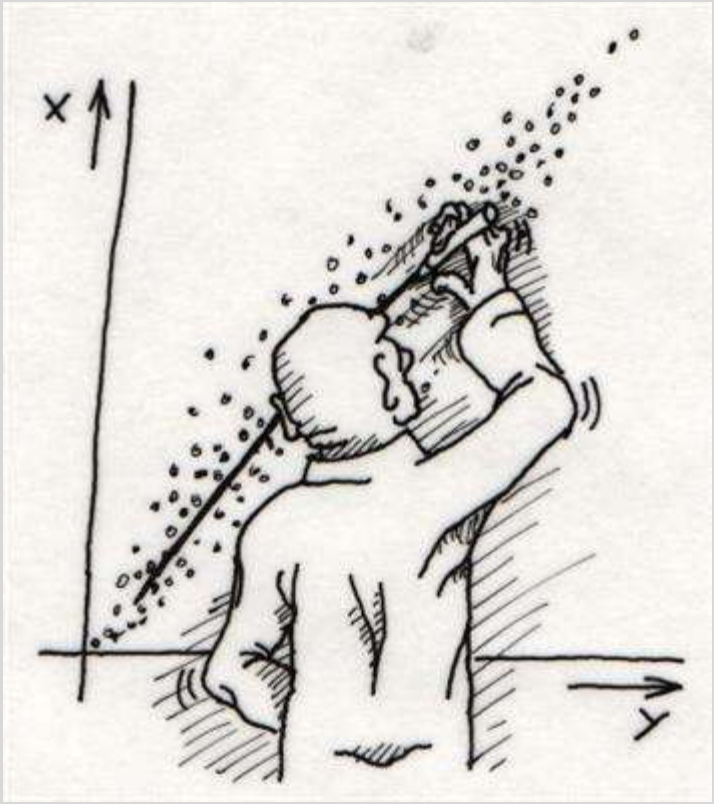



DigitalHouse > Coding School

# DATA SCIENCE

UNIDAD 1
MÓDULO 1

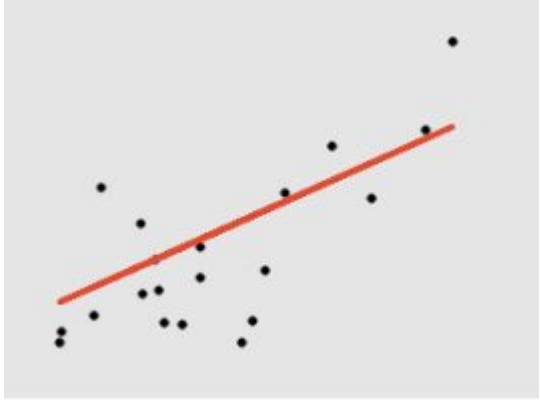
Presentación del programa

Agosto 2017

# PRESENTACIÓN DEL PROGRAMA

# OBJETIVOS DE LA CLASE

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**

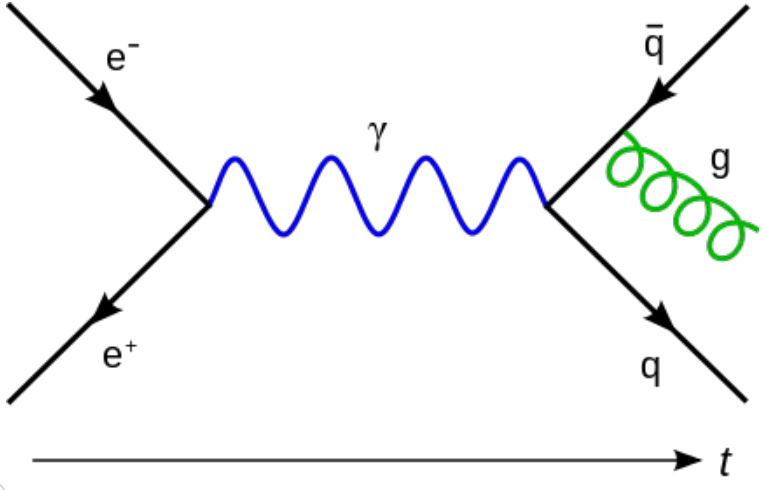


1. Aprender las bases
2. Aprender a pensar
3. Aprender haciendo
4. Aprender a aprender

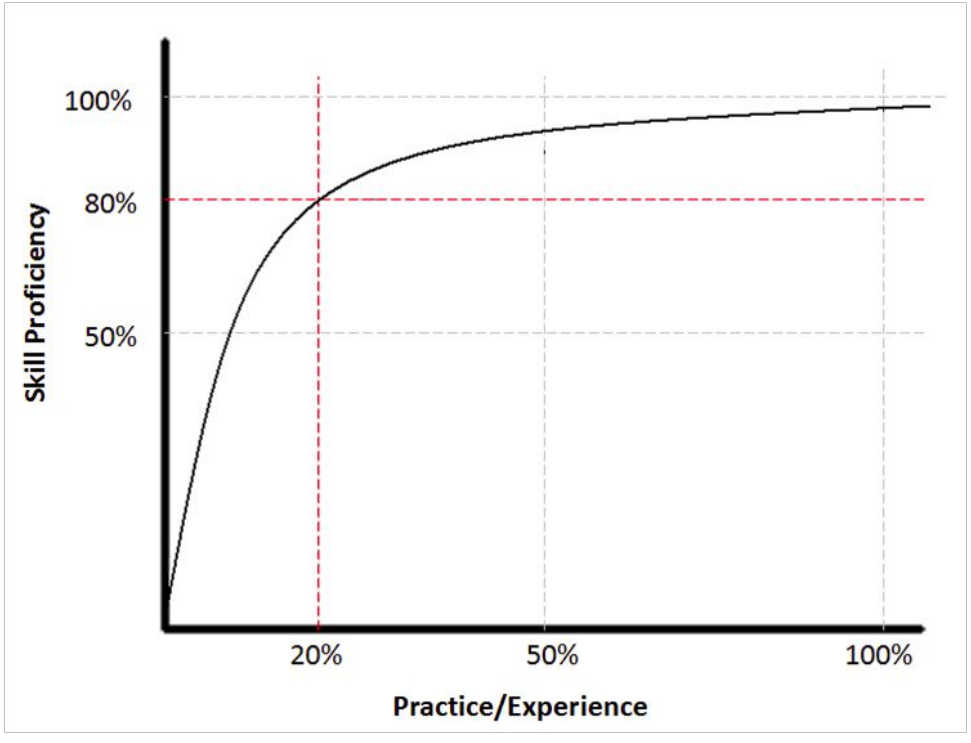
# FILOSOFIA DEL PROGRAMA

- 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 puedes explicarlo bien, entonces vuelve al libro.
  - Paso 4. Simplificá tu lenguaje.

**Quantum field theory**
Feynman diagram



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



# El 20% de los algoritmos que genera el 80% del impacto en el negocio

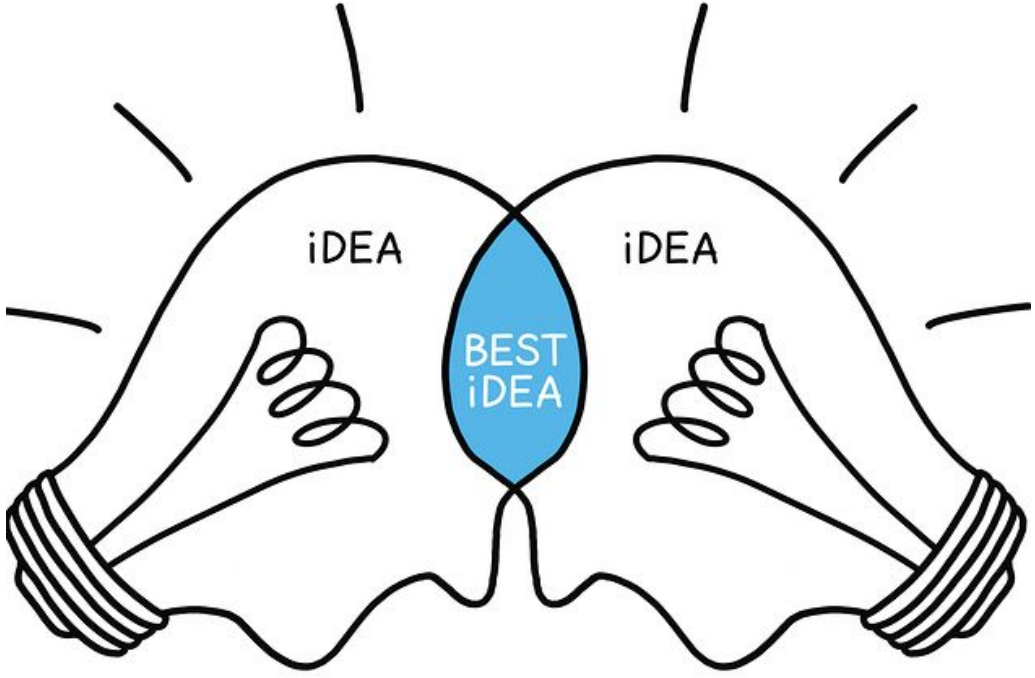


Table 4: KDnuggets 2016 Poll: Algorithms Used by Data Scientists

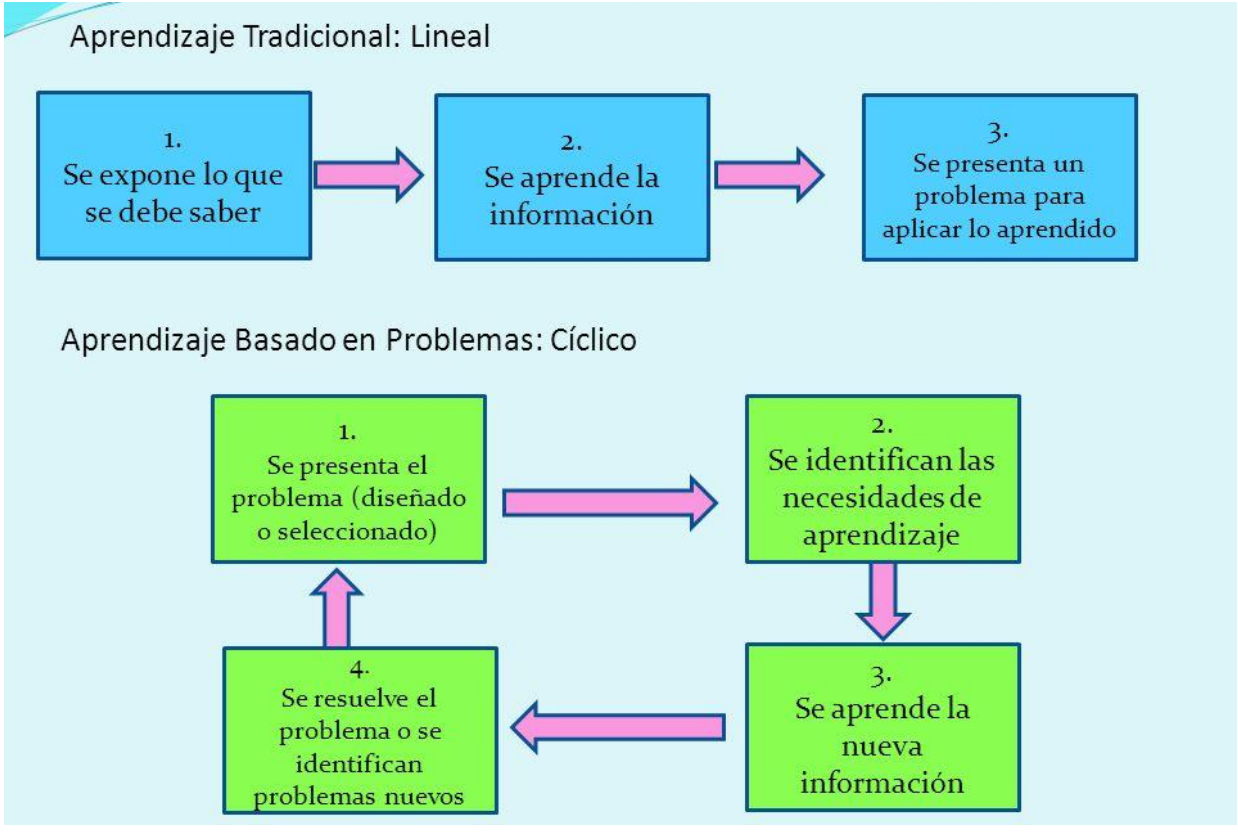
| N | Algorithm | Type | 2016 % used | 2011 % used | % Change | Industry Affinity |
|---|---|---|---|---|---|---|
| 1 | Regression | S | 67% | 58% | 16% | 0.21 |
| 2 | Clustering | U | 57% | 52% | 8.7% | 0.05 |
| 3 | Decision Trees/Rules | S | 55% | 60% | -7.3% | 0.21 |
| 4 | Visualization | Z | 49% | 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 |

- 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



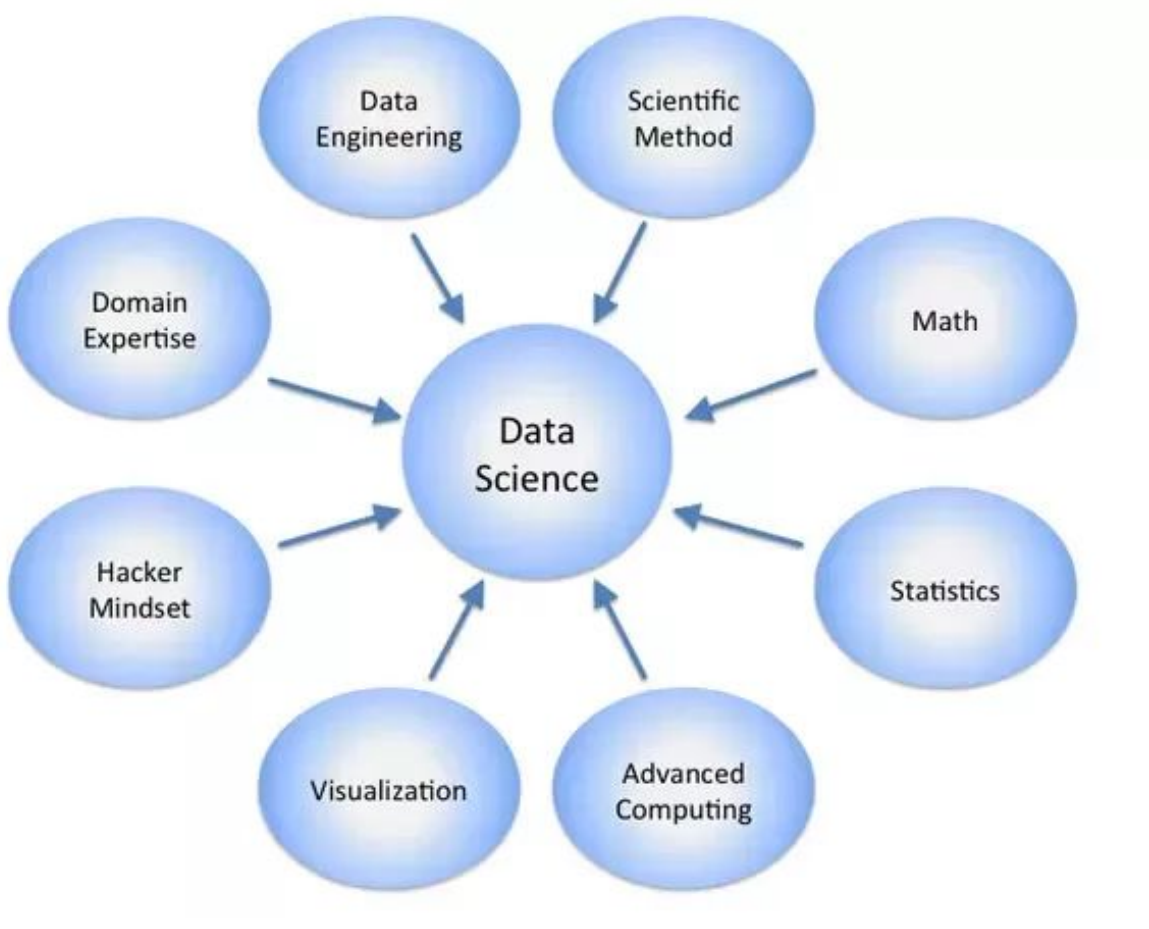
# LINEAMIENTOS DE CLASE

- Sistemas a utilizar:
  - Todos **Open Source**
  - Sistema Operativo: **Ubuntu**
  - Lenguaje: **Python 3.6** (distribución Anaconda)
  - **Librerías de Python**: muchas, scypy, numpy, pandas, sklearn, BeatifulSoup, MatPlotLib, Bokeh
  - Bases de datos: **PostgreSQL**

- **Comunicación** entre participantes
  - **Slack es nuestro principal medio de comunicación**
  - Vía correo electrónico: subject específico
  - Personalmente en el espacio de colearning


dsdh-curso.slack.com

# MÓDULOS


Data Engineering
Scientific Method
Domain Expertise
Math
Data Science
Hacker Mindset
Statistics
Visualization
Advanced Computing

# PROGRAMA - MÓDULOS

DigitalHouse > Coding School

01 Fundamentos: Numpy, Stats y Visualización

02 EDA, Pandas & Limpieza de datos

03 Regresión Lineal, y Sklearn

04 Regresión Logística y Web Scraping

05 SQL, Bases de Datos, & Clasificación

06 APIs, Árboles y Métodos de Ensamble

07 PCA, Clustering, Manifold, MCA

PROYECTO INTEGRADOR

DigitalHouse > Coding School

**Fundamentos: Numpy, Stats 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.

SQL, Bases de Datos, & Clasificación

05

- Bases de Datos
- Intro a SQL
- 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
- Análisis Correspondencias Múltiples
- 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 proyecti
  - Objetivos
  - Datasets
  - Métodos
  - Visualizaciones
  - Storytelling

# DESAFÍOS Y PROYECTO INTEGRADOR

PROYECTO INTEGRADOR

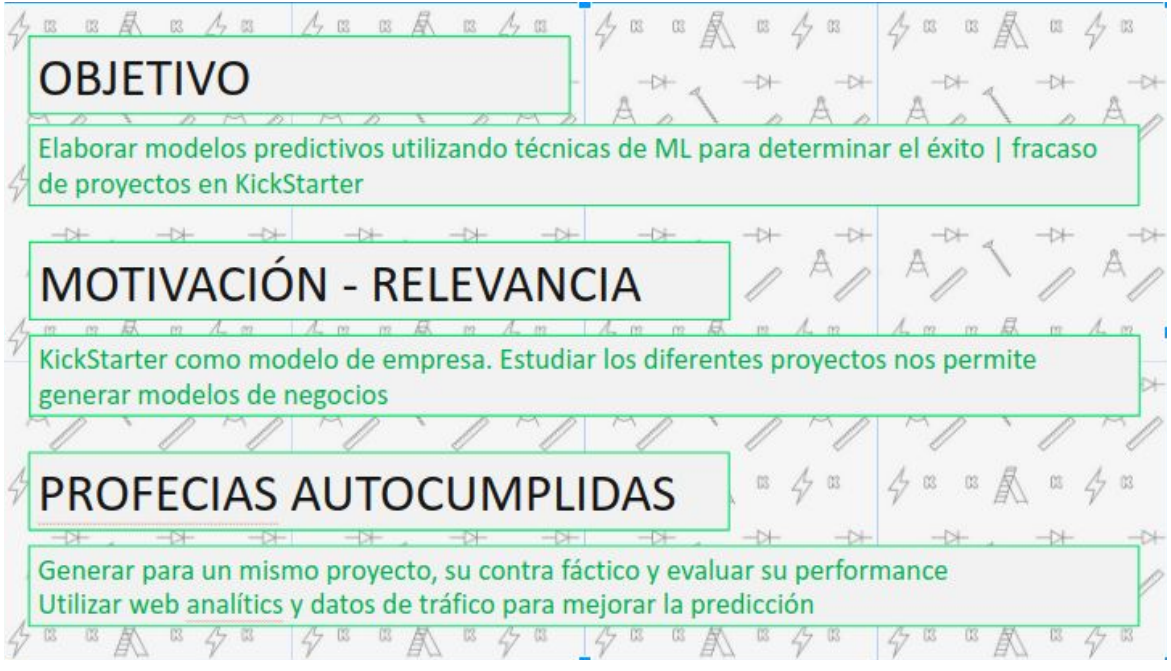
- **Desafíos y proyectos - objetivos generales:**
  - Resolver un problema práctico
  - Generar un reporte técnico (con código y análisis)
  - Generar un reporte para una audiencia no técnica
- **Desafíos Final: Proyecto Integrador: recorrer todo el Flujo de Trabajo de Data Science**
  - Planteo y fundamentación 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

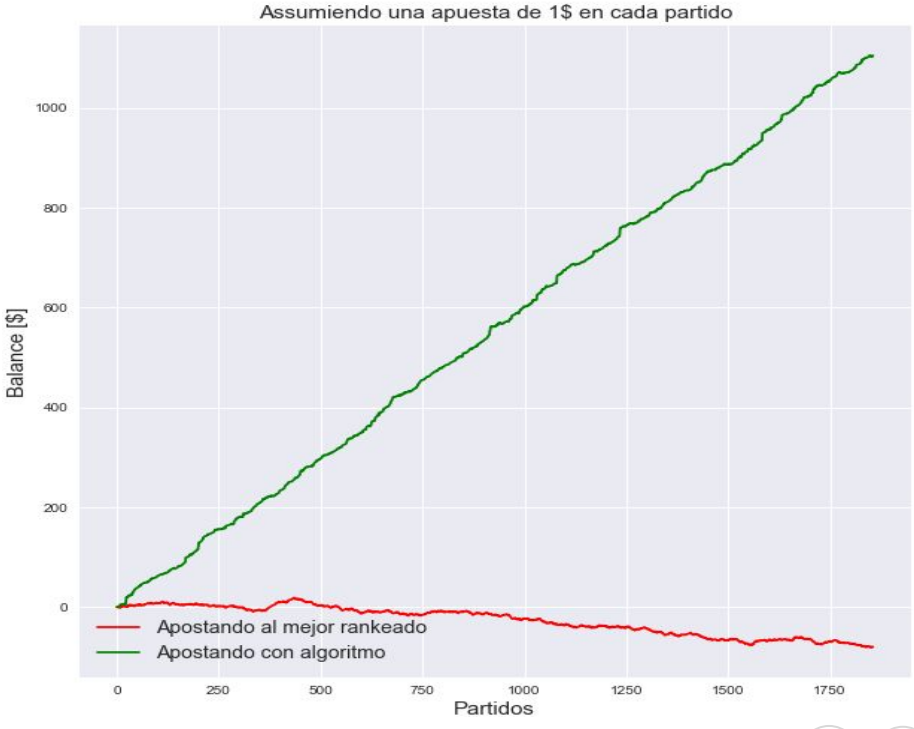
# KICK-ASS MACHINE LEARNING: ¿QUÉ DETERMINA EL EXITO DE PROYECTOS EN LA PLATAFORMA KICKSTARTER?

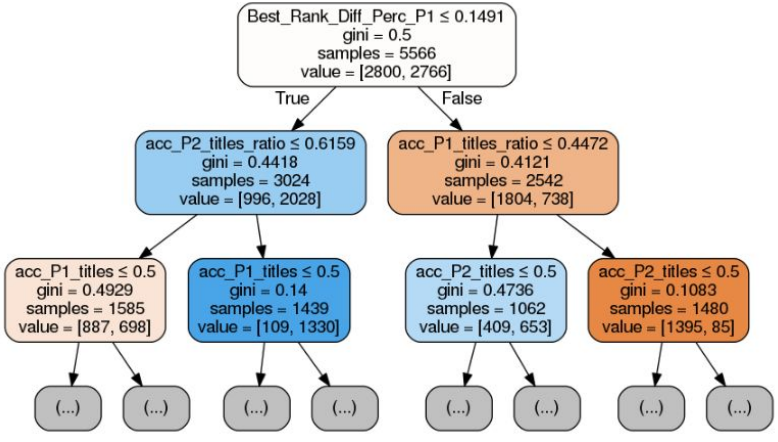
José SANCHEZ, Jonatah COHEN



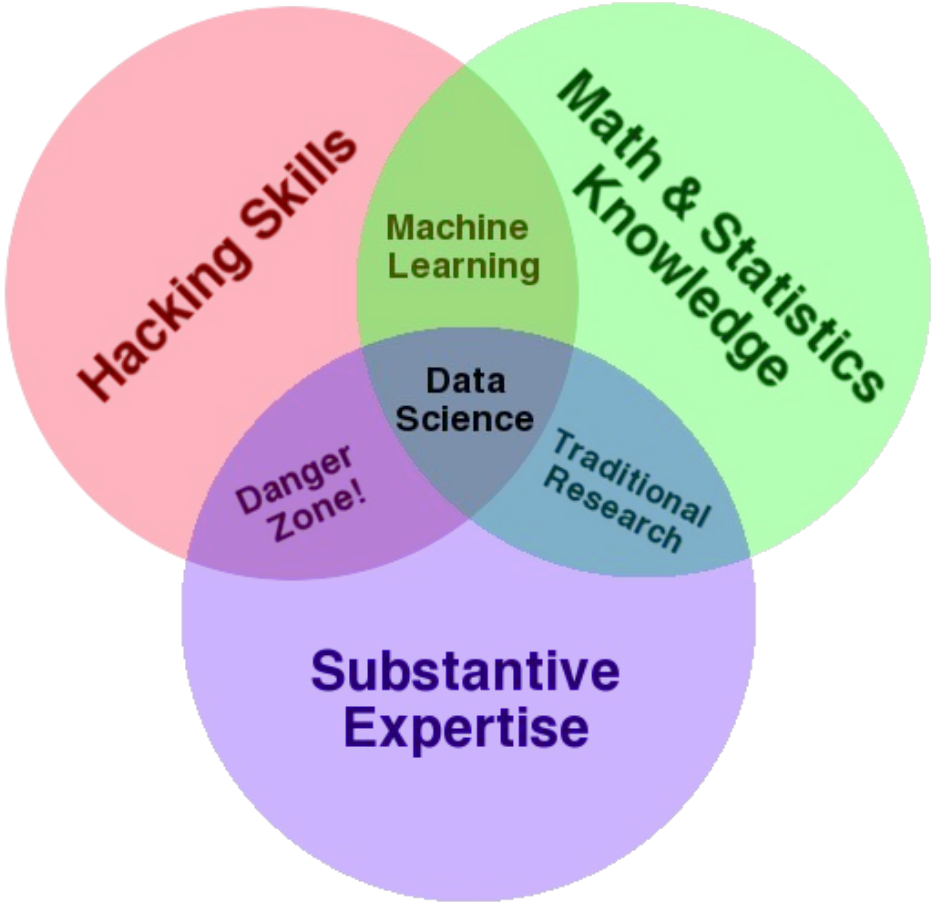
# PREDICCIÓN DE RESULTADOS EN PARTIDOS DE TENIS DEL CIRCUITO ATP

Carlos RAMIRO, Franco CATANIA

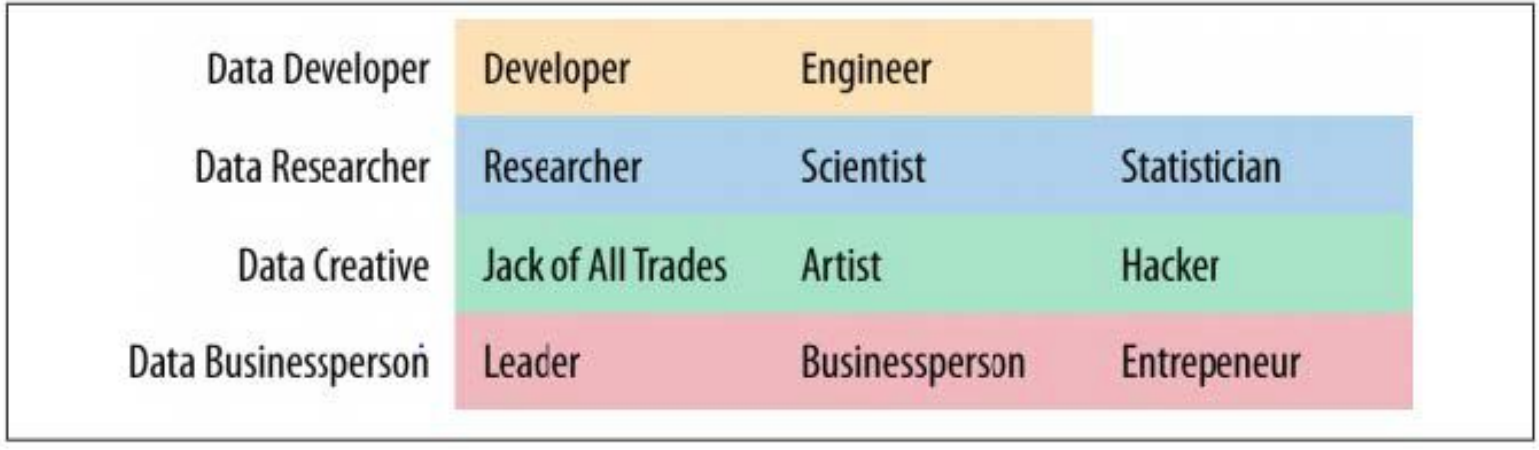


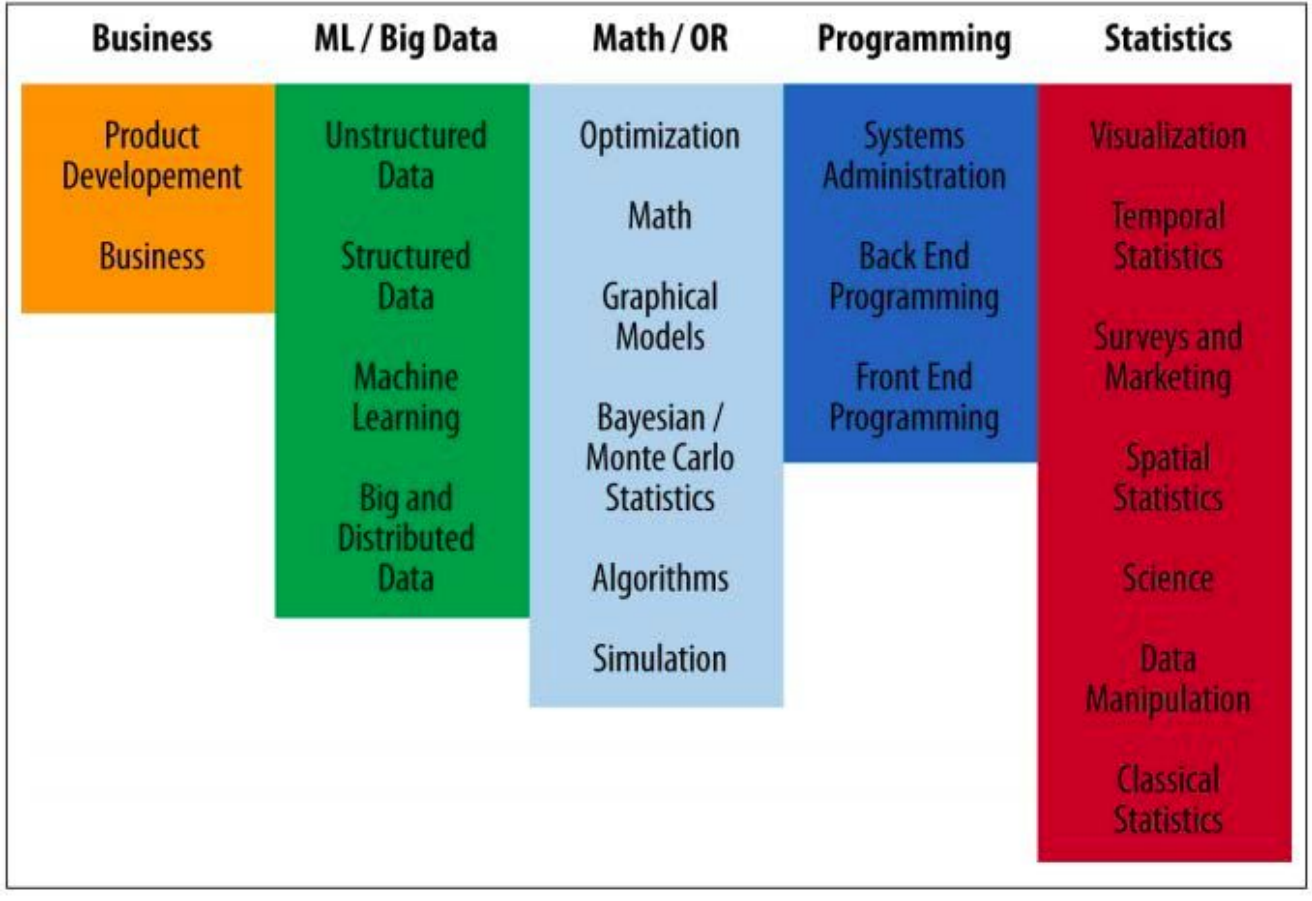



# ¿QUÉ ES DATA SCIENCE?

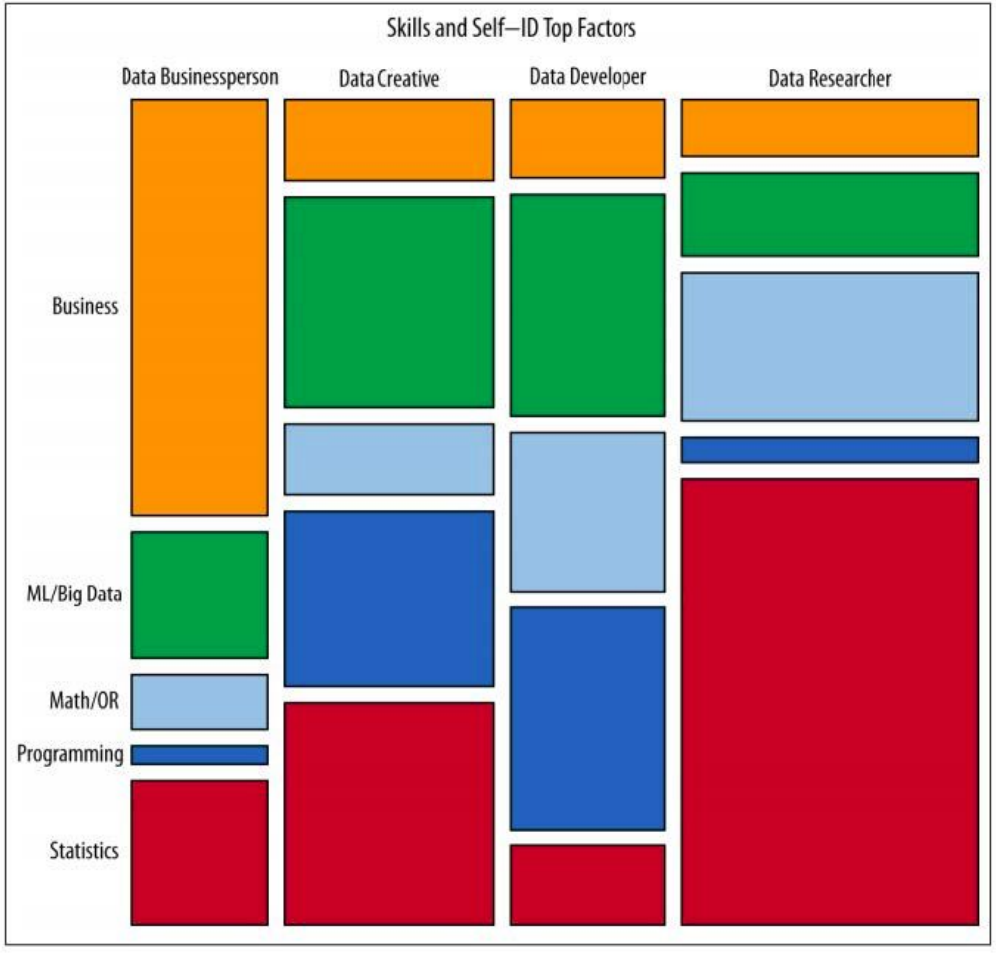


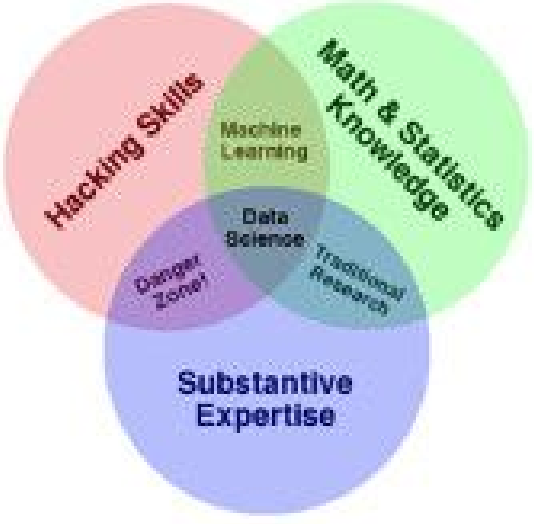

- Un set de herramientas y técnicas para extraer información útil de los datos
- Una práctica interdisciplinaria orientada a **resolver problemas**
- La aplicación de técnicas científicas a problemas prácticos
- ¿Quién usa Data Science?
  - Recomendaciones de películas Netflix
  - Algoritmo Amazon: “si 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

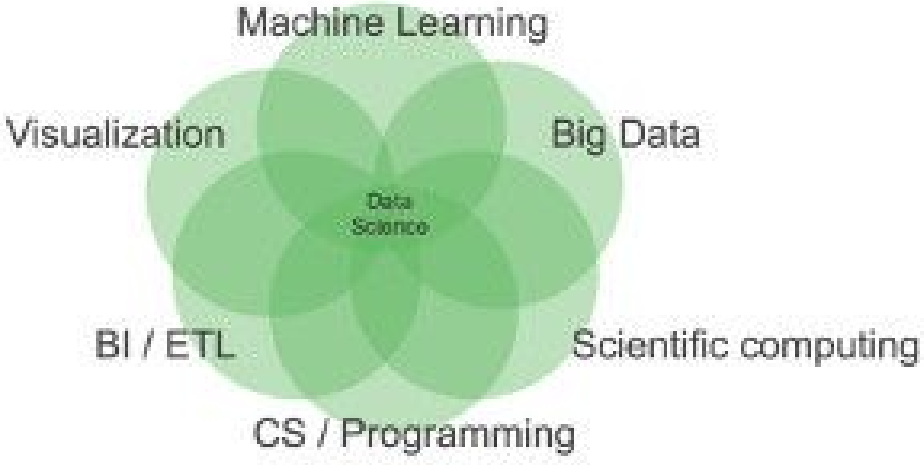

Business
ML / Big Data
Math / OR
Programming
Statistics
Product Developement
Business
Unstructured Data
Structured Data
Machine Learning
Big and Distributed Data
Optimization
Math
Graphical Models
Bayesian / Monte Carlo Statistics
Algorithms
Simulation
Systems Administration
Back End Programming
Front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

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

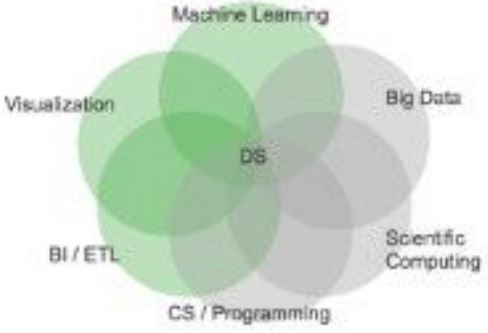




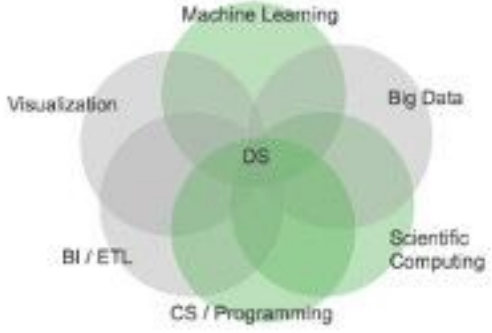

Traditional Data Science Venn Diagram



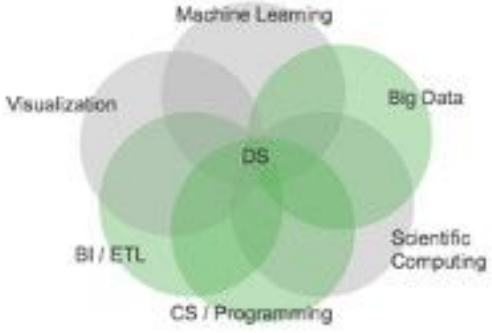

Revisited Data Science Venn Diagram




Statistician / Analyst



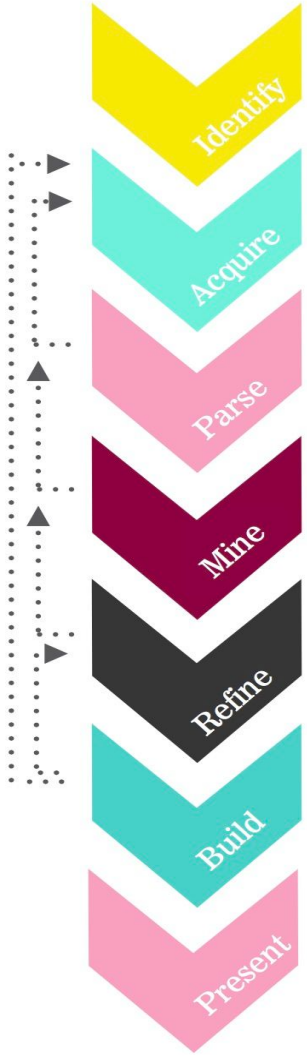

Research / Computational Scientist




Developer / Engineer

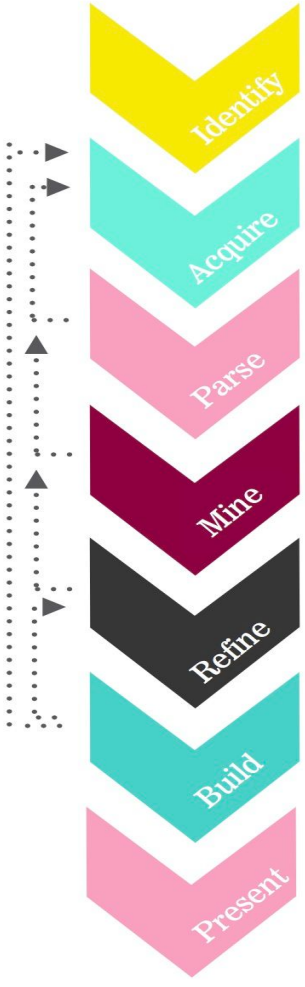
- A lo largo de las clases seguiremos el “Flujo de trabajo de Data Science”. Nos servirá para generar resultados confiables y reproducibles.
  - “confiables” = precisos
  - “reproducibles” = otros pueden replicar lo realizado y obtener resultados similares

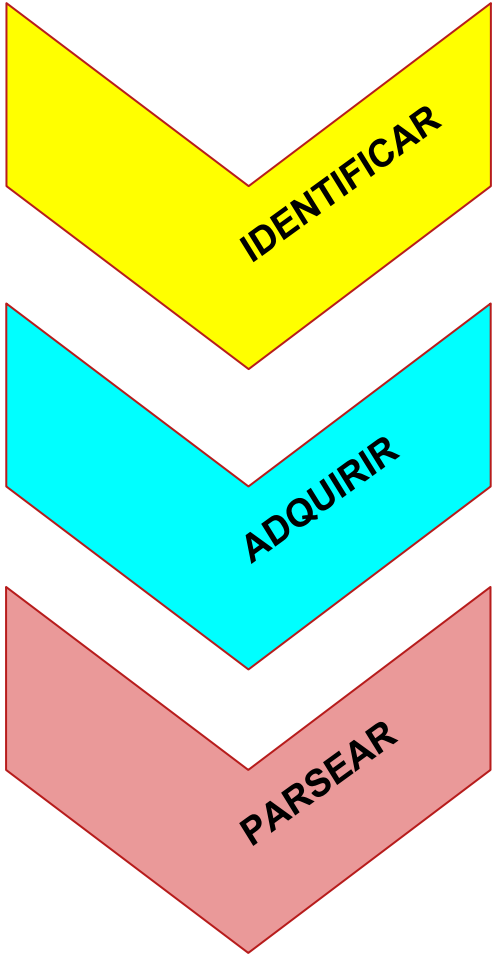
- En cualquier punto del proceso, puede ser necesario repetir pasos previos para iterar a lo largo del flujo. Esto dependerá de
  - la aparición de nuevos datos,
  - la necesidad de corregir errores,
  - el cambio acerca de las preguntas y objetivos, etc.



- 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**

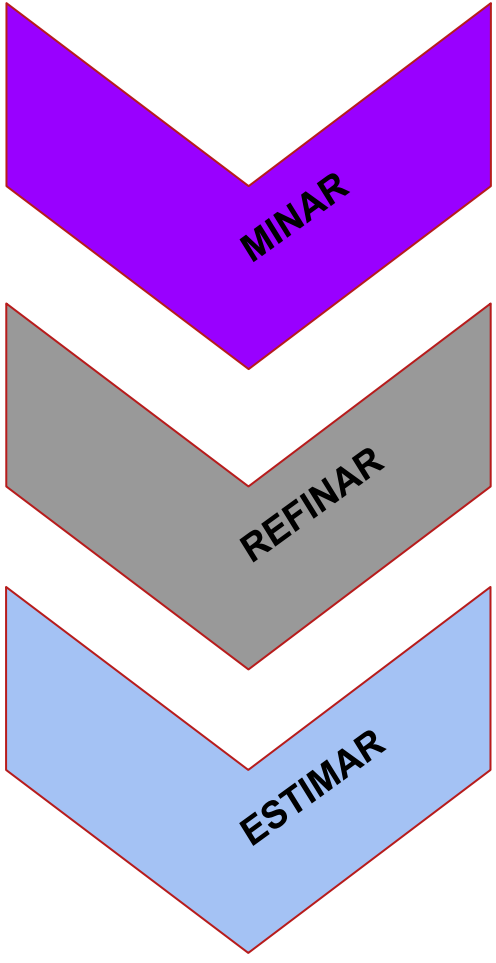
- Identificar los objetivos del producto/negocio/problema
- Identificar y generar hipótesis sobre metas y criterios para el éxito del análisis
- Generar un set de preguntas para identificar el dataset "correcto".

**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**

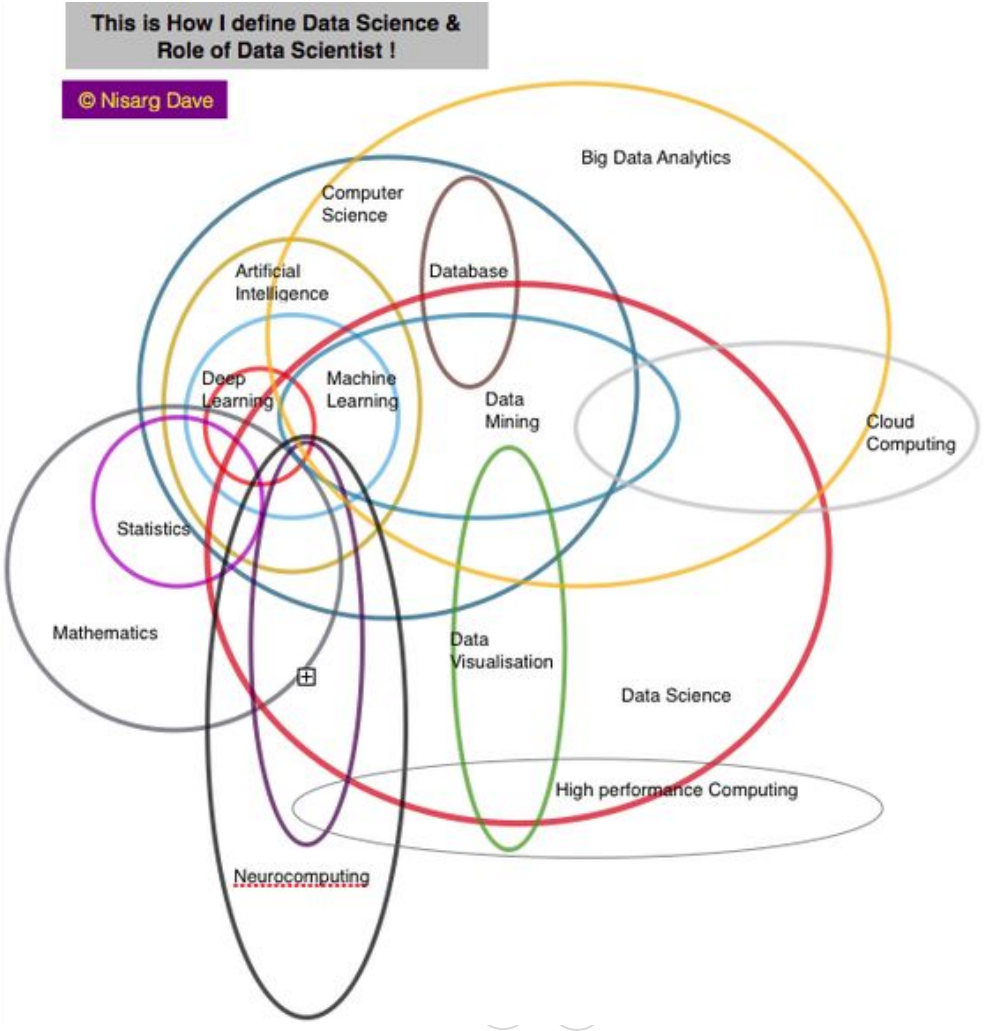
- Identificar tendencias y outliers
- Aplicar y calcular estadísticos descriptivos e inferenciales
- Documentar y transformar los datos

**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


This is How I define Data Science & Role of Data Scientist !
© Nisarg Dave
Big Data Analytics
Computer Science
Artificial Intelligence
Database
Deep Learning
Machine Learning
Data Mining
Cloud Computing
Statistics
Mathematics
Data Visualisation
Data Science
High performance Computing
Neurocomputing


FAIL


DigitalHouse > Coding School

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

DigitalHouse >
Coding School

# 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.

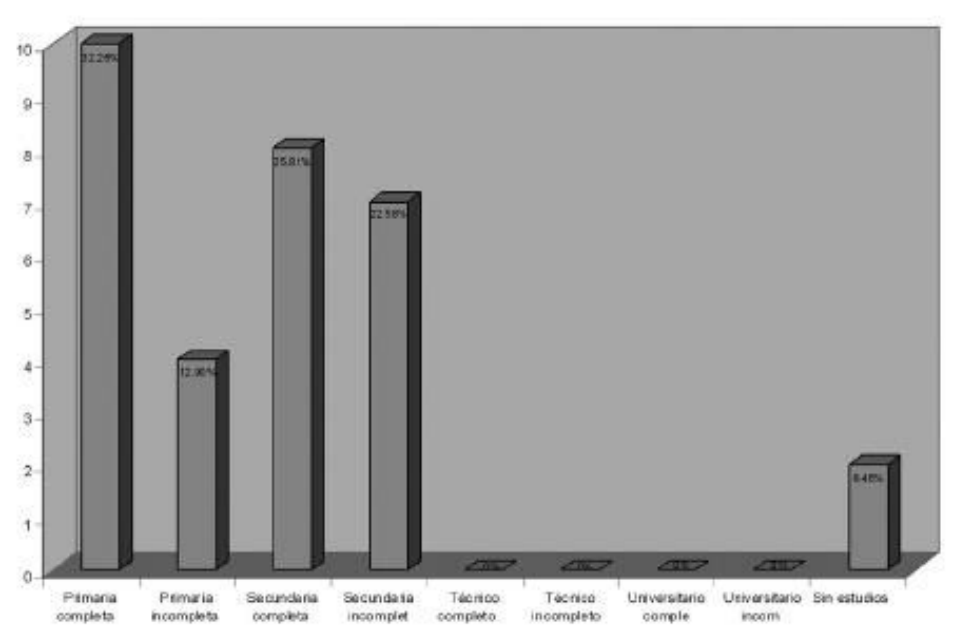
DigitalHouse >
Coding School

# 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



DigitalHouse >
Coding School

# 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 | 10 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.