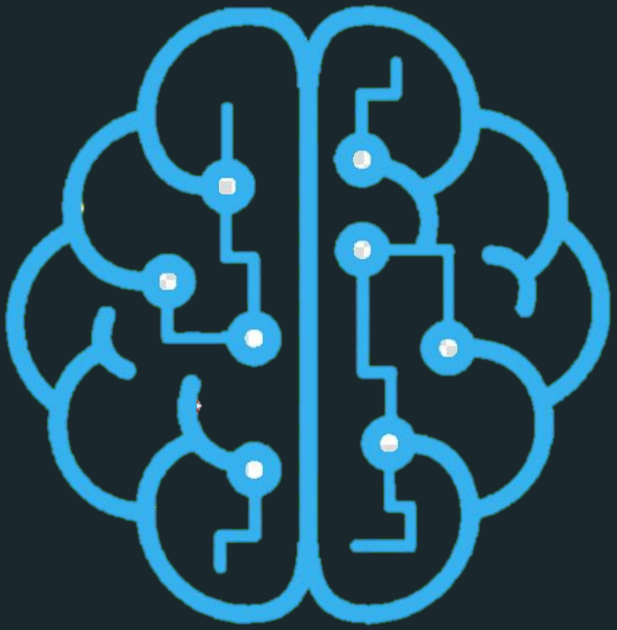
APRENDIZAJE DE MÁQUINA

TEC.AZUAY

INSTITUTO UNIVERSITARIO

Preprocesamiento



Preprocesamiento

Introducción, **Data cleaning:** Valores faltantes y duplicados, **Data transformation:** Codificación de los datos Escalamiento de los datos Normalización de los datos, **Data reduction:** Selección de características Extracción de característica, **Conclusiones**

#1. Introducción

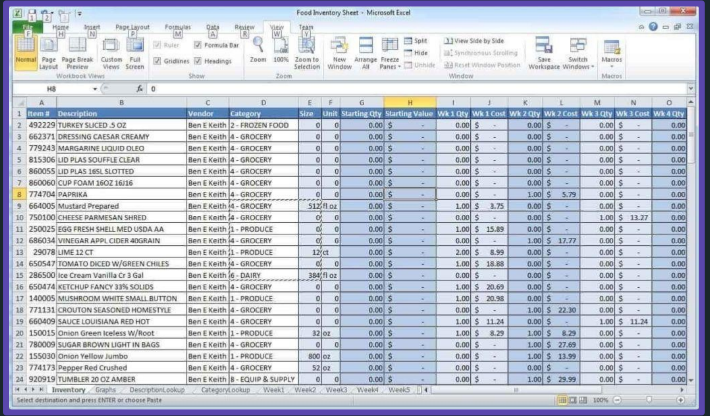
Datooos!!! Muchos datos!!

- Recurso más valioso hoy en el mundo actual.
- Según el Foro Económico Mundial, para el 2025 estaremos generando alrededor de **463 exabytes de datos a nivel mundial por día**.
- Pero:

¿Todos estos **datos son lo suficientemente adecuados** para ser utilizados por algoritmos de aprendizaje automático?

Hablando de datos pensamos …

- Grandes conjuntos de datos con una gran cantidad de filas y columnas.



Food Inventory Sheet - Microsoft Excel

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Item #	Description	Vendor	Category	Size	Unit	Starting Qty	Starting Value	Wk 1 Qty	Wk 1 Cost	Wk 2 Qty	Wk 2 Cost	Wk 3 Qty	Wk 3 Cost	Wk 4 Qty
2	492229	TURKEY SLICED .5 OZ	Ben E Keith	2 - FROZEN FOOD	0	0	0.00	$ -	0.00	$ -	0.00	$ -	0.00	$ -	0.00
3	662371	DRESSING CAESAR CREAMY	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	0.00	$ -	0.00	$ -	0.00	$ -	0.00
4	779243	MARGARINE LIQUID OLEO	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	0.00	$ -	0.00	$ -	0.00	$ -	0.00
5	815306	LID PLAS SOUFFLE CLEAR	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	0.00	$ -	0.00	$ -	0.00	$ -	0.00
6	860055	LID PLAS 16SL SLOTTED	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	0.00	$ -	0.00	$ -	0.00	$ -	0.00
7	860060	CUP FOAM 16OZ 16J16	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	0.00	$ -	0.00	$ -	0.00	$ -	0.00
8	774704	PAPRIKA	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	0.00	$ -	1.00	$ 5.79	0.00	$ -	0.00
9	664005	Mustard Prepared	Ben E Keith	4 - GROCERY	512	fl oz	0.00	$ -	1.00	$ 3.75	0.00	$ -	0.00	$ -	0.00
10	750100	CHEESE PARMESAN SHRED	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	0.00	$ -	0.00	$ -	1.00	$ 13.27	0.00
11	250025	EGG FRESH SHELL MED USDA AA	Ben E Keith	1 - PRODUCE	0	0	0.00	$ -	1.00	$ 15.89	0.00	$ -	0.00	$ -	0.00
12	686034	VINEGAR APPL CIDER 40GRAIN	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	0.00	$ -	1.00	$ 17.77	0.00	$ -	0.00
13	29078	LIME 12 CT	Ben E Keith	1 - PRODUCE	12	ct	0.00	$ -	2.00	$ 8.99	0.00	$ -	0.00	$ -	0.00
14	650547	TOMATO DICED W/GREEN CHILES	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	1.00	$ 18.88	0.00	$ -	0.00	$ -	0.00
15	286500	Ice Cream Vanilla Cr 3 Gal	Ben E Keith	6 - DAIRY	384	fl oz	0.00	$ -	0.00	$ -	0.00	$ -	0.00	$ -	0.00
16	650474	KETCHUP FANCY 33% SOLIDS	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	1.00	$ 20.69	0.00	$ -	0.00	$ -	0.00
17	140005	MUSHROOM WHITE SMALL BUTTON	Ben E Keith	1 - PRODUCE	0	0	0.00	$ -	1.00	$ 20.98	0.00	$ -	0.00	$ -	0.00
18	771131	CROUTON SEASONED HOMESTYLE	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	0.00	$ -	1.00	$ 22.30	0.00	$ -	0.00
19	660409	SAUCE LOUISIANA RED HOT	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	1.00	$ 11.24	0.00	$ -	1.00	$ 11.24	0.00
20	150015	Onion Green Iceless W/Root	Ben E Keith	1 - PRODUCE	32	oz	0.00	$ -	1.00	$ 8.29	1.00	$ 8.29	0.00	$ -	0.00
21	780009	SUGAR BROWN LIGHT IN BAGS	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	0.00	$ -	1.00	$ 27.69	0.00	$ -	0.00
22	155030	Onion Yellow Jumbo	Ben E Keith	1 - PRODUCE	800	oz	0.00	$ -	0.00	$ -	1.00	$ 13.99	0.00	$ -	0.00
23	774173	Pepper Red Crushed	Ben E Keith	4 - GROCERY	52	oz	0.00	$ -	0.00	$ -	0.00	$ -	0.00	$ -	0.00
24	920919	TUMBLER 20 OZ AMBER	Ben E Keith	8 - EQUIP & SUPPLY	0	0	0.00	$ -	0.00	$ -	1.00	$ 29.99	0.00	$ -	0.00

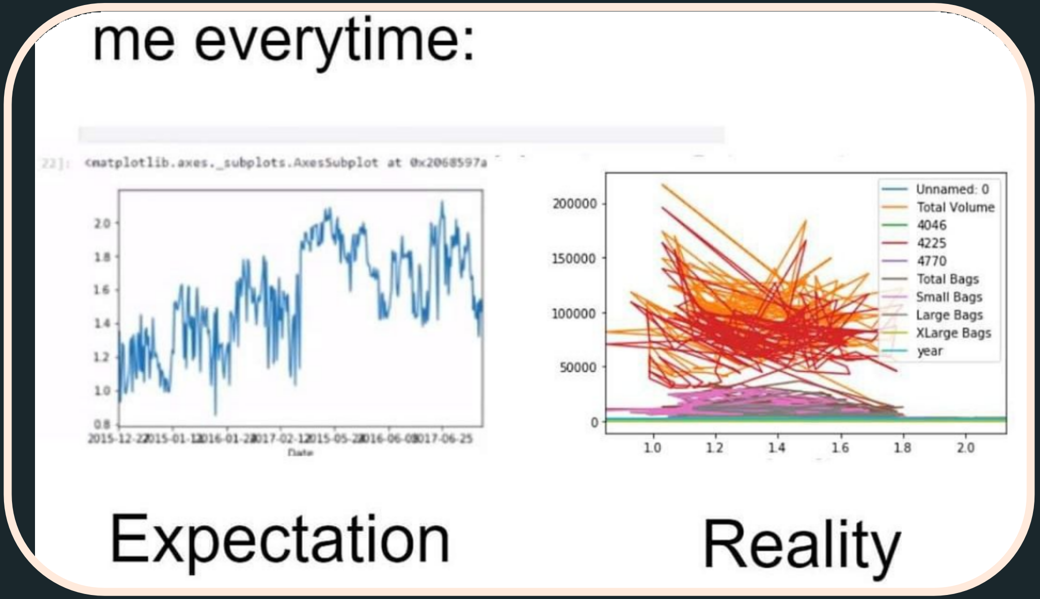
Select destination and press ENTER or choose Paste

- No siempre es el caso: los **datos pueden estar en muchas formas diferentes:** tablas estructuradas, imágenes, archivos de audio, videos, etc.

- Las **máquinas no entienden** el texto libre, las imágenes o los datos de video tal como están, **entienden los 1 y 0.**

Tenemos un problema

- Los datos del mundo real a menudo son:
  - incompletos,
  - inconsistentes,
  - contienen muchos errores.

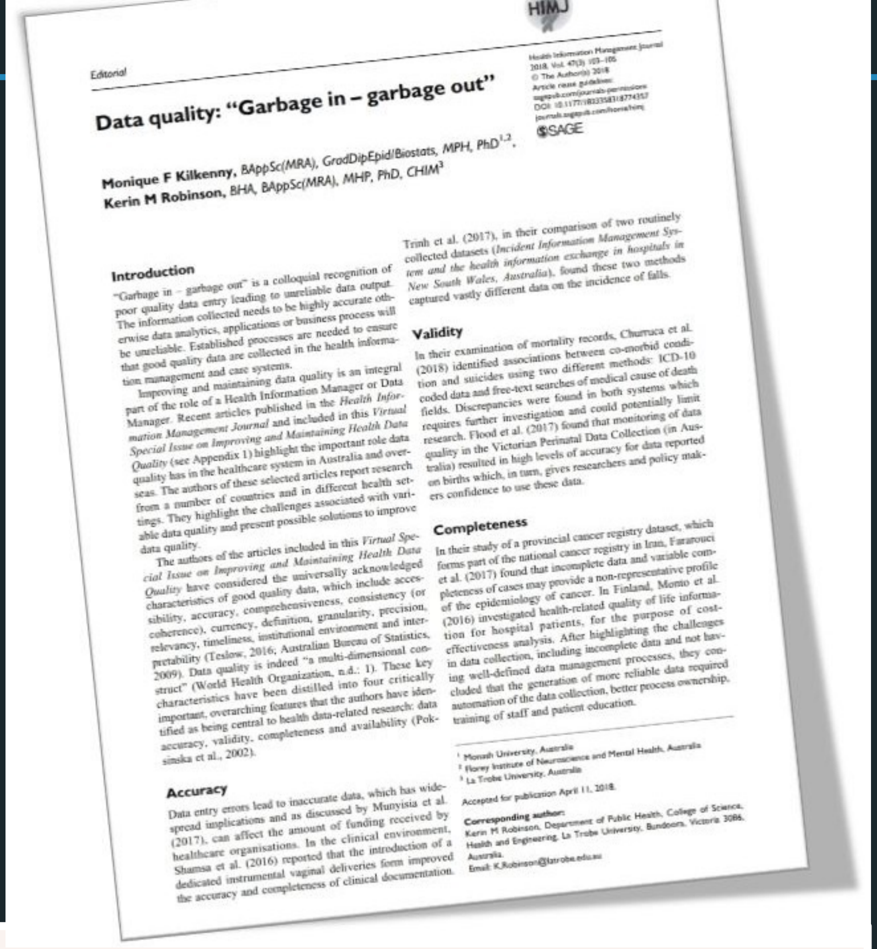


```
<matplotlib.axes._subplots.AxesSubplot at 0x2068597a
```

La calidad de los datos afecta directamente la capacidad de nuestro modelo para aprender.

Y una necesidad

- Contar con datos que sean
  - completos,
  - precisos,
  - válidos



HIMJ

Editorial

Health Information Management Journal
2018, Vol. 47(3) 103–105
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1833358318774357
journals.sagepub.com/home/himj
SAGE

Data quality: "Garbage in – garbage out"

Monique F Kilkenny, *BAppSc(MRA), GradDipEpid/Biostats, MPH, PhD*[1,2], **Kerin M Robinson**, *BHA, BAppSc(MRA), MHP, PhD, CHIM*[3]

Introduction

"Garbage in – garbage out" is a colloquial recognition of poor quality data entry leading to unreliable data output. The information collected needs to be highly accurate otherwise data analytics, applications or business process will be unreliable. Established processes are needed to ensure that good quality data are collected in the health information management and care systems.

Improving and maintaining data quality is an integral part of the role of a Health Information Manager or Data Manager. Recent articles published in the *Health Information Management Journal* and included in this *Virtual Special Issue on Improving and Maintaining Health Data Quality* (see Appendix 1) highlight the important role data quality has in the healthcare system in Australia and overseas. The authors of these selected articles report research from a number of countries and in different health settings. They highlight the challenges associated with variable data quality and present possible solutions to improve data quality.

The authors of the articles included in this *Virtual Special Issue on Improving and Maintaining Health Data Quality* have considered the universally acknowledged characteristics of good quality data, which include accessibility, accuracy, comprehensiveness, consistency (or coherence), currency, definition, granularity, precision, relevancy, timeliness, institutional environment and interpretability (Teslow, 2016; Australian Bureau of Statistics, 2009). Data quality is indeed "a multi-dimensional construct" (World Health Organization, n.d.: 1). These key characteristics have been distilled into four critically important, overarching features that the authors have identified as being central to health data-related research: data accuracy, validity, completeness and availability (Pokinska et al., 2002).

Accuracy

Data entry errors lead to inaccurate data, which has widespread implications and as discussed by Munyisia et al. (2017), can affect the amount of funding received by healthcare organisations. In the clinical environment, Shamsa et al. (2016) reported that the introduction of a dedicated instrumental vaginal deliveries form improved the accuracy and completeness of clinical documentation.

Trinh et al. (2017), in their comparison of two routinely collected datasets (*Incident Information Management System and the health information exchange in hospitals in New South Wales, Australia*), found these two methods captured vastly different data on the incidence of falls.

Validity

In their examination of mortality records, Churruca et al. (2018) identified associations between co-morbid condition and suicides using two different methods: ICD-10 coded data and free-text searches of medical cause of death fields. Discrepancies were found in both systems which requires further investigation and could potentially limit research. Flood et al. (2017) found that monitoring of data quality in the Victorian Perinatal Data Collection (in Australia) resulted in high levels of accuracy for data reported on births which, in turn, gives researchers and policy makers confidence to use these data.

Completeness

In their study of a provincial cancer registry dataset, which forms part of the national cancer registry in Iran, Fararouci et al. (2017) found that incomplete data and variable completeness of cases may provide a non-representative profile of the epidemiology of cancer. In Finland, Momo et al. (2016) investigated health-related quality of life information for hospital patients, for the purpose of cost-effectiveness analysis. After highlighting the challenges in data collection, including incomplete data and not having well-defined data management processes, they concluded that the generation of more reliable data required automation of the data collection, better process ownership, training of staff and patient education.

[1] Monash University, Australia
[2] Florey Institute of Neuroscience and Mental Health, Australia
[3] La Trobe University, Australia

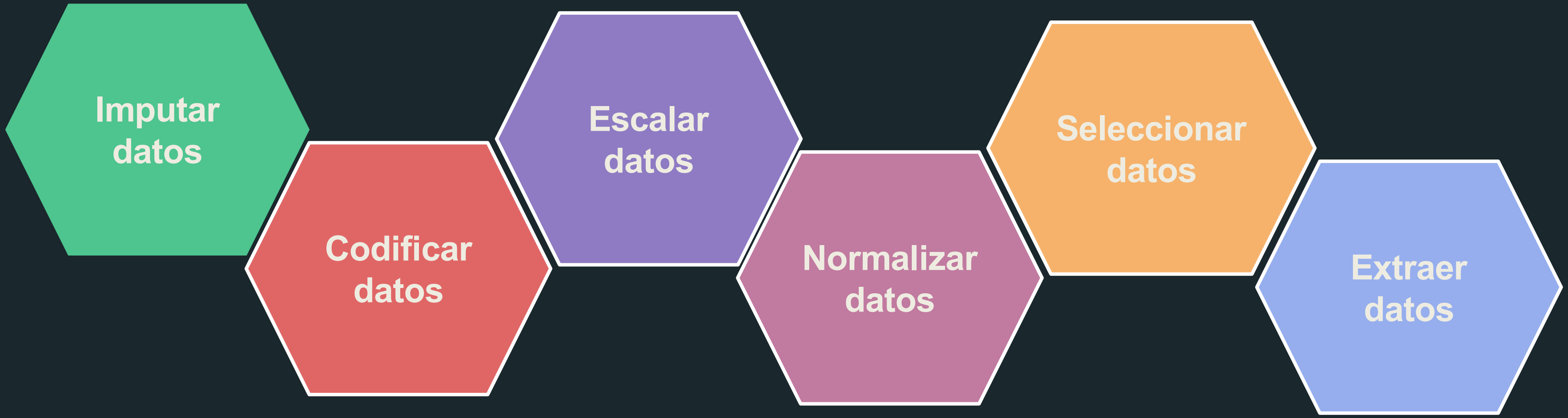
Accepted for publication April 11, 2018.

Corresponding author:
Kerin M Robinson, Department of Public Health, College of Science, Health and Engineering, La Trobe University, Bundoora, Victoria 3086, Australia.
Email: K.Robinson@latrobe.edu.au

Las características universalmente reconocidas de los datos de buena calidad, que incluyen: accesibilidad, exactitud, exhaustividad, consistenci (o coherencia), definición a vigencia, , institucional e interpretabilidad (Teslow, 2016). granularidad, precisión, relevancia, entorno oport

Data preprocessing

Es el proceso de **aplicar transformaciones a los datos** para llevarlos a un estado que la máquina pueda **analizarlos fácilmente.**



Imputar datos

Codificar datos

Escalar datos

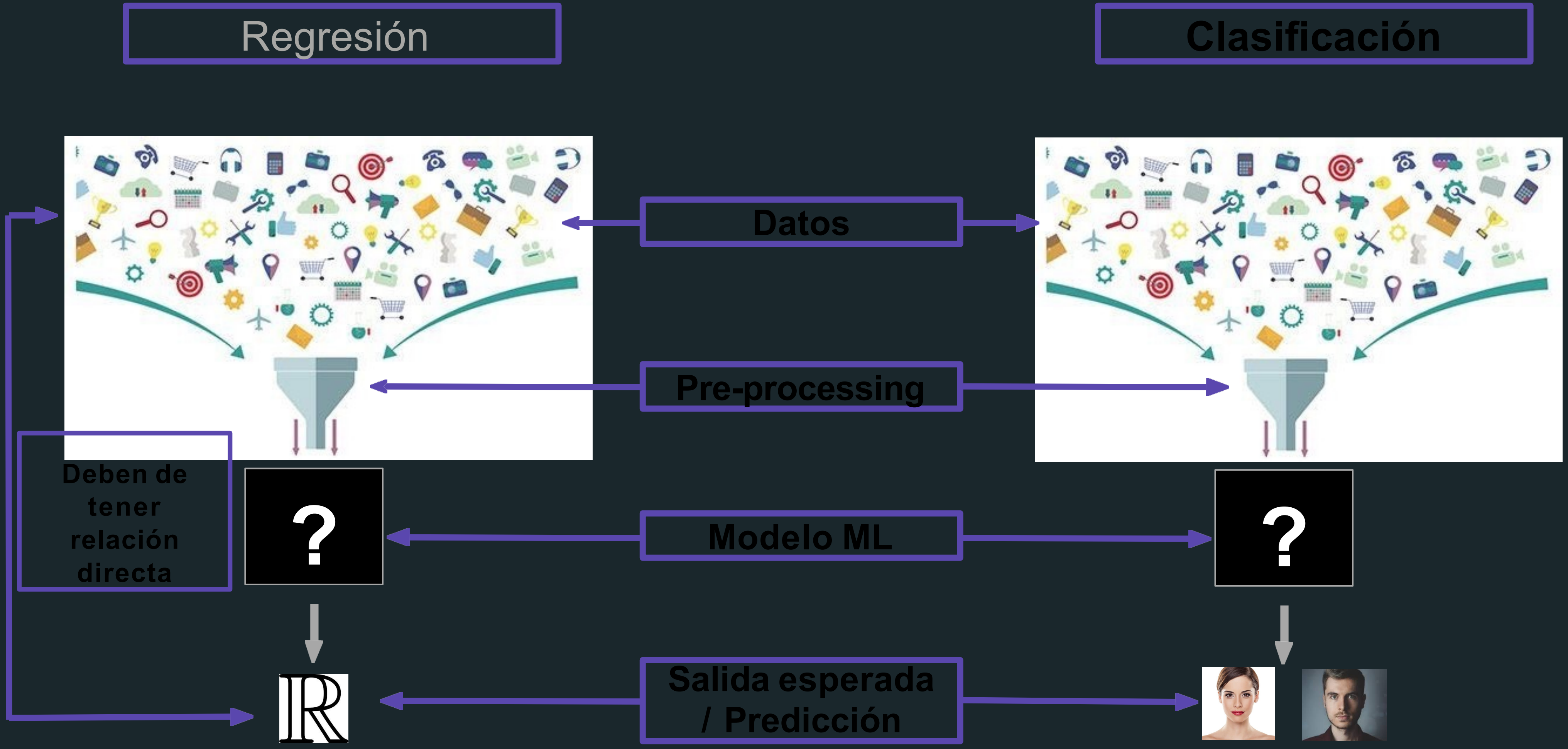
Normalizar datos

Seleccionar datos

Extraer datos

Paso muuuy importante en todo proceso de Machine Learning!!!

Su objetivo es **disponer de datos de calidad** previo al modelado utilizando algoritmos.



Regresión

Clasificación

Datos

Pre-processing

Deben de tener relación directa

?

Modelo ML

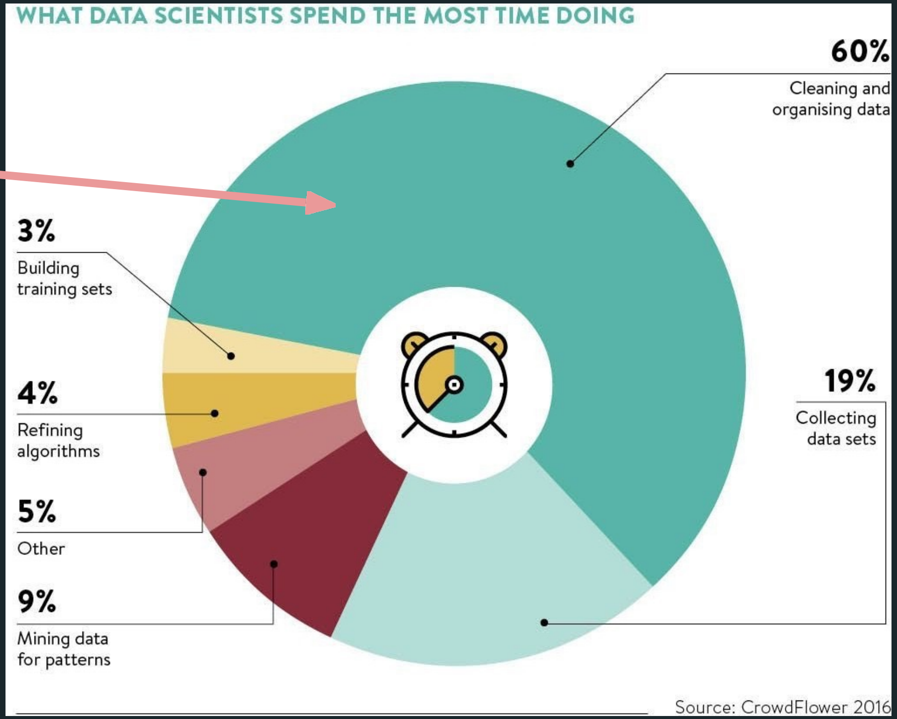
?

ℝ

Salida esperada / Predicción

Data preprocessing

Alrededor del **60% del tiempo** de los científicos de datos es empleado aquí, **preparando los datos para el modelado.**



Source: CrowdFlower 2016

Fuente: Forbes (2016).

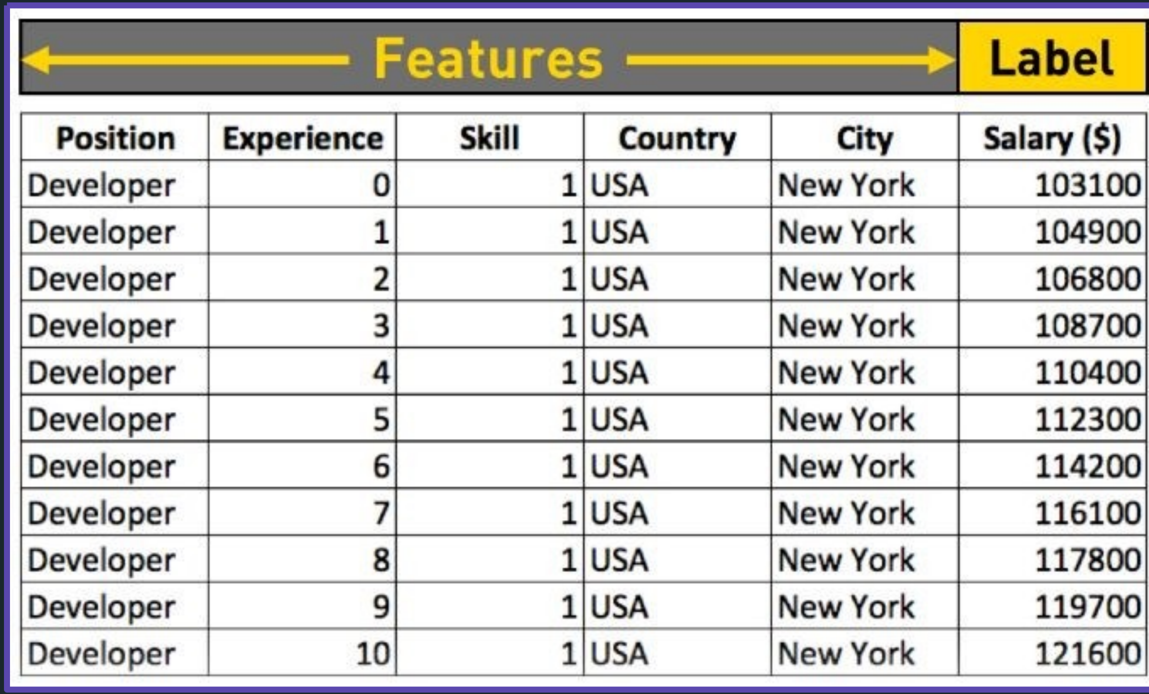
Paso que requiere mucho tiempo!!!

Definiciones

Dataset:

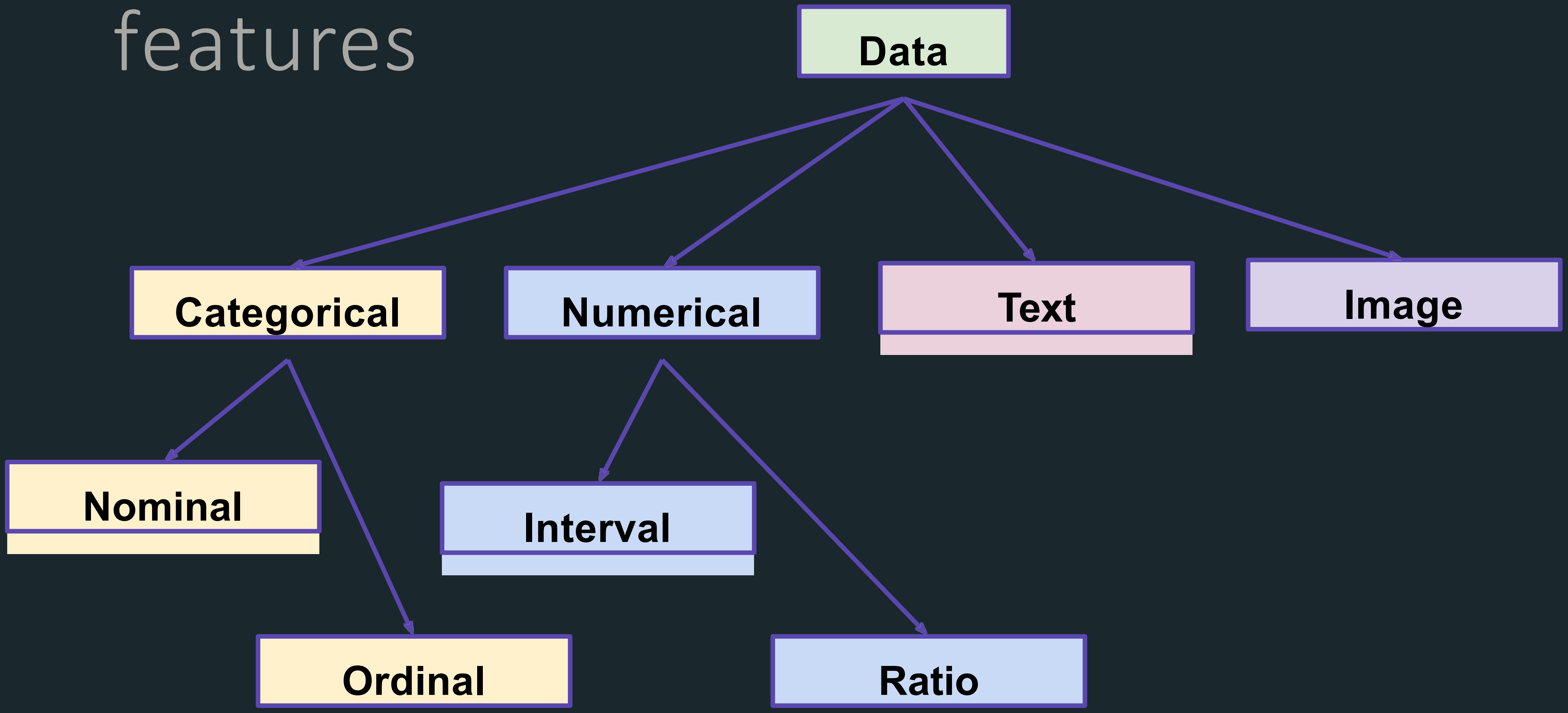
- Colección de datos.
  - **Ejemplos:** registros de interacciones, eventos, observaciones.
- Descritos mediante una serie de características o features.
  - **Ejemplos**: la masa de un objeto físico o el momento en que ocurrió un evento, etc.

Feature vector



Features					Label
Position	**Experience**	**Skill**	**Country**	**City**	**Salary ($)**
Developer	0	1	USA	New York	103100
Developer	1	1	USA	New York	104900
Developer	2	1	USA	New York	106800
Developer	3	1	USA	New York	108700
Developer	4	1	USA	New York	110400
Developer	5	1	USA	New York	112300
Developer	6	1	USA	New York	114200
Developer	7	1	USA	New York	116100
Developer	8	1	USA	New York	117800
Developer	9	1	USA	New York	119700
Developer	10	1	USA	New York	121600

Definiciones: Tipos de features



Tipos de features : Categorical

- Características cuyos valores se toman de un **conjunto definido de valores.**
  - Ejemplos?…. En el chat :D

Nominal

- Variables categóricas **sin un orden implícito.**
- **Ejemplo:** Los colores de un carro: negro, morado, rosa.

Ordinal

- Variables categóricas **con un orden natural implícito.**
- **Ejemplo:** Los tamaños de la ropa: chico, mediano, grande.

Tipos de features : Numerical

- Características representadas por **números cuyos valores son continuos o discretos.**
  - Ejemplos?…. En el chat :D

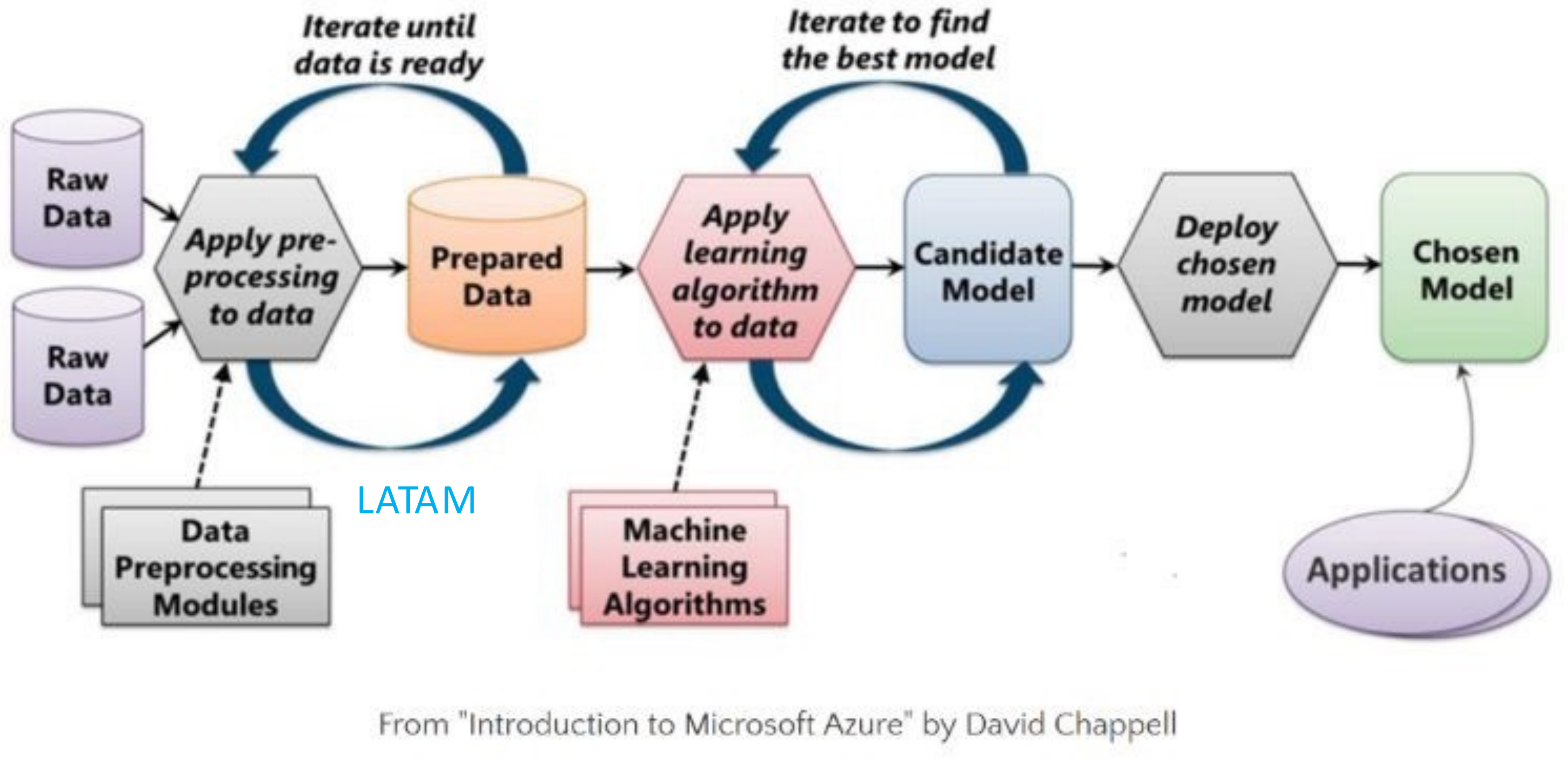
Interval

- Con una unidad de medida definida.
- Representa valores como 0 y menores que 0.
- **Ejemplo:** temperatura en Celsius

Ratio

- Con una unidad de medida definida.
- Representa valores de 0 y mayores a 0.
- **Ejemplo:** estatura y peso.

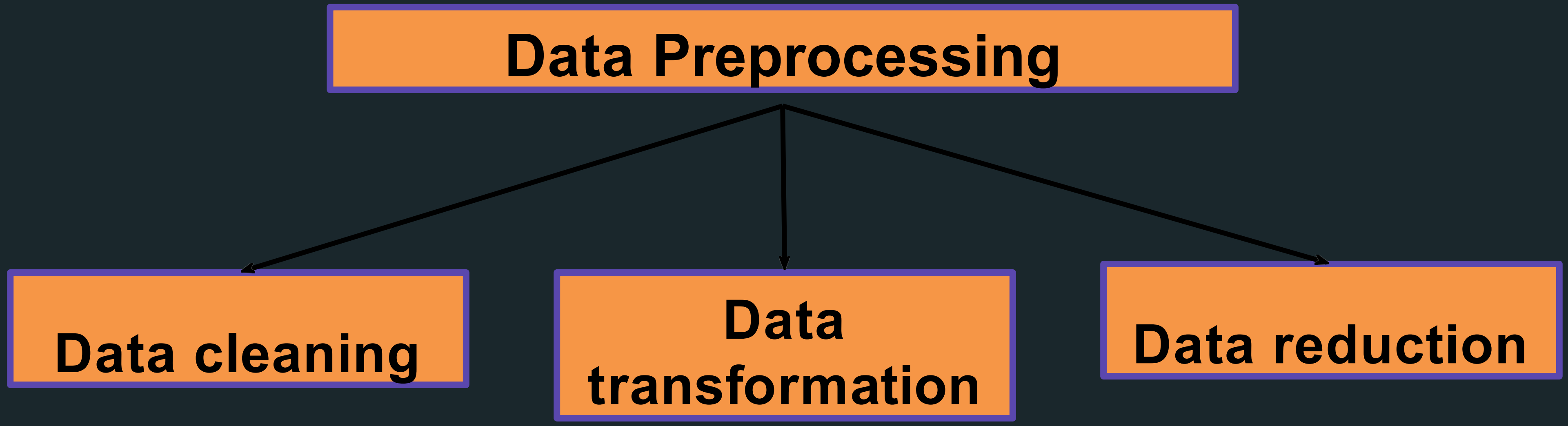
Metodología de Ciencia de Datos.


Iterate until data is ready
Iterate to find the best model
Raw Data
Raw Data
Apply pre-processing to data
Prepared Data
Apply learning algorithm to data
Candidate Model
Deploy chosen model
Chosen Model
Data Preprocessing Modules
LATAM
Machine Learning Algorithms
Applications

From "Introduction to Microsoft Azure" by David Chappell

Saturdays.AI

Data Preprocesing



Data Preprocessing

Data cleaning

Data transformation

Data reduction

LATAM

CONCLUSIONES!!!