# What Can  Big Data and Cloud Computing do for Scientits?

Salima Benbernou
Université Paris Descartes
LIPADE-Data Managment  and Mining Group
Salima.benbernou@parisdescartes.fr

# A Golden Era in Computing

Explosion of domain applications

Powerful multi-core processors

General purpose graphic processors

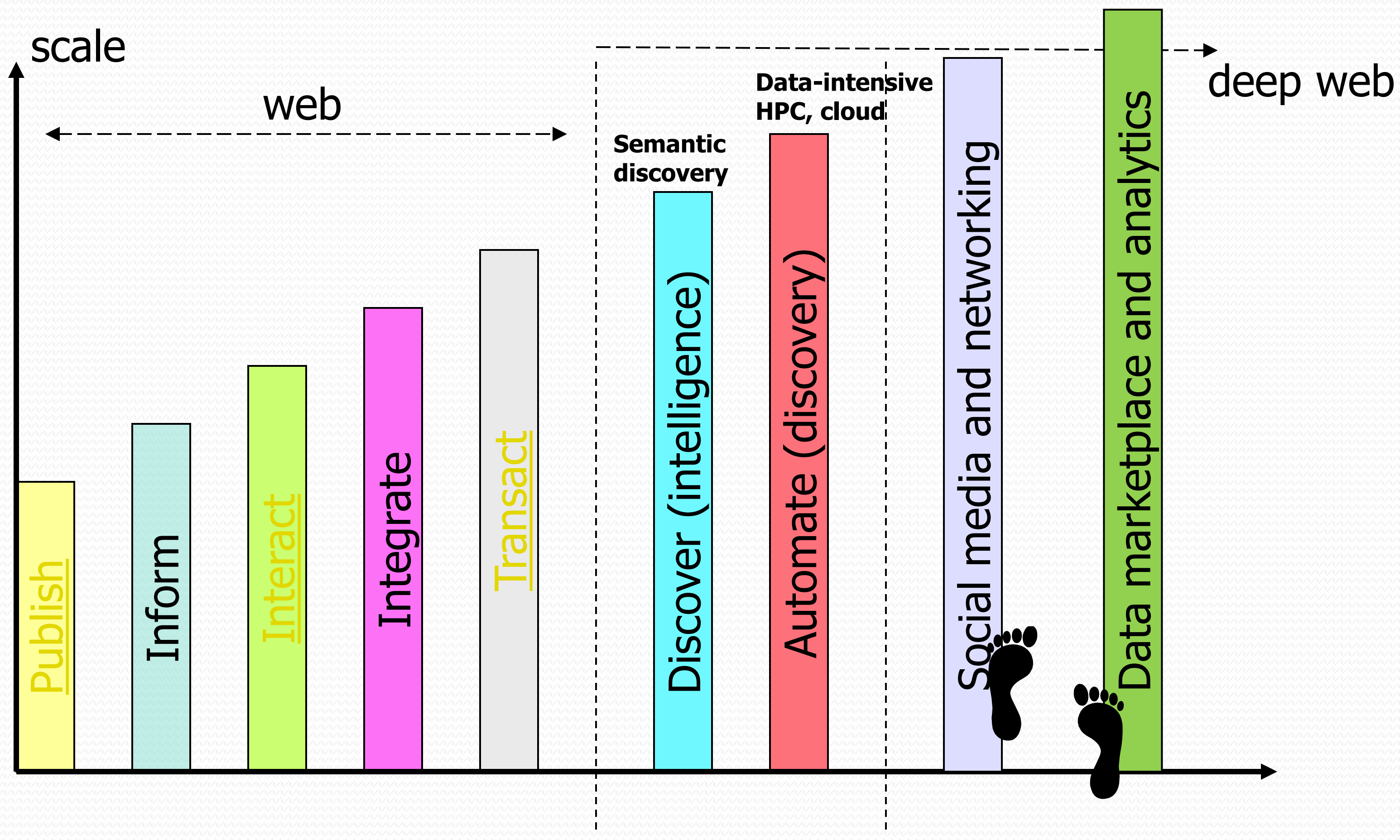
Superior software methodologies

Proliferation of devices

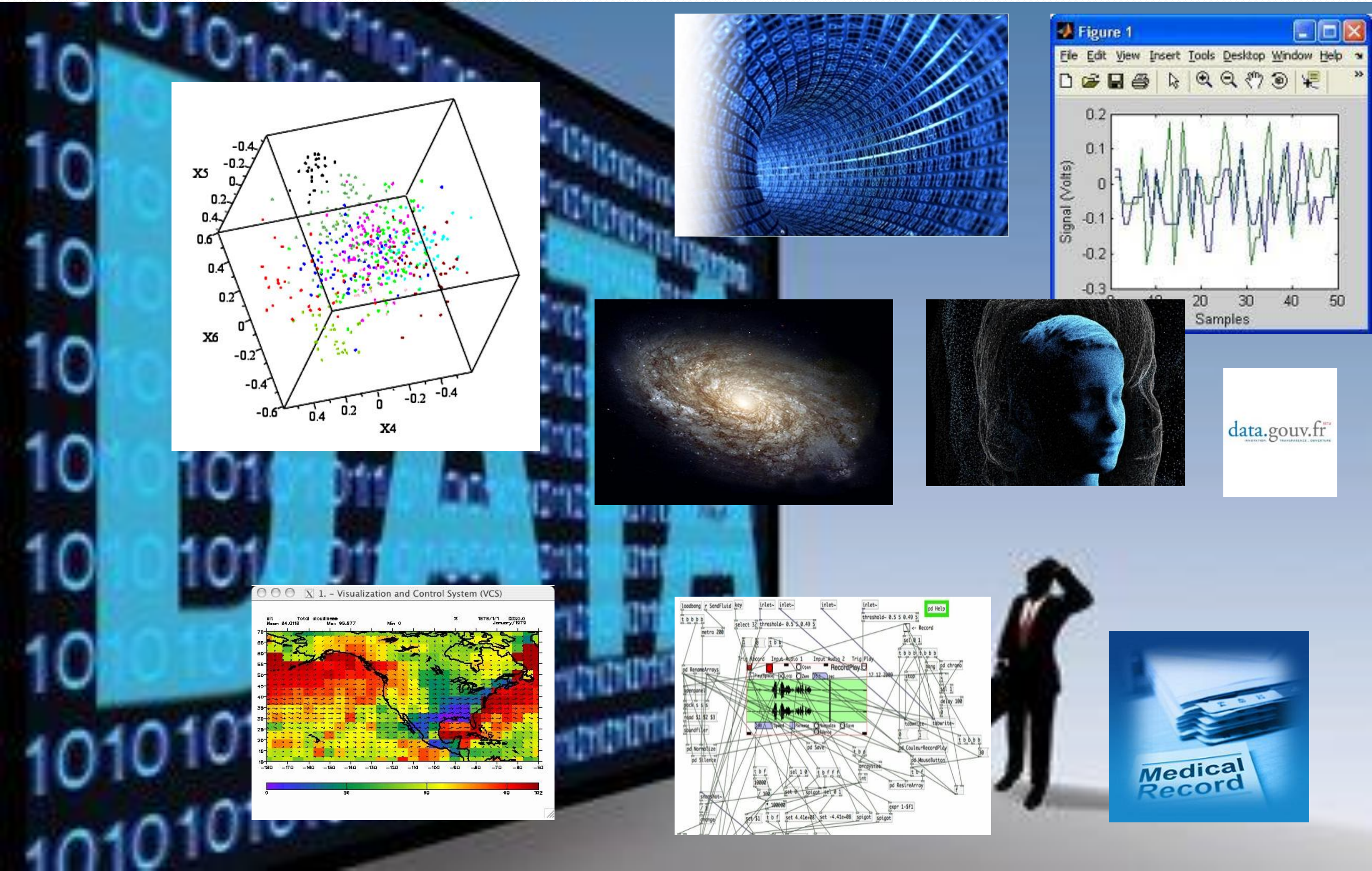
Wider bandwidth for communication

Virtualization leveraging the powerful hardware

# Evolution of Internet Computing



# Big Data in the world

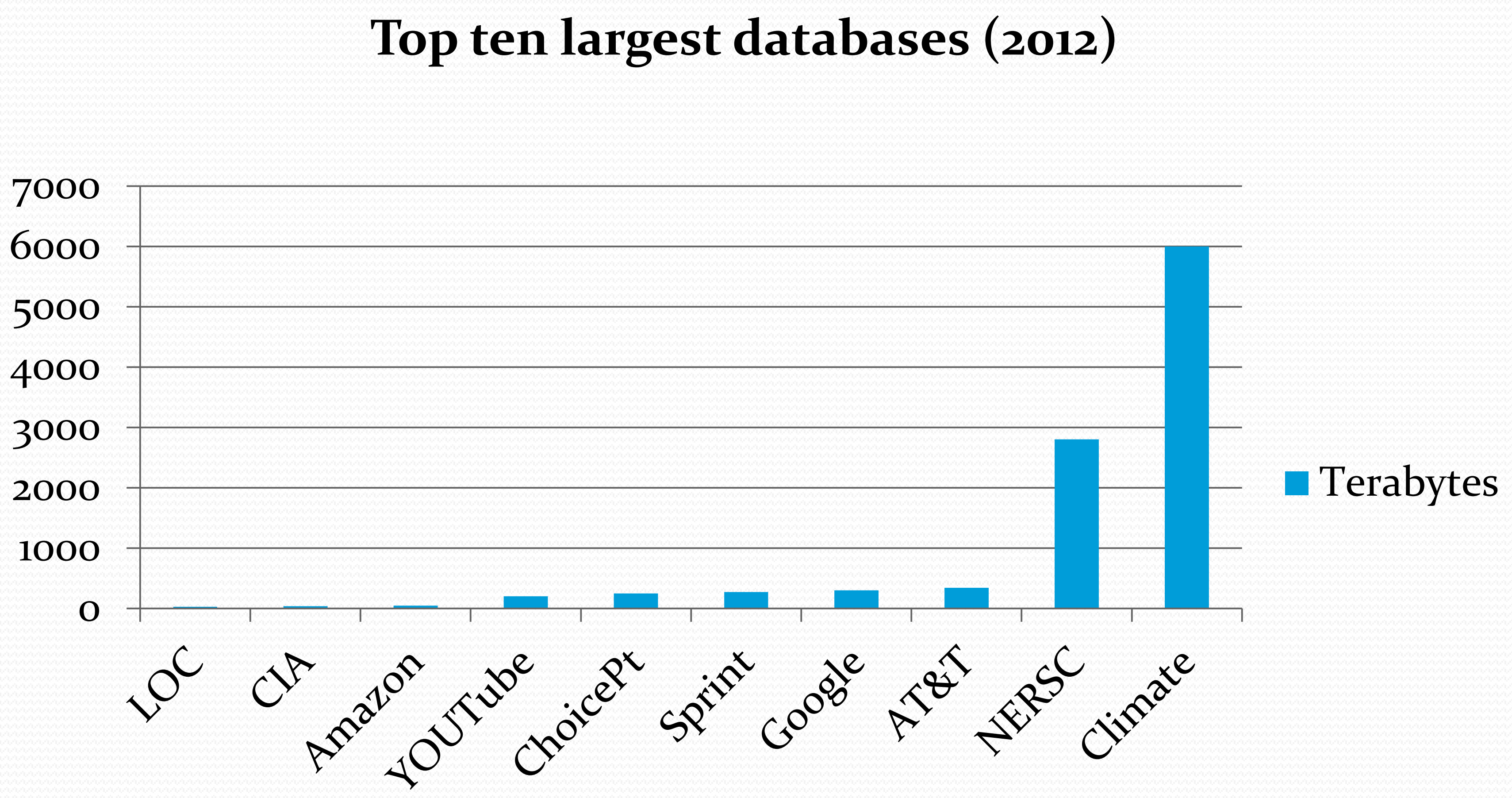


# Big data: Some applications

| Application | Big Data | Algorithms | Compute Style |
|---|---|---|---|
| Scientific study (e.g. earthquake study) | Ground model | Earthquake simulation, thermal conduction, … | HPC |
| Internet library search | Historic web snapshots | *Data mining* | MapReduce |
| Virtual world analysis | Virtual world database | *Data mining* | TBD |
| Language translation | Text corpuses, audio archives,… | Speech recognition, machine translation, text-to-speech, … | MapReduce & HPC |
| Video search | Video data | Object/gesture identification, face recognition, … | MapReduce |

# Why? WEB is replacing the Desktop

# Paradigm in Computing


Top ten largest databases (2012)
7000
6000
5000
4000
3000
2000
1000
0
LOC
CIA
Amazon
YOUTube
ChoicePt
Sprint
Google
AT&T
NERSC
Climate
Terabytes

# What is Cloud Computing?

- **Cloud computing** is Internet-based computing, whereby shared resources, software and information are provided to computers and other devices on-demand, like the electricity grid.
- The cloud computing  is a culmination of numerous attempts at large scale computing with seamless access to virtually limitless resources.

# What is Cloud Computing?

- Delivering applications  and services over the Internet:
  - Software as a service (SaaS)
- Extended to:
  - Infrastructure as a service: Amazon EC2 (IaaS)
  - Platform as a service: Google AppEngine, Microsoft Azure (PaaS)
- Utility Computing: pay-as-you-go computing
  - Illusion of infinite resources
  - No up-front cost
  - Fine-grained billing (e.g. hourly)

# Essential Characteristics

- On-demand self-service
- Broad network access
- Resource pooling
- Rapid elasticity
- Measured Service

## Service Models

- Cloud Software as a Service (SaaS)
- Cloud Platform as a Service (PaaS)
- Cloud Infrastructure as a Service (IaaS)

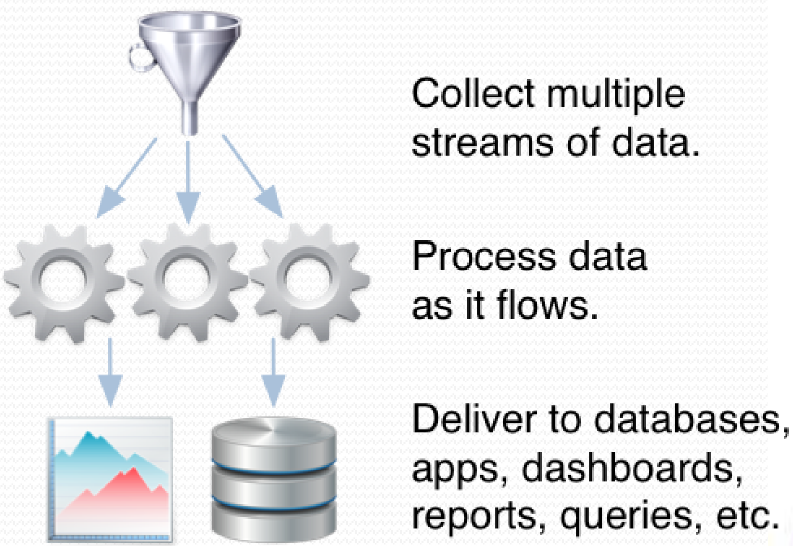
### Deployment Models

- Private cloud
- Community cloud
- Public cloud
- Hybrid cloud

Source: NIST - http://csrc.nist.gov/groups/SNS/cloud-computing/index.html

# More in cloud …

- Data as a Service (DaaS)



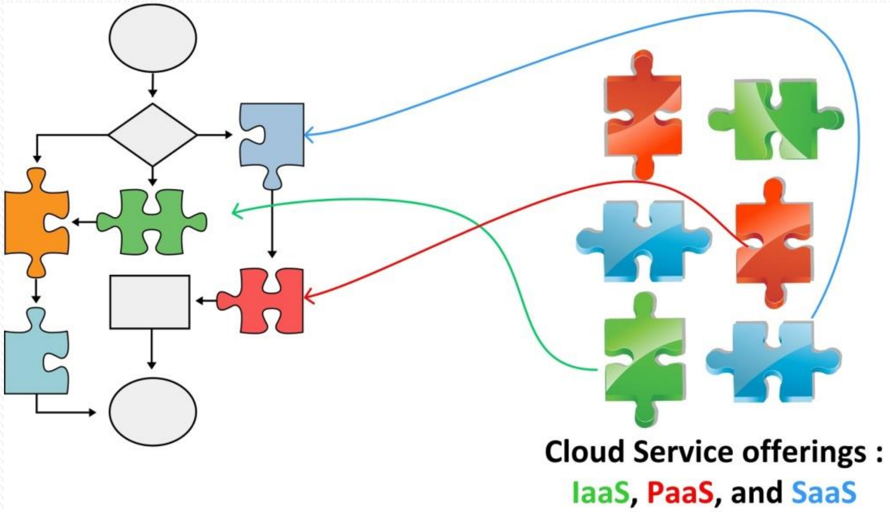

Data Delivery as service

Figure 2: Basic data value chain


Source: Liaison Technologies

# What is Cloud Computing?

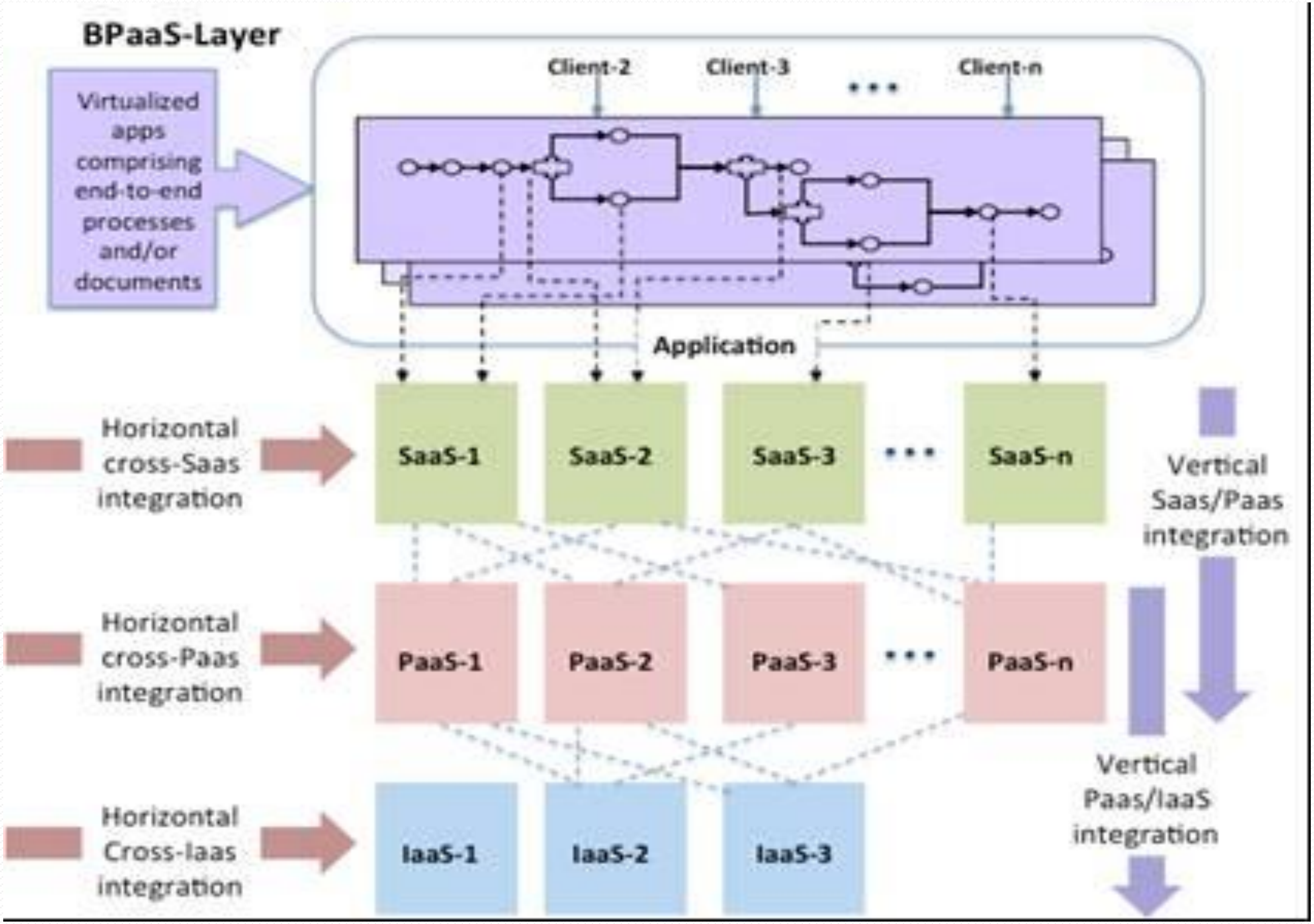
- Cloud federation, Business Process as a Service (BPaaS) (Benbernou et al Cloud-I@VLDB2012, ICWS2012) and workflow



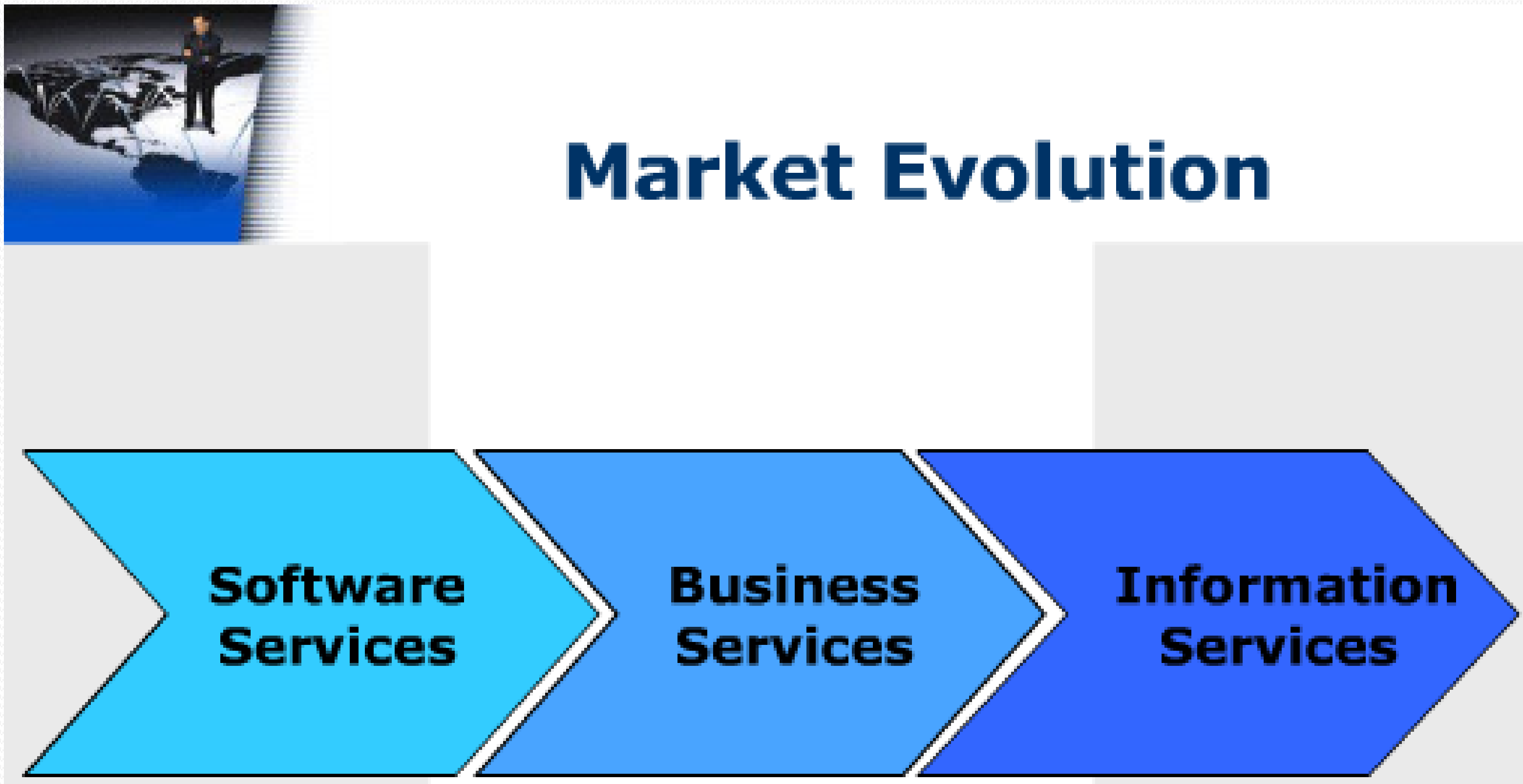

Compose and mashup

The next step forward in the evolution of cloud computing

# Syndicated mixed-channel cloud delivery model



# Market moves to « Everything as a Service » !

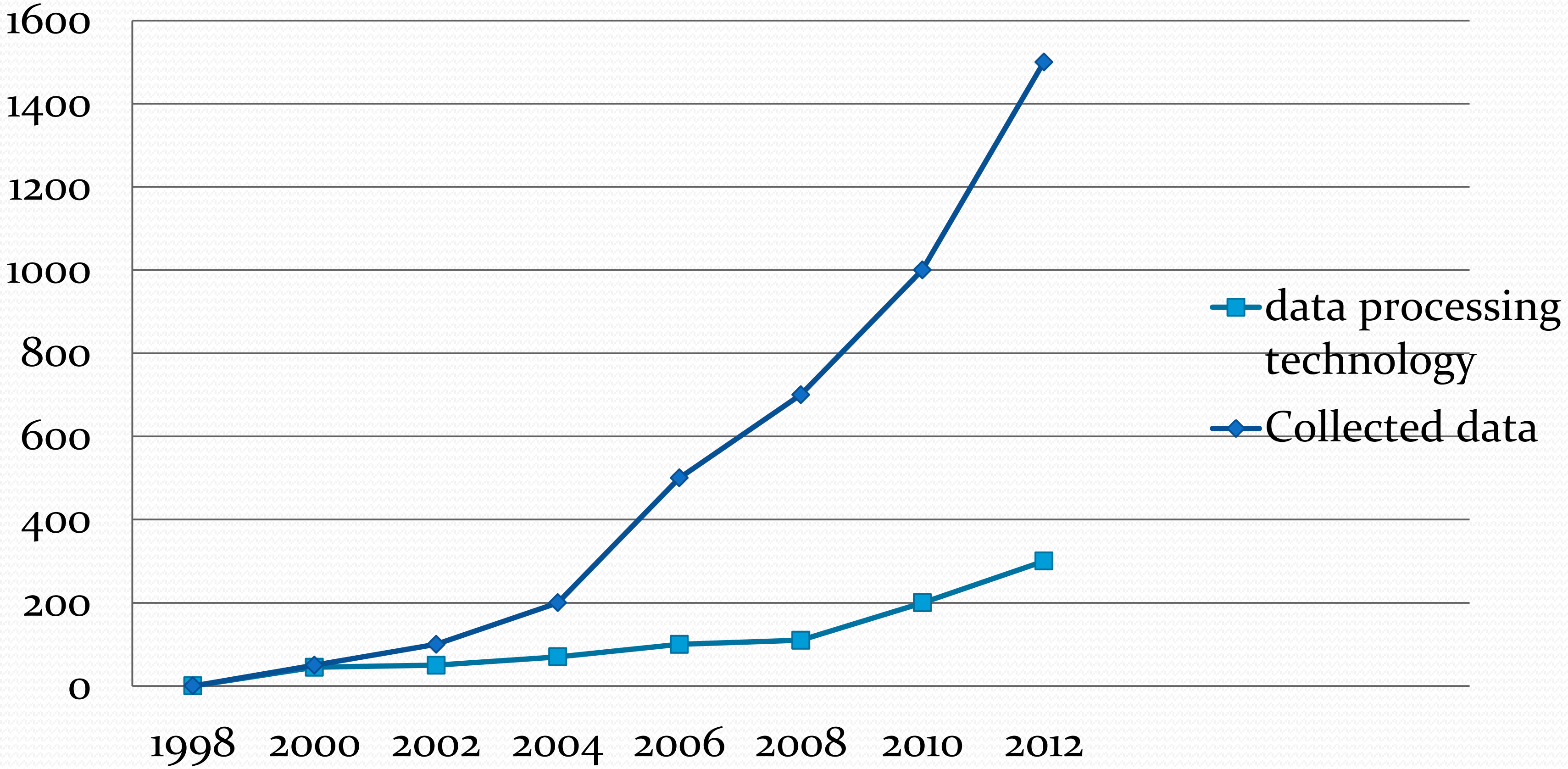



# Exploring Cloud for Scientific missions

- Gaining traction in commercial world (Amazon, Google, Yahoo, ..) offering pay as you go cycles for extra computing power in organisations.

- Does the approach meet the computing and data storage demands of the nation's scientific community?

# Scientific data grows much faster than technology




Wintercorp Survey

# Scientific managment  now

- Legacy software
- In main memory of supercomputers
- Database too rigid to use

As data grows, problem changes

- Difficult  and slow
- Some data discarded

Bridge CS and domain sciences

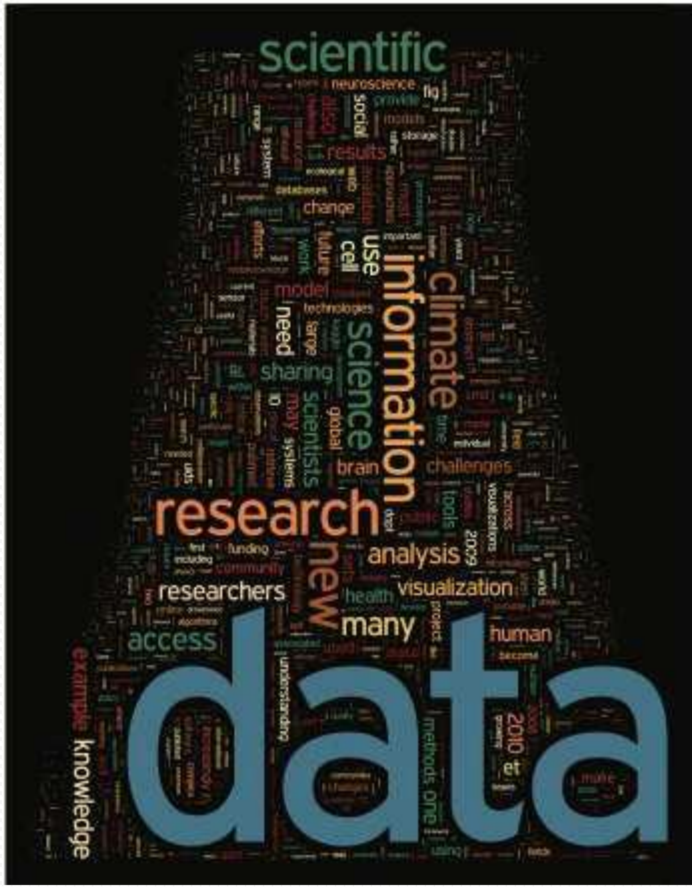
# Data-driven science

Past:

- Theory
- Simulation
- Experiments

The « fourth paradigm »

Scientific breakthrough computing on massive data

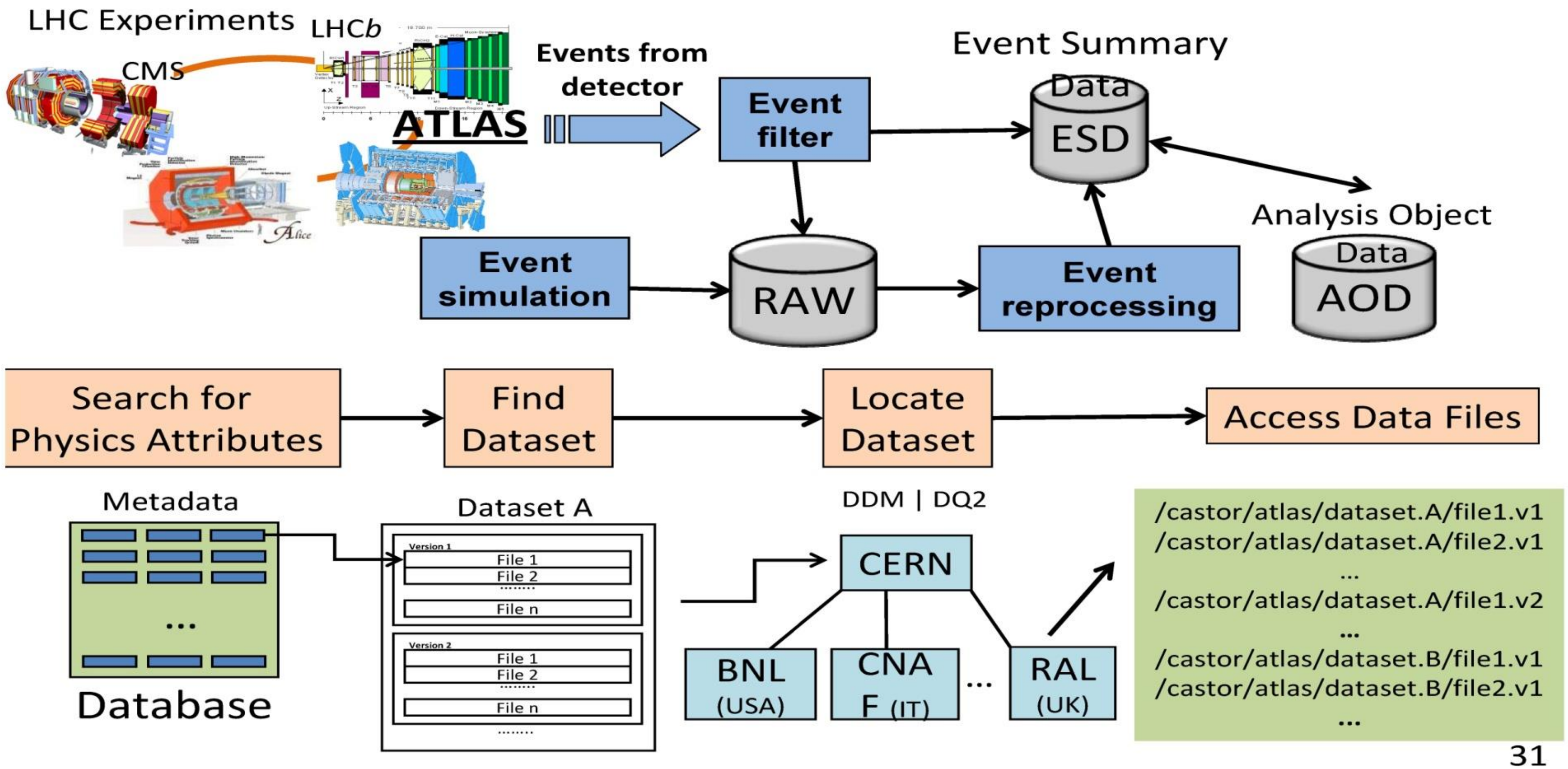



From Anastasia Alaimaki

# The CERN large hadron collider, now

100 M sensors/dectection
40 M detecttions/sec

# ATLAS experiment (simplified)



# Some current projects

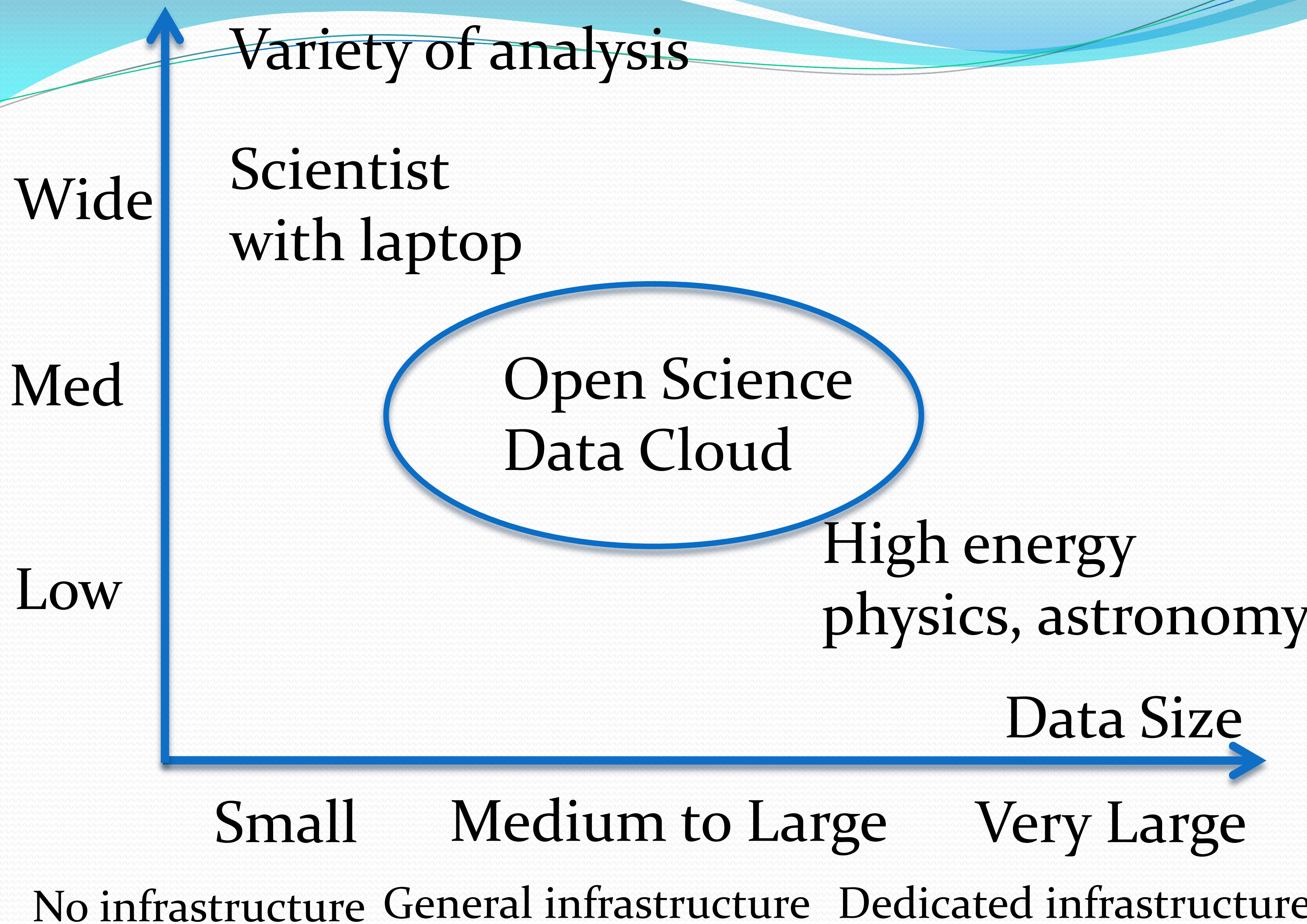
- The Magellan project

•Serving the needs of mid- range computing and future data-intensive computing workloads.

•A set of research questions was formed to probe various aspects of cloud computing from performance, usability, and cost.

# Open Science Data Cloud

The OCC is a not-for-profit supporting the scientific community by operating cloud infrastructure.


Variety of analysis
Wide
Med
Low
Scientist
with laptop
Open Science
Data Cloud
High energy
physics, astronomy
Data Size
Small
Medium to Large
Very Large
No infrastructure
General infrastructure
Dedicated infrastructure

# Project Bionimbus

Institute for Genomics & Systems Biology

## Bionimbus Cloud

Bionimbus is a cloud-based system for managing, analyzing and sharing genomic data.

News | About Bionimbus | Public Data | Using Bionimbus | Registered Users | Support | Sponsors

Search Bionimbus Cloud

Search »

### Complete Genomics Chooses the Bionimbus as Mirror Site for CGI 60 Genomes Release

Edit

Published on 2011/02/03 in Uncategorized. Closed

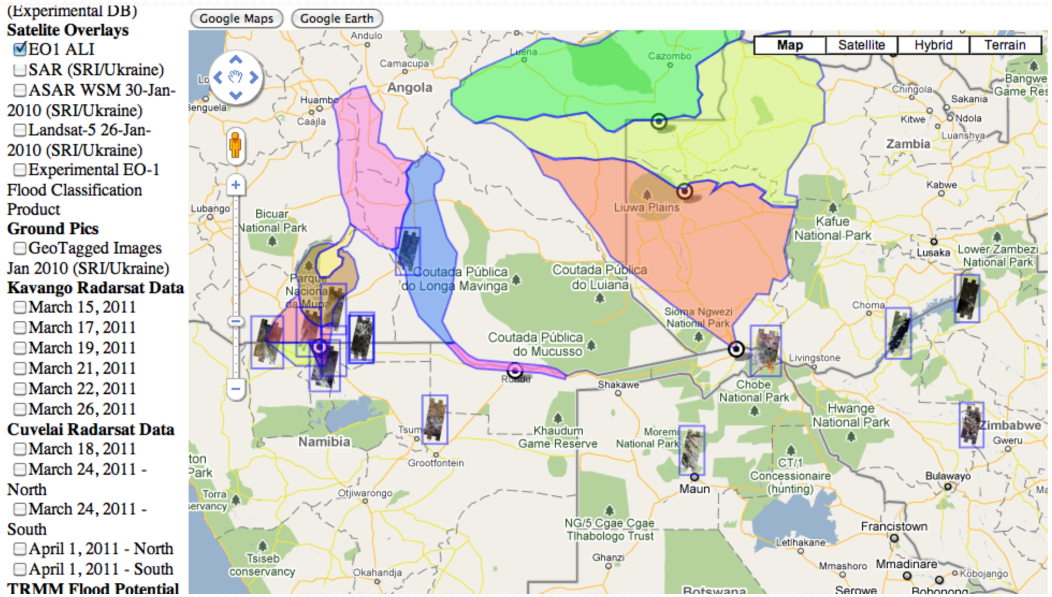
Complete Genomics Inc. has chosen the Bionimbus Community Cloud as a mirror site for their 60 Genomes dataset.

The 60 Genomes dataset can be found here, as part of the public data that Bionimbus makes available to researchers. With the Bionimbus Community Cloud, the data is available via both the commodity Internet, as well as via high performance research networks, such as the National LambdaRail and Internet2.

The genomes in the dataset have on average more than 55x mapped read coverage, and the sequencing of these 60 genomes generated more than 12.2 terabases (Tb) of total mapped reads. This dataset will complement other publicly available whole genome data sets, such as the 1000 Genomes Project's recent publication of six high-coverage and 179 low-coverage human genomes. Forty of the sixty genomes are available now and the remainder will be available at the end of March.

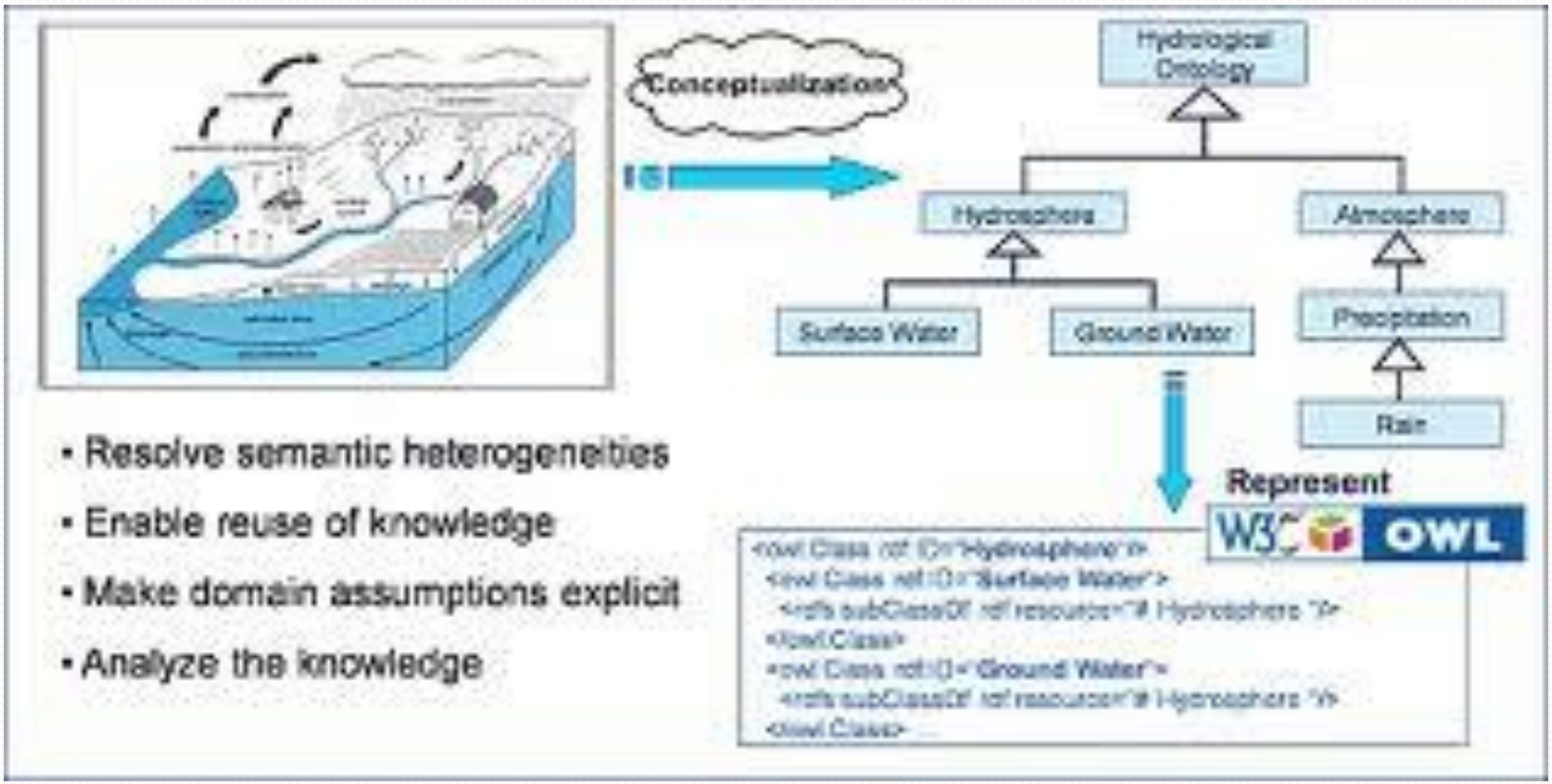
www.bionimbus.org (biological data)

# Project Matsu 2:
# An Elastic Cloud For Earth Science Data (& disaster relief)




matsu.opencloudconsortium.org

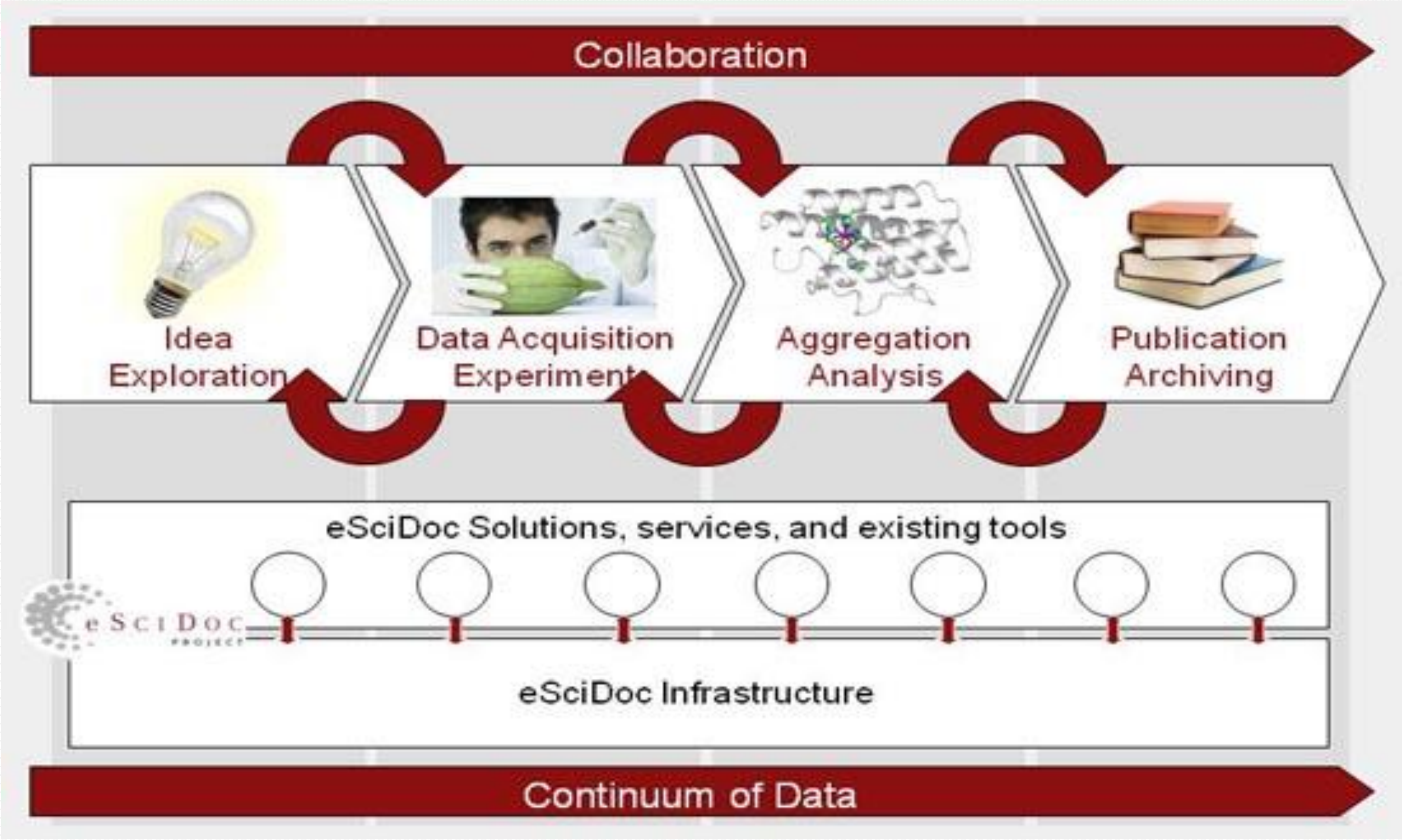
# Issues: Semantic and heterogeneities



# Meta data templates

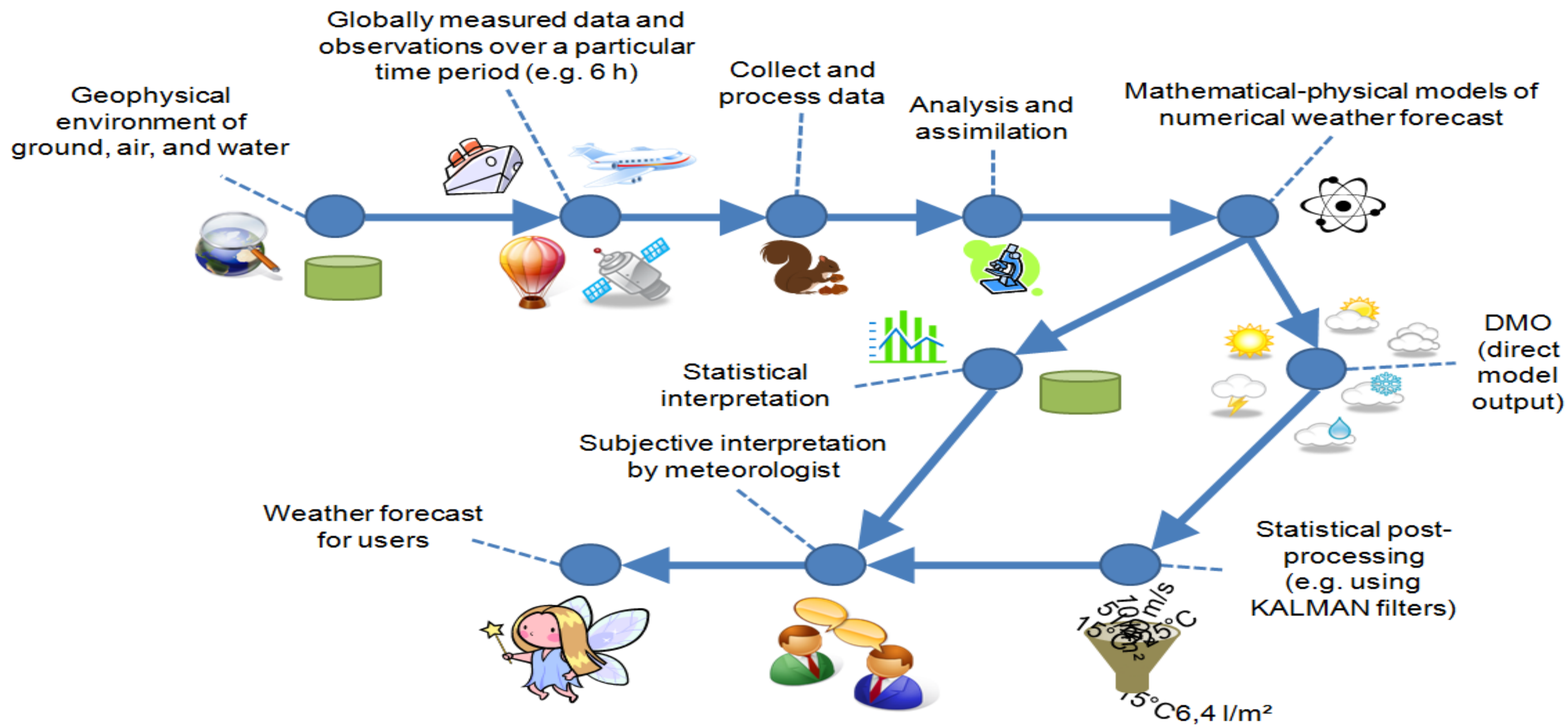
- The need of templates describing how a cloud offering is presented & consumed.
- The offering is abstracted from the specific resources offered.
- The provider uses service template to describe in a general form what a cloud service can offer.

# Issue : Scientific workflows

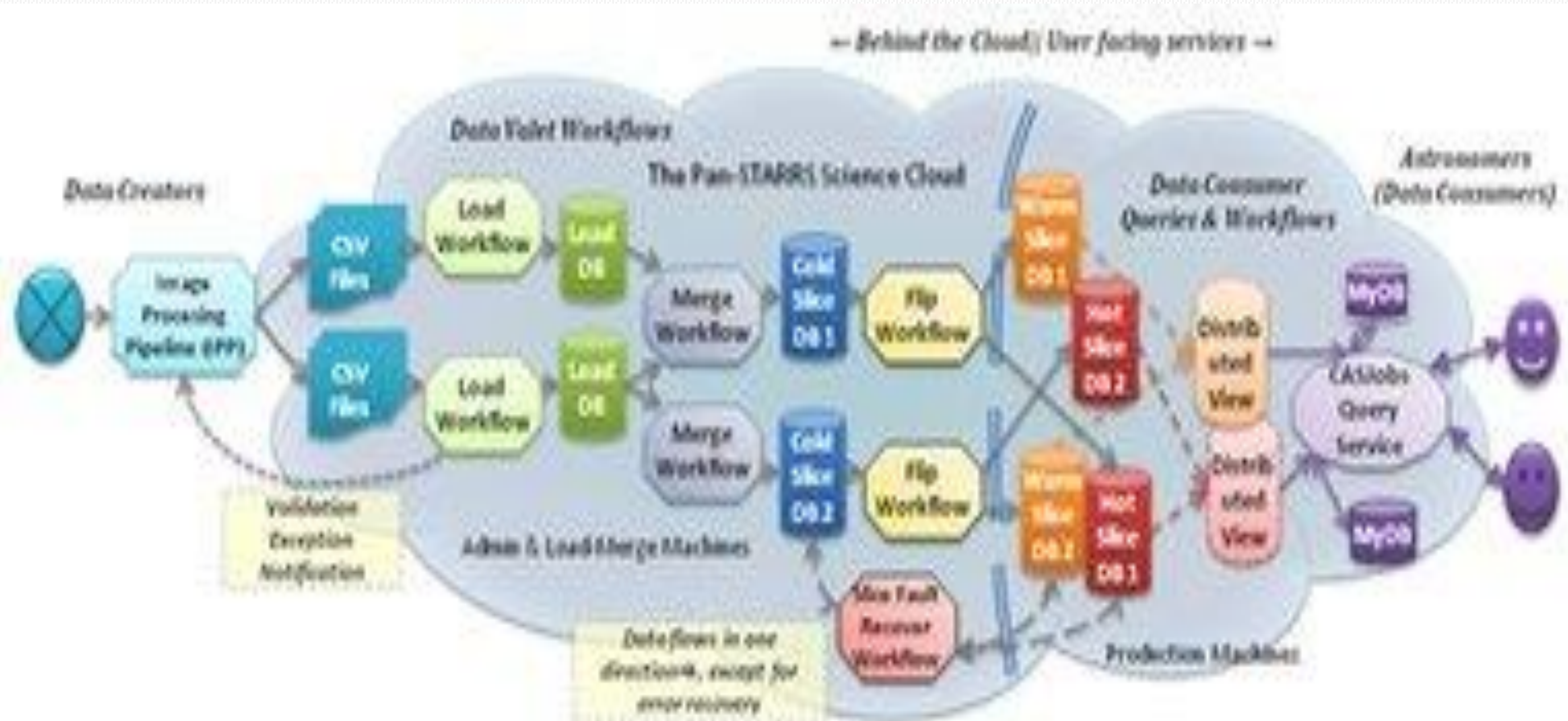


# What are scientific workflows?

- Scientific experiments/computations/simulations modeled and executed as wokflows
- Characteritics :deal with huge mouts of data, are often long running, usually data driven, can integrate muliple data sources (i.e. sensors)



# Scientific workflow:Trident




The Panoramic Survey Telescope and Rapid Response helps to detect objects in the solar system that might pose a threat to Earth.

# Sharing scientific workflows


The myExperiment social web site was launched in November 2007 and with over 1100 workflows

# Issue: scientific workflows and the clouds

- Workflow technology can be applied to improve the IT support for scientific experiments and simulations
  - Provide an end-to-end support for experiments
    - Automate all phases of an experiment – pre-, post-processing, execution, visualization - by a single workflow
  - and business processes
    - That may also require support for simulations
  - Parallel execution of experimental runs
- Clouds will have an even more important role for scientific experiments and simulations

# Evolution for the workflow

- Workflow are already used in E-science
- Some workflow systems in e-science: Kepler, Taverna, Pegasus, Trident, Simulink, …

- To be improved
  - Robustness, fault handling
  - Flexibility and adaptability
  - Reusability
  - Scalability
  - Interaction with users, user-friendliness of tools
  - science skills required from scientist…

# Issue: Querying and processsing big data

MapReduce

- A computing model based on heavy  distribution that scales huge volumes of data (data-intensive computing on commodity clusters)
  - 2004: google publication
  - 2006:open source implementation, Hadoop.
- Data distributed on a large number of shared nothing machine
- To process and to analze large quantities of data
  - Use parallelism
  - Push data to machines.

# What is MapReduce Used For?

- At Google:
  - Index building for Google Search
  - Article clustering for Google News
  - Statistical machine translation
- At Yahoo!:
  - Index building for Yahoo! Search
  - Spam detection for Yahoo! Mail
- At Facebook:
  - Data mining
  - Ad optimization
  - Spam detection

# What is MapReduce Used For ?

- In research:
  - Analyzing Wikipedia conflicts (PARC)
  - Natural language processing (CMU)
  - Climate simulation (Washington)
  - Bioinformatics (Maryland)
  - Particle physics (Nebraska)
  - **<Your application here>**

# Issue: privacy preserving


- Privacy aware outsourcing the data
- Privacy aware reusing fragment from scientific worflows
- Privacy aware crowdsourcing the data (expertise people)

# Research questions:

Scientific data managment - essential technology for accelerating scientific discoveries

1. Develop technology to encapsulate a scientist’s data and analysis tools and to export, save and move these between clouds.
2. Develop protocols, utilities, and applications so that new racks and containers can be added to data clouds with minimal human involvement.
3. Develop technology to support the long term, low cost preservation of data in clouds.

# Human problem

- Pushing the collaboration between scientists and computer science

- Avoid more than one year to get data and learn more about scientific applications and datasets.


MERCI!
THANK YOU!
FRAPAR.