



UNIMORE
 UNIVERSITÀ DEGLI STUDI DI
 MODENA E REGGIO EMILIA

Il ruolo dei dati nella trasformazione della società

INTELLIGENZA ARTIFICIALE, LAVORO, IMPRESA
 TRA APPLICAZIONI PRATICHE E REGOLAZIONE EUROPEA Tavola rotonda
 Dipartimento di Giurisprudenza – Aula Magna 8 giugno 2022, ore 14.30 – 18.30
Prof. Sonia Bergamaschi

Department of Engineering “Enzo Ferrari”

sonia.bergamaschi@unimore.it

www.dbgroup.unimore.it



- Chi sono ?
- Dati e Nuove Tecnologie Informatiche: Cyber Physical Systems, Cloud Computing, Cognitive Computing (o Artificial intelligence)
- Big Data, Big Data Integration, Scienza dei Dati
- Il valore economico dei dati
- Quale sarà l'impatto delle nuove tecnologie sull'occupazione?
- Issues in Ethical Data Management
- Privacy e uso dei dati, come fare?

Prof. Sonia Bergamaschi

- Professore Ordinario
dip. Ingegneria “Enzo Ferrari – Università
Di Modena e Reggio Emilia (UNIMORE)
- Email: sonia.bergamaschi@unimore.it
- Leader del Database Research Group
([DBGGroup](#))
- Direttore del dottorato in ICT [ICT Doctorate](#)
- Delegato ai Sistemi Informativi e Informatici (UNIMORE)
- ACM Distinguished Researcher
- IEEE Senior Member
- >300 pubblicazioni in conferenze e riviste internazionali
su gestione, integrazione analisi di Big Data
[DBLP](#) · [Google Scholar](#) · [Scopus](#)



- Membri:
 - 6 Faculty
 - [Prof. Sonia Bergamaschi](#)
 - [Prof. Domenico Beneventano](#)
 - [Prof. Francesco Guerra](#)
 - [Prof. Laura Po](#)
 - [Prof. Maurizio Vincini](#)
 - [Giovanni Simonini, PhD](#) (RTDB)
 - 1 Postdoc
 - [Luca Gagliardelli, PhD](#) (RTDA)
 - 4 ICT PhD Students
 - Luca Zecchini (2nd year, *Task-driven Big Data Integration*)
 - Adeel Aslam (1st year, *Big Data and Artificial Intelligence for the enhancement of energy virtuosity*, ER Grant)
 - Giulio De Sabbata (1st year, *Big Data and Artificial Intelligence to the efficiency of production processes in industrial manufacturing*, DataRiver)
 - Ambra Di Piano (1st year, *Deep learning in real-time on the astrophysical data obtained from the Cerenkov CTA Observatory*, INFN)
- Referente UNIMORE del [CINI Big Data Lab](#)
- Co-fondatrice di [DataRiver](#): ex spin-off (ora impresa innovativa) distributrice della versione open source del sistema di integrazione dati MOMIS





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

- *Un Cyber Physical System* (CPS) o *IOT* è un sistema controllato o monitorato da algoritmi basati su computer, strettamente integrato con Internet e i suoi utenti.
 - componenti fisici e software interagiscono (sistemi di guida autonoma, sistemi robotici, monitoraggio medico, ecc.)
- *Cloud Computing* (in italiano **nuvola informatica**) indica la fornitura di risorse informatiche preesistenti e configurabili, quali l'archiviazione, l'elaborazione o la trasmissione di dati, disponibili on demand tramite Internet.
- *Cognitive Computing* Cognitive Computing è la tecnologia che ci permetterà di interagire con i computer praticamente "parlando" alle macchine e sfruttando la loro capacità di apprendere dall'esperienza (Intelligenza Artificiale/ Machine Learning).
- *Big Data Integration* è la tecnologia che ci permette di integrare informazioni relativamente allo stesso oggetto del mondo da sorgenti dati eterogenei e distribuiti.

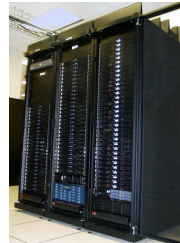
Big Data



PC



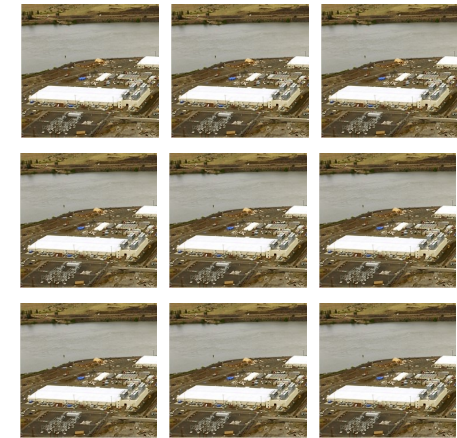
Server



Cluster



Data center



Network of data centers

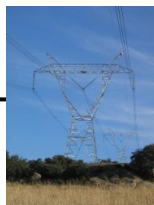
Google data center location (inferred):



- Analogia delle centrali elettriche



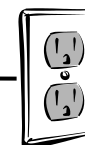
Power source



Network



Metering device



Customer

- Il cloud computing è un modello per consentire un comodo accesso di rete su richiesta a un pool condiviso di risorse informatiche configurabili (ad es. reti, server, storage, applicazioni e servizi) di cui è possibile eseguire rapidamente il provisioning e il rilascio con il minimo sforzo di gestione o fornitore di servizi interazione
- Caratteristica:
 - Self service su richiesta
 - Servizio misurato
 - Rapida elasticità (le risorse possono essere fornite in modo rapido ed elastico per aumentare rapidamente e rilasciarle rapidamente per ridurle rapidamente)
 - Condivisione delle risorse

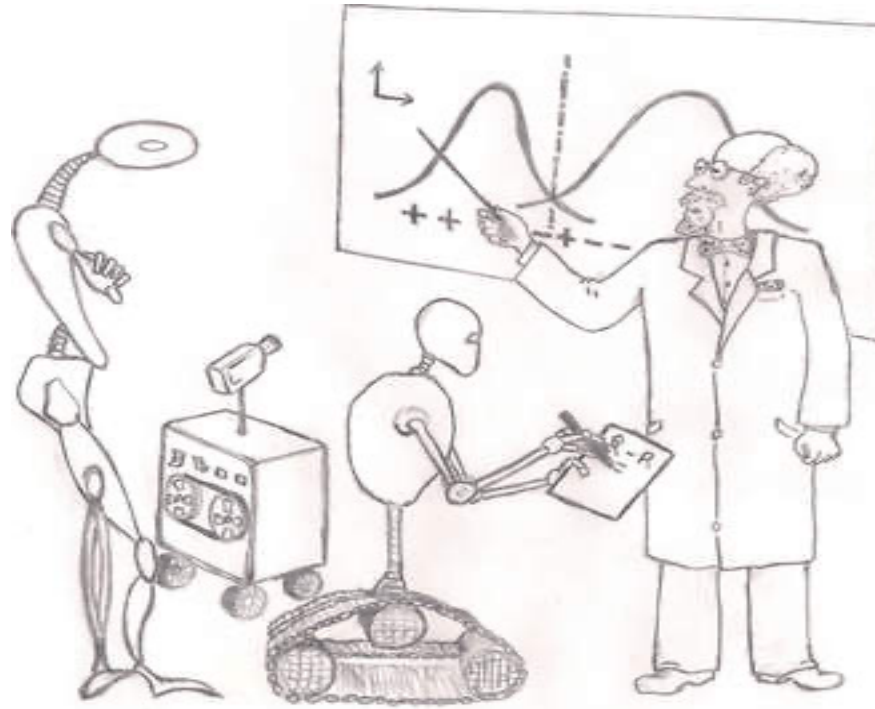
Cognitive Computing –Turing Test



Alan Turing



Turing Award



Machine learning gives computers the ability to learn without being explicitly programmed (Arthur Samuel, 1959) .. Now algorithms which can learn and make predictions on data.

Big Data – Piena fiducia nel potere dei dati

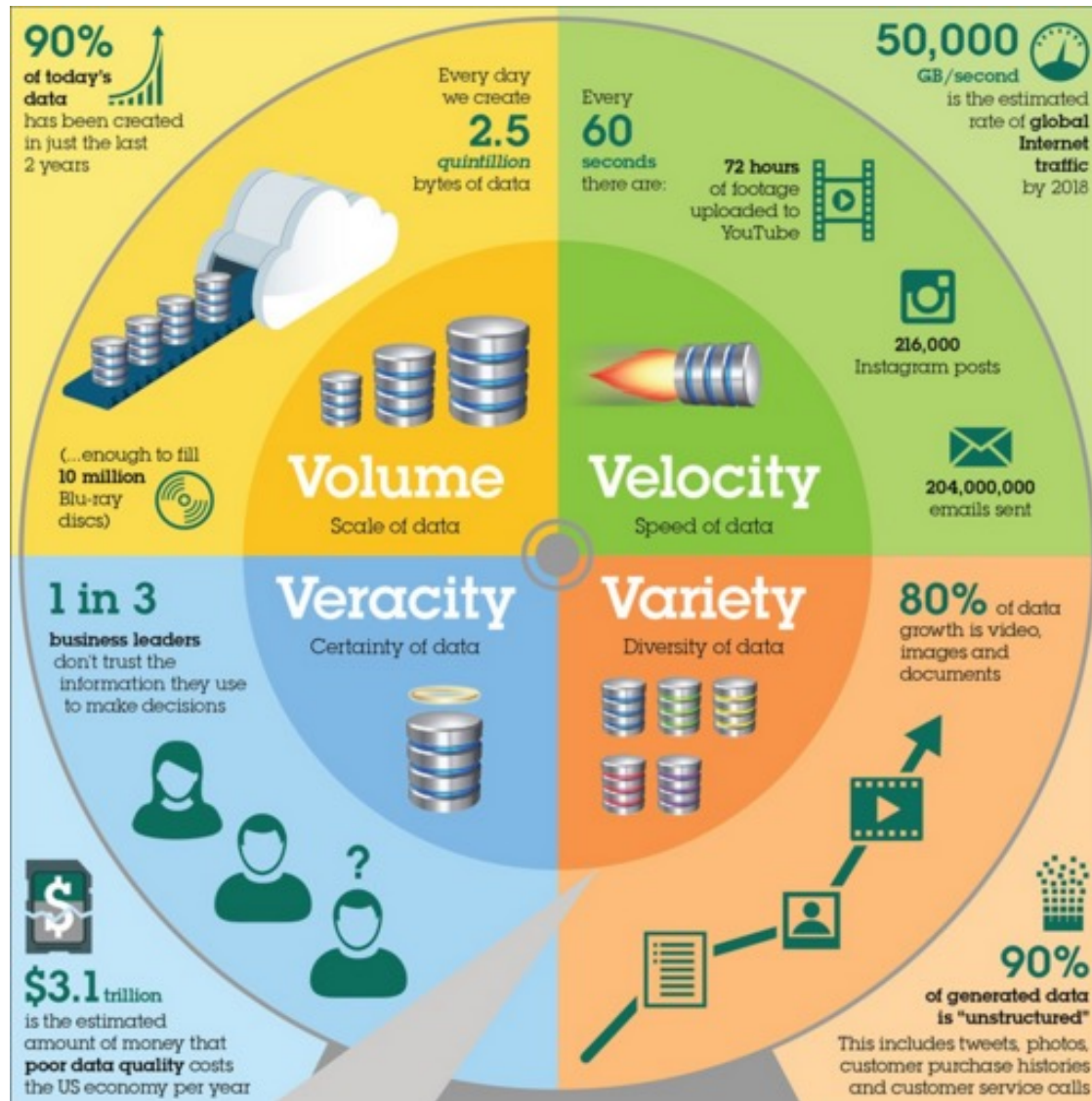
Dati = nuovo petrolio

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



1 Petabyte è l'equivalente di 20 milioni di schedari alti o 500 miliardi di pagine di testo stampato standard.

The FOUR V's of Big Data



The production of data is expanding at an astonishing pace. Experts now point to a 4300% increase in annual data generation by 2020. Drivers include the switch from analog to digital technologies and the rapid increase in data generation by individuals and corporations alike.

2020: MORE THAN 1/3 OF THE DATA PRODUCED WILL LIVE IN OR PASS THROUGH THE CLOUD.

Size of Total Data
Enterprise Created Data
Enterprise Managed Data

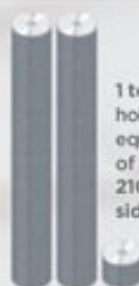
Only 0.5% to 1% of the data is used for analysis.

2012: CUSTOMERS WILL START STORING 1 EB OF INFORMATION.



WHAT IS A ZETTABYTE?

1,000,000,000,000	gigabytes
1,000,000,000,000,000	terabytes
1,000,000,000,000,000,000	petabytes
1,000,000,000,000,000,000,000	exabytes
1,000,000,000,000,000,000,000,000	zettabyte



1 terabyte holds the equivalent of roughly 210 single-sided DVDs.

It took roughly 1 petabyte of local storage to render the 3D CGI effects in Avatar.



In 2007, the estimated information content of all human knowledge was 295 exabytes.

DATA PRODUCTION WILL BE 44 TIMES GREATER IN 2020 THAN IT WAS IN 2009

More than 70% of the digital universe is generated by individuals. But enterprises have responsibility for the storage, protection and management of 80% of it.*

Valore – La V più importante di tutte!



La quinta V : Valore!

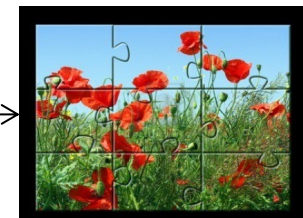
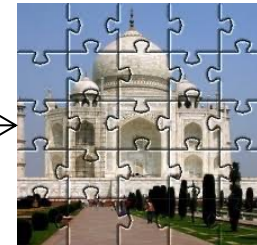
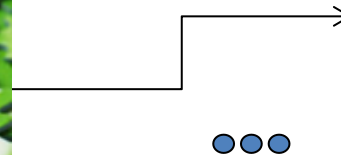
- Avere accesso ai big data è inutile a meno che non possiamo trasformarli in valore.

Per estrarre Valore



- **(Big) Data Integration**
- **Data Analysis (including Cognitive Computing)**

- La Data Integration comprende le pratiche, le tecniche e gli strumenti dell'architettura che acquisiscono, trasformano, combinano e forniscono dati attraverso lo spettro dei tipi di informazioni nell'azienda e oltre, al fine di soddisfare i requisiti di consumo dei dati di tutte le applicazioni e i processi aziendali.
- Applicazioni di Data Integration
 - Affari, scienza, governo, Web, salute... praticamente ovunque
 - Data Integration = risolvere puzzles
 - Ogni puzzle (e.g., Taj Mahal) è un' **entità integrata**
 - Ogni pezzo di un puzzle proviene da qualche fonte dati



La torta della Scienza dei Dati



Ingredients:

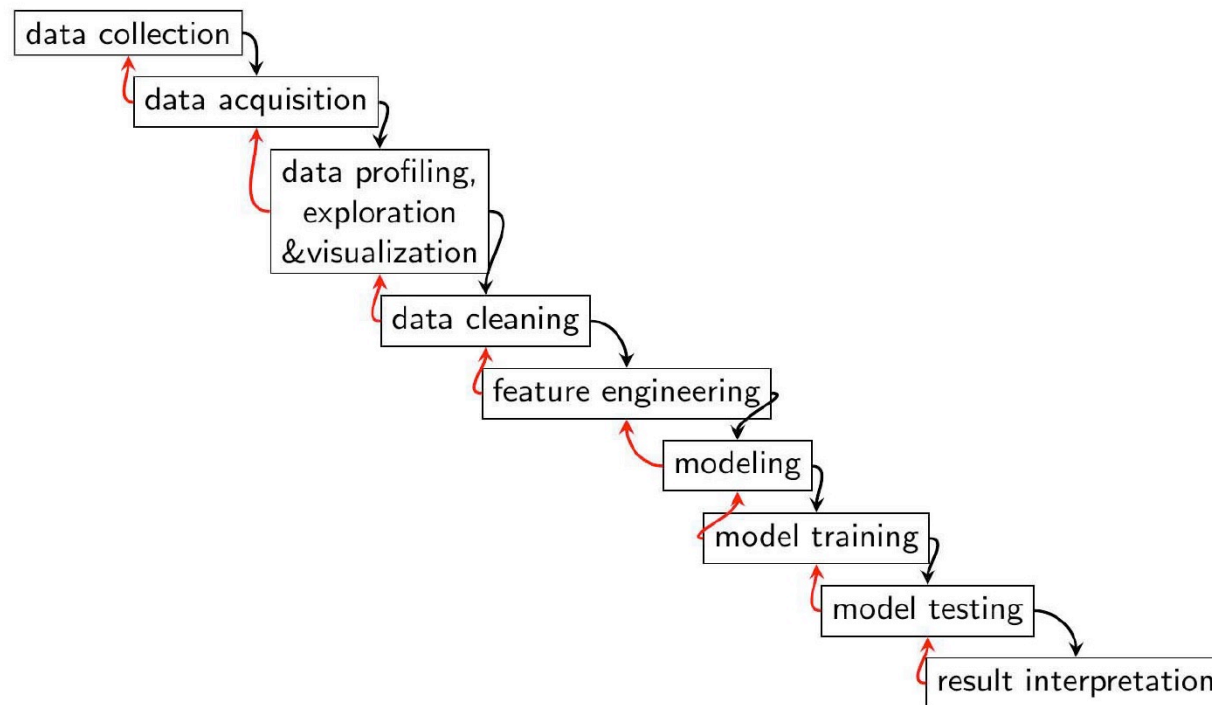
50g statistics
120g linear algebra
200g programming
1kg visualisation
300g software engineering

Additional skills:

creativity
out of the box thinking
grit
team spirit

© [istock.com](https://www.istock.com) sasilsolutions

Il modello di pipeline/cascata della scienza dei dati



This is **at the same time** a process model **and** a dataflow.

From Jens Dittrich (Saarland University)

Data Integration +

Data Analysis

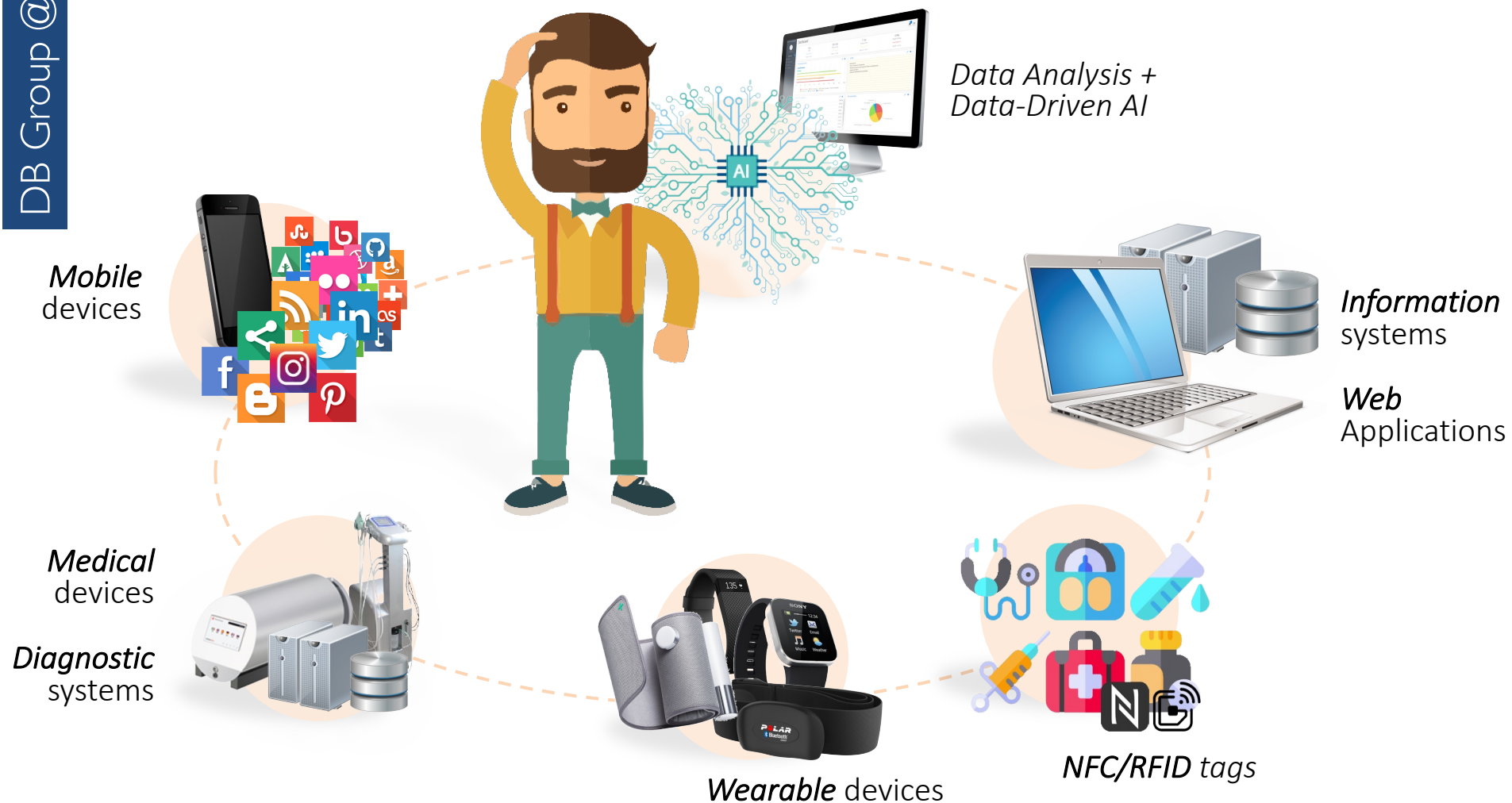
(Business Intelligence, Statistics, Data Mining, Math)

+

Data-Driven Artificial Intelligence

(i primi 4 task della pipeline sono indispensabili)

L'integrazione è un prerequisito per attuare la scienza dei dati: esempio Sanità



General Data Protection Regulation (GDPR)

European Data Market



Data workers

6.16 million in 2016



10.43 million by 2020



Data companies

255,000 in 2016



359,050 by 2020



Data economy value

Almost € 300 billion in 2016 → € 739 billion by 2020



European
Commission

Source: European Data Market study

Chi ha i dati in cloud per fare soldi?

40%

16%

10%

8%

7%

amazon

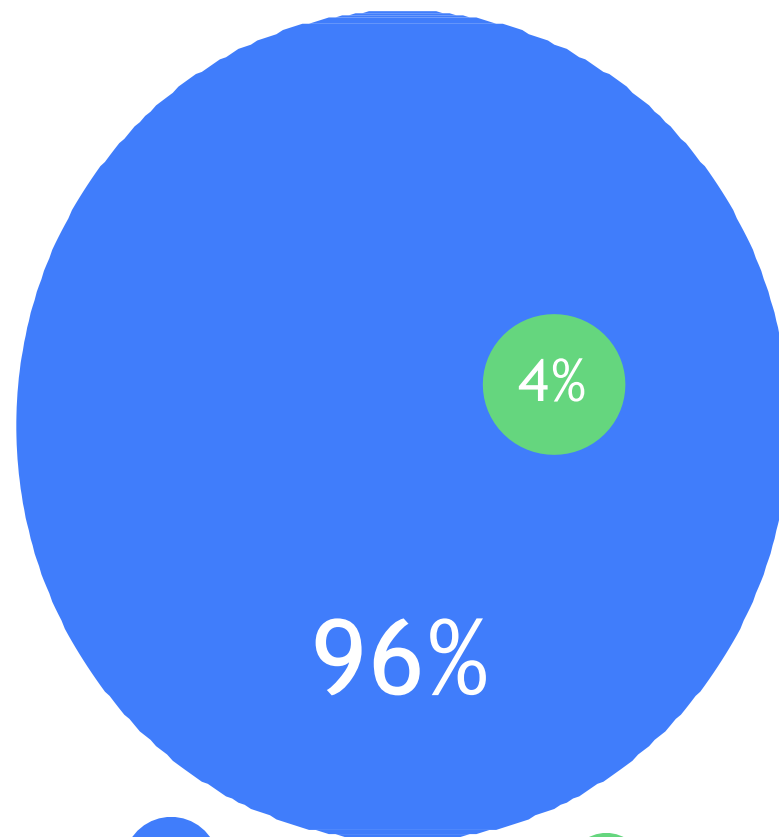
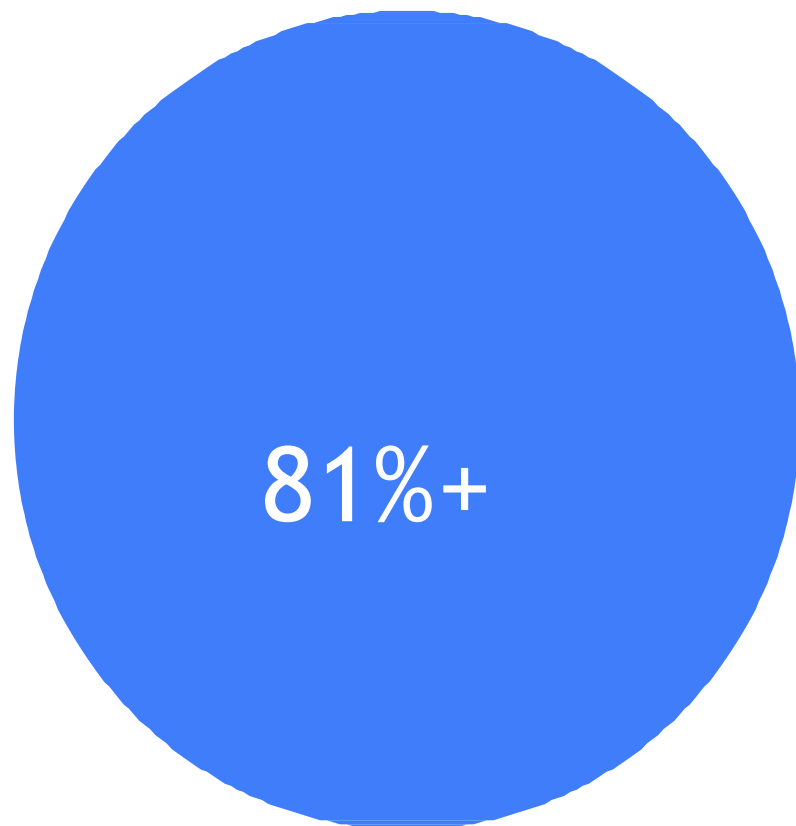
 Microsoft

IBM

Google

salesforce

Dove sono i dati?



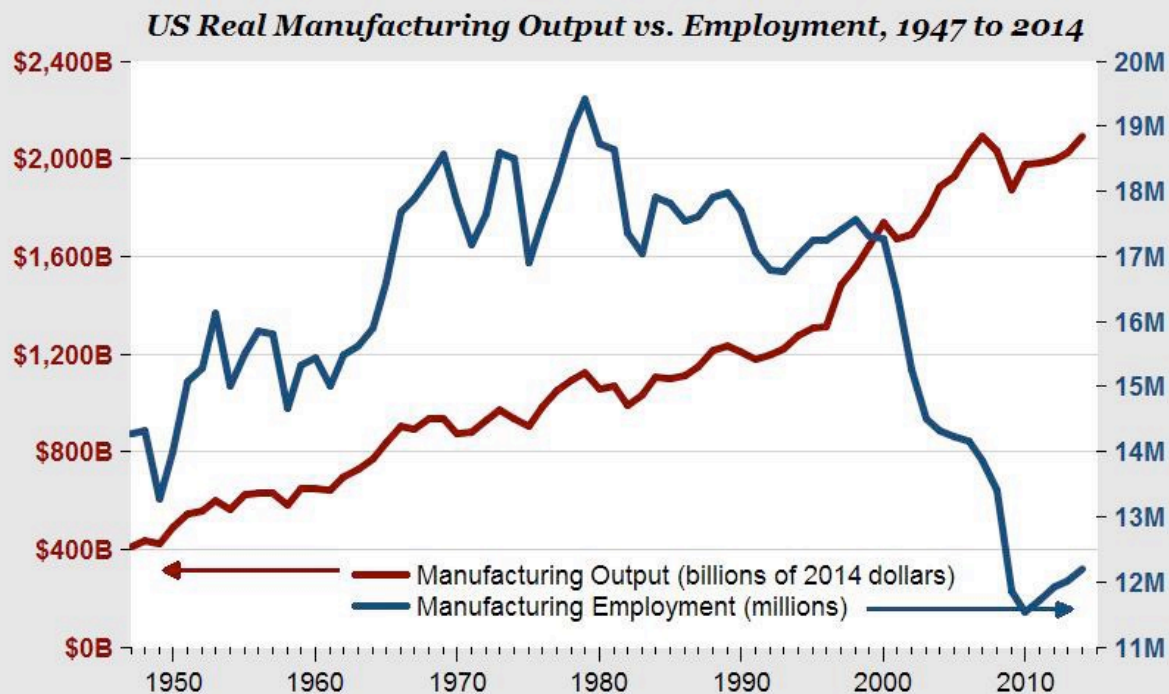
Foreign



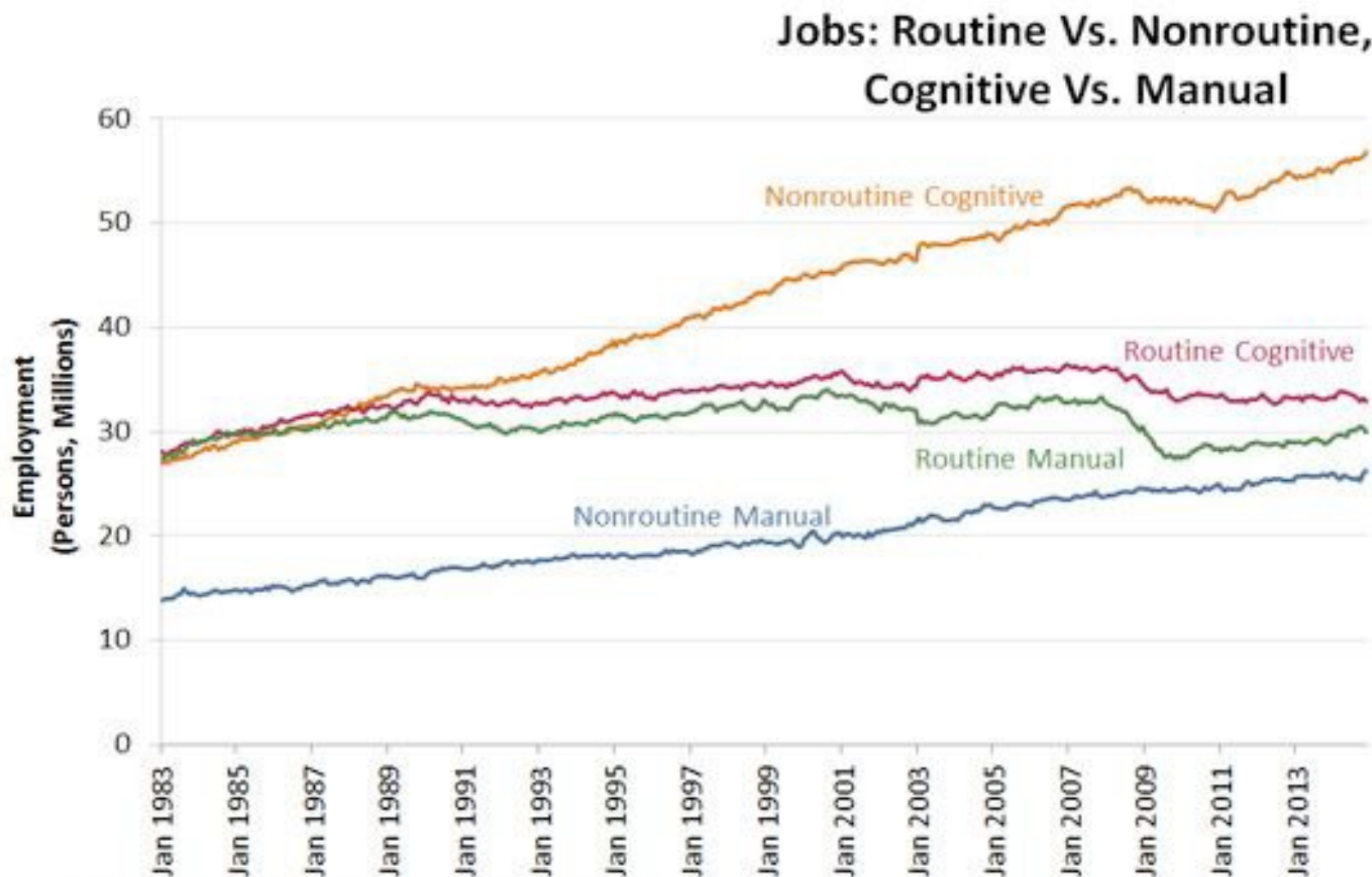
US Citizen

Sfide : Evoluzione del Mondo del Lavoro

- McKinsey: il 45% dei posti di lavoro sarà sostituito da tecnologia già disponibile
- Gartner: 1 lavoro su 3 sarà sostituito dalla tecnologia nel 2025
- ... Ma la tecnologia creerà nuovi posti di lavoro



Lavoro ripetitivo vs. non ripetitivo, intellettuale vs. manuale

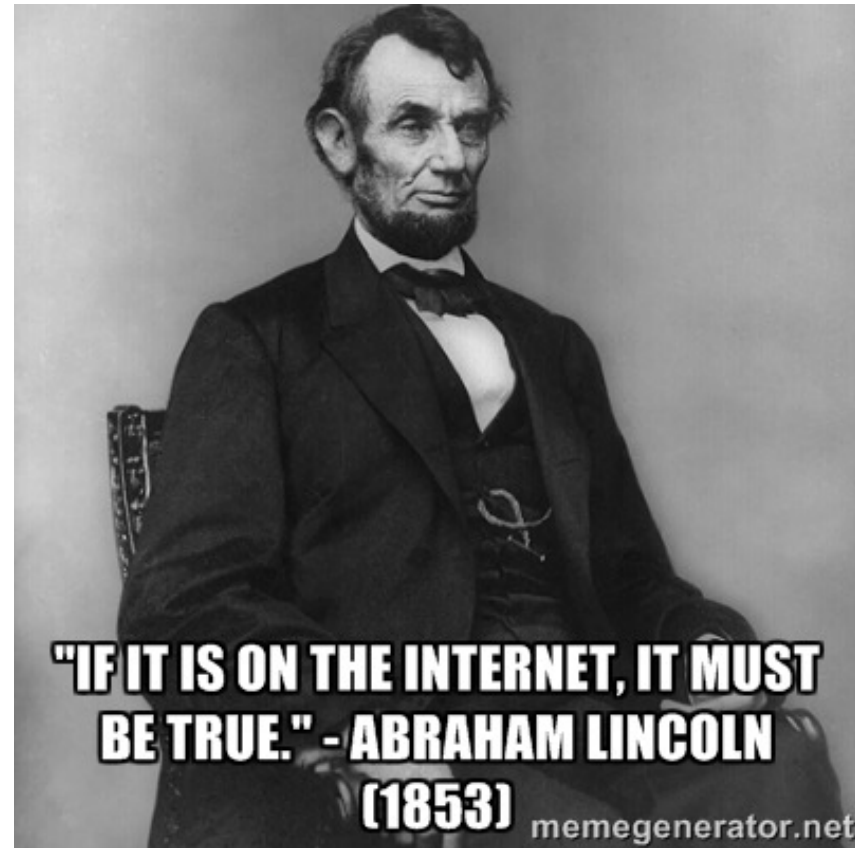


SOURCE: Current Population Survey and author's calculations.

FEDERAL RESERVE BANK of ST. LOUIS

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

General Data Protection Regulation (GDPR)

- In molti domini applicativi vengono creati dati personali sensibili sugli individui, ogni volta che questi dati devono essere integrati, è necessario considerare la privacy e la riservatezza.
- La protezione dei dati in Europa è sancita dal Regolamento generale europeo sulla protezione dei dati (GDPR) entrato in vigore a maggio 2018 ed è un quadro giuridico completo che definisce le linee guida per la raccolta e il trattamento delle informazioni personali delle persone che risiedono nell'Unione Europea (UE).
- **La misura appropriata per attuare in modo efficace i principi di protezione dei dati è la *Pseudonimizzazione*.**
- Ciò vale per l'uso di tecniche tolleranti di tutela della privacy per creare forme anonime dei dati da integrare (*pseudonimo*).



Classificazione di tipi di dati (con attenzione alla privacy)

I database possono contenere diversi tipi di dati che devono essere gestiti in modo diverso per garantire la protezione dell'identità e dei dati sensibili delle persone:

- **Informazioni di identificazione personale (PII):** uno/più attributi che costituiscono un identificatore univoco per ciascuna entità di una popolazione..

Es. SPID

- **Quasi-identifiers (QID):** un attributo che contiene informazioni che potenzialmente identificano i proprietari di record quando unite ad altre informazioni. Ex. nomi, date di nascita, indirizzi.

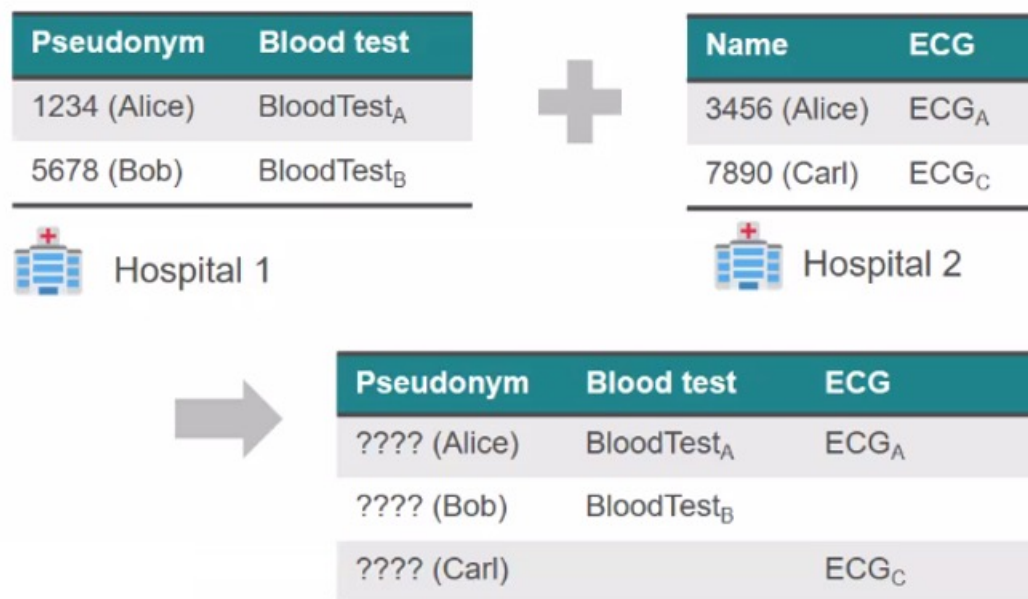
- **Dati Sensibili:** un attributo che rappresenta informazioni personali sensibili che devono essere protette dalla divulgazione della privacy. Ex. Malattia o reddito, religione o opinioni politiche.

- **Dati Non-Sensibili:** un attributo che contiene informazioni che non meritano protezione. Ex. metadati.

Classificare i dati in base all'identificabilità in uno scenario reale non è un compito facile perché diversi tipi di dati possono sovrapporsi.

Privacy-Preserving Record Linkage (PPRL)

Privacy-Preserving Record Linkage (PPRL) è il compito di identificare e collegare i **record (profili di entità)** che si riferiscono allo **stesso oggetto del mondo reale (entità)**, attraverso diverse fonti di dati detenute da parti diverse, in modo da impedire sia il calcolo e l'output del calcolo dalla rivelazione di informazioni riservate riservate sulle entità rappresentate nei dati.



Si basa sulla gestione dell'identità della protezione dei dati per la pseudonimizzazione specifica del contesto in cui deve essere istituito un **Istituto virtuale (VI)** che funge da dominio di integrazione per facilitare lo scambio di dati tra i sistemi in un ambiente informatico distribuito.

