

# 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*



- 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

# Nuove sfide per gli Istituti di Statistica

# 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.**

## 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.

# 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**.



# 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.**



# 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  $\neq$  POPOLAZIONE AMMINISTRATIVA

DEFINIZIONI E CLASSIFICAZIONI POSSONO NON COINCIDERE CON QUELLI UTILIZZATI DALLA STATISTICA UFFICIALE (AD ES. UNITÀ AMMINISTRATIVA  $\neq$  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.**



# 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  $\neq$  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

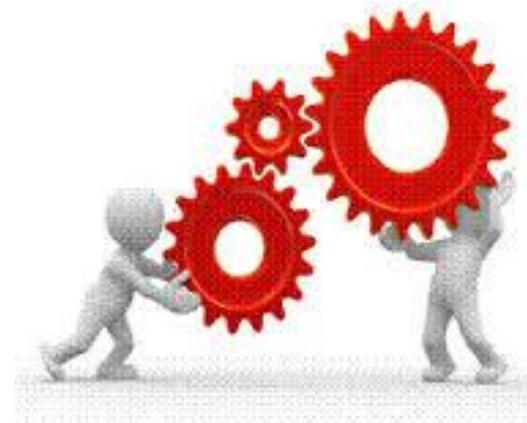
...

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



# 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 dei 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.

# Gli Istituti di Statistica alla prova dei Big Data

## I fattori critici

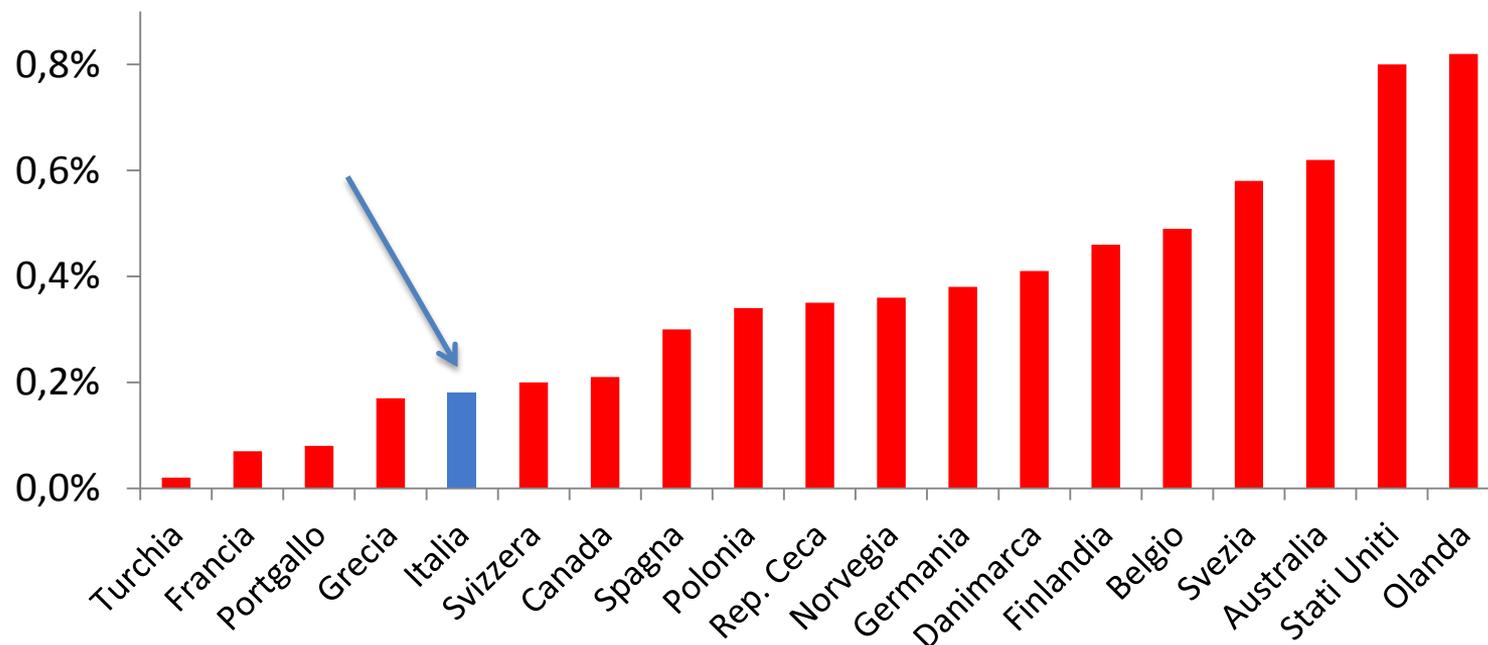
- L'ACCESSO AI DATI
- PRIVACY E CONFIDENZIALITÀ
- NUOVE INFRASTRUTTURE:
  - METODOLOGICHE
  - TECNOLOGICHE
  - ORGANIZZATIVE
- NUOVE COMPETENZE



## 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)

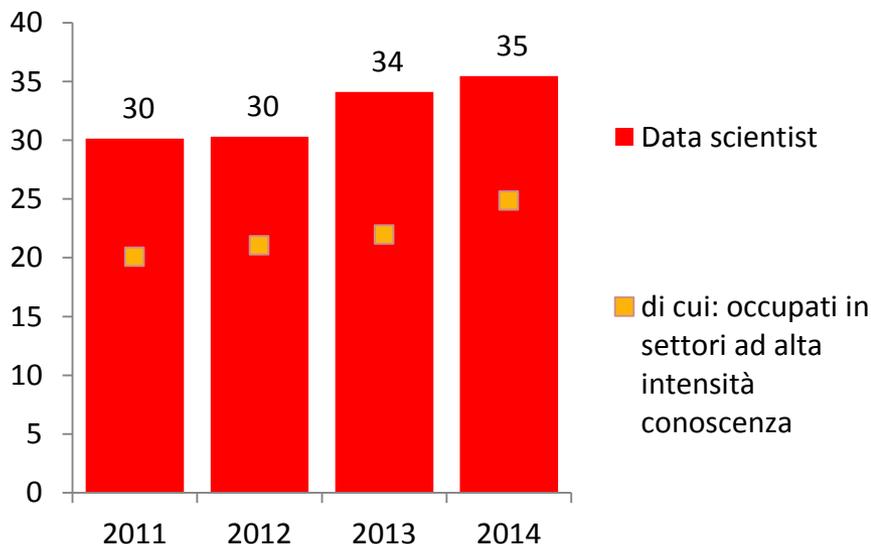


FONTE: EUROSTAT, STATISTICS CANADA, AUSTRALIAN BUREAU OF STATISTICS LABOUR FORCE SURVEYS AND US CURRENT POPULATION SURVEY, MARCH SUPPLEMENT, FEBRUARY 2015.

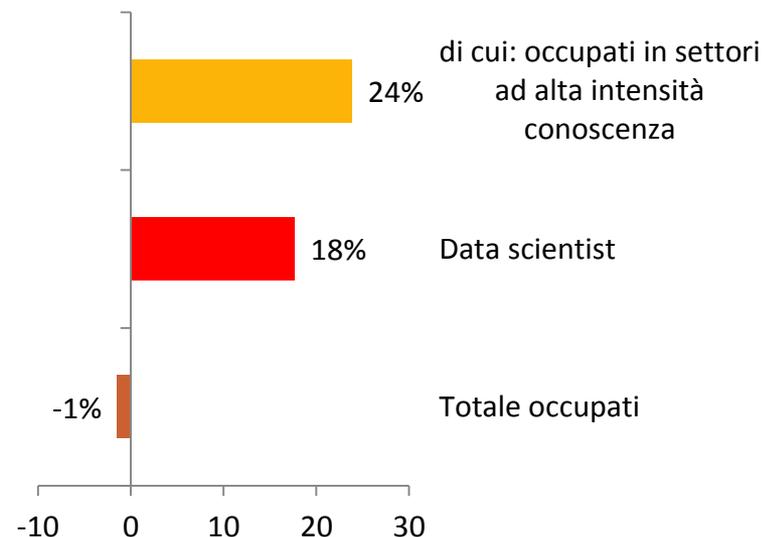
# Le competenze. I "data scientist" in Italia

- 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.**

# 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 CLLOUDERA.

I PROGETTI VEDONO LA COLLABORAZIONE DI IMPRESE, UNIVERSITÀ, CENTRI DI RICERCA.

**PASSARE DALLA SPERIMENTAZIONE ALLA PRODUZIONE!**

# 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

# 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?



# 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:**

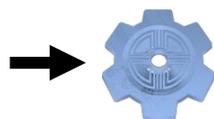
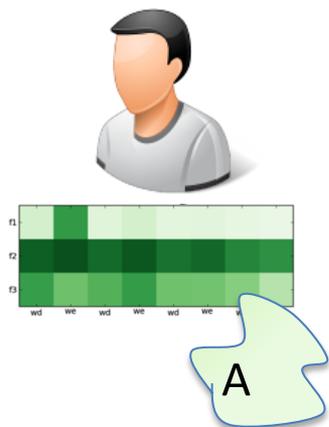
**Le esperienze dell'Istat sull'uso dei Big  
Data**

# Il progetto "Persons and Places"

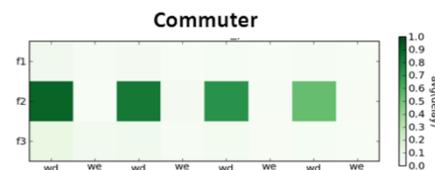
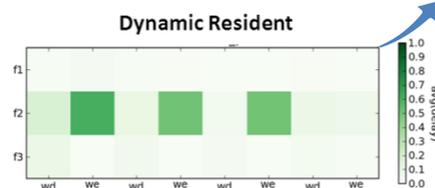
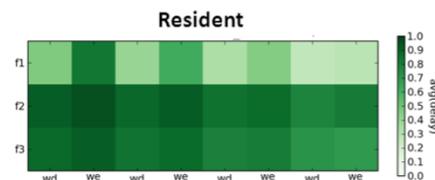
- LA FINALITÀ DEL LAVORO È QUELLA DI **INTEGRARE L'USO DI DATI ANONIMIZZATI DI TELEFONIA MOBILE** NEL PROCESSO STATISTICO DI STIMA DI FLUSSI DI POPOLAZIONE INTERCOMUNALE, UTILIZZANDO I COSIDDETTI **CALL DATA RECORD (CDR)** FORNITI DALLE COMPAGNIE TELEFONICHE.
- LE POTENZIALITÀ SONO ENORMI:
  - **AUMENTARE** 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.

# Il progetto pilota "Persons and Places"

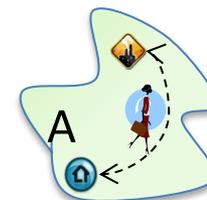
## PROFILO DI CHIAMATA INDIVIDUALE



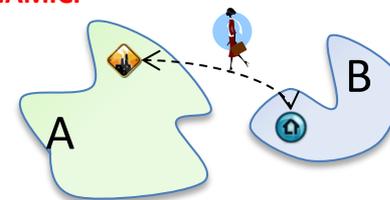
ALGORITMO DI CLASSIFICAZIONE



RESIDENTI STATICI

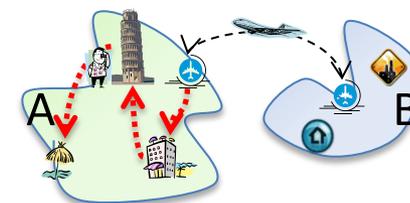


RESIDENTI DINAMICI



PENDOLARI

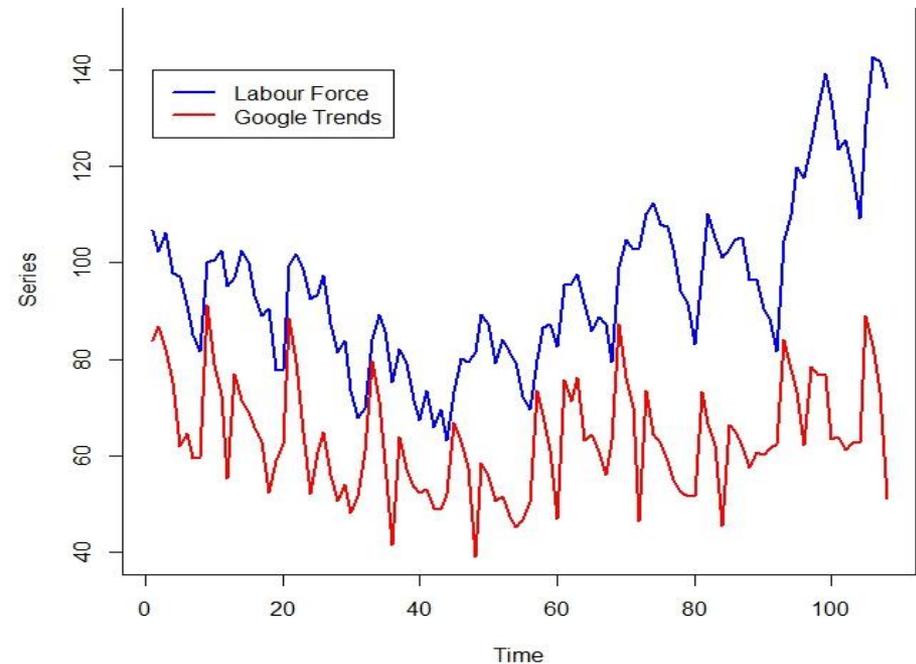
VISITATORI



**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**



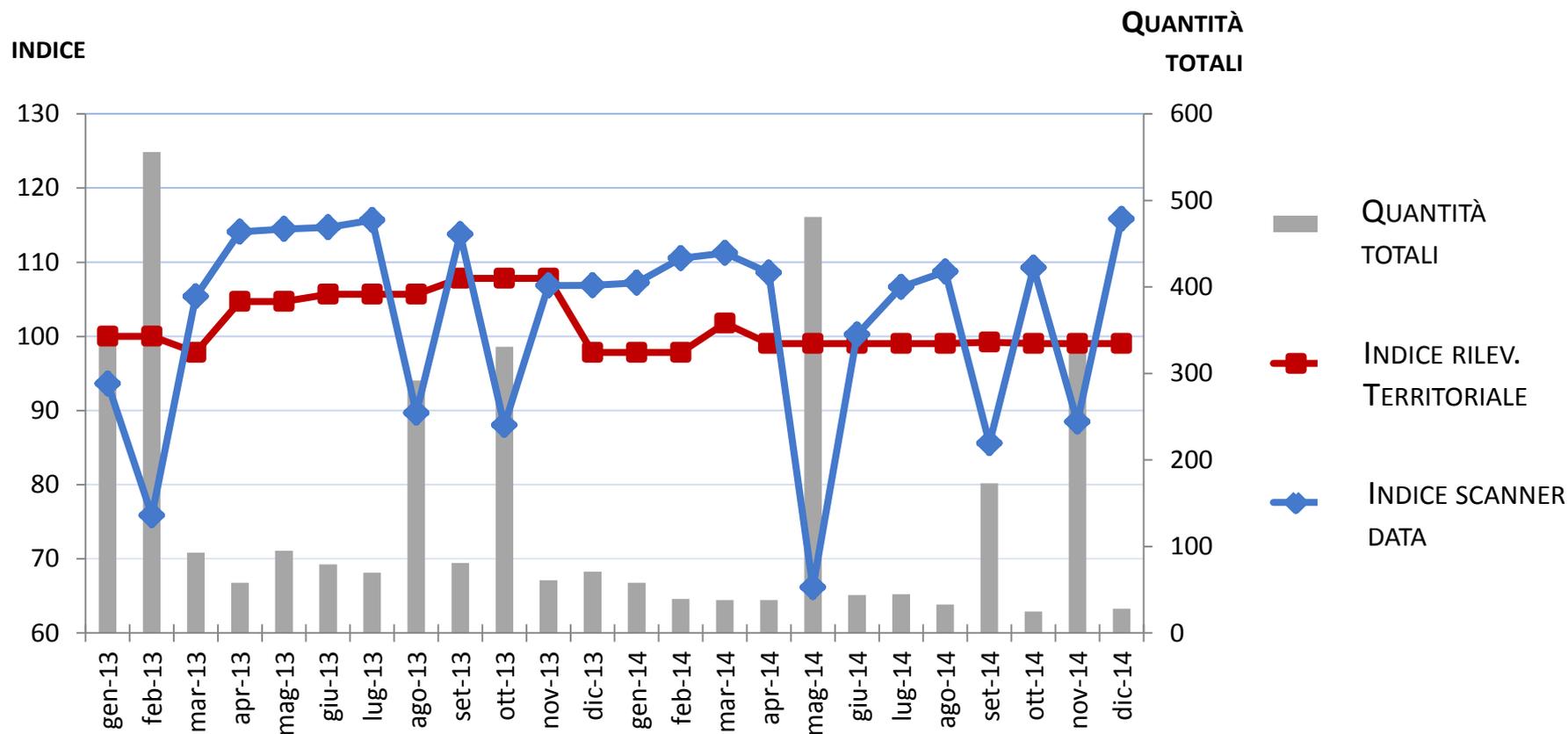
## Scanner data

- **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**.

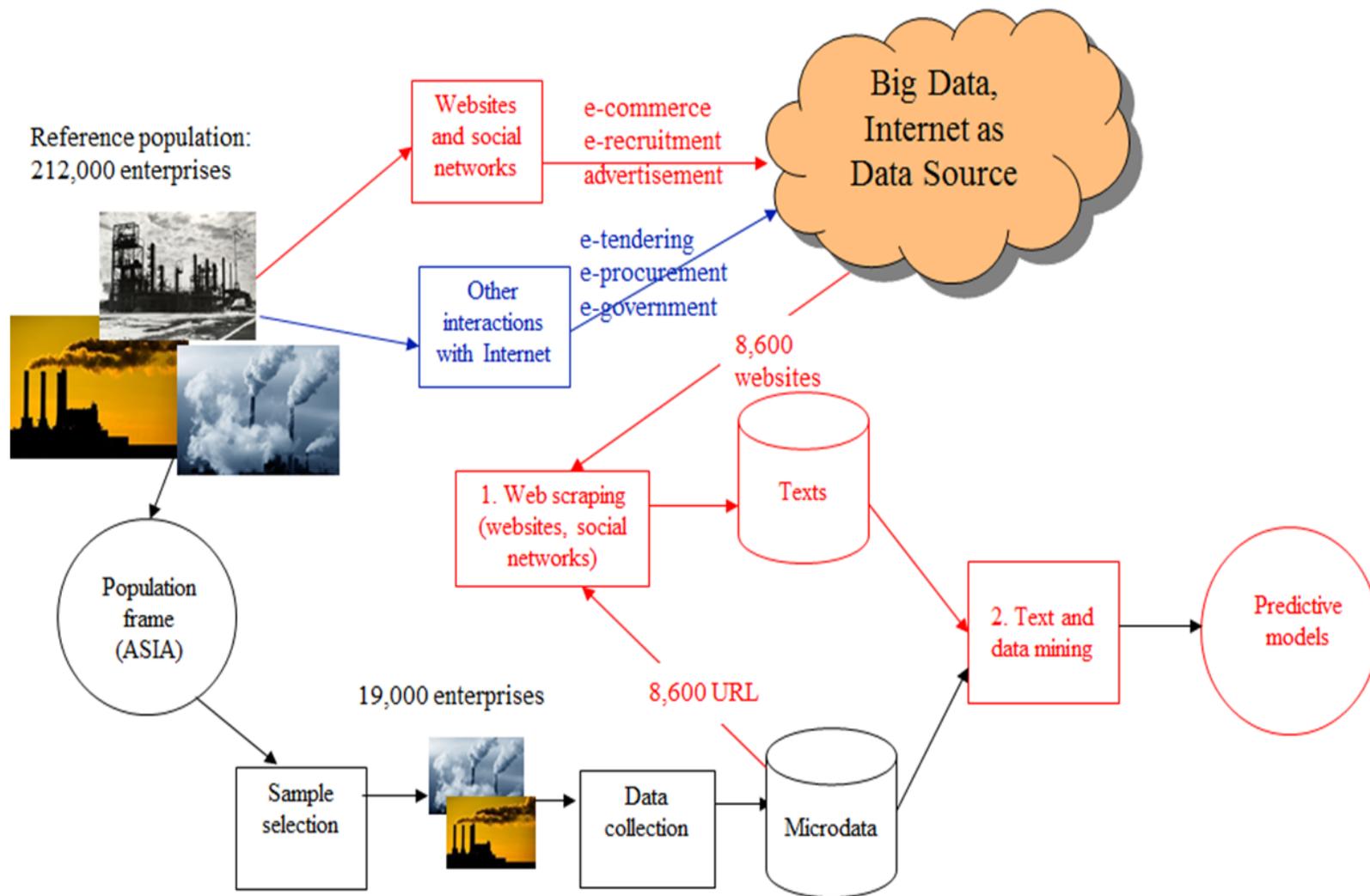


# Scanner data. Il prezzo del caffè

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

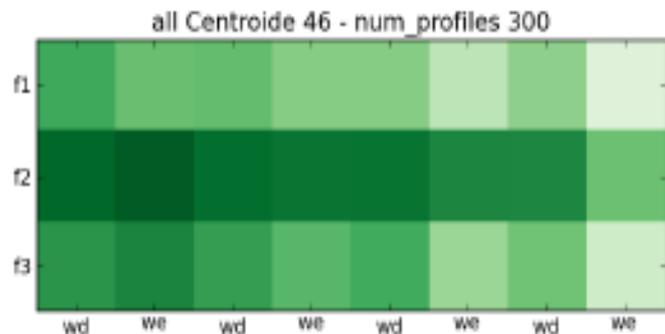


# ICT nelle imprese: tecniche di Web Scraping e Text Mining

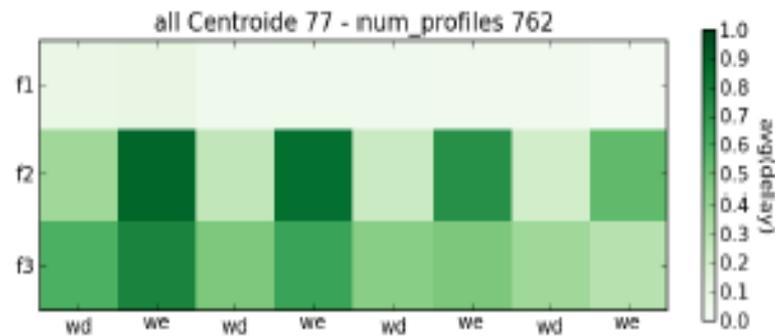




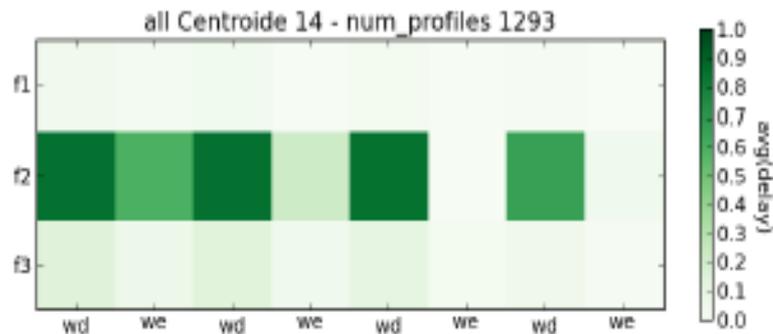
# Il progetto pilota "Persons and Places"



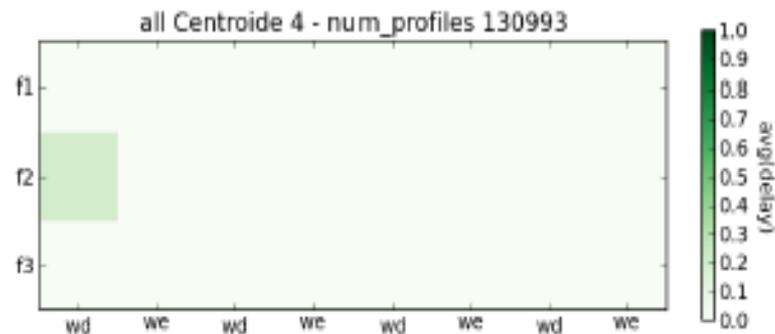
(a) Residents



(b) Dynamic residents



(c) Commuters



(d) Visitors