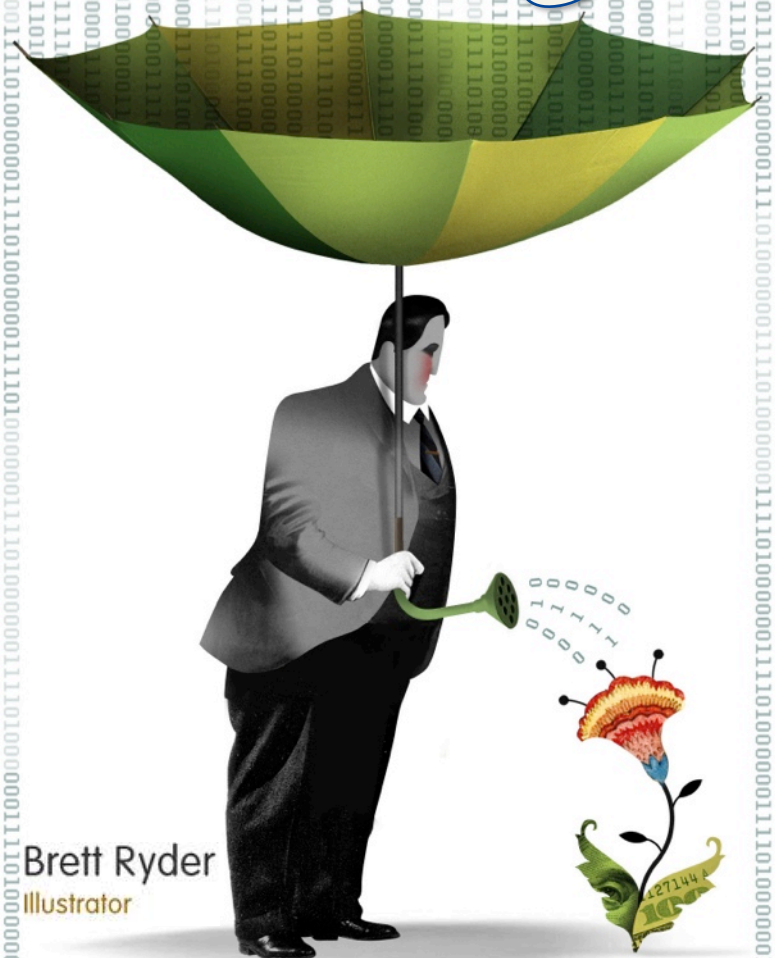
¿Qué es Big Data?




SG
VIRTUAL
CONFERENCE

http://en.wikipedia.org/wiki/File:Blind_men_and_elephant3.jpg

#sgvirtual

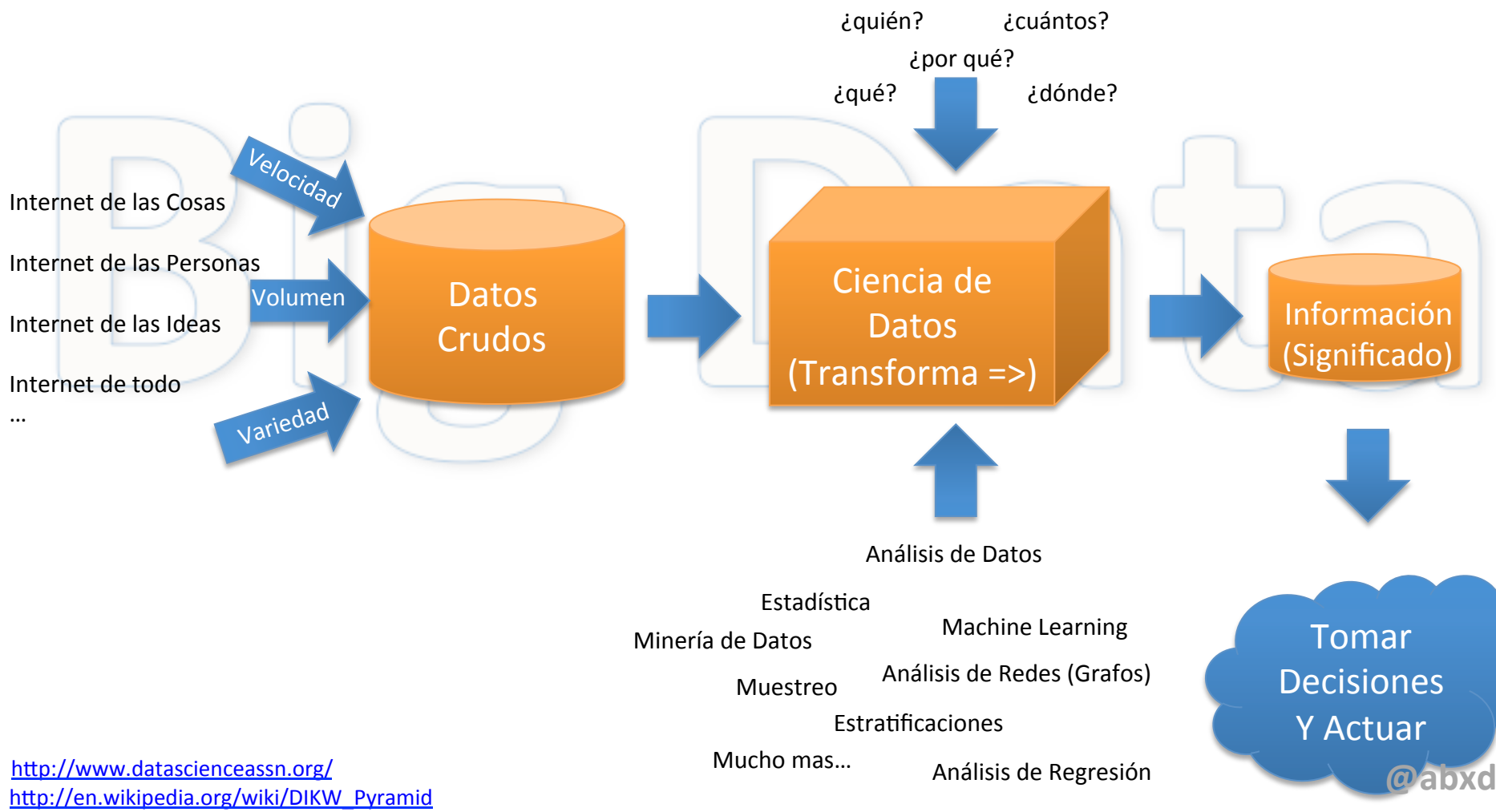
@abxda

Según Gartner:

Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced ***insight*** and decision making.

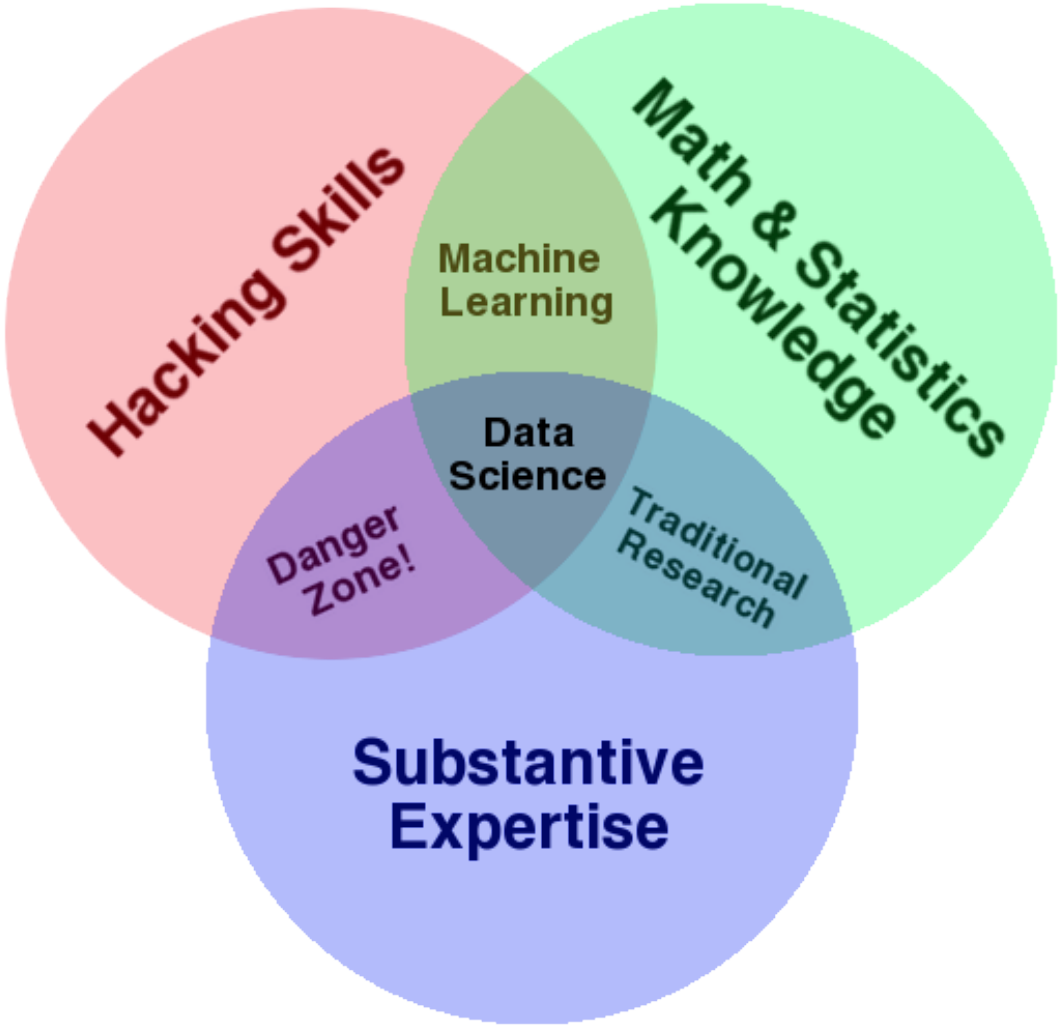
http://www.ft.com/intl/cms/e91a32d0-2bac-11e3-bfe2-00144feab7de.pdf

Big Data y Ciencia de Datos



http://www.datascienceassn.org/
http://en.wikipedia.org/wiki/DIKW_Pyramid

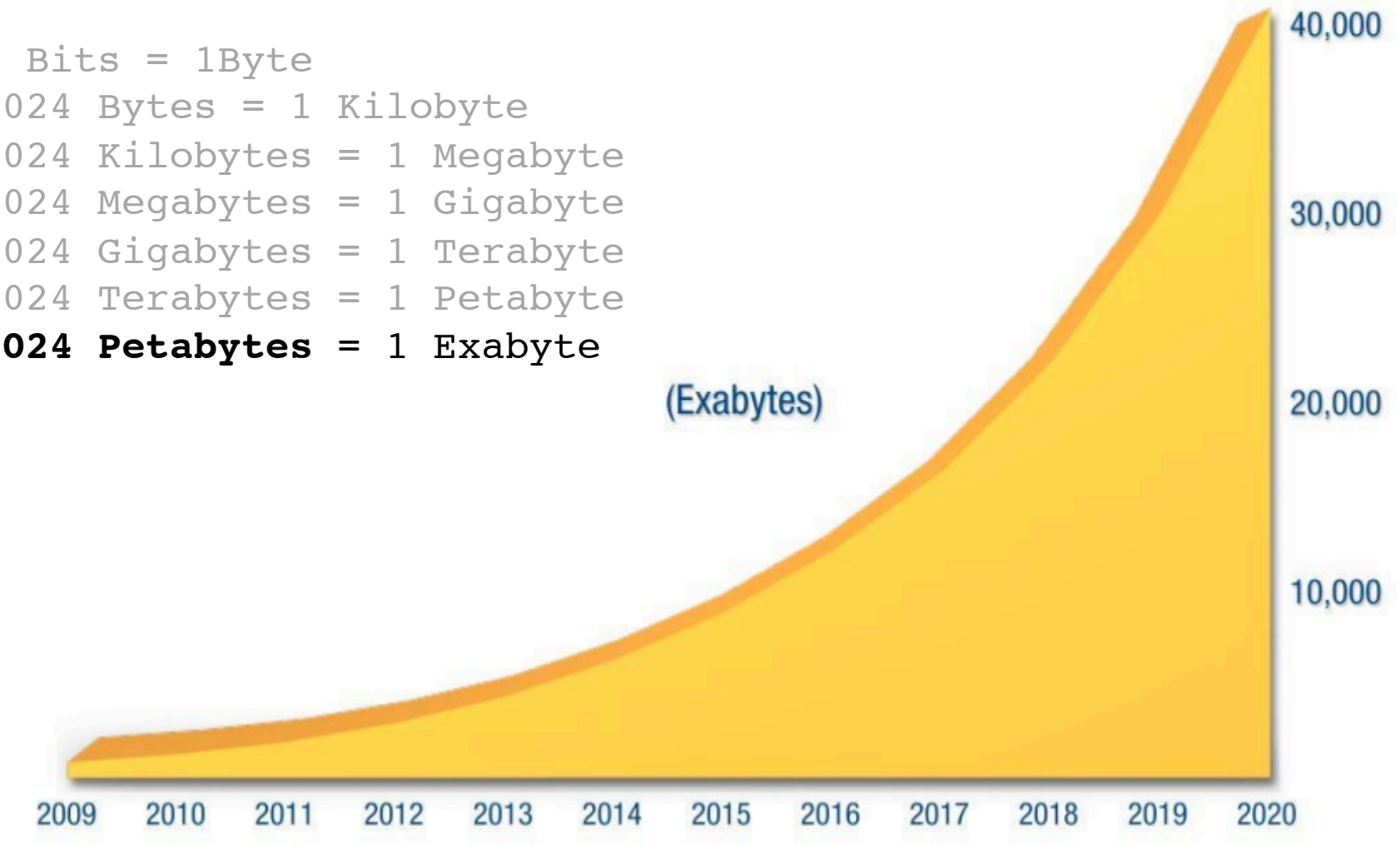
Científico de Datos



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

¿Qué tanto es tantito?

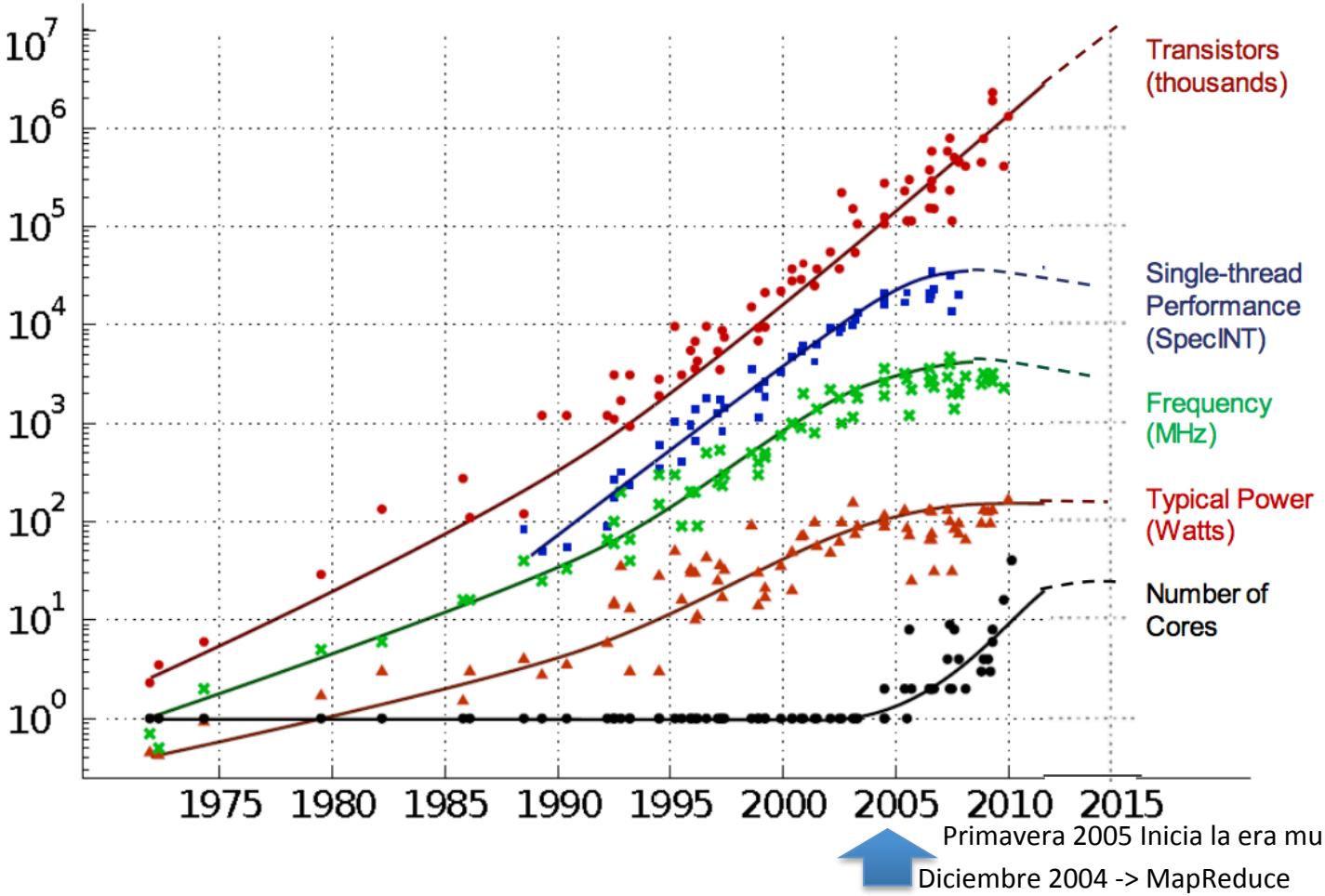
```
8 Bits = 1Byte
1024 Bytes = 1 Kilobyte
1024 Kilobytes = 1 Megabyte
1024 Megabytes = 1 Gigabyte
1024 Gigabytes = 1 Terabyte
1024 Terabytes = 1 Petabyte
1024 Petabytes = 1 Exabyte
```



Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

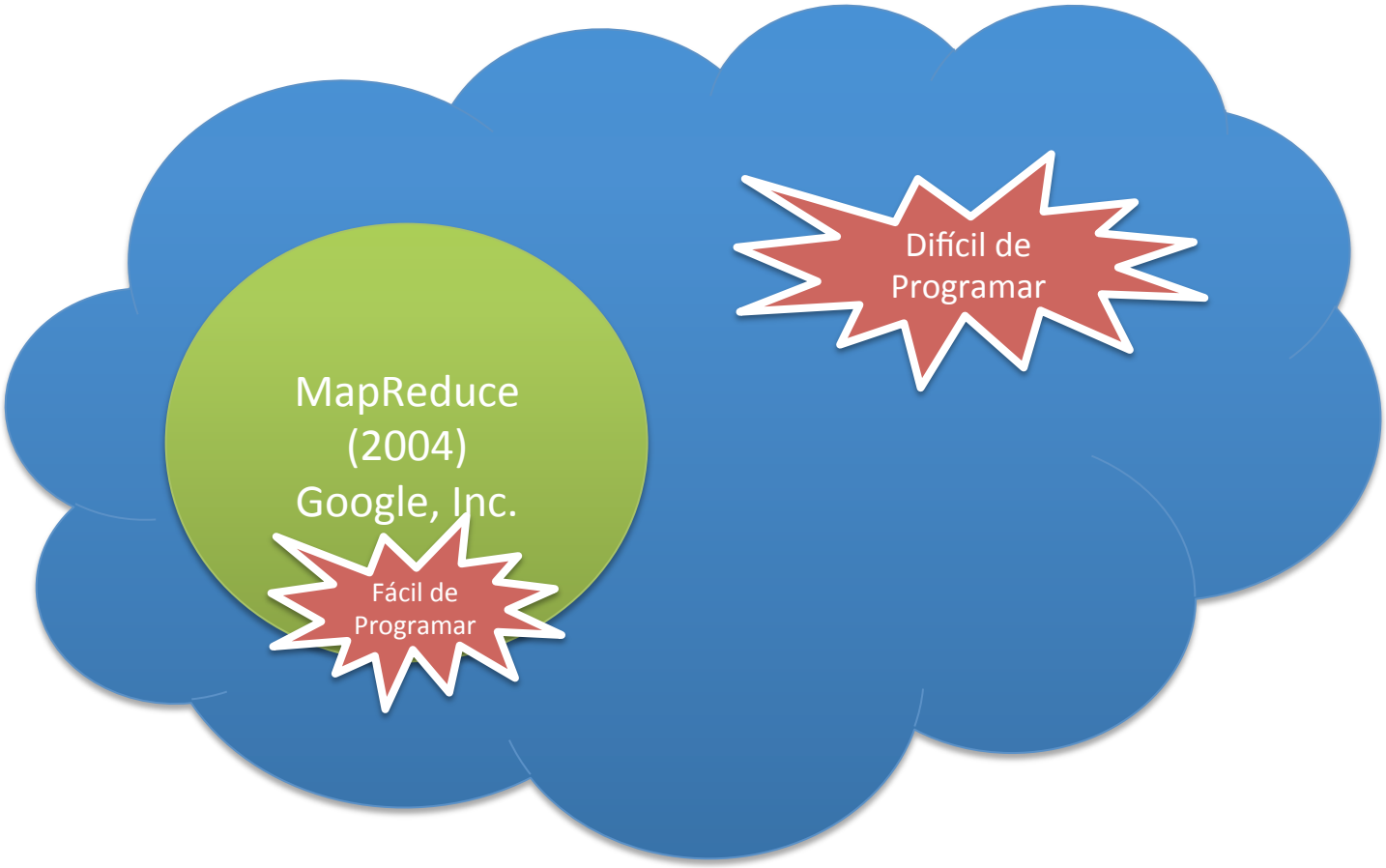
Era Multicore

35 años de Historia del Microprocesador



Original data collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond and C. Batten
Dotted line extrapolations by C. Moore

Computo en Paralelo

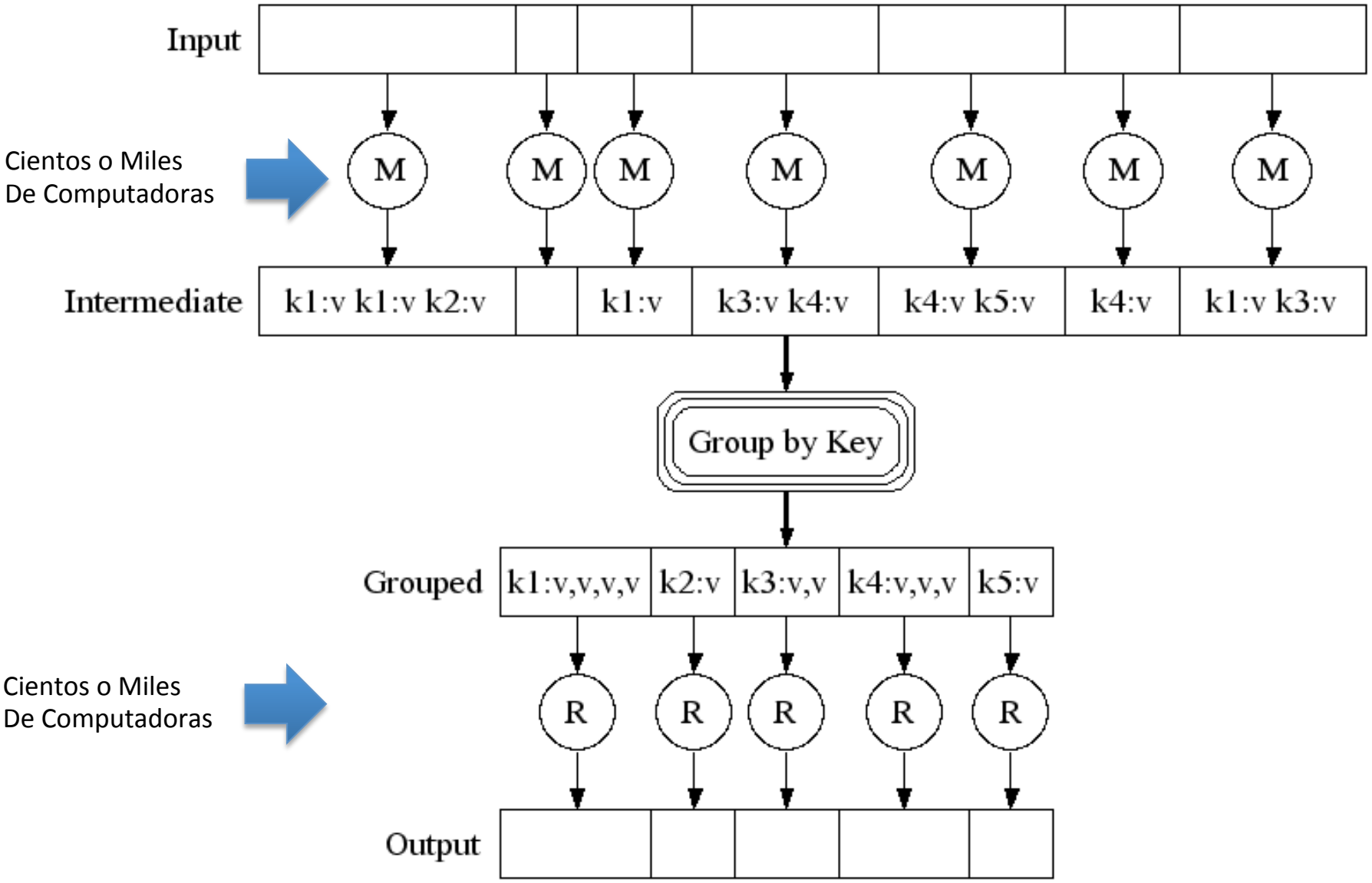


http://theory.stanford.edu/~sergei/papers/soda10-mrc.pdf

http://www.sciencedirect.com/science/article/pii/S1877050912001470

http://research.google.com/archive/mapreduce.html

MapReduce

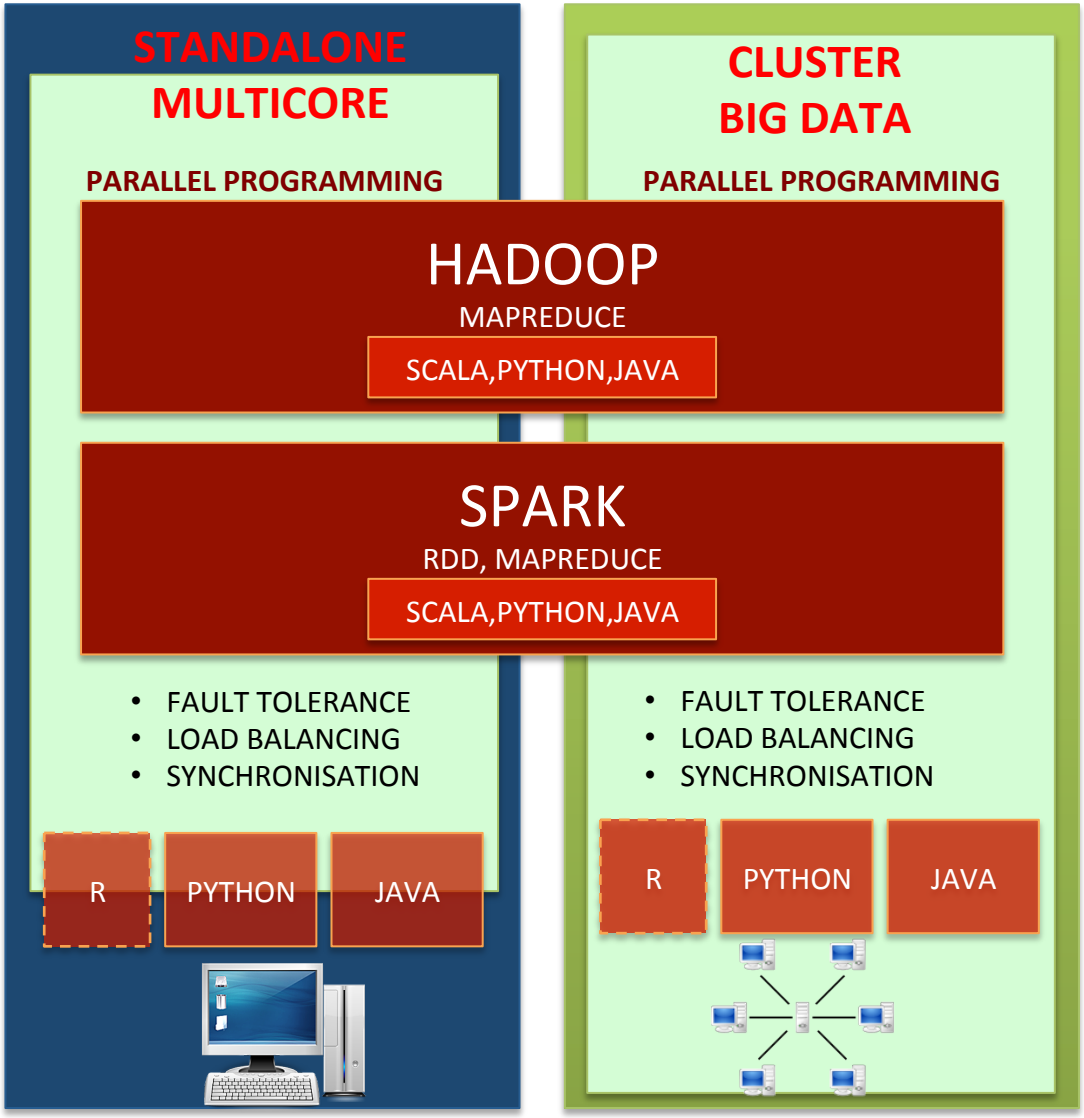


http://research.google.com/archive/mapreduce.html

MapReduce
(Pseudocódigo para contar palabras)

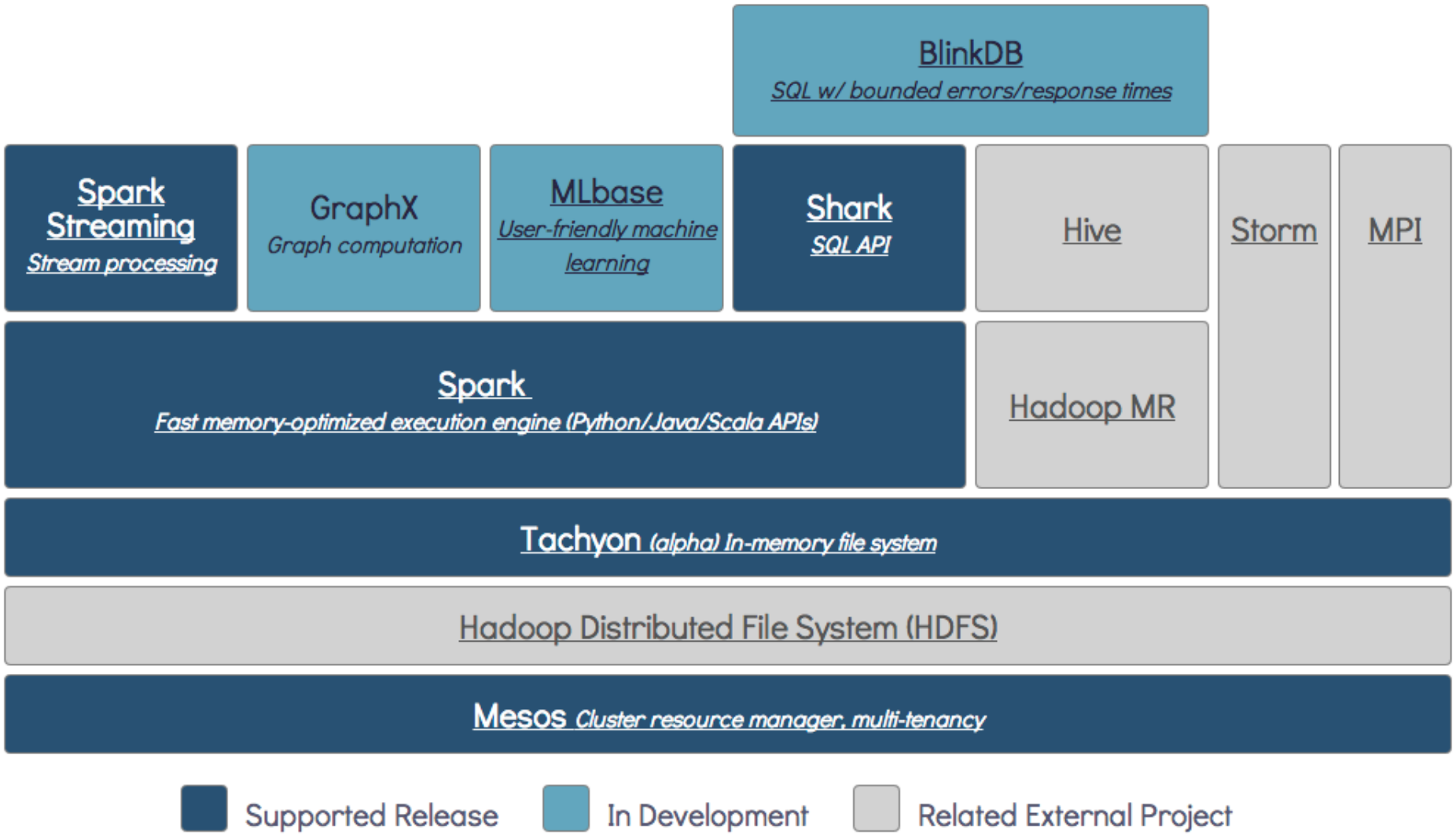
```
map(String input_key, String input_value):
    // input_key: document name
    // input_value: document contents
    for each word w in input_value:
      EmitIntermediate(w, "1");


reduce(String output_key, Iterator intermediate_values):
    // output_key: a word
    // output_values: a list of counts
    int result = 0;
    for each v in intermediate_values:
      result += ParseInt(v);
    Emit(AsString(result));
```

Herramientas



http://hadoop.apache.org/
http://spark.incubator.apache.org/

Spark una plataforma Big Data

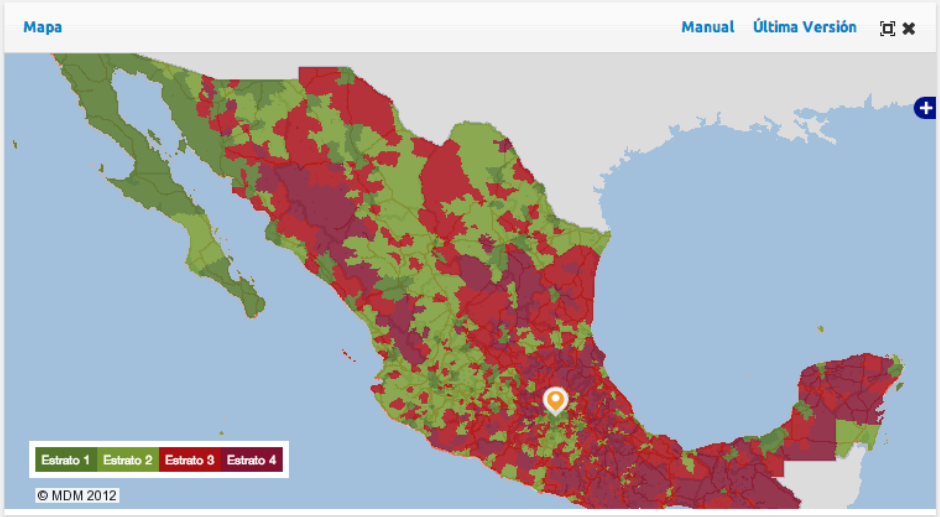


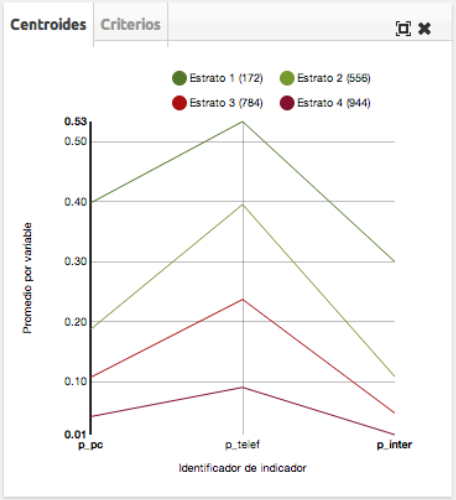
https://amplab.cs.berkeley.edu/software/

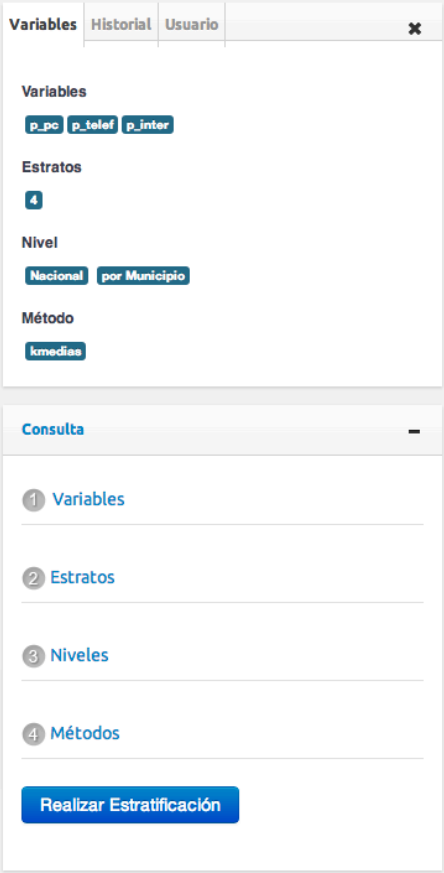
http://strataconf.com/strata2013/public/schedule/detail/27438

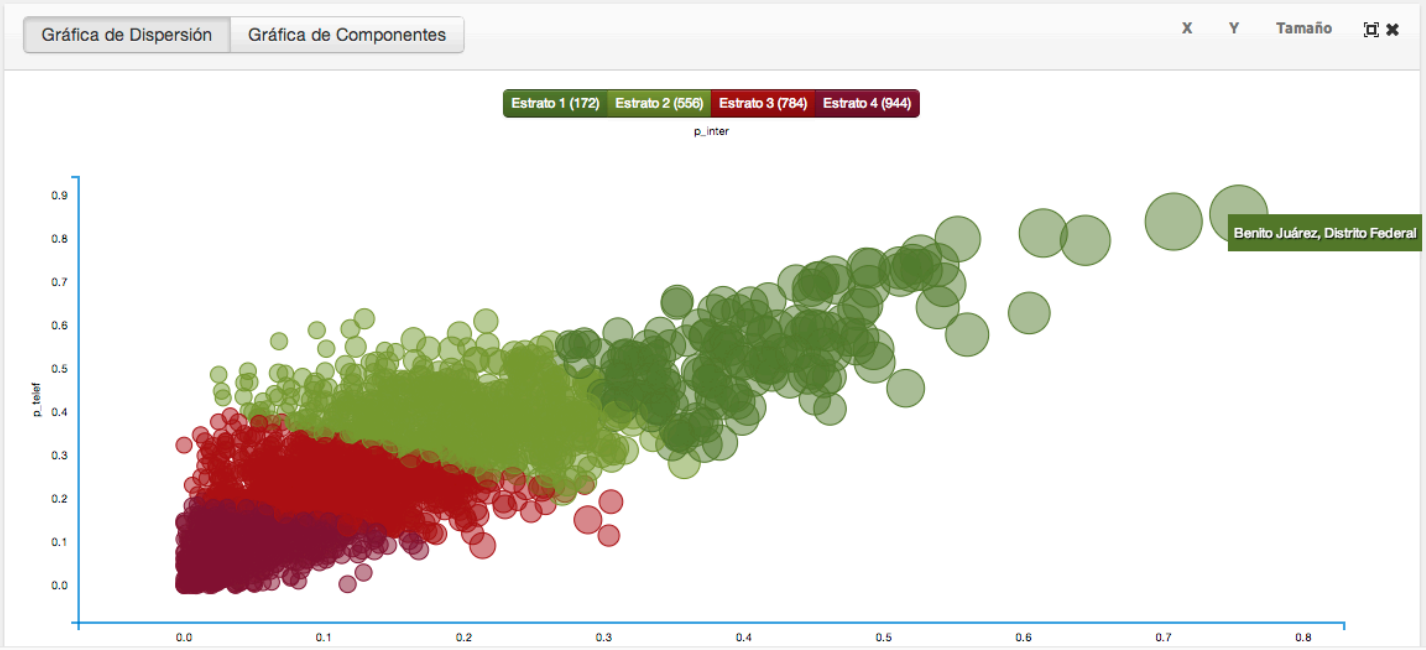
#sgvirtual

Ciencia de Datos en Acción





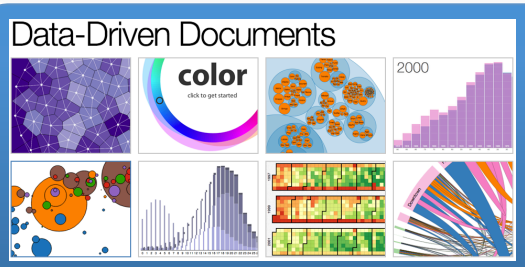




Créditos
www.inegi.org.mx/est/contenidos/Proyectos/estratificador/

@abxda

Tecnologías Involucradas



D3.js Librería JavaScript para creación de los gráficos vectoriales interactivos.

ANGULARJS by Google

Librería JavaScript facilita la incorporación del patrón MVC en aplicaciones web de una sola página.

HTML JS CSS B

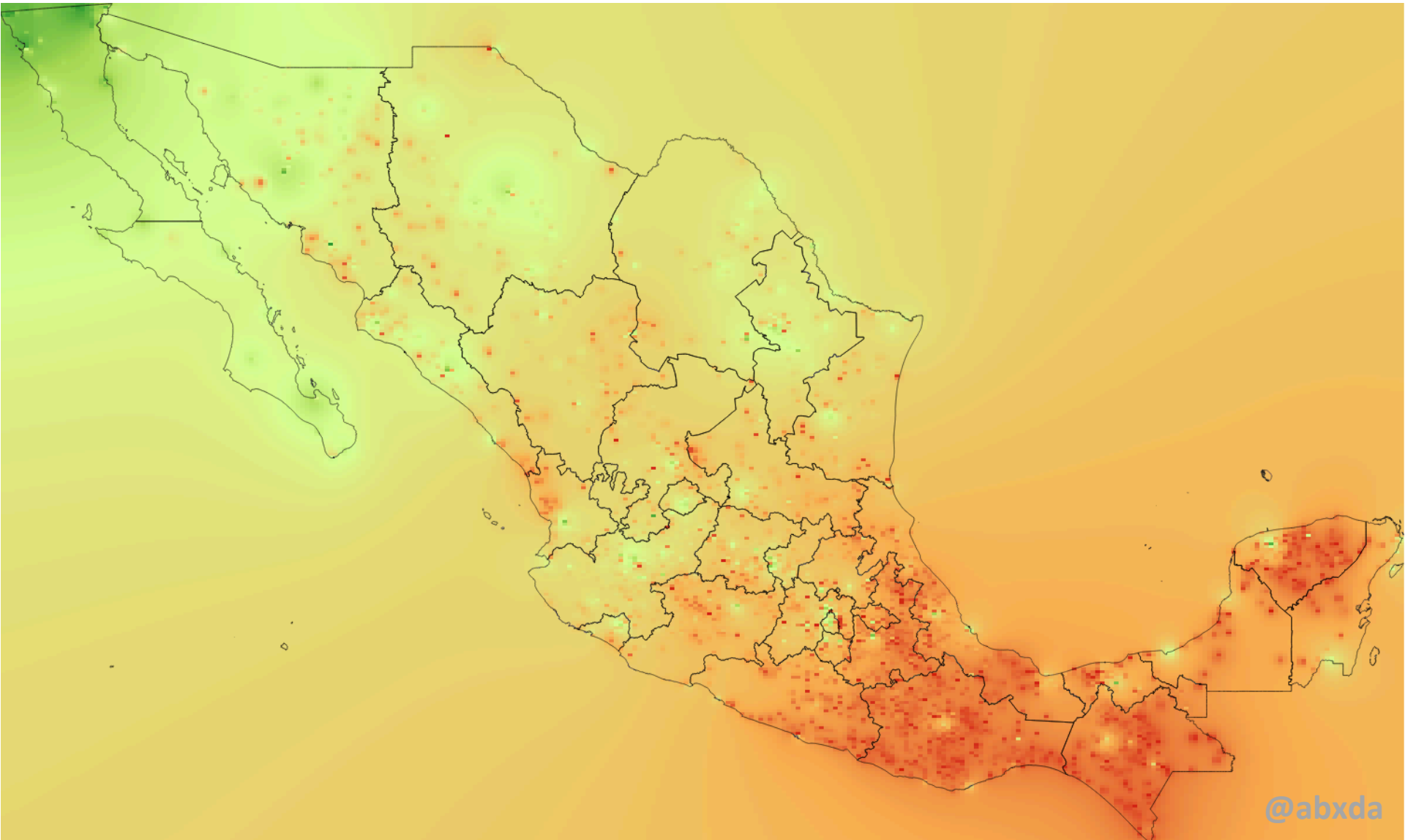
Diseño de estructura de la página y habilitación responsiva via Twitter Bootstrap.

{ JSON }

JSON formato de intercambio de datos.

Motor de análisis estadístico, habilitador de la inteligencia estadística.

Big Data en el Laboratorio



Tecnologías Involucradas

Qgis, Sistema de información geográfico Open Source.

Shark, SQL sobre Spark, compatible con Hive.

Mlbase, Librería para el desarrollo de algoritmos de Aprendizaje estadístico que corren sobre Spark

Spark, Plataforma Open Source de cómputo en cluster.

Scala, Lenguaje de programación funcional y orientado a objetos. Apto para paralelismo. Akka, habilita concurrencia.

C,S,V

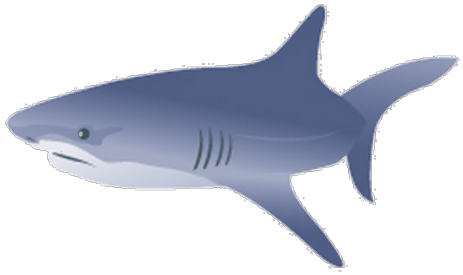
Archivo CSV con 1.2 millones de registros, 1 por manzana en el país. Cada uno con 168 variables censales.

Spark y MLBase

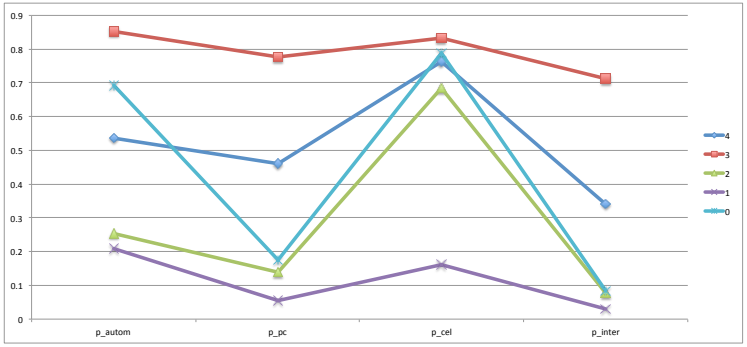

```
import org.apache.spark.mllib.clustering._

val manzanas = sc.textFile("/Users/abxda/…/datos.csv")
val subconjunto = manzanas.map(manzana => extractColumn(manzana))
points_nacional.cache
var modelo = KMeans.train(subconjunto, k=5, maxIterations=10)
val out = new PrintWriter("/Users/abxda/…/salida.csv")
subconjunto.collect.foreach(x => out.println(modelo.predict(x)))
out.close()
```

8 seg

Shark

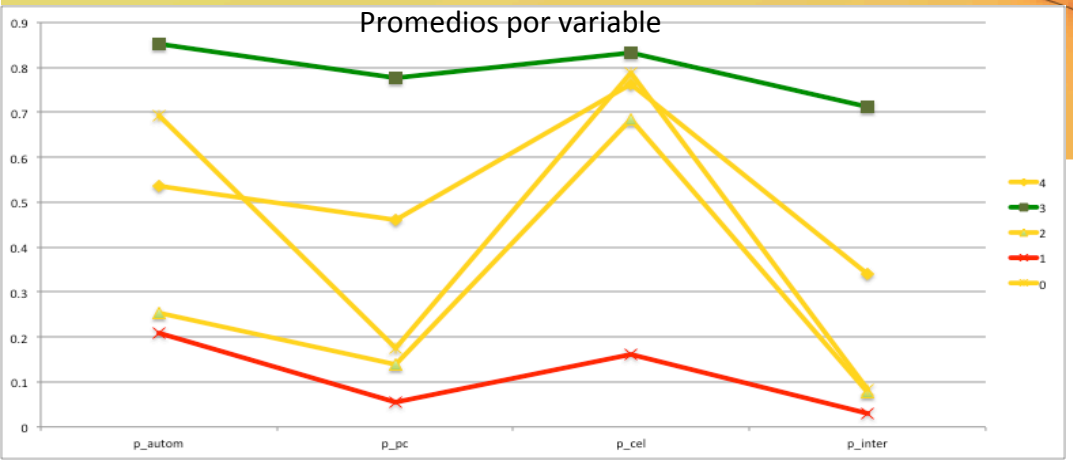


```
select
    estrato,
    avg(p_autom),
    avg(p_pc),
    avg(p_cel),
    avg(p_inter),
    count(*)
from salida group by estrato;
```



estrato	p_autom	p_pc	p_cel	p_inter	count(*)
4	0.536577059	0.46087735	0.76176366	0.340057367	308206
3	0.851219807	0.777557128	0.833951292	0.712273104	192934
2	0.254049418	0.139711048	0.683405158	0.076031984	376060
1	0.20981258	0.055136755	0.160281722	0.030043591	169243
0	0.693759231	0.176546203	0.788936165	0.084017414	174737
					1’221,180

#sgvirtual



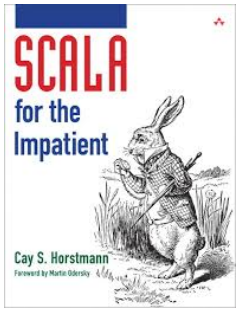
@abxda

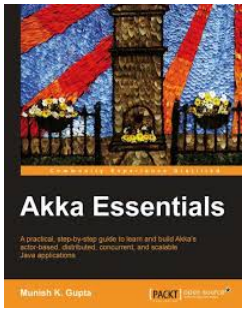
Equipo Big Data

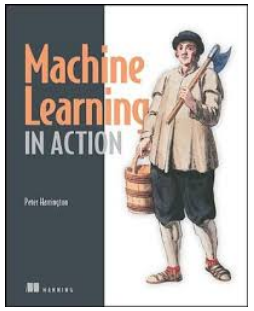
- **Cientificos de Datos**, expertos en integracion de soluciones Big Data (MapReduce, Scala, Machine Learning, Spark, R, Estadística).
- **Estadísticos**, expertos en modelado estadistico, enfoque en aprendizaje estadístico (R).
- **Desarrolladores de Software**, expertos en desarrollo de software (JavaScript, Arquitecturas de Software, Patrones de Diseño, Api's REST).
- **Diseñadores Gráficos**, expertos en presentación de información (HTML5, CSS3, JavaScript, Twitter Bootstrap).
- **Administradores de Sistemas**, expertos en arquitecturas de computo, infraestructura. Desde redes a clusters de computadoras (Linux).

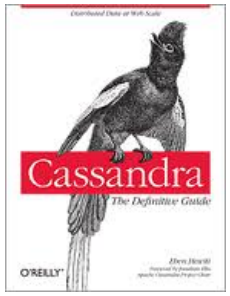
La tarea

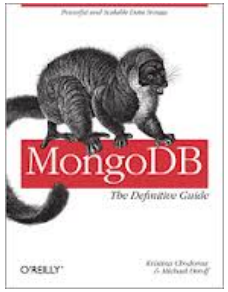
- Programación funcional
  - Scala
  - Akka
- Estadística
  - Probabilidad y Estadística
  - Muestreo
  - Machine Learning
  - R
- Almacenes de Datos NoSQL
  - Cassandra
  - MongoDB
  - Hbase
  - Neo4j
- Plataformas Big Data
  - Hadoop
  - Spark
- Visualización de Datos
  - D3.js

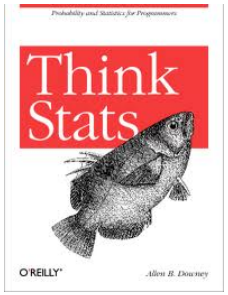


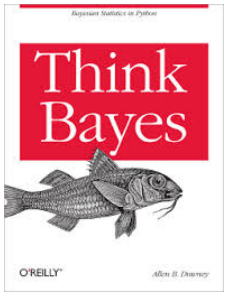


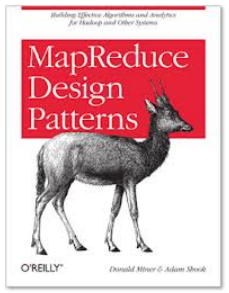


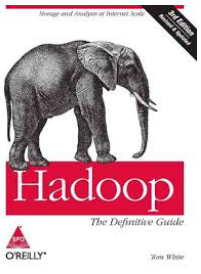


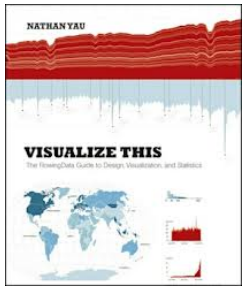


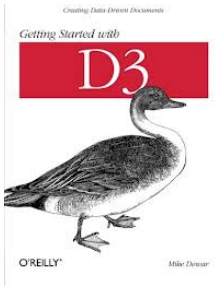


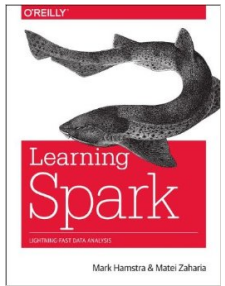


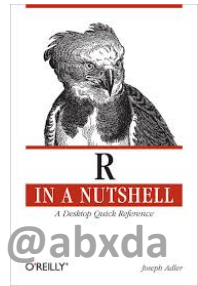












GRACIAS

Abel Alejandro Coronado Iruegas

Twitter : @abxda

http://abxda.wordpress.com/