



Preparazione dati



Preparazione dati

Premessa

Ogni architettura, modello, rete neurale...nasce per svolgere un compito, per risolvere un problema...è naturale quindi chiedersi:

Cosa voglio ottenere?

Vorrei riconoscere l'erba cattiva dal giardino ed eliminarla...

Vorrei sapere in anticipo se il motore della macchina sta per rompersi...



Preparazione dati

Premessa

Dopo avere chiaro cosa si vuole ottenere, è necessario identificare il tipo di problema che si dovrà affrontare:

Che tipo di problema è?

Problema di regressione, problema di classificazione...

Object detection, instance segmentation...



Preparazione dati

Premessa

I **dati** saranno il nostro punto di partenza ma, per partire, bisognerà conoscere il tipo e il formato di quelli con cui avremo a che fare:

Di che dati necessito per risolverlo?

Immagini, testo, segnali, dati tabellari, serie temporali...

Classi, valori di regressione, segmentazioni, immagini, segnali...



Preparazione dati

Premessa

Dopo aver chiaro l'obiettivo, il tipo di problema e i dati che ci servono per risolverlo, l'ultima domanda da porsi è:

Come lo risolvo?

Linear regression, SVM, k-nearest neighbor...

Artificial Neural Network, Convolutional Neural Network, Generative Adversarial Network...



Preparazione dati

Raccolta

Per prima cosa si identifica quale sarà la fonte:



Audio



Dispositivi di misura



Raccolte di immagini



File csv

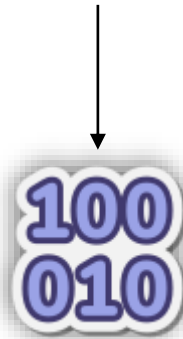


Video



Tabelle dati Excel

I modi in cui i dati si vanno a distribuire, vengono rappresentati o sono raccolti sono svariati ma a livello software vale un principio fondamentale.



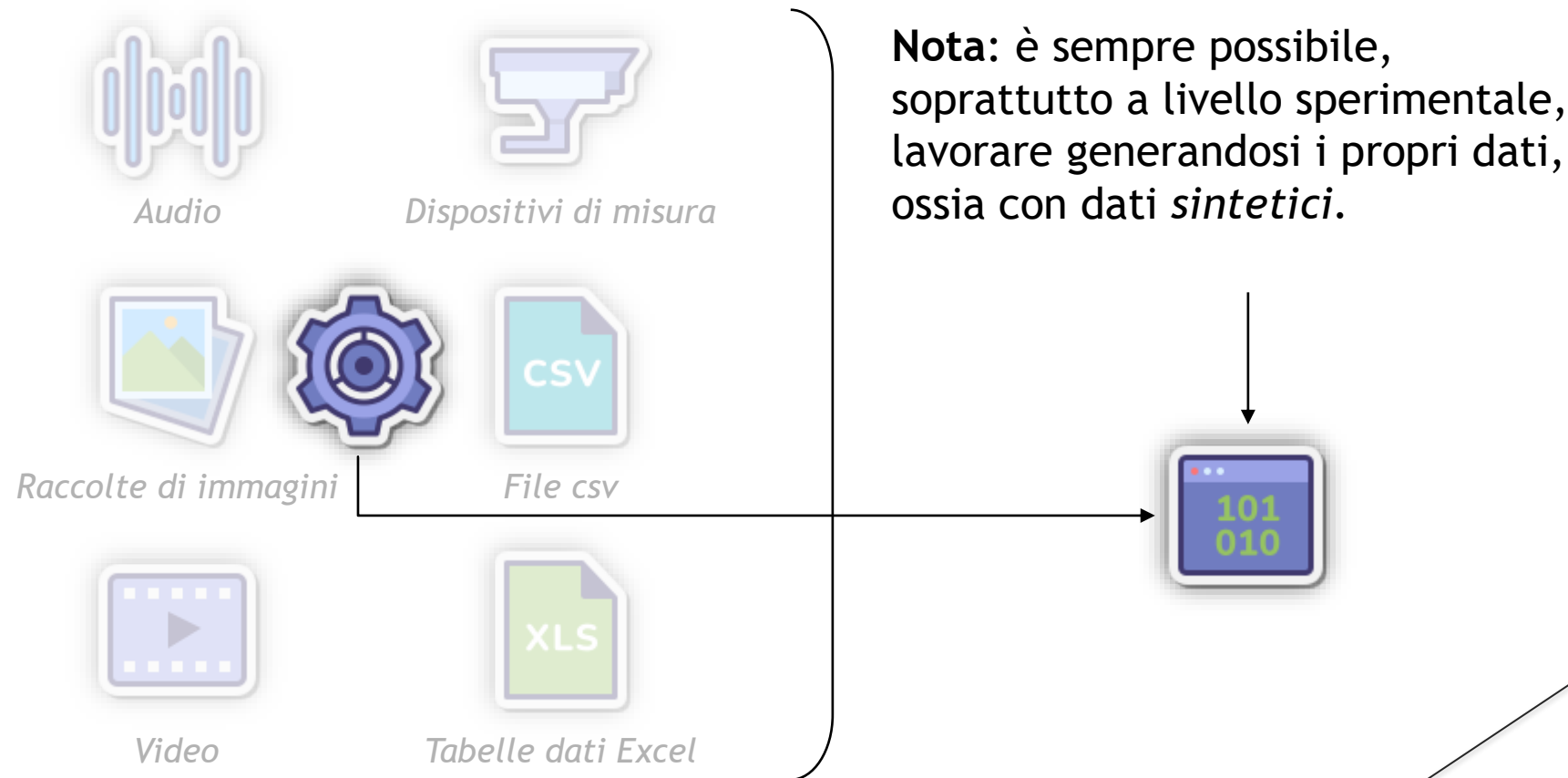
Tutti sono numeri.



Preparazione dati

Raccolta

Per prima cosa si identifica quale sarà la fonte:

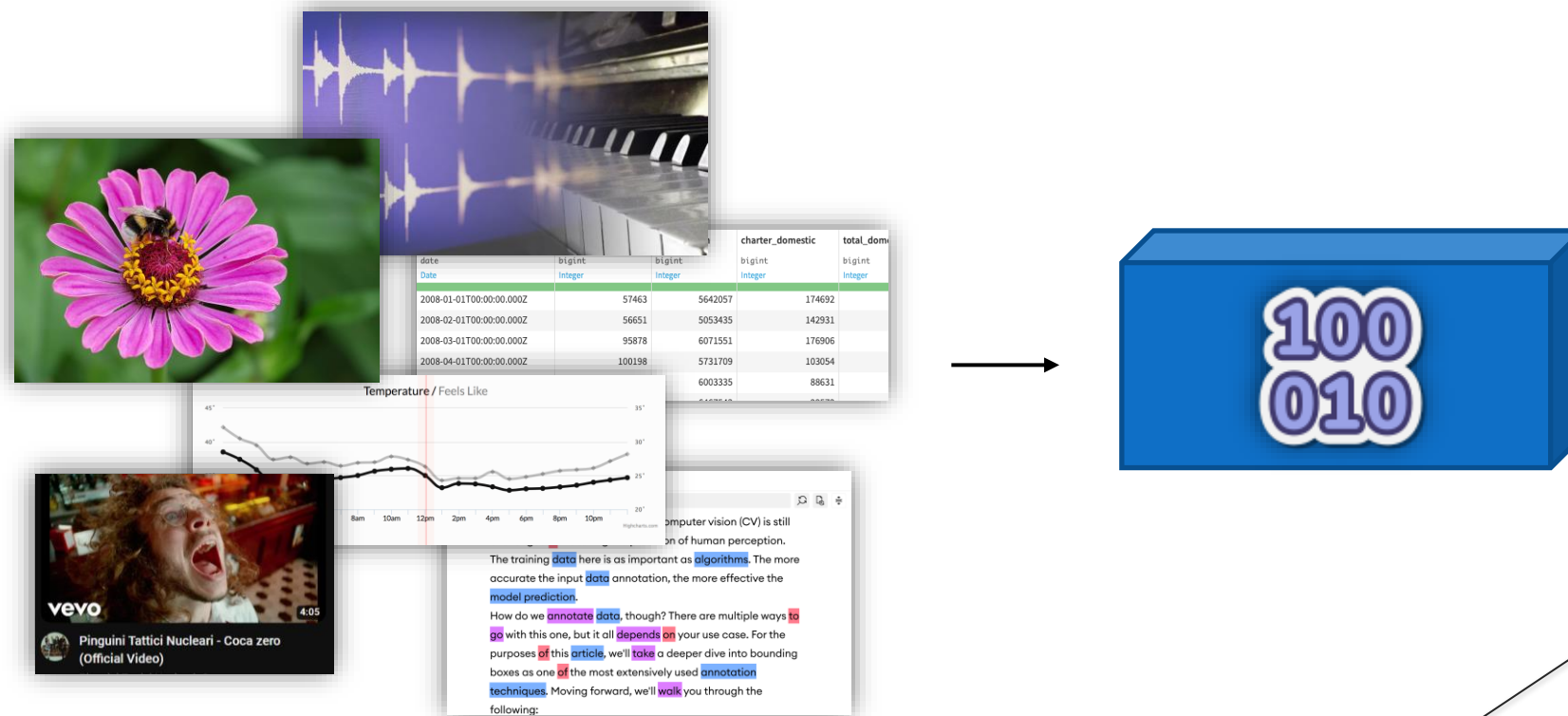




Preparazione dati

Da input ad output: numerical encoding

Immagini, audio, video, testo...sono codificabili come **numeri**: tensori.

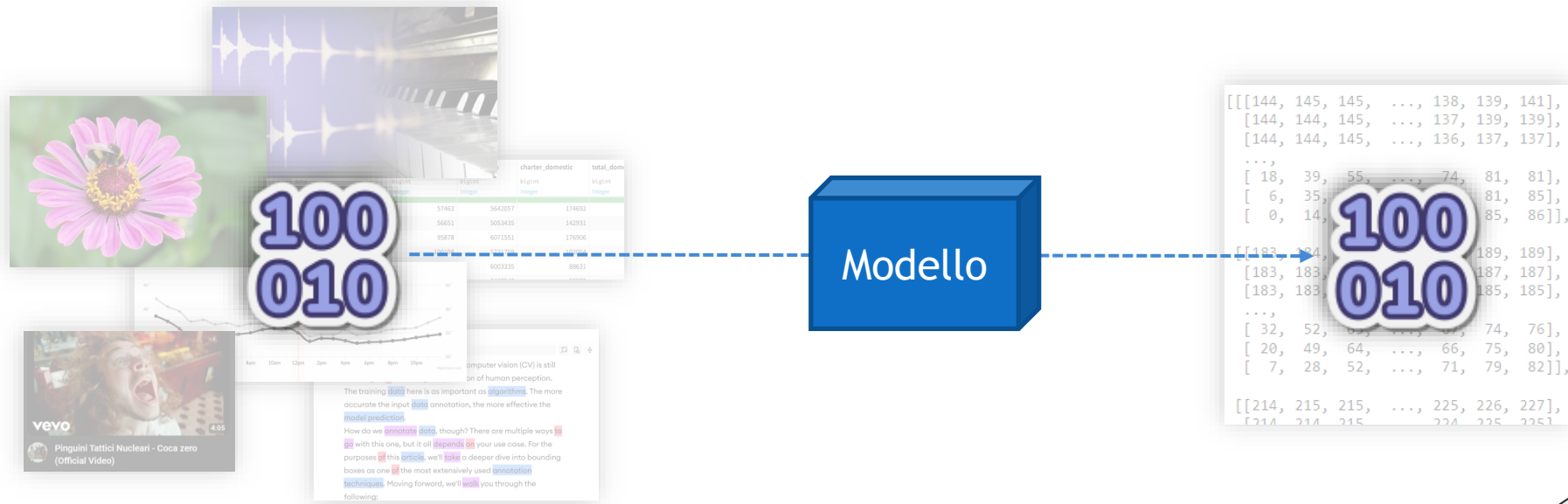




Preparazione dati

Da input ad output: passaggio nel modello

Ciò che entra in un modello è **numero**, un tensore. Ciò che esce da un modello è **numero**, un tensore.

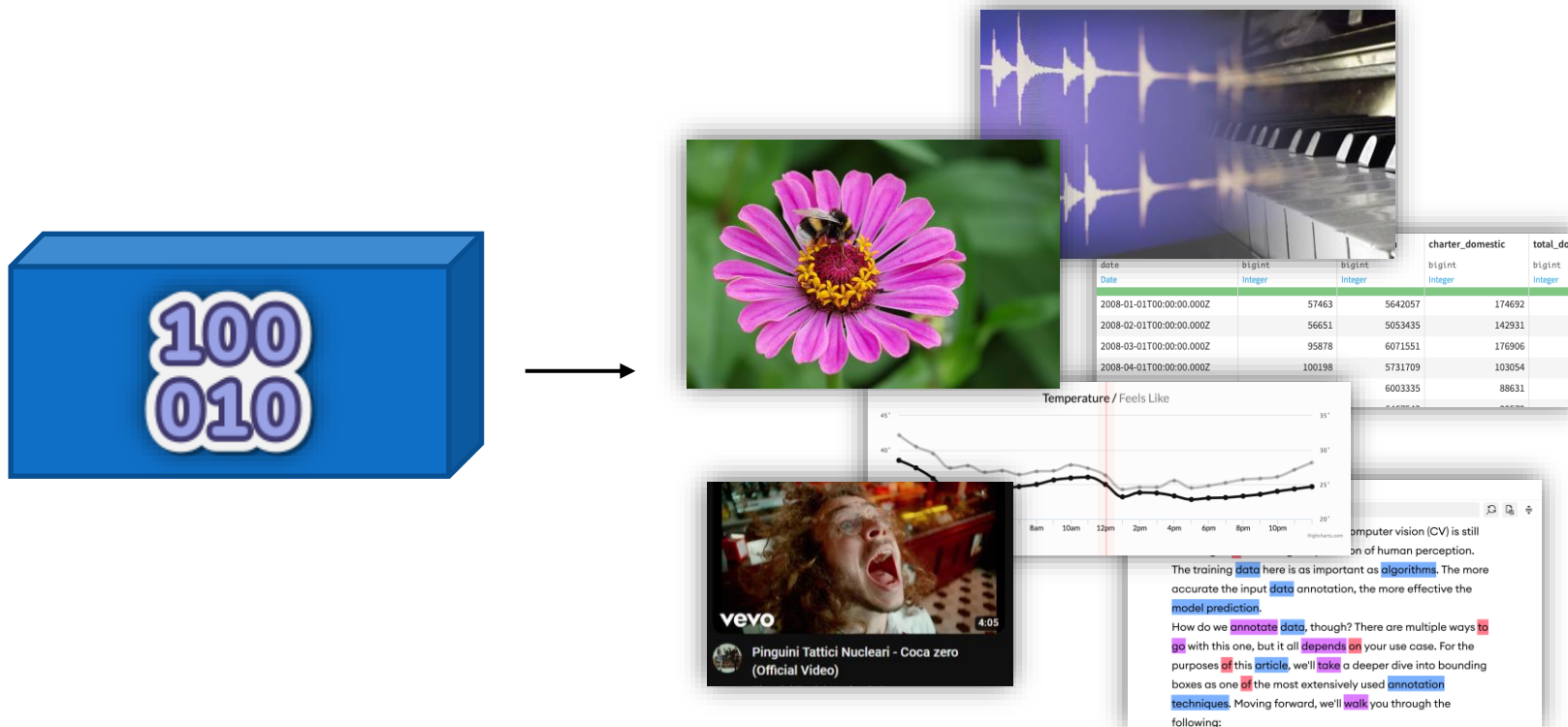




Preparazione dati

Da input ad output: numerical decoding

L'output di una rete, infine, può tornare ad avere una rappresentazione.
...Tornare ad essere un **dato**.





Preparazione dati

Manipolazione

Ottenuti i dati, da fonti sintetiche o no, si avrà a che fare con **tensori**. Rari però sono i casi in cui, questi ultimi sono pronti all'uso:

- ▶ Pre-processamenti.
- ▶ Trasformazioni.
- ▶ Normalizzazioni.

Questi, potenzialmente, sono alcuni degli step da applicare prima di ottenere un dato pulito e adeguato al passaggio in una rete neurale.



Preparazione dati

Manipolazione: esempi

Alcuni step, ad esempio, divisi per caso d'uso:

Classificazione immagini.

- Resize a dimensione fissa.
- Normalizzazione/standardizzazione.
- Data augmentation.
- ...

Processamento del linguaggio.

- Da testo a parole e sotto-parole.
- Creazione di frasi di lunghezza fissa.
- Creazione di un vocabolario.
- ...

Analisi di serie temporali.

- Normalizzazione/standardizzazione.
- Rimozione del rumore.
- Sequenze dello stesso periodo.
- ...

Localizzazione di oggetti.

- Data augmentation.
- Resize.
- Creazione di box associate agli oggetti.
- ...

Vediamone degli esempi.



Preparazione dati

Manipolazione: normalizzazione e standardizzazione

Normalizzazione:

Riscalda i valori in un range [min, max], generalmente [0, 1].

$$x_{normal} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Non è nota la distribuzione dati e possiedono un'ampia scala valori. Riduce i valori in un range molto più piccolo, avvicina i campioni e rende più regolare e robusto l'addestramento delle reti neurali.

Standardizzazione:

Riscalda i valori per ottenere media 0 e deviazione standard 1.

$$x_{standard} = \frac{x - \mu}{\sigma}$$

I dati sono normalmente distribuiti o contengono outliers. Riduce l'impatto degli outliers e allo stesso tempo normalizza i dati. Aiuta l'addestramento di reti neurali facilitando la convergenza.



Preparazione dati

Manipolazione: bilanciamento campioni

Nei problemi di machine learning e deep learning, in particolare quelli di classificazione, è spesso importante avere una raccolta dati bilanciata dove il numero di campioni fra le classi è quindi uguale o molto simile.

Le principali tecniche di bilanciamento dati sono quattro:

► *Undersampling.*

► *Oversampling.*

► *Classweight.*

► *Spostamento della soglia.*

Campioni sbilanciati



Queste tecniche sono legate prevalentemente alla rete e non ai dati...



Preparazione dati

Manipolazione: bilanciamento campioni

L' **undersampling** riduce il numero di campioni di ogni classe fino a quello delle classe *minima*. L' **oversampling** aumento il numero di campioni di ogni classe fino a quello della classe *massima*.

Undersampling



Oversampling





Preparazione dati

Manipolazione: data augmentation

Rappresenta un insieme di tecniche utilizzate per aumentare la quantità di dati a disposizione: applica ai dati già esistenti dei cambiamenti casuali controllati, realizzandone delle nuove versioni. Agevola l'*oversampling*.



Le principali tecniche sono: *flip*, *resize*, *rotazioni*, *cambi di luminosità*, *ritagli*, *ridimensionamenti*...



Preparazione dati

Manipolazione: data augmentation

Note importanti:

- ▶ Crea variabilità nella insieme dei dati, evita che l'addestramento di una rete si focalizzi nell'apprendere dalla raccolta originale, problematica chiamata *overfitting*.
- ▶ L'utilizzo di questa tecnica deve avvenire solo in caso di necessità. Aumentare la quantità di dati quindi, di conseguenza, la conoscenza da apprendere, potrebbe rendere difficile l'addestramento portando l'effetto inverso, l'*underfitting*.



Preparazione dati

Dataset

Presi i dati e, manipolatili nella maniera più consona, si ha un **dataset**.

Un **dataset**, nella sua definizione più semplice, non è nient'altro che la raccolta di campioni che saranno utilizzati per...

- ▶ Addestrare.
- ▶ Validare.
- ▶ Testare.

...un modello di rete neurale.

Ogni **azione** richiederà parte del dataset inizialmente raccolto.



Preparazione dati

Dataset: separazione dei campioni

È buona norma, per ottenere solidi risultati nello sviluppo di una rete neurale, sfruttare ognuna delle tre parti del dataset durante la fase di creazione del modello.

Il numero di campioni per gruppo non è fisso ma, in genere, si seguono questi ordini di grandezza:



Training
60-80%

La rete **vedrà** questi campioni e li **utilizzerà**, in fase di addestramento, **per apprendere** la conoscenza necessaria.



Validation
10-20%

La rete **utilizzerà** questi campioni, in fase di addestramento, per **testare** la conoscenza mano a mano appresa.



Test
10-20%

La rete **utilizzerà** questi campioni, in fase di test, per **testare** la conoscenza raggiunta al termine dell'addestramento.



Preparazione dati

Dataset: i gruppi



Training
60-80%

- ▶ Costituito da campioni usati per l'addestramento della rete.
- ▶ Questi campioni possono venir trattati, bilanciati, aumentati.
- ▶ Deve essere sufficientemente ampio, perché il modello possa avere dati sufficienti ad apprendere.
- ▶ Deve essere vario per permette che l'addestramento sia meno specializzato sui dati di training, generalizzando la conoscenza appresa.



Validation
10-20%

- ▶ Costituito da campioni mai visti in addestramento.
- ▶ I campioni sono usati per la verifica di quello che la rete ha appreso in un particolare fase di quest'ultimo.
- ▶ Non necessita di essere bilanciato o aumentato.



Test
10-20%

- ▶ Costituito da campioni mai visti in addestramento.
- ▶ I campioni sono usati per testare quello che la rete ha imparato al termine dell'addestramento.
- ▶ Non necessita di essere aumentato o bilanciato.



Preparazione dati

Dataset: campioni ed etichette

Esistono diversi modi di addestrare una rete neurale ma, la principale distinzione, che impatta anche il modo in cui i campioni saranno rappresentati è fra: addestramento **supervisionato** e **non supervisionato**.

Semplicemente, la differenza fra i due è la presenza o meno di campioni **etichettati**.

Apprendimento supervisionato:

La rete impara a raggiungere il proprio obiettivo ricevendo campioni i quali possiederanno, ognuno, una propria etichetta ad indicare l'output atteso come risposta dalla rete.

Apprendimento non supervisionato:

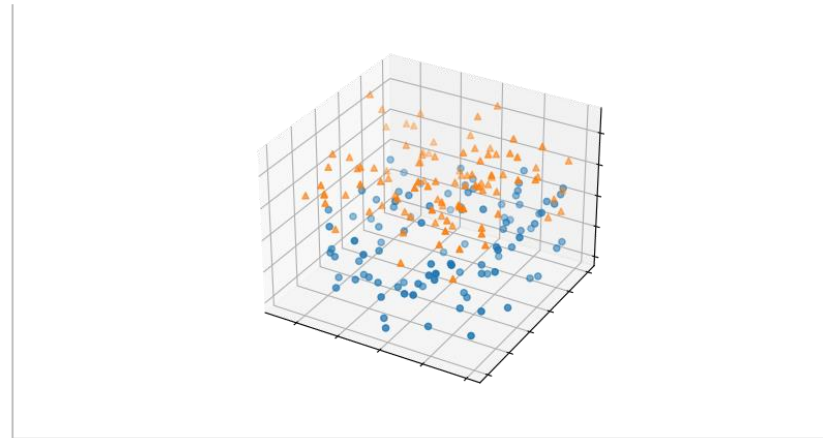
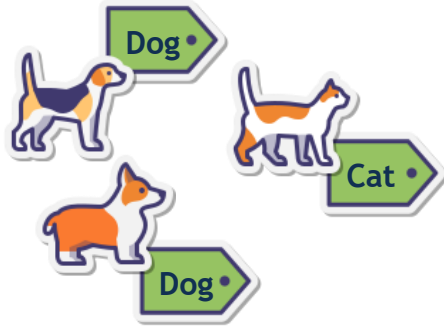
I campioni sono a sé, non hanno una etichetta a descriverli e spetta alla rete scoprire se qualcosa fra loro li distingue.



Preparazione dati

Dataset: campioni ed etichette

Nei dataset etichettati si distinguerà, in genere, fra **data** e **labels**, oppure, **x** (i dati) e **y** (le etichette).



x_n {  , iris }







x_{n+1} {  , girasole }

x_n { {capelli bianchi, artrite...}, 90 }

x_{n+1} { {capelli biondi, acne... }, 18 }

Preparazione dati

Dataset: dove trovarli

-  Google Dataset Search Beta
-  **FiveThirtyEight**
-  DATA.GOV
-  kaggle
-  ucimlr
-  Papers With Code



Proviamo?

