# Riduzione della dimensionalità

# Riduzione della dimensionalità

## Premessa

Per **dimensionalità** dei dati si intende il numero di variabili o features che descrivono ogni punto o campione di un dataset.

Indica, cioè:

*La quantità di informazioni che vengono prese in considerazione per rappresentare ciascun campione.*

Un dataset di dimensionalità cinque, ad esempio:

| Modello | Anno immatricolazione | Costo | Città | Numero incidenti | Numero porte |
|---|---|---|---|---|---|
| A | 2013 | 5000 | FC | 0 | 3 |
| B | 2014 | 8000 | RA | 3 | 5 |
| C | 2020 | 10000 | BO | 0 | 3 |
| D | 2019 | 5000 | RM | 2 | 3 |

# Riduzione della dimensionalità

## Premessa

La gestione adeguata della dimensionalità è un aspetto importante nel processo di analisi dei dati e nella costruzione di modelli di machine learning efficaci:

- *Alta dimensionalità può portare a **overfitting**, ridondanza delle informazioni…*
- *Bassa dimensionalità può portare a una rappresentazione insufficiente dei dati…*

**Si pensi, ad esempio, ci venga chiesto di replicare un quadro:**

Replicare la gioconda con solo una matita non permetterà di descrivere ogni sfumatura, ogni tratto, ogni colore dell'opera.

*Si è di fronte ad un limite intrinseco nella quantità di informazioni che si potrà rappresentare.*

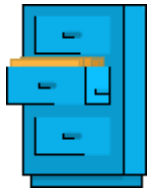
# Riduzione della dimensionalità

## Premessa

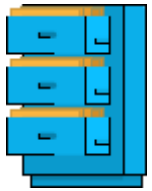
La gestione adeguata della dimensionalità è un aspetto importante nel processo di analisi dei dati e nella costruzione di modelli di machine learning efficaci:

- *Alta dimensionalità può portare a **overfitting**, ridondanza delle informazioni…*
- *Bassa dimensionalità può portare a una rappresentazione insufficiente dei dati…*

**Si pensi, ad esempio, ci venga chiesto di trovare un documento:**



*Conosciamo scaffale e cassetto. Rimane da esplorare una sola dimensione: il cassetto in profondità.*



*Conosciamo lo scaffale. Si dovrà esplorare ogni cassetto, dall'alto in basso, in profondità.*

*Non conosciamo nemmeno lo scaffale. Si dovrà esplorare tre dimensioni: ogni cassetto di ogni scaffale.*

# Riduzione della dimensionalità

## Premessa

Trattando l'ambito del machine learning e del deep learning si ha modo di affrontare entrambi questi problemi.

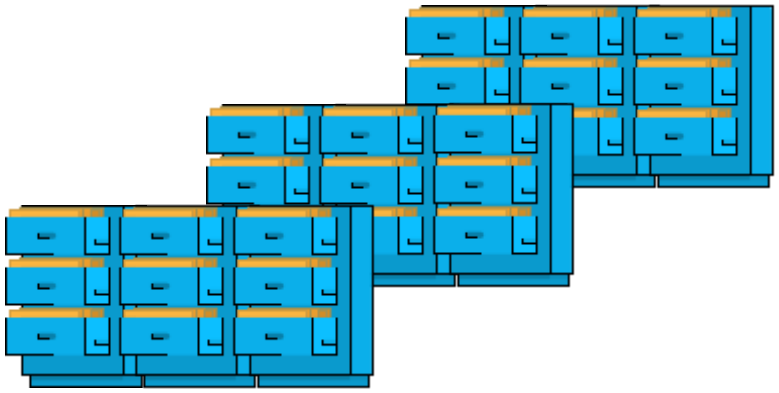
Vale la pena però porre una maggiore attenzione all'importanza di **condensare le informazioni**, **la conoscenza appresa** e apprendere come distillare le sole informazioni di cui un problema necessita per essere affrontato.

In breve, come affrontare correttamente la riduzione della dimensionalità.

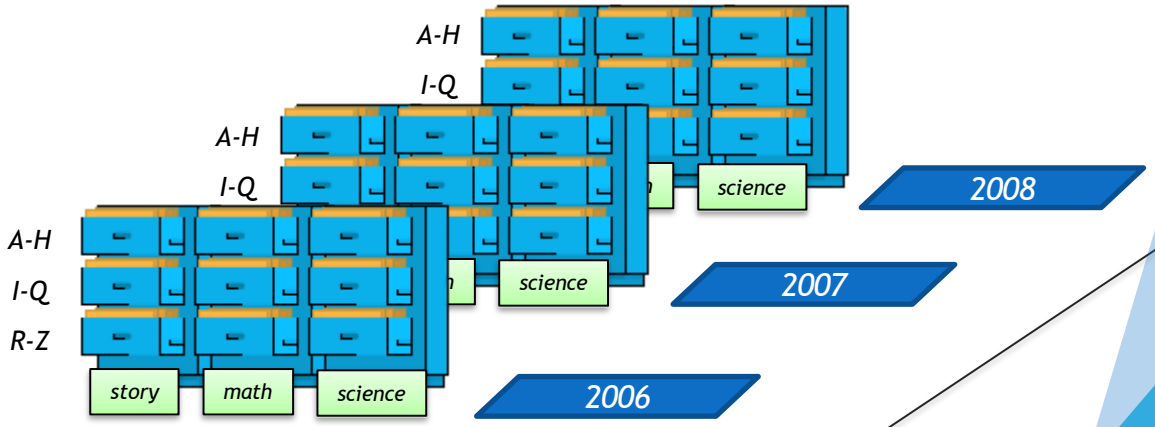
# Riduzione della dimensionalità

## Premessa

Ritornando al precedente esempio, ad una rete neurale sarebbe richiesto il compito di trovare quel condensato di informazioni, ridotte, che semplifica la gestione dei dati…senza andarvene a precludere il significato.



L'accesso ad un documento, prima complesso, può essere semplificato con sole tre informazioni: anno, categoria, iniziale…



# Riduzione della dimensionalità

## Course of dimensionality

Con il termine **Course of Dimensionality**, ‘maledizione della dimensionalità’, ci si riferisce ad una serie di problemi che si vengono a verificare quando la dimensionalità dei dati è troppo alta.

Si identificano fra questi:

1. *Sparsità dei dati.*
2. *Crescita del numero di parametri.*
3. *Ridondanza delle variabili.*
4. *Complessità del modello.*

# Riduzione della dimensionalità

## Course of dimensionality

Spesso, volendo affrontare uno specifico task/problema, ci si rende conto che il numero di informazioni legate ai dati è immotivatamente alto. Alcune informazioni sono effettivamente non necessarie.

*La riduzione, in questo caso, può avvenire direttamente a livello di campioni.*

| Modello | Anno immatricolazione | Costo | Città | Numero incidenti | Numero porte |
|---|---|---|---|---|---|
| A | 2013 | 5000 | FC | 0 | 3 |
| B | 2014 | 8000 | RA | 3 | 5 |
| C | 2020 | 10000 | BO | 0 | 3 |
| D | 2019 | 5000 | RM | 2 | 3 |

| Modello | Anno immatricolazione | Città | Numero incidenti |
|---|---|---|---|
| A | 2013 | FC | 0 |
| B | 2014 | RA | 3 |
| C | 2020 | BO | 0 |
| D | 2019 | RM | 2 |

# Riduzione della dimensionalità

## Vantaggi

Ridurre la dimensionalità significa:

*Estrarre la quantità minima di informazioni che permettono di risolvere un problema trascurando tutto il resto.*

Da un punto di vista pratico questo porta a:

- *Ridurre i tempi di calcolo.*
- *Evitare che un modello si concentri erroneamente su dati che non sono importanti.*
- *Migliorare le prestazioni.*
- *Agevolare il raggiungimento della convergenza.*

# Riduzione della dimensionalità

## Variabili latenti

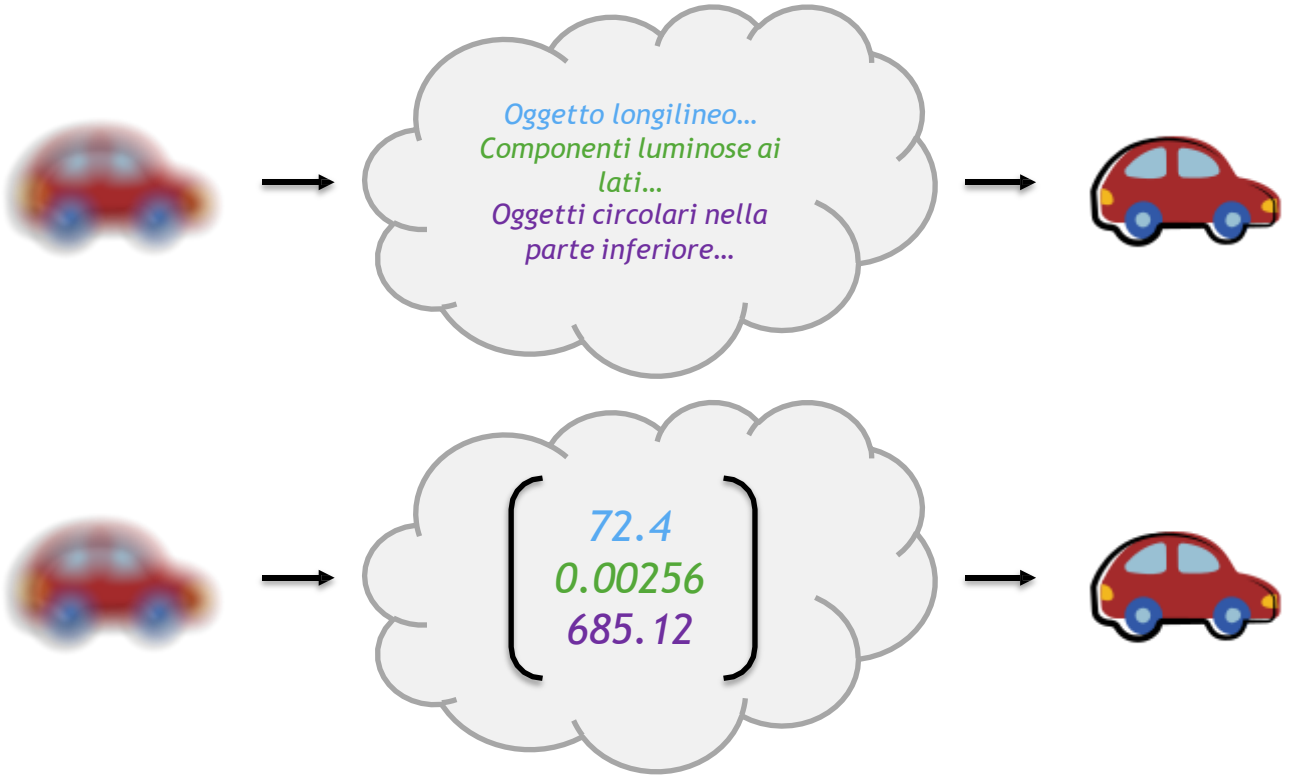
Per *variabile latente*, si intende una variabile 'nascosta', non direttamente osservabile nei dati di input, ma utilizzata per rappresentare o spiegare determinati fenomeni/caratteristiche che portano l'input all'output.

*Queste variabili catturano informazioni complesse, pattern, relazioni presenti nei dati e nella variabili osservabili.*

# Riduzione della dimensionalità

## Variabili latenti

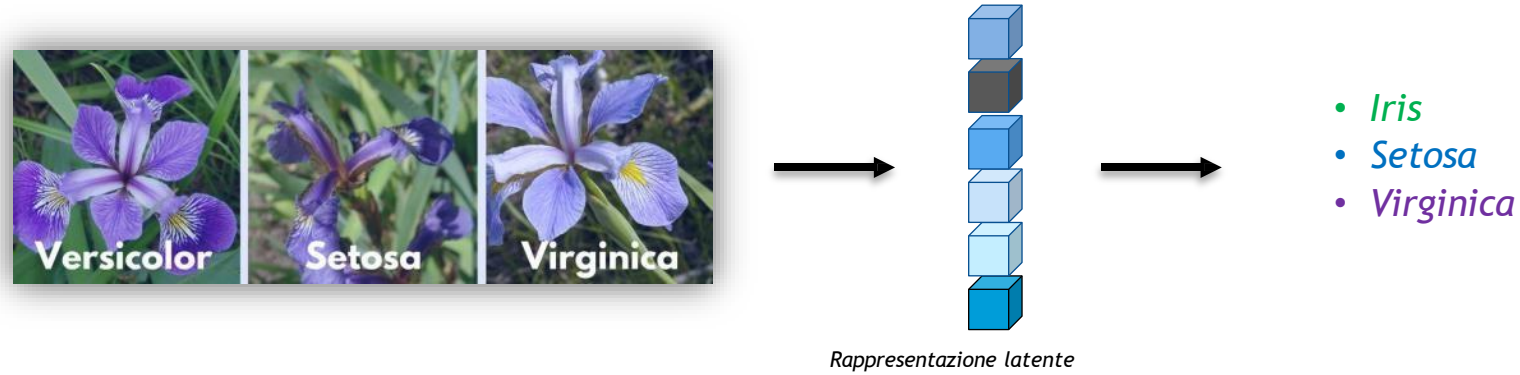
In termini pratici, la rappresentazione latente di una codifica di informazioni, non sarà nient'altro che un vettore di numeri, il cui significato è intrinsecamente appreso dalla rete.



# Riduzione della dimensionalità

## Variabili latenti

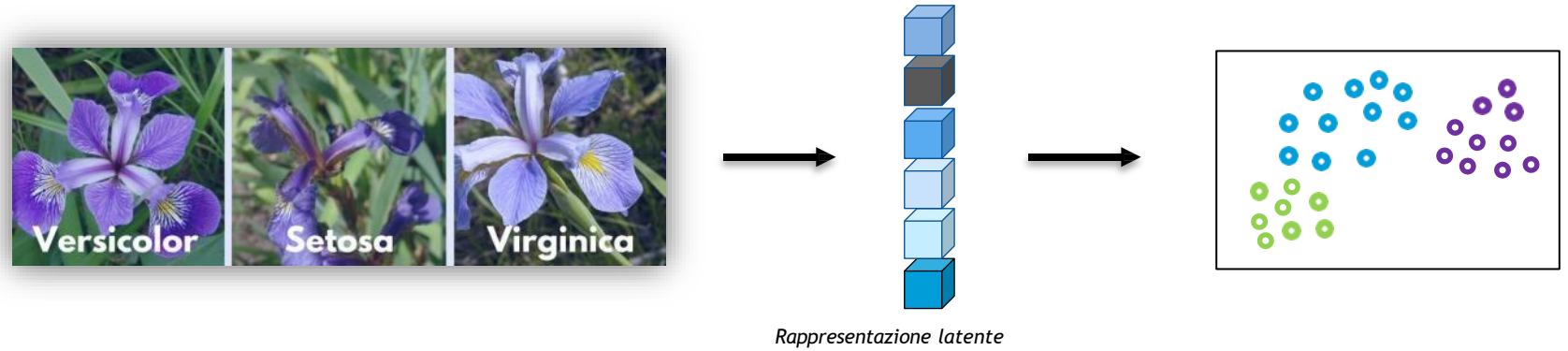
Nell'ambito del **ML** e dell'apprendimento supervisionato le variabili latenti possono essere utilizzate per modellare correlazioni complesse tra le variabili di input e l'output desiderato.



# Riduzione della dimensionalità

## Variabili latenti

Nell'ambito del **ML** e dell'apprendimento **non** supervisionato (*clusterizzazione o riduzione della dimensionalità*) le variabili latenti possono essere utilizzate per estrarre strutture nascoste o pattern nei dati di input.




*Rappresentazione latente*

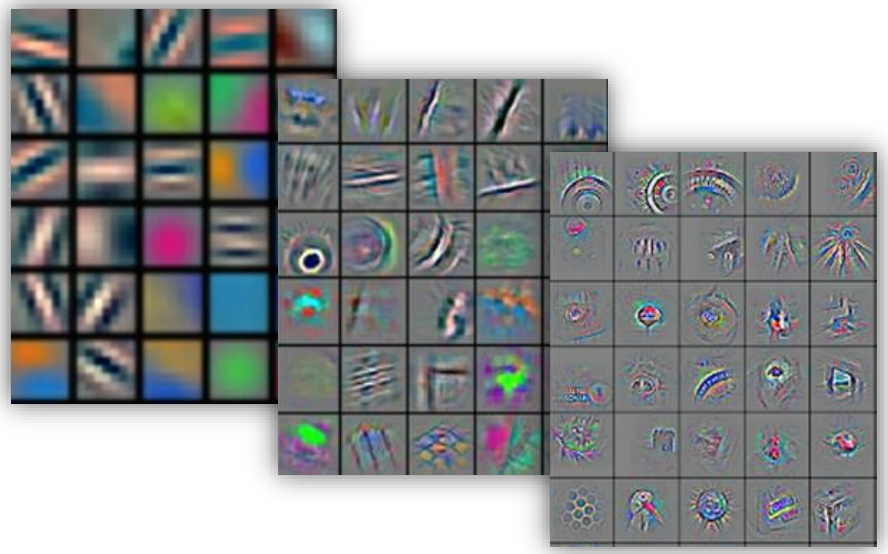
# Riduzione della dimensionalità

## Variabili latenti

Nei modelli di **DL**, nelle reti neurali, le variabili latenti vengono spesso chiamate ***latent features*** o ***latent representation***.

Queste caratteristiche latenti vengono **apprese automaticamente durante il processo di addestramento** e condensano in esse, informazioni rilevanti e significative nei dati di input.

In una **CNN**, ad esempio:



*Al crescere della profondità dei layer, la features latenti apprendono nuove relazioni, combinano pattern precedenti…*

# Riduzione della dimensionalità

## Riferimenti di interesse

Di seguito alcuni riferimenti a visualizzatori e ‘play-ground’ dove vedere l’azione e la generalizzazione delle variabili latenti e dei condensati di informazioni:

- *An Interactive Node-Link Visualization (adamharley.com)*
- *CNN Explainer (poloclub.github.io)*
- *ConvNetJS: Deep Learning in your browser (stanford.edu)*

# Proviamo?