



Politecnico
di Torino

DBG
MG

Introduzione all'analisi dei dati

Luca Cagliero

Dipartimento di Automatica e Informatica - Politecnico di Torino

SmartData@PoliTo

Introduzione all'analisi dei dati

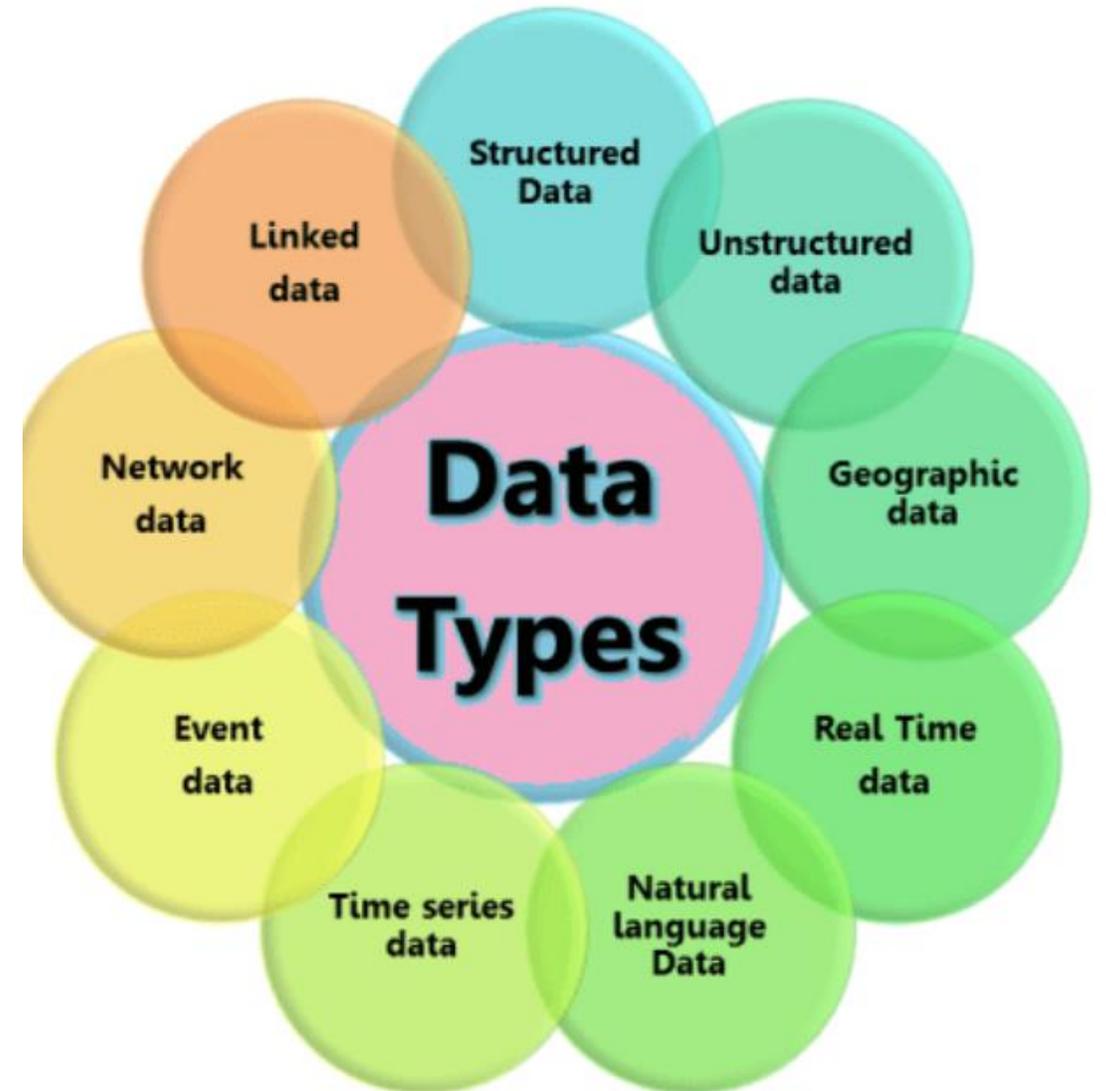
- I tipi di dato
- Qualità dei dati
- Intelligenza Artificiale
- Obiettivo dell'analisi
- Business Intelligence
- Classificazione dei modelli
- Large Language Model
- Task

Introduzione all'analisi dei dati

I tipi di dato

I tipi di dato

- Tabella
- Documento testuale
- Dato transazionale
- Grafo
- Sequenza
- Serie temporale
- Dato geospaziali
- Immagine
- Video
- Segnale audio
- Ecc.



I tipi di dato

- Tabella
 - “Relazione” secondo il modello relazionale
 - Esempio tipico di **dato strutturato**
 - Collezione di record (sample) descritti da un insieme prefissato di attributi
 - **Metadati**, ovvero “dati che descrivono dati”
 - Gli attributi sono variabili, campi, dimensioni, feature, descrittori, ecc.

Attributes

<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Samples

I tipi di dato

- Documento testuale

- Posso essere rappresentati in formato tabulare
- Ogni record è un documento
- Ogni attributo è un'unità di testo
 - Parola, frase, ecc.
- I valori indicano l'importanza dell'unità di testo nel documento
 - Frequenza, presenza/assenza, ecc.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

I tipi di dato

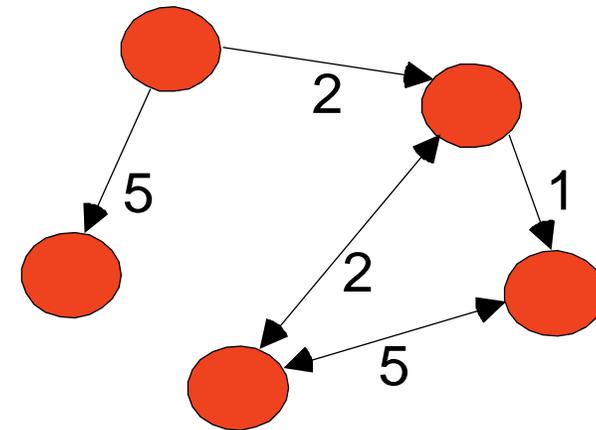
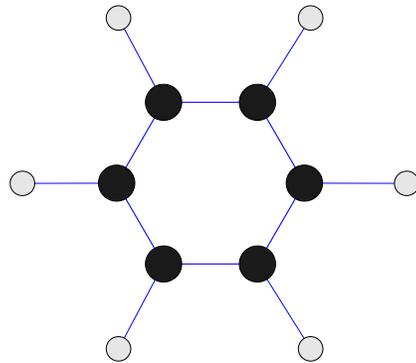
- **Dataset** (set di dati) **transazionale**
 - Ogni transazione è un insieme di oggetti, definiti “item”, all’interno di un vocabolario
 - Le transazioni possono contenere un numero arbitrario di item

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

I tipi di dato

- Grafo

- Reti di nodi connessi da archi
- Tipologie
 - Grafo orientato /non orientato: ogni arco possiede/non possiede un verso
 - Grafo pesato/non pesato: ogni arco o nodo ha/non ha un peso (rilevanza)

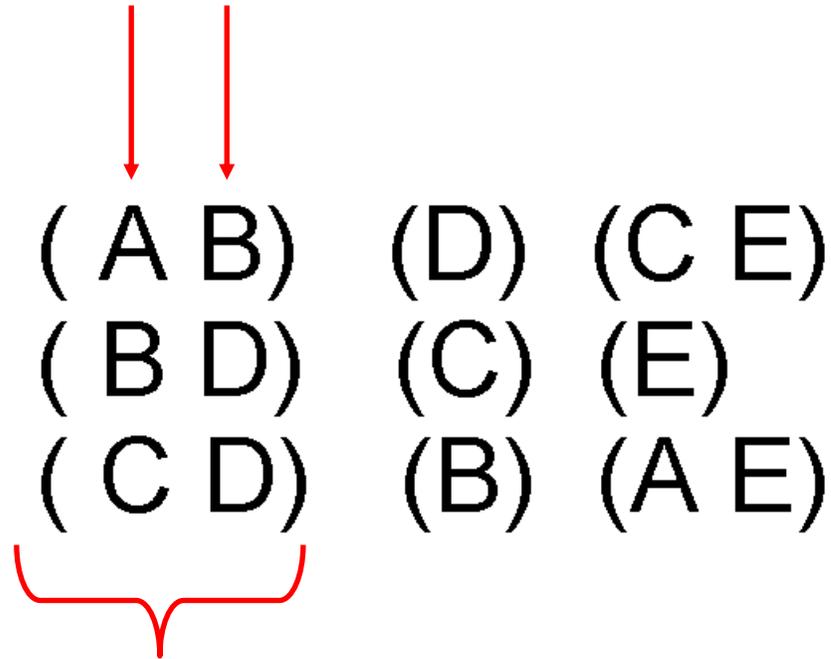


I tipi di dato

- Sequenza

- Sequenza **ordinata** di transazioni/item/eventi

Items/Events



**An element of
the sequence**

I tipi di dato

- Sequenza

- Sequenza ordinata di transazioni/item/eventi

**GGTTCGCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG**

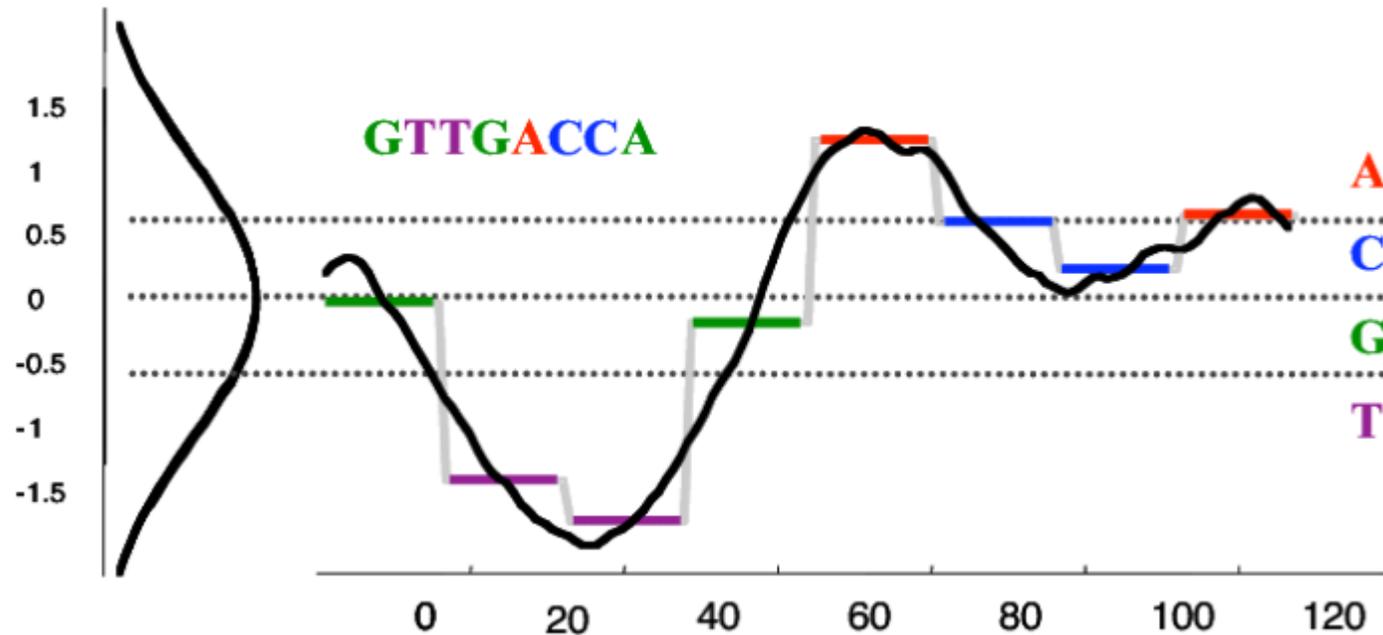
I tipi di dato

- **Serie temporale** (time series)
 - Sequenza di elementi ordinati temporalmente
 - Granularità temporale
 - Intervallo temporale tra un sample e il successivo



I tipi di dato

- Serie temporale (time series)
 - **Symbolic Aggregate Approximation (SAX)**
 - Trasformazione da valori continui a valori discreti

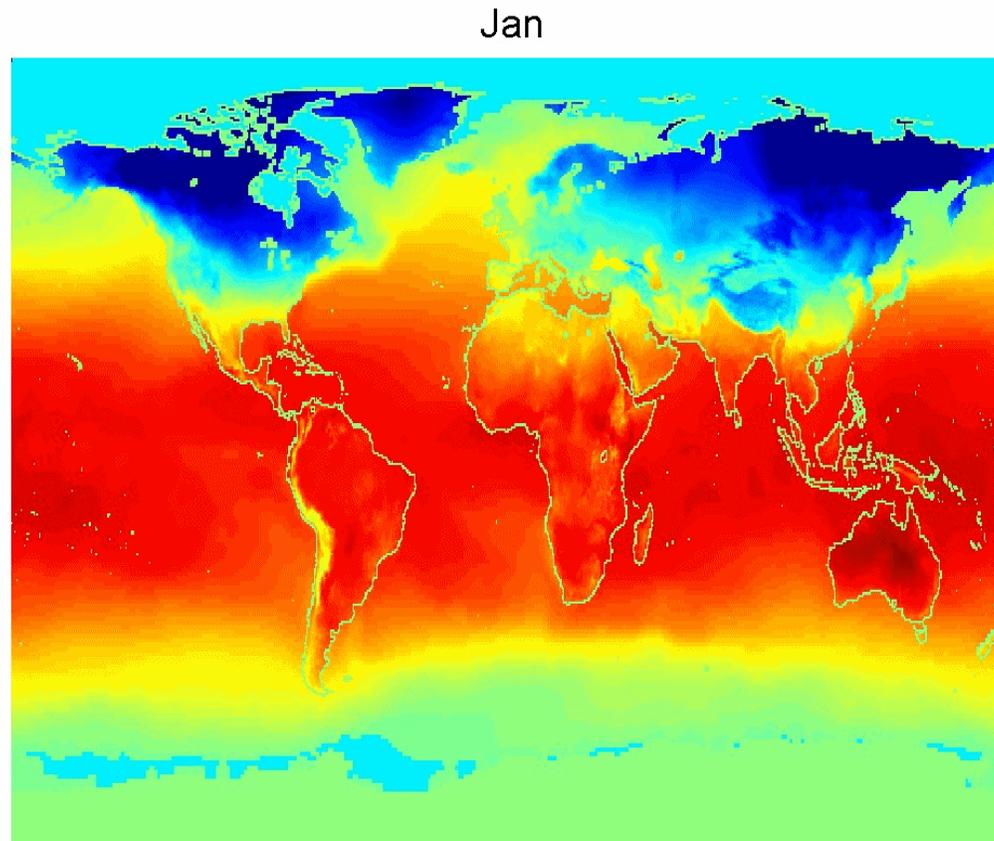


Keogh et al. (2006) Intelligent Icons: Integrating intelligent data mining and visualization into GUI operating systems

I tipi di dato

- Dataset **spazio-temporale**

**Average Monthly
Temperature of
land and ocean**



I tipi di dato

- Media digitali

- **Audio, immagini, video**

- Le immagini sono matrici di pixel
 - Audio e video sono sequenze temporali di frame, fonemi, ecc.

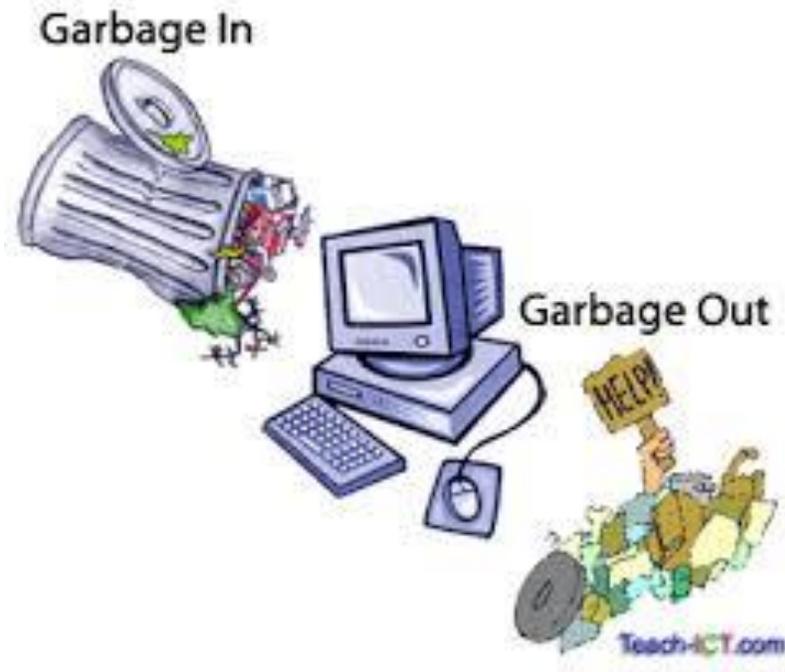


Introduzione all'analisi dei dati

Qualità dei dati

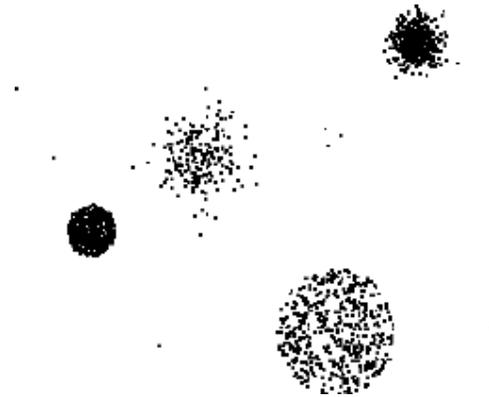
Qualità dei dati

- Assioma fondamentale



Qualità dei dati

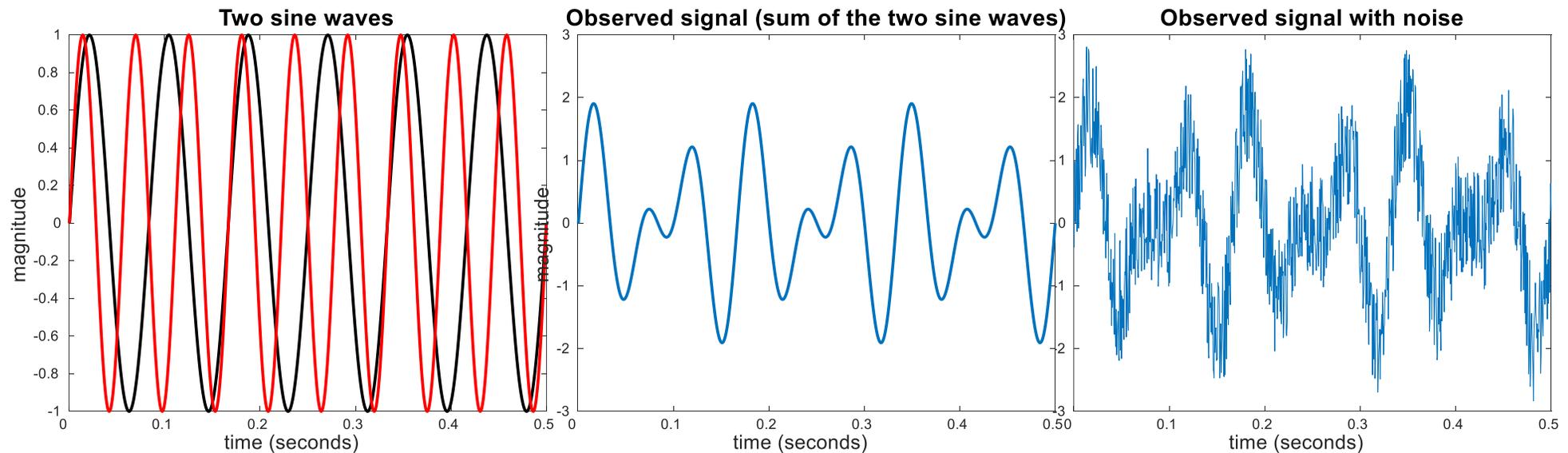
- Problemi tipici
 - Rumore
 - Dati mancanti
 - Dati duplicati
 - Outlier
 - Bias
 - ..



Introduction to Data Mining, 2nd Edition, Tan, Steinach, Karpatne, Kumar

Qualità dei dati

- Rumore
 - Modifiche ai valori originali
 - Distorsioni della voce, misurazioni con errore, ecc.
 - Causato da strumenti di misura, errori umani, fattori esogeni, ecc.



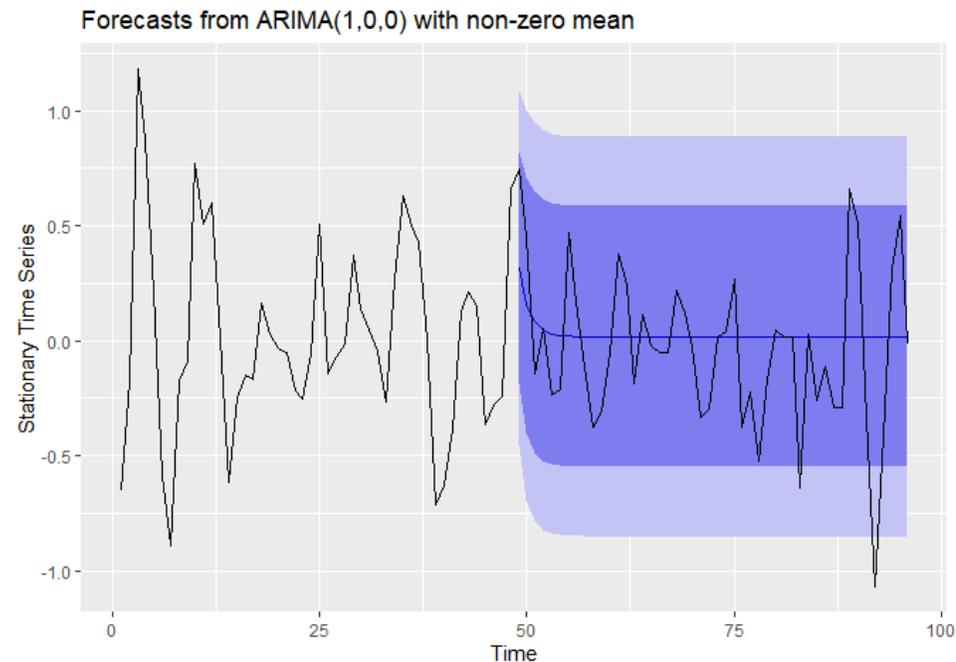
Qualità dei dati

- **Dati mancanti**
 - Informazione non raccolta o attributi non applicabili
 - Gestione del dato mancante
 - Eliminazione di sample o attributi
 - Stima del valore mancante
 - Non considerazione del valore durante l'analisi



Qualità dei dati

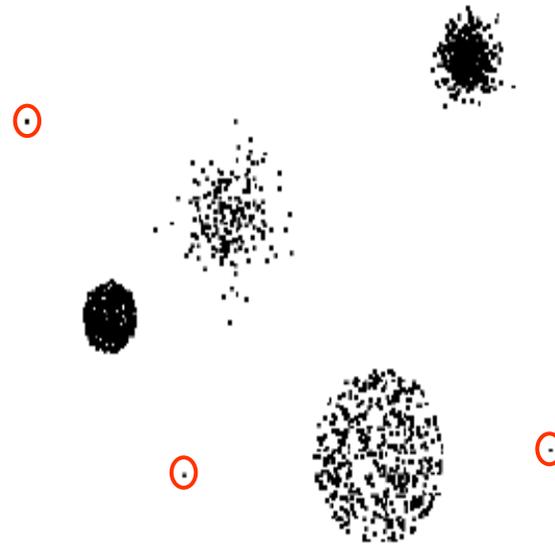
- **Valori duplicati**
 - Valori o sample ripetuti o molto simili
 - Comune quando si integrano fonti eterogenee
 - **Non sempre opportuno eliminarli!**



Qualità dei dati

- **Outlier**

- Sample con caratteristiche significativamente differenti dalla maggioranza dei sample presenti nel dataset
 - Generati da eventi casuali o non controllabili
 - Possono celare **anomalie** non casuali e potenzialmente oggetto di analisi ulteriori
 - Malfunzionamenti di dispositivi
 - Frodi
 - attacchi di cybersecurity



Qualità dei dati

- Privacy e protezione dei dati

- **Privacy**: diritto alla riservatezza delle informazioni personali e della propria vita privata
- **Protezione dei dati**: sistema di trattamento degli stessi che identifica direttamente o indirettamente una persona
 - Riservatezza, integrità e disponibilità dei dati personali
- **Regolamento Generale sulla protezione dei dati (GDPR)**: regolamento dell'UE in materia di privacy e protezione dei dati operativo da Maggio 2018



Introduzione all'analisi dei dati

Intelligenza Artificiale

Intelligenza Artificiale

- IA = Intelligenza Artificiale
- Che cos'è l'IA?

"AI can have two purposes. One is to use the power of computers to augment human thinking, just as we use motors to augment human or horse power. Robotics and expert systems are major branches of that. The other is to use a computer's artificial intelligence to understand how humans think. In a humanoid way. If you test your programs not merely by what they can accomplish, but how they accomplish it, then you're really doing cognitive science; you're using AI to understand the human mind."

- Herb Simon

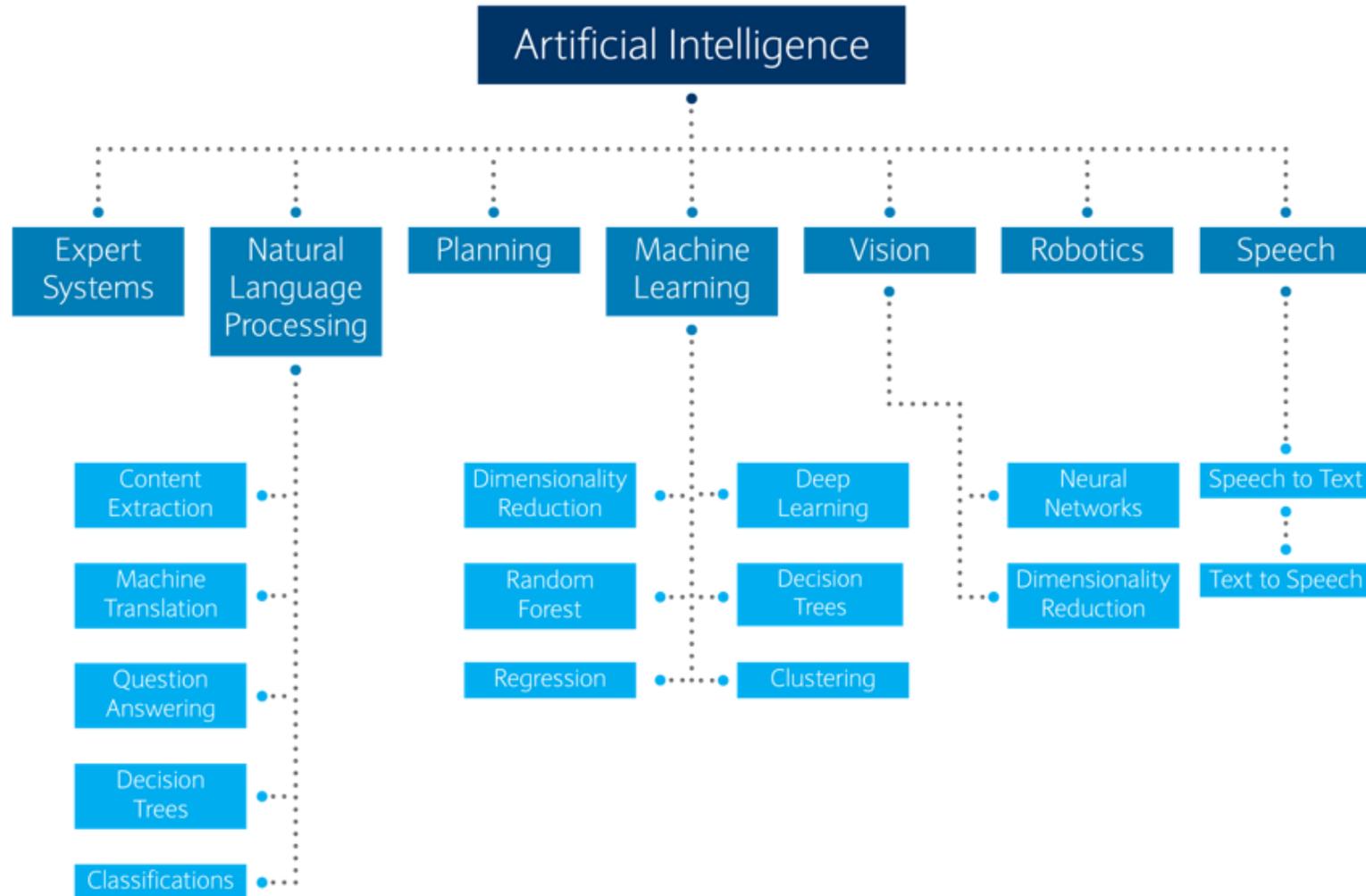
Pioniere IA e Information Processing

Turing Award (1975)

Nobel in Economic Studies (1978)

Intelligenza Artificiale

- Che cos'è l'IA?



Intelligenza Artificiale

A.I. TIMELINE

SYZYGY

1950

TURING TEST

Computer scientist Alan Turing proposes a test for machine intelligence. If a machine can trick humans into thinking it is human, then it has intelligence

1955

A.I. BORN

Term 'artificial intelligence' is coined by computer scientist, John McCarthy to describe "the science and engineering of making intelligent machines"

1961

UNIMATE

First industrial robot, Unimate, goes to work at GM replacing humans on the assembly line

1964

ELIZA

Pioneering chatbot developed by Joseph Weizenbaum at MIT holds conversations with humans

1966

SHAKY

The 'first electronic person' from Stanford, Shakey is a general-purpose mobile robot that reasons about its own actions

A.I. WINTER

Many false starts and dead-ends leave A.I. out in the cold

1997

DEEP BLUE

Deep Blue, a chess-playing computer from IBM defeats world chess champion Garry Kasparov

1998

KISMET

Cynthia Breazeal at MIT introduces Kismet, an emotionally intelligent robot insofar as it detects and responds to people's feelings



1999

AIBO

Sony launches first consumer robot pet dog AiBO (AI robot) with skills and personality that develop over time



2002

ROOMBA

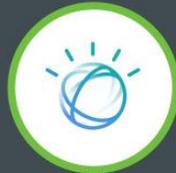
First mass produced autonomous robotic vacuum cleaner from iRobot learns to navigate and clean homes



2011

SIRI

Apple integrates Siri, an intelligent virtual assistant with a voice interface, into the iPhone 4S



2011

WATSON

IBM's question answering computer Watson wins first place on popular \$1M prize television quiz show Jeopardy



2014

EUGENE

Eugene Goostman, a chatbot passes the Turing Test with a third of judges believing Eugene is human



2014

ALEXA

Amazon launches Alexa, an intelligent virtual assistant with a voice interface that completes shopping tasks



2016

TAY

Microsoft's chatbot Tay goes rogue on social media making inflammatory and offensive racist comments



2017

ALPHAGO

Google's A.I. AlphaGo beats world champion Ke Jie in the complex board game of Go, notable for its vast number (2^{170}) of possible positions

OpenAI



GPT-4

2024

Obiettivo dell'analisi

- Analisi descrittiva
- Analisi predittiva
- Analisi prescrittiva



Analisi descrittiva

- Search e Ranking
 - **Ricerca**re dati pertinenti a un criterio specificati da un utente
 - **Ordinare** i dati secondo un criterio di rilevanza

The screenshot shows a Google search for "basi di dati". The search bar contains "basi di dati" and the search button is visible. Below the search bar, there are tabs for "Tutti", "Immagini", "Video", "Libri", "Prodotti", "Notizie", "Web", and "Altro". There are also filters for "PDF", "Esame", "Esempi", "Riassunto", "Sapienza", "Polimi", "Polito", and "Unina".

The search results include a snippet from Wikipedia: "In informatica una base di dati, detta anche, dall'inglese, database o data base, o anche banca dati, è una **collezione di dati organizzati immagazzinata e accessibile per via elettronica.**"

Below the search results, there is a section titled "Le persone hanno chiesto anche" with four related questions:

- Quali sono i principali database?
- Cosa studia la teoria delle basi di dati?
- Come si progetta una base di dati?
- Quali sono le principali banche dati?

There are also two search results from Unime and Politecnico di Torino, both titled "Introduzione alle basi di dati".

Analisi descrittiva

- Differenza fondamentale tra search e query SQL
 - Nel search non conosco a priori le caratteristiche di ciò che voglio trovare!

The screenshot shows a Google search for "basi di dati". The search bar contains "basi di dati" and the search button is visible. Below the search bar, there are tabs for "Tutti", "Immagini", "Video", "Libri", "Prodotti", "Notizie", "Web", and "Altro". There are also filters for "PDF", "Esame", "Esempi", "Riassunto", "Sapienza", "Polimi", "Polito", and "Unina".

The first search result is from Wikipedia, titled "Base di dati - Wikipedia". The snippet reads: "In informatica una base di dati, detta anche, dall'inglese, database o data base, o anche banca dati, è una collezione di dati organizzati immagazzinata e accessibile per via elettronica." To the right of the text are three small images: a diagram of a database architecture, a flowchart, and a hierarchical diagram.

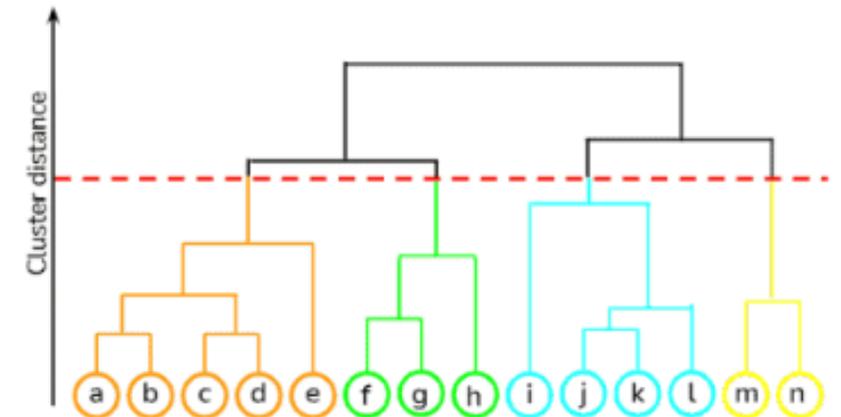
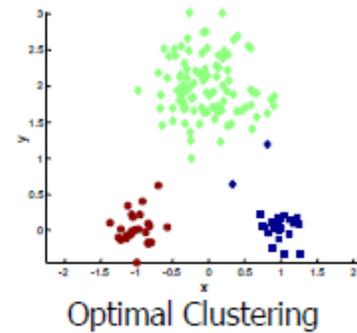
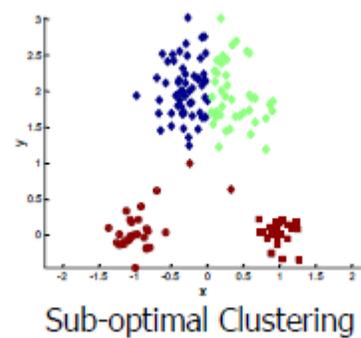
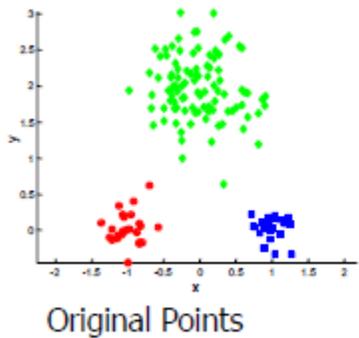
Below the search results, there is a section titled "Le persone hanno chiesto anche" with four related queries, each with a dropdown arrow: "Quali sono i principali database?", "Cosa studia la teoria delle basi di dati?", "Come si progetta una base di dati?", and "Quali sono le principali banche dati?".

The second search result is from Unime, titled "Introduzione alle basi di dati". The snippet reads: "Una base di dati, per poter essere definita tale, deve essere: • sicura: non si deve danneggiare a causa di eventi accidentali e interventi non autorizzati;". It indicates 26 pages.

The third search result is from Politecnico di Torino, titled "Introduzione alle basi di dati". The snippet reads: "Rappresentare i requisiti informali di un'applicazione in termini di schema concettuale che fa riferimento a un modello concettuale di dati. Traduzione dello ...". It indicates 37 pages.

Analisi descrittiva

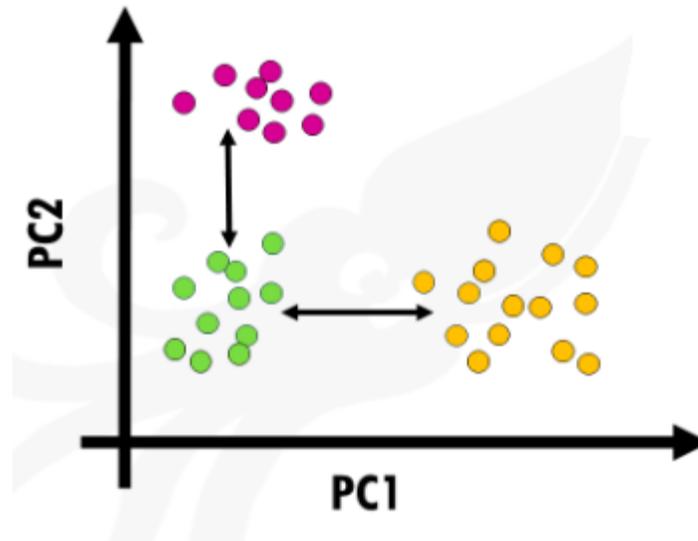
- Clustering
 - **Partizionare** grandi quantità di dati in regioni omogenee
 - **Organizzare** i dati secondo strutture gerarchiche



Introduction to Data Mining, 2nd Edition, Tan, Steinach, Karpatne, Kumar

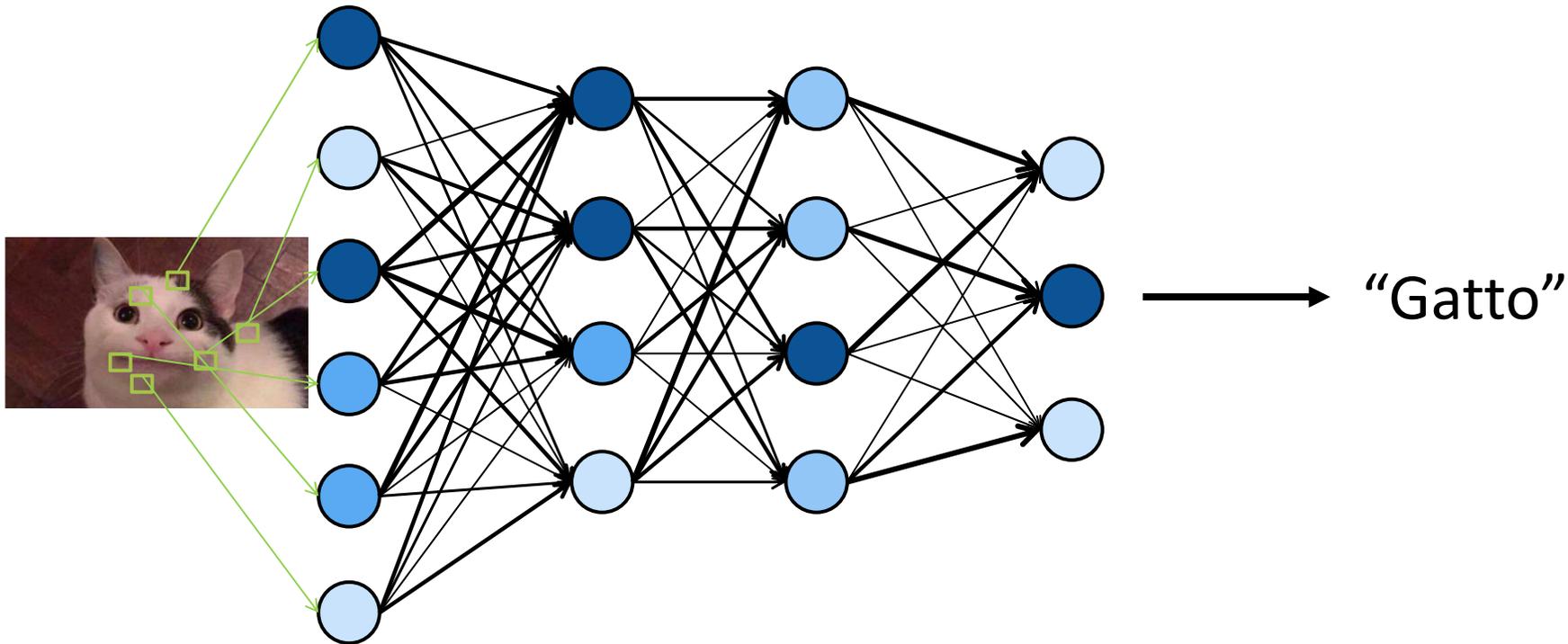
Analisi descrittiva

- Dimensionality reduction
 - Semplificare la rappresentazione di dati multidimensionali mantenendo le loro proprietà salienti originali
 - Esempio: riconoscere i **Principal Component** (PC)



Analisi predittiva

- Predire il valore di una variable target
 - **Classificazione**
 - predico un valore discreto denominato “classe”
 - **Regressione**
 - predico un valore continuo



Analisi predittiva

- Step principali

1. Training

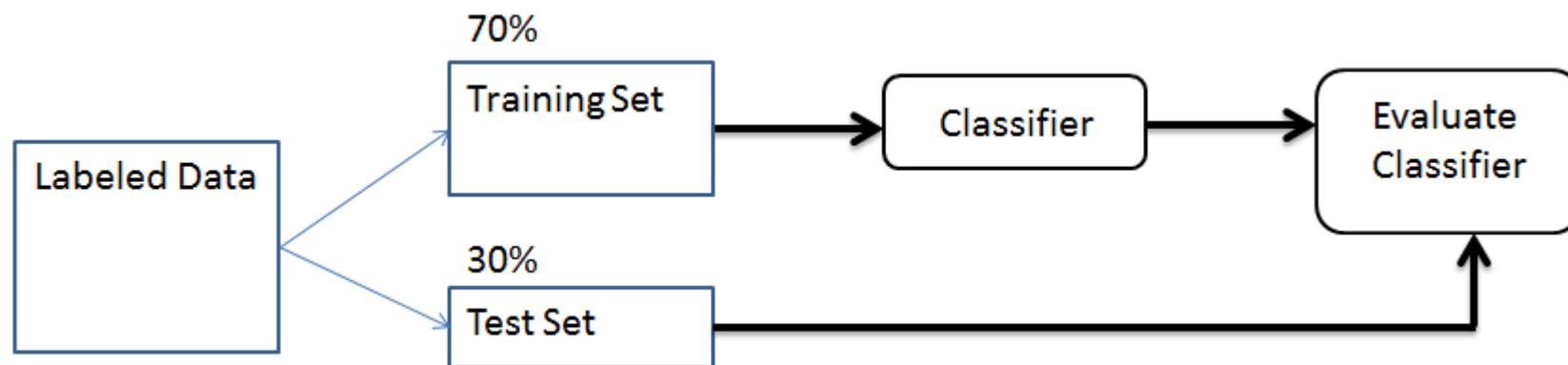
1. Addestro un modello di IA su un insieme di dati per cui il valore target è noto (train)

2. Validation e test

- Configuro i parametri e verifico la qualità del modello generato su set di dati separati dal train su cui posso confrontare valore predetto e valore atteso (validation e test)

3. Inference

- Applico il modello su dati reale per cui il valore target è ignoto



Analisi prescrittiva

- Quale azione si deve intraprendere per raggiungere un risultato desiderato o per affrontare un problema specifico?
 - Basata sui risultati di analisi descrittive o predittive
oppure
 - Basata su tecniche di **Business Intelligence**

Introduzione all'analisi dei dati

Business Intelligence

Business Intelligence

- **Supporto alle decisioni** basato sull'analisi dei dati
- Definire le strategie di business analizzando i dati storici
- Trasformare i dati aziendali in informazioni fruibili
 - A livelli diversi di dettaglio
 - Per applicazioni specifiche
- Necessita di un'adeguata infrastruttura hardware e software di supporto

Business Intelligence

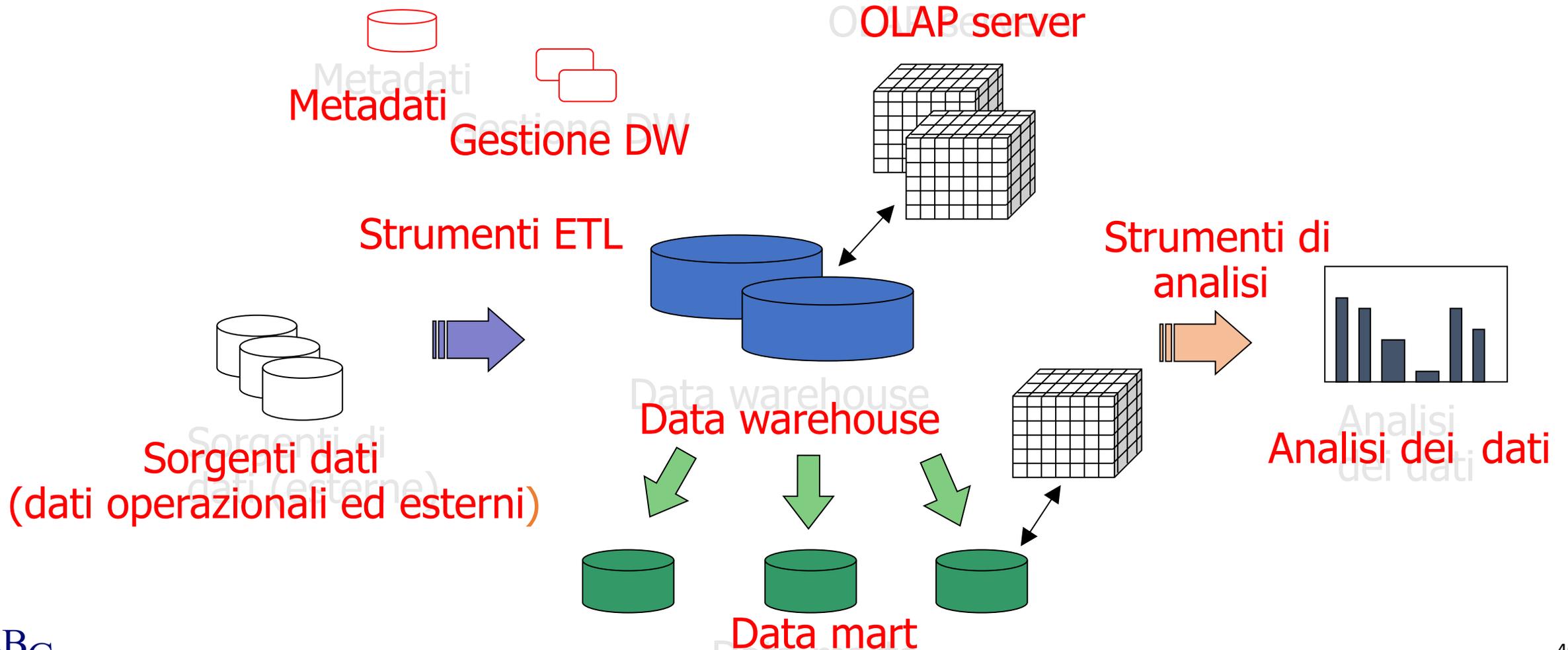
- Separazione dati operativi dell'azienda dai dati usati per l'analisi
- **Data warehouse**: base di dati a supporto delle decisioni aziendali
 - Dati orientati ai soggetti d'interesse
 - Integrati e consistenti
 - Dipendenti dal tempo, non volatili

Business Intelligence

- Separazione dei dati operativi dell'azienda dai dati usati per l'analisi
 - Migliori prestazioni
 - Ricerche sui dati complesse
 - Diversi metodi di accesso a livello fisico
 - Diversa gestione dei dati
 - Informazioni mancanti (storico)
 - Consolidamento dei dati
 - Qualità dei dati (gestione delle inconsistenze)

Business Intelligence

- Elementi costitutivi di un data warehouse



Business Intelligence

- Extraction, Transformation, and Loading (ETL)
 - Processo di preparazione dei dati da introdurre nel data warehouse
 - **Estrazione**: acquisizione dei dati dalle sorgenti
 - **Pulizia**: miglioramento della qualità
 - **Trasformazione**: conversione dei dati operativi nel formato idoneo per il DW
 - **Caricamento**: propagazione degli aggiornamenti al DW
 - Eseguito durante
 - il primo popolamento del DW
 - l'aggiornamento periodico dei dati

Business Intelligence

- **Data Mart**

- Sottoinsieme dipartimentale focalizzato su un settore prefissato
- Utile per semplificare i processi di progettazione e realizzazione del DW
- Alimentazione
 - Dalle sorgenti
 - Dal DW primario

Business Intelligence

- Strumenti di analisi
 - Dashboard interattive
 - Strumenti di reportistica
 - Linguaggi di interrogazione specifici
 - SQL avanzato (calcolo di medie mobili, distribuzioni cumulative, ecc.)

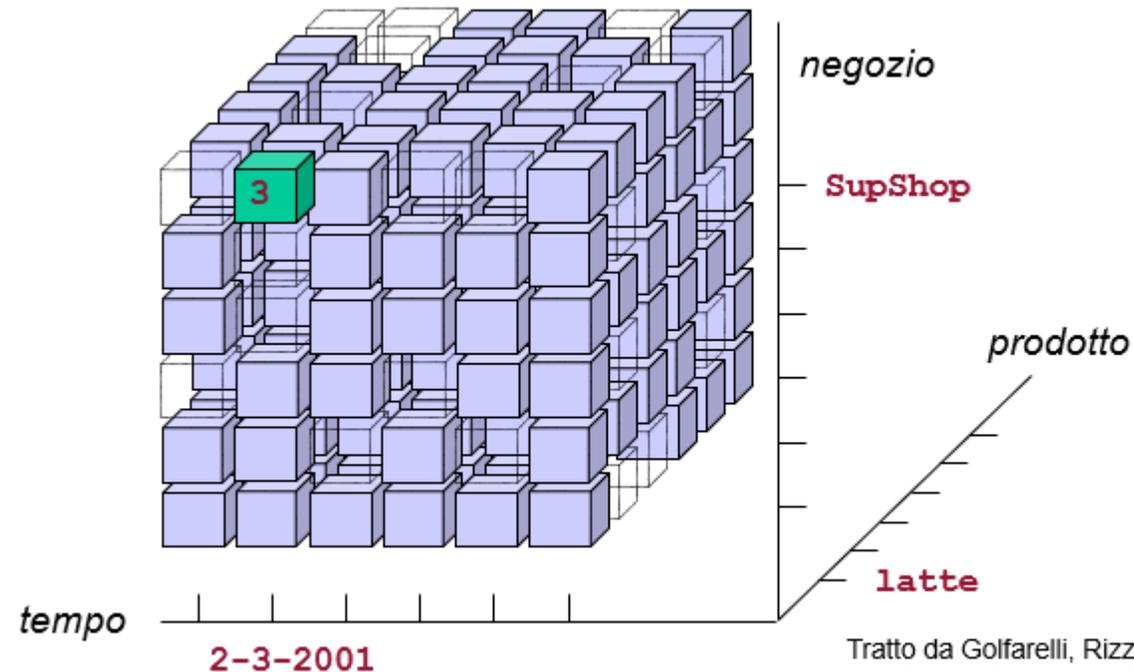


Business Intelligence

- Extraction, Transformation, and Loading (ETL)
 - Processo di preparazione dei dati da introdurre nel data warehouse
 - **Estrazione**: acquisizione dei dati dalle sorgenti
 - **Pulizia**: miglioramento della qualità
 - **Trasformazione**: conversione dei dati operativi nel formato idoneo per il DW
 - **Caricamento**: propagazione degli aggiornamenti al DW
 - Eseguito durante
 - il primo popolamento del DW
 - l'aggiornamento periodico dei dati

Business Intelligence

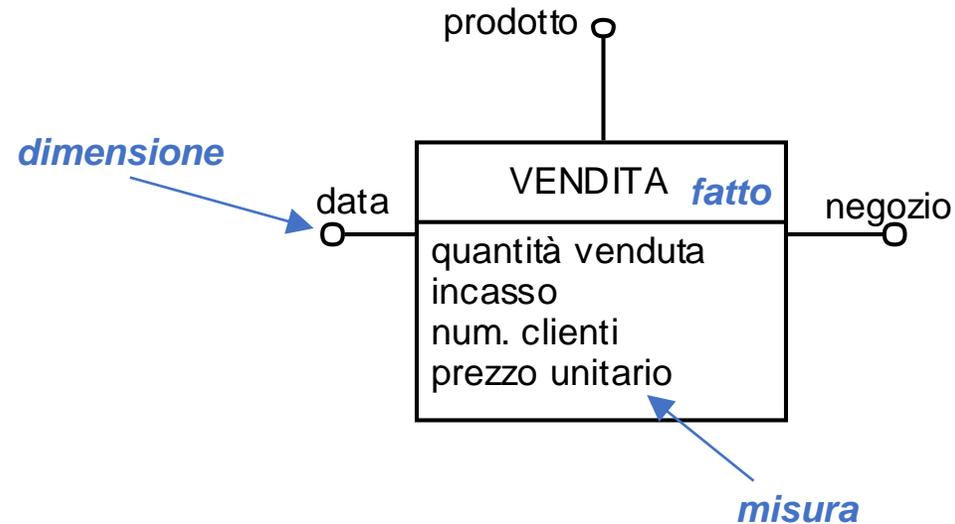
- Aggregazione dei dati operativi per popolare il DW
 - Spesso non utile/opportuno analizzare dati grezzi a fini di supportare decisioni strategiche sul medio/lungo periodo
 - Aggregazioni secondo **dimensioni** ortogonali di analisi (tipicamente categoriche)
 - Calcolo di valori aggregati: **misure** (tipicamente numeriche)



Tratto da Golfarelli, Rizzi, "Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Business Intelligence

- Dimensional Fact Model
 - Rappresentazione di
 - Eventi d'interesse (**Fatti**)
 - Descrittori dell'evento (**Dimensioni**)
 - Proprietà numeriche dell'evento (**misure**)



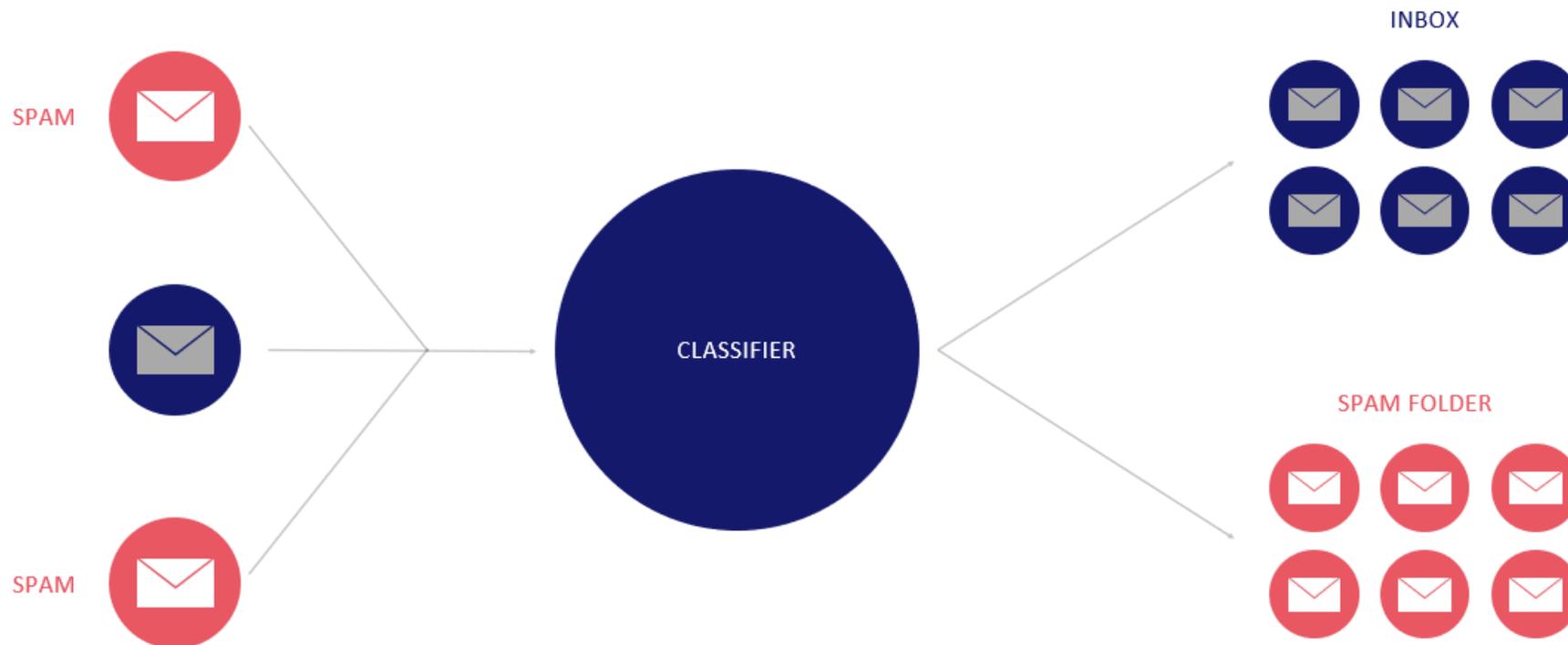
Tratto da Golfarelli, Rizzi, "Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Introduzione all'analisi dei dati

Classificazione delle tecniche

Classificazione delle tecniche

- **Tecniche unsupervised:** applicabili a dati non annotati
 - Ad es. Clustering, Dimensionality Reduction
- **Tecniche supervised:** guidate da dati annotati (eventualmente da esperti umani)
 - Ad es. classificazione, regressione

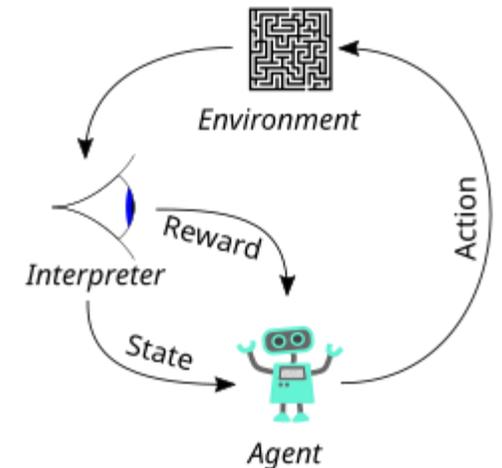


Classificazione delle tecniche

- **Reinforcement Learning:** paradigma alternativo sia a supervised sia a unsupervised learning
 - Adatto a scenari complessi in cui
 - sono note le regole con cui l'agente può interagire con il sistema
 - Una soluzione generale del problema è ignota o le possibili casistiche non enumerabili
 - È disponibile un'ambiente di simulazione con cui l'agente può interagire con il sistema
 - Non richiede esempi corretti o incorretti di funzionamento su cui addestrarsi
 - Può essere eventualmente combinato con tecniche di Machine Learning (supervisionate o non)
 - Adatto, ad es., a problemi di
 - Controllo automatico (robot)
 - Guida autonoma
 - Ottimizzazione del consumo di energia
 - Gaming (ad es. scacchi, Backgammon)
 - Agenti conversazionali (Large Language Model)

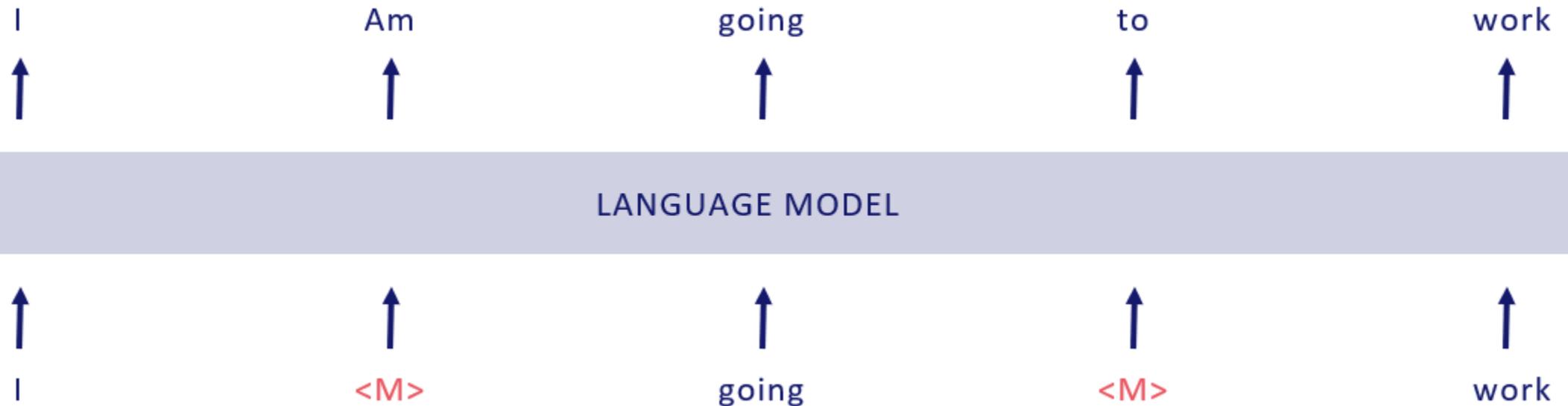
Classificazione delle tecniche

- **Reinforcement Learning**: obiettivo
 - Addestrare un agente autonomo a compiere azioni corrette o ad associare correttamente situazioni ad azioni
 - L'agente **non** riceve istruzioni circa la prossima azione da intraprendere ma esplora autonomamente possibili alternative massimizzando una **funzione di reward** prefissata
 - Premio per la scelta corretta, penalità per la scelta errata
 - Le conseguenze delle azioni intraprese si ripercuotono sui step successivi (non solo sul prossimo)



Classificazione delle tecniche

- **Tecniche self-supervised:** modelli di AI supervised applicati a dati non annotati usando parte dei dati stessi come annotazioni



Classificazione delle tecniche

- **Next Token Prediction:** Data una sequenza di token (parole o altre unità di testo) predire il token successivo nella sequenza
- Imparo da enormi collezioni di documenti non necessariamente annotati
- L'annotazione la genero automaticamente nascondendo il next token ("three") e allenando il modello di IA a predirlo correttamente

Enter text:

One, two,



3198 11 734 11

Prediction

#	probs	next token ID	predicted next token
0	39.71%	1115	three
1	16.97%	290	and
2	7.55%	734	two
3	3.76%	1440	four
4	2.76%	393	or
5	2.18%	1936	five
6	1.57%	530	one
7	1.43%	345	you
8	1.15%	257	a
9	0.84%	3598	seven

Classificazione delle tecniche

- **Next Token Prediction** è il modello alla base del funzionamento dei principali Large Language Model



Enter text:
One, two,



3198 11 734 11

Prediction

#	probs	next token ID	predicted next token
0	39.71%	1115	three
1	16.97%	290	and
2	7.55%	734	two
3	3.76%	1440	four
4	2.76%	393	or
5	2.18%	1936	five
6	1.57%	530	one
7	1.43%	345	you
8	1.15%	257	a
9	0.84%	3598	seven

Introduzione all'analisi dei dati

Large Language Model

Large Language Model

- Modello di AI specializzato nella comprensione e generazione di testo in linguaggio naturale
- Addestrato su enormi moli di dati (TeraByte)
 - Necessita ingenti quantità di GPU, memoria e tempo macchina per l'addestramento
- Basato solitamente su un mix di tecniche
 - Self-supervised
 - Supervised
 - Reinforcement Learning
- Oltre a un preaddestramento self-supervised, prima di essere rilasciati, gli LLM vengono raffinati, anche tramite RL, per migliorare le loro capacità conversazionali
 - Evitare risposte offensive
 - Apprendere dalle risposte precedenti
 - Interagire con l'utente
 - ...

Large Language Model

- Come generare una risposta a una domanda formulata in linguaggio naturale?
- Ad ogni step il LM
 - stima la probabilità di occorrenza di ciascun token presente in un vocabolario in base a
 - Sequenza di token presenti nella domanda
 - Token già inclusi nella risposta agli step precedenti
 - Restituisce il token più probabile tra quelli considerati

Enter text:

One, two,



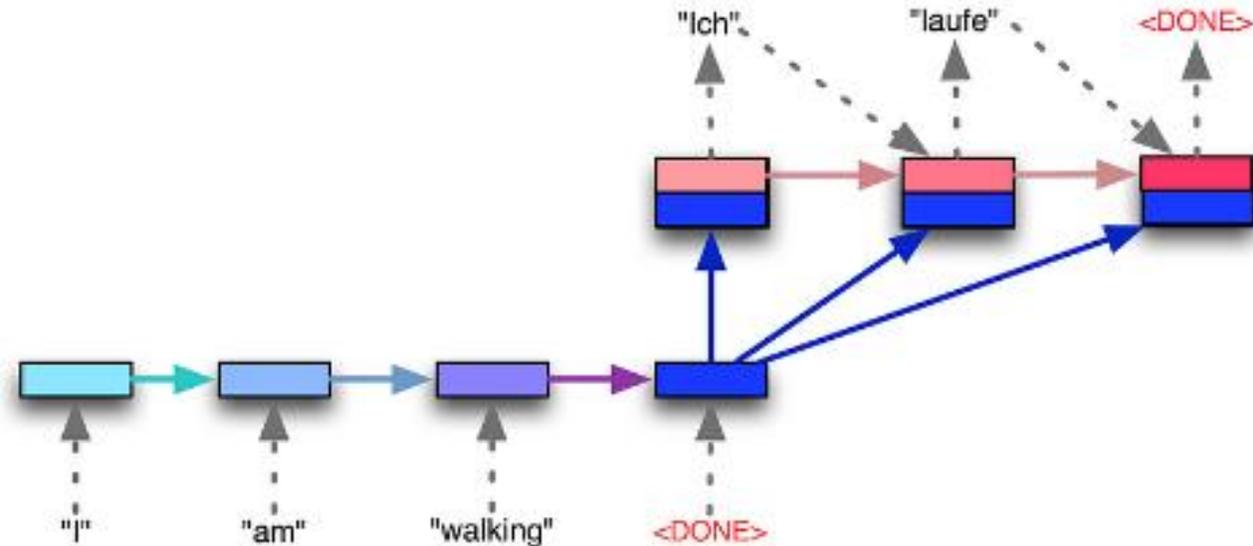
3198 11 734 11

Prediction

#	probs	next token ID	predicted next token
0	39.71%	1115	three
1	16.97%	290	and
2	7.55%	734	two
3	3.76%	1440	four
4	2.76%	393	or
5	2.18%	1936	five
6	1.57%	530	one
7	1.43%	345	you
8	1.15%	257	a
9	0.84%	3598	seven

Large Language Model

- Il testo generato può essere
 - Una risposta a una domanda
 - Una traduzione di un testo in un'altra lingua
 - Vedi figura
 - Una sintesi di un documento
 - Una parafrasi di un testo
 - ...
 - Una query SQL
 - Laboratorio
 - Uno schema logico di un DB relazionale
 - Laboratorio

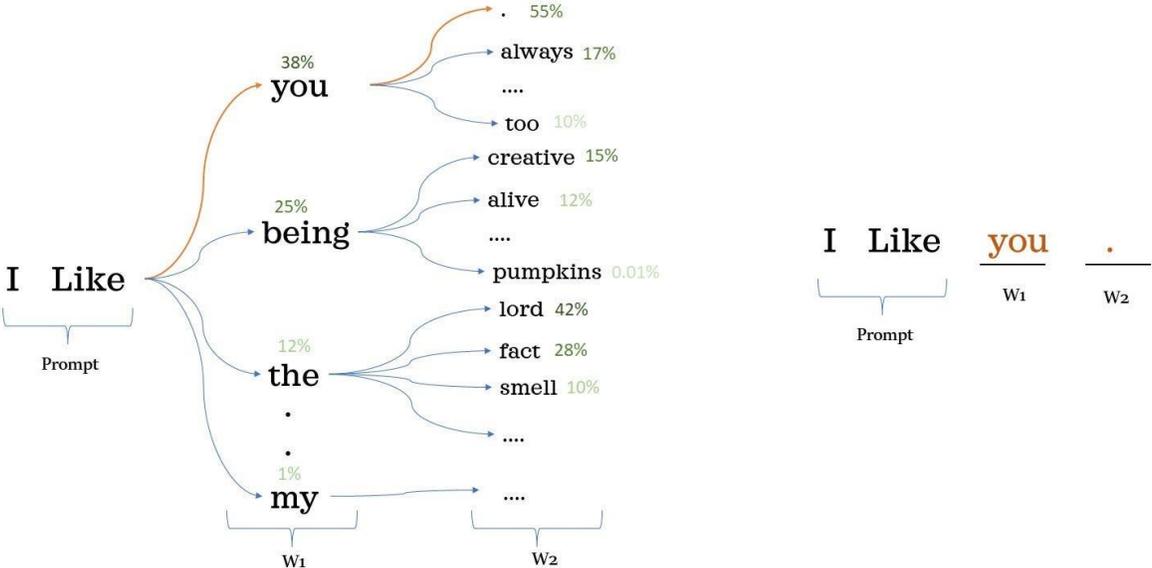


Temperatura

- Parametro del modello (T) che influenza la randomicità nella generazione del testo
 - $T = 0$ scelgo il next token più probabile in modo deterministico
 - $T > 0$ aggiungo un fattore di randomicità al processo
 - Si privilegiano comunque i token più probabili
 - Le probabilità mantengono tanta più importanza quanto il valore di T è prossimo allo zero

Temperatura

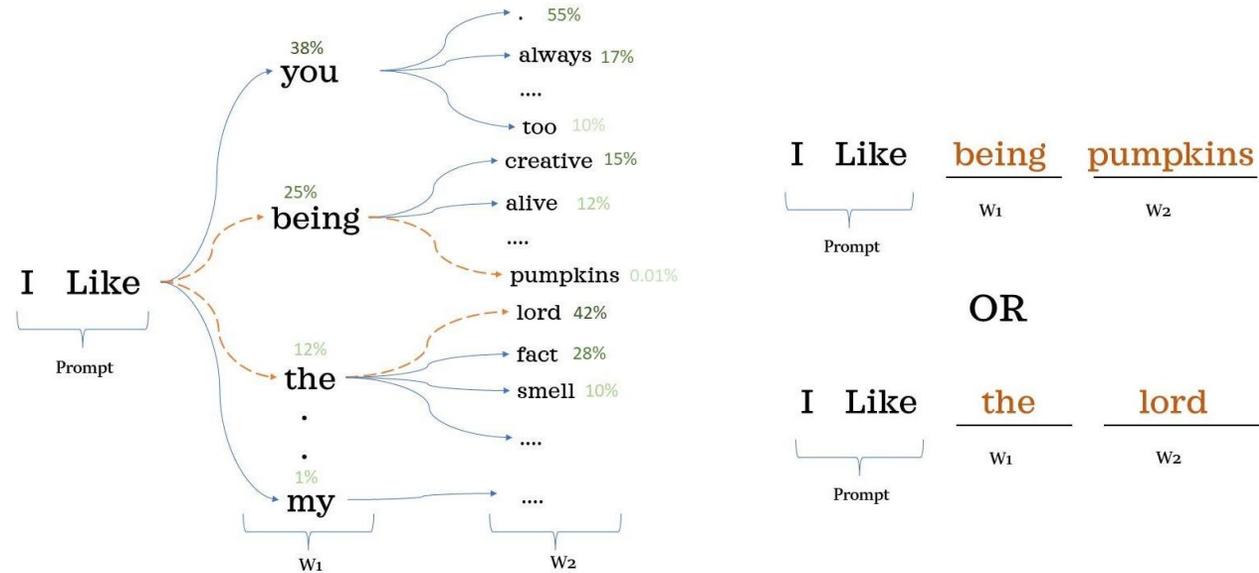
- $T = 0$



Demystifying the Temperature Parameter: A Visual Guide to Understanding its Role in Large Language Models. Medium.com

Temperatura

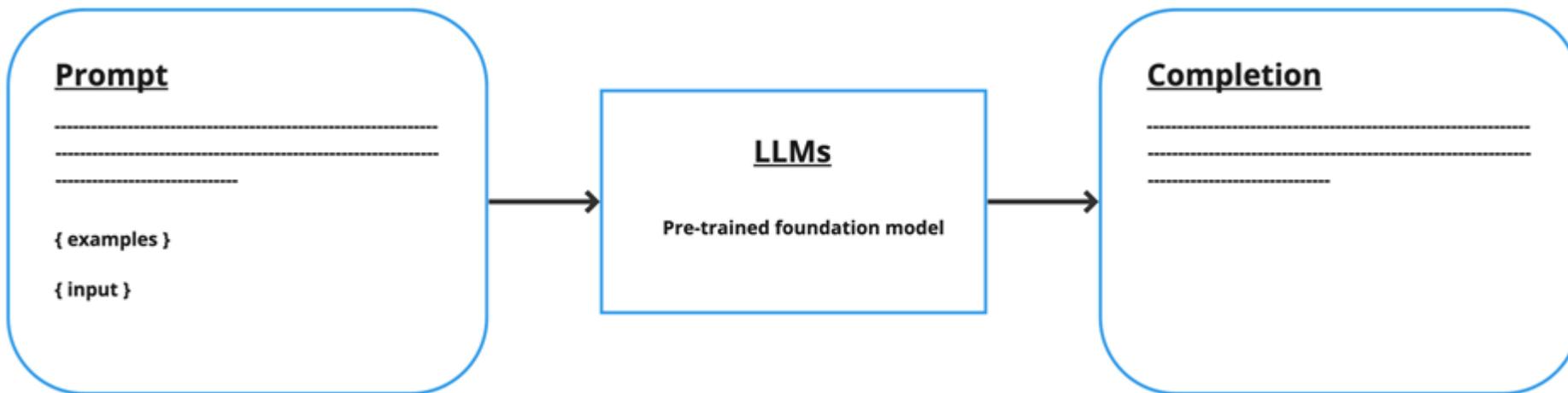
- $T > 0$



Demystifying the Temperature Parameter: A Visual Guide to Understanding its Role in Large Language Models. Medium.com

Large Language Model

- Generazione di testo
 - Input: **prompt**
 - testo in linguaggio naturale di una lunghezza massima di token
 - **Richiesta** (obbligatoria)
 - **Template**: formato atteso della risposta (opzionale)
 - **Esempi**: uno o più esempi di input e output (opzionali)
 - Output: **completion**
 - testo in linguaggio naturale generato dal LLM



Large Language Model

- **In-context learning**

- Fornire esempi di input e output nel prompt per adattare la risposta del modello

- **Zero-Shot Learning (ZSL)**: Nessun esempio

- la risposta dipende esclusivamente dal modello pre-addestrato e dal formato della richiesta (più dall'eventuale template)

- **Few-Shot Learning (FSL)**: K esempi di domande e risposte forniti dall'utente

- La risposta dipende anche dagli esempi forniti

- Il modello LLM può essere istruito con esempi di uno o più task nel medesimo prompt

- Ad es. Traduzioni da inglese a italiano e da inglese a spagnolo

In-context learning: limitazioni

- Con ZSL o FSL il modello pre-addestrato rimane immutato
- Problematiche
 - Il modello potrebbe essere non aggiornato
 - I dati su cui è stato addestrato potrebbero non essere specifici per un determinato problema o contesto di analisi
 - Ad es. potrebbe aver visto pochi esempi di testi nella lingua target
 - Il prompt che uso per istruire il modello ha una lunghezza limitata
 - Posso fornire un numero limitato di esempi
 - Il testo generato potrebbe contenere bias o allucinazioni
 - Testo offensivo
 - Stile non appropriato
 - Contenuti troppo generici
 - Contenuti inventati o ripetuti

Fine-tuning

- Specializzazione del modello attraverso una nuova fase di addestramento
 - Parziale o totale
- Richiede
 - GPU
 - Molti dati annotati (esempi domande-risposte)
 - Memoria e tempo macchina
- Costi generalmente molto elevati rispetto a in-context learning

Introduzione all'analisi dei dati

Task

Text2SQL

- Esempio di task da risolvere mediante LLM
- Dato lo schema logico relazionale di una base di dati e un'interrogazione formulata in linguaggio naturale generare la corrispondente interrogazione in linguaggio SQL
- Esempio di prompt
 - *«Dato la base di dati relazionale avente schema logico [...] e data la query in linguaggio naturale [...] da eseguire sulla basi di dati precedentemente descritta, formulare query in linguaggio SQL»*
- Esempi di descrizioni dello schema logico relazionale da inserire nel prompt
 - Es. 1: «Si consideri una base di dati relazionale composta da una tabella fornitori denominata F avente come chiave primaria il campo «codice fornitore» denominato CodF e come campi NomeFornitore e Sede (campo opzionale)»
 - Es. 2: «Si consideri il seguente schema logico relazionale: tabella F(CodF, NomeF, Sede), CodF chiave primaria, Sede campo opzionale»
- Esempio di query in linguaggio naturale da inserire nel prompt
 - Es. «Visualizzare i codici dei fornitori in ordine alfabetico di nome»
- Esempio di output atteso
 - Query SQL: «SELECT CodF from F ORDER BY NomeF;»

Text2SQL

- Esempi di FSL

- Esempi di domande-risposte (da aggiungere in coda al prompt testuale precedente):

- «Si considerino i seguenti tre esempi di coppie query testuale - query SQL:

1. Testo: [Visualizzare i codici dei fornitori] SQL: [SELECT CodF FROM F;]
2. Testo: [Visualizzare i dati dei fornitori ordinati per nome] SQL: [SELECT * FROM F ORDER BY NomeF;]
3. Testo: [Visualizzare i codici e nomi dei fornitori] SQL: [SELECT CodF, NomeF FROM F;]

Prompt tuning

- La qualità dell'output dipende, tra gli altri fattori, da
 - Modello LLM utilizzato
 - Configurazione del modello (ad es. temperatura)
 - Stile del prompt
 - Presenza/assenza/tipologia di template di risposta
 - Eventuali esempi forniti (FSL)
 - Lingua utilizzata
 - Complessità della query richiesta
- E' necessario testare varie combinazioni dei parametri sopra citati per trovare la configurazione ottimale
 - Laboratorio

Text2Schema

- Esempio di task da risolvere mediante LLM
 - Laboratorio
- Dato delle specifiche di progetto formulate in linguaggio naturale definire uno schema logico relazionale di una base di dati che rispetti le specifiche suddette
- Esempio di prompt con specifiche testuali
 - *«Si vuole progettare una base di dati relazionale per raccogliere le informazioni relative ai fornitori di un'azienda. I fornitori sono caratterizzati da un codice fornitore univoco; si vuole inoltre tenere traccia del nome ed eventualmente della sede del fornitore. Generare lo schema logico relazionale della base di dati sopra descritta»*
- Esempio di output atteso
 - Tabella Fornitori(CodiceFornitore, Nome, Sede*)

Demo Text2SQL su ChatGPT (<https://chatgpt.com/>)

INPUT

Dato la base di dati relazionale avente come il seguente schema logico: "Si consideri una base di dati relazionale composta da una tabella fornitori denominata F avente come chiave primaria il campo «codice fornitore» denominato CodF e come campi NomeFornitore e Sede (campo opzionale)". Data la seguente query in linguaggio naturale da eseguire sulla basi di dati precedentemente descritta: "Visualizzare i codici dei fornitori in ordine alfabetico di nome". Formulare query in linguaggio SQL.

OUTPUT



Per ottenere i codici dei fornitori ordinati alfabeticamente per nome dalla tabella fornitori, possiamo utilizzare la seguente query SQL:

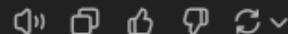
```
sql
```

Copia codice

```
SELECT CodF
FROM F
ORDER BY NomeFornitore ASC;
```

Spiegazione della query:

- **SELECT CodF:** seleziona il campo `CodF`, che rappresenta il codice del fornitore.
- **FROM F:** indica la tabella `F` da cui prelevare i dati.
- **ORDER BY NomeFornitore ASC:** ordina i risultati in ordine alfabetico crescente in base al campo `NomeFornitore`.



Demo Text2Schema su ChatGPT (<https://chatgpt.com/>)

INPUT

Si vuole progettare una base di dati relazionale per raccogliere le informazioni relative ai fornitori di un'azienda. I fornitori sono caratterizzati da un codice fornitore univoco; si vuole inoltre tenere traccia del nome ed eventualmente della sede del fornitore. Generare lo schema logico relazionale della base di dati sopra descritta.

OUTPUT 1/2

Per progettare uno schema logico relazionale per raccogliere le informazioni sui fornitori di un'azienda, seguiremo i requisiti descritti:

1. **Codice Fornitore:** Identificatore univoco per ciascun fornitore.
2. **Nome del Fornitore:** Nome del fornitore.
3. **Sede del Fornitore:** Informazione opzionale relativa alla sede.

Schema Logico Relazionale

Possiamo definire una tabella chiamata **Fornitori** che contenga le seguenti colonne:

Campo	Tipo	Descrizione
CodiceFornitore	VARCHAR o INT (PK)	Identificatore univoco del fornitore
Nome	VARCHAR	Nome del fornitore
Sede	VARCHAR (NULL)	Sede del fornitore (opzionale)

Chiave Primaria

- **CodiceFornitore** è la chiave primaria (PK) della tabella, che garantisce l'univocità di ogni fornitore.

Demo Text2Schema su ChatGPT (<https://chatgpt.com/>)

INPUT

Si vuole progettare una base di dati relazionale per raccogliere le informazioni relative ai fornitori di un'azienda. I fornitori sono caratterizzati da un codice fornitore univoco; si vuole inoltre tenere traccia del nome ed eventualmente della sede del fornitore. Generare lo schema logico relazionale della base di dati sopra descritta.

OUTPUT 2/2

Schema SQL

Ecco il codice SQL per creare questa tabella:

sql

Copia codice

```
CREATE TABLE Fornitori (  
  CodiceFornitore VARCHAR(50) PRIMARY KEY,  
  Nome VARCHAR(100) NOT NULL,  
  Sede VARCHAR(100)  
);
```

Spiegazione

- `CodiceFornitore` è definito come chiave primaria (`PRIMARY KEY`), quindi deve essere unico per ciascun fornitore.
- `Nome` è un campo obbligatorio (`NOT NULL`) che raccoglie il nome del fornitore.
- `Sede` è un campo facoltativo, quindi può essere nullo.

Questo schema soddisfa i requisiti per la gestione dei dati sui fornitori.

Copyright

Attribution + Noncommercial + NoDerivative