

DigitalHouse >
Coding School

DATA SCIENCE

MÓDULO 1

Presentación del
programa

1

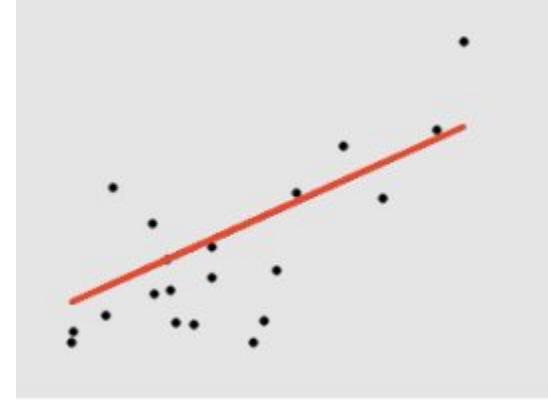
Revisar el concepto de Data Science

2

Desarrollar lineamientos de clase

3

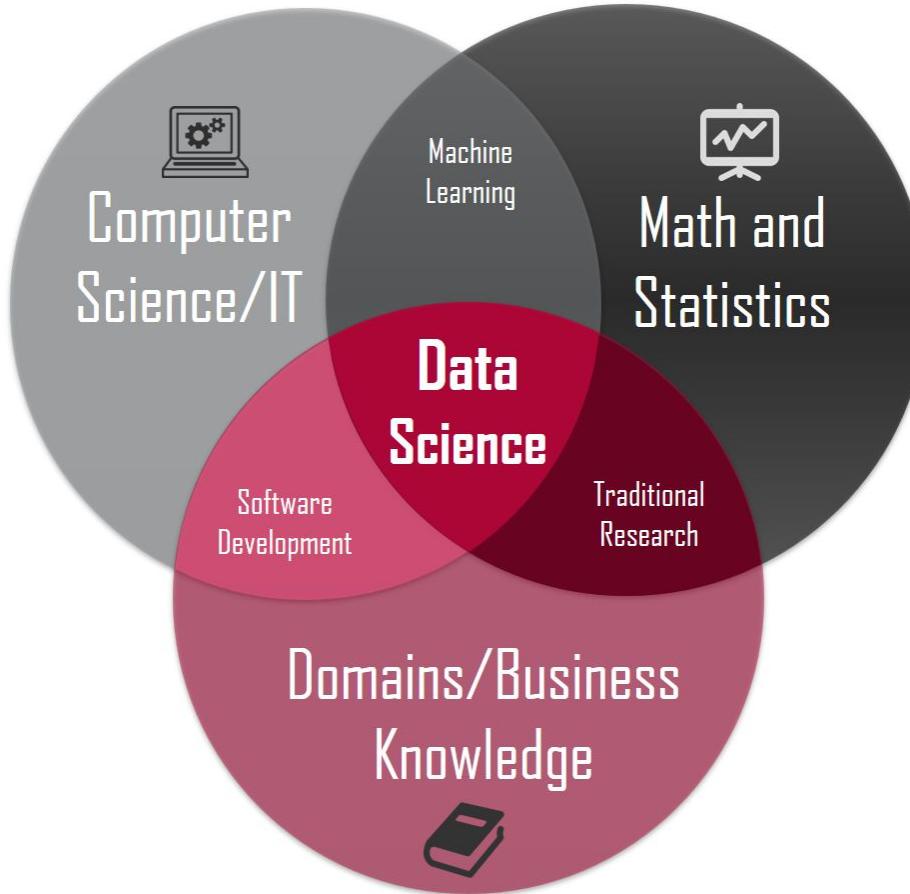
Discutir nuestra propuesta abordar el aprendizaje



¿QUÉ ES DATA SCIENCE?

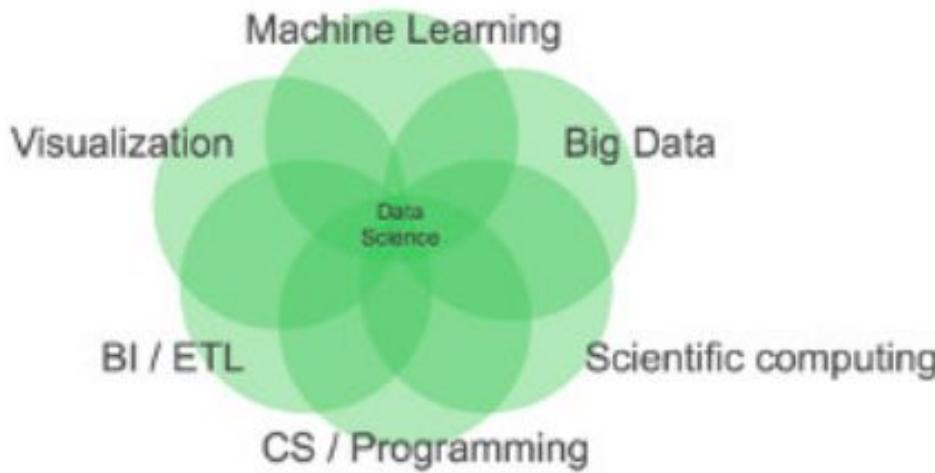


Traditional Data Science Venn Diagram



- Un set de herramientas y técnicas para extraer información útil de los datos
- Una práctica interdisciplinaria orientada a **resolver problemas**
- La aplicación de técnicas científicas a problemas prácticos
- ¿Quién usa Data Science?
 - Recomendaciones de películas Netflix
 - Algoritmo Amazon: “si te gustó X, quizás te guste Y”
 - Five Thirty Eight: cobertura electoral y de deportes

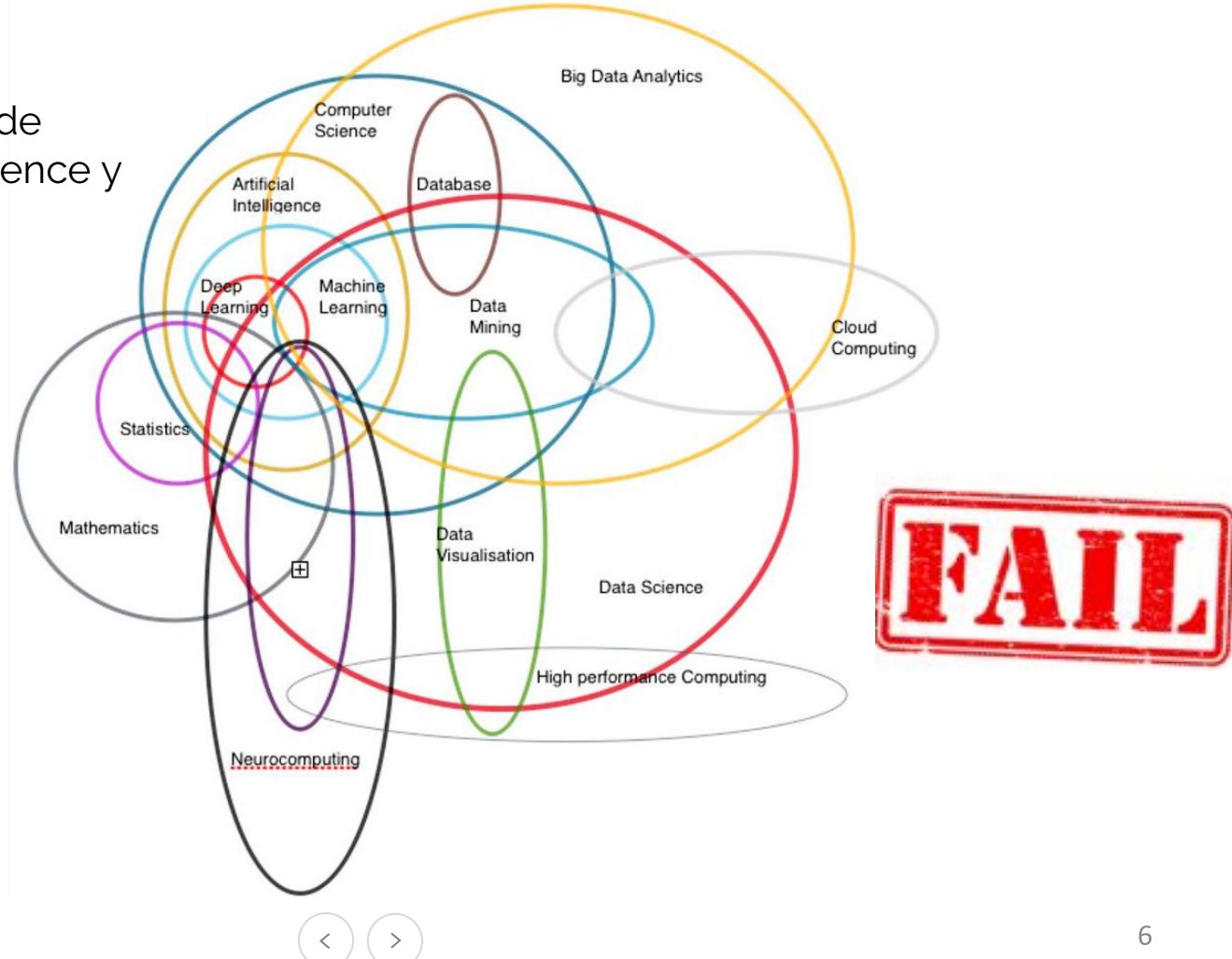
Revised Data Science Venn Diagram

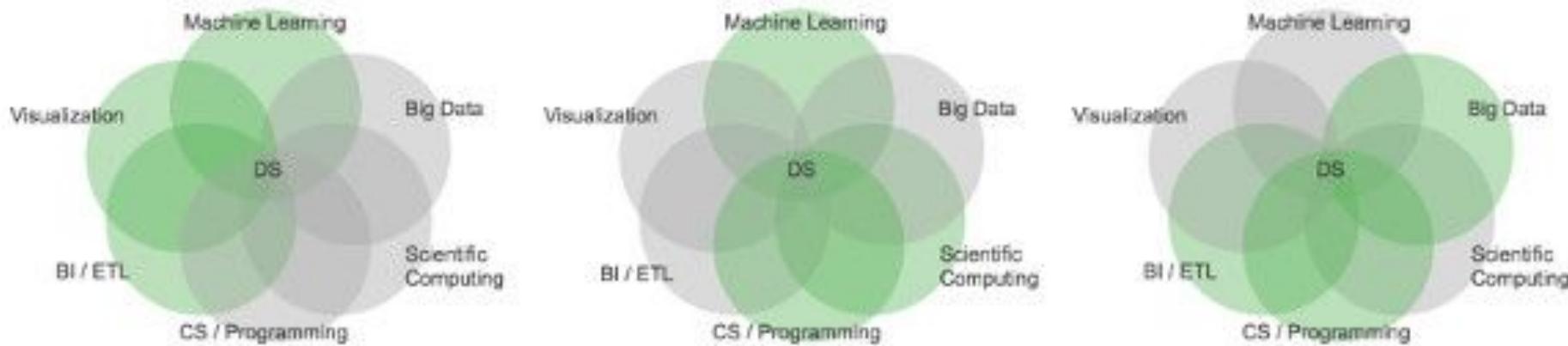


- *Visualización y presentación*
- *Inteligencia de Negocios (BI)*, *Extract-Transform-Load (ETL)* y *bases de datos*
- *Machine Learning, Inteligencia Artificial y Estadística.*
- *CS / Programming (Implementacion)*
- *Computación científica (Bibliotecas) y de alta performance.*
- *Big Data*

Segun Nisarg Dave:

Este complejo diagrama de Venn representa Data Science y el rol de Data Scientist...





**Analista /
Estadística**



**Investigador /
Computación**



**Desarrollador /
Ingeniería**

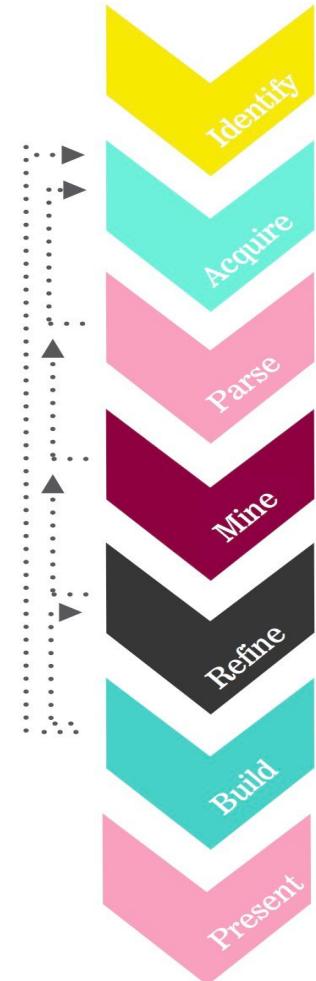
WORKFLOW DE DATA SCIENCE



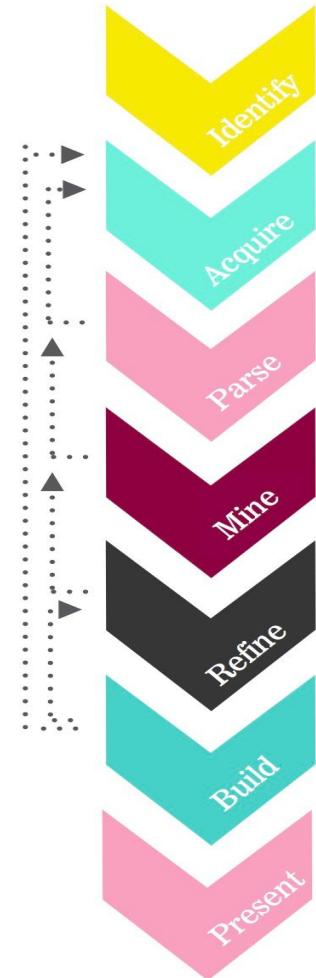
- El “Flujo de trabajo de Data Science” nos sirve para generar resultados confiables y reproducibles.
 - “confiables” = precisos
 - “reproducibles” = otros pueden replicar lo realizado y obtener resultados similares
- **En cualquier punto del proceso, puede ser necesario repetir pasos previos** para iterar a lo largo del flujo.

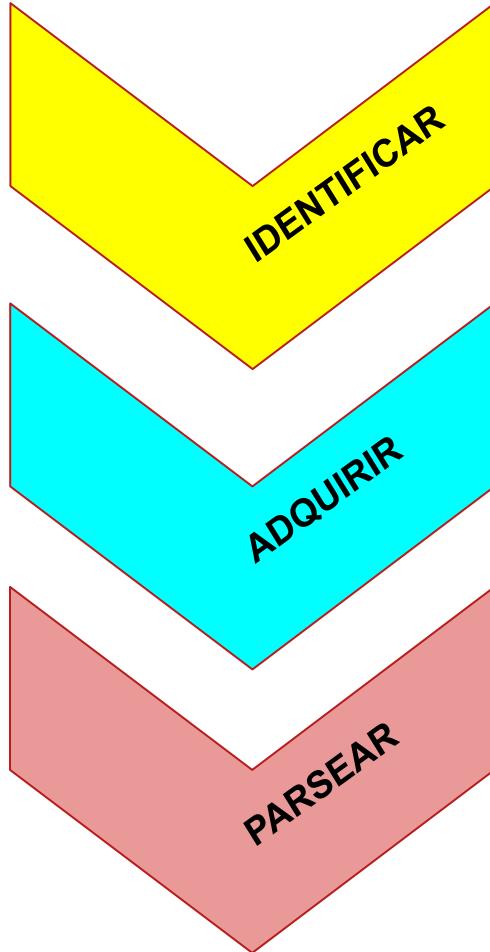
Esto dependerá de:

- la aparición de nuevos datos,
- la necesidad de corregir errores,
- el cambio acerca de las preguntas y objetivos, etc.



- El “Flujo de trabajo de Data Science” constituye, en última instancia, un set de standards sumamente útil y una referencia para tener en cuenta en los **desafíos del curso**.
- Repasemos las diferentes etapas, que están explicadas en detalle en el documento “**Flujo de Trabajo en Data Science.pdf**”





IDENTIFICAR EL PROBLEMA

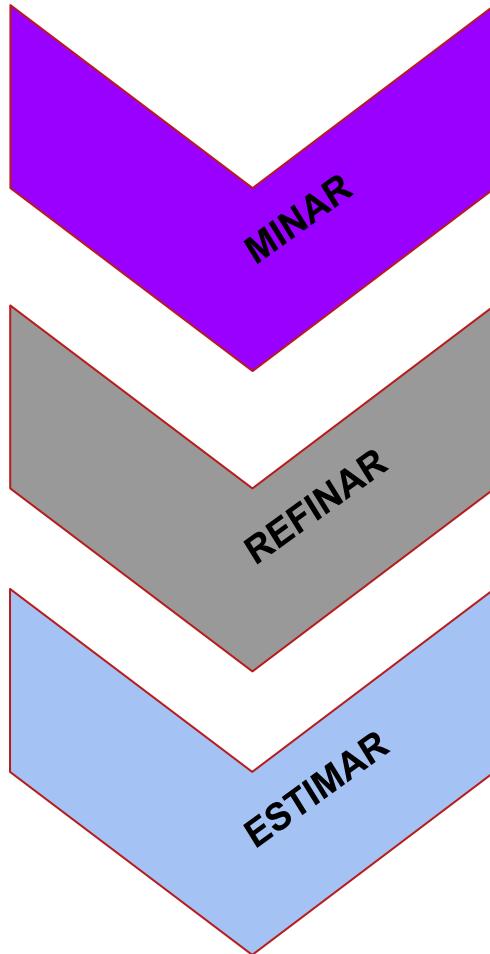
- Identificar los objetivos del producto/negocio/problema.
- Identificar y generar hipótesis sobre metas y criterios para el éxito del análisis.
- Generar un set de preguntas para identificar el dataset "correcto".

ADQUIRIR LOS DATOS

- Identificar el dataset "correcto".
- Importar los datos y generar las estructuras de datos adecuadas.
- Determinar las herramientas más apropiadas para trabajar con los datos.

PARSEAR LOS DATOS

- Explorar toda la documentación relacionada con los datos.
- Realizar Análisis Exploratorio de los Datos (AED).
- Verificar la calidad de los datos.



MINAR LOS DATOS

- Dar formato, limpiar, homogeneizar y filtrar los datos
- Crear nuevas columnas derivadas de los datos originales (recodificaciones, cálculos, etc.)

REFINAR LOS DATOS

- Identificar tendencias y outliers
- Aplicar y calcular estadísticos descriptivos e inferenciales
- Documentar y transformar los datos

ESTIMAR UN MODELO

- Seleccionar un modelo apropiado (forma funcional, estimación, etc.)
- Estimar el modelo
- Evaluar y refinar el modelo



PRESENTAR LOS RESULTADOS

- Resumir los resultados del análisis con alguna narrativa o historia
- Presentar las limitaciones, los supuestos y las fortalezas del/los modelo/s estimados
- Identificar preguntas derivadas y nuevos problemas para seguir profundizando el análisis



FILOSOFÍA DEL PROGRAMA

LINEAMIENTOS DE LA CLASE



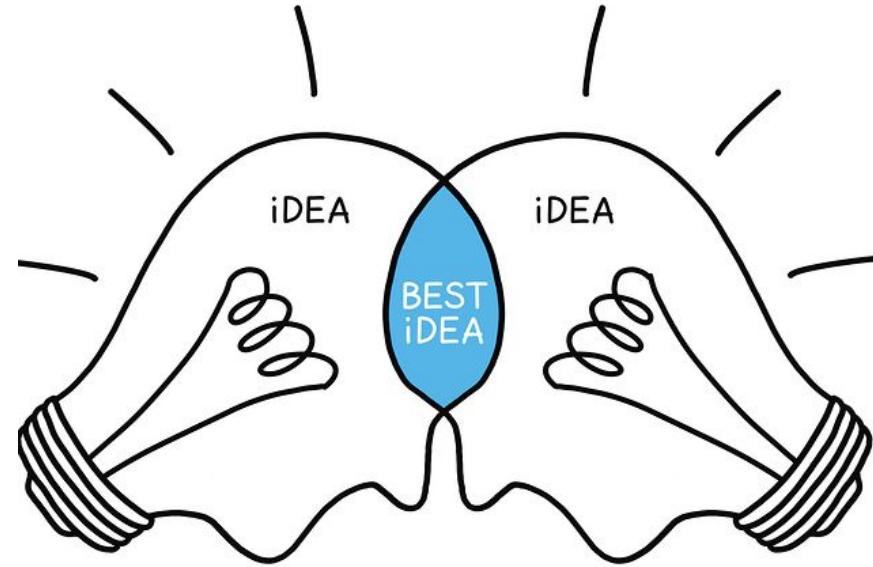
- 1. Aprender las bases**
- 2. Aprender a pensar**
- 3. Aprender haciendo**
- 4. Aprender a aprender**



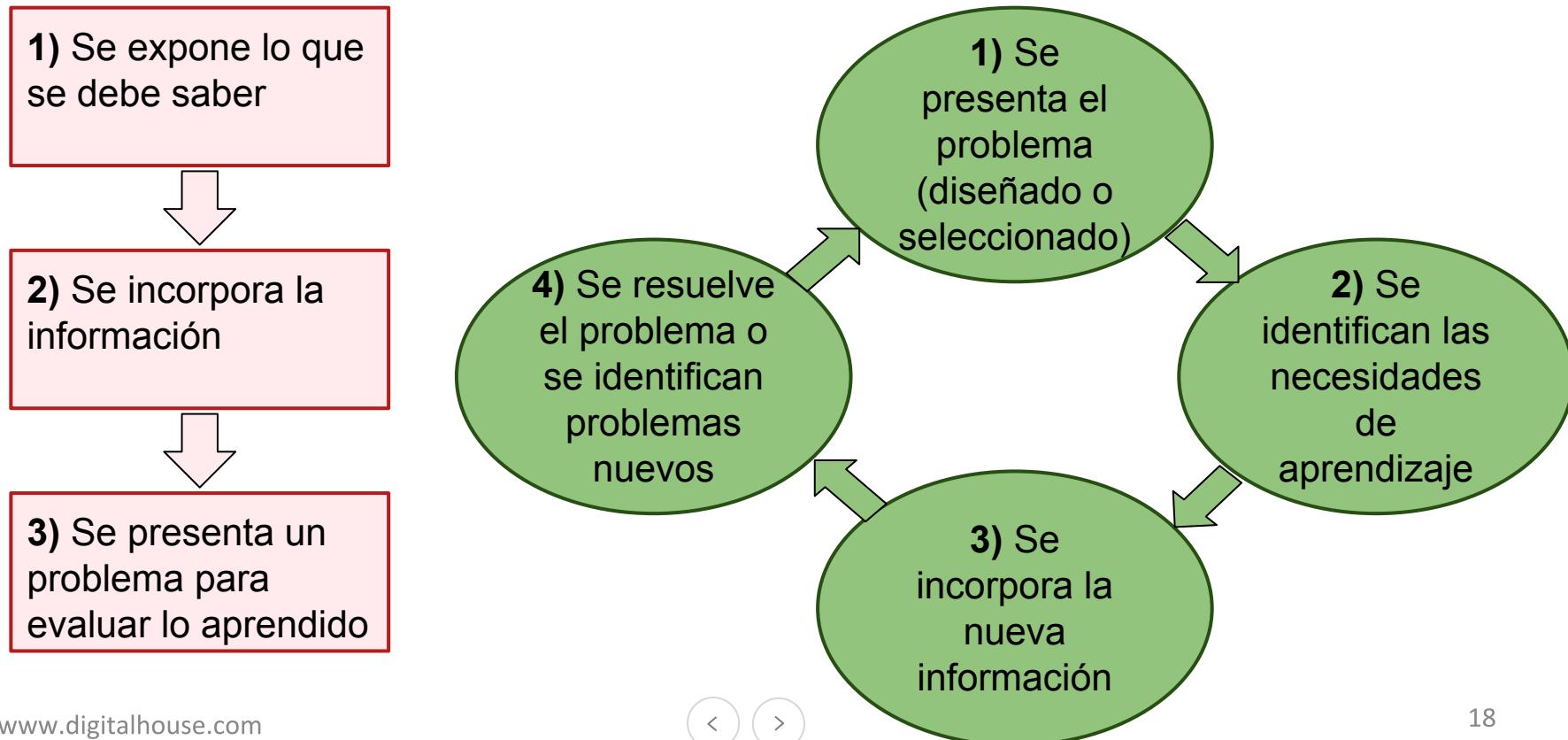
- Fomentar y trabajar en un **entorno diverso**
- Encontrar el **ritmo de aprendizaje óptimo** para cada uno
- **Comunicar** pronto y frecuentemente
- El **éxito** en este curso no se obtiene por comparación. “There is nothing noble in being superior to your fellow man; true nobility is being superior to your former self.” Ernest Hemingway.



- La **dedición**, más importante que el conocimiento previo
- Hacé **preguntas**, todas las veces que consideres necesarias
- **Ayudá** a tus compañeros
- Sé **paciente** con vos mismo



APRENDIZAJE BASADO EN PROBLEMAS



MÓDULOS



- **Los módulos del curso están organizados de manera tal que los asistentes sean capaces de**
 - Extraer, consultar, limpiar y agregar datos para su análisis.
 - Construir, implementar y evaluar problemas de Data Science usando los algoritmos apropiados de machine learning.
 - Usar las herramientas de visualización adecuadas para comunicar sus conclusiones.
 - Investigar, modelar y validar procesos de resolución de problemas aplicados a datasets provenientes de diversas industrias para proveer experiencias en distintos tipos de problemas y soluciones del mundo real.



>

Fundamentos:
POO, Numpy,
Pandas,
estadística

01



>

Clustering y text
mining

05



>

EDA, Limpieza de
datos, Inferencia
Estadística, PCA

02



>

Árboles,
Métodos de
Ensamble y
boosting

06



>

Intro a ML:
Regresión Lineal,
Regularización,
Validación de
Modelos , Pickle y
Flask

03



>

PROYECTO
INTEGRADOR

07



>

Problemas de
Clasificación,
GridSearch

04



>

Fundamentos:
POO, Numpy,
Pandas,
estadística

01

- Introducción al programa y a la disciplina
- Repaso de Python / POO
- Estadística Descriptiva con Numpy
- Pandas



EDA, Limpieza de
datos, Inferencia
Estadística, PCA

02

- Limpieza de datos
- Estadística inferencial
- Visualización
- Variables Dummies
- Datos Faltantes
- Joins con Pandas
- GeoPandas
- PCA y T-SNE

Desafío del Módulo

Usando un dataset crudo de Properati usarán Pandas para limpiar los datos, plantearán formalmente un problema y realizarán análisis exploratorio.



Intro a ML: Regresión
Lineal, Regularización,
Validación de Modelos ,
Web Scraping, Pickle y
Flask

03

- Introducción a Machine Learning
- Intro a Stats Models & Sklearn
- Regresión Lineal
- Separación Entrenamiento/Test
- Regularización & Sobreajuste (Overfitting)
- APIs, Pickle y Flask

Desafío del Módulo

Los participantes construirán un modelo para valuar propiedades en base al dataset de Properati.



>

Problemas de
Clasificación,
GridSearch,
series de tiempo
y text mining

04

- Intro a Clasificación y KNN
- Regresión Logística
- Naive Bayes Classifiers
- Evaluación de modelos
- Feature selection





> Clustering, sistemas
de recomendación,
procesamiento
distribuido, grafos

05

- Clustering
- Text mining

Proyecto Integrador

El Proyecto Integrador (PI) debería representar un aporte original y significativo, aplicando técnicas de data science a un problema interesante.

Charla relámpago:

- Planteo del problema
- Selección de datasets



>

Árboles y
Métodos de
Ensamble

06

- Intro a CARTS
- Árboles de Decisión y Bagging
- Random Forests y Boosting
- XGBoost
- Evaluación de Modelos y Feature Importance

Proyecto Integrador

Informe de avance:

- Análisis Exploratorio
- Primeros intentos con el/los algoritmo(s) seleccionado(s)
- Resultados preliminares



Proyecto Integrador

Entrega Final

- Reporte técnico detallado con todos los análisis desarrollados (en formato notebook)
- Presentación de 10-15 minutos con los insights más relevantes del proyecto
 - Objetivos
 - Datasets
 - Métodos
 - Visualizaciones
 - Storytelling

DESAFÍOS Y PROYECTO INTEGRADOR





— Desafíos y proyectos - objetivos generales:

- Resolver un problema práctico
- Generar un reporte técnico (con código y análisis)
- Generar un reporte para una audiencia no técnica

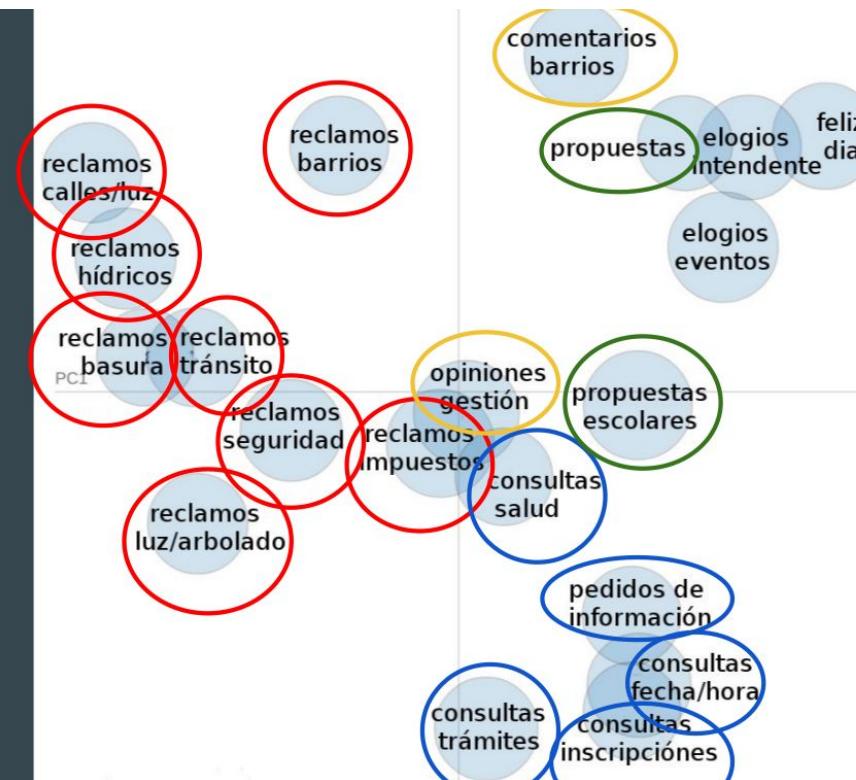
— Desafíos Final: Proyecto Integrador (recorrer todo el Flujo de Trabajo de Data Science)

- Planteo y fundamentación de un problema
- Generación/adquisición de un dataset apropiado para el problema
- Análisis, modelado y visualización de resultados
- Presentación técnica y no técnica de hallazgos y conclusiones

ESCUCHA DE REDES SOCIALES PARA LA GESTIÓN PÚBLICA

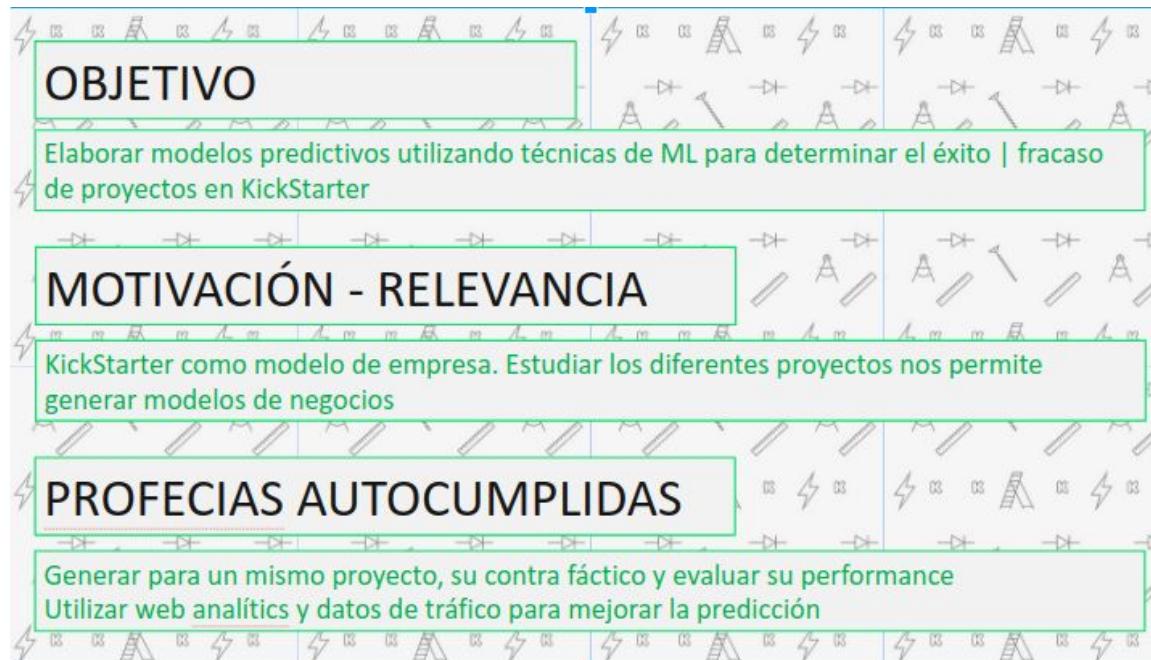
Francisco PENSA

- Reclamos
 - Consultas
 - Propuestas
 - Críticas



KICK-ASS MACHINE LEARNING: ¿QUÉ DETERMINA EL ÉXITO DE PROYECTOS EN LA PLATAFORMA KICKSTARTER?

José SANCHEZ, Jonathan COHEN



Sentiment Analysis y Topic Modeling en Twitter

**Juan ARANGUREN
Jose Luis FINOCCHIARO**



Promesas del Fútbol Mundial

Roberto DI LISIO

Guido BOZZANO

Benjamin BELLOT, Natalia MORAN

| NOMBRE ED DC | ED. | OVR | PO | EQUIPO & CONTRATO 2009 ~ 2021 | BÁSICOS | | VISITAS 2026 COM |
|----------------------------|-----|-----|----|---|----------|--------|---------------------|
| | | | | | VALUE | WAGE | |
| Cristiano Ronaldo ED DC | 32 | 94 | 94 | Real Madrid Club de Fútbol 2009 ~ 2021 | \$107M | \$633K | 2109 |
| L. Messi ED DC SD | 30 | 93 | 93 | Fútbol Club Barcelona 2004 ~ 2021 | \$117.6M | \$633K | 2109 |
| Neymar EI | 25 | 92 | 94 | Paris Saint-Germain 2017 ~ 2022 | \$137.8M | \$314K | 2105 |
| L. Suárez DC | 30 | 92 | 92 | Fútbol Club Barcelona 2014 ~ 2021 | \$108.6M | \$571K | 2291 |
| M. Neuer POR | 31 | 92 | 92 | FC Bayern Munich 2011 ~ 2021 | \$68.3M | \$258K | 1493 |
| De Gea POR | 26 | 91 | 93 | Manchester United 2011 ~ 2019 | \$83.4M | \$330K | 1463 |
| R. Lewandowski DC | 28 | 91 | 91 | FC Bayern Munich 2014 ~ 2021 | \$103M | \$398K | 2150 |
| K. De Bruyne MCO MC | 26 | 90 | 92 | Manchester City 2015 ~ 2021 | \$104.2M | \$319K | 2183 |
| E. Hazard EI MCO | 26 | 90 | 91 | Chelsea 2012 ~ 2020 | \$101.4M | \$330K | 2109 |

Objetivos



Clasificar Jugadores en...



Crack



Promesa



Normal



Predecir Precio de Jugadores

Al final del curso, ustedes serán capaces de:

- Extraer, consultar, limpiar y agregar datos para su análisis.
- Realizar análisis visuales y estadísticos de datos, usando Python y sus bibliotecas asociadas.
- Construir, implementar y evaluar problemas de Data Science usando los algoritmos apropiados de machine learning.
- Usar las herramientas de visualización adecuadas para comunicar sus conclusiones.

Al final del curso, ustedes serán capaces de:

- Crear reportes claros y reproducibles para los stakeholders.
- Investigar, modelar y validar procesos de resolución de problemas aplicados a datasets provenientes de diversas industrias para proveer experiencias en distintos tipos de problemas y soluciones del mundo real.



DigitalHouse >
Coding School

**Conociendo a los
participantes del
programa
usando Data Science
(40 minutos)**

Te proponemos

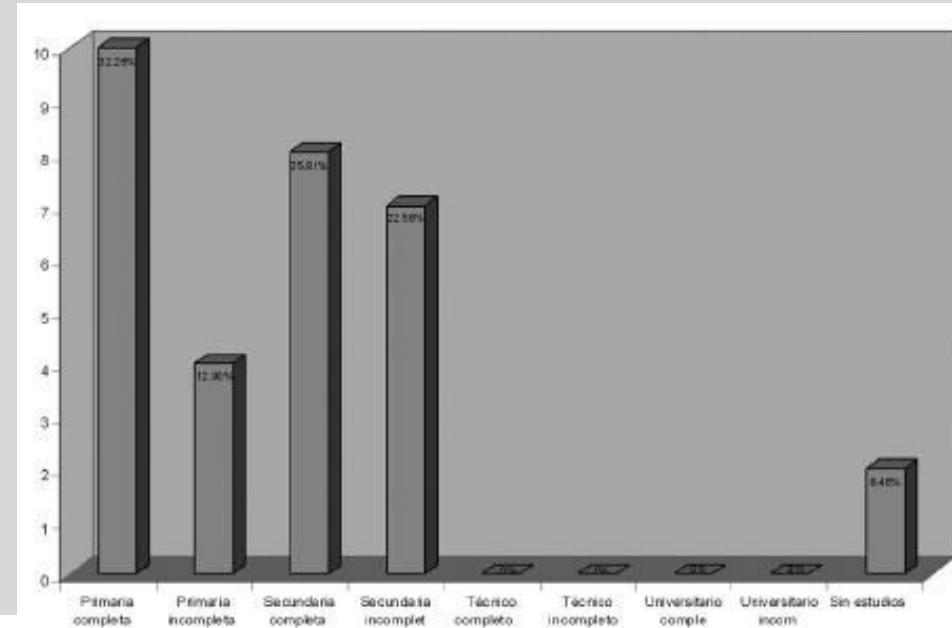
- Que todos los participantes del programa se conozcan mutuamente usando algunos pasos del Flujo de Trabajo de Data Science.
- Que formen grupos de 4 a 6 personas
- Que cada grupo defina **una** pregunta sobre algún aspecto que le interese conocer acerca de los compañeros (motivación, formación, etc.)
- Que a partir de la Encuesta Introductoria al curso puedan abordar las preguntas planteadas.

La idea es que...

- Cada grupo defina los siguientes roles:
 - 1 Project Manager (PM) - Data Business Person: responsable del cumplimiento de los tiempos, de facilitar la comunicación y hacer seguimiento del flujo de trabajo.
 - 1 a 3 Researchers: encargados de adecuar la pregunta a los datos disponibles y de resumir la información para obtener la respuesta. Arman visualizaciones lo más claras y sintéticas posibles de la pregunta en cuestión.
 - 1 a 2 Comunicadores-Creativos: encargados de resumir y presentar los hallazgos y conclusiones a los participantes.

Por Ejemplo

- ¿Cuál es el perfil educativo del curso de Data Science-2017?
 - Primario incompleto
 - Primario completo
 - Secundario incompleto
 - Secundario completo
 - Universitario/Terciario incompleto
 - Universitario/Terciario completo
 - Posgrado o superior
 - Sin Estudios



Cronograma

| Actividad | Tiempo | Responsable |
|---|------------|-----------------------------|
| Formación de grupos y distribución de roles | 5 minutos | Equipo |
| Diseño de la pregunta | 5 minutos | Equipo |
| Resumen y visualizaciones de la información | 15 minutos | Analistas, Presentadores |
| Presentación de resultados | 10 minutos | Presentadores |