

INTRODUCCIÓN A LA CIENCIA DE DATOS



LINKS

≡ [kaggle](#)

 Search

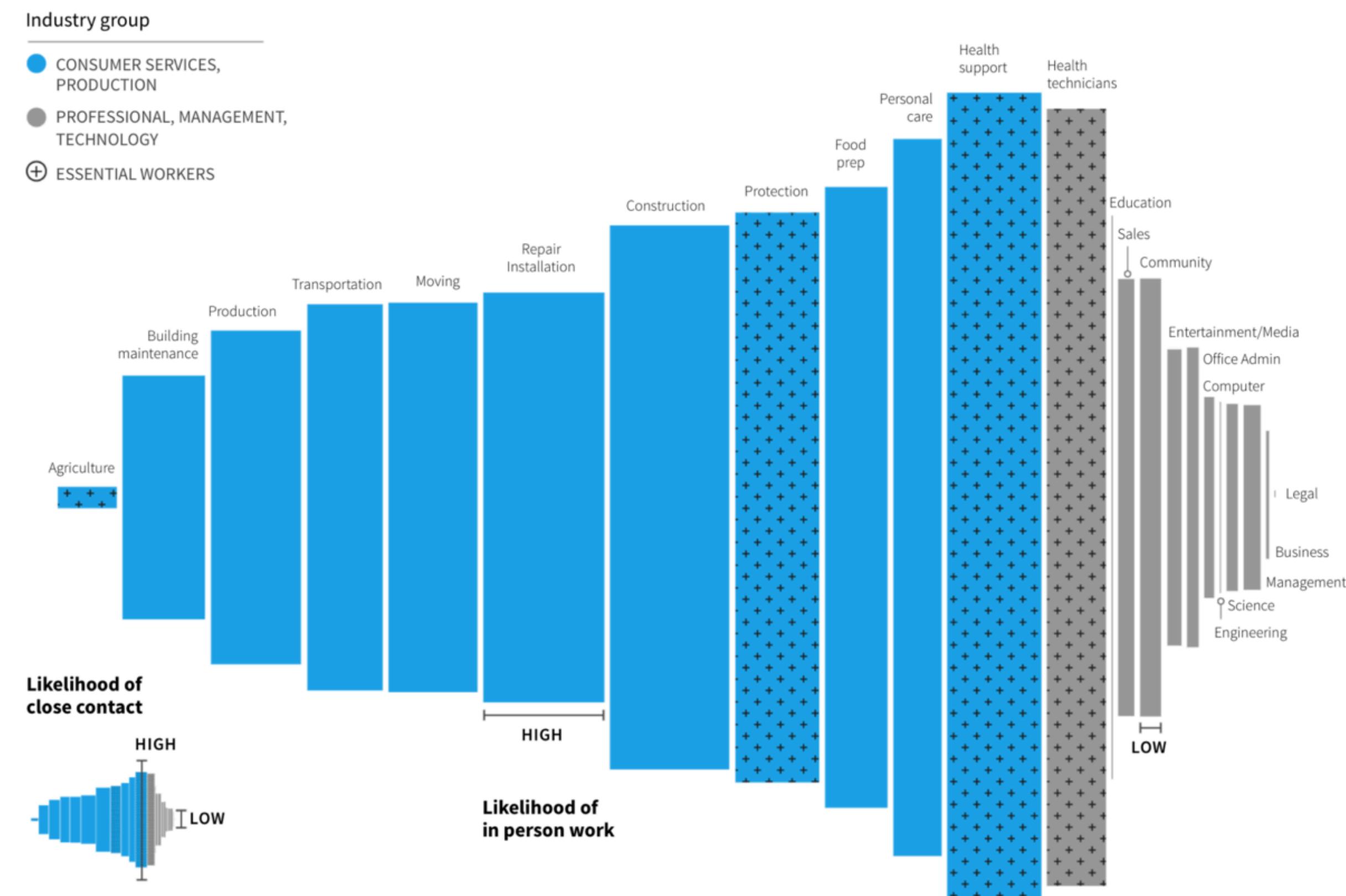
[Sign In](#) [Register](#)

-  [Home](#)
-  [Compete](#)
-  [Data](#)
-  [Notebooks](#)
-  [Discuss](#)
-  [Courses](#)
-  [More](#)

[Active](#) [Completed](#) [InClass](#) [All Categories ▾](#) [Default Sort ▾](#)

	OSIC Pulmonary Fibrosis Progression Predict lung function decline Featured • 2 months to go • Code Competition • 788 Teams	\$55,000
	SIIM-ISIC Melanoma Classification Identify melanoma in lesion images Featured • 7 days to go • 3257 Teams	\$30,000
	Google Landmark Retrieval 2020 Given an image, can you find all of the same landmarks in a dataset? Research • 7 days to go • Code Competition • 485 Teams	\$25,000
	Cornell Birdcall Identification Build tools for bird population monitoring Research • a month to go • Code Competition • 738 Teams	\$25,000
	Google Landmark Recognition 2020 Label famous (and not-so-famous) landmarks in images Research • 2 months to go • Code Competition • 155 Teams	\$25,000
	Halite by Two Sigma Collect the most halite during your match in space Featured • a month to go • Simulation Competition • 935 Teams	Swag

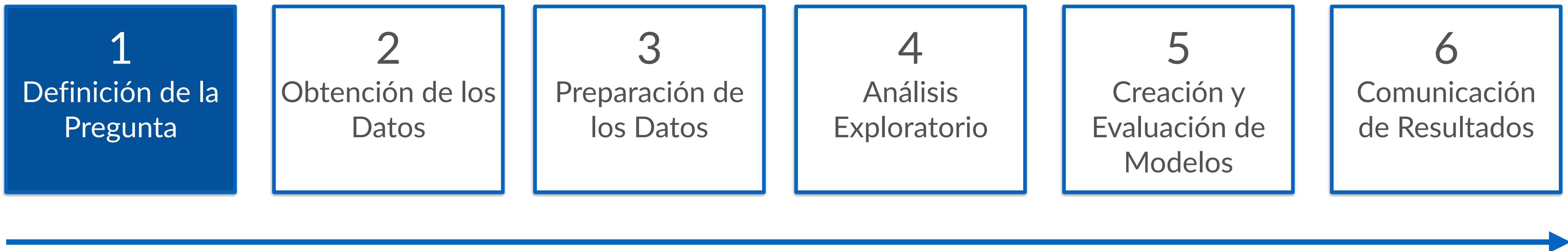
Remote work and industry



QUIZ

TIPOS DE PREGUNTAS

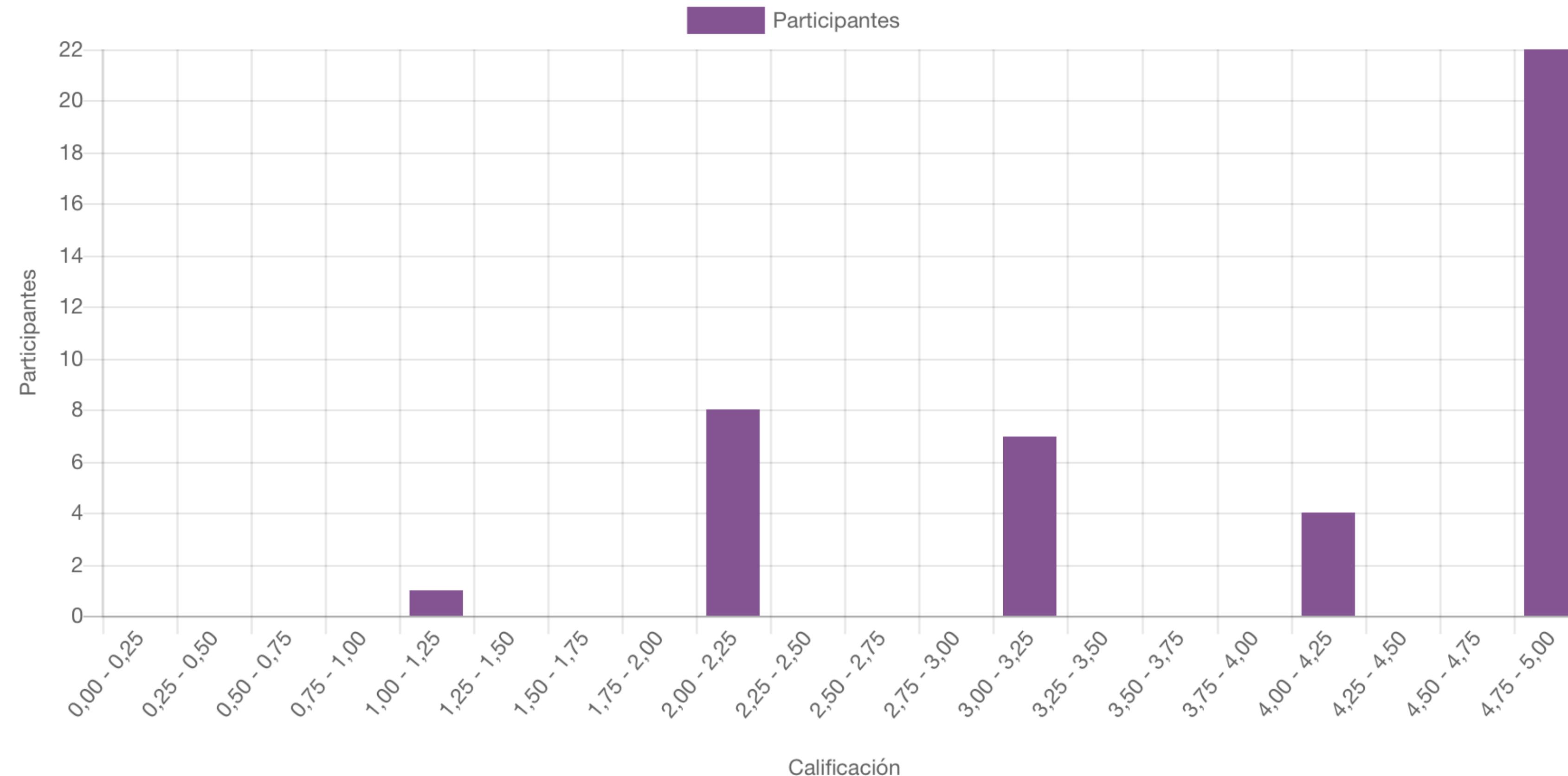
Data Science Work Flow



Distribución de Notas

Pregunta 01

Gráfico de barras del número de estudiantes que alcanzan los rangos de calificación



Pregunta 1

Quiz

Cuántos alumnos entraron a Moodle antes de comenzar el curso?

Seleccione una:

- a. Causal
- b. Inferencial
- c. Descriptiva
- d. Exploratoria
- e. Predictiva

Pregunta 2

Quiz

Cuánto se va a vender el mes que viene del producto A?

Seleccione una:

- a. Predictiva
- b. Causal
- c. Inferencial
- d. Exploratoria
- e. Descriptiva

Pregunta 3

Quiz

Hacer ejercicio cardiovascular ayuda a bajar de peso?

Seleccione una:

- a. Causal
- b. Exploratoria
- c. Predictiva
- d. Inferencial
- e. Descriptiva

Pregunta 4

Quiz

Cuántos enfermos de COVID-19 son personal del sistema de Salud?

Seleccione una:

- a. Causal
- b. Exploratoria
- c. Predictiva
- d. Descriptiva
- e. Inferencial

Pregunta 5

Quiz

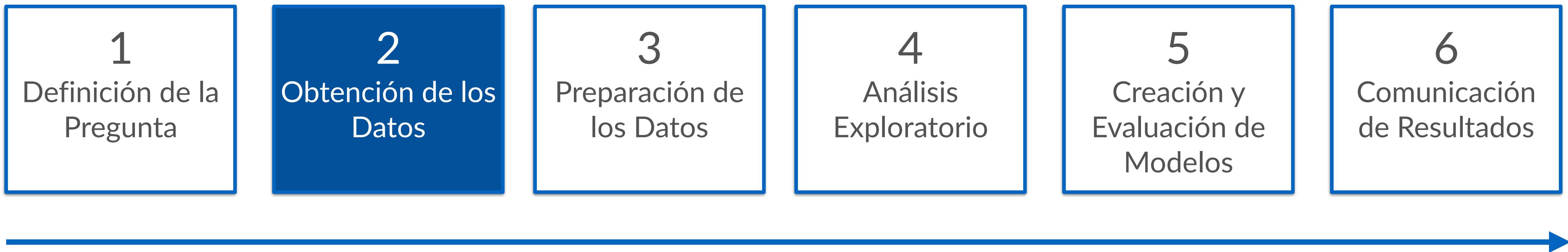
Cuántos montevideanos piensan a votar a Laura Raffo en las próximas elecciones departamentales?

Seleccione una:

- a. Descriptiva.
- b. Causal
- c. Exploratoria
- d. Inferencial
- e. Predictiva

DATA SCIENCE WORK FLOW

Data Science Work Flow



Definición

Obtención de Datos

La recopilación de datos es el **proceso de recopilar y medir información sobre variables específicas**, que luego permite responder preguntas relevantes y evaluar resultados.

El **objetivo** de la recopilación de datos es capturar evidencia de calidad que permita que el análisis conduzca a la formulación de respuestas convincentes y creíbles a las preguntas que se han planteado.

Wikipedia

Integridad de los datos

Obtención de Datos

1. Quality assurance - acciones previas a la recolección de los datos

- Identificación de la persona responsable.
- Lista de variables necesarias para recolectar.
- Descripción de los instrumentos de recopilación de datos.
- Instrucciones para usar, hacer ajustes y calibrar el equipo de recolección de datos.
- Mecanismo predeterminado para documentar los cambios en los procedimientos que ocurren durante la investigación.

2. Quality control - acciones durante y luego de la recolección de los datos.

Primeros Pasos

Obtención de Datos

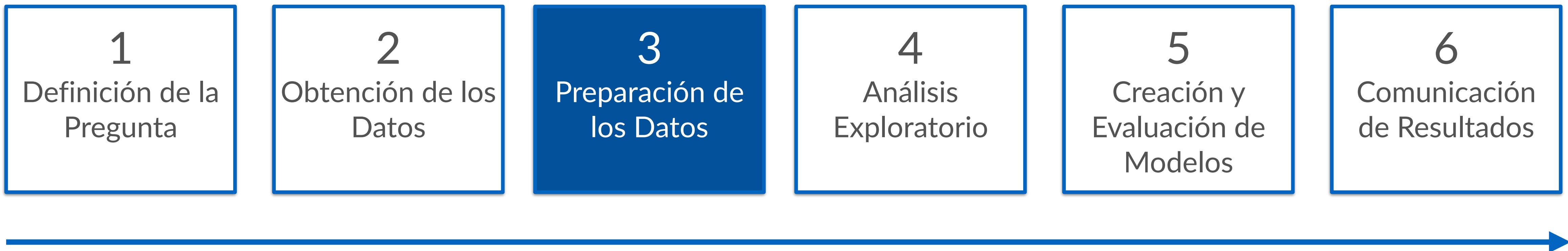
1. Qué tipo de datos preciso?

2. Cómo puedo acceder a ellos?

1. Qué tipo de datos preciso?

- Internos
- Externos (datos de mercado, competencia, clientes, reviews)
 - Públicos o privados.

Data Science Work Flow

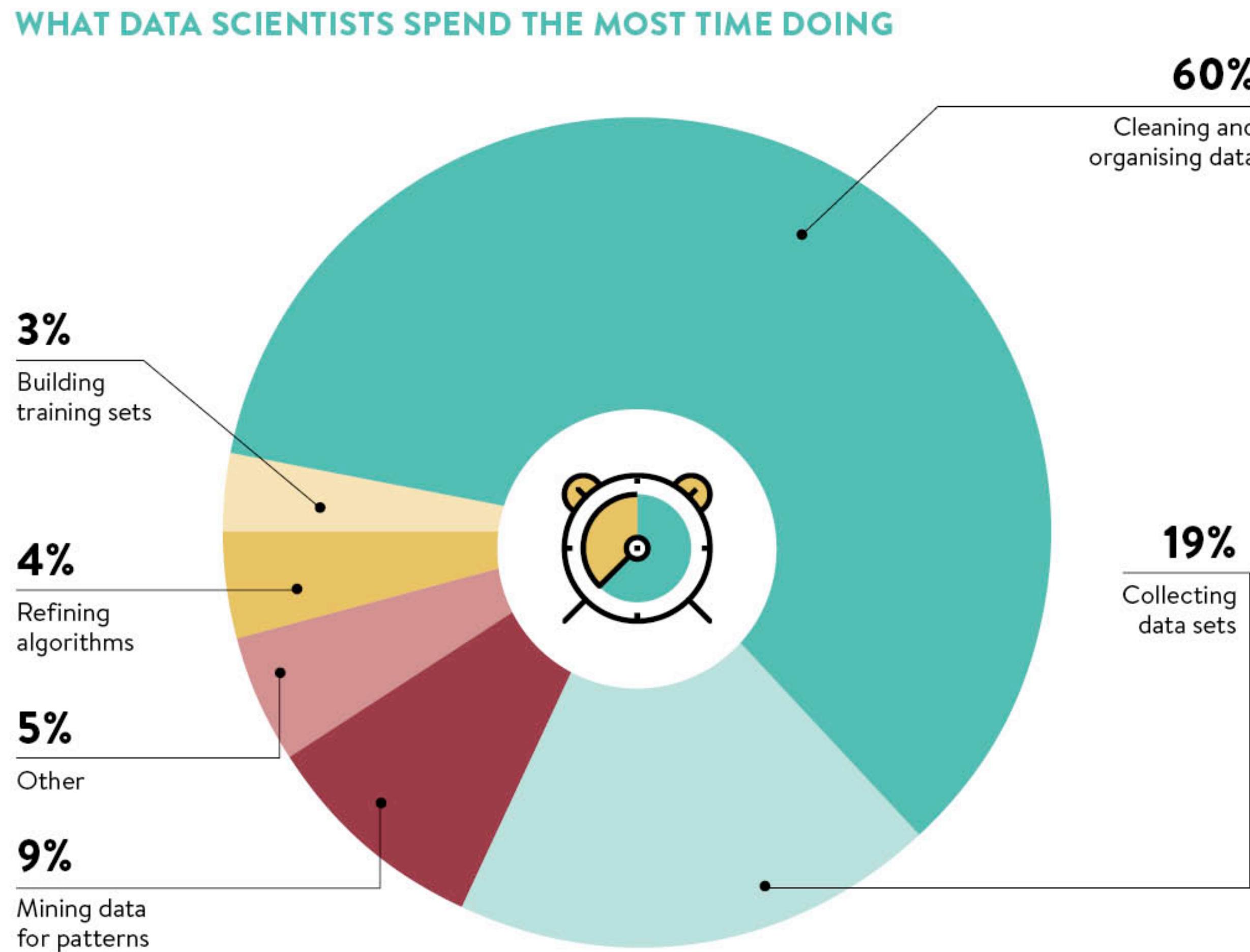


Cleaning the Data



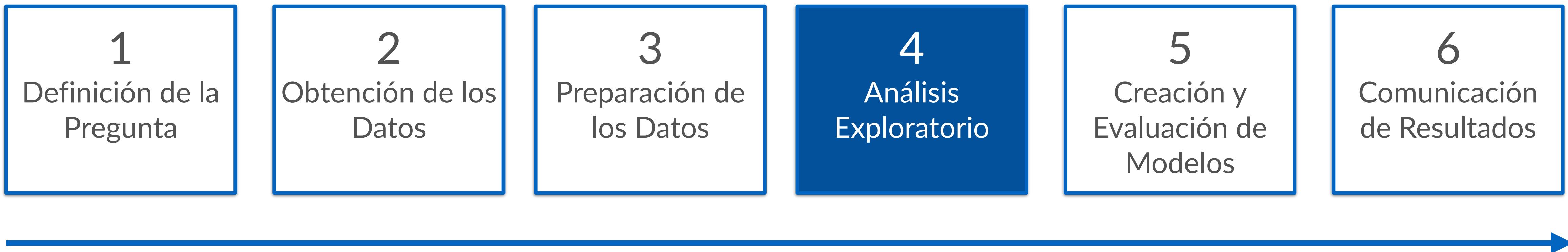
Tiempo dedicado

Preparación de los datos



1. Eliminar observaciones no deseadas.
2. Corregir errores estructurales.
3. Filtrar valores atípicos no deseados.
4. Tratar datos faltantes.

Data Science Work Flow



Qué es?

Análisis Exploratorio

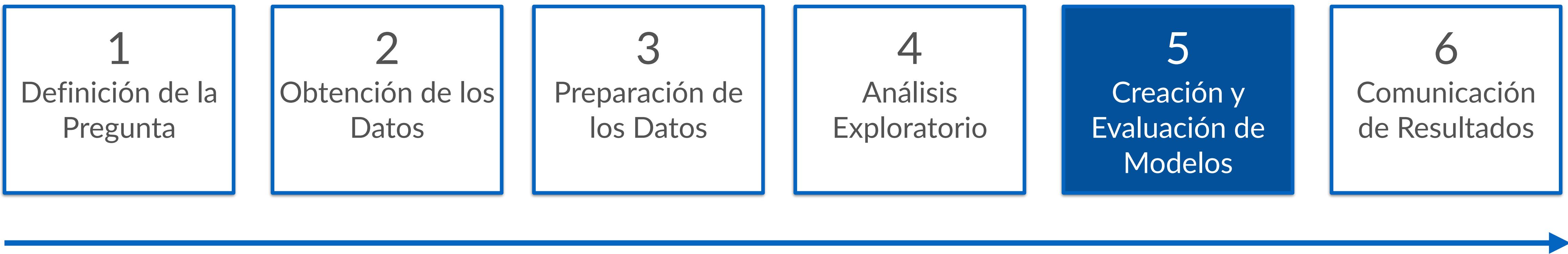
- A menudo es el **primer paso** en el análisis de datos, implementado antes de que se aplique cualquier técnica estadística formal.
- Se trata de **conocer los datos**, adquirir cierta familiaridad antes de comenzar a extraer conocimientos de ellos.
- El análisis exploratorio de datos se refiere al proceso crítico de realizar investigaciones iniciales sobre los datos para descubrir patrones, detectar anomalías, probar hipótesis y verificar suposiciones con la ayuda de estadísticas resumidas y representaciones gráficas.

Herramientas

Análisis Exploratorio

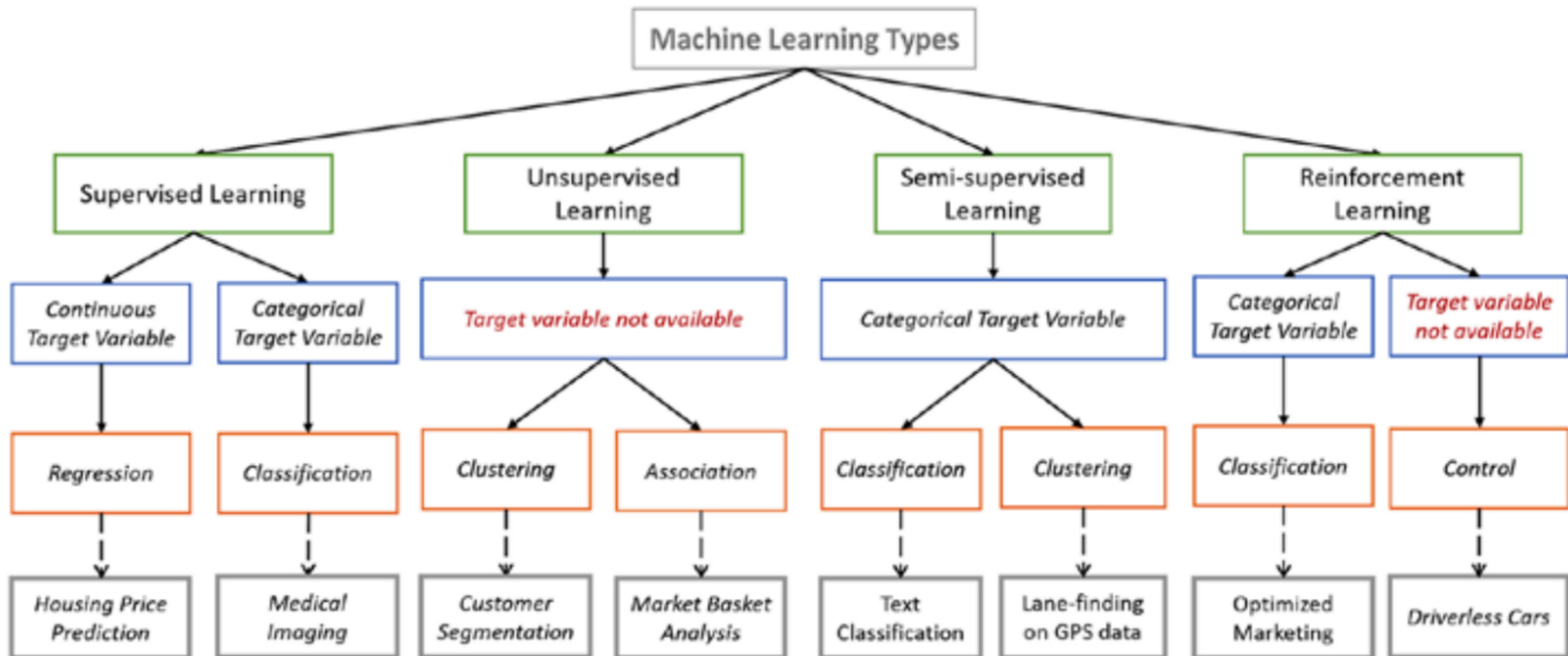
- **Resúmenes numéricos.**
 - Tendencia central, dispersión, sesgo, correlación.
- **Visualizaciones de datos.**
 - Histogramas, box-plots, heatmaps, barcharts.

Data Science Work Flow

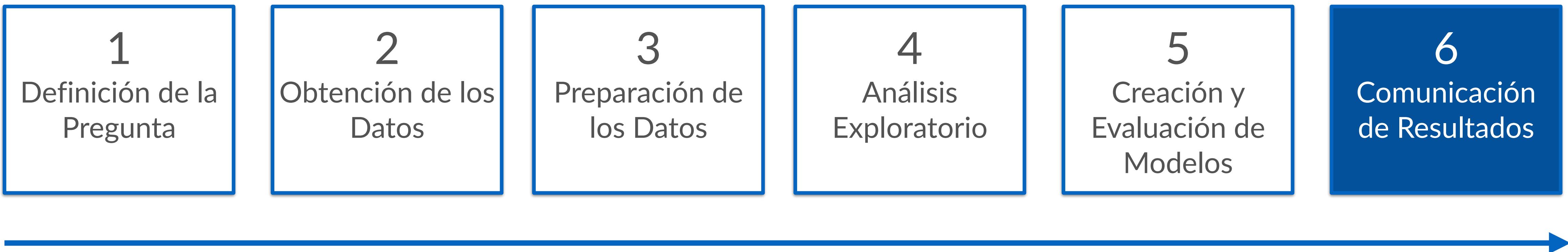


Tipos de Modelos

Creación y Evaluación de Modelos



Data Science Work Flow



- **Conocimientos complejos => buena historia genera mayor aceptación.**
- Definición de la audiencia: quién es nuestro público.
- Contexto: establecer el contexto para que la audiencia comprenda la relevancia de la historia.
- Transmisión: usar una narrativa interesante y emocionante para transmitir el mensaje.
- Resumir: reiterar los aspectos más destacados o la “moraleja de la historia” al final.

- Importancia de la comprensión y el conocimiento empresarial.
- Resaltar el impacto y la oportunidad.
- Identificar correctamente el plan de acción.

TIPOS DE DATOS

Definición

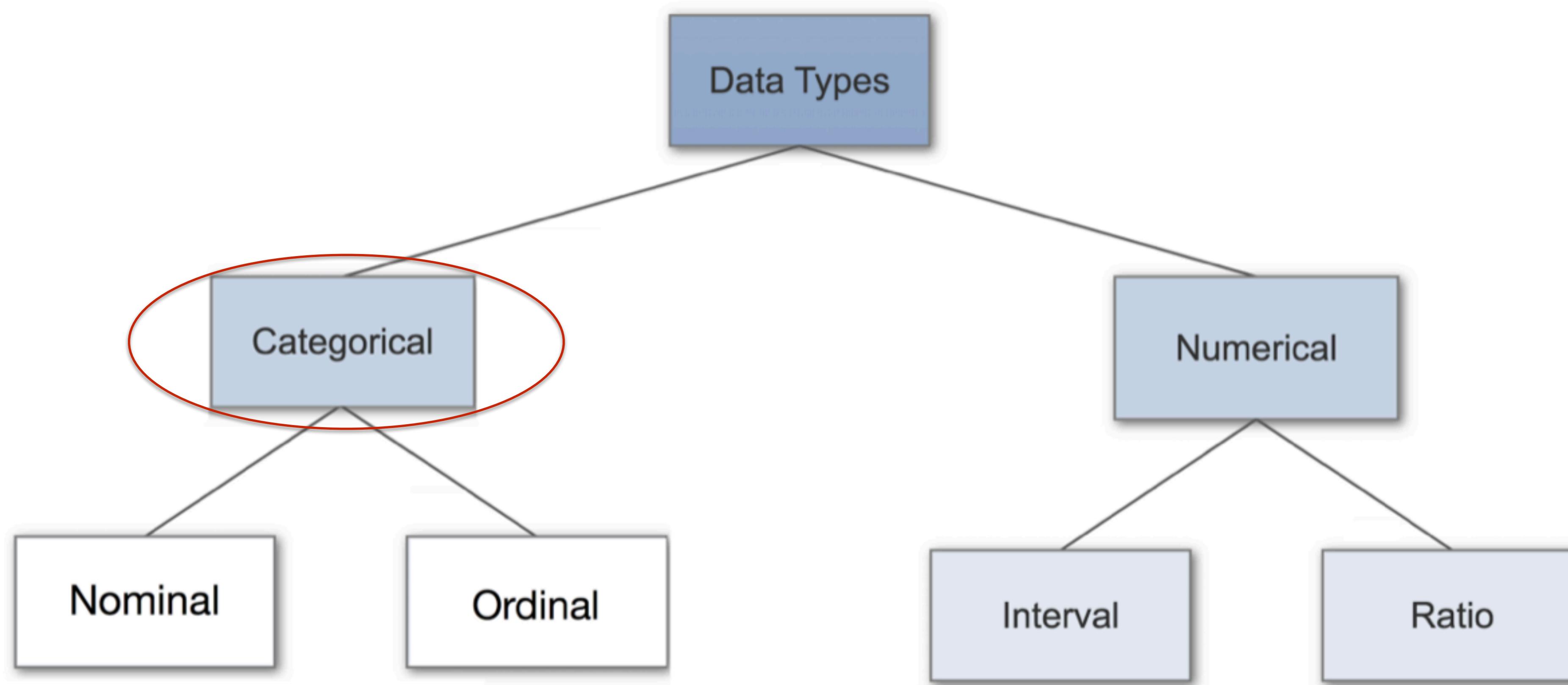
Datos

Un dato es una representación simbólica de un atributo o variable cuantitativa o cualitativa.

Wikipedia

Categóricos o Numéricos

Tipos de Datos



Categóricos o Numéricos

Tipos de Datos

Datos categóricos

- Representan características de los datos.
- Pueden tomar valores numéricos: 0 mujer - 1 hombre, pero no tienen un significado matemático.

Nominales

Are you married?

Yes

No


Dicotómico

Solamente dos categorías

What languages do you speak?

Englisch

French

German

Spanish

Ordinales

What Is Your Educational Background?

1 - Elementary

2 - High School

3 - Undegraduate

4 - Graduate



- El orden importa.
- La diferencia entre los valores no es medible.
- Otros ejemplos: happiness, customer satisfaction.

Categóricos o Numéricos

Tipos de Datos

Attribute Type	Description	Examples	Operations
Categorical (Qualitative)	Nominal		
	Ordinal		
Numeric (Quantitative)	Interval		
	Ratio Scale		

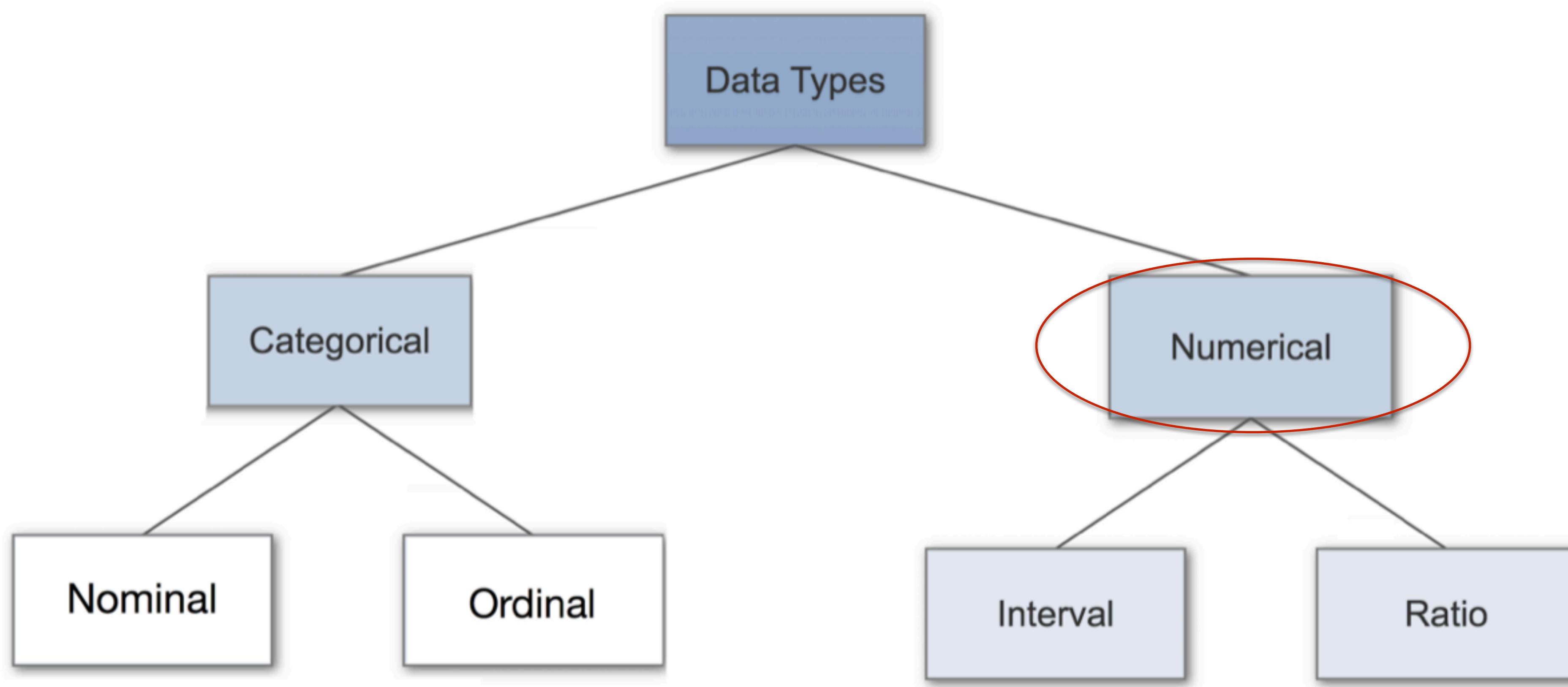
Categóricos o Numéricos

Tipos de Datos

Attribute Type	Description	Examples	Operations
Categorical (Qualitative)	Nominal	No natural ordering; only distinguishing information ($=, \neq$)	favorite color, gender, ID number
	Ordinal		
Numeric (Quantitative)	Interval		
	Ratio Scale		

Categóricos o Numéricos

Tipos de Datos



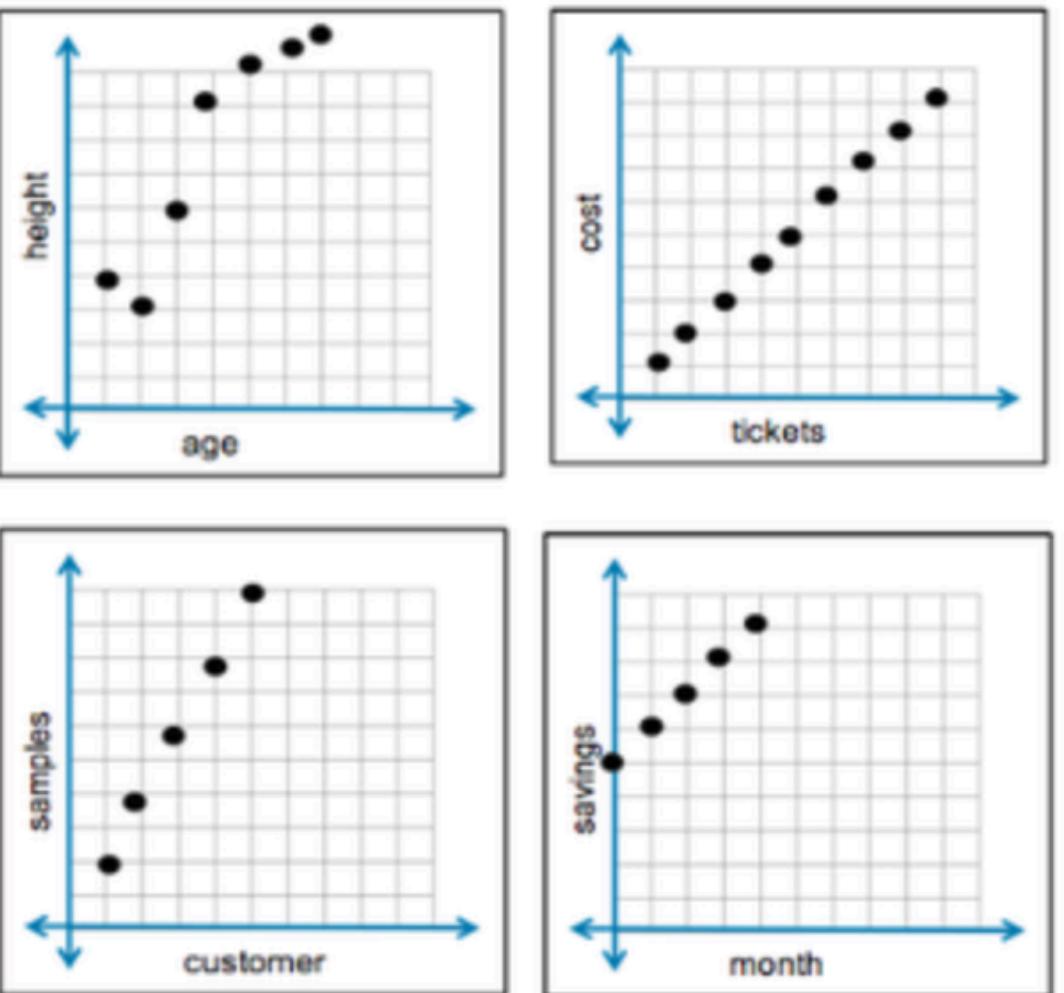
Categóricos o Numéricos

Tipos de Datos

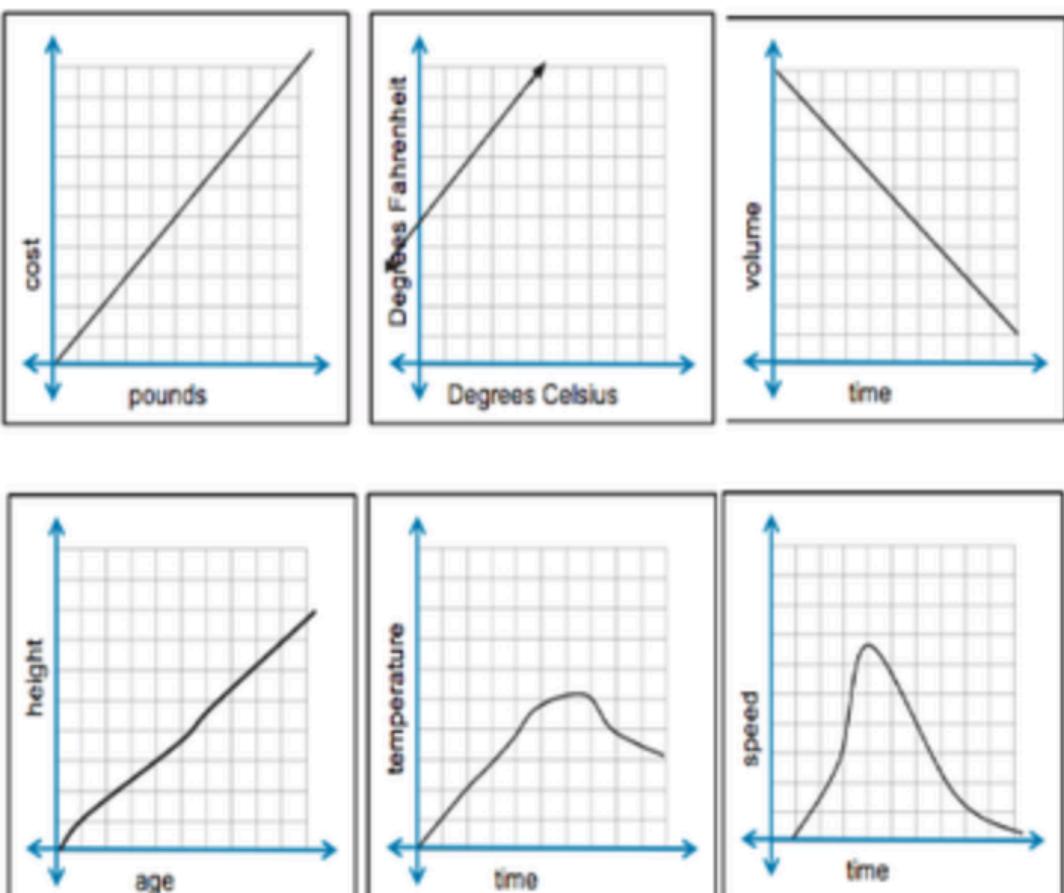
Datos numéricas

- **Discretos:** se dice que un conjunto de datos es discreto si los valores que pertenecen al conjunto son distintos y separados (valores no conectados).
 - Ejemplo: número de alumnos en la clase.
 - Pueden contarse.
 - Números enteros.
 - Solamente pueden tomar determinados valores en un intervalo.
- **Continuos:** se dice que un conjunto de datos es continuo si los valores que pertenecen al conjunto pueden tomar cualquier valor dentro de un intervalo finito o infinito.
 - Ejemplo: tiempo que demoran los alumnos en terminar el examen.

Discretos



Continuos



Categóricos o Numéricos

Tipos de Datos

Attribute Type	Description	Examples	Operations	
Categorical (Qualitative)	Nominal	No natural ordering; only distinguishing information ($=, \neq$)	favorite color, gender, ID number	mode, entropy, correlation, χ^2 test
	Ordinal	categories can be ordered ($>, <$)	t-shirt size, attitude	median, percentiles, rank correlation
Numeric (Quantitative)	Interval	numerical, equidistant measure, differences are meaningful ($+, -$)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, t and F tests
	Ratio Scale			

- **Medición:** los datos de intervalo se miden utilizando una escala de intervalo, que no solo muestra el orden y la dirección, sino que también muestra la diferencia exacta en el valor. Por ejemplo, las marcas de un termómetro o una regla son equidistantes, en palabras más simples miden la misma distancia entre las dos marcas.
- **Diferencia de intervalo:** las distancias entre cada valor en los datos de intervalo son iguales. Por ejemplo, la diferencia entre 10 cm y 20 cms es igual a 20 cms y 30 cms.
- **Cálculo:** en datos de intervalo, uno puede sumar o restar valores pero no puede dividir ni multiplicar. Casi todos los análisis estadísticos son aplicables al calcular datos de intervalo, media, moda, mediana, etc.
- **Punto cero:** el punto cero absoluto es arbitrario, lo que significa que una variable se puede medir incluso si tiene un valor negativo, como la temperatura puede ser -10 por debajo de cero, pero la altura no puede estar por debajo de cero.

Categóricos o Numéricos

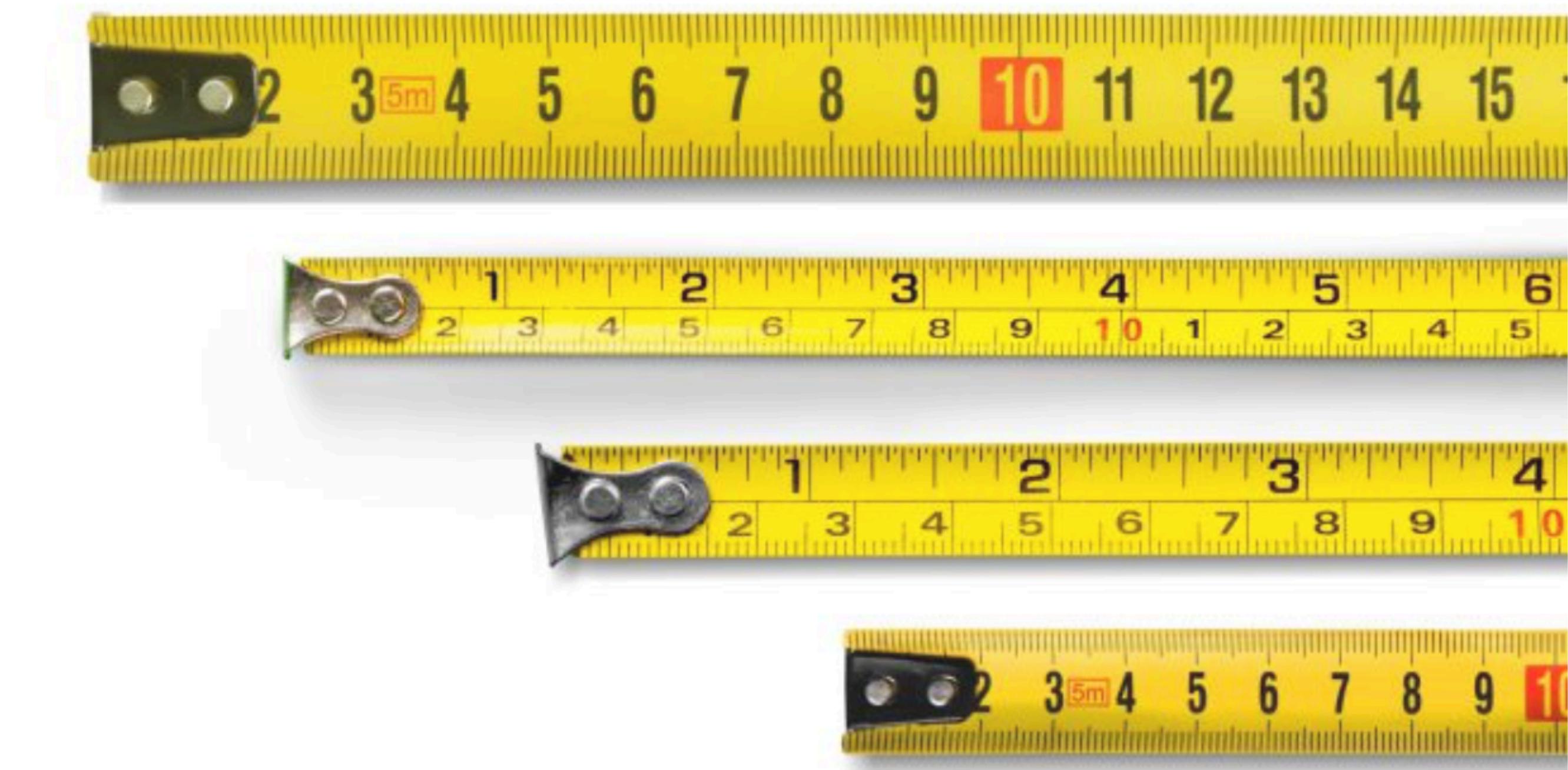
Tipos de Datos

Attribute Type	Description	Examples	Operations	
Categorical (Qualitative)	Nominal	No natural ordering; only distinguishing information ($=, \neq$)	favorite color, gender, ID number	mode, entropy, correlation, χ^2 test
	Ordinal	categories can be ordered ($>, <$)	t-shirt size, attitude	median, percentiles, rank correlation
Numeric (Quantitative)	Interval	numerical, equidistant measure, differences are meaningful (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, t and F tests
	Ratio Scale	Interval Scale, both difference and ratios are meaningful (*, /)	age, temperature in Kelvin, True ratios exist (e.g. I swim half as fast as Phelps)	geometric/harmonic mean, percent variation

Ratio Scale

Datos Numéricos

- **Tiene cero absoluto.** La característica del punto cero hace que sea relevante o significativo decir, "un objeto tiene el doble de largo que el otro" o "es el doble de largo".
- La escala **no tiene un números negativos.**
- Las variables se pueden **sumar, restar, multiplicar y dividir** sistemáticamente.
- **Ejemplos:** altura, peso, horas de TV por día.



Primarios o Secundarios

Tipos de Datos

Datos primarios

- Obtenidos directamente de la fuente.

PROS

- Se adapta a las necesidades del análisis.
- Más precisa. No está sujeta a sesgos.
- Generalmente está actualizada.
- Se pueden elegir los métodos de recolección.

CONS

- Más costosa.
- Requiere más tiempo.
- Proceso más complejo.

Datos secundarios

- Recolectada en el pasado para otros propósitos.

PROS

- Más accesible.
- Bajo costo.
- Requiere menos tiempo.
- Facilita estudios longitudinales sin tener que esperar.

CONS

- Hay que evaluar la veracidad de los datos.
- Hay que adaptarla a las necesidades actuales.
- En casos puede estar desactualizada.

Estructurado o No Estructurado

Tipos de Datos

Datos
Estructurados

Datos Semi
Estructurados

Datos No
Estructurados

Datos Estructurados

Tipos de Datos

Datos Estructurados

Los datos estructurados están altamente organizados y formateados de manera que se pueden buscar fácilmente en bases de datos relacionales.

Ejemplo: Relational Databases (mySQL, PostgreSQL, etc.), Planillas de Excel.

A	B	C	D	E	F	G	H	I
Number	GivenName	MiddleInitial	Surname	Gender	StreetAddress	City	State	ZipCode
1	Bruce	R	Bloch	male	3151 Ferrell Street	Argyle	MN	56713
2	Marie	E	Humphreys	female	3062 Bond Street	Woonsocket	RI	2895
3	Sylvia	H	Carter	female	1481 Lakeland Terrace	Westland	MI	48185
4	William	E	Bentz	male	3318 Briercliff Road	New York	NY	10011
5	Shelly	R	Preston	female	3592 Todd's Lane	San Antonio	TX	78212
6	Chad	P	Henry	male	3553 Grant Street	Tyler	TX	75702
7	David	L	Richardson	male	1289 Metz Lane	Marlton	NJ	8053
8	Stephen	A	Pond	male	4316 Bridge Avenue	Lafayette	LA	70503
9	Jenny	P	Thomas	female	2941 Harron Drive	Baltimore	MD	21202
10	William	V	Fries	male	4300 Tanglewood Road	Jackson	MS	39201
11	Julio	D	Bessette	male	4177 Lauren Drive	Madison	WI	53718
12	Jerry	J	Nicholas	male	2722 Elk Street	Irvine	CA	92718
13	Thomas	A	Hunter	male	4112 Stadium Drive	Franklin	MA	2038
14	Edmund	C	Chagoya	male	3685 Essex Court	Brattleboro	VT	5301
15	David	E	Meador	male	1215 Stratford Drive	Kona	HI	96740
16	Joan	L	Mayfield	female	3137 Pin Oak Drive	Whittier	CA	90603

Estructurado o No Estructurado

Tipos de Datos

Datos Semi Estructurados

Los datos semiestructurados son una forma de datos estructurados que no obedece a la estructura tabular de los modelos de datos asociados con bases de datos relacionales u otras formas de tablas de datos.

Ejemplo: CSV, XML, JSON, NoSQL.

```
If DataGridView5.RowCount > 0 Then
    Dim save As New SaveFileDialog
    save.Filter = "ARCHIVO XML (*.xml)|*.xml"
    save.FileName = "nombre_de_archivo_generado.xml"
    If save.ShowDialog = Windows.Forms.DialogResult.OK Then
        Dim fs As System.IO.FileStream
        Dim xtw As System.Xml.XmlTextWriter
        miDataTable.TableName = "items"
        fs = New System.IO.FileStream(save.FileName, IO FileMode.Create)
        xtw = New System.Xml.XmlTextWriter(fs, System.Text.Encoding.Unicode)
        xtw.WriteProcessingInstruction("xml", "version='1.0'")
        xtw.WriteProcessingInstruction("mso-application", "progid='Excel.Sheet'")
        miDataTable.WriteXml(xtw)
        xtw.Close()
        MsgBox("La exportación ha sido realizado, debes abrir el archivo .XML con Excel")
    End If
Else
    MsgBox("No existen datos a exportar", MsgBoxStyle.Critical, "Atención")
    Exit Sub
End If
```

Estructurado o No Estructurado

Tipos de Datos

Datos No Estructurados

Los datos no estructurados no tienen un formato u organización predefinida, lo que hace que sean mucho más difíciles recopilar, procesar y analizar.

Son los datos con mayor crecimiento.

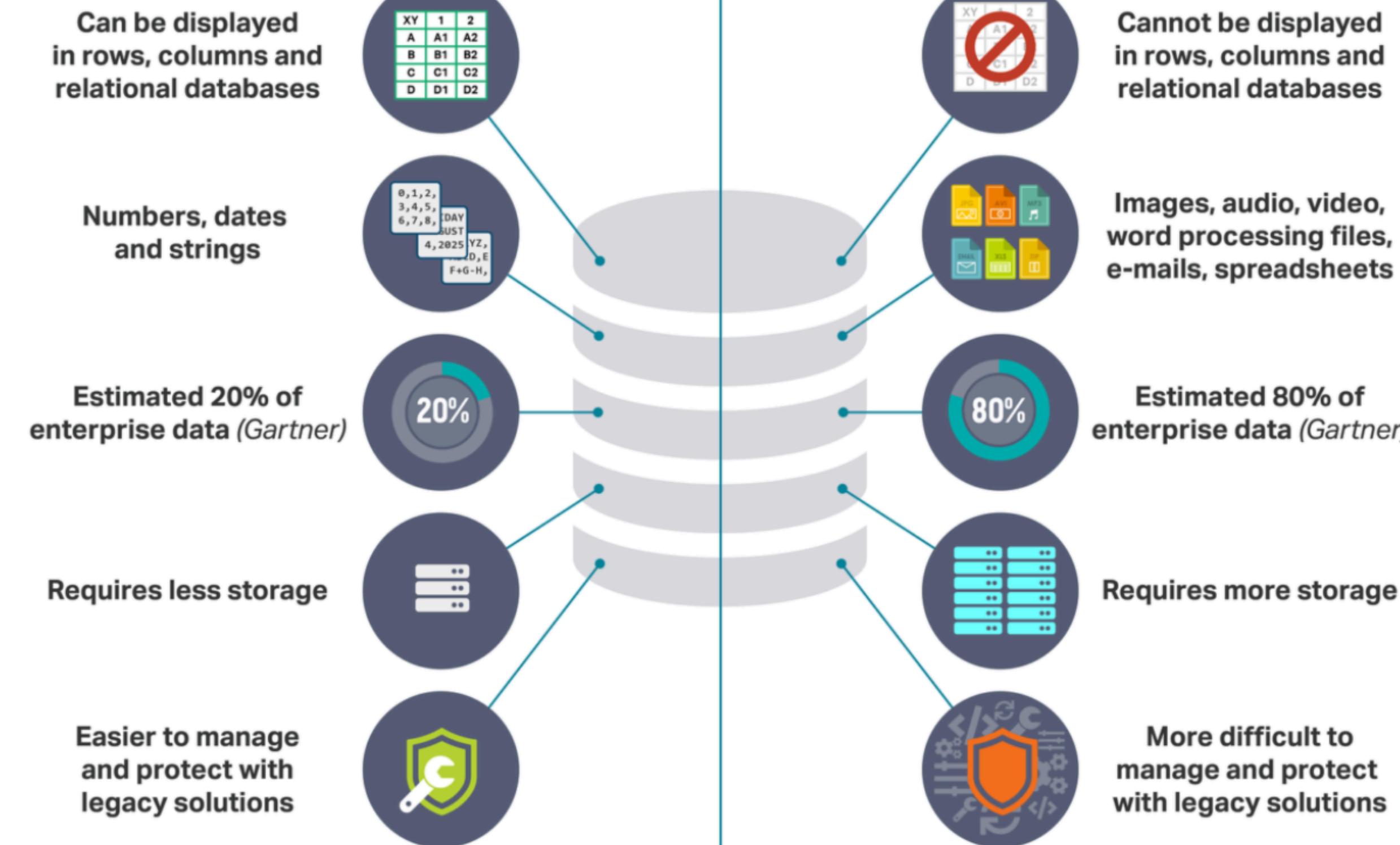
Ejemplo: text, multimedia.



Estructurados o No Estructurados

Tipos de Datos

Structured Data vs Unstructured Data



TIDYING THE DATA

Objetivo

Tidying the Data

El objetivo de crear un conjunto de datos ordenado es obtener los datos en un formato que pueda compartirse y analizarse fácilmente.

- El trabajo de convertir los datos de forma cruda a una forma directamente analizable es el **primer paso** de cualquier análisis de datos.
- Es importante ver los datos sin procesar y comprender los pasos en el proceso de procesamiento.
- Los pasos del procesamiento deben estar bien documentados y estandarizados.

Componentes

Tidying the Data

Componentes de un dataset ordenado

1. Los datos en bruto.
2. Un conjunto de datos ordenado.
3. Un code book que describe cada variable y sus valores en el conjunto de datos ordenado.
4. Una receta explícita y exacta para ir del paso 1 al 2.

Raw Data

Tidying the Data

Características

- Los datos brutos están en el formato correcto si no se ejecutó ningún software en los datos, no se manipuló ninguno de los números, no eliminó ningún dato y no se resumieron los datos de ninguna manera.
- Es fundamental que se incluya la forma más cruda de los datos utilizados en el entregable.
- Ejemplos: el archivo Excel sin formato con 10 hojas de trabajo que la compañía que te contrató te envió. Los datos JSON que surgen de la API de Twitter.
- Los datos brutos son relativos, hay procesamientos inevitables antes de adquirir los datos.

```
array (size=1)
  0 =>
    object(stdClass)[2]
      public 'location' => string 'Poulton-Le-Fylde, England' (length=25)
      public 'profile_sidebar_fill_color' => string 'C0DFEC' (length=6)
      public 'verified' => boolean false
      public 'name' => string 'James Mallison' (length=14)
      public 'time_zone' => string 'London' (length=6)
      public 'follow_request_sent' => boolean false
      public 'default_profile_image' => boolean false
      public 'lang' => string 'en' (length=2)
      public 'notifications' => boolean false
      public 'favourites_count' => int 123
      public 'id' => int 70512422
      public 'profile_background_color' => string '022330' (length=6)
      public 'status' =>
        object(stdClass)[3]
          public 'in_reply_to_status_id' => int 345499014994726913
          public 'in_reply_to_screen_name' => string 'omer_janjua' (length=11)
          public 'source' => string 'web' (length=3)
          public 'created_at' => string 'Fri Jun 14 11:20:59 +0000 2013' (length=30)
          public 'entities' =>
            object(stdClass)[4]
              ...
              public 'favorited' => boolean false
              public 'place' => null
              public 'geo' => null
              public 'id' => int 345501116127137792
              public 'contributors' => null
              public 'truncated' => boolean false
              public 'id_str' => string '345501116127137792' (length=18)
              public 'in_reply_to_user_id' => int 205270138
              public 'in_reply_to_status_id_str' => string '345499014994726913' (length=18)
              public 'retweet_count' => int 0
              public 'text' => string '@omer_janjua Already played that game. Ninjas win.' (length=50)
              public 'coordinates' => null
              public 'in_reply_to_user_id_str' => string '205270138' (length=9)
              public 'retweeted' => boolean false
              public 'is_translator' => boolean false
              public 'profile_background_image_url' => string 'http://a0.twimg.com/images/themes/theme15/bg.png' (length=48)
              public 'geo_enabled' => boolean true
              public 'profile_link_color' => string '0084B4' (length=6)
              public 'profile_background_image_url_https' => string 'https://twimg0-a.akamaihd.net/images/themes/theme15/bg.png' (length=58)
              public 'following' => boolean false
              public 'utc_offset' => int 0
              public 'profile_image_url' => string 'http://a0.twimg.com/profile_images/3777720764/17dec790156f89fc990c3696db6cb2ff_normal.jpeg' (length=90)
        
```

Características

- Cada variable debe estar en una columna.
- Cada observación debe estar en una fila diferente.
- Debe haber una tabla para cada "tipo" de variable.
- Si tiene varias tablas, deben incluir una columna en la tabla que les permite vincularse.

Tips

- Incluir los nombres de las variables en la primera fila del archivo.
- Nombres que seas entendibles. Ejemplo: AgeAtDiagnosis vs. AgeDX
- Evitar planillas de Excel con muchas hojas. Es mejor una hoja por archivo.

Code Book

Tidying the Data

Características

Las variables y medidas calculadas deben describirse con más detalle en el Code Book.

Como mínimo debe contener:

- Información sobre las variables, incluyendo sus unidades.
- Información sobre el método de resumen o agregación de las variables.
- Información sobre el diseño del estudio experimental.

Reproducible Data Science

Tidying the Data

Definición

Los proyectos reproducibles son aquellos que permiten a otros recrear y construir sobre el análisis, así como reutilizar y modificar fácilmente el código.

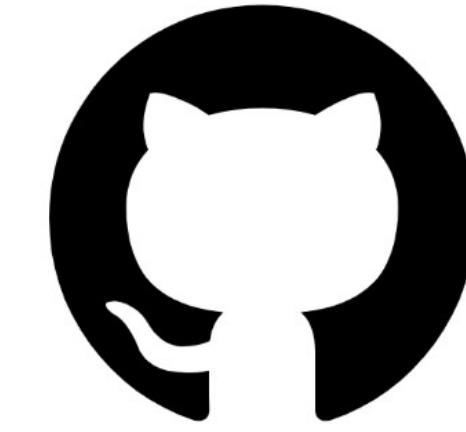
Relevancia

- La empresas puede solicitar al científico de datos que **repita el análisis con diferentes parámetros**. Si el código no se adapta fácilmente, esto impedirá cumplir con los nuevos requisitos dentro de un plazo razonable.
- Si un proyecto obtiene buenos resultados, la empresa querrá **producirlo a escala**. En la mayoría de las empresas, esto significa entregar el proyecto a un equipo de ingeniería para que lo implemente. El código de producción bien documentado hará que esta transición sea mucho más fluida.
- **La reproducibilidad genera confianza**, es más probable que las partes interesadas confíen en un modelo si pueden recorrer el análisis ellos mismos. El código bien probado también es más preciso y es menos probable que contenga errores de programación obvios.
- La reproducibilidad permite **compartir conocimientos** entre los científicos de datos y los aspirantes a científicos de datos en la empresa. La buena documentación permite a otros comprender las técnicas utilizadas y el código reproducible les permite construir y reutilizar partes del proyecto de su equipo.

Reproducible Data Science

Tidying the Data

- Idealmente un script.
- Jupyter Notebook, R Markdown.
- El input deben ser los datos brutos.
- El output deben ser los datos procesados.



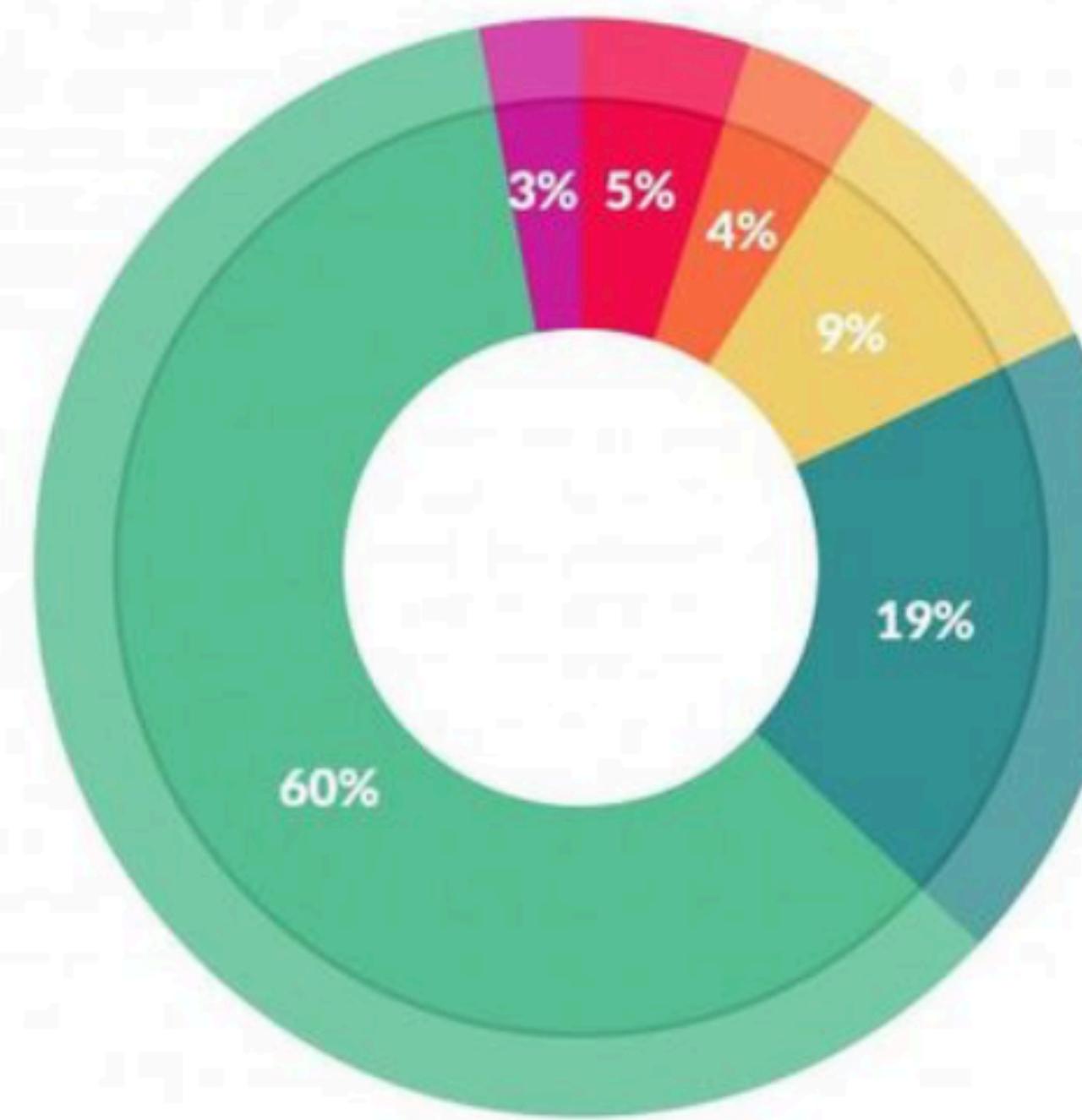
CLEANING THE DATA



Importancia

Better data beats fancier algorithms

- Garbage-in, garbage-out.
- La limpieza adecuada de los datos puede definir el éxito o el fracaso del proyecto.
- Los científicos de datos dedican el **60% de su tiempo a limpiar y organizar datos**.
- El 76% de los científicos de datos ven la preparación de datos como la parte menos agradable de su trabajo.



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

Importancia

Cleaning the data

Data quality is one of the most important problems in data management, since dirty data often leads to inaccurate data analytics results and incorrect business decisions.

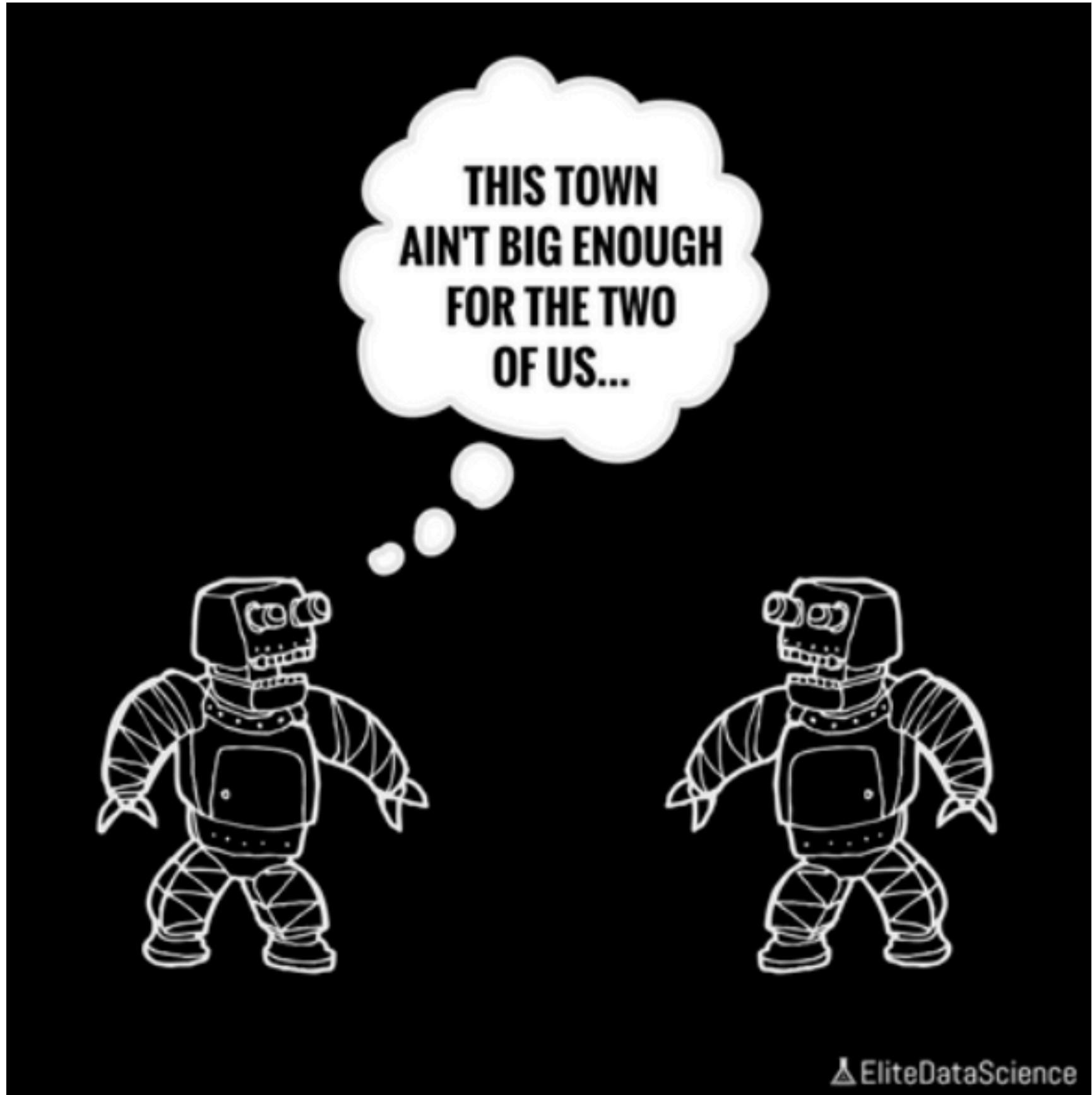
Eliminar duplicados

Cleaning the data

Las **observaciones duplicadas** surgen con mayor frecuencia durante la recopilación de datos.

Por ejemplo:

- Al combinar conjuntos de datos de múltiples lugares.
- Scrappear datos. Páginas de supermercados con productos repetidos en diferentes categorías.
- Recibir datos de clientes / otros departamentos.



Eliminar datos irrelevantes

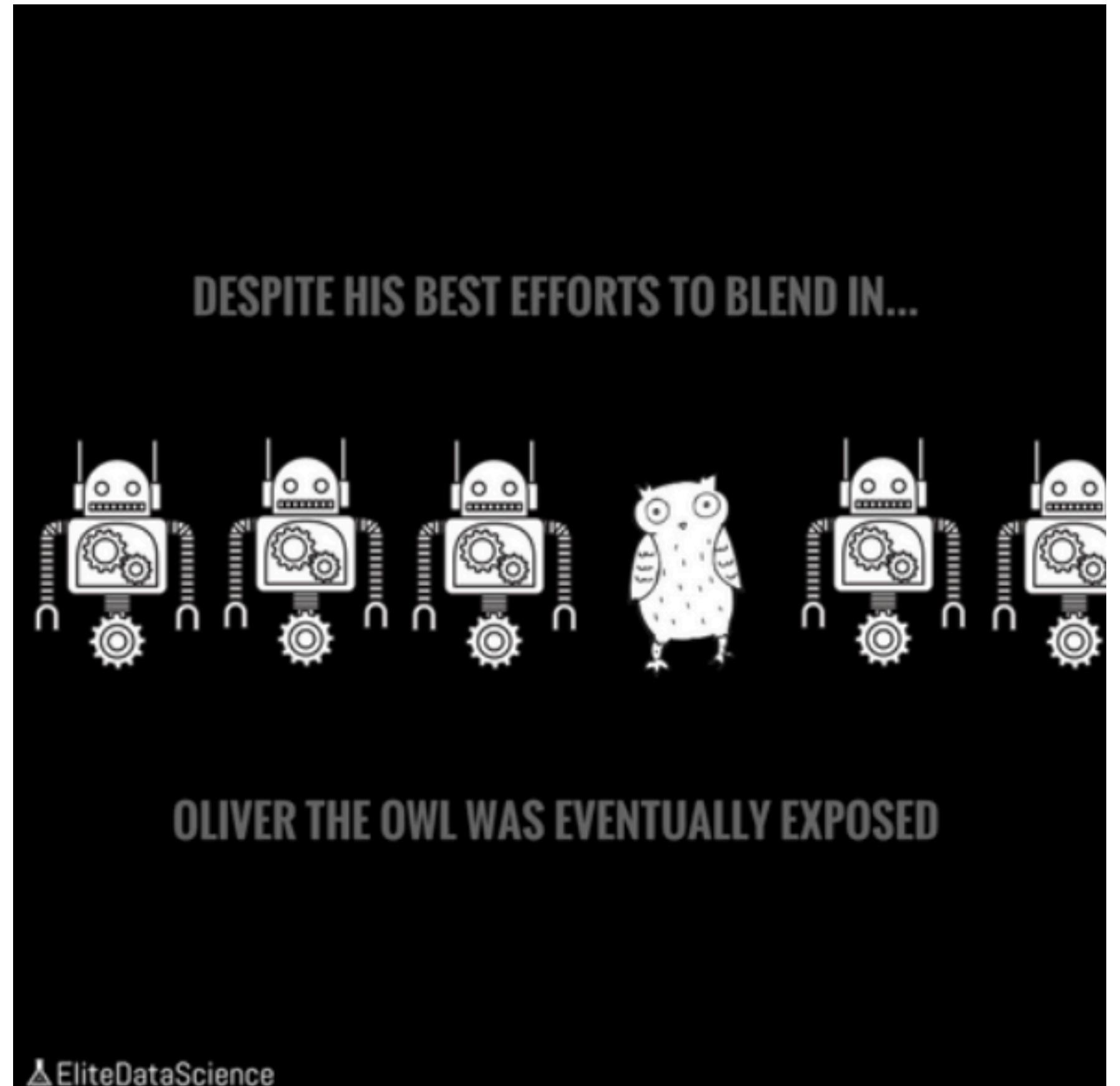
Cleaning the data

Las **observaciones irrelevantes** son aquellas que no se ajustan al problema específico que se está tratando de resolver.

- Buscar observaciones irrelevantes antes de realizar feature engineering puede ahorrar muchos problemas en el futuro.
- Es útil visualizar las variables categóricas y evaluar si hay clases que no deberían estar allí.

Por ejemplo:

- Bolsas de plástico en datos de venta de un supermercado.



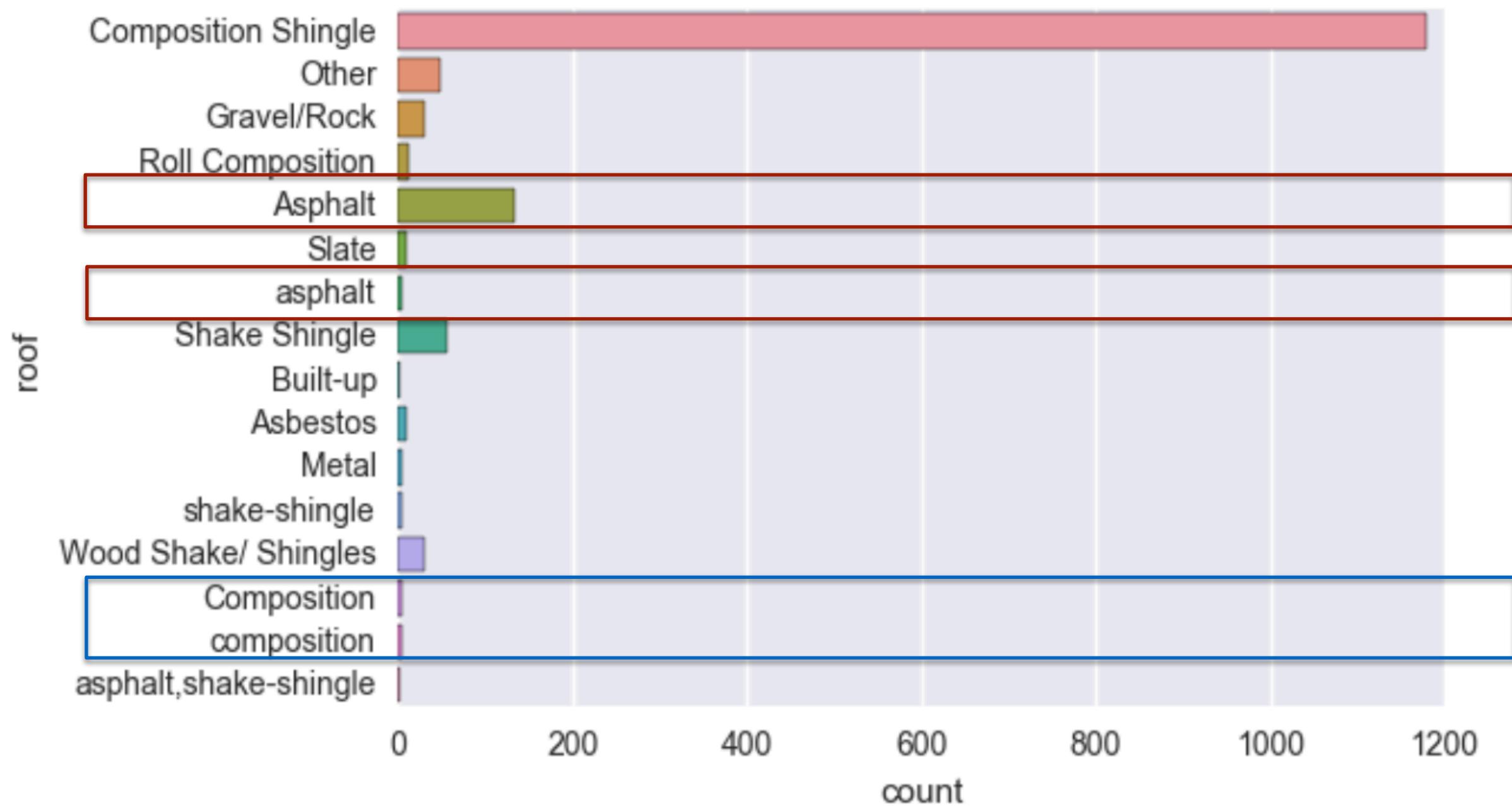
Corregir errores estructurales

Cleaning the data

Los errores estructurales son aquellos que surgen durante la medición, la transferencia de datos u otros tipos de "mala limpieza".

Por ejemplo:

- Verificar errores tipográficos o mayúsculas inconsistentes.
- Verificar variables categóricas con listado de valores únicos o gráficos de barras.

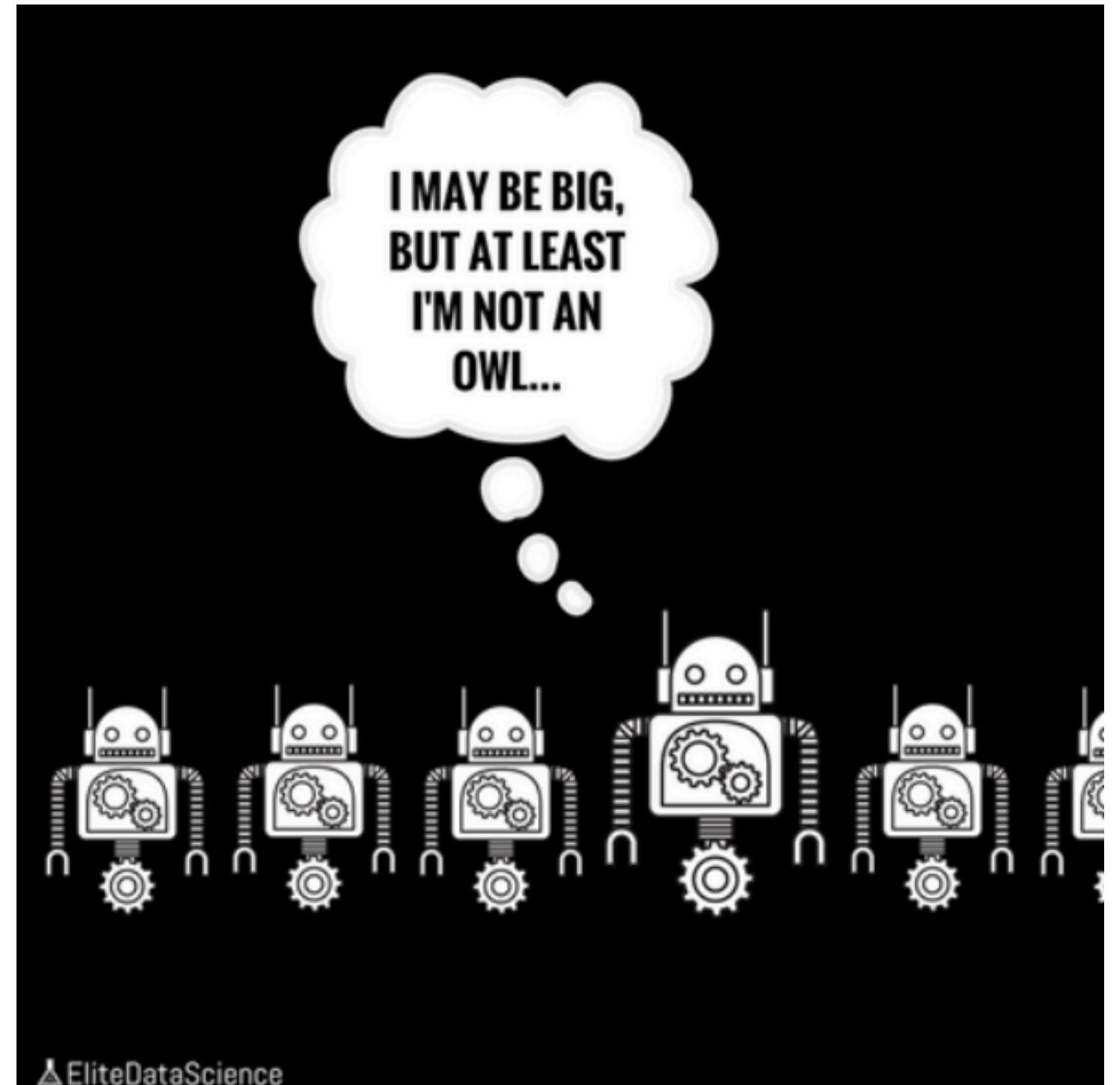


Outliers

Cleaning the data

Los valores atípicos pueden causar problemas con ciertos tipos de modelos. Por ejemplo, los modelos de regresión lineal son menos robustos a los valores atípicos que los modelos de árbol de decisión.

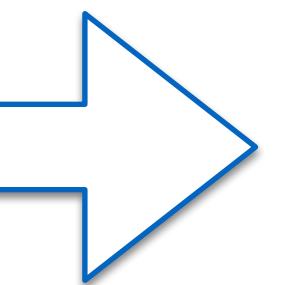
- En general, si se tiene un motivo legítimo para eliminar un valor atípico, ayudará al rendimiento de su modelo.
- Sin embargo, **los valores atípicos son inocentes hasta que se demuestre lo contrario.** Nunca se debe eliminar un valor atípico solo porque es un "gran número". Ese gran número podría ser muy informativo para el modelo.
- Debe haber una buena razón para eliminar un valor atípico, como mediciones sospechosas que probablemente no sean datos reales.



Outliers

Cleaning the data

To drop or not
to drop?



Analizar la naturaleza del
outlier.

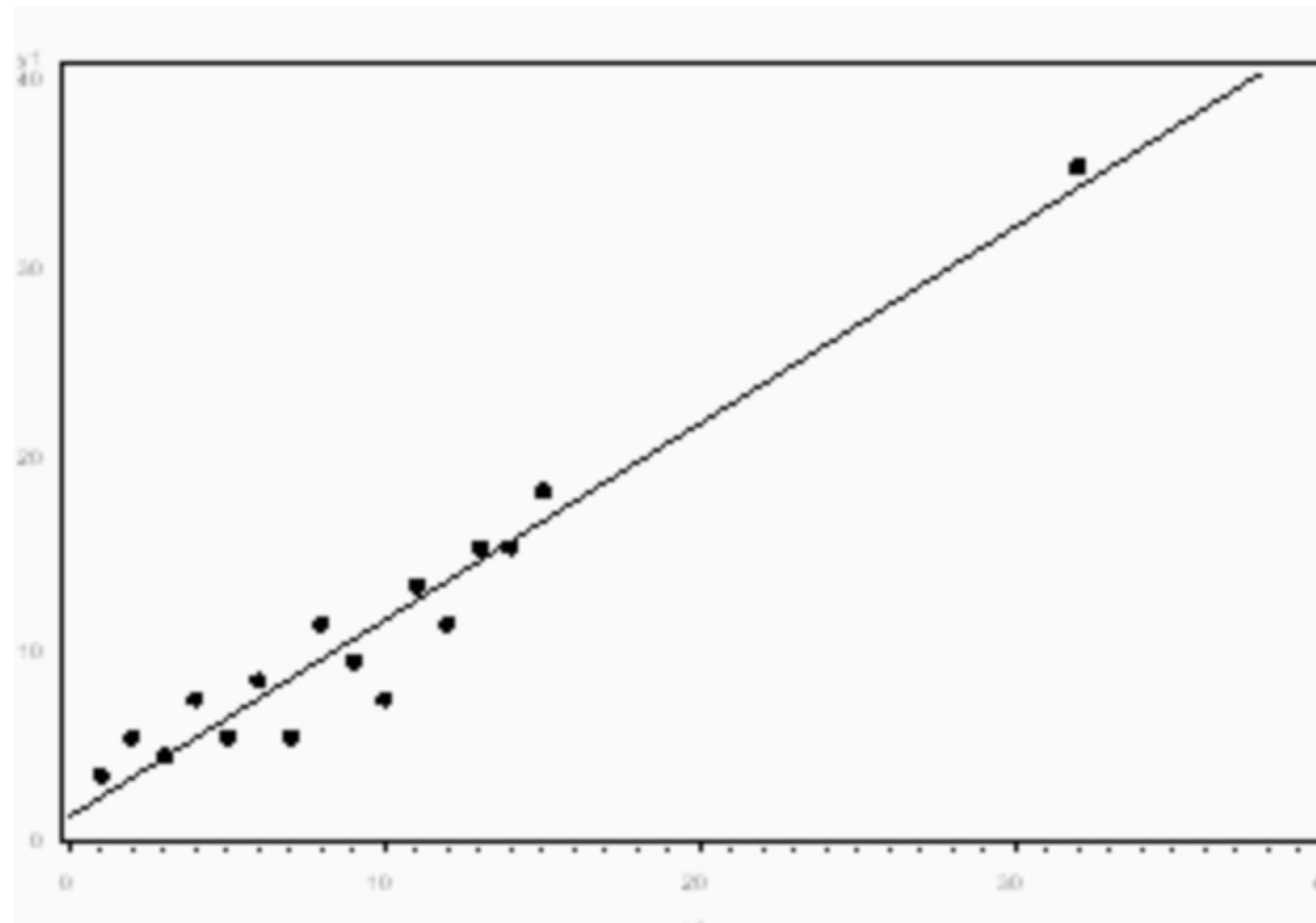
Outliers

Cleaning the data

1. El valor es claramente un error → Eliminar el dato.

Ejemplo: Variable: altura de personas / Valor: 11.6 m

2. No afecta los resultados → Eliminar el dato y dejar registro.

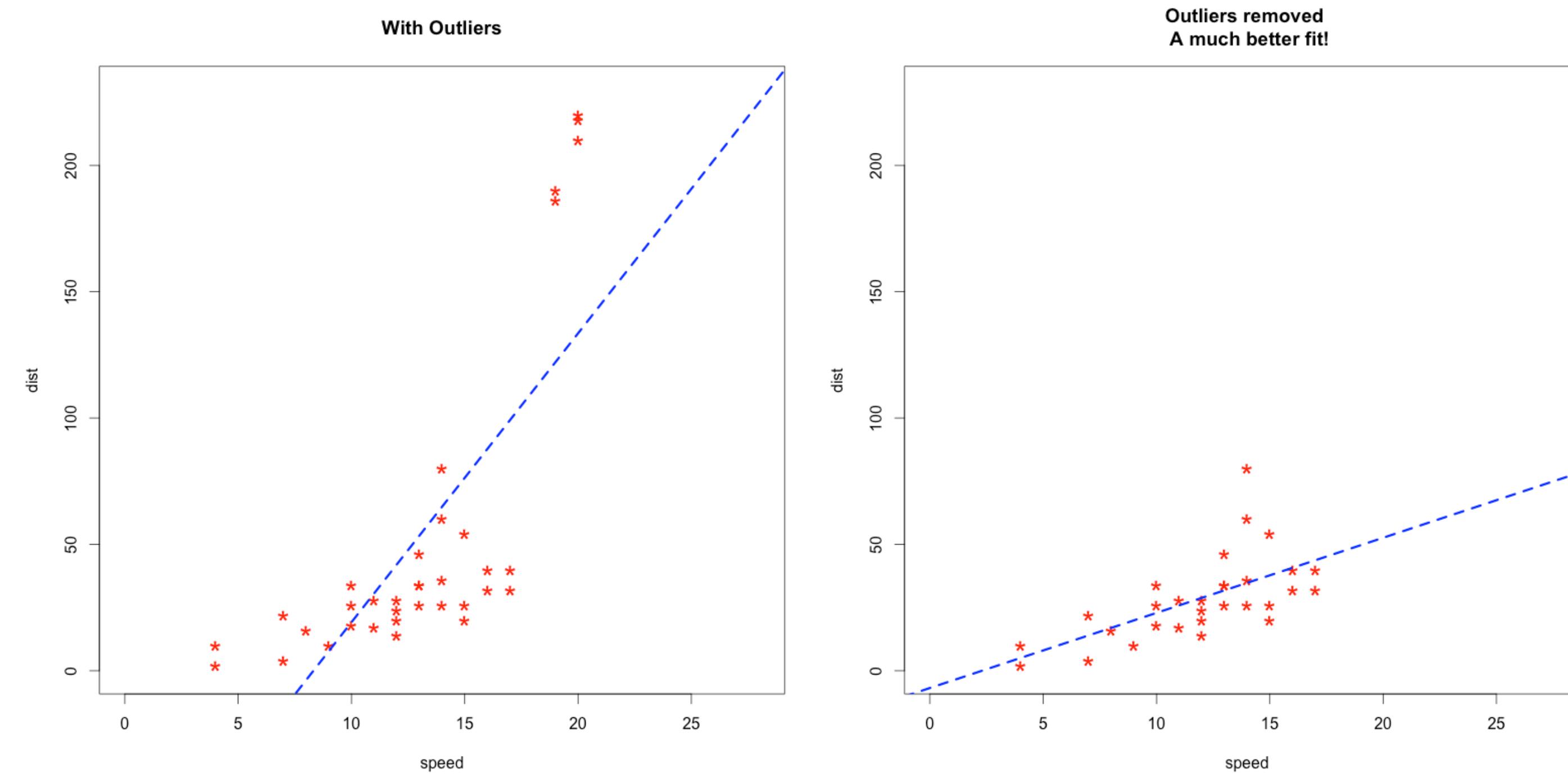


La presencia o ausencia
del outlier no cambia la
regresión.

Outliers

Cleaning the data

3. Afecta los resultados → Realizar el análisis con y sin outliers, evaluar resultados.
Dejar constancia.



Outliers

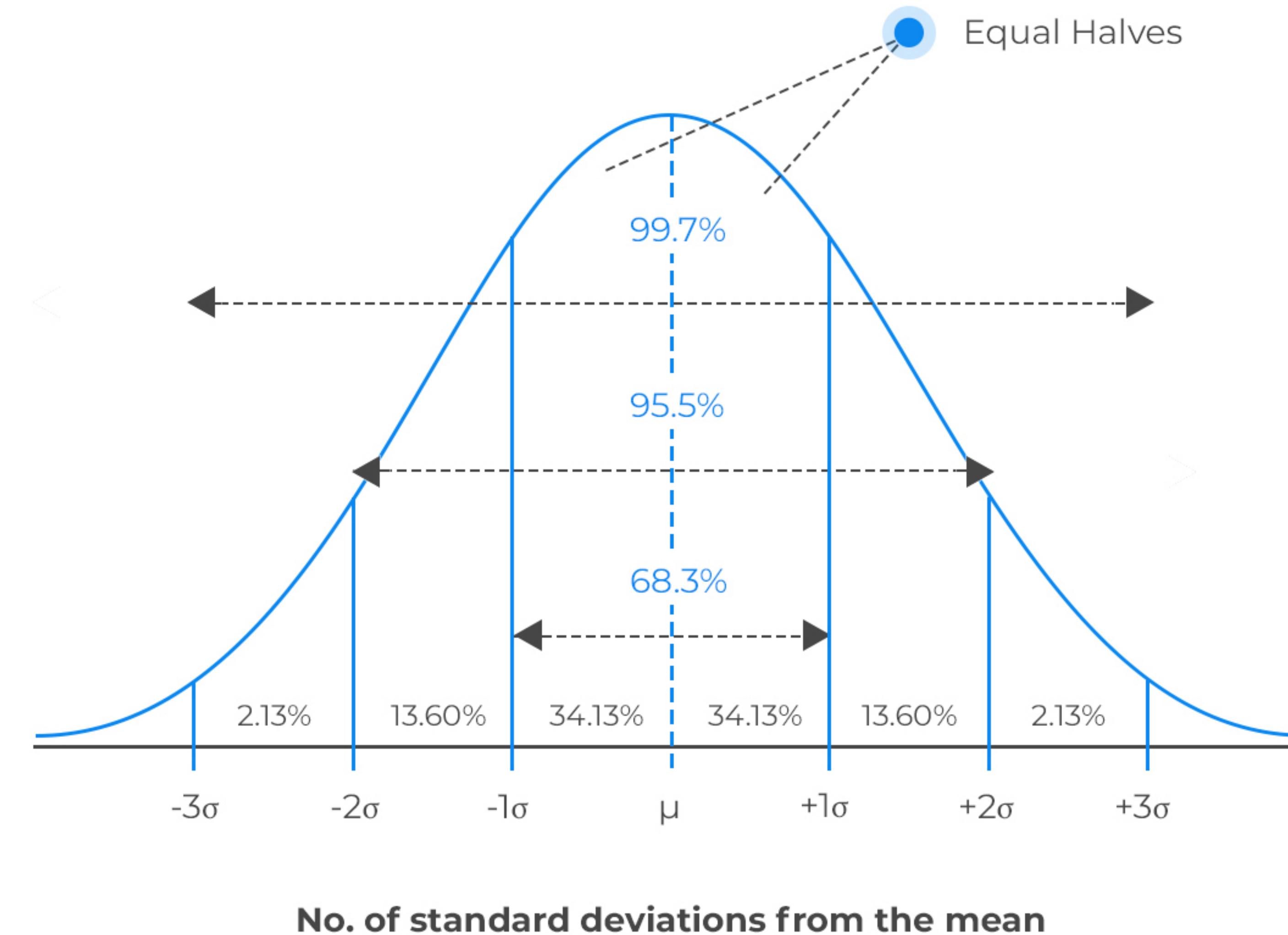
Cleaning the data

Standard Deviation Method

- Distribución normal -> podemos usar la desviación estándar de la muestra como punto de corte para identificar valores atípicos.
 - 1 desviación estándar de la media: 68%
 - 2 desviaciones estándar de la media: 95%
 - 3 desviaciones estándar de la media: 99,7%
- Tres desviaciones estándar de la media es un límite común para identificar valores atípicos en una distribución gaussiana o similar a la gaussiana. Para muestras más pequeñas de datos, quizás se pueda usar un valor de 2 desviaciones estándar (95%), y para muestras más grandes, quizás se pueda usar un valor de 4 desviaciones estándar (99,9%).

Outliers

Cleaning the data



Outliers

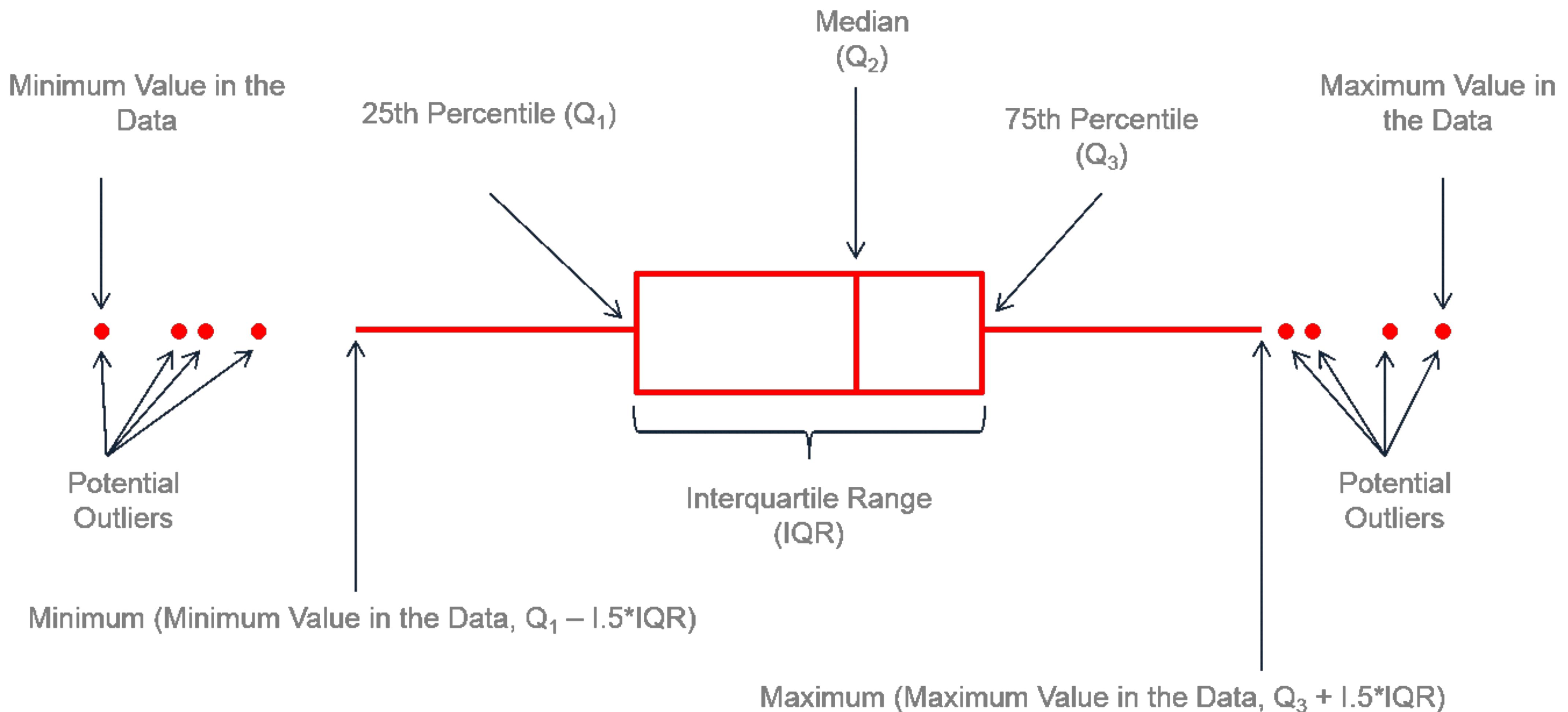
Cleaning the data

Interquartile Range Method

- Una buena estadística para resumir una muestra de datos de distribución no gaussiana es el rango intercuartílico (IQR).
- El IQR se calcula como la diferencia entre los percentiles 75 y 25 de los datos y define la caja en un boxplot.
- El IQR se puede utilizar para identificar valores atípicos definiendo límites en los valores de la muestra que son un factor k del IQR por debajo del percentil 25 o por encima del percentil 75. El valor común del factor k es el valor 1,5. Se puede utilizar un factor k de 3 o más para identificar valores que son valores atípicos extremos o "muy alejados".

Outliers

Cleaning the data



Outliers

Cleaning the data

Interquartile Range Method

- Una buena estadística para resumir una muestra de datos de distribución no gaussiana es el rango intercuartílico (IQR).
- El IQR se calcula como la diferencia entre los percentiles 75 y 25 de los datos y define la caja en un boxplot.
- El IQR se puede utilizar para identificar valores atípicos definiendo límites en los valores de la muestra que son un factor k del IQR por debajo del percentil 25 o por encima del percentil 75. El valor común del factor k es el valor 1,5. Se puede utilizar un factor k de 3 o más para identificar valores que son valores atípicos extremos o "muy alejados".

Missing Data

Cleaning the data

Los datos perdidos se definen como el valor de los datos que no se almacenan para una variable en la observación de interés.

Características

- Los datos faltantes ocurren en casi todas los análisis.
- Los datos faltantes pueden reducir el poder estadístico de un análisis y pueden producir estimaciones sesgadas, lo que lleva a conclusiones no válidas.

Missing Data

Cleaning the data

Tipos de missing values

Missing completely at random (MCAR)

La probabilidad de que falten los datos no está relacionada ni con el valor específico que se supone que se obtendrá ni con el conjunto de respuestas observadas.

- Ejemplo: fallas del equipo.
- El resultado se mantiene unbiased.

Missing at random (MAR)

ocurre cuando la falta no es aleatoria, pero donde la falta se puede explicar completamente por variables donde hay información completa.

- Ejemplo: es menos probable que los hombres completen una encuesta sobre depresión, pero esto no tiene nada que ver con su nivel de depresión, después de tener en cuenta la masculinidad.
- En principio sesga los resultados. Se deben estimar los datos faltantes.

Missing Data

Cleaning the data

Tipos de missing values

Missing not at random (MNAR)

- La única forma de obtener una estimación insesgada de los parámetros en tal caso es modelar los datos faltantes. Luego, el modelo puede incorporarse a uno más complejo para estimar los valores faltantes.

Missing Data

Cleaning the data

Cómo tratar Missing Data?

- Eliminación de observaciones que tienen missing values.
 - Descartar missing values es subóptimo porque cuando se descartan observaciones, se descarta información.
- Imputar los missing values en base a otras observaciones.
 - La imputación de missing values es subóptima porque el valor originalmente continua faltando, lo que siempre conduce a una pérdida de información, sin importar cuán sofisticado sea el método de imputación.