LavoroDirittiEuropa
Rivista nuova di Diritto del Lavoro

**Numero 3 / 2022**

Sonia Bergamaschi

# Il ruolo dei Dati nella trasformazione della Società

# Il ruolo dei Dati nella trasformazione della Società

Sonia Bergamaschi

*Dipartimento di Ingegneria "Enzo Ferrari" Università di Modena e Reggio Emilia*

**Sommario**

L'obiettivo di questo breve intervento è quello di contribuire a definire il concetto di *Scienza dei Dati,* vero strumento di trasformazione della società che include tecniche di Intelligenza Artificiale ma deve avvalersi anche di altre importantissime nuove tecnologie informatiche e, in primo luogo, di tecniche di gestione, integrazione e arricchimento dei dati (il loro significato) provenienti da diverse sorgenti, eterogenee per tipo e affidabilità. La realtà è che la trasformazione della società può avvenire grazie alla enorme disponibilità di dati generata negli ultimi anni (i cosiddetti *Big Data*), dall'evoluzione delle tecnologie informatiche e anche (ma non solo) dai nuovi progressi dell'informatica nell'area dell'Intelligenza Artificiale, con particolare riferimento al *Machine Learning* (apprendimento automatico). Cercherò di illustrare il concetto di *Scienza dei Dati* e di introdurre gli ingredienti indispensabili per realizzarla concretamente, avvalendomi prevalentemente di immagini.[1]

Comincerò introducendo nella sezione 2 le competenze del mio gruppo di ricerca in ambito di gestione, analisi e integrazione di Big Data; nella sezione 3 verrà fornita una panoramica delle nuove Tecnologie Informatiche abilitanti la Scienza dei Dati: Cyber Physical System (Sistema Fisico Cibernetico), Cloud Computing (Computazione nella Nuvola), Cognitive Computing (Computazione Cognitiva). La sezione 4 illustrerà i

[1] Dovrò fare uso di termini inglesi, che tradurrò in italiano.

concetti di Big Data, Integrazione di Big Data, mentre, la sezione 5 è dedicata a chiarire il concetto di Scienza dei Dati ed il processo a cascata (pipeline) per realizzarla.

Visti i limiti di spazio, non potrò parlare di temi molto importanti, quali: chi possiede e fa mercato dei nostri dati (i grandi attori di Internet), quale sarà l'impatto delle nuove tecnologie sull'occupazione, come è possibile utilizzare i dati seguendo principi etici e preservando la privacy. Questi temi sono stati da me brevemente trattati nella tavola rotonda "Intelligenza Artificiale Lavoro Impresa tra applicazioni partiche e regolazione europea" tenutasi presso il Dipartimento di Giurisprudenza l'8 Giugno 2022. Allego il link alla mia presentazione.

## 1. Il Database Group

Il gruppo è stato da me fondato nel 1992, in qualità di neoprofessore associato presso la facoltà di Ingegneria di Modena, ed è cresciuto in questi anni arrivando a costituire un'unità di ricerca di notevoli dimensioni: 6 professori, un tecnico laureato, un giovane ricercatore e una decina di dottorandi (una parte del gruppo è mostrata nella foto di **Error! Reference source not found.**).

L'attività di ricerca è attualmente focalizzata sui Big Data e Integrazione di Big Data:

- Big Data: Memorizzazione, Gestione, Interrogazione e Analisi di Big Data eterogenei (strutturati, testuali, multimediali) e distribuiti in rete in modalità *batch* o *streaming* (batch = già archiviati, streaming = in flusso).
- Integrazione di Big Data e, in particolare, *record linkage* (il processo per identificare la stessa entità del mondo reale in diverse fonti di dati).

L'approccio all'analisi dei dati si basa su tecniche quali la *Business Intelligence* (un insieme di tecnologie per raccogliere dati ed analizzare informazioni strategiche utilizzate dalle aziende tradizionalmente) e l'Intelligenza Artificiale (IA). In particolare, il focus è sulla cosiddetta *IA Data-Centric,* il nuovo approccio all'utilizzo dell'IA promosso recentemente dal ricercatore Andrew Ng. Andrew Ng è direttore del laboratorio di IA di Stanford, Co-Fondatore di Coursera nel 2012 con Daphne Koller, fondatore di DeepLearning.AI (creatore dei programmi didattici di IA di Coursera). Il DBGroup fa parte del nuovo movimento di ricerca internazionale su questo tema, che propone una nuova prospettiva di utilizzo del *Machine Learning (ML),* ovvero *MLOps*. L'obiettivo di MLOps è rendere disponibili dati di alta qualità in tutte le fasi del ciclo di vita di un progetto ML. Gli strumenti MLOps sono necessari per rendere l'IA basata sui dati un processo efficiente e sistematico.

Il DBGroup ha sviluppato una serie di tecniche teoriche di integrazione dei dati pubblicate in prestigiose sedi internazionali e implementato applicazioni prototipali a supporto della ricerca. MOMIS, un sistema di integrazione dati sviluppato da DBGroup, è stato industrializzato ed è ora commercializzato da DataRiver, una startup fondata dai membri del gruppo.

Chi fosse interessato ai progetti di ricerca, all'attività didattica, ai rapporti di collaborazioni internazionali, può visitare il sito: www.dbgroup.unimore.it.

## 2. Le nuove tecnologie informatiche

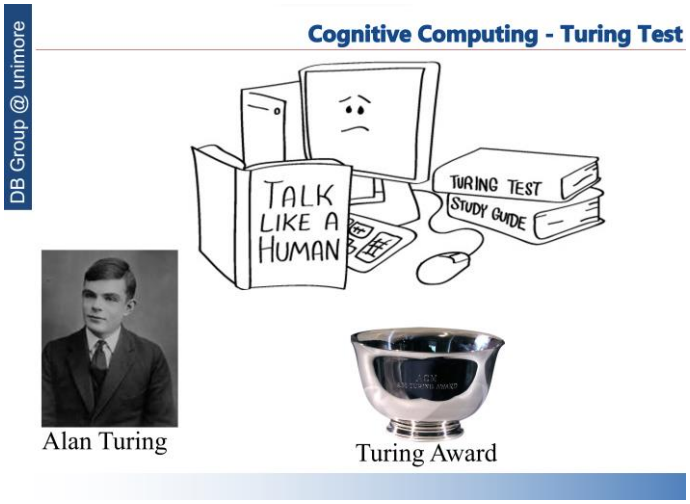


*Figura 1*

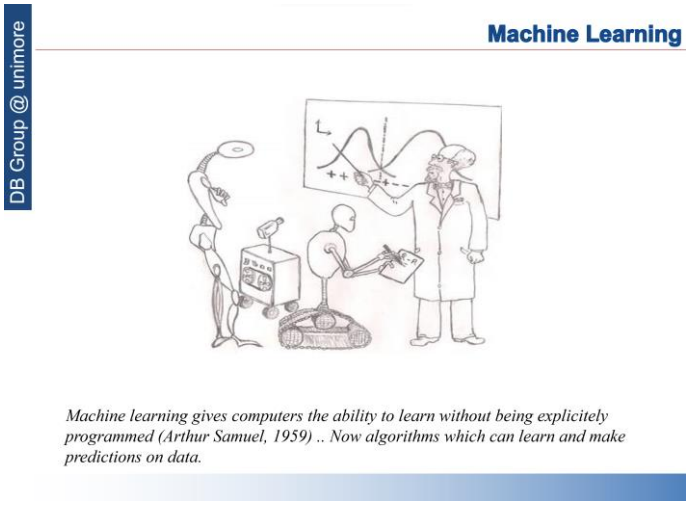


*Figura 2*

Introduciamo le nuove tecnologie informatiche che rendono possibile realizzare la *Scienza dei Dati.*

- *Cyber Physical System (CPS)*, 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.)
  - è un concetto più ampio di Internet of Things - IOT (Internet delle Cose).

- *Cloud Computing* 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* è la tecnologia che ci permetterà di interagire con i computer praticamente "parlando" alle macchine e sfruttando la loro capacità di apprendere dall'esperienza (Figura 1 e Figura 2).
  - è un concetto più ampio di Machine Learning, che è stato introdotto dal padre dell'Informatica *Alan Turing*. Non a caso il Nobel dell'Informatica è il Turing Award.
- *Big Data Integration* è la tecnologia fondamentale che ci permette di integrare informazioni relativamente allo stesso oggetto del mondo da sorgenti dati Big, eterogenee e distribuite (Figura 3, Figura 4 e Figura 5 per Big Data; Figura 6 e Figura 7 per Data Integration e sua rilevanza nel dominio della medicina).


*Figura 3*

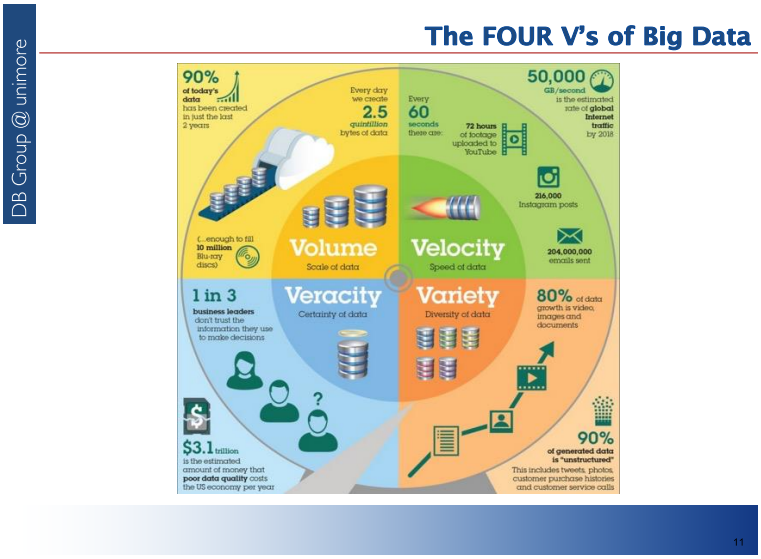


*Figura 4*

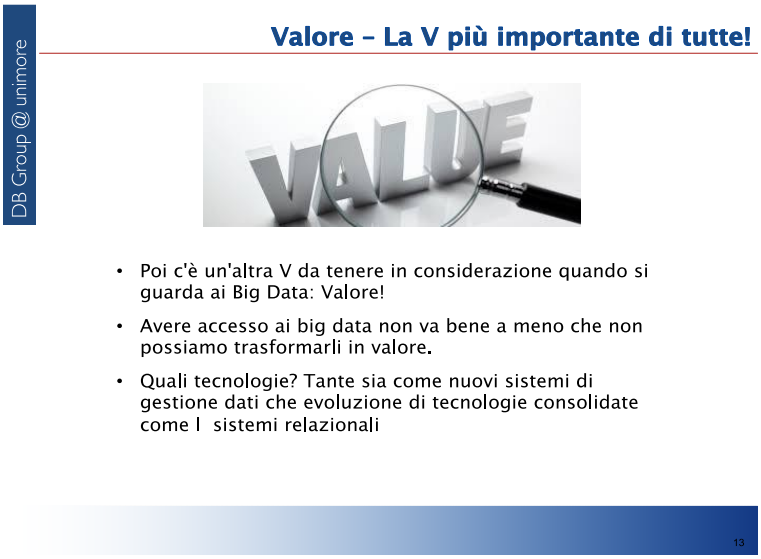


*Figura 5*

## 3. Big Data Integration

I dati non sono solo il nuovo petrolio, ma un elemento indispensabile per il progresso della scienza, come evidenziato dalla copertina di Wired in Figura 3. Nella stessa immagine si rende evidente che, per estrarre quello che serve dai dati è necessario dotarsi di *strumenti* (l'ombrello per la persona in figura che deve avere l'acqua necessaria per la sua pianta): *le giuste*

*tecnologie e l'intelligenza umana* per estrarre informazioni utili alle decisioni per la nostra società.

In Figura 4 si evidenziano le famose *4 V (Volume, Veridicità, Varietà e Variabilità)* dei Big Data e in Figura 5, la *V* più importante, il *Valore* che possiamo estrarre *con l'utilizzo intelligente delle tecnologie.*


*Figura 6*


*Figura 7*

La Figura 6 dà un'intuizione del concetto di *Integrazione dei Dati*. Il tema è stato oggetto di ricerche teoriche e di sviluppo di tecnologie a

partire dagli anni 90, ma recentemente, proprio la disponibilità dei Big Data e l'obiettivo di utilizzarli per migliorare le decisioni in ogni ambito della società, lo hanno reso un ingrediente fondamentale per rendere effettiva la scienza dei dati. La Figura 7 dà un'intuizione di come l'obiettivo della medicina personalizzata sia raggiungibile solo attraverso l'integrazione dei dati.

## 4. La Scienza dei Dati


*Figura 8*

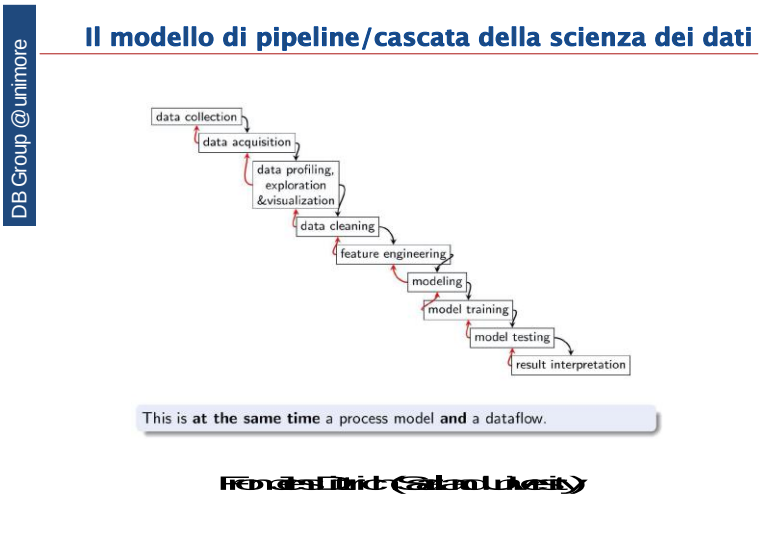


*Figura 9*

La Scienza dei Dati, come si vede in Figura 8, può essere rappresentata come una *torta* fatta da più ingredienti, ma anche come una *pipeline*, cioè un processo che utilizza strumenti per raccogliere dati grezzi da più fonti, integrarli, analizzarli e presentare i risultati in un formato comprensibile (vedi Figura 9).

In questo processo, l'intervento umano (dell'esperto del dominio) è indispensabile e, come è ben rappresentato dalle frecce che tornano indietro, le varie attività rappresentate nei rettangoli devono spesso essere ripetute.

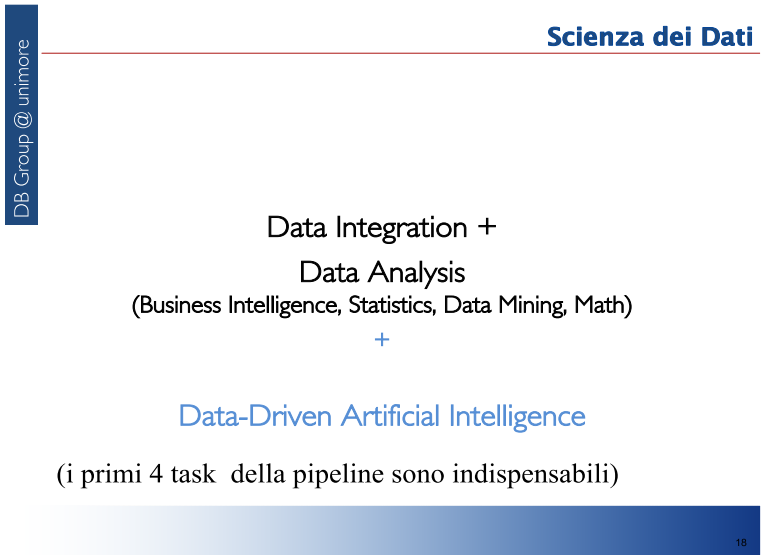


*Figura 10*

La Figura 10 evidenzia che le prime 4 attività (task), che vengono assieme indicate come *preparazione dei dati* costituiscano un requisito indispensabile per ogni progetto che abbia l'obiettivo di trarre *valore* dai dati. Data-Driven Artificial Intelligence o il suo sinonimo Data-Centric Artificial Intelligence è la giusta direzione per applicare e rendere efficace l'IA.

## 5. Conclusioni

In conclusione, una società in cui le decisioni vengano guidate dai dati reali, che garantisca i diritti dell'individuo e rispetti i principi etici è possibile. Abbiamo le tecnologie, ma queste devono essere utilizzate a partire da dati di qualità e di cui è noto il significato. Il risultato di ogni attività della pipeline deve essere valutato dall'esperto del dominio ed eventualmente rivisto. Non si può pensare ad un processo che non preveda l'intervento umano in ogni fase.