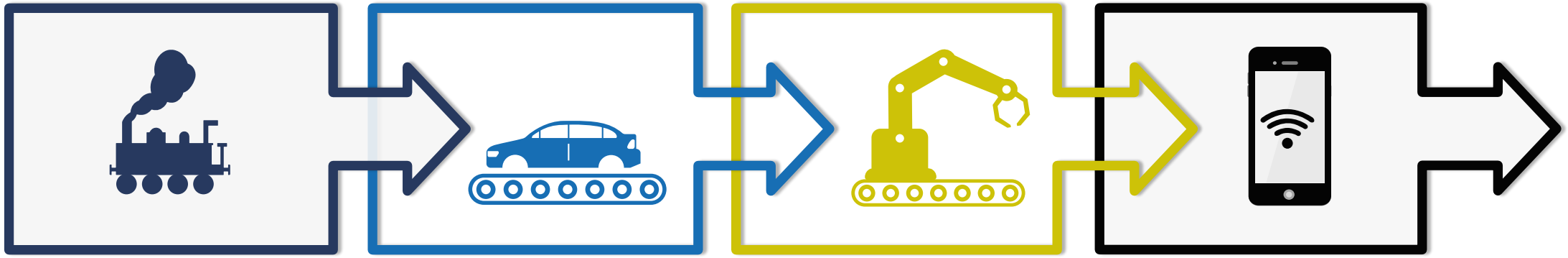
MoF-DAC
Ministry of Finance
Data Analytics Community

EDA

by Sindhu Wardhana & Rahmat Rusfandi

We are riding the data analytic wave right now…

Industrial Revolutions



1.0

STEAM POWER (STEAM ENGINE)

Around 1784 steam engine was invented, evolved the use of physical strength to machine power,

2.0

ASSEMBLY LINE (ELECTRICITY)

Around 1870 electrical technology developed and used in production. Efficiency and productivity of industry increased sharply.

3.0

AUTOMATION (COMPUTERS)

Around 1969 the emergence of computers were used in production process and further use of electronics and IT assisted the production automation and reduced human-power in life.

4.0

INTERNET OF THINGS (HIGH SPEED INTERNET)

The machines are integrated into the computer, communicating with each other with intelligent connectivity that enables intelligent decision.

WINIX (2018)

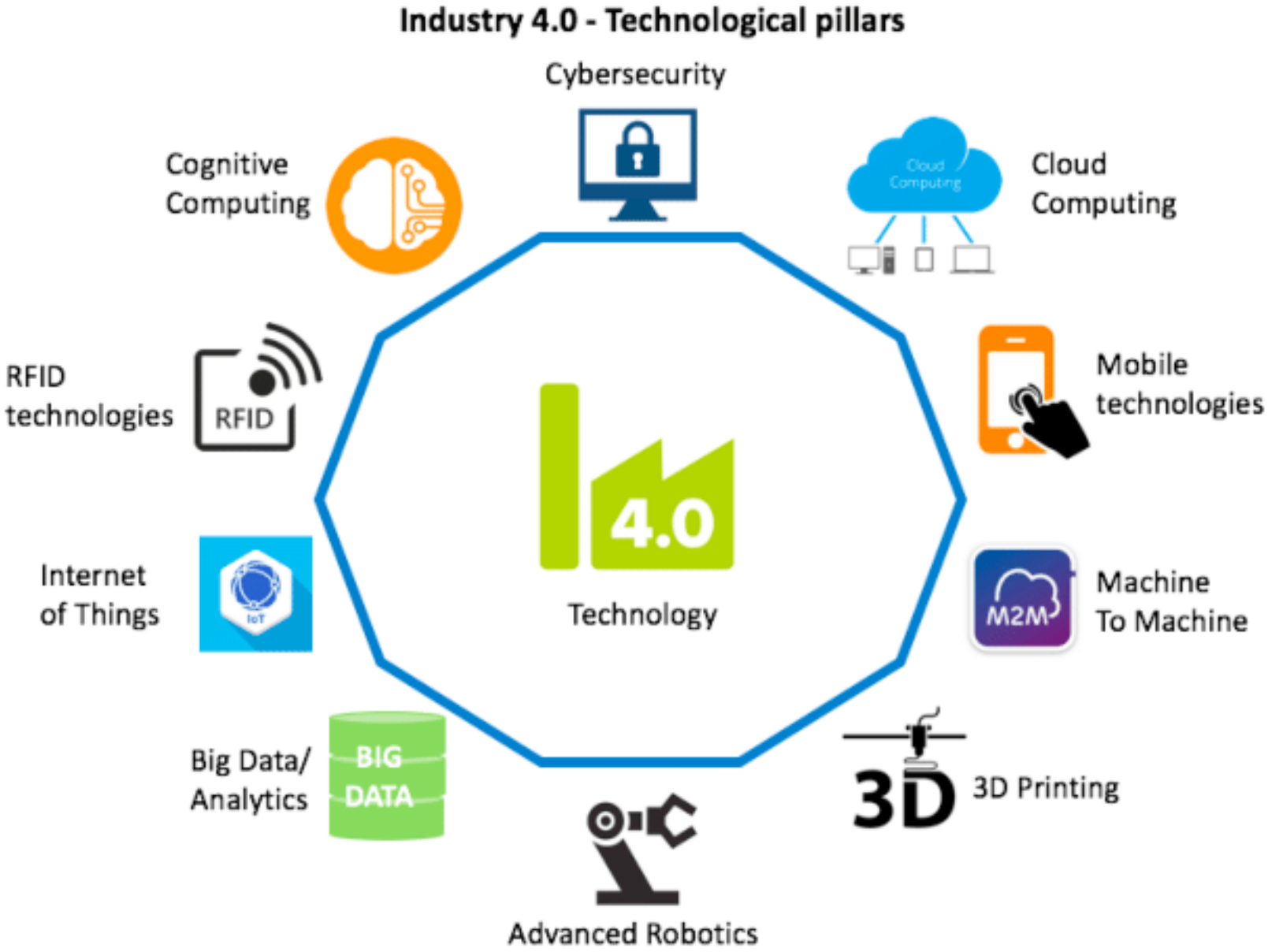
Klaus Schwab:

“Technological revolution that is blurring the lines between the physical, digital, and biological spheres”

Forbes

Forbes:

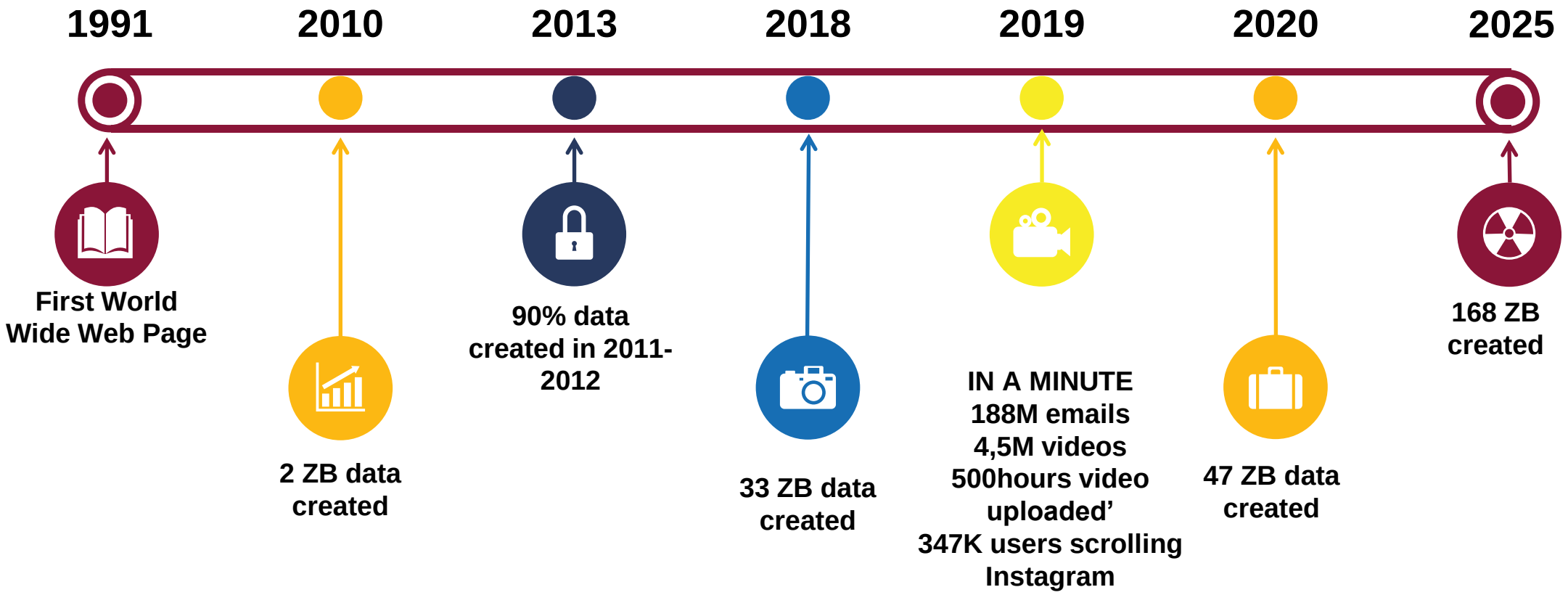
“Industry 4.0 optimizes the computerization of Industry 3.0”



Saturno et al - 2017


BIG DATA

Data Explosion



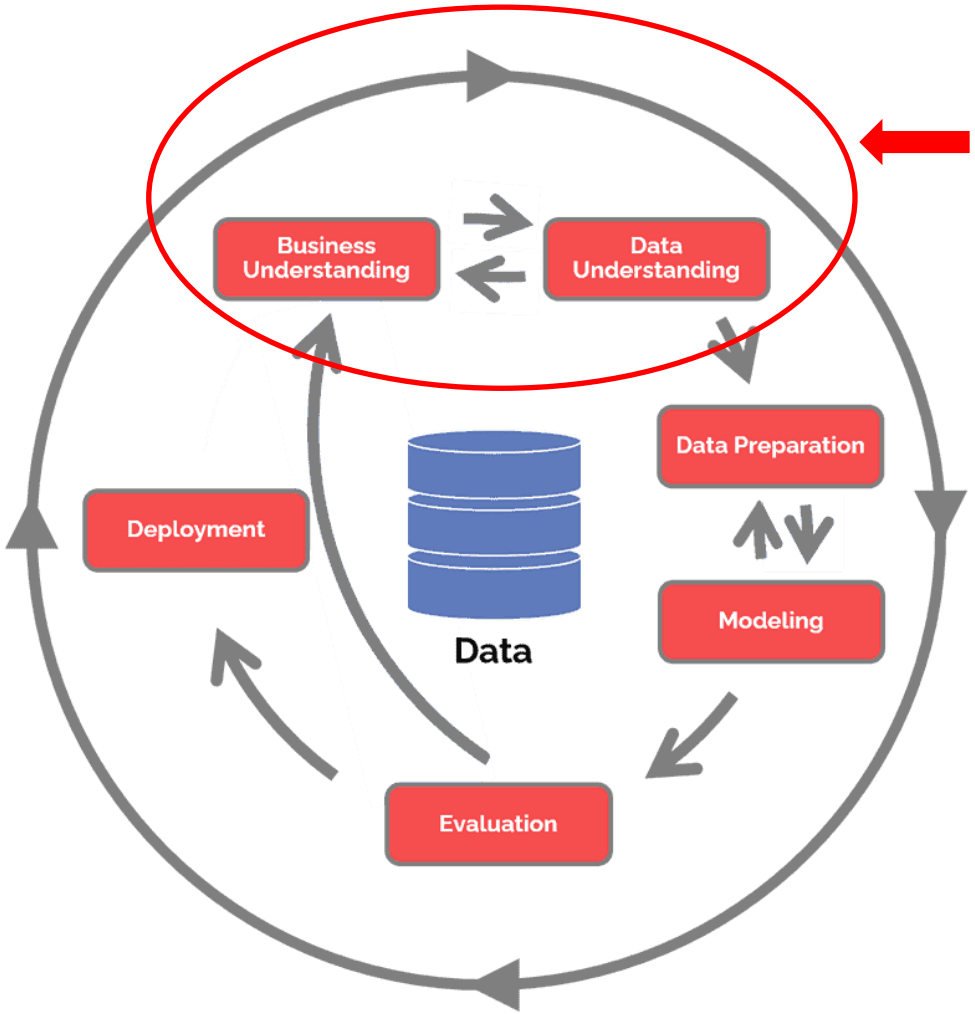
***1 ZB = 1,000,000,000,000 TB**

Seedscientific (2020) & statista (2019)

Types of Data

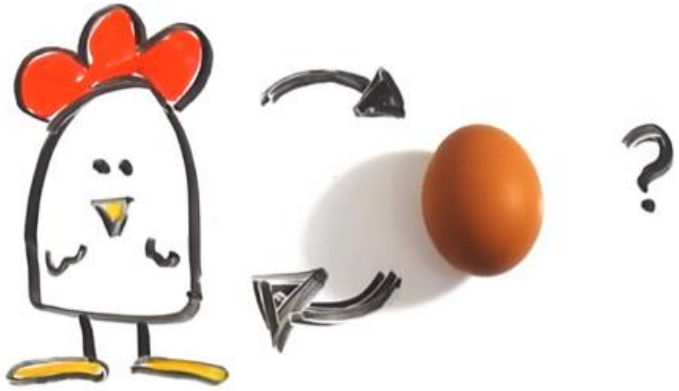
FEATURES	STRUCTURED	SEMI STRUCTURED	UNSTRUCTURED
Format Type	Relational Database	HTML, XML, JSON	Binary, Character
Version Management	Rows, columns, tuples	Not as common – graph is possible	Whole data
Implementation	SQL	Anonymous nodes	-
Robustness	Robust	Limited robustness	-
Storage Requirement	Less	Significant	Large
Applications	DBMS, RDF, ERP system, Data Warehouse, Apache Parquet, Financial Data, Relational Table	Server Logs, Sensor Output	No SQL, Video, Audio, Social Media, Online Forums, MRI, Ultrasound

Data Mining Lifecycle



Kita beriterasi di sini

THE ALL KNOWN CRISP-DM



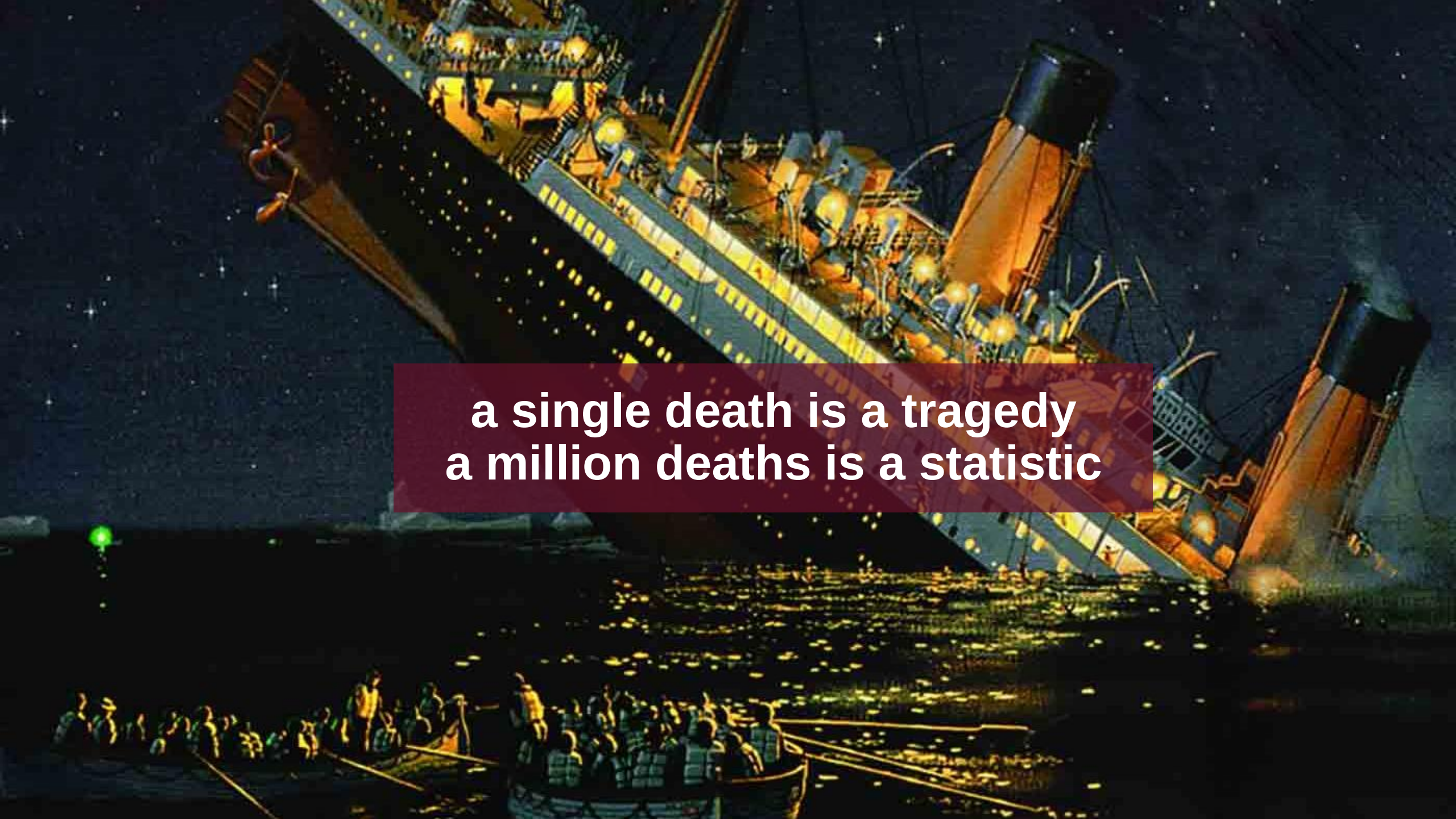
The Big Question: Business or data first first?

Business Understanding

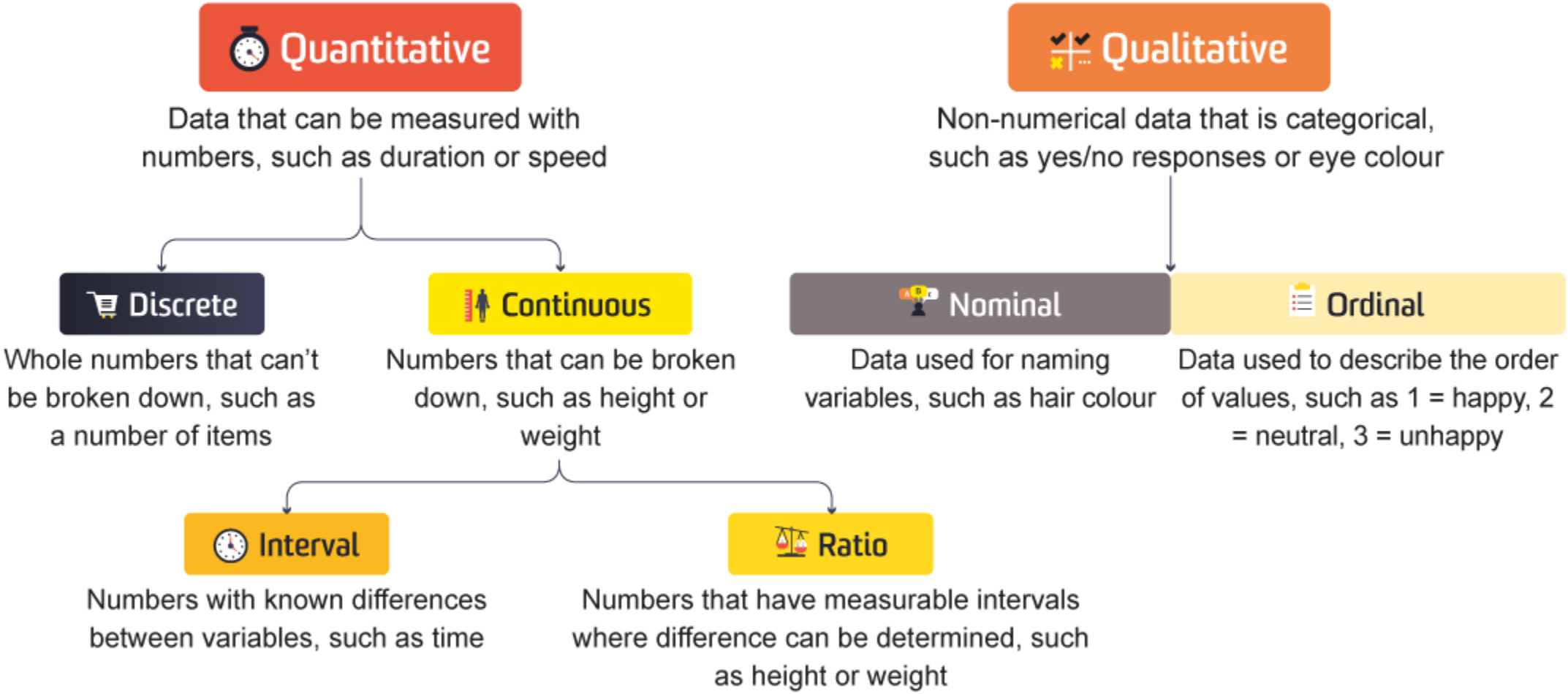
- Determine business objectives
- Asses Situation
- Determine Data Mining Goals
- Produce Project Plan
- It is a first challenge
- Sometimes, you need to learn new field/domain problem

Data Understanding

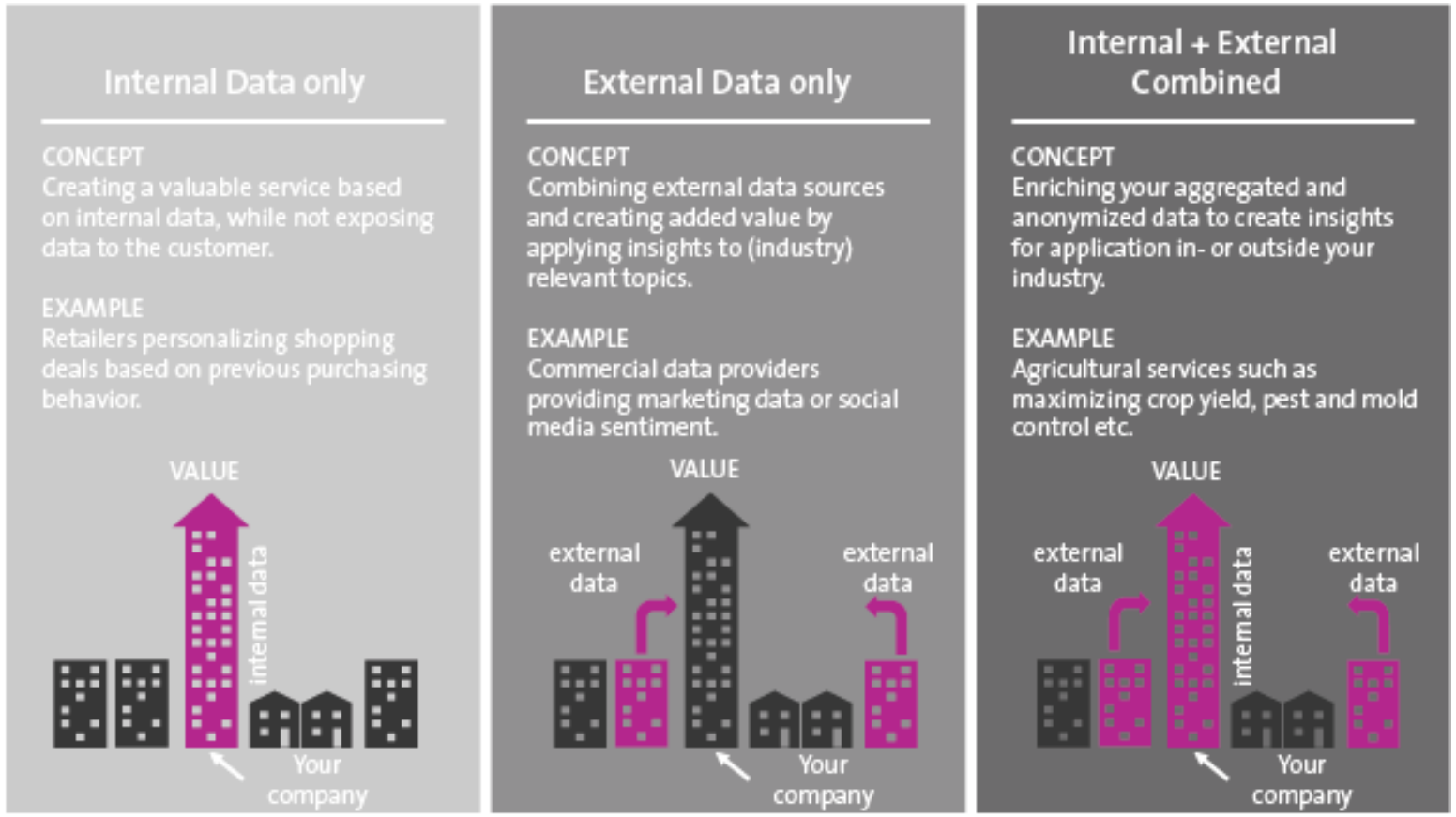
- Collect Initial Data
- Describe Data
- Explore Data
- Verify Data Quality
- Spot and Understands Anomalies and Outliers
- Understanding Variables, discovering relationship
- EDA and Visualizations


a single death is a tragedy
a million deaths is a statistic

Types of Data



Internal and External Data Sources



That hierarchy you should look at the data:

1. Internal structured data
2. Internal semi-structured
3. Internal unstructured
4. External structured
5. External unstructured.

Seven main ways of collecting data

1. **Created data:** market research surveys, focus groups or employee surveys, loyalty programs, usually structured or semi-structured and can be internal or external.
2. **Provoked data:** it wouldn't exist unless you invited people to express their views, like five star system rating, usually structured or semi-structured and can be internal or external.
3. **Transaction data:** generated every time a customer buys something, is usually internal structured data
4. **Compiled data:** from the giant databases that that compile vast amounts of data from different sources, usually external structured data.
5. **Experimental data:** hybrid of created and transacted data, different customer sets receive different marketing treatments (created) and observing the results in the real world (transaction), usually structured or semi-structured and can be internal or externa
6. **Captured data:** GPS data, sensors, IoT, usually unstructured and can be internal or external
7. **User-generated data:** individuals and companies generate consciously – or at least knowingly, usually unstructured and can be internal or external.

Data Quality Attributes

Attribute	What it means	Example of good practice	Example of bad practice	Metrics
Consistency	No matter where you look in the database, you won't find any contradictions in your data.	Your payment system shows that Jane Brown has made 5 purchases this month, and CRM system contains the same information.	Your payment system shows that Jane Brown has made 5 purchases this month, while CRM system shows she has made only 4.	The number of inconsistencies.
Accuracy	The information your data contains corresponds to reality.	Your customer's name is Jane Brown. And this is exactly how it's reflected in your CRM.	In your CRM, the customer's name is spelled Jane Brawn, though her actual name is Jane Brown.	The ratio of data to errors.
Completeness	All available elements of the data have found their way to the database.	You know that Jane Brown is born on 11/04/1975.	You have no idea how old Jane Brown is, as the date of birth cell is empty.	The number of missing values.
Auditability	Data is accessible and it's possible to trace introduced changes.	You can track down the changes made in Jane's data record. For example, on 12/5/2018, her phone number was changed.	It's impossible to trace down the changes in Jane's record.	% of cells where the metadata about introduced changes is not accessible.

Attribute	What it means	Example of good practice	Example of bad practice	Metrics
Orderliness	The data entered has the required format and structure.	The entry for December 11, 2018 is in the format 12/11/2018.	The entry for December 11, 2018 is in the format 12/11/18, 12/11/2018 and even 11/12/18 (in your European stores).	The ratio of data of inappropriate format.
Uniqueness	A data record with specific details appears only once in the database.	You have only one record for Jane Brown, born on 11/04/1975, who lives in Seattle.	You have multiple duplicate records for Jane Brown.	The number of duplicates revealed.
Timeliness	Data represents reality within a reasonable period of time or in accordance with corporate standards.	On 02/15/2018, the customer informed you that her name is misspelled in the emails you send her. The customer's name was corrected the next day.	On 02/15/2018, the customer informed you that her name is misspelled in the emails you send her. Her name was corrected only in a month.	Number of records with delayed changes.

Guide to Data Quality Management: Metrics, Process and Best Practices (scnsoft.com)

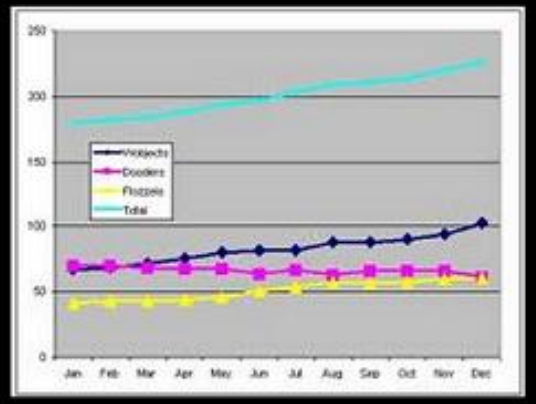
Praktik Tipis2
Jupyter Notebook

Data Scientist

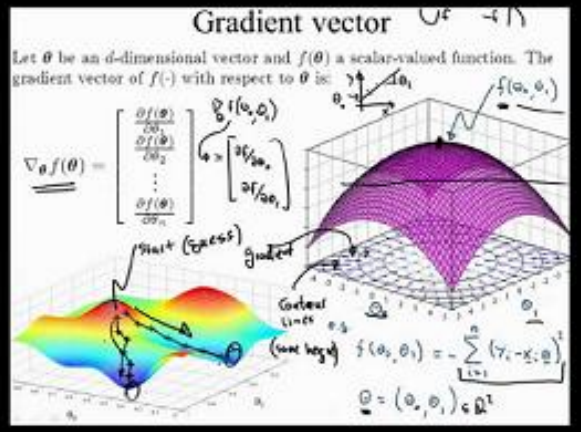
What my friends think I do

What my mom thinks I do

What society thinks I do



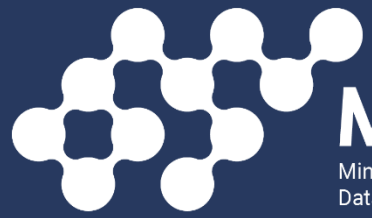
What my boss thinks I do



What I think I do

What I actually do

The learning curve is quite high, but it will be the keystone of data driven organization.



MoF-DAC

Ministry of Finance
Data Analytics Community

Thanks