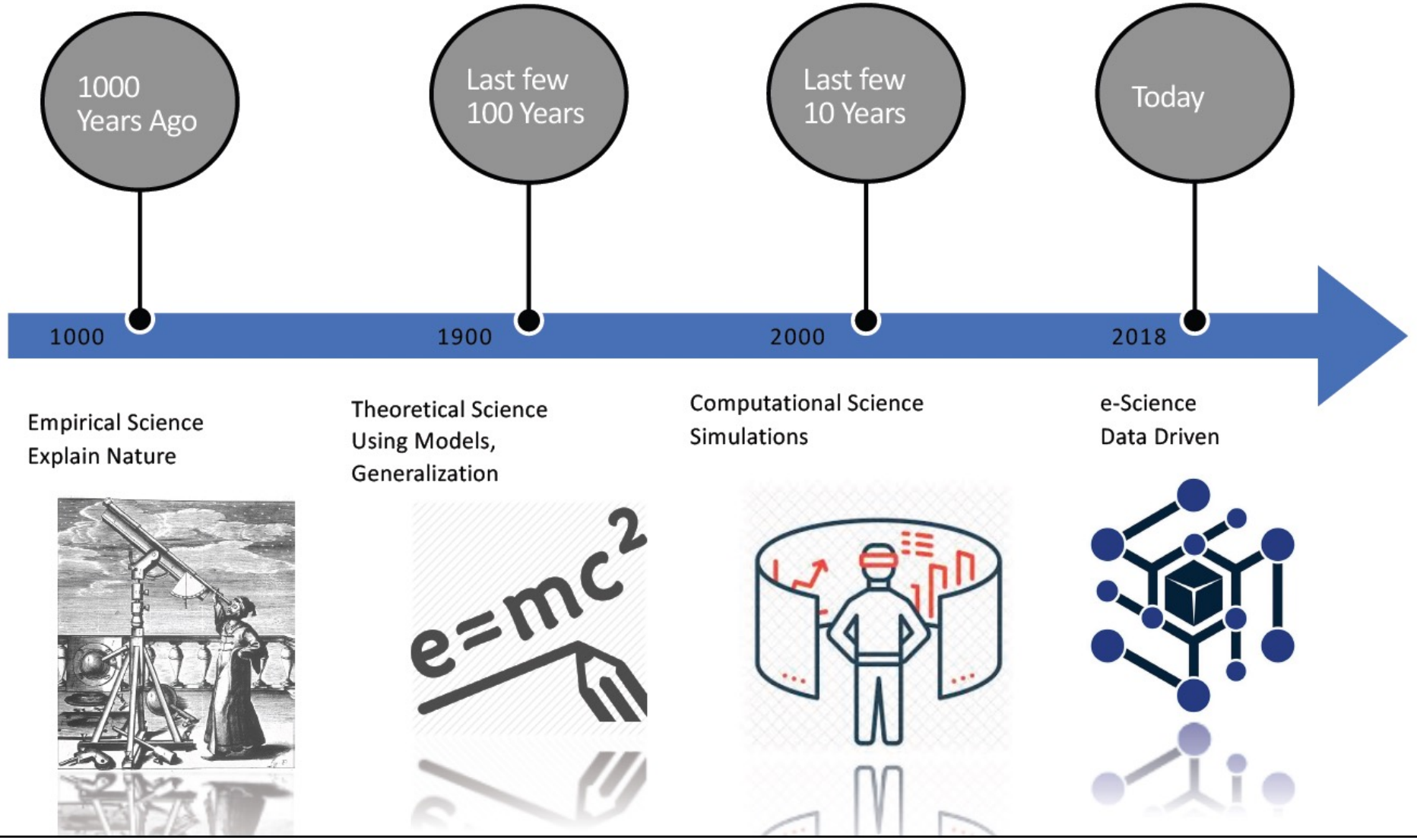
Introduction to Data Analytics

Contents

- Why data analytics
- Data driven decision making
- 4 steps of data analytics
- 4 Data types
- Data wrangling
- 7 steps of machine learning
- Hands-on

The New Paradigm

Jim Gray



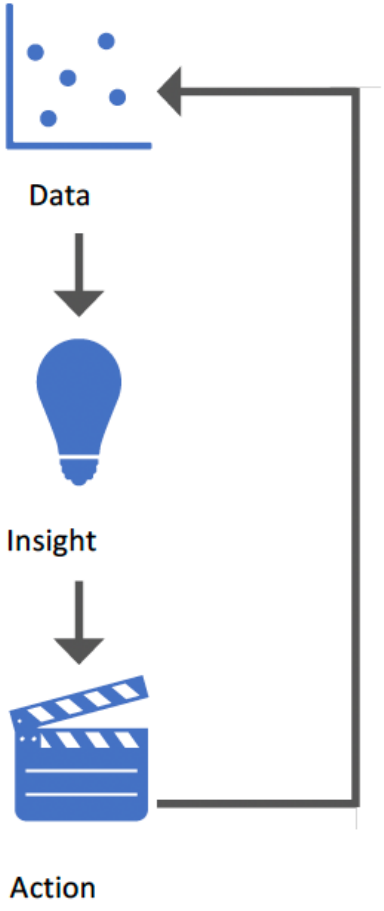
Let data drive decisions, not the Highest Paid Person's Opinion.

#HowGoogleWorks

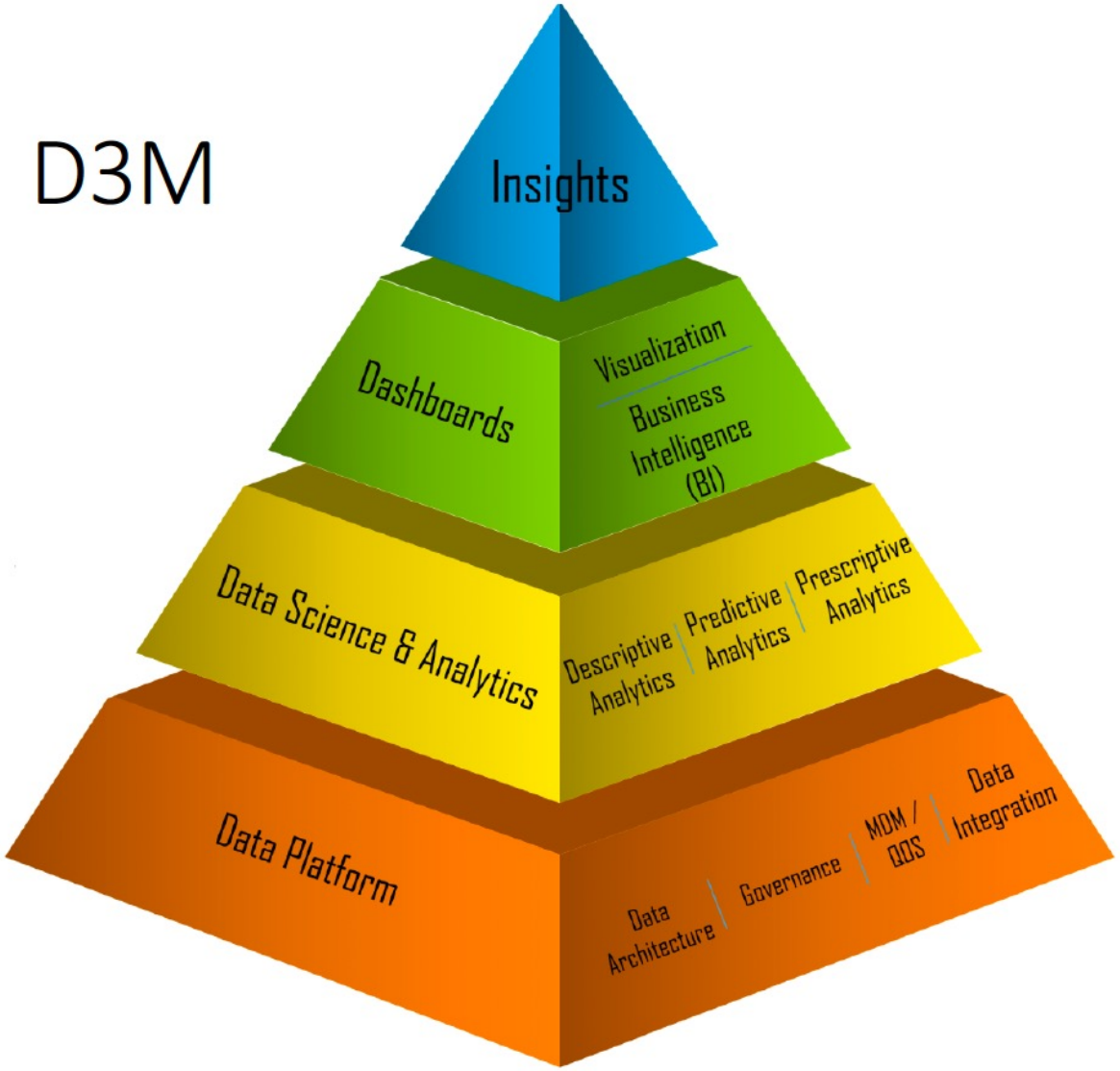
HowGoogleWorks.net

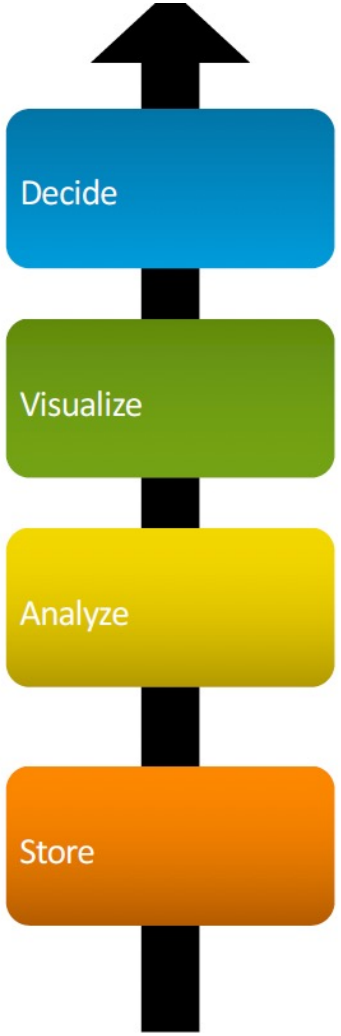
Data Driven Decision Making (D^3M)

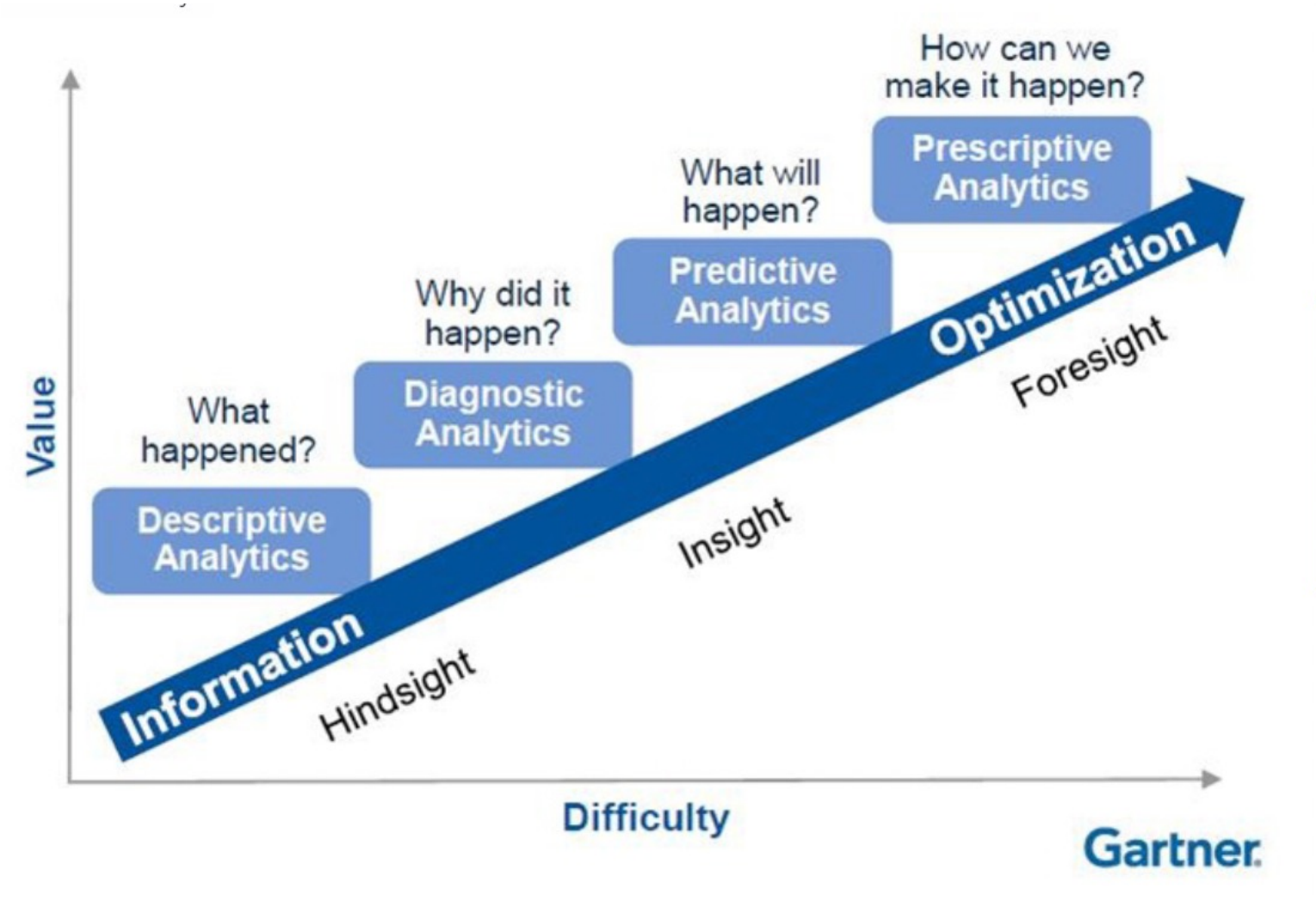
- Data-driven decision making (DDDM) involves making decisions that are backed up by hard data rather than making decisions that are intuitive or based on observation alone.
- As business technology has advanced exponentially in recent years, data-driven decision making has become a much more fundamental part of all sorts of industries.
  - Medicine
  - Transportation
  - Manufacture



D3M


Insights
Dashboards
Visualization
Business Intelligence (BI)
Data Science & Analytics
Descriptive Analytics
Predictive Analytics
Prescriptive Analytics
Data Platform
Data Architecture
Governance
MDM / QDS
Data Integration


Decide
Visualize
Analyze
Store



Gartners Maturity Model

4 Step Analytical Process

https://www.youtube.com/watch?v=tIVXbHFnaVw

Examples

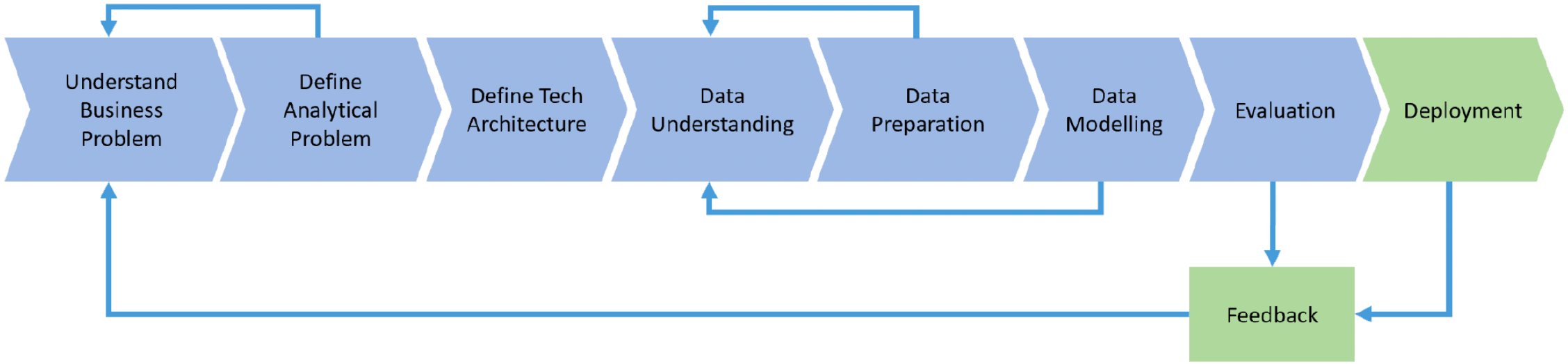
Activity 1.1.a: Can you find such examples in your org ?

- Descriptive
  - How many car did we sell last Q ?
  - How many patients were diagnosed with HBP last year ?
- Diagnostic
  - Why did we only sell 10 mid size cars last year ?
  - Why did these patients developed HBP ?
- Predictive
  - If I run paper adds how many mid size car can I sell ?
  - What are the chances John's HBP will result in stroke?
- Prescriptive
  - What do we need to do to sell 100 mid size cars ?
  - What John should do to avoid stroke ?

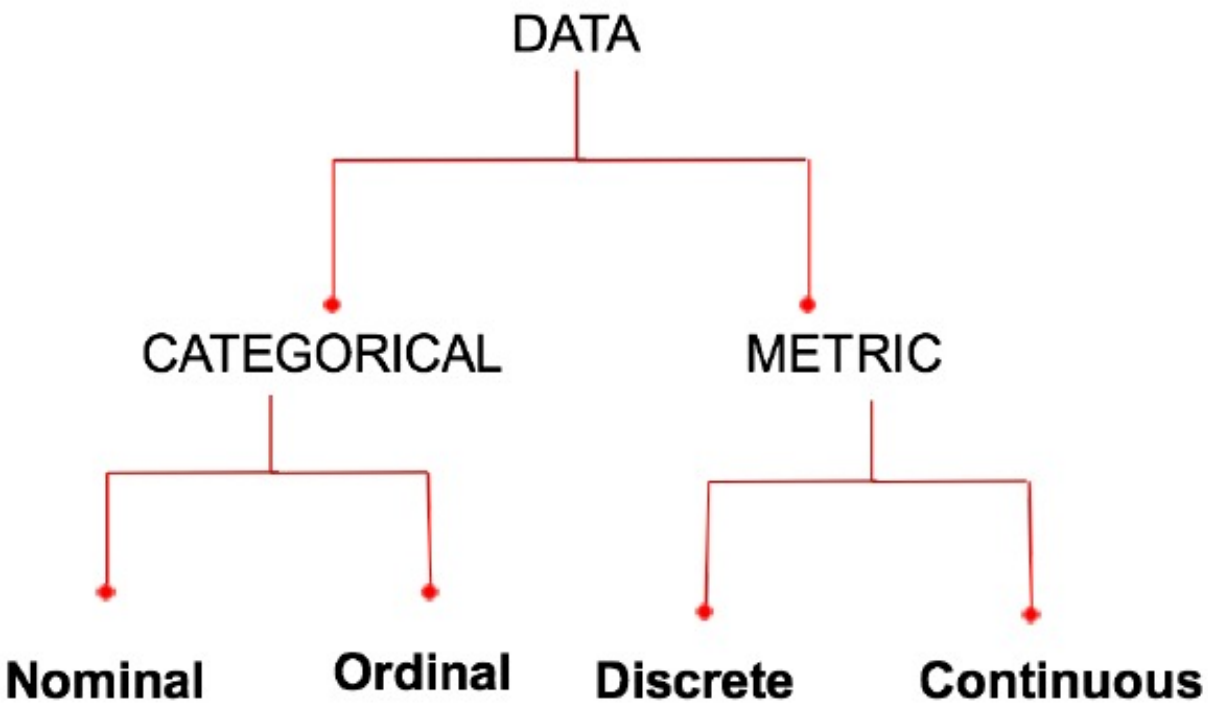
Examples

- AETNA hospital developed a process for first time patients that improves their condition for 60%.
- Amex can identify 24% of their customers who will close their accounts.
- Atlanta Falcons American football team create game plan using the GPS traces from players (uses a chip)
- Google Working with the U.S. Centers for Disease Control, tracks when users are inputting search terms related to flu topics, to help predict which regions may experience outbreaks.
- Netflix produce entertainment based on viewership
- Uber is cutting the number of cars on the roads of London by a third through UberPool that cater to users who are interested in lowering their carbon footprint and fuel costs.
- Wal-Mart says adding semantic search has improved online shoppers completing a purchase by 10% to 15%. In Wal-Mart terms, that is billions of dollars.

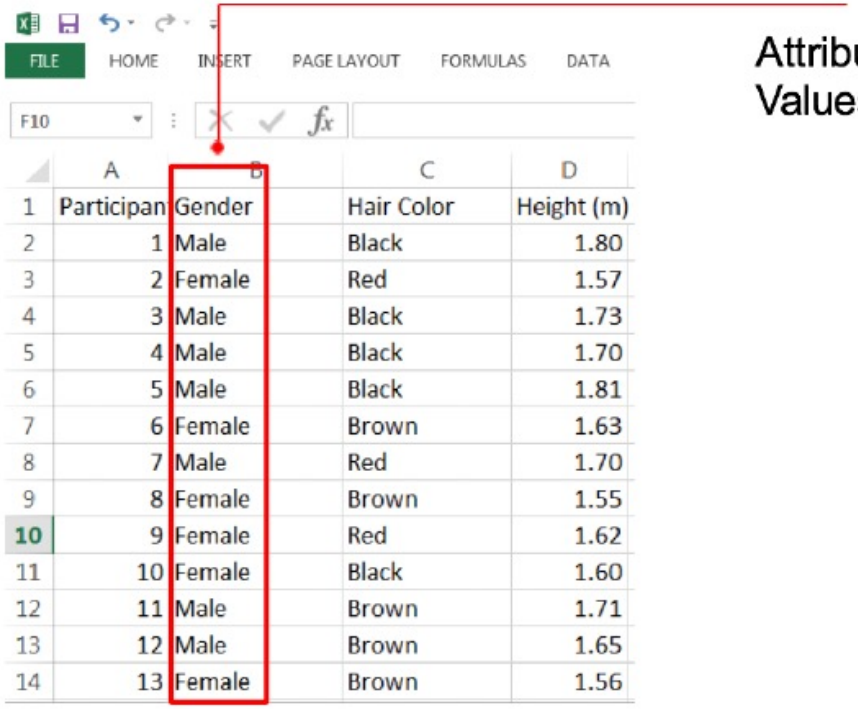
DataScience Process



Fundamental Data Types



Nominal (Categorical Variable)

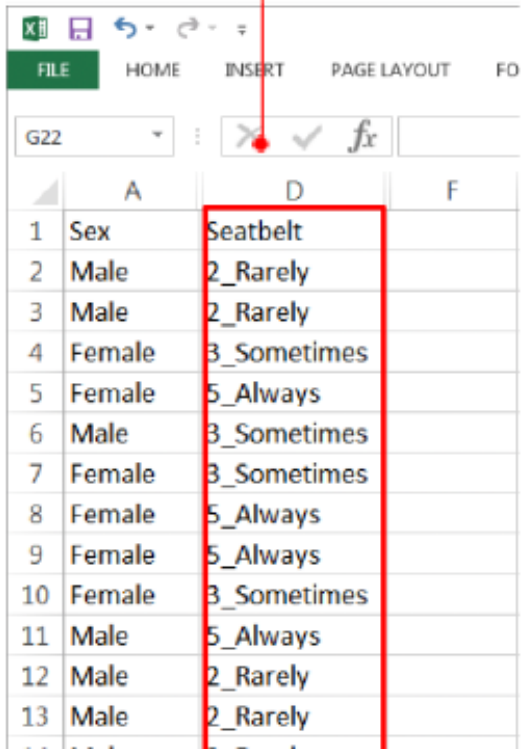


	A	B	C	D
1	Participan	Gender	Hair Color	Height (m)
2	1	Male	Black	1.80
3	2	Female	Red	1.57
4	3	Male	Black	1.73
5	4	Male	Black	1.70
6	5	Male	Black	1.81
7	6	Female	Brown	1.63
8	7	Male	Red	1.70
9	8	Female	Brown	1.55
10	9	Female	Red	1.62
11	10	Female	Black	1.60
12	11	Male	Brown	1.71
13	12	Male	Brown	1.65
14	13	Female	Brown	1.56

Attribute: Gender
Values: {Male, Female}

NONIMAL DATA

Ordinal (Categorical Variable)

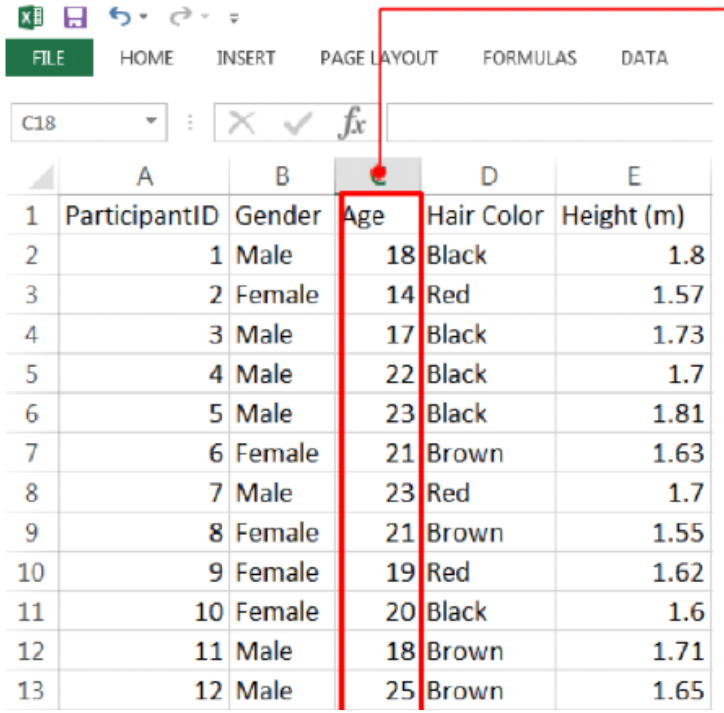


	A	D	F
1	Sex	Seatbelt	
2	Male	2_Rarely	
3	Male	2_Rarely	
4	Female	3_Sometimes	
5	Female	5_Always	
6	Male	3_Sometimes	
7	Female	3_Sometimes	
8	Female	5_Always	
9	Female	5_Always	
10	Female	3_Sometimes	
11	Male	5_Always	
12	Male	2_Rarely	
13	Male	2_Rarely	

Attribute: Seatbelt
Values: {Always, Mosttimes, Sometimes, Rarely, Never}

ORDINAL DATA

Discrete (Metric Variable)



	A	B	C	D	E
1	ParticipantID	Gender	Age	Hair Color	Height (m)
2	1	Male	18	Black	1.8
3	2	Female	14	Red	1.57
4	3	Male	17	Black	1.73
5	4	Male	22	Black	1.7
6	5	Male	23	Black	1.81
7	6	Female	21	Brown	1.63
8	7	Male	23	Red	1.7
9	8	Female	21	Brown	1.55
10	9	Female	19	Red	1.62
11	10	Female	20	Black	1.6
12	11	Male	18	Brown	1.71
13	12	Male	25	Brown	1.65

Attribute: **Age**
Values: 1,2,3,4,5,6,7,…..,125

DISCRETE DATA

Continuous (Metric Variable)

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA

F10

	A	B	C	D
1	Participan	Gender	Hair Color	Height (m)
2	1	Male	Black	1.80
3	2	Female	Red	1.57
4	3	Male	Black	1.73
5	4	Male	Black	1.70
6	5	Male	Black	1.81
7	6	Female	Brown	1.63
8	7	Male	Red	1.70
9	8	Female	Brown	1.55
10	9	Female	Red	1.62
11	10	Female	Black	1.60
12	11	Male	Brown	1.71
13	12	Male	Brown	1.65
14	13	Female	Brown	1.56

Attribute: Height
Values: 1.8, 1.57, …

CONTINUOUS DATA

Data Wrangling

Data wrangling is the process of cleaning, structuring and enriching raw data into a desired format for better decision making.

Data Wrangling Tasks

DISCOVERY STRUCTURING CLEANING ENRICHING VALIDATING PUBLISHING

Discovering

In this step, the data is to be understood more deeply.

Before implementing methods to clean it, you will definitely need to have a better idea about what the data is about.

Wrangling needs to be done in specific manners, based on some criteria which could demarcate and divide the data accordingly – these are identified in this step.

Structuring

Raw data is given to you in a haphazard manner, in most cases – there will not be any structure to it.

This needs to be rectified, and the data needs to be **restructured** in a manner that better suits the analytical method used.

Based on the criteria identified in the first step, the data will need to be **separated** for ease of use.

One column may become two, or rows may be split – whatever needs to be done for better analysis.

Cleaning

All datasets are sure to have some outliers, which can skew the results of the analysis.

These will have to be cleaned, for the best results.

In this step, the data is cleaned thoroughly for high-quality analysis.

Null values will have to be changed, and the formatting will be standardized in order to make the data of higher quality.

Enriching

After cleaning, it will have to be enriched – this is done in the fourth step.

This means that you will have to take stock of what is in the data and strategise whether you will have to augment it using some additional data in order to make it better.

You should also brainstorm about whether you can derive any new data from the existing clean data set that you have.

Predictive analytics

7-steps of machine learning

- https://www.youtube.com/watch?v=nKW8Ndu7Mjw

- In class activity:
  - Choose a prediction problem and prepare slides that shows the associated 7 steps.

Data Wrangling Process – Hands On

- **Data exploration** — columns, unique values in a column, describe, duplicates
- **Dealing with missing values** — quantifying missing values per column, filling & dropping missing values
- **Reshaping data** — one hot encoding, pivot tables, joins, grouping and aggregating
- **Filtering data**
- **Other** — Making descriptive columns, element-wise conditional operations

Hands-on Session

- https://colab.research.google.com/drive/1JWmLeulLLXx0da5TLbWv3J5diujpBMvk
- https://colab.research.google.com/drive/1W3aZyBhYO1IqphQp0SeJqlcI7sj4j0pA
- https://colab.research.google.com/drive/1Jolii3He8rQWmdFxwTXrbgRrYJ3yooBO
- https://colab.research.google.com/drive/1c3HWNy3yjDYp7UykRQL67iF4XjfGKlL8

- Exercise
- https://colab.research.google.com/drive/1NpPseHuQ7Tbyyj3h_g6TMpQH6tmv7nom
- https://colab.research.google.com/drive/1rtgUIrm7xH5L9tjL1hLhhXoyAWLfC2KG