Studio e approfondimento sul metodo computazionale UMAP per la riduzione di dimensionalità e la visualizzazione di dati clinici

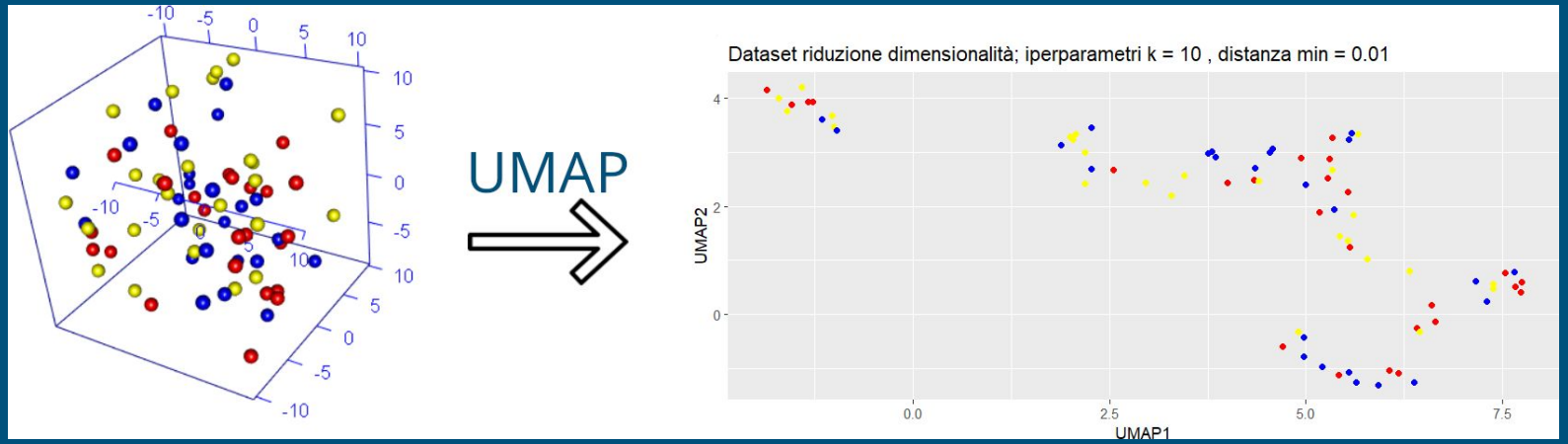
Relatore: Dott. Davide Chicco
Correlatore: Prof. Francesca Gasparini

[Laurea Triennale in Informatica] Giulio Riggio
23/07/2024 844901 g.riggio@campus.unimib.it

Riduzione di dimensionalità

Lo scopo del mio lavoro di tesi è capire il metodo UMAP

- Il problema è testare le capacità, la velocità, capire pregi e difetti del metodo
- Per cercare le soluzioni ho creato datasets artificiali, usato datasets reali, per fare molti diversi test
- Riguardo i risultati, ho compreso caratteristiche interessanti



Revisione letteratura … con Google Scholar

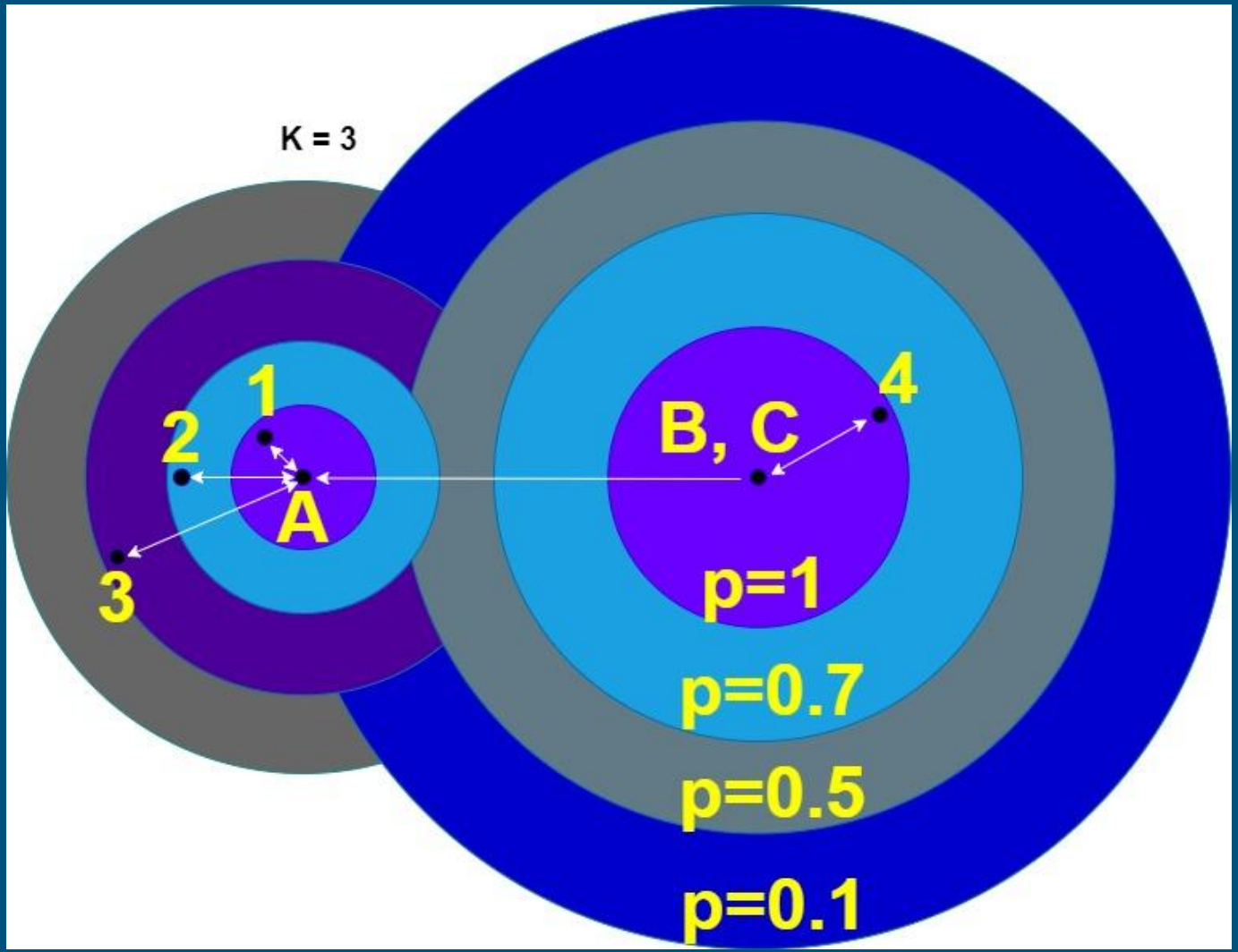
Tipo di dato
- EHRs
- scRNA-seq

Patologia
- Neuroblastoma
- Insufficienze d’organo
- Depressione
- Virus respiratori
- Perdita di gravidanza

Uso UMAP
- Pre Clustering
- Visualizzazione

Confronto con altri metodi
- UMAP è più veloce
- UMAP esegue proiezioni migliori
- Parametrizzazione più chiara per la struttura locale o la struttura globale

Casi particolari
- UMAP può creare falsi cluster
- Difficoltà nel rimuovere gli outliers
- Applicazione di UMAP a due punti che hanno distanza nulla tra loro
- Dataset senza struttura globale significativa

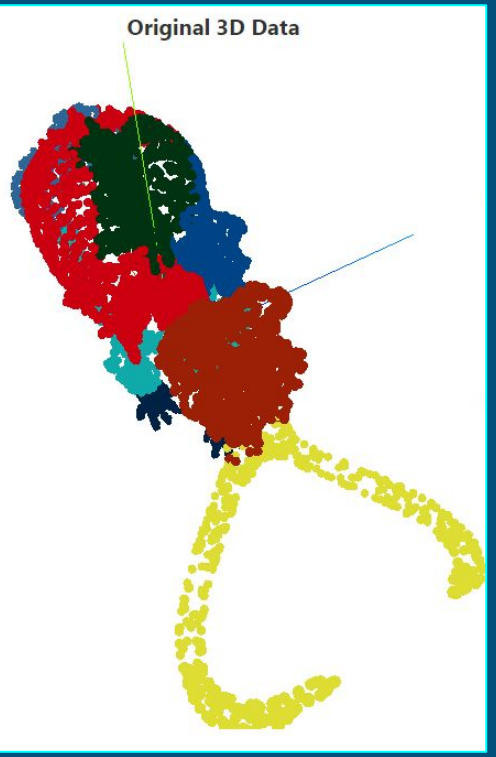


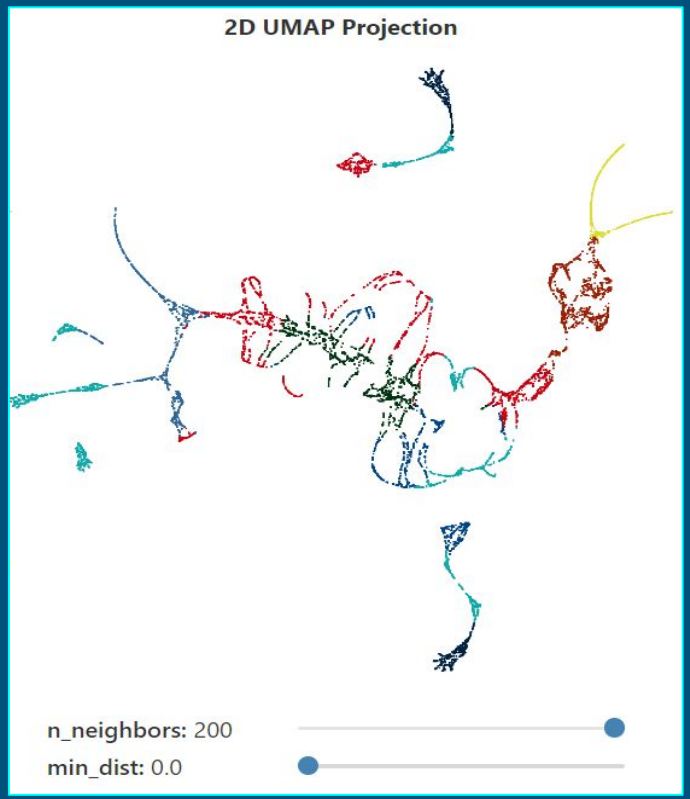
Parametri, Grafo pesato

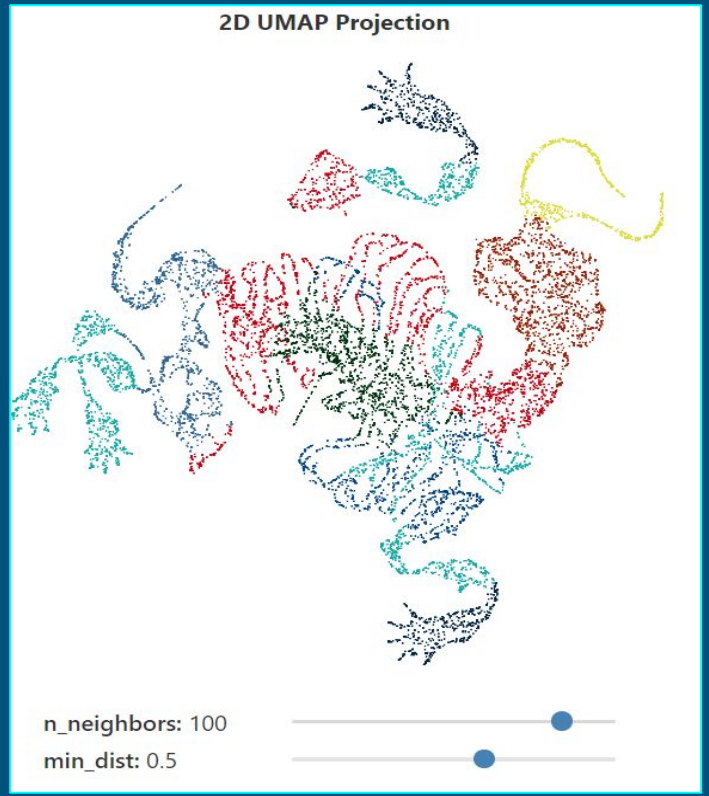
Consideriamo l'insieme dei punti {A,B,C,1,2,3,4} focalizzando l'attenzione solo sugli archi in uscita ed entrata di A e B

B,C hanno le stesse coordinate nello spazio ad alta dimensionalità → causa errore in letteratura

Per vedere l'effetto degli iperparametri **numero vicini** e **distanza minima** utilizzo il sito *Understanding UMAP*. Esempio consideriamo un'immagine 3D di un mammut.







Coenen a., Pearce a. (2024). Understanding UMAP. GitHub. https://pair-code.github.io/understanding-umap/

A sinistra si vede la struttura globale. A destra si vede la struttura locale.

UMAP è un algoritmo con due parti stocastiche

- Individuazione dei k vicini più prossimi di ogni punto
- Ottimizzazione della proiezione a bassa dimensionalità

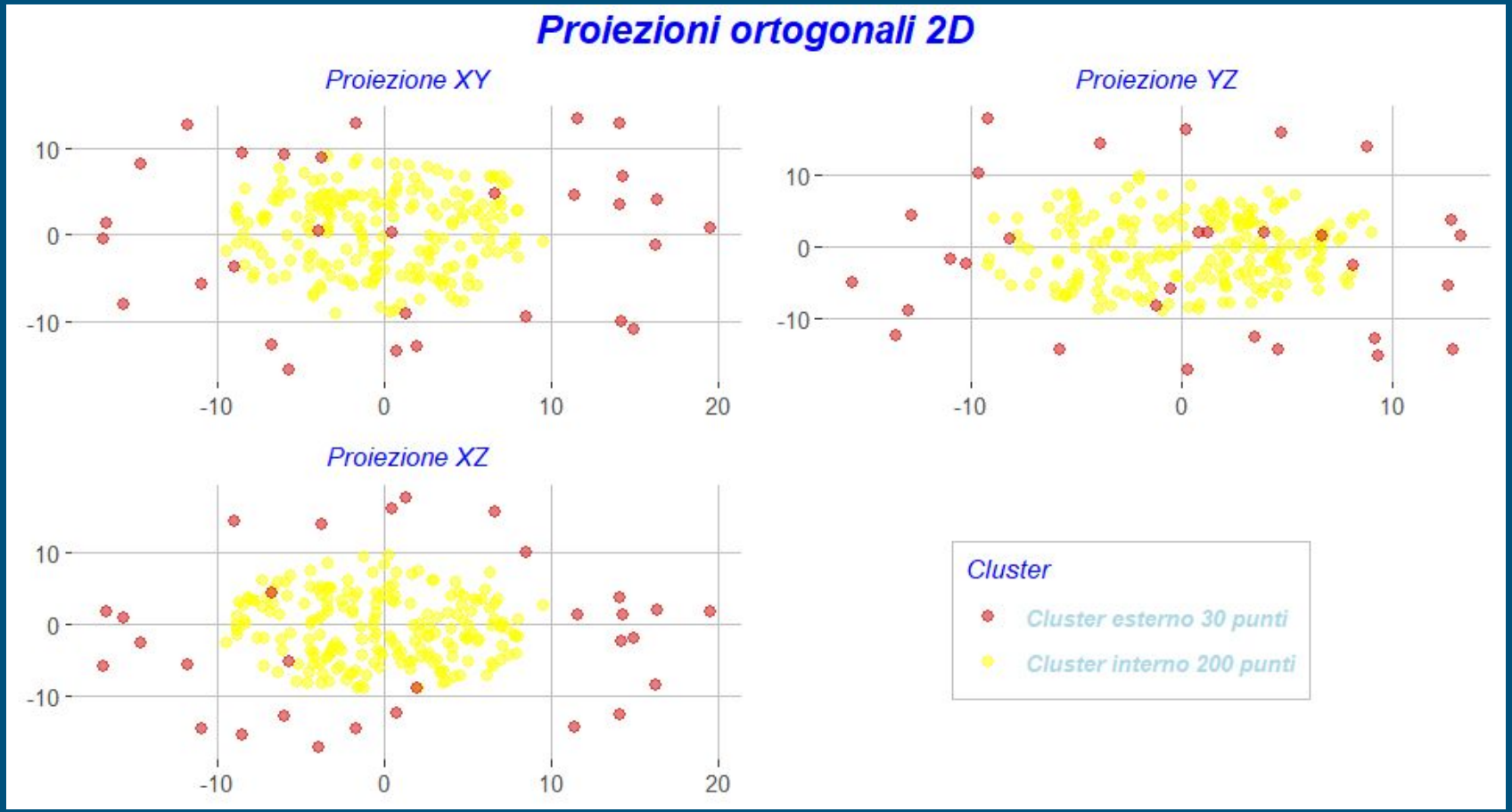
Grazie a queste approssimazioni l'algoritmo ha delle ottime performance nei tempi di computazione.

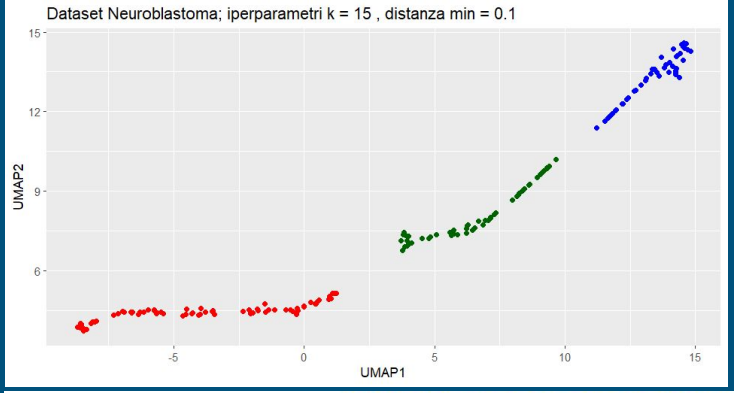
UMAP è disponibile in diversi linguaggi, originariamente sviluppato in *Python*.
Io ho scelto la versione del pacchetto di *Bioconductor*, linguaggio R.

Nel contesto di 

si utilizza il pacchetto **densvis**.

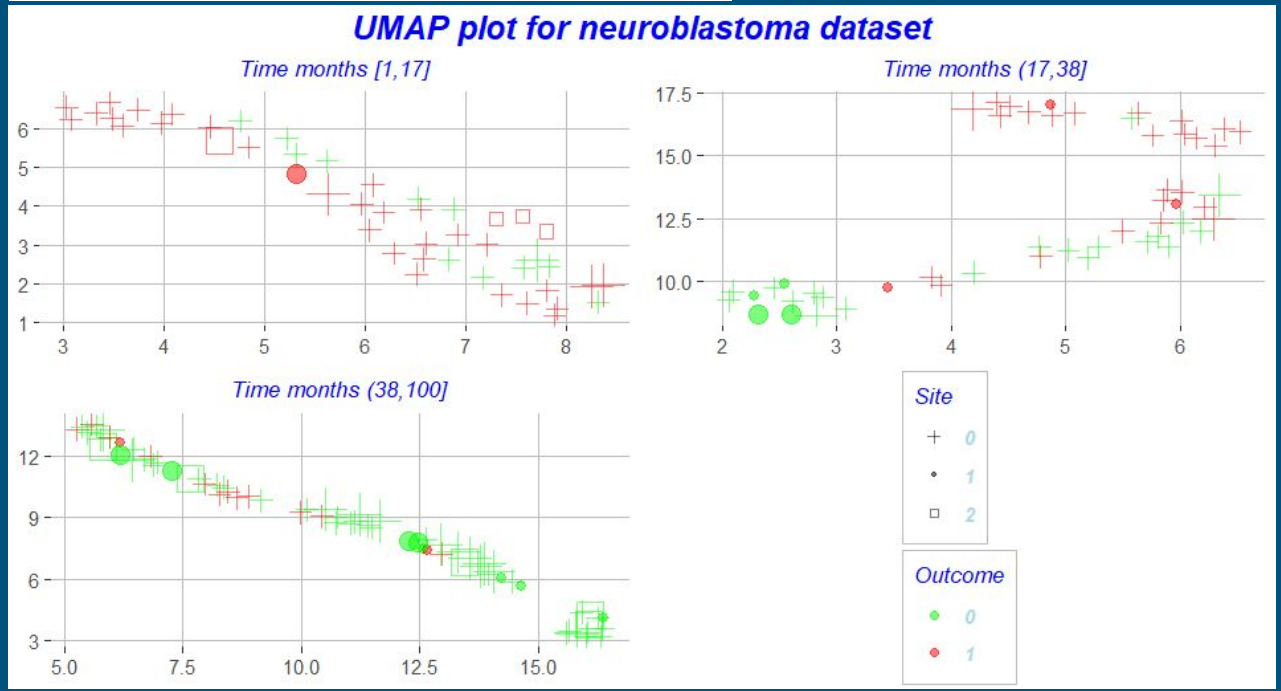
Outliers o cluster utile? creo un dataset artificiale. Ho scoperto che in questo caso le proiezioni UMAP sono di difficile interpretazione.





Dataset Neuroblastoma

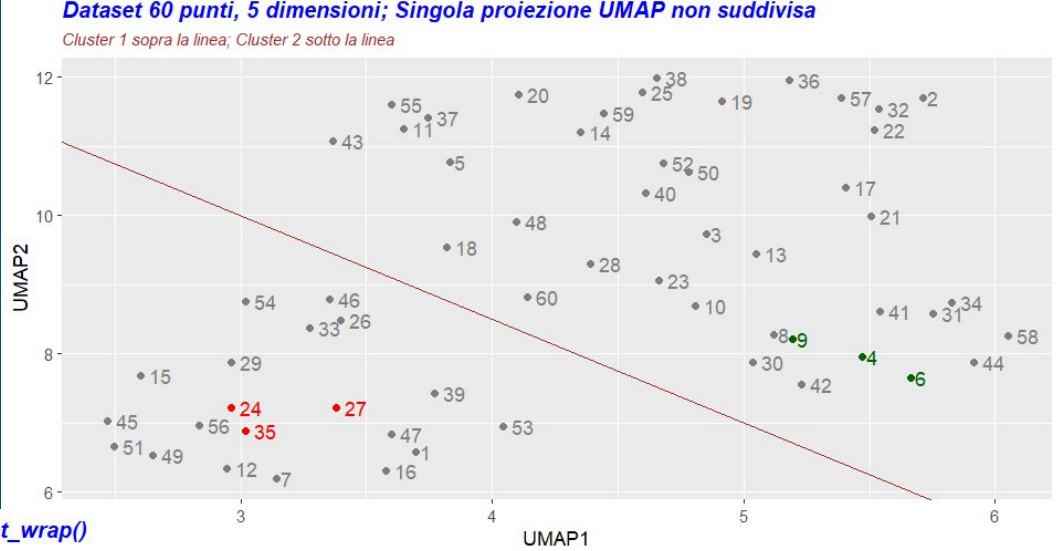
Dataset di 169 pazienti in 13 dimensioni.
Per capire la struttura del dataset e ottenere la proiezione ho fatto molti test.

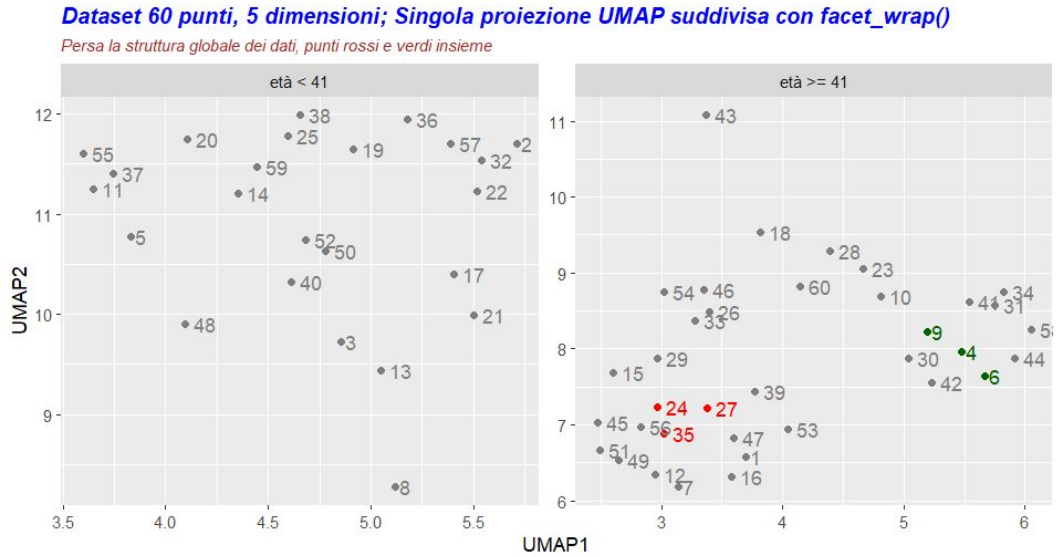


Controprova

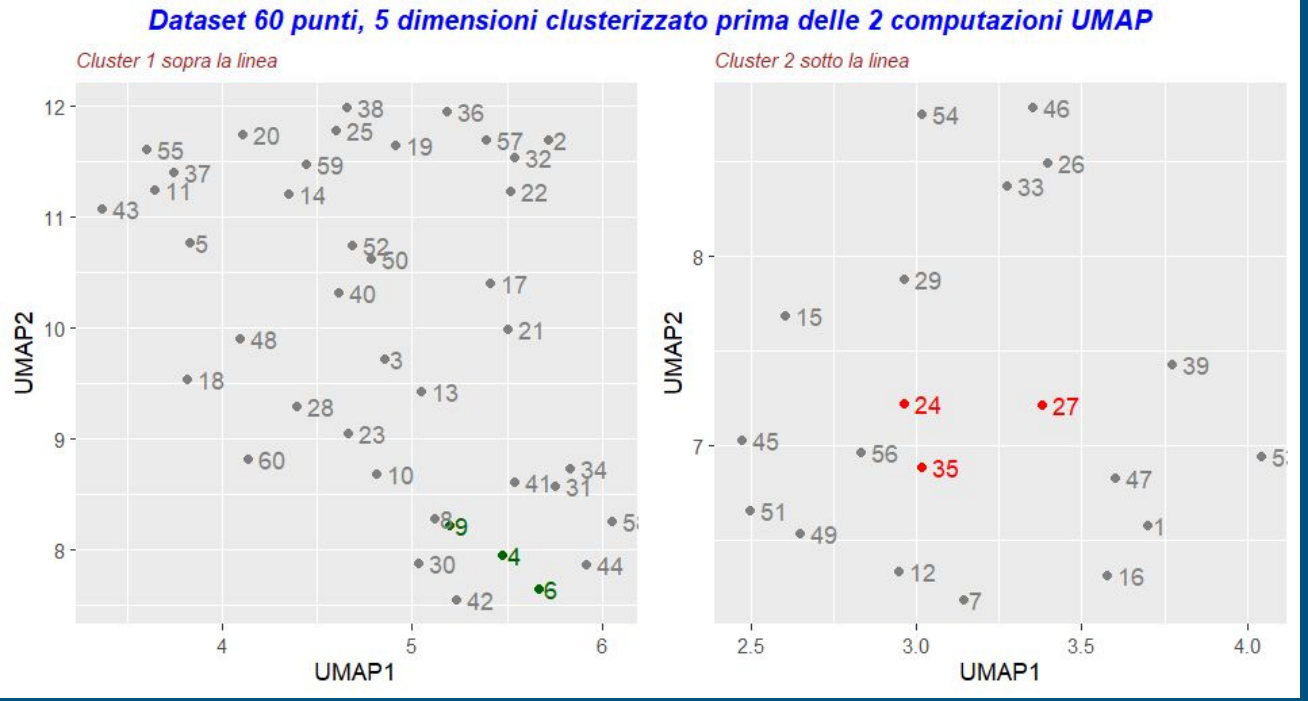
- Forme
- Colori
- Dimensioni
- Partizioni

Potrebbe essere errato avvalersi di funzioni che partizionano le proiezioni prima di fare i grafici





In R ad esempio riguarda la funzione facet_wrap().



Volendo partizionare per una qualunque variabile è sensato prima partizionare il dataset ed in seguito computare il metodo UMAP per ogni partizione.

Le proiezioni di UMAP sono molto variabili, occorre manualmente valorizzare gli iperparametri molte volte per trovare la proiezione che visualizza in modo sensato le informazioni.

Studio e approfondimento sul metodo computazionale UMAP per la riduzione di dimensionalità e la visualizzazione di dati clinici

In conclusione da questa tesi…

- UMAP è usato principalmente prima della clusterizzazione
- L'utilizzatore di UMAP deve conoscere il contesto applicativo
- Il pacchetto di *Bioconductor* **densvis** è molto veloce.
- Aspetto da ricordare è la migliore separazione dei cluster
- Future work: metrica di valutazione; tecniche di pre-elaborazione dati

Grazie dell'attenzione!

[Laurea Triennale in Informatica] Giulio Riggio

844901 g.riggio@campus.unimib.it