POLITECNICO DI MILANO

door data mining and optimization research group

Big data analytics

Come orientarsi nel labirinto di dati

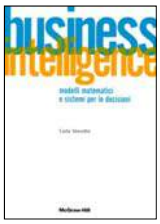
12° Forum europeo "Manfredo Golfieri"
L'innovazione per la competitività

Reggio Calabria, 15 dicembre 2015

door@polimi.it - www.door.polimi.it - via Lambruschini 4b, 20156 Milano

Le competenze del team door

- **World-class team nella realizzazione di predictive analytics e ottimizzazione per progetti di:**
  - **Big data analytics & business intelligence**
  - **ottimizzazione di azioni di targeting per campagne di marketing**
  - **market segmentation, customer profitability e basket analysis**
  - **web mining e social network analysis**
  - **demand forecasting**
  - **identificazione di frodi in ambito creditizio, assicurativo, fiscale, monetario, sanitario**
  - **credit scoring e risk management**
  - **supply chain optimization**



Carlo Vercellis
Full Professor
of Computer Science
Politecnico di Milano
School of Management
Via Lambruschini 4b
20156 Milano

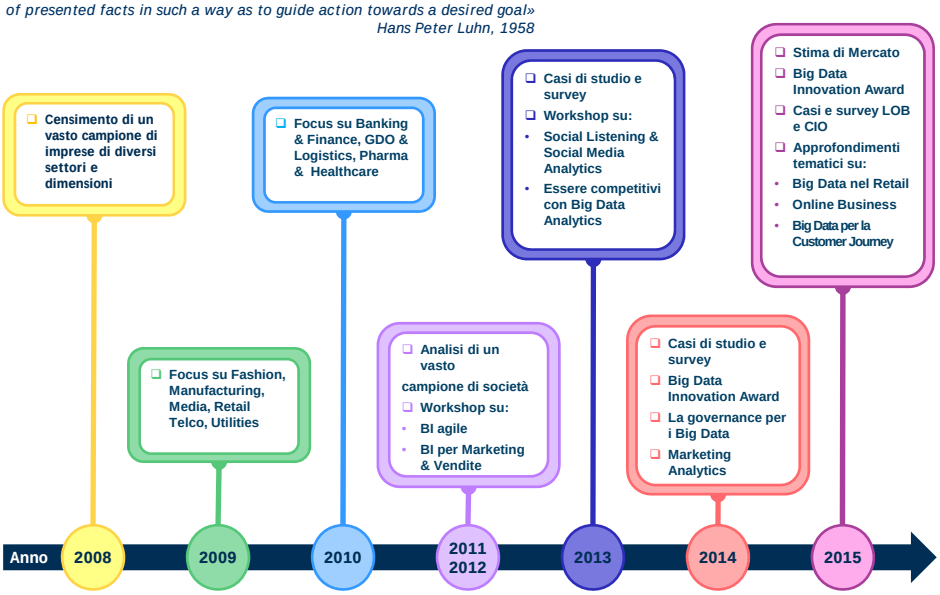
carlo.vercellis@polimi.it
www.door.polimi.it

door group

door@polimi.it
www.door.polimi.it

L'Osservatorio Big Data Analytics & Business Intelligence

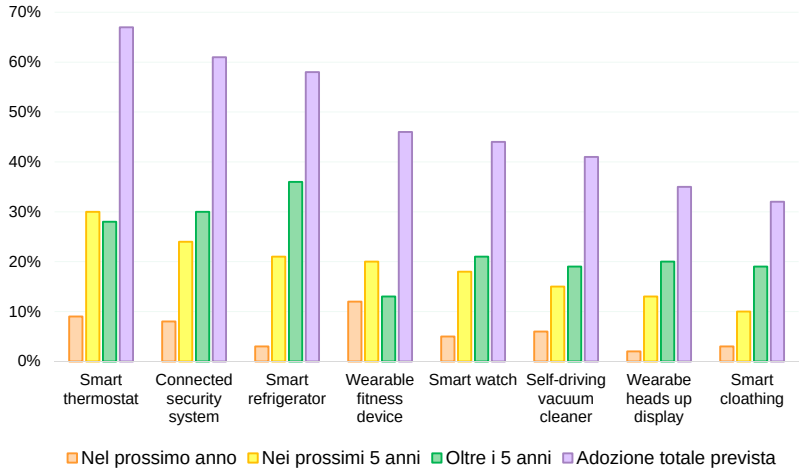
Business Intelligence: «the ability to apprehend the interrelationships of presented facts in such a way as to guide action towards a desired goal»
Hans Peter Luhn, 1958



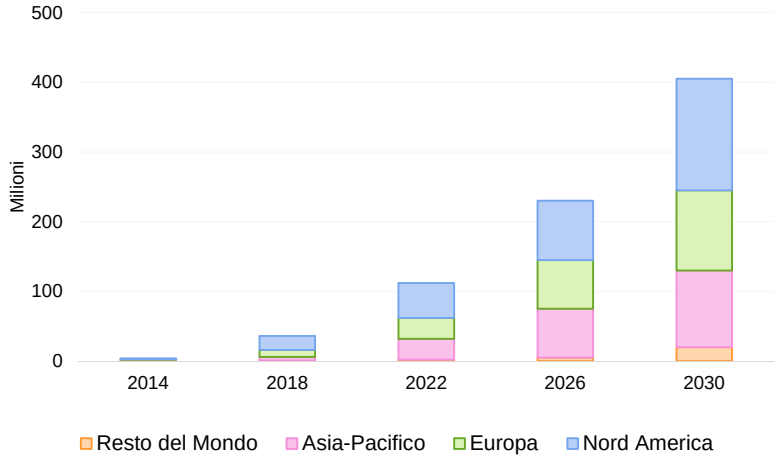
La formazione

2014 : i mobile devices hanno superato la popolazione

30 trilioni di smart devices per il 2020

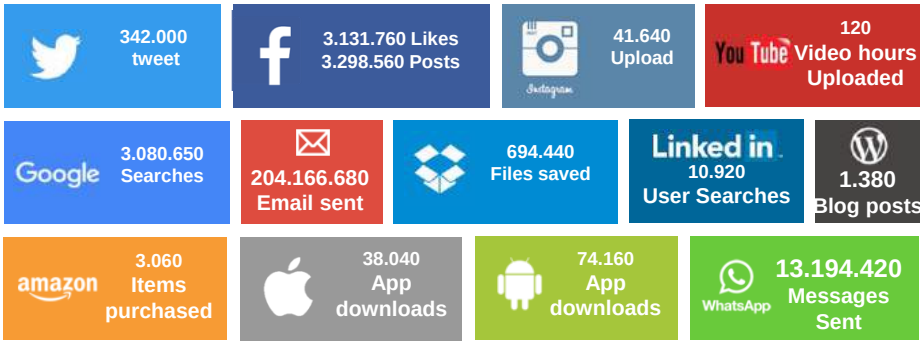


…il mercato «connected cars»



Il 40% della popolazione mondiale è connesso ad internet

In un minuto nel mondo…



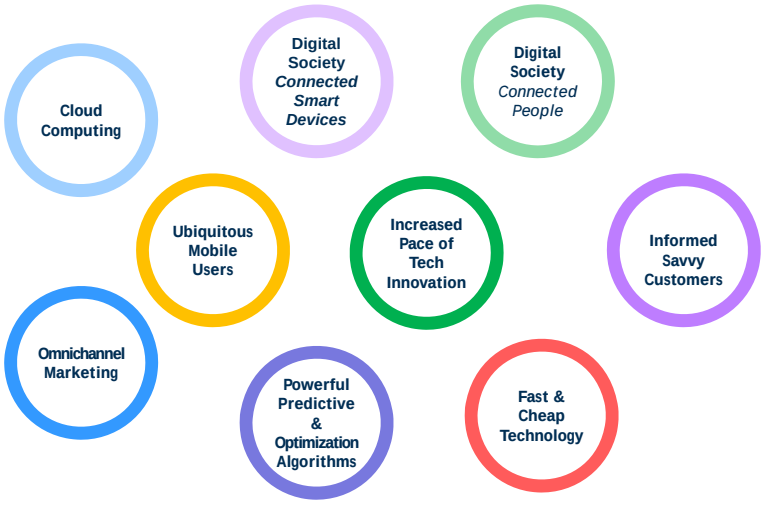
44 Zettabyte di dati online saranno generati entro il 2020

Big Data: le caratteristiche

door

Le tendenze in atto

door



La percezione delle imprese

13

door

44%

La Business Intelligence e i Big Data Analytics rappresentano la **principale priorità di investimento** per la digitalizzazione delle imprese secondo i CIO italiani

22%

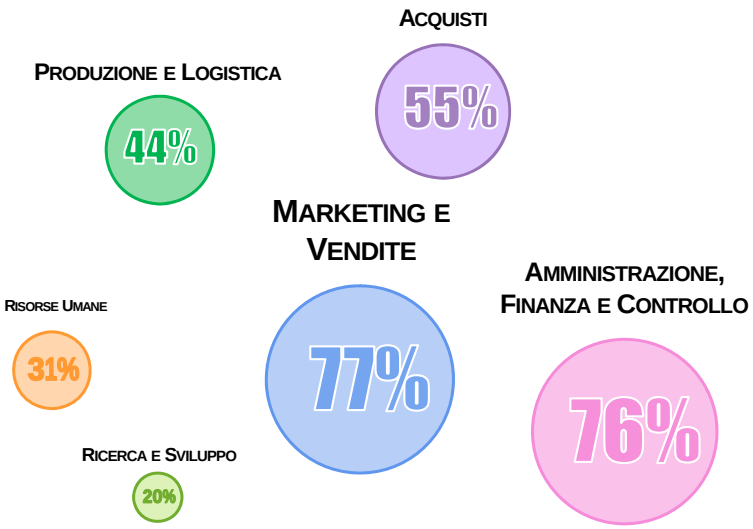
Per il 22% dei CIO italiani la gestione dei Big Data Analytics rappresenta la **principale sfida** per la trasformazione digitale dell'impresa

Campione: 231 CIO di medie e grandi organizzazioni italiane

I principali fruitori e le maggiori potenzialità

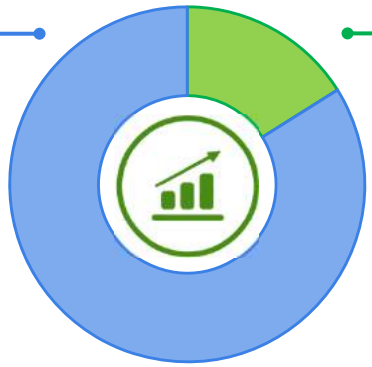
14

door



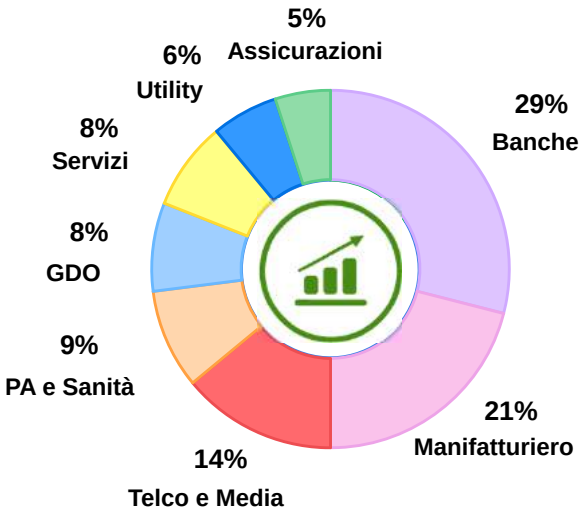
Campione: 91 CIO di medie e grandi organizzazioni italiane

Stima Mercato Analytics 2015: 790 milioni di € (+14% rispetto 2014)



BUSINESS INTELLIGENCE
84%
+11% vs 2014

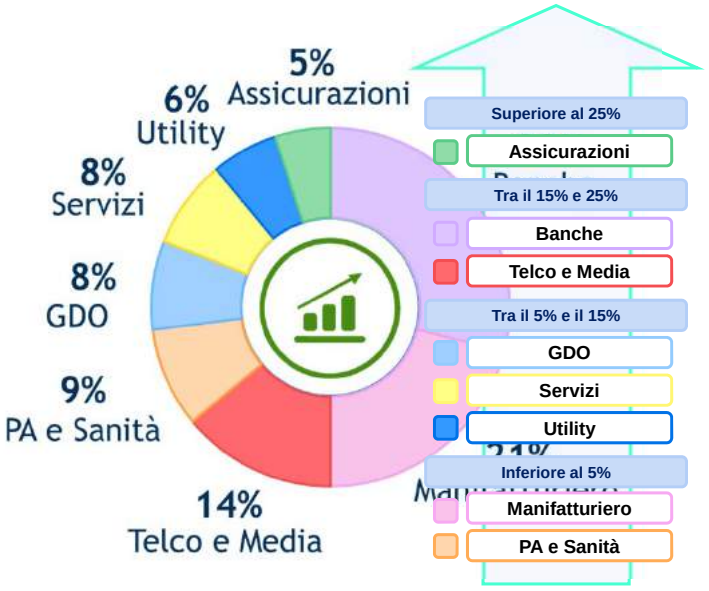
BIG DATA
16%
+34% vs 2014



Il mercato suddiviso per settore

17

door

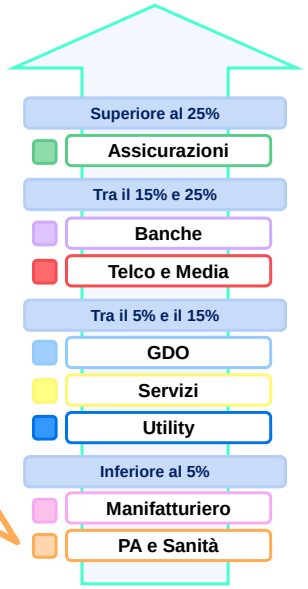


Il mercato suddiviso per settore

18

door

- Analisi frodi ed errori
- Gestione infrastrutture e asset demaniali
- Ottimizzazione ricoveri e degenza
- Supporto decisionale clinico
- Valutazione condizioni mediche

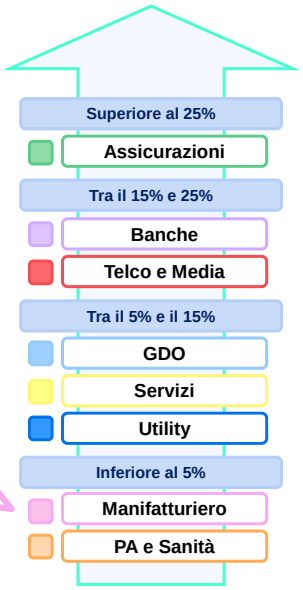


Il mercato suddiviso per settore

19

door

- Gestione asset e operations
- Gestione ciclo di vita del prodotto
- Gestione qualità
- Manutenzione predittiva
- Previsione domanda
- Ricerca & Sviluppo

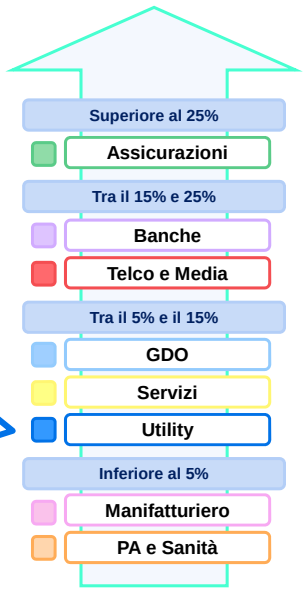


Il mercato suddiviso per settore

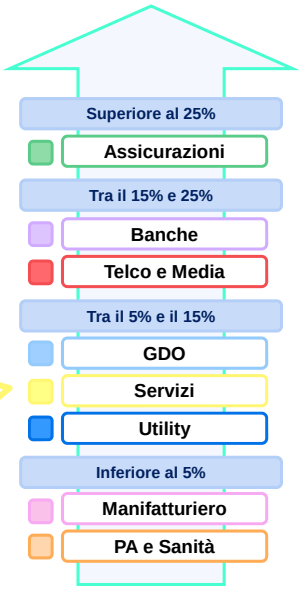
20

door

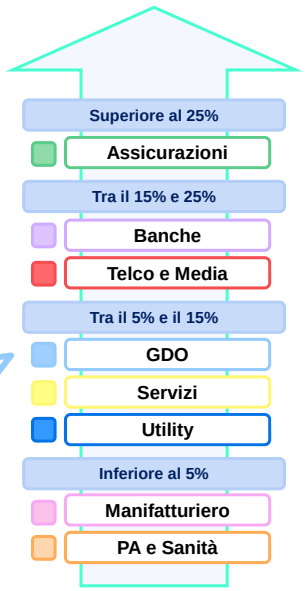
- Analisi consumi
- Analisi churn
- Manutenzione predittiva della rete
- Marketing geolocalizzato
- Pianificazione capacità rete
- Valutazione rischio



- Offerte personalizzate
- Carte fedeltà
- Ottimizzazione percorsi
- Ottimizzazione flotta



- Analisi comportamento clienti
- Cross & Up Selling
- Gestione scorte
- Ottimizzazione pricing
- Programmi loyalty

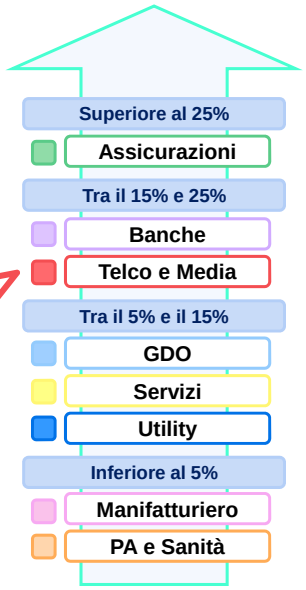


Il mercato suddiviso per settore

23

door

- Analisi profili e consumi
- Marketing geolocalizzato
- Miglioramento customer experience
- Personalizzazione offerta
- Pianificazione capacità rete e manutenzione
- Riduzione errori e malfunzionamenti
- Advertising mirato
- Identificazione argomenti virali
- Supporto servizio clienti

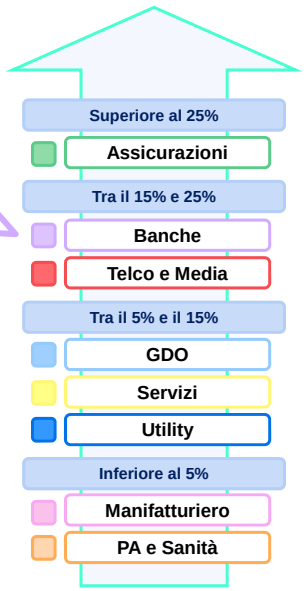


Il mercato suddiviso per settore

24

door

- Analisi andamento indici azionari
- Analisi comportamenti clienti
- Gestione rischio creditizio
- Gestione frodi

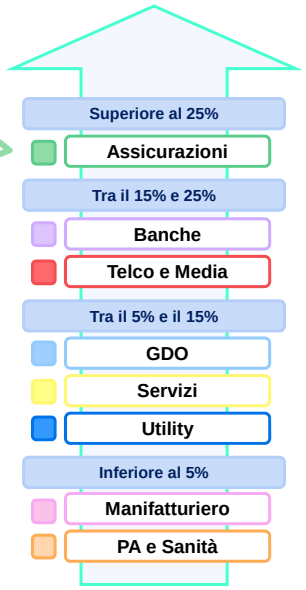


Il mercato suddiviso per settore

25

door

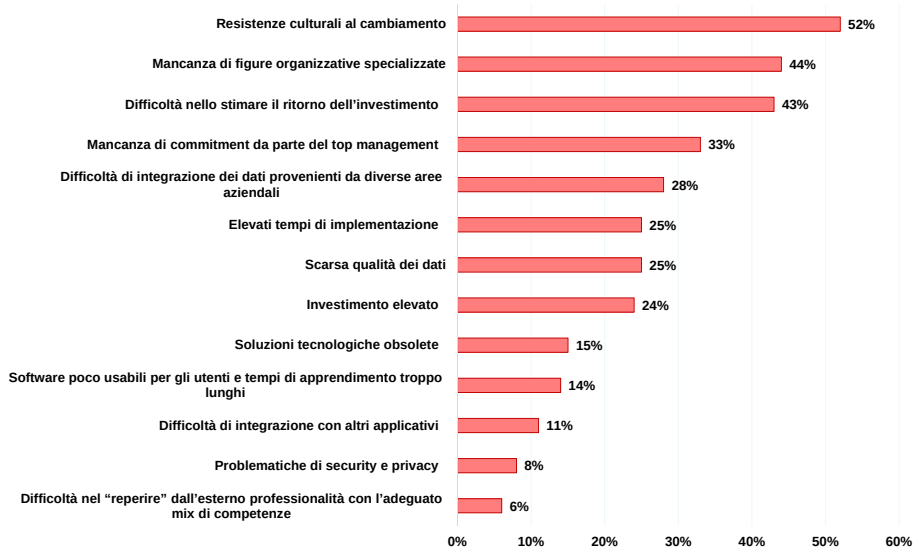
- Gestione frodi
- Profilazione servizi
- Pricing polizze



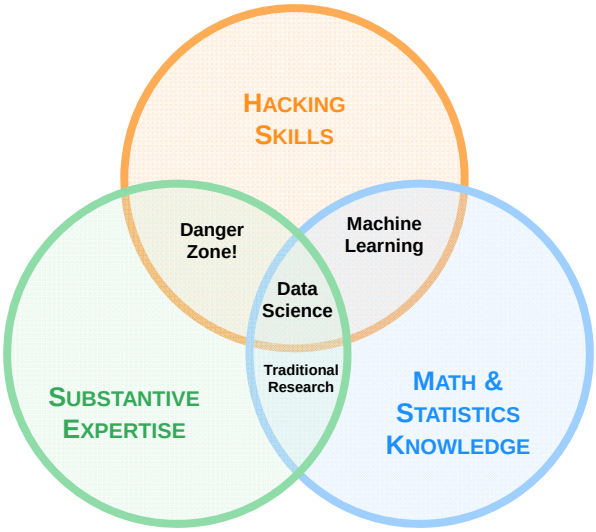
I freni all'adozione

26

door



Campione: 91 CIO di medie e grandi organizzazioni italiane


HACKING SKILLS
Danger Zone!
Machine Learning
Data Science
Traditional Research
SUBSTANTIVE EXPERTISE
MATH & STATISTICS KNOWLEDGE

LA DIMENSIONE STRATEGICA DEI BIG DATA

DAL CLOUD COMPUTING ALLA TUTELA DELLA PRIVACY

FORUM
INNOVAZIONE
PER LA COMPETITIVITA' RC 15-12-2015

DOTT. FRANCECO VITALI - GPDP

BIG DATA

Ghiaia

Nuvole tempestose

Partite a poker

Lettura del pensiero

Soldi e diritti

BIG DATA DIAMO I NUMERI ?

QUANTI SONO I DATI NEL MONDO?
800 Terabytes nel 2000
160 Exabytes nel 2006 (1EB = 10^{18}B)
4.5 Zettabytes nel 2012 (1ZB = 10^{21}B)
44 Zettabytes stimati nel 2020

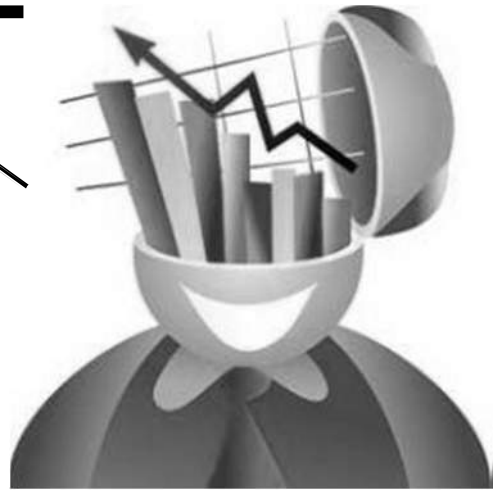
COS'È UN ZETTABYTE?
1,000,000,000,000,000,000,000 bytes
Una pila di hard disk da 1TB alta 25,400 km

Oltre il 90% dei dati nel mondo sono stati generati negli ultimi due anni!

Non è solo una questione di Zettabytes…

Morte del "campione" statistico

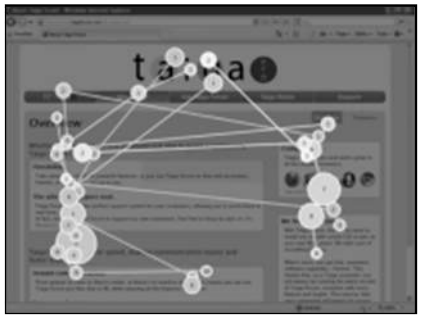
Dati destrutturati



"Lettura del pensiero"

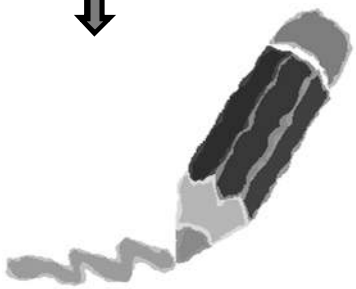
ANALISI COGNITIVA…
IN TEMPO REALE – DI LUNGO TERMINE

parametri salute

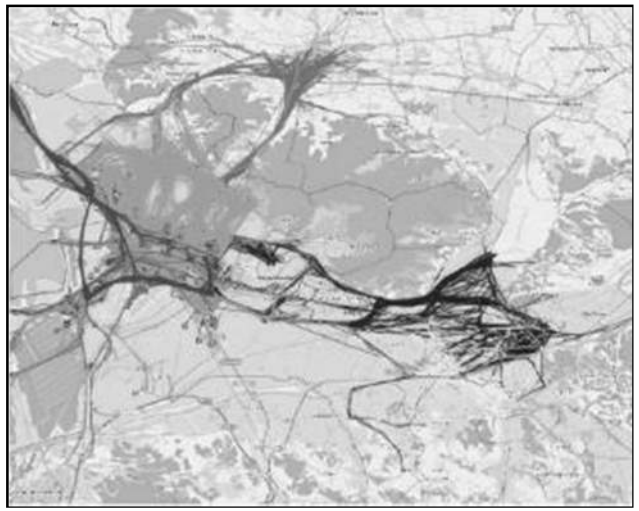


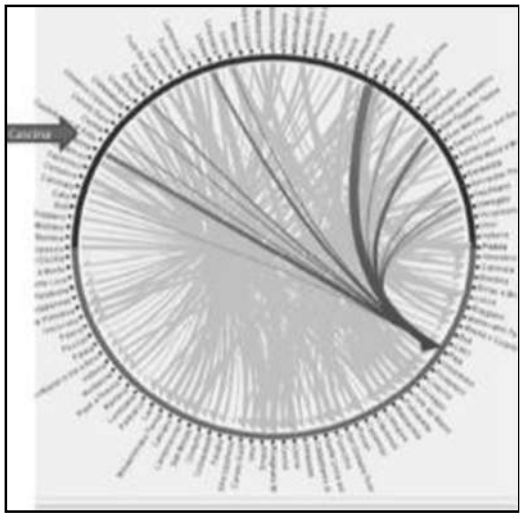
movimenti oculari

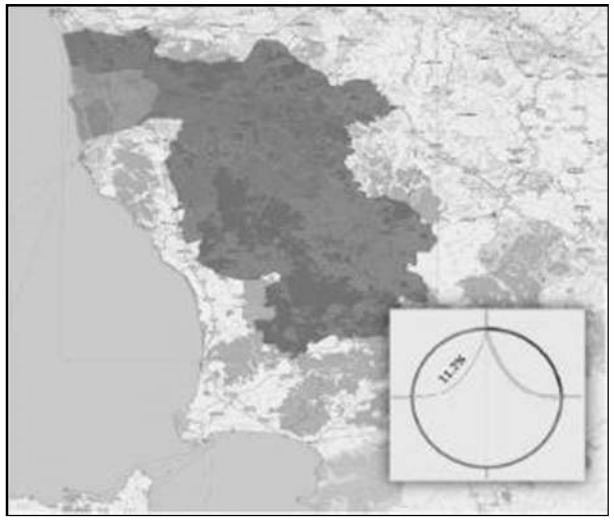
- Lettura delle espressioni facciali
- Lettura prossemica e dei movimenti
- Analisi dei processi attentivi
- Identificazione stimoli collativi
- …..



NUOVI CONCETTI SPAZIALI.. e nuovi modelli di business

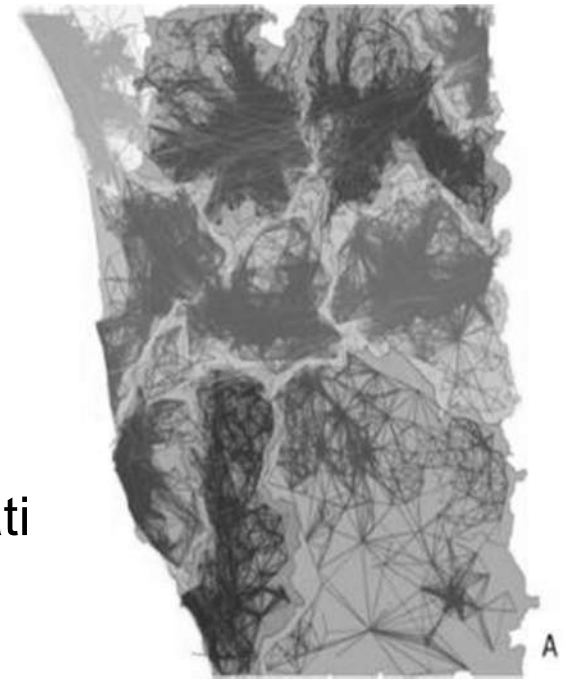






- Relazioni
- Scambi
- Movimenti
- Interessi condivisi
- Snodi di potere
- Grado di influenza…

- Servizi geolocalizzati
- Smart cities
- Campagne mirate



FINANZA

CAMPO POLITICO

MEDICINA

TRASPORTI

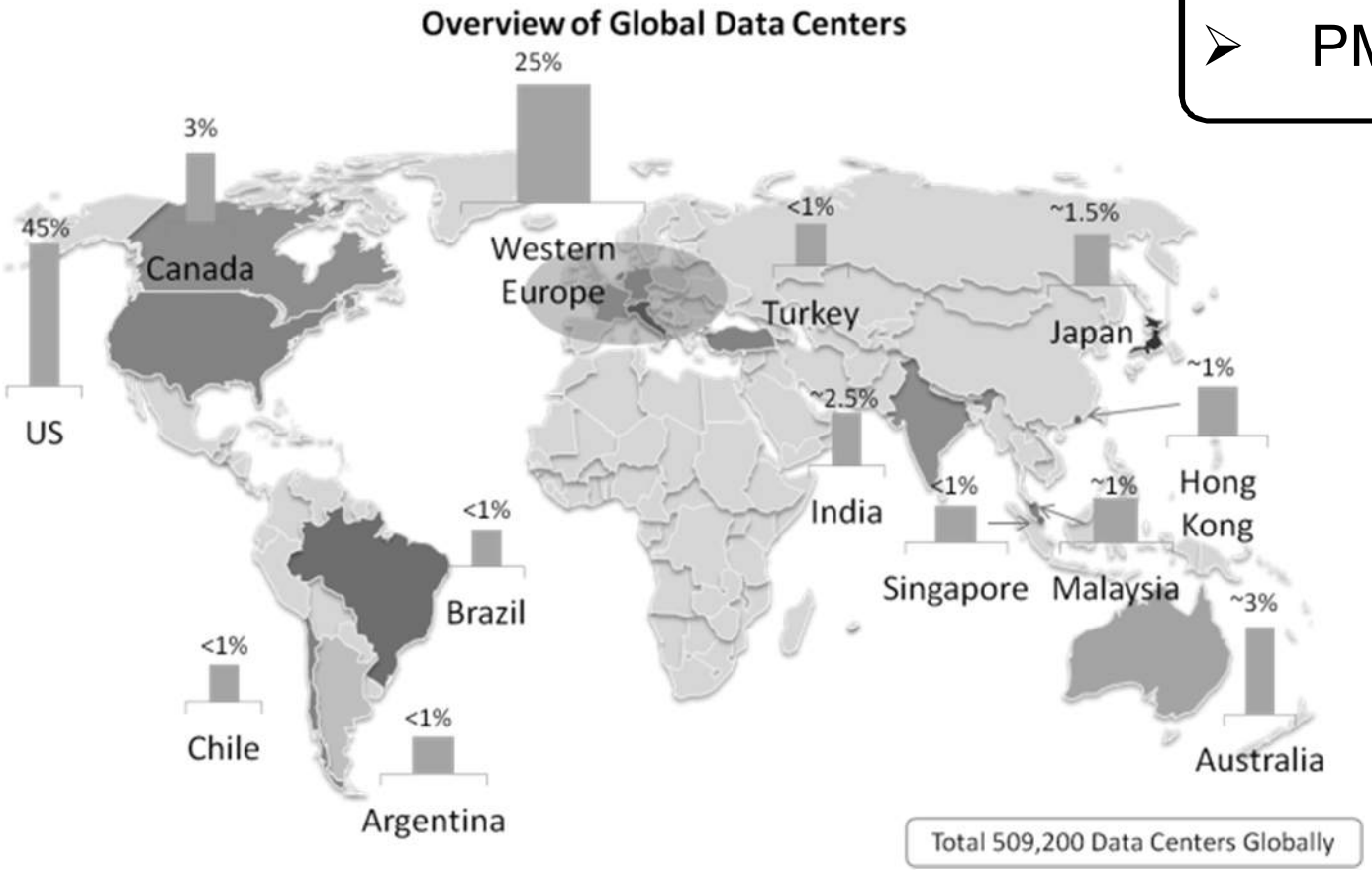
TURISMO

….Amazon ha depositato un brevetto di "Anticipatory shipping"

CHI PUO' SFRUTTARE I BIG DATA?

- SOLO GRANDI SOCIETA'?

- PMI?



IL SOGNO GRATUITO DELLE NUVOLE… per le piccole, medie e grandi imprese

IL CLOUD COMPUTING

E’ GRATIS?!

N.B. ***VEDI IL CONTRATTO UTENTE***

LA PRIVACY … PER GIOCARE CON LE REGOLE

1. **PROFILAZIONE SENZA LIMITI E DE-ANONIMIZZAZIONE DEI DATI**

2. **REAL TIME BIDDING E L'ASIMMETRIA INFORMATIVA**

3. **FILTER BUBBLE OVVERO IL BOZZOLO VIRTUALE**

4. **DATI SENSIBILI… spesso bistrattati**

LA PRIVACY E' MORTA?
W LA PRIVACY!

Catena di comando

- Chi comanda in un'impresa?
- Chi è il titolare?

Finalità

- Accessibilità/possesso VS libero utilizzo

"Proprietà" VS furto dei dati

- Potere al popolo… degli interessati

Il valore dei dati

- Tutela del proprio patrimonio aziendale

Ruolo sociale dell'impresa

- Tutela della democrazia

MONDO DELL'IMPRESA E PROTEZIONE DEI DATI
… trasformare il problema in opportunità

SENTENZA DIRETTIVA FRATTINI –
Data Retention - sicurezza

SENTENZA GOOGLE SPAIN - *territorialità*

SENTENZA SCHREMS /FACEBOOK- *Safe Harbour … problema Cloud Computing*

CONCLUSIONE 2

E ORA CHE ABBIAMO TUTTI QUESTI DATI…?

Francesco Vitali
fv @futurevision .it
f. vitali @ garanteprivacy . it

Grazie!

Ministero dello Sviluppo Economico

L'Agenda digitale e le startup tecnologiche come elementi cardine di una nuova politica industriale

Enrico Martini
Segreteria tecnica del Ministro

agenda digitale

DIAMO ALL'ITALIA UNA STRATEGIA DIGITALE

Strategia per la banda ultralarga

il piano strategico per la banda ultralarga si pone l'obiettivo di:

- **massimizzare entro il 2020 la copertura della popolazione con una connettività ad almeno 100 Mbps** (ad oggi la copertura è del 5%)
- comunque **garantire a tutti i cittadini almeno 30 Mbps in download** (ad oggi la copertura è del 20%)

Strategia per la crescita digitale

- **Servizio Pubblico d'Identità Digitale**
- **Anagrafe Nazionale della Popolazione Residente**
- **Sistema pagamenti PA**

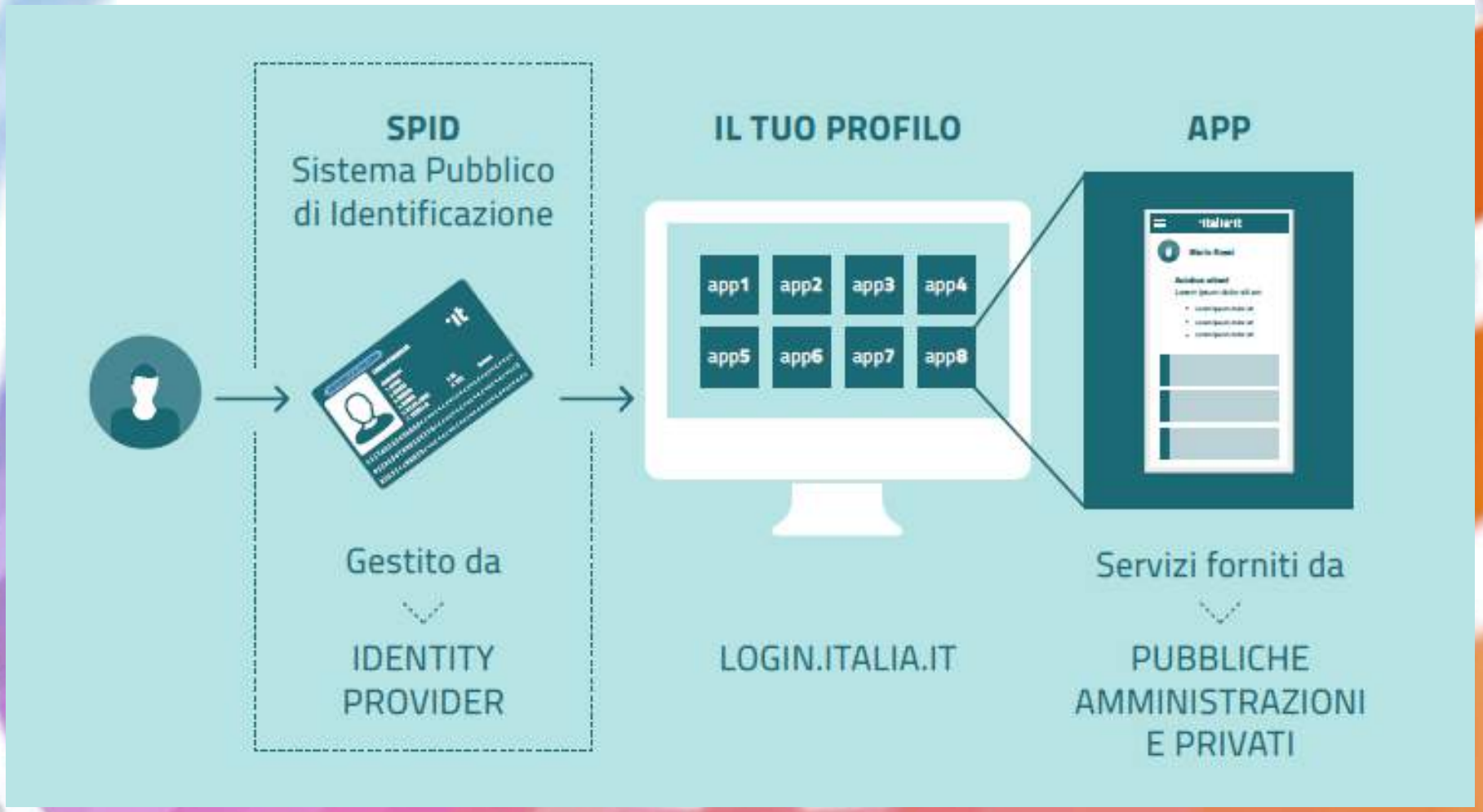
Un piano di investimenti pubblici fino a **12 miliardi €** in 7 anni

4.4 MLD
FESR/FEASR

5 MLD
FSC

Altre risorse
Fondo Juncker, «Sblocca italia»,
economie SPC

Sistema Pubblico per la gestione dell'Identità Digitale (SPID)



Servizio Pubblico d'Identità Digitale (SPID)

Perchè

- **Incrementare l'utilizzo dei servizi on line**, specialmente quelli dispositivi e il commercio elettronico, in maniera da beneficiare dell'utilizzo di internet e delle nuove tecnologie in tutti i settori economici
- **Semplificare l'accesso ai servizi digitali** senza penalizzare la sicurezza e la privacy
- **Proteggere il cittadino/consumatore**

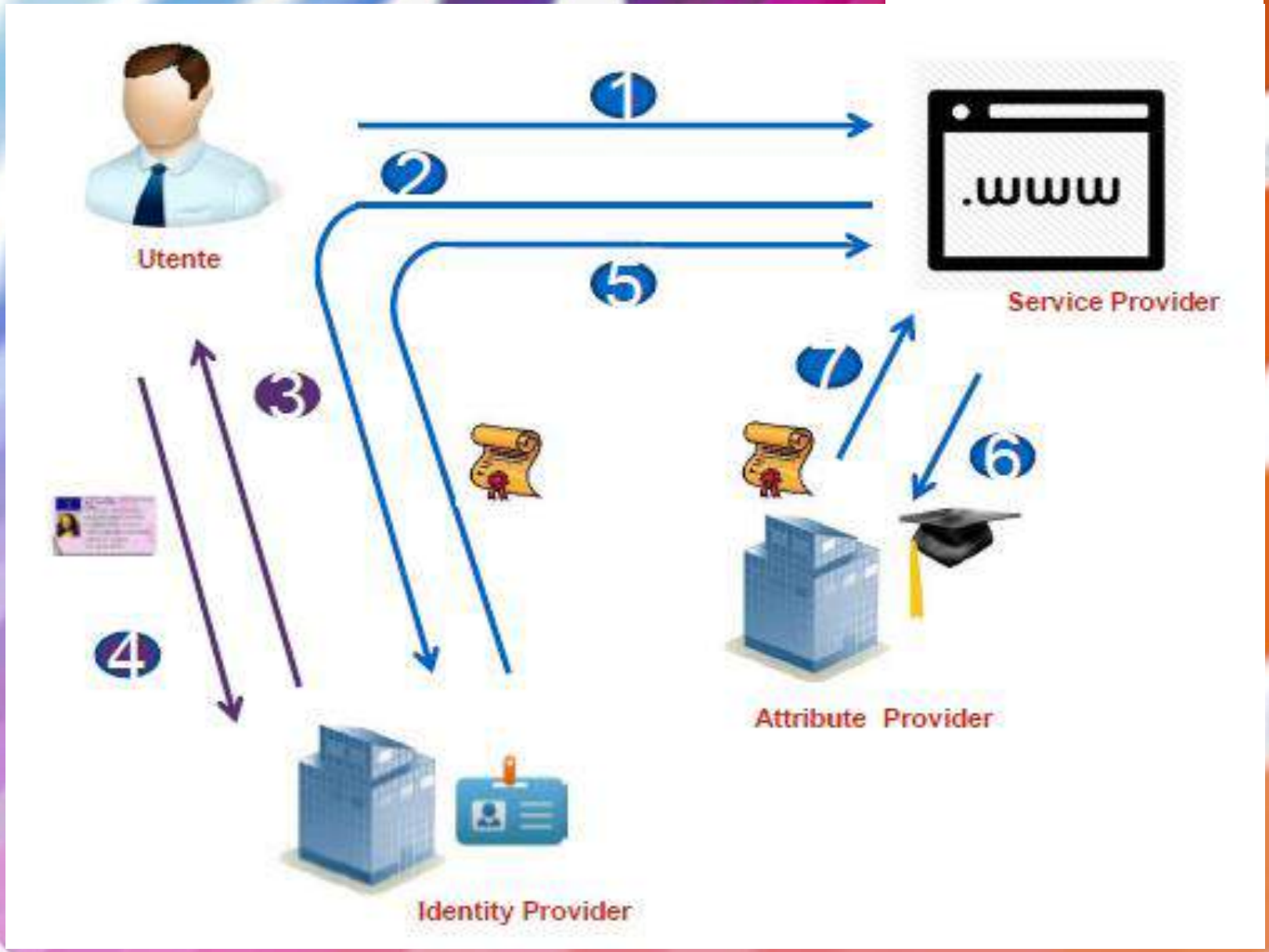
Cosa

- **Sistema che consente agli utenti di essere riconosciuti e di ricevere credenziali**, per accedere con le medesime a tutti i servizi pubblici e privati il cui livello di accesso sia compatibile con quello della credenziale presentata (livelli 1, 2, 3)
- **In SPID i fornitori di attributi qualificati**, su richiesta del fornitore di servizi, attestano in rete il possesso degli attributi o qualifiche necessari per accedere ad un determinato servizio

I soggetti coinvolti

- **Service Provider**: i soggetti pubblici e privati che utilizzano SPID per il controllo delle credenziali di accesso ai propri servizi
- **Identity provider**: I soggetti che, previo accreditamento da parte AgID e nel rispetto dei regolamenti, attribuiscono l'identità digitale ai soggetti che la richiedono, fornendo la relativa credenziale e garantendo ai service provider la verifica della credenziale medesima;
- **Attribute provider**: i soggetti titolati che, previo accreditamento AgID e nel rispetto dei regolamenti, forniscono prova del possesso di determinati attributi e qualifiche
- **AgID**: svolge il ruolo di vigilanza sui soggetti accreditati ed il ruolo di garante della federazione, gestendo il registro che rappresenta l'insieme dei soggetti che hanno sottoscritto un rapporto di fiducia

Funzionamento



Monitoraggio

SPID

sistema pubblico di identità digitale

amministrazioni pilota	10
richieste accreditamento identity provider	4

Dicembre 2015

Si chiudono le prime procedure di accreditamento IdP

Gennaio 2016

Partenza di 300 servizi disponibili tramite SPID

Dicembre 2017

Adesione di tutta la PA a SPID

I target

- **Avvio dei servizi almeno delle PA pilota (INPS, Inail, Agenzia delle entrate e le maggiori Regioni) entro il 2015**, massimizzando il numero di identità SPID rilasciate ai cittadini
- **10 milioni di utenti SPID entro il 2017**
- **30 milioni di utenti SPID entro il 2020**
- **Avvio utilizzo di SPID in significativi settori privati entro il 2017**
- **Attivazione degli attribute provider dal 2016**

Anagrafe Nazionale della Popolazione Residente (ANPR)

- **Unica banca dati con le informazioni anagrafiche della popolazione residente** a cui faranno riferimento non solo i Comuni, ma l'intera Pubblica amministrazione e tutti coloro che sono interessati ai dati anagrafici, in particolare i gestori di pubblici servizi

- Allineando i dati toponomastici, **permetterà di concretizzare l'Anagrafe nazionale dei numeri civici e delle strade urbane** (ANNCSU), strumento necessario a completare la **riforma del Catasto**

- **Assicurerà ai Comuni un sistema di controllo, gestione e interscambio, puntuale e massivo, di dati, servizi e transazioni** necessario ai sistemi locali per lo svolgimento delle funzioni istituzionali di competenza comunale

Monitoraggio

ANPR

anagrafe nazionale popolazione residente

comuni pilota	26
milioni di cittadini coinvolti	6.5

Dicembre 2015

Partenza dei primi 2 comuni pilota, Cesena (FC) e Bagnacavallo (RA)

Febbraio 2016

subentro dei rimanenti comuni del gruppo pilota

Dicembre 2016

Completamento dell'Anagrafe Unica per tutti i Comuni

Progetto «Pago la Pubblica Amministrazione»

- **Progetto strategico che consente a cittadini ed imprese di eseguire pagamenti in modalità elettronica** scegliendo liberamente:
  - **il prestatore di servizio,**
  - **gli strumenti di pagamento**
  - **il canale tecnologico preferito**

  … e alle **pubbliche amministrazioni** di:

  - **velocizzare la riscossione dei crediti** (esito in tempo reale e riconciliazione certa e automatica)
  - **ridurre i costi e uniformare i servizi agli utenti**

- **Risultano aderenti 494 Enti Creditori** che coprono tutte le possibili tipologie della PA e la sua complessità operativa
- **La popolazione coperta dalle adesioni dei Comuni è di almeno 4 milioni di cittadini**
- **Per fine 2015 si prevede di traguardare un numero di Enti aderenti pari a 600**
- **Tra quelli più rilevanti si evidenzia**: ACI, Regione Puglia e Umbria, Equitalia, Città Metropolitane di Milano e Napoli

PagoPA
sistema dei pagamenti elettronici

enti creditori aderenti	494
enti in esercizio	190
prestatori servizi di pagamento	40
transazioni totali	109050

Altri progetti strategici

Fatturazione elettronica

milioni di fatture gestite	17
uffici fatturazione elettronica su IPA	54775
amministrazioni su IPA	22875

Open data

i dati aperti della pubblica amministrazione

dataset	10348
amministrazioni	76

Competenze digitali

coalizione per le competenze digitali

membri della coalizione	126
progetti	79

Fse

fascicolo sanitario elettronico

regioni operative	4

Open e Big Data

- Attraverso l'uso delle tecnologie e con metodi innovativi, **il Governo persegue le politiche di open data**, anche nell'ambito della Open Government Partnership, promuovendo la cultura della trasparenza nella pubblica amministrazione

- **Trasparenza, accountability e partecipazione** sono infatti obiettivi fondamentali dell'azione del Governo italiano

- **Qualunque dato trattato da una pubblica amministrazione deve essere reso accessibile e fruibile**, fermo restando il rispetto della normativa in materia di protezione dei dati personali

- **10.348 dataset prodotti da 76 amministrazioni tra cui Dati Geografici e 695 Dati Statistici (www.dati.gov.it)**:

  - 629 dataset comunali, 353 nazionali, 203 provinciali

Rimettere in moto
l'ingranaggio della crescita

Le imprese innovative

Perché le startup innovative sono importanti

Negli ultimi dieci anni, nei settori non-finanziari le imprese giovani (fino 5 anni di vita), pur impiegando soltanto il 20% dell'occupazione complessiva, hanno generato quasi la metà del totale di nuovi posti di lavoro

(indagine Ocse su 15 Paesi membri)

"the Internet economy in the developed markets of the G-20 will to grow at an annual rate of 8% over the next five years"

(the Boston Consulting Group)

"the number of applications developers in Europe is set to rise from 1 million in 2013 to 2.8 million in 2018. Support and marketing staff, meanwhile, accounted for a total of 1.8 million jobs in 2013, and this number is set to grow to 4.8 million by 2018"

(Gigaom Research)

Benefici per le startup innovative (dl 179/2012)

- **Costituzione societaria e successive modificazioni anche con un modello standard tipizzato senza passare dal notaio (dl 3/2015);** le startup innovative sono registrate presso una sezione speciale del Registro delle imprese delle CCIAA tramite **autocertificazione**
- **Gestione aziendale estremamente flessibile** su capitale e dei diritti di voto dei soci
- **Disapplicazione fiscalità su società di comodo** e in perdita sistematica
- **Start-up "soggetti non fallibili"**, introdotti meccanismi di Fail-Fast e riduzione stigma da fallimento

❖ **Robusti sgravi fiscali a chi investe nel capitale della startup**

❖ **Possibilità di raccogliere fondi attraverso portali web di equity-crowdfunding**

❖ **Garanzia pubblica gratuita, veloce e semplificata** sui finanziamenti bancari tramite intervento del FCG

✓ **Diritto del lavoro flessibile**: liberalizzazione del contratto a termine applicabile per l'intero ciclo di vita della startup

✓ **Possibilità di prevedere una retribuzione variabile** a seconda della performance dell'impresa

✓ **Possibilità di remunerare lavoratori e consulenti con stock option e work for equity** (tassate come capital gain!)

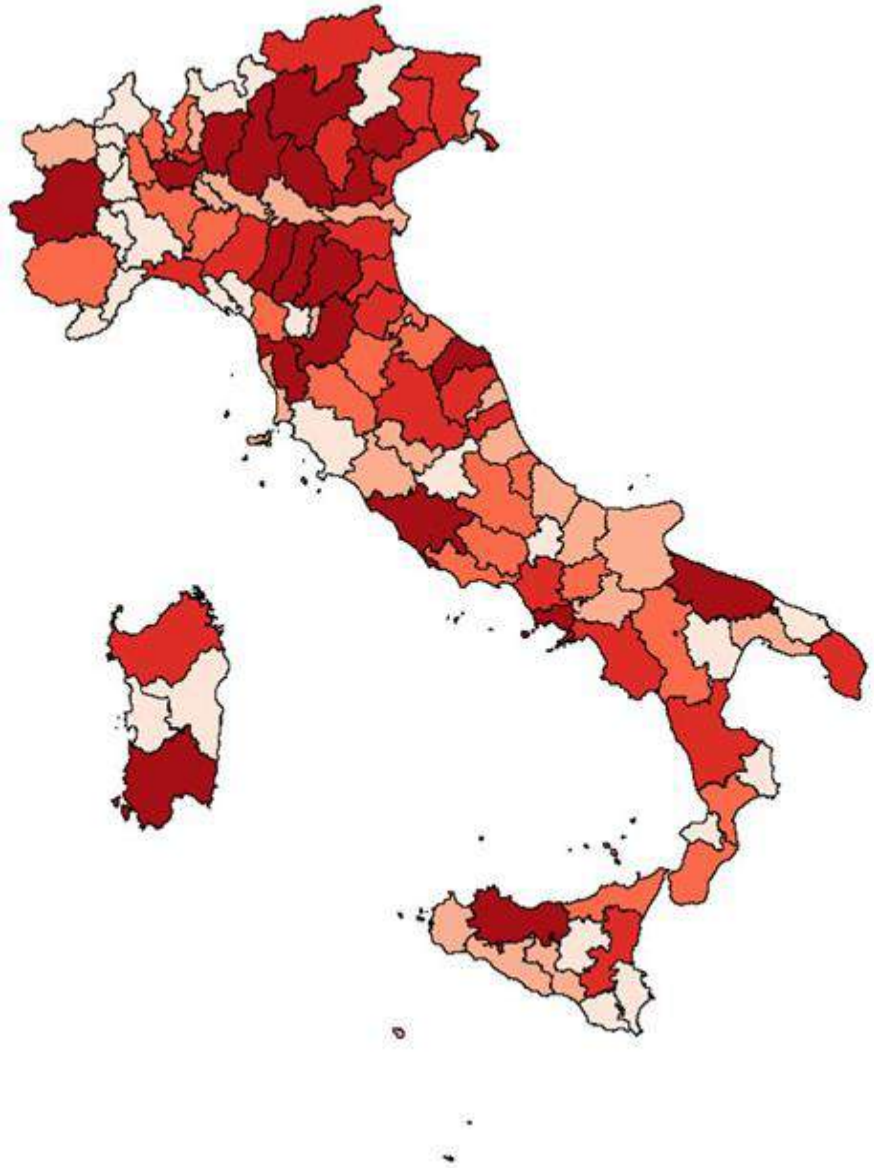
Le protagoniste dell'ecosistema italiano

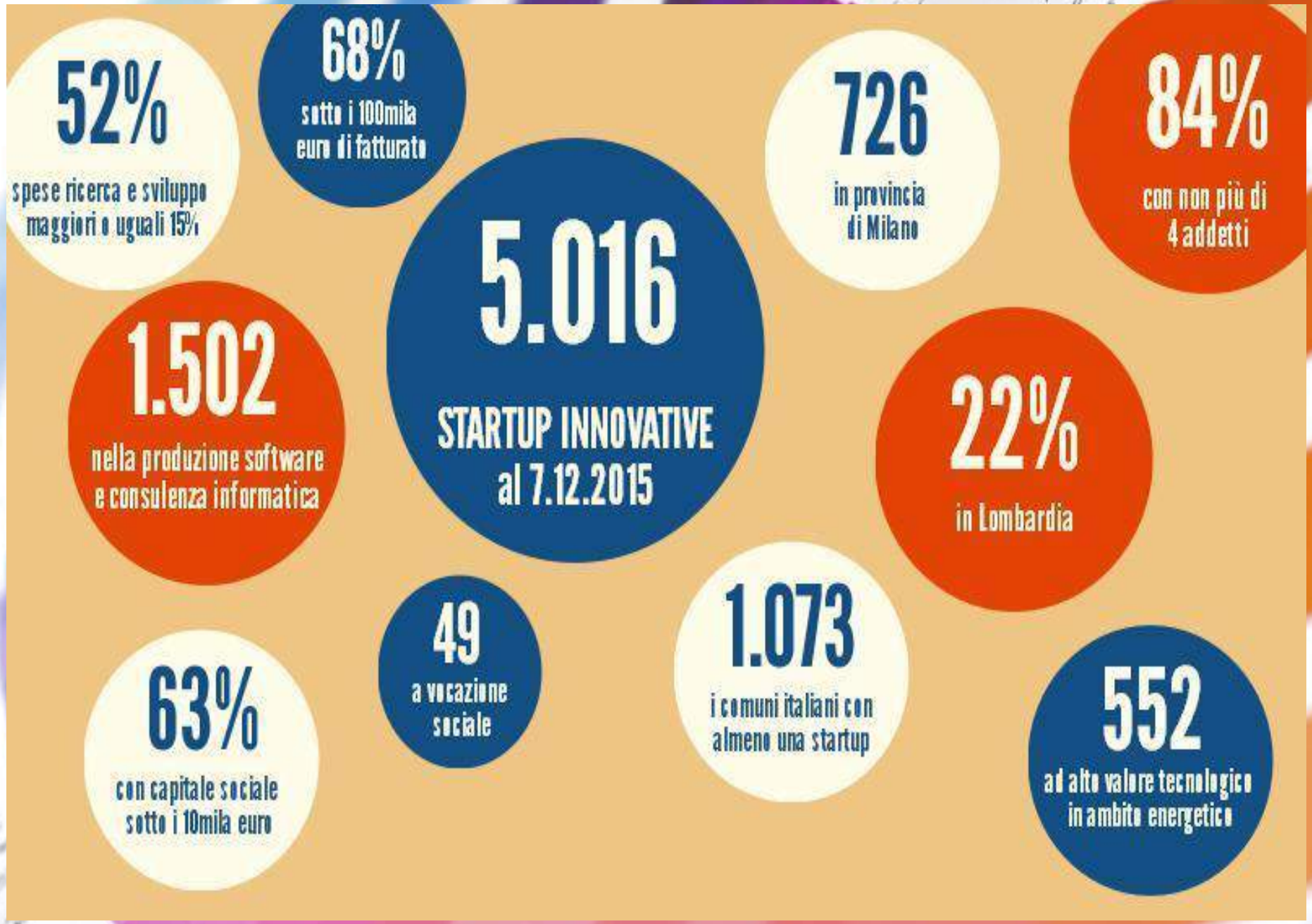
Sono oltre **5.000** le **imprese iscritte.** Il 56% delle quali si localizza al Nord, il 23% nel Mezzogiorno, il 21% al Centro

A **livello regionale** in testa c'è la Lombardia con oltre mille imprese, seguono l'Emilia-Romagna (565) e il Lazio (484)

Quasi l'80% delle startup opera nei **servizi (soprattutto produzione di software e attività di R&S)**, il 18% nell'**industria (prodotti elettronici e macchinari)**, il 4% nel **commercio**

Il fenomeno assume dimensioni interessanti sotto il profilo occupazionale: le startup impiegano oltre 20mila persone tra soci e dipendenti




52%
spese ricerca e sviluppo maggiori o uguali 15%
68%
sotto i 100mila euro di fatturato
726
in provincia di Milano
84%
con non più di 4 addetti
1.502
nella produzione software e consulenza informatica
5.016
STARTUP INNOVATIVE al 7.12.2015
22%
in Lombardia
49
a vocazione sociale
1.073
i comuni italiani con almeno una startup
63%
con capitale sociale sotto i 10mila euro
552
ad alto valore tecnologico in ambito energetico

Altri attori dell'ecosistema

36 incubatori certificati nel Registro delle imprese

La grande novità del dl 3/2015: la PMI innovativa

I benefici previsti:

- Le PMI innovative sono registrate presso le CCIAA tramite **autocertificazione**
- **Gestione aziendale estremamente flessibile** su capitale e dei diritti di voto dei soci **(la struttura finanziaria della s.r.l. si avvicina a quella della s.p.a.)**
- **Disapplicazione fiscalità su società di comodo** e in perdita sistematica
- Possibilità di remunerare lavoratori e consulenti con **stock option e work for equity**
- Possibilità di raccogliere fondi attraverso **portali web di crowdfunding**
- **Garanzia pubblica** gratuita e semplificata sui finanziamenti bancari tramite l'intervento del Fondo Centrale (D.M. MiSE-MEF)
- Robusti **sgravi fiscali a chi investe nel capitale** (D.M. MEF-MiSE previa notifica alla CE)

Le PMI innovative iscritte al Registro sono 85

#ItalyFrontiers

- **Una vetrina ufficiale su startup.registroimprese.it per le startup e le PMI innovative** che vogliono farsi conoscere da imprese e investitori italiani e internazionali
- La piattaforma coniuga i dati disponibili nelle sezioni speciali del Registro Imprese con **un ricco set di informazioni inserite volontariamente dalle imprese con firma digitale**
- Attraverso **un motore di ricerca**, le imprese potranno essere filtrate dall'utente per **settore di attività, area geografica, classe dimensionale**
- **Frutto della collaborazione** tra Ministero dello Sviluppo Economico, Giovani Imprenditori di Confindustria e Unioncamere, la piattaforma è stata **realizzata da InfoCamere**

"InfoCamere"

L'Agenda digitale e le startup tecnologiche come elementi cardine di una nuova politica industriale

GRAZIE PER L'ATTENZIONE

Big data e prospettive per la statistica ufficiale: la qualità dell'informazione

Stefano De Francisci
Istituto nazionale di statistica - ISTAT
Direzione centrale per le tecnologie dell'informazione e della comunicazione

Istat

Sommario

1. Big Data e statistica ufficiale
2. La qualità dei Big Data: l'altra faccia della medaglia
3. Possibili fonti, possibili usi, possibili effetti
4. Strategie e azioni Istat per l'utilizzo di Big Data
5. Quadro riassuntivo

Statistica ufficiale e Big Data

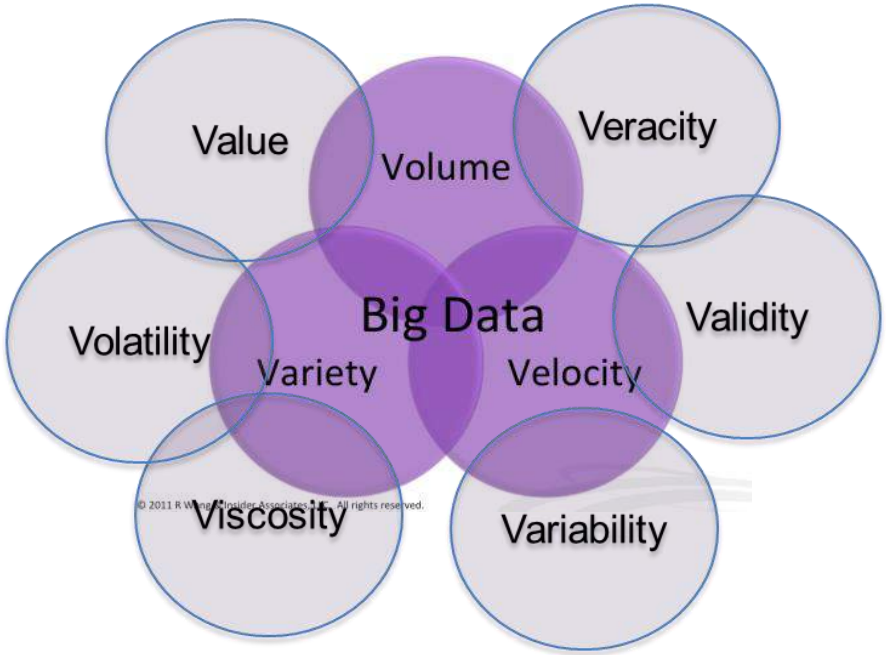
I caratteri…

…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

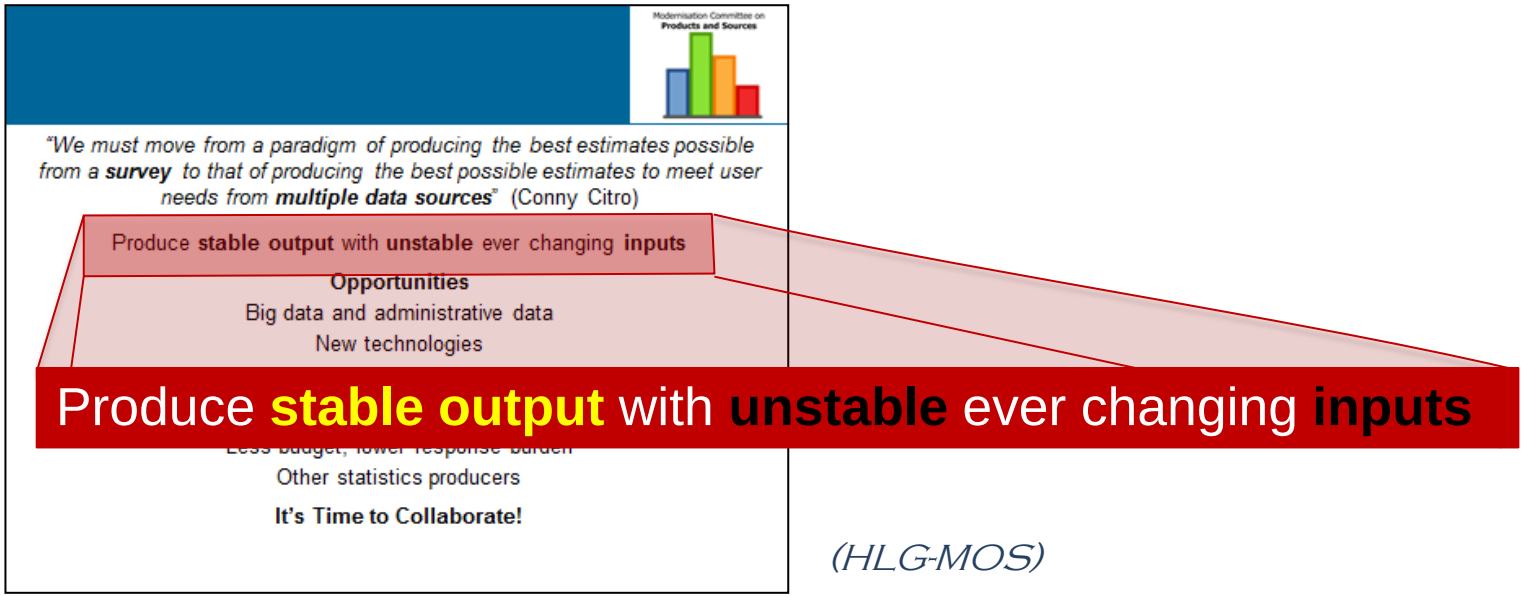


« … Big Data is also potentially very interesting as an input for Official Statistics; either for use on its own, or in combination with more traditional data sources such as sample surveys and administrative registers» *(**H**igh-**L**evel **G**roup for the **M**odernisation of **O**fficial **S**tatistics , gennaio 2013)*

Cosa comporta fare statistica ufficiale con Big Data?

Qualità come stabilità di dati e trattamento

1



(HLG-MOS)

Qualità come valore dell'informazione

2

"With the advent of big data, data quality management has become more important than ever. Typically, volume, velocity and variety are used to characterize the key properties of big data. **But to extract value and make big data operational, the importance of the fourth 'V' of big data, veracity, is increasingly being recognized. Veracity directly refers to inconsistency and data quality problem.**"

(B. SAHA, D. SRIVASTAVA)

I rischi della mancanza di qualità nei Big Data

Effetto palla di neve

…ovvero quando anche un errore minore può crescere via via lungo il trattamento e diventare un errore non più correggibile

http://mitiq.mit.edu/IQIS/Documents/CDOIQS_201177/Papers/01_04_T2A_+Sarsfield.pdf

Effetto farfalla

…ovvero quando un piccolo errore iniziale può far scaturire enormi problemi

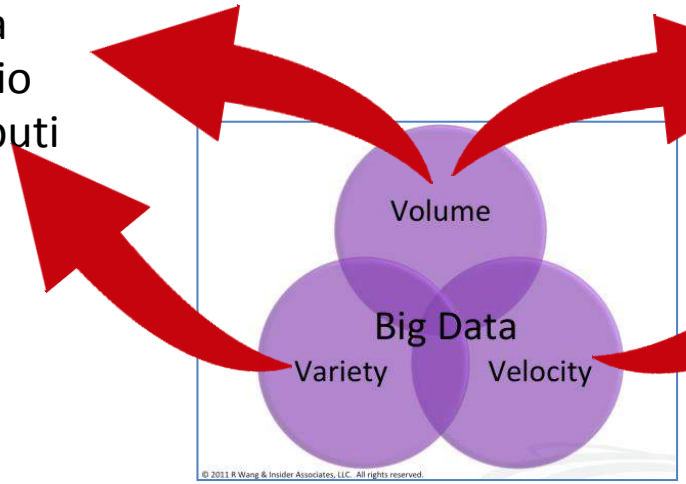
http://www.theserverside.com/feature/Handling-the-four-Vs-of-big-data-volume-velocity-variety-and-veracity

Qualità dei Big Data: vincoli e opportunità

Difficoltà di determinare la semantica dei dati e lo studio delle correlazioni tra gli attributi

Specifiche semantiche globali → **impossibili**

Specifiche semantiche locali → **non sufficienti per intero set di dati**



Difficoltà di capire o eliminare i dati errati in modo adeguatamente veloce

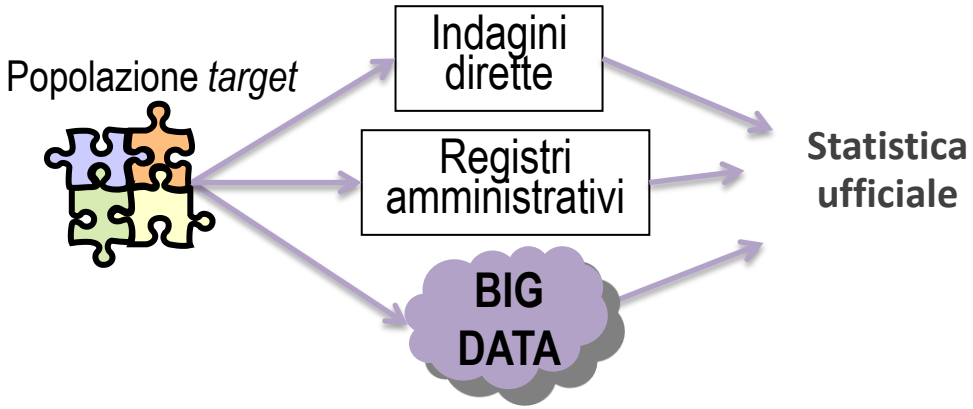
Necessità di imparare le regole di controllo e correzione dagli stessi dati *sporchi*

Dal mondo chiuso dei data base alla visione di un mondo informativo aperto

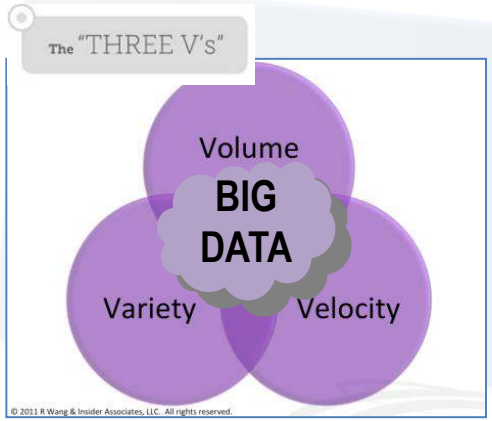
- **Regole di qualità dei dati *sensibili al contesto***
- **Regole apprese dai dati via via che vengono raccolti**
- **Dati validati e aggiornati in modo incrementale e sulla base dei dati più recenti**

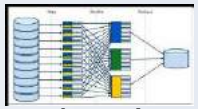
Oltre l'*infanzia* dei Big Data

Big data come **fonte addizionale** per le indagini statistiche

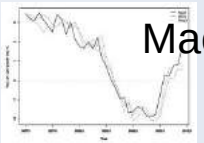


Da sperimentazione su caratteristiche base… …a consolidamento con specifiche tecniche … …a uso *organico* nel ciclo di vita dei processi statistici

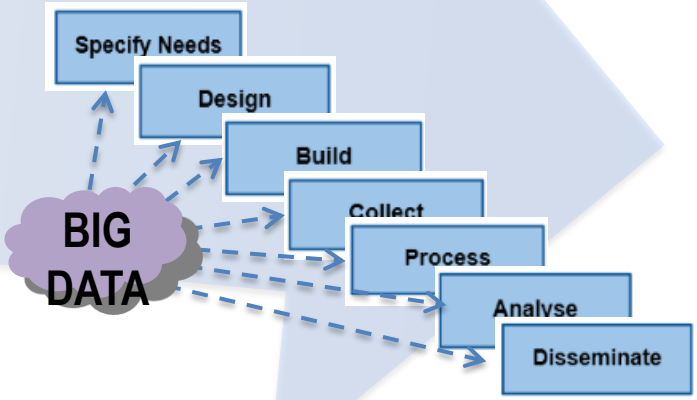



Map/Reduce

Machine learning


Nowcasting


Data Mining



Classificare le 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

Traditional Business systems

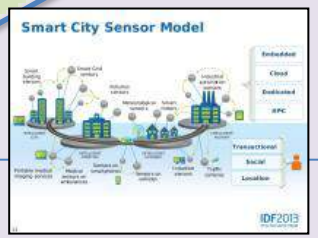
Dati prodotti da sistemi transazionali tradizionali e in modo passivo:
Scanner data
Log ricerca,
Record medici,
Transazioni commerciali e bancarie

HUMAN-SOURCED INFORMATION

PROCESS-MEDIATED DATA

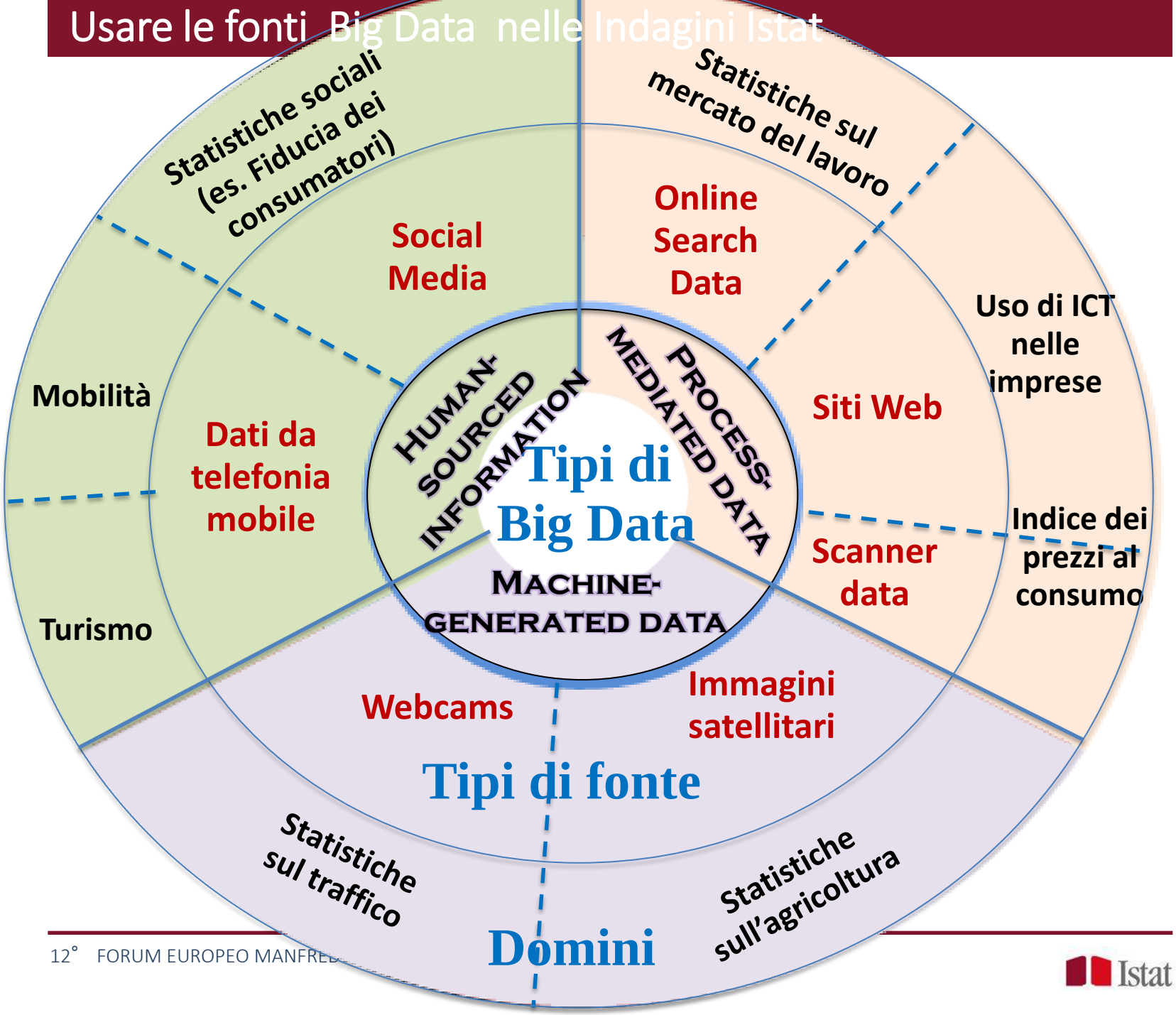
MACHINE-GENERATED DATA

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

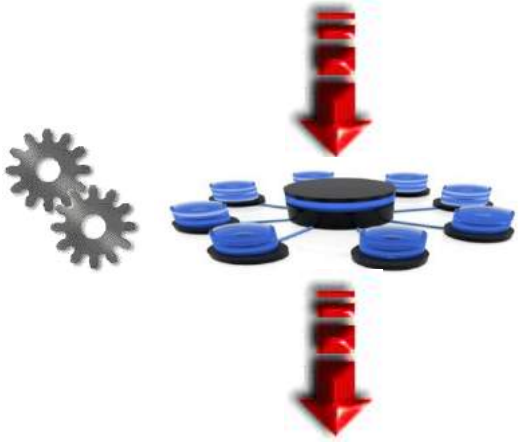
Usare le fonti Big Data nelle Indagini Istat



Istat

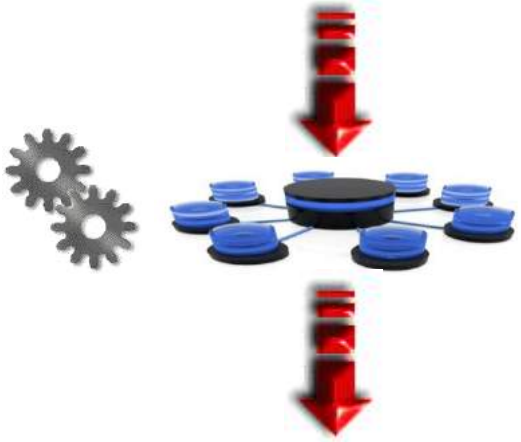
Indice prezzi al consumo

- Scopo: Innovare il disegno dell'indagine utilizzando anche fonti non tradizionali
- Fonti utilizzate: Uso alternativo di tre differenti canali
  - ✓ Collezione diretta CAPI tramite PC/tablet
  - ✓ Scanner data
  - ✓ Web scraping su siti Web per alcuni prodotti (Hi-Tech, Mobile, IT, ecc.)

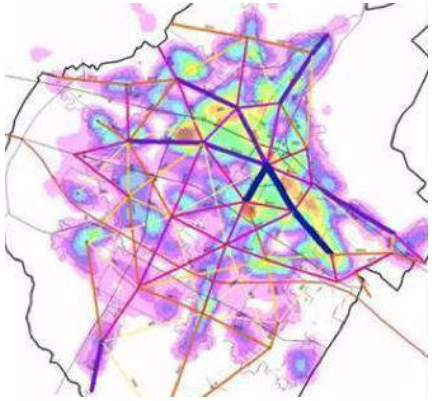


Indice prezzi al consumo

- Gruppi della grande distribuzione: Coop, Conad, Selex, Esselunga, Auchan, Carrefour
- Prodotti: alimentari e grocery
- Mercati
  - Primo invio ottobre 2014: Torino, Ancona, Palermo, Piacenza, Cagliari.
  - In seguito: Ravenna, Roma, Bari, Bergamo, Perugia, Napoli, Catania, ecc.
- Record: Punti Vendita della Grande Distribuzione
- Variabili: Identificativo, Ragione sociale, Indirizzo, Partita IVA, Ean-code (European Article Number), Quantità venduta, Fatturato (IVA inclusa) , ecc.



Persons & Places



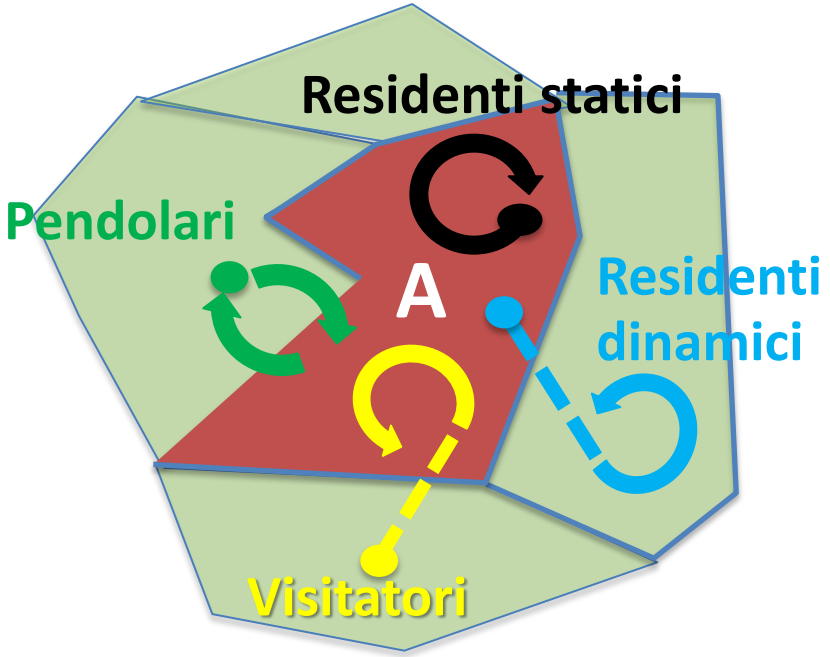
- Scopo: produzione di stime sulla matrice O/D della mobilità giornaliera per motivi di studio e lavoro a livello di comune, partendo da dati di telefonia mobile.

  Lo scopo è ottenere una affidabilità comparabile a quella ottenuta con dati provenienti dal censimento e da registri amministrativi

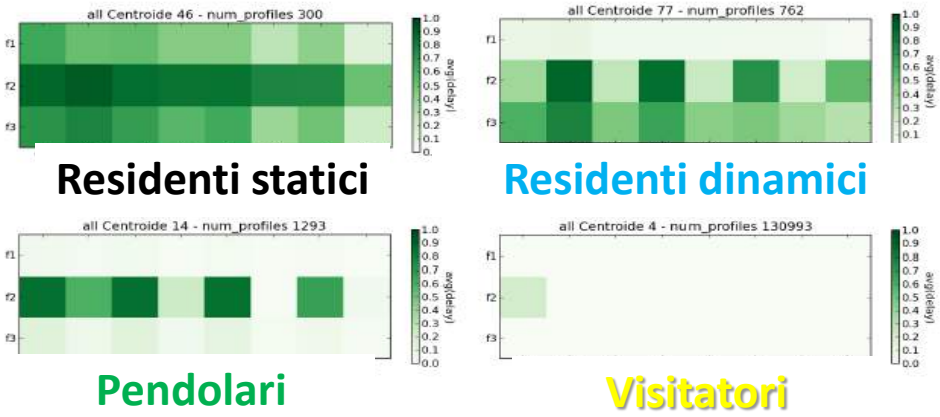
  Raggiungere questo risultato significa essere in grado di integrare in modo sicuro le statistiche esistenti della popolazione e dei flussi con le stime aggiornate continuamente ottenibili dai dati GSM

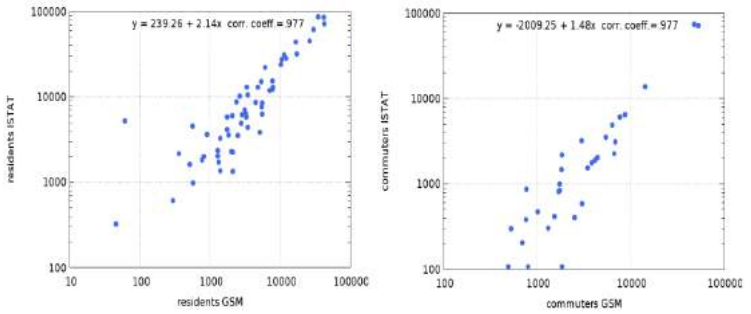
- Fonti: uso integrato di Big Data, fonti censuarie e fonti amministrative
- Attori: CNR, Università di Pisa, Istat
- Metodologia
  - Inferenza sui profili di mobilità della popolazione attraverso GSM Call Deail Records (CDRs), acquisendo informazioni sul riferimento temporale di inizio/fine della chiamata e sulla sua localizzazione territoriale
  - Applicazione di metodi di classificazione automatica (cluster analysis non supervisionata)

Persons & Places: popolazione che *insiste* su un territorio



- Utilizzando solo dati amministrativi non è possibile distinguere tra residenti e pendolari dinamici
- Ciò è possibile utilizzando modelli ottenuti dai dati GSM

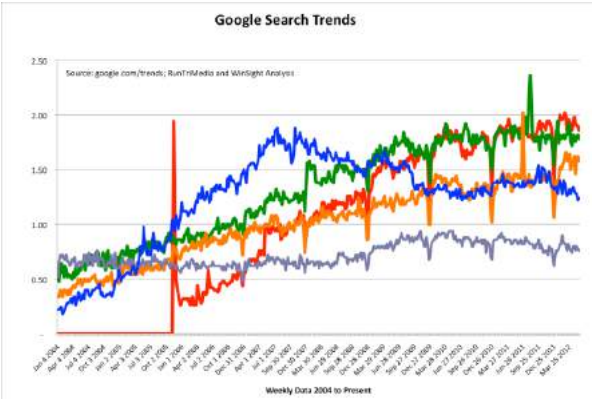




Correlazione tra stime ottenute da indagini Istat di fonte amministrativa o dei censimenti e dati da GSM

Stime sul mercato del lavoro

- Scopo: Verificare l'utilizzo di Google Trends nell'indagine sulle forze di lavoro per la produzione di stime integrate per la previsione mensile e il *nowcasting* per piccole aree (miglioramento delle stime a livello territoriali accedendo serie GT a granularità più fine, ad esempio, province).

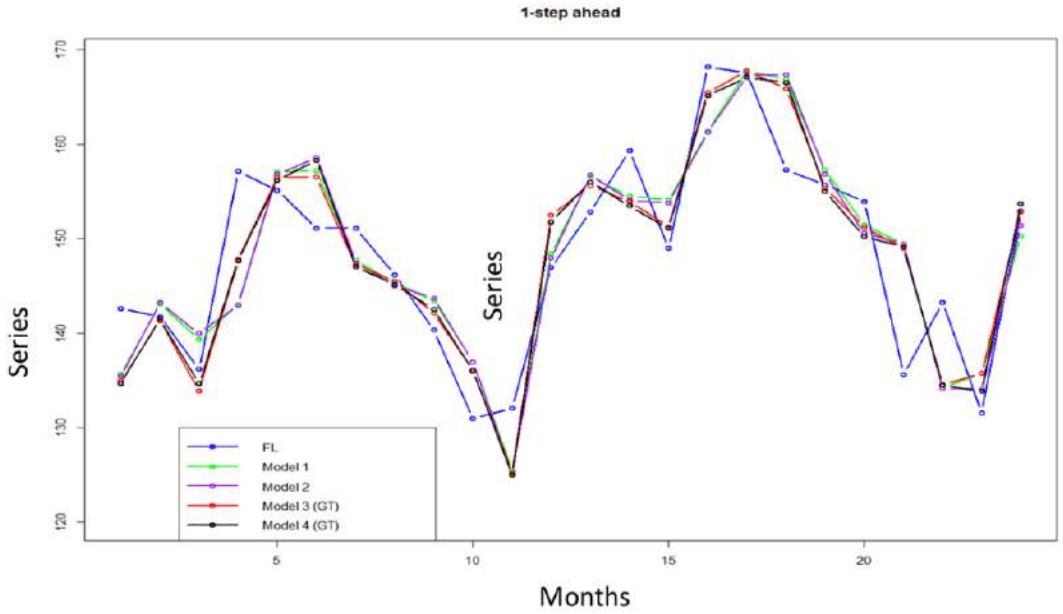


- Tipo di processo: uso di query sulle serie storiche da Google Trends, come variabili ausiliari per migliorare le stime prodotte dall'Istat attraverso l'utilizzo di modelli basati su metodi di stima.

- Metodologia
  - Modelli autoregressivi vs utilizzo di Google Trends, come modelli di previsione
  - Confronto esteso ai modelli di previsione macroeconomica

Stime sul mercato del lavoro : principali risultati

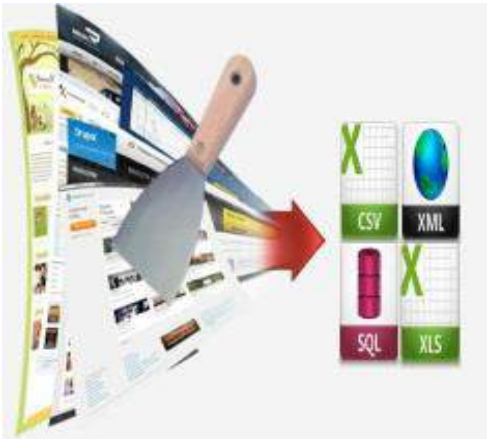
- Modellazione delle serie storiche delle forze di lavoro e selezione dei due modelli ARIMA (Modello autoregressivo integrato a media mobile) con la migliore performance (modelli 1 e 2), adottato come benchmark



- Aggiunta dei risultati di Google Trends sul termine di ricerca «offerte di lavoro» nei modelli 1 e 2 per ottenere i modelli 3 e 4
- modelli diagnostici in esperimenti preliminari hanno indicato una potenziale utilità di Google Trends per aumentare l'accuratezza delle previsioni

Uso di ICT nelle imprese

- Scopo: Valutare la possibilità di adottare tecniche di **Web scraping** e **text mining** per stimare l'uso di ICT da parte delle imprese e delle pubbliche amministrazioni tramite il reperimento di alcune variabili del questionario direttamente dal Web in sostituzione delle risposte al questionario

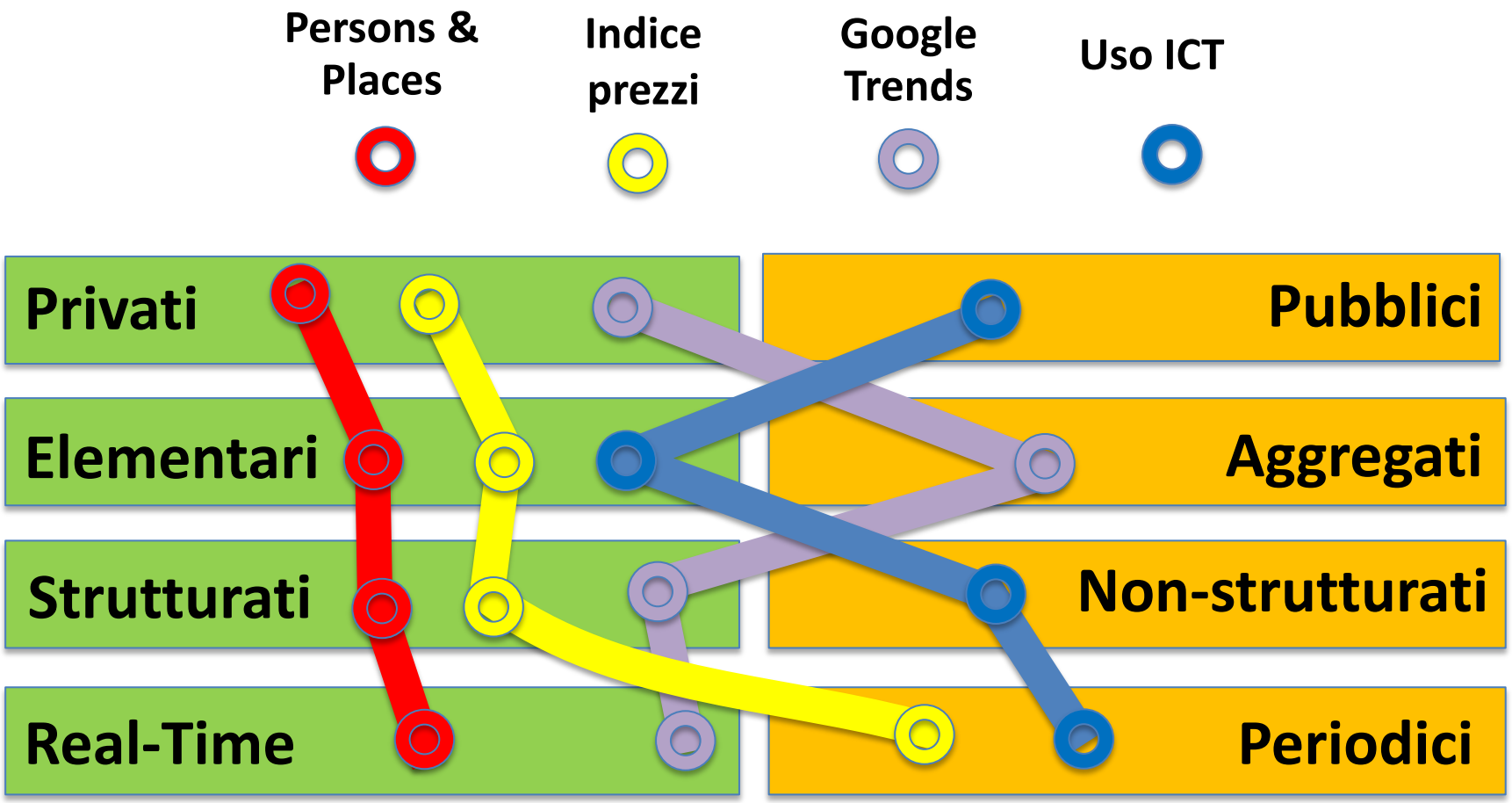


- Attori coinvolti nel progetto: Istat, Cineca
- Stato: Analizzati 8.600 website (campione rispondenti indagine ICT che hanno dichiarato di avere siti web).
- Tecnologia: Hadoop/Nutch (90min). In futuro previsto il passaggio all'analisi di 200.000 imprese (ca. 100.000 con sito Web)
- Metodologia:
  - Scraping dei siti Web per estrarre dati riferibili ad alcune domande del questionario (ad es. E-commerce)
  - Tecniche di classificazione supervisionata

Uso di ICT nelle imprese: effetto dell'uso di Big Data

	Azioni		Effetti
1	Sostituire le tecniche tradizionali di *data collection* baste su questionari con tecniche basate su *Internet as Data Source*, per tutte i quesiti idonei	→	Riduzione dell'onere sui rispondenti
2	Integrare le informazioni raccolte via questionario con quelle generate via IaD	→	Aumentare l'accuratezza delle stime
3	Raccolta di informazioni tradizionali	→	Aumentare l'offerta statistica

Caratteristiche dei dati



Caratteristiche dei Big Data utilizzati nelle sperimentazioni

Quadro di sintesi

		Persons &Places	Google Trends	Uso ICT
Fonti		Machine-generated data	Human-sourced information	Traditional Business Systems
Problematiche	IT	Applicazioni Smart sensing Identificazione di Pattern su tracking data	Acquisizione e trattamento di Search records	Web Scraping Meta-searching
	Statistiche	Record linkage e Statistical matching Popolazione target non-omogenea Controllo di qualità sui risultati	Migliorare le perfromance delle previsioni	Tecniche di Text mining
	Organizzzative	Privacy	Accesso ai risultati delle Web search	Accesso a Web sites
Impatto su processi di produzione		Notevole impatto sul processo di produzione: le fonti sostituiscono il campionamento e la raccolta tradizionale	Impatto limitato sul processo di produzione: integrazione della fase di stima	Impatto limitato o considerevole: stesso processo di produzione, sottoinsieme di dati raccolti tramite Internet o di metodi di stima basati sulla popolazione

Lato legislativo e organizzativo

- **Aspetti legislativi e di regolazione dell'accesso ai dati**
  - ✓ Possibili legislazioni differenti per i vari paesi
- **Privacy**
  - ✓ Possibili strategie di privacy-by-design
- **Aspetti finanziari**
  - ✓ Providers di Big data privati
- **Management**
  - ✓ Necessità di Training specifici

Lato statistico

- **Linkage** (con un grado di incertezza noto o stimato) degli eventi ai quali i Big Data si riferiscono, alle unità di popolazione di interesse per la statistica ufficiale (individui, famiglie, imprese o istituzioni)
- **Processare** i dati raccolti con l'obiettivo di renderli compatibili con il framework statistico di interesse (concetti, definizioni, classificazioni)
- **Attribuire pesi** (con incertezza nota o stimata) ai dati, in modo da garantire rappresentatività nei confronti della popolazione target
- **Stimare** aggregati di interesse fornendo misure della loro qualità basate sull'incertezza delle misurazioni negli step precedenti

Big Data nella statistica: principali opportunità e sfide (3)

Lato ICT : tecnologie

- **Migliorare l'efficienza nel processamento di dati, anche non-Big**
  - Parallelismo nell'esecuzione, processamento in-memory etc.
- **Abilitare l'uso di Big Data**
  - Necessità di avere tecnologie dedicate per consentire il trattamento dei Big Data

Lato ICT : metodologie

- **Estrazione semantica**
- **Aspetti legati ai processi:**
  - **Dimensioni big**, risolvibili tramite architetture dedicate, e.g. map-reduce/hadoop, cluster di db NoSQL, etc.
  - **Assenza di modelli**, risolvibile potenzialmente da tecniche di *machine-learning*
  - **Vincoli di Privacy**, risolvibili da tecniche di *privacy-preserving*
  - **Dati in streaming**, risolvibili da *event data management systems* e *OLAP in real-time*

FINE

Questa presentazione è stata realizzata anche grazie al materiale fornito da:

Giulio Barcaroli

Paolo Righi

Monica Scannapieco

(ISTAT)

Grazie dell'attenzione

stefano.defrancisci@istat.it