Analisis de datos Masivos usando Python

Dr. Edgar Acuna
Departamento de Ciencias Matematicas
Universidad de Puerto Rico-Mayaguez
http://academic.uprm.edu/eacuna

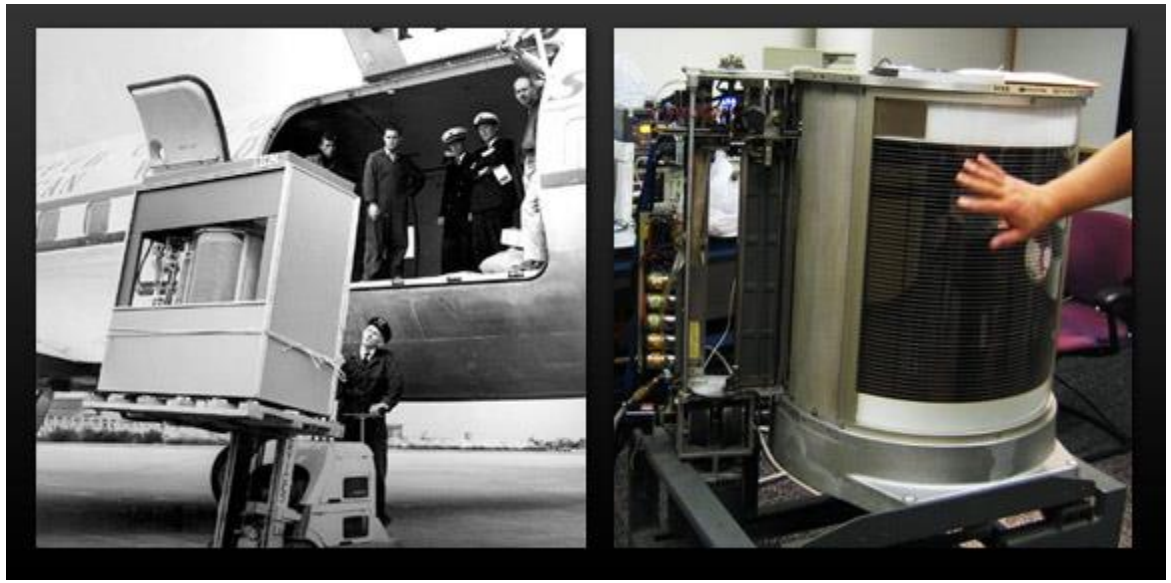
E-mail: edgar.acuna@upr.edu , eacunaf@gmail.com
Github: github.com/eacunafer
Noviembre 2019

Motivacion

Los mecanismos para coleccion automatica de datos y el desarrollo de la tecnologia de bases de datos ha generado que se puedan almacenar grandes cantidades de datos en bases de datos, almacenes de datos y otros depositarios de informacion.

Hay la necesidad de convertir esos 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 y me costo $1000

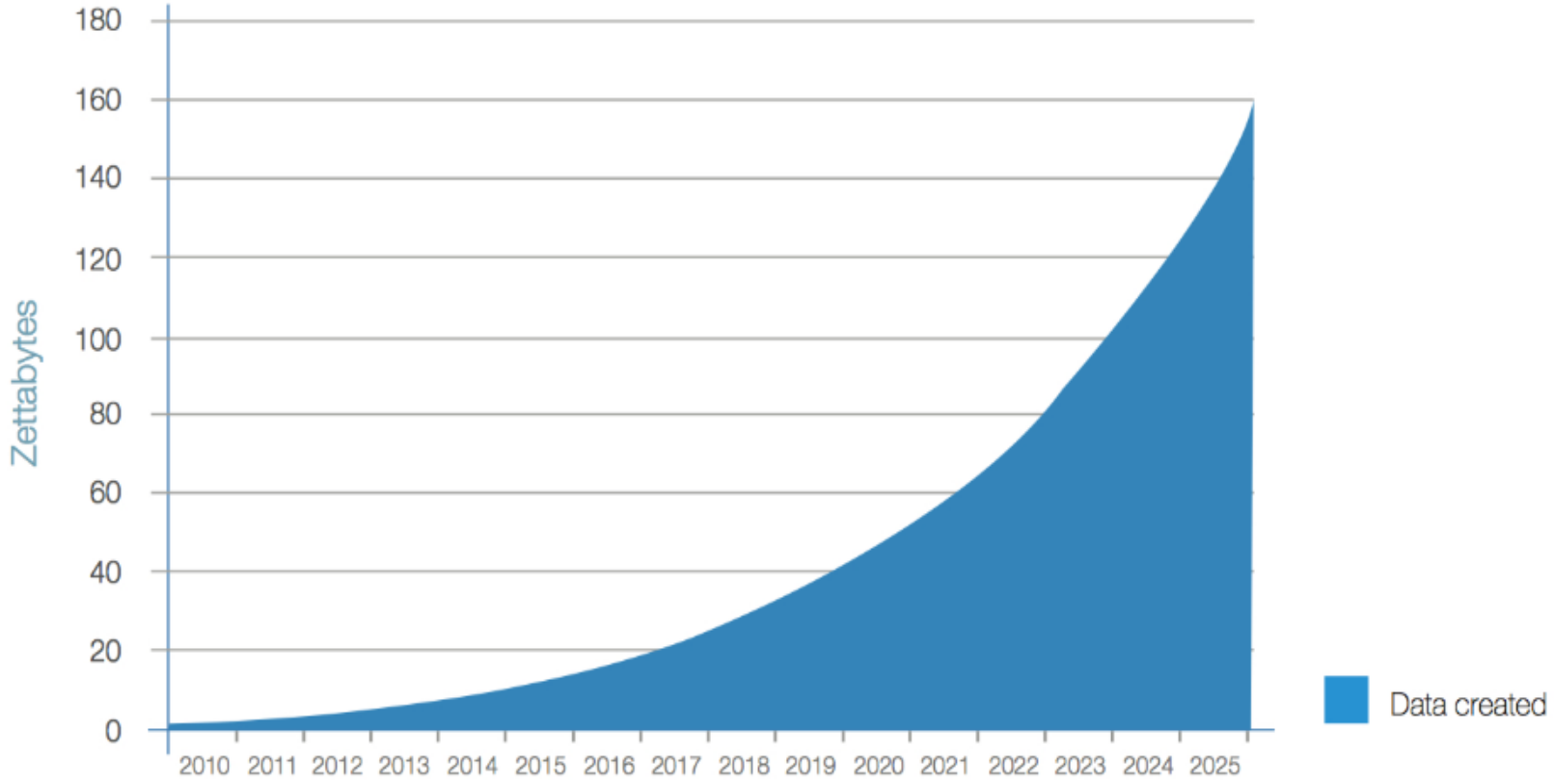
Tamano de datasets (en Bytes)

Description	Size	Storage Media
Very small	10^2	Piece of paper
Small	10^4	Several sheets of paper
Medium	10^6 (megabyte)	Floppy Disk
Large	10^9(gigabite)	USB/Hard Disk
Massive	10^{12}(Terabyte)	Hard disk/USB
Super-massive	10^{15}(Petabyte)	File of distributed data
Exabyte(10^{18}), Zettabytes(10^{21}), Yottabytes(10^{24})		

Ejemplos de grandes bases de datos hasta el 2016

- Hasta 2016, conjunto de datos de demora en los vuelos era de approx 25 Gigabytes.
- Amazon.com 45 TB de informacion de 60 millones de clientes,
- En 2010, la base de datos de llamadas de ATT era de 323 Terabytes.
- Hasta el 2016,Google busca en mas de 130 trillones de paginas, que representa mas de 390 Petabytes.
- El telescopio Large Hadron Collider (LCH) almacena al año cerca de 600 Petabytes de datos de sensores.
- El 2014, se construyo el centro de datos de la NSA seria capaz de almacenar 5 zettabytes (1,000 exabytes).

Cantidad estimada de Datos disponible en las computadoras del mundo

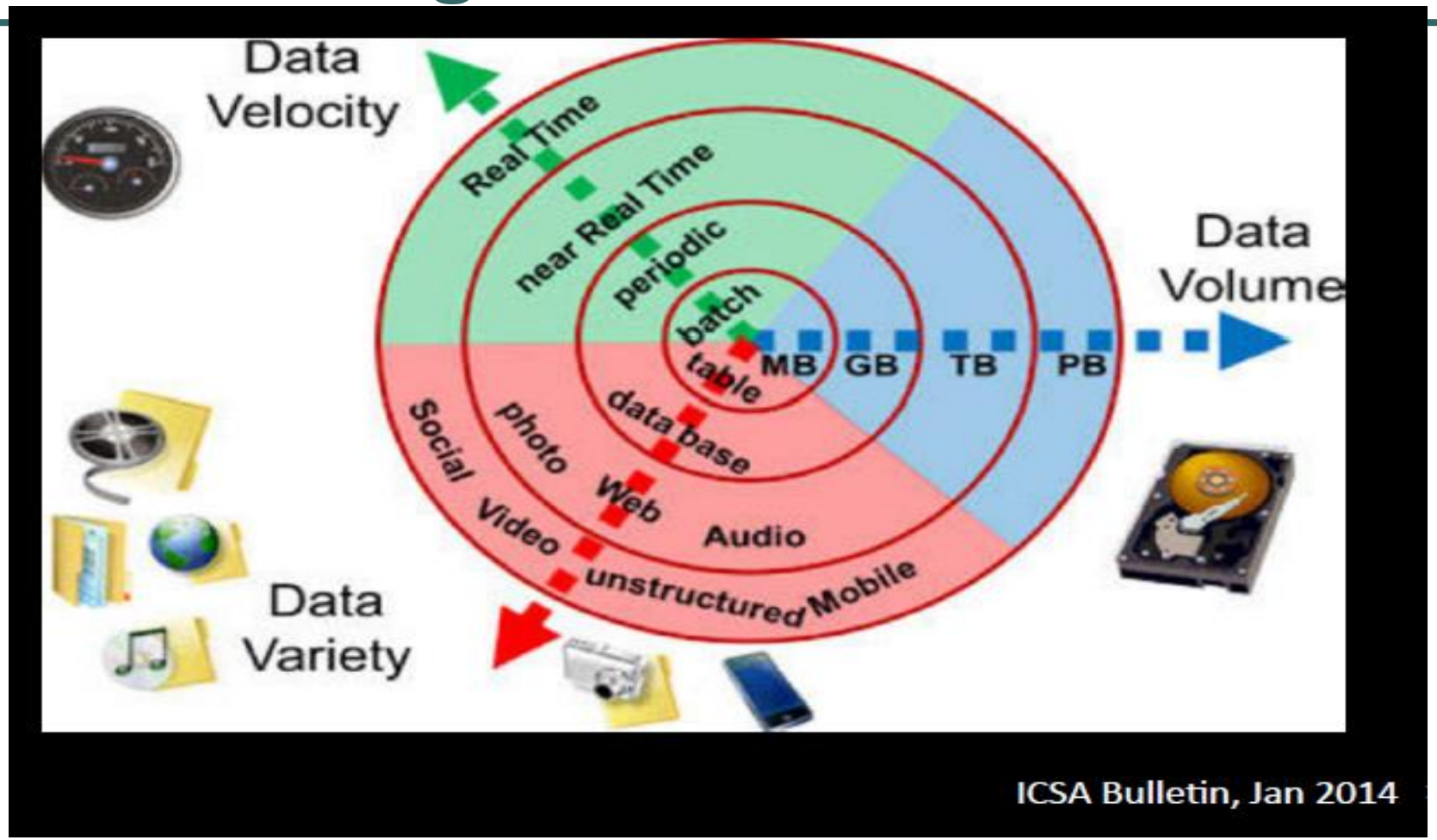


Data Revolution

Year	Digital	Analog	Amount
2000	25%	75%	2 Exabytes
2007	93%	7%	300 Exaby
2013	98%	2%	1,200 Exaby

Source: Viktor Mayer-Schönberger and Kenneth Cukier: Big Data: A Revolution that will Transform how We Live, Work and Think 2013)

Las 3 V's de Big data



Tipos de Datos

- **Datos estructurados:** Son datos faciles de organizar y usualmente estan disponible en bases de datos. Solo cuenta por un 20% de los datos disponibles actualmente. Datos de sensores, datos de llamadas telefonicas, datos de tarjetas de credito, datos de cajeros automaticos.
- **Datos No estructurados:** No estan organizados en un formato de estructura predefinida. No estan almacenados en una base de datos relacional. Puede ser textual o non-textual, y generado por humano o maquina. Generalmente son datos provenientes de las redes sociales:Twitter, Facebook, LinkedIn, etc. Tambien mensajes de texto por e-mail, imagenes, videos, files de audio, etc.
- Puede ser almacenado dentro de una base de datos no-relacional como NoSQL.

Tipos de datos (cont)

Tambien hay **datos semi-estructurados**, en donde se mantienen tags para identficar elementos de datos por separado, que permite despues hacer agrupamientos y jerarquias. Los mensajes Email pueden ser considerados como datos semi-estructurados. La metadata asociada a los email's permite clasificar y hacer busqueda de los mensajes. Ejemplos de datos semi-estructurados incluyen documentos en Markup Language XML, en JSON (JavaScript Object Notation) y bases de datos NoSQL (MongoDB, Hive, Cassandra, Couch DB, etc). Estas bases de datos son usadas mayormente en Big Data cuando se transmite en tiempo real datos de aplicaciones web.
En mineria de datos se trabaja mayormente con datos estructurados. Ciencias de datos trabaja con cualquier tipo de datos. El proceso de convertir datos no estructurados en datos capaces de ser analizados es llamado "Wrangling".

Que es Mineria de Datos?

- Es el descubrimiento de conocimiento en un conjunto de datos enormemente grande. El conocimiento que se obtiene viene dado en forma de caracteristicas(patrones) que no son triviales, que son previamente desconocidas y que tienen bastante posibilidades de ser utiles.
- Otros nombres: Descubrimiento de conocimiento en bases de datos (KDD), extraccion de conocimiento, analisis inteligente de datos.

Data Mining no es …

- Buscar un numero en una guia telefonica
- Buscar una definicion en Google.
- Generar histogramas de salarios por grupos de edad.
- Hacer una consulta en SQL y leer la respuesta de la consulta.

Data mining es ...

- Hallar grupos de personas que padecen las mismas enfermedades.
- 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.
- Recommendar productos a un cliente basado en su historial de compras.
- Determinar las caracteristicas de los clientes que abandonan la subscripcion a un servicio.

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.

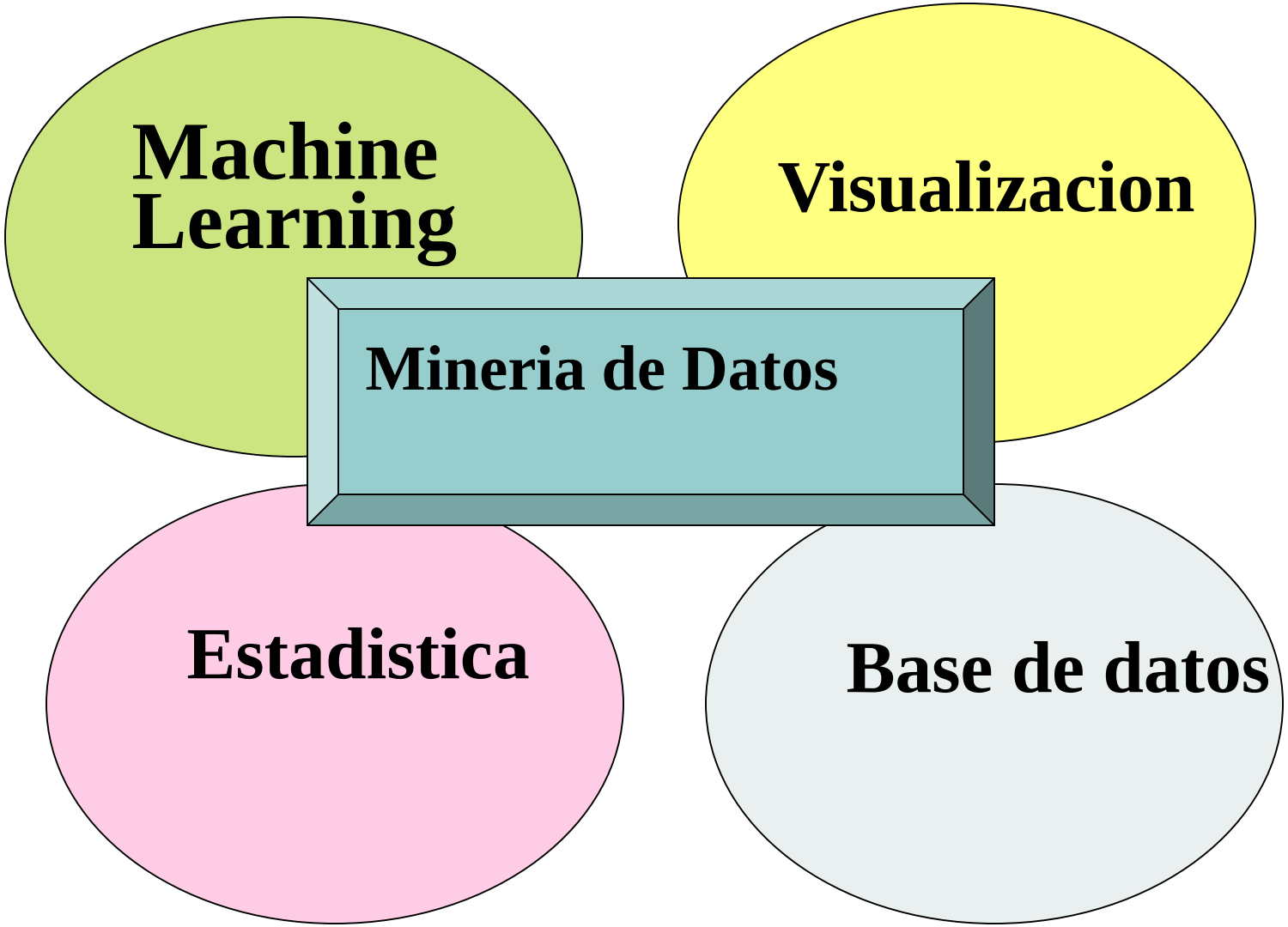
Text Mining: Extraer informacion previamente desconocida de diversas fuentes escritas (e-mails)

Web mining: E-comerce (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.

Areas relacionadas



Estadistica, Machine Learning

- Estadistica (~30% 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.
  - Estimacion de modelos.
- Machine learning (~30 % de DM)
  - Parte de Inteligencia Artificial. Machine es equivalente a un modelo en estadistica.
  - Mas heuristica que Estadistica.
  - Se enfoca en mejorar el rendimiento de un clasificador basado en sus experiencias pasadas.Tambien considera el tiempo que dura el proceso de aprendizaje.

Visualizacion, base de datos, etc

- Base de datos relacionales (~20% de DM)
  - Una base de datos relacional es un conjunto de tablas conteniendo datos de una categoria predeterminada. Cada una de las tablas (llamada relacion) contiene una o mas columnas de datos las cuales representan ciertos attributos. Cada una de las filas de la tabla contiene datos de las categorias definidas en las columnas.
  - El interface entre el usuario y la base de datos relacional mas usado es SQL(structured query laguage).
  - Una base de datos relacional puede ser agrandada facilmente
- Visualizacion (~10 % 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 (~ 10%): Pattern recognition, expert systems, High Performance Computing.

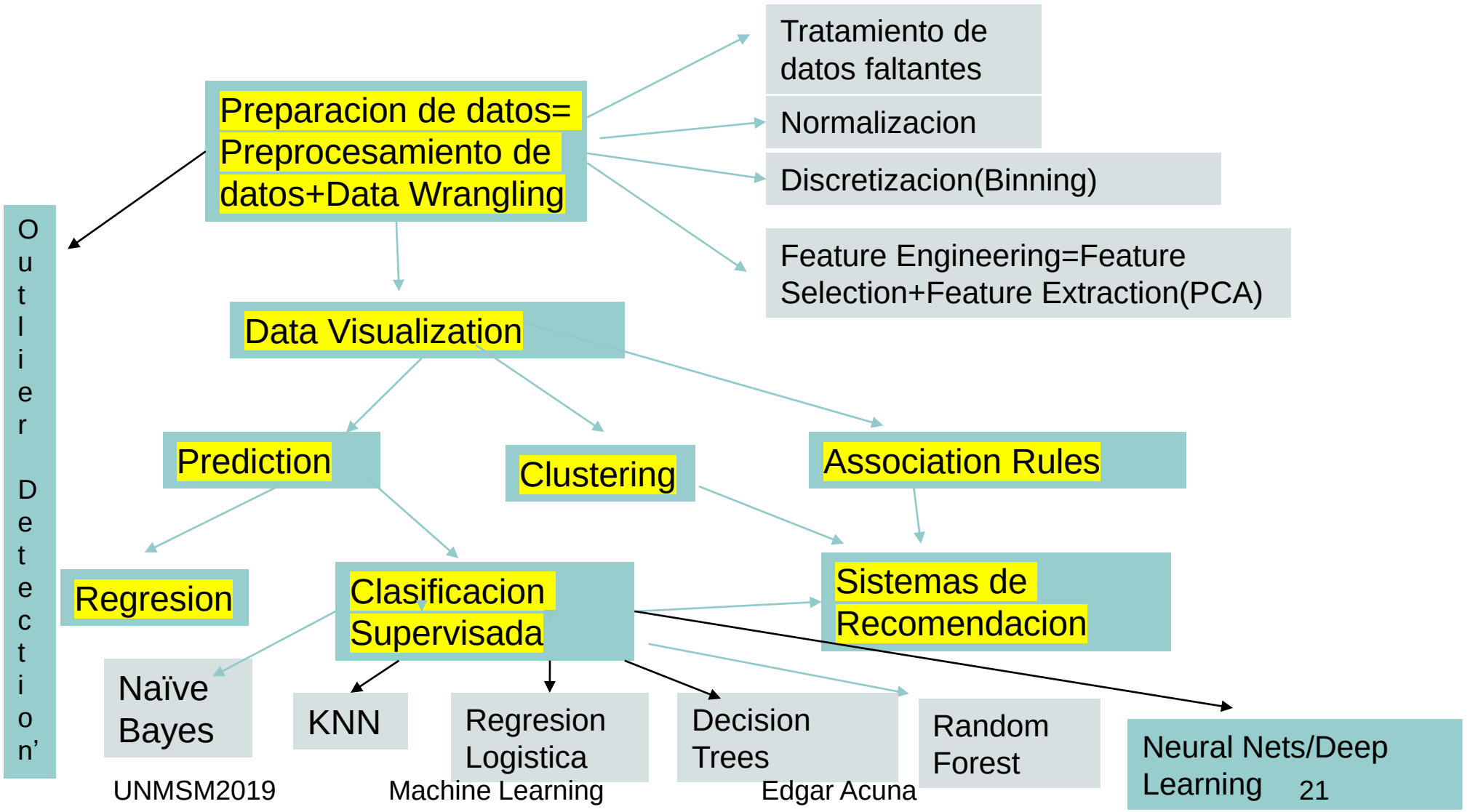
Tipos de tareas en data mining

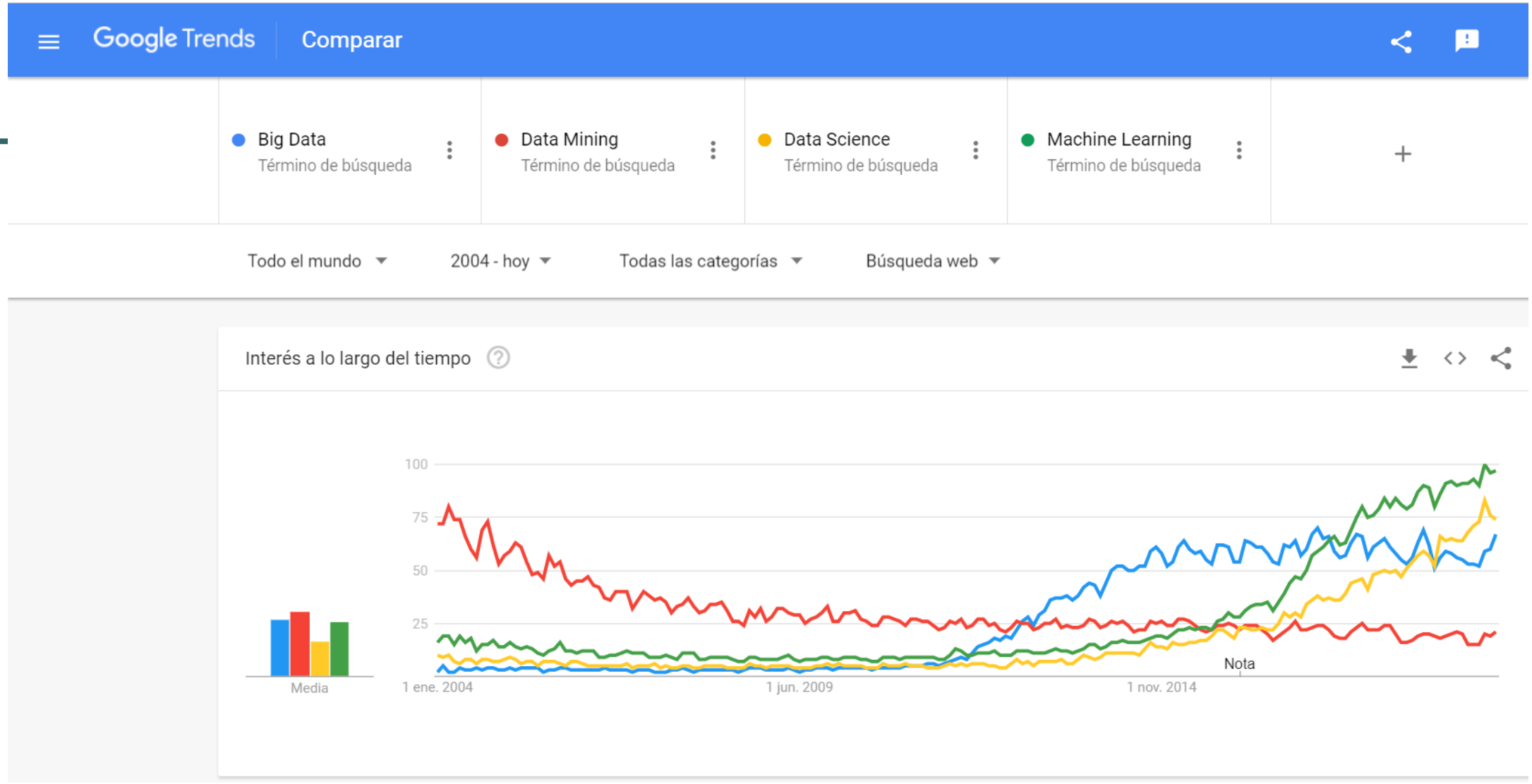
- Descriptivas: Se encuentra las propiedades generales de la base de datos. Se descubre las caracteristicas mas importantes de la base de datos.
- Predictivas: Se entrena (estima) un modelo usando los datos recolectados para hacer predicciones futuras. Nunca es 100% precisa y lo que mas importa es el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

Tareas en data mining

- Regresion (Predictiva)
- Clasificacion (Predictiva)
- Clasificacion No supervisada –Clustering (descriptiva)
- Reglas de Asociacion (descriptiva)
- Deteccion de Outliers (descriptiva)
- Visualizacion (descriptiva)
- Sistemas de Recomendacion (Predictiva)
- Analisis de Sentimientos (Descriptiva/Predictiva)

Diagrama de Flujo Machine Learning




Google Trends
Comparar
Big Data
Término de búsqueda
Data Mining
Término de búsqueda
Data Science
Término de búsqueda
Machine Learning
Término de búsqueda
Todo el mundo
2004 - hoy
Todas las categorías
Búsqueda web
Interés a lo largo del tiempo
100
75
50
25
Nota
Media
1 ene. 2004
1 jun. 2009
1 nov. 2014

Software

- **Gratuitos:**
- R (cran.r-project.org). Inclinado a la estadistica (48.5% de usuarios segun kdnuggets, Mayo 2018)
- Python (python.org 65.6% de usuarios)
- Rapidminer (rapidminer.com). (52.7% de usuarios)
- **Comerciales:** Microsoft SQL (39.6%), (Excel (39.1%), KNIME (12.3%) , SAS Enterprise Miner (4.3%), IBM Watson(3.1%).

El uso de modulos en Python

- Un programa en Python inicialmente solo tiene acceso a unas funciones basicas.

  (“int”, “dict”, “len”, “sum”, “range”, ...)
- dir(__builtins__) da una lista de las funciones disponibles.
- El uso de “Modulos” le anade mas funcionalidad a Python. Para cargar un modulo se usa el comando “import” .
- Ejemplos

>>> **import math #Calculo de funciones matematicas**

Usando el modulo math

```
>>> import math
>>> math.pi
3.141592653589793 1
>>> math.cos(0)
1.0
>>> math.cos(math.pi)
-1.0
>>> dir(math)
['__doc__', '__file__', '__name__', '__package__', 'acos', 'acosh',
'asin', 'asinh', 'atan', 'atan2', 'atanh', 'ceil', 'copysign', 'cos',
'cosh', 'degrees', 'e', 'exp', 'fabs', 'factorial', 'floor', 'fmod',
'frexp', 'fsum', 'hypot', 'isinf', 'isnan', 'ldexp', 'log', 'log10',
'log1p', 'modf', 'pi', 'pow', 'radians', 'sin', 'sinh', 'sqrt', 'tan',
'tanh', 'trunc']
>>> help(math)
>>> help(math.cos)
```

“import” y “from … import …”

```
>>> import math as m
m.cos
>>> from math import cos, pi
cos
>>> from math import *
```

Mas modulos

>>> **import Numpy #Contiene calculos y graficas estadisticos.**

>>>import scipy # Para hacer calculos cientificos tales como integracion numerica y optimizacion. El submodulo scipy.stats hace computes estadisticos

>>> import matplotlib # Para hacer graficas al estilo de matlab

>>> import pandas #Para hacer analisis estadistico al estilo de R

>>> import statsmodels #Para hacer regresion y series de tiempos

>>> import sklearn #Para ejecutar algoritmos de Machine Learning

>>> import h2o #Para ejecutar algoritmos de Machine Learning

>>> import keras #Para ejecutar algoritmos de Deep Learning

Leyendo datos con Pandas

```
# desde un file en su propia computadora
import pandas as pd
train=pd.read_csv('c:/Users/edgar2017/Downloads/titanic.csv')
#directamente de la internet
#na_values representa los valores faltantes(missing values)
breastdf=pd.read_csv("https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data",header=None, sep=",",na_values=['?'])
```

Ejemplos a ser mostrado en este taller

- Arboles y Random Forest
- Redes Neuronales y DeepLearning
- Deteccion de outliers usando autoencoders
- Sistemas de Recomendacion usando deep learning