Gestione e integrazione di dati (grandi e piccoli): la sfida è nel significato

Paolo Atzeni

Dipartimento di Ingegneria

Università Roma Tre

Gestione e integrazione di dati (grandi e piccoli): la sfida è nel significato

Dati e Big Data

- Perché ci interessano?

The New York Times — March 18th 1923

WHY CLIMB MOUNT EVEREST?
"BECAUSE IT'S THERE"
SAID MALLORY

George H.L. Mallory
(Mobberley, 1886 – Everest, 1924)

Big Data

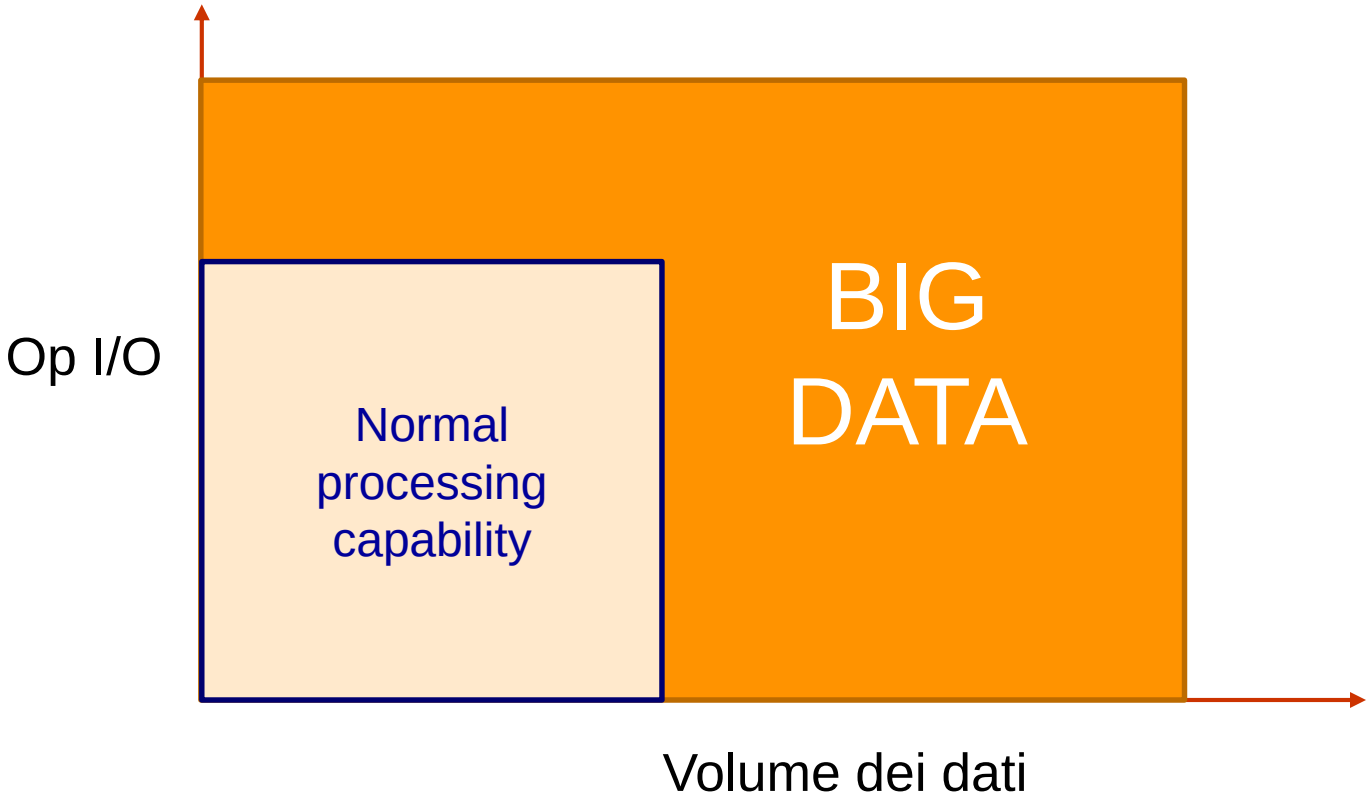
- Varie definizioni

"Big data exceeds the reach of commonly used hardware environments and software tools to capture, manage, and process it with in a tolerable elapsed time for its user population." - Teradata Magazine article, 2011

"Big data refers to data sets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze." - The McKinsey Global Institute, 2012

"Big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools." - Wikipedia, 2014

Big data

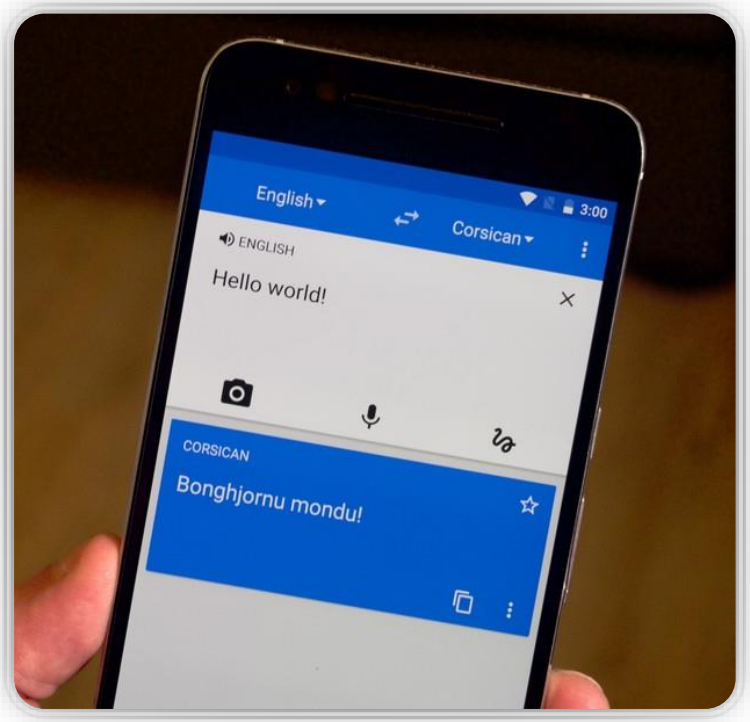

Op I/O
Normal processing capability
BIG DATA
Volume dei dati

Tradizionali operazioni su basi di dati

- Semplici o anche complesse, ma di solito ben definite:
  - Il saldo di un conto corrente è dato dal saldo inziale più la somma delle operazioni a credito meno le operazioni a debito
  - viene confermata la prenotazione su un treno solo se c'è un posto disponibile (che non viene poi assegnato ad altri)

Operazioni sui Big Data

- Con tanti dati, si può provare qualcosa di più ambizioso
- Un esempio: Google Translate
  - Raccolta traduzioni (note) di frasi brevi.
  - Confronto tra frasi del testo da tradurre con frasi raccolte.
  - Aggiornamento continuo dell'archivio delle traduzioni.
- Funziona
  - Non perfettamente, ma in modo molto soddisfacente
  - Migliora nel tempo



Operazioni sui Big Data

- Tante possibilità
  - operazioni tradizionali su grandi volumi (di solito non si tratta davvero di big data, ma solo di tanti dati)
  - modelli matematici complessi:
    - pseudo-deterministici (o comunque algoritmici)
    - euristici (ad esempio, le previsioni del tempo)
  - machine learning

Approssimare serve

- "All models are wrong, but some are useful" (G. Box e altri)
- La mappa dell'impero in scala 1:1 (Borges, Eco)

La mappa dell'impero (Borges)

- *“…In quell'Impero, l'Arte della Cartografia giunse a una tal Perfezione che la Mappa di una sola Provincia occupava tutta una Città, e la Mappa dell'Impero tutta una Provincia. Col tempo, queste Mappe smisurate non bastarono più. I Collegi dei Cartografi fecero una Mappa dell'Impero che aveva l'Immensità dell'Impero e coincideva perfettamente con esso. Ma le Generazioni Seguenti, meno portate allo Studio della Cartografia, pensarono che questa Mappa enorme era inutile e non senza Empietà la abbandonarono alle Inclemenze del Sole e degli Inverni. Nei Deserti dell'Ovest sopravvivono lacerate Rovine della Mappa, abitate da Animali e Mendichi; in tutto il Paese non c'è altra Reliquia delle Discipline Geografiche.”* *(Da* Viajes de Varones Prudentes *di Suàrez Miranda, libro IV, cap. XIV, Lérida, 1658. Citato da Jorge Luis Borges,* Storia universale dell'infamia, “Etc.”.)

I problemi principali

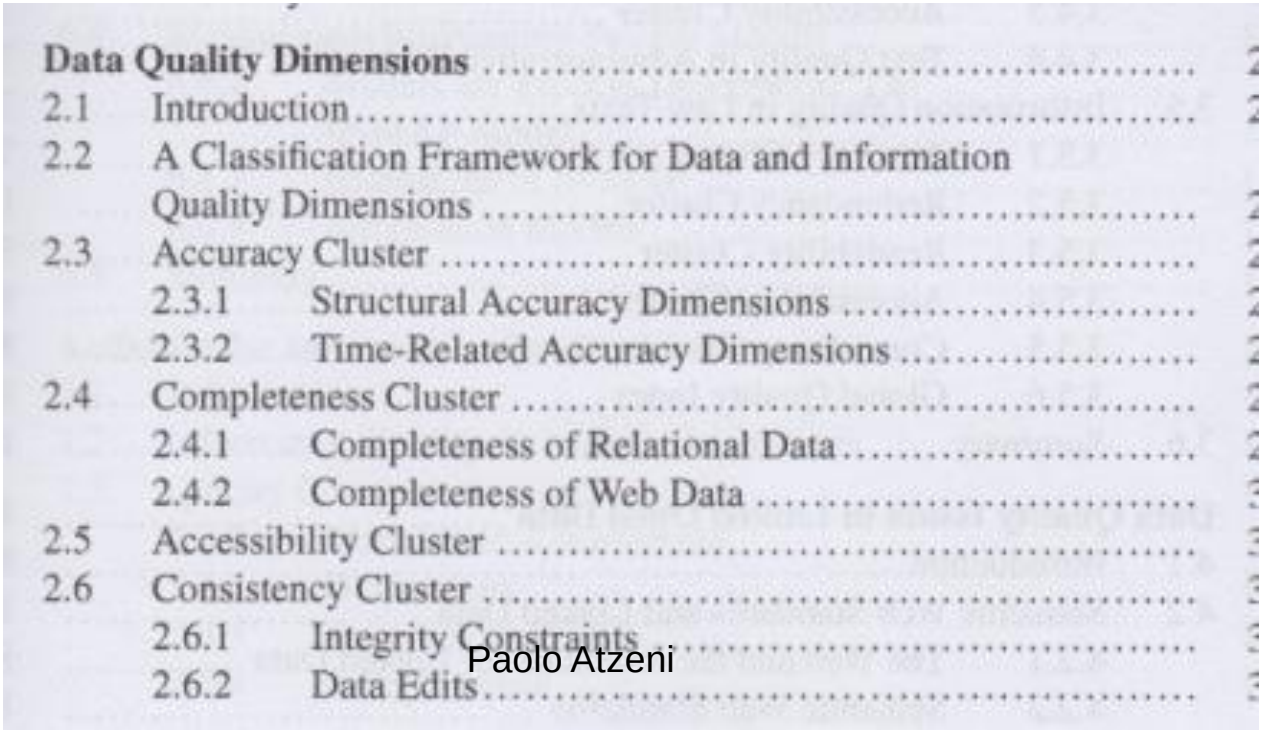
- Sui dati
- Sulle tecniche

Problemi sui dati

- Qualità dei dati
- Interpretazione

Una grande classe di problemi sui dati

- Dati "sbagliati" (imprecisi, incompleti, …)
  - … la qualità dei dati è un problema molto articolato
    - si cita spesso il "fat-finger error" …
  - un libro recente di 480 pagine (C. Batini, M. Scannapieco. Data and Information Quality. Springer-Verlag, 2016)



Data Quality Dimensions 2
2.1 Introduction 2
2.2 A Classification Framework for Data and Information Quality Dimensions 2
2.3 Accuracy Cluster 2
2.3.1 Structural Accuracy Dimensions 2
2.3.2 Time-Related Accuracy Dimensions 2
2.4 Completeness Cluster 2
2.4.1 Completeness of Relational Data 2
2.4.2 Completeness of Web Data 3
2.5 Accessibility Cluster 3
2.6 Consistency Cluster 3
2.6.1 Integrity Constraints 3
2.6.2 Data Edits 3

Interpretazione dei dati

Interpretazione dei dati

7–17

White or black figures without brackets indicate weekdays, except weekdays before Sundays or public holidays.

(11-14)

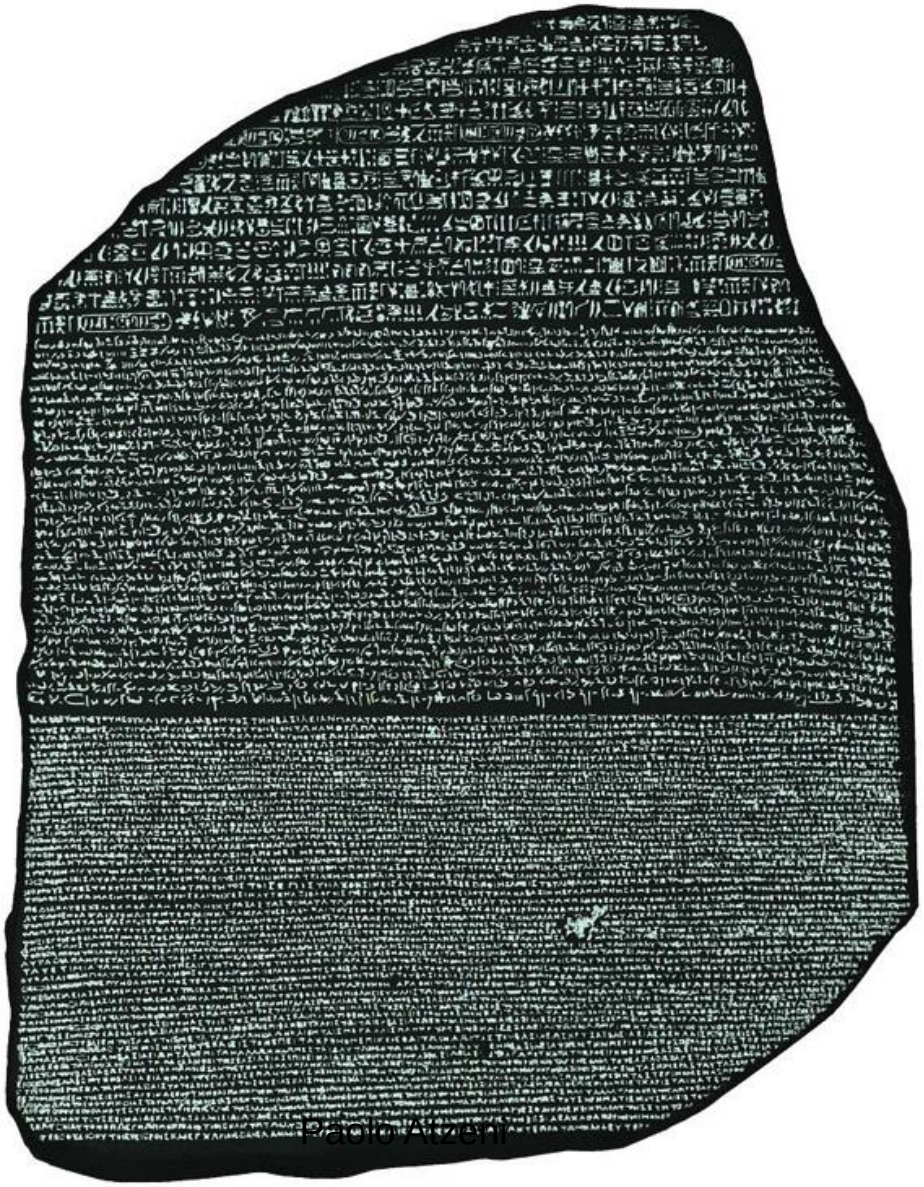
White or black figures in brackets indicate weekdays, before Sundays or public holidays.

11-14

Red figures indicate Sundays or public holidays.

Torsdag: giovedì

Un famoso esempio per l'interpretazione



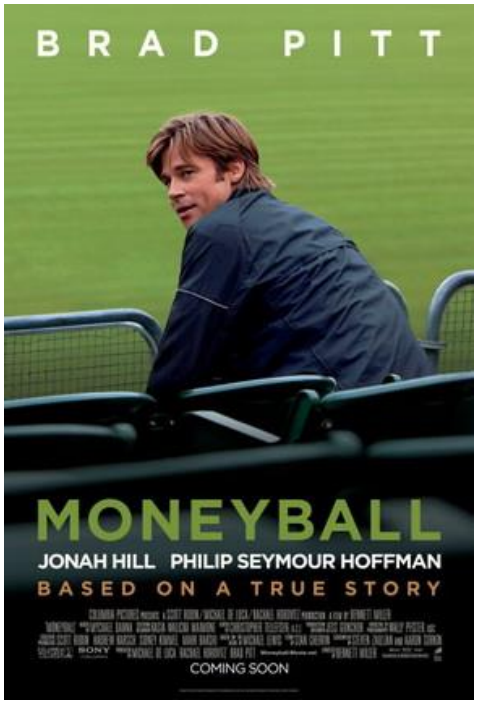
Un altro aspetto dell'interpretazione dei dati

- Quanti sono gli studenti iscritti al Dipartimento di Ingegneria?
  - Dipende da quale criterio usiamo
    - Iscritti al 5 novembre
    - Iscritti ad oggi (in uno specifico giorno)
    - Iscritti in corso
    - Iscritti in corso appartenenti ad una coorte "pura"
    - Iscritti "regolari" (cioè da un numero di anni non superiore alla durata legale del corso)
    - …

Problemi sulle tecniche

Dati e indicatori

- I dati si utilizzano per costruire modelli:
  - un modello descrive un fenomeno e utilizziamo i dati come input per il modello
- Esempio tipico: le previsioni del tempo
- La forma più semplice di modello:
  - gli indicatori
- Un caso interessante:
  - Moneyball, libro e film



Moneyball

- Le squadre di baseball, fino al 2000 circa, utilizzavano, per valutare i giocatori, alcune statistiche
  - percentuale di battute valide (batting average)
  - numero di punti segnati (runs batted in, RBI)
  - fuori campo (home run)
- I grandi giocatori del passato (Babe Ruth, Ted Williams, Lou Gehrig, Joe Di Maggio) sono famosi per i loro record di battute valide, home run, RBI
- I giocatori riuscivano ad avere ingaggi strettamente correlati a questi "indicatori" (che si riteneva descrivessero l'utilità del giocatore, cioè il contributo alla vittoria delle partite)

Moneyball, l'idea chiave

- Il protagonista è un allenatore che
  - intuisce (e verifica analiticamente) che altri indicatori (slugging percentage, on-base percentage) sono più strettamente correlati al risultato
  - ingaggia giocatori che hanno ottimi valori su questi indicatori e valori meno buoni sugli indicatori tradizionali e quindi chiedono ingaggi più bassi
  - ottiene risultati straordinari spendendo poco
- Morale:
  - per trarre valore dai dati, si deve sapere quali utilizzare e che cosa significano

I dati si usano per costruire indicatori

- Gli indicatori debbono "indicare" qualcosa, cioè rappresentare il fenomeno:
  - gli indicatori di Moneyball misurano l'aspettativa di vittoria
- Attenzione:
  - Non è ovvio che esistano sempre indicatori idonei
- Ad esempio, gli indicatori relativi ai prodotti della ricerca …
  - In un incontro di settore, criticai l'uso estremo della bibliometria nella VQR e un collega autorevole (?) mi disse: meglio questi indicatori che niente …
  - mi spiace, ma dissento: gli indicatori possono essere utilizzati se sono realmente correlati al fenomeno

Tecniche discutibili

- Washington, D.C., 2007:
  - Valutazione delle scuole, anzi degli insegnanti, sulla base di test standardizzati, utilizzati per "misurare le competenze"
  - "Misurando" gli studenti alla fine di ogni anno, si riteneva di poter misurare il miglioramento (più o meno significativo) e quindi la "qualità" del docente

E il modello o algoritmo?

- Spesso non è noto!
- Una docente licenziata a Washington chiese informazioni
- … e le fu detto che l'algoritmo era stato predisposto da una società di consulenza, molto specializzata
- … e che l'algoritmo era molto sofisticato e teneva conto di tanti parametri
- … e che quindi era certamente corretto!
- E comunque non poteva essere spiegato in dettaglio perché di proprietà della società di consulenza!

Due forti dubbi

- È appropriato il modello?
  - Cioè, descrive correttamente il fenomeno?
- È corretto l'algoritmo?
  - Una volta descritto compiutamente il problema, l'algoritmo trova davvero il risultato (definito secondo le regole del problema)?

Dati e tecniche sono correlati (in modo potenzialmente perverso)

- Washington, D.C., 2007:
  - …
  - Si è scoperto che molti docenti "aiutavano" gli studenti nei test, per migliorare la propria "prestazione", penalizzando, e molto, i docenti cui gli stessi studenti venivano affidati l'anno successivo
- Qualunque uso semplicistico dei dati ingenera circuiti viziosi
  - … pubblicazioni e citazioni nell'accademia italiana negli ultimi anni …

Indicatori sbagliati e reazioni negative

- Autorevoli fisici, studiosi delle onde gravitazionali (fra i quali due premi Nobel, annunciati qualche mese dopo),
  - hanno scritto alla Ministra segnalando che i docenti italiani della loro area non hanno alcuna possibilità di carriera perché surclassati, negli indicatori, da studiosi di altre aree dello stesso settore concorsuale che lavorano in gruppi più grandi, pubblicano di più e sono citati di più …
- Il risultato netto, segnalano i firmatari, è che gli studiosi italiani saranno spinti ad abbandonare queste tematiche
- … oppure ad abbandonare l'Italia

Gestione e integrazione di dati (grandi e piccoli): la sfida è nel significato

"A ten-year goal for database research"

- The “Asilomar report”
  (Bernstein et al. Sigmod Record 1999 www.acm.org/sigmod):
  - ***The information utility:***
    make it easy for everyone to store, organize, access,
    and analyze the majority of human information online

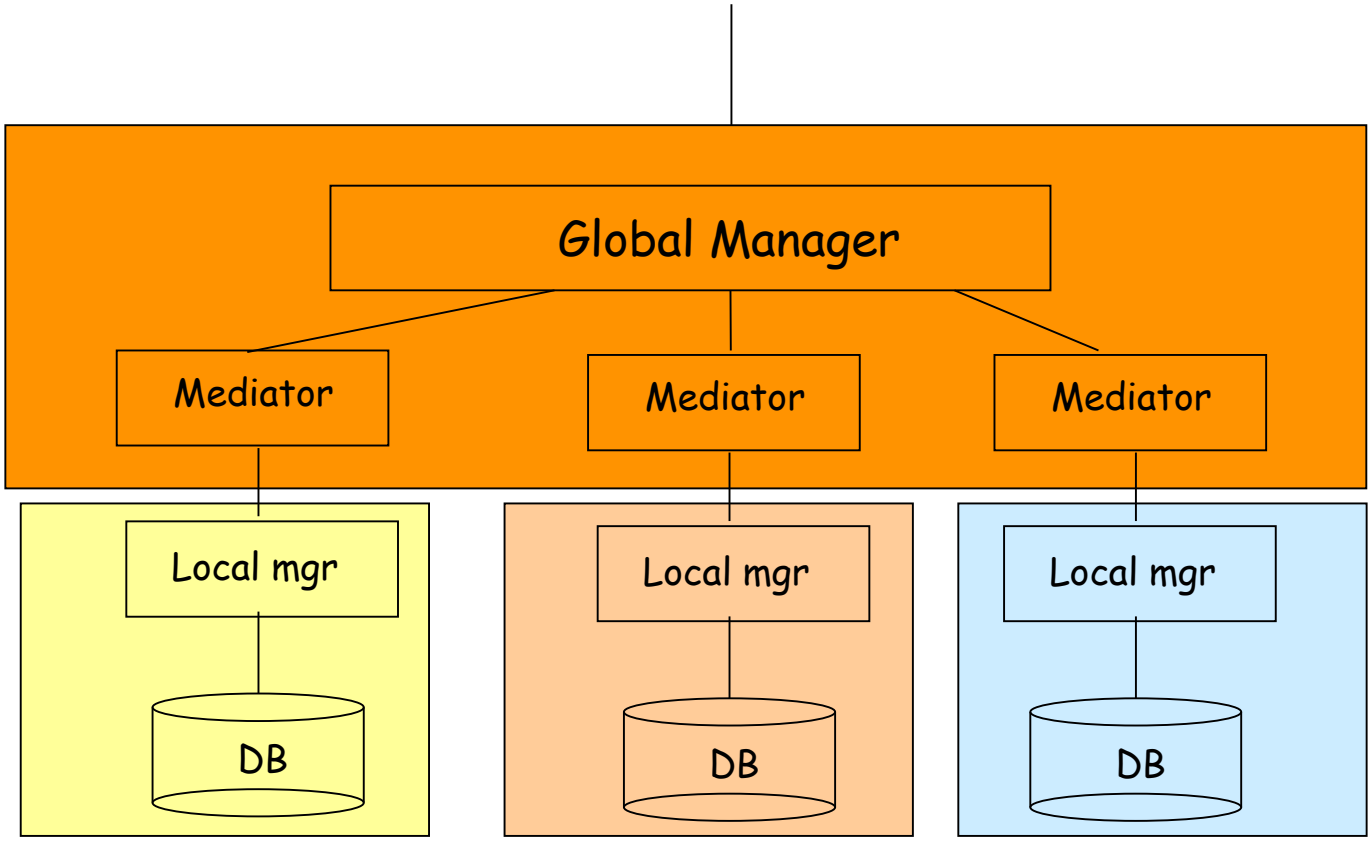
- Molti risultati interessanti sono stati ottenuti, ma…
- …integrazione, traduzione, interscambio di dati restano difficili…
- **… sono passati 20 anni, il 2019 è arrivato … siamo in ritardo**
- **… e abbiamo anche i Big data**

Big data integration

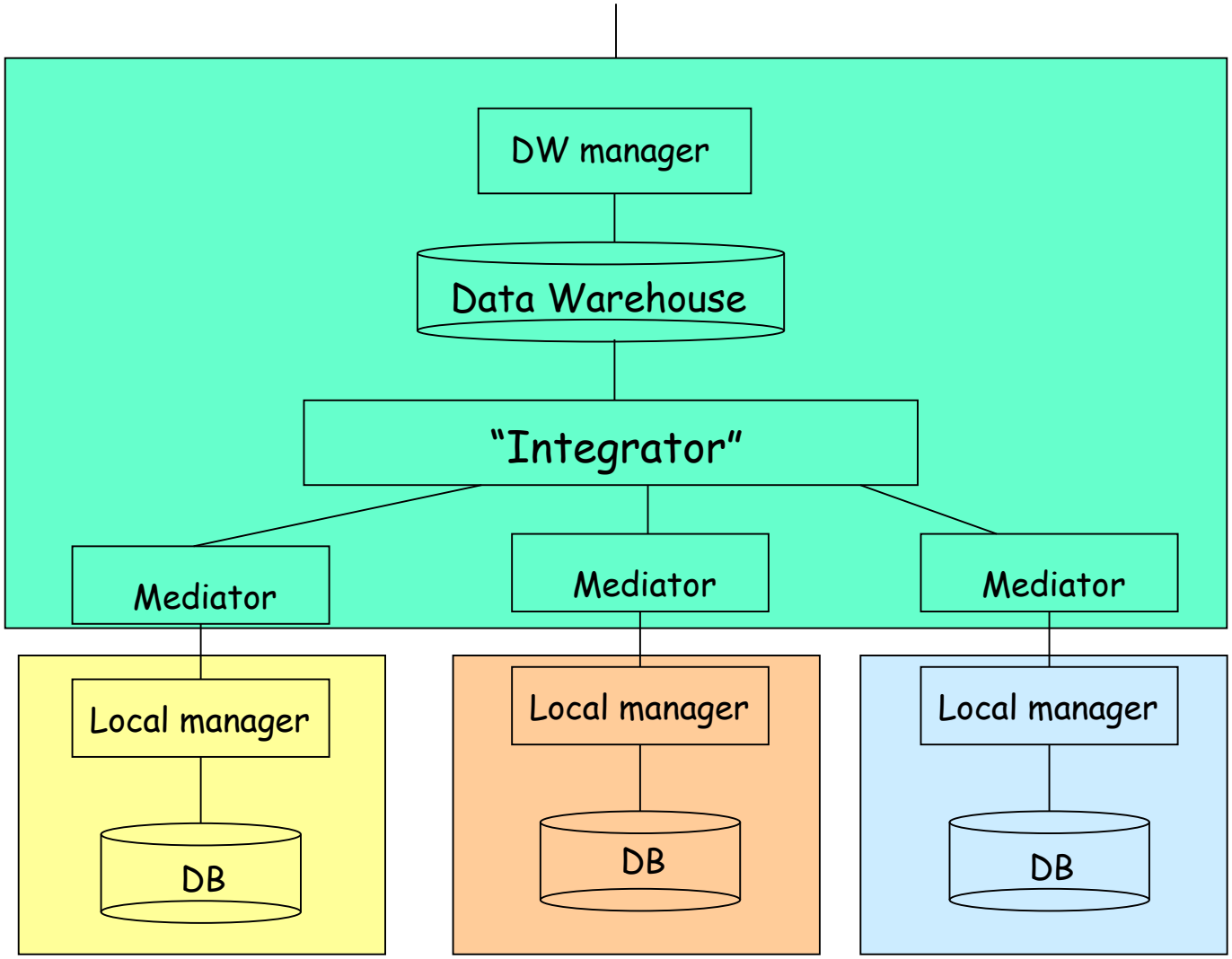
- Big data integration = data integration + big data

- Integrazione: accesso a dati di molte sorgenti
  - Due approcci
    - Virtuale ("multidatabase")
    - Data Warehousing

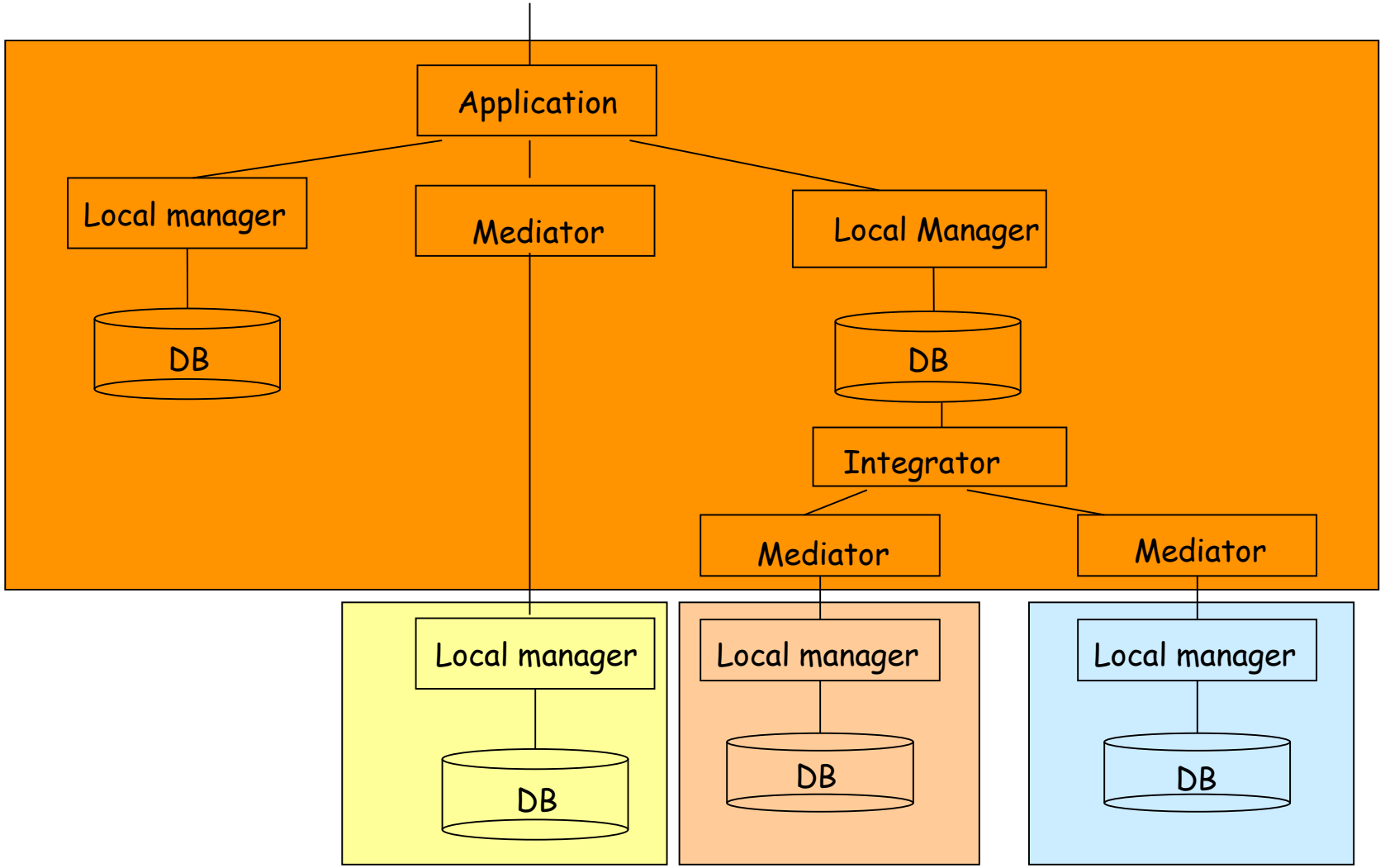
Multidatabase



Data Warehousing System



Soluzioni intermedie



Tipico approccio all'integrazione

- Tre fasi (Dong e Srivastava)
  - Schema alignment: individuare corrispondenze fra gli schemi
    - mediated schema
    - attribute matching
    - schema mapping
  - Record linkage: individuare i record, che, nelle varie sorgenti, corrispondono alle stesse entità
  - Data fusion: individuare valori corretti per i vari campi

Integrazione di big data: le difficoltà

- Le quattro V
  - Volume
  - Velocità
  - Varietà
  - Veracità

- Ognuna aggiunge difficoltà

Integrazione di big data: le opportunità

- Ridondanza
- Evoluzione lenta

Integrazione di big data: tecniche

- Schema alignment
  - valorizzazione della ridondanza ("voting", tolleranza agli errori)
- Record linkage
  - paradigmi map reduce
  - tecniche di blocking (per ridurre la complessità del clustering)
- Data fusion
  - valutazione dell'affidabilità della sorgente e dell'accuratezza
  - individuazione dei "copiatori"

- Tecniche basate sulla statistica in tutte le fasi

Conclusioni

- I dati sono disponibili in misura sempre maggiore e questo è certamente interessante
- Non costituiscono però un valore sempre e comunque
- Molti aspetti meritano attenzione
  - Qualità dei dati
  - Comprensione dei dati
  - Appropriatezza dei modelli e qualità degli algoritmi
  - Trasparenza dei dati
  - Trasparenza dei modelli e degli algoritmi
  - Integrazione di big data