Dal Dato ai Big Data

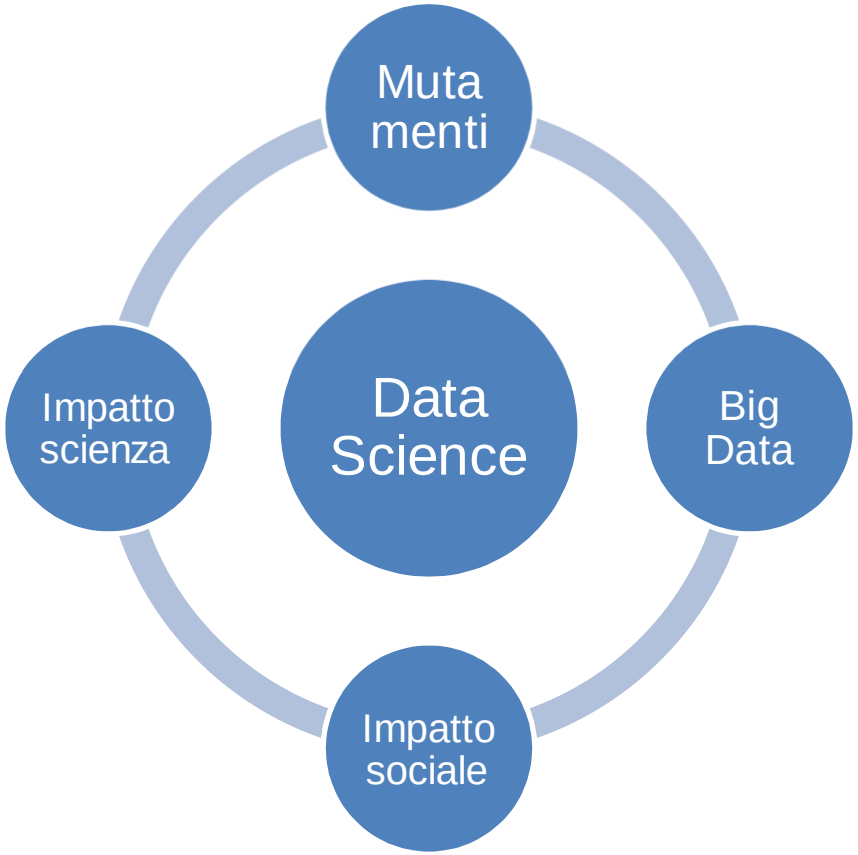
Prof. Michelangelo Ceci

UNIBA – Dipartimento di Informatica
CINI – Laboratorio Big Data

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

cini consorzio interuniversitario nazionale per l'informatica

Di cosa parleremo?



UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

consorzio interuniversitario nazionale per l'informatica

Una stima di crescita …

- I dati crescono in media del 30-40% annuo
- Ogni 2,5 anni si raddoppia il volume

Quanti dati fra venti anni?

- Oggi X
- Fra 2,5 anni $X \cdot 2 = X \cdot 2^1$
- Fra 5 anni $X \cdot 2 \cdot 2 = X \cdot 2^2$
- Fra 7,5 anni $X \cdot 2 \cdot 2 \cdot 2 = X \cdot 2^3$
- Fra 10 anni $X \cdot 2 \cdot 2 \cdot 2 \cdot 2 = X \cdot 2^4$

…

- Fra 20 anni $X \cdot 2^8 = 256 \cdot X$

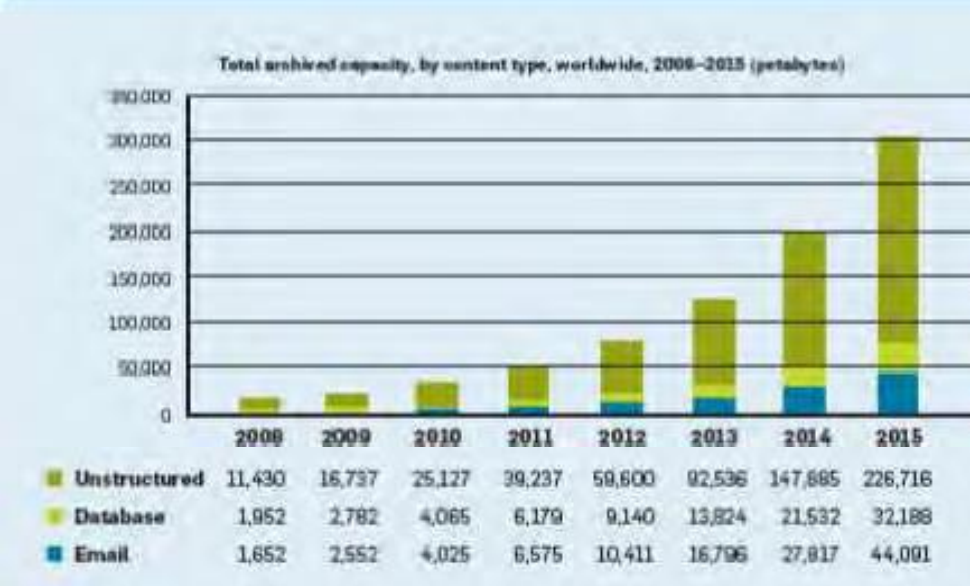
… quando andrò in pensione (spero) ☺

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

cini consorzio interuniversitario nazionale per l'informatica

Crescita esponenziale

- Crescita esponenziale dei dati
  - 2,7 ZB (10^{21} bytes) nel 2012!
  - 35 ZB nel 2020



Nome	Simbolo	Multiplo
kilobyte	kB	10^3
megabyte	MB	10^6
gigabyte	GB	10^9
terabyte	TB	10^{12}
petabyte	PB	10^{15}
exabyte	EB	10^{18}
zettabyte	ZB	10^{21}
yottabyte	YB	10^{24}

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

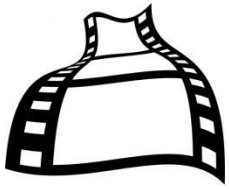
cini consorzio interuniversitario nazionale per l'informatica

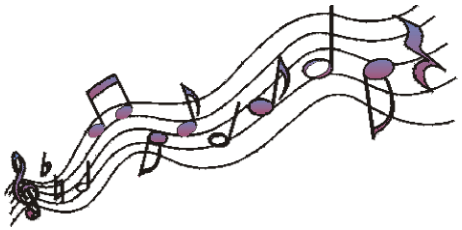
Crescita esponenziale

- La *Divina Commedia* di Dante Alighieri è composta da 671.447 caratteri
- 1 carattere = 1 byte
- 670 Kb = 1 Divina Commedia

Datizzazione (Datification)

- Neologismo che indica la conversione in formato digitale (dati) di:
- Film, musica, libri, etc. (contenuti che fino a qualche anno fa viaggiavano su pellicole, carta, vinili e altri supporti)
- Conversazioni telefoniche, mail, trasmissioni televisive e radiofoniche





UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

cini consorzio interuniversitario nazionale per l'informatica

Datizzazione

- *Facebook* ha “datizzato” le relazioni,

- *Twitter* ha reso possibile la “datizzazione” dei sentimenti,

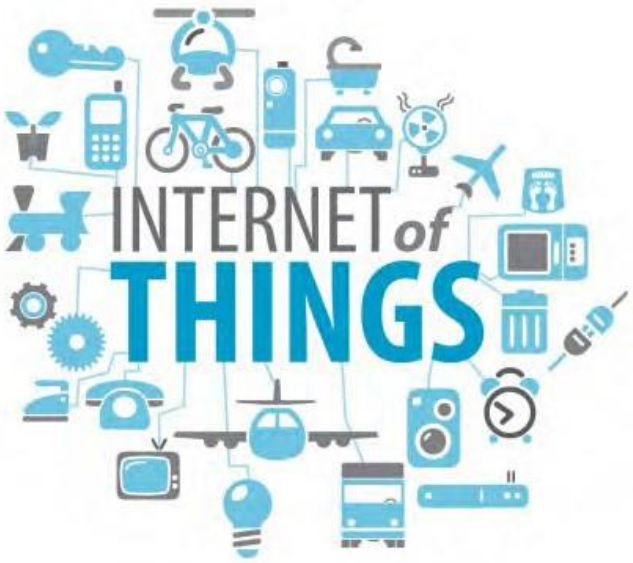
- *LinkedIn* ha “datizzato” le nostre esperienze professionali

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

cini consorzio interuniversitario nazionale per l'informatica

I dispositivi generano dati …

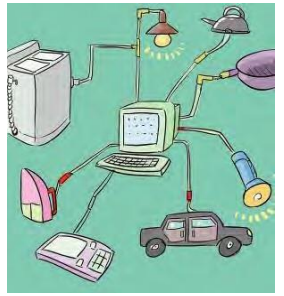
- **Internet of Things** (IoT): Gli oggetti si rendono riconoscibili e acquisiscono intelligenza grazie al fatto di poter comunicare dati su se stessi e accedere ad informazioni aggregate da parte di altri.



UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

cini consorzio interuniversitario nazionale per l'informatica

I dispositivi generano dati …



Esempi:

- Le sveglie suonano prima in caso di traffico,
- Le piante comunicano all'innaffiatoio quando è il momento di essere innaffiate,
- i vasetti delle medicine avvisano i familiari se un loro parente dimentica di prendere il farmaco.

Tutti gli oggetti possono acquisire un <u>ruolo</u> <u>attivo</u> grazie al collegamento a Intenet.

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

consorzio interuniversitario nazionale per l'informatica

Noi generiamo dati …

- Grazie alla nostra forte simbiosi con le tecnologie digitali, siamo diventati dei "sensori" viventi.
- 7 miliardi di persone e 6,8 miliardi di cellulari

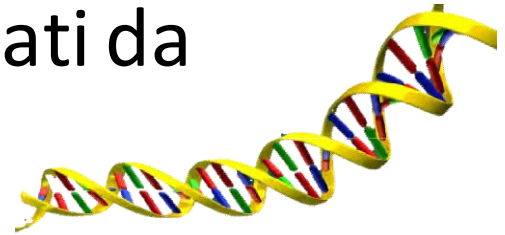
- «***Siamo Pollicini digitali, ci lasciamo dietro briciole di informazioni, tracce di noi stessi***». (Dino Pedreschi)

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

cini consorzio interuniversitario nazionale per l'informatica

La scienza genera dati …

- Le tecnologie digitali hanno permesso di fare passi da gigante, in questi anni, nel campo della **genomica**, dove le moli di dati da analizzare sono enormi.



- mappatura del DNA di un individuo

  da 3 miliardi di dollari e 13 anni di ricerca (1990-2003) ➔ poche migliaia di dollari per un processo che dura un paio di settimane.

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

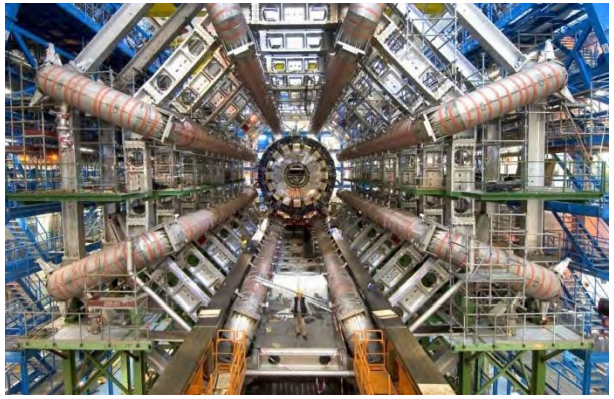
cini consorzio interuniversitario nazionale per l'informatica

La scienza genera dati …

- ***Human Brain Project***
- Un osservatorio del cervello che monitora 1 milione di neuroni (o 100.000 neuroni in 10 soggetti) per 1 volta al secondo genererebbe
  - 1 gigabyte di dati al secondo,
  - 4 terabytes all'ora,
  - 100 terabytes al giorno
  - 4 petabyte all'anno (ipotizzando un fattore di compressione di 1/10).

La scienza genera dati …

- Il ***Large Hadron Collider*** (LHC) acceleratore di particelle situato presso il CERN di Ginevra, utilizzato per ricerche sperimentali nel campo della fisica delle particelle, può produrre 30 petabyte di dati l'anno.
- L'Agenzia spaziale europea genera più di un petabyte di dati all'anno.



esa
European Space Agency

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

cini consorzio interuniversitario nazionale per l'informatica

Le Pubbliche Amministrazioni generano dati (in formato aperto)

- Open Data
  - dati liberamente accessibili a tutti, privi di brevetti o altre forme di controllo che ne limitino la riproduzione
  - gli eventuali copyright eventualmente si limitano all'obbligo di citazione della fonte o al rilascio delle modifiche con stesso copyright.
  - Su dati.gov.it sono disponibili 20.851 Dataset

dati.gov.it
I dati aperti della Pubblica Amministrazione

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

cini consorzio interuniversitario nazionale per l'informatica

Le aziende generano dati …

- Oggi ogni grande business è un digital business:
  - **Alibaba** è il più grande negozio al mondo, ma non ha nemmeno un magazzino.
  - **Uber** è la più grande compagnia di noleggio veicoli, ma non possiede nemmeno un'auto.
  - **Airbnb** è il più esteso network dedicato alla ricettività, ma è del tutto privo di strutture.

Le aziende generano dati …

- Ordini, acquisti, vendite, spedizioni, difetti di produzione, …
- I dati sono raccolti nei sistemi informatici delle aziende. Sono considerati un *asset (intangibile).*
- *Facebook:* dichiara *asset* (tangibili) per 6,3 miliardi ma venne valutata in Borsa 104 miliardi il giorno del suo debutto.

Le aziende generano dati …

- Nonostante i dati siano un asset, oggi viene elaborato solo il **5‰** dei dati aziendali
- Perché?
  - mancanza di competenze sull'analisi computazionale dei dati;
  - sovversione dei poteri generati da un'informazione così tempestiva.

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

consorzio interuniversitario nazionale per l'informatica

Il Diluvio dei Dati

- Il termine "**diluvio dei dati**" si riferisce alla situazione in cui le incredibili dimensioni dei dati generati sta sopraffacendo la capacità delle istituzioni nel gestirli e dei ricercatori nel farne uso nei loro studi.

Febbraio 2010

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

cini consorzio interuniversitario nazionale per l'informatica

Big Data

- Raccolta di dati così estesa in termini di volume, velocità e varietà da richiedere **strumenti non convenzionali** per estrapolare, gestire e processare informazioni entro un tempo ragionevole.

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

cini consorzio interuniversitario nazionale per l'informatica

Big Data: una rivoluzione?

- *La vera rivoluzione non sta nelle tecnologie per elaborare i dati, ma nei dati in sé e nel modo in cui li usiamo.*

- *Aumentando la scala dei dati con cui si lavora, si possono fare cose nuove che non sono possibili con minori quantità dei dati.*

Big Data vs. Data Science

- Data Science
  - *La scienza dei dati studia i metodi per estrarre la conoscenza dai dati.*
    - *Dati di qualunque natura*
  - Un approccio olistico alla creazione di prodotti e servizi basati sull'estrazione di conoscenza dai dati
    - La conoscenza estratta è immediatamente utilizzabile (*actionable*) nei processi decisionali.

Big Data vs. Data Science

- Data Science vs. Big Data
  - Data Science non necessita sempre di Big Data, tuttavia la costante crescita dei dati fa si che i Big Data siano un aspetto importante della Data Science.

Data Scientist

- Data Scientist
  - Figura professionale dotata di abilità “integrate” che spaziano dalla matematica, all’apprendimento automatico, alla statistica, al data mining, ai database e all’ottimizzazione
  - capace di ingegnerizzare delle soluzioni efficaci alla creazione di nuovi prodotti e servizi.

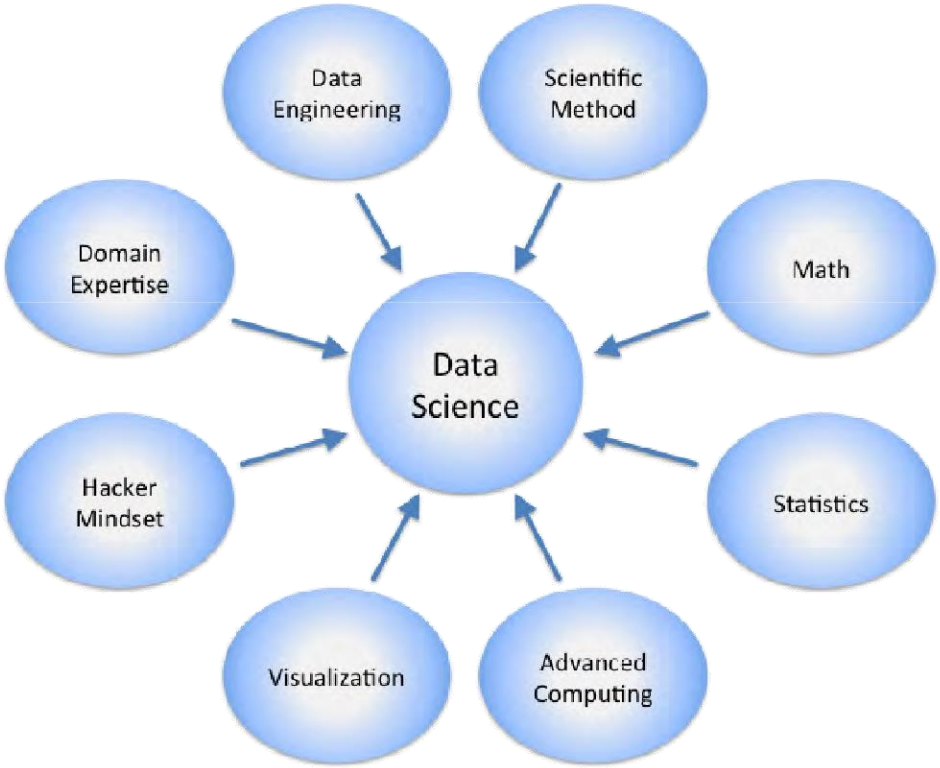
Carenza di data scientist

- C'è una **carenza di professionalità**:
  150.000 data scientist richiesti solo negli USA
- Osservatorio Big Data Analytics & Business Intelligence - anno 2017

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

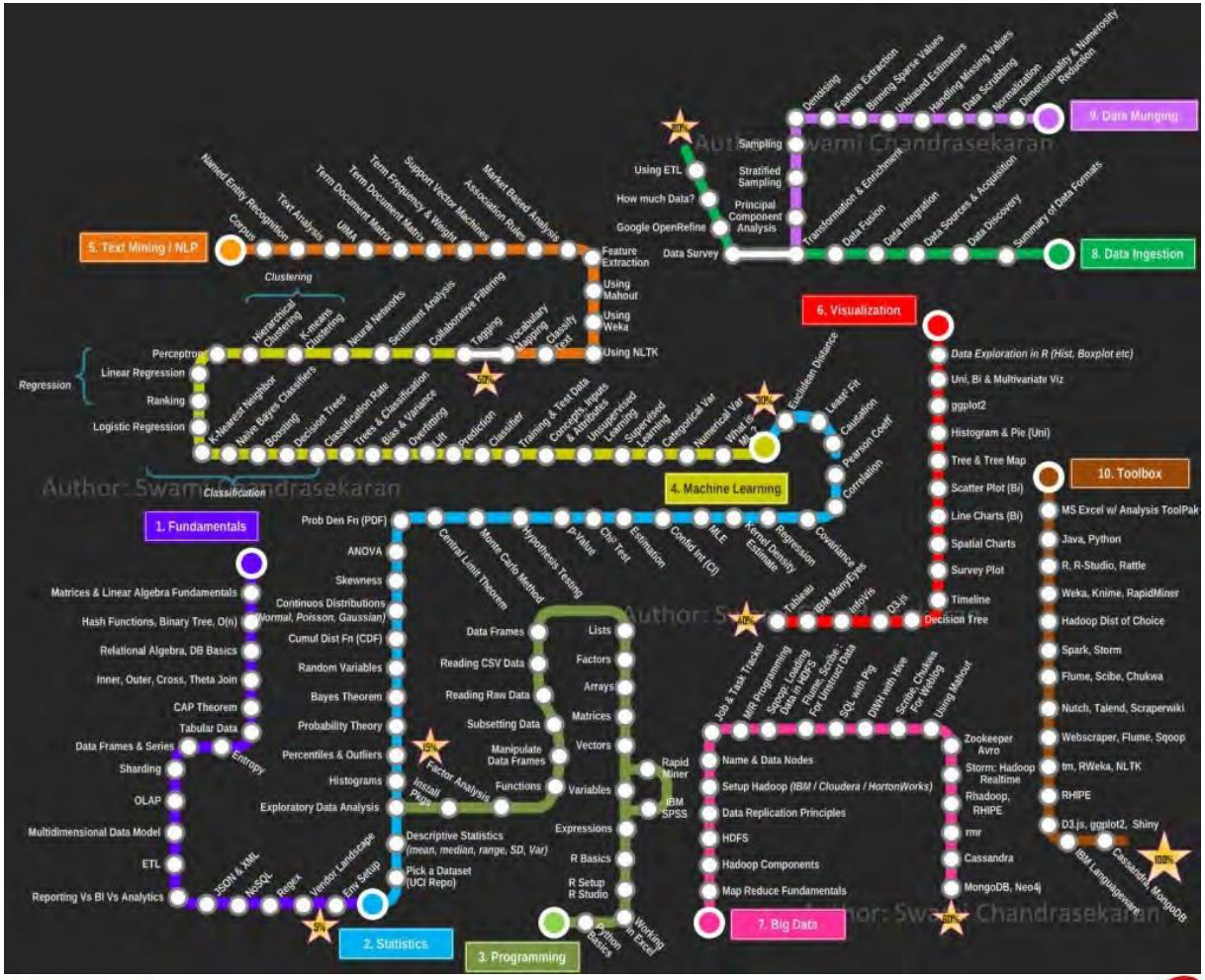
cini consorzio interuniversitario nazionale per l'informatica

Quali competenze professionali?



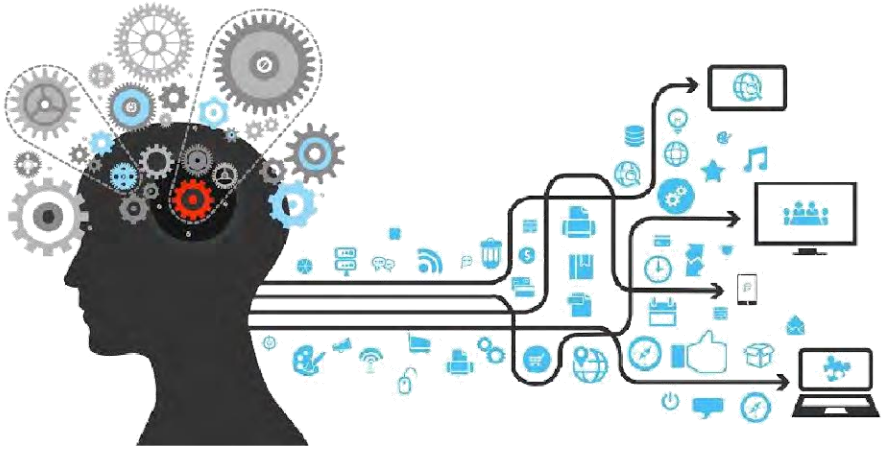
Calvin Andrus - Creative Commons

Quali competenze professionali?



Data Science vs. Artificial Intelligence

- IA indaga sistemi che mostrano un comportamento intelligente analizzando il proprio **ambiente** e compiendo **azioni**, con un certo grado di autonomia, per raggiungere specifici **obiettivi**.
- Agiscono nel mondo virtuale o reale.



UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

cini consorzio interuniversitario nazionale per l'informatica

Data Science vs. Artificial Intelligence

- DS: indaga come produrre **intuizioni** dai dati
- IA produce **azioni** dai dati

Esempio:

Un'auto a guida autonoma usa sistemi di IA per prendere l'azione di applicare i freni.

La data science associa le basse prestazioni nei test su strada a un segnale di stop (intuizione).

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

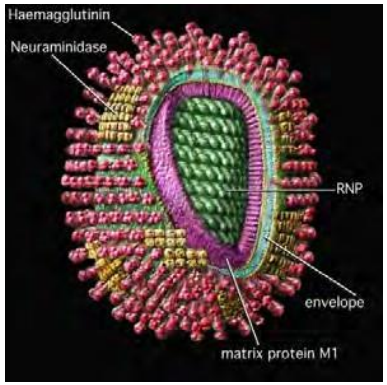
cini consorzio interuniversitario nazionale per l'informatica

Big Data*: Un esempio*

- 2009: Pericolo di pandemia da virus A/H1N1 (febbre suina)

  USA Centers for Disease Control and Prevention:

  raccolta delle segnalazioni di nuovi casi di influenza da parte dei medici + modelli epidemiologici ➔ due settimane di sfasamento



UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

cini consorzio interuniversitario nazionale per l'informatica

Big Data*: Un esempio*

Google Flu Trends: previsione in base all'oggetto delle ricerche condotte con *Google search* ➔ ugualmente accurate ma in tempo reale

Detecting influenza epidemics using search engine query data. ***Nature 457, 1012-1014 (19 February 2009)***

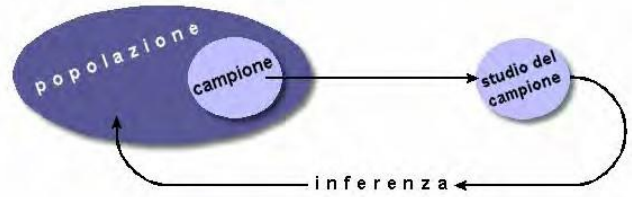
UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

consorzio interuniversitario nazionale per l'informatica

Big Data*: Un cambio di prospettiva*

- L'ascesa dei Big Data evidenzia tre mutamenti nel modo in cui analizziamo le informazioni:
  1. Analizzare tutti i dati disponibili
  2. Accantonare il desiderio di esattezza
  3. Abbandonare la tendenza a ricercare la causalità

Big Data: *Di più*



- Analizzare tutti i dati disponibili
  - Assuefazione al campionamento statistico ➔ autolimitazione nell'uso delle informazioni
  - Il campionamento casuale è solo un ripiego
  - È poco utile quando si vuole scavare in profondità
  - Il campionamento trascura i dettagli !!
  - L'identità "N = tutti" non comporta necessariamente l'analisi di una gran massa di dati

Big Data: *Confusione*

- Rinunciare all'esattezza
  - L'incremento dei volumi ➔ inesattezza
  - È importante avere *small data* (campioni) accurati
  - L'esattezza può essere sacrificata in favore dell'ampiezza o della frequenza
  - Accettare l'inesattezza dei modelli estratti dai dati o nella struttura dei dati

Big Data: *Confusione*

- L'esattezza vs. ampiezza / frequenza

Vigna wireless della Intel. La rete di sensori wireless rileva la presenza di parassiti e permette di selezionare l'insetticida.

[*J. Burrell, T. Brooke, and R. Beckwith, "Vineyard computing: Sensor networks in agricultural production," IEEE Pervasive Computing, vol. 3, no. 1, pp. 38–45, 2004.*]

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

cini consorzio interuniversitario nazionale per l'informatica

Big Data: *Confusione*

- Indice dei prezzi al consumo: PriceStats

Usa solo prezzi online

Big Data: *Confusione*

- Nell'epoca dei Big Data, ***la quantità è più importante della qualità***.
- L'abbondanza permette di tollerare un certo livello di imprecisione, di confusione

Big Data: *Confusione*

- Il traduttore di Google prende le informazioni di cui ha bisogno per le sue traduzioni da pagine web non filtrate, piene di errori ortografici, sintattici e a volte incomplete, ma la sterminata quantità di dati a disposizione gli permette di essere più affidabile di tutti i suoi predecessori, che si basavano su dizionari corretti e redatti da esperti, ma con il limite di contenere un numero limitato di informazioni.

Big Data: *Correlazione*

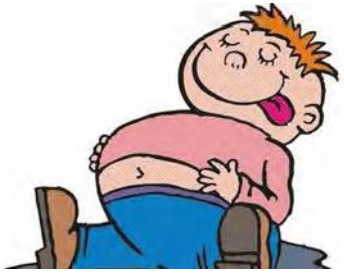
- Rinunciare alla causalità in favore della correlazione
  - Non conta sapere perché (*why*) vendo un libro online, ma cosa (*what*) fa aumentare le vendite
    - In previsione di un uragano aumentano le vendite di torce elettriche, ma anche di merendine e dolci

Big Data: *Correlazione*

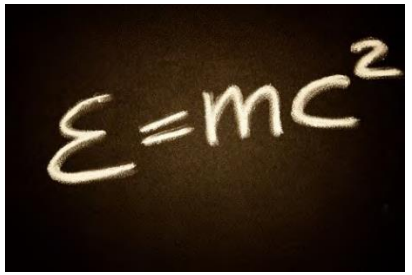
- Rinunciare alla causalità in favore della correlazione
  - La dimostrazione di una causalità è molto più costosa della individuazione di una correlazione.

Big Data: *Correlazione*

Esempio: il peso dei bambini di scuola elementare è correlato positivamente al quoziente intellettivo



peso ↔ QI



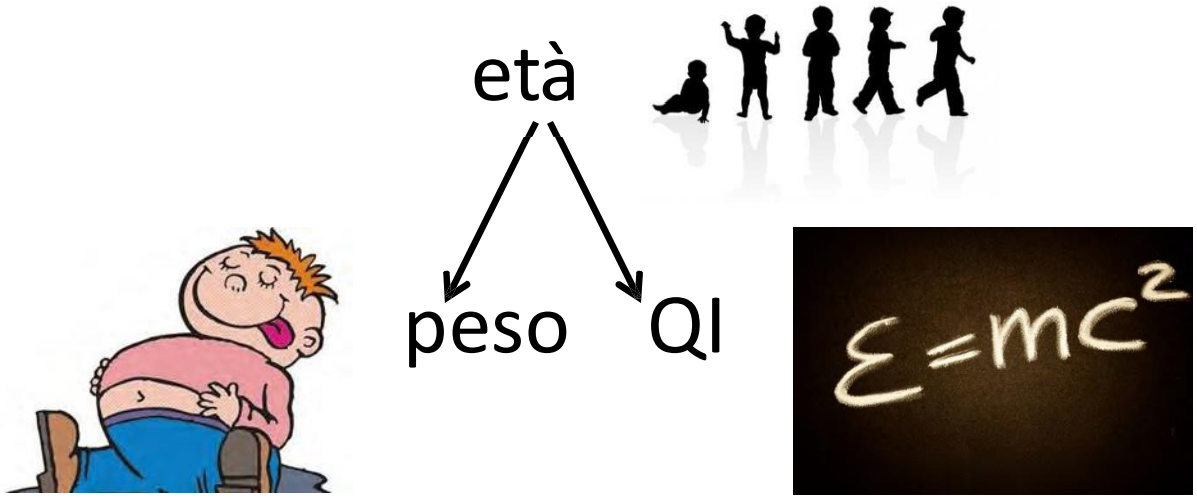
Facile da scoprire.

Ma direste che mangiare fa aumentare il QI?

O che il QI influisce sul peso?

Big Data: *Correlazione*

Esempio: Se tenessimo sotto controllo il fattore età giungeremmo a conclusioni diverse.

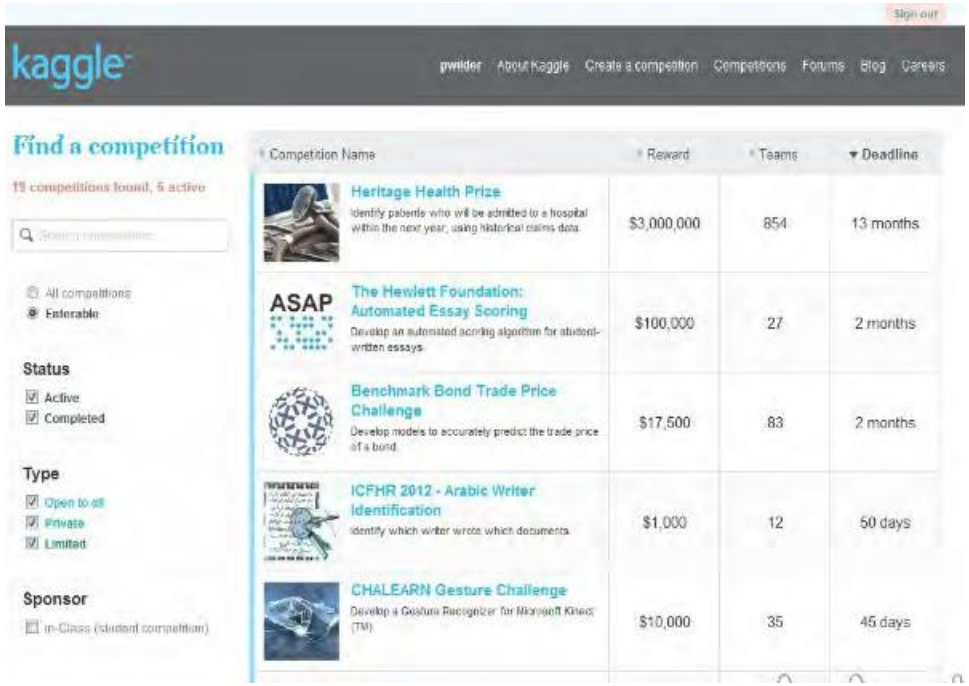


Tenere sotto controllo un singolo fattore costa

… e non sempre è possibile

Big Data: *Correlazione*

- Se scopriste che le auto usate arancioni sono meno soggette ad avere difetti, che auto usata comprereste?
- Vi porreste il problema di spiegare perché avete fatto quell'acquisto?



UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

cini consorzio interuniversitario nazionale per l'informatica

Big Data: *Correlazione*

- Avvertimento: Avere una gran quantità di dati a disposizione non significa saper comprendere la realtà.
- Le correlazioni ci dicono cosa, ma non perché.
- I Big Data ci aiuteranno a individuare il colore che andrà di moda il prossimo anno, ma non sono in grado di spiegarci perché.

I Big Data per il bene comune

- I big data ci danno anche la possibilità di osservare la rete delle relazioni sociali e dei movimenti, mettendo a nudo il tessuto sociale in cui siamo immersi.

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

cini consorzio interuniversitario nazionale per l'informatica

I Big Data per il bene comune

- Esperimento di Nathan Eagle (Science, 2010) basato sui big data telefonici.
- Obiettivo: misurare la diversificazione sociale di ciascun utente in base alle sue telefonate.
- Un utente che chiama sempre le stesse, poche, persone ha una bassa diversificazione, al contrario di un utente che chiama una vasta rete di contatti.

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

cini consorzio interuniversitario nazionale per l'informatica

I Big Data per il bene comune

- Ad ogni utente possiamo associare un indice numerico di diversità sociale, e possiamo poi aggregare questo valore su un determinato territorio, (comune, provincia, etc.) calcolando la media della diversità sociale degli abitanti quel territorio.
- Otteniamo così un unico indicatore per ciascun territorio, che può essere confrontato con altri indicatori ottenuti ad esempio attraverso indagini di istituti di statistica, sondaggi, etc.

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

cini consorzio interuniversitario nazionale per l'informatica

I Big Data per il bene comune

- Eagle ha confrontato la diversità sociale delle comunità locali dell'Inghilterra con un indicatore di benessere calcolato periodicamente dal governo britannico, che tiene conto di vari fattori oltre al reddito (educazione, salute, occupazione, etc.).
- Il risultato non lascia spazio a dubbi: **il benessere è strettamente legato alla diversità, i due indicatori crescono (o calano) insieme**.

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

cini consorzio interuniversitario nazionale per l'informatica

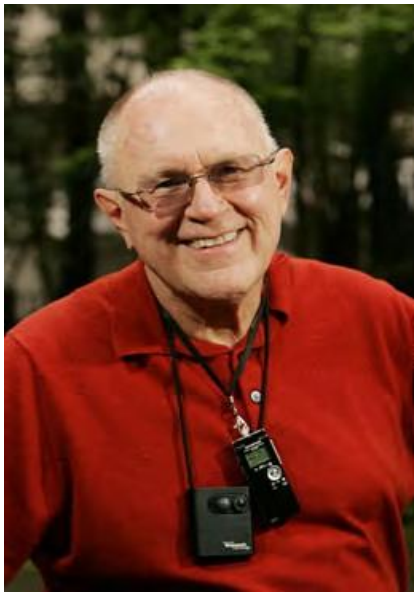
Big Data per una scienza *data-driven*

- I Big Data sono alla base anche di un cambiamento epocale nel modo di fare scienza.

Un quarto paradigma?

- Recentemente **Gordon Bell**, ricercatore emerito a Microsoft Research, che come Ken Wilson ha sostenuto il riconoscimento del terzo paradigma, ha indicato un nuovo paradigma alla base della scienza moderna: la **data intensive scientific discovery**.

http://research.microsoft.com/en-us/collaboration/fourthparadigm/



UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

cini consorzio interuniversitario nazionale per l'informatica

Un quarto paradigma?

- In molte discipline si raccolgono grandi collezioni di dati che non verranno mai analizzati. La raccolta di dati non è più finalizzata a un esperimento, come nel caso del paradigma della sperimentazione o nella simulazione. In molti progetti, la raccolta dei dati diventa un fine e non un mezzo per produrre conoscenza.
- Purtroppo molti dei dati raccolti non serviranno mai a produrre nuova conoscenza.
- Se negli ultimi anni sono stati fatti dei progressi nella costruzione di strumenti di simulazione, siamo ancora agli inizi nella produzione di efficaci strumenti di analisi.

Un quarto paradigma?

- Il quarto paradigma promette la produzione di nuove ipotesi a partire da queste vaste raccolte di dati (Big Data), ipotesi che potranno essere poi verificate mediante strumenti analitici, sperimentali e di simulazione.
- Quindi, dopo l'osservazione empirica, volta alla descrizione dei fenomeni naturali; la riflessione teorica, che mira a generalizzare i risultati dell'osservazione e costruire modelli; e negli ultimi decenni, l'approccio computazionale, che costruisce la simulazione dei fenomeni complessi, si sta recentemente affermando l'esplorazione e la manipolazione di grandi quantità di dati, come un nuovo paradigma.

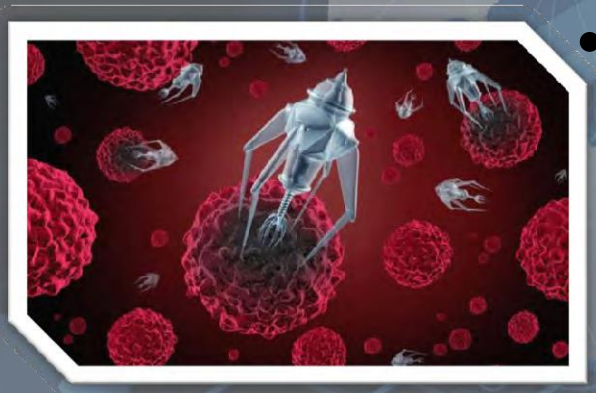
X-informatics: le nuove scienze

- Sono sorte nuove discipline scientifiche, le cosiddette **X-informatics** (bioinformatics, astroinformatics, etc.), fondate sul "**quarto paradigma**".
- Si parte dai dati, non dai modelli o dalle teorie.

Quella informazionale non è la sola grande rivoluzione tecnologica in corso

- Genetica

- Nanotecnologica



- Neurocognitivo

Conclusioni

I Big Data:

- cambiano il modo in cui analizziamo le informazioni
- presentano un enorme potenziale
  - nella società

    democrazia, beni culturali, cultura, sport, …
  - nella scienza
  - nella economia

    il 91% delle Fortune 100 companies hanno almeno una iniziativa big data in corso

Conclusioni

In un mondo guidato dai dati,
che spazio resta per le persone?

Laboratorio CINI su *Big Data*

http://www.consorzio-cini.it/

- **CINI**
  - 45 università aderenti al consorzio
- **Nodi partecipanti al Lab CINI su Big Data:**
  - 32 unità.
- **Personale interessato**
  - Circa 300 docenti

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

consorzio interuniversitario nazionale per l'informatica

Insegnamenti e Corsi di Laurea

Big Data, Artificial Intelligence, Machine Learning, ecc.

Corso di Laurea Magistrale in Computer Science (indirizzo Artificial Intelligence)

Big Data Analytics, Data Mining, ecc.

Corso di Laurea Magistrale in Data Science