



SAPIENZA
UNIVERSITÀ DI ROMA

Tecniche Innovative per la Produzione di Dati

Sinergie con i Metodi
Donatella Firmani **Statistici**

donatella.firmani@uniroma1.it

<https://sites.google.com/uniroma1.it/donatellafirmani>

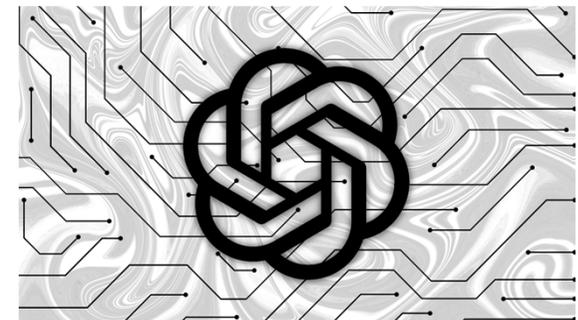
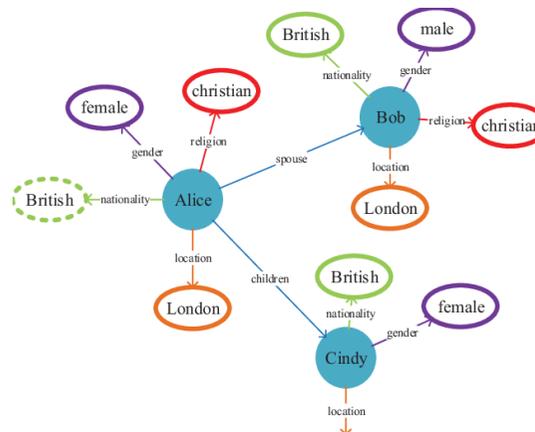
Introduzione

- Abbiamo recentemente assistito ad una esplosione – in quantità e complessità - di dati generati da dispositivi e media
- Molte decisioni si affidano ai dati (data-driven) con effetti sulle politiche e sull'individuo
- E' naturale considerare sfide e opportunità delle nuovi fonti nel contesto della produzione di dati



Fonti secondarie

- Tipicamente fonti Big Data
- Esempio principale: Web
- Altri esempi: Database pubblici / Deep Web, Knowledge Graphs, e più recentemente Large Language Models



Web

- Diversi modi per estrarre dati dal Web
- Cercare dati su un motore di ricerca e accedere alle informazioni integrate
- Google Knowledge Graph
- Scraping di pagine curate manualmente come Wikipedia
- Scraping di siti personali e aziendali

The image shows a Google search interface for 'barack obama'. The search results include a Knowledge Graph card with a Wikipedia snippet. The snippet provides biographical information: Barack Hussein Obama (Honolulu, 4 agosto 1961) è un ex politico e avvocato statunitense, 44° presidente degli Stati Uniti d'America dal 2009 al 2017, prima persona di origini afroamericane a ricoprire tale carica. It also lists his vice president (Joe Biden), duration of mandate (20 gennaio 2009 -), title of study (Bachelor of Arts; Juris Doctor), and universities (Columbia University; Università...). Below the snippet are sections for 'Le persone hanno chiesto anche' (Why did Obama win the Nobel Peace Prize? How many languages does Barack Obama speak? Who is Michelle Obama? Where did Barack Obama originate?), 'Notizie principali' (Barack Obama speaks to 10,000 people in Berlin...), and 'Altre ricerche degli utenti' (Michelle Obama, Joe Biden, Donald Trump, George W. Bush).

Large Language Models

- Anche qui ci sono diversi modi
- Il modo principale per estrarre dati è il prompting, ovvero una richiesta in linguaggio naturale

D genera una tabella con attributi e informazioni su Barack Obama, ad esempio Data Di Nascita: Giorno

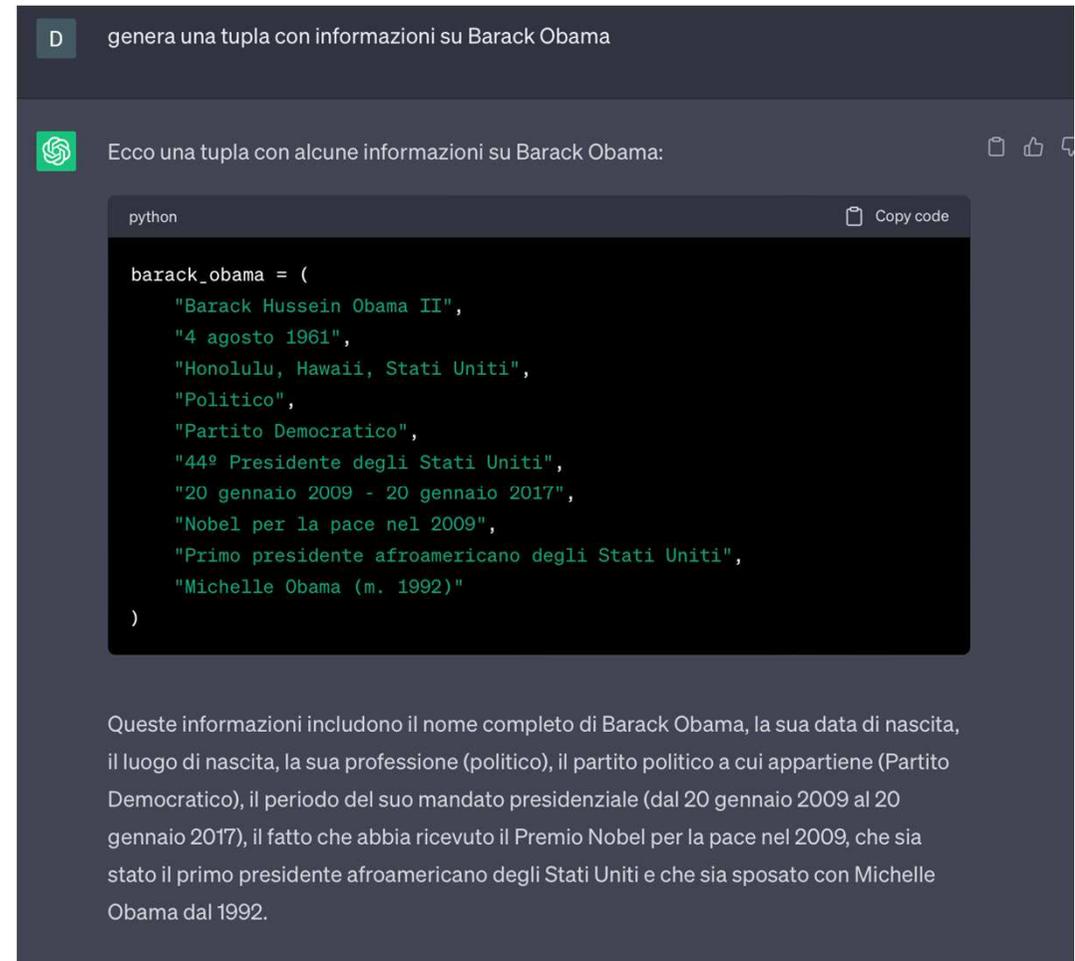
Ecco una tabella con alcuni attributi e informazioni su Barack Obama:

Attributo	Informazione
Nome completo	Barack Hussein Obama II
Data di nascita	4 agosto 1961
Luogo di nascita	Honolulu, Hawaii, Stati Uniti
Nazionalità	Statunitense
Professione	Politico, avvocato
Partito politico	Democratico
Carica	44° presidente degli Stati Uniti
Periodo di mandato	20 gennaio 2009 - 20 gennaio 2017
Coniuge	Michelle Obama
Figlie	Malia Ann e Natasha Marian (Sasha) Obama
Istruzione	Columbia University (BA), Harvard Law School (JD)
Premi e riconoscimenti	Premio Nobel per la pace (2009)
Pagine ufficiali	Sito Web

Si prega di notare che queste informazioni sono aggiornate fino a settembre 2021, quindi alcune informazioni più recenti potrebbero non essere incluse nella tabella.

Large Language Models

- Anche qui ci sono diversi modi
- Diversi prompt, diverse risposte



D genera una tupla con informazioni su Barack Obama

Ecco una tupla con alcune informazioni su Barack Obama:

```
python Copy code

barack_obama = (
    "Barack Hussein Obama II",
    "4 agosto 1961",
    "Honolulu, Hawaii, Stati Uniti",
    "Politico",
    "Partito Democratico",
    "44° Presidente degli Stati Uniti",
    "20 gennaio 2009 - 20 gennaio 2017",
    "Nobel per la pace nel 2009",
    "Primo presidente afroamericano degli Stati Uniti",
    "Michelle Obama (m. 1992)"
)
```

Queste informazioni includono il nome completo di Barack Obama, la sua data di nascita, il luogo di nascita, la sua professione (politico), il partito politico a cui appartiene (Partito Democratico), il periodo del suo mandato presidenziale (dal 20 gennaio 2009 al 20 gennaio 2017), il fatto che abbia ricevuto il Premio Nobel per la pace nel 2009, che sia stato il primo presidente afroamericano degli Stati Uniti e che sia sposato con Michelle Obama dal 1992.

Opportunità e Rischi

A seconda del tipo di fonte secondaria, diverse opportunità e diversi rischi ma in generale si possono identificare I seguenti.

Opportunità

- Aumentare righe e colonne a disposizione
- Indentificare nuove categorie

Rischi

- Varietà: Formati inconsistenti → errori del processo di trasformazione
 - Es: June 4th 1961 → 06/04/1961
- Veridicità: Informazioni inaccurate → errori del dato in sé
 - Es: Pagina web personale riportata come ufficiale

Questo Intervento

BigData 4 Stat

Come possono le nuove tecniche basate su grandi quantità di dati aiutarci a cogliere le opportunità delle fonti secondarie mitigando i rischi

- “Tecniche Big Data” per brevità
- Esempi: Natural Language Processing, Artificial Intelligence, Machine Learning, Clustering, Time Series.

Stat 4 BigData

Come possono i metodi statistici aiutarci a rendere tali tecniche più trasparenti e affidabili, in modo che i ricercatori possano fidarsi dei dati prodotti.



DEPARTMENT OF STATISTICAL SCIENCES

www.dss.uniroma1.it



Pausa

Big Data 4 Stat

- D: Come possono le tecniche BigData aiutarci a produrre nuovi dati?
- R: Fornendoci strumenti per questi, ed altri, problemi
 - Schema Mapping
 - Record Linkage
 - Data Fusion
 - Data Imputation
 - Data Categorization

Nelle prossime diapositive

- Illustrazione ad alto livello di idee e tecniche di **Data Integration**, basate sul concetto di **Embedding**

- Schema Mapping
- Record Linkage
- Data Fusion

Data Integration

Data Integration

	Manufacturer	Model	Resolution
#1	Canon	4000D	18.0Mp
#2	Canon	250D	24.1Mp

Sorgente A

	Brand	Model	Megapixel	Sensor
#1	Canon	4000D	17.9	CMOS

Sorgente B

Soluzione Naïve

	A.Manufactur	A.Model	A.Resolution	B.Brand	B.Model	B.Megapixel	B.Sensor
A.1	Canon	4000D	18.0Mp	NULL	NULL	NULL	NULL
A.2	Canon	250D	24.1Mp	NULL	NULL	NULL	NULL
B.1	NULL	NULL	NULL	Canon	4000D	17.9	CMOS

Data Integration

	Manufacturer	Model	Resolution
#1	Canon	4000D	18.0Mp
#2	Canon	250D	24.1Mp

Sorgente A

	Brand	Model	Megapixel	Sensor
#1	Canon Inc.	4000D	17.9	CMOS

Sorgente B

Vista Unificata

Brand	Name	Megapixel	Sensor
Canon	4000D	18.0	CMOS
Canon	250D	24.1	Null

Embedding: Intuizione

String Similarity: Similarità sintattica di diverse rappresentazioni della stessa entità

- Canon Inc. ~ Canon
- Molto usata in **tecniche classiche** di Data Integration
 - Levenshtein, Edit distance, Jaro-Winkler...

Word Embedding: Similarità semantica!

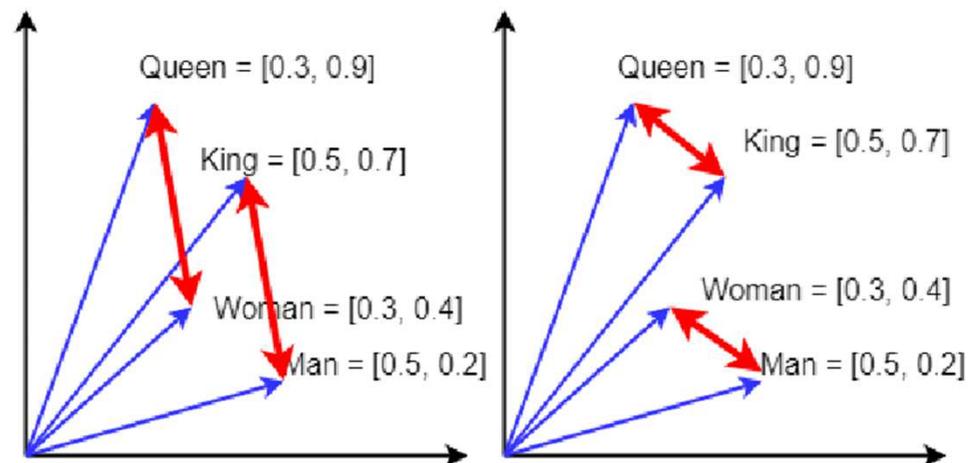
- Canon Inc. ~ Kyanon kabushiki gaisha
- **Nuove tecniche** basate su Language Models

Definizione di Word Embedding

- Embedding: Rappresentazione in uno spazio vettoriale
- Word Embedding: Elemento cruciale di un Language Model, rappresentazione di elementi del linguaggio (parole, verbi, nomi) in uno spazio vettoriale
- Obiettivo in Data Integration: Identificare concetti identificativi su righe e colonne
- Soluzione: Usa una rappresentazione vettoriale che codifica la similarità
 - Canon Inc. → [0.1 0.4 0.5 0.2 0.1 0.3]
 - Kyanon kabushiki gaisha → [0.2 0.8 1.1 0.4 0.3 0.6]
 - Cosine similarity → 0.997 😊
- Problema: Come imparo il mapping tra elementi e vettori?

Tecniche di Word Embedding

- L'idea generale è di rappresentare una parola in modo che sia facile prevederla a partire dal contesto
- Contesti simili, concetti analoghi
 - Il re è stato incoronato
 - La regina è stata incoronata
- Addestramento a partire da enormi quantità di dati testuali (corpora)



Language Models

- Oggetto di attiva ricerca negli ultimi anni
- Word2Vec, FastText e Glove tra I primi
- Recente rivoluzione basata sull'architettura Transformer per le reti neurali profonde (NeurIPS 2017)
 - BERT model and altri (RoBERTa, DistilBERT...)
 - ChatGPT
- Decine di implementazioni disponibili pubblicamente
 - Es. Hugging Face



DEPARTMENT OF STATISTICAL SCIENCES

www.dss.uniroma1.it



Pausa

Embedding e Data Integration

Tipicamente i metodi di Data Integration possono avere

- Una parte non-supervisionata con word embedding (pre-addestrati) per rappresentare la semantica
- Una parte supervisionata, con dati etichettati (labelled) per l'effettivo svolgimento delle operazioni

Dati Etichettati

- Es. Record Linkage → Esempi di coppie di record che rappresentano la stessa entità
- Produrre dati etichettati di etichette quindi si cerca di fare meno ricorso possibile ad esse

Schema Mapping (Columns)

	Manufacturer	Model	Resolution
#1	Canon	4000D	18.0Mp
#2	Canon	250D	24.1Mp

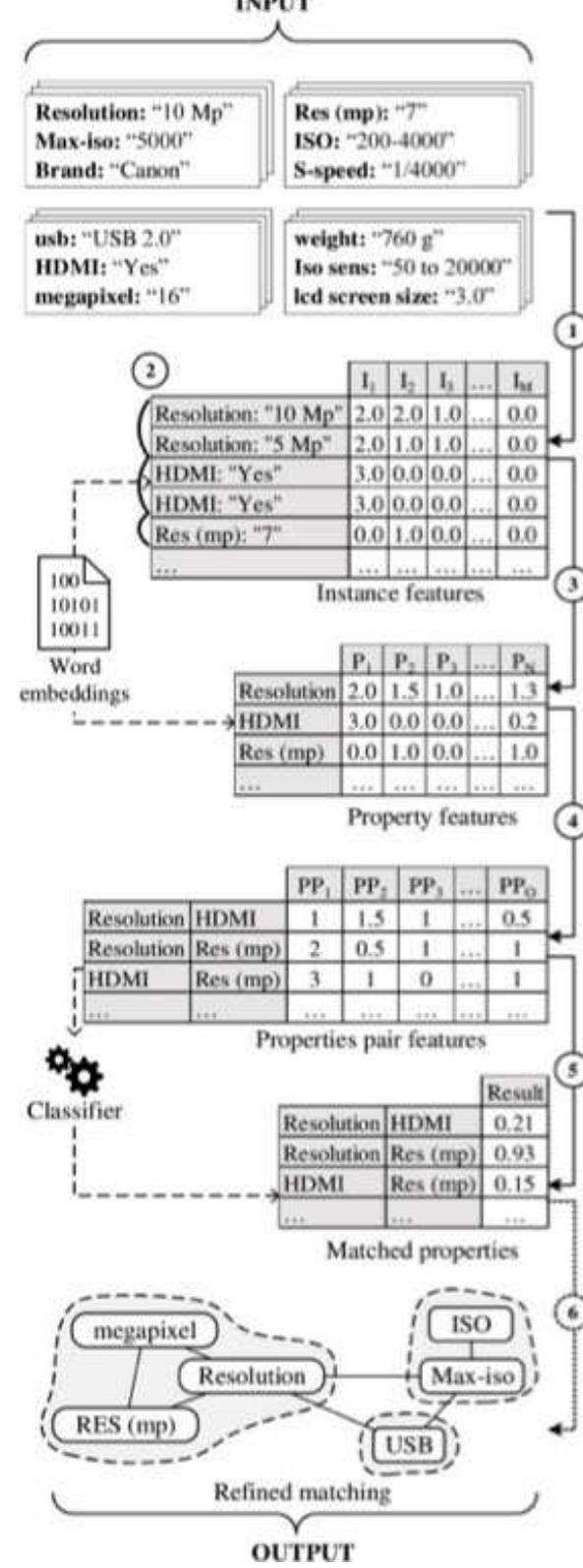
Source A

	Brand	Model	Megapixel	Sensor
#1	Canon	4000D	17.9	CMOS

Source B

	Brand	Model	Resolution	Sensor
A.1	Canon	4000D	18.0Mp	NULL
A.2	Canon	250D	24.1Mp	NULL
B.1	Canon	4000D	17.9	CMOS

Pipeline “Leapme”



Ayala, D., Hernández, I., Ruiz, D., & Rahm, E. (2022). Leapme: Learning-based property matching with embeddings. *Data & Knowledge Engineering*, 137, <https://doi.org/10.1016/j.datak.2021.101943>

Record Linkage (Rows)

	Manufacturer	Model	Resolution
#1	Canon	4000D	18.0Mp
#2	Canon	250D	24.1Mp

Source A

	Brand	Model	Megapixel	Sensor
#1	Canon	4000D	17.9	CMOS

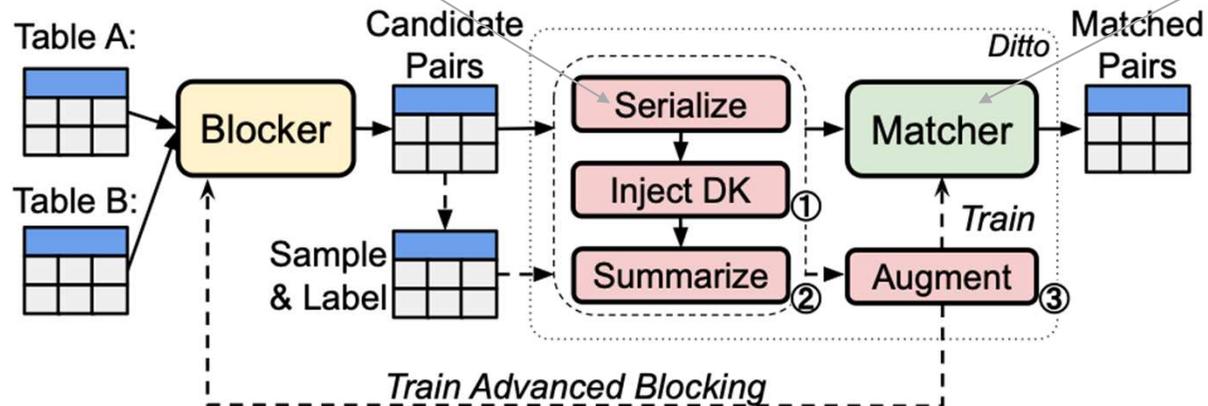
Source B

	Brand	Model	Resolution	Sensor
A.1 B.1	Canon	4000D	18.0Mp 17.9	CMOS
A.2	Canon	250D	24.1Mp	NULL

Pipeline “Ditto”

[COL] title [VAL] instant immers spanish dlux 2 [COL]
manf./modelno [VAL] NULL [COL] price [VAL] 36.11

Fine-Tuned LM + fully
connected layer and a
softmax output layer



Data Fusion (Cells)

	Manufacturer	Model	Resolution
#1	Canon	4000D	18.0Mp
#2	Canon	250D	24.1Mp

Source A

	Brand	Model	Megapixel	Sensor
#1	Canon	4000D	17.9	CMOS

Source B

	Brand	Model	Resolution	Sensor
A.1 B.1	Canon	4000D	18.0Mp 17.9	CMOS
A.2	Canon	250D	24.1Mp	NULL

Multiple Pipelines

Tipicamente si usano diverse pipeline a seconda del dominio

- Dati biomedici
- Dati IoT
- Dati anagrafici

Si possono fondere dati con più modalità, ad esempio l'immagine del logo di una azienda e il testo con il nome



DEPARTMENT OF STATISTICAL SCIENCES

www.dss.uniroma1.it



Pausa

Data Imputation

	Manufacturer	Model	Resolution
#1	Canon	4000D	18.0Mp
#2	Canon	250D	24.1Mp

Source A

	Brand	Model	Megapixel	Sensor
#1	Canon	4000D	17.9	CMOS

Source B

	Brand	Model	Resolution	Sensor
A.1 B.1	Canon	4000D	18.0Mp 17.9	CMOS
A.2	Canon	250D	24.1Mp	NULL

Approcci principali

Imputare I dati da quelli che ci sono

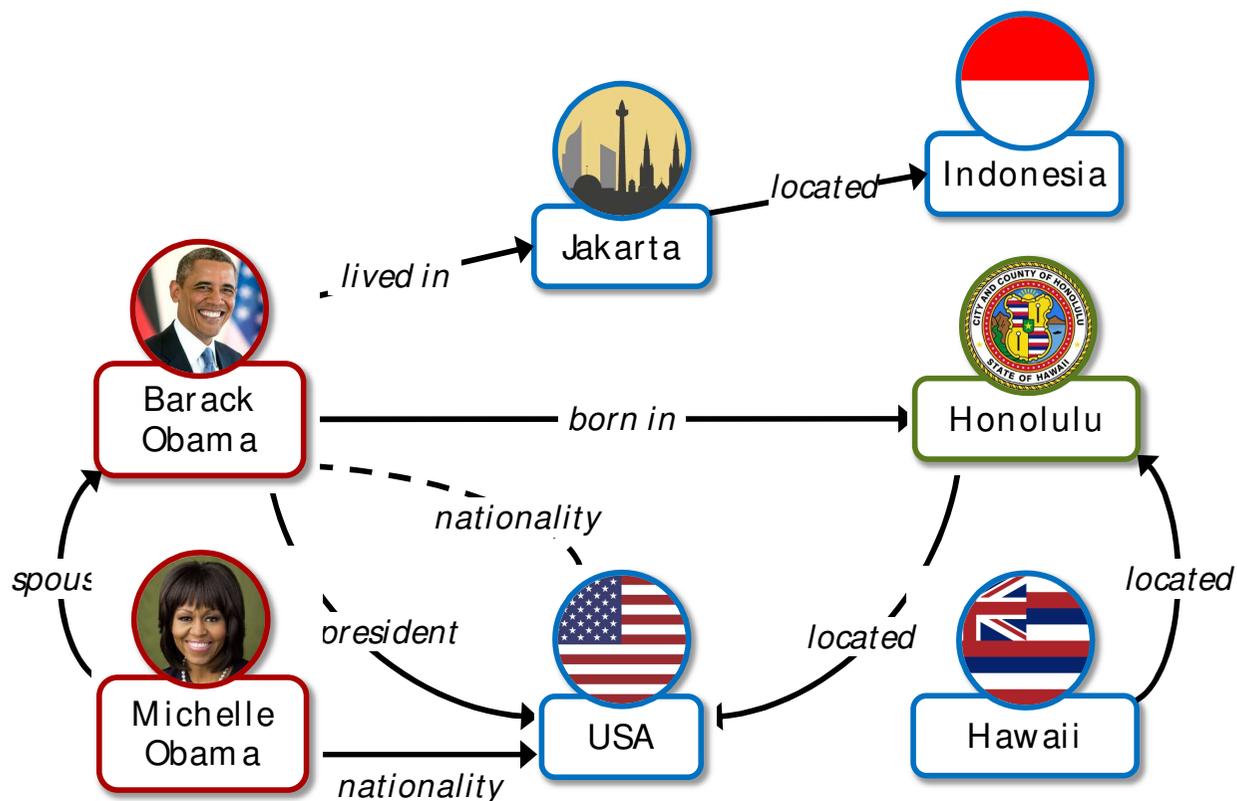
- ML: Random Forests, Compression
- Deep Learning: Autoencoder, GAN
- Modelli di errore: missing completely at random, missing at random, and missing not at random

Imputare i dati tramite ragionamento

- Es. Sorgenti esterne come I Knowledge Graphs
- **Non proprio imputazione → Predizione**

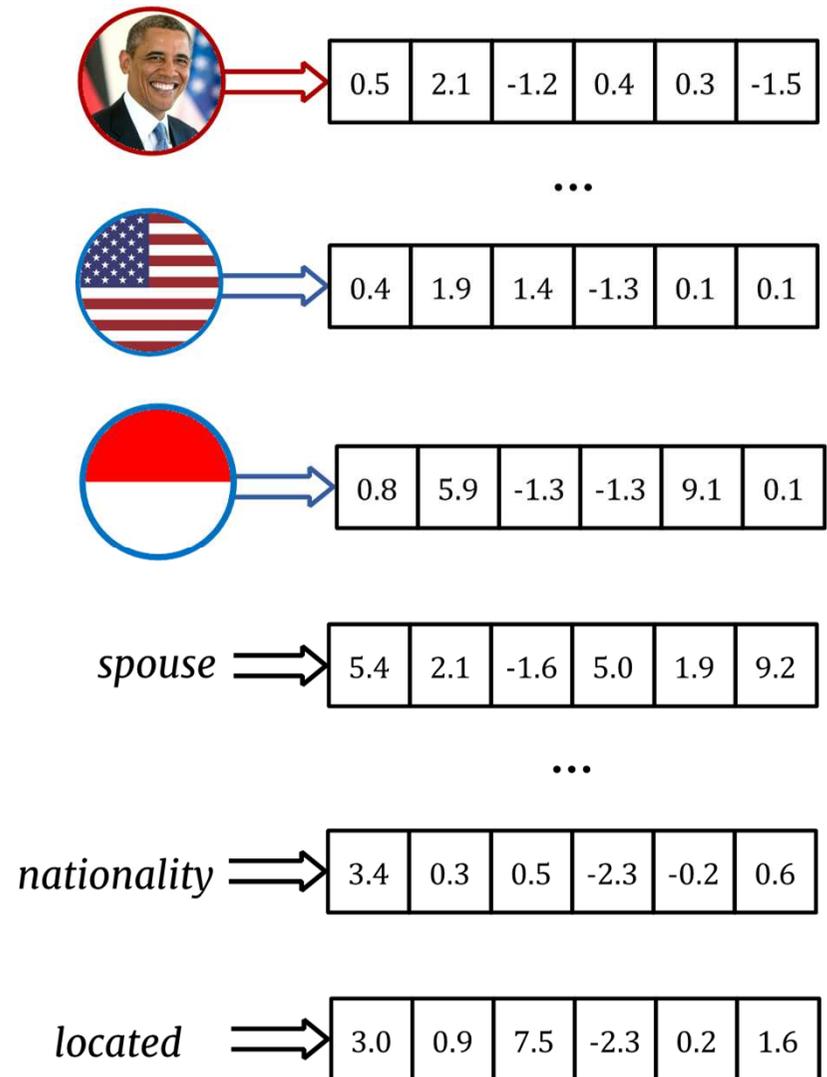
Knowledge Graphs

- Un fatto mancante che possiamo predire è <Barack Obama, Nationality, USA>



KG Embedding

- Entities and relations are mapped to vectors or matrices
- A scoring function ϕ estimates the plausibility of a triple $\langle h, r, t \rangle$ by combining embeddings of h , r and t
- The loss function used for learning embeddings is
 - related to ϕ
 - minimized for facts that are already in the KG





DEPARTMENT OF STATISTICAL SCIENCES

www.dss.uniroma1.it

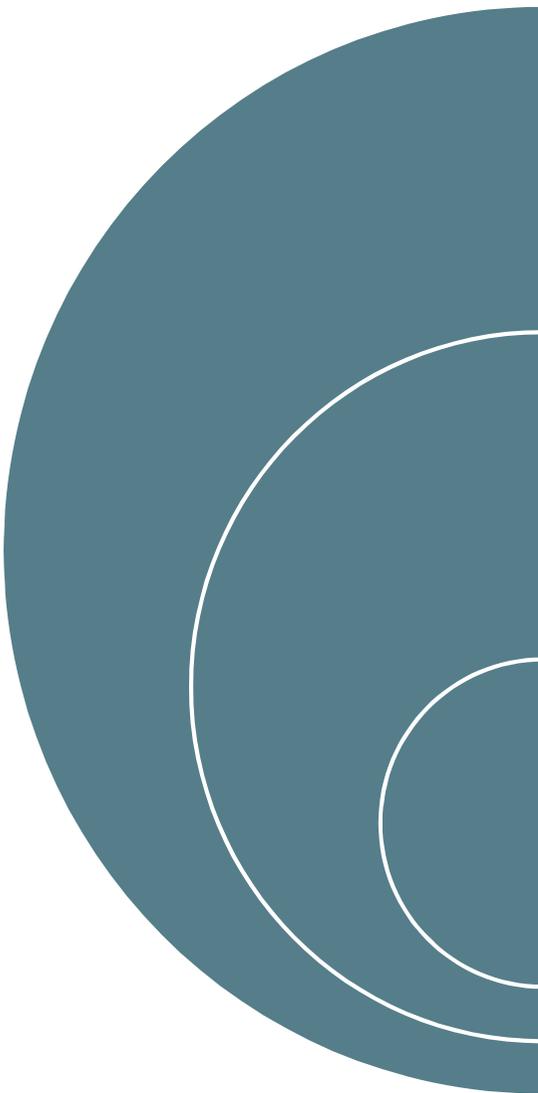


Pausa

Stat 4 Big Data

- D: Come possono le tecniche statistiche aiutarci a fidarci delle tecniche BigData?
- R: Questo è oggetto di dibattito attuale, ma in generale è noto che le tecniche Big Data possono soffrire di
 - Opacità
 - Bias
 - Dipendenza dal dominio

Opacità



Sfruttando i dati, le tecniche viste fin ora possono risolvere un compito specifico senza l'uso di istruzioni esplicite.

Questo permette di raggiungere un'accuratezza senza precedenti in molti scenari applicativi.

Explainable AI: Nuova disciplina interessata alla domanda "Perché" anziché "Quanto accurato"?

Opacità nel Record Linkage

- Prendiamo DeepMatcher, una pipeline simile a quella vista finora
- F-Measure = 94.9 %

	Album	Artist	Copyright	Genre	Price	Date	Song Name	Time
iTunes	Flo Rida	Flo Rida	Atlantic Recording	Hip-Hop/Rap , Music , Dirty South	\$ 1.79	17-mar-08	Elevator (feat . Timbaland)	3:55
Amazon	Flo Rida	Flo Rida	2008 Atlantic Recording Corporation for the ...	Hip-Hop & Rap	1.9 USD	03/17/2008	Elevator	3:55

Risultato: Match

Opacità nel Record Linkage

- Intuizione: perturbazioni randomiche
- Le variabili con più probabilità di cambiare il risultato sono più utili a capire il perché della classificazione

	Album	Artist	Copyright	Genre	Price	Date	Song Name	Time
iTunes	Flo Rida	Flo Rida	Atlantic Recording	Hip-Hop/Rap , Music , Dirty South	\$ 1.79	17-mar-08	Elevator (feat . Timbaland)	3:55
Amazon	Flo Rida	Flo Rida	2008 Atlantic Recording Corporation for the ...	Hip-Hop & Rap	██████████	03/17/2008	Elevator	3:55

Risultato? No Match

Opacità nel Record Linkage

- Esperimenti con perturbazioni randomiche ripetute hanno identificato la durata della canzone (Time)
- Necessaria selezione più accurata dei dati di training

	Album	Artist	Copyright	Genre	Price	Date	Song Name	Time
iTunes	Flo Rida	Flo Rida	Atlantic Recording	Hip-Hop/Rap , Music , Dirty South	\$ 1.79	17-mar-08	Elevator (feat . Timbaland)	3:55
Amazon	Flo Rida	Flo Rida	2008 Atlantic Recording Corporation for the ...	Hip-Hop & Rap	1.9 USD	03/17/2008	Elevator	3:55



Conclusioni

Scenario

- L'aumento della quantità e della complessità dei dati nel mondo odierno presenta sia sfide che opportunità per la produzione di dati di alta qualità

Soluzione

- L'integrazione delle tecniche di big data nei metodi statistici consolidati offre una soluzione efficace per ottenere dati comprensivi e flessibili
- Queste tecniche si basano sui recenti progressi dell'intelligenza artificiale, in particolare dei modelli del linguaggio e delle reti neurali profonde
 - abilitate proprio dai big data!

Spunti

- È fondamentale rendere le tecniche big data più trasparenti e affidabili.
- Una direzione promettente è quella dell'utilizzo di metodi statistici, come in un circolo virtuoso.

Ulteriori riferimenti

- Problemi Aperti
 - Trasparenza
 - Auto configurabilità
 - Conoscenza del dominio
- Ulteriori riferimenti
 - Chengliang Chai, Nan Tang, Ju Fan, and Yuyu Luo. 2023. Demystifying Artificial Intelligence for Data Preparation. In SIGMOD/PODS '23 Companion. ACM, New York, NY, USA, 13–20.
<https://doi.org/10.1145/3555041.3589406>



DEPARTMENT OF STATISTICAL SCIENCES

www.dss.uniroma1.it



Grazie per l'Attenzione