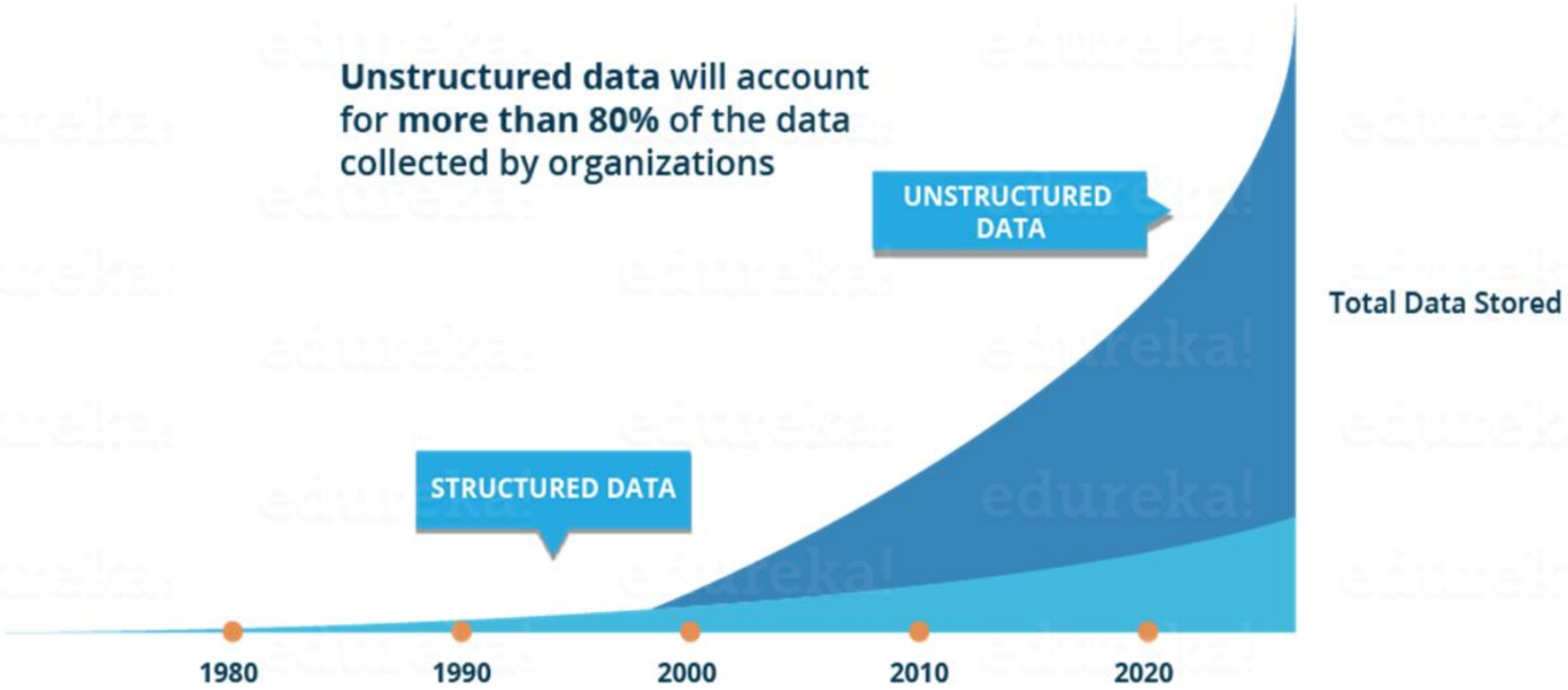
Data Analysis

THE BRIDGE
DIGITAL TALENT ACCELERATOR

Origen


Unstructured data will account for more than 80% of the data collected by organizations
UNSTRUCTURED DATA
STRUCTURED DATA
Total Data Stored
1980
1990
2000
2010
2020

Origen

Tech Tuesdays: Why a Shocking 80% of Data Science ...

Sep 8, 2020 — Trifacta, a company that tries to make the **data cleaning** process easier, has this to say: **80**% of the **time** spent on data analytics is allocated to ...

www.reddit.com › datascience › comments › bupmyf

Data Scientists spend up to 80% of time on "data cleaning" in ...

May 30, 2019 — Data Scientists spend up to **80**% of **time** on "**data cleaning**" in preparation for data analysis, statistical modeling, & machine learning. Post Credit: Igor Korolev.

Data Cleaning 80% time? : datascience - Reddit Nov 13, 2020
How much **time** do you spend on **data** wrangling ... - Reddit Aug 2, 2018
More results from www.reddit.com

www.infoworld.com › Data Science › Analytics

The 80/20 data science dilemma | InfoWorld

Sep 26, 2017 — Most **data** scientists spend only 20 percent of their **time** on actual **data** analysis and **80** percent of their **time** finding, **cleaning**, and reorganizing ...

www.datanami.com › 2020/07/06 › data-prep-still-dom...

Data Prep Still Dominates Data Scientists' Time, Survey Finds

Jul 6, 2020 — "**Data** preparation and **cleansing** takes valuable **time** away from real ... in the past, **data** prep tasks have occupied upwards of 70% to **80**% of a ...

www.forbes.com › sites › gilpress › 2016/03/23 › data-pr...

Cleaning Big Data: Most Time-Consuming, Least Enjoyable ...

Cleaning Big **Data**: Most **Time**-Consuming, Least Enjoyable **Data** ... The survey of about **80** **data** scientists ...

THE BRIDGE

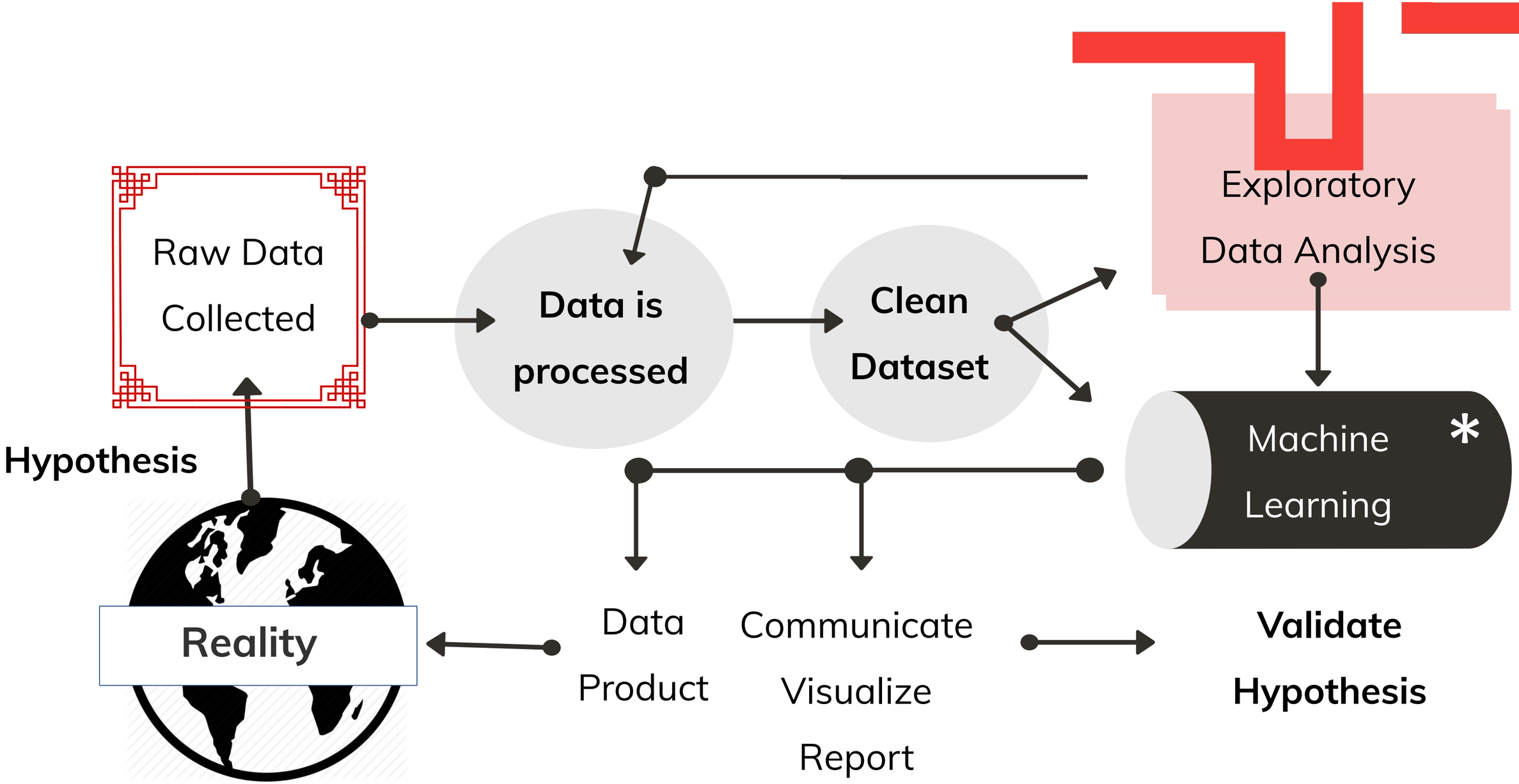
¿Qué es Data Analysis?

Data Analysis

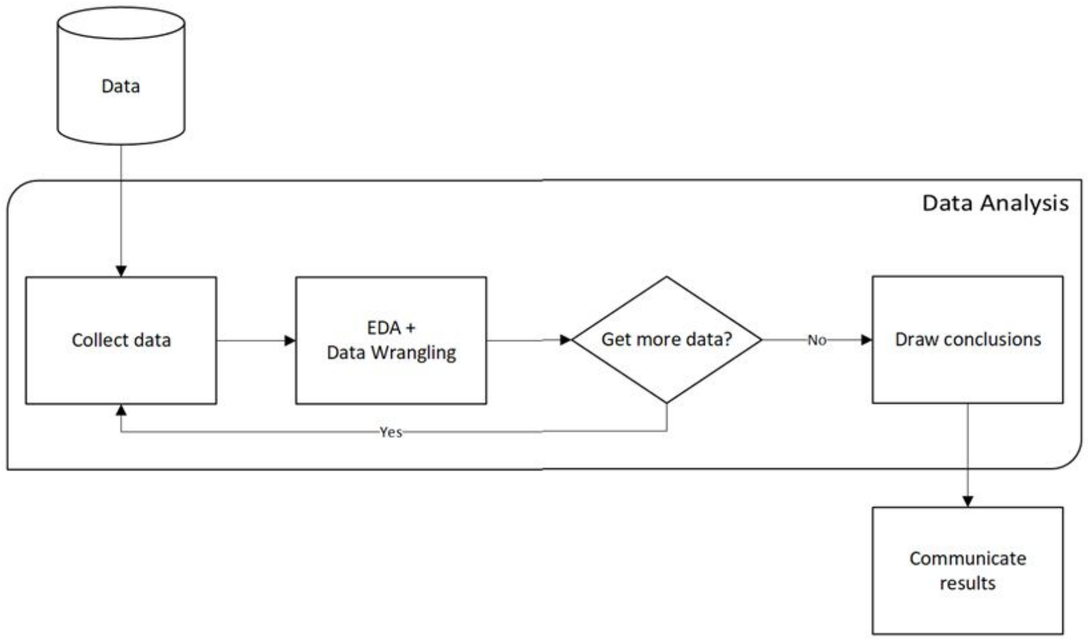
El proceso de Análisis de Datos conlleva la recolección, transformación, limpieza y modelado de datos para descubrir la información útil y de interés para una organización. Todos los datos obtenidos se transforman en conclusiones y se usan para la toma de decisiones.

¿Dónde sitúo todo esto en el conjunto del Bootcamp? ¿Y de la industria? ¿Y de Data Science?

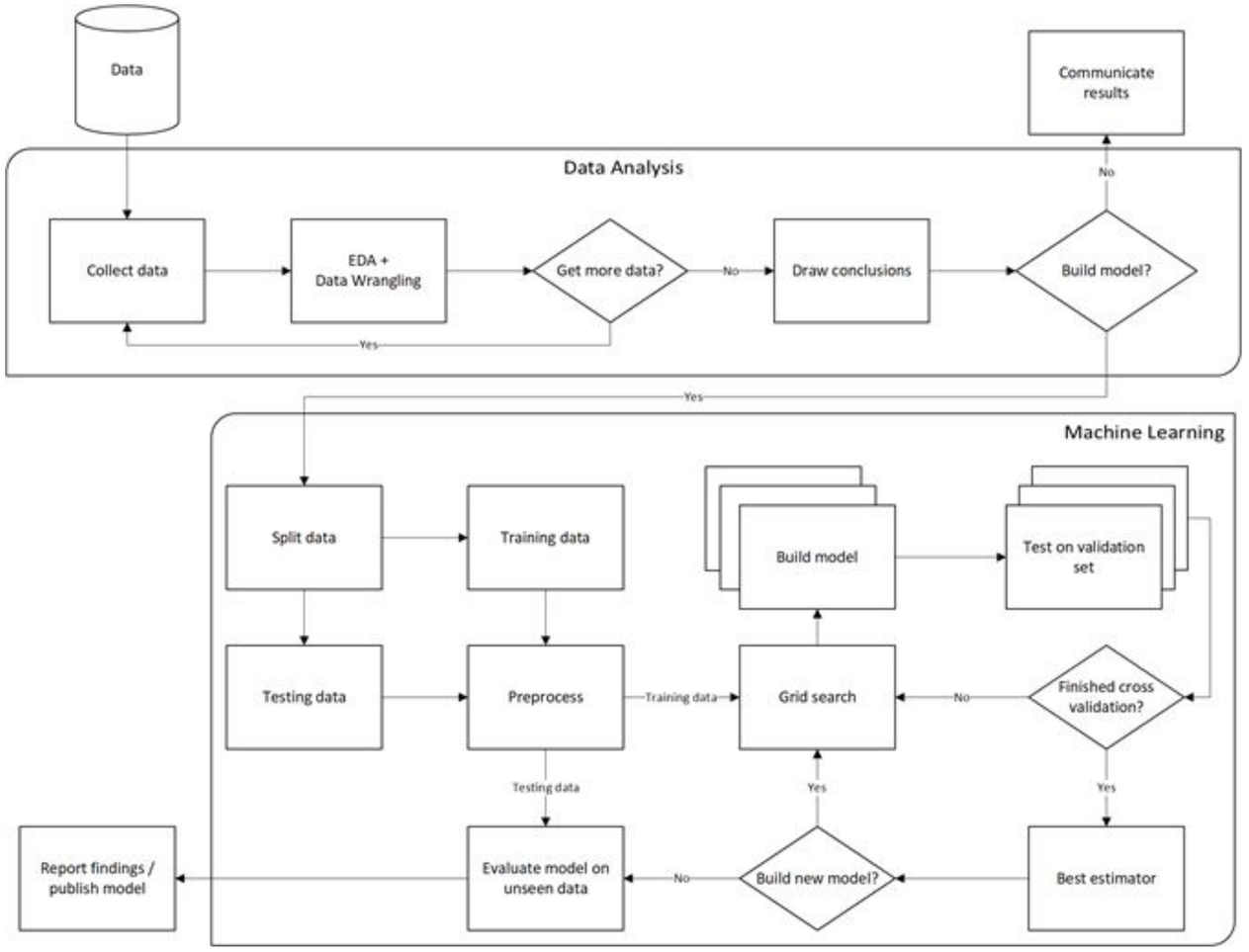
THE BRIDGE


Raw Data
Collected
Data is
processed
Clean
Dataset
Exploratory
Data Analysis
Machine
Learning
*
Hypothesis
Reality
Data
Product
Communicate
Visualize
Report
Validate
Hypothesis

Data Analysis

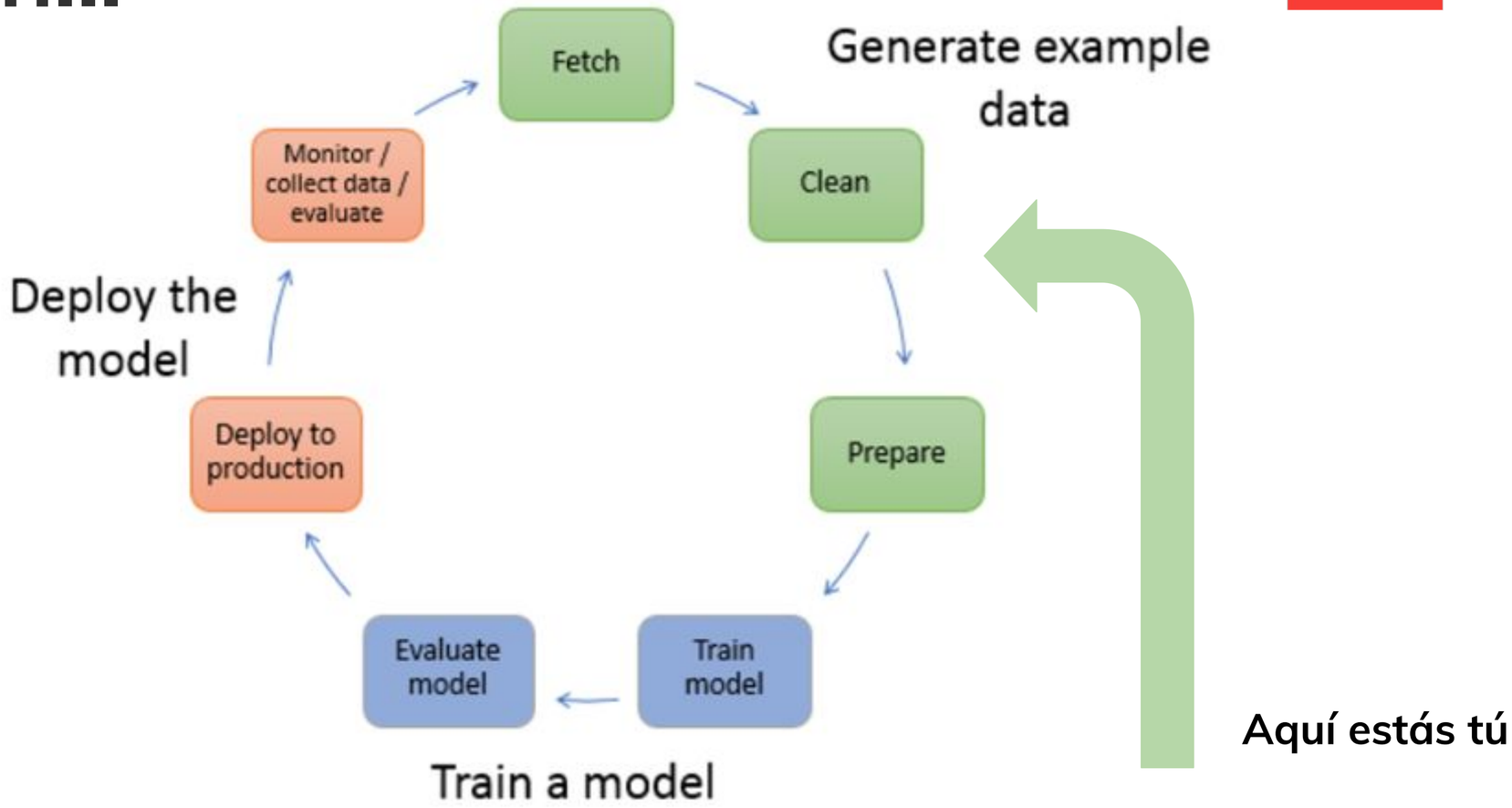


Data Science



THE BRIDGE

En resumen...


Fetch
Generate example data
Clean
Prepare
Train model
Evaluate model
Deploy to production
Deploy the model
Monitor / collect data / evaluate
Train a model
Aquí estás tú

Exploratory Data Analysis

THE BRIDGE

Exploratory Data Analysis

¿De qué se compone un EDA?

Hipótesis
Cuestiones de negocio como puntos débiles, oportunidades

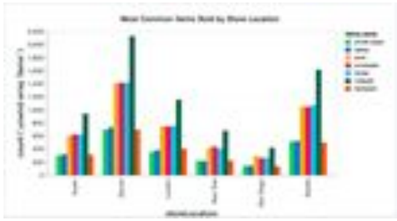
Estructura
Cómo son tus datos, de qué tipo son, calidad del dato

Limpieza
Datos duplicados, missings, features correlacionadas.

Outliers
Técnicas de detección e imputación de outliers

Feature Engineering
Transformación de variables y creación de nuevas a partir de las originales

Relación entre datos
Qué variables están correlacionadas

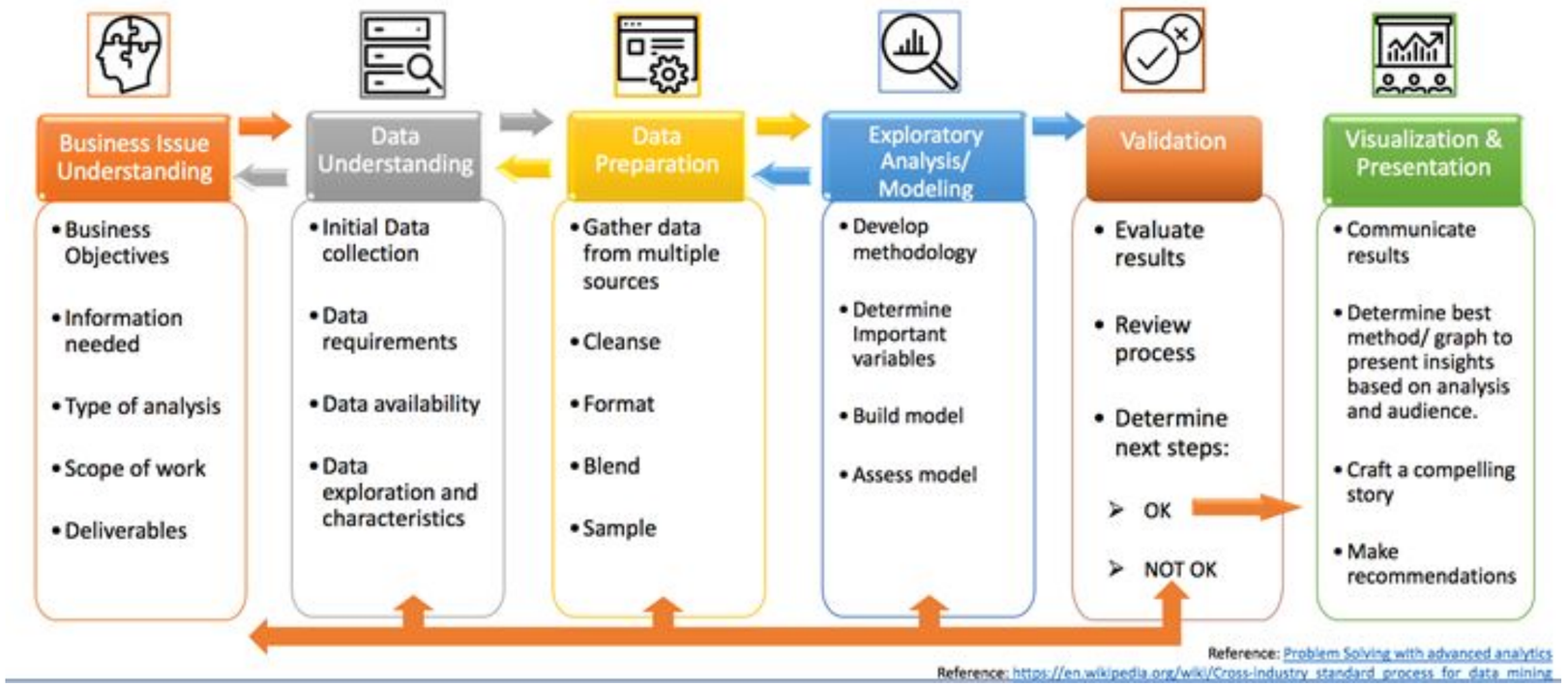


Visualización
Representación de relaciones y conclusiones sacadas de los datos.

THE BRIDGE

Proyecto EDA

¿En qué momento se hace un EDA?



Preparamos los datos para entrenar un modelo de machine learning

Análisis de datos de la empresa para entender el negocio y los clientes

Análisis de la calidad de los datos

Reportes

¿Ha tenido éxito el nuevo producto lanzado? ¿Con qué producto estamos teniendo más pérdidas? ¿O más ganancias? ¿Cómo es el cliente que compra esos productos?

Validación de hipótesis que pretenden observar la realidad

THE BRIDGE

Contexto Bootcamp

THE BRIDGE

Bootcamp

Lo que ya hemos visto

- *Variables, tipos de datos*
- *Sentencias if/else*
- *Bucles: for, while*
- *Try/Except*
- *Funciones*
- *Clases y objetos*
- *Módulos y paquetes*
- *Clean Code*

- *Concepto de Array*
- *Atributos del array*
- *Indexing*
- *Slicing*
- *Reshape*
- *Tipos de los datos en numpy*
- *Sustitución*
- *Copias*
- *Splitting*
- *Agregaciones*
- *Máscaras*
- *Operaciones*

- *Estructuras de datos*
- *Series*
- *DataFrame*
- *Index*
- *Selección e indexing*
- *Exploración DataFrame: head, describe, info*
- *Lectura de datos: read_csv*
- *Filtrado de filas*
- *Missings*
- *Uniendo tablas*
  - *Concat*
  - *Merge*
- *Agregaciones: groupby*

THE BRIDGE

Bootcamp

Lo que viene

Web scraping
APIs

y más...

Semana 7/8 | Semana 9 | Semana 10 | Semana 11

Feature Engineering

THE BRIDGE

Web Scraping

Anexo:Países por PIB (nominal) per cápita

Esta es una lista de países del mundo ordenada según su producto interno bruto (PIB) a precios nominales per cápita, significando la suma de todos los bienes y servicios finales producidos por un país en un año, dividido por la población estimada para mediados del mismo año.

Lista según el Fondo Monetario Internacional (*Estimado* 2019)[1]

Pos.	País	USD
1	Luxemburgo	113,196
2	Suiza	83,716
—	Macao	81,151
3	Noruega	77,975
4	Irlanda	77,771
5	Catar	69,687
6	Islandia	67,037
7	Estados Unidos	65,111
8	Singapur	63,987
9	Dinamarca	59,795
10	Australia	53,825

Lista según el Banco Mundial (2019)[2]

Pos.	País	USD
1	Mónaco (2018)	185,741
2	Liechtenstein (2017)	173,356
3	Luxemburgo	114,705
—	Bermudas (RU) (2013)	85,748
—	Islas Caimán (RU) (2018)	85,477
—	Macao (China)	84,096
4	Suiza	81,994
—	Isla de Man (RU) (2017)	80,989

Lista según la ONU (2018)[3] nota 2

Pos.	País	USD
1	Mónaco	190 532
2	Liechtenstein	178 799
—	*Bermudas*	117 768
3	Luxemburgo	115 481
—	*Islas Caimán*	92 692
4	Suiza	85 135
—	*Macao*	84 097
5	Irlanda	81 637

	Paises	USD	Euro
0	Luxemburgo	113196	93952.68
1	Suiza	83716	69484.28
2	Macao	81151	67355.33
3	Noruega	77975	64719.25
4	Irlanda	77771	64549.93
...	...	...	...
188	Niger	405	336.15
189	Malaui	370	307.10
190	Eritrea	342	283.86
191	Burundi	309	256.47
192	Sudán del Sur	275	228.25

THE BRIDGE

Fuentes de datos

kaggle

UCI Machine Learning Repository

DATAMINER

+ https://www.paperswithcode.com/datasets
+ https://ec.europa.eu/eurostat/data/database
+ Open Data [Inserte aquí lo que le interese] (ejemplo: Ayuntamiento Madrid)

THE BRIDGE

Como véis, esto no ha hecho
más que empezar...