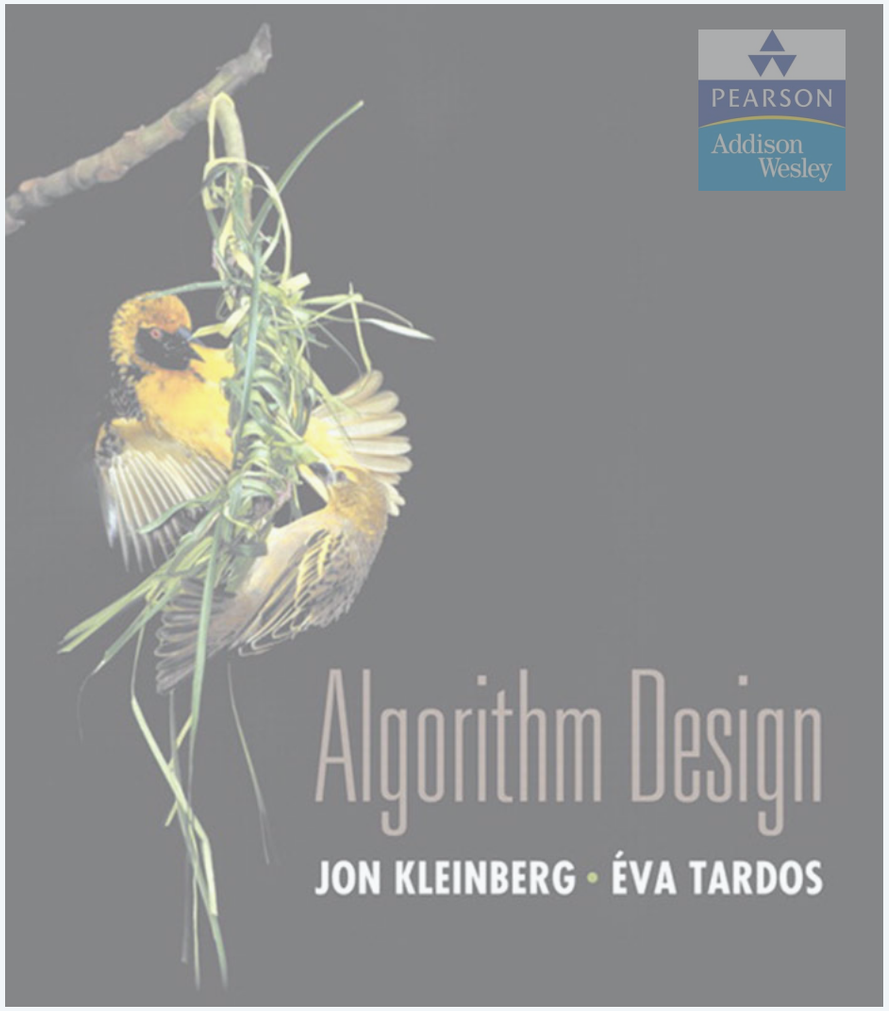


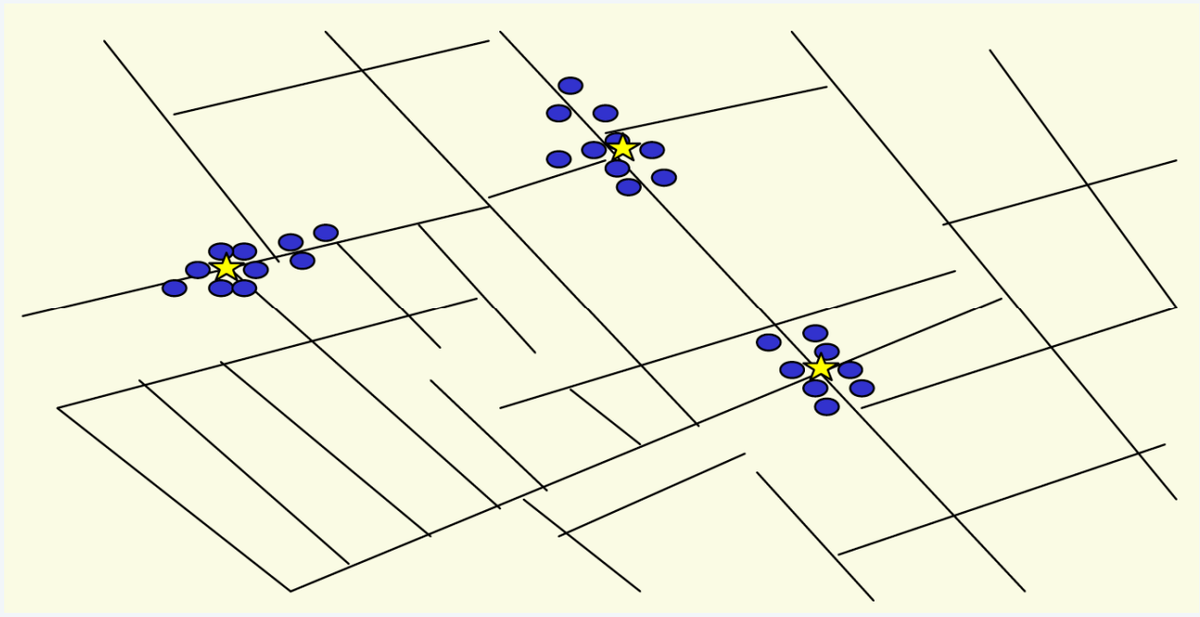
SECTION 4.7

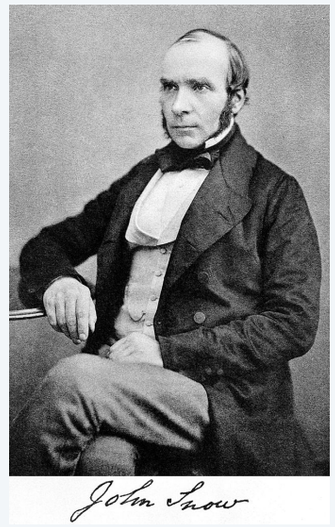
4. ALGORITMI AVIDI II

- *Dijkstra's algorithm*
- *minimum spanning trees*
- *Prim, Kruskal, Boruvka*
- *clustering a massima spaziatura*
- *min-cost arborescences*

Clustering [raggruppamento]

Scopo. Dato un insieme U di n oggetti etichettati $p_1, \ldots, p_n$, partizionarli in gruppi (cluster) in modo che oggetti di gruppi distinti siano lontani tra loro.





epidemia di morti per colera nella Londra del 1850 (fonte: Nina Mishra)

Applicazioni.

- Instradamento in reti mobili ad-hoc.
- Categorizzazione di documenti per la ricerca web.
- Ricerche di similarità in basi di dati di immagini mediche
- Raggruppamento di corpi celesti in stelle, quasar, galassie
- ...

Clustering a massima spaziatura

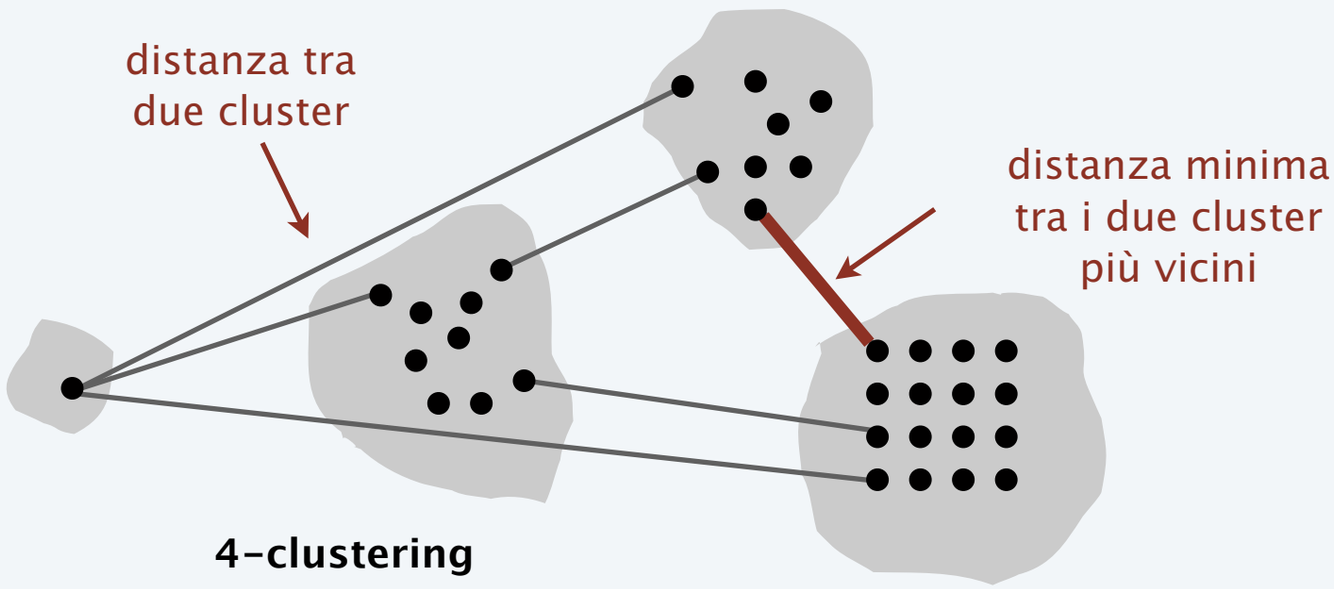
k-clustering. Suddividere gli oggetti in k gruppi non vuoti.

Funzione distanza. Valore numerico per la "vicinanza" tra due oggetti.

- $d(p_i, p_j) = 0$ sse $p_i = p_j$ [identità degli indiscernibili]
- $d(p_i, p_j) \geq 0$ [non-negatività]
- $d(p_i, p_j) = d(p_j, p_i)$ [simmetria]

Spaziatura. Distanza minima tra coppie di punti in cluster distinti.

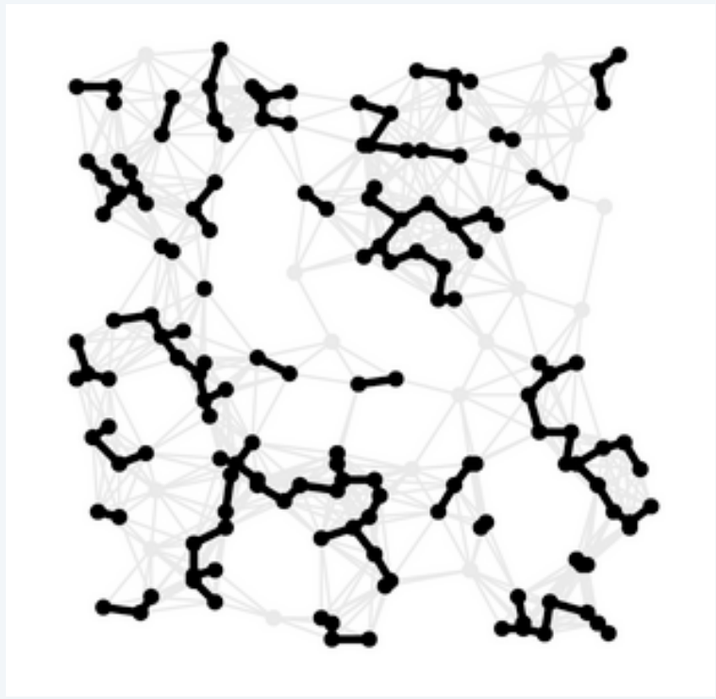
Scopo. Dato un intero k, trovare un k-clustering a massima spaziatura.



Algoritmo avido per il clustering

Algoritmo "ben noto" in letteratura scientifica per il k-clustering:

- Forma un grafo sull'insieme dei nodi U, corrispondenti a n cluster.
- Trova la coppia più vicina di oggetti in cluster diversi, e aggiungi un arco tra i due nodi corrispondenti.
- Ripeti $n - k$ volte (finché non si arriva a k cluster).



Osservazione chiave. Questa procedura è la stessa dell'algoritmo di Kruskal (eccetto che ci si ferma quando si arriva a k componenti connesse).

Alternativa. Trova un MST e cancella i $k - 1$ archi più lunghi.

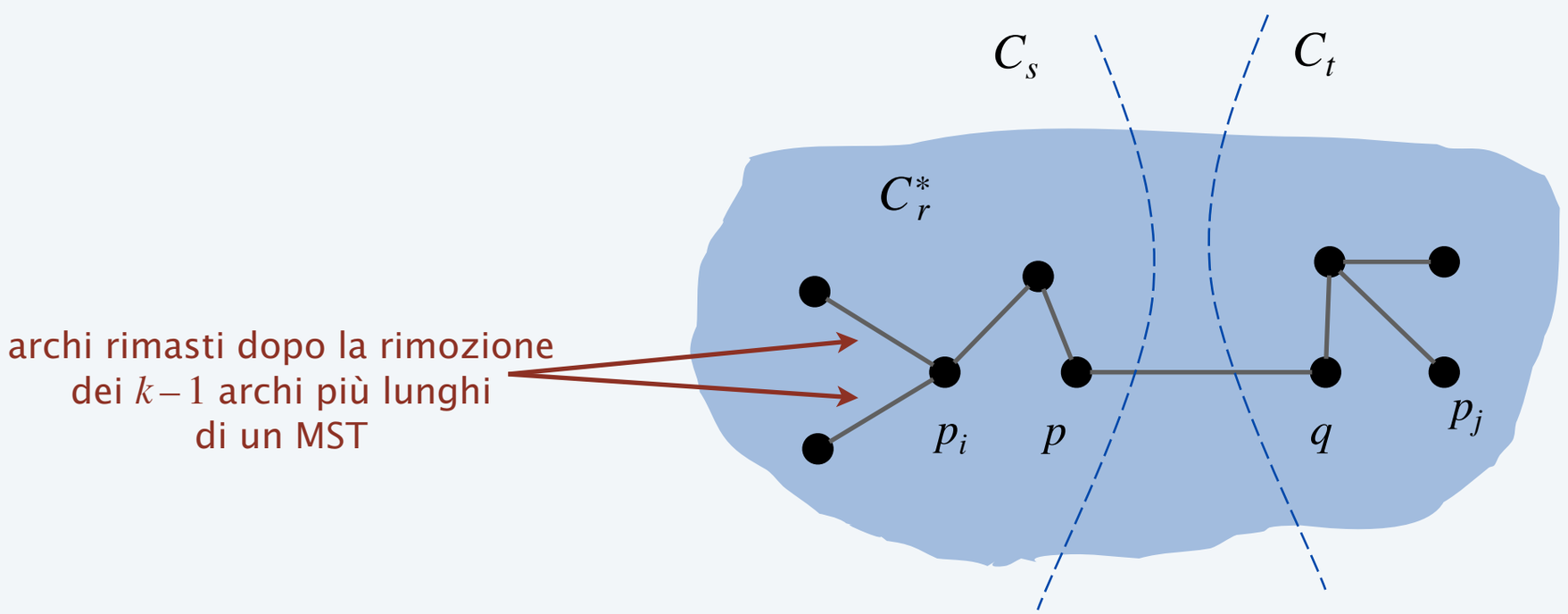
Algoritmo avido per il clustering: analisi

Teorema. Sia C^* un clustering $C_1^*, \ldots, C_k^*$ formato rimuovendo i $k-1$ archi più lunghi di un MST. Allora C^* è un k-clustering a massima spaziatura.

Dim. Sia C un qualunque altro clustering $C_1, \ldots, C_k$.

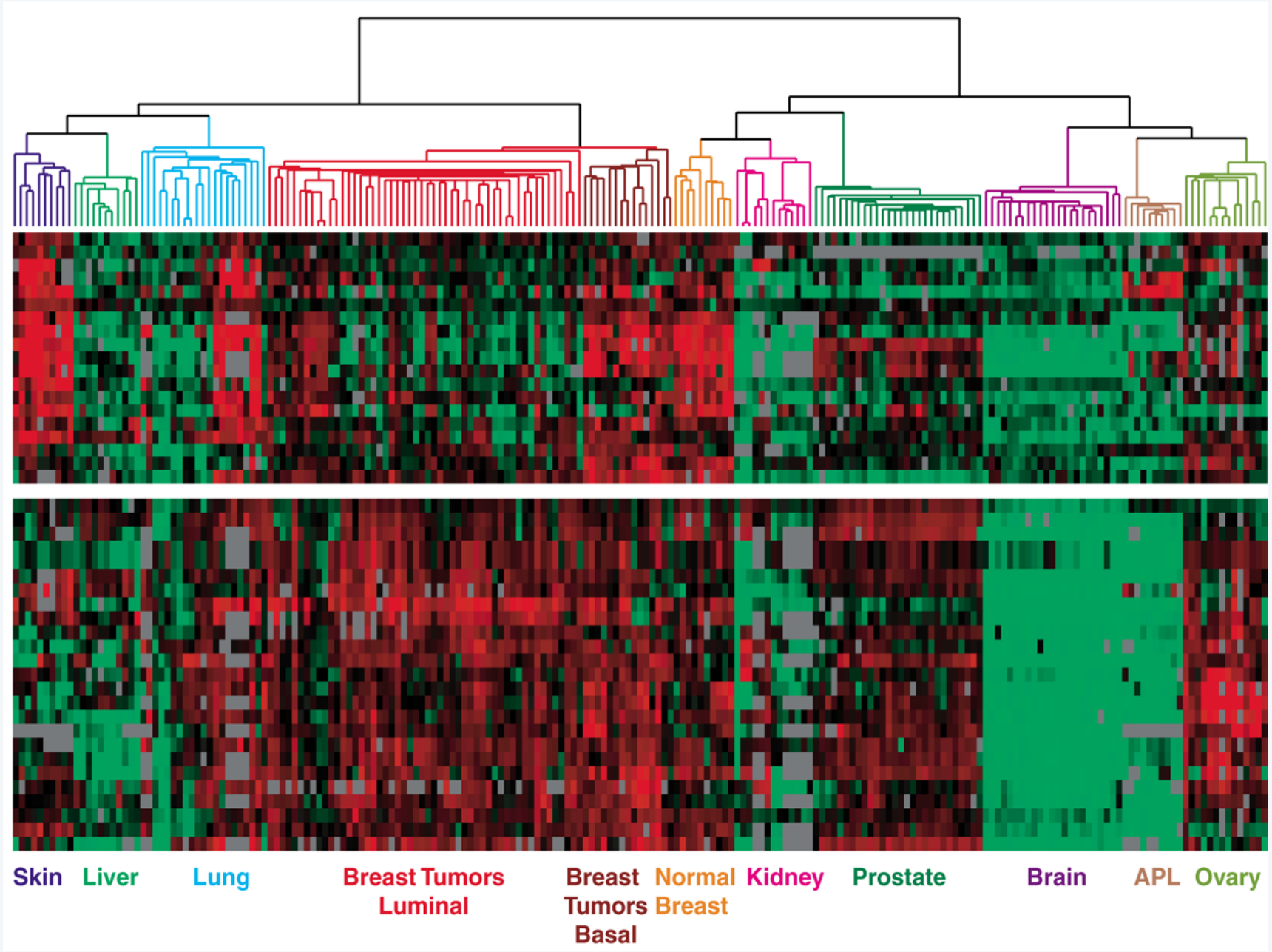
- Siano p_i e p_j nello stesso cluster di C^*, sia esso C_r^*, ma in diversi cluster di C, siano essi C_s e C_t.
- Qualche arco (p, q) sul cammino $p_i - p_j$ in C_r^* interseca due cluster diversi di C.
- Spaziatura di C^*= lunghezza d^* del $(k-1)$esimo arco più lungo di un MST.
- L'arco (p, q) ha lunghezza $\le d^*$ poiché è stato scelto da Kruskal.
- La spaziatura di C è $\le d^*$ poiché p e q sono in cluster diversi. ▪

questo è l'arco che Kruskal avrebbe aggiunto al passo successivo se avessimo proseguito



Dendrogramma dei tumori umani

Tumori in tessuti simili tendono a raggrupparsi assieme nel clustering.



Riferimento: Botstein & Brown group

gene espresso

gene non espresso