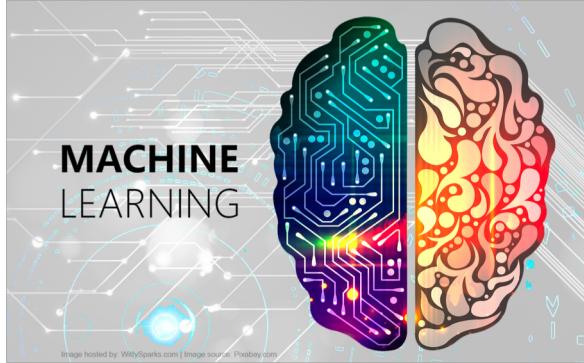


# Conceptos Básicos



*Desde la Estadística hasta la  
Ciencia de Datos, pasando  
por el concepto de "Big Data"*

# AGENDA

1. Historia del Concepto: Ciencia de Datos
2. Evolución del Concepto: Ciencia de Datos
3. Estadística, Machine-Learning, Data-Mining, Big Data, Ingeniería de Datos y Ciencia de Datos
4. Herramientas más utilizadas en Ciencia de Datos.

# Conceptos en Ciencia de Datos



PROMIDAT  
IBEROAMERICANO

# Evolución del concepto “Ciencia de Datos”

- **1750 – 1970:** Probabilidad y Estadística
- **1970 – 1990:** Análisis de Datos (Exploratory Data Analysis), SQL
- **1990 – 2000:** OLAP (Online Analytical Processing)
- **2000 – 2005:** Minería de Datos, Business Intelligence
- **2005 – 2011:** Modelos Predictivos, Analytics, Machine Learning
- **2011 – 2015:** Big Data, Big Data Analytics.
- **2015 – presente:** Ciencia de Datos o Ingeniería de Datos

# ★ ¿Qué es Estadística?

## ★ Según el diccionario:

- Ciencia que utiliza conjuntos de datos numéricos para obtener, a partir de ellos, inferencias basadas en el cálculo de probabilidades.
- Estudio que reúne, clasifica y recuenta todos los hechos que tienen una determinada característica en común, para poder llegar a conclusiones a partir de los datos numéricos extraídos.





# ¿Qué es Estadística?

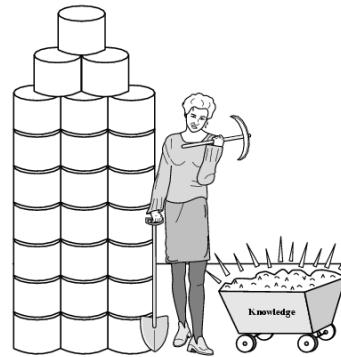
## Según el Wikipedia:

- La estadística es una rama de las matemáticas y una herramienta que estudia usos y análisis provenientes de una muestra representativa de datos, que busca explicar las correlaciones y dependencias de un fenómeno físico o natural, de ocurrencia en forma aleatoria o condicional.



# 🌟 ¿Qué es Minería de Datos?

- Extracción de información o de patrones (no trivial, implícita, previamente desconocida y potencialmente útil) de grandes bases de datos.





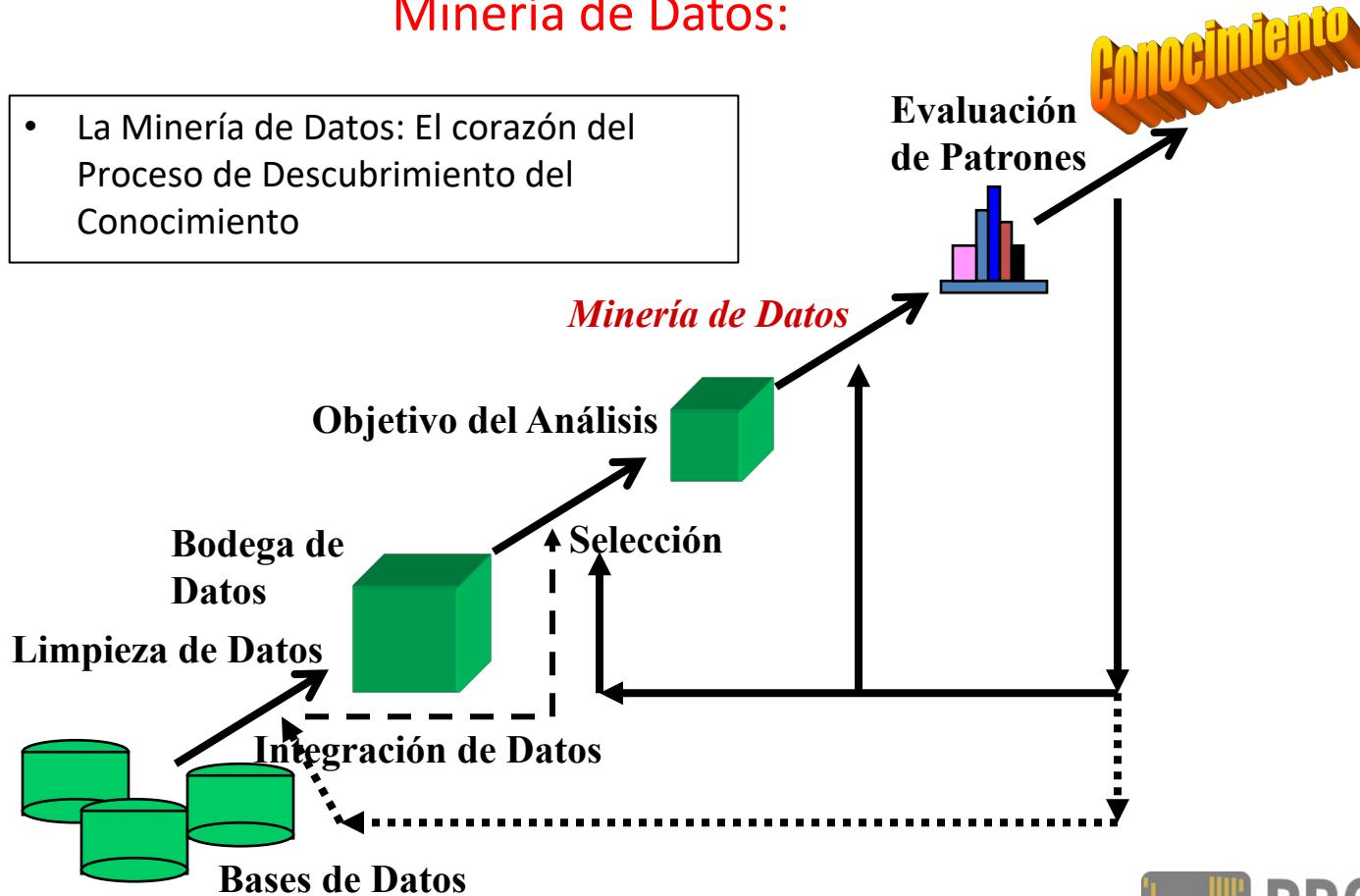
# ¿Qué es Minería de Datos?

- Es analizar datos para encontrar patrones ocultos usando medios automatizados.



# Minería de Datos:

- La Minería de Datos: El corazón del Proceso de Descubrimiento del Conocimiento



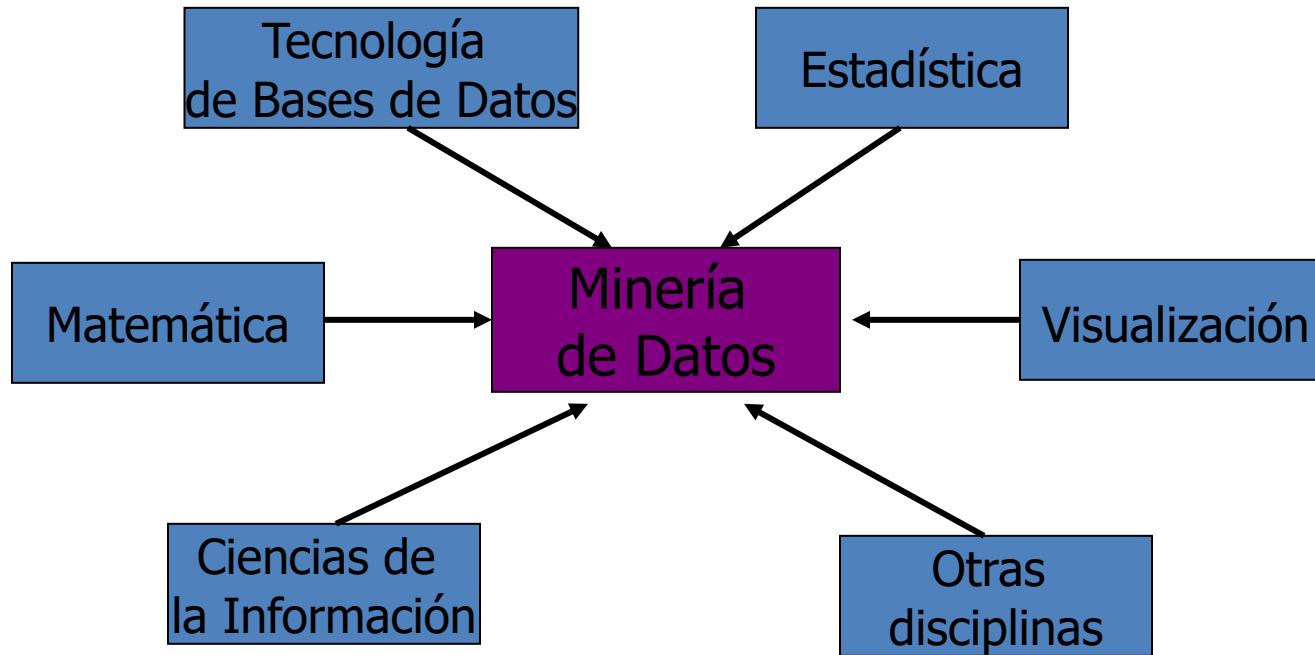
## Minería de Datos versus Estadística

- ✓ La estadística generalmente analiza muestras de datos para luego hacer inferencia a toda la población, mientras que la minería de datos pretende buscar información útil usando toda la base datos.
  
- ✓ La estadística en la mayoría de los casos supone que los datos se comportan de acuerdo a ciertas distribuciones de probabilidad (normal, binomial, geométrica, Poisson, etc), mientras que la minería de datos usa técnicas mucho más exploratorias que vienen de la IA, o del “Analyse des Données”.

- Minería de Datos versus Machine Learning

- “*Machine Learning*”: es un área de la Inteligencia Artificial (IA) que trata sobre como escribir programas puedan aprender.
- En “Data Mining” es usualmente usado para predicción y clasificación.
- Se divide en dos: aprendizaje supervisado (learns by example) y aprendizaje no supervisado.

# La Minería de Datos: Confluencia de Múltiples Disciplinas



# Tareas de la Minería de Datos

- **Exploratorias:**

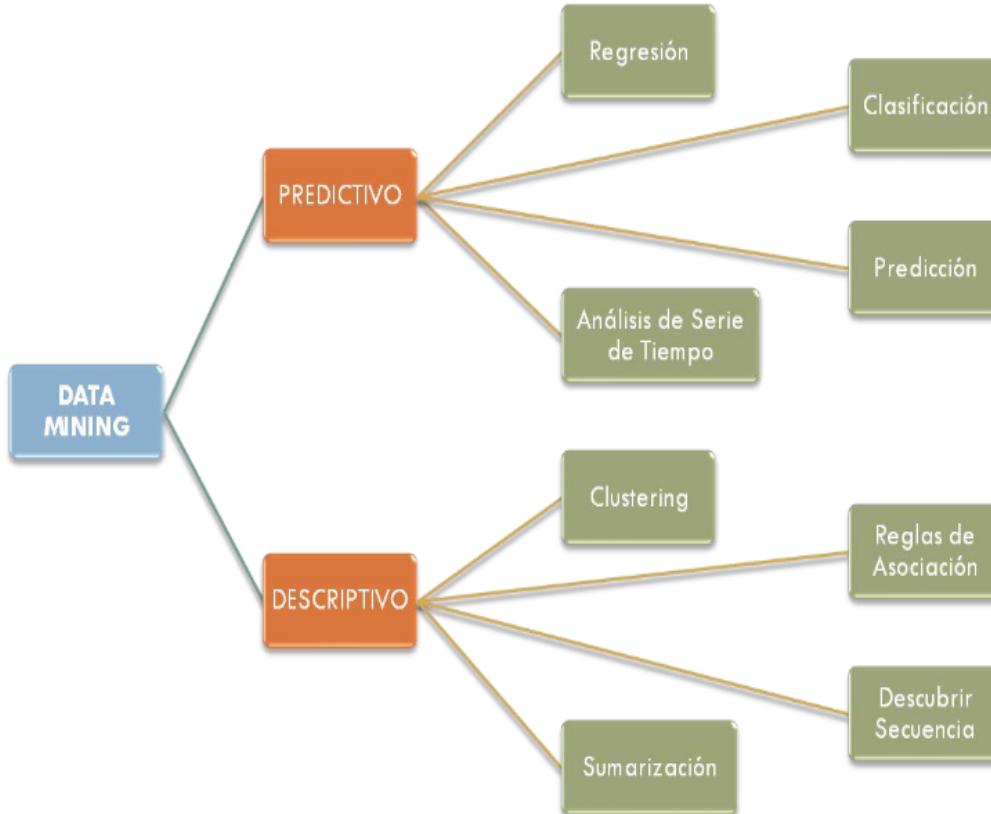
- Buscar patrones humano-interpretables que describen los datos



- **Predictivas:**

- Utiliza algunas de las variables para predecir los valores futuros desconocidos de la misma variable o bien de otras variables





# Minería de Datos: ¿En qué tipo de datos?

- Bases de datos relacionales
- Bodegas de datos
- Bases de datos transaccionales
- Bases de datos orientadas a objetos y simbólicas
- Bases de datos espaciales Sistemas de Información Geográfica - GIS
- Series cronológicas de datos y los datos temporales
- Bases de datos de texto
- Bases de datos multimedia
- www (web mining)

# ¿Qué NO es Minería de Datos?



# ¿Qué NO es Minería de Datos?

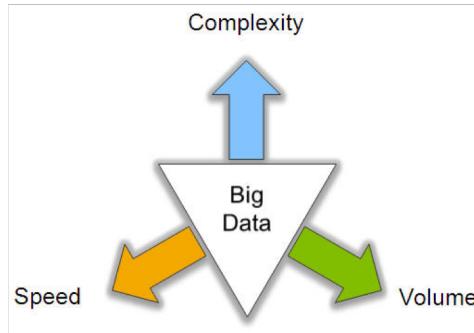
- En general la Minería de Datos NO se basa en modelos ***Determinísticos***.
- Un modelo ***Determinístico*** es un modelo matemático donde las mismas entradas producirán invariablemente las mismas salidas, no contemplándose la existencia del azar ni el principio de incertidumbre.

# ¿Qué NO es Minería de Datos?

- En general la Minería de Datos se basa en modelos **Probabilísticos**.
- Un modelo **Probabilístico** es un modelo matemático que nos ayuda a predecir la conducta de futuras repeticiones de un experimento aleatorio mediante la estimación de una probabilidad de ocurrencia de dicho evento concreto.

# ¿Qué es Big Data?

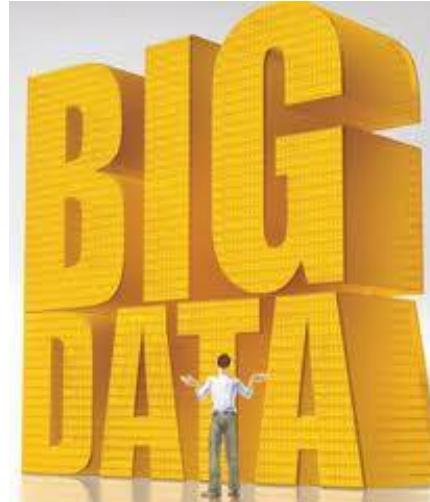
- **Big Data** es el término que popularmente designa un crecimiento, disponibilidad y uso exponenciales de la información estructurada y no estructurada.
- Debido al gran **Volumen, Variedad y Velocidad** que tienen los datos actualmente (las 3 V's)



# La Traducción Correcta

## ➤ Big Data

- Datos Masivos
- Datos Gigantes
- Macrodatos
- Datos demasiado grandes
- Grandes volúmenes de datos



**BigData,** = información para  
BI y analítica      tomar decisiones

# ¿Es una moda?

- **Big Data:** el término de moda en el mundo de la informática y de la Administración de Negocios (MBA)
- **Big Data** no es fácil de definir ya que, en mi opinión, es un nuevo término “inventado para el marketing”
- Durante 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 aquéllos que no se sumen a este nuevo movimiento se quedarán “obsoletas” en cuanto a la capacidad de reacción en sus decisiones, perdiendo competitividad y oportunidades de negocio contra su competencia.

# Breve Historia de Big Data

- Siempre es difícil identificar como surgen los conceptos y paradigmas. “Big Data” no supone una excepción a esta regla, siendo difícil identificar si surge como consecuencia o acompañante de otros conceptos como el “Open Data”
- La popularización del término ***Big Data*** viene, sin duda, ligada al documento del concepto publicado por McKinsey Global Institute en Junio de 2011, en el cual se define como “*conjuntos de datos cuyo tamaño va más allá de la capacidad de captura, almacenado, gestión y análisis de las herramientas de base de datos tradicionales*”

# Breve Historia de Big Data

May 2011

## McKinsey Report

**Big data: The next frontier  
for innovation, competition,  
and productivity**



# Breve Historia de Big Data

## Box 1. What do we mean by "big data"?

"Big data" refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze. This definition is intentionally subjective and incorporates a moving definition of how big a dataset needs to be in order to be considered big data—i.e., we don't define big data in terms of being larger than a certain number of terabytes (thousands of gigabytes). We assume that, as technology advances over time, the size of datasets that qualify as big data will also increase. Also note that the definition can vary by sector, depending on what kinds of software tools are commonly available and what sizes of datasets are common in a particular industry. With those caveats, big data in many sectors today will range from a few dozen terabytes to multiple petabytes (thousands of terabytes).



# Los datos (la vida) en la nube: Big Data y Cloud Computing

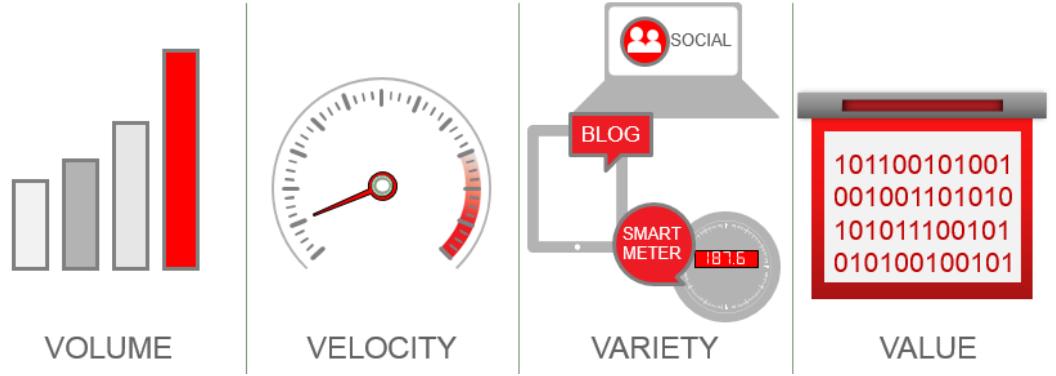
La computación en la nube, concepto conocido también bajo los términos servicios en la nube, informática en la nube, nube de cómputo o nube de conceptos, del inglés "cloud computing", es un paradigma que permite ofrecer servicios de computación a través de Internet.



- ◀ ★ Favorites
- Desktop
- Downloads
- Recent places
- SkyDrive
- Google Drive
- Dropbox

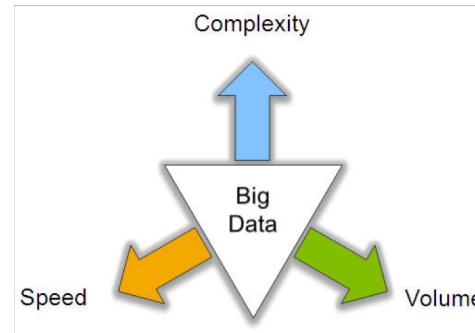
# ¿Qué es Big Data?

“Big data” son activos de información caracterizados por su alto **volumen, velocidad y variedad**, que demandan soluciones innovadoras y eficientes de procesado para la mejora del conocimiento y toma de decisiones en las organizaciones.



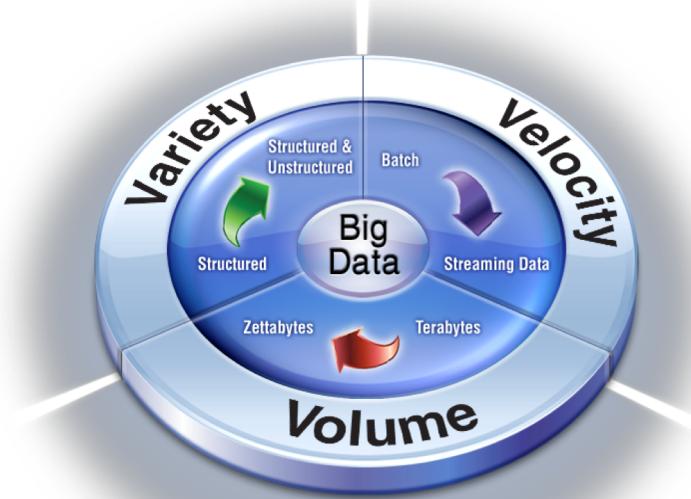
# ¿Qué es Big Data?

- **Big Data** es el término que popularmente designa un crecimiento, disponibilidad y uso exponenciales de la información estructurada y no estructurada.
- Debido al gran **Volumen, Variedad y Velocidad** que tienen los datos actualmente (las 3 V's)



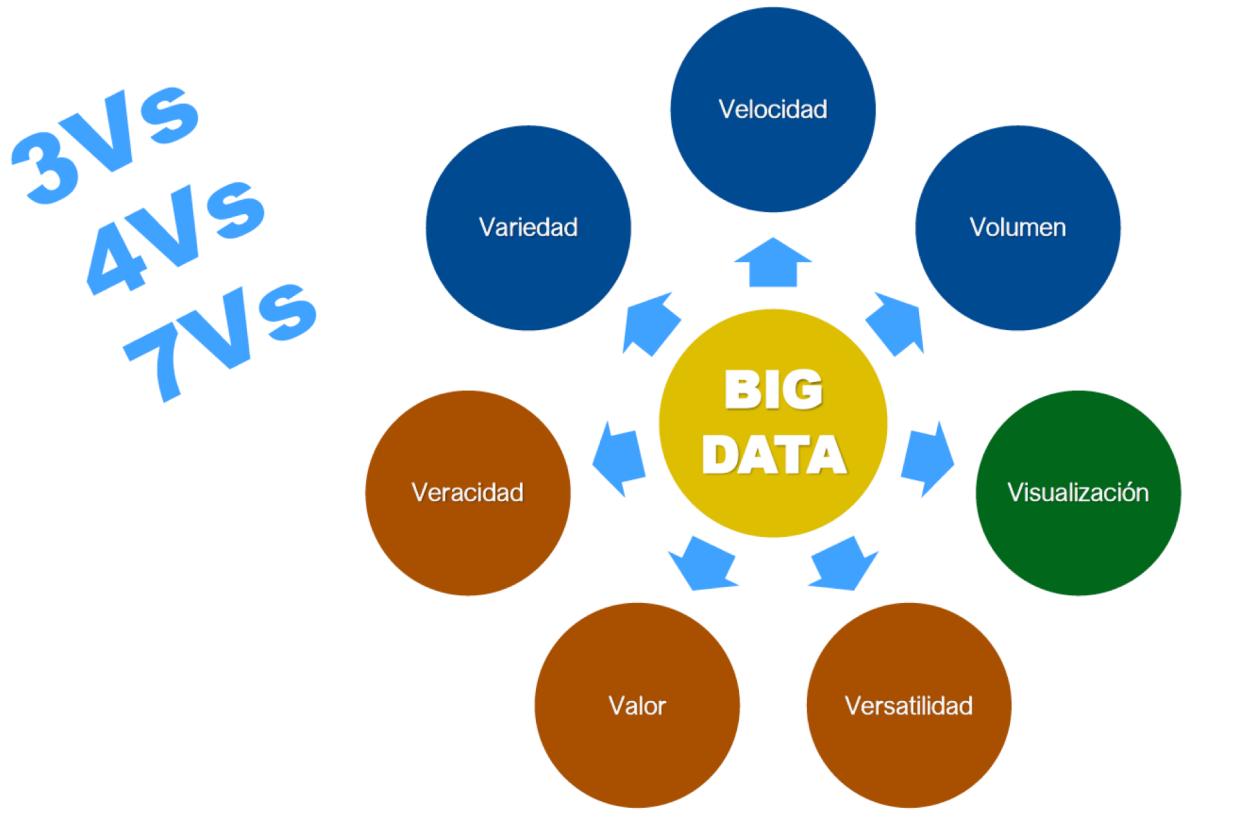
# El significado de Big Data - 3 V's

- Big Volume
- Big Velocity
- Big Variety



# Las 4 dimensiones de “Big Data”

- Volumen: ...
- Velocidad: ...
- Variedad: ...
- **Veracidad:** Por último el ***Big Data*** ha de ser capaz de tratar y analizar inteligentemente este inmenso volumen de datos con la finalidad de obtener una información verídica y útil que nos permita mejorar la toma de decisiones en las organizaciones.



Fuente: Dr. Carlos González

Programa Iberoamericano de Formación en Minería de Datos

# Características de Big Data

¿Qué tan grande es Big Data?

***Lo que es grande hoy en día  
tal vez no lo sea mañana***

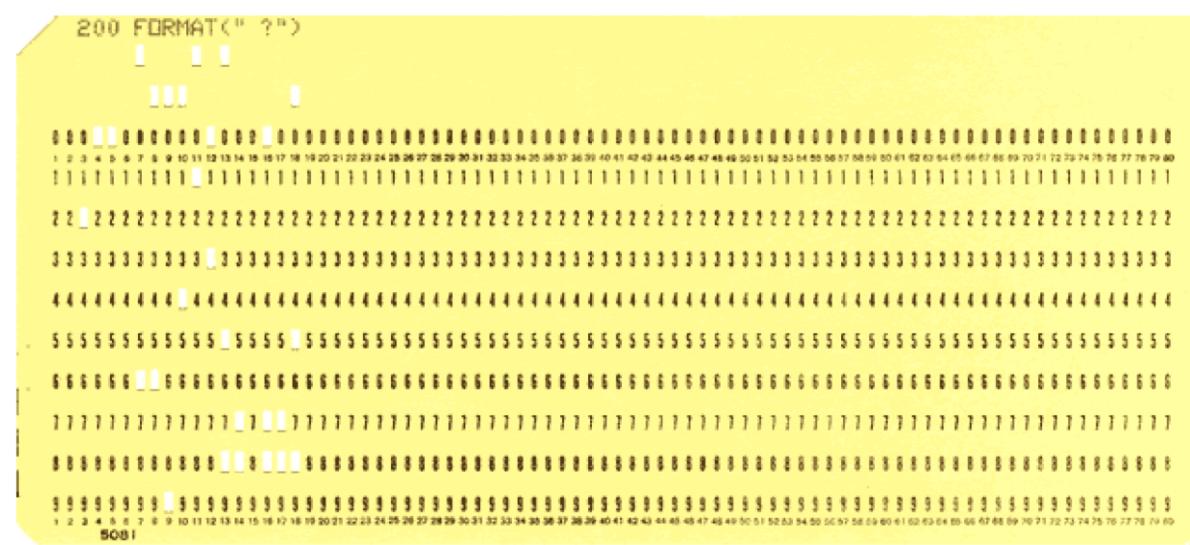
Memory unit	Size	Binary size
kilobyte (kB/KB)	$10^3$	$2^{10}$
megabyte (MB)	$10^6$	$2^{20}$
gigabyte (GB)	$10^9$	$2^{30}$
terabyte (TB)	$10^{12}$	$2^{40}$
petabyte (PB)	$10^{15}$	$2^{50}$
exabyte (EB)	$10^{18}$	$2^{60}$
zettabyte (ZB)	$10^{21}$	$2^{70}$
yottabyte (YB)	$10^{24}$	$2^{80}$

# Dos Eventos Importantes

- Evolución de la capacidad de almacenamiento
- Explosión en los datos



# Evolución de la Capacidad de Almacenamiento

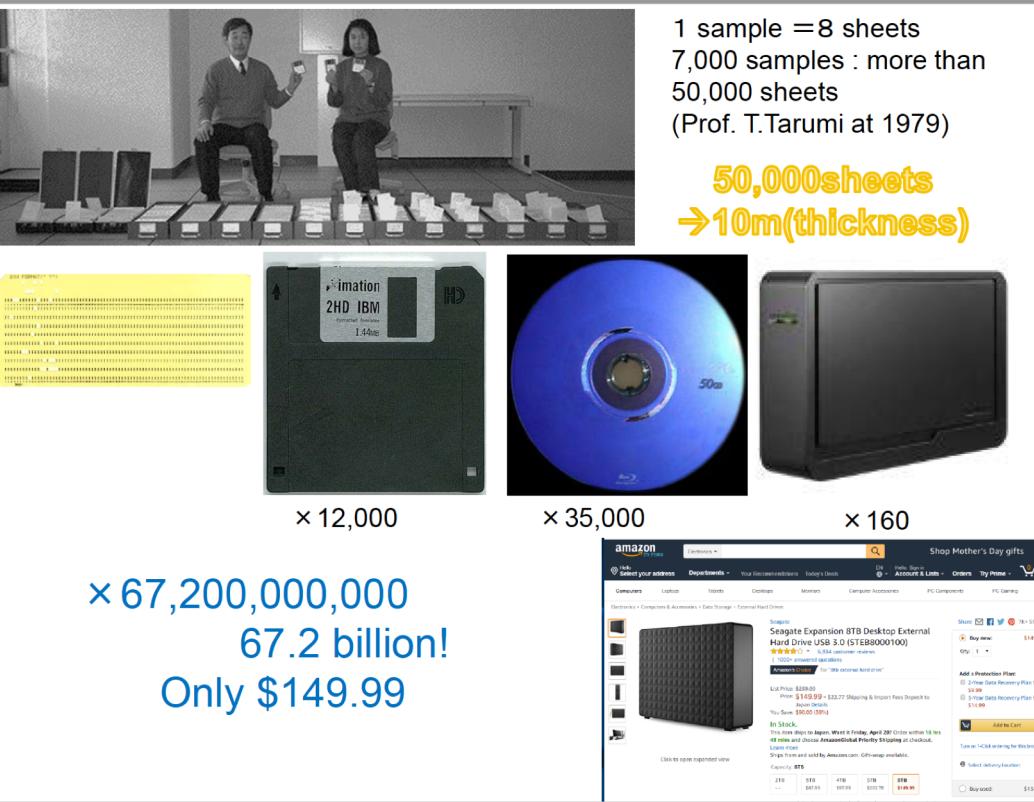


punched card

**120 BYTES**

Fuente: Masahiro Mizuta

# Evolución de la Capacidad de Almacenamiento



Fuente: Masahiro Mizuta

Programa Iberoamericano de Formación en Minería de Datos



**PROMiDAT**  
IBEROAMERICANO

# Evolución de la Capacidad de Almacenamiento

67,200,000,000sheets  
67.2 billion sheets  
Only \$149.99



100sheets  
→2cm(thickness)



8TB  
→13,440km(thickness)

13,268 km

Distance between  
Costa Rica and Japan



<https://www.ebay.com/itm/100pcs-VINTAGE-MAINFRAME-COMPUTER-PUNCH-CARDS-IBM-80-column-card-format-70-80s-/191211291509>

Fuente: Masahiro Mizuta

# Explosión en los Datos

## INFORMATION UNTIL 2003



5 EXABYTES

Fuente: Masahiro Mizuta

# Explosión en los Datos

**THERE WAS 5 EXABYTES OF INFORMATION  
CREATED BETWEEN THE DAWN OF  
CIVILIZATION THROUGH 2003,**



**BUT THAT MUCH INFORMATION IS NOW  
CREATED EVERY 2 DAYS --> EVERY FEW HOURS, NOW**



(Eric Schmidt, Google, CEO, 2010)

*Fuente: Masahiro Mizuta*

Programa Iberoamericano de Formación en Minería de Datos



**PROMiDAT**  
IBEROAMERICANO

# ¿Quién genera “Big Data”?

facebook

twitter

WORDPRESS

YouTube

flickr

Redes Sociales



Bancos



Instrumentos científicos



Dispositivos Móviles



Tecnología de sensores y redes



Comercio

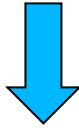
# Big Data - Big Analytics

- Se requieren de complejas operaciones matemáticas (machine learning, statistical elearning , clustering, trend detection, ....)
  - Análisis Exploratorio
  - Modelos Predictivos
  - Modelos de sumarización
  - Modelos simbólicos
- ***Mayor poder computacional y algoritmos eficientes*** para poder ejecutar operaciones como (en matrices gigantes):
  - Matrix multiply
  - QR decomposition
  - SVD decomposition
  - Linear regression

# Plataforma de código abierto

## “*Hadoop*”

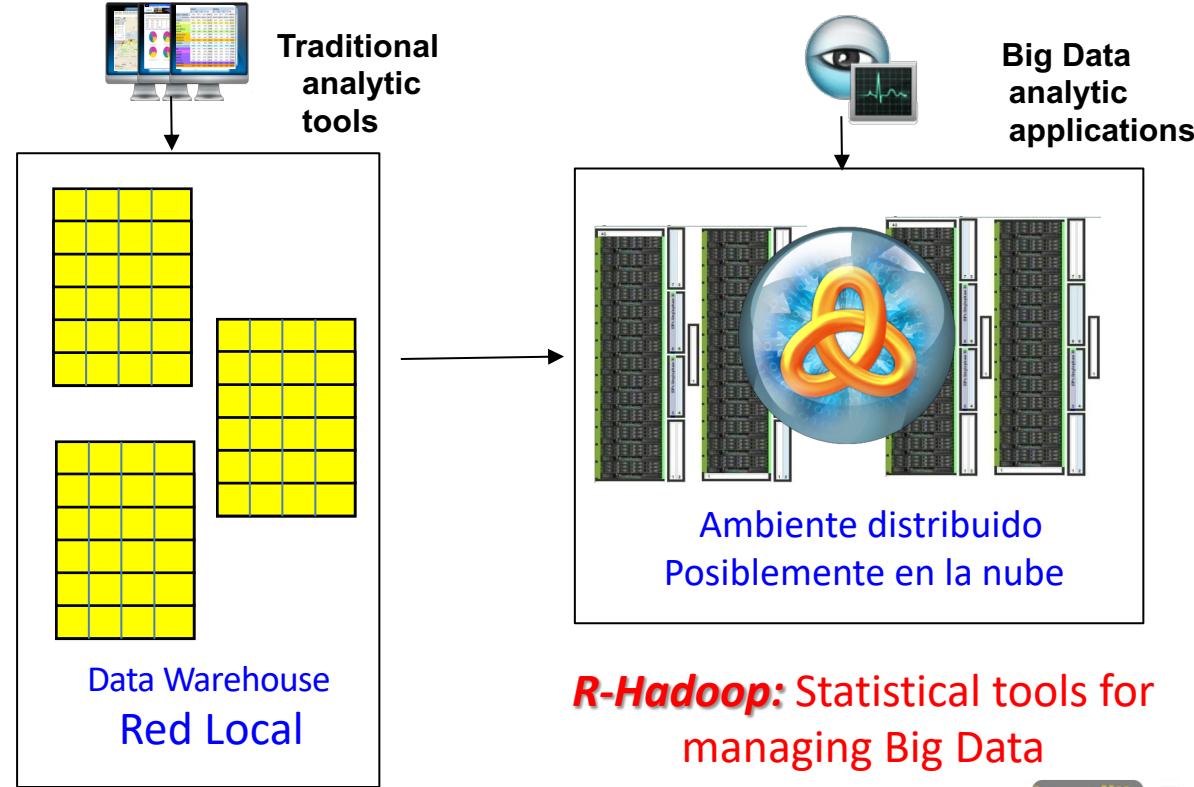
- Open-source software framework de Apache
- Inspirada en:
  - Google Map-Reduce
  - GFS (Google File System)



- **HDFS**
- **Map/Reduce**

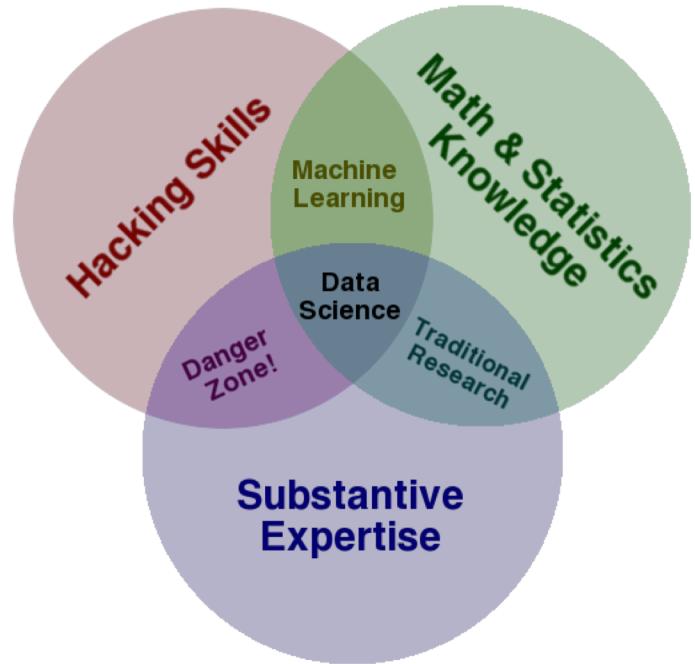


# Plataforma de código abierto “*Hadoop*”

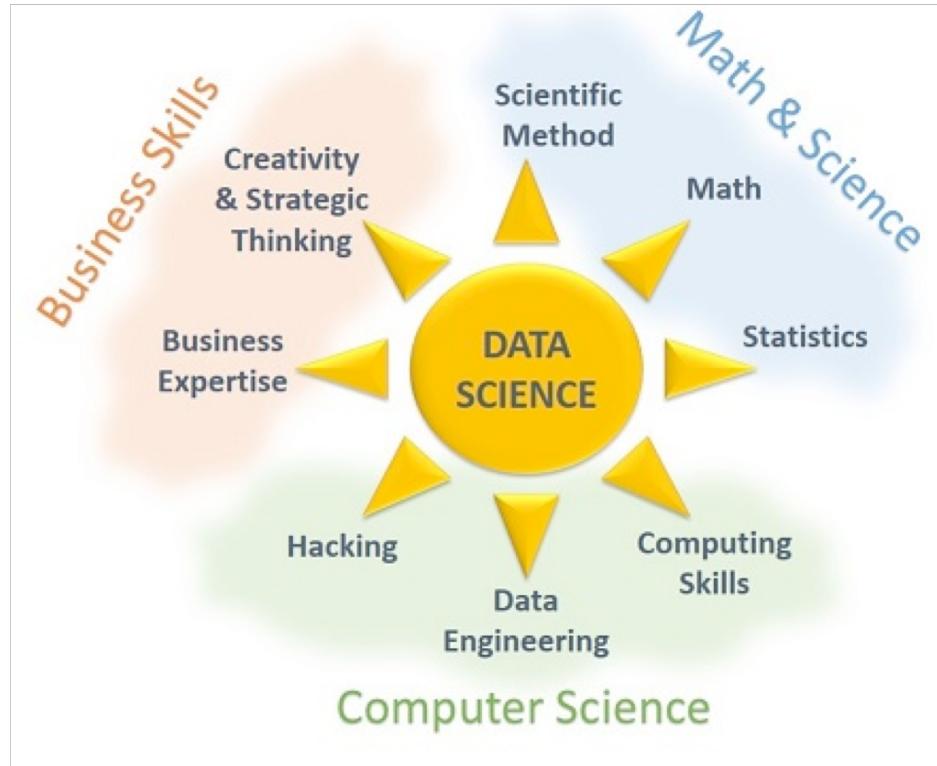


# ¿Qué es Ciencia de Datos?

- [Wikipedia] La ciencia de datos es un campo interdisciplinario que involucra métodos científicos, procesos y sistemas para extraer conocimiento o un mejor entendimiento de datos en sus diferentes formas, ya sea estructurados o no estructurados, lo cual es una continuación de algunos campos de análisis de datos como la estadística, la minería de datos, el aprendizaje automático y la analítica predictiva.

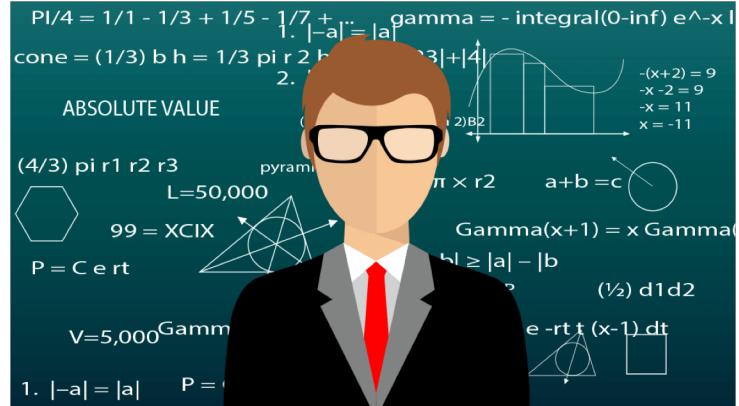


# ¿Qué es Ciencia de Datos?



# ¿Qué es un Científico(a) de Datos?

- [Josh Wills] "*Científico de datos*: Persona que sabe más de estadística que cualquier programador y que a la vez sabe más de programación que cualquier estadístico".
- Un científico de datos es sencillamente un profesional dedicado a analizar e interpretar grandes bases de datos.
- Científico capacitado para crear sus propios modelos dado un juego de datos, es decir, conoce los fundamentos de los métodos y no solo usa "cajas negras".



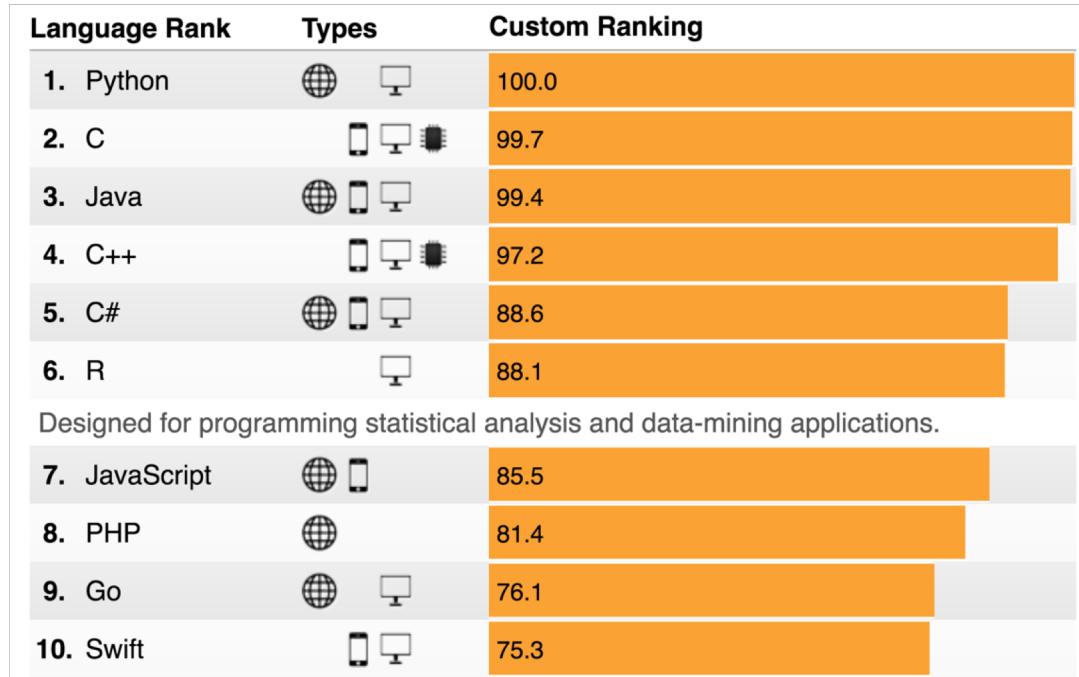


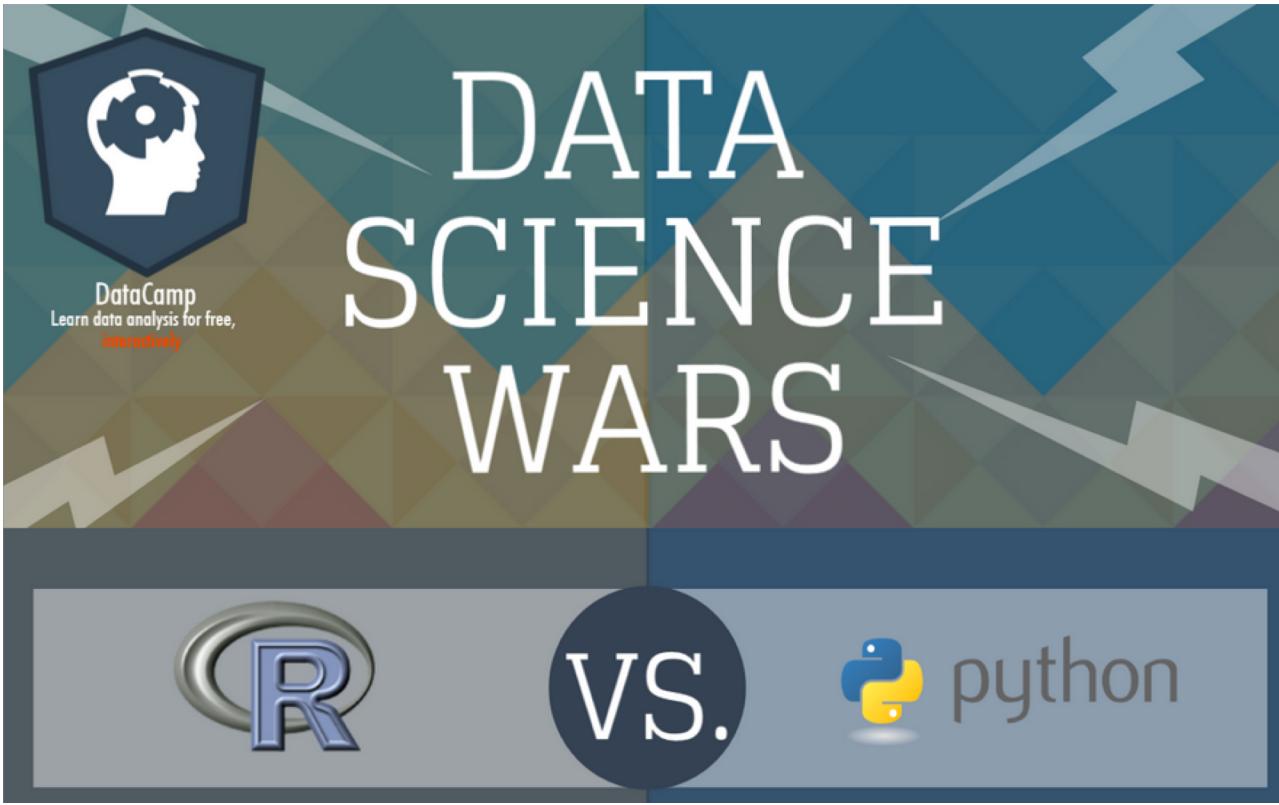
# Herramientas en Ciencia de Datos



**PROMIDAT**  
IBEROAMERICANO

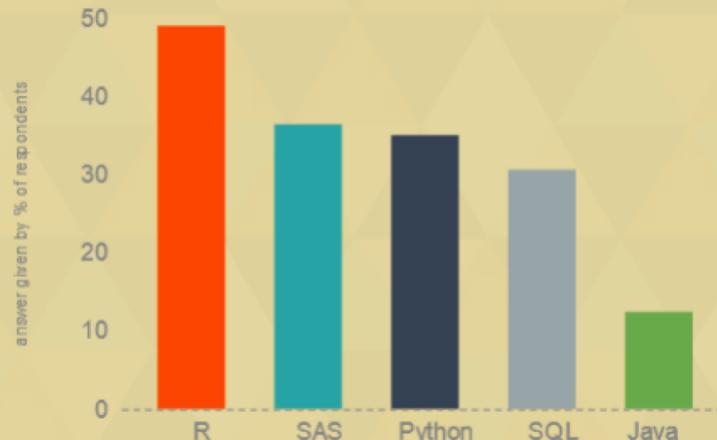
# Python en la actualidad



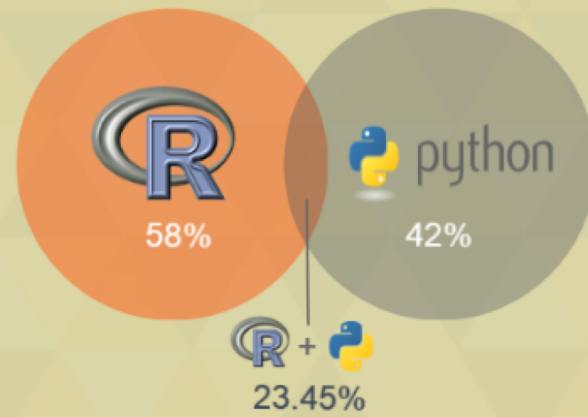


## General

Languages for data analysis used in 2014 (KDnuggets polls)



Analysis of R and Python used together in 2014 (KDnuggets polls)





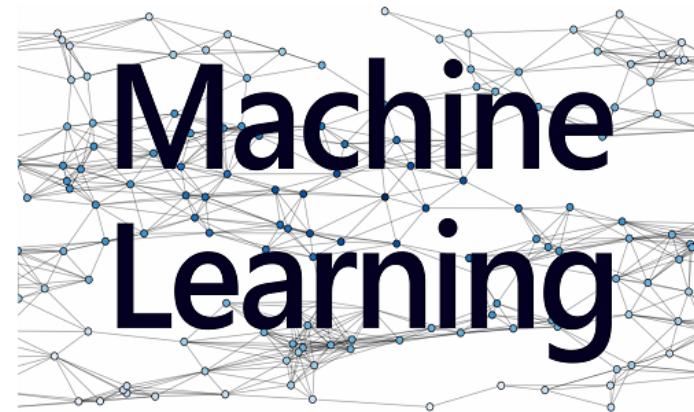
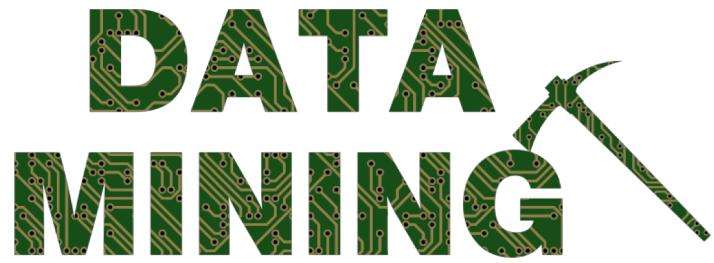
- R es un lenguaje enfocado al análisis de datos, la estadístico, la minería de datos, la visualización de modelos.



- John Chambers: “R es lenguaje creado por un estadístico para hacer estadística”.



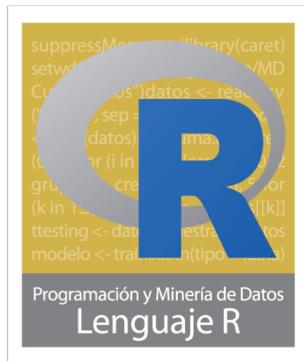
- Python es un lenguaje de propósito general enfocado a la reutilización de código.
- Python es un lenguaje orientado a objetos que permite una fácil integración a los sistemas de una organización.



Programa Iberoamericano de Formación en Minería de Datos

 **PROMiDAT**  
IBEROAMERICANO

# Entonces... ¿con cuál me quedo?



- Esta pregunta quizás sea la más complicada a la que dar respuesta.
- El uso de uno y otro debe por tanto estar motivado por la respuesta a la siguiente pregunta: ***¿Qué tipo de problema quiero resolver?***
- Optaremos por uno u otro en función del tipo de análisis de datos que queramos llevar a cabo, ya sea Machine Learning, Data Mining, analítica web, etc.
- Así, R es una muy buena opción cuando el análisis de datos requiere una computación independiente o un análisis individual en los servidores.
- Mientras que Python lo usaremos cuando el análisis de datos requiera ser integrado con las aplicaciones web o si necesitamos incorporar el código de análisis estadístico en una base de datos de producción.

**Muchas gracias  
por su atención....**

