https://giaruffo.github.io/

giancarlo.ruffo@uniupo.it

https://arc2s.github.io/

Giancarlo Ruffo - Università degli Studi del Piemonte Orientale

Come (non) mentire con i dati

Distorsioni e Scienza dei dati:
bias percettivi, cognitivi, statistici e metodologici che non ci consentono di distinguere il vero dal falso

02 Dicembre 2024
Laurea in ICT (Innovazione sociale, Comunicazione e nuove Tecnologie)
Università degli Studi di Torino

Agenda

- La pipeline della Scienza dei Dati
- Il test dello Scimpanzé
- Distorsioni nella visualizzazione dei dati
- Misurare le bugie nei grafici
- Distorsioni statistiche e nell'interpretazione statistica
- Riferimenti

The Data Science pipeline

- Vedremo che le distorsioni che possono inficiare il processo raffigurato a dx sono presenti in tutti questi blocchi concettuali
- Forse la frase più abusata nel settore: *"if you torture the data long enough, it will confess to anything*", Ronald Coase, premio Nobel per l'Economia
- Andando indietro nel tempo (Alfred Sauvy, 1961): "*Les chiffres sont des innocents qui, sous la sollicitation, sous la torture, avouent très vite ce qu'on leur demande, quitte à se rétracter plus tard*"

Il test dello scimpanzè

The best stats you've ever seen - Hans Rosling

https://tinyurl.com/jvt4mgd

Il test dello scimpanzè

- 13 domande di geo politica
- 3 risposte possibili per ognuna

Test dello scimpanzé

Prova la tua cultura generale sul mondo!

RICORDA:
Non sarai giudicato se sbagli.
Rispondi con la risposta che credi sia giusta, evitando di "cercare su Internet".
Quando tutti avrete terminato vedremo insieme le risposte corrette e faremo qualche riflessione sui risultati.

1. Nelle nazioni che oggi nel mondo sono considerate a basso reddito, quante ragazze terminano le scuole elementari? *

- 20%
- 40%
- 60%

2. Dove vive la maggioranza della popolazione *

- Nelle nazioni a basso reddito
- Nelle nazioni a reddito medio
- Nelle nazioni ad alto reddito

3. Negli ultimi 20 anni, la proporzione della popolazione mondiale che vive in *

- quasi raddoppiata
- rimasta più o meno la stessa
- quasi dimezzata

147 responses

Accepting responses

Summary | Question | Individual

Insights

Average	Median	Range
3.99/13 points	4/13 points	1-10 points

Total points distribution

Frequently missed questions

Question	Correct responses
1. Nelle nazioni che oggi nel mondo sono considerate a basso reddito, quante ragazze terminano le scuole elementari?	20/147
2. Dove vive la maggioranza della popolazione mondiale?	55/147
3. Negli ultimi 20 anni, la proporzione della popolazione mondiale che vive in estrema povertà è	29/147
4. Qual è oggi l'aspettativa di vita nel mondo?	66/147
5. Ci sono circa 2 miliardi di bambini oggi nel mondo, tra i 0 ed i 15 anni. Quanti bambini ci saranno nel 2100, secondo le Nazioni Unite?	49/147

Cosa ci dice il test dello scimpanzè

- I nostri pregiudizi peggiorano la nostra capacità di analisi
- Dobbiamo approcciare la visualizzazione e la comunicazione scientifica nel modo più “onesto” possibile: prima ancora di mentire agli altri, rischiamo di mentire a noi stessi.
- … non abbiamo neanche iniziato a parlare dei limiti della nostra percezione.

Lezione #1

- Non fidatevi dei vostri pregiudizi e fate parlare i dati (senza torturarli)

"LET MY DATASET CHANGE YOUR MINDSET"

Hans Rosling
(1948-2017)

Distorsioni nella visualizzazione dei dati

Il potere della visualizzazione dati

Le tabelle piene di informazioni funzionano perfettamente se abbiamo dei valori precisi e rendono facile l'identificazione di dati individuali

2007 Sales Revenue
(U.S. dollars in thousands)

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Domestic	1983	2343	2593	2283	2574	2838	2382	2634	2938	2739	2983	3493
International	574	636	673	593	644	679	593	139	599	583	602	690
	$2,557	$2,979	$3,266	$2,876	$3,218	$3,517	$2,975	$2,773	$3,537	$3,322	$3,585	$4,183

Cosa fare quando vogliamo confrontare i dati, trovare "pattern", tendenze, relazioni, etc.?

2007 Sales Revenue (esempio)

John Snow's data journalism: the cholera map that changed the world

http://www.theguardian.com/news/datablog/2013/mar/15/john-snow-cholera-map

Le “visualizzazioni” ci aiutano ad interpretare dati e numeri.

I numeri sono “fatti”, “osservazioni”. Non possono mentire

Ma…

Usare le visualizzazione in modo efficace

"We must do more than simply display data graphically: we must understand how visual perception works and then present data visually in ways that follow the rules"

Stephen Few, Now You See It, Analytics Press

Chapter 3: Thinking with our eyes

Come funziona la percezione visiva

Cosa possiamo percepire?

- percepiamo un intero oggetto nella sua totalità, ma possiamo anche percepirlo come composto di diverse proprietà visive

- posizione 2D, lunghezze, ampiezze, profondità, forma, colore orientamento, …

- Possiamo potenzialmente “catturare” molti valori in una sola volta (i valori possono essere combinati in “pattern”)

Quanti 3 vedete nella sequenza sottostante?

2481348118711671554138819844377134791564153184530584864 1
2347578941148412223881469161354804840789087707867875121 1
8658423404437713479156415318453058486412347578941148412 2
2388146916135480484078908770786787512118658423401887427 6

Quanti 3 vedete nella sequenza sottostante?

2481348118711671554138819844377134791564153184530584864123475789411484122388146916135480484078908770786787512118658423401887427623475789411484122238814691613548048407890877078678751211865842340443771347915641531845305848641234757894114841222388146916135480484078908770786787512118658423401887427623881469161354804840789087707867875121186584234018874276

Lezione #2

“Non limitiamoci a tutto quello che possiamo vedere. La percezione visiva è selettiva ed è giusto che sia cosí, dato che ci obbliga a concentrarci su ciò che potrebbe sopraffarci. La nostra attenzione si focalizza spesso su ciò che si differenzia dall’ordinario."

Cosa vedete qui?

“Our eyes are drawn to familiar patterns.
We see what we know and expect”

Cos’altro vedete qui?

“Our eyes are drawn to familiar patterns.
We see what we know and expect”

Okay, c'è una rosa. Cos'altro vedete dentro la rosa?

"Our eyes are drawn to familiar patterns.
We see what we know and expect"

Okay, facciamo tutti uno sforzo per vedere un **delfino** dentro la rosa…

"Our eyes are drawn to familiar patterns.
We see what we know and expect"

Lezione #3

“I nostri occhi sono “guidati” da figure familiari. Vediamo ciò che conosciamo e ci aspettiamo di vedere”

https://www.youtube.com/watch?v=oH6yOc606uQ

Lezione #4

“La memoria gioca un ruolo fondamentale nella cognizione umana, ma la memoria utile è estremamente limitata”

Attributi “pre-attentive”

- Quattro categorie
  - Colori
  - Forme
  - Posizioni
  - Animazioni

- Associa valori e dati ad attributi “pre-attentive” con cautela:
  - Non creare DISTRAZIONI dai dati
  - Tieni in considerazione la memoria a breve termine

Form
Orientation
Line Length
Line Width
Size
Shape
Curvature
Added Marks
Enclosure
Color
Intensity
Hue
Spatial Position
2-D Position

Esempio

Forme

https://makingmaps.net/2007/08/28/perceptual-scaling-of-map-symbols/

Forme

- Dimensioni effettive e percepite nei cerchi

Forme

- Dimensioni effettive e percepite nei cerchi

https://en.wikipedia.org/wiki/Ebbinghaus_illusion

Animazioni

- Due attributi
  - Intermittenza (compare/scompare)
  - Movimento (da una posizione all'altra)
- Uno dei modi più efficaci per attirare l'attenzione
- Molto spesso abusato nel marketing

Attenzione e movimento

https://www.youtube.com/watch?time_continue=2&v=vJG698U2Mvo

https://www.youtube.com/watch?v=IGQmdoK_ZfY

Contorni e margini

https://www.csc2.ncsu.edu/faculty/healey/PP/

Contorni e margini (2)

Un contorno definito da due o più attributi diventa difficile da identificare

Gestalt

Tendenza a percepire oggetti come parti di un tutto

Chiusura

Figure e sfondo

Pareidolia

Effetto spesso usato per supportare credenze, superstizioni o teorie del complotto.

Riferimenti

https://en.wikipedia.org/wiki/Pareidolia

http://gestaltrevision.be/en/master-index/66-what-we-do/overview/research-areas/mid-level/multistability/71-pareidolia

Lezione #5

“Usare attributi pre-attentive per valorizzare ed esaltare strutture che emergono dai dati, ma non esagerare. Inoltre, non fidatevi troppo di quello che l’occhio vede e vuole vedere”

Raccontare storie diverse con stessi dati

anche "mappe di calore" e dotmap hanno problemi

gerrymandering

Congressional District 17

nationalatlas.gov™

17 Congressional District

Fulton County

Illinois (19 Districts)

Congressional District 2

nationalatlas.gov™

2 Congressional District

Grand County

Utah (3 Districts)

Congressional District 22

nationalatlas.gov™

22 Congressional District

Harris County

Texas (32 Districts)

Congressional District 12

nationalatlas.gov™

12 Congressional District

Rowan County

North Carolina (13 Districts)

Valutare grafici e figure

Information visualization: come farlo bene

The Visual Display of Quantitative Information
Edward R. Tufte

https://www.edwardtufte.com/tufte/books_vdqi

Integrità grafica

Connecticut Traffic Deaths,
Before (1955) and After (1956)
Stricter Enforcement by the Police
Against Cars Exceeding Speed limit

Integrità grafica

- Niente contesto
- Da quale pattern viene il grafico precedente?

Integrità grafica

Lie Factor

$$\text{Lie Factor} = \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}}$$

Lie factor should be close to one

i.e. $1.05 > \text{Lie Factor} > 0.95$

$$\frac{27.5-18.0}{18.0} = 0.53 \quad \frac{5.3-0.6}{0.6} = 7.83 \quad \text{Lie Factor} = \frac{7.83}{0.53} = 14.8$$

This line, representing 18 miles per gallon in 1978, is 0.6 inches long.

This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.

Integrità grafica = no lies!

Lezione #6

“Non introdurre bugie nascondendosi dietro l’idea di migliorare l’estetica della infografica o di semplificarne la lettura. Gli errori sono misurabili, prima o poi vengono a galla e la reputazione si perde in un baleno”

Bias statistico

Il bias statistico

- Definizione standard: *Il bias statistico è un termine utilizzato per descrivere le statistiche che non offrono una rappresentazione accurata della popolazione. Alcuni dati sono difettosi perché il campione di persone preso in esame non rappresenta accuratamente la popolazione.*
- I nostri modelli sono addestrati su dati che troppo spesso accettiamo 'come atto di fede'
- Bisogna conoscere il dataset sul quale lavoriamo, il modo in cui i dati sono stati raccolti, quanto è rappresentativo, etc.
  - Conoscere le limitazioni del nostro lavoro: tutto il resto è fuffa!

The Washington Post
Democracy Dies in Darkness

Racial bias in a medical algorithm favors white patients over sicker black patients

By Carolyn Y. Johnson

October 24, 2019 at 2:00 p.m. EDT

epic.org | ELECTRONIC PRIVACY INFORMATION CENTER

ISSUES + OUR

Artificial Intelligence and Human Rights

AI in the Criminal Justice System

BACKGROUND

Automated decision-making tools are used widely and opaquely both directly in the criminal justice system and in ways that directly feed the criminal justice cycle in the U.S.

Figure 1: Left: Percentage of explained variance in PCA in the embedding differences. Right: Selected words projecting to the first two principle components where the blue dots are the sentences with male context and the orange dots are from the sentences with female context.

Gender Bias in Word Embeddings: A Comprehensive Analysis of Frequency, Syntax, and Semantics

Authors: Aylin Caliskan, Pimparkar Parth Ajay, Tessa Charlesworth, Robert Wolfe, Mahzarin R. Banaji

Authors Info & Claims

AIES '22: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society • July 2022 • Pages 156–170 • https://doi.org/10.1145/3514094.3534162

Published: 27 July 2022 Publication History

Check for updates

8 1,932

eReader PDF

Distorsioni dell'interpretazione statistica

Correlazioni spurie

http://www.tylervigen.com/spurious-correlations

Correlazioni spurie

http://www.tylervigen.com/spurious-correlations

La media statistica

- Distribuzione delle *altezze* negli individui: la media ci aiuta a fare predizioni
- Ma attenzione, anche rimanendo alla statistica elementare, non c'è solo la media!

La media statistica

esempio: Distribuzione della ricchezza

La media in questo caso non ha alcun valore statistico!

In questo caso gli eventi estremi sono poco probabili, ma comunque MOLTO più probabili che con una curva a "campana"

Esempio: climate change

- "Quest'anno anno abbiamo avuto le minime più basse degli ultimi anni: con questo freddo, come si fa a parlare di *riscaldamento globale*?"

Capire un fenomeno complesso da pochi eventi locali è **sbagliato**

Abbiamo bisogno di molto più contesto e di tante tante tante analisi

Immagine prodotta da Olaf Jensen - "Chi ha paura del buio?"

Lezione #7

“Non torturare i dati con una matematica sbagliata. Indici, misure, rappresentazioni e modelli per fare predizioni sono spesso utili, ma il rischio di usare strumenti non fondati matematicamente è altissimo e quasi mai i risultati sono ovvi”

Conclusioni

Basta cosí? Per nulla…

- Per natura siamo soggetti a moltissime *distorsioni cognitive*: bias di conferma, astrazione selettiva dell'informazione, ipergeneralizzazione, etc.
- Viviamo in camere d'eco (echo chamber), siamo soggetti alle filter bubble indotte dagli algoritmi di selezione e personalizzazione delle informazioni, le società hanno una tendenza naturale alla polarizzazione e tutti questi fenomeni sono correlati tra di loro!
- Se siete *scienziati sociali e della comunicazione*, bisogna stare attenti ad idolatrare le **opinioni**: dovrebbero essere subordinate ai fatti e non viceversa.
- Se siete *scienziati "duri"* e avete elaborato un **modello fisico** (deterministico o stocastico): non innamoratevene mai! il modello deve essere sempre **falsificabile** (Aristotele, Popper, etc.) e deve essere validato. Il modello è utile fino a quando non sarà falsificato e **prima o poi accadrà**.
- Se siete degli esperti di **machine learning** e avete addestrato il vostro modello: attenzione al bias nel data set iniziale, all'overfitting. Questi modelli sono una scatola nera: anche se le predizioni sono ottime, non capiamo davvero cosa succede dentro. Usare con **cautela** e **competenza**
- Qualsiasi cosa siate, colti o meno colti che siate: avete dei **pregiudizi**. Dovete iniziare ad ammetterlo a voi stessi e ad esserne consapevoli quando consumate o producete informazione.…

Scoraggiati?

- "*Questo secolo oramai alla fine*
  Saturo di parassiti senza dignità
  Mi spinge solo ad essere migliore
  Con più volontà" (F. Battiato, "E ti vengo a cercare", 1988)
- Mai scoraggiarsi! C'è tanto buon giornalismo basato sui fatti e gli amici sui social possono fare (consapevolmente o meno) ottime rassegne stampa, ma bisogna impare l'arte dello **scetticismo**.
- Internet può essere il vostro nemico, ma anche il vostro migliore amico: https://www.generazioniconnesse.it/site/_file/documenti/Comunicazione/Fake_news/Dispensa_Docenti_decalogo_bastabufale.pdf

Come diventare un detective anti-bufala di Paolo Attivissimo

1. Non condividere notizie che non hai verificato
2. Usa gli strumenti di Internet per verificare le notizie
3. Chiedi le fonti e le prove
4. Chiedi aiuto a una persona esperta o a un ente davvero competente
5. Ricorda che anche Internet e i social network sono manipolabili
6. Riconosci i vari tipi e gli stili delle notizie false
7. Hai un potere enorme: usalo bene
8. Dai il buon esempio: non lamentarti del buio, ma accendi una luce

   Inoltre (mie aggiunte):

9. Se la notizia è riportata in un articolo scientifico, controlla se non sia stato ritratto o se non sia stato smentito da una ricerca successiva: nelle Scienze Dure non esistono Verità assolute ed eterne come in matematica!
10. Se la notizia è in linea con quella che credi sia la verità: applica le regole di sopra DUE volte

ARC²S

Applied Research on Computational Complex Systems

In Science we trust, though Science doesn't always reciprocate!

Get Started

Complex Networks

Modeling. Structural analysis. Prediction and process dynamics.

Data Science

Analyzing. Representation. Understanding. Discovery.

Computational Social Science

Investigating. Societal dynamics. Computational methods.

Data Visualization

Visualizing. Data fusion. Data interaction. Visual analytics.

ARC²S @ UPO UNIVERSITÀ DEL PIEMONTE ORIENTALE

Home | About us | Team | Projects | Papers | Talks | Contact

TEAM

We are spread between University of Eastern Piedmont (UPO) and University of Turin (UniTO). However, we like to travel and visit each other wherever we are.

Current Members

Giancarlo Ruffo
Associate Professor @UPO

Anna Sapienza
Tenure-Track Assistant Professor
RTD-B @UPO

Mirko Lai
Tenure-Track Assistant Professor
RTD-B @UPO

Salvatore Vilella
Postdoctoral Researcher
Assegno di ricerca @UPO

Former Members @UniTO

Rossano Schifanella
Associate Professor
University of Turin

Arthur Capozzi
Postdoctoral researcher
ETH Zürich

Alfonso Semeraro
Senior Data Scientist
Wind 3

Alessandra Urbinati
Postdoctoral Researcher
Northeastern University

Narges Azizfard
Postdoctoral Researcher
University of Helsinki

Edoardo Galimberti
Senior Data Scientist
Allianz

Marcella Tambuscio
Postdoctoral Researcher
Karl-Franzens-Universität Graz

Emilio Sulis
Short-Term Assistant Professor
RTD-A
University of Turin

Martina Deplano
Freelancer

Luca Maria Aiello
Associate Professor
ITU Copenhagen

André Panisson
Principal Researcher
CENTAI

Salvatore Spoto
Game Developer
34BigThings

Marco Milanesio
Research Engineer
Université Côte d'Azur

Alessandro Basso
Software engineer
Ahold Delhaize

https://arc2s.github.io/

Thanks!