Il valore dei dati nell'era dei Big Data

Giorgio Alleva

Presidente dell'Istituto Nazionale di Statistica

Università di Napoli Federico II
Dipartimento di Scienze Politiche
Aula Spinelli

Istat

Outline

- Nuove sfide per gli istituti di statistica
- Prospettive di integrazione e uso di nuove fonti di dati
- Gli istituti di statistica alla prova dei big data
- Oltre la statistica: le questioni aperte e il "valore" dei dati
- Conclusioni

Istat

Nuove sfide per gli Istituti di Statistica

Istat

Nuove sfide per gli Istituti di Statistica

- "MISURARE" LA SOCIETÀ E L'ECONOMIA È UN COMPITO SEMPRE PIÙ COMPLESSO.
- AL CONTEMPO È IN AUMENTO LA DOMANDA DI INFORMAZIONE STATISTICA NELLA SOCIETÀ.
- CRESCE LA CAPACITÀ DI ARCHIVIARE, PROCESSARE E ANALIZZARE QUANTITÀ SEMPRE MAGGIORI DI DATI.
- IL SETTORE PRIVATO INVESTE UNA QUANTITÀ CRESCENTE DI RISORSE PER ELABORARE DATI E INFORMAZIONI.
- È ESSENZIALE CONTENERE IL FASTIDIO STATISTICO SUI RISPONDENTI E RIDURRE I COSTI COMPLESSIVI DELLA PRODUZIONE STATISTICA UFFICIALE.

CRESCE IL "VALORE" DEI DATI NELLA SOCIETÀ.

CRESCONO ANCHE LE SFIDE CUI GLI ISTITUTI DI STATISTICA DEVONO FAR FRONTE.

GROVES (2011) "A POSSIBLE DATA FUTURE FOR THE OBSERVATIONAL SOCIAL SCIENCES"

Istat

L'integrazione delle fonti

- LA CAPACITÀ DI **ESTRARRE VALORE DAI DATI** È LEGATA ALLA CAPACITÀ DI **INTEGRARE** DATI CHE PROVENGONO DA FONTI DIFFERENTI.

- SI TRATTA DI UN **PERCORSO INTRAPRESO** DA MOLTI ISTITUTI DI STATISTICA DEI PAESI AVANZATI.

- METTERE A CONFRONTO FONTI DIFFERENTI GARANTISCE GUADAGNI IN TERMINI DI **ACCURATEZZA**, **COERENZA**, **COMPLETEZZA** DELLE INFORMAZIONI STATISTICHE PRODOTTE.

Istat

Il processo di modernizzazione dell'Istat

- DA UN MODELLO «**TRADIZIONALE**», BASATO SULL'ACQUISIZIONE DIRETTA DEI DATI, AD UN MODELLO BASATO SULL'UTILIZZO DEI **REGISTRI STATISTICI**, ESSENZIALMENTE DERIVATI DALLE FONTI AMMINISTRATIVE E ALIMENTATI NEL CONTINUO DA FLUSSI TELEMATICI.

- SEBBENE IL PROCESSO DI MODERNIZZAZIONE **CAPITALIZZI ESPERIENZE** GIÀ COMPIUTE DALL'ISTAT SUL FRONTE DELL'INTEGRAZIONE DEI MICRODATI, ESSO RICHIEDE **RILEVANTI CAMBIAMENTI ORGANIZZATIVI** SUL FRONTE INTERNO.

- SUL FRONTE ESTERNO, È INVECE INDISPENSABILE **UN'INTENSA COLLABORAZIONE CON TUTTI I SOGGETTI** CHE RACCOLGONO INFORMAZIONI DI TIPO AMMINISTRATIVO.

- ACCELERARE IL PROCESSO DI EVOLUZIONE DEI MECCANISMI DI PRODUZIONE DELLE STATISTICHE È ESSENZIALE PER **AUMENTARE LA TEMPESTIVITÀ NELLA PRODUZIONE DEI DATI E L'ACCESSO DA PARTE DEI CITTADINI**.

Istat

Prospettive di integrazione e uso di nuove fonti di dati

La natura dei dati. Le survey

SURVEY (CAMPIONARIA O CENSUARIA)

INDAGINI STATISTICHE PIANIFICATE AD HOC

SPECIFICA POPOLAZIONE OBIETTIVO

DEFINIZIONI, CONCETTI E CLASSIFICAZIONI DEFINITE EX-ANTE

QUESITI MIRATI

STIME BASATE SUL PARADIGMA INFERENZIALE TRADIZIONALE (NEL CASO DI CAMP.)

TECNOLOGIE E STRUMENTI DI ANALISI NON PARTICOLARMENTE SOFISTICATI

MA…

COSTI ELEVATI

ELEVATA PRESSIONE STATISTICA SUI RISPONDENTI

NEL TEMPO I TASSI DI RISPOSTA DELLE SURVEY SONO PROGRESSIVAMENTE DIMINUITI.

Istat

La natura dei dati. I dati amministrativi

ARCHIVI AMMINISTRATIVI (ANAGRAFI, BANCHE DATI REDDITUALI, ARCHIVI MINISTERI, ETC.)

RIDUZIONE DEI COSTI E DEL FASTIDIO STATISTICO

AUMENTO DEL DETTAGLIO (SOTTO-POPOLAZIONI E LIVELLI TERRITORIALI)

COERENZA DEL CONTESTO IN CUI VENGONO PRODOTTI I DATI

MA…

POPOLAZIONE OBIETTIVO ≠ POPOLAZIONE AMMINISTRATIVA

DEFINIZIONI E CLASSIFICAZIONI POSSONO NON COINCIDERE CON QUELLI UTILIZZATI DALLA STATISTICA UFFICIALE (AD ES. UNITÀ AMMINISTRATIVA ≠ UNITÀ STATISTICA)

L'ACCESSO AI DATI PUÒ ESSERE PROBLEMATICO

VALUTARE DISPONIBILITÀ E QUALITÀ DEI DATI AMMINISTRATIVI

È NECESSARIO TRADURRE IL SEGNALE AMMINISTRATIVO IN INFORMAZIONE STATISTICA DI QUALITÀ!

L'USO DI DATI AMMINISTRATIVI VIENE FORTEMENTE RACCOMANDATO DAL SISTEMA STATISTICO EUROPEO.

Istat

La natura dei dati. I Big Data

BIG DATA (DATI ORIGINATI DALL'USO DEGLI STRUMENTI DIGITALI)

- REGISTRANO EVENTI, SPESSO REGISTRANO "COMPORTAMENTI" (SPONTANEI)
- AMPLIANO LE OPPORTUNITÀ DI ANALISI E LE INFORMAZIONI DISPONIBILI
- DATI TEMPESTIVI, GENERATI AD UN COSTO ESTREMAMENTE CONTENUTO

MA…

- POPOLAZIONE OBIETTIVO ≠ POPOLAZIONE BIG DATA
- DEFINIZIONI E CLASSIFICAZIONI DI SOLITO NON COINCIDONO CON QUELLI UTILIZZATI DALLA STATISTICA UFFICIALE
- L'ACCESSO AI DATI PUÒ ESSERE PROBLEMATICO
- VALUTARE DISPONIBILITÀ E QUALITÀ DEI DATI
- PROBLEMI TECNOLOGICI DOVUTI AL TRATTAMENTO DI INGENTI QUANTITÀ DI DATI
- DIFFICOLTÀ NELL'ESTRARRE L'INFORMAZIONE RILEVANTE
- …

È NECESSARIO UN GRANDE IMPEGNO PER ESTRARRE VALORE DAI BIG DATA! I METODI FINORA UTILIZZATI NON SONO SUFFICIENTI!

Istat

I vantaggi dell'integrazione

- L'UTILIZZO DI **DATI AMMINISTRATIVI** E LA LORO INTEGRAZIONE PERMETTERÀ DI:
  - AUMENTARE IL DETTAGLIO DI ANALISI
  - METTERE INSIEME I PERCORSI SOCIALI ED ECONOMICI DI INDIVIDUI E IMPRESE ("SCRIVERE" LE STORIE INDIVIDUALI)
  - CONNETTERE A LIVELLO MICRO I FENOMENI ECONOMICI E SOCIALI.

- LE **SURVEY** CONTINUERANNO AD ESSERE UTILIZZATE PER COMPLETARE IL QUADRO INFORMATIVO, ANALIZZARE FENOMENI SPECIFICI, FORNIRE RISPOSTE A DETERMINATE CHIAVI DI LETTURA, INDIVIDUARE NUOVI TREND.

- NEL FUTURO I **BIG DATA** SARANNO UTILI PER AMPLIARE LE OPPORTUNITÀ DI ANALISI, AUMENTARE LA TEMPESTIVITÀ DELLE INFORMAZIONI, CONTRIBUIRE A MIGLIORARE LA QUALITÀ DELLE STIME.

L'importanza del microdati. Esplorare l'eterogeneità

- LA MAGGIORE DISPONIBILITÀ DI MICRODATI INTEGRATI GARANTIRÀ **NUOVE OPPORTUNITÀ DI RICERCA PER LA STATISTICA UFFICIALE**.

- IL MICRO-DATO DARÀ ANCHE A RICERCATORI E POLICY MAKERS L'OPPORTUNITÀ DI STUDIARE RELAZIONI PIÙ COMPLESSE, VERIFICARE L'IMPATTO DELLE POLITICHE, ANALIZZARE L'EVOLUZIONE DEI FENOMENI SOCIALI.

- GLI ISTITUTI DI STATISTICA STANNO ESPLORANDO **NUOVE STRATEGIE PER DARE ACCESSO AI MICRODATI** SENZA INCORRERE IN QUESTIONI DI PRIVACY E CONFIDENZIALITÀ.

Istat

Gli Istituti di Statistica alla prova dei Big Data

Istat

I fattori critici

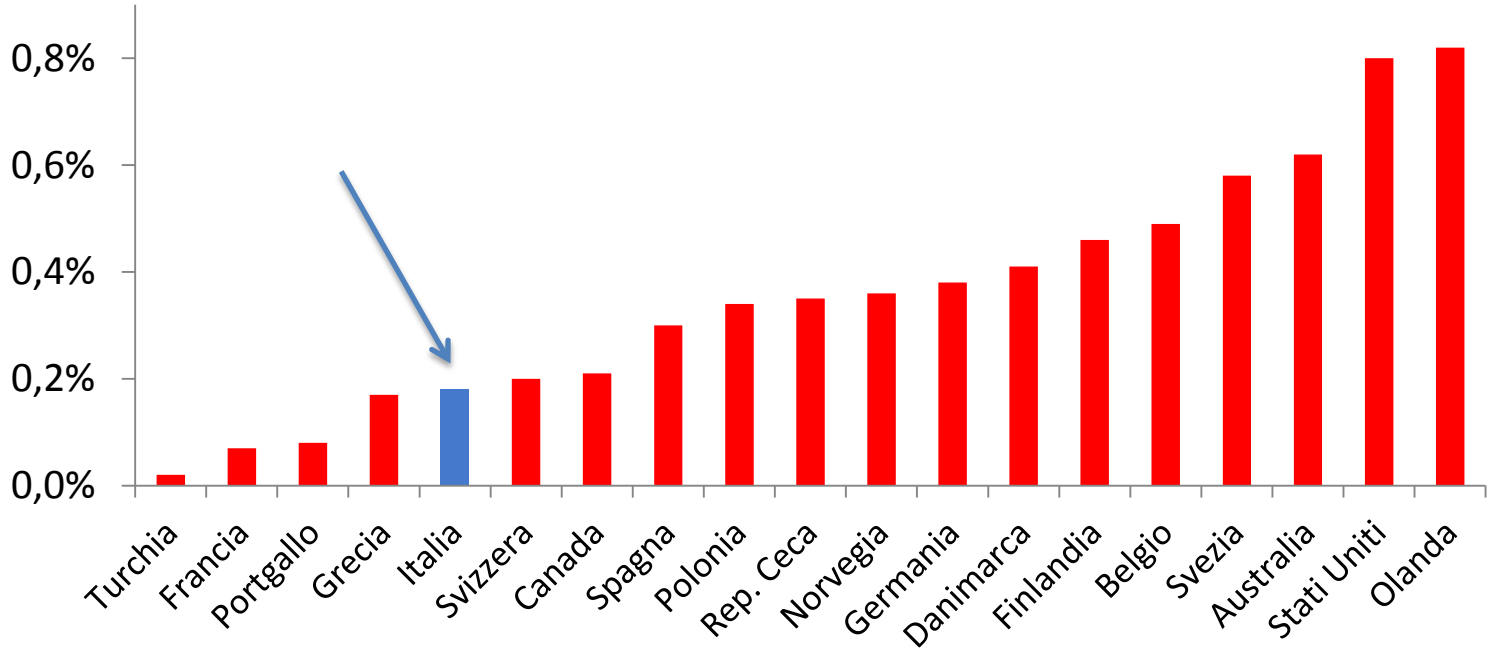
- L'ACCESSO AI DATI
- PRIVACY E CONFIDENZIALITÀ
- NUOVE INFRASTRUTTURE:
  - METODOLOGICHE
  - TECNOLOGICHE
  - ORGANIZZATIVE
- NUOVE COMPETENZE

Istat

Le competenze. I "data scientist" nel mondo

L'OCSE STIMA CHE NEL 2013 IL NUMERO DI "DATA SCIENTIST" ERA INFERIORE ALL'1% DELL'OCCUPAZIONE NELLA MAGGIOR PARTE DEI PAESI. PER L'ITALIA TALE QUOTA È DELLO 0,2% (2014).

QUOTA DATA SCIENTIST SUL TOTALE DELL'OCCUPAZIONE NEI PRINCIPALI PAESI OCSE – ANNO 2013
(VALORI PERCENTUALI)

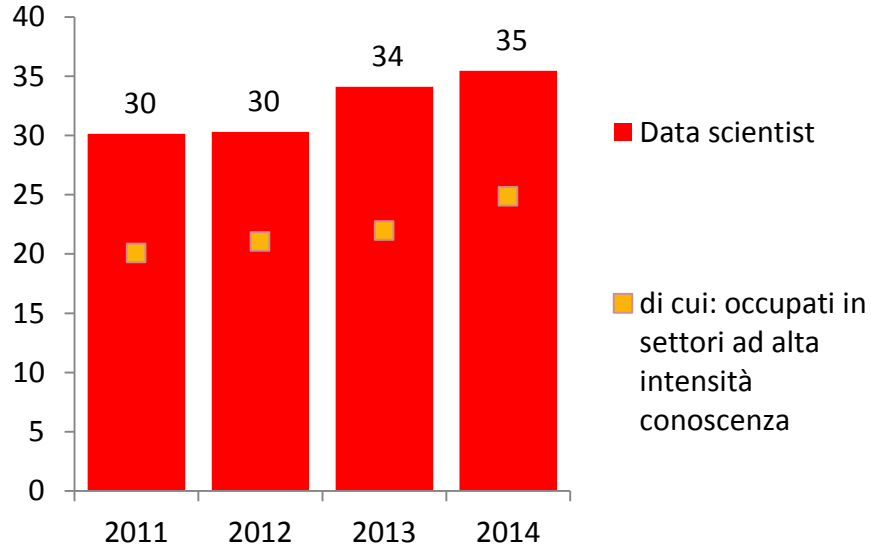


FONTE: EUROSTAT, STATISTICS CANADA, AUSTRALIAN BUREAU OF STATISTICS LABOUR FORCE SURVEYS AND US CURRENT POPULATION SURVEY, MARCH SUPPLEMENT, FEBRUARY 2015.

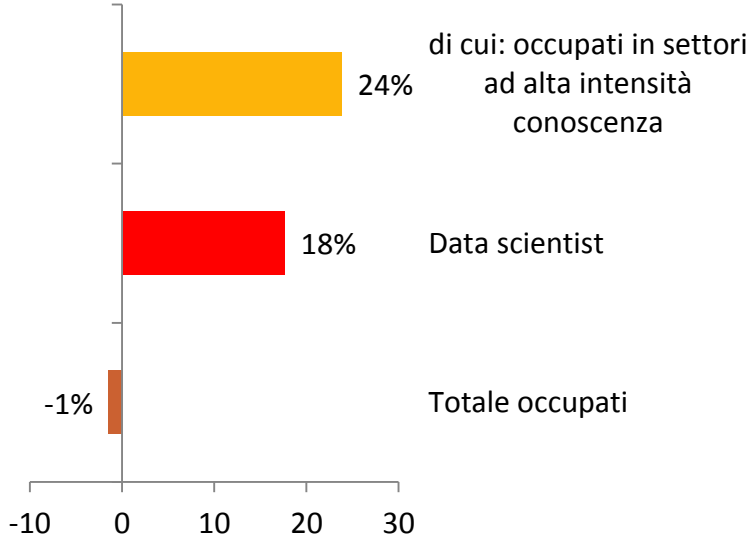
Le competenze. I "data scientist" in Italia

- In Italia il numero di "data scientist" è in ascesa.
- I settori di attività nei quali risultano maggiormente occupati sono la produzione di software e consulenza informatica e il settore pubblico.

NUMERO DI "DATA SCIENTIST" IN ITALIA
2011-2014 – VALORI IN MIGLIAIA



TASSO DI VARIAZIONE DELL'OCCUPAZIONE
2011-2014 – VALORI PERCENTUALI



FONTE: ISTAT, RILEVAZIONE FORZE LAVORO.

A che punto sono gli Istituti di Statistica con i Big Data?

Survey UNECE (90 paesi + Eurostat, 115 progetti)

United Nations
Statistics Division

- **Fonti utilizzate**: scanner data, satellite imagery, web-scraping data
- **Principali ragioni per l'utilizzo dei Big Data**: Faster statistics, reduce response burden, modernize production
- **Collaborazioni**: government institutes, satellite provider, research and academics
- **Need for guidance**: Skills and training, quality framework, access

Tuttavia, negli istituti di statistica è ancora assente una visione di lungo periodo sull'utilizzo dei Big Data.

R. Jansen (2015) "Global assessment of Big Data for Official Statistics", UN Statistics Division.

Istat

L'esperienza dell'Istat nell'uso dei Big Data

Dal 2013 l'Istat ha avviato vari progetti sull'uso dei Big Data:

- Persons and Places (Mobile Phone Data)
- Labour Market Estimation (Google Trends)
- Scanner data
- ICT Usage by enterprises and "internet as a data source" (web-scraping)
- Social media (Twitter, Facebook)

Istat ha implementato infrastrutture e software per il trattamento dei Big Data: Sandbox e Cloudera.

I progetti vedono la collaborazione di imprese, università, centri di ricerca.

Passare dalla sperimentazione alla produzione!

Istat

Un nuovo framework per valutare la qualità dei Big Data

- I QUALITY FRAMEWORK TRADIZIONALI NON SONO SUFFICIENTI AD AFFRONTARE LA COMPLESSITÀ DEI BIG DATA! È NECESSARIO RIVISITARE LE USUALI "DIMENSIONI" DELLA QUALITÀ E PROPORNE DI NUOVE.
  - L'AMBIENTE IN CUI SONO PRODOTTI I DATI
  - LA PRIVACY E LA SICUREZZA DEI DATI
  - LA COMPLESSITÀ DEI DATI (STRUTTURA, FORMATO,...)
  - L'UTILIZZABILITÀ
  - LA RAPPRESENTATIVITÀ
  - LA "LINKABILITÀ"
  - LA VALIDITÀ

UNECE (2014) "A SUGGESTED FRAMEWORK FOR THE QUALITY OF BIG DATA"

Istat

Oltre la statistica: le questioni aperte e il "valore" dei dati

Oltre la statistica: le questioni aperte e il "valore" dei dati

PRIVACY

QUALI LIMITI ALL'UTILIZZO DEI BIG DATA?

DISCRIMINAZIONE

DISUGUAGLIANZE NELL'ACCESSO AI DATI?

CONTROLLO

INTERESSI COLLETTIVI VS INTERESSI PRIVATI?

DEMOCRATIZZAZIONE NELL' UTILIZZO

RISCHI DI CONFUSIONE?

BENESSERE E IMPATTO SUI CITTADINI

COME UTILIZZARE I BIG DATA PER AUMENTARE IL BENESSERE INDIVIDUALE E COLLETTIVO?

Istat

Conclusioni

Conclusioni

- I DATI: INFRASTRUTTURA CHIAVE PER IL XXI SECOLO.
- È IMPORTANTE CHE I DECISORI PUBBLICI BASINO LE LORO SCELTE SU DATI E ANALISI DI QUALITÀ.
- NUOVO RUOLO E SFIDE URGENTI PER LA STATISTICA UFFICIALE CON L'ASCESA DEI BIG DATA: DALLA SPERIMENTAZIONE ALLA PRODUZIONE.
- EDUCARE ALLA STATISTICA E AL VALORE DEI DATI, COINVOLGENDO I CITTADINI NEL CICLO DI PRODUZIONE DELLA STATISTICA UFFICIALE.

Extra-slide:
Le esperienze dell'Istat sull'uso dei Big Data

Il progetto "Persons and Places"

- La finalità del lavoro è quella di **integrare l'uso di dati anonimizzati di telefonia mobile** nel processo statistico di stima di flussi di popolazione intercomunale, utilizzando i cosiddetti **Call Data Record (CDR)** forniti dalle compagnie telefoniche.

- Le potenzialità sono enormi:

  - **Aumentare** l'efficienza dei sistemi urbani e promuovere la loro integrazione

  - **Anticipare** la domanda sociale di infrastrutture e servizi di trasporto.

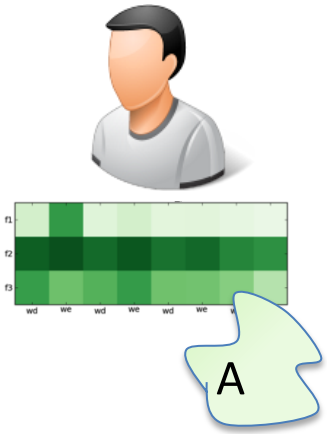
- Attori coinvolti nel progetto pilota:

  Istat, Cnr, Università di Pisa.

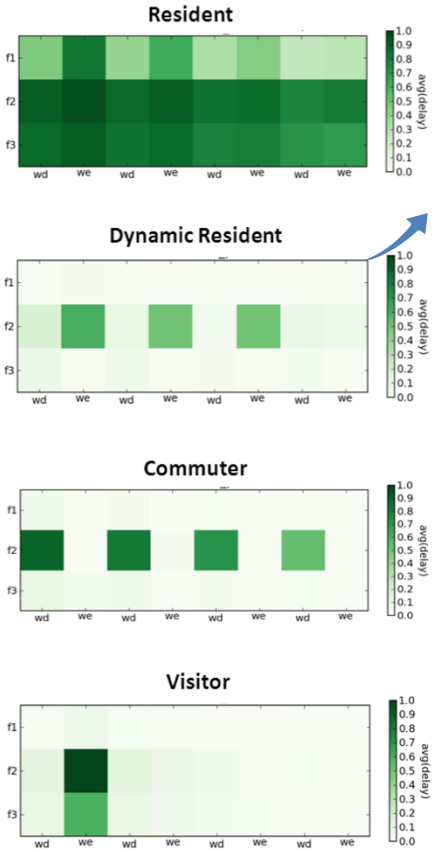
BACK

Il progetto pilota "Persons and Places"

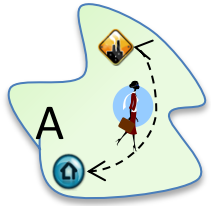
PROFILO DI CHIAMATA INDIVIDUALE



ALGORITMO DI CLASSIFICAZIONE

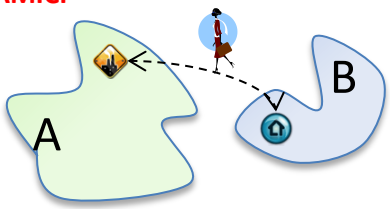


RESIDENTI STATICI

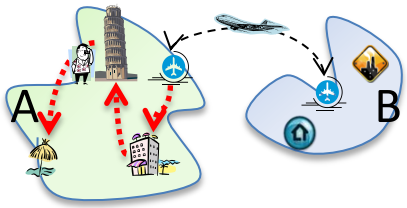


RESIDENTI DINAMICI

PENDOLARI



VISITATORI



BACK

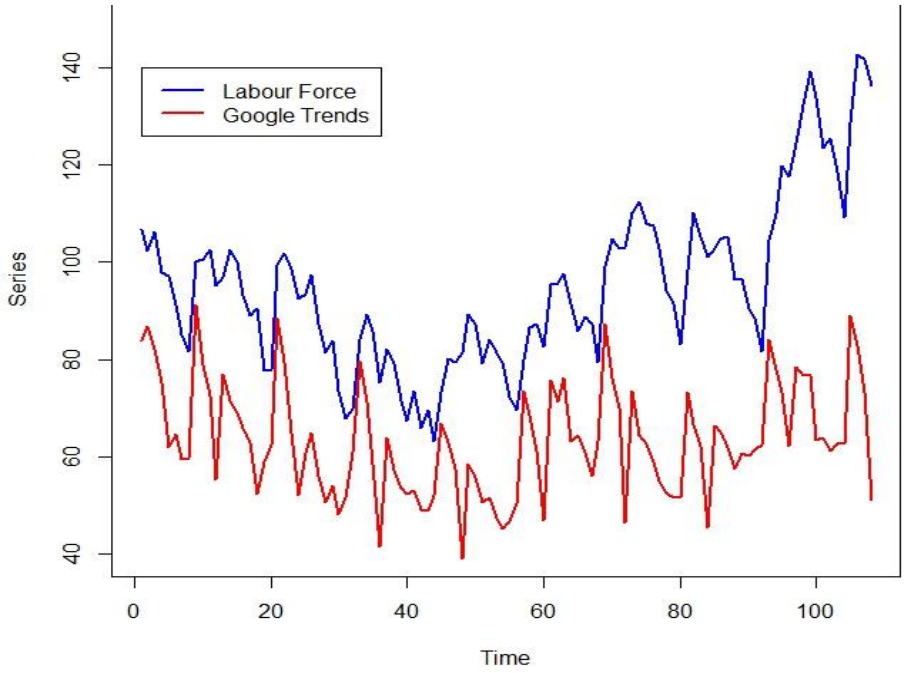
Istat

Google Trends

GOOGLE TRENDS PUÒ ESSERE UTILIZZATO PER MIGLIORARE LE STIME PRODOTTE DALL'ISTAT SUL MERCATO DEL LAVORO IN TERMINI DI PREVISIONI E NOWCASTING.

- SI AVVICINA IL CICLO DEI DATI A QUELLO DELLE **DECISIONI.**
- SI AMPLIA LA CAPACITÀ DI **DETTAGLIO TERRITORIALE** DEGLI INDICATORI SUL LAVORO.
- SI ATTENUA IL TRADE-OFF TRA **ACCURATEZZA** E **TEMPESTIVITÀ.**

TASSO DI DISOCCUPAZIONE MENSILE (RFL) E OFFERTA DI LAVORO (GOOGLE TRENDS) - INDICE 2004=100



BACK

Istat

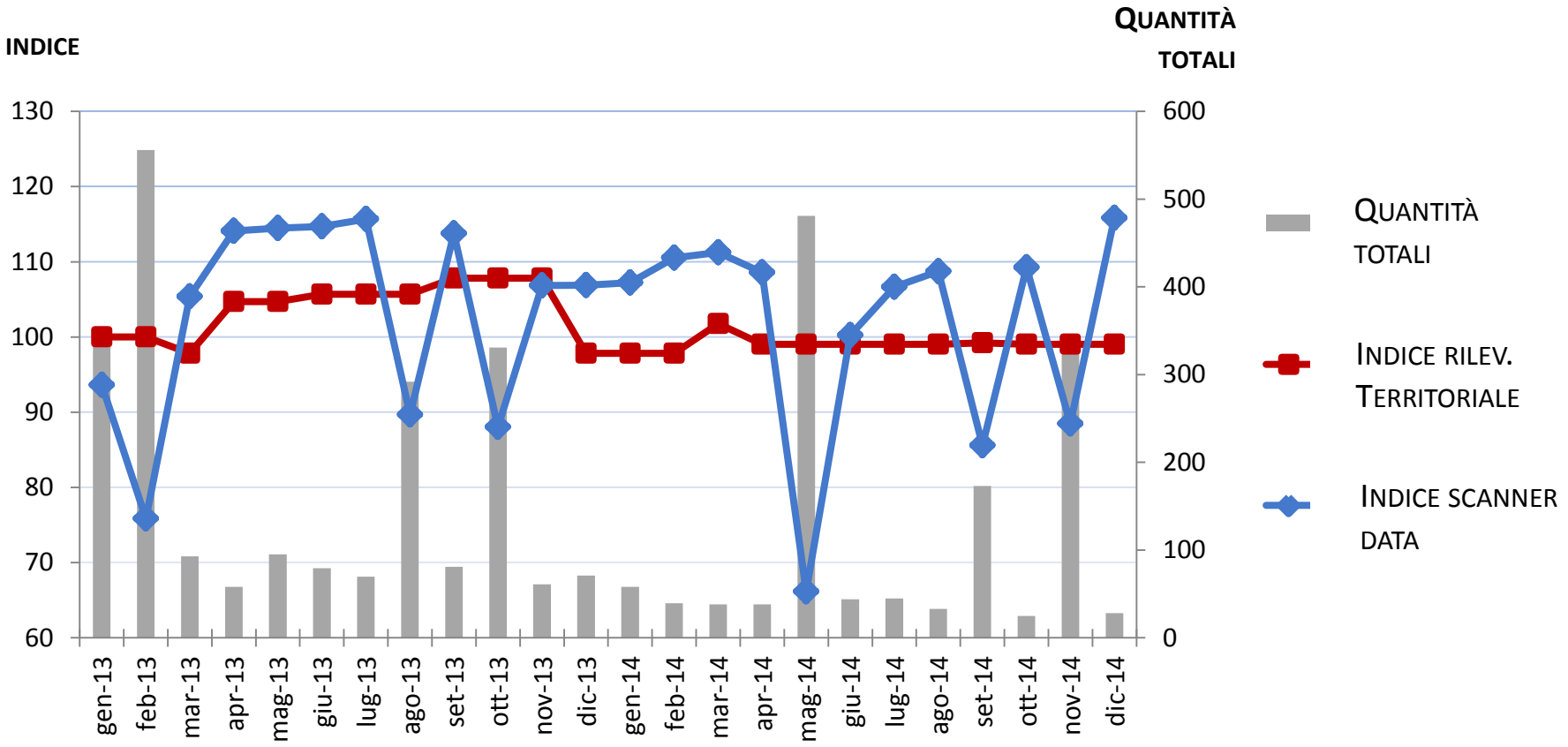
Scanner data

- **REVISIONE** IN CORSO DELL'ORGANIZZAZIONE DELLA RILEVAZIONE SUI PREZZI AL CONSUMO A PARTIRE DALLA STRATEGIA CAMPIONARIA DELL'INDAGINE.
- L'OBIETTIVO È QUELLO DI UTILIZZARE LE NUOVE FONTI DI DATI (SCANNER DATA E WEB SCRAPING) PER **COLMARE IL GAP INFORMATIVO** E RISPONDERE ALL'ULTERIORE E CRESCENTE ARTICOLAZIONE DELLA DOMANDA DI INFORMAZIONE STATISTICA SUI PREZZI AL CONSUMO, SOPRATTUTTO A LIVELLO TERRITORIALE.
- DALLA FINE DEL 2013, CON **ADM** E **GDO** È STATO AVVIATO UN TAVOLO INFORMALE PER L'ACQUISIZIONE DEGLI **SCANNER DATA**.

BACK

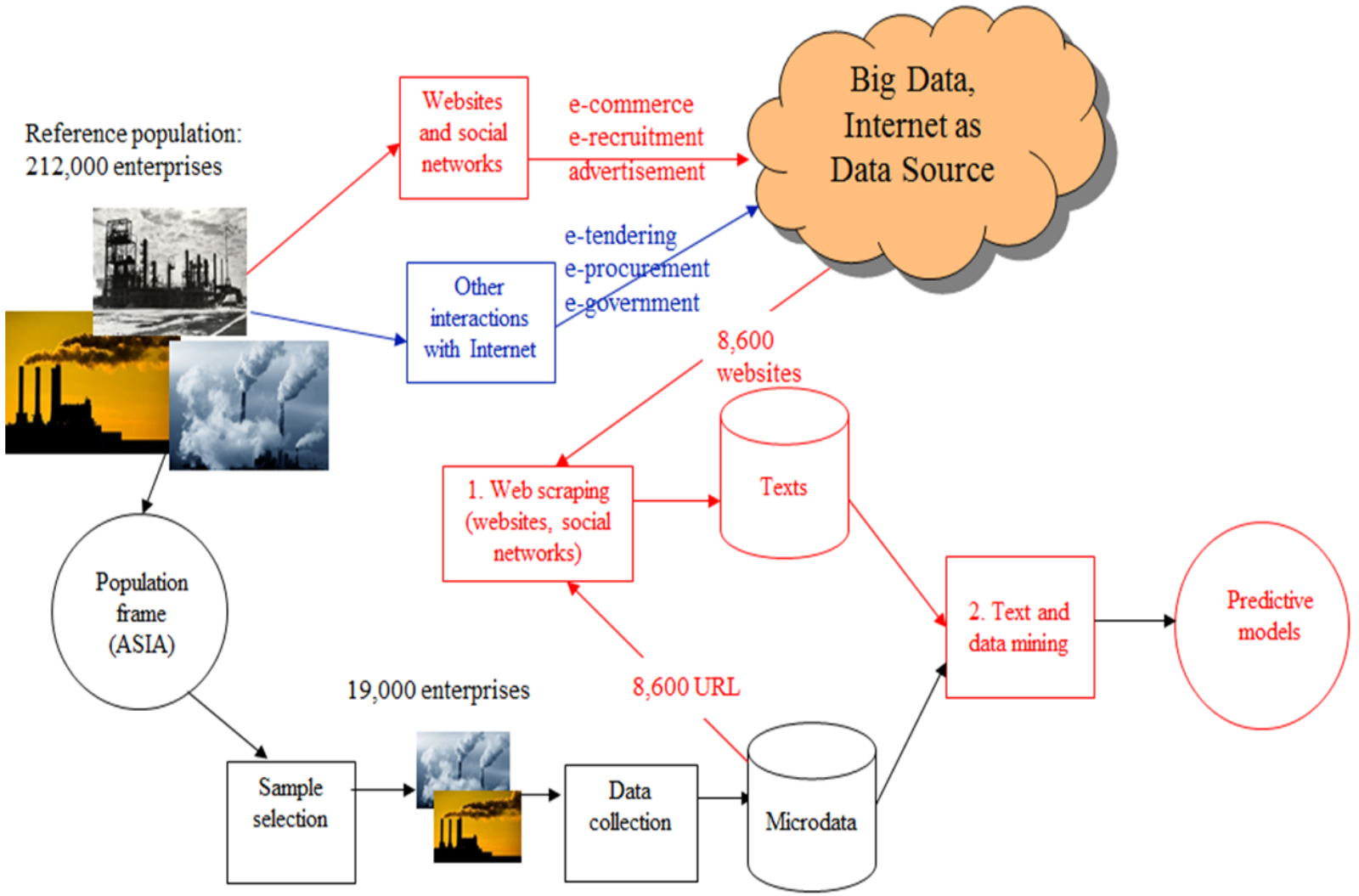
Scanner data. Il prezzo del caffé

INDICE DEI PREZZI AL CONSUMO DI SINGOLA REFERENZA DEL CAFFÈ TOSTATO E QUANTITÀ VENDUTE NEL MESE. COMPARAZIONE TRA INDICE SCANNER DATA E INDICE CALCOLATO SULLA BASE DEI DATI DELLA RILEVAZIONE TERRITORIALE. GEN 2013 – DIC 2014



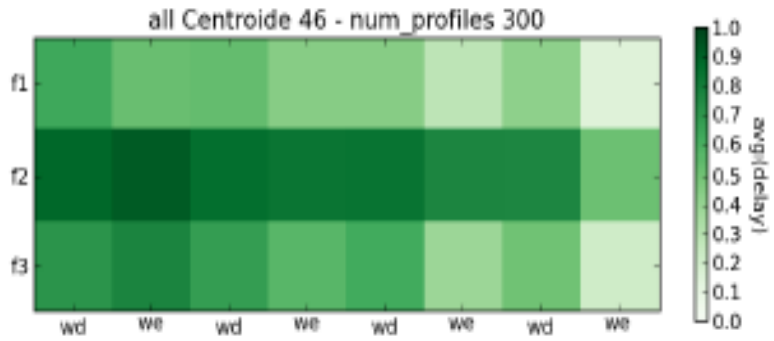
BACK

ICT nelle imprese: tecniche di Web Scraping e Text Mining

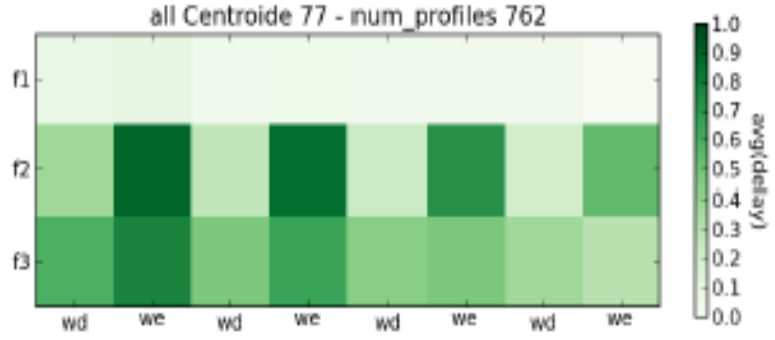


BACK

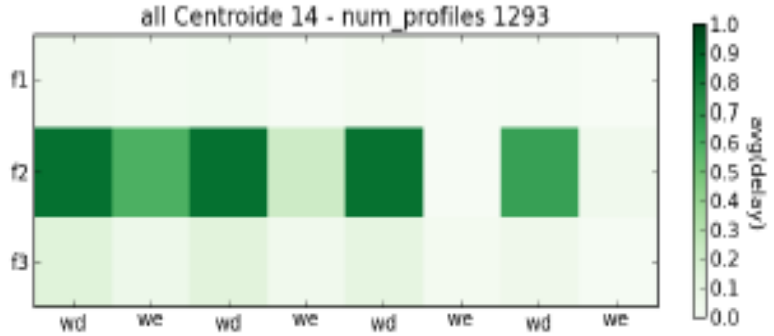
Il progetto pilota "Persons and Places"



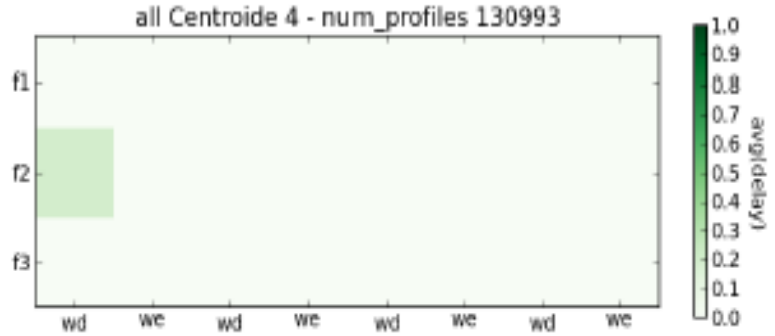
(a) Residents



(b) Dynamic residents



(c) Commuters



(d) Visitors

BACK

Istat