De la Estadistica a Ciencias de Datos

Dr. Edgar Acuña
Departamento de Ciencias Mathematicas y Programa Doctoral en Ciencias e Ingenieria de la Información y Computación
Universidad de Puerto Rico-Mayaguez

E-mail: edgar.acuna@upr.edu , eacunaf@gmail.com

Website: academic.uprm.edu/eacuna

Cuando y Porque decidi estudiar Estadistica?

Tenia 15 años, en el verano antes de empezar mi ultimo año de secundaria

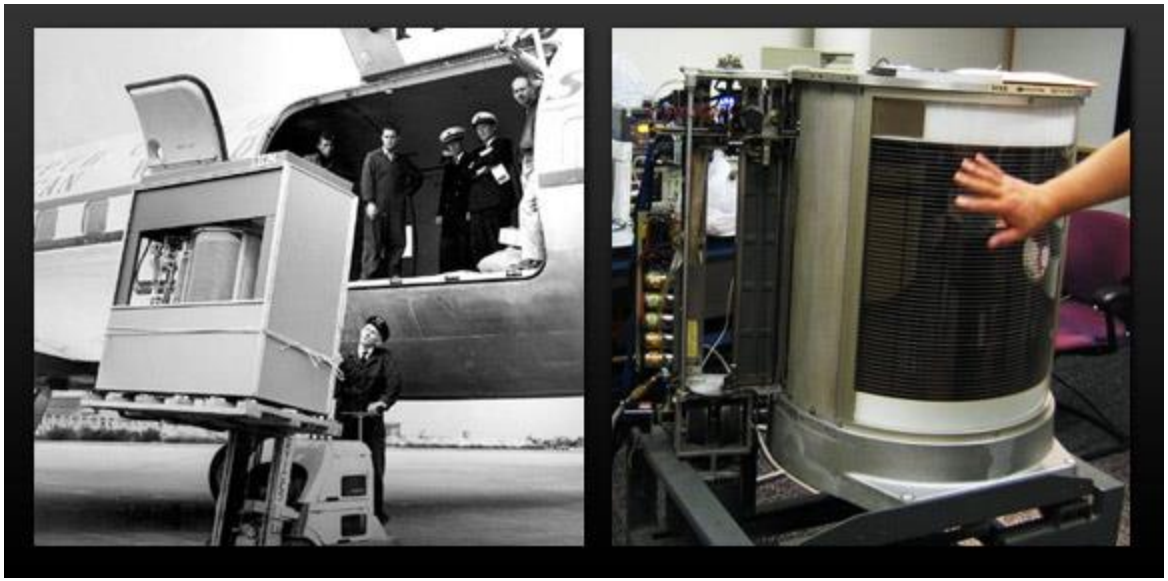
Queria hacer predicciones basado en datos: Quien campeonaria ese año en el futbol peruano, Como ganarse la polla (acertar 10 carreras) de los caballos. Hacia esto basado en reglas o patrones no usaba ecuaciones matematicas.

Mineria de Datos [1]: Motivacion

Los mecanismos para la recoleccion automatica de datos (por ejemplo, sensores) y el desarrollo de tecnologia para bases de datos ha hecho posible que una gran cantidad de datos pueda estar disponibles en bases de datos, almacenes de datos y otros repositorios de informacion.

Hoy dia, hay la necesidad de convertir estos datos en conocimiento e informacion.

El primer disco duro, 1956



IBM 350, tenia el tamano de dos refrigeradoras y una capacidad de alrededor de 5MB. Costaba aprox 50,000 dolares. Mi laptop tiene 100,000 veces mas de esa capacidad

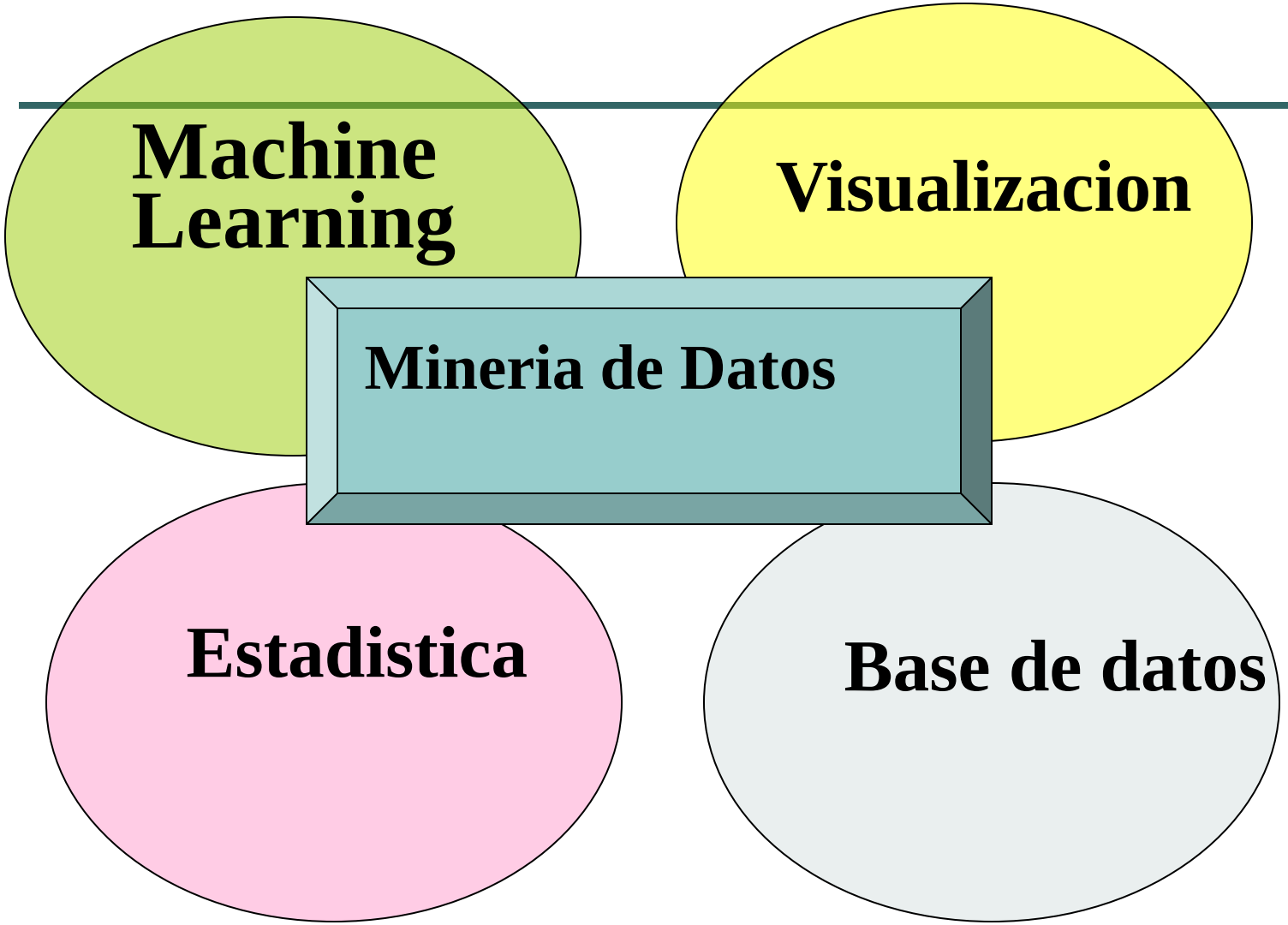
Mineria de datos [2]Que es Mineria de Datos?

- Es el proceso de extraer conocimiento/informacion valida de una base de datos bien grande. El conocimiento es dado como caracteristicas (“patrones”)y reglas que son no-triviales, previamente desconocidas, comprensibles y con alto potencial de ser utiles.
- Otro nombres: Descubrimiento de Conocimiento en bases de datos (KDD), Analisis Inteligente de datos, Ingenieria de datos, Inteligencia de negocios.
- El primer articulo en Mineria de datos aparecio en 1993.

Ejemplos de grandes bases de datos

- El telescopio Large Hadron Collider (LCH) almacena al año cerca de 25 Petabytes de datos de sensores.
- En 2010, la base de datos de llamadas de ATT era de 323 Terabytes.
- El 2008, los almacenes Walmart almacenaba 10 Terabytes de transacciones por dia.
- El 2011, Google buscaba en mas de 20 billones de paginas, que representa mas de 400 TB.
- Se estima que en el 2009 habia 500 exabytes (1,000 petabytes) de informacion en la internet.
- El 2013, se anuncio que el centro de datos de la NSA seria capaz de almacenar 5 zettabytes (1,000 exabytes).

Mineria de datos[5]. Areas relacionadas



Estadistica, Machine Learning

- Estadistica (~35% de DM)
  - Se basa mas en teoria. Asume propiedades distribucionales de las variables que estan siendo consideradas.
  - Se enfoca mas en probar hipotesis y en estimacion de parametros.
  - Se consideran efiicientes estrategias de recolectar datos.
  - Estimacion de modelos.
- “Machine learning” (Aprendizaje Automatico) (~30 % de DM)
  - Parte de Inteligencia Artificial. Machine es equivalente a un modelo en estadistica.
  - Mas heuristica que Estadistica.
  - Incluye a: Redes Neurales, arboles de decision, algoritmos geneticos.

Base de datos, visualizacion, etc

- Base de datos relacionales (~25% de DM)
  - Conjunto de tablas conteniendo datos de una categoria predeterminada. Cada una de las tablas (llamada relacion) contiene un o mas columnas de datos las cuales representan ciertos atributos.
  - Fue introducida por E. F. Codd de IBM en 1970.
  - El interface entre el usuario y la base de datos relacional mas usado es SQL(structured query language).
- Visualizacion (~5 % de DM)
  - Se explora la estructura del conjunto de datos en forma visual.
  - Puede ser usado en la etapa de pre o post procesamiento del KDD.

Otras Areas: (~5%): Pattern Recognition, Expert Systems, High Performance Computing.

Data mining es ...

- Determinar las caracteristicas de personas a las que se puede hacer un prestamo bancario.
- Detectar intrusos (casos anomalos) en un sistema.
- Determinar las caracteristicas de los clientes de un banco que pueden cometer fraude.
- Determinar las caracteristicas de los clientes que abandonan la subscripcion a un servicio.
- Hallar ciudades de un pais que tienen caracteristicas similares (regionalizacion).

Aplicaciones de DM

Administracion de negocios: Investigacion de mercados, relacion de los clientes con la gerencia, deteccion de Fraudes, Telecomunicaciones, etc.

Gobierno: deteccion de evasores de impuestos, terrorismo.

Ciencias: Astronomia, Bioinformatica (Genomics, Proteonomics, Metabolomics), decubrimiento de medicinas.

Text Mining: Extraer informacion previamente desconocida de diversas fuentes escritas (e-mails)

Web mining: Recomendar lista de productos que un usuario puede comprar. (Amazon.com)

Tipos de tareas en data mining

- Descriptivas: Se encuentra las propiedades generales de la base de datos. Se descubre las caracteristicas mas importantes de la base de datos.
- Predictivas: Se entrena (estima) un modelo usando los datos recolectados para hacer predicciones futuras. Nunca es 100% precisa y lo que mas importa es el rendimiento del modelo cuando es aplicado a nuevos datos.

Tareas en data mining

- Regresion (Predictiva)
- Classificacion (Predictiva)
- Classificacion No supervisada – Clustering (descriptiva)
- Reglas de Asociacion (descriptiva)
- Deteccion de Outliers (descriptiva)
- Visualizacion (descriptiva)

Retos de Data Mining

- Escalabilidad
- Dimensionalidad
- Datos complejos y Heterogeneos.
- Calidad de datos
- Preservacion de privacidad
- Analisis en tiempo real (streaming data)

Data Mining Software

- **Open source**
- R (cran.r-project.org). Relacionado a Estadistica (38.5% de usuarios, Kdnuggets.com, agosto 2014).
- Weka (http://www.cs.waikato.ac.nz/ml/weka/) (17.0%): Relacionado a Machine Learning. En Java.
- RapidMiner (http://rapidminer.com). (44.2%) Relacionado a la comunidad de base de datos. Parcialmente gratis.
- Python y Hadoop estan subiendo ultimamente.
- **Comercial:** SAS Enterprise Miner(10.5%), KMINE(15.0%), XLMiner(25.8%), Microsoft SQL(10.5%), Oracle(2.2%).

Data Mining segun Google trends

Big Data [1]: Antecedentes

Febrero 1977, L. Breiman organizo la conferencia "Analysis of large complex data sets". Patrocinada por la ASA y IMS Dallas, USA.
1994 COMPSTAT, Proceedings of Computational Statistics. Part I. Treatment of "huge" Data sets.
Mayo1997. The 29th Symposium on the Interface (Houston, TX) " Data Mining and the analysis of large data sets".
Octubre 1997, M. Cox y D. Ellsworth usaron el termino "big data" en una conferencia sobre visualization organizada por la la IEEE.
1998, Workshop on Massive data sets. Committee on Applied and Theoretical Statistics. NRC, USA.
Abril 1998, John Massey, jefe cientifico de SGI, presenta un paper "Big data… and the next wave of Infrastress" en una reunion de la USENIX.

Big Data[2]: Antecedentes

En Febrero 2001, Doug Laney, analista para el Meta Group, menciono que los retos del analisis de datos va en tres direcciones: Volumen creciente (cantidad de datos), velocidad (rapidez de los datos que entran y salen), y variedad (una gran gama de tipos de datos y fuentes de datos).

Septiembre 2008, La revista Nature publica una edicion especial acerca de “big data”.

Mayo del 2011, investigadores del McKinsey Global Institute publicaron: “Big data: the next frontier for innovation, competition and productivity”.

Big Data: Definicion

En 2012, Gartner actualizo la definicion de “Big data” como sigue: "Big data es informacion que se recibe en alto volumen, con alta velocidad, y/o alta variedad y que requiere nuevas formas de procesamiento para mejorar la toma de decisiones, entender mejor los descubrimientos y el proceso de optimizacion”.

Big data usualmente incluye bases de datos con tamanos mas alla de la capacidad del “software” comun para capturar, administrar, y procesar los datos dentro de un tiempo tolerable.

Los tamanos de Big data cambian constantemente. En 2002, era aprox 100GB, en 2012 de 10TB hasta petabytes para una sola base de datos.

Big data segun Google trends

Big data es ...

Big data is like teenage sex:

everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it,

so everyone claims they are doing it...

Dan Ariely (2013), Professor of Psicology. Duke University.

Big data segun Google trends

Big data es ...

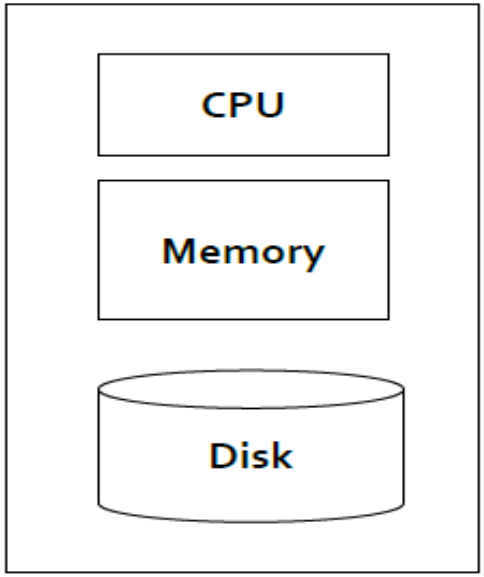
Big data is like teenage sex:

everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it,

so everyone claims they are doing it...

Dan Ariely (2013), Professor of Psicology. Duke University.

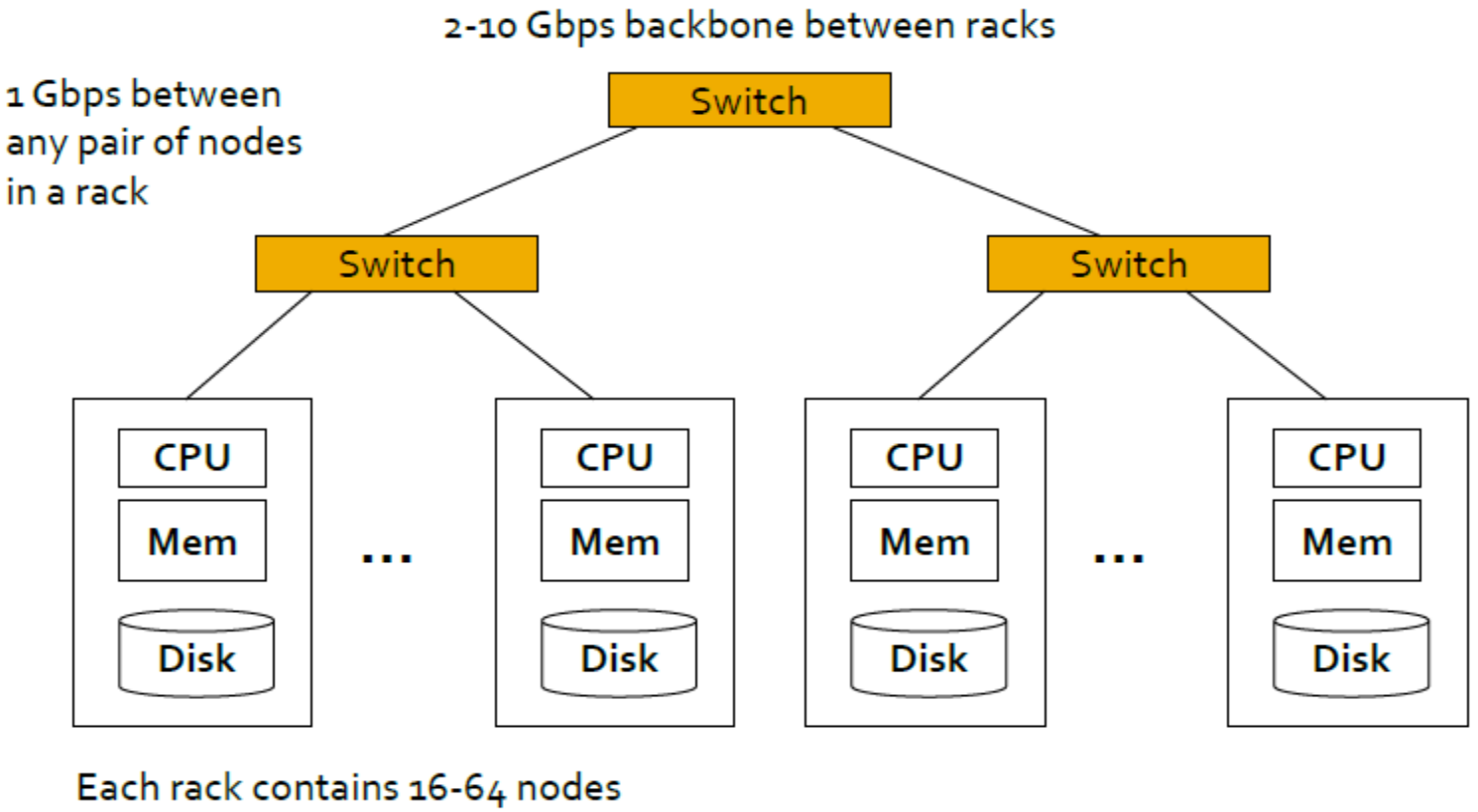
Arquitectura de un solo nodo



Machine Learning, Statistics

"Classical" Data Mining

Arquitectura de clusters



Uso de R en computacion paralela y distribuida

RMPI: Paquete para correr R con MPI
Snow: Paquete para correr R con MPI
Gputools: Paquete para correr R con CUDA
Rmr: Interface entre R y Hadoop

Elio Lozano y Edgar Acuna: 2003-2006, varios articulos en aplicacion de R en computacion en paralelo y distribuida para calculo de metaclassifiers, metaclustering, boosting, bagging, kernel density estimation, ensembles and outlier detection.

Motivacion: Ejemplo de Google

Google busca en mas de 20 billion de paginas web x20KB= 400+ TB. Una computadora lee del disco con una velocidad de 30-35 MB/sec.
Le tomaria aprox 4 meses en leer la web.

Seria necesario aprox 1000 hard drives de 400GB para leer la web.
Seria necesario aun mas disco duro para analizar los datos.

Hoy en dia, una arquitectura estandar esta siendo usada. Consiste de
-Un cluster de nodos Linux
- Una red de ethernet para conectar los nodos.

Retos en computacion en gran-escala para mineria de datos

Como distribuir los computos?
Como escribir facilmente programas para computacion distribuida?
Ls computadoras fallan!.
Una computadora puede funcionar sin problemas por tres anos (1000 dias)
Si hay 1000 servidores, se espera que en un dia cualquiera uno de ellos no funcione.
En 2011, se estimaba que Google tenia 1 million de computadoras, asi que en un dia cualquiera 1000 servidores podian fallar.

Que es Hadoop?

- En 2004, J. Dean y S. Ghemawhat, empleados de Google escribieron un articulo explicando el modelo de programacion Mapreduce y una estructura para almacenamiento de grandes conjuntos de datos (sistema de archivos) llamado Google File System (GFS).
- MapReduce es open source pero GFS no lo es.
- En 2006, Doug Cutting de Yahoo! , creo un open source GFS y lo llamo Hadoop Distributed File System (HDFS). En el 2009, el dejo Yahoo! y se fue a Cloudera.
- La estructura de software que apoya HDFS, MapReduce y otros programas relacionados es llamado el Proyecto Hadoop o simplemente Hadoop.
- Hadoop es distribuido por Apache Software Foundation.

Hadoop

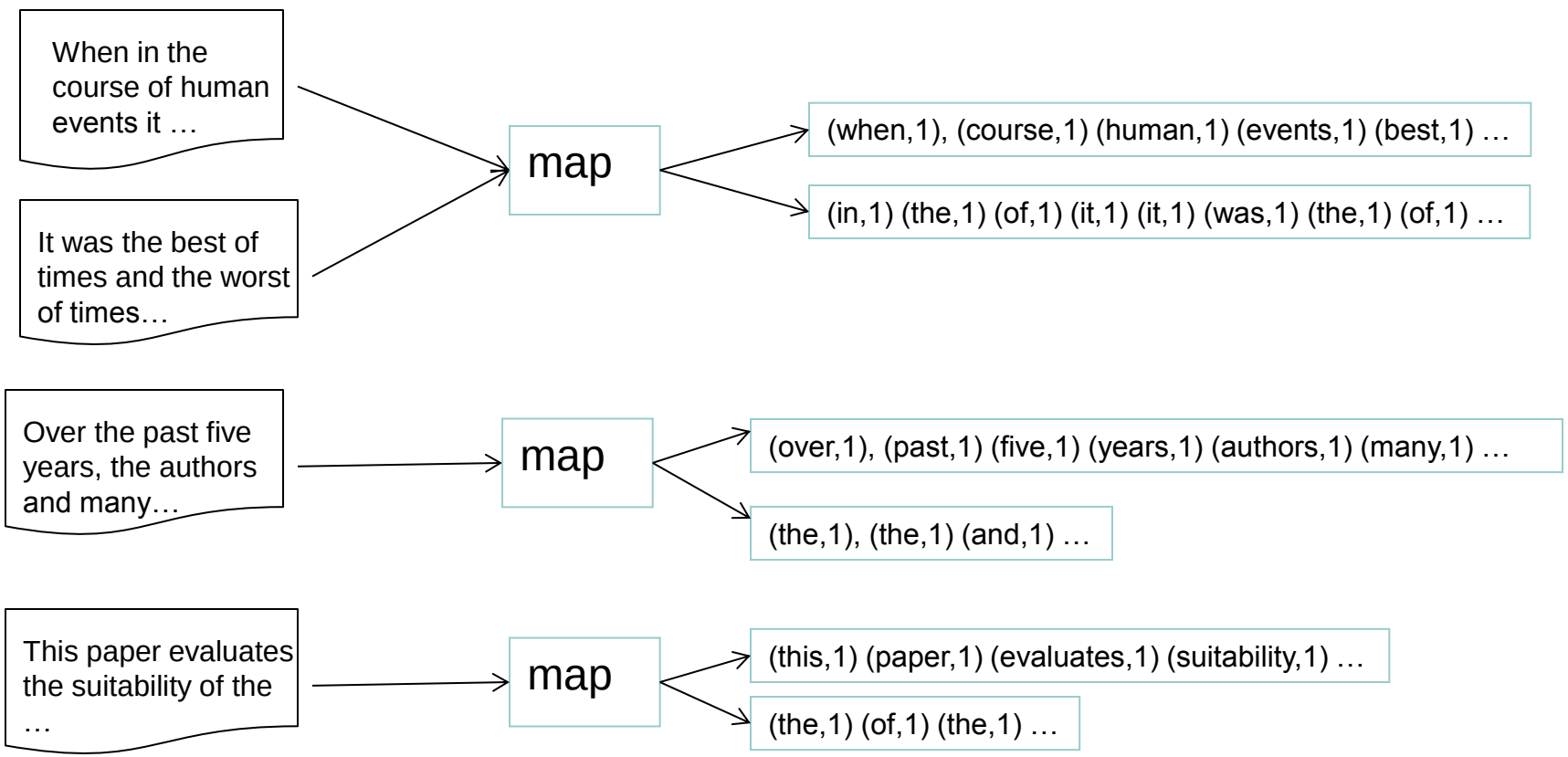
Hadoop incluye:

Distributed Files System(HDFS) –que distribuye los datos.
Map/Reduce-que distribuye la aplicacion

Esta escrito en Java

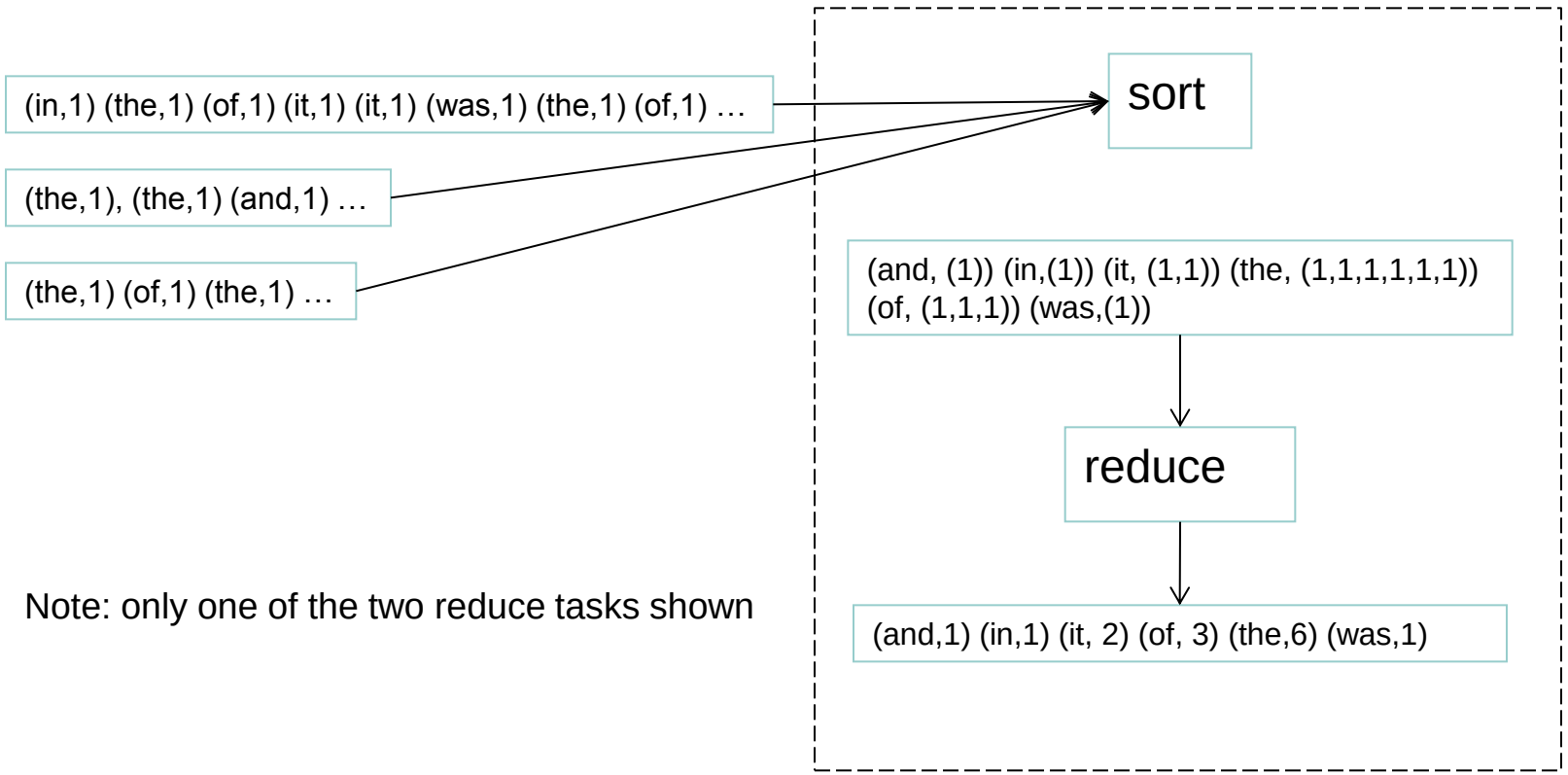
Corre en
-Linux, MacOS/X, Windows, y Solaris
-Usa hardware comun

El Map (Ejemplo: Wordcount)



El Reduce (Ejemplo:Wordcount)

partition (intermediate files) (R=2)



Note: only one of the two reduce tasks shown

Interfaces de otros programas con Hadoop

Hadoop streaming: C++,Python, perl,ruby,
Rhadoop (R and Hadoop), Weka(Mark Hall esta trabajando en esto), Radoop (rapidminer and hadoop, comercial)
Hadoop Pipes: (C++) no es recomendable.

Donde se puede correr hadoop?

En su computadora personal usando hadoop para un cluster de un solo nodo. Si usa Windows instale a maquina virtual (Vmware Player) donde corra Ubuntu(Ver el website de Michael Noll)

Gratis:
En el cluster Gordon del San Diego Supercomputer Center (1024 nodes) a traves del proyecto XSEDE.
En varios clusters: India(128 nodes), Sierra(84 nodes) y Alamo del proyecto Futuregrid liderado por la Universidad de Indiana.
Lamentablemente el proyecto se esta terminando.

No-Gratis, pero no muy costoso
Amazon Elastic Compute Cloud (EC2)

Quienes estan usando Hadoop?

Yahoo
Facebook
Amazon
Google
IBM
Netflix
Ebay
LinkedIn
Twitter

Falla de Big data

En Febrero del 2009 , investigadores de Google trataron de predecir la propagacion de la Influenza construyendo un modelo basado en las busquedas que hacia la gente en Google sobre 45 terminos relacionados a Influenza. (Nature, February 2009).

En los primeros dos anos el modelo tuvo un aceptable nivel de confiabilidad. Pero a partir del 2011 el modelo comenzo a fallar el principio en un 50% llegando hasta un 92% de error (Science, March 2014).

Otras herramientas computacionales

Spark del laboratorio AMP de UC Berkeley. Hay que programar en Scala, Java o Python. Gratis.

Watson Analytics de IBM (dueno de SPSS). Gratis para las universidades.

Azure de Microsoft Research, software hecho en Linux. Gratis para las universidades.

Percepcion de la gente acerca de la Estadistica y de los estadisticos

Un estadistico es aquel que lleva cuenta de los numeros (similar a un contable).

Un estadistico hace tablas y graficas para reportar y/o resumir informacion.

Un estadistico hace calculo (promedios, porcentajes y desviaciones estandar) para resumir informacion.

Un estadistico ayuda a los politicos o al gobierno a enganar a la gente (Mentiras, malditas mentiras y estadisticas).

Antecedentes de Ciencias de Datos

John Tukey (Annals of Mathematical Statistics", 1962). "The Future of Data Analysis". Tukey se declara analista de datos y duda que sea un estadistico matematico.

En 1996, la International Federation of Classification Societies (IFCS) hacen una conferencia en "Data Science,Classification and Related Methods".

J. Wu (1997). Propone cambiar el nombre de Estadistica a Ciencias de Datos. Ademas de que deberia haber un enfoque en el analisis de datos grandes y complejos. Estadistico=Cientifico de datos.

Antecedentes de Ciencias de Datos

J. H. Friedman (2000). “The role of the statistics in the data revolution”. Stanford University Tech Report. Propuso absober Mineria de Datos dentro de Estadistica.

L. Breiman (Statistical Science, 2001). Statistical Modelling: The two cultures. Afirma que: “Si el objetivo de la estadistica es usar datos para resolver problemas entonces deberiamos alejarnos de usar solamente “data models” y comenzar a usar “algorithmic models”.

Antecedentes de Ciencias de Datos

W. Cleveland (2001).”Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics”. Propone la creacion de la la carrera de Ciencias de Datos, similar a Estadistica multidisciplinaria.

En el ano 2002, comienzan a publicarse “Data Science Journal” y el “Journal of Data Science”.

En el 2008, D.J.Patil (LinkedIn) y Jeff Hamerbacher (Facebook) introdujeron el termino cientifico de datos.(????)

Ciencias de Datos como disciplina

En Enero del 2009. Hail Varian, el Jefe economista de Google afirma que el trabajo mas excitante por los proximos 10 anos sera ser estadistico!!!

Entre 2009 y 2010, mucha gente de ciencias de computacion y de administracion de empresas comenzaron a empujar la idea de la estadistica a la que se referia Hal Varian no era la tradicional sino un nuevo campo llamado “Data Science”.

T.H Davenport y J.H. Patil (2012, Harvard Business Review) inclinaron la balanza a favor de Ciencias de Datos

Estadistica versus Ciencias de Datos

	Estadistico	Cientifico de Datos
Imagen	Basket/Futbol	Trabajo mas sexy del siglo
Trabaja	Solo	En equipo
Datos	Preparados, limpios	Distribuidos, Sucios, No estructurados.
Tamano de datos	Kilobytes	Gigabytes
Tools	SAS, un CPU	R, Python, Hadoop, Linux
Focus	Inferencia	Prediccion
tiempo	semanas	segundos

Programas en Ciencias de Datos

Bachillerato: University of Rochester, University of San Francisco, College of Charleston, Nothern Kentucky University, Case Western Universty (mas de una decena en USA). Warwick University (UK).

Maestria: Carnegie Mellon University, Stanford University, Columbia University, Indiana University, NYU, SMU, Virginia. Mas de 30 en USA, Cabada, UK, Espana y resto del mundo.

Doctorado: Edimburgh University (UK), Rochester, Columbia. No pasan la decena aun.

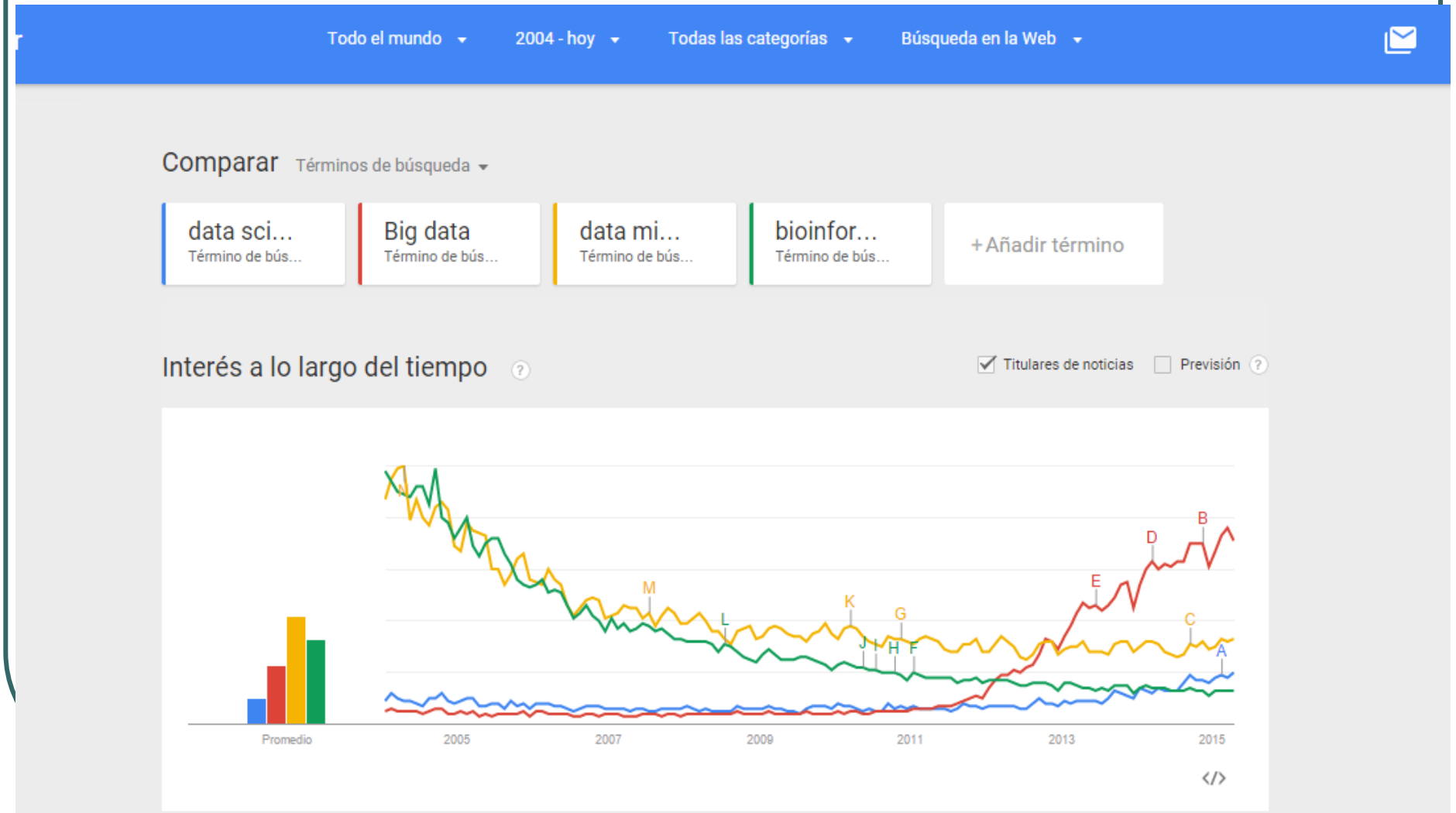
Latinoamerica: Solo Mexico (ITAM desde el 2014. Ni Brazil ni Chile.

Curriculum de un Bachillerato en Ciencias de Datos

Calculo Multivariado
Algebra lineal (matrices)
Dos cursos de probabilidad y estadistica
Introduccion a Ciencias de datos (estadistica aplicada actualizada)
Data Mining
Machine Learning (aprendizaje automatico)
Regresion
Muestreo
Estadistica Bayesiana
Estadistica Noparametrica
Visualizacion de datos
Bases de Datos
Algoritmos
Computacion en Paralelo (computacion de alto rendimiento)
Cursos de Economia/Administracion de Empresas/Teoria de Juegos

Data Science segun Google Trends

Comparacion de Data Mining, Big Data, Bioinformatica y Data Science segun Google Trends



Mi investigacion en progreso

Objetivo; Evitar el GIGO(Basura que entra-Basura que sale)
Estamos implementando varias tecnicas de preprocesamiento de datos en Mapreduce y Hadoop.

-The ReliefF:Metodo de seleccion de atributos (Carlos Gomez)
- Deteccion de Outlier (tesis doctoral de Carlos Gomez)
- Reduccion de la dimensionalidad
- Tratamiento de valores faltantes
- Tratamiento de datos de distintos tipos (cuantitativos y nominales)
- Extendiendo R usando Hadoop