DB Group @ unimore

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Big Data: Opportunità e Sfide

Accademia Nazionale di Scienze, Lettere e Arti

23 Novembre 2017

Prof. Sonia Bergamaschi

Department of Engineering "Enzo Ferrari"

Università di Modena e Reggio Emilia

DB Group UNIMORE

Outline

- Chi sono e cosa è il DBGROUP?
- Cosa sono i Big Data?
- Nuove Tecnologie Informatiche per Big Data Management & Analysis
- Le Nuove Sfide Tecnologiche: Big Data Integration, Cognitive Computing
- Rischi sull'Occupazione?
- UNIMORE come si colloca rispetto a queste sfide scientifiche, tecnologiche e occupazionali?
  - Il contributo del gruppo di ricerca sui database (DBGROUP)
  - Il contributo del Dipartimento di Ingegneria "Enzo Ferrari" e le scelte dei nostri giovani
  - Il contributo della Regione Emilia Romagna
- E le donne?
- Gli aspetti etici?

DBGroup at Glance

Prof. Sonia Bergamaschi

Leader of the Database research group (DBGroup)

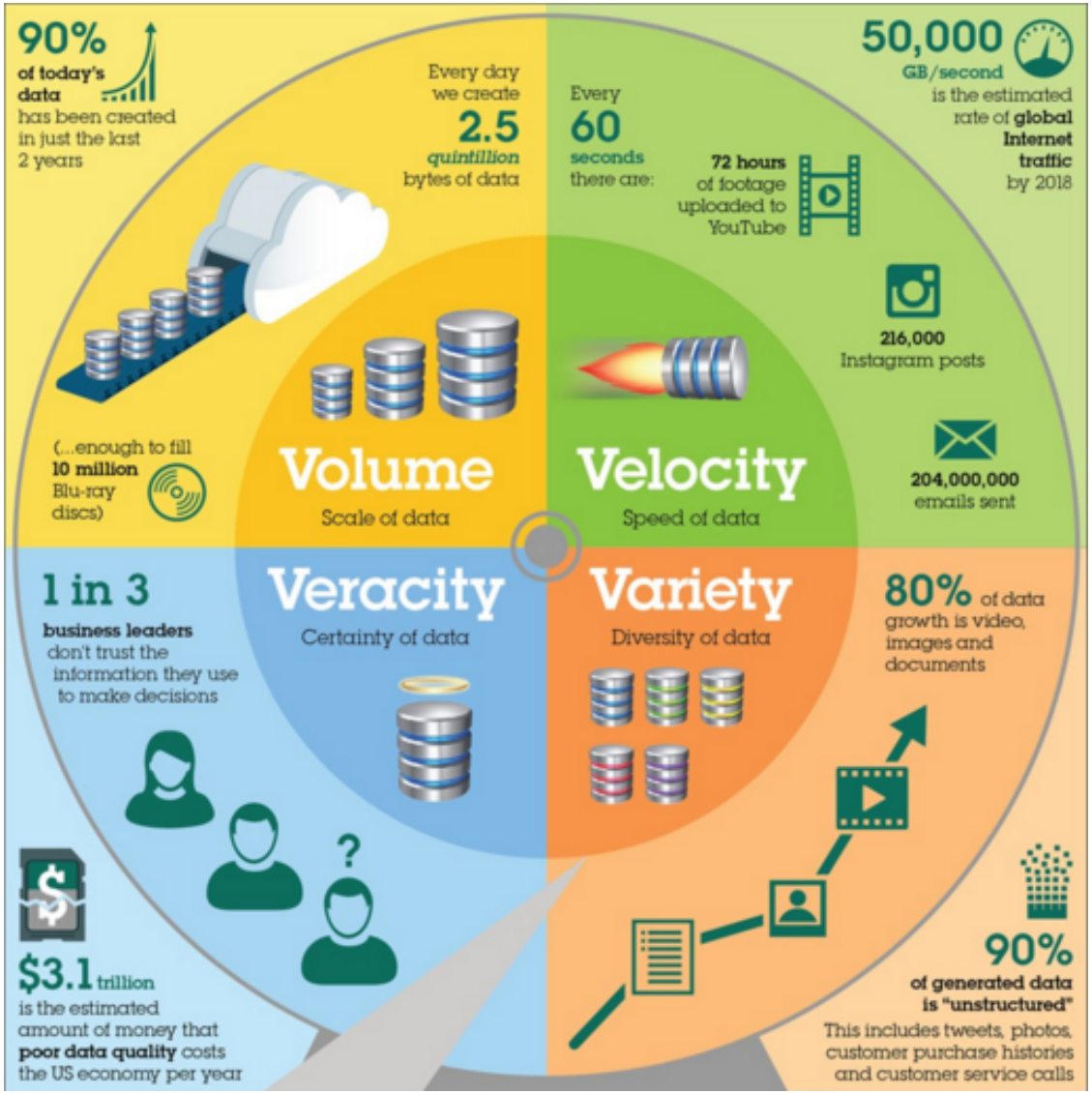
Dean of the ICT doctorate
(www.ict.unimore.it)

- 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
- ***e…. Allieva del prof. Paolo Tiberio***

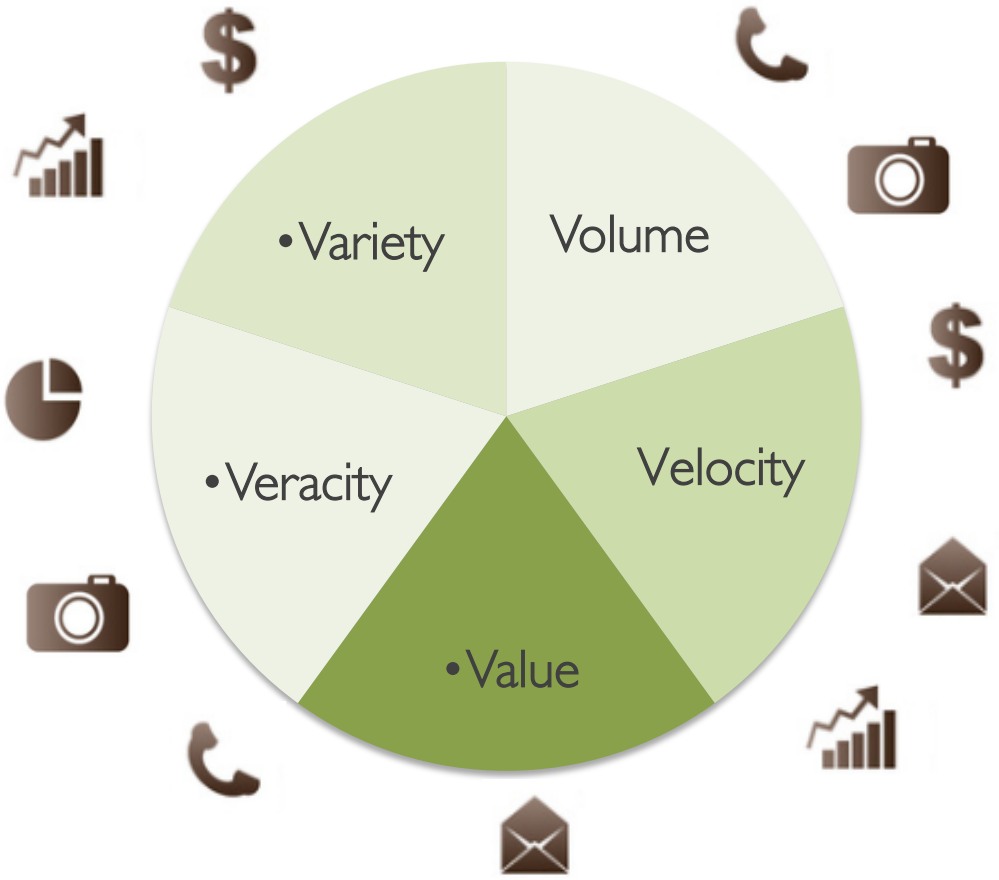
DBGroup at Glance

- Current Members:
  - 5 faculty
    - Sonia Bergamaschi
    - Domenico Beneventano
    - Maurizio Vincini
    - Francesco Guerra
    - Laura Po
  - 1 postdoc
    - Giovanni Simonini (IEEE best Computer Science phd thesis award 2017)
  - 6 ICT PhD students
    - Zhu Song (*Big Data Management* – 3rd year)
    - Luca Magnotta (industrial phd DATARIVER on *Big Data Integration & Analysis* –2nd year)
    - Gagliardelli Luca (Emilia-Romagna phd scholarship on *Big Data Integration & Analysis*– 2nd year)
    - Giueppe Fiameni (CINECA –*Big Data Management* –2nd year)
    - Giovanni Morrone (*Cognitive Computing* phd at Doctorate School Industria 4.0– 1st year)
    - Nicolò Parmiggiani (INAF-Astronomical Data –1st year)
- Reference person for UNIMORE **CINI Big Data Lab**
- 1 spin-off DATARIVER (now innovative SME) to deploy the MOMIS Data Integration System www.datariver.it

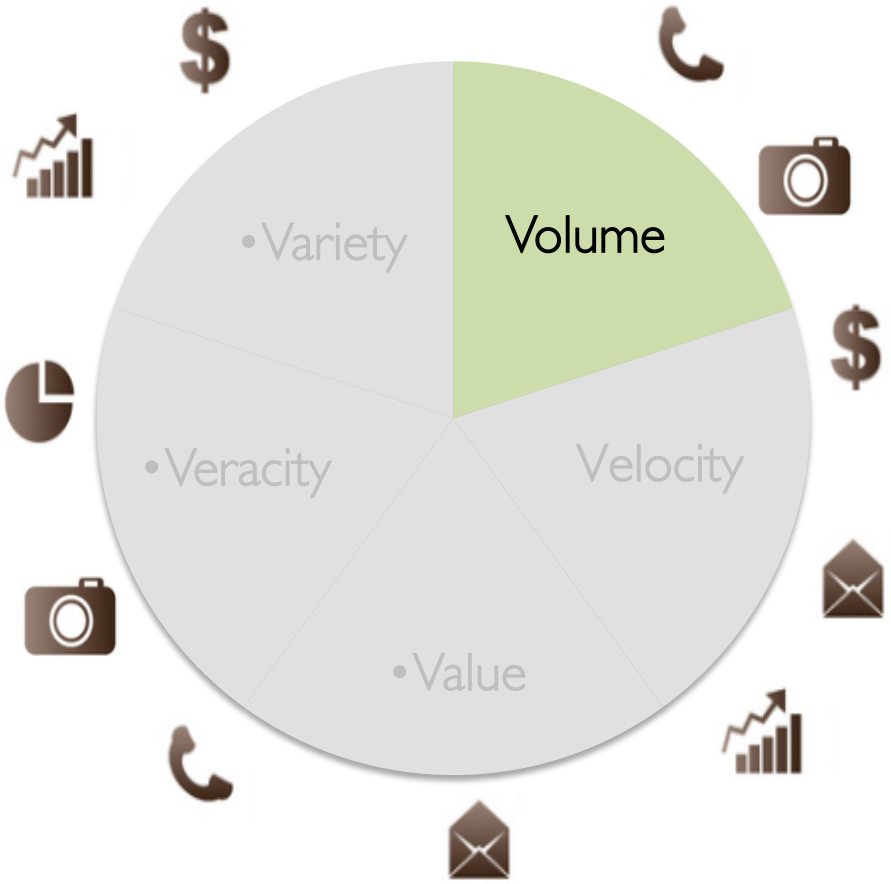
The FOUR V's of Big Data



Big Data have Five V's!



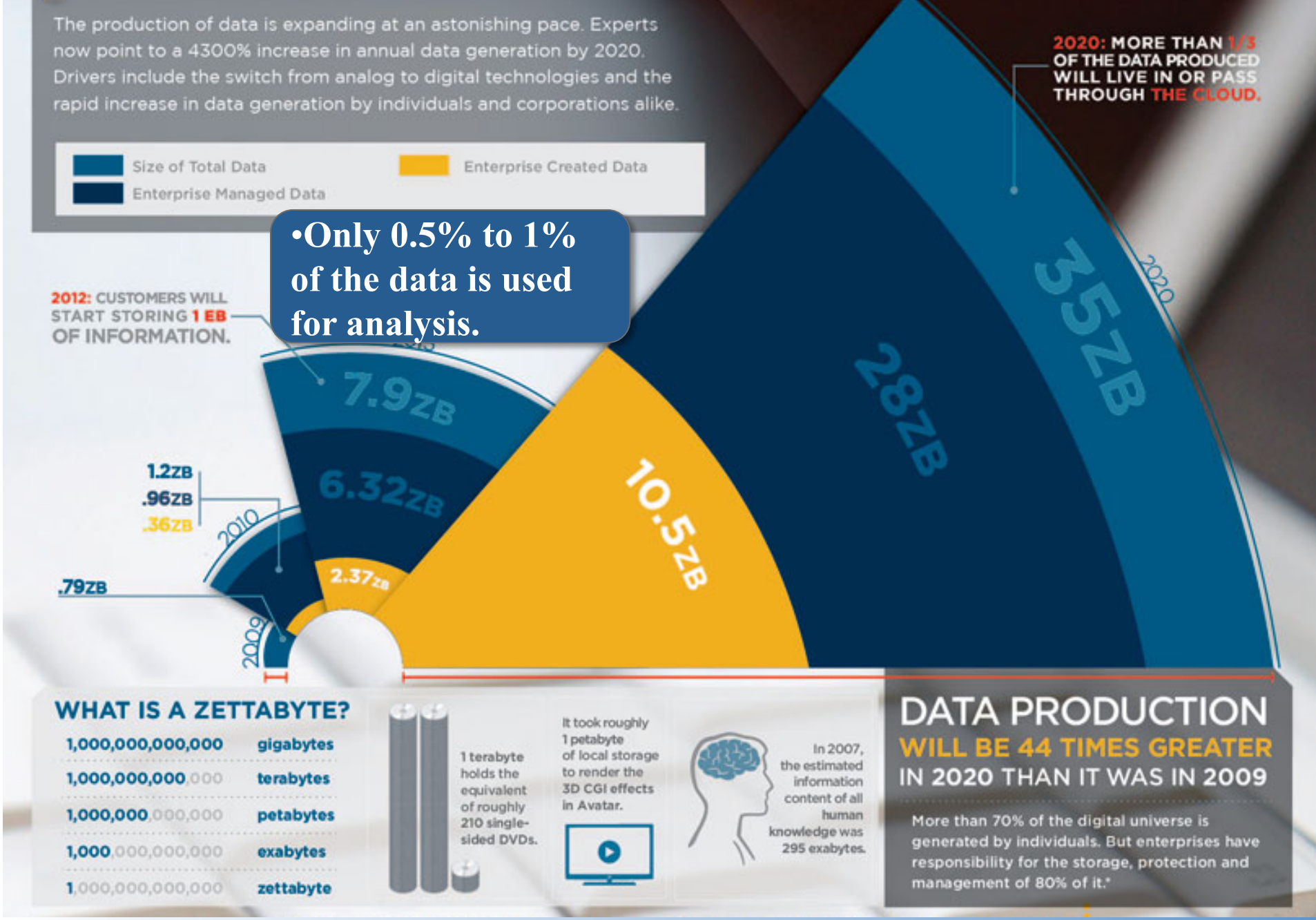
Increasing volumes of data, that grow at exponential rates



The increase in data volume is due to many factors:

- transaction based data stored through the years
- text data constantly streaming in from social media
- increasing amounts of sensor data being collected, etc.

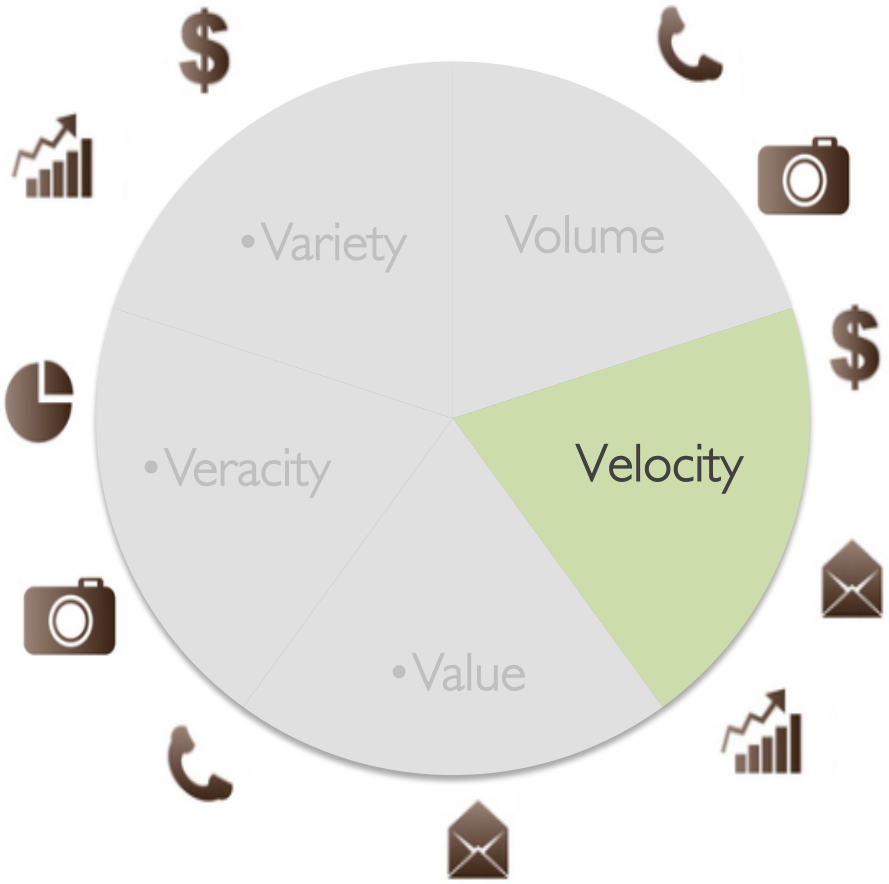
In the past, excessive data volume created a storage issue, but with today's decreasing storage costs, other issues emerge, including how to determine *relevance* amidst the large volumes of data and how to create *value* from data that is relevant



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

Increasing velocity at which data changes, travels or increases



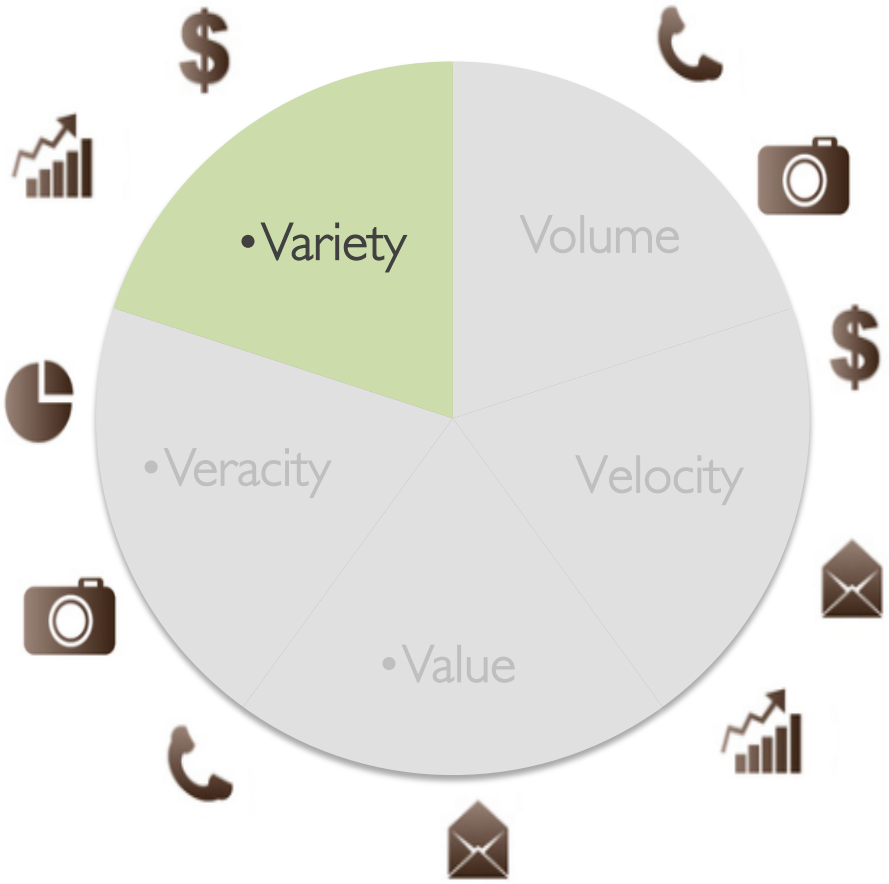
- According to Gartner, velocity means both:
- how fast data is being produced
- how fast the data must be processed to meet demand

Reacting quickly enough to deal with velocity is a challenge to most organizations

Velocity

Current application examples: financial services, stock brokerage, weather tracking, movies/entertainment and online retail

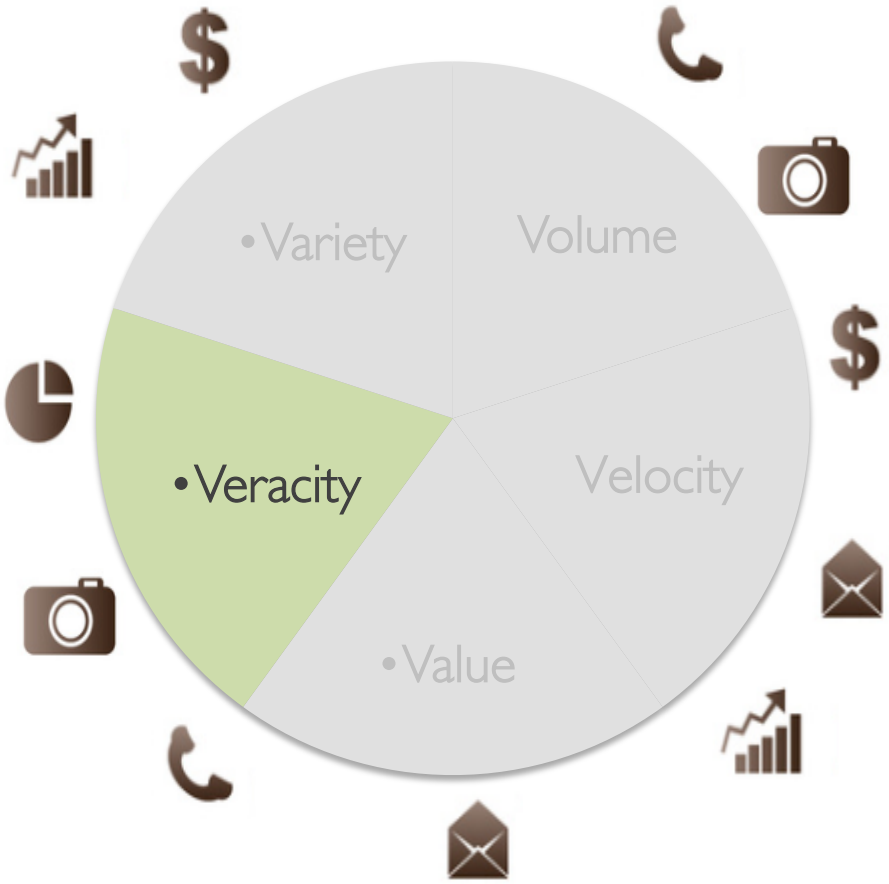
Increasing Variety of data types



Data today comes in all types of formats:

- from traditional databases to RDF data stores created by end users and OLAP systems
- to text documents, email, meter-collected data, video, audio, stock ticker data and financial transactions.

We see increasing veracity (or accuracy) of data



Refers to the ***messiness*** or ***trustworthiness*** of the data. With many forms of big data ***quality*** and ***accuracy*** are ***less controllable***

(just think of Twitter posts with hash tags, abbreviations, typos and colloquial speech as well as the reliability and accuracy of content)

but technology now allows us to work with this type of data.

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
- Companies are starting to generate amazing value from their big data

Questions typically asked on Big Data

- What if your data volume gets so large and *varied* you don't know how to deal with it?
- Do you store all your data?
- Do you analyze it all?
- What is coverage, skew, quality?
- How can you find out which data points are really important?
- How can you use it to your best advantage?

State-of-the-Art & Current Focus on Big Data

- Focus on verticals
  advertising, social media, retail, financial services, telecom and healthcare

  – Aggregate data, focused on transactions, *limited integration (limited complexity)*, analytics to find (simple) patterns

  – Emphasis on technologies to handle volume/scale, and to lesser extent velocity: Hadoop, NoSQL, MPP (Massive Parallel Processing) for data warehouse: DWA (Data Warehousing Appliance), ….

  – Full faith in the power of data (no hypothesis), bottom up analysis

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**!

Technologies for Big Data

- Managing Big Data
- Analyzing Big Data

Database Philosophy

God made integers,

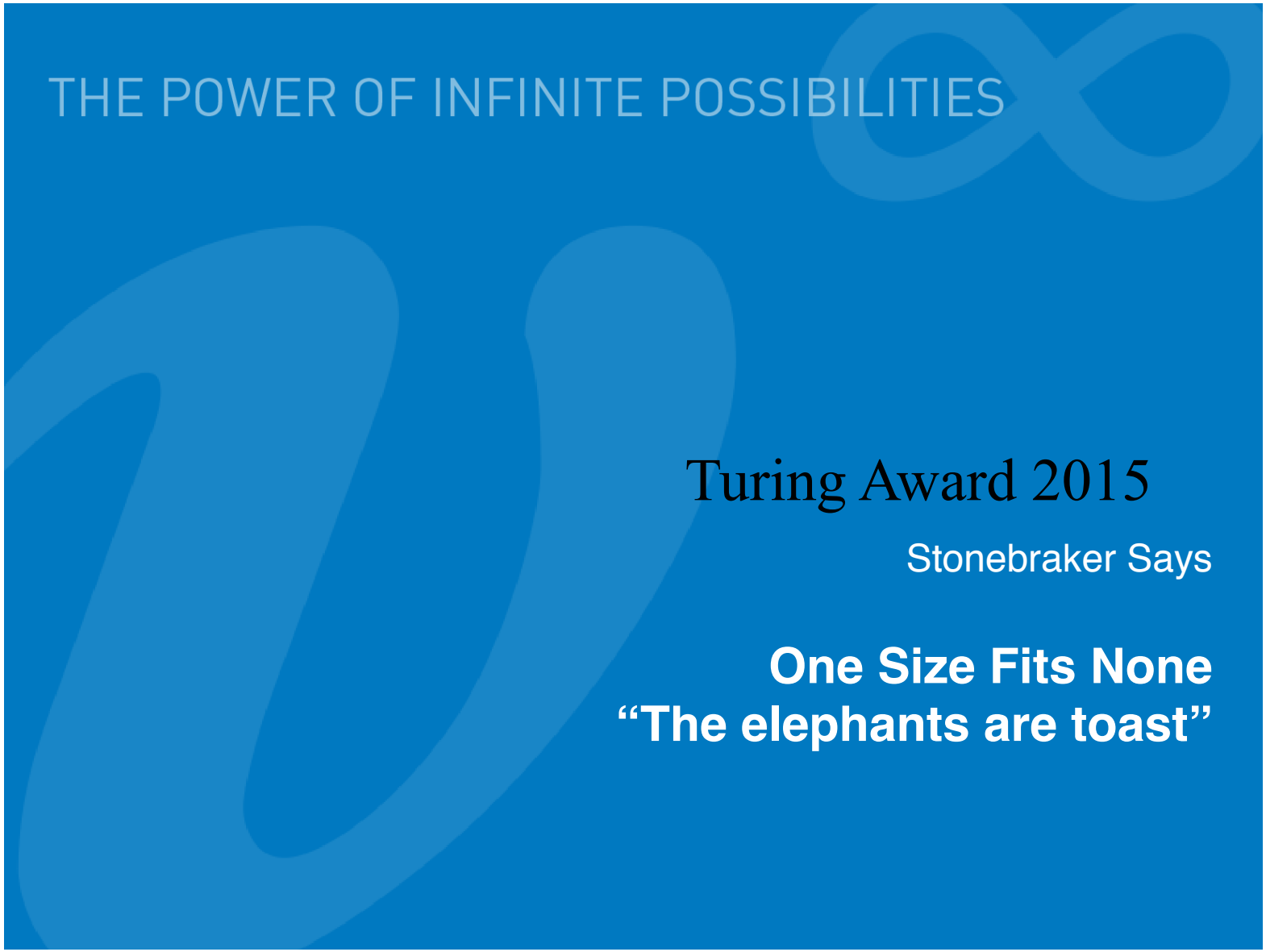
all else is the work of man.

(Leopold Kronecker, 19th Century Mathematician)

Codd made relations,

all else is the work of man.

(Raghu Ramakrishan, DB text book author)

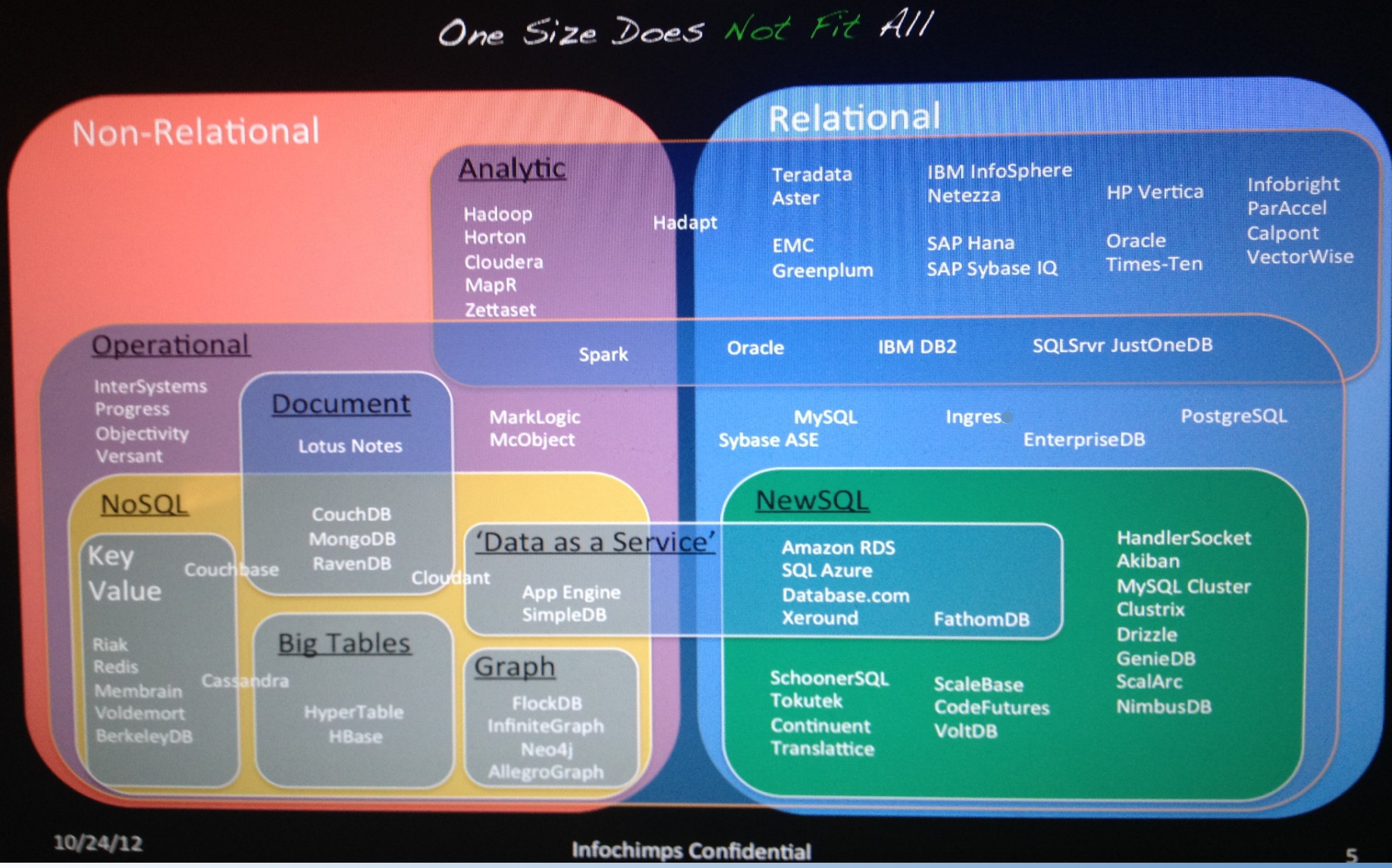

THE POWER OF INFINITE POSSIBILITIES
Turing Award 2015
Stonebraker Says
One Size Fits None
“The elephants are toast”

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)

Variety of Data Analytics Enablers



http://techcrunch.com/2012/10/27/big-data-right-now-five-trendy-open-source-technologies/

Challenges (1) – Selection of the Big Data Technology

- **Volume, Velocity**

  Calling for new **Big Data systems:**

  - **Big Data Management Systems:**

    • ***Many more…***

  - **Big Data Analysis Systems:**

    • **Batch + Streaming**

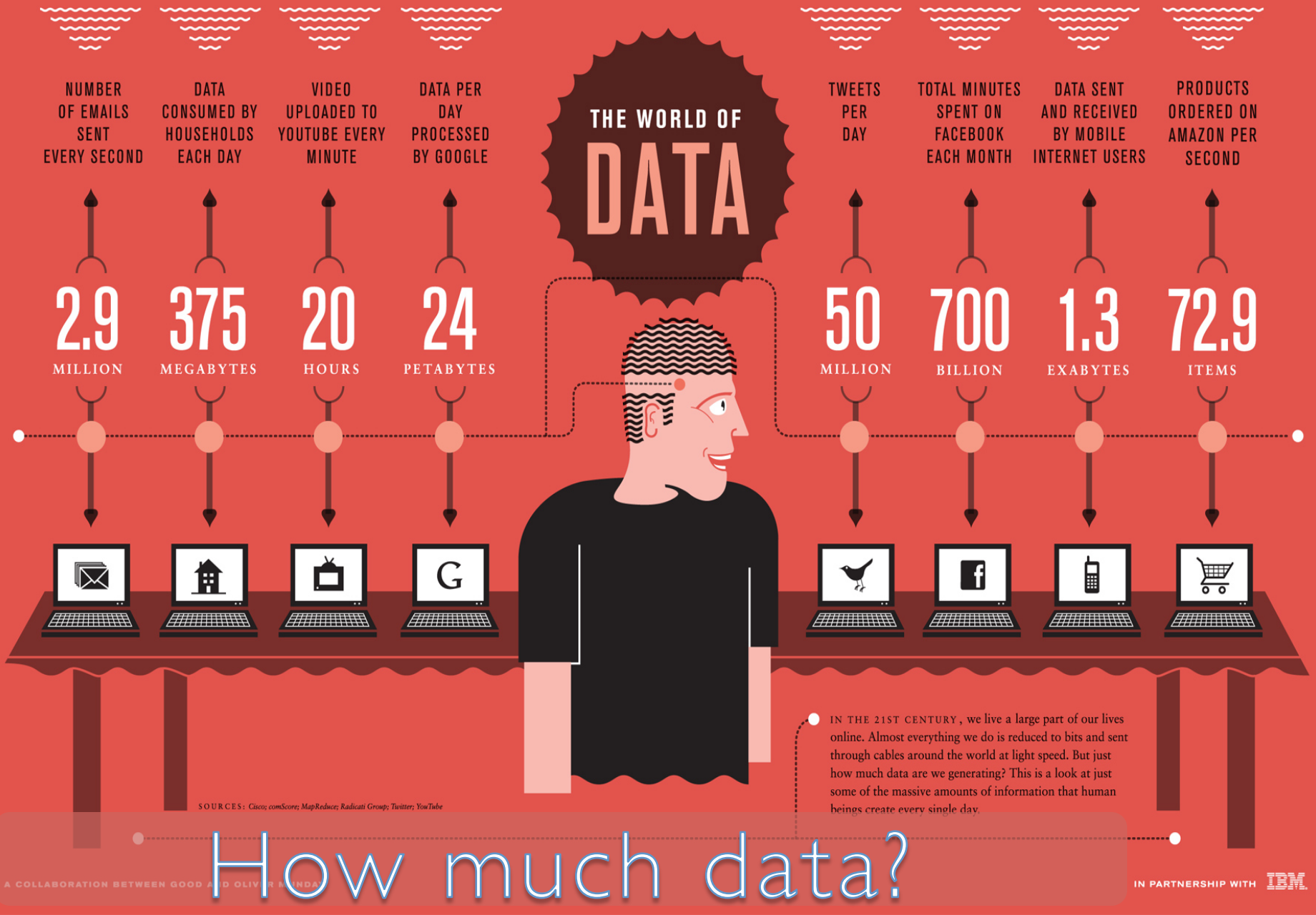
    

    • ***Many more…***

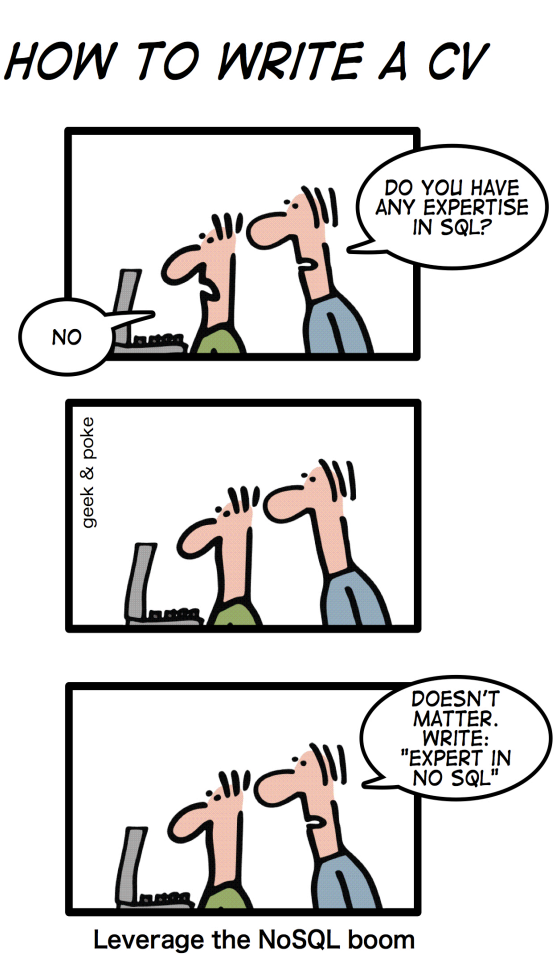
- *Not only Relational Database Management Systems & Business Intelligence*



How much data?

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

- NOSQL stands for "Not Only SQL" where SQL doesn't really mean the query language, but instead it denotes relational DBMS.
- Google, Facebook, Linkedin, eBay, Amazon, etc. did not use 'traditional' RDBMS for Big Data. They need:
  - To perform a massive number of Simple Operations very quickly on a variety of data types
- They inspired many NOSQL systems:
  - Memcached demonstrated that in-memory indexes can be highly scalable, distributing and replicating objects over multiple nodes
  - Dynamo (Amazon) pioneered the idea of *eventual consistency* as a way to achieve higher availability and scalability
  - BigTable, HDFS (Google), demonstrated that persistent record storage could be scaled to thousands of nodes
  - Map-Reduce (Google) paradigm for parallel processing



Technologies for Big Data

- Managing Big Data
- Analyzing Big Data

Moving the computation near the data

- Moore's Law has held firm for over 40 years
  - processing power doubles every two years
  - Processing speed is no longer the problem

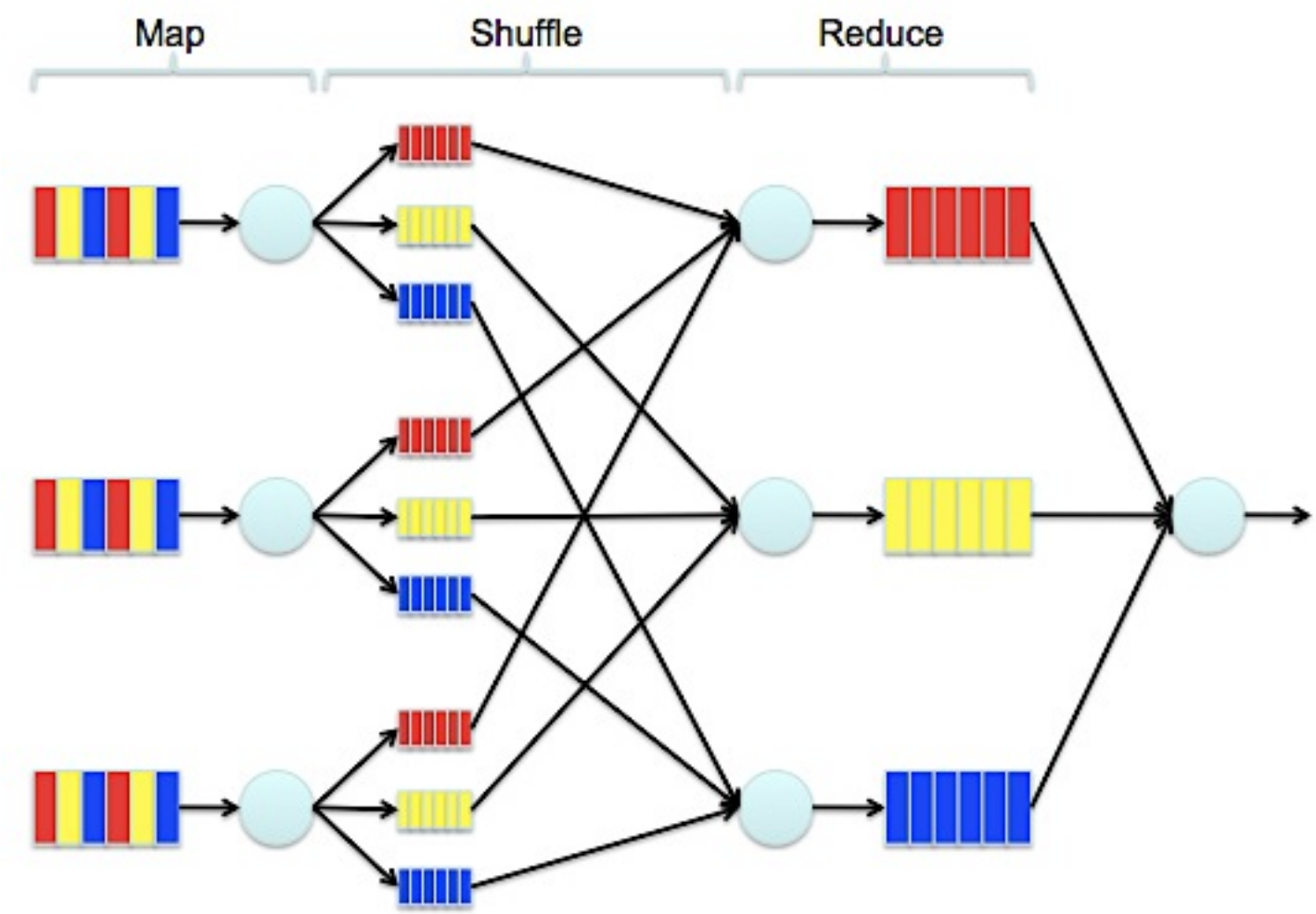
- Getting the data to the processor becomes the bottleneck

- Quick calculation:
  - Typical disk data transfer rate: 75MB/sec
  - Time taken to transfer "only" 100GB of data to the processor: ~ 22minutes !
    - Actual time will be worse, if servers have less than 100GB of RAM available

- MapReduce (invented by Google) solution: move the computation near the data

note that often the data transfer over the network is still the bottleneck!

MapReduce



http://www.slideshare.net/lynnlangit/hadoop-mapreduce-fundamentals-21427224

Big Data Solutions: Big Vendors Business Intelligence Platforms

BIG VENDORS: Oracle, IBM, Teradata, Sap, HP, Microsoft,…

Data Warehouse Appliance (DWA)

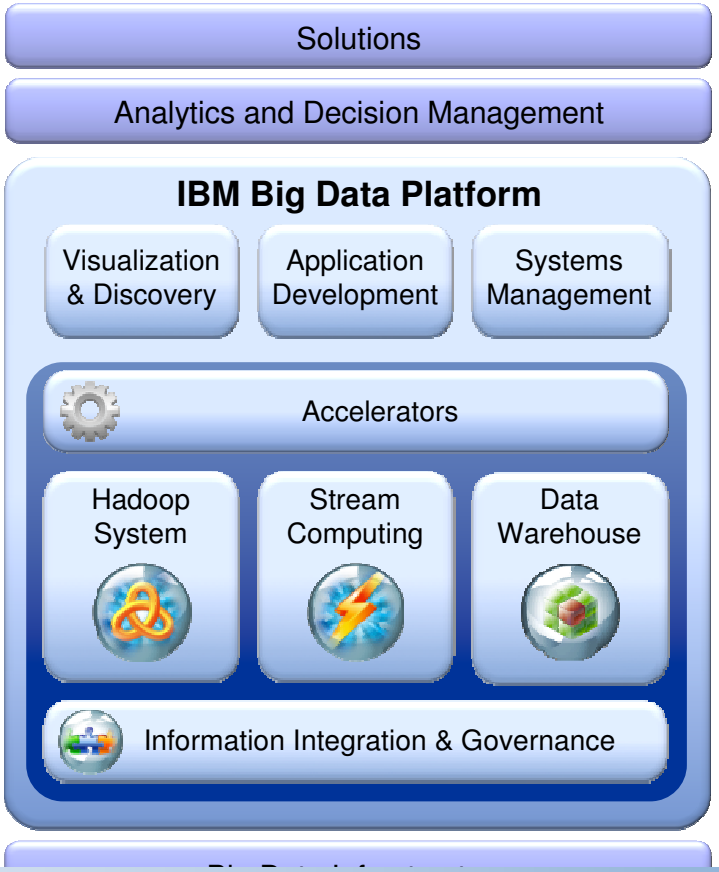
A new category of computer architecture for data warehousing (DW) specifically targeted for Big Data Analytics and Discovery that is:

- simple to use (not a pre-configuration) and very high performance for this workload.
- A DWA includes an integrated set of servers, storage, operating system(s), and DBMS.
- New Database Solutions (based on: exploiting main memory, combined row and column databases, enforcing MPP (Massive Parallel Processing))

Appliance: Netezza

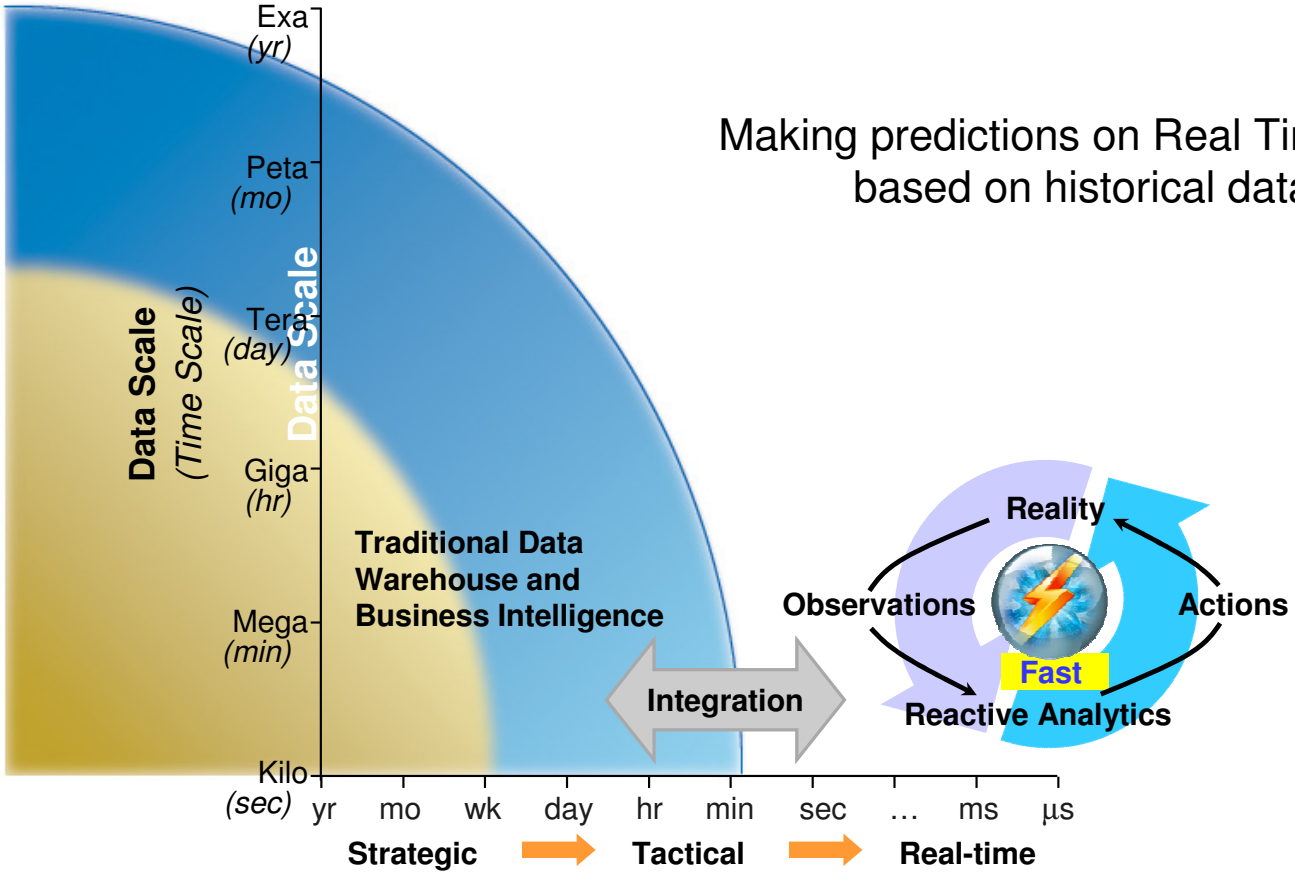
The IBM Big Data Platform

- Process any type of data
  - Structured, unstructured, in-motion, at-rest
- Built-for-purpose engines
  - Designed to handle different requirements



- Analyze data in motion
- Manage and govern data in the ecosystem
- Enterprise data integration
- Grow and evolve on current infrastructure

Combining Deep and Reactive Analytics



Making predictions on Real Time data based on historical data

Le Nuove Sfide Tecnologiche

- Cognitive Computing

Le nuove sfide tecnologiche

- Un *Cyber- Physical System (CPS)* è un sistema controllato o monitorato da algoritmi basati su computer, strettamente integrati con Internet e i suoi utenti.
  - ✓ (*Internet of Things – IOT*)

- *Cloud Computing* (in italiano **nuvola informatica**) indica l' erogazione di risorse informatiche pre-esistenti e configurabili, come l'archiviazione, l'elaborazione o la trasmissione di dati, disponibili *on demand* attraverso Internet.
-

- *Cognitive Computing* è la tecnologia che ci consentirà di interagire con i computer praticamente "parlando" alle macchine e sfruttando la loro capacità di imparare dall'esperienza (**Artificial Intelligence, Machine Learning**).

✓ I vantaggi principali si avranno in tutti I campi in cui è necessario elaborare contemporaneamente e velocemente grandi quantità di dati (*BIG Data*) di formati diversi (***Big Data Integration***).

Challenges(2) Cognitive Computing–Turing Test

Alan Turing

Turing Award

Industria 4.0

Industria 4.0: La 4° rivoluzione industriale

1° Rivoluzione industriale

Utilizzo di macchine azionate da energia meccanica

Introduzione di potenza vapore per il funzionamento degli stabilimenti produttivi

Fine 18° secolo

2° Rivoluzione industriale

Produzione di massa e catena di montaggio

Introduzione dell'elettricità, dei prodotti chimici e del petrolio

Inizio 20° secolo

3° Rivoluzione industriale

Robot industriali e computer

Utilizzo dell'elettronica e dell'IT per automatizzare ulteriormente la produzione

Primi anni '70

4° Rivoluzione industriale

Connessione tra sistemi fisici e digitali , analisi complesse attraverso Big Data e adattamenti real-time

Utilizzo di macchine intelligenti, interconnesse e collegate ad internet

Oggi - prossimo futuro

- www.sviluppoeconomico.gov.it/images/stories/documenti/Industria_40%20_conferenza_21_9

Artificial Intelligence: Veicoli a Guida Autonoma

Si prevede l'automazione dell'intera catena di approvvigionamento: navi da carico, porti, camion, magazzini, consegna, ...

Machine Learning 2016: AlphaGo batte il campione di Go Lee Sedol

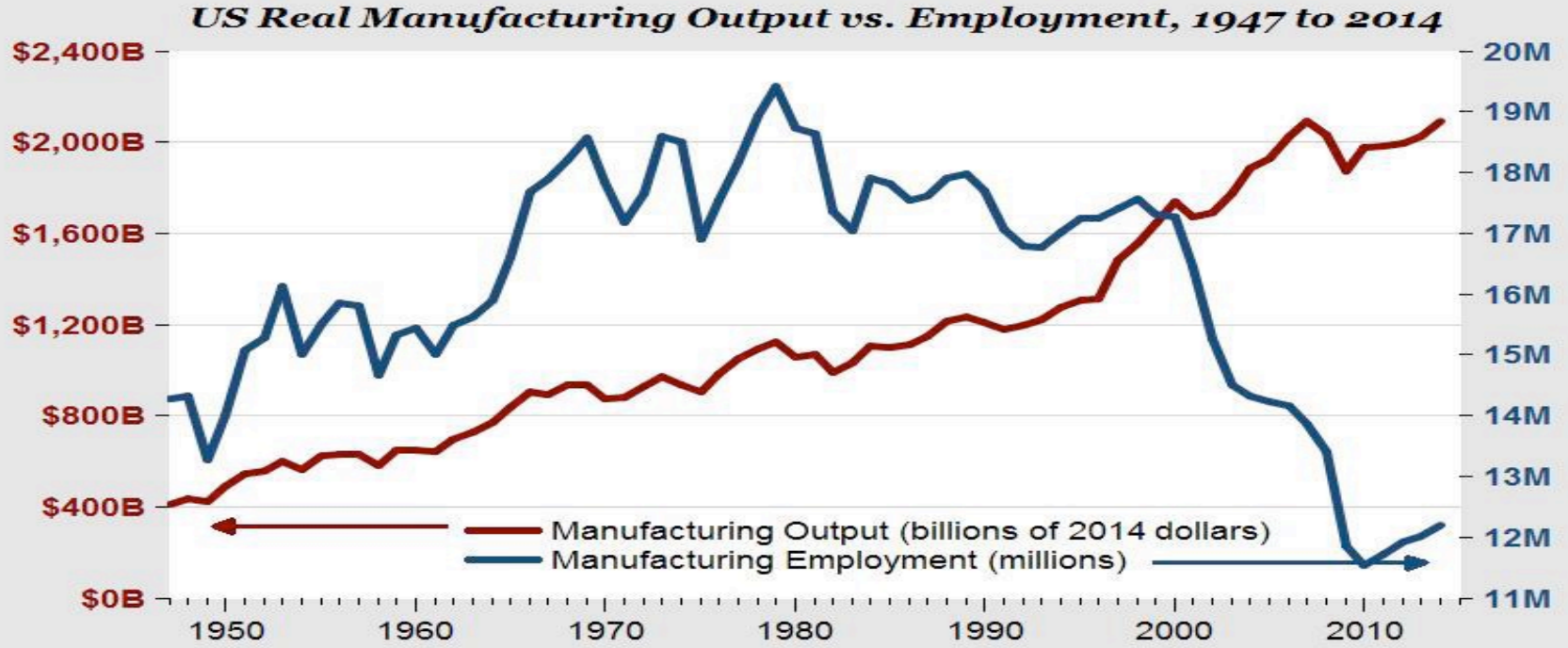
Grazie a tecniche di *Machine Learning* AlphaGo può sviluppare «intuizioni» per il gioco del Go.

Challenges (4) Evoluzione del Mondo del Lavoro

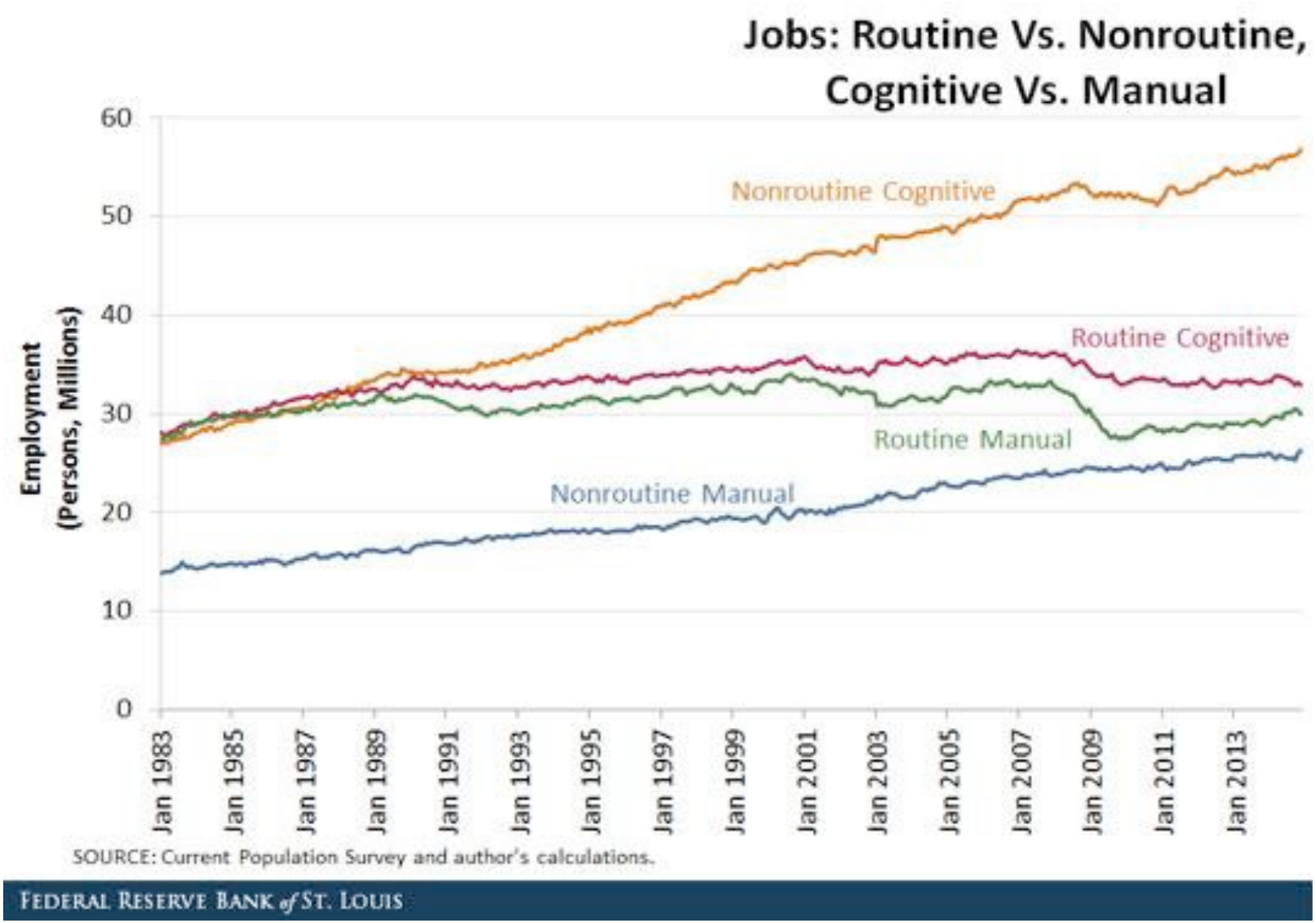
McKinsey: 45% dei posti di lavoro saranno sostituiti da tecnologia già disponibile

Gartner: 1 posto di lavoro su 3 sarà sostituito dalla tecnologia nel 2025

Ma nuovi lavori verranno creati dalla tecnologia



Lavori ripetitivi vs. non ripetitivi, intellettuali vs. manuali



Contribution of the DBGROUP

DBGROUP? Research, Dissemination & Teaching on Big Data

- Research: Big Data Management and Analysis, Big Data Integration (see www.dbgroup.unimore.it)

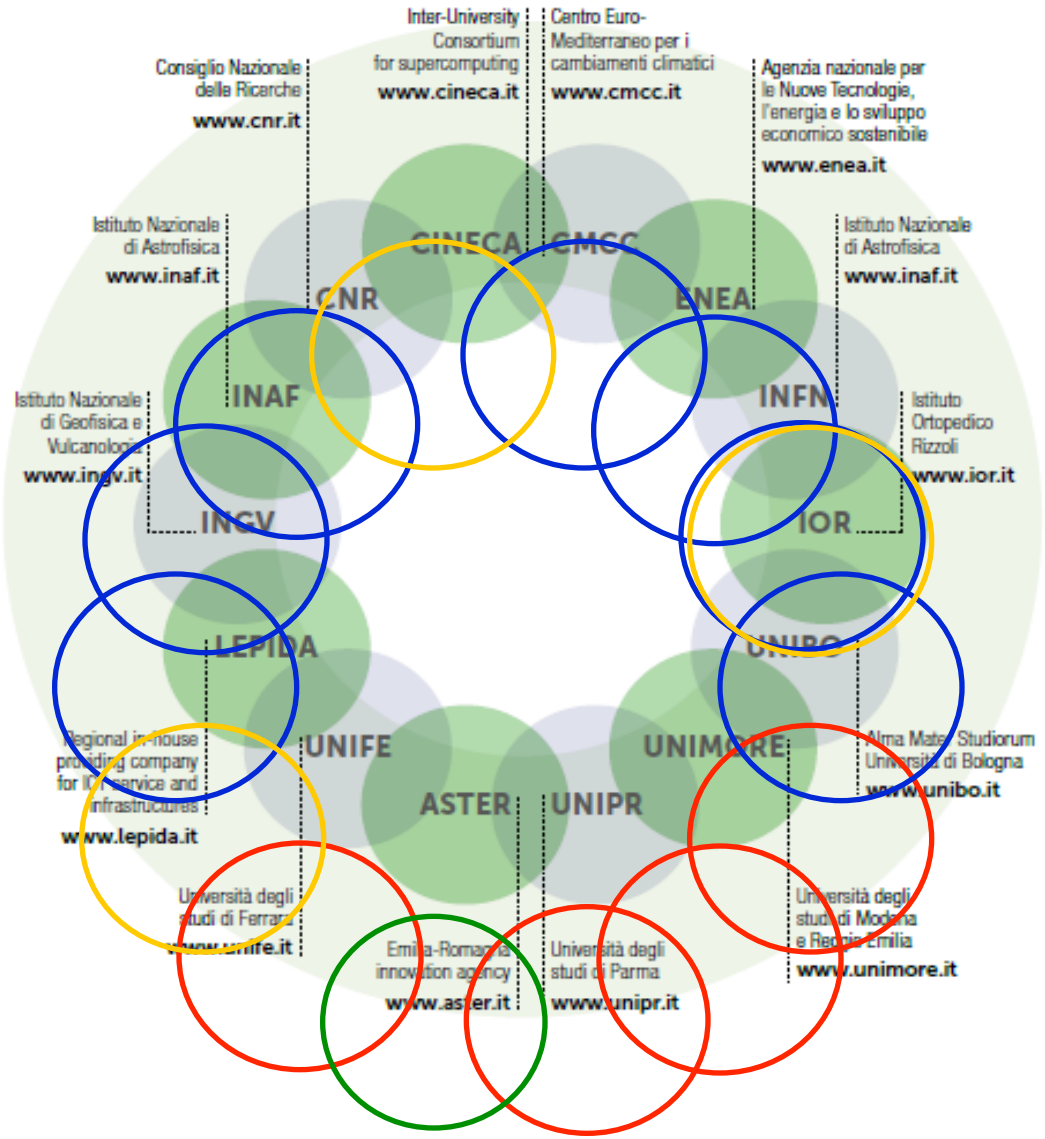
- Scientific Dissemination
  - "Big Data panel" at SEBD 2016; invited paper at CLADAG 2015 (8-10 October 2015)
  - Workshop "PICO: the CINECA solution for Big Data management" @ headquarters of Casalecchio on December 5th 2014;
  - IC3K 2014 (http://www.ic3k.org/KeynoteSpeakers.aspx) – lecture title"Big Data integration – State of the Art &Challenges" – Roma 21-24 October 2014;
  - 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.

Academic & Advanced Training COURSES

- Corsi: "Data Management and Governance", "Big Data Analysis" , Rappresentazione della Conoscenza"–Laurea Magistrale Ingegneria Informatica.
- *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.

Cosa fa la Regione Emilia Romagna sui 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

DBGROUP & Regione Emilia Romagna

- Academy "Metodologie, tecniche e tool per l'analisi dei Big Data"
  - **DBgroup** Università di Modena e Reggio Emilia & **CINECA**
  - 50 ore di formazione:
    - 30 ore didattica
    - 20 ore laboratorio big data con infrastruttura CINECA
  - 20 posti disponibili
  - 10 borse di studio da 1,500 €
  - Iscrizione: 3,000 €

- Assegni a favore delle imprese modenesi (bandi a breve)
  - **BPER**: "Big Data e Analytics per lo sviluppo del comportamento digitale del cliente da prospect ad acquisito "
  - **DOXEE**: "Metodologia di progettazione di applicazioni sui Big Data basata su tecnologia Amazon Web Services "
  - **Expert System**: "Data Scientist per supportare il processo di produzione di intelligence (Corporate Intelligence Data Scientist)"

  Delibera n. 554 del 28/04/2017

Cosa fanno i dipartimenti di Ingegneria di Modena e Reggio Emilia ?

- Conferenze Internazionali
- Alta Formazione
  - Corsi di Laurea, di Laurea Magistrale
- Altissima Formazione
  - Corsi di dottorato di ricerca di base, di ricerca industriale, In alto apprendistato

(da coordinatore del corso di dottorato in ICT www.ict.unimore.it contattatemi per informazioni)

Conferenze Internazionali promosse dal DIEF

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

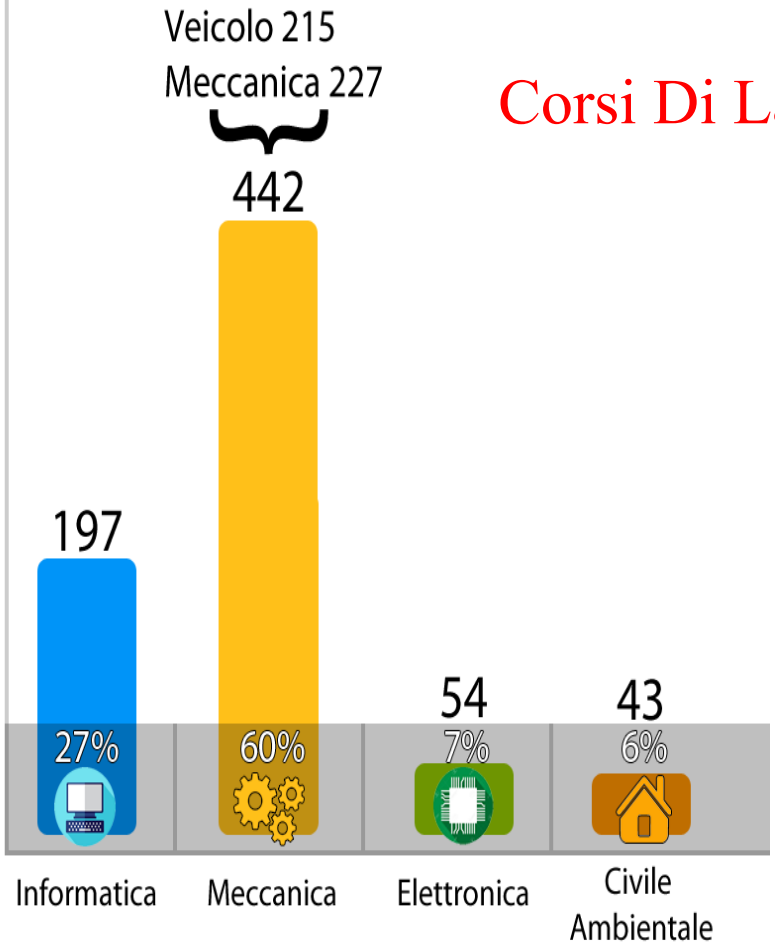
UNIVERSITAS STUDIORUM MUTINENSIS ET REGIENSIS 1175

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

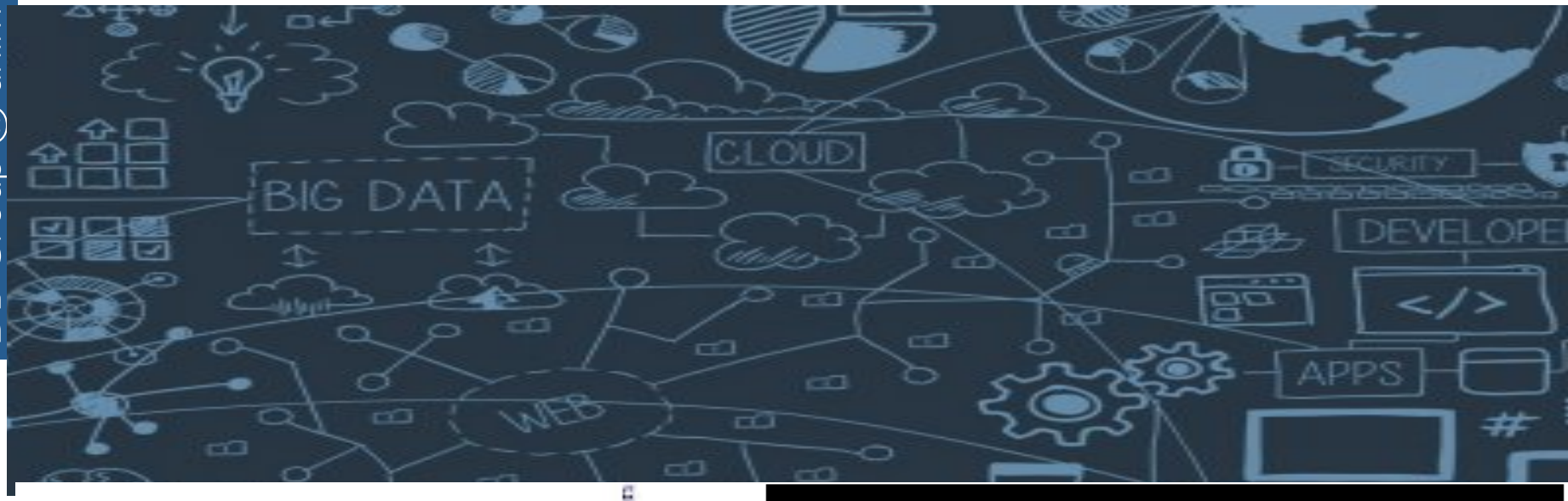
Industry 4.0 Smart healthcare Smart mobility

Corsi Di Laurea Triennali

Dati in divenire

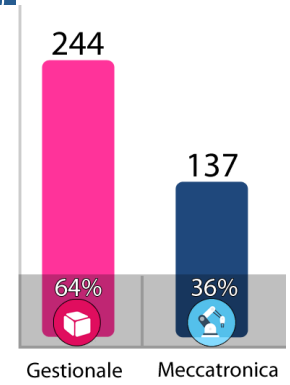


Corsi offerti alla Laurea Magistrale in Ingegneria Informatica

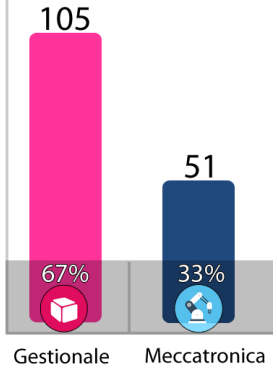


Immatricolazioni

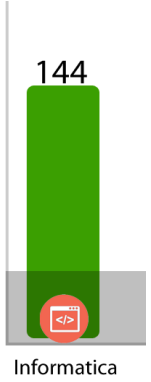
•2016/2017



Triennali



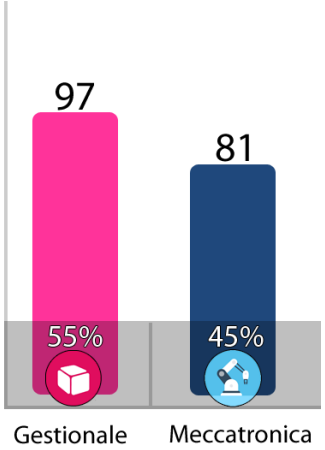
Magistrali



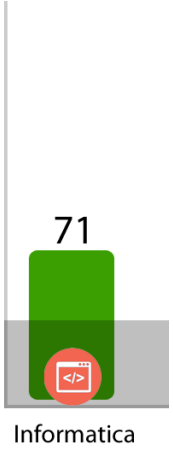
Triennale

Magistrale

•2017/2018



Triennali



Triennale

*dato temporaneo – le immatricolazioni terminano a **dicembre 2017***

Progetto Ragazze Digitali

- **4° Edizione *(dal 2014)***
- 12 Giugno – 7 Luglio 2017

- ***Summer camp gratuito di 4 settimane***
- Dedicato alle ***studentesse*** di 3° e 4° superiore

- Attività di laboratorio – ***Learn by doing***
  - Sviluppo di videogiochi in Python
  - Team working

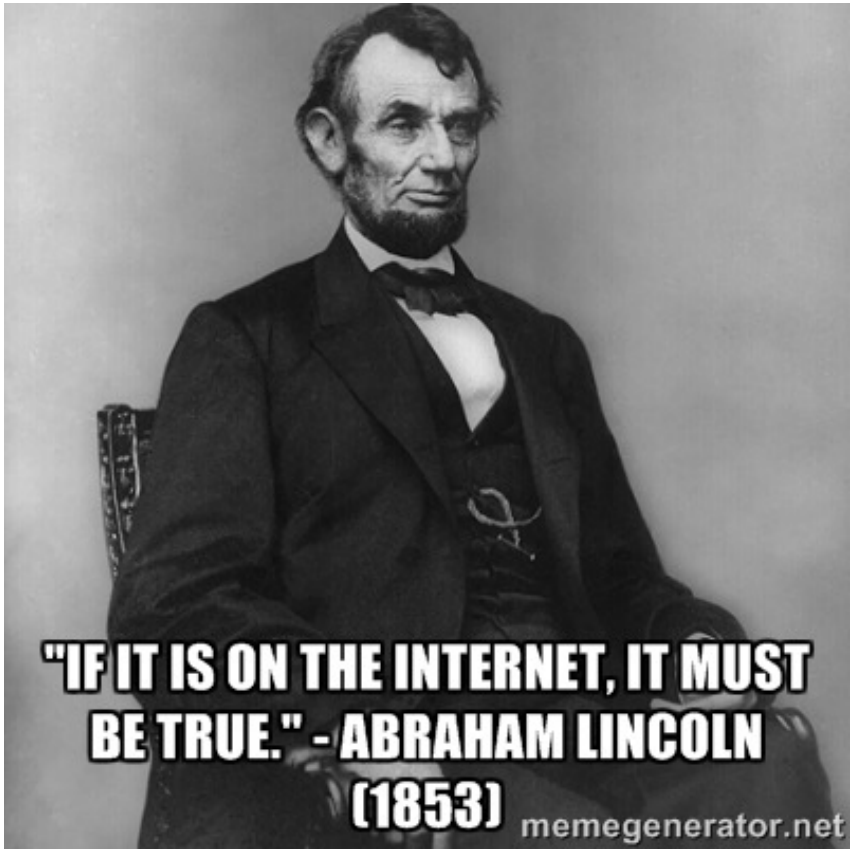
- ***Role model femminili*** in campo ICT
- ***#teamwork #creativity #coding #empowerment***

Grazie!

E gli aspetti etici del trattamento dei dati?

Issues in Ethical Data Management

Serge Abiteboul



Promises and risks of massive data

- Improve people's lives, e.g., recommendation
- Accelerate scientific discovery, e.g., medicine
- Boost innovation, e.g., autonomous cars
- Transform society, e.g., open government
- Optimize business, e.g., advertisement targeting

Growing resentment

- Against bad behaviors: racism, terrorist sites, pedophilia, identity theft, cyberbullying, cybercrime
- Against companies: intrusive marketing, cryptic personalization and business decisions
- Against governments: NSA and its European counterparts

Increasing awareness of the dissymmetry between what these systems know about a person, and what the person actually knows

Future challenges in data management

An opinion:

- In the past, the field was driven by
  - Company data
  - Data model & performance & reliability
- In the future
  - **Personal and social data**
  - **Ethical issues**

Ethics: concepts and principles that guide us in determining what behavior helps or harms us

GRAZIE DELL'ATTENZIONE

GLI ASPETTI ETICI DEL TRATTAMENTO DEI BIG DATA MERITANO UN'ALTRA RELAZIONE