I Big Data
Inforav 4 Ottobre 2013

Carlo Batini
Universita' di Milano Bicocca
batini@disco.unimib.it

Indice

1. Una definizione e un esempio di Big Data
2. Le 3 (+2) V che caratterizzano i Big Data
3. Tecniche e Tecnologie per Big Data
4. Il valore dei Big data nelle amministrazioni pubbliche
5. Aspetti metodologici nelle statistiche pubbliche connessi ai Big Data

Appendici

- How Big are Data?
- Myths around Big Data
- To Hadoop or not To Hadoop?

1. Una definizione di Big Data

Definizione di Big Data (Fonte: O' Really)

- *"Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the structures of your database architectures. To gain value from this data, you must choose an alternative way to process it.*
- The definition can vary by sector, depending on what kinds of software tools are commonly available and what sizes of datasets are common in a particular industry.
- With those caveats, big data in many sectors today will range from a few dozen terabytes to multiple petabytes (thousands of terabytes).

Valutazioni contrastanti

- Molti analisti, es. McKinsey "Big data is the next frontier for innovation, competition, and productivity"

- Nate Silver (statistico che ha indovinato I risultati elezioni USA in 49/50 stati): "Big Data e' uno slogan di marketing, una bolla"
- "Insieme alle informazioni, aumentano solo i rumori e gli errori, **la verita' resta la stessa**"

da La Lettura, Corriere della Sera, 29 settembre

Una classificazione dei tipi di big data da Istat 2013

Human-sourced information (People to People)

- Social Networks (Facebook, Twitter, LinkedIn, etc.);
- Blogs and comments;
- Internet Searches on search engines (Google, etc.);
- Videos loaded in the Internet (Youtube, etc.);
- User-generated maps;
- Picture archives (Instagram, Flickr, Picasa, etc.);
- Data and contents from mobile phones (text messages, etc.);
- E-Mail;
- E-books;
- Personal documents.

Process-mediated data (People to Machine)

- Data produced by Public Bodies and Institutions (medical records, etc.);
- Data produced by the Private sector (commercial transactions, banking/stock records, e-commerce, credit cards, etc.).

Machine-generated data (Machine to Machine; Internet of Things)

- Data from sensors
  - fixed sensors (home automation, weather/pollution sensors, traffic sensors/webcams, scientific sensors, security/surveillance videos/images, etc.);
  - mobile sensors – for tracking (satellite images, GPS, mobile phone location, car devices, etc.);
- Data from computer systems (log files, web logs, etc.)

Le 3 (+ 3) V che caratterizzano i Big Data

Data are becoming increasingly complex: the first 3 V's of Big data

- Volume: size
- Velocity: real-time streams
- Variety
  - structure (linguistic/media)
  - semantic heterogeneity
  - uncertainty

The other 3 V's

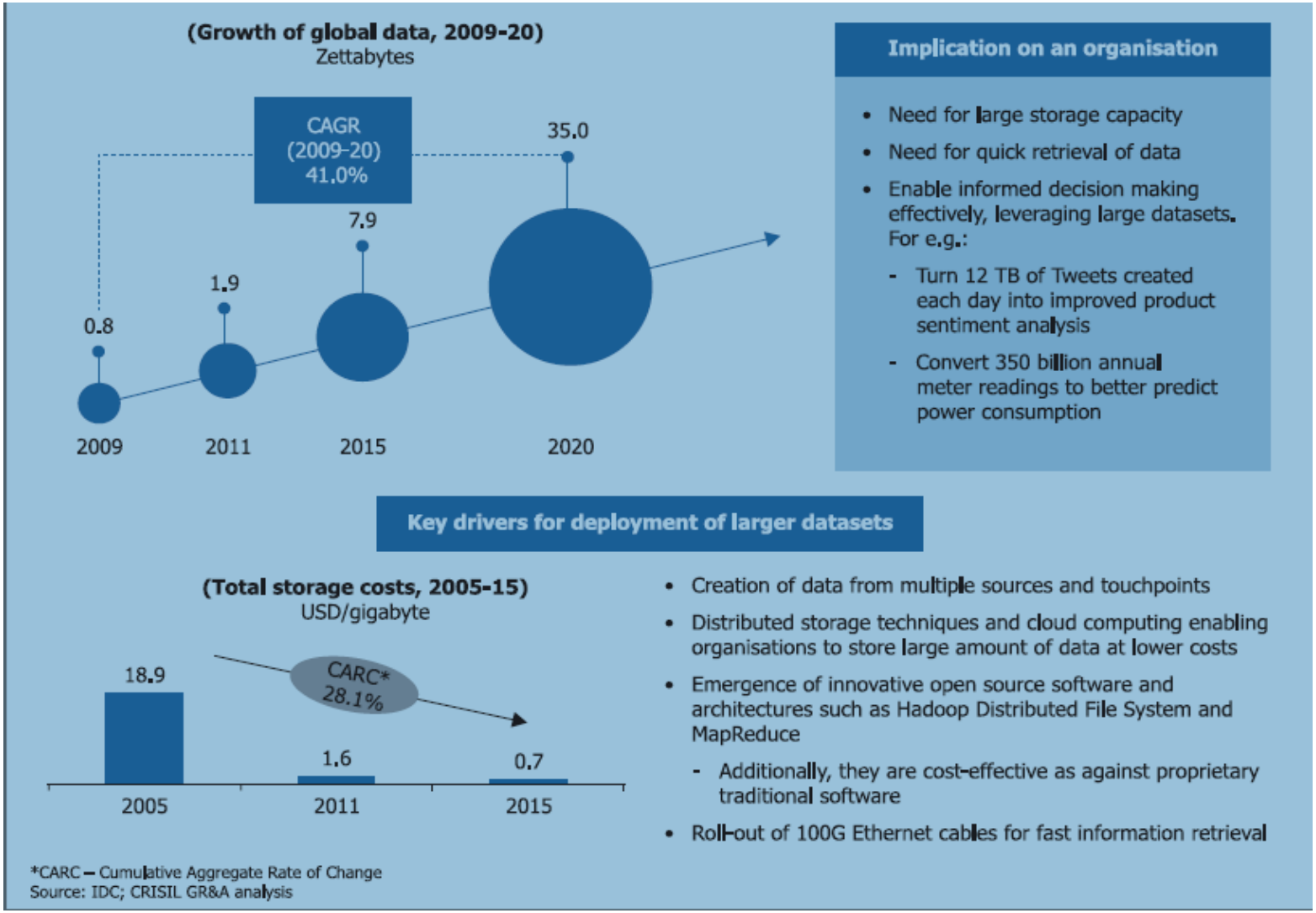
- Value → Utility
- Veracity → Quality
- Visualization → Usability

Esempio: La settimana della moda a Milano →

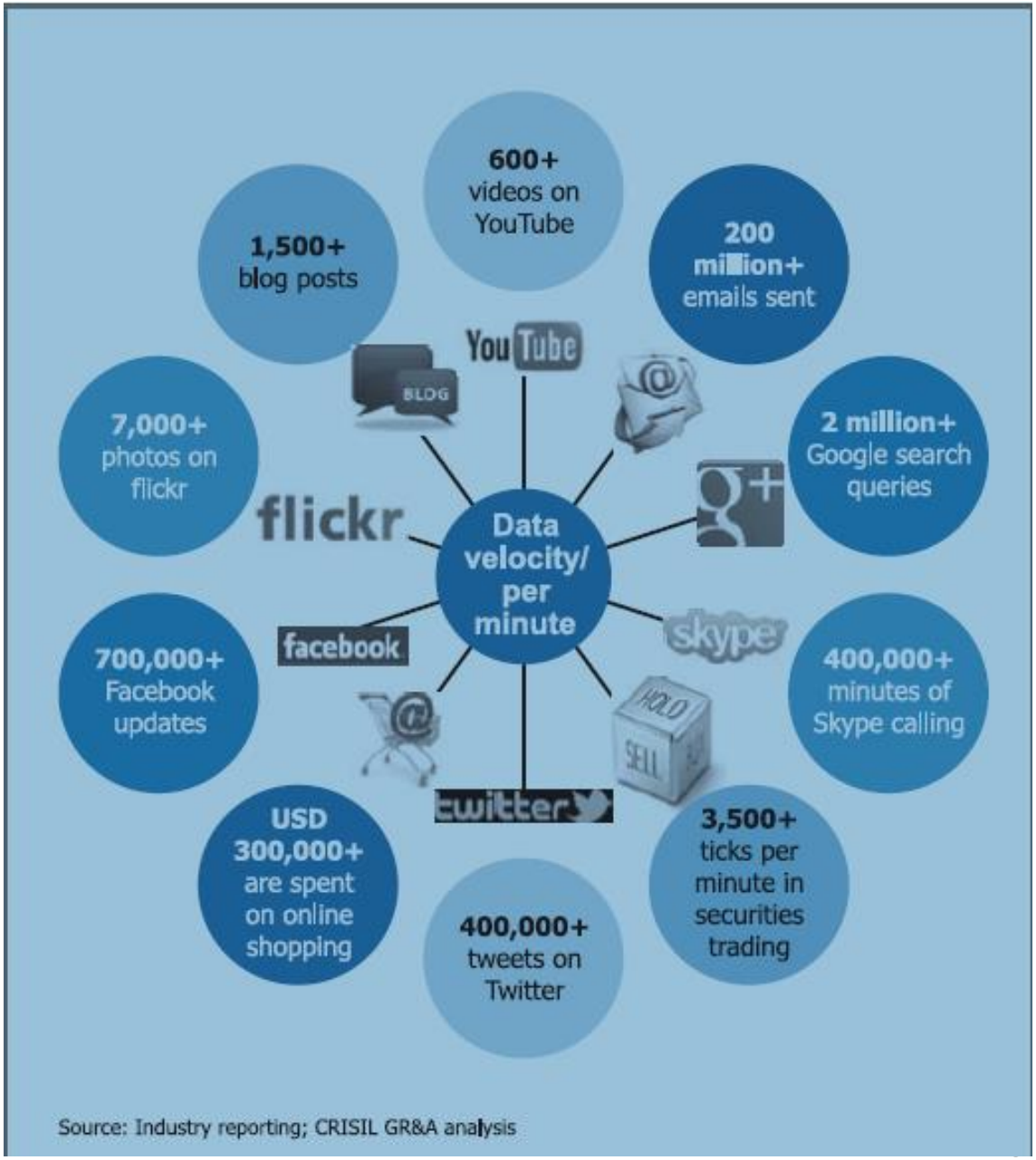
General figures
Moore's law is valid also for Big Data…

- Volume → Wordlwide information volume is growing at a rate of 60% annually (Source: Gartner)
- Velocity → 90% of data in the world today has been created during the last two years
- Variety → Individuals are generating huge amount of big data → 63% of the data (Gartner)

Volume – 2
(Fonte: NASSCOM)



Data Velocity per minute



More on Velocity

- Velocity is about the rate of change in the data and how quickly it must be used to create real value.
- Traditional technologies are especially poorly suited to storing and using high-velocity data.
- So new approaches are needed.

Variety

Variety

Place	Country	Population	Main economic activity
Portofino	Italy	700.000	Tourism

Structured data

Image

Portofino

Map

Text

Dear Laure, I try to describe the wonderful harbour of Potfino as I have seen this morning a boat is going in, other boats are along the wharf. Small pretty buildings and villas are looking on to the harbour.

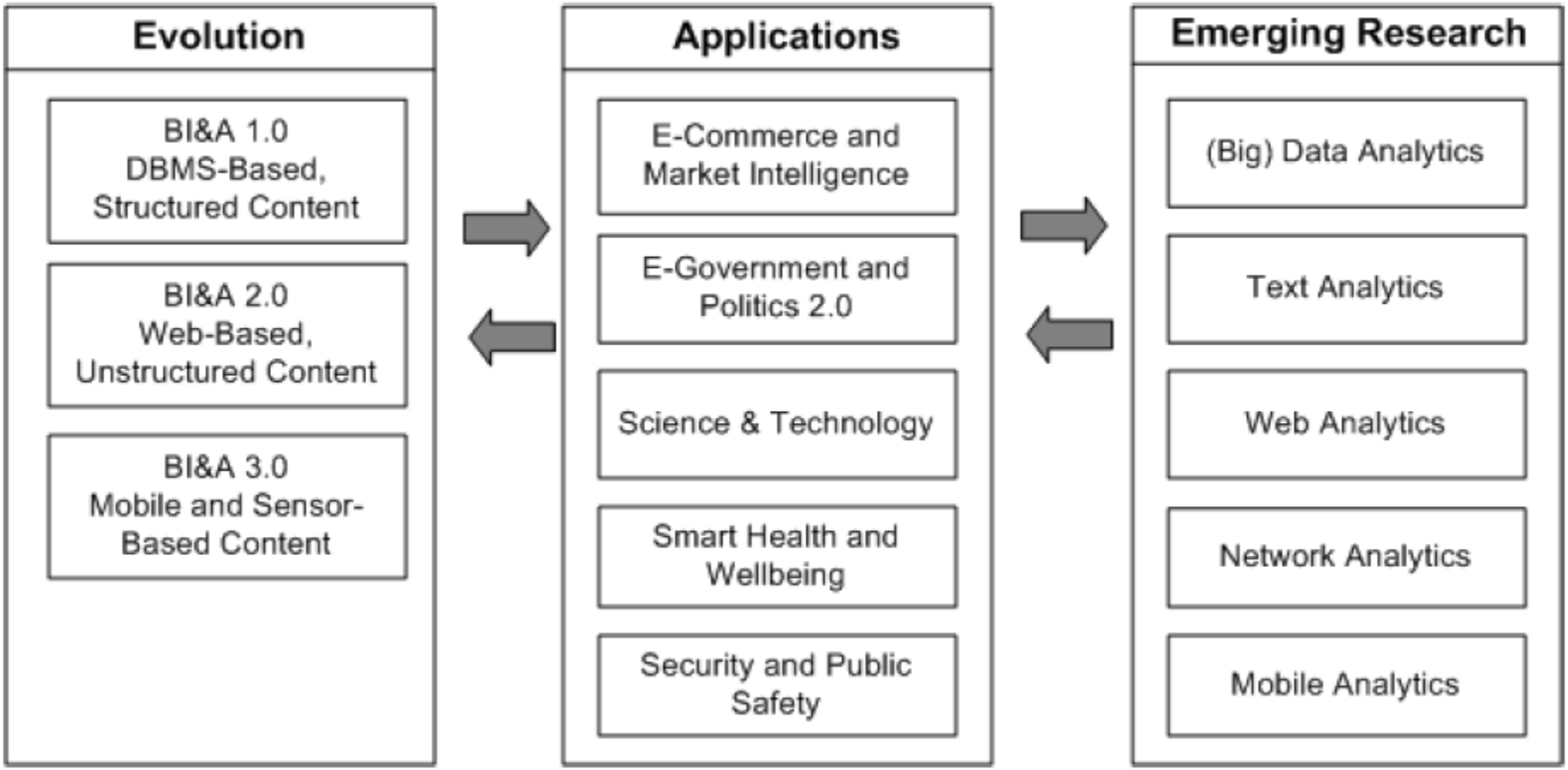
3. Tecniche e Tecnologie per i Big Data

Non solo data
(fonte M. Fatallah)

Corporate intelligence quotient =

f(data, algorithms)

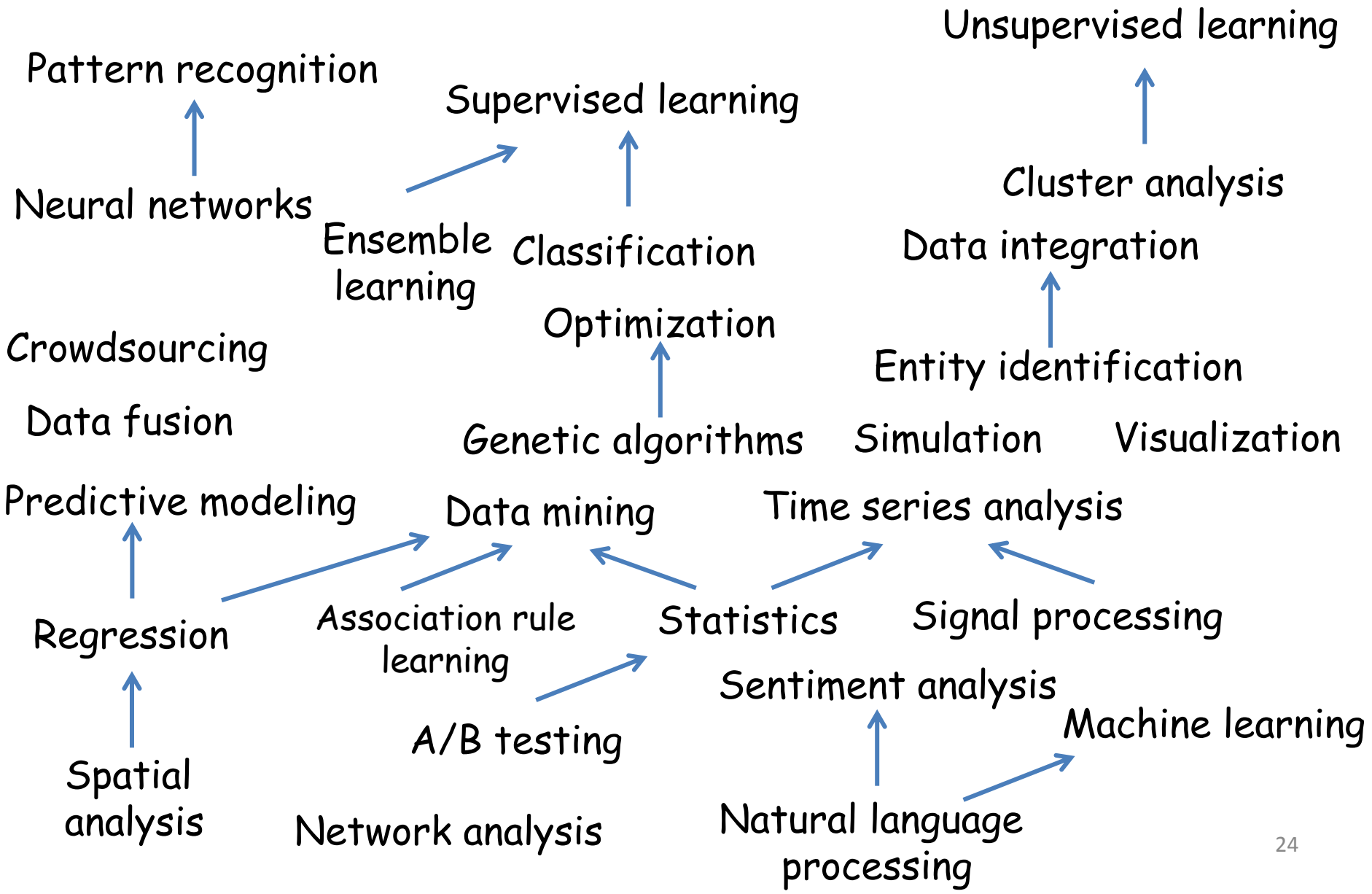
Data&Algorithms: the three phases of Business Intelligence & Analytics (fonte Chen 2012)



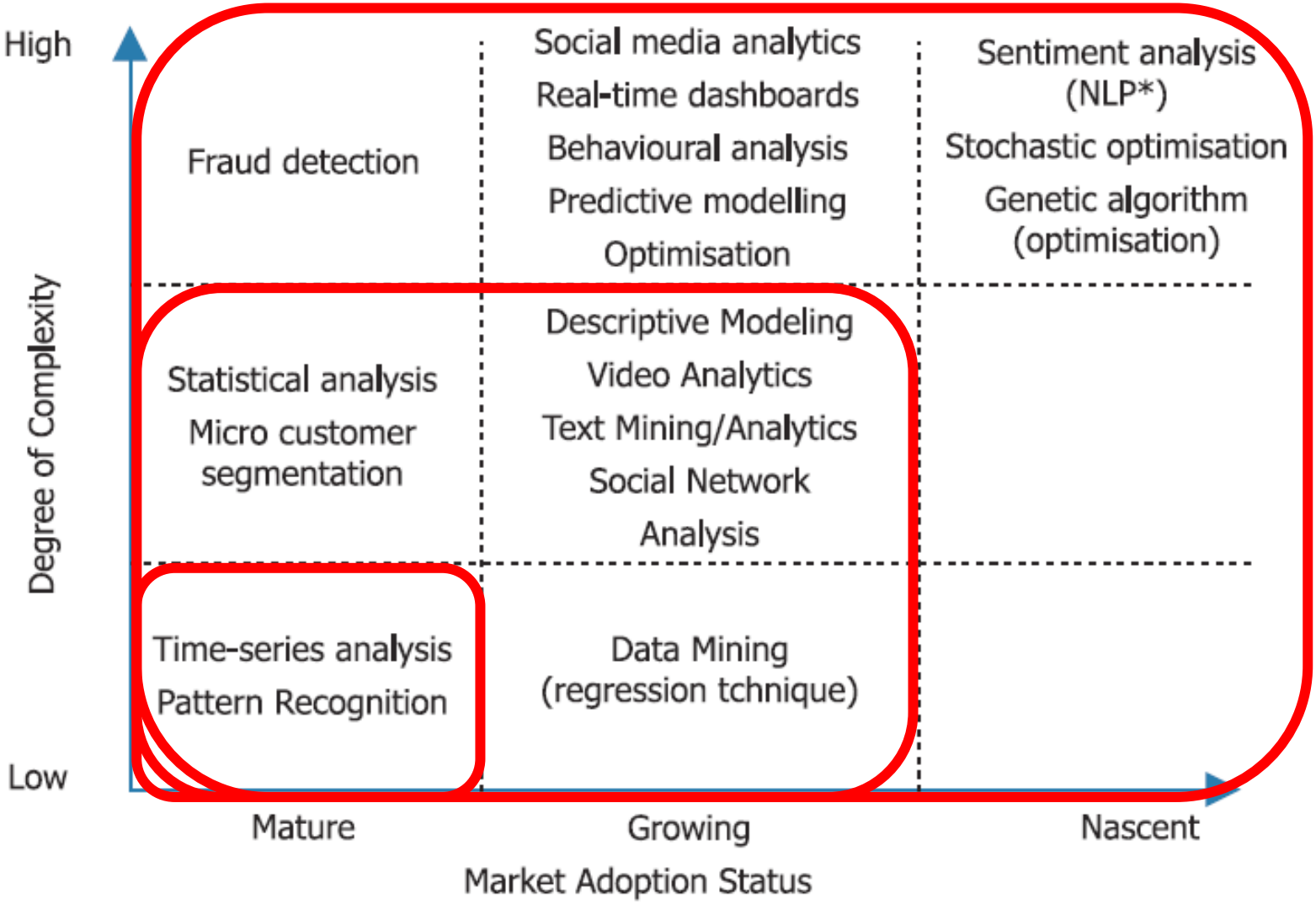
BI&A evolution (fonte Chen 2012)

	Key Characteristics	Gartner BI Platforms Core Capabilities	Gartner Hype Cycle
BI&A 1.0	DBMS-based, structured content • RDBMS & data warehousing • ETL & OLAP • Dashboards & scorecards • Data mining & statistical analysis	• *Ad hoc* query & search-based BI • Reporting, dashboards & scorecards • OLAP • Interactive visualization • Predictive modeling & data mining	• Column-based DBMS • In-memory DBMS • Real-time decision • Data mining workbenches
BI&A 2.0	Web-based, unstructured content • Information retrieval and extraction • Opinion mining • Question answering • Web analytics and web intelligence • Social media analytics • Social network analysis • Spatial-temporal analysis		• Information semantic services • Natural language question answering • Content & text analytics
BI&A 3.0	Mobile and sensor-based content • Location-aware analysis • Person-centered analysis • Context-relevant analysis • Mobile visualization & HCI		• Mobile BI

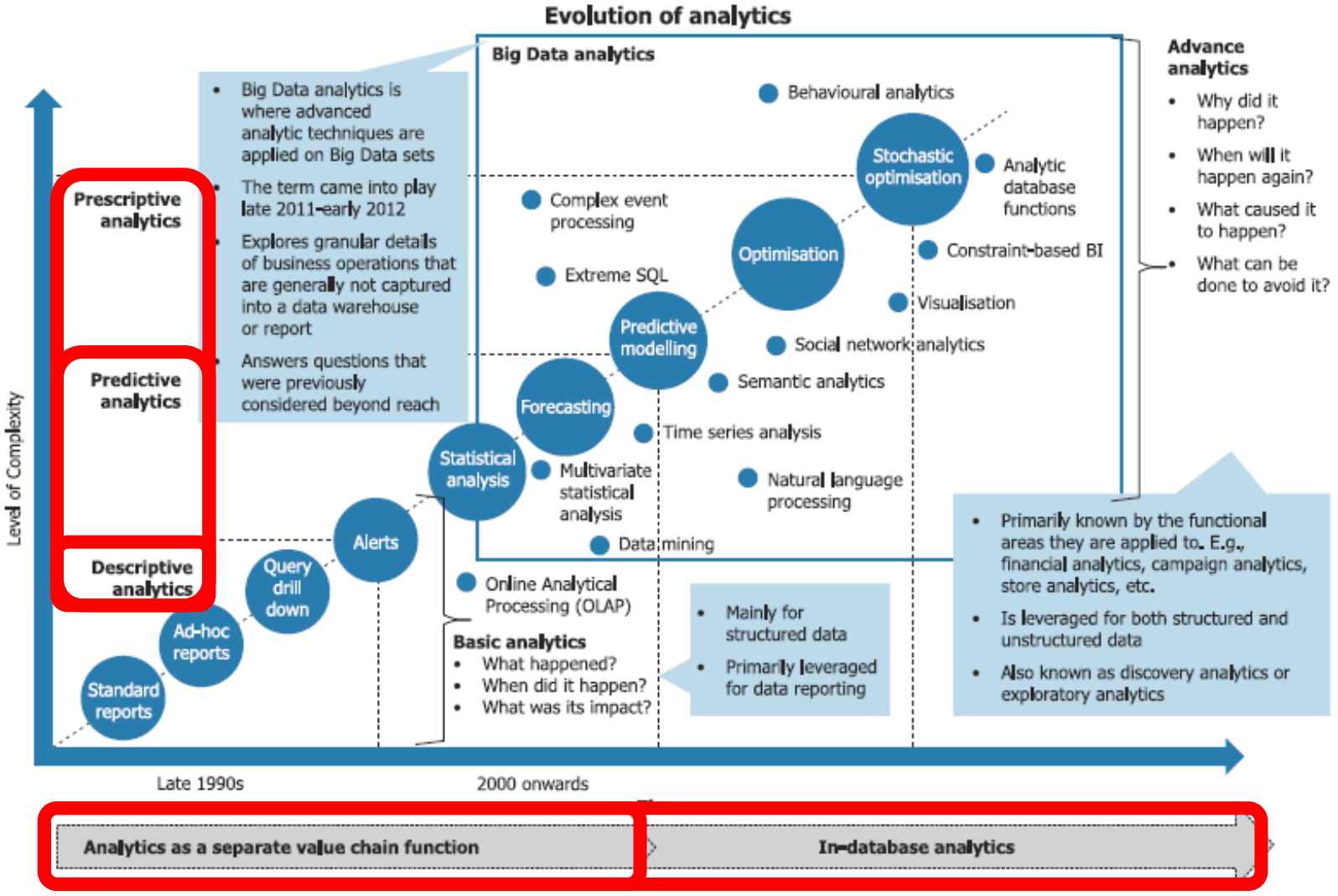
Big Data Techniques (fonte: McKinsey)



Degree of complexity of applications vs market adoption status of techniques (fonte NASSCOM)

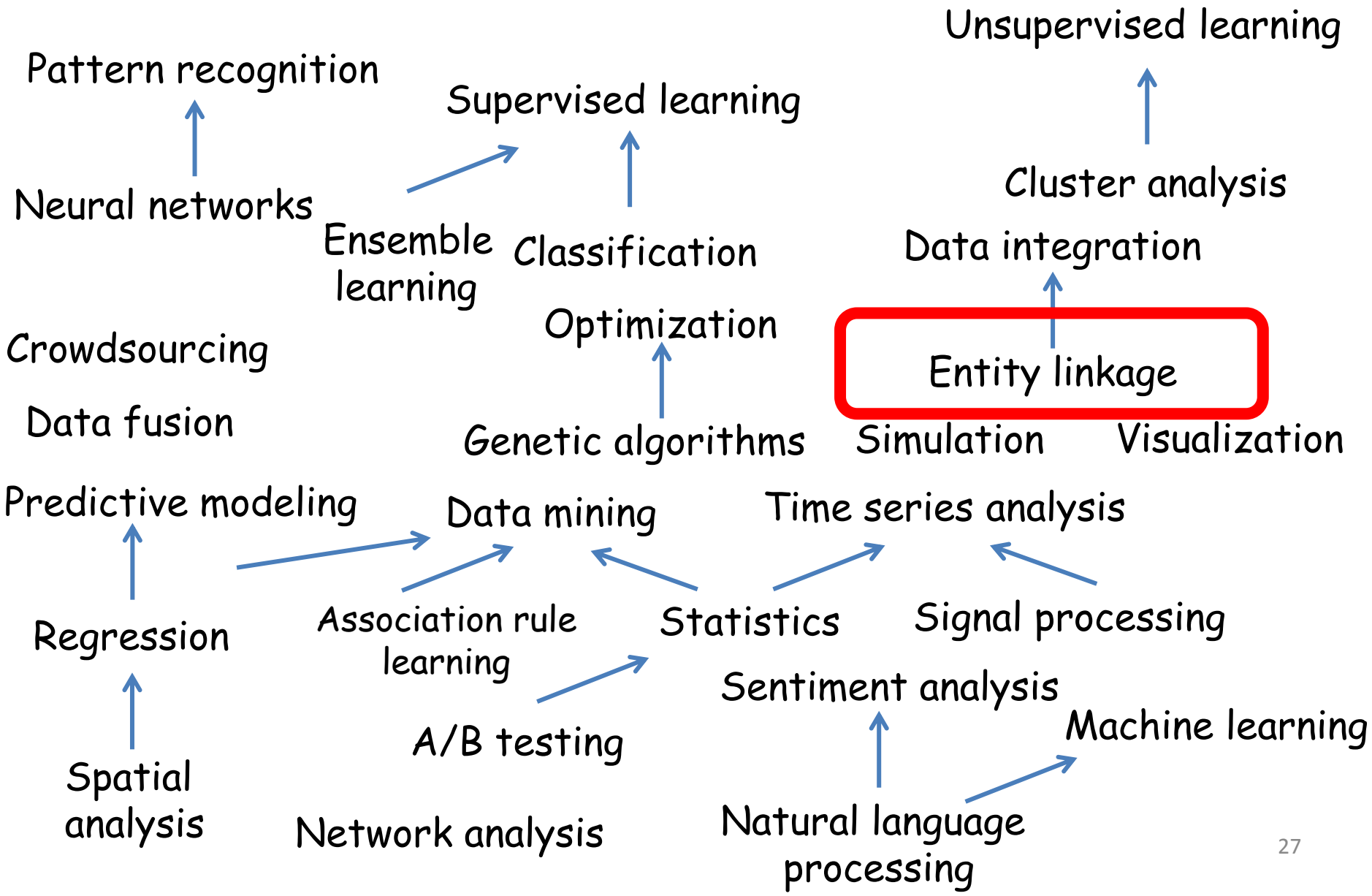


Complexity of analytics vs Evolution in time (Fonte NASSCOM)



Exponential growth in data volume, variety and velocity has facilitated the progression of analytics for better and more informed business insights

Big Data Techniques



Entity Identification

- Traditional techniques
  - Record linkage → Trovare tutti i record che fanno riferimento alla stessa entita' del mondo reale

- New techniques
  - Temporal linkage → Linkage di records con differenti time stamps
  - Statistical matching (sotto certe condizioni)

r1: Xin Dong
R. Polytechnic Institute

r2: Xin Dong
University of Washington

r3: Xin Dong
University of Washington

r4: Xin Luna Dong
University of Washington

r5: Xin Luna Dong
AT&T Labs-Research

r6: Xin Luna Dong
AT&T Labs-Research

-How many authors?

-What are their authoring histories?

1991 2004 2005 2006 2007 2008 2009 2010 2011

r11: Dong Xin
Microsoft Research

r8:Dong Xin
University of Illinois

r12: Dong Xin
Microsoft Research

r9: Dong Xin
Microsoft Research

r10: Dong Xin
University of Illinois

r7: Dong Xin
University of Illinois

r1: Xin Dong
R. Polytechnic Institute

r4: Xin Luna Dong
University of Washington

r2: Xin Dong
University of Washington

r5: Xin Luna Dong
AT&T Labs-Research

r3: Xin Dong
University of Washington

r6: Xin Luna Dong
AT&T Labs-Research

-Ground Truth

3 authors

r11: Dong Xin
Microsoft Research

r8:Dong Xin
University of Illinois

r12: Dong Xin
Microsoft Research

r9: Dong Xin
Microsoft Research

r10: Dong Xin
University of Illinois

r7: Dong Xin
University of Illinois

r1: Xin Dong
R. Polytechnic Institute

r2: Xin Dong
University of Washington

r3: Xin Dong
University of Washington

r4: Xin Luna Dong
University of Washington

r5: Xin Luna Dong
AT&T Labs-Research

r6: Xin Luna Dong
AT&T Labs-Research

-Solution 1:
-requiring high value consistency

1991 2004 2005 2006 2007 2008 2009 2010 2011

5 authors
false negative

r7: Dong Xin
University of Illinois

r8:Dong Xin
University of Illinois

r9: Dong Xin
Microsoft Research

r10: Dong Xin
University of Illinois

r11: Dong Xin
Microsoft Research

r12: Dong Xin
Microsoft Research

r1: Xin Dong
R. Polytechnic Institute

r4: Xin Luna Dong
University of Washington

r2: Xin Dong
University of Washington

r5: Xin Luna Dong
AT&T Labs-Research

r3: Xin Dong
University of Washington

r6: Xin Luna Dong
AT&T Labs-Research

-Solution 2:
-matching records w. similar names

1991 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011

2 authors
false positive

r11: Dong Xin
Microsoft Research

r8:Dong Xin
University of Illinois

r12: Dong Xin
Microsoft Research

r9: Dong Xin
Microsoft Research

r10: Dong Xin
University of Illinois

r7: Dong Xin
University of Illinois

Big Data Technologies

Unstructured data

Data Warehouse

Distributed system

NoSQL

Visualization

Hadoop
MapReduce

Business intelligence

SQL

Mashup

Cassandra

Stream processing

Semi-Structured data

ETL

Metadata

Cloud computing

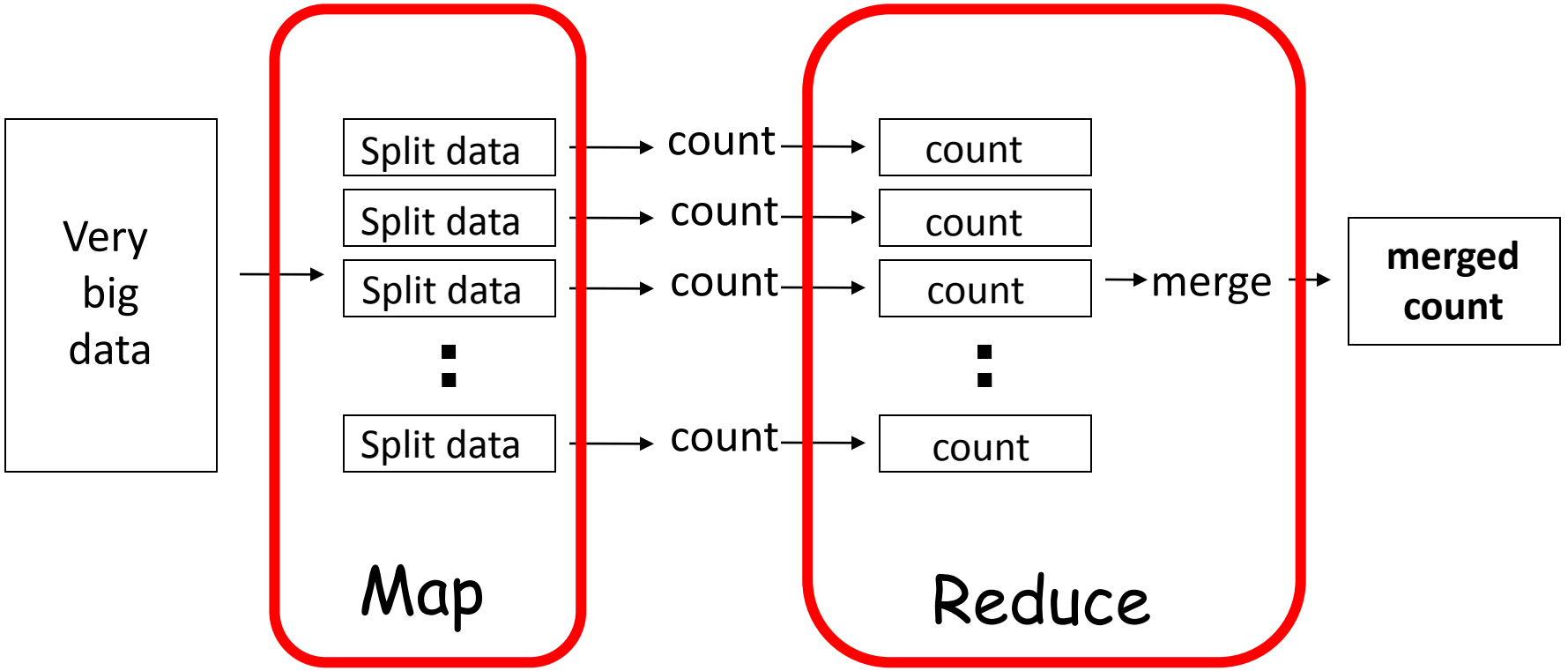
System R

NonRelational database

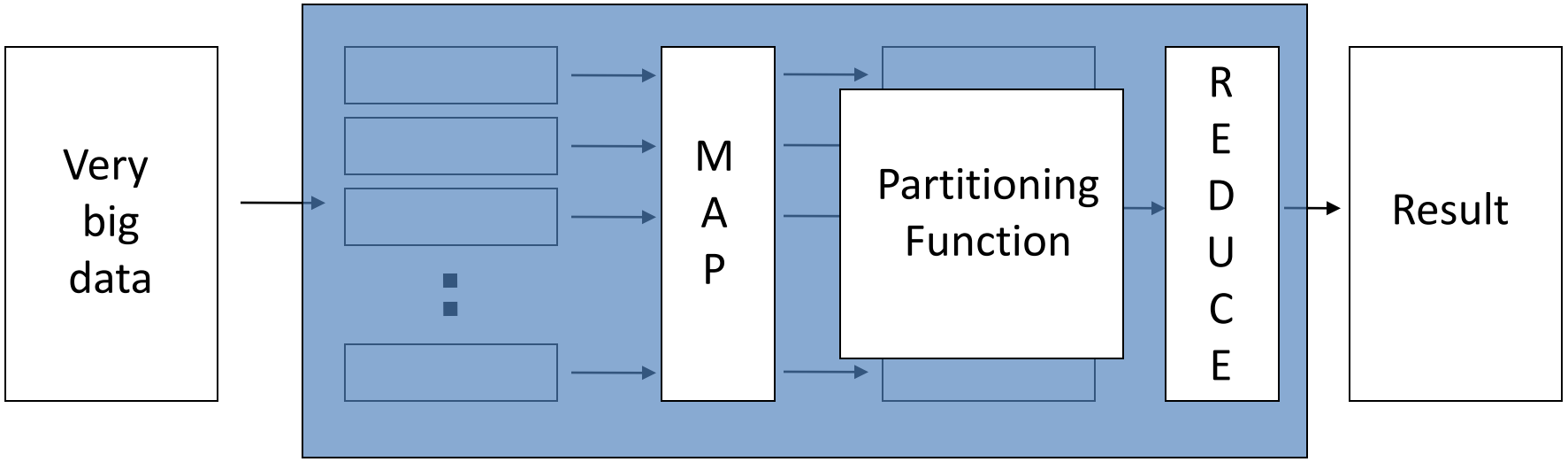
Relational database

Data Mart

Esempio: Distributed Word Count

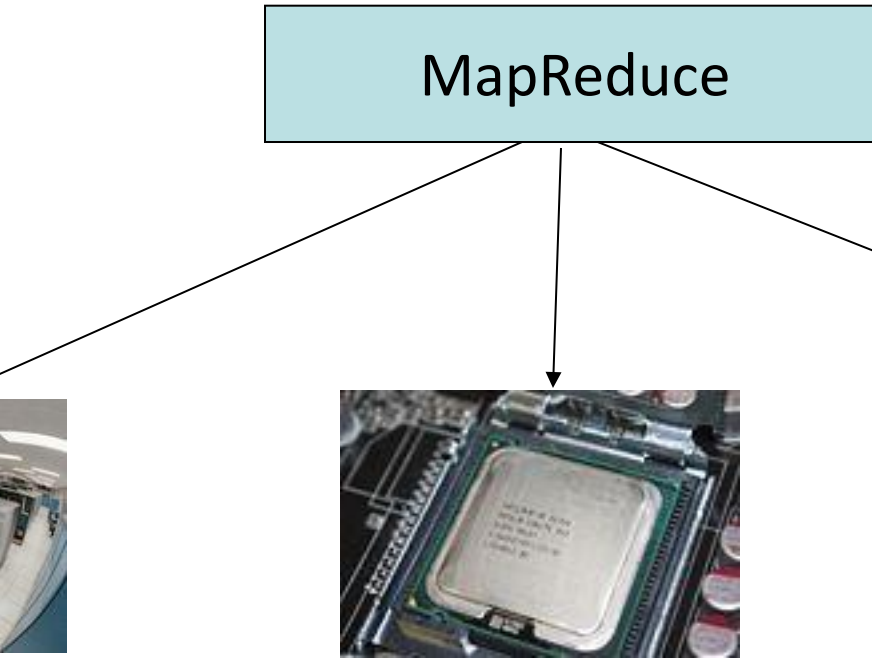


Map+Reduce



- Map:
  - Accepts *input* key/value pair
  - Emits *intermediate* key/value pair
- Reduce :
  - Accepts *intermediate* key/value* pair
  - Emits *output* key/value pair

MapReduce Implementations

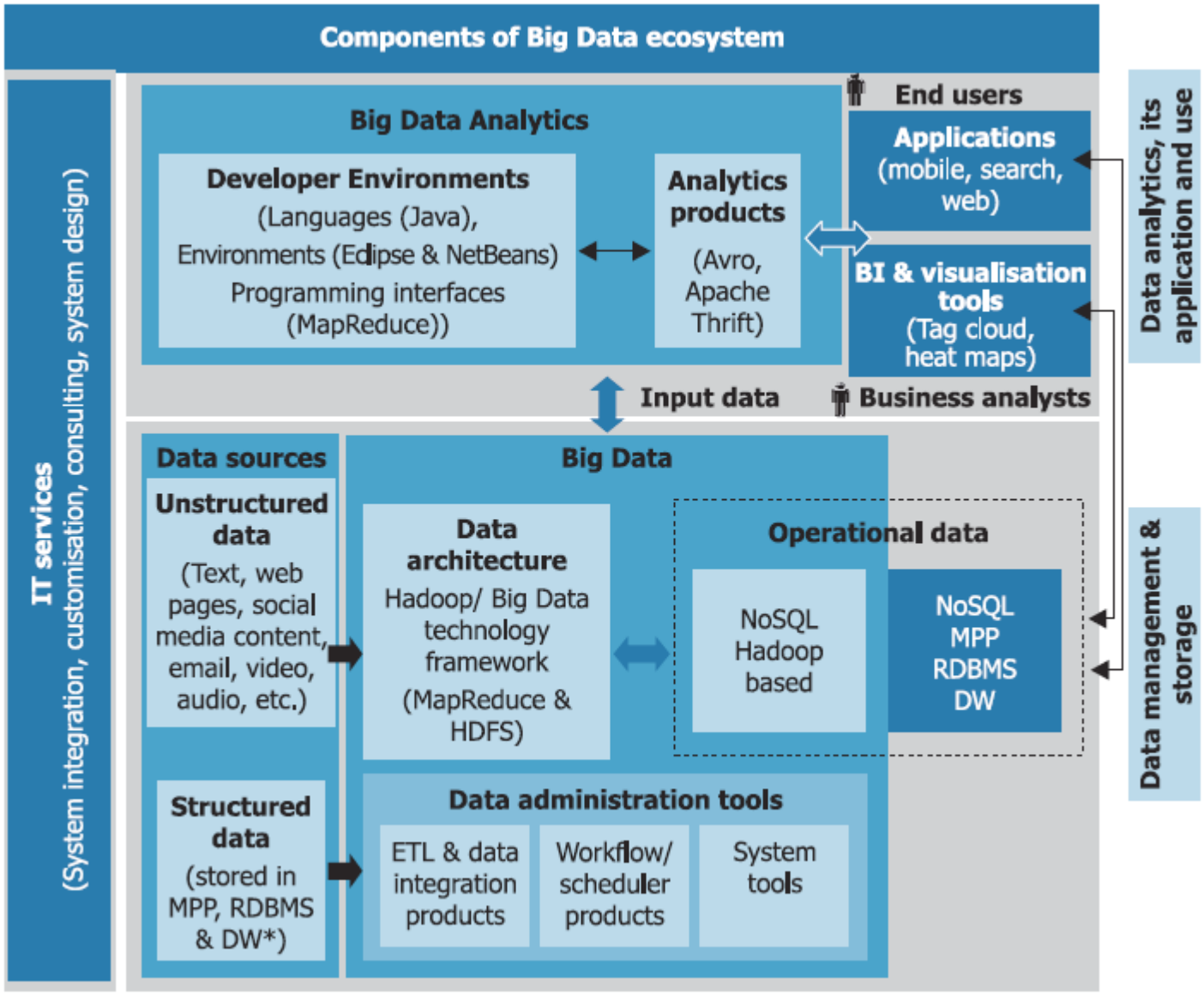


Cluster,
1, Google
2, Apache Hadoop

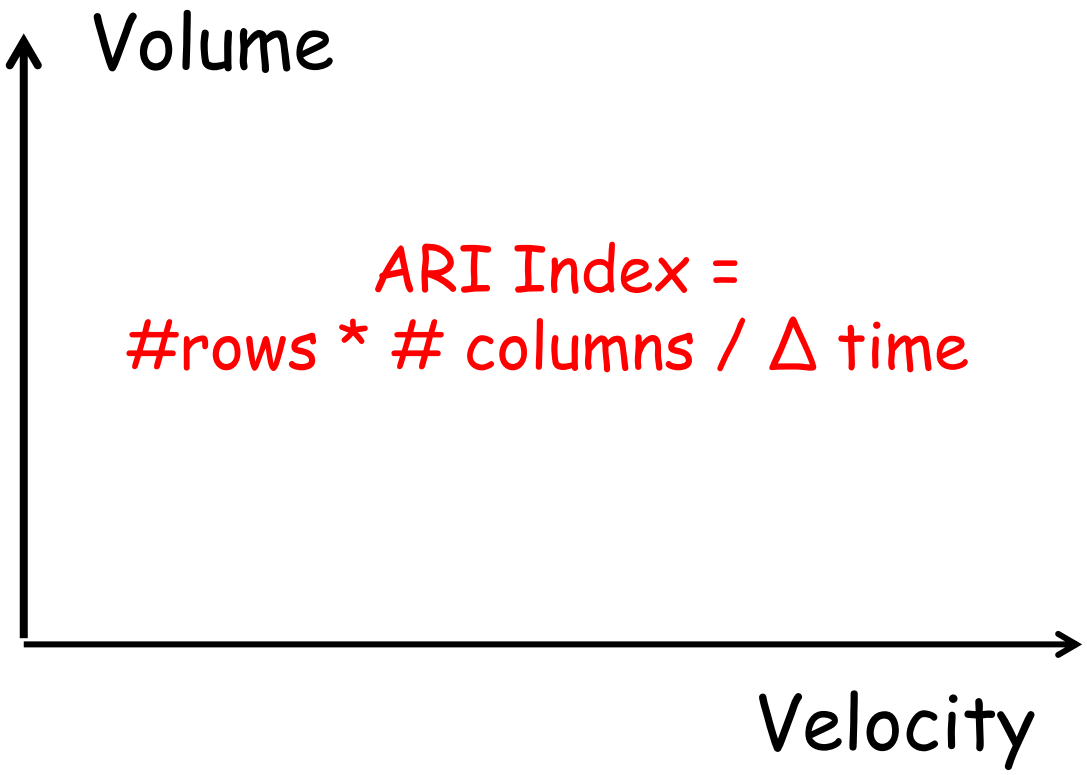
Multicore CPU,
Phoenix @ stanford

GPU,
Mars@HKUST

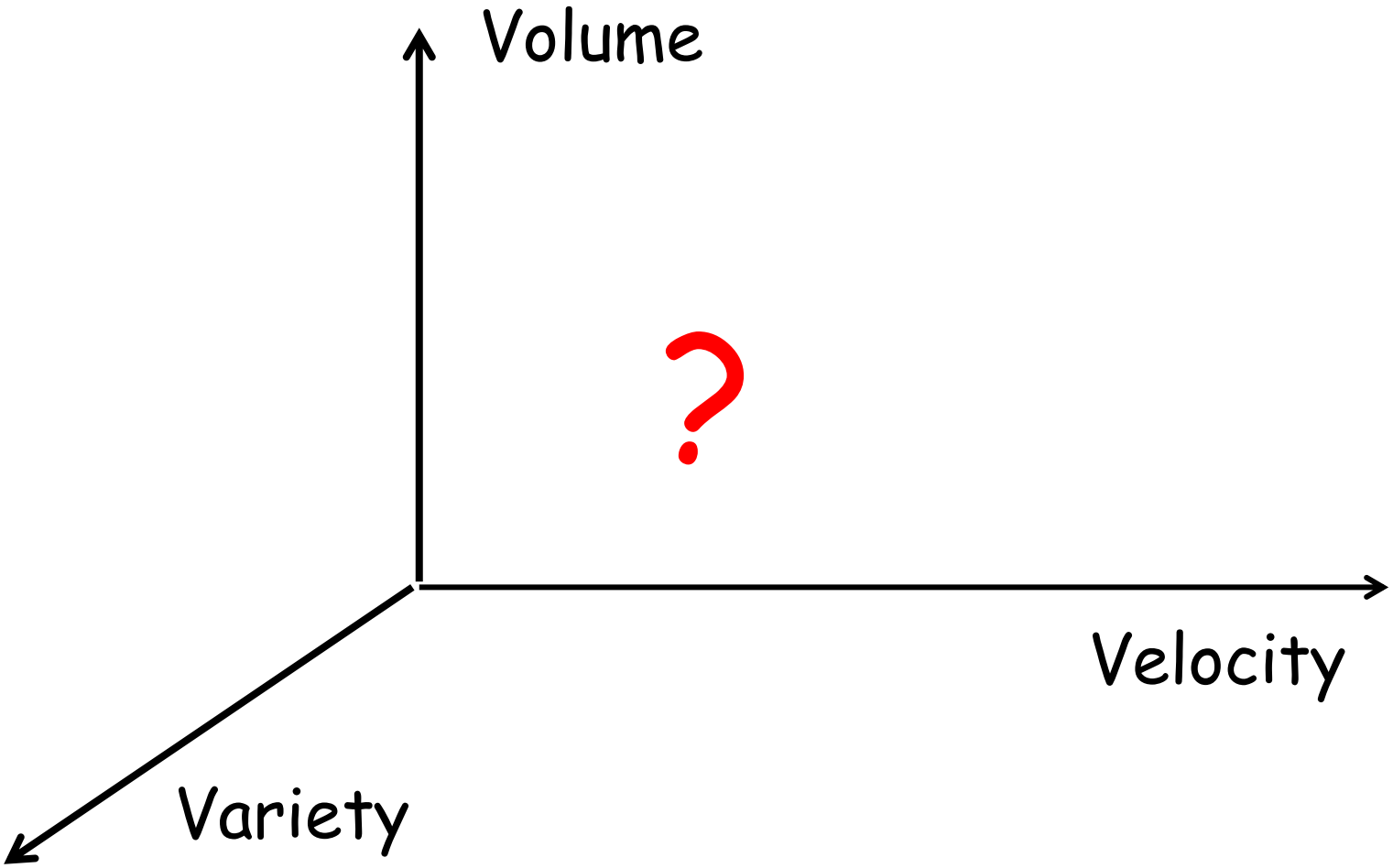
Big Data Ecosystem



Come stimare la complessita' nei Big Data? - 1



Come stimare la complessita' nei Big Data? - 2



Value

Applications of Big Data

	E-Commerce and Market Intelligence	E-Government and Politics 2.0	Science & Technology	Smart Health and Wellbeing	Security and Public Safety
Applications	• Recommender systems • Social media monitoring and analysis • Crowd-sourcing systems • Social and virtual games	• Ubiquitous government services • Equal access and public services • Citizen engagement and participation • Political campaign and e-polling	• S&T innovation • Hypothesis testing • Knowledge discovery	• Human and plant genomics • Healthcare decision support • Patient community analysis	• Crime analysis • Computational criminology • Terrorism informatics • Open-source intelligence • Cyber security
Data	• Search and user logs • Customer transaction records • Customer-generated content	• Government information and services • Rules and regulations • Citizen feedback and comments	• S&T instruments and system-generated data • Sensor and network content	• Genomics and sequence data • Electronic health records (EHR) • Health and patient social media	• Criminal records • Crime maps • Criminal networks • News and web contents • Terrorism incident databases • Viruses, cyber attacks, and botnets
	Characteristics: Structured web-based, user-generated content, rich network information, unstructured informal customer opinions	Characteristics: Fragmented information sources and legacy systems, rich textual content, unstructured informal citizen conversations	Characteristics: High-throughput instrument-based data collection, fine-grained multiple-modality and large-scale records, S&T specific data formats	Characteristics: Disparate but highly linked content, person-specific content, HIPAA, IRB and ethics issues	Characteristics: Personal identity information, incomplete and deceptive content, rich group and network information, multilingual content
Impacts	Long-tail marketing, targeted and personalized recommendation, increased sale and customer satisfaction	Transforming governments, empowering citizens, improving transparency, participation, and equality	S&T advances, scientific impact	Improved healthcare quality, improved long-term care, patient empowerment	Improved public safety and security

Quanti tipi di valore?

- Private providers → Value in exchange
  → Economic value
- Final Users → Value in use
  - Citizens → Quality of Life
  - Companies → Economic value
- Public Administration →
  - Public Value
  - Social Value

Leve per la creazione di valore nel settore pubblico (Source McKinsey) - 1

1. Creating transparency
   - Non far perdere tempo
   - Chiedere solo le informazioni che non si hanno gia'
   - Scambiare dati tra amministrazioni, per aumentarne il valore
   - Aprire i dati, per aumentare il valore d'uso e favorire il riuso.

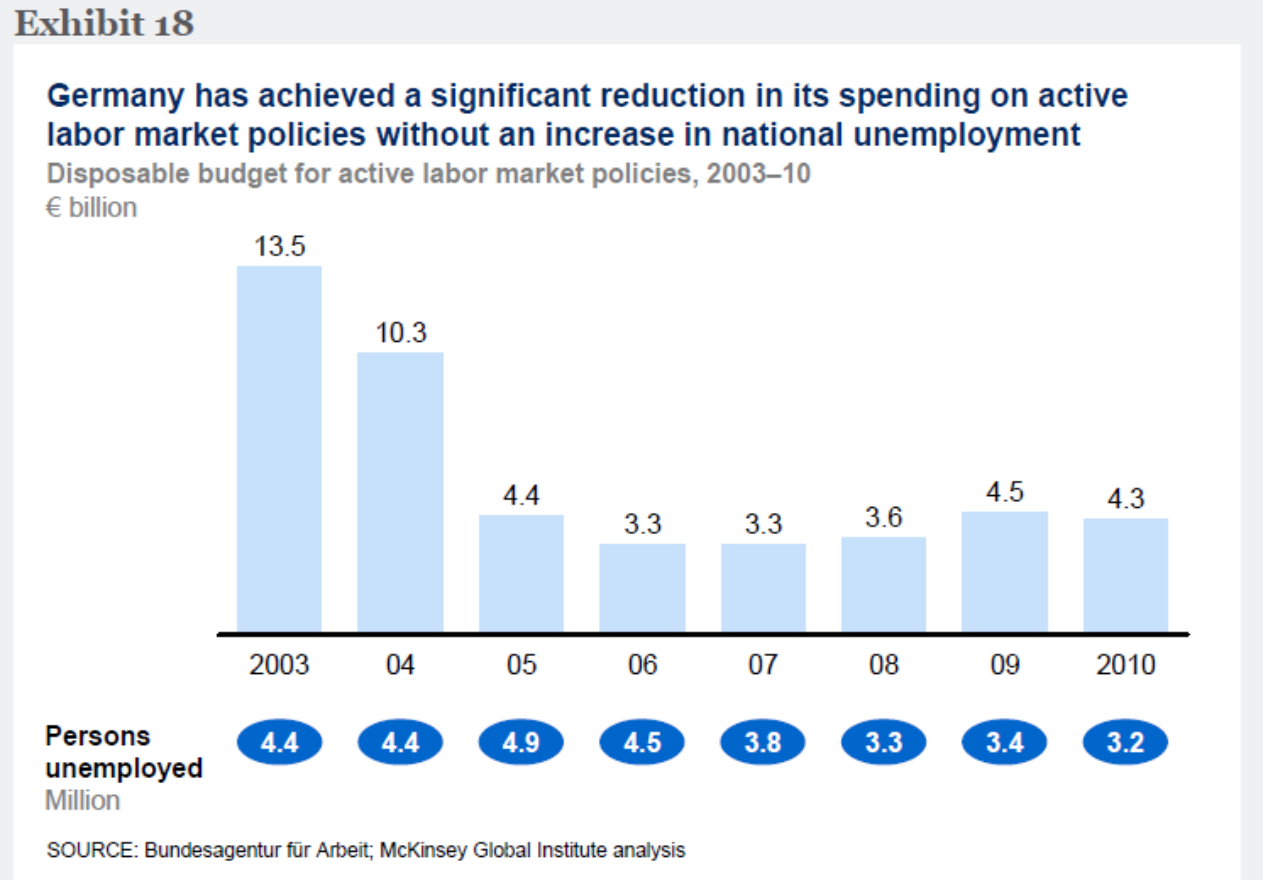
Leve per la creazione di valore nel settore pubblico (Source McKinsey)

2. Enabling experimentation to discover needs, expose variability, and improve performance
   - Aumentare la produttivita' → Fare benchmarking per favorire la competizione tra unita' organizzative
   - Spending review → Favorire il procurement di beni e servizi
   - Coinvolgere i cittadini nella scoperta dei bisogni inespressi e non soddisfatti dai servizi amministrativi basici
   - Tagliare i costi → prossima trasparenza

Leve per la creazione di valore nel settore pubblico (Source McKinsey)

3. Segmenting populations to customize actions

- Differenziare, come nel privato, i rapporti, in quel caso con i clienti (valore di scambio), in questo caso con i cittadini in cerca di lavoro, aumentando il rapporto <u>efficacia/efficienza</u> →



Leve per la creazione di valore nel settore pubblico (Source McKinsey)

4. Replacing/supporting human decision making with automated algorithms
   - Scambio di dati tra amministrazione per il contrasto alla evasione fiscale e contributiva
   - Contrasto alle frodi
   - Correzione degli errori nelle procedure amministrative.

Leve per la creazione di valore nel settore privato – es. Salute

5. Innovating new business models, products and services – Ex. The Daily Bread in UK

Conclusioni

Big data has the potential to create €150 billion to €300 billion or more in value across the OECD-Europe public sector

		Total base[1] € billion	×	Addressable %	×	Reduction %	=	Total value € billion
Operational efficiency savings	**Operating expenditure**	4,000		20–25		15–20		120–200
Reduction in fraud and error	**Transfer payment**	2,500		1–3[2]		30–40		7–30
Increase in tax collection	**Tax revenue**	5,400		5–10[3]		10–20		25–110
								150–300+

Social value

Social value – Economist October 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

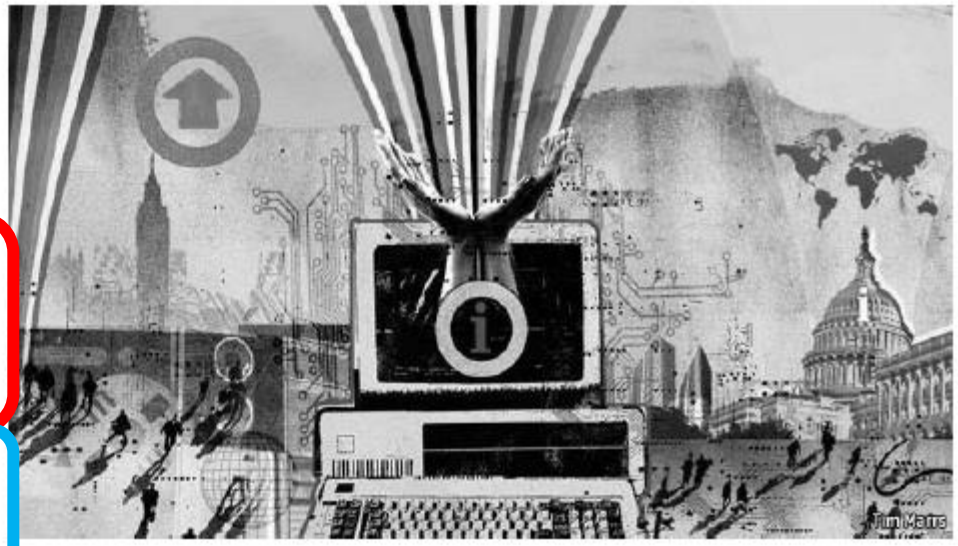
The parting of the red tape

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

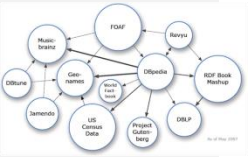
UGANDA is not best known as a testbed for new ideas in governance. But research there by Jakob Svensson at the University of Stockholm and colleagues suggested that giving people health-care performance data and helping them organise to submit complaints cut the death rate in under-fives by a third.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



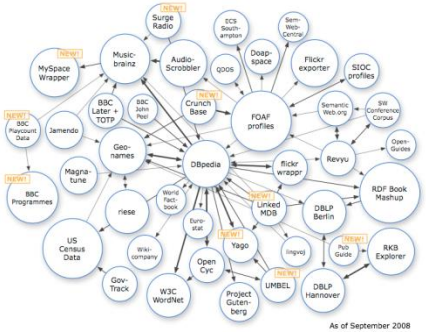
Tipi di Big Data su Web di interesse pubblico

- Open data
- Open linked Data

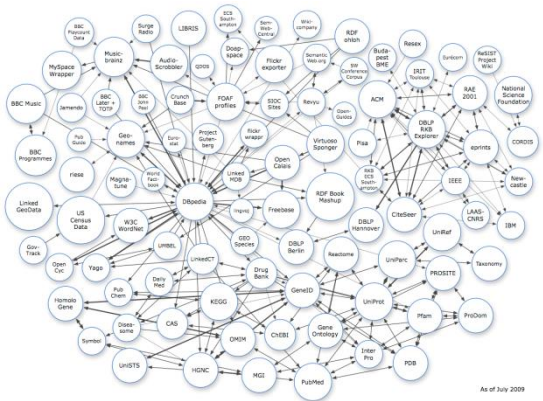
LOD Datasets on the Web



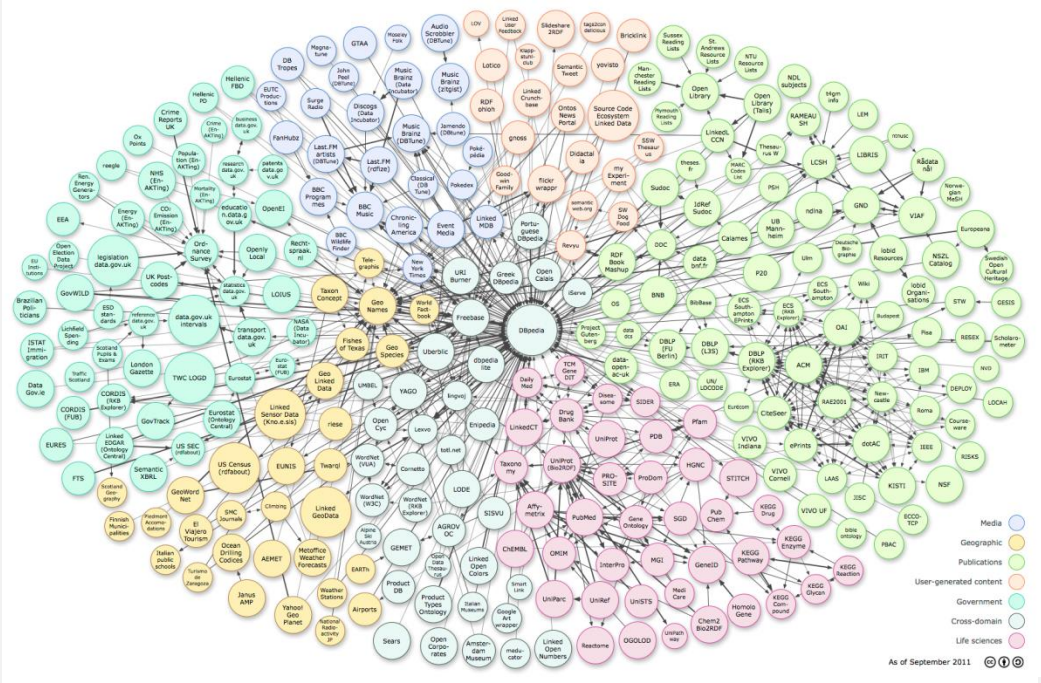
May 2007



September 2008

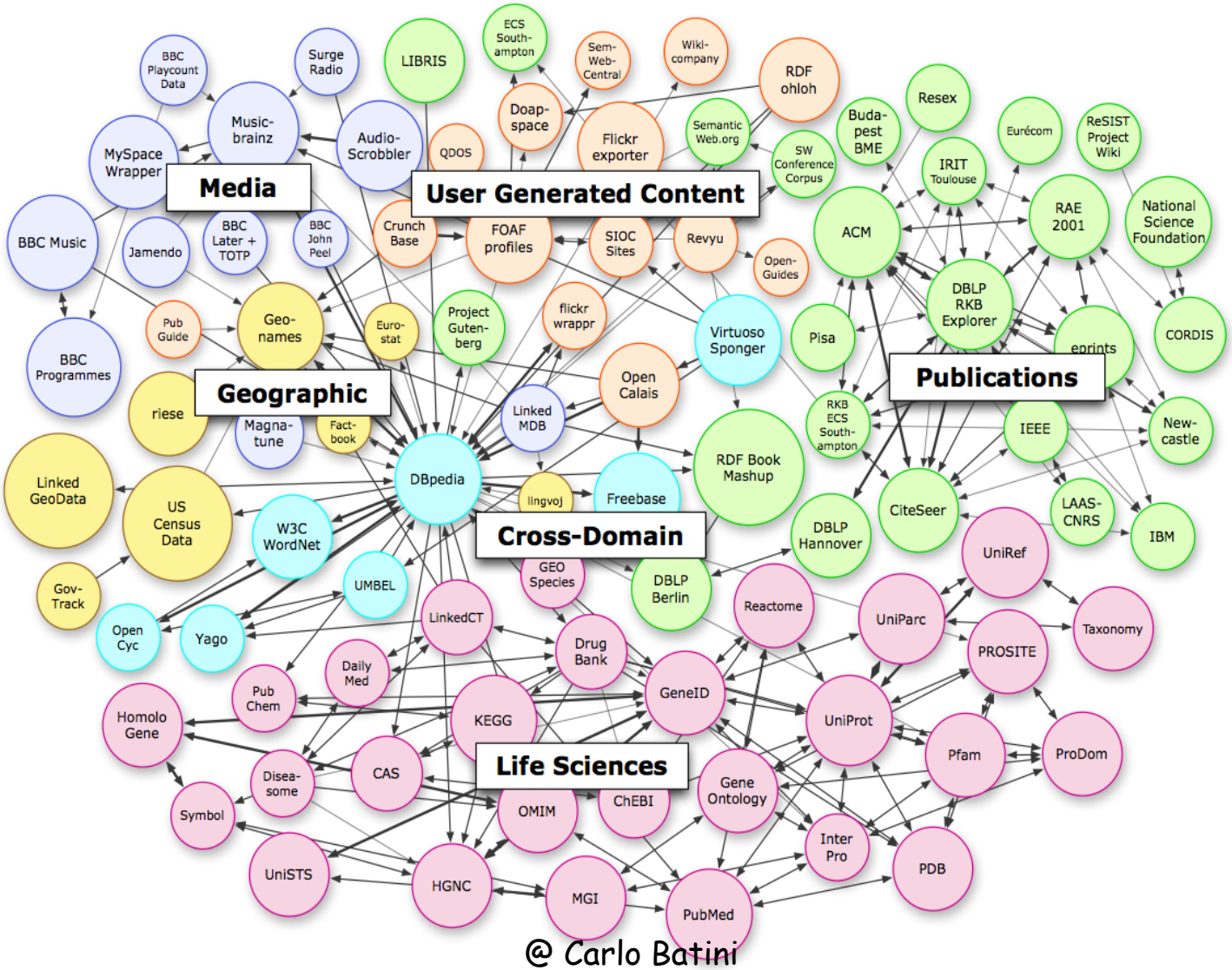


July 2009



September 2011

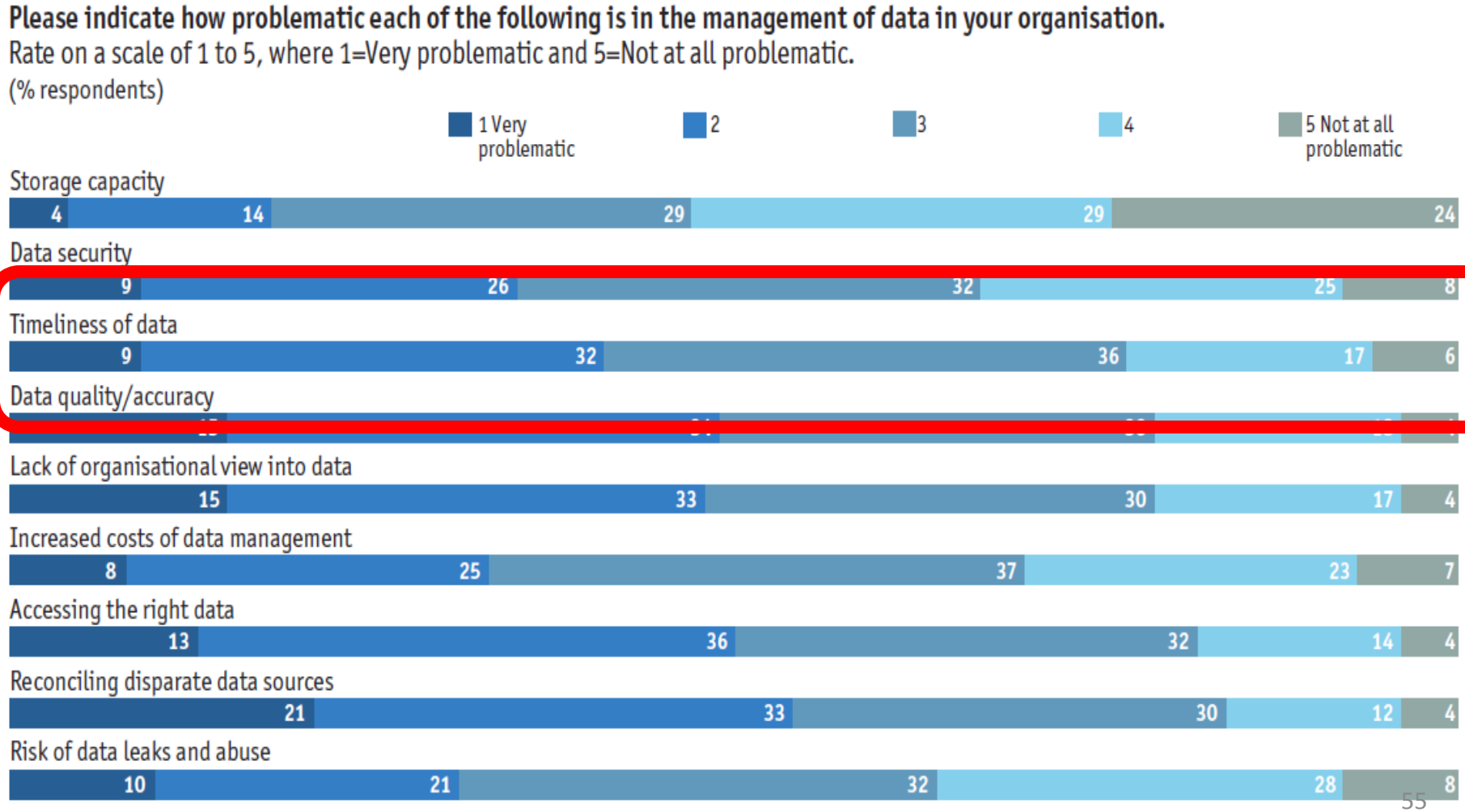
Linked Open Data: interconnettere dati sul Web



@ Carlo Batini

Is Data Quality a problem in Big Data?

Fonte Big Data
The Economist Intelligence Unit, sponsored by SAS
Responses from the Survey



Le dimensioni di qualita' dei dati

Dati strutturati in sistemi informativi tradizionali

- Accuracy
- Completeness
- Currency and Timeliness
- Consistency

Big Data

- Accuracy
- Completeness
- Currency and Timeliness
- Consistency
- Relevance
- Trustworthiness
- Provenance

Variety

Place	Country	Population	Main economic activity
Portofino	Italy	700.000	Tourism

Structured data

Image

Portofino

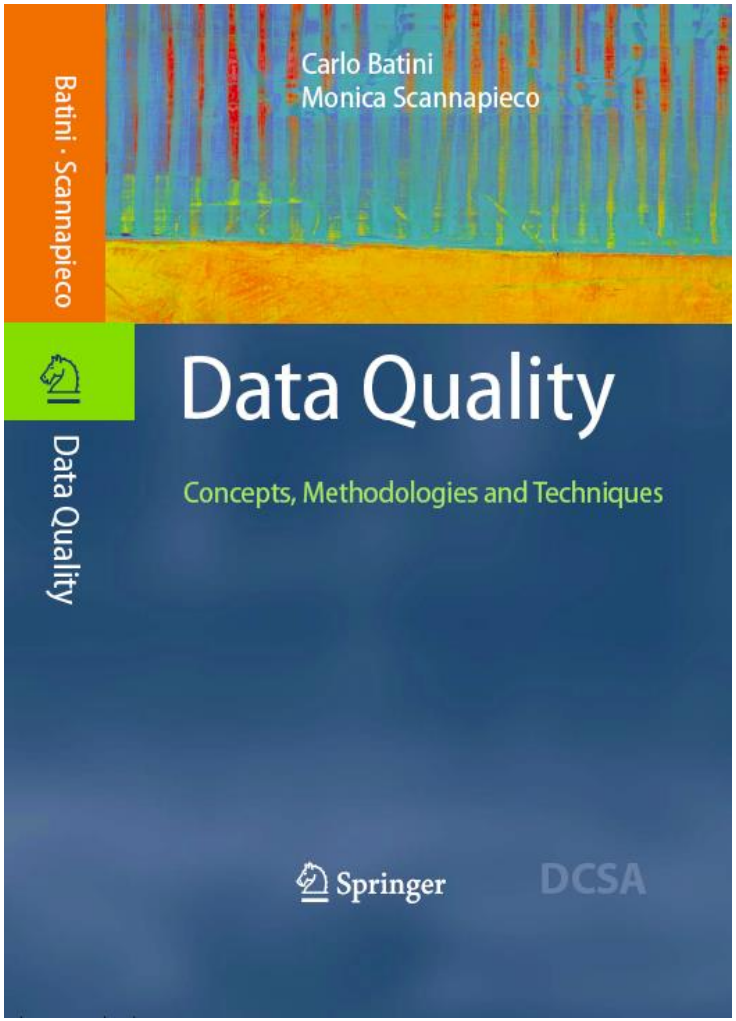
Map

Text

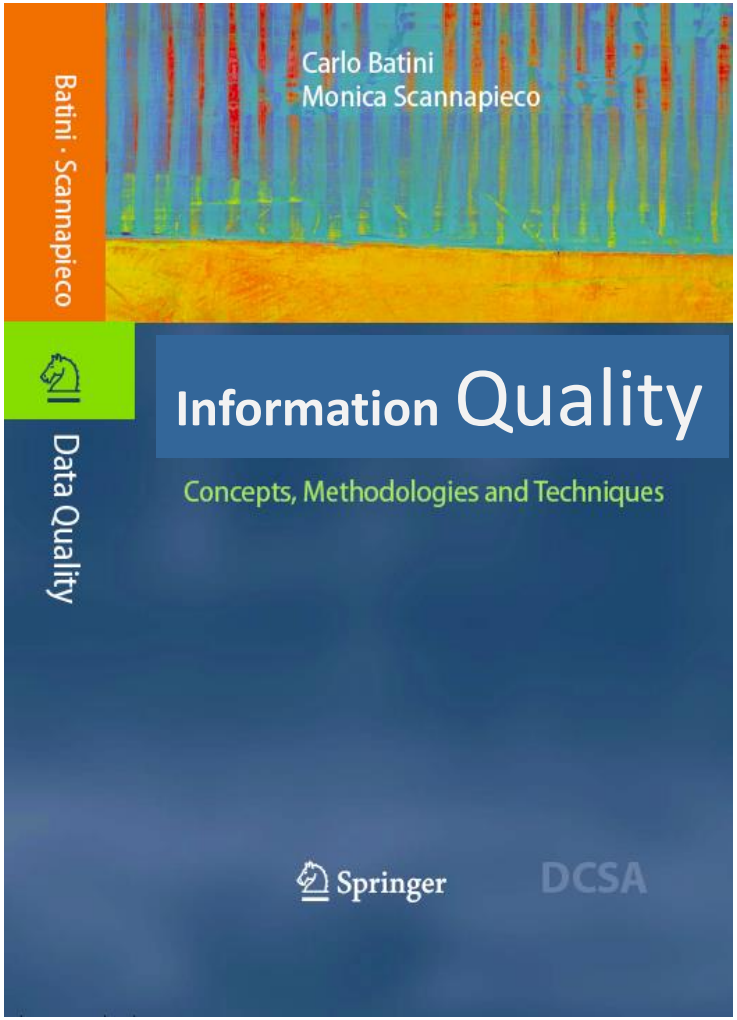
Dear Laure, I try to describe the wonder-ful harbour of Potfino as I have seen this morning a boat is going in, other boats are along the wharf. Small pretty buildings and villas are looking on to the harbour.

Un po' di pubblicita'……

2006



2014



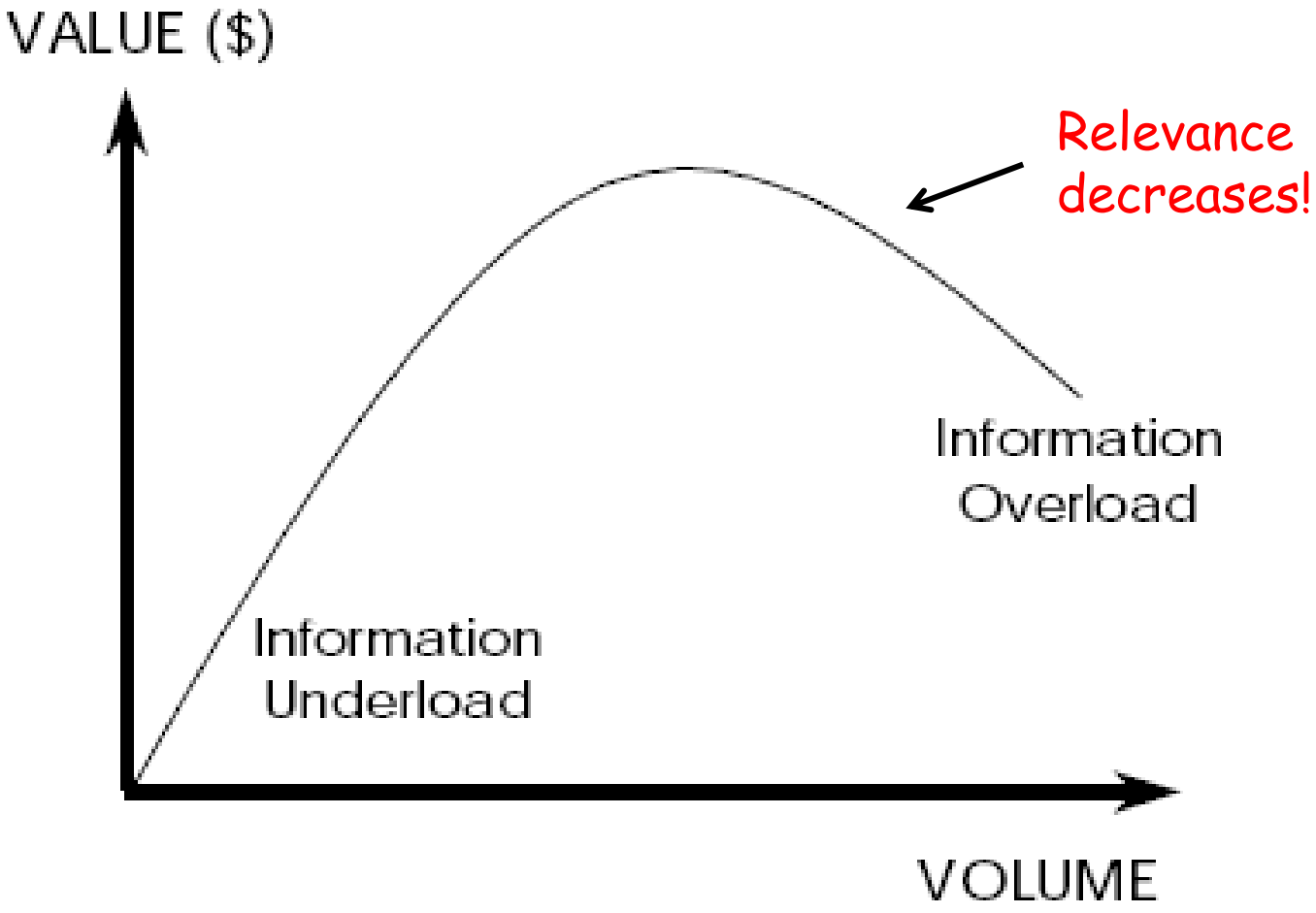
58

Relevance
Fonte: Indagine SAP, Caise Conference, Riga, Lettonia, 2004

- L'80% dei dati che mi servono per prendere le decisioni non li ho
- L'80% dei dati che mi forniscono per prendere le decisioni non mi servono

Moody & Walsh Law 6

More is not necessarily better



Value in terms of Utility
Utility in terms of Abstraction and Codification

- In the Boisot' approach, the utility that can be extracted from an information good is a function of codification and abstraction, two of the three dimensions in the I-Space.
- The third dimension of the I-Space, diffusion, allows us to establish the scarcity of information products.

The I-Space (Figure from Boisot 1995)

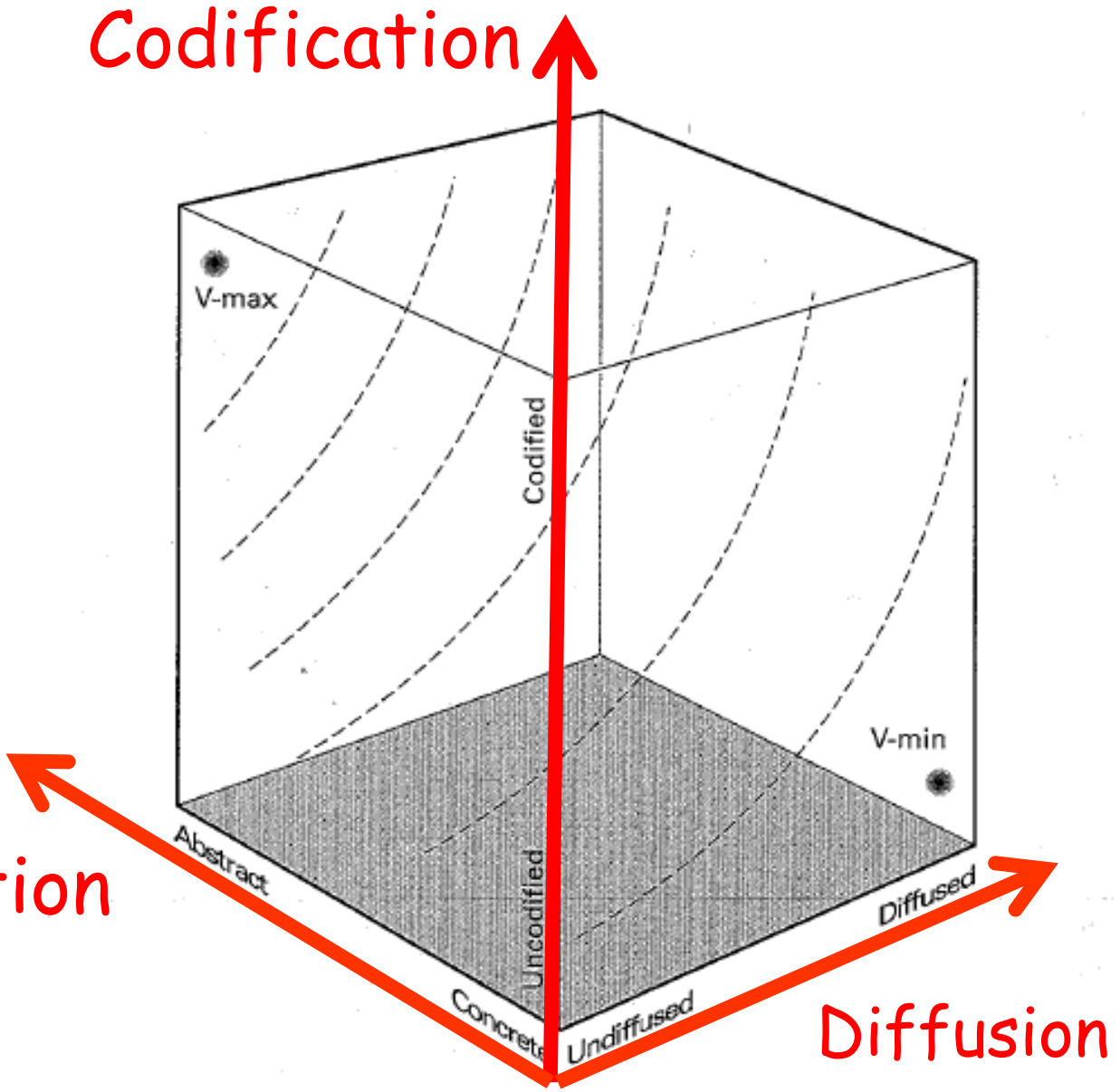
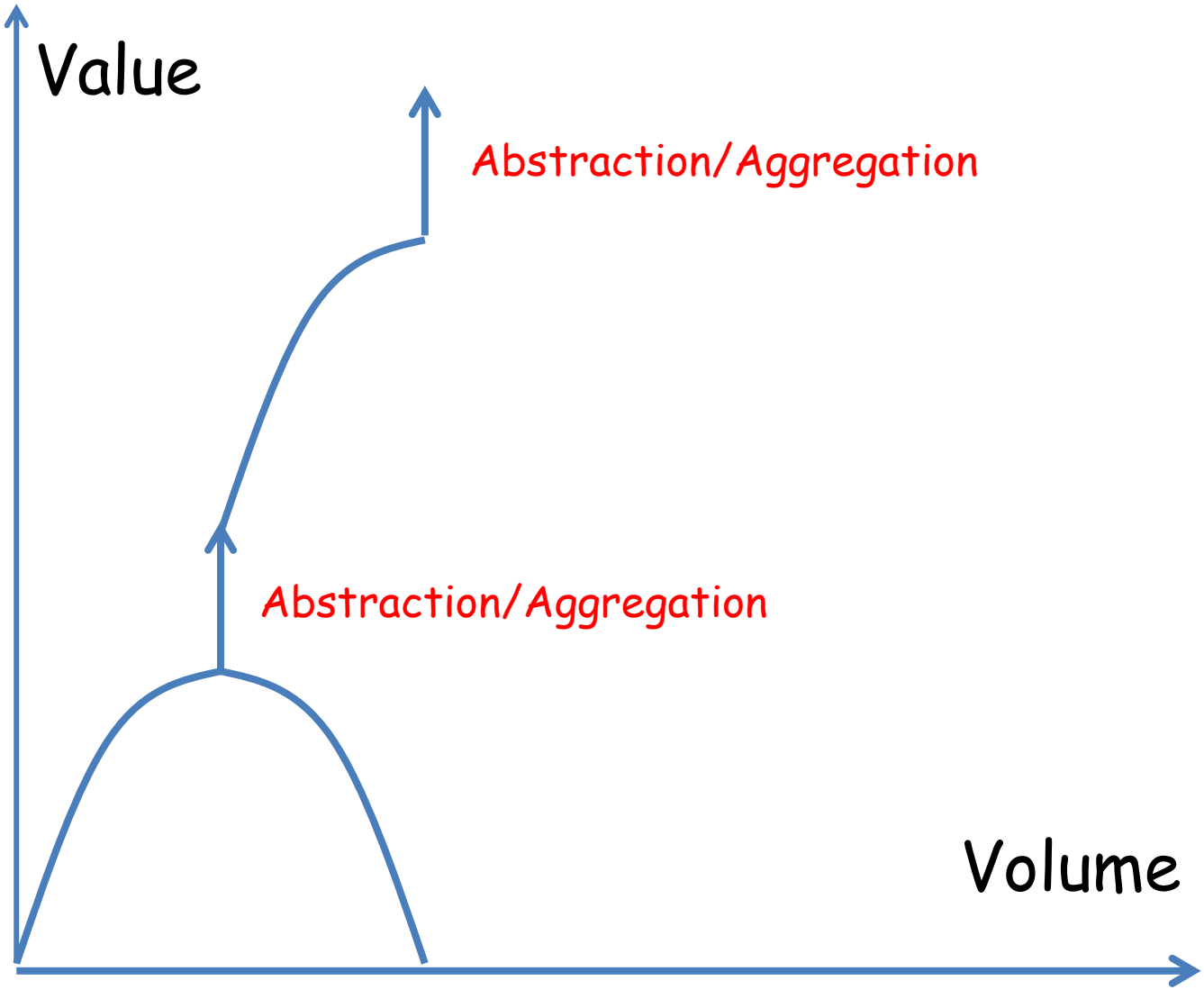


FIG. 4.1 Maximum and Minimum Value in the I-Space: V-max and V-min

Value & Abstraction



Aspetti metodologici delle statistiche pubbliche con fonti estese ai Big Data (fonte Barcaroli (Istat), 2013)

Tradizionali fonti delle indagini statistiche

- Fonti censuarie o a campionarie
- Fonti amministrative

Nuove fonti
delle indagini statistiche

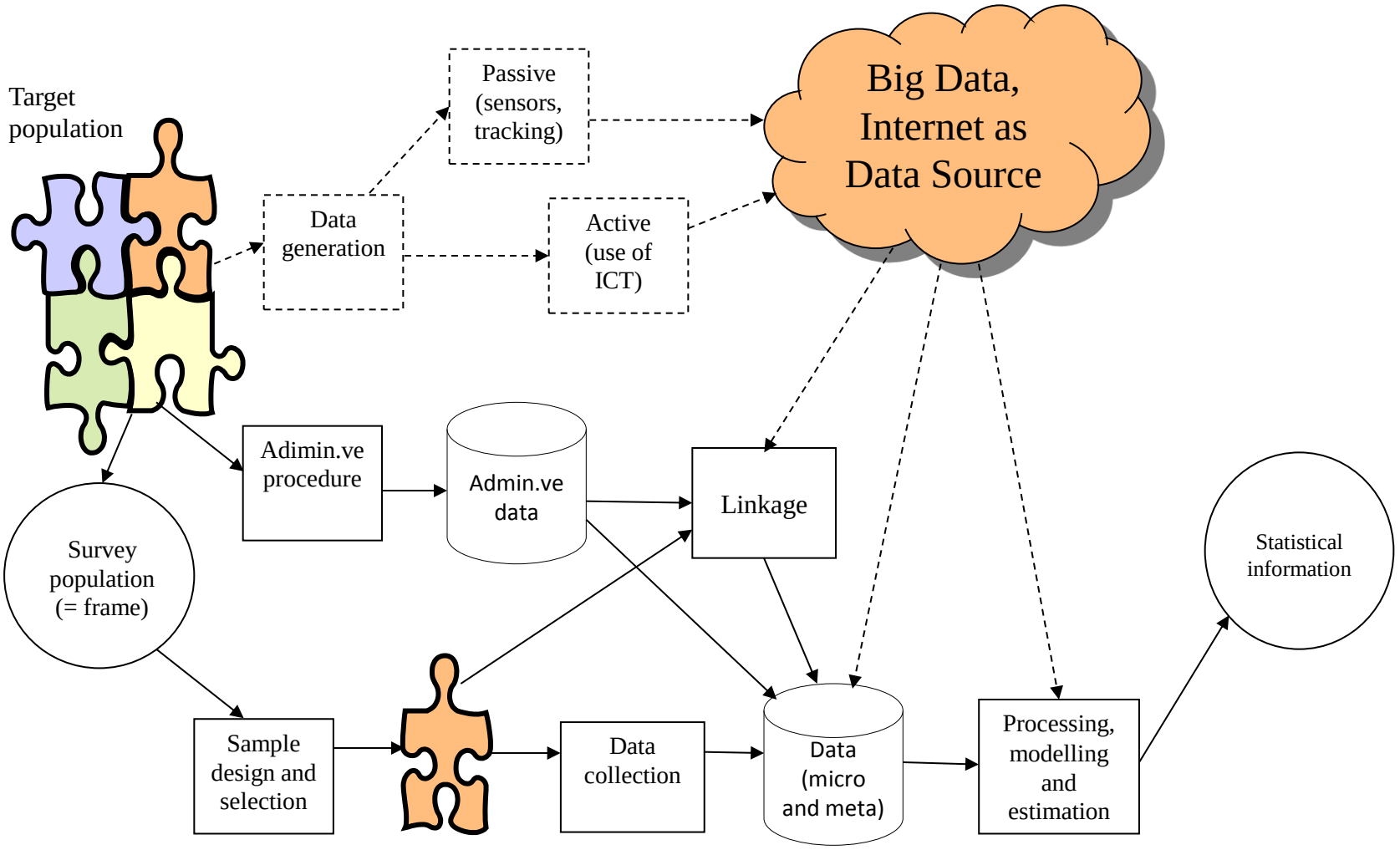
- Fonti censuarie o a campionarie
- Fonti amministrative
- Big Data

The methodological challenge

Big Data are generated by non planned (at statistical purposes) events. For a statistical use of Big Data, a suitable methodology must be able to:

1. **link** (with a known or estimated uncertainty) events to which Big Data refer to, to units of the population of interest for official statistics (individuals, households, businesses or institutions);
2. **process** collected data with the aim to make them coherent with the desired statistical framework (concepts, definitions, classifications);
3. give **weights** (with a known or estimated uncertainty) to data so as to guarantee **representativeness** with respect to the target population;
4. **estimate** aggregates of interest and accompany them with a measure of their quality, based on the uncertainty measures in previous steps.

A general framework



Riferimenti

- G. Barcaroli et al.- Using Big Data for Statistical Purposes – conference on Big data, social mining & social indicators, Pisa 2013.
- H. Chen et al. – Business Intellligence and Analytics: from big Data to Big Impact, MIS Quartelry, December 2012.
- Istat 2013 Documento non pubblicato, 2013.
- M. Fatallah – How can companies leverage big data? – World summit on Organization Design and big Data, Paris, May 2013.
- McKinsey & Company - Big data: The next frontier for innovation, competition, and productivity, 2011
- D. Moody, and P. Walsh, P. (1999) Measuring the value of Information: An asset valuation approach, ECIS 1999.
- Nasscom – Big Data: The Next Big Thing, 2012.
- O' Reilly Media (2012). Big Data Now
- The Economist Intelligence Unit - Big Data, sponsored by SAS, 2011.

Appendici

- How Big are Data?
- Myths around Big Data
- To Hadoop or not To Hadoop?

How big are data?

- "Big Data" is an imprecise term, and is even less precise when you consider "Big Data Analytics."
- How big is "big", exactly? Skytree introduced the Analytics Requirements Index (ARI) in order to eliminate the confusion and help organizations quantify their analytics requirements. It is a simple formula that can be used to estimate the scale of the analytics problem and thus the power of the analytics engine required to solve the problem

Analytics Requirements Index (ARI) - 1

Definition

The Analytics Requirements Index (ARI) has a simple definition:

Analytics Requirements Index = $\frac{\text{\# Rows} \times \text{\# Columns}}{\text{Time (secs)}}$

Where	# Rows =	Number of records being analyzed
	# Columns =	Number of variables captured in each record
	Time (secs) =	The timeframe within which to complete the analysis

---

A simple way to interpret this definition is to say that a certain number of data elements (ie. # Rows × # Columns) must be processed within a given period of time.

Analytics Requirements Index (ARI) - 2

Example

To bring the ARI to life, consider a recommendation engine that suggests additional purchases to a user before checkout from an online store. In this case:

# Rows =	2MM previous purchases at the store
# Columns =	33 fields including previous items, price, etc.
Time (secs) =	The recommendation should be generated in half a second

---

The resulting ARI is 132MM per second.

In effect this means that the data analytics engine must process 132MM data elements per second to optimally address this problem.

Analytics Requirements Index (ARI) - 3

Scale

Skytree Server scales to handle virtually any ARI – contact us and we can help you understand your implementation options for your Big Data Analytics challenge.

	Low	Medium	High
Volume (# rows)	< 10M	< 100M	> 100M
Velocity (time)	Hours	Minutes	Real-time (Sub-Seconds)
Variety/Complexity (# columns)	< 100	< 1000	> 1000

Myths around Big Data

- With the amount of hype around Big Data it's easy to forget that we're just in the first inning. More than three exabytes of new data are created each day, and market research firm IDC estimates that 1,200 exabytes of data will be generated this year alone.
- The expansion of digital data has been underway for more than a decade and for those who've done a little homework, they understand that Big Data references more than just Google, eBay, or Amazon-sized data sets. The opportunity for a company of any size to gain advantages from Big Data stem from data aggregation, data exhaust, and metadata — the fundamental building blocks to tomorrow's business analytics. Combined, these data forces present an unparalleled opportunity.
- Yet, despite how broadly Big Data is being discussed, it appears that it is still a very big mystery to many. In fact, outside of the experts who have a strong command of this topic, the misunderstandings around Big Data seem to have reached mythical proportions. Here are the top five myths

1. Big Data is Only About Massive Data Volume

- Volume is just one key element in defining Big Data, and it is arguably the least important of three elements. The other two are variety and velocity. Taken together, these three "Vs" of Big Data were originally posited by Gartner's Doug Laney in a 2001 research report.
- Generally speaking, experts consider petabytes of data volumes as the starting point for Big Data, although this volume indicator is a moving target. Therefore, while volume is important, the next two "Vs" are better individual indicators.
- Variety refers to the many different data and file types that are important to manage and analyze more thoroughly, but for which traditional relational databases are poorly suited. Some examples of this variety include sound and movie files, images, documents, geo-location data, web logs, and text strings.
- Velocity is about the rate of change in the data and how quickly it must be used to create real value. Traditional technologies are especially poorly suited to storing and using high-velocity data. So new approaches are needed. If the data in question is created and aggregates very quickly and must be used swiftly to uncover patterns and problems, the greater the velocity and the more likely that you have a Big Data opportunity.

2. Big Data Means Hadoop

- Hadoop is the Apache open-source software framework for working with Big Data. It was derived from Google technology and put to practice by Yahoo and others. But, Big Data is too varied and complex for a one-size-fits-all solution. While Hadoop has surely captured the greatest name recognition, it is just one of three classes of technologies well suited to storing and managing Big Data. The other two classes are NoSQL and Massively Parallel Processing (MPP) data stores. (See myth number five below for more about NoSQL.) Examples of MPP data stores include EMC's Greenplum, IBM's Netezza, and HP's Vertica.

  Plus, Hadoop is a software framework, which means it includes a number of components that were specifically designed to solve large-scale distributed data storage, analysis and retrieval tasks. Not all of the Hadoop components are necessary for a Big Data solution, and some of these components can be replaced with other technologies that better complement a user's needs. One example is MapR's Hadoop distribution, which includes NFS as an alternative to HDFS, and offers a full random-access, read/write file system.

3. Big Data Means Unstructured Data

- The term "unstructured" is imprecise and doesn't account for the many varying and subtle structures typically associated with Big Data types. Also, Big Data may well have different data types within the same set that do not contain the same structure.
- Therefore, Big Data is probably better termed "multi-structured" as it could include text strings, documents of all types, audio and video files, metadata, web pages, email messages, social media feeds, form data, and so on. The consistent trait of these varied data types is that the data schema isn't known or defined when the data is captured and stored. Rather, a data model is often applied at the time the data is used.

4. Big Data is for Social Media Feeds and Sentiment Analysis

- Simply put, if your organization needs to broadly analyze web traffic, IT system logs, customer sentiment, or any other type of digital shadows being created in record volumes each day, Big Data offers a way to do this.
- Even though the early pioneers of Big Data have been the largest, web-based, social media companies — Google, Yahoo, Facebook — it was the volume, variety, and velocity of data generated by their services that required a radically new solution rather than the need to analyze social feeds or gauge audience sentiment.
- Now, thanks to rapidly increasing computer power (often cloud-based), open source software (e.g., the Apache Hadoop distribution), and a modern onslaught of data that could generate economic value if properly utilized, there are an endless stream of Big Data uses and applications. A favorite and brief primer on Big Data, which contains some thought-provoking uses, was published as an article early this year in *Forbes.*

5. NoSQL means No SQL

- NoSQL means "not only" SQL because these types of data stores offer domain-specific access and query techniques in addition to SQL or SQL-like interfaces. Technologies in this NoSQL category include key value stores, document-oriented databases, graph databases, big table structures, and caching data stores. The specific native access methods to stored data provide a rich, low-latency approach, typically through a proprietary interface. SQL access has the advantage of familiarity and compatibility with many existing tools. Although this is usually at some expense of latency driven by the interpretation of the query to the native "language" of the underlying system.
- For example, Cassandra, the popular open source key value store offered in commercial form by DataStax, not only includes native APIs for direct access to Cassandra data, but CQL (it's SQL-like interface) as its emerging preferred access mechanism. It's important to choose the right NoSQL technology to fit both the business problem and data type and the many categories of NoSQL technologies offer plenty of choice.

To Hadoop or Not to Hadoop?

Not to Hadoop
1. Big Data cravings

- While businesses like to believe that they have a Big Data dataset, sadly, it seems that is often not the case.
- Regarding data volume and common perceptions that one possesses "Big Data", a research article, Nobody Ever Got Fired For Buying a Cluster, reveals that while Hadoop was designed for tera/petabyte scale computation, majority of real world jobs process less than 100 GB of input (with median jobs at Microsoft & Yahoo under 14 GB and 90% of jobs at Facebook being well under 100GB) and hence, puts forth the case for a single "scale-up" server over a "scale-out" setup running Hadoop.

2. You are in the queue

- When submitting jobs, Hadoop's minimum latency is about a minute. This means that it takes the system a minute or more to respond, and provide recommendations, to the customer's purchase. It would be a loyal and patient customer who would stare at the screen for 60+ seconds waiting for a response.
- An option is to pre-compute related items for every item in the inventory a priori using Hadoop, and provide the web site or mobile app immediate, one-second-or-less access to the stored result. Hadoop is an excellent Big Data pre-computation engine. Of course, as the nature of your response gets more complicated complete pre-computation is very inefficient.

3. Your call will be answered in ..

- Hadoop has not served businesses requiring real-time responses to their queries. Jobs which go through the map-reduce cycle also spend time in the shuffle cycle. None of these are time-bound making developing real-time applications on top of Hadoop, very difficult. Volume-weighted average price trading is an example where responses need to be time-bound to place buys.
- Analysts sorely miss SQL. Hadoop doesn't function well for random access to its datasets (even with Hive, which basically makes MapReduce jobs of your query). Google's Dremel (and by extension, BigQuery) architecture is designed to support ad-hoc queries over huge row-sets in under seconds. And SQL lets you do joins. Shark from University of California, Berkeley's AmpLab and the Stinger initiative led by Hortonworks are other alternatives to look out for.

Continues

- Let's say it together: Hadoop works in batch mode. That means as new data is added the jobs need to run over the entire set again. Hence, analyses time keeps increasing. Chunks of fresh data, mere updates or small changes might flow in real-time. Often, businesses need to make decisions based on these events.
- However rapidly the incoming data is ingested Hadoop would still process them in batch mode. YARN promises to address this in the future. Twitter's Storm is already popular & an available alternative. The case for combining Storm with a distributed messaging system like Kafka opens up a variety of use cases for stream aggregation and processing. But load balancing is sorely missing in Storm while available in Yahoo's S4

Continues

- Real-time advertisements and monitoring sensor data mandate real-time processing of streaming input. But Hadoop or tools built on tops of them are not the only alternatives.
- SAP's HANA in-memory database was used in the McClaren team's ATLAS suite of analytics tools during the recent Indy 500 along with MATLAB to run simulations and respond to telemetry during the race.
- Many analysts opine that the future of Hadoop is interactive and real-time.

4. I just broke up with my social network..

- Hadoop, especially MapReduce, is best suited for data that can be decomposed to key-value pairs without fear of losing context or any implicit relationship. Graphs possess implicit relationships (edges, sub-trees, child and parent relationships, weights, etc.) and not all of them will exist on a node. This attribute requires most graph algorithms to carry a portion or the entire graph through each iteration. This is often not feasible or at least convoluted to realize in MapReduce.
- There is also the problem of strategy of data partitioning across nodes. If your primary data structure is a graph or a network, then you are probably better off using a graph database like Neo4J or Dex or you could explore recent entries on the scene like Google's Pregel or Apache Giraph.

5. The mold of MapReduce

- Some tasks/jobs/algorithms simply do not yield to the programming model of MapReduce. One such set of problems was touched upon in the previous paragraph. Tasks that need the results of intermediate steps to compute results of current step would be another category (an academic example is the Fibonacci series computation).
- Some machine learning algorithms (gradient-based learning or expectation maximization) too do not fall well into the MapReduce paradigm.
- There are specific optimisations/strategies (global state, passing along data structures for reference, etc.) for each of these issues that have been suggested by researchers but it still makes the implementation more non-intuitive & complicated than is necessary.

Continues

- Added to these are business cases where the data is not significantly large or the total data set is large but made up of billions of small files (e.g. many image files which need to be scanned for a particular shape) which can't be concatenated.
- As we already mentioned, jobs which do not lend themselves to the MapReduce paradigm of divide and aggregate also make adopting Hadoop contrived.

Hadoop!

Does your organization...

- Want to extract information from piles of text logs?
- Want to transform largely unstructured or semi-structured data into some other useable and structured format?
- Have tasks that can run over the entire set of data, overnight (like credit card companies do with the day's transactions)?
- Treat conclusions drawn from a single processing of data as valid till the next scheduled processing (unlike stock market prices which definitely change between end of day values)?

Then, most certainly you should explore Hadoop.

Continues

- These represent a sizeable list of categories of business problems which fit well into the Hadoop model (although reports suggest that even on those, taking it to production is a non-trivial challenge). Typical jobs that have to go over huge quantities of unstructured or semi-structured data and either summarise the contents or transform relevant observations into a structured form to be utilised by other components in the system, are very well suited for the Hadoop model.
- If your collected data has elements that can easily be captured as an identifier with its corresponding value (which in Hadoop-speak is key-value pairs), you can utilise that simple association to perform several kinds of aggregations.
- At the end of the day, the key is to recognise the business resources available and understand the nature of the problem you wish to solve. That and the elaboration above would help you choose the best tools for your business. And it may very well be Hadoop

Resti

Big Data sources

- Commercial or transactional: (arising from the transaction between two entities), e.g. credit card transactions, on-line transactions (including from mobile devices), etc.
- From sensors, e.g. satellite imaging, road sensors, climate sensors, etc.
- From tracking devices, e.g. tracking data from cells, GPS, etc.
- Behavioural, e.g. online searches (about a product, a service or any other type of information), online page view, etc.
- Opinion, e.g. comments on social media, etc.

Esempio di Big Data
il sito www.flightradard24.com

Tecnologie per le fonti di Big Data

ADS-B

The primary technology that we use to receive flight information is called automatic dependent surveillance-broadcast (ADS-B). The ADS-B technology itself is best explained by the image to the right.

1. Aircraft gets its location from a GPS navigation source (satellite)
2. The ADS-B transponder on aircraft transmits signal containing the location (and much more)
3. ADS-B signal is picked up by a receiver connected to Flightradar24
4. Receiver feeds data to Flightradar24
5. Data is shown on www.flightradar24.com and in Flightradar24 apps

Today, roughly 60% of all passenger aircraft (70% in Europe, 30% in the US) are equipped with an ADS-B transponder. This percentage is steadily increasing as ADS-B is set to replace radar as the primary surveillance method for controlling aircraft.

Flightradar24 has a network of about 2000 ADS-B receivers around the world that receives plane and flight information from aircraft with

More on Volume

- Volume is just one key element in defining Big Data, and it is arguably the least important of three elements.
- Generally speaking, experts consider petabytes of data volumes as the starting point for Big Data, although this volume indicator is a moving target. Therefore, while volume is important, the next two "Vs" are better individual indicators.