I Big Data nelle scienze sociali e in geografia

Mario Verdicchio
Università degli Studi di Bergamo
Project Work Geourbanistica
23 aprile 2020


Participants (3)
Panelists (3)
Viewers (0)
Joanna Jiang (Host, me, participant ID:18)
JS
Jill Sanders
Leo Wang
Mute All
Unmute All
More
Chat
From Me to Everyone:
How is everyone?
Let me know if you have any questions. I would love to answer.
From Me to Leo Wang: (Privately)
Can you hear me?
Let's keep the conversation going?
From Me to All Panelists:
Great job, guys
More
To: All Panelists
Type message here...
End Meeting
Mute
Stop Video
Participants
3
Q&A
Polling
Share Screen
Chat
More

Sommario 1/2

- La vita sociale sempre più
  - viene mediata da sistemi digitali
  - si svolge in ambienti digitali
- Che siano “big” o meno, sono generati dati da questa digitalizzazione della vita sociale, in cui distinguiamo
  - vita digitale
  - tracce digitali
  - vita digitalizzata

Sommario 2/2

- Molti sono convinti del potenziale di questi dati per lo studio di una varietà di fenomeni altrimenti difficili da osservare
- Ci sono, però,
  - numerose vulnerabilità che ricorrono in questo tipo di uso di dati
  - questioni etiche nuove che emergono e devono ancora essere codificate dalle istituzioni responsabili
  - nuovi trend nell'uso dei dati che non sono necessariamente influenzati dalle suddette questioni

Archivi

- Archivi sulle attività umane esistono da millenni
- La novità di questi ultimi decenni (o ultimo decennio, se vogliamo concentrarci sull'avvento degli smartphone) è data da una combinazione di
  - archivi digitali
  - crescita esponenziale della potenza di calcolo dei computer
- Il potenziale di trasformazione di discipline come la sociologia e la geografia è dato dalla possibilità di studiare la connettività di intere società, in termini di:
  - chi comunica con chi
  - su che cosa si sta comunicando
  - come si muovono le persone negli spazi
  - chi dice che cosa
  - chi compra che cosa
  - etc.
- Il tutto con una granularità temporale di secondi o minuti

Problemi degli archivi digitali

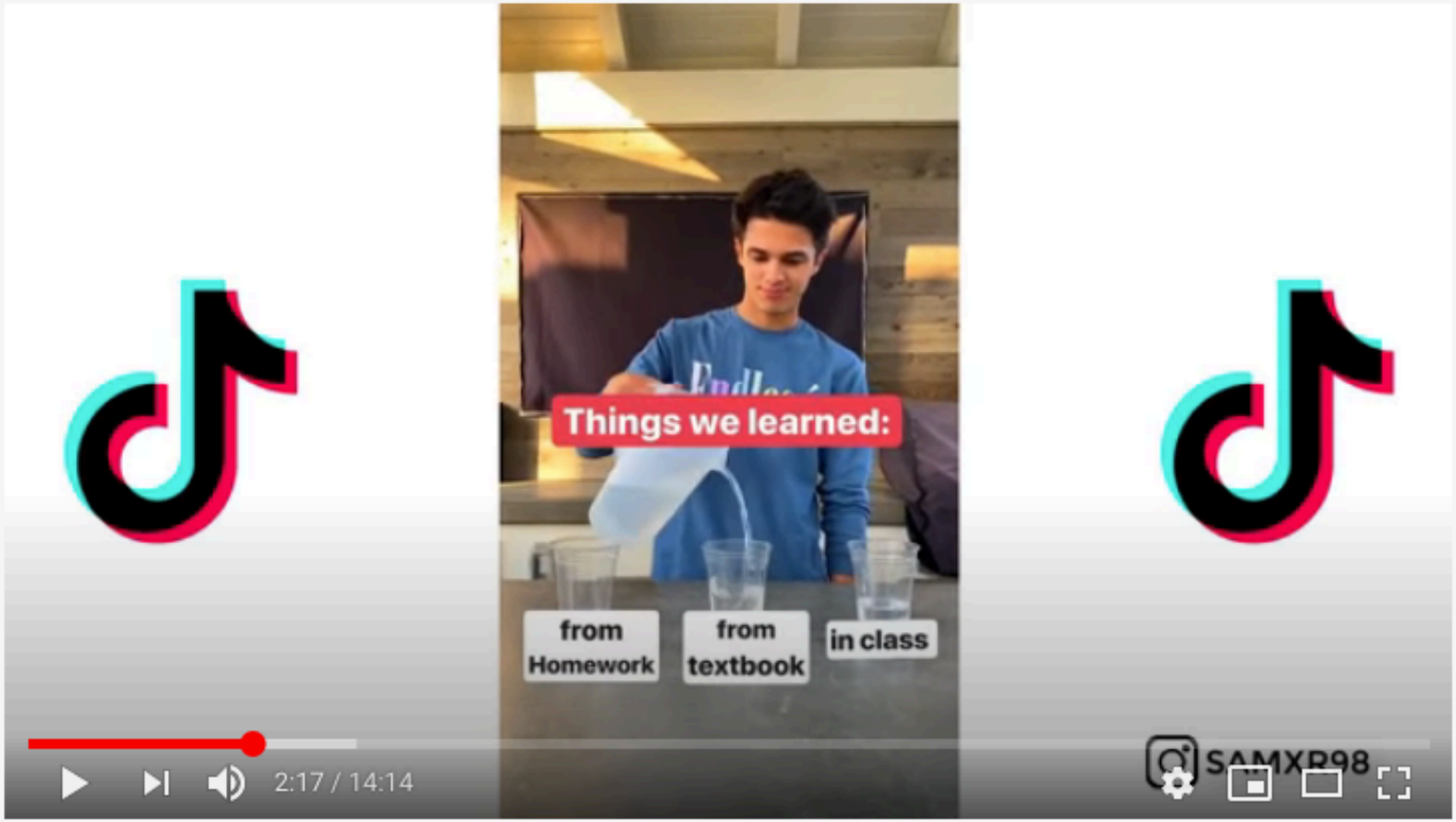
- Non sono concepiti e realizzati con criteri scientifici
- I dati contenuti in questi archivi non sono quelli che un sociologo o un geografo sceglierebbero per le loro indagini
- I dati raccolti cambiano di tipologia, formato, caratteristiche in continuazione, e a volte in maniera improvvisa
- I dati sono passibili di manipolazione, a volte accidentale, a volte dolosa
- I comportamenti sociali di interesse sono spesso divisi su diversi archivi di dati, senza modi pratici e rigorosi di combinarli (ad esempio, quasi tutte le ricerche su dati telefonici sono basate su dati di un singolo operatore)

Big Data

- L’aggettivo “big”, popolarizzato da un rapporto della società di consulenza McKinsey, si riferisce a:
  - il volume dei dati
  - la velocità con cui vengono prodotti
  - la varietà che li caratterizza
- Tutte caratteristiche che richiedono un nuovo insieme di strumenti per l’elaborazione dei dati
- Sul contenuto non c’è restrizione: si spazia da dati astronomici a versioni digitali di biblioteche

Fonti di dati

- Fare un elenco di possibili punti di generazione di big data è molto difficile, non solo per la loro quantità ma anche per la loro continua evoluzione (ad esempio, il video social network Tik Tok è una fonte neonata)
- È però possibile e utile classificare le fonti in 3 categorie



The Best TikTok Compilation of April 2020

868,947 views • Apr 2, 2020

5.2K 559 SHARE SAVE ...

Categorie di fonti di dati digitali

- Vita digitale
  - acquisizione di comportamenti sociali che sono mediati digitalmente
- Tracce digitali
  - prodotti di “scarto” dell’organizzazione digitale, che a loro volta possono formare un archivio
- Vita digitalizzata
  - spostamento di comportamenti intrinsecamente fisici* verso una forma digitale

*c’è chi li chiama “analogici”, sbagliando.

Vita digitale 1/2

- Non siamo (ancora?) esseri digitali che vivono in un mondo virtuale, ma una parte crescente della nostra vita è mediata in maniera intrinseca da mezzi digitali
- I comportamenti su Twitter, Facebook, e Wikipedia sono tutti online
- Si possono riferire a eventi nel mondo fisico, ma comportamenti come twittare sono intrinsecamente digitali

Donald J. Trump @realDonaldTrump · Apr 21

96% Approval Rating in the Republican Party. Thank you! This must also mean that, most importantly, we are doing a good (great) job in the handling of the Pandemic.

35.1K 32.7K 198.5K

Vita digitale 2/2

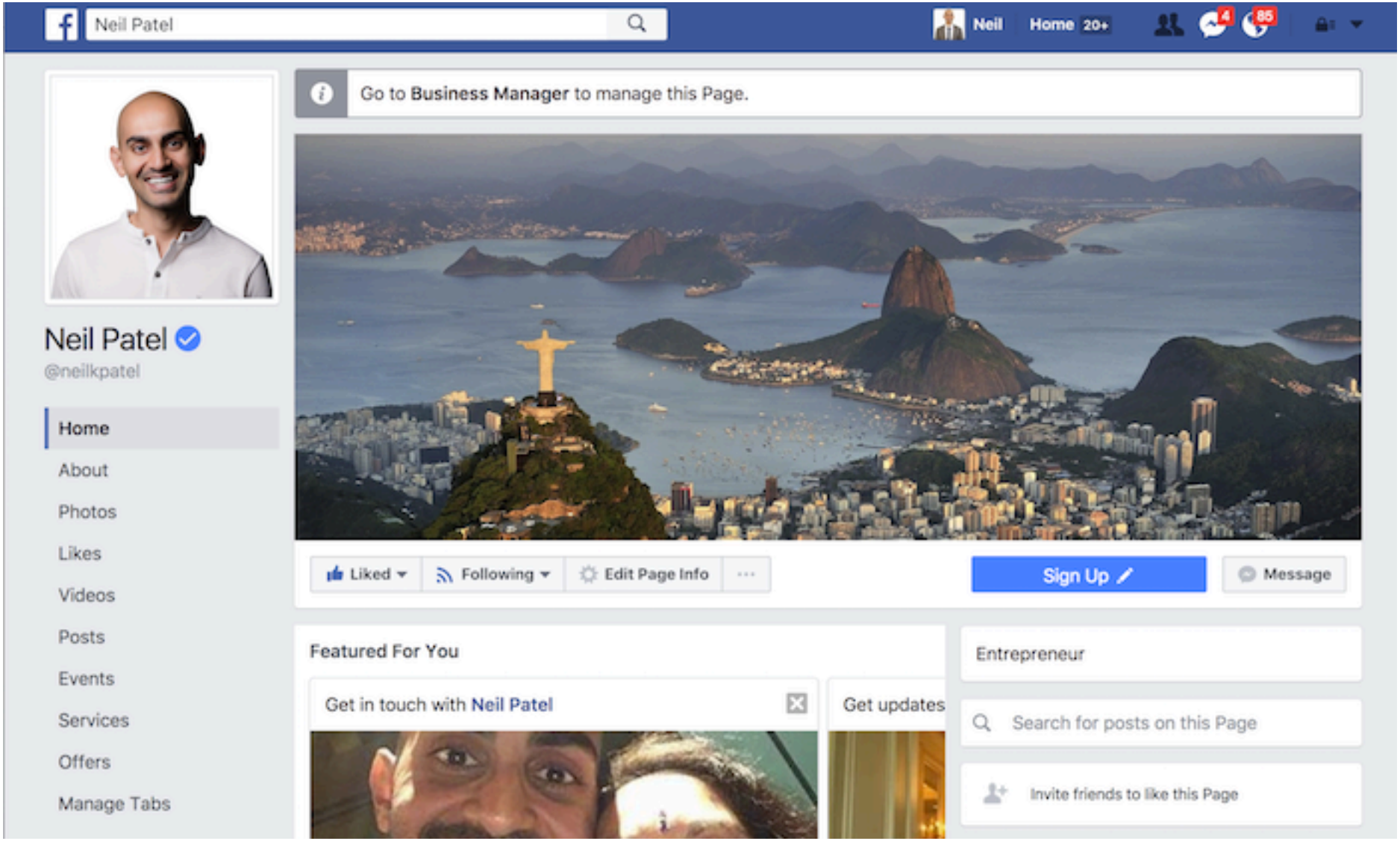
- Tipicamente, i comportamenti su queste piattaforme sono acquisiti dai proprietari delle piattaforme, perché il loro modello di business si basa sulle inferenze fatte su questi dati (ad esempio per pubblicità personalizzata)
- Inoltre, terzi possono accedere ai dati su queste piattaforme:
  - Facebook permette agli utenti di fare il download di porzioni dei loro dati
  - l'intera storia delle modifiche delle pagine di Wikipedia può essere scaricata per essere analizzata
  - Google permette parziale accesso ai dati sul volume delle ricerche
  - i dati di Twitter sono i più usati dai ricercatori perché i più accessibili

Due interpretazioni

- Ci sono due modi diversi di interpretare i dati che provengono dalle piattaforme digitali, basati su come si considerano tali piattaforme:
  - esse sono microcosmi generalizzabili della società
  - esse sono delle sfere distinte dove negli ultimi anni si è trasferita una parte significativa dell'esperienza umana

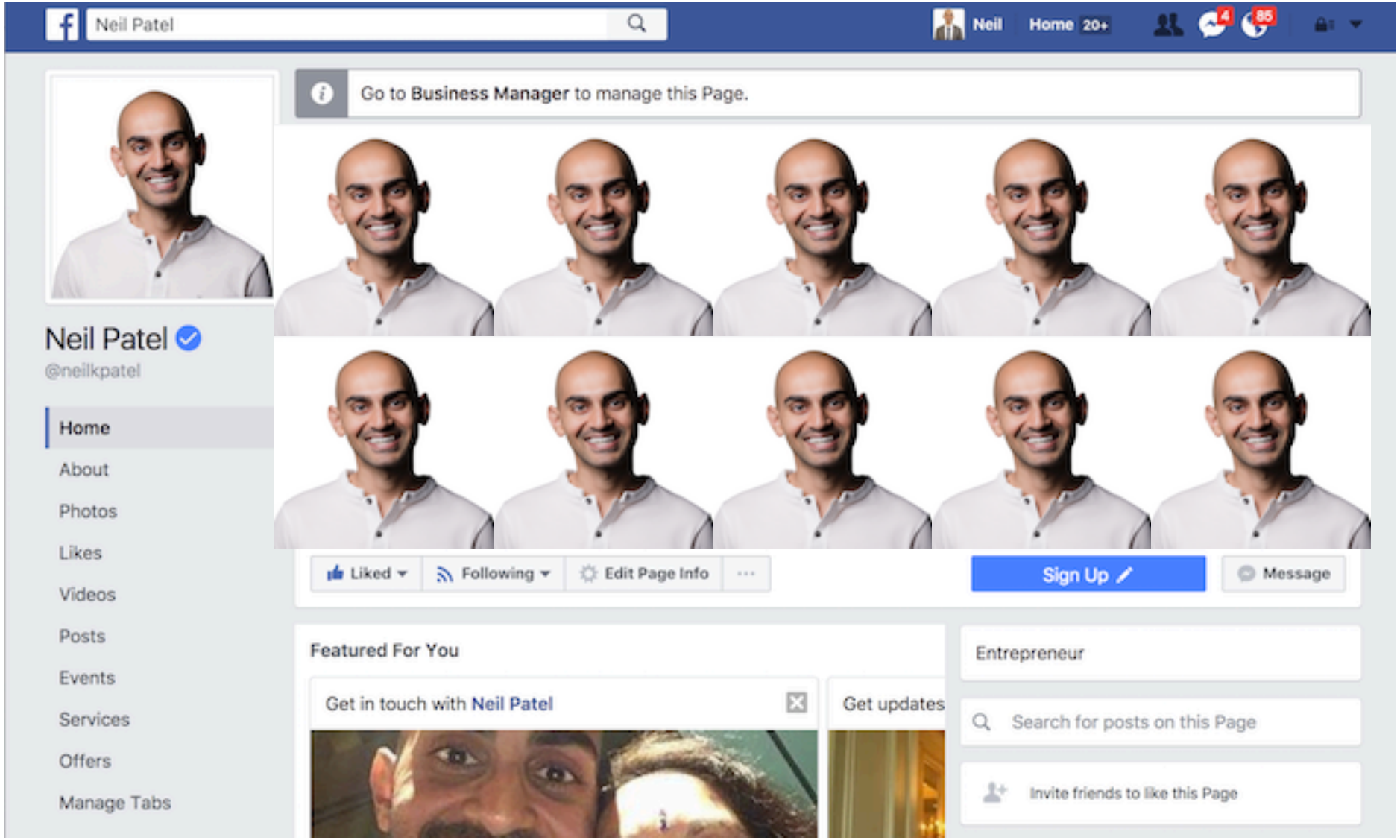
Microcosmi: esempi di studio

- Studio delle email per verificare la teoria dei “social foci” nella formazione di reti sociali
- Analisi di Twitter per studiare la mobilitazione politica
- Studio di Facebook e Wikipedia per analizzare la diffusione di notizie e rumors
- Analisi di mercati online come Airbnb e Kickstarter per studiare pattern di discriminazione sociale
- Estrazione di dati da negozi online per tenere traccia dell’inflazione
- Analisi delle ricerche su Google per mappare le zone di maggiore diffusione dell’influenza stagionale


Neil Patel
Neil
Home 20+
Go to Business Manager to manage this Page.
Neil Patel
@neilkpatel
Home
About
Photos
Likes
Videos
Posts
Events
Services
Offers
Manage Tabs
Liked
Following
Edit Page Info
Sign Up
Message
Featured For You
Get in touch with Neil Patel
Get updates
Entrepreneur
Search for posts on this Page
Invite friends to like this Page

Sfere distinte: esempi di studio

- Studio su Facebook per verificare l'ipotesi secondo cui la piattaforma crea o accentua un filtro per l'informazione attorno ai suoi utenti, in modo che le persone finiscano per vedere solo contenuti che siano ideologicamente compatibili con le loro convinzioni
- Questo tipo di analisi parte dal presupposto che il modo di fruire delle notizie su una piattaforma digitale sia nuovo e diverso rispetto a quanto tradizionalmente presente in società

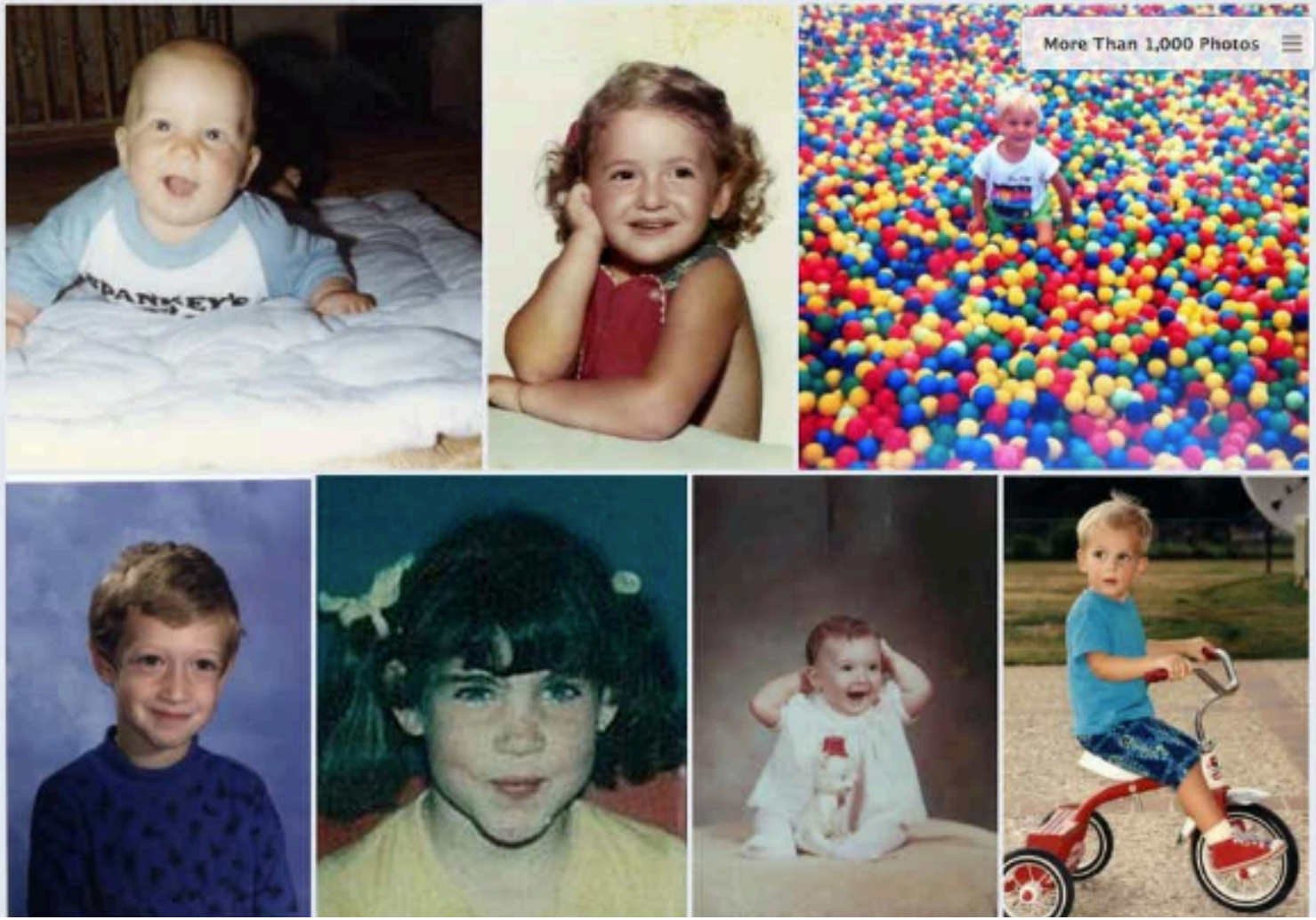

Neil Patel
Neil
Home 20+
Go to Business Manager to manage this Page.
Neil Patel
@neilkpatel
Home
About
Photos
Likes
Videos
Posts
Events
Services
Offers
Manage Tabs
Liked
Following
Edit Page Info
Sign Up
Message
Featured For You
Get in touch with Neil Patel
Get updates
Entrepreneur
Search for posts on this Page
Invite friends to like this Page

Caveat per entrambe le interpretazioni

- Se anche vediamo le piattaforme digitali come uno specchio della società, registrare quanto avviene su Internet è tutt'altro che banale, e generalmente queste raccolte sono istantanee di particolari momenti di particolari piattaforme
- Inoltre, la corrispondenza tra quanto avviene su Internet e i fenomeni di interesse sociologico e geografico potrebbe essere tenue: non tutti gli amici sono amici su Facebook, non tutti gli amici su Facebook sono amici

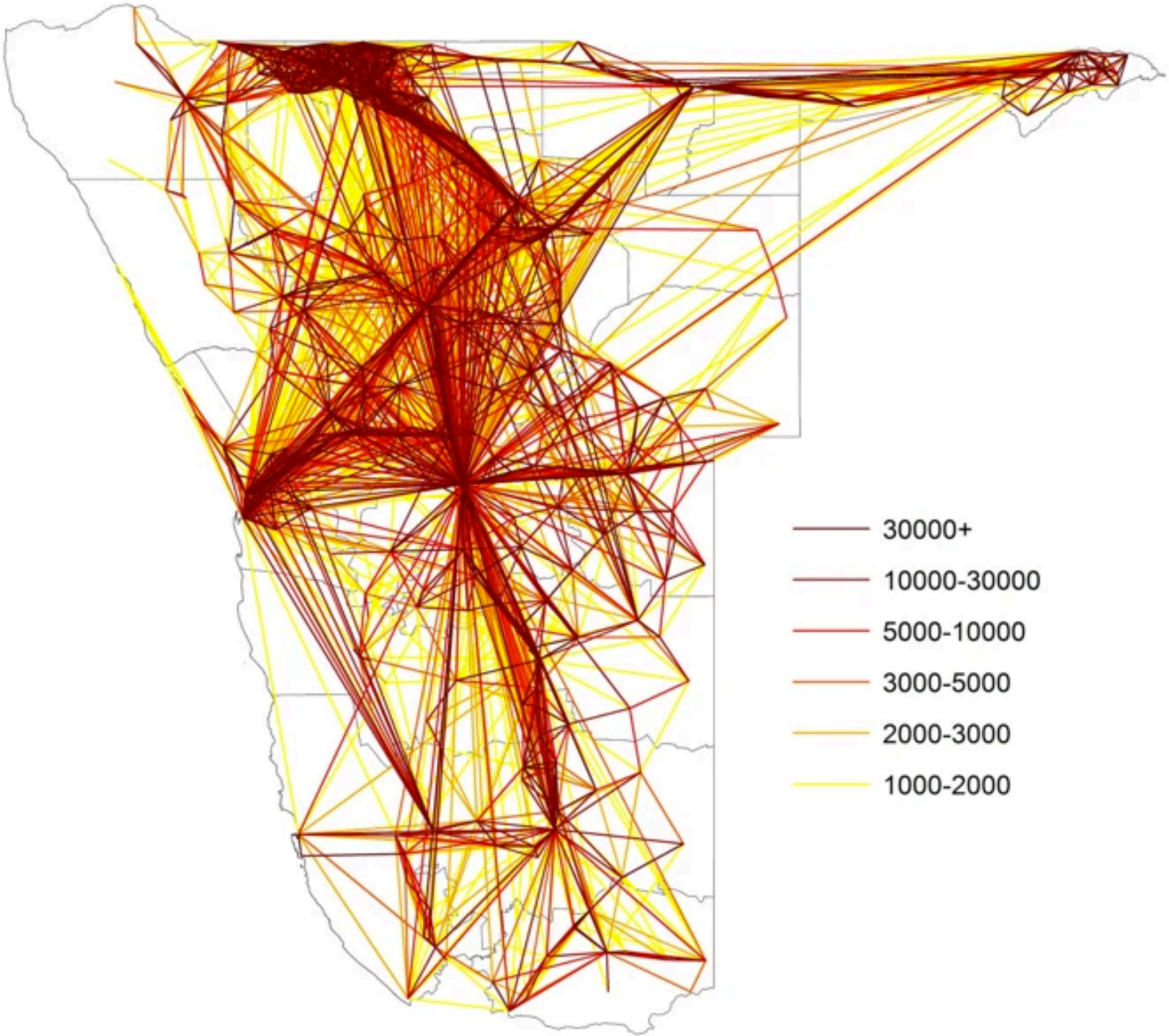
Photos of my friends before 1999

Home



Tracce digitali

- Le tracce digitali sono un effetto della vita digitale, ma ne sono ben distinte:
  - la vita digitale è fatta di azioni digitali (ad es. twittare)
  - le tracce sono solo la registrazione di tali azioni, ma non le azioni stesse
- La complessa organizzazione delle piattaforme digitali crea un output continuo di tracce, note anche come metadati
- Esempio tipico di metadati sono quelli legati alle chiamate telefoniche:
  - identificatore del chiamante
  - identificatore del chiamato
  - identificatori delle torri cellulari usate durante la chiamata
- Questo tipo di tracce è stato usato in passato per fare analisi di
  - forza di legami interpersonali
  - livello di disoccupazione in specifiche aree
  - modelli di diffusione della malaria
- Dati governativi come registri di voto, donazioni ai partiti e dati fiscali sono altri esempi di tracce digitali


30000+
10000-30000
5000-10000
3000-5000
2000-3000
1000-2000

Vita digitalizzata

- Per vita digitalizzata si intende registrare in forma digitale la parte non digitale della vita
- Esempi di questo tipo di digitalizzazione:
  - misurare la prossimità degli individui programmando gli smartphone per riconoscere apparati con Bluetooth nelle vicinanze
  - registrazione delle interazioni umane su supporto digitale attraverso telecamere sparse per la città o negli edifici
  - digitalizzazione mediante scansione di oggetti informativi dell'era pre-digitale, come libri e giornali

The Folger SHAKESPEARE

PLAYS & POEMS ABOUT DOWNLOAD CITE LISTEN

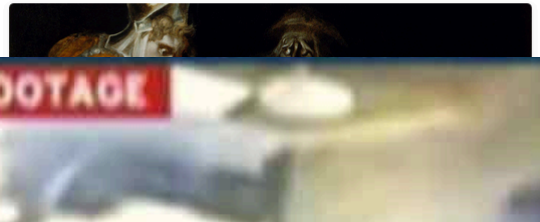
Search Folger Shakespeare

Welcome to The Folger Shakespeare

Enjoy Shakespeare's plays, sonnets, and poems for free from the Folger Shakespeare Library! While you're here, you can also read more about Shakespeare's language, life, and the world he knew.

NEW: Listen to free audio recordings of seven complete plays through July 1, 2020.

Explore Shakespeare's Plays



Opportunità e vulnerabilità

- I big data della vita digitale, delle tracce digitali e della vita digitalizzata sembrano offrire enormi possibilità di analisi per sociologi e geografi
- È però interessante notare come ogni esperimento, anche tra quelli di maggiore successo, mostri contemporaneamente i vantaggi e gli svantaggi di questo tipo di approccio

The Copenhagen Network Study 1/3

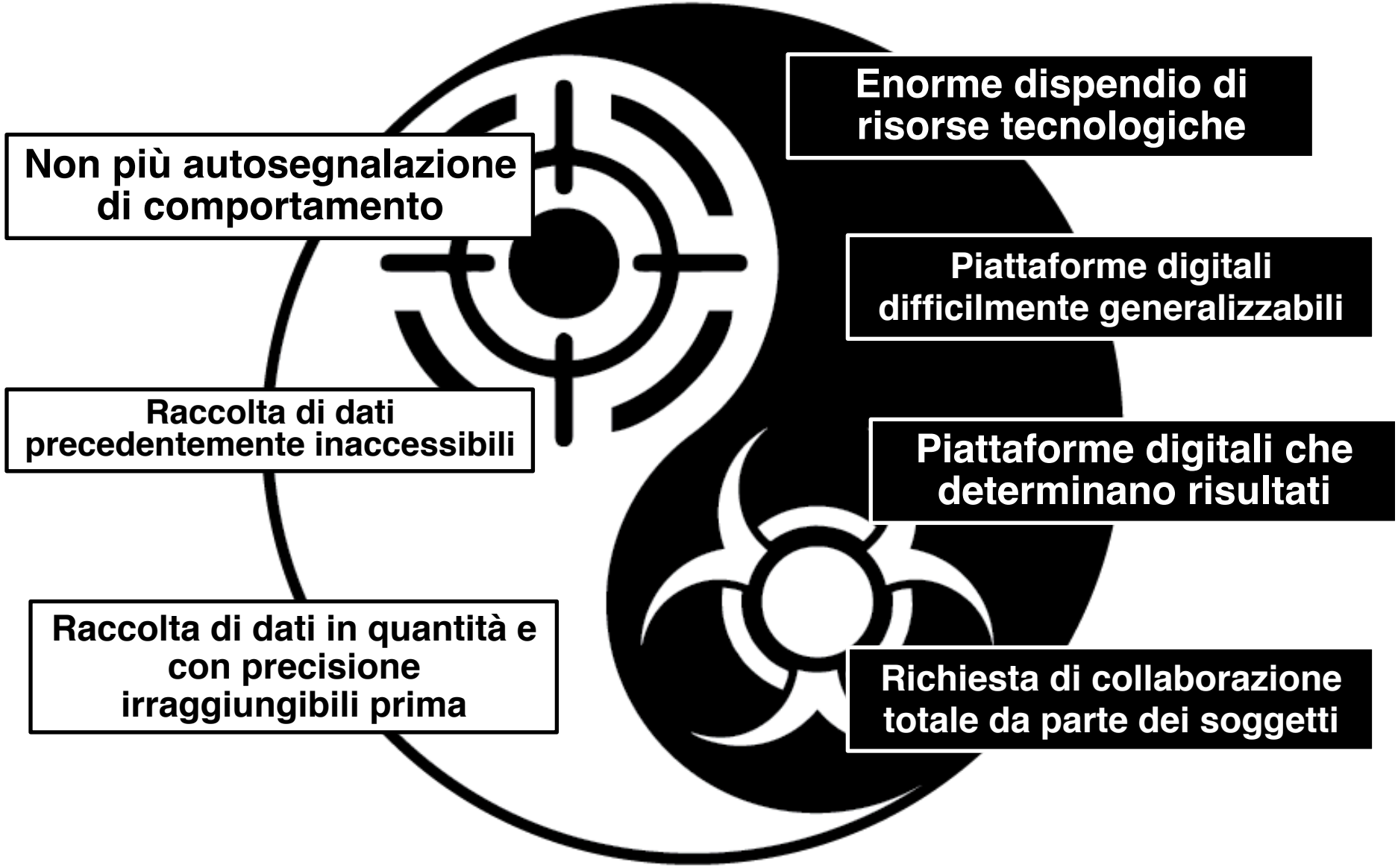
- I ricercatori hanno dato 1000 cellulari alle matricole del Politecnico di Danimarca a Copenhagen nel 2012 e nel 2013
- Hanno usato i cellulari per inferire prossimità via Bluetooth, prossimità geografica via GPS, e interazioni per mezzo di chiamate e messaggi
- Hanno combinato questi dati con i dati degli account Facebook degli studenti, la loro vicinanza ai router e la osservazioni qualitative di un antropologo sul campo

The Copenhagen Network Study 2/3

- Risultati dell'analisi dei dati:
  - quasi tutto il call network (rete di amici che si telefonano) è individuato dai dati di prossimità Bluetooth
  - l'80% delle amicizie su Facebook è anch'esso individuato dalla prossimità Bluetooth
  - solo il 20% delle amicizie su Facebook viene individuato dalle chiamate telefoniche

The Copenhagen Network Study 3/3

- Interpretazione dei risultati:
  - le misure di comportamenti non sono intercambiabili: sistemi digitali diversi portano a reti sociali diverse con differenti caratteristiche
  - non c'è una singola rete sociale per tutti, ma una serie di reti che cambiano a seconda delle organizzazioni e tecnologie che l'individuo usa per formare e mantenere relazioni
  - è probabile che la scelta dei ricercatori su quali sistemi usare per raccogliere dati e su come integrarli influenzi i risultati degli studi


Enorme dispendio di risorse tecnologiche
Piattaforme digitali difficilmente generalizzabili
Piattaforme digitali che determinano risultati
Richiesta di collaborazione totale da parte dei soggetti
Raccolta di dati in quantità e con precisione irraggiungibili prima
Raccolta di dati precedentemente inaccessibili
Non più autosegnalazione di comportamento

Proteste a confronto 1/4

- Ricercatori hanno usato dati estratti da Twitter per confrontare i pattern di mobilitazione tra:
  - la protesta al Taksim Gezi Park (Istanbul, Turchia) nel 2013 durante la Primavera Araba
  - le proteste di Occupy Wall Street (New York, USA) e quelle degli Indignados (Madrid, Spagna) nel 2012
- L'indagine ha lo scopo di capire perché la protesta in Turchia ha avuto successo mentre le altre due no

Proteste a confronto 2/4

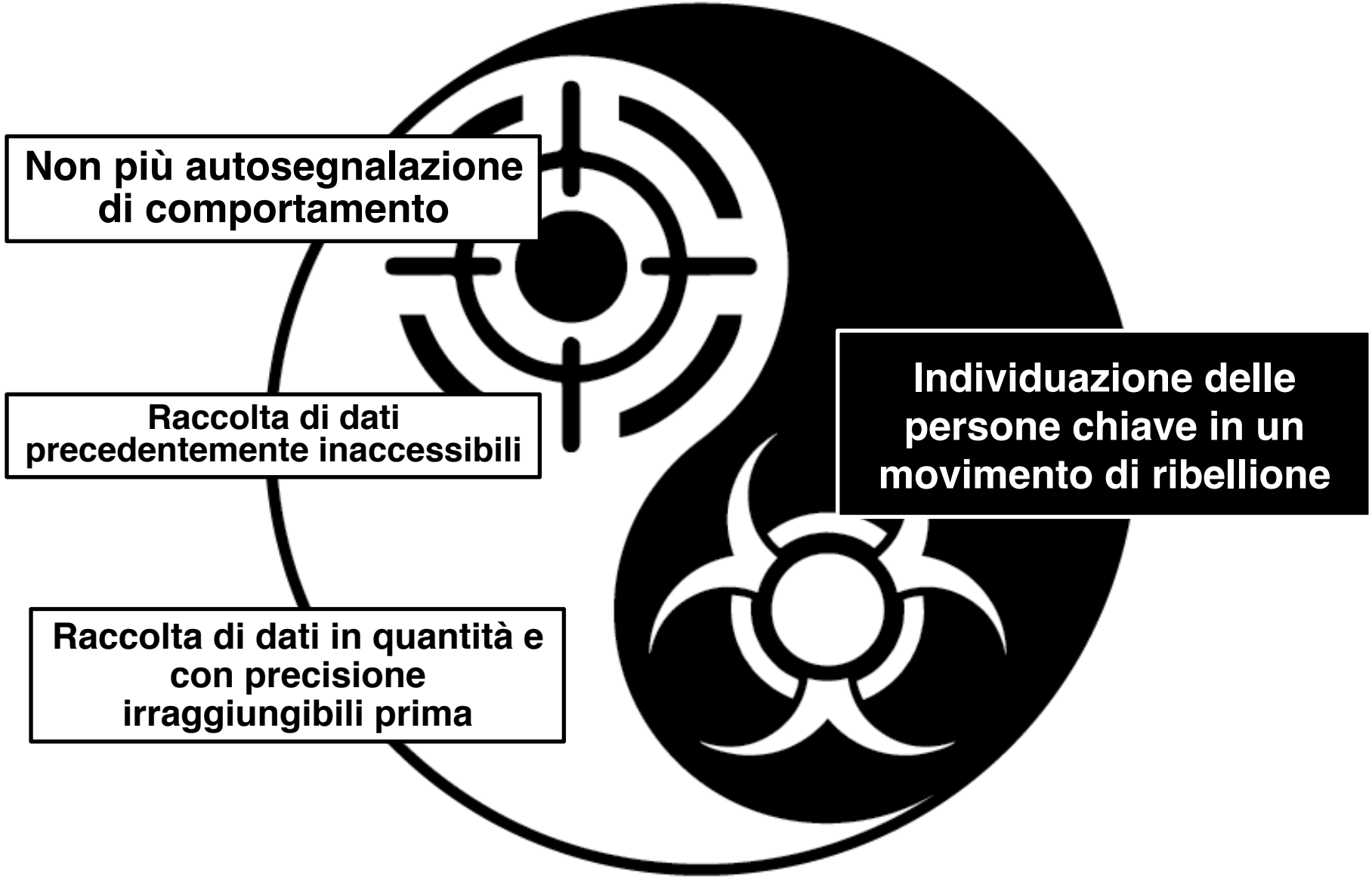
- Attraverso l'interfaccia di programmazione fornita da Twitter, i ricercatori hanno cercato parole chiave e hashtag per raccogliere campioni di tweet dai tre movimenti
- Hanno ricostruito network di mobilitazione tra utenti che hanno pubblicato messaggi contenenti tali parole chiave e utenti che hanno ripubblicato (con retweet) tali messaggi

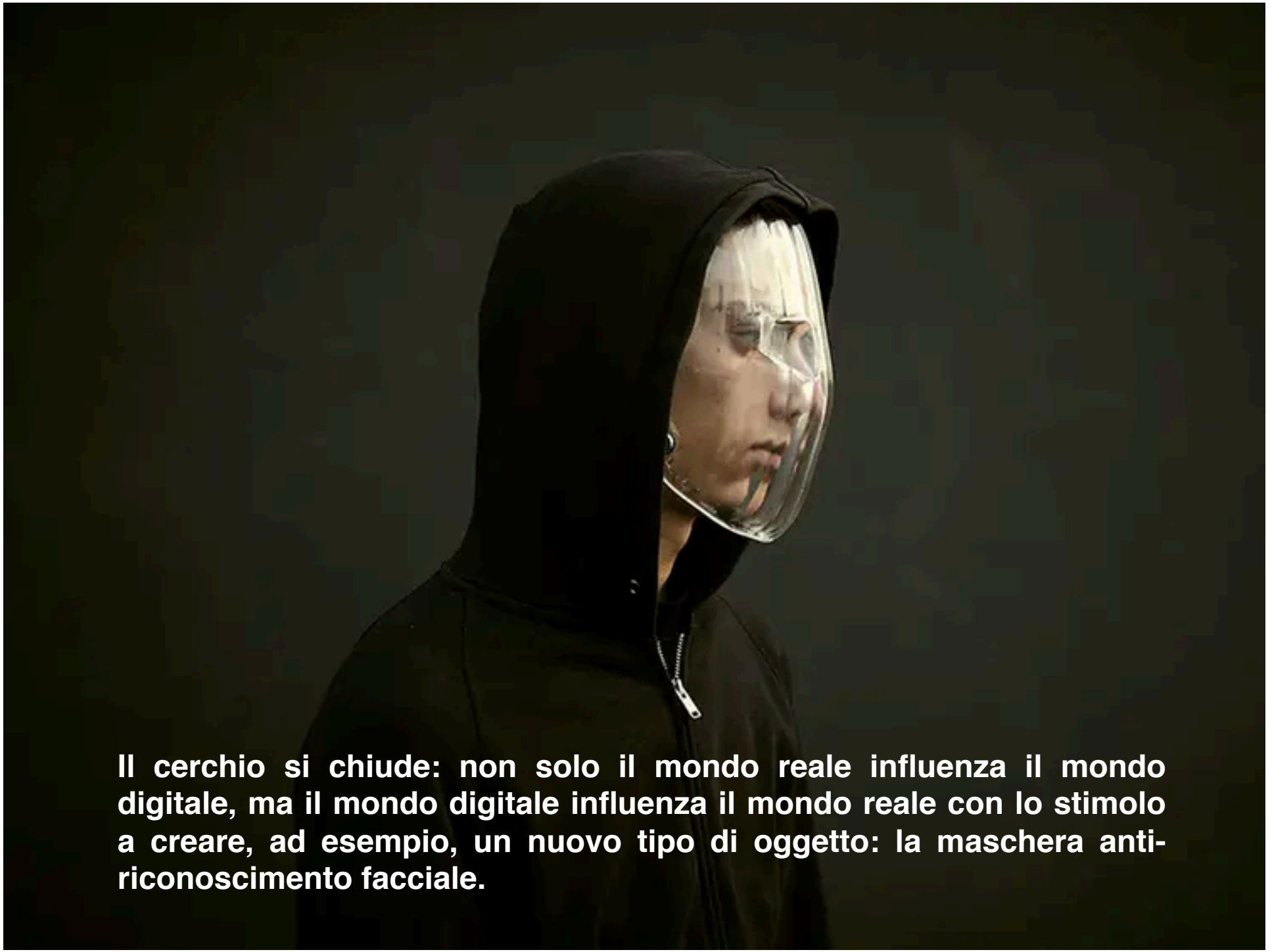
Proteste a confronto 3/4

- I ricercatori hanno scoperto che:
  - i membri periferici del network della protesta al Gezi Park hanno mobilitato più persone che i membri del nucleo dei network di Occupy o degli Indignados
  - in network di controllo non politici (persone che su Twitter scrivevano degli Oscar 2014 oppure che scrivevano sul salario minimo negli USA), non hanno riscontrato pattern di interazione tra periferia e centro simili

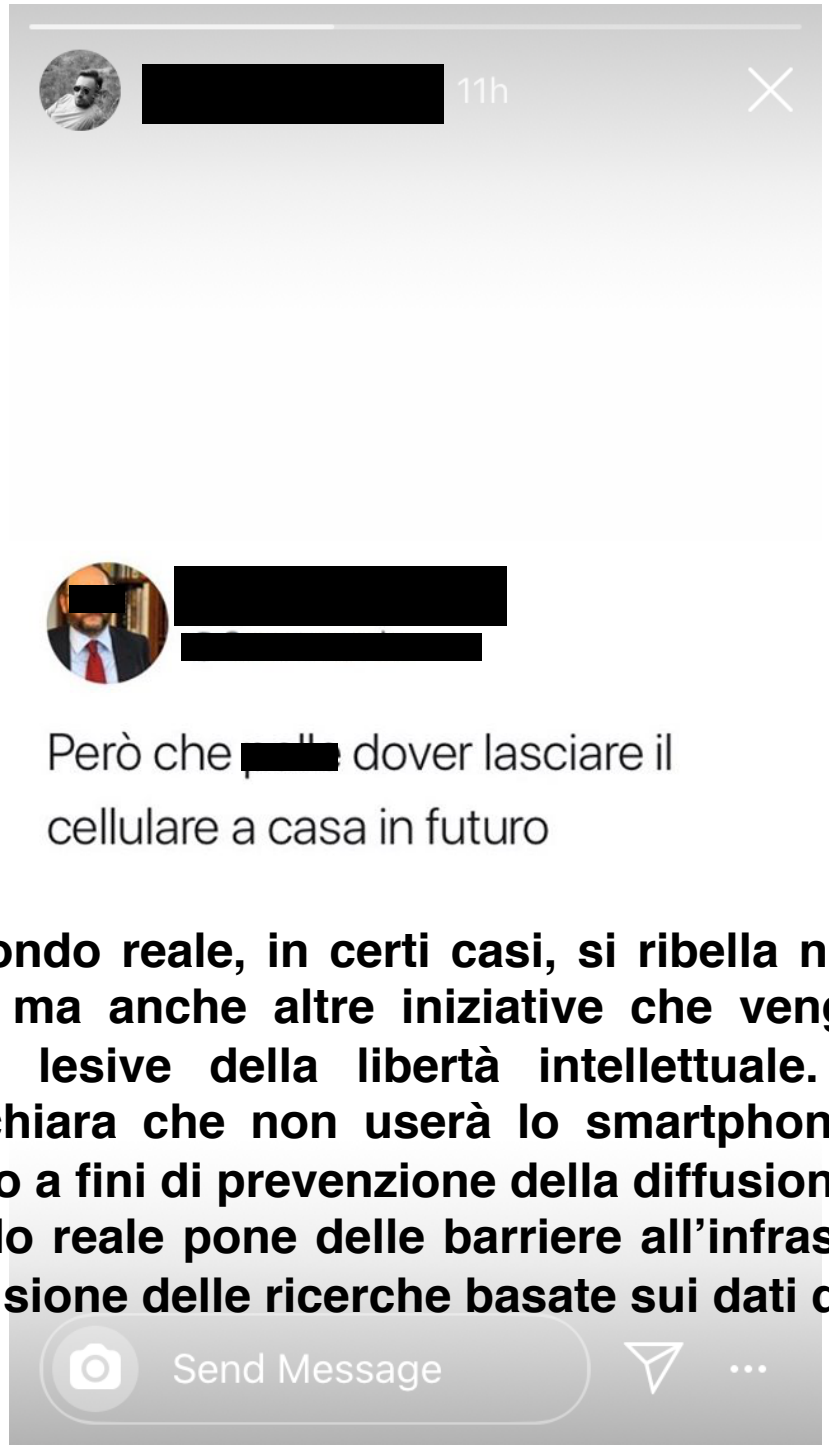
Proteste a confronto 4/4

- Interpretazione dei risultati:
  - in una mobilitazione di successo, i membri periferici del network sociale fanno ulteriore proselitismo su utenti ancora più periferici
  - i mobilizzati devono mobilizzare ulteriormente
  - in linea con le teorie attuali dei movimenti sociali, le risorse e la capacità organizzativa non sono sufficienti a garantire una mobilitazione significativa
  - la diffusione della rete sociale della protesta può essere una discriminante tra successo e fallimento


Non più autosegnalazione di comportamento
Raccolta di dati precedentemente inaccessibili
Raccolta di dati in quantità e con precisione irraggiungibili prima
Individuazione delle persone chiave in un movimento di ribellione



Il cerchio si chiude: non solo il mondo reale influenza il mondo digitale, ma il mondo digitale influenza il mondo reale con lo stimolo a creare, ad esempio, un nuovo tipo di oggetto: la maschera anti-riconoscimento facciale.



Attenzione: il mondo reale, in certi casi, si ribella non solo contro i regimi totalitari, ma anche altre iniziative che vengono comunque concepite come lesive della libertà intellettuale. In figura: una persona che dichiara che non userà lo smartphone per evitare di essere localizzato a fini di prevenzione della diffusione del Covid-19. Laddove il mondo reale pone delle barriere all'infrastruttura digitale, lì si ferma l'estensione delle ricerche basate sui dati digitali.

Grazie per l’attenzione.

Bibliografia

- Manyika J, Chui M, Brown B, Bughin J, Dobbs R, et al. 2011. Big data: the next frontier for innovation, competition, and productivity. Rep., McKinsey Global Inst. http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation

- Gartner. 2011. Gartner says solving “big data” challenge involves more than just managing volumes of data. News Release, June 27.

- Manovich L. 2012. Trending: the promises and the challenges of big social data. In Debates in the Digital Humanities, Vol. 2, ed. MK Gold, pp. 460–75. Minneapolis, MN: Univ. Minn. Press

- Tufekci Z. 2014. Big questions for social media big data: representativeness, validity and other methodological pitfalls. arXiv: 1403.7400 [cs.SI]

- Bakshy E, Messing S, Adamic LA. 2015. Exposure to ideologically diverse news and opinion on Facebook. Science 348(6239): 1130–32

- Toole JL, Lin Y-R, Muehlegger E, Shoag D, González MC, Lazer D. 2015. Tracking employment shocks using mobile phone data. J. R. Soc. Interface 12(107): 20150185

- Wesolowski A, Eagle N, Tatem AJ, Smith DL, Noor AM, et al. 2012. Quantifying the impact of human mobility on malaria. Science. 338(6104): 267–70

- Stopczynski A, Pietri R, Pentland A, Lazer D, Lehmann S. 2014a. Privacy in sensor-driven human data collection: a guide for practitioners. arXiv: 1403.5299 [cs.CY]

- Barberá P, Wang N, Bonneau R, Jost JT, Nagler J, et al. 2015. The critical periphery in the growth of social protests. PLOS ONE 10(11): 1–15