Big Data Analytics and Apache Spark

Spark meetup BARCELONA

Jordi Torres

15/06/2015, UPC - BSC

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Support:

informàTICs
Enginyers en informàtica de Catalunya

FIB Alumni

tni)

tecnonews
la tecnología a un click

HiPEAC

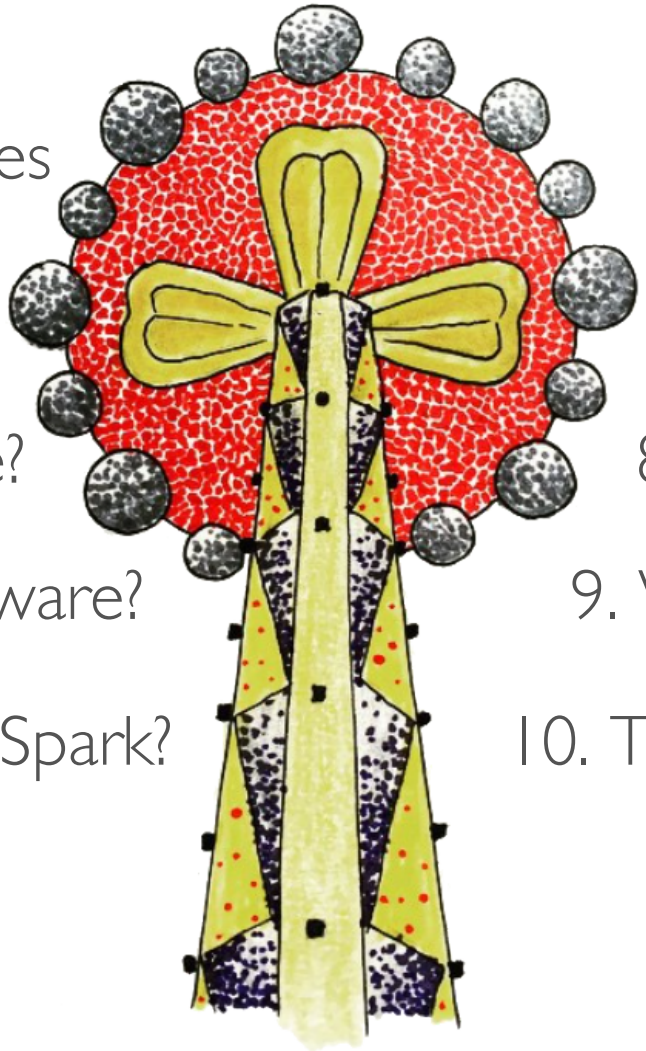
TARGET?


QUAHOG'S
Most Wanted!
Learning
Spark


O'REILLY
Learning
Spark

TALK OUTLINE

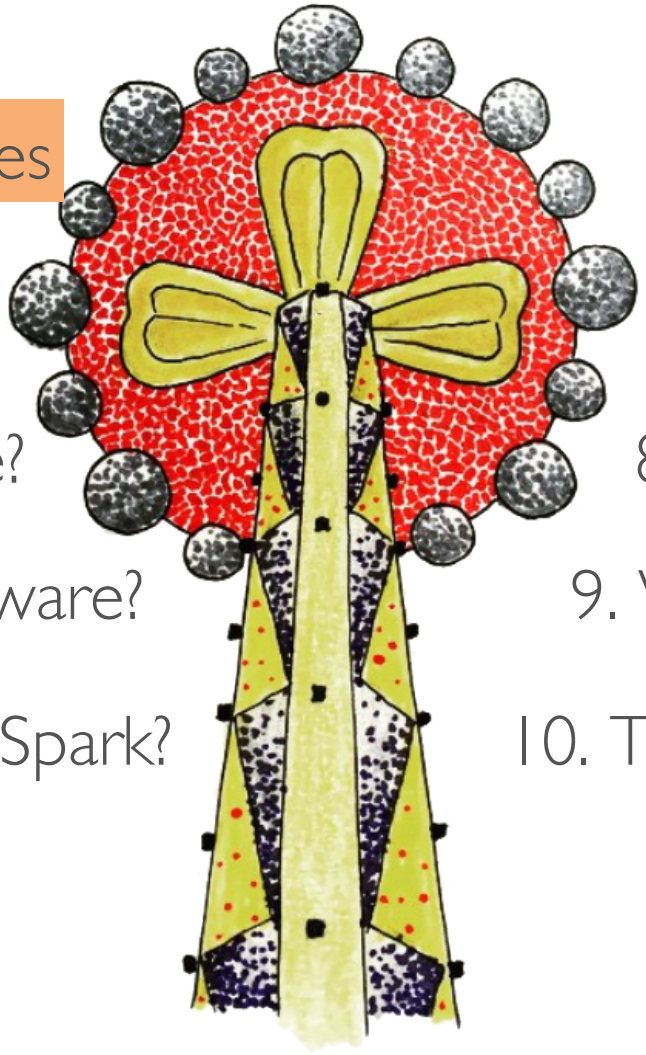
1. Computing Waves
2. Why now?
3. Future Hardware?
4. Software Middleware?
5. What is Apache Spark?
6. Spark Basics
7. Spark Ecosystem
8. Spark & Marenostrum
9. What next?
10. To learn more …



Spark Meetup BARCELONA

TALK OUTLINE

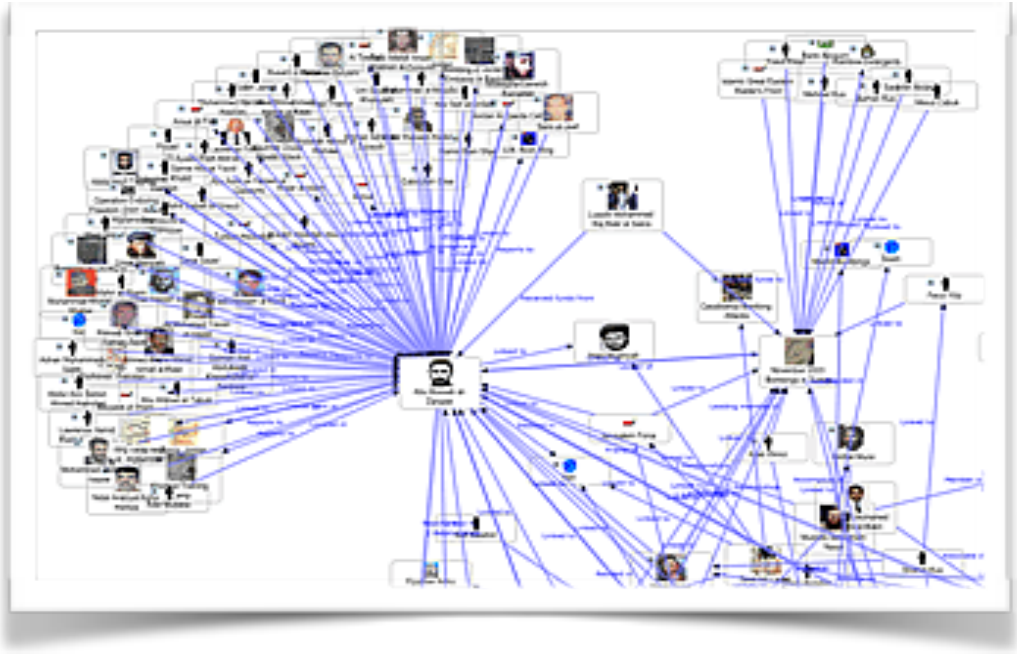
1. Computing Waves
2. Why now?
3. Future Hardware?
4. Software Middleware?
5. What is Apache Spark?
6. Spark Basics
7. Spark Ecosystem
8. Spark & Marenostrum
9. What next?
10. To learn more …



Spark Meetup BARCELONA

COMPUTING WAVES

The first wave of computing made numbers computable



The second wave has made text and rich media computable and accessible digitally

COMPUTING WAVES

We are in the next wave that will also make context computable

Source: www.DesignedInBarcelona.com

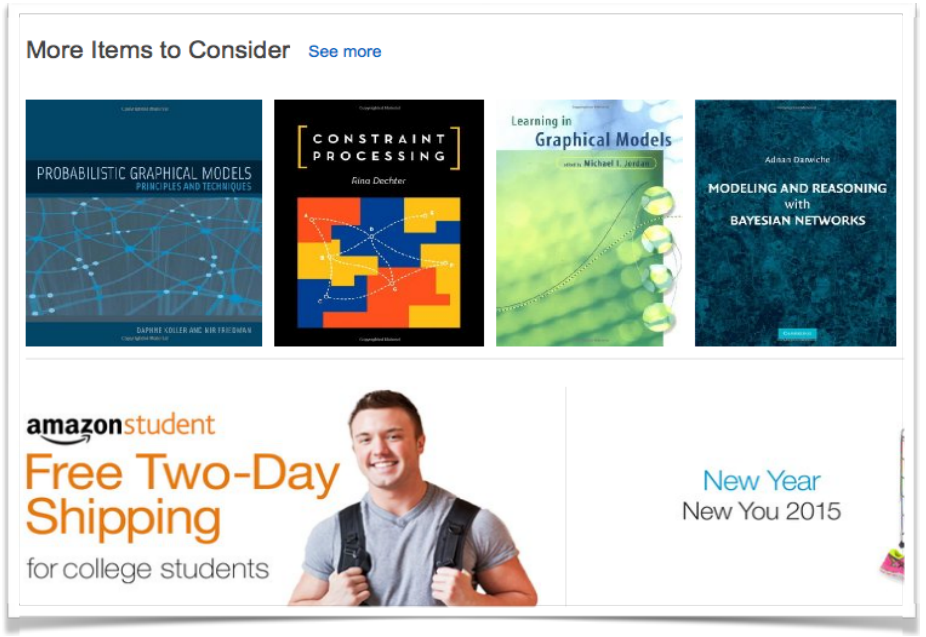
TECHNOLOGY IS OFTEN SO SUBTLE

The consumer have no idea that computation is actually helping make their decisions!

Source: www.DesignedInBarcelona.com

EXAMPLE: ONLINE SHOPPING

Amazon's recommendation engine (over 250M customers)



Amazon predict items and send it to your nearest delivery hub

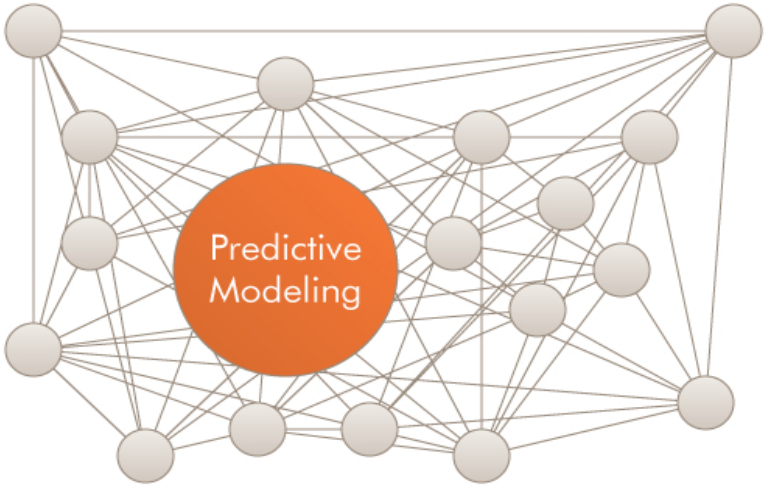
EXAMPLE: SHOPPING

In 2012 Target identified a pregnant teenager before her family knew about her condition.

They've identified 25 items, purchased in a particular order, …

HOW THEY DO IT?

Using predictive models



They can be achieved with programs that use machine learning algorithms.

```
input : {D_{1:N}, ε1, ε2, ε3, Minpts, k, u}
output: {Cluster_Labels}

Means = list();
Temporal_Label = −2;
Cluster_Label = −1;
T_{1:N} = FindTopicDistributions(D_{1:N});
for d ← D_{1:N} do
    if d ∈ Cluster then
        next;
    end
    Neigh = FindNeighbors(d, D_{1:N}, T_{1:N}, ε1, ε2, ε3);
    if |Neigh| < Minpts or diversity < u then
        Mark d as Noise;
    else
        Mark d with Temporal_Label;
        Cand.Push(Neigh)
        while |Cand| > 0 do
            o = Cand.Pop();
            if o ∉ Cluster then
                Mark o with Temporal_Label;
                NewNeigh = FindNeighbors(o, D_{1:N}, T_{1:N}, ε1, ε2,
                ε3);
                if |NewNeigh| ≥ Minpts and diversity ≥ u then
                    Cand.Push(NewNeigh)
                end
            end
        end
    end
    mean = Mean(T_{1:N}[Temporal_Label]);
    if ∃ FindSimilarCluster(k) then
        Cluster_Label =FindSimilarCluster(k);
        Means[FindSimilarCluster(k)] = mean;
    else
        Cluster_Label = max(Cluster_Label) + 1;
        Means.Push(mean);
    end
    Mark D_{1:N}[Temporal_Label] with Cluster_Label;
end
```

Source: http://bigsonata.com/wp-content/uploads/2014/07/PredictiveModeling.jpg

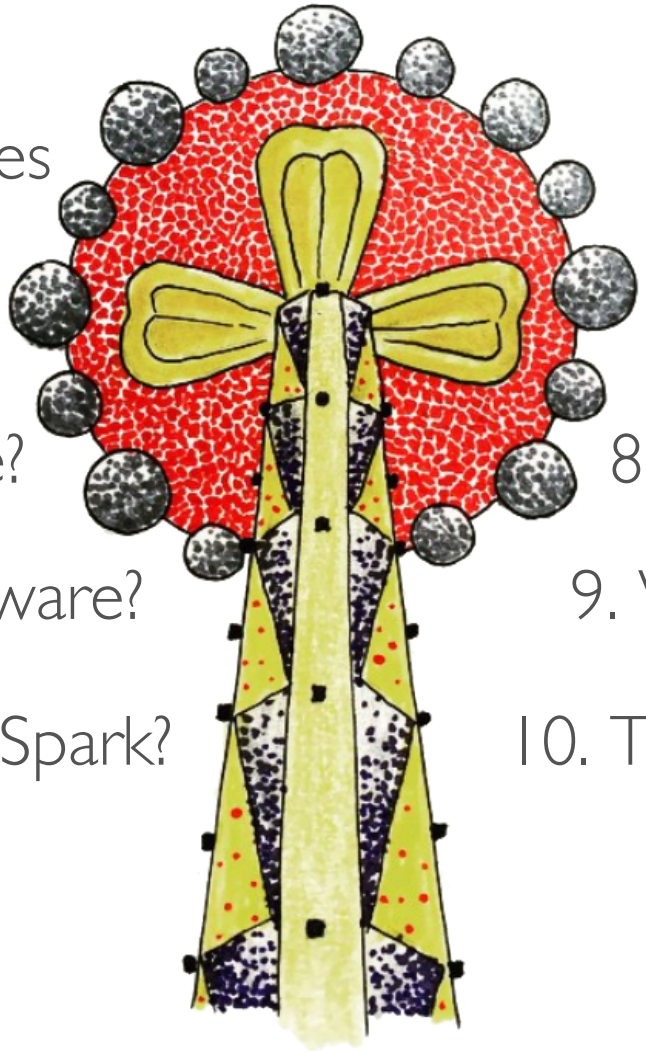
OLD WINE IN A NEW BOTTLE

Artificial Intelligence plays an important role

- the term itself dates from the 1950s.
- periods of hype and high expectations alternating with periods of setback and disappointment.

TALK OUTLINE

1. Computing Waves
2. Why now?
3. Future Hardware?
4. Software Middleware?
5. What is Apache Spark?
6. Spark Basics
7. Spark Ecosystem
8. Spark & Marenostrum
9. What next?
10. To learn more …



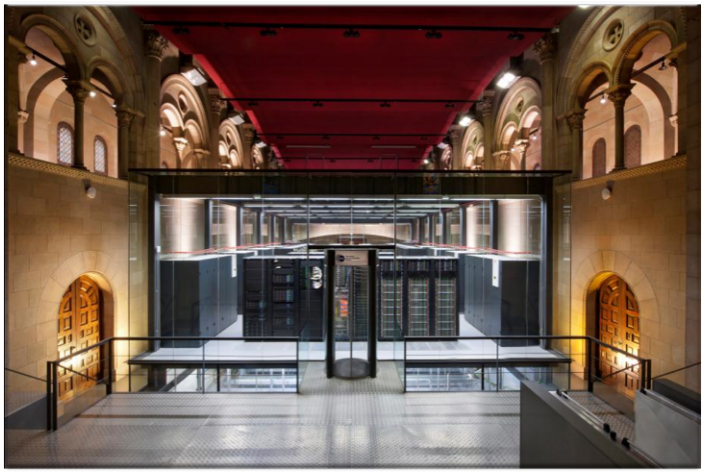
Spark meetup BARCELONA

WHY NOW?

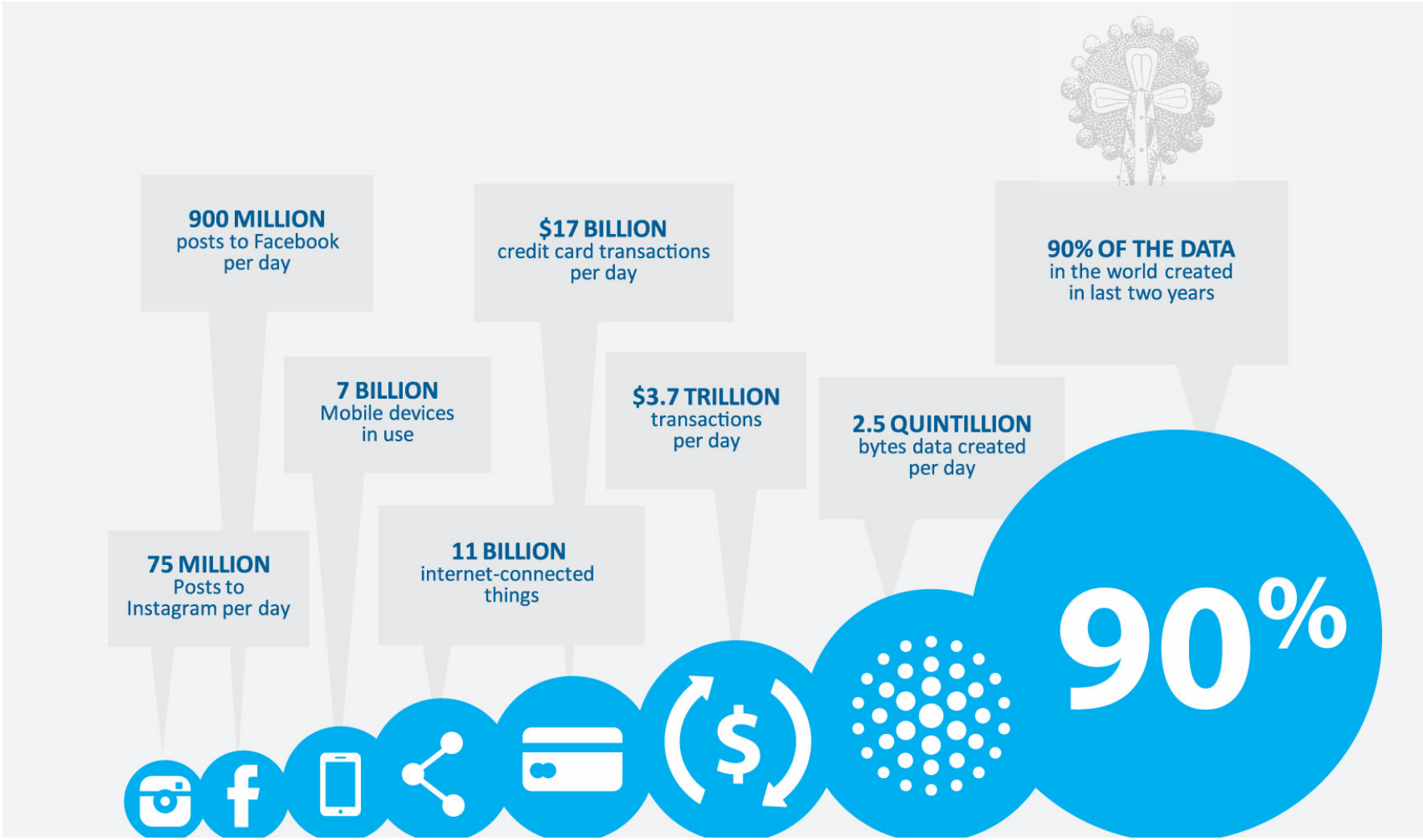
1. Along the explosion of data …

now algorithms can be "trained" by exposing them to large data sets that were previously unavailable.

2. And the computing power necessary to implement these algorithms are now available

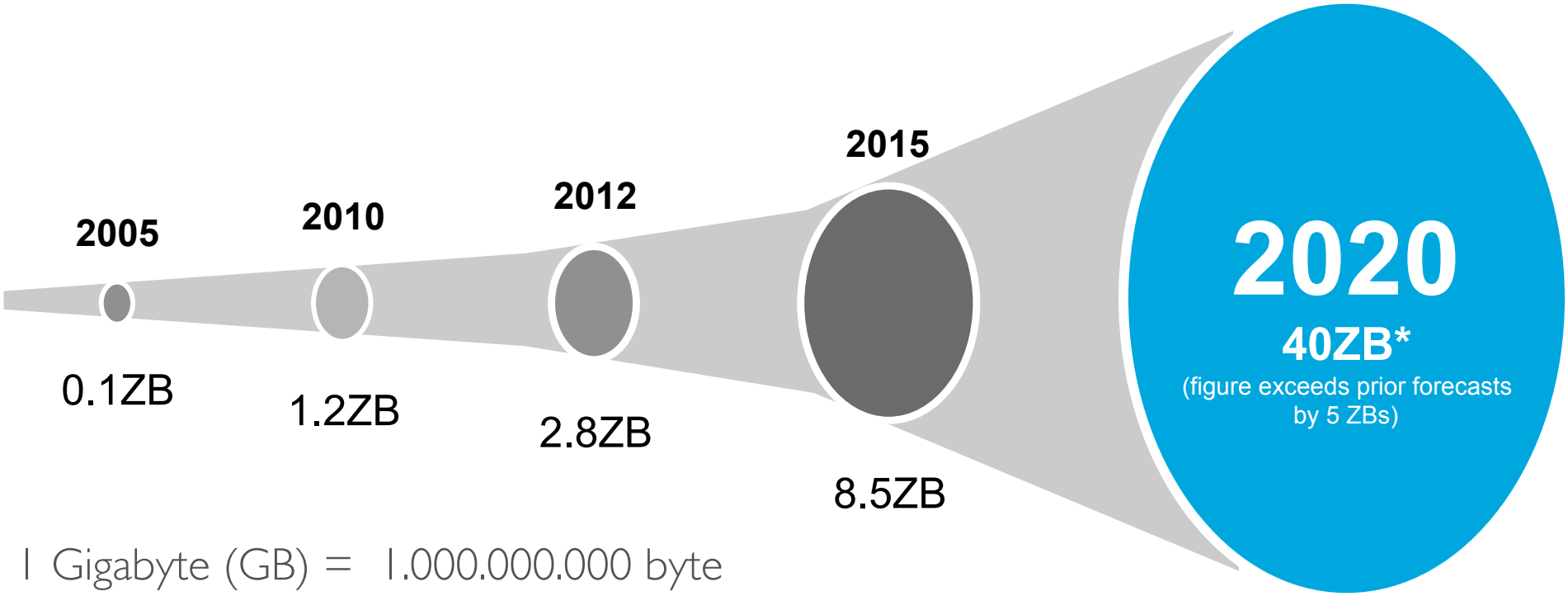


THE DATA DELUGE



Source: http://sysorexhosting.com/wp-content/uploads/big-data-infographic.jpg

THE DATA DELUGE



- 1 Gigabyte (GB) = 1.000.000.000 byte
- 1 Terabyte (TB) = 1.000 (GB)
- 1 Petabyte (PB) = 1.000.000 (GB)
- 1 Exabyte (EB) = 1.000.000.000 (GB)
- 1 Zettabyte (ZB) = 1.000.000.000.000 (GB)

Source: http://sysorexhosting.com/wp-content/uploads/big-data-infographic.jpg

EVOLUTION OF COMPUTING POWER

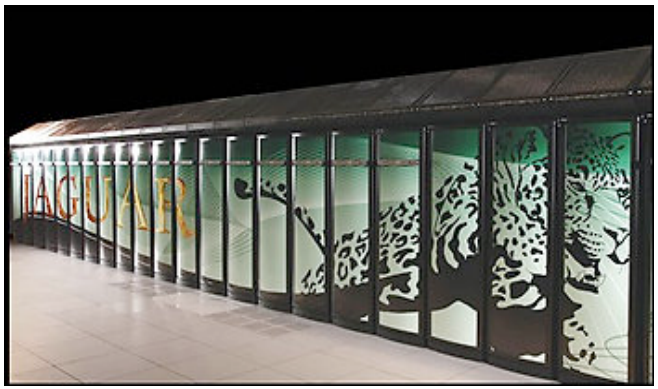
FLOP/second

E P T G

1000000000000000000

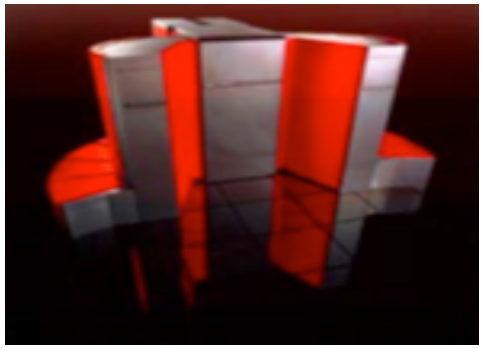
~2019

? ($1x10^7$ processadors



2008

Cray XT5 (15000 processadors)



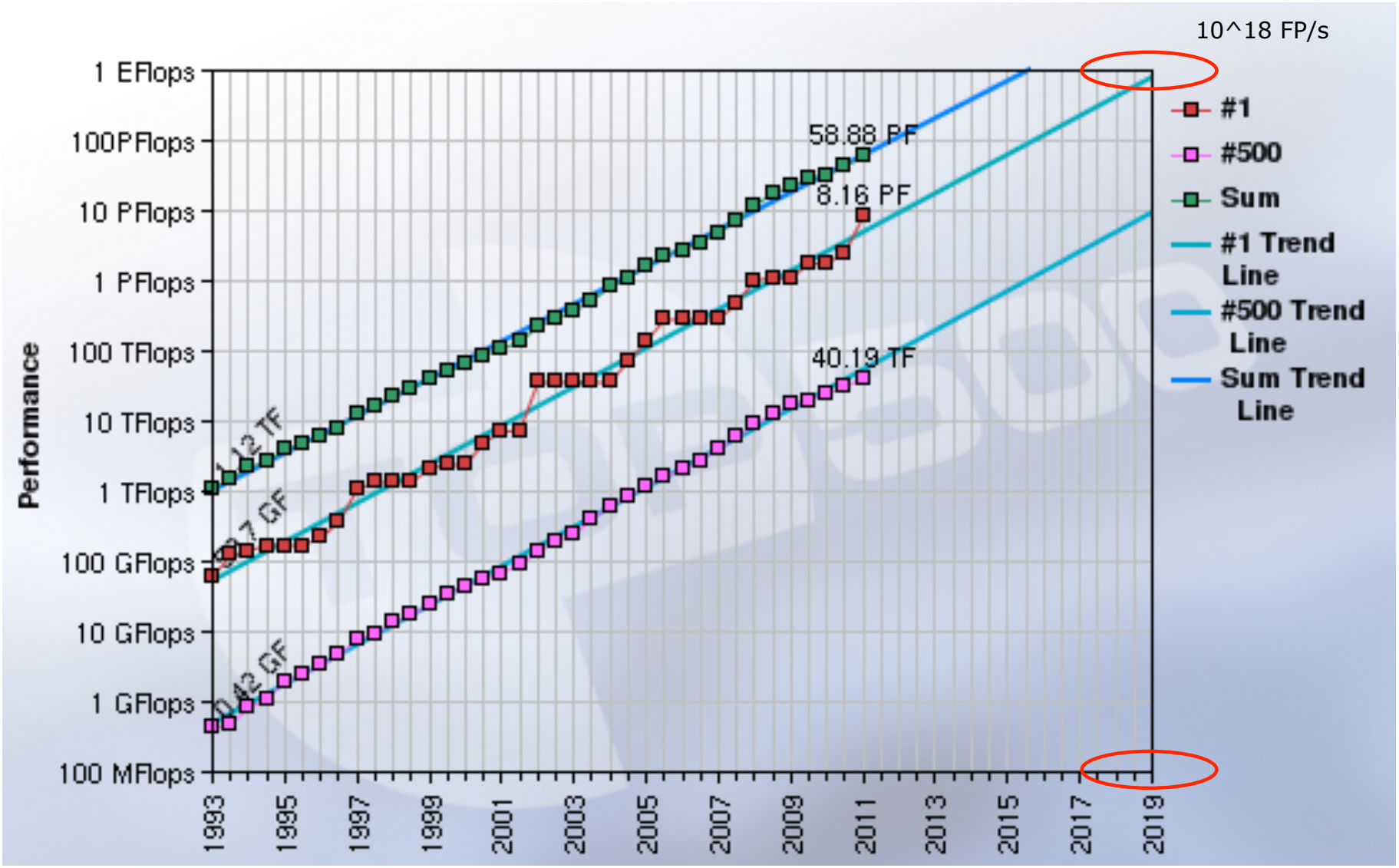
1988

Cray Y-MP (8 processadors)

1998

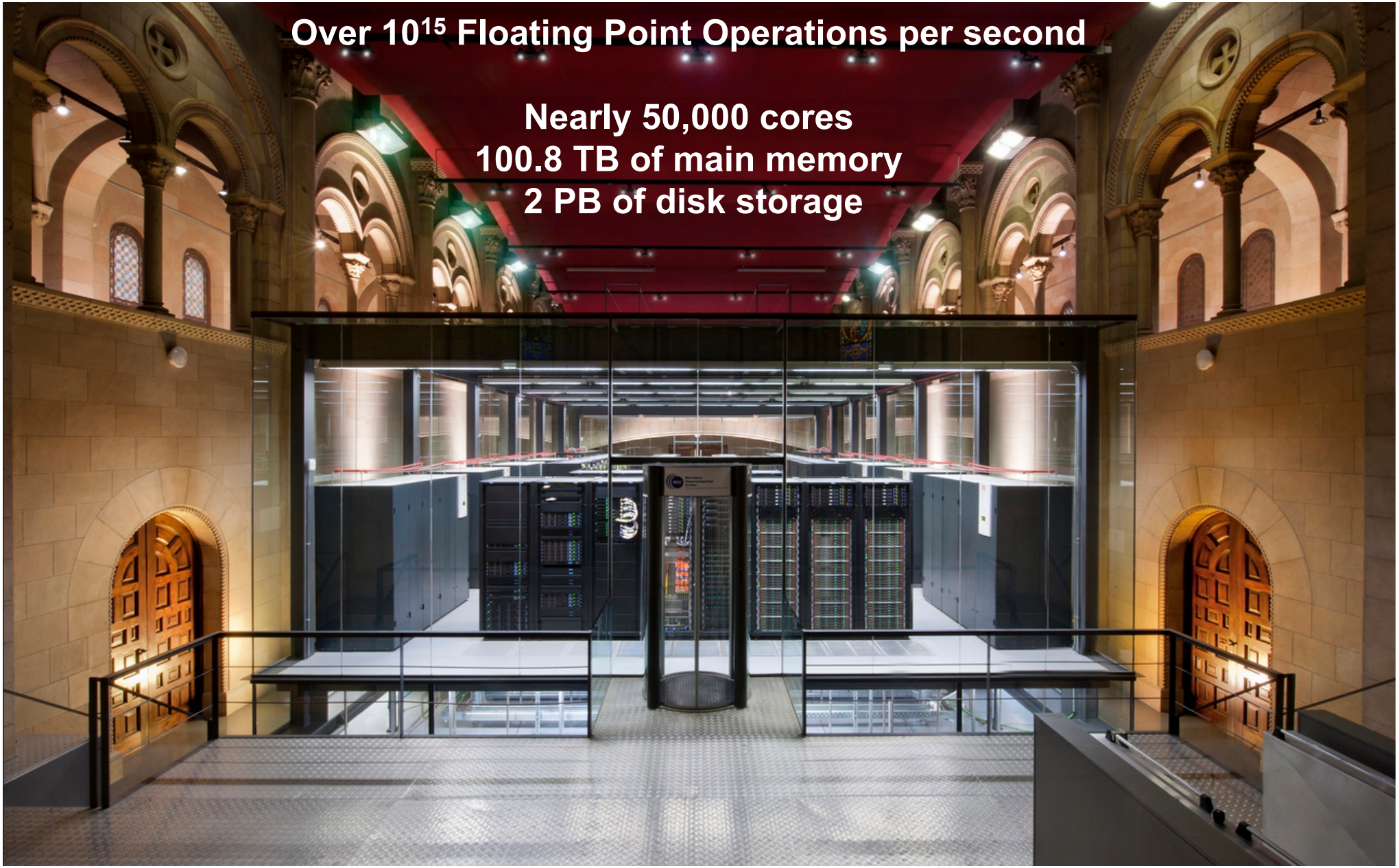
Cray T3E (1024 processadors)

TOP 500 "FORMULA 1" OF COMPUTING



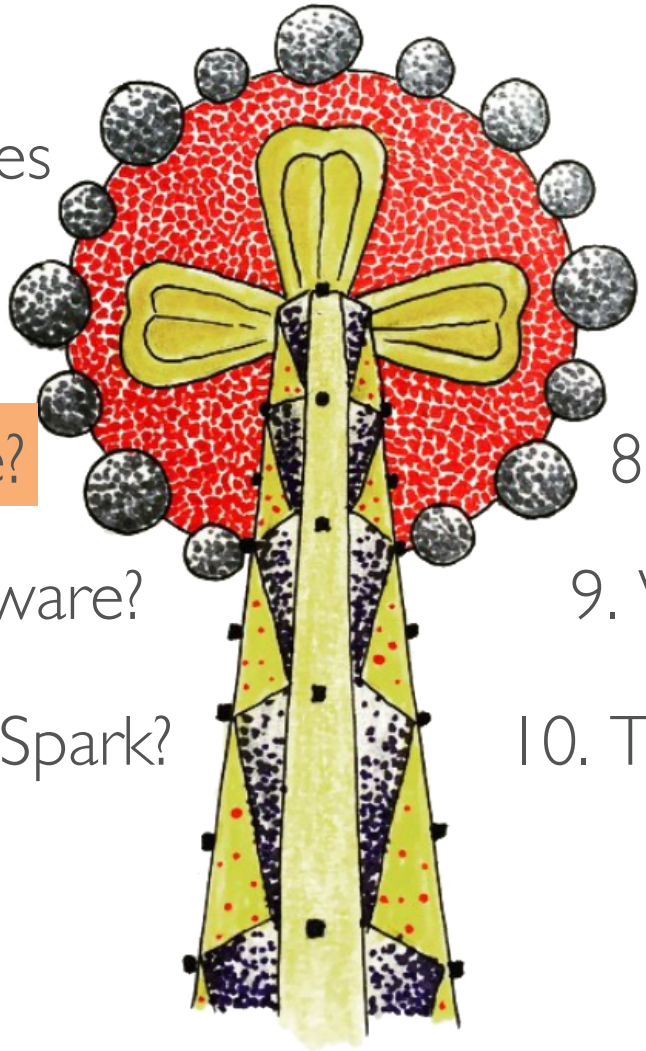
Source: top500.org

"EL NOSTRE" FORMULA I



TALK OUTLINE

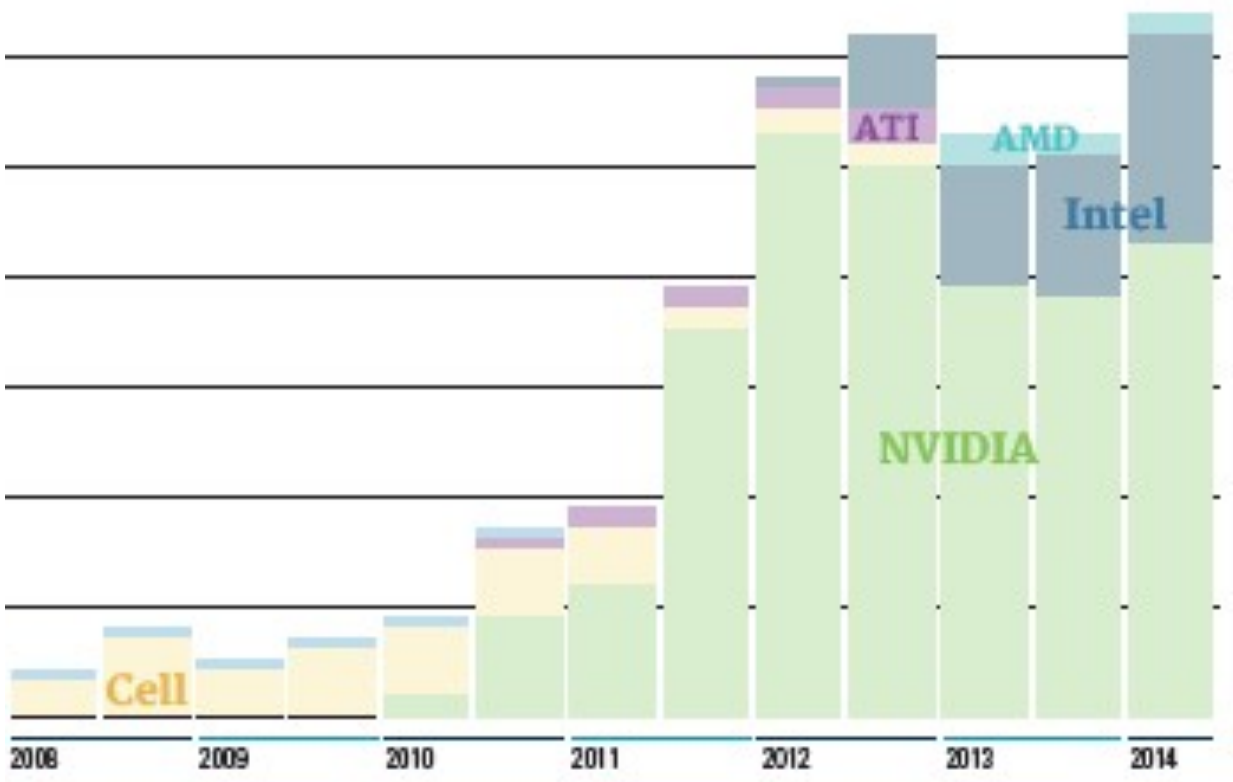
1. Computing Waves
2. Why now?
3. Future Hardware?
4. Software Middleware?
5. What is Apache Spark?
6. Spark Basics
7. Spark Ecosystem
8. Spark & Marenostrum
9. What next?
10. To learn more …



Spark meetup BARCELONA

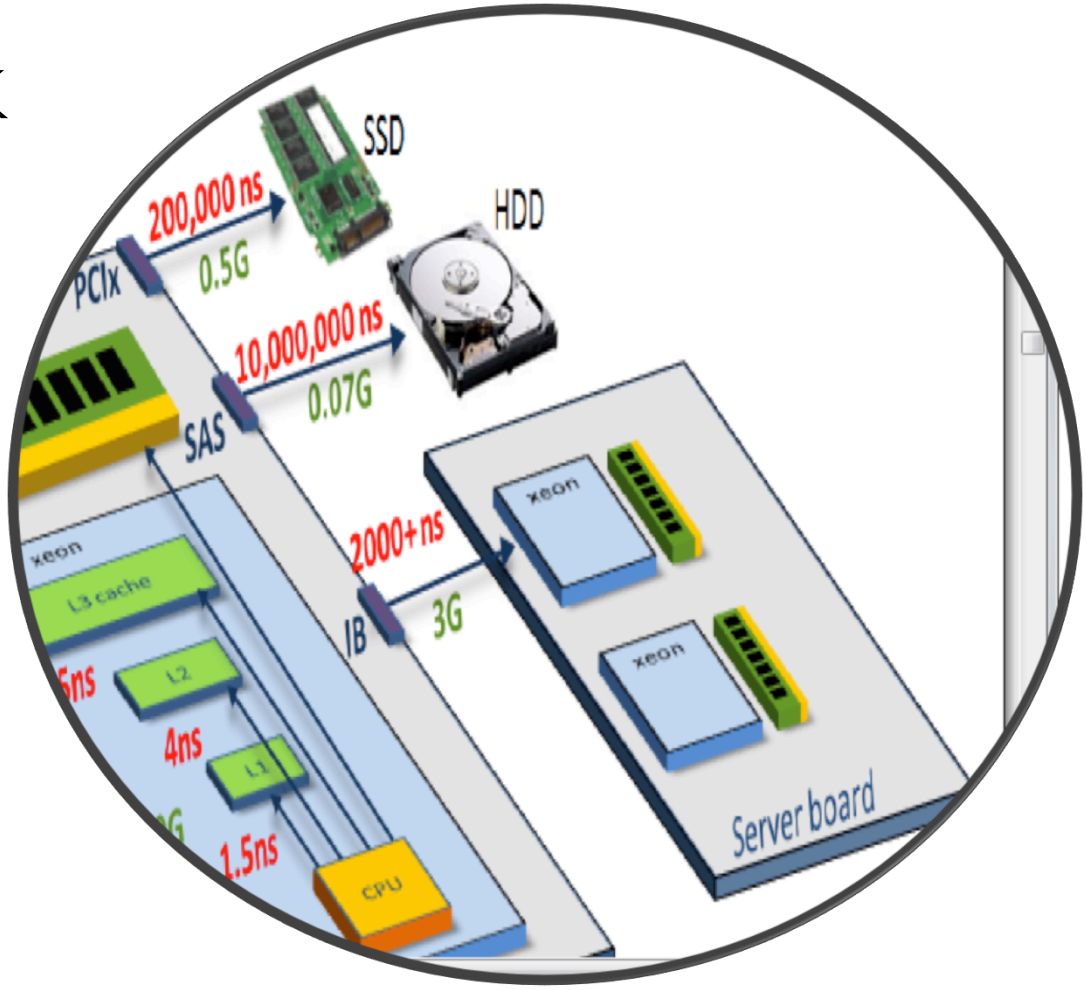
FUTURE HARDWARE?

Multicore and accelerators technology



FUTURE HARDWARE?

Increased network bandwidth

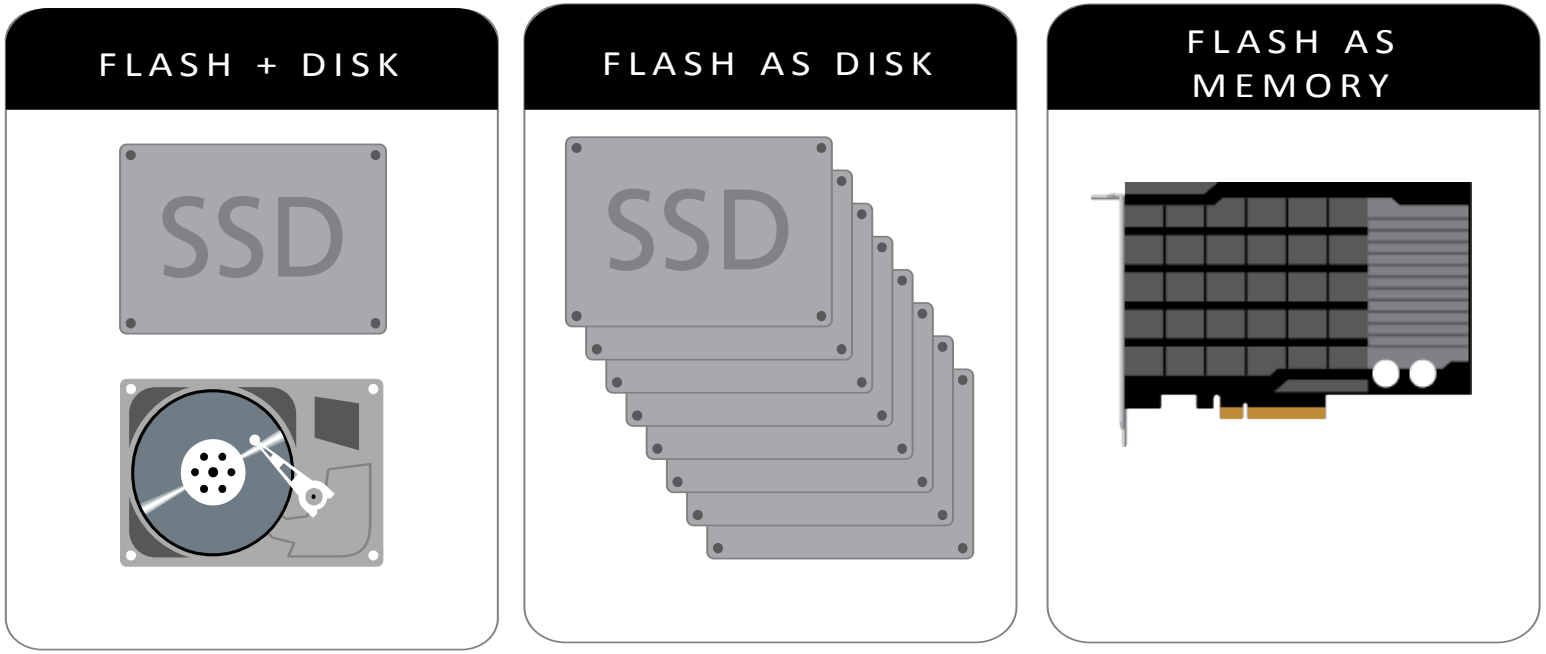


IB latency 2000ns is only 20x slower than RAM and is 100x faster than SSD

Source: http://www.slideshare.net/blopeur/hecatonchire-kvm-forum2012benoithudzia

FUTURE HARDWARE?

Flash Technology



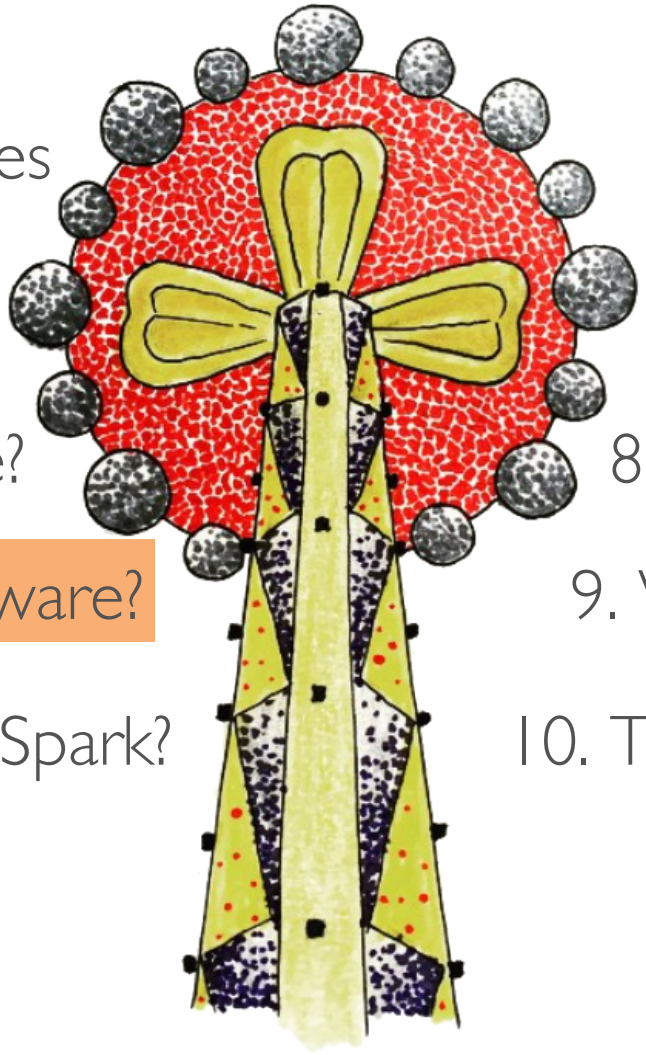
Source: David Carrera , BSC-UPC

Non Volatile Memory evolution

(*) HHD 100 cheaper than RAM . But 1000 times slower

TALK OUTLINE

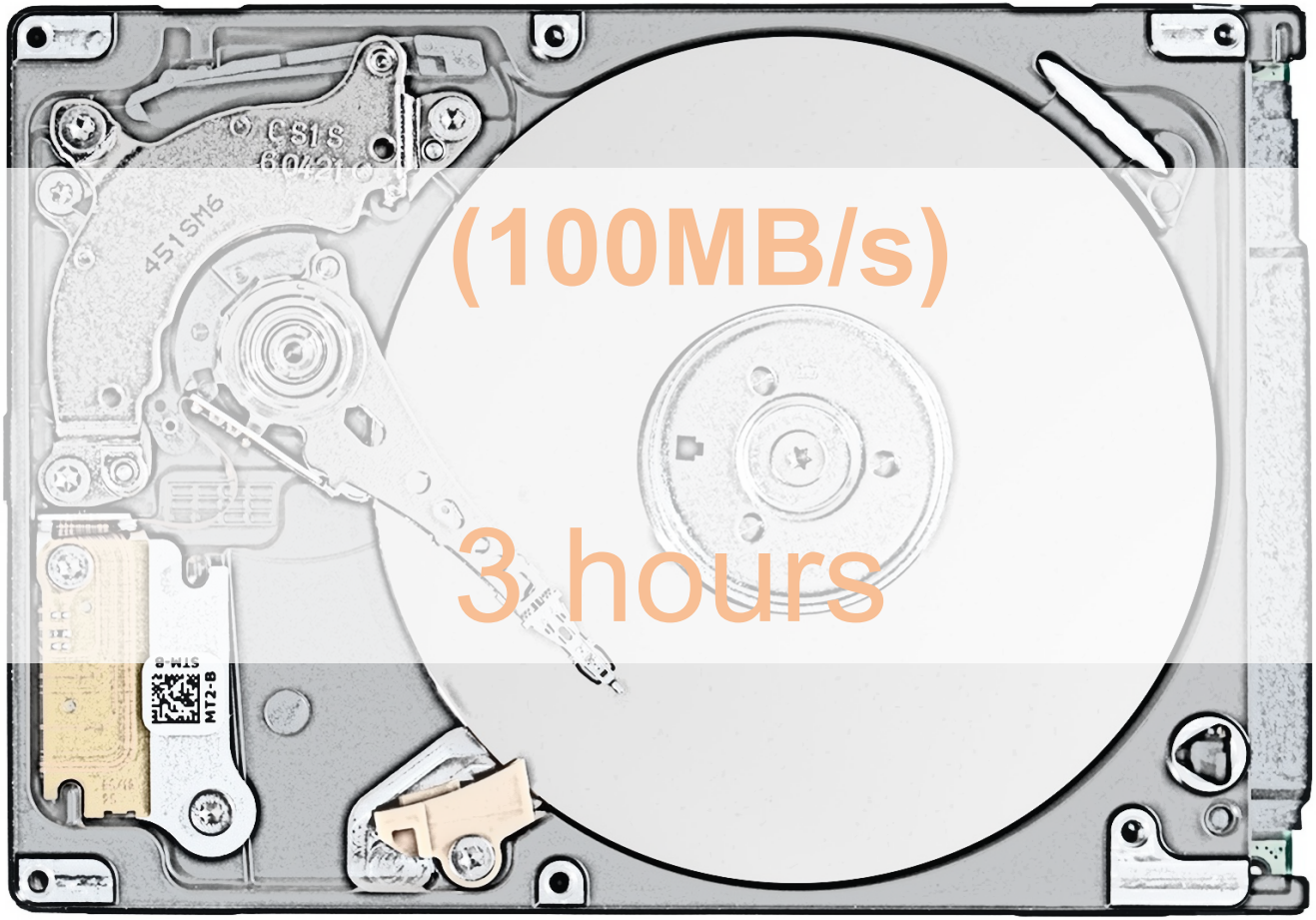
1. Computing Waves
2. Why now?
3. Future Hardware?
4. Software Middleware?
5. What is Apache Spark?
6. Spark Basics
7. Spark Ecosystem
8. Spark & Marenostrum
9. What next?
10. To learn more …



Spark Meetup BARCELONA

THE BIG DATA PROBLEM

Scanning **1** Terabyte:



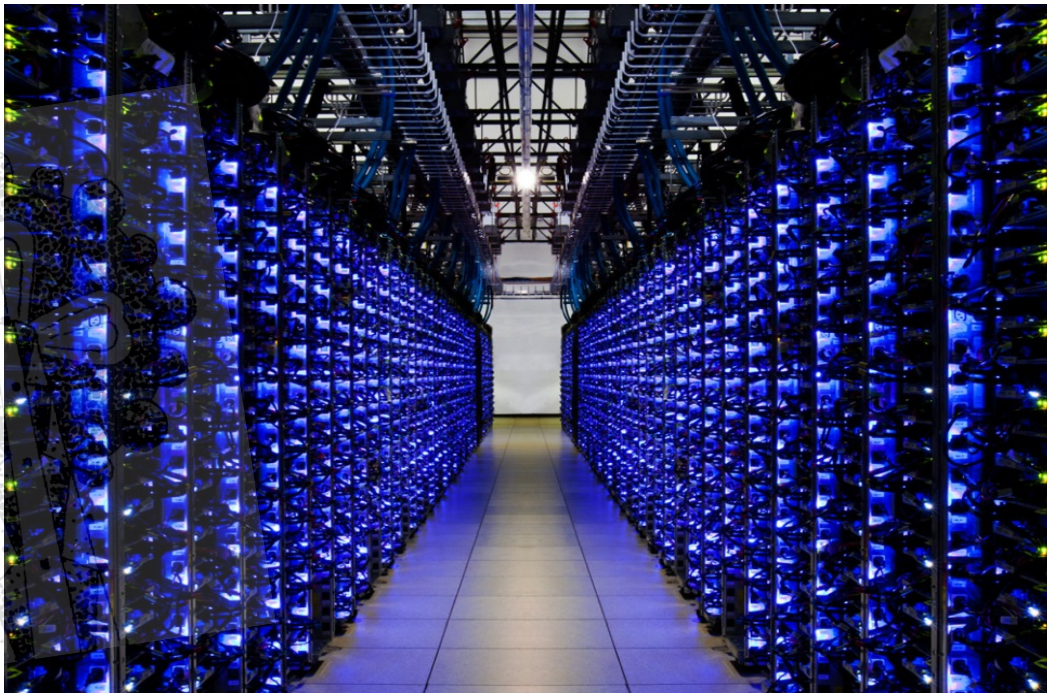
A single machine can no longer process all the data

THE BIG DATA PROBLEM

Solution: distribute data over large clusters

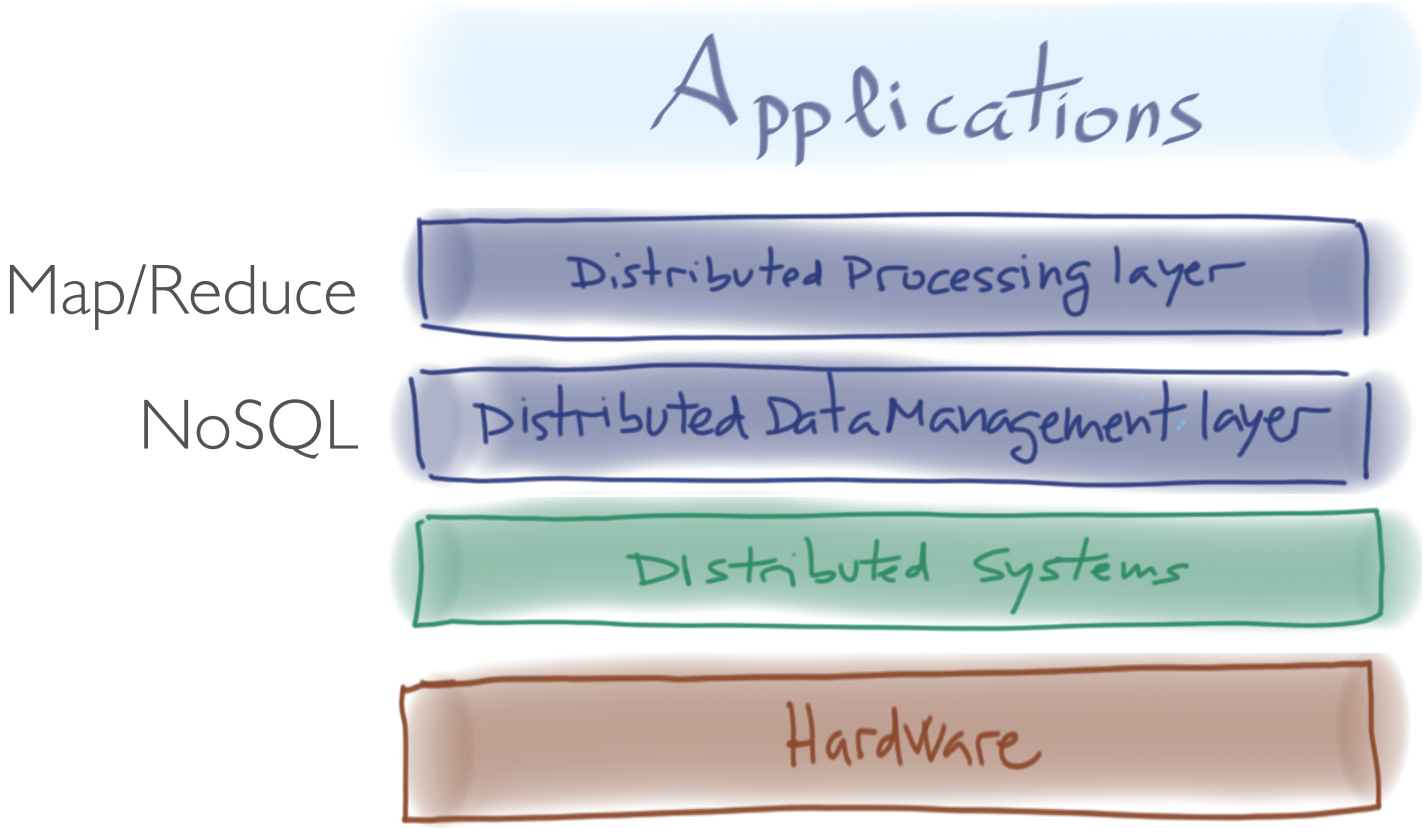
Problem:

- How do we split work across servers?
- How do we program this "thing"?



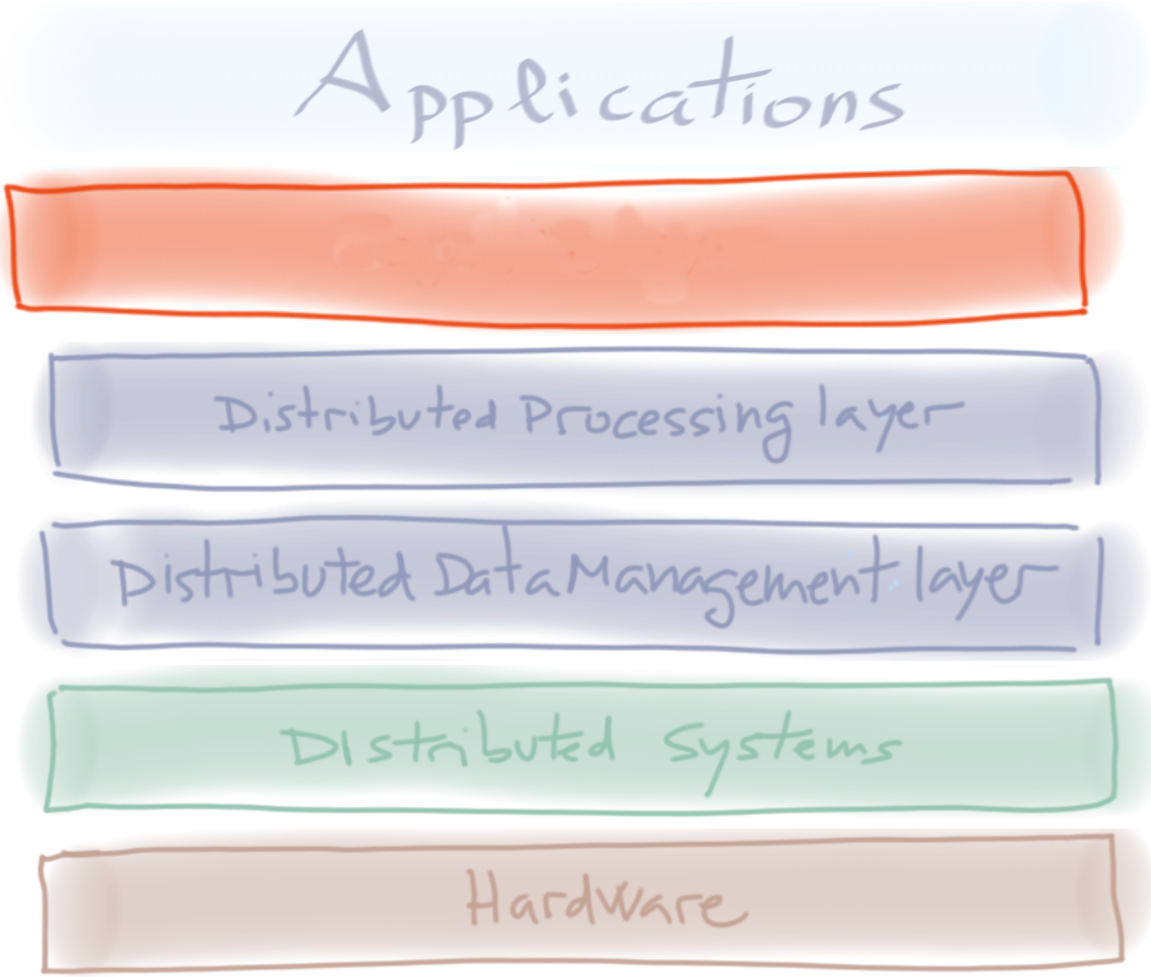
Source: http://www.google.com/about/datacenters/gallery/images/_2000/IDI_018.jpg

NEW COMPUTER MIDDLEWARE FOR BIG DATA



AND … DATA ANALYTICS?

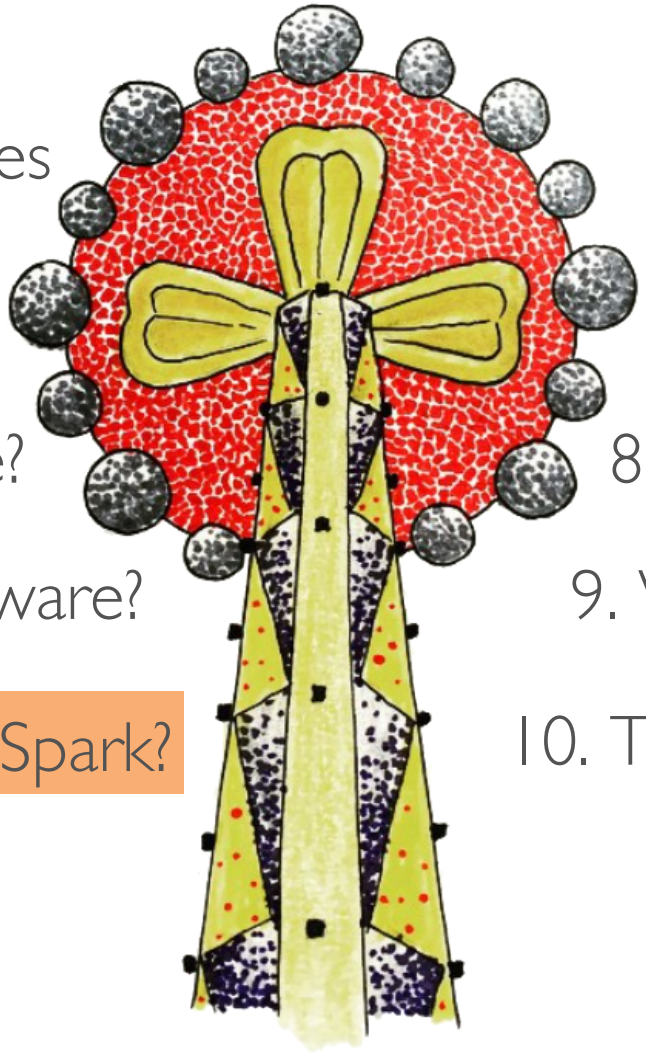
- New abstraction layer required



Distributed Analytic layer?

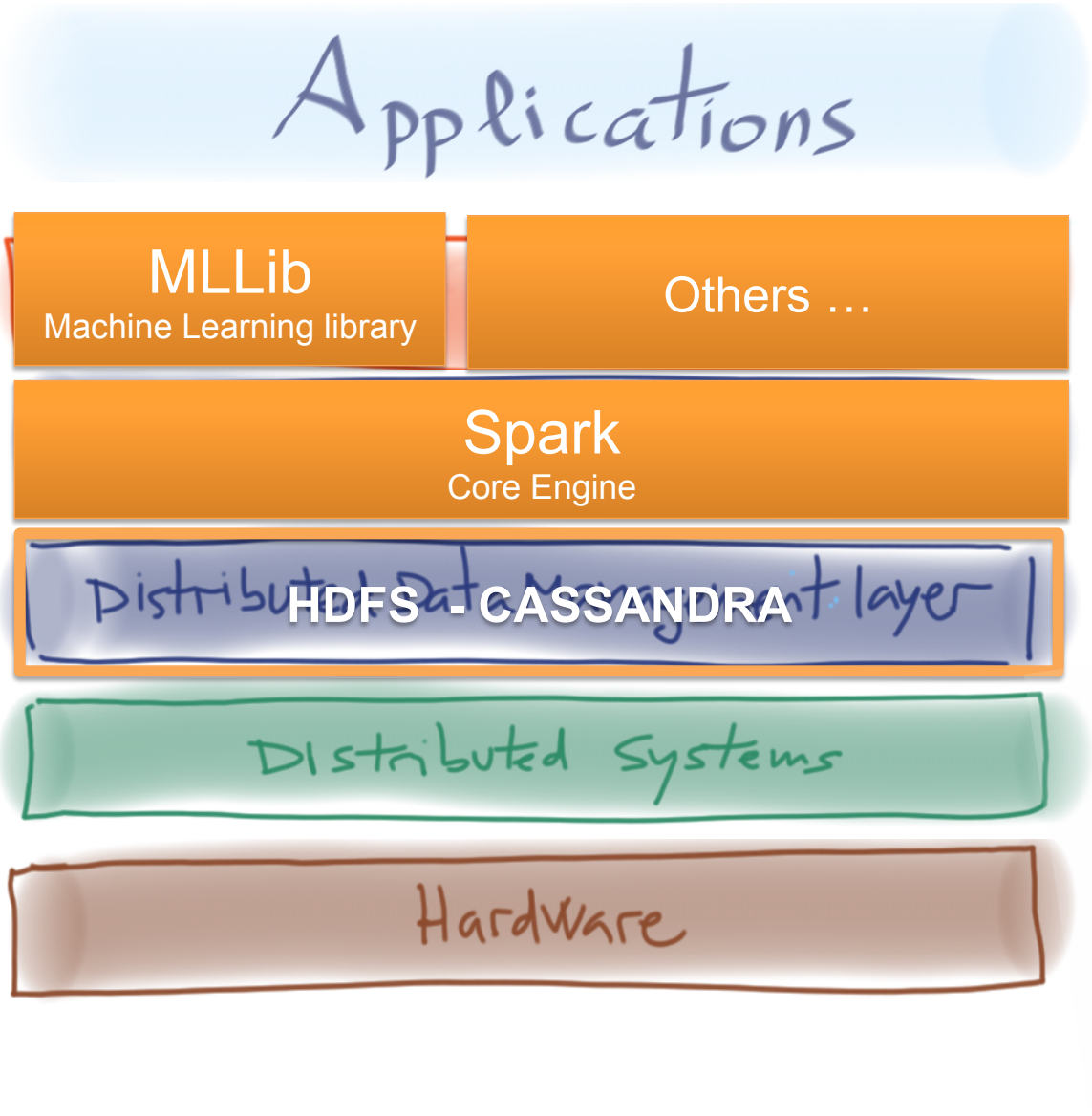
TALK OUTLINE

1. Computing Waves
2. Why now?
3. Future Hardware?
4. Software Middleware?
5. What is Apache Spark?
6. Spark Basics
7. Spark Ecosystem
8. Spark & Marenostrum
9. What next?
10. To learn more …



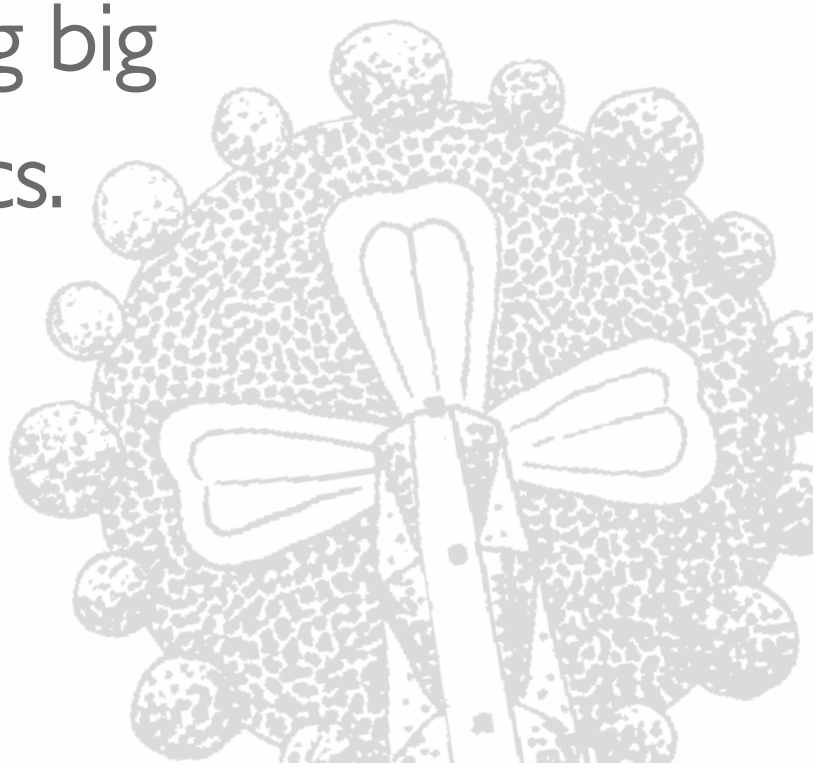
Spark meetup BARCELONA

WHAT IS APACHE SPARK?



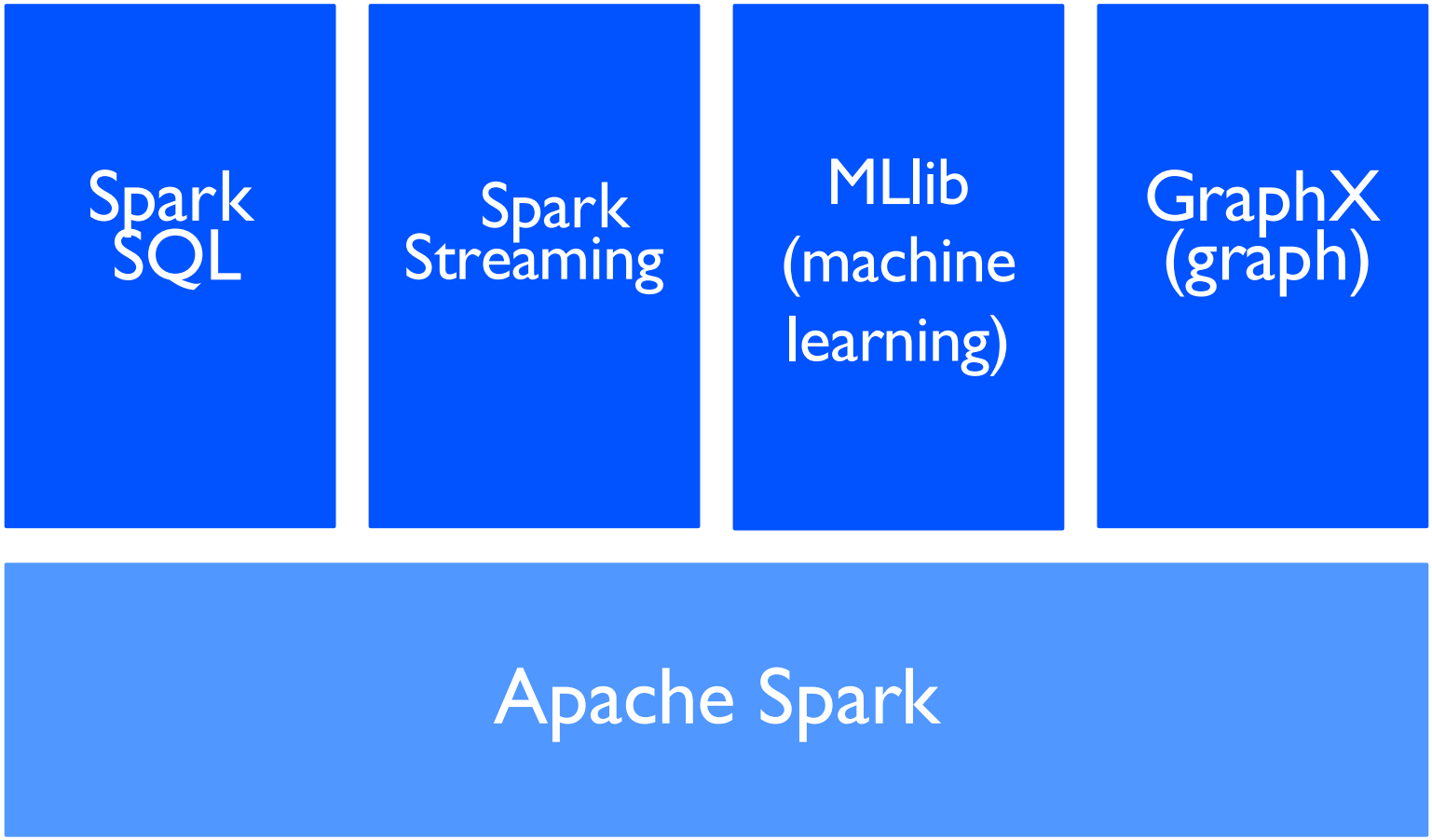
WHAT IS APACHE SPARK?

In brief, Spark is a unified platform for cluster computing, enabling big data management and analytics.



- Keep more data in-memory

UNIFIED PLATFORM: SPARK ECOSYSTEM



A BRIEF HISTORY

- Developed in 2009 at UC Berkeley AMPLab, then open sourced in 2010
- In 2013 Databricks was founded by Spark creators
- It is an Apache Project and its current version is 1.4 (last week)
- Spark has become one of the largest communities in big data, with over 500 contributors in 200 organizations

SOME KEY POINT ABOUT SPARK

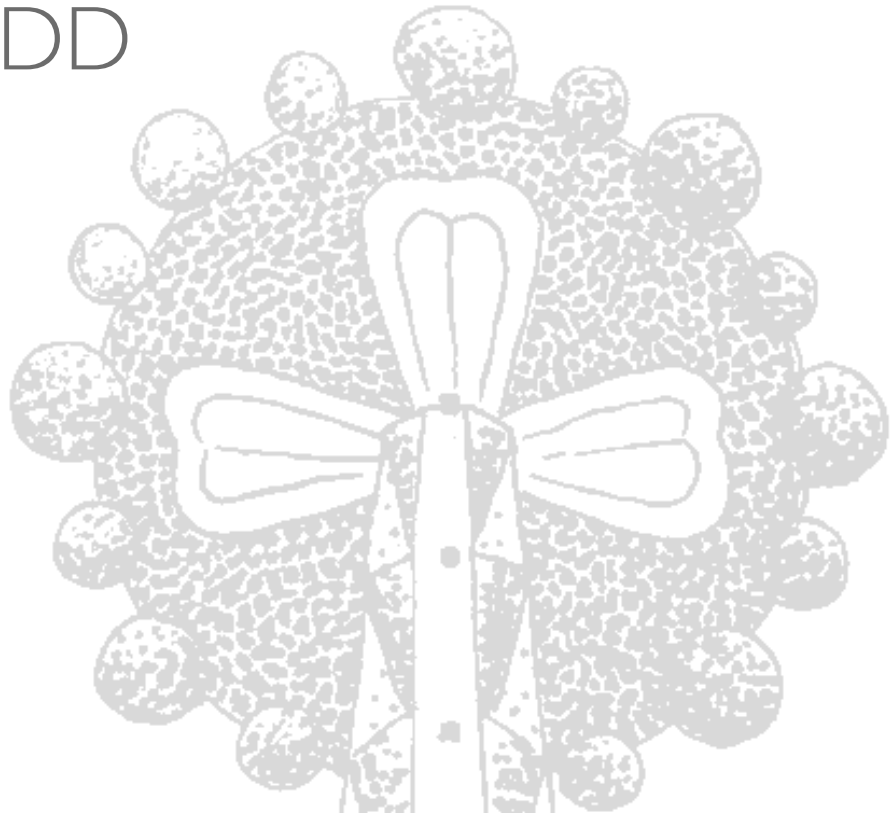
- Handles batch, interactive and real-time within a single framework
- Works with any Hadoop-supported storage system

SOME KEY POINT ABOUT SPARK

- Native integration with Java, Python and Scala
    - Interactive Shell in Scala and Python
- Programming at a higher level of abstraction
    - Map/Reduce is just one supported constructs

SOME KEY POINT ABOUT SPARK

- Two important internal elements:
  - fast data sharing with RDD
  - general DAGs



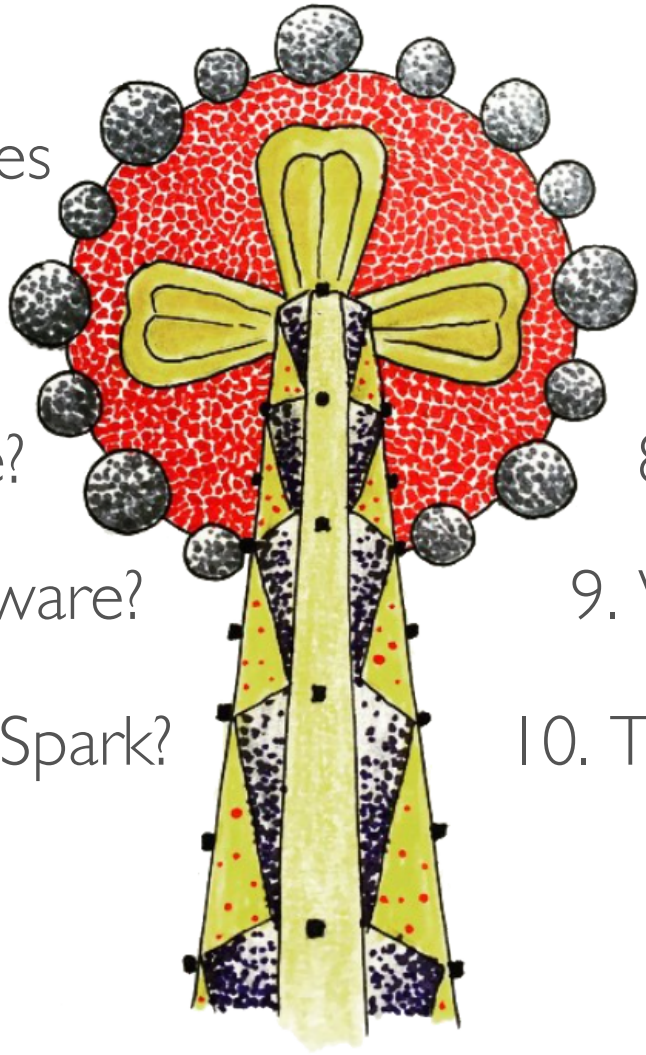
SPARK IS BUILD UPON AKKA

- Akka is an open-source toolkit and runtime simplifying the construction of concurrent and distributed applications on the JVM.

- Akka is written in Scala
  - Object oriented + functional programming
  - High-level language for the JVM
  - Interoperates with Java

TALK OUTLINE

1. Computing Waves
2. Why now?
3. Future Hardware?
4. Software Middleware?
5. What is Apache Spark?



6. Spark Basics
7. Spark Ecosystem
8. Spark & Marenostrum
9. What next?
10. To learn more …

Spark meetup BARCELONA

SPARK BASICS

Spark *Lightning-fast cluster computing*

Download | Libraries | Documentation | Examples | Community | FAQ

Download Spark

The latest release of Spark is Spark 1.2.0, released on December 18, 2014 (release notes) (git tag)

1. Chose a Spark release: 1.2.0 (Dec 18 2014)
2. Chose a package type: Pre-built for Hadoop 2.4 and later
3. Chose a download type: Direct Download
4. Download Spark: spark-1.2.0-bin-hadoop2.4.tgz
5. Verify this release using the 1.2.0 signatures and checksums.

Note: Scala 2.11 users should download the Spark source package and build with Scala 2.11 support.

Link with Spark

Spark artifacts are hosted in Maven Central. You can add a Maven dependency with the following coordinates:

```
groupId: org.apache.spark
artifactId: spark-core_2.10
version: 1.2.0
```

Latest News

Spark Summit East agenda posted, CFP open for West (Jan 21, 2015)

Spark 1.2.0 released (Dec 18, 2014)

Spark 1.1.1 released (Nov 26, 2014)

Registration open for Spark Summit East 2015 (Nov 26, 2014)

Archive

Download Spark

Built-in Libraries:

- Spark SQL
- Spark Streaming
- MLlib (machine learning)
- GraphX (graph)

http://spark.apache.org

SPARK'S SHELLS

Python: ./bin/pyspark

```
lines = sc.textFile("README.md") # Create an RDD called lines
lines.count() # Count the number of items in this RDD
```

Scala: ./bin/spark-shell

```
val lines = sc.textFile("README.md") // Create an RDD called lines
lines.count() // Count the number of items in this RDD
```

DEMO

SPARK CORE

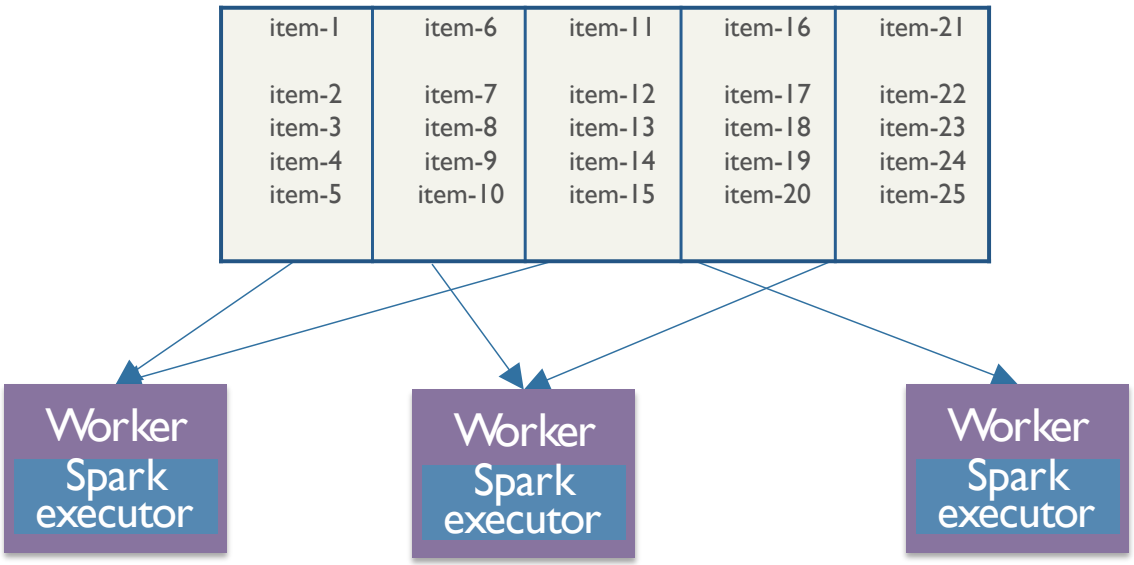
- Basic functionality /components for:
  - Task Scheduling
  - Memory Management
  - Fault Recovery
  - Interacting with Storage Systems
  - …
- Resilient Distributed Datasets (RDD) are the primary abstraction in Spark

  represents a collection of items distributed across many compute nodes that can be manipulated in parallel.

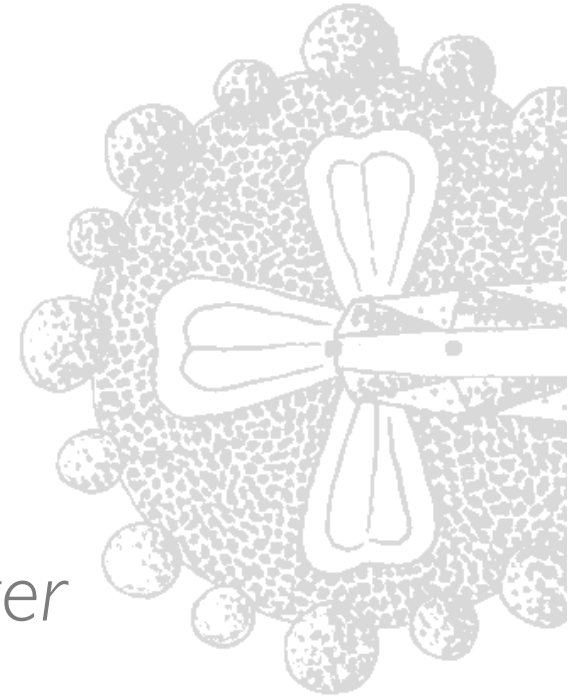
RDDS

- RDDs are distributed across workers (servers)

  Programmer specifies number of partitions for an RDD



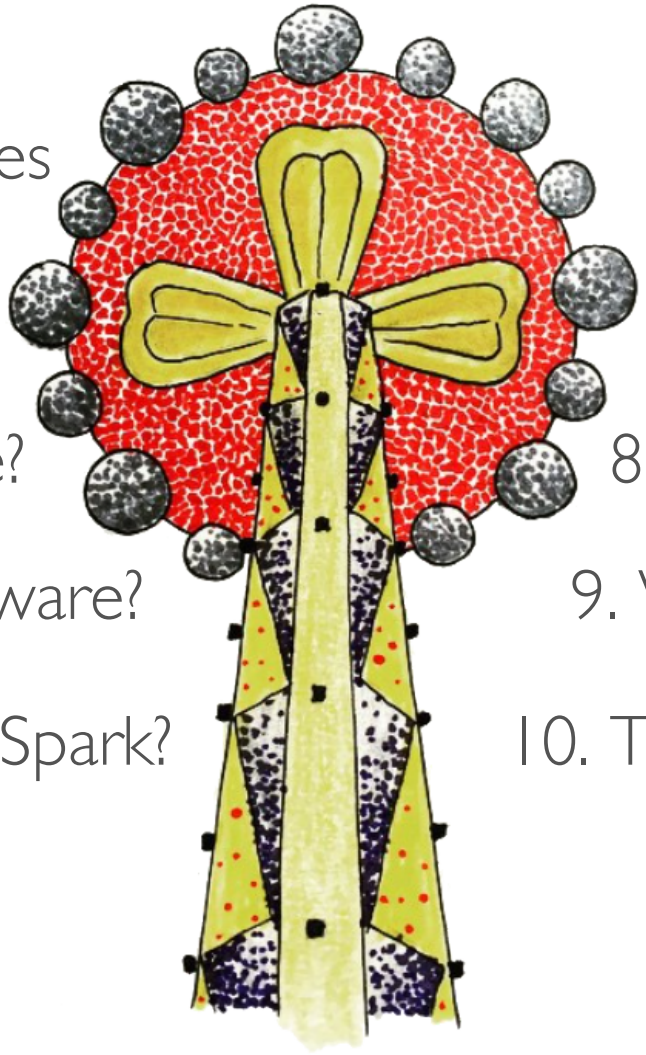
RDDS



- Two types of operations:
    - Transformations: *e.g. Map, Filter*
    - Actions: *e.g. Count, Reduce*
- Transformations are lazy (not computed immediately)
    - Transformated RDD is executed when action runs on it
- Persist (cache) RDDS in memory or disk

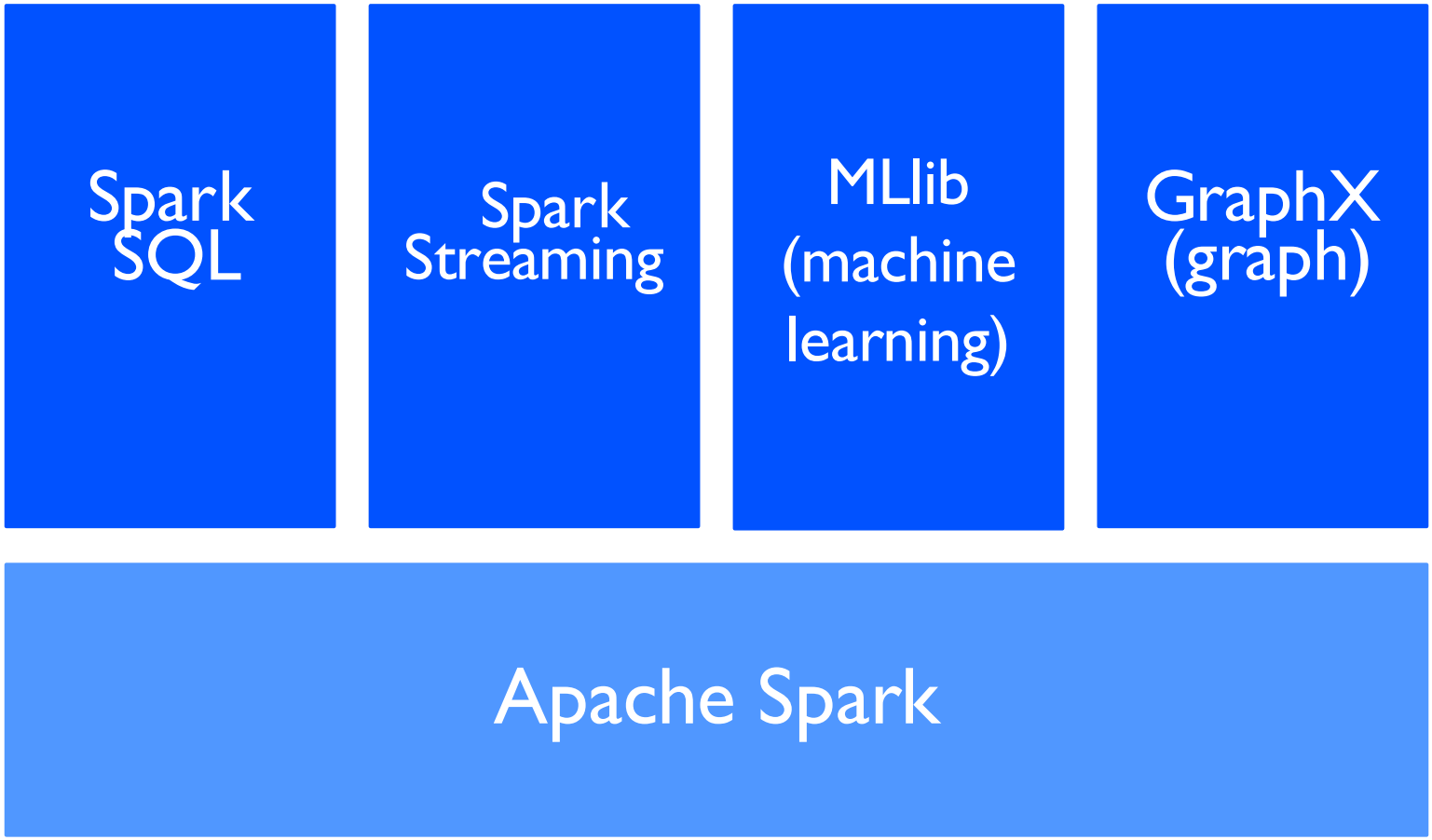
TALK OUTLINE

1. Computing Waves
2. Why now?
3. Future Hardware?
4. Software Middleware?
5. What is Apache Spark?
6. Spark Basics
7. Spark Ecosystem
8. Spark & Marenostrum
9. What next?
10. To learn more …



Spark Meetup BARCELONA

SPARK ECOSYSTEM



SPARK SQL

Interface for working with structured and semistructured data (data that has a schema, a known set of fields):

Load data from a variety of structured sources: JSON, Hive, and Parquet.

Rich integration between SQL and Python/ Scala/Java

SPARK SQL

- Both loading data and executing queries return SchemaRDDs (now called Data Frame's):

  - SchemaRDDs are similar to tables in a traditional database.

  - Under the hood, a SchemaRDD is an RDD composed of Row objects with additional schema information of the types in each column.

EXAMPLE SPARK SQL

- Scala SQL imports

```
import org.apache.spark.sql.hive.HiveContext
```

- Constructing a SQL context (in Scala)

```
val sc = new SparkContext(...)

val hiveCtx = new HiveContext(sc)
```

EXAMPLE SPARK SQL

- Example Loading and quering tweets in Scala

```
val input = hiveCtx.jsonFile(inputFile)

// Register the input schema RDD

input.registerTempTable("tweets")

// Select tweets based on the retweetCount

val topTweets = hiveCtx.sql("SELECT text, retweetCount
FROM tweets ORDER BY retweetCount LIMIT 10")
```

SPARK STREAMING

- Data can be ingested from many sources and results can be pushed out to many formats

- Spark's built-in machine learning algorithms and graph processing algorithms can be applied to data streams

SPARK STREAMING

- Spark Streaming extends the core API to allow high-throughput, fault-tolerant stream processing of live data streams

- Streaming computation treated as series of batch computations

  In each time interval

  - Data collected into a batch
  - At the end of interval RDD is created
  - Processed using Spark jobs

SPARK STREAMING EXAMPLE

- Obtain popular hashtags/topics

sbt 'run <master> " + "consumerKey consumerSecret accessToken accessTokenSecret" + " [filter1] [filter2] ... [filter n]"+ """)

DEMO

```
import org.apache.spark.streaming.{Seconds, StreamingContext}
import StreamingContext._
import org.apache.spark.SparkContext._
import org.apache.spark.streaming.twitter._
import org.apache.log4j.Logger
import org.apache.log4j.Level

object TwitterPopularTags {

  def main(args: Array[String]) {

    val (master, filters) = (args(0), args.slice(5, args.length))

    // Twitter Authentication credentials
    System.setProperty("twitter4j.oauth.consumerKey", args(1))
    System.setProperty("twitter4j.oauth.consumerSecret", args(2))
```

MLLIB

MLlib is Spark's library of machine learning functions:

- It provides us with machine learning algorithms to run on RDDs.
- It only contains parallel versions of algorithms.

MLLIB

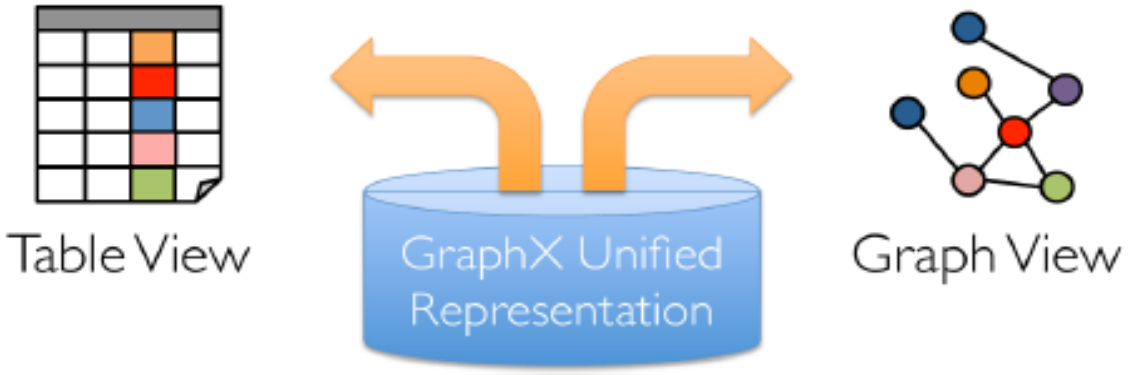
- Example of training ang test

```
// http://spark.apache.org/docs/latest/mllib-guide.html!
val train_data = // RDD of Vector!
val model = KMeans.train(train_data, k=10)!
// evaluate the model!
val test_data = // RDD of Vector!
test_data.map(t =>
model.predict(t)).collect().foreach(println)
```

GRAPHX

- Extends Spark with recent advances in graph systems to enable users to easily and interactively build, transform, and reason about graph structured data at scale.

- Unifies Data-Parallel & Graph-Parallel systems.



SUMMARY

News room > News releases >

IBM Announces Major Commitment to Advance Apache®Spark™, Calling it Potentially the Most Significant Open Source Project of the Next Decade

IBM Joins Spark Community, Plans to Educate More Than 1 Million Data Scientists

Select a topic or year

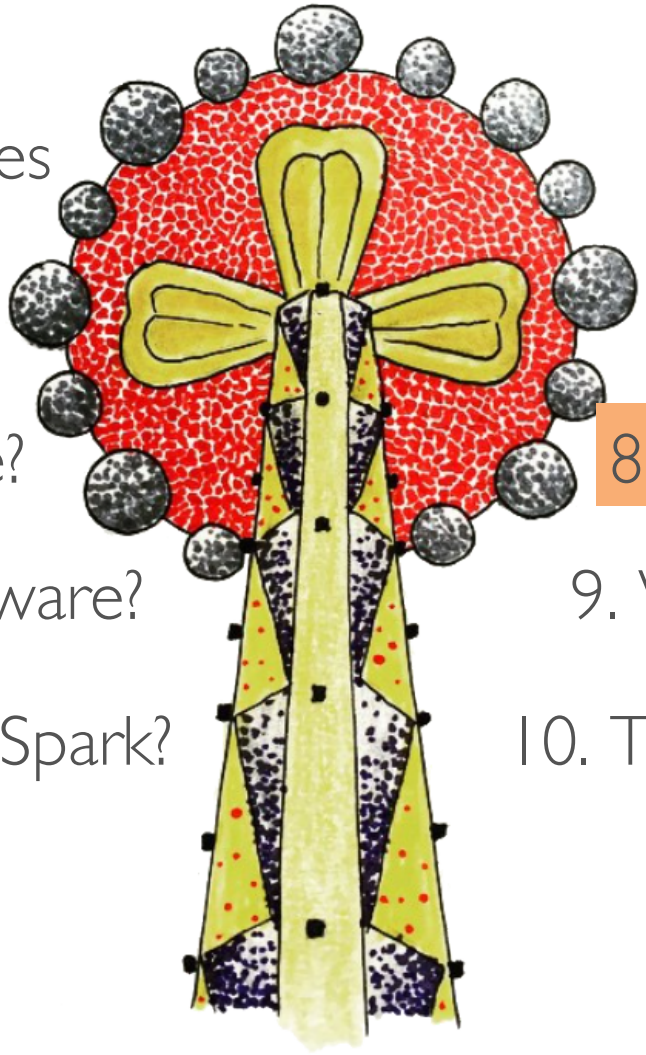
↓ News release ↓ Contact(s) information

↓ Related XML feeds ↓ Related resources

ARMONK, NY - 15 Jun 2015: IBM (NYSE:IBM) today announced a major commitment to Apache®Spark™, potentially the most important new open source project in a decade that is being defined by data. At the core of this commitment, IBM plans to embed Spark into its industry-leading Analytics and Commerce platforms, and to offer Spark as a service on IBM Cloud. IBM will also put more than 3,500 IBM researchers and developers to work on Spark-related projects at more than a dozen labs worldwide; donate its breakthrough IBM SystemML machine learning technology to the Spark open source ecosystem; and educate more than one million data scientists and data engineers on Spark.

TALK OUTLINE

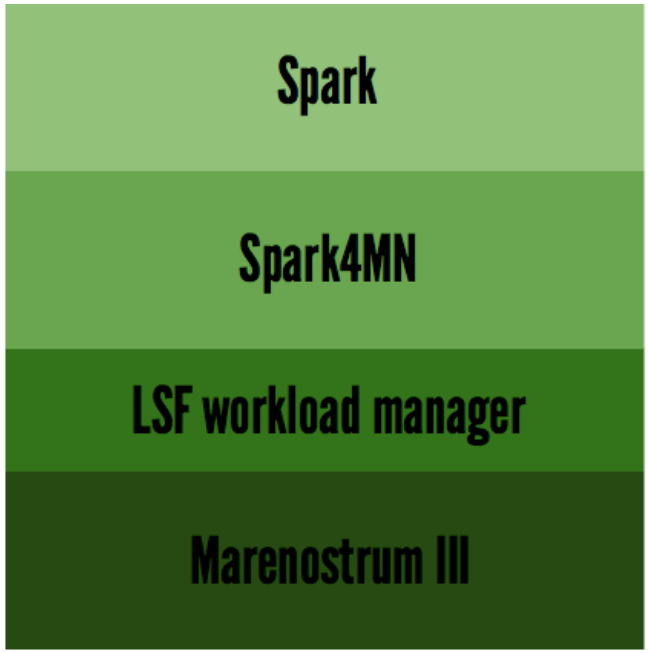
1. Computing Waves
2. Why now?
3. Future Hardware?
4. Software Middleware?
5. What is Apache Spark?
6. Spark Basics
7. Spark Ecosystem
8. Spark & Marenostrum
9. What next?
10. To learn more …



Spark meetup BARCELONA

NEW MIDDLEWARE @ MNIII

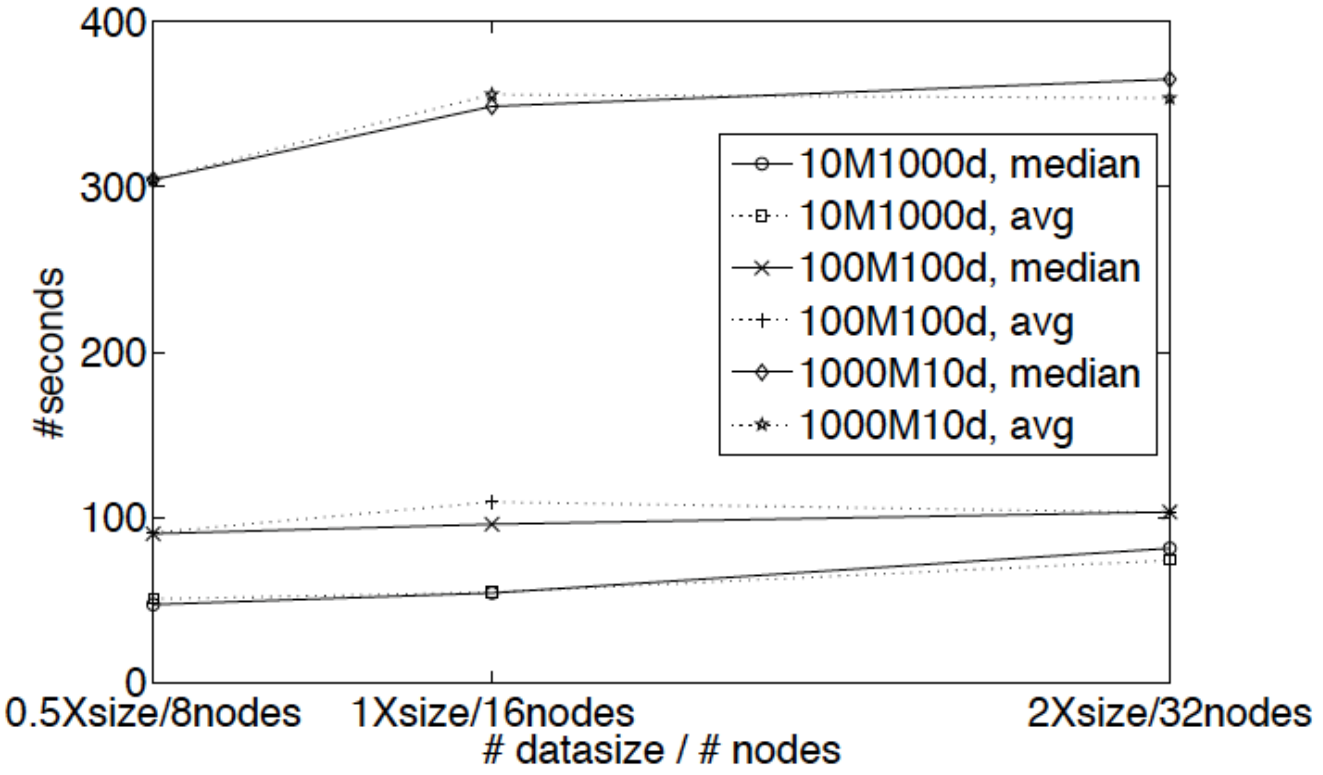
- SPARK4MN: framework to enable Spark worklods over IBM LSF Platform workload manager on MNIII



NEW MIDDLEWARE @ MNIII

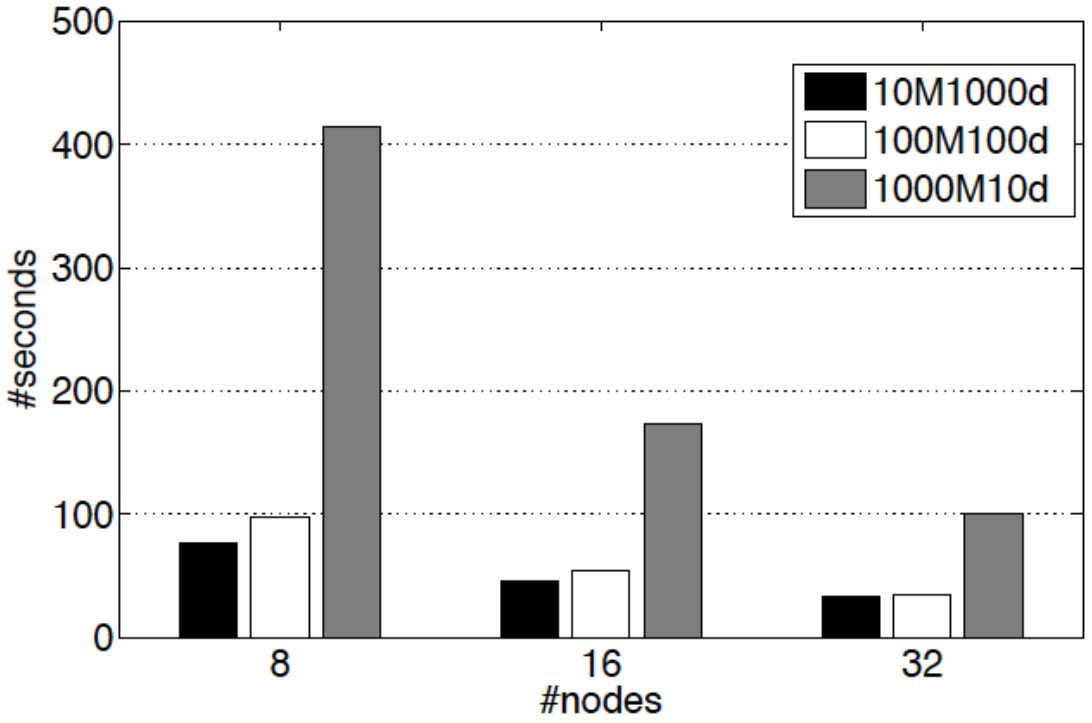
- Allow us to configure and test parameters as
    - cluster geometry (number of nodes, number of workers per node, etc)
    - memory/node ,
    - storage (local, HDFS or GPFS),
    - network interface,
    - CPU affinity,
    - etc

EXAMPLE: K-MEANS SCALE-UP



(*) ideal: horizontal plots

EXAMPLE: K-MEANS SPEED-UP



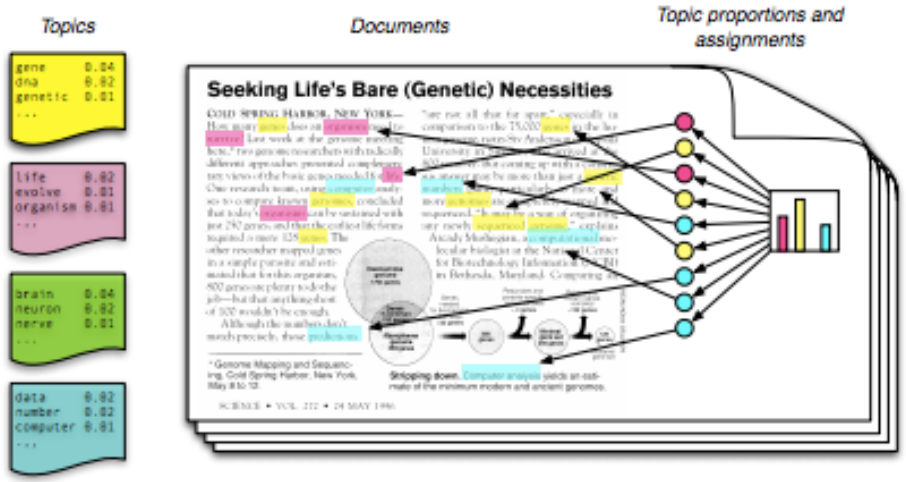
(*)

Problem size constant = 100GBs (10M1000D = 10M vectors of 1000 dimensions)

More dimensions → smaller speed-up due the increased shuffling (same number of centroids to shuffle but bigger)

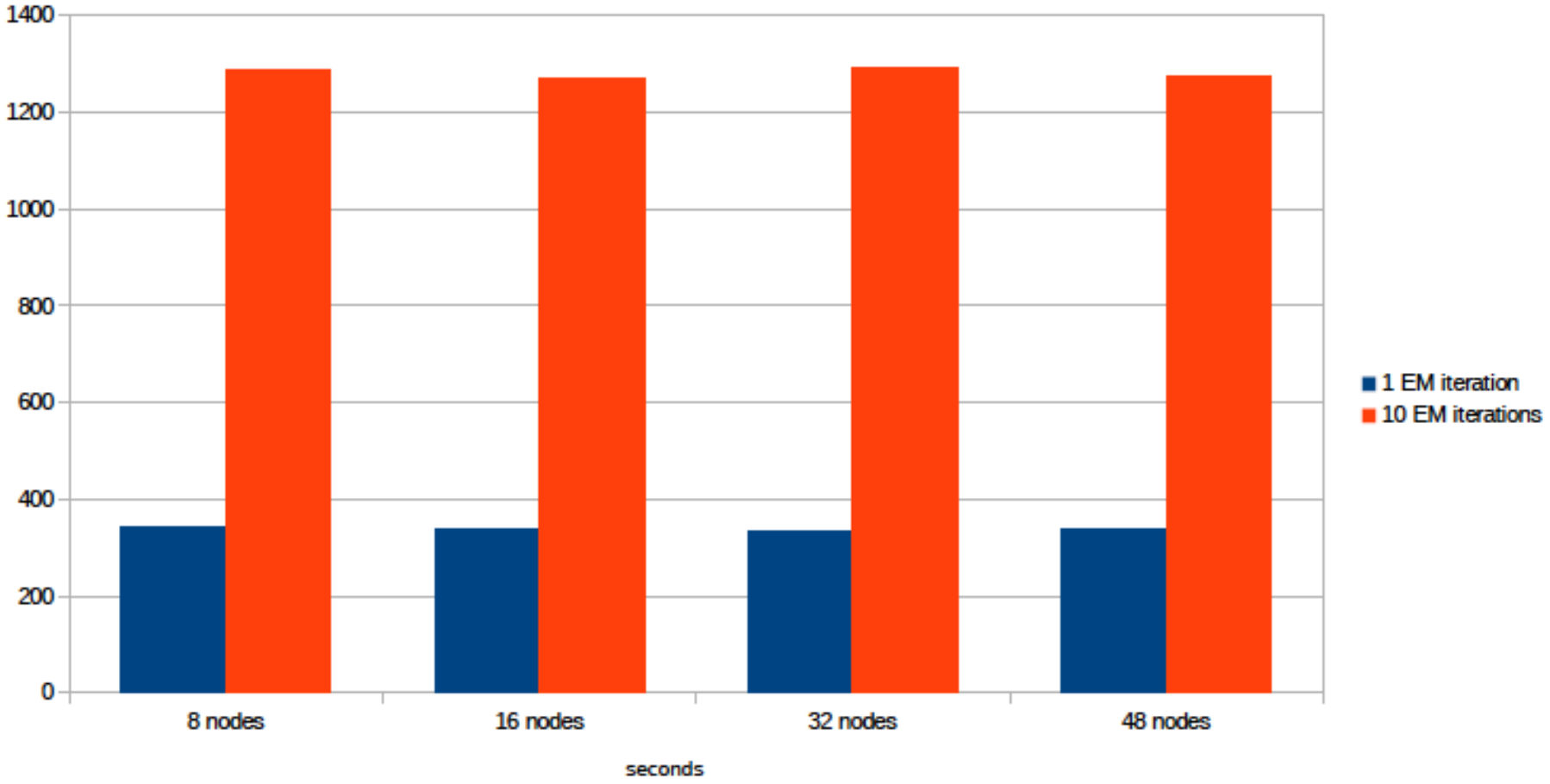
ANOTHER EXAMPLE: LDA

- Latent Dirichlet Allocation (LDA)
- LDA is known to be hard to parallelize efficiently



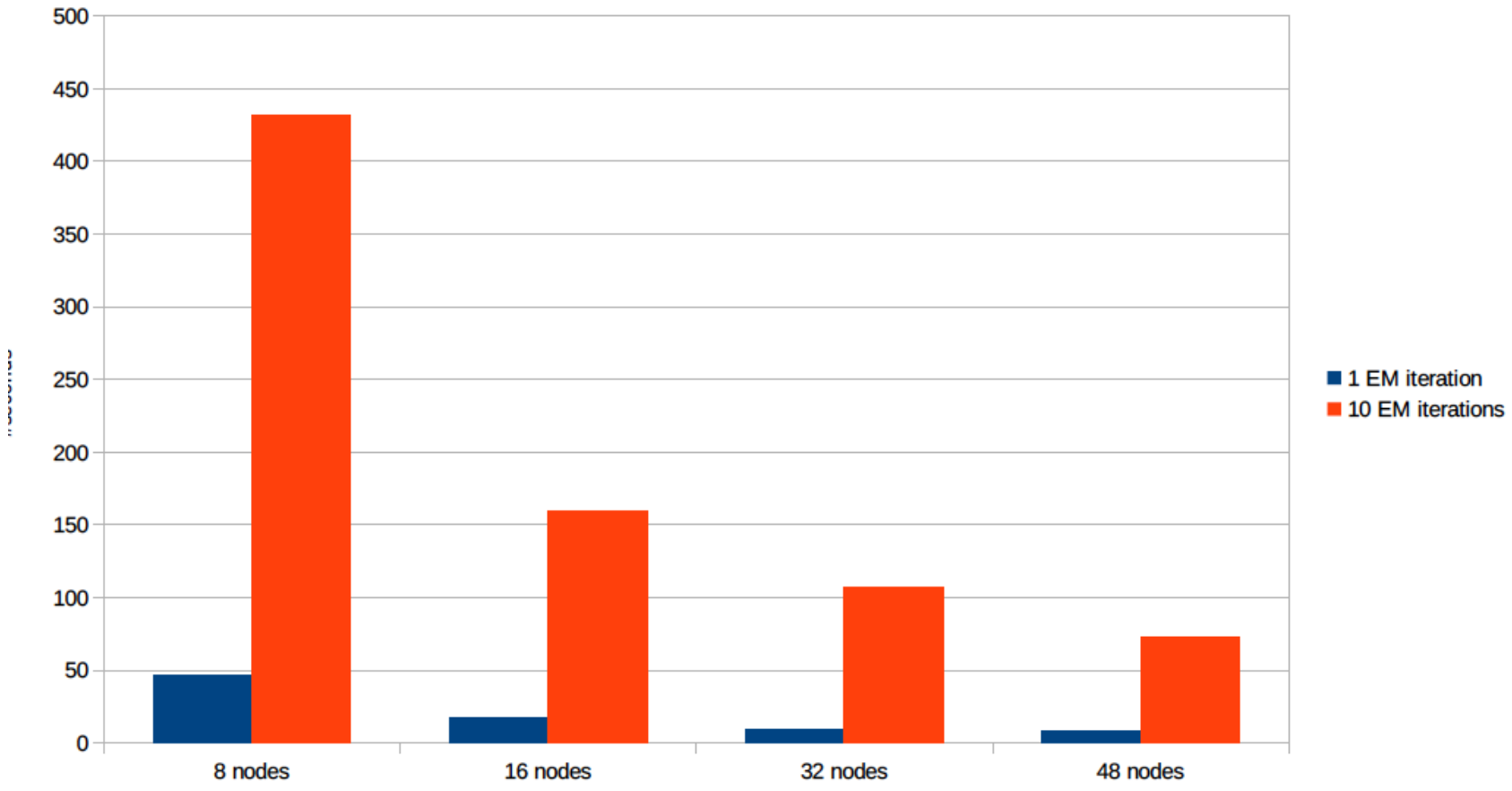
INITIAL RESULTS (ON MARENOSTRUM III)

- Not automatic scalability



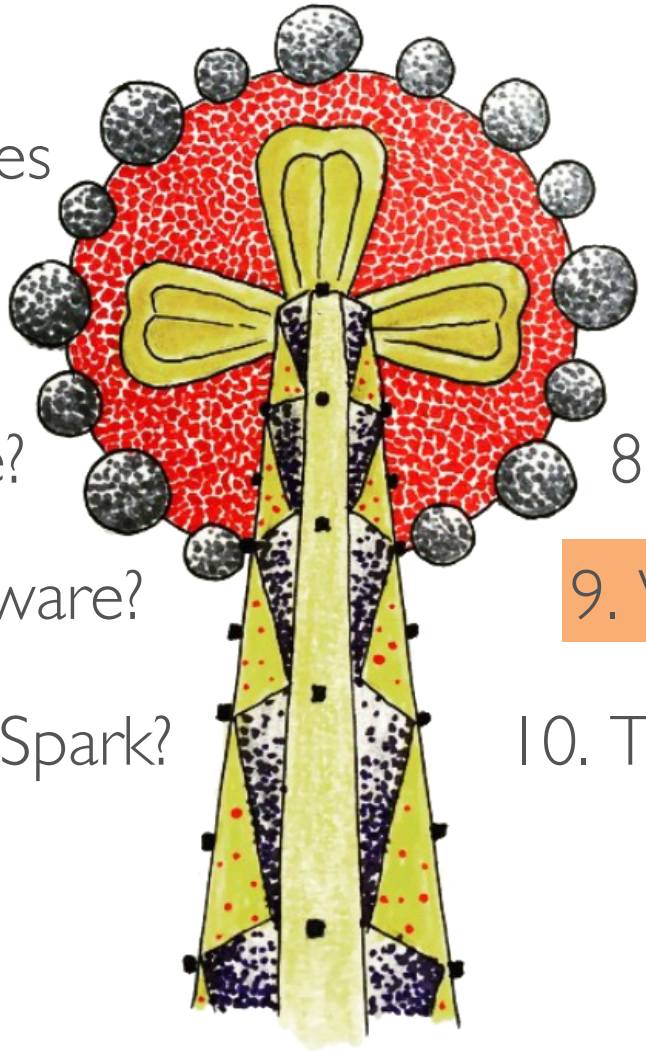
RESULTS
AFTER TUNNING PARAMETERS

- Nice scalability through proper paralelization



TALK OUTLINE

1. Computing Waves
2. Why now?
3. Future Hardware?
4. Software Middleware?
5. What is Apache Spark?
6. Spark Basics
7. Spark Ecosystem
8. Spark & Marenostrum
9. What next?
10. To learn more …



Spark Meetup BARCELONA

WHAT NEXT?

Jordi Torres y Mateo Valero

Ordenadores más sabios

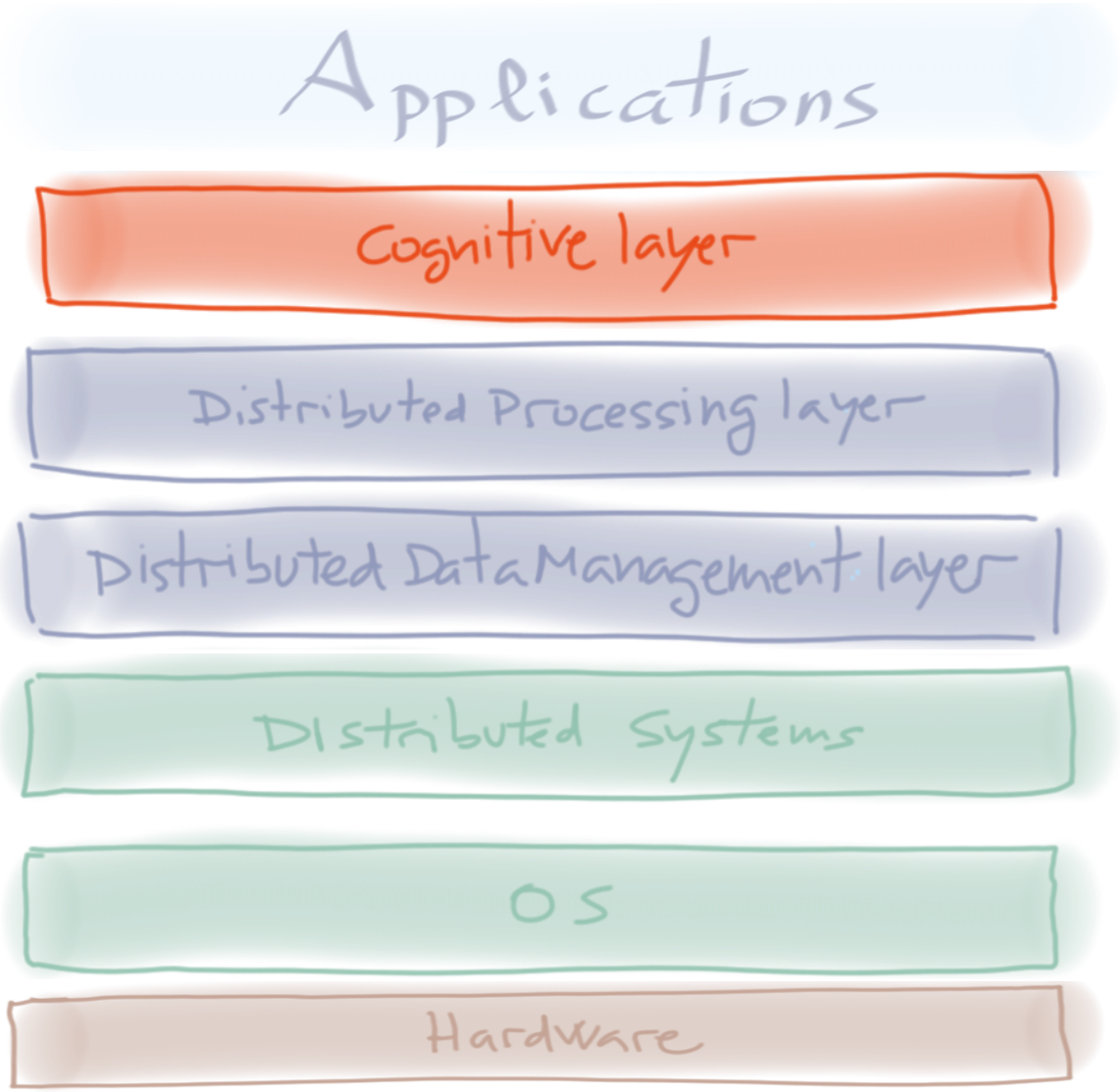
Ya hace tiempo que hemos pasado de una era tecnológica basada en el procesamiento de números a una en que los textos y contenidos multimedia también son computables y, al mismo tiempo, accesibles digitalmente desde cualquier lugar y en cualquier momento. Los dispositivos móviles interaccionan con los usuarios y lo hacen entre sí. De ahora en adelante, la computación también hará el contexto computable, incorporará capacidades predictivas y de aprendizaje, proporcionando la funcionalidad correcta y el contenido en el instante adecuado, para la persona correcta, prediciendo lo que esta necesitará. No será extraño que le pidamos al asistente de voz de nuestro móvil: "Necesito un vuelo a Nueva York vía Londres", y que nos muestre las mejores opciones y nos reserve el billete según las preferencias en los asientos o el número de viajero frecuente. Ya se trabaja en una nueva familia de supercomputadores capaces de tratar situaciones complejas caracterizadas por la ambigüedad. Máquinas cada vez más sabias, con algoritmos de aprendizaje automático para extraer conocimiento del gran volumen de datos disponible y capaces de predecir y autoaprender. Como el supercomputador Watson, que ganó a los dos mejores concursantes de la historia del popular concurso de televisión norteamericana *Jeopardy*. La computación cognitiva se empieza a aplicar sobre un gran número de datos sanitarios para identificar pacientes con más riesgo de enfermedad o readmisión. Así, se mejora la atención preventiva y se hace un uso más eficiente de los recursos sanitarios.

Esta tecnología mejorará nuestras vidas y permitirá controlar lo que estamos a punto de hacer, con algoritmos que pueden predecir, como los que ya se aplican para conocer las preferencias de los usuarios en las compras por internet. La importancia de la privacidad pronto pasará a segundo plano cuando el reto sea salvaguardar la capacidad individual para decidir. Este nuevo estadio representa al tiempo un desafío a los trabajadores de cuello blanco de la sociedad del conocimiento, de la misma manera que la automatización de las fábricas en el siglo XX fue una revolución para los trabajadores de mono azul en las cadenas de montaje. Nos hace falta un debate social para prepararnos para la llegada de esta nueva era tecnológica que transformará profundamente la manera en que vivimos, trabajamos y pensamos.●

J. TORRES Y M. VALERO, *investigador de la UPC y del Barcelona Supercomputer Center, e investigador de la UPC y director del BSC-CNS, respectivamente*

NEW SOFTWARE STACK

Systems will have a new cognitive abstraction layer in the software stack



offering learning tools, but at the same time, abstracting lower layers to simplify the big data software tack.

COGNITIVE LAYER INCLUDES

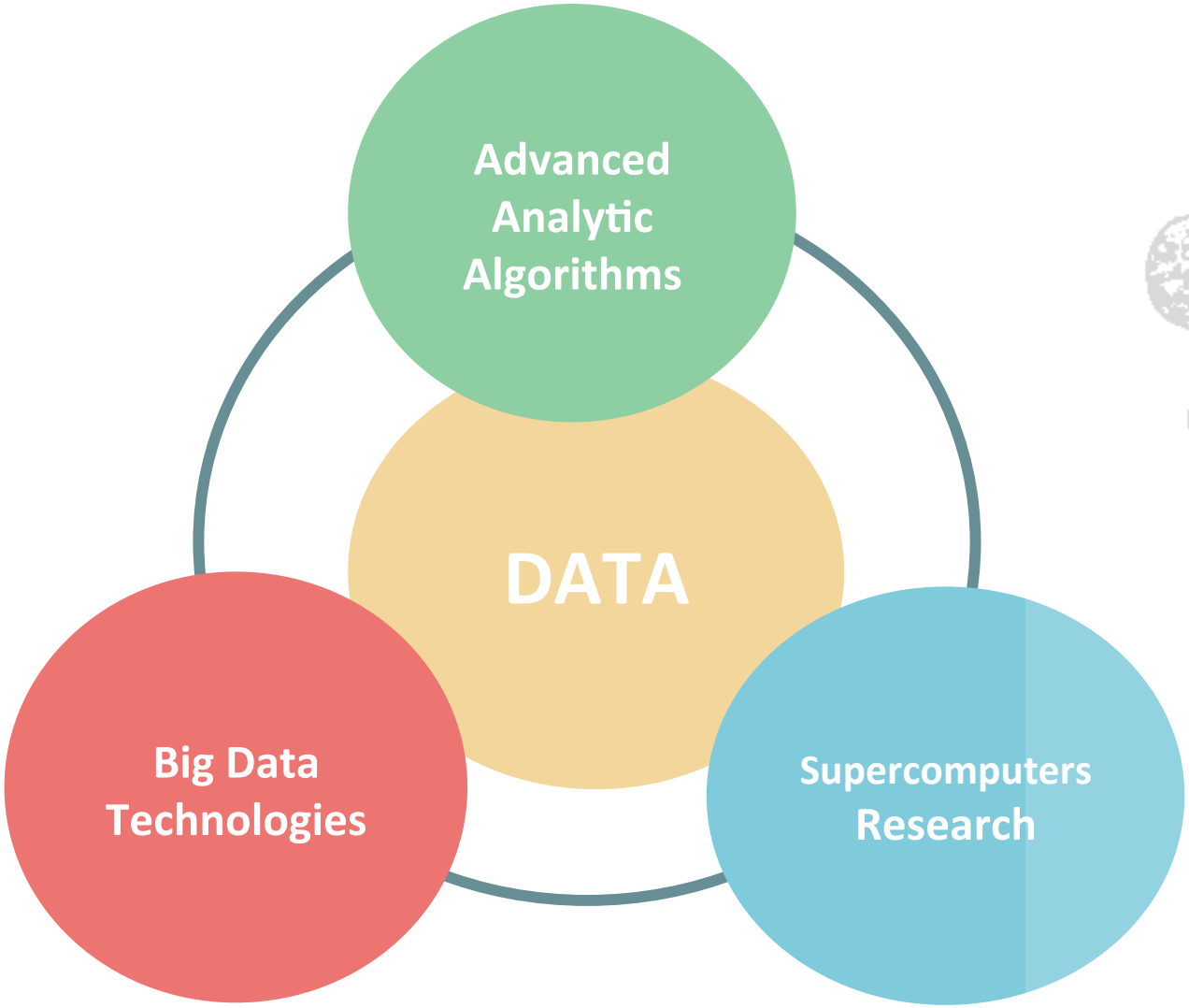
- Machine Learning algorithms
- Statistics
- Technologies enabled by Artificial Intelligence as

Computer Vision

Speech Recognition

Natural Language Processing

SUMMARY: FOUNDATIONAL BUILDING BLOCKS



ALREADY IN ENTERTAINMENT…

In 2011 IBM Watson computer defeated two of Jeopardy' s greatest champions

ALREADY IN BUSINESS …

Booking a flights …

"I want a flight to
MWC Barcelona with
a return five days later
via London."

Just closed on $12.5 M in venture capital funding.

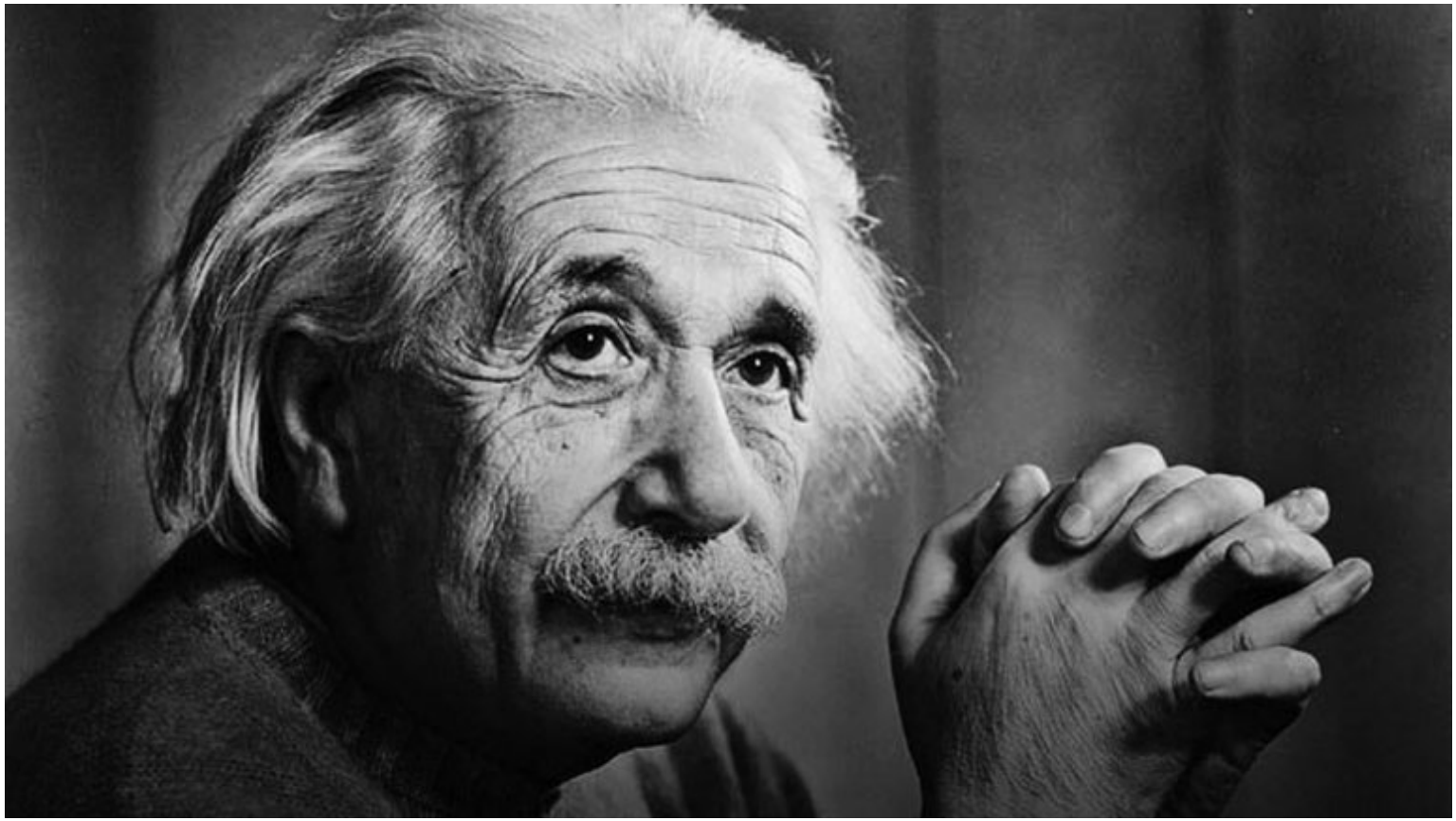
ALREADY IN RESEARCH…

Baylor College of Medicine (Houston, Texas)

The solution analyzed 70.000 scientific articles on p53.

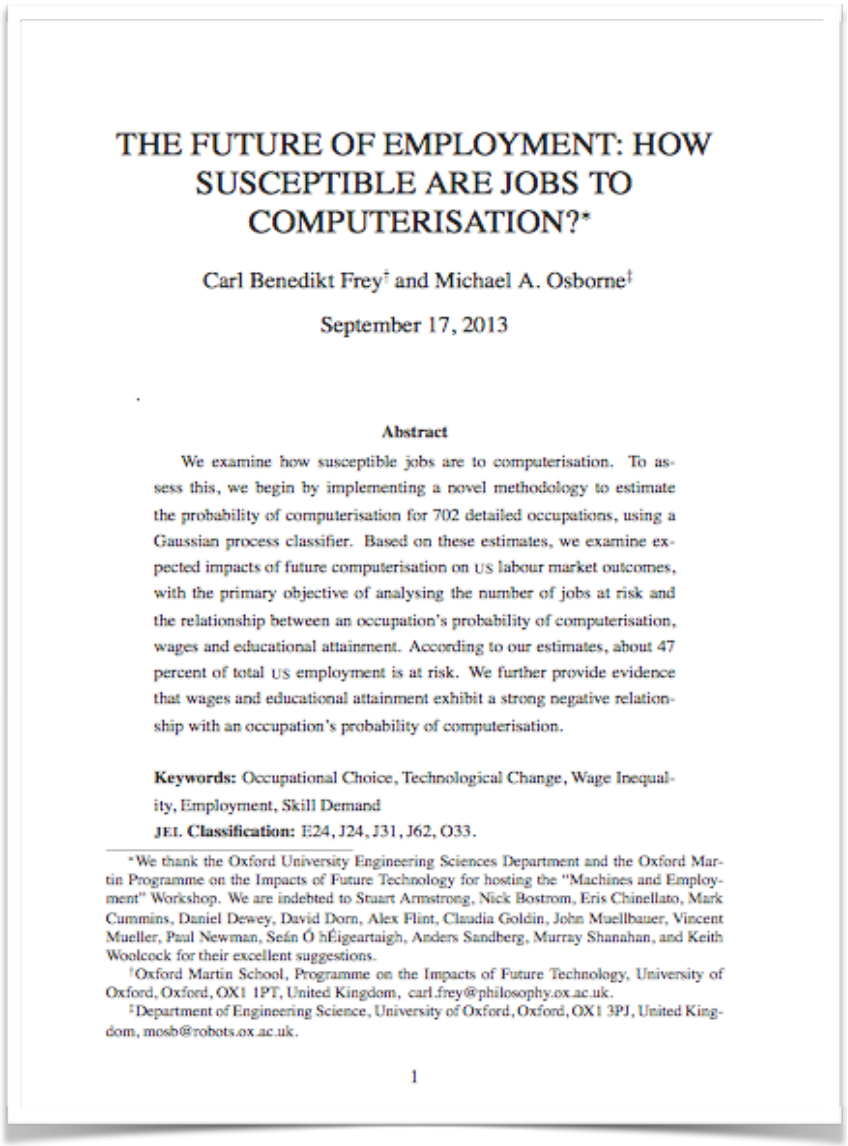
Identified proteins that modify p53, an important protein related to may cancers

IMPLICATIONS?



Sourcehttp://cdn01.am.infobae.com/adjuntos/163/imagenes/011/380/001380705.jpg

EMPLOYMENT?



THE FUTURE OF EMPLOYMENT: HOW SUSCEPTIBLE ARE JOBS TO COMPUTERISATION?*

Carl Benedikt Frey† and Michael A. Osborne‡

September 17, 2013

Abstract

We examine how susceptible jobs are to computerisation. To assess this, we begin by implementing a novel methodology to estimate the probability of computerisation for 702 detailed occupations, using a Gaussian process classifier. Based on these estimates, we examine expected impacts of future computerisation on US labour market outcomes, with the primary objective of analysing the number of jobs at risk and the relationship between an occupation's probability of computerisation, wages and educational attainment. According to our estimates, about 47 percent of total US employment is at risk. We further provide evidence that wages and educational attainment exhibit a strong negative relationship with an occupation's probability of computerisation.

Keywords: Occupational Choice, Technological Change, Wage Inequality, Employment, Skill Demand

JEL Classification: E24, J24, J31, J62, O33.

*We thank the Oxford University Engineering Sciences Department and the Oxford Martin Programme on the Impacts of Future Technology for hosting the "Machines and Employment" Workshop. We are indebted to Stuart Armstrong, Nick Bostrom, Eris Chinellato, Mark Cummins, Daniel Dewey, David Dorn, Alex Flint, Claudia Goldin, John Muellbauer, Vincent Mueller, Paul Newman, Seán Ó hÉigeartaigh, Anders Sandberg, Murray Shanahan, and Keith Woolcock for their excellent suggestions.

†Oxford Martin School, Programme on the Impacts of Future Technology, University of Oxford, Oxford, OX1 1PT, United Kingdom, carl.frey@philosophy.ox.ac.uk.

‡Department of Engineering Science, University of Oxford, Oxford, OX1 3PJ, United Kingdom, mosb@robots.ox.ac.uk.

Researchers at Oxford published a study estimating that 47 percent of total US employment is "at risk" due to the automation of cognitive tasks.

LV, 14/05/2015 (MAR GALTÉS)

EN PORTADA

ENTRE LA FANTASÍA Y EL FUTURO

Disruptivo y exponencial

La Singularity University, impulsada por Google y la NASA, promueve el poder transformador de las nuevas tecnologías ante los grandes retos de la humanidad

Mar Galtés

Hace unos 60 millones de años, un asteroide impactó en la Tierra y causó un cambio dramático: los dinosaurios no lograron sobrevivir y fueron los mamíferos, más pequeños, ágiles y adaptables, quienes se convirtieron en la especie predominante. "Ese asteroide son hoy las tecnologías exponenciales", dice en sus presentaciones Peter Diamandis, biólogo molecular, ingeniero aeroespacial, doctorado en Medicina, y cofundador de la Singularity University.

Silicon Valley es otro mundo, a medio camino entre la fantasía y el futuro, y la Singularity University es la máxima expresión de la voluntad de que las nuevas tecnologías tengan poder de transformación y un impacto global y transversal. Se trata de una organización impulsada por la NASA y Google, que organiza programas de formación y alianzas con empresas e instituciones, con la misión de aplicar los últimos avances tecnológicos a solucionar los grandes retos de la humanidad: educación, energía, medio ambiente, alimentación, salud, pobreza, seguridad, espacio y agua.

La ambición de la Singularity es ser el motor donde se piensa y se origina el cambio. Un cambio que se define con dos adjetivos: disruptivo (brusco) y exponencial (en contraposición a la evolución incremental). Los cursos de la Singularity –diez días cuestan unos 16.000 dólares, en el campus situado en las instalaciones de la NASA en el Silicon Valley, que han sido alquiladas por Google– son un lujo intelectual y aspiracional para directivos, emprendedores, líderes y mentes abiertas de todas partes.

Porque el futuro llega a una velocidad sin precedentes y que el cerebro humano ya no es capaz de comprender: desde el origen del mundo hasta 2003, se crearon 5 exabytes de información. En 2010 eran 5 exabytes cada 48 horas. En 2015 se crean cada 120 segundos. (Y, ¿quién puede asimilar qué es un exabyte?).

Pero no hace falta ir tan lejos: Google ya tiene 5 productos con más de mil millones de usuarios cada uno. Y LinkedIn ya tiene 350 millones de usuarios y gana 173.000 nuevos cada día (2 por segundo). Pero al mismo tiempo, en la Singularity se preguntan: ¿qué pasa con los 4.000 millones de personas de los países todavía en desarrollo que aspiran algún día a tener también su oportunidad? ¿Con los empleados cuyos puestos de trabajo se sustituyen por robots? ¿Con los ordenadores que se autoprogramarán y pensarán por sí solos? Y hay dudas muy cercanas: si los coches circularán sin conductor, ¿quién se hará responsable del seguro? Las respuestas, como no puede ser de otra manera en este lugar, no tienen fronteras –"Push the boudaries" es otro de los lemas de la cultura común del Valley–. Aquí nada impide pensar en cultivar plantas enriquecidas con proteína animal, con unos costes de producción inferiores a los que requiere la ganadería, y que permitirían alimentar a millones de personas que ahora no tienen acceso a recursos mínimos.

El impacto de la transformación tecnológica es inmenso. "¿Cómo se reequilibra el hecho que el trabajo de 17 ingenieros destruya un negocio que tenía 200.000 empleados?" se pregunta Pau Guinart, doctorando en Humanidades en Stanford. No es una pregunta retórica: Kodak tuvo 50.000 empleados, en Instagram trabajaban 13 personas cuando fue adquirida por Facebook. "Disrupt, or you will be disrupted", es una de las grandes lecciones de la Singularity. Y Washington, la administración, el Gobierno, las normas, quedan lejos. Silicon Valley quiere cambiar el mundo, lo quiere mejorar, de manera. Pero hay voces que alertan de que el liberalismo extremo agrava las desigualdades, del riesgo de que el mundo sea aún más injusto.

Para los empresarios catalanes de FemCat, el viaje tuvo un mensaje transformador y esperanzador: viniendo de un país más acostumbrado a utilizar las cosas que a crearlas, Silicon Valley es la demostración de que con la tecnología al alcance de cualquiera, nada impide apuntarse a la revolución.

Moffett Field, la instalación de la NASA que Google ha arrendado por 60 años, y donde tiene su campus la Singularity ARCHIVO

¿qué pasa con los 4.000 millones de personas de los países todavía en desarrollo que aspiran algún día a tener también su oportunidad? ¿Con los empleados cuyos puestos de trabajo se sustituyen por robots? ¿Con los ordenadores que se autoprogramarán y pensarán por sí solos?

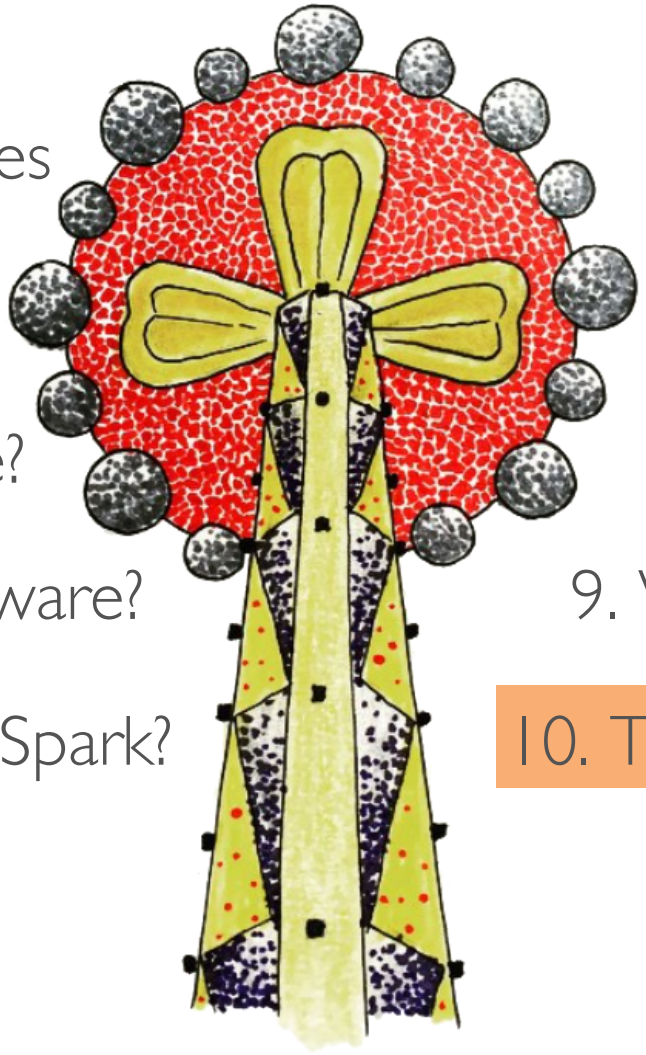
MAY WE BE CONTROLLED?

We may be controlled by algorithms that are likely to predict what we are about to do

- Privacy was the central challenge in the second wave era
- In the next ware of Cognitive Computing, the challenge will be safeguarding free will.

TALK OUTLINE

1. Computing Waves
2. Why now?
3. Future Hardware?
4. Software Middleware?
5. What is Apache Spark?

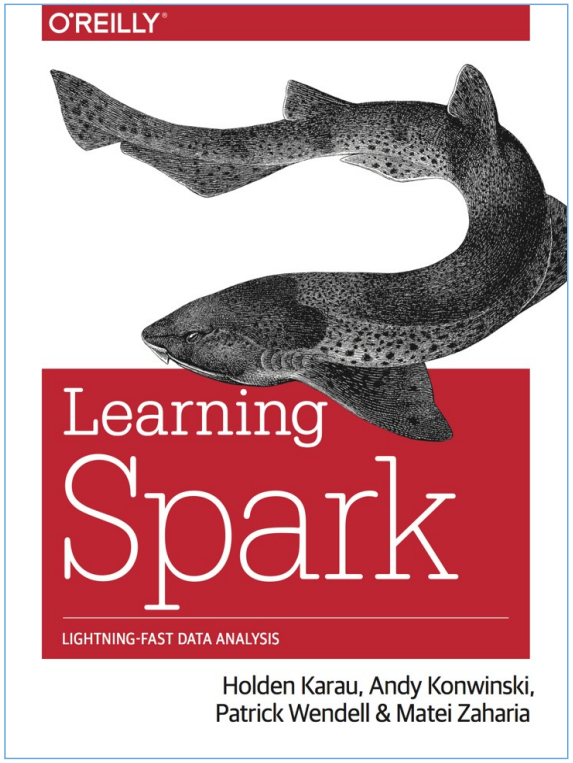


6. Spark Basics
7. Spark Ecosystem
8. Spark & Marenostrum
9. What next?
10. To learn more …

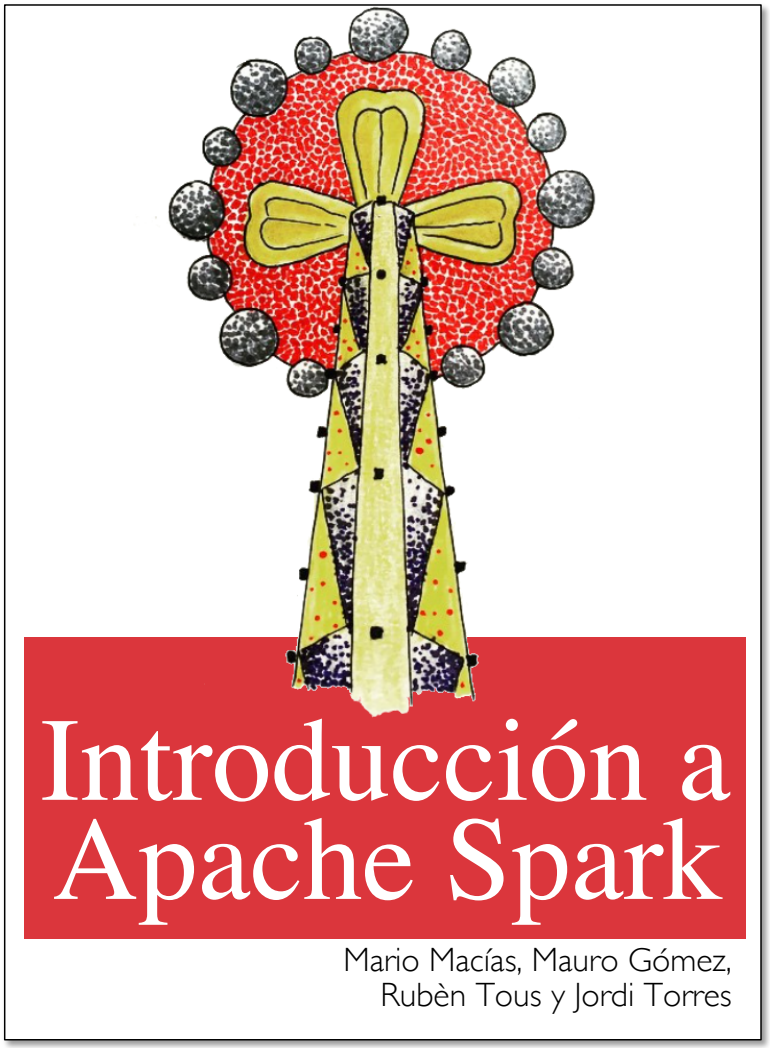
Spark meetup BARCELONA

TO LEARN MORE …

- http://spark.apache.org/docs
- https://databricks.com/spark/developer-resources
-



AND NEXT FALL…



Thank you for your attention!

JORDI TORRES

@JordiTorresBCN www.JordiTorres.eu