MASTER IN ANALISI DATI PER LA BUSINESS INTELLIGENCE E DATA SCIENCE

Il Machine Learning: motore di innovazione guidato dai dati

Rosa Meo

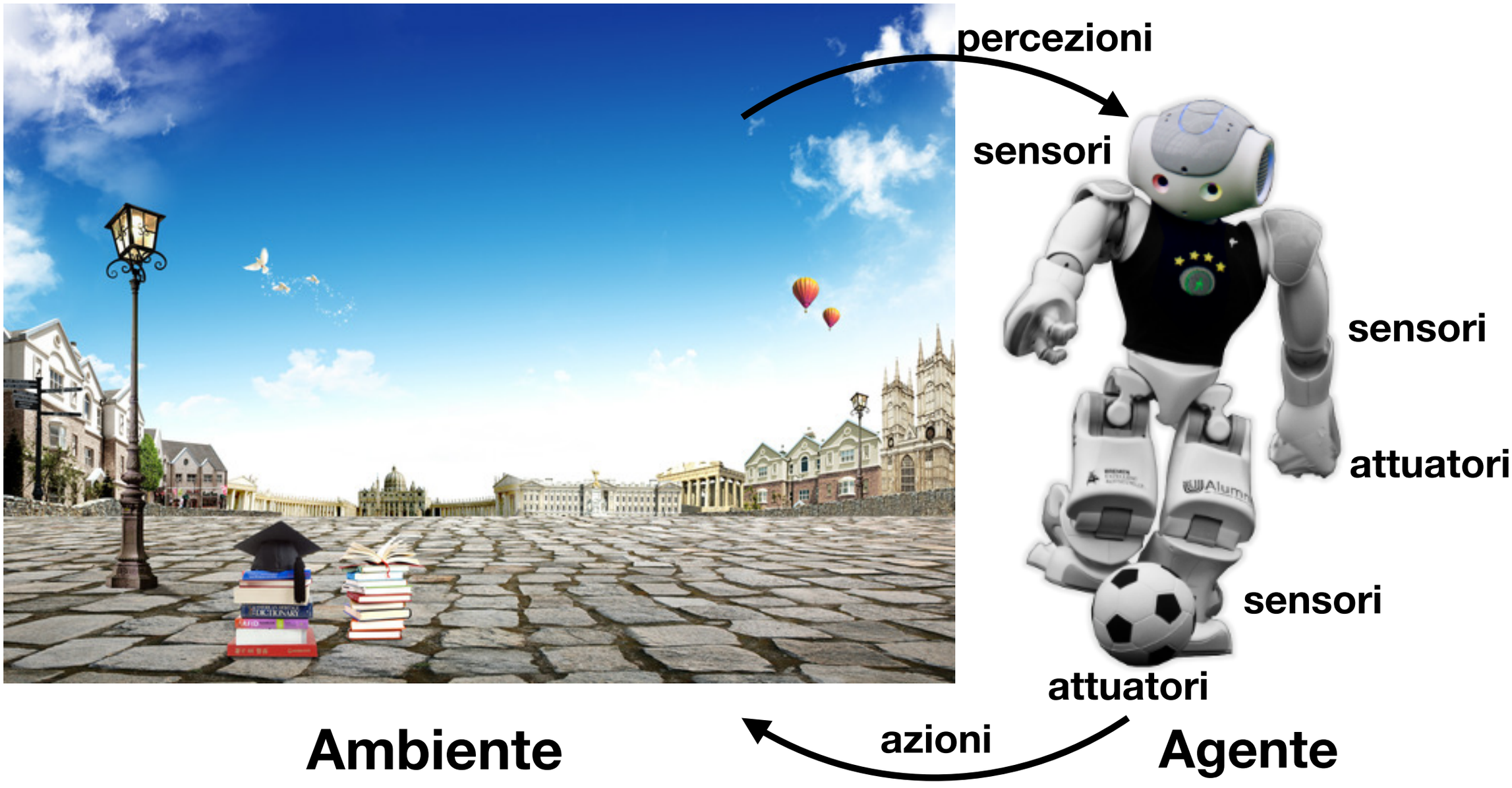
di.unito.it
DIPARTIMENTO DI INFORMATICA

Di cosa parleremo…

- Il Machine Learning e l'Intelligenza Artificiale
- I principali modelli di apprendimento del Machine Learning
- Recenti sviluppi con il Deep Learning
- I dati e gli open data
- Librerie software Open source
- Problemi e criticità
- Conclusioni

Il Machine Learning e l'Intelligenza Artificiale

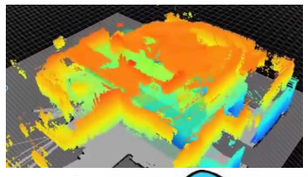
- Il Machine learning è una componente della filiera dei task che servono per realizzare l'Intelligenza Artificiale



La filiera dei task dell'IA

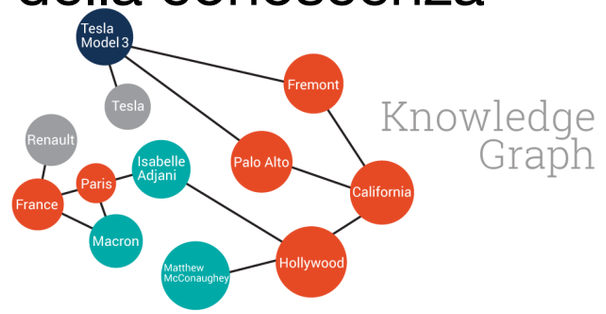
Percezione dell'ambiente

visione
comprensione
linguaggio naturale



Modellazione dell'ambiente

rappresentazione della conoscenza



Ragionamento

basato sulla conoscenza

Dati
+
Algoritmi
+
Potenza Computazionale

Azioni fisiche

robotica

Adattamento a situazioni nuove

apprendimento automatico
(machine learning)

Decisioni sulle azioni

decisione
pianificazione
schedulazione

I modelli del Machine Learning

- Apprendimento Supervisionato (Predittivo)
- Apprendimento Non Supervisionato (Descrittivo)
- Apprendimento Semi-Supervisionato
- Apprendimento per Rinforzo
- Apprendimento di Topics (Modelli semantici dei testi)

Apprendimento Supervisionato

- I sistemi di apprendimento automatico "apprendono" i concetti considerando esempi, senza essere programmati in base a regole specifiche del dominio

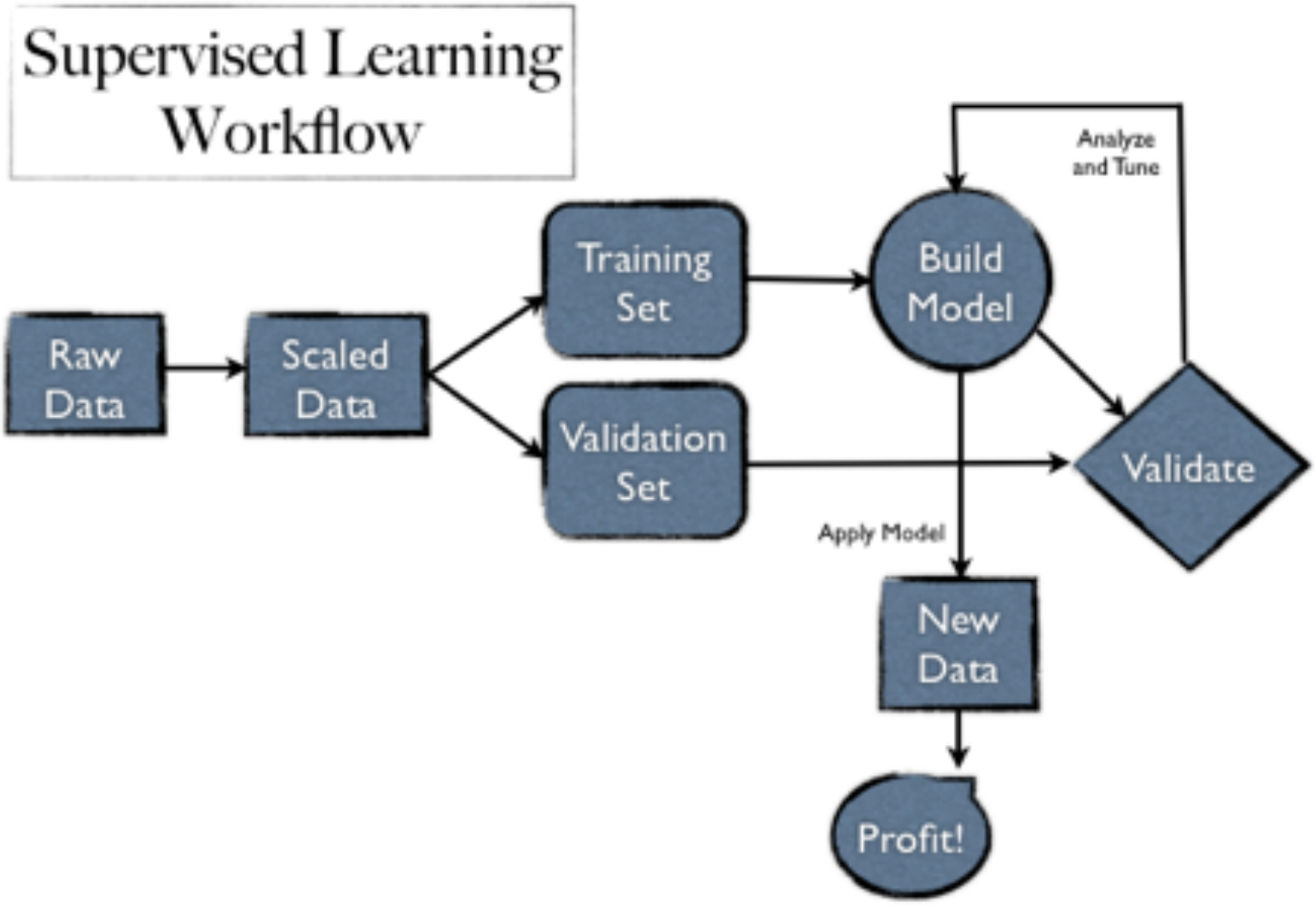
- Per esempio, nel riconoscimento degli oggetti in una immagine, riconoscono i gatti analizzando molte immagini che contengono gatti e che sono state manualmente etichettate come "gatto" o "non gatto"

- Usano poi i risultati per riconoscere i gatti in altre immagini

- Non hanno nessuna conoscenza a priori di come è fatto un gatto (baffi, pelo, coda, zampe, occhi)

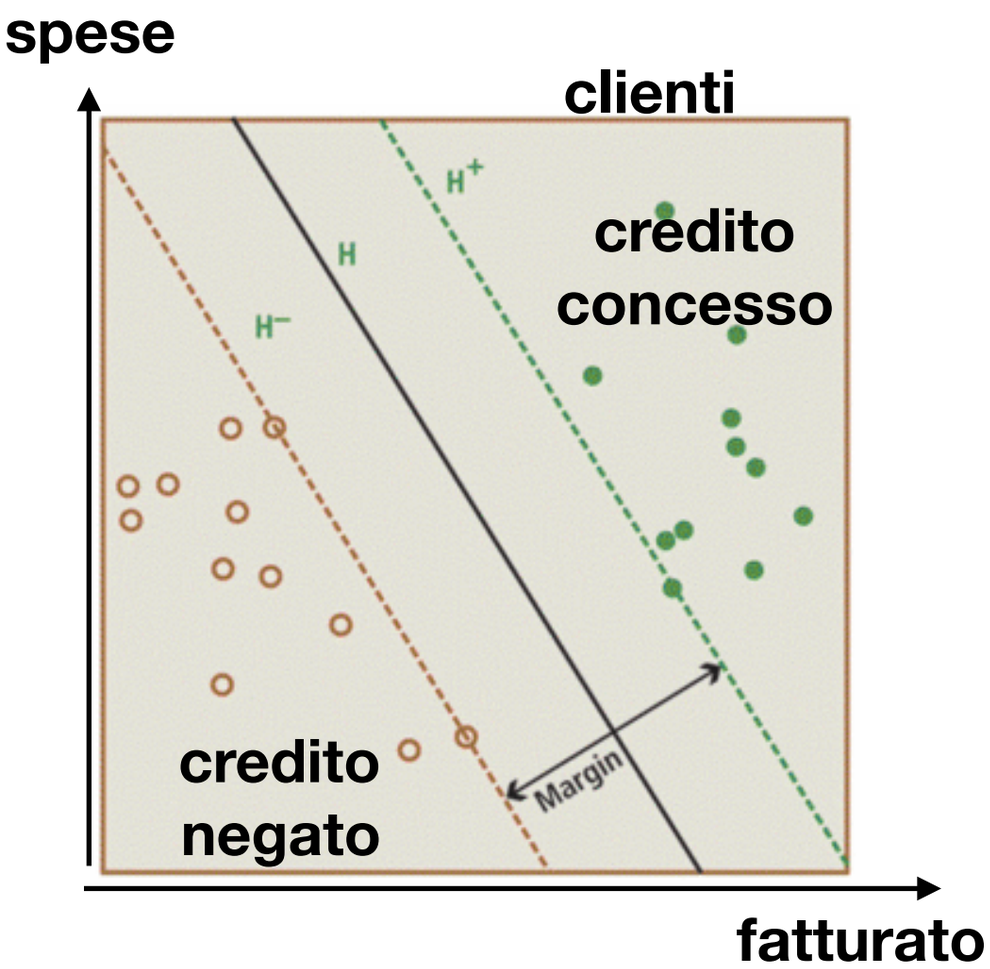
- Generalizzano le caratteristiche dei gatti dal materiale di apprendimento che hanno processato

Apprendimento Supervisionato

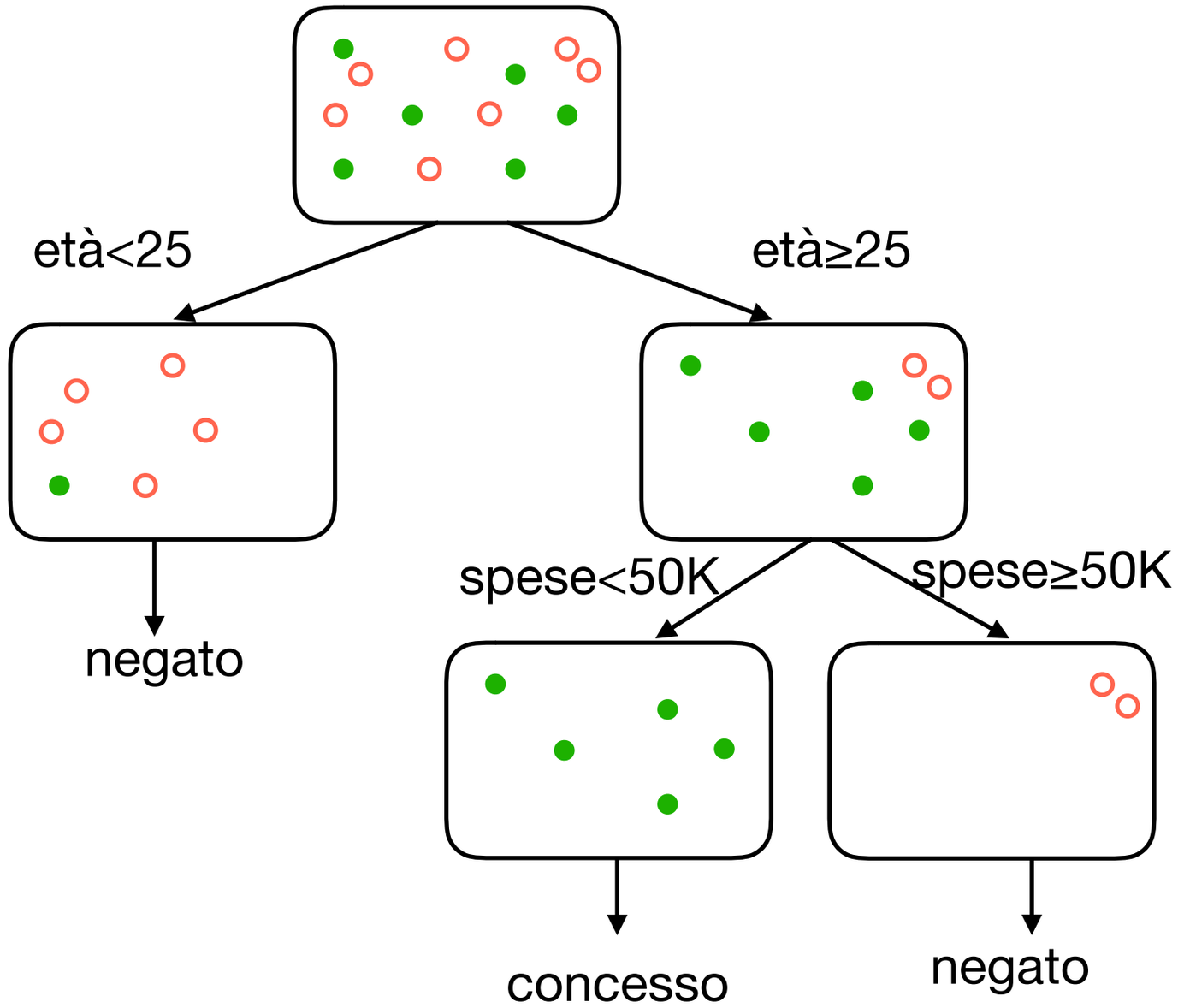


Modelli Supervisionati

Predizione della classe (valore di variabile target di interesse)

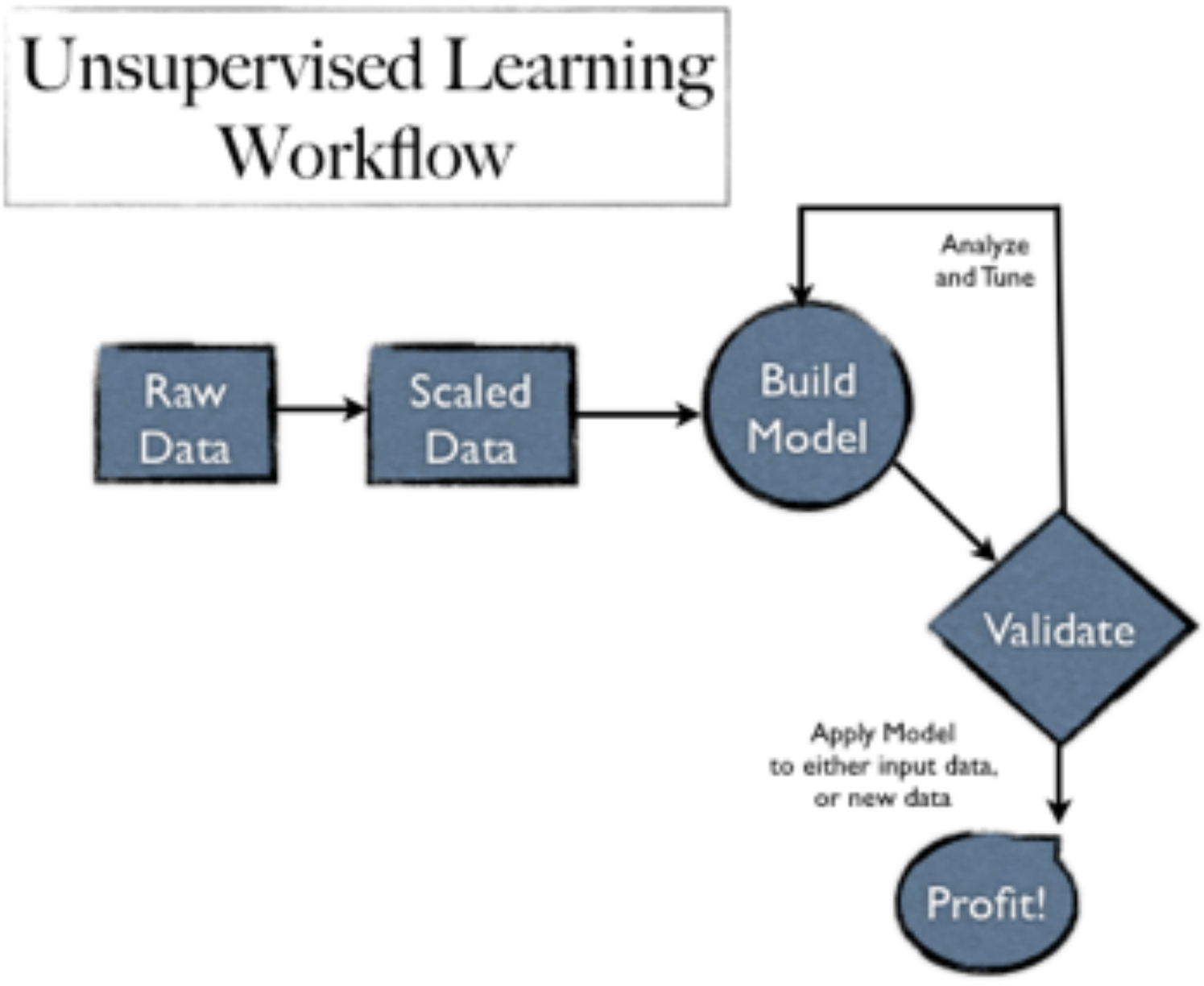


Support Vector Machines



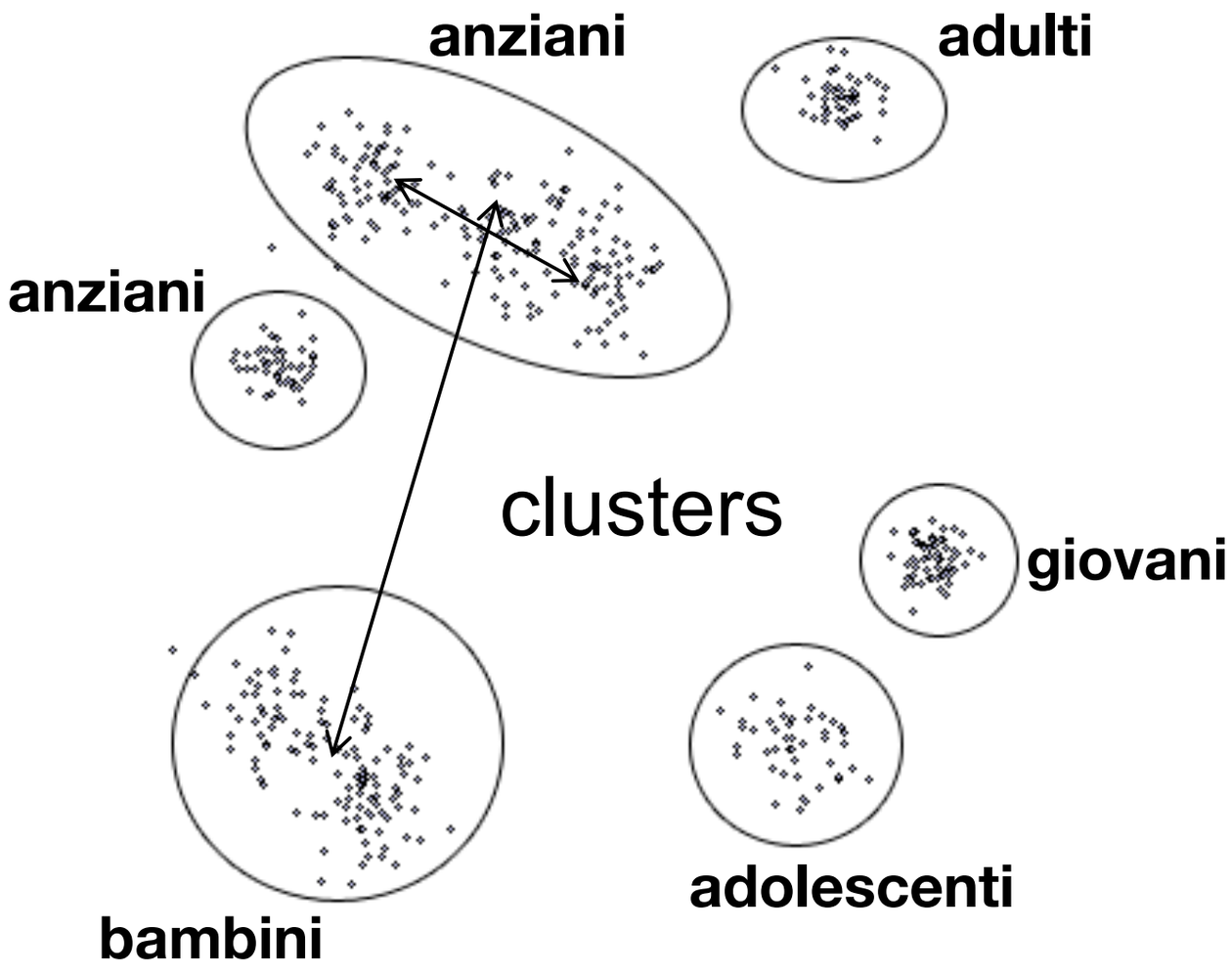
Alberi di decisione

Apprendimento Non Supervisionato



Modelli Non Supervisionati

Segmentazione e descrizione per analisi esplorativa



Apprendimento Semi-Supervisionato

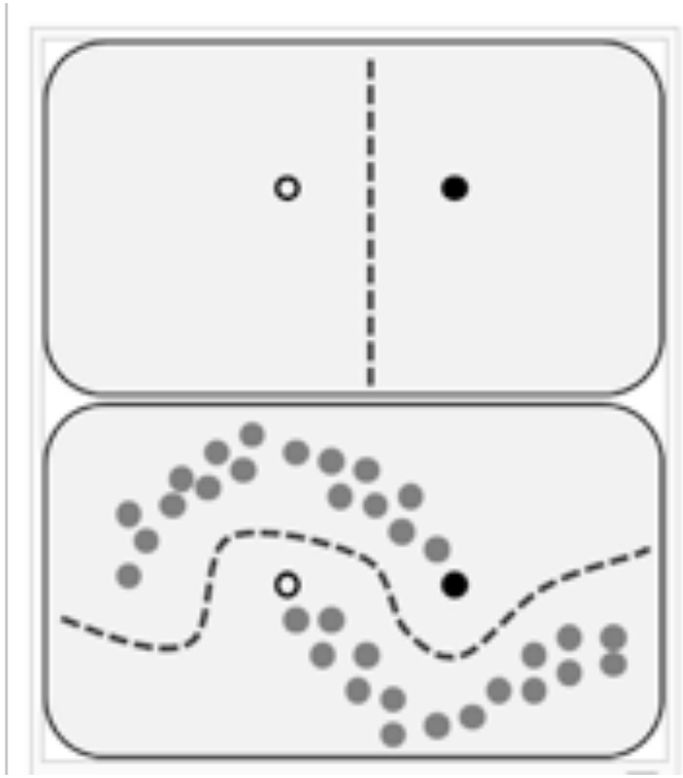
- Quando molti esempi non sono etichettati
- Differenze tra i modelli (bordo) appresi in:

A. Considera solo gli esempi etichettati

B. Considera tutti gli esempi.

   Crea cluster di esempi simili in base alla vicinanza o densità.

   Decide della classe degli esempi non etichettati in un cluster assegnando l'etichetta più verosimile (l'etichetta degli esempi vicini o della maggioranza)



A.

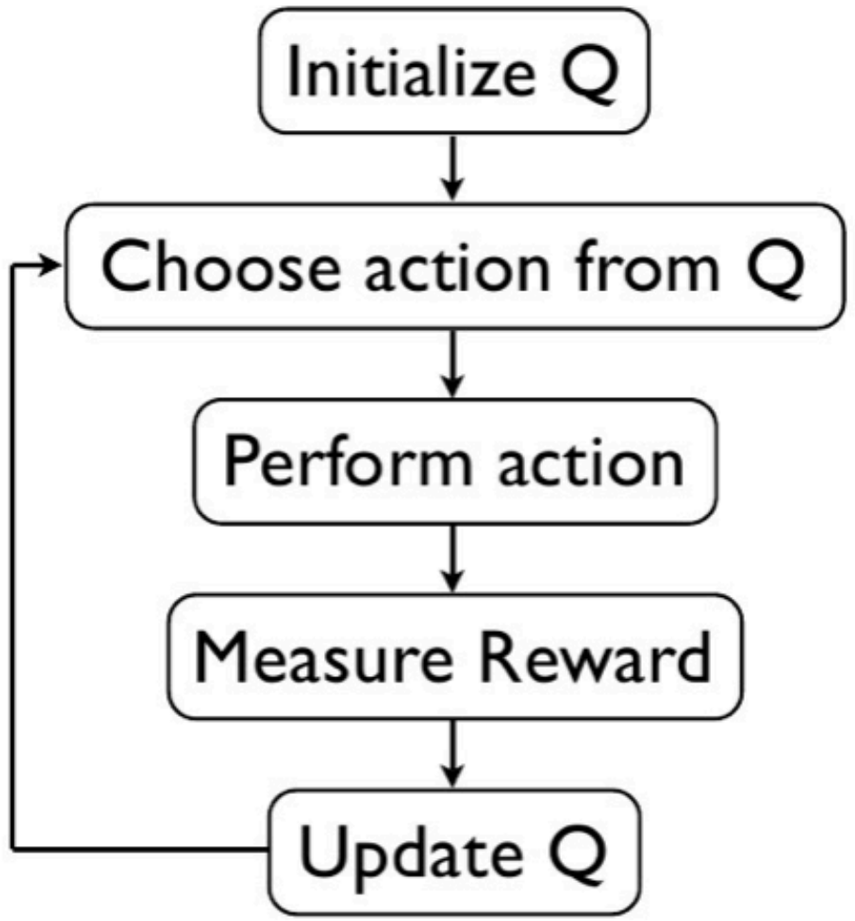
B.

Apprendimento per rinforzo

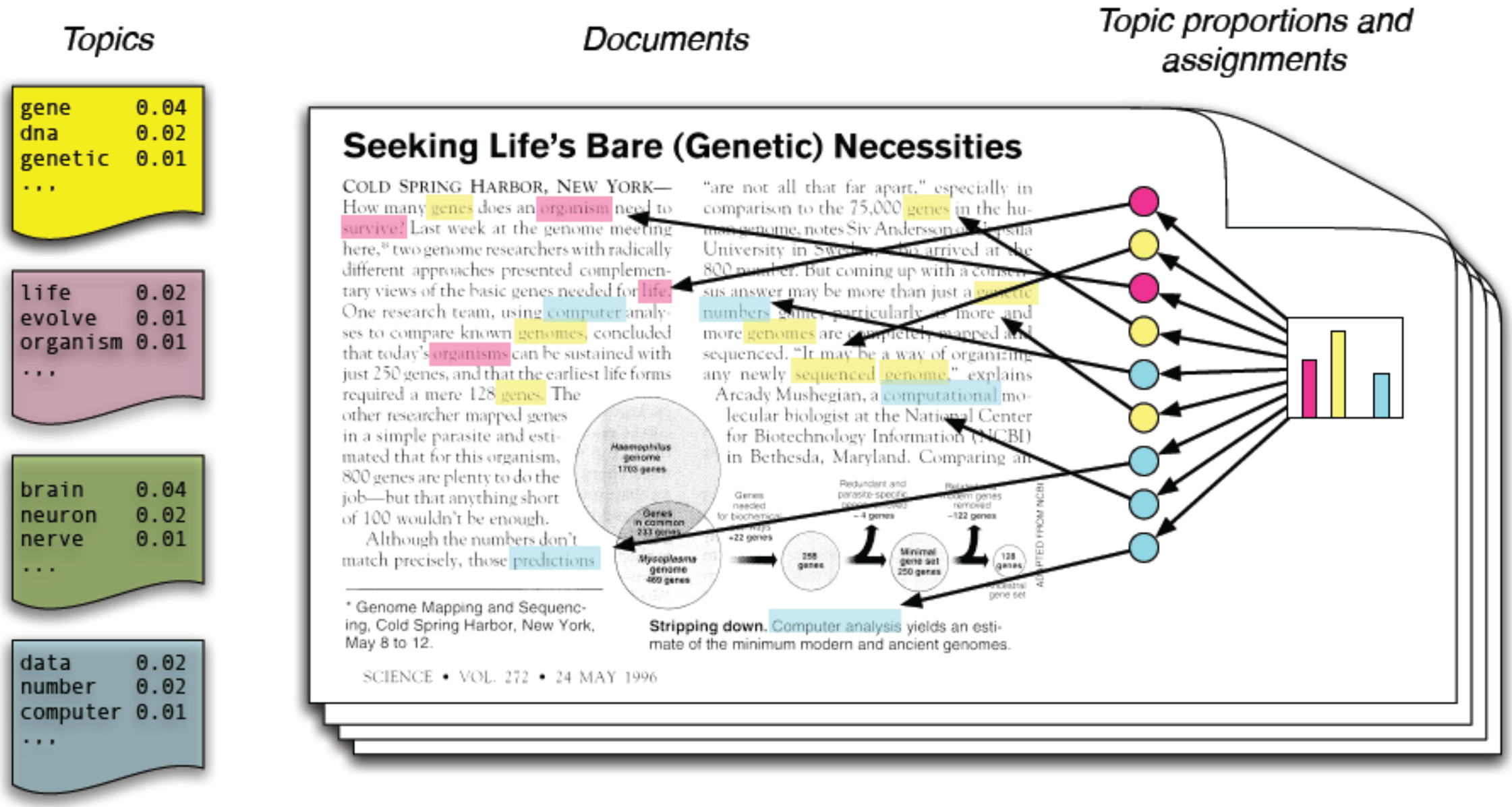
- E’ un metodo di apprendimento di un agente che interagisce con l’ambiente e deve raggiungere un obiettivo

- L’agente ha un insieme di stati (**s**) e di azioni (**a**) possibili

- Viene guidato nella scelta dell’azione in un certo stato da una funzione di **premio Q(s, a)** che riceve dall’ambiente in risposta alle proprie azioni



Apprendimento di Argomenti (Topics) Latenti nei Testi



Modello degli Argomenti

	w 1	w 2	w j	w m
doc 1	f_{11}	f_{12}	…	f_{1m}
doc 2	f_{21}	f_{22}	…	f_{2m}
doc 3	…	…	…	…
doc i	f_{i1}	f_{i2}	…	f_{im}
…	…	…	…	…
doc n	f_{n1}	f_{n2}	…	f_{nm}

(documenti , parole)

=

	t 1	t 2	t j	w k
doc 1	f_{11}	f_{12}	…	f_{1k}
doc 2	f_{21}	f_{22}	…	f_{2k}
doc 3	…	…	…	…
doc i	f_{i1}	f_{i2}	…	f_{ik}
…	…	…	…	…
doc n	f_{n1}	f_{n2}	…	f_{nk}

(documenti , topic)

x

	w 1	w 2	w j	w m
t 1	f_{11}	f_{12}	…	f_{1m}
t 2	f_{21}	f_{22}	…	f_{2m}
t .	…	…	…	…
t k	f_{k1}	f_{k2}	…	f_{km}

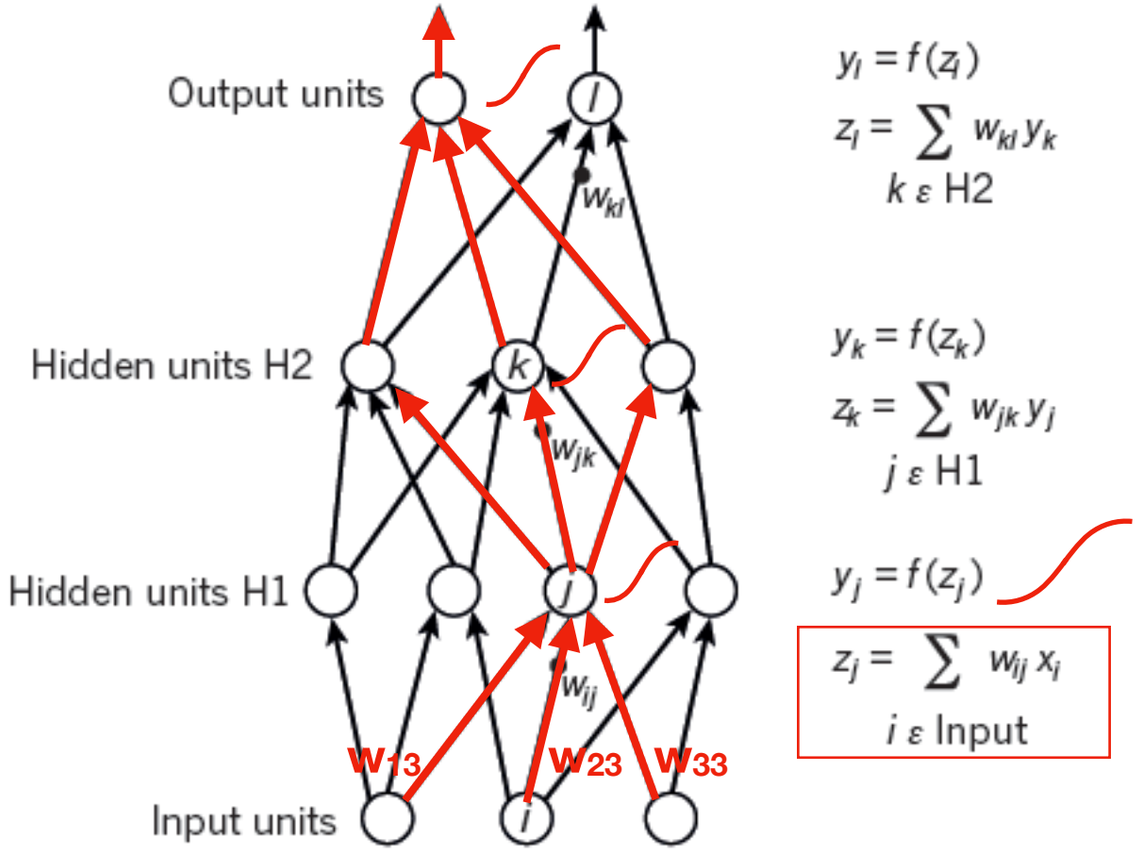
(topic, parole)

descrive i documenti in termini di argomenti

descrive gli argomenti in termini di parole

Le reti neurali artificiali

- Le reti neurali artificiali sono uno dei primi modelli dell'Intelligenza Artificiale del funzionamento del cervello
- sono costituite da neuroni artificiali collegati tramite archi che rappresentano le sinapsi



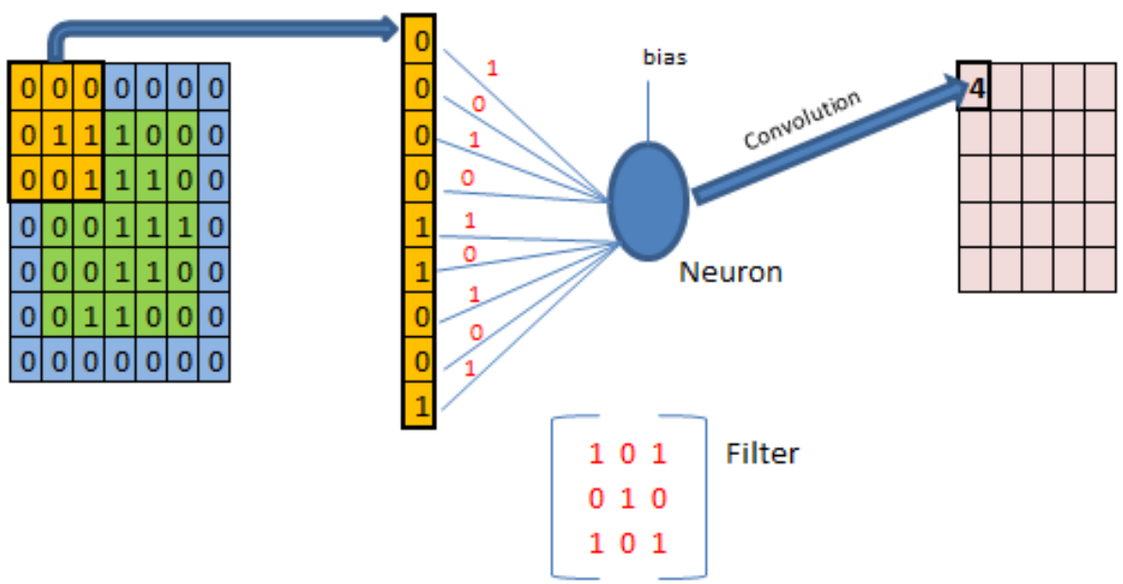
Il Deep Learning

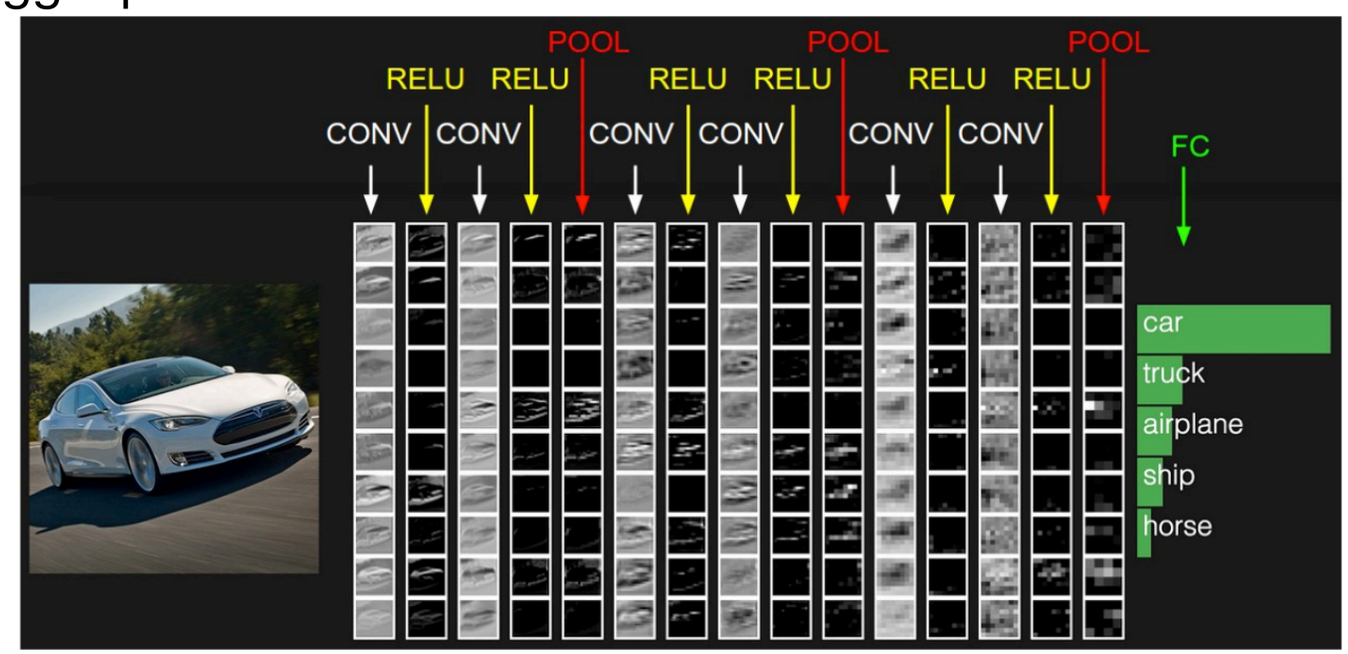
- E’ l’applicazione delle reti neurali artificiali molto profonde (dalle decine alle centinaia di livelli). Per efficienza sono anche implementate con algoritmi paralleli su GPU. (*)

- Le reti convoluzionali (CNN) hanno ottenuto successo in competizioni di apprendimento (*ImageNet* dal 2012 (**)) su:

    - riconoscimento di oggetti in immagini
    - processamento di video, audio, linguaggio parlato

ApprendimentoAutomatico.it

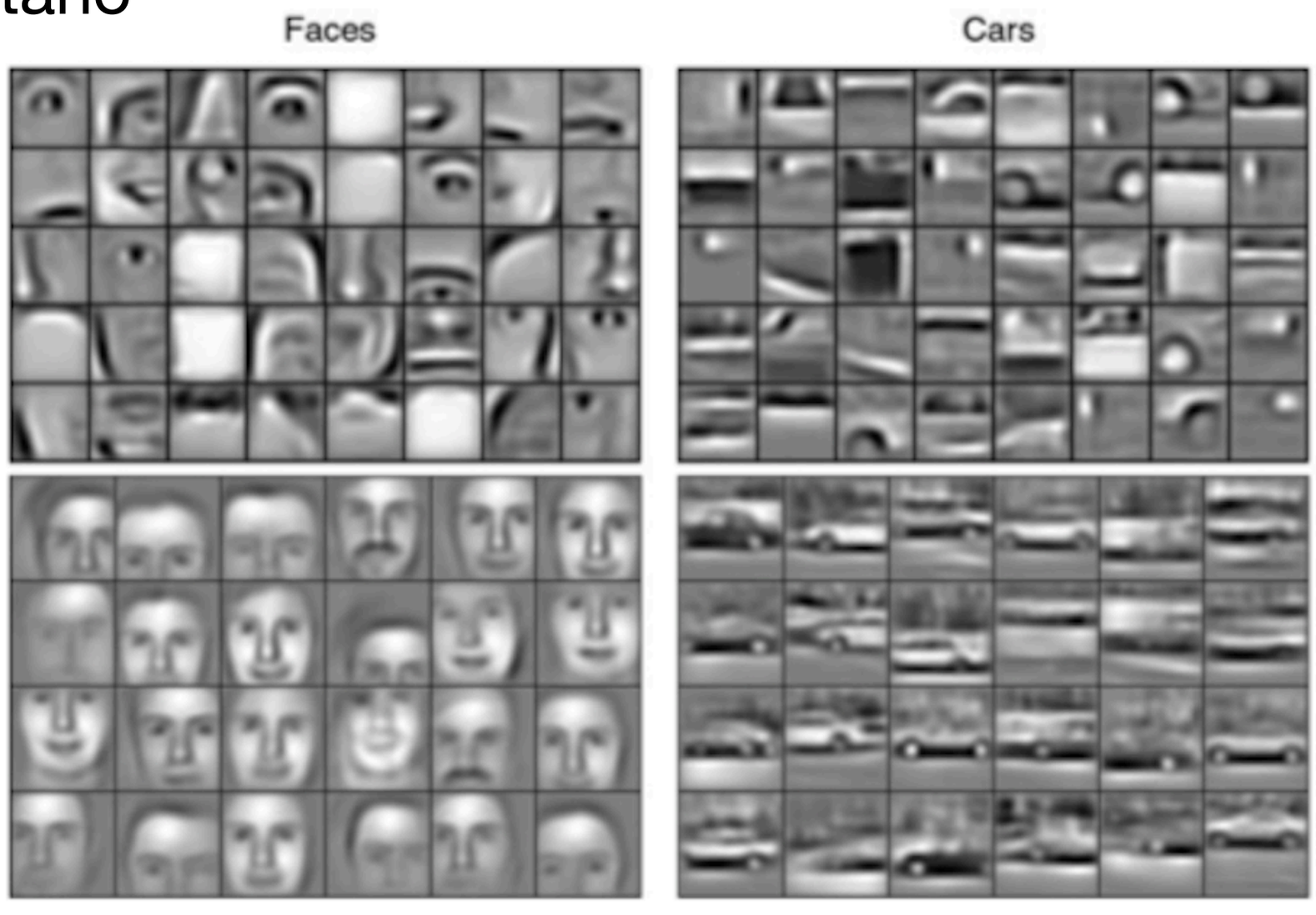




(*) *Deep learning,* di Yann LeCun, Yoshua Bengio & Geoffrey Hinton, in *Nature,* vol. 521, Maggio 2015
(**) 5247 concetti (synsets da WordNet), 500-1000 immagini ciascuna, 3.2 milioni di immagini totali

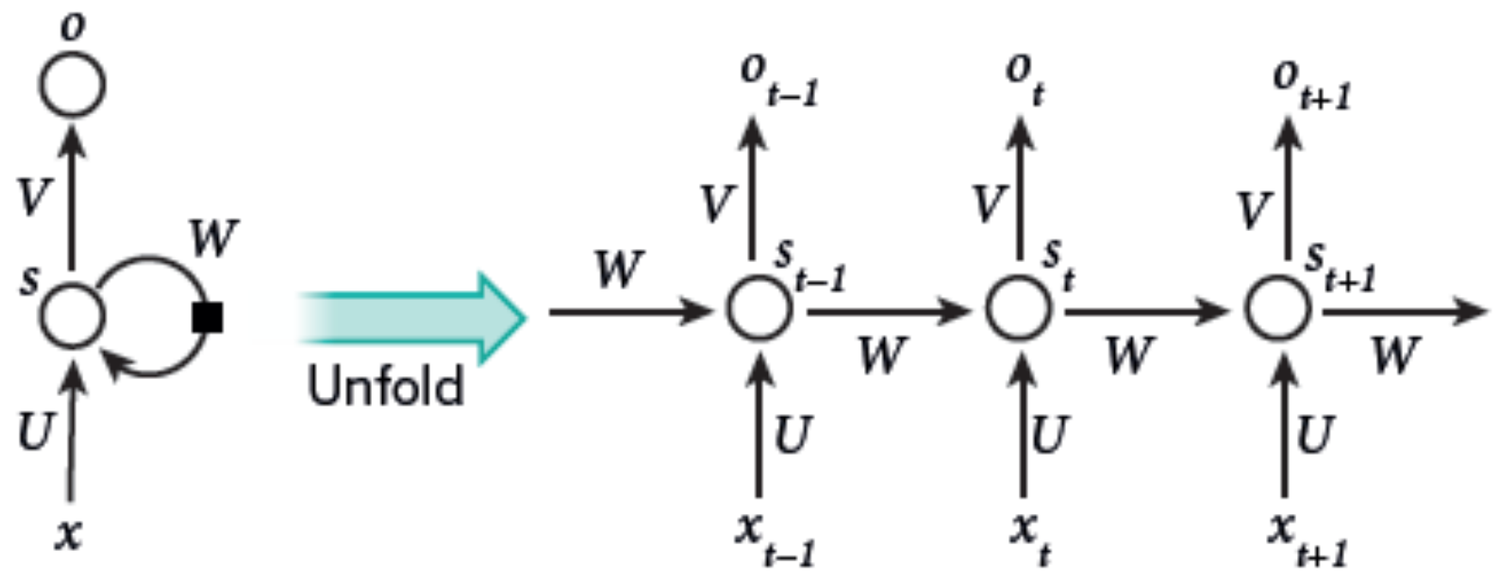
Astrazione

- La sequenza di livelli costruisce delle caratteristiche astratte e composite che servono per il riconoscimento degli oggetti nei livelli finali, in maniera indipendente dal contesto e dalle modalità nel quale gli oggetti si presentano



Le reti ricorrenti (RNN)

- Le reti ricorrenti hanno un arco di retroazione che permette di apprendere le sequenze senza limiti di lunghezza della sequenza



- Hanno ottenuto successo nel riconoscimento di dati sequenziali come linguaggio parlato e testuale

Alcuni risultati sorprendenti

- Immagini associate automaticamente alla loro descrizione da una rete ricorrente (generazione dei testi) associata a una rete convoluzionale (riconoscimento oggetti in immagini)

A woman is throwing a **frisbee** in a park.

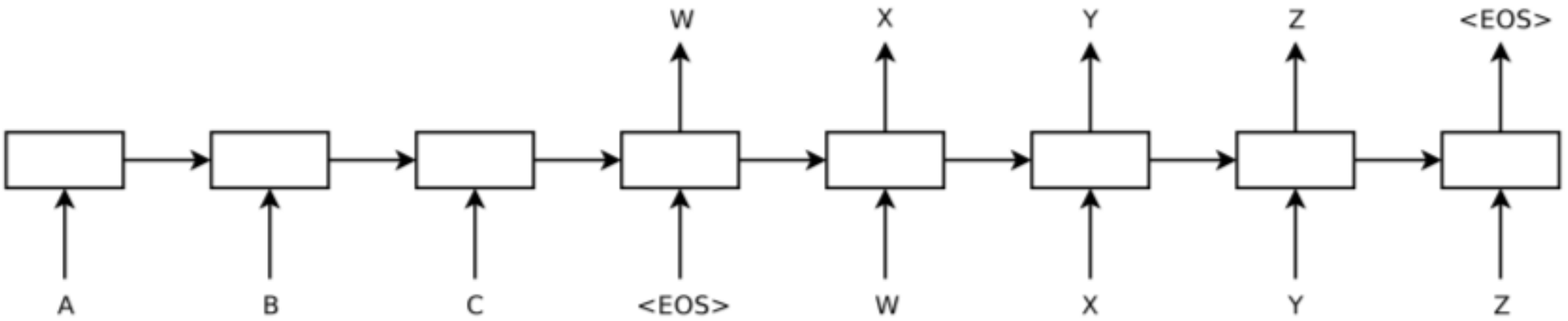
A **stop** sign is on a road with a mountain in the background

Vinyals, O., Toshev, A., Bengio, S. & Erhan, D. Show and tell: a neural image caption generator. In *Proc. International Conference on Machine Learning* http:// arxiv.org/abs/1502.03044 (2014).

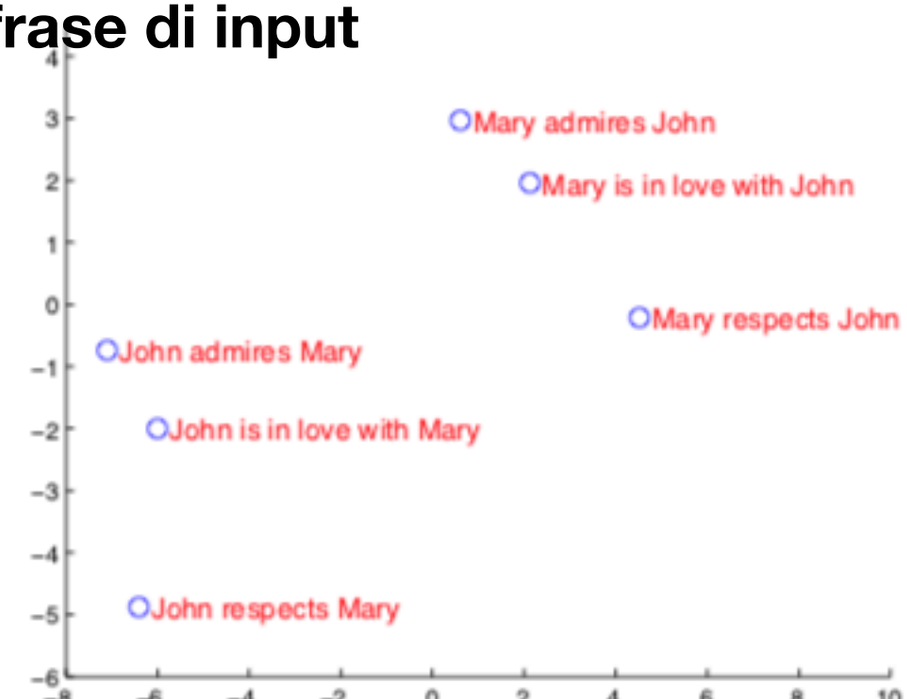
Traduzione automatica

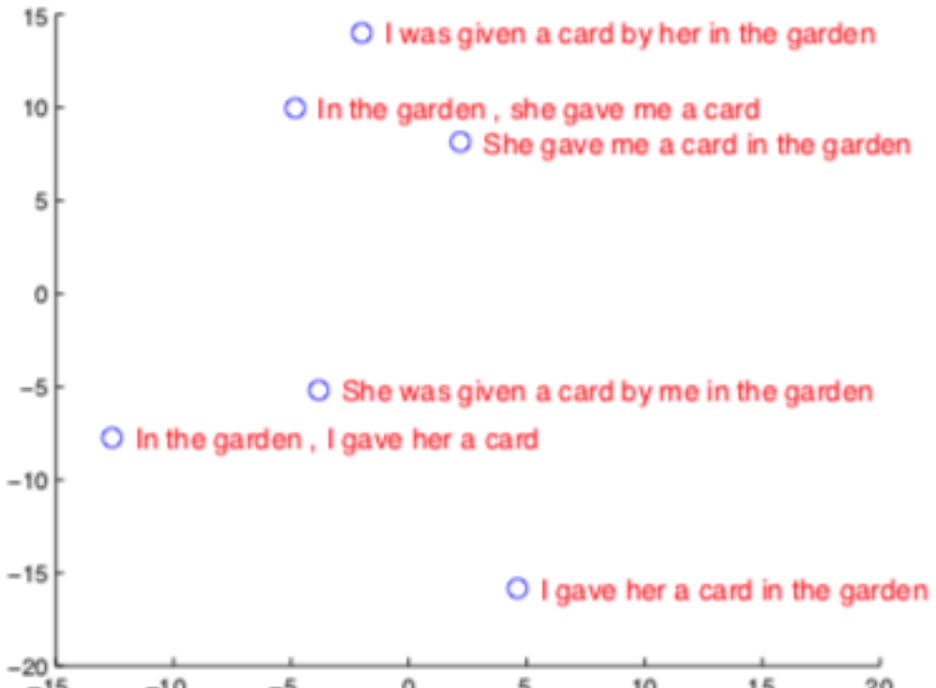
- Una versione di una deep neural network (Long Short-Term Memory model)

sequenza di parole della frase di output



sequenza di parole della frase di input





Gli open data

- La comunità del Semantic Web e dei Linked Open Data ha creato una base di conoscenza condivisa che rappresenta il vocabolario di molti domini e le relazioni semantiche di molti concetti del nostro sapere

- schema.org
  - Creative works: CreativeWork, Book, Movie, MusicRecording, Recipe, TVSeries ...
  - Embedded non-text objects: AudioObject, ImageObject, VideoObject
  - Event
  - Organization
  - Person
  - Place, LocalBusiness, Restaurant ...
  - Product, Offer, AggregateOffer
  - Review, AggregateRating

Il progetto NELL

- NELL (Never Ending Language Learning) è un progetto di Tom Mitchell (CMU) su agenti intelligenti
- Da Gennaio 2010, 24x7, gli agenti apprendono concetti semantici dal WEB e li integrano nella base di conoscenza:
  - il metodo è un ibrido con molte tecniche esistenti
  - 50 milioni di *beliefs* corredati da livelli di confidenza
  - http://rtw.ml.cmu.edu/rtw/kbbrowser/
- alcune relazioni vengono inferite dal sistema, in modo autonomo.

Recently-Learned Facts twitter Refresh

instance	iteration	date learned	confidence
xbox_360_cheats is a video game system	1111	06-jul-2018	98.2
usability_professionals___association is a non-profit organization	1111	06-jul-2018	98.9
desert_patchnose_snake is a reptile	1111	06-jul-2018	100.0
collins_family_markets is a retail store	1111	06-jul-2018	99.6
privet_thrips is an invertebrate	1111	06-jul-2018	100.0
jobs is a person who moved to the state kansas	1115	03-sep-2018	99.9

Open source

- Ci sono molte librerie di open software di Apprendimento Automatico che si possono utilizzare:

- http://data.dmlc.ml (MxNet di Apache) si possono scaricare reti CNN già addestrate su ImageNet

- Si veda la piattaforma: kdnuggets.com

- UC Irvine ha un archivio di datasets reali e pubblici

- Kaggle è una piattaforma per condividere problemi e sfide su datasets reali

  - Theano in Python

  - Pytorch

  - Ad esempio, a partire dai modelli di traduttori di frasi si può arrivare a modelli di chat-bots che conversano generando la frase successiva del dialogo date le precedenti frasi del dialogo

Problemi e criticità

- **Privacy** delle persone:
  i dati sono il nuovo petrolio;
  le analisi dei dati “consumano” la privacy delle persone

  - Sono stati realizzati varianti dei metodi di apprendimento che preservano la privacy:
    si aggiunge *un po’ di rumore* nei punti giusti del modello;
    non si pregiudica la sua utilità ma ostacola l’identificazione delle persone

- **Discriminazione:**
  alcuni modelli replicano i pregiudizi delle persone e le disuguaglianze sociali presenti nei dati

- **Etica:**
  le eccellenti performance nel riconoscimento in immagini, video e parlato, hanno permesso di costruire veicoli autonomi. In caso di incidente, occorre decidere chi salvare:
  pedone o passeggero?

Conclusioni

- E’ un momento entusiasmante per l’apprendimento automatico
- Si sono ottenuti molti risultati sorprendenti, specialmente con le reti neurali profonde, nel campo del riconoscimento delle immagini e del linguaggio
- Avanzamenti che potranno portare innovazione alle imprese, ad esempio:
  - riconoscimento di situazioni anomale,
  - allerta per rischio,
  - ottimizzazione del processo produttivo
  - riduzione dei prodotti difettosi,
  - manutenzione predittiva,
  - riduzione dei costi