

Big Data

E' già troppo tardi oppure no?

Andrea Isidori

Società Generale d'Informatica SOGEI S.p.A.

Città, gg mese anno

BIG DATA un paio di definizioni

I **Big Data** sono risorse informative di grande **Volume**, di grande **Velocità** e grande **Varietà** che richiedono forme innovative ed economiche di elaborazione delle informazioni, che consentono una migliore comprensione e abilitano il processo decisionale e l'automazione dei processi. (Gartner)

“**Big Data** è un termine generico per insiemi di dati così grandi o complessi che le tradizionali applicazioni di elaborazione dati sono inadeguate.” (Wikipedia)



Momento Amarcord

Bytes

Kilobytes 10^3

Megabytes 10^6

Gigabytes 10^9

Terabytes 10^{12}

Petabytes 10^{15}

Exabytes 10^{18}

Zettabytes 10^{21}

Yottabytes 10^{24}

I floppy disk da 8 pollici nacquero nel 1967



Singola faccia,
capacità
formattata 110 kB

I floppy disk da 5 ¼ 1978 costo 10-15.000Lire

Una crescita esponenziale



2000: 800 Terabytes
2006: 160 Exabytes
2009: 500 Exabytes
2012: 2,7 Zettabytes
2020: 40 Zettabytes



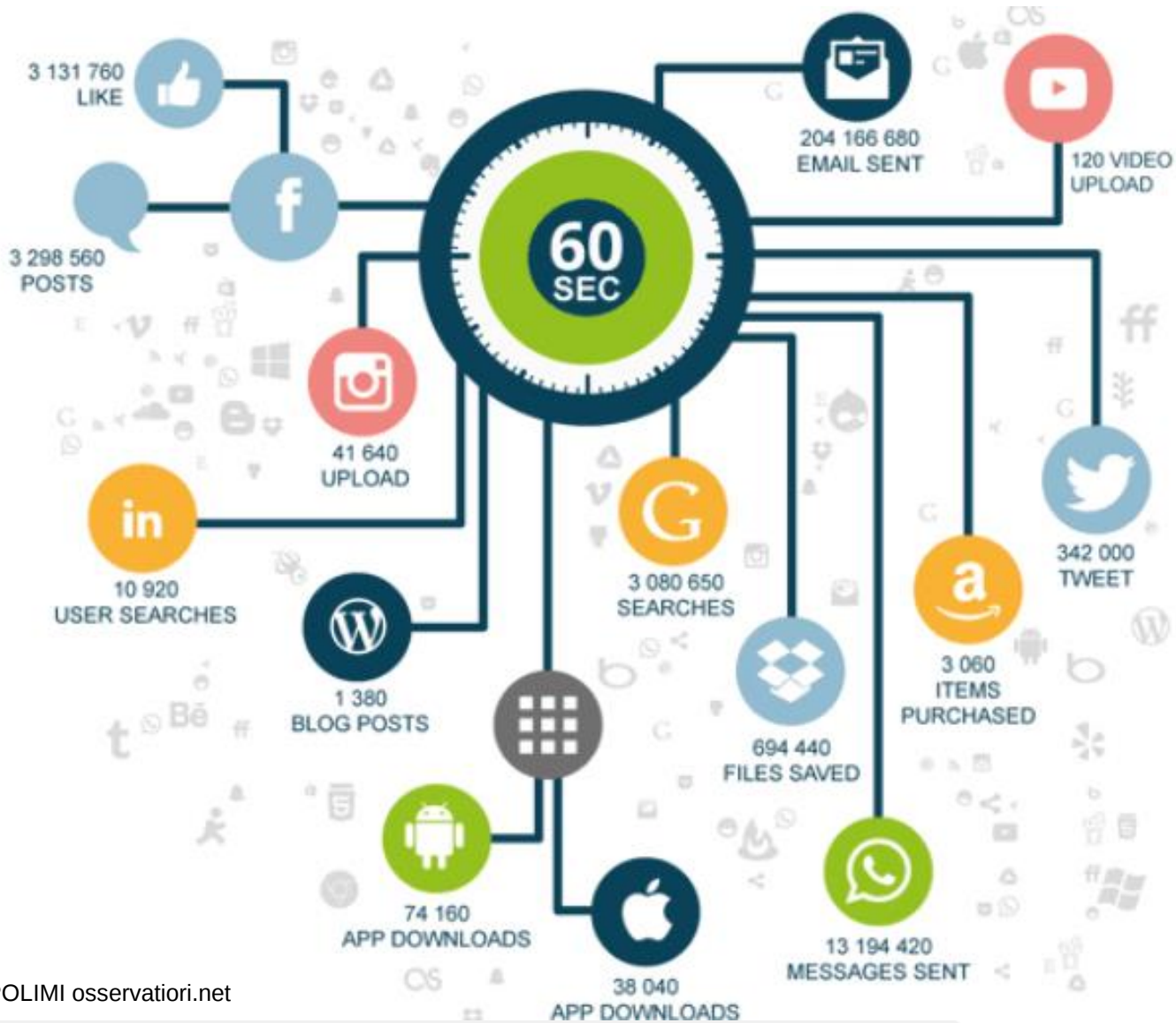
7 TB Twitter
10 TB Facebook



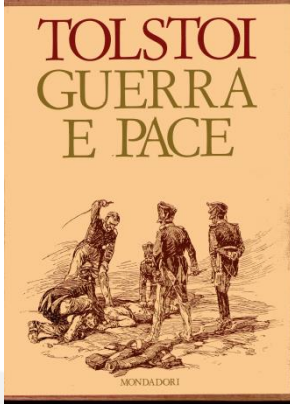
12 gigabyte di dati generati
da ognuno di noi ogni giorno

sogei

Ogni minuto nel mondo



Dal dato al dado!



561.304 parole

L'intero testo



Una immagine media

3 MB



1 dado standard da 16 mm

31 TeraBytes: uno scuolabus pieno di dadi



sogei

Sede Legale Via M. Carucci n. 99 - 00143 Roma

3 ExaBytes: lo stadio dei DallasCowboys



sogei

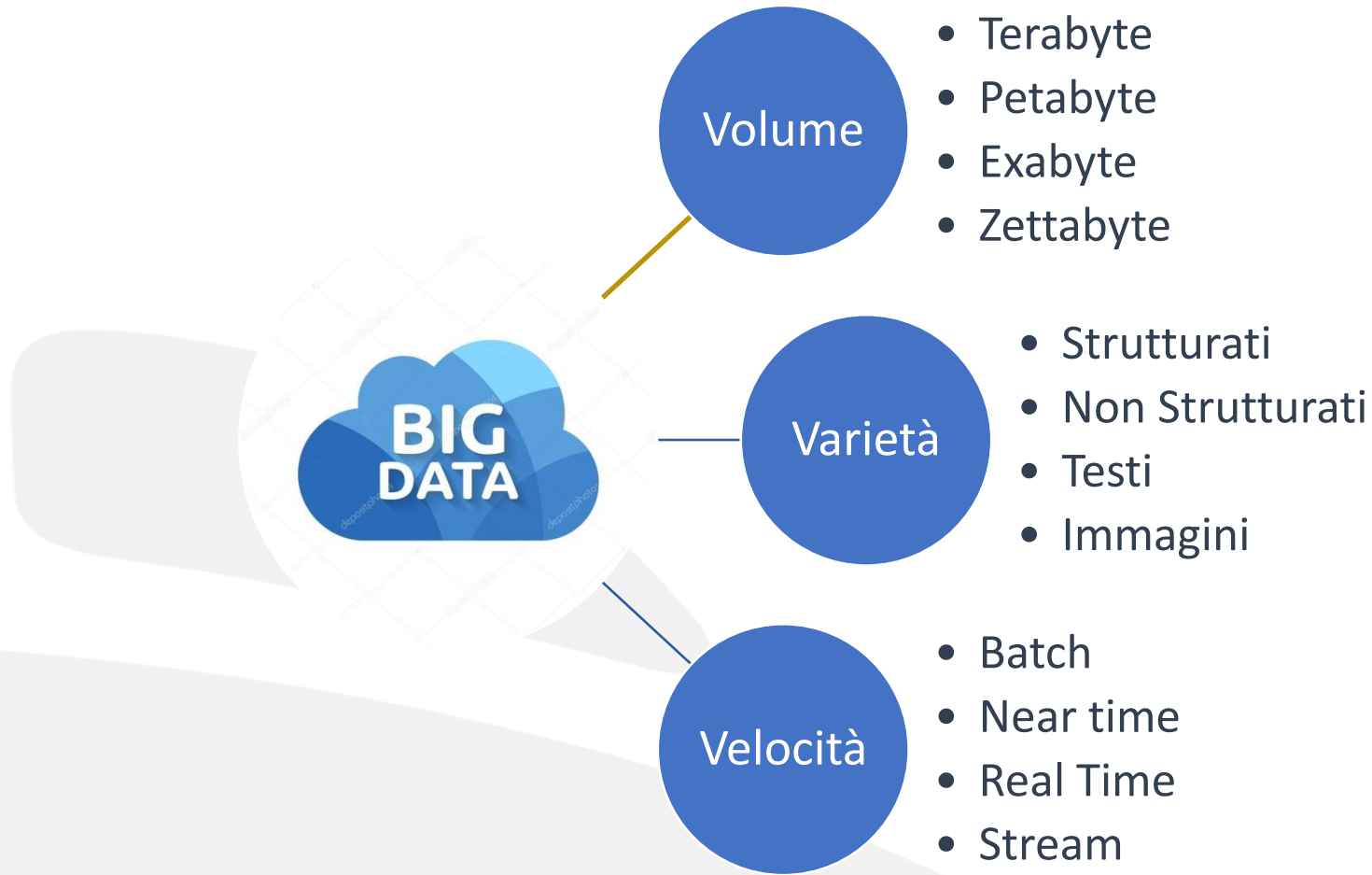
Sede Legale Via M. Carucci n. 99 - 00143 Roma

40 ZettaBytes: l'Everest



sogei

Le 3 V



Volume dei dati

40 ZETTABYTES

[43 TRILLION GIGABYTES]

of data will be created by 2020, an increase of 300 times from 2005



It's estimated that

2.5 QUINTILLION BYTES

[2.3 TRILLION GIGABYTES]

of data are created each day



6 BILLION PEOPLE

have cell phones



WORLD POPULATION: 7 BILLION

**Volume
SCALE OF DATA**



Most companies in the U.S. have at least

100 TERABYTES

[100,000 GIGABYTES]

of data stored

sogei

Varietà dei dati

As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES

[161 BILLION GIGABYTES]



**30 BILLION
PIECES OF CONTENT**

are shared on Facebook
every month



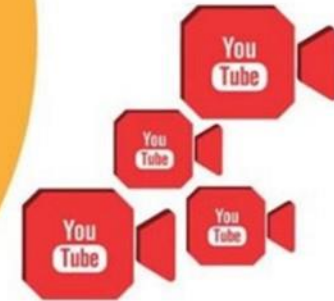
By 2014, it's anticipated
there will be

**420 MILLION
WEARABLE, WIRELESS
HEALTH MONITORS**



**4 BILLION+
HOURS OF VIDEO**

are watched on
YouTube each month



400 MILLION TWEETS

are sent per day by about 200
million monthly active users



Variety
**DIFFERENT
FORMS OF DATA**

sogei

Velocità dei dati

The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

during each trading session



By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

– almost 2.5 connections per person on earth



Modern cars have close to

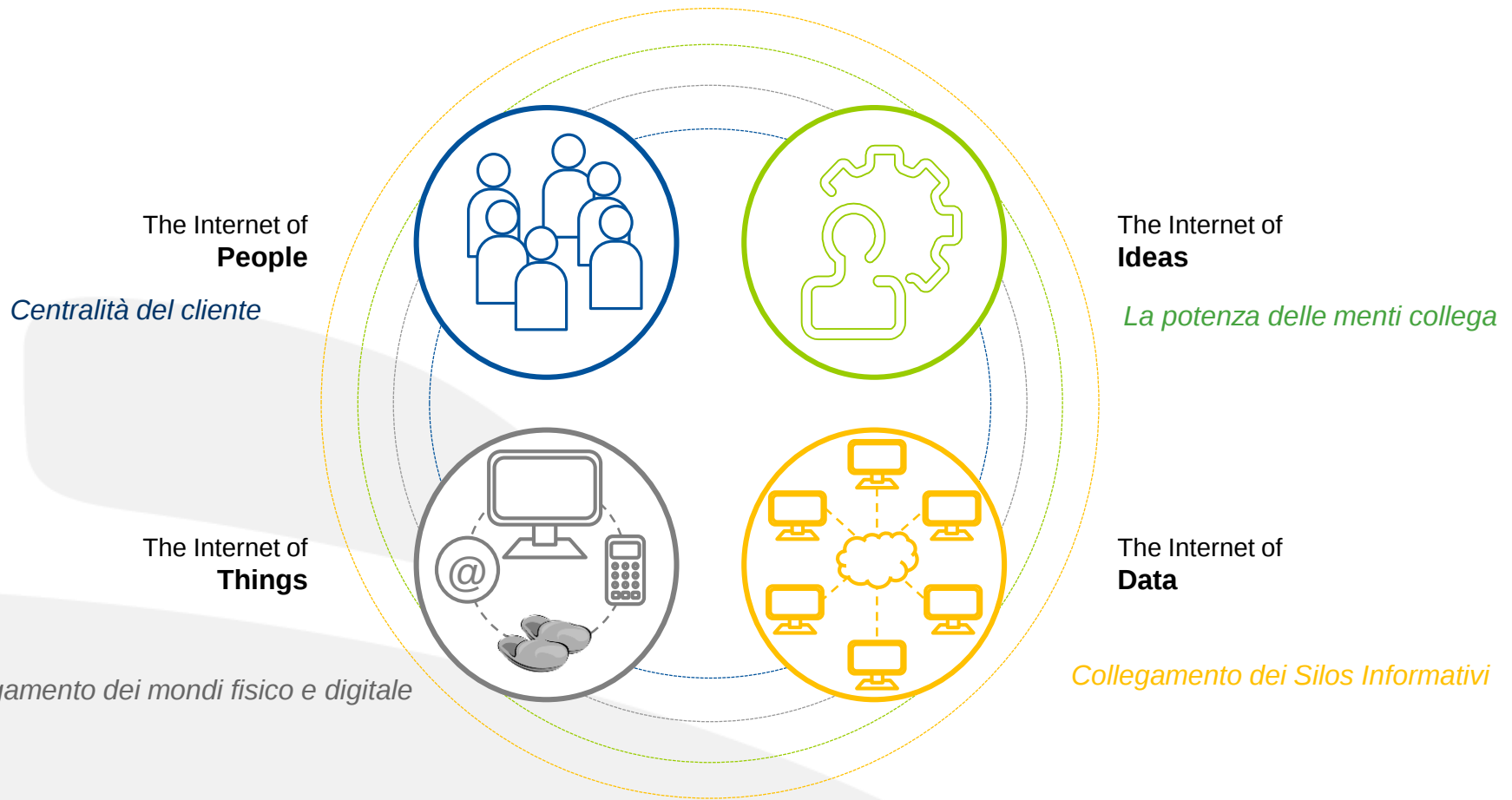
100 SENSORS

that monitor items such as fuel level and tire pressure

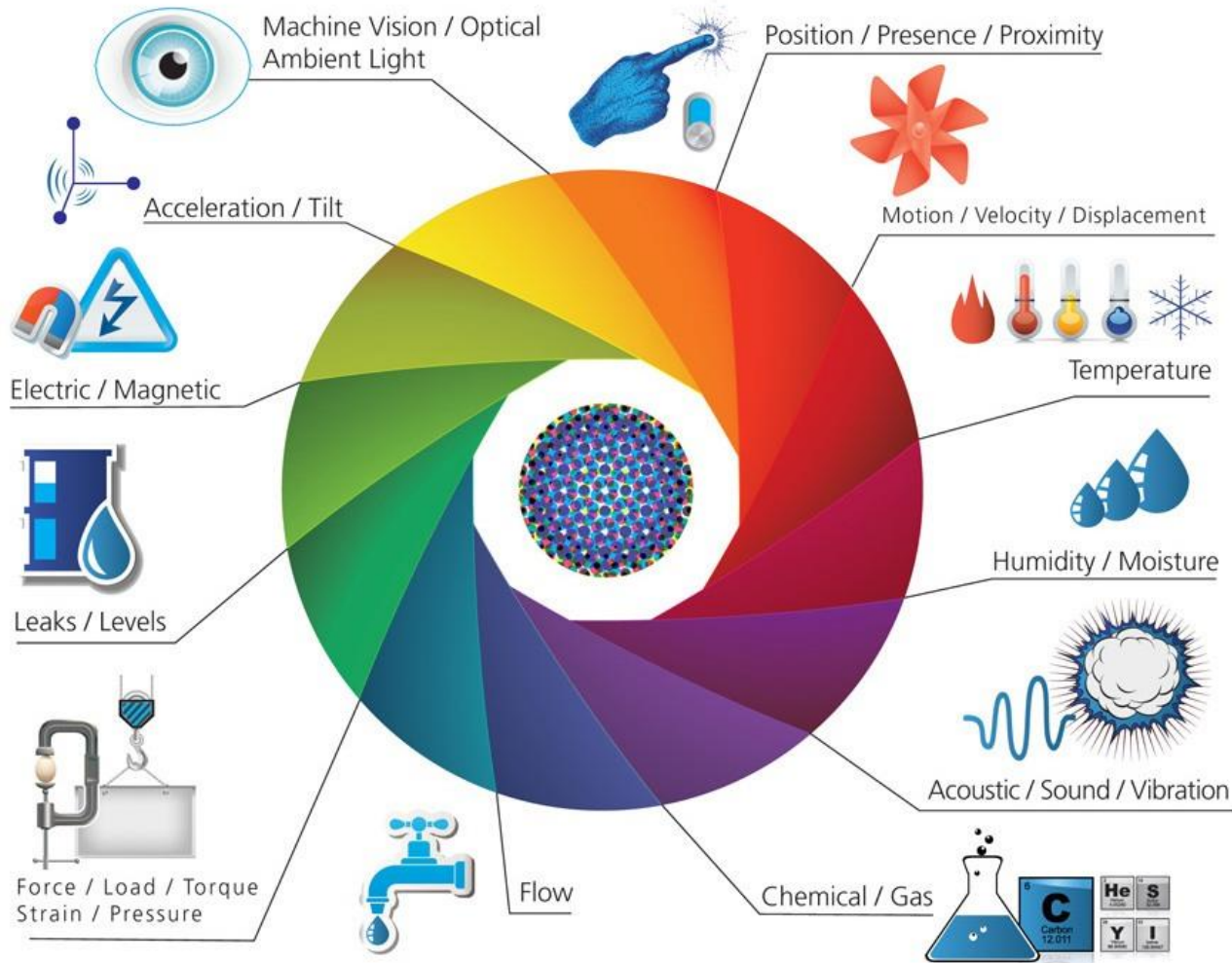
Velocity
ANALYSIS OF STREAMING DATA

sogei

The Internet of Everything!



Tutto è Smart



sogei

Cosa è un sensore

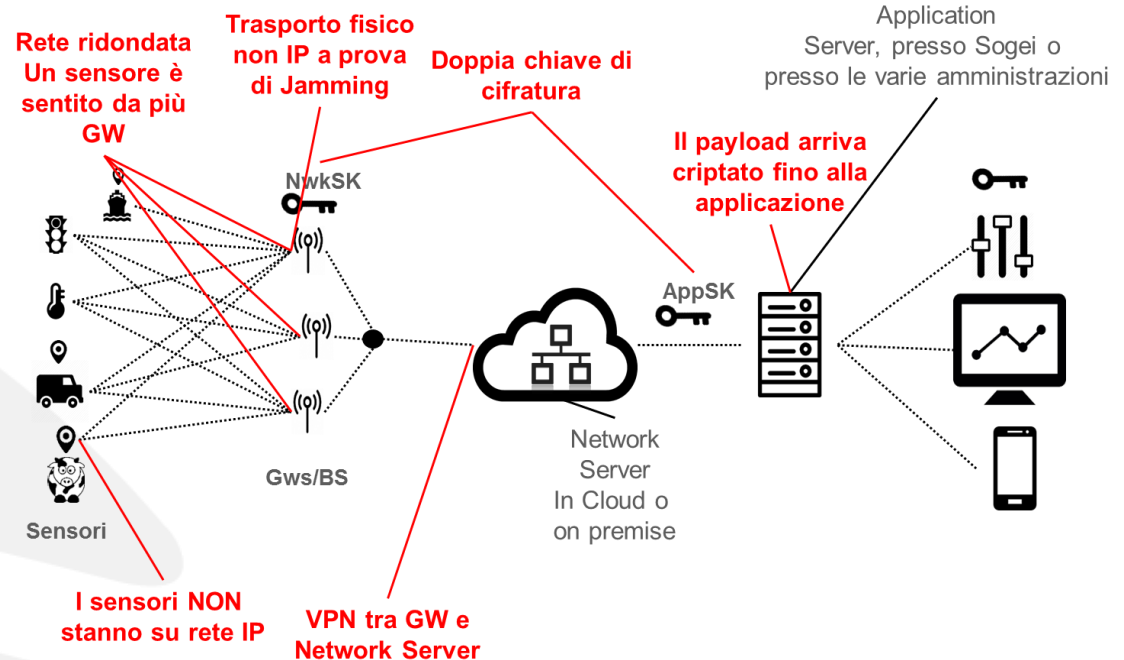


Xignal

Dimosystems BV

Xignal mouse-traps and rat-traps are LoRaWAN connected devices that monitor and communicate the trap-status: armed or un-armed, with or with-out catch...

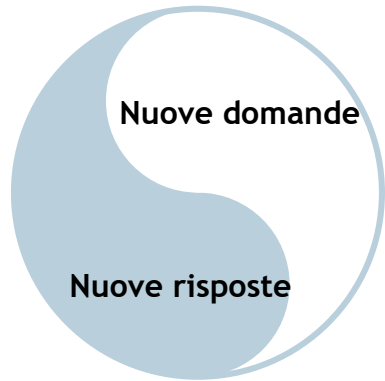
Product details >



sogei

Sede Legale Via M. Carucci n. 99 - 00143 Roma

Ma cos'altro è cambiato?



Persone + Processi = Nuove Domande

Persone + Processi + Tecnologia = Nuove domande



- Aspettative sui tempi di risposta
- Aggiornamento dati



- Tempi di elaborazione dei dati

1990

Canned Report
La BI su gestionali
La BI su ODS

2000

Primi Data Warehouse
OLAP e MOLAP

2010

Gara tra i vendor di BI sulla Performance
Data mining

Oggi

Machine Learning
Real Time
Data Lake

sogei

I dati cresceranno ancora



La sensoristica

- costa sempre meno
- consuma sempre meno
- già oggi, non ha bisogno di reti cellulari
- fa risparmiare tempo e denaro

Le richieste saranno più stringenti

Gli utenti vogliono

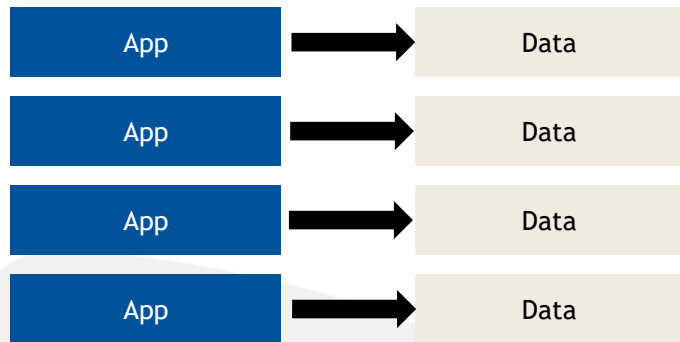
- monitorare qualsiasi dato
- incrociare qualsiasi informazione
- prevedere qualsiasi evento
- effettuare tutto questo in real time



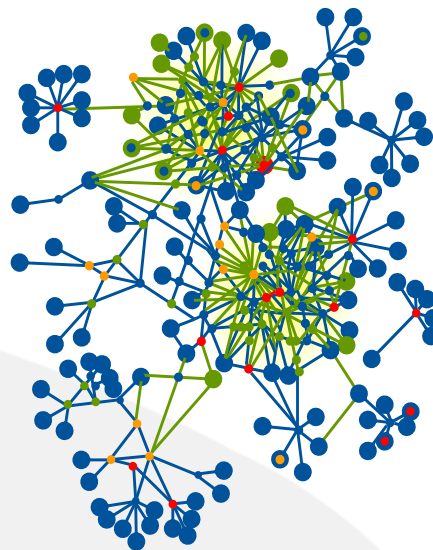
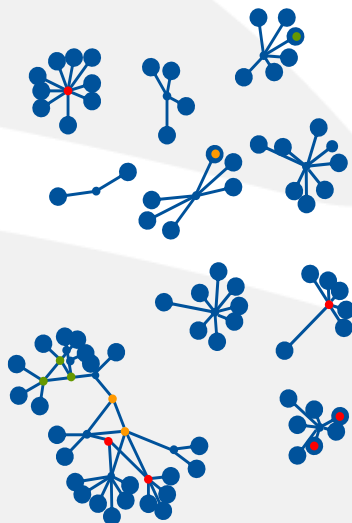
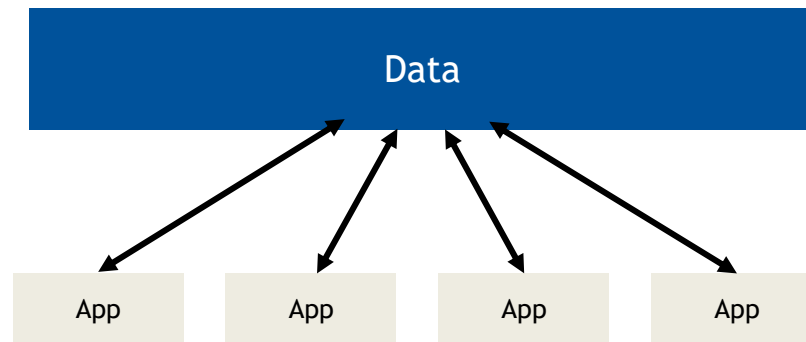
sogel

Dati sempre più connessi

Passato

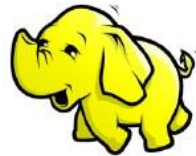


Presente



E necessario un nuovo approccio

Big Data describes the growing ability to answer questions with data that have been too expensive to answer before



**Gestire «Big Data»
necessita di
architetture e
strumenti differenti**

La tecnologia: Hadoop



“**Apache Hadoop** è un framework che consente l'**elaborazione distribuita** di grandi set di dati su **cluster di computer** utilizzando modelli di programmazione semplici. È progettato per **scalare** da singoli server a migliaia di macchine, ognuna delle quali offre calcolo e storage locale.”

Apache Hadoop è open source.

Apache Hadoop è lo standard de facto delle piattaforme Big Data come:

- Cloudera, Hortonworks, Oracle e IBM

Ma è un mondo in continua evoluzione...

HDFS: Hadoop Distributed File System

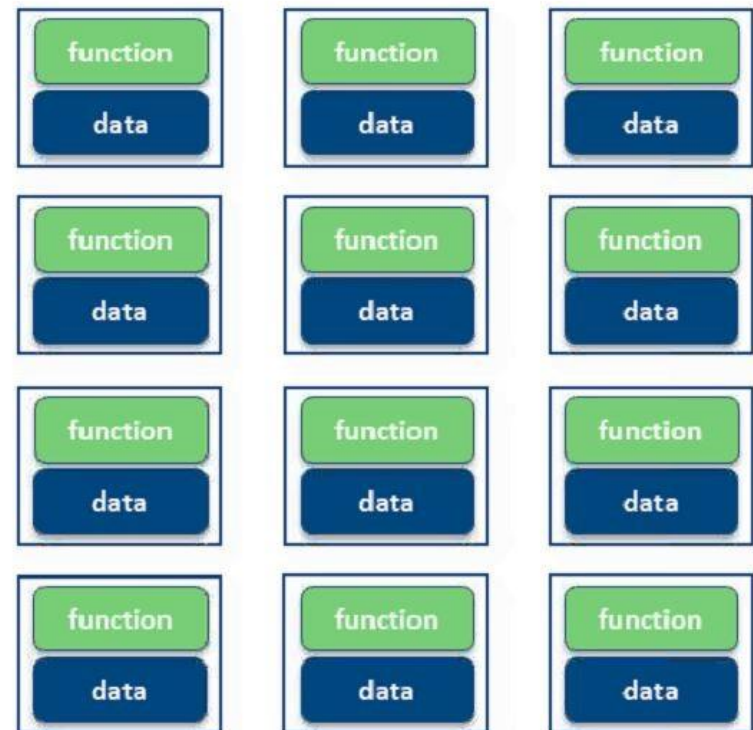
- I file sono suddivisi in **blocchi**, tipicamente di 64 MB
- I **blocchi**, sono distribuiti e replicati su differenti “nodi” (almeno 3 nodi di default)
- Un nodo speciale memorizza per ogni file, la posizione dei suoi **blocchi**
- Ogni nodo sa dove trovare il nodo “master”
- HDFS permette la creazione, cancellazione e lo spostamento dei file, ma non la loro modifica. Questo permette di superare i problemi tipici della modifica concorrente e di ottenere prestazioni eccellenti

Le differenze

Traditional Architecture



Hadoop



Hadoop MapReduce e Spark

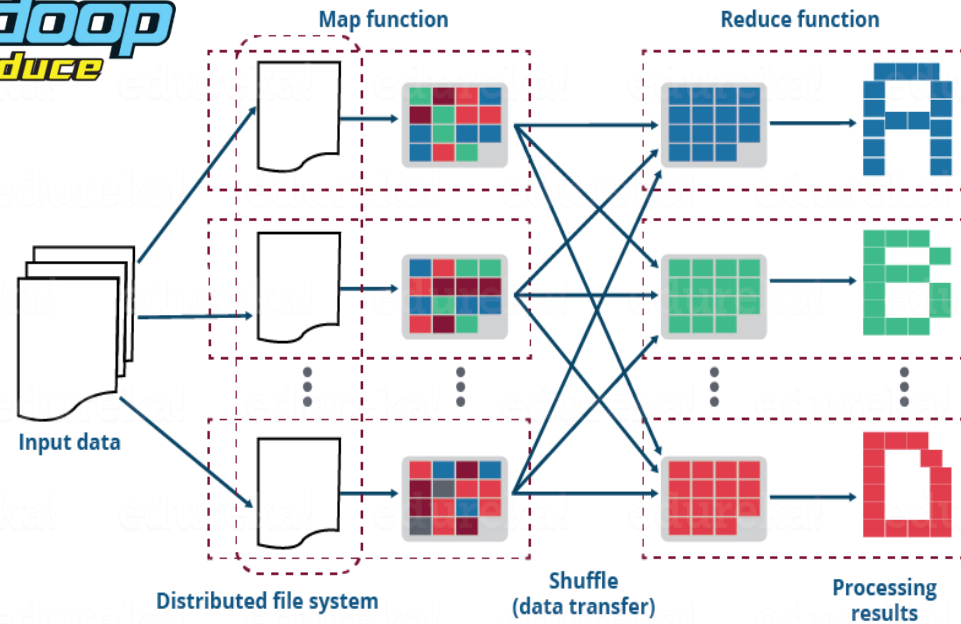


MAP:

- Iterazione di un gran numero di record in parallelo
- Estrarre qualcosa di interessante da ogni iterazione
- Mescolare e ordinare i risultati intermedi di diverse iterazioni simultanee

REDUCE:

- Risultati intermedi aggregati
- Generare l'output finale



MapReduce è un **modello di programmazione** progettato per l'elaborazione di **grandi volumi** di dati in parallelo, dividendo il lavoro in una serie di compiti indipendenti.

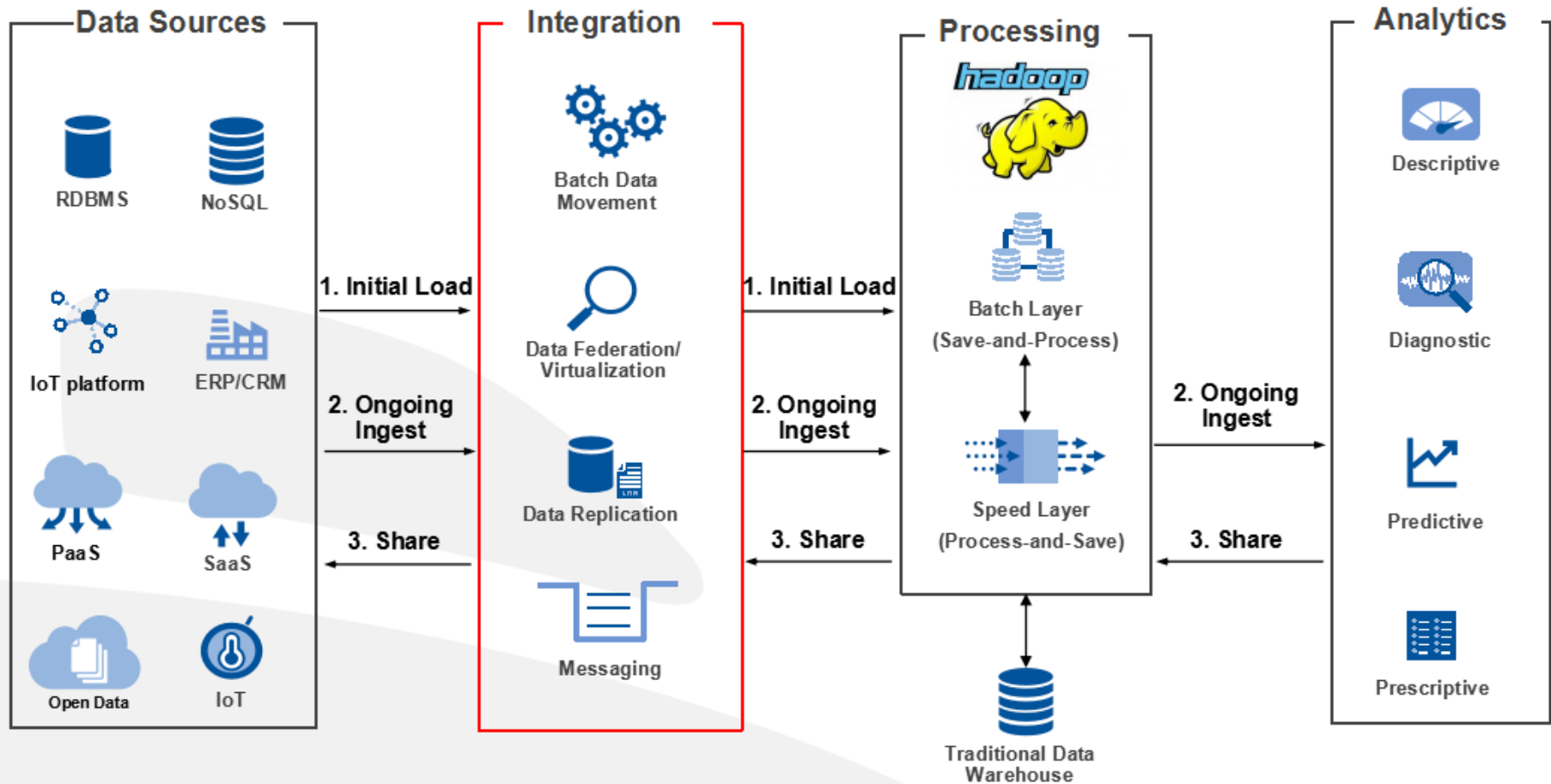
L'infrastruttura orchestra l'elaborazione eseguendo i vari compiti in **parallelo**, gestendo tutte le comunicazioni e i trasferimenti di dati tra le varie parti del sistema e prevedendo la **fault tolerance**

Spark è (anche!) un motore di calcolo clusterizzato in memoria che può sostituire MapReduce.

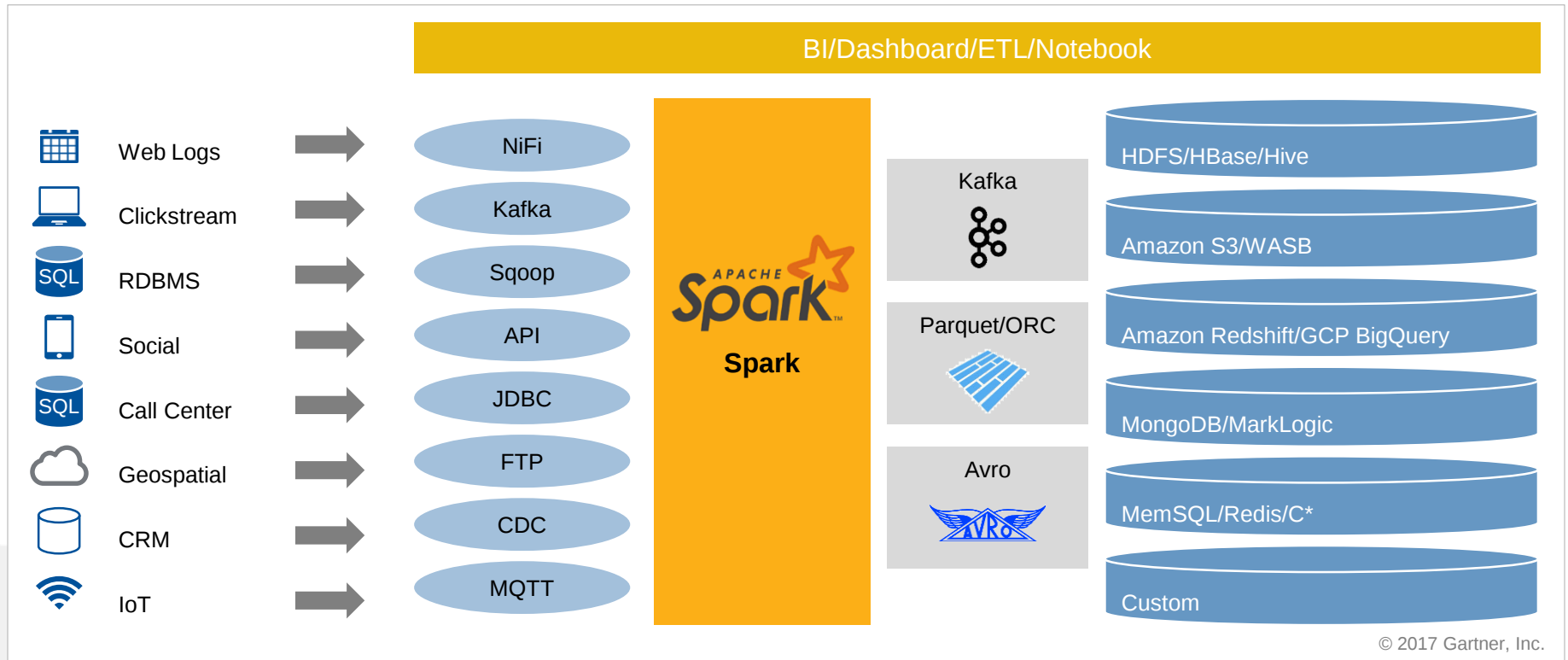


sogei

Un esempio di architettura Hadoop



Un esempio di architettura Spark



Chi utilizza piattaforme BigData?

facebook

twitter

NETFLIX

amazon

ebay



Google

YAHOO!

IBM

The New York Times

ORACLE®

LinkedIn

CISCO

sogei

sogei

Inutile zoommare...Ci sono proprio tutti!

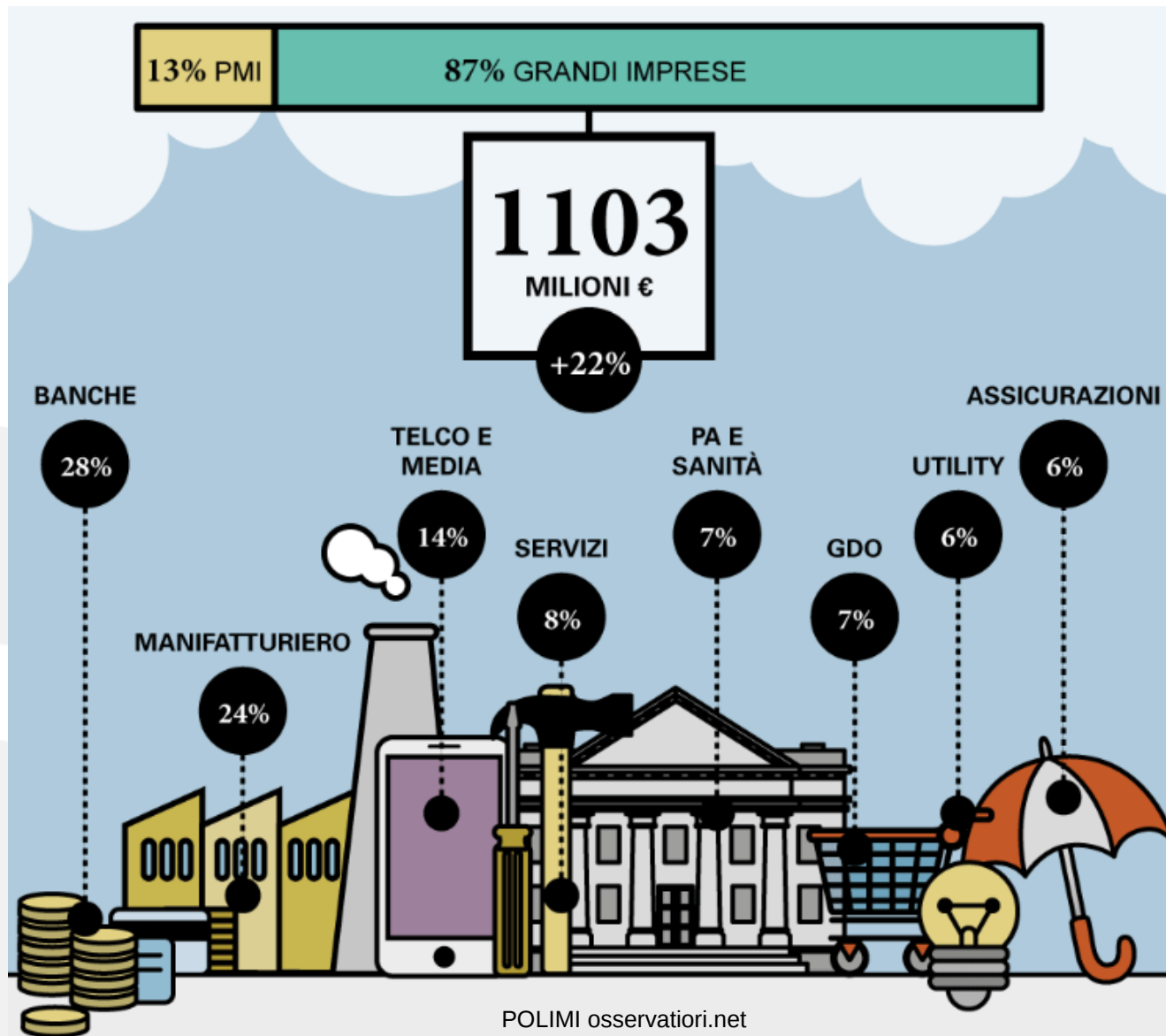
BIG DATA LANDSCAPE 2017



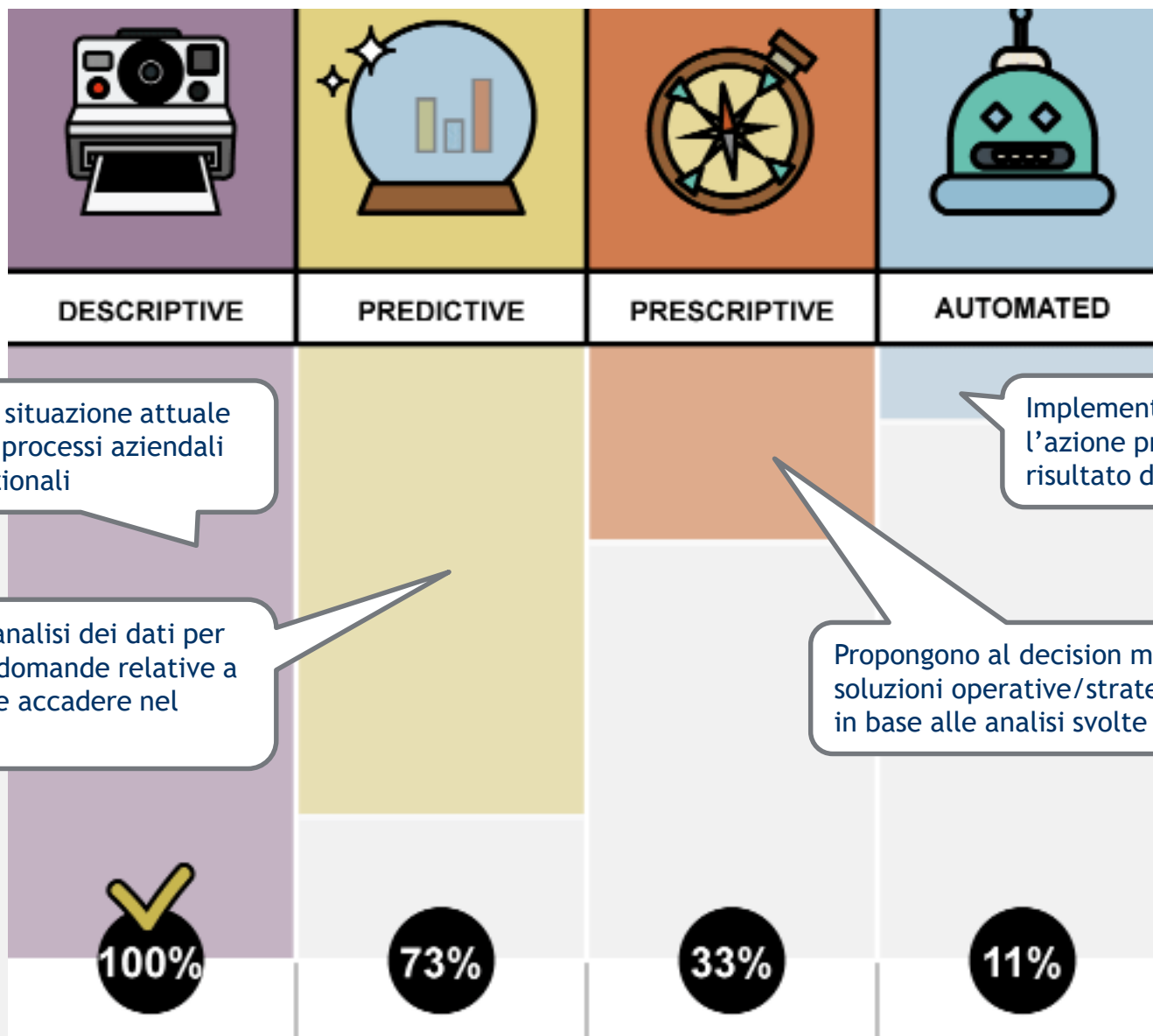
sogei

Sede Legale Via M. Carucci n. 99 - 00143 Roma

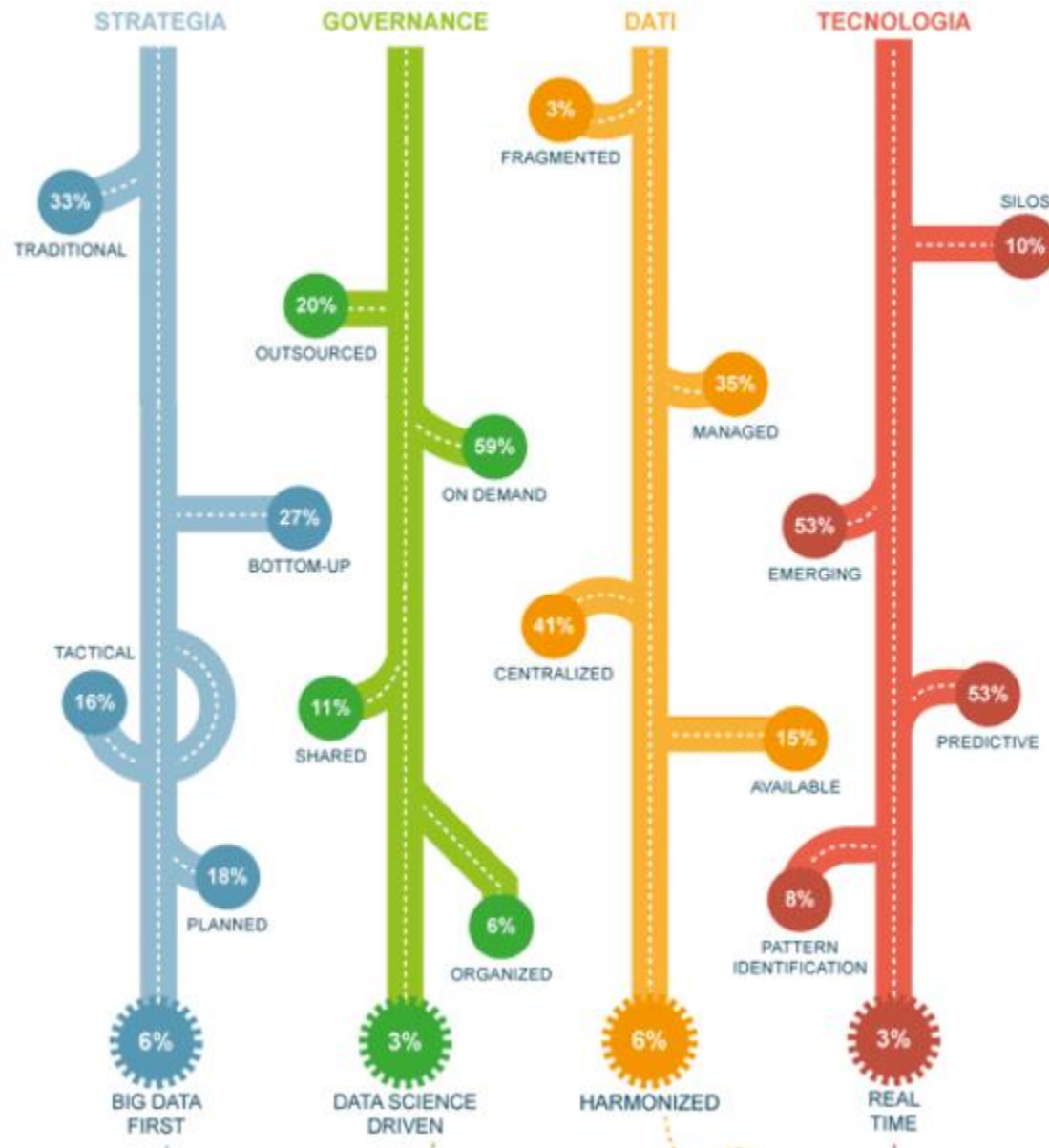
Lo scenario



L'adozione degli analytics nell grandi imprese



La maturità delle organizzazioni italiane



Tecniche per l'analisi dei big data

- **Data Mining**
 - Association rule learning
 - Classification
 - Cluster analysis
 - Regression
- **Crowdsourcing**
- **Data fusion and data integration**
- **Genetic algorithms**
- **Machine Learning**
 - Supervised Learning
 - Unsupervised Learning
- **Natural Language Processing**
- **Neural Networks**
- **Network Analysis**
- **Pattern Recognition**
- **Predictive modeling**
- **Sentiment analysis**
- **Signal processing**
- **Spatial analysis**
- **Simulation**
- **Time series analysis**
- **Visualization**
- **Chat bot**

Come vengono utilizzati oggi?

Quali funzioni

64%

CRM ANALYTICS

- analisi comportamentali
- preferenze utenti

56%

FINANCE ANALYTICS

- gestione rischio credito
- analisi indici azionari
- gestione frodi

41%

TOP MANAGEMENT DASHBOARD

44%

SOCIAL & WEB ANALYTICS

39%

CUSTOMER EXPERIENCE ANALYTICS

Quali aree in crescita?

57%

MOBILE

49%

ANALYTICS IN MEMORY

42%

SOCIAL ANALYTICS

36%

NEAR/REAL TIME ANALYTICS

Quali settori?

>25%



ASSICURAZIONI



BANCHE

15%-25%



UTILITY



MANIFATTURIERO

10%-15%



SERVIZI



TELCO E MEDIA

< 10%



PA E SANITÀ



GDO

Quali possibilità nella PA? Dite la vostra!

Sfide odierne	Dati	Possibilità
Sanità: Controllo della spesa	Fascicolo Sanitario elettronico	Cure preventive, riduzione delle ospedalizzazioni
Supporto alla gestione	Sensori su «macchine» (IoT)	Auto diagnosi, supporto alla manutenzione, Prevenzione dei guasti, ottimizzazione interventi
Standardizzazione dei servizi	Soddisfazione dei cittadini	Servizi su misura, riduzione dei costi
Orientamento al cliente Comportamento del cliente	Social Media	Analisi del sentimento degli utenti, verifica della soddisfazione monitoraggio canali social
Fraud Management	Giochi, Scommesse, Tasse	Analisi delle frodi, analisi delle transazioni
Sanità prescrizioni	Ricetta elettronica	Analisi delle prescrizione di farmaci
Sicurezza dati, applicazione normative e conformità (GDPR)	Dati SIEM	archiviare informazioni sulla sicurezza e sugli eventi (SIEM) per lunghi periodi di tempo invece di scaricarli dopo diversi mesi
Gestione del rischio e conformità	ERM	Analisi in real time, analisi di impatto

Tante opportunità per la PA



TAVOLA PERIODICA DEGLI ELEMENTI

Gestiamo un numero enorme di
informazioni
che crescono esponenzialmente.
Molte sono ancora da scoprire,
collegare, arricchire
A noi trasformarle
in valore e crescita per il paese

sogei

E quindi Big Data si o no nella PA?

Sono già qui. Le tecnologie esistono e creano valore

I punti di attenzione

- Dati prevalentemente testuali o numerici
- Occorre contenere le spese nell'interoperabilità dei dati
- Grande disponibilità di dati ma spesso non analizzati

I punti critici

- Livelli di adozione di soluzioni molto bassi
- Spesso non possono essere pubblicati in trasparenza

In generale

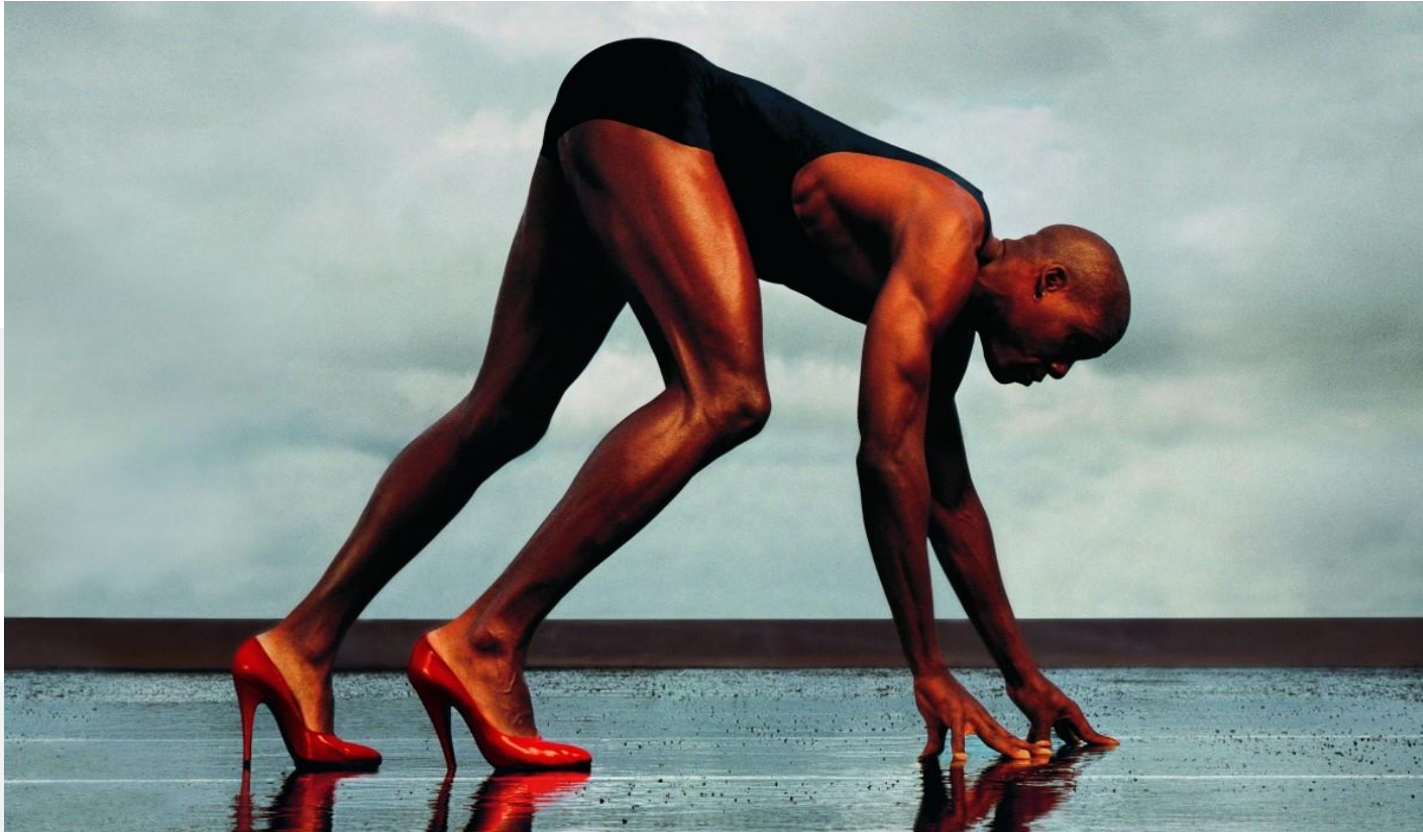
- E' necessario attivarci
- Ma siamo certi di comunicare ed analizzare al meglio i dati che già abbiamo, con gli strumenti attuali?

Inutile avere una informazione in real time se poi la osserviamo una volta al mese (o mai!)



sogei

La potenza è nulla senza il controllo

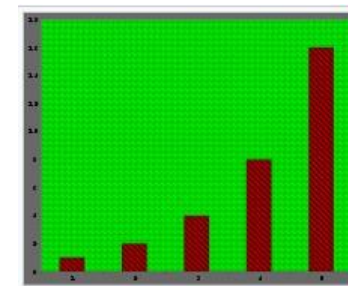
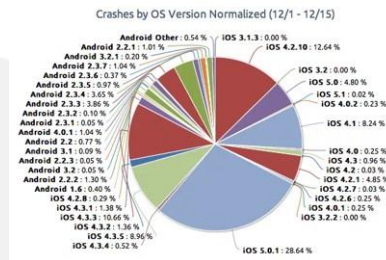
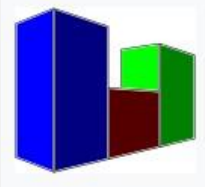
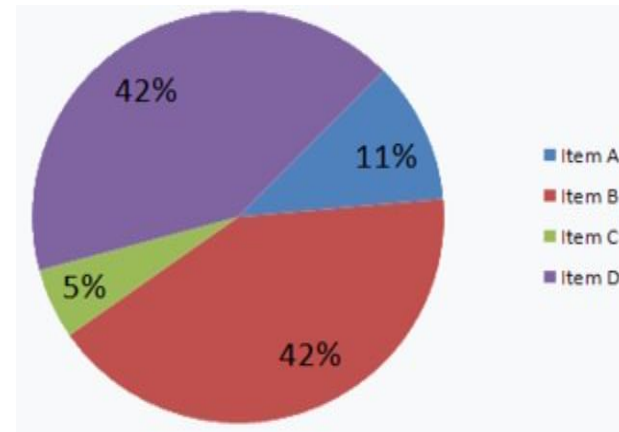
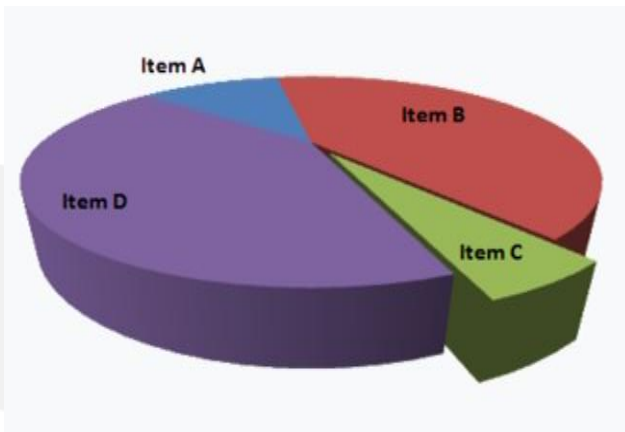


Imponete la semplicità



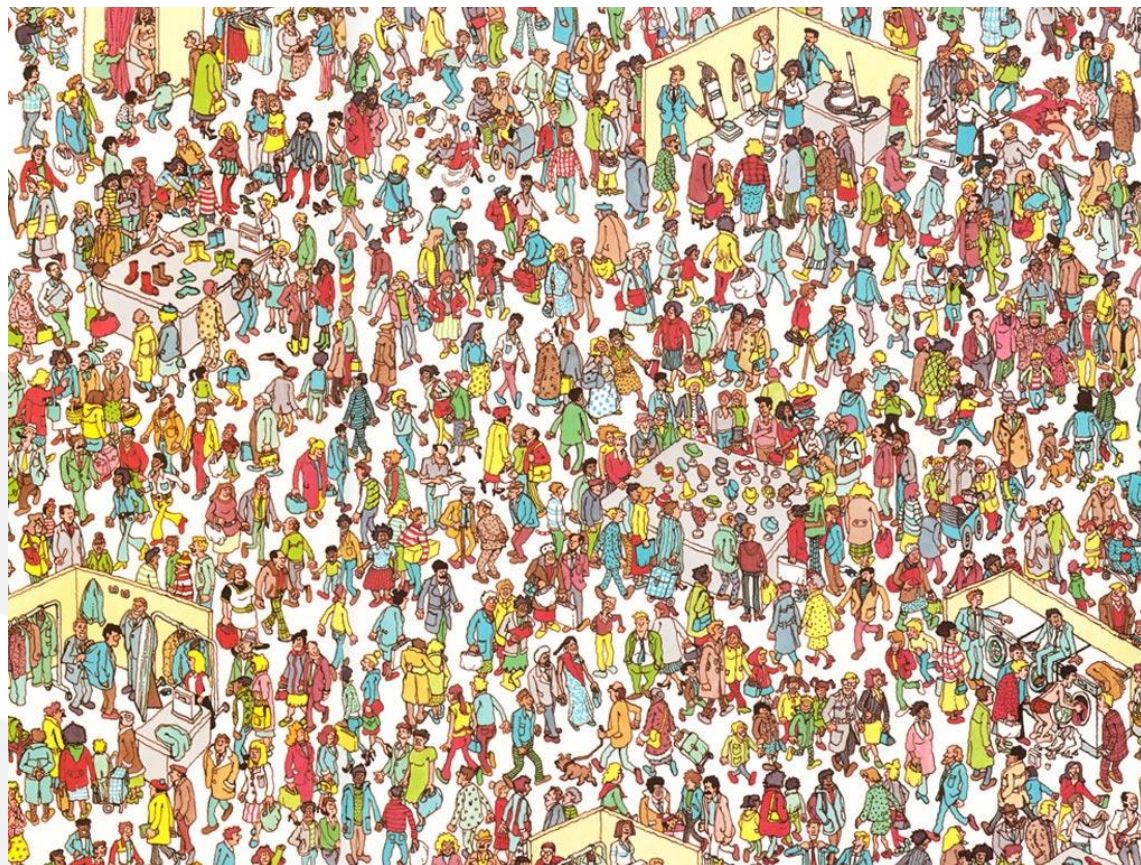
A nulla servono i Big Data se poi li presentiamo male

3D, Misleading graphs, chartjunk ed altri errori/errori



sogei

Cruscotti e pagine stile “Where is Wally style”

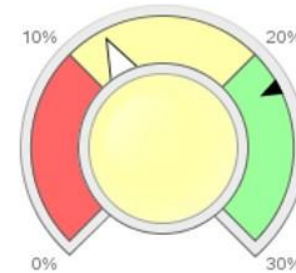
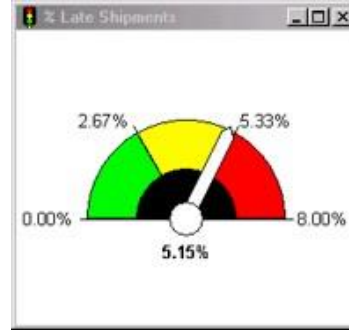


Non sovraffollare

sogei

Sede Legale Via M. Carucci n. 99 - 00143 Roma

Gauges, tachimetri ecc



Con moderazione!

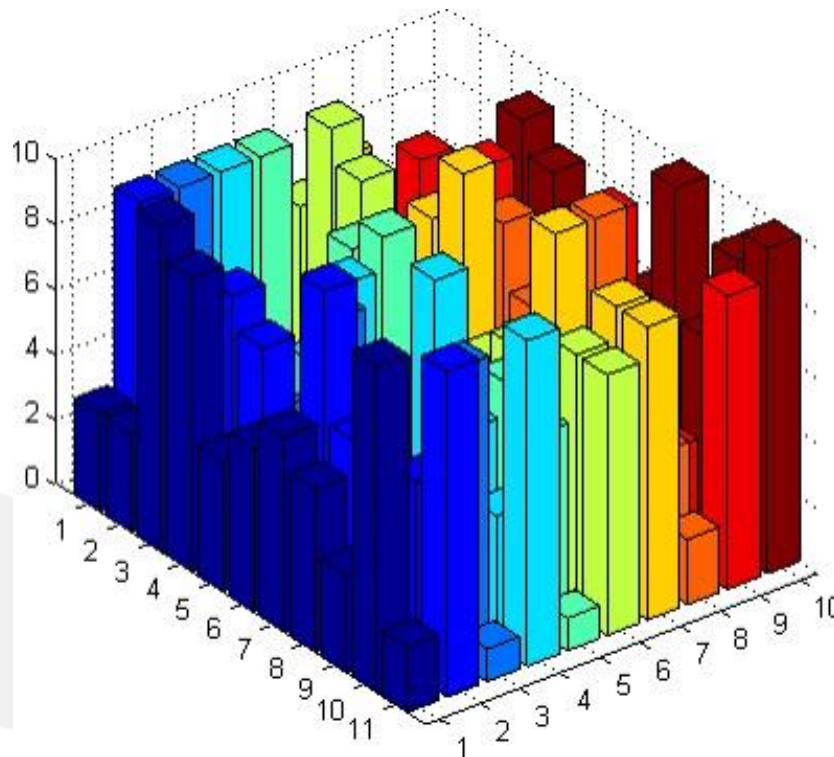
No custom: attenzione alla manutenibilità, le soglie

Nascondono informazioni ad esempio i trend

Rubano spazio: meglio un numero, scorecard, micrograph, istogrammi..



«Manhattan report» il fascino incomprensibile del 3D



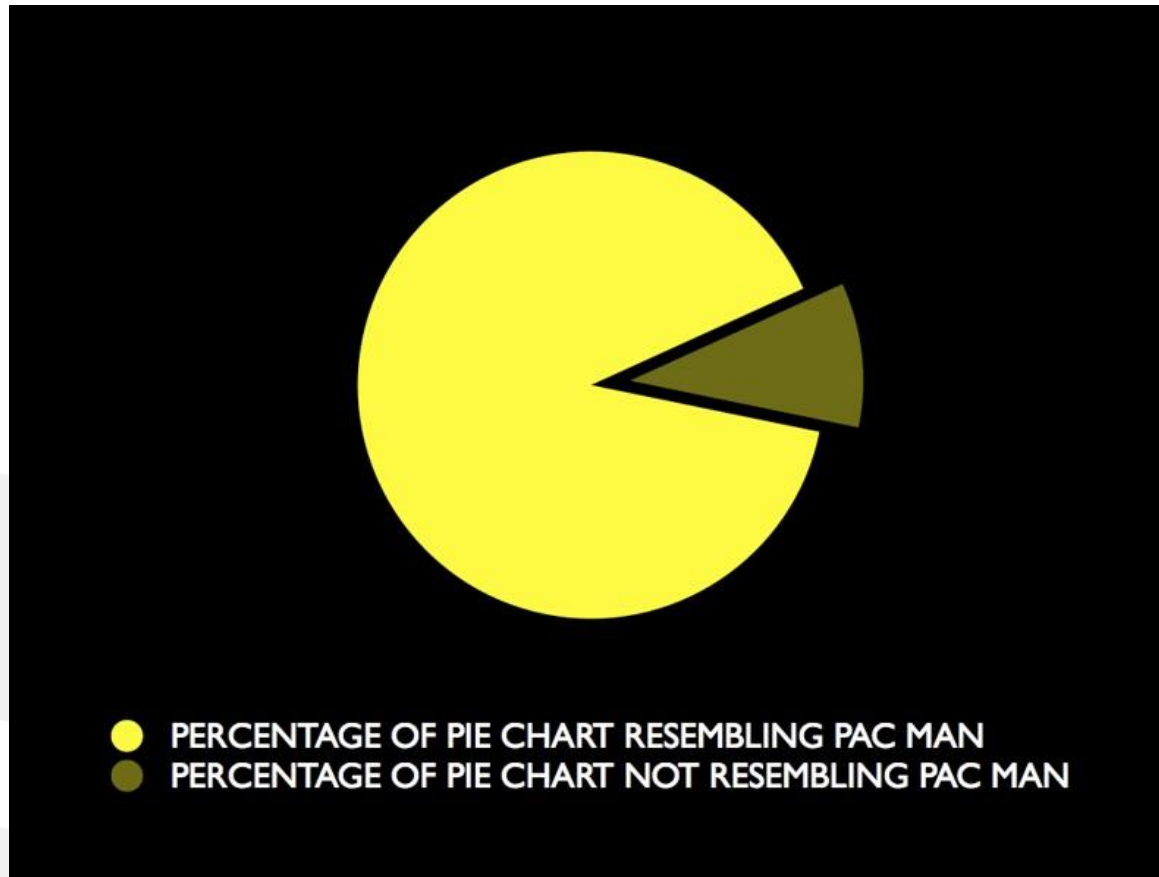
**CRIME
ALERT**



C'è stato un omicidio tra la Quinta e la Terza...
Ma voi non lo saprete mai!
Mai il 3D...E' un vero crimine!

sogei

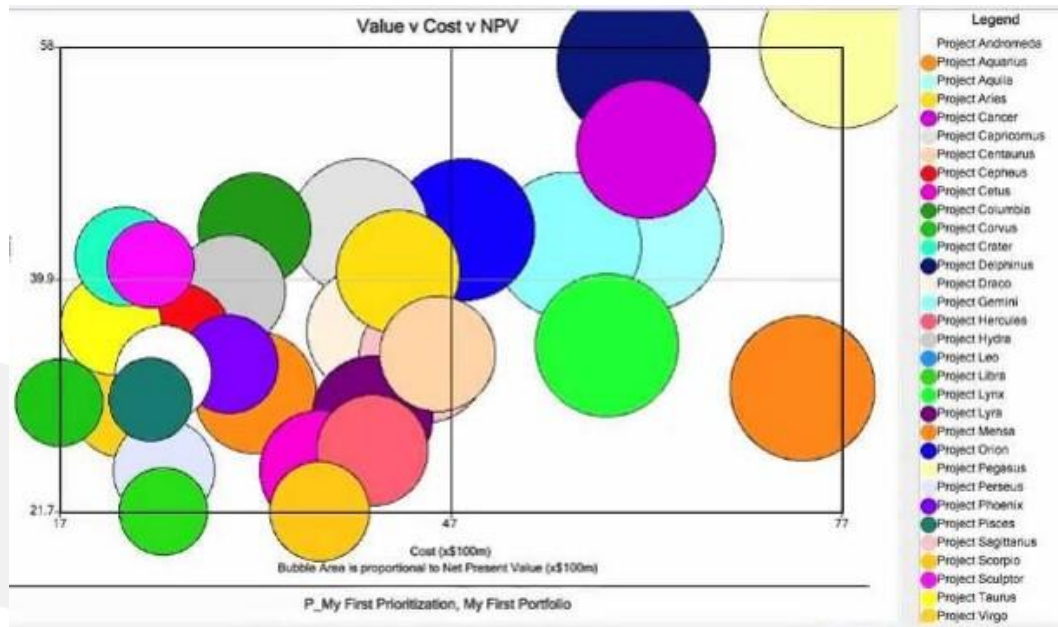
Il report «Pac Man»



Report che non cambiano mai o che
cambiano quando è ormai troppo tardi...

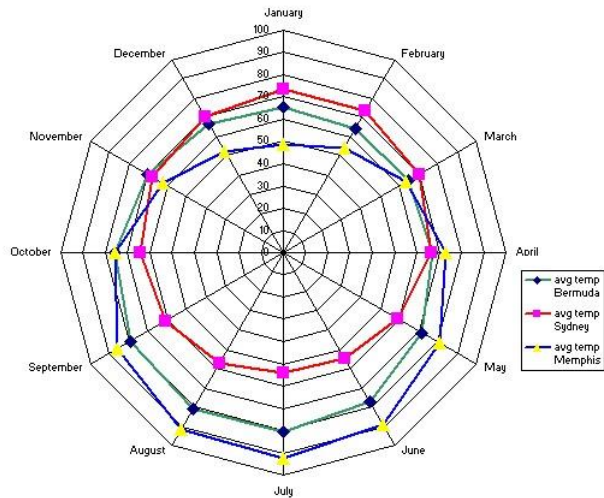
sögei

Report «busta di coriandoli» (Bubble chart)

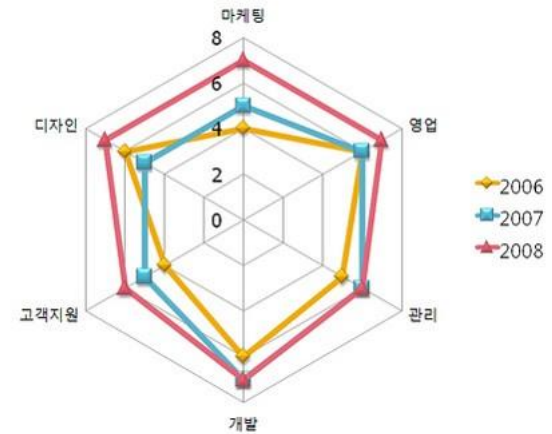


Perfetto... se vuoi nascondere un fenomeno
Bolle nascoste, colori simili, confronti impossibili
Meglio «mappe di calore», o istogrammi

Lo «Spiderman Chart» Il mitico Radar



2008 부서별 만족도 조사 결과



Lo so affascina molto gli statistici...ma
...Ci sono altre 5 rappresentazioni per
renderle i dati più comprensibili

Utilizzo improprio di report



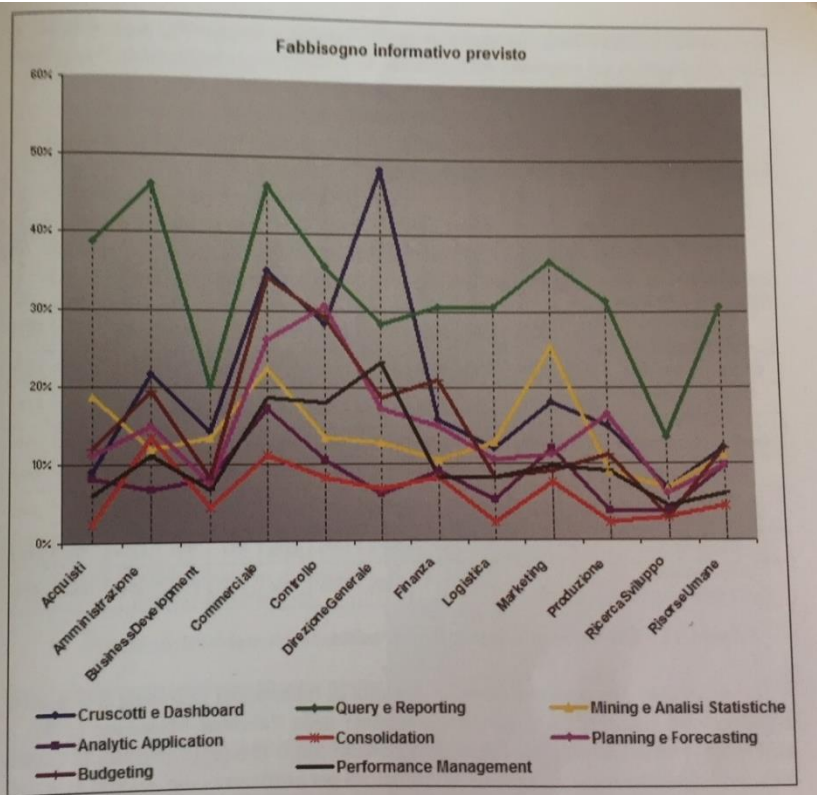
Andavamo bene in
Amministrazione con
Cruscotti&dashboard

ma poi abbiamo avuto
un calo improvviso nella
BusinessDevelopment

per fortuna ci siamo
ripresi con il
Commerciale

Abbiamo poi chiuso
abbastanza bene con
Risorse umane

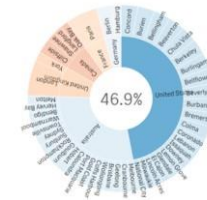
Le linee solo per il tempo, unico colore per la
misura e la dimensione



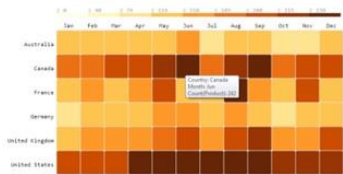
Ma allora cosa si può fare?



			Monthly	3-Mon Rolling	Budget/Actual	Variance to Budget
All	● \$	1,652,992				-2%
Chewbacca	● \$	116,303				-7%
Hansolo	● \$	489,337				22%
James Kirk	● \$	480,777				-6%
Luke Skywalker	● \$	556,575				-14%



Total Sales ⓘ
555.4k ▼ 104.9M ✓
Sales



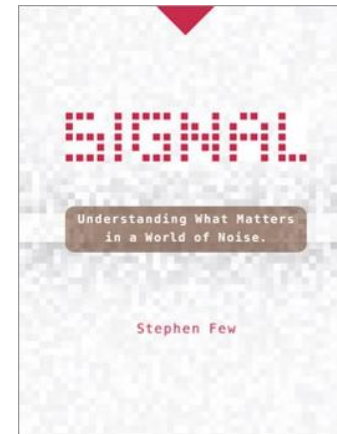
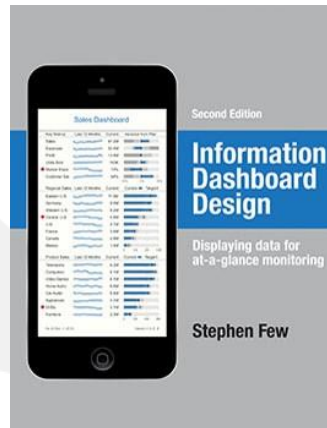
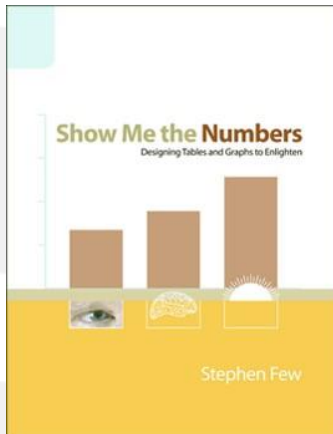
Lasciatevi guidare da quello che già esiste...

sogei

Comunicare bene nel mondo dei Big Data (ma anche gli small data!)

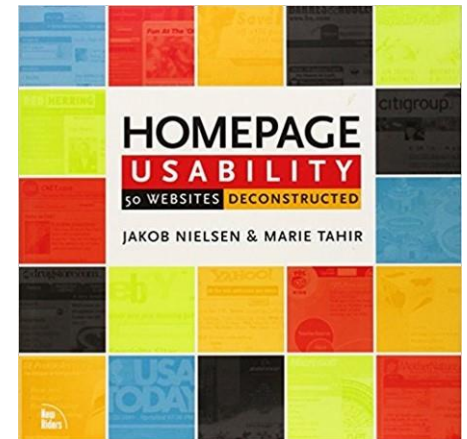
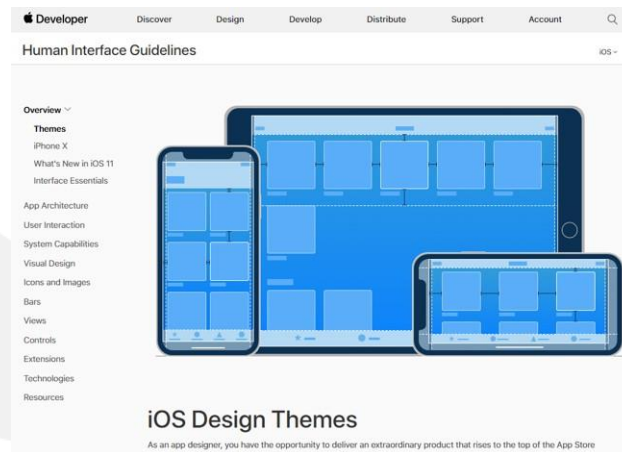
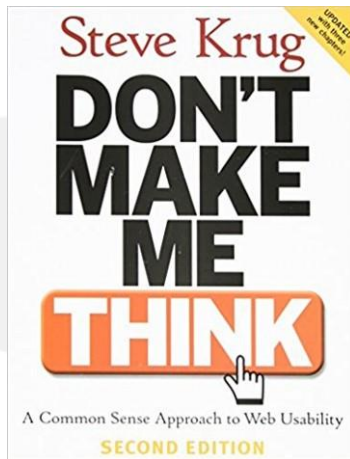
Stephen Few

perceptualedge.com



sogei

Dalla User Experience alla «Citizen Experience» per essere Digitali



Sull'usabilità nel quotidiano

Donald A. Norman

«La caffettiera del masochista» 1988

