GARANTE PER LA PROTEZIONE DEI DATI PERSONALI

GDPR AWARENESS RAISING CAMPAIGN FOR SMEs

Intelligenza Artificiale e Profilazione

ovvero

(un) mondo come necessità e non discriminazione

ing. Stefano Rinauro Ph.D.
Dip. Tecnologie Digitali e Sicurezza Informatica

Firenze, 22 gennaio 2024

Who am I

Ph.D., Academic research on statistical signal processing (2006-2013)

Data Protection and Cyber Security advisor (2013 -2023)

National Data Protection Officer (2018-2023)

Information Technology Officer (2023-)

Agenda

- Di cosa parliamo quando parliamo di…
- Profilazione: rischi e opportunità
- Caso di studio: necessità e non discriminazione
- Come la normativa aiuta a comprendere necessità e non discriminazione della profilazione AI

Di cosa parliamo quando parliamo di…

Intelligenza

[Treccani]: *Complesso di facoltà psichiche e mentali che consentono all'uomo di pensare, comprendere o spiegare i fatti o le azioni, elaborare modelli astratti della realtà, intendere e farsi intendere dagli altri, giudicare, e lo rendono insieme capace di adattarsi a situazioni nuove e di modificare la situazione stessa quando questa presenta ostacoli all'adattamento*

***NB**: non esiste una definizione di «Intelligenza» comunemente accettata dalla comunità scientifica*

***NB2**: intelligenza ≠ QI*

Di cosa parliamo quando parliamo di…

Intelligenza Artificiale

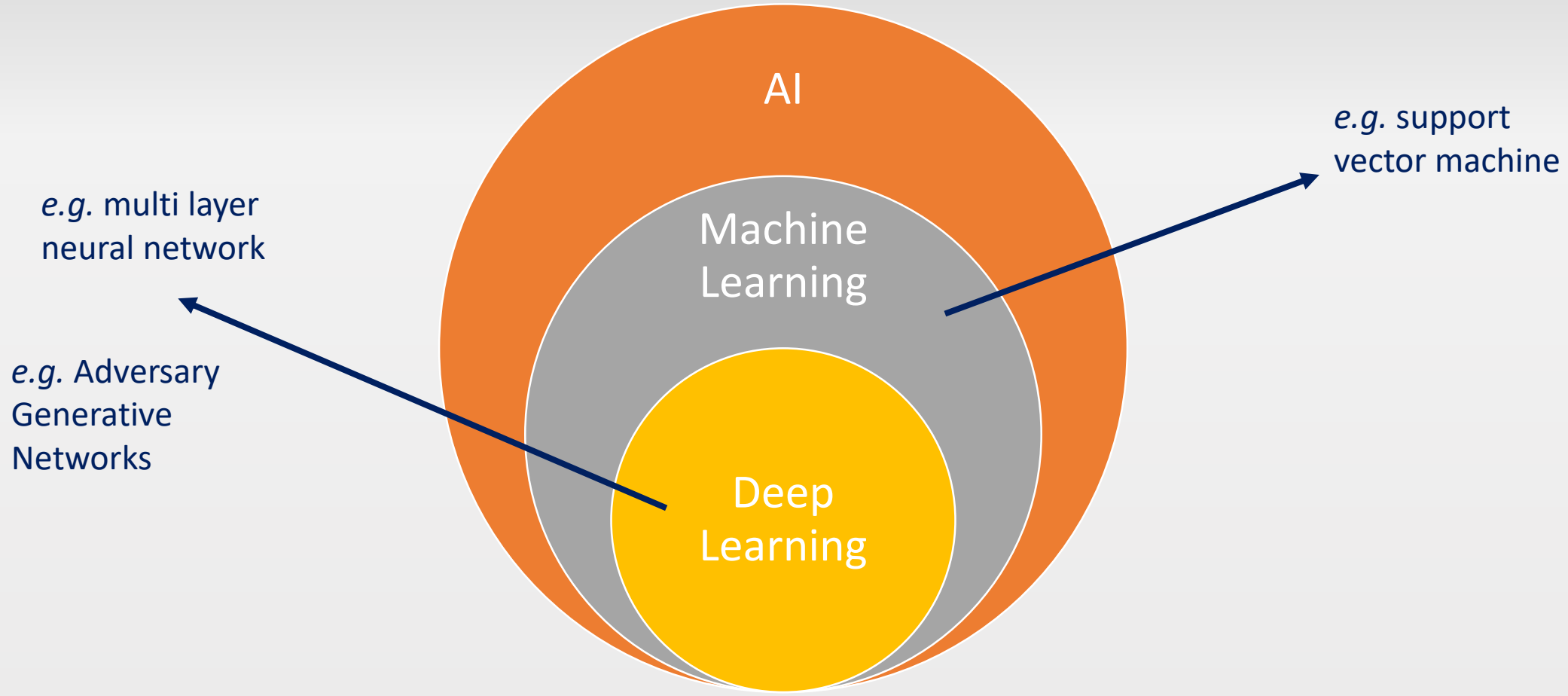
[ISO/IEC 42001:2023]: *capacità di un sistema di mostrare capacità umane quali il ragionamento, l'apprendimento, la pianificazione e la creatività.*

Sistemi in grado di

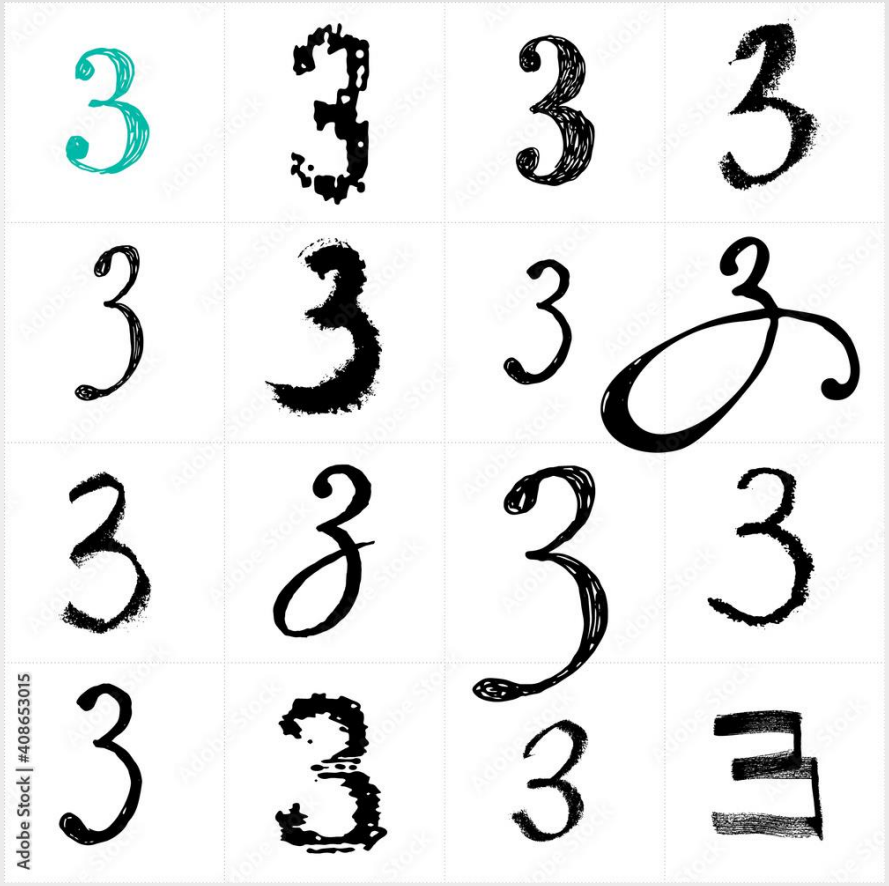
- apprendere e distillare significato e conoscenza dai dati presi in input;
- applicare tali conoscenza per risolvere nuovi problemi in analogia con quanto farebbe un essere umano;
- *e.g.* conversare con un essere umano rispondendo a quesiti, generare immagini e testi originali, prendere decisioni basate sui dati in ingresso e sul contesto di riferimento

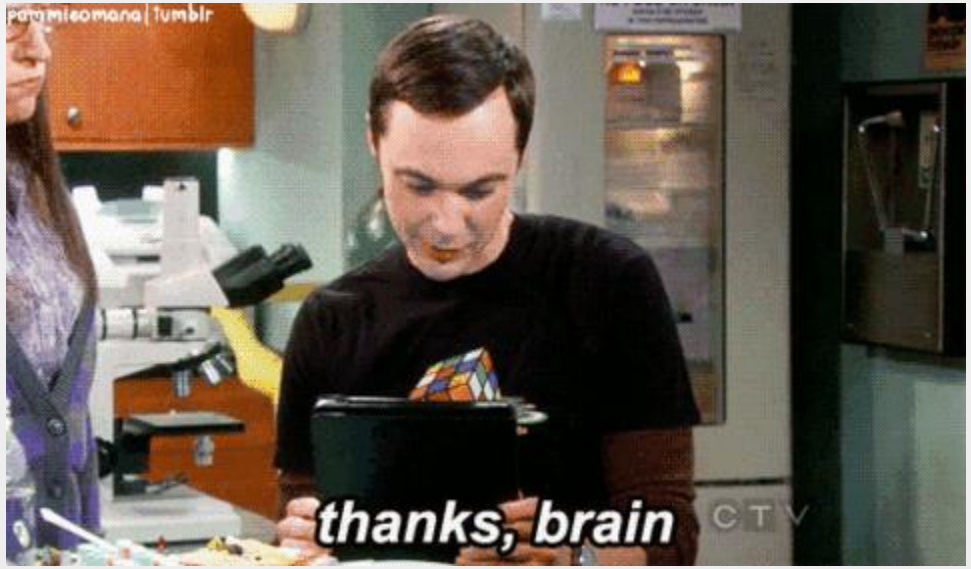
NB. IA Generativa (cfr. GPT) o non generativa

Di cosa parliamo quando parliamo di…

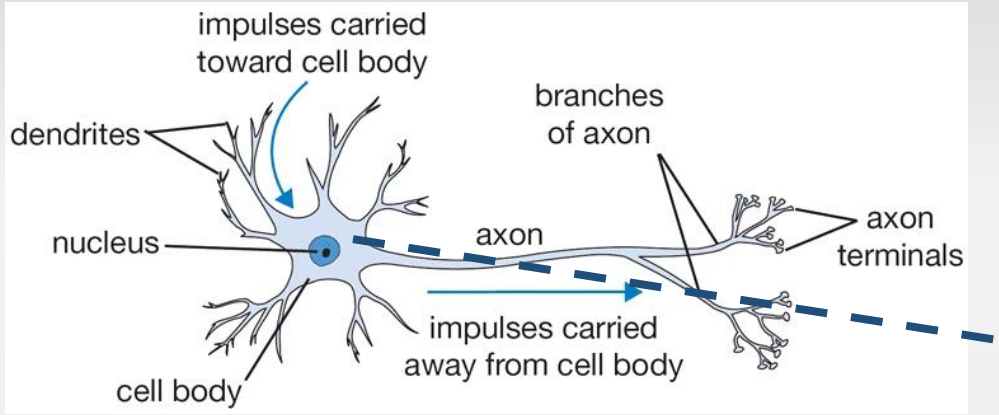


Riconoscimento e classificazione

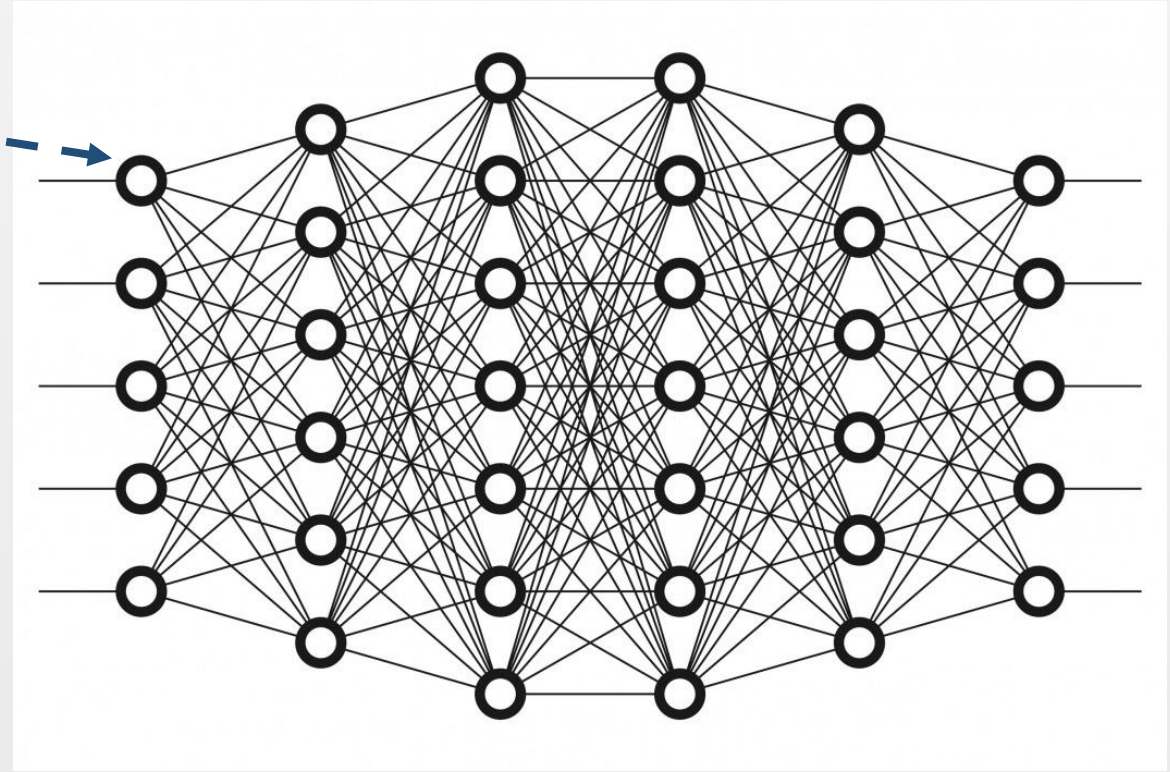




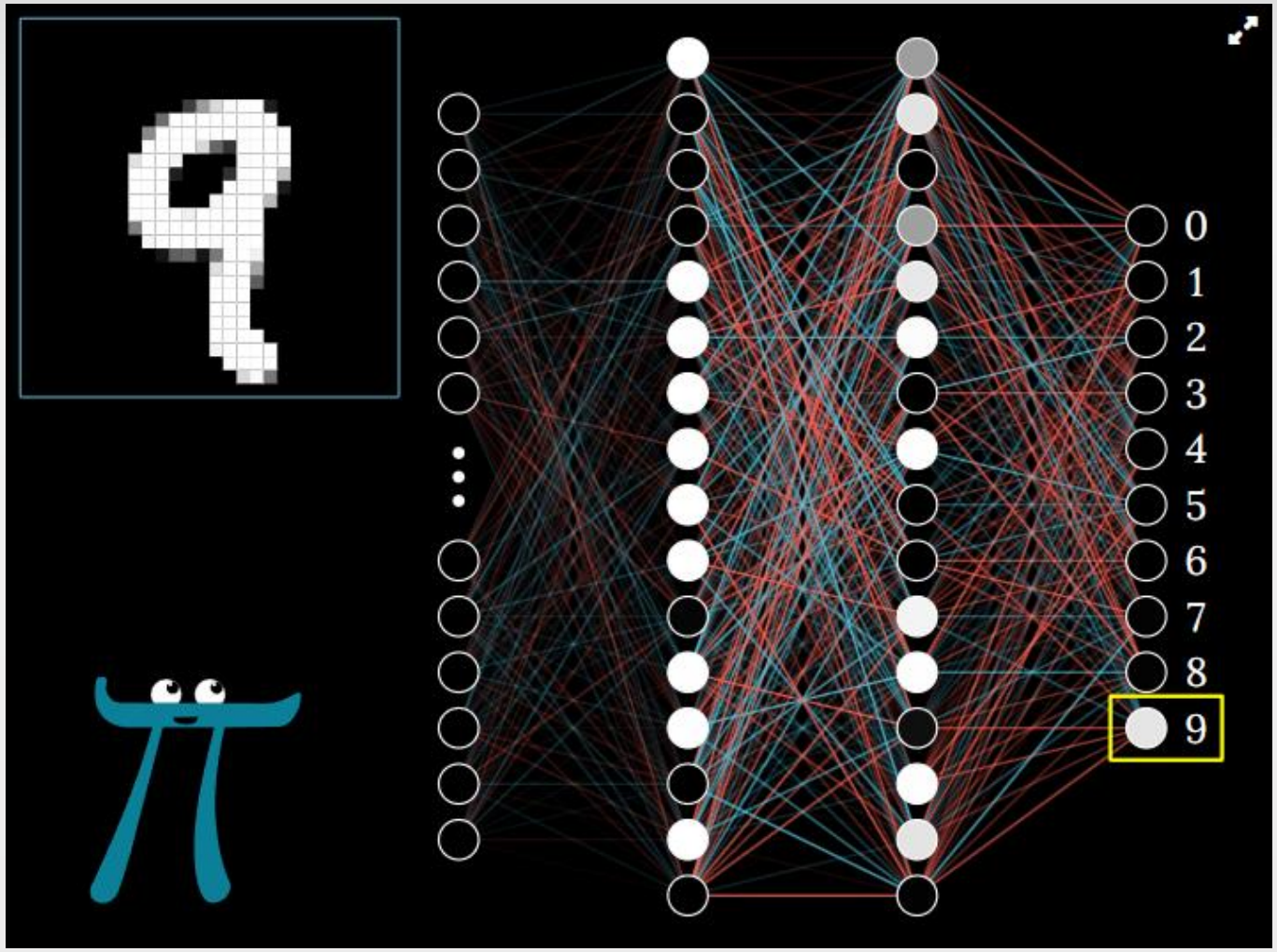
Reti Neurali multi-strato



Source:
gadictos.com

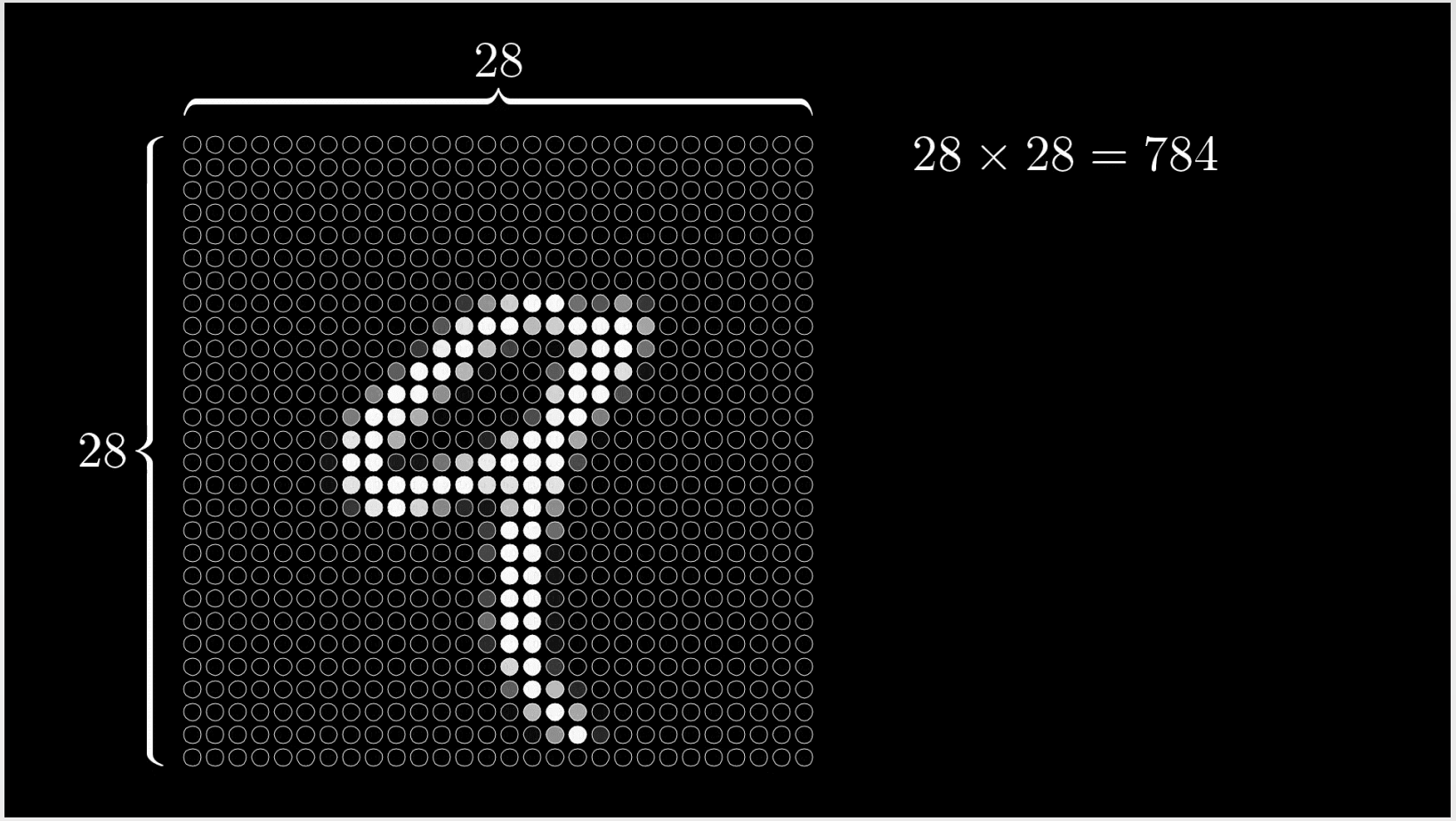


Reti Neurali multi-strato: esempio

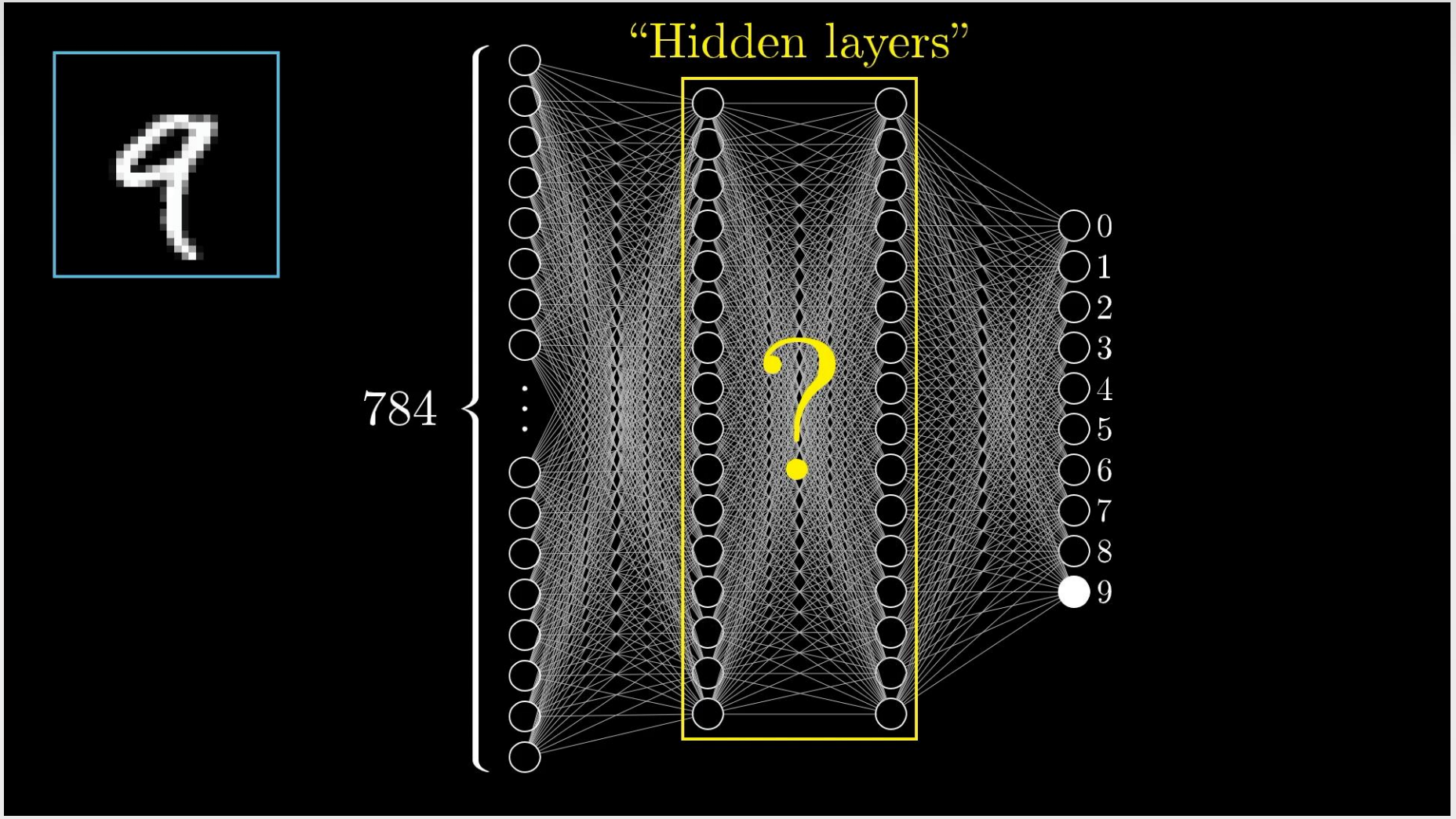


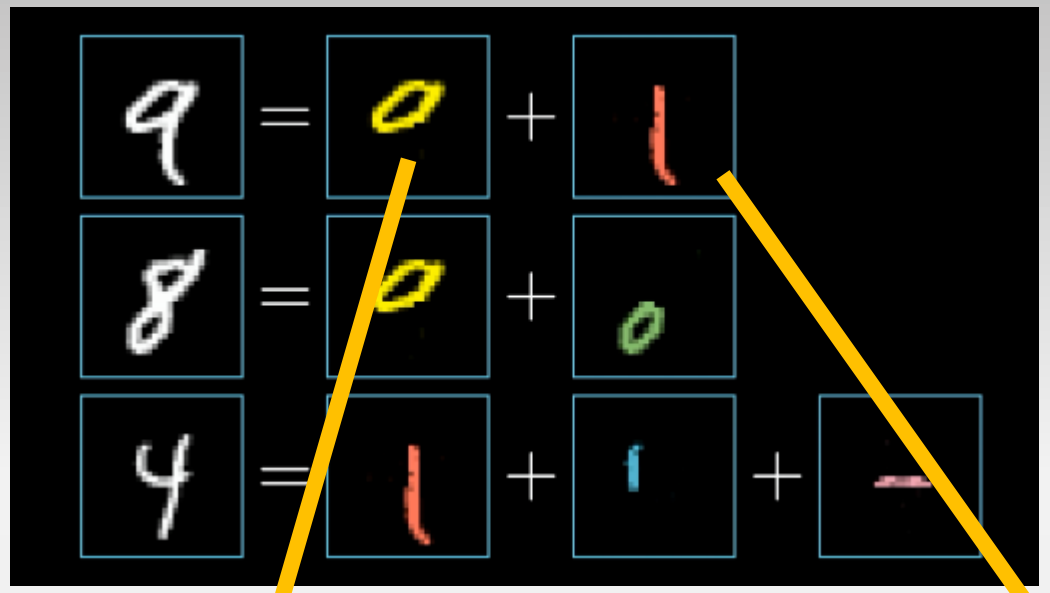
Source:
3Blue1Brown.com

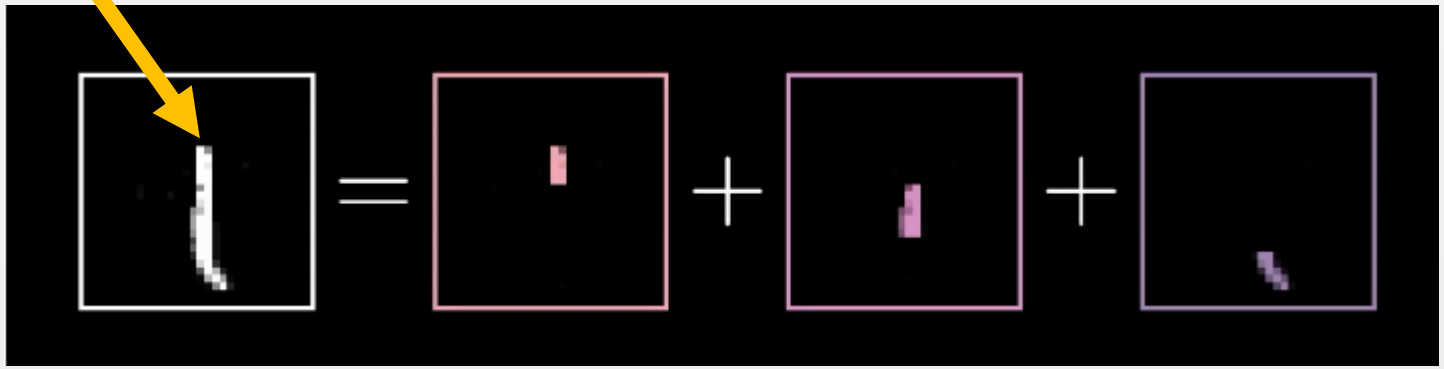
Reti Neurali multi-strato: esempio

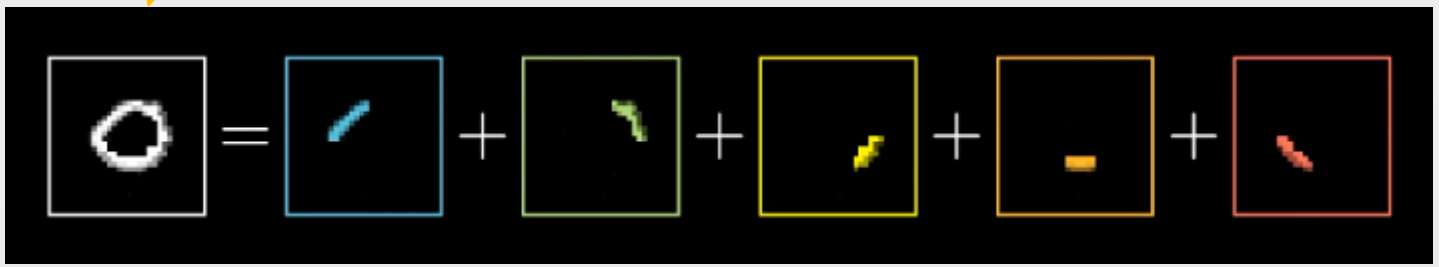


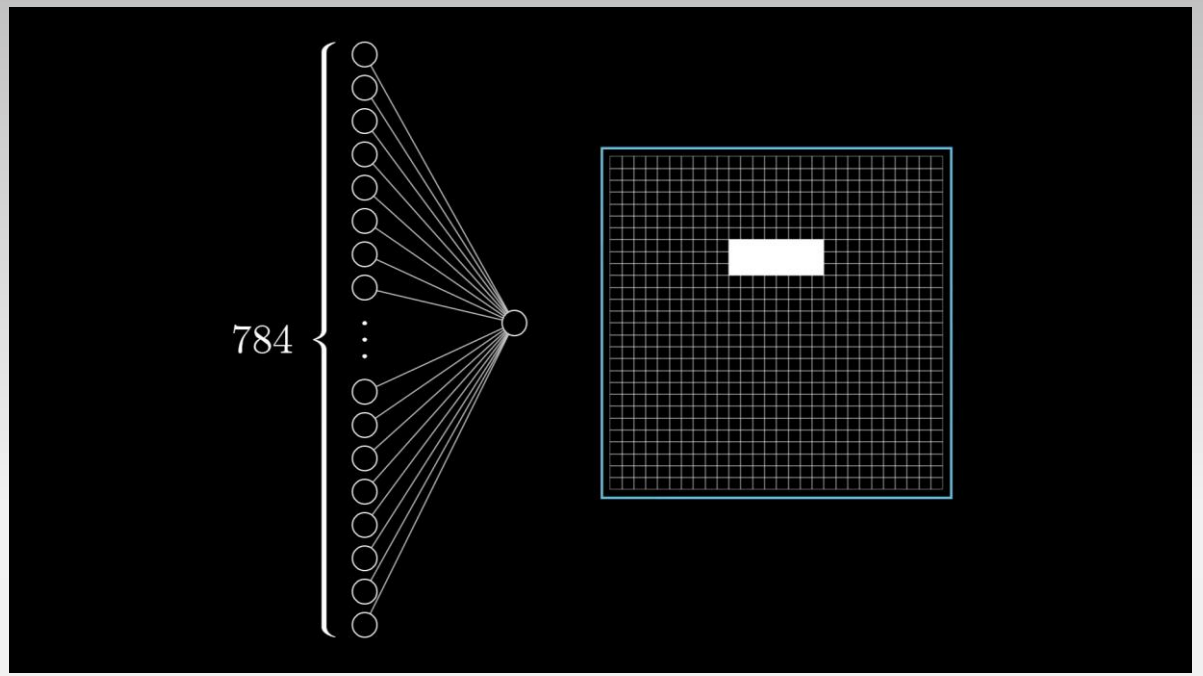
Reti Neurali multi-strato: esempio

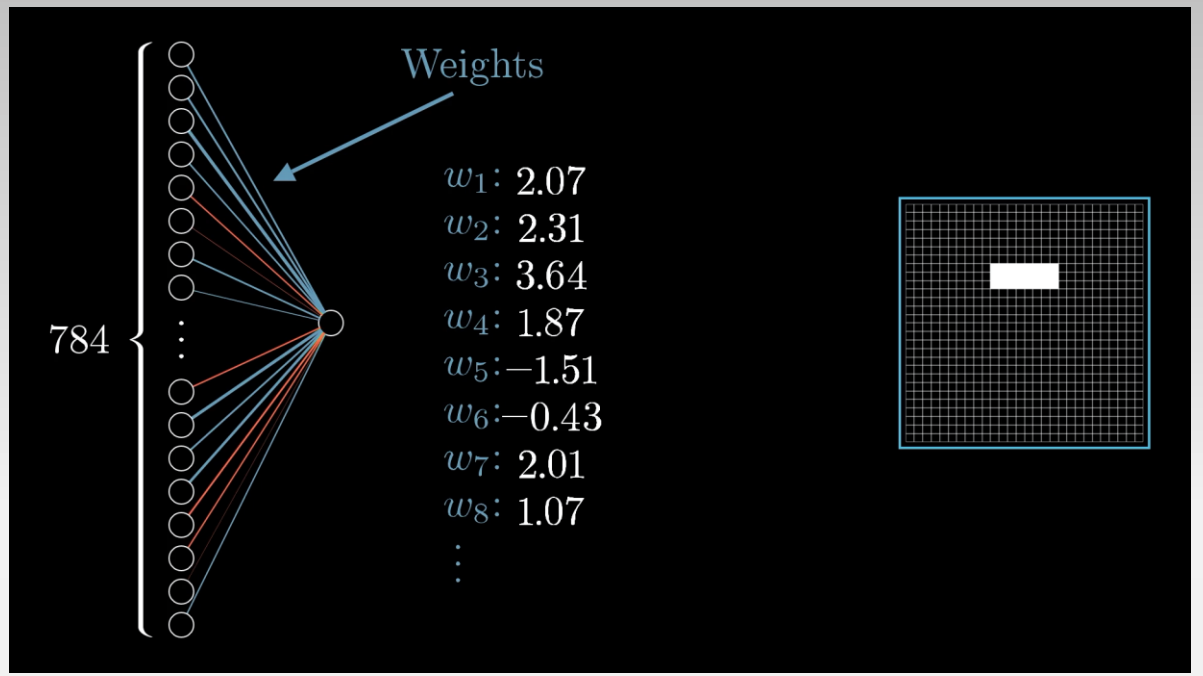


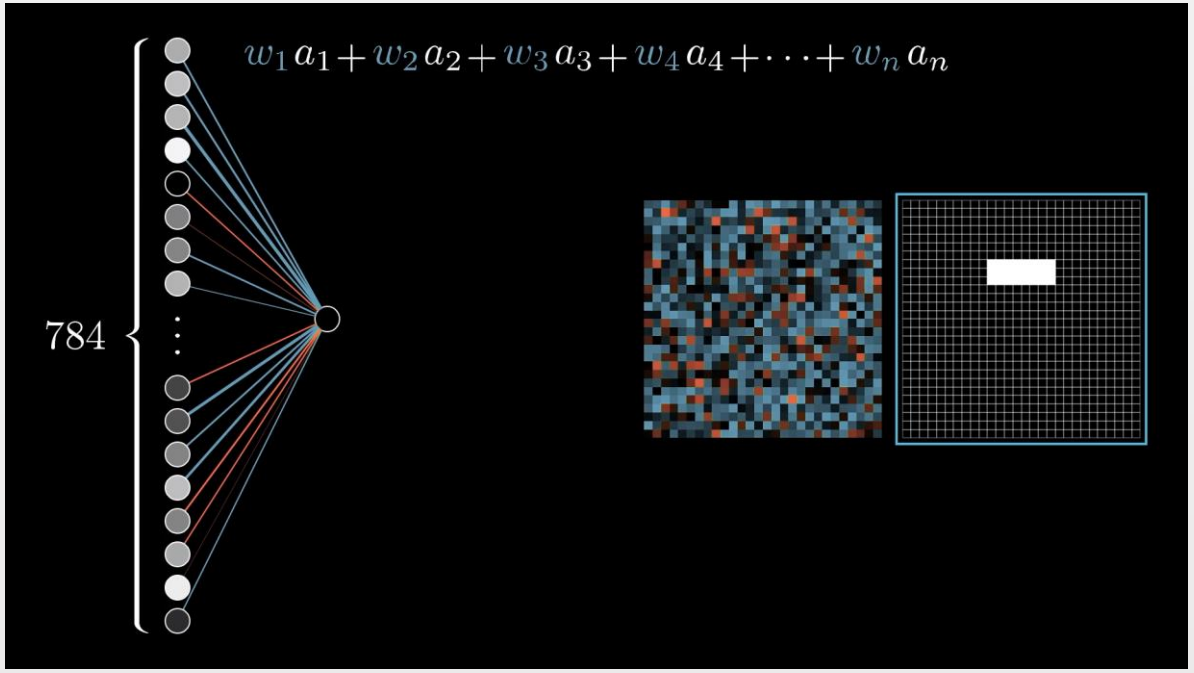


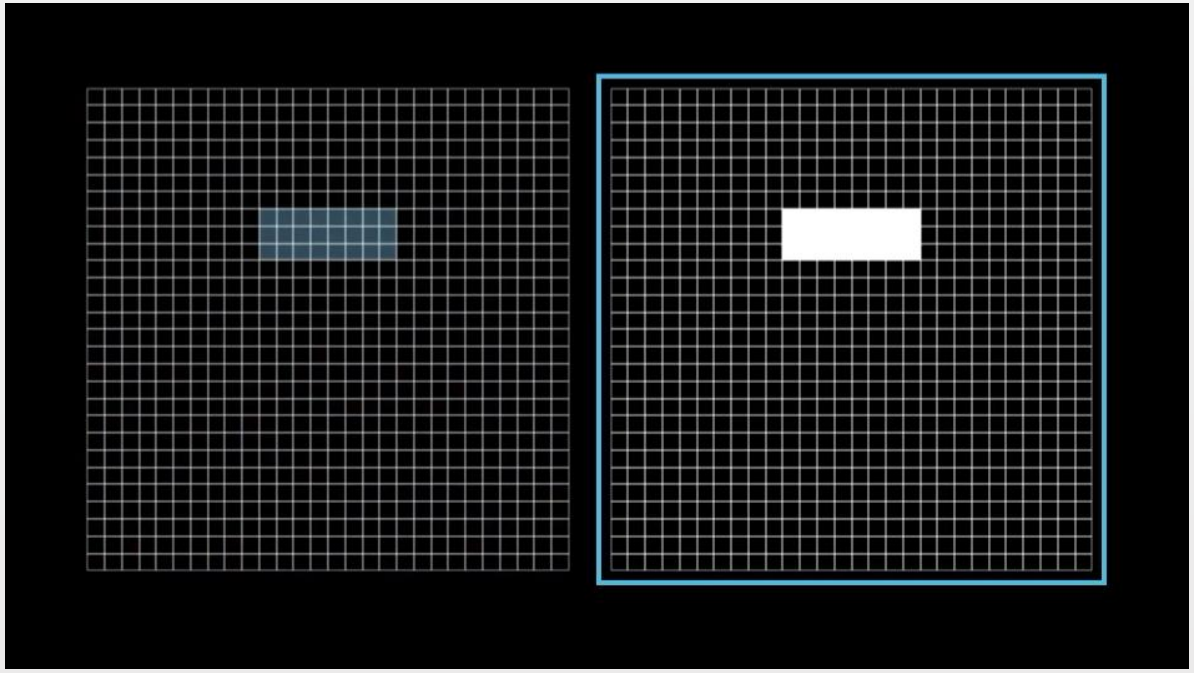




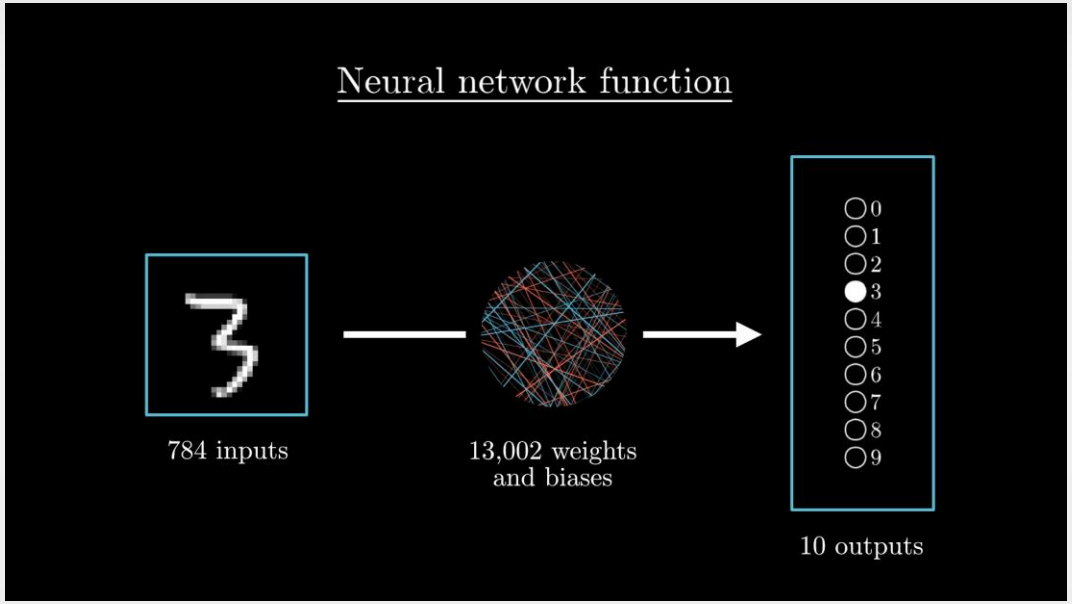

784

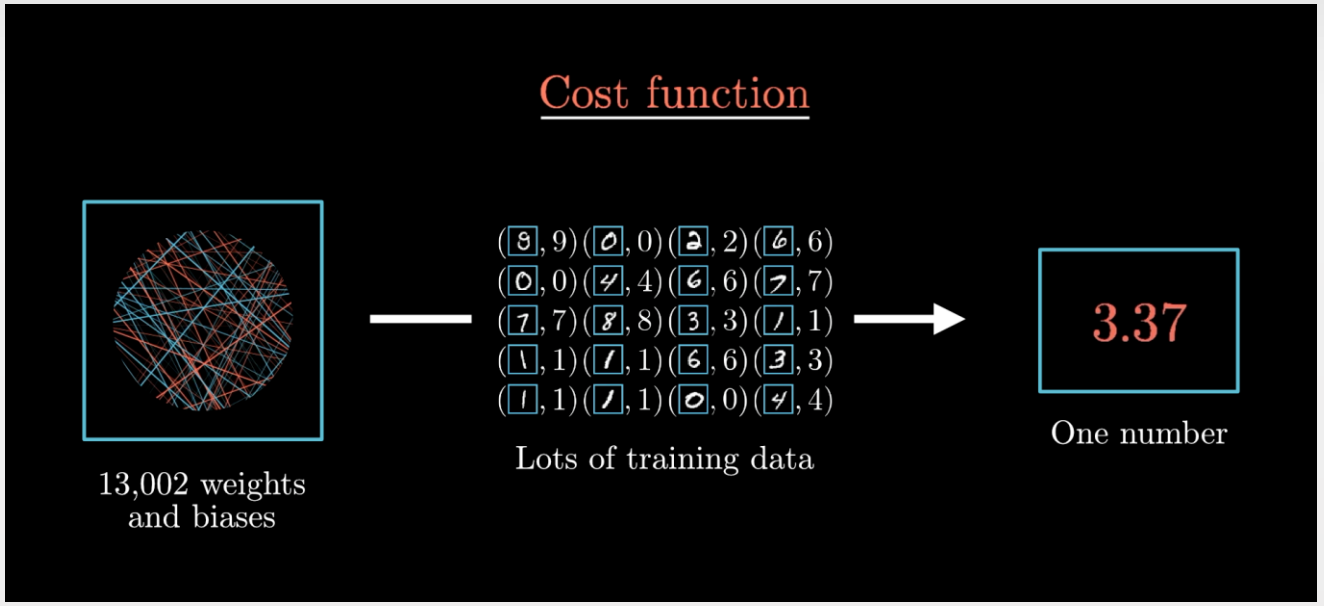

Weights
784
w_1: 2.07
w_2: 2.31
w_3: 3.64
w_4: 1.87
w_5: −1.51
w_6: −0.43
w_7: 2.01
w_8: 1.07


$w_1 a_1 + w_2 a_2 + w_3 a_3 + w_4 a_4 + \cdots + w_n a_n$
784



Come un sistema AI impara…





Di cosa parliamo quando parliamo di…

Algoritmi

- [Treccani]: *procedimento sistematico di calcolo, oggi per lo più destinato a essere eseguito da un automa esecutore quale un computer.*
- Passi operativi per trasformare un input nell'output desiderati
- Procedure deterministiche e neutrali
- Anche al di fuori dell'informatica, molte decisioni sono prese su base algoritmica (*e.g.* medicina, radar, tlc…)
- Un algoritmo classico attua le competenze di chi lo ha scritto

Di cosa parliamo quando parliamo di…

Algoritmi AI

- [Treccani]: *procedimento sistematico di calcolo, oggi per lo più destinato a essere eseguito da un automa esecutore quale un computer.*
- Passi operativi per trasformare un input nell'output desiderati
- Non solo attua le competenze di chi lo ha scritto, ma è in grado di acquisire esperienza
- Si modifica in relazione ai dati elaborati
- Comportamenti non più deterministici o predicibili (ancora neutrali?)

NB. differenze tra algoritmo classico e algoritmo IA

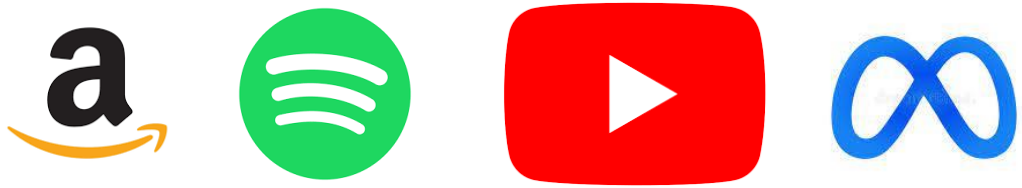
[cfr Consiglio di Stato, Sent. 7892, 2021]: *«mentre nell'algoritmo tradizionale la sequenza di istruzioni è ben definita […] e produce un determinato risultato, nel caso dell'IA […] l'algoritmo elabora ulteriori criteri di inferenza e assume decisioni sulla base di un apprendimento automatico»*

Di cosa parliamo quando parliamo di…

Profilazione

- [GDPR, Art. 4]: *qualsiasi forma di trattamento automatizzato di dati personali consistente nell'utilizzo di tali dati personali per valutare determinati aspetti personali relativi a una persona fisica, in particolare per analizzare o prevedere aspetti riguardanti il rendimento professionale, la situazione economica, la salute, le preferenze personali, gli interessi, l'affidabilità, il comportamento, l'ubicazione o gli spostamenti di detta persona fisica.*

Nella vita di tutti i giorni



Il Buono, il Brutto e il Cattivo

- Ci semplifica la vita (?) (the good)
  - Il sogno dei commerciali e dei mass media anni 90: programmazione ad hoc con pubblicità sartorialmente cucita sul cliente/spettatore;
  - E' un mero scambio commerciale e potrebbe anche andarci bene così… (If you're not paying for the product, then you are the product)
  - …purché non sia usata «contro» di noi

- Ci preclude potenzialità di crescita, di ampliare i nostri orizzonti (forse anche di «svolgere la nostra personalità»?) (the ugly)

- Potrebbe essere alla base di discriminazioni (the bad)

Profilazione e discriminazione

Amazon ditched AI recruiting tool that favored men for technical jobs

Specialists had been building computer programs since 2014 to review résumés in an effort to automate the search process

Amazon's automated hiring tool was found to be inadequate after penalizing the résumés of female candidates. Photograph: Brian Snyder/Reuters

Caso di studio

Julia Dressel and Hany Farid, "The accuracy, fairness and limits of predicting recidivism", Science Advances, 2018 Jan; 4(1)

COMPAS

DOC uses the Correctional Offender Management Profiling for Alternative Sanctions tool, commonly known as COMPAS, for criminogenic risk and needs assessments and unified case planning. This actuarial risk assessment system contains offender information specifically designed to determine their risk and needs and inform dynamic case plans that will guide the offender throughout his or her lifecycle in the criminal justice system.

Caso di studio: contesto

- Sistema penale statunitense
- Software di profilazione per stabilire se lasciare libera su cauzione la persona incriminata o meno
- La decisione del giudice, in diverse corti (*e.g. Wisconsin*) viene supportata da un software di profilazione

`[input]` informazioni personali relative alla persona incriminata

`[output]` fattore di rischio legato alla verosimiglianza che la persona commetta altri crimini in futuro

Caso di studio: criticità

Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[prestazioni software profilazione]	Falso positivo (alto fattore di rischio erroneamente stimato)	Falso negativo (basso fattore di rischio erroneamente stimato)
afrodiscendenti	44.9%	28.1%
caucasici	23.5%	47.7%

Caso di studio: questioni rilevanti

- L'algoritmo mostra polarizzazione razziale

Q1. Perché? Da dove proviene questa polarizzazione?

1. Logica algoritmica
   - `if` nationality == mexican `then` risk+=10

2. Popolazione del campione usato per l'addestramento
   - dati scelti ad hoc con postura discriminante?
   - dati scelti con imparzialità, ma riflesso di un società discriminatoria (o di una storia di discriminazione)?

Q2. L'algoritmo funziona davvero meglio di una persona? E' davvero più neutrale?

Caso di studio: Esperimento per testare Q2

Campione di analisi: 1000 persone delle quali erano note

- informazioni demografiche
- storia criminale
- risk ranking del software
- osservazione del comportamento criminale nei due anni successivi

Partecipanti all'esperimento: 400 persone selezionate in maniera casuale online

Setting dell'esperimento:

`[input]`: lettura di una descrizione recante 7 informazioni su ogni persona nel campione (nessuna informazione sulle origini razziali)

`[output]`: risposta alla domanda «il soggetto commetterà altri crimini nel futuro prossimo?»

Caso di studio: Esperimento per testare Q2

Human assessment

A descriptive paragraph for each of 1000 defendants was generated:

> The defendant is a [SEX] aged [AGE]. They have been charged with: [CRIME CHARGE]. This crime is classified as a [CRIMINAL DEGREE]. They have been convicted of [NON-JUVENILE PRIOR COUNT] prior crimes. They have [JUVENILE- FELONY COUNT] juvenile felony charges and [JUVENILE-MISDEMEANOR COUNT] juvenile misdemeanor charges on their record.

In a follow-up condition, the defendant's race was included so that the first line of the above paragraph read, "The defendant is a [RACE] [SEX] aged [AGE]."

There were a total of 63 unique criminal charges including armed robbery, burglary, grand theft, prostitution, robbery, and sexual assault. The crime degree is either "misdemeanor" or "felony." To ensure that our participants understood the nature of each crime, the above paragraph was followed by a short description of each criminal charge:

[CRIME CHARGE]: [CRIME DESCRIPTION]

After reading the defendant description, participants were then asked to respond either "yes" or "no" to the question "Do you think this person will commit another crime within 2 years?" The participants

The participants were recruited through Amazon's Mechanical Turk, an online crowdsourcing marketplace where people are paid to perform a wide variety of tasks (Institutional Review Board guidelines were followed for all participants). Our task was titled "Predicting Crime" with the description "Read a few sentences about an actual person and predict if they will commit a crime in the future." The keywords for the task were "survey, research, and criminal justice." The participants were paid $1.00 for completing the task and a $5.00 bonus if their overall accuracy on the task was greater than 65%. This bonus was intended to provide an incentive for participants to pay close attention to the task. To filter out participants who were not paying close attention, three catch trials were randomly added to the subset of 50 questions. These questions were formatted to look like all other questions but had easily identifiable correct answers. A participant's response was eliminated from

Caso di studio: risultati esperimento

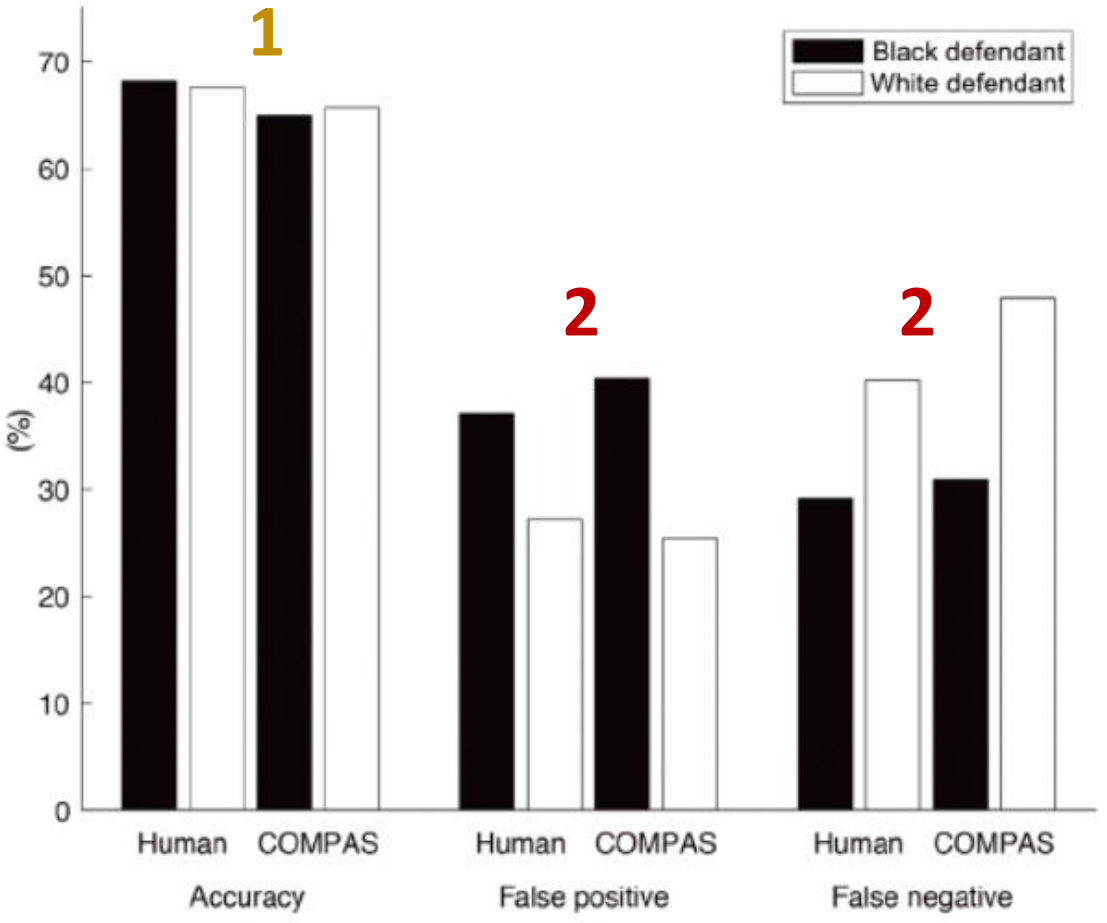


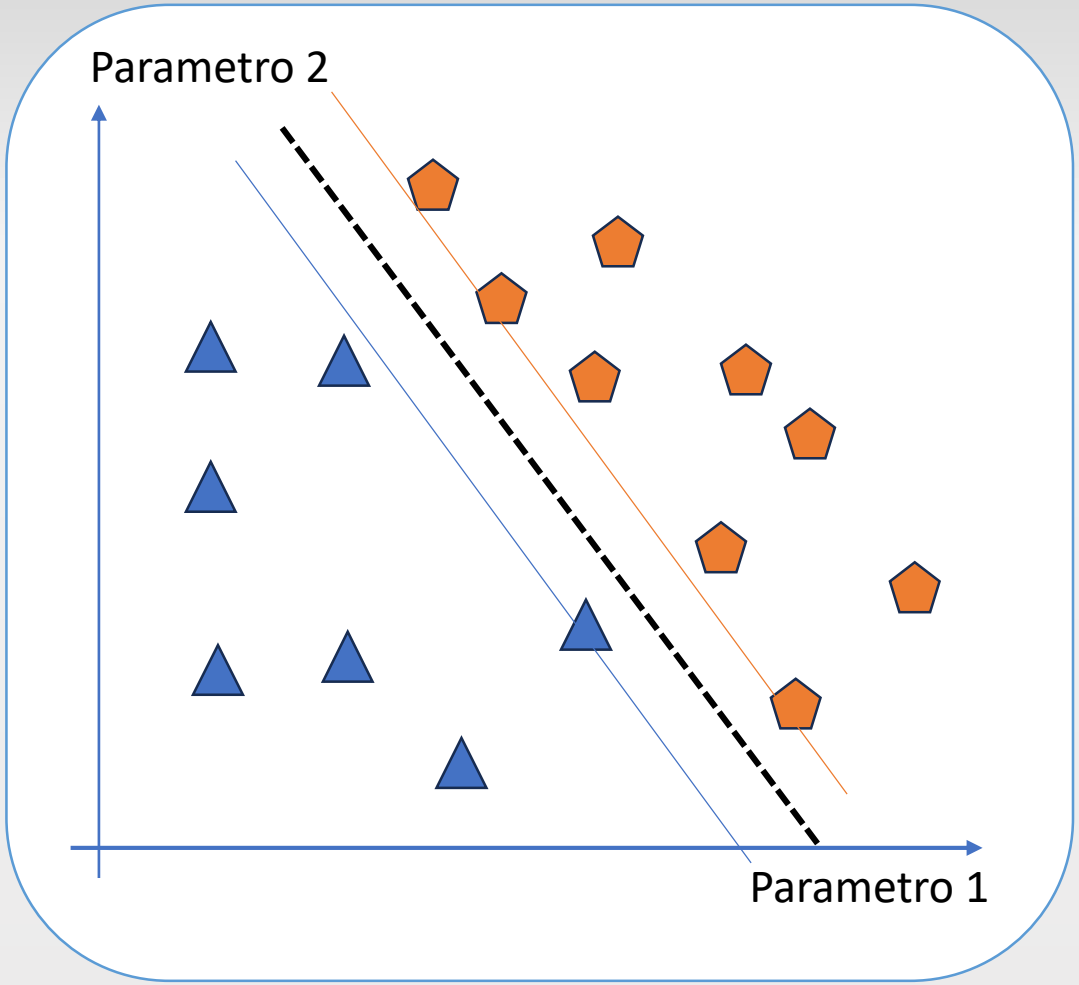
Fig. 1. Human (no-race condition) versus COMPAS algorithmic predictions (see also Table 1).

1. **400 persone selezionate a caso, pagate al massimo 6$, prendono decisioni con la medesima accuratezza del software di profilazione**
2. **I partecipanti mostrano gli stessi errori di pregiudizio razziale del software**

NB. né il software né i partecipanti conoscevano le origini razziali dei soggetti nel campione

Caso di studio: dietro ai risultati

- Per comprendere i risultati è necessario comprendere la logica algoritmica del software di profilazione
- non possibile (software proprietario)
- **Soluzione**: retroingegnerizzazione della logica algoritmica
- **Procedura**: costruzione di classificatori dei soggetti nel campione sulla base dei parametri della descrizione



Caso di studio: dietro ai risultati

- Vengono prodotti classificatori basati su ogni possibile sottoinsieme dei 7 descrittori usati nell'esperimento **Q2**.
- Di ogni classificatore vengono calcolate le prestazioni in termini di accuratezza di classificazione

and the winner is…

Caso di studio: dietro ai risultati

… un classificatore basato solo su due varabili di classificazione:

Età del soggetto e numero di precedenti penali

Accuratezza software commerciale di profilazione	Accuratezza classificatore basato su età del soggetto e numero di precedenti penali
65.2%	66.8%

Caso di studio: dietro ai risultati

- *[cfr. slide n.22]* **Q1**. Perché? Da dove proviene questa polarizzazione?

Analizzando le variabili individuate per la classificazione:

Età del soggetto : non ha nulla a che vedere con questioni razziali

Numero di precedenti penali : in US è noto (per quanto il trend sia in diminuzione) che un afrodiscendente ha molta più probabilità di essere condannato rispetto ad una persona caucasica

Quindi

- L'algoritmo commerciale è **neutrale** (classifica sostanzialmente sulla base dell'età e del numero di precedenti)
- Il set di dati di addestramento è scelto in modo neutrale (e condivisibile)

ma è la logica stessa della profilazione a cristallizzare uno status quo discriminatorio

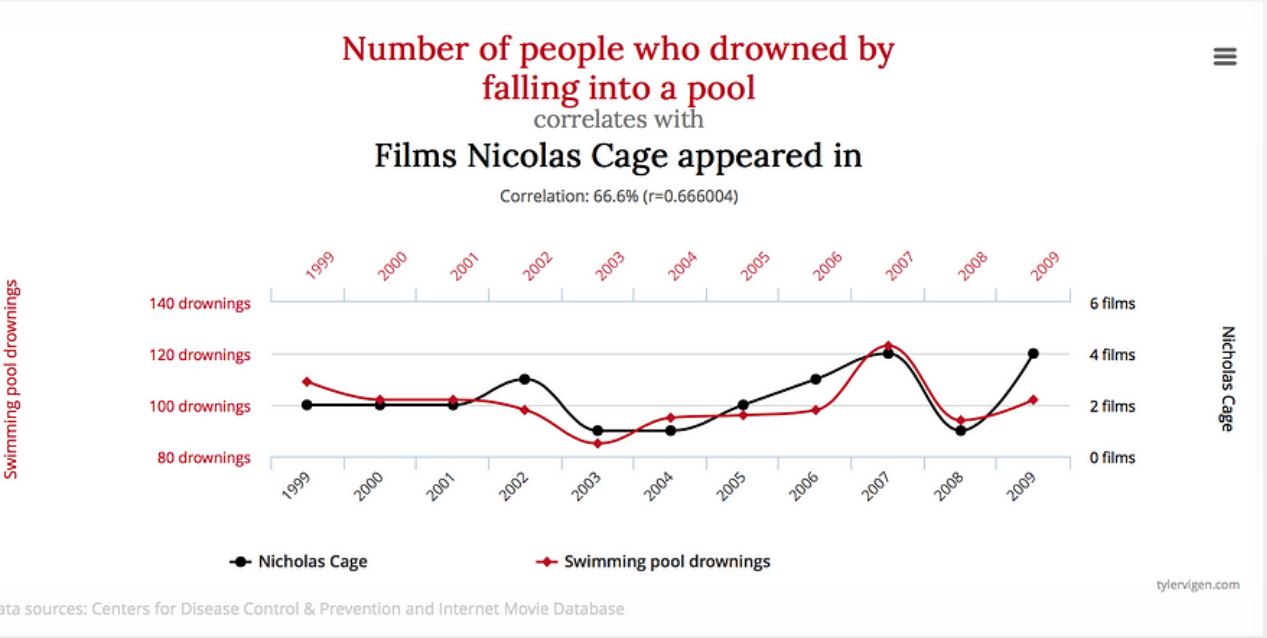
in quanto l'insieme di dati di addestramento è espressione di una società discriminatoria e della storia discriminatoria di quella società

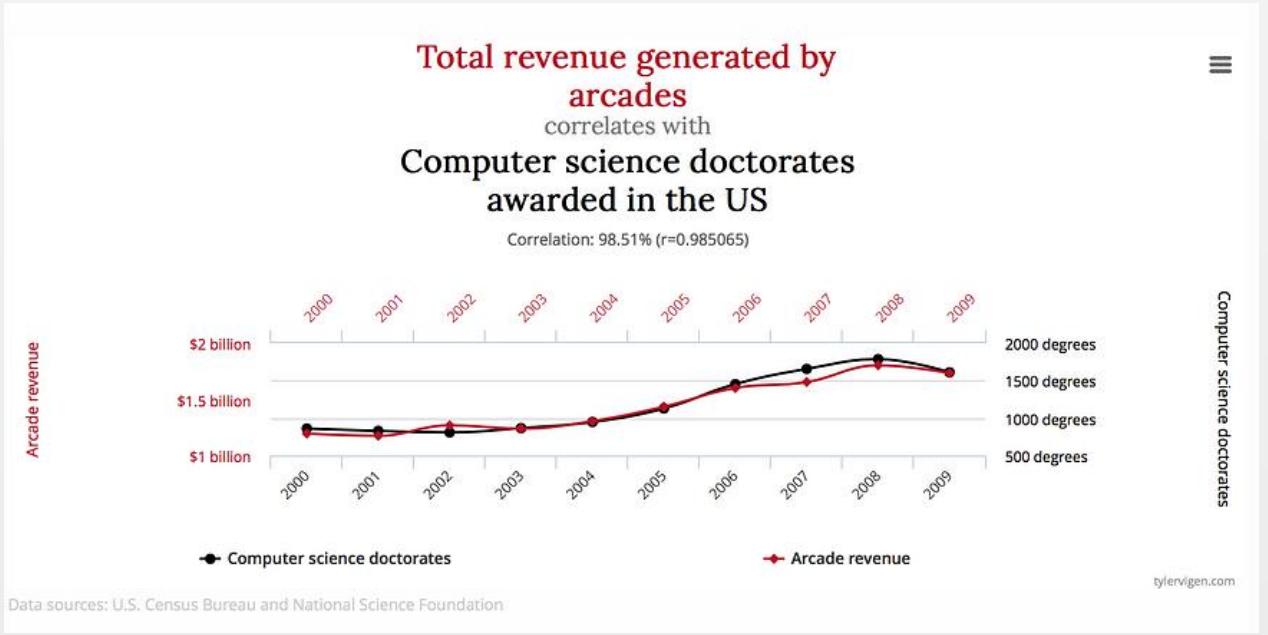
It’s all about the data.

- Insieme di addestramento di dati che potrebbe essere discriminatorio ex ante
- Insieme di addestramento di dati che potrebbe riflettere e cristallizzare fenomeni discriminatori attualmente in essere (è una foto fedele delle realtà, ma se la realtà è discriminatoria…)
- Insieme di addestramento di dati corrotti (data poisoning) da attacchi hacker

It’s all about the data. But data it’s not enough

- ***NB***. «correlation is not causation, nor it implies it»





Source: https://medium.com/@andrewhayes/correlation-does-not-imply-causation-21472b85630f

Intelligenza Artificiale, ovvero non umana

«ma è la logica stessa della profilazione a cristallizzare uno status quo discriminatorio»

La profilazione AI sembra in effetti essere destinata a replicare con alta probabilità schemi sociali e culturali già in essere

Come sfuggire allora da un continuo perpetrarsi di dinamiche discriminatorie?

Con l'intervento umano.

La decisione umana è l'unica, attualmente, in grado di inserire **l'empatia** nel calcolo e pertanto di inserire un **errore di calcolo**, una **mutazione evoluzionistica** in grado di evolvere la società.

Cosa insegna il caso di studio…

E’ **davvero necessario** dotarsi di uno strumento di profilazione AI (costoso, complesso, oscuro, etc..) che sostanzialmente ha performance analoghe a persone prese a caso online e che estrinseca il suo criterio di selezione sulla più banale delle logiche? ***[necessità]***

Non è sempre così semplice capire se e in che misura un insieme di dati di addestramento può o meno risultare discriminatorio. ***[non discriminazione]***

e il GDPR come ci aiuta?

- Valutazione di impatto
- Protezione dei dati personali sin dalla progettazione e per impostazione predefinita
- Processo di decisione basato esclusivamente su un trattamento automatizzato
- Qualità ed integrità dei dati
- Trasparenza
- Analisi dei rischi

Analisi del rischio

Rischio

[ISO]: *effetto dell'incertezza sugli obiettivi*

NB. Rischio ≠ Pericolo

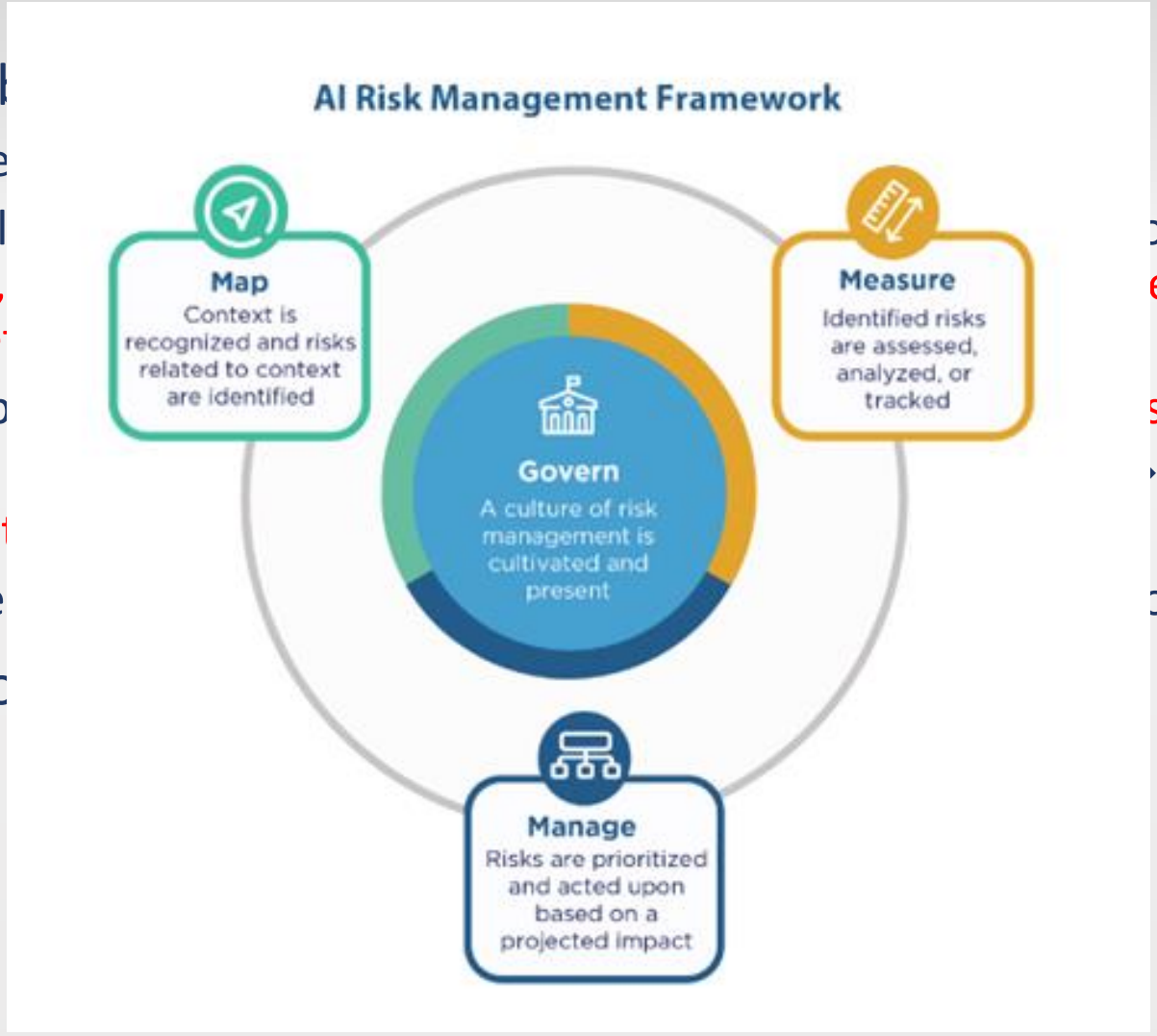
NB2. Rischio = $f\ (prob, impatto)$

- Decidere sulla base dei rischi implica:
  - Comprendere il contesto
  - Identificare gli scenari di rischio (sorgenti di rischio e come si attivano)
  - Valorizzare i possibili impatti
  - Valorizzare la probabilità (verosimiglianza statistica) dello scenario
  - Ottenuti i livelli di rischio dei vari scenari capire come gestire il rischio

Analisi del rischio

- Decidere sulla l
  - Comprendere
  - Identificare gl
    bias cognitivi,
    sui risultati, e
  - Valorizzare i p
  - Valorizzare la
    storiche, valu
  - Ottenuti i live

NB. Il rischio, in c



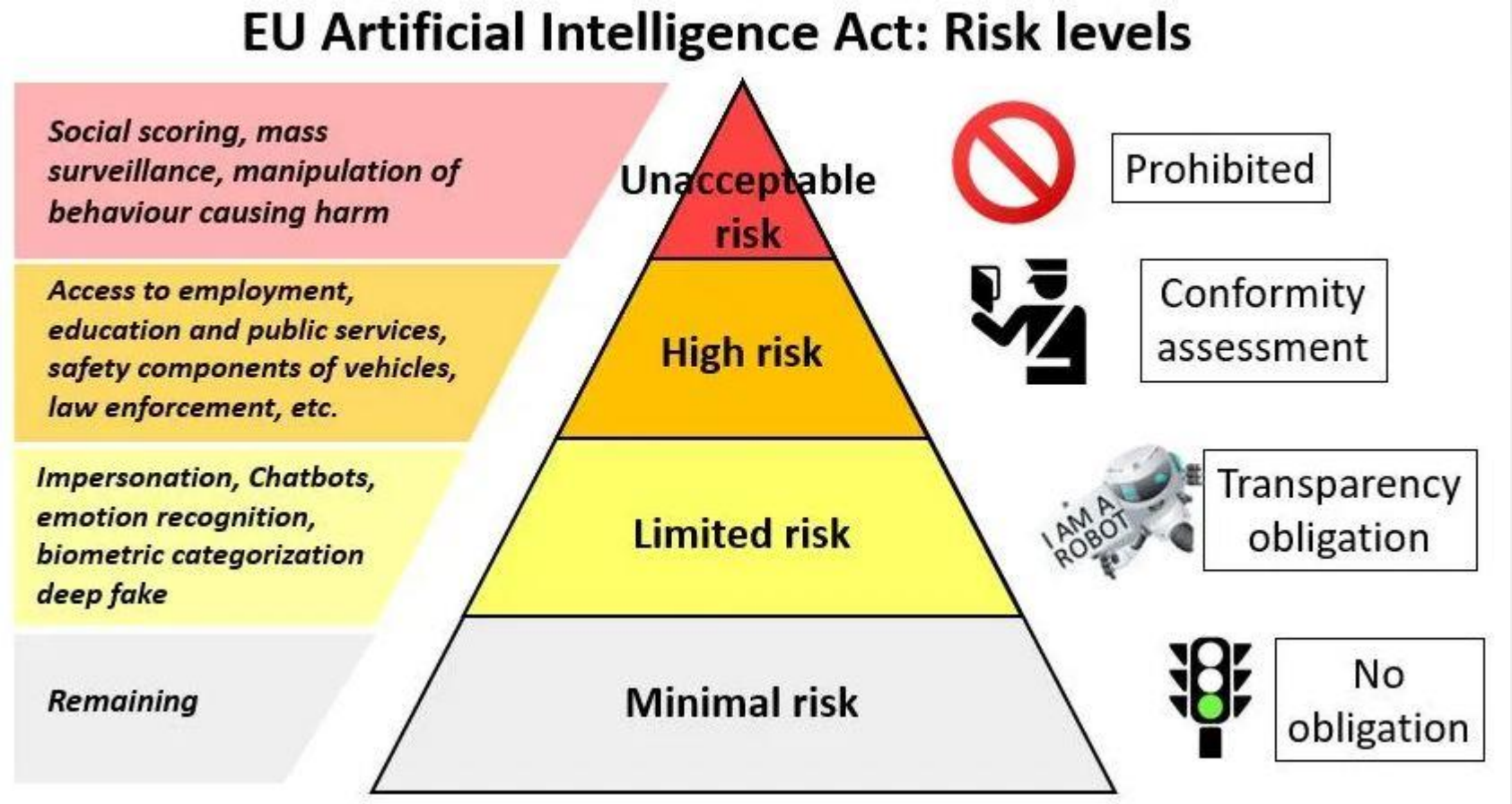
o) → tecnologie in uso,
esso, perdita di controllo

ssati

studi e ricerche, serie

o

e l’AI ACT?


EU Artificial Intelligence Act: Risk levels
Social scoring, mass surveillance, manipulation of behaviour causing harm
Access to employment, education and public services, safety components of vehicles, law enforcement, etc.
Impersonation, Chatbots, emotion recognition, biometric categorization deep fake
Remaining
Unacceptable risk
High risk
Limited risk
Minimal risk
Prohibited
Conformity assessment
I AM A ROBOT
Transparency obligation
No obligation

Not every IA is bad

INTELLIGENZA ARTIFICIALE

DeepMind svela la struttura delle proteine

WILL DOUGLAS HEAVEN • 24 LUGLIO 2021 • 4 MINUTE READ

L'azienda ha già utilizzato la sua intelligenza artificiale per il ripiegamento delle proteine, AlphaFold, per generare strutture per il proteoma umano, oltre a lieviti, moscerini della frutta, topi e altro ancora.

DANNY GOLD SECURITY AUG 16, 2018 6:00 AM

Saving Lives With Tech Amid Syria's Endless Civil War

The Bashar al-Assad regime's indiscriminate air strikes have terrorized civilians for years. Now a small band of activist-entrepreneurs is building a sensor network that listens for warplanes and warns people when and where the bombs will fall.

Conclusioni

- Da grandi poteri derivano grandi responsabilità
- Valutare la necessità di dotarsi di strumenti di intelligenza artificiale per effettuare valutazioni
- Il GDPR ha già in sé strumenti e approcci per deflazionare rischi di discriminazione legati alla profilazione AI
- AI Act come prossimo pilastro di regolamentazione basato sul rischio