RESEARCH WITH PLYMOUTH UNIVERSITY

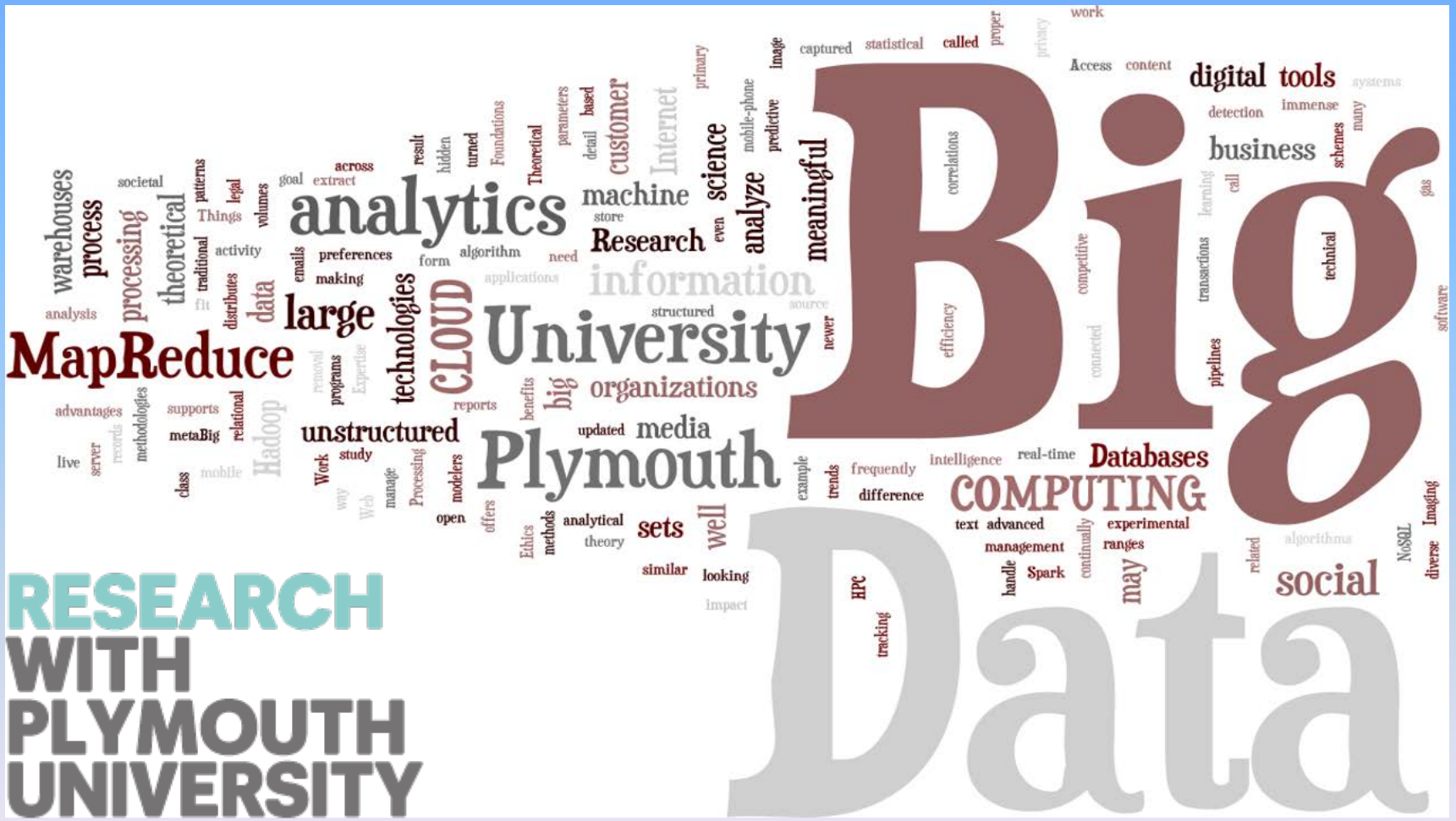
An overview of BIG DATA analysis

Dr. Giovanni Masala

Plymouth University, UK

giovanni.masala@plymouth.ac.uk

Big data group in Plymouth



https://www.plymouth.ac.uk/research/big-data-group

Big data analytics

- **Big Data Science and Foundations**
- Novel Theoretical Models for Big Data
- Machine Learning – Pattern recognition
- Statistics and Data Science
- **Big Data Infrastructure**
- Cloud/Grid/Stream Computing for Big Data
- High Performance/Parallel Computing Platforms for Big Data

Big data analytics

Big Data Management

- Search and Mining of variety of data including scientific and engineering, social, sensor/IoT/IoE, and multimedia data
- Algorithms and Systems for Big Data
- Search Distributed, and Peer-to-peer Search
- Big Data Search Architectures, Scalability and Efficiency
- Data Acquisition, Integration, Cleaning, and Best Practices

- **Big Data Search and Mining**
- Social Web Search and Mining
- Web Search Algorithms
- Systems for Big Data Search

Big data analytics

Big Data Security, Privacy and Trust

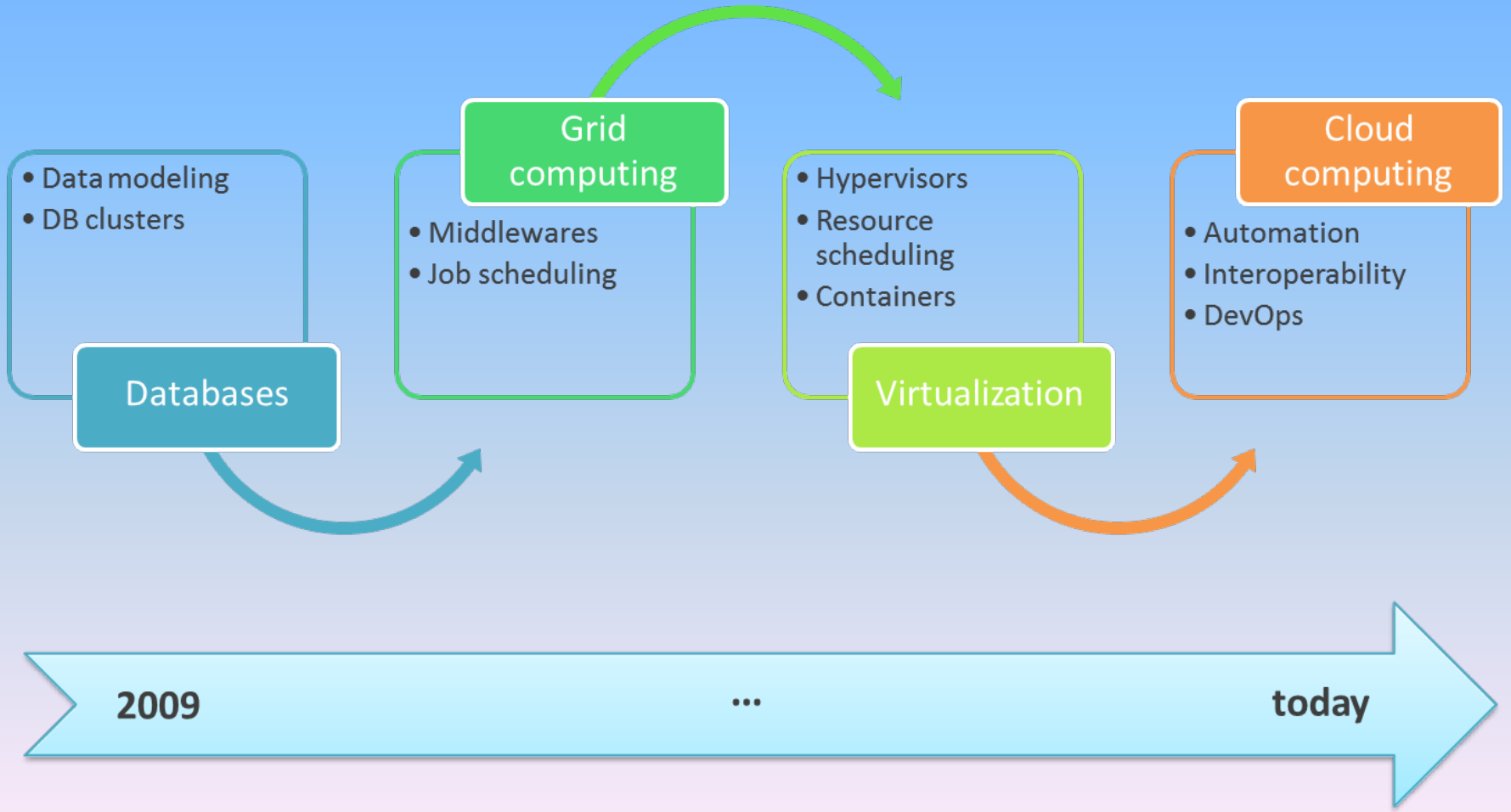
- Sociological Aspects of Big Data Privacy
- Trust management in IoT and other Big Data Systems
- Privacy Threats of Big Data

Big Data Applications

- Complex Big Data Applications in Science, Engineering, Medicine, Healthcare, Finance, Business, Law, Education, Transportation, Retailing, Telecommunication
- Big Data Analytics in Small Business Enterprises (SMEs)
- Big Data Analytics in Government, Public Sector and Society in GeneralReal-life Case Studies of Value Creation through
- Big Data Analytics
- Big Data as a Service

Big Data Infrastructure

Short history



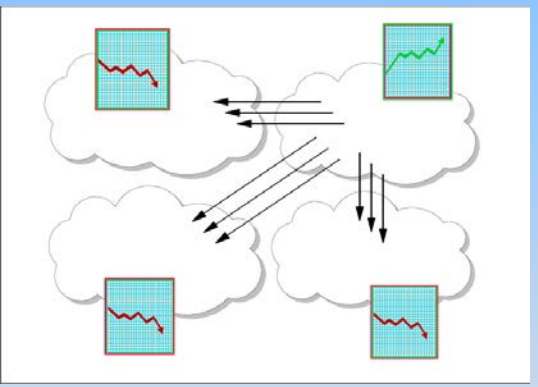
What is Grid Computing?

The term *Grid computing* originated in the early 1990s as a metaphor for making computer power as easy to access as an electric power grid.

- The definitive definition of a Grid is provided by Ian Foster in his article "What is the Grid?
  - Computing resources are not administered centrally.
  - Open standards are used.
  - Non-trivial quality of service is achieved.
- IBM : "A Grid is a type of parallel and distributed system that enables the sharing, selection, and aggregation of resources distributed across multiple administrative domains based on the resources availability, capacity, performance, cost and users' quality-of-service requirements"

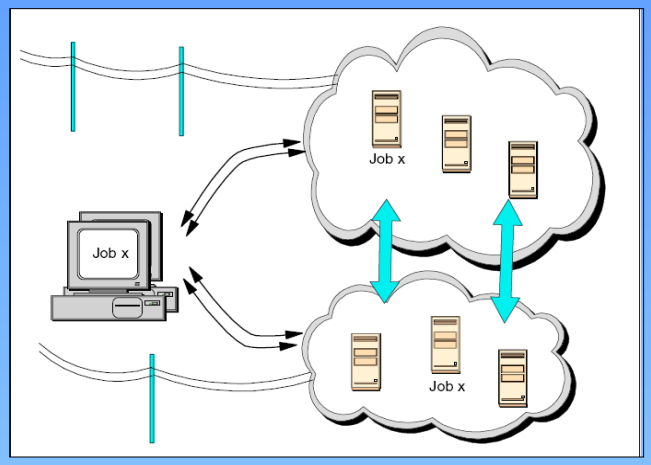
- **Exploiting under utilized resources**
  - Computing:
    - Desktop: less than %5
    - Even servers in many organizations
  - Unused disk capacity
  - Implications:
    - without undue overhead.
    - remote machine must meet any special hardware, software, or resource requirements
- **Parallel CPU capacity**
  - Submiting jobs on different machines
  - Barriers often exist to perfect scalability.
- **Applications**
  - Grid-enabled applications
  - no practical tools for transforming arbitrary applications to exploit the parallel capabilities of a grid.

What grid computing can do



- Virtual resources and virtual organizations for collaboration
  - More capable than distributed computing
    - Wider audience
    - Open standards, hence highly heterogeneous systems
  - Data, equipment, software, services, licenses,…
  - Several real and virtual organizations
  - Access to additional resources
    - special equipment, software, licenses, and other services
    - Resource balancing

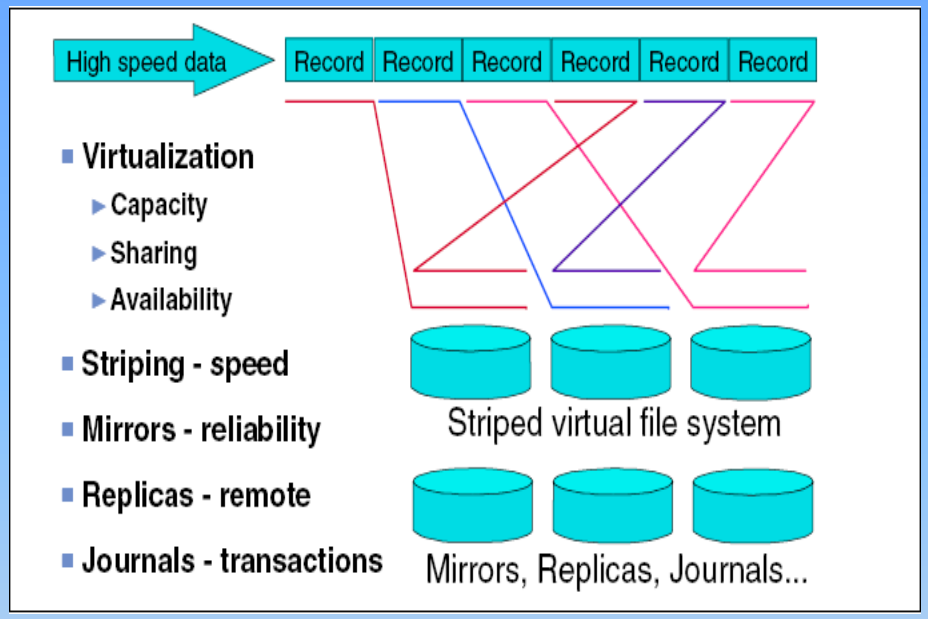
What grid computing can do



- **Reliability**
  - Now: redundancy in hardware
  - Future: Software

- **Management**
  - More disperse IT infrastructure
  - Priority among projects

What grid computing can do

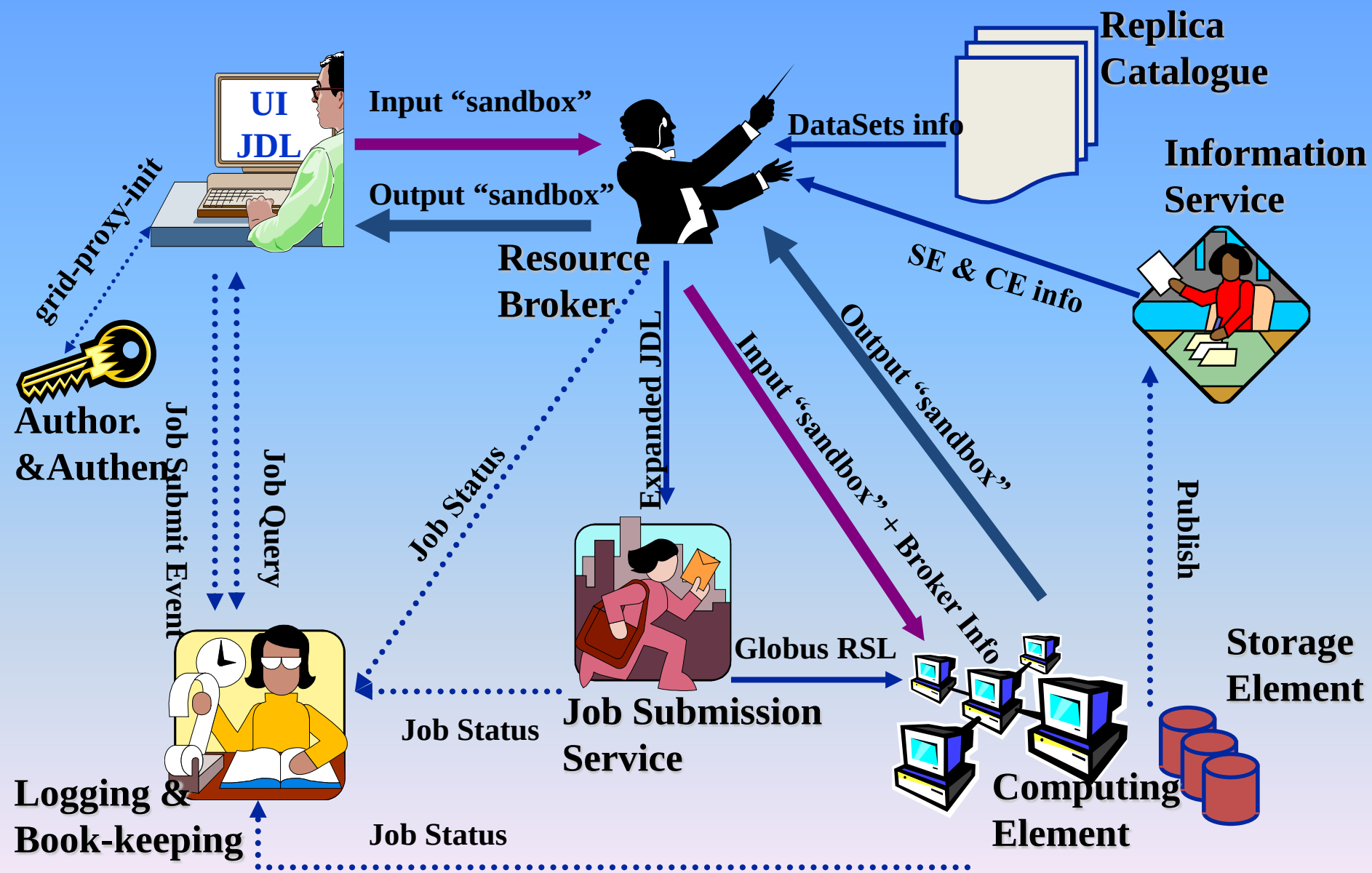


- Computation
- Storage
  - Primary/secondary storage
  - Mountable networked filed system
    - AFS, NFS, DFS, GPFS
  - Capacity increase
  - Uniform name space
  - Data Stripping

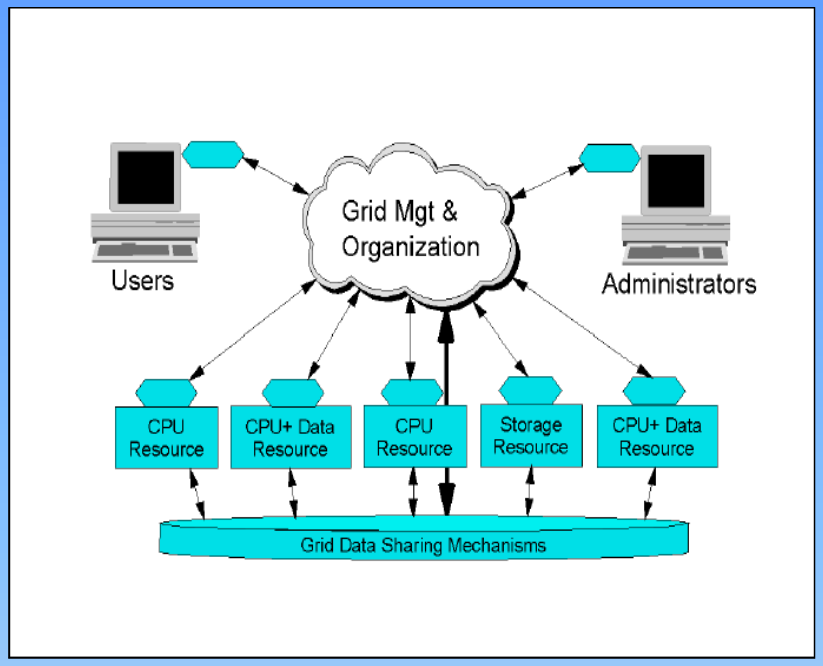
Grid concepts and components

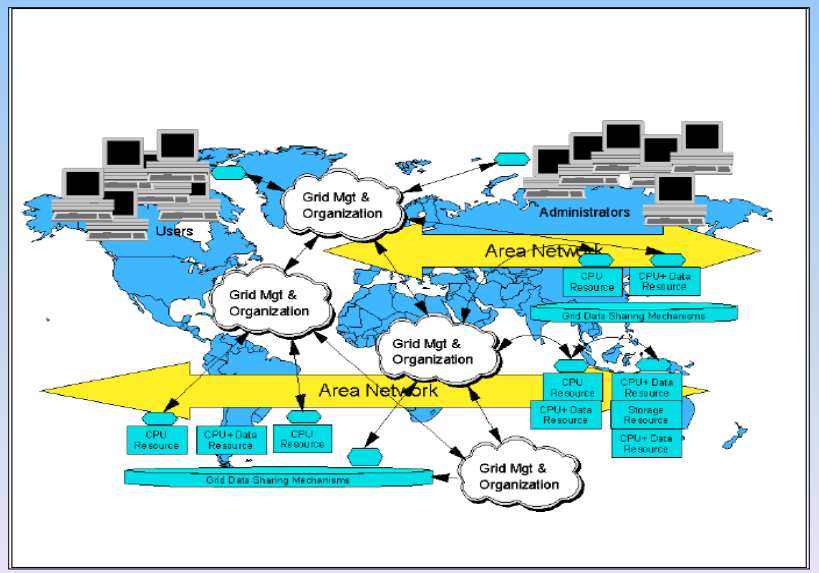
Types of resources

The lifecycle of an EGEE job



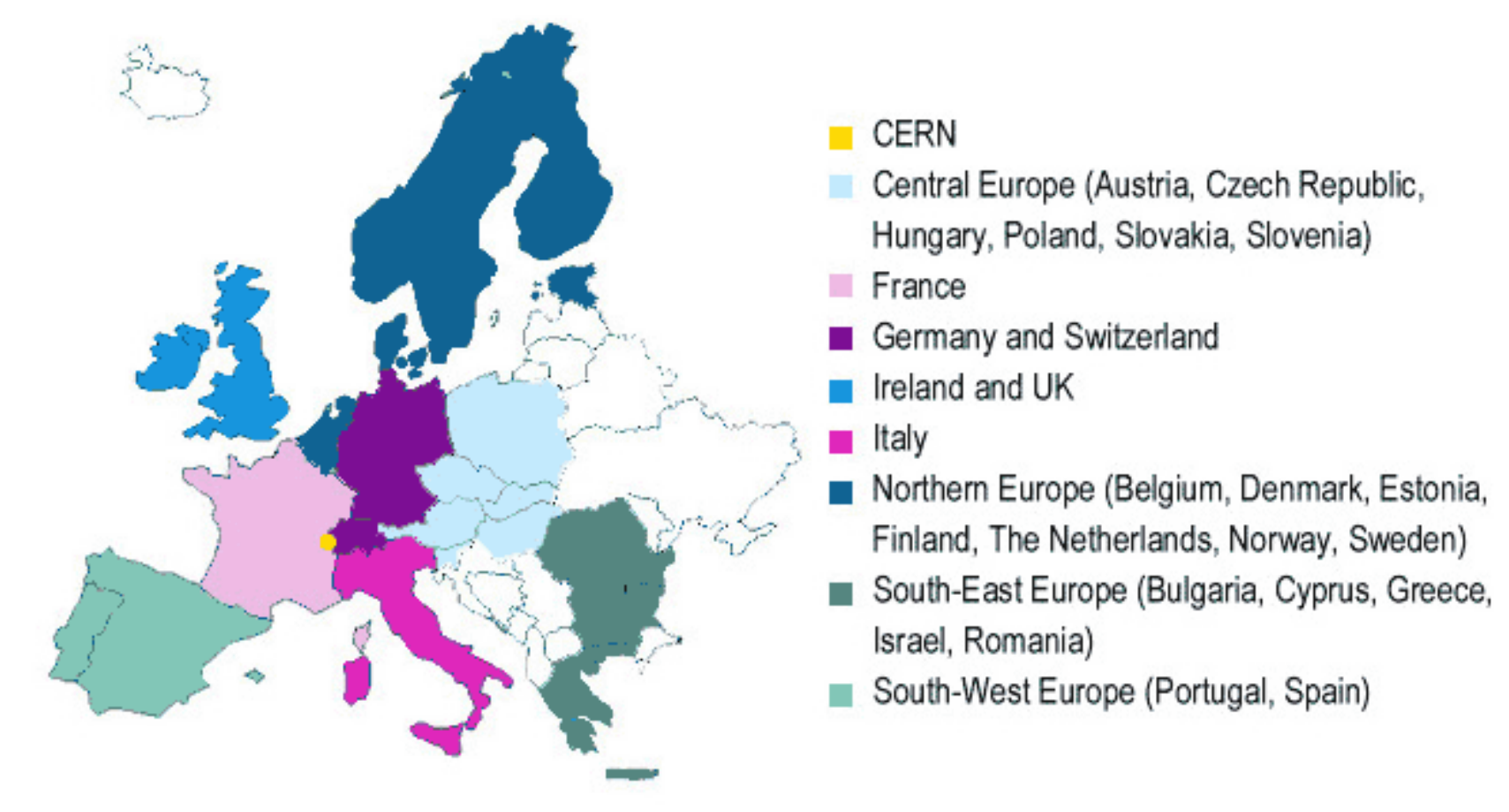
- **Intragrid to Intergrid**
  - **cluster**
    - **same hardware/software**
  - **Intragrid**
    - **heterogeneous machines/software**
    - **multiple department/same organization**
  - **Intergrid**
    - **heterogeneous machines/software**
    - **multiple department/multiple organization**



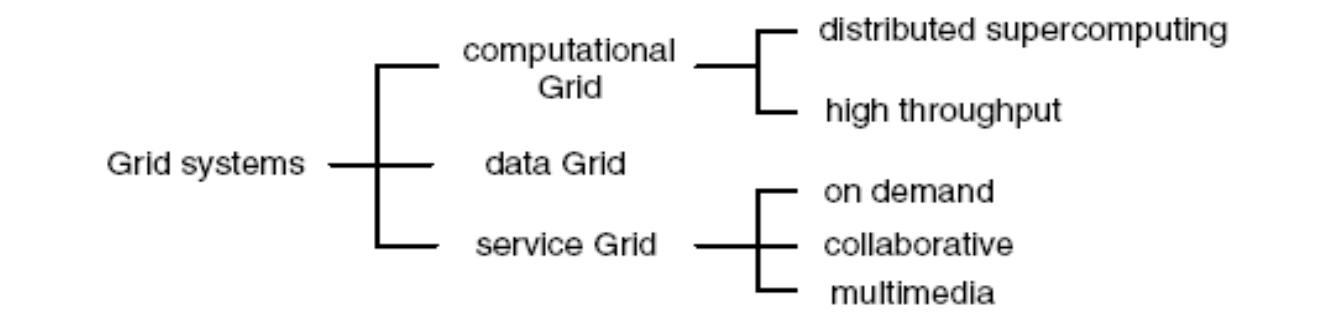


Grid concepts and components

Total of 70 full partners covering entire EU and beyond
Total budget: ~32 M€



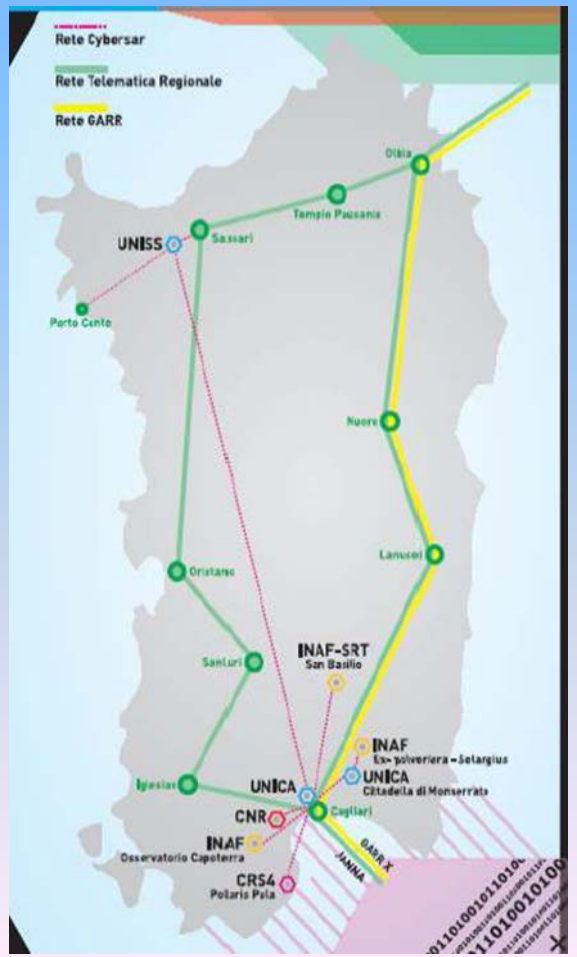
Enabling Grids for E-sciencE (EGEE) Consortium



- CERN's new particle accelerator
  - 15 petabytes(15 million gigabytes) a year
    - stack of CDs more than 20 km high!!!
  - 200 sites around the globe
  - Over 20 000 computers
  - Runing up to 30 000 jobs per day
- Has already served for:
  - 300 000 chemical compounds in search of potential drugs for Flu
  - Simulations of over 40 million potential drug molecules against malaria

Enabling Grids for E-sciencE (EGEE)

Project CYBERSAR - COSMOLAB supercomputing, computational modelling and management of larges database National Operational Program 2007-2013



Services:

- **HPC**
- **GRID (EGEE)**

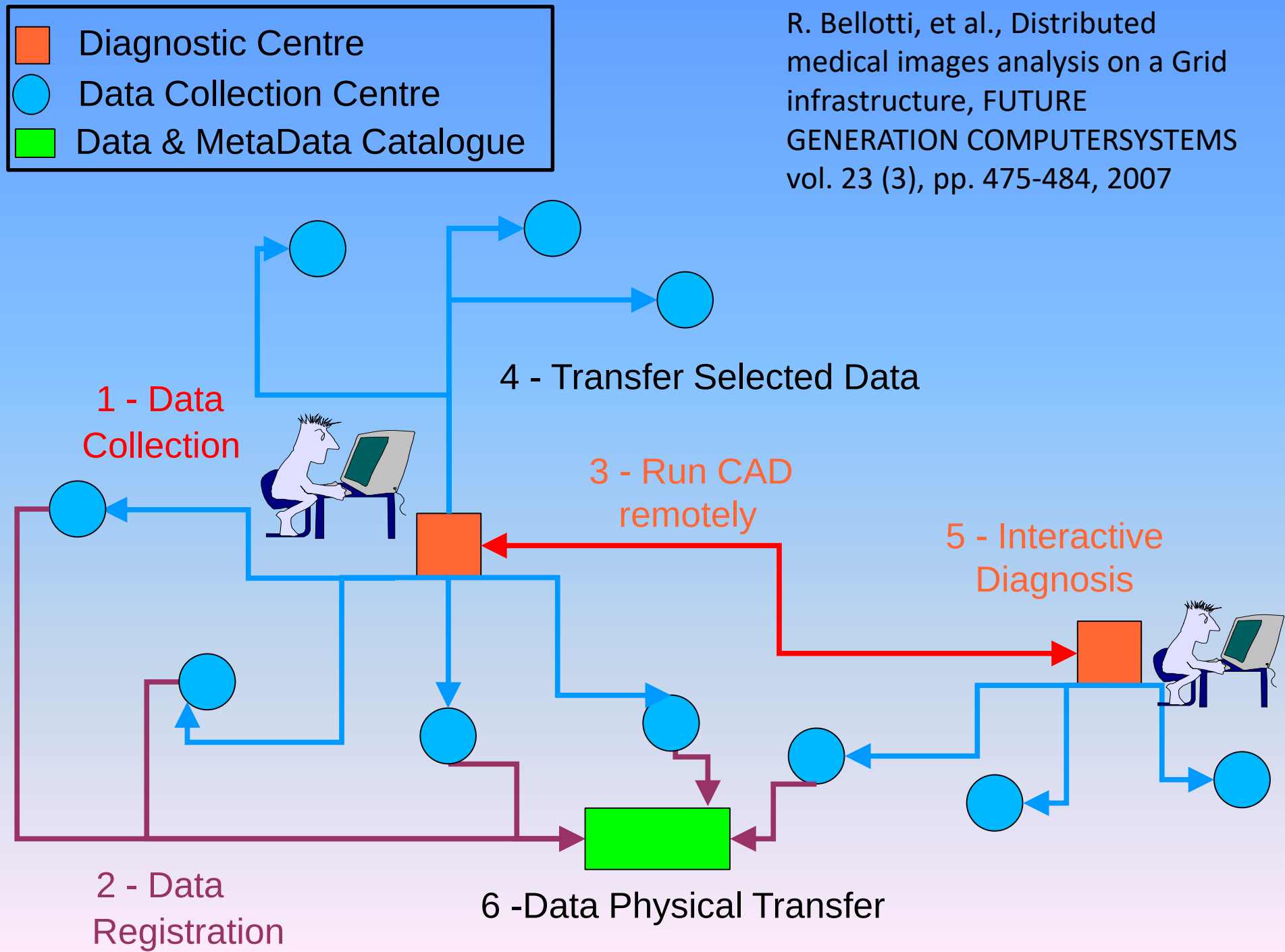
Cybersar GRID (320)

Cybersar-SS GRID (50)

CYBERSAR is an italian consortium for supercomputing, computational modelling and management

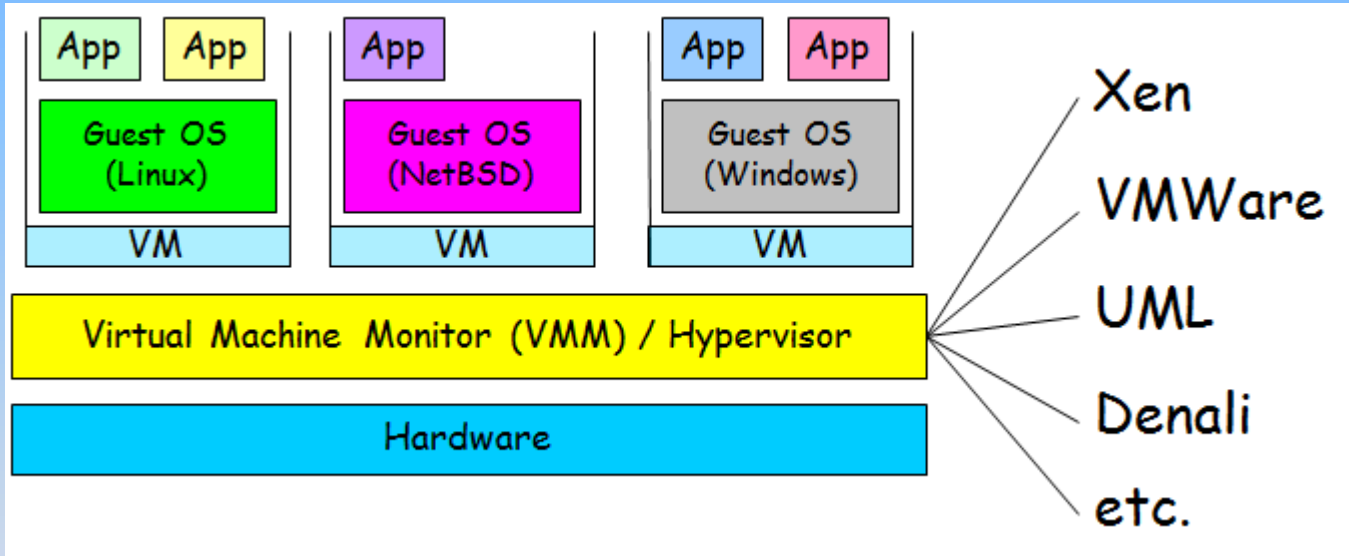
GRID in Sardinia (Italy)

ISI Italian Grid → EGEE

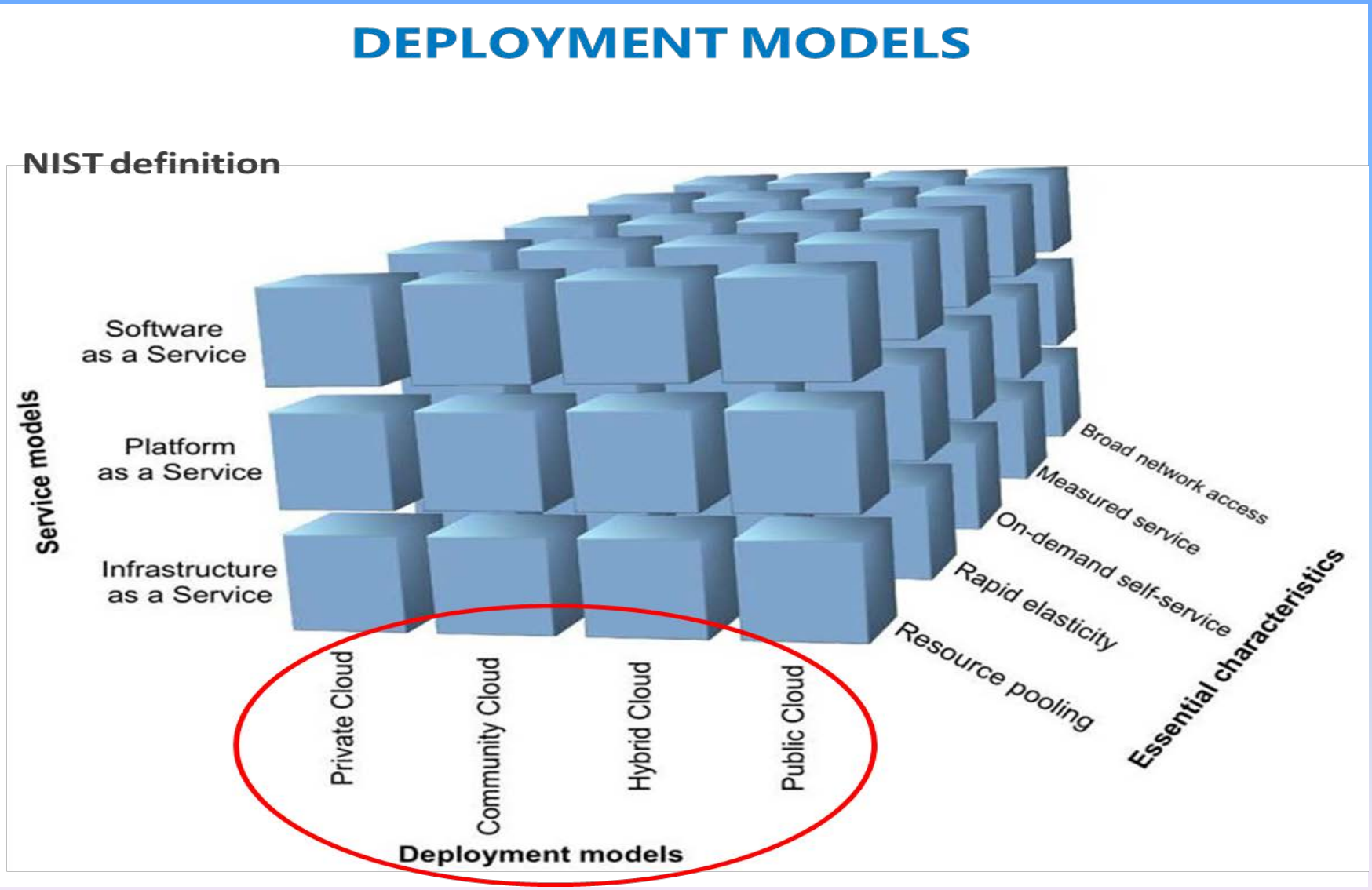

Diagnostic Centre
Data Collection Centre
Data & MetaData Catalogue
R. Bellotti, et al., Distributed medical images analysis on a Grid infrastructure, FUTURE GENERATION COMPUTERSYSTEMS vol. 23 (3), pp. 475-484, 2007
4 - Transfer Selected Data
1 - Data Collection
3 - Run CAD remotely
5 - Interactive Diagnosis
2 - Data Registration
6 -Data Physical Transfer

Virtualization examples

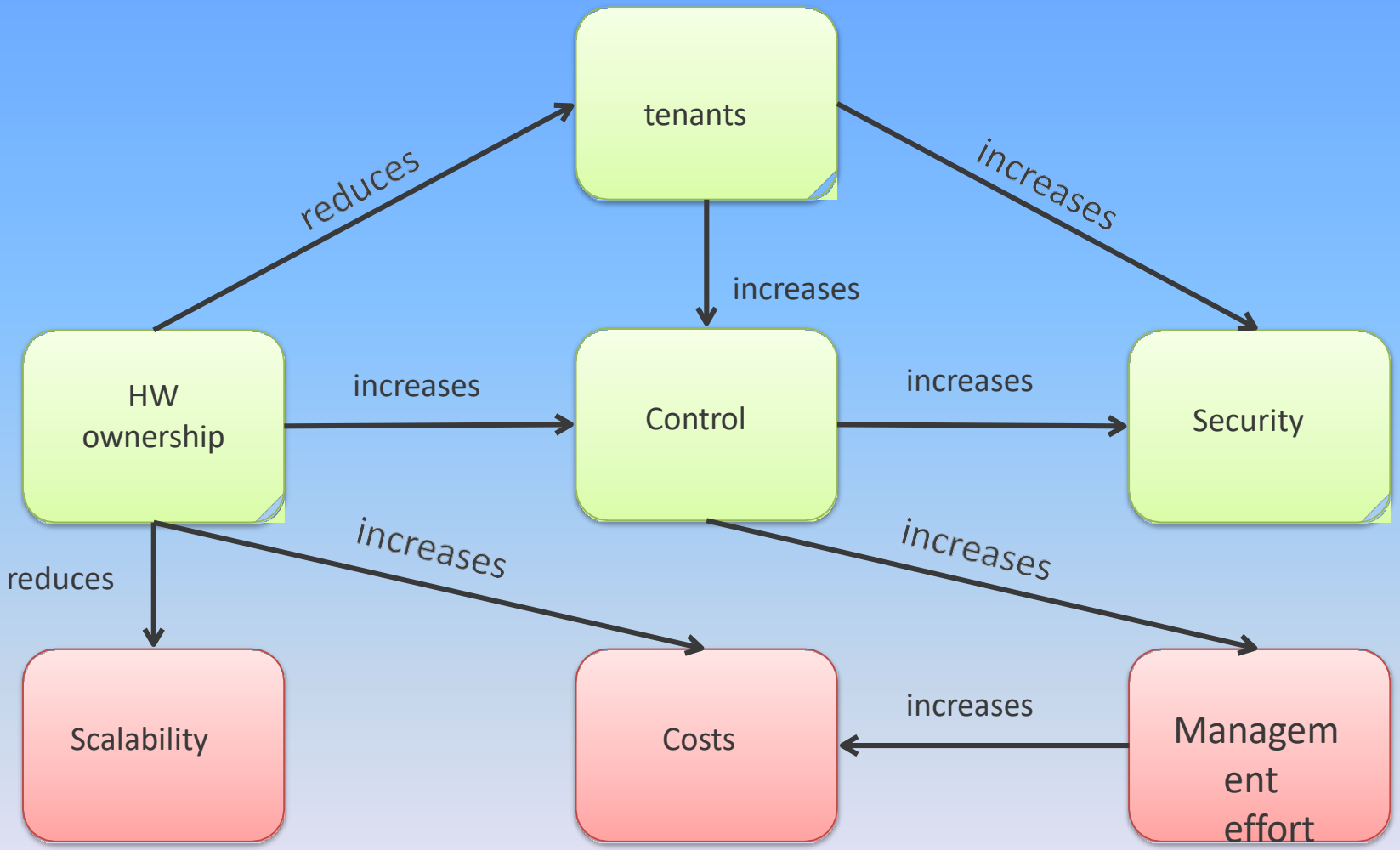
VM technology allows multiple virtual machines to run on a single physical machine

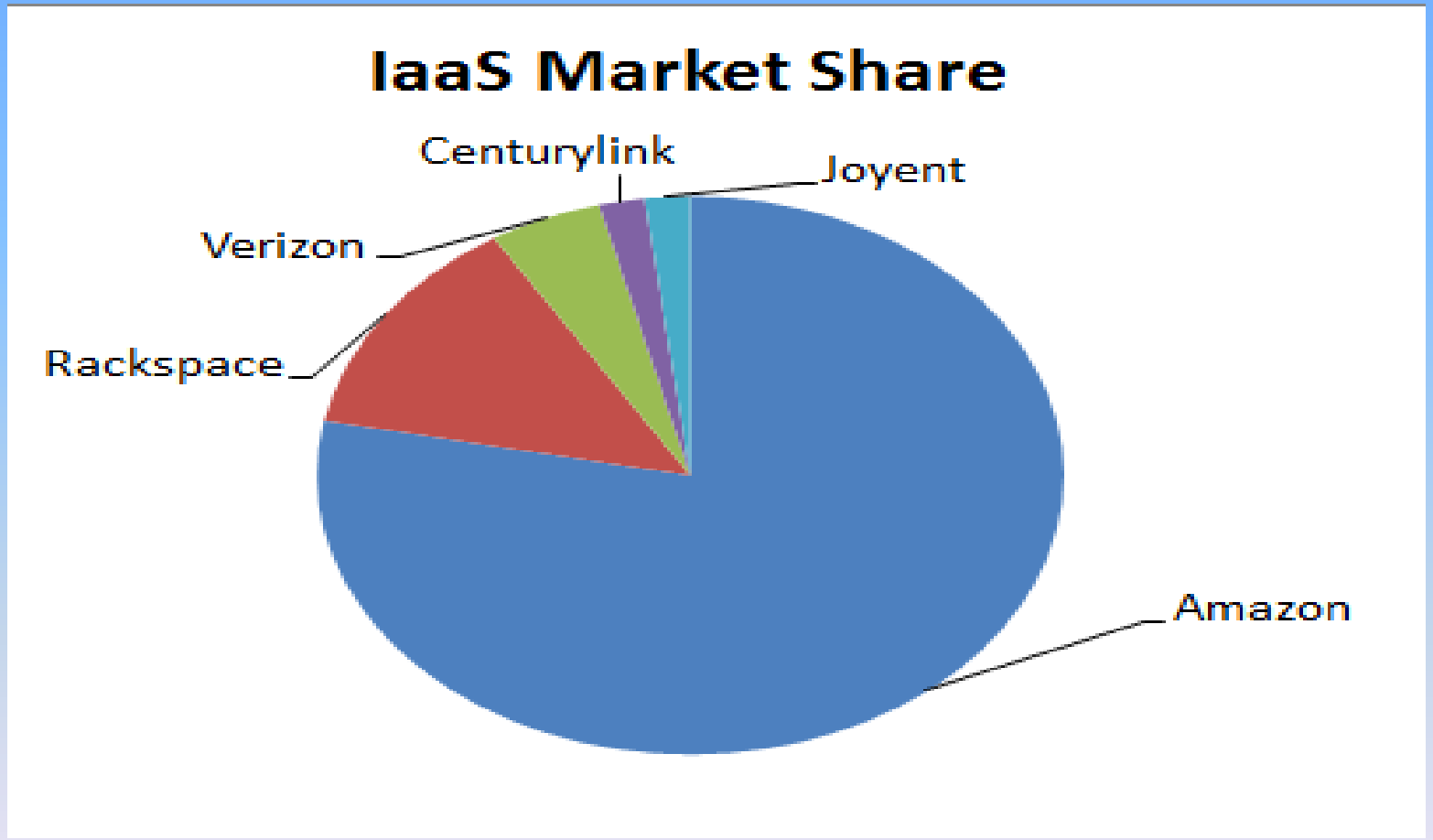


Cloud definition



PRIVATE VS PUBLIC




IaaS Market Share
Centurylink
Joyent
Verizon
Rackspace
Amazon

Cloud Characteristics

- On demand provisioning - you don't need to phone someone up to arrange a VM for next week
- Customer has no knowledge of location of resources - except at a very coarse level: "eu-west"
- Resource pooling
- Rapid elasticity - both up and down

SaaS

- Use applications running on cloud infrastructure
- Google Docs provides common business applications online
- Salesforce.com provide Customer Relationship Management software following this model

PaaS

Delivery of a virtualized application runtime platform that has a software stack for developing applications or application services.

PaaS applications and infrastructure are run and managed by the services vendor.

- e.g. platform provides databases or support to bill customers in many countries
- Windows Azure, Google App Engine

https://azure.microsoft.com

http://www.slideshare.net/MiguelFierro1/leveraging-data-driven-research-through-microsoft-azure-71465022

IaaS

- Provide access to raw computing systems and the customer can run arbitrary software
- Amazon played a key role in the development of cloud computing by providing external access to their data systems by way of Amazon Web Services on a utility computing basis.
- Provides dynamical scaling facilities
- Amazon EC2, Rackspace Openstack

Simple Comparison

	In-house Server	Cloud Server
Cost/hr (over 3 yrs)	$0.37	$0.45

- Seems that the cloud server is more expensive!

Include Efficiency

	In-house Server	Cloud Server
Cost/hr (3 yrs)	$0.37	$0.45
Efficiency	40%	80%
Cost/effective hr	$0.92	$0.56

- In house server is under-utilised: say 40% Cloud
- server can be scaled: utilisation say 80% Now the
- cloud server is cheaper

Include Other Costs

	In-house Server	Cloud Server
Eff. Cost/hr (3yrs)	$0.92	$0.56
Power&cooling	$0.37	
Management	$0.10	$0.01
Total	$1.39	$0.57

- Power and cooling at least as much as amortised purchase cost
- Management more for in-house A
- significant difference

Cloud Security

- Technical level of security typically improves due to centralization of data, increased security-focused resources, etc.
- But customers are concerned about loss of control over sensitive data - related to legal standards for securing personal data

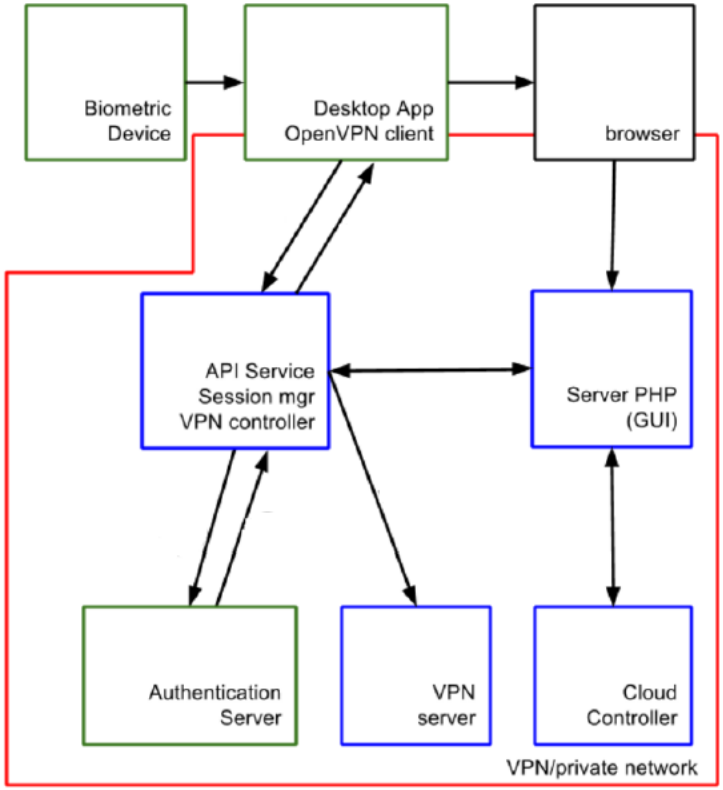
Issues: Security

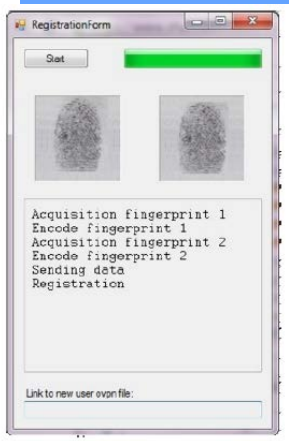
- Privileged user access—who has access to data and how are such administrators hired and managed?
- Regulatory compliance—is the cloud vendor willing to undergo external audits and/or security certifications?
- Data location—do I retain any control over the location of data? Data segregation—is encryption used at all stages and is the security protocol designed and tested by professionals?
- Recovery—what happens to data in the case of a disaster; do they offer complete restoration and, if so, how long does it take
- Investigative Support—does the cloud vendor have the ability to investigate any inappropriate or illegal activity
- Long-term viability—what will happen to data if the company goes out of business; how will data be returned?

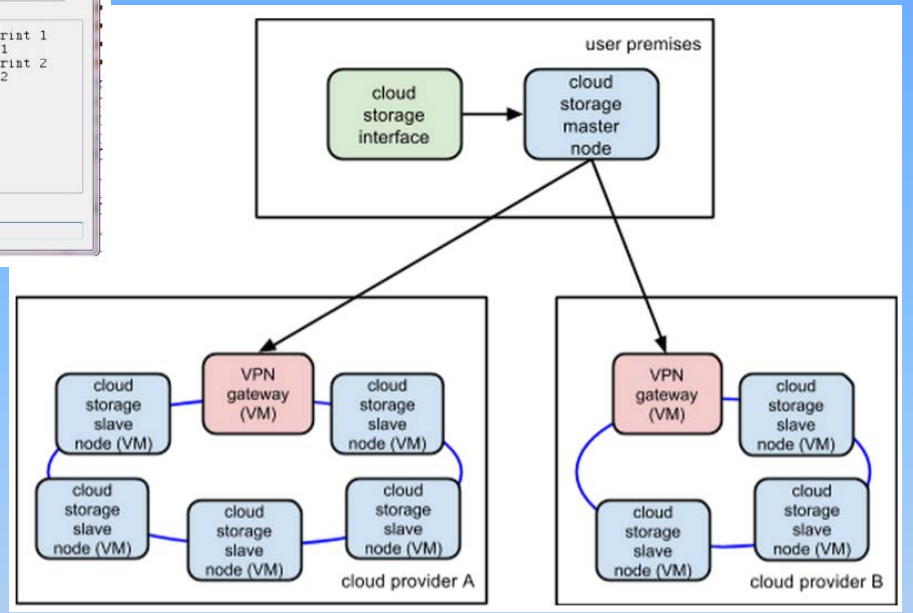
Issues: Legal Jurisdiction

- The cloud spans many borders and is subject to complex geopolitical issues: providers must satisfy a myriad of regulatory environments in order to deliver service to a global market.
- Despite efforts (such as US-EU Safe Harbour) to harmonise the legal environment, providers like Amazon Web Services cater to the major markets by deploying local infrastructure and allowing customers to select "availability zones."

CLOUD SECURITY







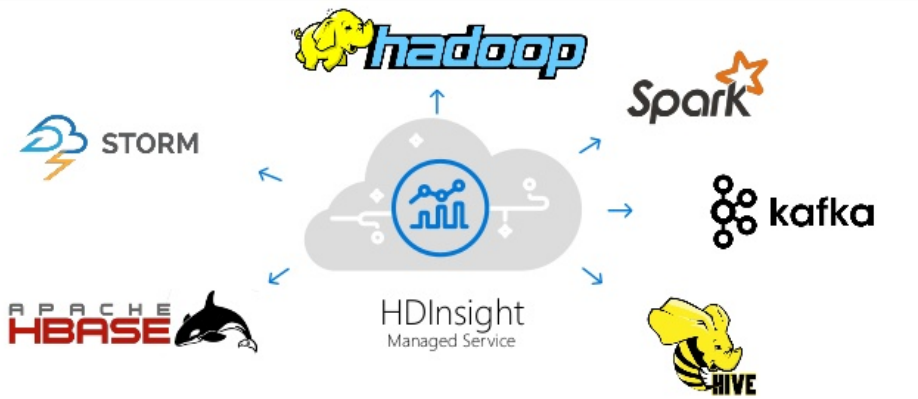
DIGITAL IDENTITY: Integration of biometric recognition with cloud platform (*OPENSTACK*) using a client desktop on the user side and a dedicated authentication server connected to the Keystone module.

DATA SECURITY: Novel data chunking solution based on innovative distributed cloud storage architecture (*MongoDB*). The basic idea is to share data in small chunks and spread them on different VMs hosted on cloud computing.

Microsoft Azure

WHAT IS HDINSIGHT

Microsoft

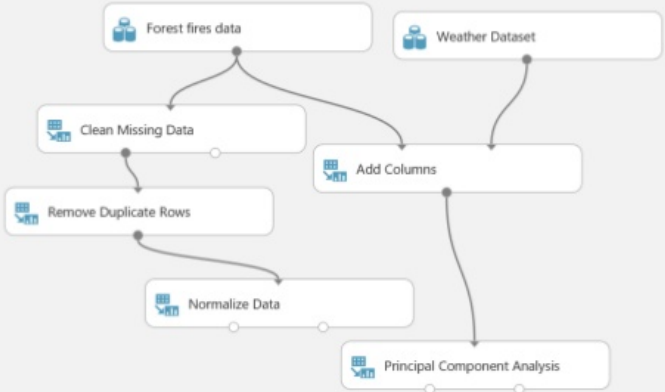


PYTHON & R SCRIPTS

Microsoft

DATA MANIPULATION

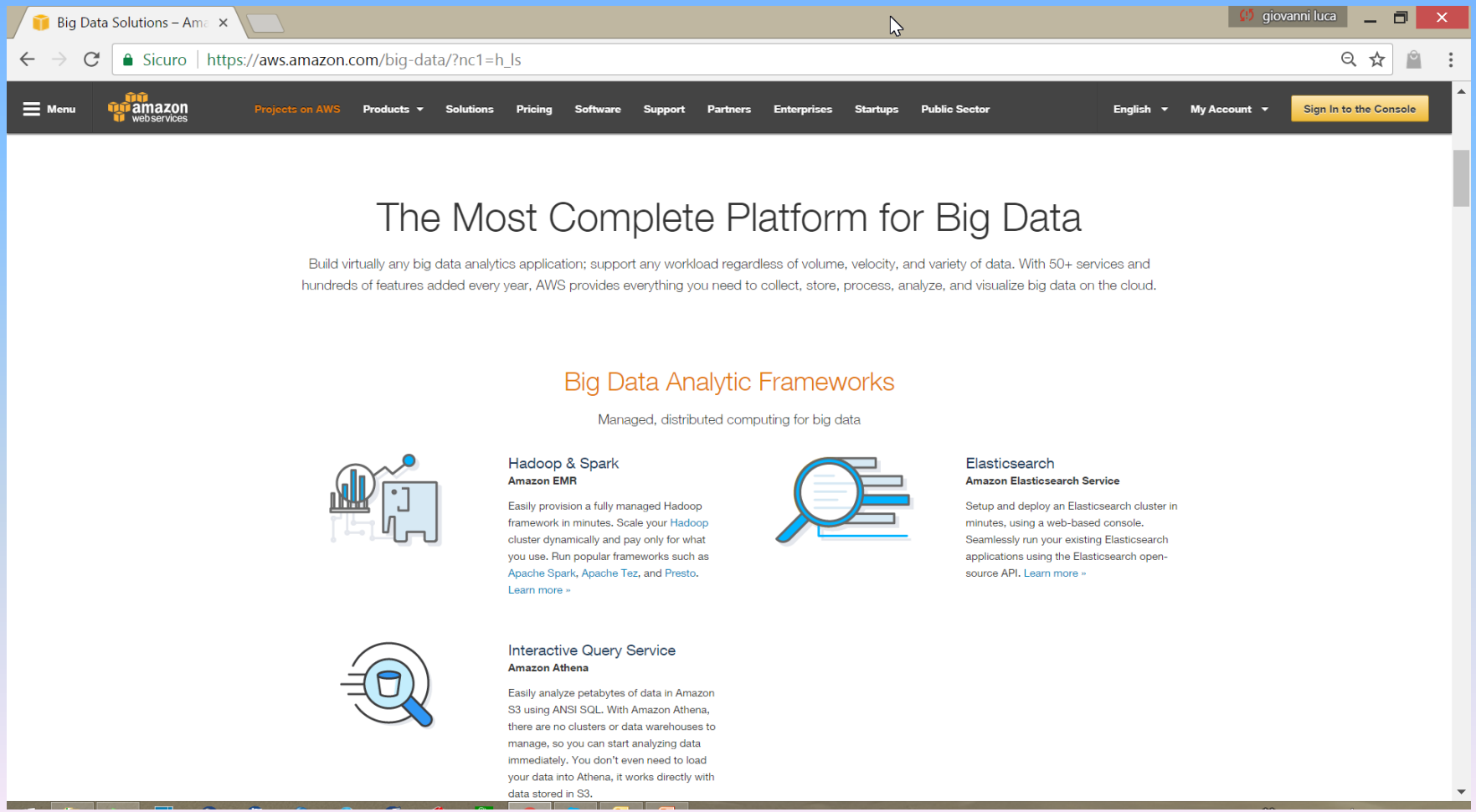
Microsoft



CLASSIFICATION & REGRESSION

Microsoft

Amazon AWS


Big Data Solutions – Ama
giovanni luca
Sicuro
https://aws.amazon.com/big-data/?nc1=h_ls
Menu
amazon web services
Projects on AWS
Products
Solutions
Pricing
Software
Support
Partners
Enterprises
Startups
Public Sector
English
My Account
Sign In to the Console
The Most Complete Platform for Big Data
Build virtually any big data analytics application; support any workload regardless of volume, velocity, and variety of data. With 50+ services and hundreds of features added every year, AWS provides everything you need to collect, store, process, analyze, and visualize big data on the cloud.
Big Data Analytic Frameworks
Managed, distributed computing for big data
Hadoop & Spark
Amazon EMR
Easily provision a fully managed Hadoop framework in minutes. Scale your Hadoop cluster dynamically and pay only for what you use. Run popular frameworks such as Apache Spark, Apache Tez, and Presto. Learn more »
Elasticsearch
Amazon Elasticsearch Service
Setup and deploy an Elasticsearch cluster in minutes, using a web-based console. Seamlessly run your existing Elasticsearch applications using the Elasticsearch open-source API. Learn more »
Interactive Query Service
Amazon Athena
Easily analyze petabytes of data in Amazon S3 using ANSI SQL. With Amazon Athena, there are no clusters or data warehouses to manage, so you can start analyzing data immediately. You don't even need to load your data into Athena, it works directly with data stored in S3.

Apache Hadoop

- A framework for processing large data sets across clusters (Big data analysis).
- High availability and potential scaling
- Move the computation to the data - Hadoop relies on:
  - The distributed cluster file system HDFS
  - The programming paradigm “mapreduce”

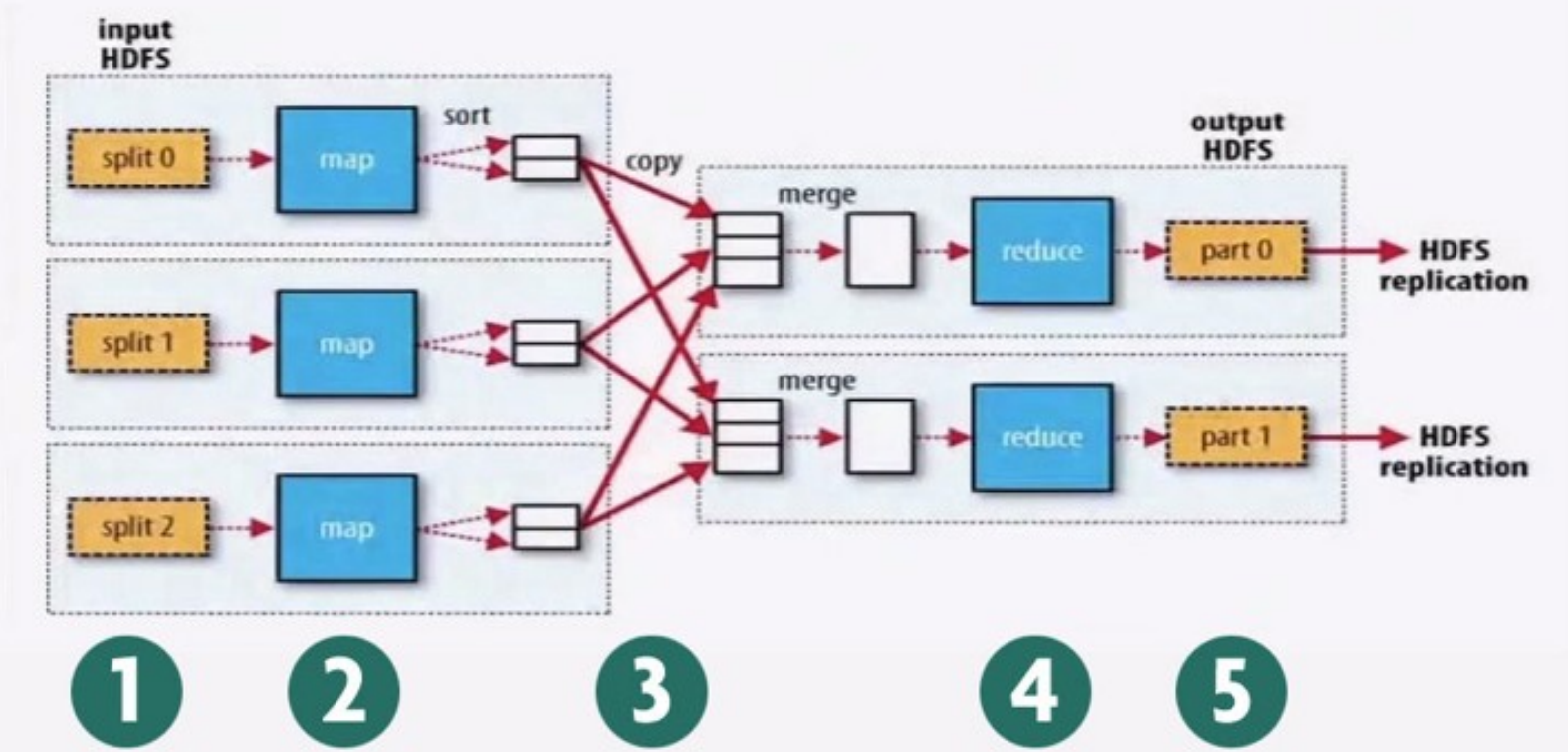
HDFS

https://www-01.ibm.com/software/data/infosphere/hadoop/hdfs/

- Special Purpose cluster file system - not POSIX and not part of the OS
- For local storage associated with each server
- Compare with Google File System GFS (library)
- For big (multi-GB) files that will be read but rarely written
- Replicated Data Blocks - designed to be fault resilient on commodity hardware
- Provides high throughput to processing system

MapReduce steps

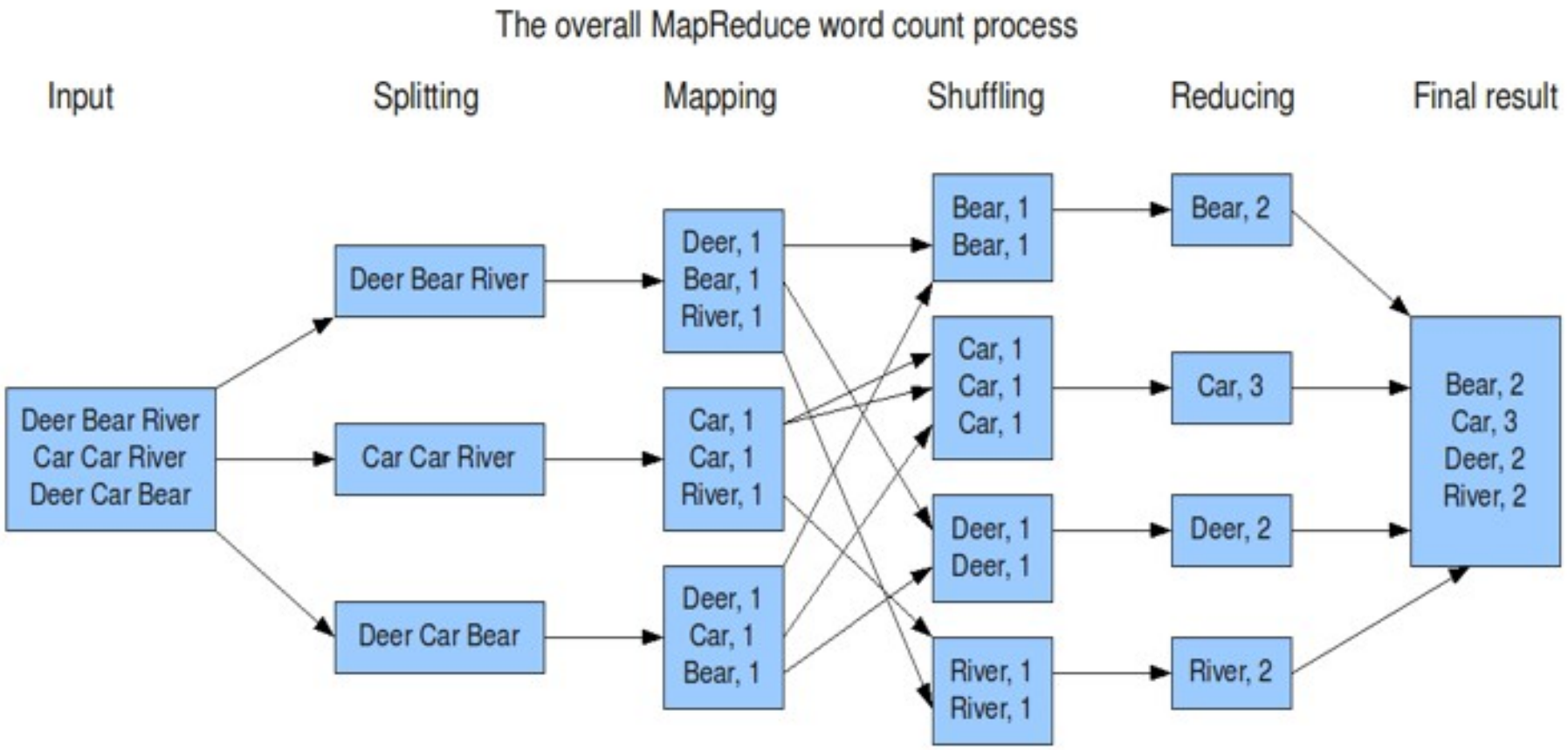
1. Input data split on HDFS
2. Map independently on each node
3. Shuffle
4. Reduce on reduce workers
5. Output to HDFS



WordCount example

- Given a text document
- Map reads each word and outputs `(word, 1)`
- Reduce is called for each word, and adds up the `1's`
- Gives the number of times each word appears in the document

MapReduce for Word Count



Some topics for the discussion

What is the size of BIG data to obtain correct prevision?

What the Latest Nate Silver Controversy Teaches Us About Big Data
All the data in the world can't eliminate uncertainty—yet.

http://fortune.com/2016/11/06/nate-silver-controversy-big-data/

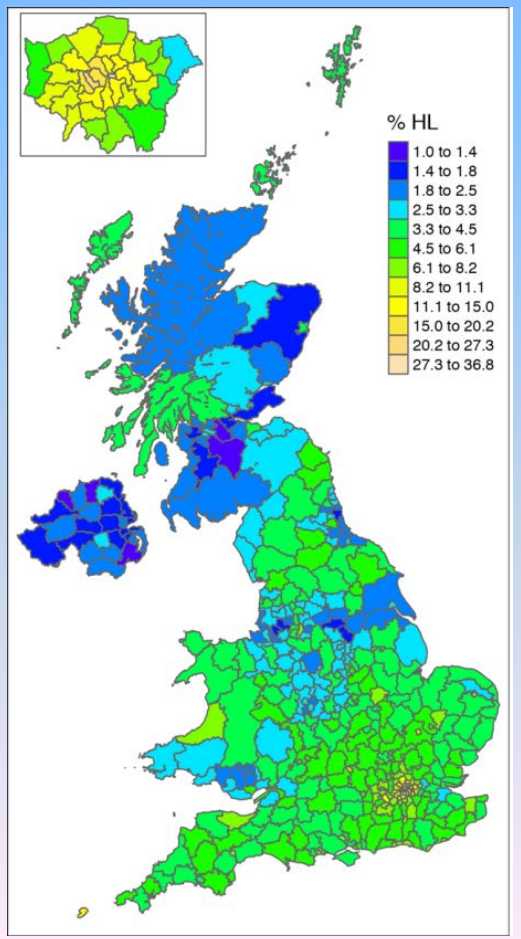
Are we able to predict complex events?

Biological Sciences - Ecology:
Kevin M. Bakker, Micaela Elvira Martinez-Bakker, Barbara Helm, and Tyler J. Stevenson *Digital epidemiology reveals global childhood disease seasonality and the effects of immunization*
PNAS 2016 113 (24) 6689-6694; published ahead of print May 31, 2016, doi:10.1073/pnas.1523941113

Big Data Science and Foundations

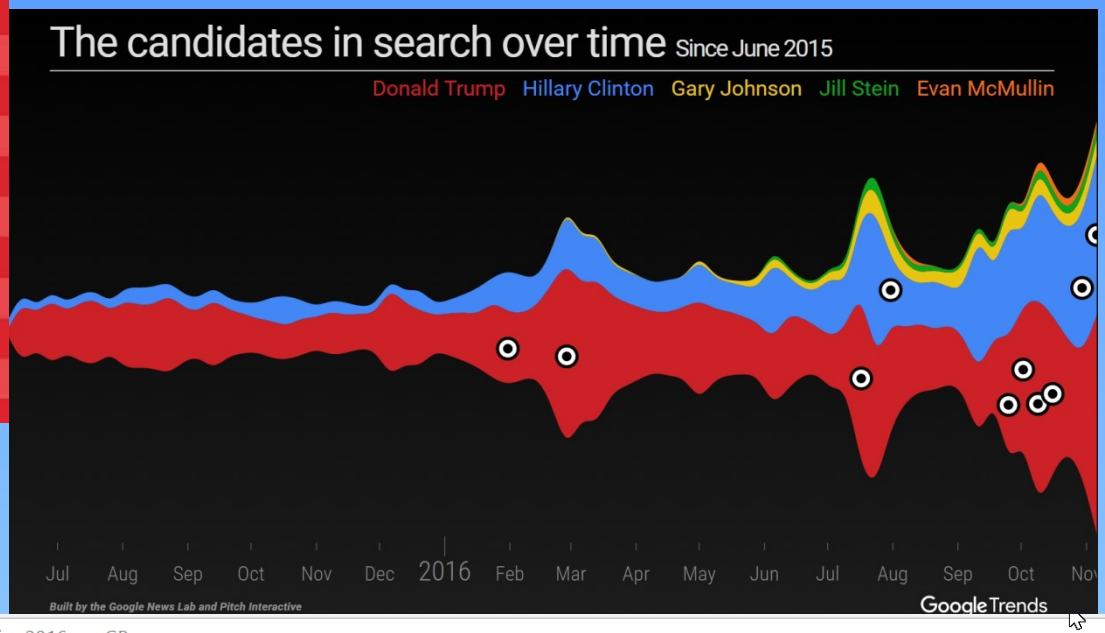
Social networks and sentiment analysis:

-Facebook
-Twitter
-Google trends



Cortina Borja, M., Stander, J. and Dalla Valle, L. *The EU referendum: surname diversity and voting patterns*, **SIGNIFICANCE**, August 2016

https://www.google.co.uk/trends/story/election2016_en-GB



https://www.google.co.uk/trends/story/election2016_en-GB

Google Trends

Cerca tra gli argomenti

Top UK questions on the US election as Trump wins US election

1. Who won the US election?
2. When is the US election?
3. Is Donald Trump president?
4. When is the next US election?
5. How old is Donald Trump?

UK search interest in Clinton and Trump

Hillary Clinton Donald Trump

100 75 50 25

3 nov alle... 8 nov alle ore 17:00

Top UK questions on Hillary Clinton as Trump wins US election

1. How old is Hillary Clinton?
2. Who is winning, Trump or Clinton?
3. Who won, Trump or Clinton?
4. Where is Hillary Clinton?
5. Why don't people like Hillary Clinton?

Top UK questions on Donald Trump as he wins the US election

1. How old is Donald Trump?
2. Is Donald Trump president?
3. What does Trump Pence mean?
4. Who is Donald Trump?
5. When does Trump become president?

Sentiment Analysis

- We can download **unstructured text data** from *Facebook* and *Twitter*
- We can summarize word frequencies by producing **Word Clouds**
- We can perform **Sentiment Analyses**
- We have published light articles using data from **Facebook** about
  - the EU Referendum
  - Brexit
- We'll now see a example from **Twitter**, run on Tuesday, 15 November 2016, which we'll develop using Facebook data

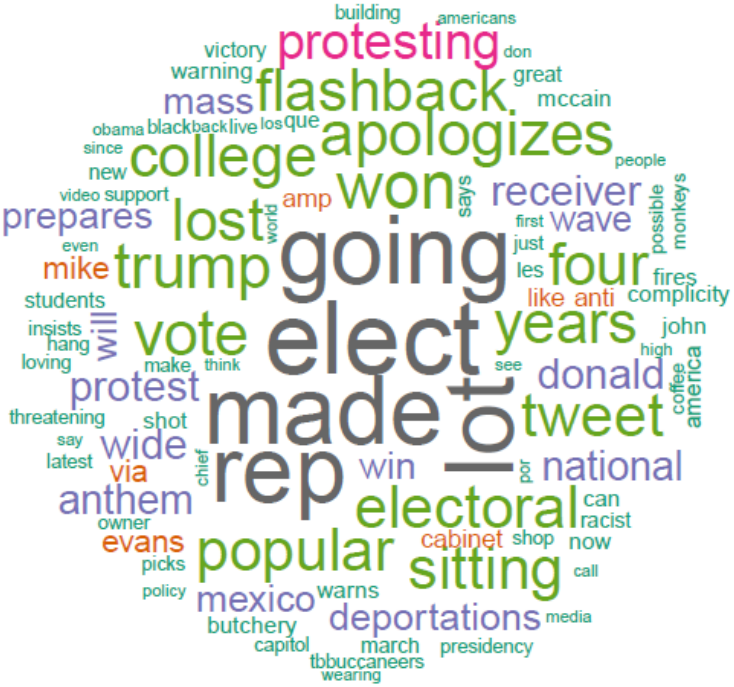
Sentiment Analysis (US election)

Stander, J., Dalla Valle, L. and Contina Borja, M. (2016)
Sentiments, surnames and so long EU.
Communicator Autumn 2016 - Special Supplement Science Communication. Pages 19-23.

- We downloaded, cleaned and processed 5000 tweets containing
  - `#HilaryClinton`, and separately
  - `#DonaldTrump`
- We created a **Corpus**
  - This is similar to a *book* (all tweets)
  - with *chapters* (individual tweets)
- We created a **Document Term Matrix**
  - The entries of this matrix tell us how many times each word appears in each tweet
- From this we can calculate **word counts**

Word Cloud for Tweets Containing #HilaryClinton

Word Cloud for Tweets Containing #DonaldTrump

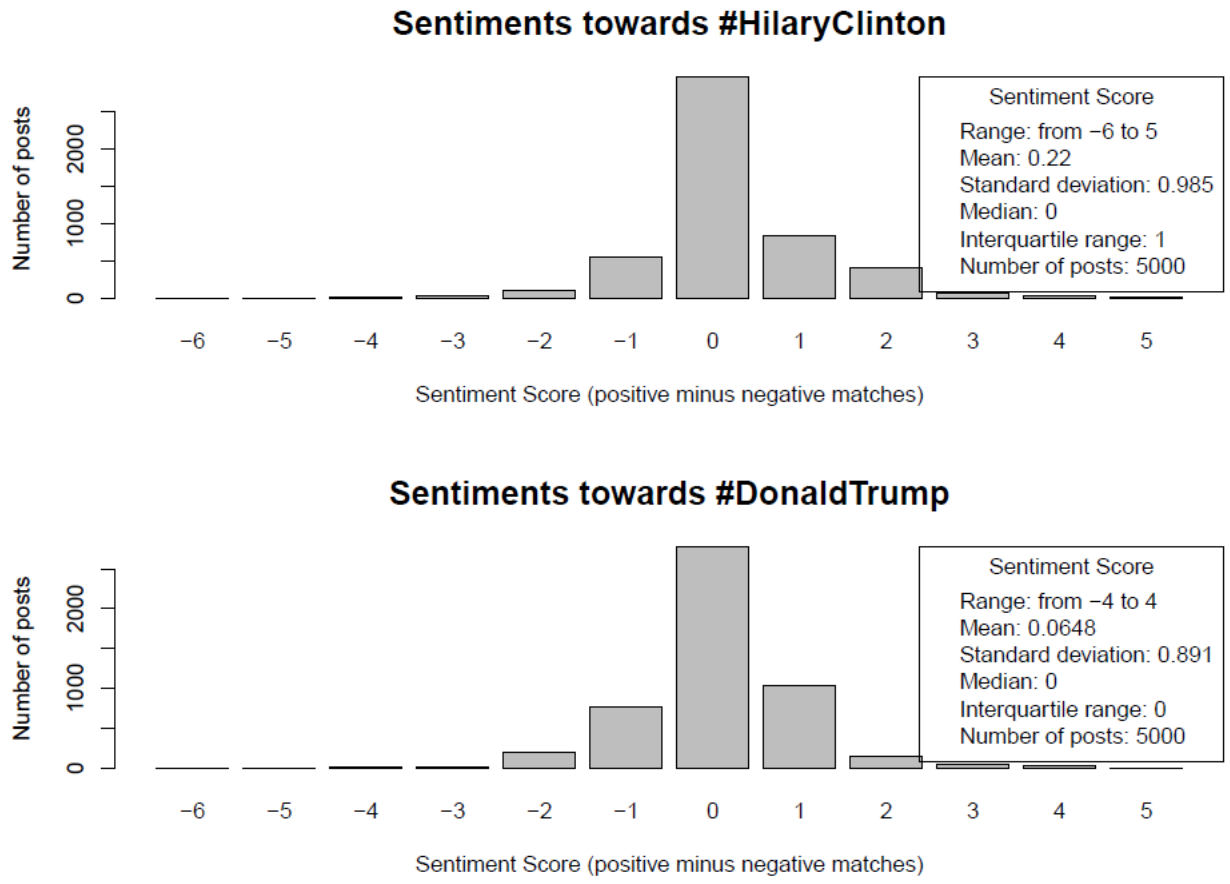


Assessing Posts

- The basic task in Sentiment Analysis is to determine whether opinions expressed are **positive**, **negative** or **neutral**
- Our approach is based on matching words in posts to dictionaries of **positive** and **negative** words supplied by Cheng, Hu and Liu
- The **sentiment score** is

  the number of positive matches – the number of negative matches

Graphs of the Sentiment Scores

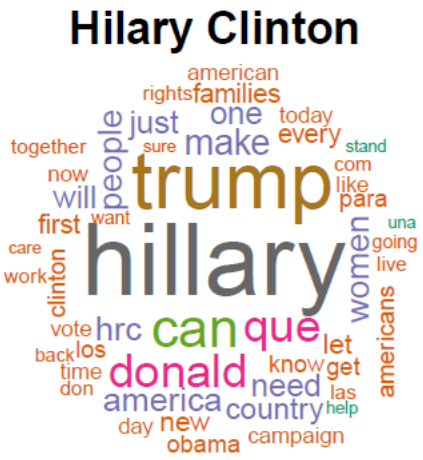


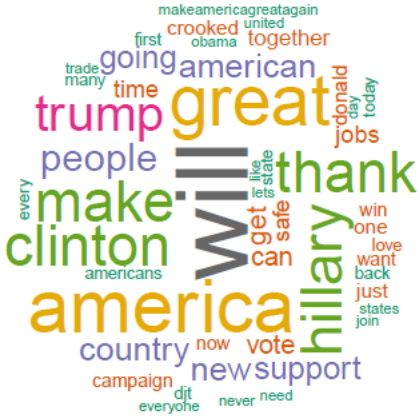
- Similar, with quite a lot of negativity

Pages Analyzed

- We now consider the **main Facebook pages** for Hilary Clinton and Donald Trump
  - `hillaryclinton`
  - `DonaldTrump`
- The difference with Twitter is that Clinton and Trump essentially control what is posted on these Facebook pages

Word Clouds





- Please remember that Clinton and Trump essentially control what is posted on these pages

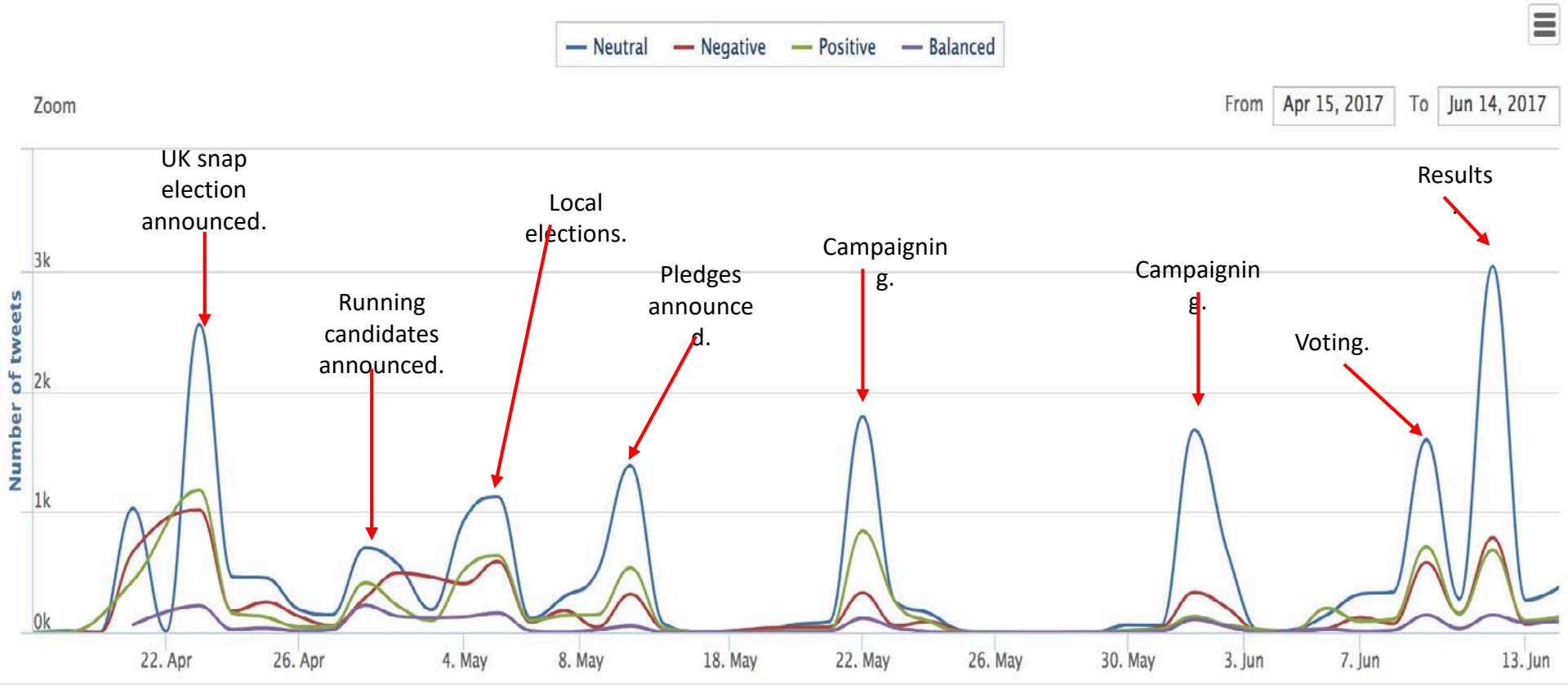
Example Tweets


```
_id: ObjectID('58f9e7eff5e7659eafe09e88')
simplifiedDate: "21-04-2017"
tweetDate: "21-04-2017 12:07:01"
tweetText: "He isn't standing for the UK so we don't need to worry. The truth is his passion brought us to this won
          derful Brexi… https://t.co/Tom3ajUc8c"
sentimentFound: "passion,wonderful"
overallSentiment: 2
tweetPolarity: "Positive"
```

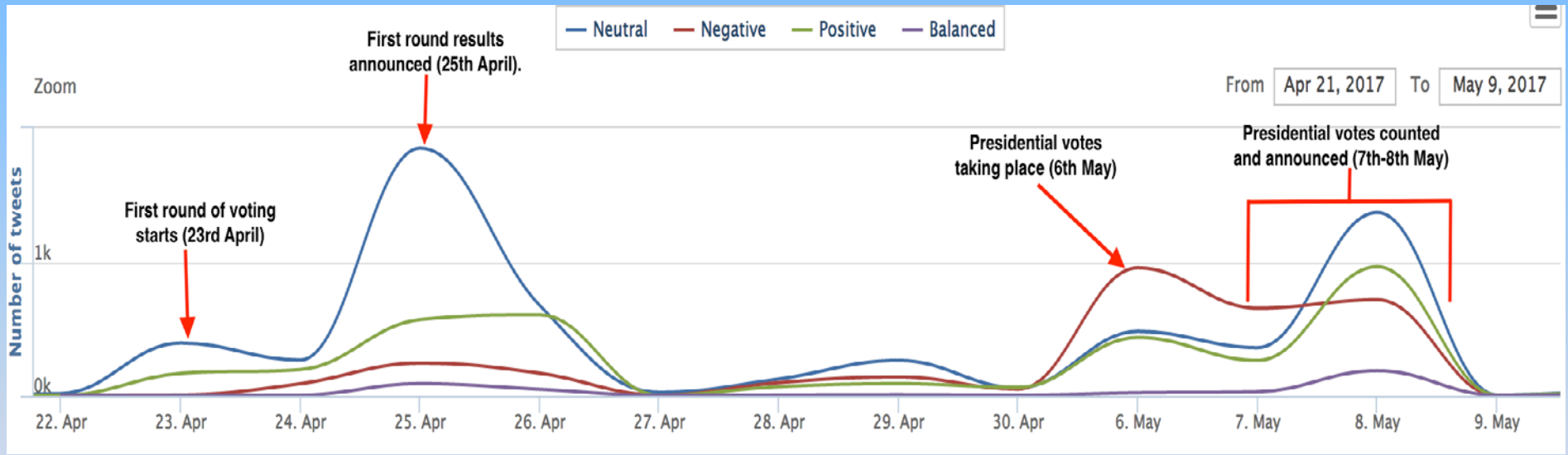
```
_id: ObjectID('58f9e4faf5e7659eafe09c54')
simplifiedDate: "21-04-2017"
tweetDate: "21-04-2017 11:54:48"
tweetText: "Just as false but even more disastrous for far more people should she be elected and the EU fall. Brexi
          t was just t… https://t.co/IuttirsFmk"
sentimentFound: "false,disastrous"
overallSentiment: -2
tweetPolarity: "Negative"
```

```
_id: ObjectID('58f9e6c4f5e7659eafe09d9e')
simplifiedDate: "21-04-2017"
tweetDate: "21-04-2017 12:02:00"
tweetText: "UKIP is in disarray but Nigel could win a seat if the 'establishment' played fair. That said he can do
          much in Euro… https://t.co/g4S4jMzmRH"
sentimentFound: "win,disarray"
overallSentiment: 0
tweetPolarity: "Balanced"
```

Results (#UKGeneralElection)



Results (#FrenchElection)



MASALA G.L.C, Ruiu P, and Grosso E, Biometric Authentication and DataSecurity in Cloud Computing, Computer and Network Security Essentials,Springer, in press 2017