IMT2200

INTRODUCCIÓN A CIENCIA DE DATOS

2022-2

INTRODUCCIÓN A CIENCIA DE DATOS

CLASE 1

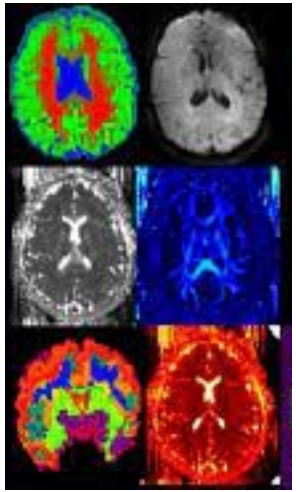
AGENDA DE HOY

- Bienvenida
- ¿De qué se trata el curso?
- Programa y planificación del curso
- Próximos pasos

¿Qué es la Ciencia de Datos?

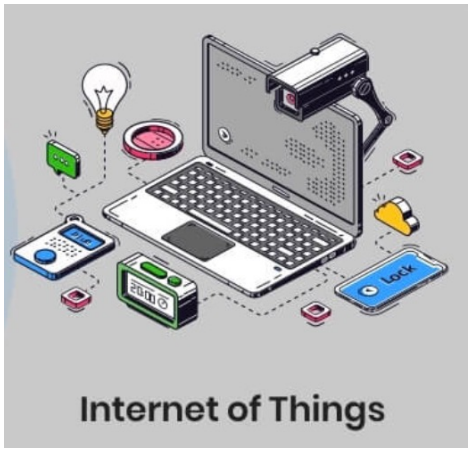
¿POR QUÉ?

DATA AGE: LA DIGITIZACIÓN DEL MUNDO



#US44413318

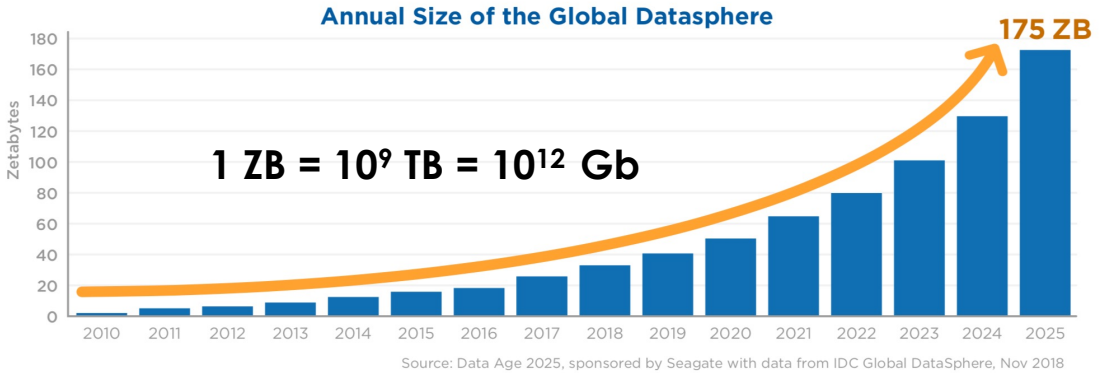
The data-driven world will be **always on**, **always tracking**, **always monitoring**, **always listening** and **always watching** – **because it will be always learning**.



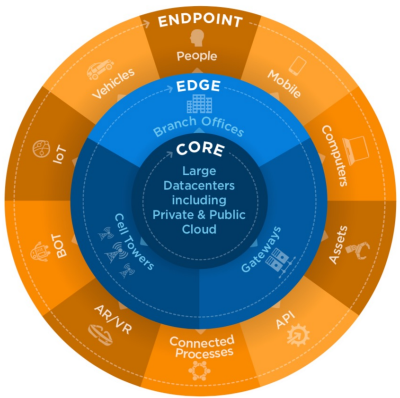
Reinsel, D.; Gantz, J.; Rydning, J.; (2018) "The Digitization of the World: From Edge to Core", IDC White Paper, #US44413318

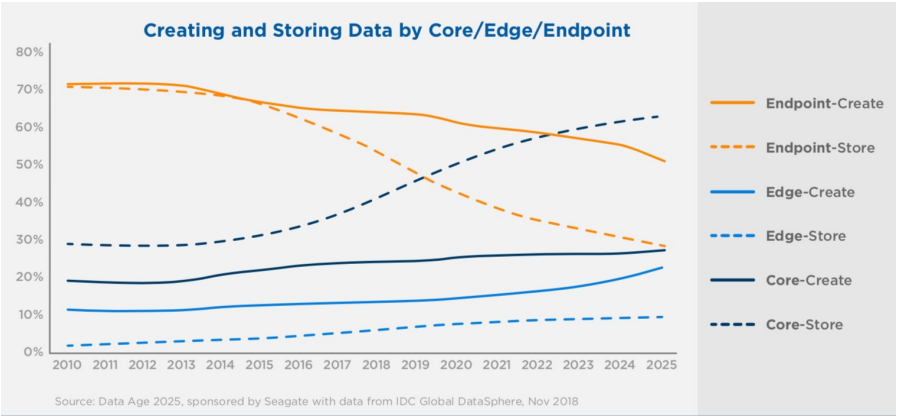
DATA AGE: LA DIGITIZACIÓN DEL MUNDO

"Datósfera Global":
Una medida de toda la nueva data que es capturada, creada y replicada en un cierto año en el planeta.



¿DÓNDE SE GENERAN Y ALMACENAN ESTOS DATOS?





Mayor volumen, variedad, velocidad y acceso a datos.

Reinsel, D.; Gantz, J.; Rydning, J.; (2018) "The Digitization of the World: From Edge to Core", IDC White Paper, #US44413318

¿QUÉ?

CIENCIA DE DATOS: DEFINICIONES SENCILLAS

- "Es la ciencia de extraer información significativa de los datos"

- Procesos y técnicas involucrados en la transformación de estos recursos (datos) en información y conocimiento.

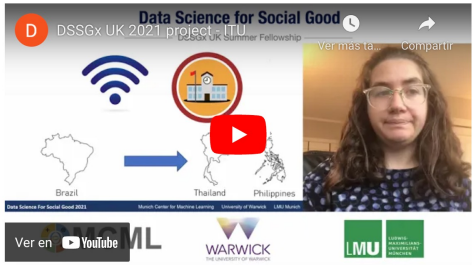
- Es el proceso de formular una pregunta cuantitativa, que puede ser respondida con datos,
  - Recolectando y limpiando los datos
  - Analizando los datos, y
  - Comunicando la respuesta a la pregunta a una audiencia relevante.

- Foco está en correlaciones y patrones, más que en causalidad.

EJEMPLO DEL DÍA: DATA SCIENCE FOR SOCIAL GOOD

https://www.datascienceforsocialgood.org/
https://warwick.ac.uk/research/data-science/warwick-data/dssgx/

Mapping the world's offline population



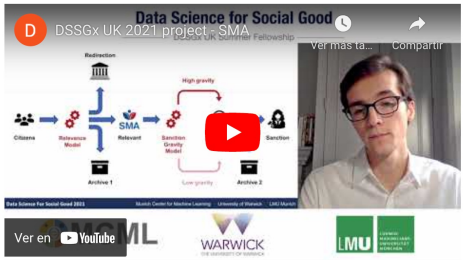
The project is in collaboration with the International Telecommunication Union (ITU), a specialised agency of the United Nations responsible for all matters related to information and communication technology.

ITU's mission is to facilitate connecting all the world's people and to protect and support everyone's right to communicate. ITU and UNICEF have joined forces in the Giga initiative in a bid to connect every school to the Internet by 2030. Aside from directly benefitting children, schools often act as community hubs and so connecting a school often connects the surrounding community too.

This project aims to integrate a number of different data sources to provide real-time estimates of the number of offline people in local areas that would benefit from extending Internet connectivity. The new model will inform the prioritisation of infrastructure projects within Giga as well as policy and decision making at the community level more broadly, and enhance the understanding and use of national and local level internet connectivity data.

Project in partnership with ITU.

Prioritising environmental complaints



This project is a collaboration with the Superintendency of the Environment (SMA) in Chile. SMA is a public service responsible for conducting environmental inspections and ensuring compliance of thousands of facilities to protect the environment and public health.

Due to its limited budget, SMA cannot afford to respond to all citizen demands so it must prioritize where to allocate its efforts. Thousands of citizen complaints are received yearly on a wide range of environmental problems, and this figure has quadruplicated the past year with the launch of online complaints. Each citizen complaint needs to be analysed, and where relevant, lead to inspection and sanction processes.

The project will use a mix of structured (e.g., facility information) and unstructured (e.g., description of environmental problems) data to attempt to prioritise complaints by identifying those that are more likely to lead to grave sanctions as well as those that are outside of the SMA remit. This will help SMA attend to some of the environmental and public health issues that require pressing attention.

Project in partnership with SMA.

SMA
Superintendencia del Medio Ambiente
Gobierno de Chile

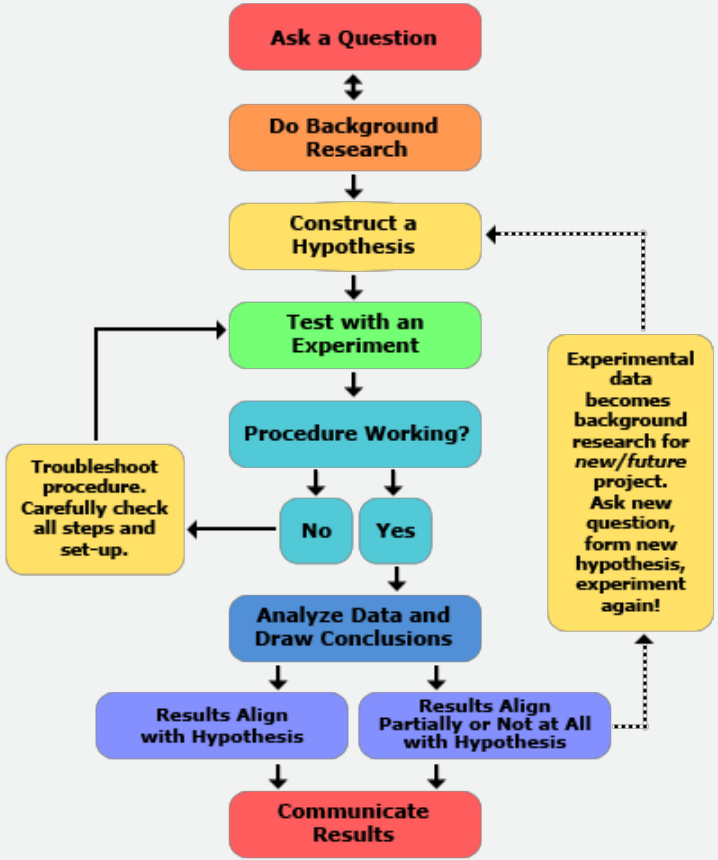
CIENCIA Y CIENCIA DE DATOS

¿Qué entendemos por ciencia?

- Estudio sistemático de la estructura y comportamiento del mundo físico y natural, mediante la observación y experimentación.

- Foco: relaciones de causalidad ➔ ¿Por qué?

- En la medida que abordamos problemas cada vez más complejos, la ciencia requiere un enfoque interdisciplinario para identificar las hipótesis, análisis y relaciones de causalidad correctas.



IA y CIENCIA DE DATOS

¿Es Ciencia de Datos lo mismo que trabajar con "Big Data"?

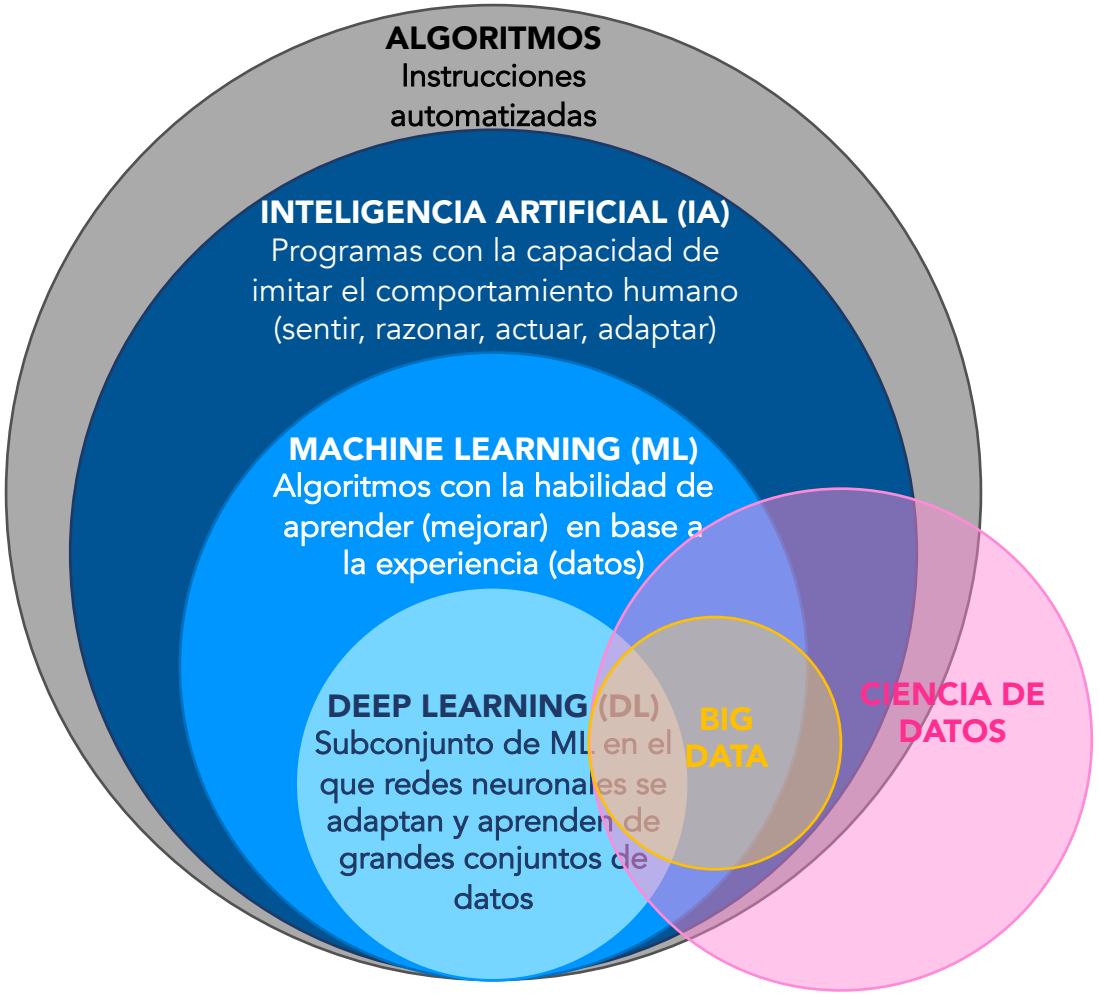
BIG DATA:

Algunas propiedades fundamentales (VVV):

- Gran volumen
- Gran velocidad (ej: tiempo real)
- Gran variedad: estructurada y no estructurada
- Veracidad, valor….

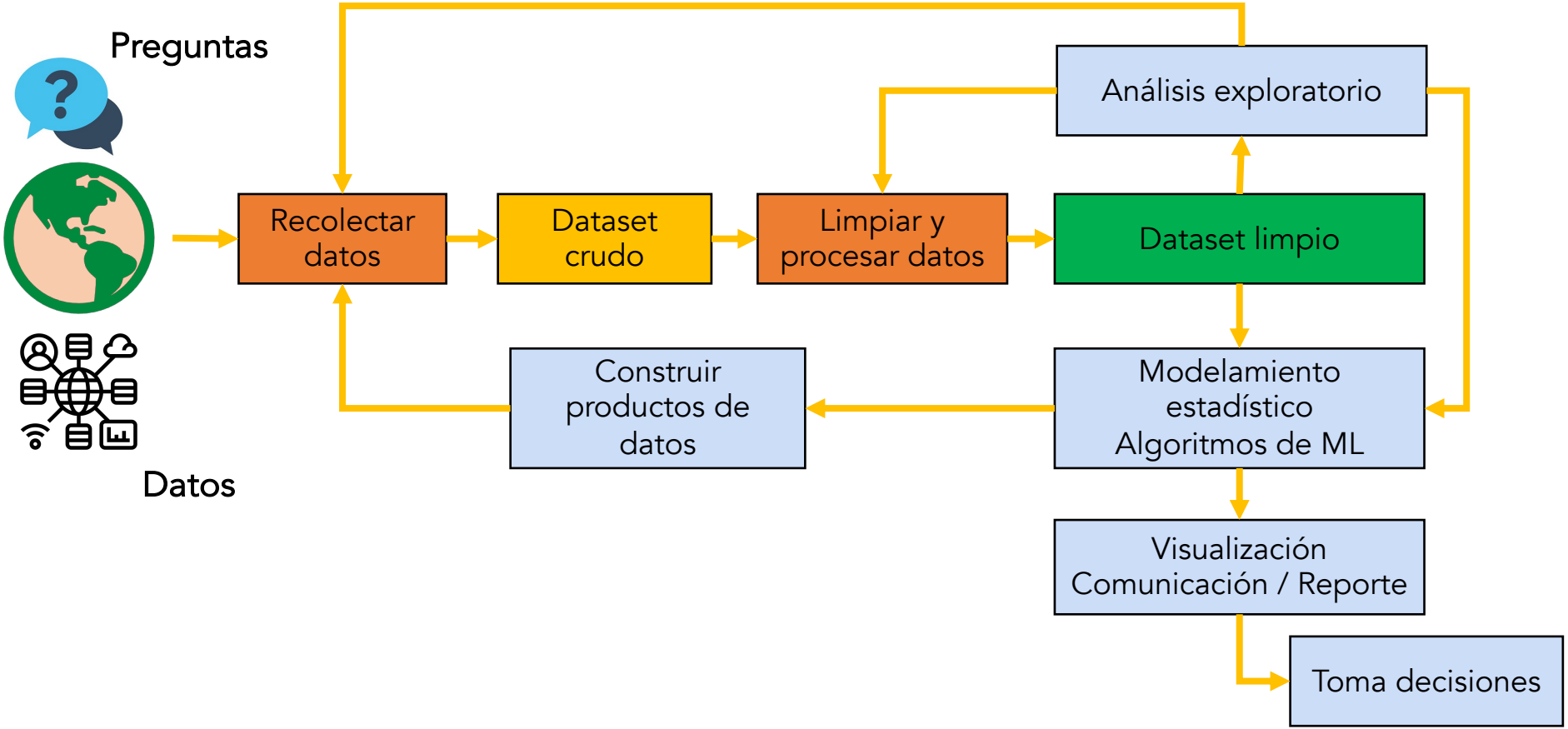
- Es una definición relativa: "es Big Data cuando no puede ser contenida en una sola unidad de cómputo"

  (David Cranshaw, Google)



¿CÓMO?

EL PROCESO DE CIENCIA DE DATOS



¿QUÉ ES UN CIENTISTA DE DATOS?

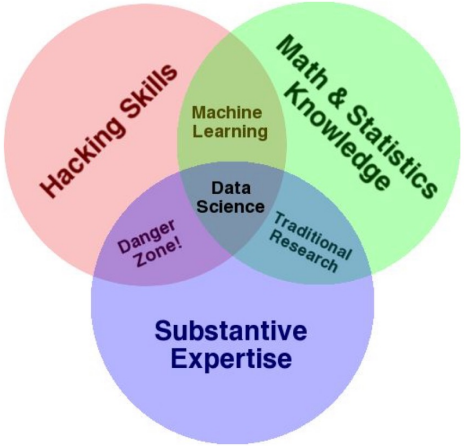
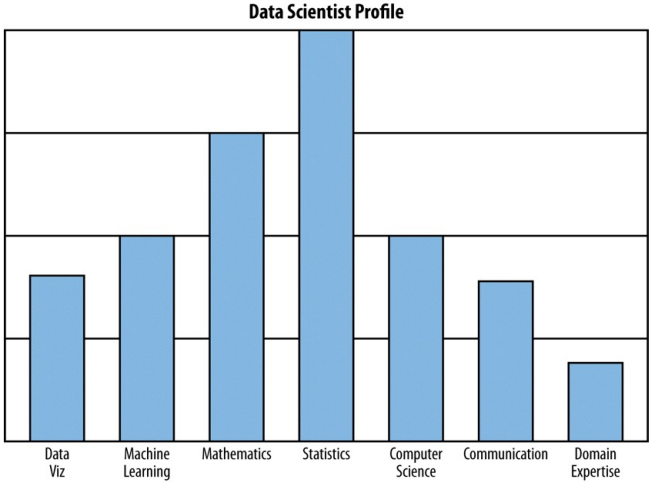


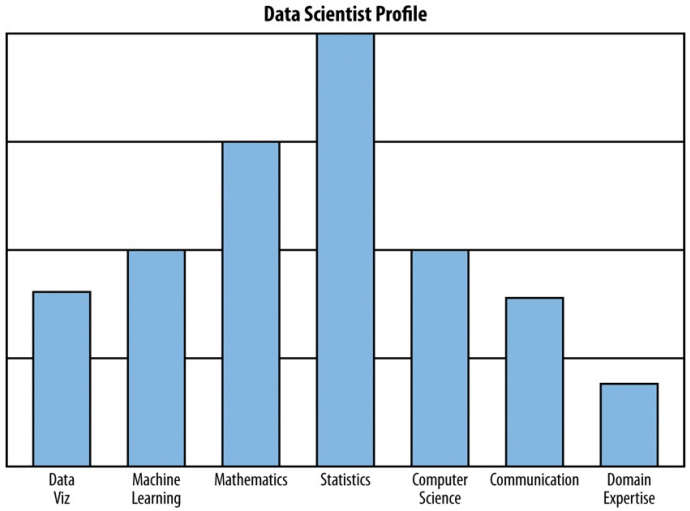
Figure 1-1. Drew Conway's Venn diagram of data science

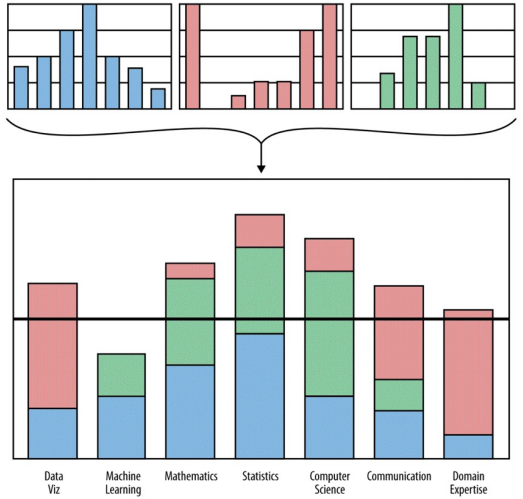


Se requiere una combinación de competencias o conocimientos

- Estadística y modelación matemática
- Computación / programación
- Visualización y comunicación
- Conocimiento específico del campo de estudio
- Trabajo colaborativo

¿QUÉ ES UN CIENTISTA DE DATOS?





Perfiles individuales

Equipo

Se requiere una **combinación de competencias** o conocimientos

- Estadística y modelación matemática
- Computación / programación
- Visualización y comunicación
- Conocimiento específico del campo de estudio
- Trabajo **colaborativo**

¿QUÉ ES UN CIENTISTA DE DATOS?

http://nirvacana.com/thoughts/2013/07/08/becoming-a-data-scientist/

PROGRAMA DEL CURSO

INFORMACIÓN GENERAL

PÁGINA WEB:

https://imt2200-2022.github.io

- Programa del curso
- Calendario
- Material de clases
- Material complementario

➔ REVISAR CON FRECUENCIA

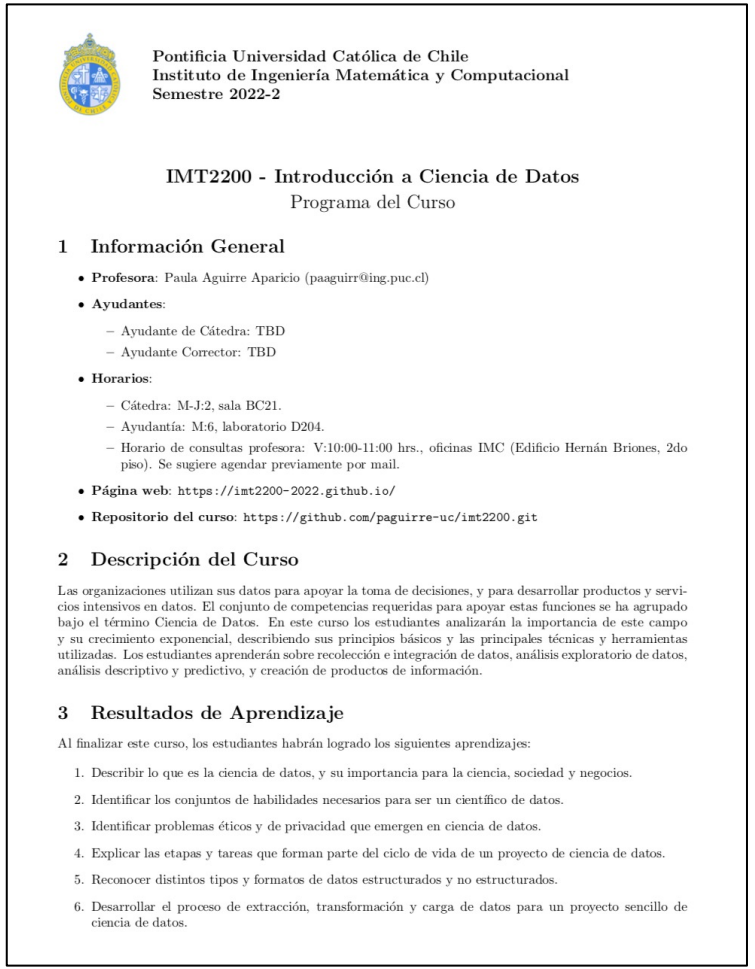
GITHUB:

https://github.com/paguirre-uc/imt2200

- Tareas y proyecto

CANVAS:

- Anuncios
- Calificaciones
- Foros



Pontificia Universidad Católica de Chile
Instituto de Ingeniería Matemática y Computacional
Semestre 2022-2

IMT2200 - Introducción a Ciencia de Datos
Programa del Curso

1 Información General

- **Profesora**: Paula Aguirre Aparicio (paaguirr@ing.puc.cl)
- **Ayudantes**:
  - Ayudante de Cátedra: TBD
  - Ayudante Corrector: TBD
- **Horarios**:
  - Cátedra: M-J:2, sala BC21.
  - Ayudantía: M:6, laboratorio D204.
  - Horario de consultas profesora: V:10:00-11:00 hrs., oficinas IMC (Edificio Hernán Briones, 2do piso). Se sugiere agendar previamente por mail.
- **Página web**: `https://imt2200-2022.github.io/`
- **Repositorio del curso**: `https://github.com/paguirre-uc/imt2200.git`

2 Descripción del Curso

Las organizaciones utilizan sus datos para apoyar la toma de decisiones, y para desarrollar productos y servicios intensivos en datos. El conjunto de competencias requeridas para apoyar estas funciones se ha agrupado bajo el término Ciencia de Datos. En este curso los estudiantes analizarán la importancia de este campo y su crecimiento exponencial, describiendo sus principios básicos y las principales técnicas y herramientas utilizadas. Los estudiantes aprenderán sobre recolección e integración de datos, análisis exploratorio de datos, análisis descriptivo y predictivo, y creación de productos de información.

3 Resultados de Aprendizaje

Al finalizar este curso, los estudiantes habrán logrado los siguientes aprendizajes:

1. Describir lo que es la ciencia de datos, y su importancia para la ciencia, sociedad y negocios.
2. Identificar los conjuntos de habilidades necesarios para ser un científico de datos.
3. Identificar problemas éticos y de privacidad que emergen en ciencia de datos.
4. Explicar las etapas y tareas que forman parte del ciclo de vida de un proyecto de ciencia de datos.
5. Reconocer distintos tipos y formatos de datos estructurados y no estructurados.
6. Desarrollar el proceso de extracción, transformación y carga de datos para un proyecto sencillo de ciencia de datos.

CLASES Y AYUDANTÍAS

CLASES

- Horario: M-J:2 (10:00-11:20)

AYUDANTÍAS

- Ayudante: TBD
- Horario: M:6

- Todas las actividades son presenciales (excepto en situaciones de fuerza mayor).

- Se requiere un computador portátil ➔ disponibles para préstamo en IMC (contactar a Zunilda Alcántara)

EVALUACIONES

TAREAS (NT):

- 4 tareas **individuales** de programación enfocadas en las distintas etapas del proceso de CD.
- Entrega a través de repositorio GitHub personal de cada estudiante.

$$NT = \frac{T_1 + T_2 + T_3 + T_4}{4}$$

ACTIVIDADES EN CLASE (NA):

- ~8 actividades prácticas en clase o ayudantía (ejercicios de programación y/o teóricos).
- Fechas previamente anunciadas.
- Nota = {1, 5, 7}
- Se eliminan 2 notas.
- No se recuperan inasistencias.

$$NA$$

EVALUACIONES

INTERROGACIONES (NI):

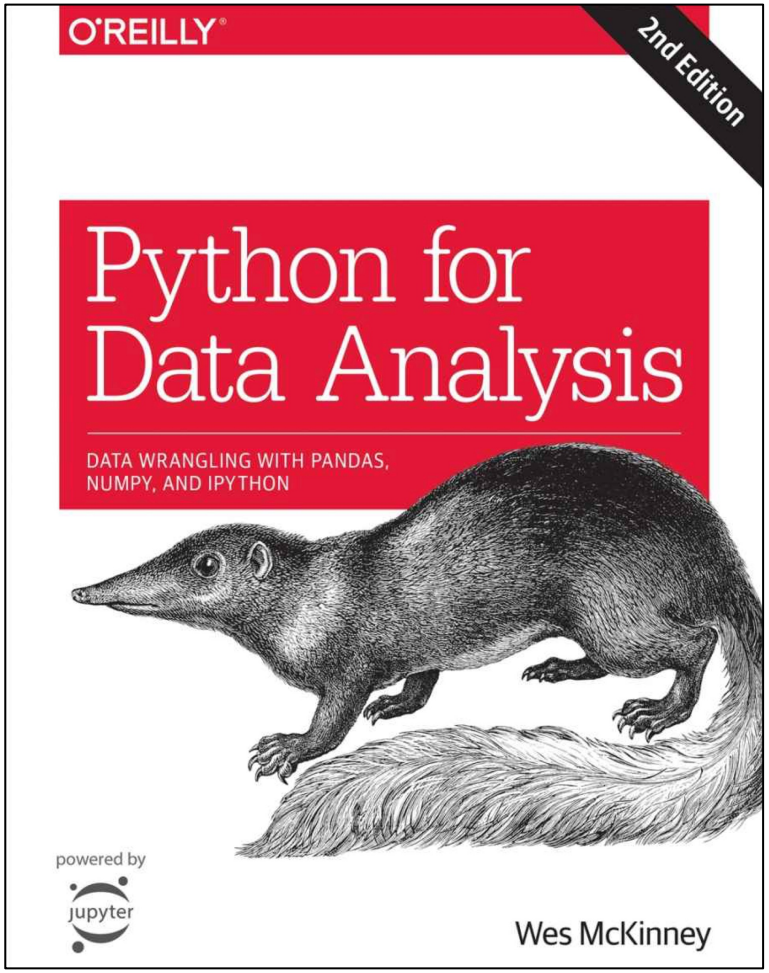
- 2 interrogaciones: 15/09, 15/11, 18:30.

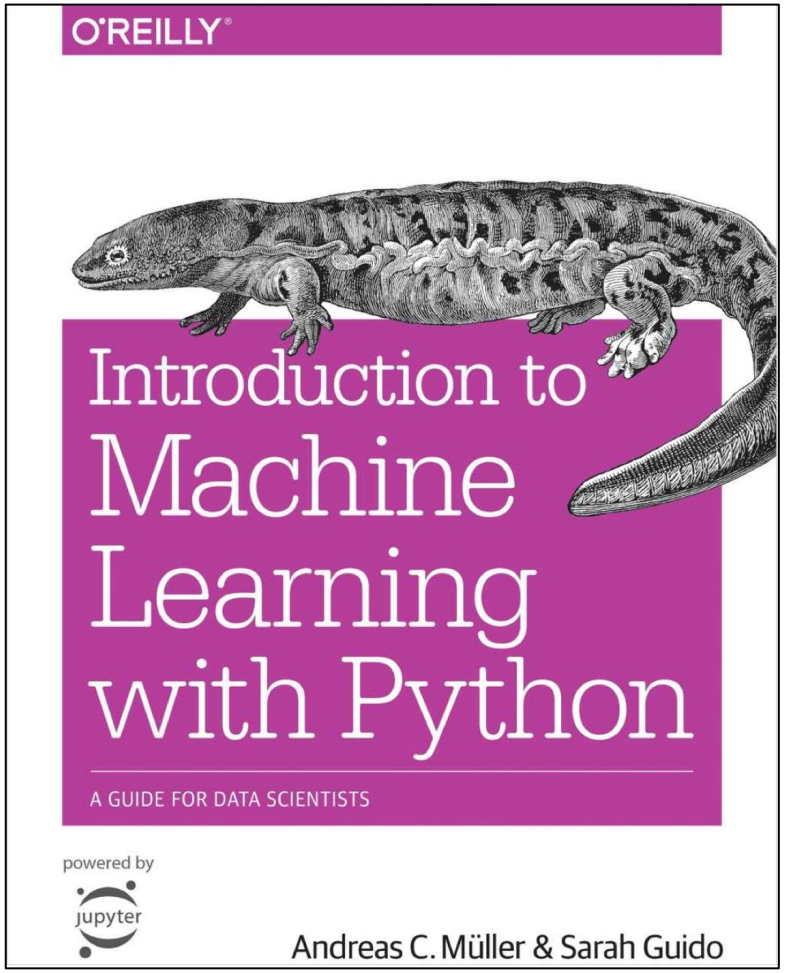
PROYECTO (NP):

- Proyecto grupal de aplicación, en el cual los estudiantes deberán desarrollar el proceso completo de ciencia de datos para resolver una pregunta sobre algún tema a su elección.
- Tres entregables:
  - Propuesta (20%)
  - Repositorio y pagina GitHub (60%)
  - Presentación visual (20%)
- Entrega de enunciado y detalles próxima semana.

NOTA FINAL (NF): $$NF = 0.3 \cdot NP + 0.3 \cdot NT + 0.3 \cdot NI + 0.1 \cdot NA$$

BIBLIOGRAFÍA





PRÓXIMOS PASOS

- Hoy / mañana:
  - Crear cuenta personal en GitHub
  - Enviar información de la cuenta a través de formulario en Canvas (disponible desde las 15:00 hrs.)
  - Instalar herramientas de programación:
    - Python (recomendado: Anaconda)
    - Jupyter Lab
    - Librerías: numpy, matplotlib, pandas, sklearn, geopandas
  - Recibirá una invitación a unirse al repositorio privado del curso.
- Jueves: Clases, Actividad 1.