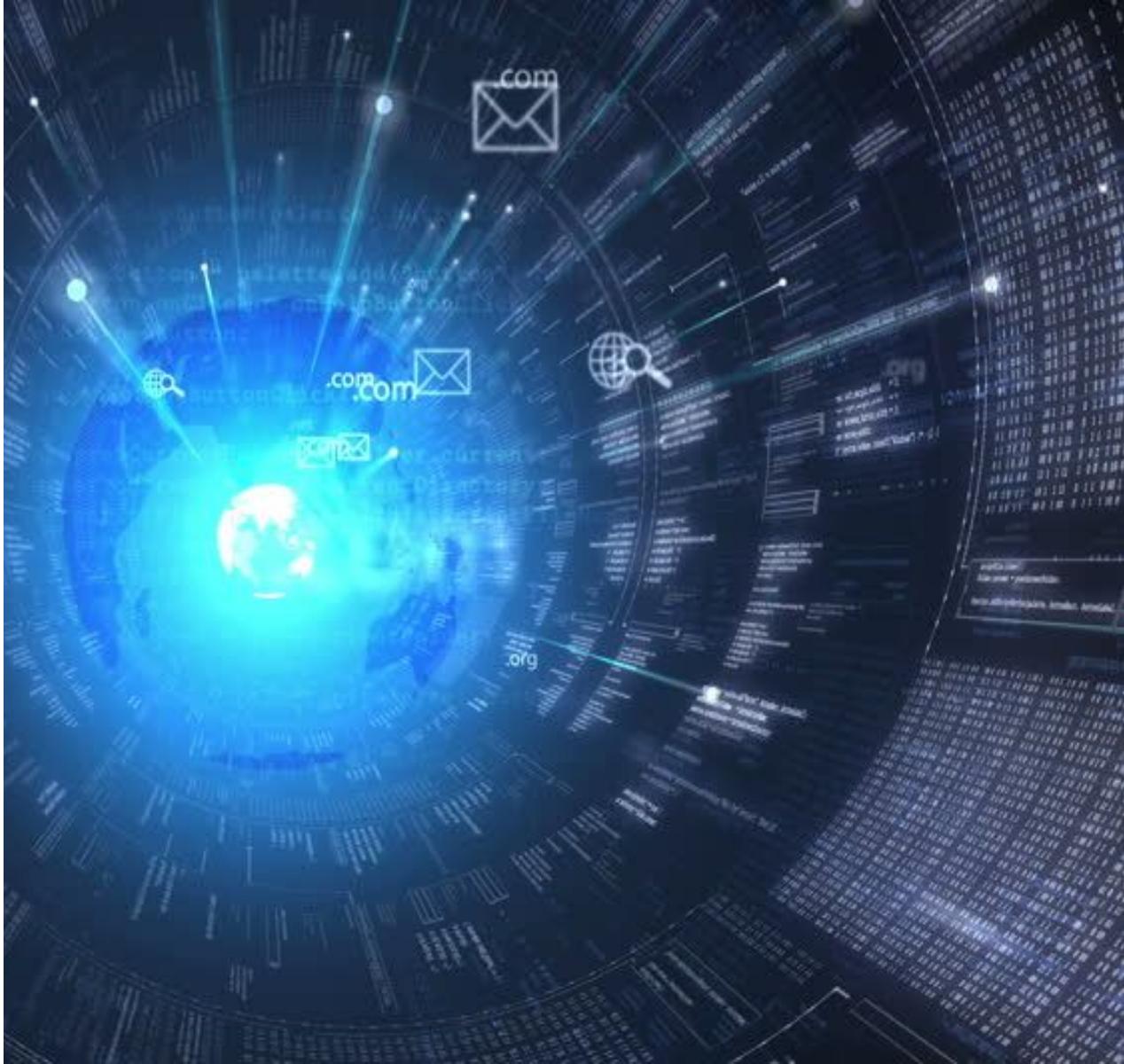




BIENVENIDOS A

Data Science



Presentación del equipo



María Frances Gaska

Programadora orientada al Data Mining. Licenciada en Economía Magna Cum Laude por la Universidad de Buenos Aires. Especializada en Explotación de Datos y Descubrimiento del Conocimiento en la Facultad de Ciencias Exactas.



Germán Rosati

Sociólogo, Doctor en Ciencias Sociales (UBA) y Master en Generación y Análisis de Información Estadística (UNTREF). Especializado en la aplicación de métodos de machine learning/ data mining a las ciencias sociales.



Matías Grinberg

Programador, Data Scientist Jr, Estudiante de Neurociencias y Músico. Estudiante de Psicología con orientación en Neurociencias (U. Favaloro), complementado con estudios en Informática, y Data Science.



Marianela Sarabia

Investigadora en desarrollo económico y consultora independiente. Lic. en Economía (UBA), MSc en Economía Laboral Aplicada (SciencesPo y U.Turín) y doctoranda en Economía (UADE)

Presentación del equipo



Martín Ríos

Estudiante avanzado de Ingeniería Civil. Especializado en Data Science también cuenta con experiencia en desarrollo y diseño Web.



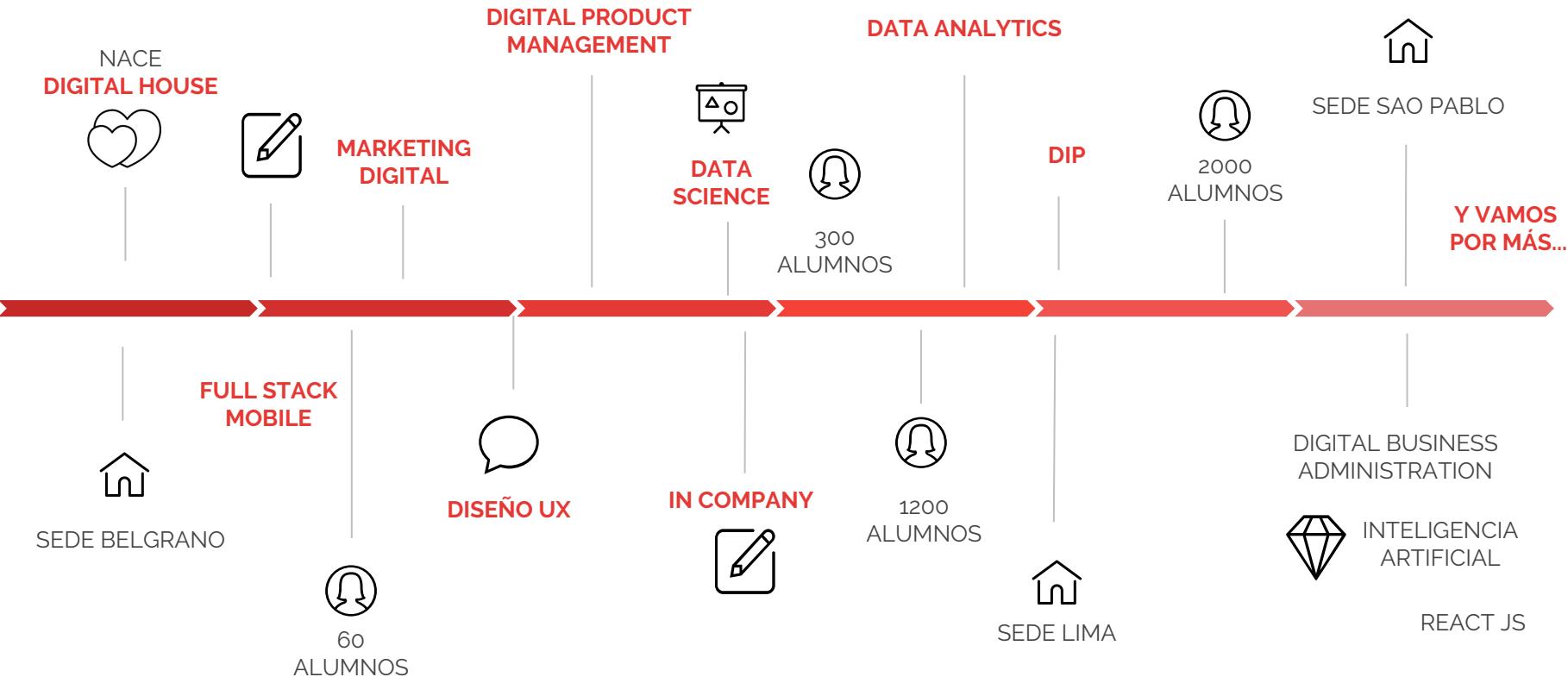
Pablo Roccatagliata

Maestría en Estadística Matemática (en curso).
.Posgrado en Economía.
Universidad Torcuato Di Tella.
de maestría. Licenciado en Economía. Universidad de Buenos Aires. Magna Cum Laude.



Demian Avendaño

Estudiante de Licenciatura de Ciencias Biológicas con orientación en Bioinformática, cursando materias de Licenciatura en Ciencias de la Computación.



¿Qué es el co learning?

Es un espacio para:

- Aprender junto a otros y trabajar en equipo.
- Hacer consultas puntuales de temas desarrollados durante las clases.
- Pensar y concretar nuevos proyectos.



Monroe: Lunes a Viernes 14 a 18hs

Lima: Lunes, Miércoles y Viernes: 18hs a 22hs

¿Qué no es el co learning?

- No es un espacio de clases particulares, ni de consultoría.
- En el co learning no se recuperan las clases en las que el alumno estuvo ausente.
- En caso de ausencia, el alumno puede acceder al material en el Campus Virtual, y asistir al co learning para aclarar dudas puntuales.





Horario de atención: 9.00 a 20.00 hs.



alumnos@digitalhouse.com



- Acompañamiento y seguimiento durante la cursada
- Soporte para acceso a campus virtual
- Asesoramiento sobre beneficios de Comunidad Digital House



¡Nuestras acciones comienzan a **mitad de semestre!**



graduados@digitalhouse.com



- Entrevistas de **Coaching** personalizado
- **Talleres** con estrategias para Cv Tech, **Linkedin**, Comunicación no Verbal en Entrevistas.
- Estrategias a **freelancers** y **emprendedores**.





Horario de atención: Lunes a
Jueves de 9.00 a 20.00 hs.
Viernes de 9:00 a 18:00 hs.



cobranzas@digitalhouse.com



Asesoramiento sobre pago de cuotas,
vencimientos, medios de pago.



- Importantes
- **Periódicas (2 o 3 semanas)**
- Detectar problemas en el curso y poder corregirlos



Eventos para alumnos



SOCIAL DAY: SAVE THE DATE!



SOCIAL DAY

AUDITORIO
12:30 A 13:30 HS /
17.30 a 19.30 hs.



**NETWORKING
DAY**

AUDITORIO
17.30 a 18.30 hs



**RECRUITING
DAY**

AUDITORIO + BAR +
ESPACIO CO
LEARNING
16:00 a 19:00 hs



SOCIAL DAY: SAVE THE DATE!



SOCIAL DAY

ESPAZIO CO
LEARNING 2DO PISO
17.30 a 19.30 hs.



**NETWORKING
DAY**

ESPAZIO CO
LEARNING 2DO PISO
17.30 a 18.30 hs



**RECRUITING
DAY**

AUDITORIO + BAR +
ESPAZIO CO
LEARNING
16:00 a 19:00 hs



Sumate a nuestras redes



<https://www.facebook.com/groups/dhcomunidad/>

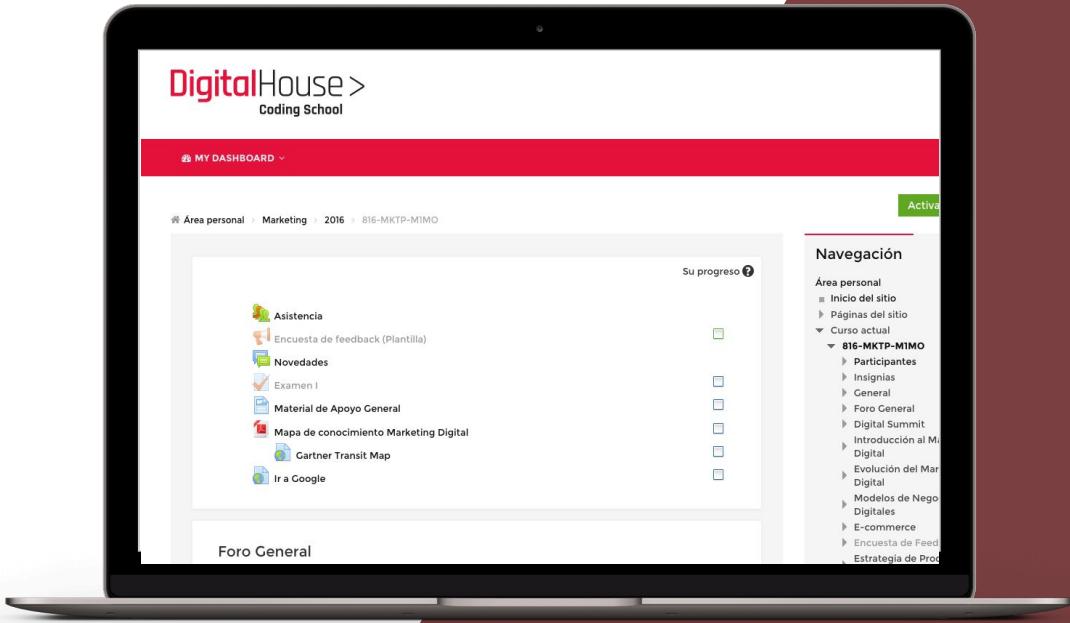
<https://www.linkedin.com/groups/12026442>

Comunidad Digital House

FORMAS DE COMUNICACIÓN

Qué canales vamos a usar y con qué finalidad cada uno

CÓMO NOS COMUNICAMOS?



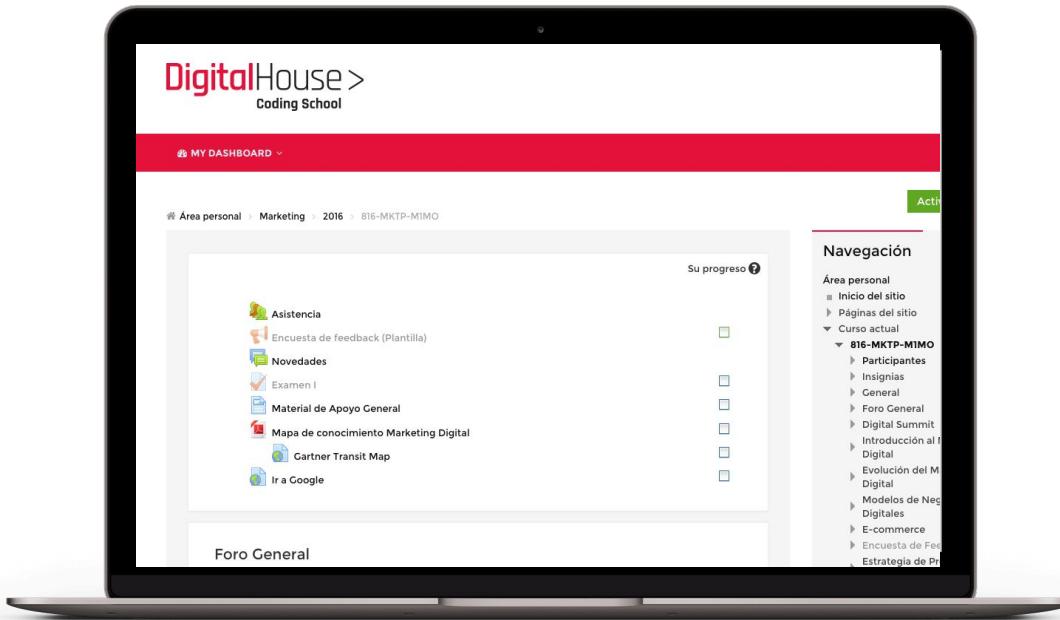
CAMPUS VIRTUAL

PARA QUÉ SIRVE?

- VER MATERIAL DE CLASES
- ENCUESTAS
- ENTERARSE DE EVENTOS, CHARLAS, ETC
- FORO: CONSULTAS O COMUNICADOS
- ACCEDER A LA BOLSA DE TRABAJO
- ENTREGA FORMAL DE TRABAJOS

WIFI: alumnos-digitalhouse

CLAVE: alumnos-digitalhouse



1

INGRESAR EN:
<http://campus.digitalhouse.com>

2

DARSE DE ALTA CON:
EMAIL:
CLAVE: Digitalhouse860!

3

SUBIR FOTO! :)

Ante cualquier inconveniente, enviar un mail a
campus@digitalhouse.com

Aclarar:

Nombre y apellido.

Curso y comisión.

Describir brevemente cuál es el problema.



SLACK

dsdh-curso.slack.com

Para estar comunicados entre nosotros, compartir
novedades e información relevante



90% DE ASISTENCIA



PARTICIPACIÓN EN CLASE



TRABAJO EN PROYECTO INTEGRADOR DURANTE TODA LA CURSADA



ENTREGA A TIEMPO DE LOS TRABAJOS



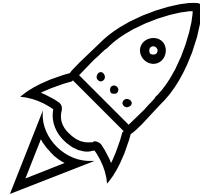
CUIDAR LOS ESPACIOS Y RECURSOS DE TRABAJO



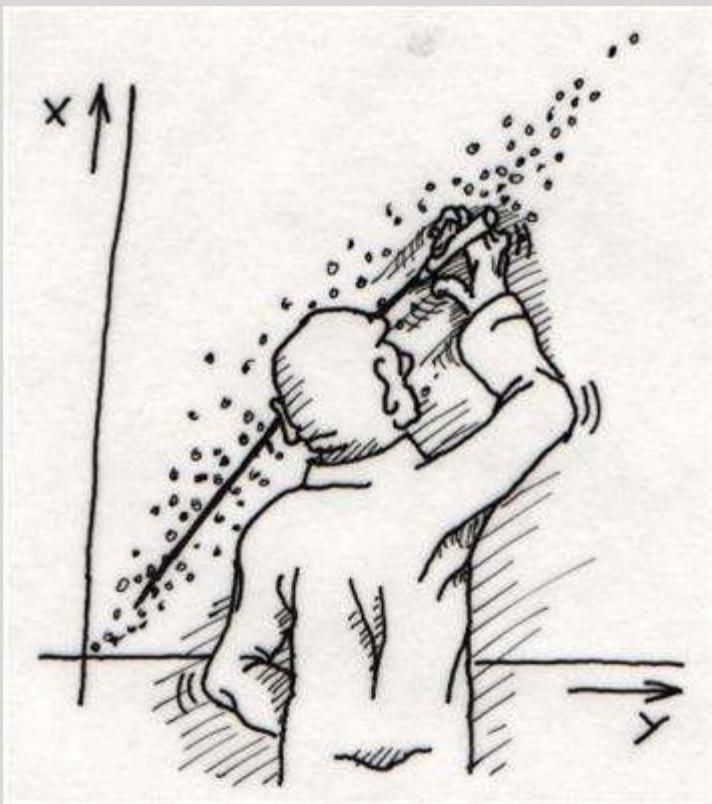
APROVECHAR EL CO LEARNING



GENERAR UN BUEN CLIMA DE TRABAJO



¡Muchas gracias!

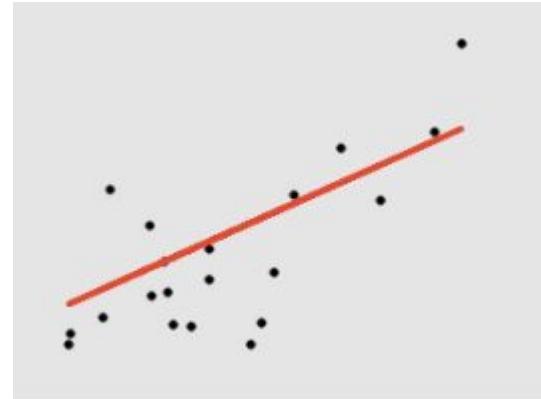


DigitalHouse >
Coding School
DATA SCIENCE

UNIDAD 1
MÓDULO 1

Presentación del
programa

- 1 Presentar la filosofía y los objetivos del programa de Data Science
- 2 Desarrollar lineamientos de clase
- 3 Discutir sobre la naturaleza de la Ciencia de Datos
- 4 Lograr que los participantes del programa se presenten y se conozcan usando el flujo de trabajo de Data Science





anokas

Mikel Bober-Irizar

Guildford, England, United Kingdom

Joined 2 years ago · last seen in the past day



<http://mxbi.net>

Followers 269

Following 26



Competitions
Master

With his skill, **enthusiasm**, and **cooperative attitude** within the community, Mikel is very much the template of a **rising star** in the Kaggle and greater AI communities.

Except for one thing: Mikel is just 16 years old.

"I don't know all the math behind the algorithms, but in terms of actually using it, I think it's much more important to have a logical understanding of how it works. Even if I can't write it from scratch, I still know what it does, and that helps me to understand where it might be useful."

1

Aprender las bases

2

Aprender a pensar

3

Aprender haciendo

4

Aprender a aprender

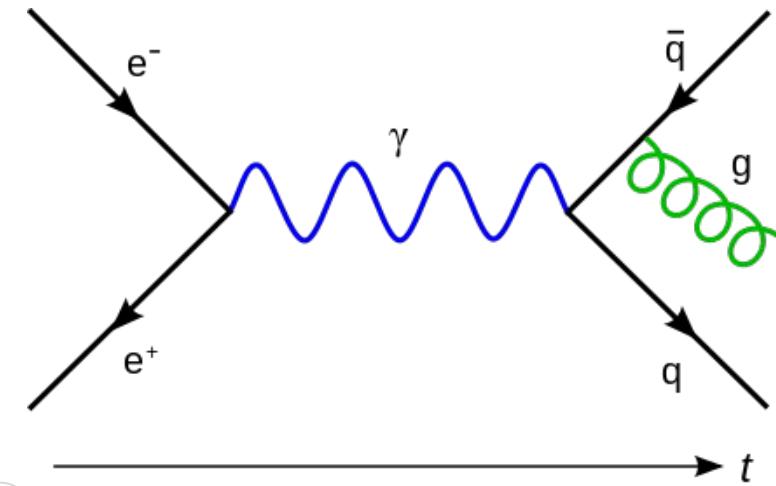


FILOSOFIA DEL PROGRAMA

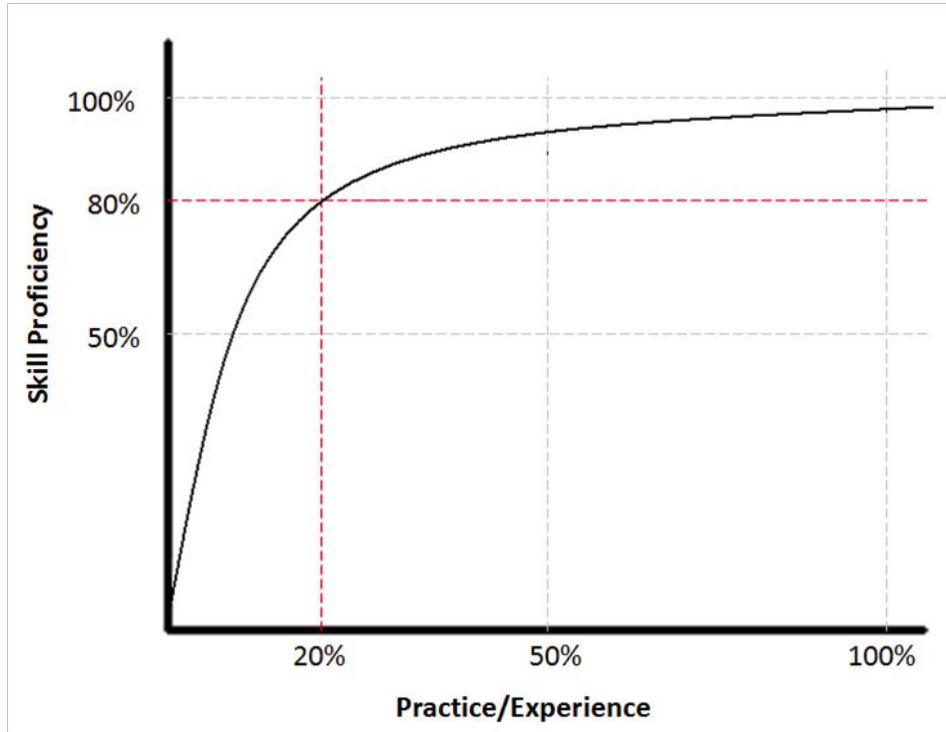


- La verdad se reconoce por su belleza y simplicidad
 - Richard Feynman
- Con el método Feynman se puede aprender de todo, desde tocar el tambor hasta física cuántica.
- La técnica de Feynman:
 - Paso 1. Elegí el concepto que quieras entender.
 - Paso 2. Simulá que le estás enseñando esta idea a alguien más.
 - Paso 3. Si no podés explicarlo bien, entonces volvé al libro.
 - Paso 4. Simplificá tu lenguaje.

Quantum field theory
Feynman diagram



- Fomentar y trabajar en un **entorno diverso**



KDnuggets™ Subscribe to KDnuggets News | [Twitter](#) | [f](#) | [in](#) | Contact

[SOFTWARE](#) | [NEWS](#) | [Top stories](#) | [Opinions](#) | [Tutorials](#) | [JOBS](#) | [Companies](#) | [Courses](#) | [Datasets](#) | [EDUCATI](#)

Table 4: KDnuggets 2016 Poll: Algorithms Used by Data Scientists

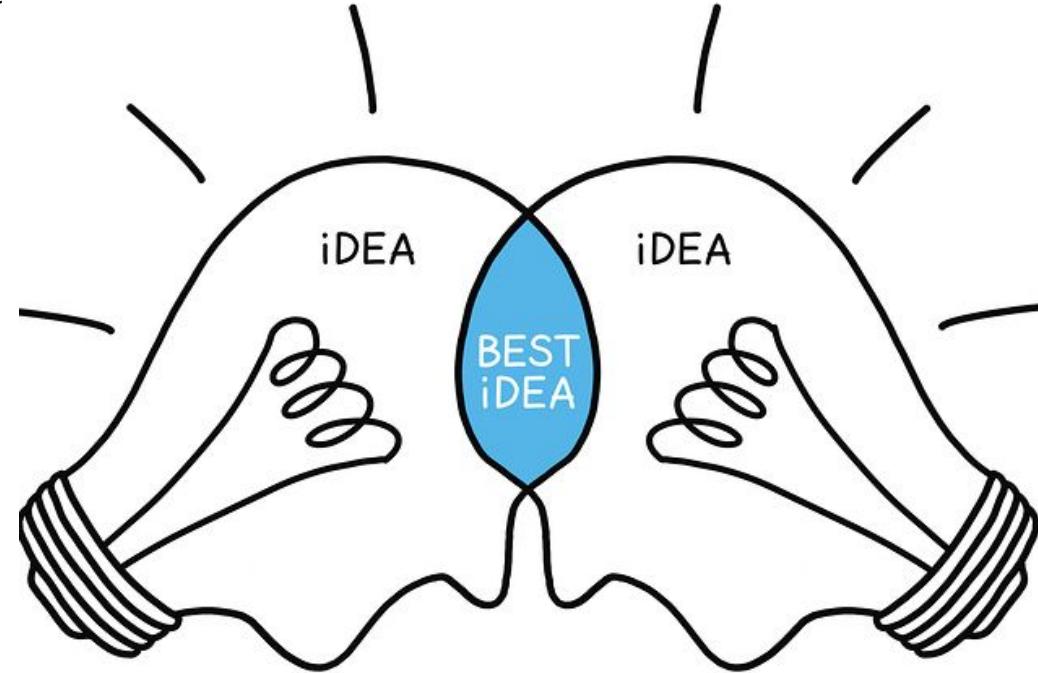
N	Algorithm	Type	2016 % used	2011 % used	% Change	Industry Affinity
1	Regression	S	67%	58%	16%	0.21
2	Clustering	U	57%	52%	8.7%	0.05
3	Decision Trees/Rules	S	55%	60%	-7.3%	0.21
4	Visualization	Z	49%	38%	27%	0.44
5	K-nearest neighbors	S	46%			0.32
6	PCA	U	43%			0.02
7	Statistics	Z	43%	48%	-11.0%	1.39
8	Random Forests	S	38%			0.22
9	Time series/Sequence analysis	Z	37%	30%	25.0%	0.69
10	Text Mining	Z	36%	28%	29.8%	0.01
11	Ensemble methods	M	34%	28%	18.9%	-0.17
12	SVM	S	34%	29%	17.6%	-0.24
13	Boosting	M	33%	23%	40%	0.24
14	Neural networks - regular	S	24%	27%	-10.5%	-0.35
15	Optimization	Z	24%			0.07
16	Naive Bayes	S	24%	22%	8.9%	-0.02
17	Bagging	M	22%	20%	8.8%	0.02
18	Anomaly/Deviation detection	Z	20%	16%	19%	1.61
19	Neural networks - Deep Learning	S	19%			-0.35
20	Singular Value Decomposition	U	16%			0.29

Subscribe to KDnuggets News | [Twitter](#) | [f](#) | [in](#)

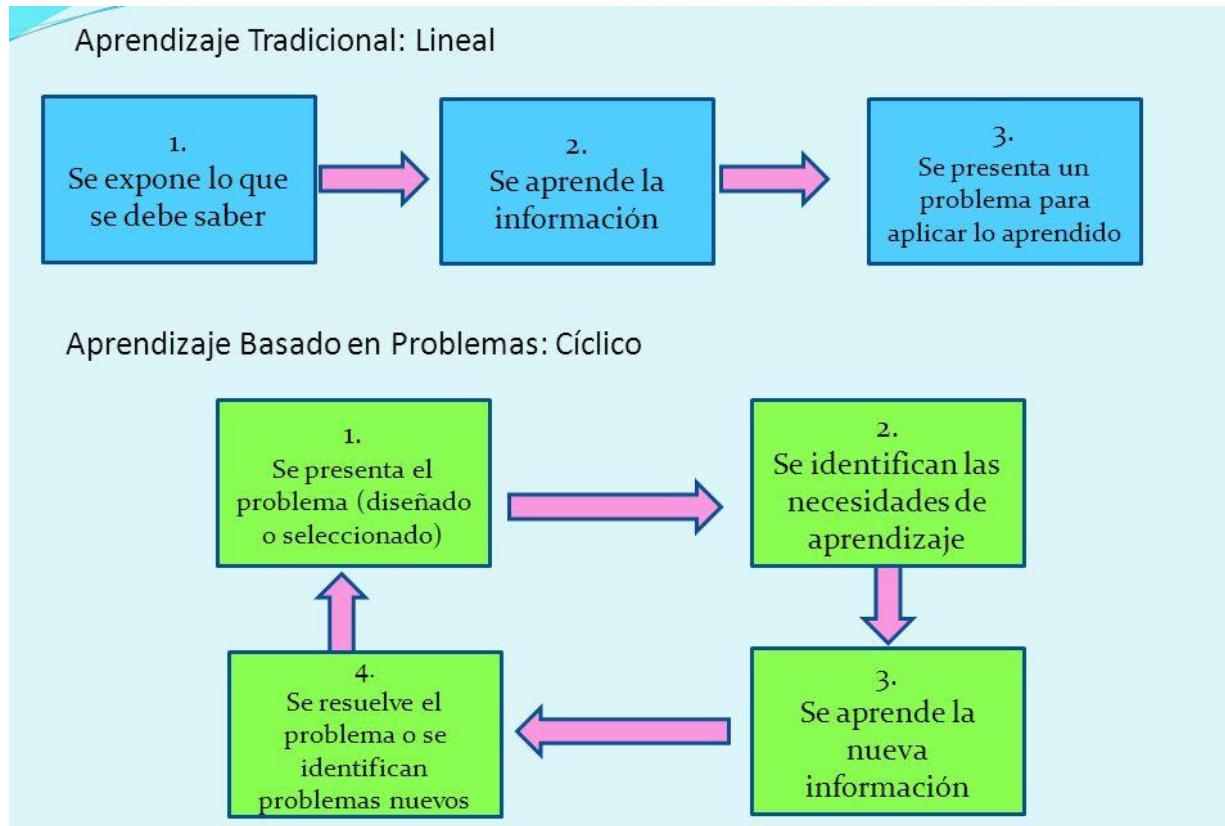
- Fomentar y trabajar en un **entorno diverso**
- Encontrar el **ritmo de aprendizaje óptimo** para cada uno
- **Comunicar** pronto y frecuentemente
- El **éxito** en este curso no se obtiene por comparación. “There is nothing noble in being superior to your fellow man; true nobility is being superior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



- La **dedicación**, más importante que el conocimiento previo
- Hacé **preguntas**, todo el tiempo, por default
- **Ayudá** a tus compañeros
- Sé **paciente** con vos mismo



APRENDIZAJE BASADO EN PROBLEMAS



PRESENTACIÓN DEL PROGRAMA

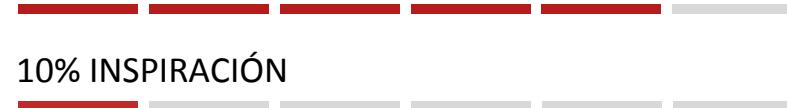


HOW TO BECOME A ROCKSTAR ON XXX

APROBACIÓN DEL CURSO

90% ESFUERZO

10% INSPIRACIÓN





90% DE
ASISTENCIA

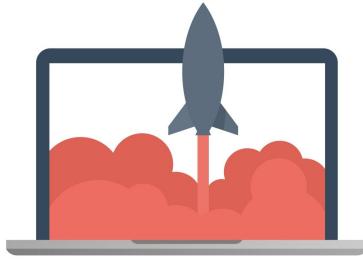


APROBACIÓN DE
TRABAJOS
PARCIALES



APROBACIÓN
DEL PROYECTO
INTEGRADOR



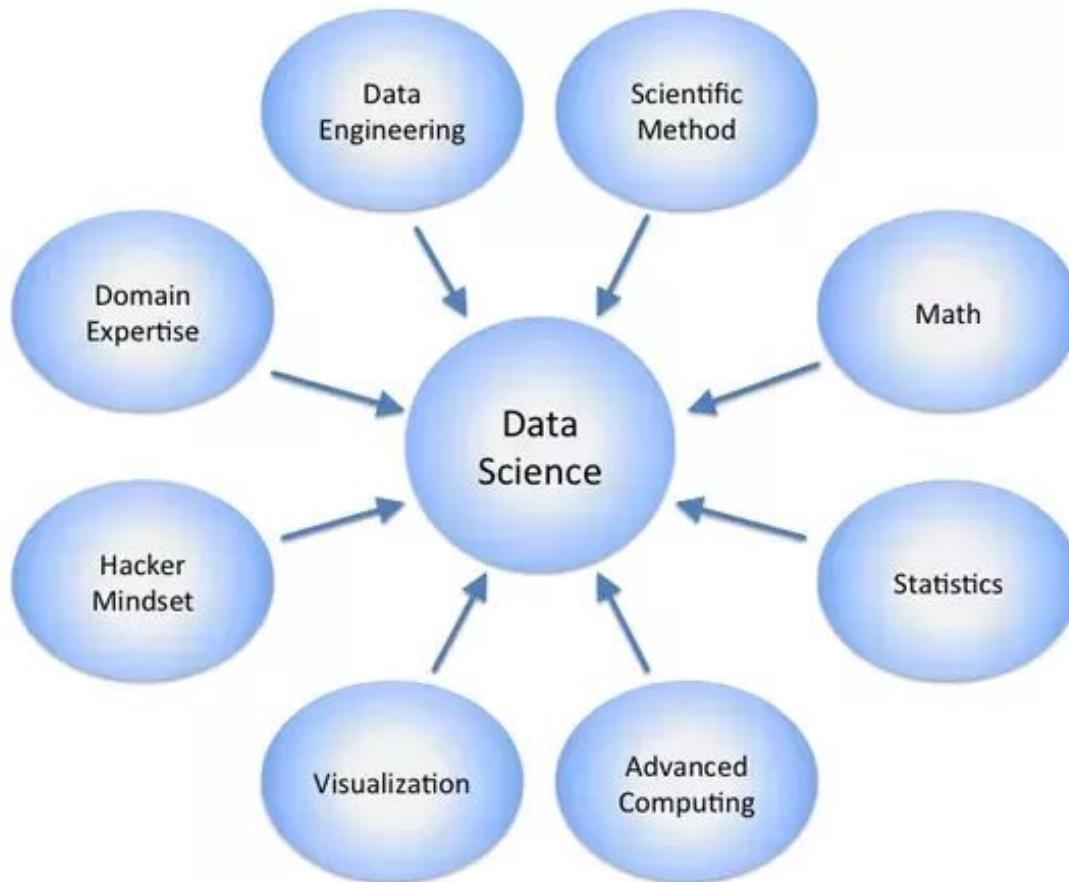


OBJETIVOS DEL CURSO

- Conocer los fundamentos de estadística y programación en Python.
- Comprender las mejores prácticas para cada parte del Workflow de un científico de datos.
- Comprender e implementar las principales técnicas de Aprendizaje Supervisado y No Supervisado

MÓDULOS







>

Fundamentos:
Numpy, SQL y
Visualización

01



>

Text Mining, Time
Series &
Clasificación

05



>

EDA, Pandas &
Limpieza de
datos

02



>

APIs, Árboles y
Métodos de
Ensamble

06



>

Regresión Lineal,
y Sklearn

03



>

PCA, Clustering,
Manifold

07



>

Regresión
Logística y Web
Scraping

04



>

PROYECTO
INTEGRADOR



>

Fundamentos:
Numpy, SQL y
Visualización

01

- Introducción al programa y a la disciplina
- Resumen Python
- Numpy
- Estadística Descriptiva con Numpy
- Introducción a la Visualización de Datos
- Presentaciones de resultados del Desafío 1

Desafío del Módulo

Provistos de un dataset de puntajes SAT de todo EEUU, los participantes realizarán un análisis exploratorio utilizando Numpy y Matplotlib aplicando técnicas básicas de estadística descriptiva.



EDA, Pandas &
Limpieza de
datos

02

- Pandas
- Probabilidad
- Limpieza de Datos
- Variables Dummies
- Datos Faltantes
- Estadística Inferencial
- Joins con Pandas
- Visualización

Desafío del Módulo

Usando un dataset crudo de hits de Billboard, los participantes usarán Pandas para limpiar los datos, plantearán formalmente un problema y realizarán análisis exploratorio para un publisher de música



>

Regresión Lineal,
y Sklearn

03

- Introducción a Machine Learning
- Regresión Lineal
- Intro a Stats Models & Sklearn
- Descomposición Bias-Varianza
- Regularización & Sobreajuste (Overfitting)
- Separación Entrenamiento/Test
- Lab: Workflow de Datos Limpieza y Optimización

- Métricas de Regresión & Funciones de Pérdida (Loss Functions)
- Descenso del gradiente
- Feature Scaling (Normalización)

Desafío del Módulo

Dado el acceso a datos de ventas, los participantes tendrán la posibilidad de elegir entre realizar una investigación de marketing o llevar adelante una auditoría impositiva, usando Pandas, Statsmodels y sklearn para transformar los datos, realizar una regresión lineal y visualizar los resultados



>

Regresión
Logística y Web
Scraping

04

- Intro a Clasificación
- Lab: Práctica de Scraping
- Intro a Regresión Logística
- Support Vector Machines
- Naive Bayes Classifiers
- Evaluación de modelos

Desafío del Módulo

Actuando como un federal contractor, los participantes van a hacer scraping de un sitio web y utilizar Pandas, Statsmodels y Sklearn para analizar datos, realizar regresión logística y evaluar coeficientes de correlación.



Text Mining, Time
Series &
Clasificación

05

- Text Mining
- Time Series
- Pipelines
- Feature Selection

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



>

APIs, Árboles y
Métodos de
Ensamble

06

- Intro a CARTS
- APIs & JSON
- Joins SQL
- Árboles de Decisión y Bagging
- Random Forests y Boosting
- Evaluación de Modelos y Feature Importance

Proyecto Integrador

Informe de avance:

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



>

PCA, Clustering,
Manifold, MCA

07

- Intro a Clustering
- K-means
- Intro a Clustering Jerárquico
- DBSCAN
- PCA
- Manifold Learning
- Pipeline de Detección de Caras

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

DESAFIOS Y PROYECTO INTEGRADOR



PROYECTO INTEGRADOR



TRABAJO EN
EQUIPO



ELECCIÓN DEL
PROYECTO



APLICAR LOS
CONOCIMIENTOS
APRENDIDOS



PRESENTACIÓN DEL
PROYECTO



— 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

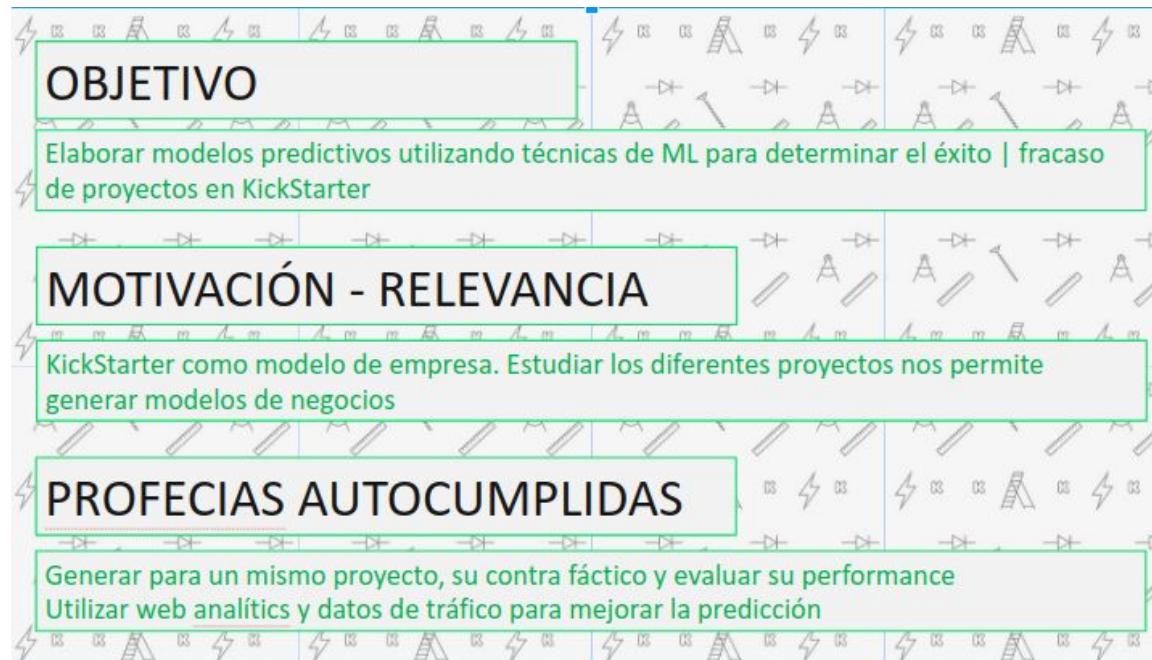
ESCUCHA DE REDES SOCIALES PARA LA GESTIÓN PÚBLICA

Francisco PENSA



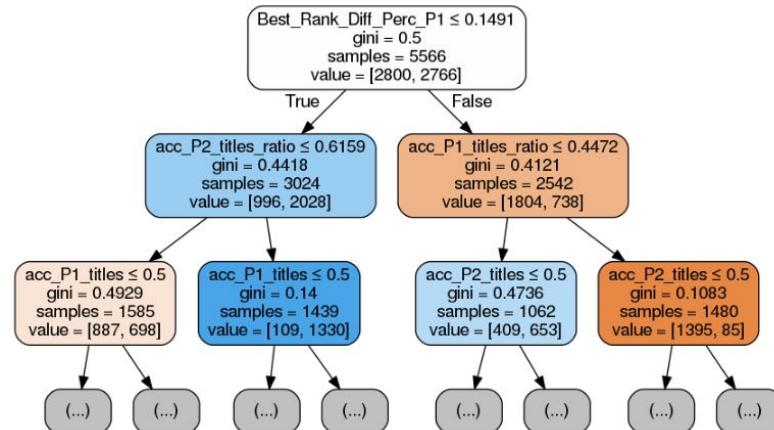
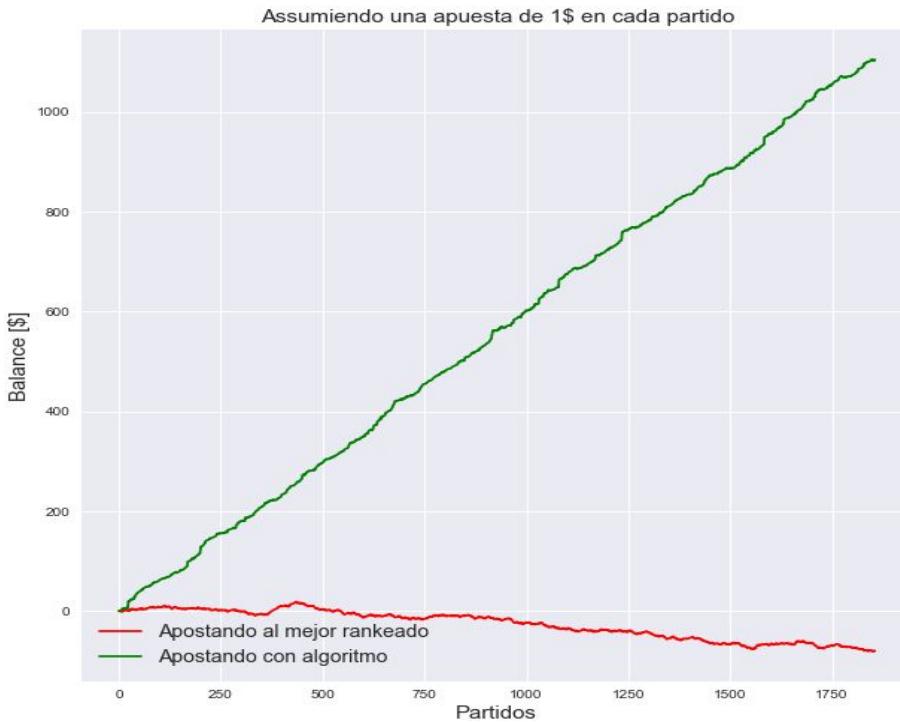
KICK-ASS MACHINE LEARNING: ¿QUÉ DETERMINA EL EXITO DE PROYECTOS EN LA PLATAFORMA KICKSTARTER?

José SANCHEZ, Jonatah COHEN



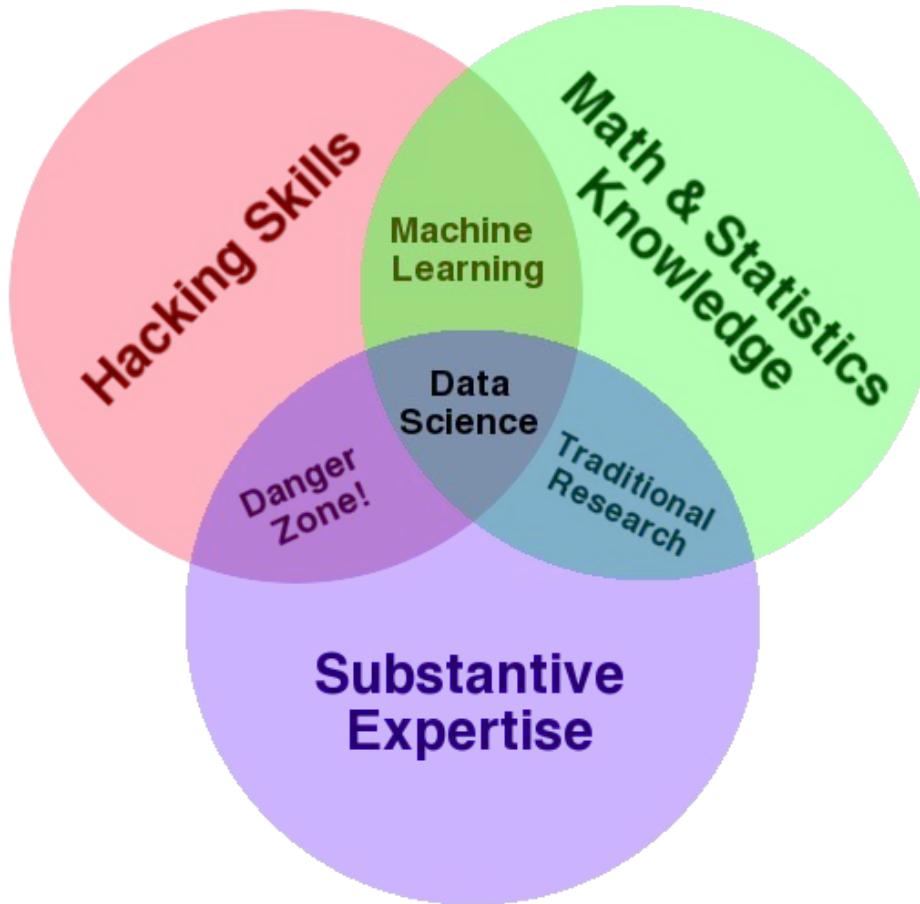
PREDICCIÓN DE RESULTADOS EN PARTIDOS DE TENIS DEL CIRCUITO ATP

Carlos RAMIRO, Franco CATANIA



¿QUÉ ES DATA SCIENCE?



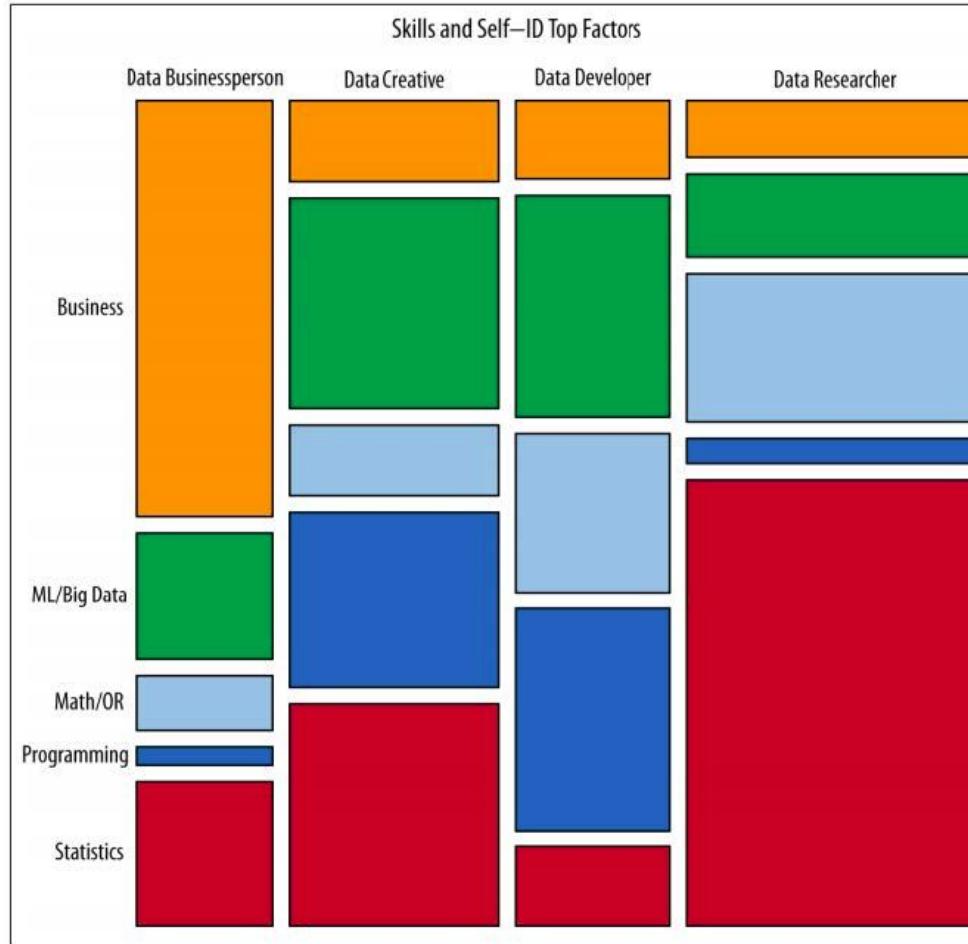


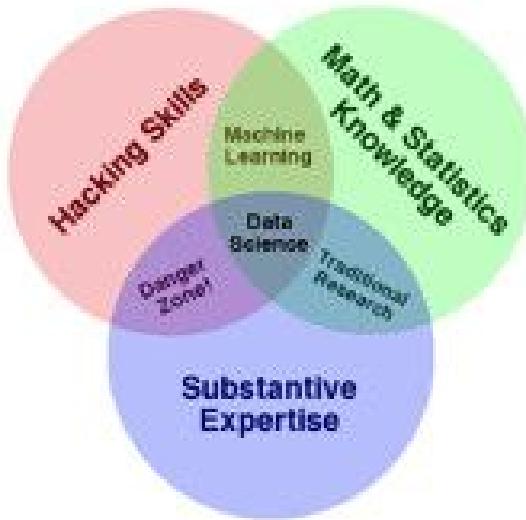
- Un set de herramientas y técnicas para extraer información útil de los datos
- Una práctica interdisciplinaria orientada a **resolver problemas**
- La aplicación de técnicas científicas a problemas prácticos
- ¿Quién usa Data Science?
 - Recomendaciones de películas Netflix
 - Algoritmo Amazon: "si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

Data Developer	Developer	Engineer	
Data Researcher	Researcher	Scientist	Statistician
Data Creative	Jack of All Trades	Artist	Hacker
Data Businessperson	Leader	Businessperson	Entrepeneur

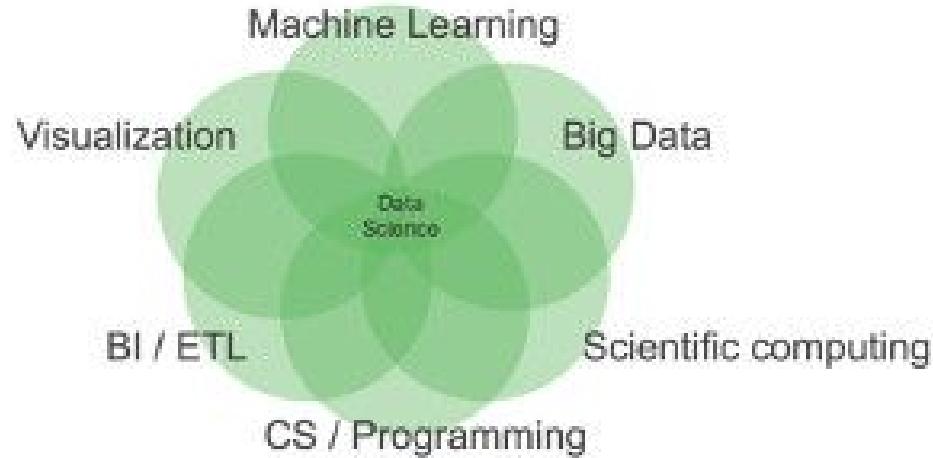
Business	ML / Big Data	Math / OR	Programming	Statistics
Product Development	Unstructured Data	Optimization	Systems Administration	Visualization
Business	Structured Data	Math	Back End Programming	Temporal Statistics
	Machine Learning	Graphical Models	Front End Programming	Surveys and Marketing
	Big and Distributed Data	Bayesian / Monte Carlo Statistics		Spatial Statistics
		Algorithms		Science
		Simulation		Data Manipulation
				Classical Statistics

¿QUÉ HABILIDADES DEBERÍA TENER CADA ROL EN UN EQUIPO DE DATA SCIENCE?

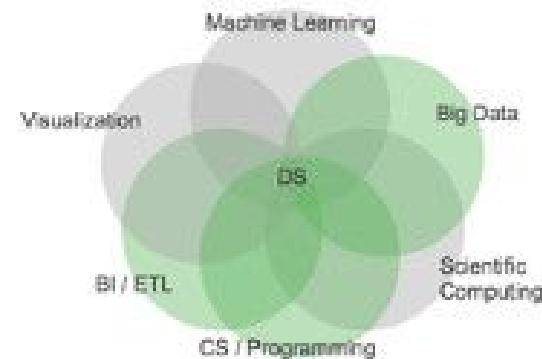
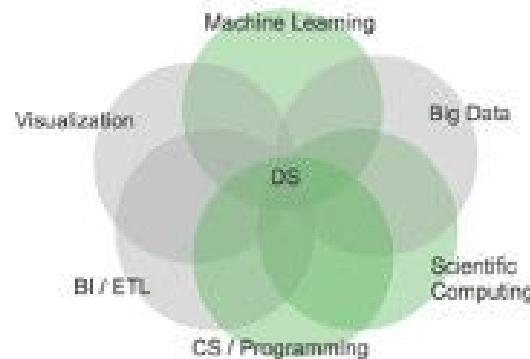
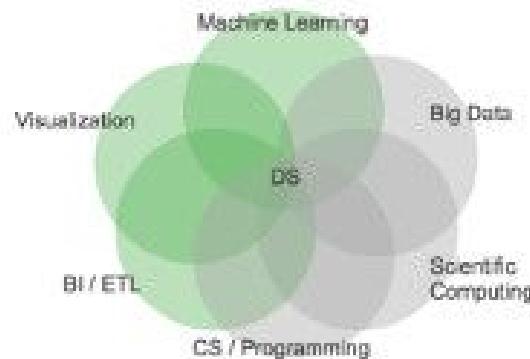




Traditional Data Science Venn Diagram



Revisited Data Science Venn Diagram



Statistician / Analyst

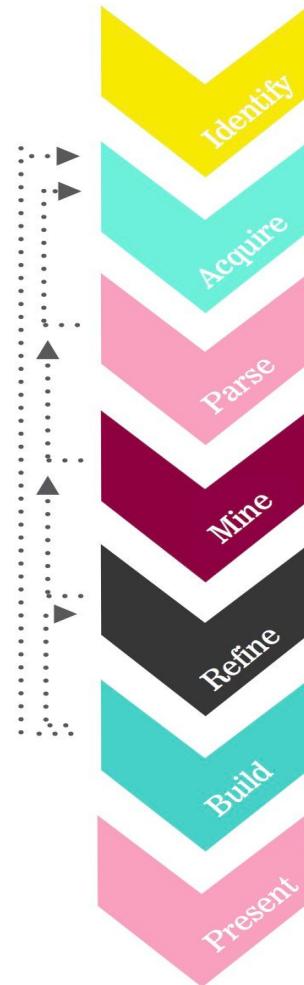


Research / Computational
Scientist

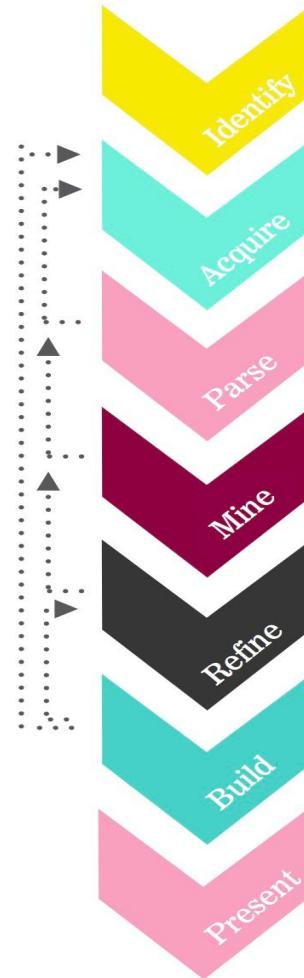


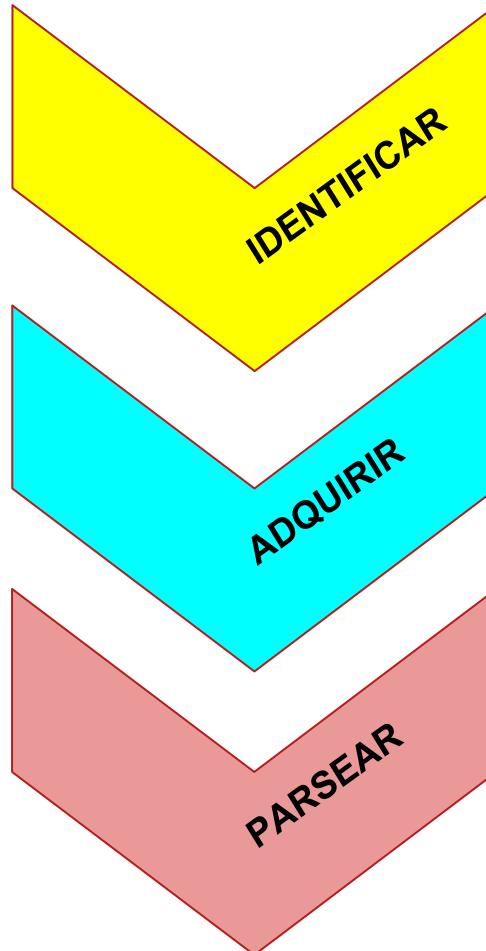
Developer / Engineer

- A lo largo de las clases seguiremos el “Flujo de trabajo de Data Science”. Nos servirá 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

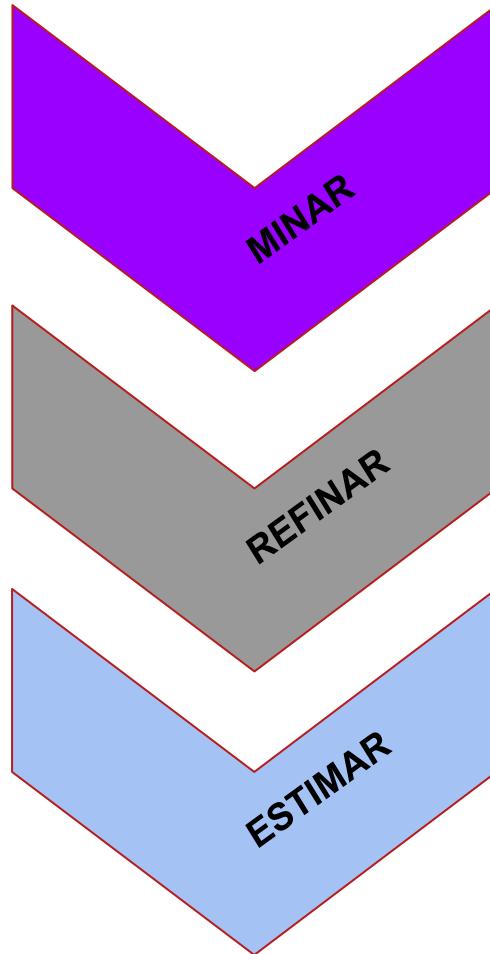
- Identificar los objetivos del producto/negocio/problema
- Identificar y generar hipótesis sobre metas y criterios para el éxito del análisis
- Generar un set de preguntas para identificar el dataset “correcto”.

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

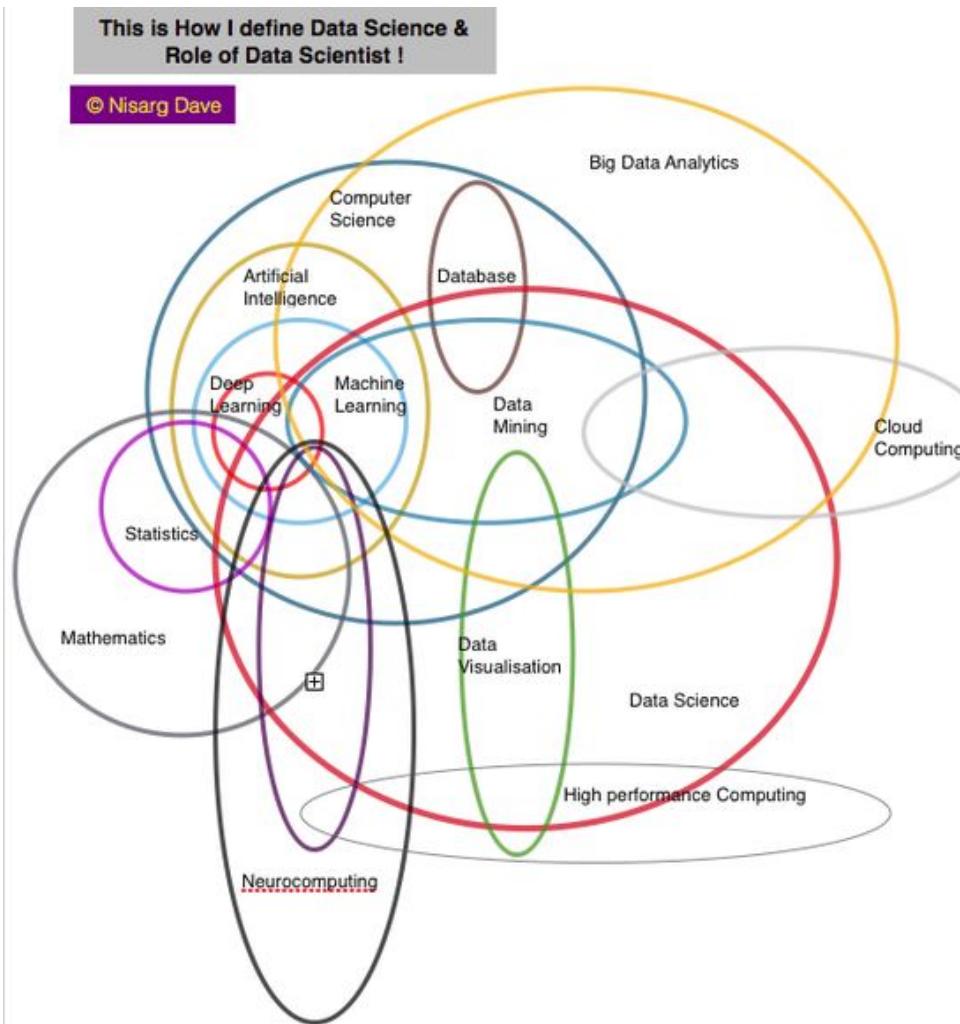
ESTIMAR UN MODELO

- 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



FAIL



DigitalHouse >
Coding School

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

Te proponemos

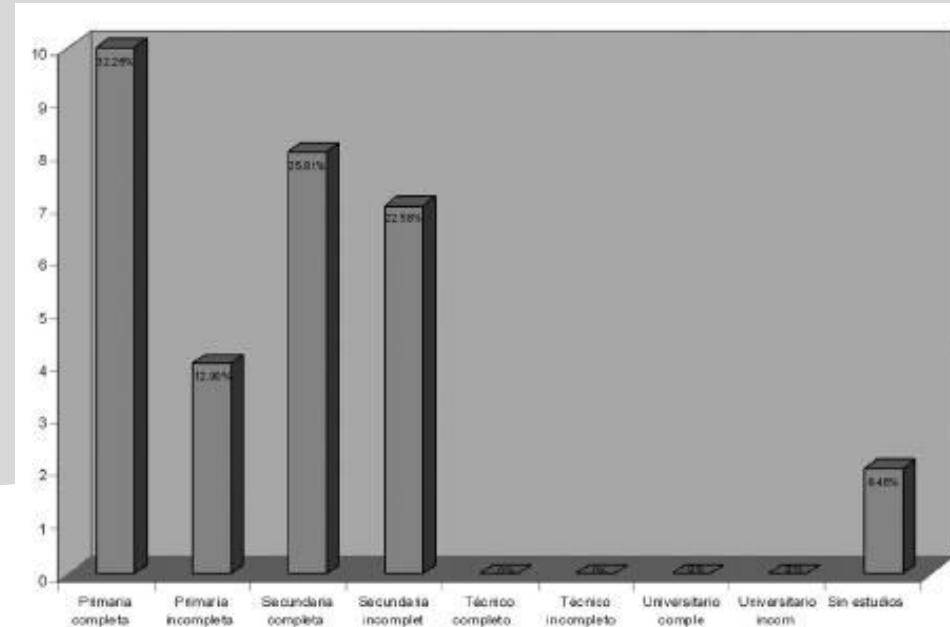
- Que todos los participantes del programa se conozcan mutuamente usando algunos pasos del Flujo de Trabajo de Data Science.
- Que formen grupos de 4 a 6 personas
- Que cada grupo defina **una** pregunta sobre algún aspecto que le interese conocer acerca de los compañeros (motivación, formación, etc.)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

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.

- 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.