Big Data e Linked Open Data per la statistica ufficiale: verso nuove forme di conoscenza generata dai dati

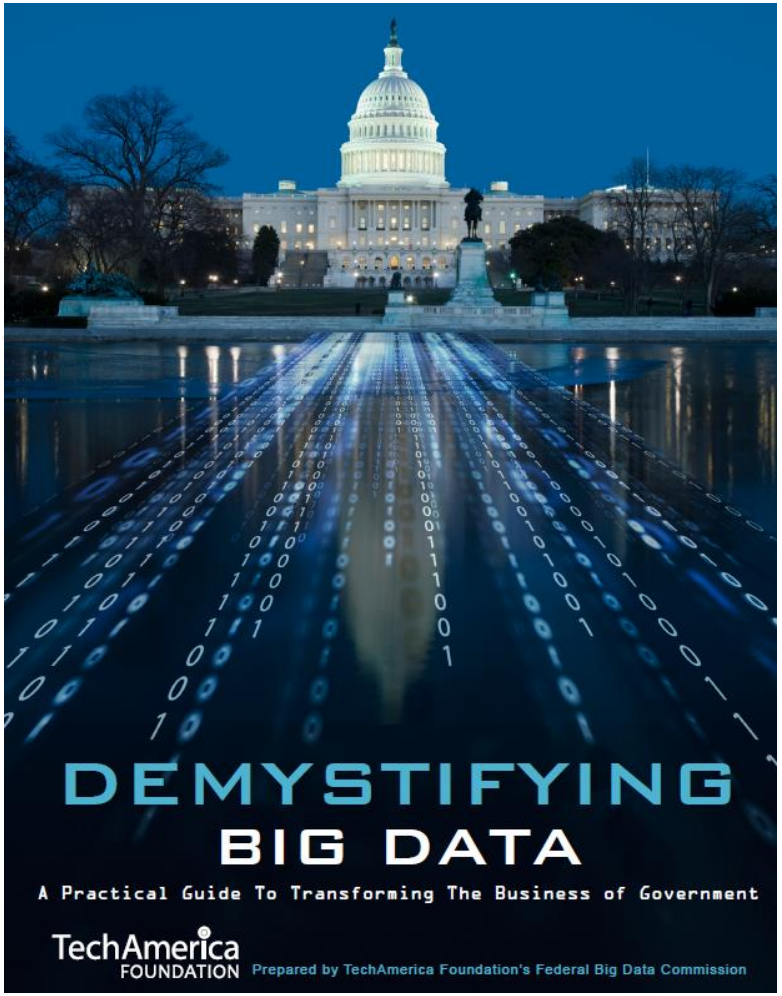
Stefano De Francisci

25 maggio 2017

Istat

Indice

Indice



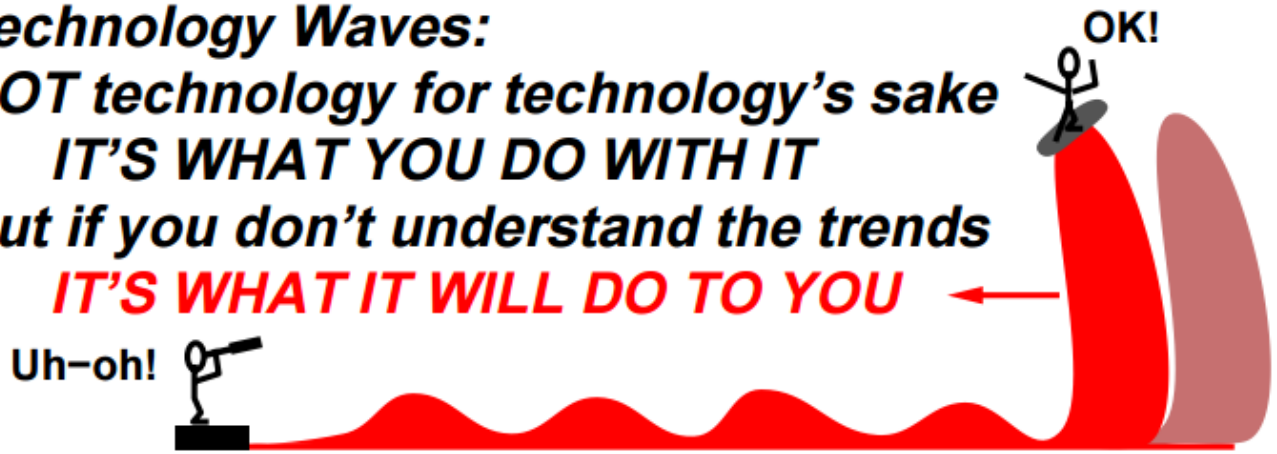
Origine dei Big Data

sgi

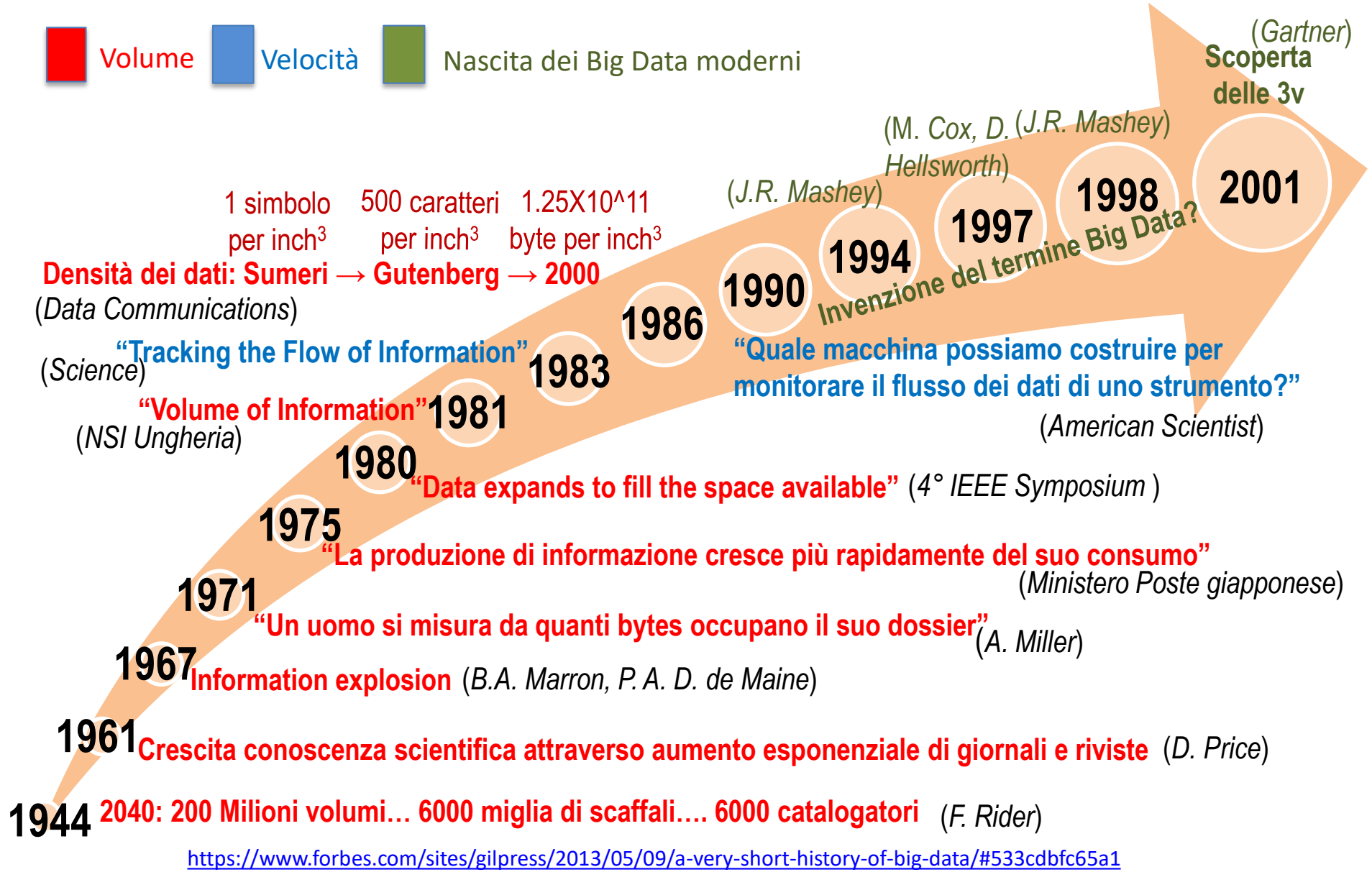
Big Data ...
and the Next Wave of InfraStress

John R. Mashey
Chief Scientist, SGI

Technology Waves:
NOT technology for technology's sake
IT'S WHAT YOU DO WITH IT
But if you don't understand the trends
IT'S WHAT IT WILL DO TO YOU

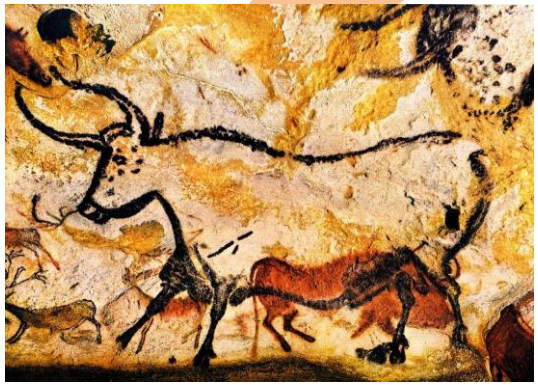


Breve storia dei Big Data secondo Gil Press

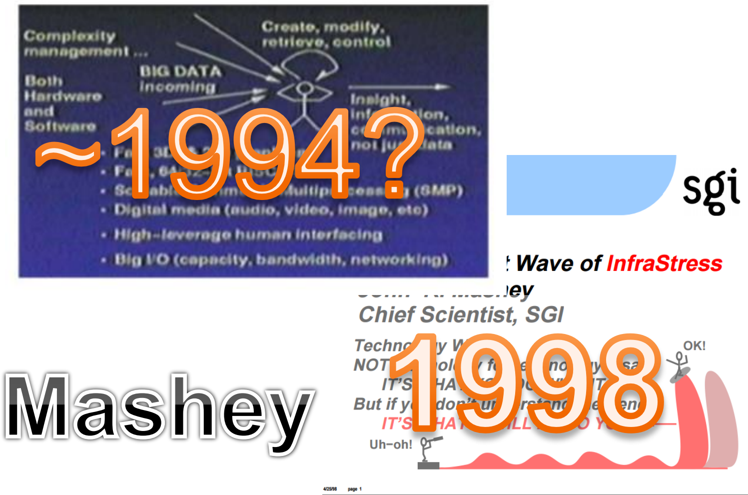


https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#533cdbfc65a1

Breve storia dei Big Data secondo Gil Press



Quadro di contesto dei Big Data. Chi fu il primo?



Application-Controlled Demand Paging for Out-of-Core Visualization

1997

Michael Cox
MRJ/NASA Ames Research Center
Microcomputer Research Labs, Intel Corporation
<mbc@nas.nasa.gov>

David Ellsworth
MRJ/NASA Ames Research Center
<ellswort@nas.nasa.gov>

Cox & Ellsworth

Abstract

In the area of scientific visualization, input data sets are often very large. In visualization of Computational Fluid Dynamics (CFD) in particular, input data sets today can surpass 100 Gbytes, and are expected to scale with the ability of supercomputers to generate them. Some visualization tools already partition large data sets into segments, and load appropriate segments as they are needed. However, this does not remove the problem for two reasons: 1) there are data sets for which even the individual segments are too large for the

1 Introduction

Visualization provides an interesting challenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of *big data*. When data sets do not fit in main memory (*in core*), or when they do not fit even on local disk, the most common solution is to acquire more resources. This *write-a-check* algorithm has two drawbacks. First, if visualization algorithms and tools are worth developing, then they are worth deploying to more production-oriented scientists and engineers who may have on their

A Personal Perspective on the Origin(s) and Development of "Big Data": The Phenomenon, the Term, and the Discipline*

→ **1997**

Diebold

Francis X. Diebold
University of Pennsylvania
fdiebold@sas.upenn.edu

First Draft, August 2012
This Draft, November 26, 2012

Application Delivery Strategies — META Group

Date: 6 February 2001 File: 949
Author: Doug Laney

3D Data Management: Controlling Data Volume, Velocity, and Variety. Current business conditions and mediums are pushing traditional data management principles to their limits, giving rise to novel, more formalized approaches.

META Trend: During 2001/02, leading enterprises will increasingly use a centralized data warehouse to define a common business vocabulary that improves internal and external collaboration. Through 2003/04, data quality and integration woes will be tempered by data profiling technologies (for generating metadata, consolidated schemas, and integration logic) and information logistics agents. By 2005/06, data, document, and knowledge management will coalesce, driven by schema-agnostic indexing strategies and portal maturity.

2001

The effect of the e-commerce surge, a rise in merger/acquisition activity, increased collaboration, and the drive for harnessing information as a competitive catalyst is driving enterprises to higher levels of consciousness about how data is managed at its most basic level. In 2001/02, historical, integrated databases (e.g., data warehouses, operational data stores, data marts), will be leveraged not only for intended analytical ... ical, and collaborative consistencies, changing economic conditions have made this job more difficult. E-commerce, in particular, has exploded data management challenges along three dimensions: volumes, ... In 2001/02, IT organizations must compile various approaches to have at their disposal for dealing with each.

Laney

Quadro di contesto dei Big Data. Voci critiche

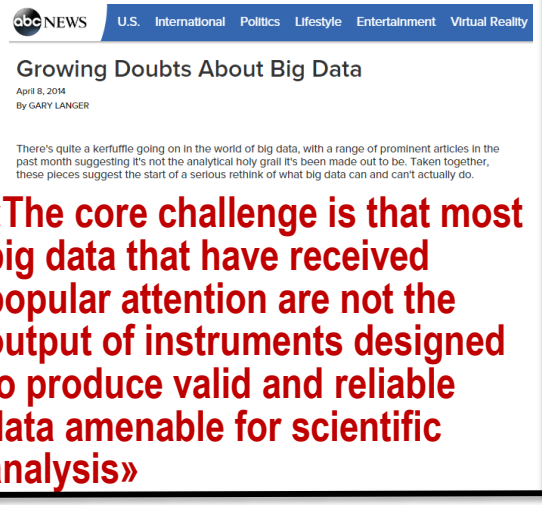
«È difficile dare un senso a grandi quantità di dati in modo significativo»

«È difficile che le imprese IT tradizionali adottino i Big Data»

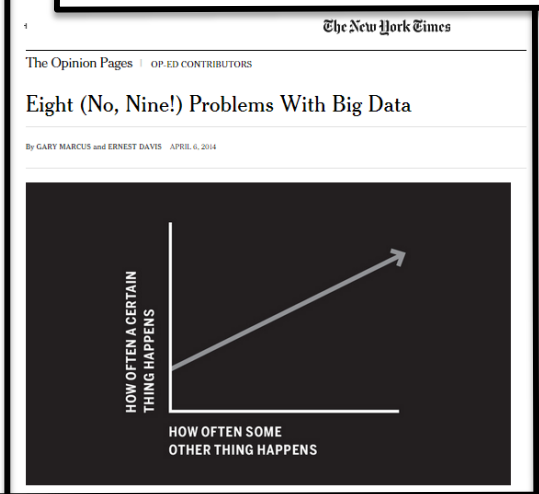
WIRED

«I Big Data sono solo una grande quantità di dati»

«As with so many buzzwords, "big data" is a vague term, often thrown around by people with something to sell»



«The core challenge is that most big data that have received popular attention are not the output of instruments designed to produce valid and reliable data amenable for scientific analysis»



«Non è vero che grandi moli di dati comportino cambiamenti nei modi in cui interagiamo con essi per esplorarli e dare loro un senso»

«Non è vero che le nuovi fonti di dati siano davvero nuove»

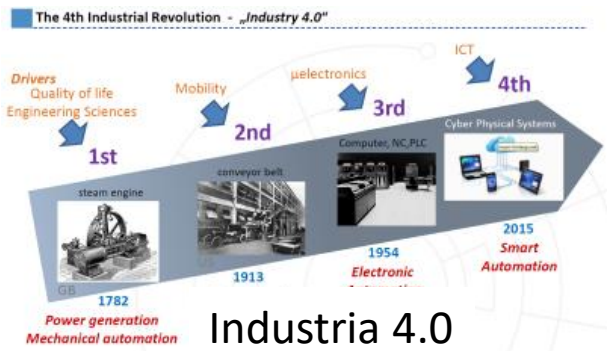
http://www.perceptualedge.com/articles/visual_business_intelligence/big_data_big_ruse.pdf; http://www.julianbrowne.com/article/viewer/big-data-deception
http://www.wired.com/2014/04/big-data-big-hype/; http://www.ft.com/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html#axzz2yQ2QQfQX
http://www.nytimes.com/2014/04/07/opinion/eight-no-nine-problems-with-big-data.html?_r=1; http://abcnews.go.com/blogs/politics/2014/04/growing-doubts-about-big-data/; https://hbr.org/2014/03/google-flu-trends-failure-shows-good-data-big-data

Cosa è cambiato dal 2001?

Statistica ufficiale

Scienza

Business

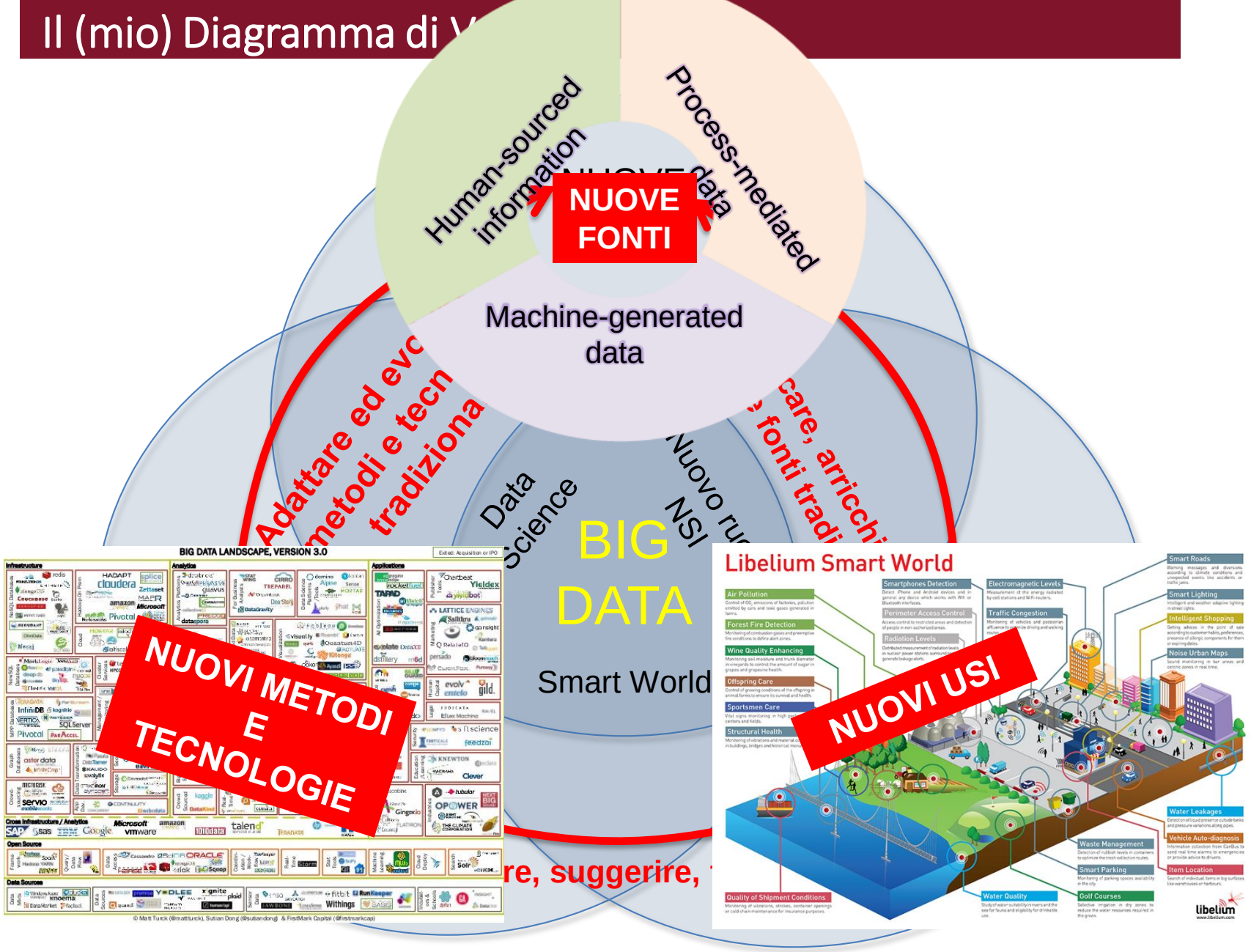


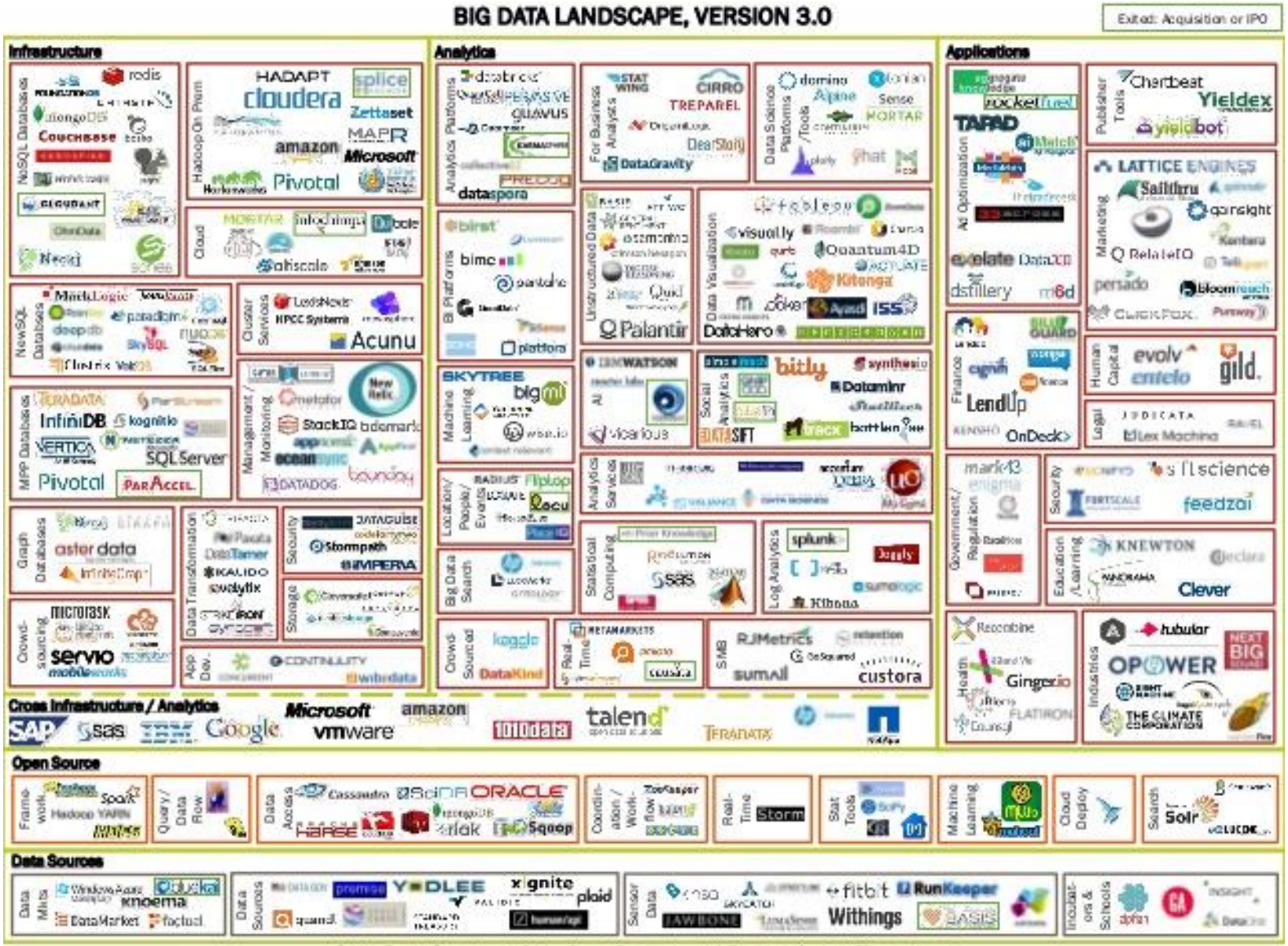
Industria 4.0

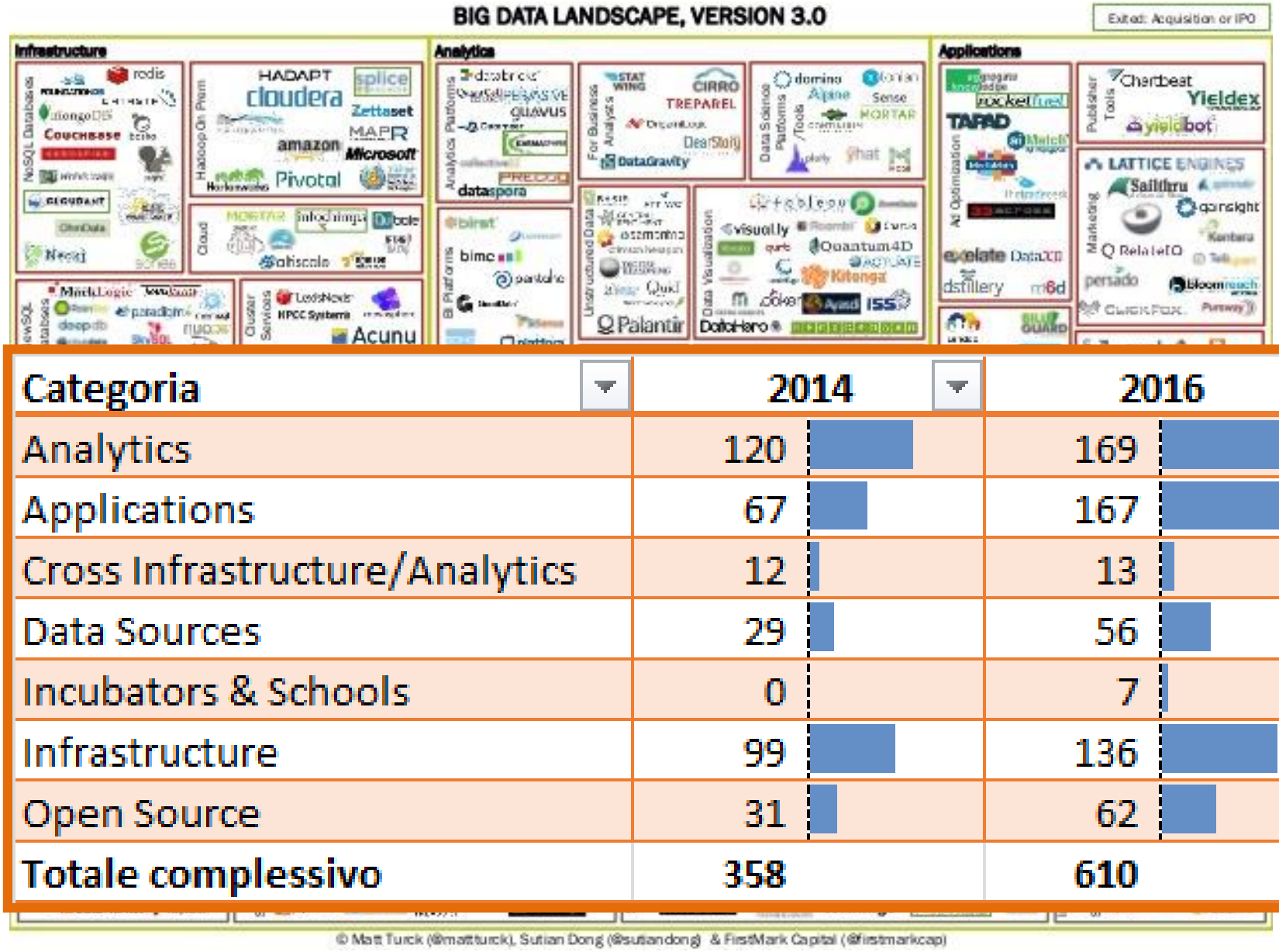
Smartness

- **NUOVE FONTI che affiancano, arricchiscono e sostituiscono le fonti tradizionali**
- **NUOVI METODI E TECNOLOGIE per adattare ed evolvere metodi e tecniche tradizionali**
- **Sono suggeriti, sostenuti e favoriti NUOVI USI dell'informazione**

Il (mio) Diagramma di V

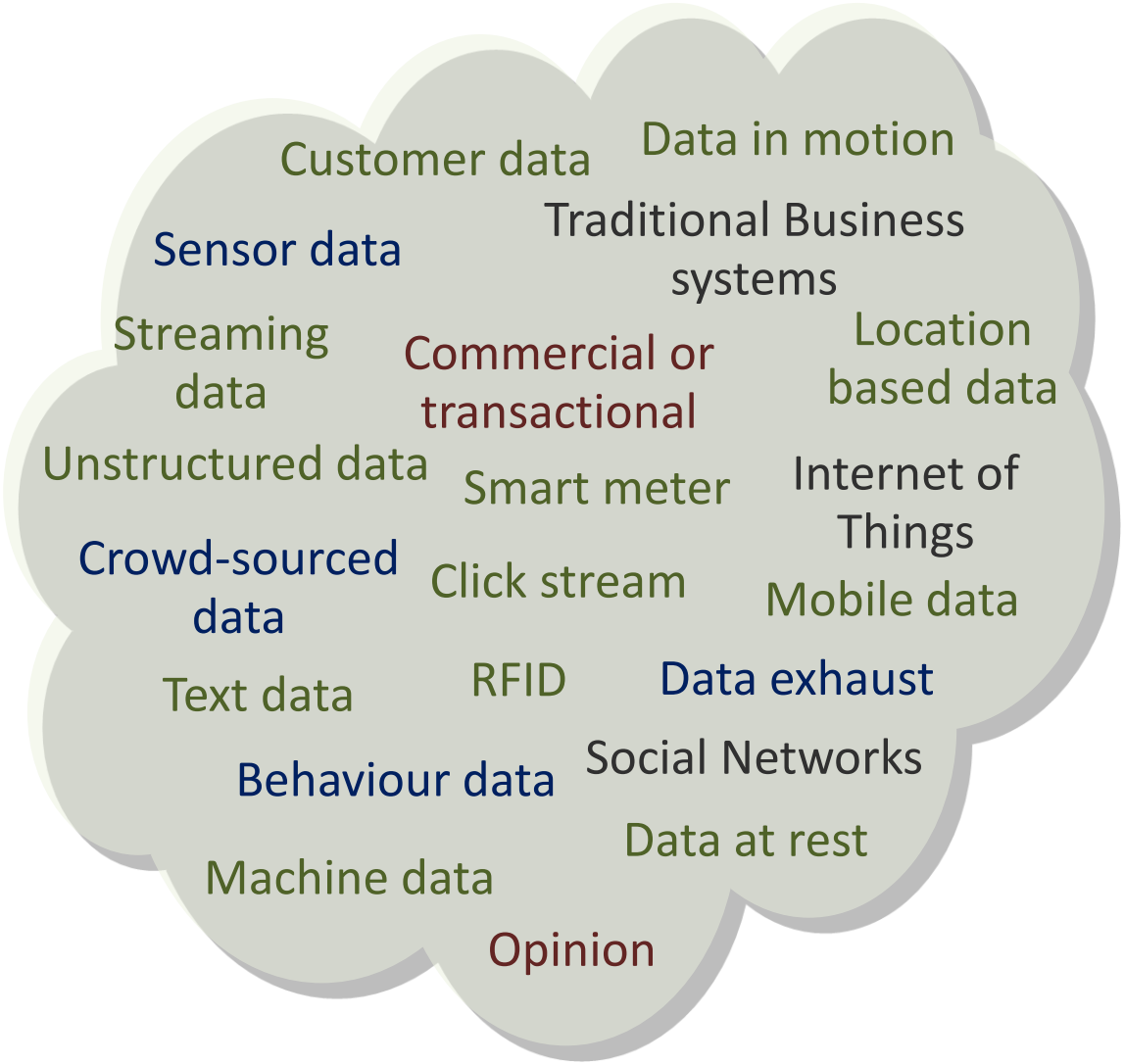



BIG DATA LANDSCAPE, VERSION 3.0
Exited: Acquisition or IPO
Infrastructure
Analytics
Applications
Cross Infrastructure / Analytics
Open Source
Data Sources
© Matt Turck (@mattturck), Sutian Dong (@sutiandong) & FirstMark Capital (@firstmarkcap)



Categoria	2014	2016
Analytics	120	169
Applications	67	167
Cross Infrastructure/Analytics	12	13
Data Sources	29	56
Incubators & Schools	0	7
Infrastructure	99	136
Open Source	31	62
Totale complessivo	**358**	**610**

Classificazione dei Big Data



Proposta nata in ambito UNECE

Process-mediated data

Human-sourced information

Machine-generated data

Classificazione delle nuove fonti Big Data

Social Networks

Dati prodotti dall'interazione con mezzi di informazione e social media o tramite dispositivi (anche mobili)
Blog, Twitter, Facebook
User-generated maps

Human-sourced information

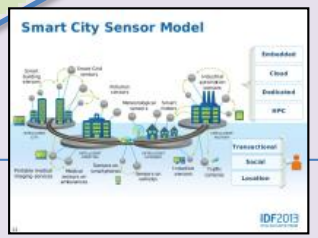
Traditional Business systems

Process-mediated data

Dati prodotti da sistemi transazionali tradizionali e in modo passivo:
Scanner data
Log ricerca,
Record medici,
Transazioni commerciali e bancarie

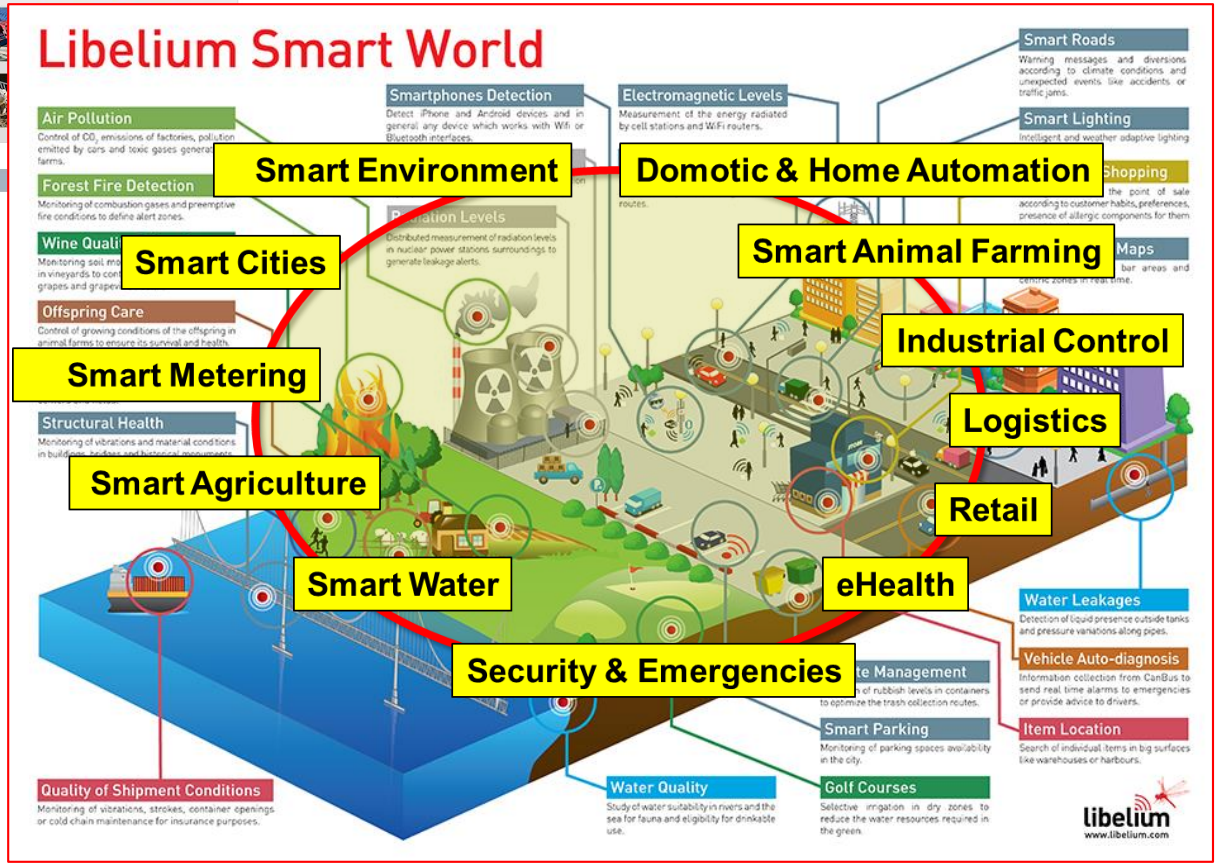
Machine-generated data

Internet of Things



Dati prodotti da sensori e macchinari utilizzati per misurare e registrare eventi e situazioni nel mondo fisico: immagini satellitari, sensori stradali e di traffico, sensori climatici e ambientali, ecc

Possibili usi «locali» dell'IoT



1. Demistificare i Big Data

2. Il contesto dei Big Data nella statistica ufficiale

3. Altri punti di vi

4.

5.

« … Big Data is also potentially very interesting as an input for Official Statistics; either for use on its own, or in combination with more traditional data sources such as sample surveys and administrative registers»

Gennaio 2013

High **L**evel **G**roup for Modernization of Statistical Production and Services

La sfida dei Big Data nella statistica ufficiale

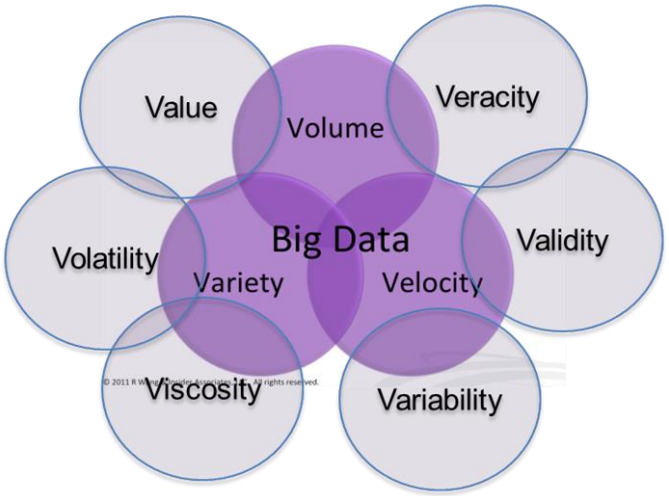
I caratteri

…della statistica ufficiale

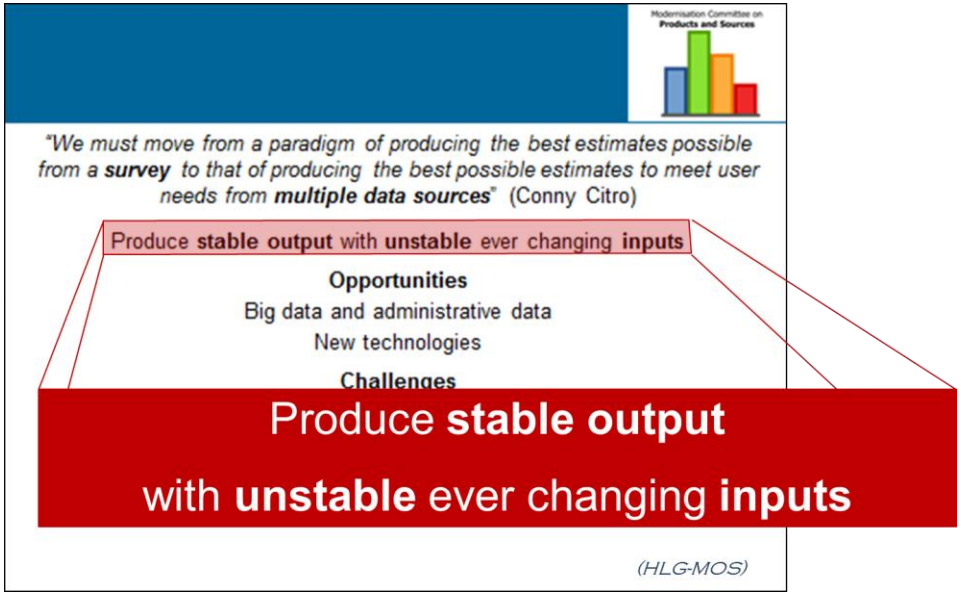
- imparzialità
- affidabilità
- obiettività
- indipendenza scientifica
- efficienza economica
- riservatezza statistica
- non comporta oneri eccessivi per gli operatori economici

CARTA DEI DIRITTI FONDAMENTALI
Art. 338 del trattato sul funzionamento dell'UE
http://europa.eu/pol/pdf/consolidated-treaties_it.pdf

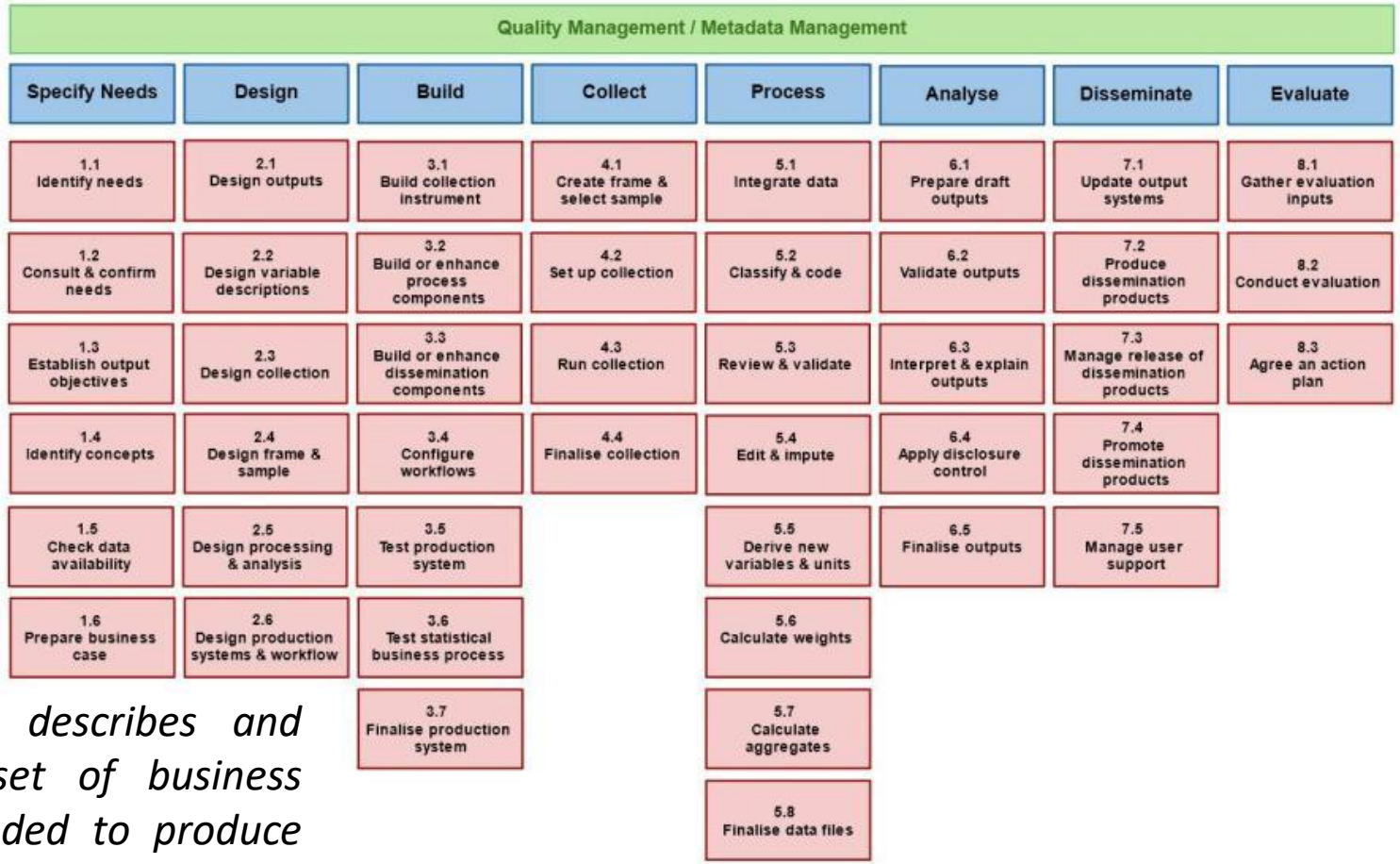
…dei Big Data



La sfida



Big Data e GSBPM

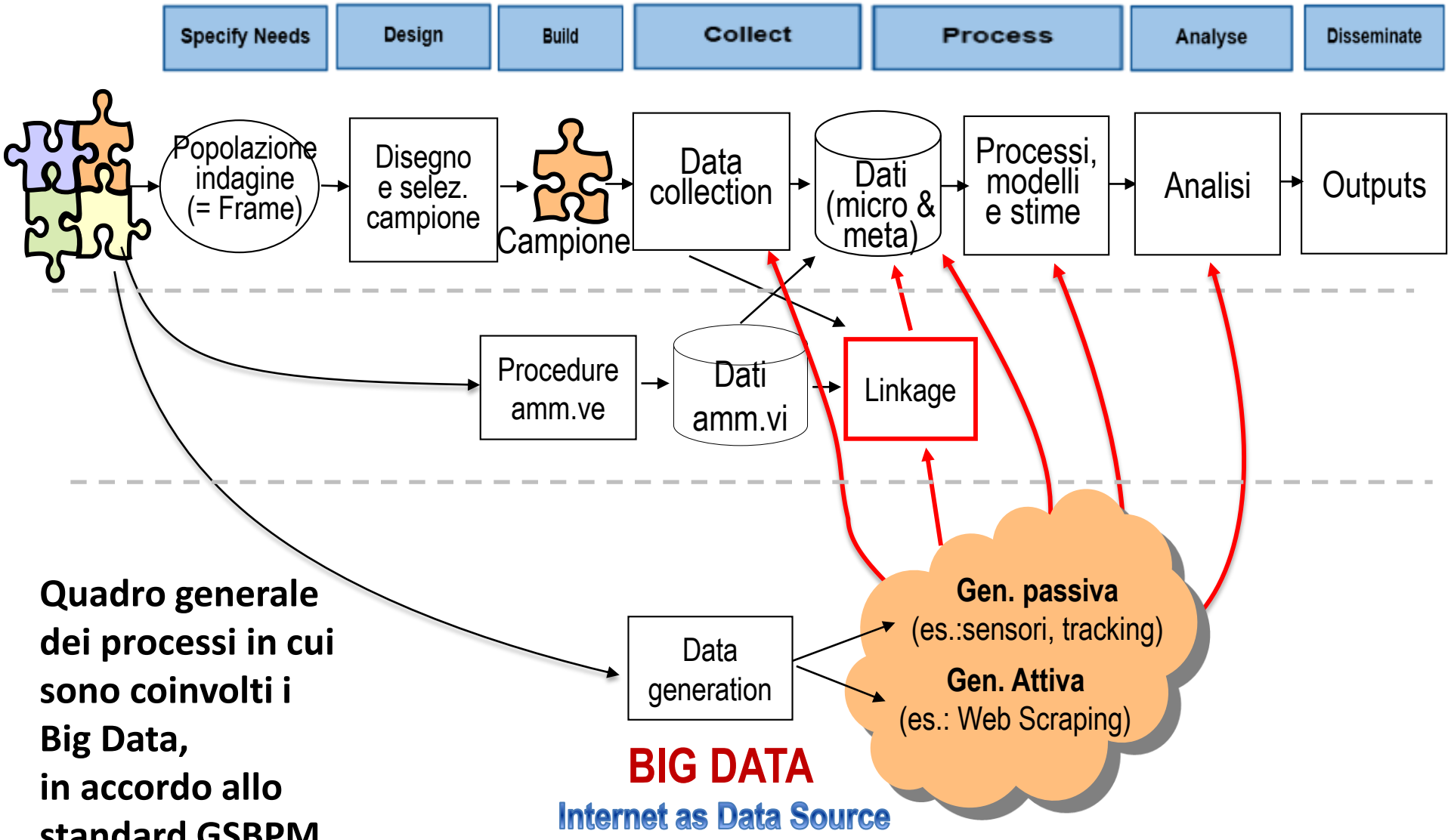


"The GSBPM describes and defines the set of business processes needed to produce official statistics.

It provides a standard framework and harmonised terminology to help statistical organisations to modernise their statistical production processes, as well as to share methods and components."

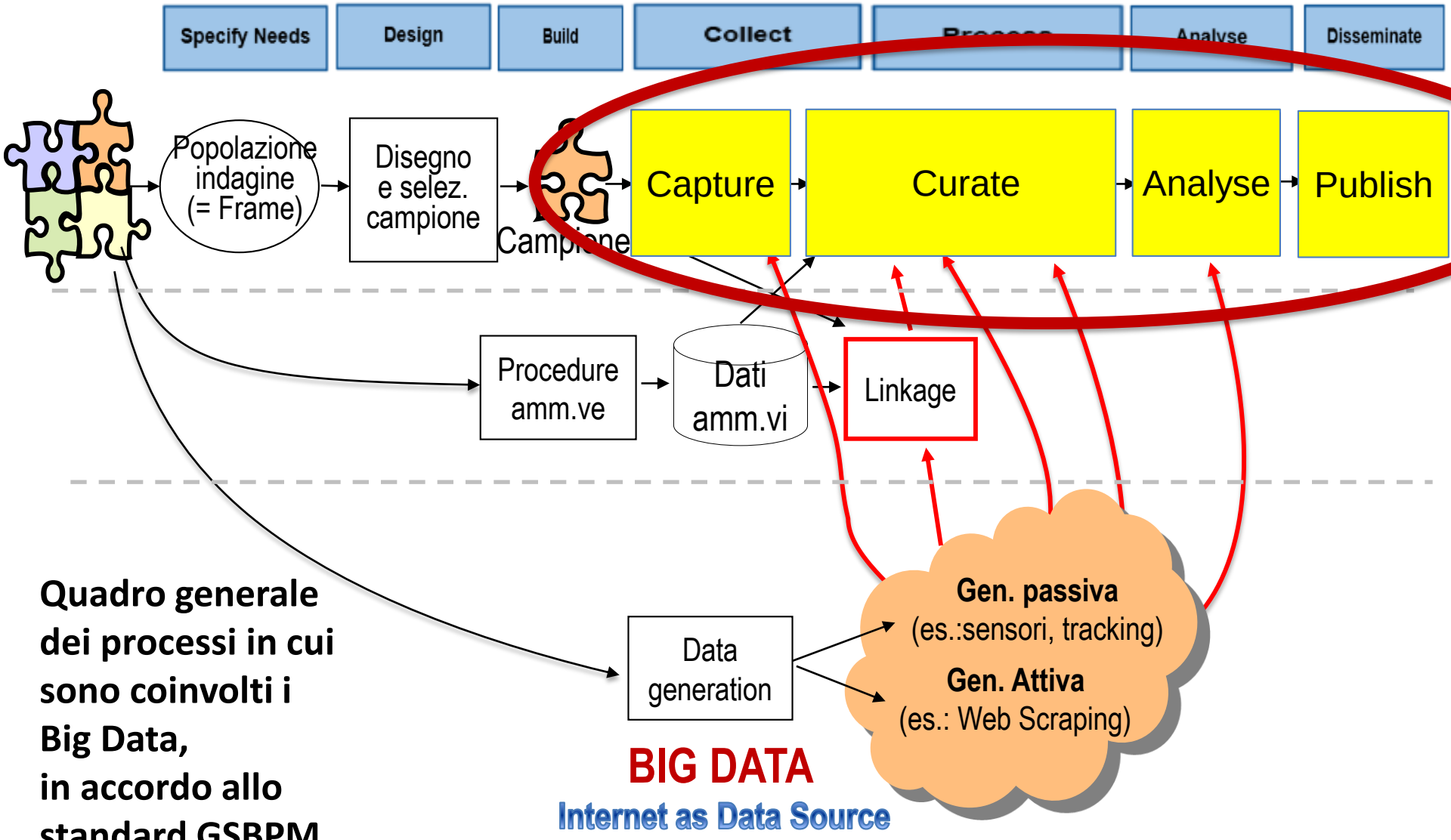
http://www1.unece.org/stat/platform/display/GSBPM/I._Introduction#I._Introduction-_Toc375051192

Big Data: possibile uso nelle fasi del processo statistico



Quadro generale dei processi in cui sono coinvolti i Big Data, in accordo allo standard GSBPM

Big Data: possibile uso nelle fasi del processo statistico



Quadro generale dei processi in cui sono coinvolti i Big Data, in accordo allo standard GSBPM

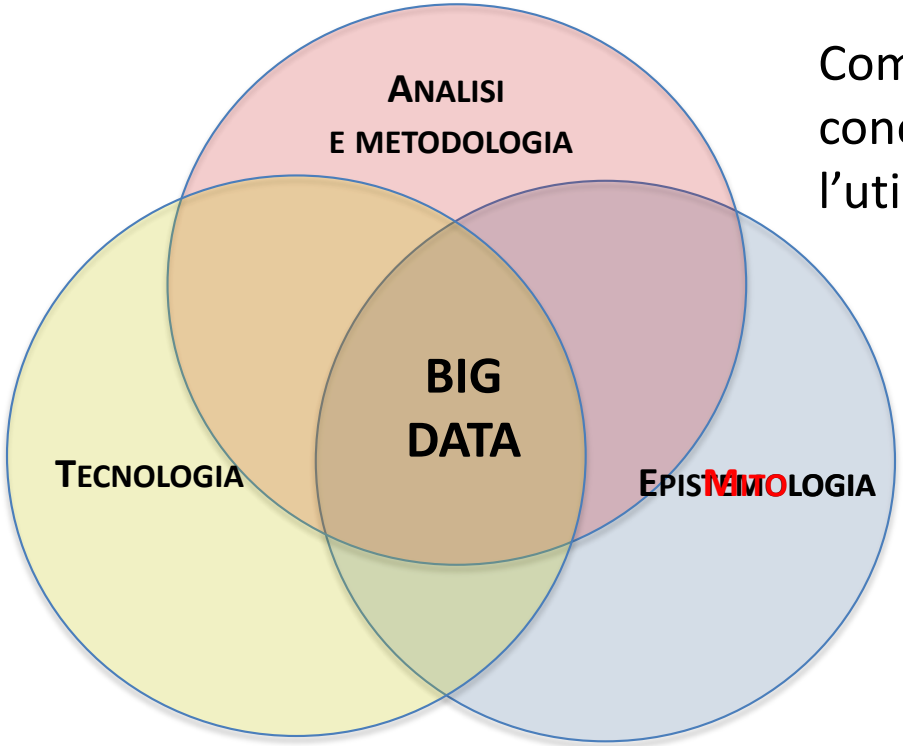
Indice

1. Demistificare i Big Data
2. Il contesto dei Big Data nella statistica ufficiale
3. **Altri punti di vista**
4. Problemi aperti
5. Esperienze correnti in Istat

- **La ricerca sociale**

- **Il legame col territorio**

La tematica dei Big Data nella ricerca sociale



Come metodi e condizioni della conoscenza scientifica cambino con l'utilizzo di nuove fonti informative

La diffusa convinzione che i grandi set di dati offrono una forma più alta di Intelligenza e conoscenza che possono generare intuizioni in precedenza impossibili, con l'aura della verità, dell'obiettività e dell'accuratezza

[BOYD, CRAWFORD]

https://www.danah.org/papers/2012/BigData-ICS-Draft.pdf

La tematica dei Big Data nella ricerca sociale

«I dati non richiedono sforzi specifici per essere raccolti, essendo il sottoprodotto digitale di operazioni di routine svolte entro il sistema»

- Gli algoritmi come principi ordinatori dei più diversi ambiti della vita sociale
- I dati come scatola nera da aprire

- Trasformare molti aspetti della vita delle persone in dati digitali
- Trasformare queste informazioni in nuove forme di valore

IL POTERE DEI GRANDI NUMERI E LA GOVERNANCE BY NUMBERS

IL PROBLEMA DELLA DISINTERMEDIAZIONE

Big Data nella ricerca sociale

LA DATIFICAZIONE DEL MONDO

Mise en données du monde

RAW DATA VS. COOKED DATA

CAUSALITÀ VS. CORRELAZIONE

«Raw data is both an oxymoron and a bad idea; to the contrary, data should be cooked with care»

«Correlation doesn't mean causation»

La tematica dei Big Data nella ricerca sociale

Use of more than one method of data collection or research.
Mixed methods research is more specific in that it includes the mixing of qualitative and quantitative data, methods, methodologies, and/or paradigms.

MULTIMETHODOLOGY E MIXED METHODS

GROUNDED THEORY

Osservazione ed elaborazione teorica procedono di pari passo, in un'interazione continua. Il ricercatore scopre la teoria nel corso della ricerca empirica

Big Data nella ricerca sociale

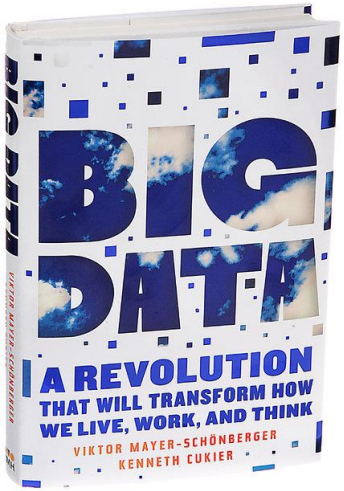
PARADIGMA INTERPRETATIVO

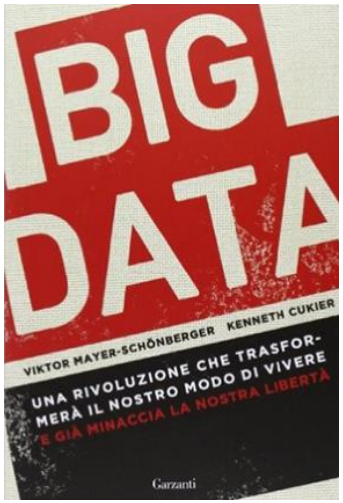
SCIENCE AND TECHNOLOGY STUDIES

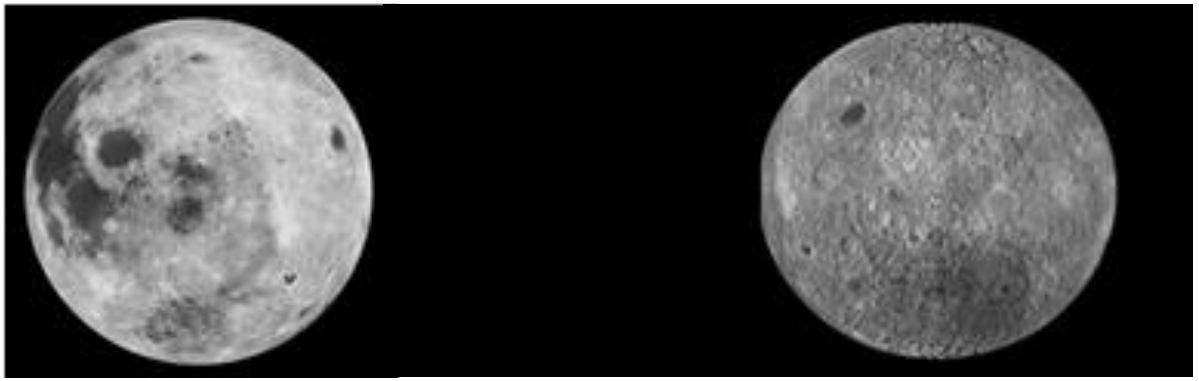
La realtà sociale non può essere semplicemente osservata ma necessita di interpretazione.
Comprendere significa cogliere l'intenzionalità dell'agire umano, attraverso il senso soggettivo attribuito dall'individuo al proprio comportamento.

- Relazioni fra innovazioni scientifiche e tecnologiche, partendo dal presupposto che ambedue sono socialmente costruite e che la società è essa stessa un aggregato sociotecnico.
- Analisi degli effetti, rischi, ridefinizione dei parametri sociali

Le due facce dei Big Data







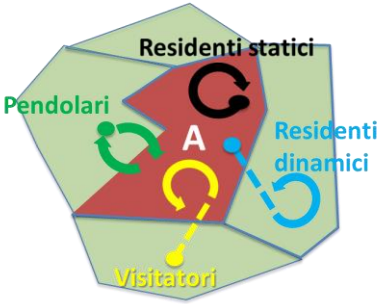
- Le persone come "somma" delle relazioni sociali, delle interazioni online e delle connessioni con i *contenuti* delle azioni che svolgono
- La conoscenza di una persona passa attraverso la penombra (il più larga possibile) di dati che la circonda

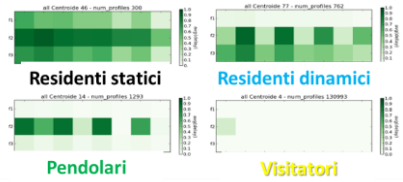
- **BIG BROTHER**:
  La privatezza diventa più difficile da gestire
- **MINORITY REPORT**:
  "[…] predictions seem so accurate that people can be arrested for crimes before they are committed"

Big Data e territorio

Persons & Places

Popolazione che insiste su un territorio





Utilizzo di dati GSM e applicazione di modelli distinguere tra residenti e pendolari dinamici (non possibile con i soli dati amministrativi)

Disegno di nuovi «territori» in base alla mobilità della popolazione

Analisi degli spostamenti tramite GPS per delimitare le aree prevalenti entro cui si svolgono le attività. La densità di traffico permette di costruire confini geografici

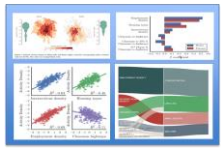
[Rinzivillo et al. KI-Künstliche Intelligenz, 26 (2012)]

Fonti nuove e tradizionali per analizzare le condizioni della vita urbana sulla base del grado di *vita pedonale*

Death vs. **Life** (Jacobs, 1961)

La struttura urbana «spiega» la vita urbana (77%)

1) *Mixed land uses*
2) *Small blocks*
3) Diversificazione edilizia
4) Concentrazione equilibrata di persone ed edifici



[Fondazione Bruno Kessler, University of Trento, Sorbonne Universités, Telecom Italia, Bell Labs Cambridge]

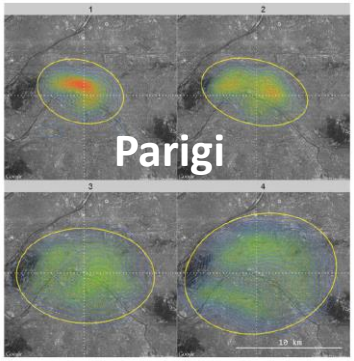
Dati di sensori di telefonia mobile per monitorare l'attività umana nelle città

Nuovi indicatori riferiti alla vita nelle città

Tempo sociale delle attività

Differenze spaziali per misurare il «giorno attivo»

Mappe di densità calcolate per quartili di durata della vita sociale (approssimata dalla *vitalità* delle celle di telefonia mobile)



Indice

Six Provocations for Big Data

1. Automating Research Changes the Definition of Knowledge

2. Claims to Objectivity and Accuracy are Misleading

3. Bigger Data are Not Always Better Data

4. Not All Data Are Equivalent

5. Just Because it is Accessible Doesn't Make it Ethical

6. Limited Access to Big Data Creates New Digital Divides

Six provocations for Big Data

Danah Boyd, Kate Crawford
Samira Shaikh, Veena Ravishankar

Big data is at its best when analyzing things that are extremely common, but often falls short when analyzing things that are less common

Although big data is very good at detecting correlations, especially subtle correlations that an analysis of smaller data sets might miss, it never tells us which correlations are meaningful

Big data can work well as an adjunct to scientific inquiry but rarely succeeds as a wholesale replacement

Big data is prone to giving scientific-sounding solutions to hopelessly imprecise questions.

Many tools that are based on big data can be easily gamed

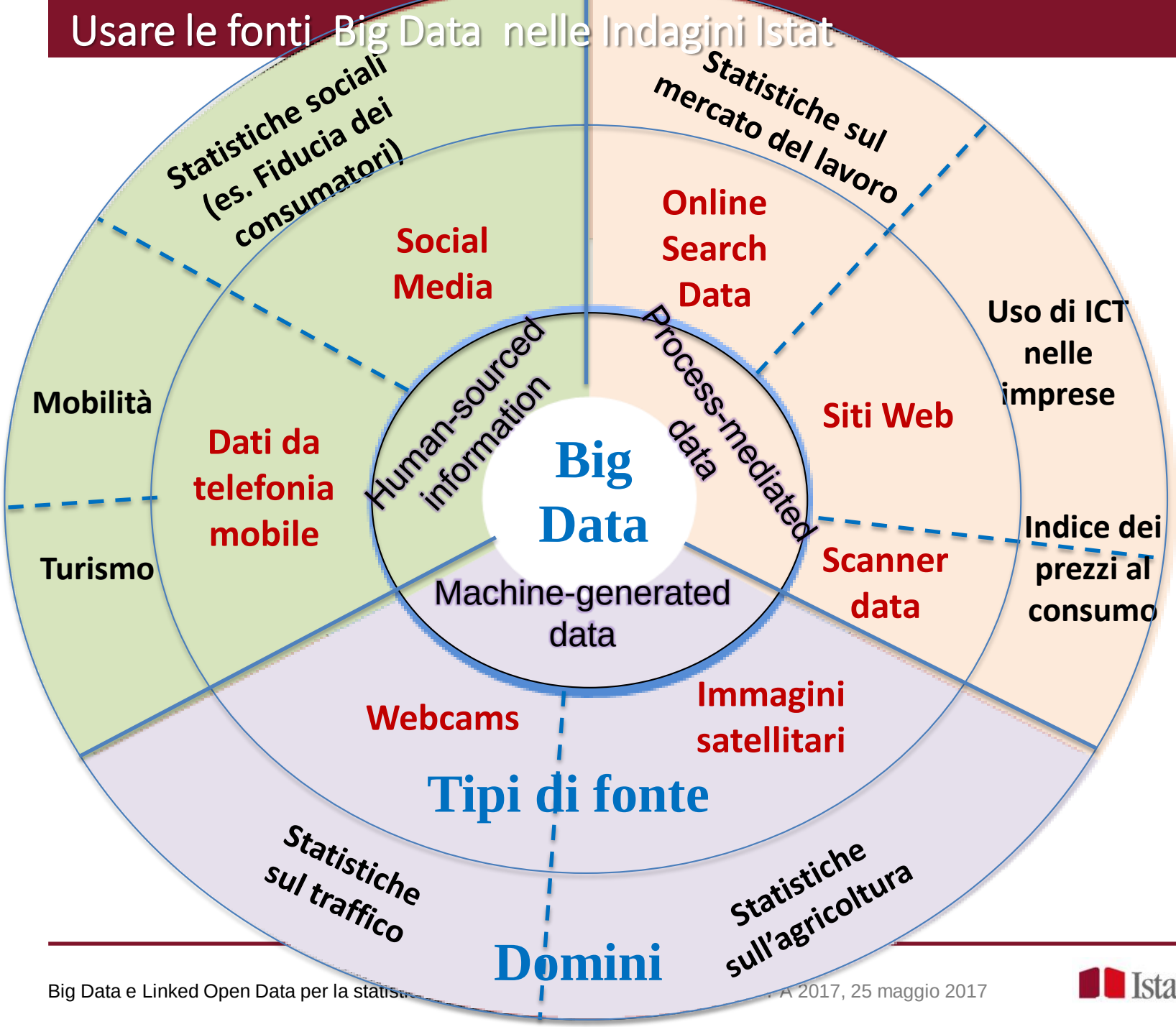
Absent careful supervision, the magnitudes of big data can greatly amplify such errors.

Whenever the source of information for a big data analysis is itself a product of big data, opportunities for vicious cycles abound

Even when the results of a big data analysis aren't intentionally gamed, they often turn out to be less robust than they initially seem

Indice

Usare le fonti Big Data nelle Indagini Istat



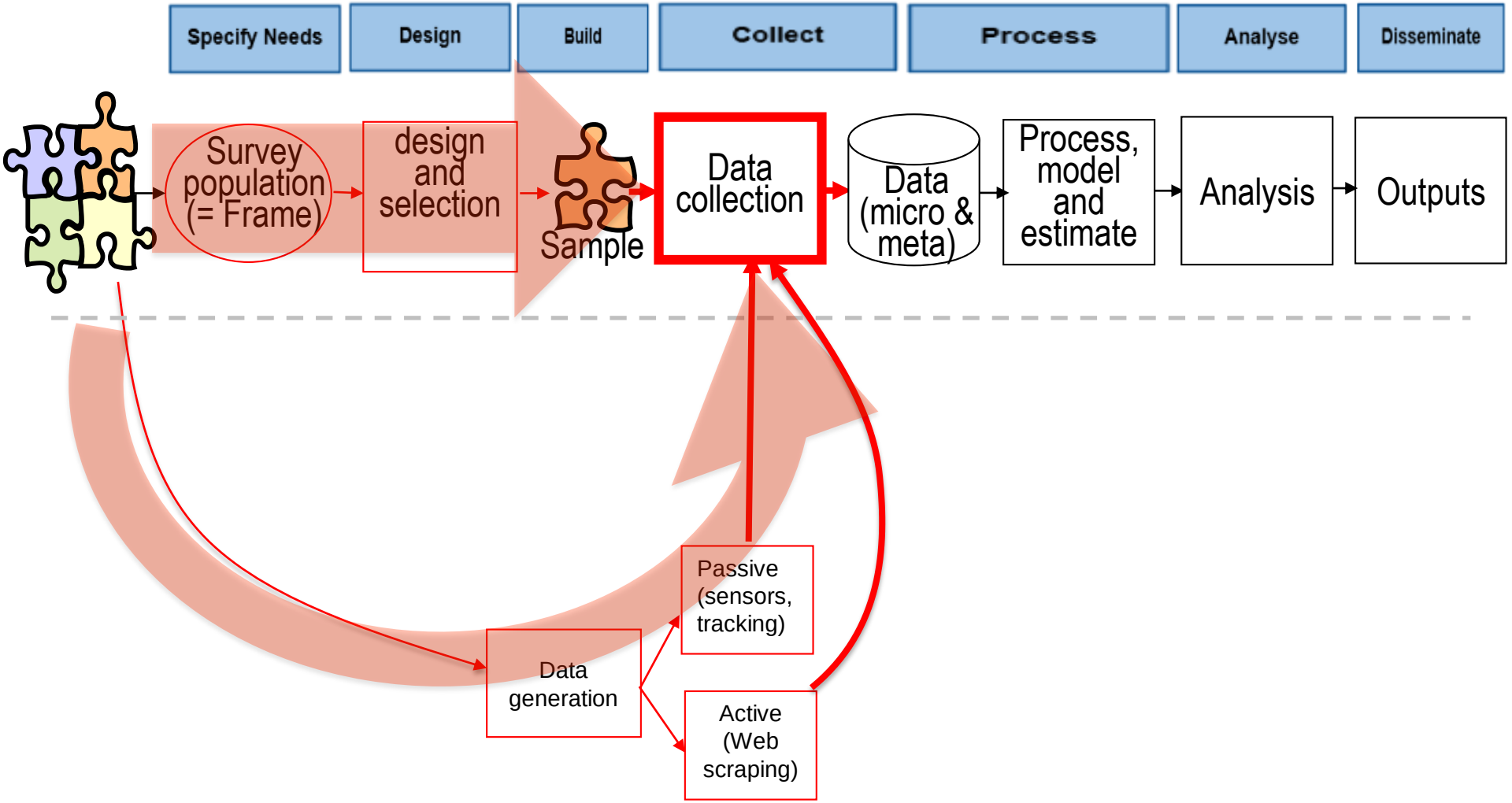
Big Data: possibili scenari di utilizzo nel processo

- Scenario 1: **Uso delle fonti Big limitato alla raccolta dati**
  Vengono raccolte le stesse informazioni utilizzate nel processo statistico tradizionale, ma accedendo direttamente alla fonte Big e senza interventi significativi su approcci/ tecniche di analisi

- Scenario 2: **Uso delle fonti Big in combinazione o integrato con le altre fonti di dati (da indagine e amministrativi)**
  Per le stime si utilizzano sia i dati da fonte tradizionale sia Big Data, dopo opportuno e specifico passo di integrazione (RL) e/o trattamento (NLP, Text Mining, ML, ecc.)

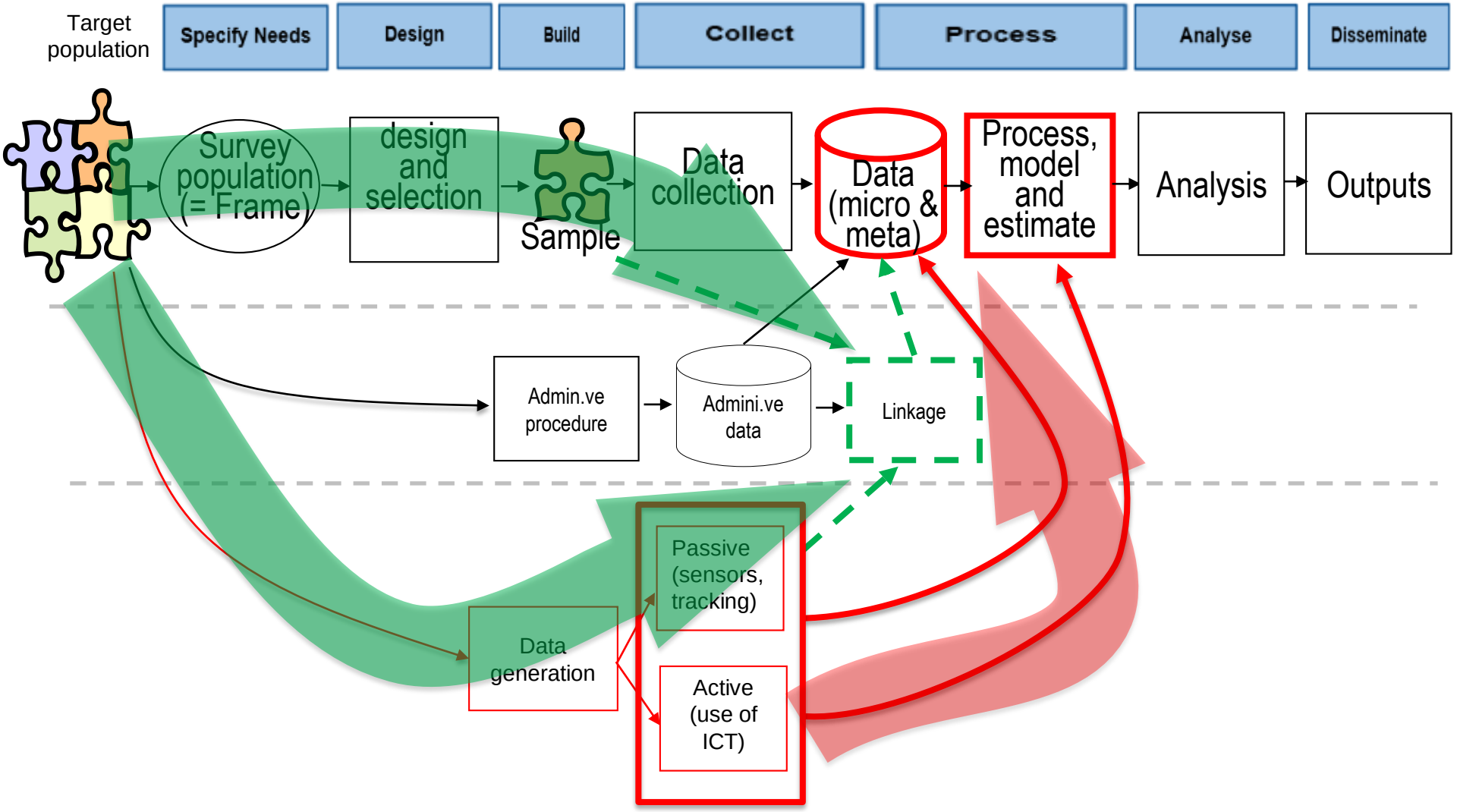
- Scenario 3: **Uso delle fonti Big in sostituzione delle (o alternativo rispetto alle) fonti tradizionali**
  Per le stime si usano solo Big Data e non dati di indagine, con individuazione di specifiche tecniche e nuovi metodi di analisi/trattamento lungo tutto il processo statistico (da RD in poi)

Scenario 1: tecniche alternative di data collection



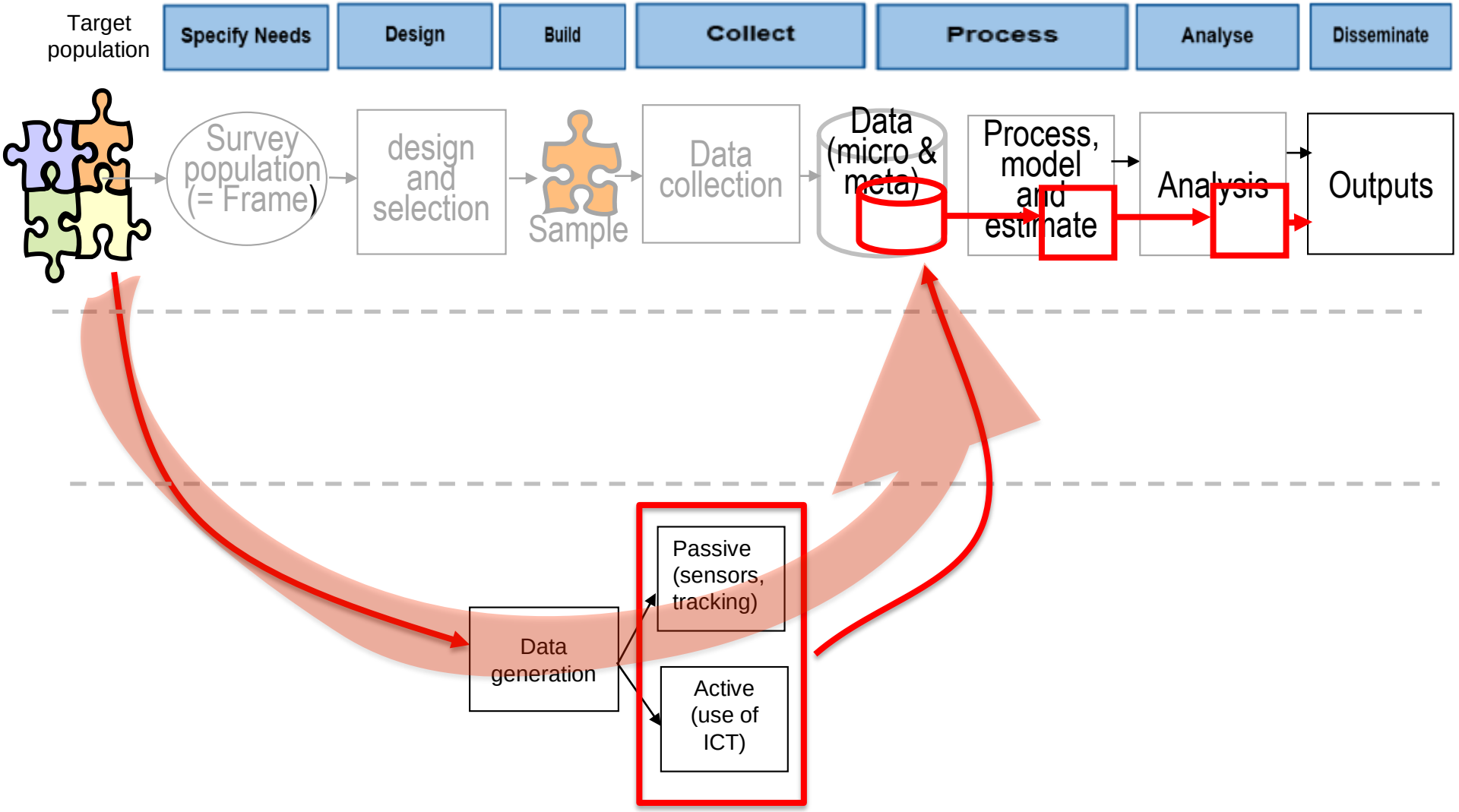
Istat

Scenario 2: uso integrato

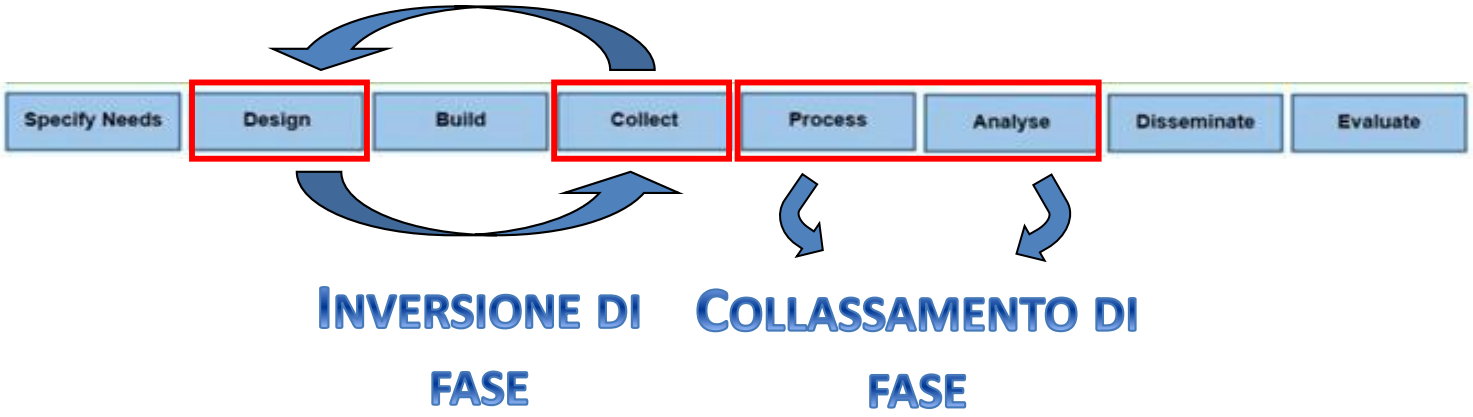


Istat

Scenario 3: uso in sostituzione delle fonti tradizionali



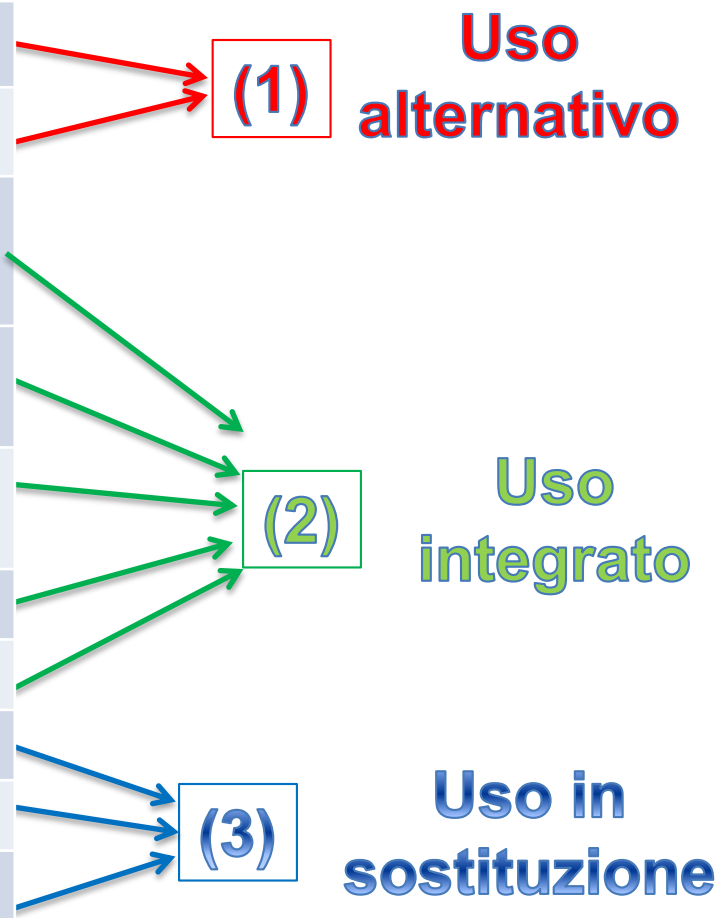
Impatto dei Big Data nelle fasi del processo statistico



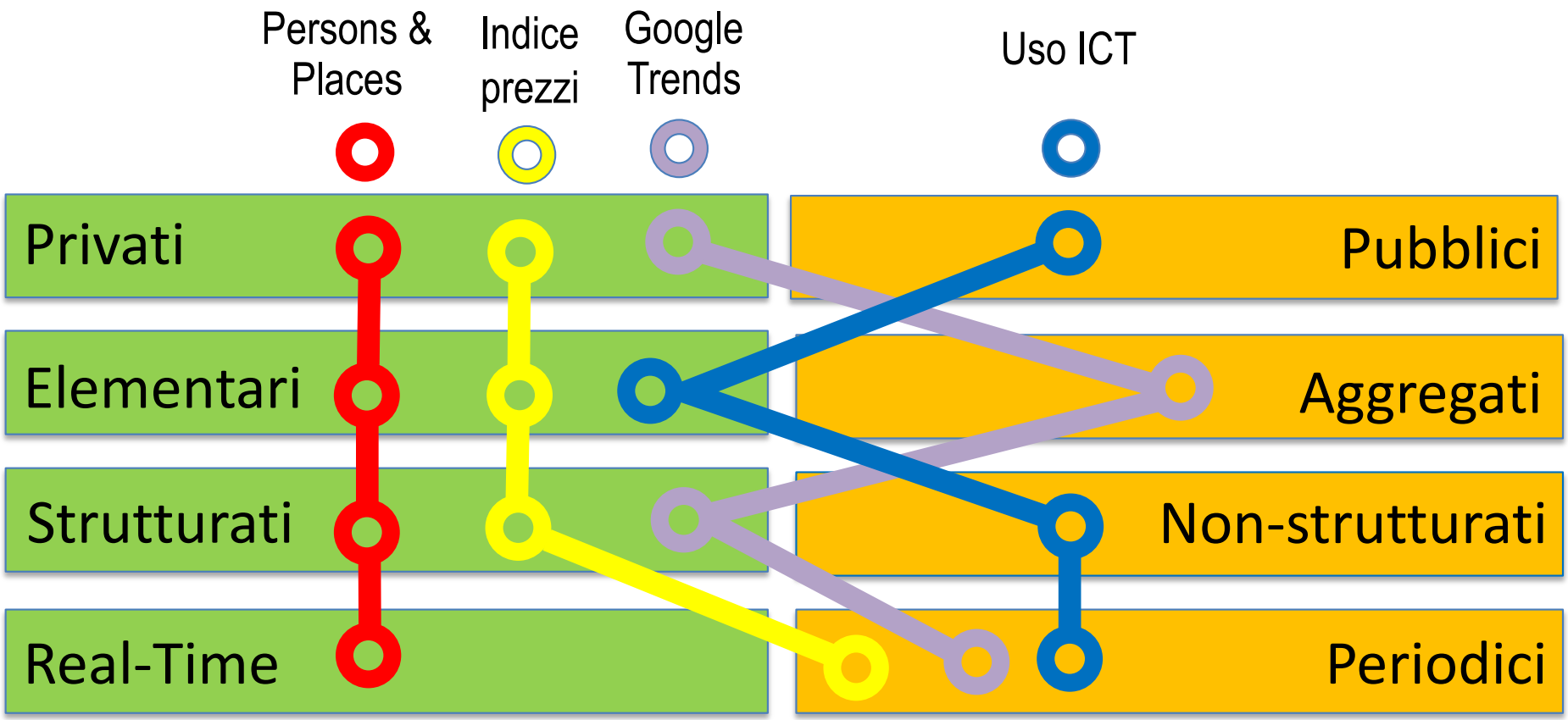
- Possibile inversione di alcune fasi (*Design* e *Collect*)
- La fase di *collezione dati* può a volte essere sostituita da quella di *generazione dati*
- Possibile collassamento delle fasi di *Process* e *Analyse* (possono avvalersi degli stessi metodi)
- Altre fasi (ad es. *Dissemination)* non sono ancora coinvolte

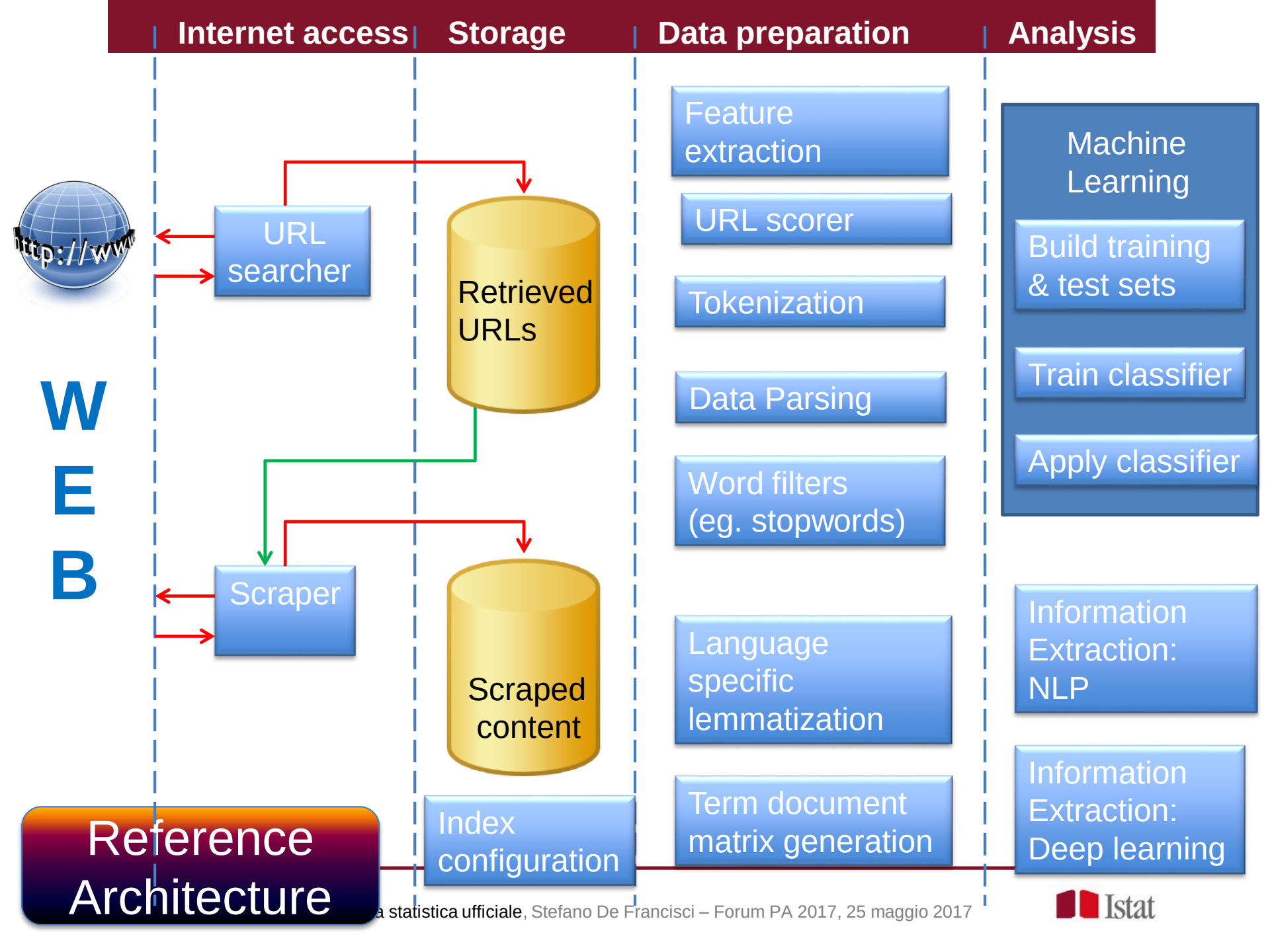
Big Data: possibili scenari e applicazioni

Domini di interesse	Tipi di fonte
Indice prezzi al consumo	Scanner data
	Internet data (web scraping)
ICT nelle Imprese	Internet data (web scraping)
Agriturismo	
Mercato del lavoro	Online Search data (Google queries)
Grado di fiducia dei consumatori	Social Media
Statistiche sull'agricoltura	Immagini satellitari
Mobilità e popolazione	Dati di telefonia cellulare
Turismo	Dati di telefonia cellulare
Traffico	Traffic webcams
Spese e consumi energetici	Smart meters

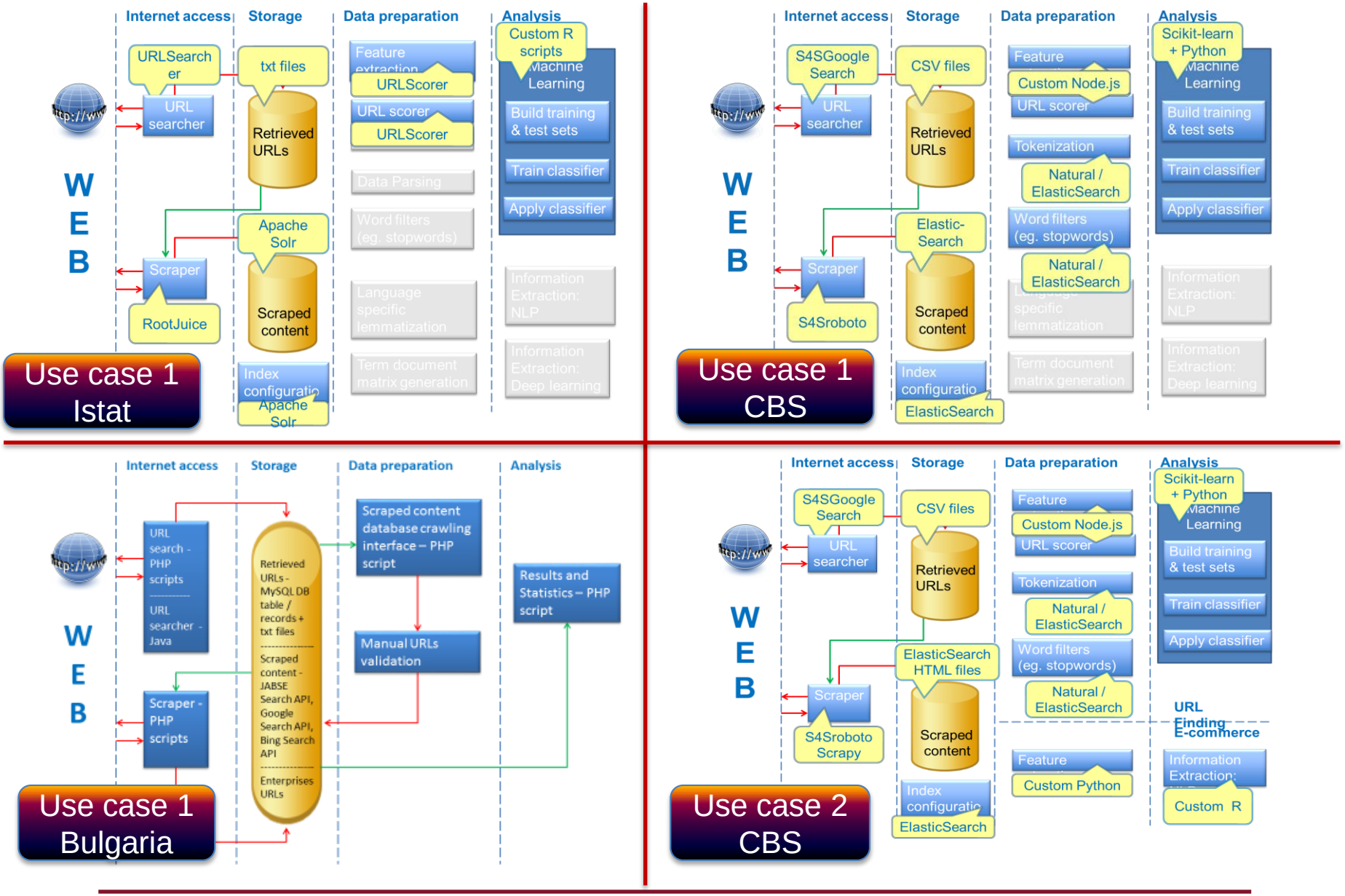


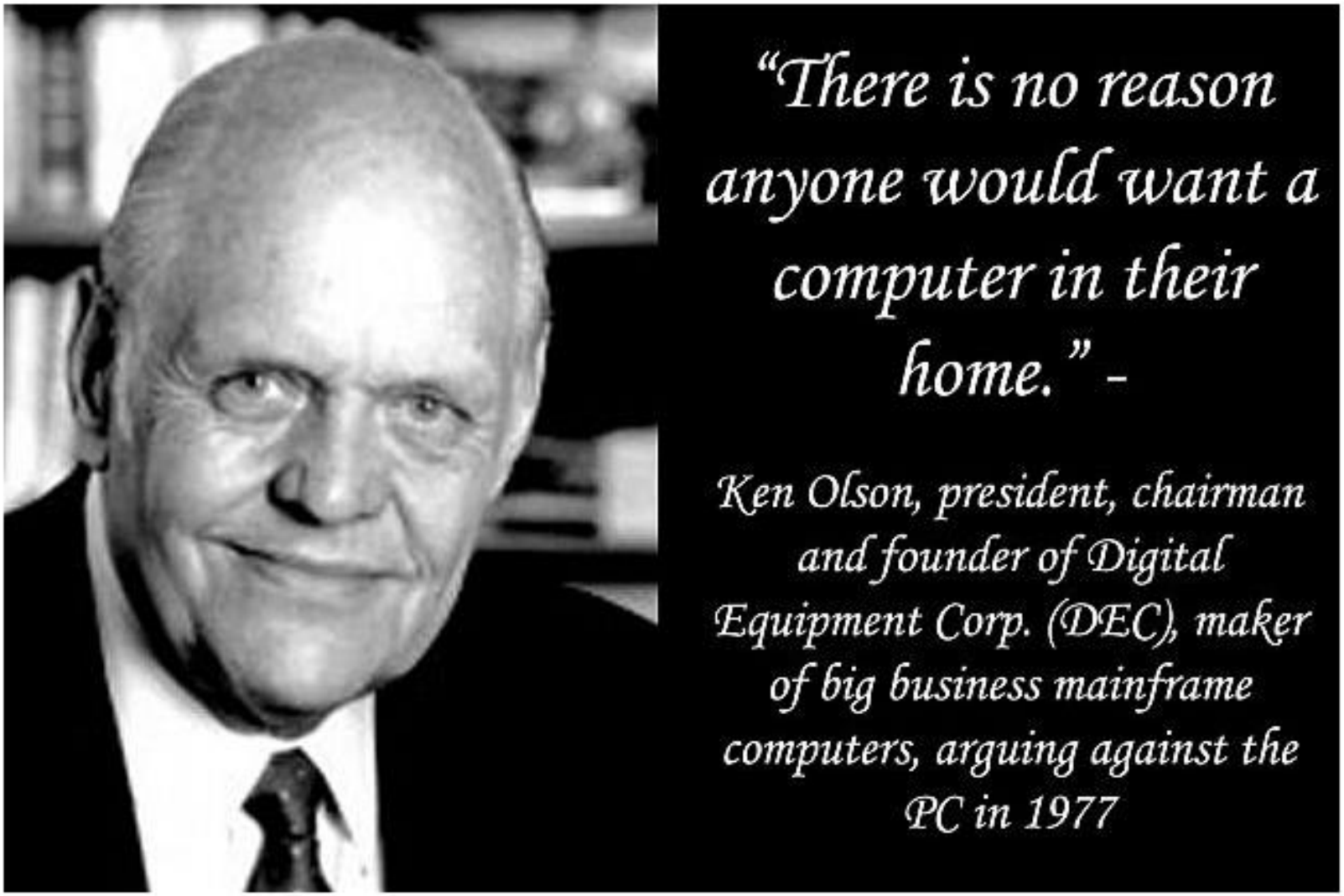
Caratteristiche dei Big Data utilizzati nelle sperimentazioni




Internet access
Storage
Data preparation
Analysis
WEB
URL searcher
Retrieved URLs
Scraper
Scraped content
Index configuration
Feature extraction
URL scorer
Tokenization
Data Parsing
Word filters (eg. stopwords)
Language specific lemmatization
Term document matrix generation
Machine Learning
Build training & test sets
Train classifier
Apply classifier
Information Extraction: NLP
Information Extraction: Deep learning
Reference Architecture

Istat


Internet access
Storage
Data preparation
Analysis
URLSearcher
URL searcher
txt files
Retrieved URLs
Apache Solr
Scraper
RootJuice
Scraped content
Index configuratio
Apache Solr
Feature extraction
URLScorer
URL scorer
URLScorer
Data Parsing
Word filters (eg. stopwords)
Language specific lemmatization
Term document matrix generation
Custom R scripts
Machine Learning
Build training & test sets
Train classifier
Apply classifier
Information Extraction: NLP
Information Extraction: Deep learning
W E B
Use case 1 Istat
S4SGoogle Search
URL searcher
CSV files
Retrieved URLs
Elastic-Search
Scraper
S4Sroboto
Scraped content
Index configuratio
ElasticSearch
Feature
Custom Node.js
URL scorer
Tokenization
Natural / ElasticSearch
Word filters (eg. stopwords)
Natural / ElasticSearch
Scikit-learn + Python
Machine Learning
Build training & test sets
Train classifier
Apply classifier
Information Extraction: NLP
Information Extraction: Deep learning
Use case 1 CBS
URL search - PHP scripts
URL searcher - Java
Scraper - PHP scripts
Retrieved URLs - MySQL DB table / records + txt files
Scraped content - JABSE Search API, Google Search API, Bing Search API
Enterprises URLs
Scraped content database crawling interface – PHP script
Manual URLs validation
Results and Statistics – PHP script
Use case 1 Bulgaria
S4SGoogle Search
URL searcher
CSV files
Retrieved URLs
ElasticSearch HTML files
Scraper
S4Sroboto Scrapy
Scraped content
Index configuratio
ElasticSearch
Feature
Custom Node.js
URL scorer
Tokenization
Natural / ElasticSearch
Word filters (eg. stopwords)
Natural / ElasticSearch
Feature
Custom Python
Scikit-learn + Python
Machine Learning
Build training & test sets
Train classifier
Apply classifier
URL Finding E-commerce
Information Extraction:
Custom R
Use case 2 CBS


"There is no reason anyone would want a computer in their home." -
Ken Olson, president, chairman and founder of Digital Equipment Corp. (DEC), maker of big business mainframe computers, arguing against the PC in 1977

FINE

Grazie

stefano.defrancisci@istat.it