UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Challenges and Opportunities of a Data Driven Culture

Federmanager Bologna 22 Maggio 2019

Prof. Sonia Bergamaschi
Department of Engineering "Enzo Ferrari"
sonia.bergamaschi@unimore.it
www.dbgroup.unimore.it

Outline

- Who am I?
- New Computer Engineering and Science Technologies: Big Data , Big Data Integration, Cyber Physical Systems, Cognitive Computing
- Occupational Challenges?

- The DBGROUP contribution
- Engineering UNIMORE departments contribution
- Emilia Romagna region contribution
- And the women?

DBGroup at Glance

Prof. Sonia Bergamaschi

Leader of the Database research group (DBGroup)

Dean of the ICT doctorate (www.ict.unimore.it)

- ACM distinguished researcher
- IEEE senior member
- Email: sonia.bergamaschi@unimore.it
- www.dbgroup.unimore.it

>200 publications in international conference and journals

DBLP

Google Scholar

Scopus

DBGroup at Glance

- Current Members:
  - 4 faculty
    - Sonia Bergamaschi
    - Domenico Beneventano
    - Francesco Guerra
    - Laura Po
  - 1 postdoc
    - Giovanni Simonini (IEEE best Computer Science phd thesis award 2017)
  - 4 ICT PhD students
    - Luca Magnotta (industrial phd DATARIVER on *Big Data Integration &Analysis* 3rd year)
    - Gagliardelli Luca (Emilia-Romagna phd scholarship on *Big Data Integration & Analysis* 3rd year)
    - Giuseppe Fiameni (CINECA –*Big Data Management* 3rd year)
    - Giovanni Morrone (*Cognitive Computing* phd at Doctorate School Industria 4.0 1st year)
- Member of the Italian **CINI Big Data Lab**
- 1 spin-off (now innovative SME) to deploy the MOMIS data integration system www.datariver.it

New Technological & Scientific Challenges

- ✓ Cyber Physical Systems (CPS)/ Internet of Things (IOT)
- ✓ Cloud Computing
- ✓ Big Data & Big Data Integration
- ✓ Cognitive Computing
- ✓ Blockchain

New Technological & Scientific Challenges

- A *Cyber Physical System (CPS)* is a system controlled or monitored by computer-based algorithms, tightly integrated with the Internet and its users..
  - ✓ physical components and software interact (autonomous driving systems, robotics systems, medical monitoring, etc.)
  - ✓ it is similar to the Internet of Things (*Internet of Things – IOT*)

- *Cloud Computing* (in italian **nuvola informatica**) indicates the provision of preexisting and configurable IT resources, such as archiving, processing or transmission of data, available on demand through the Internet.

- *Cognitive Computing* it is the technology that will allow us to interact with computers practically "talking" to machines and exploiting their ability to learn from experience (Artificial Intelligence, Machine Learning).

  Big Data Integration is a fundamental technique in this context.

Cyber Physical Systems (IOT)

- Internet of Things (IoT) merges physical and virtual worlds, creating smart environments. (IoT) represents the next step towards the digitisation of our society and economy, where objects and people are interconnected through communication networks and report about their status and/or the surrounding environment.

Cloud Computing: Scaling Up

PC

Server

Cluster

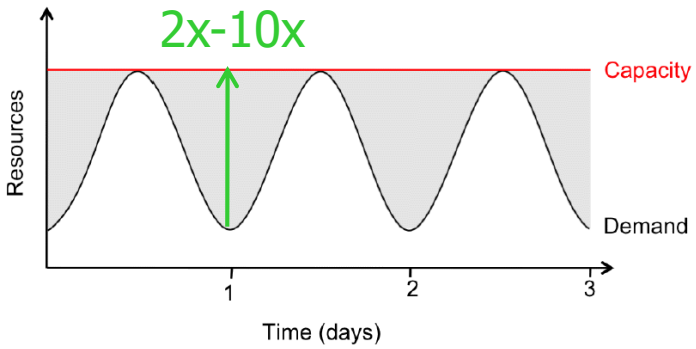
Data center

Network of data centers

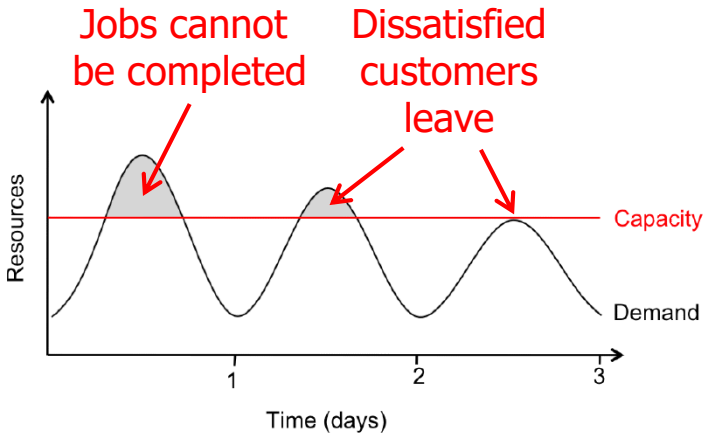
Google data center location (inferred):

Cloud Computing - Problem with "classical" scaling techniques

- Difficult to dimension because **load can vary considerably**
  - **Peak load** can exceed average load by factor 2x to 10x
  - **Result**: server utilization in existing data centers is around 5%-20%
  - **Dilemma**: waste resources or loose customers



Provisioning for the peak load



Provisioning below the peak

- Expensive
- Scaling up is difficult
  - Planning and setting up a large cluster is highly nontrivial
  - Need to order new machines, install them, integrate with existing cluster
  - Large scaling factors may require major redesign, e.g., new storage system, new interconnect, new special software, new building (!)
- Need maintenance

Solutions? The Cloud

- The power plant analogy:

- Cloud computing is a **model** for enabling convenient, **on-demand** network access to a **shared pool** of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction
- Characteristic:
  - On-demand self service
  - Measured service
  - Rapid elasticity (resources can be rapidly and elastically provisioned to quickly scale up and rapidly released to quickly scale down)
  - Resource pooling

The Cloud SPI model

- The cloud provides three types of commonly distinguished service:
  - **Software as a service** (SaaS)
    - The cloud provides an entire application (word processor, calendar, **data storage**, etc)
    - Customer pays cloud provider
    - Example: Google Apps, Saleforce.com
  - **Platform as a service** (PaaS)
    - The cloud provides just the middleware/infrastructure
    - Customer pays SaaS provider for the service, SaaS provider pays the cloud for the infrastructure
    - Example: Windows Azure, Google App Engine
  - **Infrastructure as a service** (IaaS)
    - Virtual machine, blade server, hard disk
    - Customer pays SaaS provider for the service, SaaS provider pays the cloud for the resources
    - Example: Amazon Web Services (AWS), Rackspace Cloud, GoGrid
- Who can became a customer of the cloud? Three types of cloud
  - **Public cloud**: commercial service, open to (almost anyone)
    - Amazon Web Services , Windows Azure, Google App Engine
  - **Community cloud**: shared by several similar organizations
    - Google's "Gov Cloud"
  - **Private cloud**: shared within a single organization
    - Internal datacenter of a large company

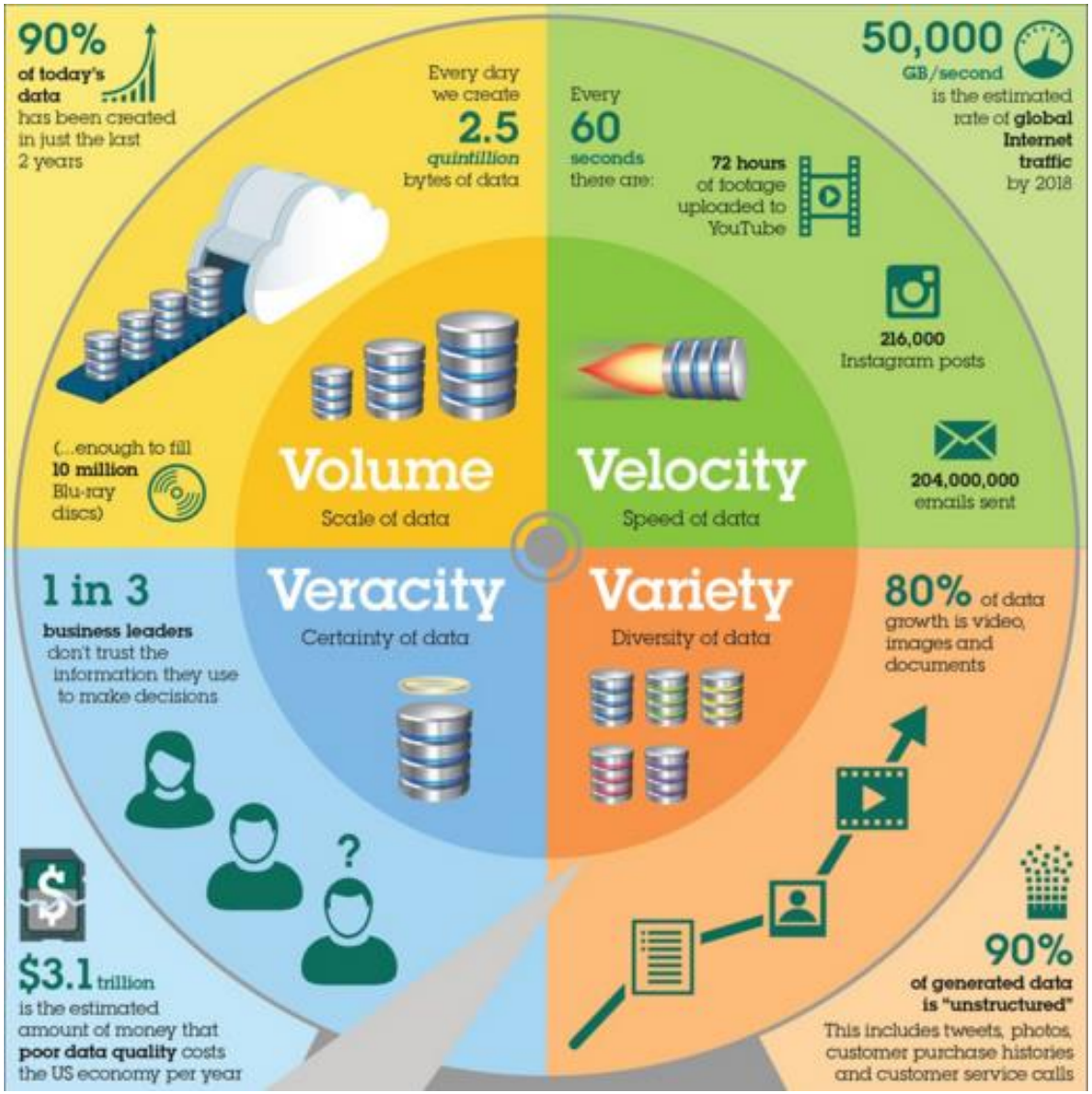
10 Obstacles and Opportunities for Cloud Computing [14]

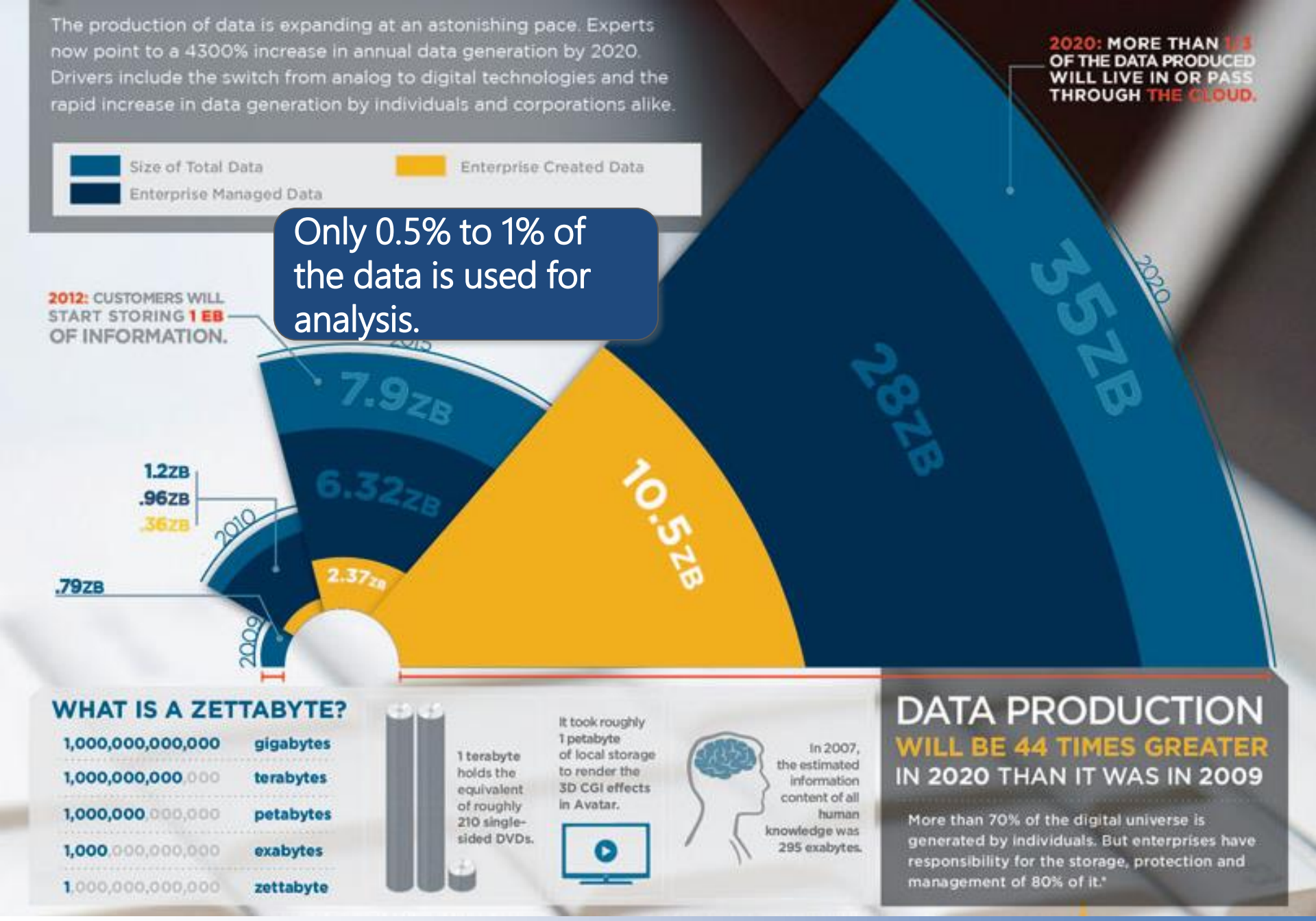
1. Availiability
2. Data lock-in
   - How do I move my data from one cloud to another?
3. Data confidentiality
   - How do I make sure that the cloud doesn't leak my confidential data?
4. Data transfer bottlenecks
   - How do I copy large amounts of data from/to the cloud?
5. Performance unpredictability
   - Remember, you are sharing resources with other customers
6. Scalable storage
   - The cloud model (short-term usage, infinite capacity on demand) does not fit well with persistent storage
7. Bugs in large distributed systems
8. Scaling quickly
9. Reputation fate sharing
   - One customer's bad behavior can affect the reputation of others using the same cloud
10. Software licensing

Big Data - Full faith in the power of data

The quest for knowledge used to begin with grand theories. Now it begins with massive amounts of data. **Welcome to the Petabyte Age!**

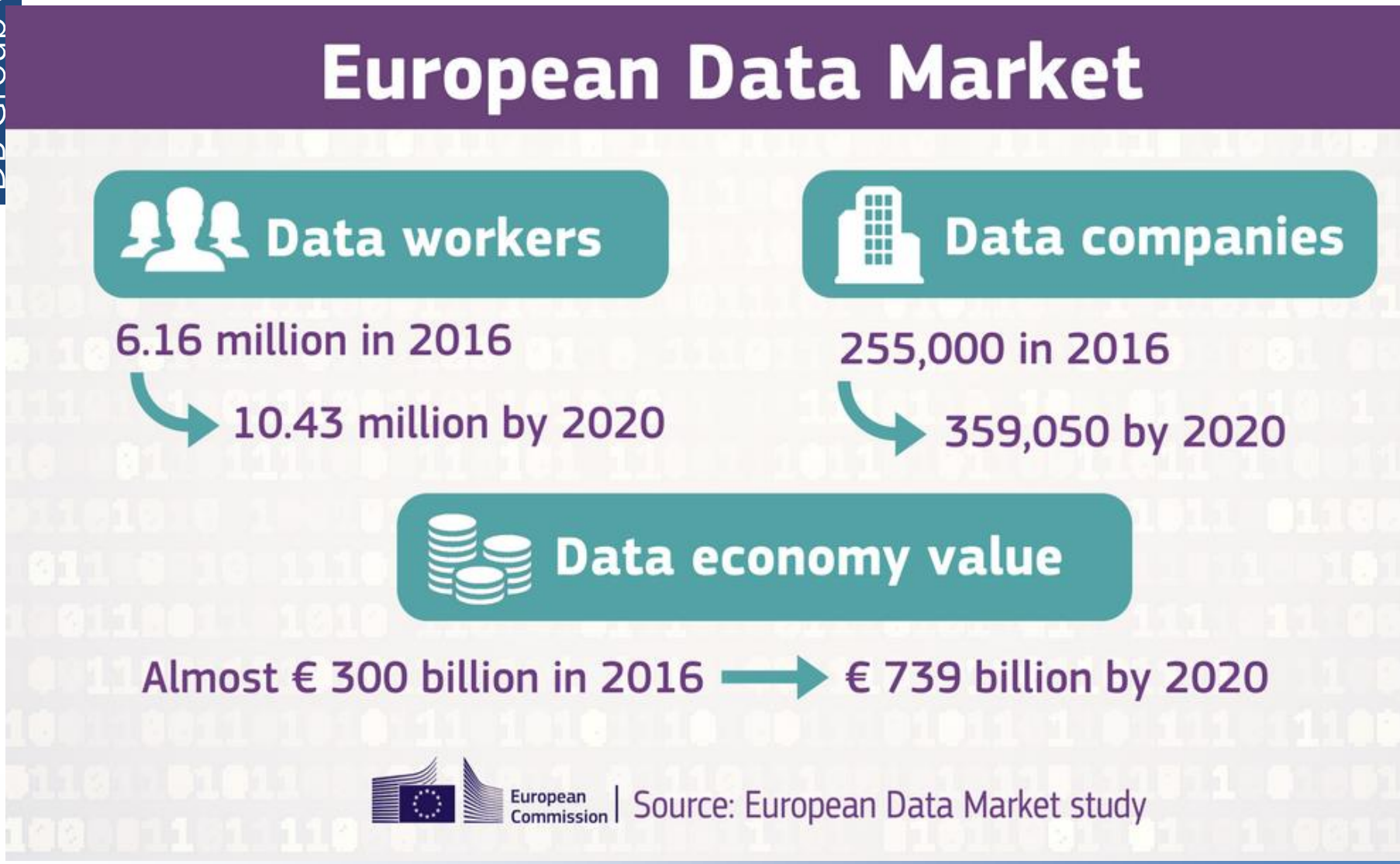
The FOUR V's of Big Data





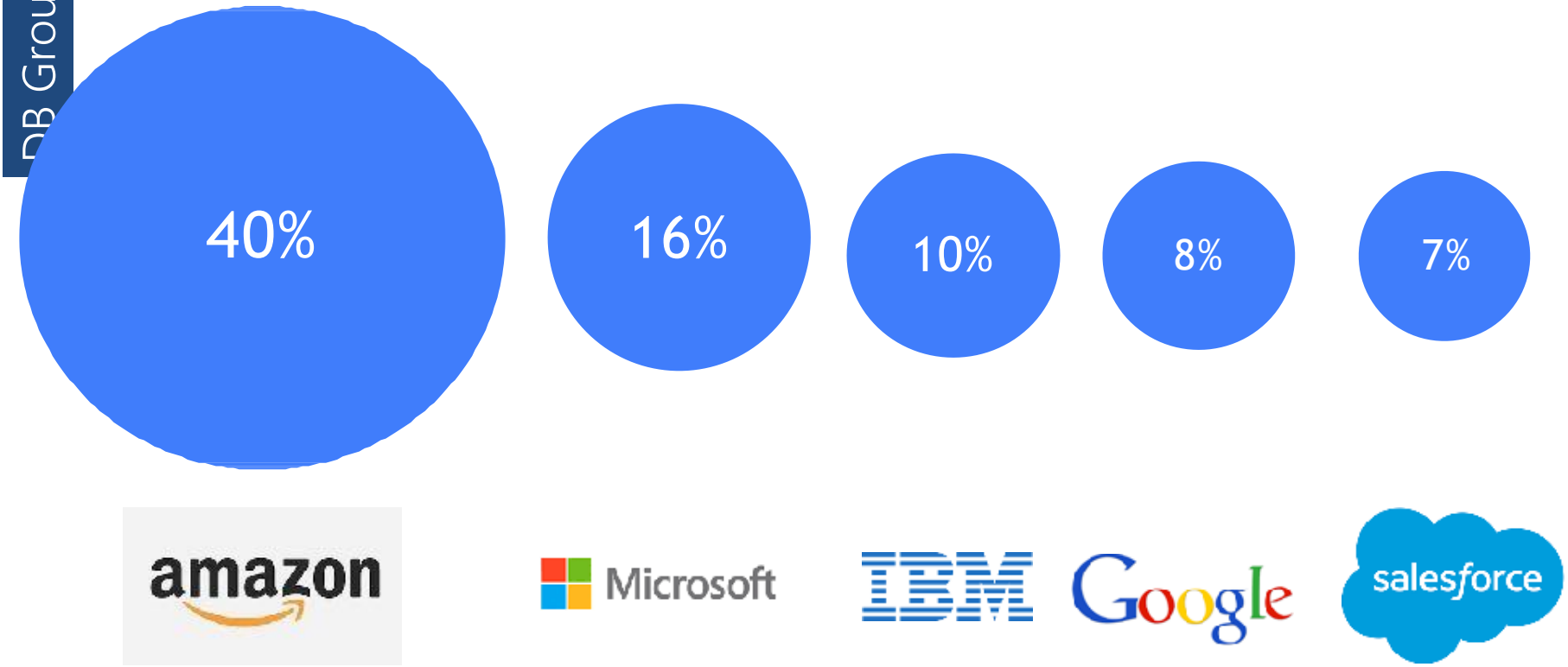
http://www.csc.com/insights/flxwd/78931-big_data_growth_just_beginning_to_explode
http://www.guardian.co.uk/news/datablog/2012/dec/19/big-data-study-digital-universe-global-volume

General Data Protection Regulation (GDPR)



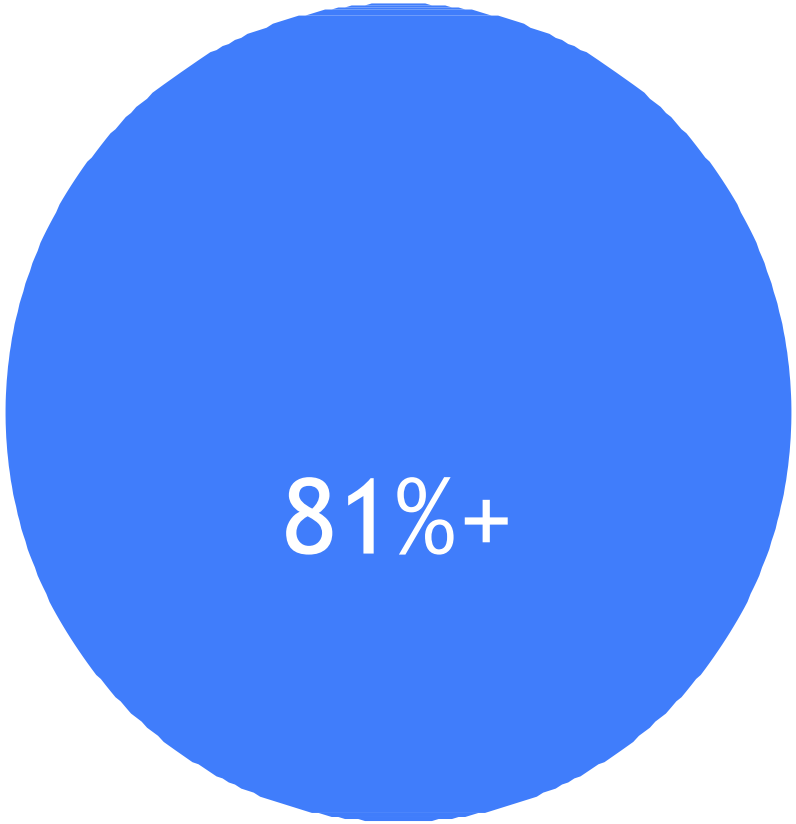
WHERE ARE THE DATA?

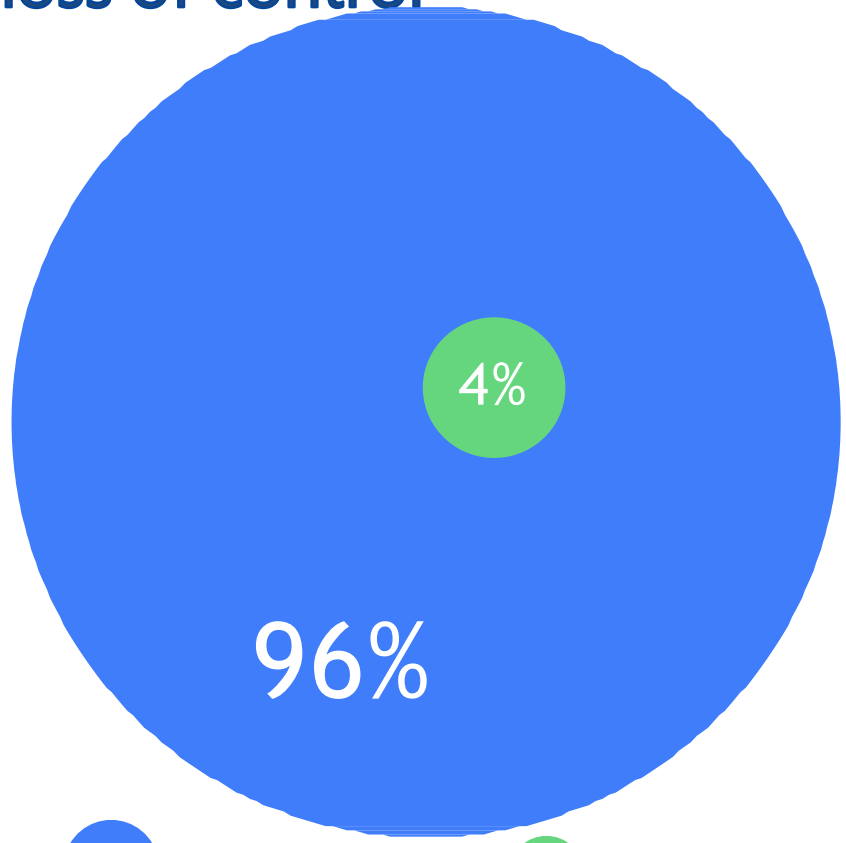
Volume and Variety: the loss of control



WHERE ARE THE DATA ?

The loss of control





Foreign

US Citizen

Value – The most important V of all!

- Then there is another V to take into account when looking at Big *Data: Value!*
- Having access to big data is no good unless we can turn it into value.
- What technologies?

Database Philosophy

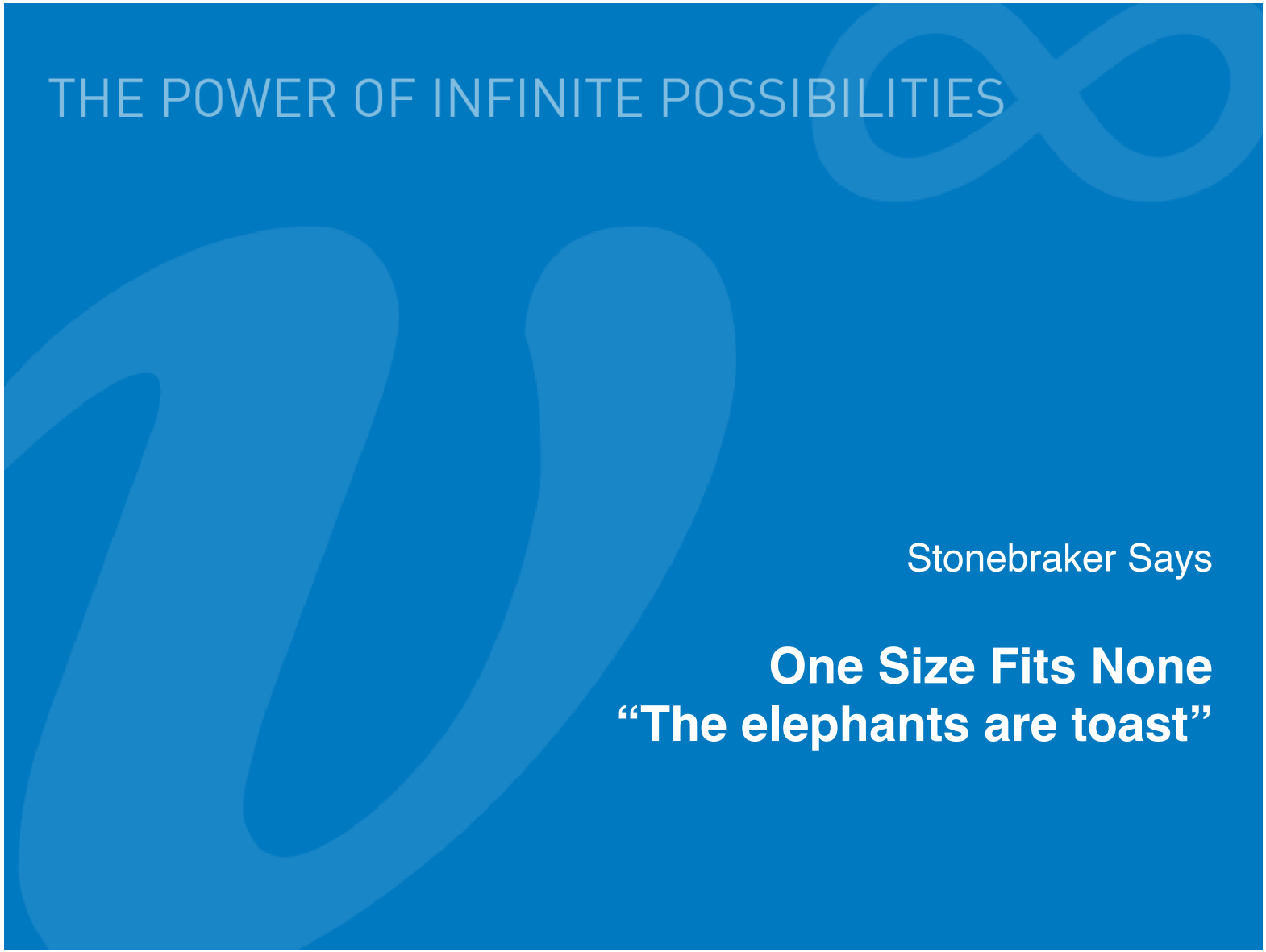
God made integers,
all else is the work of man.

(Leopold Kronecker, 19th Century Mathematician)

Codd made relations,
all else is the work of man.

(Raghu Ramakrishan, DB text book author)

Traditional RDBMS



At This Point, RDBMS is “long in the tooth”'

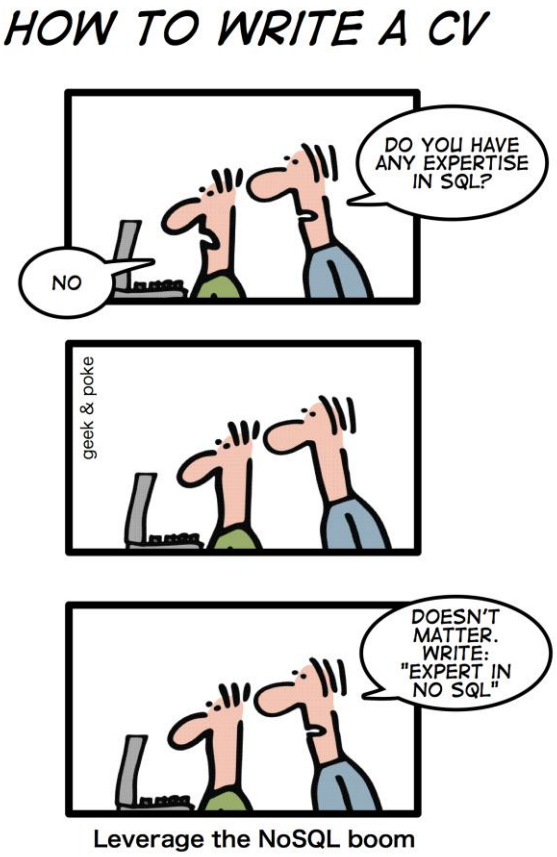
There are at least 6 (non trivial) markets where a row store can be clobbered by a specialized architecture !

- Warehouse (Vertica, Red Shift, Sybase IQ, DW Appliances)
- OLTP (VoltDB, HANA, Hekaton)
- RDF (Vertica, et. al.)
- Text (Google, Yahoo, …)
- Scientific data (R, MatLab, SciDB)
- Data Streaming (Storm, Spark Streaming, InfoSphere)

An emerging "movement" around non-relational software for Big Data

- NOSQL stands for "Not Only SQL" (but is not entirely agreed upon), where SQL doesn't really mean the query language, but instead it denotes the traditional relational DBMS.

- Google **Bigtable**, **Memcached**, and Amazon's **Dynamo** are the "proof of concept" that inspired many of the data stores we will see:
  - Memcached demonstrated that in-memory indexes can be highly scalable, distributing and replicating objects over multiple nodes
  - Dynamo pioneered the idea of *eventual consistency* as a way to achieve higher availability and scalability [5]
  - BigTable demonstrated that persistent record storage could be scaled to thousands of nodes [4]



Challenges (1) – Selection of the Big Data Technology

- **Volume, Velocity**

  Calling for new **Big Data systems:**

  – **Big Data Management Systems:** ***NOSQL & more***

  Many more…

  – **Big Data Analysis Systems:**

  - **Batch + Streaming**

  Many more…

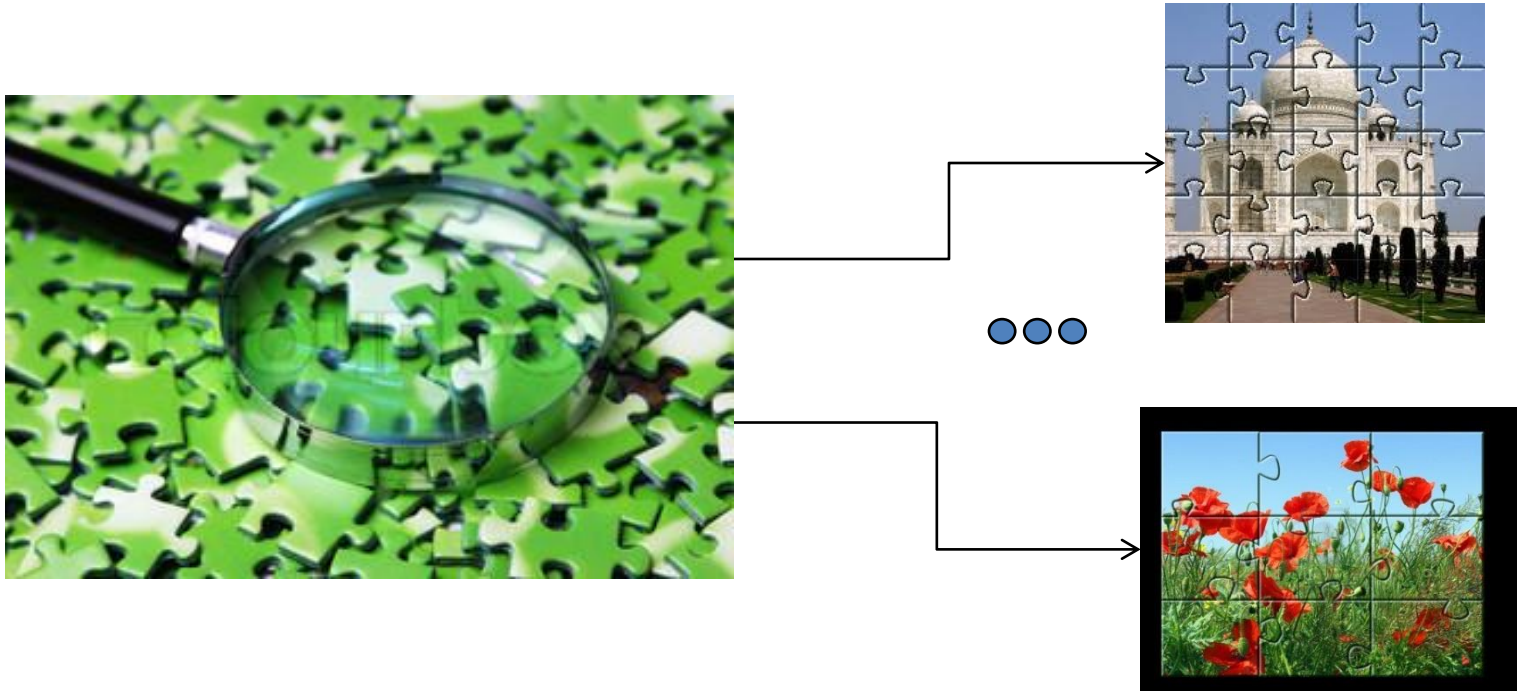
Not only Relational Database Management Systems
and Business Intelligence

Challenges (3)- Value Extraction

- Big Data Integration & Data Science
- Cognitive Computing

"Small" Data Integration: What Is It?

- Data integration = solving lots of puzzles
  - Each puzzle (e.g., Taj Mahal) is an integrated entity
  - Each piece of a puzzle comes from some source
  - Small data integration → solving small puzzles



Data integration as a new abstraction

- Databases are great: they let us manage huge amounts of data
  - Assuming you've put it all into your schema.
- In reality, data sets are often created independently
  - Only to discover later that they need to combine their data!
  - At that point, they're using different systems, different schemata and semantics
- The goal of data integration: tie together different sources, controlled by many people, und
- DBMS: it's all about a
  - Logical vs. Physical;
  - What vs. How.

Students:

SSN	Name	Category
123-45-6789	Charles	undergrad
234-56-7890	Dan	grad
	...	...

Courses:

CID	Name	Quarter
CSE444	Databases	fall
CSE541	Operating systems	winter

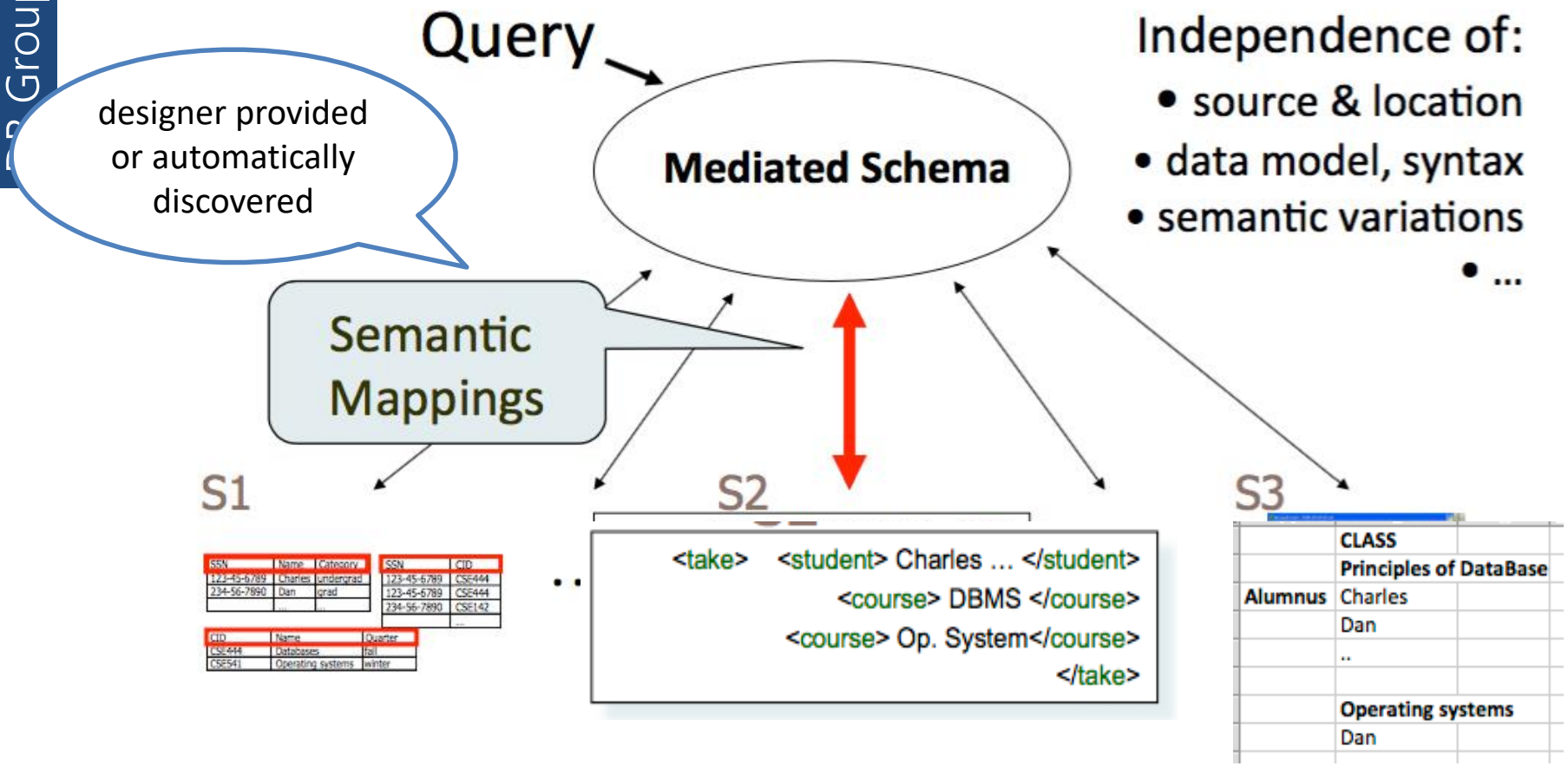
Takes:

SSN	CID
123-45-6789	CSE444
123-45-6789	CSE444

```
SELECT C.name
FROM Students S, Takes T, Courses C
WHERE S.name="Mary" and
      S.ssn = T.ssn and T.cid = C.cid
```

Data integration as a new abstraction (2)

- Data Integration: A Higher-level Abstraction



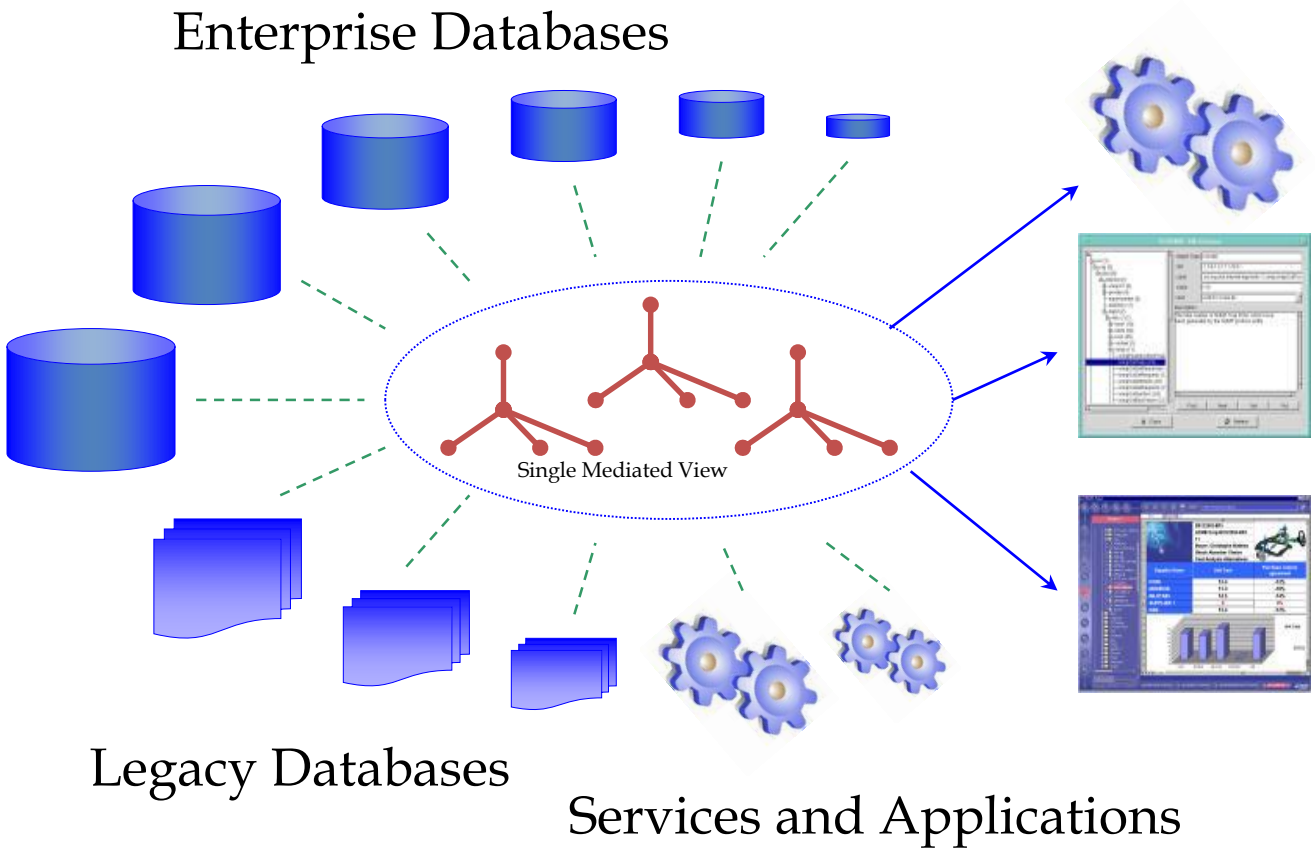
Data integration as a new technological solution

- Modern enterprises are often organized as "virtual networks", where enterprises operate through inter-enterprise cooperative processes
- To manage inter-enterprise processes and data exchange a key issue is to mediate among the heterogeneity of different information systems:

  Data Integration is a technological solution to build a shared and integrated knowledge base

- Applications of Data Integration
  - Business
  - Science
  - Government
  - The Web
  - Pretty much everywhere

Application Area 1: Business



EII Apps:

CRM
ERP
Portals
…

50% of all IT $$$ spent here!

Data integration as a new commercial software

- ✓ **According to Gartner:**
  Gartner estimates that the data integration tool market generated more than $2.7 billion in software revenue (in constant currency) at the end of 2016.

- ✓ A projected five-year compound annual growth rate of 6.32% will bring the total market revenue to around $4 billion in 2021 (see "Forecast: Enterprise Software Markets, Worldwide, 2014-2021, 2Q17 Update")

- ✓ The discipline of data integration comprises the practices, architectural techniques and tools that ingest, transform, combine and provision data across the spectrum of information types in the enterprise and beyond in order to meet the data consumption requirements of all applications and business processes.

Market Overview:

- ✓ The biggest change in the market from 2016 is the pervasive yet elusive demand for metadata-driven solutions.
- ✓ Consumers are asking for hybrid deployment not just in the cloud and on-premises (which is metadata-driven combined with services distribution), but also across multiple data tiers throughout broad deployment models, plus the ability to blend data integration with application integration platforms (which is metadata driven in combination with workflow management and process orchestration) and a supplier focus on product and delivery initiatives to support these demands.

Data integration in the research community

- The research community has been investigating data integration for more than 20 years: different research communities (database, artificial intelligence, semantic web) have been developing and addressing issues related to data integration:
  - Definitions, architectures, classification of the problems to be addressed
  - Different approaches have been proposed and benchmarks developed
- (see See tutorial: S. Bergamaschi, A. Maurino for an extensive description of data integration/ontology alignement techniques at:

http://www.dbgroup.unimo.it/site2012/index.php/component/content/article/2-uncategorised/93-momis)

- Future directions
  - Uncertainty, **Provenance**, and Cleaning
  - Lightweight Integration
  - Visualizing Integrated Data
  - Integrating Social Media
  - **Big Data Integration**

What is "Big Data Integration?"

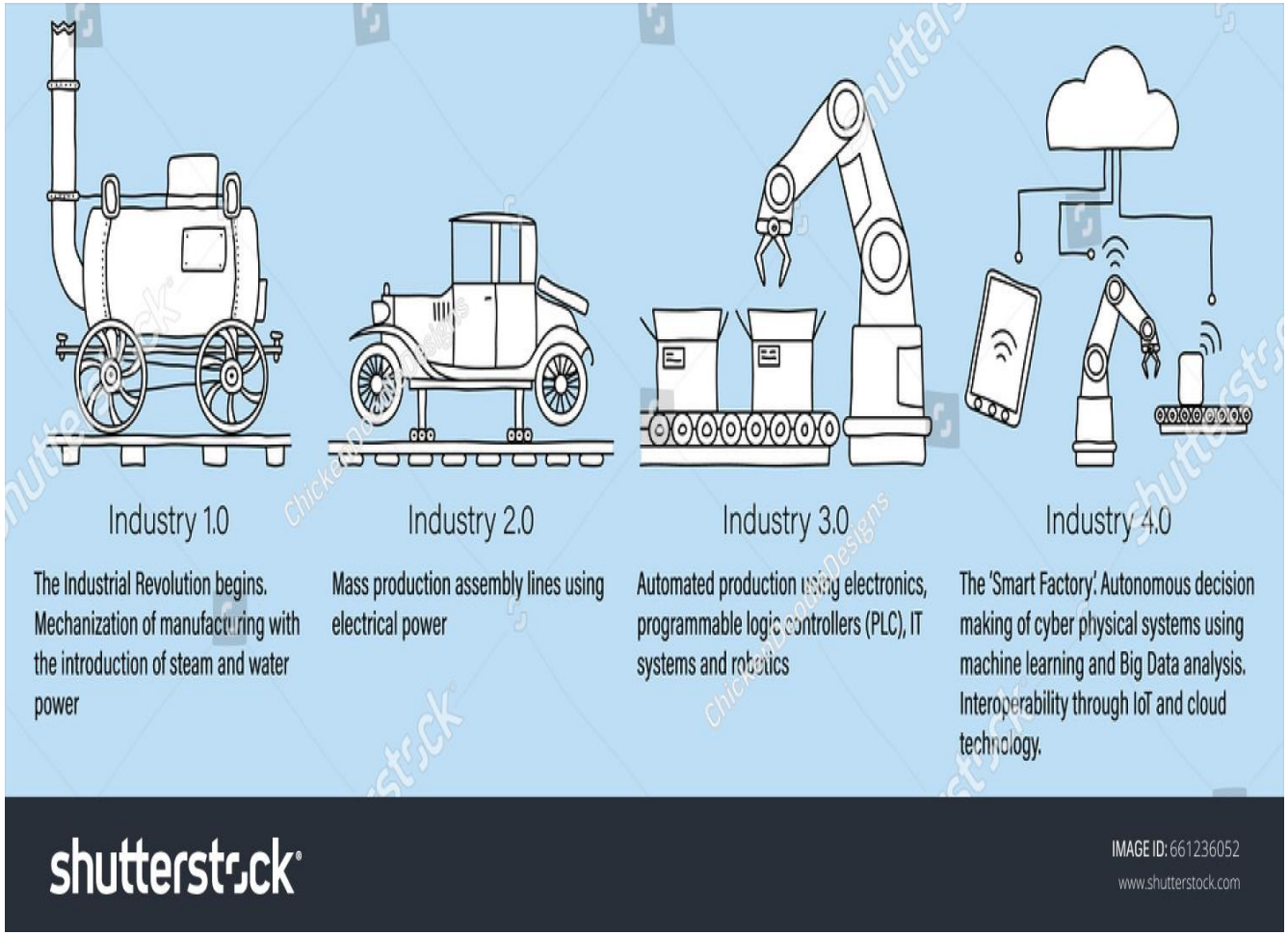
- Big Data Integration (BDI)= Big Data + Data Integration
- Data Integration: easy access to multiple data sources
  - Virtual: mediated schema, query reformulation, link + fuse answers
  - Warehouse: materialized data, easy querying, consistency issues
- Big data in the context of data integration: still about the V's ☺
  - Size: large volume of sources, changing at high velocity
  - Complexity: huge variety of sources, of questionable veracity
  - Utility: sources of considerable value

Challenges shared with Big Data

Interconnection among all physical devices (IoT):

- Volume Velocity Veracity Variety
  - ….

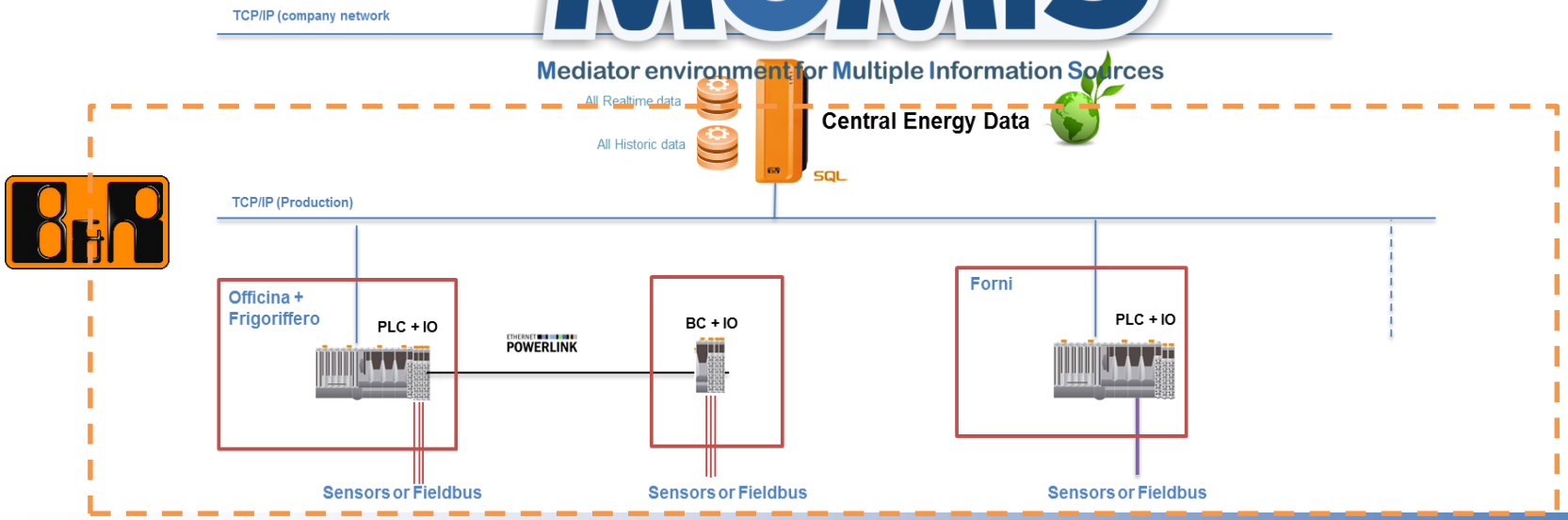
Industry 4.0


Industry 1.0
The Industrial Revolution begins. Mechanization of manufacturing with the introduction of steam and water power
Industry 2.0
Mass production assembly lines using electrical power
Industry 3.0
Automated production using electronics, programmable logic controllers (PLC), IT systems and robotics
Industry 4.0
The 'Smart Factory'. Autonomous decision making of cyber physical systems using machine learning and Big Data analysis. Interoperability through IoT and cloud technology.
shutterstock
IMAGE ID: 661236052
www.shutterstock.com

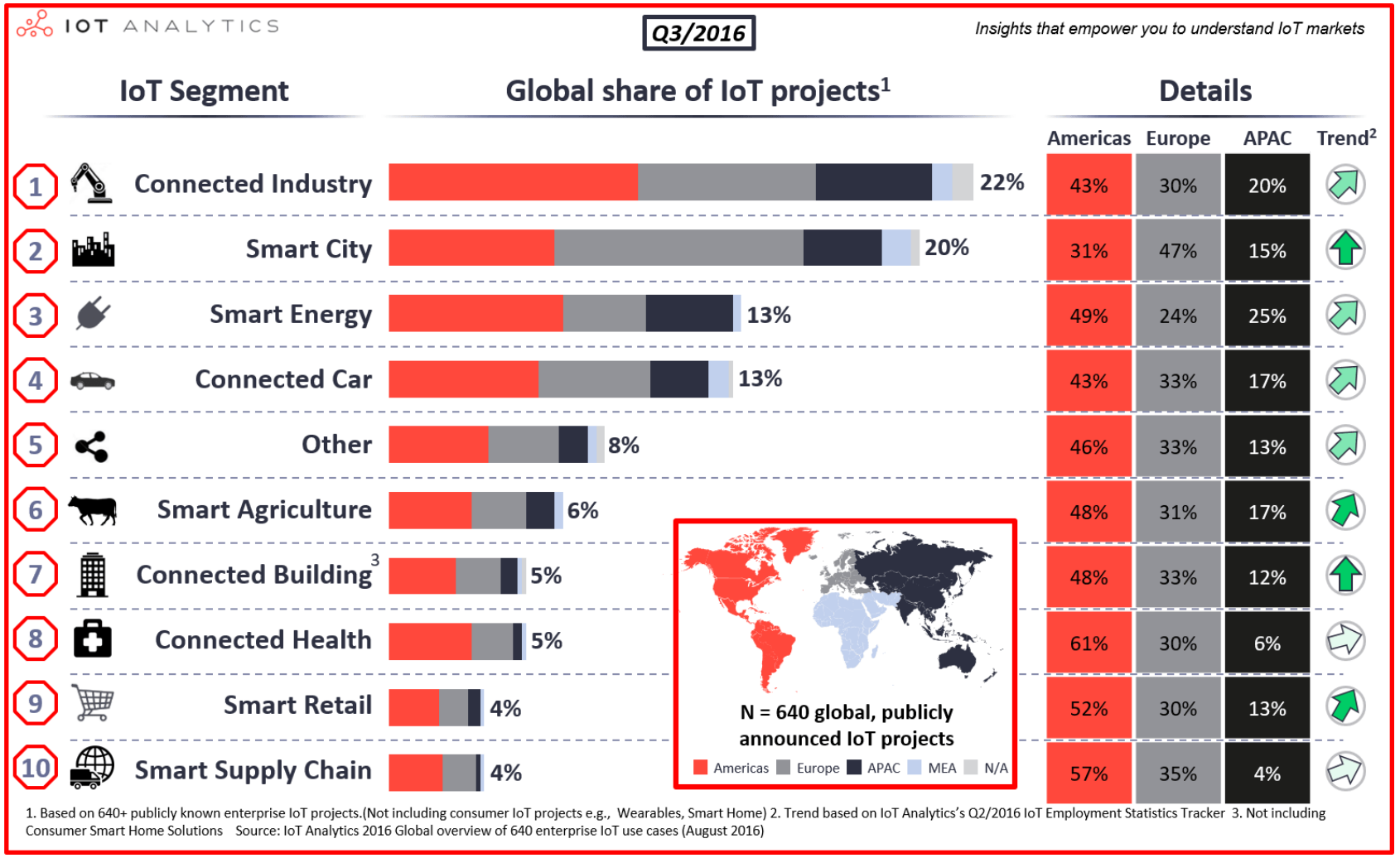
Data Integration Example

MoMIS

Mediator environment for Multiple Information Sources



Internet of Things(IoT) & Industry 4.0



Smart Cities

BDI: Why is it Challenging?

- Data integration = solving lots of puzzles
  - Big data integration → big messy puzzles
  - E.g., missing, duplicate, damaged piece

Cognitive Computing -Turing Test

Alan Turing

Turing Award

Machine Learning

Cognitive Computing - Autonomous vehicles

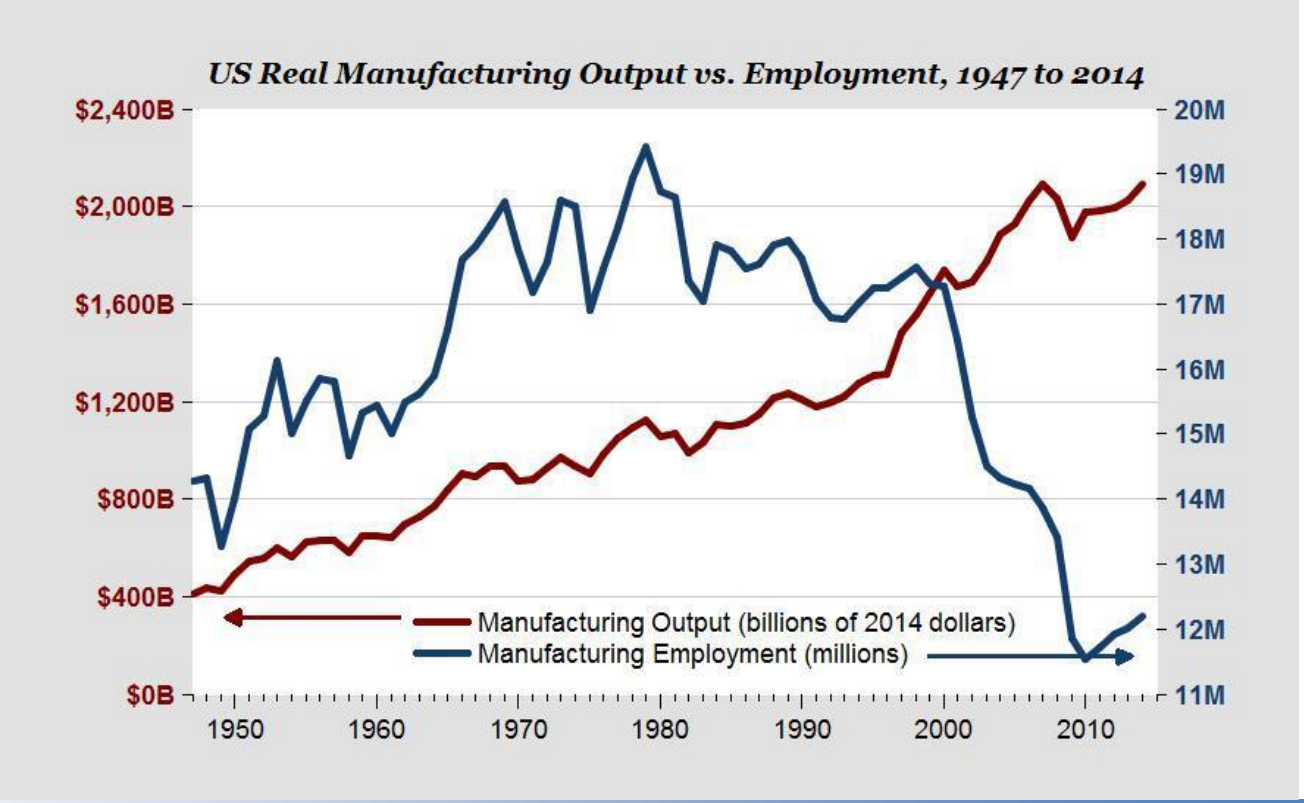
Automation of the entire supply chain is expected: cargo ships, ports, trucks, warehouses, delivery, ...

2016: AlphaGo beats the champion of Go Lee Sedol

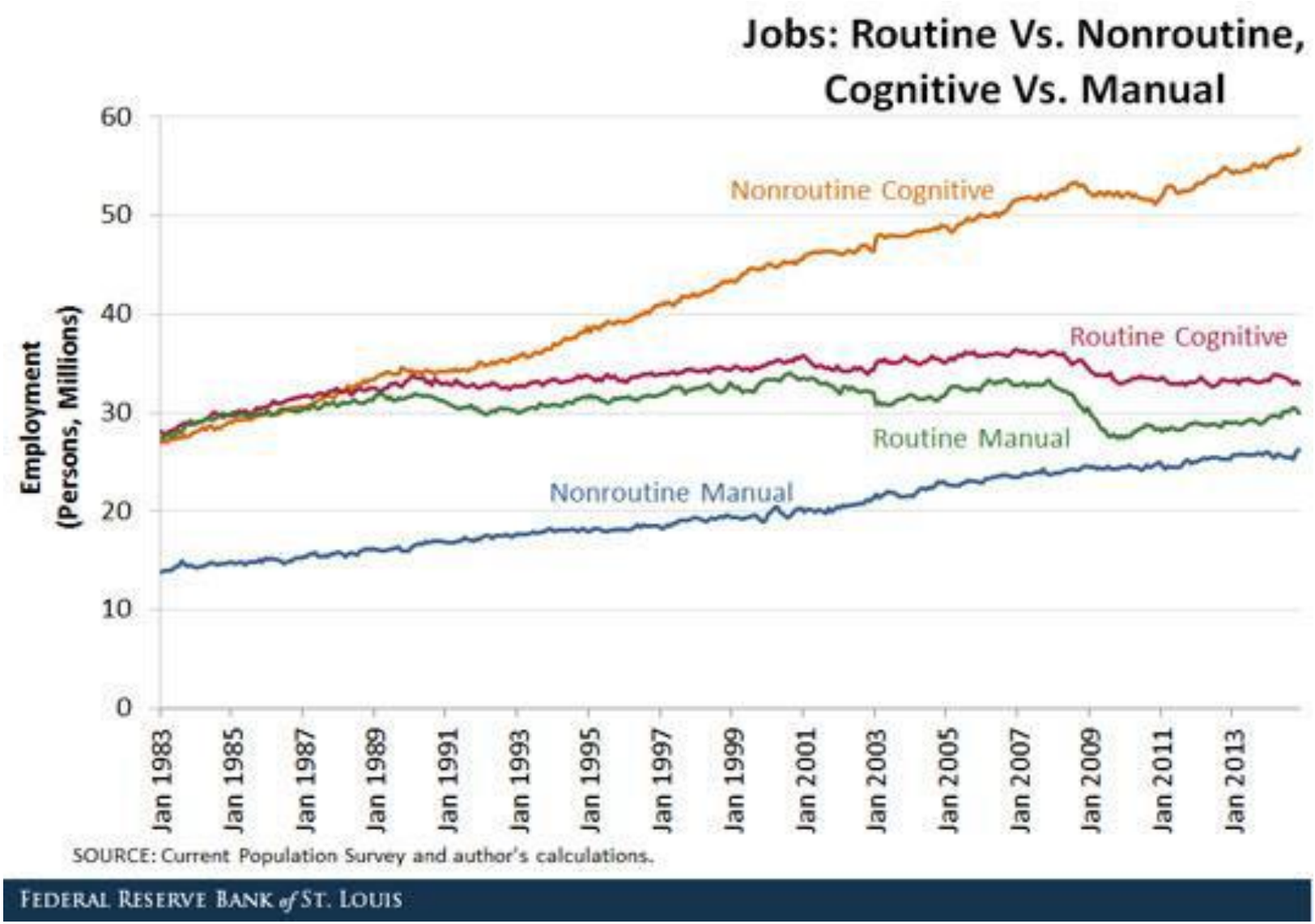
Thanks to machine learning techniques AlphaGo can develop «intuitions» for the game of Go.

Challenges : Evolution of the World of Work

- McKinsey: 45% of jobs will be replaced by already available technology
- Gartner: 1 of 3 jobs will be replaced by technology in 2025
- … But new jobs will be created by technology



Repetitive work vs. not repetitive, intellectual vs. manuals



What does the Dbgroup?

- **Big Data Management**
  the DBGroup is adopting and improving cutting-edge technologies to manage Big Data (e.g. Apache Hadoop, Apache Spark, NoSQL/NewSQL DBMS)

- **Big Data Analysis**
  how to get valuable insight from the data, and how to extract information to drive decision making process, when the data is too large for traditional algorithms (including traditional BI and machine learning techniques)

- **Big Data Integration**
  to extract the real value of data by integrating different and heterogeneous data sources (structured, semi-structured, unstructured)

What does the DBGROUP do? Dissemination & Teaching on Big Data Management, Integration and Analysis

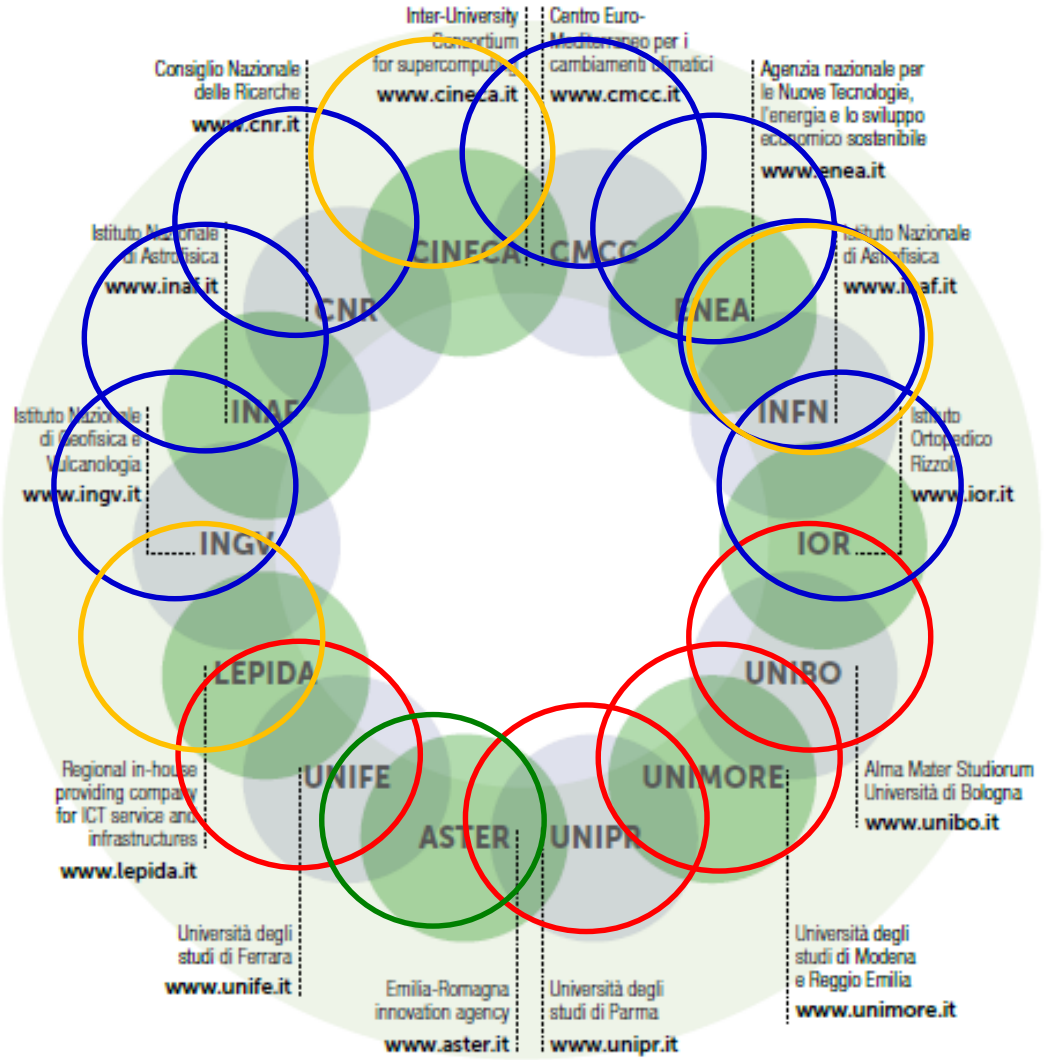
- Sonia Bergamaschi - ”Big Data panel" at SEBD 20yy
- Sonia Bergamaschi - invited paper at CLADAG 2015 (8-10 October 2015)
- Sonia Bergamaschi - Workshop "PICO: the CINECA solution for Big Data management" @ headquarters of Casalecchio on December 5th 2014
- Sonia Bergamaschi - IC3K 2014 (http://www.ic3k.org/KeynoteSpeakers.aspx) - lecture title"Big Data integration - State of the Art &Challenges" - Roma 21-24 October 2014.
- Sonia Bergamaschi - BDAA 2014 - lecture title " Big Data Analysis: Trends & Challenges" [IEEE Proceedings of the International Conference on High Performance Computing & Simulation (HPCS 2014), pag. 303 – 304.

Advanced Training COURSES

- *Corso di Formazione per l'Ordine degli Ingegneri*: “Metodi e Tecniche per l'Analisi di Big Data” DIEF UNIMORE (15 ore) – Aprile 2017.
- Several courses "Tools and techniques for massive data analysis" promoted in conjunction with Cineca for the scientific research community on 2015, 2016,2017.

What does the Emilia Romagna region do on Big Data?

R&I Public stakeholders involved



CONNECTIVITY
LEPIDA, GARR

INFRASTRUCTURES
HW
CINECA, INFN, LEPIDA,
SW
CINECA, INFN, UNIMORE, UNIBO, UNIFE, INAF, CNR, ENEA

END USERS
UNIMORE, CINECA, INFN, UNIBO, UNIFE, INAF, CNR, IOR, UNIPR, LEPIDA, ENEA, CMCC, INGV

Academy & research grants- Emilia Romagna region

Academy "Methods, techniques and tools for the analysis of Big Data"

- **DBgroup** Università di Modena e Reggio Emilia & **CINECA**
- 50 hours of training:
  - 30 hours of teaching
  - 20 hours big data lab with CINECA infrastructure
- 20 places available
  - 10 € 1,500 scholarships
  - Registration: € 3,000

Research grants for companies in Modena

- **BPER**: "Big Data and Analytics for the development of the customer's digital behavior from prospect to acquired "
- **DOXEE**: "Methodology of designing Big Data applications based on Amazon Web Services technology"
- **Expert System**: "Data Scientist to support the intelligence production process (Corporate Intelligence Data Scientist"

UNIMORE Engineering Departments

What do the Engineering departments of Modena and Reggio Emilia do?

- International Conferences
- University Education
  - Master's Degree Couses
- University Top Education
  - PhD courses of basic research, industrial research,
  - High apprenticeship
  - (as coordinator of the PhD course in ICT (Information and Communication Technologies) www.ict.unimore.it
    please contact me for information)

FAIM2017
27th International Conference on Flexible Automation and Intelligent Manufacturing

June 27-30, 2017
Modena, Italy

IEEE RTSI 2017 RESEARCH AND TECHNOLOGIES FOR SOCIETY AND INDUSTRY 3° INTERNATIONAL FORUM

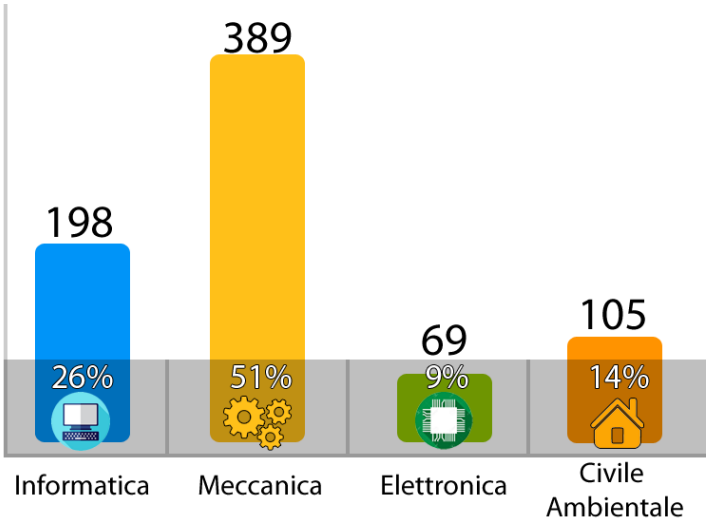
IEEE Italy Section

UNIVERSITAS STUDIORUM MUTINENSIS ET REGIENSIS 1175

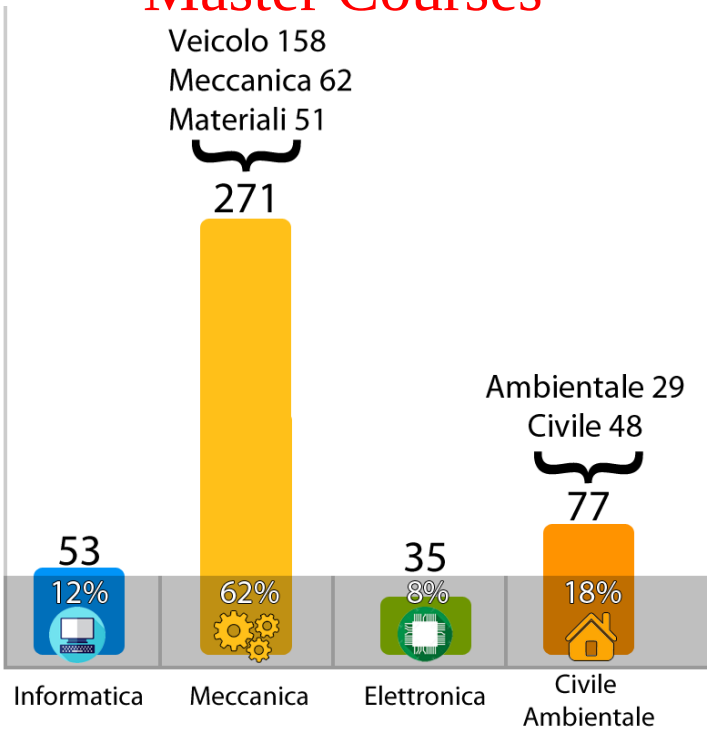
UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Industry 4.0 Smart healthcare Smart mobility

Bachelor Courses

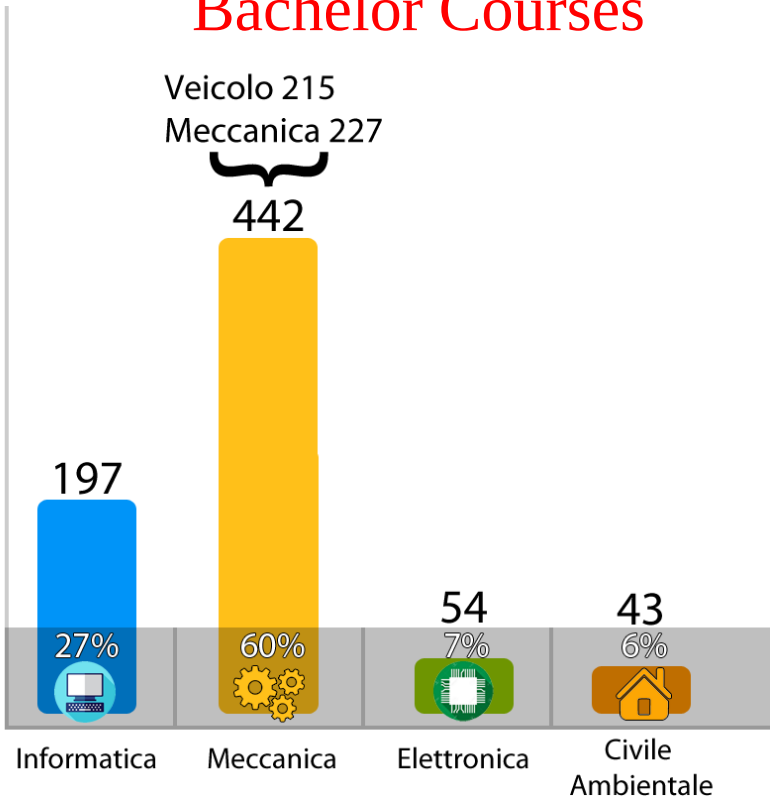


Master Courses



Enrolled students 2016/2017

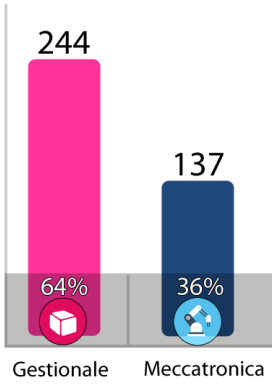
Bachelor Courses



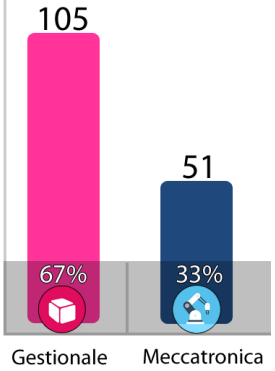
Enrolled Students 2017/2018

Computer Engineering training offer

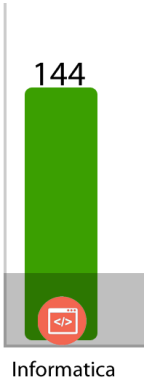
Enrolled Students



Triennali

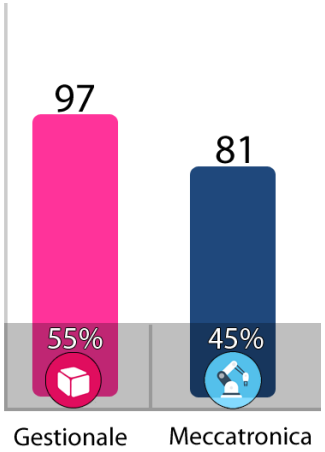


Magistrali

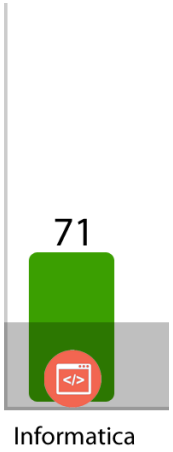


Triennale

Magistrale



Triennali



Triennale

And women?

Are women involved?

Project Digital Girls

5° Edizione *(dal 2014)*
June – July 2018

4-week free summer camp
Dedicated to the students of 3rd and 4th high school

Laboratory activities– ***Learn by doing***
- Development of video games in Python
- Team working

female role models in ICT
#teamwork #creativity #coding #empowerment

Thank you!