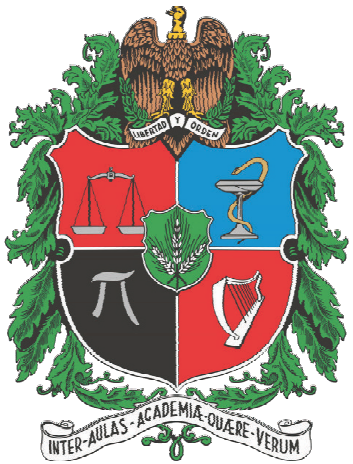
Ciencia de los datos
(Aprendizaje de máquina)

A. M. Alvarez-Meza, Ph.D.
amalvarezme@unal.edu.co

Departamento de ingeniería eléctrica, electrónica y computación
Universidad Nacional de Colombia-sede Manizales



Contenido

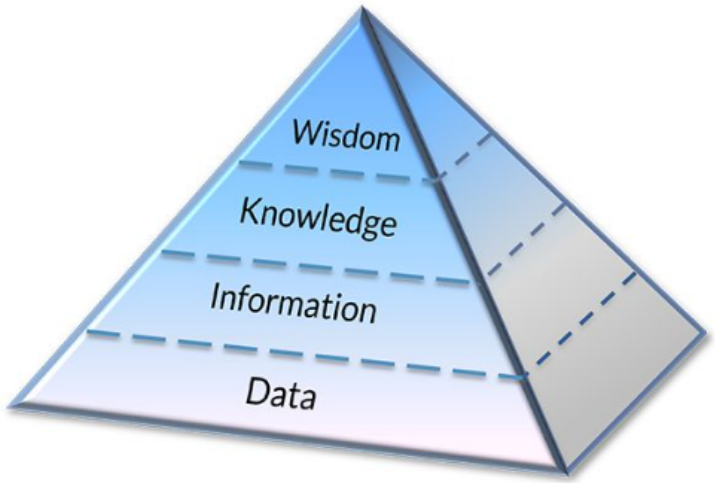
Contenido

Datos $\neq$ Información

- Los datos se pueden encontrar "fácilmente" en todos lados
  - Evolución del precio de las acciones de una empresa en bolsa
  - Estadísticas de resultados deportivos
  - Históricos de consumo de ciertos productos
  - Precios de mercado de bienes y/o servicios
  - ...

- La información, sin embargo, hay que saber cómo y dónde buscarla
  - Normalmente subyace escondida detrás los datos
  - Obtenerla, requiere del procesamiento y del análisis de los datos
  - *Soft information*, *Hard information*

DIKW - *Data, Information, Knowledge and Wisdom I*



- *Data:* Tener las cifras en crudo de un determinado fenómeno
- *Information:* Poder extraer de esas cifras relaciones, dependencias, influencias, causas y posibles consecuencias
- *Knowledge:* Saber cómo hacer frente a la información obtenida
- *Wisdom:* Tener el poder para hacerlo

DIKW - *Data, Information, Knowledge and Wisdom II*

Ejemplo de DIKW - Calentamiento global

- *Data:* Las cifras históricas de la temperatura en el mundo en los últimos cien años
- *Information:* Descubrir que la temperatura global va en aumento
- *Knowledge:* Saber qué estrategias deben seguirse para reducir la producción de gases de efecto invernadero
- *Wisdom:* Tener la capacidad y el poder para implementar acuerdos como el de Kyoto (1997) o el de París (COP21 - 2015)

DIKW - *Data, Information, Knowledge and Wisdom III*

El caso (o mito) de la cerveza y los pañales[1]

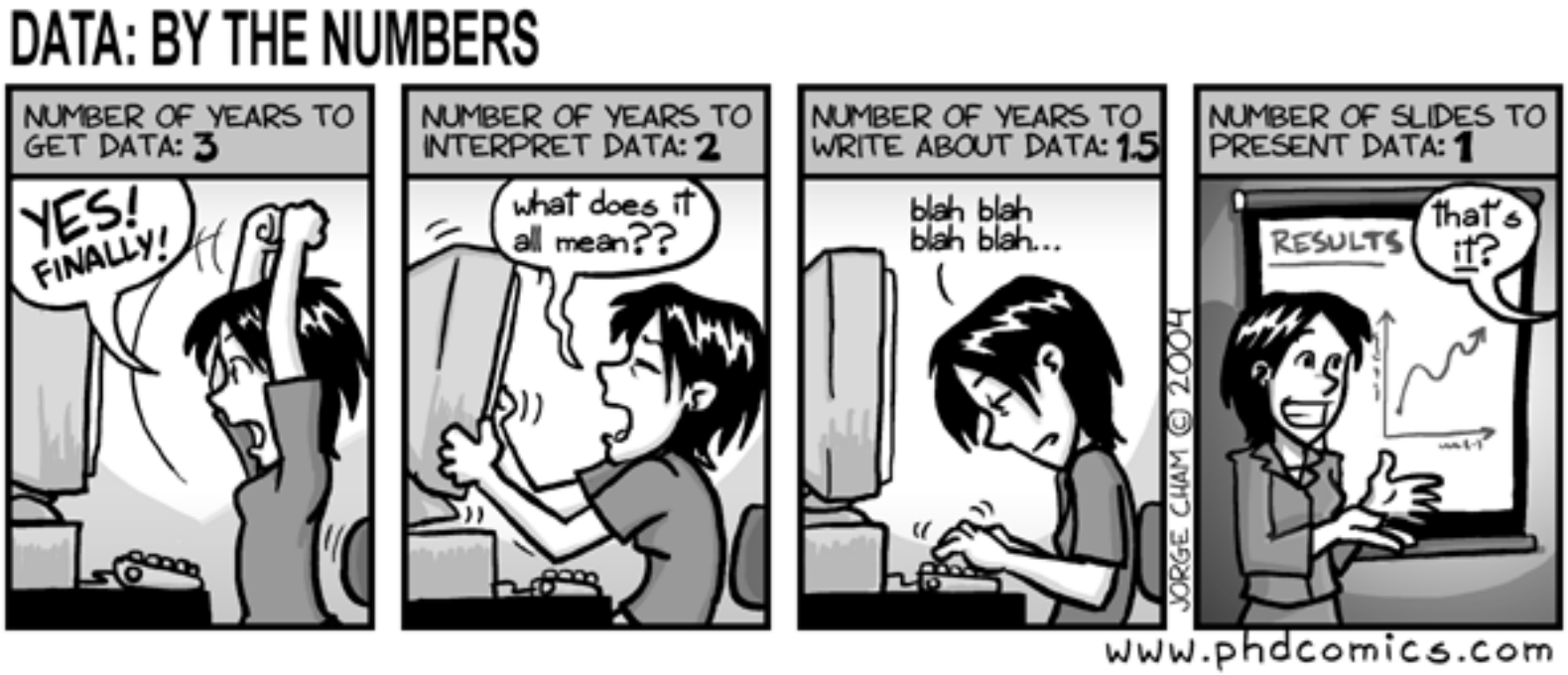
En una cadena de almacenes (Wal-Mart o Costco) analizaron los datos de compras de sus clientes

- *Data:* Los registros de los artículos que habían comprado, junto con datos relativos a la hora, el género del comprador y la edad
- *Information:* Se descubrió una alta correlación entre: *compradores hombres*, *compras entre 5pm y 7pm*, *pañales* y *cervezas*
- *Knowledge:* Saber que los padres, después de salir del trabajo, suelen comprar pañales y también cervezas.
- *Wisdom:* Implementar nuevas estrategias de publicidad y mercadeo.

[1] `http://web.onetel.net.uk/~hibou/Beer%20and%20Nappies.html`

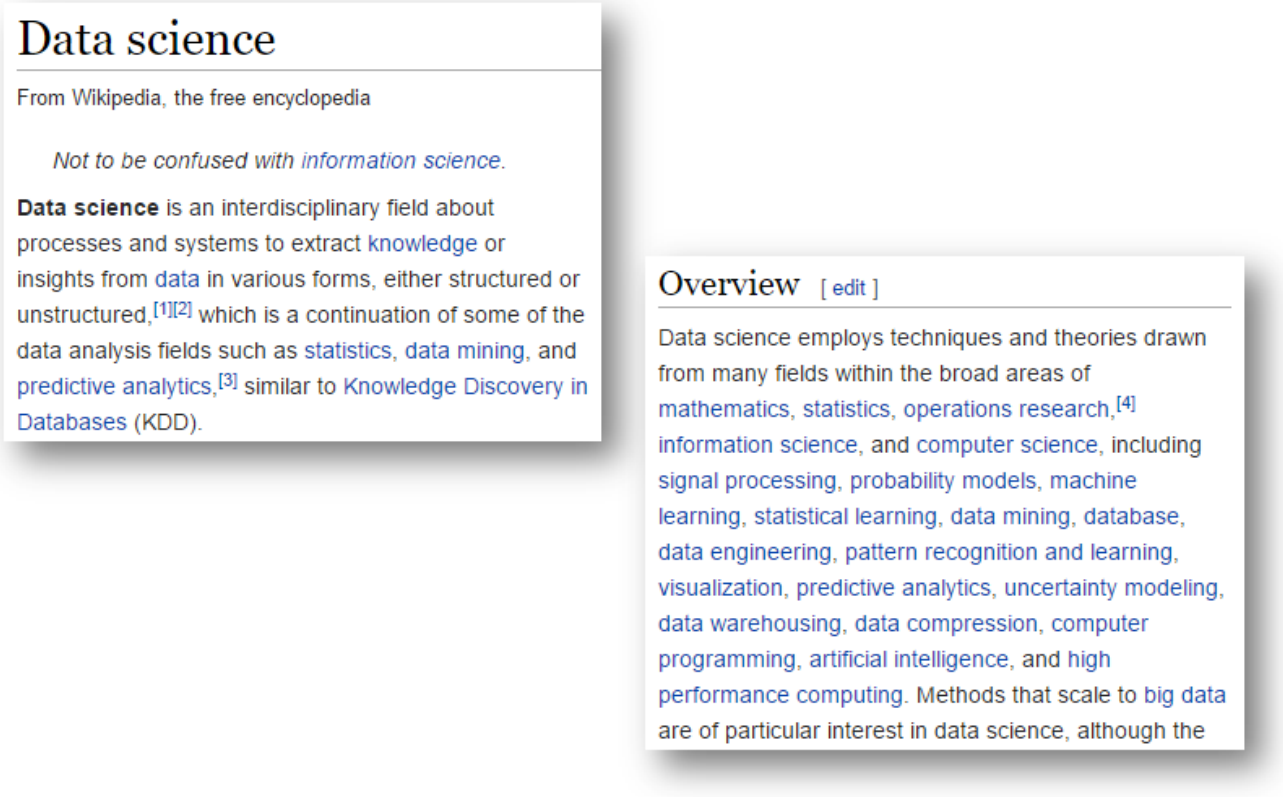
Contenido

Básicamente...[2]



[2] `http://phdcomics.com/comics.php`

Ciencia de los datos - *Data Science* II



Data science

From Wikipedia, the free encyclopedia

Not to be confused with information science.

Data science is an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured,[1][2] which is a continuation of some of the data analysis fields such as statistics, data mining, and predictive analytics,[3] similar to Knowledge Discovery in Databases (KDD).

Overview [edit]

Data science employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, operations research,[4] information science, and computer science, including signal processing, probability models, machine learning, statistical learning, data mining, database, data engineering, pattern recognition and learning, visualization, predictive analytics, uncertainty modeling, data warehousing, data compression, computer programming, artificial intelligence, and high performance computing. Methods that scale to big data are of particular interest in data science, although the

¿La Ciencia de los Datos es "eso" que hacen google y facebook?

Antes de profundizar en *¿qué es la Ciencia de los Datos?*, entendamos primero un poco los conceptos que la acompañan

Alrededor de la Ciencia de los Datos...

La Ciencia de los Datos está relacionada con áreas tan diversas (y a la vez tan afines) como son:

- *Big data*
- *Machine learning*
- *Data mining*
- *Business intelligence*
- *Data analytics*
- ...

Big data I

¿Qué es *big data*? y ¿qué relación tiene con la Ciencia de los Datos?

¿La Ciencia de los Datos es la ciencia del *big data*?

¿Por qué cuando se habla de *big data* se habla de varias disciplinas (finanzas, mercados, astronomía, tecnología) y cuando se habla de la ciencia de datos se habla sólo del campo de desarrollo de software?

Big data II

Veamos una definición tomada de *Doing Data Science*[3]

> Big Data is a vague term, used loosely, if often, these days. But put simply, the catchall phrase means three things. First, it is a bundle of technologies. Second, it is a potential revolution in measurement. And third, it is a point of view, or philosophy, about how decisions will be—and perhaps should be—made in the future.
>
> — Steve Lohr
> *The New York Times*

[3] Schutt R. and O'Neil K. *Doing Data Science*, Ed. O'Reilly, 2013

Big data III

big data busca recoger y gestionar grandes cantidades de datos para alimentar, principalmente, aplicaciones web $\Rightarrow$ (debido a tamaños de almacenamiento y poder de cómputo)

La Ciencia de los Datos busca crear modelos que capturen los patrones ocultos (subyacentes) en sistemas complejos

Si bien lo dos tienen el potencial de generar valor añadido a partir de los datos, la diferencia se podría resumir en: *Collecting Does Not Mean Discovering*[4]

[4] `http://www.kdnuggets.com/2015/07/data-science-big-data-different-beasts.html`

Big data IV

Big data vs. Small data

Si bien no hay cómo cuantificarlo a ciencia cierta, los más conservadores hablan de *big data* $\Rightarrow$ petabytes o exabytes

Procesar cantidades de información a estas escalas es costoso y requiere de un esfuerzo considerable

Hoy en día se habla de *Small data*[5] $\Rightarrow$ cantidades de información que maximizan la relación costo-beneficio

En comparación, representan pequeñas fracciones de lo que podría representar *big data*

[5] `https://www.bbvaopenmind.com/en/small-data-vs-big-data-back-to-the-basics`

Aprendizaje de máquina

En una frase: *aprendizaje de máquina* es el conjunto de los **algoritmos** y las **técnicas** que se usan para diseñar sistemas que aprendan a partir de los datos

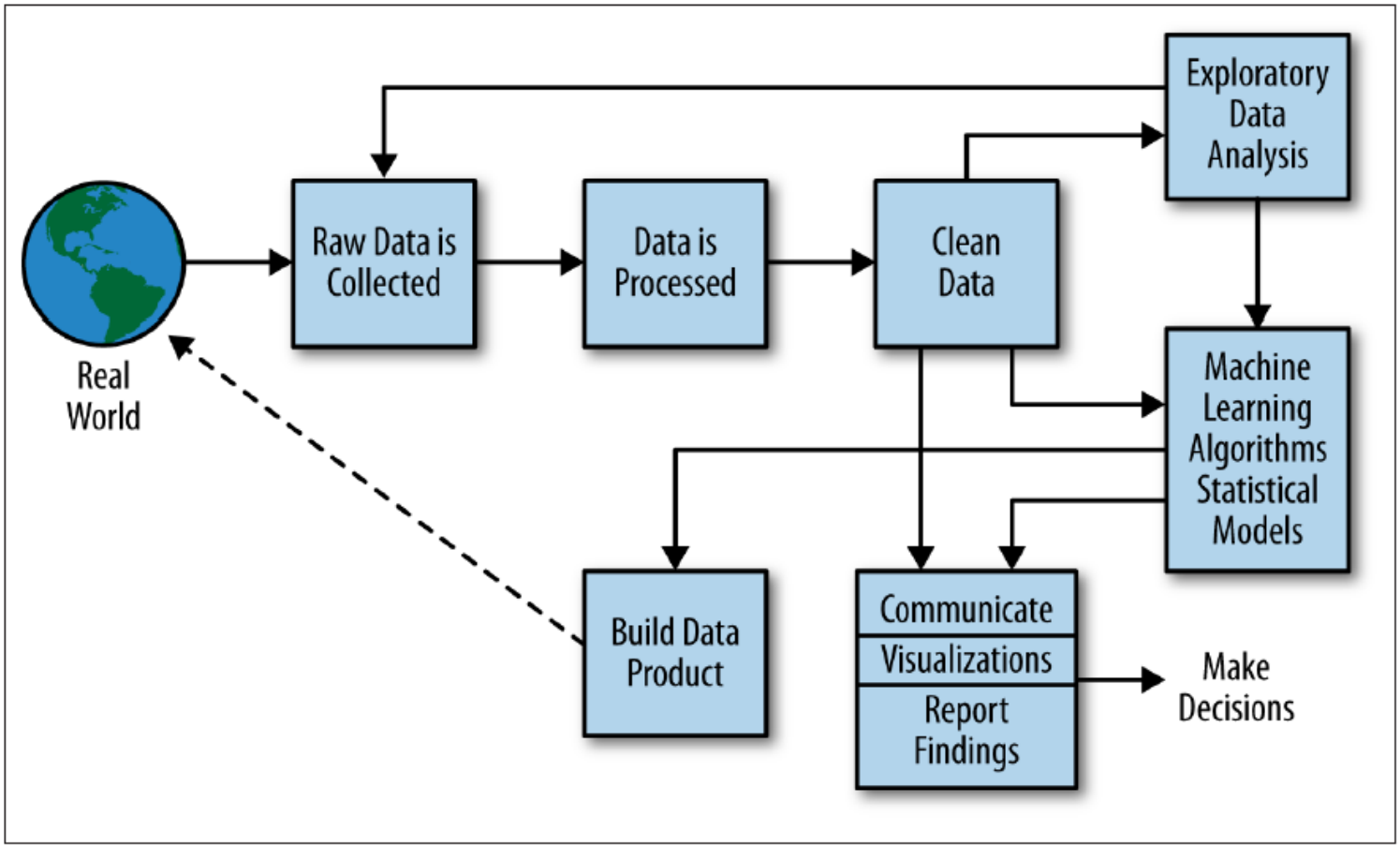
Los fundamentos del *aprendizaje de máquina* se basan en las **matemáticas** y la **estadística**

De forma general, no tienen en cuenta el conocimiento del dominio y el pre-procesamiento de los datos

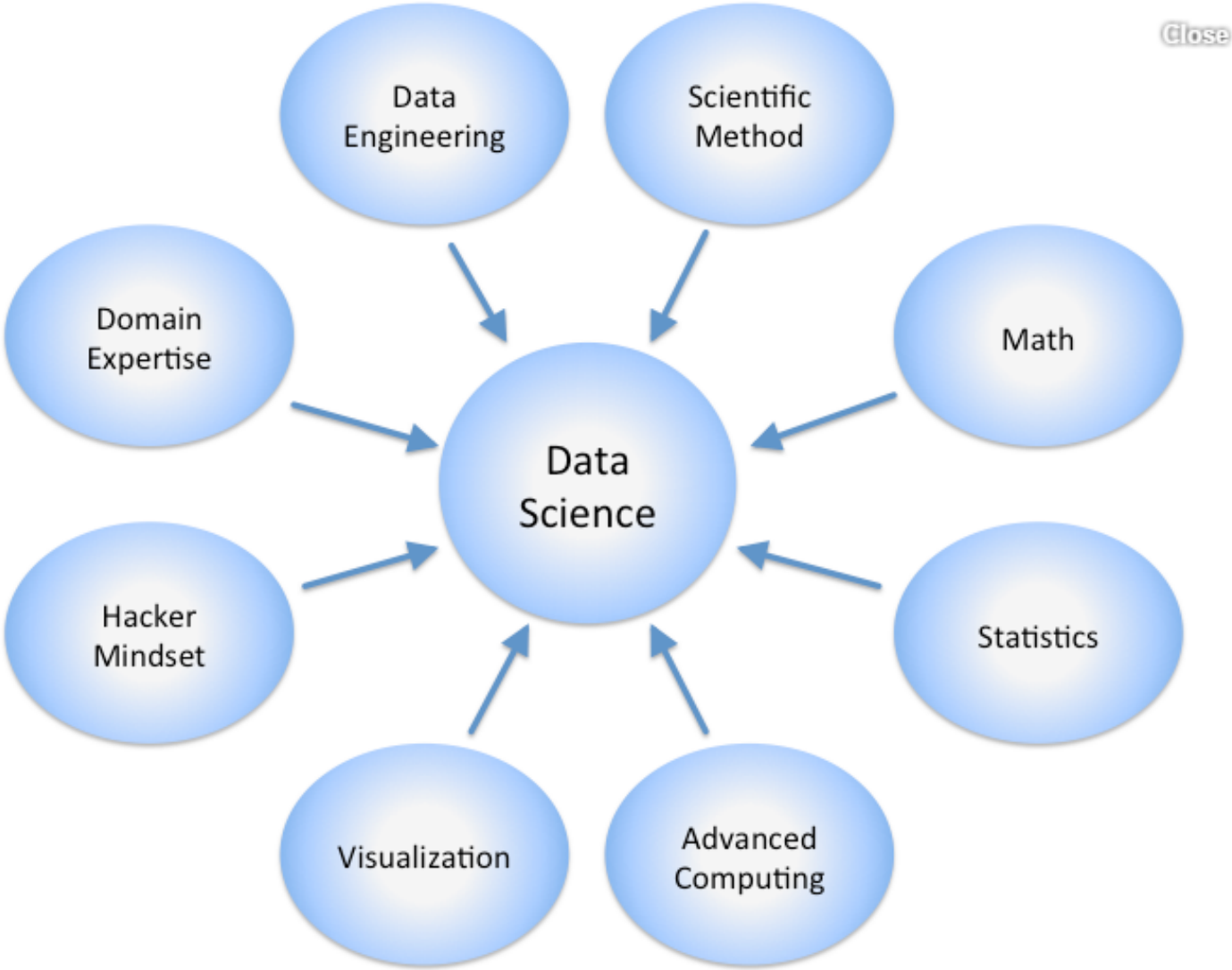
Míneria de datos

Minería de datos $\Rightarrow$ Proceso de descubrir patrones en los datos

Ciencia de los datos I



Ciencia de los datos II



Contenido

En Ingeniería I

Las Smart Grids (Redes Eléctricas Inteligentes) integran las tecnologías de la información y la comunicación (TICs) dentro del negocio de la energía eléctrica

La AMI (*Advanced Metering Infrastructure*) permite realizar lecturas en tiempo real del consumo energético de cada uno de los abonados

En Ingeniería II

Algunos investigadores concuerdan en que de haber aplicado la Ciencia de los Datos al monitoreo del estado y del desgaste de sensores y actuadores se habrían podido evitar desastres como los de Deepwater Horizon, Exxon Valdez o Fukushima[6]

[6] http://www.mastersindatascience.org/industry/energy/

En la Inteligencia de Negocios

La inteligencia de negocios (*Business Intelligence*) se ha abierto campo como la disciplina encargada de involucrar el análisis cuantitativo de datos en la toma de decisiones.

Ejemplos:

- Tarjetas de fidelización de clientes (por medio de éstas se obtienen datos de edad, género, ubicación geográfica, entre otros)
- Segmentación de mercados regionales (hacer más inversiones en publicidad dependiendo de los artículos más vendidos por regiones)
- Mejorar la logística y los canales de distribución de bienes y servicios

En sistemas de recomendación - (*Association Rules*)

Desarrollo de sistemas de recomendación personalizada.

Generación de perfiles de usuario (caso Netflix)[7]

[7] Maheshwari A., *Data analytics made accesible*, 2014

En minería de datos - (*data mining*)

Portales de noticias como `https://news.google.com/`

Contenido

Motivación I

- En los últimos años ha habido un *boom* relacionado con el *big data* y la Ciencia de los Datos
- Las fuentes de datos se han multiplicado y diversificado (Internet, dispositivos móviles, sensores, transacciones comerciales, etc.)
- Se han reducido los costos en la obtención de los datos
- Estamos experimentando un cambio de paradigma en la forma como se analizan los datos y se extrae información de ellos
- La Ciencia de los Datos es un área aún por explorar y con grandísimas capacidades de expansión y desarrollo

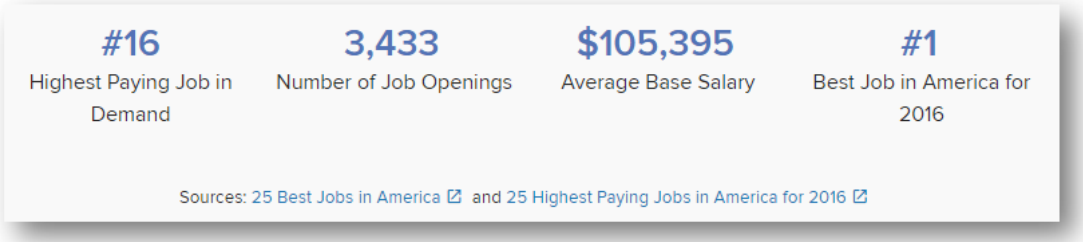
Motivación II

Colciencias en el 2014 lanzó la Convocatoria de Centros de Excelencia en Apropiación de *Big Data* y *Data Analytics* (recursos de $3.500.000.000)

Motivación III

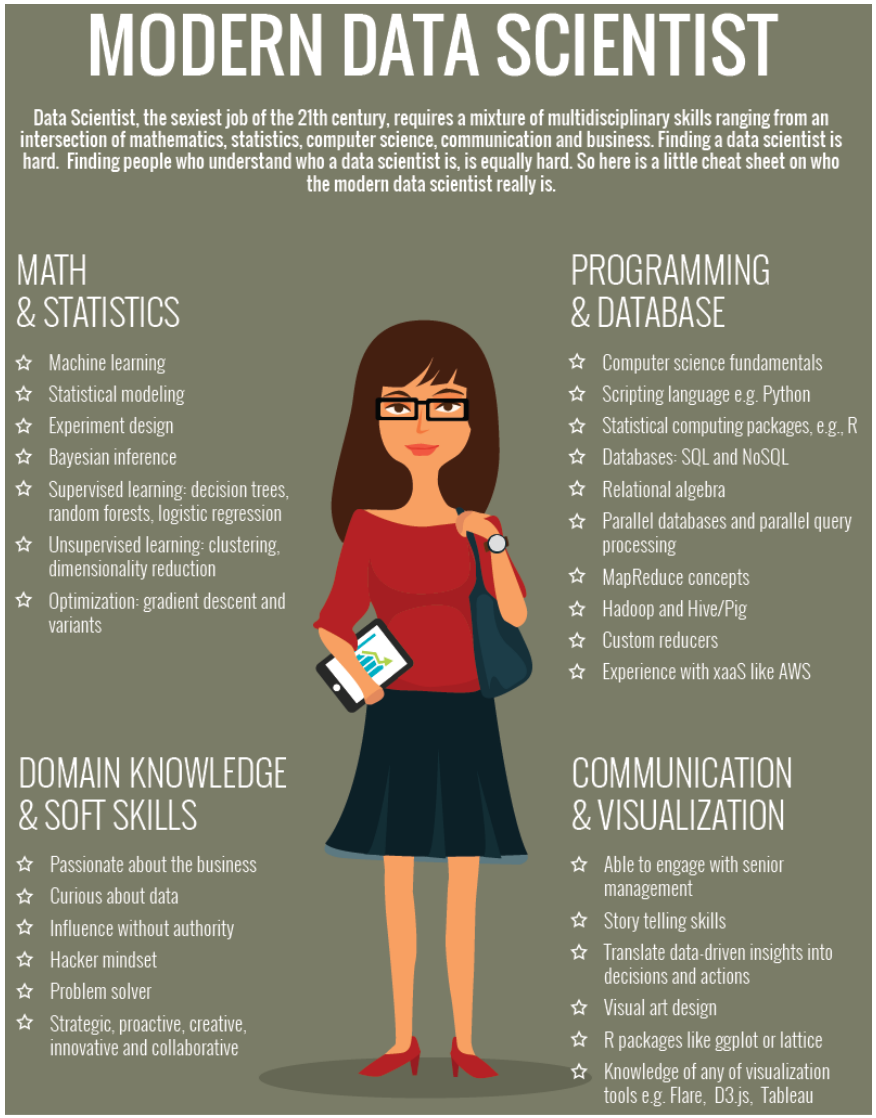
De acuerdo al Harvard Business Review[8]

De acuerdo a la *School of Information* de la Universidad de Berkeley[9]



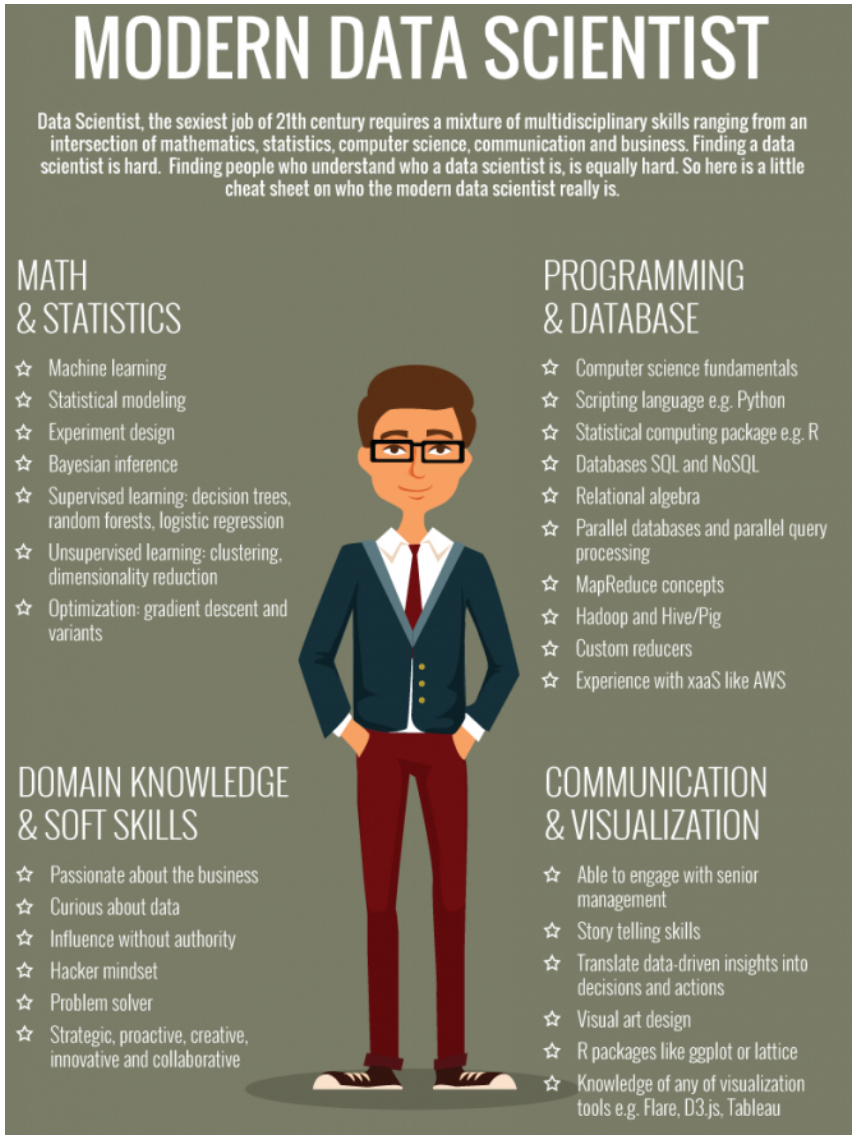
[8] https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century
[9] https://datascience.berkeley.edu/about/what-is-data-science/

Perfil de la científica de datos



http://www.marketingdistillery.com/2014/11/29/is-data-science-a-buzzword-modern-data-scientist-defined/

Perfil del científico de datos



http://www.marketingdistillery.com/2014/11/29/is-data-science-a-buzzword-modern-data-scientist-defined/

En resumen...

Científico de datos: "Persona que sabe más de **estadística** que cualquier programador y que a la vez sabe más de **programación** que cualquier estadístico". Necesitamos:

- Álgebra lineal
- Teoría de probabilidades
- Optimización
- Programación (Matlab, R, **Python**)
- En conclusión necesitamos del aprendizaje estadístico (aprendizaje de máquina)