Big data e statistica: il lago e l'iceberg

Giovanni A. Barbieri, Istat

StatCities Trento

La vendemmia statistica

Linee operative e prospettive di riforma del sistema statistico nazionale a livello locale

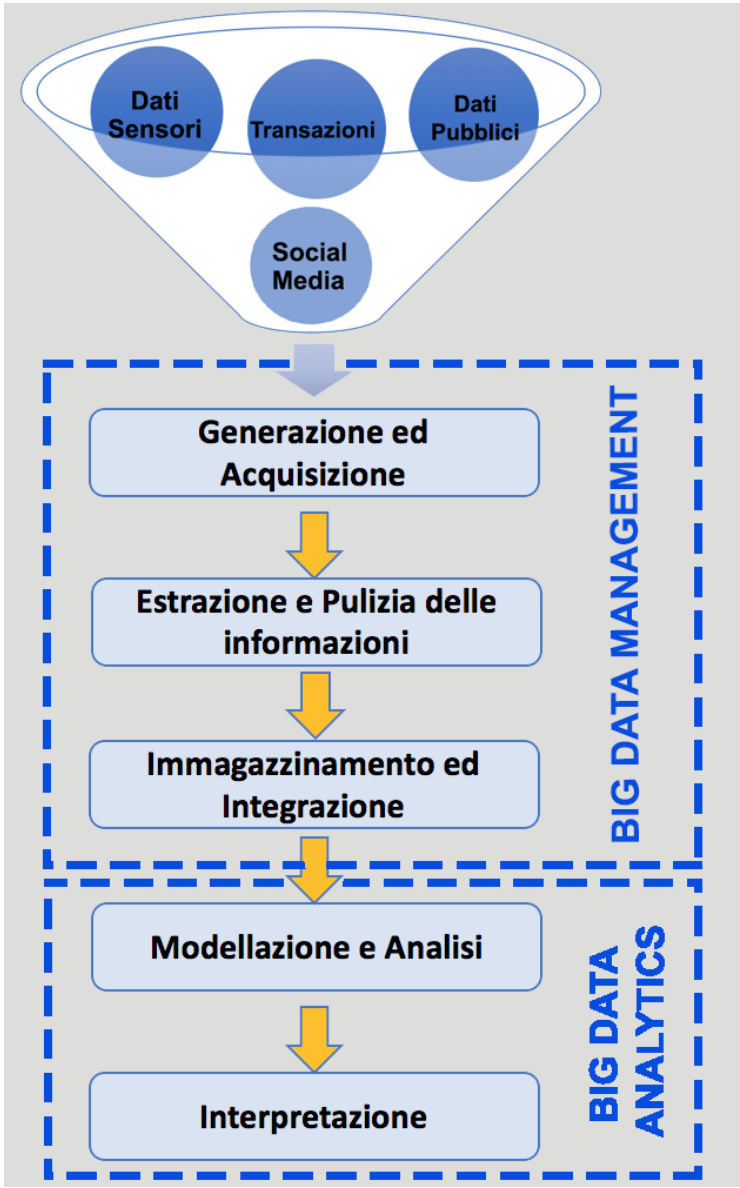
Trento, 15 settembre 2017

Indice

- Che cosa sono i big data?
- I big data sono qui e non se ne andranno
- Quantità e qualità
- Esagerazioni e sciocchezze
- Che cos'è il lago?
- Che si può fare?
- Compagni di strada?

Che cosa sono i big data?

- Sono grossi
  - La tecnologia c'entra
    - Richiedono tecnologie e metodi specifici
- Sono dati
  - Il portato della società dell'informazione
    - Datafication
    - Entropia e informazione



I big data sono qui e non se ne andranno

- Diluvio dei dati e cambiamento climatico
  - Non più se o quando, ma come
  - Dalla mitigazione all'adattamento
- Complesso di superiorità
  - Solo le nostre statistiche sono belle e ben fatte
  - Atteggiamento difensivo e irrazionale…

Quantità e qualità

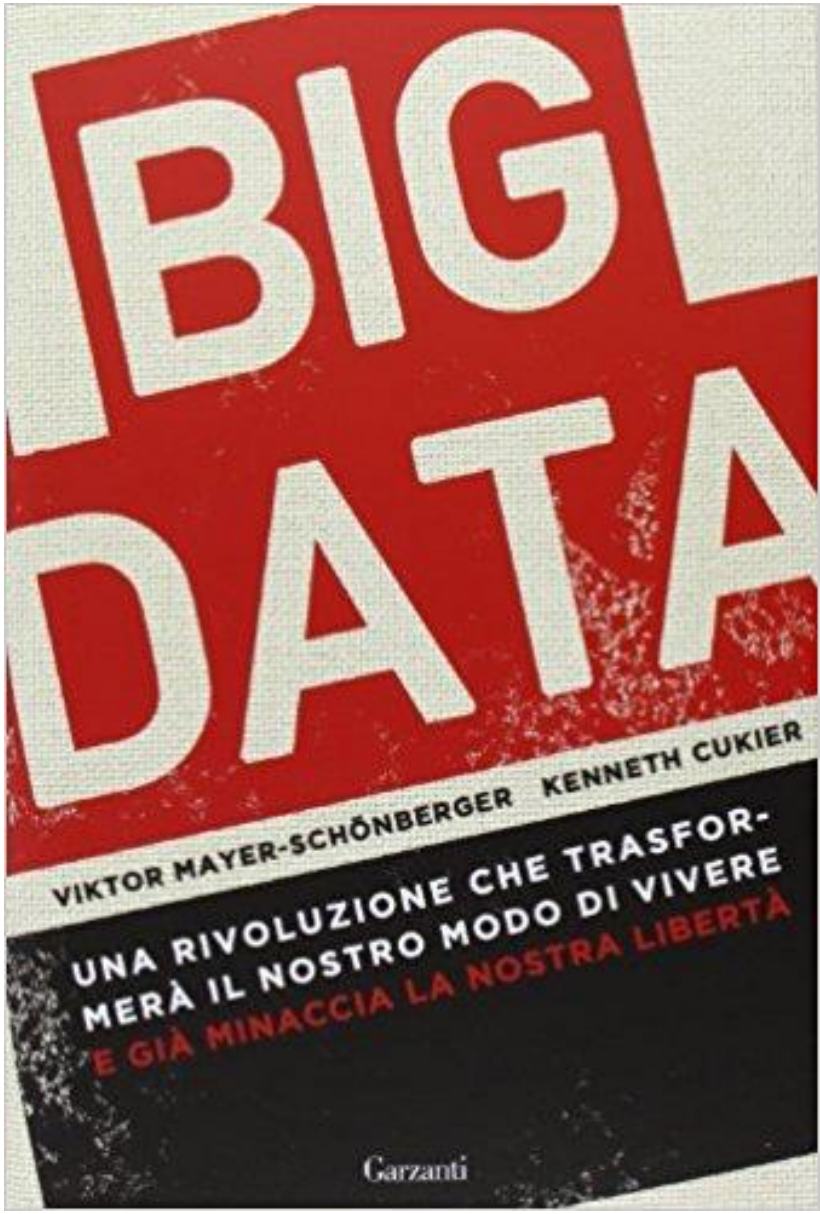
- La dimensione è sufficiente a ridefinire i termini del problema?
  - Il «salto di qualità» è frutto della crescita dimensionale?
- Sì, sotto il profilo pratico (la tecnologia c'entra)
- No, sotto quello concettuale
  - Se definiamo «dati amministrativi» quelli generati nell'ambito del funzionamento di un processo (e non soltanto quelli generati da una PA), i *big data* ricadono in larghissima misura in questa definizione

Esagerazioni e sciocchezze

- La fine della teoria (Chris Anderson su Wired, 2008)
  - Una fondamentale incomprensione del ruolo dei modelli in statistica, George Box 1976-1978

Esagerazioni e sciocchezze

- N = all
  - Non servono più i campioni
  - E non servono più neppure gli esperti…



Che cos'è il lago?

- *Processed data storage* contro *raw data storage*
  - Non è acqua in bottiglia
  - Superamento dei silos
- Importanza dei metadati a corredo

DATA WAREHOUSE	vs.	DATA LAKE
structured, processed	DATA	structured / semi-structured / unstructured, raw
schema-on-write	PROCESSING	schema-on-read
expensive for large data volumes	STORAGE	designed for low-cost storage
less agile, fixed configuration	AGILITY	highly agile, configure and reconfigure as needed
mature	SECURITY	maturing
business professionals	USERS	data scientists et. al.

©KDnuggets

Che si può fare?

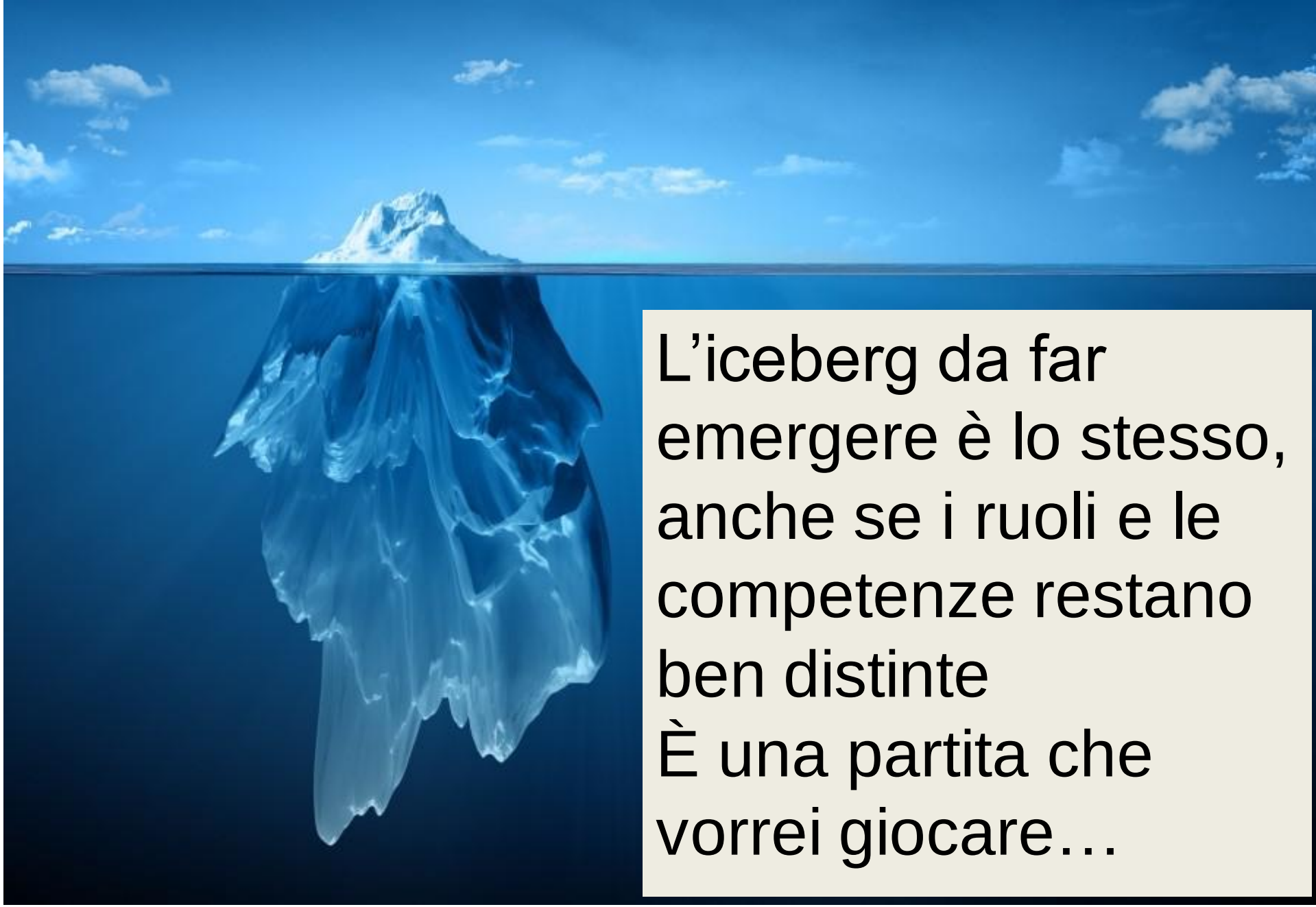
- Modelli di interconnessione di fonti di dati diverse
- Analisi dei dati
- Sviluppo di modelli di *machine learning*
- Produzione di *data applications*

▼

Veicolare in modo efficace
(e rendere, quindi, più accessibile)
l'informazione presente nei dati

Sounds familiar?

- Differenze e analogie
  - Non si sta parlando di informazione statistica, ma di valorizzazione di un'informazione che nasce e resta «amministrativa»
  - Ma ci sono forti punti di contatto sia nei mezzi (integrazione delle fonti, analisi dei dati, modelli, *data applications*) sia nei fini (veicolare l'informazione)



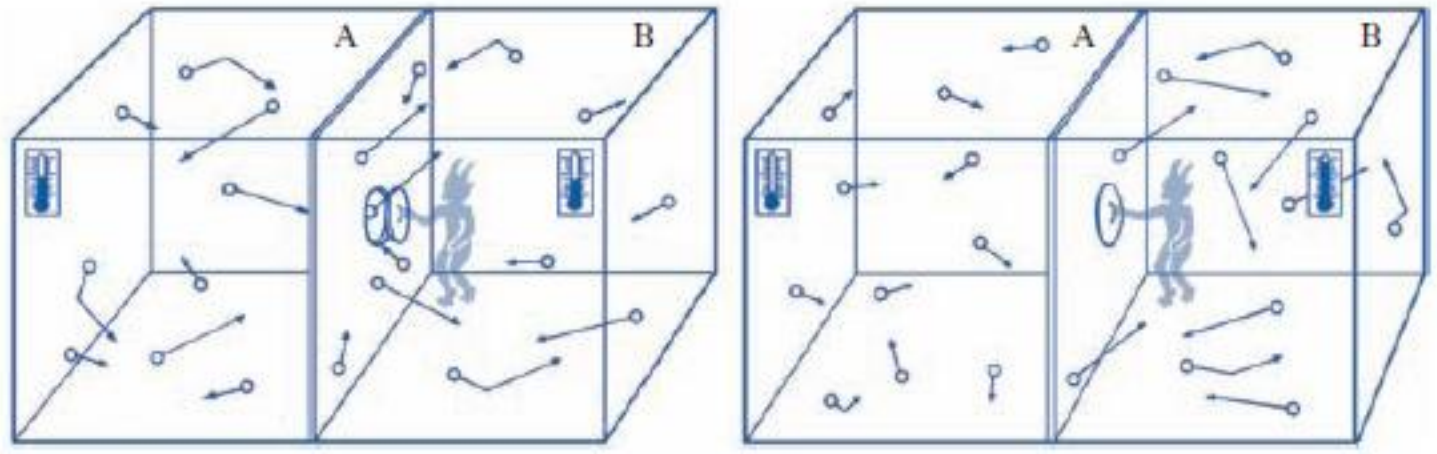
L’iceberg da far emergere è lo stesso, anche se i ruoli e le competenze restano ben distinte
È una partita che vorrei giocare…

Datafication

- Taking information about all things under the sun – including ones we never used to think of as information at all, such as a person's location, the vibrations of an engine, or the stress on a bridge – and transforming it into a data format to make it quantified
- New ways to use the information
- Unlock the implicit, latent value of the information
- [Mayer-Schonberger & Cukier. Big Data: The Essential Guide to Work, Life and Learning in the Age of Insight]

Entropia 1

- Boltzmann
  - Meccanica statistica: entropia correlata a ordine (differenti probabilità degli stati del sistema)
  - Entropia espressione del «grado di disordine» di un sistema



Entropia 2

- Shannon
  - Livello di imprevedibilità di una fonte d'informazione
  - Quanta «informazione» in un messaggio (segnale/rumore)?
  - Grado di ignoranza equiparato a disordine: messaggio come quantità di informazione che fa passare il ricevente da incertezza a ordine (minor incertezza)

Atteggiamenti difensivi e irrazionali

- 1° gennaio 2005, fine dell'accordo multifibre:
  - «… i nostri scaffali saranno invasi da merci cinesi…»
    - Vero non solo per il tessile-abbigliamento
    - Ma la capacità di esportazione italiana è aumentata e abbiamo tenuto o guadagnato quote di mercato in un commercio mondiale in espansione
- Padova 2007:
  - Ci interrogavamo sulla fine del monopolio naturale della statistica pubblica
  - Le risposte che siamo stati capaci di dare, dopo 10 anni, sono sotto i nostri occhi

Un vecchio equivoco

- Hegel, Scienza della logica #776:
  - «i mutamenti dell'essere, in generale, non sono soltanto il passaggio di una grandezza in un'altra grandezza, ma sono il passaggio dal qualitativo al quantitativo e viceversa»
  - Sta parlando in realtà dei «cambiamenti di fase»
  - Non può essere interpretata come una legge generale

I modelli in statistica

- George Box 1976:
  - «Since all models are wrong the scientist cannot obtain a ‘correct’ one by excessive elaboration»
  - Invocava il rasoio di Occam
- 1978:
  - *All models are wrong but some are useful* è il titolo di un paragrafo
  - «…parsimonious models often do provide remarkably useful approximations»