Data Science

FOR

~~DUMMIES~~

A Wiley Brand

MANAGERS

(Y GRUPOS AUTO-ORGANIZADOS)

About me...

- Ingeniero en Sistemas de Información
- **Nerd...**
- Trabajo en Pi Data Strategy and Consulting
- Individuals and Interactions over almost everything...

@pdelboca

Agenda

- Objetivo
- Data Scientist y Data Science
- Casos de uso típicos
- Ciclo de vida de un Proyecto
  - Objetivo del Proyecto
  - Recolección de Datos
  - Construcción de un Modelo
  - Evaluación del Modelo
  - Presentación de Resultados
  - Implementación del Modelo
- Thumb Rules
- Conclusiones
- Preguntas

objetivo

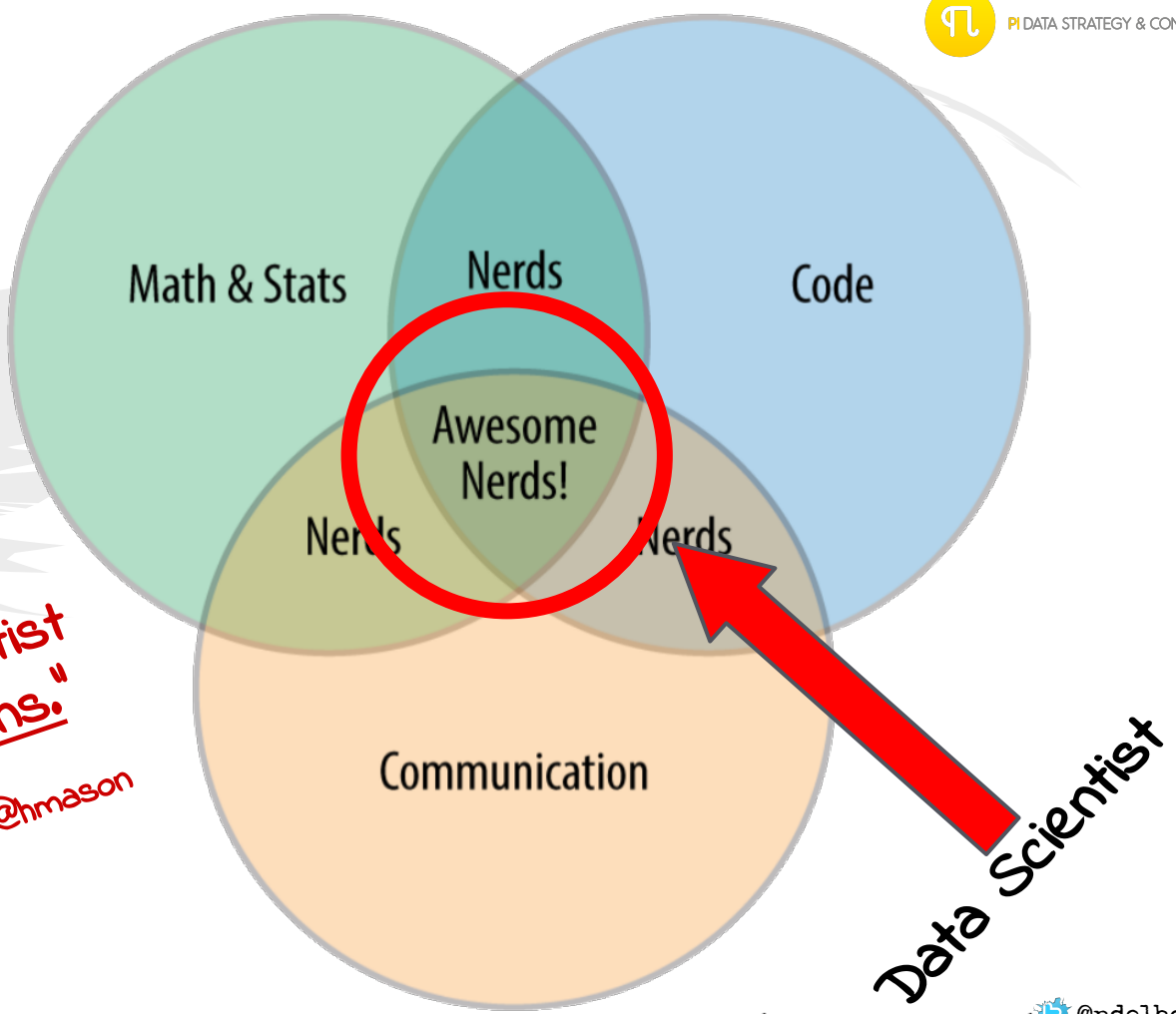
Entender los **conceptos fundamentales** y **características básicas** de un Proyecto de Data Science.

Data Science

"we define Data Science as managing the process that can transform hypothesis and data into actionable predictions."

-Practical Data Science with R-

Data Scientist



"The job of the Data Scientist is to ask the right questions."

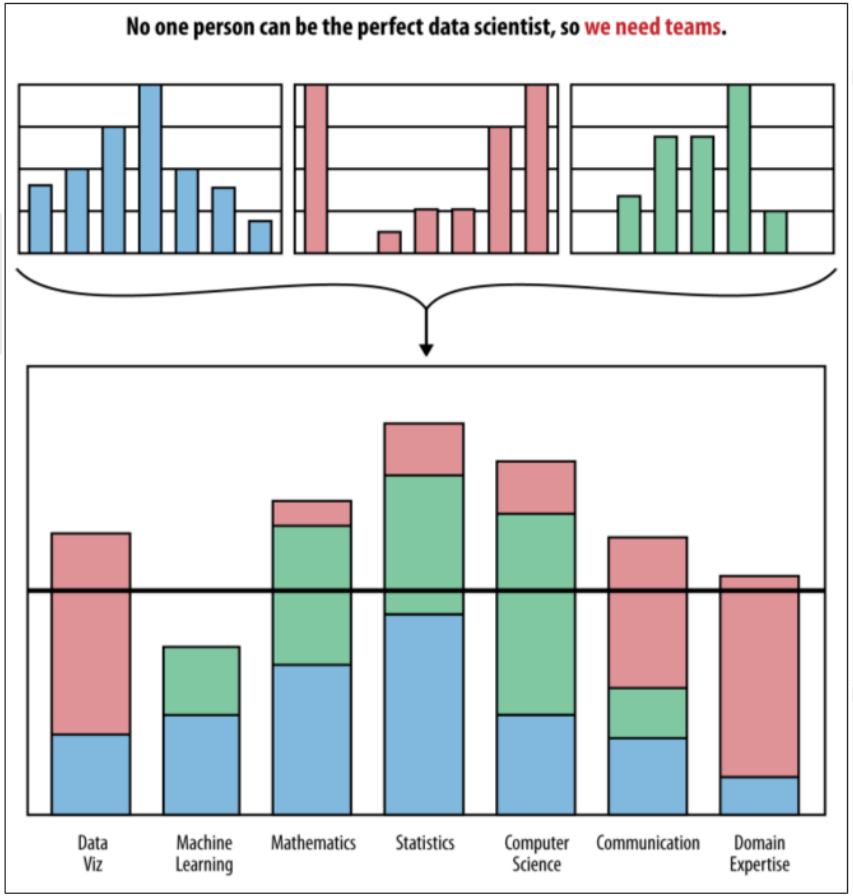
@hmason

Data Driven - Hilary Mason & Dj Patil

Data Scientist

(Equipos Interdisciplinarios)

"Unicorn Data Scientist vs Data Scientist Team"



-Doing Data Science, Rachel Schutt and Cathy o'Neil-

Casos de Uso más Frecuentes

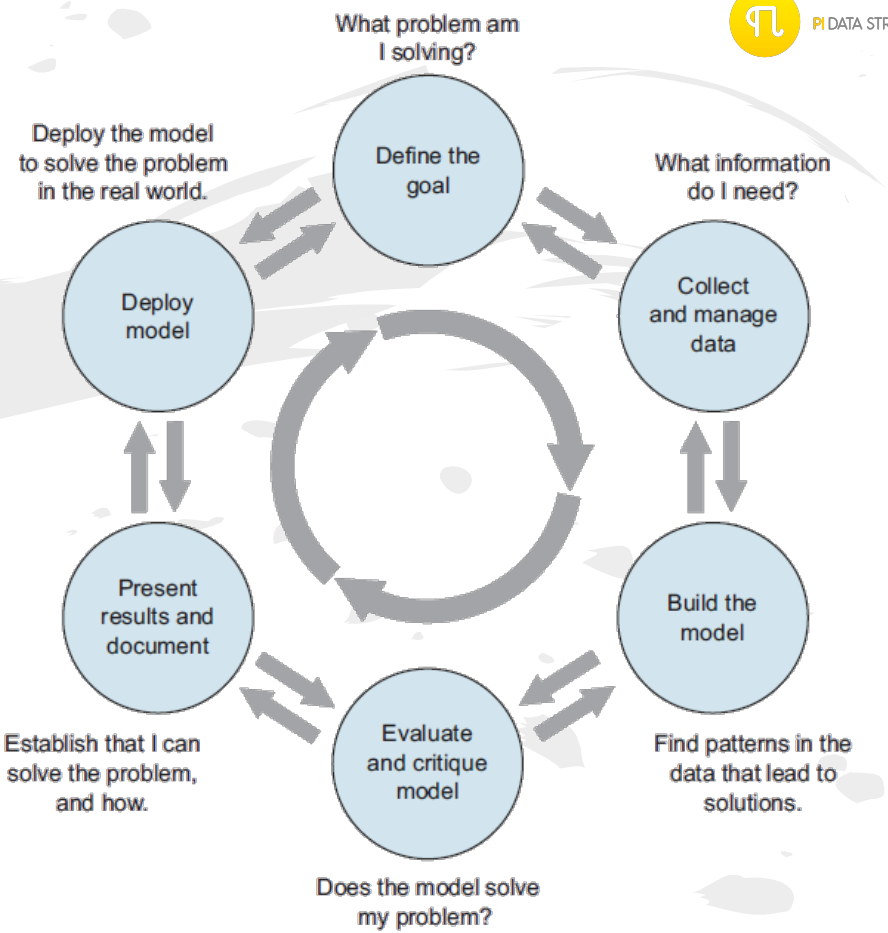
- Encontrar un modelo que nos permita explicar la relación entre un conjunto de variables de entrada y una variable objetivo ya conocida (supervisado) con el objetivo de predecir que valores tomará dicha variable objetivo en un futuro:
  - Clasificación
  - Regresión
  - Ranking
- Encontrar patrones y estructuras hasta ahora desconocidas (no supervisado) en un conjunto de datos que nos permitan explicar el comportamiento de los datos
  - Clustering
  - Pattern Finding
  - Characterization

Agenda

- Objetivo ✓
- Data Scientist y Data Science ✓
- Casos de uso típicos ✓
- Ciclo de vida de un Proyecto
  - Objetivo del Proyecto
  - Recolección de Datos
  - Construcción de un Modelo
  - Evaluación del Modelo
  - Presentación de Resultados
  - Implementación del Modelo
- Thumb Rules
- Conclusiones
- Preguntas

Ciclo de vida de un Proyecto

Individuals and Interactions
>
Processes and Tools



Practical Data Science with R - Nina Zumel and John Mount

Definición del objetivo

"Identify 90% of accounts that will go into default at least two months before the first missed payment with a false positive rate of no more than 25%..."

Definición del objetivo

(Tips & Tricks)

Customer Collaboration
>
Contract Negotiation

- Baselines!!

- Cómo mediría el éxito el cliente?

- **C**ontext, **N**eed, **V**ision, **O**utcome
  -Thinking with data - Max Shron-

Definición del objetivo

(Vocabulario)

" In common usage, prediction means to forecast a future event. In Data Science, prediction more generally means to **estimate an unknown value.** "

-Data Science for Business - Foster Provost & Tom Fawcett-

Definición del objetivo

(Dude's Law)

"At the end of the day, it is usually **how we frame the problem**, not the **tools and techniques** that we use to answer it, that determine how **valuable** our work is."

-Thinking with data - Max Shron-

Obtención y Gestión de Datos

(Obtención)

- Primera barrera a la hora de encarar un proyecto.
- Insume mucho tiempo
- Fomentar grupos y actividades de openData

@opendatacba

Construcción del Modelo

(E.D.A)

- Búsqueda de variables con mayor "poder predictivo"
- Apoyo en herramientas de BI e infraestructura actual
- Mucho ida y vuelta entre el Data Scientist y el Negocio

Construcción del Modelo (Modelado)

"Essentially, all models are wrong, but some are useful."

- George E. P. Box -

- Representación simplificada de la realidad
- Generalización que busca comprender y explicar el comportamiento de los datos

Construcción del Modelo

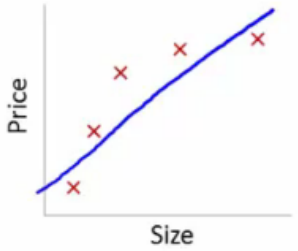
(Entrenamiento)

Evaluación del Modelo

- Resuelve el problema...?
- Diferentes técnicas:
  - Precision & Recall,
  - Matriz de Confusión
  - Error Cuadrático
  - PRESS,
  - etc
- Dependen de:
  - Tipo de Problema
  - Conjunto de Datos
- Error Analysis

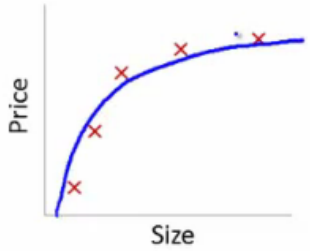
Evaluación del Modelo

(overfitting)



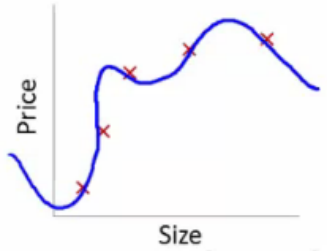
$\theta_0 + \theta_1 x$

High bias
(underfit)



$\theta_0 + \theta_1 x + \theta_2 x^2$

"Just right"



$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

High variance
(overfit)

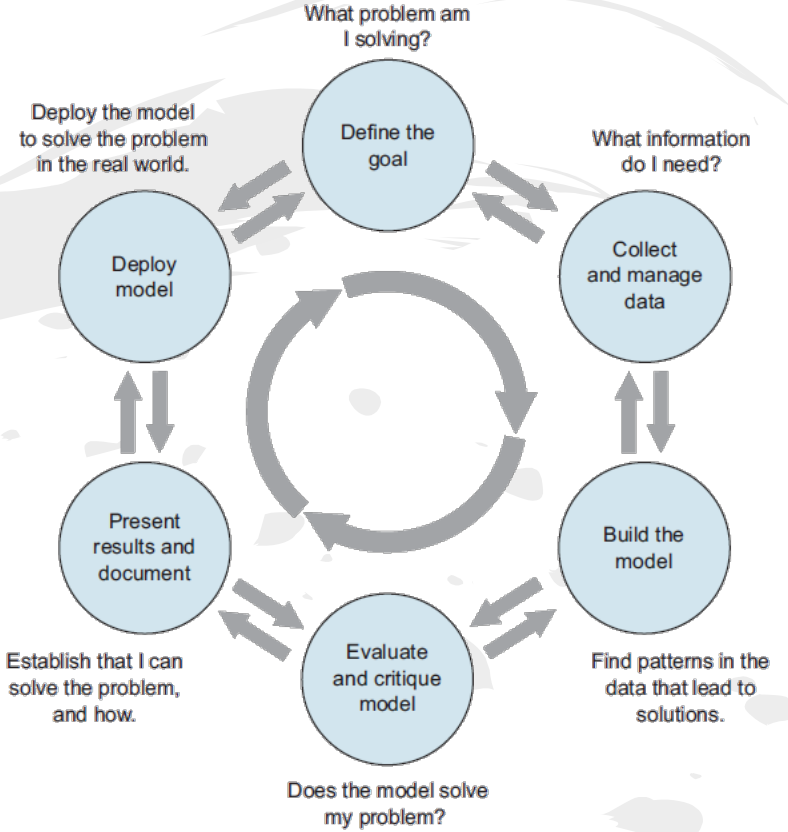
Presentación del Modelo

- Reunión con los principales stakeholders
  - Usuarios finales
  - Sponsors
  - Gente de IT
- Tres audiencias, tres lenguajes

Implementación del Modelo

- Tarea no tan trivial.
- Data Scientist <-> Software Engineer
- Cloud, web-service, batch?
- Integración a otros sistemas.
- Cambios en el día a día del usuario

Mejora Continua



Practical Data Science with R - Nina Zumel and John Mount

Agenda

- Objetivo ✓
- Data Scientist y Data Science ✓
- Casos de uso típicos ✓
- Ciclo de vida de un Proyecto ✓
  - Objetivo del Proyecto ✓
  - Recolección de Datos ✓
  - Construcción de un Modelo ✓
  - Evaluación del Modelo ✓
  - Presentación de Resultados ✓
  - Implementación del Modelo ✓
- Thumb Rules
- Conclusiones
- Preguntas

Thumb Rules

- ¿Tienen datos etiquetados? ¿Cuántos? (1000 poco, +10 millones requiere consideración especial)
- Qué tipos de datos? Numérico, texto, imágenes?
- Garbage in, garbage out.
- Una prueba de concepto previa puede ayudar a ajustar muchísimo las estimaciones y las expectativas del cliente.
- ¿Está el cliente comprometido con un trabajo de Calidad?

Conclusiones

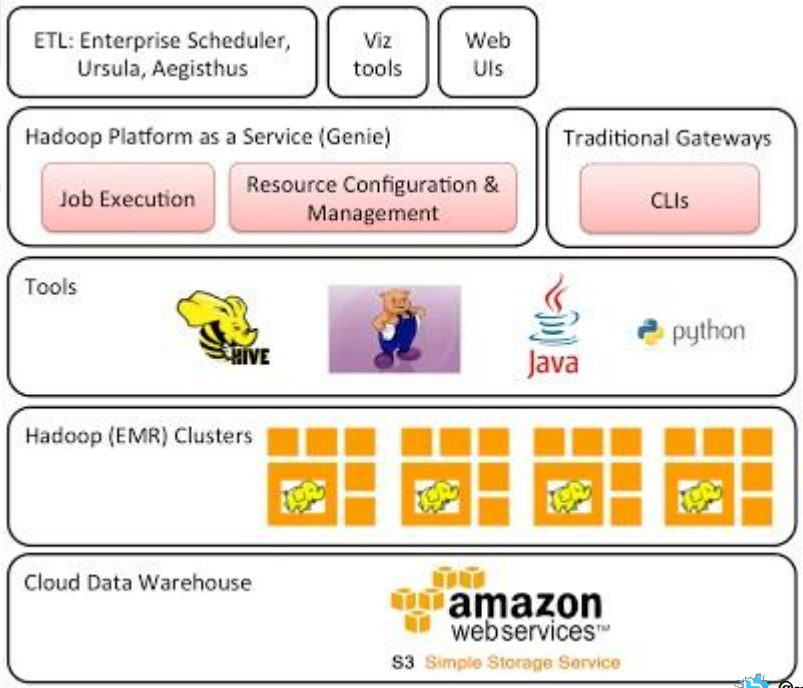
- Data Science es un proceso Iterativo que requiere de muchas idas y vueltas entre cada etapa

- Lo más importante en un Proyecto es la pregunta que se hace para abordar una problemática concreta, más aún que las herramientas que se usen.

- Un proyecto de Data Science involucra un equipo multidisciplinario, con todas las ventajas/desventajas que eso conlleva.

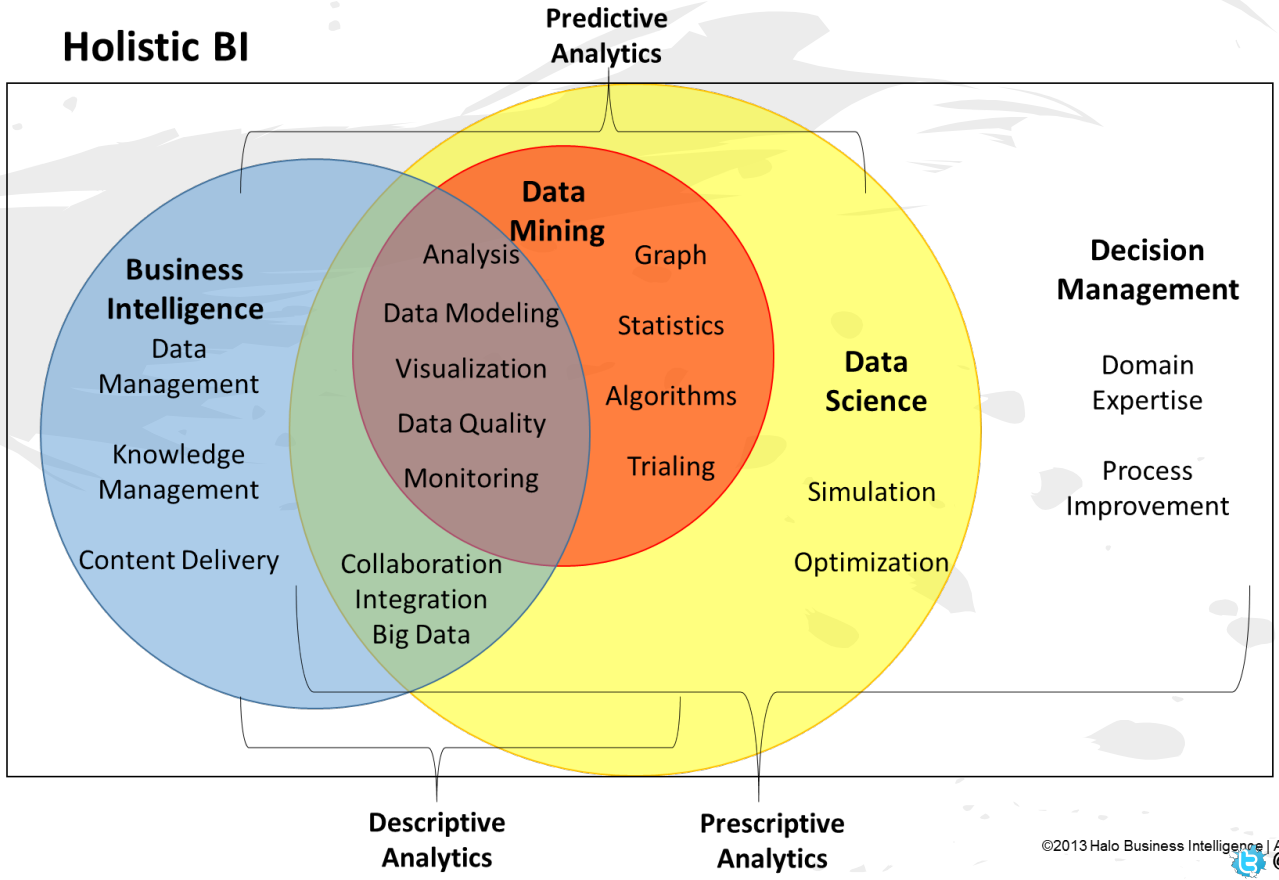
Big Data..?

- "Big" is when you can't fit it on one machine. -



BI..?

Holistic BI



@pdelboca

Preguntas...?

PI DATA STRATEGY & CONSULTING

Muchas Gracias!


@pdelboca


pdelboca@piconsulting.com.ar


patriciodelboca@gmail.com


@ds_cba