

TEC

Tecnológico
de Costa Rica

CIENCIA DE DATOS: MARCO DE REFERENCIA Y TENDENCIAS

M.Sc. Esteban Ballesterero

Instituto Tecnológico de Costa Rica
Escuela de Ciencias Naturales y Exactas



ALGUNAS ESTADÍSTICAS IMPORTANTES...

- A diario en el mundo se generan 2.5 trillones de bytes de información
- El 90% de los datos a nivel mundial se han creado solamente en los últimos 2 años
- Empresas como Google cuenta con al menos 600 personas dedicadas al estudio del Big Data
- The Economist Intelligence Unit (EIU) entrevistó a 600 ejecutivos globales y el 54% de los empresarios estadounidenses aseguraron que encontrar a los profesionales adecuados para un exitoso proyecto de Big Data es el obstáculo más importante para no hacerlo

ALGUNAS ESTADÍSTICAS IMPORTANTES...

- El mayor número de profesionales en el área de Big Data provienen de países asiáticos
- De 2012-2013 más del **60% de los artículos de opinión de tecnología avanzada hablan de Big Data** como la nueva estrategia indispensable para las empresas de cualquier sector, declarando, poco menos , que aquellos que no se sumen a este nuevo movimiento, **quedarán obsoletos**.
- Cada minuto, enviamos 204 millones de correos electrónicos
- Cada hora la información consumida a través de Internet equivale a 7 millones de DVD's. Si apilásemos estos discos lado a lado, uno sobre otro, podríamos escalar el Monte Everest 95 veces.

Source: Big Data, Bernard Marr

If data is 'the new oil', then the data scientist functions much like an oil refinery, converting data into insights that can both save money and generate capital

— Eva Short

**GENERAL
MANAGEMENT
PROGRAM**

APPLY NOW

≡ MENU

**Harvard
Business
Review**

DECISION MAKING

Big Data: The Management Revolution

by **Andrew McAfee** and **Erik Brynjolfsson**

FROM THE OCTOBER 2012 ISSUE

DATA

Data Scientist: The Sexiest Job of the 21st Century

10 profesiones que serán más solicitadas en el futuro (pero aún no existen)



Con la colaboración de
Dinero en Imagen

02 may 2017

Redacción



Últimos Artículos



¿Cómo ordenar la circulación de los coches autónomos? Los pájaros dan la solución

EL PAÍS 13 nov 2018



El país latino con más "ransomware" y otros 6 clics tecnológicos en América

Alejandro Rincón Moreno - eldiario.es

El 75% de las profesiones del futuro no existen actualmente, según expertos. Es decir, tres de cada cuatro carreras que se estudian en las universidades podrían quedar desfasadas en cuestión de años.

Aún es complicado saber cuáles serán esas nuevas profesiones del futuro, pero hay quienes se atreven a acercarse a esta realidad.

Recogemos algunas de estas profesiones que aún no existen --o de las que apenas escuchamos hablar-- pero que serán clave en el futuro, así como un par de opciones para poder formar con éxito a los profesionales del futuro.

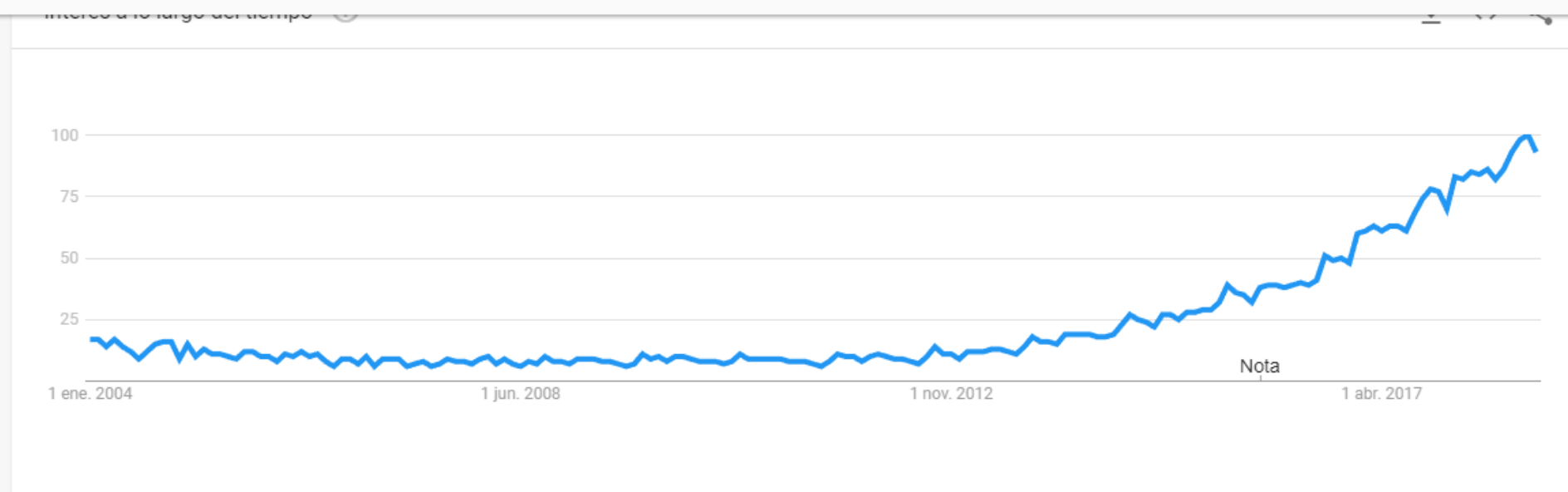
1. Científico de datos

Big Data, Smart Data, Fast Data... La tecnología hace posible recopilar millones de datos sobre cada persona, ya sea a través del móvil, de Internet o de los dispositivos conectados. El análisis en tiempo real sobre datos filtrados será imprescindible para cualquier empresa que quiera ofrecer productos o servicios personalizados a sus clientes y poder mantenerse en el negocio.

Para ello será necesario que existan científicos capaces de dar valor a la información que nos proporcionen las máquinas. Empresas como Paradigma ayudan a sus clientes ya a transformarse digitalmente y aprovechar la inteligencia que proporcionan los datos para adaptarse al mercado.

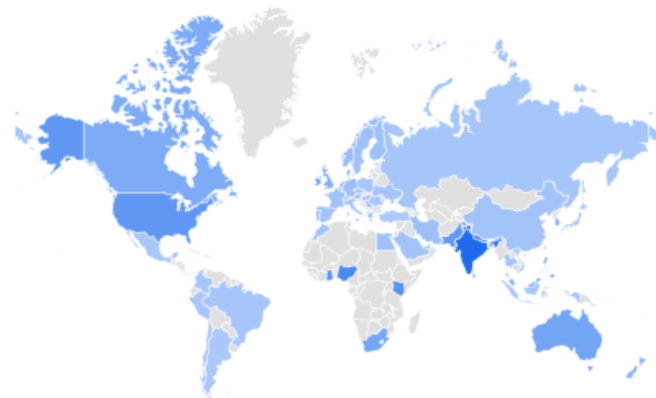
• Data Science

Todo el mundo, 2004 - hoy



Interés por región ?

Región ▼ ⬇ ⌂



1	Santa Elena	100	<div></div>
2	India	83	<div></div>
3	Singapur	78	<div></div>
4	Nepal	65	<div></div>
5	Nigeria	58	<div></div>

☐ Incluir regiones con un volumen de búsquedas bajo

< Mostrando 1-5 de 61 regiones >

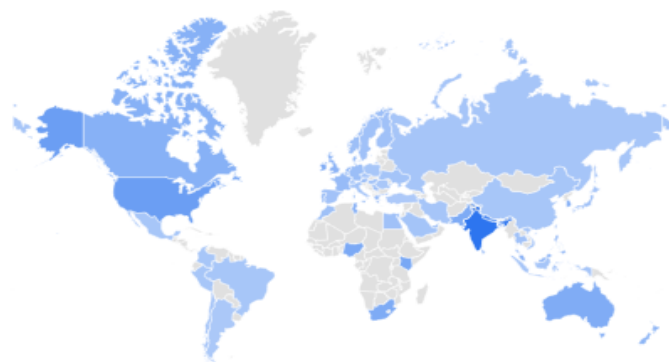
Todo el mundo ▼ 2004 - hoy ▼ Todas las categorías ▼ Búsqueda web ▼

Interés a lo largo del tiempo ?



Interés por región ?

Región ▼



1	Santa Elena	100	<div></div>
2	Singapur	83	<div></div>
3	India	75	<div></div>
4	Nepal	57	<div></div>
5	Estados Unidos	40	<div></div>

☐ Incluir regiones con un volumen de búsquedas bajo

< Mostrando 1-5 de 61 regiones >

MARCO DE REFERENCIA...

Algunos conceptos importantes a considerar



TIPOS DE DATOS

Estructurados



Referencia	Fecha Alta	Tipo	Operación	Provincia	Superficie	Precio Venta	Fecha Venta	Vendedor
1	01/01/17	Parking	Alquiler	Lleida	291 m2	2.133.903,00 €	19/06/17	Carmen
2	01/01/17	Local	Venta	Girona	199 m2	1.945.424,00 €	19/04/17	Pedro
3	01/01/17	Oficina	Alquiler	Girona	82 m2	712.416,00 €	08/11/17	Joaquín
4	02/01/17	Parking	Alquiler	Girona	285 m2	1.815.450,00 €	27/04/17	Jesús
5	02/01/17	Suelo	Venta	Tarragona	152 m2	1.138.024,00 €	10/07/17	María
6	03/01/17	Industrial	Alquiler	Girona	131 m2	953.156,00 €	05/09/17	Pedro
7	03/01/17	Parking	Alquiler	Tarragona	69 m2	406.686,00 €	07/06/17	Pedro
8	03/01/17	Oficina	Venta	Girona	235 m2	2.158.475,00 €	31/10/17	Jesús
9	04/01/17	Piso	Alquiler	Lleida	108 m2	1.024.380,00 €	28/12/17	Jesús
10	04/01/17	Parking	Venta	Lleida	299 m2	2.042.768,00 €	06/10/17	Joaquín

NO
Estructurados



¿QUÉ ES BIG DATA?



PARA ALGUNOS...

Big Data es: lo que no cabe en Excel



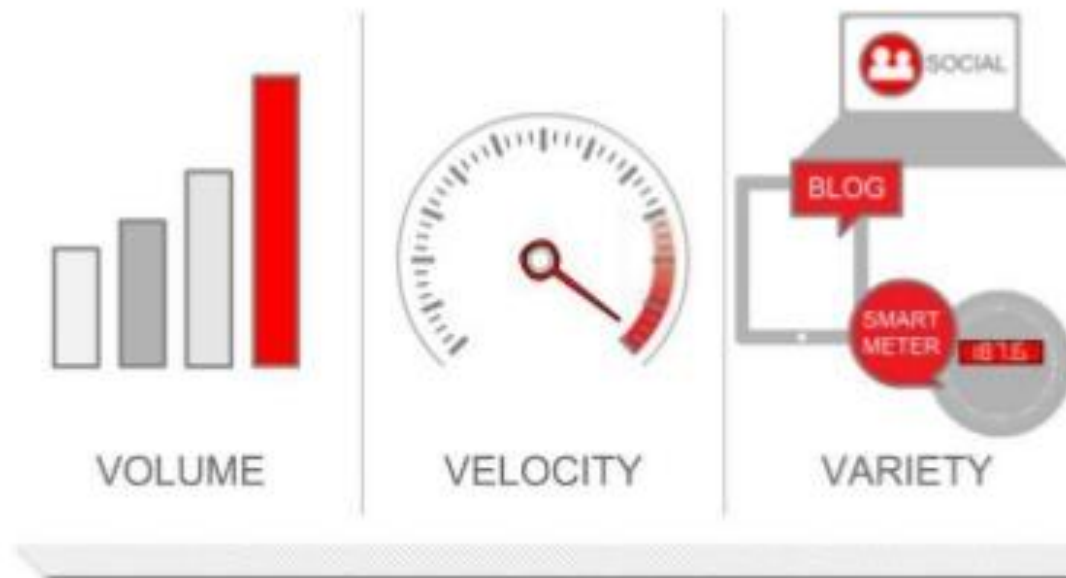
PARA ALGUNOS...

Big Data es: Datos medianos (10 GB - 1 TB), Big data más de 1 TB ■



BIG DATA, ¿QUÉ ES?

- Son activos de información caracterizados por su alto Volumen, Velocidad y Variedad (3 V's), que demandan soluciones innovadoras y eficientes.
- Más que un todo, podría verse como una suma de partes



¿QUIÉN GENERA BIG DATA?

- Sensores que recogen información climática
- Publicaciones en las redes sociales
- Imágenes y vídeos digitales
- Registros de compra y transacciones
- Señales de GPS de los móviles, entre otros
- Internet de las Cosas (IoT)
- Smart Cities





Políticas públicas

Construcción

Industria

Ventas al detalle

Telecomunicaciones

Logística

Servicios
financieros

BIG DATA

Ventas al detalle

Transporte aéreo

Salud

Producción de alimentos

Manufactura

PROBLEMAS DEL BIG DATA

- Análisis de la información (Variedad y cantidad de datos)
- Almacenamiento de los datos (Volumen de datos)
- Velocidad de acceso a los datos (Infraestructura)

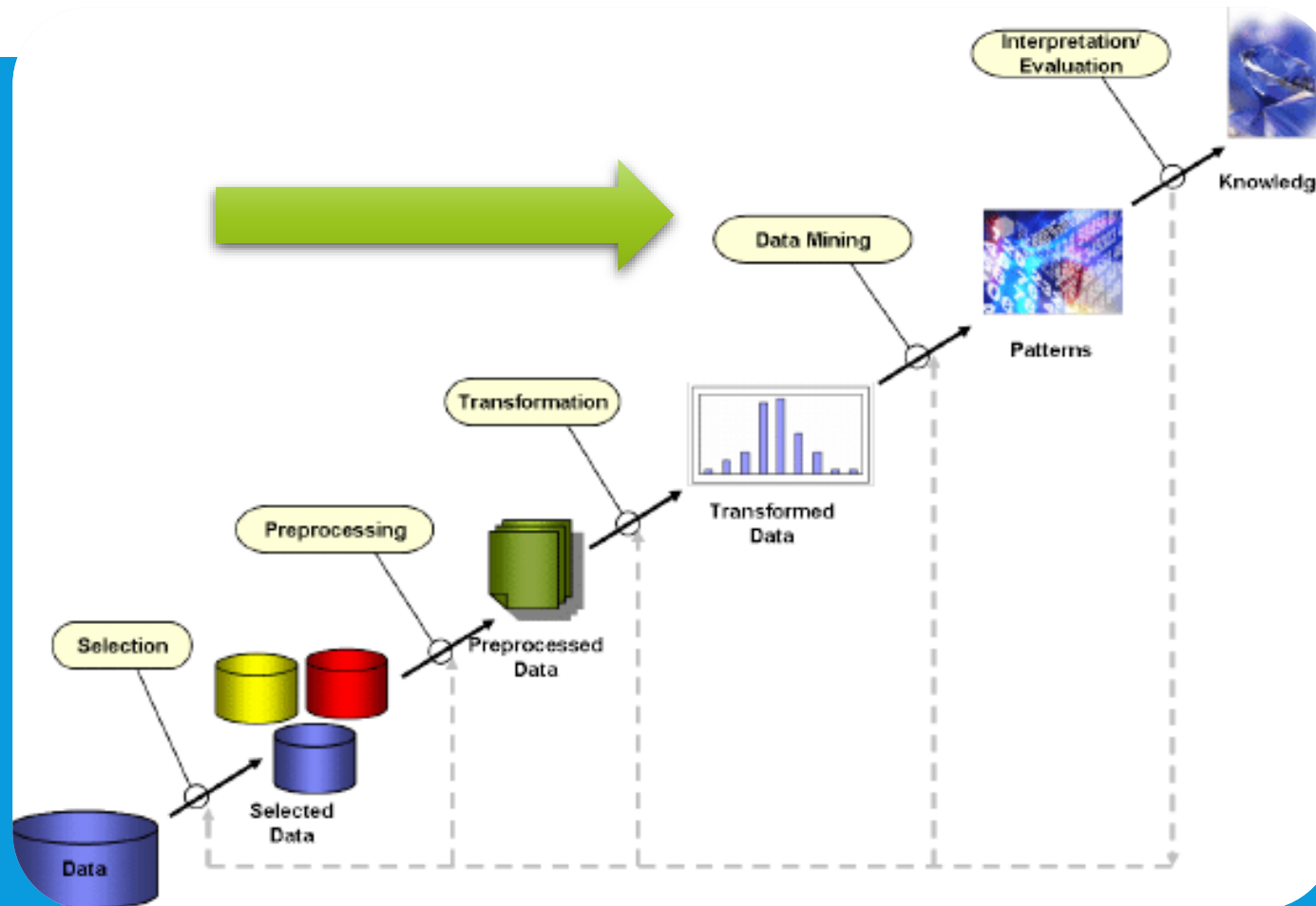
¿CÓMO SE ANALIZA EL BIG DATA?

- Se requiere de Minería de Datos, pero...

¿Qué es Minería de Datos?

Extracción de patrones o de información no trivial, implícita, previamente desconocida y potencialmente útil de las grandes bases de datos

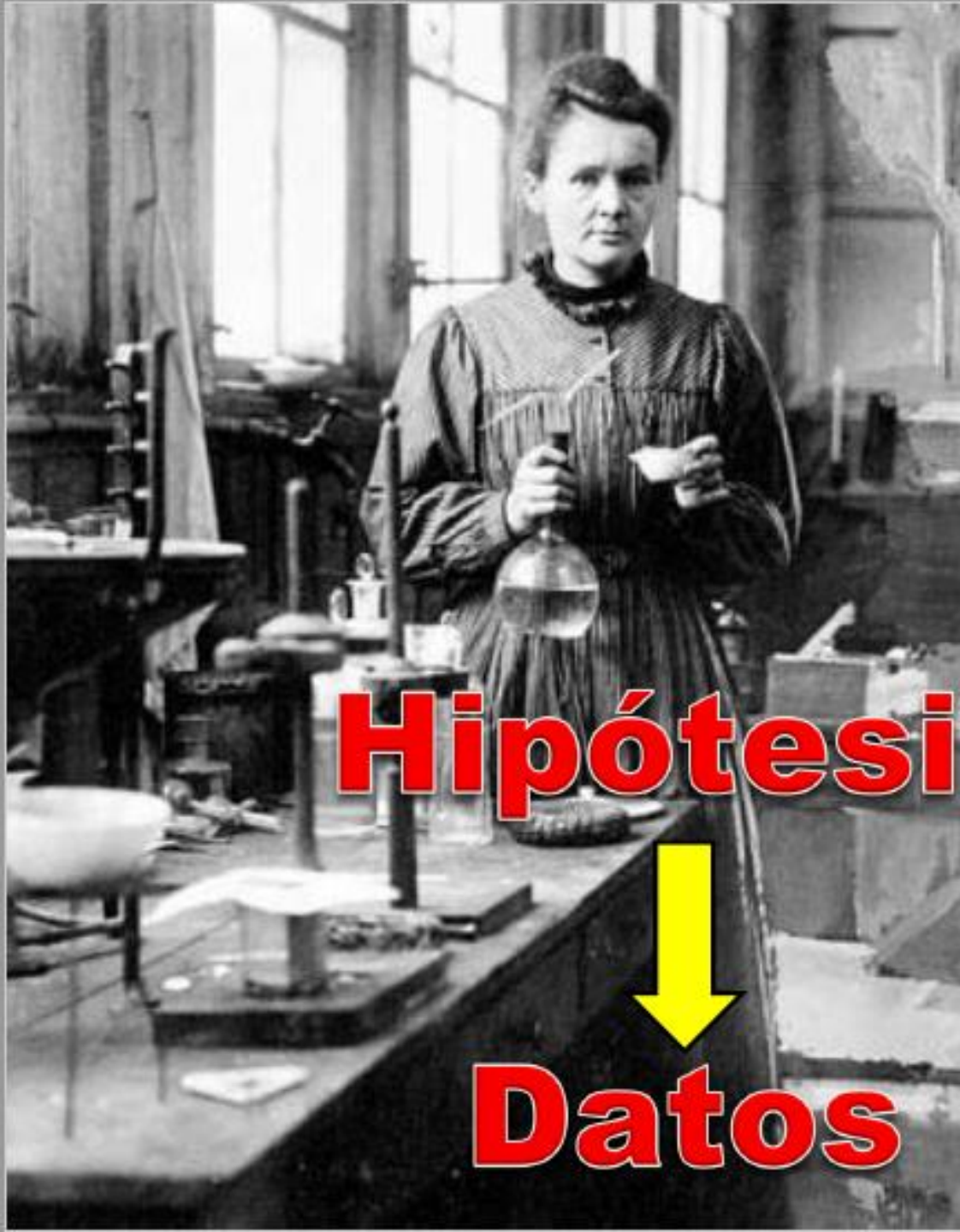
PROCESO DE KDD: KNOWLEDGE DISCOVERY IN DATABASES



Fuente: Explorations of the BDI Multi-agent support for the Knowledge Discovery in Databases Process - Scientific Figure on ResearchGate.
Available from: https://www.researchgate.net/Steps-in-the-KDD-process_fig1_236373188 [accessed 16 Nov, 2018]

MINERÍA DE DATOS VS ESTADÍSTICA

- La estadística generalmente analiza **Muestras** de datos, para luego hacer inferencias a toda la población, mientras que la minería de datos pretende buscar información útil usando **TODA** la base de datos.
- La estadística en la mayoría de los casos, supone que los datos se comportan de acuerdo a ciertas distribuciones (Normal, binomial, geométrica, Poisson, etc), mientras que la MD usa técnicas más exploratorias que vienen de la IA.



Hipótesis



Datos



Hipótesis



Datos

MINERÍA DE DATOS VS MACHINE LEARNING

- “Machine Learning” es un área de la Inteligencia Artificial (IA) que trata sobre escribir programas que puedan aprender.
- El enfoque de “Machine Learning” está más enfocado a la eficiencia de los algoritmos, mientras que el enfoque en MD es la interpretación de los resultados.
- Ambas, se dividen en dos: aprendizaje supervisado (learns by example) y aprendizaje no supervisado

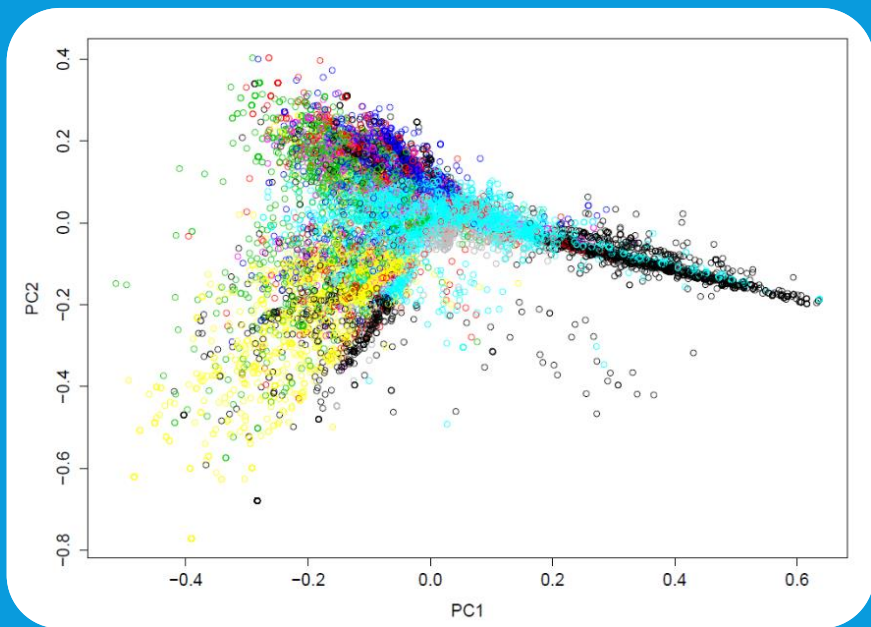
MACHINE LEARNING VS INTELIGENCIA ARTIFICIAL

- Machine Learning: Un experto toma datos y los clasifica, el **modelo** aprende con estos datos (hay métodos), con nuevos datos se genera una predicción.
- Inteligencia Artificial: ya no es un modelo que reconozca un patrón, sino un agente que tiene objetivo (manejar una cartera), son **modelos** con una conexión con el entorno y actuaciones (respuestas de actores), que toma el agente. Es inteligencia aumentada (scoring de crédito) para ayudar a alguien a colocar créditos. Otro ejemplo es google, que hace más fácil la búsqueda para la eficiencia del usuario.

¿QUÉ TIPO DE TAREAS SE ABORDAN DESDE LA MINERÍA DE DATOS?

Descriptivas

- ACP
- Clustering
- K-means
- Cubos OLAP



Predictivas

- Redes Neuronales
- Árboles de decisión
- Bayesiano
- SVM
- Bosques Aleatorios
- KNN



APLICACIONES DE LA MINERÍA DE DATOS

- Retención de clientes: ¿Cuáles son clientes potenciales a irse para la competencia?.
- Patrones de compra: cuando un cliente compra un producto, ¿Cuál otro producto le podría interesar?.
- Detección de fraude: ¿Cuáles transacciones podría ser parte de un fraude?
- Manejo del riesgo: ¿A qué cliente se le podría aprobar un préstamo?.
- Segmentación de clientes: ¿Quiénes son mis clientes?.
- Predicción de ventas: ¿Cuánto voy a vender el próximo mes?

APLICACIONES DE LA MINERÍA DE DATOS

- Análisis de opinión: ¿quién lidera las encuestas?.
- Análisis del sentimiento: ¿Qué opina la gente sobre mi producto con respecto al de la competencia?.
- Análisis de imagen: ¿está enferma o sana la planta?
- Salud: ¿este paciente está potencialmente predispuesto a padecer de cáncer?.
- Educación, la bolsa de valores, desastres naturales, reconocimiento de voz, reconocimiento facial, etc.

¿QUÉ TIPO DE DATOS SE PODRÍAN ANALIZAR CON MD?

- Bases de datos relacionales
- Bodegas de datos
- Bases de datos transaccionales
- Bases de datos orientadas a objetos y simbólicas
- Bases de datos espaciales: Sistema de Información Geográfica (GIS)
- Series cronológicas de datos y datos temporales
- Bases de datos de texto
- Bases de datos de multimedia

¿DÓNDE ALMACENAR BIG DATA? CLOUD COMPUTING



Amazon **Redshift**

AMAZON Redshift



dashDB

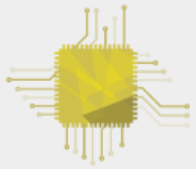


Azure

¿QUE ES CIENCIA DE LOS DATOS?

Data science creates meaning from vast amounts of complex data.

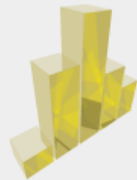
Using automated analytical methods, it reveals patterns humans alone might never see. Data science combines aspects of:



COMPUTER
SCIENCE



APPLIED
MATHEMATICS



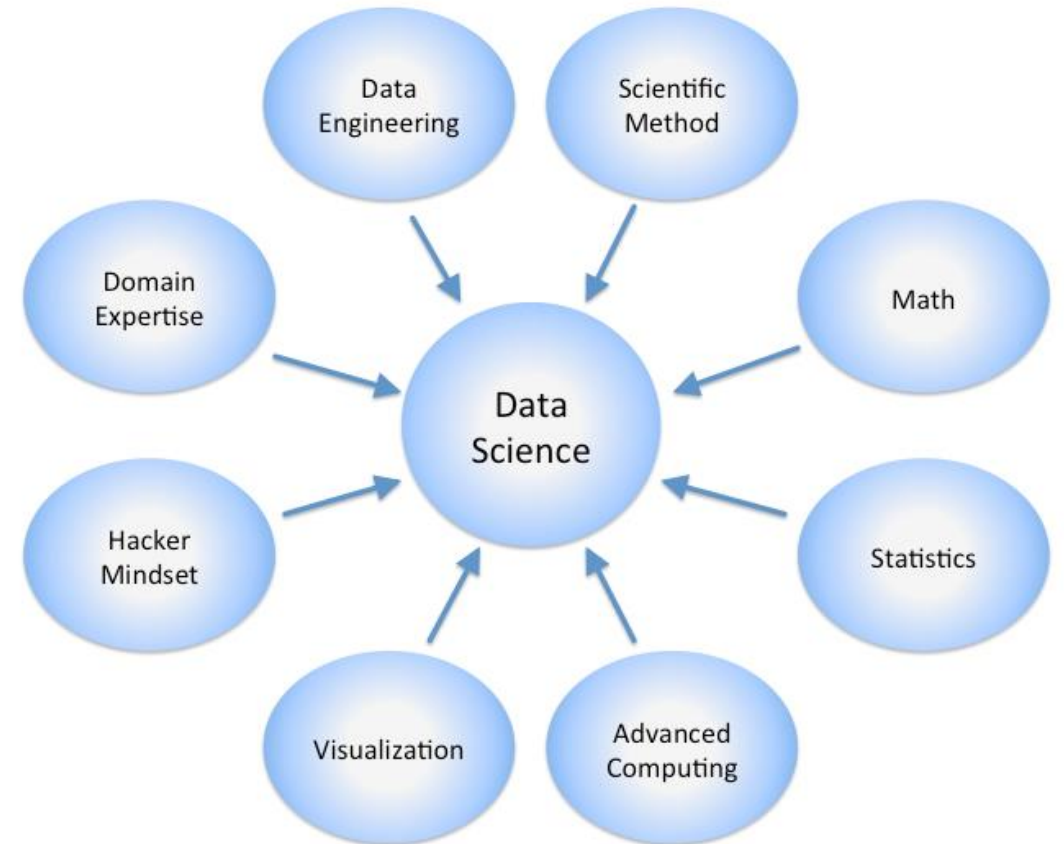
STATISTICS








MACHINE
LEARNING



VISUALIZATION









Employer type	Salary/Income
Self (12)	 \$141K
Company (303)	 \$107K
Government/Non-profit (21)	 \$90K
Academic/University (34)	 \$70K
Student (4)	 \$24K



INGRESO PROMEDIO
PARA UN CIENTÍFICO
DE DATOS, SEGÚN
EMPLEADOR

Ingreso promedio
para un Científico de
Datos, según puesto
















Analytic Role	Salary
Manage teams which analyze data (60)	 \$158K
other role (16)	 \$116K
Data Scientist/Data Miner - analyze data yourself (206)	 \$101K
Academic Researcher (23)	 \$75K
Data Analyst/Business Analyst (support data analysis) (62)	 \$72K
Student (21)	 \$26K

Fuente: <https://www.kdnuggets.com/polls/2015/salary-analytics-data-science-data-mining.html>

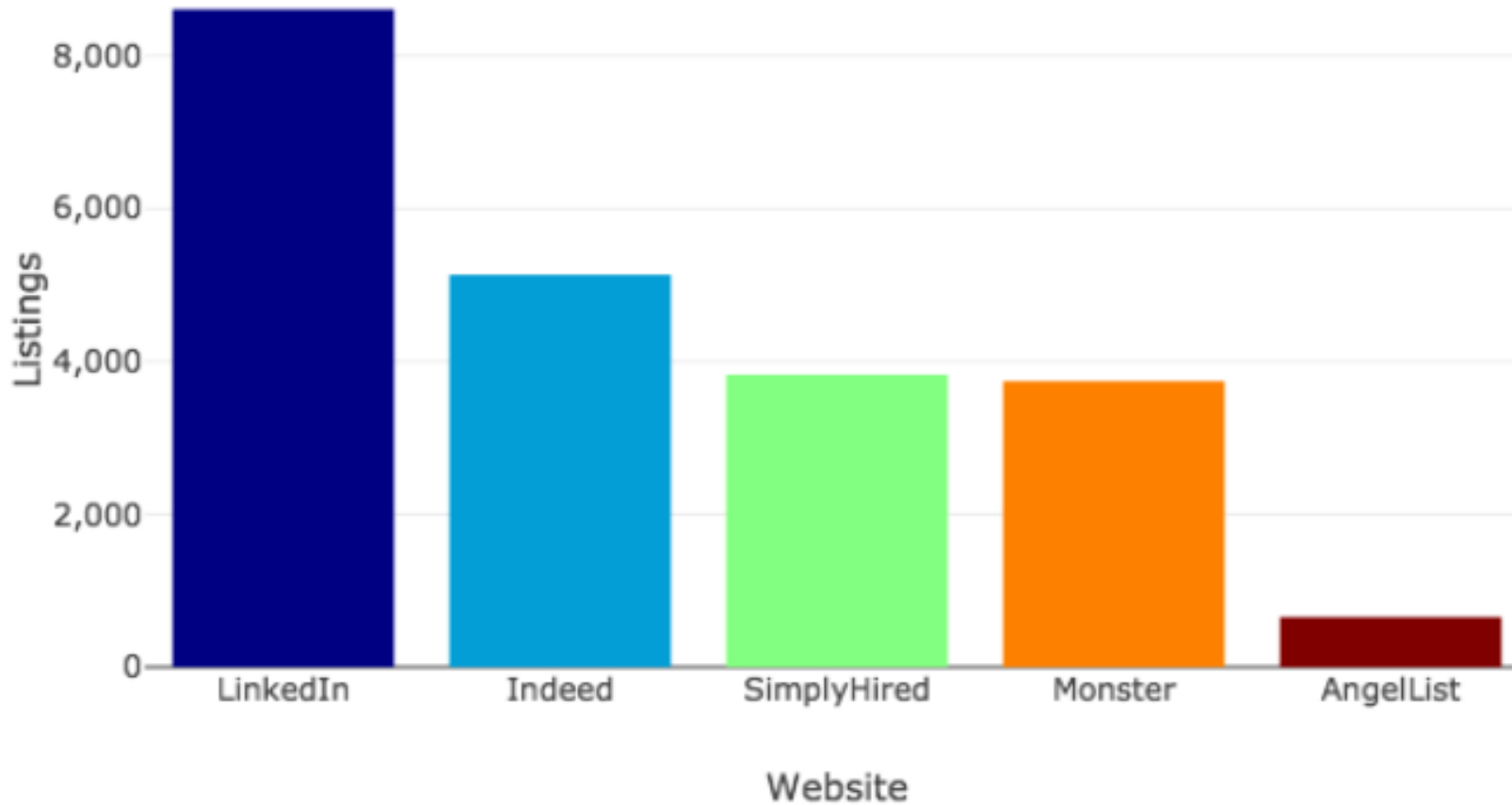
SALARIO (INGRESO) PROMEDIO PARA UN CIENTÍFICO DE DATOS, POR REGIÓN



Region	Employer Type	Salary or Income
US/Canada (221)	Company/Self	 \$131K
	Academic, Government	 \$103K
Australia/NZ (7)	Company/Self	\$172K
	Academic, Government	 \$60K
W. Europe (78)	Company/Self	 \$82K
	Academic, Government	 \$54K
E. Europe (23)	Company/Self	 \$34K
	Academic, Government	 \$24K
Africa/Middle East (7)	Company/Self	 \$62K
	Academic, Government	 \$105K
Latin America (12)	Company/Self	 \$59K
	Academic, Government	 \$58K
Asia (22)	Company/Self	 \$47K
	Academic, Government	 \$59K

Fuente: <https://www.kdnuggets.com/polls/2015/salary-analytics-data-science-data-mining.html>

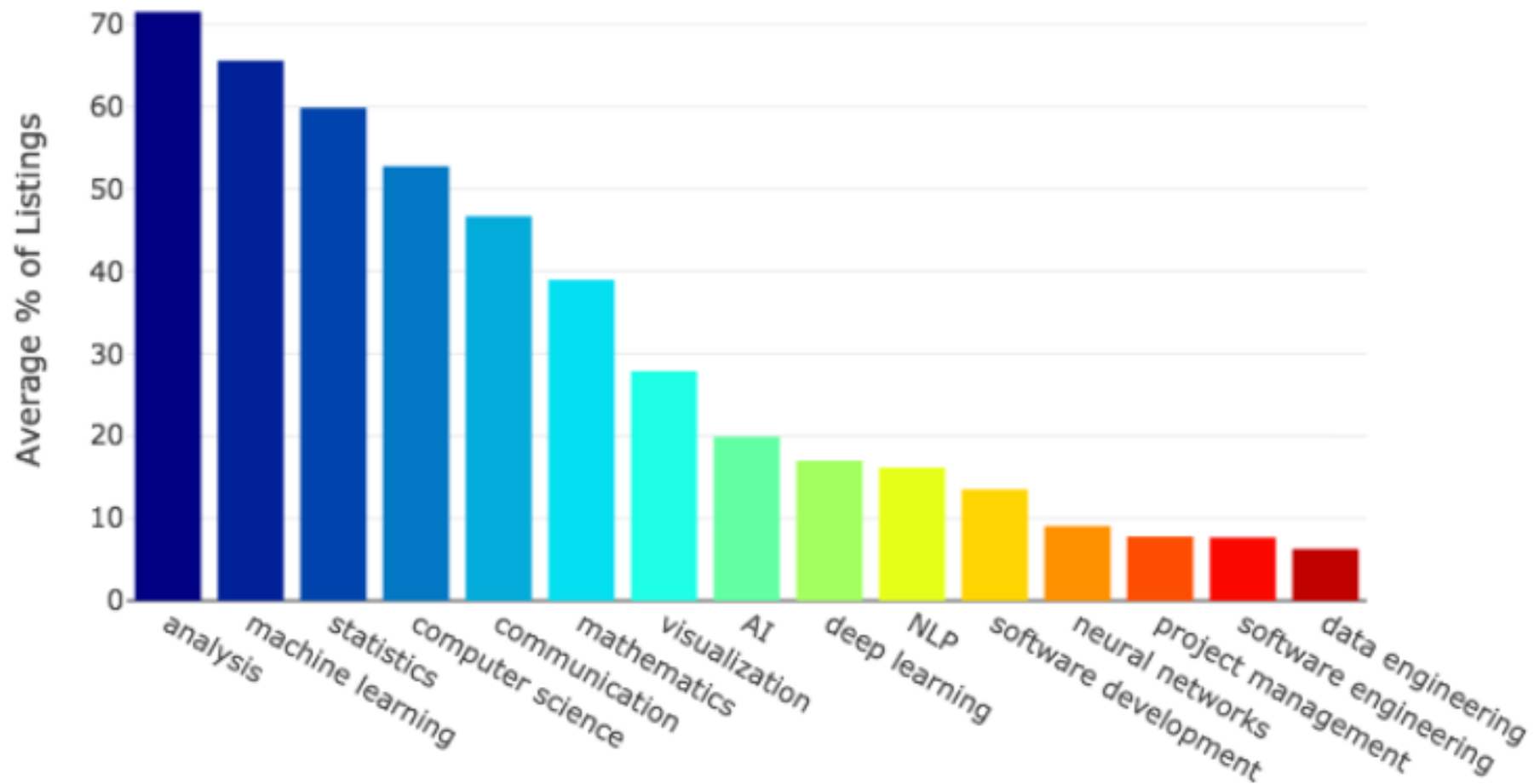
¿DÓNDE BUSCAR UN EMPLEO EN CIENCIA DE DATOS?



Fuente: <https://www.kdnuggets.com/polls/2015/salary-analytics-data-science-data-mining.html>

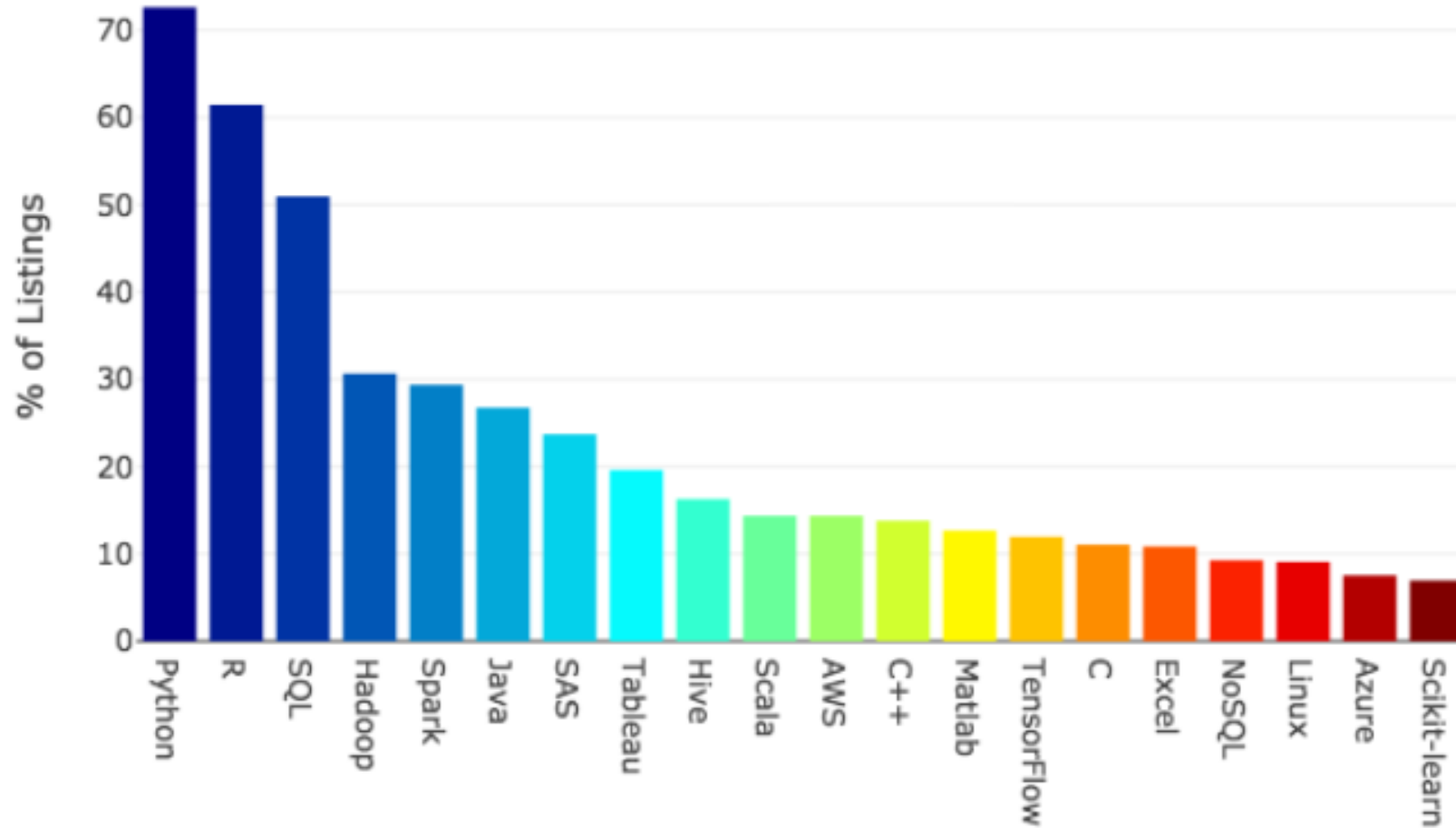
Website

HABILIDADES GENERALES EN EL TRABAJO PARA UN CIENTÍFICO DE DATOS



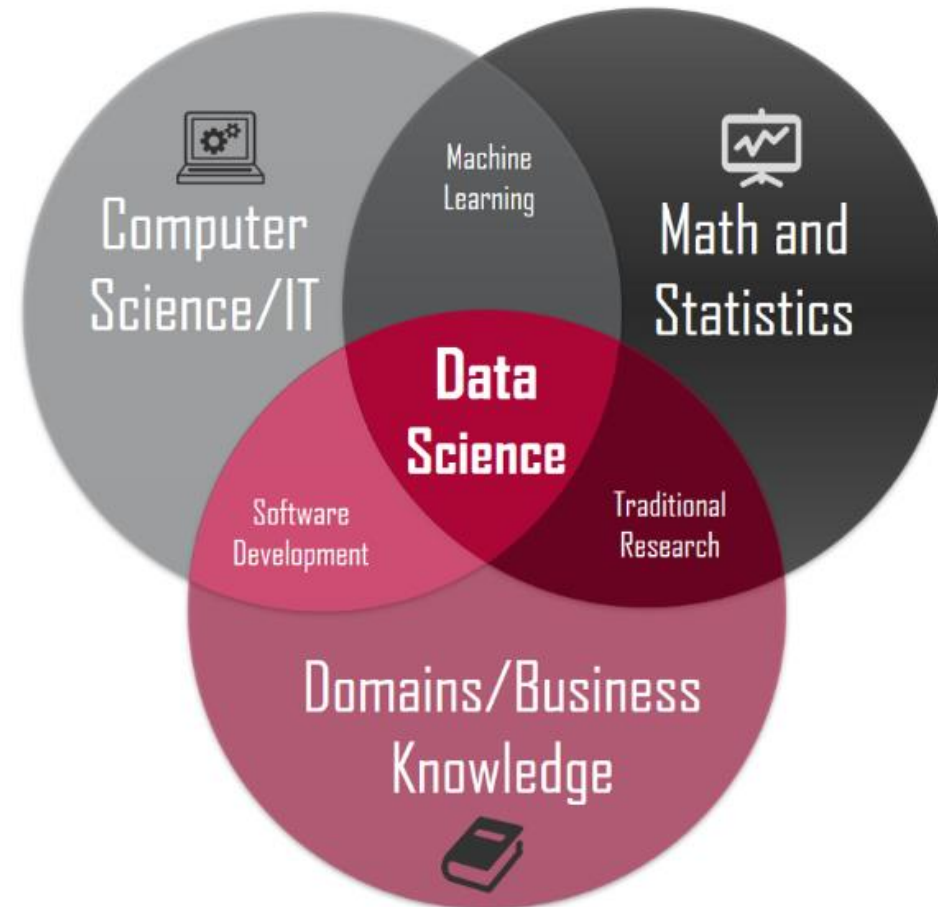
Fuente: <https://www.kdnuggets.com/polls/2015/salary-analytics-data-science-data-mining.html>

TOP 20: HABILIDADES TECNOLÓGICAS REQUERIDAS EN TRABAJOS PARA UN CIENTÍFICO DE DATOS



Fuente: <https://www.kdnuggets.com/polls/2015/salary-analytics-data-science-data-mining.html>

DATA SCIENTIST SKILLS



¿QUÉ ES UN CIENTÍFICO DE DATOS (DATA SCIENTIST)?

- Los científicos de datos tienen habilidades **matemáticas, de computación y analíticas** estelares. La **curiosidad** también es una característica esencial de un científico de datos. Se acercan a los datos estructurados y no estructurados que ingresan desde fuentes como sensores, la Web, teléfonos inteligentes y lectores de tarjetas de crédito, por nombrar solo algunos, con el deseo del explorador de **descubrir** lo que otros podrían no ver. También se destacan en la presentación de hallazgos complicados para que tanto expertos como no expertos puedan hacer uso de los conocimientos adquiridos a partir de los datos.
- Los científicos de datos trabajan en una amplia gama de sectores: tecnología, salud, finanzas, gestión, gobierno. También tienen una gran demanda. Un estudio del McKinsey Global Institute, por ejemplo, proyecta que para 2018 los Estados Unidos enfrentarán una escasez de aproximadamente **140,000 a 190,000** personas con las habilidades de los científicos de datos.

Fuente: CDS, NY University <https://cds.nyu.edu/academics/>



OTROS ACERCAMIENTOS

Jonathan Ma Swag, youtuber de Tecnología (Joma Tech), quien ha trabajado en LinkedIn, Facebook, Microsoft and BuzzFeed dice:

“Ciencia de datos no es sobre crear modelos complejos, no es sobre hacer sorprendentes visualizaciones, no es sobre escribir código, sino, es acerca de usar “La Data” para crear impacto en su compañía, tanto como sea posible. Probablemente se requiera escribir código o crear modelos complejos, pero el trabajo del Científico de Datos es resolver un problema para su compañía usando los datos y las herramientas que se deban utilizar, no debería ser el foco de la preocupación”



CIENCIA DE DATOS VS INGENIERÍA DE DATOS

DATA Engineer

Develops, constructs, tests, and maintains architectures. Such as databases and large-scale processing systems.



DataCamp
Learn Data Science By Doing

DATA Scientist

Cleans, massages and organizes (big) data. Performs descriptive statistics and analysis to develop insights, build models and solve a business need.

\$124,000

MEDIUM MARKET

MIN \$34k

MAX \$341k

\$135,000

MEDIUM MARKET

MIN \$43k

MAX \$364k

Languages, Tools & Software



DATA ENGINEERING: EL PRIMO MÁS CERCANO DE DATA SCIENCE

THE DATA SCIENCE HIERARCHY OF NEEDS

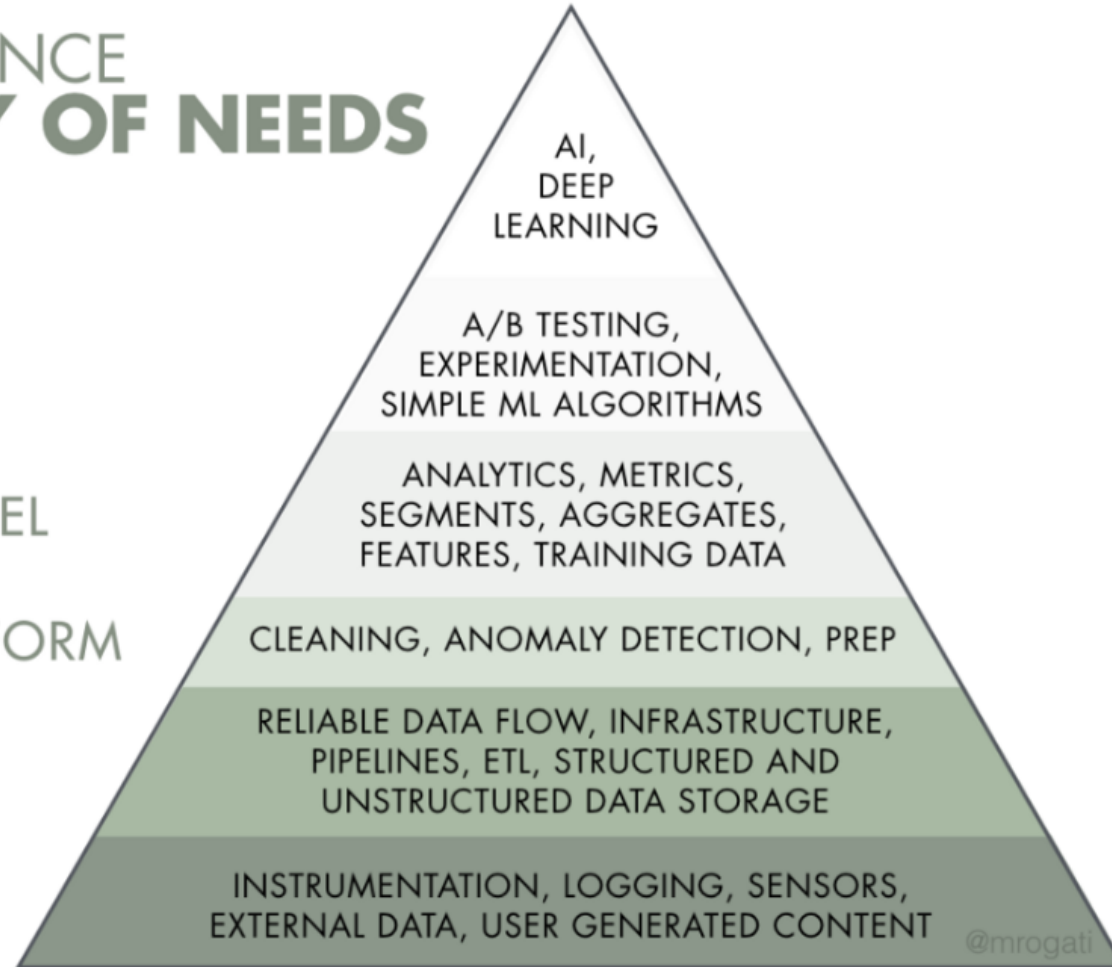
LEARN/OPTIMIZE

AGGREGATE/LABEL

EXPLORE/TRANSFORM

MOVE/STORE

COLLECT



Ciencia de datos

Ingeniería de datos

Ingeniería de software

Fuente: Monica Rogati's fantastic Medium post "The AI Hierarchy of Needs"

The most desired skill in data science



¿CUÁL ES LA MAYOR VACÍO DE HABILIDADES EN LA CIENCIA DE DATOS SEGÚN LOS GERENTES DE CONTRATACIÓN QUE BUSCAN CONTRATAR LOS RECIENTES GRADUADOS? UNA PISTA: NO ES LA CODIFICACIÓN.

The most desired skill in data science



RESPUESTA: PENSAMIENTO CRÍTICO

1. Habilidad para generar las preguntas
2. Capacidad de cuestionar a los datos



All you've done is chisel all day! Do something useful,
like helping your brother drag those rocks up the hill.

Fuente: <https://writereflections4u.com/wp-content/uploads/Creative-vs.-Critical-Thinker-1080x675.png>

LINK'S DE INTERÉS

<http://www.datalatam.com/>



<http://www.conectar2019.ucr.ac.cr/>

LINK'S DE INTERÉS

<https://www.meetup.com/es-ES/San-Carlos-R-User-Group/>

SCRUG
San Carlos R User Group

Webinarios (en línea)
cada último Jueves del Mes
19:00 - 20:00 UTC-06

<https://meetup.com/san-Carlos-R-User-Group>



San Carlos R User Group
San Jose, Costa Rica · 506 miembros · Grupo público

 Organizado por
Frans van Dunné

Compartir:    

[Sobre nosotros](#) [Meetups](#) [Miembros](#) [Fotos](#) [Conversaciones](#) [Eres miembro](#)

Meetup

Data  Latam

www.datalatam.com

Data Latam Meetup

San Jose, Costa Rica · 365 miembros · Grupo público



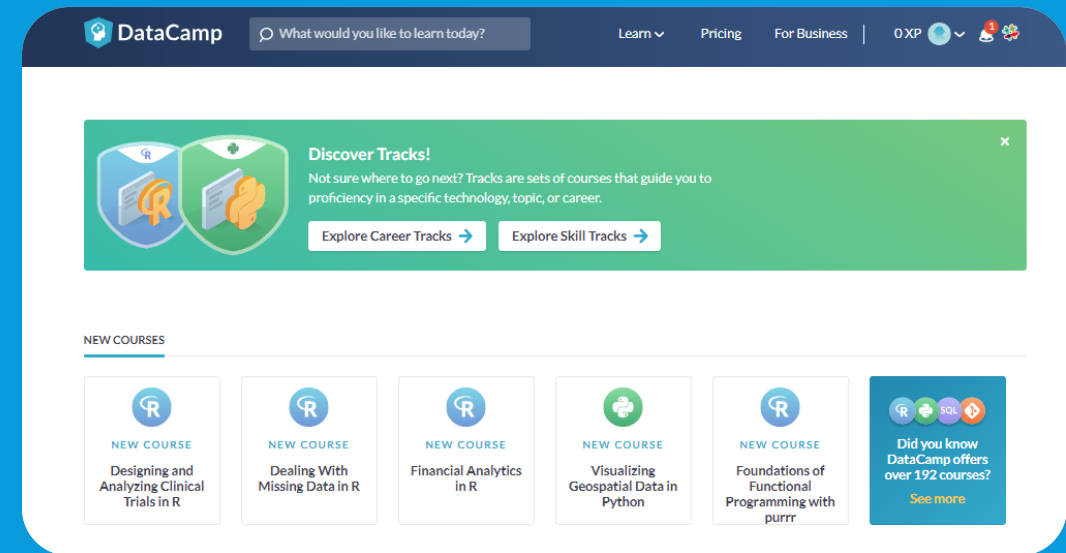
Organizado por
Frans van Dunné y otras 2 personas

Compartir:    

<https://www.meetup.com/es-ES/DataLatam/>

LINK'S DE INTERÉS

<https://www.datacamp.com/home>



<https://www.kdnuggets.com/>

