

Introduzione al machine learning e all'interpretabile AI



Giro presentazioni

Programma di oggi

Mattina:

Introduzione a concetti chiave machine learning, explainable AI e active learning

Pomeriggio:

Sessione pratica con esempi Python, discussione progetto AITravel

Materiali utilizzati durante la giornata disponibili via Github.

Introduzione

Prima parte:

Introduzione divulgativa all'intelligenza artificiale e all'apprendimento automatico.

Seconda parte:

Aspetti pratici legati alla messa in produzione di un algoritmo di apprendimento automatico.

Prima parte

Introduzione all'AI e al machine
learning

Macchine pensanti e il gioco dell'imitazione

La questione filosofica

Il concetto di macchina pensante e' stato affrontato nel corso dei secoli in **ambito filosofico**.

La nascita dei computer moderni ha naturalmente portato la questione alla ribalta.

Alan Turing fu il primo a teorizzare l'architettura di un computer moderno e in seguito si soffermò sul quesito se tale macchina potesse eventualmente pensare.

COMPUTING MACHINERY AND INTELLIGENCE

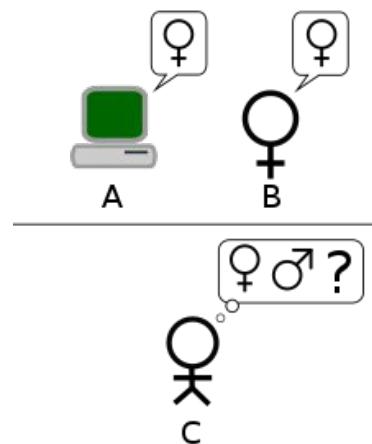
By A. M. Turing

1. The Imitation Game

I propose to consider the question, "Can machines think?" This should begin with definitions of the meaning of the terms "machine" and "think." The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words "machine" and "think" are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, "Can machines think?" is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.



Alan Turing
(1912-1954)

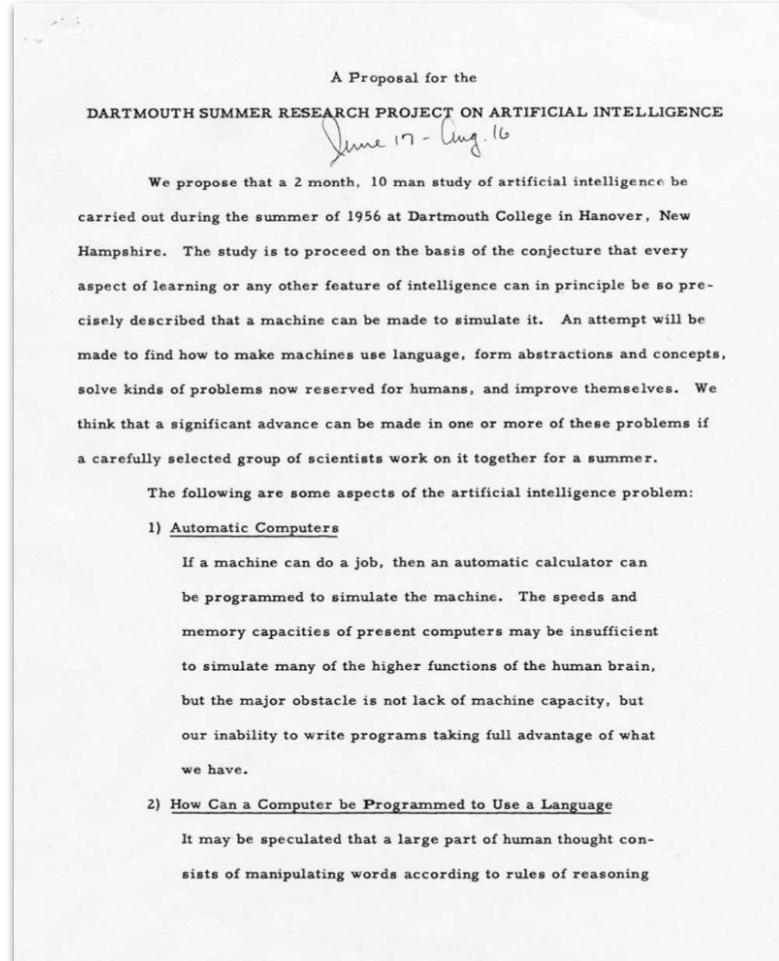


Da macchine pensanti a intelligenza artificiale

La conferenza di Dartmouth del 1956

È considerato l'evento che segna la nascita della nozione moderna di intelligenza artificiale.

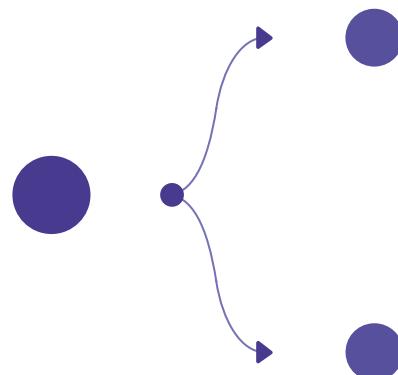
«Lo studio procederà sulla base della congettura per cui, in linea di principio, ogni aspetto dell'apprendimento o una qualsiasi altra caratteristica dell'intelligenza possano essere descritte così precisamente da poter costruire una macchina che le simuli.»



Una distinzione

Tipi di intelligenza artificiale

Intelligenza artificiale



Generale (forte)

Insegnare ad una macchina a ragionare come una persona

Debole

Insegnare ad una macchina a risolvere problemi specifici

Primi tentativi e fallimenti

L'inverno dell'IA

Dopo la conferenza di Dartmouth:

Entusiasmo e ottimismo nel potenziale della tecnologia accompagnato da risorse computazionali decisamente limitate.

Negli anni '70 si venne a creare l'idea che l'IA fosse una tecnologia applicabile solamente a problemi giocattolo → Perdita di fondi e investimenti in ricerca.

“In no part of the field have the discoveries made so far produced the major impact that was then promised”

Lighthill Report
(1973)

L'avvento dei sistemi esperti

Anni '80: Primi utilizzi commerciali dell'IA

Sistema esperto:

Programma che cerca di replicare le prestazioni di una o più persone esperte in un certo dominio usando ad esempio regole.

Primi successi commerciali, nel 1986 *Digital Equipment Corporation* dichiarò di risparmiare tramite un sistema di questo tipo 40 milioni dollari all'anno

Principale problema:

Per disegnare e modificare un sistema esperto è necessaria la conoscenza di una persona esperta del settore in cui si opera.

Esempio regola:

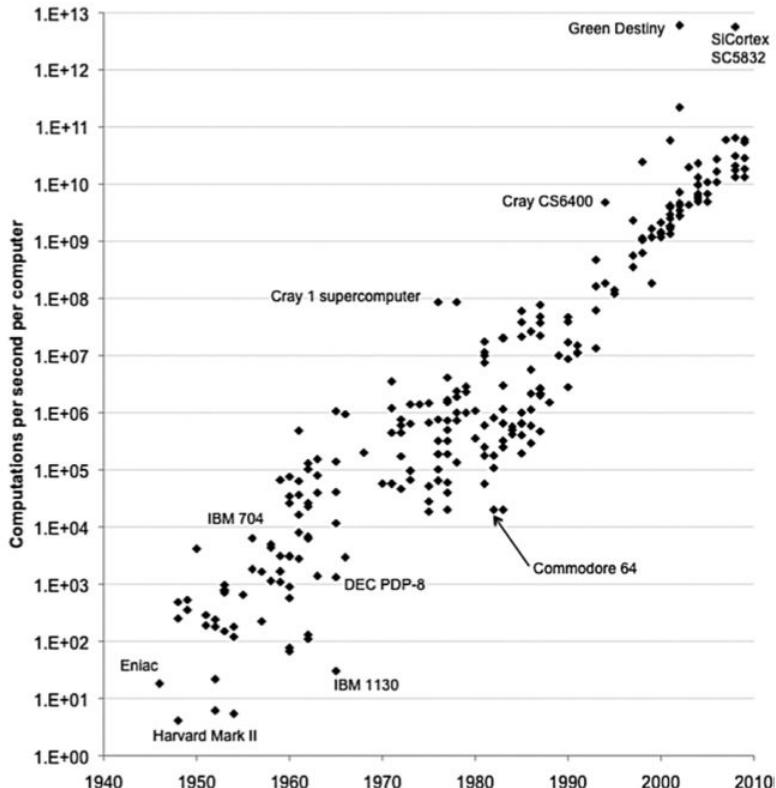
```
IF ((mal di testa)
    AND (raffreddore)
    AND (temperatura >= 38))
THEN (INFLUENZA).
```

Cambiamento di paradigma

Risorse computazionali

Capacità di calcolo è mediamente raddoppiata ogni 18 mesi negli ultimi 60 anni.

Utilizzo di schede grafiche ha ulteriormente accelerato questo fenomeno.



<https://ourworldindata.org/technological-progress>

Cambiamento di paradigma

Dati

Avvento di internet, social media, smartphones e IoT ha causato un'esplosione dei dati che ciascuna persona genera mediamente.

Si stima che ogni individuo generi mediamente 1,7 MB al secondo.

Il 90% dei dati mondiali sono stati creati negli ultimi 2 anni.

<https://techjury.net/blog/how-much-data-is-created-every-day/>

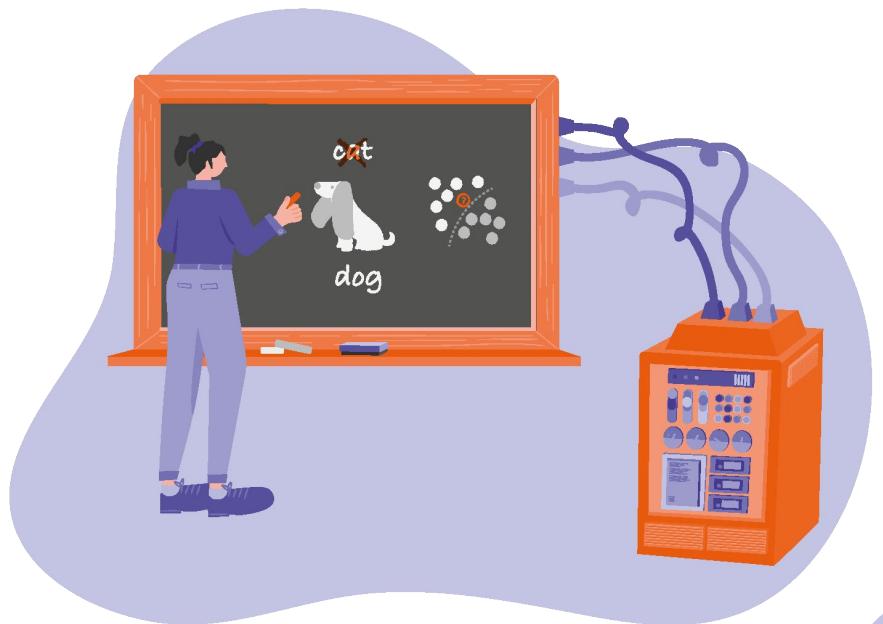


L'apprendimento automatico (machine learning)

Il machine learning è lo studio di algoritmi disegnati per imparare a risolvere compiti specifici attraverso l'**esperienza**.

Dati \Rightarrow esperienza

Risorse computazionali \Rightarrow abilità di imparare



L'apprendimento automatico supervisionato

Apprendimento supervisionato, o apprendimento tramite **esempi**, rappresenta applicazione più comune dell'apprendimento automatico.

X



Y

L'apprendimento automatico supervisionato

| Input X | Output Y |
|------------------|------------------------------|
| email | spam/non spam |
| radiografia | paziente malato/non malato |
| testo in inglese | testo in italiano |
| dati cliente | prestito accettato/rifiutato |

Esempio pratico

Problema: vogliamo sviluppare un modello che aiuti un'agenzia immobiliare a stabilire il prezzo di vendita di un immobile che sta per essere immesso nel mercato

Input: Dati immobile

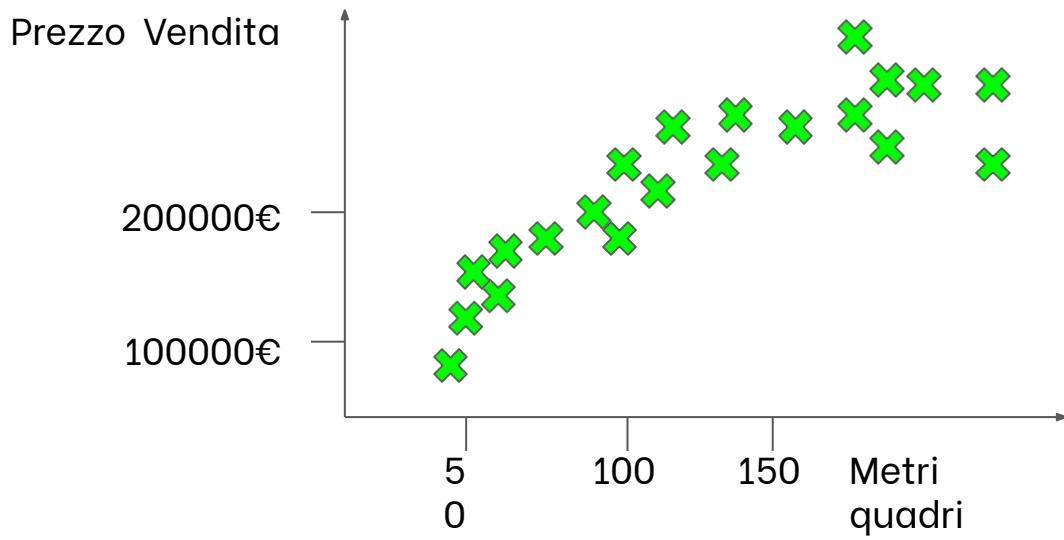
Output: Prezzo vendita

| ID immobile | locali | tipo | Superficie (m2) | bagni | giardino | Prezzo |
|-------------|--------|-------|-----------------|-------|----------|--------|
| 1 | 4 | villa | 95 | 1 | si | ??? € |

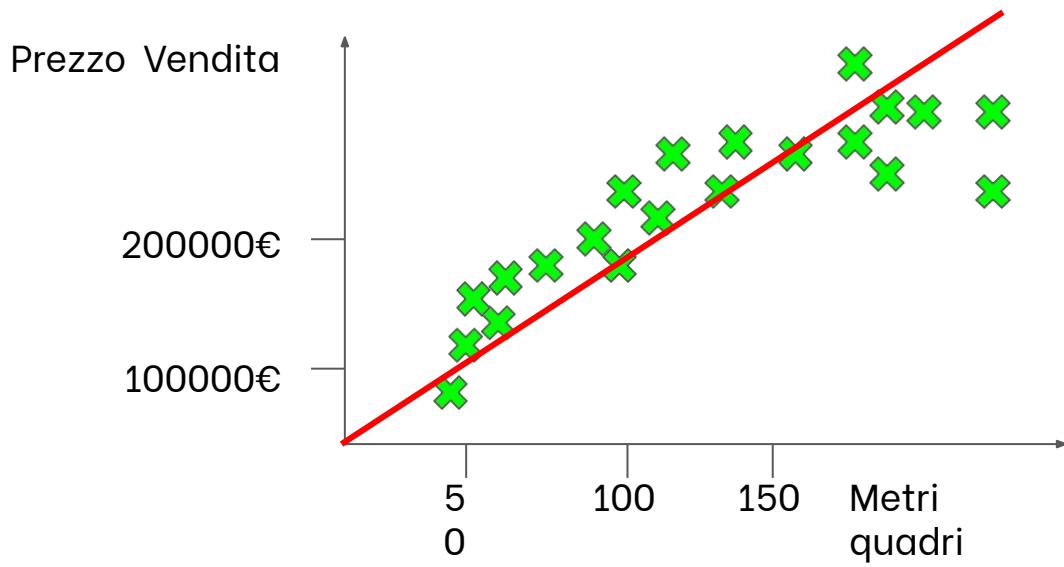
Base di partenza: **dati storici**

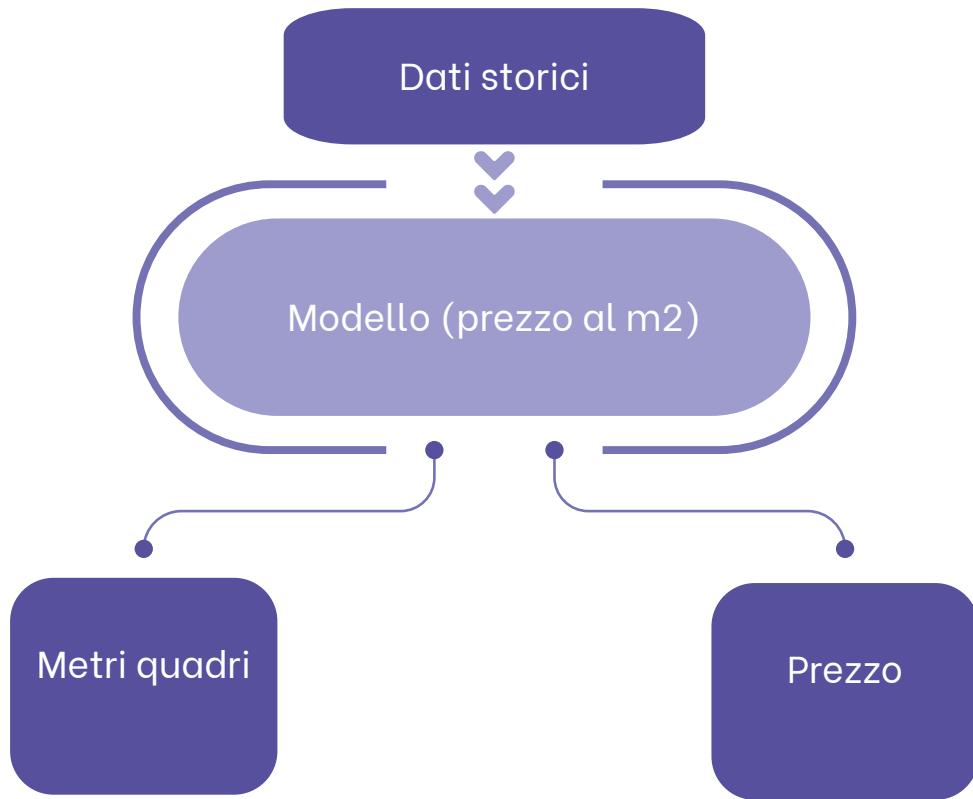
| ID immobile | tipo | locali | superficie (m2) | bagni | giardino | venduto a |
|-------------|--------------|--------|-----------------|-------|----------|----------------|
| 1 | villa | 4 | 120 | 2 | si | 350000€ |
| ... | appartamento | 3 | 100 | 1 | no | 200000€ |
| N | appartamento | 2 | 60 | 1 | no | 90000€ |

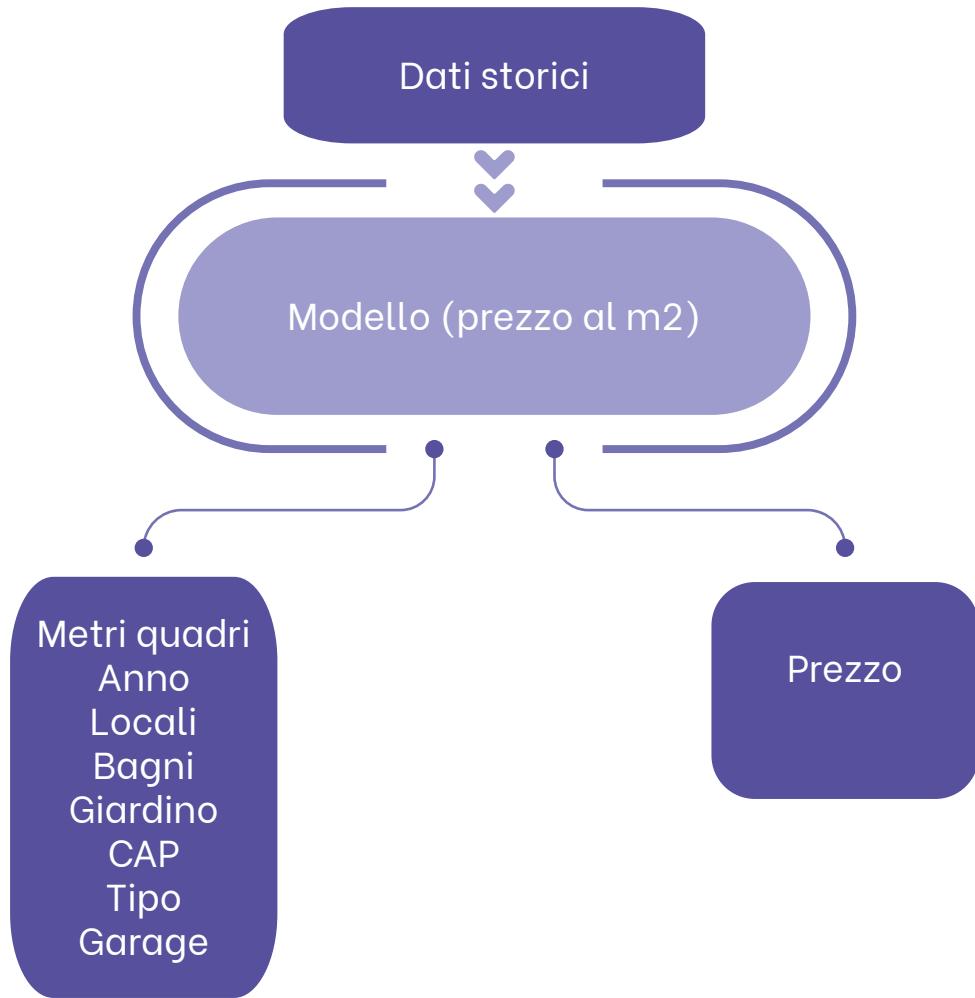
Prima intuizione: relazione metri quadri → prezzo vendita



$$\text{Prezzo vendita} = (\text{Prezzo al m}^2) * \text{m}^2$$







Modelli lineari

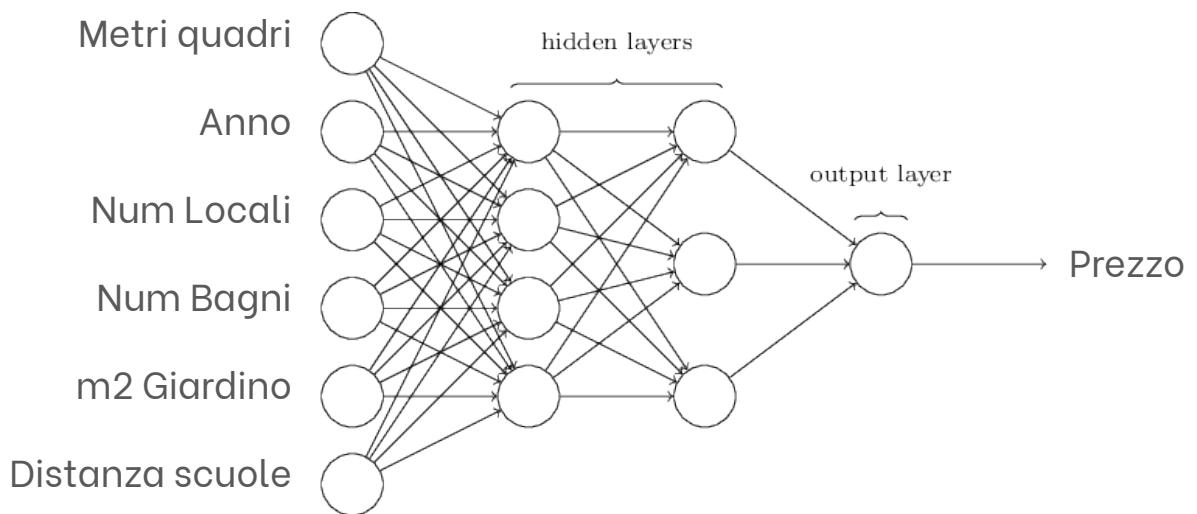


Ogni collegamento rappresenta la pendenza della retta associata alla rispettiva variabile.

“**Addestrare**” un modello di questo tipo significa trovare il valore di questi coefficienti che approssimino meglio i dati storici.

Nonostante la relativa semplicità, modelli come quelli lineari sono **tra i più utilizzati** quando si parla di machine learning in produzione.

Reti neurali

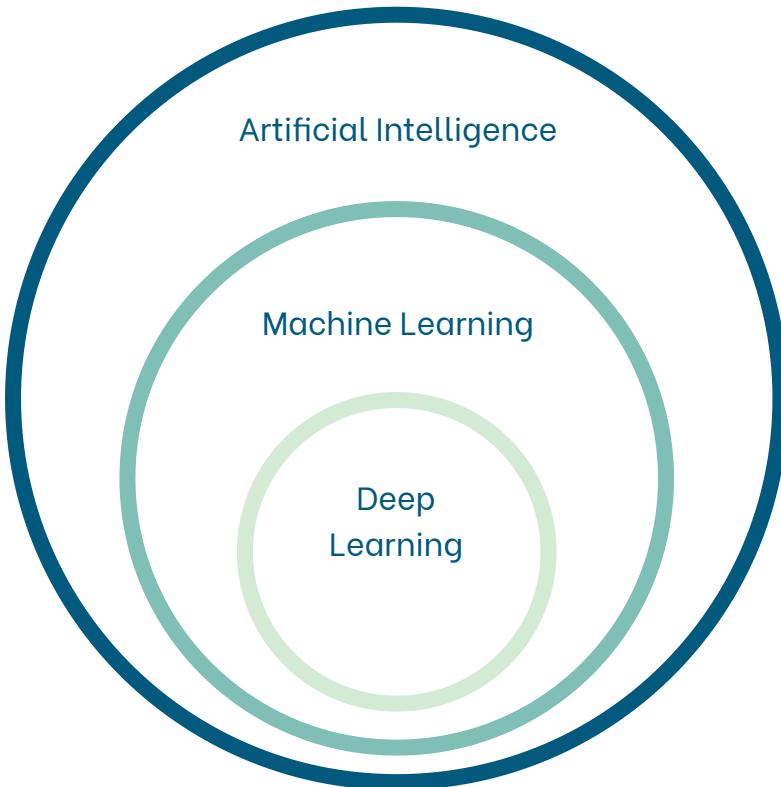


Deep learning

“Le reti profonde”

Il termine deep learning viene usato quando in presenza di reti neurali caratterizzate da un alto numero di strati.

Salito alla ribalta a partire dal 2010 in seguito a successi nel campo della **computer vision** e del **processamento del linguaggio naturale**



Tipi di dati/problemi

Dati tabellari

Input del modello è la riga di una tabella. Esempi:

- Credit scoring
- Behavioural policy pricing
- Medicina predittiva

Computer vision

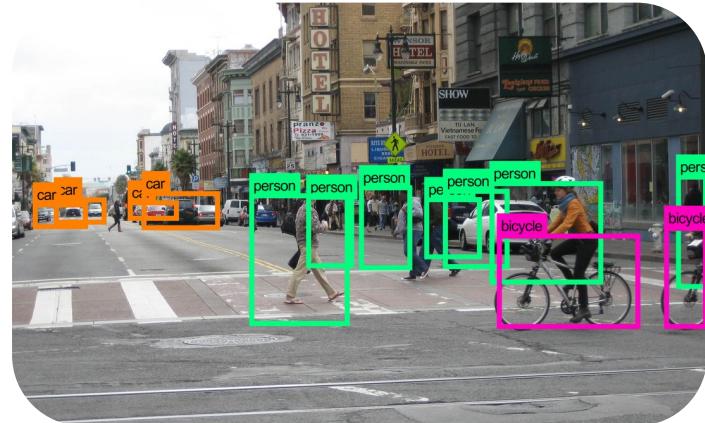
Input del modello è un'immagine o un video.

- Riconoscimento facciale
- Analisi scansioni mediche
- Guida autonoma

Natural Language Processing

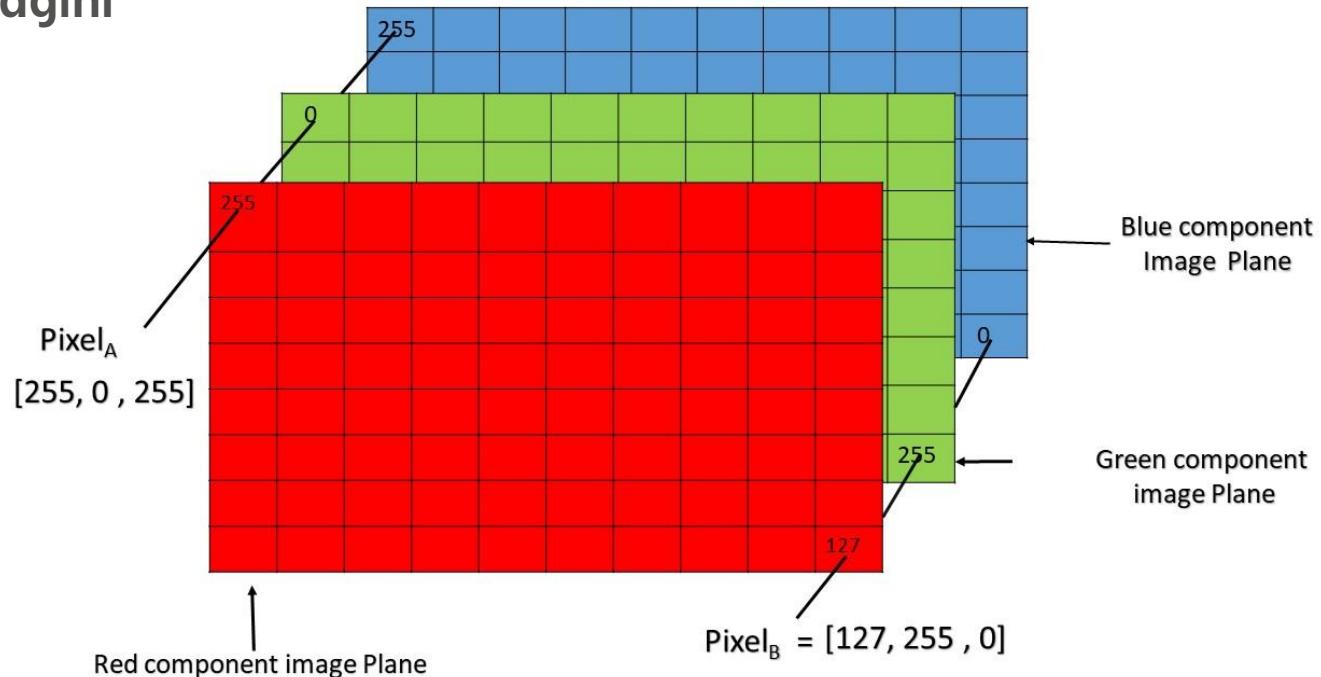
Input del modello è una sequenza di parole

- Analisi sentimento
- Traduzione macchina
- Chatbots



A screenshot of the Google Translate mobile application. At the top, it shows the language pair 'TURKISH' to 'ENGLISH'. Below this, the input text 'o bir doktor' is shown in the Turkish section. The English translation 'she is a doctor (feminine)' is displayed in the English section. There are also buttons for audio playback and a share icon.

Immagini



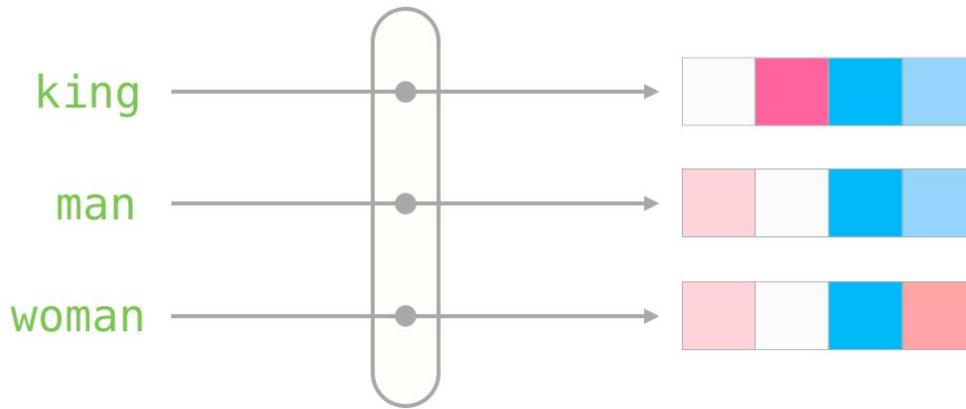
Pixel of an RGB image are formed from the corresponding pixel of the three component images

<https://www.geeksforgeeks.org/matlab-rgb-image-representation/>



Linguaggio

Word2vec



$$\text{king} - \text{man} + \text{woman} \approx \text{queen}$$

<http://jalammar.github.io/illustrated-word2vec/>



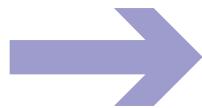
Altri tipi di apprendimento automatico

Apprendimento non supervisionato

(unsupervised learning)

Apprendimento non supervisionato permette di effettuare analisi sulle relazioni fra i dati. Viene effettuato senza 'etichette' Y.

X



~~Y~~

Apprendimento non supervisionato

(unsupervised learning)

Apprendimento non supervisionato permette di effettuare analisi sulle relazioni fra i dati. Viene effettuato senza ‘etichette’ Y.



Esempio:



Storia



Saggistica



Romanzo



Storia

Apprendimento non supervisionato

(unsupervised learning)

Apprendimento non supervisionato permette di effettuare analisi sulle relazioni fra i dati. Viene effettuato senza ‘etichette’ Y.

Applicazioni più legate all’analisi dei dati.



Esempio:



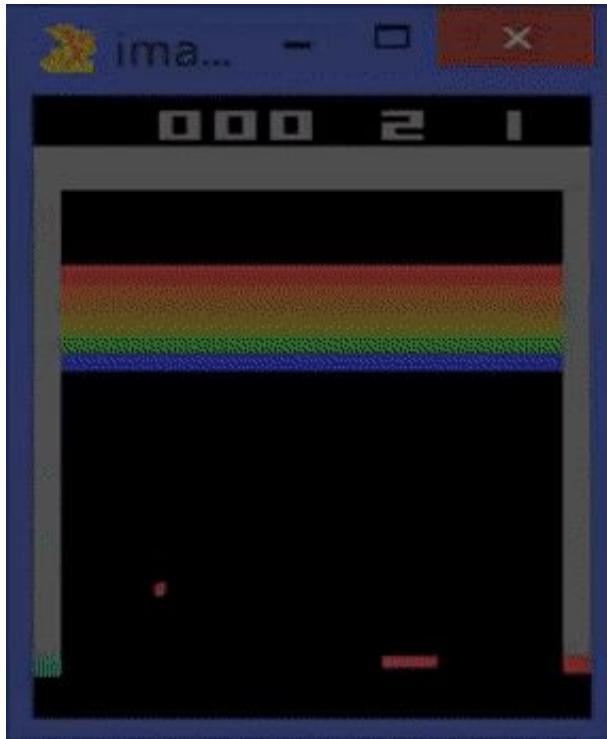
Apprendimento per rinforzo

(reinforcement learning)

Il “reinforcement learning” punta ad addestrare agenti in grado di raggiungere obiettivi imparando attraverso l’interazione in un determinato ambiente.

I modelli sono disegnati per imparare a migliorarsi in base ai propri successi o errori.

Prime applicazioni in ambito di automazione industriale.



Sviluppi recenti in ambito IA

Intelligenza artificiale nelle news

AlphaGo

DeepMind

AlphaGo è un programma sviluppato da **DeepMind** per giocare al **gioco del go**.

Il go è considerato molto più complesso degli scacchi in termini di numero di possibili movimenti a disposizione dei giocatori anche in stadi avanzati della partita..

Nel marzo 2016 sconfisse uno dei giocatori di go più forti al mondo.

Combina apprendimento supervisionato, per rinforzo, e algoritmi di ricerca.



Midjourney, DALL-E, Stable Diffusion

Modelli generativi per creazione di immagini



<https://www.aiartdigest.com/midjourney-examples/>



ThisPersonDoesNotExist.com

Large Language Models

OpenAI

Modelli in grado di leggere un testo in input e restituire un testo in output.

Trasformazione input-output ottenuta per mezzo di miliardi di parametri.

Input Prompt:

Recite the first law of robotics



Output:



AI e regolamentazione

GDPR ha introdotto (articolo 22) regole sui diritti di individui affetti da decisioni prese da algoritmi.

AI Act: proposta di regolamentazione Europea in ambito AI. Bozza approvata da Parlamento Europeo, dovrebbe entrare in vigore nel 2024.

Divide sistemi AI in base a categorie di rischio e definisce azioni necessarie per poter utilizzare sistemi nelle varie categorie di rischio.



AI Act

Classificazione modelli AI basata sul rischio, quattro categorie:

- **Rischio inaccettabile:** applicazioni di IA proibite a causa di violazioni della dignità umana, della democrazia o dei diritti fondamentali, come i sistemi di valutazione sociale o l'identificazione biometrica in tempo reale negli spazi pubblici.
- **Alto rischio:** applicazioni di IA soggette a rigorosi obblighi prima che possano essere messe in servizio. Include l'IA utilizzata in infrastrutture critiche, istruzione, occupazione, salute, giustizia o forze dell'ordine.
- **Rischio limitato:** Le applicazioni di IA in questa categoria, come i chatbot, devono attenersi ai requisiti di trasparenza.
- **Rischio minimo:** Le applicazioni di IA con rischio minimo non affrontano obblighi specifici, come l'IA utilizzata per scopi di intrattenimento o personali.



Fine prima parte

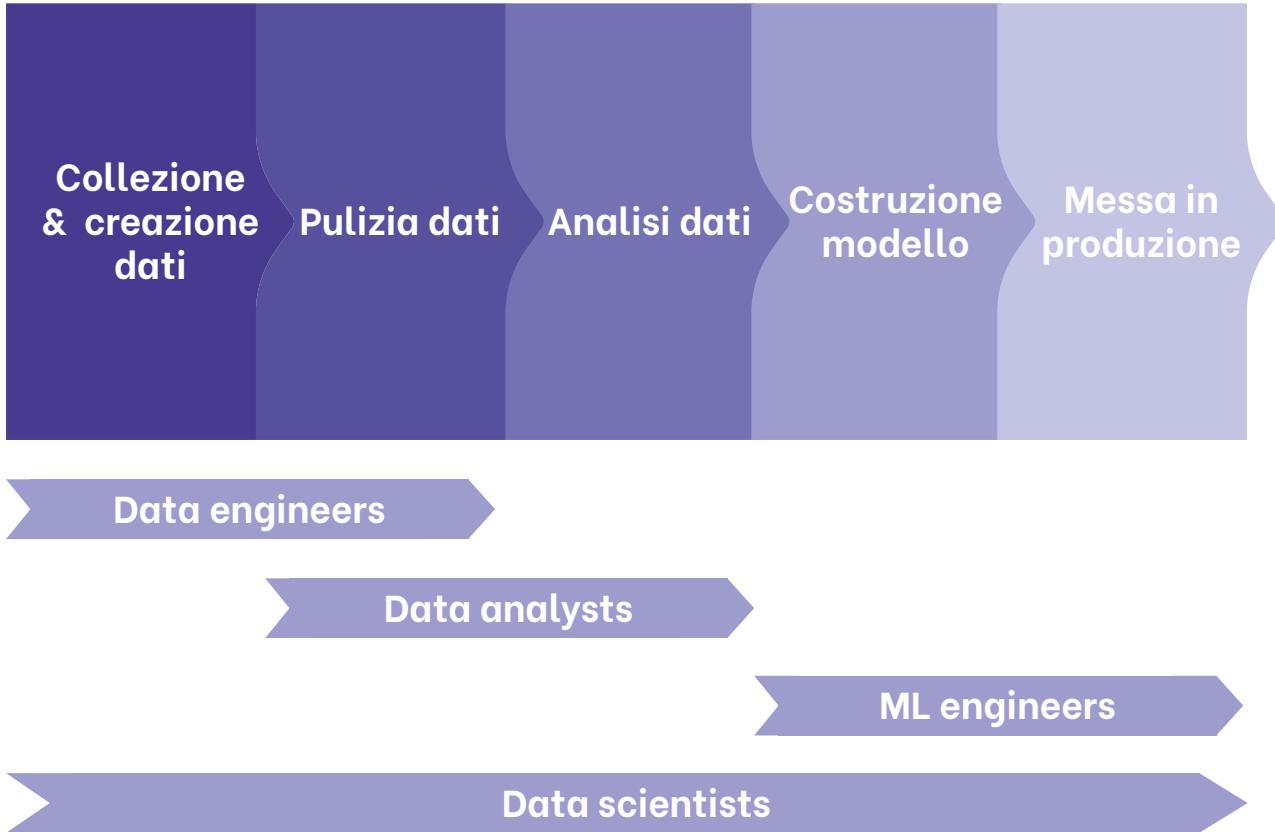
Seconda parte

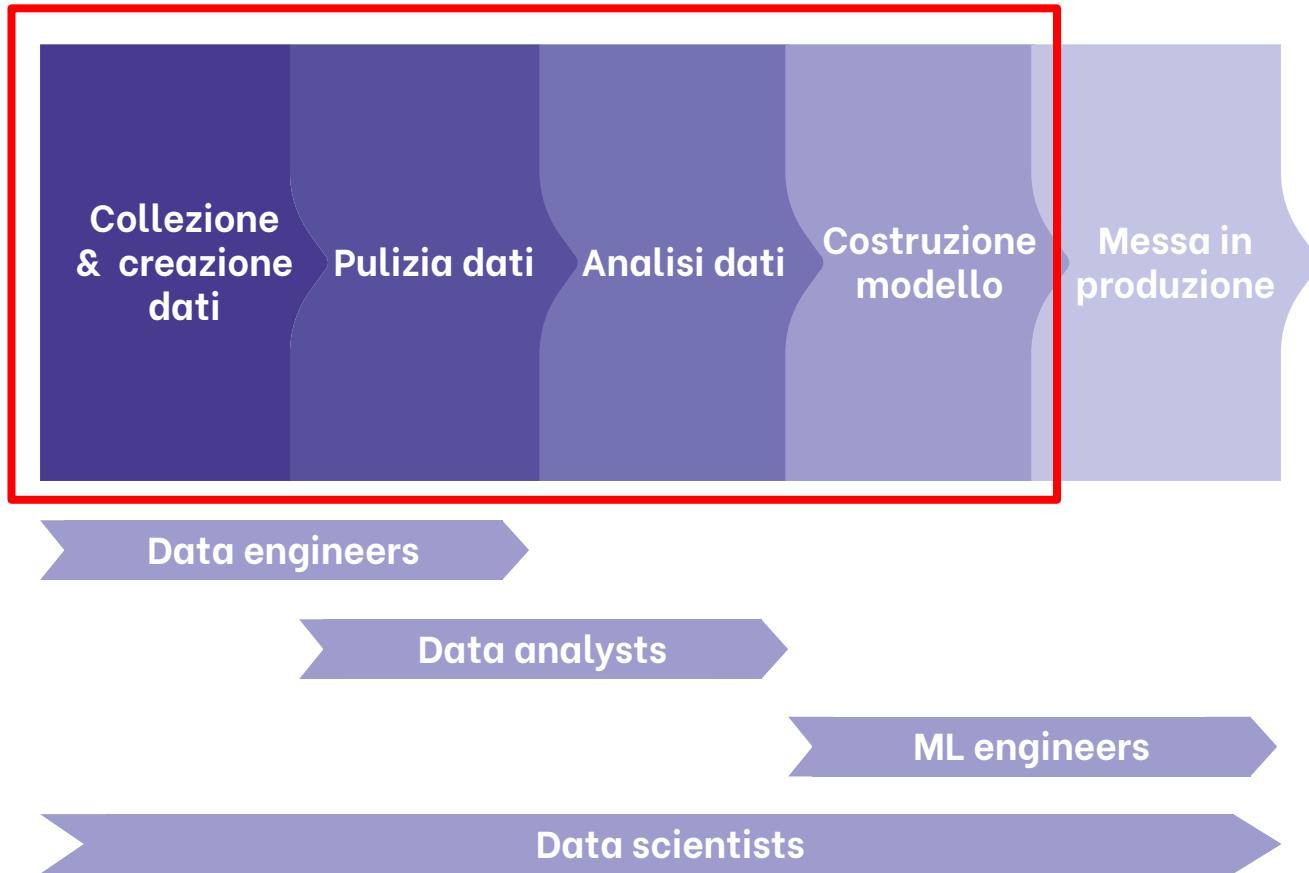
Aspetti pratici di un progetto di
machine learning

L'apprendimento automatico

- 1 - **Definizione problema:** che cosa vogliamo automatizzare?
- 2 - **Raccolta e preparazione dati:** abbiamo esempi di combinazione domanda/risposta?
- 3 - **Creazione e addestramento modello:** come creiamo un modello che generi risposte a nuove domande?
- 4 - **Validazione e produzione:** come lo trasformiamo in uno strumento utile?







Raccolta dati

Il primo ostacolo

Problema:

I modelli di machine learning più performanti, come ad esempio quelli basati sul deep learning hanno bisogno di grandi quantità di dati per essere addestrati.

Oltre ad essere “raccolti” e puliti questi dati devono essere anche **etichettati**. Processo molto dispendioso in termini di tempo e denaro.



<https://www.nytimes.com/2018/11/25/business/china-artificial-intelligence-labeling.html>

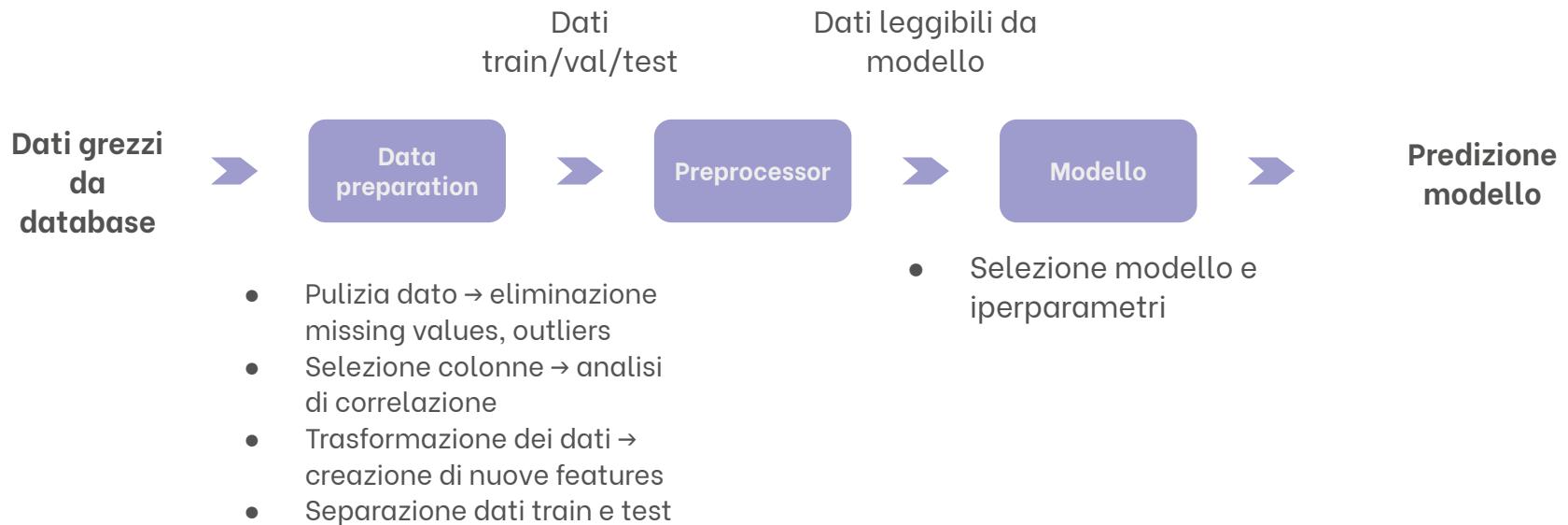
Esempio

Problema giocattolo → Classificazione binaria (Adult Income Dataset)

| age | work_class | education | marital_status | occupation | relationship | race | sex | capital_gain | capital_loss | hours_per_week | native_country | income |
|-----|------------------|------------|--------------------|-------------------|---------------|-------|--------|--------------|--------------|----------------|----------------|--------|
| 39 | State-gov | Bachelors | Never-married | Adm-clerical | Not-in-family | White | Male | 2174\$ | 0\$ | 40 | United-States | <=50K |
| 50 | Self-emp-not-inc | Bachelors | Married-civ-spouse | Exec-managerial | Husband | White | Male | 0\$ | 0\$ | 13 | United-States | <=50K |
| 38 | Private | HS-grad | Divorced | Handlers-cleaners | Not-in-family | White | Male | 0\$ | 0\$ | 40 | United-States | <=50K |
| 53 | Private | 11th | Married-civ-spouse | Handlers-cleaners | Husband | Black | Male | 0\$ | 0\$ | 40 | United-States | <=50K |
| 28 | Private | Bachelors | Married-civ-spouse | Prof-specialty | Wife | Black | Female | 0\$ | 0\$ | 40 | Cuba | <=50K |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 27 | Private | Assoc-acdm | Married-civ-spouse | Tech-support | Wife | White | Female | 0\$ | 0\$ | 38 | United-States | <=50K |
| 40 | Private | HS-grad | Married-civ-spouse | Machine-op-inspct | Husband | White | Male | 0\$ | 0\$ | 40 | United-States | >50K |
| 58 | Private | HS-grad | Widowed | Adm-clerical | Unmarried | White | Female | 0\$ | 0\$ | 40 | United-States | <=50K |
| 22 | Private | HS-grad | Never-married | Adm-clerical | Own-child | White | Male | 0\$ | 0\$ | 20 | United-States | <=50K |
| 52 | Self-emp-inc | HS-grad | Married-civ-spouse | Exec-managerial | Wife | White | Female | 15024\$ | 0\$ | 40 | United-States | >50K |

Esempio

Problema giocattolo → Classificazione binaria (Adult Income Dataset)



Data processing

| id | color |
|----|-------|
| 1 | red |
| 2 | blue |
| 3 | green |
| 4 | blue |

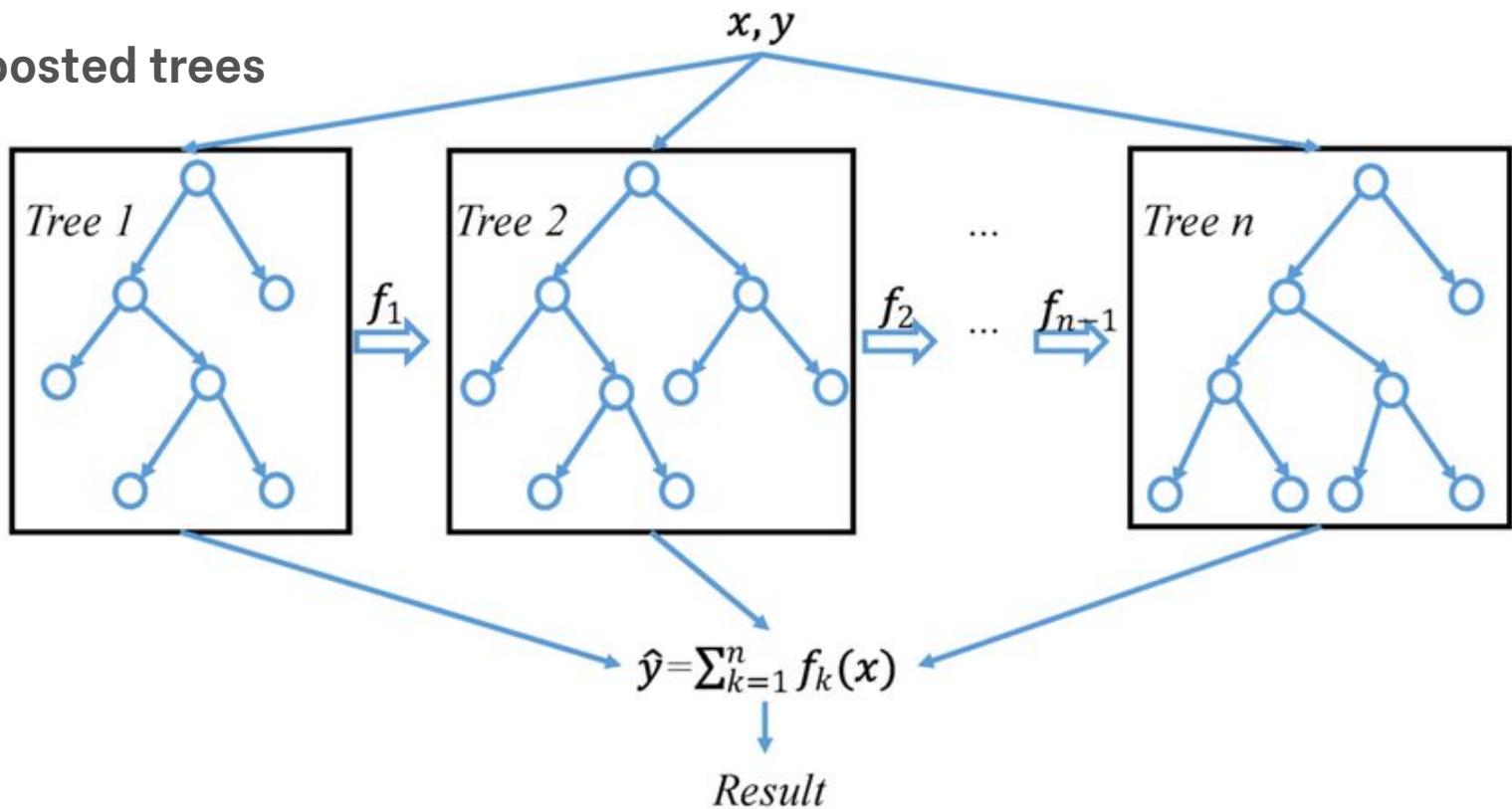


| id | color_red | color_blue | color_green |
|----|-----------|------------|-------------|
| 1 | 1 | 0 | 0 |
| 2 | 0 | 1 | 0 |
| 3 | 0 | 0 | 1 |
| 4 | 0 | 1 | 0 |

<https://towardsdatascience.com/building-a-one-hot-encoding-layer-with-tensorflow-f907d686bf39>



Boosted trees



Metriche di valutazione: esempio classificazione

| | | Real label |
|-----------------|--|--|
| Predicted label | True Positive | False Positive |
| | False Negative | True Negative |
| |  |  |
| |  |  |

Accuracy: $(TP + TN) / (TP+FP+TN+FN)$

Quanti cani e gatti ho classificato correttamente?

Precision: $TP / (TP+FP)$

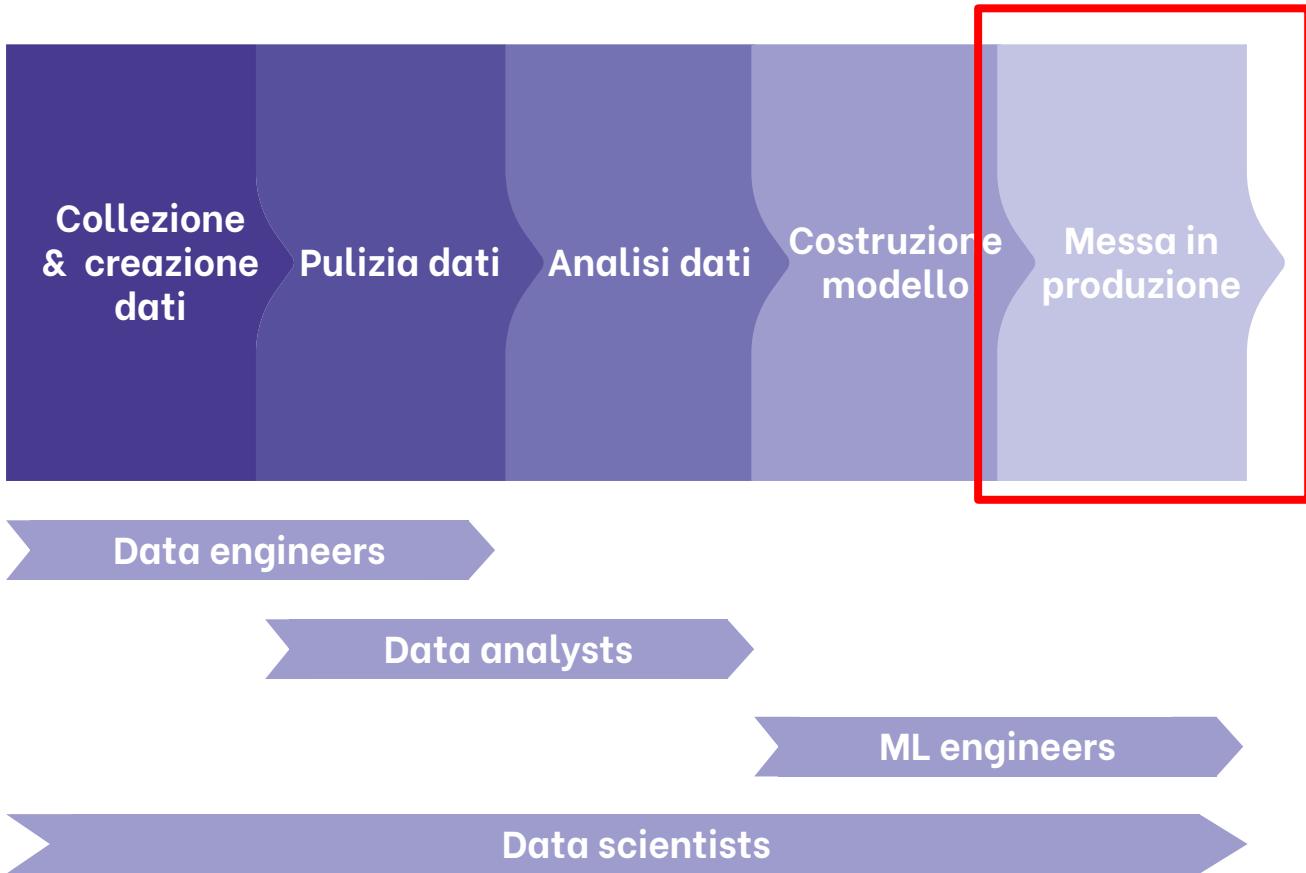
Quanti gatti ho classificato come cani?

Recall: $TP / (TP+FN)$

Quanti cani mi sono perso?

F1 Score: $2 * \text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall})$

È la media armonica tra la precisione e la recall



**“Although more than 90%
of organizations are
investing in AI today, only
17% have been able to
scale AI.”**

Secondo KPMG

Aspetti importanti della messa in produzione di modelli

- Interpretabilità
- Monitoraggio e miglioramento continuo

Interpretabilità

Fiducia nell'intelligenza artificiale

Algoritmi di IA principalmente usati per compiti molto specifici, solitamente in sinergia con umani.

Intelligenza artificiale ⇒ **Intelligenza aumentata**

Obiettivo e' quello di efficientare un processo, molto spesso però avviene il contrario.

Esiste un problema definito come la **mancanza di fiducia negli algoritmi**. Il motivo principale e' la difficoltà da parte degli umani di comprendere come funziona l'algoritmo stesso.

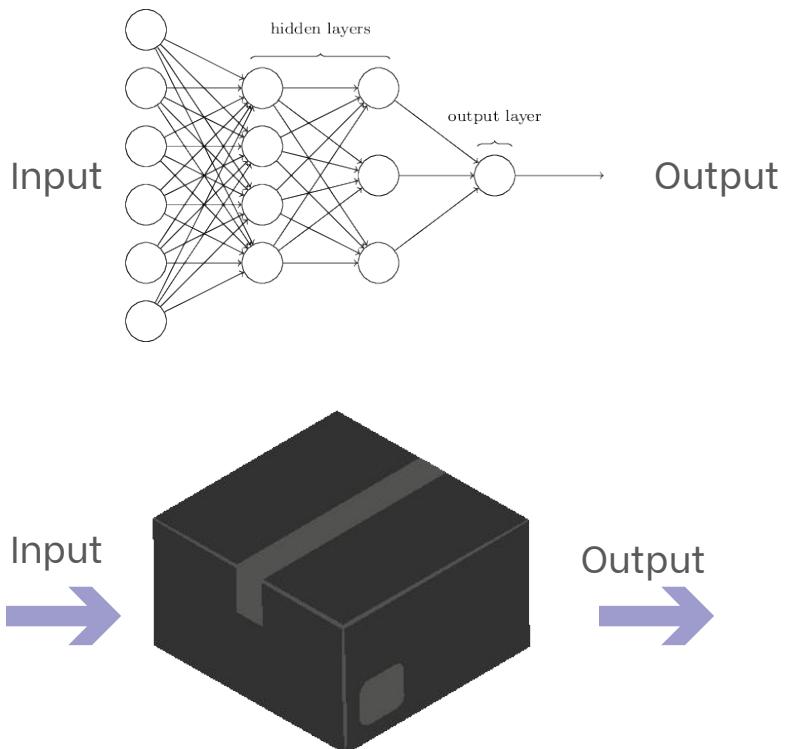
Loan application



Il problema della “black box”

Algoritmi più performanti tendono ad essere **meno spiegabili**, difficile stabilire il perché un modello abbia deciso una cosa piuttosto che un'altra

Questo crea problemi di trasparenza, fiducia e sicurezza e spinge organizzazioni a preferire modelli più semplici e meno performanti.



Come aumentare la fiducia nell'IA

Andando ad arricchire l'output dei modelli con spiegazioni

Spiegabilità modelli

 IEEE
SPECTRUM

Engineering Topics ▾ Special Reports ▾ Blogs ▾ Multimedia ▾ The Magazine ▾ Professional Resources ▾ Search ▾

Feature | Biomedical | Diagnostics

02 Apr 2019 | 15:00 GMT

How IBM Watson Overpromised and Underdelivered on AI Health Care

After its triumph on Jeopardy!, IBM's AI seemed poised to revolutionize medicine. Doctors are still waiting

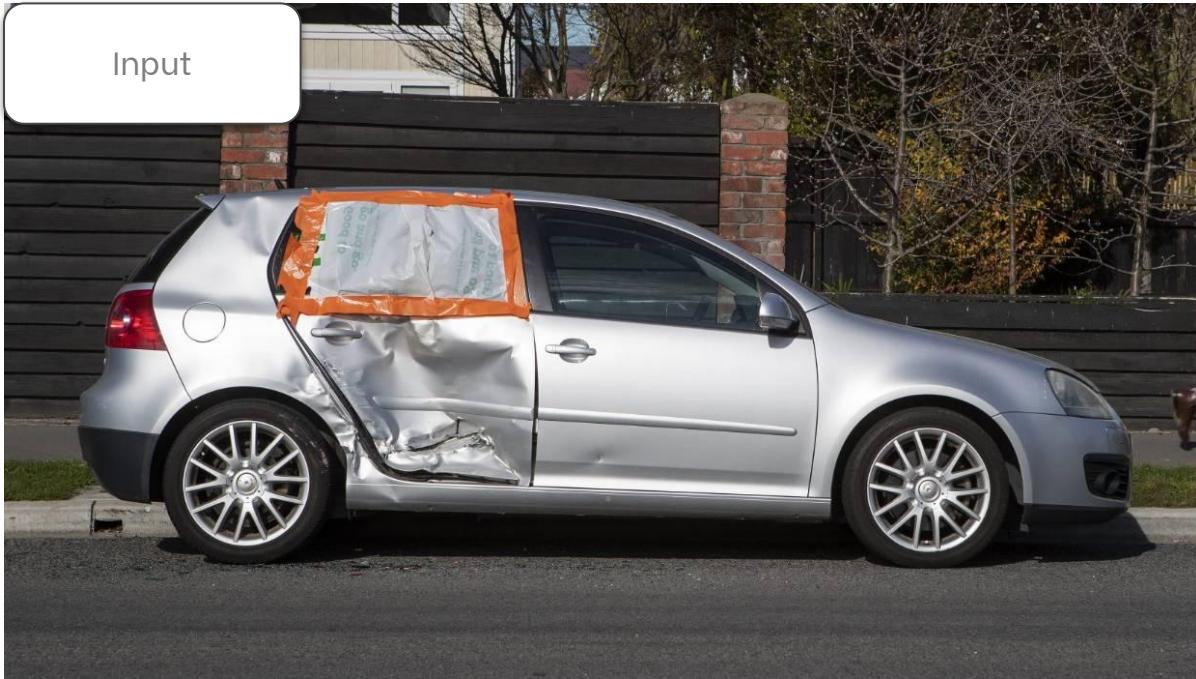
By Eliza Strickland

In 2014, IBM opened swanky new headquarters for its artificial intelligence division, known as **IBM Watson**. Inside the glassy tower in lower Manhattan, IBMers can bring prospective clients and visiting journalists into the "immersion room," which resembles a miniature planetarium. There, in the darkened space, visitors sit on swiveling stools while fancy graphics flash around the curved screens covering the



Spiegabilità modelli

Input

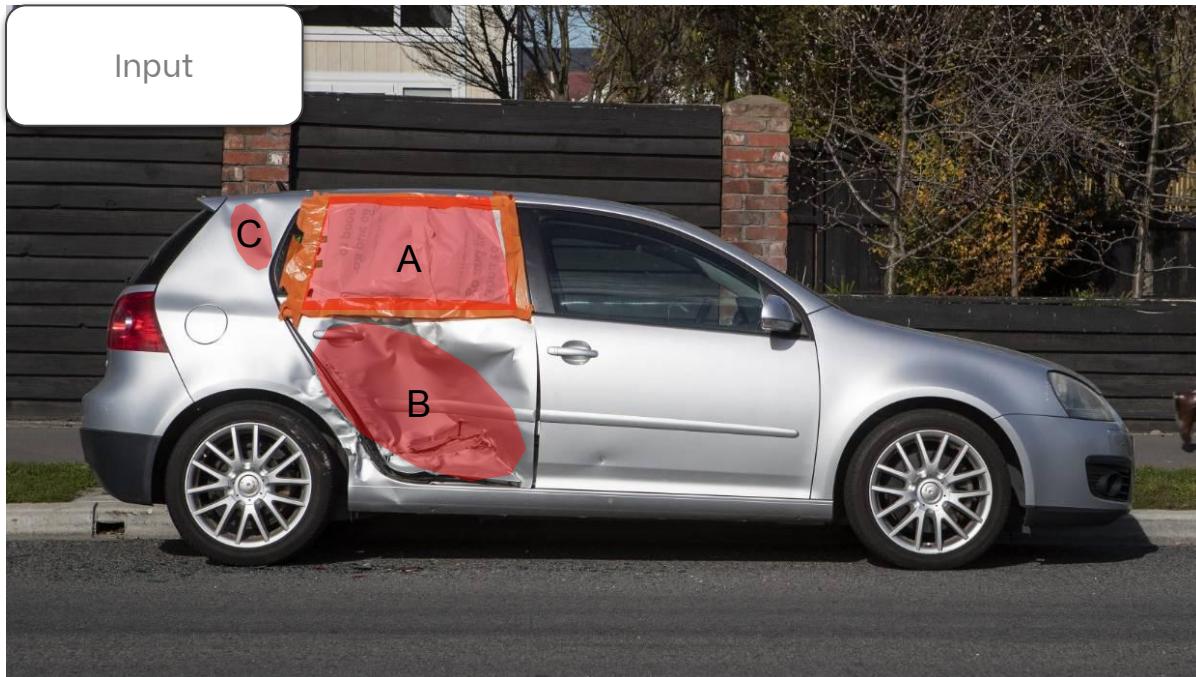


Output:

Damage = 2000€

Spiegabilità modelli

oltre al debugging

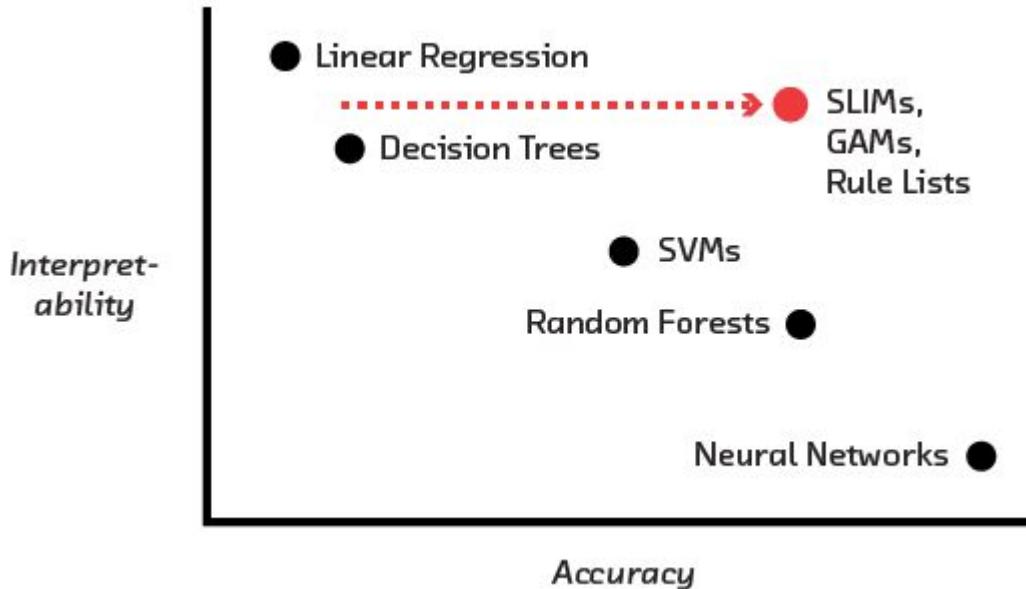


Output:
Damage = 2000€
Breakdown:
A = 300€
B = 1200€
C = 500€

Historical Example for B
Label = 2500€

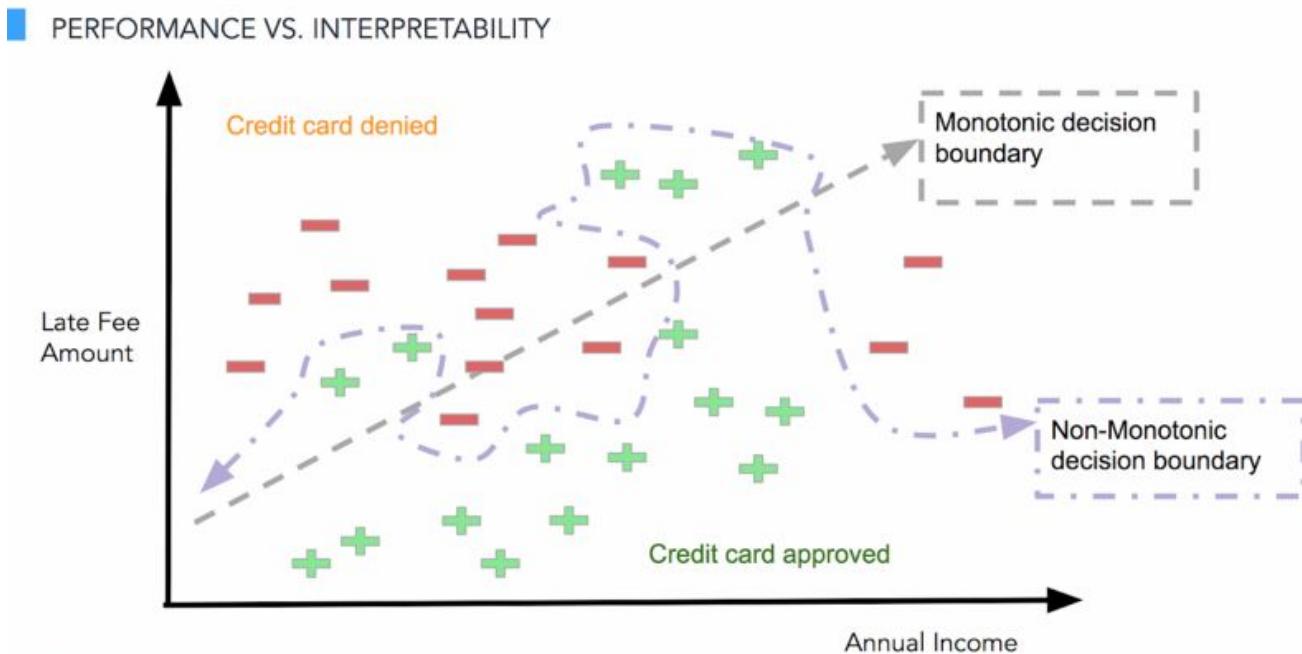


Spiegabilità modelli



<https://ff06-2020.fastforwardlabs.com/>

Tassonomia eXplainable AI

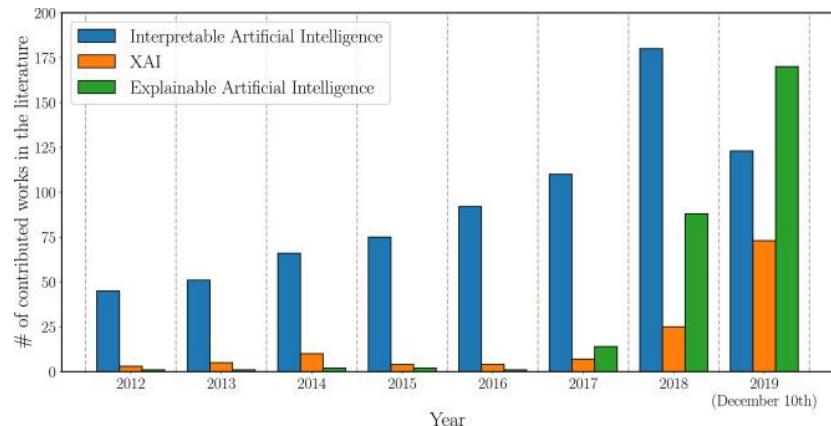


<https://www.kdnuggets.com/2018/12/explainable-ai-model-interpretation-strategies.html>

L'eXplainable AI (XAI)

Ambito di ricerca il cui scopo è quello di aiutare umani a **interpretare** decisioni prese da modelli.

Interesse accademico è in costante crescita → Sempre più metodi e librerie.

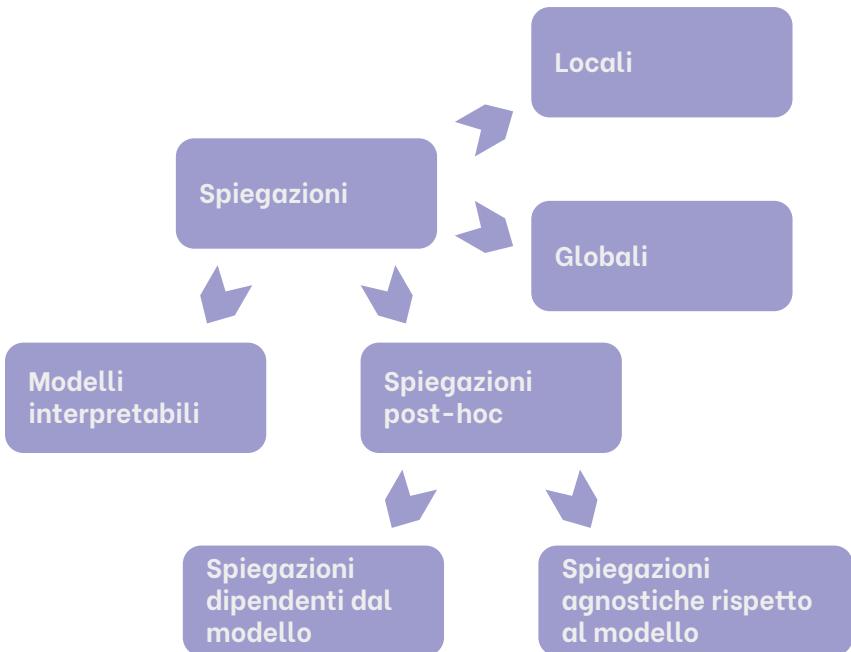


Arrieta et al, Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion, 58, 2020

Tassonomia eXplainable AI

Modi di spiegare i modelli possono essere molteplici.

Nelle prossime slides ci focalizzeremo su spiegazioni **locali**, agnostiche e non.

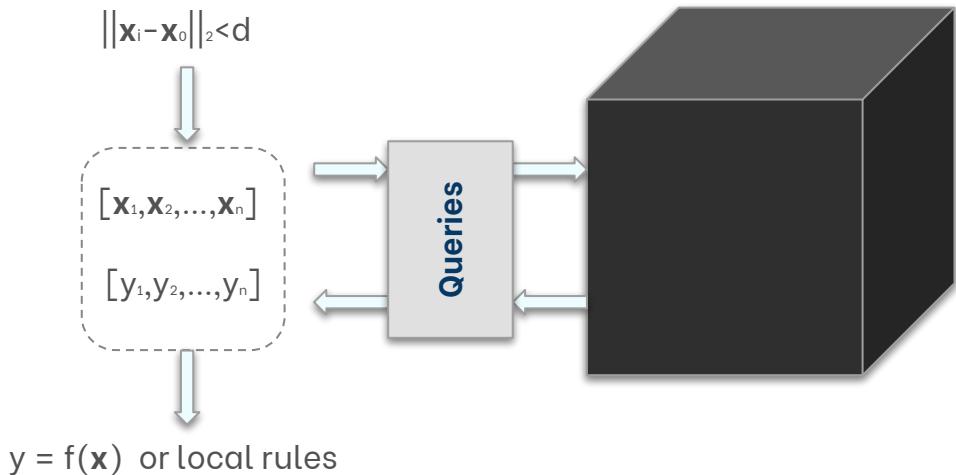


Metodi perturbativi

Agnostici rispetto al tipo di modello

Metodi disegnati per generare spiegazioni **locali** in maniera **agnostica**.

Spiegazioni generate ricostruendo un **modello semplificato** che approssimi il modello originale nelle vicinanze del punto da spiegare.

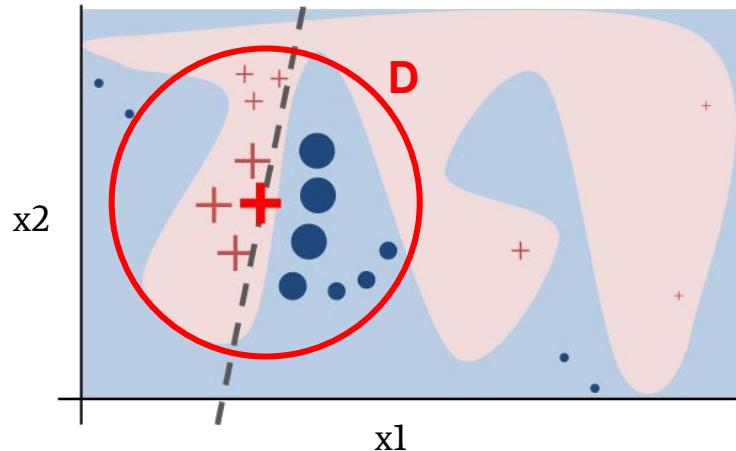


Local Interpretable Model-agnostic Explanations (LIME)

Rientra tra i metodi più popolari, precursore approcci perturbativi.

Metodo:

- Genera **N** punti in un intorno **D** del punto da spiegare.
- Ottieni la risposta del modello per questo insieme di punti.
- Costruisci un classificatore lineare usando le x,y ottenute nei passi precedenti.



<https://arxiv.org/pdf/1602.04938.pdf>

LIME

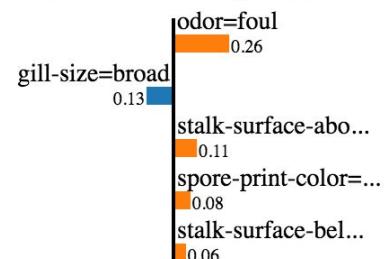
Esempio

Prediction probabilities



edible

poisonous



Feature

Value

| | |
|--------------------------------|------|
| odor=foul | True |
| gill-size=broad | True |
| stalk-surface-above-ring=silky | True |
| spore-print-color=chocolate | True |
| stalk-surface-below-ring=silky | True |

- Spiegazioni generate usando diversi iperparametri (N punti, distanza D, etc)
- Può essere applicato a problemi su dati strutturati e non strutturati
- Spiegazioni possono non convergere.
- Spiegazioni non sono **prescrittive**.

Shapley values

Teoria dei giochi

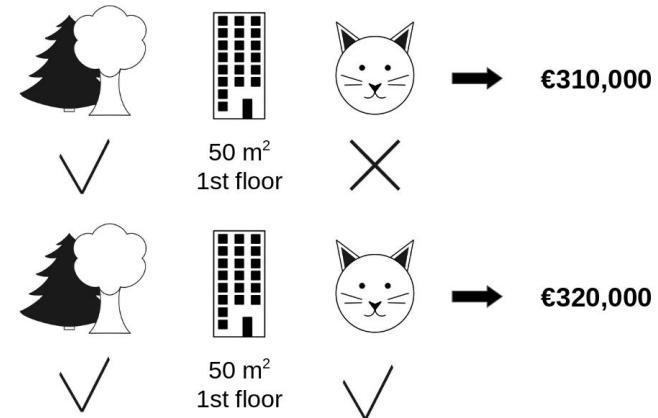
Concetto proveniente dalla teoria dei giochi: *come ridistribuire la ricompensa di un gioco a cui ha partecipato un gruppo di giocatori in maniera cooperativa?*

Coefficienti di Shapley definiscono una maniera per distribuire ricompensa tra partecipanti.

Applicato al machine learning:

Giocatori → Features

Ricompensa → Output del modello



<https://christophm.github.io/interpretable-ml-book/>

SHAP

Approssimazione Shapley Values

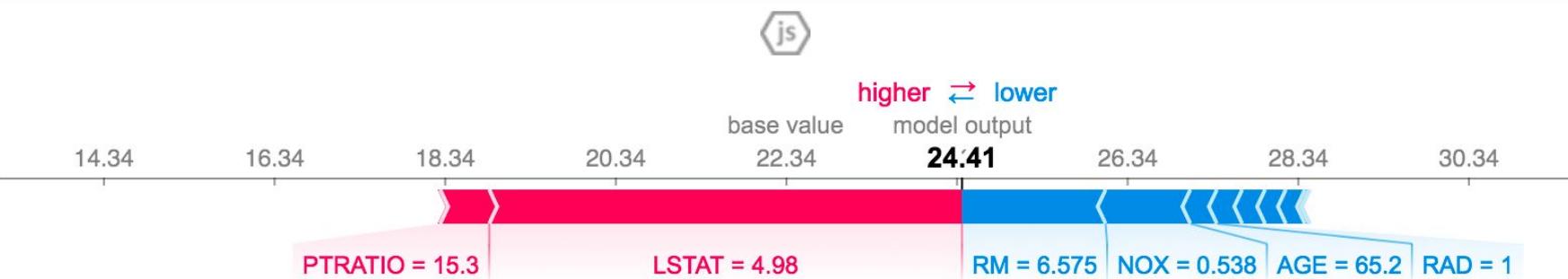
Metodo definisce una strategia per ottenere un'approssimazione dei valori di Shapley usando un approccio molto simile a LIME.

KernelSHAP → Metodo completamente agnostico rispetto al modello

TreeSHAP → Implementazione per modelli basati su alberi decisionali, molto più veloce

SHAP

Approssimazione Shapley Values



- Libreria SHAP offre oltre al metodo stesso librerie di visualizzazione molto curate
- KernelSHAP e' decisamente lento.
- Così come per LIME la generazione di punti sintetici non tiene conto delle dipendenze statistiche tra features.

Spiegazioni immagini

Features modello → superpixels che possono essere attivati o disattivati.

Pixels disattivati: trasformati in pixel neri o associati a rumore (come gaussian blur)



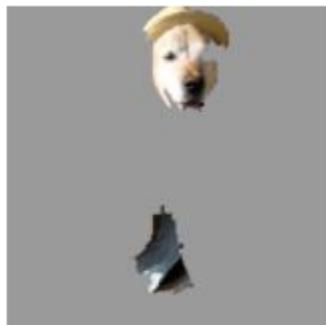
(a) Original Image



(b) Explaining *Electric guitar*



(c) Explaining *Acoustic guitar*



(d) Explaining *Labrador*

Spiegazioni tramite esempi

Saper indicare quali dati di allenamento hanno influito su una particolare predizione.

| ID immobile | locali | tipo | Superficie (m2) | bagni | giardino | Venduto a |
|-------------|--------|-------|-----------------|-------|----------|-----------|
| n | 3 | villa | 150 | 2 | si | 315000 € |

Spiegazioni tramite esempi

Saper indicare quali dati di allenamento hanno influito su una particolare predizione.

| ID immobile | locali | tipo | Superficie (m2) | bagni | giardino | Venduto a |
|-------------|--------|-------|-----------------|-------|----------|-----------|
| n | 3 | villa | 150 | 2 | si | 315000 € |

Questo immobile proveniente dai dati di allenamento è molto simile:

| ID immobile | locali | tipo | Superficie (m2) | bagni | giardino | Venduto a |
|-------------|--------|-------|-----------------|-------|----------|-----------|
| 424 | 3 | villa | 140 | 2 | si | 310000 € |

Spiegazioni tramite esempi

Saper indicare quali dati di allenamento hanno influito su una particolare predizione.

Problema: come definire similitudini tra punti?

Approccio più immediato → Nearest Neighbours utilizzando input stessi o rappresentazioni intermedie interne ai modelli (es: attivazioni in reti neurali)



https://beenkim.github.io/papers/KIM2016NIPS_MMD.pdf

Esempi controfattuali

ragionamento tramite scenari ipotetici

Esempi che **cambiano la predizione** in una determinata direzione stando all'interno di **vincoli** specifici.

Es: *Se avessi avuto 5 anni in più il prestito sarebbe stato accettato.*

Impostati come problema di ottimizzazione vincolata nello spazio delle features.

$$\|\mathbf{x} - \mathbf{x}'\| \leq \delta$$

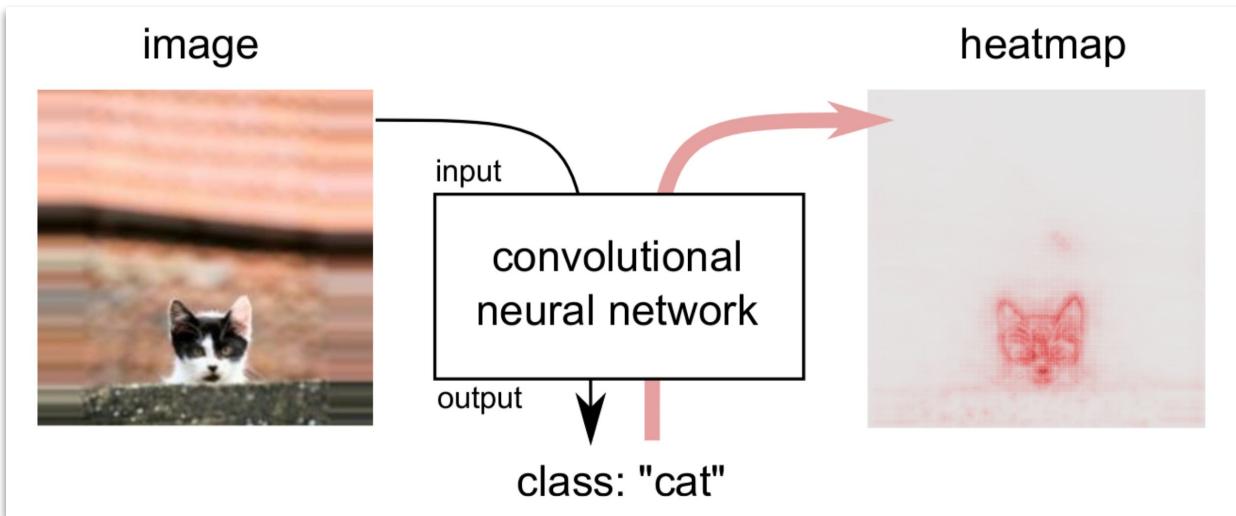


$$|f(\mathbf{x}) - f(\mathbf{x}')| > \epsilon$$

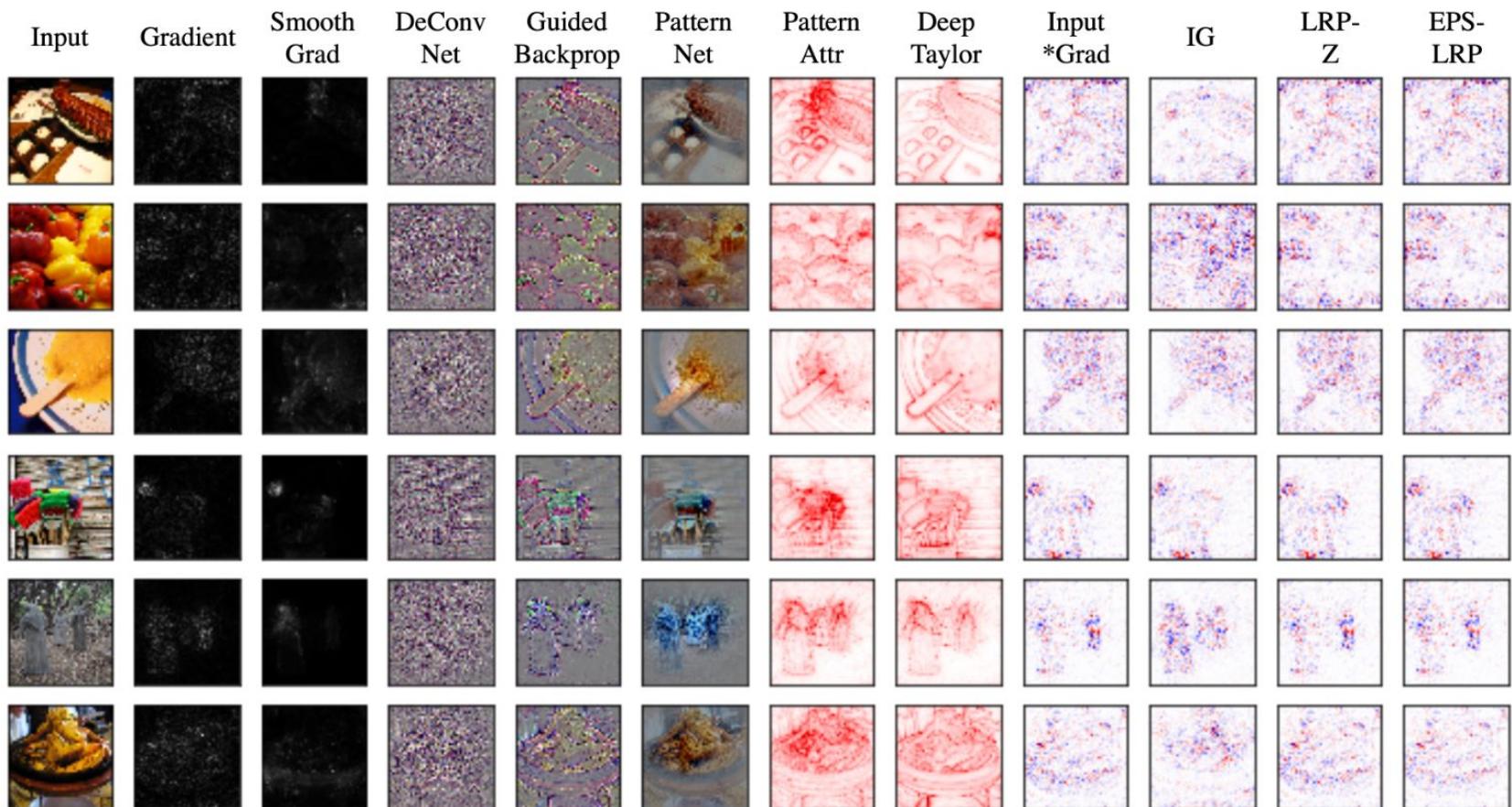
Metodi intrusivi

Aprire la black-box

Esempio: Deep Taylor Decomposition



Source: Montavon et al. (ICML 2016)



Singh A, Sengupta S, Lakshminarayanan V. Explainable Deep Learning Models in Medical Image Analysis. *Journal of Imaging*. 2020; 6(6):52. <https://doi.org/10.3390/jimaging6060052>

Oltre l'interpretabilità

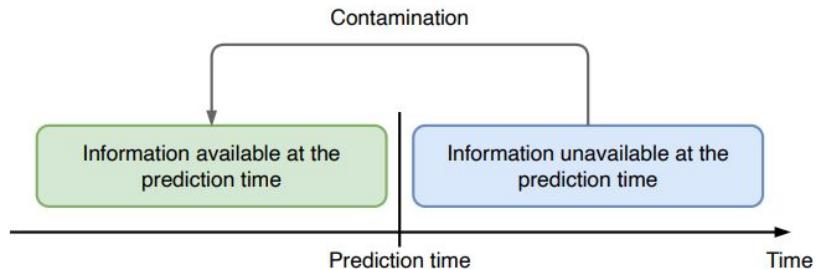
L'interpretabilità dei modelli, come si è visto, è un problema di fondamentale importanza per i modelli di IA. Altri aspetti da tenere in considerazione, prima della messa in produzione dei modelli sono:

- Data leakage
- Bias e fairness
- Robustezza modelli

Data leakage

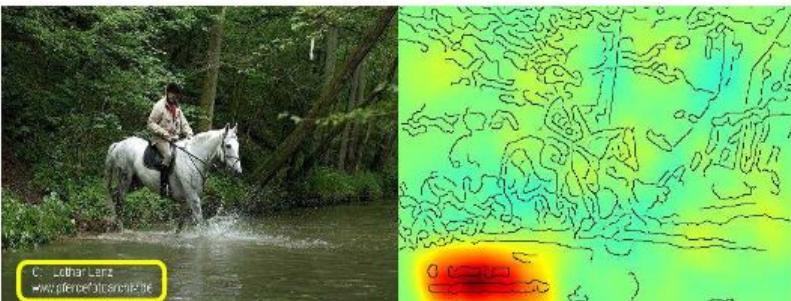
Il modello vede in fase di allenamento informazione che non è presente in fase di inferenza. Motivi tipici:

- Una delle features ‘nasconde’ il target
- Una delle features viene dal ‘futuro’



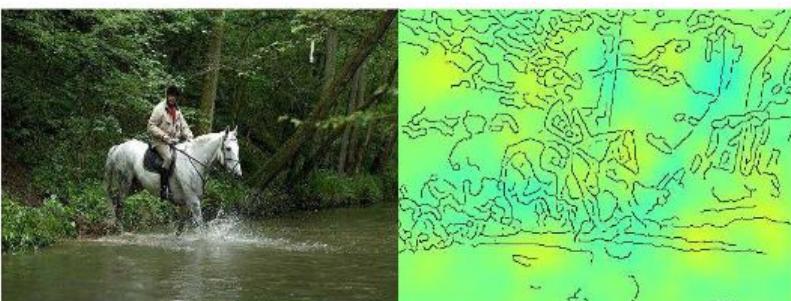
Data leakage

Horse-picture from Pascal VOC data set



Source tag present

Classified as horse



No source tag present

Not classified as horse

Unmasking Clever Hans Predictors and Assessing What Machines Really Learn. Lapuschkin et al. Nature comm (2019)

Bias e Fairness

Modelli imparano da dati → Dati possono contenere diversi tipi di bias che potrebbero non risultare accettabili in un modello in produzione.

Bias e Fairness in generative AI



“An impressionist painting of a data scientist working on their laptop”



“An impressionist painting of a person sweeping the floor”

Bias e Fairness

The screen shows a blue male icon at the top left. The title "Your Results" is centered above a text area. Below the text area are two cards, each containing a numbered list item and a small icon.

Some of the symptoms you reported might need emergency treatment. If things feel serious, your safest option is to call an ambulance.

Based on the information you gave, some possible causes are listed below.

1 Pericarditis
An inflammation of the thin membranes surrounding the heart.
This is usually treated at the emergency department.

2 Unstable angina
A lack of blood supply to the heart muscle, causing unpredictable chest pain.
This is usually treated at the emergency department.

The screen shows a red female icon at the top left. The title "Your Results" is centered above a text area. Below the text area are two cards, each containing a numbered list item and a small icon.

Some of the symptoms you reported might need to be checked out by a GP within the next 6 hours.

Based on the information you gave, some possible causes are listed below.

If symptoms persist, worsen or you are concerned, seek further medical advice.

1 Panic attack
A sudden period of intense fear and anxiety.
This can usually be treated at home.

2 Depression
A persistently low mood causing sadness and a loss of interest in doing things.
This usually requires seeing a GP.

Bias e Fairness

Apple's 'sexist' credit card investigated by US regulator

11 November 2019

f t w m Share



Steve Wozniak

@stevewoz

Replying to [@dhh](#)

The same thing happened to us. We have no separate bank accounts or credit cards or assets of any kind. We both have the same high limits on our cards, including our AmEx Centurion card. But 10x on the Apple Card.

7:58 AM · Nov 10, 2019 · Twitter Web App



DHH
@dhh

The [@AppleCard](#) is such a fucking sexist program. My wife and I filed joint tax returns, live in a community-property state, and have been married for a long time. Yet Apple's black box algorithm thinks I deserve 20x the credit limit she does. No appeals work.

9:34 PM · Nov 7, 2019 · Twitter for iPhone

12.8K Retweets 28.6K Likes



DHH @dhh · Nov 7, 2019

Replying to [@dhh](#)

I'm surprised that they even let her apply for a card without the signed approval of her spouse? I mean, can you really trust women with a credit card these days??!

86 270 4.4K

DHH @dhh · Nov 7, 2019

It gets even worse. Even when she pays off her ridiculously low limit in full, the card won't approve any spending until the next billing period. Women apparently aren't good credit risks even when they pay off the fucking balance in advance and in full.

Robustezza modelli

Robustezza di un modello definita come la **stabilità rispetto a piccole perturbazioni**.

In determinati contesti modelli poco robusto possono avere conseguenze legate alla sicurezza.

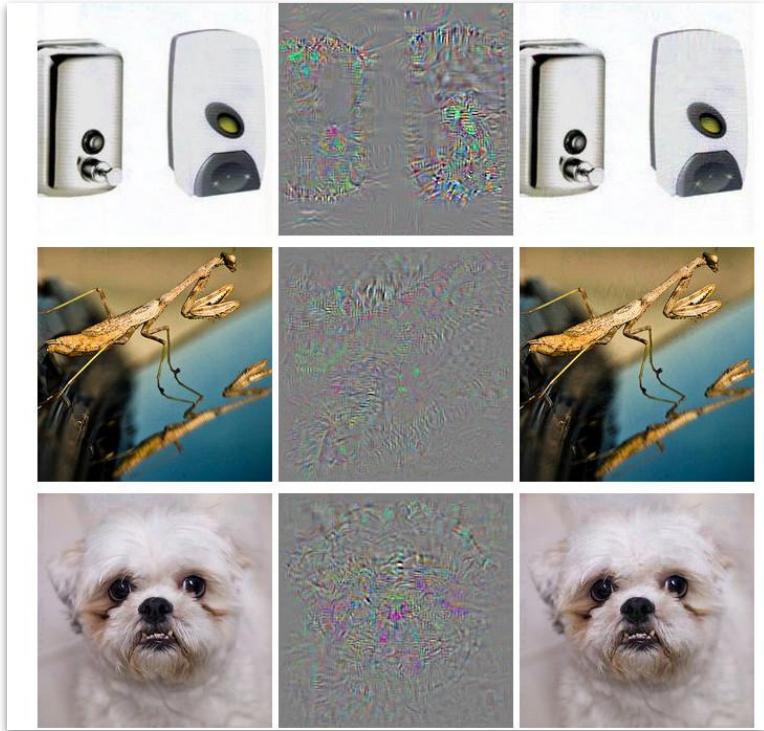
Es: Modello in ambito computer vision che commette errori quando in presenza di artefatti legati alla compressione di un'immagine

Adversarial robustness

Fenomeno legato alla robustezza di un modello →
Perturbazioni **mirate** che portano a errori importanti.



Source: Berkeley AI Research (BAIR)



“Intriguing properties of neural networks”, Szegedy et.al, 2013

Monitoraggio e miglioramento continuo

Analisi degli errori: motivazione

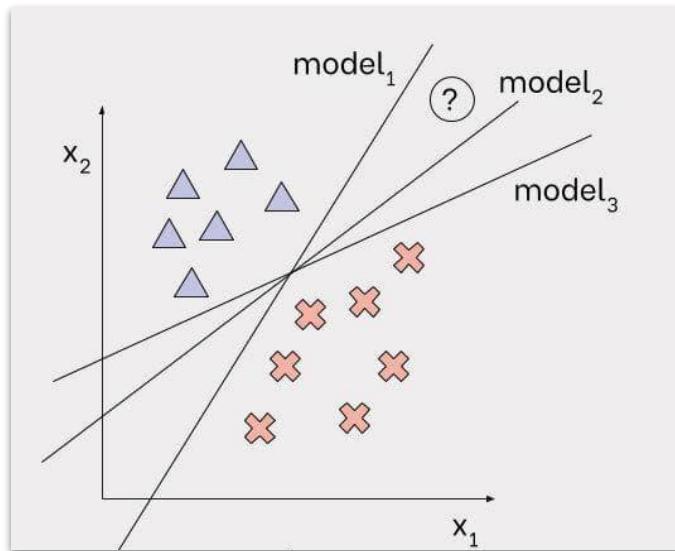
Un modello validato ed interpretabile sarà comunque soggetto ad errori.

Dobbiamo essere in grado di stimare la **probabilità di incorrere** in errore e di permettere a modelli di optare di non dare una risposta.

Incertezza epistemica o riducibile

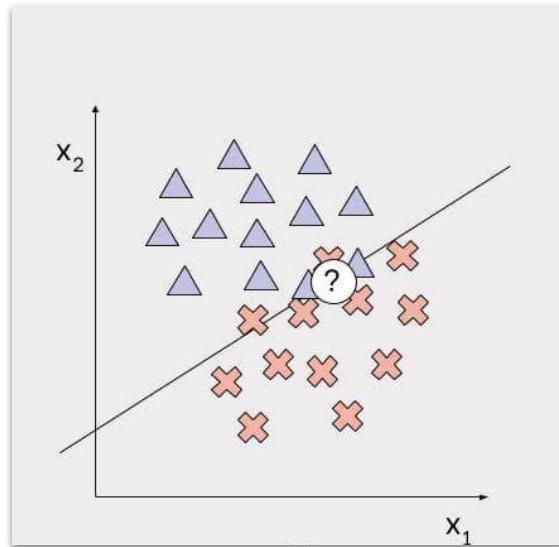
Incertezza epistemica rappresenta l'incertezza dovuta ad una **conoscenza incompleta** del problema analizzato.

In ambito machine learning è solitamente associata a mancanza di informazione a livello di dati.



Incertezza aleatoria o irriducibile

Questo tipo di incertezza e' associata ad una presenza di **rumore** a livello di dati che non può essere ridotto tramite feature engineering o raccolta dati.



Calibrazione modelli

Definizione

Approccio comune in ambito machine learning:
incorporare le varie sorgenti di incertezza
all'interno di un'unica quantità, la **confidenza
del modello**.

Problema sorge quando in presenza di modelli
cosiddetti non calibrati:

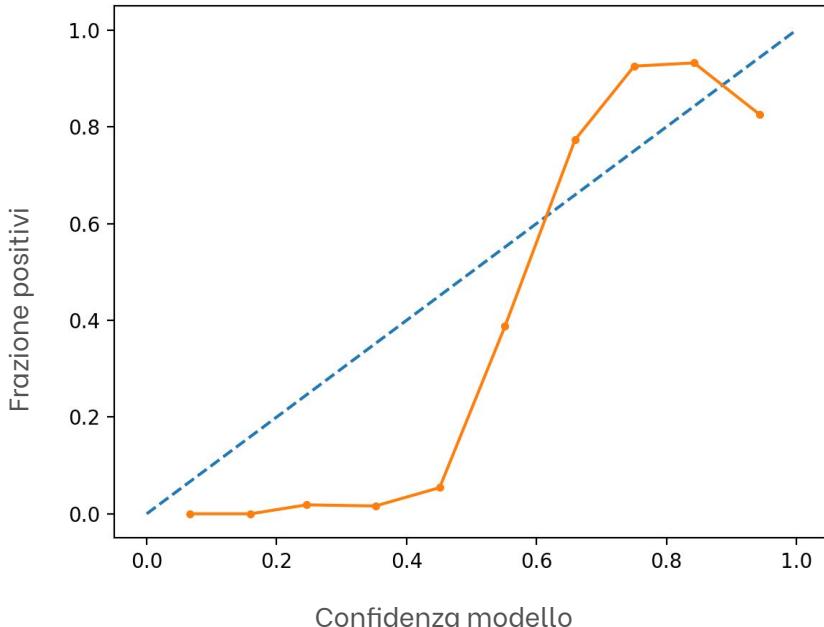
*es. una confidenza del 90% in un'etichetta deve
corrispondere dal vivo in un rateo di successo
analogo.*

Calibrazione modelli

Metodi

Calibrazione di un modello può essere quantificata tramite la curva di calibrazione (o reliability diagram)

Metodi per migliorare calibrazione di un modello non calibrato:



- **Platt scaling:** fit di una funzione di tipo sigmoide da aggiungere all'output del modello
- **Isotonic regression:** utilizzo di una funzione di regressione isotonica

Anomaly detection

Definizione limiti operativi modello

Prima di spostarsi nella fase di produzione bisogna definire un **range operativo** per un dato modello.

Modelli sono allenati su determinate distribuzioni di dati, devo assicurarmi che i dati in produzione provengano da distribuzioni simili, altrimenti il modello starebbe lavorando in un range sconosciuto.

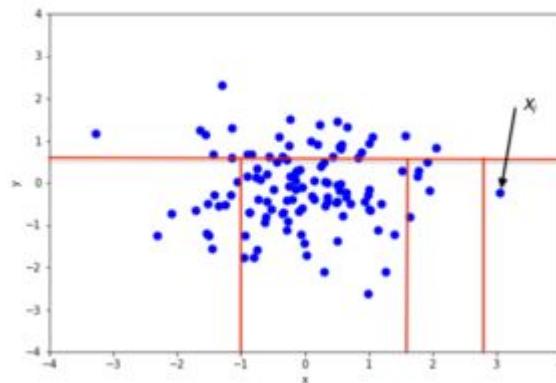
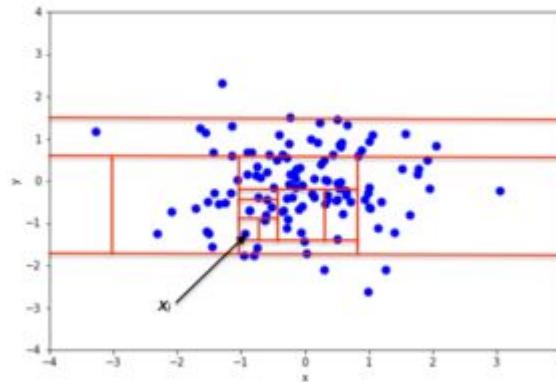
Come automatizzare questo processo di **rejection** di eventuali punti mai visti? Tramite **anomaly detection**.

Isolation Forests

Trovare anomalies tramite alberi decisionali

Idea: quando creiamo un albero decisionale i punti più anomali sono i più facili da isolare tramite splitting. (Foglie contenenti anomalie tendono ad essere vicine alla radice dell'albero)

Un Isolation forest consiste in un numero di isolation trees dove i punti anomali sono quelli isolati usando il percorso più breve.



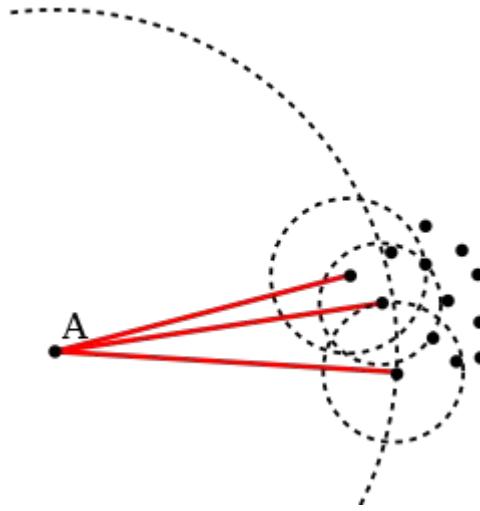
Local Outlier Factor

Analisi densità locali

Idea: Comparare la densità locale di un punto con densità locale di punti vicini.

Così come isolation forest e' basata su alberti decisionale, LOF e' basato su tecniche di tipo **Nearest Neighbours**.

Caratterizzato da stesse limitazioni associate a kNN: quale **metrica** usare per calcolare distanze tra punti?



Monitoraggio

Modelli in produzione devono essere tenuti sotto controllo e ri-allenati con una certa periodicità.

Motivo: dati cambiano nel tempo → Modelli rischiano di finire in condizioni operative diverse da quelle in cui sono stati allenati.

Due principali tipi di fenomeno:

1. **Data** drift
2. **Concept** drift

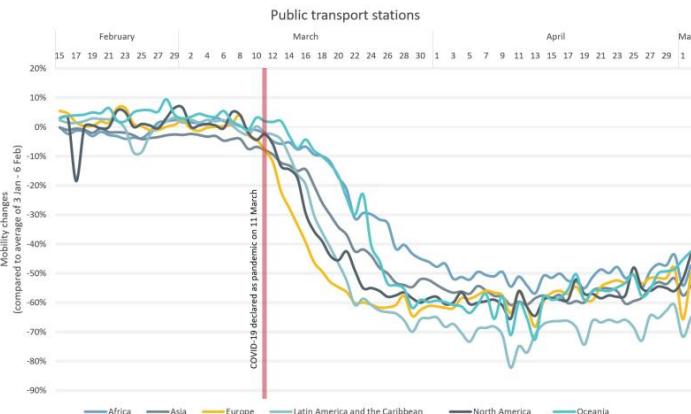
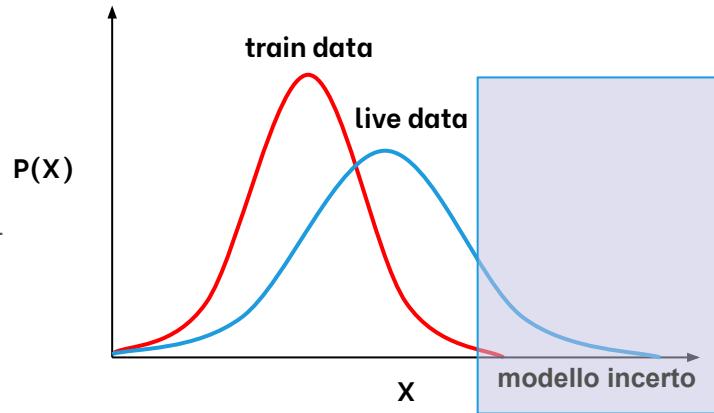
Data drift

Distribuzioni di probabilità delle features in input cambiano nel tempo.

Motivi possono essere molteplici.

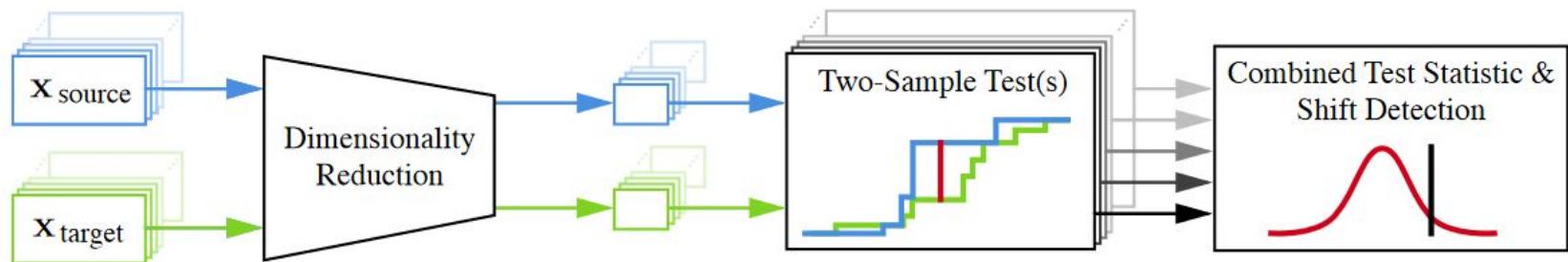
Statisticamente misurabile usando test come:

- Chi-square
- Maximum Mean Discrepancy (MMD)
- ...



Data drift

Problema: come misurare drift in combinazioni di variabili o in features nominali?

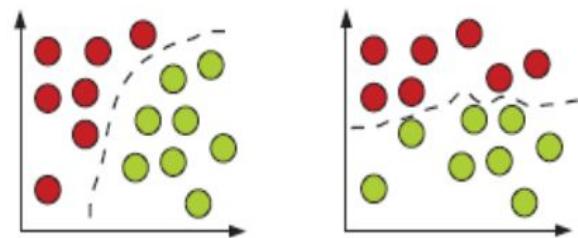


Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift, <https://arxiv.org/pdf/1810.11953.pdf>

Concept drift

Cambiamento nella relazione tra distribuzioni in input e **output** può cambiare considerevolmente nel tempo a seconda del problema in esame.

Distribuzioni nello spazio delle features possono rimanere simili quello che cambia in questo caso è il decision boundary del problema



HAL Id : hal-02062610, version 1

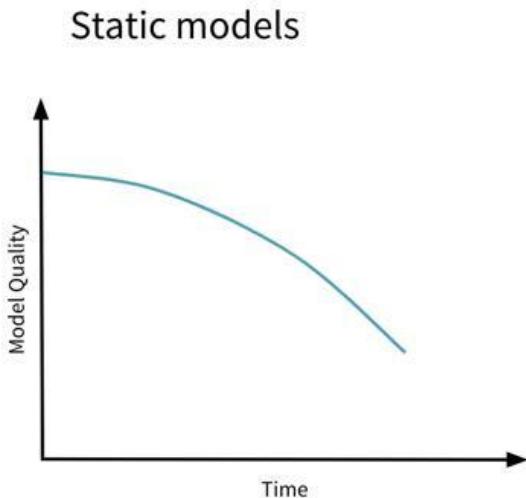
Concept drift

Esempi:

Churn prediction: nuovo provider telefonico low-cost entra nel mercato → churn rate cambia drasticamente a parità di distribuzioni in input

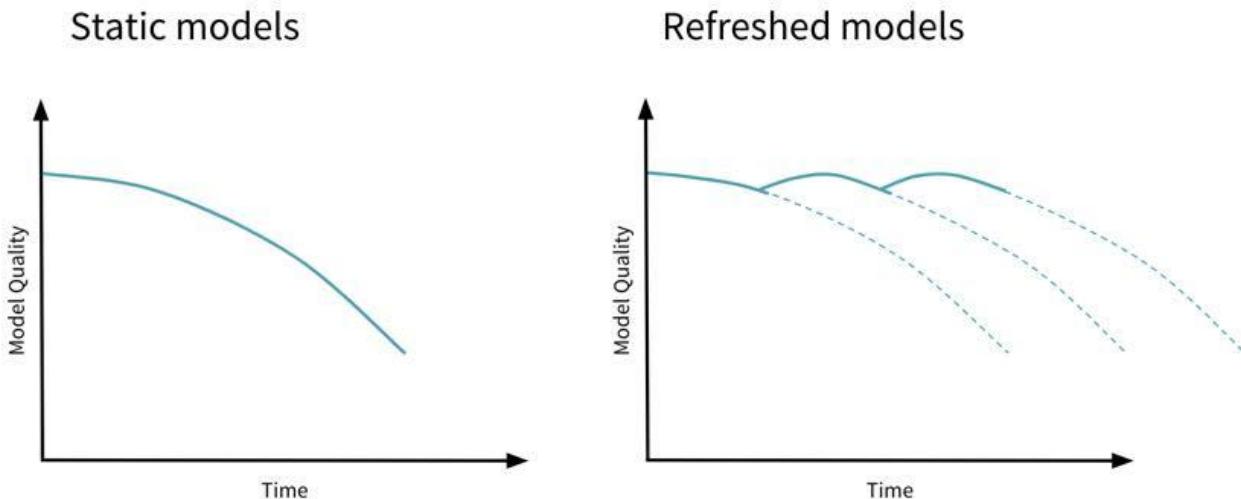
Sentiment analysis: ‘*This song is sick!*’ avrebbe assunto un significato completamente diverso prima degli anni ‘90.

Concept e data drift



<https://databricks.com/blog/2019/09/18/productionizing-machine-learning-from-deployment-to-drift-detection.html>

Concept e data drift



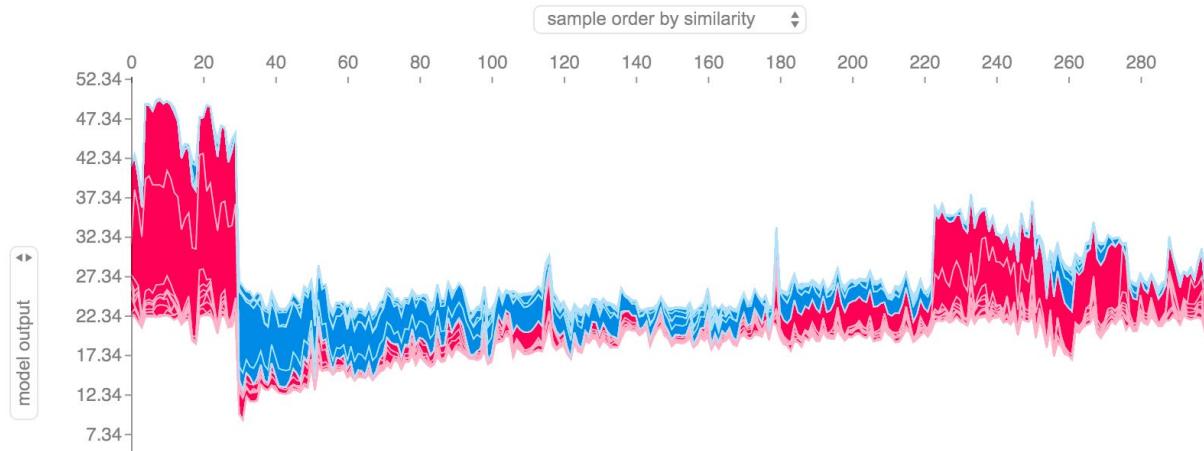
<https://databricks.com/blog/2019/09/18/productionizing-machine-learning-from-deployment-to-drift-detection.html>

Monitoraggio delle spiegazioni

delle decisioni del modello

Spiegazioni delle decisioni del modello possono essere viste come un modo di descrivere quantitativamente il suo comportamento.

Lo stesso discorso può essere affrontato in ambito di monitoraggio: **come e quanto sta cambiando il comportamento del modello nel tempo?**



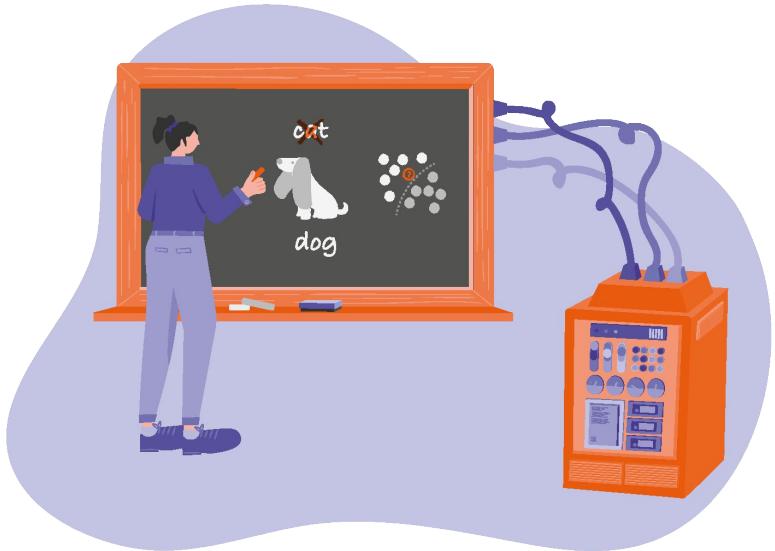
Monitoraggio performance nel tempo

= Bisogno di etichettatura continua

Monitorate performance di modelli (accuracy, F1, etc) richiede la conoscenza della ground truth relativa ai dati che arrivano dal vivo.

Due possibilità:

- Processo automatico
→ problemi di forecasting
- Processo richiede etichettatura umana
→ Esempio, decision support system

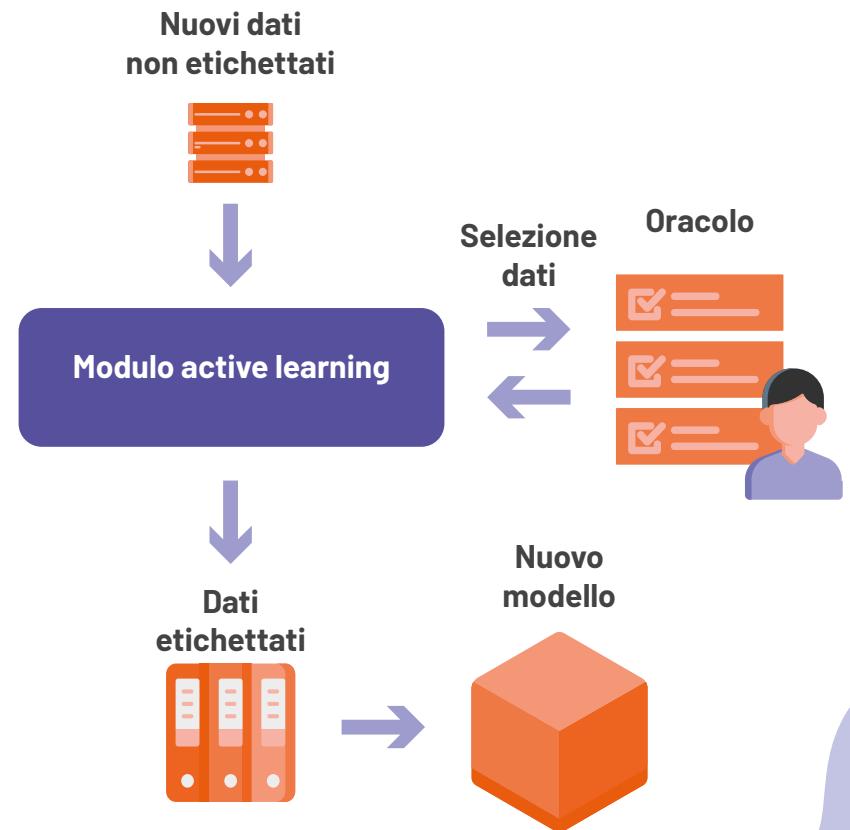


Active Learning

In molti casi etichettare dati può essere costoso in termini di tempo e denaro.

Lo scopo dell'active learning è di definire strategie atte a migliorare il processo di ri-allenamento dei modelli usando il minor numero di nuove etichette possibile.

In fisica lo stesso tipo di problema viene definito come '*Design Of Experiments*'



Active Learning

Come scegliere i punti da etichettare?

Diverse strategie possibili. Tra le più popolare strategie basate sulla riduzione dell'incertezza del modello che si vuole migliorare (**uncertainty sampling**)

| Prediction # | Prediction |
|--------------|------------|
| 1 | 0.99 |
| 2 | 0.8 |
| 3 | 0.1 |
| 4 | 0. |
| 5 | 0.6 |
| 6 | 0.4 |
| 7 | 1 |
| ... | ... |





Thanks for Reading

Feel free to contact us:



www.clearbox.ai



support@clearbox.ai



@ClearboxAI

Introduzione a pandas

- Introduzione Pandas
- Esempio pratico pipeline ML
- Esempio applicazione SHAP
- Discussione progetto AITravel



Introduzione a pandas



Libreria per manipolazione e analisi dati, la più usata quando si ha a che fare con dati strutturati e serie temporali.

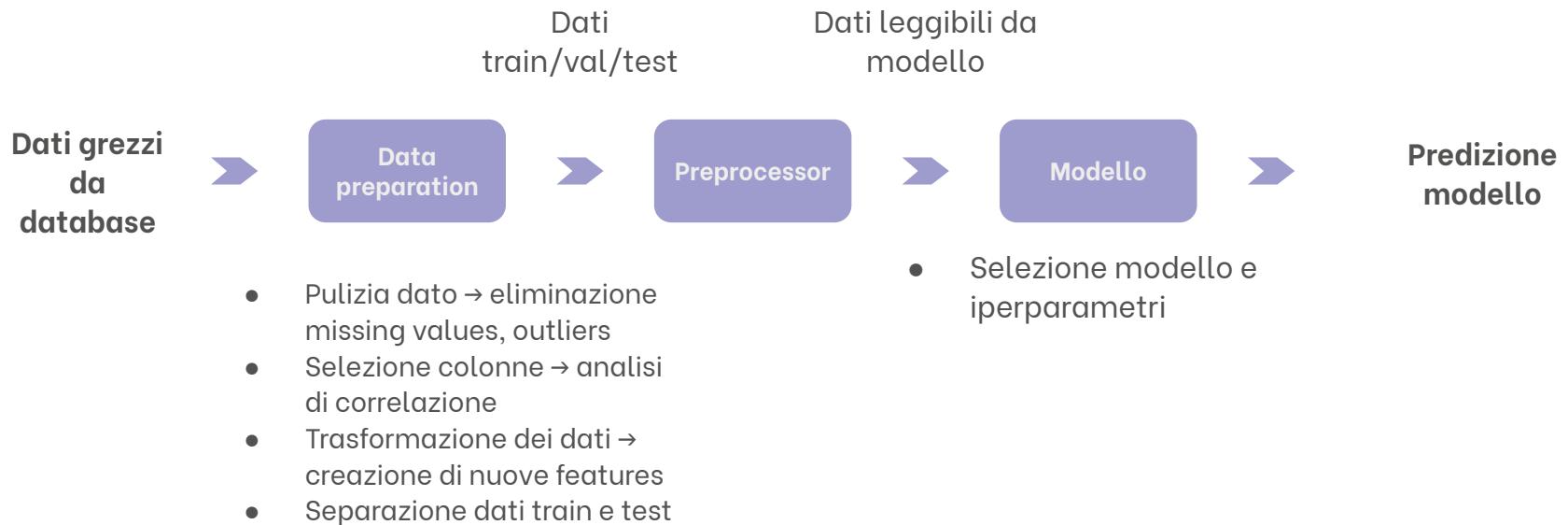
Scritta in Python, Cython e C, molto efficiente quando non si devono fare calcoli paralleli. Per processamento dati su larga scala non è l'alternativa migliore (in questo caso meglio usare Spark o PySpark)

pandas è la libreria più usata in ambito machine learning per la **pulizia, preparazione e preprocessamento** dei dati per il modelli.

Esempio Pandas

Esempio

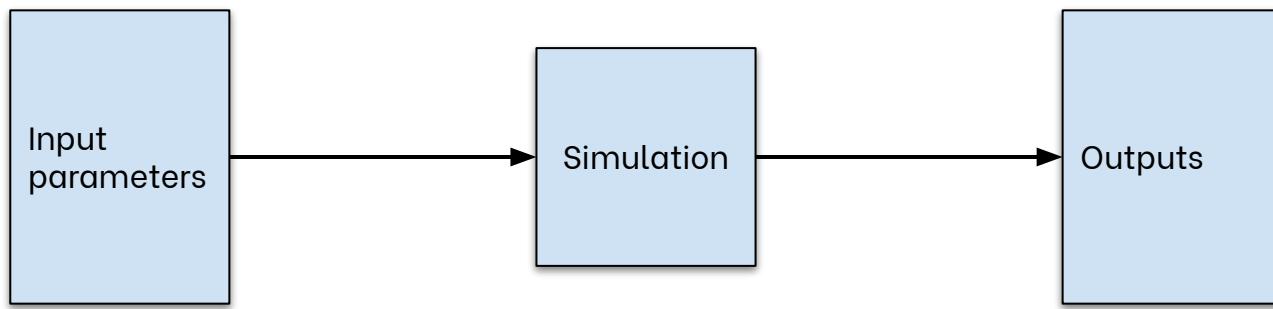
Problema giocattolo → Classificazione binaria (Adult Income Dataset)



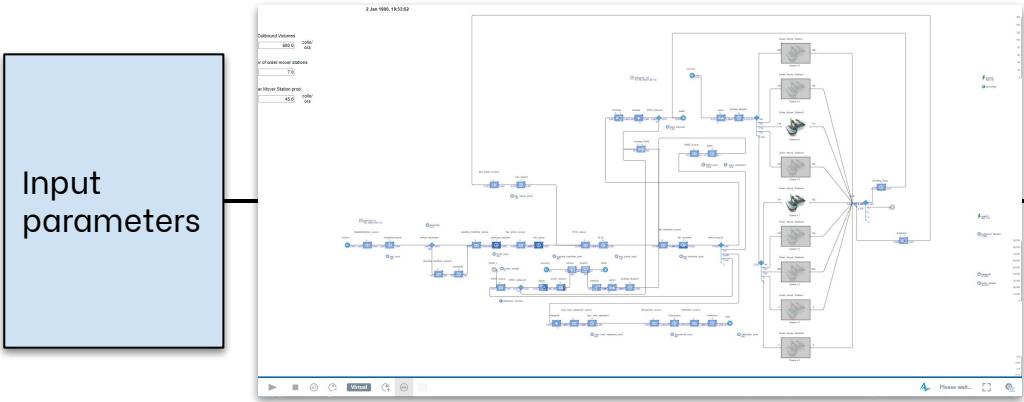
Esempio pipeline ML

Esempio SHAP

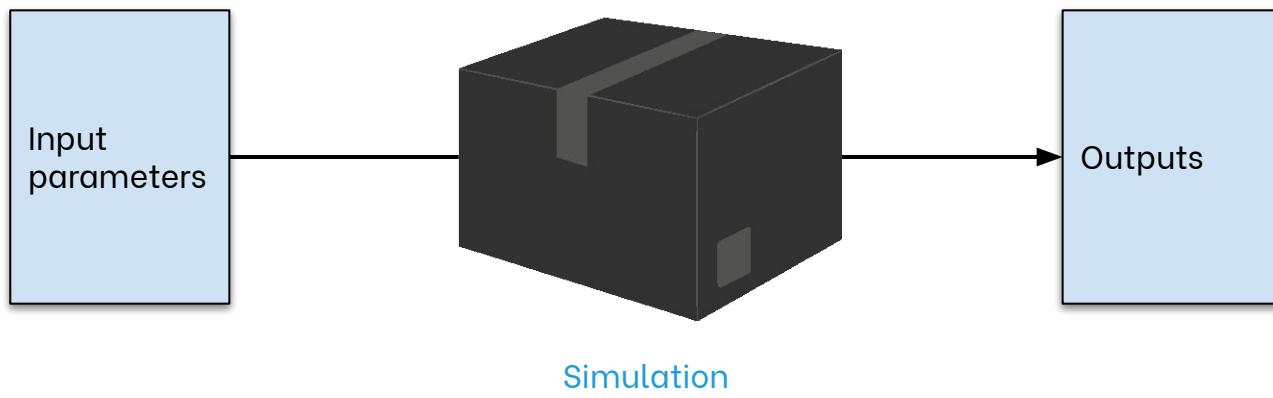
Progetto AITravel

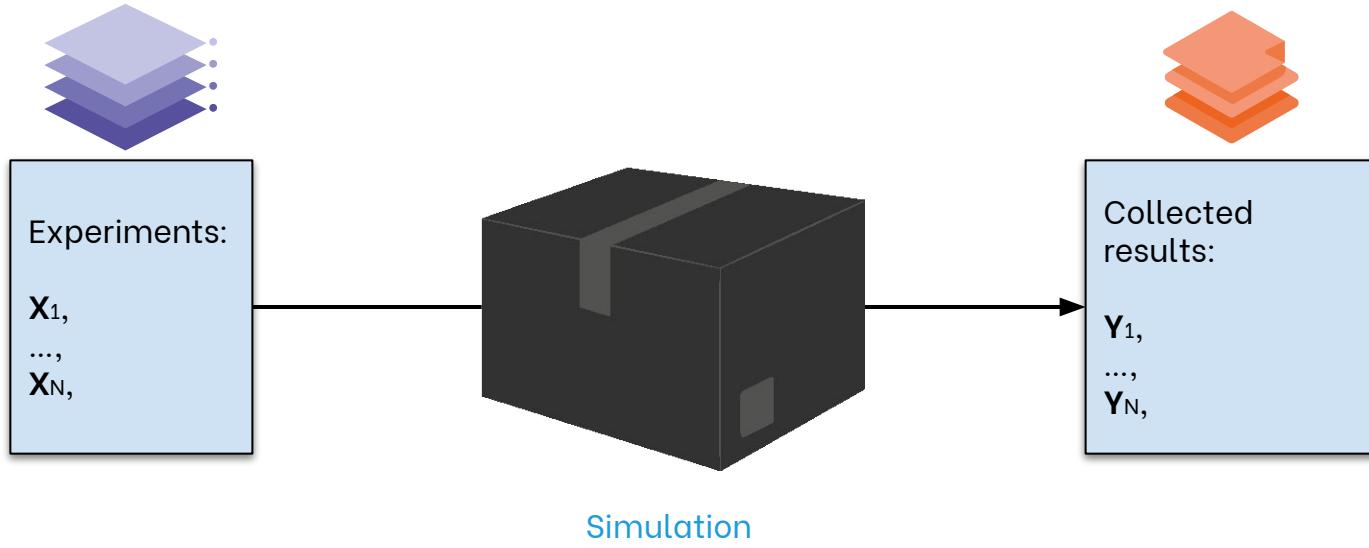


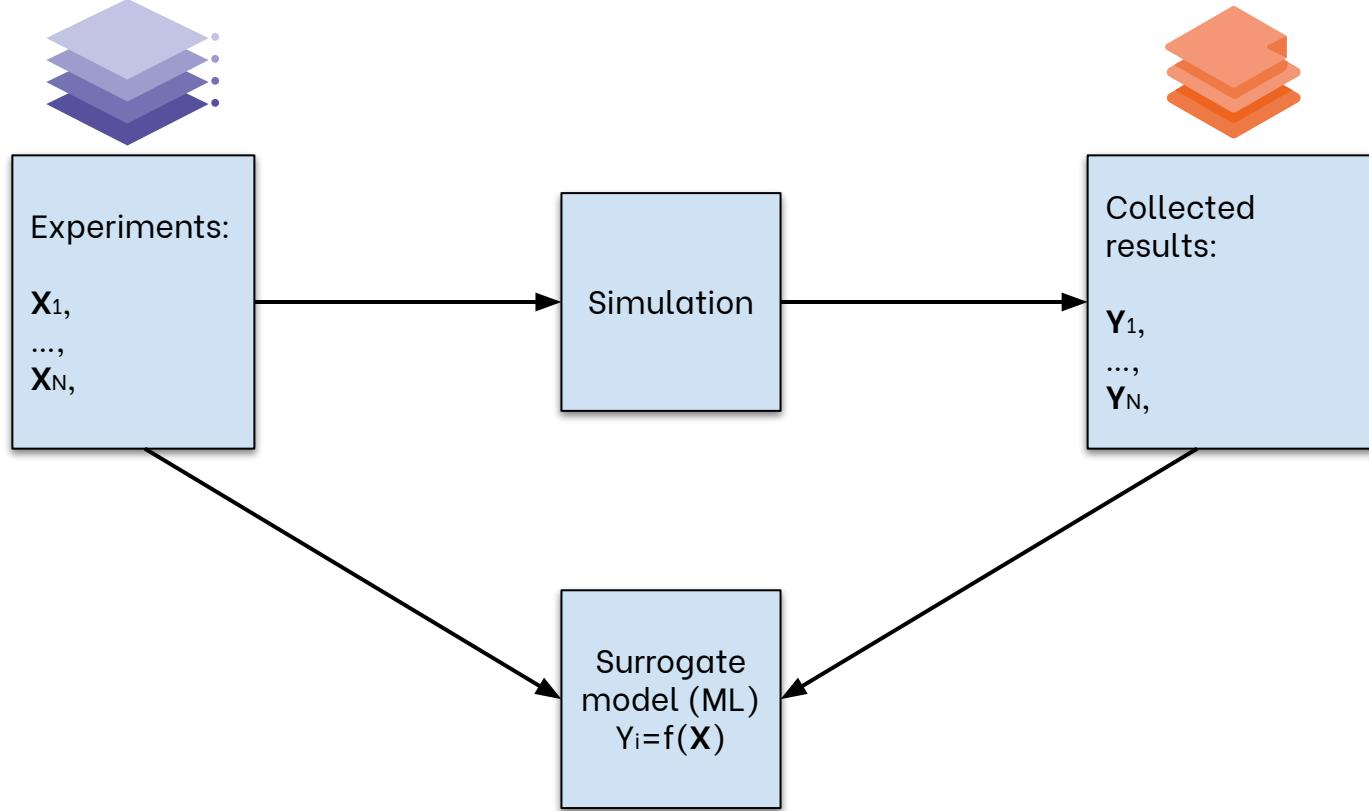
Input
parameters

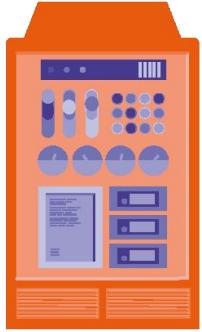


→
Outputs

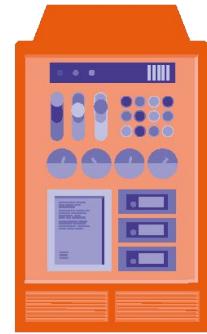






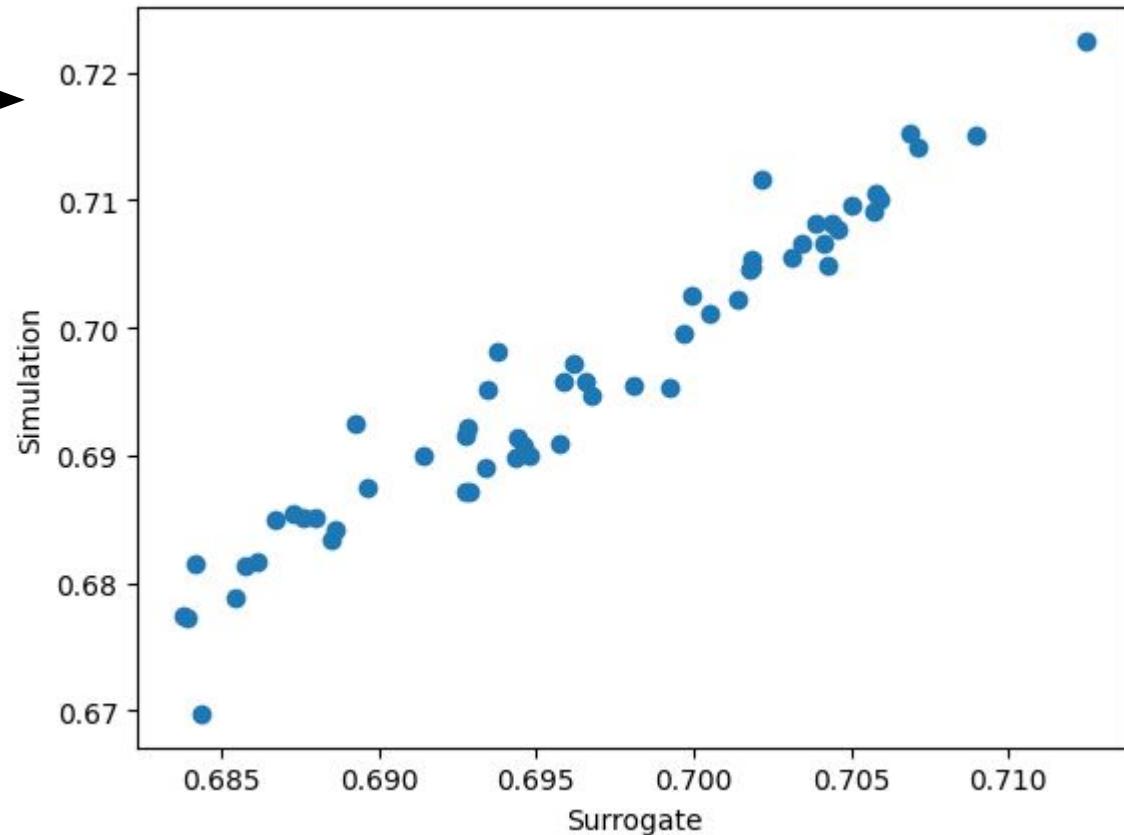


Surrogate model

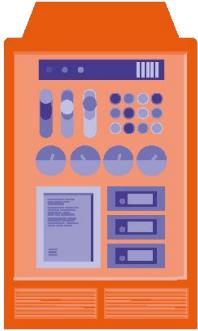


Surrogate model

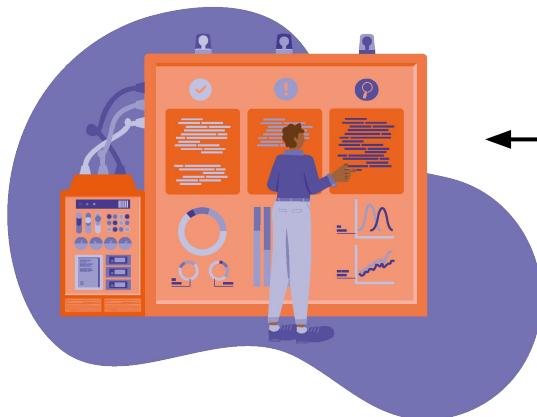
Surrogate model performance analysis



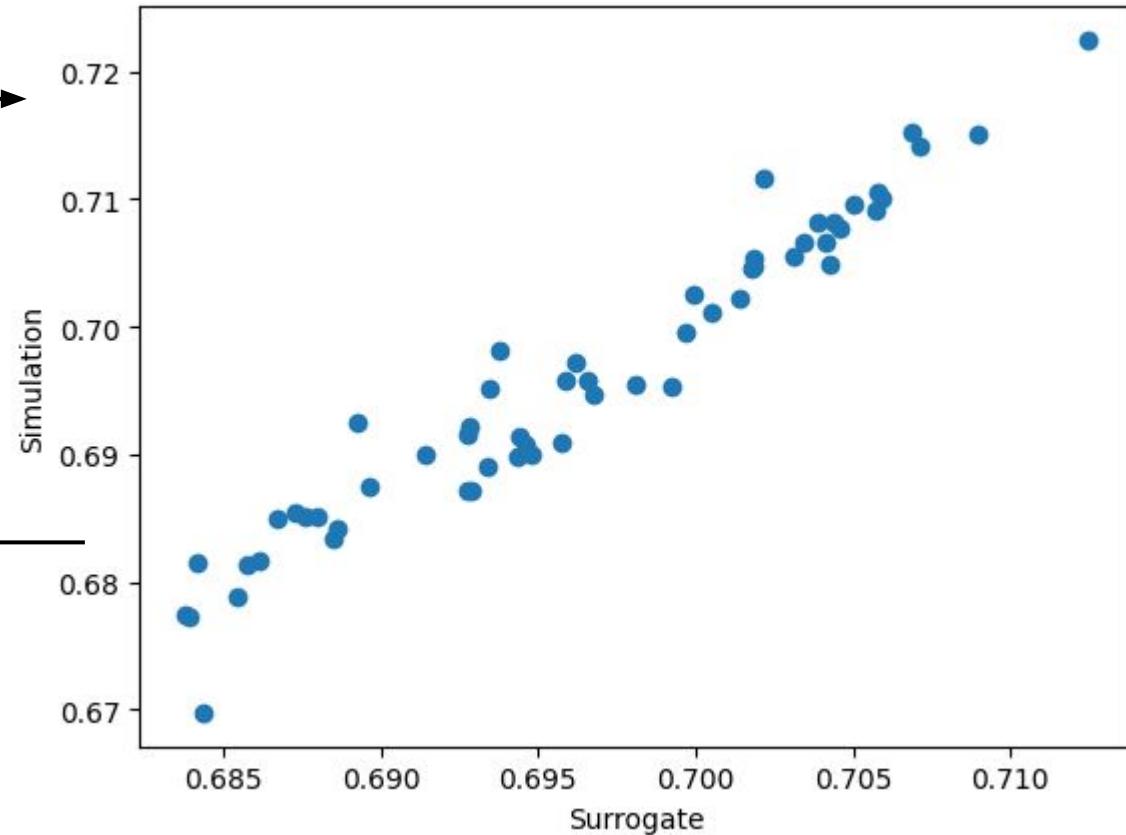
Surrogate model performance analysis

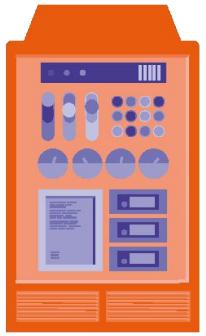


Surrogate model

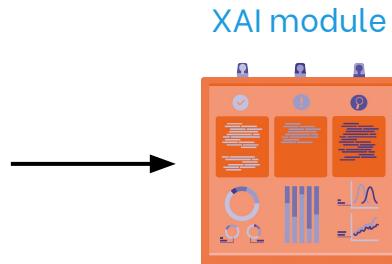


Active learning: defining new experiments

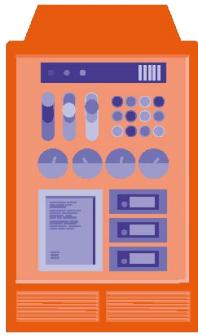




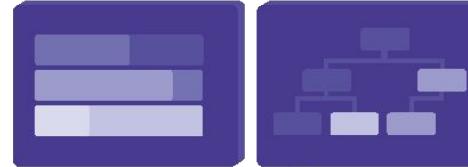
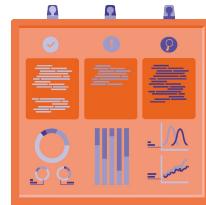
Surrogate model



XAI module

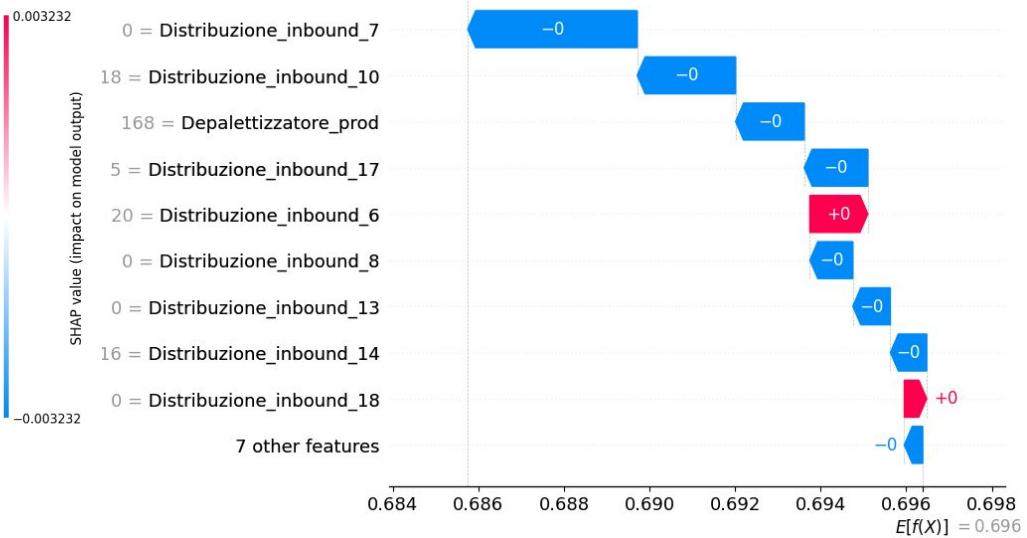
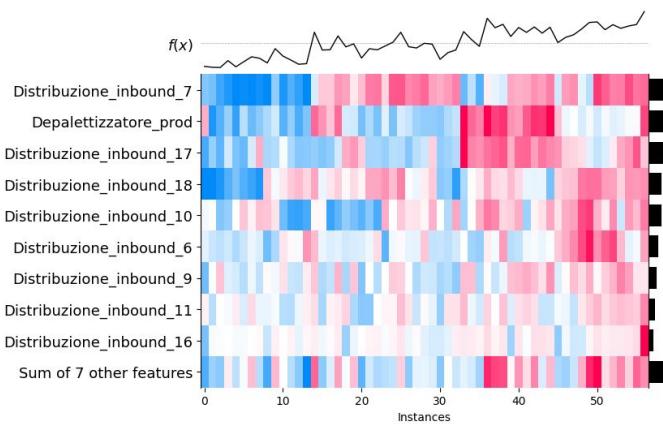


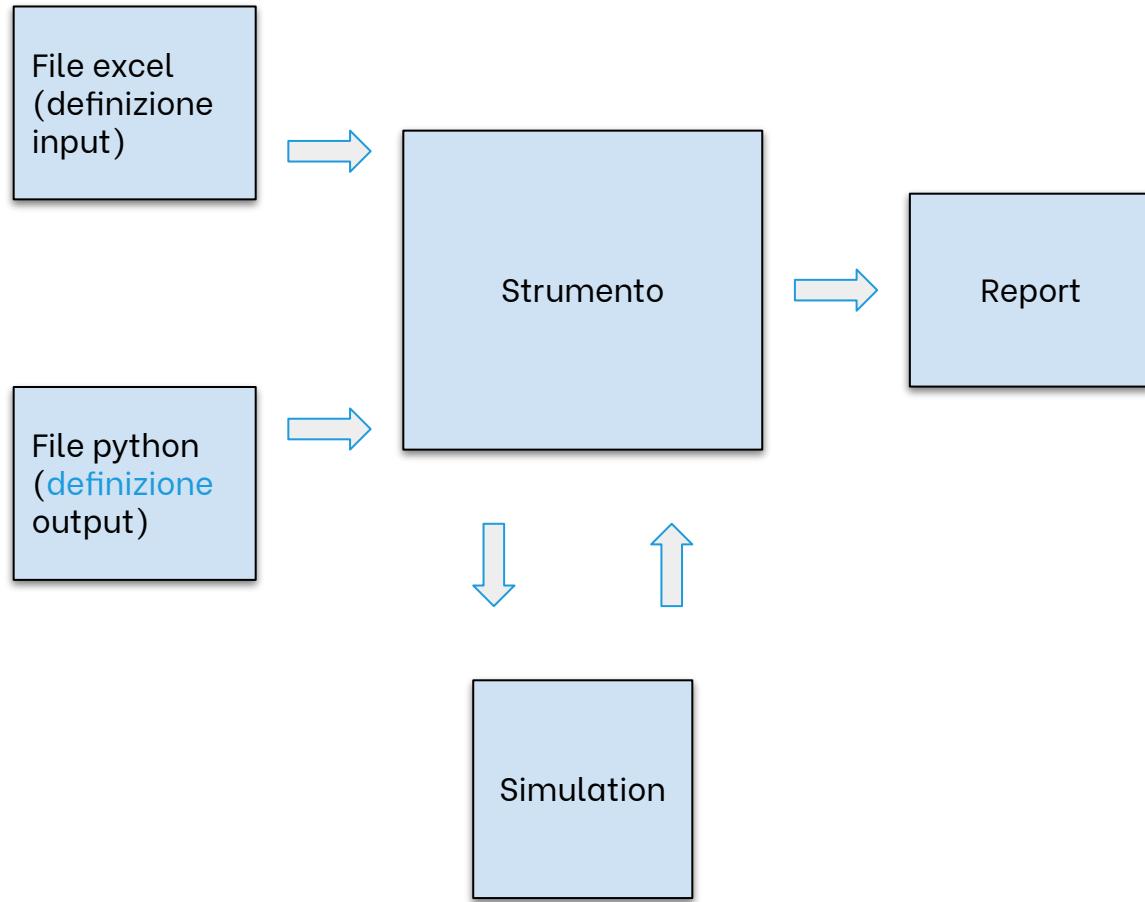
XAI module



Local explanations

Surrogate model





Input

- Idea: permettere all'utente di definire input simulazione direttamente da file excel contenente placeholders.

| A | B | C |
|-------------|-------------|---------------|
| Data i | Data f | Distribuzione |
| 12:00:00 am | 1:00:00 am | 0% |
| 1:00:00 am | 2:00:00 am | 0% |
| 2:00:00 am | 3:00:00 am | 0% |
| 3:00:00 am | 4:00:00 am | 0% |
| 4:00:00 am | 5:00:00 am | 0% |
| 5:00:00 am | 6:00:00 am | 0% |
| 6:00:00 am | 7:00:00 am | input_1, 0-20 |
| 7:00:00 am | 8:00:00 am | input_2, 0-20 |
| 8:00:00 am | 9:00:00 am | input_3, 0-20 |
| 9:00:00 am | 10:00:00 am | input_4, 0-20 |
| 10:00:00 am | 11:00:00 am | input_5, 0-20 |

Output

- Utente definisce output rilevanti modello tramite un insieme di funzioni python.

```
def output1():
    folder_path = 'c:\\\\Users\\\\gillu\\\\stam\\\\Modello AnyLogic_AITrawel_v0\\\\Risultati\\\\'

    all_files = [os.path.join(folder_path, f) for f in os.listdir(folder_path)
                if os.path.isfile(os.path.join(folder_path, f))]

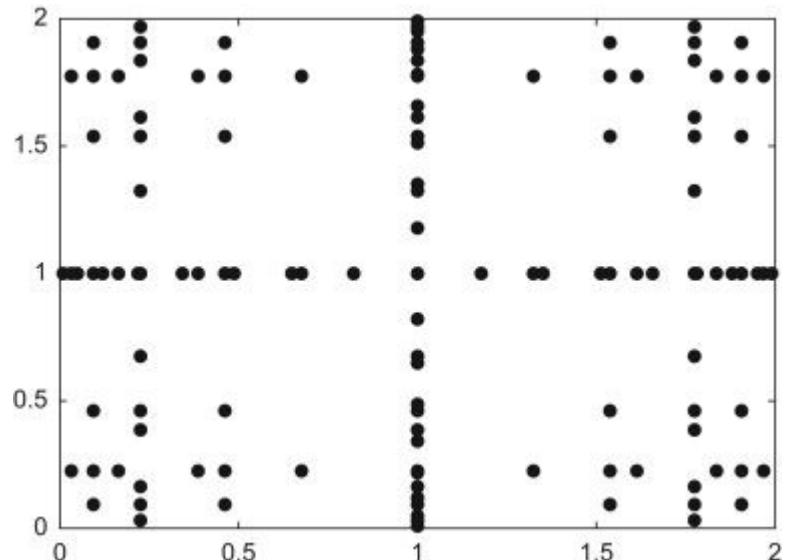
    all_files.sort(key=os.path.getctime, reverse=True)

    outputs = all_files[:7]
    x = pd.read_csv(outputs[0], sep='\t')
    print(x)
    y = {}
    y['Promessa_outbound'] = x.iloc[:,1].mean()

    return y
```

Active learning

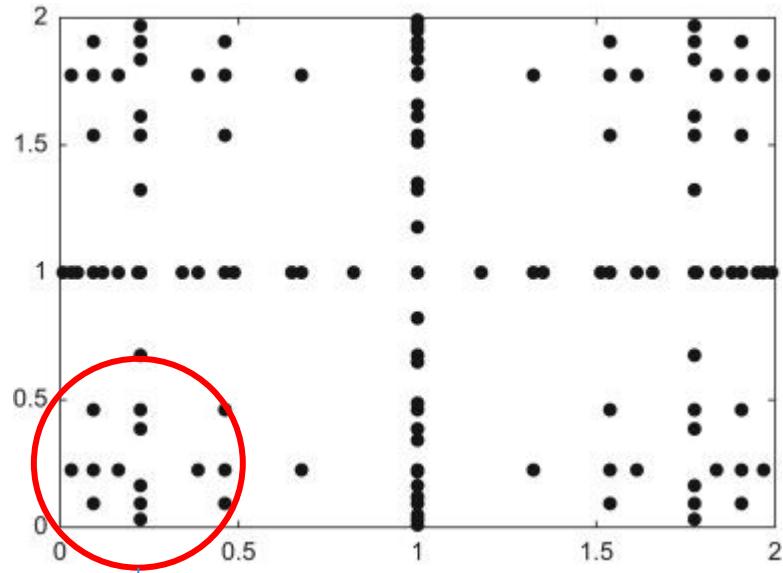
- Implementato approccio adattivo basato su concetto di Polynomial Chaos
- Batch di queries per simulatore fino al raggiungimento di convergenza.



Interpretable ML

Analizzate diverse alternative per generazione spiegazioni

- SHAP per spiegazioni locali e globali di modello
- Alberi decisionali locali per spiegazioni di tipo IF-THEN



Allenamento decision tree su sottoinsieme di punti (determinati su grid)

Esempio dal vivo

Large Language Models

Large Language Models

Cosa sono i Large Language Models?

Modelli in grado di leggere un testo in input e restituire un testo in output.

Trasformazione input-output ottenuta per mezzo di miliardi di parametri.

Input Prompt: Recite the first law of robotics



Output:

Attention is all you need

Attention Is All You Need

Articolo del **2017** di Google introduce architettura creata per migliorare le performance dei modelli di traduzione. Consente ai modelli di avere rappresentazioni delle parole in funzione del contesto in cui sono scritte.

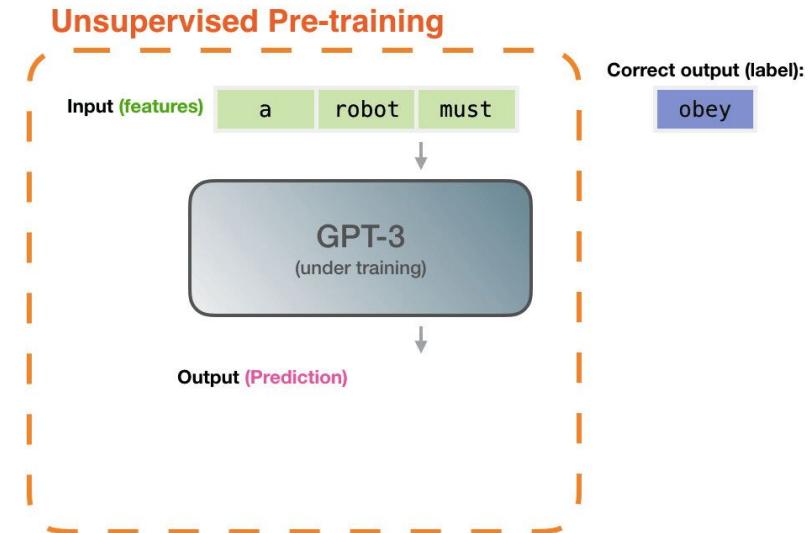
- Pesca una carta e provo ad indovinarla
- Sono andato al supermercato per comprare qualche pesca
- E' necessaria una licenza di pesca

Prime applicazioni dell'architettura effettuate in ambito **supervisionato**: modelli allenati utilizzando una serie di esempi **input→output** da cui apprendere.

Unsupervised pre-training

Idea di fondo: possiamo allenare modelli di linguaggio senza task di apprendimento specifici?

Allenando il modello a prevedere la prossima parola in frasi provenienti da un corpus non etichettato.



GPT-1

Modello presentato da OpenAI Giugno **2018**.

- 116 Milioni di parametri
- Allenato utilizzando 5Gb di dati
- 8 GPU utilizzate per un mese

Completamente open-source e riproducibile.

Dimostra che pre-allenamento consente di migliorare performance di modelli supervisionati.

Improving Language Understanding by Generative Pre-Training

Alec Radford
OpenAI
alec@openai.com

Karthik Narasimhan
OpenAI
karthikn@openai.com

Tim Salimans
OpenAI
tim@openai.com

Ilya Sutskever
OpenAI
ilyasu@openai.com

Abstract

Natural language understanding comprises a wide range of diverse tasks such as textual entailment, question answering, semantic similarity assessment, and document classification. Although large unlabeled text corpora are abundant, labeled data for learning these specific tasks is scarce, making it challenging for discriminatively trained models to perform adequately. We demonstrate that large gains on these tasks can be realized by *generative pre-training* of a language model on a diverse corpus of unlabeled text, followed by *discriminative fine-tuning* on each specific task. In contrast to previous approaches, we make use of task-aware input transformations during fine-tuning to achieve effective transfer while requiring minimal changes to the model architecture. We demonstrate the effectiveness of our approach on a wide range of benchmarks for natural language understanding. Our general task-agnostic model outperforms discriminatively trained models that use architectures specifically crafted for each task, significantly improving upon the state of the art in 9 out of the 12 tasks studied. For instance, we achieve absolute improvements of 8.9% on commonsense reasoning (Stories Cloze Test), 5.7% on question answering (RACE), and 1.5% on textual entailment (MultiNLI).

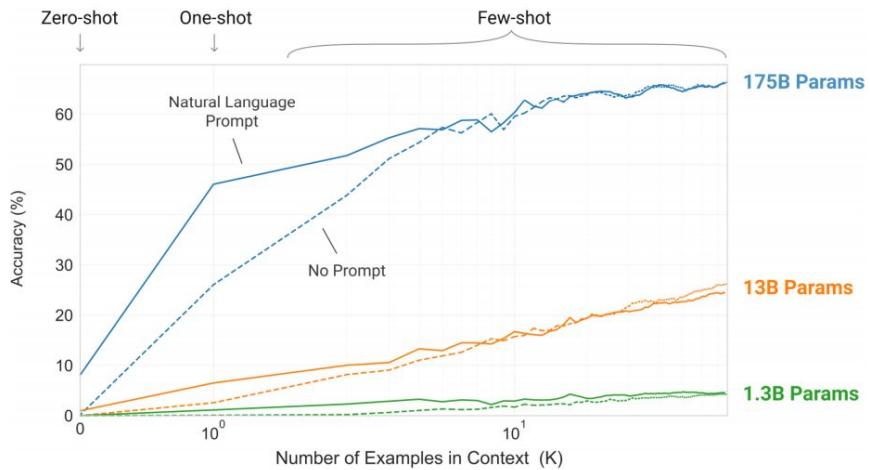


GPT-3

Presentato da OpenAI nel Giugno 2020.

Pre-training su centinaia di miliardi di parole → ottimi risultati su task supervisionati anche solo fornendo pochi esempi.

- 45 TB
- 175B parametri
- 4.5 M\$



GPT-3

Presentato da OpenAI nel Giugno 2020.

Pre-training su centinaia di miliardi di parole → ottimi risultati su task supervisionati anche solo fornendo pochi esempi.

- 45 TB
- 175B parametri
- 4.5 M\$



<https://arxiv.org/abs/2005.14165>

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.



One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.



Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

