SOSE WEST

Workshop Economico, Statistico e Tecnologico

DATA SCIENCE

PER LE POLITICHE

PUBBLICHE

Roma - 30 OTTOBRE 2019

BIG DATA E STATISTICA UFFICIALE

Monica Pratesi- Università di Pisa

Presidente della

Società Italiana di Statistica

sose

SFIDE PER LE SCIENZE STATISTICHE E LA STATISTICA UFFICIALE

- Dati da usare
- Competenze da attivare
- Metodi da innovare
- Governance dei dati da organizzare

naturale evoluzione?

John Tukey (1962, 1977), Jean-Paul Benzécri (1979), Leo Breiman (1996, 2001), Eurostat (2013), UNECE (2013), Jerome H. Friedman (2001), Piercesare Secchi (2018), Giorgio Alleva (2019)

SFIDE PER LE SCIENZE STATISTICHE E LA STATISTICA UFFICIALE

Nuovi elementi?

1 Interesse di industrie, governi e media: crescita di responsabilità, necessarie risposte giuste a domande ben poste

2 Opportunità e i grandi rischi: straordinaria disponibilità di grandi quantità di dati, dati amministrativi, fenomeni inesplorati, “big datum” di grandi società private fuori dai sistemi statistici nazionali (SSN)

3 Mercato dei dati: ingenti risorse investite, diverso quadro giuridico (privati, ISN, Agenzie nazionali); fiducia dei cittadini

SFIDE PER LE SCIENZE STATISTICHE E LA STATISTICA UFFICIALE

Rischi?

5 Il ruolo di spicco dell'informatica e dell'architettura dei dati in questa fase

6 Il rischio di perdita della centralità di alcuni principi fondamentali per trarre conclusioni scientifiche da i dati, come fornire una misura dell'incertezza per le dichiarazioni scientifiche basate su dati

7 Inferenza induttiva: dai dati ai modelli e alle conclusioni scientifiche, anche con l'avvento di enormi set di dati. Intelligenza Artificiale, machine learning, (statistical) deep learning, gestione dell'incertezza

SFIDE PER LE SCIENZE STATISTICHE E LA STATISTICA UFFICIALE

Dibattiti tra culture diverse:

- Il dualismo di approcci alla scienza basati sui dati (data driven) e basati sui modelli (model based)
- Molti strumenti diversificati e necessità di utilizzare i dati (tanti, da integrare) per risolvere i problemi (story telling, data4policy)
- Riconoscimento e gestione dell'errore e dell'incertezza : approcci basati su modelli stocastici o procedure algoritmiche nell'ambito dell'intelligenza artificiale (AI) e dell'ingegneria della conoscenza (KE)

SFIDE PER LE SCIENZE STATISTICHE E LA STATISTICA UFFICIALE

Coesistenza di fonti diverse:

- Dati statistici raccolti mediante indagini tradizionali : dati raccolti in diretto contatto con le unità selezionate nella popolazione e trattati al fine di produrre stime;
- Dati amministrativi : dati frutto di procedure amministrative (sicurezza sociale, sanità, istruzione, carte d'identità, dati contabili interni, etc.);
- Big Data : dati originati dall'uso di dispositivi digitali, nel senso più ampio del termine.

SFIDE PER LE SCIENZE STATISTICHE E LA STATISTICA UFFICIALE

- Nuovi termini per **modern antiques**: *Descriptive, predictive, prescriptive, automated analytics*
- Riflessione sulla scientificità delle conclusioni: replicabilità e stabilità

Ruolo delle società scientifiche - Statement della SIS su Statistica, Scienza dei dati e Big data

http://sis-statistica.it/upload/contenuti/2018/files/SIS_-_Statement_su_Data_Science_%28con_firmatari%29%281%29.pdf

DATA (PLURALE DI *DATUM*)

- **Banche dati a disposizione delle pubbliche amministrazioni in Italia**. Ogni interazione del cittadino o dell'impresa con la pubblica amministrazione (PA) genera dati amministrativi (generalmente strutturati)
- **Open data**. I dati raccolti quotidianamente dalla PA liberi da copyright, brevetti o altre forme di controllo, messi liberamente a disposizione dei cittadini
- **Big Data**. Solo quei dati che soddisfano almeno uno dei seguenti criteri: **volume** elevato (più di 50 Terabyte o crescita annua maggiore del 50%), **velocità** (raccolta e analisi dei dati real-time) e **varietà** (fonti eterogenei e dati non solo strutturati).

DATI: CLASSIFICHIAMO I BIG DATA

UNECE (Commissione Economica per l'EU delle Nazioni Unite)tassonomia:

- **Human-generated data** (ad es.: dati da Social Media, Blog, SMS, e-mail, User generated contents e maps, ecc.);
- **Process-mediated data** (quali Sistemi transazionali, commerciali e bancari tradizionali, e-commerce, carte di credito, dati prodotti da Enti Pubblici e/o privati;
- **Machine-generated data** (tipicamente ciò che va sotto il nome di *Internet of Things* , come sensori fissi (home-automation, sensori ambientali/meteorologici, sistemi per il controllo del traffico, ecc.) e mobili (dispositivi mobili, sensori su automezzi, immagini satellitari).

DATI: Caratterizzazione dei Big Data in ambito statistico (con riferimento ad aspetti di qualità)

qualità “specifica per le fonti”

- fonte Big come i sensori avrà una qualità specifica che dipende dal fatto che i dati da sensori sono spesso mancanti, soggetti a rumore o ad effetti di calibrazione degli strumenti di misura.

qualità delle fonti Big “per dominio”

- Se il dominio di interesse è quello della statistica ufficiale, alcune dimensioni di qualità particolarmente rilevanti sono la “rappresentatività” di una fonte, fondamentale per poter produrre stime affidabili, l’accuratezza in termini di qualità intrinseca dei valori acquisiti e l’affidabilità della fonte.
- Integrazione tra archivi amministrativi, qualità statistica degli archivi amministrativi

DATI: STATISTICA UFFICIALE

Trusted Smart Statistics is the European Statistical System strategy to face with Big Data
(*Bucharest Memorandum*, adopted in October 2018 DGINS).

Eurostat: the NSIs will have to go through a **'playground' phase**, which is necessary to understand the potential and limits of Big Data sources, as well as the methods necessary to treat them, **to a phase of mature use** of these sources, called Trusted Smart Statistics.

ESEMPI

Smart Data Platforms, integrazioni delle informazioni provenienti dall'Internet of Things (sensori, oggetti interconnessi), dall'Internet of People (social network) e gli open data, ovvero quei dati resi pubblici dalle stesse PA.

Mettere insieme i dati, analizzarli, fornire ai decisori pubblici delle visualizzazioni dei risultati semplici e immediate, rendendo data-driven, più efficaci, tempestive ed efficienti le azioni delle pubbliche amministrazioni

Esempi: conoscere in tempo reale la situazione del traffico cittadino, controllare dati ambientali o di consumo energetico….

coordinamento di attori: PA, Agenda digitale, Istat e Sistan

DATI: TRUSTED SMART STATISTICS

Smart Data Sources: include Big Data sources, mainly belonging to the category of “Internet of Things/machine generated data”, in charge of data providers out of NSSs (business\partnership models are needed).

Smart Methods Design and Execution

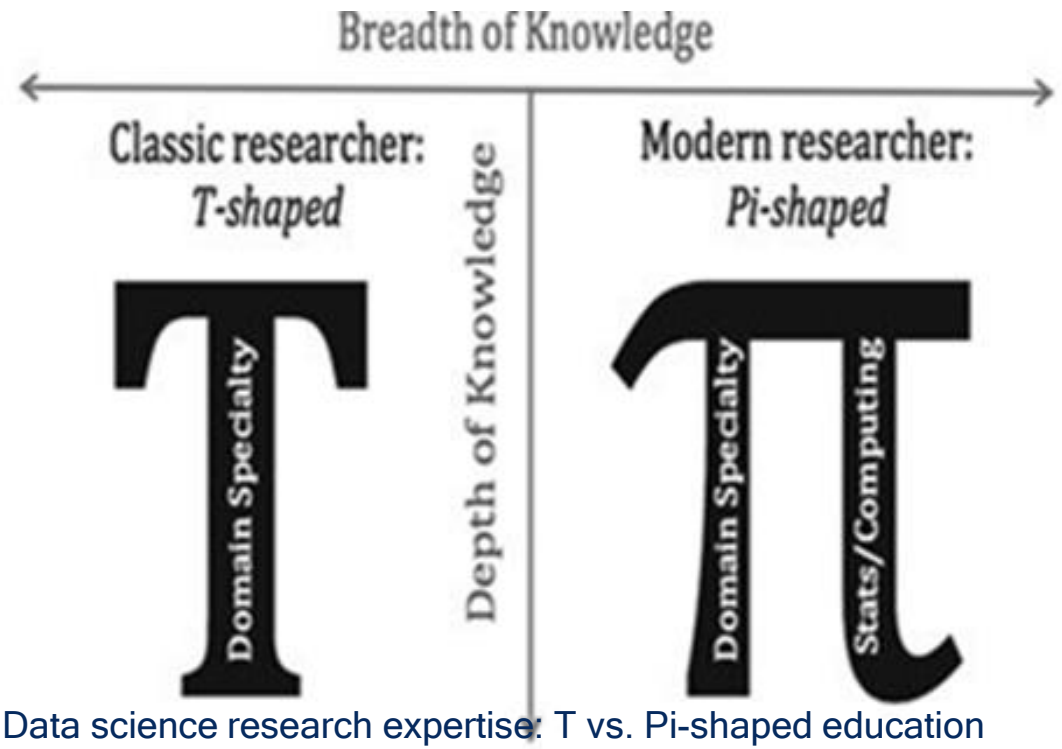
The design of processing will result from both the traditional statistical processing framework (design- and model-based frameworks) and the algorithmic-based processing framework (mainly machine-learning framework). A specific feature for smart statistical processing is the geographically distributed computation.

The design of methods/algorithms is the responsibility of the NSIs.

The execution of methods/algorithms can take place either at the data provider side or at the premises of the NSIs.

Output: Trust. Statistical products need to have *specific trust guarantees* established in the design phase.

CAPACITA'E ABILITA' NECESSARIE



Data science research expertise: T vs. Pi-shaped education (DrewConway, 2015)

La vita professionale richiede in qualsiasi dominio la capacità di trarre informazioni da grandi set di dati, utilizzando un insieme più ampio di competenze.

Teams interdisciplinari

Questo è un importante riconoscimento della Statistica come utile in tutti i campi.

CAPACITA'E ABILITA' NECESSARIE

Come sviluppare le nuove competenze, abilità e professionalità? Il sistema universitario sta sperimentando nuovi percorsi in questa direzione. C'è bisogno di formare statistici con nuove abilità o costruire un nuovo profilo professionale? È possibile per un singolo ricercatore mantenere una competenza sufficiente sia in statistica che in informatica per modellare da soli problemi complessi?

A mio avviso, accanto alla figura dei data scientist, l'elemento decisivo è saper dialogare e interagire tra comunità diverse, tra esperti con competenze diverse. Non solo nella progettazione di nuovi corsi di formazione, ma anche nella ricerca di nuovi metodi di analisi. Pertanto, non solo l'integrazione dei dati, ma anche l'integrazione tra le competenze e le diverse comunità professionali.

Ricerca sui Metodi di Rilevazione ed analisi

Situazione:

- l'alto grado di eterogeneità e dati non Indipendenti Identicamente Distribuiti
- dati altamente non strutturati (inclusi immagini, testo, suono, altre nuove forme) da comprendere insieme,
- dati osservati attraverso diversi tipi di dispositivi di misurazione,
- non necessariamente derivanti da esperimenti progettati, ma corrispondenti a una miscela di molte popolazioni eterogenee,
- con osservazioni mancanti e distorte (facendo deduzioni valide da campionamenti non probabilistici).

Ricerca sui Metodi di Rilevazione ed analisi

Replicabilità e stabilità delle conclusioni e dei risultati nell'"ecosistema" dei dati.

Le complessità dei dati richiedono una sostanziale pre-elaborazione (Reid, 2018) e nuovi approcci a concetti teorici e sviluppi metodologici.

Integrazione dei dati (linkage, statistical matching, …) e riduzione dei dati (PCA, samples)

Inferenze valide con modelli molto complessi (inferenza ad alta dimensione) su reti, forme, immagini, dati spaziali che si evolvono nel tempo, anche con variabile di interesse multidimensionale.

Divario tra modelli statistici e inferenza algoritmica e IA ("opacità" , equità delle decisioni algoritmiche (in particolare con algoritmi di deep learning), (Shah, 2017; O'Neil, 2016).

Ricerca sui Metodi di Rilevazione ed analisi

Statistica ufficiale:

- nuovo ruolo per le indagini campionarie nel nuovo ecosistema multi-sorgente (Alleva, 2017);
- l'incertezza delle informazioni prodotte sfruttando il nuovo ecosistema di dati (ad esempio il sistema integrato italiano di registri statistici, ISSR)
- Big Data per produrre dati ufficiali: un percorso che parte dalla sperimentazione del loro trattamento e valutazione rispetto alla qualità statistica (statistiche sperimentali)
- Valorizzazione degli ambiti tematici nella produzione dati (sociale, economico) e nella loro diffusione (disintermediazione della comunicazione)

Ricerca sui Metodi di Rilevazione ed analisi

Statistica ufficiale: una scelta difficile

"rendere l'uso dell'ISSR limitato e consentire la diffusione solo degli output pianificati con un livello di accuratezza certificato"

"rendere il sistema più flessibile, consentendo agli utenti di produrre le proprie statistiche dall'ISSR"

Gestione dell'incertezza? una questione strategica per gli INS, per la fiducia e la trasparenza.

GOVERNANCE DEI DATI

- Il punto non riguarda solo i progressi nei metodi, nelle tecnologie e nelle competenze, ma anche nella governance dei dati, in termini di politica, apertura, privacy e fiducia: la quarta sfida per il futuro delle statistiche è la governance dei dati, la loro produzione, elaborazione e comunicazione .

- Il tema trasversale più importante relativo ai big data è la privacy, che copre tutti gli aspetti del ciclo di vita dei dati. Le nuove normative mirano non solo a proteggere la nostra privacy e il modo in cui archiviamo le informazioni su noi stessi, ma anche a includere tutto il processo e anche ciò che ci è consentito fare con tali dati (EU-GDPR, 2016).

CONCLUSIONI

Integrazione e partenariato sono i due punti chiave:

- integrazione delle fonti di dati,
- integrazione tra modellazione stocastica e procedure algoritmiche,
- combinazione di approcci frequentista e bayesiano,
- integrazione tra competenze e profili professionali per l'istruzione e formazione,
- partenariato tra Statistica Ufficiale e ricercatori di Academia e altri soggetti pubblici e istituzione e società private,

CONCLUSIONI

- partnership tra INS e titolari privati di dati,
- partenariato tra INs e autorità sulla privacy.
- mostrare successi e innovazioni e anche spazio per la pubblicazione di risultati negativi, Big data for Small Areas?;
- denunciare un uso errato da parte dei titolari dei dati o dei media; questo significa essere in grado di comunicare e utilizzare strumenti di comunicazione;
- promuovere queste fonti come beni pubblici, in primo luogo per gli INS e gli NSS, ma anche per i ricercatori….non è facile accedere ai Big data!

CITAZIONI

Alleva G. (2019) The future of Statistics: challenges for understanding new phenomena in Keynote speech, ITACOSM 2019, 5-7 June, Florence,

Benzécri J.P. (1979) L'analyse des données, Tome I Taxinomie, Tome II Correspondances, Dunod.

Breiman L. (1996) Heuristics of instability and stabilization in model selection. Ann. Statist. 24, 2350-2383.
Breiman L. (2001) Statistical Modeling: The Two Cultures, *Statistical Science,* 2001, Vol. 16, No. 3, 199-231

Secchi P. (2018) On the role of statistics in the era of big data: a call for a debate. Statist. Probab. Lett. 136, 10–14.

Tukey J.W. (1962) The future of data analysis, Ann. Statist. 33, 1-67
Tukey J.W. (1977) Exploratory data analysis. Reading, PA: Addison-Wesley.

Eurostat (2013) *Scheveningen Memorandum* on *Big Data for Official Statistics*, DGINS 2013.
Friedman J.H. (2001) The Role of Statistics in the Data Revolution?, International Stat Rev., Vol. 69, No. 1, pp. 5-10

UNECE (10 March 2013) CONFERENCE OF EUROPEAN STATISTICIANS **WHAT DOES "BIG DATA" MEAN FOR OFFICIAL STATISTICS?**

https://www.agid.gov.it/it/agenzia/strategia-quadro-normativo