I big Data – dietro le quinte di Google e Facebook

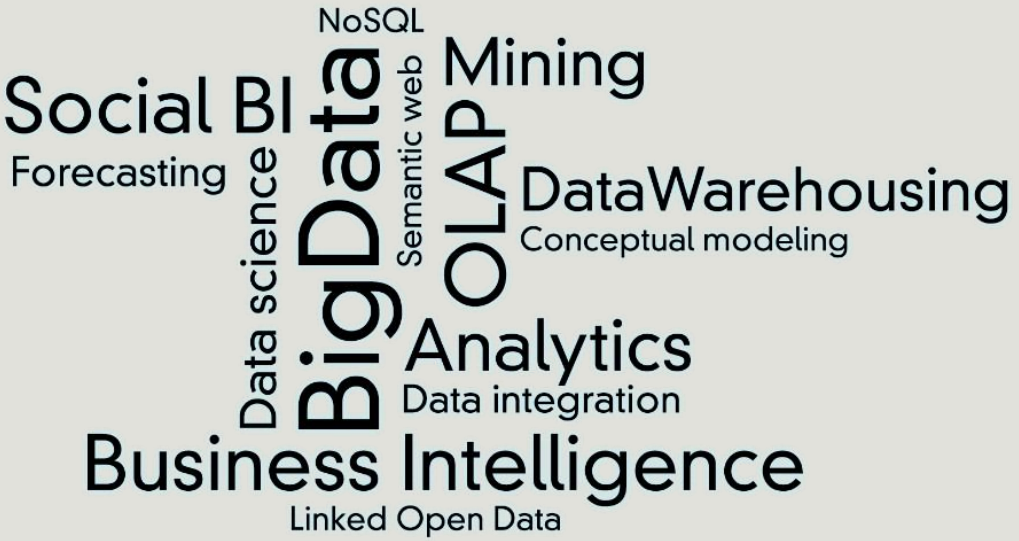
TECNOLOGIA, SCIENZA ED ETICA

Business Intelligence Group

Matteo Golfarelli

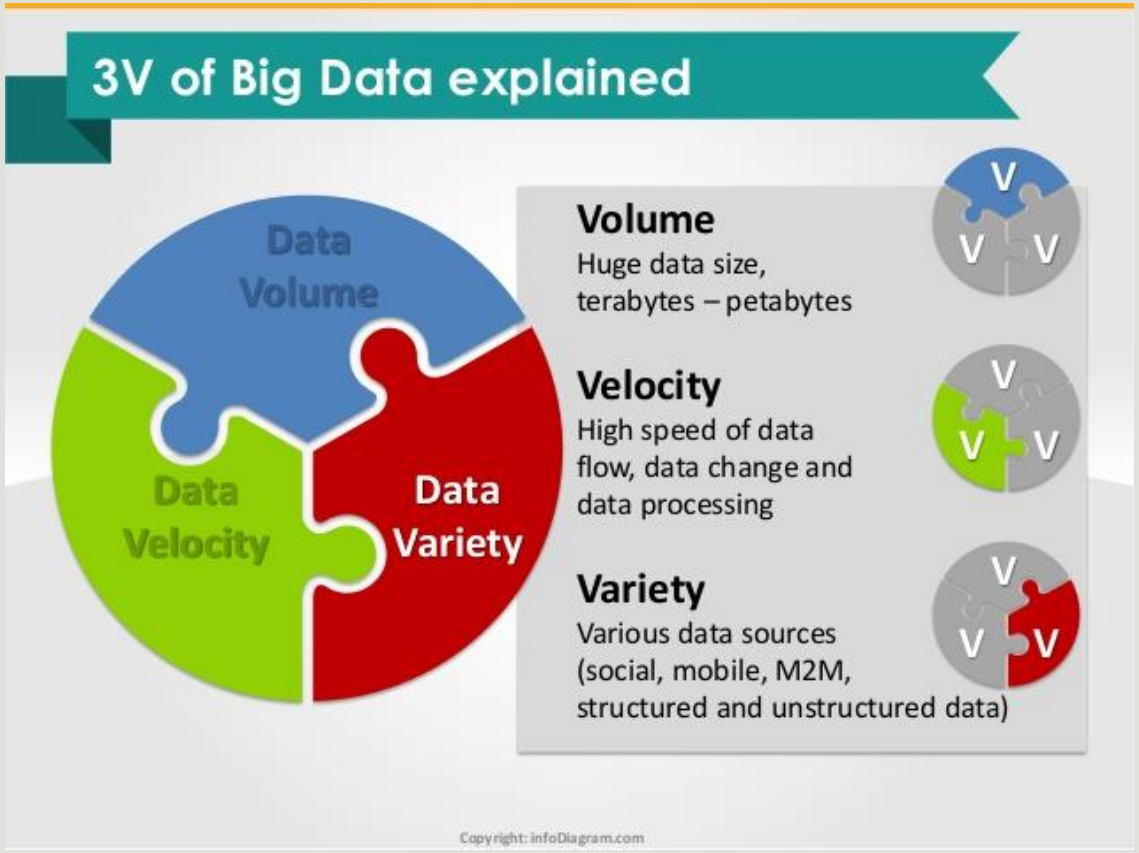
Il Business Intelligence Group

Il Business Intelligence Group dal 1997 svolge ricerche legate alle metodologie, tecniche e Tecnologie nell'ambito dell'analisi dati.



Big Data: una definizione

Big Data sono dataset con le seguenti caratteristiche, ***non necessariamente tutte!***



Big Data: volume

Volume: Terabyte o Petabyte tali da superare il limite di processamento dei sistemi tradizionali.

- Alcuni esempi:
  - Walmart: 1 milione di transazioni per ora (2010)
  - eBay: il data throughput ha raggiunto **100 PB** per giorno (2013)
  - Facebook: 40 miliardi di foto (2010); **250PB** data warehouse con **600TB** aggiunti ogni giorno (2013)
  - 500 milioni di tweet al giorno (in 2013)
  - You tube gestisce un traffico mensile di circa **27 PB**

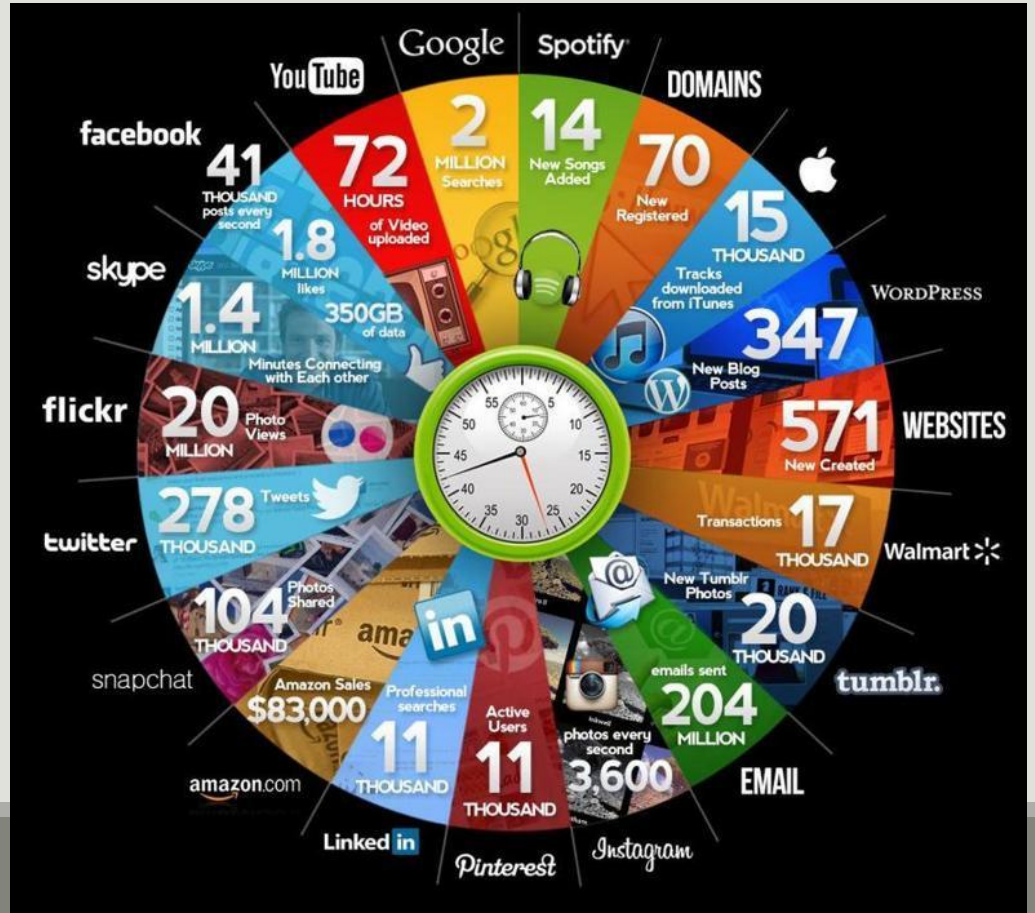
Il disco di un buon notebook o desktop PC contiene 1 TB

terabyte	TB	10^{12}	1 disco
petabyte	PB	10^{15}	1000 dischi
exabyte	EB	10^{18}	1 milione di dischi
zettabyte	ZB	10^{21}	1 miliardo di dischi

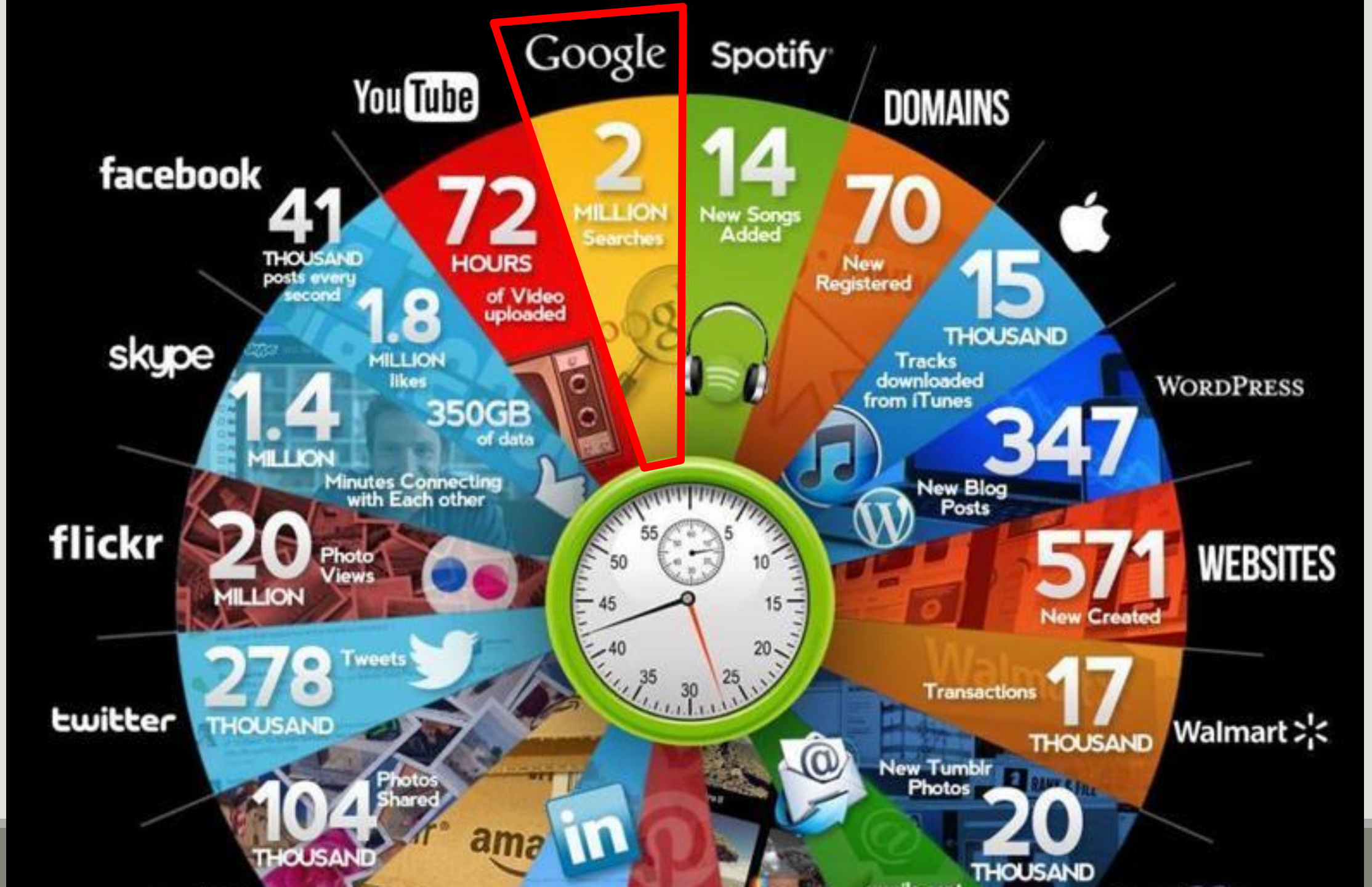
Big Data: velocità

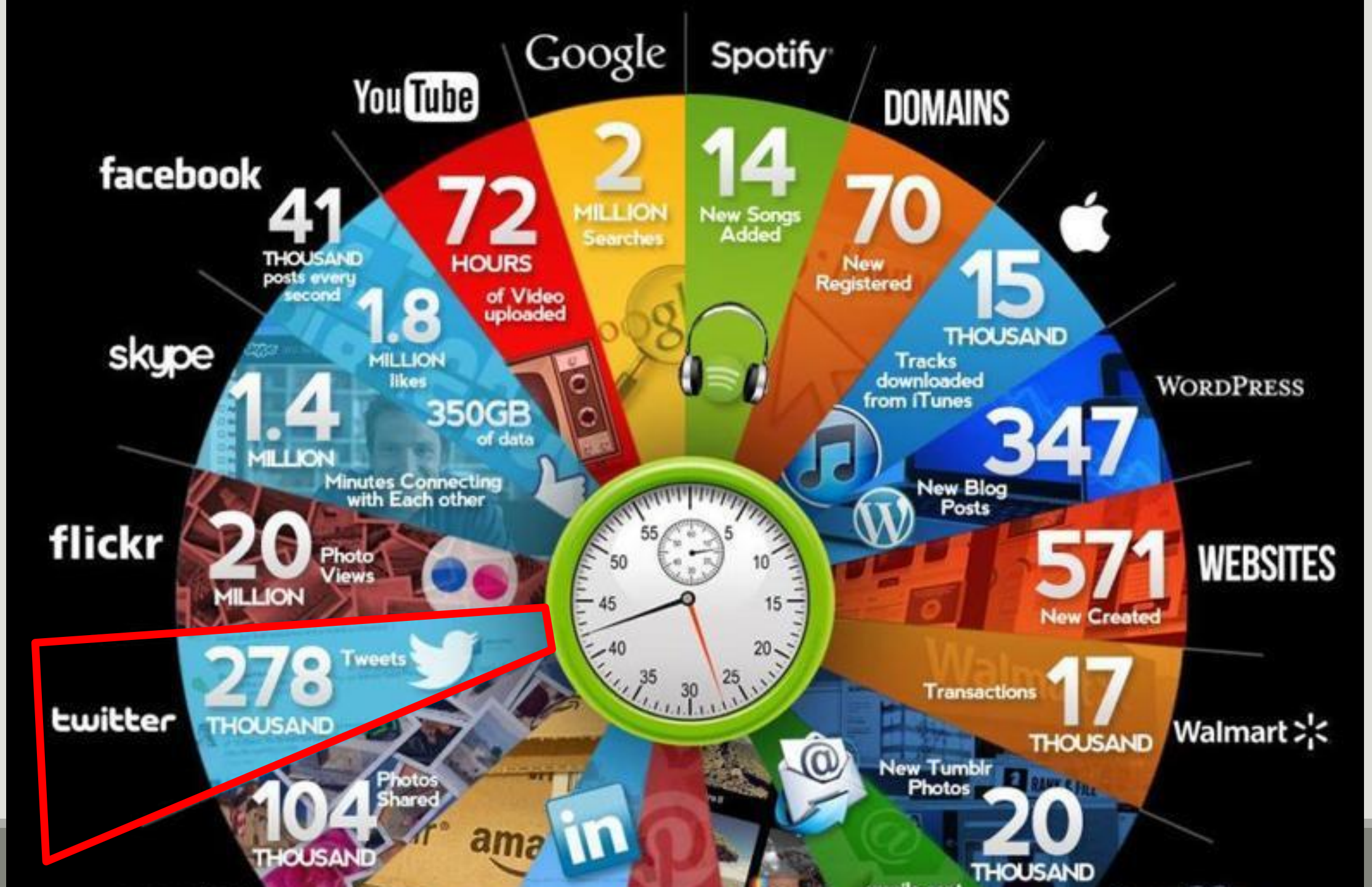
Velocità: device mobili e transazioni IoT producono dati con una frequenza superiori a quella dei sistemi informativi tradizionali.

Cosa accade ogni 60 secondi in rete?

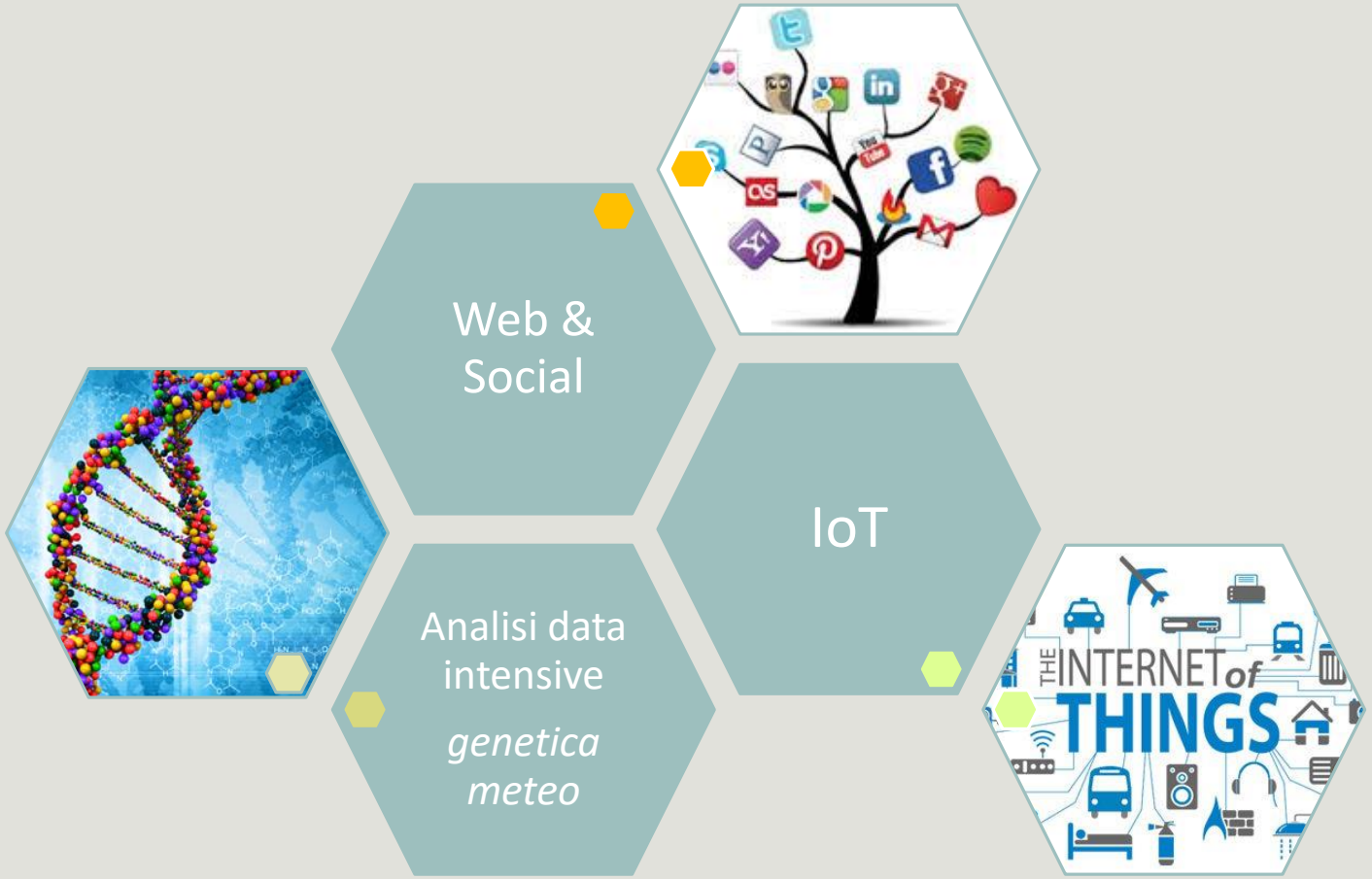



facebook
41 THOUSAND posts every second
1.8 MILLION likes
1.4 MILLION Minutes Connecting with Each other
YouTube
72 HOURS of Video uploaded
Google
2 MILLION Searches
Spotify
14 New Songs Added
DOMAINS
70 New Registered
15 THOUSAND Tracks downloaded from iTunes
WordPress
347 New Blog Posts
WEBSITES
571 New Created
Walmart
Transactions 17 THOUSAND
New Tumblr Photos 20 THOUSAND
Photos Shared 104 THOUSAND
twitter
278 THOUSAND Tweets
flickr
20 MILLION Photo Views
skype
350GB of data


Google
2
MILLION
Searches
YouTube
72
HOURS
of Video
uploaded
Spotify
14
New Songs
Added
DOMAINS
70
New
Registered
facebook
41
THOUSAND
posts every
second
1.8
MILLION
likes
skype
1.4
MILLION
350GB
of data
Minutes Connecting
with Each other
15
THOUSAND
Tracks
downloaded
from iTunes
WordPress
347
New Blog
Posts
flickr
20
MILLION
Photo
Views
571
New Created
WEBSITES
278
THOUSAND
Tweets
twitter
Transactions
17
THOUSAND
Walmart
104
THOUSAND
Photos
Shared
New Tumblr
Photos
20
THOUSAND


Google
Spotify
YouTube
DOMAINS
facebook
41
THOUSAND
posts every second
72
HOURS
of Video uploaded
2
MILLION
Searches
14
New Songs Added
70
New Registered
15
THOUSAND
Tracks downloaded from iTunes
skype
1.8
MILLION
likes
1.4
MILLION
Minutes Connecting with Each other
350GB
of data
WordPress
347
New Blog Posts
flickr
20
MILLION
Photo Views
571
New Created
WEBSITES
twitter
278
THOUSAND
Tweets
Transactions
17
THOUSAND
Walmart
104
THOUSAND
Photos Shared
New Tumblr Photos
20
THOUSAND

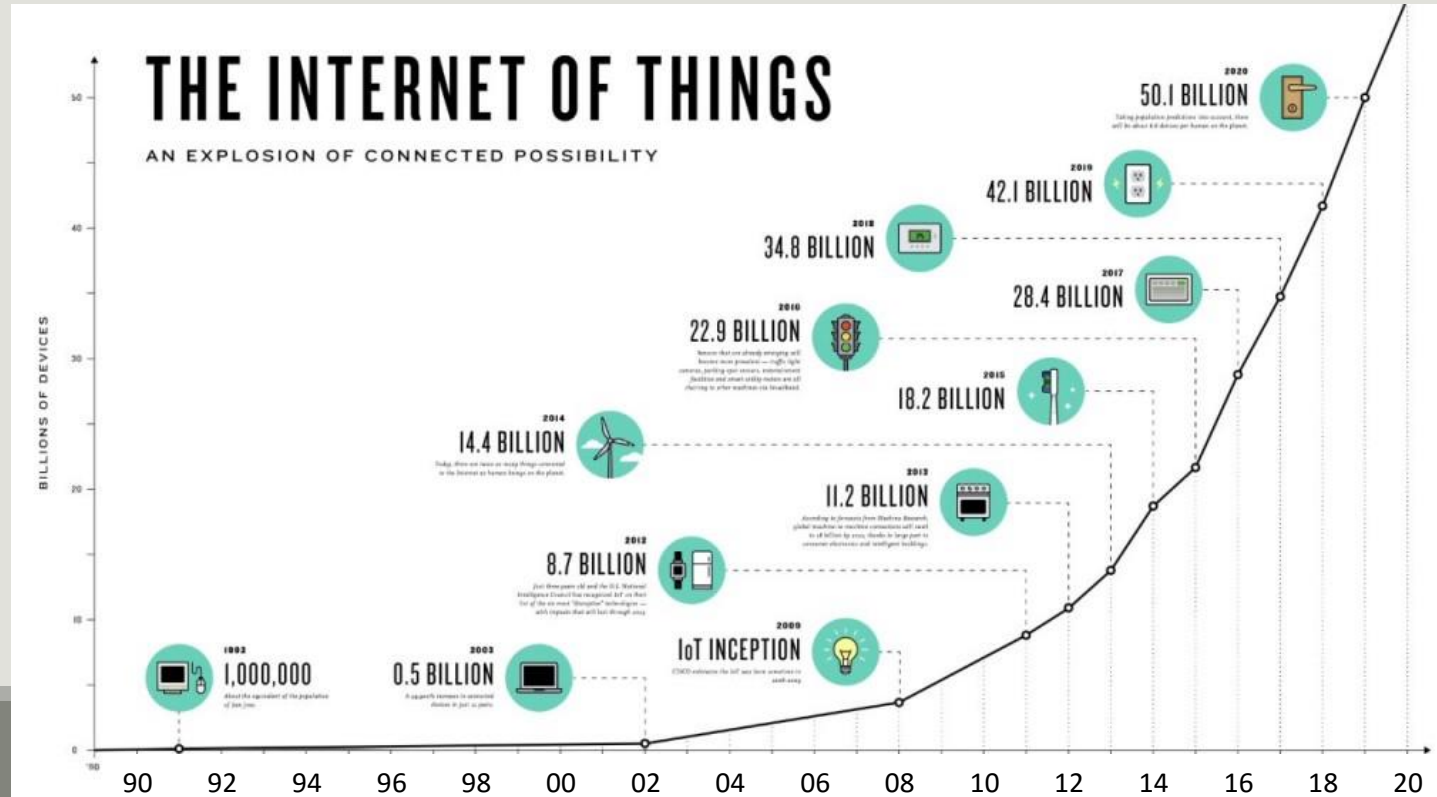
Chi genera i big data?



Chi genera i big data?

IoT – Internet of Things è l'evoluzione dell'uso di Internet in cui 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 Volume

- La sveglia suona prima in caso di traffico
- Il surgelato nel freezer del supermercato segnala che sta per scadere
- L'impianto industriale segnala un'alta probabilità di guasto entro le prossime ore

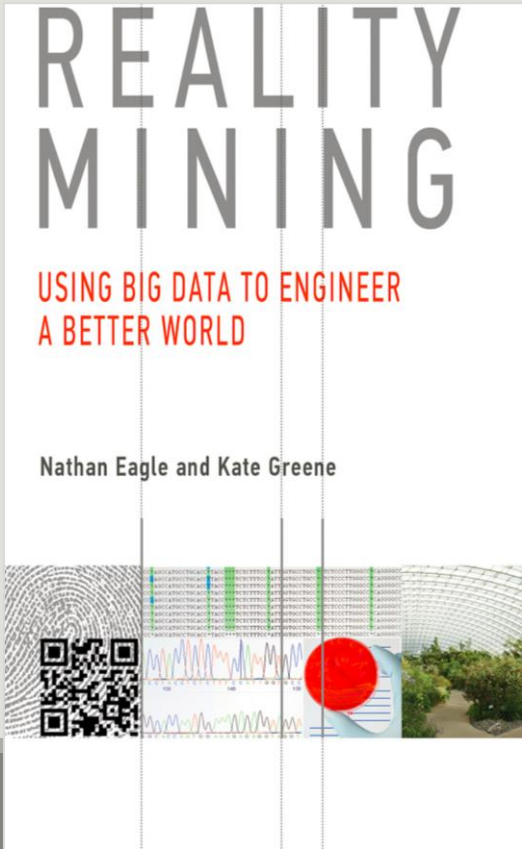


Reality Mining

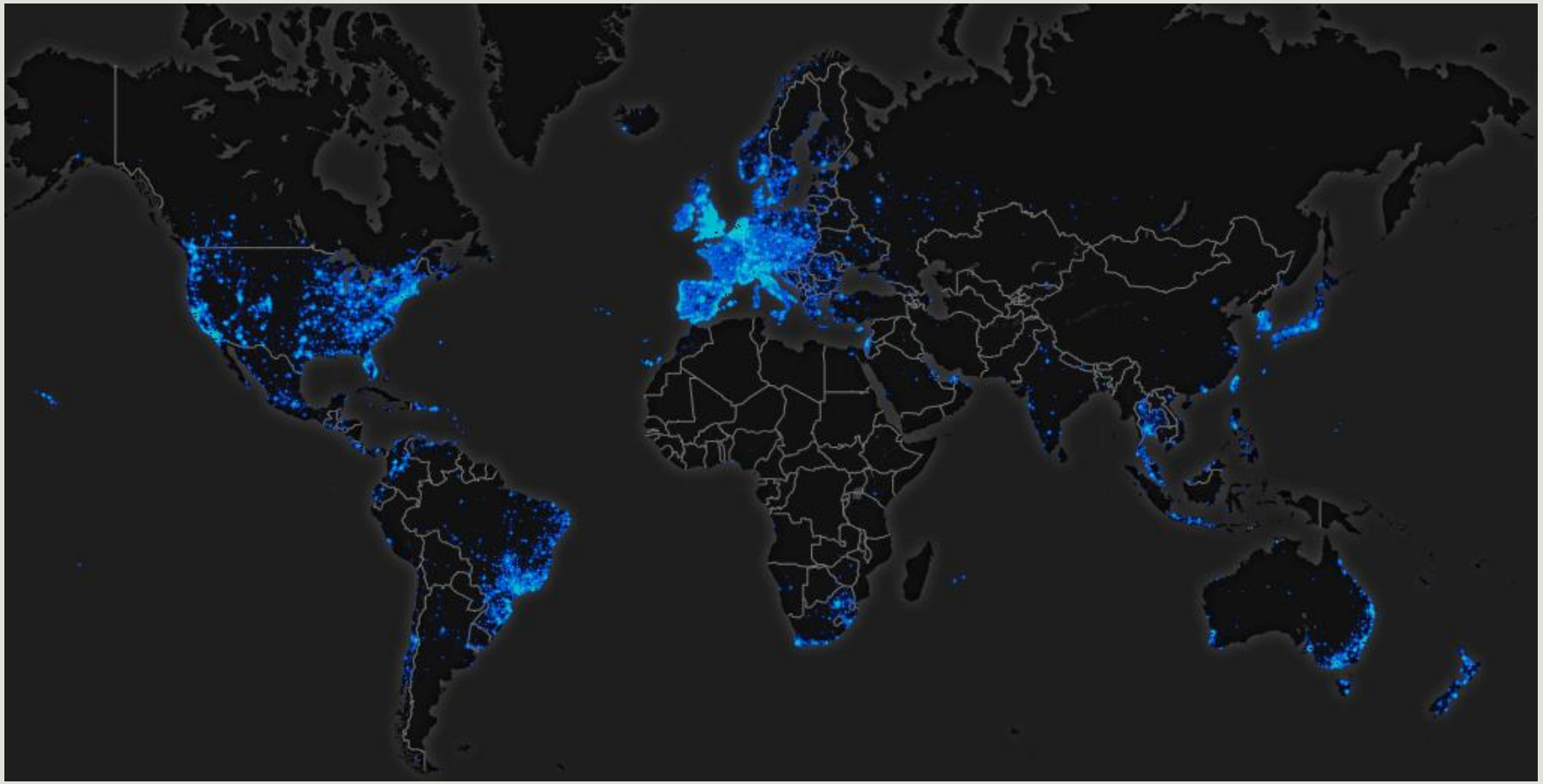
L’ampia disponibilità di sensori wearable (smartphones, activity trackers) rende possibile il monitoraggio di sistemi sociali complessi

Riconoscere pattern di comportamento quotidiano

- Inferire Relazioni tra le persone
- Identificare luoghi socialmente rilevanti
- Identificare i ritmi di comportamento della società



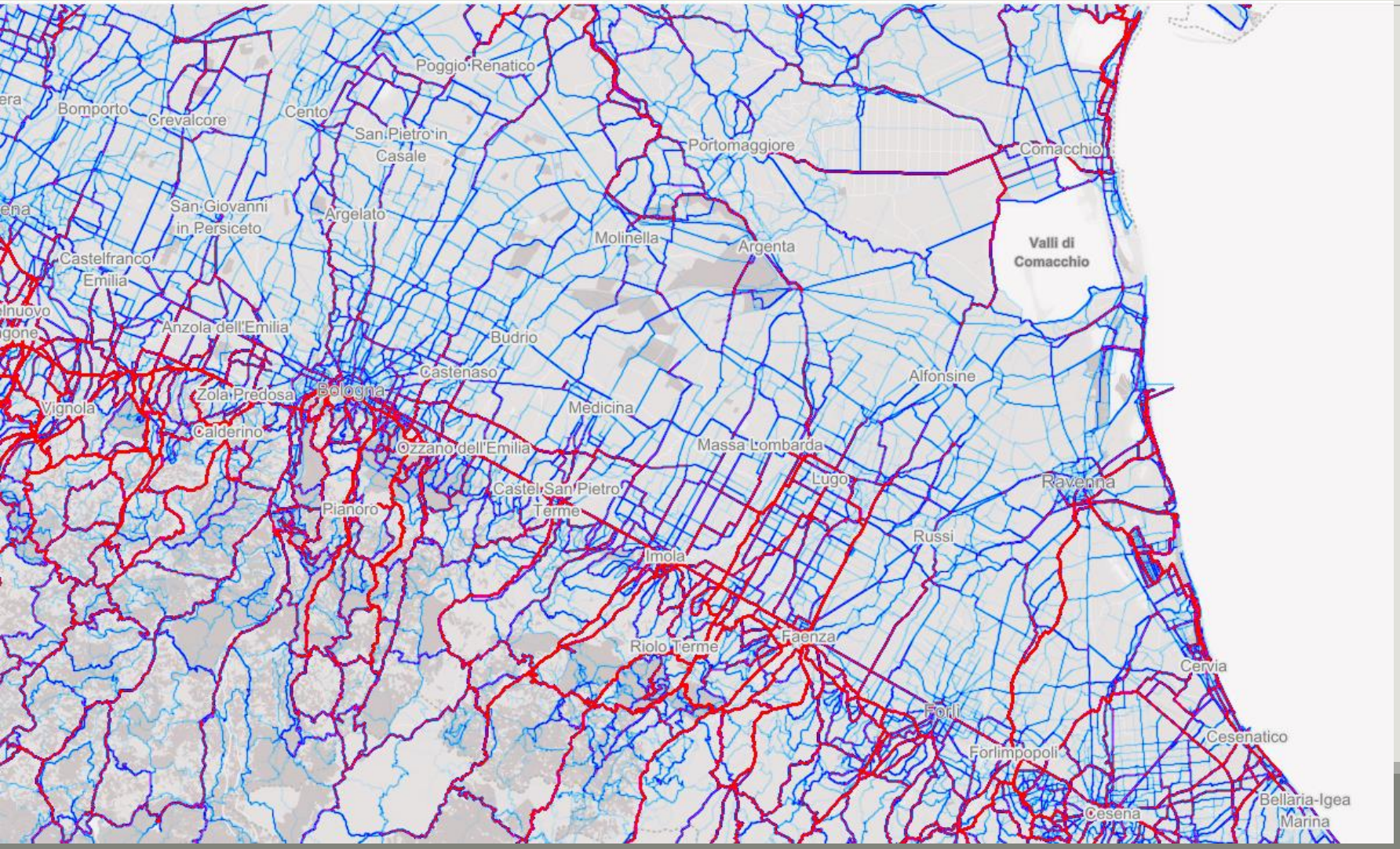
Reality Mining: Quantified Self



Community Tracing

http://labs.strava.com/heatmap/

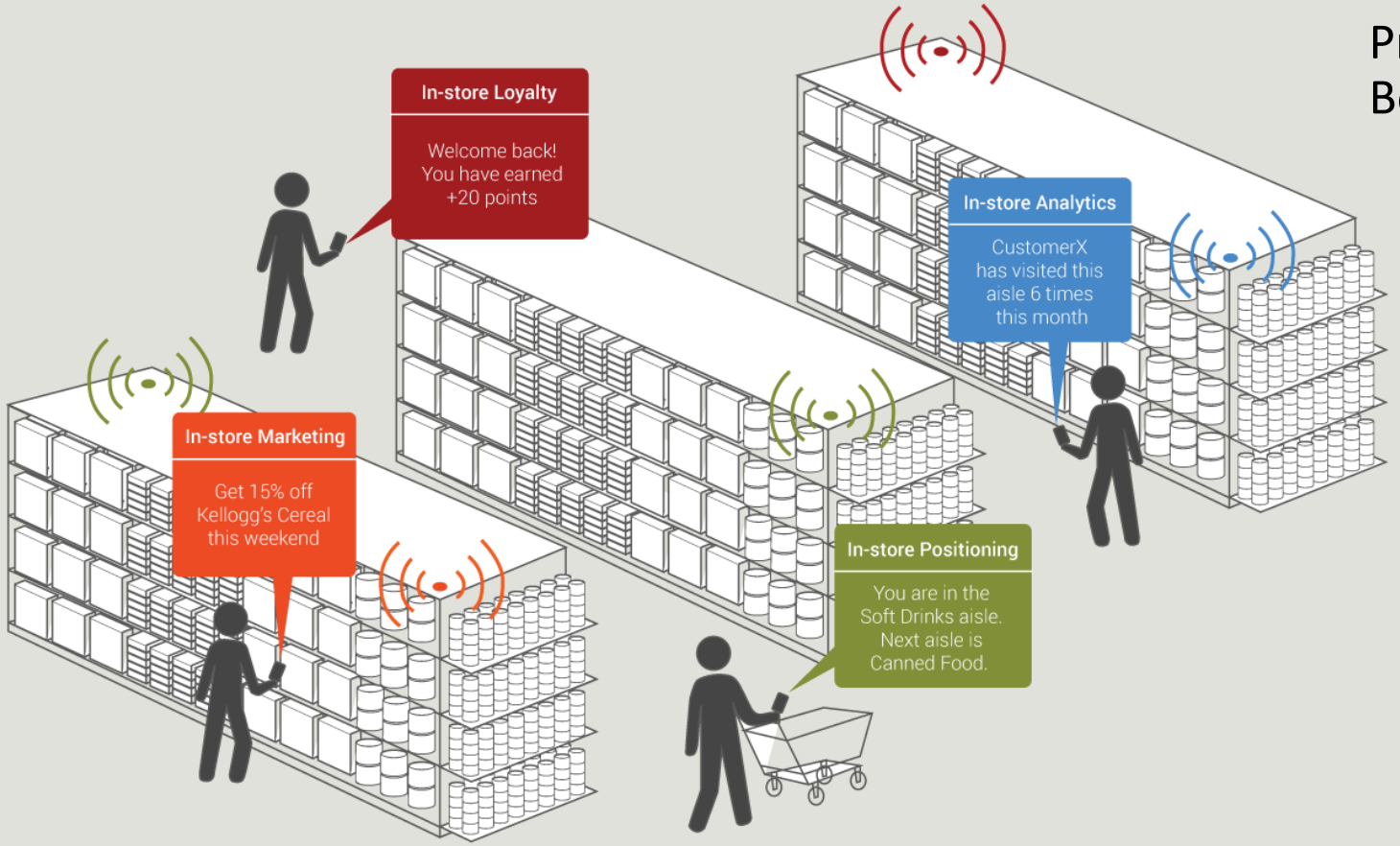
Reality Mining: Quantified Self



Community Tracing

http://labs.strava.com/heatmap/

Reality Mining: Proximity Marketing



Proximity marketing based on WIFI or Beacons

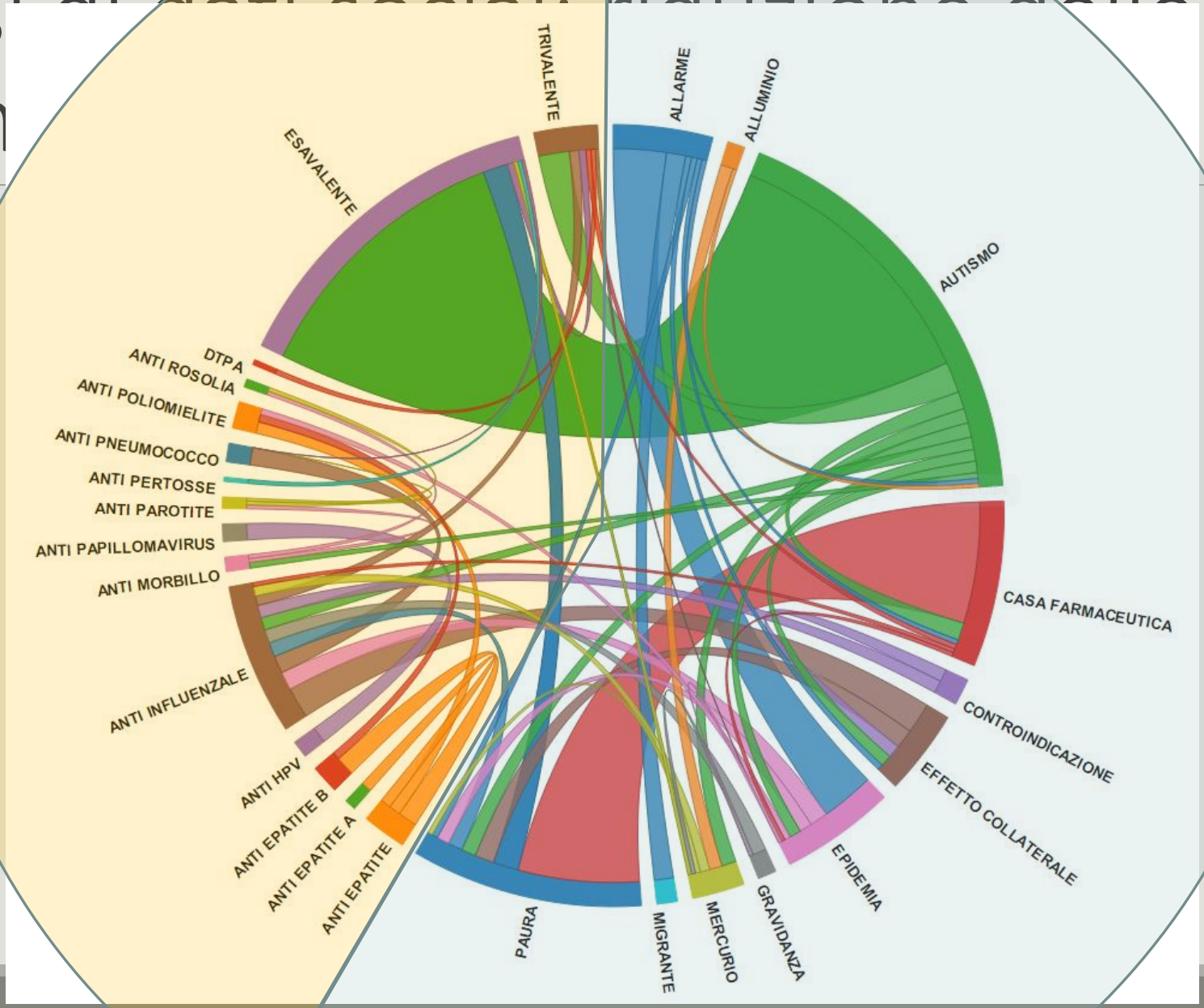
Analisi di dati social: riduzione delle vaccinazioni (regione Veneto)

Il ruolo social del testimonial e la forza dell'immagine hanno reso vincente questa campagna

Analisi di dati social riduzione delle vaccin

Vaccini

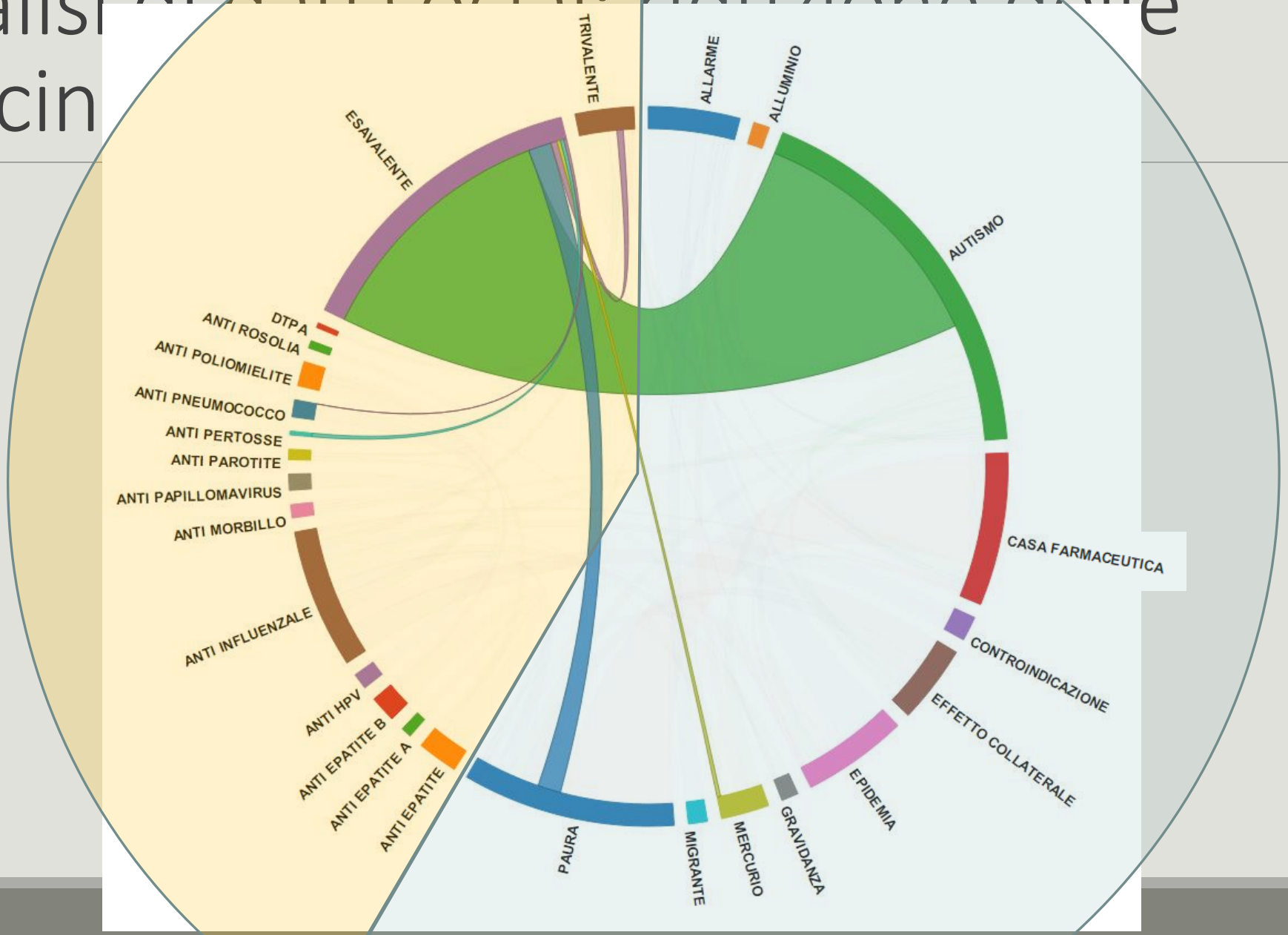
Paure



Analisi di dati social: riduzione delle vaccin

Vaccini

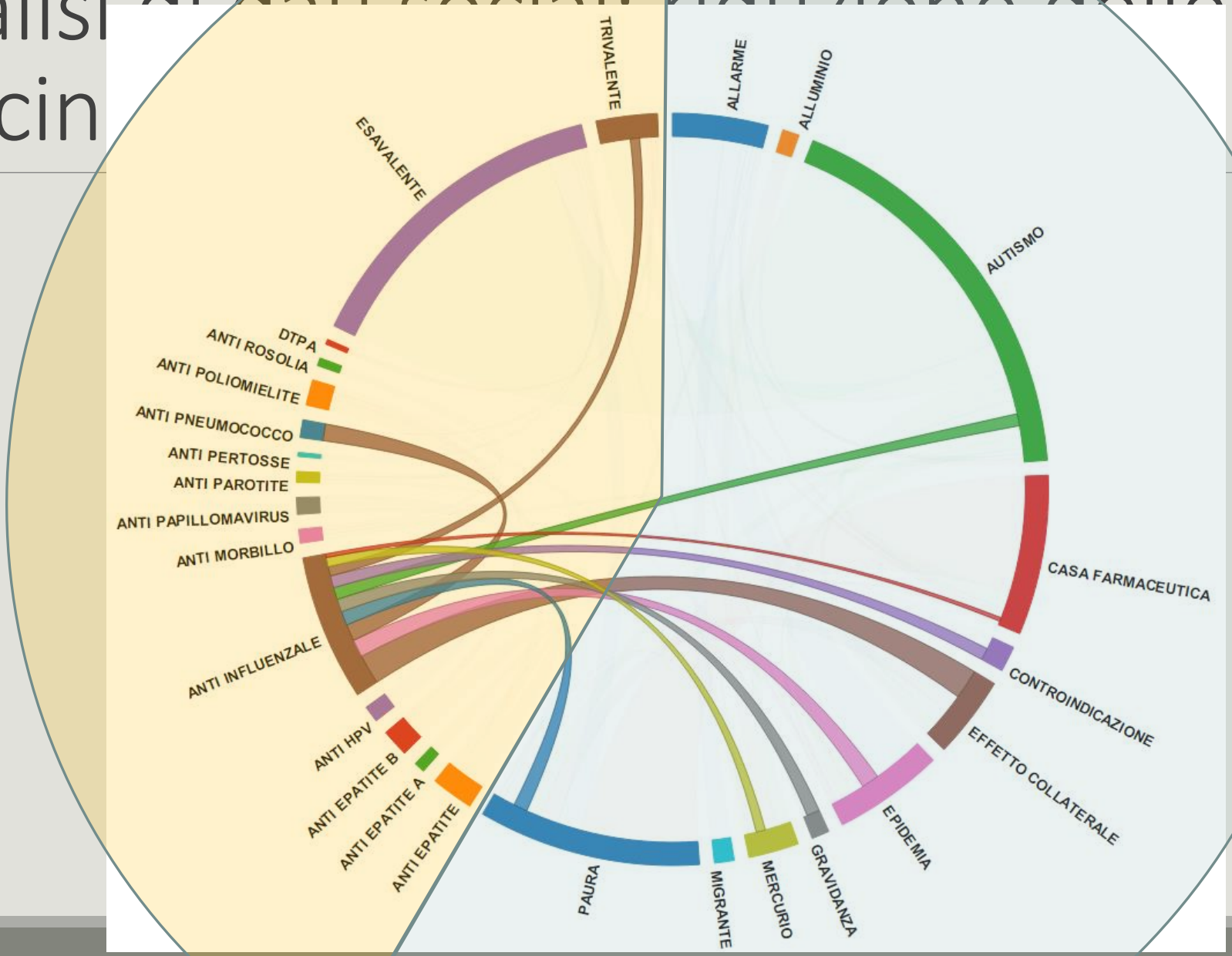
Paure



Analisi di dati social: riduzione delle vaccin

Vaccini

Paure



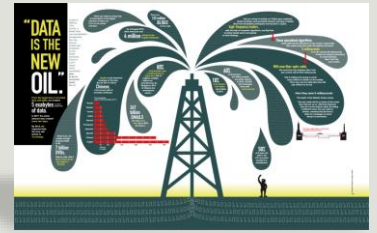
Hawk: Harnessing Wellness Knowledge

L’impatto sulla nostra vita: potenzialità e rischi

La disponibilità di grandi moli di dati relative alla nostra vita quotidiana e delle tecniche necessarie ad elaborarli ha ripercussioni sulla vita della comunità e del singolo sia in ambito private sia in ambito lavorativo

I big data sono una tecnologia, di per sé né buona né cattiva, ma dobbiamo essere consapevoli di come sono utilizzati e del potenziale impatto che hanno sul nostro modo di vivere

- **Quali nuovi servizi e funzionalità**
- **Quali nuove professioni e quali impatti su quelle esistenti**
- **Quali nuovi business**



- **Quali interferenze con le nostre scelte**
- **Quali violazioni della nostra privacy**

In questa rivoluzione possiamo essere soggetti *passivi*, *informati* o *attivi*

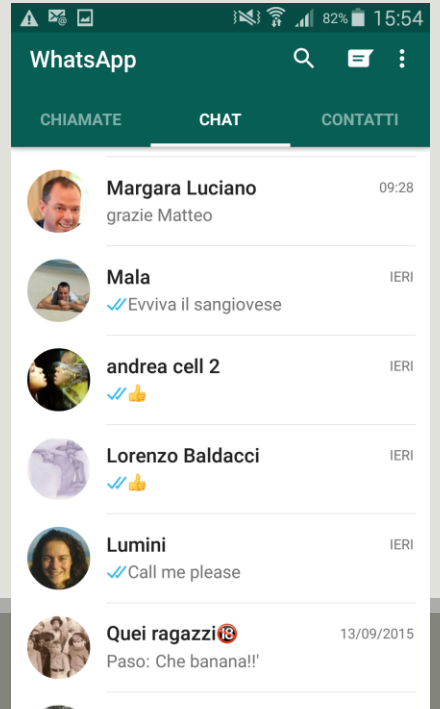
Opportunità o Abuso?

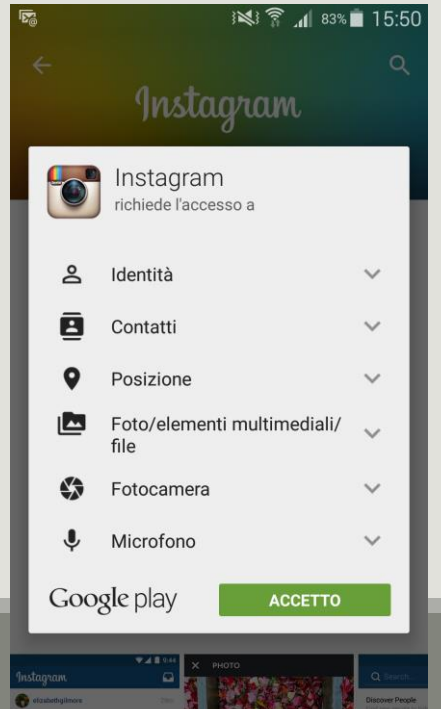
… la tecnologia garantisce enormi opportunità…

- Nuovi servizi

… ma non ci mette al riparo dai rischi!

- Violazione della privacy

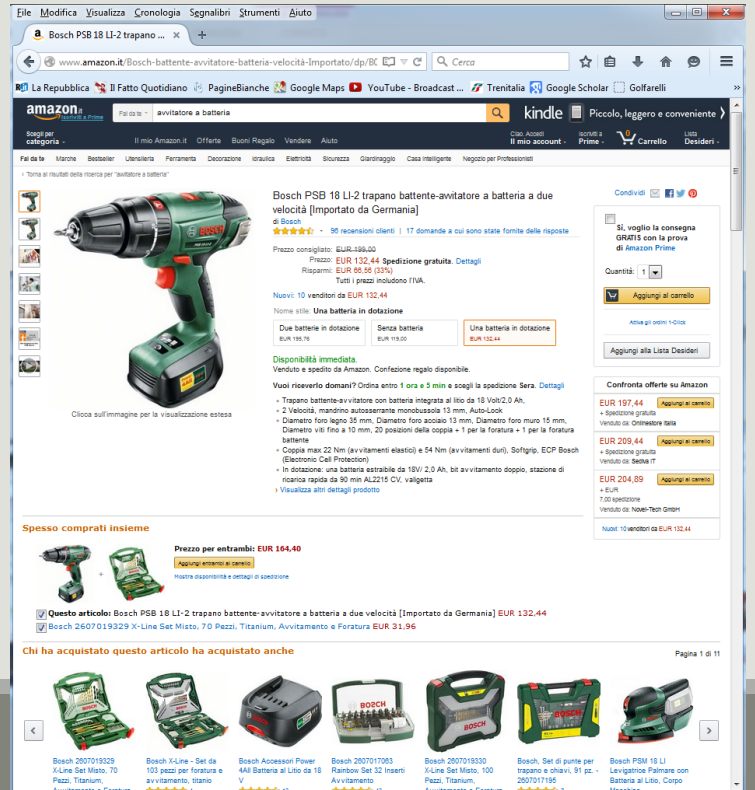




Opportunità o Abuso?

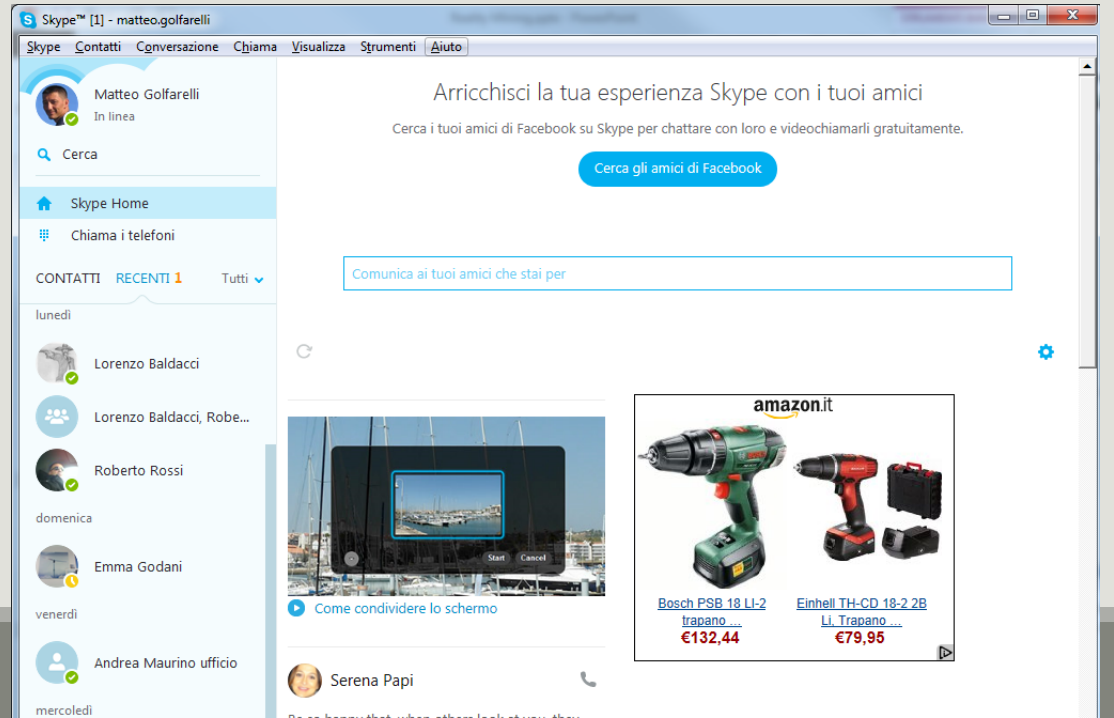
… la tecnologia garantisce enormi opportunità…

- Nuovi servizi
- **Servizi personalizzati**



… ma non ci mette al riparo dai rischi!

- **Violazione della privacy**
- **Marketing invasivo**



Opportunità o Abuso?

… la tecnologia garantisce enormi opportunità…

- Nuovi servizi
- Servizi personalizzati
- Una maggiore attenzione al cittadino
- Maggiore sicurezza

… ma non ci mette al riparo dai rischi!

- Marketing invasivo
- Violazione della privacy
- **Manipolazione delle opinioni**

Opportunità o Abuso?

... la tecnologia garantisce enc n ci mette al riparo dai rischi!

- Nuovi servizi rketing invasivo
- Servizi personalizzati lazione della privacy
- Una maggiore attenzion nipolazione dell'opinione pubblica
- Maggiore sicurezza

Il dietro le quinte dei Big data

La memorizzazione e l'analisi dei Big Data richiede una grande potenza di calcolo

- Il cluster di Yahoo! conta 100,000 CPU montate su 40,000 server

Un server ha in media un fault ogni 3 anni

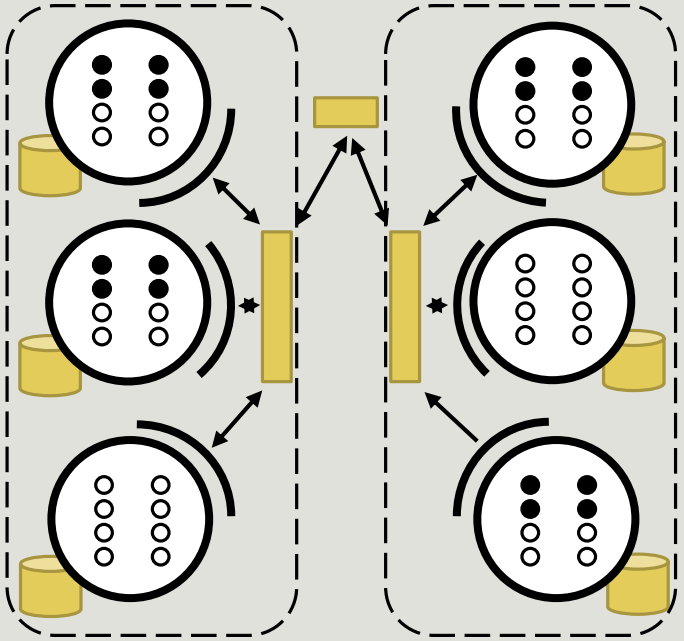
P(rottura oggi)=1/1095=0,00091

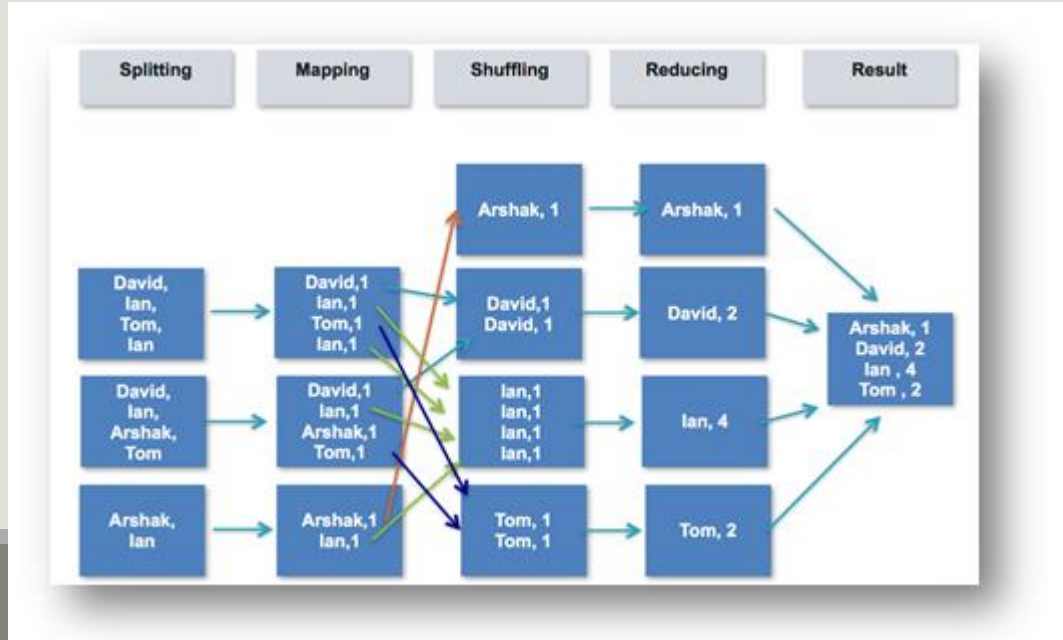
Ogni giorno si rompono 36 server

Apache Hadoop

E’ il principale framework di calcolo in ambito Big Data

- Nasce nel 2005 da progetti originatisi in Google tra il 2002 e il 2004
- E’ **robusto** rispetto a guasti HW e fallimenti di processi
- E’ basato su HW standard
- Permette l’elaborazione di enormi moli di dati per mezzo del **calcolo parallelo** su **cluster** di calcolatori
- Adotta un meccanismo di programmazione parallela semplice e nativamente parallelizzabile denominato **Map-Reduce**

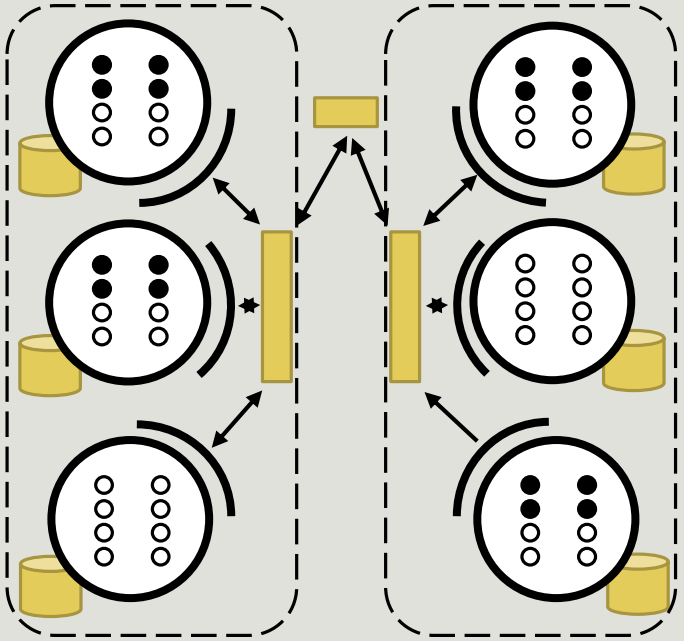


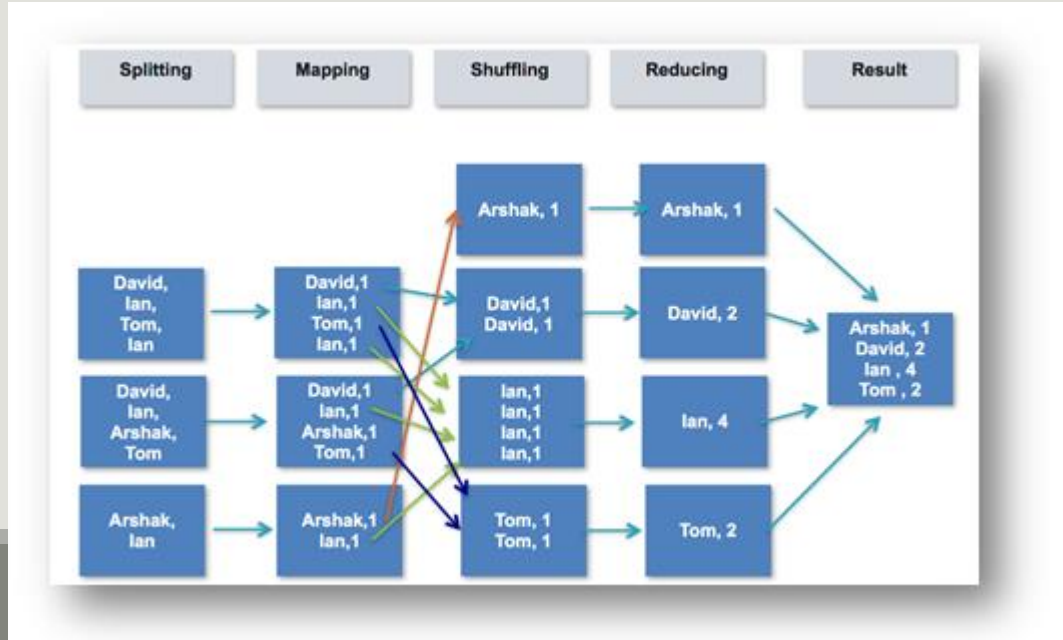


Apache Hadoop

E’ il principale framework di calcolo in ambito Big Data

- Nasce nel 2005 da progetti originatisi in Google tra il 2002 e il 2004
- E’ **robusto** rispetto a guasti HW e fallimenti di processi
- E’ basato su HW standard
- Permette l’elaborazione di enormi moli di dati per mezzo del **calcolo parallelo** su **cluster** di calcolatori
- Adotta un meccanismo di programmazione parallela semplice e nativamente parallelizzabile denominato **Map-Reduce**





Il grafo di Facebook

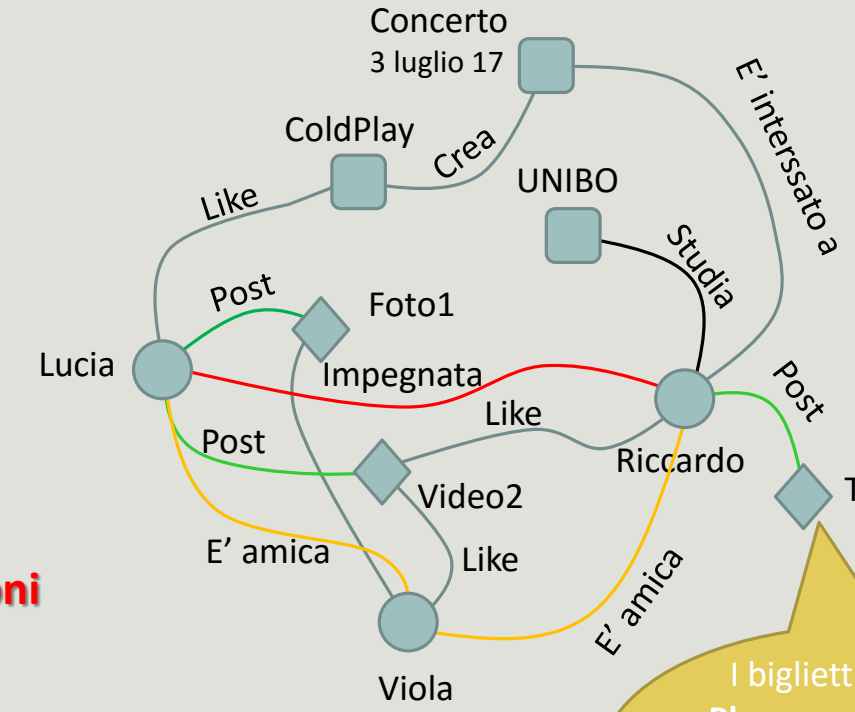
Il modello di dati di Facebook può essere concettualizzato tramite un grafo i cui i nodi modellano

- Utenti
- Pagine/Eventi
- Post/Foto/Video

… e in cui gli archi modellano interazioni

- Relazioni di Amicizia/Lavoro/Relazione sentimentale
- Like/Commenti
- Visualizzazioni

Ogni nodo e ogni arco è caratterizzato da ulteriori informazioni



I biglietti dei **Cold Play** sono finiti in un minuto! Non riuscirò mai ad andare a quel concerto!!

Le News Feed e l'algoritmo di *EdgeRank*

Se Viola ha 100 amici e ognuno di loro fa 2 post al giorno Viola rischia di essere sommersa dalle informazioni!! Ci vuole una strategia per selezionare le news più interessanti

- Per ogni utente *U* e per ogni post *P*, Facebook calcola la Relevance(*P*,*U*)

$$\text{Relevance}(P,U)^{*} = \text{Affinity}(P,U) \times \text{Performance}(P) \times \text{Type}(P) \times \text{Recency}(P)$$

- Affinity(*P*,*U*) = L'affinità tra *U* e l'utente che ha creato il post. Il peso è calcolato in base al numero di amici comuni e in base al numero delle interazioni
- Performance(*P*) = La performance del post rispetto su altri utenti in base al numero di Like e Condivisioni
- Type(*P*) = Status / Photo /Link / Video. Il peso è maggiore per i video
- Recency(*P*) = Più il post è recente maggiore sarà la sua relevance

*La formula è semplificata e ha un ruolo puramente esemplificativo (Facebook non ha reso pubblica la formula completa)

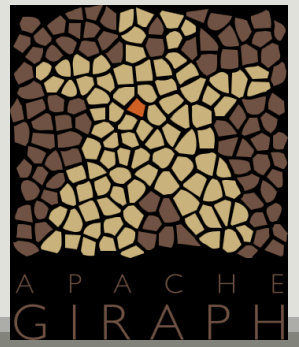
Le News Feed e l'algoritmo di *EdgeRank*

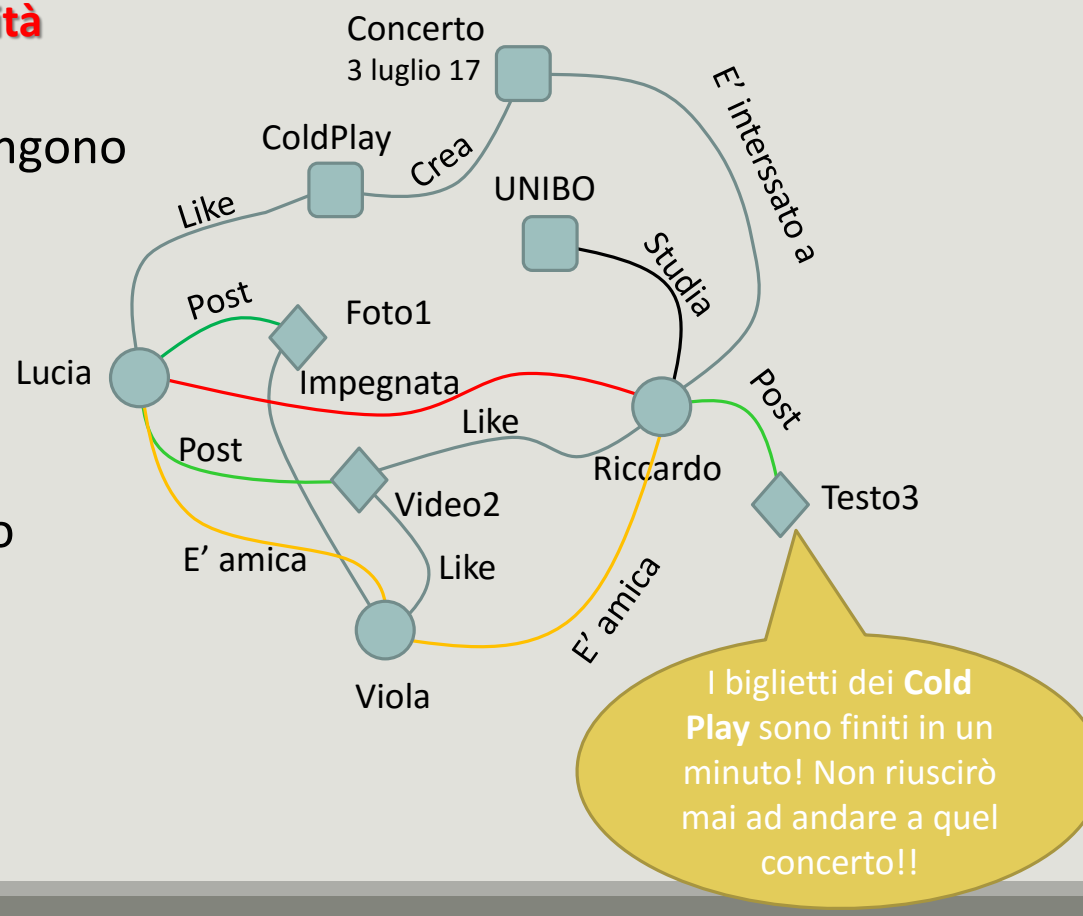
Viola ha buone possibilità che il post di Riccardo sia pubblicato sulla sua pagina di News perché riguarda una tematica di interesse per utenti con cui Viola ha alta affinità

Tutte le informazioni per il calcolo della relevance si ottengono visitando il grafo dei dati ma con:

- 2,5 M di post al minuto
- 1.8 M di like al minuto
- 1,18 B di utenti connessi ogni giorno

Solo un sistema di Big Data può mantenere FB aggiornato





Cosa ci insegna EdgeRank?

La tecnologia è utile quando ci fornisce un servizio nascondendo la complessità che serve a realizzarlo. La società moderna è pervasa da tali servizi.

- **Dal punto di vista scientifico** la codifica di un comportamento intelligente (‘*fatemi vedere solo ciò che è rilevante*’) richiede uno sforzo di modellazione, astrazione e quantificazione del concetto che si traduce in un algoritmo basato su una struttura dati

- **Dal punto di vista tecnologico** l’implementazione dell’algoritmo comporta un’enorme sforzo di implementazione e di ottimizzazione

- **Dal punto di vista etico:** percepiamo una realtà distorta. FB enfatizza in modo autoreferenziale le idee e gli interessi delle comunità di amici. Più la comunità afferma (post) e accredita (like) che una cosa è giusta/bella/importante, più saremo spinti a pensare che quella idea sia condivisa da tutti.

Un consiglio per il futuro…

La scelta del Corso di Studi da frequentare avrà un impatto fortissimo sulla vostra vita e la decisione deve essere un mix di due elementi: l'interesse verso l'area disciplinare e le prospettive di lavoro che quell'area offre.

Sulle prospettive di lavoro per i laureati in Ingegneria e Scienze Informatiche posso garantire….

…. valutate voi la vostra passione verso l'Informatica

Volendola mettere in formule:

$$\text{CourseRank}(U,C) = \text{Interesse}(U,C) \times \text{ProspettiveDiLavoro}(C)$$

Matteo Golfarelli (Phd)
Computer Science & Engineering
University of Bologna
Tel: +39 0547 338 862
e-mail: matteo.golfarelli@unibo.it
skype: matteo.golfarelli

www: http://bias.csr.unibo.it/golfarelli/
BIG: http://big.csr.unibo.it/

Business Intelligence Group