

INTRODUCCIÓN AL CURSO

ALAN REYES-FIGUEROA

INTRODUCCIÓN A LA CIENCIA DE DATOS

(AULA 01) 11.ENERO.2021

Motivación

El curso de ciencia de datos es una introducción a los métodos estadísticos, matemáticos y computacionales para extraer información basada en datos. Incluye técnicas provenientes áreas como: estadística, reconocimiento estadístico de patrones (*pattern recognition*), aprendizaje estadístico o aprendizaje de máquina (*machine learning*), ciencia de datos.

Este es un curso integrador. Haremos uso de

- estadística e inferencia estadística,
- álgebra lineal (espacios, autovalores, descomposición matricial),
- optimización continua,
- reconocimiento de patrones y aprendizaje estadístico,
- programación y algoritmos.

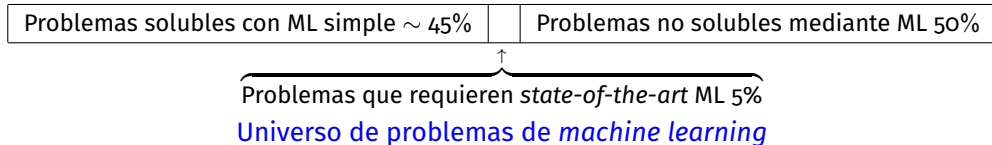
¿Qué no es ciencia de datos?

Ciencia de datos \neq *machine learning* (ML)

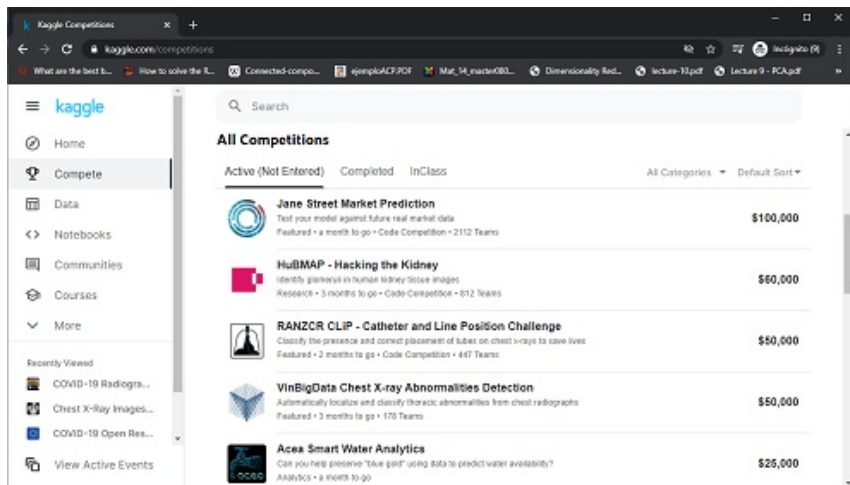
- El aprendizaje automático involucra, matemática, computación y estadística, pero tradicionalmente no trata sobre cómo resolver preguntas científicas.

El aprendizaje automático tiene un enfoque más de algoritmos.






- Algunas veces, la mejor forma de resolver un problema es visualizando los datos.



¿Qué no es ciencia de datos?



The screenshot shows the Kaggle website's 'All Competitions' page. The left sidebar contains navigation links: Home, Compete (selected), Data, Notebooks, Communities, Courses, and More. Below these are 'Recently Viewed' items: COVID-19 Radiographs, Chest X-ray Images, and COVID-19 Open Results. The main content area lists competitions with their icons, titles, descriptions, and prize amounts.

Competition Icon	Competition Title	Description	Prize Amount
	Jane Street Market Prediction	Test your model against future real market data. Featured • a month to go • Code Competition • 2112 Teams	\$100,000
	HuBMAP - Hacking the Kidney	Identify glomeruli in human kidney tissue images. Research • 3 months to go • Code Competition • 612 Teams	\$60,000
	RANZCR CLIP - Catheter and Line Position Challenge	Classify the presence and correct placement of tubes on chest x-rays to save lives. Featured • 2 months to go • Code Competition • 447 Teams	\$50,000
	VinBigData Chest X-ray Abnormalities Detection	Automatically localize and classify thoracic abnormalities from chest radiographs. Featured • 3 months to go • 176 Teams	\$50,000
	Acea Smart Water Analytics	Can you help preserve "blue gold" using data to predict water availability? Analytics • a month to go	\$25,000

¿Qué no es ciencia de datos?

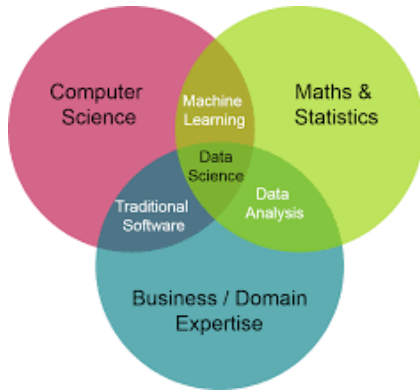
Data science \neq competencias o concursos.

- Concursos de ciencia de datos, *e.g.* Kaggle, usualmente requieren optimizar una métrica sobre un conjunto de datos fijo.
- Esto, en última instancia, no resuelve un problema científico o aplicado.
- La ciencia de datos es un ciclo iterativo en el que se plantea un problema, y se busca diseñar mecanismos o algoritmos para resolverlo (o determinar que no es posible), y evaluar qué aportes pueden generar estos algoritmos sobre la pregunta en consideración.

¿Qué no es ciencia de datos?

Ciencia de datos \neq estadística

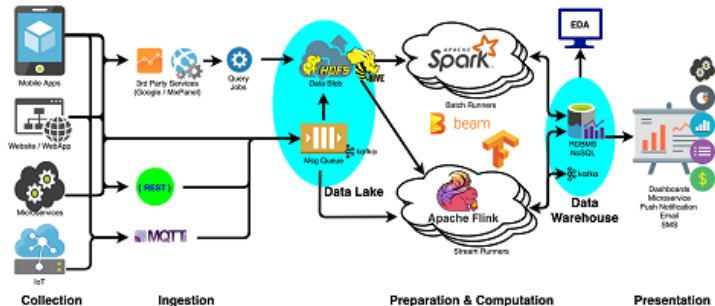
- Estadística (al menos en un sentido académico), ha evolucionado al punto de probar teoremas. Hacer teoría estadística.
- En este curso veremos algunos pocos teoremas, pero no vamos a hacer teoría. La idea principal es que este sea un curso aplicado.



¿Qué no es ciencia de datos?

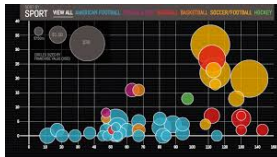
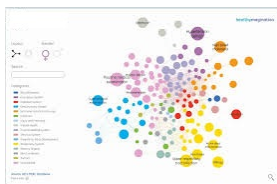
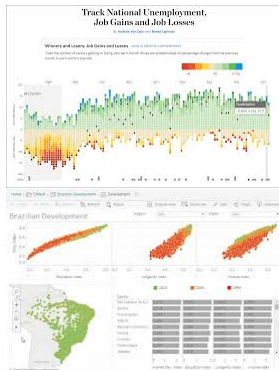
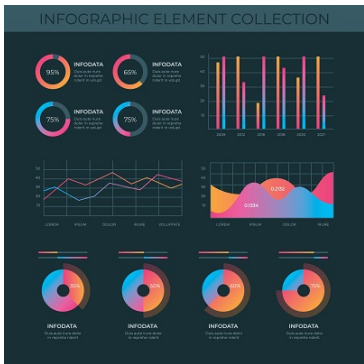
Ciencia de datos \neq *big data*

- El término *big data* está más relacionado con la ingeniería de software. Se refiere más al tratamiento de grandes cantidades de datos, o a las técnicas, metodologías o desarrollo de *pipelines* o *workflows* para el procesamiento de datos.



¿Qué no es ciencia de datos

Ciencia de datos \neq visualización



«The greatest value of a picture is when it forces us to notice what we never expected to see.» –John Tukey

¿Qué es ciencia de datos?

Algunas posibles definiciones.

- Es la aplicación de técnicas estadísticas y computacionales para obtener o ganar entendimiento de un problema en el mundo real, mediante datos.
- Ciencia de datos = estadística + procesamiento (minería) de datos + aprendizaje automático + investigación científica + visualización de datos + inteligencia de negocio + *big data* + ...
- A criterio personal, aún no hay una definición concreta, cada persona hace su propia definición según su experiencia y punto de vista.
- Lo que está claro, es que es un tema que mezcla y usa herramientas de muchas áreas del conocimiento.

¿Qué es ciencia de datos?

- Recientemente hay mucha demanda por científicos de datos.
- En 2018, US experimentará una demanda de 190,000 científicos de datos, y 1.5 millones de gerentes y analistas capaces de generar información útil mediante datos.

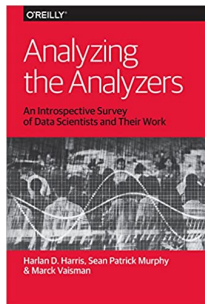
Ref. Susan Lund *et al.*, “Game Changers: Five Opportunities for US Growth and Renewal,” McKinsey Global Institute Report, July 2013.



¿Qué es ciencia de datos?

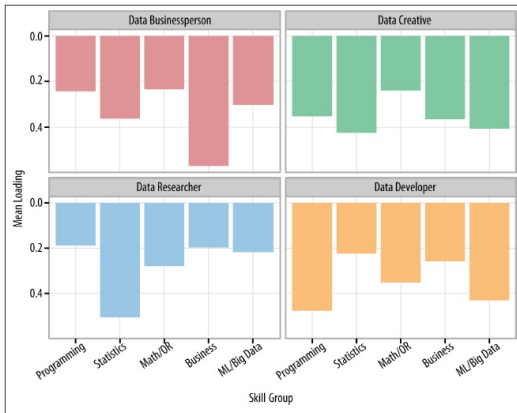
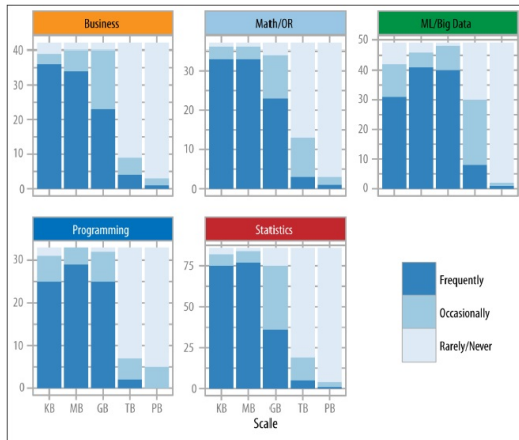
- La ciencia de datos y el aprendizaje automático no son nada nuevo, pero la tendencia actual continúa impulsando las tecnologías hacia el centro de atención.
- Creciente interés (y exageración) en torno a la inteligencia artificial (IA), impulsado por el *marketing* y combinada con la comprensible confusión de términos: IA, ML, DC.
- Escasez de talento en ciencia de datos y aprendizaje automático.
- Aumento de la capacidad y potencia informática y la disponibilidad de arquitecturas avanzadas. (Estos avances han alimentado la publicidad y el interés en torno al aprendizaje profundo (*deep learning*)).
- Aumento y popularidad de herramientas y bibliotecas de código abierto para ciencia de datos y aprendizaje automático.

¿Qué hace un científico de datos?

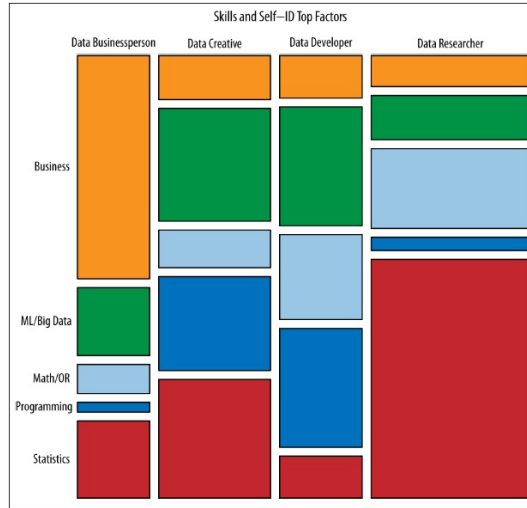


Data Developer	Developer	Engineer	
Data Researcher	Researcher	Scientist	Statistician
Data Creative	Jack of All Trades	Artist	Hacker
Data Businessperson	Leader	Businessperson	Entrepreneur

¿Qué hace un científico de datos?



Habilidades



Habilidades

Business	ML / Big Data	Math / OR	Programming	Statistics
Product Development	Unstructured Data	Optimization	Systems Administration	Visualization
Business	Structured Data	Math	Back End Programming	Temporal Statistics
	Machine Learning	Graphical Models	Front End Programming	Surveys and Marketing
	Big and Distributed Data	Bayesian / Monte Carlo Statistics		Spatial Statistics
		Algorithms		Science
		Simulation		Data Manipulation
				Classical Statistics

Tareas

		Hacker																Scripter					Application User											
		Analytics	Biotech	Datamart	Finance	Finance	Healthcare	Healthcare	Healthcare	Insurance	Marketing	Marketing	News	Retail	Retail	Social Networking	Social Networking	Visualization	Web	Web	Analytics	Analytics	Analytics	Finance	Healthcare	Media	Retail	Finance	Insurance	Retail	Retail	Sports	Web	Security
Process	Discovery	Locating Data	x	x	x	x	x	x	x					x		x	x	x	x															
		Field Definitions	x	x	x	x	x	x	x	x	x	x				x	x	x		x														
	Wrangle	Data Integration	x	x	x	x		x	x	x	x	x			x	x	x	x		x														
		Parsing Semi-Structured	x	x	x			x	x	x	x	x				x	x	x	x	x														
		Advanced Aggregation and Filtering	x			x		x	x	x				x	x		x	x	x	x														
	Profile	Data Quality	x	x						x	x	x			x	x	x	x	x	x														
		Verifying Assumptions	x	x	x	x		x	x	x	x	x			x	x	x	x	x	x														
	Model	Feature Selection	x	x	x						x	x	x			x	x	x	x	x														
		Scale	x	x	x	x	x		x	x	x	x	x			x	x	x	x	x														
		Advanced Analytics	x			x					x	x	x	x			x	x		x														
Report	Communicating Assumptions							x	x		x	x			x	x	x		x															
	Static Reports		x	x		x		x	x	x	x				x	x			x															
Workflow	Data Migration	x	x	x	x	x		x	x				x		x	x	x	x	x															
	Operationalizing Workflows		x	x		x	x			x	x			x	x	x	x		x															
Tools	Database	SQL	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x								
		Hadoop/Hive/Pig	x		x				x						x		x	x		x														
		MongoDB										x																						
		CustomDB	x					x	x	x																								
	Scripting	Java	x	x		x			x	x	x			x	x	x	x	x		x														
		Perl																																
		Python	x	x	x	x	x		x	x	x					x	x	x		x														
		Clojure											x				x	x																
		Visual Basic		x																														
	Modeling	R	x		x						x	x			x	x	x	x		x														
		Matlab				x									x					x														
		SAS	x																															
		Excel		x		x	x			x	x	x	x		x	x	x			x														



Data Science Programming Languages



Python



Scala



R



SAS



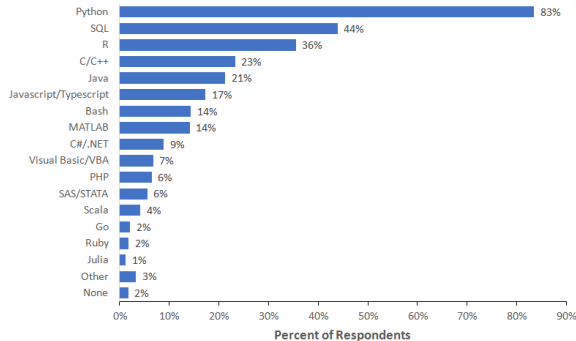
SQL



Julia

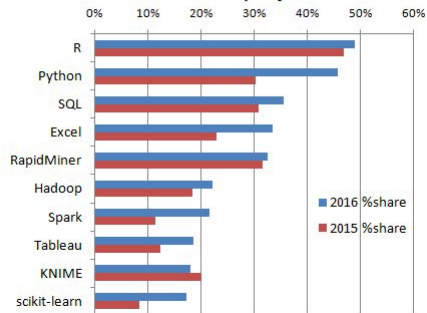
Lenguajes y herramientas

What programming language do you use on a regular basis?



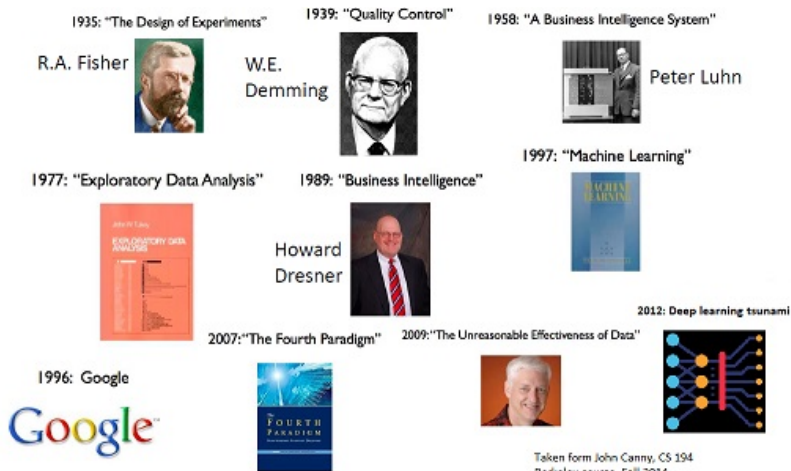
Note: Data are from the 2018 Kaggle Machine Learning and Data Science Survey.

KDnuggets Analytics/Data Science
2016 Software Poll, top 10 tools



Un poco de historia

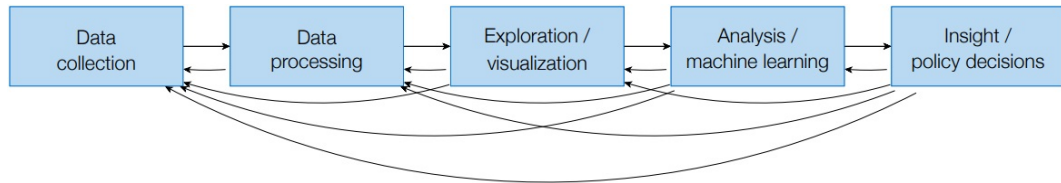
La ciencia de datos no es un tema nuevo.



Taken from John Canny, CS 194
Berkeley course, Fall 2014.

Haciendo ciencia de datos

Hacer ciencia de datos es un proceso que conlleva varias etapas y que integra habilidades diversas, y colaboración entre disciplinas, profesionales y enfoques diversos.



Por ejemplo, Ben Fry, propone el siguiente modelo de ciencia de datos:

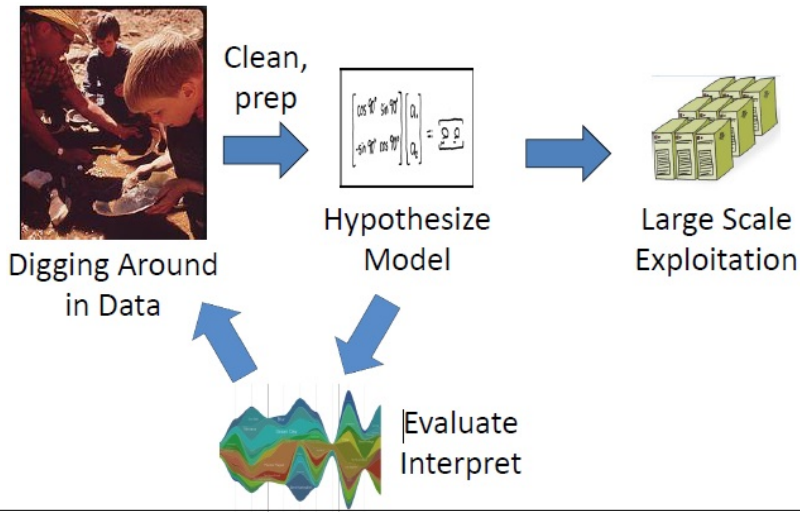
1. Acquire
2. Parse
3. Filter
4. Mine
5. Represent
6. Refine
7. Interact

Haciendo ciencia de datos

En contraste, Jeff Hammerbacher propone este esquema para hacer ciencia de datos:

1. Identify problem
2. Instrument data sources
3. Collect data
4. Prepare data (integrate, transform, clean, filter, aggregate)
5. Build model
6. Evaluate model
7. Communicate results

Haciendo ciencia de datos

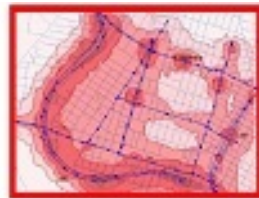
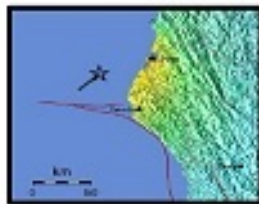


La parte difícil

¿Qué parte es difícil a la hora de hacer ciencia de datos?

- Overcoming assumptions
- Making ad-hoc explanations of data patterns
- Overgeneralizing
- Communication
- Not checking enough (validate models, data pipeline integrity, etc.)
- Using statistical tests correctly
- Mathematical models fail (who do you ask?)
- Prototype - Production transitions
- Data pipeline complexity (who do you ask?)

En resumen



Crowdsourcing + physical modeling + sensing + data assimilation

to produce:



Algunos ejemplos

¿De qué va este curso?

Como este es un curso de matemática, la idea es hacer una introducción a la ciencia de datos, desde un punto de vista más matemático.

- Más orientado a *machine learning*, patrones y análisis de datos.
Veremos algoritmos, y su fundamento matemático (no vamos a hacer teoría, pero sí vamos a mencionar teoremas importantes, y mostrar algunos de ellos).
Fundamentos en optimización, estadística, cálculo y álgebra lineal (herramientas).
- Veremos una parte computacional: implementar algoritmos.
Laboratorios
Ejercicios sobre algoritmos (teórico), analizar datos (aplicado).
- Análisis de datos reales.
Proyectos aplicados

Detalles importantes

- Requisitos:
 - Cálculo, álgebra lineal
 - Al menos un curso de estadística
 - Al menos un curso de programación (Python)
- Horario de laboratorio.
- Horario de atención.
- ¿Qué han visto en otros cursos? (*e.g. big data*)
- Seminarios