Uomini, dati e SALAMI: una questione scientifica aperta

Maria Chiara Pievatolo

Università di Pisa

pievatolo@dsp.unipi.it

27 settembre 2024

Incontro su AI e ricerca - Milano Statale

Outline

UPDATED: Let's forget the term AI. Let's call them Systematic Approaches to Learning Algorithms and Machine Inferences (SALAMI).

24 Comments / November 24, 2019

-(A sunday morning provocation)

Artificial Intelligence is a name born in 1956. It has gained a lot of attention thanks to its resonance with humans, and favoured the devlopment of an imaginative Hollywood production stream.

The name "artificial intelligence" has an implicit bias that does not allow for a cognitive perception adherent to reality.

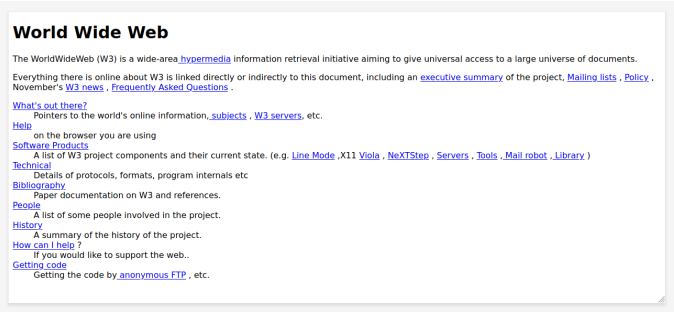
On the contrary, the name favours the suggestion of the possibility of machines to develop some form of consciouness, emotions, acquire a "personality" similar to humans' and, ultimately overcome human limitations and developing a self superior to humans.

You've seen the movies, you know the narrative… But they are only devices that extract correlations from data and use those correlations to make predictions and a load of very useful things. And as having calculators doing square roots, they can do it at a scale and speed far exceeding human's performances. By throwing in some logic and randomness they can also exhibit some interesting and original behaviors. Yet machines have no clue of what reality is. At best, they mimic a model of the reality, so they are two steps away from reality.

After a conference on AI at the Pontifical Academy Of Science in Rome, discussing with some friends (among them Aimee van Wynsberghe), we argued that **the first and foremost AI bias is its name**. It induces analogies that have limited adhrence to reality and it generates infinite speculations (some of them causing excessive expectations and fears).

https://blog.quintarelli.it/2019/11/lets-forget-the-term-ai-lets-call-them-systematic-approaches-to-learning-algorithms-and-machine-inferen

Una rivoluzione di 30 anni fa



World Wide Web

The WorldWideWeb (W3) is a wide-area hypermedia information retrieval initiative aiming to give universal access to a large universe of documents.

Everything there is online about W3 is linked directly or indirectly to this document, including an executive summary of the project, Mailing lists , Policy , November's W3 news , Frequently Asked Questions .

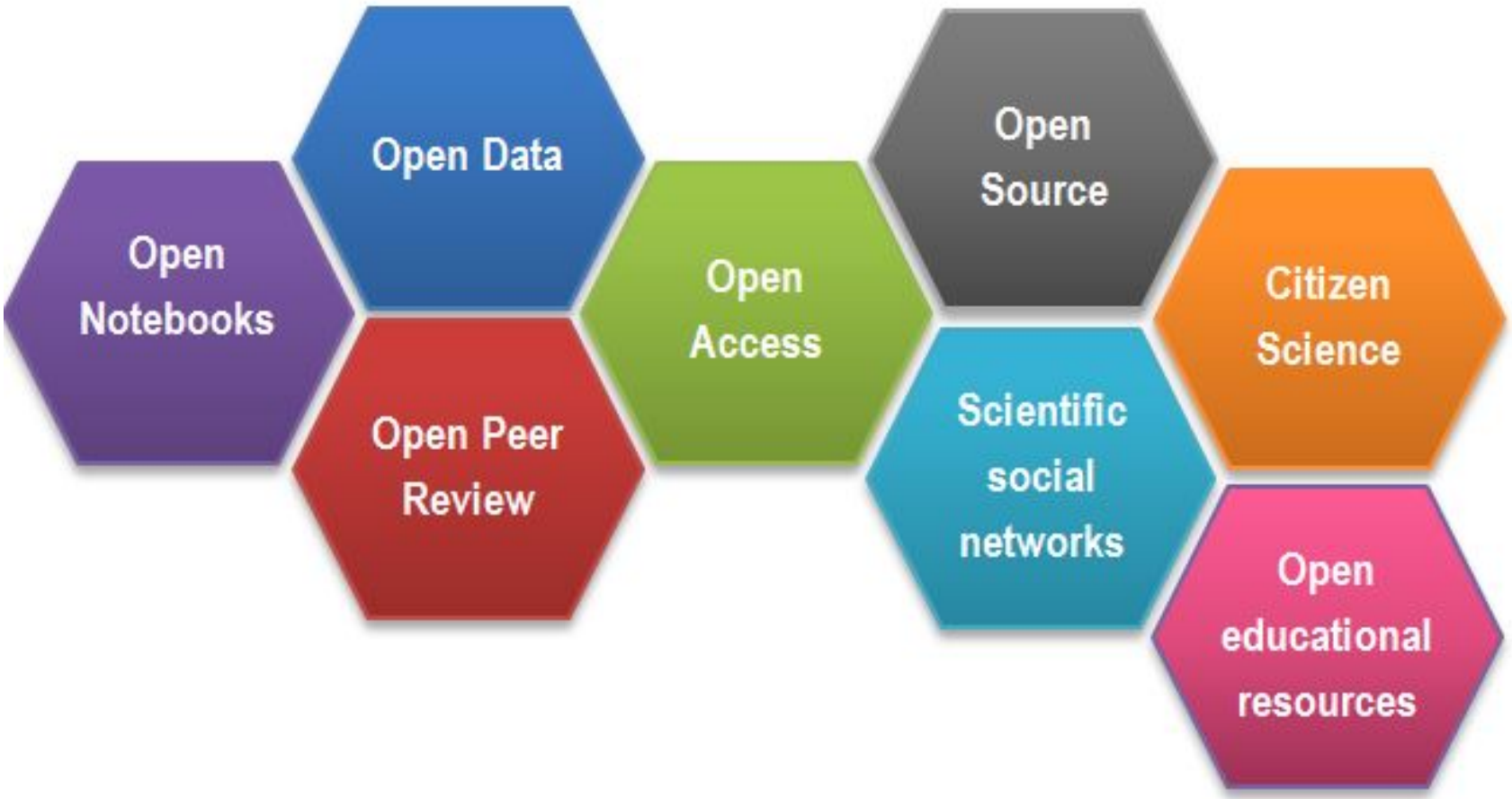
What's out there?
Pointers to the world's online information, subjects , W3 servers, etc.
Help
on the browser you are using
Software Products
A list of W3 project components and their current state. (e.g. Line Mode ,X11 Viola , NeXTStep , Servers , Tools , Mail robot , Library)
Technical
Details of protocols, formats, program internals etc
Bibliography
Paper documentation on W3 and references.
People
A list of some people involved in the project.
History
A summary of the history of the project.
How can I help ?
If you would like to support the web..
Getting code
Getting the code by anonymous FTP , etc.

La macchina: il World Wide Web (1989-1991) 1991: ArXiv

Le persone: la comunità scientifica

I testi: le licenze libere (1989, 2001)

L'alveare dell'open science



https://www.fosteropenscience.eu/content/what-open-science-introduction

Scienza aperta o, semplicemente, scienza (1610)


SIDEREVS
NVNCIVS
MAGNA, LONGEQVE ADMIRABILIA
Spectacula pandens, suspiciendoque proponens
vnicuique, præsertim verò
PHILOSOPHIS, necnon ASTRONOMIS, quæ à
GALILEO GALILEO
PATRITIO FLORENTINO
Patauini Gymnasij Publico Mathematico
PERSPICILLI
QVATVOR PLANETIS
MEDICEA SIDERA
NVNCVPANDOS DECREVIT.
VENETIIS, Apud Thomam Baglionum. M DC X.
M VIIII

"Parmi necessario, per aumentare il grido di questi scoprimenti, il fare che con l'effetto stesso sia veduta et riconosciuta la verità da più persone che siia possibile..." (Lettera a B. Vinta, 1610)

La rivista scientifica commerciale come dato valutativo inaggirabile

How to make open access the natural choice for researchers

Many who advocate open access envisage the development of a new publishing environment—new journals, new ways of operating—in which researchers can eventually be resettled. But it may be preferable to work with the publishing habitat that has evolved organically and bring open access into it. This could be achieved by transforming the existing core journals' business models while simultaneously maintaining their function of providing quality assurance through peer review, publishing services and brand value.

This would enable a large-scale shift to open access while still providing researchers with the services and functions of the journal publishing system in which they are comfortable. The beauty of this idea is that the disruption would be perceptible only in the organisational domain in which the money is managed; since this side of business is typically hidden from researchers, authors would not experience any disturbance to their ordinary publishing activity.

Ralf Schimmer is head of scientific information provision at the Max Planck Digital Library in Munich.

"Research Europe"

2015

https://www.mpdl.mpg.de/images/documents/Nachrichten/schimmer_ResearchEurope.pdf

Doppia alienazione bibliometrica: valutare senza sapere, sapere senza valutare

> Unlike traditional practices of evaluation that, like peer review, are not just qualitative but craft-based, metrics cannot be produced by a single scholar but are instead obtained, typically for a fee, from large data analytics corporations - yet another example of today's monetization of data. The introduction of quantitative and automated methodologies has thus introduced a **new separation between the producer and the user of the evidence on which the evaluation rests** - two roles that were traditionally folded into the same person: the scholar who read and judged. Metrics are therefore a "doubly alien" form of knowledge: both produced and used by people who are not practitioners of the field to which the publications belong. (Mario Biagioli, 2018 `https://www.journals.uchicago.edu/doi/10.1086/699152`)

Scienziati statisticamente fungibili

Chi vuole intraprendere strade non ancora accettate dalla comunità in primo luogo ha difficoltà a pubblicare, scontrandosi con un **muro omogeneo e anonimo**. Se anche, come supponiamo per comodità di argomentazione, riuscisse nell'intento di inaugurare una scuola di pensiero alternativa sarebbe ovviamente poco citato, perché sarebbero ben rari i ricercatori che sceglierebbero di entrare in un gruppo minoritario, sapendo che il meccanismo quantitativo di valutazione, basato sul numero di citazioni, attribuirebbe ai loro risultati certamente un valore minimo.**Il meccanismo per sua natura evidentemente si autoalimenta, generando automaticamente omogeneità**. Un cambiamento di opinione è reso possibile solo da una transizione di fase che cambi contemporaneamente l'opinione di tutti gli specialisti. È ciò che avviene effettivamente con il rapido susseguirsi delle mode. Le qualità che vengono così selezionate sono la **repentinità dell'informazione e la prontezza di riflessi che permettono di trovarsi sempre dalla parte maggioritaria**. (Lucio Russo, *La cultura componibile*, 2008)

Data o capta?

Data

"Data" has become a bloodless word; it disguises both its material origins and its ends. And if data is seen as abstract and immaterial, then it more easily falls outside of traditional understandings and responsibilities of care, consent, or risk. As researchers Luke Stark and Anna Lauren Hoffman argue, metaphors of data as a "**natural resource**" just lying in wait to be discovered are a well-established rhetorical trick used for centuries by colonial powers. [Crawford, 2021]

Capta

I dati sono sempre **presi**: "How data are construed, recorded, and collected is the result of human decisions — decisions about what exactly to measure, when and where to do so, and by what methods." [Barrowman, 2019]

...e **cotti secondo le nostre ricette**: "Raw data is both an oxymoron and a bad idea; to the contrary, data should be cooked with care." [Bowker, 2005]

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification

Joy Buolamwini, Timnit Gebru *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, PMLR 81:77-91, 2018.

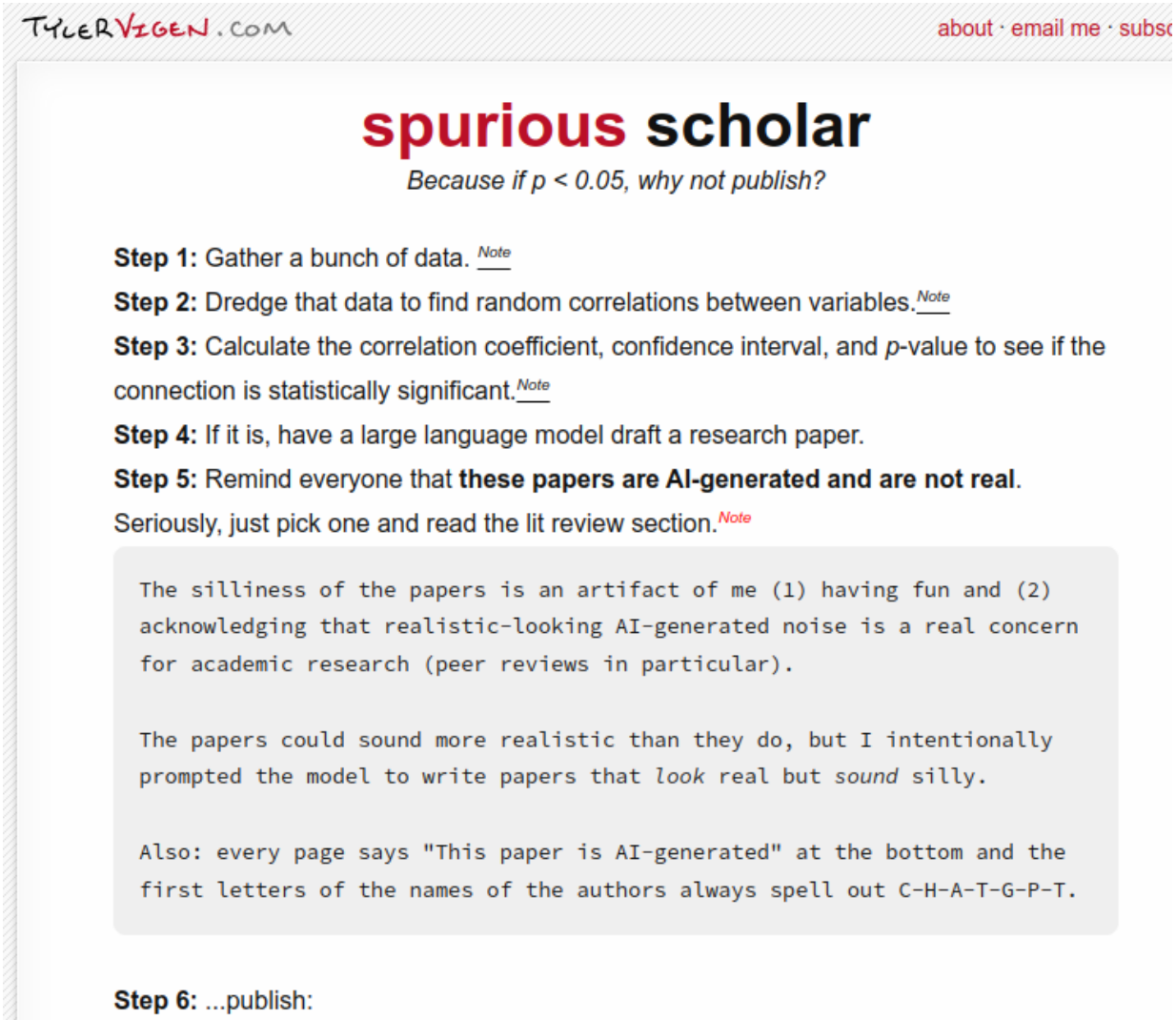
Accesso critico ai dati di addestramento

Abstract

Recent studies demonstrate that machine learning algorithms can discriminate based on classes like race and gender. In this work, we present an approach to evaluate bias present in automated facial analysis algorithms and datasets with respect to phenotypic subgroups. Using the dermatologist approved Fitzpatrick Skin Type classification system, we characterize the gender and skin type distribution of two facial analysis benchmarks, IJB-A and Adience. We find that these datasets are overwhelmingly composed of lighter-skinned subjects (79.6% for IJB-A and 86.2% for Adience) and introduce a new facial analysis dataset which is balanced by gender and skin type. We evaluate 3 commercial gender classification systems using our dataset and show that darker-skinned females are the most misclassified group (with error rates of up to 34.7%). The maximum error rate for lighter-skinned males is 0.8%. The substantial disparities in the accuracy of classifying darker females, lighter females, darker males, and lighter males in gender classification systems require urgent attention if commercial companies are to build genuinely fair, transparent and accountable facial analysis algorithms.

https://proceedings.mlr.press/v81/buolamwini18a.html

Correlazioni spurie: uso acritico dei dati



TYLERVIGEN.COM about · email me · subsc

spurious scholar

Because if p < 0.05, why not publish?

Step 1: Gather a bunch of data. Note

Step 2: Dredge that data to find random correlations between variables. Note

Step 3: Calculate the correlation coefficient, confidence interval, and *p*-value to see if the connection is statistically significant. Note

Step 4: If it is, have a large language model draft a research paper.

Step 5: Remind everyone that **these papers are AI-generated and are not real**. Seriously, just pick one and read the lit review section. Note

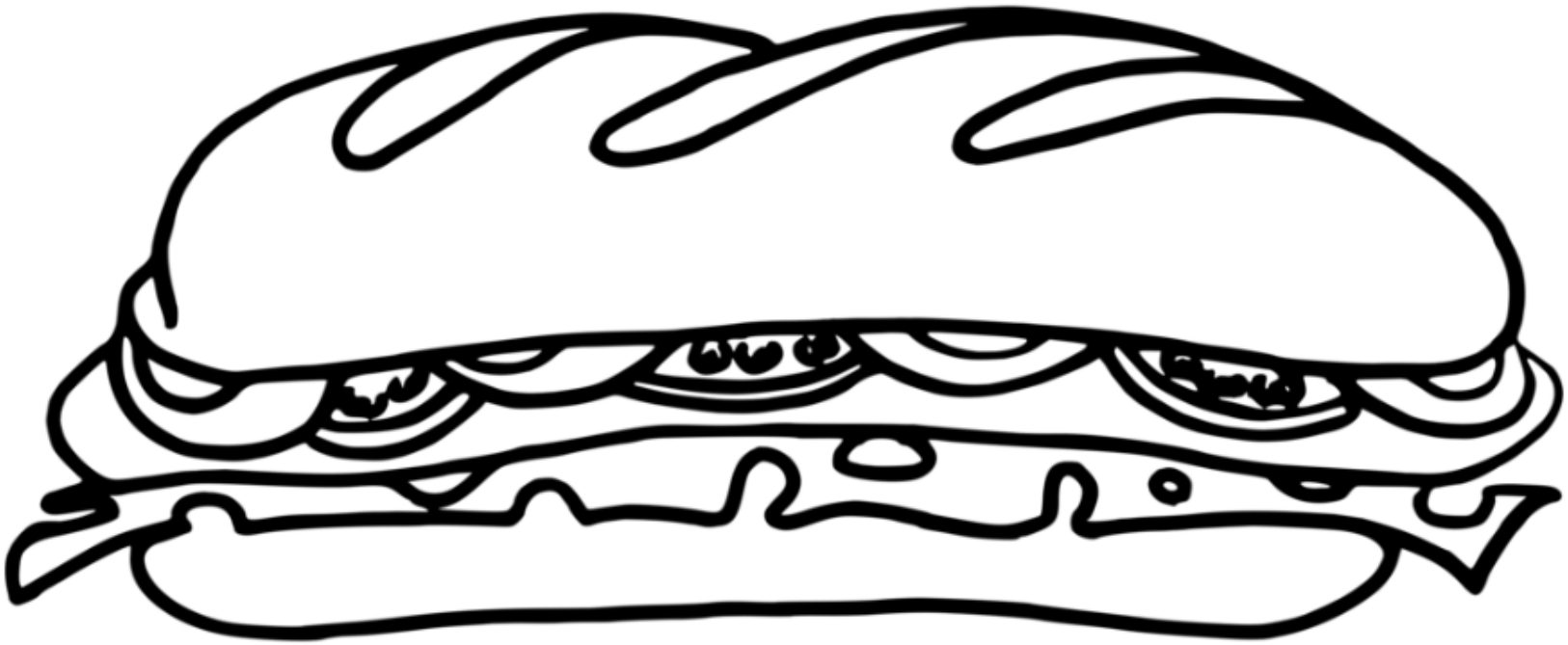
> The silliness of the papers is an artifact of me (1) having fun and (2) acknowledging that realistic-looking AI-generated noise is a real concern for academic research (peer reviews in particular).
>
> The papers could sound more realistic than they do, but I intentionally prompted the model to write papers that *look* real but *sound* silly.
>
> Also: every page says "This paper is AI-generated" at the bottom and the first letters of the names of the authors always spell out C-H-A-T-G-P-T.

Step 6: ...publish:

https://tylervigen.com/spurious-scholar

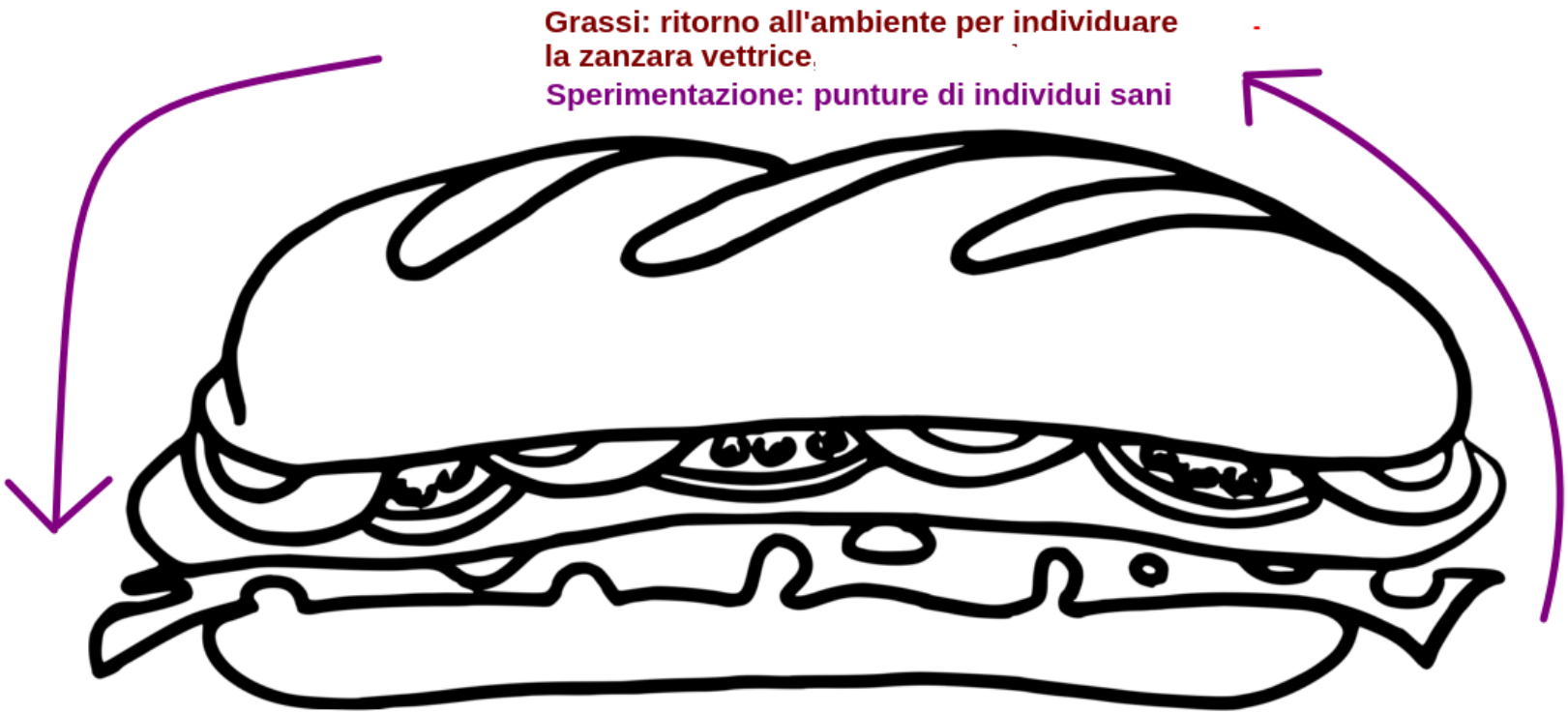
SALAMI, dati e ricerca

Sei sicuro di non aver trovato correlazioni spurie? Sei in grado di ipotizzare la causa sottostante? Sei capace di provarla/falsificarla sperimentalmente?

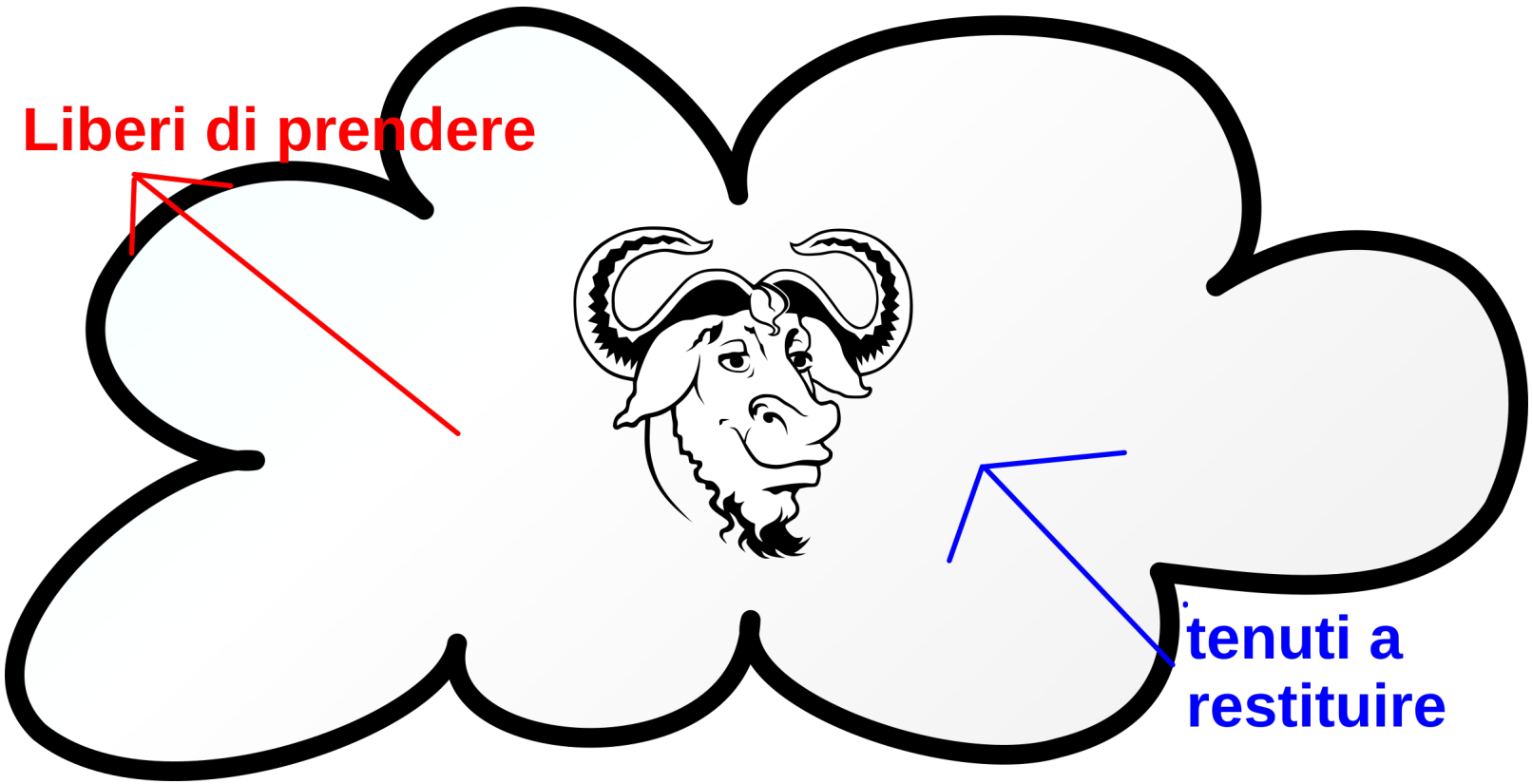


Che dati usi? Come li hai presi? Come li hai selezionati? Perché li hai presi?

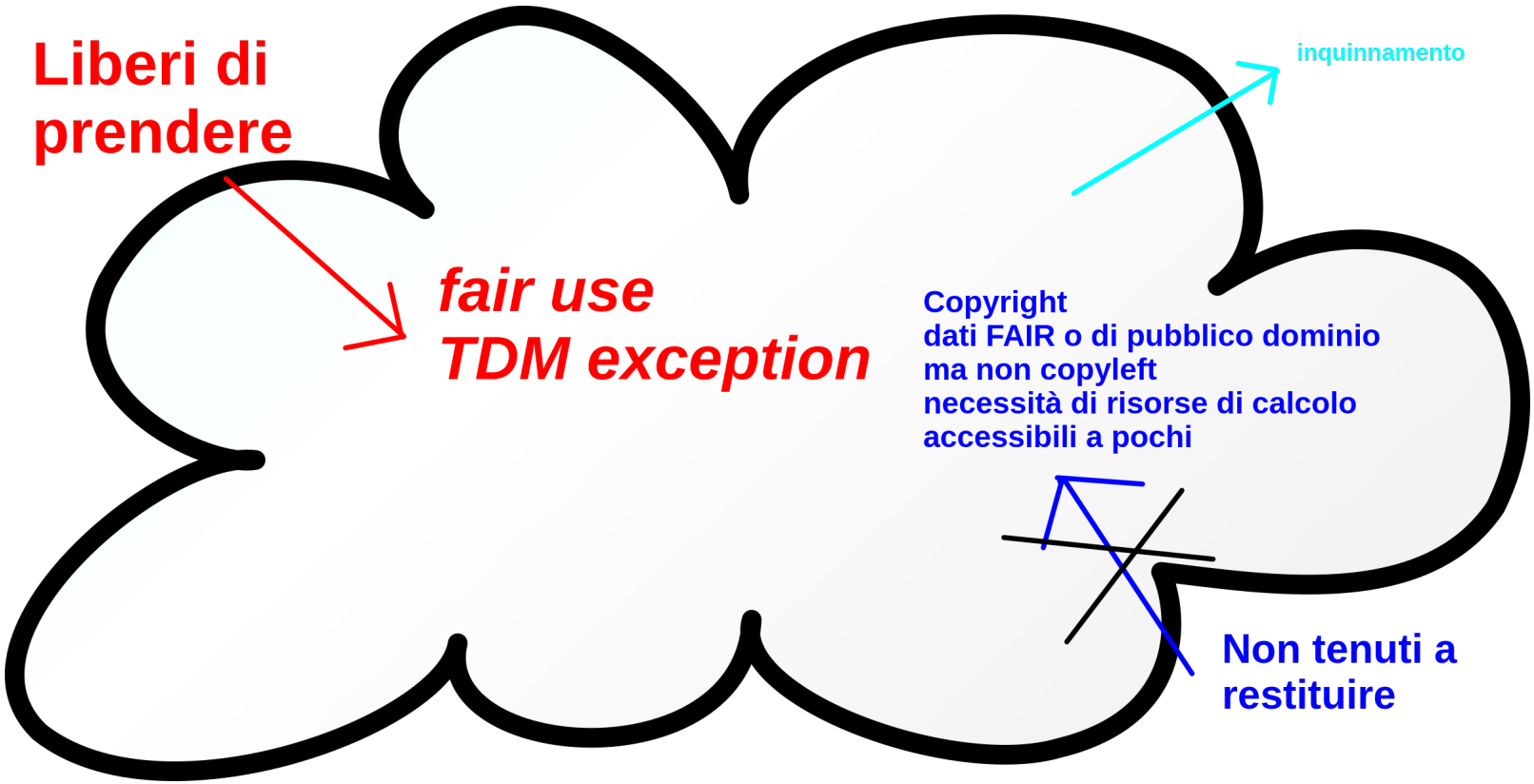
Malaria: correlazioni, dati e ricerca



La commedia dei beni comuni



Una nuova tragedia dei beni comuni (Felix Reda)



https://opensource.org/blog/copyright-law-makes-a-case-for-requiring-data-information-rathe

Due uscite alternative

tutela dei beni comuni mettere il copyright al servizio della scienza: copyleft su tutto ciò che è sotto copyright ed è usato come dato d'addestramento

pretendere le briciole del banchetto di Big Tech un nuovo diritto sui generis su ciò che è usato come dato di addestramento ed è sotto copyright

SALAMI e ricerca

- I SALAMI non correggono le distorsioni (bias): le esaltano, avendo tutti i problemi della blbliometria proprietaria (dati invisibili, valorizzazione della ricerca alla moda, esposizione al gaming, formazione di monopoli e oligopoli) e anche qualcuno in più ("allucinazioni", plagiarism machine)
- la scienza aperta potrebbe confezionarli per un uso scientifico, ma non senza una riforma della valutazione della ricerca e del copyright
- i giovani, prima di confrontarsi con i SALAMI, dovrebbero emanciparsi dal pensiero magico. Per farlo dovrebbero vivere in un ambiente telematico libero,[Levi,2022] a infrastruttura federativa e decentrata, basato su free software, controllo dei propri dati personali e robuste regole antitrust e di tutela dell'ambiente. [Strether, 2024]


Mel Andrews et al. (2024)
The reanimation of pseudoscience in machine learning and its ethical repercussions
https://linkinghub.elsevier.com/retrieve/pii/S2666389924001600


Maria Chiara Pievatolo, (2021)
I custodi del sapere
https://btfp.sp.unipi.it/it/2021/05/i-custodi-del-sapere/


Lucio Russo (2008)
La cultura componibile
https://ipfs.io/ipfs/bafykbzacearoado5rqfbgi74u4f5dx6752dnpqxk4zgeukz2fltlztcaetlqq?filename=la-cultura-componibile-dalla-frammentazione-alla--annas-archive.djvu


Michael Hagner (2018)
Open access, data capitalism and academic publishing
https://doi.org/10.4414/smw.2018.14600


Nick Barrowman (2019)
Why Data Is Never Raw
https://www.thenewatlantis.com/publications/why-data-is-never-raw


Rob Kitchin (2018)
The Data Revolution. Big Data, Open Data, Data Infrastructures and their Consequences
https://pdfs.semanticscholar.org/c5ff/796807fc22db9037ae779e60b9e3c305909e.pdf


Geoffrey C. Bowker (2005)
Memory practices in the sciences
https://www.morgan-klaus.com/readings/memory-practices.html

Lai Ma (2023

The Platformisation of Scholarly Information and How to Fight It
https://doi.org/10.53377/lq.13561

Maria Chiara Pievatolo, (2017)

La bilancia e la spada: scienza di stato e valutazione della ricerca
https://commentbfp.sp.unipi.it/
maria-chiara-pievatolo-la-bilancia-e-la-spada-scienza-di-stato-e-valutazione-della-ricerca/

Kate Crawford, 2021

Atlas of AI. Power, Politics, and the Planetary Costs of Artificial Intelligence
https://www.agendadigitale.eu/cultura-digitale/
atlas-of-ai-power-politics-and-the-planetary-costs-of-artificial-intelligence/

Simona Levi, 2022

Democratic Digitalisation - Proposal for a Sovereign and Democratic Digitalisation of Europe
https://digitalizacion-democratica.xnet-x.net/

David Gray Widder, Meredith Whittaker, Sarah Myers West, 2023

Open (for business): Big Tech , Concentrated Power ,and the political economy of Open AI
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4543807

Lambert Strether,

Why Must Humans Compete for Electric Power with AI Bullshit Generators Programmed by Ritual Incantations?
https://www.nakedcapitalism.com/2024/09/
why-must-humans-compete-for-electric-power-with-bullshit-generators-programmed-by-ritual-incantat
html