SWAP researchgroup

Università degli Studi di Bari Aldo Moro
Dipartimento di Informatica

Anno Accademico 2013/2014

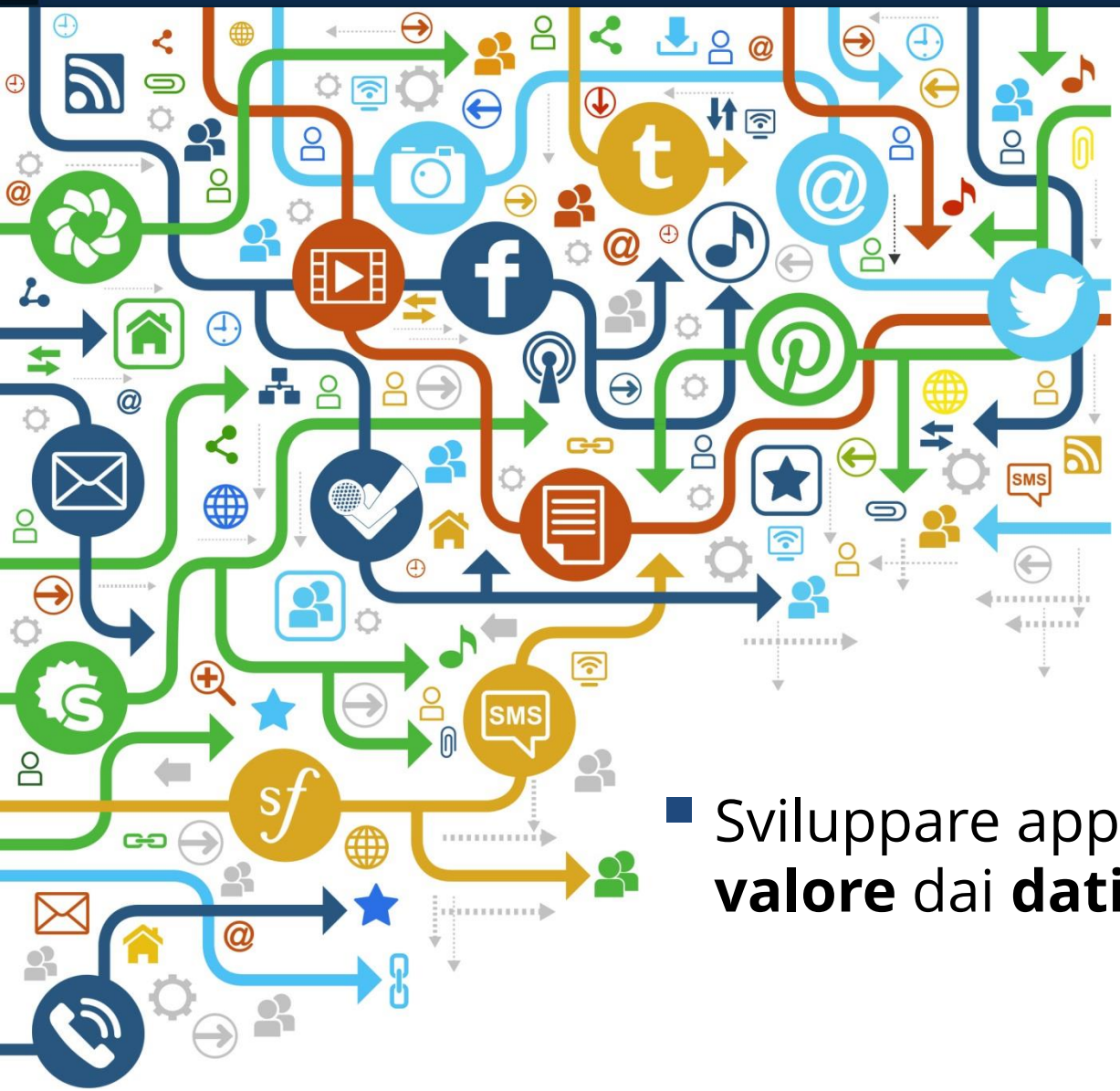
Piattaforma cloud per l'analisi di big data provenienti da social network

Tecnologie per l'individuazione di contenuti a sfondo discriminatorio

Relatori:
Prof. Pasquale Lops
Dott. Pierpaolo Basile

Laureando:
Gianvito Taneburgo

Scenario della ricerca



- Grande mole di dati **non strutturati**:
  - video
  - immagini
  - log
  - email
  - ecc.

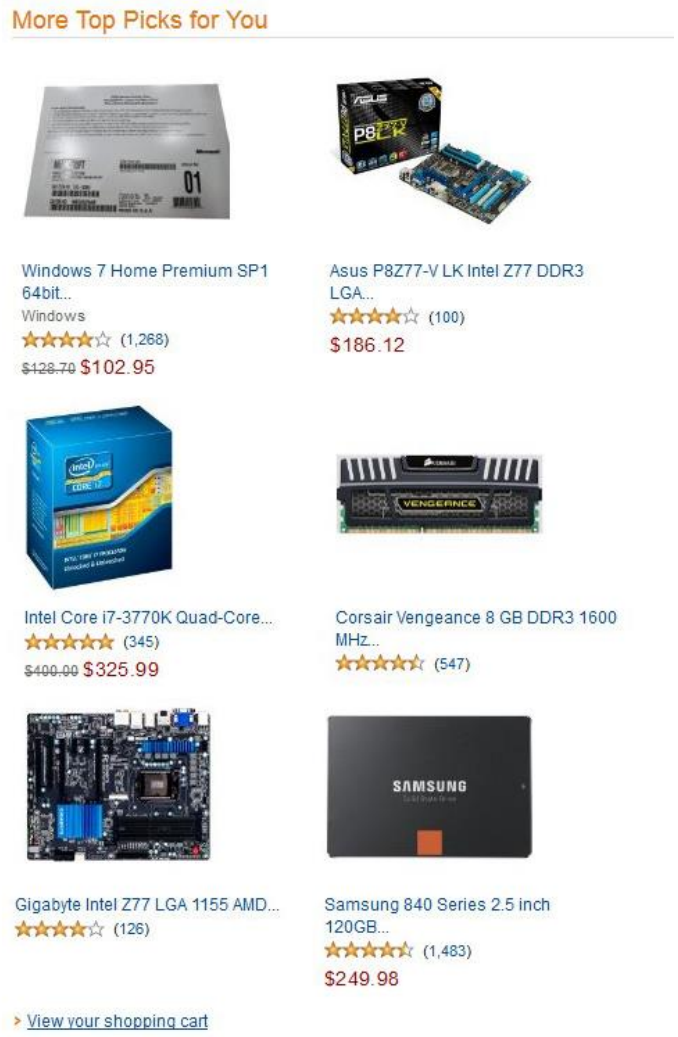
- Sviluppare applicazioni per estrarre **valore** dai **dati grezzi**

Opportunità

- **Brand reputation analysis**
- Recommender system
- Targeted advertising
- Query log mining

Opportunità

- Brand reputation analysis
- **Recommender system**
- Targeted advertising
- Query log mining



Opportunità

- Brand reputation analysis
- Recommender system
- **Targeted advertising**
- Query log mining

Opportunità

- Brand reputation analysis
- Recommender system
- Targeted advertising
- **Query log mining**

Problematiche

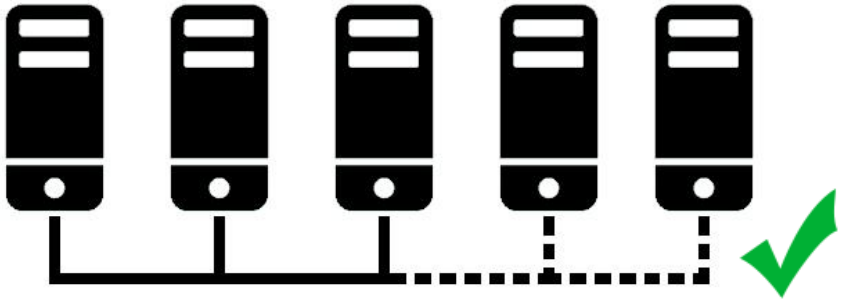
Strumenti adatti ai dati

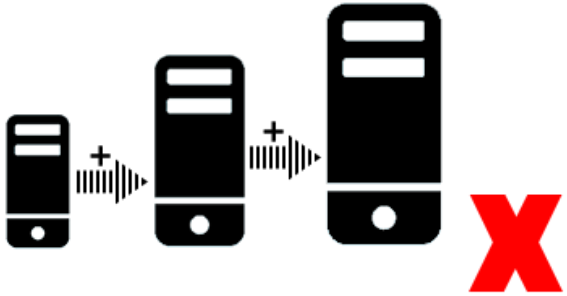
Memorizzare big data

Elaborare big data

Soluzioni

Tecnologie **scalabili**





Piattaforme di **cloud** computing

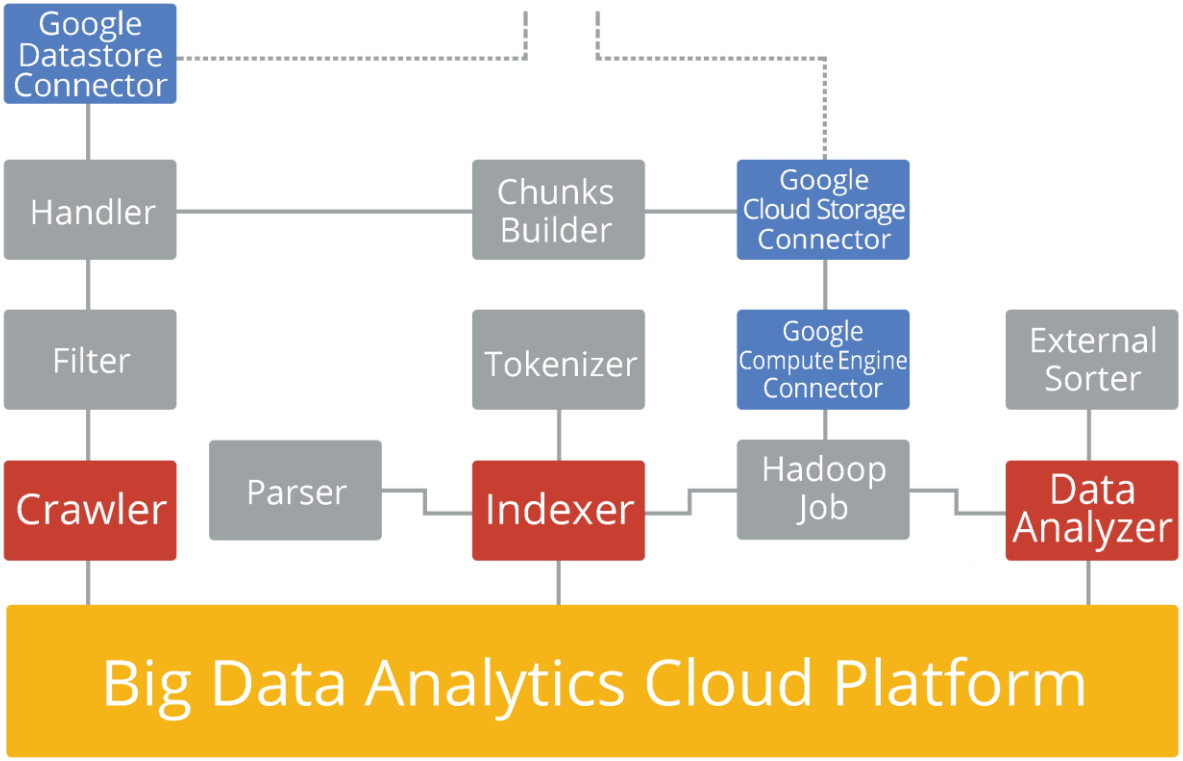
amazon web services™

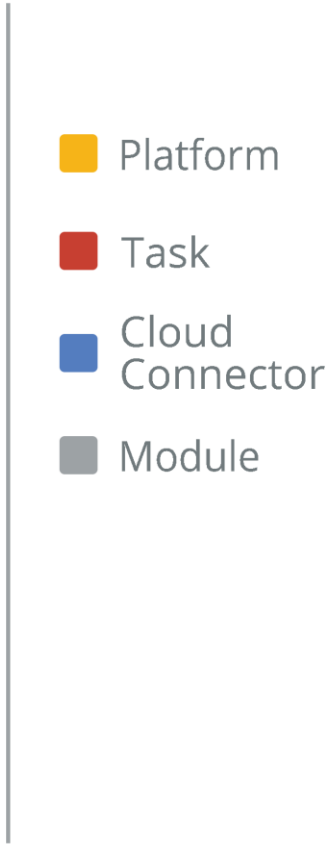
Google Cloud Platform

SOFTLAYER® an IBM Company

Architettura della piattaforma

Google Cloud Platform

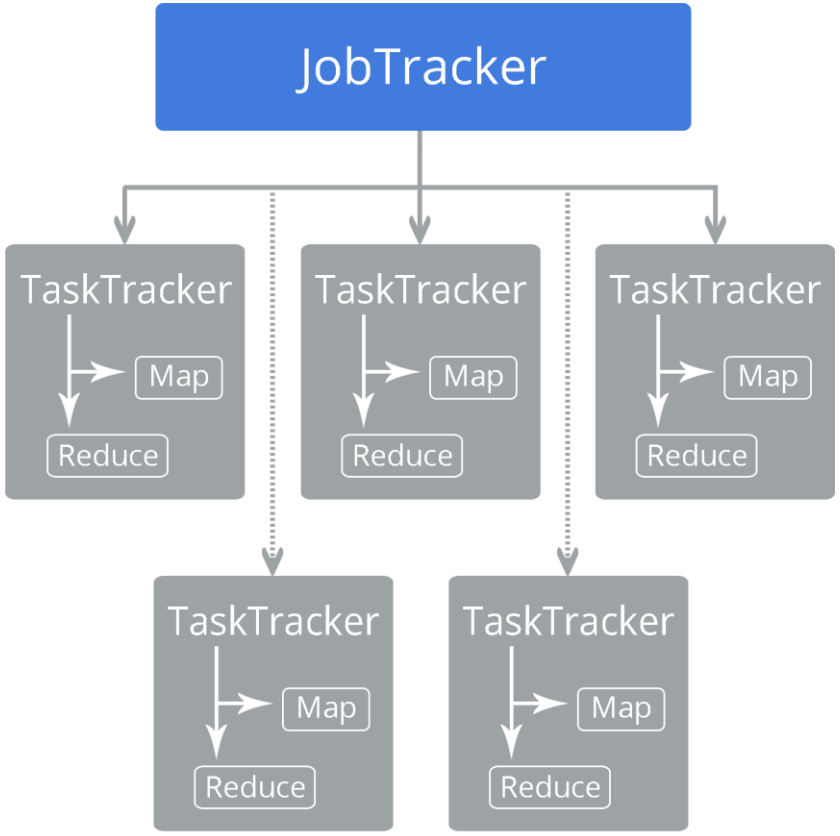




Apache Hadoop

Framework per il **calcolo distribuito** tra nodi di un cluster

- Architettura master/slave
- Modello di programmazione MapReduce
  - **map**: elaborazione locale
  - **reduce**: aggrega i risultati
- Hadoop Distributed File System
- Commodity hardware
- Scalabilità orizzontale



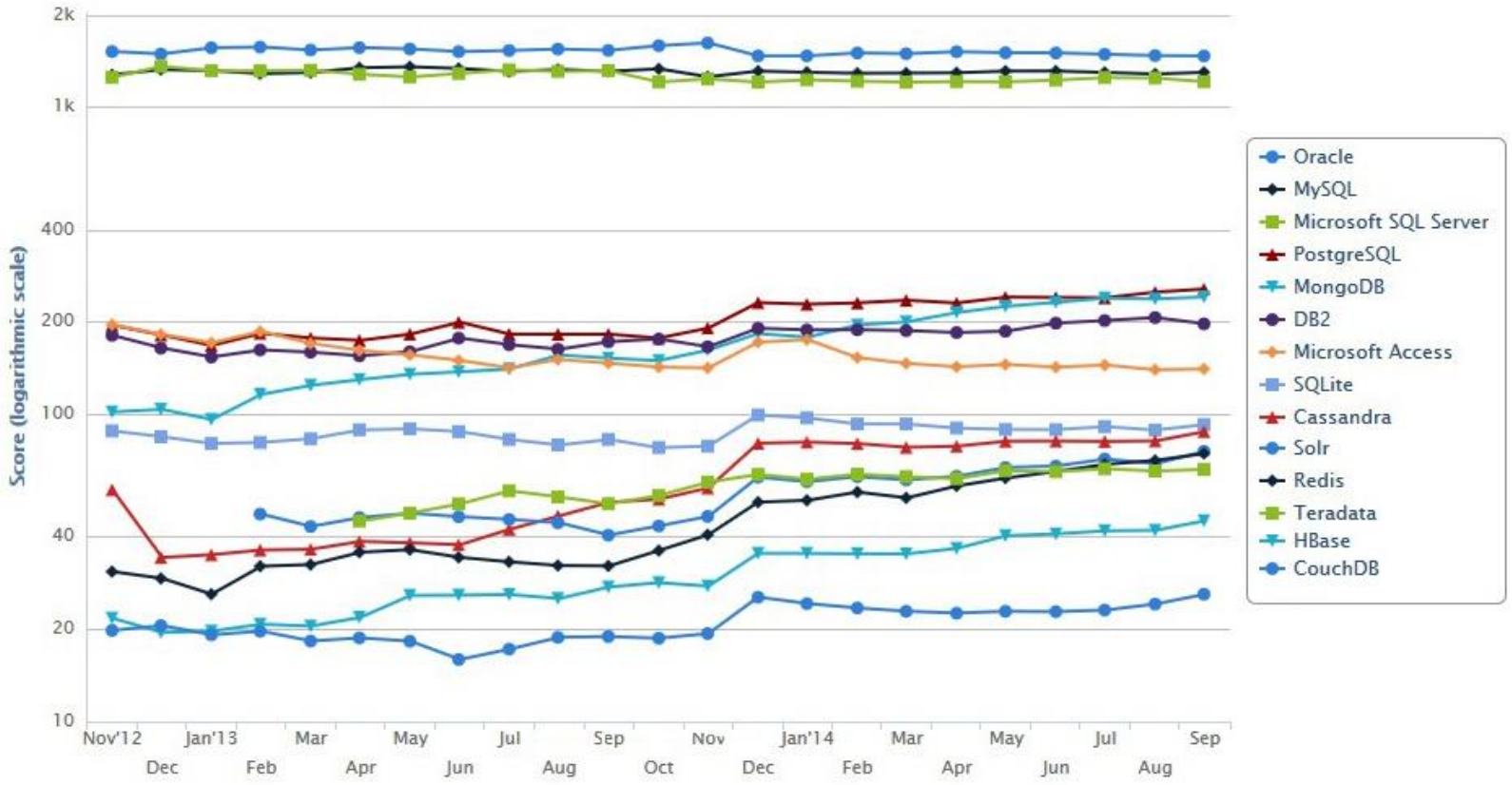
Database NoSQL

Vantaggi

- scalabilità orizzontale
- modello flessibile dei dati

Svantaggi

- transazioni ACID → BASE
- teorema CAP



Classifica dei DBMS per numero di installazioni

Cloud provider

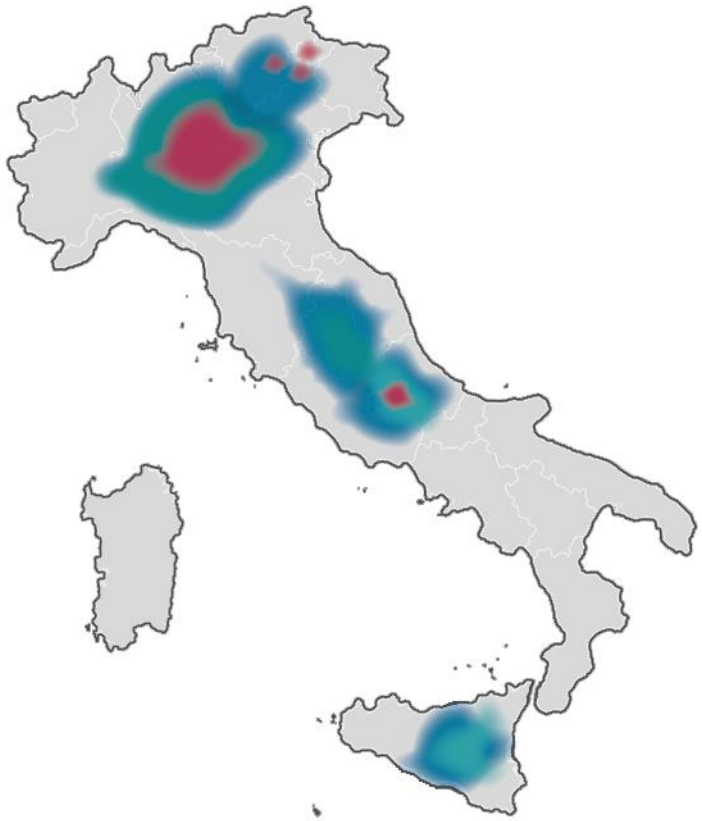
- Infrastructure-as-a-Service (IaaS)
- Platform-as-a-Service (PaaS)
- Software-as-a-Service (SaaS)
- Database-as-a-Service (DaaS)

Cloud provider

- Compute Engine: IaaS
- Cloud Storage: IaaS
- Datastore: DaaS

La «Mappa dell'Intolleranza»

- Quanto siamo razzisti?
- Quanto siamo omofobi?
- Quanto discriminiamo il prossimo?



Nel 2013:

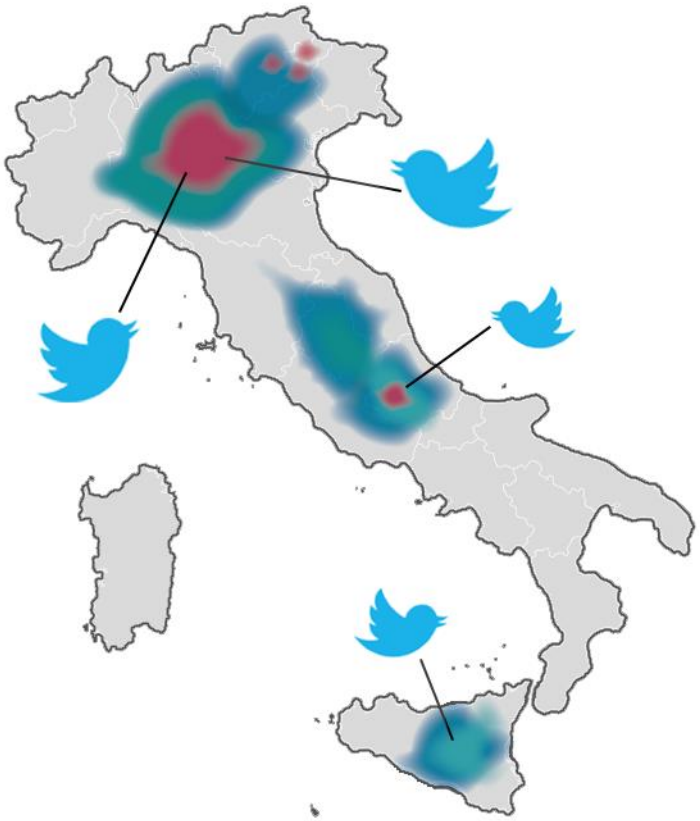
- 45% dei giovani si è dichiarato xenofobo o diffidente degli stranieri
- 92% delle persone LGBT discriminate per l'orientamento sessuale
- 25% degli omosessuali e 6.743.000 donne vittime di violenze

La «Mappa dell'Intolleranza»

- Quanto siamo razzisti?
- Quanto siamo omofobi?
- Quanto discriminiamo il prossimo?

Nel 2013:

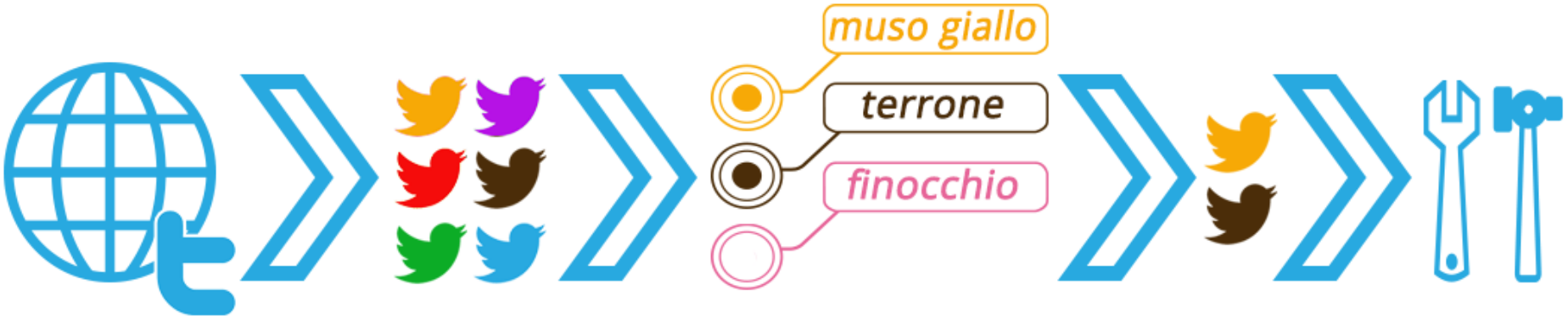
- 45% dei giovani si è dichiarato xenofobo o diffidente degli stranieri
- 92% delle persone LGBT discriminate per l'orientamento sessuale
- 25% degli omosessuali e 6.743.000 donne vittime di violenze



Raccolta preliminare dei tweet

OMOFOBIA	RAZZISMO	DISABILITÀ	MISOGINIA	ANTISEMITISMO
finocchio	neg*o	nano	baldr**ca	ebreo ai forni
ricch***e	terrone	storpio	zoc**la	rabbino
fr***o	zingaro	spastico	boc***nara	giudeo
rotti***lo	muso giallo	zoppo	tro**na	ebreo di me**a
cul***one	crucco	cerebroleso	mign**ta	

Esempi di seed per classi di discriminazione



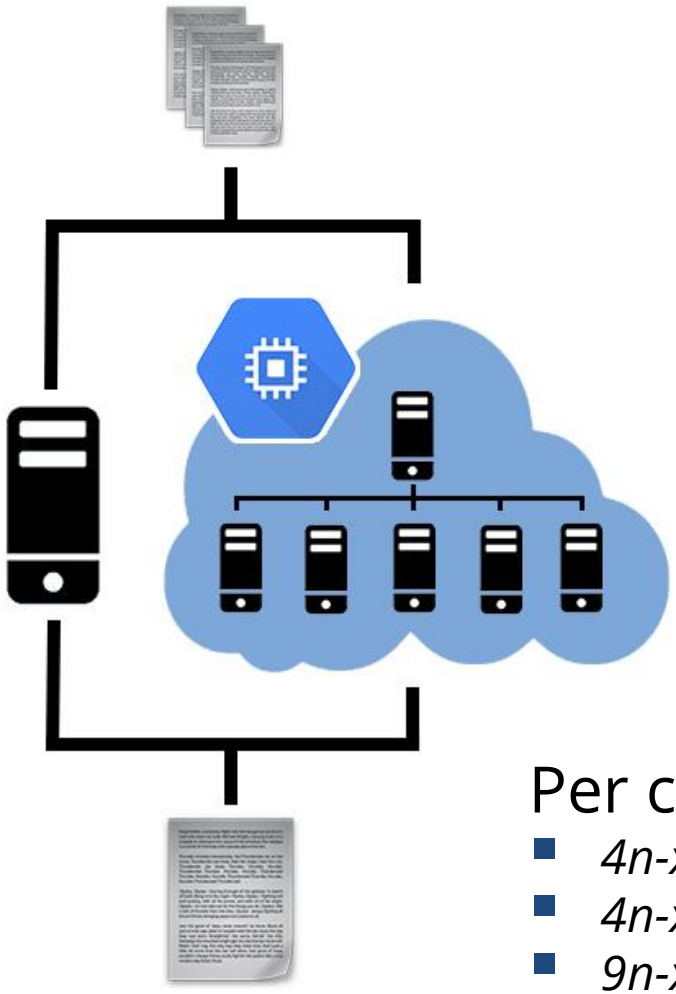
1st Research question

Quantificare vantaggi e svantaggi di una soluzione cloud per l'elaborazione distribuita di grandi e piccoli dataset

Collezione	Contenuto	Dimensione
itWaC	1.870.000 documenti	11,16 GB
omofobia	22.564 tweet	1.738 kB
disabilità	29.793 tweet	2.242 kB
misoginia	249.425 tweet	17.935 kB

Dati di input dell'esperimento

Protocollo sperimentale



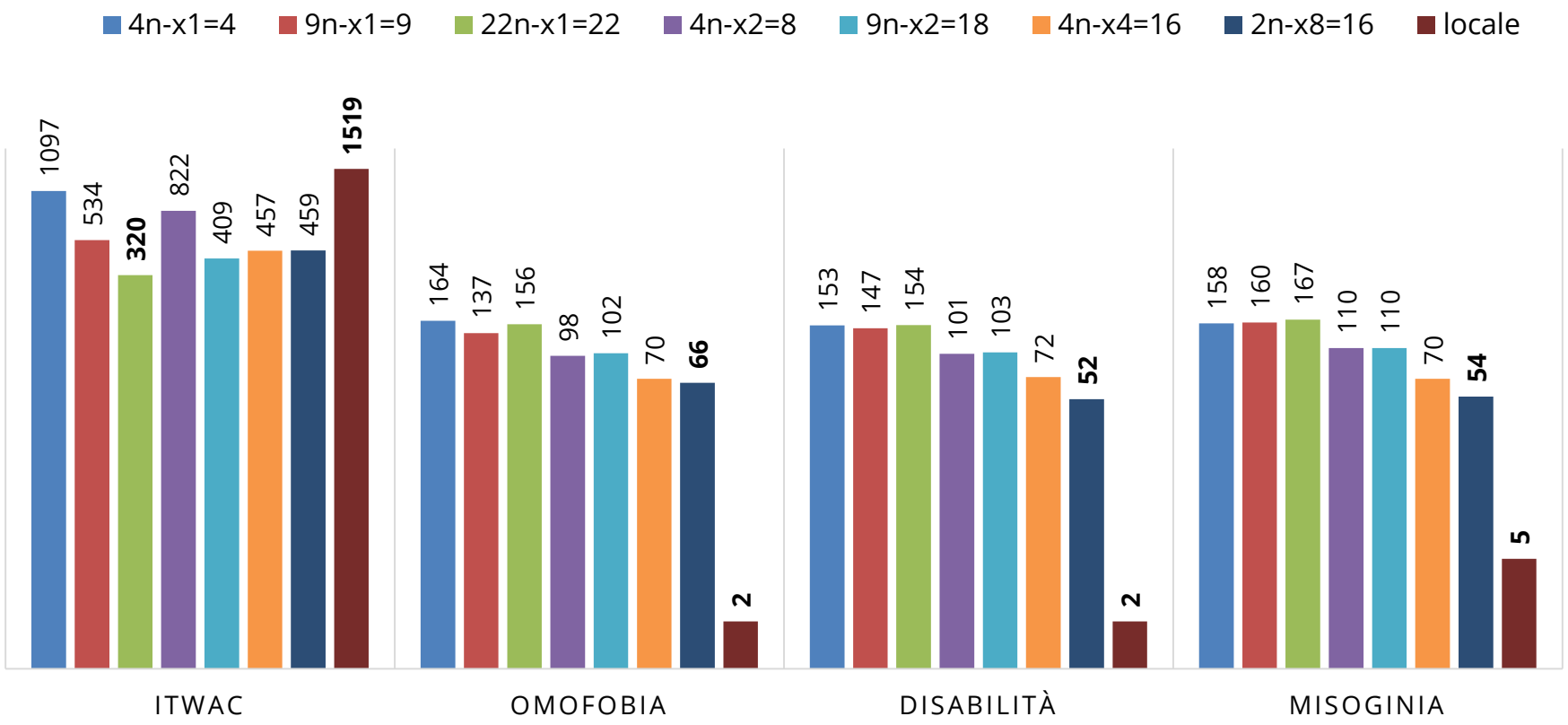
Confronto dei tempi medi di indicizzazione con Hadoop tra:

- macchina locale (cluster in modalità pseudo-distribuita)
- cluster su Compute Engine differenti tra loro per
  - numero di nodi interconnessi
  - tipo di macchine virtuali

Per convenzione:

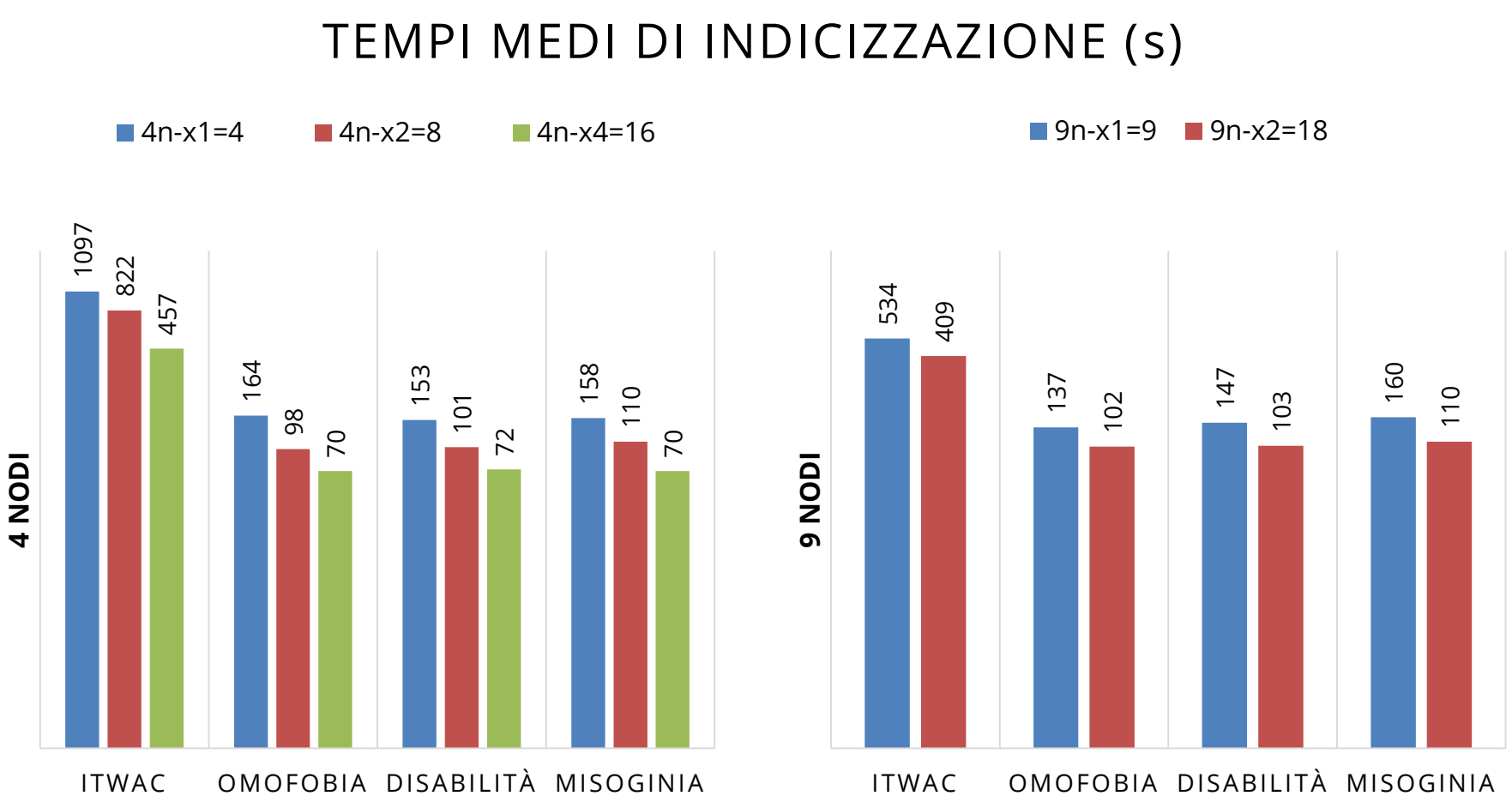
- *4n-x1=4*: cluster con 4 nodi da 1 CPU ciascuno (4 tot)
- *4n-x2=8:* cluster con 4 nodi da 2 CPU ciascuno (8 tot)
- *9n-x2=18:* cluster con 9 nodi da 2 CPU ciascuno (18 tot)
- …
- *#nodi–tipo_VM=CPU_cluster*

Risultati



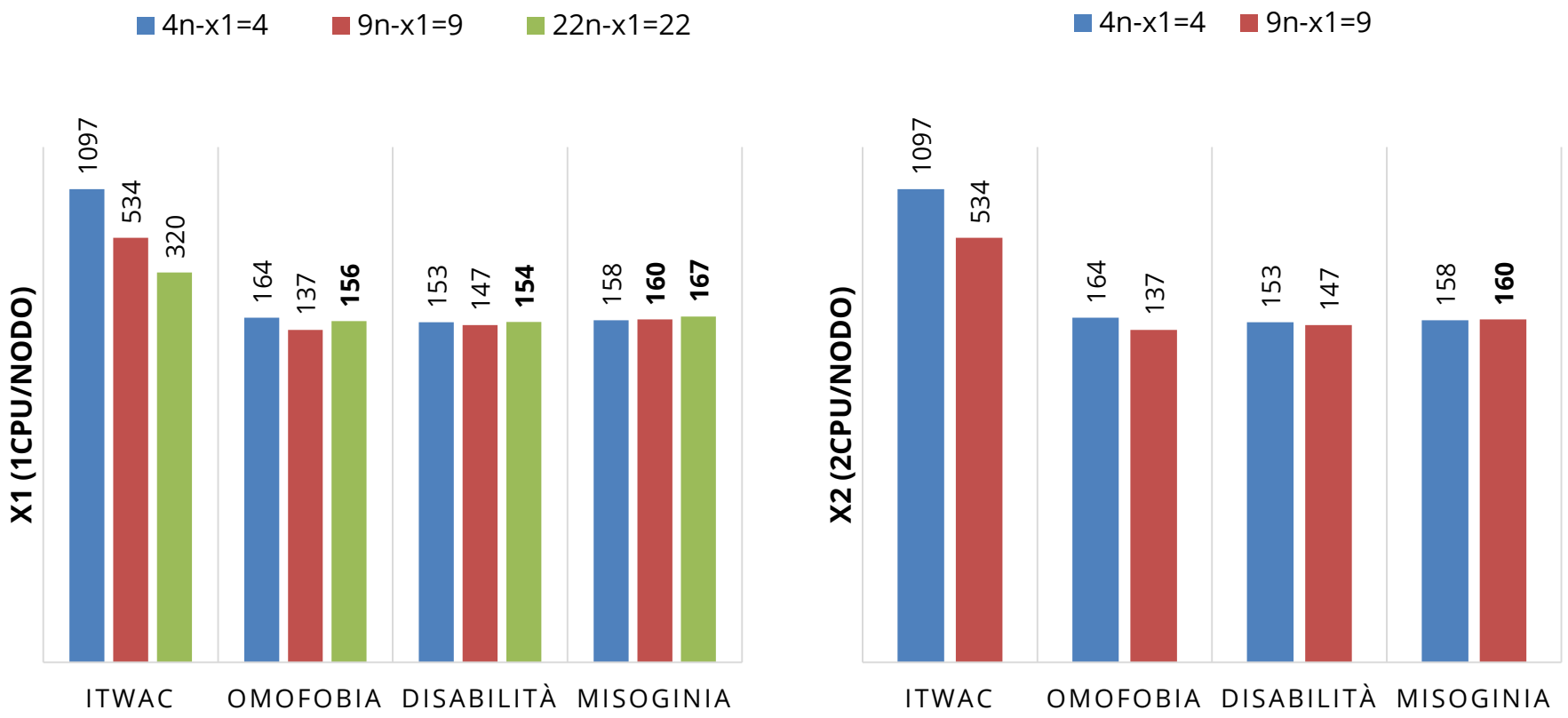
Grandi dataset: miglior cluster outperforms macchina locale
Piccoli dataset: macchina locale outperforms miglior cluster

Risultati



Mantenendo costante il numero di nodi e migliorando il tipo di macchina virtuale, i tempi di indicizzazione diminuiscono sempre

Risultati



Mantenendo costante il tipo di macchina virtuale ed aumentando il numero di nodi, i tempi di indicizzazione diminuiscono solo su grandi dataset

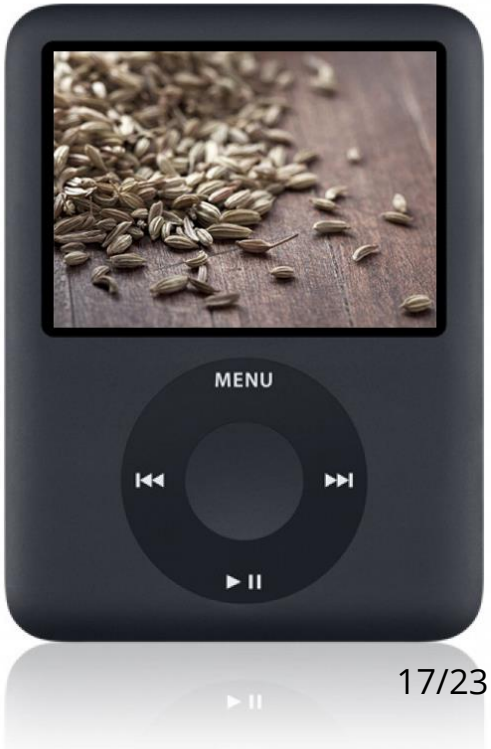
Problematiche riscontrate

Language model

- *RT #Vogliadi #solo5ingredienti @CraftMarmalade: Frollini all'olio d'oliva e semi di finocchio ;D http://t.co/GZd7loEcbx*
  - *retweet, hashtag, menzioni, emoticon, URL, abbreviazioni, ecc.*

Falsi positivi

- *Angolo del risparmio: iPod **nano** con il 30% di sconto! http://t.com/VwnCRlhCLH*
- *Cuoco omofobo espelle dalla cucina anche i semi di **finocchio***



2nd Research question

Migliorare i seed ricercando ed analizzando
i termini più significativi nei tweet

Collezione	Cardinalità vocabolario
itWaC	4.431.080 termini
omofobia	54.507 termini
disabilità	72.569 termini
misoginia	280.412 termini

Dati di input dell'esperimento

Protocollo sperimentale

- **Divergenza di Kullback-Leibler**: misura la prossimità tra due distribuzioni di probabilità discrete:
  - *C*: corpus itWaC
  - *T*: collezione di tweet di una classe
- **Pointwise Kullback-Leibler divergence**: è il contributo dato da una parola *s* alla divergenza complessiva di T da C:

$$\delta_S(T\|C) = P_T(s) \ln \frac{P_T(s)}{P_C(s)}$$

  - Correzione di Laplace: $P_D(i) = \frac{f_{D,i} + 1}{f_D + |D|}$
  - Calcolo distribuito con Hadoop: **14 milioni di probabilità**!
- Le parole con PKLD maggiore sono quelle più significative nella collezione *T* quando questa viene confrontata con *C*

Risultati

Nelle prime posizioni dei ranking emergono:

- bestemmie, imprecazioni, turpiloquio
- personaggi famosi (politici, giovani cantanti, ecc…)
- trasmissioni televisive
- termini complementari ai seed in espressioni d'uso comune (nano: Biancaneve, iPod, giardino; finocchio: semi, tisana)

- almeno **due nuovi seed** tra i primi 300 termini più significativi per ogni classe

Conclusioni

- Sistemi distribuiti per memorizzare ed elaborare dati:
  - controproducenti per piccoli dataset
  - indispensabili per big data

- Piattaforme di cloud computing:
  - migliorano le performance delle tecnologie scalabili
  - facili da configurare ed utilizzare
  - economicamente vantaggiose (costo esperimenti: 21€)

- Raccolta ed analisi dei tweet:
  - rimozione di seed poco significativi e filtraggio in base al contesto
  - aggiunta di nuovi seed

Sviluppi futuri

- Piattaforma d'analisi dei dati:
  - aggiungere task di data visualization
  - integrare altri prodotti cloud di Google (es: Big Query)
- Elaborazione dei tweet:
  - indagare sulla semantica dei messaggi (es: sentiment analysis)
  - apprendere automaticamente language model dei tweet
- Riconoscimento di contenuti discriminatori:
  - costruire classificatori per i tweet
  - studiare la personalità dei discriminatori e predire quella degli individui per cercare corrispondenze
  - analizzare la struttura del grafo del social network

Grazie per l’attenzione