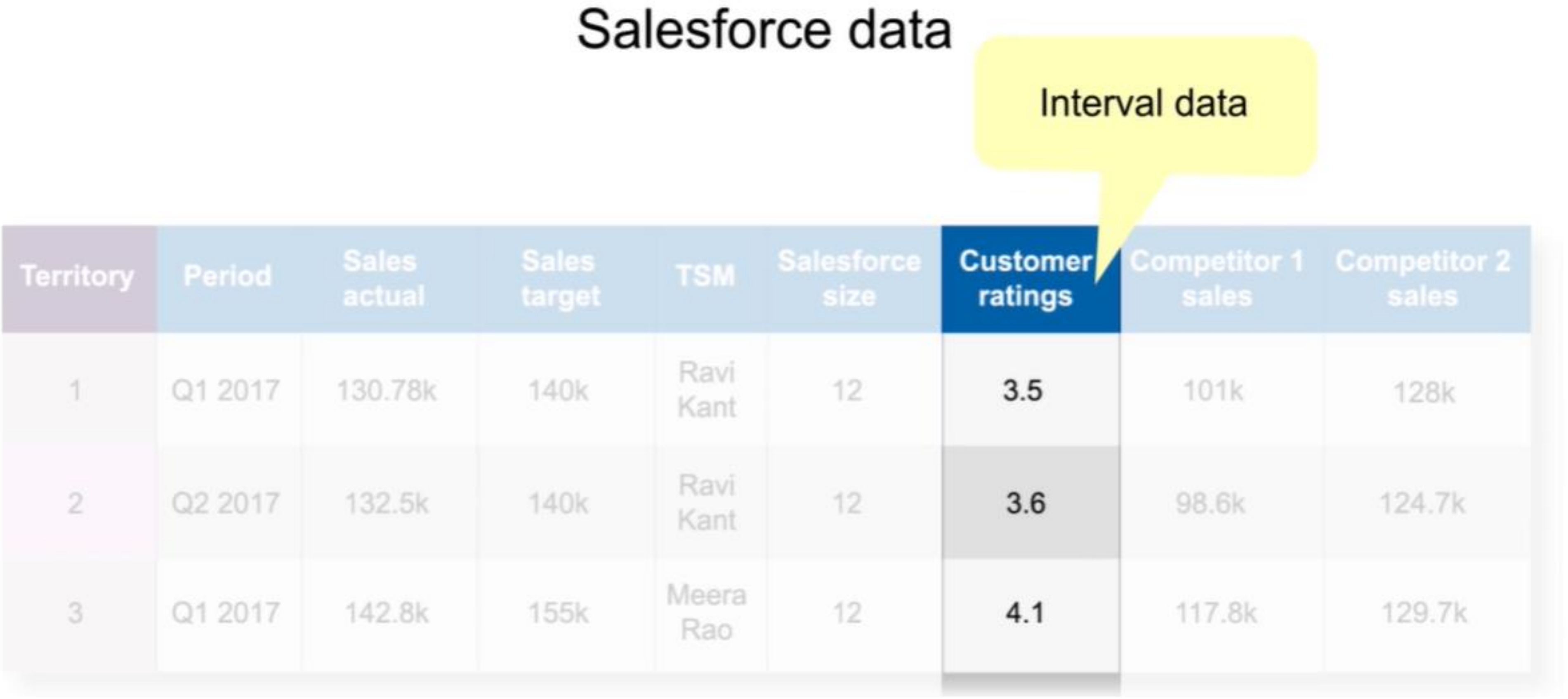
Salesforce data



Metric ratio data

Territory	Period	Sales actual	Sales target	TSM	Salesforce size	Customer ratings	Competitor 1 sales	Competitor 2 sales
1	Q1 2017	130.78k	140k	Ravi Kant	12	3.5	101k	128k
2	Q2 2017	132.5k	140k	Ravi Kant	12	3.6	98.6k	124.7k
3	Q1 2017	142.8k	155k	Meera Rao	12	4.1	117.8k	129.7k

Example: Metric vs Non-metric

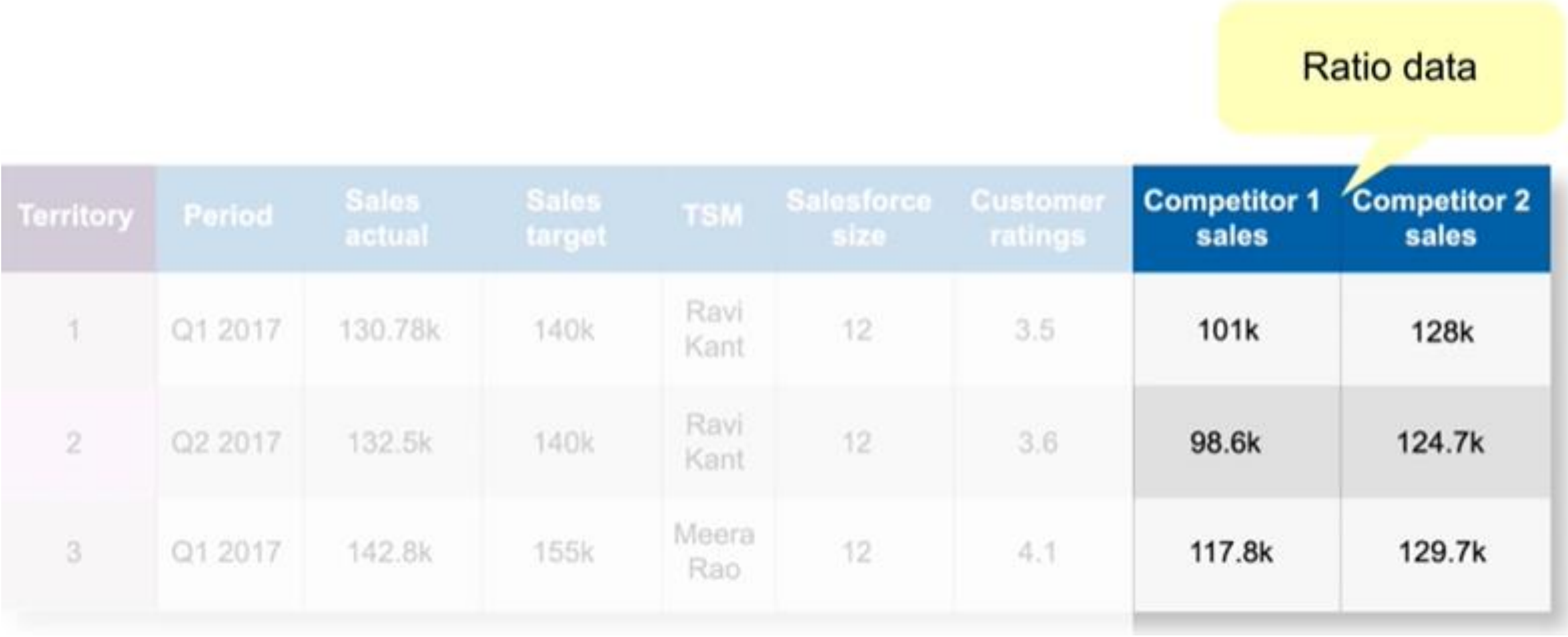


Salesforce data

Interval data

Territory	Period	Sales actual	Sales target	TSM	Salesforce size	Customer ratings	Competitor 1 sales	Competitor 2 sales
1	Q1 2017	130.78k	140k	Ravi Kant	12	3.5	101k	128k
2	Q2 2017	132.5k	140k	Ravi Kant	12	3.6	98.6k	124.7k
3	Q1 2017	142.8k	155k	Meera Rao	12	4.1	117.8k	129.7k

Example: Metric vs Non-metric



Ratio data

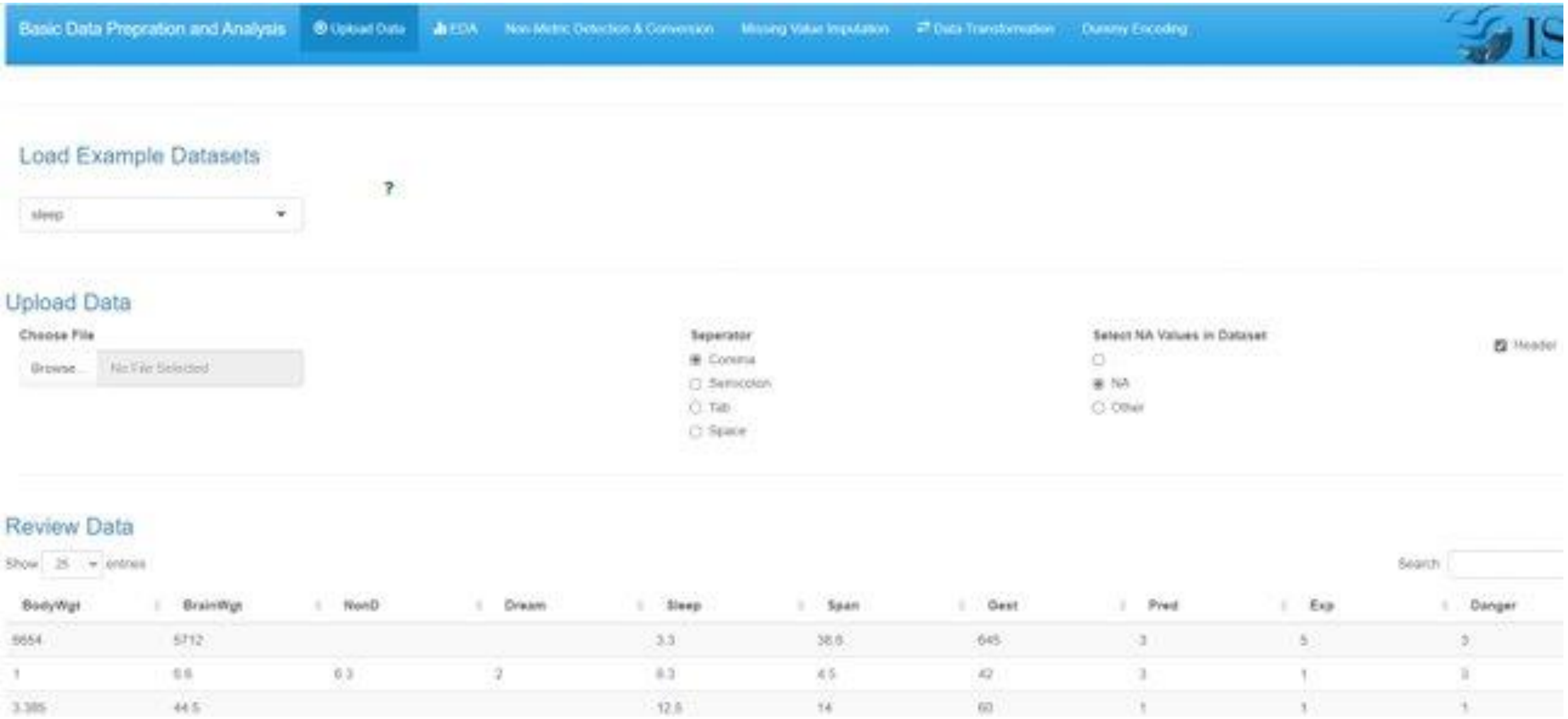
Territory	Period	Sales actual	Sales target	TSM	Salesforce size	Customer ratings	Competitor 1 sales	Competitor 2 sales
1	Q1 2017	130.78k	140k	Ravi Kant	12	3.5	101k	128k
2	Q2 2017	132.5k	140k	Ravi Kant	12	3.6	98.6k	124.7k
3	Q1 2017	142.8k	155k	Meera Rao	12	4.1	117.8k	129.7k

Why Care about Data Pre-processing?

- Because to be useful, data must first be *usable* for analysis
- To be **usable**, data must be **clean** and **consistent**, that is ...
  - Have no lost or missing values (hence, data *imputation*)
  - Have no mis-identified columns (e.g., nonmetric variable mistakenly used as metric)
  - Have sufficient variance in every variable (variance implies informativeness)
  - Have adequate transformations (e.g., *re-scaling* of variables, creation of dummy variables), etc.
- The **Data-Preproc App** provides a one-stop (small-sample) way forward to us.

Data-PreProc App Layout

- Let’s first examine the App layout, in particular…
  - Input UI elements
  - Output tabs
  - Then, we will go into each element and examine its workings



Load Example Datasets

sleep ▾ ?

Upload Data

Choose File

Browse... No File Selected

Seperator

● Comma
○ Semicolon
○ Tab
○ Space

Select NA Values in Dataset

○
● NA
○ Other

☑ Header

Review Data

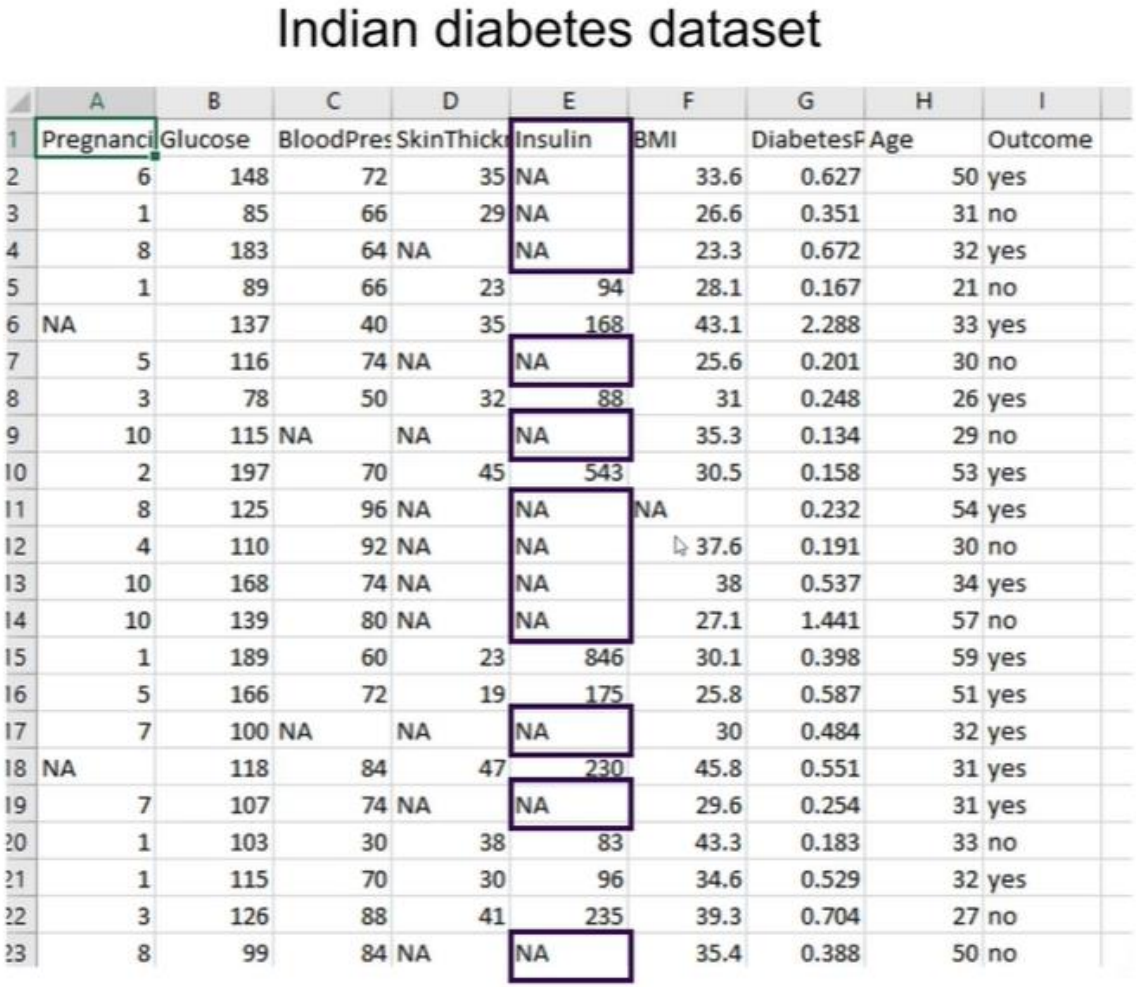
Show 25 ▾ entries

Search:

BodyWgt	BrainWgt	NonD	Dream	Sleep	Span	Gest	Pred	Exp	Danger
6654	5712			3.3	38.6	645	3	5	3
1	6.6	6.3	2	8.3	4.5	42	3	1	3
3.385	44.5			12.5	14	60	1	1	1

Data Walkthrough for the Data-PreProc App

First, examine the Diabetes Dataset (*diabetes.csv*).



Indian diabetes dataset

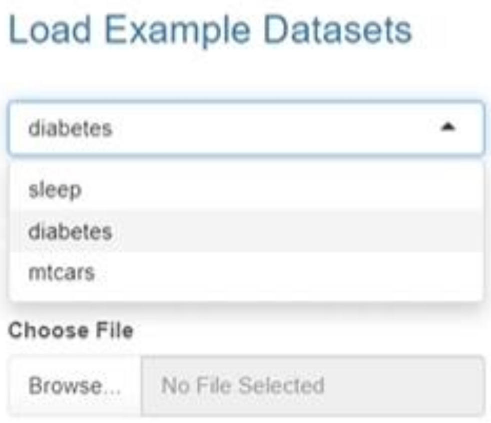
	A	B	C	D	E	F	G	H	I
1	Pregnanci	Glucose	BloodPres	SkinThick	Insulin	BMI	DiabetesF	Age	Outcome
2	6	148	72	35	NA	33.6	0.627	50	yes
3	1	85	66	29	NA	26.6	0.351	31	no
4	8	183	64	NA	NA	23.3	0.672	32	yes
5	1	89	66	23	94	28.1	0.167	21	no
6	NA	137	40	35	168	43.1	2.288	33	yes
7	5	116	74	NA	NA	25.6	0.201	30	no
8	3	78	50	32	88	31	0.248	26	yes
9	10	115	NA	NA	NA	35.3	0.134	29	no
10	2	197	70	45	543	30.5	0.158	53	yes
11	8	125	96	NA	NA	NA	0.232	54	yes
12	4	110	92	NA	NA	37.6	0.191	30	no
13	10	168	74	NA	NA	38	0.537	34	yes
14	10	139	80	NA	NA	27.1	1.441	57	no
15	1	189	60	23	846	30.1	0.398	59	yes
16	5	166	72	19	175	25.8	0.587	51	yes
17	7	100	NA	NA	NA	30	0.484	32	yes
18	NA	118	84	47	230	45.8	0.551	31	yes
19	7	107	74	NA	NA	29.6	0.254	31	yes
20	1	103	30	38	83	43.3	0.183	33	no
21	1	115	70	30	96	34.6	0.529	32	yes
22	3	126	88	41	235	39.3	0.704	27	no
23	8	99	84	NA	NA	35.4	0.388	50	no

Data Walkthrough for the Data PreProc App

The variables are self-explanatory, but descriptions can be found in the pre-loaded dataset.

Indian diabetes dataset

- The variables are self-explanatory



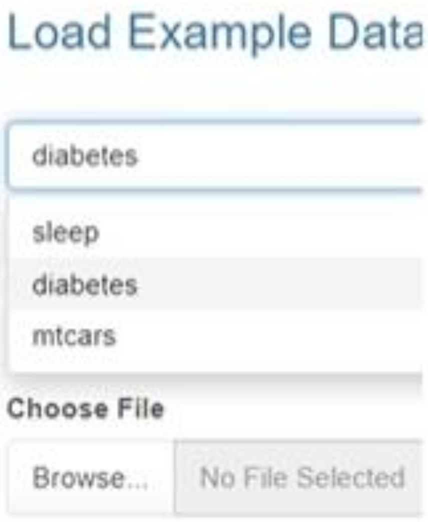
Load Example Datasets

diabetes

sleep
diabetes
mtcars

Choose File

Browse... No File Selected



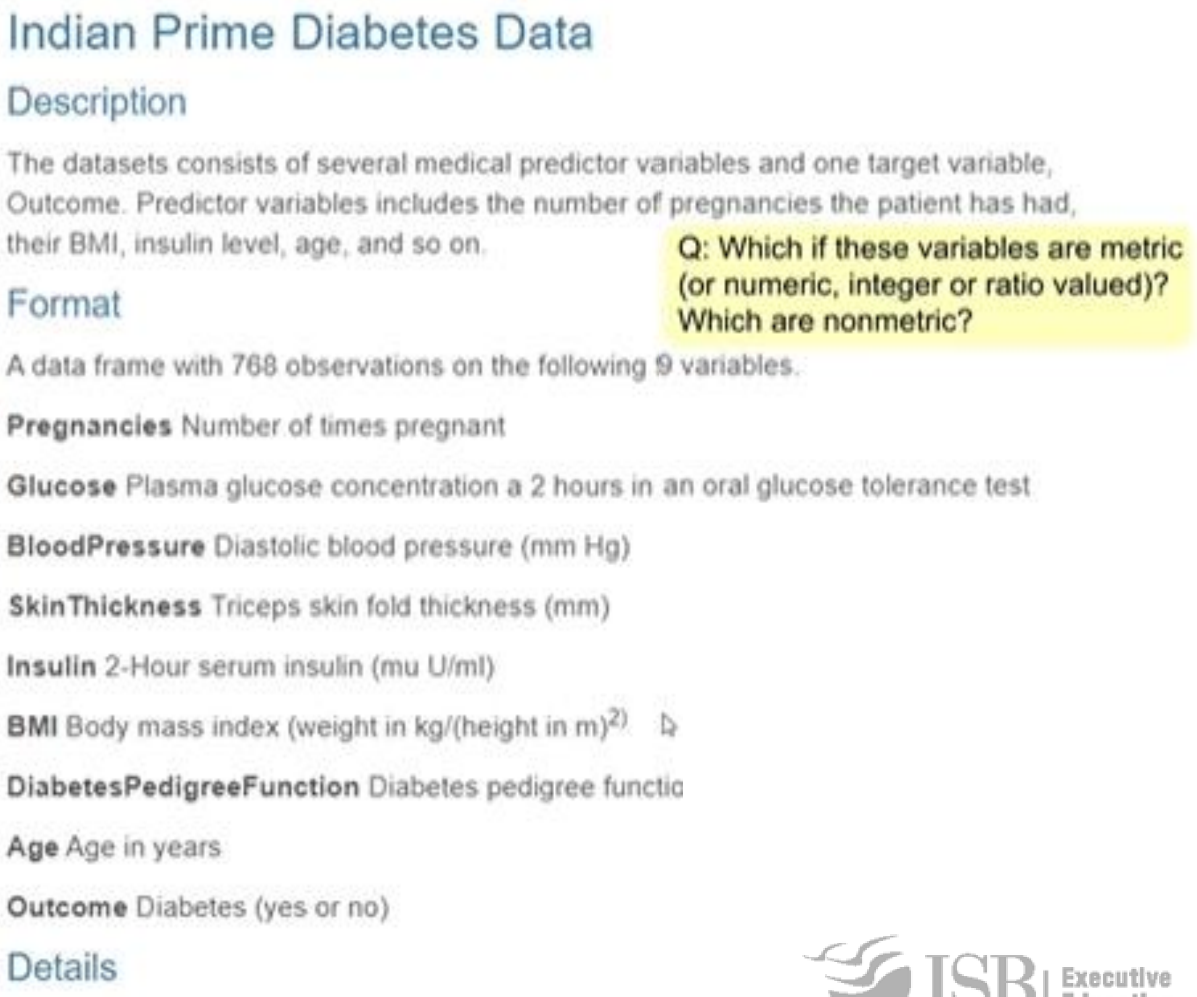
Load Example Data

diabetes

sleep
diabetes
mtcars

Choose File

Browse... No File Selected



Indian Prime Diabetes Data

Description

The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

Q: Which if these variables are metric (or numeric, integer or ratio valued)? Which are nonmetric?

Format

A data frame with 768 observations on the following 9 variables.

Pregnancies Number of times pregnant

Glucose Plasma glucose concentration a 2 hours in an oral glucose tolerance test

BloodPressure Diastolic blood pressure (mm Hg)

SkinThickness Triceps skin fold thickness (mm)

Insulin 2-Hour serum insulin (mu U/ml)

BMI Body mass index (weight in kg/(height in m)$^{2)}$

DiabetesPedigreeFunction Diabetes pedigree functio

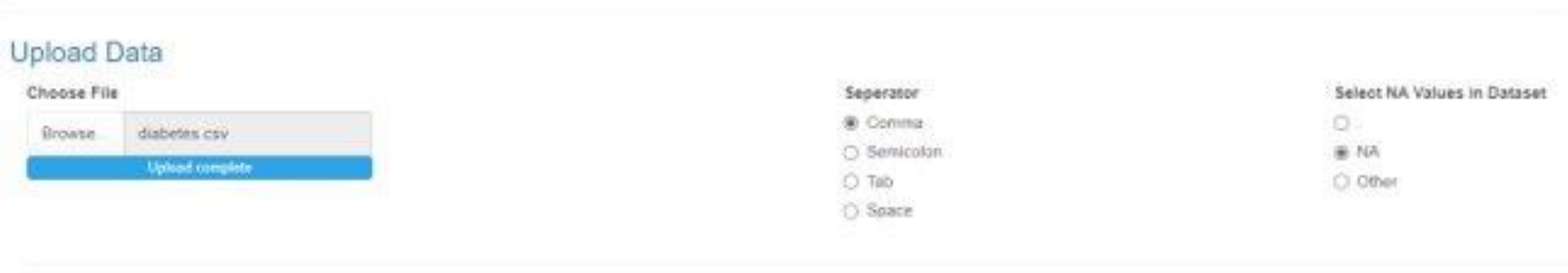
Age Age in years

Outcome Diabetes (yes or no)

Details

Data Walkthrough for the Data PreProc App

Now read-in the data using the file input field.



Upload Data

Choose File

Browse... diabetes.csv

Upload complete

Seperator

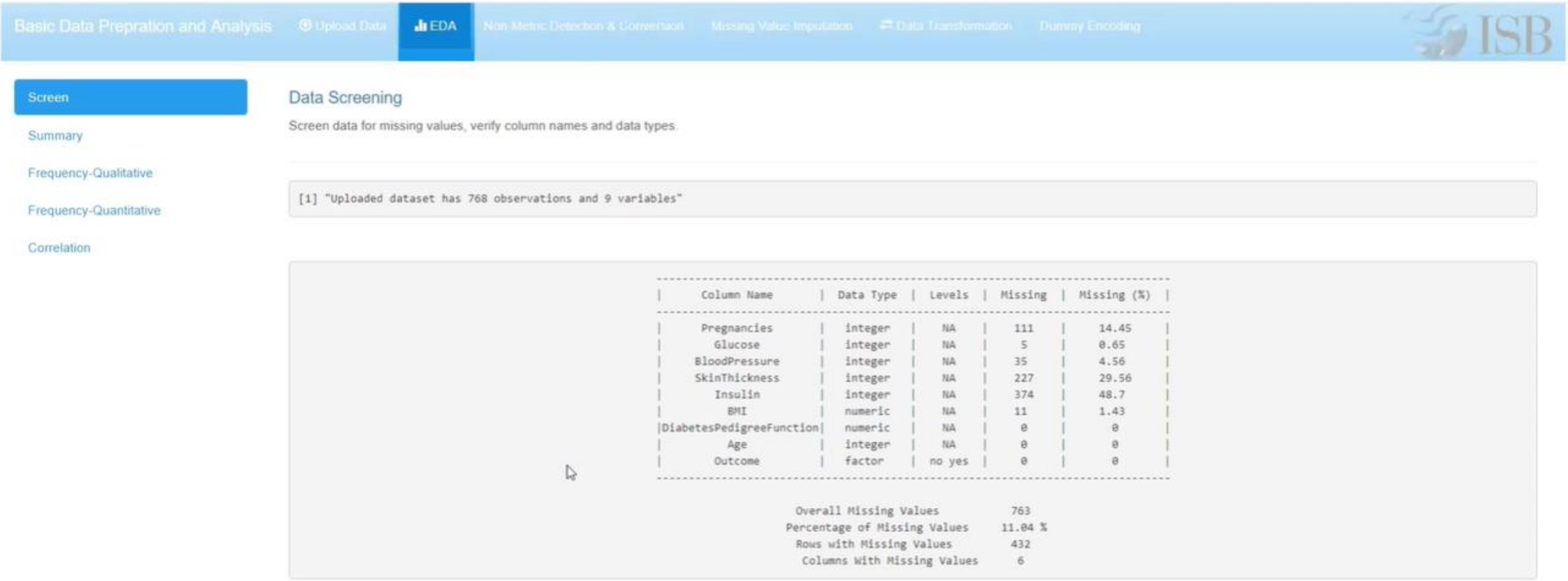
● Comma
○ Semicolon
○ Tab
○ Space

Select NA Values in Dataset

○ -
● NA
○ Other

Data Walkthrough for the Data PreProc App

Look for data issues and resolve them. Note that some data are missing, some variables are non-metric, etc.



Data Screening

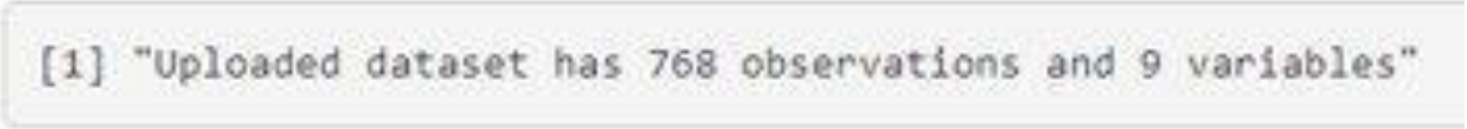
Screen data for missing values, verify column names and data types.

```
[1] "Uploaded dataset has 768 observations and 9 variables"
```

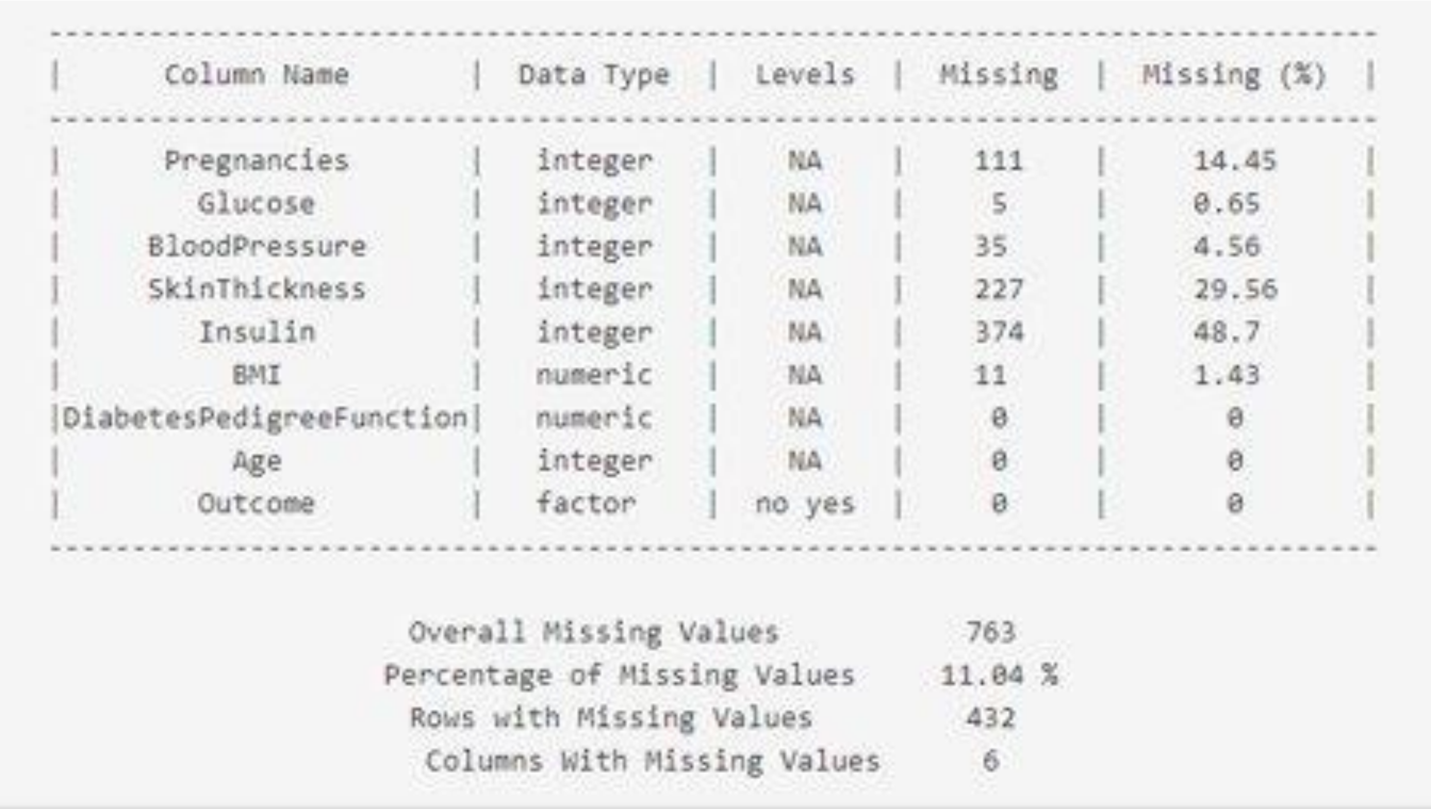
```
------------------------------------------------------------------------
|       Column Name        | Data Type | Levels | Missing | Missing (%) |
------------------------------------------------------------------------
|       Pregnancies        |  integer  |   NA   |   111   |    14.45    |
|         Glucose          |  integer  |   NA   |    5    |    0.65     |
|      BloodPressure       |  integer  |   NA   |   35    |    4.56     |
|      SkinThickness       |  integer  |   NA   |   227   |    29.56    |
|         Insulin          |  integer  |   NA   |   374   |    48.7     |
|           BMI            |  numeric  |   NA   |   11    |    1.43     |
|DiabetesPedigreeFunction|  numeric  |   NA   |    0    |      0      |
|           Age            |  integer  |   NA   |    0    |      0      |
|         Outcome          |  factor   | no yes |    0    |      0      |
------------------------------------------------------------------------

          Overall Missing Values          763
       Percentage of Missing Values       11.04 %
          Rows with Missing Values        432
         Columns With Missing Values       6
```

Exploring the EDA Output Tab: Screen the Data

- First, we *screen* the data for missing values and inconsistencies.
  - What is the **size** of the dataset?
  - Which variables have been identified as factor (i.e., nonmetric) versus metric?



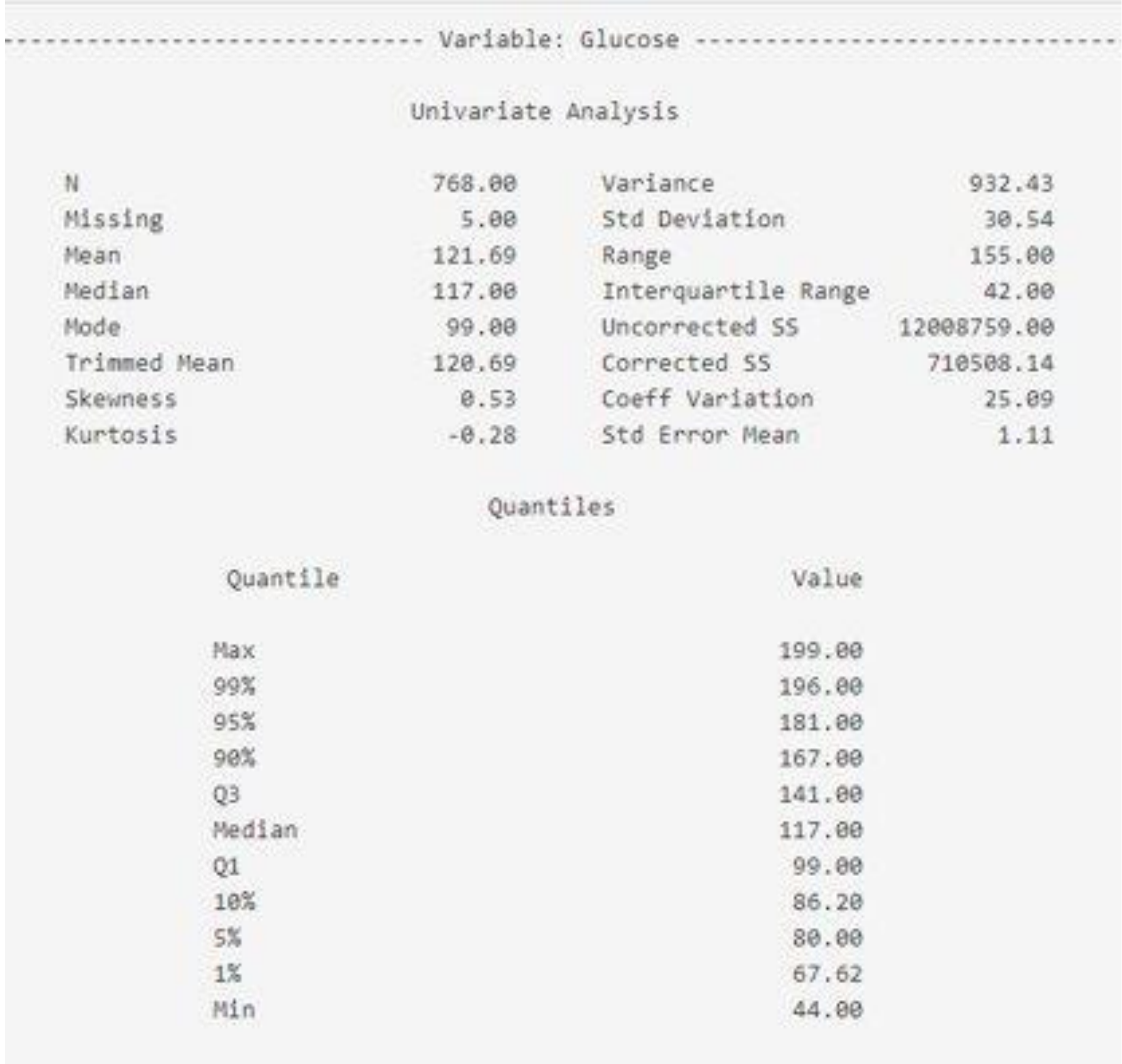
```
[1] "Uploaded dataset has 768 observations and 9 variables"
```



```
-------------------------------------------------------------------------------
|      Column Name        | Data Type  |  Levels  |  Missing  |  Missing (%)  |
-------------------------------------------------------------------------------
|      Pregnancies        |  integer   |    NA    |    111    |     14.45     |
|        Glucose          |  integer   |    NA    |     5     |     0.65      |
|     BloodPressure       |  integer   |    NA    |    35     |     4.56      |
|     SkinThickness       |  integer   |    NA    |    227    |     29.56     |
|        Insulin          |  integer   |    NA    |    374    |     48.7      |
|          BMI            |  numeric   |    NA    |    11     |     1.43      |
|DiabetesPedigreeFunction|  numeric   |    NA    |     0     |       0       |
|          Age            |  integer   |    NA    |     0     |       0       |
|        Outcome          |  factor    |  no yes  |     0     |       0       |
-------------------------------------------------------------------------------

                    Overall Missing Values          763
                  Percentage of Missing Values    11.04 %
                    Rows with Missing Values        432
                    Columns With Missing Values      6
```

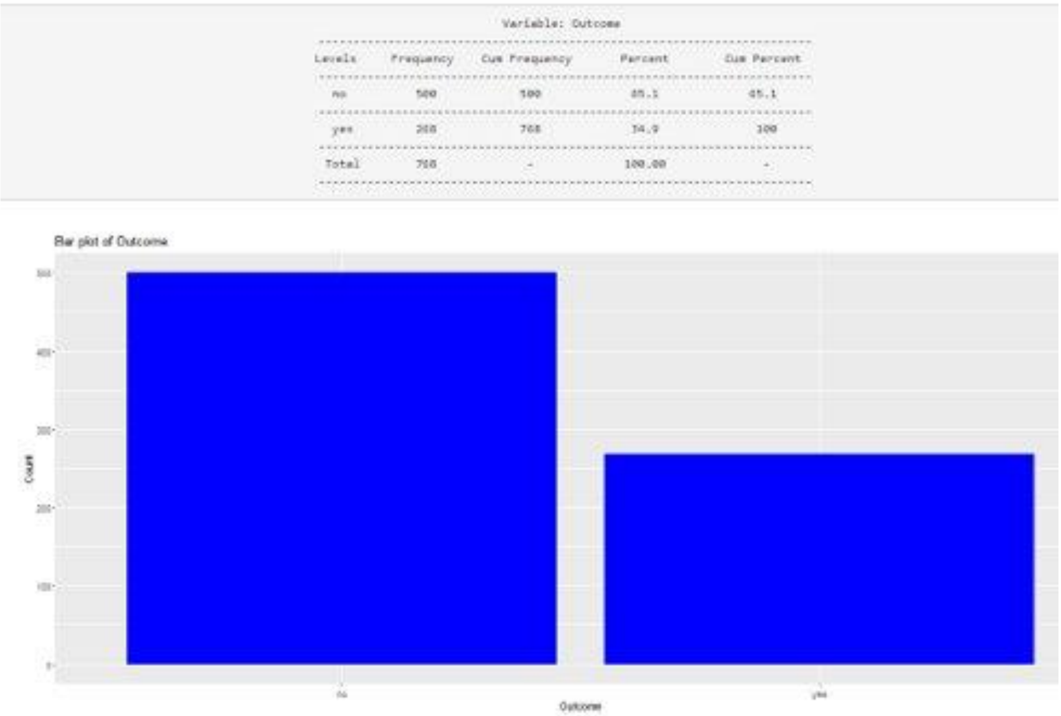
Exploring the EDA Output Tab: Summary Statistics

For a variable, say, 'Glucose', what is the mean, standard deviation and range?



```
---------------------------- Variable: Glucose ----------------------------

                          Univariate Analysis

 N                          768.00      Variance                   932.43
 Missing                      5.00      Std Deviation               30.54
 Mean                       121.69      Range                      155.00
 Median                     117.00      Interquartile Range         42.00
 Mode                        99.00      Uncorrected SS        12008759.00
 Trimmed Mean               120.69      Corrected SS            710508.14
 Skewness                     0.53      Coeff Variation             25.09
 Kurtosis                    -0.28      Std Error Mean               1.11

                               Quantiles

             Quantile                             Value

            Max                                  199.00
            99%                                  196.00
            95%                                  181.00
            90%                                  167.00
            Q3                                   141.00
            Median                               117.00
            Q1                                    99.00
            10%                                   86.20
            5%                                    80.00
            1%                                    67.62
            Min                                   44.00
```

The *Frequency-Qualitative* and *Quantitative* sub-tabs yield histograms of variable distributions.

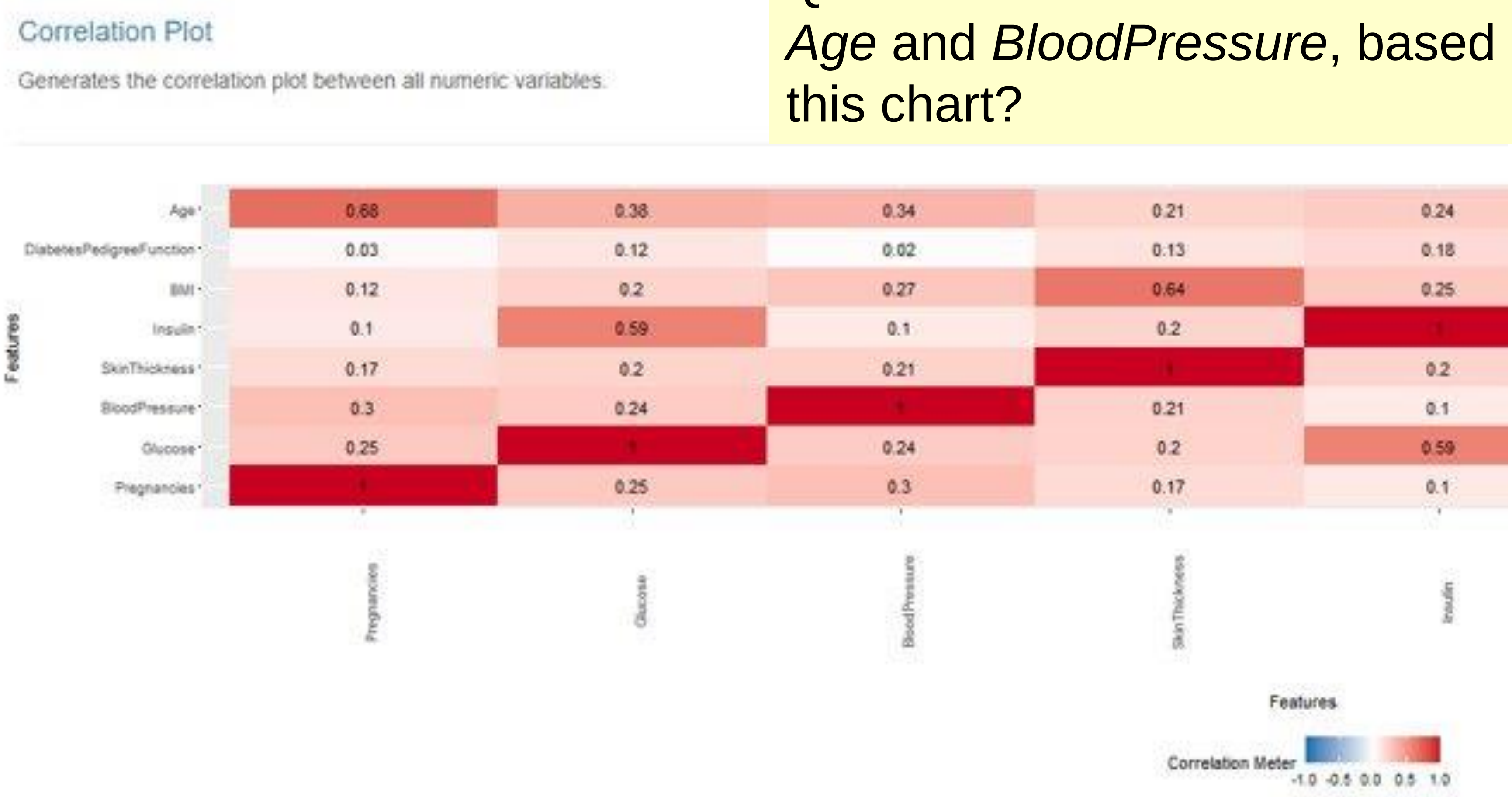


```
                         Variable: Outcome
---------------------------------------------------------------------
Levels   Frequency   Cum Frequency      Percent        Cum Percent
---------------------------------------------------------------------
  no        500           500             65.1             65.1
---------------------------------------------------------------------
  yes       268           768             34.9             100
---------------------------------------------------------------------
 Total      768            -             100.00             -
---------------------------------------------------------------------
```

Exploring the EDA Output Tab: Correlation

Finally, *correlation* shows us how the variables are inter-related.

Q: What is the correlation between *Age* and *BloodPressure*, based on this chart?



Correlation Plot

Generates the correlation plot between all numeric variables.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin
Age	0.68	0.38	0.34	0.21	0.24
DiabetesPedigreeFunction	0.03	0.12	0.02	0.13	0.18
BMI	0.12	0.2	0.27	0.64	0.25
Insulin	0.1	0.59	0.1	0.2	1
SkinThickness	0.17	0.2	0.21	1	0.2
BloodPressure	0.3	0.24	1	0.21	0.1
Glucose	0.25	1	0.24	0.2	0.59
Pregnancies	1	0.25	0.3	0.17	0.1

Non-Metric Detection and Conversion Tab

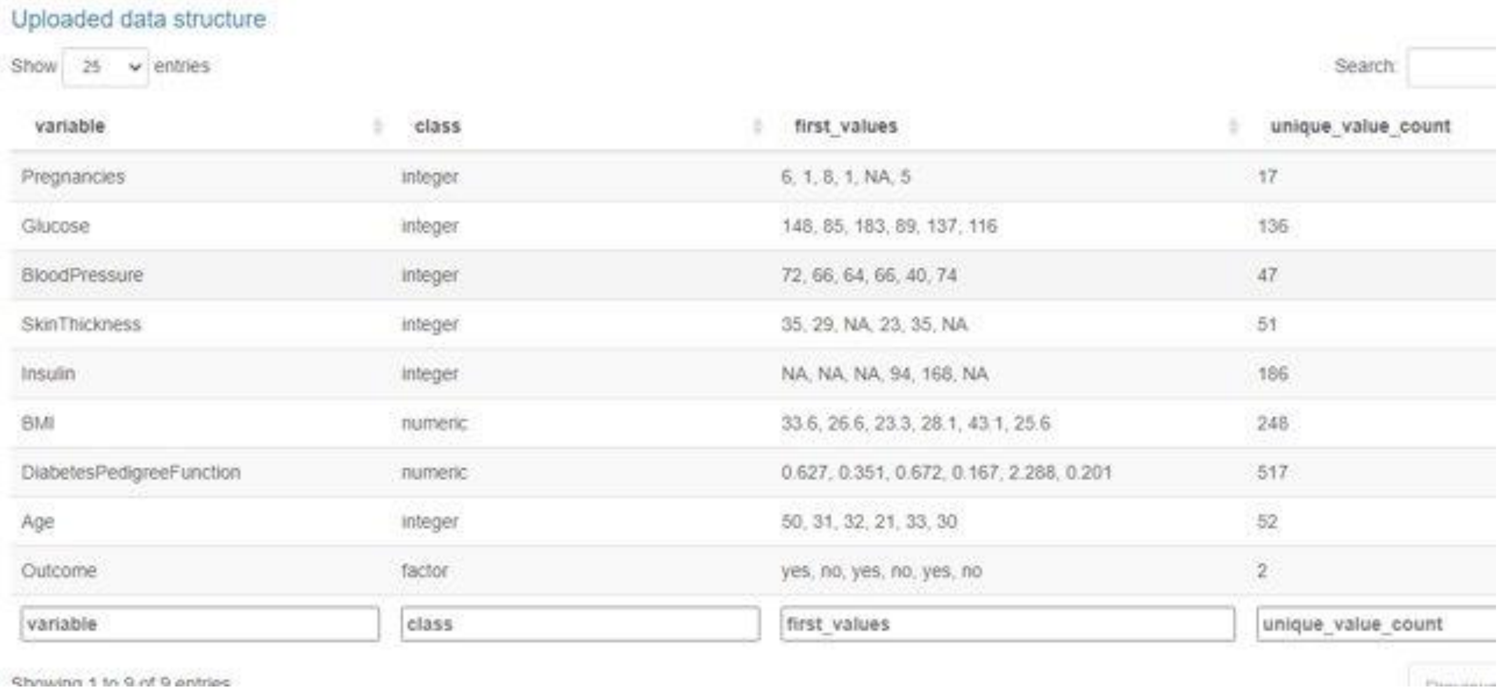
- Are there any nonmetric variables erroneously identified as metric?
  - E.g., Bus route number appears numeric but is actually categorical.

- Examine the '*unique_value_count*' for each variable. The smaller this is, more the possibility of nonmetric interpretation.
  - Which metric variable has the lowest number of unique counts?

- Say, we think the variable 'Pregnancies' is non-metric.
  - How then to convert this column's character to non-metric?

- We do *Select -> Convert -> Post-Conversion Structure*.

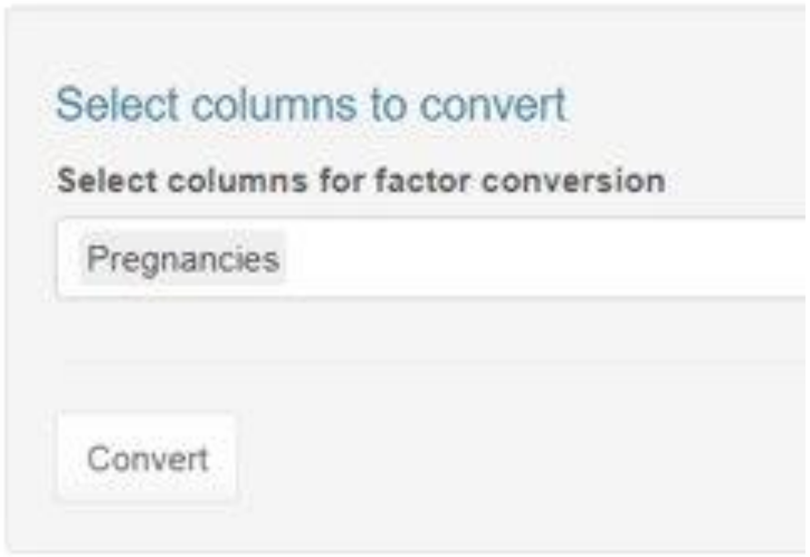
Non-Metric Detection and Conversion Tab



Uploaded data structure

variable	class	first_values	unique_value_count
Pregnancies	integer	6, 1, 8, 1, NA, 5	17
Glucose	integer	148, 85, 183, 89, 137, 116	136
BloodPressure	integer	72, 66, 64, 66, 40, 74	47
SkinThickness	integer	35, 29, NA, 23, 35, NA	51
Insulin	integer	NA, NA, NA, 94, 168, NA	186
BMI	numeric	33.6, 26.6, 23.3, 28.1, 43.1, 25.6	248
DiabetesPedigreeFunction	numeric	0.627, 0.351, 0.672, 0.167, 2.288, 0.201	517
Age	integer	50, 31, 32, 21, 33, 30	52
Outcome	factor	yes, no, yes, no, yes, no	2
variable	class	first_values	unique_value_count

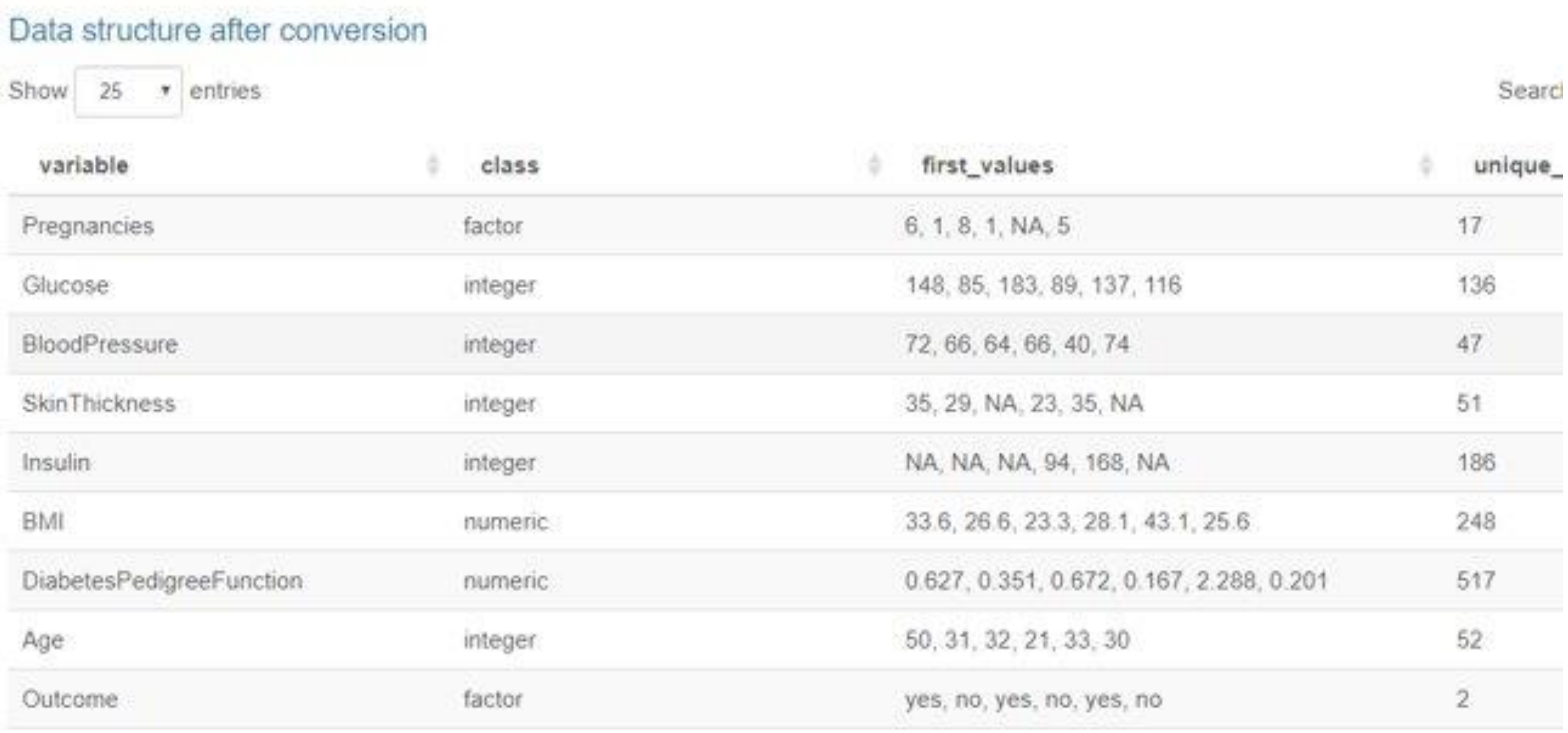
Showing 1 to 9 of 9 entries



Select columns to convert

Select columns for factor conversion: Pregnancies

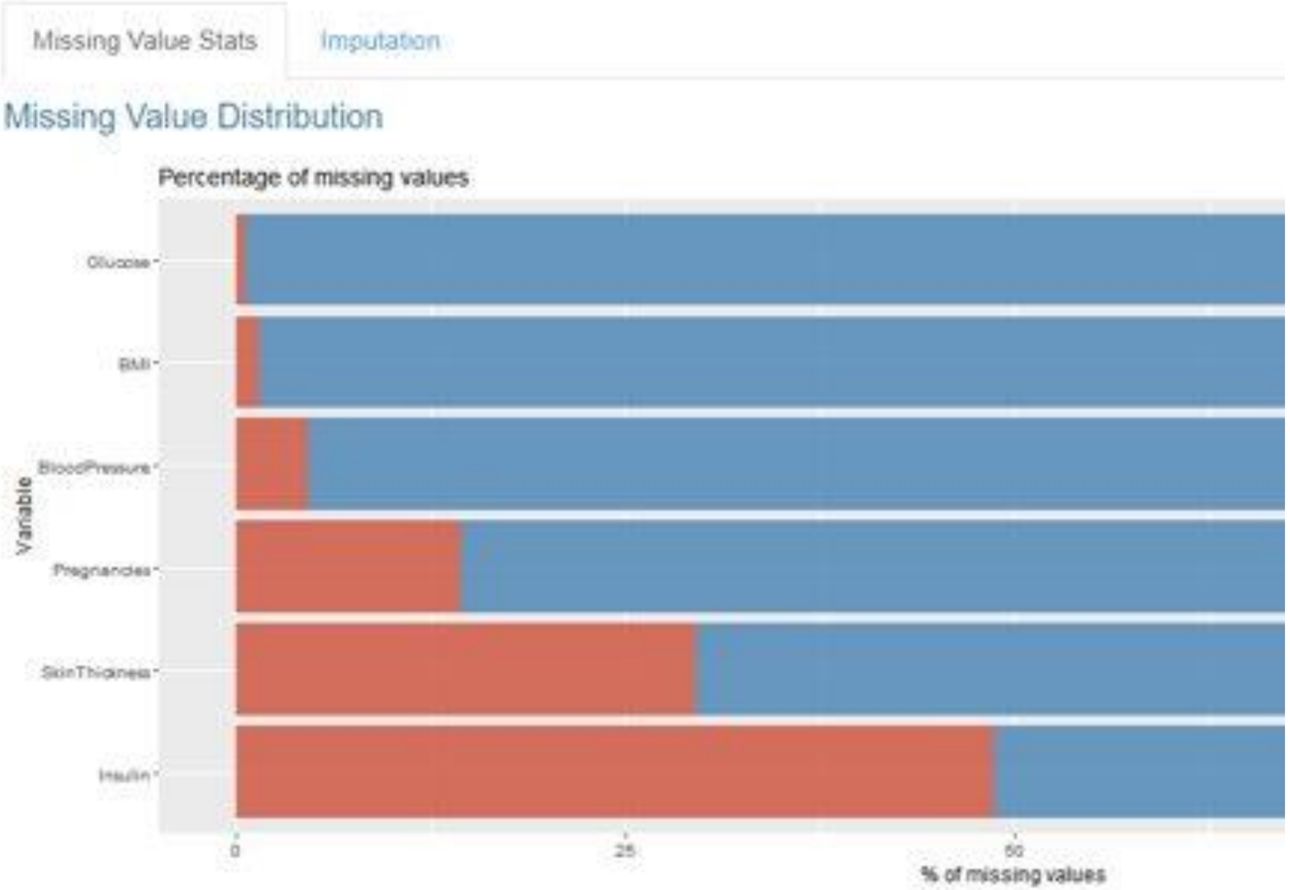
Non-Metric Detection and Conversion Tab

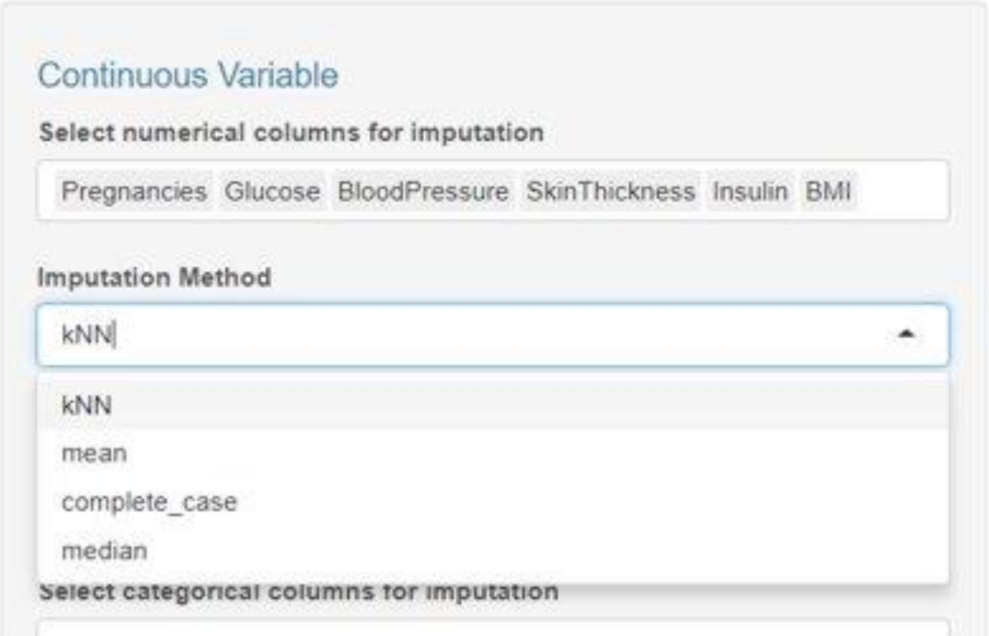


Data structure after conversion

variable	class	first_values	unique_
Pregnancies	factor	6, 1, 8, 1, NA, 5	17
Glucose	integer	148, 85, 183, 89, 137, 116	136
BloodPressure	integer	72, 66, 64, 66, 40, 74	47
SkinThickness	integer	35, 29, NA, 23, 35, NA	51
Insulin	integer	NA, NA, NA, 94, 168, NA	186
BMI	numeric	33.6, 26.6, 23.3, 28.1, 43.1, 25.6	248
DiabetesPedigreeFunction	numeric	0.627, 0.351, 0.672, 0.167, 2.288, 0.201	517
Age	integer	50, 31, 32, 21, 33, 30	52
Outcome	factor	yes, no, yes, no, yes, no	2

Exploring the Missing Value Imputation Tab





Continuous Variable

Select numerical columns for imputation

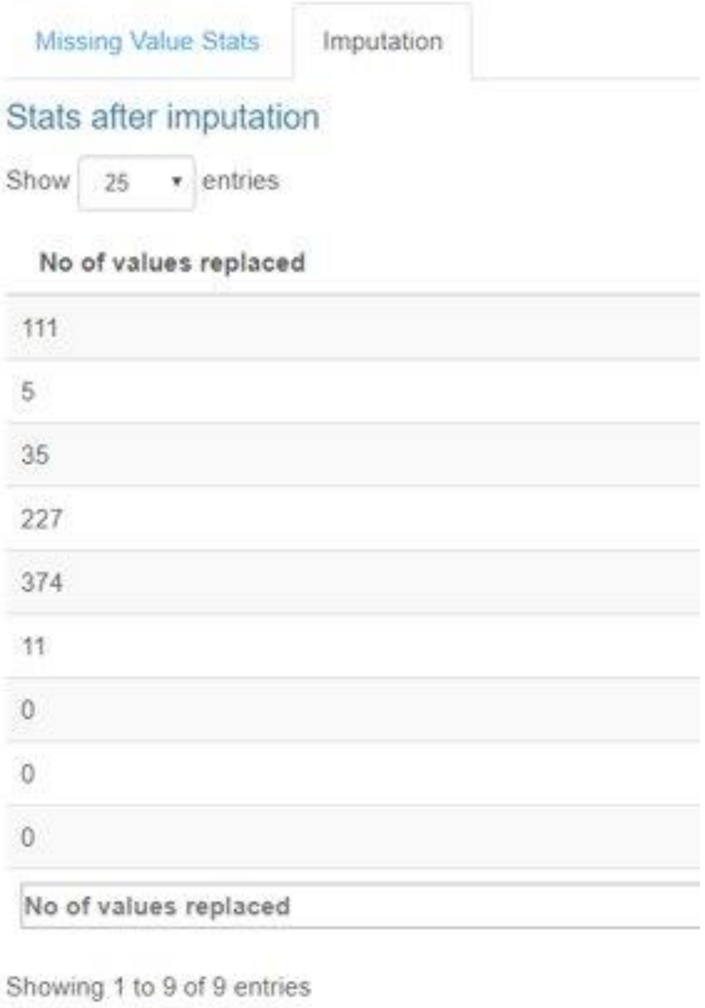
Pregnancies Glucose BloodPressure SkinThickness Insulin BMI

Imputation Method

kNN

Select categorical columns for imputation

Exploring the Missing Value Imputation Tab



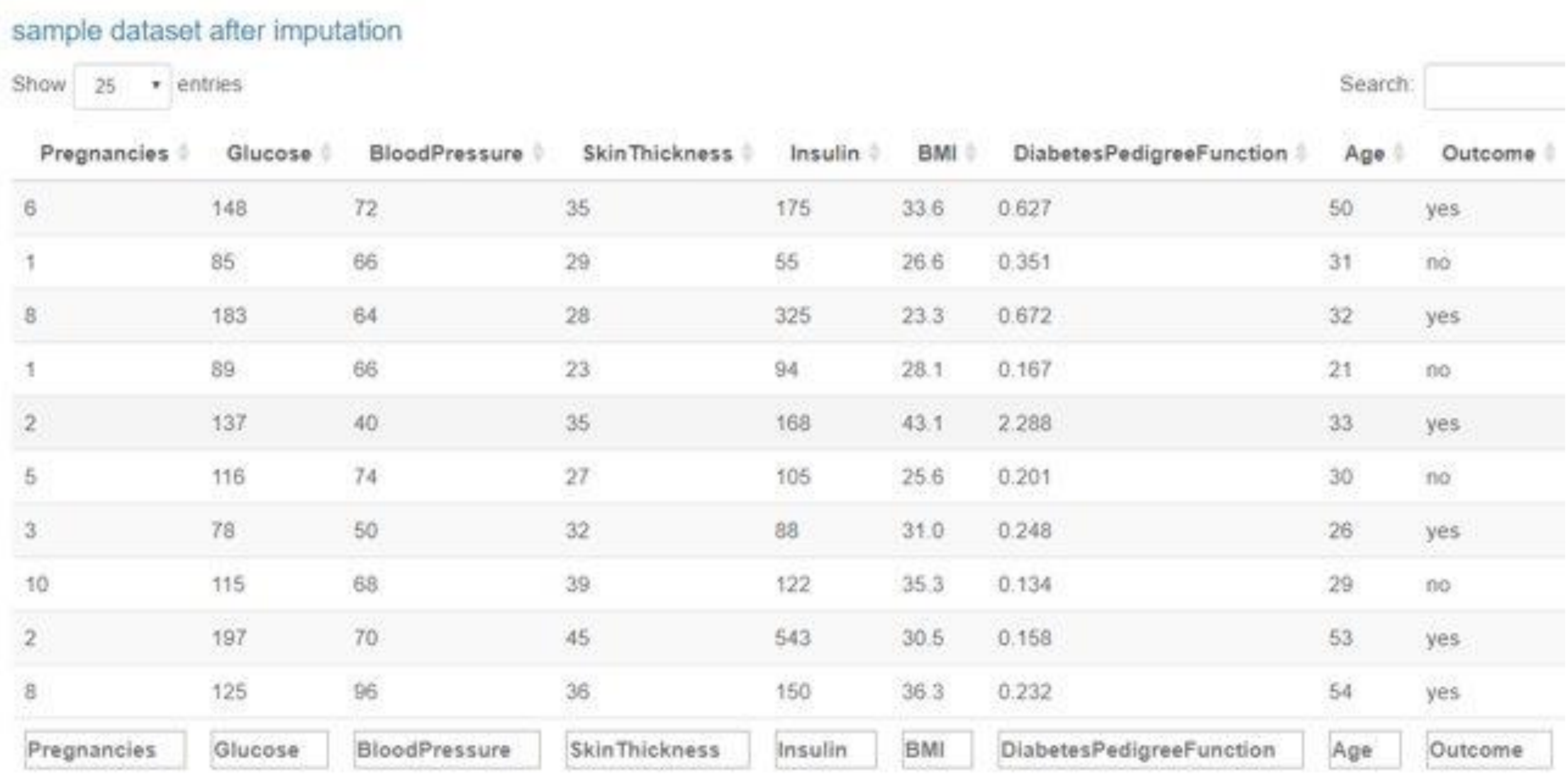
Stats after imputation

Show 25 entries

No of values replaced
111
5
35
227
374
11
0
0
0
No of values replaced

Showing 1 to 9 of 9 entries

Exploring the Missing Value Imputation Tab



sample dataset after imputation

Show 25 entries

Search:

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	175	33.6	0.627	50	yes
1	85	66	29	55	26.6	0.351	31	no
8	183	64	28	325	23.3	0.672	32	yes
1	89	66	23	94	28.1	0.167	21	no
2	137	40	35	168	43.1	2.288	33	yes
5	116	74	27	105	25.6	0.201	30	no
3	78	50	32	88	31.0	0.248	26	yes
10	115	68	39	122	35.3	0.134	29	no
2	197	70	45	543	30.5	0.158	53	yes
8	125	96	36	150	36.3	0.232	54	yes
Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome

Exploring the 'Data Transformation' Tab

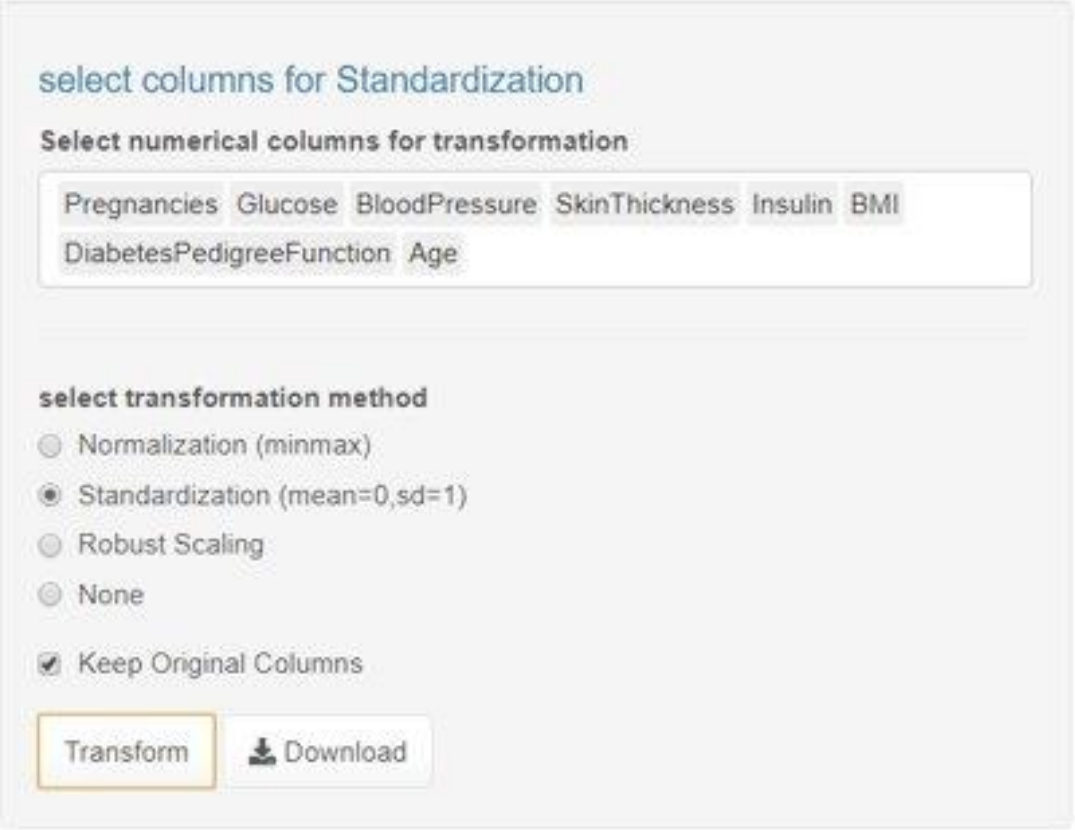
- Sometimes, we *transform* the **scale** of metric variables for better analysis and interpretation:

  - *Standardisation* brings all variables to the same scale (mean=0, std dev=1)
  - *Normalisation* brings all variables to within a [0,1] range (0=min value, 1=max value)

- We can transform the data as per our choice of scaling.

- We can then download the resulting transformed dataset for further analysis.

Exploring the 'Data Transformation' Tab



select columns for Standardization

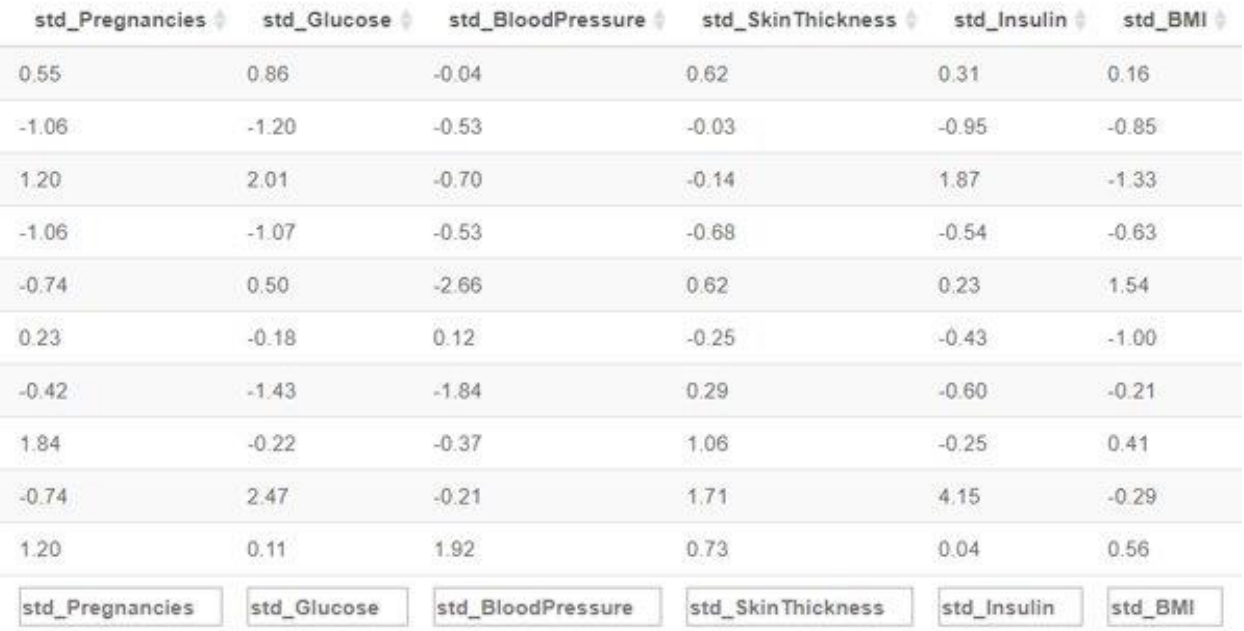
Select numerical columns for transformation

Pregnancies Glucose BloodPressure SkinThickness Insulin BMI DiabetesPedigreeFunction Age

select transformation method

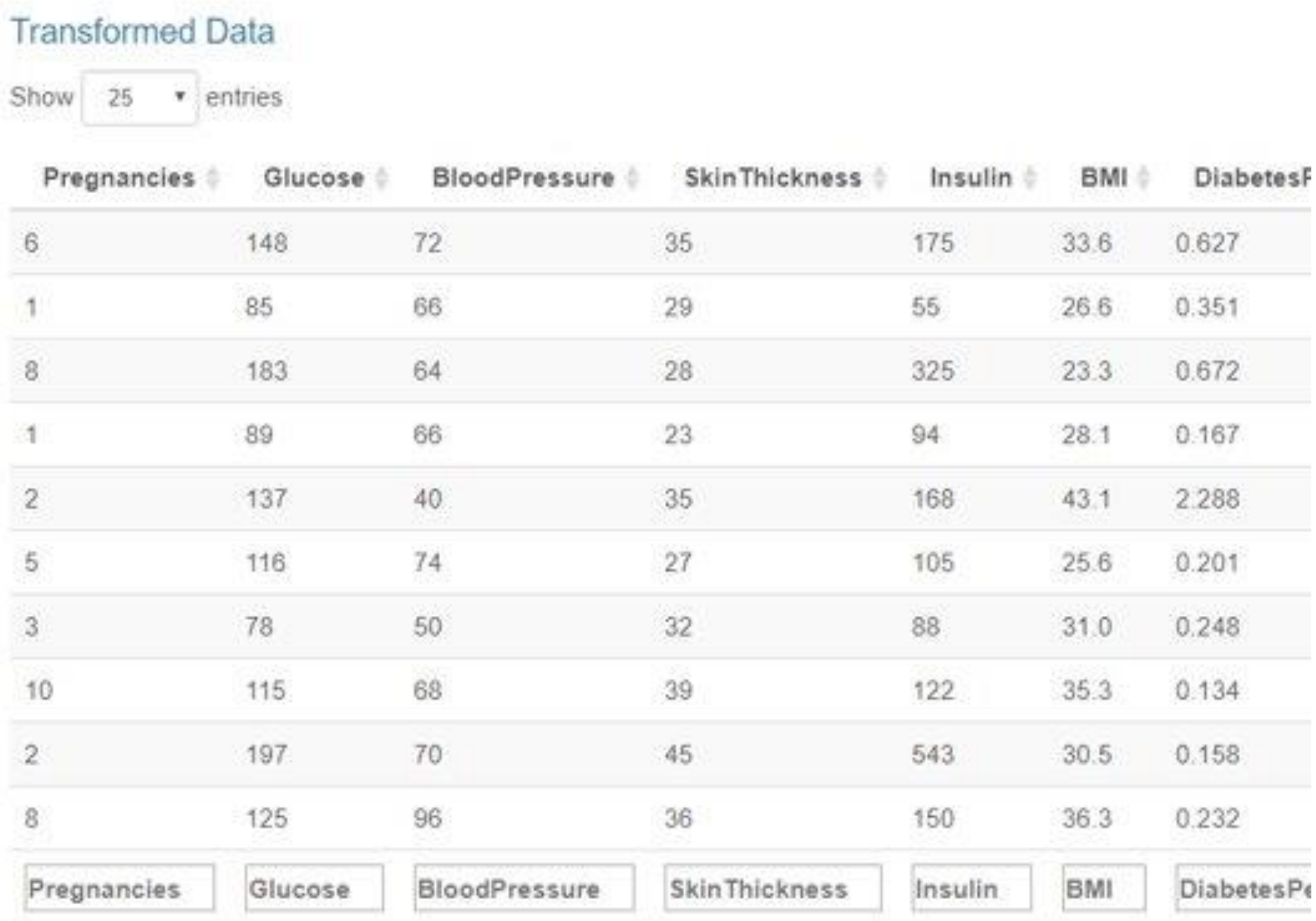
○ Normalization (minmax)
● Standardization (mean=0,sd=1)
○ Robust Scaling
○ None

☑ Keep Original Columns



std_Pregnancies	std_Glucose	std_BloodPressure	std_SkinThickness	std_Insulin	std_BMI
0.55	0.86	-0.04	0.62	0.31	0.16
-1.06	-1.20	-0.53	-0.03	-0.95	-0.85
1.20	2.01	-0.70	-0.14	1.87	-1.33
-1.06	-1.07	-0.53	-0.68	-0.54	-0.63
-0.74	0.50	-2.66	0.62	0.23	1.54
0.23	-0.18	0.12	-0.25	-0.43	-1.00
-0.42	-1.43	-1.84	0.29	-0.60	-0.21
1.84	-0.22	-0.37	1.06	-0.25	0.41
-0.74	2.47	-0.21	1.71	4.15	-0.29
1.20	0.11	1.92	0.73	0.04	0.56
std_Pregnancies	std_Glucose	std_BloodPressure	std_SkinThickness	std_Insulin	std_BMI

Exploring the 'Data Transformation' Tab



Transformed Data

Show 25 ▾ entries

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesP
6	148	72	35	175	33.6	0.627
1	85	66	29	55	26.6	0.351
8	183	64	28	325	23.3	0.672
1	89	66	23	94	28.1	0.167
2	137	40	35	168	43.1	2.288
5	116	74	27	105	25.6	0.201
3	78	50	32	88	31.0	0.248
10	115	68	39	122	35.3	0.134
2	197	70	45	543	30.5	0.158
8	125	96	36	150	36.3	0.232
Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPe

Exploring the 'Dummy Encoding' Tab

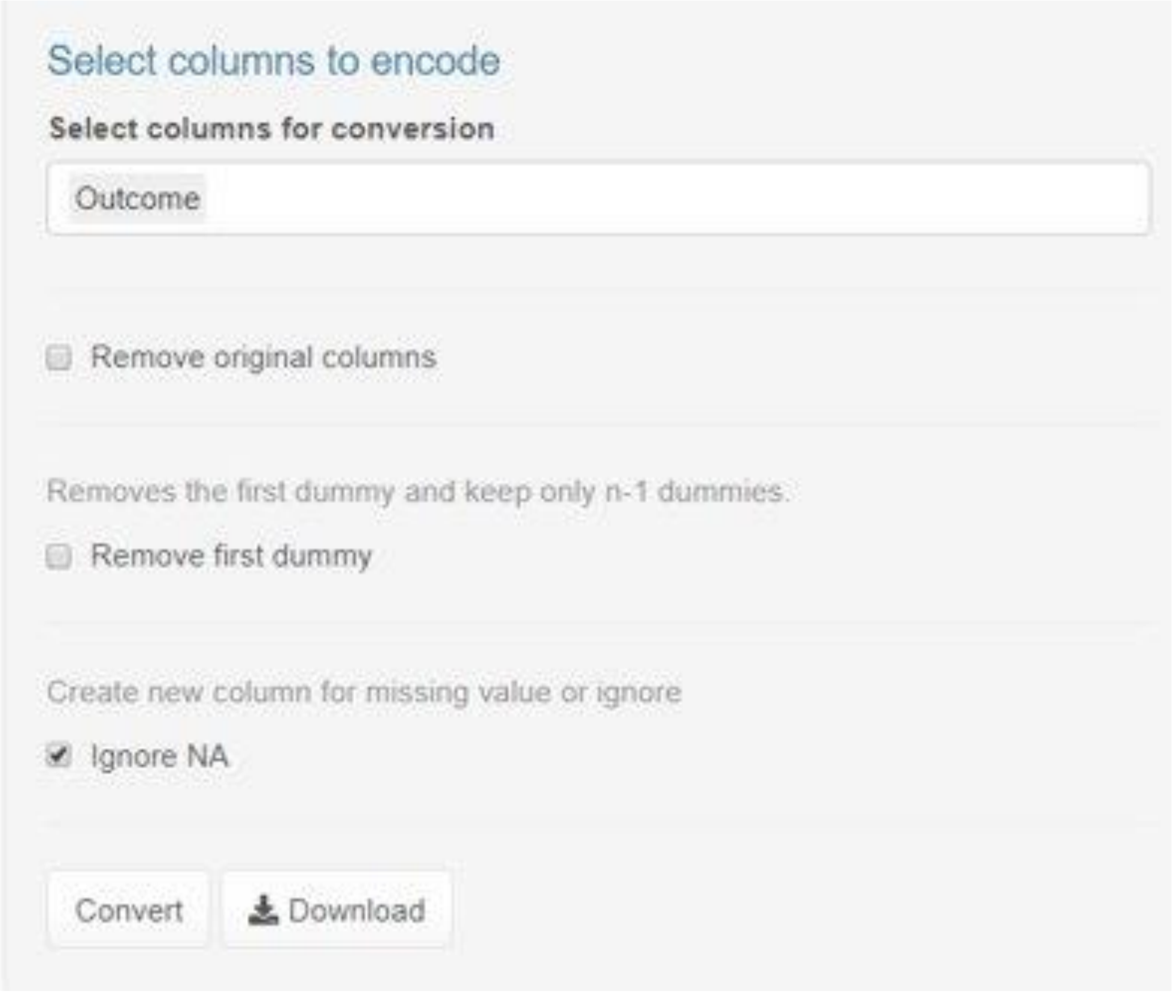
- Analysis of non-metric variables requires we create binary (or 1/0 valued) columns for each level of the variable.

- E.g., 'outcome' had 2 levels - 'yes' or 'no'

- We could create dummy (or one-hot-encoded) columns corresponding to 'outcome’.

- This transformed dataset can now be downloaded for further analysis.

Exploring the 'Dummy Encoding' Tab



Select columns to encode

Select columns for conversion

Outcome

□ Remove original columns

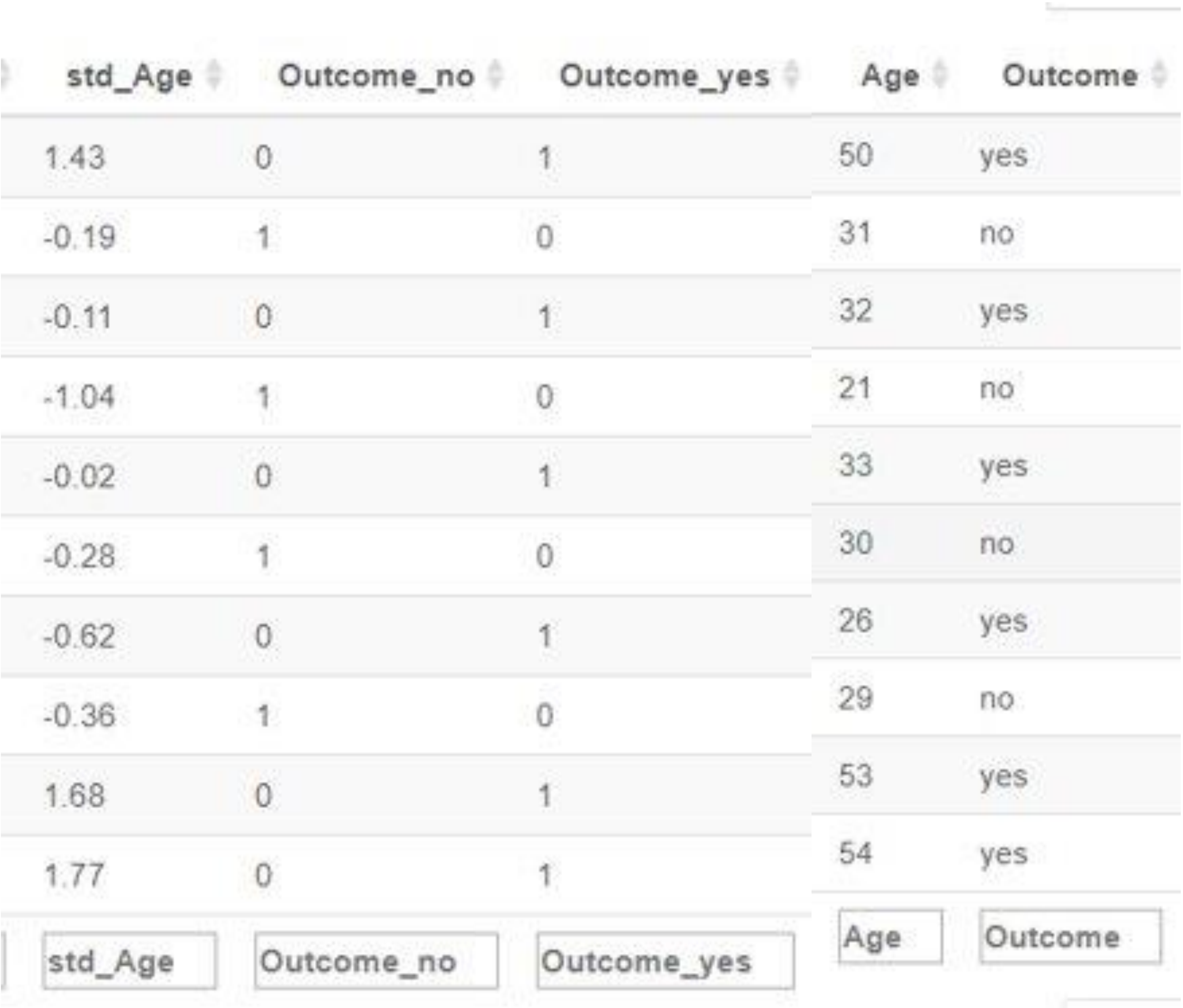
Removes the first dummy and keep only n-1 dummies.

□ Remove first dummy

Create new column for missing value or ignore

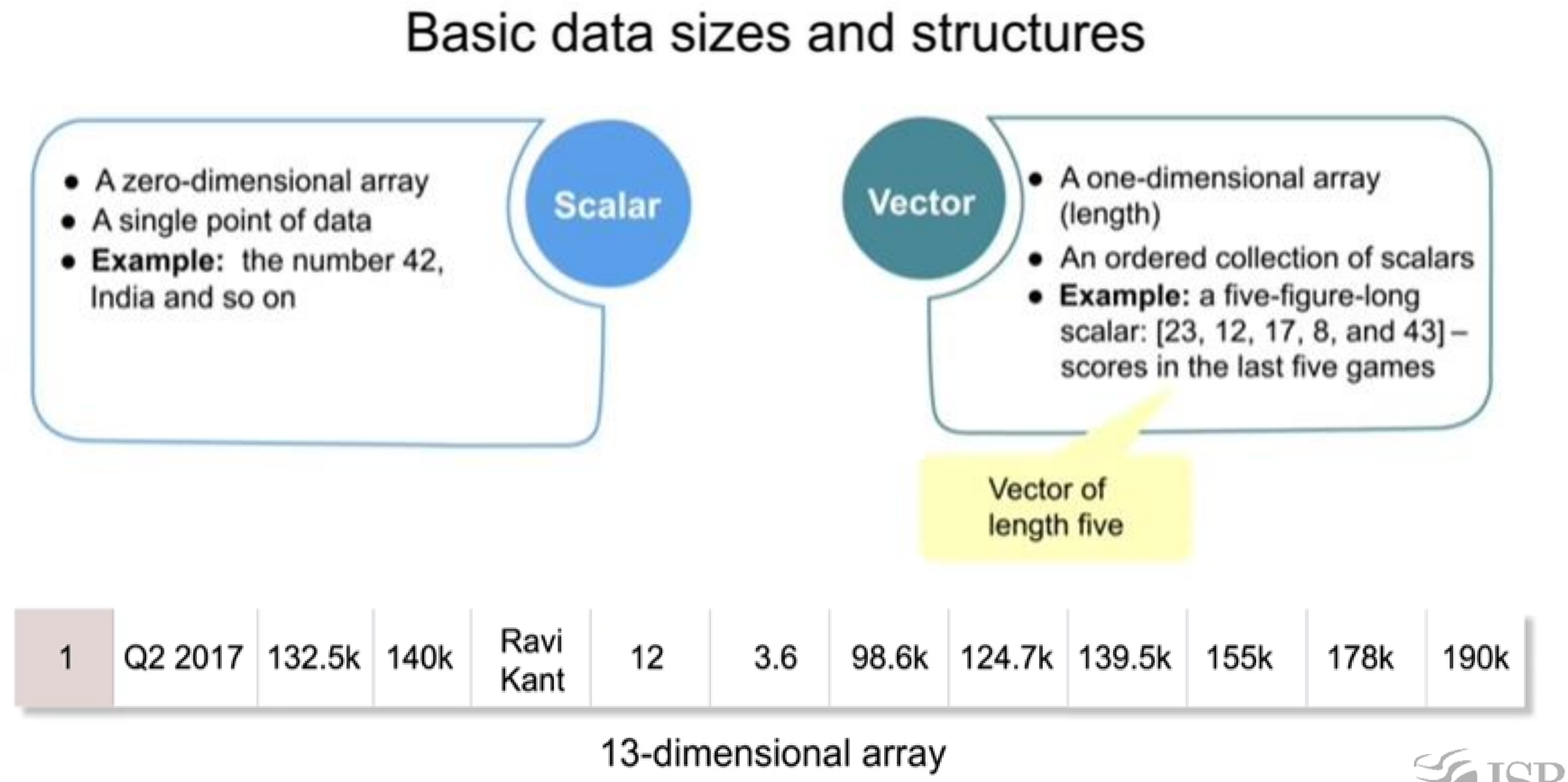
☑ Ignore NA

Exploring the 'Dummy Encoding' Tab



std_Age	Outcome_no	Outcome_yes	Age	Outcome
1.43	0	1	50	yes
-0.19	1	0	31	no
-0.11	0	1	32	yes
-1.04	1	0	21	no
-0.02	0	1	33	yes
-0.28	1	0	30	no
-0.62	0	1	26	yes
-0.36	1	0	29	no
1.68	0	1	53	yes
1.77	0	1	54	yes
std_Age	Outcome_no	Outcome_yes	Age	Outcome

Basic Data Algebra



Basic data sizes and structures

Scalar

- A zero-dimensional array
- A single point of data
- **Example:** the number 42, India and so on

Vector

- A one-dimensional array (length)
- An ordered collection of scalars
- **Example:** a five-figure-long scalar: [23, 12, 17, 8, and 43] – scores in the last five games

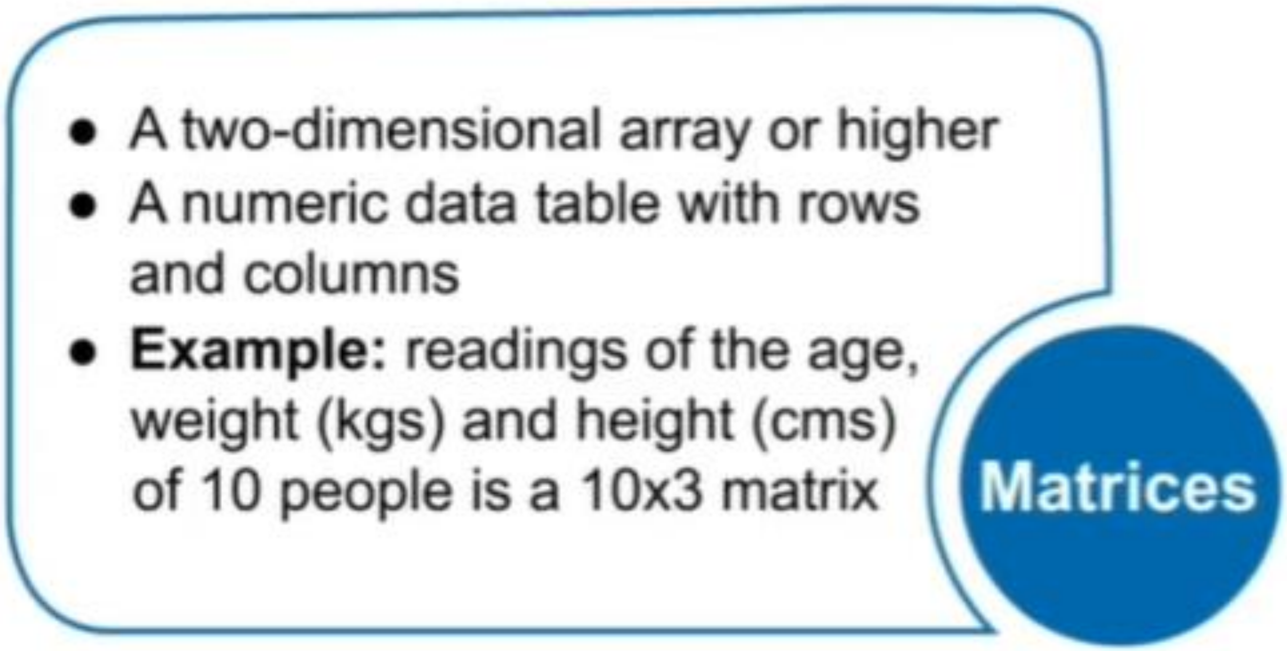
Vector of length five

1	Q2 2017	132.5k	140k	Ravi Kant	12	3.6	98.6k	124.7k	139.5k	155k	178k	190k

13-dimensional array

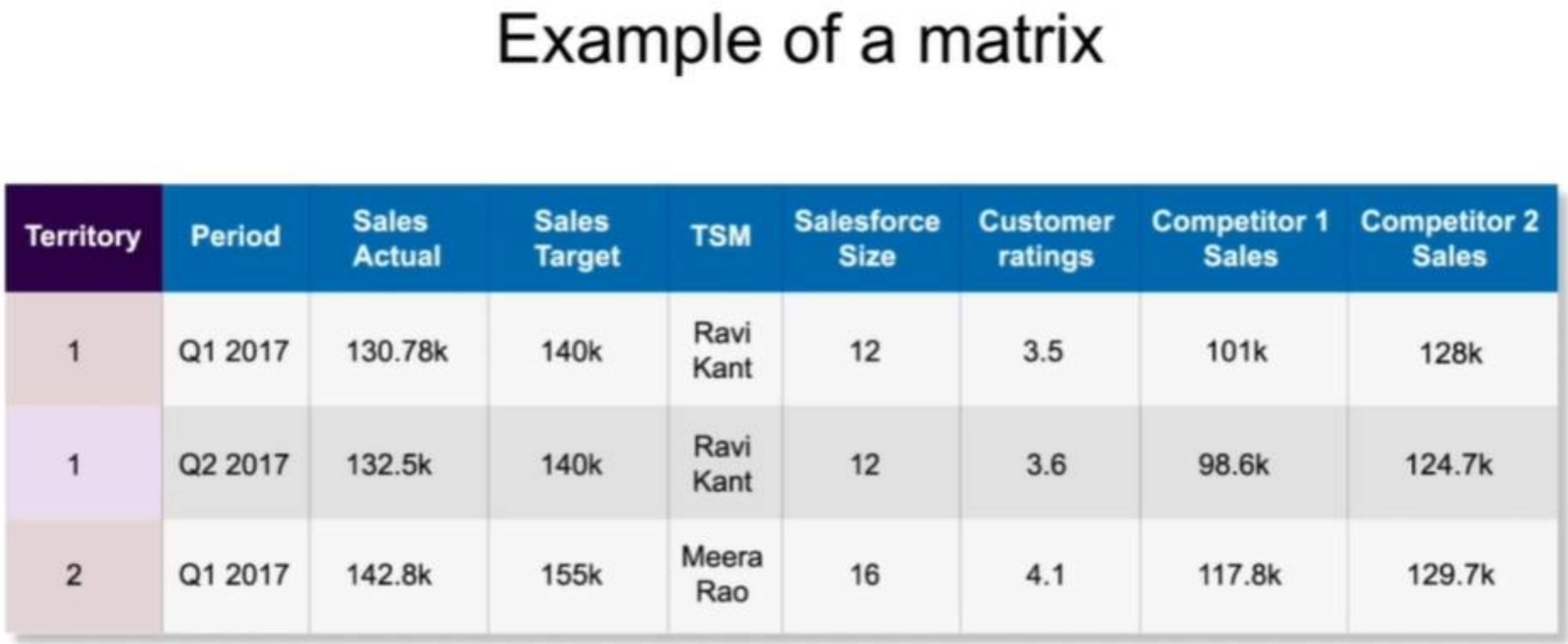
Basic Data Algebra

Basic data sizes and structures



- A two-dimensional array or higher
- A numeric data table with rows and columns
- **Example:** readings of the age, weight (kgs) and height (cms) of 10 people is a 10x3 matrix

Basic Data Algebra



Example of a matrix

Territory	Period	Sales Actual	Sales Target	TSM	Salesforce Size	Customer ratings	Competitor 1 Sales	Competitor 2 Sales
1	Q1 2017	130.78k	140k	Ravi Kant	12	3.5	101k	128k
1	Q2 2017	132.5k	140k	Ravi Kant	12	3.6	98.6k	124.7k
2	Q1 2017	142.8k	155k	Meera Rao	16	4.1	117.8k	129.7k

Example: Metric vs Non-metric

Basic data sizes and structures

Tensors

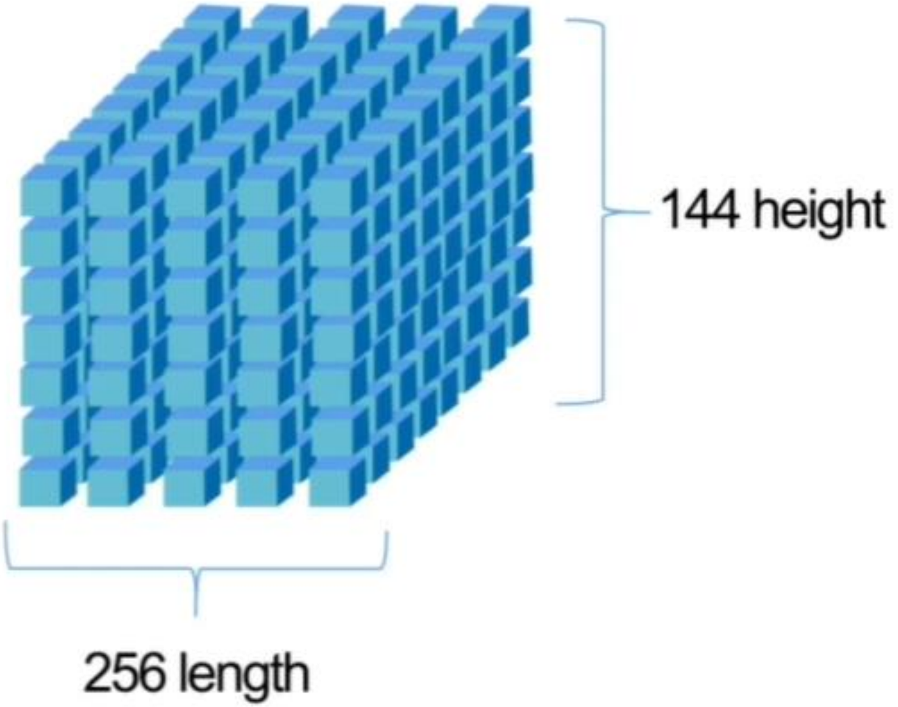
- A three-dimensional array or higher

Tensor size: 144 x 256 x 3 ←

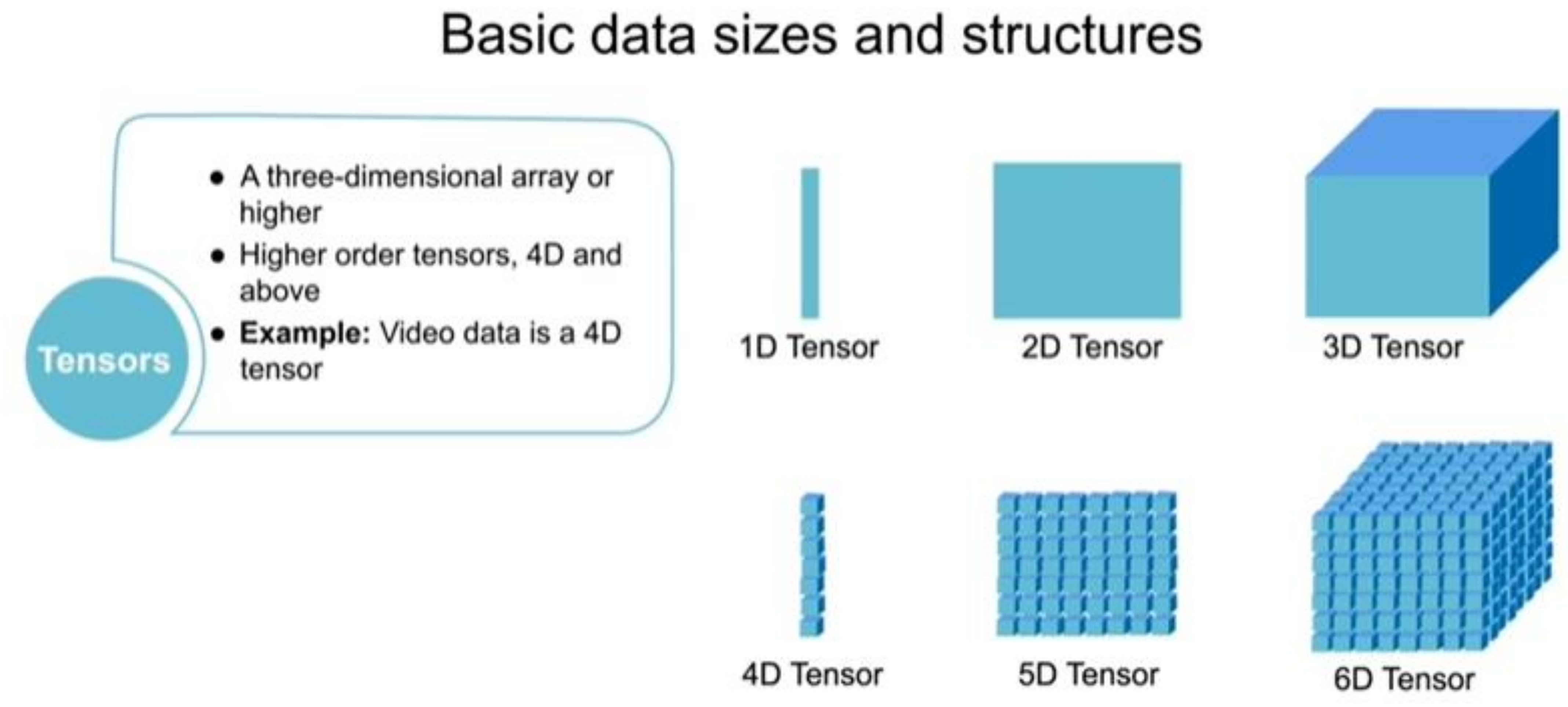
Image data includes height, breadth and colour channels

Image data

Three primary colours



Example: Metric vs Non-metric

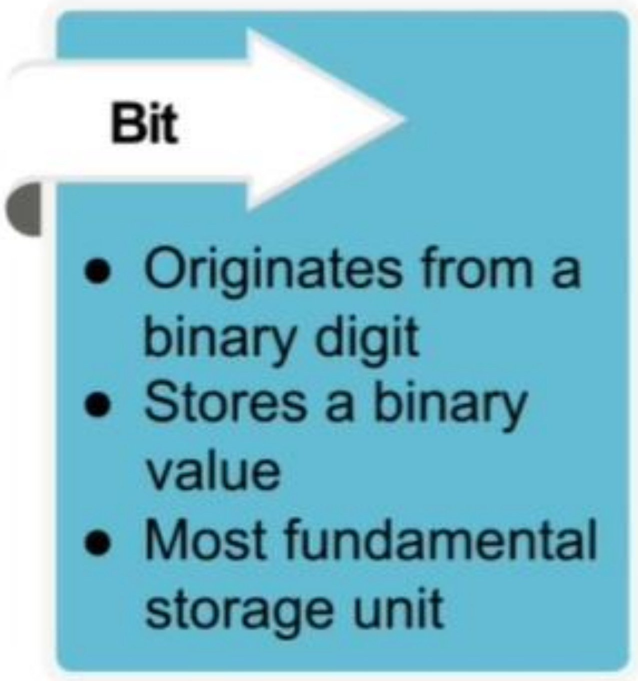


Basic data sizes and structures

Tensors

- A three-dimensional array or higher
- Higher order tensors, 4D and above
- **Example:** Video data is a 4D tensor

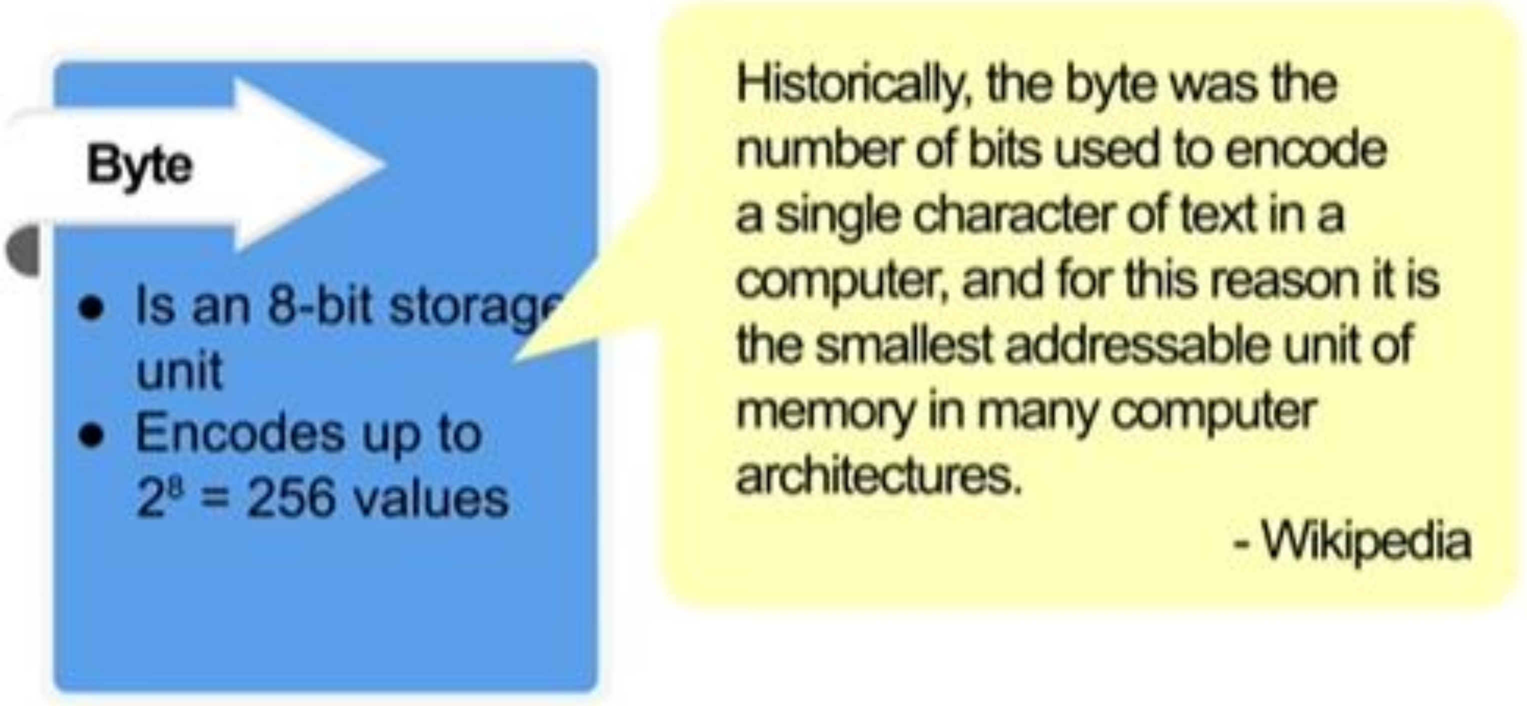
Data Storages and Sizes



Bit

- Originates from a binary digit
- Stores a binary value
- Most fundamental storage unit

Data Storages and Sizes



Byte

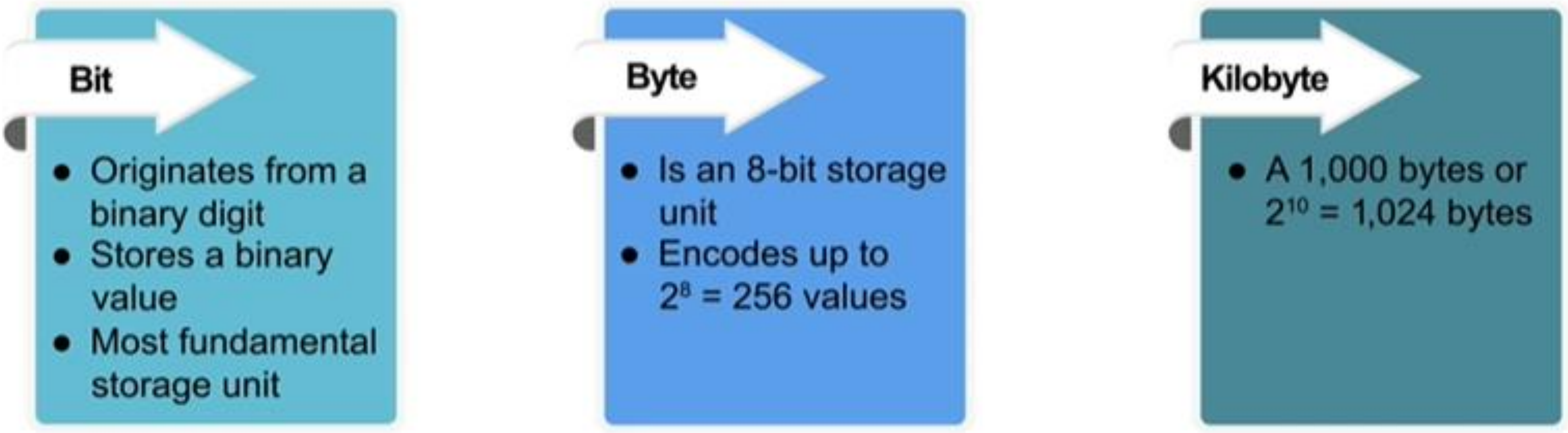
- Is an 8-bit storage unit
- Encodes up to $2^8 = 256$ values

Historically, the byte was the number of bits used to encode a single character of text in a computer, and for this reason it is the smallest addressable unit of memory in many computer architectures.

- Wikipedia

1	0	1	0	0	1	0	1

Data Storages and Sizes



Bit

- Originates from a binary digit
- Stores a binary value
- Most fundamental storage unit

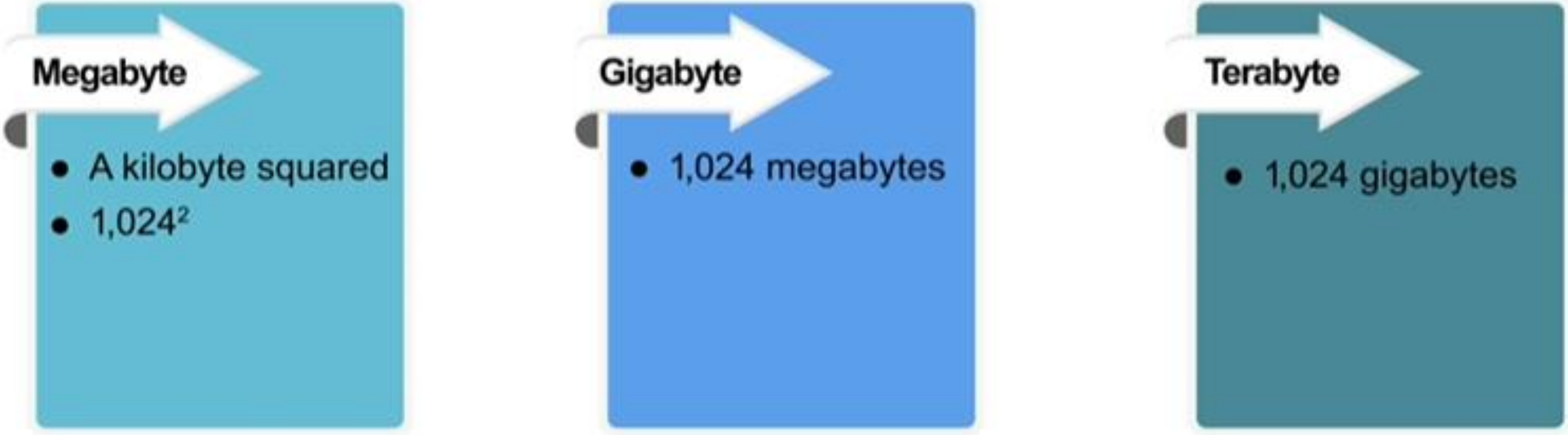
Byte

- Is an 8-bit storage unit
- Encodes up to $2^8 = 256$ values

Kilobyte

- A 1,000 bytes or $2^{10} = 1,024$ bytes

Data Storages and Sizes



Megabyte

- A kilobyte squared
- $1{,}024^2$

Gigabyte

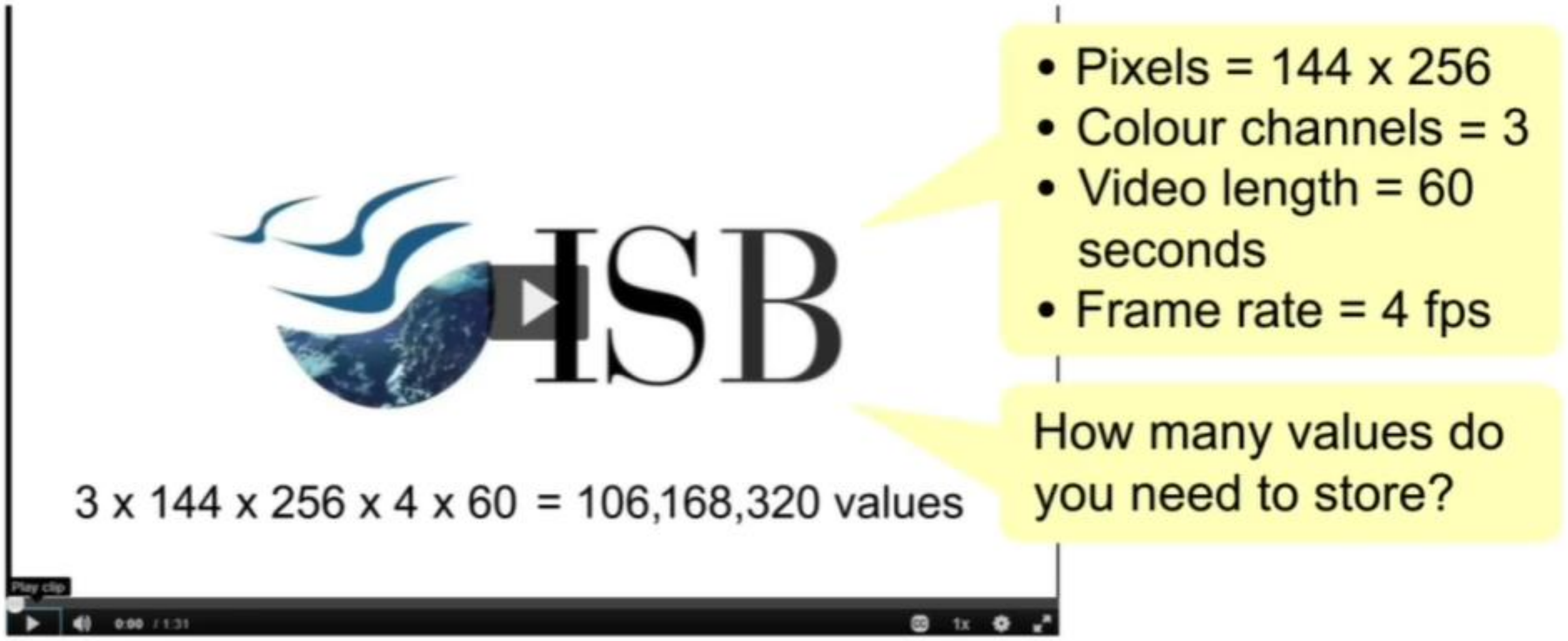
- 1,024 megabytes

Terabyte

- 1,024 gigabytes

Size and Storage Space Occupied by a 4D Tensor

Example: a still frame in a YouTube video



- Pixels = 144 x 256
- Colour channels = 3
- Video length = 60 seconds
- Frame rate = 4 fps

How many values do you need to store?

3 x 144 x 256 x 4 x 60 = 106,168,320 values

A 32-bit float per value yields 101.25 MB of storage (without compression)

Basic Data Structures, Data Storage and Sizes: Summary

Two ways of looking at data

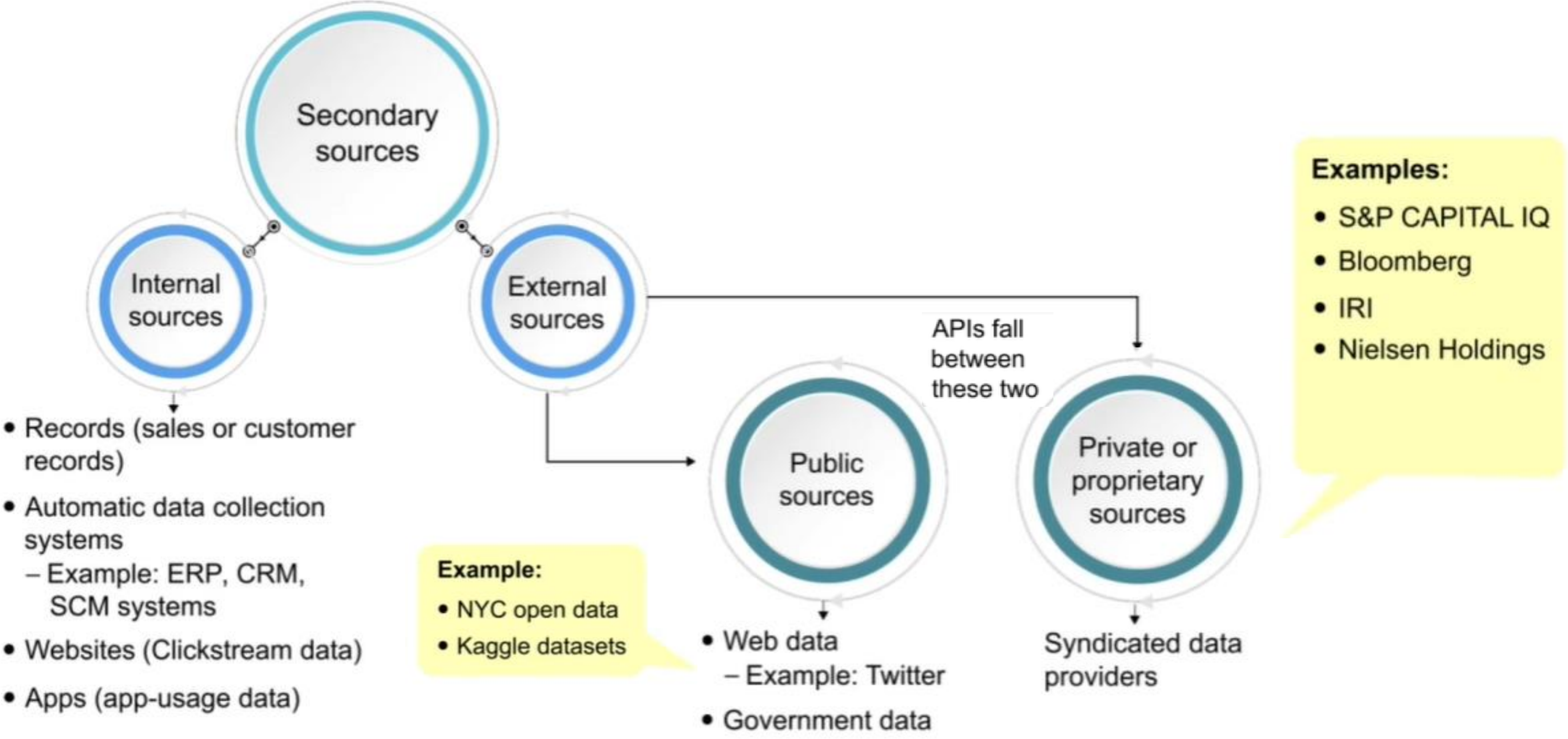
- Software

  - Deals with data structures (vectors, matrices and tensors)

- Hardware

  - Deals with data storage and sizes

Secondary Data Sources for Businesses



Secondary sources

Internal sources

- Records (sales or customer records)
- Automatic data collection systems
  - Example: ERP, CRM, SCM systems
- Websites (Clickstream data)
- Apps (app-usage data)

External sources

APIs fall between these two

Public sources

- Web data
  - Example: Twitter
- Government data

Example:

- NYC open data
- Kaggle datasets

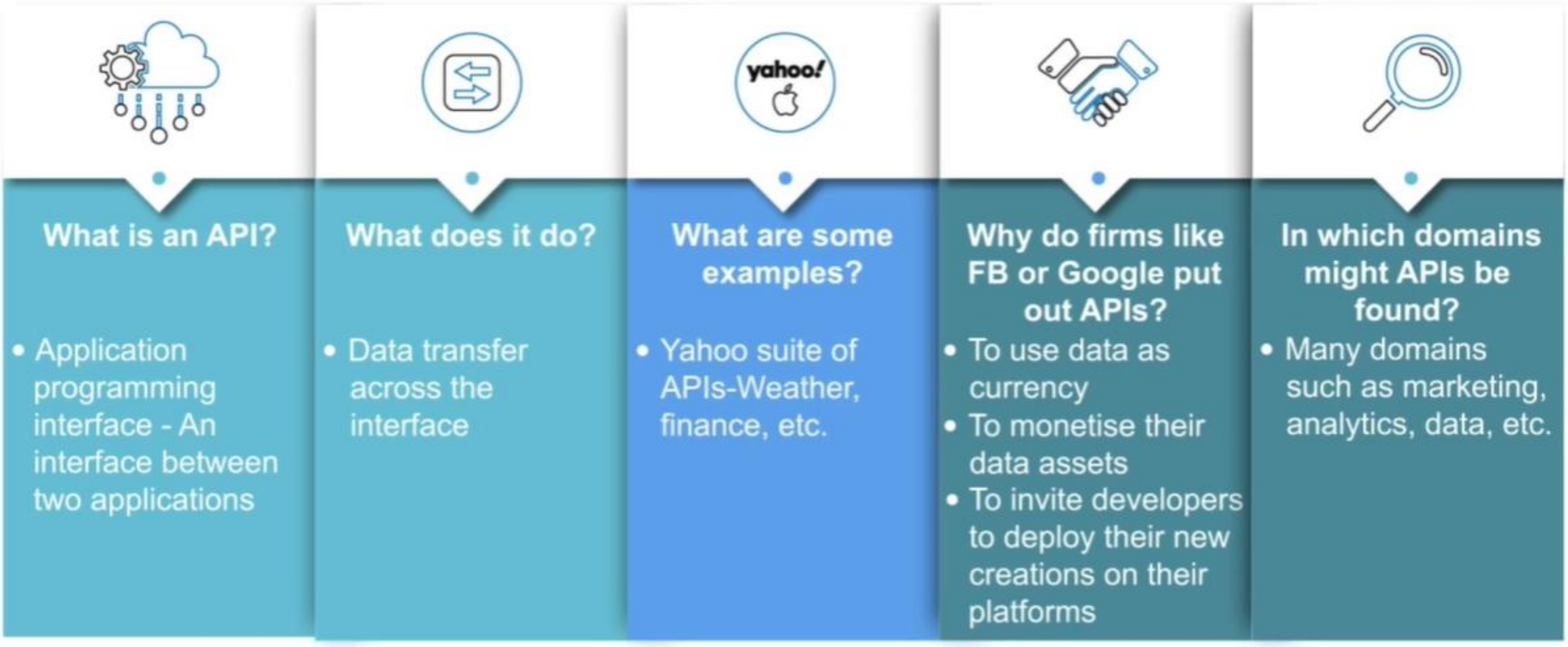
Private or proprietary sources

Syndicated data providers

Examples:

- S&P CAPITAL IQ
- Bloomberg
- IRI
- Nielsen Holdings

APIs



What is an API?

- Application programming interface - An interface between two applications

What does it do?

- Data transfer across the interface

What are some examples?

- Yahoo suite of APIs-Weather, finance, etc.

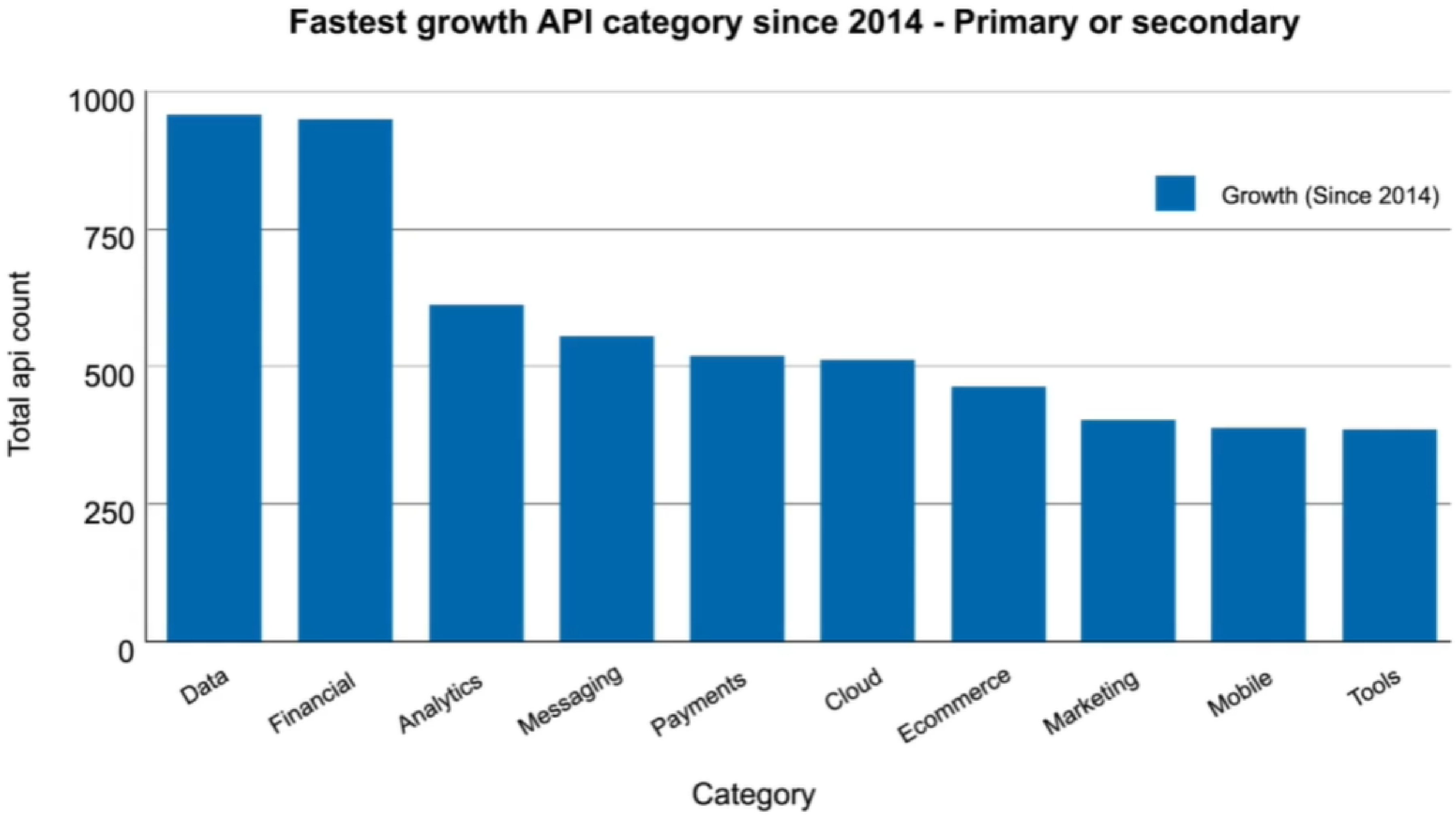
Why do firms like FB or Google put out APIs?

- To use data as currency
- To monetise their data assets
- To invite developers to deploy their new creations on their platforms

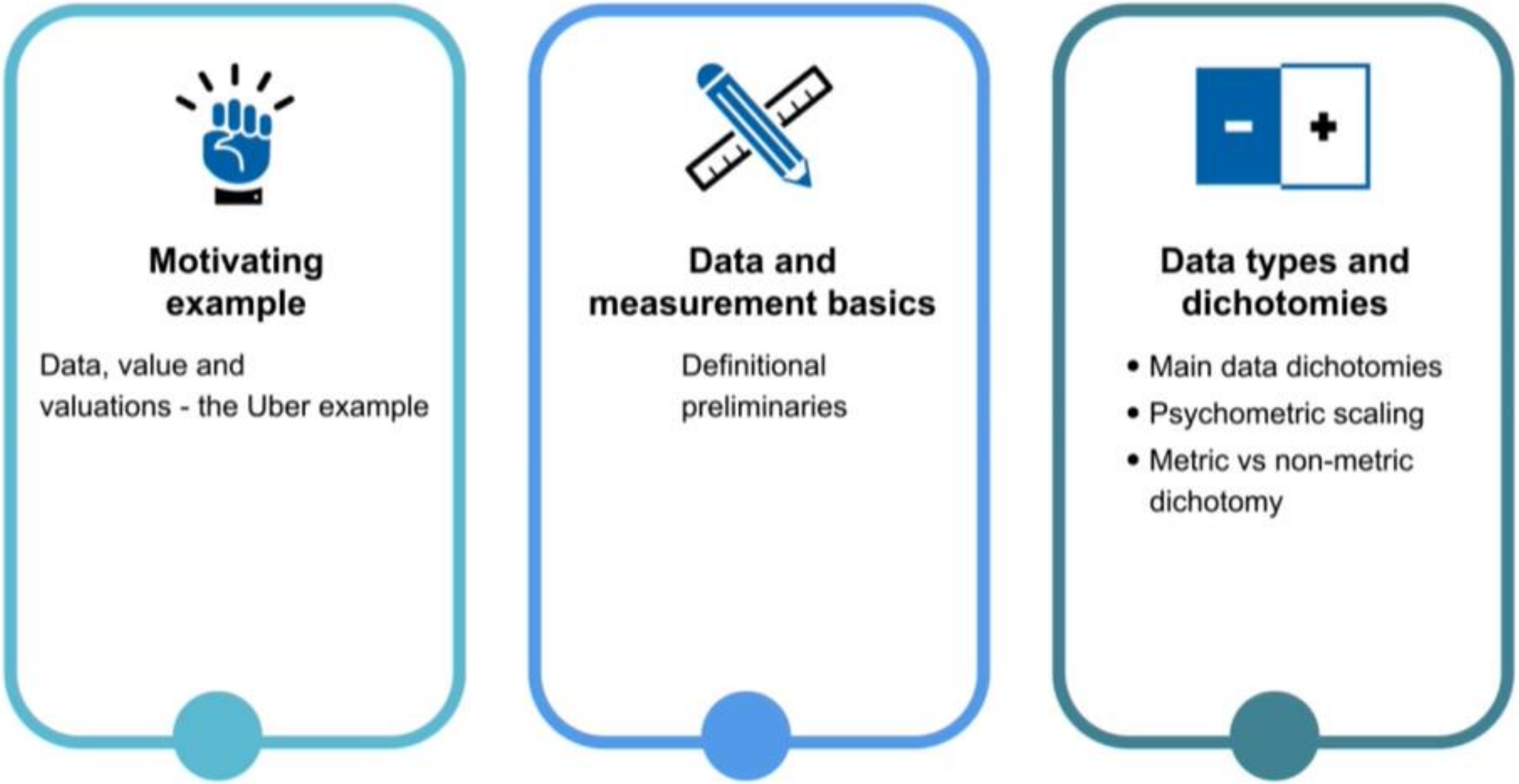
In which domains might APIs be found?

- Many domains such as marketing, analytics, data, etc.

APIs: The Growth Story across Domains



Data Preliminaries for Analytics: Summary



Motivating example

Data, value and valuations - the Uber example

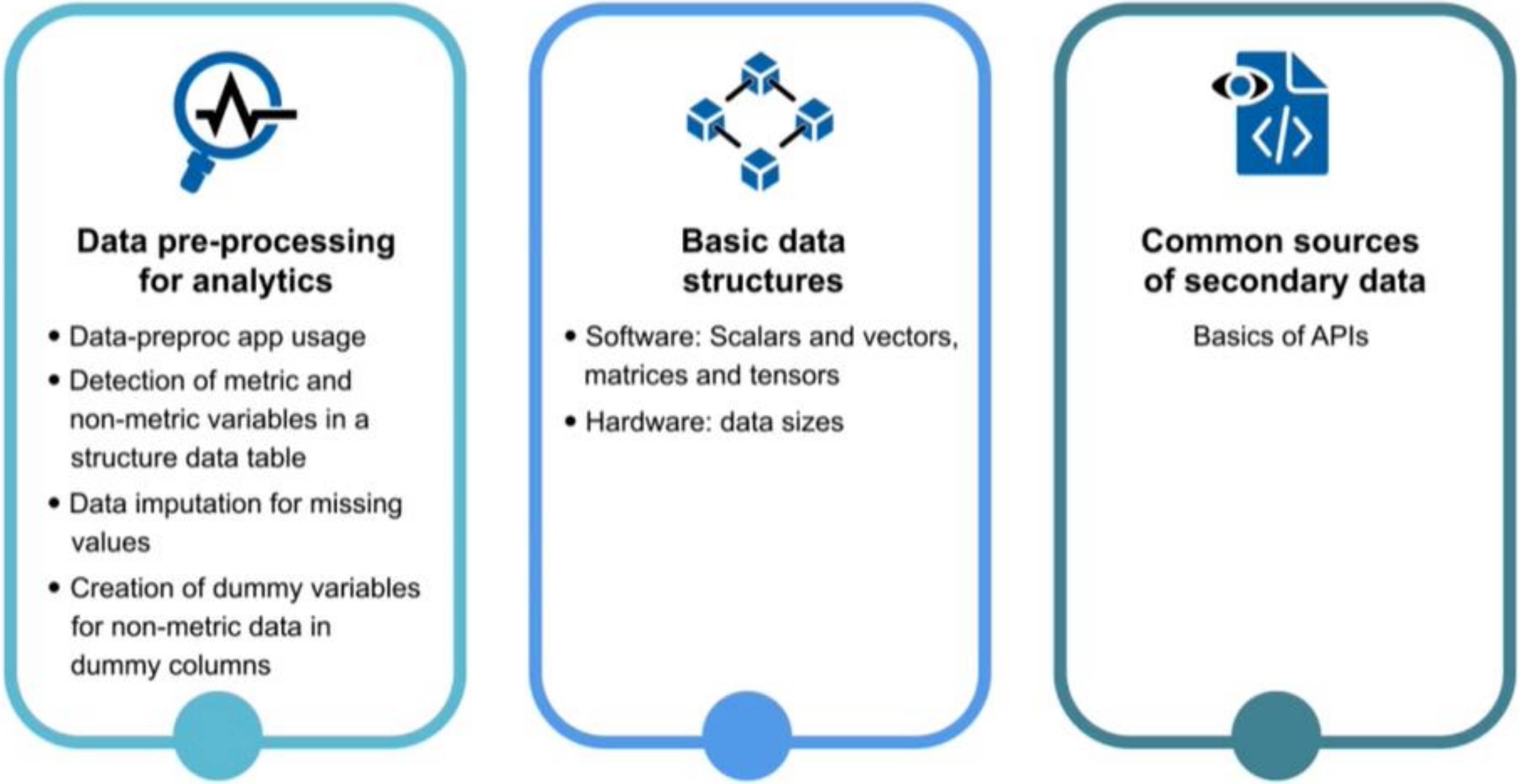
Data and measurement basics

Definitional preliminaries

Data types and dichotomies

- Main data dichotomies
- Psychometric scaling
- Metric vs non-metric dichotomy

Data Preliminaries for Analytics: Summary



Data pre-processing for analytics

- Data-preproc app usage
- Detection of metric and non-metric variables in a structure data table
- Data imputation for missing values
- Creation of dummy variables for non-metric data in dummy columns

Basic data structures

- Software: Scalars and vectors, matrices and tensors
- Hardware: data sizes

Common sources of secondary data

Basics of APIs

ISB | Executive Education