

Fondamenti di Data Science e Machine Learning

Progetti 2024-2025

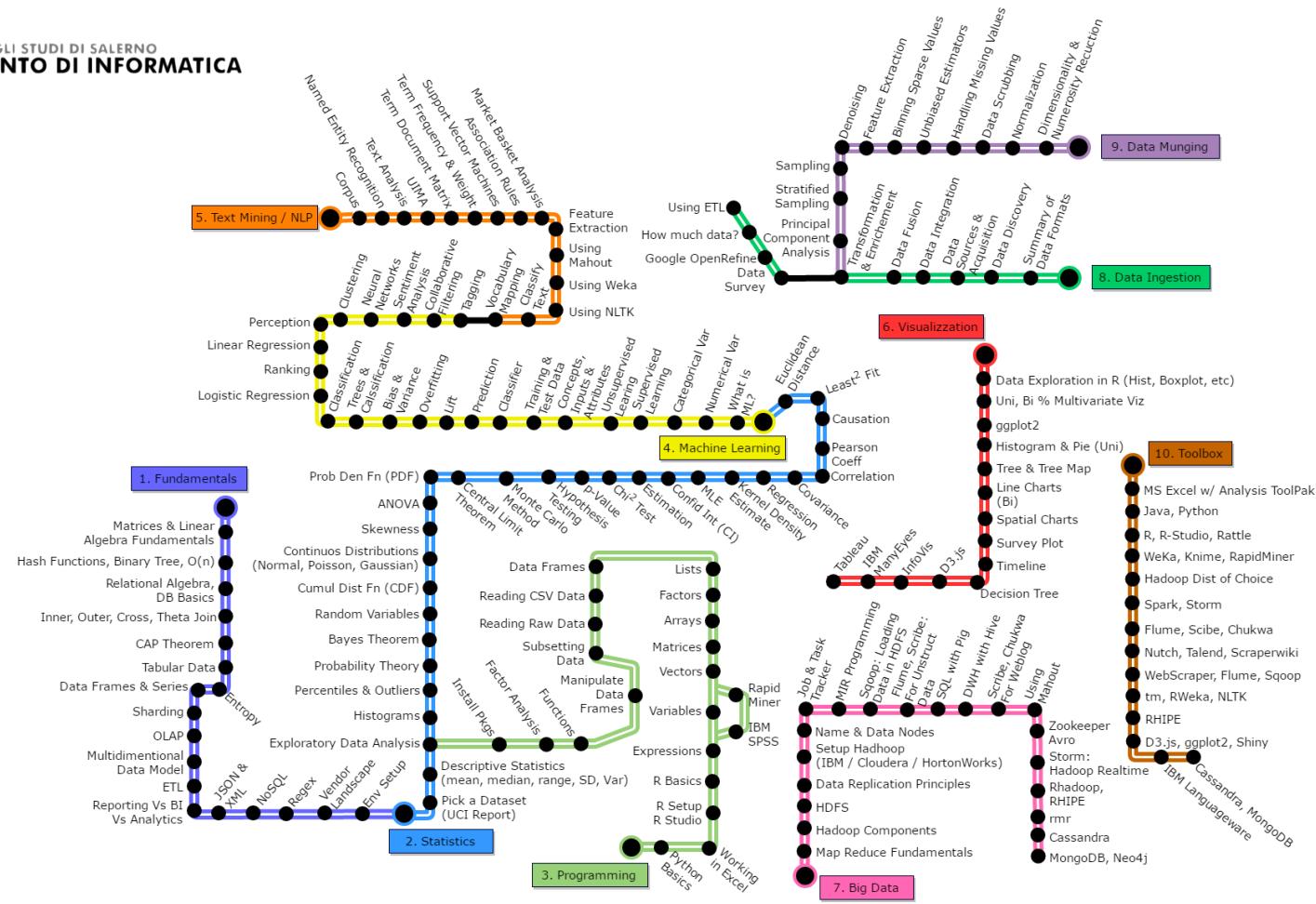
ITADATAhack



The landing page for the III Edition of ITADATAhack. It features a large banner with the event name and a city skyline background. Below the banner, text indicates the hackathon runs from August 29-31 (online) and September 9-11 (awards in Turin). A paragraph describes the event as the national hackathon of the CINI Data Science Laboratory, returning at the end of August, with winners awarded at ITADATA 2025 in Turin. A "Subscribe" button is present. At the bottom, a red box displays a timer: 127 Giorni, 20 Ore, 05 Minuti, 20 Secondi.

- L'evento, per team, sarà aperto a tutti gli studenti di università italiane e sarà focalizzato su una sfida di analisi dei dati di un database aziendale privato
- L'obiettivo è far lavorare gli studenti su dataset reali e ingaggiarli in una competizione che permetta loro di mettere in mostra le proprie competenze in ambito analisi e modellazione dei dati ed HPC
- **Partecipando** a ITADATAhack 2025 come player I team potranno analizzare dati unici sul mercato STEM italiano, cercando di collaborare in maniera innovativa con i membri del tuo team
- L'hackathon sarà composto da una gara di Data Science esplorativa con classifica temporanea e **durerà 4 giorni**
- Verrà richiesto ai team di analizzare dati provenienti da un database aziendale privato
- Le iscrizioni saranno aperte a tutti gli atenei nazionali indipendentemente dall'appartenenza al Laboratorio Data Science del CINI
- I migliori team verranno **invitati** a presentare i propri lavori durante “ITADATA 2025” a Torino



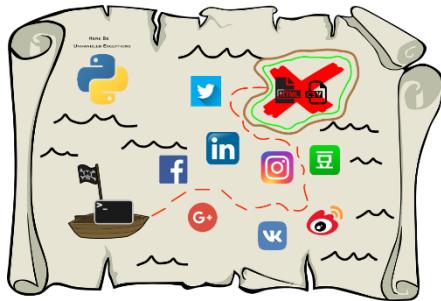


Fondamenti di Data Science e Machine Learning Privacy Preserving ML

Cybersecurity in Social Network

SOcial Mapper

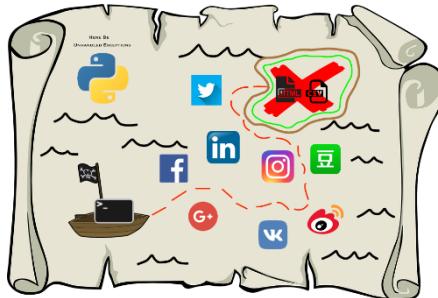
Find Social Media Profiles Using Photo Only



Cybersecurity in Social Network

Social Mapper

Find Social Media Profiles Using Photo Only



Data Analyzer



Profile: linkedin.com/in/williamhgates

Profile: facebook.com/BillGates

Profile: instagram.com/thisisbillgates

Profile: vk.com/bilgatess

Profile: twitter.com/billgates

Facebook				
	Name	Surname	Email	Phone
ID1	Bill	Gates	gates@microsoft.it	🔒
ID2	Mario	Rossi	✉️	🔒
ID3	Mary	Johnson	✉️	✉️
ID4	Emily	Scott	✉️	✉️

Linkedin				
	Job	City	Email	Phone
ID1	CEO	Washington	✉️	🔒
ID2	Developer	Rome	rossi@em.com	39135..
ID3	Professor	London	galilei@auk.it	177445...

	Name	Surname	Email	Phone	Job	City	Email_Ln	Phone_Ln
ID1	Bill	Gates	gates@microsoft.it	✉️	CEO	Washington	✉️	🔒
ID2	Mario	Rossi	✉️	✉️	Developer	Rome	rossi@em.com	39135..
ID3	Mary	Johnson	✉️	✉️	Professor	London	john1@auk.it	177445...
ID4	Emily	Scott	✉️	✉️	?	?	?	?

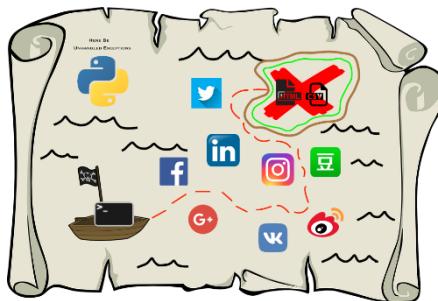
Person Not Found

Data Privatized

Data Reconstructed

Cybersecurity in Social Network

Social Mapper
Find Social Media Profiles Using Photo Only

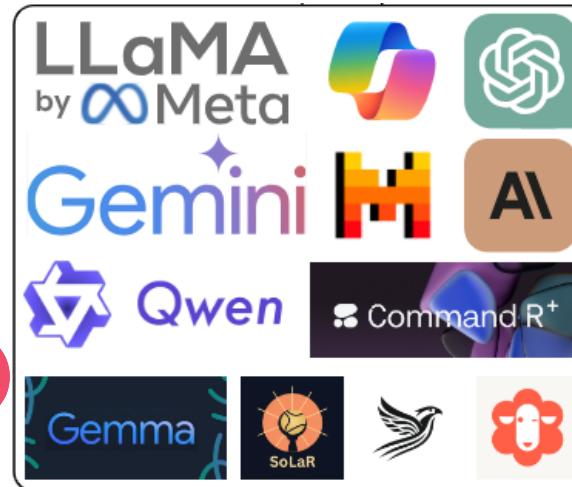


+

Data Analyzer



Large Language Model



Web Data			
	Personal Site	Newspaper	Photos
ID1	http://id1.com	NY Times 1	Photo3.png
ID2	http://id2.it	Corriere del..	Photo2.jpg
ID3	http://id3.org	Il Mattino	Photo1.jpg

Name	Surname	Email	Phone	Job	City	Email_Ln	Phone_Ln
ID1	Bill	Gates	gates@microsoft.it	CEO	Washington	✉️	✉️
ID2	Mario	Rossi	✉️	Developer	Rome	rossi@em.com	39135..
ID3	Mary	Johnson	✉️	Professor	London	john1@auk.it	177445...
ID4	Emily	Scott	✉️	?	?	?	?

❓ Person Not Found

🔒 Data Privatized

🔓 Data Reconstructed

Cybersecurity in Social Network

- Il progetto è mirato allo studio della sensibilità dei dati associati agli utenti su differenti social network (analisi cross social)
 - Sviluppo e ricerca di modelli AI e sistemi di scraping automatici
 - Migliorare il sistema di riconoscimento degli utenti sui social network tramite modelli di fare recognition avanzati
 - Utilizzare sistemi di data reconstruction per la creazione di profili/cloni digitali
 - Valutare le capacità di ricostruzione delle informazioni di Large Language Model generativi e compararle con l'attuale sistema



Sergio



Privacy Preserving Machine Learning

- ▶ La privatizzazione dei dati è il processo di mascheramento o rimozione di informazioni sensibili dai dati in modo che non possano essere utilizzate per identificare gli individui.
- ▶ Questo può essere fatto utilizzando una varietà di tecniche, tra cui:
 - ▶ **Anonimizzazione:** **Rimozione** di tutte le informazioni di identificazione personale (PII) dai dati
 - ▶ **Pseudonimizzazione:** **Sostituzione** delle informazioni PII con pseudonimi
 - ▶ **Aggregazione:** **Aggregazione** dei dati in gruppi in modo che non sia possibile identificare gli individui
 - ▶ **Generalizzazione:** **Generalizzazione** dei dati in modo che non siano più specifici



Privacy Preserving Machine Learning

ID Paziente	Sesso	Regione	Età	Data del ricovero	Città	Nome	Cognome	Sintomi	Diagnosi generica
12345	Maschio	Nord Italia	31	2020	Roma	Mario	Rossi	Dolore al petto, affanno, vertigini	Malattie del sistema cardiovascolare
67890	Femmina	Centro Italia	45	2021	Roma	Anna	Bianchi	Tosse, respiro sibilante, difficoltà respiratorie	Malattie respiratorie
23456	Maschio	Sud Italia	28	2022	Morterone	Luca	Verdi	Affaticamento, perdita di peso, febbre	Malattie neoplastiche
78901	Femmina	Nord Italia	55	2023	Salerno	Giulia	Neri	Dolore addominale, nausea, vomito	Malattie del sistema digestivo
45678	Maschio	Centro Italia	63	2023	Napoli	Marco	Bianchi	affaticamento	Malattie metaboliche



- ▶ Questi dati non sono utilizzabili per addestrare modelli predittivi
 - ▶ Contengono informazioni sensibili (Es. Nome, Cognome, Età)
 - ▶ Non possono essere condivisi



ID Paziente	Sesso	Regione	Età	Data del ricovero	Città	Nome	Cognome	Sintomi	Diagnosi generica
12345	Maschio	Nord Italia	****	2020	Roma	*****	Rossi	Dolore al petto, affanno, vertigini	Malattie del sistema cardiovascolare
67890	Femmina	Centro Italia	45	2021	Roma	*****	Bianchi	Tosse, respiro sibilante, difficoltà respiratorie	Malattie respiratorie
23456	Maschio	Sud Italia	28	2022	Morterone	*****	Verdi	Affaticamento, perdita di peso, febbre	Malattie neoplastiche
78901	Femmina	Nord Italia	****	2023	Salerno	Giulia	*****	Dolore addominale, nausea, vomito	Malattie del sistema digestivo
45678	Maschio	Centro Italia	****	2023	Napoli	Marco	*****	affaticamento	Malattie metaboliche

Privacy Preserving Machine Learning

ID Paziente	Sesso	Regione	Età	Data del ricovero	Città	Nome	Cognome	Sintomi	Diagnosi generica
12345	Maschio	Nord Italia	31	2020	Roma	Mario	Rossi	Dolore al petto, affanno, vertigini	Malattie del sistema cardiovascolare
67890	Femmina	Centro Italia	45	2021	Roma	Anna	Bianchi	Tosse, respiro sibilante, difficoltà respiratorie	Malattie respiratorie
23456	Maschio	Sud Italia	28	2022	Morterone	Luca	Verdi	Affaticamento, perdita di peso, febbre	Malattie neoplastiche
78901	Femmina	Nord Italia	55	2023	Salerno	Giulia	Neri	Dolore addominale, nausea, vomito	Malattie del sistema digestivo
45678	Maschio	Centro Italia	63	2023	Napoli	Marco	Bianchi	affaticamento	Malattie metaboliche



Morterone, il più piccolo comune di Italia, conta 33 residenti

L'unico cittadino con cognome Verdi di anni 28 si chiama Luca

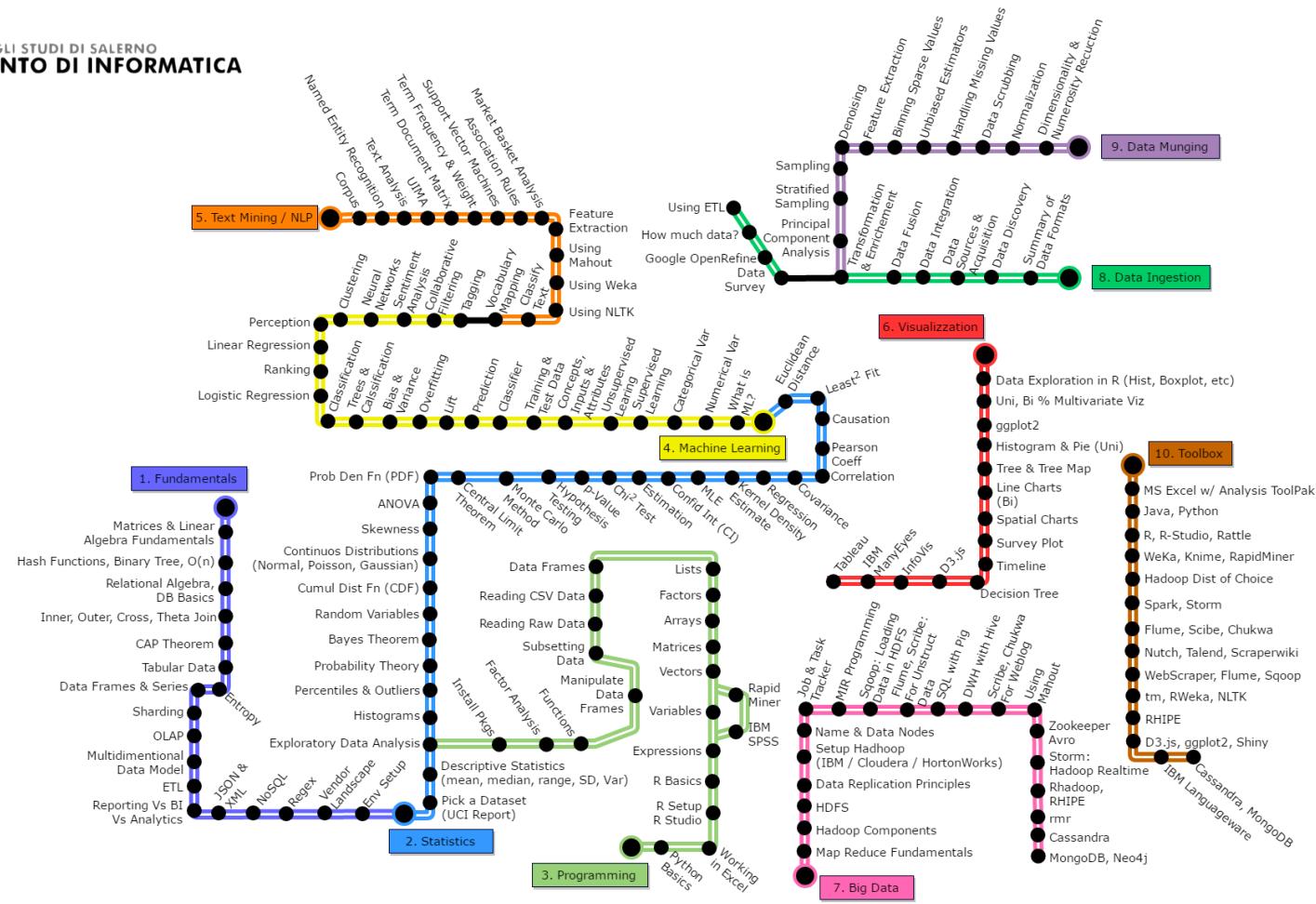
- ▶ Questi dati non sono utilizzabili per addestrare modelli predittivi
 - ▶ Contengono informazioni sensibili (Es. Nome, Cognome, Età)
 - ▶ Non possono essere condivisi



ID Paziente	Sesso	Regione	Età	Data del ricovero	Città	Nome	Cognome	Sintomi	Diagnosi generica
12345	Maschio	Nord Italia	****	2020	Roma	*****	Rossi	Dolore al petto, affanno, vertigini	Malattie del sistema cardiovascolare
67890	Femmina	Centro Italia	45	2021	Roma	*****	Bianchi	Tosse, respiro sibilante, difficoltà respiratorie	Malattie respiratorie
23456	Maschio	Sud Italia	28	2022	Morterone	*****	Verdi	Affaticamento, perdita di peso, febbre	Malattie neoplastiche
78901	Femmina	Nord Italia	****	2023	Salerno	Giulia	*****	Dolore addominale, nausea, vomito	Malattie del sistema digestivo
45678	Maschio	Centro Italia	****	2023	Napoli	Marco	*****	affaticamento	Malattie metaboliche

Data Sanitization

- ▶ La **Data Sanitization** è l'insieme delle tecniche che permettono di evitare il **disclosure** di informazioni confidenziali.
- ▶ L'obiettivo è quello di proteggere la **privacy** e garantire che i dati sensibili siano adeguatamente **trattati**.
- ▶ Alcune applicazioni sono: Data Masking, Data Encryption, Data Anonymization, Data Randomization, ...
- ▶ **Obiettivi progetto:**
 - ▶ Identificazione automatica di informazioni sensibili attraverso l'uso di tecniche di ML, NLP, o DL
 - 1. Addestramento di modelli predittivi su casi di studio reali confrontando le performance su dataset anonimizzati e non anonimizzati:
 - ▶ Sentenze Giudiziare
 - ▶ Cartelle Cliniche
 - ▶ Paper Scientifici
 - ▶ ...
 - 2. Applicazione di tecniche di ML per valutare la possibilità di estrazione o ricostruzione delle informazioni sensibili



Fondamenti di Data Science e Machine Learning

ML and AI

Generative Multimedia Contents



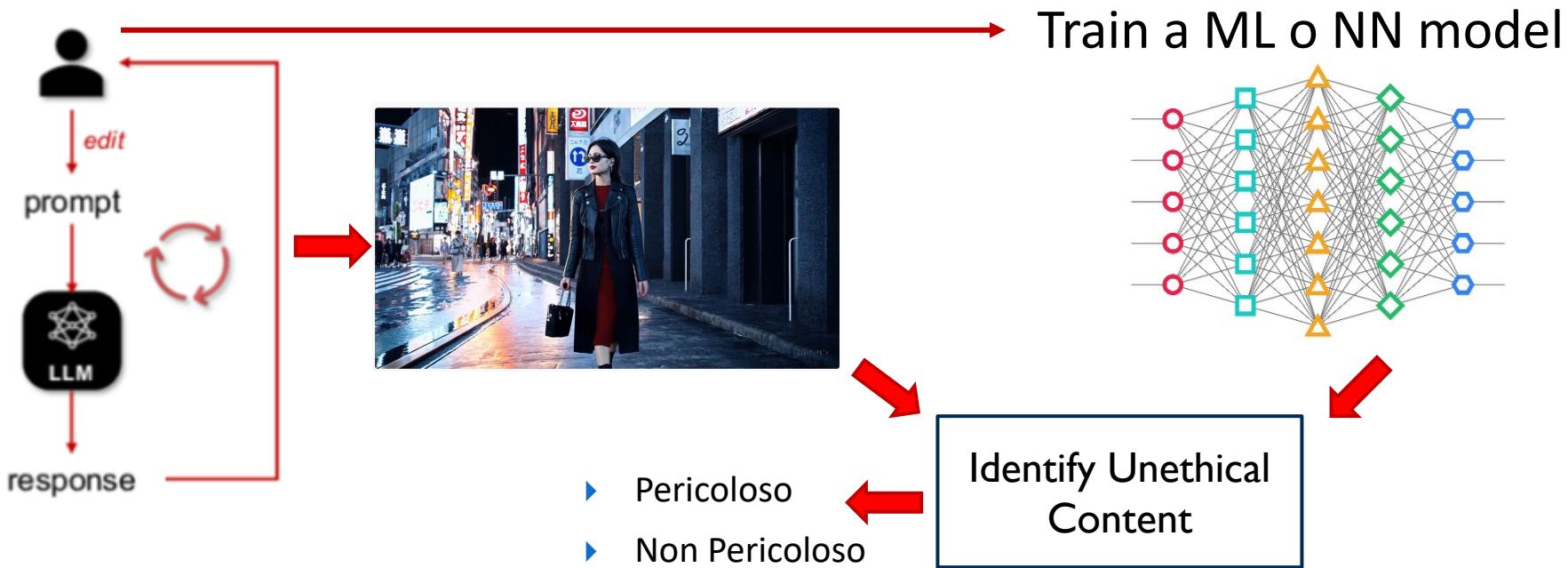
- Grazie all'AI generativa è possibile generare video solo utilizzando prompt testuali
- SORA AI rappresenta uno dei più recenti strumenti di AI generativa di contenuto multimediali

Generative (Unethical) Multimedia Contents



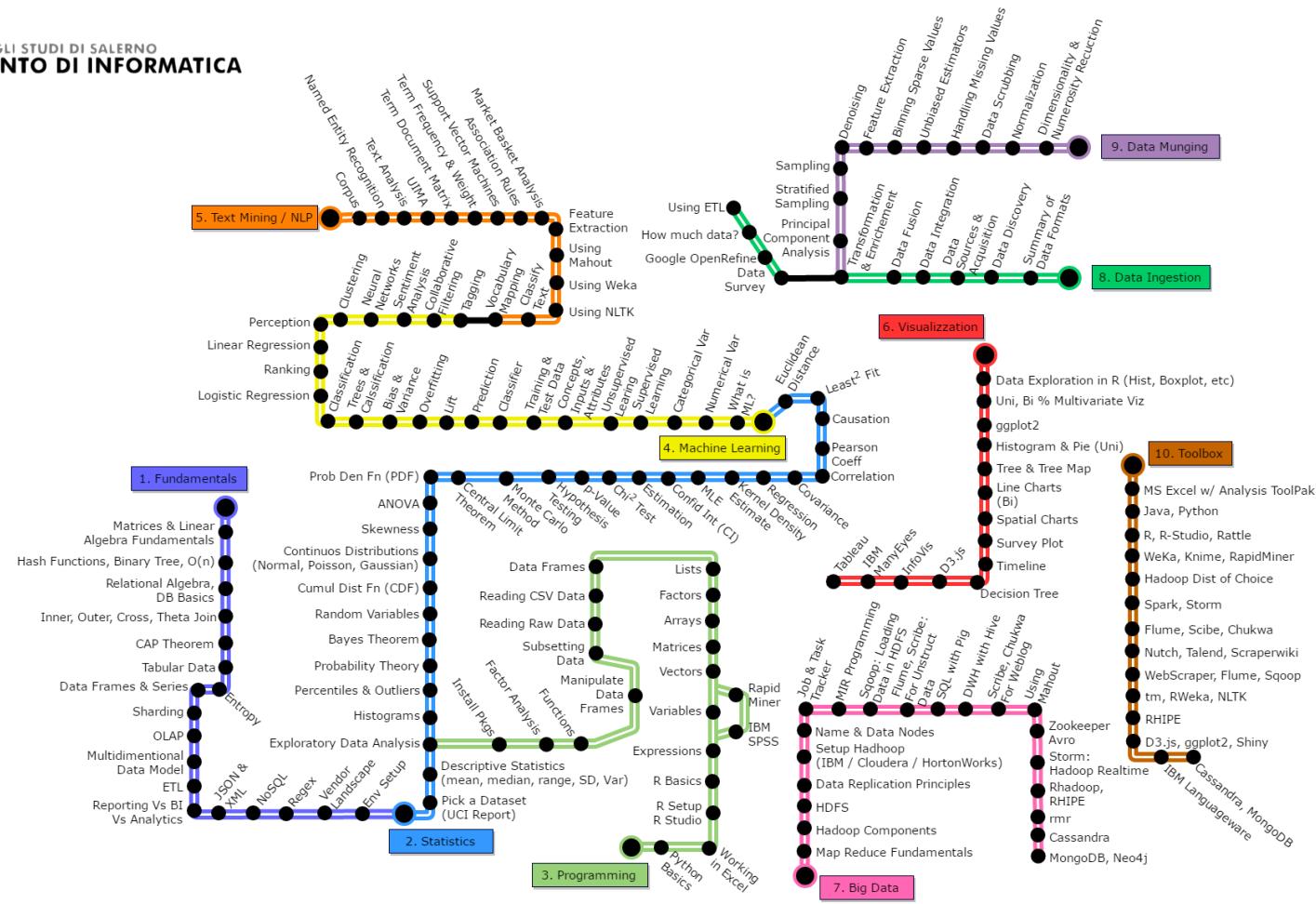
- Attualmente questa generazione di video è limitata: Si possono generare video di pochi secondi
- E' possibile nascondere messaggi pericolosi o subliminali che permettano di trasmettere comportamenti non etici?

Generative (Unethical) Multimedia Contents



▶ Obiettivi:

- ▶ Generare video includendo messaggi non-etici
- ▶ Generare immagini includendo messaggi nascosti o criptati
- ▶ Addestrare modelli di ML o NN per identificare potenziali figure/immagini/comportamenti non etici o dati criptati



Fondamenti di Data Science e Machine Learning

Smart City

Identificazione Disagi Cittadini

- ▶ Cosa sono i disagi cittadini?
 - ▶ Problemi che i cittadini incontrano nella vita quotidiana
 - ▶ Possono essere causati da una serie di fattori:
 - ▶ Infrastrutture fatiscenti
 - ▶ Servizi pubblici scadenti
 - ▶ Mancanza di opportunità
 - ▶ I disagi possono avere un impatto significativo sulla qualità della vita dei cittadini



Identificazione Disagi Cittadini

- ▶ Perché è importante identificare i disagi?
 - ▶ Per migliorare la qualità della vita dei cittadini
 - ▶ Per allocare le risorse in modo più efficiente
 - ▶ Per promuovere la giustizia sociale
 - ▶ Per costruire città più resilienti e sostenibili
- ▶ Altri Esempi di disagi cittadini:
 - ▶ Infrastrutture fatiscenti
 - ▶ Strade dissestate
 - ▶ Marciapiedi rotti
 - ▶ Illuminazione pubblica non funzionante
 - ▶ Caditoie Ostruite
 - ▶ ...
 - ▶ Servizi pubblici scadenti:
 - ▶ Raccolta dei rifiuti inefficiente
 - ▶ Abbandono dei rifiuti
 - ▶ ...



Identificazione Disagi Cittadini

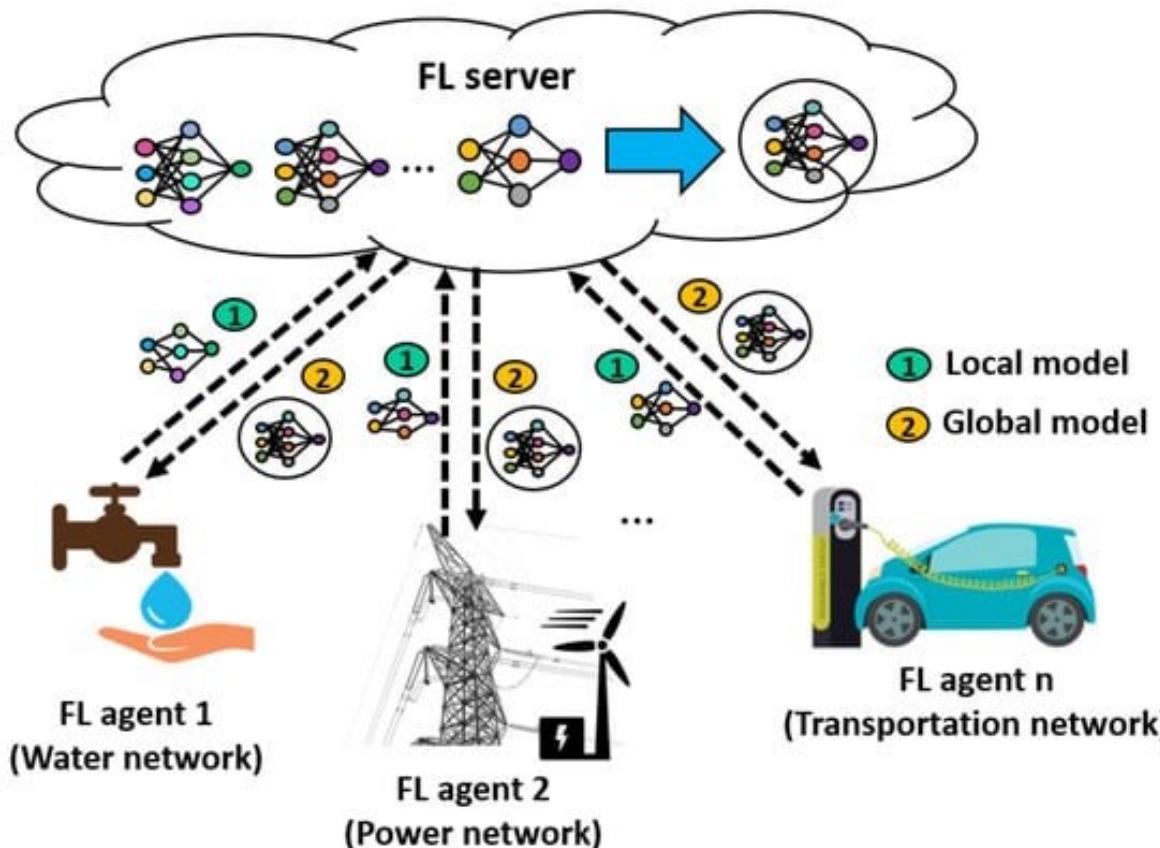
- ▶ Possiamo prevenire o aiutare a rendere più efficienti i sistemi di rilevamento di questi disagi utilizzando i mezzi pubblici?
- ▶ **Obiettivi:**
 - Addestrare almeno un modello di ML o reti neurale in grado di processare immagini provenienti da una telecamera e\o immagini e di identificare almeno uno dei disagi ambientali (Es. Buche sull'asfalto, abbandono dei rifiuti)

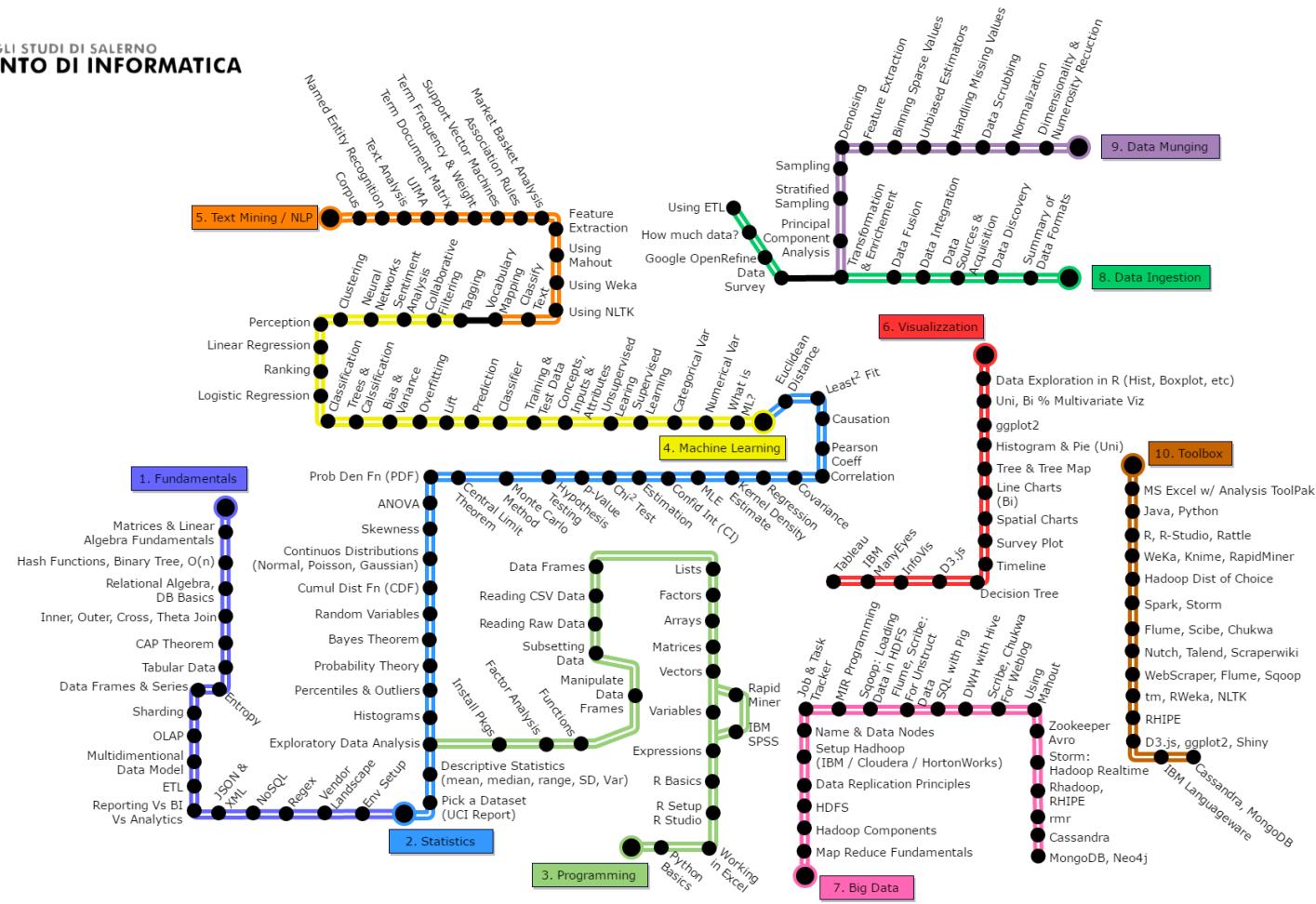


Identificazione Disagi Cittadini

► Obiettivi:

- Addestrare almeno un modello di ML o reti neurale con un approccio di apprendimento Federato per permettere l'addestramento cooperativo di modelli di ML o reti neurali per uno dei disagi cittadini





Fondamenti di Data Science e Machine Learning

Data Management e Data Profiling

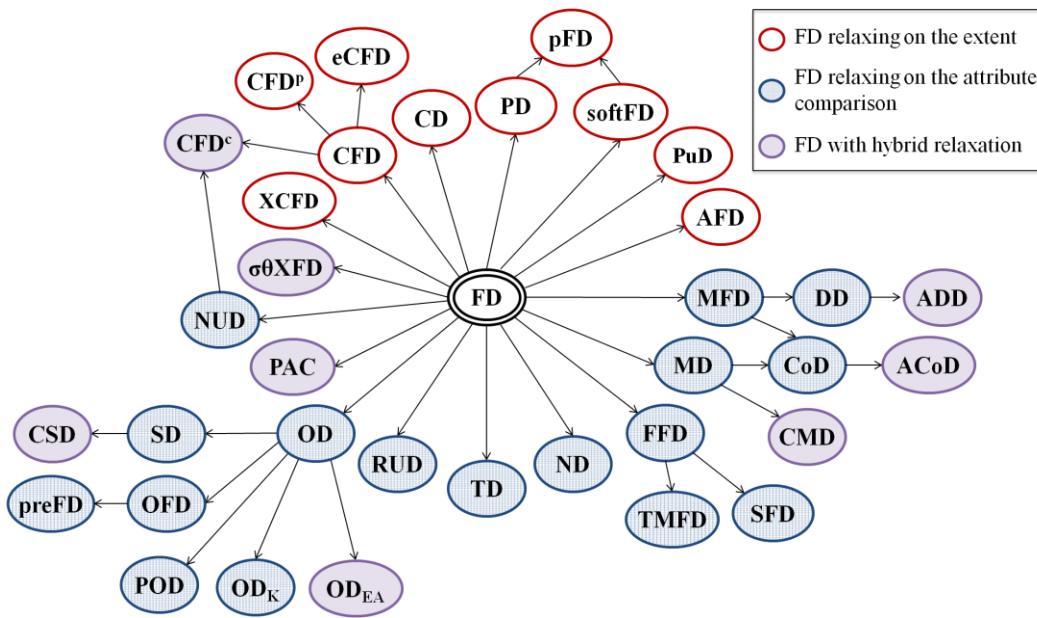
Data Profiling

- ▶ **Data Profiling** è la disciplina che si occupa di estrarre metadati utili da insiemi di dati
- ▶ Domini di applicazione del Data Profiling:
 - ▶ Machine Learning
 - ▶ Data Quality
 - ▶ Query Rewriting
 - ▶ ...
- ▶ Oltre 35 tipi di metadati differenti:
 - ▶ Dipendenze Funzionali (FD)
 - ▶ Dipendenze Funzionali Rilassate (RFD)
 - ▶ Inclusion Dependency (IND)
 - ▶ Unique Column Combination (UCC)
 - ▶ ...

Dipendenze Funzionali Rilassate

- In letteratura sono state introdotte diverse definizioni estese di FD:

Dipendenze Funzionali Rilassate (RFD)



RFD abbrev.	RFD name
ACOD	Approximate comparable dependency
ADD	Approximate differential dependency
AFD	Approximate functional dependency
COD	Comparable dependency
CFD	Conditional functional dependency
CFD ^p	CFD with built-in predicates
CFD ^c	CFD with cardinality constraints and synonym rules
CMD	Conditional matching dependency
CSD	Conditional sequential dependency
CD	Constrained functional dependency
DD	Differential dependency
eCFD	Extended conditional functional dependency
FFD	Fuzzy functional dependency
MD	Matching dependency
MFD	Metric functional dependency
ND	Neighborhood dependency
NUD	Numerical dependency
OD	Order dependency
OD _k	OD satisfied within bound k
ODEA	OD satisfied almost everywhere
OFD	Ordered functional dependency
PD	Partial determination
POD	Polarized order dependencies
preFD	Preference functional dependency
PAC	Probabilistic approximate constraint
pFD	Probabilistic functional dependency
PUD	Purity dependency
RUD	Roll-up dependency
SD	Sequential dependency
SFD	Similarity functional dependency
soft FD	Soft functional dependency
TD	Trend dependency
TMFD	Type-M functional dependency
XCFD	XML conditional functional dependency
σθXFD	XML FD with σ and θ approximation

Metadata Discovery

▶ Progetti:

1. Sviluppare delle strategie efficienti per l'estrazione di metadati da grandi dataset
 - ▶ Strategie di discovery distribuite
 - ▶ Nuove Strutture dati efficienti
 - ▶ Clustering
2. Cosa accade quando i dati evolvono e come possiamo aggiornare i metadata in modo efficiente?
 - ▶ Inserimenti
 - ▶ Cancellazioni
 - ▶ Aggiornamenti
3. Possiamo estrarre e\o definire i metadati su alberi e grafi?
4. Come possiamo utilizzare i metadati per ottimizzare le tecniche di visualizzazione dei dati?

DOMINO

- ▶ Un algoritmo di discovery per RFDs senza parametri in input

$$X_{\Phi_1} \rightarrow Y_{\Phi_2}$$

- ▶ Non è necessario passare in input thresholds da parte dell'utente
- ▶ DOMINO è basato sulla teoria dell'utilità multi-attributo, in particolare sfrutta il concetto di dominanza

Un problema decisionale multi-attributo mira a valutare un numero finito di alternative (scelte) sulla base di un numero finito di attributi (obiettivi o criteri)

Caruccio, L., Deufemia, V., Polese, G. "Discovering Relaxed Functional Dependencies Based on Multi-Attribute Dominance," *IEEE Transactions on Knowledge and Data Engineering*, 33.9 (2021):3212-3228

Reingegnerizzazione degli algoritmi

- ▶ Obiettivi:
 - ▶ Reingegnerizzazione dell'algoritmo Domino
 - ▶ Particolare attenzione deve essere rivolta all'ottimizzazione del codice, sia da un punto di vista di costo per l'esecuzione che di riusabilità del codice stesso

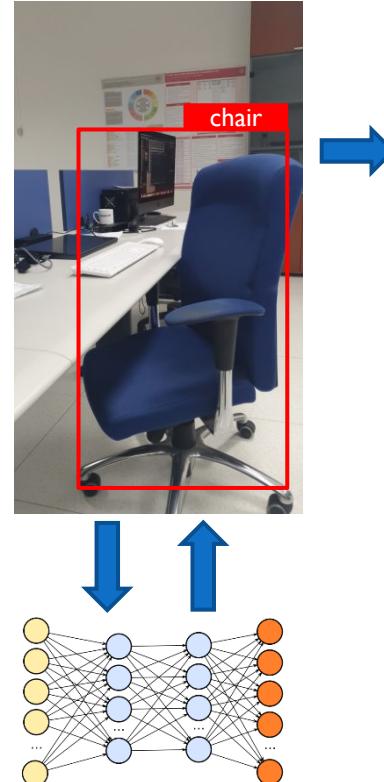
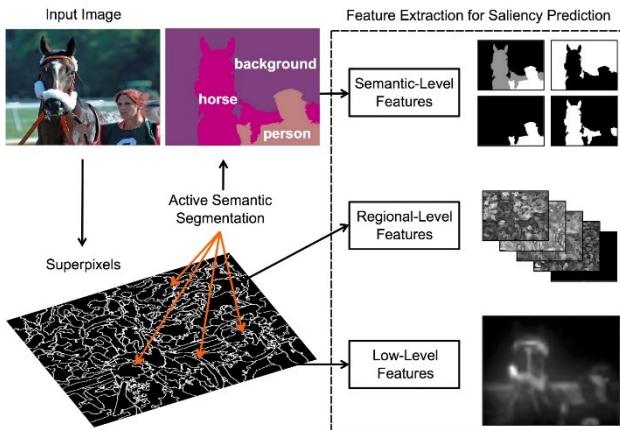
Metadata Application

▶ Open Challenge:

1. Definire tecniche di applicazione dei metadati per supportare la risoluzione di problematiche in diversi contesti d'applicazione
 - ▶ Query Optimization
 - ▶ Feature Selection
 - ▶ Data Cleaning
 - ▶ Feature Engineering
2. In che modo valutare la significatività delle dipendenze sulla base del dominio in cui vengono applicate?
 - ▶ Funzioni di utilità
 - ▶ Ranking basato su euristiche
3. L'applicazione combinata di diversi tipi di metadati può accrescere il supporto fornito negli specifici contesti di applicazione?
4. In che modo tecniche di summarization e/o di visualizzazione aiutano nella comprensione della significatività dei metadati?

Profiling di Immagini

- ▶ Le fotografie e le immagini rappresentano uno degli elementi più condivisi
- ▶ Cosa contengono?
 - ▶ Informazioni sensibili
 - ▶ Informazioni personali
 - ▶ ...
- ▶ Gli algoritmi di discovery nel data profiling esistenti non profilano immagini
- ▶ **Obiettivo:**
 - ▶ Metodologia per l'estrazione di metadati ed informazioni sensibili da immagini e fotografie a supporto dei metadati del data profiling



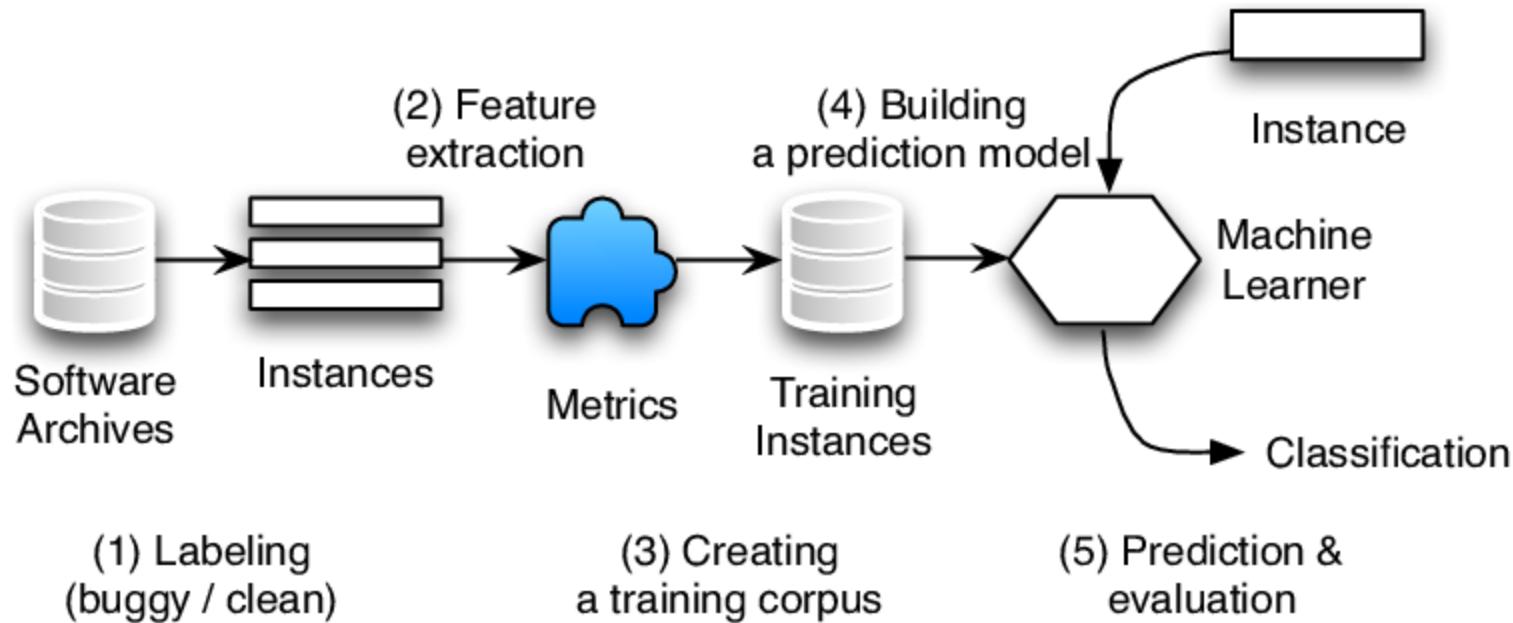
JFIF	
JFIF Version	1.01
Resolution	72 pixels/inch

File — basic information derived from the file.

File Type	JPEG
File Type Extension	jpg
MIME Type	image/jpeg
Encoding Process	Progressive DCT, Huffman coding
Bits Per Sample	8
Color Components	3
File Size	90 kB
Image Size	720 × 1,280
Y Cb Cr Sub Sampling	YCbCr4:2:0 (2 2)
Profile CMM Type	
Profile Version	2.1.0
Profile Class	Display Device Profile
Color Space Data	RGB
Profile Connection Space	XYZ
Profile Date Time	0000:00:00 00:00:00
Profile File Signature	acsp
Primary Platform	Unknown ()
CMM Flags	Not Embedded, Independent
Device Manufacturer	
Device Model	
Device Attributes	Reflective, Glossy, Positive, Color
Rendering Intent	Media-Relative Colorimetric
Connection Space Illuminant	0.9642 1 0.82491
Profile Creator	
Profile ID	0
Profile Description	sRGB
Red Matrix Column	0.43607 0.22249 0.01392
Green Matrix Column	0.38515 0.71687 0.09708
Blue Matrix Column	0.14307 0.06061 0.7141
Red Tone Reproduction Curve	(40 bytes binary data)
Green Tone Reproduction Curve	(40 bytes binary data)
Blue Tone Reproduction Curve	(40 bytes binary data)
Media White Point	0.9642 1 0.82491
Profile Copyright	Google Inc. 2016

Bug Prediction with Profiling Metadata

- ▶ **Problema:** I bug nel software possono causare malfunzionamenti, crash e persino violazioni della sicurezza
- ▶ **Soluzione:** Defect prediction, o Bug Prediction, aiuta gli sviluppatori a identificare le parti del codice soggette a bug in modo da poterli prevenire o correggerli in anticipo



Bug Prediction with Profiling Metadata

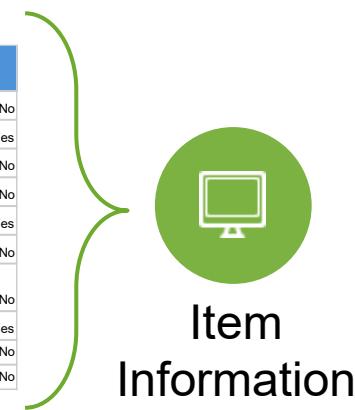
- ▶ L'approccio generico prevede l'utilizzo modelli di ML per analizzare le caratteristiche di repository e\o le azioni effettuate sul codice (es. commit) per identificare i casi in cui può esserci un bug
- ▶ I modelli vengono addestrati su set di dati con bug, imparando a riconoscere le caratteristiche del codice che li rendono più probabili
- ▶ Una volta addestrati, i modelli possono essere utilizzati per analizzare nuovo codice e prevedere la probabilità che contenga bug

Nome Repository	ID Repository	URL Repository	Linguaggio di Programmazione	Hash Commit	Data Commit	Autore Commit	Messaggio Commit	Aggiunte	Rimozioni	Modifiche	Bug
my-awesome-project	123456	https://github.com/sindresorhus/awesome	Python	abc123	19/04/2024	octocat	Aggiunta nuova funzionalità	10	2	15	No
my-awesome-project	123456	https://github.com/sindresorhus/awesome	Python	def456	18/04/2024	octocat	Risolto bug #123	5	1	8	Yes
my-awesome-project	123456	https://github.com/sindresorhus/awesome	Python	ghi789	17/04/2024	octocat	Migliorato codice di prestazioni	2	0	12	No
buggy-app	654321	https://github.com/vcrash/buggyapp	JavaScript	jkl012	19/04/2024	johndoe	Aggiunta nuova pagina	8	0	11	No
buggy-app	654321	https://github.com/vcrash/buggyapp	JavaScript	mno345	18/04/2024	janedoe	Tentativo di risolvere bug #432	4	17	6	Yes
buggy-app	654321	https://github.com/vcrash/buggyapp	JavaScript	pqr678	17/04/2024	johndoe	Aggiornamenti documentazione	3	0	5	No
simple-calculator	987654	https://github.com/topics/simple-calculator	C++	stu901	19/04/2024	gituser123	Aggiunta funzione di calcolo percentuale	15	0	22	No
simple-calculator	987654	https://github.com/topics/simple-calculator	C++	wx234	18/04/2024	gituser456	Risolto bug di divisione per zero	11	3	5	Yes
simple-calculator	987654	https://github.com/topics/simple-calculator	C++	yza567	17/04/2024	gituser789	Migliorato l'interfaccia utente	8	2	13	No
broken-library	234567	https://downdetector.com/status/github/	Java	abc123	19/04/2024	librarysupport	Aggiornamento dipendenze	6	2	4	No

Bug Prediction with Profiling Metadata

- I modelli esistenti non arrivano ad ottenere elevati risultati nella predizione dei bug
 - Se provassimo a Profilare lo sviluppatore?

Nome Repository	ID Repository	URL Repository	Linguaggio di Programmazione	Hash Commit	Data Commit	Autore Commit	Messaggio Commit	Aggiunte	Rimozioni	Modifiche	Bug
my-awesome-project	123456	https://github.com/sindresorhus/awesome	Python	abc123	19/04/2023	octocat	Aggiunta nuova funzionalità	10	2	15	No
my-awesome-project	123456	https://github.com/sindresorhus/awesome	Python	def456	18/04/2023	octocat	Risolto bug #123	5	1	8	Yes
my-awesome-project	123456	https://github.com/sindresorhus/awesome	Python	ghi789	17/04/2023	octocat	Migliorato codice di prestazioni	2	0	12	No
buggy-app	654321	https://github.com/vcrash/buggapp	JavaScript	jk1012	19/04/2023	johndoe	Aggiunta nuova pagina	8	0	11	No
buggy-app	654321	https://github.com/vcrash/buggapp	JavaScript	mno345	18/04/2023	janedoe	Tentativo di risolvere bug #432	4	17	6	Yes
buggy-app	654321	https://github.com/vcrash/buggapp	JavaScript	pqr678	17/04/2023	johndoe	Aggiornamenti documentazione	3	0	5	No
simple-calculator	987654	https://github.com/topics/simple-calculator	C++	stu901	19/04/2023	gituser123	Aggiunta funzione di calcolo percentuale	15	0	22	No
simple-calculator	987654	https://github.com/topics/simple-calculator	C++	wxx234	18/04/2023	gituser456	Risolto bug di divisione per zero	11	3	5	Yes
simple-calculator	987654	https://github.com/topics/simple-calculator	C++	yza567	17/04/2023	gituser789	Migliorato l'interfaccia utente	8	2	13	No
broken-library	234567	https://downloader.com/status/github/	Java	abc123	19/04/2023	librarysupport	Aggiornamento dipendenze	6	2	4	No



ID	Sviluppatore	Linguaggi di Programmazione	Anni di Esperienza	Livello di Esperienza	Sistema Operativo	Numero di Commit	Numero di Issue Aperte	Righe di Codice
1	octocat	Python, Java	5	Intermedio	Linux	120	7	10000
2	johndoe	C++, JavaScript	8	Esperto	Windows	180	3	5000
3	gituser123	Go, R	2	Principiante	macOS	50	15	5000
4	gituser456	Python, C#	4	Intermedio	Linux	100	6	8000
5	gituser789	Java, PHP	7	Esperto	Windows	150	4	12000
6	librarysupport	JavaScript, TypeScript	3	Principiante	macOS	70	12	6000



Bug Prediction with Profiling Metadata

► Obiettivi del progetto:

1. Addestrare almeno due modelli per Bug Prediction arricchendo le informazioni dei dataset con le informazioni degli sviluppatori
2. Estendere le informazioni creando delle nuove feature o rimuovendo quelle meno significative utilizzando i metadati di profilazione

User_F1	User_F2	Code_F1	Code_F2	Code_F3	RFD_F1	RFD_F2	Code_F1

+

-

Adding New Features

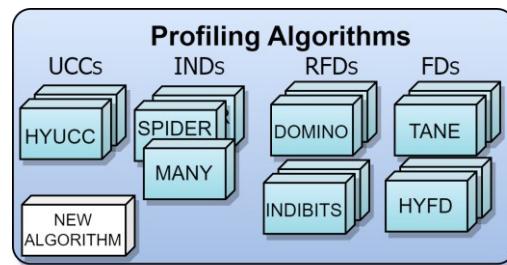
Removing Features

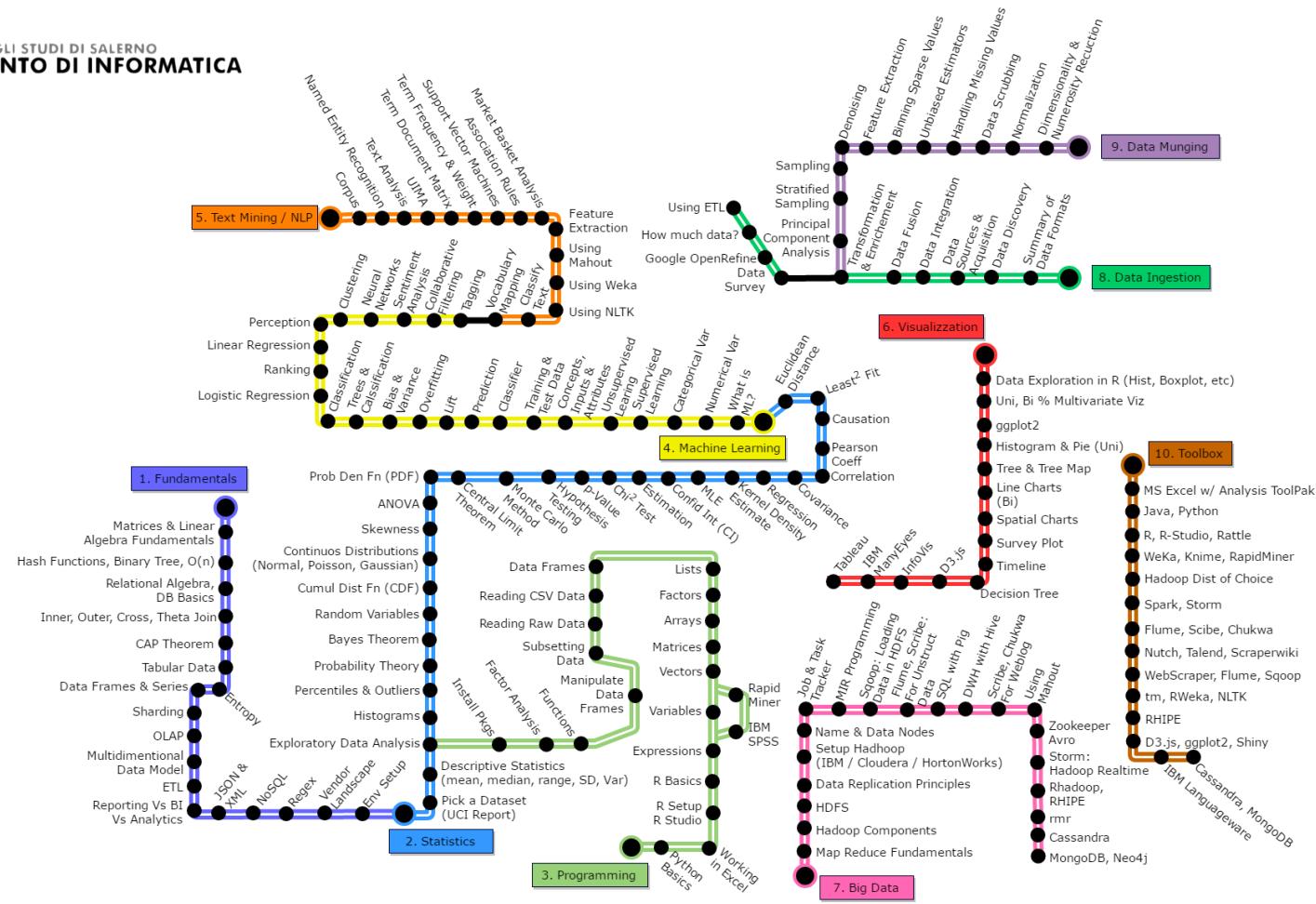


Intelligent Models



Identify Potential Bugs





Fondamenti di Data Science e Machine Learning

Speech-to-Text

Threads

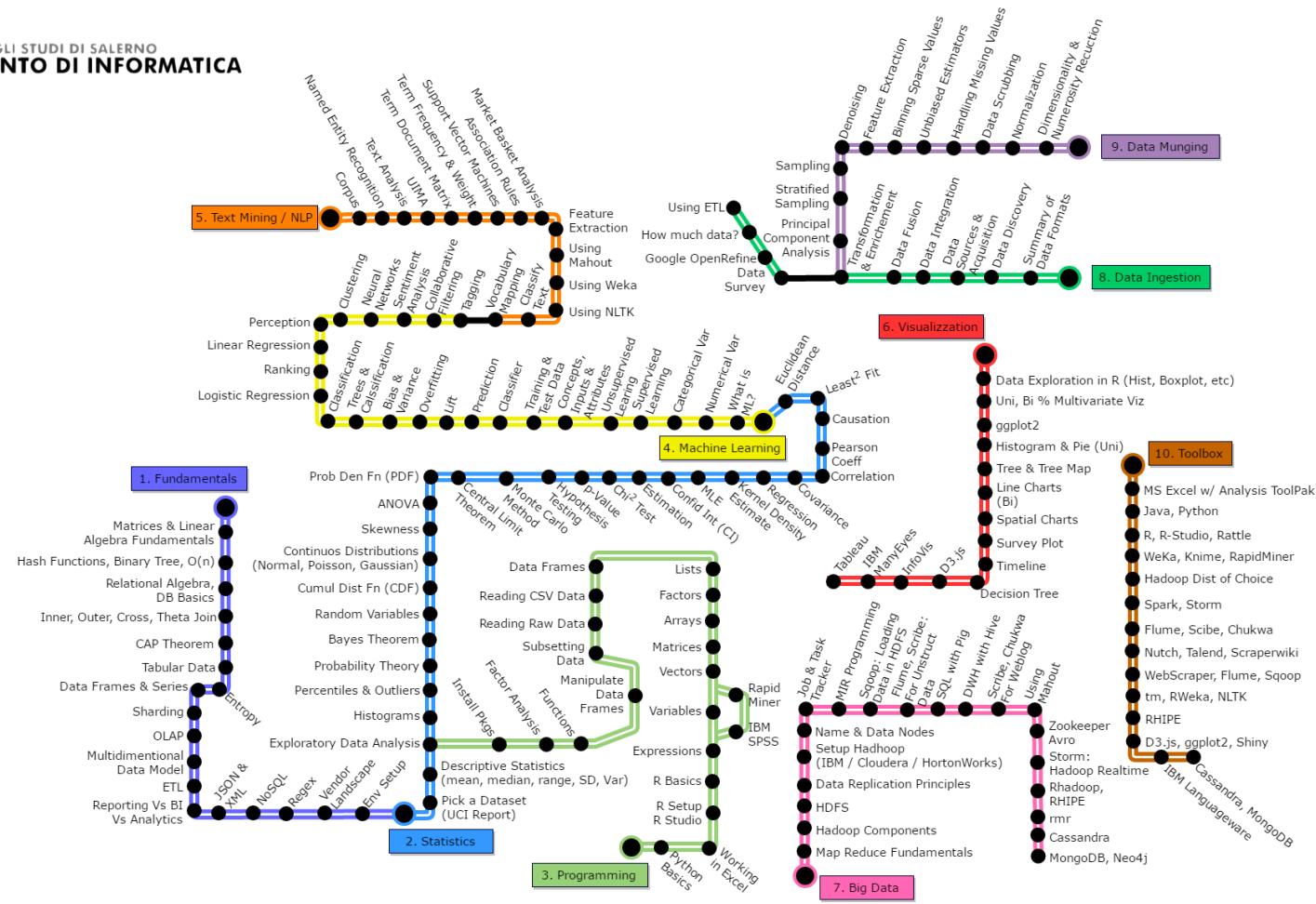
- ▶ Threads è uno dei più recenti social network
 - ▶ E' stato rilasciato da Meta nel 2023
 - ▶ Threads, il social di Meta considerato l'anti Twitter (X) ha raggiunto **130 milioni** di utenti attivi mensili
 - ▶ A differenza di molti altri social network permette di condividere anche registrazioni vocali
 - ▶ Nelle registrazioni si possono condividere messaggi di vario genere e pericolosità (es. insulti, fenomeni di razzismo, etc.)



Threads

- ▶ **Obiettivi progetto:**
 - ▶ Addestrare dei modelli per permettere di identificare fenomeni pericolosi nelle registrazioni condivise su Threads o in video TikTok, ecc
 - ▶ Razzismo
 - ▶ Hate Speech
 - ▶ Messaggi di Violenza
 - ▶ Cyberbullismo
 - ▶ ...
 - ▶ Identificare utenti pericolosi che possono diffondere messaggi errati



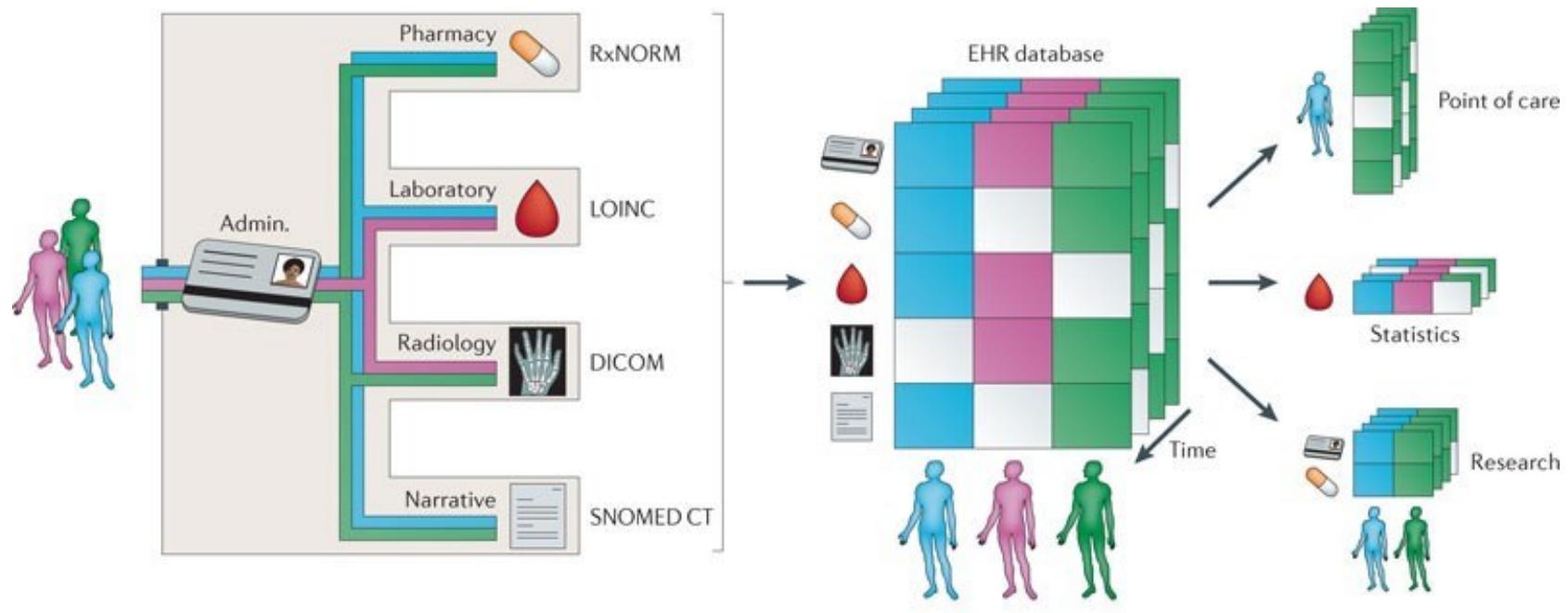


Fondamenti di Data Science e Machine Learning

ML In Health Applications

Campionamento e Pre-elaborazione di Dati Clinici

I dati clinici, pur essendo ricchi di informazioni, sono spesso caratterizzati da bassa qualità, elevata eterogeneità e mancanza di struttura – elementi che rappresentano sfide significative per le analisi successive.



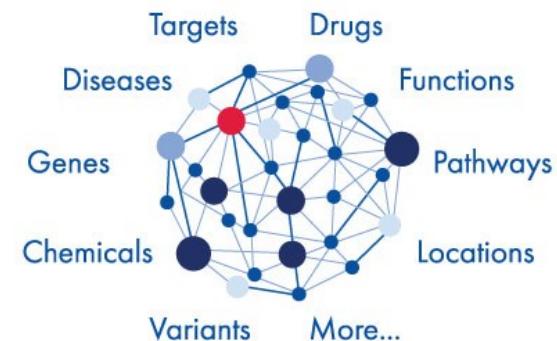
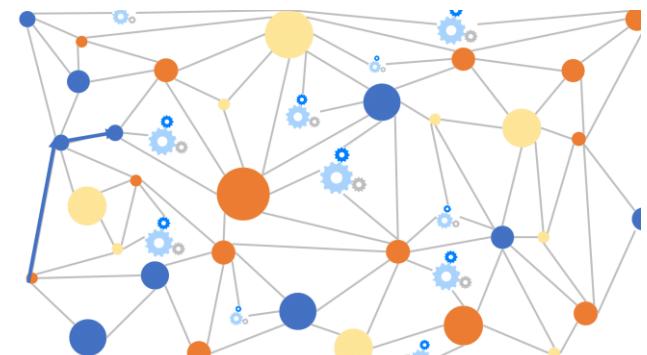
<https://www.shaip.com/blog/healthcare-datasets-for-machine-learning-projects/>

Campionamento e Pre-elaborazione di Dati Clinici

- ▶ **Definizione e progettazione di strategie di campionamento domain-oriented**
 - ▶ Selezione di sottoinsiemi di dati rappresentativi e rilevanti
 - ▶ Processi di Record Linkage
 - ▶ Analisi della presenza di bias
- ▶ **Progettazione di strategie pre-preprocessing domain-oriented**
 - ▶ Analisi e miglioramento della qualità dei dati: Gestione dei valori mancanti, formati incoerenti, record duplicati e input rumorosi
 - ▶ Normalizzazione e Standardizzazione sia di dati strutturati che non strutturati: Note cliniche, codici diagnostici e informazioni temporali
 - ▶ Utilizzo di tecniche avanzate di riconoscimento di entità e disambiguazione semantica: Allineare i dati grezzi con le ontologie mediche e trasformarli in feature leggibili da una macchina.
- ▶ **GOAL:** Definizione di dataset adatti all'inferenza clinica e alla modellazione predittiva

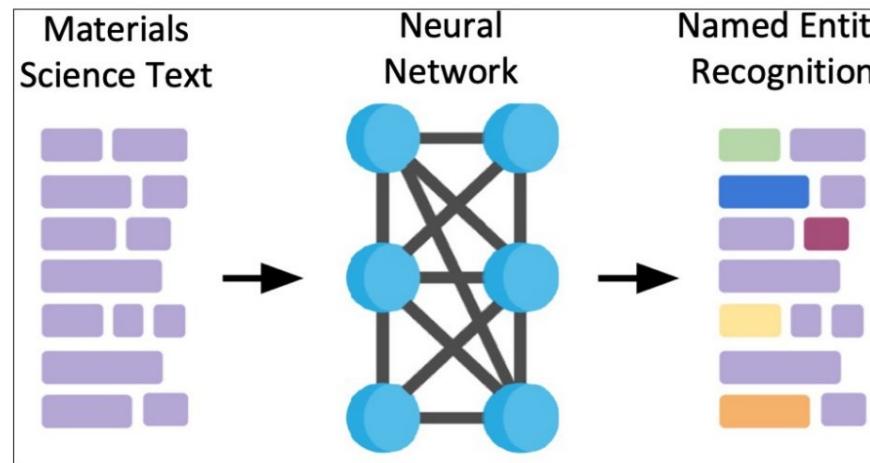
Knowledge graph

- ▶ Le informazioni estratte attraverso l'analisi semantica devono essere organizzate e rappresentate in modo tale da permettere ulteriori analisi
 - ▶ Knowledge graph
 - ▶ Ontologie
- ▶ I **Knowledge graph** sono delle strutture dati che permettono di rappresentare entità, concetti e relazioni in maniera compatta e facilmente navigabile
- ▶ In ambito sanitario, queste strutture sono usate in aree di ricerca come la **network medicine** e la **network analysis**
- ▶ Estrarre informazioni da documenti biomedici e rappresentarle utilizzando questi grafi è un task che presenta diverse criticità:
 - ▶ Qualità dei dati estratti
 - ▶ Gestione di grafi di grandi dimensioni
 - ▶ Definizione di proprietà su concetti e relazioni
 - ▶ Gestire e risolvere le query sul grafo
 - ▶



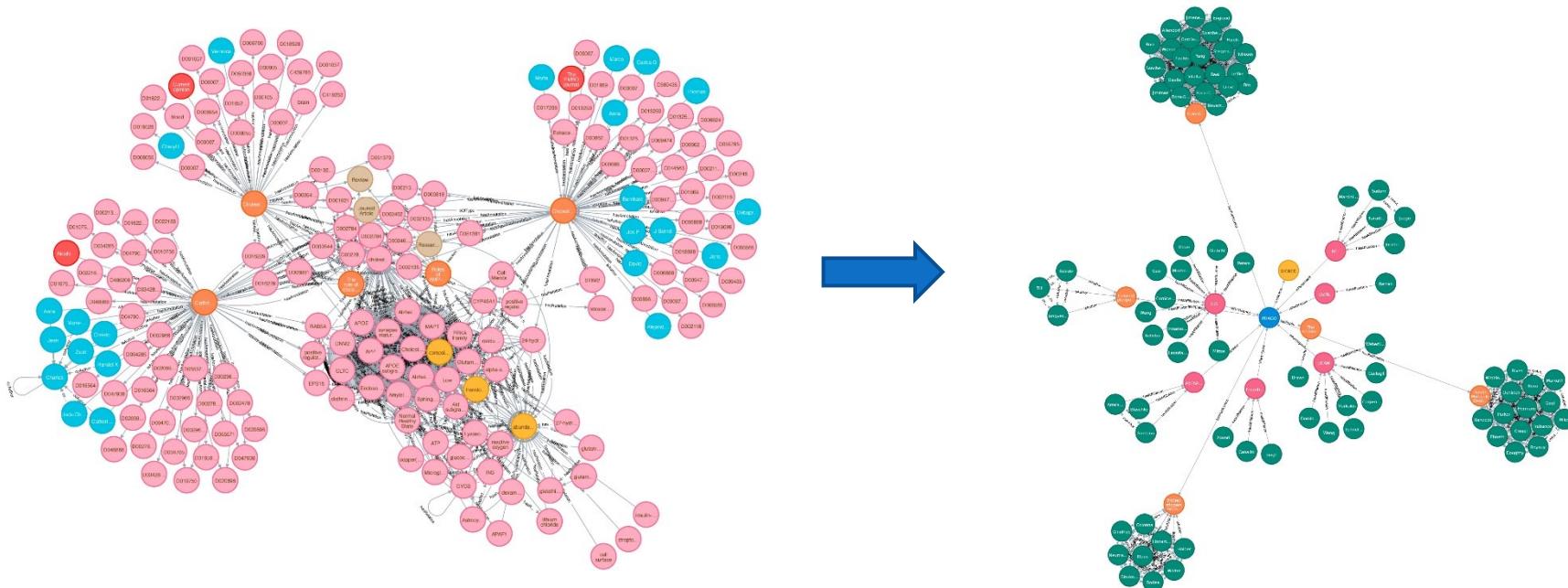
Information extraction e rappresentazione

- ▶ Abbiamo dei documenti e vogliamo estrarre keyword e altri concetti rilevanti da essi
- ▶ Successivamente, vogliamo definire relazioni tra di essi e rappresentare tutte queste informazioni sfruttando i Knowledge graph
- ▶ Attualmente, esistono diversi modelli che permettono di risolvere questi task
- ▶ **GOAL:** Analizzare i KG ottenuti dai vari modelli utilizzati e fornire una valutazione comparativa, evidenziando le differenze tra i vari modelli



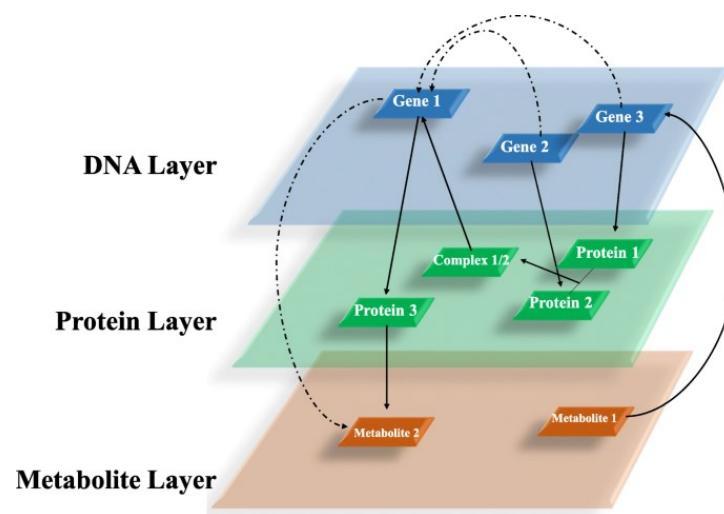
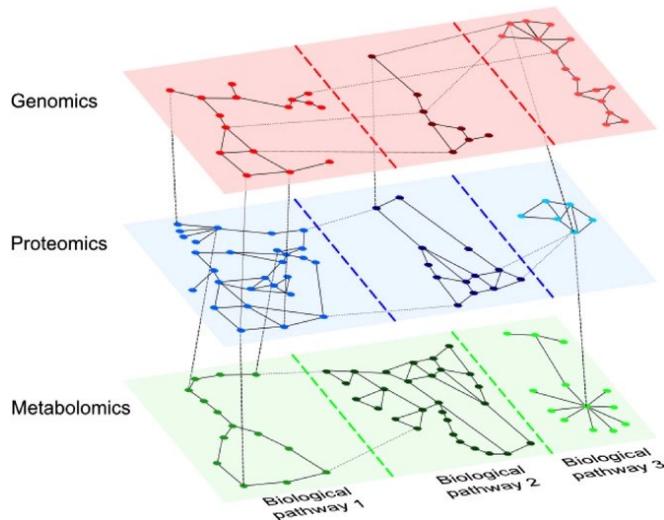
Ottimizzazione di Knowledge graph

- ▶ Con grandi quantità di dati, le dimensioni del grafo possono esplodere
- ▶ È necessario andare a ridurre queste dimensioni in modo tale da permettere una veloce interrogazione del grafo
- ▶ **GOAL:** Definizione di strategie per ottimizzare la gestione di grafi di grandi dimensioni



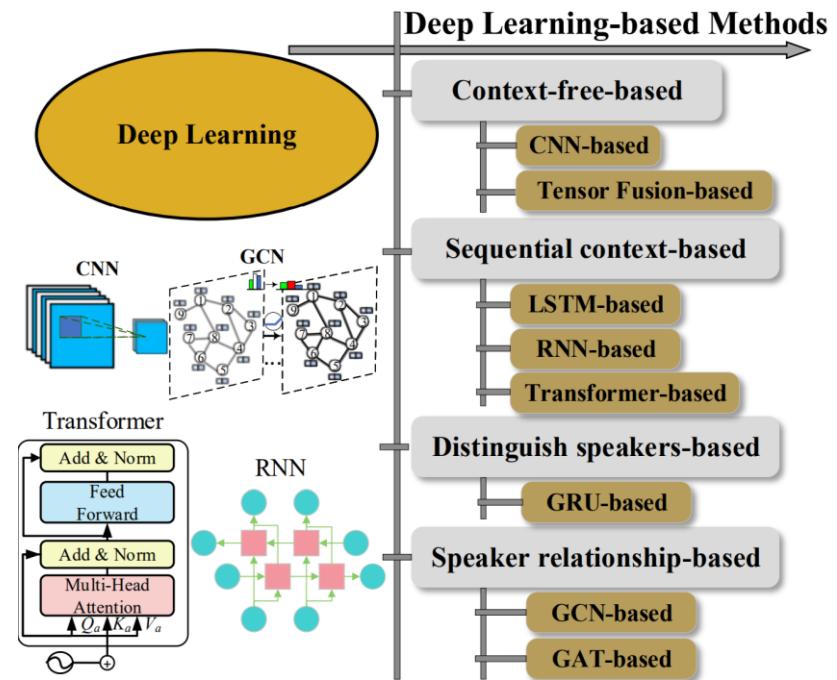
Knowledge graph multilivello

- ▶ In alcuni contesti, può essere utile avere dei grafi multilivello che ci permettano di analizzare le informazioni in base a diversi attributi
- ▶ Nella biomedicina, ad esempio, due farmaci potrebbero condividere diversi principi attivi ma avere degli effetti collaterali diversi
- ▶ Memorizzare queste informazioni attraverso l'uso di grafi multilivello può portare alla definizione di metodologie per la gestione di query molto complesse
- ▶ **GOAL:** Definizione di metodologie per la costruzione di questi grafi



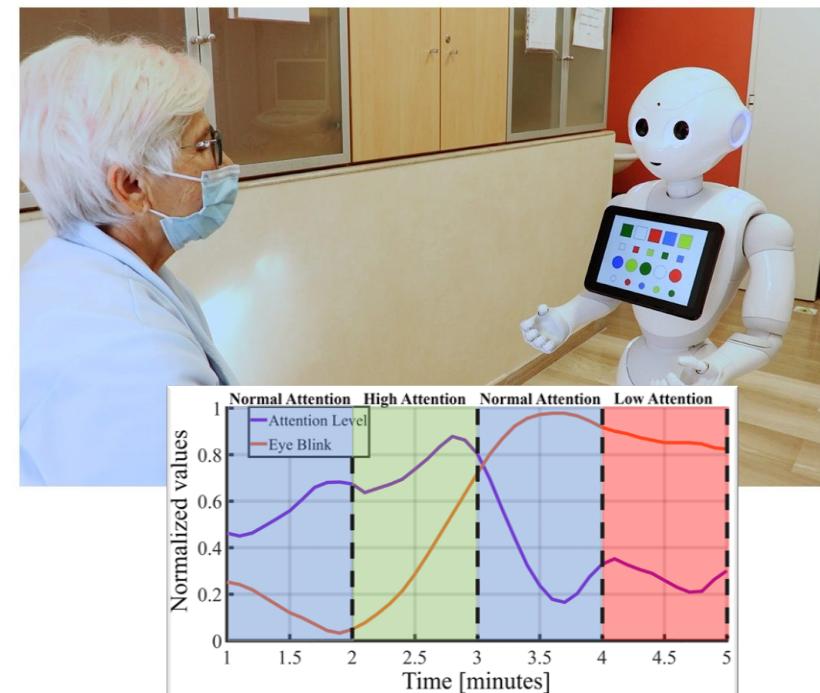
Conversational Emotion Recognition

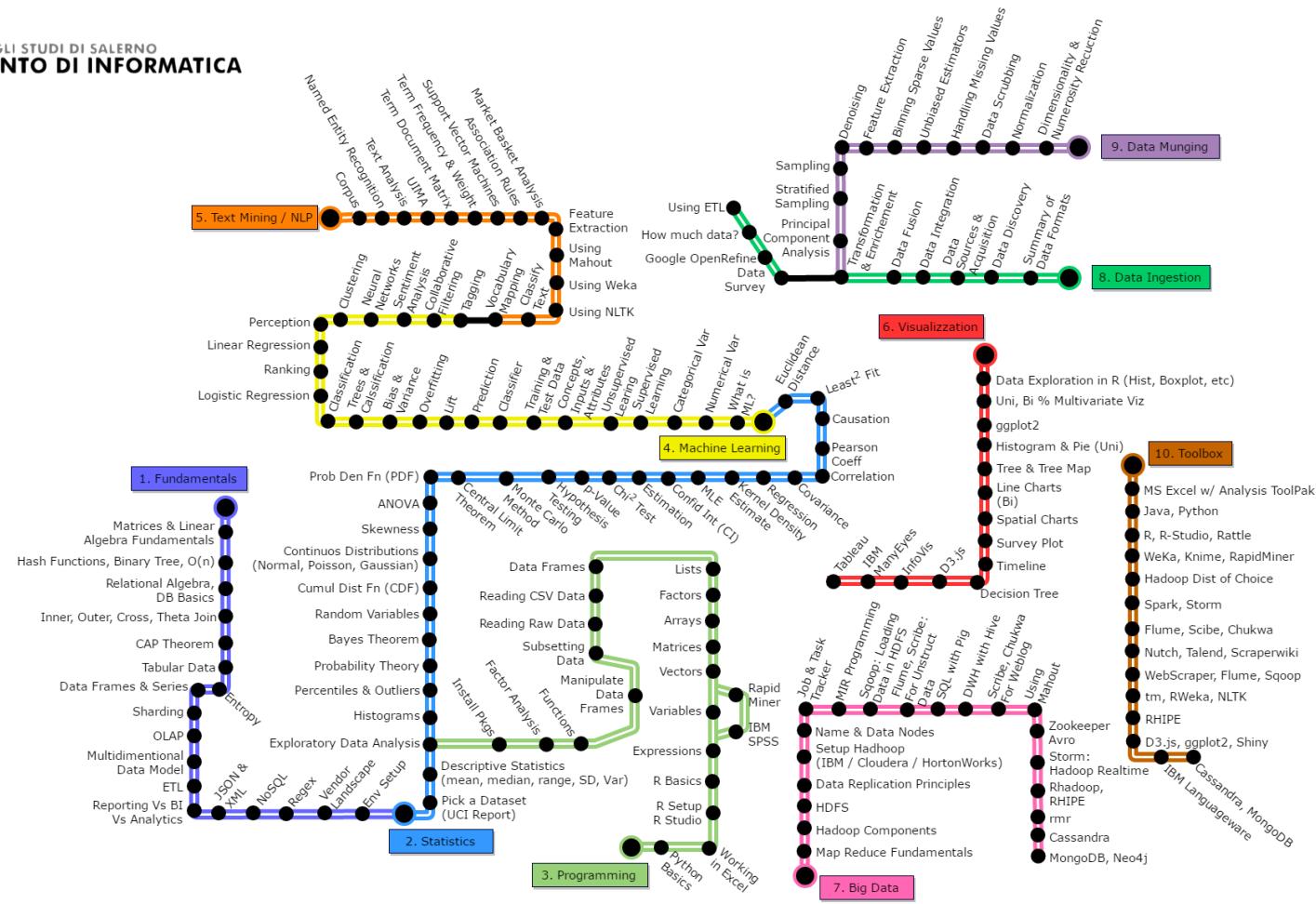
- ▶ Le informazioni estratte da conversazioni con pazienti per riconoscere quanto le indicazioni sul livello di dolore descritte non siano affette dallo status emozionale.
- ▶ Il progetto si propone per perseguire uno o più dei seguenti obiettivi:
 - ▶ Utilizzo di modelli di Deep Learning
 - ▶ Emotional detection da sorgenti video
 - ▶ Emotional detection da sorgenti testuali
 - ▶ Multi-modal Emotional Detection
 - ▶ Studio delle relazioni nei risultati e nell'elaborazione di differenti modelli
 - ▶ Definizione di meta-modelli che permettano la realizzazione di processi di feature engineering combinati
- ▶ Il progetto verrà realizzato in collaborazione con esperti di dominio, quindi medici, che lavorano nel contesto della terapia del dolore.



Analisi del Livello di Attenzione

- ▶ La capacità di concentrarsi, insieme ad altri fattori come la velocità di reazione e la memoria, è strettamente legata alla funzione cognitiva. Un calo in queste aree può indicare l'insorgenza di problemi cognitivi più ampi.
- ▶ **GOAL:** valutare i livelli di attenzione dei pazienti attraverso una serie di test interattivi
- ▶ Il progetto si propone di realizzare uno o più dei seguenti task:
 - ▶ Selezione ed implementazione di test interattivi:
 - ▶ **test sul tempo di reazione** (ad esempio, tempi di risposta a stimoli visivi o uditivi),
 - ▶ **giochi di memoria a breve termine** (come memorizzare sequenze di parole o numeri),
 - ▶ **test di comprensione** (rispondere a domande basate su brevi testi).
 - ▶ Data Representation per la definizione di training dataset
 - ▶ Emotional detection da sorgenti testuali
 - ▶ Robot-assisted Evaluation





Fondamenti di Data Science e Machine Learning

Recommender Systems

Recommender Systems

- ▶ La maggior parte dei siti di e-commerce e delle piattaforme di streaming si serve dei sistemi di raccomandazione per consigliare agli utenti prodotti simili a quelli che stanno visualizzando, ascoltando o acquistando
- ▶ Alcuni esempi di piattaforme che utilizzano tali sistemi sono Amazon, Spotify o Netflix

The screenshot shows the product page for 'Prediction Machines: The Simple Economics of Artificial Intelligence' on Amazon. It includes the book cover, title, author (Avi Agresti, Joshua Gans, Avi Goldfarb), price (\$17.90), and a 'Look Inside' button. Below the main image are smaller thumbnail images of related books like 'Applied Artificial Intelligence' and 'HUMAN + MACHINE'. A red box highlights the 'Customers who bought this item also bought' section at the bottom left, which lists several other AI-related books.

The screenshot shows the Netflix homepage. At the top, a banner says 'Because you watched Virgin River' and displays several movie and TV show thumbnails. Below this, another banner says 'Watch It Again' and shows more content. A red box highlights the 'Popular on Netflix' section, which lists shows like 'Dead to Me', 'Blacklist', and 'Outlander'. At the bottom, there are three sections: 'Scelto per te' (Selected for you) featuring a 'GENERAZIONE' playlist, 'Creato per' (Created for) featuring 'Daily Mix' playlists for 'è tornato holden', 'PSICOLOGI', and 'Lady Gaga, Taylor Swift, Avril Lavigne e...', and 'Mostra tutto' (Show all).

Recommender Systems

- ▶ L'obiettivo principale di questi sistemi è quello di aumentare l'interazione dell'utente con la piattaforma:
 - ▶ nel caso dell'e-commerce, incentivando l'acquisto di nuovi prodotti potenzialmente rilevanti
 - ▶ nel caso dello streaming, proponendo musica o film affini ai gusti dell'utente per mantenerlo coinvolto e prolungare la sua permanenza o abbonamento
- ▶ Tra le varie tecniche impiegate nei sistemi di raccomandazione, alcune piattaforme fanno uso di algoritmi di clustering per segmentare contenuti o utenti e generare suggerimenti più mirati.
 - ▶ Ad esempio Netflix e Amazon

Recommender Systems: Clustering

- ▶ Esistono tre possibili soluzioni per l'implementazione di algoritmi cluster-based
- ▶ User-Based Clustering
 - ▶ Il clustering basato sull'utente raggruppa gli utenti con comportamenti e preferenze simili
 - ▶ Una volta che gli utenti sono raggruppati si possono formulare raccomandazioni considerando le preferenze di altri utenti dello stesso cluster
 - ▶ Un possibile dataset utilizzabile in questo contesto è [Netflix Movie Rating](#)
- ▶ Item-based Clustering
 - ▶ Il clustering basato su elementi si concentra sul raggruppamento di elementi che condividono caratteristiche simili o con cui un utente interagisce frequentemente
 - ▶ Un dataset utilizzabile in questo contesto è [Netflix Movie Rating](#)
- ▶ Hybrid Clustering
 - ▶ Il clustering ibrido combina il clustering basato sull'utente e su elementi per fornire raccomandazioni più accurate
 - ▶ Tre dataset utilizzabili in questo contesto sono disponibili sul sito [HetRec 2011 | GroupLens](#)

Recommender Systems

- ▶ **GOAL:** valutare l'efficacia di un sistema di raccomandazione basato sul clustering con l'obiettivo di individuare le strategie più performanti per migliorare la qualità dei suggerimenti.
- ▶ A supporto di questi metodi, è possibile implementare regole automatiche derivate da tecniche di rule mining (ad esempio, regole di associazione, pattern mining, regole di primo ordine)
- ▶ Obiettivi specifici del progetto:
 - ▶ Applicare e confrontare diverse tecniche di clustering sul dominio di applicazione scelto
 - ▶ Estrarre regole automatiche per migliorare i cluster o le raccomandazioni
 - ▶ Analizzare l'overlap della conoscenza estraibile dalle regole inferite e i cluster definiti
 - ▶ Definire approcci ibridi che combino regole e modelli predittivi
 - ▶ Confrontare le performance del sistema con e senza l'utilizzo delle regole

Focus on: Association Rules

► Se A, allora B

(formalmente: $A \Rightarrow B$)

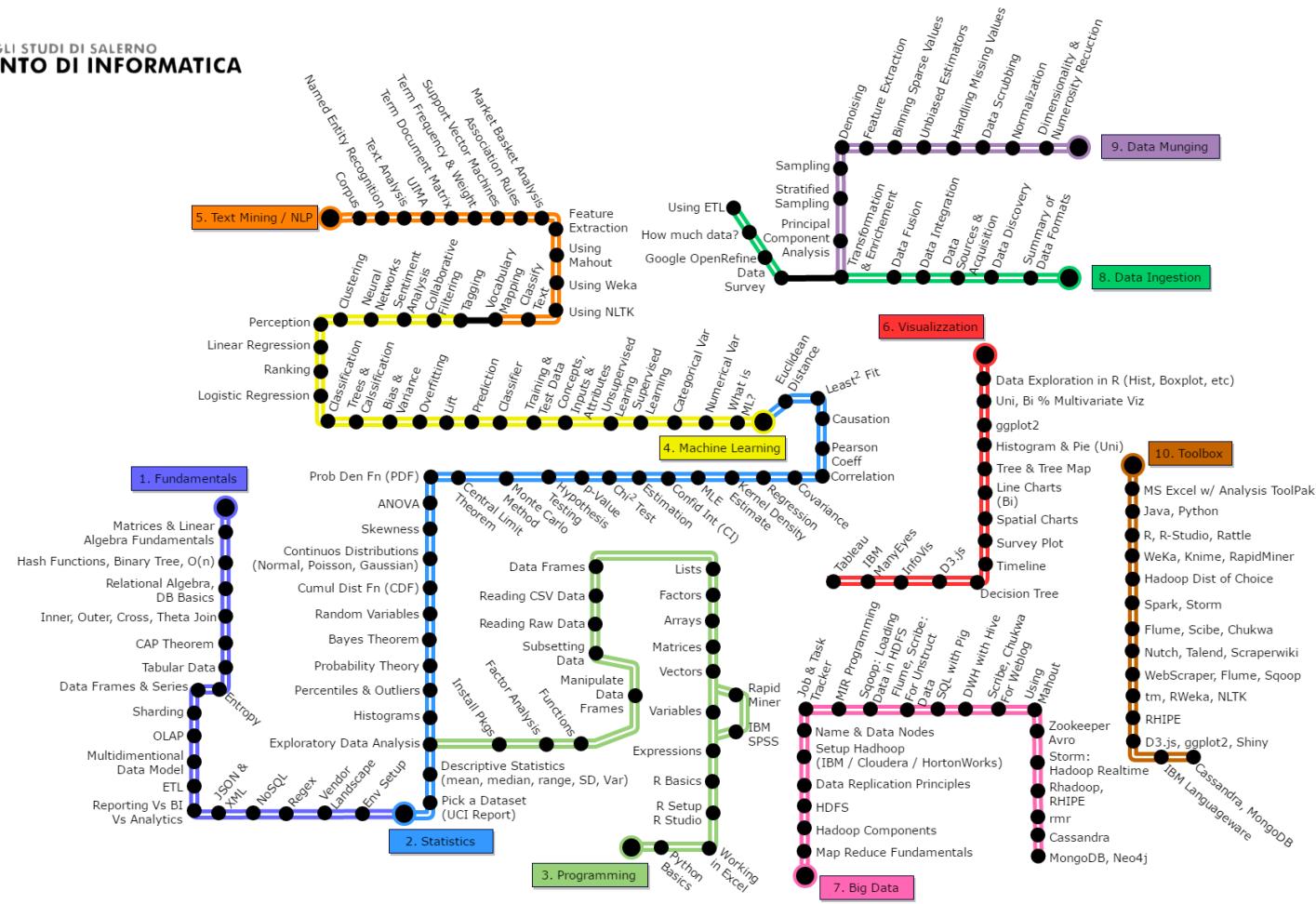
- ▶ A e B sono insiemi di elementi (itemset),
- ▶ La presenza di A implica una certa probabilità della presenza di B.

► Esempio

- ▶ Se un utente ha apprezzato "The Matrix" e "Inception", allora apprezza anche "Interstellar"

Formalmente:

- ▶ $\{\text{The Matrix}, \text{ Inception}\} \Rightarrow \{\text{Interstellar}\}$
- ▶ Ogni regola può essere valutata con metriche standard:
 - ▶ Supporto: percentuale di utenti che hanno visto e apprezzato tutti i film nella regola.
 - ▶ Confidenza: probabilità che un utente che ha visto "The Matrix" e "Inception" abbia visto anche "Interstellar".
 - ▶ Lift: misura di quanto la presenza di A aumenta la probabilità di B (valori >1 indicano correlazione positiva).

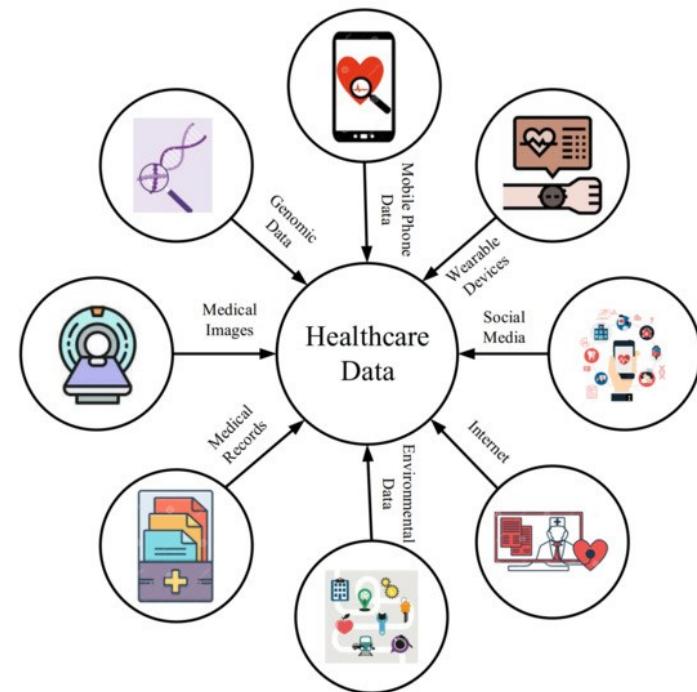


Fondamenti di Data Science e Machine Learning

Progetto di corso

Ambito Medico

- ▶ La **medicina** sta diventando sempre più data-driven.
- ▶ Differenti fonti di dati:
 - ▶ Cartelle cliniche elettroniche (EHR)
 - ▶ Imaging medico (radiografie, TAC, MRI)
 - ▶ Cartelle cliniche elettroniche (EHR)
 - ▶ Wearables e sensori remoti
 - ▶ Note cliniche/testi non strutturati
- ▶ Obiettivi principali:
 - ▶ Diagnosi precoce
 - ▶ Personalizzazione delle terapie
 - ▶ Prevenzione e monitoraggio
 - ▶ Ottimizzazione delle risorse sanitarie



Challenge

- **Qualità e completezza dei dati:**
 - Dati mancanti, rumorosi o non strutturati
- **Privacy e sicurezza:**
 - Regolamenti come GDPR e HIPAA
- **Interpretabilità dei modelli:**
 - I clinici devono poter fidarsi dei modelli → Explainable AI
- **Bias nei dati:**
 - Disparità tra gruppi demografici
- **Validazione clinica:**
 - Difficoltà a trasferire un modello da laboratorio a pratica clinica

Classificazione multimediale

- **Problematica:**

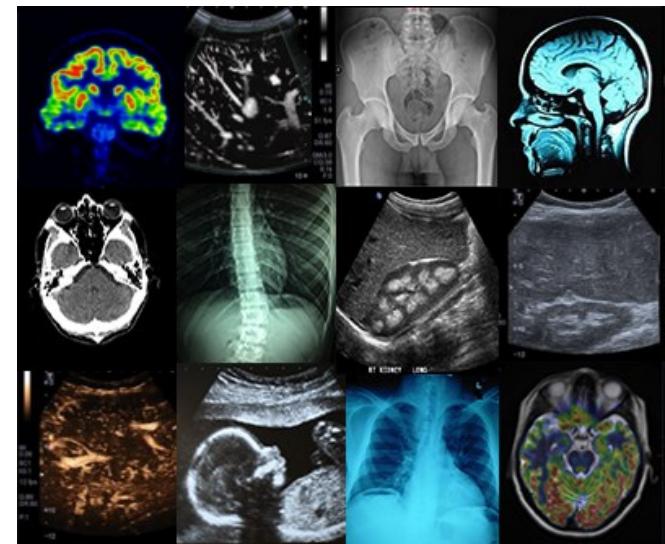
- Il processo di diagnosi è spesso svolto sia su dati clinici in formato numerico sia su dati in forma visuale che testuale.

- **Obiettivo:**

- Effettuare un analisi multimediale ed analizzare tecniche di feature engineering per ottenere diagnosi più accurate.
- Personalizzazione delle terapie (medicina di precisione)

- **Tecniche da utilizzare:**

- Reti neurali
 - Applicazione di tecniche di feature engineering
 - Applicazione di tecniche di fine tuning
- Reti Ibride (che combinano più tipologie di reti)
 - Applicazione di tecniche di fine tuning
 - Confronto su più reti



Anonimizzazione Dati Sensibili

- **Problematica:**
 - I dati medici sono spesso corredati da informazioni sensibili sul paziente finale.
- **Obiettivo:**
 - Analizzare e testare più tecniche di anonimizzazione dei dati sensibili.
 - Verificare come questi processi influiscano sulle performance dei modelli predittivi.
- **Tecniche da utilizzare:**
 - Modelli di ML, DL e NLP.



Fairness dei modelli di ML

- **Problematica:**

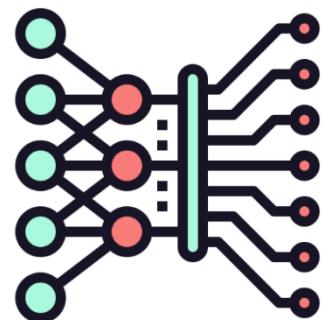
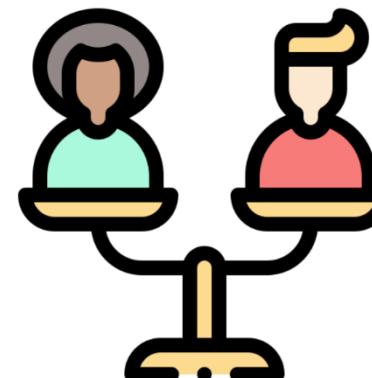
- I modelli di ML potrebbero produrre risultati meno accurati o ingiusti nei confronti di gruppi di pazienti facenti parte di minoranze.
- L'analisi della Fairness in questo ambito può ridurre diagnosi errate e disuguaglianze nell'accesso alle cure.

- **Obiettivo:**

- Rilevare bias di genere, età, etnia nei modelli predittivi.
- Valutazione tramite metriche di equità.
- Applicazione di strategie di mitigazione del bias.

- **Task da svolgere:**

- Applicazione di strategie di mitigazione del bias (reweighting/sampling, fairness-aware models).



Esempi di Datasets

Link	Tipologia
DiaTrend: https://www.synapse.org/Synapse:syn38187184/wiki/619490	Testo
Hupa-UCM: https://data.mendeley.com/datasets/3hbcs cwz44/1	Testo
UniSR	Multimodale
Sclerosis: https://data.mendeley.com/datasets/8bcts m8jz7/1	Multimodale
Mammografie	Immagini
...	...

Task del Progetto

- **Consegna Progetto e Documentazione**
 - Ogni gruppo/singolo dovrà:
 - Consegnare un documento PDF che dovrà:
 - Contenere una discussione introduttiva sul dominio del problema affrontare
 - Identificare delle Research Question (RQ) relative al problema affrontato
 - Per questioni di strutturazione del documento, vi consigliamo di utilizzare LaTeX (non obbligatorio)
 - Di seguito il riferimento ad un template pubblico disponibile su Overleaf:
Formato Documento: <https://www.overleaf.com/latex/templates/math-notes-template/kfqdrzrpvvk>

Formato Paper: <https://www.overleaf.com/latex/templates/elsevier-physics-open-journal-template/ryznykpwwmmc>
 - Consegnare il codice o i notebook del progetto opportunamente commentati e suddivisi in base delle analisi effettuate
- Nota 1: Eventuali progetti potranno essere utilizzati o continuati come lavoro di tesi
- Nota 2: Eventuali progetti potranno essere finalizzati a pubblicazioni scientifiche, qualora gli studenti siano interessati

Scelta del Progetto

La scelta del progetto può essere fatta compilando il file excel:

- Progetti*
- Il file dovrà essere riempito con:
 - Nomi partecipanti al progetto
 - Progetto scelto (Inserire solo l'ID corrispondente)
- Dopo aver compilato i campi, un membro del gruppo dovrà inviare una mail a gsolimando@unisa.it:
 - Oggetto: [FDSML] Scelta Progetto FDSML
 - Nel messaggio indicare:
 - I membri del gruppo, e il progetto scelto.

N.B. Una volta scelto il Progetto sarete affiancati da un tutor esperto nel progetto scelto.

- Il Progetto, con la relativa documentazione e codice commentato, dovranno essere consegnati al massimo 1 Settimana prima dell'appello.
- Non potranno prenotarsi all'appello d'esame coloro che non hanno ricevuto una valutazione per il Progetto.