Introducción al análisis de datos

Semana - 1

¿Qué es?

- Data Science
- Data Mining
- Big Data
- Business Intelligence

¿Tienen algo en común?

Definición de análisis avanzado de datos

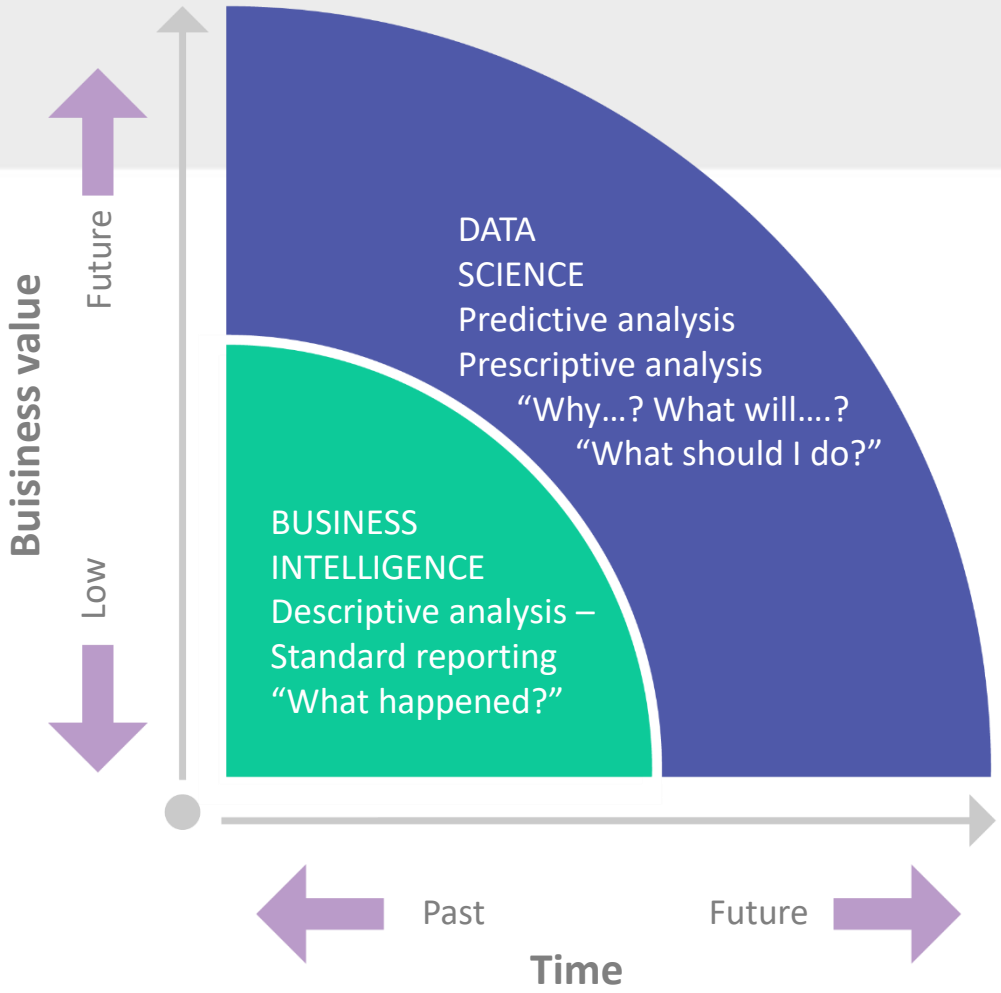
Advanced Analytics is the autonomous or semi-autonomous examination of data or content using sophisticated techniques and tools, typically beyond those of traditional business intelligence (BI), to discover deeper insights, make predictions, or generate recommendations.

- Gartner Group

Foto tomada de: https://www.istockphoto.com/photo/top-view-of-office-desk-with-data-chart-and-pie-chart-3d-rendering-gm925236050-253906983?irgwc=1&esource=AFF_IS_IR_SP_FreeImages_246195&asid=FreeImages&cid=IS&utm_medium=affiliate_SP&utm_source=FreeImages&utm_content=246195&clickid=SZazepUIPxyJR9RwUx0Mo3YyUklQ3PX-dV3NSU0

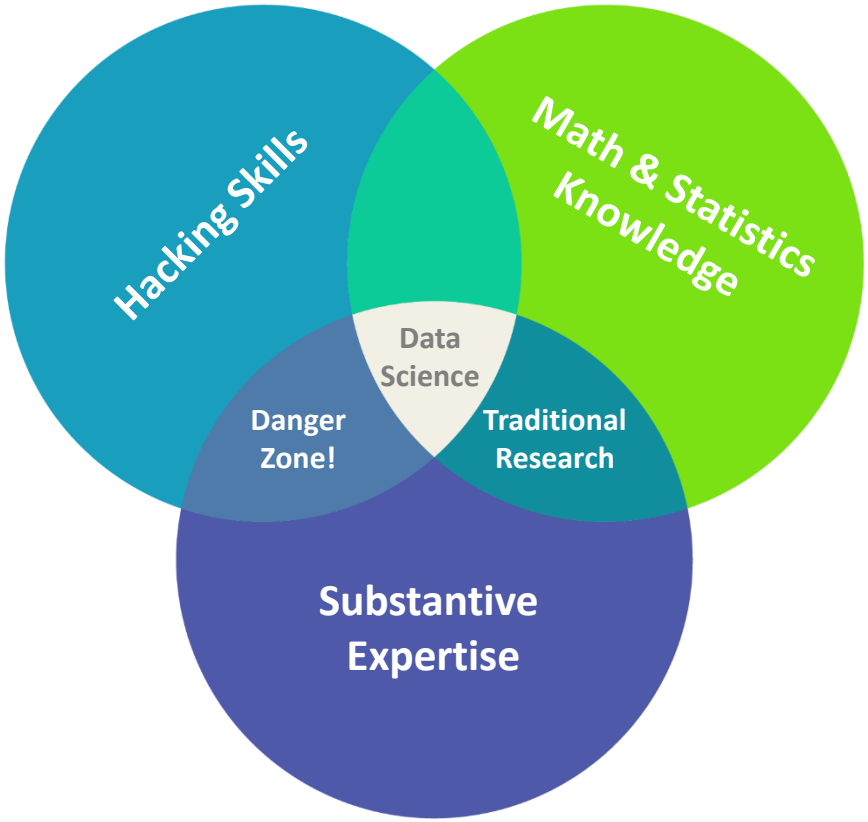
BI vs Data Science

- La principal diferencia está en las preguntas que responde.
- BI responde a la pregunta de ¿qué pasó?
- Data Science responde a las preguntas
  - ¿Por qué?
  - ¿Qué pasará?
  - ¿Qué debo hacer?

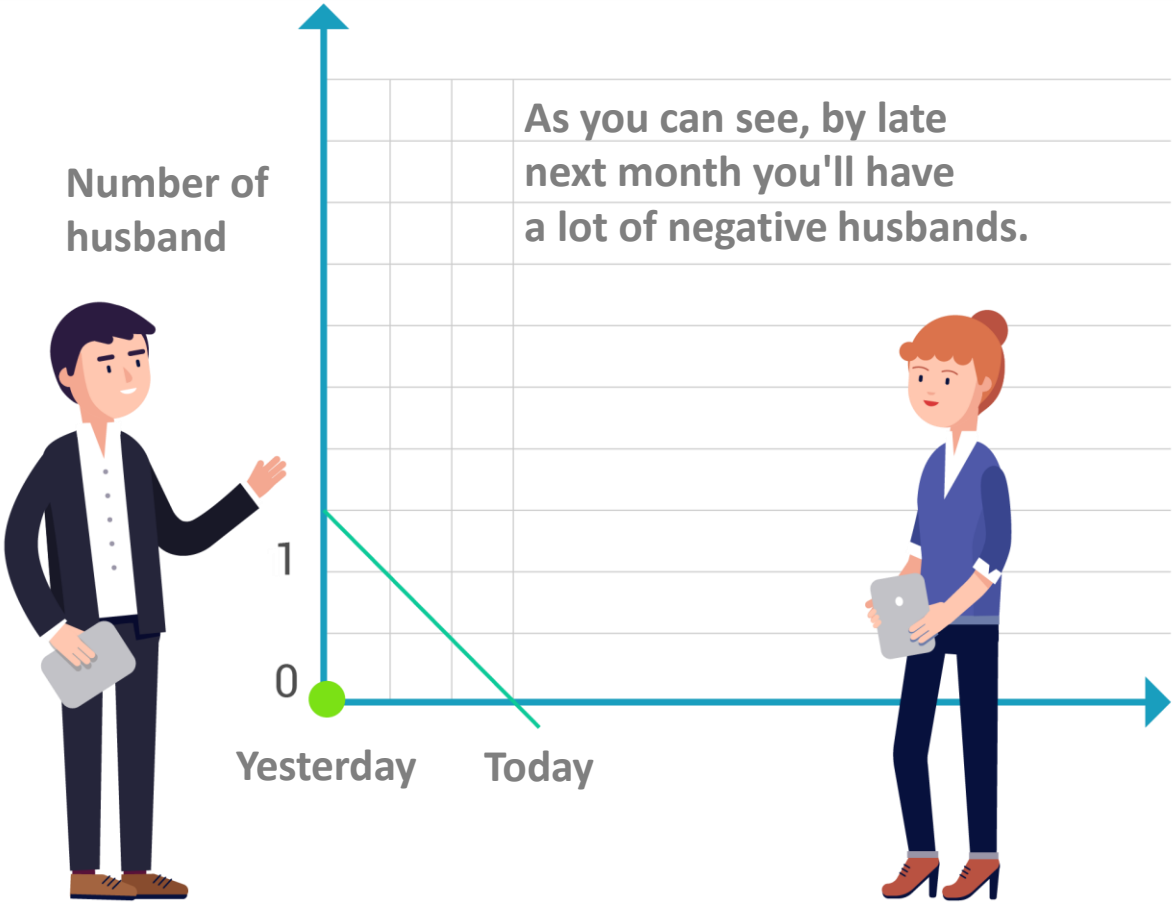


Data Science, Diagrama de Venn

- Data science se compone de:
  - Hacking Skills – Habilidades de cómputo e informática
  - Math & Statistics – Habilidad con los números
  - Sbustantive Expertise – Dominio o expertise del tema
- ¿Por qué Danger Zone?

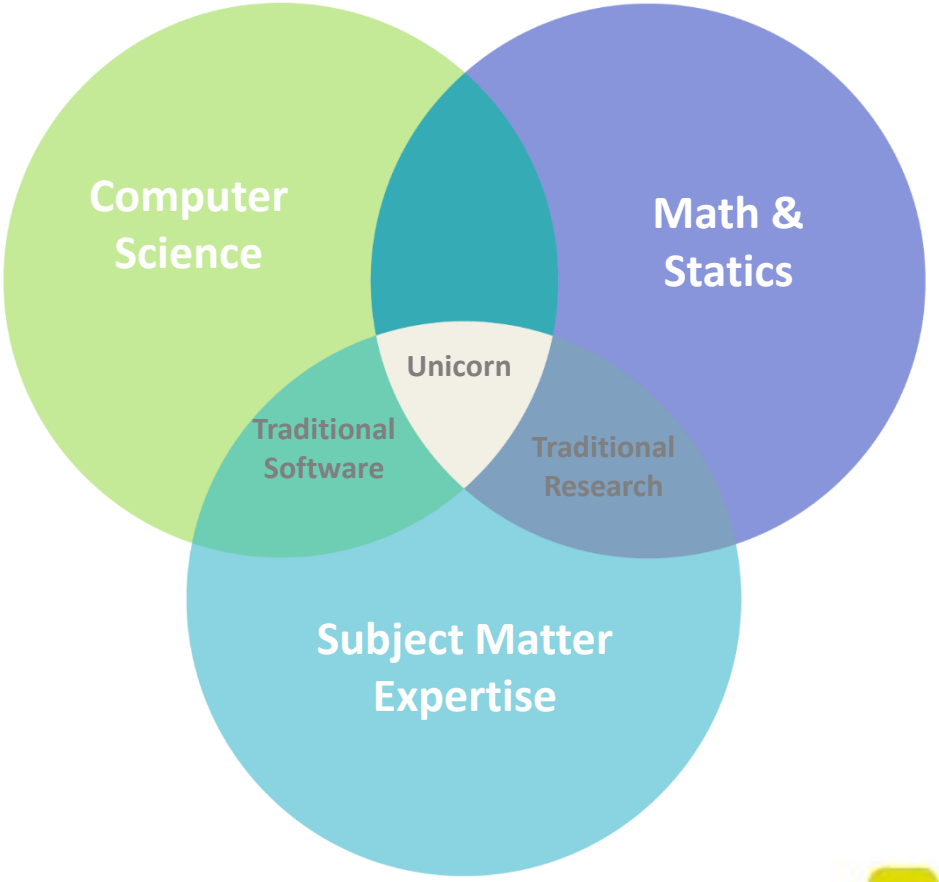


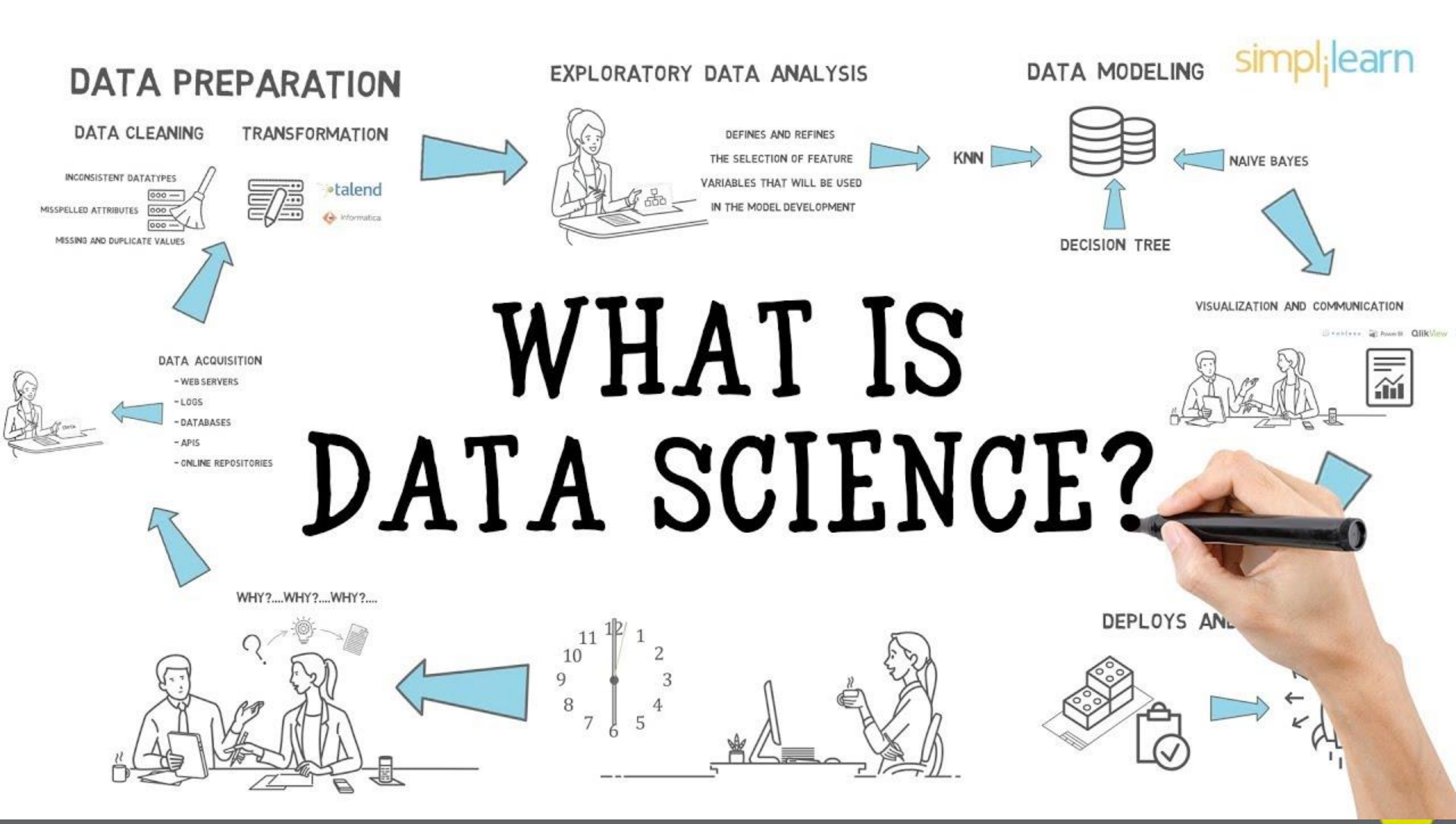
My Hobby: Extrapolating



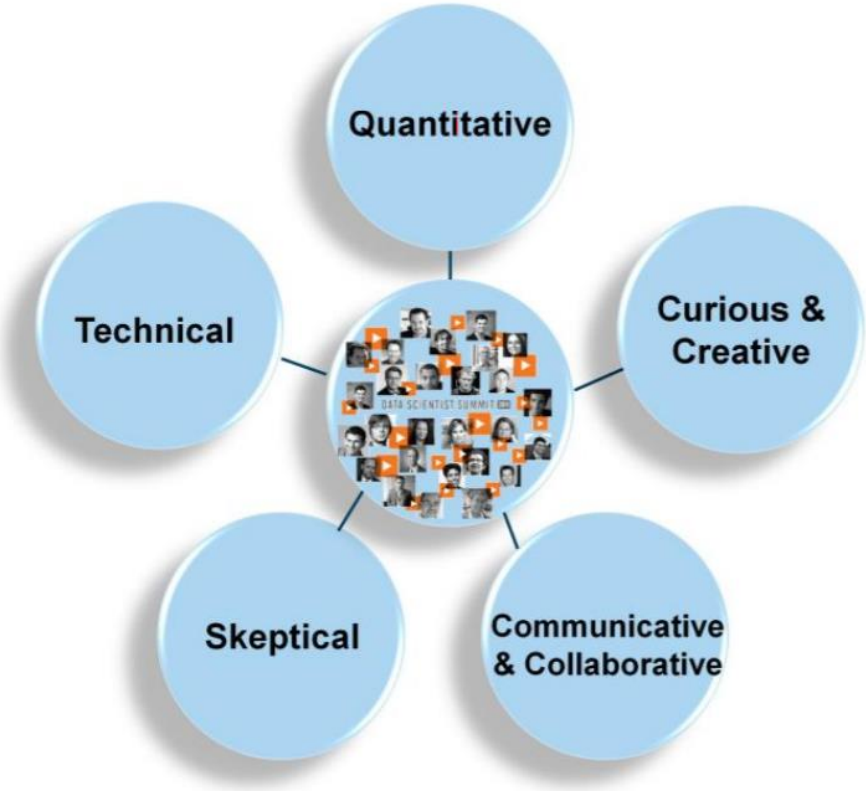
Data Science, Diagrama real

- Seamos honestos… solo los unicornios tienen las tres cosas juntas.
- Por lo cual hay que concentrarse en los dos conjuntos superiores, el inferior llega solo a través de la experiencia y el tiempo (si llega…)




WHAT IS DATA SCIENCE?
DATA PREPARATION
DATA CLEANING
TRANSFORMATION
INCONSISTENT DATATYPES
MISSPELLED ATTRIBUTES
MISSING AND DUPLICATE VALUES
talend
Informatica
EXPLORATORY DATA ANALYSIS
DEFINES AND REFINES
THE SELECTION OF FEATURE
VARIABLES THAT WILL BE USED
IN THE MODEL DEVELOPMENT
DATA MODELING
simplilearn
KNN
NAIVE BAYES
DECISION TREE
VISUALIZATION AND COMMUNICATION
DATA ACQUISITION
- WEB SERVERS
- LOGS
- DATABASES
- APIS
- ONLINE REPOSITORIES
WHY?....WHY?....WHY?....
DEPLOYS AN

¿Quién es el Data Scientist?



¿Quién es el Data Scientist?

Data Scientists

Projected U.S. talent gap: 140,000 to 190,000

Analysts & Data Savvy Managers

Projected U.S. talent gap: 1.5 million

Role	Role Description
Deep Analytical Talent	People with advanced training in quantitative disciplines, such as mathematics, statistics, and machine learning.
Data Savvy Professionals	People with a basic knowledge of statistics and/or machine learning, who can define key questions that can be answered using advanced analytics
Technology & Data Enablers	People providing technical expertise to support analytical projects. Skills sets including computer programming and database administration

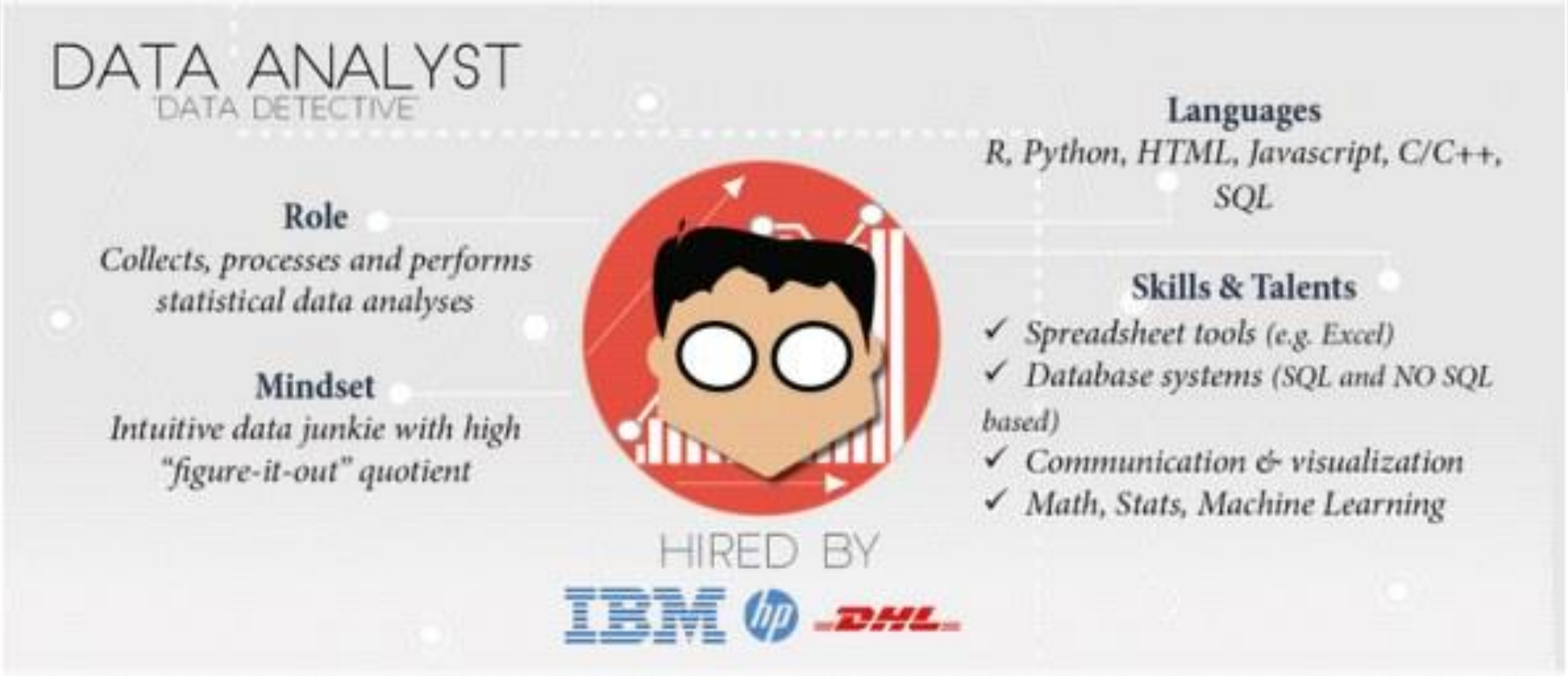
El mercado también cambia

- "… the sexy job in the next 10 years will be statisticians," Hal Varian, Google Chief Economist
- The U.S. will need 140,000-190,000 predictive analysts and 1.5 million managers/analysts by 2018. McKinsey Global Institute's June 2011
- New Data Science institutes being created or repurposed – NYU, Columbia, Washington, UCB,...
- New degree programs, courses, boot-camps:
  - e.g., at Berkeley: Stats, I-School, CS, Astronomy…
  - Maestrías en "Big Data" y "Data Science"

Diferentes Roles

https://www.kdnuggets.com/2015/11/different-data-science-roles-industry.html

Diferentes Roles



https://www.kdnuggets.com/2015/11/different-data-science-roles-industry.html

Diferentes Roles

https://www.kdnuggets.com/2015/11/different-data-science-roles-industry.html

Diferentes Roles

https://www.kdnuggets.com/2015/11/different-data-science-roles-industry.html

Diferentes Roles

DATABASE ADMINISTRATOR

'DATABASE CARETAKER'

Role

Ensures that the database is available to all relevant users, is performing properly and is being kept safe

Mindset

Master of Disaster Prevention

HIRED BY

tableau reddit

Languages

SQL, Java, Ruby on Rails, XML, C#, Python

Skills & Talents

- ✓ *Backup & recovery*
- ✓ *Data modeling and design*
- ✓ *Distributed Computing (Hadoop)*
- ✓ *Database systems (SQL and NO SQL based)*
- ✓ *Data security*
- ✓ *ERP & business knowledge*

https://www.kdnuggets.com/2015/11/different-data-science-roles-industry.html

Diferentes Roles

DATABASE ADMINISTRATOR

'DATABASE CARETAKER'

Role

Ensures that the database is available to all relevant users, is performing properly and is being kept safe

Mindset

Master of Disaster Prevention

HIRED BY

tableau reddit

Languages

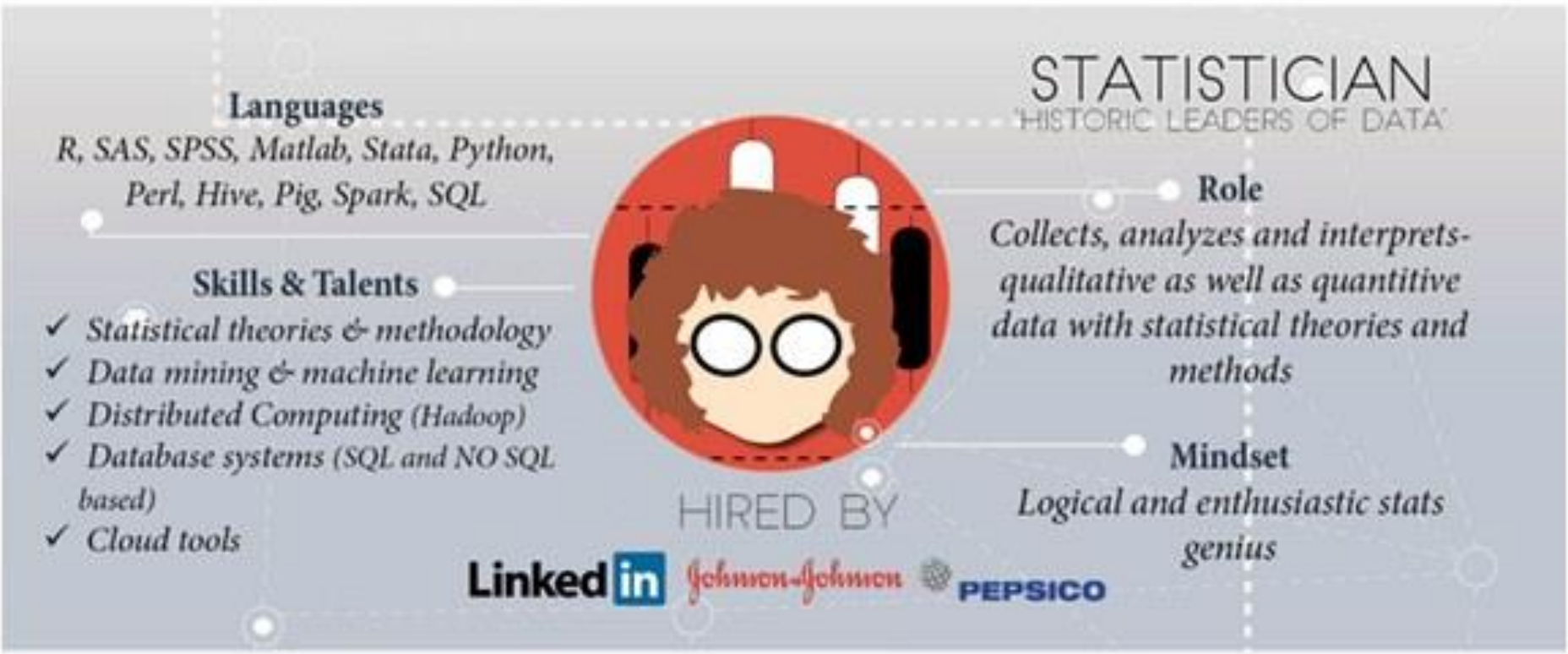
SQL, Java, Ruby on Rails, XML, C#, Python

Skills & Talents

- ✓ *Backup & recovery*
- ✓ *Data modeling and design*
- ✓ *Distributed Computing (Hadoop)*
- ✓ *Database systems (SQL and NO SQL based)*
- ✓ *Data security*
- ✓ *ERP & business knowledge*

https://www.kdnuggets.com/2015/11/different-data-science-roles-industry.html

Diferentes Roles



https://www.kdnuggets.com/2015/11/different-data-science-roles-industry.html

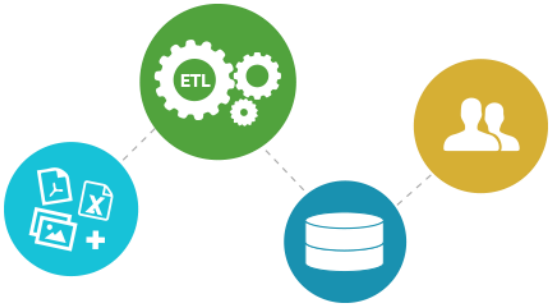
Diferentes Roles

https://www.kdnuggets.com/2015/11/different-data-science-roles-industry.html

¿Qué aprenderemos?

Associate- Data Science
Version 1.0

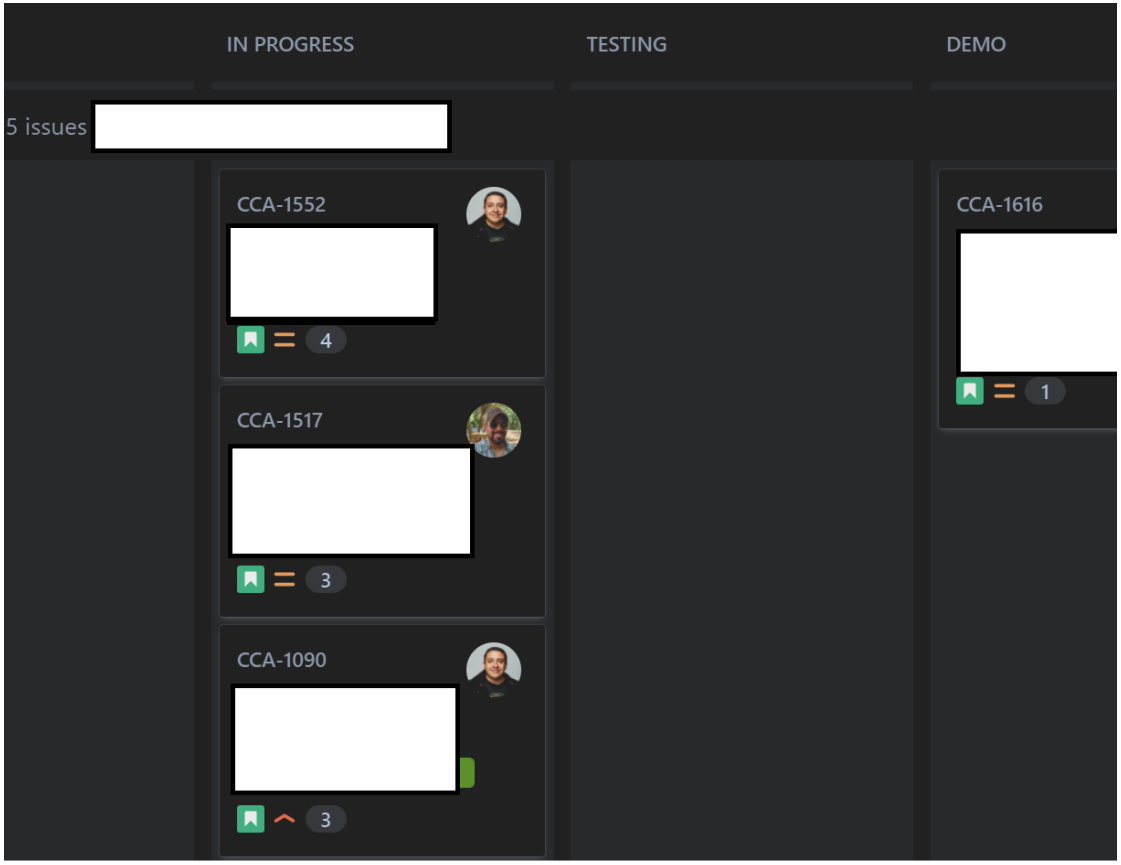
Dell EMC



Ciclo de vida del análisis de datos

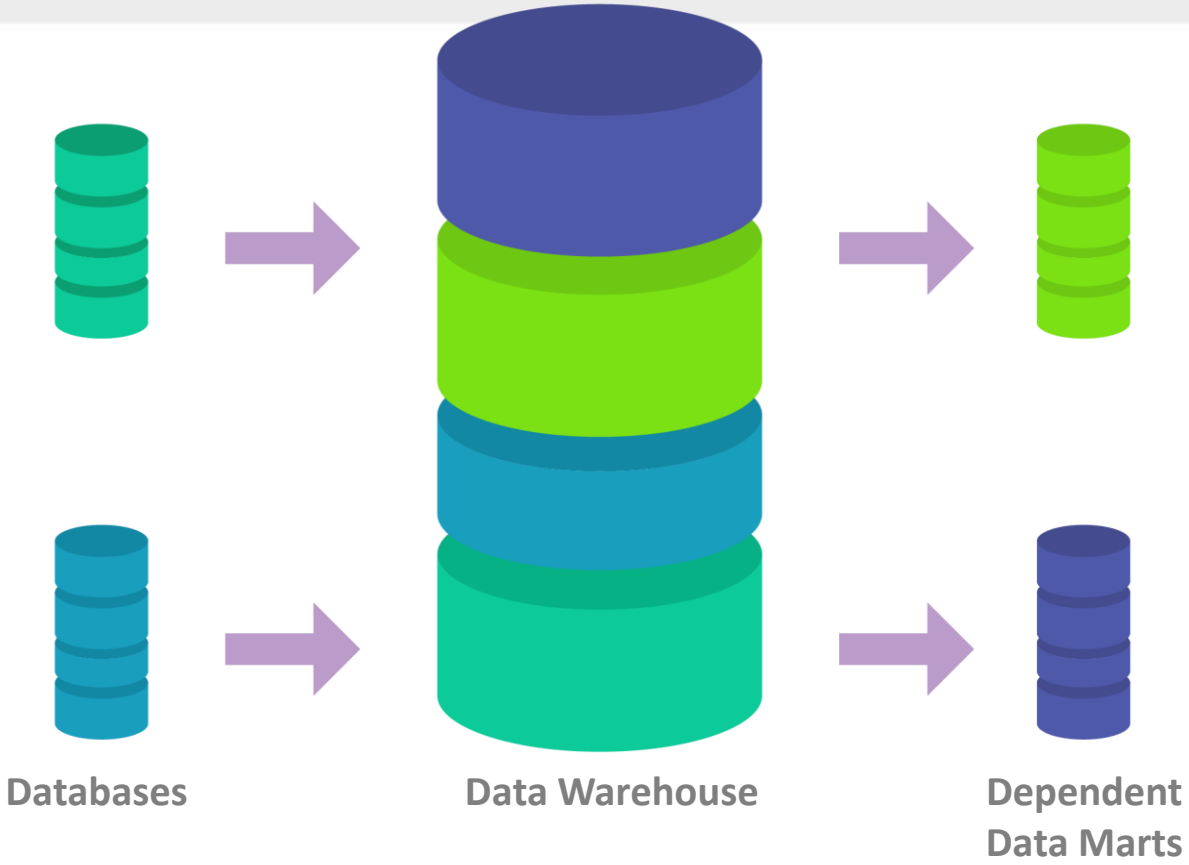
- Roles para el análisis de datos
- Fases del ciclo de vida
- Entregables
- Análisis sobre ambiente transaccional

Tareas de proyecto

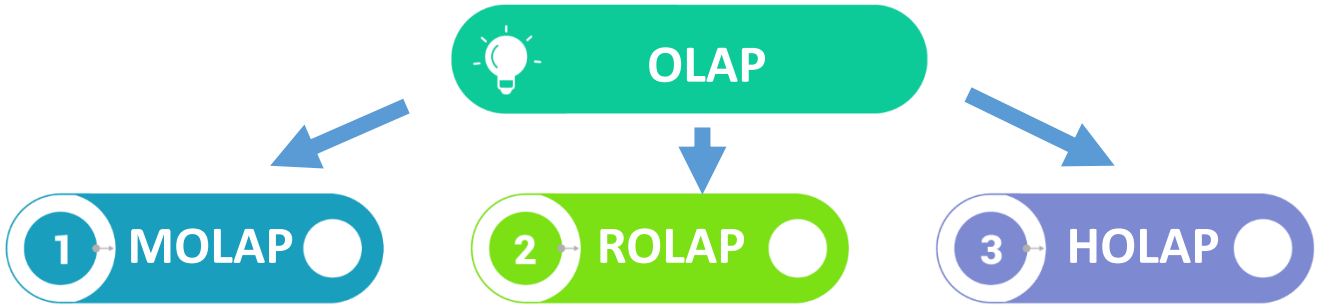


Business Intelligence

- Procesos ETL
- Esquemas multidimensionales (estrella y copo de nieve)
- DataWarehouse y DataMart
- Limpieza de datos

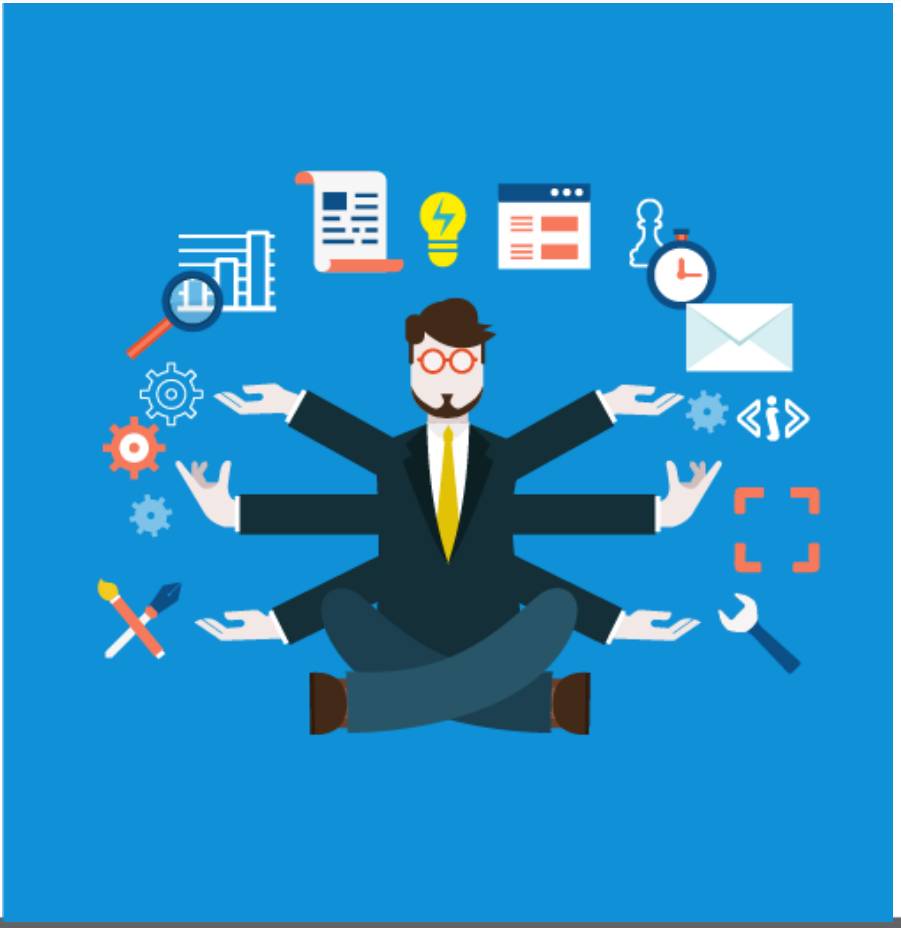


Business Intelligence

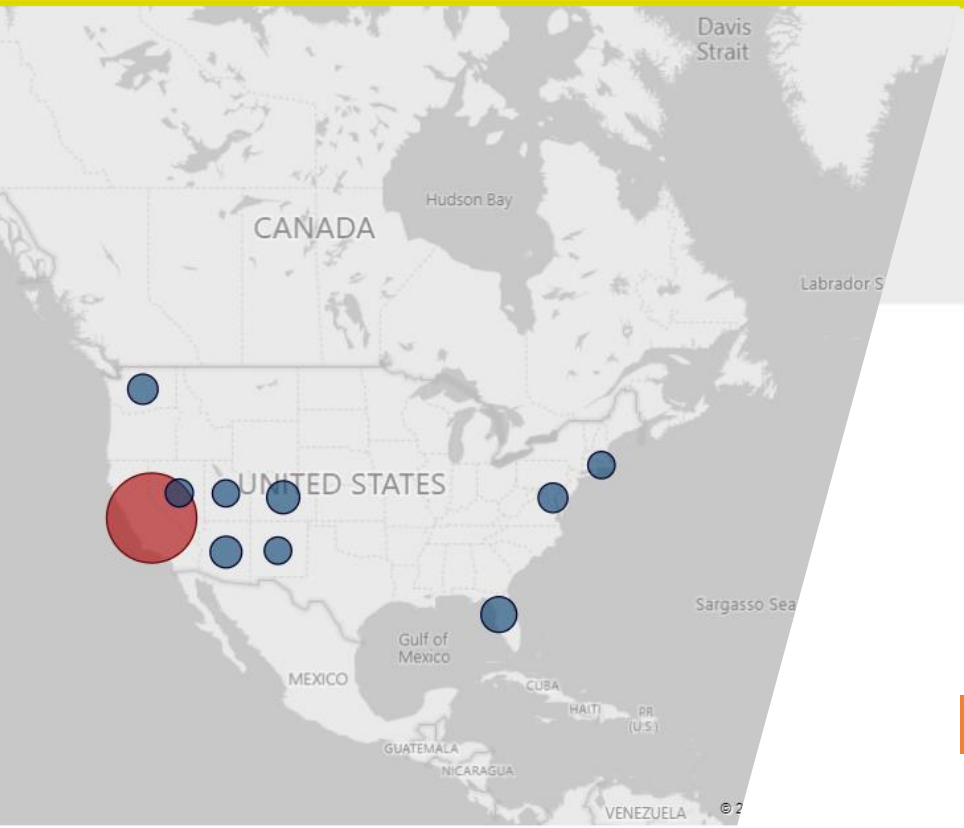


- Procesamiento transaccional
- Procesamiento analítico
- Hibrido

Visualización de datos



- Herramientas de Visualización
- “Self-Service” BI
- Reportería



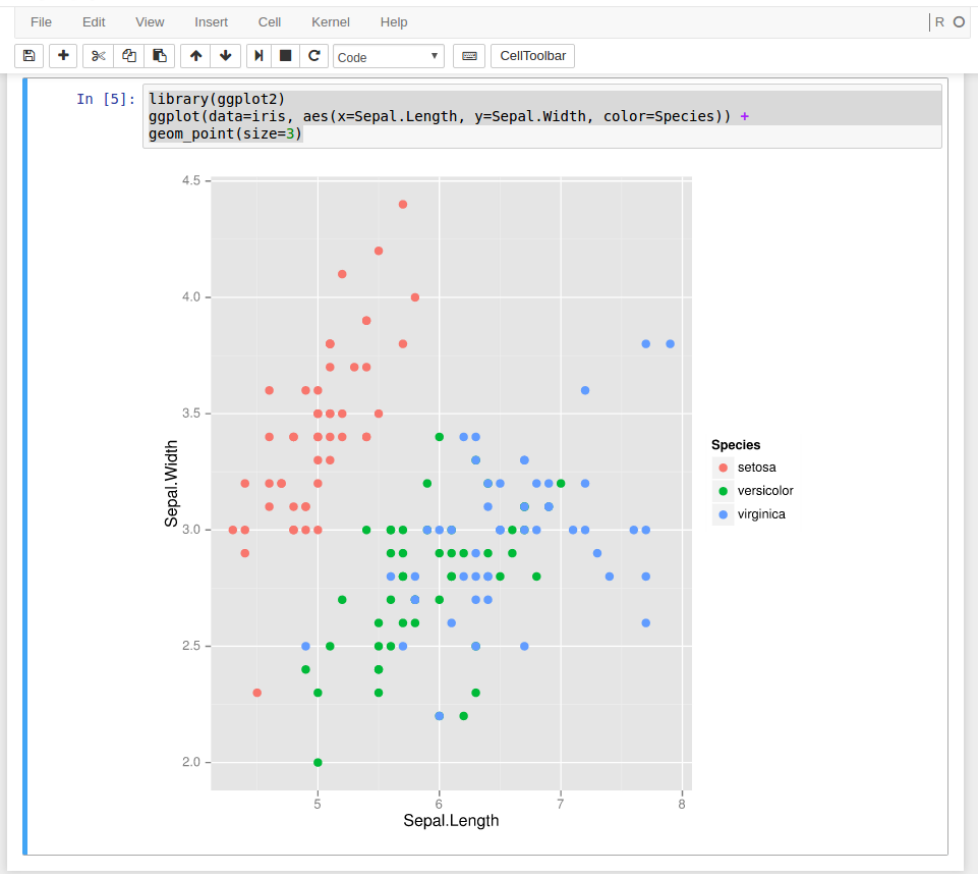
State with largest deficit
CALIFORNIA
-6445

Tablero de impacto de covid en el negocio

EJEMPLO DE TABLERO DE IMPACTO

Análisis básico de datos con R

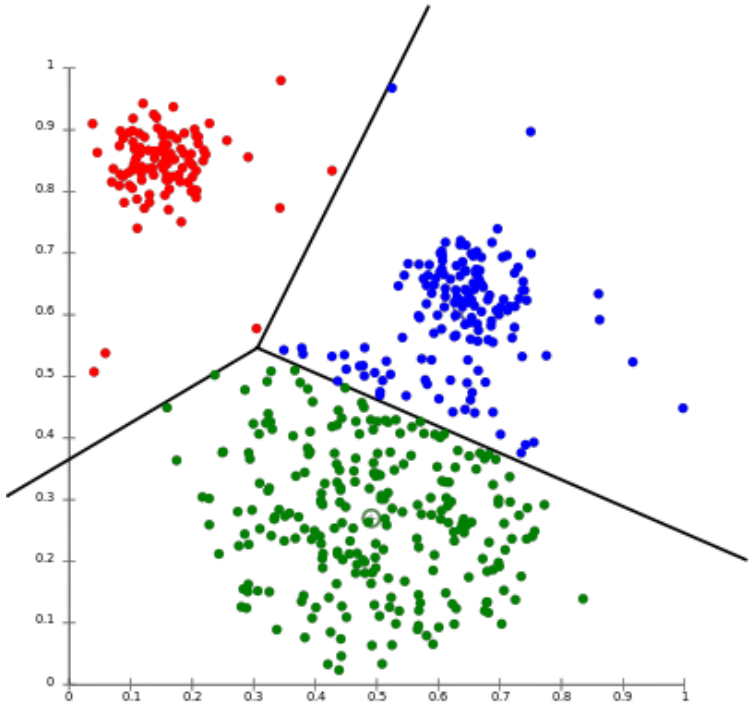
- Introducción a R
- Análisis y exploración de datos
- Construcción y evaluación de modelos estadísticos



Reglas de asociación

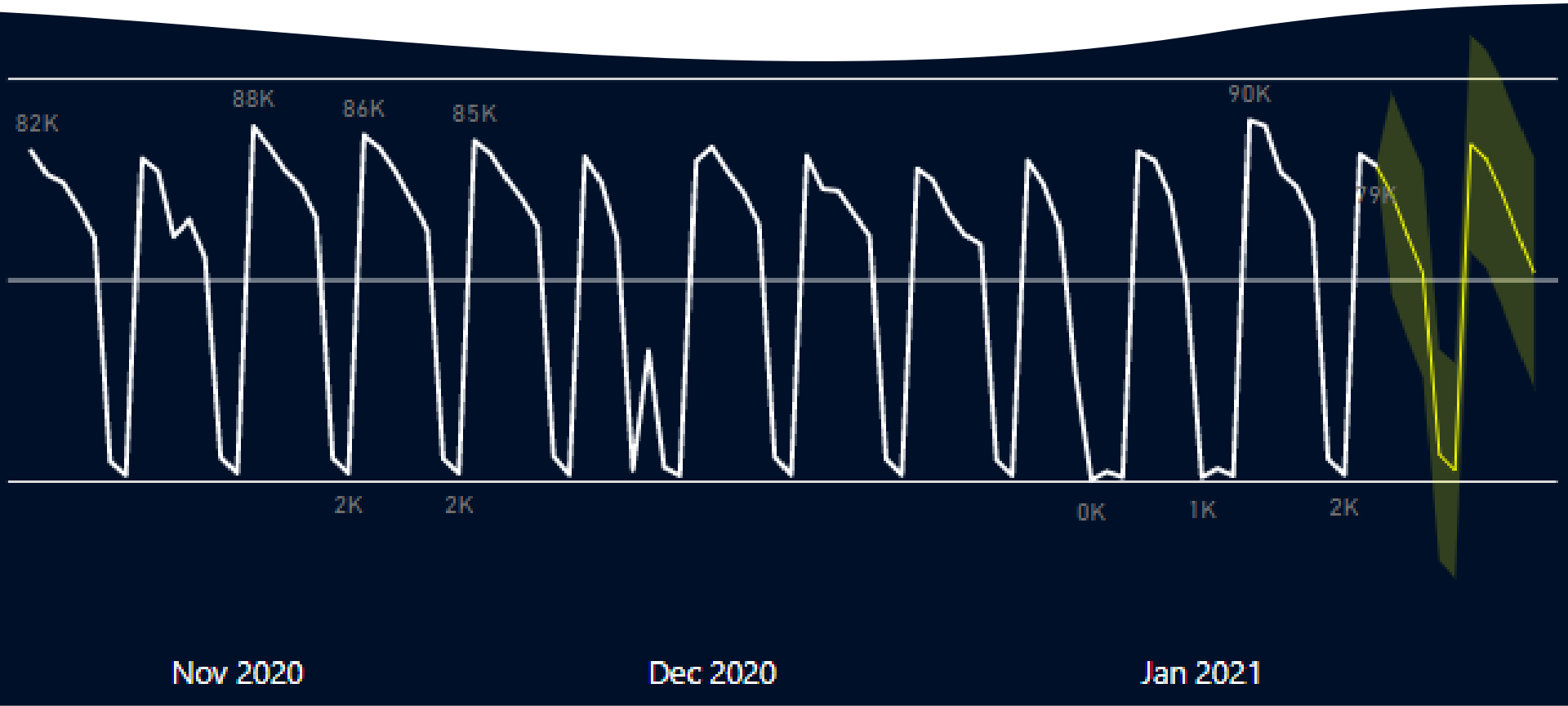
support	confidence	lift	count
0.001028418	0.9673549	290.98248	5482
0.001066313	0.9625741	290.38011	5684
0.001374539	0.9038983	272.67938	7327
0.001374539	0.9024510	271.45927	7327
0.002978323	0.8984720	270.26236	15876
0.002978323	0.8958862	270.26236	15876
0.001066313	0.8811037	265.03794	5684
0.001028418	0.8740434	263.67301	5482
0.001013035	0.5597595	12.45490	5400
0.001370974	0.5448852	12.12394	7308
0.001355779	0.5432609	12.08780	7227
0.001565327	0.5329927	50.12572	8344
0.001565327	0.5327544	50.83699	8344
0.001846538	0.5250440	49.37818	9843
0.001824589	0.5218931	49.80057	9726

Métodos de análisis avanzado a través de estadística



- Clustering y K-Means
- Reglas de asociación
- Regresión linear
- Regresión logística
- Clasificación Bayes
- Árboles de decisión
- Análisis de series de tiempo
- Análisis de texto

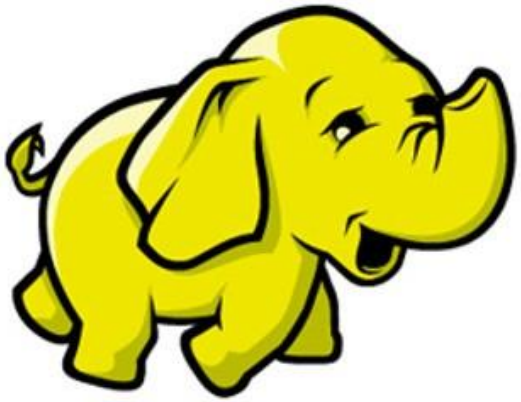
Series de Tiempo



Métodos de análisis avanzado a través de estadística

- Data Science en la Nube
  - Machine learning
  - Experimentos
- Auto ML

Herramientas avanzadas de análisis



hadoop

- Análisis de datos no estructurados
- Map Reduce
- HD insight