Economics for Data Science

Google Trends: limiti e potenzialità

F. Cordaro, L. Mandelli, S. Offredi

11 Giugno 2018

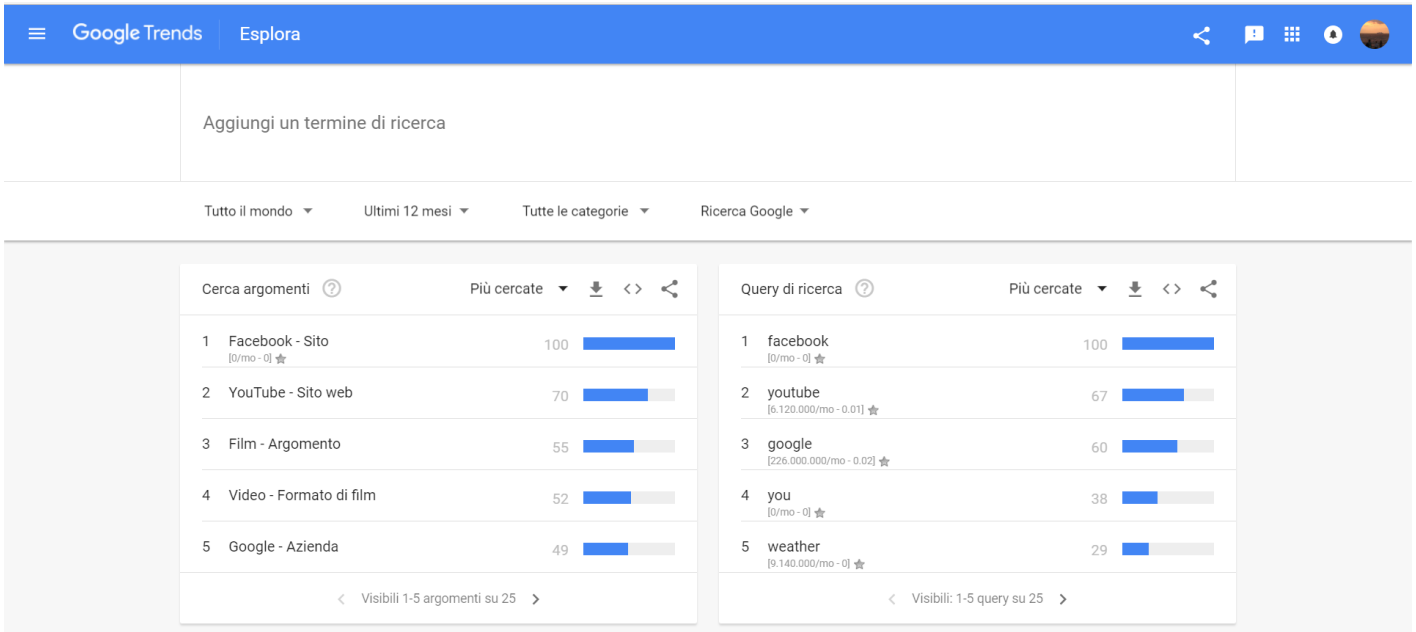
Introduzione

Scopo

Google Trends è uno strumento di *Google* ultimamente spesso usato per fare nowcasting. Abbiamo deciso di mostrare cosa sia, compresi alcuni suoi limiti, prima utilizzando un esempio *naive* supponendo di voler creare un dataset per fare nowcasting sugli incassi dei cinema in Italia, successivamente presentando due articoli che descrivono il caso reale di *Google Flu Trends*.

1. Cos'è *Google Trends*?
2. Vantaggi dei dati da *Google Trends*
3. Limiti dei dati da *Google Trends*
   - Dati riscalati
   - Dati normalizzati
   - Quali *query* usare?
   - Random sampling
4. Il caso reale di *GFT*

Google trends - Descrizione



Schermata *esplora* di Google Trends

Cos'è?

Google Trends è uno strumento di Google disponibile all'indirizzo https://trends.google.it/trends che fornisce un indice che rappresenta il numero di ricerche sul web effettuate per una particolare *query* relativamente al numero totale delle ricerche in una certa area geografica a partire da gennaio 2004.

Costruzione del dataset

Vogliamo costruire un dataset che potenzialmente ci permetta di fare nowcasting sugli incassi dei cinema in Italia.

Dataset per previsione incassi in Italia

- In Italia identifichiamo i 4 film che hanno incassato di più fino alla settimana precedente a quella attuale
- Ricerchiamo su *Google Trends* i titoli con i quali i film sono stati pubblicizzati in Italia e ne osserviamo la media delle ricerche nei giorni da giovedì a sabato per ogni settimana, in quanto i dati relativi agli incassi sono relativi a tali giorni
- Per la settimana in corso sappiamo in quante sale in Italia sono proiettati i film in quel momento al cinema, e questo ci fa identificare i 4 film più "importanti": di essi non abbiamo gli incassi ma abbiamo l'indice relativo alle ricerche su *Google* tramite *Google Trends*

Perché 4 film?

Tale scelta è stata fatta poiché 4 era il numero minimo di film per ottenere una percentuale di incassi per quei film in una certa settimana pari almeno al 40% del totale settimanale. Ci aspettiamo che un eventuale modello migliori inserendo tutti i film presenti in un istante temporale al cinema.

Google Trends - Descrizione

In quale dei servizi di Google è stata effettuata la ricerca?

- Ricerca Google
- Google immagini
- Google news
- Google shopping
- Ricerche su Youtube

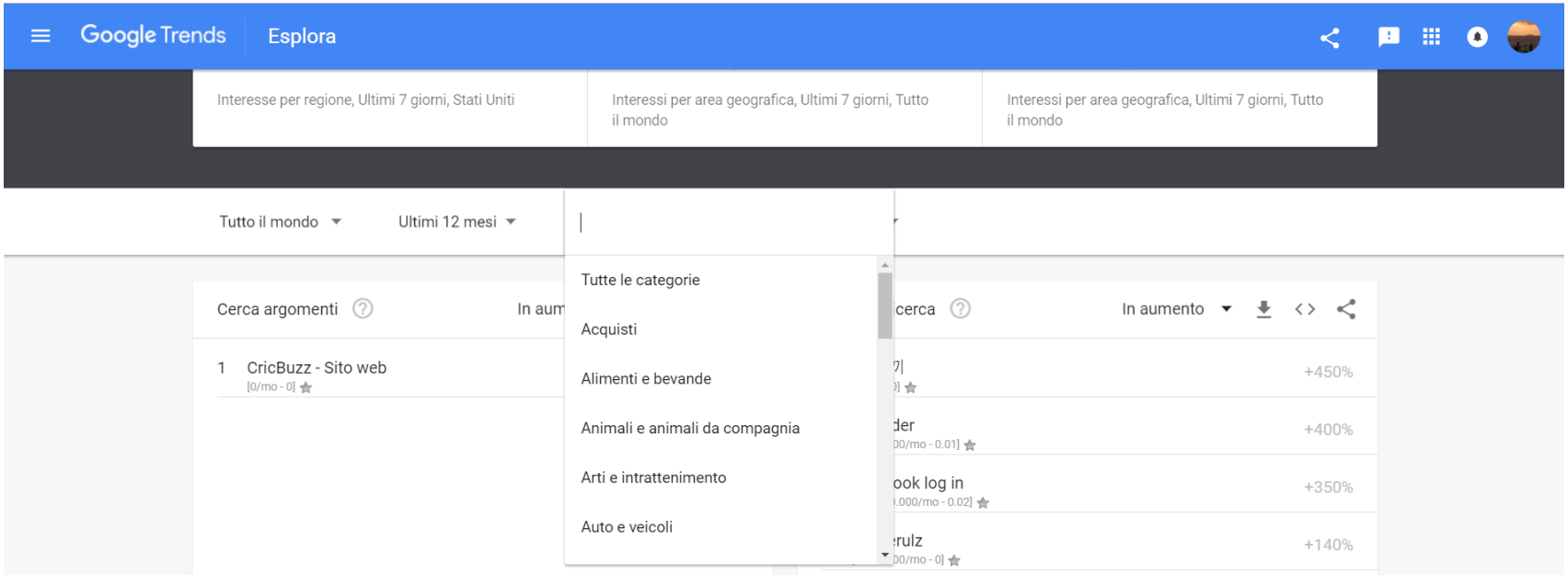
Dove sono avvenute le ricerche?

Si possono osservare le ricerche effettuate in tutto il mondo, in uno specifico stato o in un'area metropolitana (ad es. New York). Inoltre *Google trends* ci permette, ad esempio nel caso dell'Italia, di osservare anche l'interesse suddiviso per regione o nelle principali città.

Filtro arco temporale

- Ultimi 12 mesi
- Ultima ora
- Ultime 4 ore
- Ultimi 7 giorni
- Ultimi 30 giorni
- Ultimi 90 giorni
- Ultimi 5 anni
- 2004 - Presente
- Intervallo di tempo personalizzato

Google Trends - Descrizione



Filtri per categorie

Filtro per categorie

Si possono filtrare i risultati in base a una categoria specifica nella quale cercare quella parola, in quanto *Google* suddivide le query in categorie

Vantaggi dei dati da *Google Trends*

I dati di *Google Trends*

- sono gratuitamente accessibili da tutti
- vengono raccolti e aggiornati quotidianamente
- vengono raccolti solo se il numero di ricerche supera una certa soglia
- sono tali che le query ripetute da un singolo utente/IP in un breve periodo di tempo siano eliminate
- vengono eliminati i caratteri speciali, come ad esempio gli apostrofi
- provengono da tutto il mondo
- sono disponibili dal 4 gennaio 2004
- vengono suddivisi in categorie

Vantaggi (Li, 2016)

1. potrebbero essere in grado di misurare alcune attività economiche che misure o metodi tradizionali non riescono a catturare
2. riflettono gli interessi delle persone, quantità altrimenti difficile da misurare

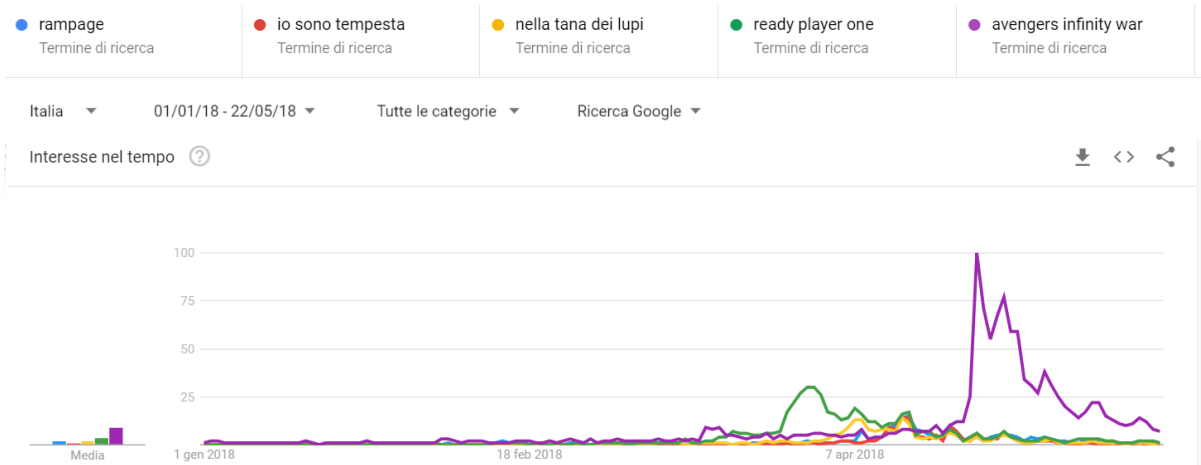
PROBLEMA 1: Dati riscalati

Dati riscalati

I valori sono riscalati in un range 0-100, in cui il numero effettivo di ricerche corrispondente al valore 100 non è noto per motivi di privacy.

Parziale soluzione

Identifichiamo la ricerca effettuata maggiormente da gennaio ad oggi e la affianchiamo alle altre ricerche in modo tale da mantenere una scala comune così da rendere confrontabili i diversi indici. Tuttavia è solo parziale come soluzione perché se ad un certo punto tale valore massimo venisse superato, si dovrebbero riscalare tutti i dati.



Esempio ricerca dei 4 film più importanti del mese con "avengers infinity war" mantenuto come termine di paragone

PROBLEMA 2: Dati normalizzati

Dati normalizzati

Il valore 100 non è attribuito al valore massimo delle ricerche, ma al valore massimo assunto dal rapporto tra il numero di ricerche della *query* e il numero totale delle ricerche.

Come avviene la normalizzazione dei dati?

Data una regione r e un giorno t, definiamo la proporzione del volume di ricerche di una certa query il giorno t nella regione r la quantità $V_{t,r}$, mentre il volume di ricerche totale è $T_{t,r}$. Vale quindi che

$$S_{t,r} = \frac{V_{t,r}}{T_{t,r}}$$

Valore settimanale dell'indice

Per calcolare il valore dell'indice settimanale *Google Trends* effettua la media dei 7 giorni (in particolare da domenica al sabato successivo).

$$S_{w,r} = \frac{1}{7} \sum_{t=domenica}^{sabato} \frac{V_{t,r}}{T_{t,r}}$$

La fonte: il *Support* di *Google Trends*

How Trends data is adjusted

Trends adjusts search data to make comparisons between terms easier. Search results are proportionate to the time and location of a query by the following process:

1. Each data point is divided by the total searches of the geography and time range it represents to compare relative popularity. Otherwise, places with the most search volume would always be ranked highest.
2. The resulting numbers are then scaled on a range of 0 to 100 based on a topic's proportion to all searches on all topics.
3. Different regions that show the same number of searches for a term don't always have the same total search volumes.

Da queste due osservazioni notiamo come i valori dell'indice di *Google Trends* non abbiano un immediato significato quantitativo (100 può corrispondere tanto a un milione di ricerche quanto a centomila), e sarebbe teoricamente possibile visualizzare un trend decrescente anche se il volume assoluto di ricerca per la *query* considerata stesse aumentando: una linea verso il basso indica che la *popolarità relativa* del termine di ricerca si sta riducendo, tuttavia non necessariamente il numero totale di ricerche ad esso corrispondente si è ridotto.

PROBLEMA 3: Quali *query* usare?

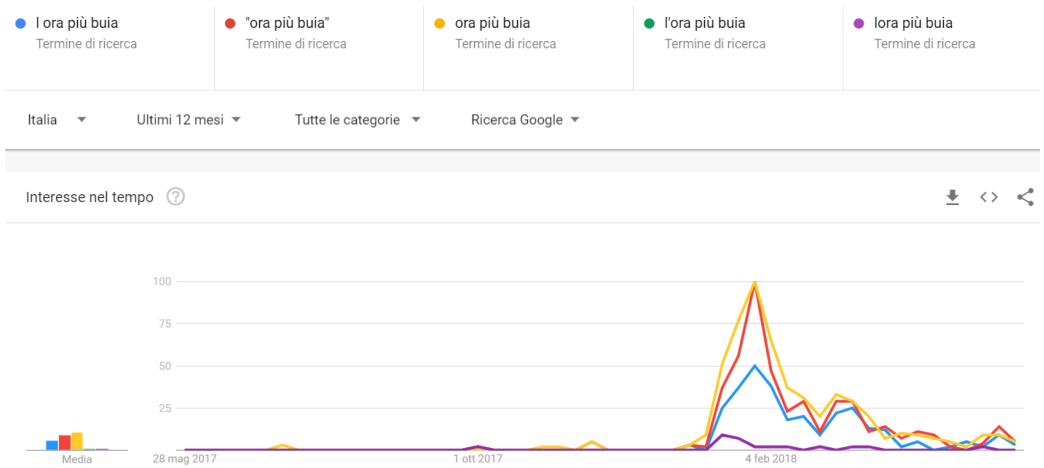
Questo problema è molto ampio e va analizzato caso per caso. Abbiamo deciso di riportare solo alcuni esempi naive di problemi riscontrati nel cercare una sola parola, scelta fatta a priori per semplicità; presenteremo poi come *Google* abbia affrontato questo problema tramite un metodo automatizzato e i problemi connessi.

Problemi emersi riguardanti le *query*

1. Diverse possibilità di cercare lo stesso film
2. Film su personaggi storici molto poco cercati: perché?
3. "Loro" e il filtro per categorie

Diverse possibilità di cercare lo stesso film

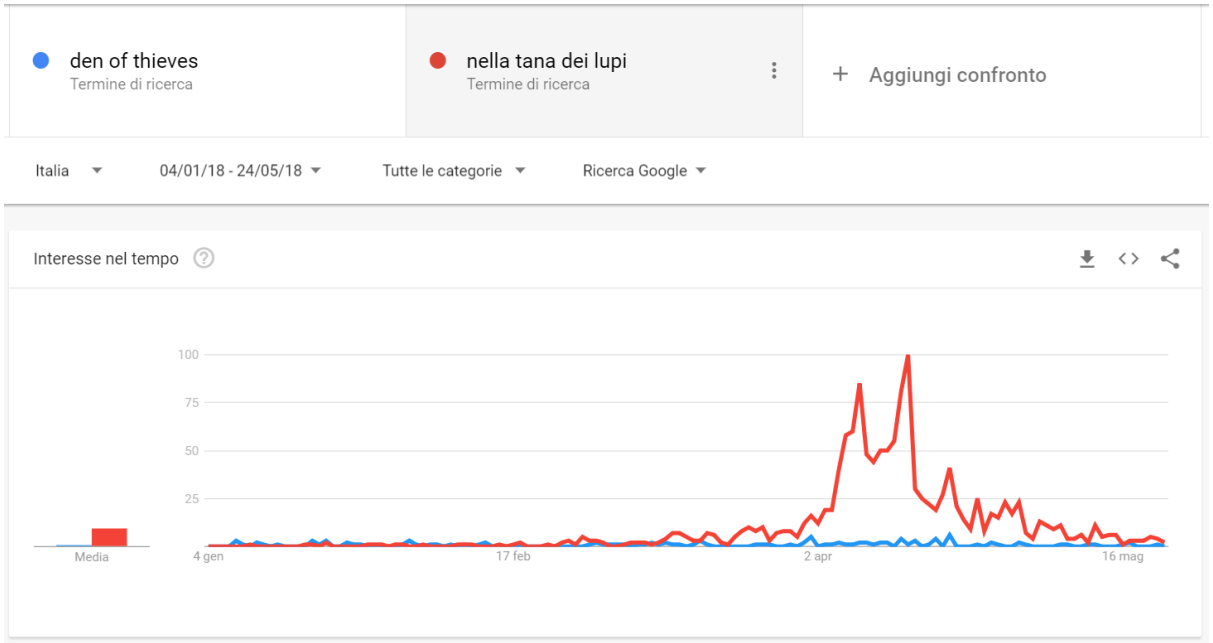
Prendiamo in considerazione il titolo del film "L'ora più buia". Osserviamo come *Google Trends*, come già detto, rimuova l'apostrofo, per cui le linee verde e viola coincidono. La ricerca in questione può essere effettuata in diversi modi, dove le virgolette indicano la sequenza di parole ricercata in un determinato ordine.



Diverse possibilità di cercare lo stesso film

Titolo originale VS titolo pubblicizzato in Italia

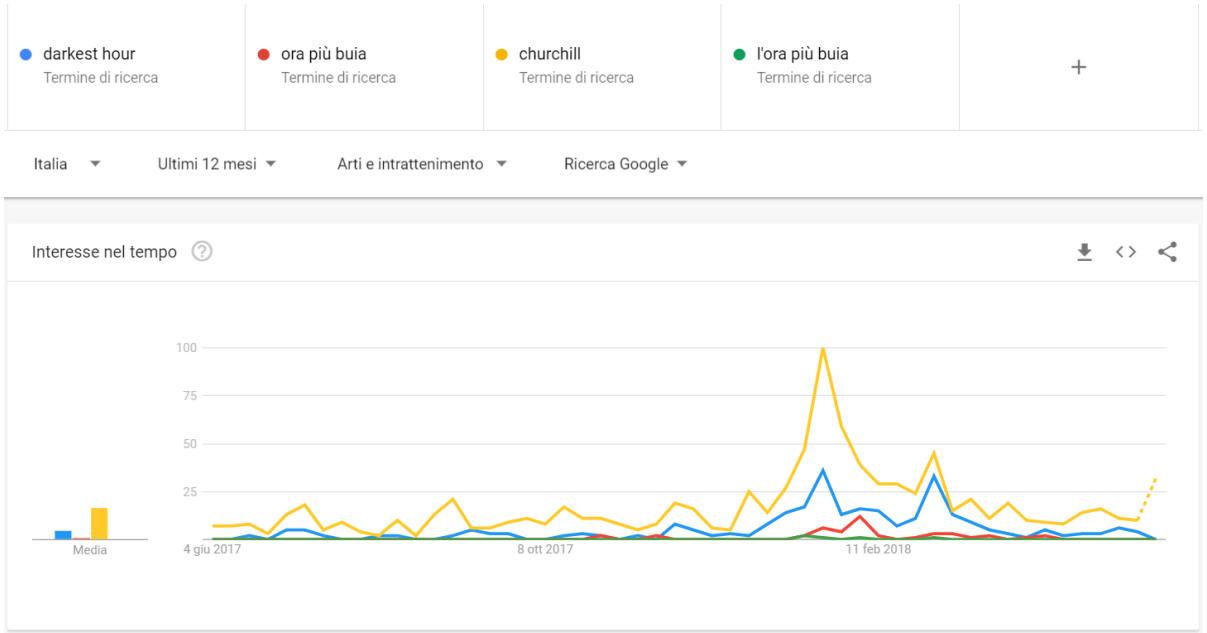
Osserviamo uno dei tanti esempi in cui si vede come sia più realistico, dal momento che stiamo costruendo un dataset per l'Italia, cercare il titolo del film così come è stato pubblicizzato in Italia piuttosto che il titolo originale.



Ricerca "nella tana dei lupi" VS "den of thieves"

Film su personaggi storici

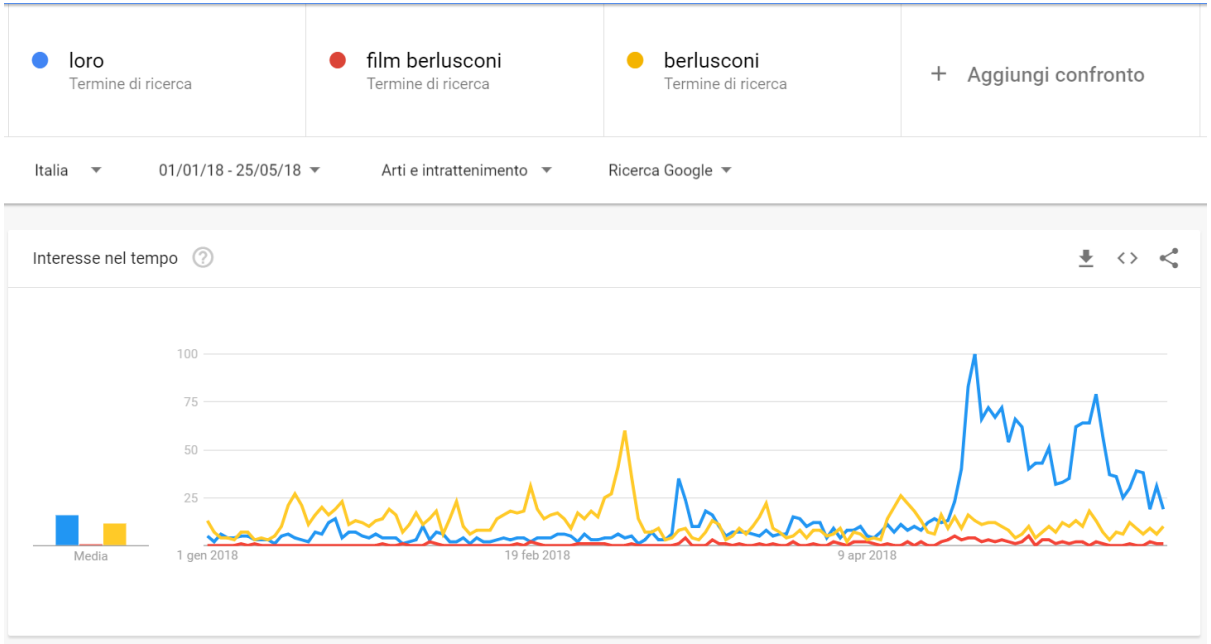
Notiamo come i film su personaggi storici siano molto poco cercati col loro titolo originale, ma molto più spesso col nome del personaggio in questione. Prendiamo, a titolo di esempio, il film "l'ora più buia", su Churchill. Osserviamo che, sebbene durante tutto l'anno vi siano più ricerche di Churchill, sebbene filtrate in "Arte e intrattenimento", c'è un picco molto alto proprio in corrispondenza del periodo nel quale è uscito il film, cioè nello stesso periodo nel quale vediamo i picchi delle serie delle altre due ricerche.



Serie storica di *Google Trends* al variare del termine cercato.

"Loro" e il filtro per categorie

"Loro", non essendo solo il titolo di un film, può essere molto più cercato di quanto lo sia effettivamente il film. Può essere, quindi, utile imporre il filtro *"arte e intrattenimento"*.

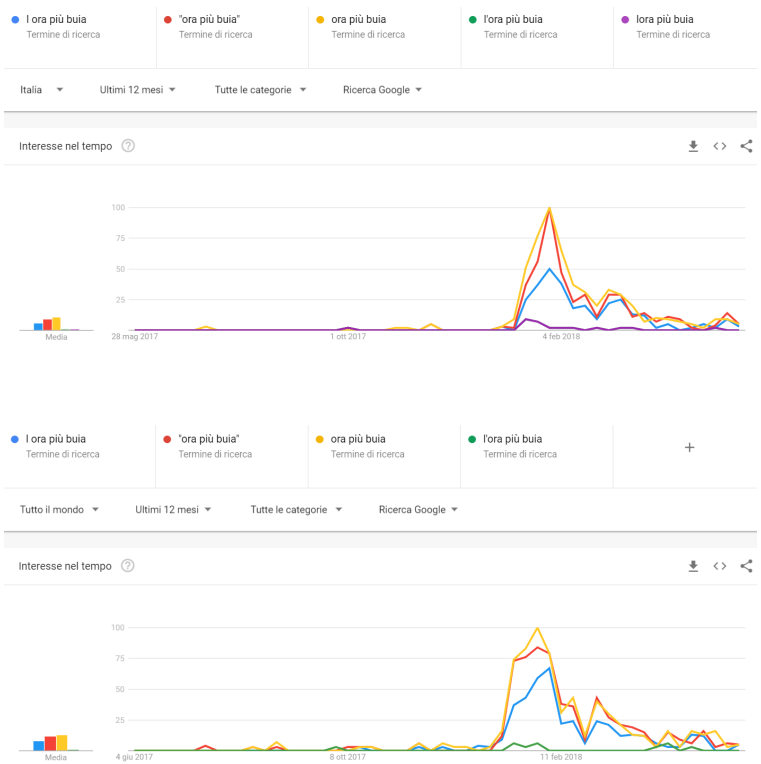


Serie storica di *Google Trends* con filtro *arte e intrattenimento*

PROBLEMA 4: Random sampling

Random sampling e errore di campionamento

Osserviamo i due grafici, fatti a distanza di una settimana dallo stesso IP, che ci testimoniano come i dati di *Google Trends* si basino solo su un *sample* delle ricerche realmente effettuate; questo fa sì che ci sia un errore di campionamento.



PROBLEMA 4: Random sampling

Secondo quanto riportato da Li (2016), questo problema è difficile da identificare ed è riportato solo in pochi articoli in quanto se un utente effettua la stessa ricerca nella stessa giornata dallo stesso account gmail e dallo stesso indirizzo IP, Google riporterà lo stesso valore dell'indice.

Choi e Varian, Carrière-Swallow e Labbé, D'Amuri e Marcucci

Choi e Varian citano solamente questo fatto: "Google Trends data is computed using a sampling method and therefore vary a few percent from day to day„. Carrière-Swallow e Labbé (2010) notano che il campionamento sembra avvenire quotidianamente, in modo tale che la richiesta di una query identica in giorni diversi restituisca serie leggermente diverse. D'Amuri e Marcucci (2015) nel loro articolo sottolineano che gli indici possono variare a seconda della data di download e dell'indirizzo IP.

Xinyuan Li e una proposta di soluzione

Nel suo articolo Li (2016) mostra due esempi nei quali è evidente che i risultati possano essere influenzati dall'errore di campionamento. Una soluzione potrebbe essere scaricare le serie da più indirizzi IP e account gmail in un breve periodo di tempo e calcolarne la media. Tuttavia questo richiederebbe un gran numero di indirizzi IP e account gmail. Inoltre, dato che non conosciamo i dati reali, non c'è modo di sapere quale sia la dimensione del campione minima per ottenere un buon risultato.

GFT - *Google Flu Trends*

Cos'è GFT?

GFT ossia *Google Flu Trends* era un servizio web creato da Google con lo scopo di rilevare l'attività influenzale monitorando le ricerche online riguardanti la salute. L'idea era quella di fornire i dati sulla diffusione della malattia in tempo reale, a differenza dei sistemi tradizionali come il *CDC* (Centers for Desease Control and Preventation) che pubblicano tali rilevazioni con lag di 1-2 settimane.

Il modello

I ricercatori hanno sviluppato un modello che stimava la probabilità che una visita medica in una certa regione fosse legata a sintomi influenzali. Il modello con un' unica variabile esplicative è il seguente:

$$logit(P) = \beta_0 + \beta_1 logit(Q) + \epsilon$$

dove

- P = percentuale di visite per influenza sul numero totale di visite
- Q = percentuale di query riguardanti l'influenza

Selezione delle query

Metodo automatizzato per la selezione delle query riguardanti l'influenza:

1. Per 9 regioni degli Stati Uniti sono state raccolte le 50 milioni di query (riguardanti qualsiasi tema) più cercate tra il 2003 e il 2008
2. Ogni query è stata testata separatamente nel modello come esplicativa Q
3. Sono stati raccolti 128 punti di training (da Settembre 2003 a Febbraio 2007) e 42 di validation (da Marzo 2007 a Maggio 2008) per ogni regione (dove un punto era una settimana)
4. E' stata implementata una 4-fold-validation, e ogni modello per-query è stato validato misurando la correlazione tra la stima del modello per i punti lasciati out-of-sample e la percentuale di regionale di ILI segnalata dal CDC per quei punti
5. Ogni query candidata otteneva così 36 diverse correlazioni (9 regioni per 4-fold-validation): per combinarle in una sola misurazione, dopo una trasformazione Z di Fisher, è stata considerata la loro media.

N top score

Della lista con gli score più alti di query cercate, sono state considerate le top N che davano un fit migliore, dove N è risultato essere 45.

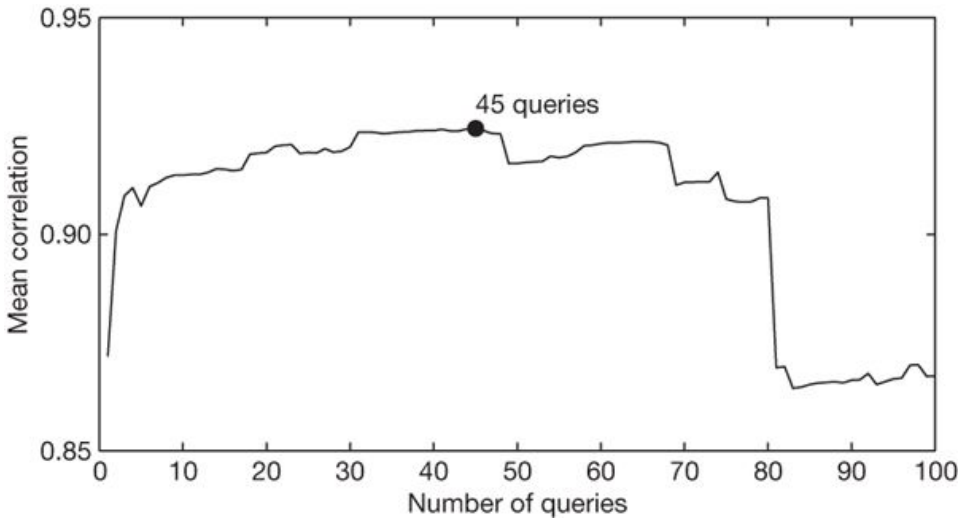


Grafico del fit del modello al variare di N

Modello finale

Il modello finale, fittato su tutte e 9 le regioni separatamente, usava come Q la frazione delle 45 query legate a ILI. Si è fatta poi una media delle correlazioni ottenute, producento un coefficiente indipendete dalle regioni.

Risultati

	Media	**Minima**	**Massima**
Correlazione train	0.9	0.8	0.96

Sintesi delle performance su 128 points per ogni regione - anni 2003-2007

	Media	**Minima**	**Massima**
Correlazione test	0.97	0.92	0.99

Sintesi delle performance su 42 points per ogni regione - anni 2007-2008

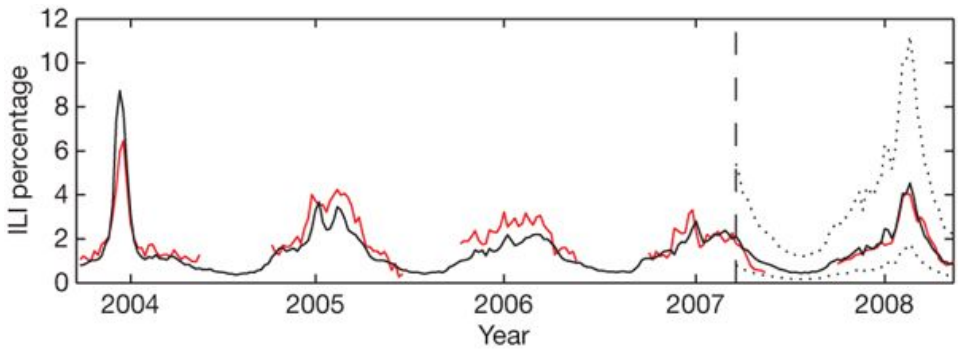


Grafico fit del modello

Big Data Hybris

- Tanti dati non sostituiscono i fondamenti teorici per l'analisi dei dati:
  - validità
  - replicabilità
  - struttura di dipendenza fra i dati
- I Big Data non nascono per produrre dati validi e replicabili per una analisi scientifica

Quantità e qualità

GFT utilizza tecniche statistiche tradizionali su Big Data

Date importanti

- 2009: pubblicazione ufficiale di GFT
  - ottima previsione sull'influenza aviaria 2005-2006
  - ottima previsione sull'influenza suina 2009
- Grandi errori di previsione dell'influenza stagionale su 2011-2012
- 2013: annunciati gli aggiornamenti avvenuti nel periodo precedente (solo tra Giugno e Luglio 2012 vi sono stati 86 aggiornamenti)

Perché non funziona?

1. Autocorrelazione tra gli errori non modellizzata: errori tra settimane diverse non sono indipendenti
2. Overfitting: costruito sui dati dal 2003 al 2008
3. Diverse query ma stesso significato: la query "indications of flu" è differente dalla query "indications of the flu"
4. Possibili distorsioni: vi possono essere fattori esterni, quali un panico mediatico, che fanno aumentare esponenzialmente il numero di ricerche oppure vi possono essere malati non per influenza che cercano le stesse query
5. Algorithm dynamics

Algorithm dynamics

Con *algorithm dynamics* si intende l'insieme dei cambiamenti fatti dagli ingegneri per migliorare il servizio commerciale e dai consumatori nell'utilizzarlo.

$\Rightarrow$

Cambia il processo di generazione dei dati

Endogeneità e esogeneità

GFT si basa sul presupposto che il volume di ricerca per certi termini si correlato a eventi esterni, tuttavia il comportamento di ricerca non è solo determinato in modo esogeno, ma è anche coltivato endogeneamente dal fornitore di servizi.

Conclusioni

Da limiti a potenzialità

Come proposto da Lazer nel suo articolo, la sfida è combinare l'utilizzo di dati estratti da fonti tradizionali con i *Big data*, come ad esempio quelli messi a nostra disposizione da *Google Trends* per fornire una più profonda e chiara comprensione del fenomeno preso in esame. Tutto ciò non può prescindere da una conoscenza approfondita dei dati di *Google Trends* con i relativi limiti e da una forte base teorica.

Bibliografia:

- Billari, F., D'Amuri, F., & Marcucci, J. (2013). Forecasting births using Google. In Annual Meeting of the Population Association of America.
- Carrière-Swallow, Y., & Labbé, F. (2013). Nowcasting with Google Trends in an emerging market. Journal of Forecasting, 32(4), 289-298.
- Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. Economic Record, 88(s1), 2-9.
- D'Amuri, F., & Marcucci, J. (2012). The predictive power of Google searches in forecasting unemployment. Bank of Italy, Temi di Discussione (Working Paper) No, 891.
- Li, X. (2016). Nowcasting with Big Data: is Google useful in Presence of other Information?. London Business School, Mimeo.
- Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. Nature, Vol 457.
- Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. "The Parable of Google Flu: Traps in Big Data Analysis", Science, Vol. 343.

Sitografia:

- https://trends.google.it/trends/?geo=IT
- http://www.boxofficemojo.com/intl/italy/
- https://support.google.com/trends/answer
- https://www.comingsoon.it/cinema/filmalcinema/
- https://www.mymovies.it/film/2018/