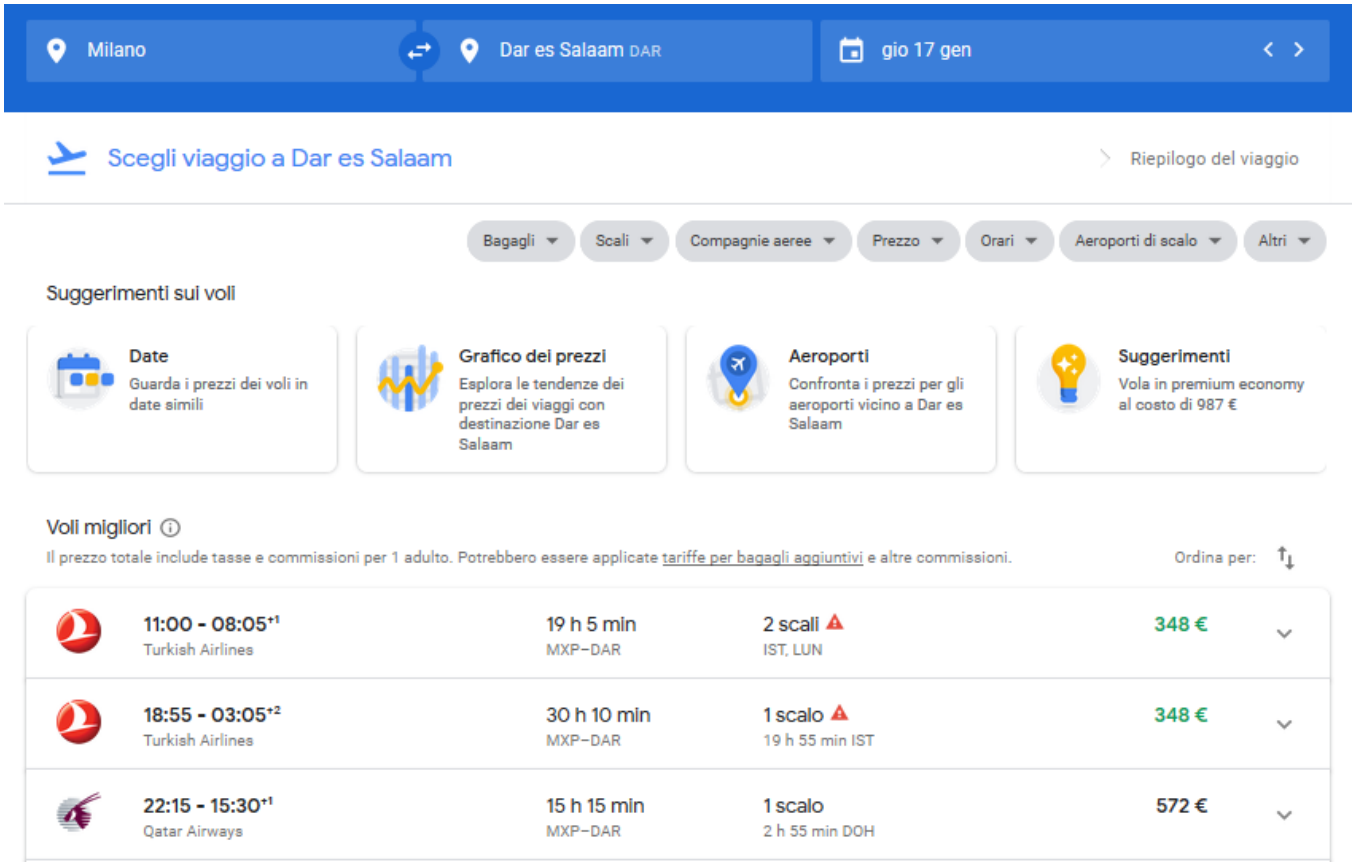
LEZIONI LINCEE DI DATA SCIENCE E SCIENZE INFORMATICHE

La nuova Scienza dei dati e le sue sfide

Carlo Batini
Università di Milano-Bicocca

Il prezzo dei biglietti aerei

Problema 0 - Fissato il giorno del viaggio, trovare il biglietto meno costoso



Problema 1 - Fissato il giorno del viaggio, scoprire **quale è il giorno** in cui il biglietto costa meno

?

Problema 2 - Prevedere quando e dove ci saranno le eclissi di sole e di luna l'anno prossimo

Problema 3 – Se oggi voglio fare un po' di corsa, quale sarà il livello di inquinamento che trovo? Verso dove conviene andare per respirare un'aria accettabile?

Problema 4 – Tradurre una frase dall'italiano in inglese, arabo, cinese, spagnolo,…

Sono problemi risolti? E da quanto tempo?

Problema 1 – **PREDIRE** Il giorno in cui acquistare un biglietto
→ Risolto **pochi anni fa**

Problema 2 – **PREDIRE** quando ci sarà una eclissi
→ Risolto dai babilonesi **oltre 2.000 anni fa** →

Problema 3 – **PREDIRE** i livelli di inquinamento
→ Risolto **pochi anni fa**

Problema 4 – **TRADURRE** un testo dall'italiano in inglese, ecc.
→ Risolto **pochi anni fa**

Problema 2 – Predire le Eclissi

L'osservazione delle eclissi all'epoca dei Babilonesi portò a scoprire il **ciclo di Saros**, che dice secondo quale scansione temporale si succedono le eclissi del sole e della luna.

Il confronto tra le previsioni fatte dai babilonesi e quelle ottenute con le attuali tecnologie mostra una precisione stupefacente per l'epoca.

Problema 4 - Tradurre un testo
Le biografie inglesi di Palazzo Chigi

Le biografie inglesi di Palazzo Chigi
(biografie riprese dal sito ufficiale del governo)


Silvio Berlusconi


Gianfranco Fini


Gianni Letta


Paolo Bonaiuti


Giuseppe Pisanu


Franco Frattini


Rocco Buttiglione

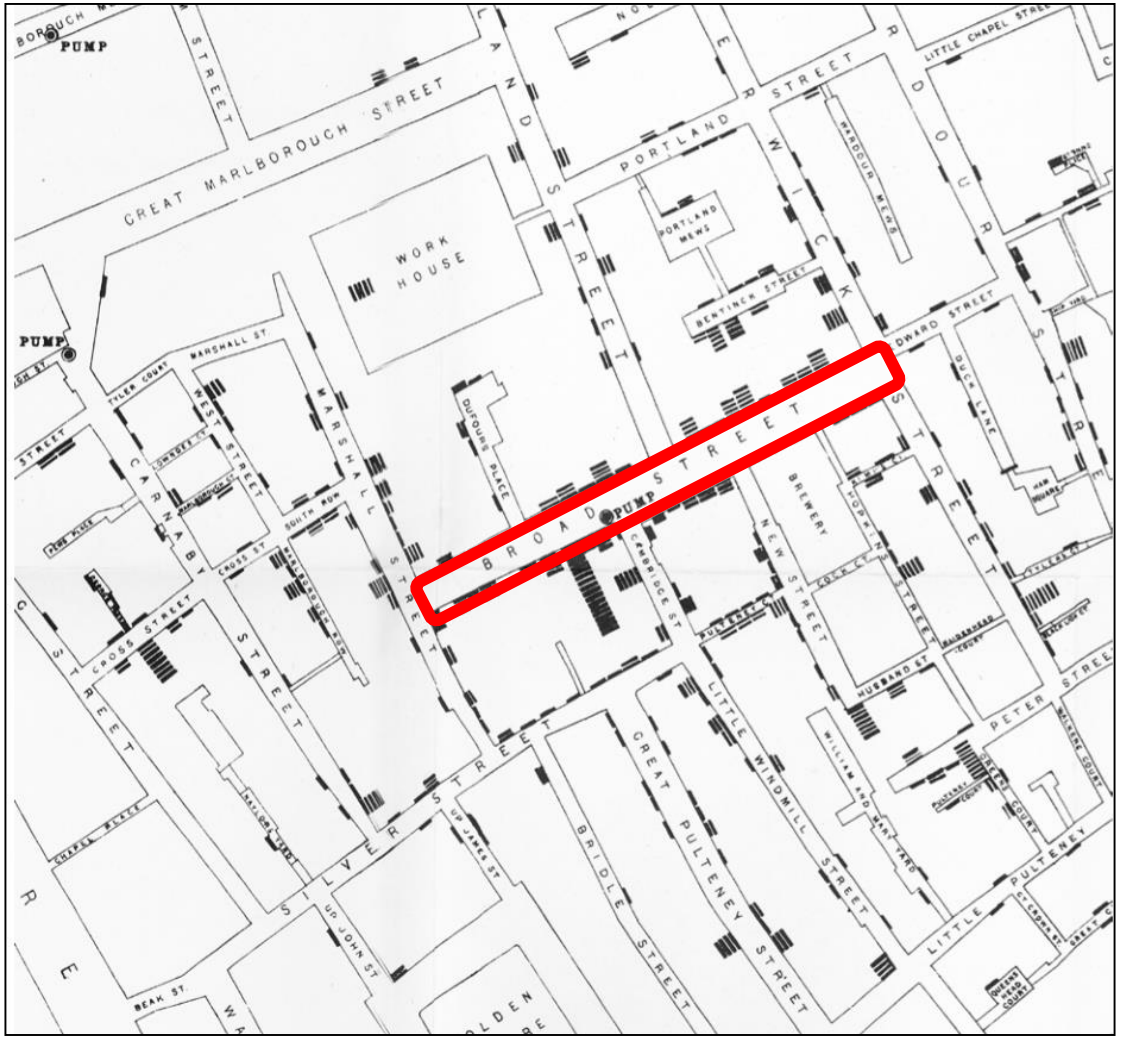

Lucio Stanca

Lucio Stanca

Been born to Lucera (Foggia) 20 October 1941. Conjugated and it has two daughters. In 1965 one has graduated in Economy near the University Mouthfuls of Milan.

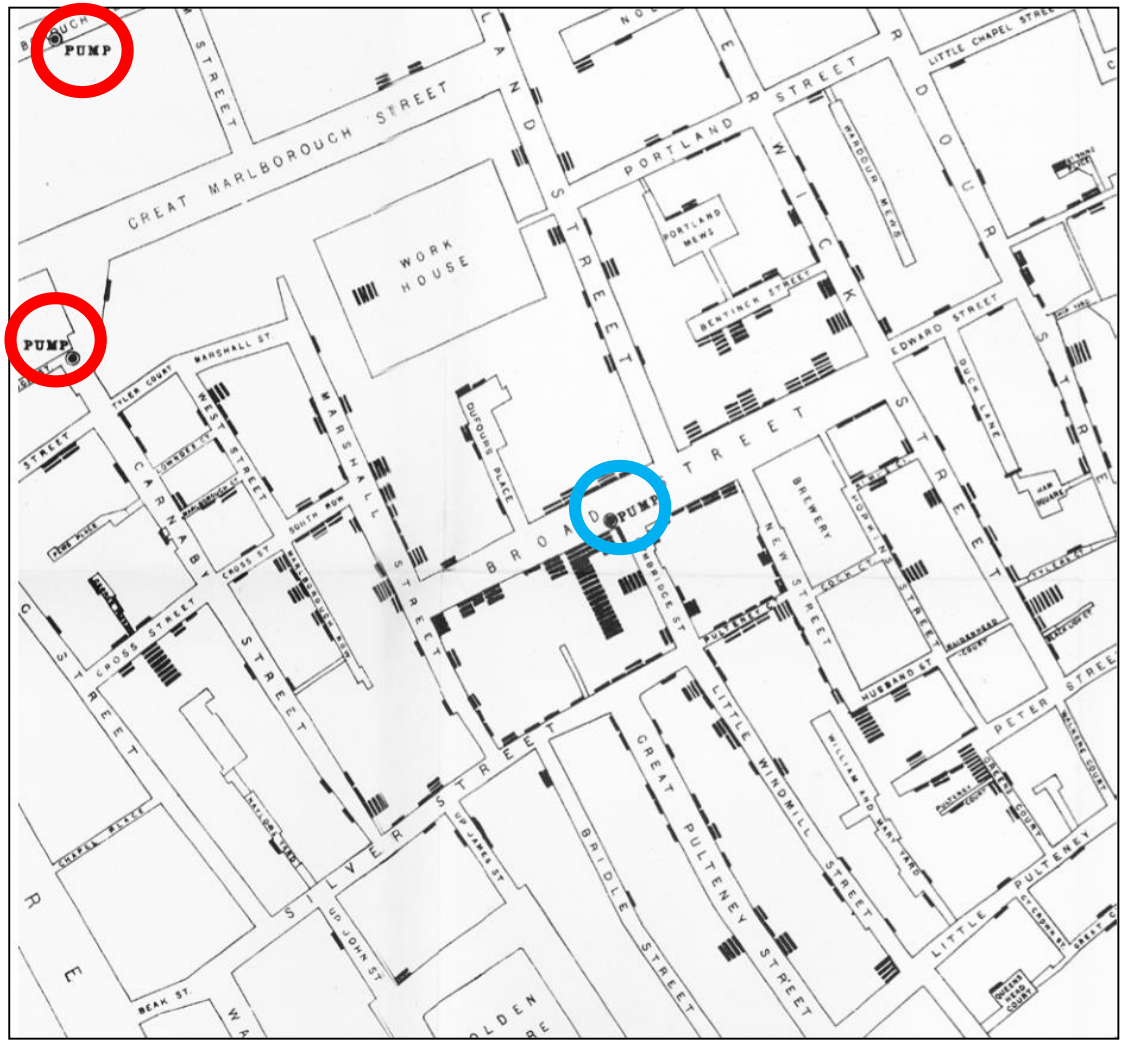
Perché gli altri problemi sono stati risolti solo pochi anni fa?

Tutto è cominciato nel 1854… - La famosa mappa disegnata da Snow dell'area di Broad Street nell'anno 1854



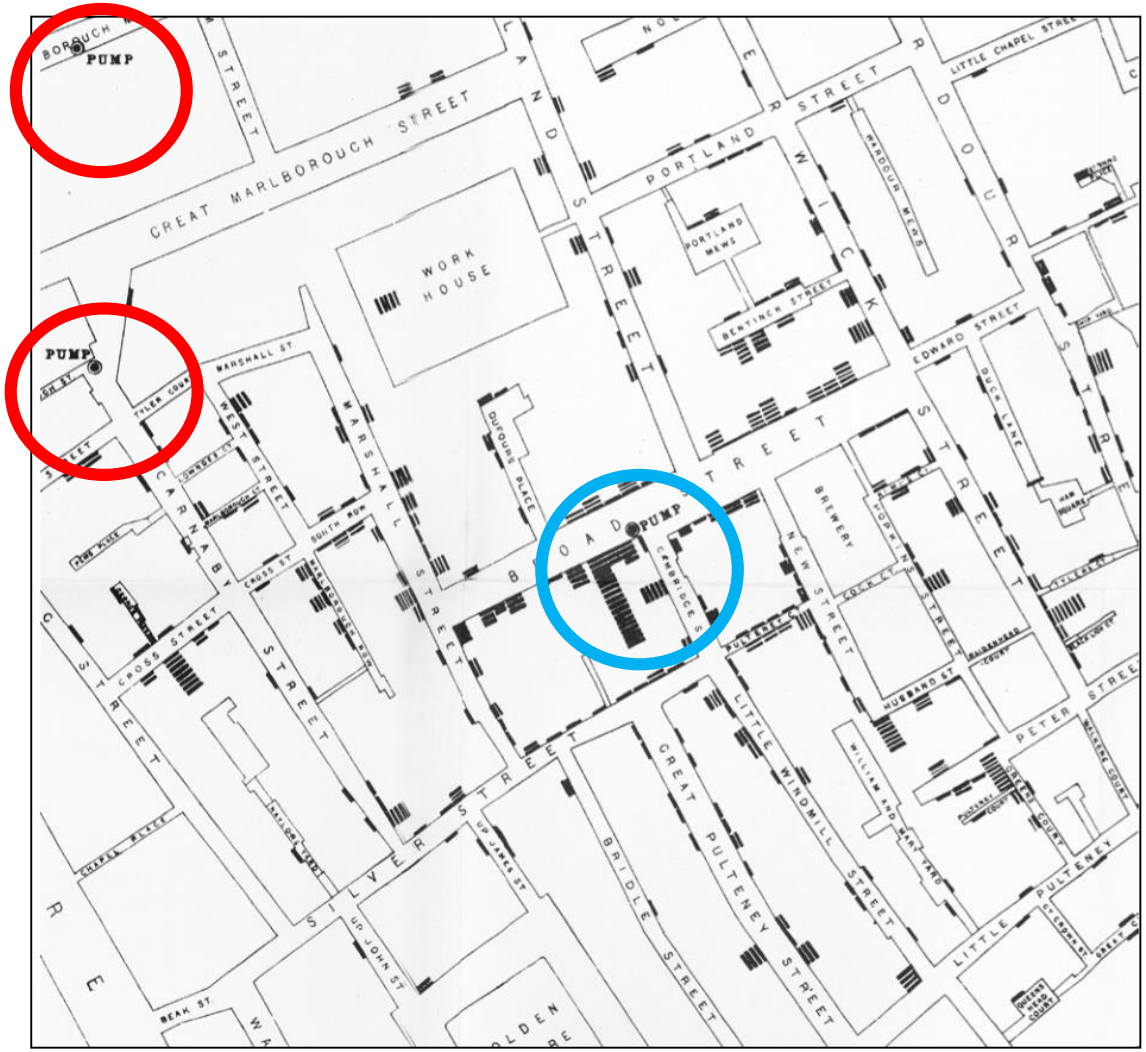
Tutto è cominciato nel 1854…

Le pompe delle diverse compagnie **rosse** e **blu**



Tutto è cominciato nel 1854…

Correlazione tra pompe e decessi

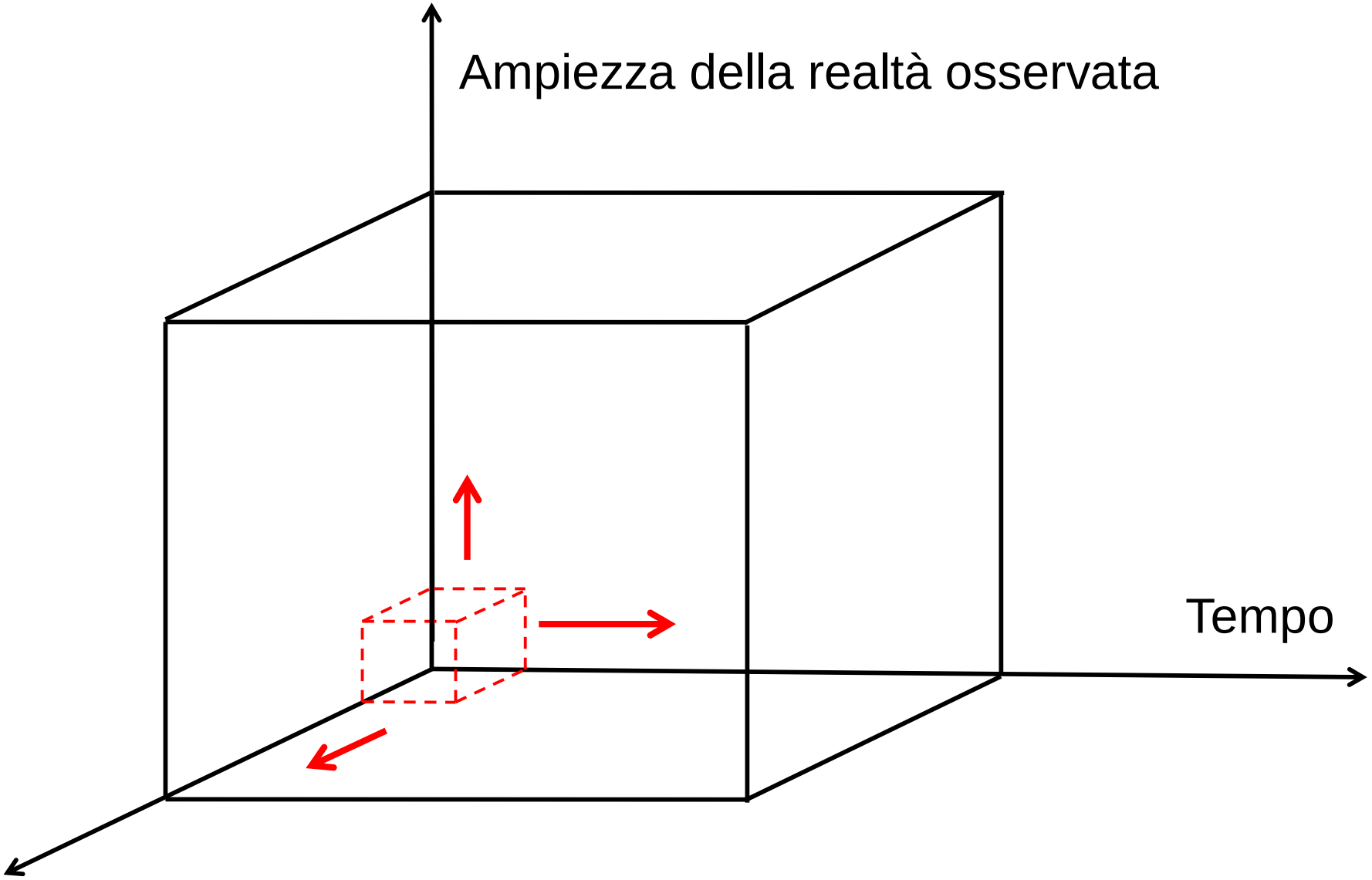


Dai piccoli dati ……

Il diluvio dei dati

- Ogni anno e mezzo raddoppia la quantità di dati scambiati sul Web
- Nel 2025 ci saranno 1.000 sensori dell'Internet delle cose per ogni essere umano

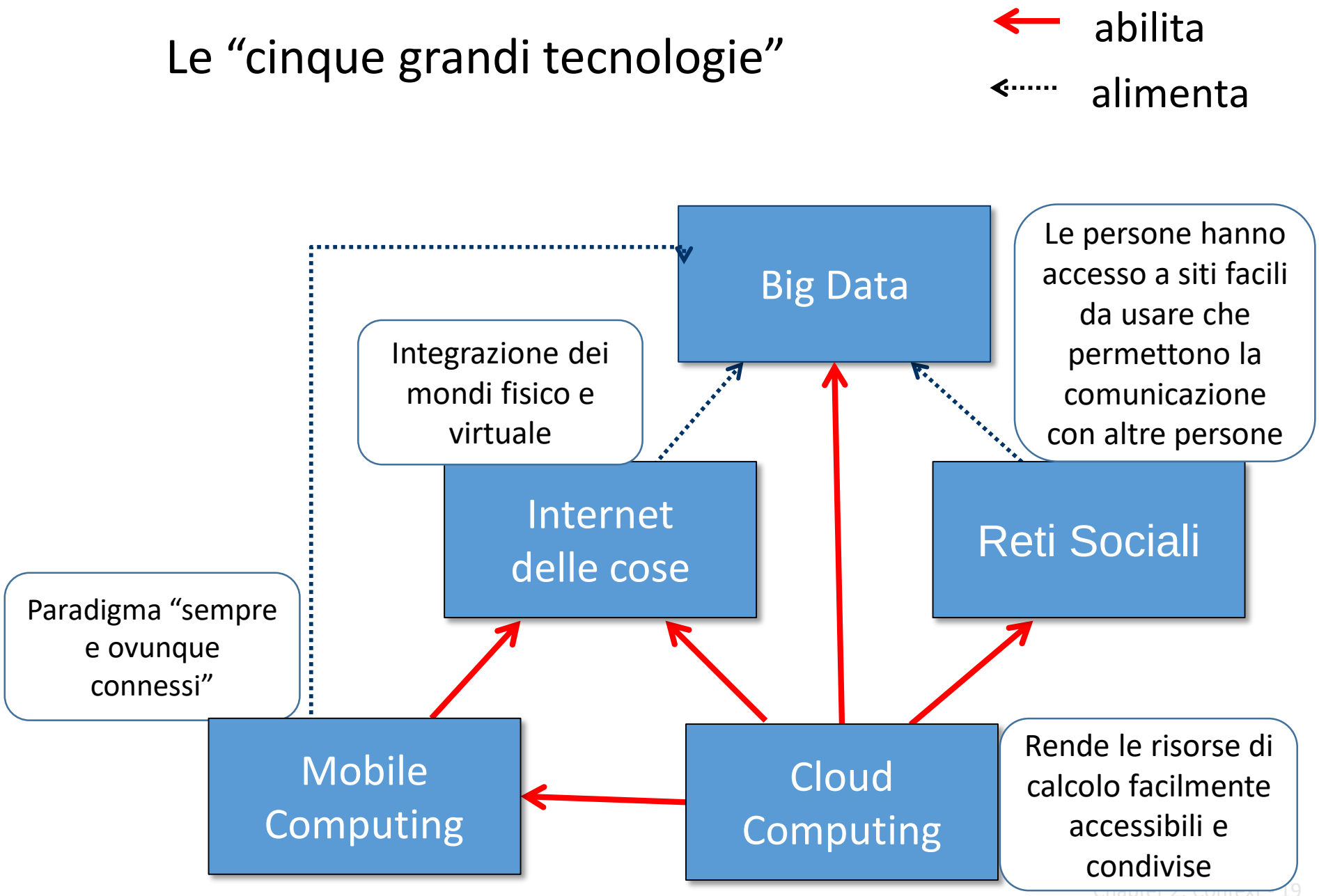
……. ai grandi dati



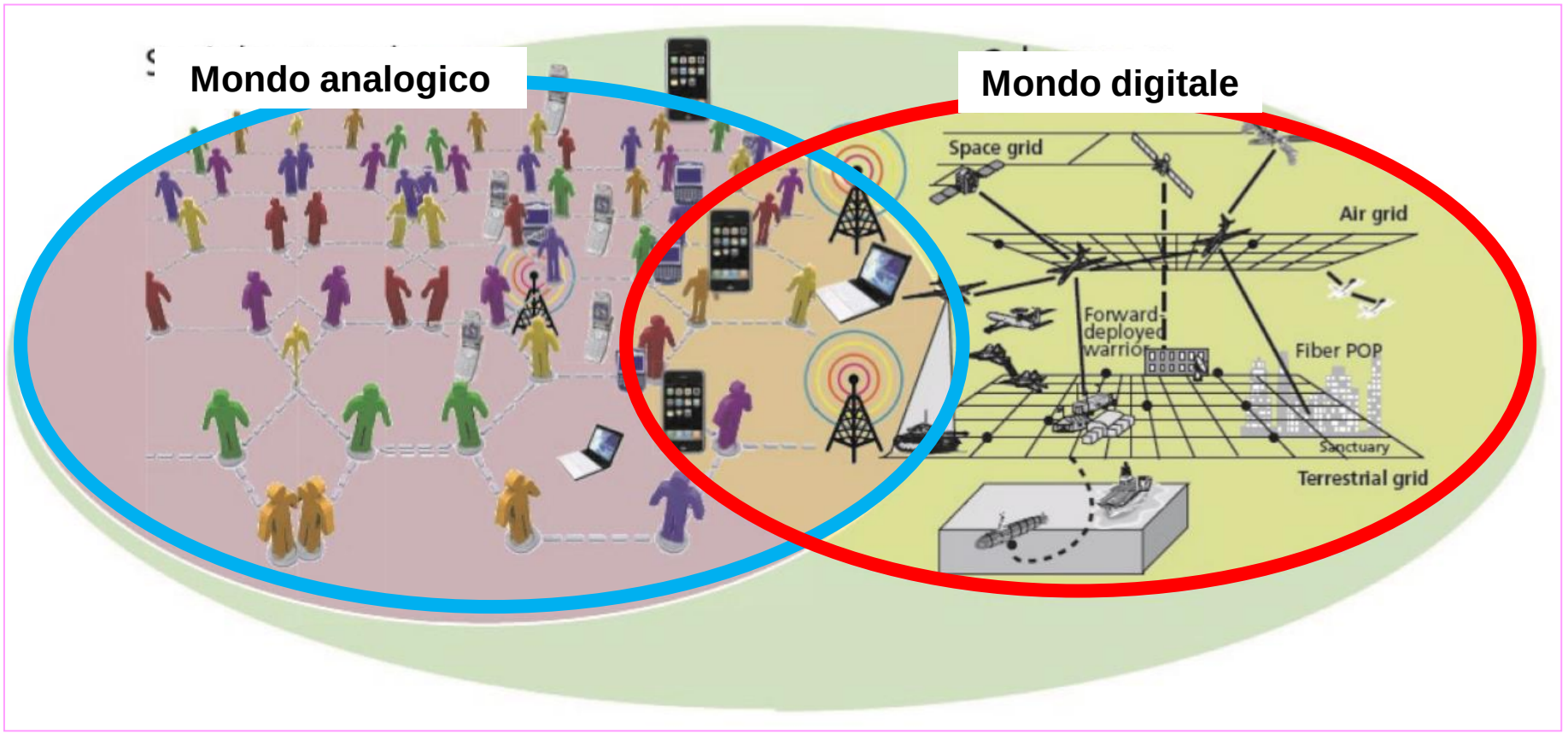
Il genoma umano

Cosa sta rendendo possibile questo «diluvio di dati»?

Le “cinque grandi tecnologie”

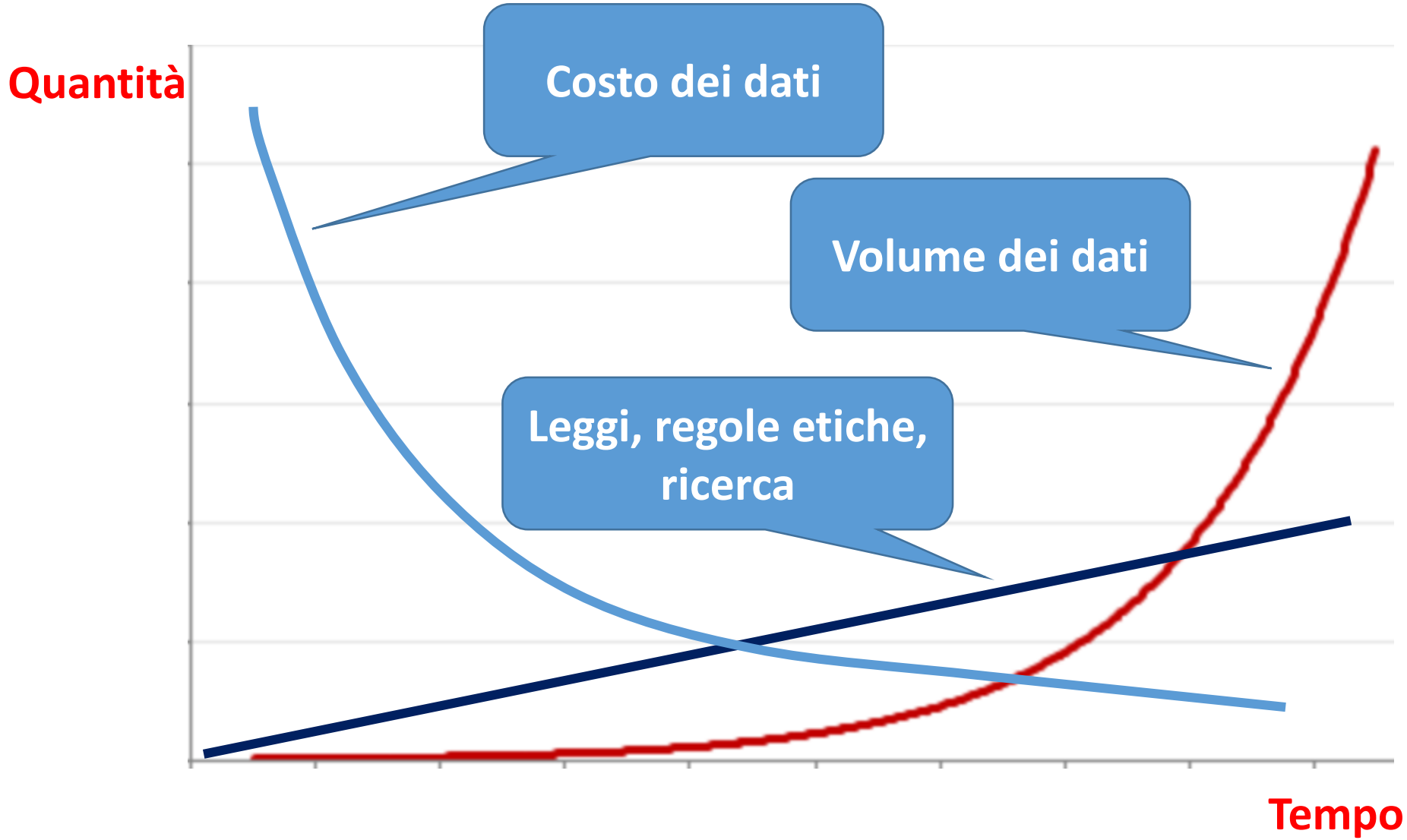


L'ecosistema dei dati
L'infosfera di Floridi
L'altro mondo di Baricco



Ready Player One, di Steven Spielberg

Crescita *esponenziale* dei dati, decrescita dei costi e crescita *lineare* della ricerca



Percentuale di aziende che assumerà per tecnologia entro il 2002

I due pilastri su cui si fonda la Scienza dei dati

Come vanno usati i dati?
Il ciclo di vita del dato digitale

1. **Scelta** delle fonti e acquisizione
2. **Preparazione** (o **Valorizzazione**) dei dati
3. **Analisi** dei dati
4. **Visualizzazione** dei risultati

Lo strumento Breezometer per prevedere i livelli di inquinamento

Dati satellitari

Fondi di dati ambientali

Flussi di trasporto

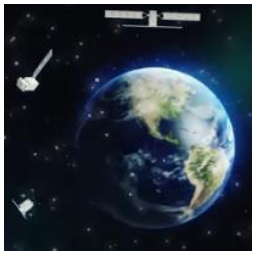
Sensori nei Comuni

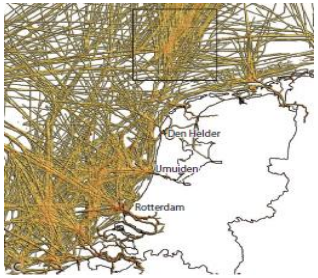
Google Maps

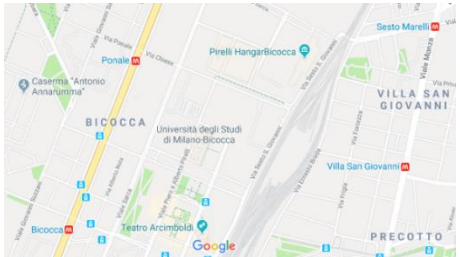
Previsione

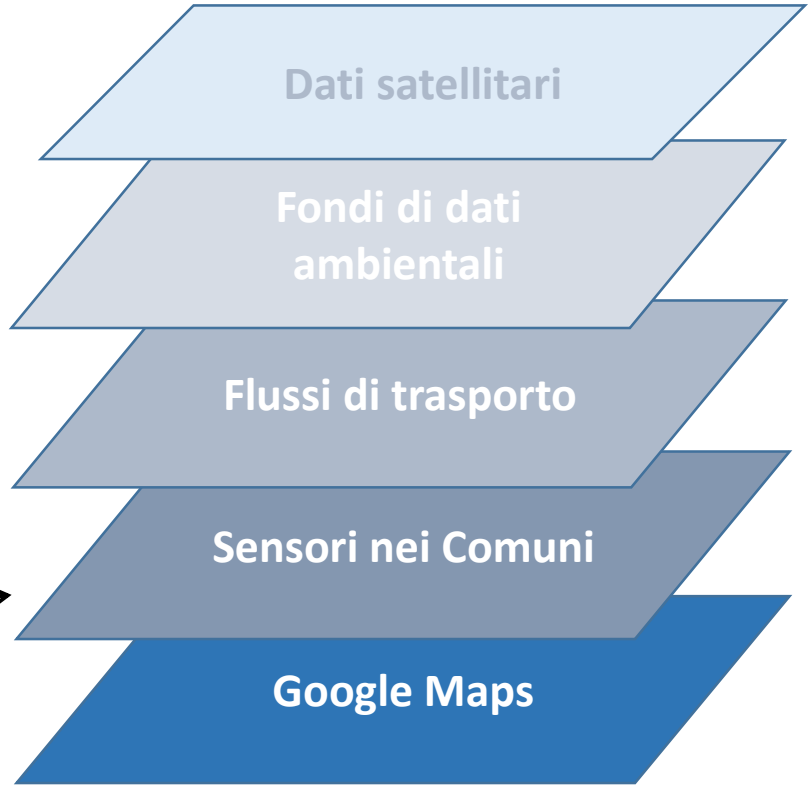
Scelta delle fonti

1. Scelta delle fonti per prevedere l'inquinamento in Breezometer









Valorizzazione

2. Valorizzazione dei dati: Quale di queste due immagini è di migliore **qualità**?

Tradeoff tra qualità

Fedeltà

Leggibilità

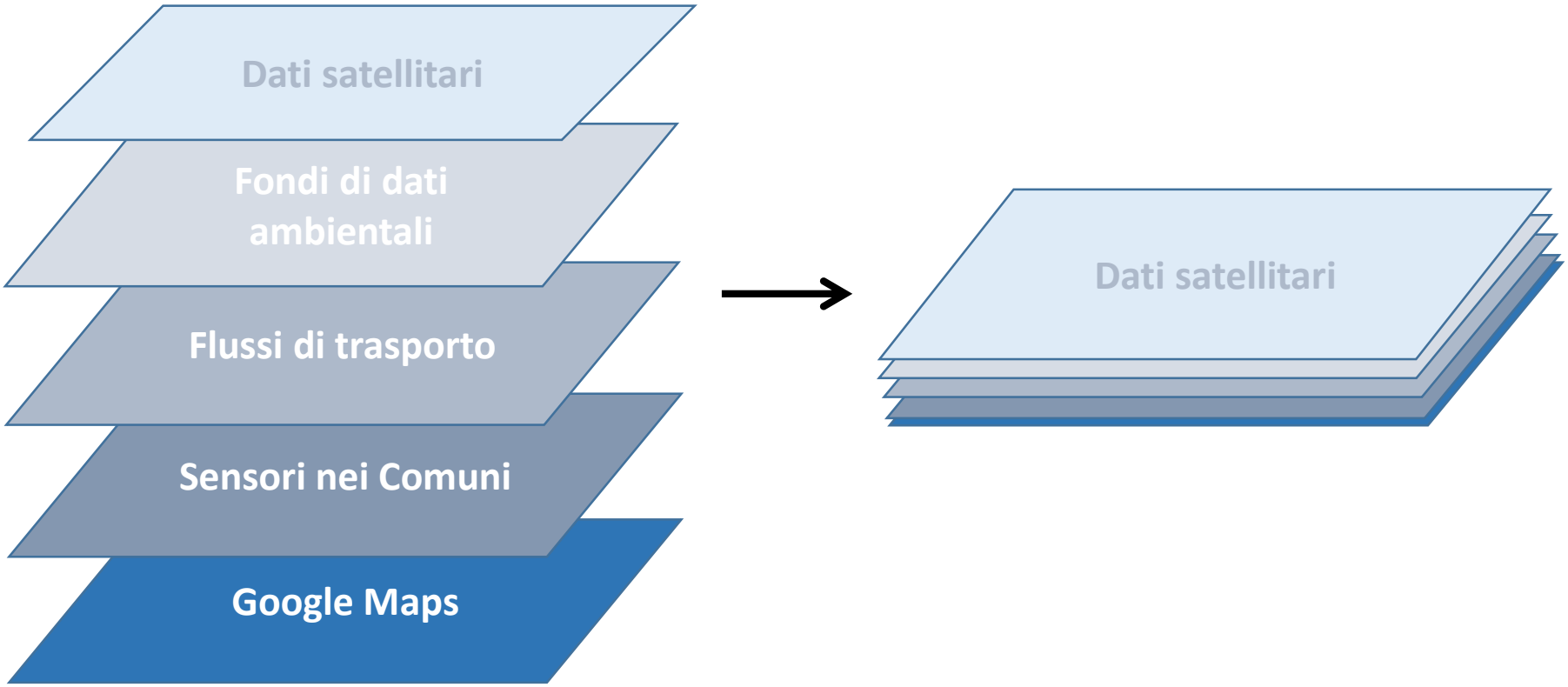
La qualità dei dati nel Web

Queste sono due immagini di Marte.
Secondo te quale è di migliore qualità?

www.hoax-slayer.com

astrobiology.nasa.gov

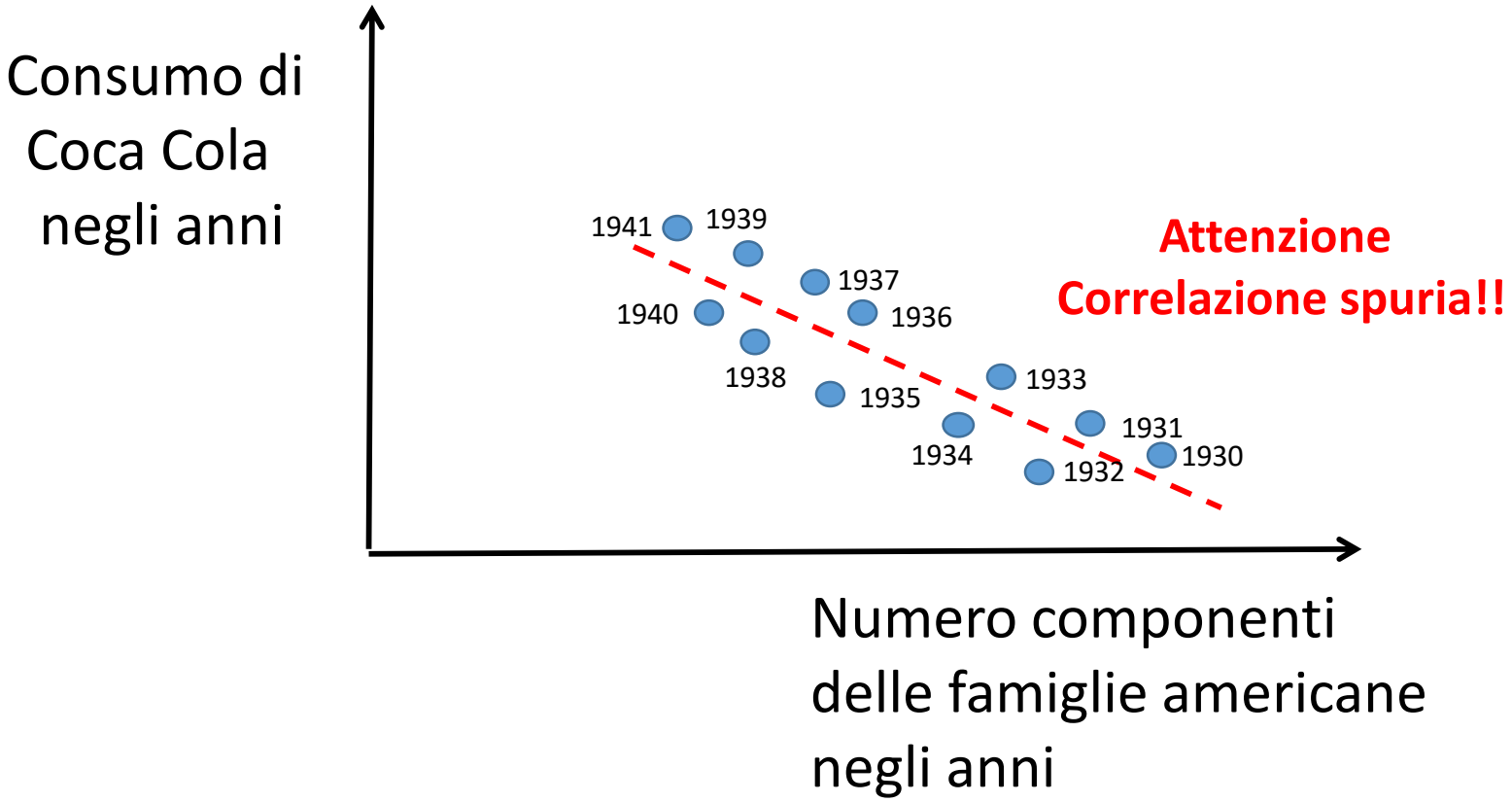
Integrazione delle fonti in Breezometer



Analisi

3. Analisi - Correlazione

Consumo di Coca Cola e numero di componenti delle famiglie americane



3. Analisi per prevedere il giorno del biglietto
Il modello predittivo di Oren Etzioni

Campione di 12.000 biglietti → $200\ 10^9$

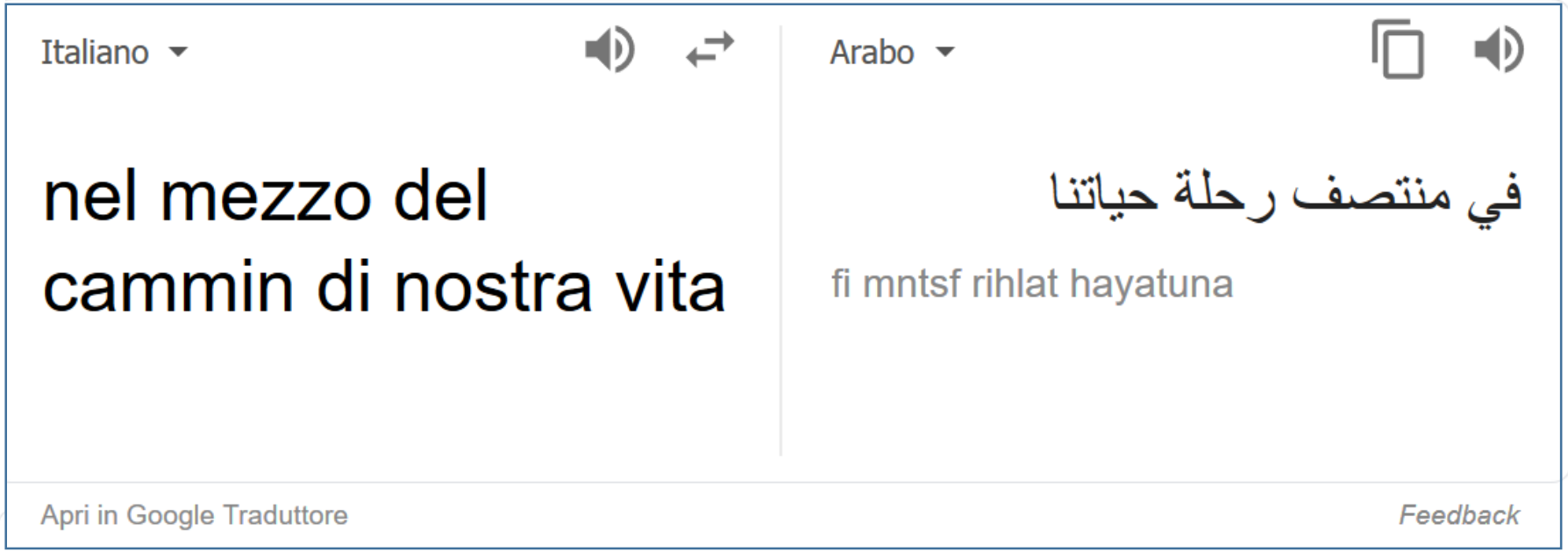
50 $ risparmio medio per biglietto
Compagnia venduta per $110\ 10^6$ $

3. Analisi dei dati in Breezometer
Tecniche di Machine Learning

Machine Learning

Modello predittivo
Inquinamento (latitudine, longitudine, tempo) =
f (ambiente, traffico, altitudine, riscaldamento, …)

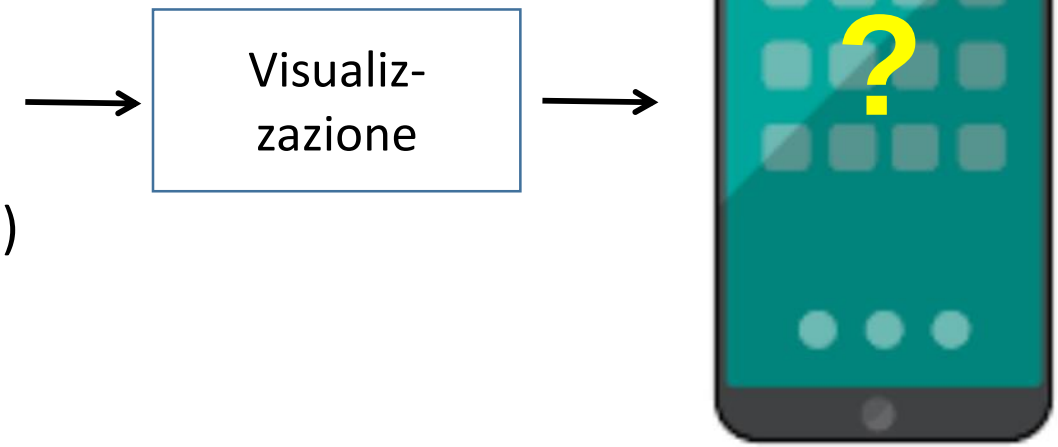
3. Analisi per traduzione - Il traduttore di Google



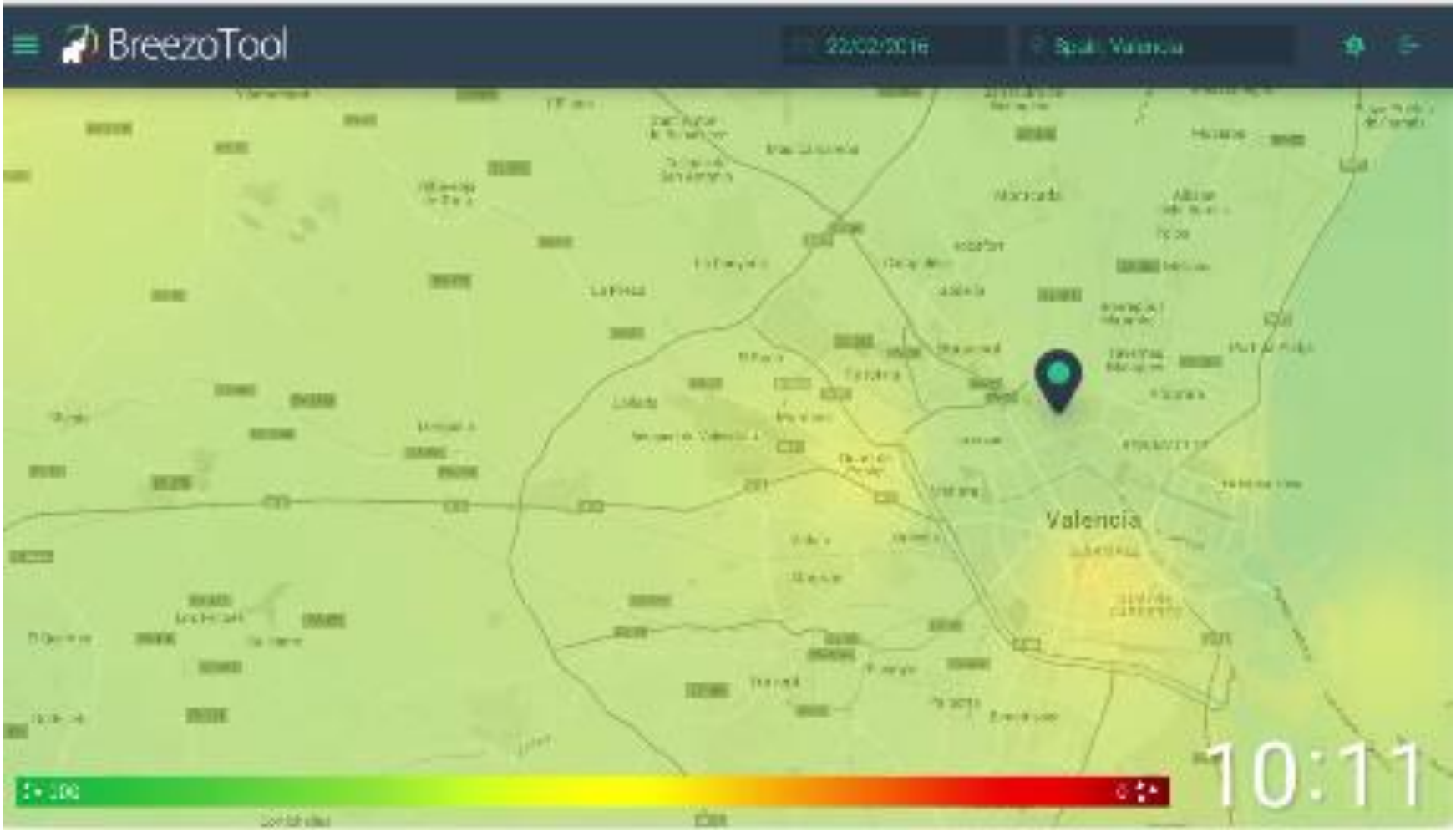
Visualizzazione

4. Visualizzazione dei dati in Breezometer

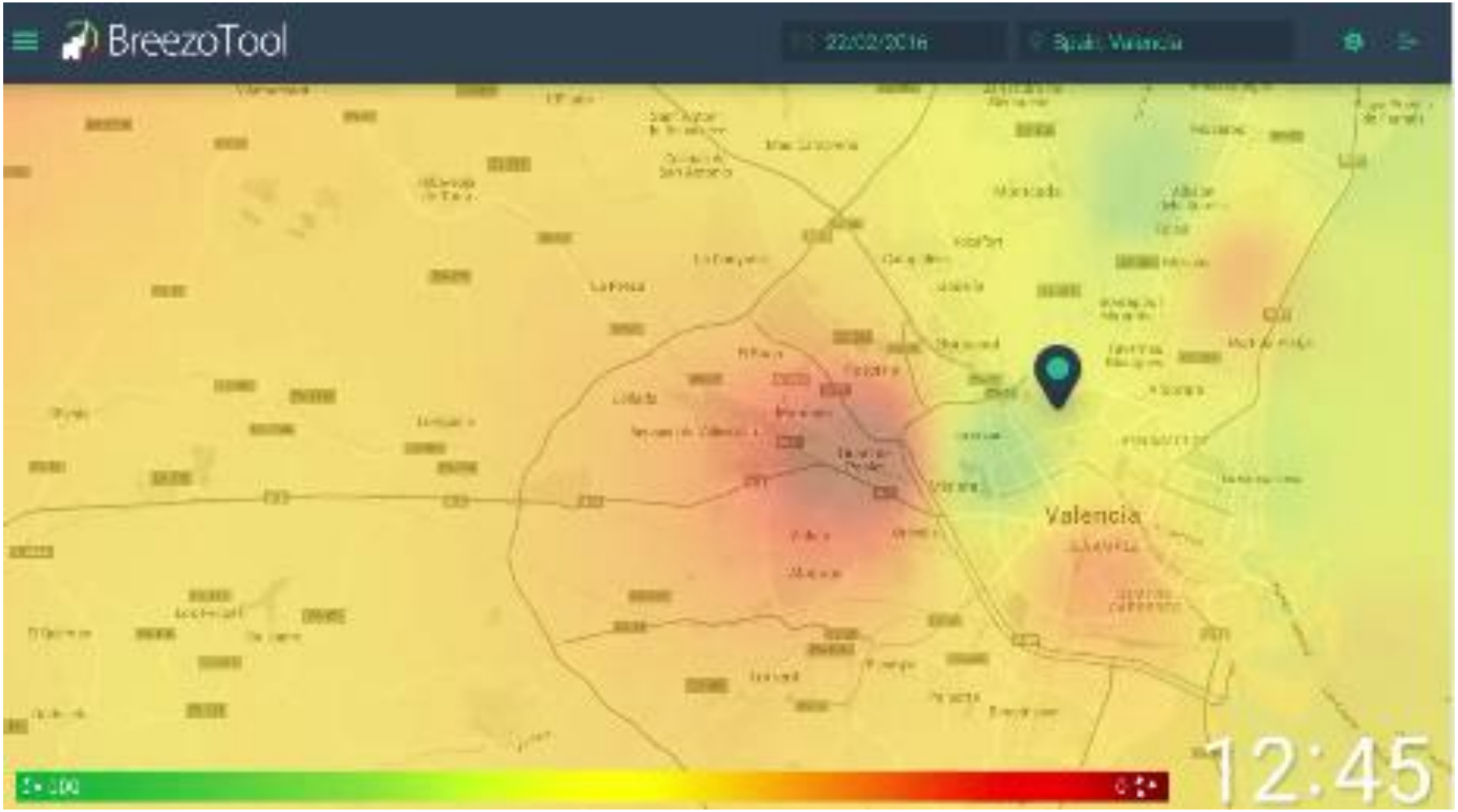
Modello predittivo
Inquinamento (latitudine, longitudine, tempo) =
f (ambiente, traffico, altitudine, riscaldamento, …)



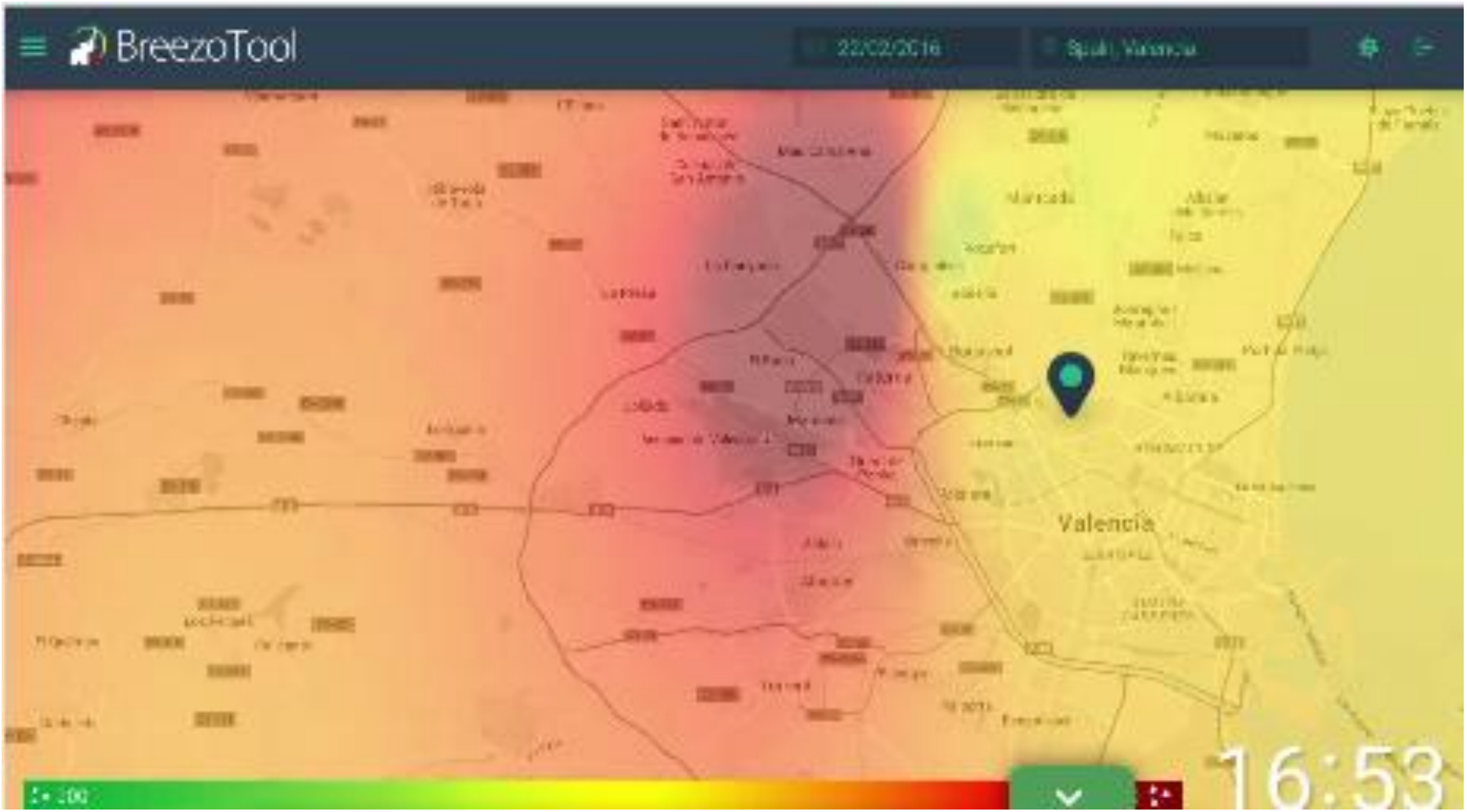
Heat Maps relative a varie ore del giorno
Ore 10:11



Ore 12.45

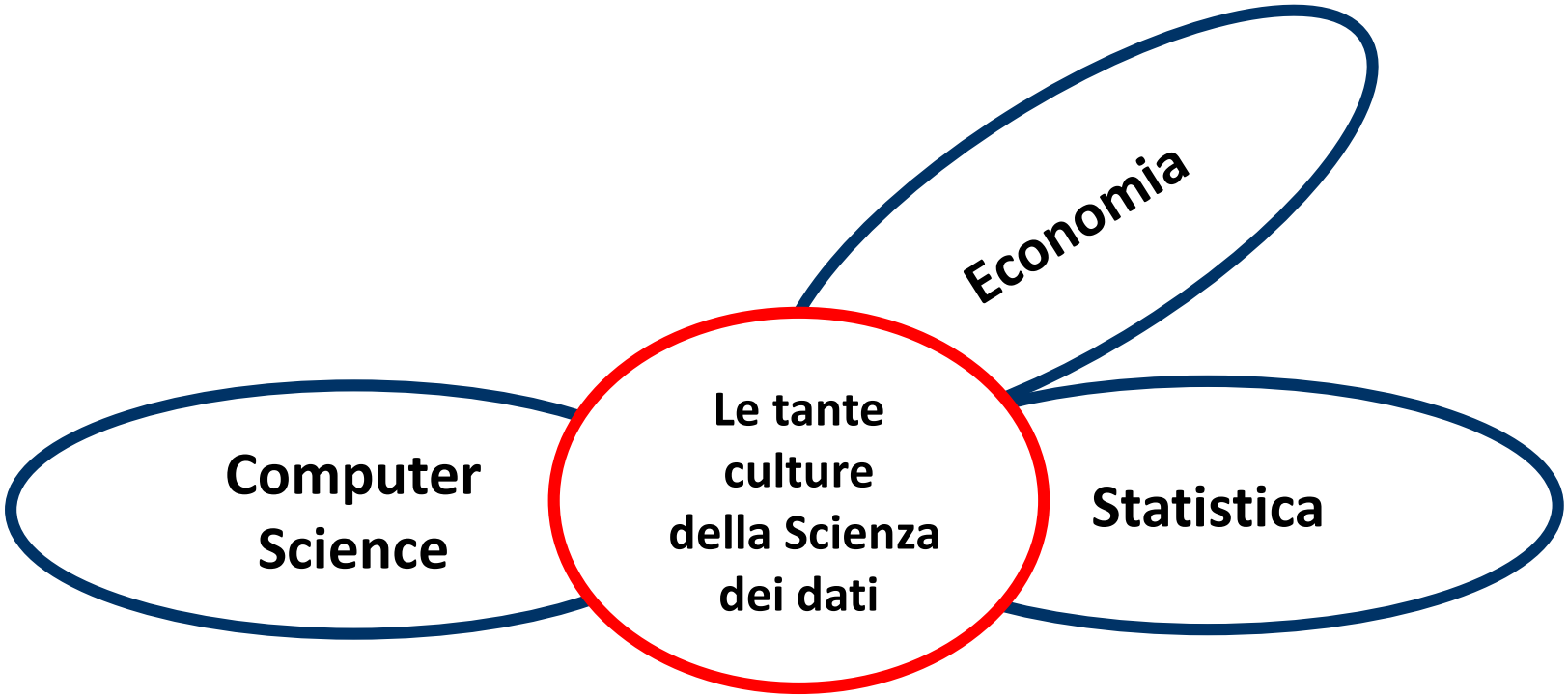


Ore 16.53



La Scienza dei Dati: una nuova Scienza sulle spalle dei giganti

Una nuova Scienza



Beni, dati, servizi

Jeans

Dati

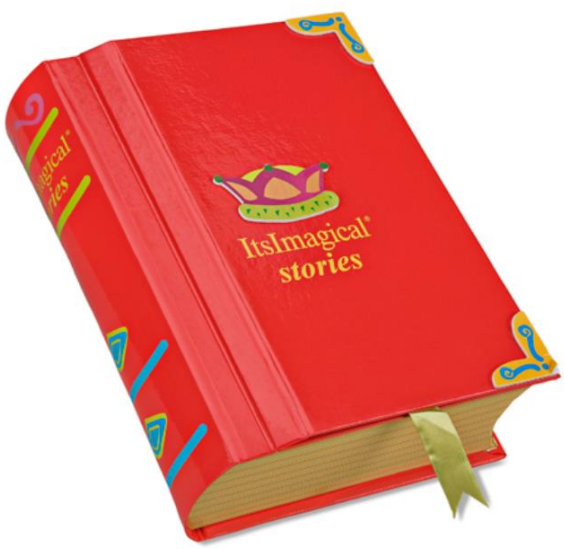
?

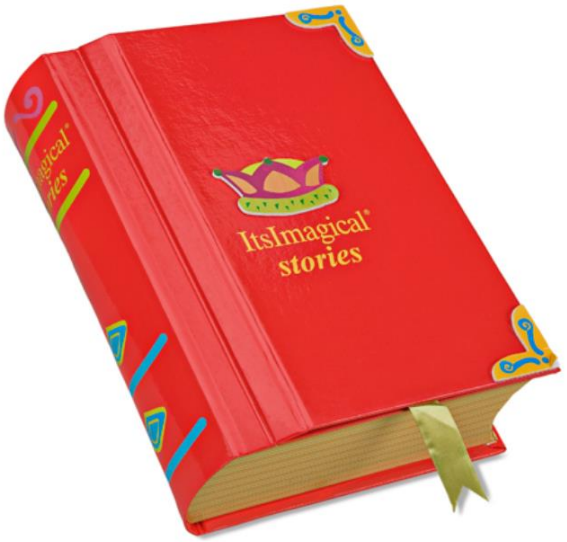
Seduta di supporto psicologico

I dati non sono materiali come i beni

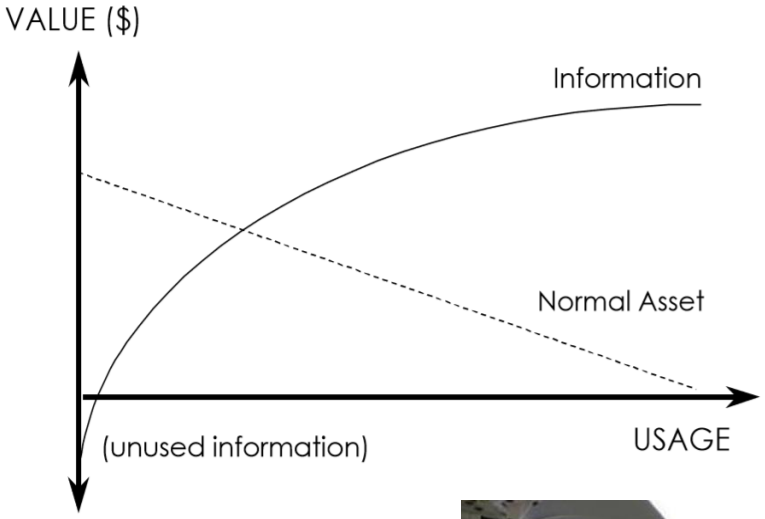
```
Source:   query [8.074e+07 x 5]
Database: spark connection master=local[8] app=sparklyr local=TRUE

          user_id    item_id rating  timestamp             category
            <chr>      <chr>  <dbl>      <int>                <chr>
1  A1EE2E3N7PW666 B000GFDAUG      5 1202256000 Amazon Instant Video
2   AGZ8SM1BGK3CK B000GFDAUG      5 1198195200 Amazon Instant Video
3  A2VHZ21245KBT7 B000GIOPK2      4 1215388800 Amazon Instant Video
4   ACX8YW2D5EGP6 B000GIOPK2      4 1185840000 Amazon Instant Video
5   A9RNMO9MUSMTJ B000GIOPK2      2 1281052800 Amazon Instant Video
6  A3STFVPM8NHJ7B B000GIOPK2      5 1203897600 Amazon Instant Video
7  A2582KMXLK2P06 B000GIOPK2      5 1205884800 Amazon Instant Video
8  A1TZCLCW9QGGBH B000GIOPK2      4 1209427200 Amazon Instant Video
9  A2E2I6B878CRMA B000GIOPK2      5 1378684800 Amazon Instant Video
10  AD5MZA8SOVMPJ B000GIOPK2      5 1218240000 Amazon Instant Video
# ... with 8.074e+07 more rows
```

Un libro, due libri

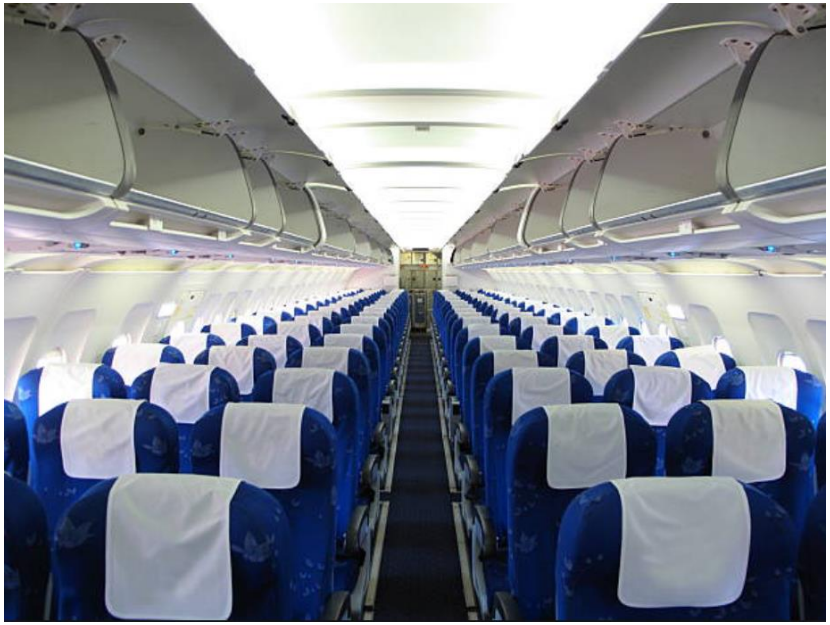




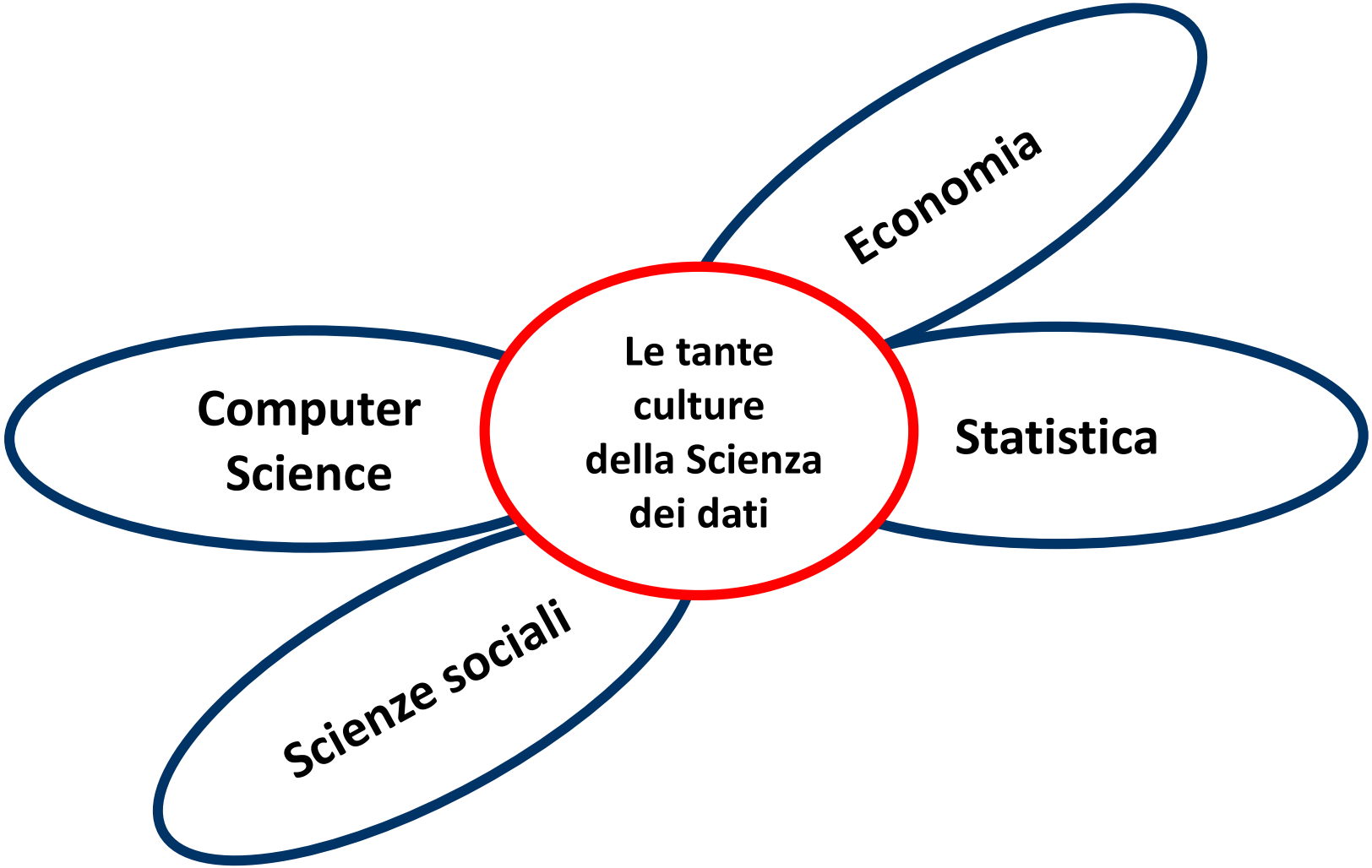
I dati non svaniscono come i servizi




```
Source:   query [8.074e+07 x 5]
Database: spark connection master=local[8] app=sparklyr local=TRUE

           user_id     item_id rating  timestamp             category
             <chr>       <chr>  <dbl>      <int>                <chr>
1  A1EE2E3N7PW666 B000GFDAUG       5 1202256000 Amazon Instant Video
2   AGZ8SM1BGK3CK B000GFDAUG       5 1198195200 Amazon Instant Video
3  A2VHZ21245KBT7 B000GIOPK2       4 1215388800 Amazon Instant Video
4   ACX8YW2D5EGP6 B000GIOPK2       4 1185840000 Amazon Instant Video
5   A9RNMO9MUSMTJ B000GIOPK2       2 1281052800 Amazon Instant Video
6  A3STFVPM8NHJ7B B000GIOPK2       5 1203897600 Amazon Instant Video
7  A2582KMXLK2P06 B000GIOPK2       5 1205884800 Amazon Instant Video
8  A1TZCLCW9QGGBH B000GIOPK2       4 1209427200 Amazon Instant Video
9  A2E2I6B878CRMA B000GIOPK2       5 1378684800 Amazon Instant Video
10  AD5MZA8SOVMPJ B000GIOPK2       5 1218240000 Amazon Instant Video
# ... with 8.074e+07 more rows
```



Una nuova Scienza



Valore sociale dei dati: Ospedali e qualità della cura in Uganda The Economist 2011

The Economist

World politics | Business & finance | Economics | Science & technology | Culture | Blogs | Debate & discuss | Multimedia | Print edition

All audio
All video
The Economist audio edition

The Open Government Partnership

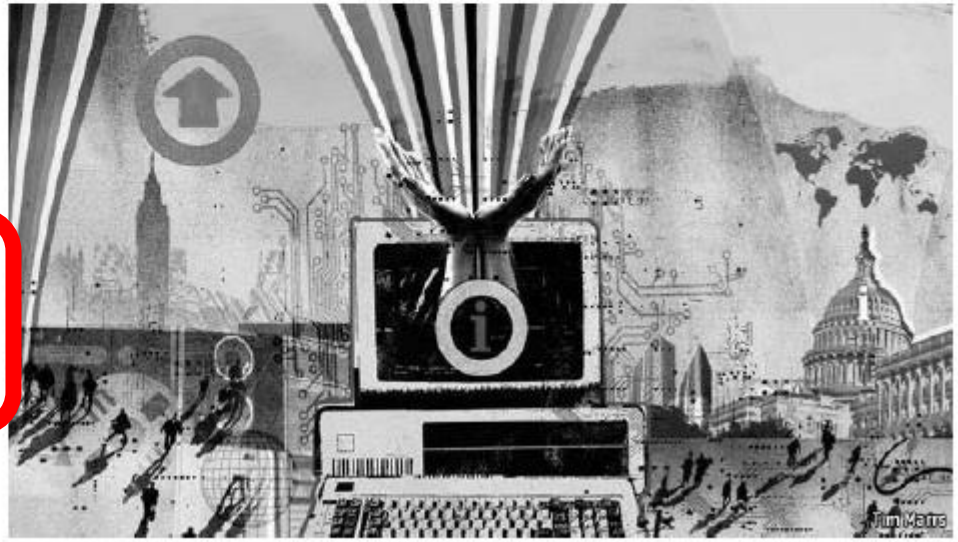
The parting of the red tape

Is it just another global talking-shop—or a fresh approach to shaking out government secrecy?

Oct 8th 2011 | NEW YORK AND TALLINN | from the print edition

Like 151 0

UGANDA is not best known as a testbed for new ideas in governance. But research there by Jakob Svensson at the University of Stockholm and colleagues suggested that giving people health-care performance data and helping them organise to submit complaints cut the death rate in under-fives by a third. Publishing data on school budgets reduced the misuse of funds and increased enrolment.

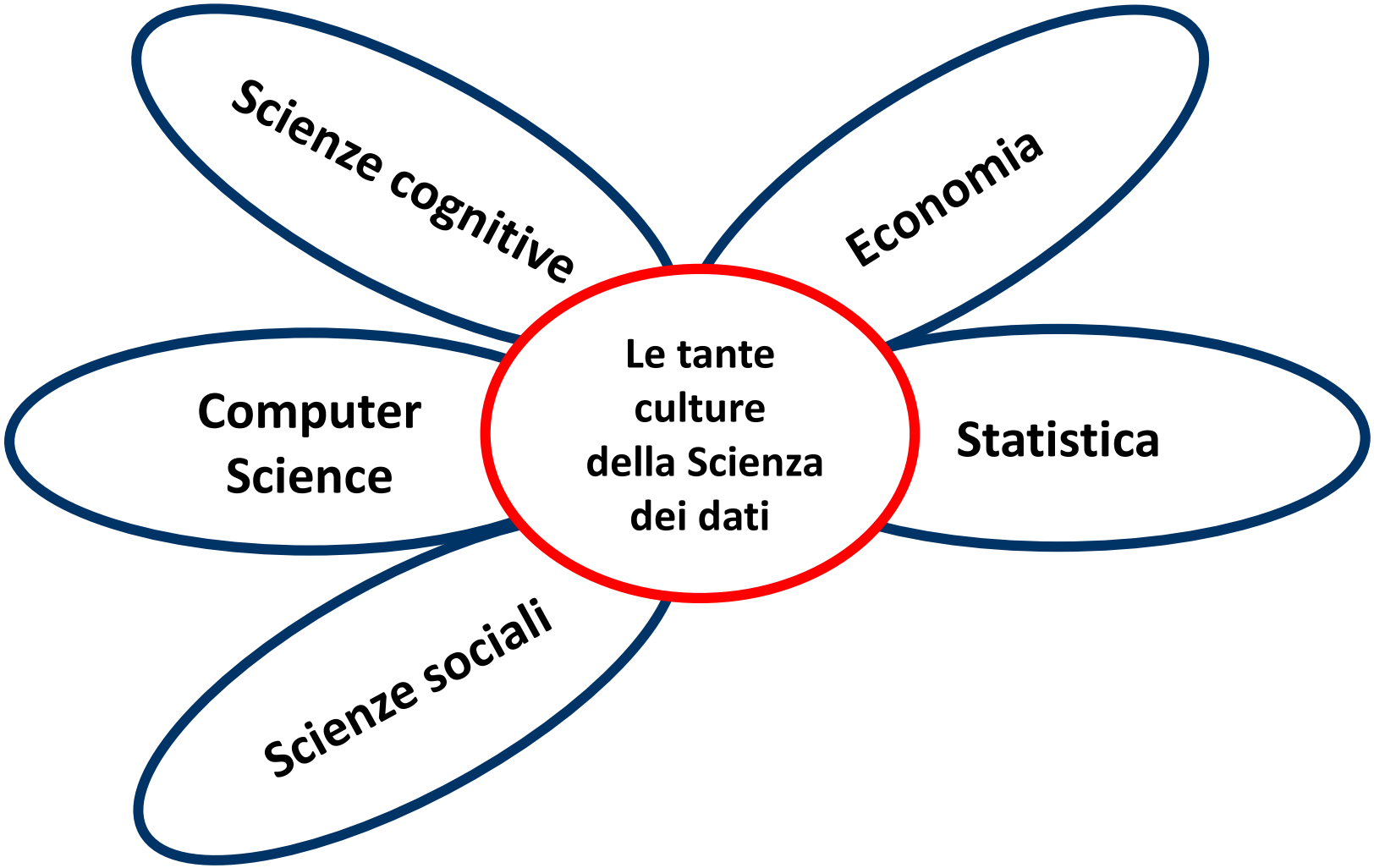


Data divide nelle mappe

Times Square, New York

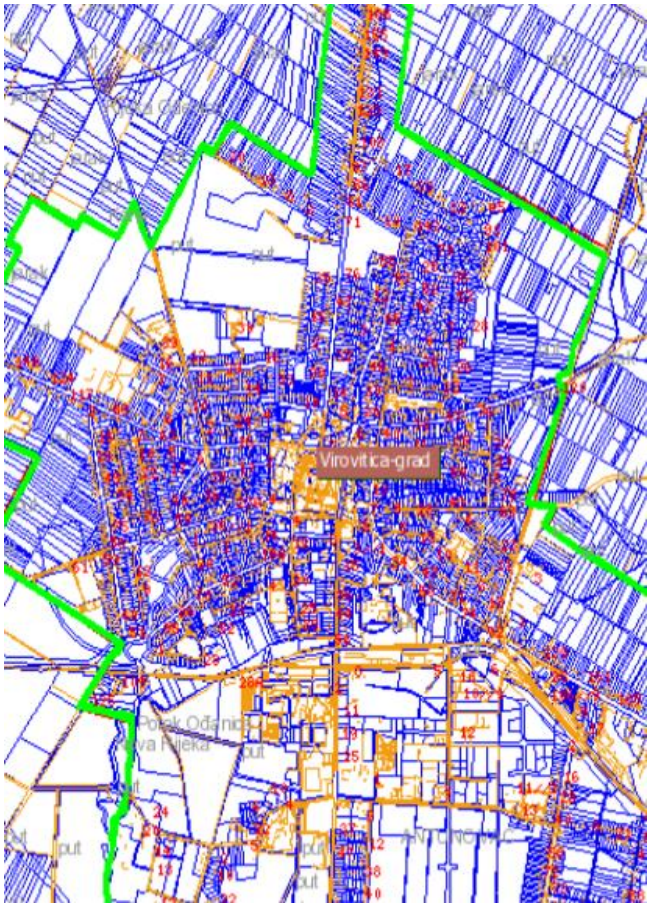
Somaliland

Una nuova Scienza

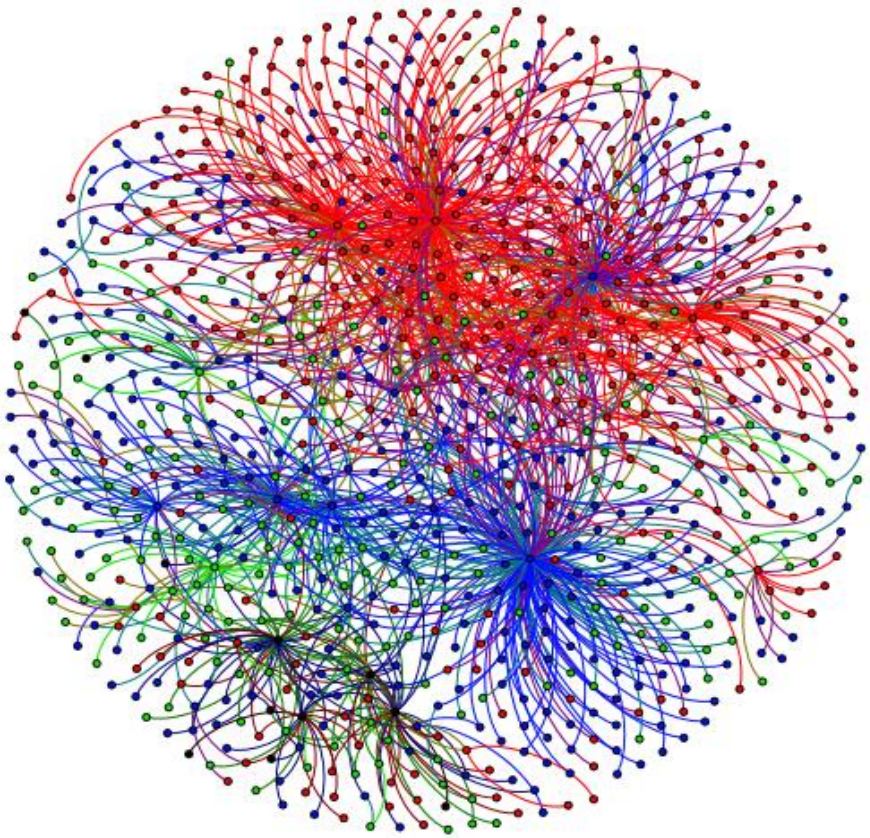


Dati catastali in India

Dati catastali in India

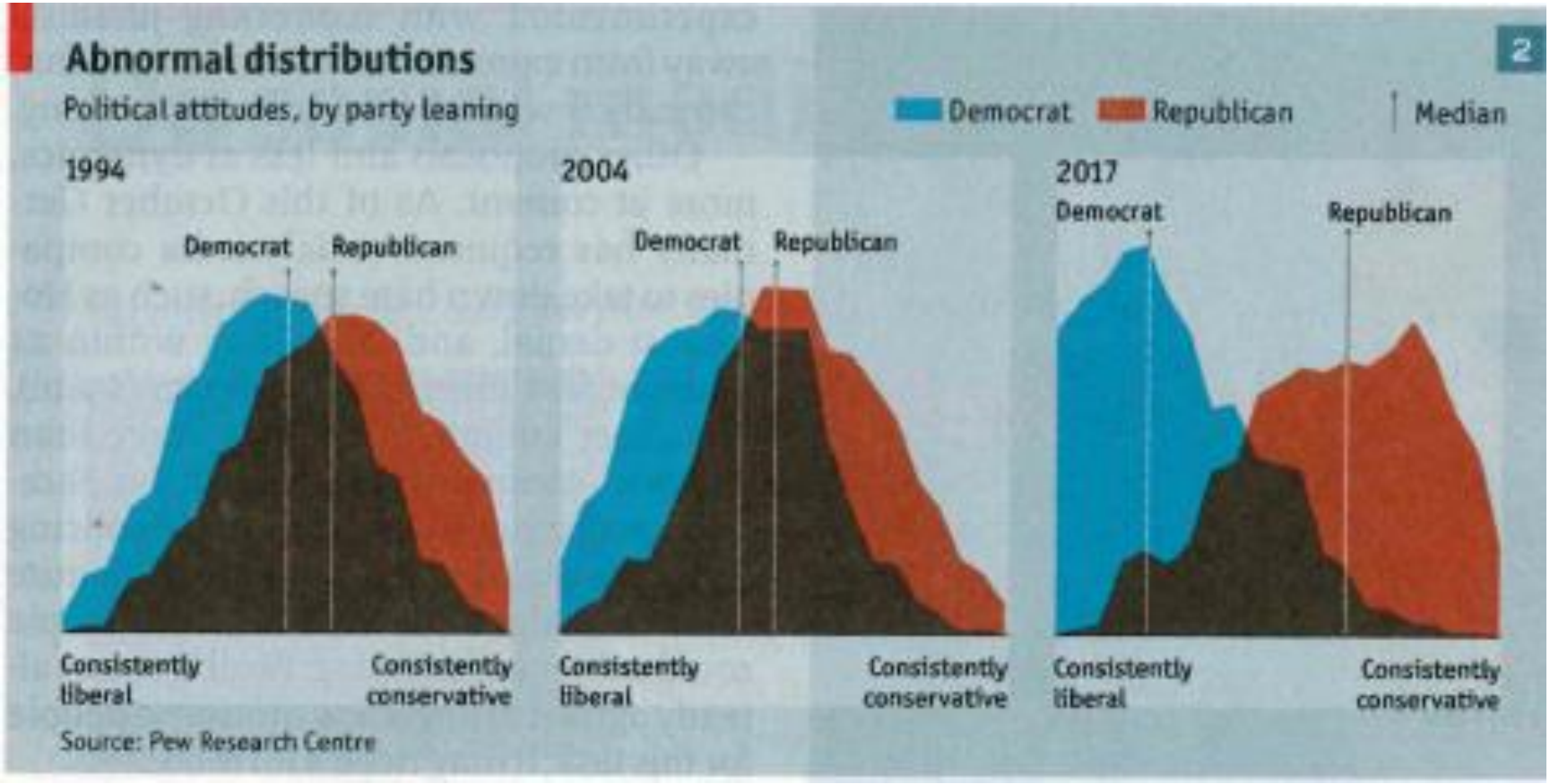


La rabbia è molto più presente della gioia



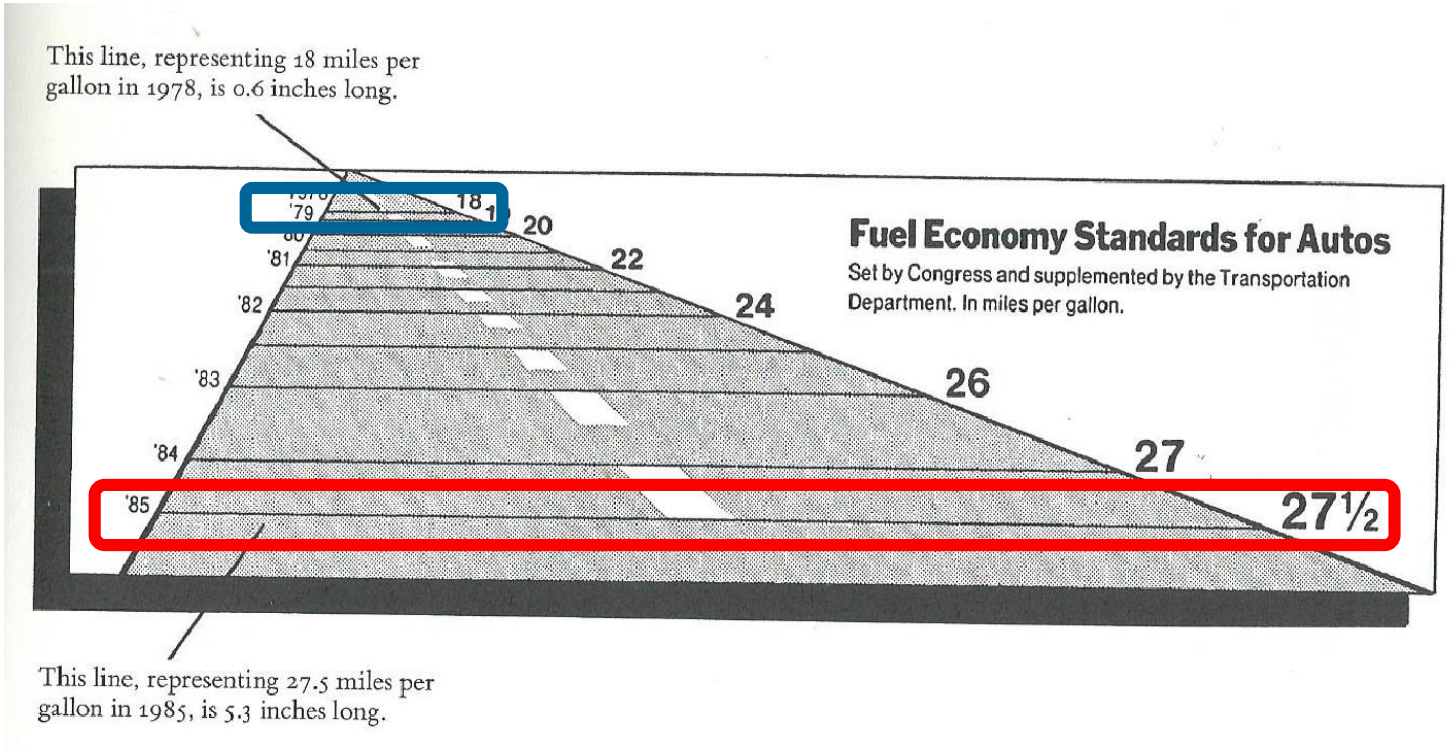
- **Rosso** sta per rabbia,
- **Verde** sta per gioia,
- **Blu** è la tristezza
- **Nero** rappresenta il disgusto.

Polarizzazione delle opinioni politiche nella popolazione USA - Economist 4/11/2017



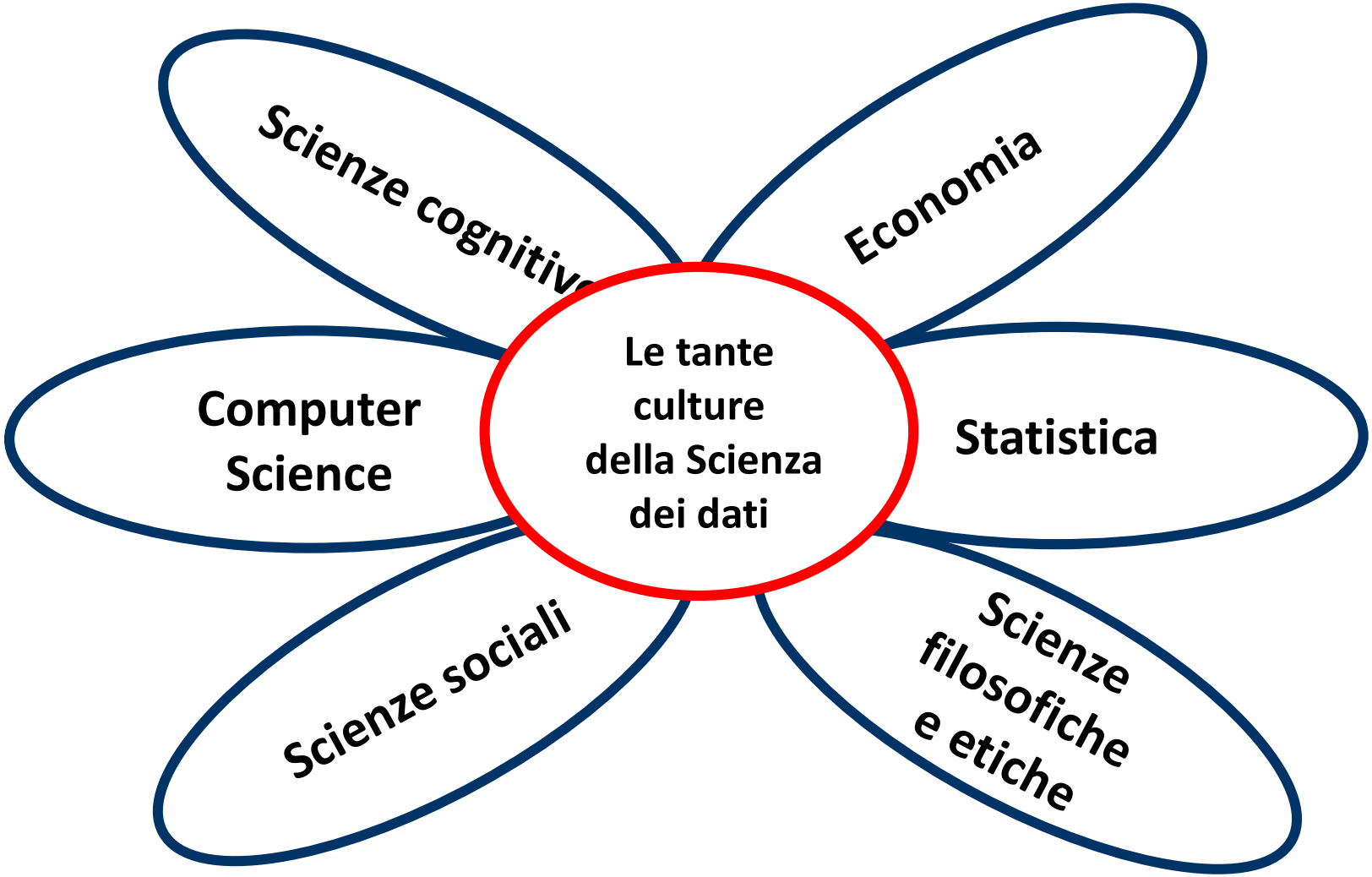
Quante bugie nelle visualizzazioni

Year	Miles per gallon
1978	18
1979	19
1980	20
1981	22
1982	24
1983	26
1984	27
1985	27,5

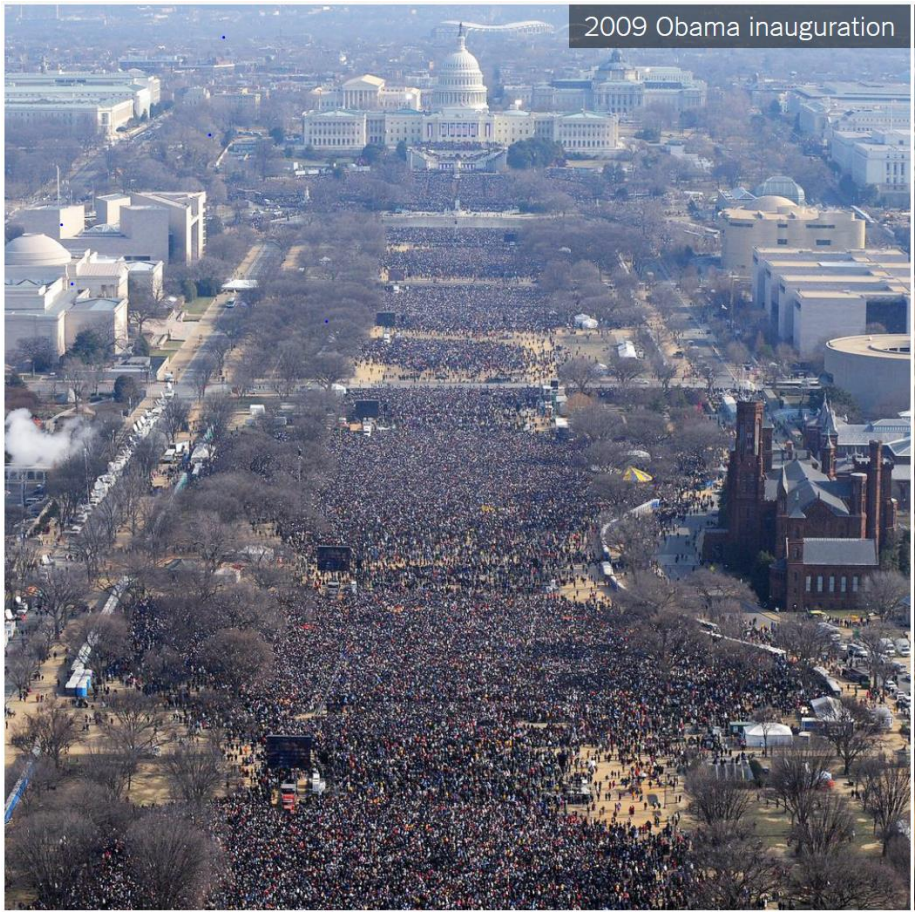


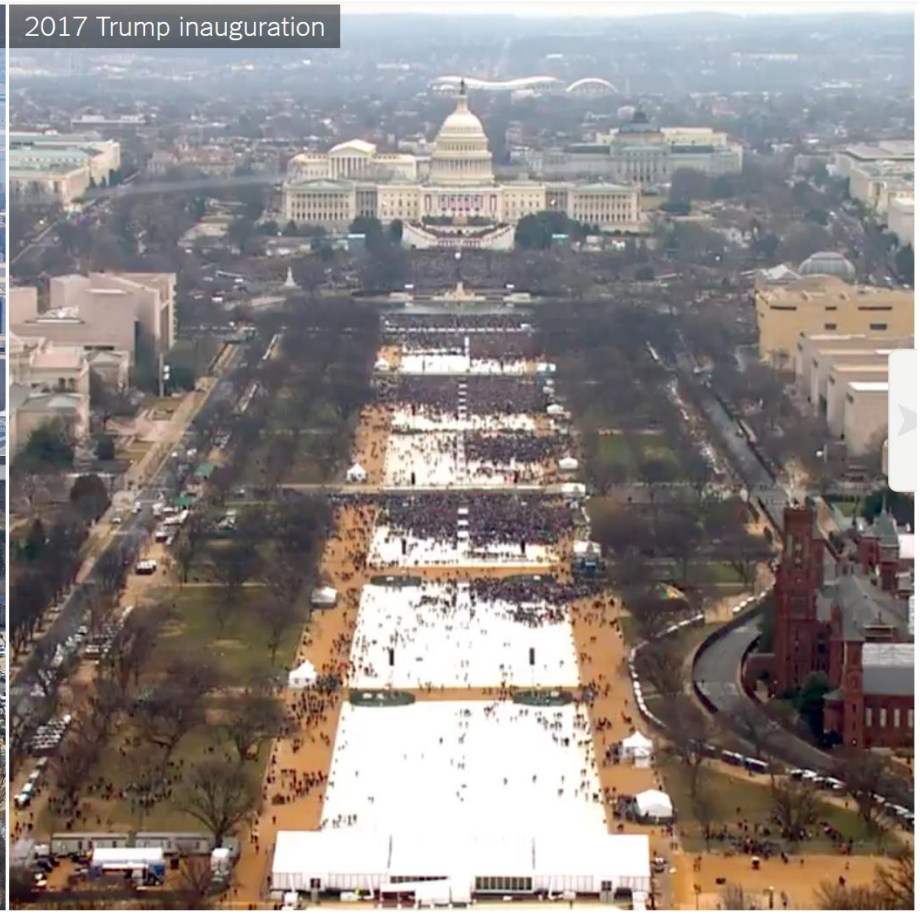
Livello di bugia = rapporto tra valori numerici nel mondo reale / rapporto tra valori numerici nella visualizzazione = 15

Una nuova Scienza



Le cerimonie di insediamento di Obama e di Trump





I fatti alternativi di Kellyanne Conway

Kellyanne Conway denies Trump press secretary lied: 'He offered alternative facts'

Dalla psicologia cognitiva non arrivano buone notizie

Non è tanto rilevante ciò che la gente pensa, ma come pensa. Riconoscere la cattiva informazione richiede processi cognitive complessi

Un semplice mito è più attrattivo cognitiva-mente di una complicata correzione

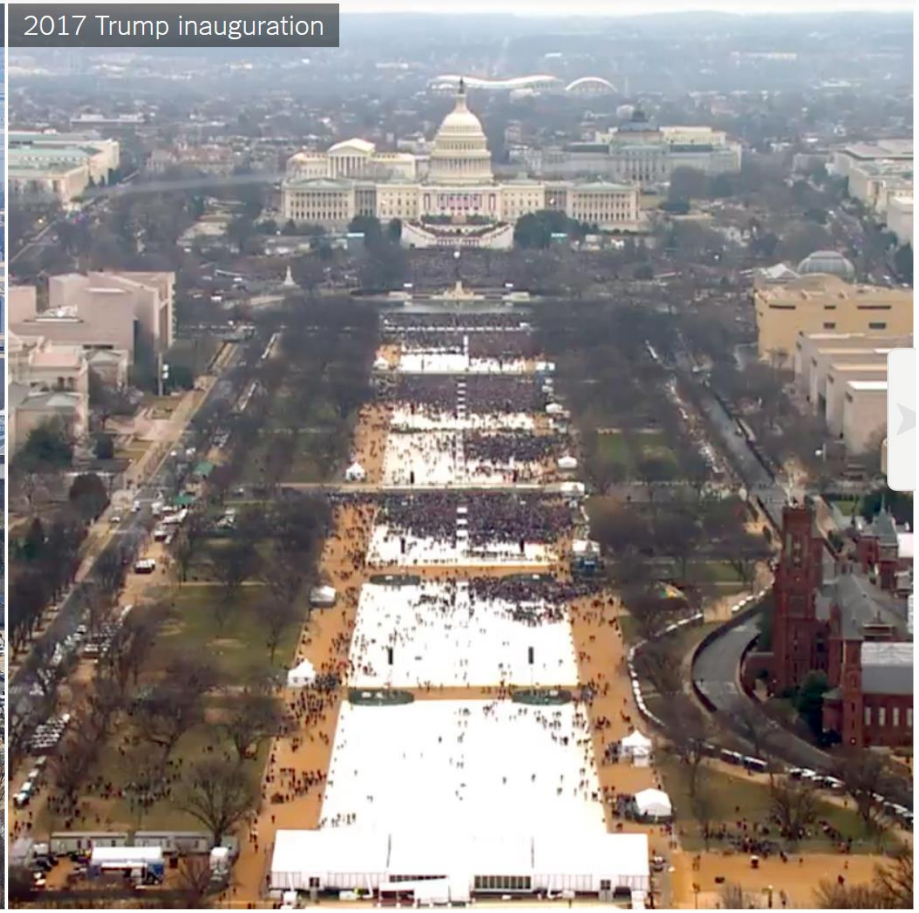
Per coloro che sono fortemente convinti delle proprie idee, gli argomenti fortemente contrari possono rafforzare le loro convinzioni

Come possiamo capire chi ha ragione?
I fatti sono testardi

Ora della foto: 11.30

Ora della cerimonia: 12

Numero biglietti metropolitana nell'ora precedente: 80.000

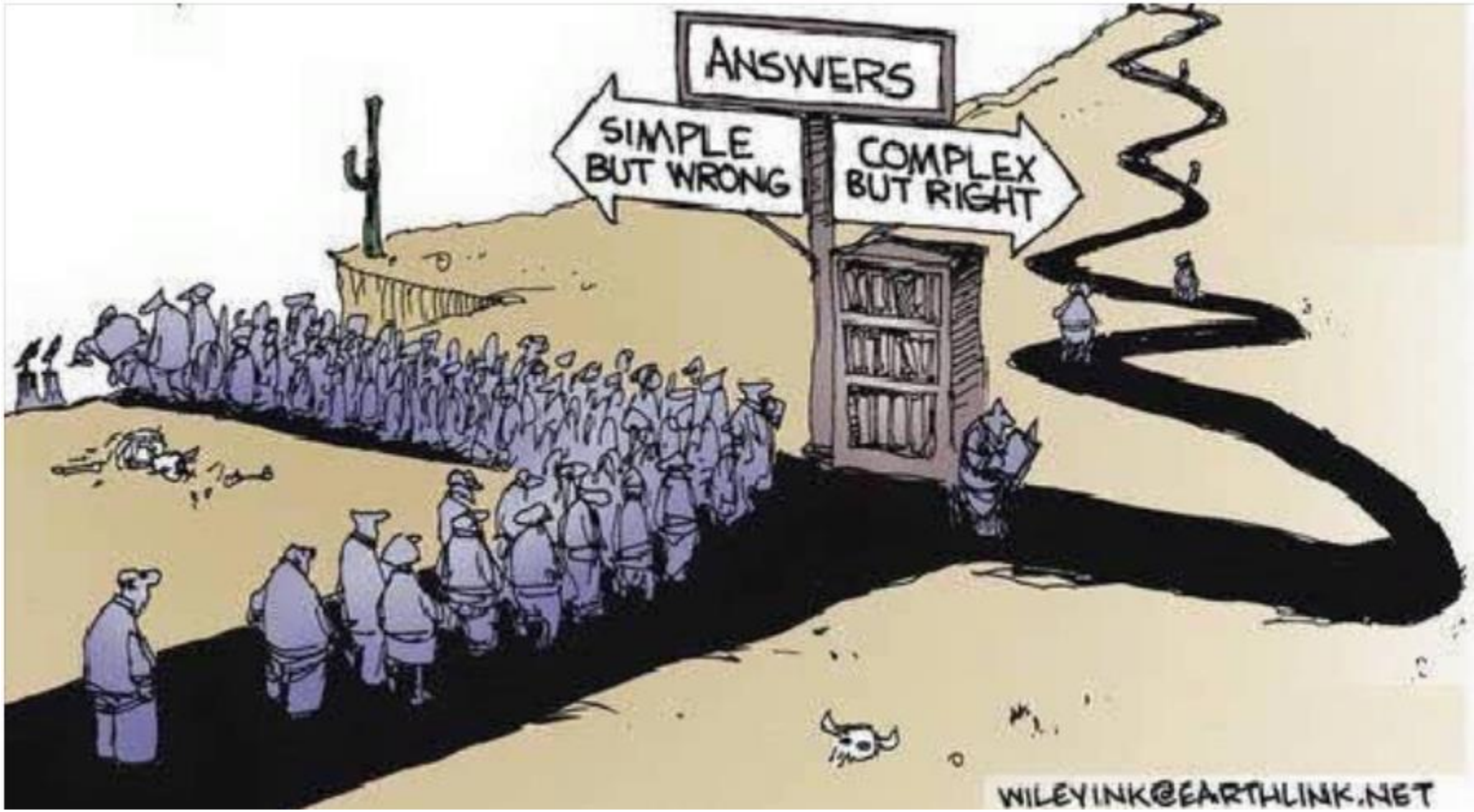


Ora della foto: 11.30

Ora della cerimonia: 12

Numero biglietti metropolitana nell'ora precedente: 20.000

Non ci sono risposte semplici a domande complicate



Literacy o alfabetizzazione

Numeracy, o capacità di far di conto

Una Definizione di Datacy: capacità di …

- applicare tecniche e modelli statistici e informatici basati su **apprendimento** per costruire modelli decisionali, interpretativi e predittivi.
- **risolvere problemi** con il supporto dei dati e prendere decisioni complesse.
- comprendere l'impatto sulla **economia** e sulla **società** del fenomeno dei dati digitali.
- analizzare i **corpi giuridici** sviluppati dalle istituzioni pubbliche in tema di dati digitali
- affrontare i nuovi temi **etici** che nascono dall'uso dei dati digitali.

L'INNOVAZIONE DIGITALE, LA CULTURA DEI DATI DIGITALI E LA SCIENZA DEI DATI
11 febbraio 2019 – 29 aprile 2019

Descrizione del progetto

- Il corso intende presentare alcune tematiche di maggiore rilevanza per l'educazione nella innovazione digitale e in particolare nella nascente disciplina della Scienza dei dati nella scuola secondaria superiore. Le lezioni saranno tenute da esperti attivi in ambito universitario e nelle aziende.

Contenuti del progetto

- Destinatari Docenti scuola secondaria II grado Durata 24 ore Per la validità del corso è necessaria la frequenza del 75% delle ore previste (minimo 15 ore) Sedi Istituto Lombardo di Cultura e Università di Milano-Bicocca

Programma del corso

Obiettivi formativi

- Gli obiettivi del corso consistono nel fornire agli animatori digitali e ai docenti di scuola secondaria di II grado i riferimenti concettuali e gli strumenti di lavoro al fine di sviluppare attività formative sui temi della cultura digitale e della scienza dei dati in ambito scolastico. Ogni lezione prevedrà una presentazione generale del tema seguita da un'illustrazione in forma di demo di idee e strumenti per attività informative e formative che potranno essere impartite indipendentemente dai docenti.
- L'insieme dei seminari costituenti il progetto non potrà affrontare tutti i temi descritti in precedenza nell'ambito della Scienza dei dati, vista la loro ampiezza e articolazione. E' possibile approfondire alcuni temi in relazione a a un possibile trasferimento in esperienze didattiche nella scuola secondaria. Essi riguardano, oltre a una introduzione generale su innovazione digitale e scienza dei dati, i linguaggi e gli strumenti per l'eLearning e per l'analisi dei dati digitali, le applicazioni in ambito scientifico e aziendale, il tema del data journalism, e approfondimenti sulla trasparenza nelle tecniche di apprendimento e sull'impatto di dati e algoritmi nella economia digitale.

Iscrizioni

- Numero massimo di corsisti: 100 (fino a esaurimento posti)
- Scadenza iscrizioni: 5 febbraio 2019

Per iscriversi al corso è necessario seguire entrambe le modalità di registrazione:

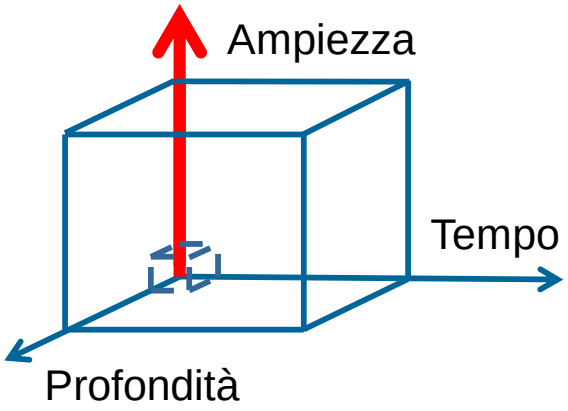
1) Compilare la scheda di iscrizione on-line: https://goo.gl/forms/OFa0B1JYgP4UI9oE3

2) Accreditarsi attraverso la piattaforma S.O.F.I.A. | Codice identificativo: 21760 | Codice identificativo: 31382

Programma di alfabetizzazione

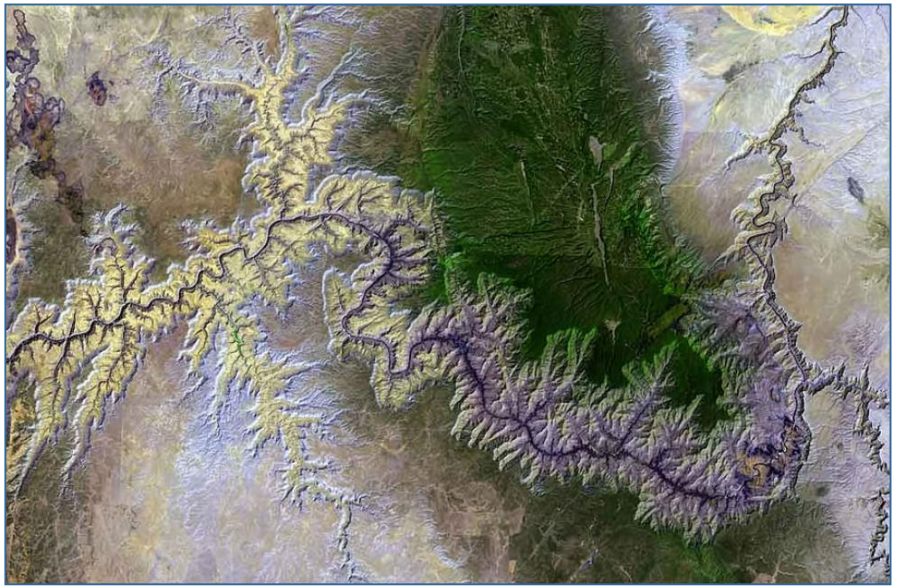
- Le **basi della Scienza dei dati**, con tecniche, applicazioni e temi metodologici
- I **linguaggi della Scienza dei dati**
  1. **Linguaggio R**
  2. **Linguaggio Python**
- **Corsi multimediali** (video + testi di approfondimento) con tutor attivo, e verifiche e certificazione finale con domande a risposta multipla

Left over



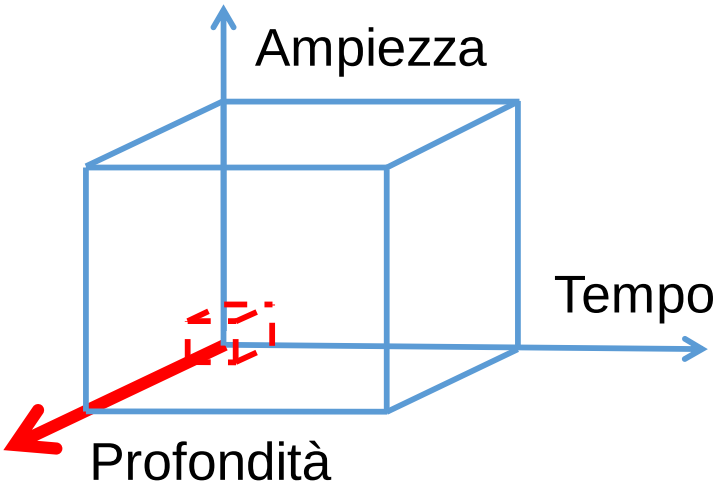
Ampiezza

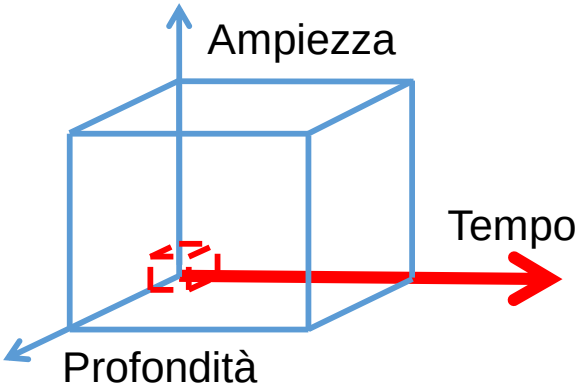
Dalla Hecateus Map (520 B.C.)…



… ai satelliti Landsat

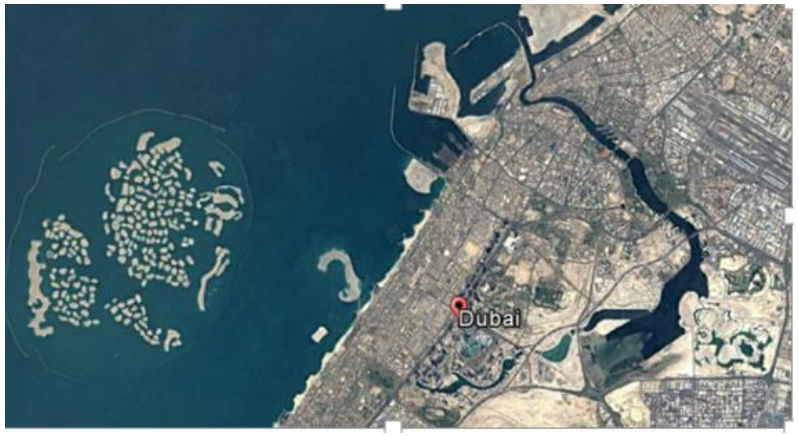
Profondità – I Pneumatici Intelligenti





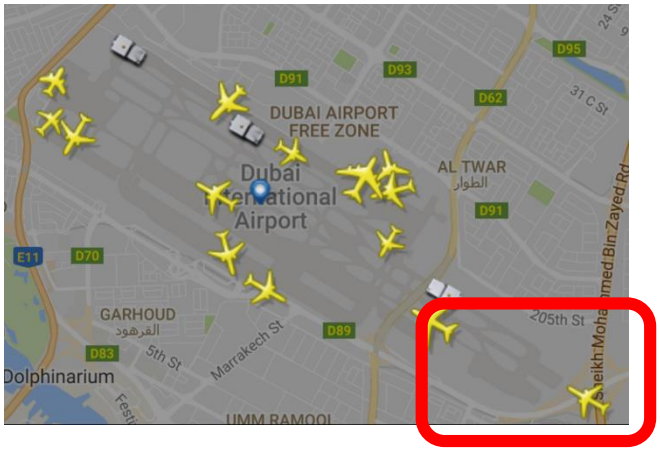
Evoluzione nel tempo

1 mese

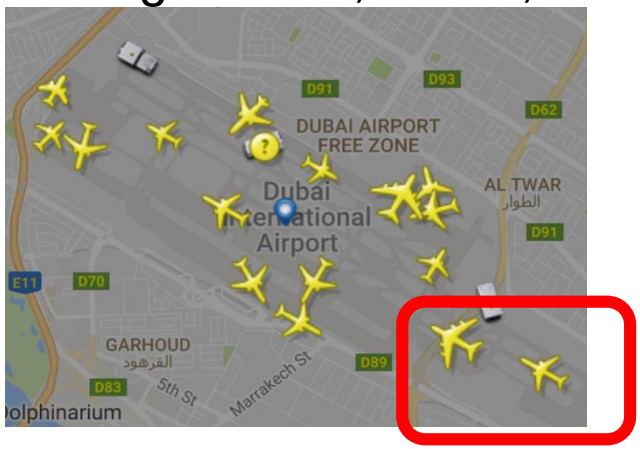


Google Earth, Dubai, **1984**

Google Earth, Dubai, **2015**

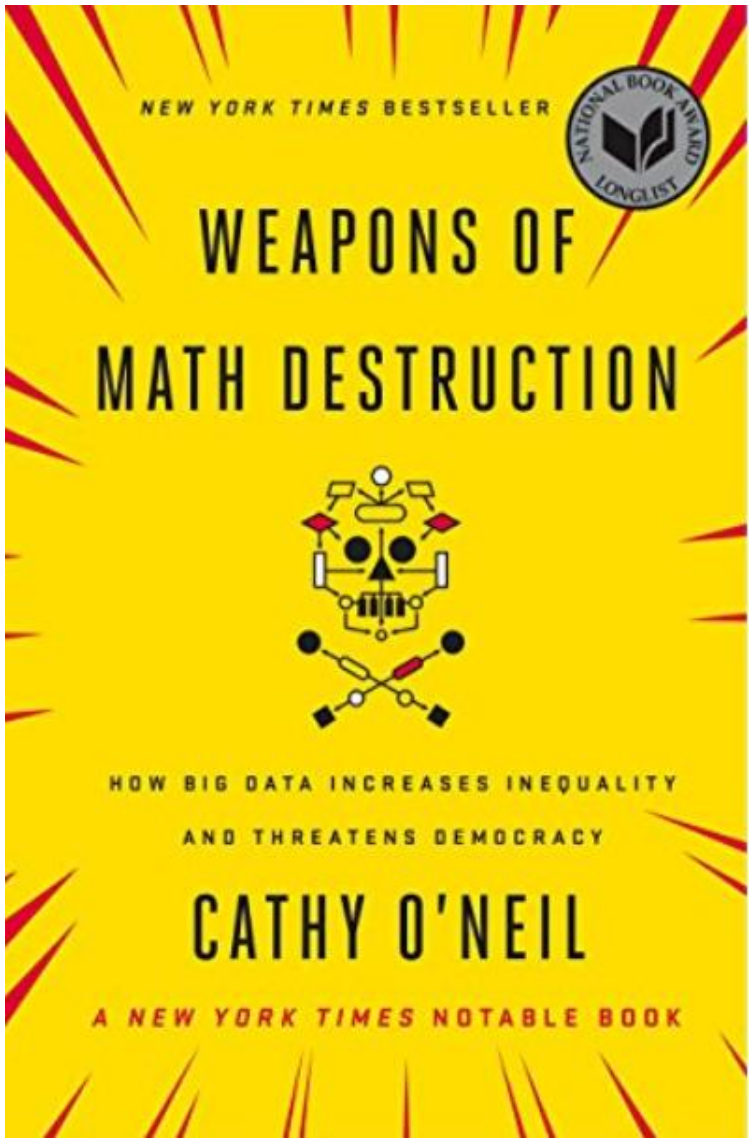


1 secondo

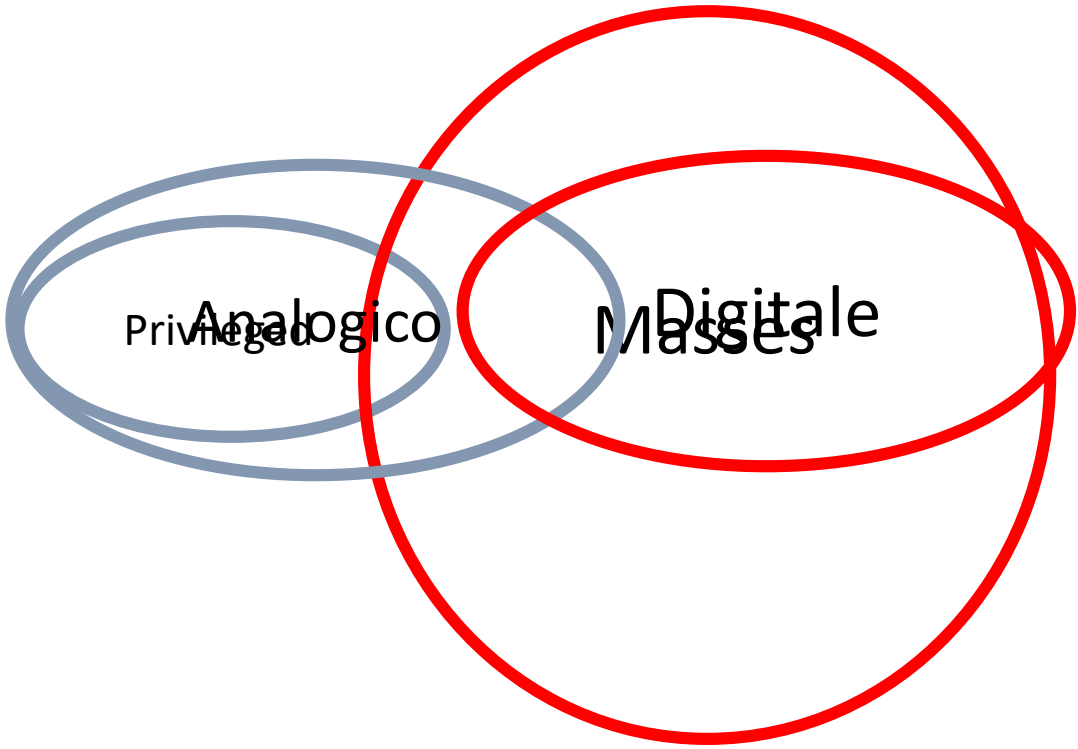


FlightRadar, Dubai 11:05:30 4:3:2017

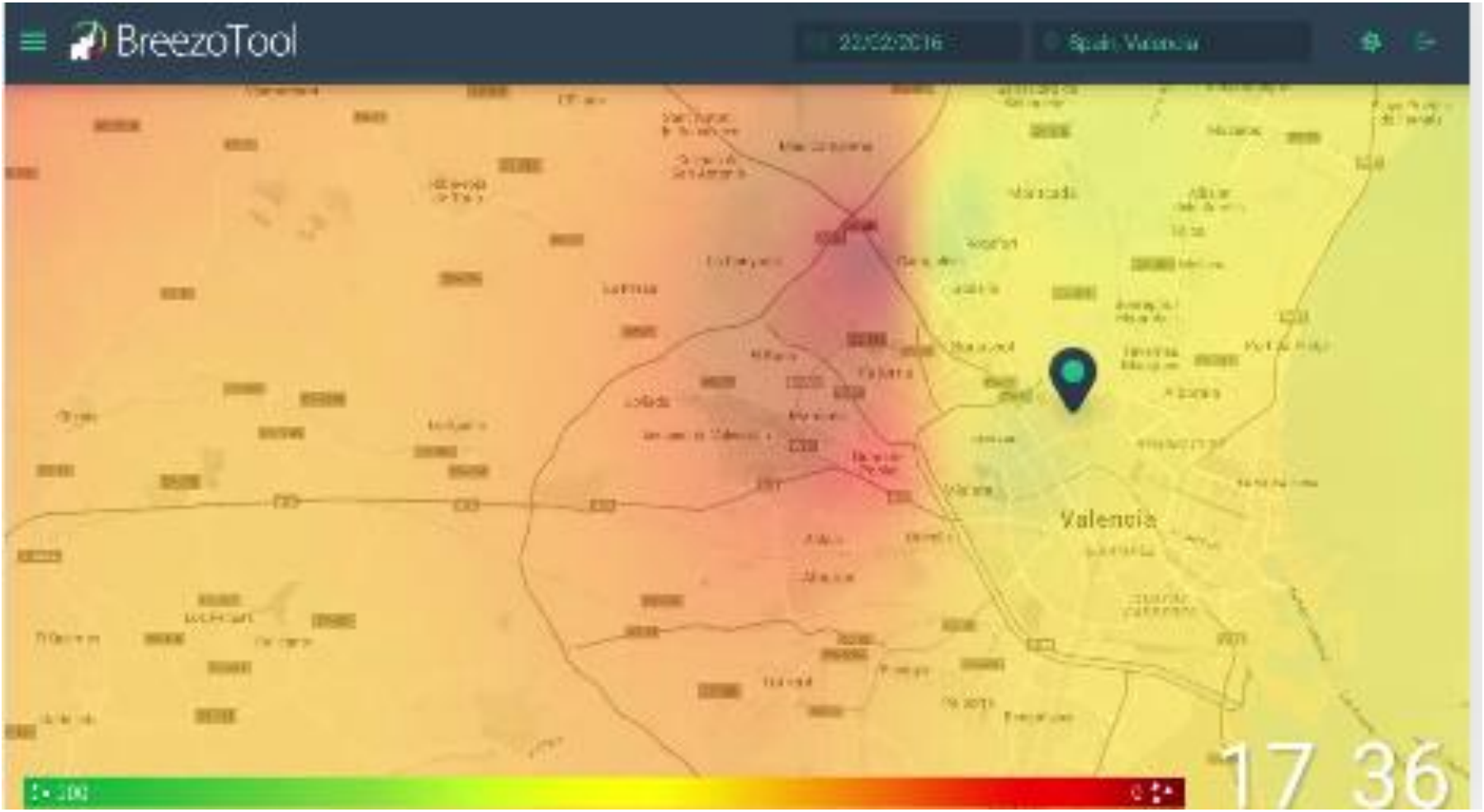
FlightRadar, Dubai 11:05:35 4:3:2017



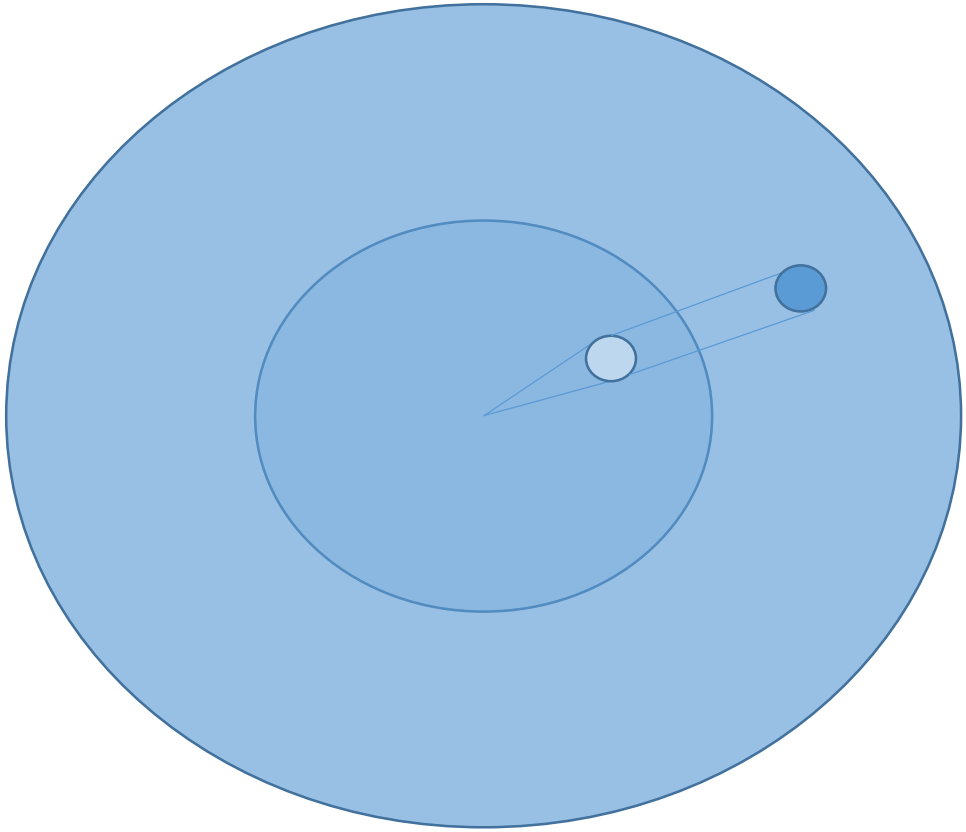
The privileged are processed more by people, the masses by machines



Ore 17:36



I limiti cognitivi



PB = 10^{15}, EB = 10^{18}, ZB = 10^{21} byte in quattro grandi domini dei Big Data nel 2025

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	[illegible]	[illegible] mining	[illegible]	[illegible]
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

Eclissi e prezzi dei biglietti

L'osservazione delle eclissi all'epoca dei Babilonesi portò a scoprire il **ciclo di Saros**, che dice secondo quale scansione temporale si succedono le eclissi del sole e della luna.

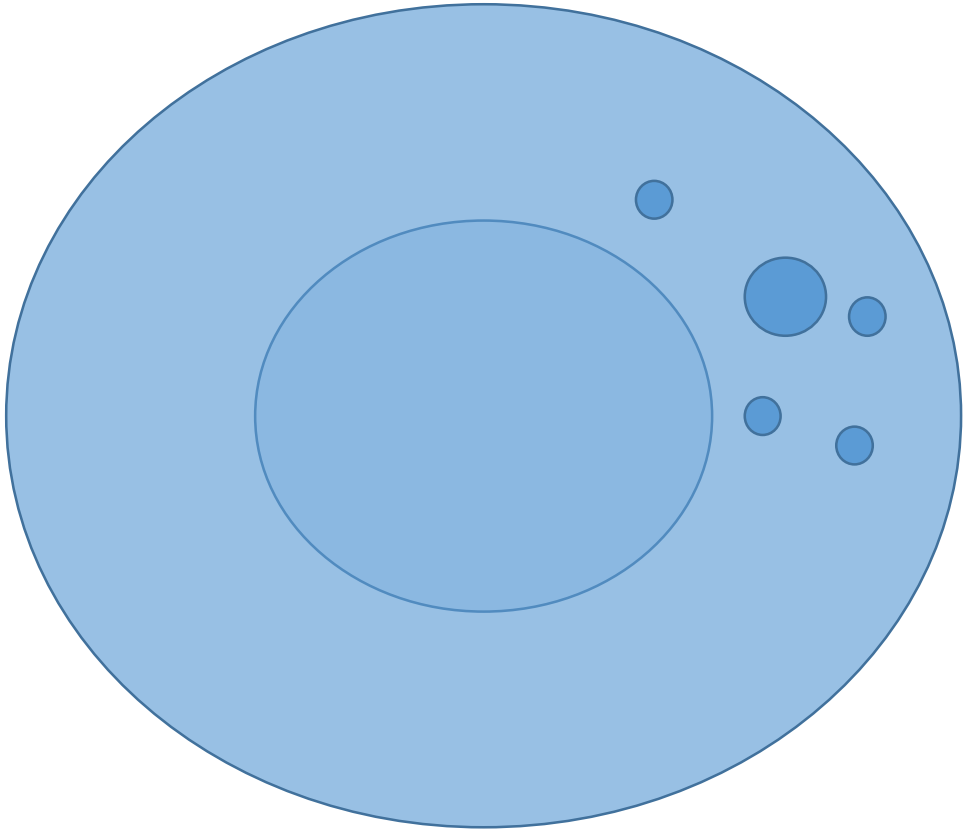
Il confronto tra le previsioni fatte dai babilonesi e quelle ottenute con le attuali tecnologie mostra una precisione stupefacente per l'epoca.

Le leggi dei prezzi dei biglietti (leggi dette di pricing) è problema più complesso, che richiede una conoscenza delle **politiche delle compagnie aeree** in genere considerate segreto aziendale.

Determinants of Ethics in Digital Data Science

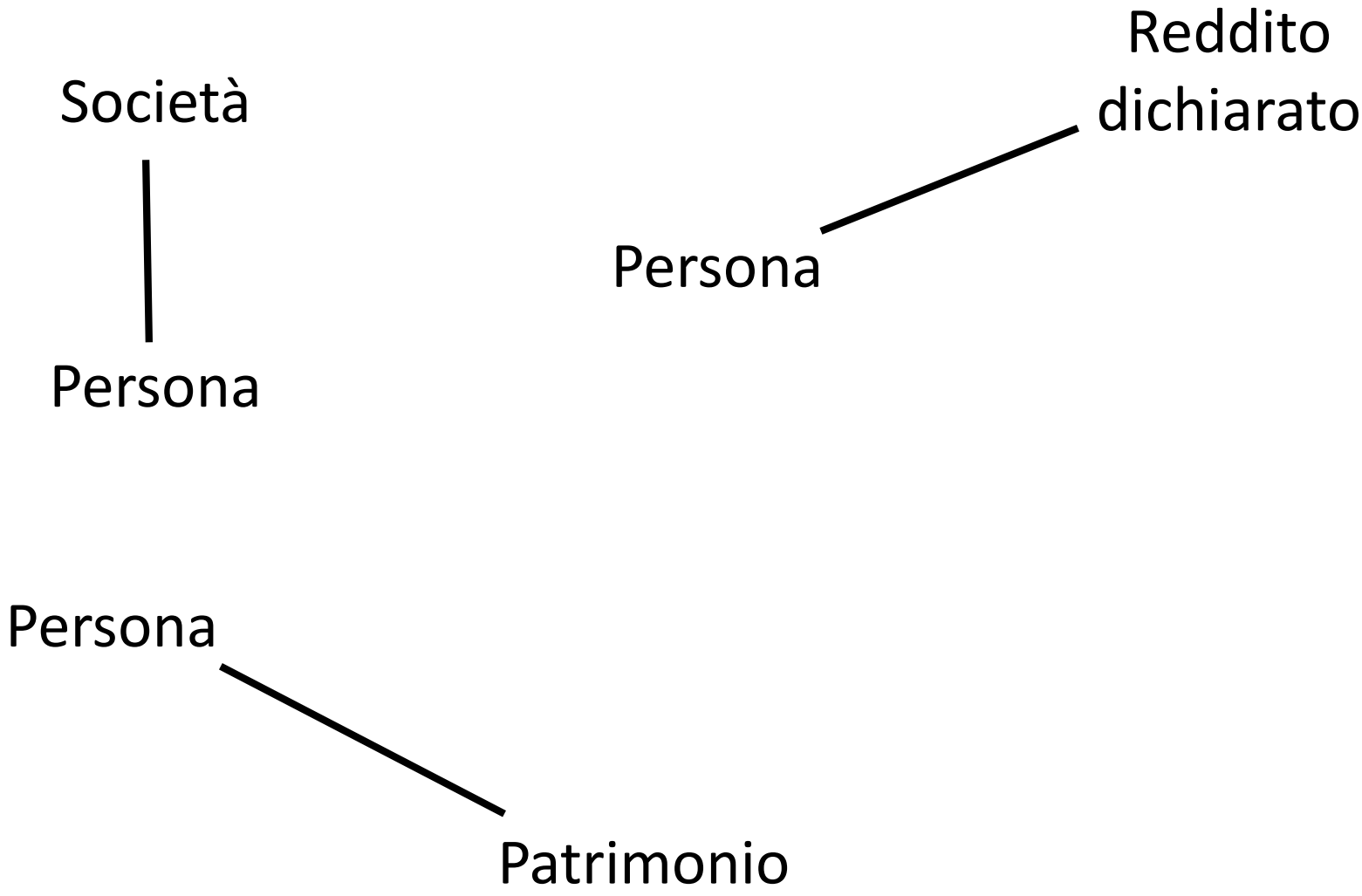
1. **Transparency** vs **Opacity**, it is easy for others to see from data what actions are performed.
2. **Accountability,** mechanisms are in place to determine who took a responsible action.
3. **Attribution of responsibility,** a process of cognition based on data by which moral account-tability is assigned to a person believed to have produced a disapproved behavior or effect
4. **Auditability,** ability of examining or evaluating (something) thoroughly
5. **Awareness,** knowledge and understanding that something is happening or exists
6. **Data Divide,** economic and social inequality with regard to use of or impact of data
7. **Equality of Opportunity** (Non Discrimination, Egalitarianism), data should enable people to compete on equal terms
8. **Explanation,** a statement made to clarify something and make it understandable
9. **Fairness,** the state, condition, or quality of being fair, or free from bias or injustice
10. **Fairness-aware (learning techniques),** classifiers to be independent of sensitive features.
11. **Generalization vs Personalization,** data for all vs tailoring data to accommodate specific individuals, sometimes tied to groups or segments of individuals
12. **Liability**, extends responsibility further to the area of laws.
13. **Objectivity,** state of being true even outside a subject's individual biases, interpretations and feelings
14. **Quality of information, (Accuracy, Veridicity etc.)** property of information to be adherent to reality
15. **Responsibility**, you accept the potential costs, and duties for the decision you made
16. **Privacy,** the state of personal data of being free from public attention
17. **Sharing,** ~~dividing~~ having data in common with others
18. **Statistical discrimination,** racial or gender inequality based on stereotypes

La parcellizzazione della conoscenza

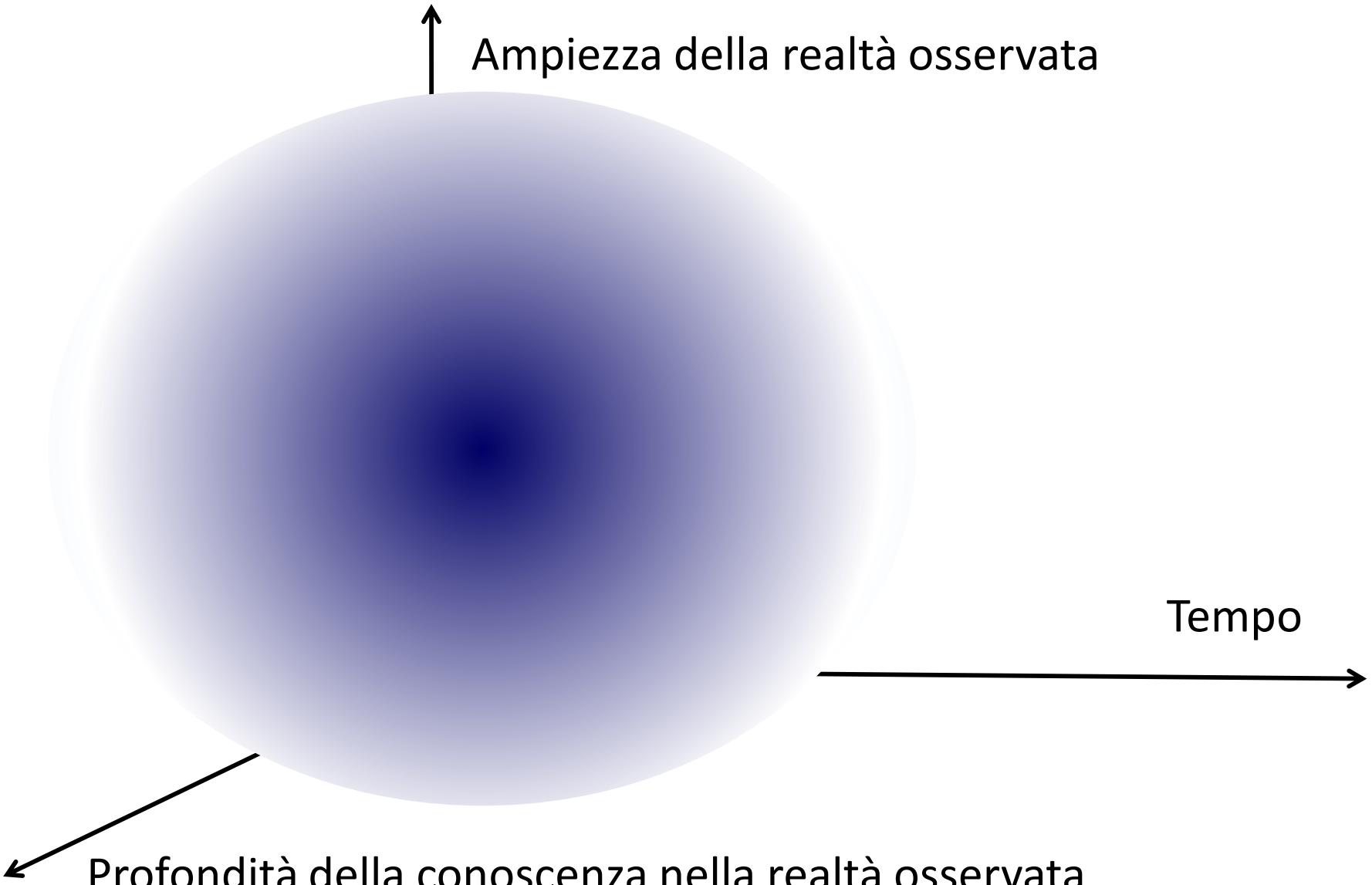


Valorizzazione

Integrare i dati porta valore

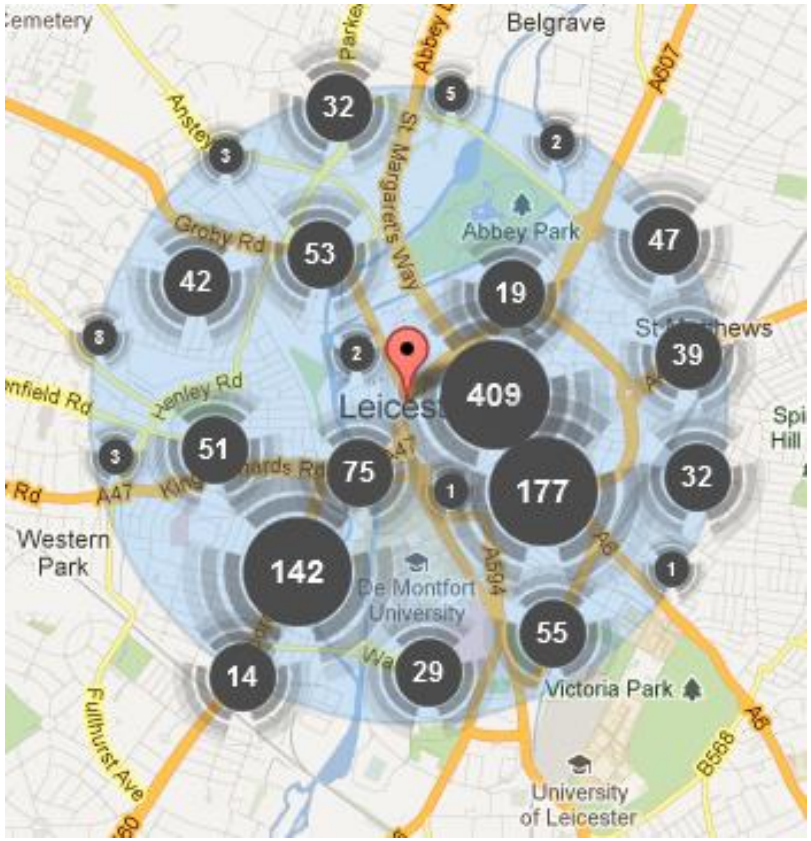


I grandi dati restituiscono una realtà opaca e confusa



Reati a Leicester: valore per una persona residente e valore per un affittuario

All crime	1241
Burglary	75
Anti-social behaviour	317
Robbery	12
Vehicle crime	57
Violent crime	181
Public disorder and weapons	45
Shoplifting	172
Criminal damage and arson	102
Other theft	178
Drugs	48
Other crime	54



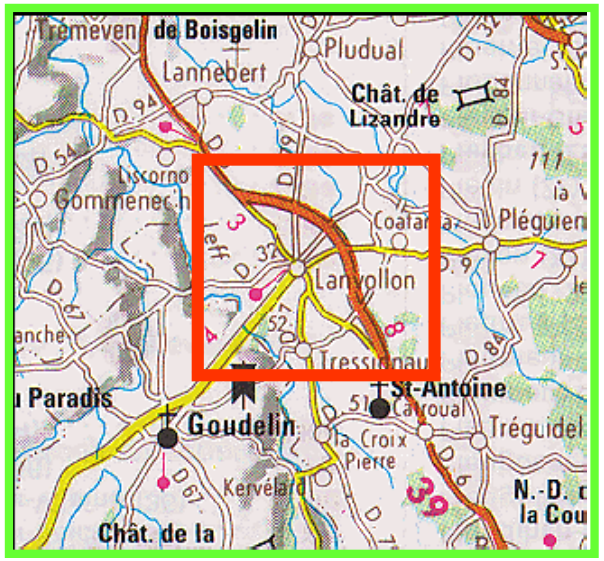
Dall' Economist, Nov. 10th, 2016
Le tecnologie costano molto meno delle persone…

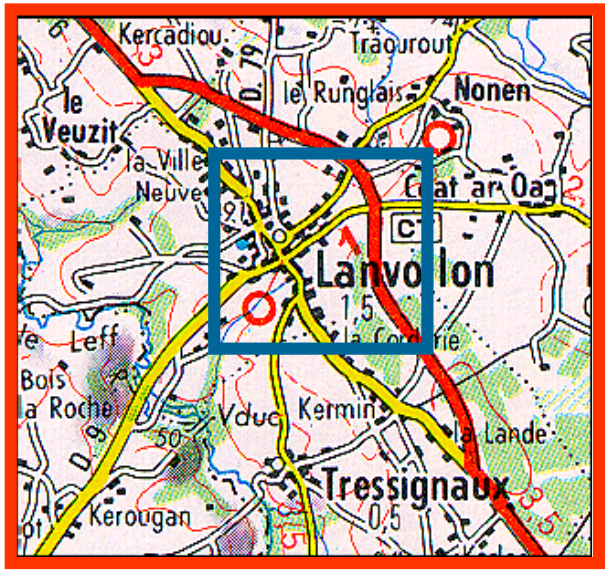
- The agencies not only do more, they also spend less.
- To **deploy agents on a tail** costs $ 175.000 a month because it takes a lot of manpower.
- To **put a GPS receiver in a someone's car** takes

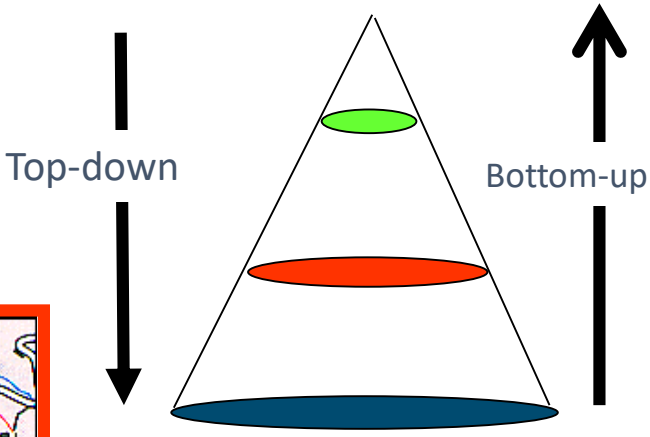
  $ 150 a month.
- To **tag a target's mobile phone**, with the help of a phone company, costs only $ 30 a month.

Questa mappa è troppo dettagliata…

Tre mappe a differenti livelli di scala







Camere dell'eco

Walter Quattrociocchi

- "Ci stiamo isolando. Per la prima volta, attraverso l'analisi di 920 agenzie di stampa e 376 milioni di utenti, abbiamo esplorato **l'anatomia del consumo di notizie** su Facebook su scala globale.
- Questi numeri ci hanno dimostrato che gli utenti tendono a focalizzare la loro attenzione su **un numero limitato di pagine**, andando a selezionare un gruppo ristretto di media da cui attingere informazioni e rafforzando così le proprie opinioni, **senza mai metterle in discussione**.
- Di fatto, **si chiudono nella loro camera dell'eco**".