

MANIPULACIÓN Y TRANSFORMACIÓN DE DATOS CON R

MG. ING. LAYLA SCHELI



PRIMER MÓDULO

Conociendo a
tu instructora

Presentación del docente

Ahora, conoceremos algunos aspectos relevantes del experto y del contenido a abordar en este curso.

Los temas que abordaremos son:

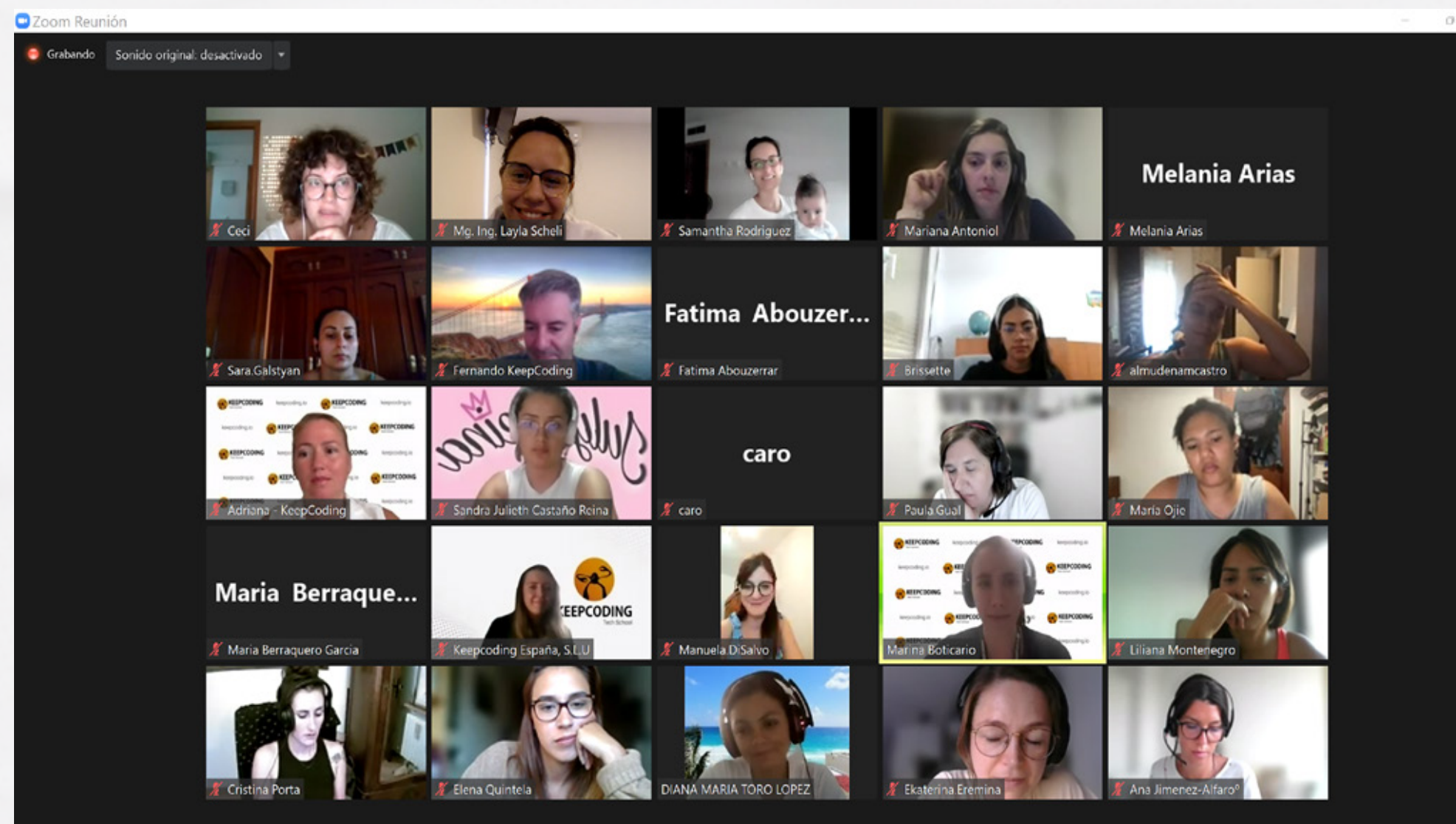
- ¿Quién soy?
- ¿Qué hago? Mi experiencia
- ¿Por qué es importante llevar este curso?
- ¿Qué aprenderán en este curso?



LAYLA SCHELI

¿Quién soy?

Ingeniera de Sistemas de Información, cuento con un Master en Big Data y BI. Poseo más de 7 años de experiencia en el mundo de la analítica de datos.



¿Qué hago?

Formadora, consultora y referente tecnológica en temáticas de Data Science, Data Analytics y Cloud Computing.



¿Por qué es importante llevar este curso?

Como bien sabemos en la actualidad el mundo de la Ciencia de Datos es sumamente solicitado a nivel laboral y conocer cómo realizar procesos de limpieza y de transformación de datos, es un skill requerido para cualquier persona que aspire a iniciar su carrera en Data Science.

El presente curso, nos ofrecerá un enfoque práctico en la aplicación y utilización de R para la manipulación de datos como así también, conocer algunas herramientas complementarias.



¿Qué aprenderán en este curso?

Aprenderás a manipular y transformar datos en R utilizando las librerías más importantes que este poderoso lenguaje de programación nos ofrece.



**Saber utilizar las herramientas
de manipulación de datos, es un
must para un Científico de Datos**





SEGUNDO MÓDULO

Fundamentos de
la Calidad de Datos

Fundamentos de la Calidad de Datos – Parte 1

Ahora, comprenderemos los fundamentos asociados a la calidad de datos para un proceso de Data Wrangling.

Los temas que abordaremos son:

- ¿Qué entendemos por calidad de datos?
- Causas de la mala calidad de datos
- Dimensiones de la calidad de datos

¿Qué entendemos por calidad de datos?

Nos hacemos las siguientes preguntas constantemente:

¿Serán correctos los datos presentados en estos dashboards?

¿Los datos de mi data warehouse, big data o modelo analítico serán consistentes?

¿Puedo utilizar estas fuentes para mis reportes?

¿Cómo reducir las incidencias de calidad de datos de nuestra organización?

¿Cómo generar confianza a través de transparentar la salud de los datos de mi organización?

¿Por qué es relevante la Calidad de Datos?

La calidad de datos confiabiliza el análisis

Si utilizamos datos no confiables podemos llegar a conclusiones erróneas.

Las fuentes son críticas en el proceso de selección de datos.

Existen métodos de control, auditoría, corrección y confiabilización de datos.

Beneficios de la Calidad de Datos

Incrementa la eficiencia y productividad en la toma de decisiones de la organización

Aumenta el nivel de confianza de los clientes y proveedores en cuanto a los datos que brindamos

Permite detectar las incidencias de datos que ocasionan multas/sanciones con los reguladores

Proporciona métodos para agilizar el lanzamiento de nuevos productos y servicios

Disminuye la cantidad de reprocesos y sobre costos en la operación

Reduce los riesgos y costos asociados a la mala calidad de datos

Causas de la mala calidad de datos



- Carga de datos en forma manual o Data Entry.
- Carga de datos externos sin los recaudos correctos para su adecuación.
- Problemas de carga originados en los sistemas transaccionales utilizados como fuente de datos.
- Implementación de nuevas aplicaciones en la organización, implica nuevos orígenes de datos, que necesitan ser congruentes con los datos ya existentes.
- Cambios en las aplicaciones existentes o migraciones de sus bases de datos.

La Calidad de Datos, debe garantizar los siguientes 3 pilares:

1

Exactitud:

- Información libre de errores materiales.
- Información consistente de diferentes períodos.
- Información registrada de forma consistente en el tiempo.

2

Suficiencia:

- Suficiente granularidad de la información.
- Suficiente información histórica para la identificación de tendencias.
- Suficiente información histórica para valorar las características de los riesgos subyacentes.

3

Adecuación:

- Información adecuada al objetivo.
- Información libre en la estimación de errores materiales.
- Información refleja la exposición al riesgo.

La información
de la
organización
debe ser:
exacta,
completa y
relevante.

Dimensiones de la Calidad de Datos

Se deben establecer las dimensiones de análisis de calidad en conjunto con las áreas técnicas y de negocio.

Compleitud	Los valores de los datos deben estar completos (tener un valor) para todos los registros. Ejemplo: Validar completitud campo celular del cliente. % de cumplimiento=93
Disponibilidad	Los datos deben estar siempre actualizados, refrescados y presentes en el repositorio que los almacena. Ejemplo: las tablas agregadas están disponibles a las 9:00 a.m. en la última semana.
Duplicidad	Identificar si los datos están duplicados en nuestro dataset. Ejemplo: el número de documento es único en la tabla de personas.
Consistencia	Los valores de un mismo dato son coherentes en todos los repositorios o sistemas donde se encuentren. Ejemplo: venta válida, cliente válido.
Conformidad	Los valores de los datos cumplen con los formatos, estándar, longitudes, rangos, normas y convenciones establecidos. Ejemplo: el correo cumple con el @ y .com como parte de su contenido.
Precisión/Exactitud	Se utiliza para validar datos sensibles, unidad de medidas, volumetría, tendencias, decimales, entre otros. Ejemplo: la venta diaria no puede ser menor que 10 mil unidades los fines de semana.

Dimensiones de la Calidad de Datos

¡IMPORTANTE!

El entendimiento de estas seis dimensiones es el primer paso para la mejora de la calidad de datos.



Ser capaz de identificar y separar los defectos de los datos clasificándolos por estas dimensiones, nos permite aplicar las técnicas adecuadas para mejorar tanto la información como los procesos que crean y manipulan la información.

¿Qué se persigue con una buena calidad de datos?

- 01** Establecer metas medibles y realistas.
- 02** Alinear las expectativas de negocio y TI y en este sentido, confirmar que la alta dirección es el sponsor del proyecto.
- 03** Entender el coste de la mala calidad de los datos
- 04** Usar una metodología de mejora continua.
- 05** Usar un calendario de despliegue por fases.
- 06** Medir el ROI.

Fundamentos de la Calidad de Datos – Parte 2

Ahora, comprenderemos los fundamentos asociados a la calidad de datos para un proceso de Data Wrangling.

Los temas que abordaremos son:

- Data Profiling
- Metodología existente/implementación
- Métricas de Calidad de Datos

Data Profiling – Perfilado de Datos

Es el análisis de estructuras y contenido utilizando técnicas estadísticas para presentar patrones de los datos.

El perfilamiento de datos debe ser el primer paso en la fase de análisis y diseño (Discovery de inputs) de un proyecto de Data.

Se recomienda automatizar las reglas tanto como sea posible.



Data Profiling – Perfilado de Datos

Profile_Customer_Demo - Column Profiling											Values Patterns Statistics				
Name	Unique V...	% Unique	NULL	% Null	Datatype	% Inferred	Documented Dat...	Minimum...	Maximu...	Last Profile Ru	Value	Frequ...	Percent	Chart	Drill down
Source Name															
CUST_ID	12	100.00	-	-	String(5)	100.00	string(8)	1234	A123	Mar 2, 2014 1:15:	NULL	4	33.33		
COMPANY	5	41.67	-	-	String(10)	100.00	string(15)	ABC INC	STOP 'N ...	Mar 2, 2014 1:15:	Springfield	2	16.67		
CONTACT	11	91.67	-	-	String(15)	100.00	string(23)	BILL WH...	WILLIAM...	Mar 2, 2014 1:15:	Denver	2	16.67		
TITLE	6	50.00	-	-	String(20)	100.00	string(30)	ANALYST SR.	DIR...	Mar 2, 2014 1:15:	San Diego	1	8.33		
ADDR1	12	100.00	-	-	String(17)	100.00	string(26)	123 MAI...	834 2nd ...	Mar 2, 2014 1:15:	SD	1	8.33		
ADDR2	7	58.33	4	33.33	String(11)	100.00	string(17)	Denver	Springfield	Mar 2, 2014 1:15:	SAN DIEGO	1	8.33		
ADDR3	6	50.00	4	33.33	Fixed Length Stri...	100.00	string(5)	MO	ca	Mar 2, 2014 1:15:	Los Angeles	1	8.33		
ADDR4	6	50.00	5	41.67	Integer(5)	100.00	number(8)	62223	92121	Mar 2, 2014 1:15:					
COUNTRY	1	8.33	-	-	Fixed Length Stri...	100.00	string(5)	USA	USA	Mar 2, 2014 1:15:					
PHONE	7	58.33	6	50.00	String(13)	100.00	string(20)	(858)55...	8585555...	Mar 2, 2014 1:15:					
EMAIL	8	66.67	3	25.00	String(20)	100.00	string(30)	JSMITH...	ww@abc...	Mar 2, 2014 1:15:					

- ¿Cuántos nulos hay?
- ¿Cuántos duplicados hay?
- ¿Cuántos valores repetidos hay?
- ¿Cuál es el valor máximo y mínimo?
- ¿Se utilizan bien los tipos de datos?
- ¿Cuál es el valor más repetido?

Estos datos sirven como input para decidir si se considera o no la fuente analizada para mi proyecto de Data.

También, nos permite tener una primera aproximación para establecer las reglas de transformación, formato y limpieza

Metodología de Implementación

La metodología la forman un flujo de procesos DQ, que consiste en un ciclo que se retroalimenta con los resultados que vamos obteniendo.

Punto 1: Iniciamos con una captura de metadatos

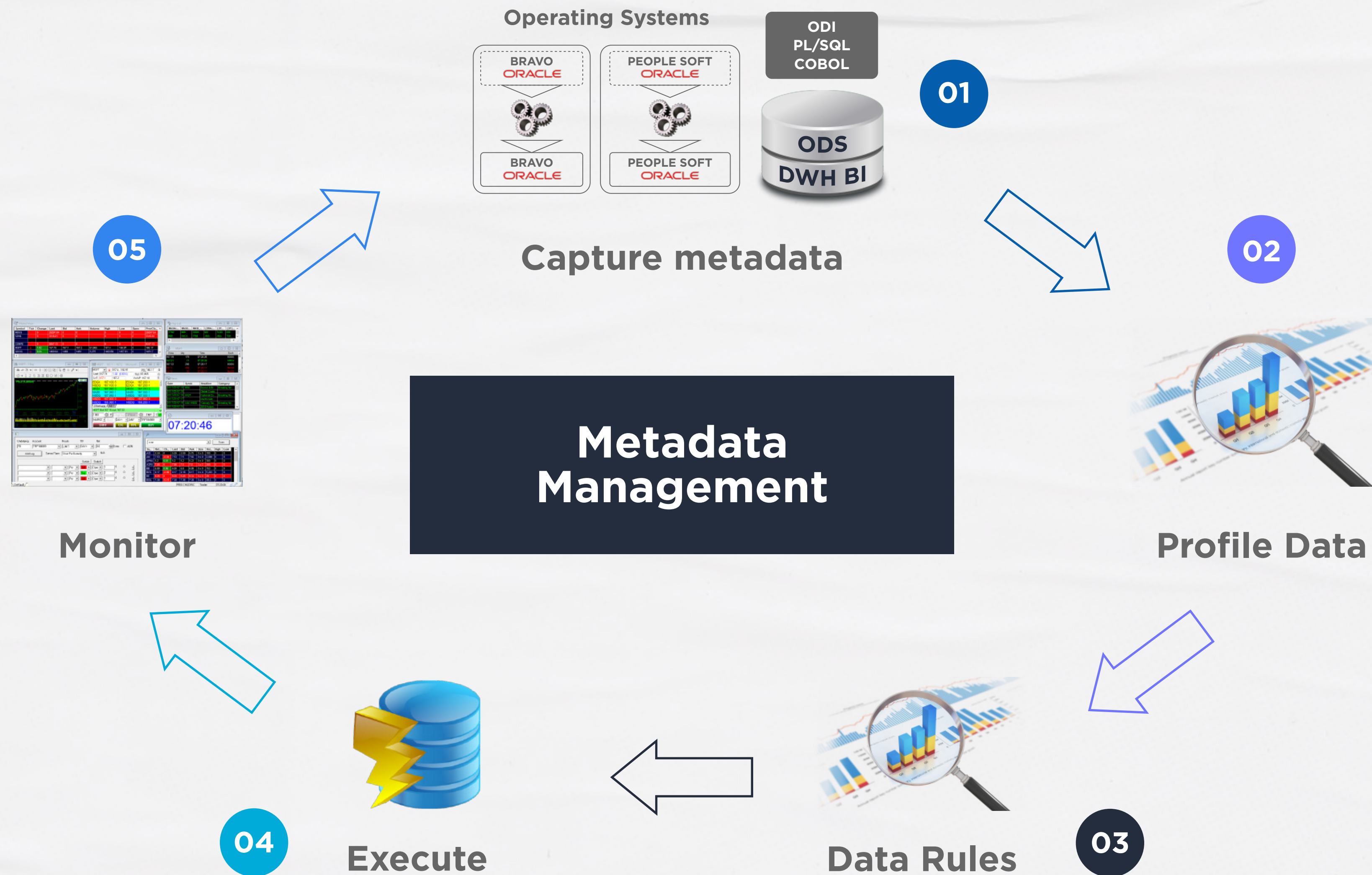
Punto 2: Analizamos la información realizando un perfilado de datos

Punto 3: Creamos reglas tanto técnicas como de negocio

Punto 4: Ejecutamos

Punto 5: Monitorizamos los resultados

De acuerdo a los resultados volvemos a aplicar todo el ciclo hasta obtener la calidad del dato deseada.



Métricas de Calidad de Datos

Las métricas acerca de la calidad de los datos tienen mucho que ver con sus atributos. Un proyecto de calidad de datos debería medir:

- **Precisión:** Exactitud general de los datos en un conjunto. Se determina comparando el conjunto de datos con una fuente de referencia fiable.
- **Compleitud:** Se trata de los datos que faltan, es decir, los campos en el conjunto de datos que se han quedado vacíos o cuyos valores predeterminados se han quedado sin cambios.
- **Conformidad:** Valores de datos de un tipo similar introducidos de una manera confusa o inutilizables, por ejemplo, números de teléfono que incluyen / omiten los códigos de área.

Métricas de Calidad de Datos

- **Consistencia:** Tipos diferentes de registros de datos en un conjunto de datos, como la combinación de la información personal y de negocios.
- **Integridad:** Tiene que ver con el reconocimiento de asociaciones significativas entre los registros de un conjunto de datos.
- **Duplicidad:** Datos que duplican entre sí la información, y que habitualmente comporta el desconocer cuál es el más actualizado.

Métricas de Calidad de Datos



La calidad del dato, siempre empieza en el origen
es decir, en el proceso de Data Acquisition.

Fundamentos de la Calidad de Datos – Parte 3

Ahora, comprenderemos los fundamentos asociados a la calidad de datos para un proceso de Data Wrangling.

Los temas que abordaremos son:

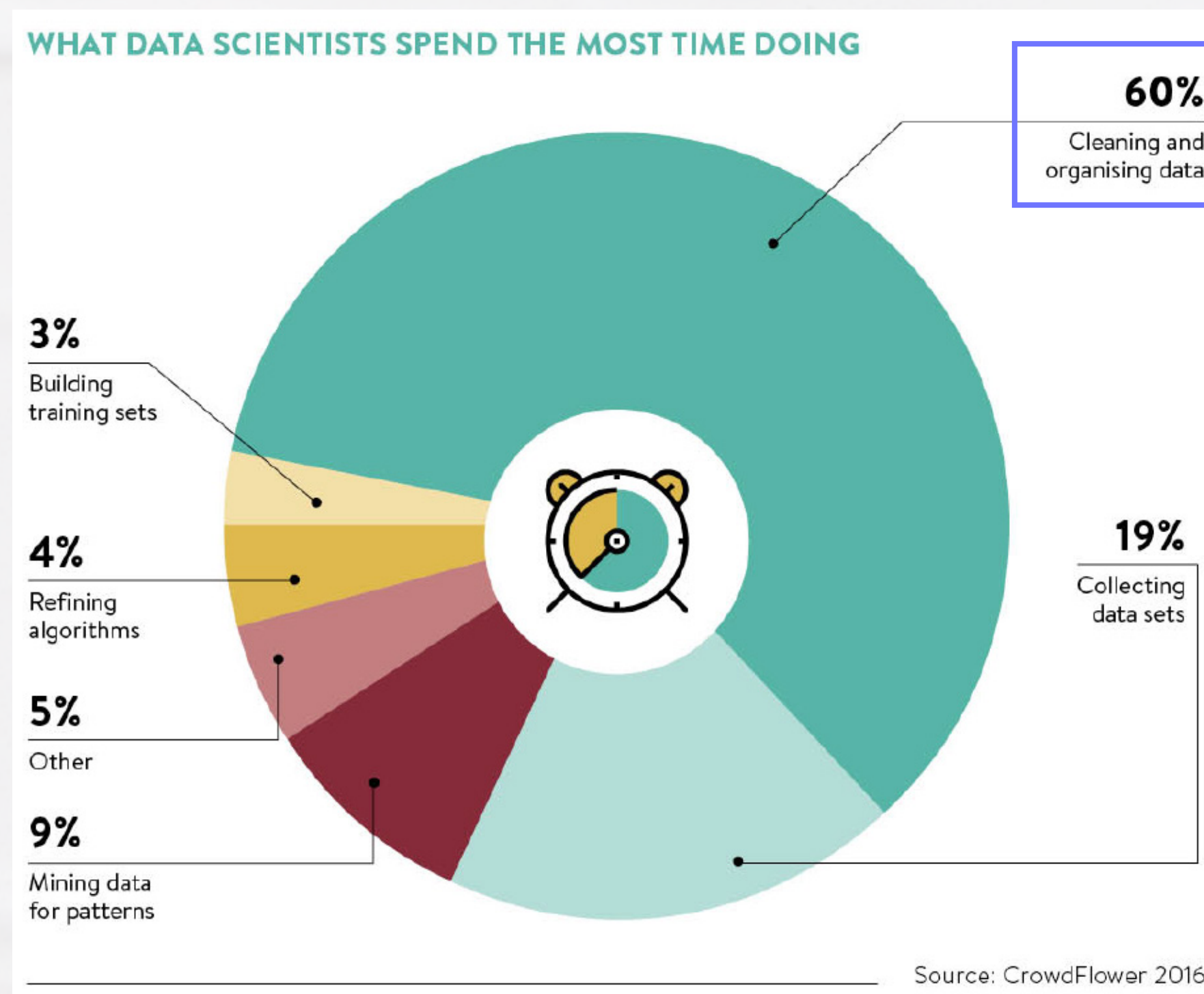
- Introducción a la manipulación de datos – Data Wrangling
- Etapas del Data Wrangling
- Recomendaciones para hacer Data wrangling

Introducción a la manipulación de datos – Data Wrangling

Es el proceso de limpieza y unificación de conjuntos de datos complejos y desordenados para facilitar el acceso, análisis y modelado. Este proceso generalmente incluye convertir y mapear los datos crudos (raw data) y dejarlos en un formato más adecuado para su uso.

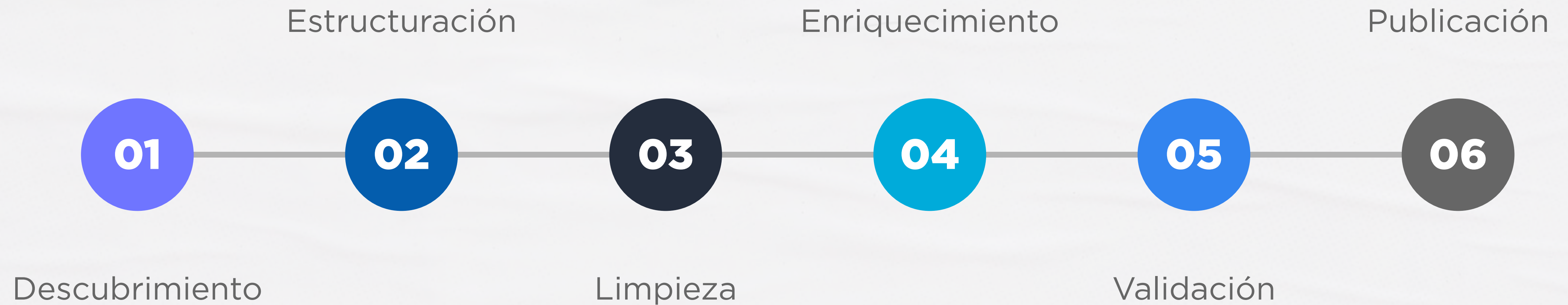
Introducción a la manipulación de datos – Data Wrangling

Recordamos las etapas de un proyecto de Data Science



La mayor cantidad de tiempo, estamos limpiando y manipulando datos para dejarlos en un formato más adecuado para su uso.

Etapas del Data Wrangling



Descubrimiento

Antes de empezar cualquier análisis, es importante comprender los datos, la estructura, tipos y cantidad. También lo es conocer por qué una compañía los utiliza y cómo. Ésto sirve para tomar decisiones posteriores con un rumbo claro.

Estructuración

La idea de esta etapa es estandarizar el formato de los datos. Dependiendo de si hay diversas fuentes u orígenes, los datos estarán en diferentes formatos y estructuras.

Limpieza

Debemos eliminar los datos que no brinden información extra como los duplicados, revisar datos faltantes, etc. Esta propiedad estandariza el formato de las columnas (float, datatimes, etc).

Enriquecimiento

Esta etapa se refiere a agregar datos extra que complementan a los que ya existen para agregar información extra al análisis.

Validación

Es muy importante para los equipos, asegurarse que los datos son precisos y que la información no se alteró durante el proceso. Esto significa asegurar la fiabilidad, credibilidad y calidad de los datos limpios debido a que van a utilizarse para tomar decisiones.

Publicación

Una vez que los datos están validados, se pueden compartir para su uso, realizar análisis exploratorios, entrenar modelos y tomar decisiones.

Recomendaciones para hacer Data wrangling

Filtrar tus datos para aligerar la carga

Considerar el resultado deseado a lo largo de la manipulación del dato

Mantener siempre la capacidad de retroceder a una versión anterior de los datos

Entender dónde y cómo están guardados los datos

Hacer un diccionario de datos

Incluir un experto en la materia siempre que sea posible





**El proceso de Data Wrangling es
el corazón de la Ciencia de Datos**





TERCER MÓDULO

Importación de múltiples
orígenes de datos

El proceso de Data Acquisition

Ahora, conoceremos el proceso de adquisición de datos en R.

Los temas que abordaremos son:

- ¿Qué es la Adquisición de Datos?
- ¿Qué hace un Científico de Datos en términos generales?
- Data Management Maturity Model

¿Qué es la Adquisición de Datos?

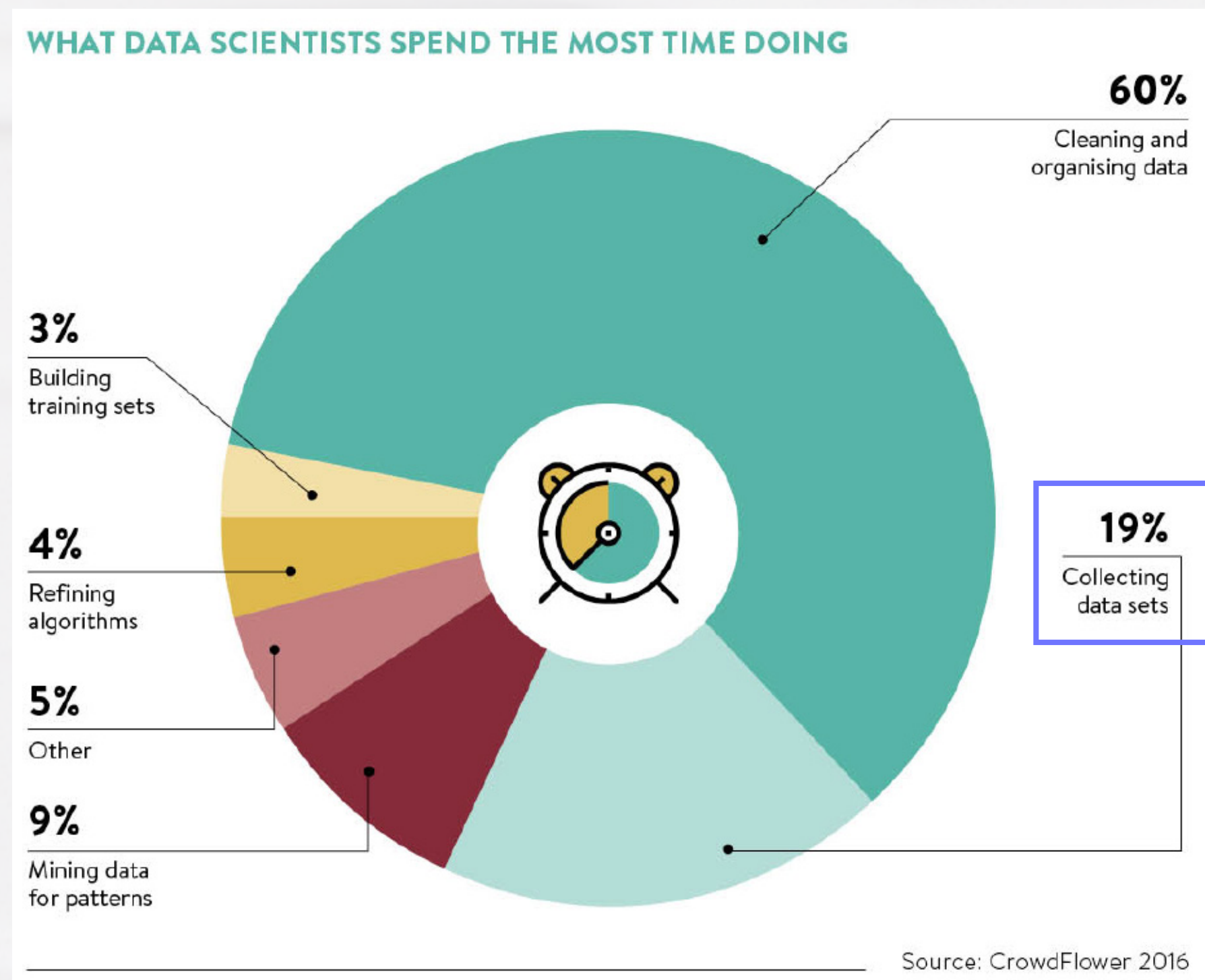
Se trata del proceso de recopilación, que incluye decidir qué datos se quiere adquirir, por qué y cómo, para luego poder explotarlos.

No hay una forma estándar de adquirirlos, debido a que hay muchos tipos de fuentes y proyectos.

Pueden ser adquiridos o heredados de la misma organización, buscados de forma externa o comprados.



¿Qué hace un Científico de Datos en términos generales?



Este número varía mucho de proyecto, pero es un indicador de que se pasa mucho tiempo en esta etapa.

Data Management Maturity Model

En base al nivel de madurez del gobierno de la compañía y su manejo de la información, se gestionará la adquisición de datos.

En la industria, la administración de datos pasa por 5 niveles que definen la madurez de esta.

Nivel 1	Nivel 2	Nivel 3	Nivel 4	Nivel 5
Poca o ninguna gobernanza	Gobierno emergente	Data vista como habilitador organizacional	Gobierno centralizado y planificado	Procesos de Alta Predicción
Roles definidos dentro de los silos	Introducción consistente de herramientas	Procesos y herramientas escalables	Gestión de Riesgos asociado a datos	Riesgo Reducido
Problemas de calidad de datos no abordados	Algunos roles y procesos definidos	Metas establecidas considerando la calidad de los datos	Métricas de Performance de Iniciativas de Datos	Métricas bien establecidas y desplegadas para medir la calidad de los datos
	Creciente conciencia del impacto de los problemas de calidad de datos	Automatización de procesos	Métricas de mejora de Calidad de Datos	

Otros marcos de trabajo

Niveles CM MI



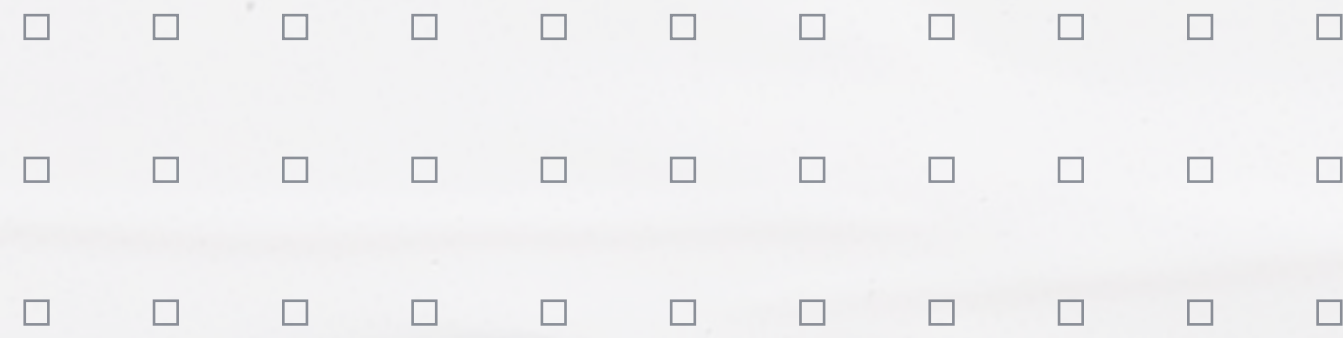
Ejercicios prácticos

Ahora, desarrollaremos ejercicios de importación de datos en R.

Los temas que abordaremos son:

- Ejemplos de aplicación

Ejemplos de aplicación



**El proceso de Data Acquisition
es clave dentro de la mejora
de la calidad de datos**





CUARTO MÓDULO

Operaciones
básicas con Dplyr

Primeros pasos con dplyr

Ahora, conoceremos la librería de Dplyr en R.

Los temas que abordaremos son:

- Instalación de la librería
- Primeros pasos con dplyr
- Interpretando la documentación técnica

Instalación de la librería



El paquete dplyr proporciona una forma bastante ágil de manejar los ficheros de datos de R.

El paquete incluye un conjunto de comandos que coinciden con las acciones más comunes que se realizan sobre un conjunto de datos como por ejemplo para seleccionar una columna podemos utilizar la función: select.

Instalación de la librería

Lo que hace que la sintaxis sea especialmente clara es la correspondencia tan nítida entre el comando y la acción. Para llevar a cabo estas acciones debemos tener en cuenta algunas características comunes:

- ➔ El primer argumento siempre es un `data.frame`
- ➔ El resto de argumentos indican lo que queremos hacer con el `data.frame`.
- ➔ El resultado siempre tiene también la estructura de `data.frame`

Primeros pasos con Dplyr

Interpretando la Documentación Técnica

<https://cran.r-project.org/web/packages/dplyr/vignettes/dplyr.html>

Introduction to dplyr

When working with data you must:

- Figure out what you want to do.
- Describe those tasks in the form of a computer program.
- Execute the program.

The dplyr package makes these steps fast and easy:

- By constraining your options, it helps you think about your data manipulation challenges.
- It provides simple “verbs”, functions that correspond to the most common data manipulation tasks, to help you translate your thoughts into code.
- It uses efficient backends, so you spend less time waiting for the computer.

This document introduces you to dplyr’s basic set of tools, and shows you how to apply them to data frames. dplyr also supports databases via the dbplyr package, once you’ve installed, read `vignette("dbplyr")` to learn more.



**Dplyr es una potente librería
para realizar transformaciones
típicas en R**





QUINTO MÓDULO

Proceso de Data
Wrangling con Tidyverse

Instalación de Tidyverse

Ahora, realizaremos una primera aproximación al proceso de Data Wrangling con Tidyverse.

Los temas que abordaremos son:

- Instalación de la librería
- Primeros pasos con Tidyverse
- Interpretando la documentación técnica

Instalación de la librería

Tidyverse

Es una colección de paquetes disponibles en R y orientados a la manipulación, importación, exploración y visualización de datos y que se utiliza exhaustivamente en ciencia de datos.

Ayuda en todo el proceso de importar transformar visualizar modelar y comunicar toda la información que normalmente utilizamos en procesos de ciencia de datos.



Instalación de la librería

Tidyverse

Permite facilitar el trabajo estadístico y la generación de trabajos reproducibles.

Está compuesto de los siguientes paquetes:

- readr
- dplyr
- ggplot2
- tibble
- tidyr
- purr
- stringr
- forcats

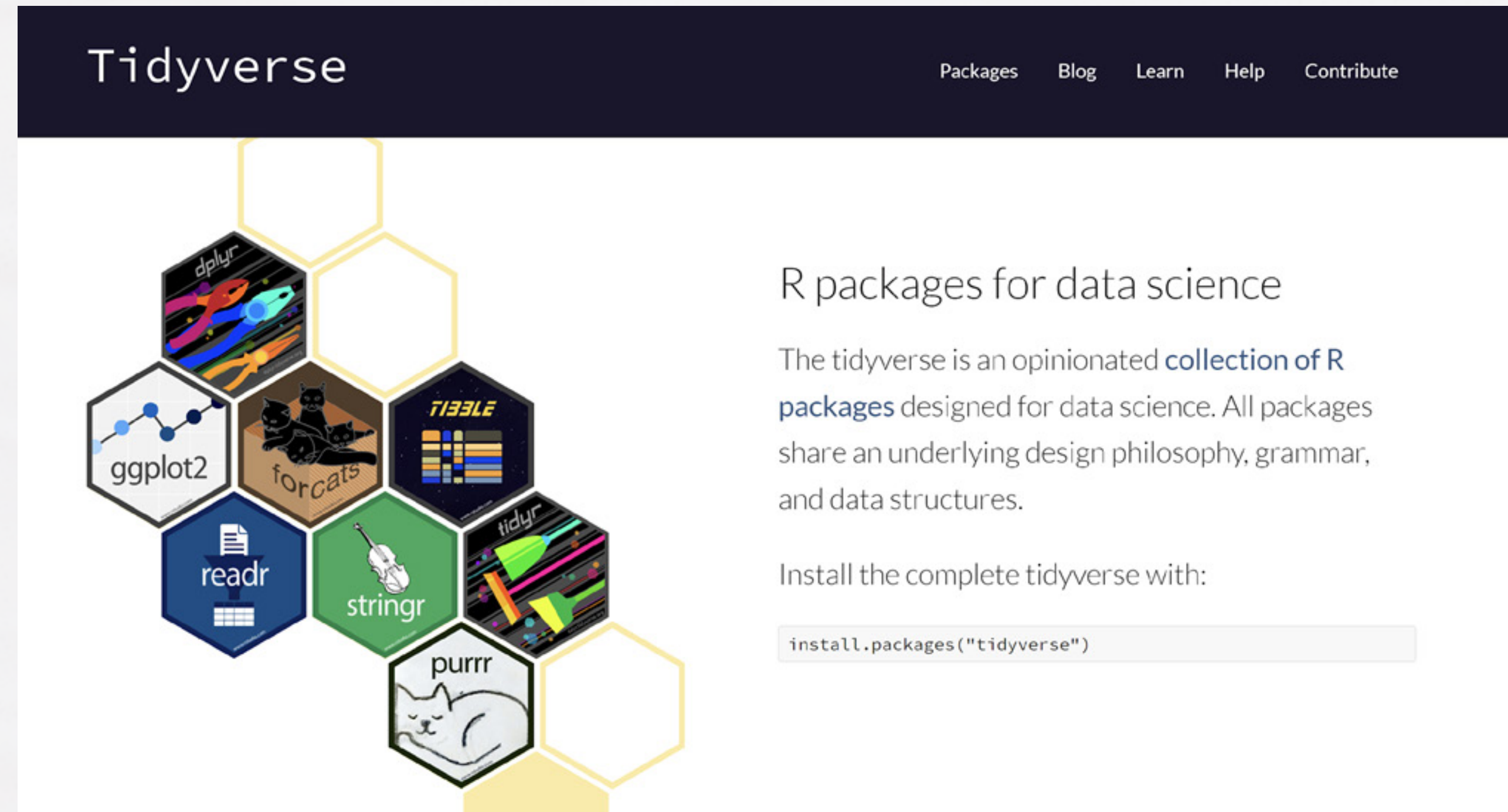
Instalación de la librería

- Básico
 - ggplot2
 - dplyr
 - tidyr
 - readr
- Intermedio
 - purrr
 - tibble
 - stringr
 - forcats

Primeros pasos con Tidyverse

Interpretando la Documentación Técnica

<https://www.tidyverse.org/>



The screenshot shows the Tidyverse website homepage. At the top is a dark blue navigation bar with the 'Tidyverse' logo on the left and links for 'Packages', 'Blog', 'Learn', 'Help', and 'Contribute' on the right. Below the navigation bar is a large graphic on the left consisting of several hexagons, each containing a different Tidyverse package logo: dplyr, ggplot2, readr, forcats, stringr, purrr, tidyr, and tibble. To the right of this graphic, the text reads 'R packages for data science'. Below this, a paragraph states: 'The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures.' Further down, it says 'Install the complete tidyverse with:' followed by a code block containing the command `install.packages("tidyverse")`.

Tidyverse

Packages Blog Learn Help Contribute

R packages for data science

The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

Install the complete tidyverse with:

```
install.packages("tidyverse")
```


Ejercicios prácticos

Ahora, desarrollaremos ejercicios de manipulación de datos en R.

Los temas que abordaremos son:

- Ejemplos de aplicación



**Tidyverse es una
herramienta muy potente en
R, la cual se integra de otros
paquetes complementarios**





SEXTO MÓDULO

Herramientas
complementarias con R

Entornos y Herramientas complementarias

Ahora, realizaremos una primera aproximación a entornos de trabajo complementarios en R.

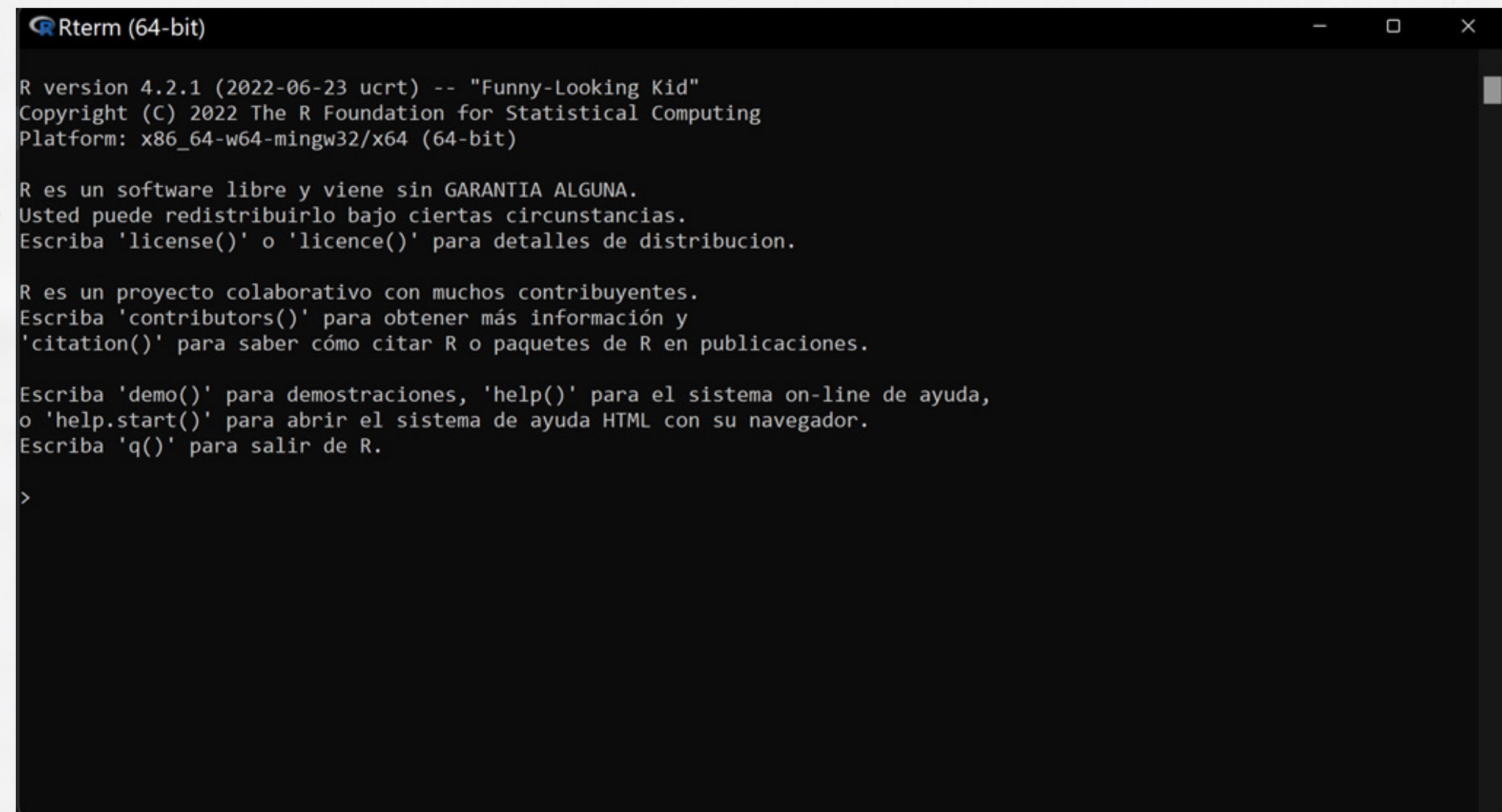
Los temas que abordaremos son:

- La terminal y consola de R
- Integración de Rstudio y Jupyter Notebook
- R en Google Colab

La terminal y consola de R

¿Cómo ingresar?

¿Por qué es relevante?



```
Rterm (64-bit)

R version 4.2.1 (2022-06-23 ucrt) -- "Funny-Looking Kid"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R es un software libre y viene sin GARANTIA ALGUNA.
Usted puede redistribuirlo bajo ciertas circunstancias.
Escriba 'license()' o 'licence()' para detalles de distribucion.

R es un proyecto colaborativo con muchos contribuyentes.
Escriba 'contributors()' para obtener más información y
'citation()' para saber cómo citar R o paquetes de R en publicaciones.

Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.
>
```


Integración de Rstudio y Jupyter Notebook



R en Google Colab





**R se integra a la perfección
con herramientas muy
potentes como Jupyter
Notebook o Google Colab**



Bibliografía

- <https://dplyr.tidyverse.org/>
- <https://gonzalezgouveia.com/que-es-tidyverse-8-paquetes-para-ciencia-de-datos/>
- <https://rsanchezs.gitbooks.io/rprogramming/content/chapter9/dplyr.html>
- <https://www.tidyverse.org/>
- <https://towardsdatascience.com/how-to-use-r-in-google-colab-b6e02d736497>

NETZUN