Universidad de Valparaíso
CHILE

Escuela Colombiana de Ingeniería Julio Garavito

III Simposio Data Analytics

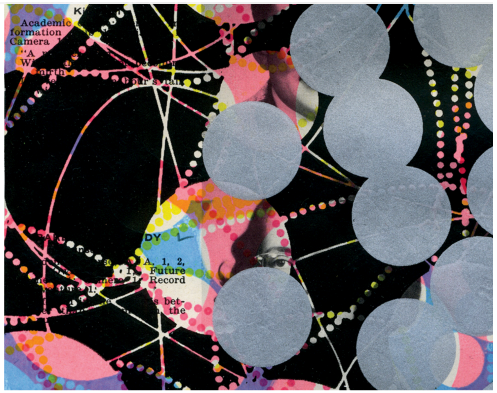
21 horas

Sistemas Inteligentes para la Toma de Decisiones"Importancia de los Datos para la Extracción de Conocimientos de una Organización"

Dr. Ing. Rodrigo Salas F. rodrigo.salas@uv.cl

Data Scientist: The Sexiest Job of the 21st Century

Harvard Business Review



DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing package e.g. R
- Databases SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

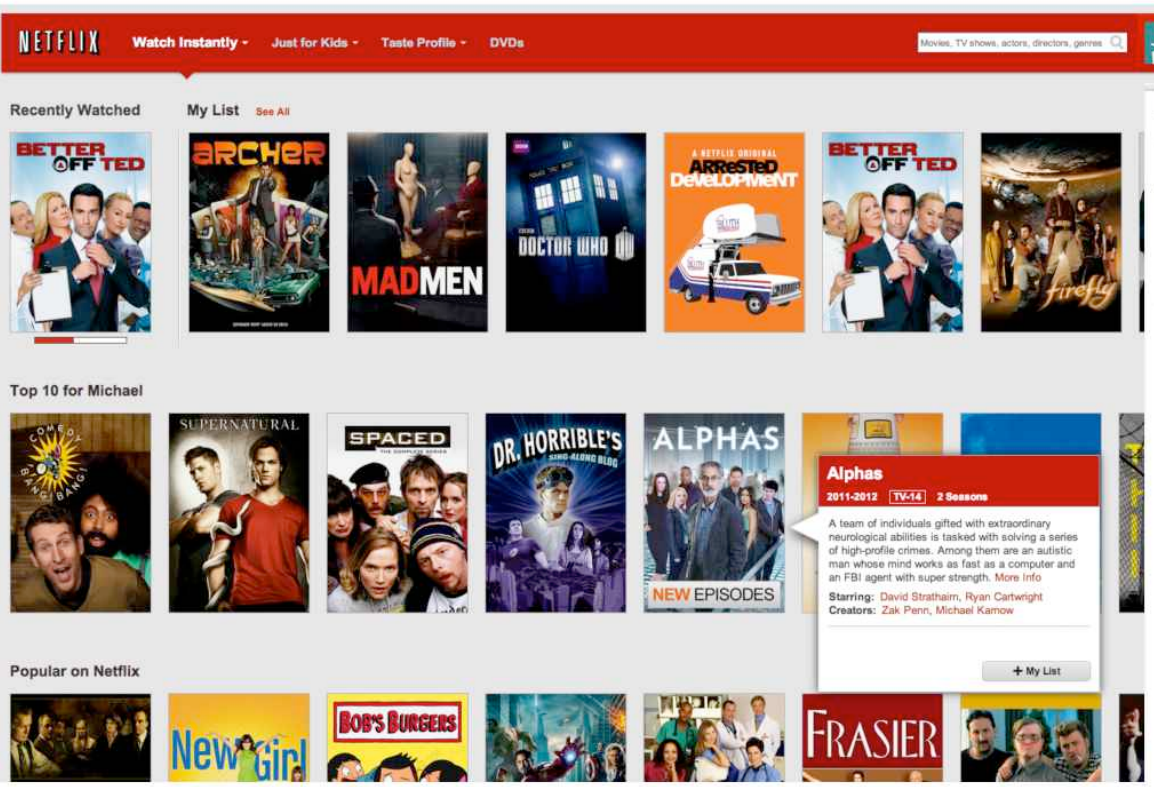
COMMUNICATION & VISUALIZATION

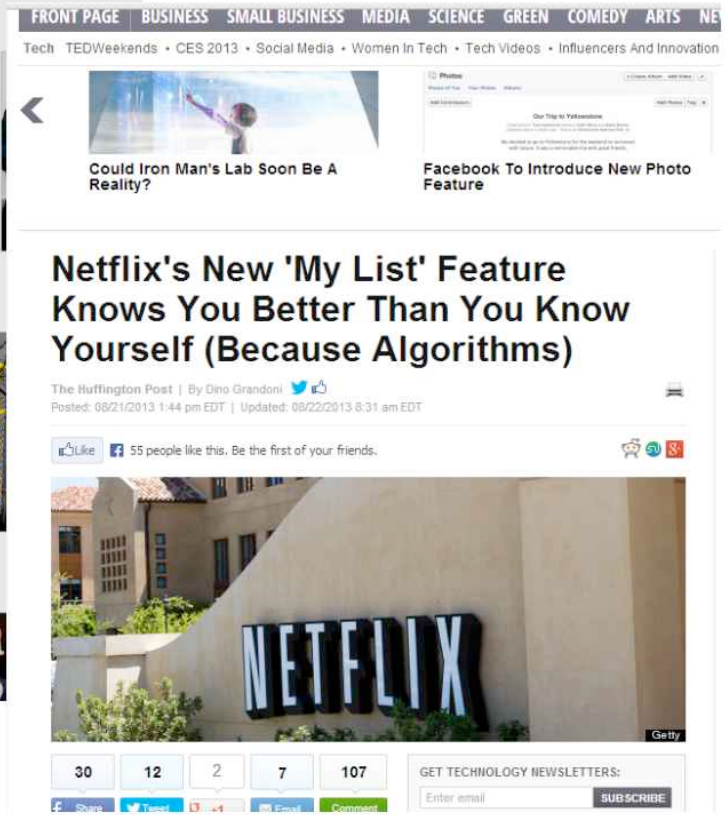
- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing DISTILLERY

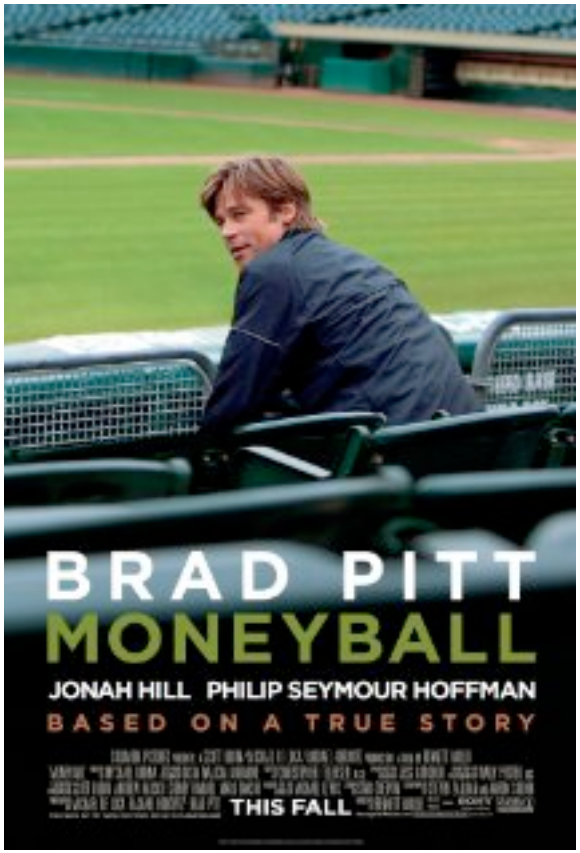
EVERYTHING is a Recommendation





http://cdn.infoq.com/statics_s1_20160217-0123u3/resource/presentations/machine-learning-netflix/en/slides/sl10.jpg

Organizaciones dirigidas por los datos



El equipo de Baseball conocido como los A's de Oakland tuvo una exitosa campaña gracias a la aplicación de la analítica computacional para identificar los jugadores sub-valorados.
El estudio fue llevado a cabo por el generen general Billy Blane y DePodesta.

Se vio reflejado el análisis del desempeño en el baseball y las lecciones aprendidas que pueden aplicarse a las organizaciones dirigidas por los datos.

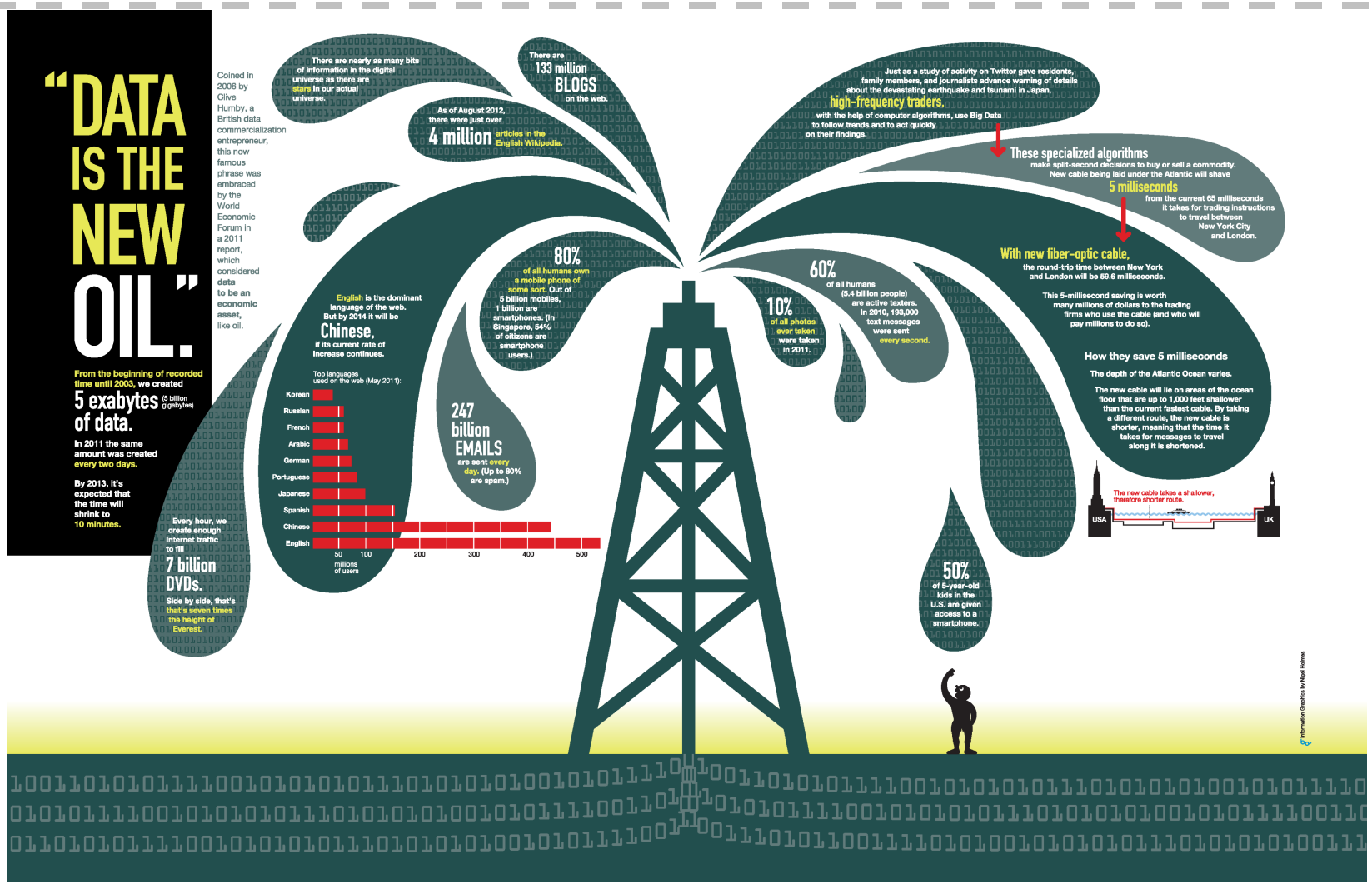
"Subjectivity ruled the day in evaluating players," he said. "We had a completely new set of metrics that bore no resemblance to anything you'd seen. We didn't solve baseball. But we reduced the inefficiency of our decision making."

PAUL DePODESTA

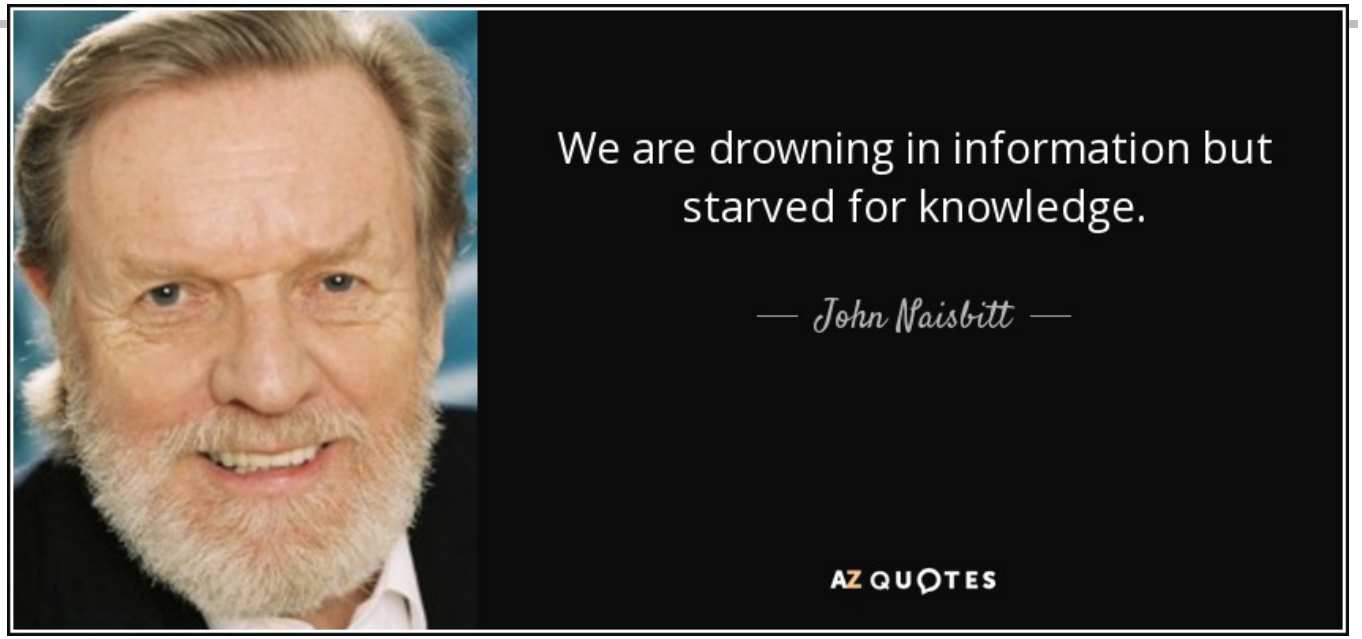
http://www.datacenterknowledge.com/archives/2011/09/23/the-lessons-of-moneyball-for-big-data-analysis/

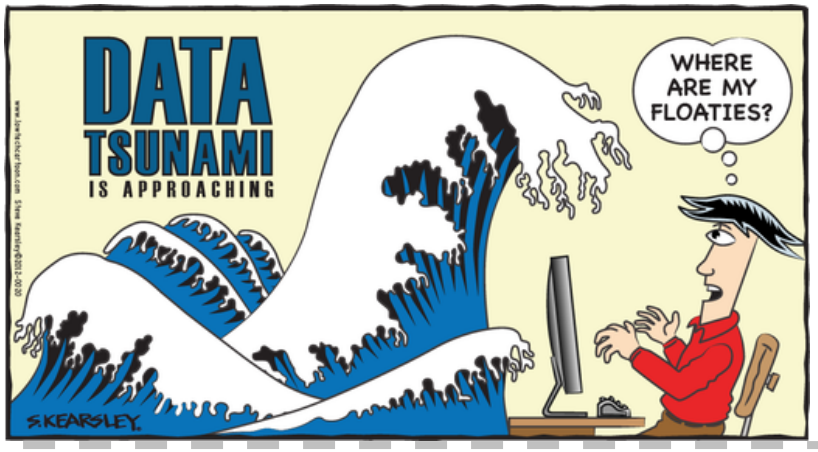
Rise of Data Analytics in Football

"Data is the New Oil" – World Economic Forum 2011



Desafíos en la Exploración de los Datos

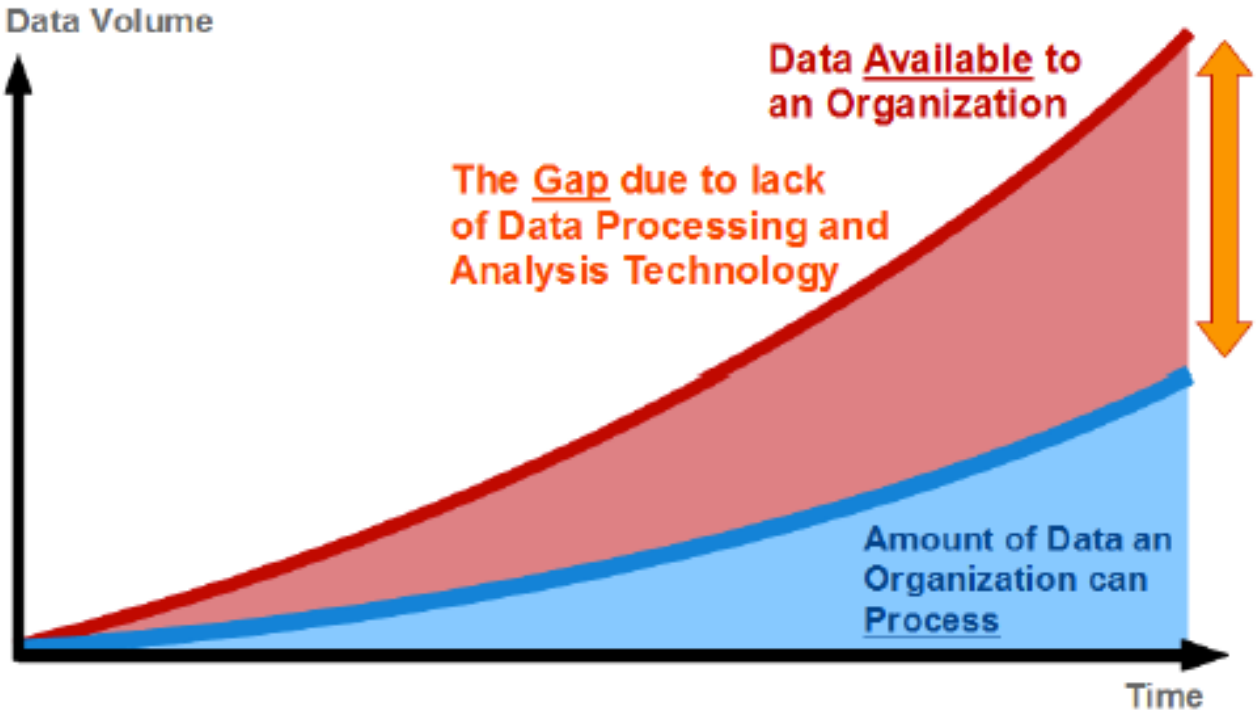




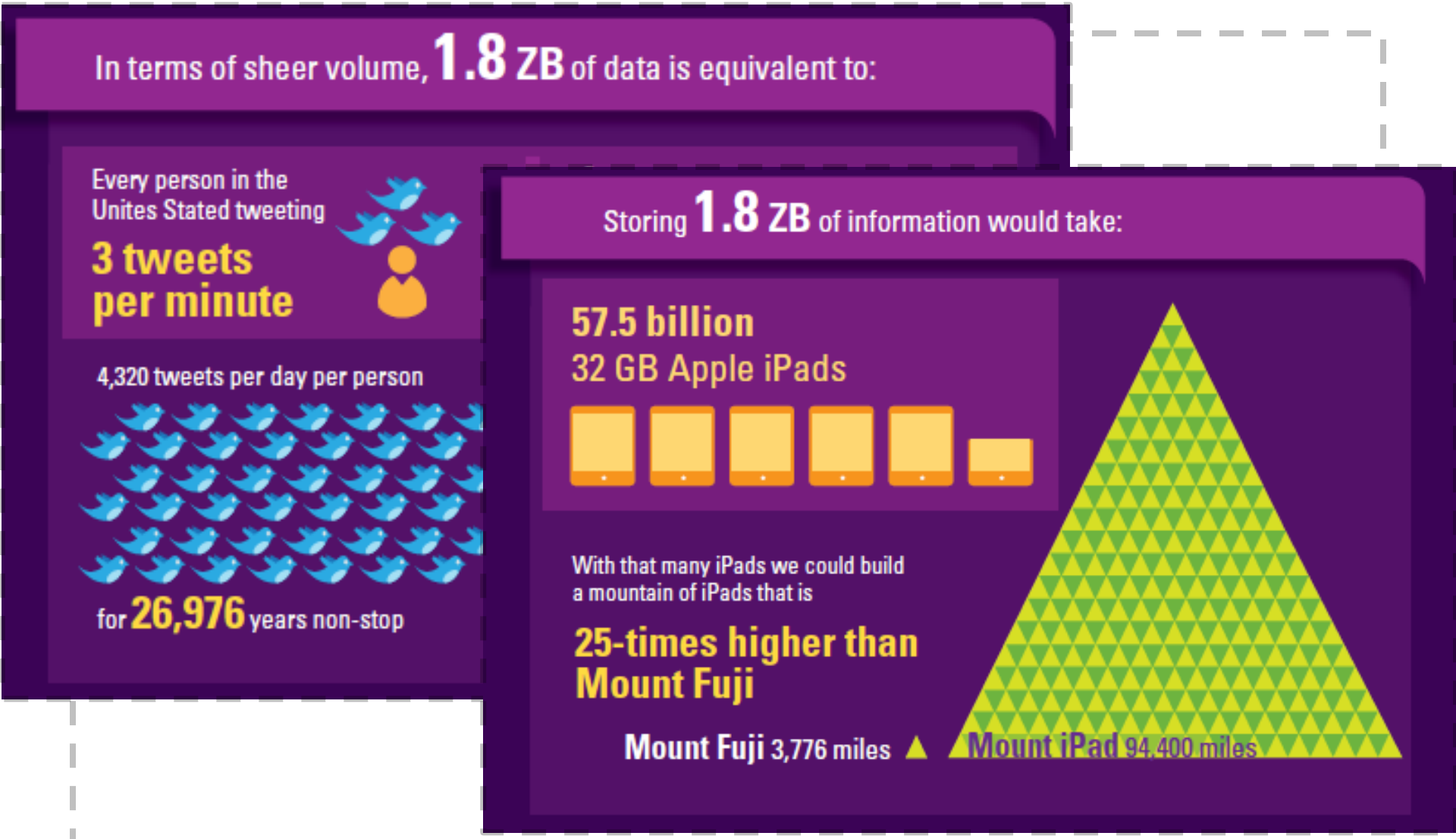
Ley de Moore en Procesamiento y Almacenamiento

THE WORLD'S INFORMATION IS DOUBLING EVERY TWO YEARS, with a collossal 1.8 zettabytes to be created and replicated in 2011.

New information being created in 2011 also includes replicated information such as shared documents or duplicated DVDs.



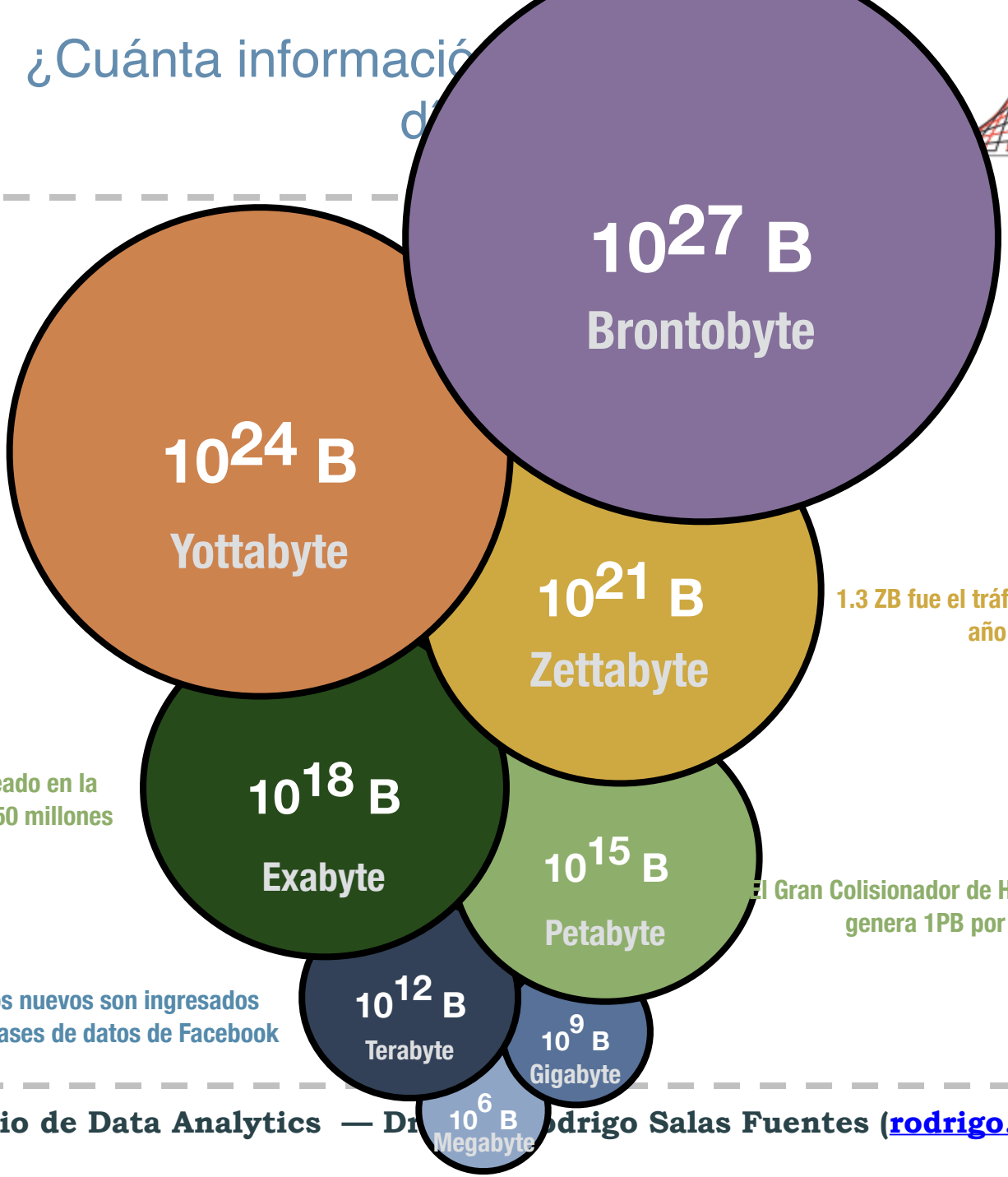
¿1.8 ZB?


In terms of sheer volume, 1.8 ZB of data is equivalent to:
Every person in the Unites Stated tweeting
3 tweets per minute
4,320 tweets per day per person
for 26,976 years non-stop
Storing 1.8 ZB of information would take:
57.5 billion
32 GB Apple iPads
With that many iPads we could build a mountain of iPads that is
25-times higher than Mount Fuji
Mount Fuji 3,776 miles
Mount iPad 94,400 miles

Universidad de Valparaíso CHILE

¿Cuánta información d

Escuela Colombiana de Ingeniería Julio Garavito



1 BB será el universo digital del futuro

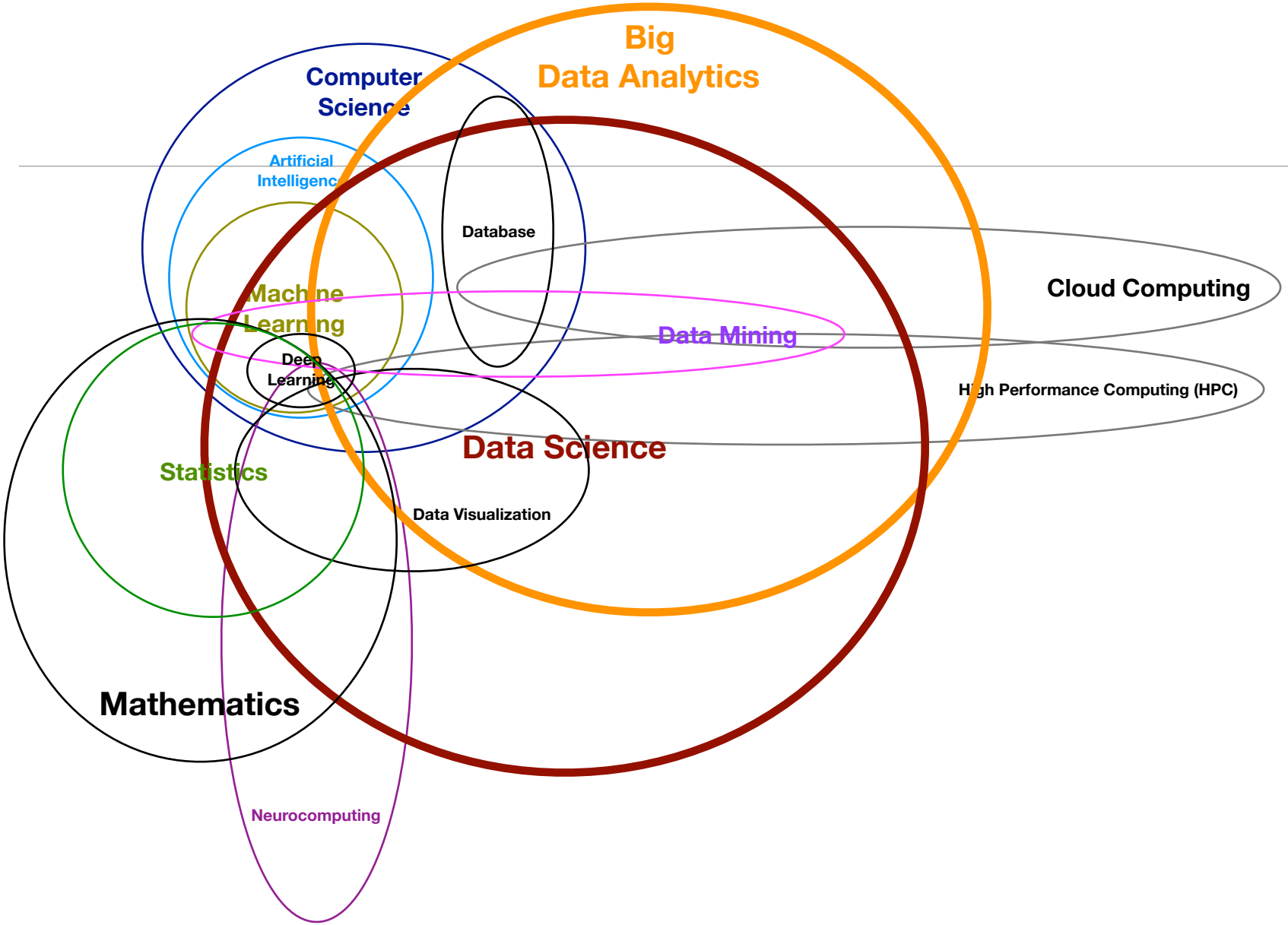
1 YB es el universo digital actual = 250 trillones de DVD

1.3 ZB fue el tráfico de la red en el año 2016

1 EB de datos es creado en la internet cada día = 250 millones de DVD

El Gran Colisionador de Hadrones del CERN genera 1PB por segundo

500TB de datos nuevos son ingresados cada día a las bases de datos de Facebook



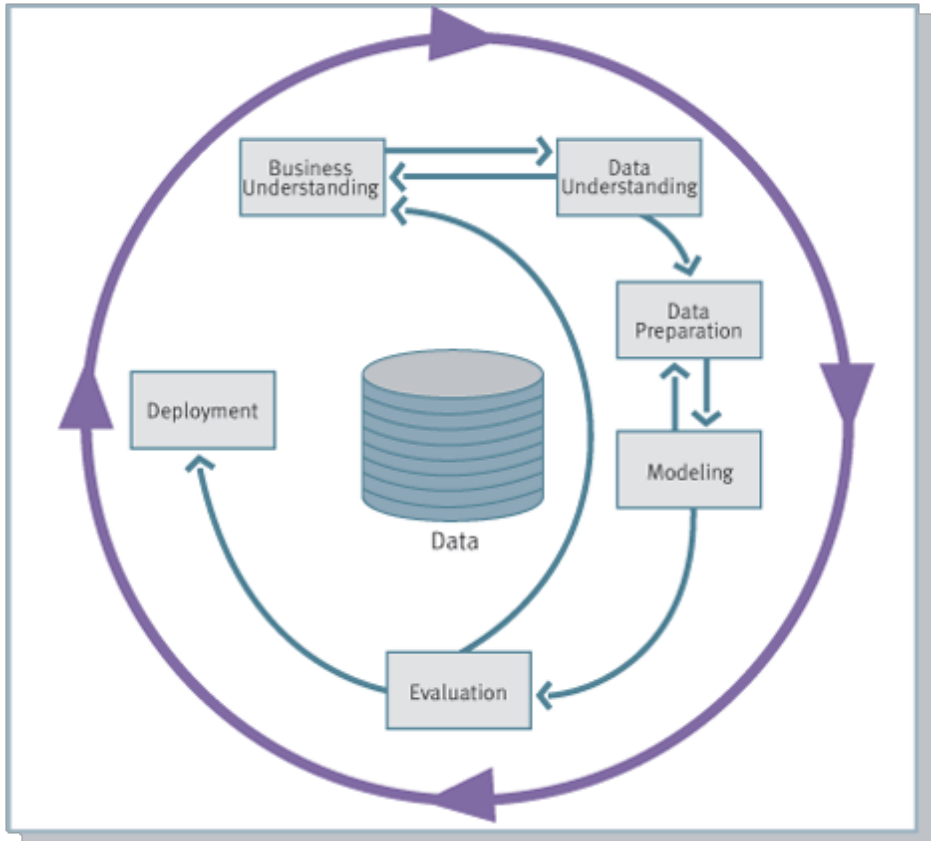
https://www.nisargdave.com/single-post/2016/10/05/Multidisciplinary-Data-Science

¿Cómo las compañías se enteran de los secretos?

- Cuando una persona va de compras, en general comparten los datos relativos a los patrones de consumos con las empresas de retail.
- La empresa TARGET descubrió cuándo una madre tendrá un bebé antes de que comience a comprar pañales.
  - "Andrew Pole analizó los datos y algunos patrones interesantes emergieron. Por ejemplo, las madres en su segundo trimestre de embarazo compran lociones sin esencias. Además durante las primeras 20 semanas compran suplementos como calcio, magnesio y zinc. Además cuando compran jabones y grandes bolsas de algodones, además de desinfectantes y toallas de mano, entonces están cerca del día del parto."
  - Además Andrew Pole identificó 25 productos que son buenos predicadores de embarazo.
  - Una familia recibió supones para ropas de bebés y el padre de una adolescente recibió la información de que su hija estaba embarazada.
  - http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=6&_r=1&hp

https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#659ddd786668

KDD - Knowledge Discovery Database



- **"Un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos"**

Fayyad, U, Piatetsky-Shapiro, G., Smyth, P. "From Data Mining to Knowledge Discovery: An Overview". Advances in Knowledge Discovery and Data Mining, pp. 1-34, AAAI/MIT Press, 1996.

Predictive Data Analytics

Figure 1.1

Predictive data analytics moving from **data** to **insight** to **decision**.

- Las organizaciones modernas recogen grandes cantidades de datos. Para que los datos sean valiosos para una organización, deben analizarse para extraer información (**insights**) que se pueda utilizar para tomar mejores decisiones.
- La extracción de información de los datos es el trabajo de análisis de datos. Este libro se enfoca en el análisis de datos predictivos, que es un sub campo importante de análisis de datos

Definición de Inteligencia de Negocios

- La ***Inteligencia de Negocios*** se define como la la combinación de tecnología, herramientas y procesos que me permiten transformar mis datos almacenados en información, esta información en conocimiento y este conocimiento dirigido a un plan o una estrategia comercial.
- La inteligencia de negocios debe ser parte de la estrategia empresarial, esta le permite optimizar la utilización de recursos, monitorear el cumplimiento de los objetivos de la empresa y la capacidad de tomar buenas decisiones para así obtener mejores resultados.

Data Warehouse Institute

Universidad de Valparaíso
CHILE

Escuela Colombiana de Ingeniería Julio Garavito

III Simposio Data Analytics

21 horas

Sistemas Inteligentes para la Toma de Decisiones

"Visualización de los Datos"

Rodrigo Salas

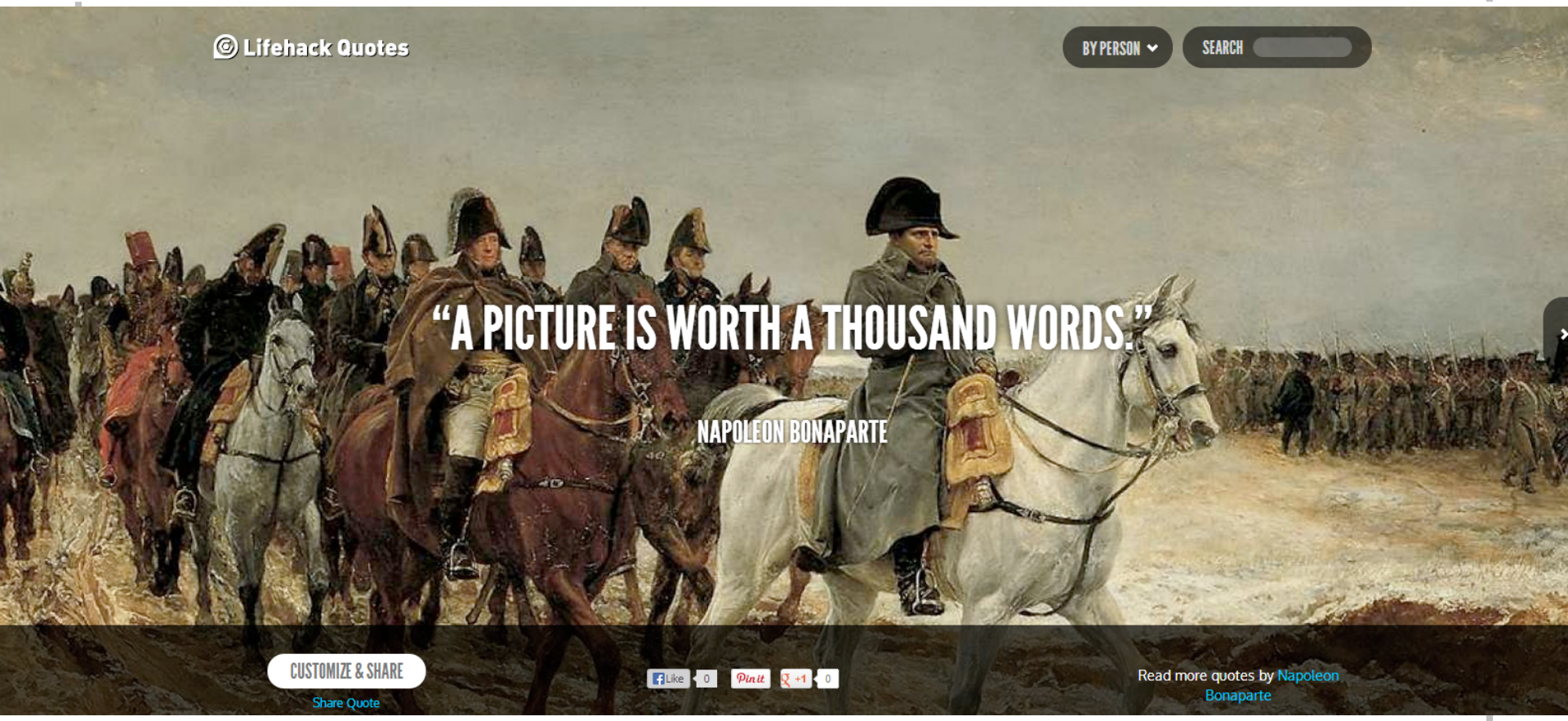
Raw Data

- Difícil de entender

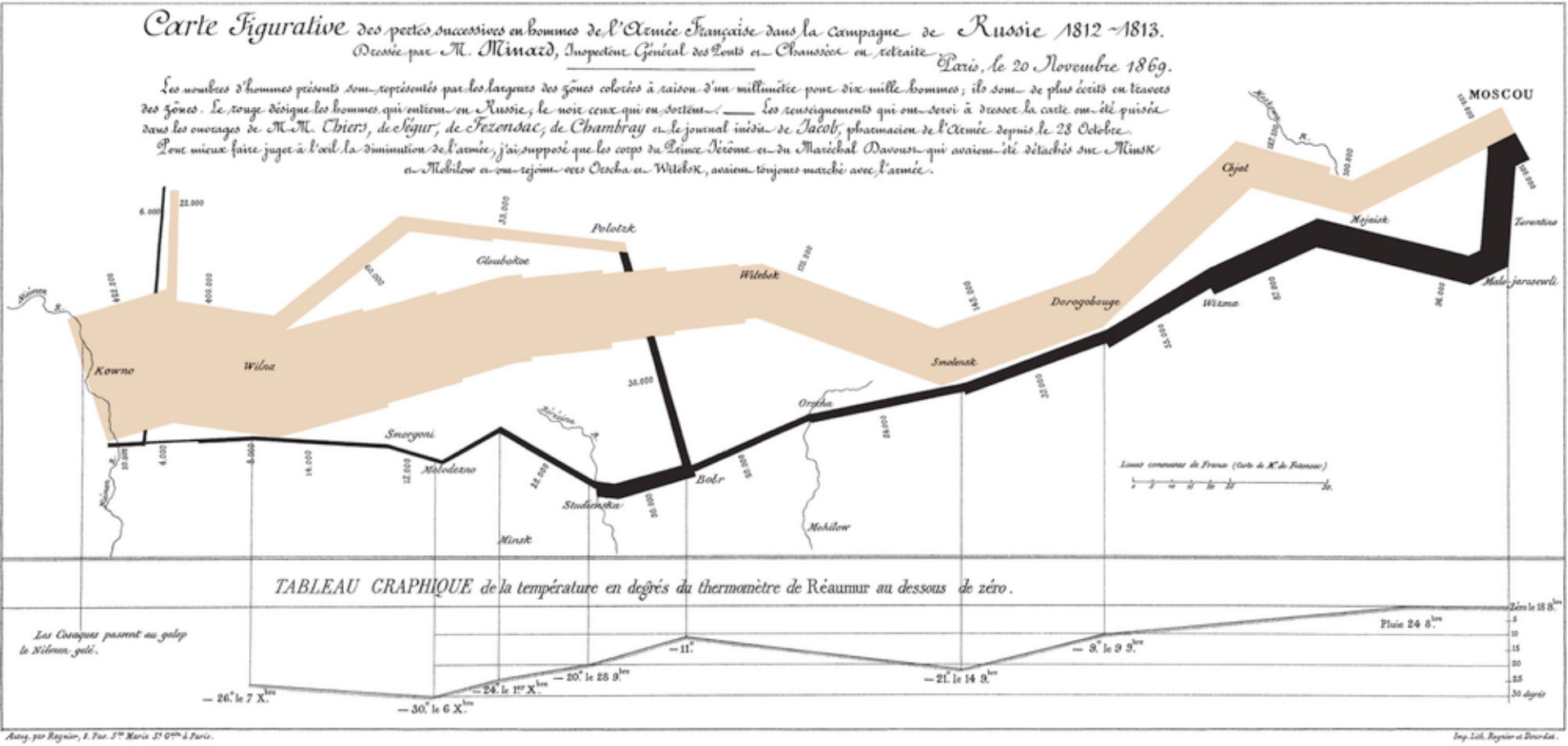
```
RespondentId,StartDate,CompletedDate,LanguageCode,Question1,Question2,Question3,Question4,Question5,Question6,Question7,Questi
27357,2006.11.27 15:6,2006.11.27 15:7,en,Denmark,Financial Services,6 - 12 months,26-100,4,4,2,"cvbcvb",2,3,3,1,Opinio,1,0,0,1
27359,2006.11.27 15:7,2006.11.27 15:8,en,Italy,Hardware Vendor,1 - 2 years,26-100,3,5,4,,1,3,3,4,Opinio,0,0,0,0,1,0,0,1,0,,,,0
27360,2006.11.27 15:8,2006.11.27 15:8,en,Lithuania,Retail,6 - 12 months,6-10,4,1,4,"this is a random other text",2,2,2,2,Opini
27361,2006.11.27 15:8,2006.11.27 15:8,en,Panama,Retail,6 - 12 months,6-10,4,1,4,"this is a random other text",2,2,2,2,Opinio,0
27362,2006.11.27 15:8,2006.11.27 15:8,en,Djibouti,Manufacturing,6+ years,101-250,0,4,0,"another random text",5,5,5,5,Opinio,1,
27363,2006.11.27 15:8,2006.11.27 15:8,en,Tanzania,Retail,1 - 2 years,1001-5000,1,1,1,"123456",2,2,2,2,Opinio,0,1,1,1,1,1,1,1,1
27364,2006.11.27 15:8,2006.11.27 15:8,en,Vanuatu,Other,1 - 2 years,1001-5000,6,5,6,"123456",6,6,6,6,Opinio,0,0,1,1,1,1,0,1,1,"
27365,2006.11.27 15:8,2006.11.27 15:8,en,Angola,Government,1 - 2 years,11-25,4,2,4,"123456",3,3,3,3,Opinio,0,0,1,1,1,1,1,1,0,,
27366,2006.11.27 15:8,2006.11.27 15:8,en,Panama,Manufacturing,<6 months,1-5,1,4,1,"hey",5,5,5,5,Opinio,0,1,0,0,0,1,0,0,0,,"hey
27367,2006.11.27 15:8,2006.11.27 15:8,en,Norway,Education,2 - 5 years,5001-10000,6,0,6,"£6{[]}+åøæ´´``*-/+\",1,1,1,1,Opinio,1,
27368,2006.11.27 15:8,2006.11.27 15:8,en,Bermuda,Software Vendor,1 - 2 years,11-25,0,2,0,"123456",3,3,3,3,Opinio,1,0,1,0,0,1,0
27369,2006.11.27 15:8,2006.11.27 15:8,en,Panama,Transportation,1 - 2 years,11-25,5,4,5,"123456",5,5,5,5,Opinio,0,1,0,0,0,1,0,0
27370,2006.11.27 15:8,2006.11.27 15:8,en,Maldives,Other,6+ years,10001 or more,2,5,2,"another random text",6,6,6,6,Network Pro
27371,2006.11.27 15:8,2006.11.27 15:8,en,Kyrgyzstan,Medical,2 - 5 years,26-100,3,5,3,"£6{[]}+åøæ´´``*-/+\",6,6,6,6,Network Pro
27372,2006.11.27 15:8,2006.11.27 15:8,en,Antigua and Barbuda,Government,6 - 12 months,501-1000,6,2,6,"this is a random other t
27373,2006.11.27 15:8,2006.11.27 15:8,en,Belarus,Financial Services,6+ years,10001 or more,2,1,2,"another random text",2,2,2,2
27374,2006.11.27 15:8,2006.11.27 15:8,en,Vatican City,Non-profit,1 - 2 years,11-25,0,0,0,"123456",1,1,1,1,Network Probe,1,0,0,
27375,2006.11.27 15:8,2006.11.27 15:8,en,Georgia,Financial Services,6+ years,10001 or more,6,1,6,"another random text",2,2,2,2
27376,2006.11.27 15:8,2006.11.27 15:8,en,Tokelau,Transportation,1 - 2 years,11-25,2,4,2,"123456",5,5,5,5,Network Probe,0,1,0,0
27377,2006.11.27 15:8,2006.11.27 15:8,en,Chad,Software Vendor,<6 months,1-5,6,2,6,"hey",3,3,3,3,Network Probe,1,1,1,1,1,1,1,1,
27378,2006.11.27 15:8,2006.11.27 15:8,en,Turkey,Software Vendor,6 - 12 months,501-1000,1,2,1,"this is a random other text",3,3
27379,2006.11.27 15:8,2006.11.27 15:8,en,East Timor,Transportation,<6 months,1-5,0,4,0,"hey",5,5,5,5,Opinio,1,1,0,0,1,0,1,1,0,
27380,2006.11.27 15:8,2006.11.27 15:8,en,Nicaragua,Medical,6 - 12 months,6-10,5,5,5,"this is a random other text",6,6,6,6,Opin
27381,2006.11.27 15:8,2006.11.27 15:8,en,Equatorial Guinea,Software Vendor,6+ years,101-250,6,2,6,"another random text",3,3,3,
27382,2006.11.27 15:8,2006.11.27 15:8,en,Zambia,Retail,<6 months,251-500,1,1,1,"hey",2,2,2,2,Surveyor,0,1,0,0,0,0,0,1,0,,"hey"
27383,2006.11.27 15:8,2006.11.27 15:8,en,French Southern and Antarctic Lands,Retail,1 - 2 years,1001-5000,2,1,2,"123456",2,2,2
27384,2006.11.27 15:8,2006.11.27 15:8,en,Guinea-Bissau,Hardware Vendor,2 - 5 years,26-100,6,3,6,"£6{[]}+åøæ´´``*-/+\",4,4,4,4,
27385,2006.11.27 15:8,2006.11.27 15:8,en,Viet Nam,Medical,2 - 5 years,26-100,4,5,4,"£6{[]}+åøæ´´``*-/+\",6,6,6,6,Opinio,1,1,1,
27386,2006.11.27 15:8,2006.11.27 15:8,en,Reunion,Medical,1 - 2 years,1001-5000,2,5,2,"123456",6,6,6,6,Opinio,1,1,1,1,1,1,1,1,1
27387,2006.11.27 15:8,2006.11.27 15:8,en,Puerto Rico,Non-profit,<6 months,1-5,0,0,0,"hey",1,1,1,1,Opinio,1,1,1,1,0,1,1,1,0,,"h
27388,2006.11.27 15:8,2006.11.27 15:8,en,East Timor,Financial Services,6 - 12 months,6-10,1,1,1,"this is a random other text",
27389,2006.11.27 15:8,2006.11.27 15:8,en,Northern Mariana Islands,Software Vendor,<6 months,1-5,2,2,2,"hey",3,3,3,3,Opinio,1,0
```

Universidad de Valparaíso CHILE

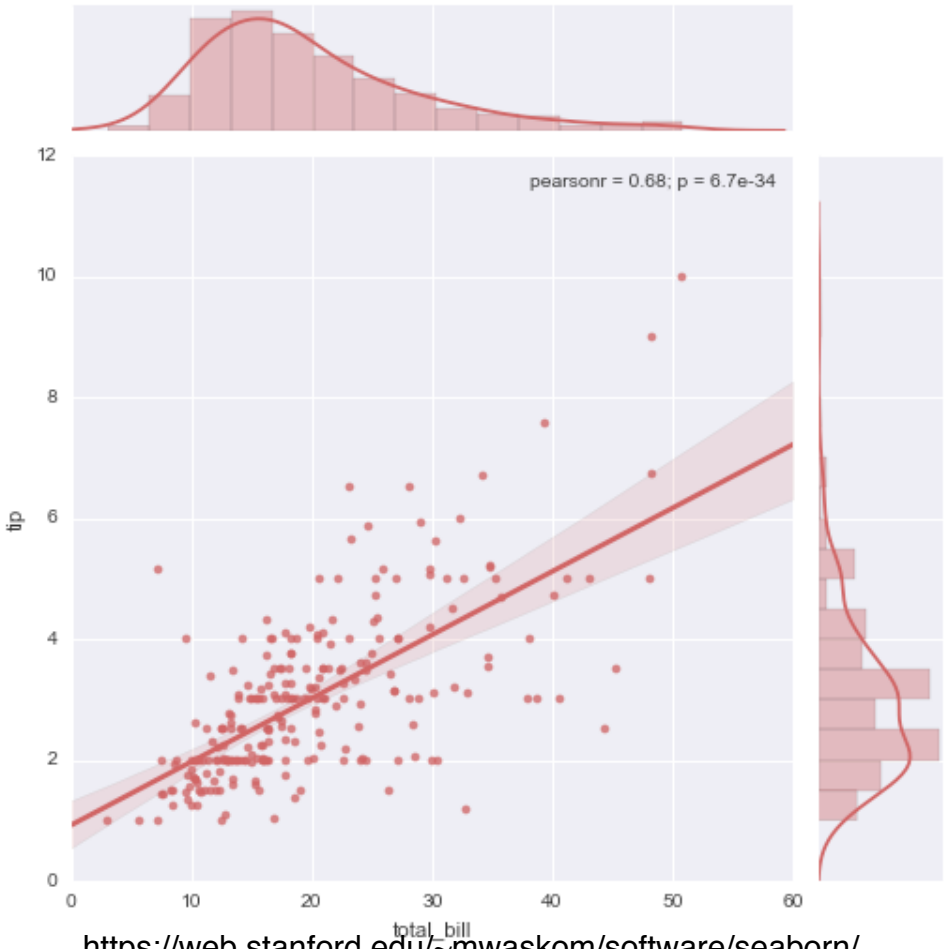
ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO



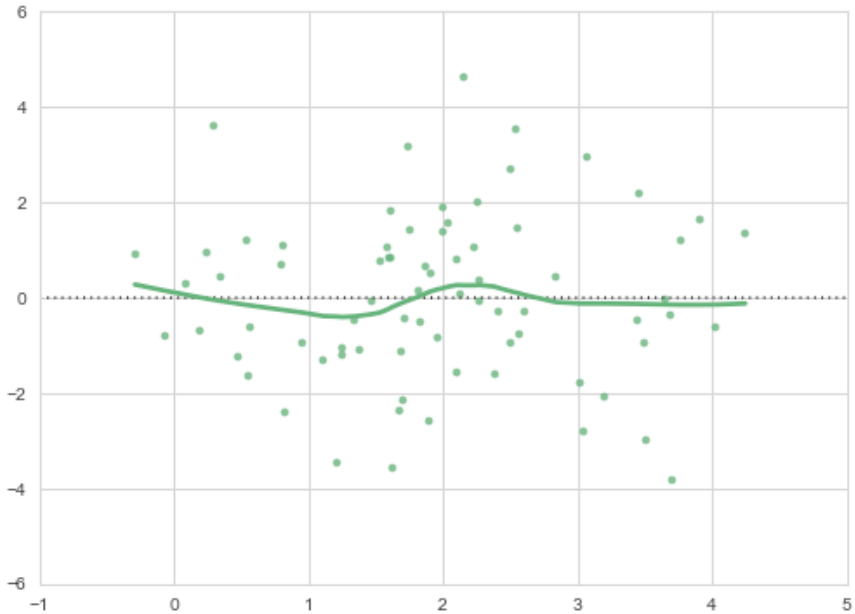
Visualización de la campaña de Napoleón a Rusia, 1812



Linear Regression - Plotting model residuals

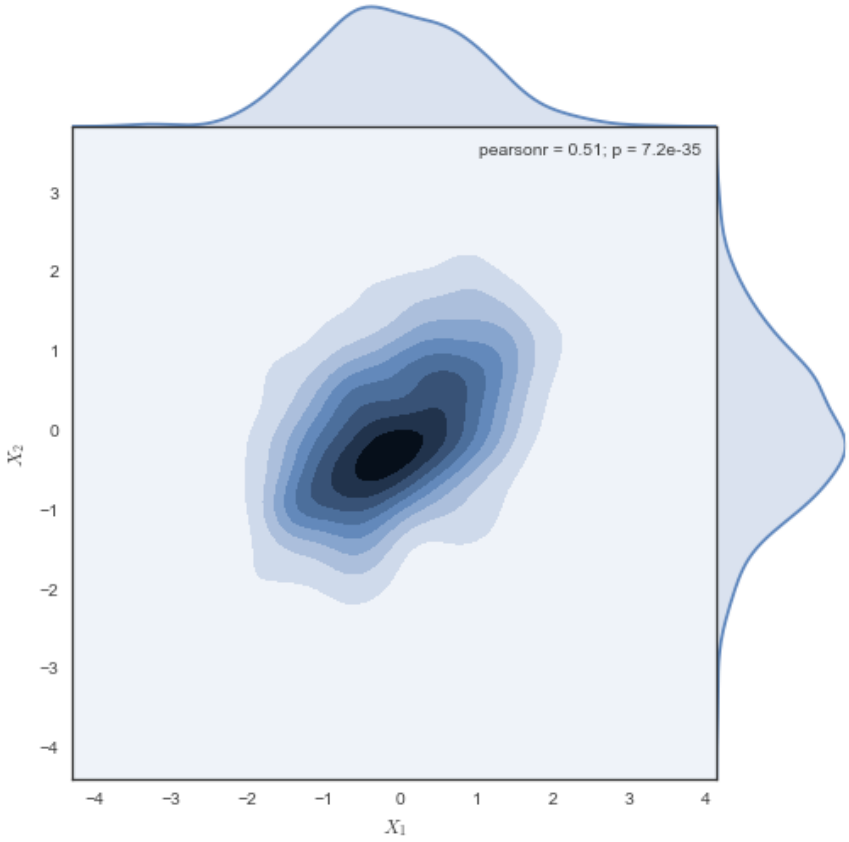


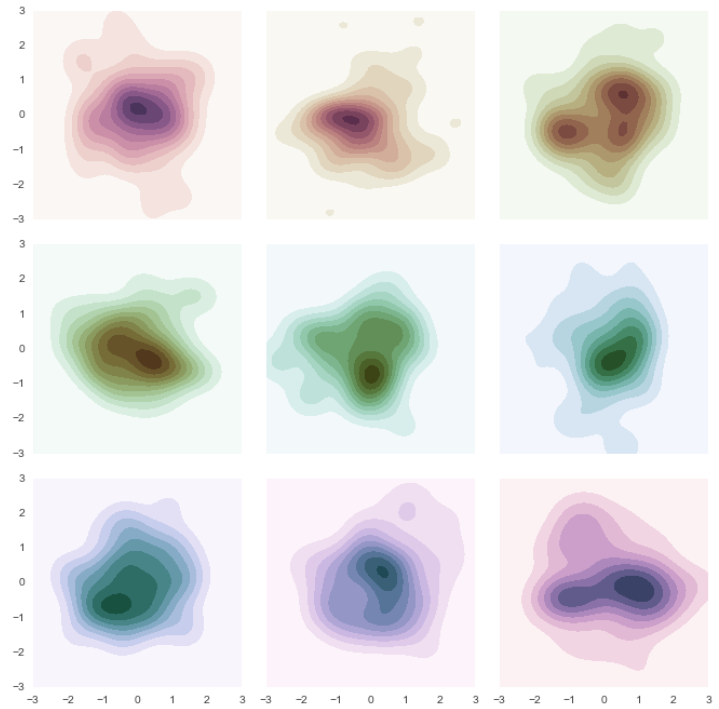
https://web.stanford.edu/~mwaskom/software/seaborn/examples/regression_marginals.html



https://web.stanford.edu/~mwaskom/software/seaborn/examples/residplot.html

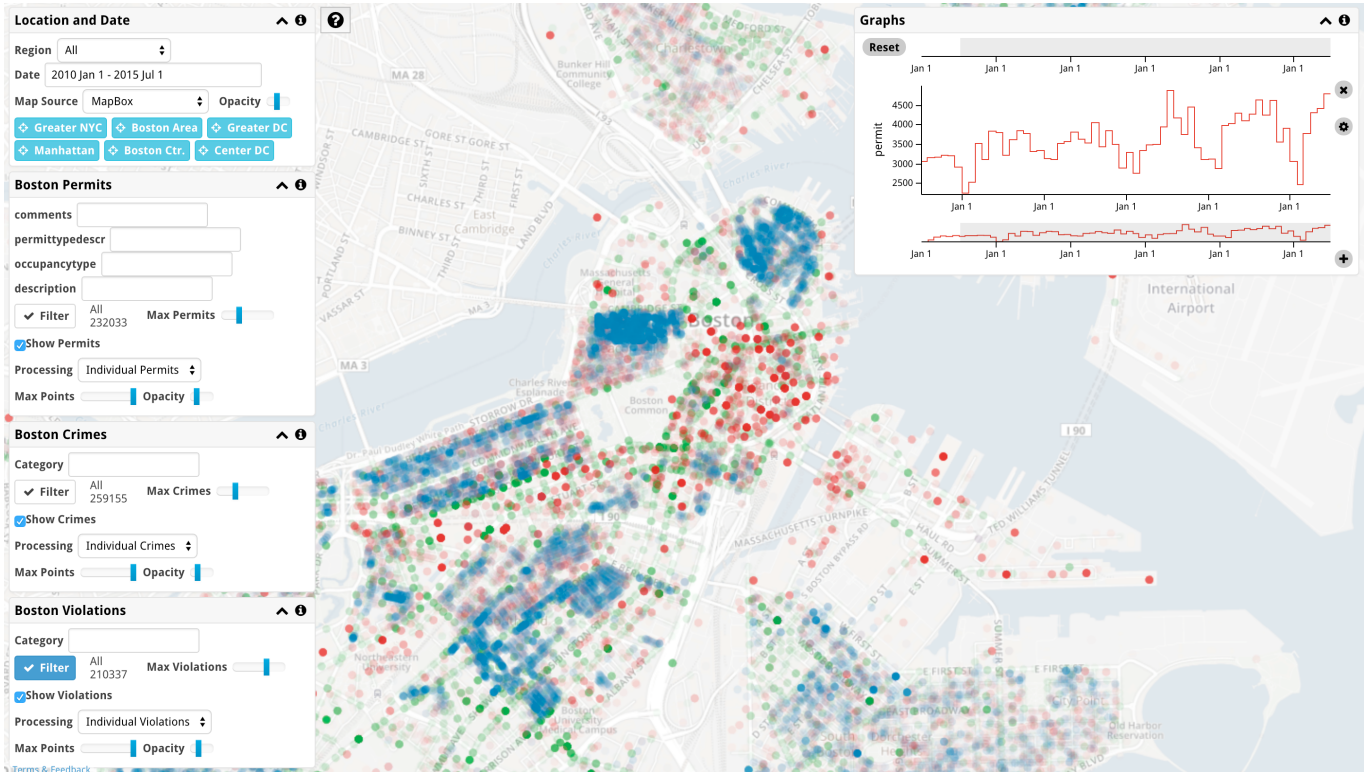
Kernel density estimate (KDE)-plot

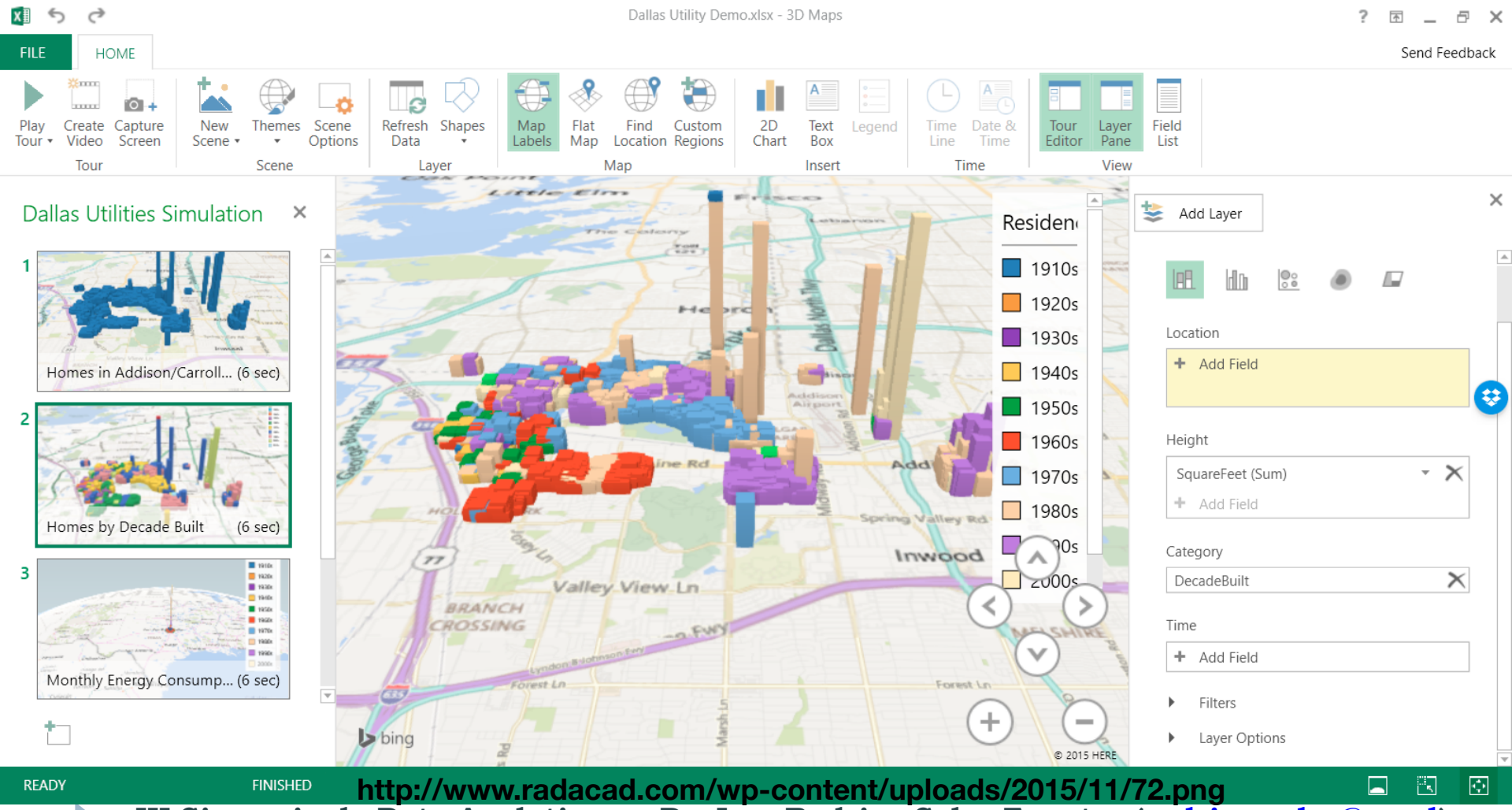




https://web.stanford.edu/~mwaskom/software/seaborn/examples/joint_kde.html

Boston Census Data





http://www.radacad.com/wp-content/uploads/2015/11/72.png

Universidad de Valparaíso
CHILE

Escuela Colombiana de Ingeniería Julio Garavito

III Simposio Data Analytics

21 horas

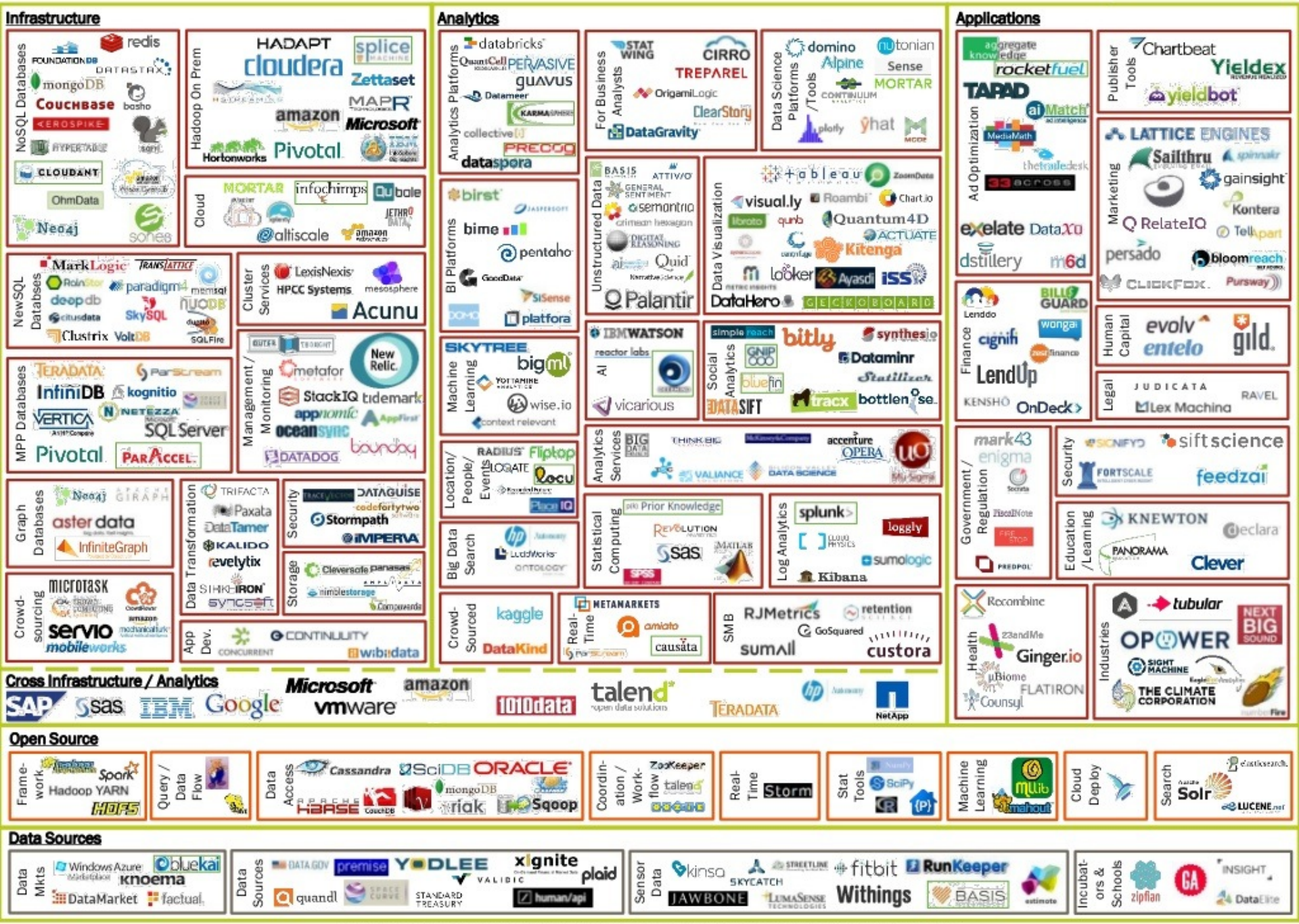
Sistemas Inteligentes para la Toma de Decisiones

"Analítica de Datos"

Rodrigo Salas

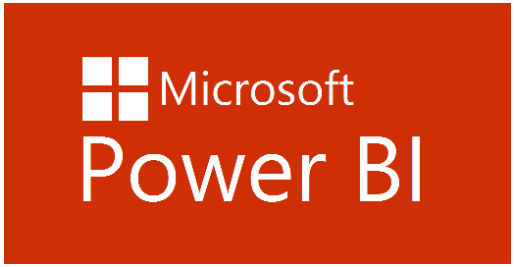
BIG DATA LANDSCAPE, VERSION 3.0

Exited: Acquisition or IPO



© Matt Turck (@mattturck), Sutian Dong (@sutiandong) & FirstMark Capital (@firstmarkcap)

Business Intelligence Platforms



Pentaho

Cuadrante de Gartner - Herramientas BI 2016

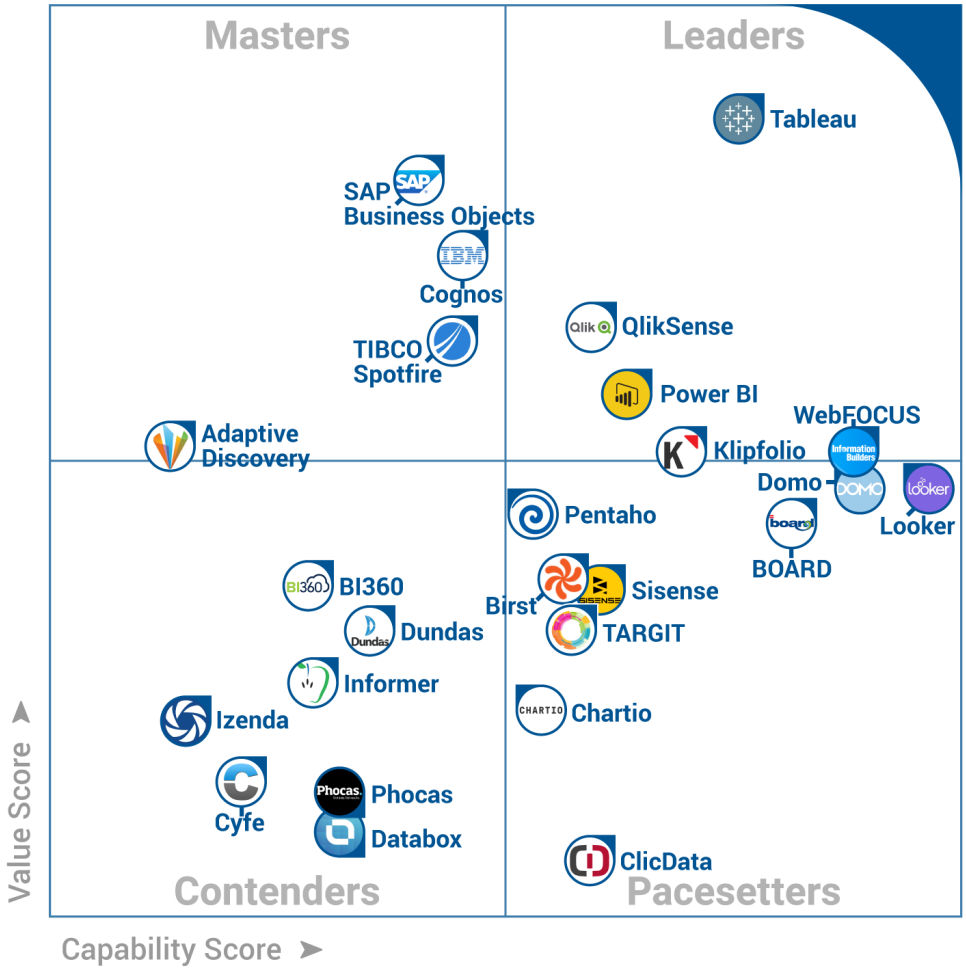
Habilidad para Ejecutar

Completitud de la Visión

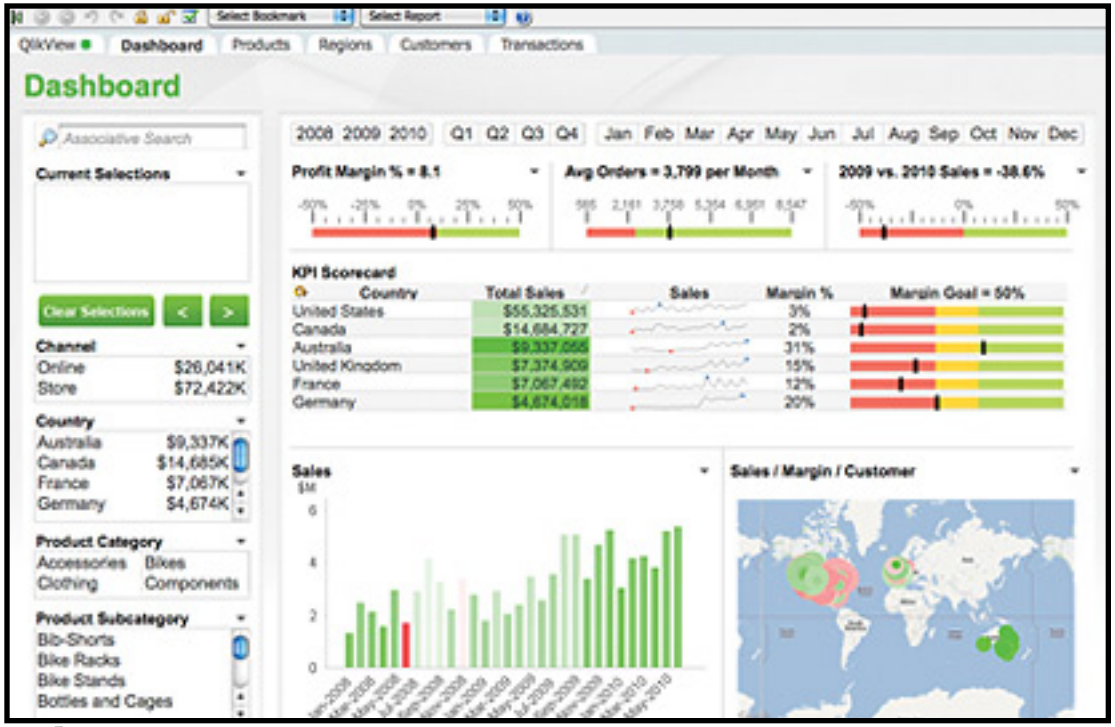
Febrero 2016

"The BI and analytics platform market's multiyear shift of focus from IT-led reporting to business-led self-service analytics has reached a tipping point. Modern BI platforms support organizational needs for greater accessibility, agility and analytical insight from a diverse range of data sources." – Gartner

Cuadrante Mágico de Gartner 2017

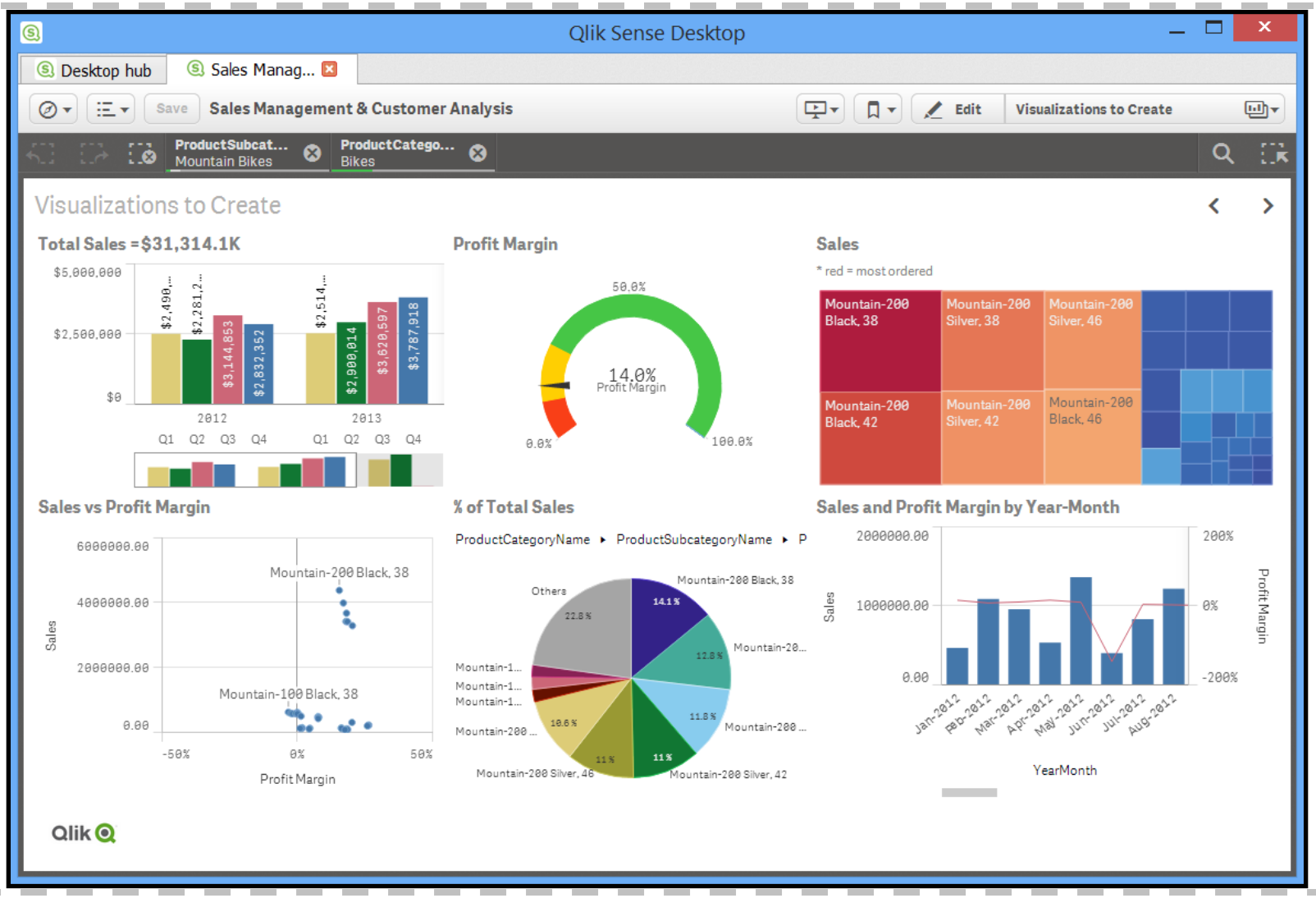


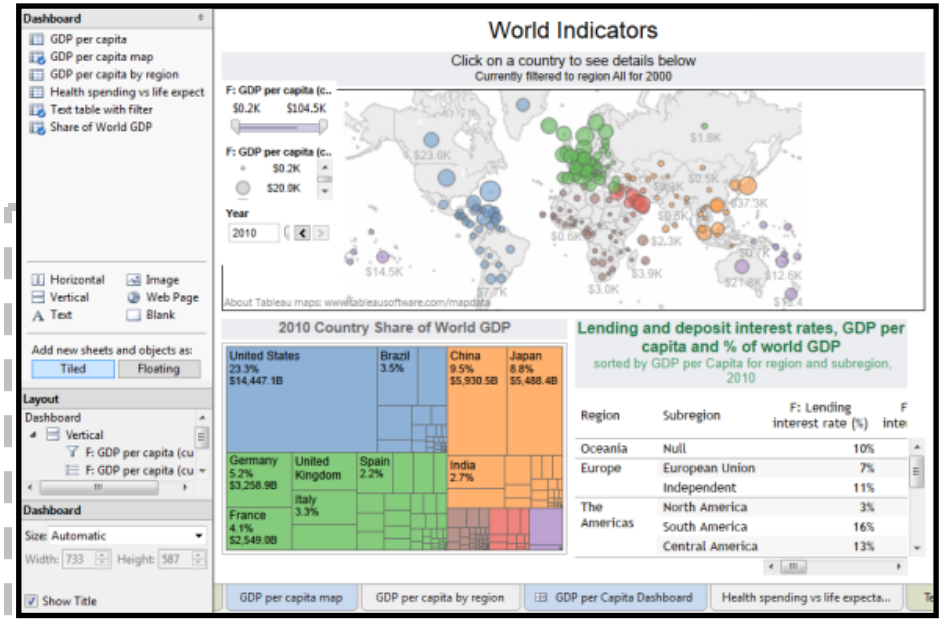
https://www.softwareadvice.com/bi/



Qlik®

http://www.qlik.com

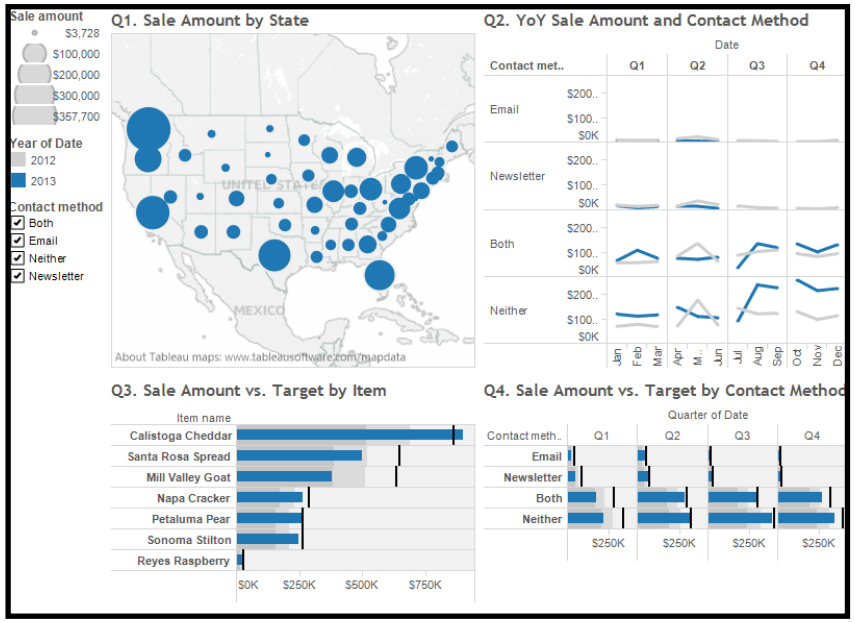

Qlik Sense Desktop
Desktop hub
Sales Manag...
Save
Sales Management & Customer Analysis
Edit
Visualizations to Create
ProductSubcat...
Mountain Bikes
ProductCatego...
Bikes
Visualizations to Create
Total Sales =$31,314.1K
$5,000,000
$2,500,000
$0
$2,490,...
$2,281,2...
$3,144,853
$2,832,352
$2,514,...
$2,900,014
$3,620,597
$3,787,918
2012
2013
Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4
Profit Margin
50.0%
14.0%
Profit Margin
0.0%
100.0%
Sales
* red = most ordered
Mountain-200 Black, 38
Mountain-200 Silver, 38
Mountain-200 Silver, 46
Mountain-200 Black, 42
Mountain-200 Silver, 42
Mountain-200 Black, 46
Sales vs Profit Margin
6000000.00
4000000.00
2000000.00
0.00
Sales
Mountain-200 Black, 38
Mountain-100 Black, 38
-50%
0%
50%
Profit Margin
% of Total Sales
ProductCategoryName ▸ ProductSubcategoryName ▸ P
Others
22.8 %
Mountain-200 Black, 38
14.1 %
Mountain-20...
12.8 %
Mountain-1...
Mountain-1...
Mountain-1...
Mountain-200 ...
10.6 %
11.8 %
Mountain-200 ...
11 %
11 %
Mountain-200 Silver, 46
Mountain-200 Silver, 42
Sales and Profit Margin by Year-Month
2000000.00
1000000.00
0.00
Sales
200%
0%
-200%
Profit Margin
Jan-2012
Feb-2012
Mar-2012
Apr-2012
May-2012
Jun-2012
Jul-2012
Aug-2012
YearMonth
Qlik

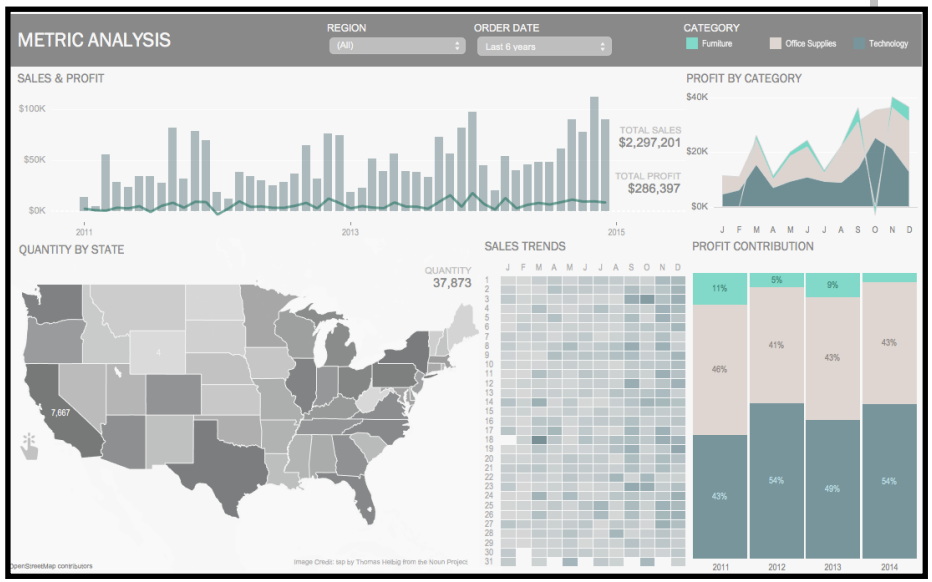


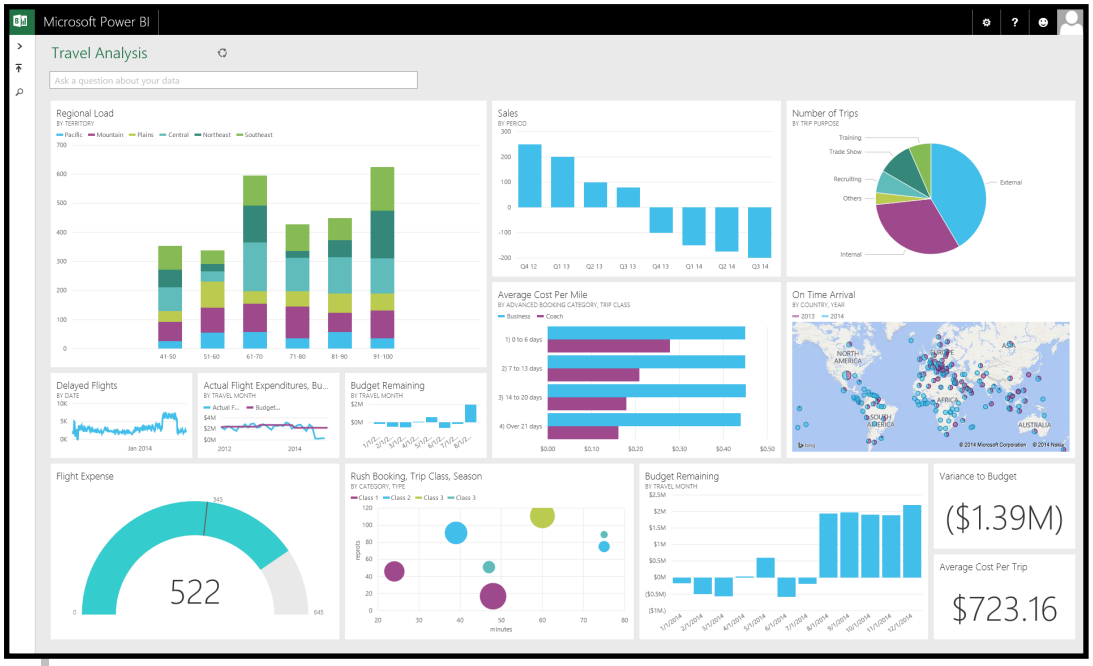
+ a b l e a u®

http://www.tableau.com

http://www.dataprix.com/blog-it/clearpeaks/obten-maximo-rendimiento-tableau-clearpeaks

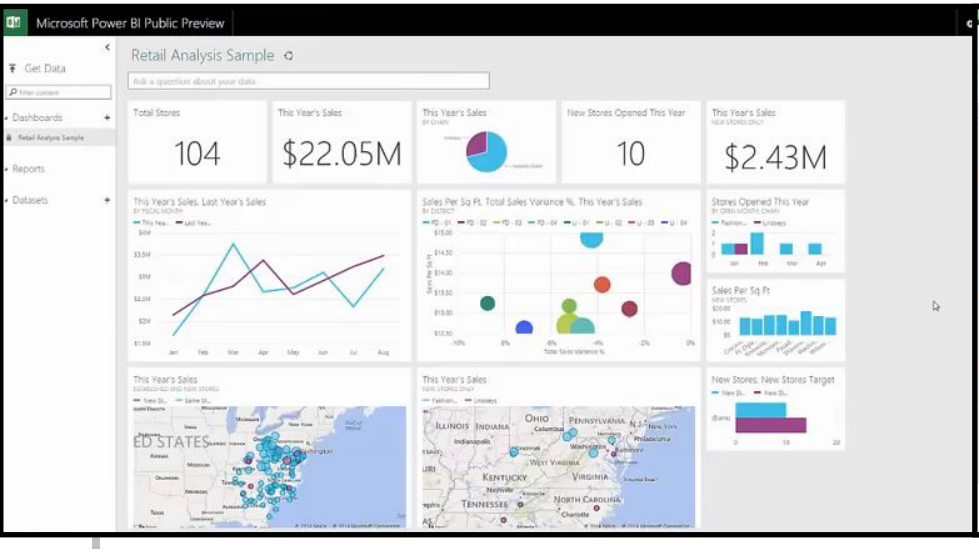


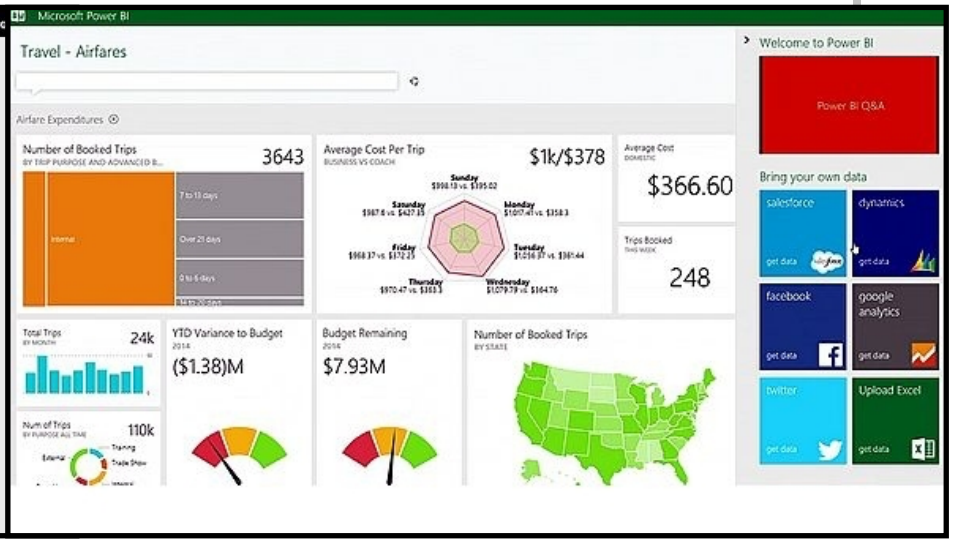




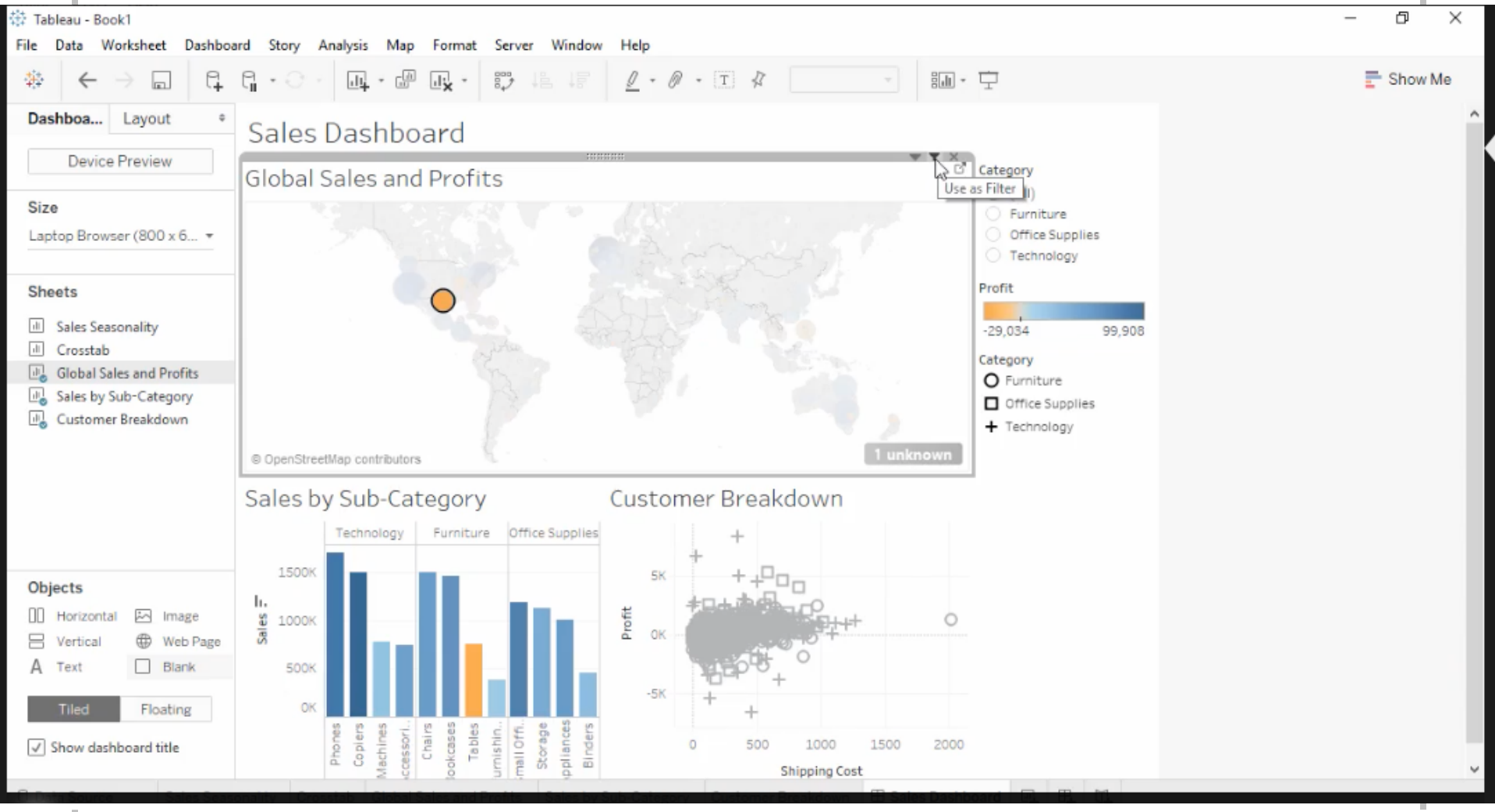
Microsoft Power BI

Power BI





Dashboard: Global Sales & Profit



libSVM

Scikit-Learn
Machine Learning con Python

Dr. Ing. Rodrigo Salas Fuentes
(rod.salas@gmail.com)

Scikit-Learn