BIG DATA: INTRODUCTION, RISKS, AND OPPORTUNITIES

SAPIENZA
UNIVERSITÀ DI ROMA

FEDERMANAGER ACADEMY

Master in Data Intelligence e Strategie Decisionali
Dipartimento di Scienze Statistiche

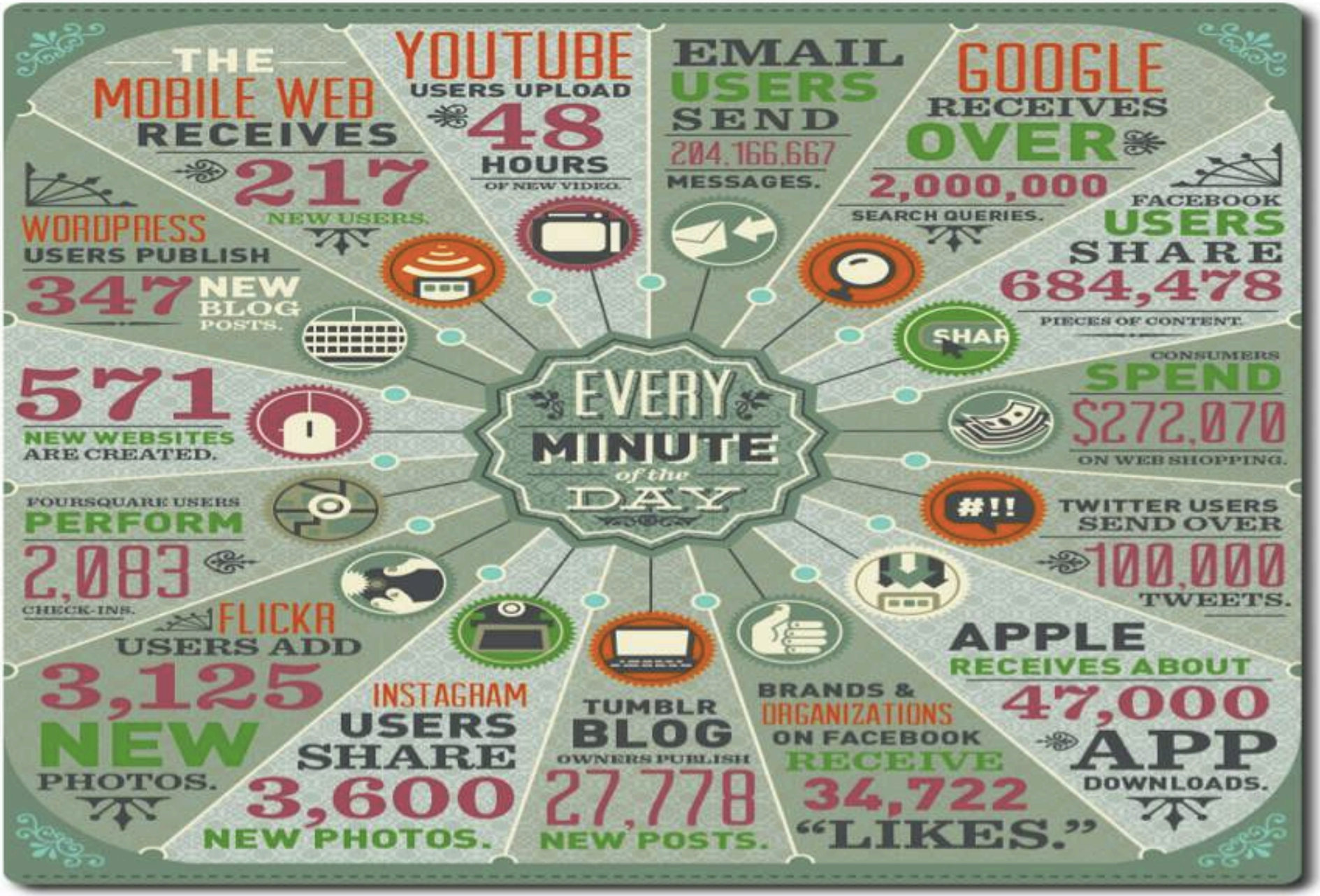
Paolo Dell'Olmo

Organizzazione delle lezioni

1. Introduzione
2. Tecniche ed esperienze
3. Tecnologie
4. Data Driven Decision Making
5. Machine Learning
6. Marketing and BI
7. Cognitive Computing
8. Progetti

Outline

1. Introduzione ai Big Data e agli Analytics
2. Small, Open e Big Data
3. Rischi
4. Opportunità
5. Esempi and Takeaways


EVERY MINUTE of the DAY
THE MOBILE WEB RECEIVES 217 NEW USERS.
YOUTUBE USERS UPLOAD 48 HOURS OF NEW VIDEO.
EMAIL USERS SEND 204,166,667 MESSAGES.
GOOGLE RECEIVES OVER 2,000,000 SEARCH QUERIES.
FACEBOOK USERS SHARE 684,478 PIECES OF CONTENT.
WORDPRESS USERS PUBLISH 347 NEW BLOG POSTS.
571 NEW WEBSITES ARE CREATED.
CONSUMERS SPEND $272,070 ON WEB SHOPPING.
FOURSQUARE USERS PERFORM 2,083 CHECK-INS.
TWITTER USERS SEND OVER 100,000 TWEETS.
FLICKA USERS ADD 3,125 NEW PHOTOS.
INSTAGRAM USERS SHARE 3,600 NEW PHOTOS.
TUMBLR BLOG OWNERS PUBLISH 27,778 NEW POSTS.
BRANDS & ORGANIZATIONS ON FACEBOOK RECEIVE 34,722 "LIKES."
APPLE RECEIVES ABOUT 47,000 APP DOWNLOADS.

Competenze

Quali sono le difficoltà percepite nel mondo Big Data

Abbiamo chiesto ai partecipanti al test di ammissione al Master in Data Intelligence e Strategie Decisionali di esprimere un grado di difficoltà (10 massimo, 1 minimo) su alcuni punti o criticità dei Big Data e Analytics sulla base della propria esperienza.

I punti attenzionati sono riportati nelle slide successive

Questionari 1/4

- Qualità dei dati
- Definizione-strutturazione del problema
- Adeguatezza dei modelli di analisi
- Dati insufficienti per lo specifico problema
- Troppi dati da gestire
- Troppi dati da analizzare

Questionari 2/4

- Funzionalità di Strumenti-Piattaforme utilizzati
- Correttezza approccio metodologico
- Competenze delle risorse umane
- Processi interni all'organizzazione
- Rappresentazione (visualizzazione) dei risultati

Questionari 3/4

- Troppe variabili descrivono il problema
- Mancanza di collegamenti espliciti tra i dati
- Troppe soluzioni da valutare
- Modelli matematici non adeguati al problema
- Modelli matematici alimentati con dati non adeguati o corretti
- Complessità computazionale per trovare/ valutare le soluzioni

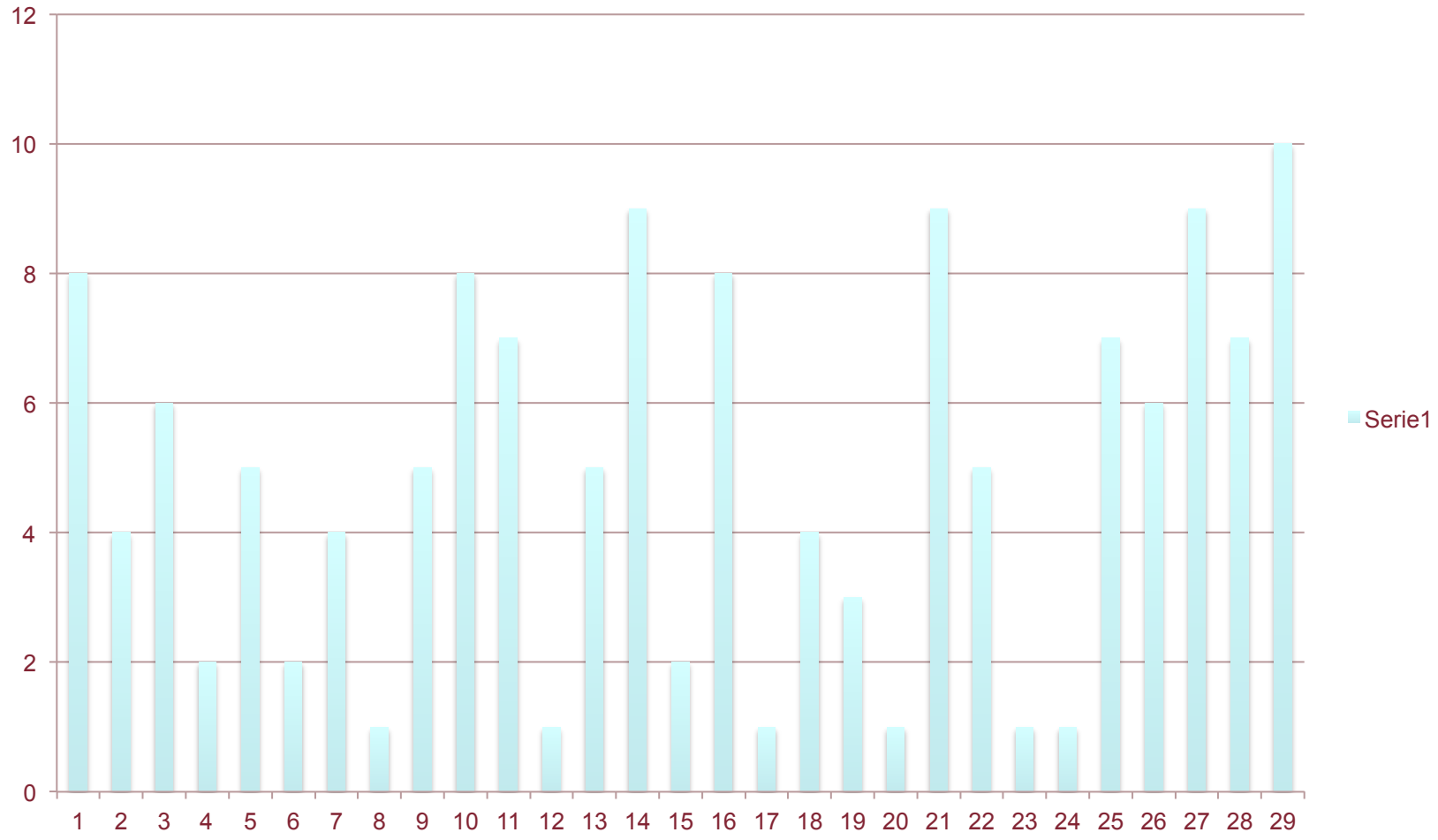
Questionari 4/4

- Uso non corretto degli strumenti
- Comunicazioni errate o poco chiare all’interno del team
- Comunicazioni errate o poco chiare con i livelli superiori
- Comunicazioni errate o poco chiare con i livelli operativi

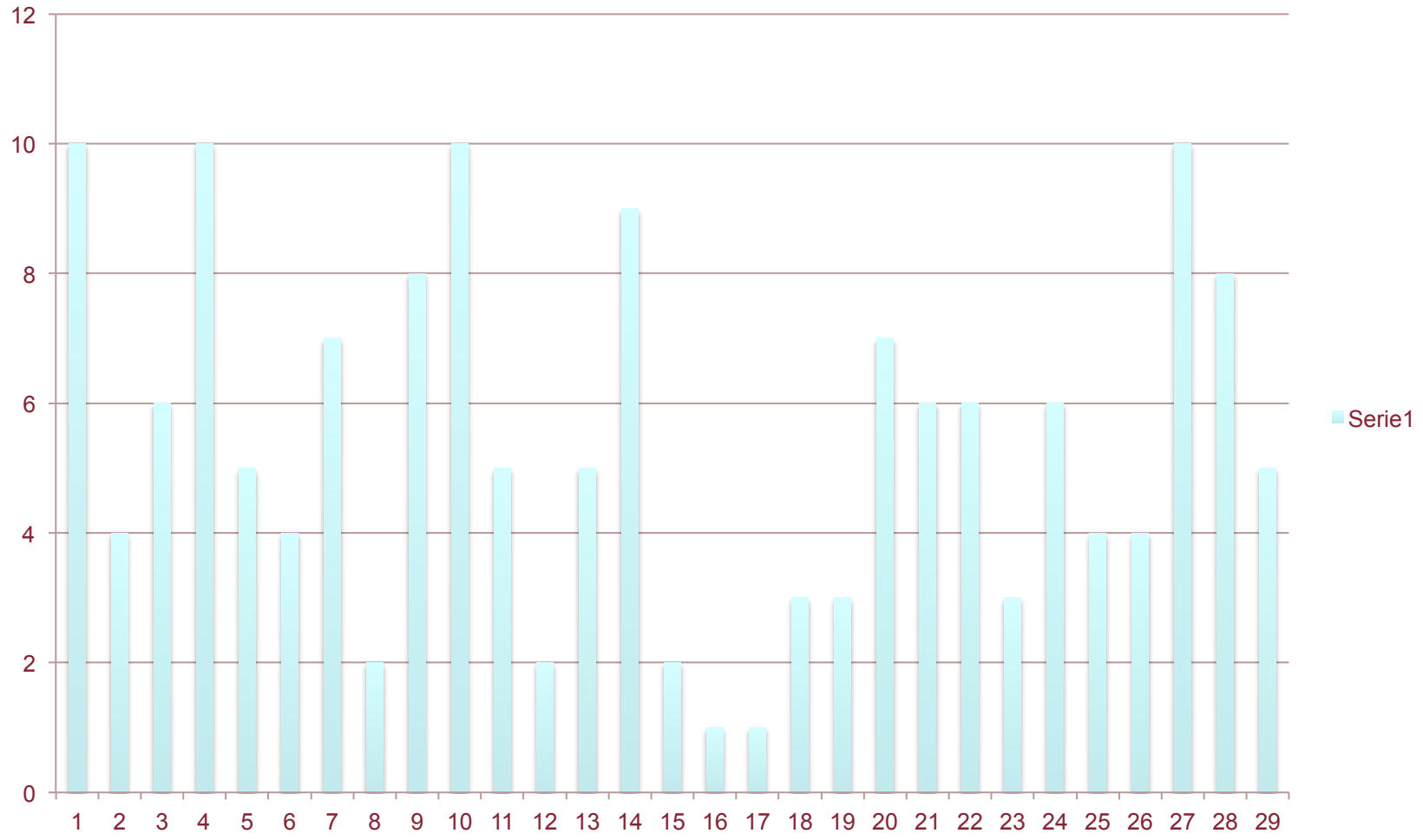
Risposte

DOMANDA	MEDIA	MEDIANA	VAR
Qualità dei dati	7,90	8	3,74
Definizione-strutturazione del problema	8,69	9	3,94
Adeguatezza dei modelli di analisi	8,72	9	1,71
Dati insufficienti per lo specifico problema	7,31	8	3,36
Troppi dati da gestire	4,86	5	8,69
Troppi dati da analizzare	4,59	5	9,11
Funzionalità di Strumenti-Piattaforme utilizzati	6,52	7	5,97
Correttezza approccio metodologico	8,41	9	2,25
Competenze delle risorse umane	7,79	8	3,38
Processi interni all'organizzazione	7,52	8	3,47
Rappresentazione (visualizzazione) dei risultati	7,14	8	4,77
Troppe variabili descrivono il problema	5,69	6	3,51
Mancanza di collegamenti espliciti tra i dati	5,72	6	5,28
Troppe soluzioni da valutare	5,38	5	7,74
Modelli matematici non adeguati al problema	6,66	7	5,95
Modelli matematici alimentati con dati non adeguati o corretti	7,41	8	5,82
Complessità computazionale per trovare/valutare le soluzioni	6,52	7	5,83
Uso non corretto degli strumenti	7,61	7,5	4,69
Comunicazioni errate o poco chiare all'interno del team	8,21	9	4,47
Comunicazioni errate o poco chiare con i livelli superiori	8,07	8	3,99
Comunicazioni errate o poco chiare con i livelli operativi	8,19	8	2,72

Troppi dati da gestire



Troppe soluzioni da valutare



Tecnologie

Progetto Grandi Attezzature TeraDrive - Sapienza

TeraDrive, la prima piattaforma disponibile in Sapienza per la gestione e l'analisi di Big Data

•Spazio di storage dell'ordine dei 256 Terabyte, espandibile sino ad oltre 1 Petabyte.

•Infiniband basata su fibra ottica a bassa latenza.

Progetto Grandi Attezzature TeraStat - Sapienza

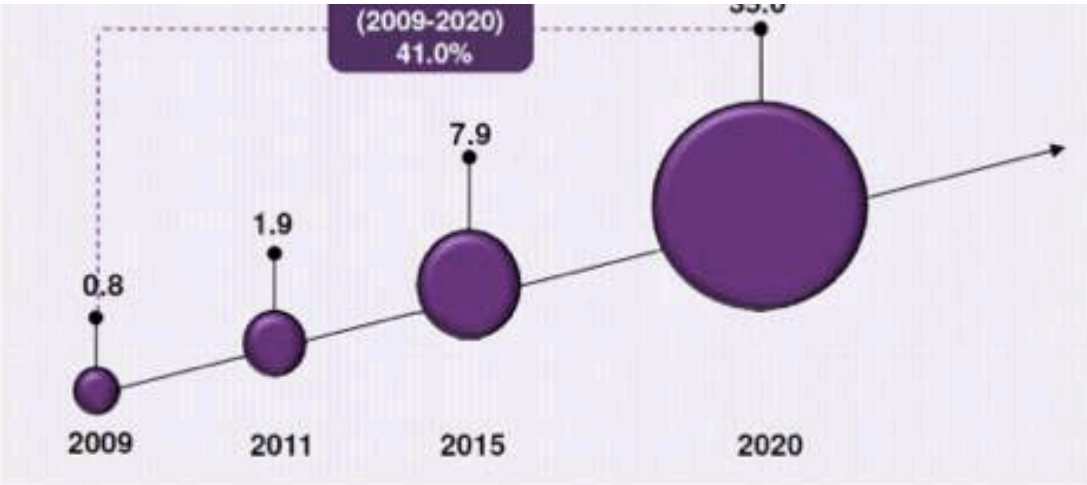
1. Prevalenza e incidenza di sclerosi multipla ad elevata risoluzione spaziale
2. Sistemi di raccomandazione su larga scala
3. Analisi della mobilità delle persone tramite Big Floating Car Data
4. Multi Objective Optimization e Big Data per la Sicurezza nei Trasporti Pubblici Locali
5. Gestione del Rischio nei mercati finanziari: nuove sfide
6. Small area estimation: measurement error, benchmarking and record linkage

Big and Small

Se non BIG … Analytics

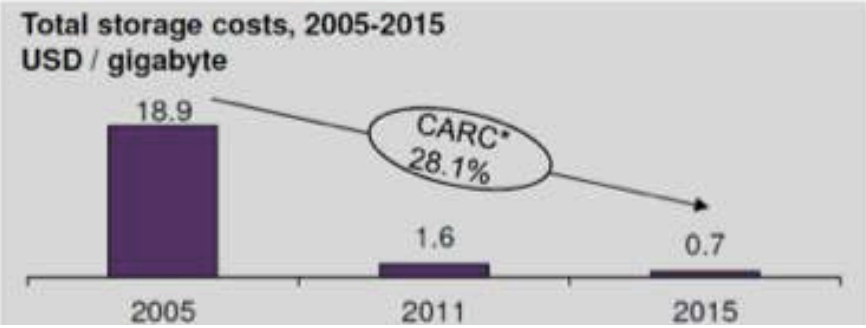
- La prevalenza dei problemi comunemente denotati come “Big Data” può essere convenientemente risolto utilizzando strumenti tradizionali

Stock di informazioni



- Need for large storage capacity
- Need for quick retrieval of data
- Enable informed decision-making effectively, leveraging large data sets. For e.g.,:
  - Turn 12 TB of Tweets created each day into improved product sentiment analysis
  - Convert 350 billion annual meter readings to better predict power consumption

Key drivers for deployment of larger datasets



- Creation of data from multiple sources and touch-points
- Distributed storage techniques and cloud computing enabling organizations to store large amount of data at lower costs
- Emergence of innovative open source software and architecture such as Hadoop Distributed File System and MapReduce
- Roll-out of 100G Ethernet cables for fast information retrieval

Stock di informazioni

•A typical PC might have had 10 gigabytes of storage in 2000.

•Today, Facebook ingests 500 terabytes of new data every day.

•Boeing 737 will generate 240 terabytes of flight data during a single flight across the US.

• The smart phones, the data they create and consume; sensors embedded into everyday objects will soon result in billions of new, constantly-updated data feeds containing environmental, location, and other information, including video.

Cambio di prospettive

- Stiamo quindi passando da un'analisi basata sulla ricerca del "**Perché**", della causa di un evento, ad un'analisi basata sulla rilevazione del "**Cosa**" stia accadendo (dalla causalità alla correlazione).
- Un altro cambio di prospettiva (Mayer-Schonberger V. & Cukier K.), è riferito ai profili professionali: nel mondo dei Big Data si attenua l'importanza degli specialisti, non c'è più la rigida ripartizione per competenze. Oggi si richiedono **conoscenze diversificate** ed accumulo di esperienze su tematiche eterogenee.

Small Data vs. Big Data

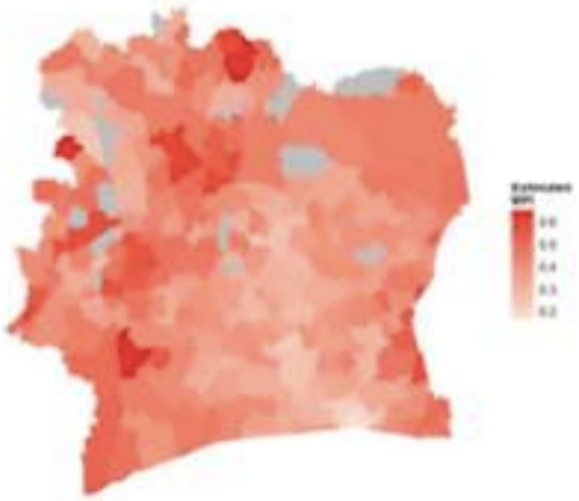
In un mondo "**Small Data**" le decisioni venivano prese sulla base di informazioni che erano:

- Limitate
- Esatte
- Di natura causale

In un mondo "**Big Data**"

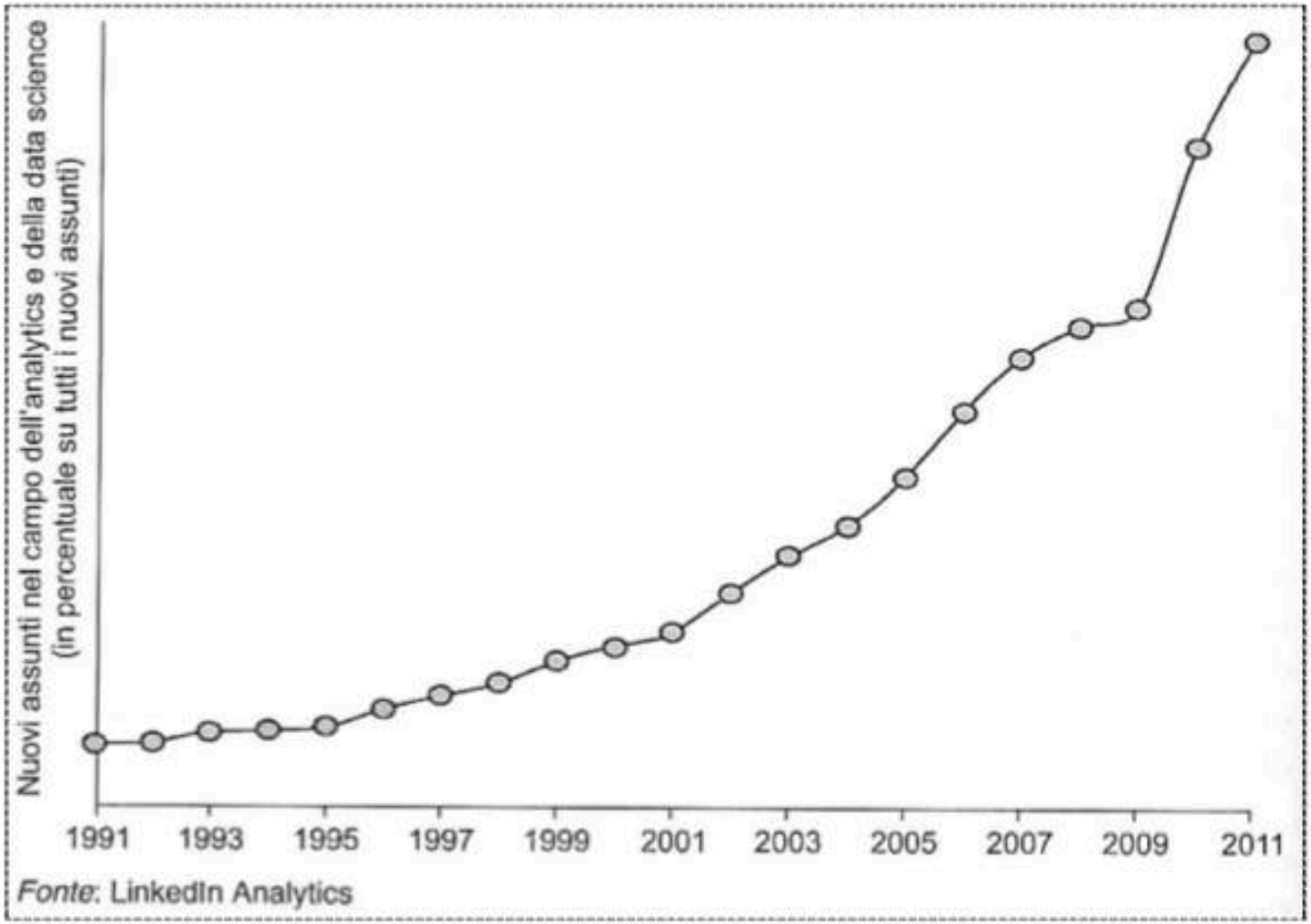
- Poco campionamento
- Rinunciare all'esattezza per la comprensione del livello macro del fenomeno
- No causalità ed analizzare correlazioni (sono più facili e soprattutto più rapide da scoprire e certamente più economiche).

Small Data vs. Big Data



Poverty map in Côte d'Ivoire
estimated by mobile phone data

Richieste di professionalità



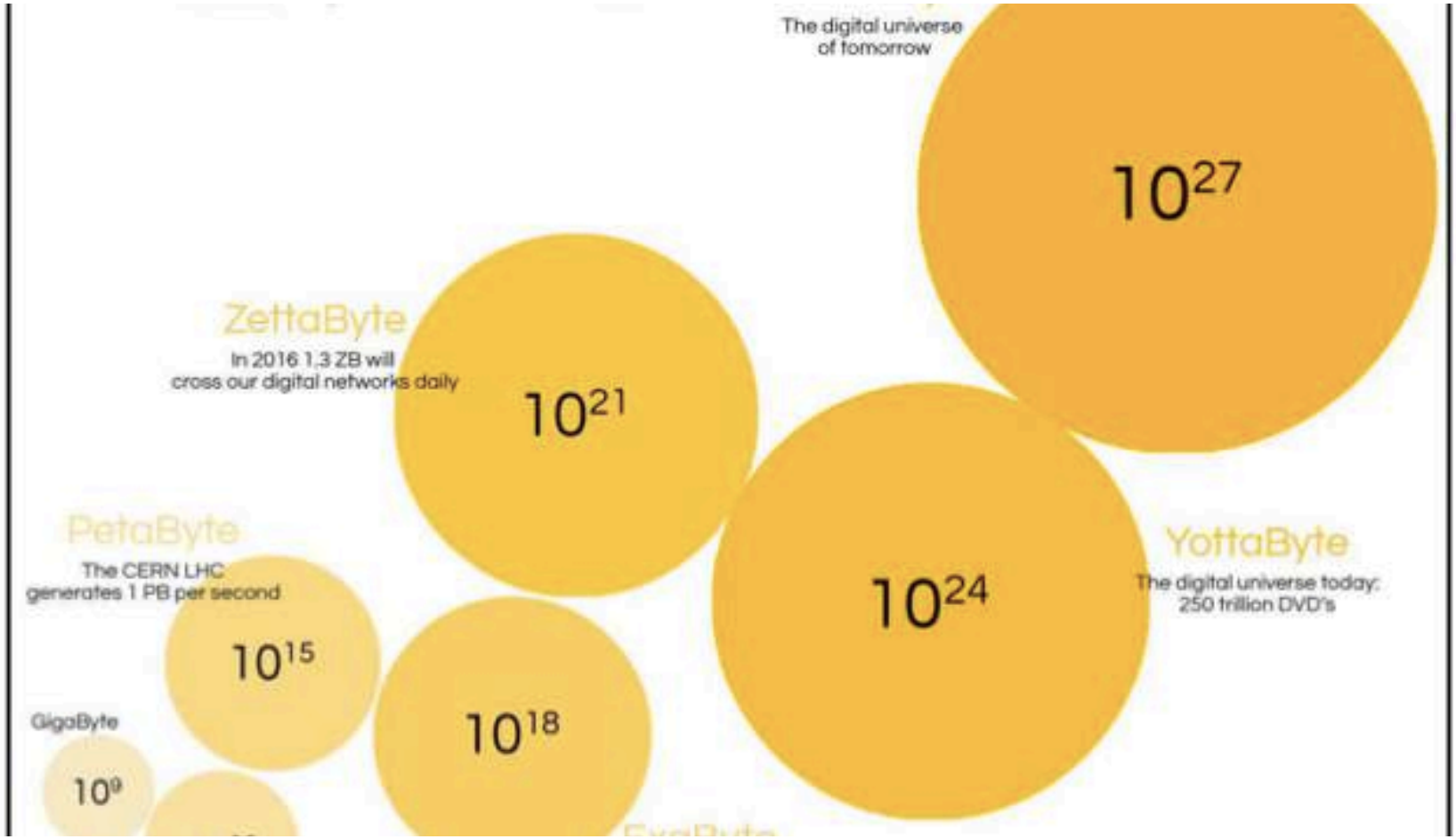
Hype Cycle 2015

Cosa si intende per Big Data (2011-14)

- **Roger Magoulas della O'Reilly Media**
- **McKinsey Global Institute (2011)**
- **Commissione Europea**
- **Gartner**
- **SDA Bocconi**

Caratteristiche

Volume (esempio): Internet of Things



Caratteristiche - Varietà

- **Dati Strutturati (es. tabelle) 5%**
- **Dati non Strutturati (es. audio e video)**
- **Dati Semi strutturati (es. XML)**

Tipi di dati

Structured Data

- Resides in formal data stores – RDBMS and Data Warehouse; grouped in the form of rows or columns
- Accounts for ~10% of the total data existing currently

RDBMS (e.g., ERP and CRM

Data Warehousing

Microsoft Project Plan File

Tipi di dati

Unstructured Data

- Comprises data formats which cannot be stored in row/ column format like audio files, video, clickstream data,
- Accounts for ~80% of the total data existing currently

Video

Audio

Text message

Blogs

Weather patterns

Location co-ordinates

Web logs & clickstreams

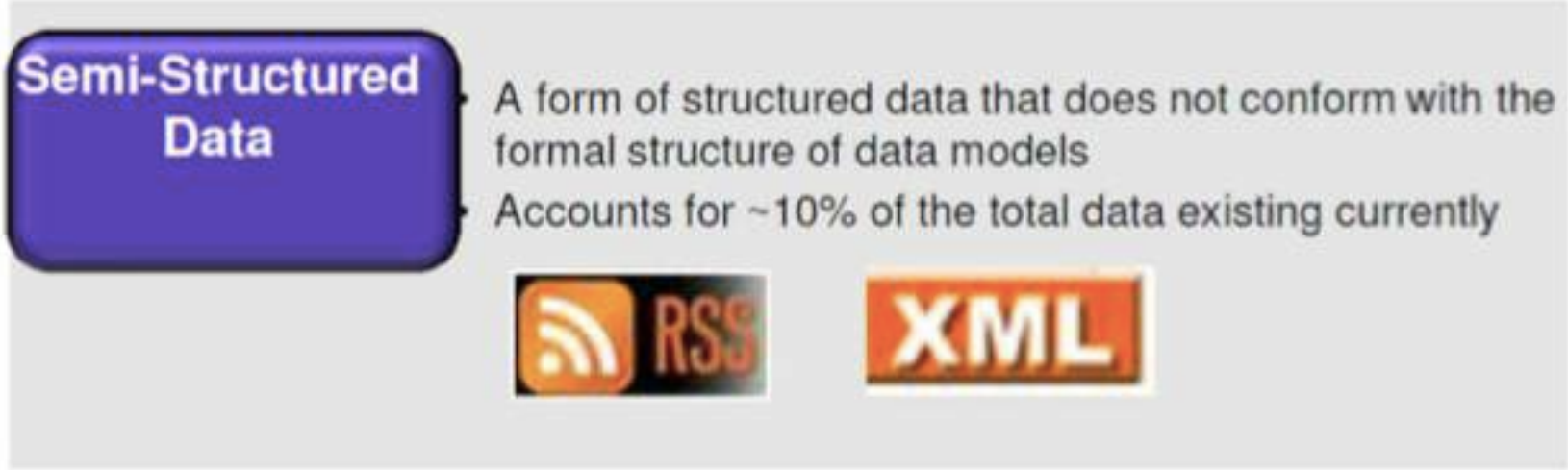
Sensor data/ M2M

Email

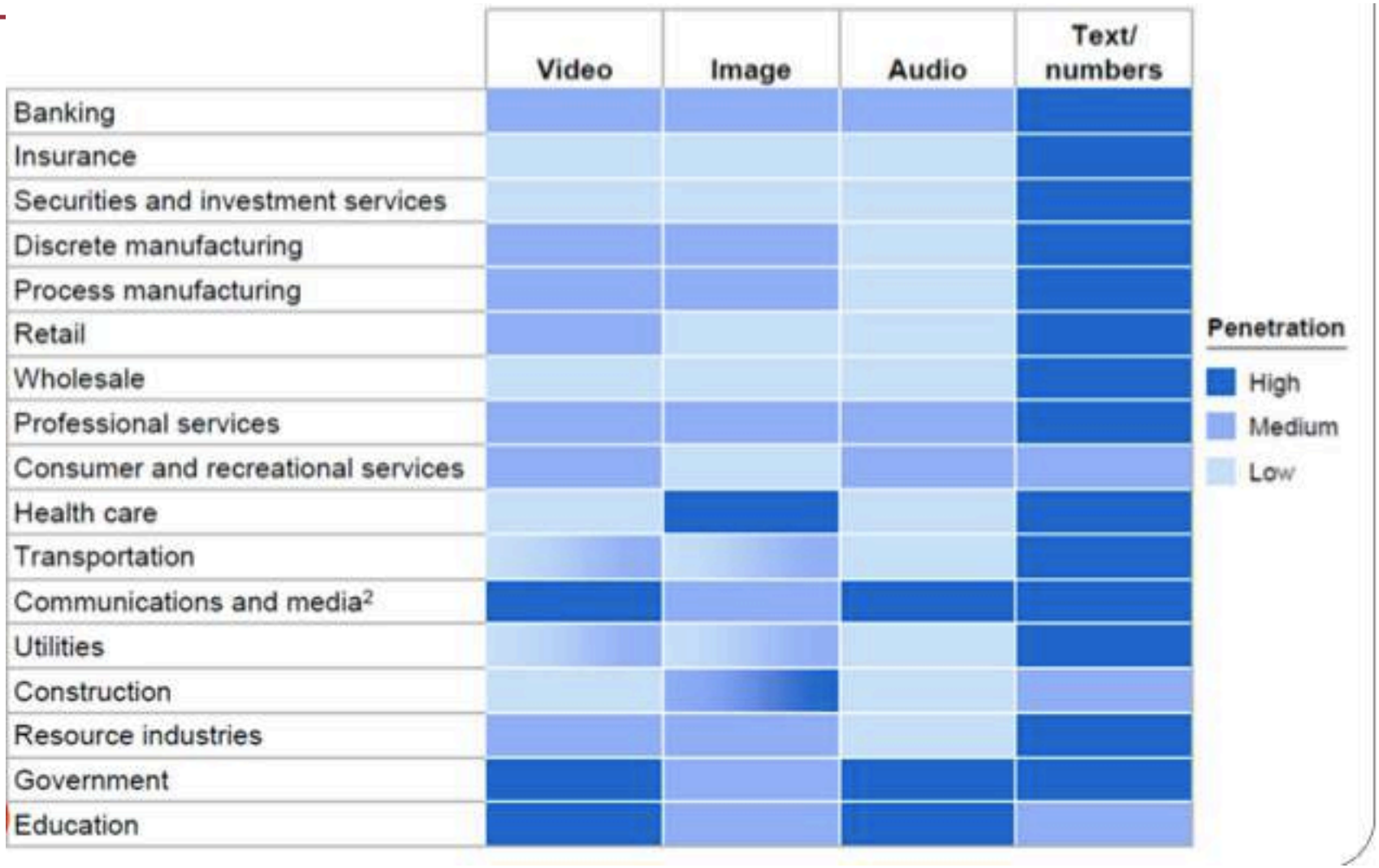
Social media

Geospatial data

Tipi di dati


Semi-Structured Data
A form of structured data that does not conform with the formal structure of data models
Accounts for ~10% of the total data existing currently
RSS
XML

Tipi di dati

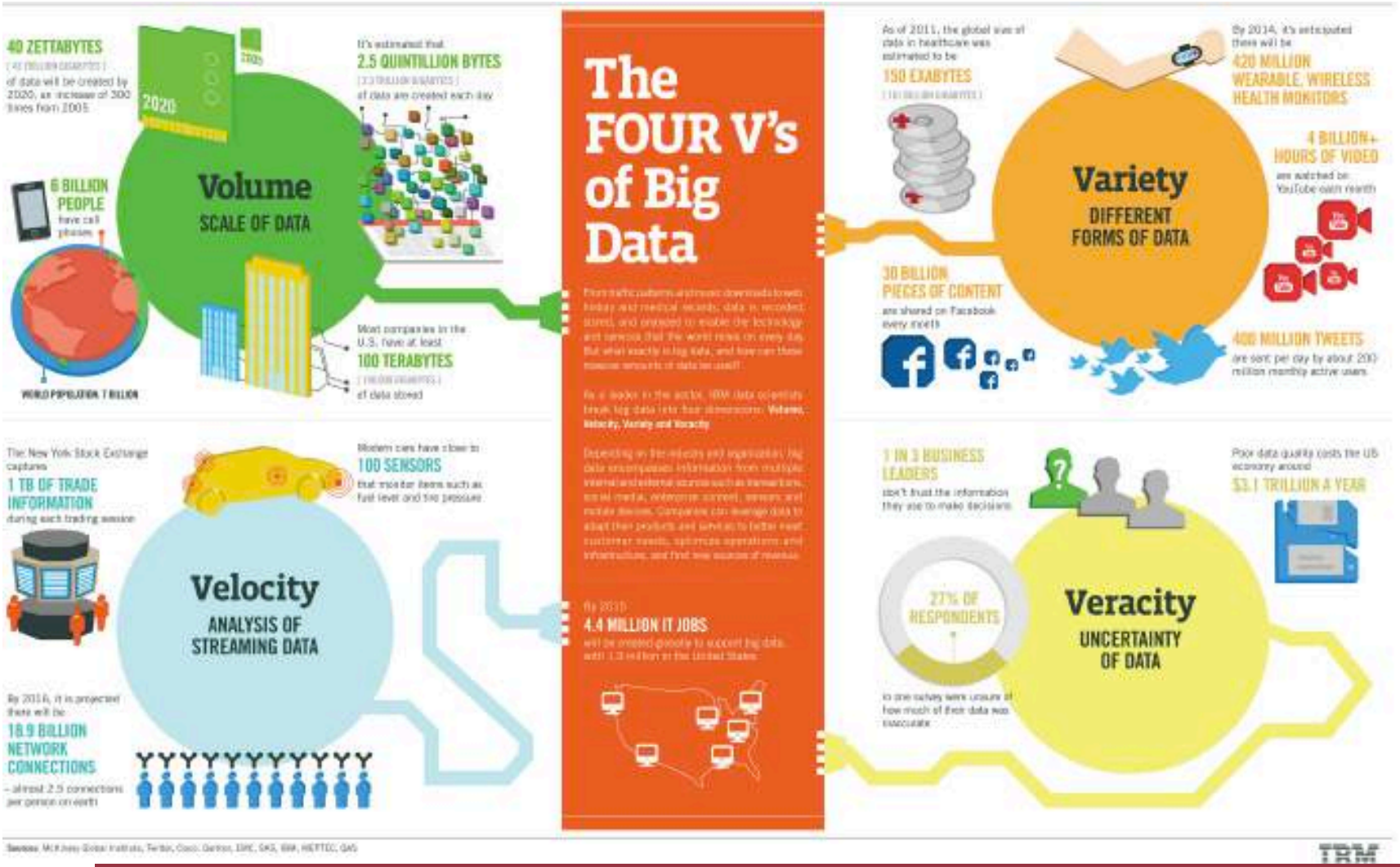

Video
Image
Audio
Text/ numbers
Banking
Insurance
Securities and investment services
Discrete manufacturing
Process manufacturing
Retail
Wholesale
Professional services
Consumer and recreational services
Health care
Transportation
Communications and media[2]
Utilities
Construction
Resource industries
Government
Education
Penetration
High
Medium
Low

Big Data: V3 + valore

Big Data: V^3+VALUE

- Volume: Gigabyte(10^9), Terabyte(10^{12}), Petabyte(10^{15}), Exabyte(10^{18}), Zettabyte (10^{21})
- Varietà: Structured, semi-structured, unstructured; Text, image, audio, video, record
- Velocità: Periodic, Near Real Time, Real Time
- Valore: Può generare grossi vantaggi competitivi!

Le “tre” V


40 ZETTABYTES
of data will be created by 2020, an increase of 300 times from 2005
2020
6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION
Volume
SCALE OF DATA
It's estimated that
2.5 QUINTILLION BYTES
of data are created each day
Most companies in the U.S. have at least
100 TERABYTES
of data stored
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session
Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure
Velocity
ANALYSIS OF STREAMING DATA
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth
The Four V's of Big Data
By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States
As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
Variety
DIFFERENT FORMS OF DATA
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
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users
1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
27% OF RESPONDENTS
Veracity
UNCERTAINTY OF DATA
IBM

Caratteristiche - Veracity

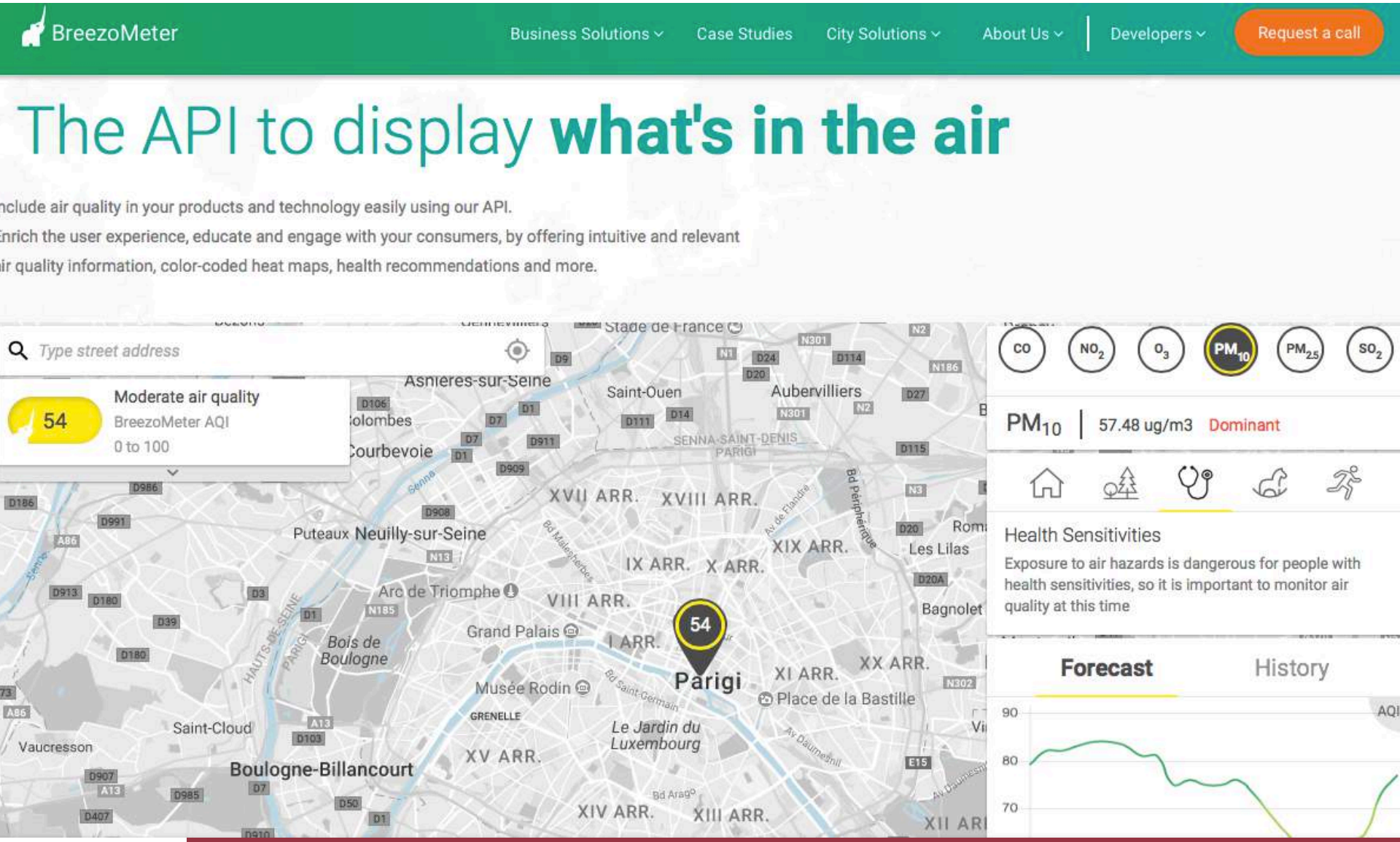
Veracity: genuinità e bontà del dataset. Questo elemento è stato introdotto da IBM, in un suo report del 2012, per porre l'accento sulla necessità di selezionare e trattare dati affidabili (esempio: i dati generati da Social Network potrebbero essere in parte inaffidabili)

Fattori abilitanti

- **Eccezionale crescita del volume dei dati accumulati**
- **Nuove “affordance” tecnologiche offerte da**
- **Affermazione di nuovi approcci teorici**

Le opportunità

Start up israeliana



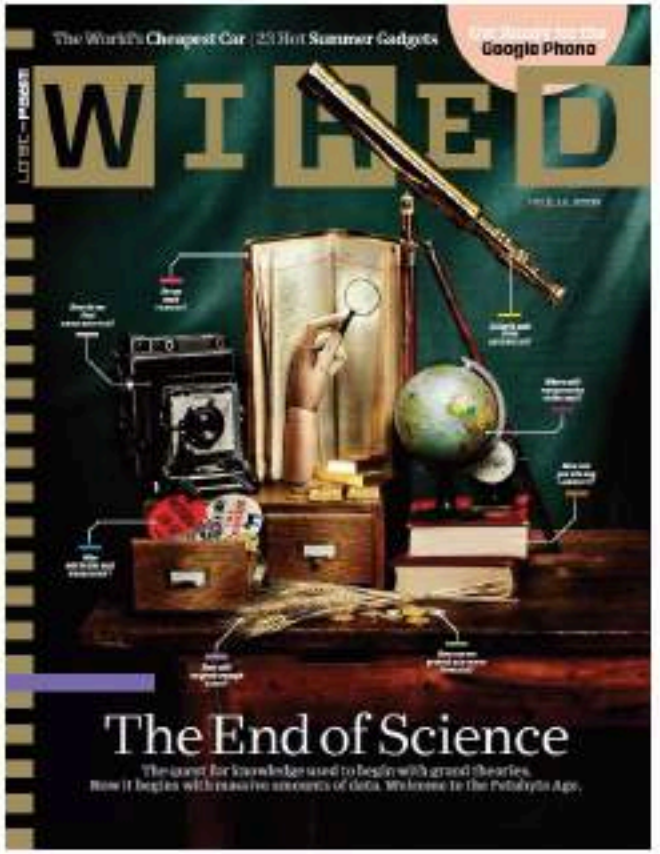
Start up israeliana (2)

AGF

Rischi

New Paradigm

"This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves." – Chris Anderson (Wired 2008)



Google Flu

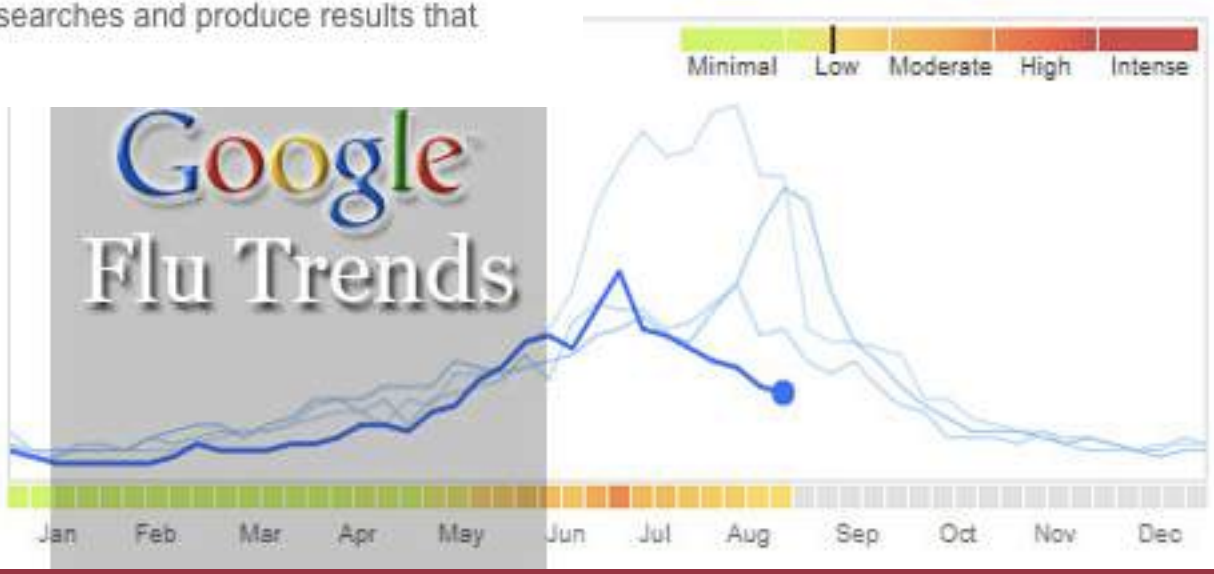
Flu outbreaks could be monitored by internet searches

Agree 2 Disagree 0

By Hicbd

Mon Nov 26 2012 11:08 am

The Centers for Disease Control and Prevention (CDC) can take up to two weeks to collect and release survey data about where and when the flu is spreading around the country. As shown by Google's Flu Trends, internet search tools could use real-time data to identify regions around the world experiencing a high number of flu-related searches and produce results that correlate closely with CDC data.



Failing of Big Data

GEORGES GOBET/AFP/Getty Images

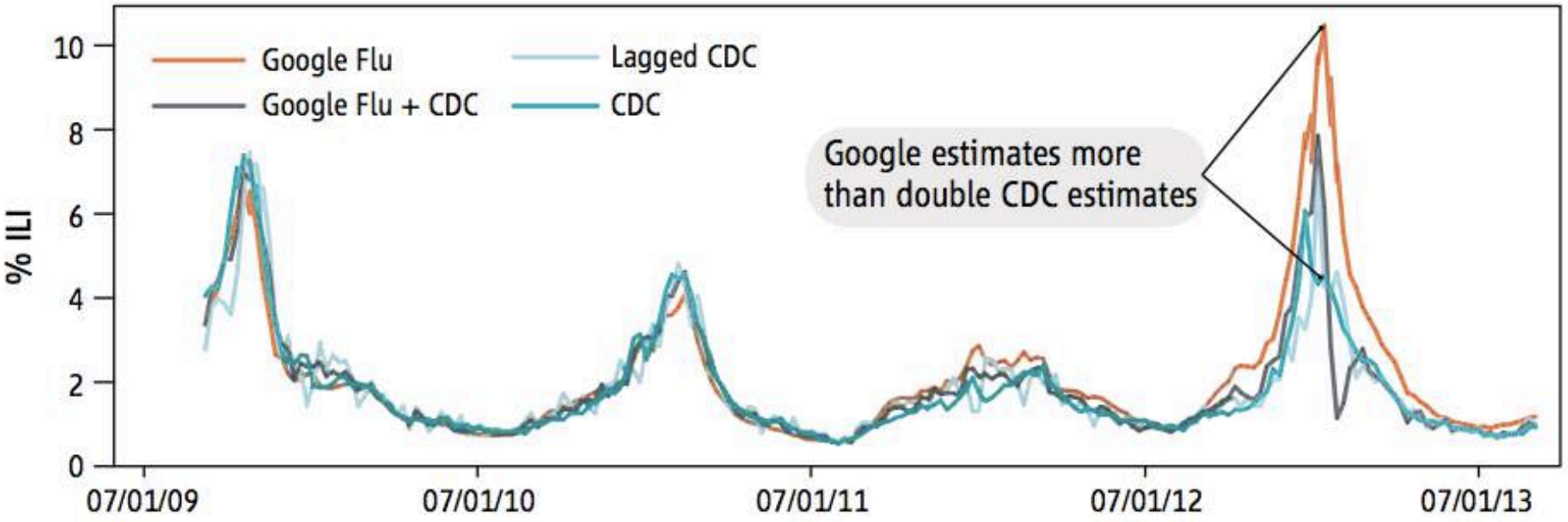
BIG DATA

Google's Flu Project Shows the Failings of Big Data

DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD

The Parable of Google Flu: Traps in Big Data Analysis

Overestimation



Grazie per l'attenzione !

paolo.dellolmo@uniroma1.it