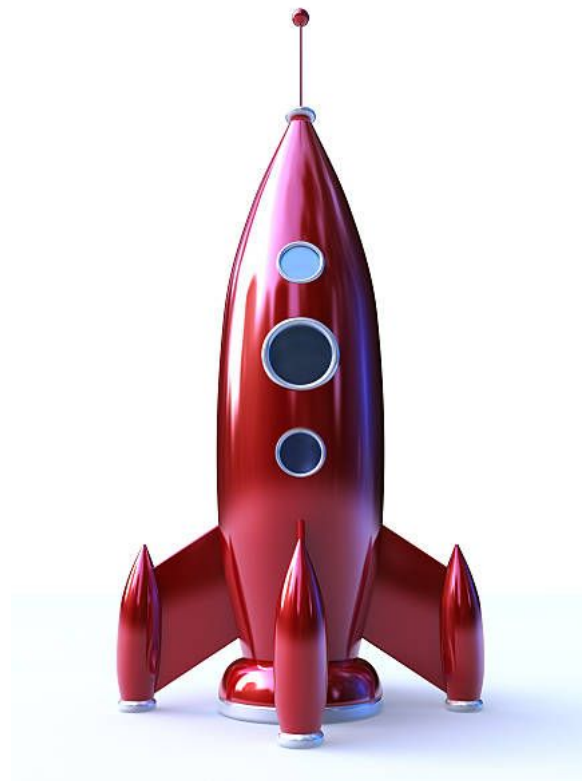


Data Science

Sergio Santoyo
Daniel Fernández

Índice

1. Información del Taller
2. ¿Qué es data science?
3. ¿Por qué el físico es un buen data scientist?
4. Ejemplos en la vida real
5. Áreas de data science
6. ¿Por qué ahora? “Big Data”
7. Procesos de data science
8. Herramientas
9. Experiencias
10. Futuras aplicaciones
11. Temario talleres
12. Notebook 0: “Hola mundo”



Taller:

- Duración: 6 sesiones
 - Sesión 1: Marzo 7
 - Sesión 2: Marzo 14
 - Sesión 3: Marzo 21
 - Sesión 4: Abril 4
 - Sesión 5: Abril 11
 - Sesión 6: Abril 21
- Autores: Sergio Santoyo, Daniel Fernández
- Salón: LPB 12
- Horario: 17 a 19
- Github del taller: <https://github.com/ldfo/ds-uia-2018>
- Canal de Slack: ds-uia.slack.com

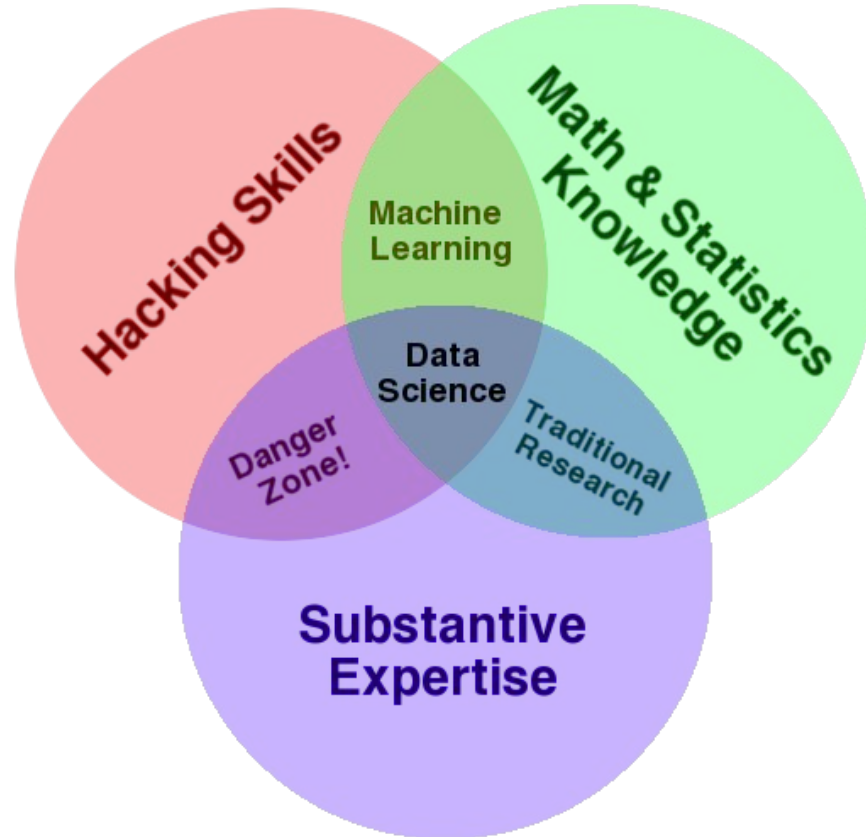
¿Qué es Data Science?

- Ayudar la toma de decisiones por medio de cómputo científico, estadística, probabilidad y visualizaciones.

Wikipedia:

- Also known as data-driven science, is an interdisciplinary field of scientific methods, processes, and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining.
- **Data mining** is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. It is an essential process where intelligent methods are applied to extract data patterns. It is an interdisciplinary subfield of computer science.

¿Por qué el físico es un buen data scientist?



Vida real

The image shows two side-by-side screenshots of web pages. The left screenshot is from Kaggle's Datasets page, displaying a list of datasets sorted by 'Hotness'. The right screenshot is from LinkedIn's Jobs page, showing search results for 'data scientist' positions.

Kaggle Datasets Page:

- Search bar: "Search kaggle"
- Navigation: Competitions, Datasets, Kernels, Discussion, Jobs, Sign In
- Section: Datasets (Learn More, New Dataset)
- Filters: Public, Your Datasets, Favorites
- Sort by: Hotness
- Results: 10,585 Datasets
- Dataset List:

Rank	Dataset Name	Description	Category	Format	Size	Views	Downloads
30	Yelp Dataset	A trove of reviews, businesses, users, tips, and check-in data! Yelp, Inc. updated 13 days ago	food and drink	CSV	5 GB	4	2
32	GitHub Repos	Code and comments from 2.8 million repos Github updated 2 months ago	programming lang...	BigQuery	3 TB	10	0
19	Hacker News	All posts from Y Combinator's social news website from 2006 to late 2017 Hacker News updated 2 months ago	journalism information techn...	BigQuery	14 GB	4	0
113	Chocolate Bar Ratings	Expert ratings of over 1,700 chocolate bars Rachael Tatman updated 6 months ago	critical theory food and drink	CSV	125 KB	60	1
37	Historical Air Quality	Air Quality Data Collected at Outdoor Monitors Across the US US Environmental Protection Agency updated 2 months ago	pollution bigquery	BigQuery	323 GB	2	1
183	H-1B Visa Petitions 2011-2016	3 million petitions for H-1B visas Sharan Naribole updated a year ago	law international relati...	CSV	409 MB	144	11

LinkedIn Jobs Page:

- Search bar: "data scientist"
- Location: "Todo el mundo"
- Section: Empleos (Fecha de publicación (1))
- Results: 30.618 resultados
- Job Listings:

Job Title	Company	Location	Details
(Erfahrener) Data Analyst/ Data Scientist (m/w) für Big Data und Data Analytics	Daimler Protics GmbH	Saarbrücken, Saarland, Deutschland	Als Data Analyst/ Data Scientist für Big Data und Data Analytics beraten Sie unsere Mitarbeiter in Projekten bezüglich des Einsatzes sowie der Integration verschiedener ...
Data Scientist (Data Analyst)	Serioplast	Bergamo Area, Italy	The Data Scientist reports to the Software Development Manager. Excellent knowledge of a programming language for data analysis (Python is the language of choice) Develop the ...
Associate Data Scientist	Prudential Corporation Asia	Hong Kong	BL business analysis, data engineering, ETL, analytics or visualisation experts without actual machine learning experience. Deep learning experience using one of the most ...
Mira empleos donde figuras entre los mejores solicitantes	Liberty Mutual Insurance	Seattle, WA, US	Imagination Data Science (IDS) is seeking a Senior Data Scientist at the Assistant Director or Director I level to join the Model Design I team and help move Business Insurance ...

The average salary for Data Scientists is \$100K.

A Data Scientist makes an average of \$100K, ranging from \$70K to \$132K based on 16K profiles. These numbers represent our estimate for potential total compensation, including Base Salary, approximate Equity, and an Annual Bonus as an aggregate of all Data Scientist salaries. Explore relevant salary ranges, open jobs, associated skills, demographics insights, and more.



How competitive are Data Scientist salaries?
The average market salary for employees is \$99.6K per year, ranging from \$69.6K to \$132K.

Last Updated on: January 26th, 2018

Job Category Trends

Mathematics

Job Category Trends January 2018

Job postings

Average job postings this month:

9,018

Change in job market share in percentage points

Compared to last month:

▲ 0.01

Compared to last year:

▲ 0.01

Job clicks over time

A line chart showing the number of job clicks over time. The y-axis ranges from 0 to 2,000,000. The x-axis shows months from Jan 2017 to Jan 2018. The line starts at approximately 1,400,000 in Jan 2017, fluctuates, and ends at approximately 1,700,000 in Jan 2018.

Top job titles per click

Data Scientist	785,886
Senior Analyst	284,806
Entry Level Analyst	189,782
Management Analyst	158,092
Junior Analyst	72,732
Associate Analyst	63,456
Financial Modeler	47,478
Biostatistician	44,642
Statistician	40,743
Data Evaluator	12,640

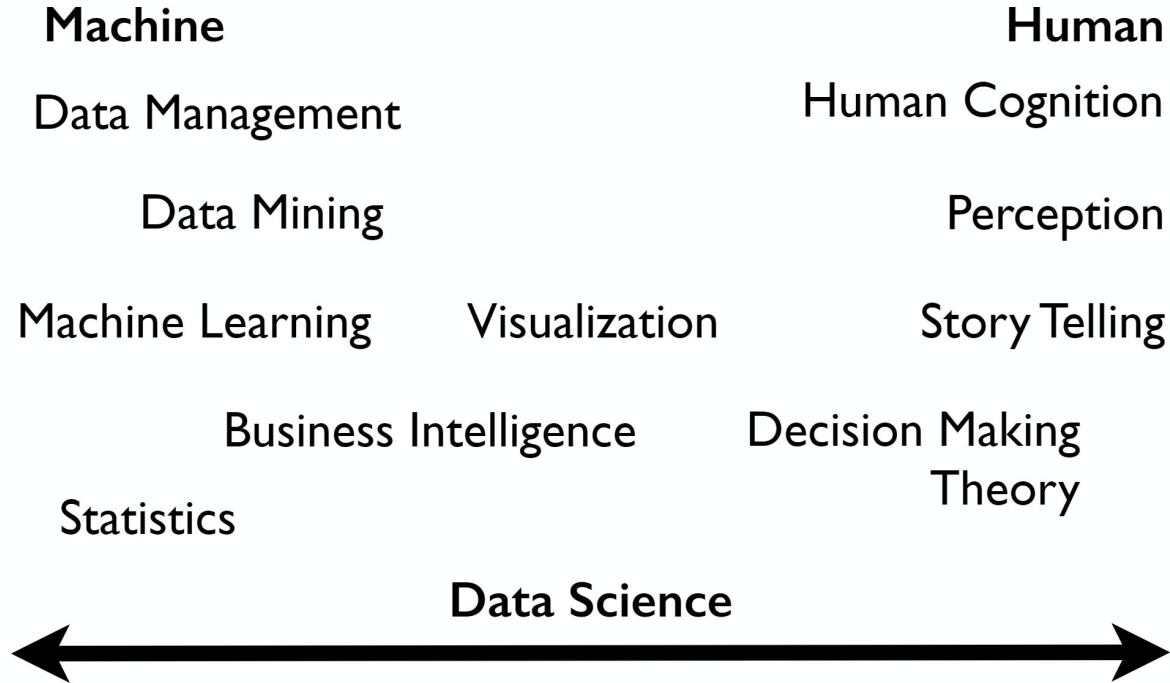
Top search terms per click

data scientist	160,425
data analyst	64,352
analyst	27,766
data science	18,330
business analyst	18,061
machine learning	17,221
statistics	14,084
statistician	13,887
sas	12,209
biostatistician	11,314

Top job locations per click

New York, NY	115,614
Washington, DC	45,641
San Francisco, CA	41,457
Chicago, IL	38,106
Dallas, TX	26,510
Atlanta, GA	24,957
Houston, TX	22,791
Boston, MA	21,943
Fort Meade, MD	21,013
Seattle, WA	18,976

Áreas de Data Science

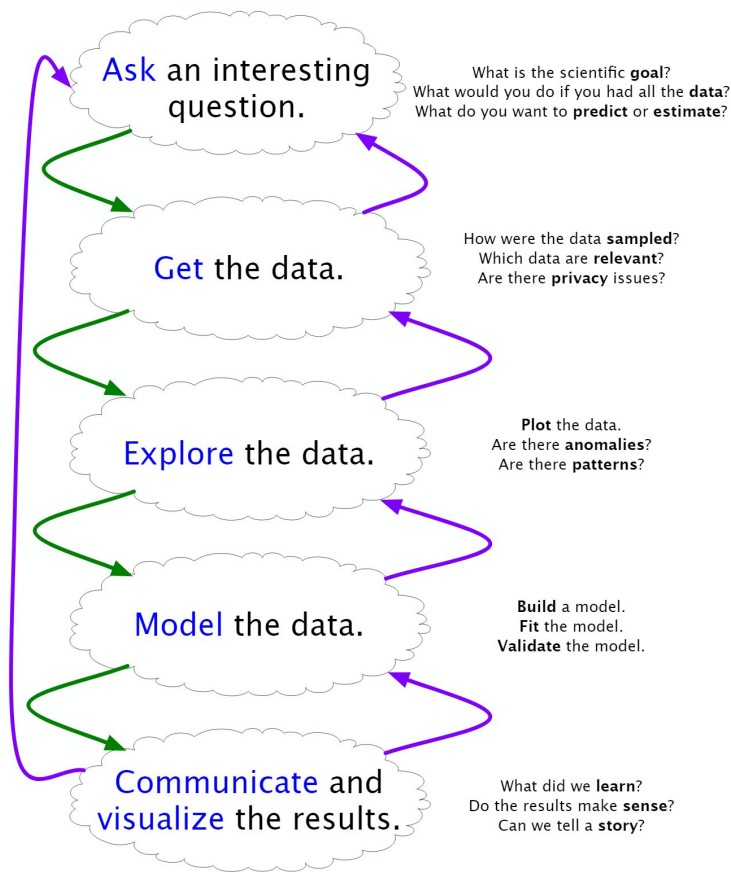


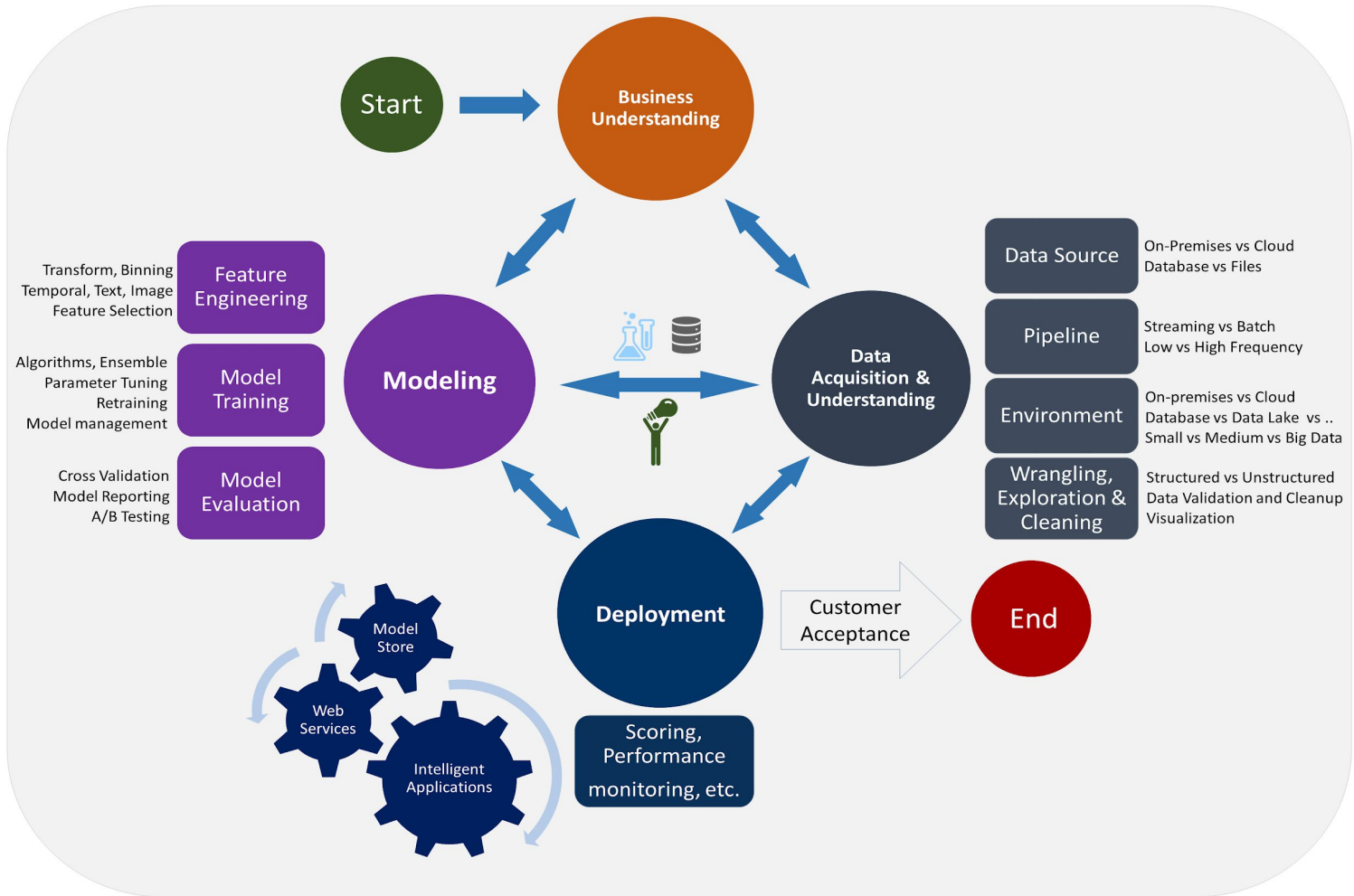
¿Por qué ahora? “Big Data”

- Para el 2020 se estima una generación de datos diaria de $35(10)^{12}$ GB.
- Gran variedad de datos (redes sociales, transacciones, videos, texto).
- 90% de los datos existentes hoy, fueron generados en los últimos 5 años.
- El contenido es generado por bots y por humanos: ~400 millones de tweets diarios.
- Más de 72 horas de video subidas a YouTube cada minuto.
- Aproximadamente 1.4 billones de usuarios diarios de Facebook.
- Internet of things.
- Industry 4.0 / 5.0

BIG DATA = volumen + velocidad + variedad


Procesos de data science





Herramientas

- El 52.6% del desarrollo se hace en Python.
- El 34.9% usan SQL para bases de datos.
- El 52.1% usan R.
- La herramienta depende del objetivo: producto, análisis, reporte, dashboard, etc.
- “No te cases con una tecnología”



GitHub ✓

Code Collaboration & Version Control

See GitHub alternatives

Favorites
★
977

Stacks
10.9K

[I Use This](#)

Fans 7.79K Votes 9.8K Jobs 2.17K



Bitbucket ✓

Code Collaboration & Version Control


See Bitbucket alternatives

Favorites
★
267

Stacks
4.32K

[I Use This](#)

Fans 3.23K Votes 2.72K Jobs 198



GitLab ✓

Code Collaboration & Version Control

See GitLab alternatives

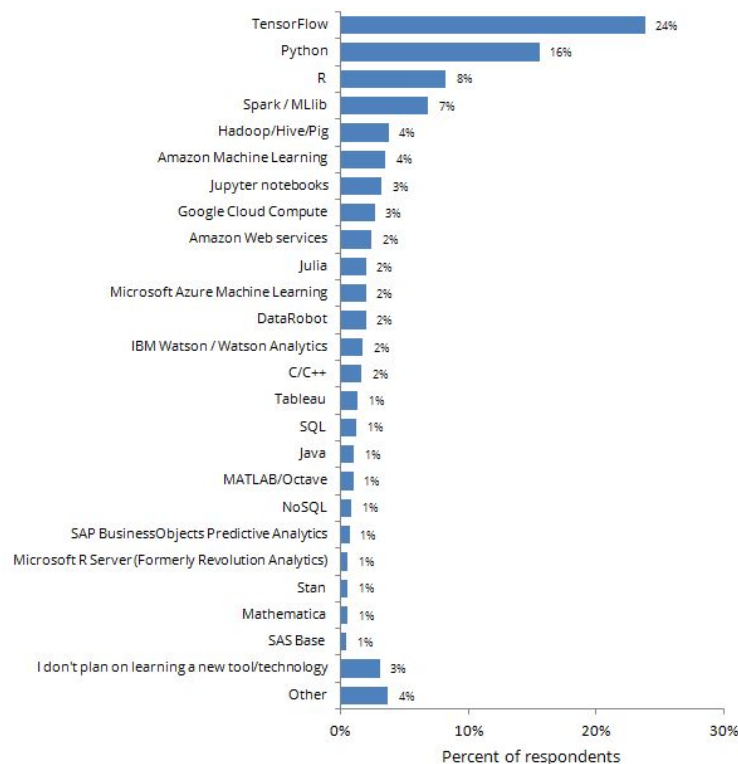
Favorites
★
267

Stacks
2.55K

[I Use This](#)

Fans 2.04K Votes 1.8K Jobs 163

Data Science Tool or Technology Data Pros are Most Excited about Learning in 2018



Data are from the Kaggle 2017 The State of Data Science and Machine Learning study. You can learn more about the study and download the data here: <https://www.kaggle.com/surveys/2017>. Respondents were asked to indicate which tool or technology they are most excited about learning in the next year. A total of 10998 respondents answered the question. Tools and technologies that were selected by less than 1% of the survey respondents are not included in the graph.



Google Cloud Platform



Source Data

Store Data

Convert & ETL

Transform Data

Exploratory Analysis

Model Build &
Generate Insights

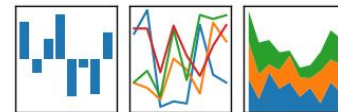
Visualisation

Model Execution in
Production



pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$

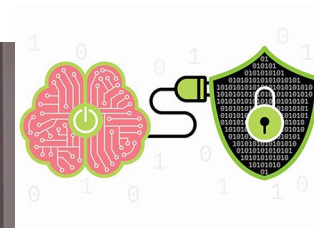
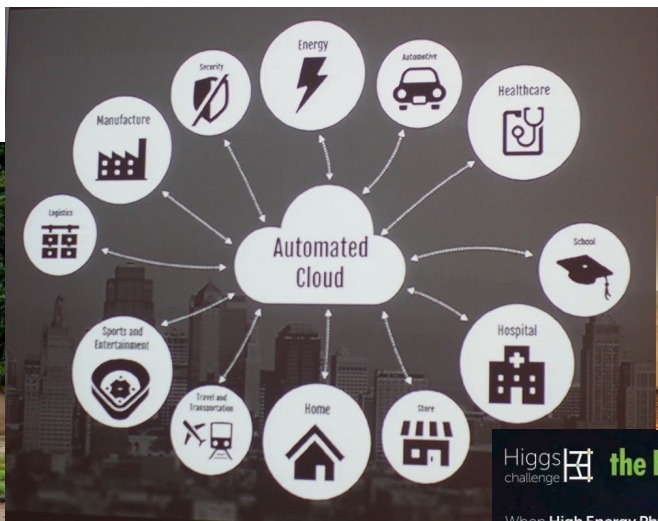



Notebook



Experiencias

- Sergio:
 - **Intelimétrica:** Real estate, desarrollar modelos, pipelines de procesamiento y manutención de bases de datos. Desarrollo de proyectos para clientes. Detección de anomalías, modelos de valuación automática, pipelines de procesamiento, proyecto de índice de valor patrimonial.
 - **Abraxas Intelligence:** Consultoría a grupo Bimbo, análisis de series de tiempo de accidentes laborales, análisis de texto de reportes de accidentes laborales, generación de reportes.
 - **Sinnia:** Redes sociales, inferencia de edades de usuarios de Twitter, clasificación de textos de noticias, análisis de polaridad política para candidatos a la presidencia, clustering de series de tiempo de trending topics para Twitter. Deployment y manutención de modelos en la nube.
- Daniel:








Higgs challenge  the HiggsML challenge

May to September 2014

When High Energy Physics meets Machine Learning

info to participate and compete : <https://www.kaggle.com/c/higgs-boson>

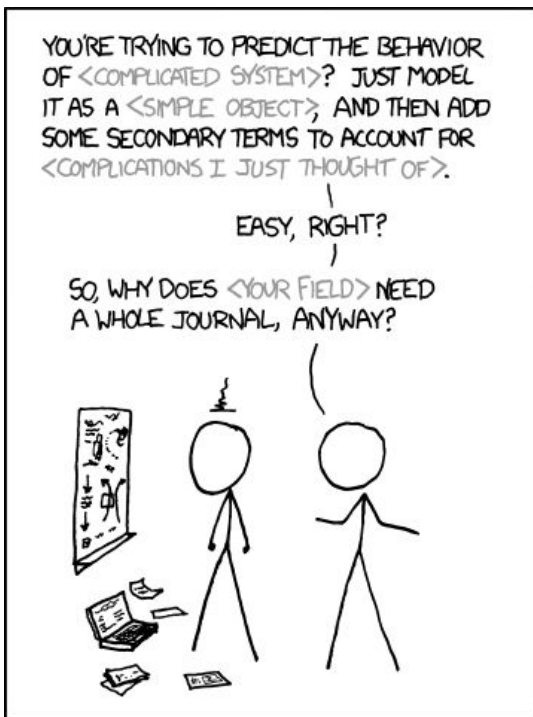
Organizing committee:
 Julien Hag, Agnèsall,
 Jack Garret, - RD-01
 David Rousseau, - RD-01
 Gisele Couder, - RD-01
 Sylvain Geyl, - RD-01
 Gisele Huet, - RD-01
 Thomas Bagnier, - RD-01
 Anne Huet, - RD-01
 Jean Sghier, - RD-01
 Ben Schwan, - RD-01

Military committee

Temario:

1. Introducción al taller
2. Manipulación y visualización de datos
3. Regresión lineal
4. Problemas de clasificación
5. Aprendizaje supervisado
6. Aprendizaje no supervisado y técnicas avanzadas

Notebook 0: “Hola Mundo”



LIBERAL-ARTS MAJORS MAY BE ANNOYING SOMETIMES, BUT THERE'S *NOTHING* MORE OBNOXIOUS THAN A PHYSICIST FIRST ENCOUNTERING A NEW SUBJECT.

