Big Data Challenges in Bioinformatics

BARCELONA SUPERCOMPUTING CENTER
COMPUTER SCIENCE DEPARTMENT
Autonomic Systems and eBusiness Platforms

Jordi Torres
Jordi.Torres@bsc.es
Talk outline

**Important Open Issues**
- Data Transfer
- Security & Privacy
- Data Store → Cloud!

**Deriving Value**
- Volume
- Variety
- Velocity

Arg Data Definition

**Computing Challenge**
- Data Processing
- Data Management
- Data Infrastructure

**Conclusion**
- Computing Central Model → Data Centric Model
Data is now considered the *Fourth Paradigm in Science*

the first three paradigms were experimental, theoretical and computational science.

This shift is being driven by the rapid growth in data from improvements in

*scientific instruments*
Scientific instruments

Physics

Large Hadron Collider produced around 15 petabytes of data in 2012.

Astronomy

Large Synoptic Survey Telescope it’s anticipated to produce around 10 petabytes per year.
Example: In Genome Research?

Cost of sequencing a human-sized genome

Source: National Human Genome Research Institute (NHGRI) http://www.genome.gov/sequencingcosts/
Data Deluge: Due to the changes in big data generation

Example: Biomedicine

Important open issues

- **Transfer of data from one location to another (*)**
  - shipping external hard disks
  - processing the data while it is being transferred
  - Future? **Data won’t be moved!**

(*) *Out of scope of this presentation*

Important open issues

• Security and privacy of the data from individuals (*)
  – The same problems that appear in other areas
  – Use advanced encryption algorithms

Important open issues

• **Increased need to store data (*)**
  – Cloud–based computing solutions have emerged
Important open issues

- Increased need to store data (*)
  - Cloud-based computing solutions have emerged
  - The most common Cloud Computing inhibitors should be tackled

- Security
- Privacy
- Lack of Standards
- Data Integrity
- Regulatory
- Data Recovery
- Control
- Vendor Maturity
- ...
The most critical open issue

DERIVING VALUE VIA HARNESSING

VOLUME, VARIETY AND VELOCITY (*)

(*) Big Data definition?

The most critical open issue

DERIVING VALUE VIA HARNESSING VOLUME, VARIETY AND VELOCITY (*)

(*) Big Data definition?
The most critical open issue

**DERIVING VALUE VIA HARNESSING VOLUME, VARIETY AND VELOCITY (**)**

(*) Big Data definition?

Source: cetemma - mataró
The most critical open issue

DERIVING VALUE VIA HARNESSING VOLUME, VARIETY AND VELOCITY (*)

The information is non actionable knowledge

(*) Big Data definition?
What is the usefulness of Big Data?

Performs predictions of outcomes and behaviors

Approach: Machine Learning “works” in the sense that these methods detect subtle structure in data relatively easily without having to make strong assumptions about parameters of distributions
Data Analytics: To extract knowledge

Big data uses inductive statistics and concepts from nonlinear system identification to infer laws (regressions, nonlinear relationships, and causal effects) from large data sets to reveal relationships, dependencies, and to perform predictions of outcomes and behaviors.

(Out of scope of this presentation) :- ( (*) Wikipedia)
Talk outline

**IMPORTANT OPEN ISSUES**
- Data Transfer
- Security & Privacy
- Data Store ➔ CLOUD!

**DERIVING VALUE**
- Volume
- Variety
- Velocity

**Computing Challenge**

**Data Processing**
- Data Management
- Data Infrastructure

**Conclusion?**

**Computing Central Model ➔ Data Central Model**
Challenges related to us?

The big data problem:

In the end it is a Computing Challenge
Researchers need to crunch a large amount of data very quickly (and easily) using high-performance computers.

Example: A de novo assembly algorithm for DNA data finds reads whose sequences “overlap” and records those overlaps in a huge diagram called an assembly graph. For a large genome, this graph can occupy many terabytes of RAM, and completing the genome sequence can require weeks or months of computation on a world-class supercomputer.
What does life science research do at BSC?

Pipeline schema:

0. Receive the data:
   Raw Genome Data: 120Gb

MareNostrum
   25-40 h / 50-100 cpus

Altix / CLL cluster
   50-60h / 1-10 cpus

Common Storage
Also data deluge appears in genomics

The DNA data deluge comes from thousands of sources

- More than 2000 sequencing instruments around the world
- more than 15 petabytes x year of genetic data.

And soon, tens of thousands of sources!!!!
The total computing burden is growing

DNA sequencing is on the path to becoming an everyday tool in medicine.

Computing, not sequencing, is now the slower and more costly aspect of genomics research.
How can we help at BSC?

- Something must be done now, or else we’ll need to put vital research on hold while the necessary computational techniques catch up—or are invented.

- What is clear is that it will involve both better algorithms and a renewed focus on such “big data” approaches in managing and processing data.

- How?

  Doing outstanding research to speed up this process
What is the time required to retrieve information?

1 Petabyte = 1000 x (1 Terabyte)
What is the time required to retrieve information?

Assume 100MB/sec
What is the time required to retrieve information?

Assume 100MB/sec

Scanning 1 Terabyte: more than 5 hours
What is the time required to retrieve information?

scanning 1 Petabyte: more than 5,000 hours
massive parallelism

not only in computation but also in storage

assume 10,000 disks: scanning 1 TB takes 1 second

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

**Important Open Issues**
- Data Transfer
- Security & Privacy
- Data Store → Cloud!

**Deriving Value**
- Volume
- Variety
- Velocity

- Big Data Definition

**Computing Challenge**
- Data Processing
- Data Management
- Data Infrastructure

**Conclusion?**
- Computing Central Model → Data Centric Model
To support this massive data parallelism & distribution it is necessary to redefine and improve:

- Data Processing across hundreds of thousands of servers
- Data Management across hundreds of thousands of data devices
- Dealing with new System Data Infrastructure
How do companies like Google read and process data from 10,000 disks in parallel?

Source: http://www.google.com/about/datacenters/gallery/images/_2000/IDI_018.jpg
To meet the challenges: MapReduce

- **Programming Model** introduced by Google in early 2000s to support **distributed computing** (special emphasis in fault-tolerance)

Ecosystem of big data processing tools

- open source, distributed, and run on commodity hardware.
The key innovation of MapReduce is
- the ability to **take a query** over a data set, **divide it**, and **run it** in parallel over many nodes.

Two phases
- Map phase
- Reduce phase
Limitations of MapReduce as a Programming model?

- MapReduce is great but not every one is a MapReduce expert
  “I am a python expert but …. ”

- There is a class of algorithms that cannot be efficiently implemented with the MapReduce programming model

- Different programming models deal with different challenges
  Example: pyCOMPSs from BSC
Big Data characteristics | Requirements from data store
--- | ---
Volume | Scalability
Variety | Scheme-less
Velocity | Relaxed consistency & capacity to digest

Relational databases are not suitable for Big Data problems
- Lack of horizontal scalability
- Complex structures to express complex relationships
- Hard consistency
- ...

Non-relational databases (NoSQL) are the alternative data store

Relaxing consistency → Eventual consistency
General view of NoSQL storage (and replication)
Big Data resource management: open issues in NoSQL

• Query performance depends heavily on data model
• Designed to support many concurrent short queries

→ Solutions:
  – automatic configuration, query plan and data organization
  – BSC-Aeneas: https://github.com/cugni/aeneas
New System Data Infrastructure required

Example: Current computer systems available at genomics research institutions are commonly designed to run general computational jobs where,

– Traditionally the limiting resource is the use of CPU.
– Also, we find a large common storage space shared for all nodes.
Example: Computers in use for bioinformatic jobs

Typical mix of such computer systems and common bioinformatics applications:

→ Bottlenecks & underutilization problems

First approach:
Big Data Rack Architecture:
“Shared Nothing”
Storage Technology: Non Volatile Memory evolution

FLASH + DISK

FLASH AS DISK

FLASH AS MEMORY

(*) HHD 100 cheaper than RAM. But 1000 times slower
**Example: Computers in use for bioinformatic jobs**

- Jobs are responsible of managing the input data: partitions, organisation, merge of intermediate results
  - Large parts of code are not functional, but housekeeping tasks

- Solutions: Active storage strategies for leveraging high-performance in-memory key/value databases to accelerate data intensive tasks
Important: Remote Nodes Have Gotten Closer

- Interconnects have become much faster
- IB latency 2000 ns is only 20x slower than RAM and is 100x faster than SSD

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

Old
Compute-centric Model

New
Data-centric Model

Massive Parallelism
Persistent Memory

Flash
Phase Change

Conclusions: How can we help?

How can IT researchers help scientists like you cope with the onslaught of data?

This is a crucial question and there is no definitive answer yet.

What is clear is that it will involve both better algorithms and a renewed focus on “big data” approaches such as: data infrastructure, data managing and data processing.
Over to you, what do you think?

Thank you for your attention! - Jordi
Thank you to ...
Updated information will be posted at www.JordiTorres.eu