



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

DOMAIN KNOWLEDGE & SOFT SKILLS

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



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

COMMUNICATION & VISUALIZATION

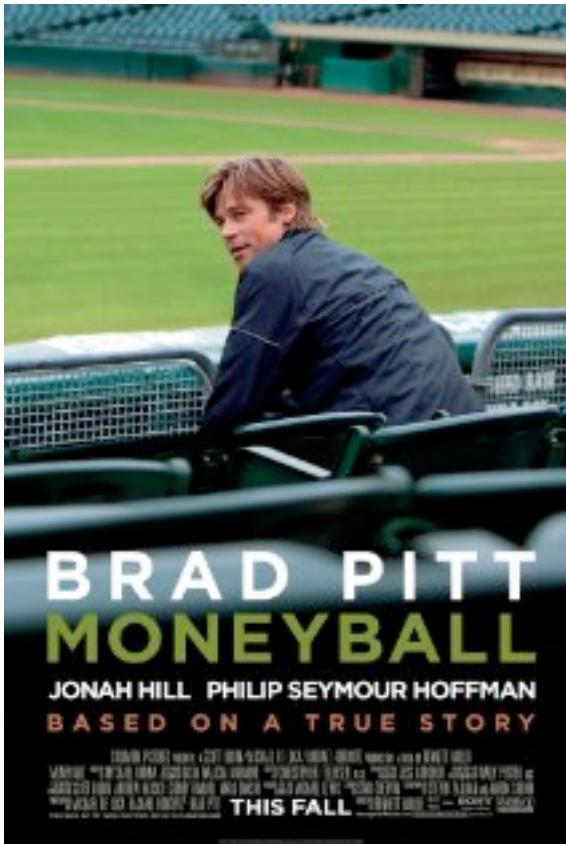
- ★ Able to engage with senior management
- ★ Story telling skills
- ★ Translate data-driven insights into decisions and actions
- ★ Visual art design
- ★ R packages like ggplot or lattice
- ★ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau



EVERYTHING is a Recommendation

The screenshot shows the Netflix homepage. At the top, there's a navigation bar with links for 'Watch Instantly', 'Just for Kids', 'Taste Profile', and 'DVDs'. A search bar is also present. Below the navigation, the 'Recently Watched' section displays thumbnails for 'BETTER OFF TED', 'ARCHER', 'MAD MEN', 'DOCTOR WHO', 'ARRESTED DEVELOPMENT', 'BETTER OFF TED', and 'firefly'. Underneath this, the 'Top 10 for Michael' section shows thumbnails for 'COMEDY BINE BANG', 'SUPERNATURAL', 'SPACED', 'DR. HORRIBLE'S AND ALONG BLOD', 'ALPHAS', and a detailed view of the 'ALPHAS' series page. The 'ALPHAS' page includes a summary, cast information (starring David Strathairn, Ryan Cartwright), and a 'Creators' section. At the bottom of the page, there are social sharing buttons and a large 'NETFLIX' logo.

A news article from The Huffington Post by Dino Grandoni. The headline reads: 'Netflix's New 'My List' Feature Knows You Better Than You Know Yourself (Because Algorithms)'. The article discusses how the 'My List' feature uses algorithms to recommend content. It includes a photo of Iron Man's lab and a photo of the Netflix headquarters building. Below the article, there are social sharing buttons for Facebook, Twitter, and Google+, and a comment section.



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 gerente general Billy Blane y DePodesta.

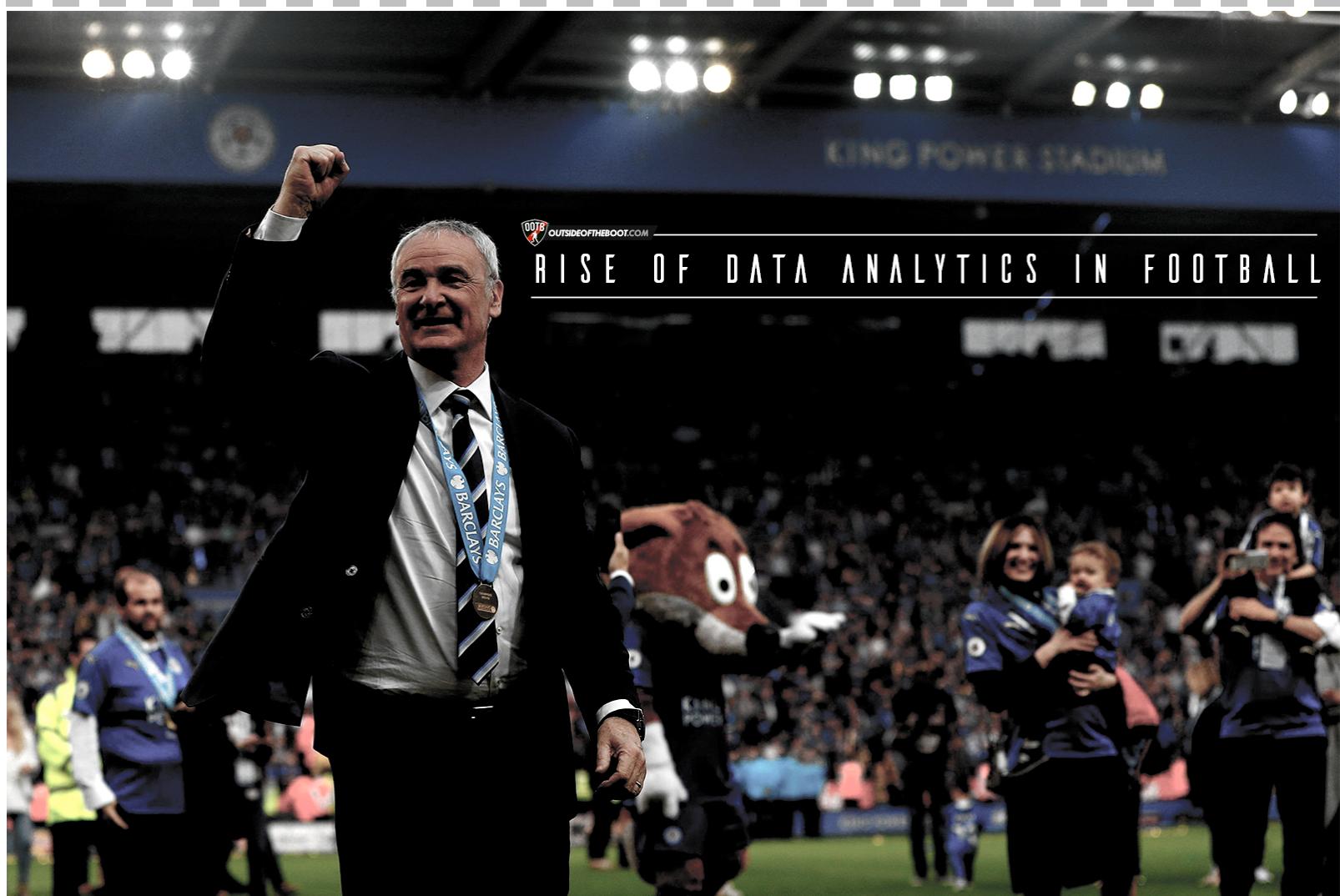
Se vio reflejado el análisis del desempeño en el baseball y las lecciones aprendidas que pueden aplicarse a las organizaciones dirigidas por los datos.

“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.”

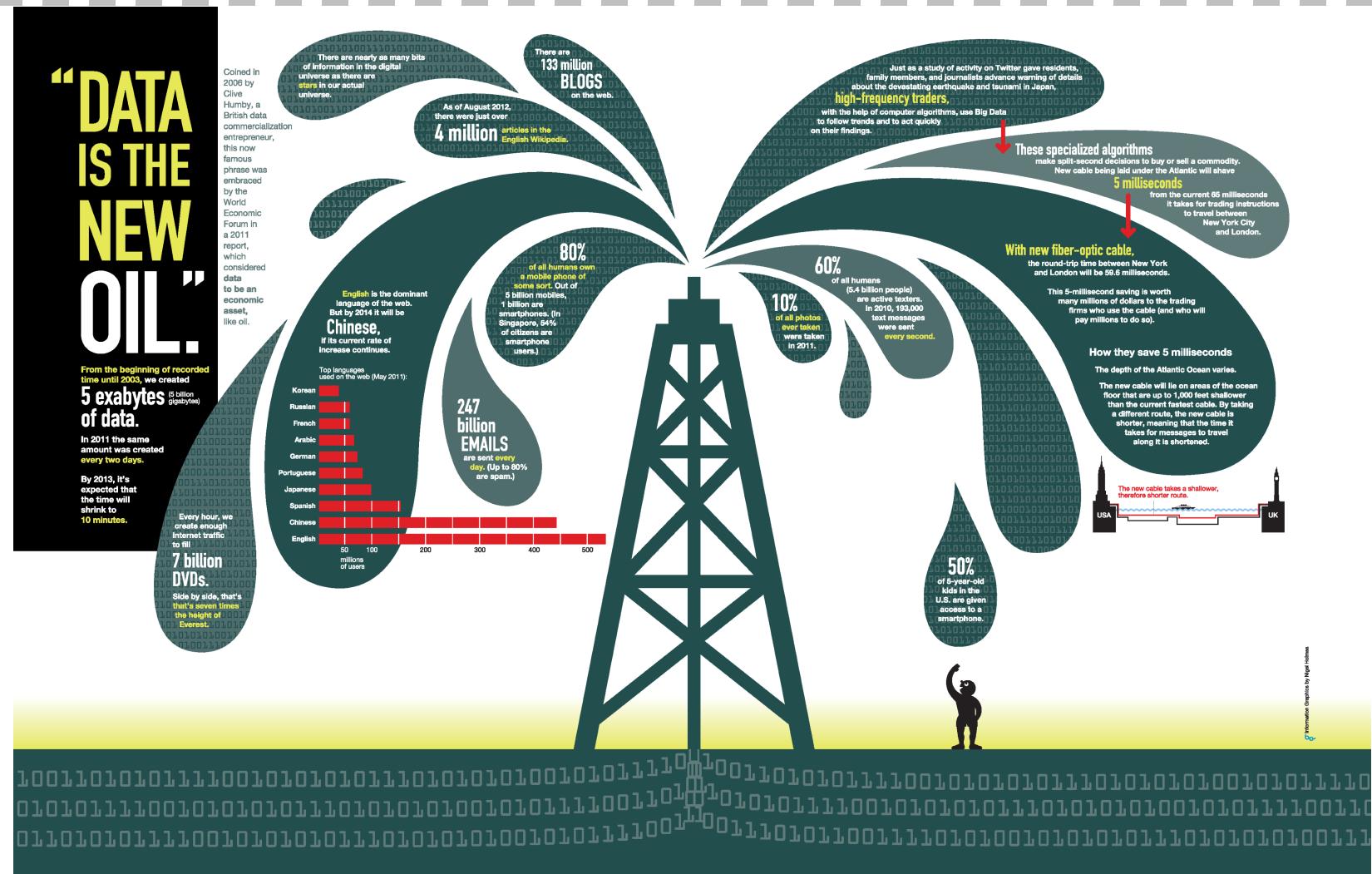


PAUL DePODESTA

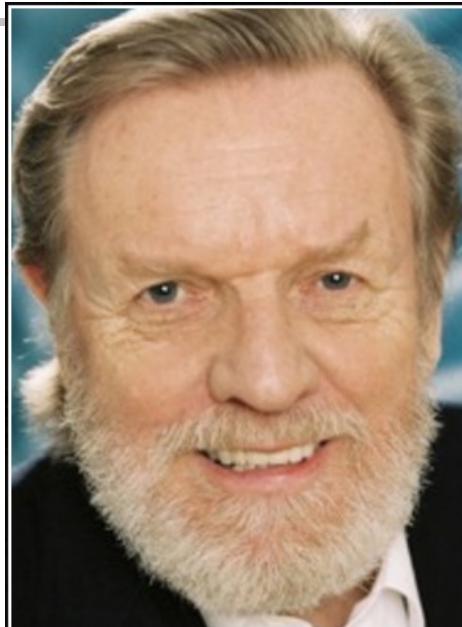
Rise of Data Analytics in Football



- III Simposio de Data Analytics — Dr. Ing. Rodrigo Salas Fuentes (rodrigo.salas@uv.cl)
<http://outsideoftheboot.com/2016/07/22/rise-of-data-analytics-in-football-3/>



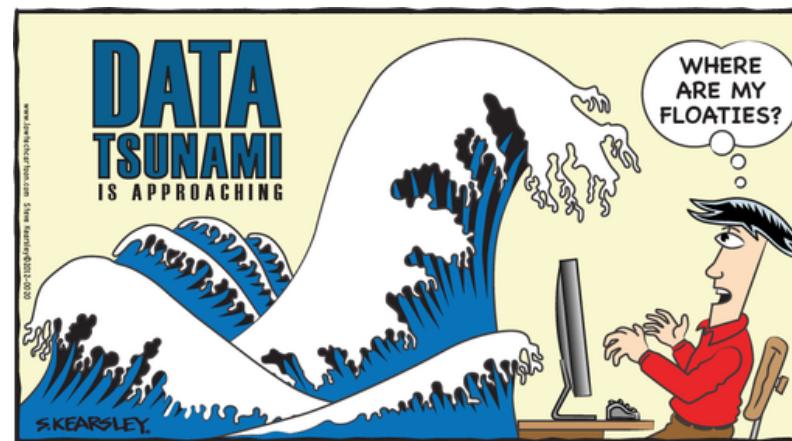
Desafíos en la Exploración de los Datos



We are drowning in information but starved for knowledge.

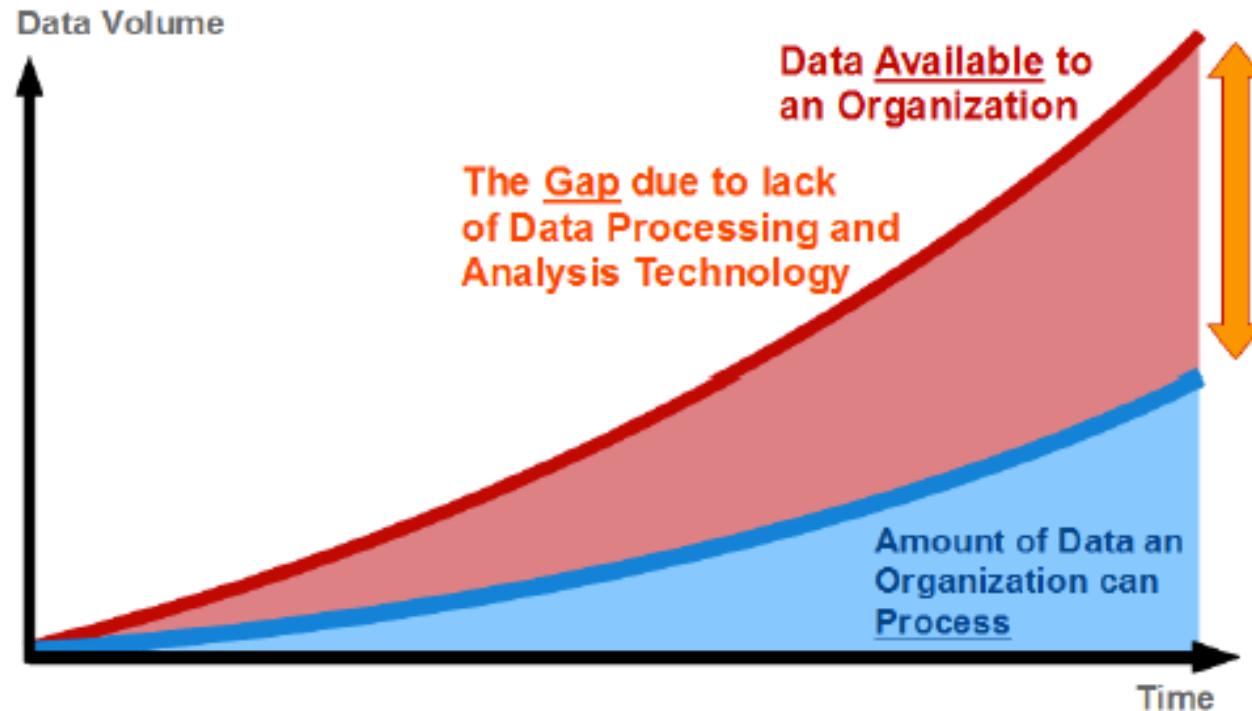
— John Naisbitt —

AZ QUOTES



**THE WORLD'S INFORMATION IS DOUBLING
EVERY TWO YEARS, with a colossal 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?

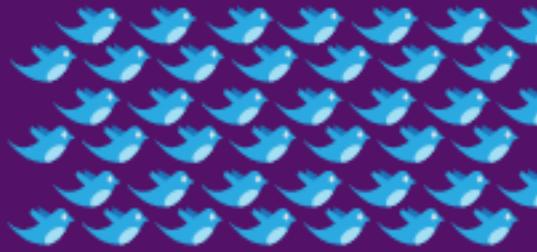
In terms of sheer volume, **1.8 ZB** of data is equivalent to:

Every person in the
United States 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

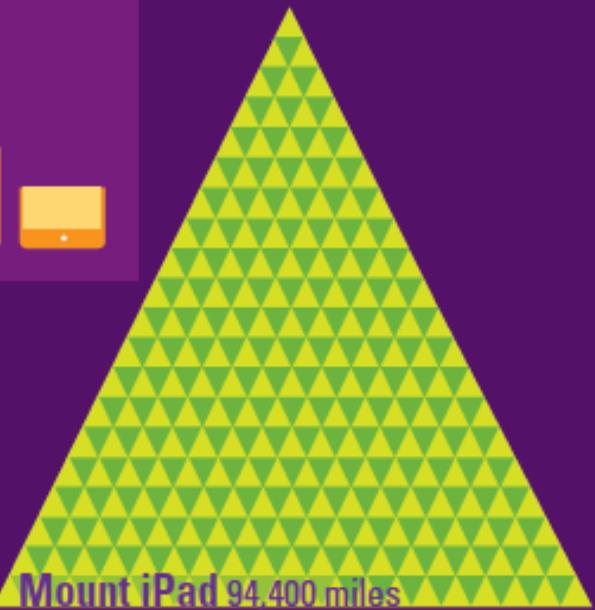


With that many iPads we could build
a mountain of iPads that is

**25-times higher than
Mount Fuji**

Mount Fuji 3,776 miles ▲

Mount iPad 94,400 miles





TBB será el
universo
digital del
futuro

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

10^{24} B

Yottabyte

10^{27} B
Brontobyte

10^{21} B
Zettabyte

10^{18} B

Exabyte

10^{15} B

Petabyte

10^{12} B

Terabyte

10^9 B

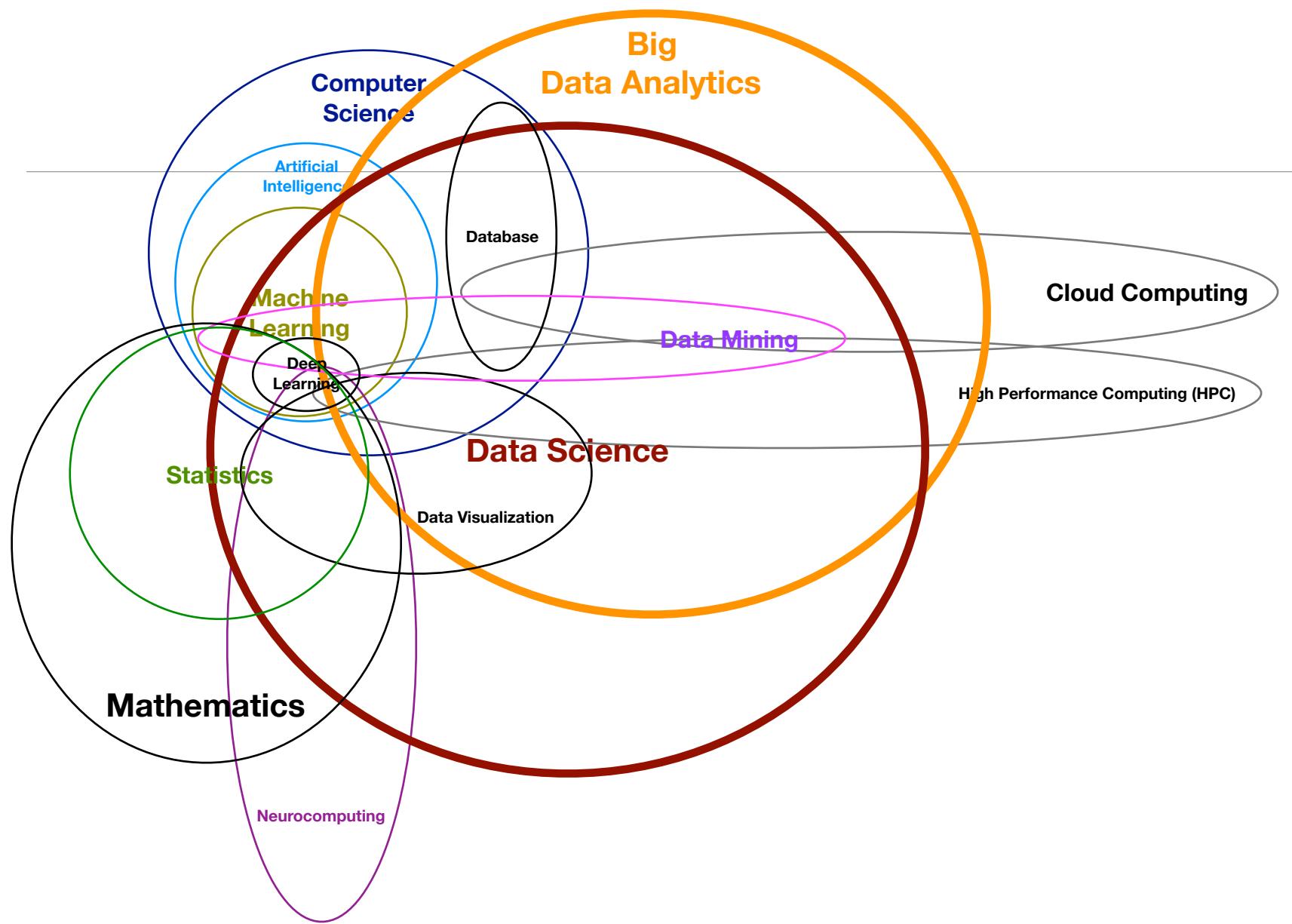
Gigabyte

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

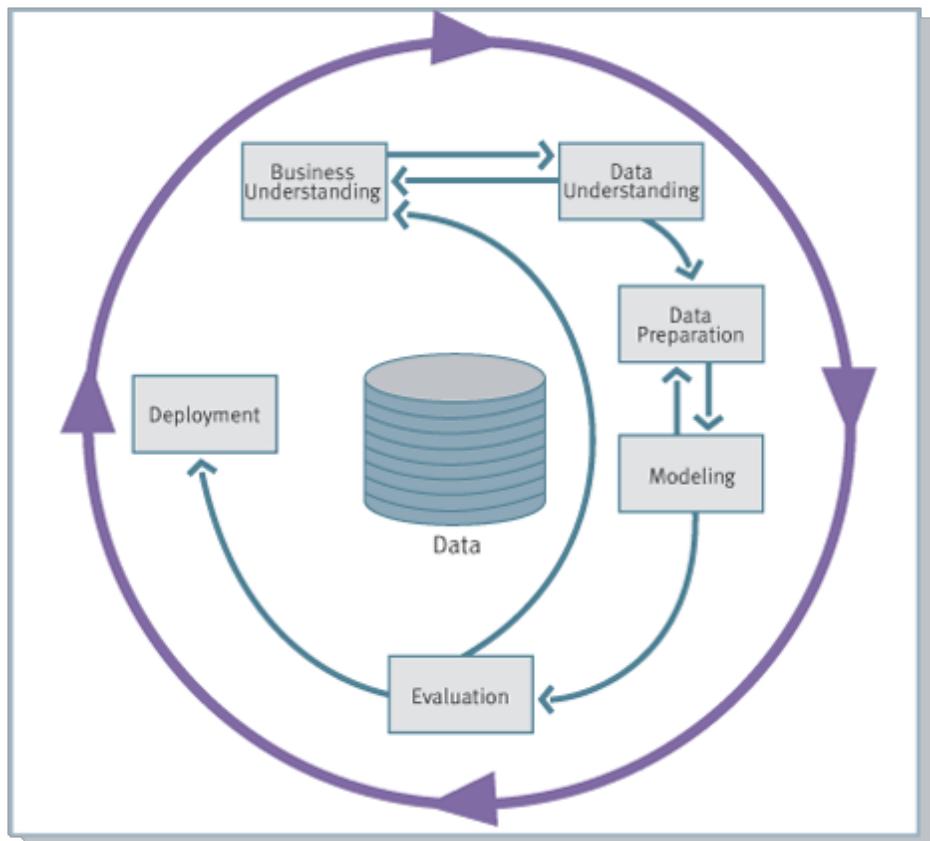
10^6 B
Megabyte

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

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



- ▶ 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 ropa 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>



► **"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.

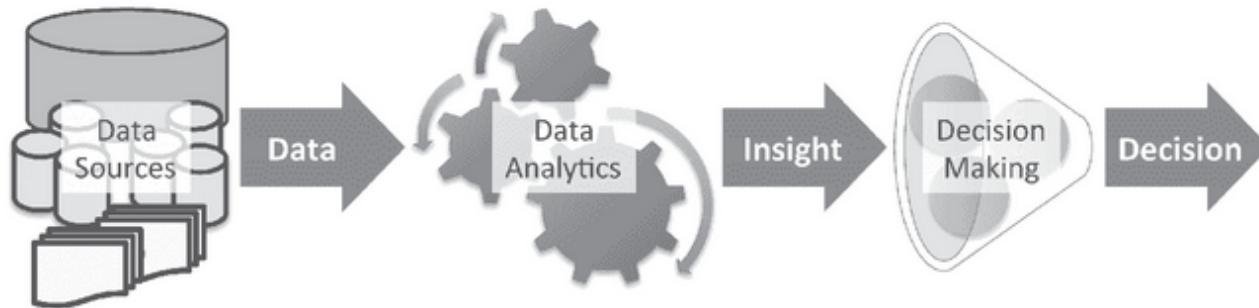


Figure 1.1

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

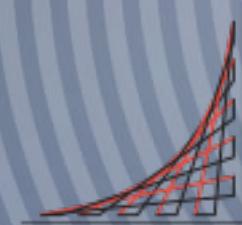
- ▶ Las organizaciones modernas recogen grandes cantidades de datos. Para que los datos sean valiosos para una organización, deben analizarse para extraer información (**insights**) que se pueda utilizar para tomar mejores decisiones.
 - ▶ La extracción de información de los datos es el trabajo de análisis de datos. Este libro se enfoca en el análisis de datos predictivos, que es un sub campo importante de análisis de datos
- ▶ **III Simposio de Data Analytics — Dr. Ing. Rodrigo Salas Fuentes (rodrigo.salas@uv.cl)**

- ▶ La **Inteligencia de Negocios** se define como 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,Question8
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,Opinio,0
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
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,"hey"
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,0,0,1,1,1,1,1,1,0,0
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,0,1,0,0,0,1,0,0,0,0
27367,2006.11.27 15:8,2006.11.27 15:8,en,Norway,Education,<2 - 5 years,5001-10000,6,0,6,"f6{[]}+âme' ''*-*/+",1,1,1,1,1,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,"f6{[]}+âme' ''*-*/+",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
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,0,1,0,0,0,0,1,0,,,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,"f6{[]}+âme' ''*-*/+",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,"f6{[]}+âme' ''*-*/+",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
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
```



“A PICTURE IS WORTH A THOUSAND WORDS.”

NAPOLEON BONAPARTE

CUSTOMIZE & SHARE

Share Quote

Like 0 Pin It +1

Read more quotes by Napoleon Bonaparte

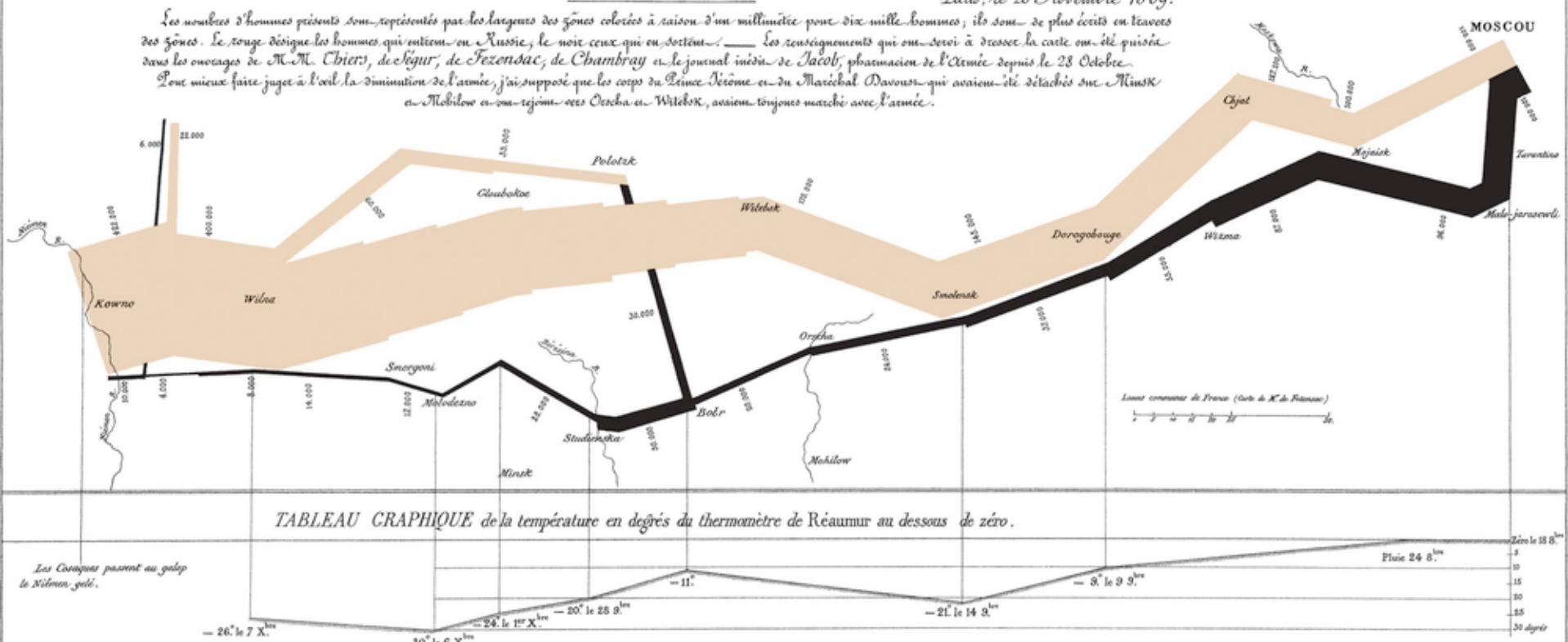


Visualización de la campaña de Napoleón a Rusia, 1812

Carte Figurative des pertes successives en hommes de l'Armée Française dans la Campagne de Russie 1812-1813.
Dessiné par M. Minard, Ingénieur Général des Ponts et Chaussées en retraite. Paris, le 20 Novembre 1869.

Les nombres d'hommes perdus sont représentés par les largures des zones colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en lettres des zones. Le rouge désigne les hommes qui entrent en Russie; le noir ceux qui en sortent. — Les renseignements qui ont servi à dresser la carte ont été prisés dans les ouvrages de M. M. Chiers, de Ségur, de Fezensac, de Chambray et le journal intérieur de Jacob, pharmacien de l'Armée depuis le 28 Octobre.

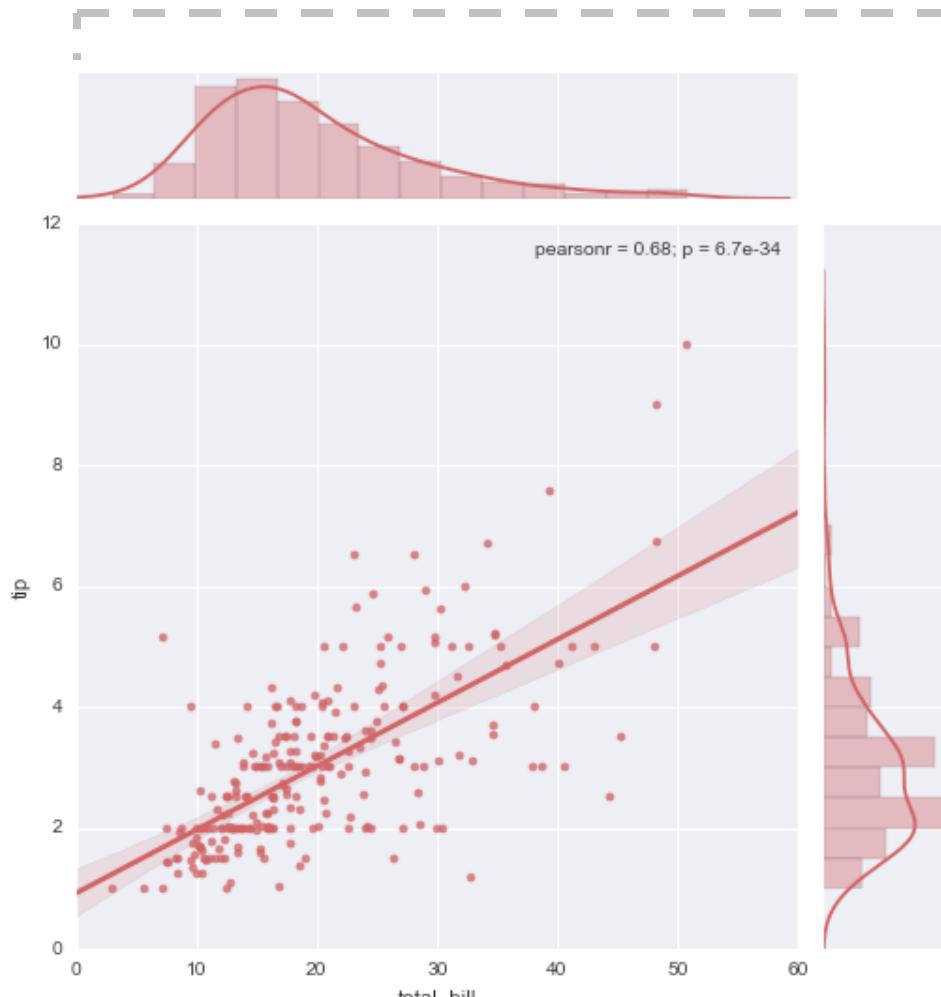
Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davout, qui avaient été détachés sur Minsk et Malibow et se rejoignirent vers Orelia et Whiteck, avaient toujours marché avec l'armée.



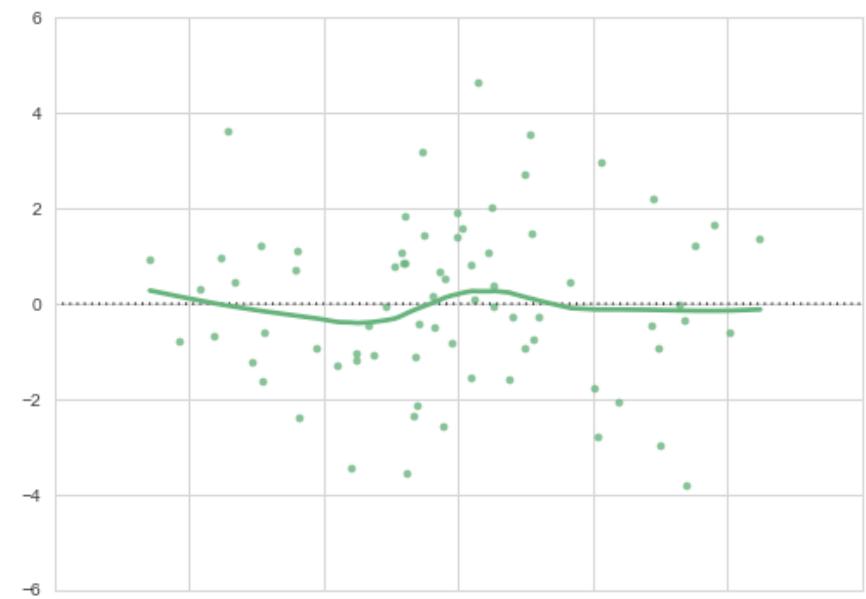
Avec la permission de l'Éditeur, E. Pau, 3^e Maré, 3^e Gén. à Paris.

Imp. Litt. Bayard et Deshayes.

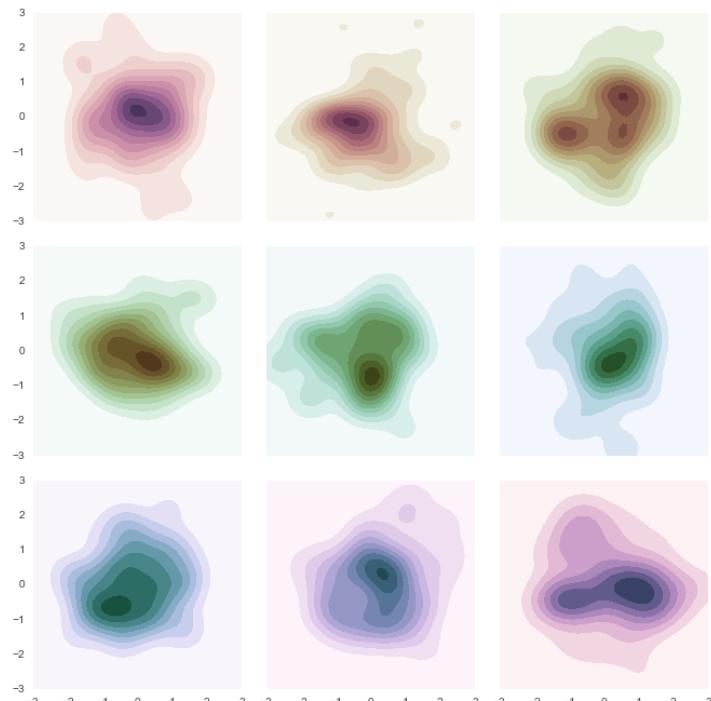
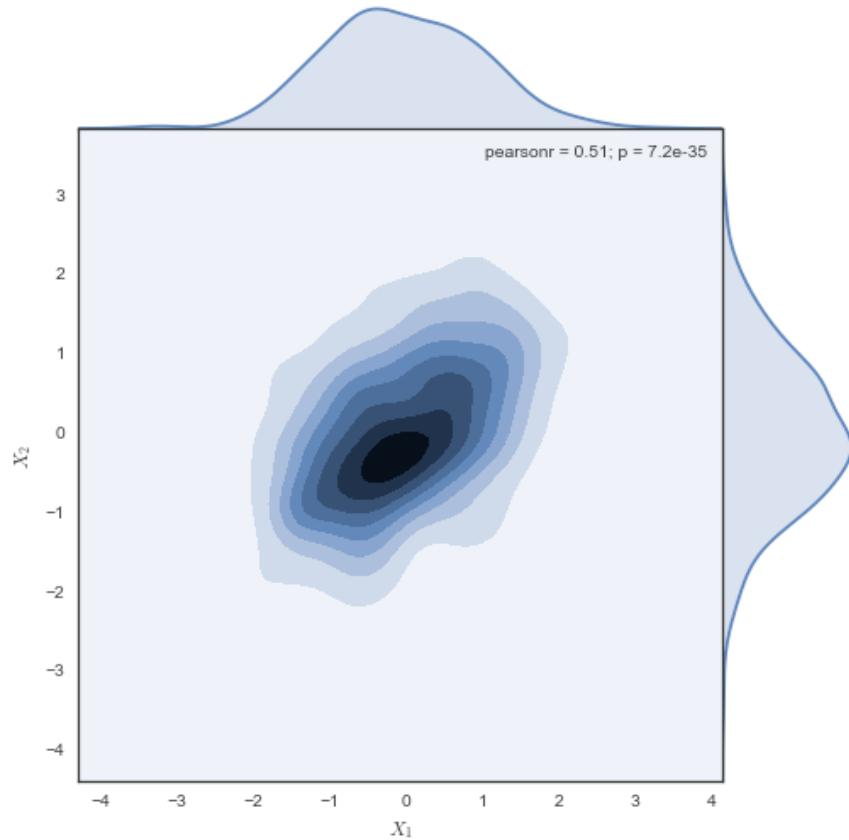




https://web.stanford.edu/~mwaskom/software/seaborn/examples/regression_marginals.html



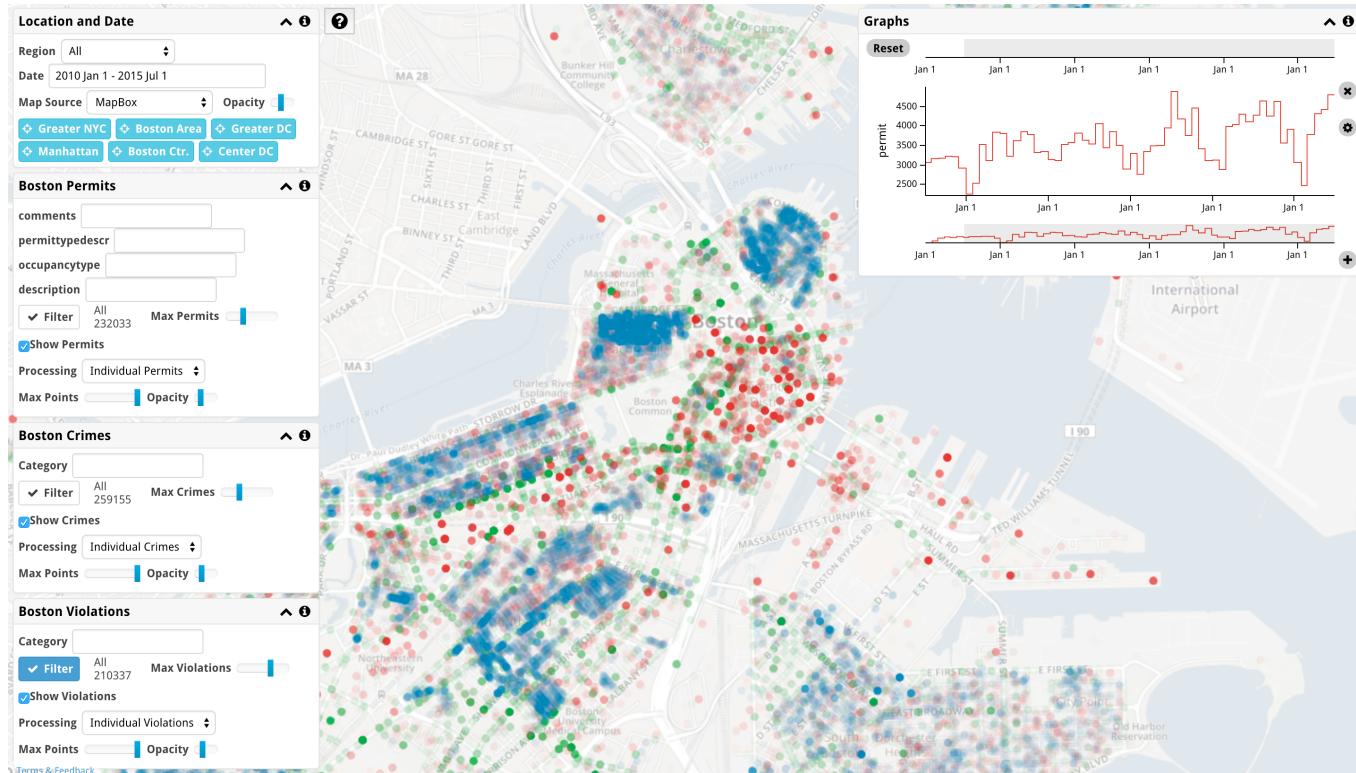
<https://web.stanford.edu/~mwaskom/software/seaborn/examples/residplot.html>

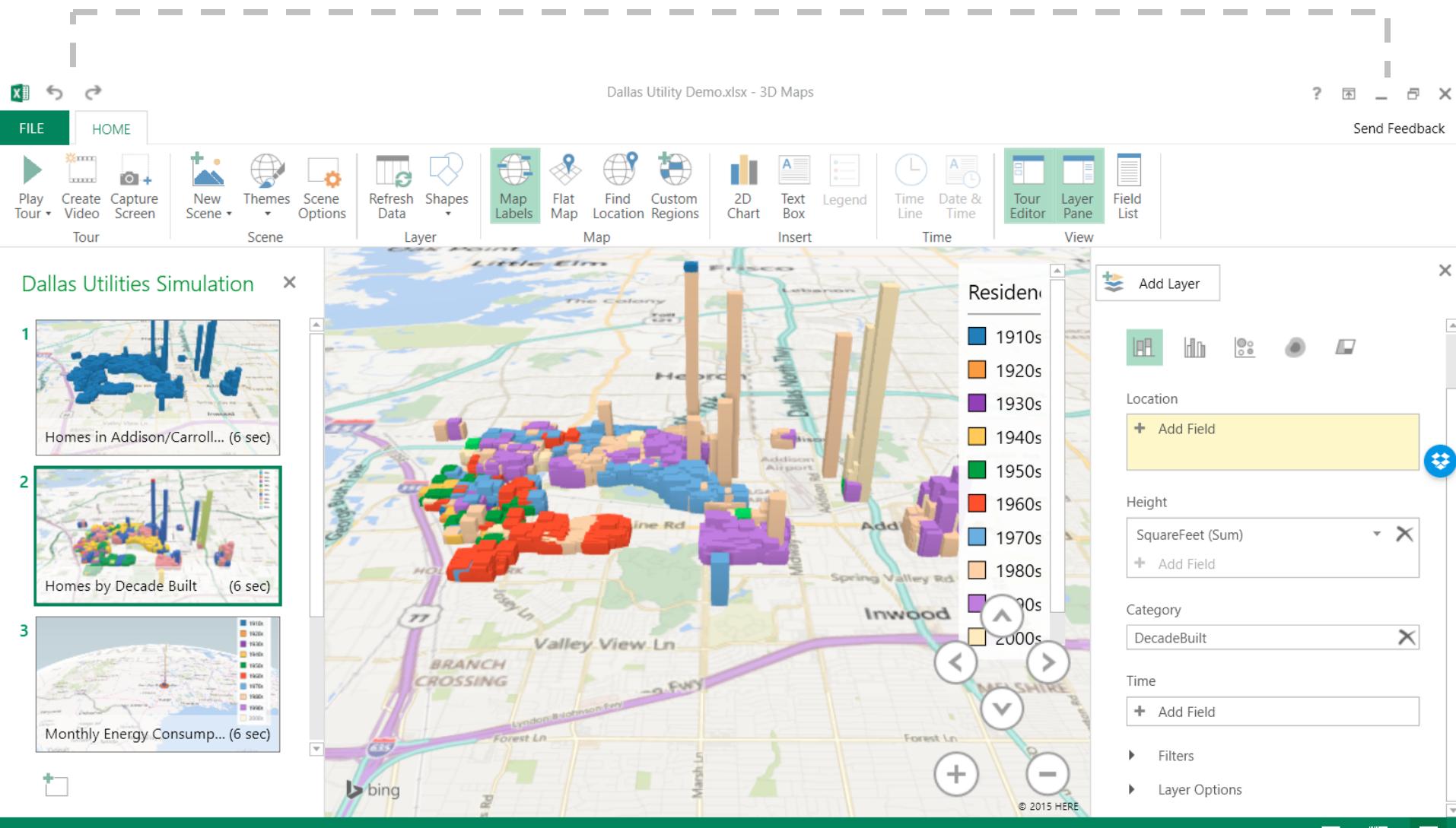


https://web.stanford.edu/~mwaskom/software/seaborn/examples/joint_kde.html



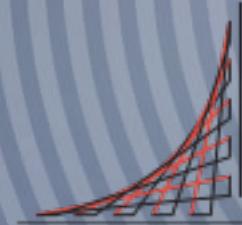
Boston Census Data







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

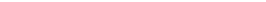
BIG DATA LANDSCAPE, VERSION 3.0

Exited: Acquisition or IPO

Infrastructure



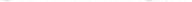
Open Source



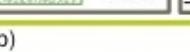
Analytics



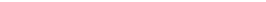
For Business



Applications



Data Sources

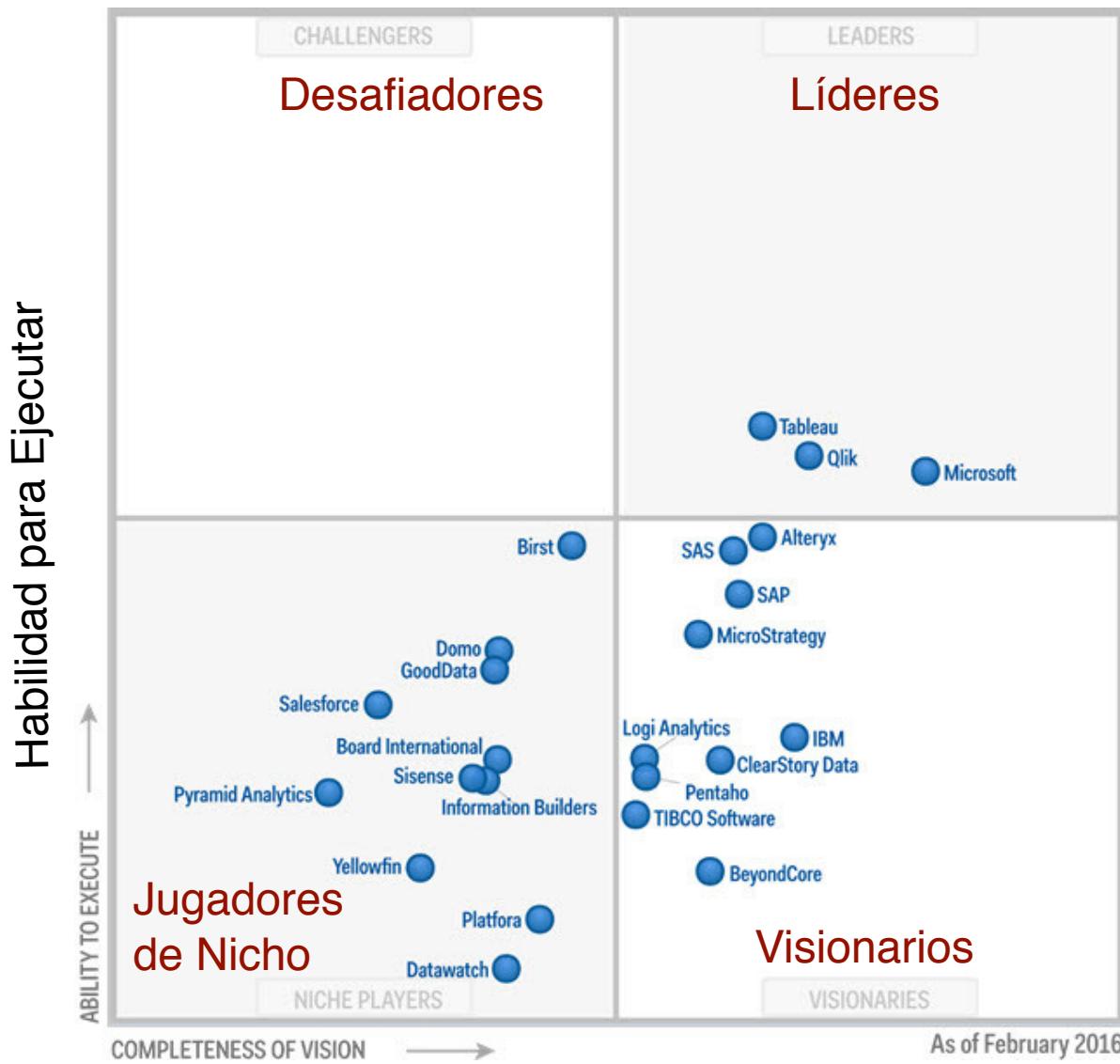




Business Intelligence Platforms



Cuadrante de Gartner - Herramientas BI 2016



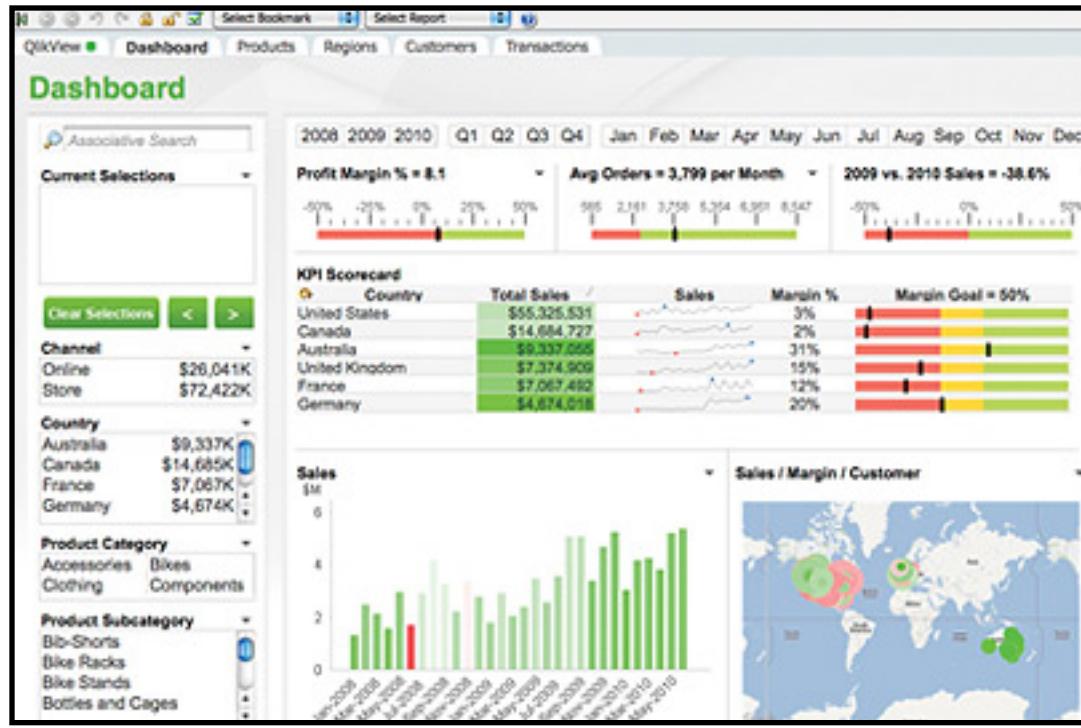
Febrero 2016

"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

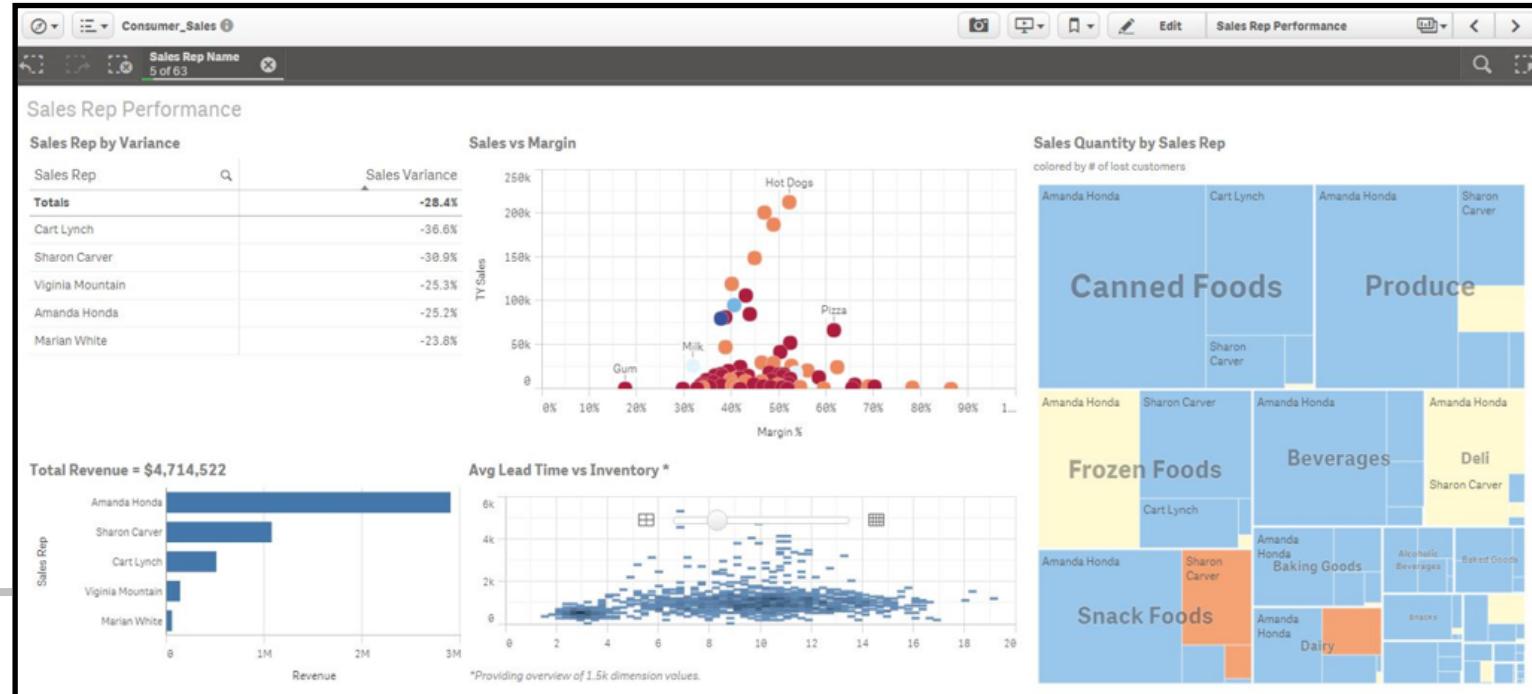


<https://www.softwareadvice.com/bi/>



Qlik

<http://www.qlik.com>

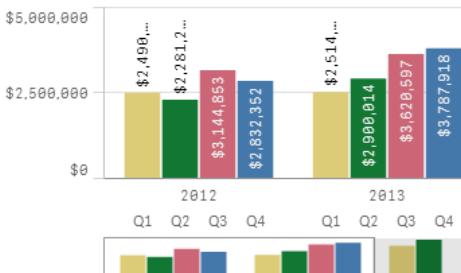


Qlik Sense Desktop

Sales Management & Customer Analysis

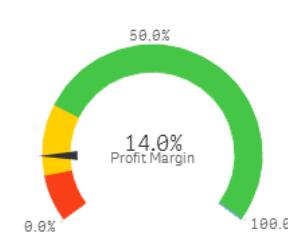
Visualizations to Create

Total Sales = \$31,314.1K



Year-Month	Sales
2012 Q1	\$2,498,125
2012 Q2	\$2,261,233
2012 Q3	\$3,144,853
2012 Q4	\$2,832,352
2013 Q1	\$2,514,104
2013 Q2	\$2,908,014
2013 Q3	\$3,620,597
2013 Q4	\$3,787,918

Profit Margin



Profit Margin Range	Percentage
0.0% - 14.0%	Red (most ordered)
14.0% - 50.0%	Yellow
50.0% - 100.0%	Green

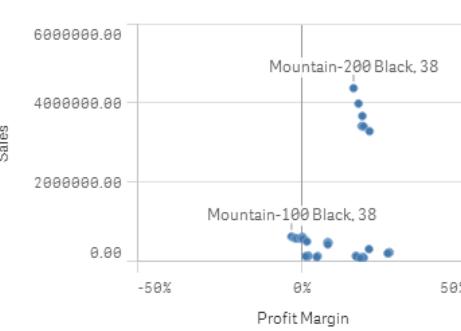
Sales

* red = mostordered



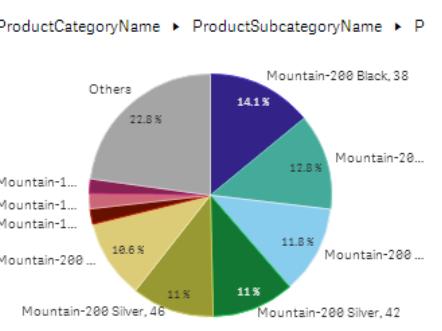
Product Category	Color	Value
Mountain-200 Black	Red	38
Mountain-200 Silver	Orange	38
Mountain-200 Silver	Blue	46
Mountain-200 Black	Red	42
Mountain-200 Silver	Orange	42
Mountain-200 Black	Blue	46

Sales vs Profit Margin



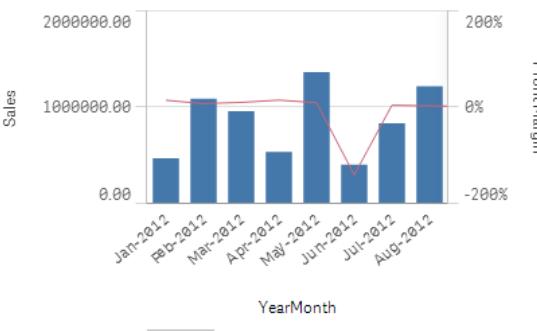
Product Category	Color	Value
Mountain-200 Black	Red	38
Mountain-100 Black	Blue	38

% of Total Sales

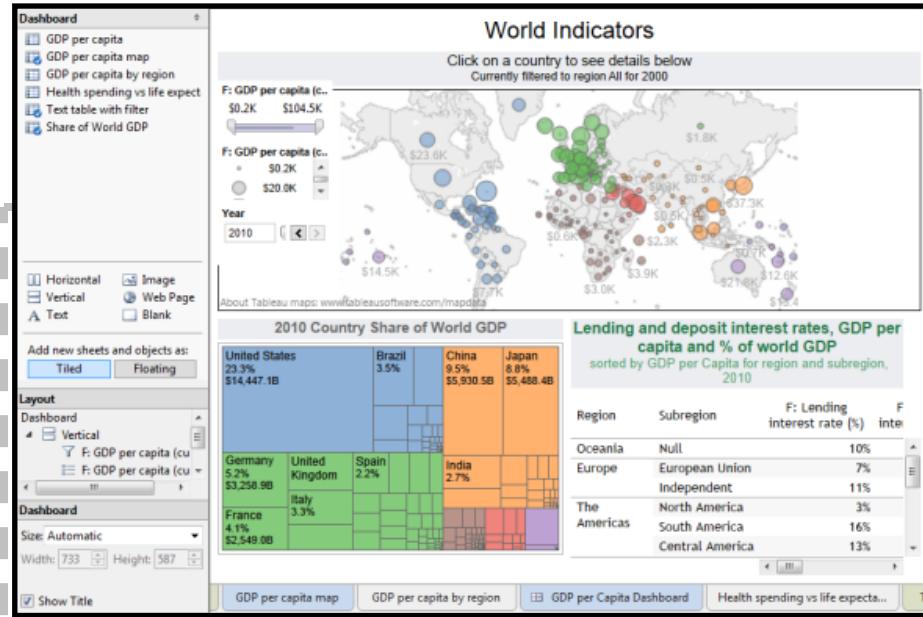


Product Category	Percentage
Others	22.8%
Mountain-200 Black, 38	14.1%
Mountain-200 Silver, 38	12.8%
Mountain-200 Silver, 46	11.8%
Mountain-200 Black, 46	11%
Mountain-100 Black, 38	10.6%
Mountain-100 Silver, 42	2.1%
Mountain-100 Silver, 46	1.1%

Sales and Profit Margin by Year-Month

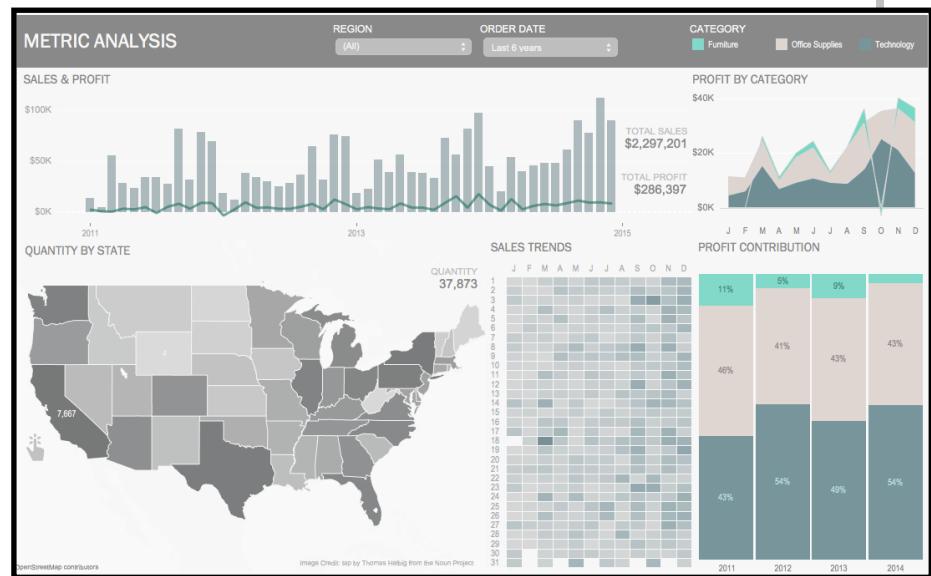
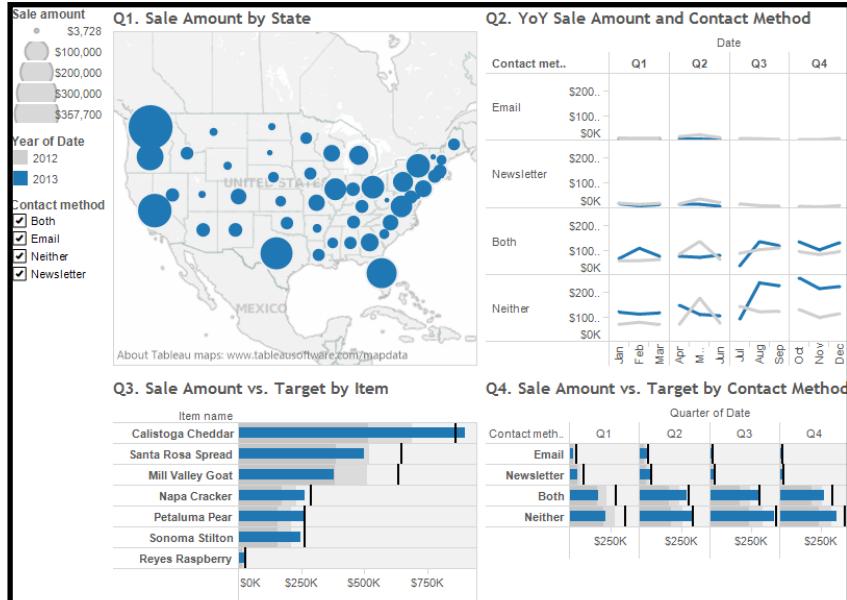


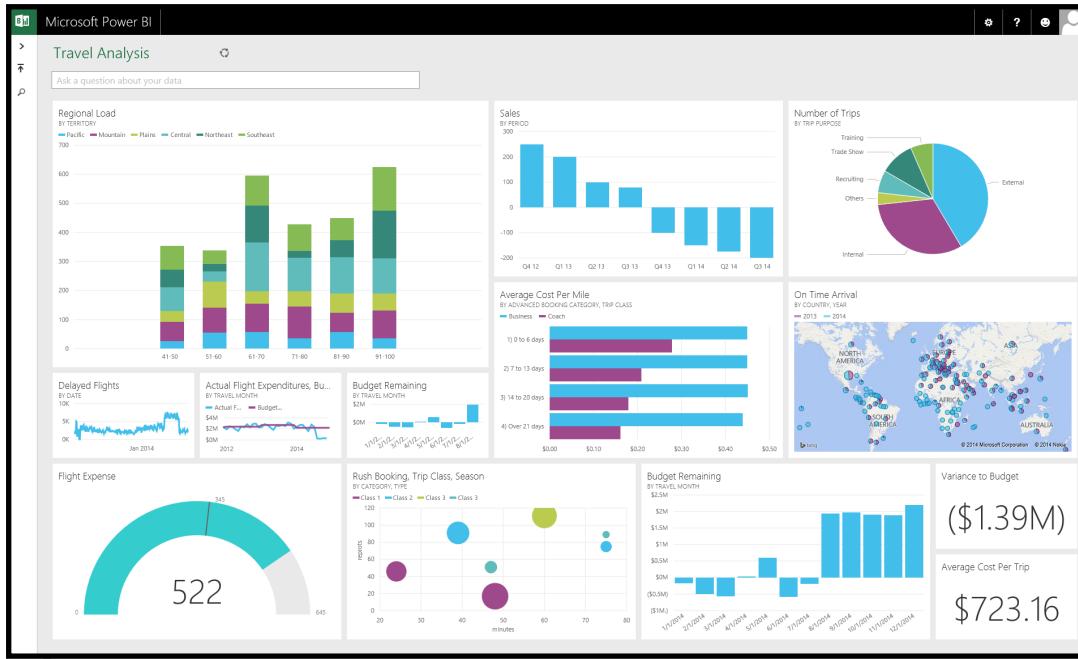
Year-Month	Sales	Profit Margin
Jan-2012	\$80,000	-10%
Feb-2012	\$100,000	-5%
Mar-2012	\$90,000	0%
Apr-2012	\$70,000	5%
May-2012	\$120,000	10%
Jun-2012	\$50,000	15%
Jul-2012	\$90,000	20%
Aug-2012	\$100,000	25%



<http://www.tableau.com>

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





Microsoft Power BI Public Preview

Get Data

Travel Analysis Sample

Ask a question about your data

Total Stores 104

This Year's Sales \$22.05M

This Year's Sales BY CATEGORIES

New Stores Opened This Year 10

This Year's Sales NEW STORES ONLY \$2.43M

This Year's Sales Last Year's Sales BY FISCAL MONTH

Sales Per Sq Ft Total Sales Variance % This Year's Sales BY CATEGORIES

Stores Opened This Year BY OPEN MONTH COUNTRY

Sales Per Sq Ft BY CATEGORIES

New Stores, New Stores Target BY CATEGORIES

Travel - Airfares

Airfare Expenditures

Number of Booked Trips BY TRIP PURPOSE AND ADVANCED B...

3643

7 to 10 days
Over 20 days
0 to 5 days
Internal

Average Cost Per Trip BUSINESS VS COACH \$1k/\$378

\$366.60

Average Cost DOMESTIC

Trips Booked this week 248

YTD Variance to Budget (\$1.38)M

Budget Remaining 2014 \$7.93M

Number of Booked Trips BY STATE

Power BI Q&A

Welcome to Power BI

Bring your own data

- salesforce
- dynamics
- facebook
- google analytics
- twitter
- Upload Excel

Dashboard: Global Sales & Profit

Tableau - Book1

File Data Worksheet Dashboard Story Analysis Map Format Server Window Help

Show Me

Dashboa... Layout

Device Preview

Size
Laptop Browser (800 x 600)

Sheets
Sales Seasonality
Crosstab
Global Sales and Profits
Sales by Sub-Category
Customer Breakdown

Objects
Horizontal Image
Vertical Web Page
Text Blank
Tiled Floating

Show dashboard title

Sales Dashboard

Global Sales and Profits

Map showing global sales and profits. A legend indicates Category (Furniture, Office Supplies, Technology) and Profit (-29,034 to 99,908). A tooltip for a point in South America shows "1 unknown".

Sales by Sub-Category

Bar chart showing sales by sub-category across three categories: Technology, Furniture, and Office Supplies.

Sub-Category	Technology	Furniture	Office Supplies
Phones	~1600K	~1500K	~1500K
Copiers	~1500K	~1500K	~1500K
Machines	~800K	~800K	~800K
Accessori...	~700K	~700K	~700K
Chairs	~1500K	~1500K	~1500K
Lookcases	~1500K	~1500K	~1500K
Tables	~800K	~800K	~800K
Urnishin...	~400K	~400K	~400K
Mail Off.	~1200K	~1200K	~1200K
Storage	~1000K	~1000K	~1000K
Appliances	~1000K	~1000K	~1000K
Binders	~400K	~400K	~400K

Customer Breakdown

Scatter plot showing Profit vs. Shipping Cost. Data points are categorized by Category (Furniture, Office Supplies, Technology).

Profit

Shipping Cost



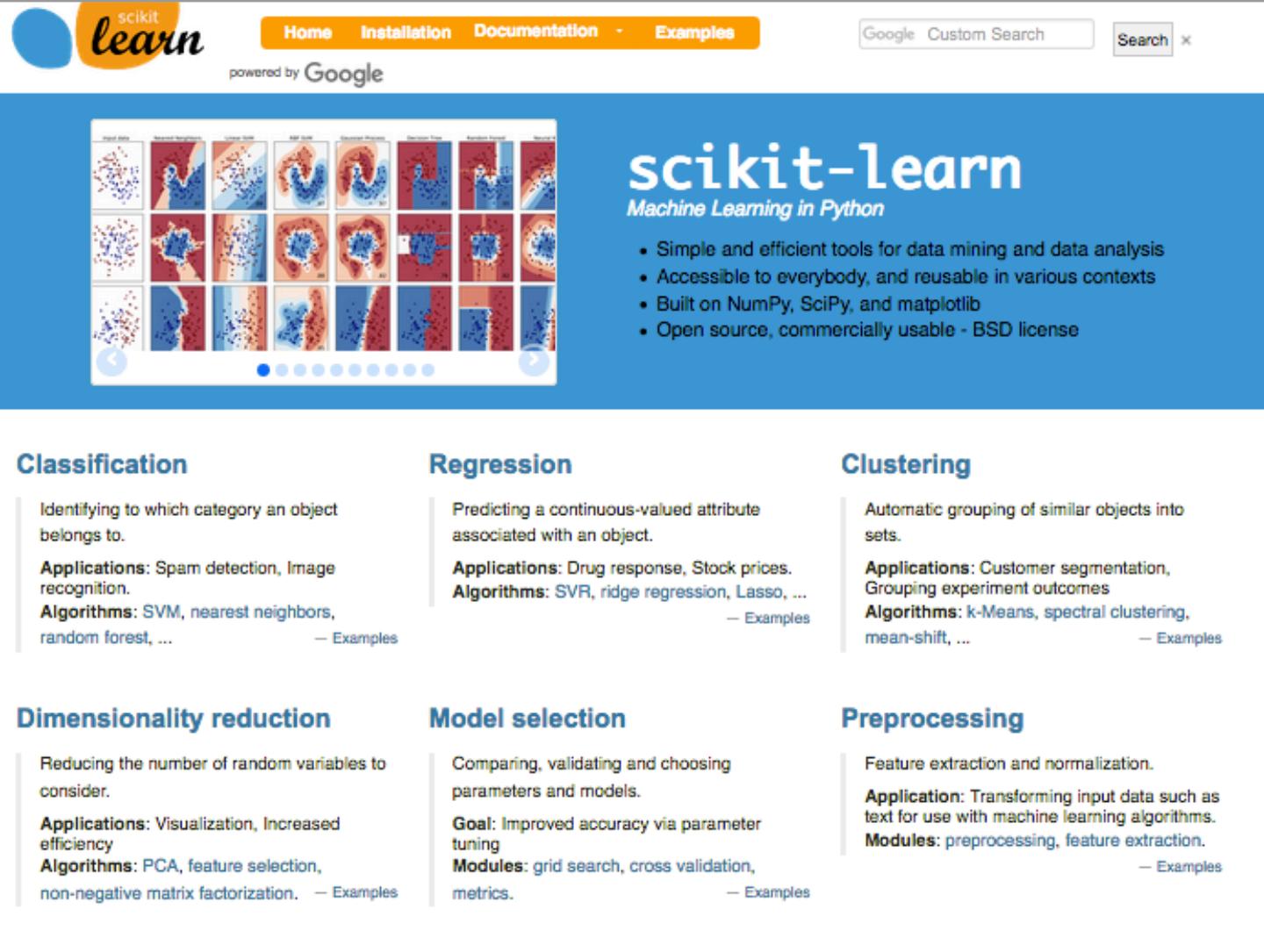
libSVM



Scikit-Learn
Machine Learning con Python

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

Scikit-Learn



The screenshot shows the official scikit-learn website. At the top, there's a navigation bar with links for Home, Installation, Documentation, Examples, Google Custom Search, and a search bar. Below the navigation is a banner featuring a grid of nine small plots illustrating various machine learning concepts like classification and regression. The main title "scikit-learn" is prominently displayed in large white letters on a blue background, with the subtitle "Machine Learning in Python" underneath. To the right of the title is a bulleted list of features:

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

The page is organized into several sections with sub-sections and examples:

- Classification**: Identifying to which category an object belongs to. Applications: Spam detection, Image recognition. Algorithms: SVM, nearest neighbors, random forest, ... [— Examples](#)
- Regression**: Predicting a continuous-valued attribute associated with an object. Applications: Drug response, Stock prices. Algorithms: SVR, ridge regression, Lasso, ... [— Examples](#)
- Clustering**: Automatic grouping of similar objects into sets. Applications: Customer segmentation, Grouping experiment outcomes. Algorithms: k-Means, spectral clustering, mean-shift, ... [— Examples](#)
- Dimensionality reduction**: Reducing the number of random variables to consider. Applications: Visualization, Increased efficiency. Algorithms: PCA, feature selection, non-negative matrix factorization. [— Examples](#)
- Model selection**: Comparing, validating and choosing parameters and models. Goal: Improved accuracy via parameter tuning. Modules: grid search, cross validation, metrics. [— Examples](#)
- Preprocessing**: Feature extraction and normalization. Application: Transforming input data such as text for use with machine learning algorithms. Modules: preprocessing, feature extraction. [— Examples](#)