

La ciencia de datos y el aprendizaje de máquina

Proyecto Integrador 1

Juan David Martínez
jdmartinev@eafit.edu.co

Paola Vallejo
pvallej3@eafit.edu.co

2023

Contenido



1. Datos e información
2. Ciencia de datos
3. Perfil del científico de datos
4. Aprendizaje de máquina
5. Cómputo en nube

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 II

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

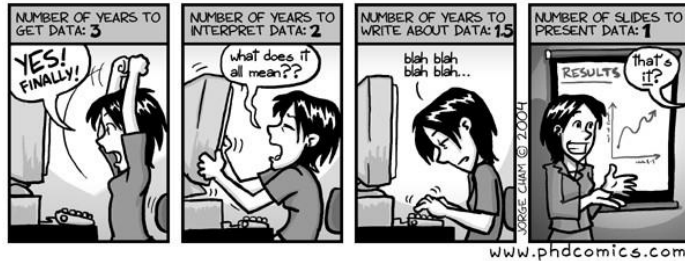
En una cadena de almacenes (Wal-Mart o Costco) analizaron los datos de compras de sus clientes

- *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.

Ciencia de datos - *Data Science I*

Básicamente...¹

DATA: BY THE NUMBERS



¹<http://phdcomics.com/comics.php>

Ciencia de 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]

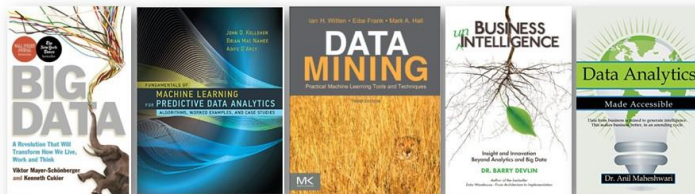
Data science employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, operations research,^[4] information science, and computer science, including signal processing, probability models, machine learning, statistical learning, data mining, database, data 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 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

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*²

²[http:](http://www.kdnuggets.com/2015/07/data-science-big-data-different-beasts.html)

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

Big data III

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

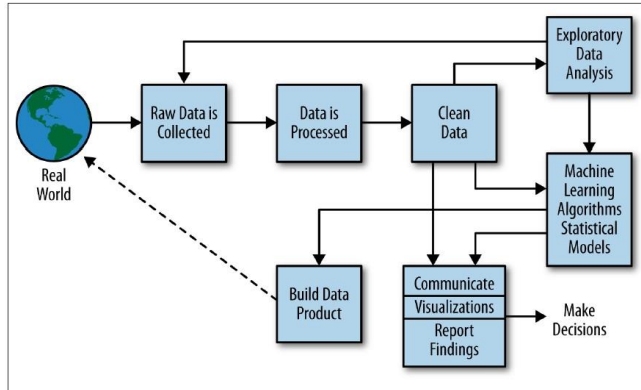
Hoy en día se habla de *Small data*³ \Rightarrow cantidades de información que maximizan la relación costo-beneficio

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

³[https:](https://www.bbvaopenmind.com/en/small-data-vs-big-data-back-to-the-basics)

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

Esquema general aplicaciones en Ciencia de los datos



En Ingeniería

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 Fukushima⁴

⁴<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)⁵



⁵Maheshwari A., *Data analytics made accesible*, 2014

En minería de datos - (data mining)

Portales de noticias como <https://news.google.com/>

Noticias destacadas



Delta Airlines sufre demoras y cancelaciones en sus vuelos por un "apagón informático"

Infobae.com - hace 1 hora

La aerolínea confirmó un fallo de su sistema, aunque no dio detalles sobre la causa ni el tiempo que demoraría en resolverlo. Mientras tanto, todos sus aviones permanecerán en tierra. La compañía opera 15.000 viajes a diario en el mundo. 8 de agosto de...

Se reanudan vuelos de Delta Airlines tras apagón informático
Noticias RCN (Comunicado de prensa) (blog)

Aerolínea Delta reanuda sus operaciones tras resolver problema técnico W Radio

De Estados Unidos: Aviones de Delta quedan varados por problema informático Mundo Hispanico

Ver las 61 fuentes »

Relacionados
Delta Air Lines »



Primera Hora Voz de Am... Yahoo! Fin... La Nación ... El Nuevo H... RCN Radio... El Nuevo D... CNNespa... E...

El boom de la ciencia de datos

- 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

El boom de la ciencia de datos

De acuerdo al Harvard Business Review⁶

**Harvard
Business
Review**

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

De acuerdo a la *School of Information* de la Universidad de Berkeley⁷

#16	3,433	\$105,395	#1
Highest Paying Job in Demand	Number of Job Openings	Average Base Salary	Best Job in America for 2016
Sources: 25 Best Jobs in America ↗ and 25 Highest Paying Jobs in America for 2016 ↗			


⁶<https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century>

⁷<https://datascience.berkeley.edu/about/what-is-data-science/>

Perfil de la científica de datos

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21st 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 packages, 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

Perfil del científico de datos

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21st 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

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**, **Cloud computing**)
- En conclusión necesitamos del aprendizaje estadístico (aprendizaje de máquina)

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.
- El aprendizaje de máquina es el eje central de la ciencia de datos y la inteligencia artificial - (IA).
- Primeros avances serios en IA:
https://www.youtube.com/watch?v=FwFduRA_l6Qabc_hannel = YannLeCun.

El renacer de la inteligencia artificial (Premio Turing 2019)

'Godfathers of AI' honored with Turing Award, the Nobel Prize of computing

Yoshua Bengio, Geoffrey Hinton, and Yann LeCun laid the foundations for modern AI

By James Vincent | Mar 27, 2019, 6:02am EDT

f t SHARE



En 2006, Geoffrey Hinton et al. publicaron un artículo⁸ que mostraba como un algoritmo de aprendizaje profundo podía reconocer dígitos a mano con una precisión $> 98\%$, llamándolo Deep Learning.

⁸ver <http://www.cs.toronto.edu/~hinton/>

El renacer de la inteligencia artificial (Aprendizaje de máquina)

- Entrenar un modelo de deep learning era considerado imposible en los 90s.
- Hinton y los demás investigadores en redes neuronales empezaron a destronar a los algoritmos clásicos de aprendizaje de máquina.
- La clave: mucho poder de cómputo y muchos datos.
- En la actualidad: aprendizaje de máquina como corazón de muchos productos de tecnología de punta (búsqueda web, teléfonos inteligentes, reconcomiento de habla, autos que se conducen solos, etc...)

Qué es aprendizaje de máquina? El renacer de la inteligencia artificial (Aprendizaje de máquina)

Básicamente...programar computadores para **aprender desde datos!**

Después de entender la importancia de la ciencia de los datos y su conexión con el aprendizaje de máquina, se busca entonces:

- Entender los modelos básicos de aprendizaje de máquina.
- Avanzar hasta los modelos más avanzados (Deep learning).
- Fortalecer las competencias en estadística y programación.
- Utilizar herramientas libres y reconocidas en Python (Pandas, SciKitlearn, TensorFlow, Keras, PyTorch).

Aprendizaje de máquina: Conceptos básicos

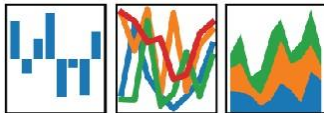
Pasos principales en un proyecto de aprendizaje de máquina.

- Funciones de costo y optimización.
- Preparación y preproceso de los datos.
- Sintonización de hyperparámetros usando validación cruzada.
- Sobreajuste (overfitting) y subajuste (underfitting)
- Algoritmos tradicionales en aprendizaje de máquina: i) regresión polinomial, ii) regresión logística, iii) k-nn, iv) SVM, v) árboles de decisión, vi) métodos de ensamble.
- Fundamentos en redes neuronales y deep learning (para más información ver curso tópicos avanzados en aprendizaje de máquina).

Nuestras librerías amigas Python - Pandas

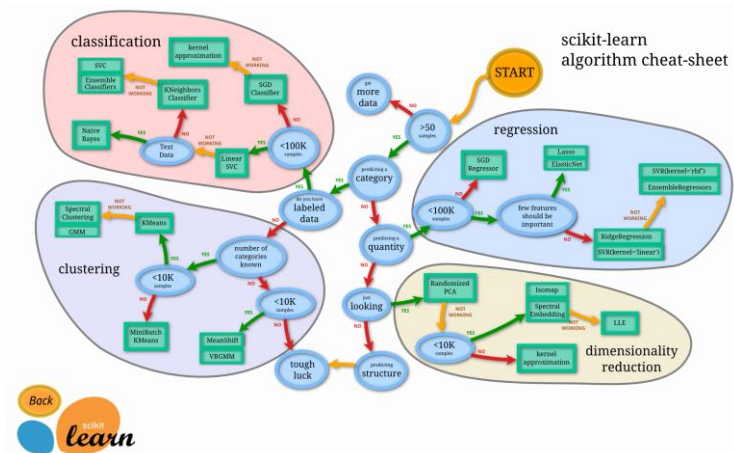
pandas

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



	BandName	WavelengthMax	WavelengthMin
0	CoastalAerosol	450	430
1	Blue	510	450
2	Green	590	530
3	Red	670	640
4	NearInfrared	880	850
5	ShortWaveInfrared_1	1650	1570
6	ShortWaveInfrared_2	2290	2110
7	Cirrus	1380	1360

Nuestras librerías amigPythonas - Scikit-learn



Nuestras librerías amigas Python – TensorFlow, Keras, PyTorch



Cómputo de alto desempeño Gratis!

No quemes más tu PC!



- [Material disponible en línea \(https://d2l.ai/\)](https://d2l.ai/)
- [Ejemplo ilustrativo - Curso Analítica de Datos UNAL \(https://github.com/amalvarezme/AnaliticaDatos\)](https://github.com/amalvarezme/AnaliticaDatos)



GRACIAS