

Analisi visuale dei dati

Jessica Leone, Luca Traini, Antinisca Di Marco, Giovanni Stilo

Università degli Studi dell'Aquila



Territori Aperti

Centro di documentazione, formazione e ricerca per la ricostruzione e la ripresa dei territori colpiti da calamità naturali.





Summary

- **Visual Data Mining-VDM**
- **Visual Graph Mining-VGM**
- **Visual Clustering**
- **Tecniche di valutazione di strumenti di VDM**




Visual Data Mining

Big Data

- ▶ Nell'era della raccolta, estrazione, e analisi dei dati, esplorare e analizzare vasti volumi di dati sta diventando sempre più difficile.
- ▶ Con i sistemi di gestione dei dati disponibili, è possibile visualizzare solo porzioni piuttosto ristrette dei dati.
- ▶ Sono necessarie quindi tecnologie di analisi solide e scalabili per estrarre informazioni significative da questi set di dati. [11]



- 
- ▶ L'analisi visiva utilizza la visualizzazione interattiva per integrare il giudizio umano nei processi algoritmici di analisi dei dati
 - ▶ Gli approcci che funzionano a livello puramente analitico o puramente visuale, non aiutano sufficientemente a filtrare informazioni sostanziali da insiemi di dati complessi in rapida crescita e a comunicarle agli esseri umani in modo appropriato



- ▶ **ANALISI VISIVA:** scienza del ragionamento analitico facilitato da interfacce visive interattive, che utilizza tecniche di interazione e visualizzazione per integrare il giudizio umano nel processo di analisi [2]

Visualizzazione
dei risultati

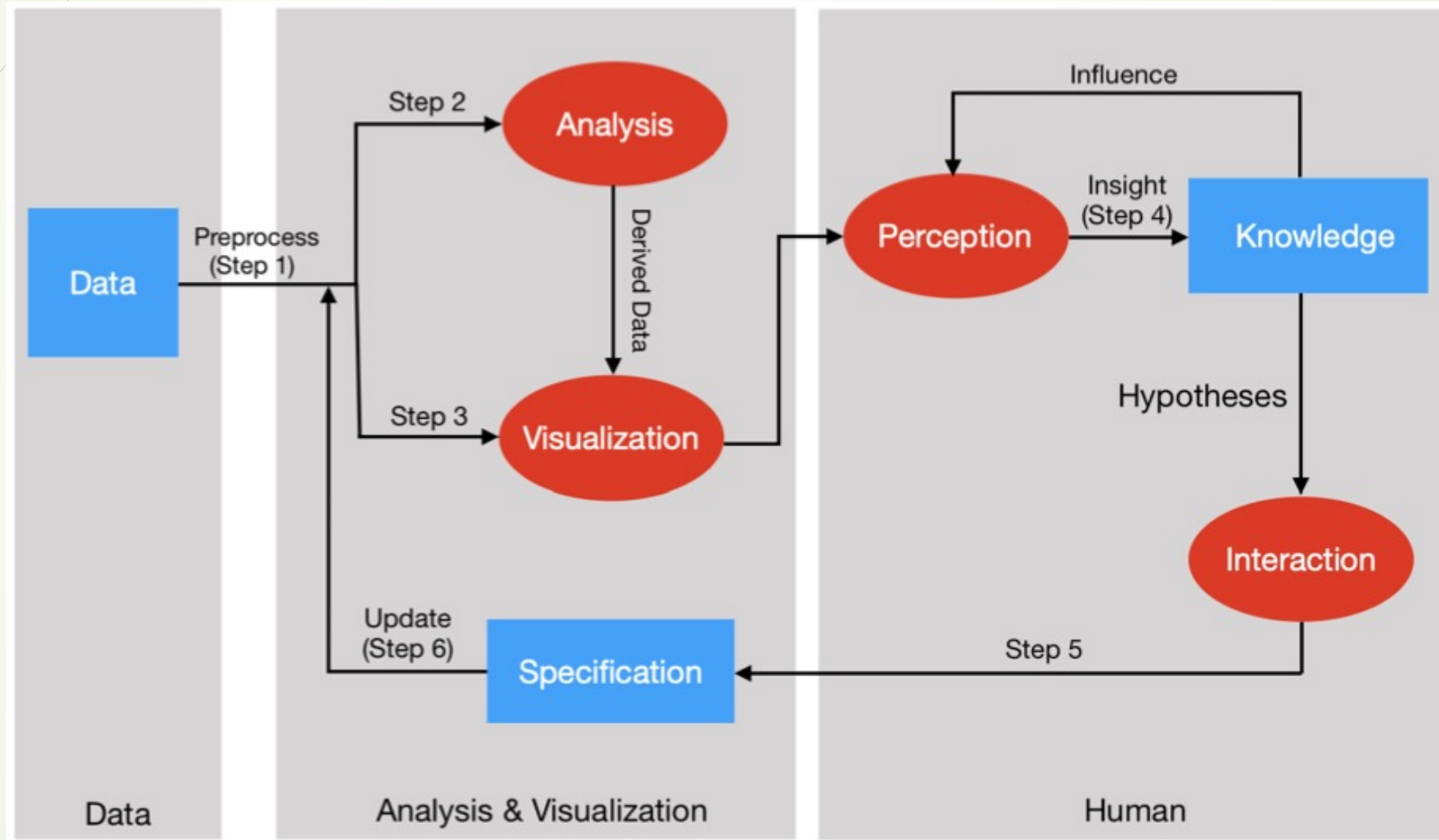


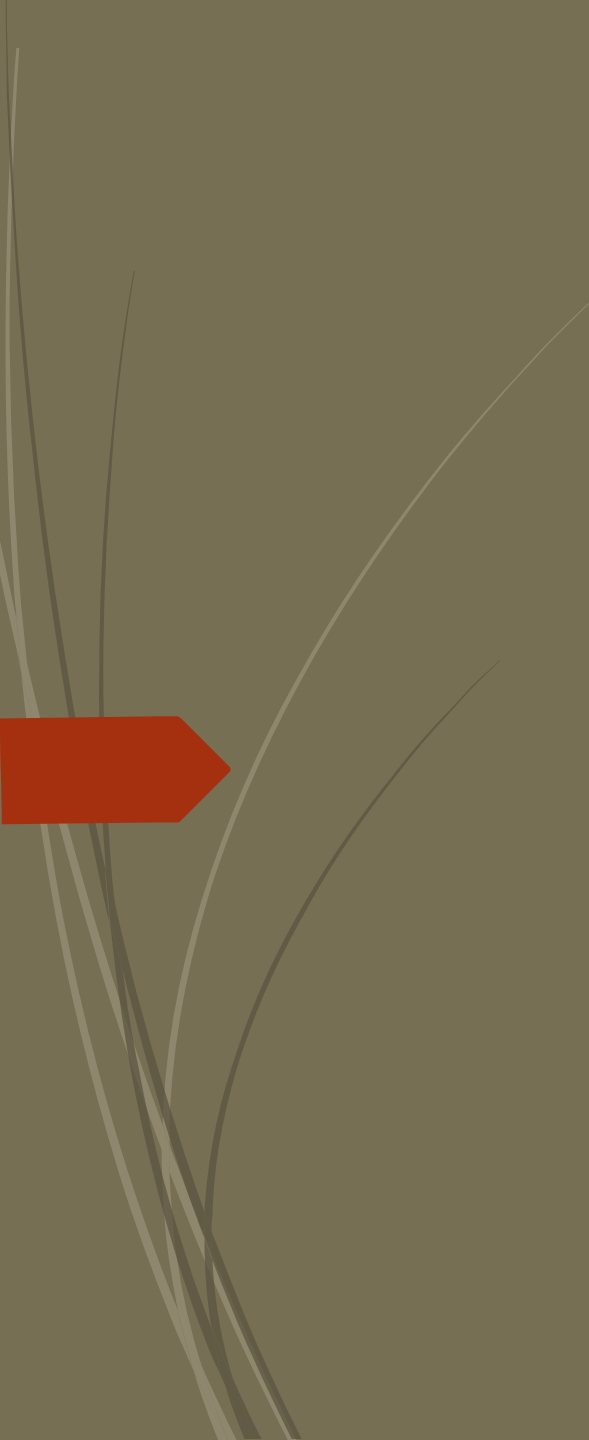
Visualizzazione per
interagire con i dati



VDM
Unione di visualizzazione,
analisi dei dati e processo
di conoscenza

Processo di analisi visiva [2]



- 
- I processi di analisi visiva comprendono diversi step e transizioni:
 1. **Raccolta dei dati**: i dati possono provenire da diverse fonti e quindi devono essere integrati prima di poter applicare ulteriori metodi di analisi. Questo passaggio comprende anche la pre-elaborazione dei dati come la normalizzazione e la pulizia degli stessi.
 2. **Visualizzazione**: l'analista applica tecniche di visualizzazione o tecniche di analisi automatizzata che possono comprendere tecniche statistiche e di Data Mining. Le visualizzazioni, quindi, vengono utilizzate non solo per valutare i risultati ma anche per perfezionare i metodi di analisi automatizzata.

L'alternanza tra analisi automatica e visualizzazione delle informazioni è la caratteristica principale dell'analisi visiva e la differenza rispetto alla semplice visualizzazione delle informazioni.

Overview first, zoom & filter, details on demand

- ▶ L'esplorazione visuale dei dati segue un processo a 3 step [4]:



OVERVIEW


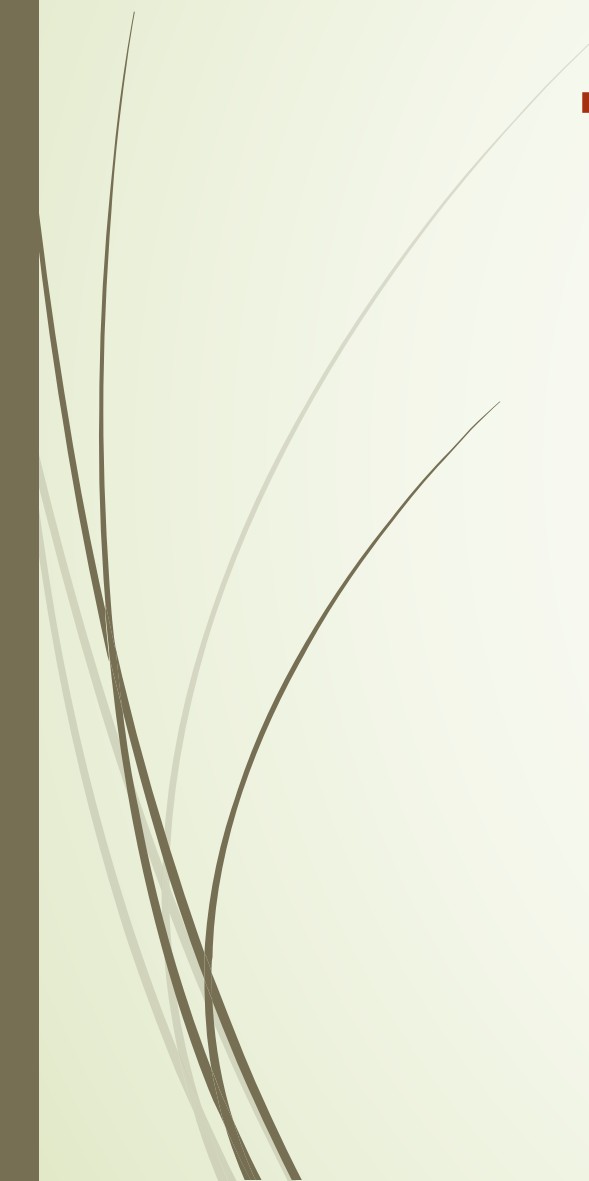
ottenere una panoramica dell'intero set di dati

ZOOM & FILTER

eseguire uno zoom sui dati interessanti e filtrare gli elementi non interessanti

DETAILS ON DEMAND

seleziona un elemento o un gruppo e ottieni i dettagli quando necessario

- 
- 
- Poiché il processo di Visual Data Mining si basa sulla visualizzazione e sull'interazione con essa, il successo del processo dipende [2]:
 1. Dall'ampiezza della raccolta di tecniche di visualizzazione
 2. Dalla coerenza del design delle visualizzazioni
 3. Dalla capacità di rimappare in modo interattivo gli attributi dei dati agli attributi di visualizzazione
 4. Dall'insieme di funzioni per interagire con le visualizzazioni e le capacità che queste funzioni offrono a supporto del processo di ragionamento.



TECNICHE DI VISUALIZZAZIONE

- Le tecniche di visualizzazione possono essere classificate in base:
 - Al **tipo di dati da visualizzare**: dati unidimensionali come i dati temporali, dati bidimensionali come mappe geografiche e tabelle relazionali, testo e ipertesto, gerarchie e grafici ecc..
 - Alle **tecniche di visualizzazione**: 2D/3D, display trasformati geometricamente, mappe ad albero ecc..
 - Alle **tecniche di interazione**: proiezione interattiva, zoom interattivo, distorsione ecc.. [13]

Si noti che le tre dimensioni della nostra classificazione - tipo di dati da visualizzare, tecnica di visualizzazione e tecnica di interazione e distorsione possono essere considerate ortogonali.

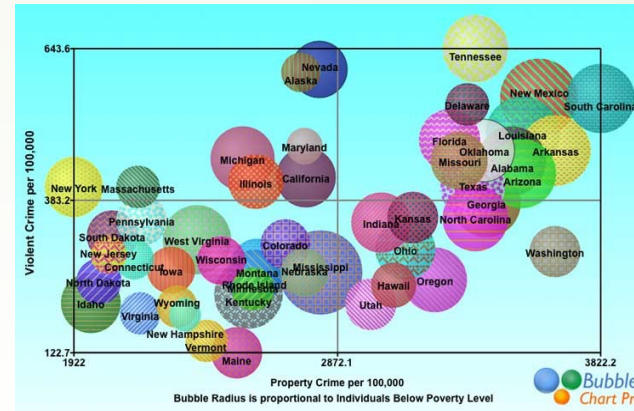
Ortogonalità significa che qualsiasi tecnica di visualizzazione può essere utilizzata insieme a qualsiasi tecnica di interazione così come qualsiasi tecnica di distorsione per qualsiasi tipo di dati.

Si noti inoltre che un sistema specifico può essere progettato per supportare diversi tipi di dati e che può utilizzare una combinazione di più tecniche di visualizzazione e interazione.

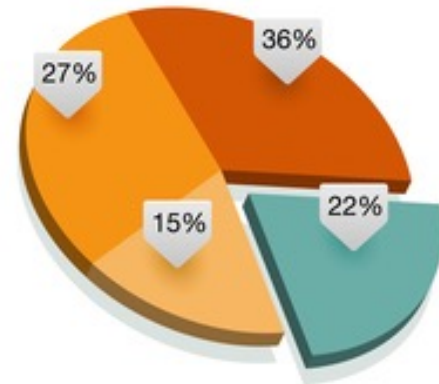
- Esistono diverse categorie di visualizzazione:
 1. Data Visualization o visualizzazione dei dati
 2. Information Visualization o visualizzazione delle informazioni

Data Visualization

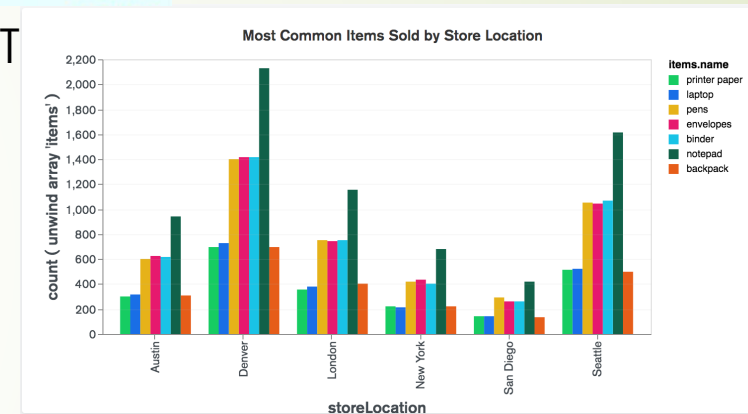
➤ È lo studio della rappresentazione dei dati in una forma sistematica. Tale rappresentazione dovrebbe essere descrittiva e interpretabile per trasmettere il messaggio al lettore in modo efficace. [16]



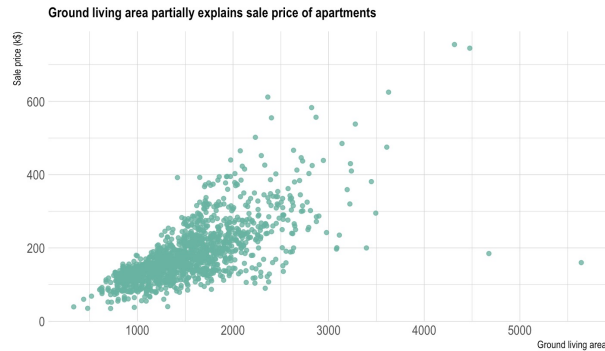
BUBBLE CHART



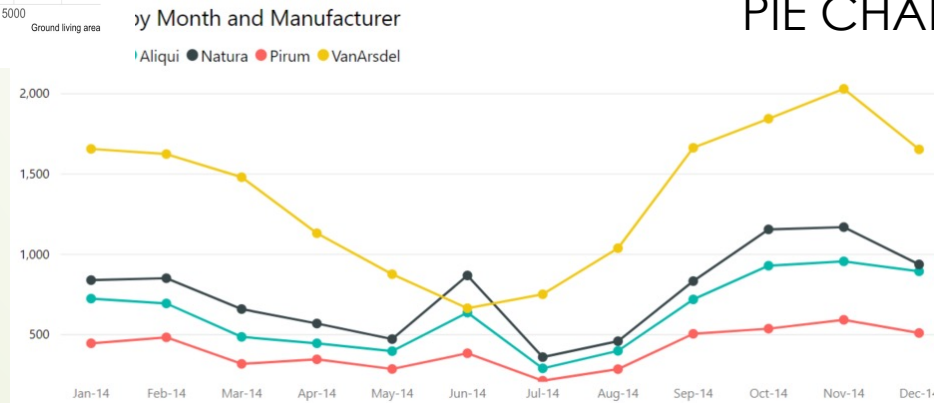
PIE CHART



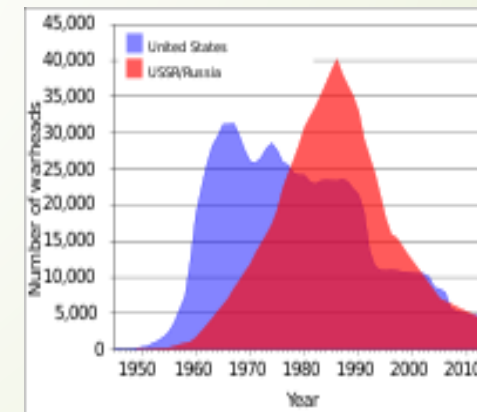
HISTOGRAM



SCATTER PLOT



LINE CHART

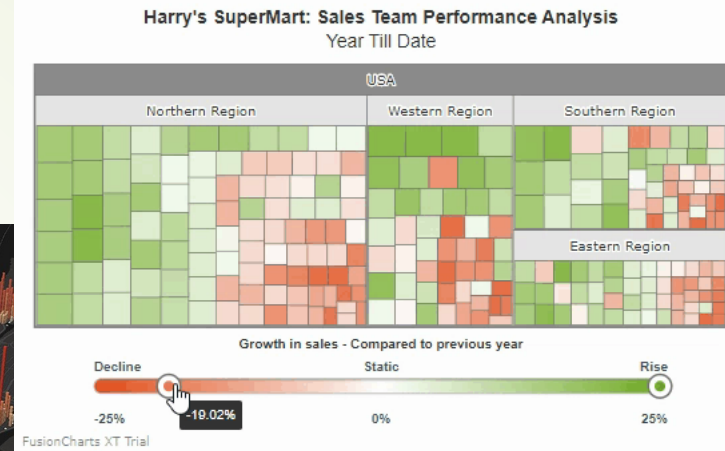
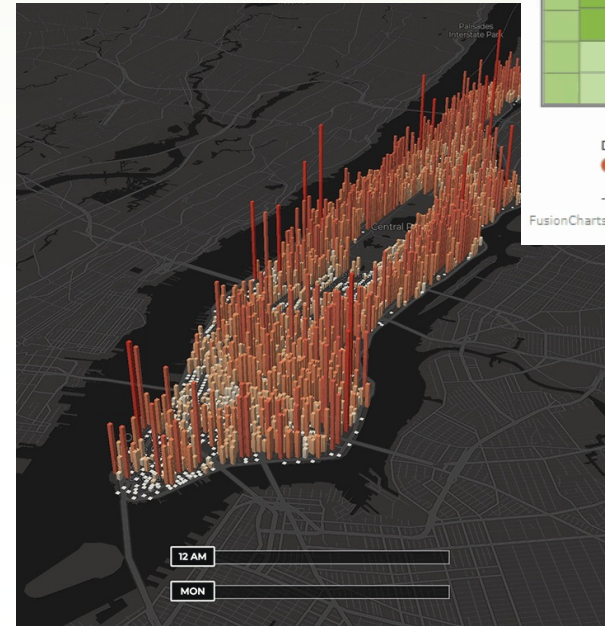


AREA CHART

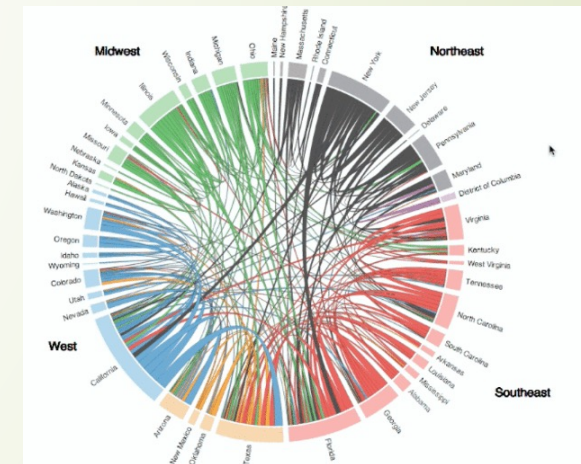
Information Visualization

- È un dominio di ricerca che si concentra sull'uso di metodi di visualizzazione per aiutare le persone a comprendere i dati e a valutarli o analizzarli [16]

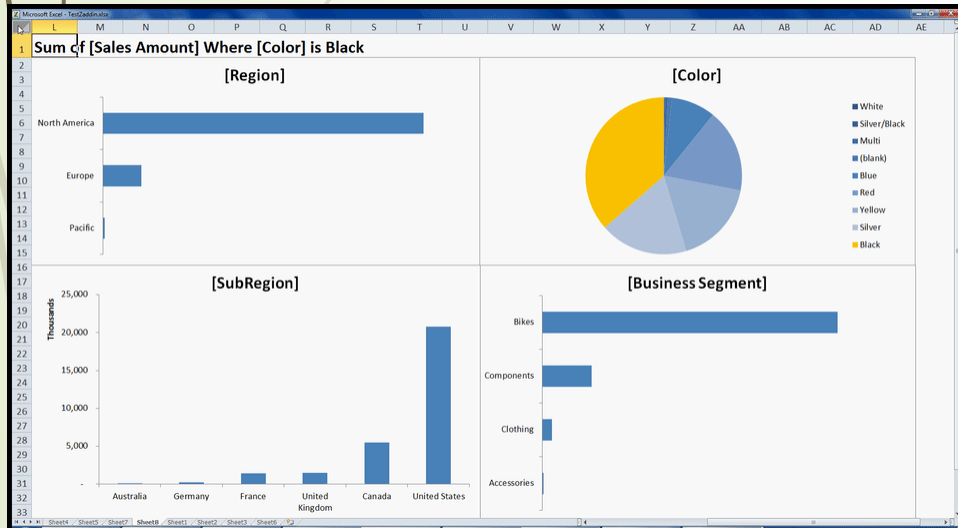
MAP



TREE MAP



INFOGRAPHIC



PIVOT TABLE

PARALLEL COORDINATES



Company	Screen Size (Inches)	Screen Resolution (Pixels)	Ram Memory (GB)	SSD Memory (GB)	HDD Memory (GB)	CPU Model	CPU Clock Rate (GHz)	GPU Model	Price (Euros)
Acer	11.6	1366x768	2	32	0	Intel Celeron	1.5	Intel HD	174
Acer	15.6	1366x768	2	16	0	Intel Celeron	1.5	Intel HD	199



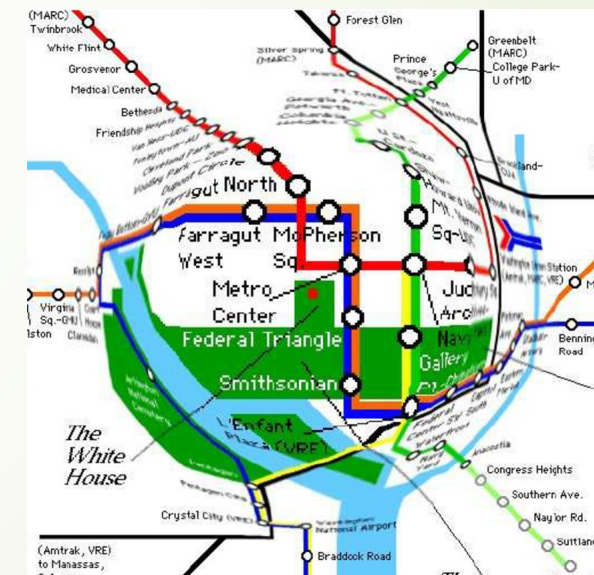
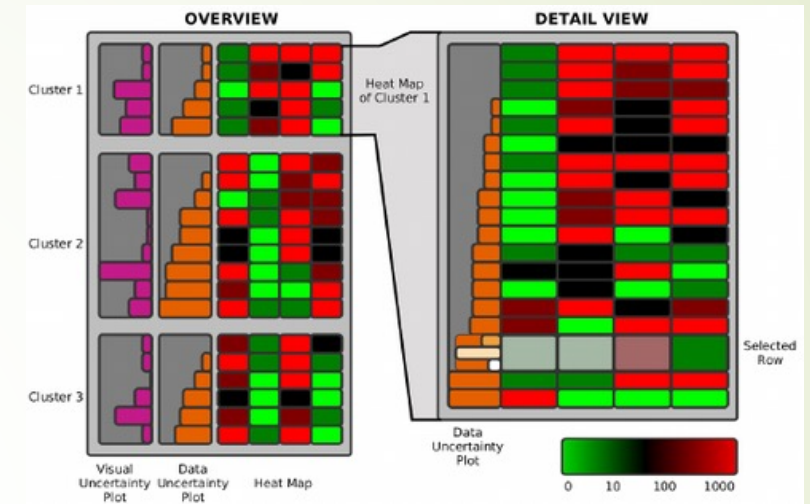
INTERAZIONE

[13]

- ▶ Poiché si è interessati a comprendere i dati, le tecniche interattive sono eccezionalmente importanti. Queste tecniche consentono di manipolare la visualizzazione in modo efficace.
- ▶ Esistono 3 approcci comuni per integrare l'essere umano nel processo di esplorazione dei dati per realizzare diversi tipi di approcci VDM:
 - ▶ **VISUALIZZAZIONE PRECEDENTE:** i dati vengono visualizzati prima di eseguire un algoritmo di Data Mining;
 - ▶ **VISUALIZZAZIONE SUCCESSIVA:** una volta eseguito un algoritmo di Data Mining, i modelli ottenuti vengono visualizzati per renderli interpretabili dall'utente. Sulla base della visualizzazione, l'utente fornisce feedback e potrebbe voler tornare indietro all'algoritmo di Data Mining per utilizzare parametri diversi e provare ad ottenere risultati migliori;
 - ▶ **VISUALIZZAZIONE INTEGRATA:** la visualizzazione viene utilizzata per mostrare risultati intermedi del processo di esplorazione. L'analista identifica sottoinsiemi interessanti su cui può applicare algoritmi di Data Mining;

Esistono diversi modi per poter interagire con la visualizzazione dei dati [16]:

1. **ZOOMING**
2. **OVERVIEW + DETAIL:** utilizza più viste contemporaneamente, ovvero visualizza una panoramica e una vista di dettaglio
3. **FISH EYE:** espande o ingrandisce un'area di messa a fuoco direttamente all'interno della vista generale.
4. **IDENTIFICAZIONE:** mostra un'etichetta identificativa al passaggio del mouse su una determinata area
5. **LINKING:** collega elementi selezionati su grafici diversi.





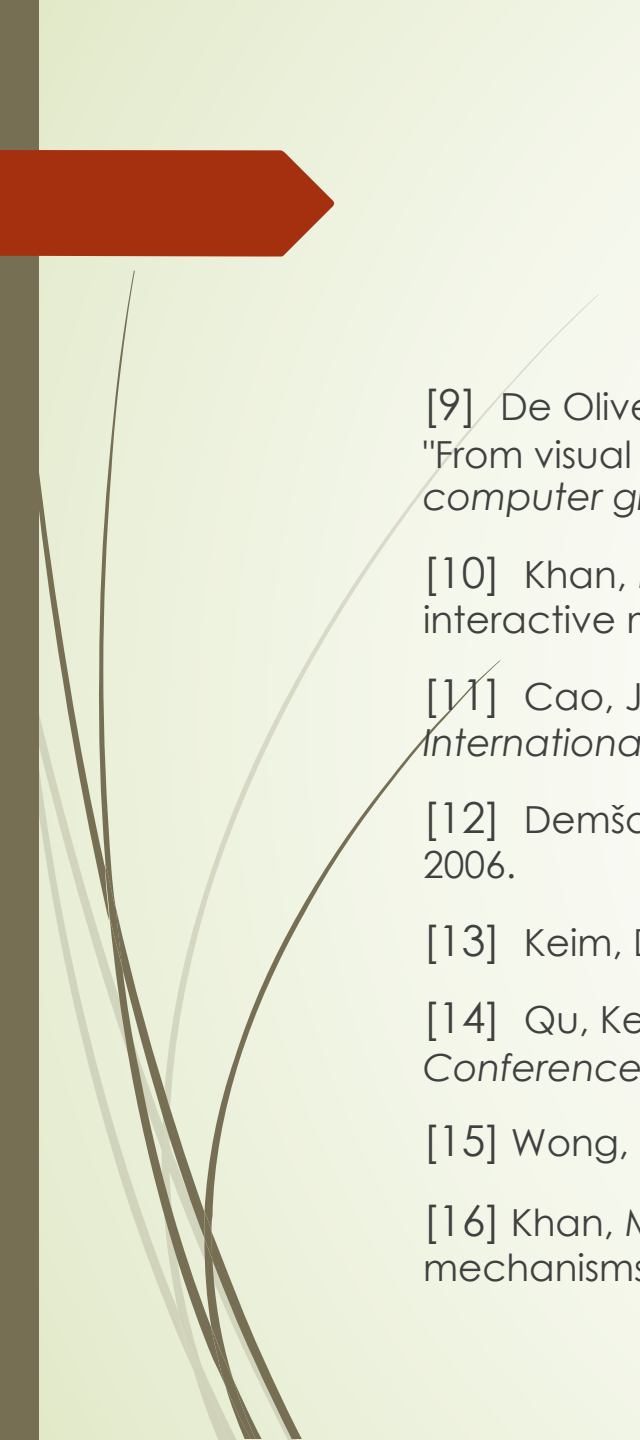
Vantaggi

- ▶ Oltre al coinvolgimento diretto dell'uomo nel processo, il VDM presenta diversi vantaggi:
 1. Permette di gestire dati altamente non omogenei e rumorosi;
 2. L'esplorazione visiva è intuitiva e non richiede la comprensione di algoritmi o parametri matematici o statistici complessi. Utilizzare grafici e immagini di facile comprensione, prima di estrarre i dati, permette di esprimere in modo intuitivo informazioni complesse sui dati;
 3. La visualizzazione può fornire una panoramica qualitativa dei dati, consentendo di isolare i fenomeni di dati per un'ulteriore analisi;
 4. Accelera il progresso e la profondità del Data Mining ed è possibile migliorarne la qualità;
 5. I risultati vengono visualizzati attraverso grafiche e immagini visive specifiche, ciò consente agli utenti di comprendere in modo chiaro e intuitivo le informazioni sui dati e le relative conoscenze e di fornire feedback di valutazione



BIBLIOGRAFIA

- [1] Stasko, John, Carsten Görg, and Zhicheng Liu. "Jigsaw: supporting investigative analysis through interactive visualization." *Information visualization* 7.2 (2008): 118-132.
- [2] Cui, Wenqiang. "Visual analytics: A comprehensive overview." *IEEE Access* 7 (2019): 81555-81573.
- [3] Brown, Eli T., et al. "Dis-function: Learning distance functions interactively." *2012 IEEE conference on visual analytics science and technology (VAST)*. IEEE, 2012.
- [4] Shneiderman, Ben. "The eyes have it: A task by data type taxonomy for information visualizations." *The craft of information visualization*. Morgan Kaufmann, 2003. 364-371.
- [5] Stahl, Frederic, et al. "An overview of interactive visual data mining techniques for knowledge discovery." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 3.4 (2013): 239-256.
- [6] Simoff, Simeon J., Michael H. Böhlen, and Arturas Mazeika. "Visual data mining: An introduction and overview." *Visual Data Mining*. Springer, Berlin, Heidelberg, 2008. 1-12.
- [7] Keim, Daniel A. "Information visualization and visual data mining." *IEEE transactions on Visualization and Computer Graphics* 8.1 (2002): 1-8.
- [8] Simoff, Simeon, Michael H. Böhlen, and Arturas Mazeika, eds. *Visual data mining: theory, techniques and tools for visual analytics*. Vol. 4404. Springer Science & Business Media, 2008.

- 
- [9] De Oliveira, MC Ferreira, and Haim Levkowitz. "From visual data exploration to visual data mining: A survey." *IEEE transactions on visualization and computer graphics* 9.3 (2003): 378-394.
- [10] Khan, Muzammil, and Sarwar Shah Khan. "Data and information visualization methods, and interactive mechanisms: A survey." *International Journal of Computer Applications* 34.1 (2011): 1-14.
- [11] Cao, Juan, and Xinying Zhang. "A Survey on Visual Data Mining Techniques and Applications." *2021 7th International Conference on Systems and Informatics (ICSAI)*. IEEE, 2021.
- [12] Demšar, Urška. *Data mining of geospatial data: combining visual and automatic methods*. Diss. KTH, 2006.
- [13] Keim, Daniel A., and Matthew O. Ward. "Visual data mining techniques." (2002): 2-27.
- [14] Qu, Kecheng, and Lina Wang. "Research on Visual Data Mining Technology." *Journal of Physics: Conference Series*. Vol. 1748. No. 3. IOP Publishing, 2021.
- [15] Wong, Pak Chung. "Visual data mining." *IEEE Computer Graphics and Applications* 19.5 (1999): 20-21.
- [16] Khan, Muzammil, and Sarwar Shah Khan. "Data and information visualization methods, and interactive mechanisms: A survey." *International Journal of Computer Applications* 34.1 (2011): 1-14.



Visual Graph Mining

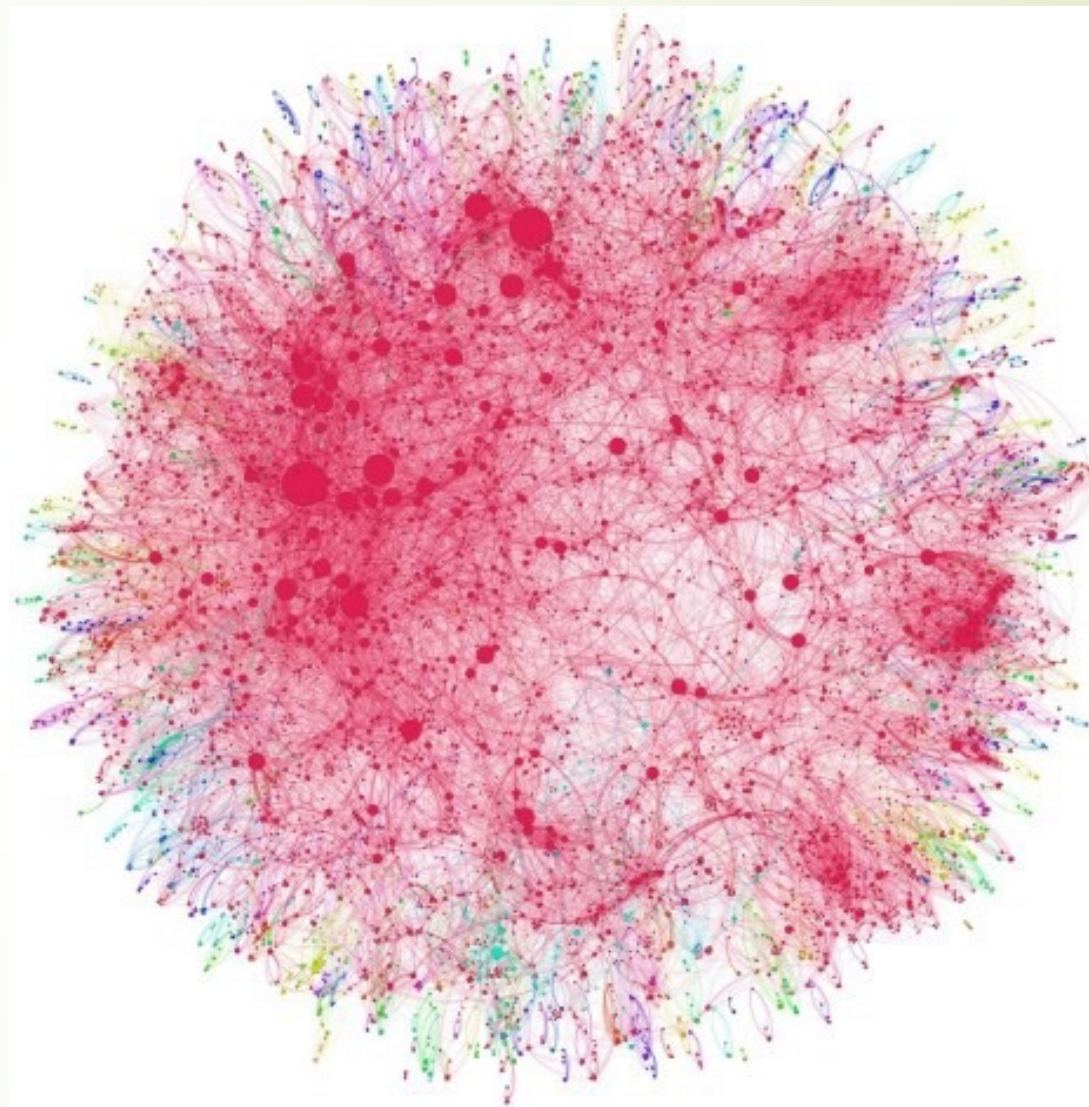
Grafi

- I grafi sono strutture relazionali altamente ottimizzate per comprendere e lavorare con le complesse relazioni tra i dati.
- Utilizzando i grafi per la visualizzazione dei dati, risulta più semplice e veloce assimilare le informazioni in quanto il cervello umano elabora le informazioni visive e strutturate più velocemente rispetto alle informazioni scritte.
- Gli strumenti di visualizzazione dei grafi sono perfetti per visualizzare le relazioni ma anche per comprendere il contesto dei dati.



Grafi e Big Data

- Tuttavia, la quantità di dati in rapida crescita non consente più una presentazione di tutti i dati
- Ad esempio Wikipedia ha milioni di articoli che formano una rete attraverso riferimenti incrociati, Facebook connette più di un miliardo di utenti in una struttura incredibilmente complessa di amici, chat, inviti di gruppi ecc.
- Diventa quindi fondamentale avere la capacità di analizzare in modo efficiente raccolte di dati così complesse. L'uso di semplici statistiche per ragionare sulla dinamica di reti così complesse non è generalmente efficace o pratico [4].



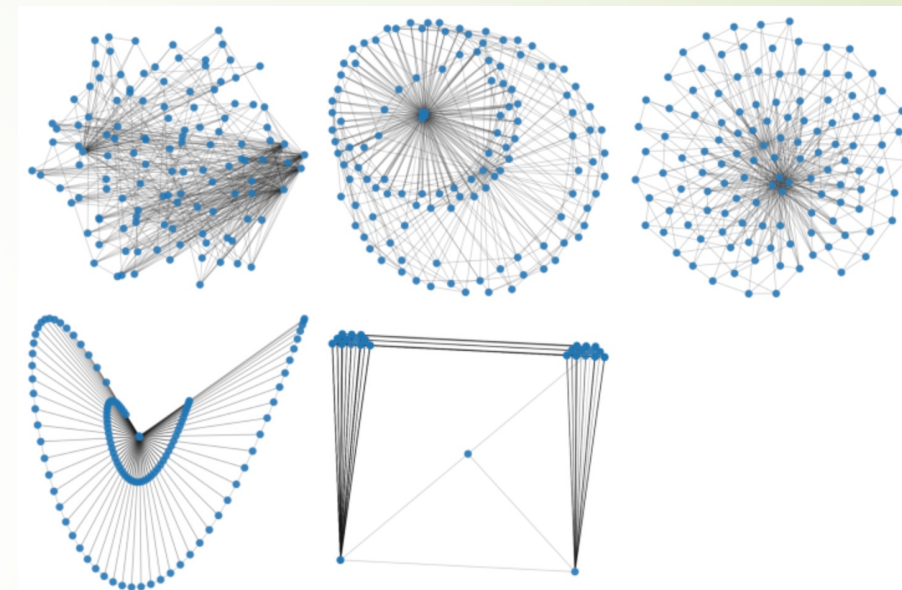


Visual Graph Mining

- ▶ Gli analisti si dirigono verso l'utilizzo della visualizzazione come metodo altamente interattivo che combinano rappresentazioni visive con analisi di rete per migliorare notevolmente la capacità di comprendere e caratterizzare le reti
- ▶ Le tre componenti principali per efficaci sistemi di Visual Graph Mining sono [1]:
 1. Visual Graph Representation
 2. Graph Algorithmic Analysis
 3. User interaction

Visual Graph Representation

- ▶ La visualizzazione è uno dei principali mezzi di analisi esplorativa dei grafi. Comprende:
 - lo sviluppo di tipi appropriati di rappresentazioni visive (ad es., diagrammi a matrice o collegamento di nodi),
 - posizionamento efficiente di elementi grafici sullo schermo e
 - mappature di attributi visivi efficienti (progettazione di elementi grafici per una migliore leggibilità del disegno).
- ▶ Nella visualizzazione grafica creata dal computer, vengono presi in considerazione diversi criteri cosiddetti estetici [2]. Ad esempio:
 1. Intersezione minima degli archi
 2. I vertici adiacenti sono più vicini l'uno con l'altro
 3. Le communities sono raggruppate in cluster
 4. Minima sovrapposizione di nodi e archi
- ▶ Oltre ai criteri estetici, la visualizzazione esplorativa dei grafi richiede più di un algoritmo di layout per rivelare diverse prospettive sulle relazioni tra i nodi.



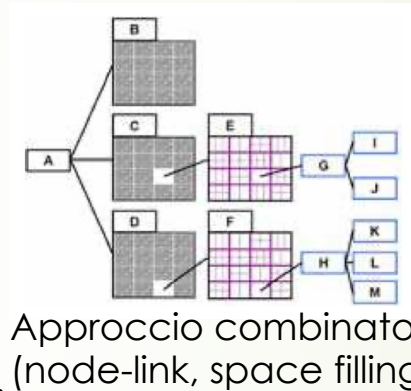
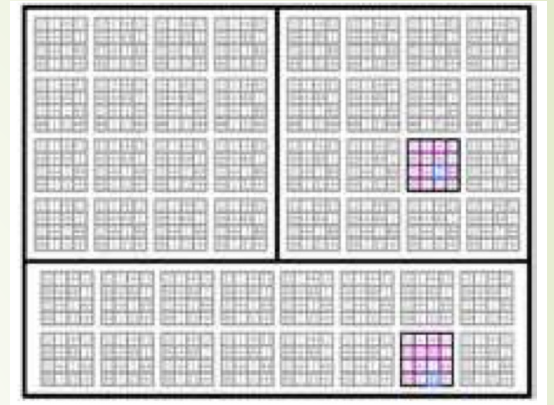
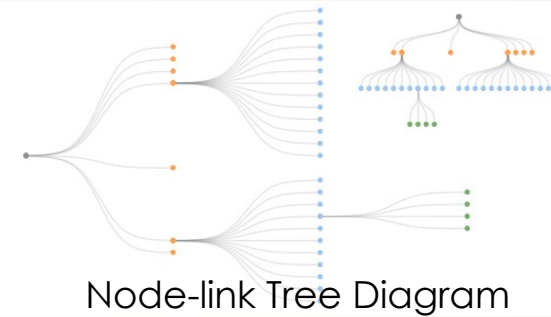
Grafi Statici [1]

ALBERI

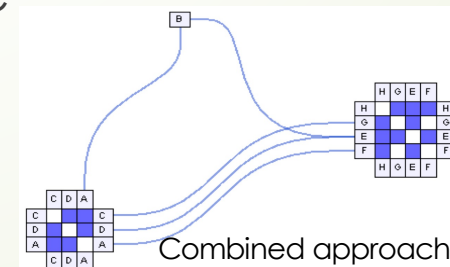
1. **Node-link:** utilizzano i collegamenti tra gli elementi per rappresentare la loro relazione.
2. **Space filling techniques:** Queste tecniche utilizzano le posizioni spaziali dei nodi, usando la vicinanza o la chiusura.
3. **Approccio combinato**

GRAFI DIRETTI E NON DIRETTI

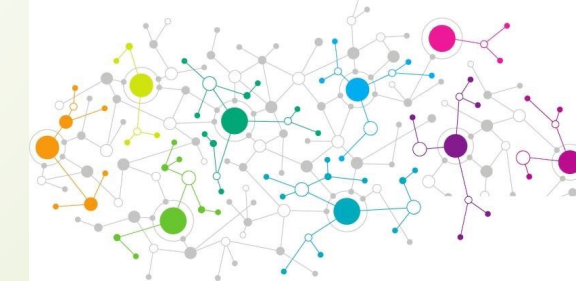
1. **Node-link:** utilizzano i collegamenti tra gli elementi per rappresentare la loro relazione.
2. **Adjacency matrix:** Queste tecniche visualizzano la matrice di adiacenza di un dato grafo, dove gli attributi degli spigoli sono codificati nelle celle della matrice.
3. **Approccio combinato**



Adjacency matrix

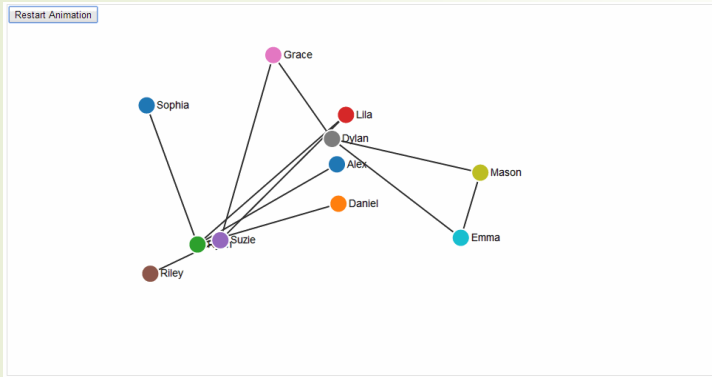


Node-link

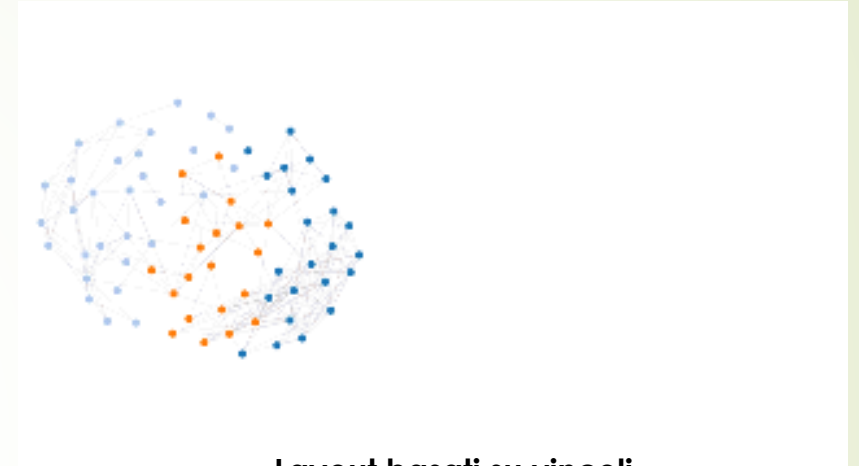
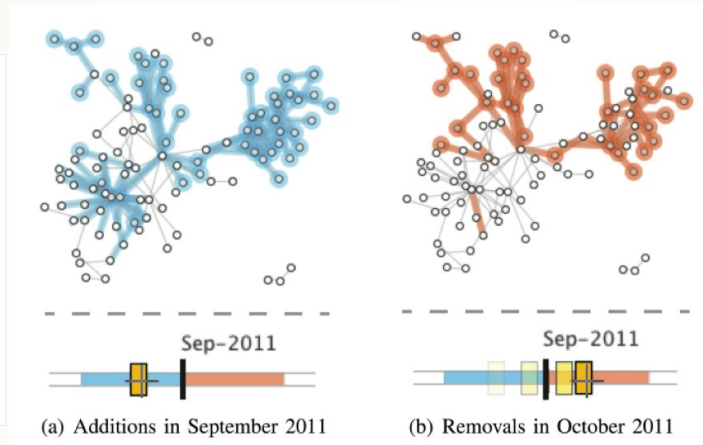


Grafi Dinamici [2]

Le visualizzazioni di grafi dinamici vengono utilizzate in molte applicazioni del mondo reale per presentare relazioni in evoluzione tra entità, ad esempio nell'analisi dei social network.

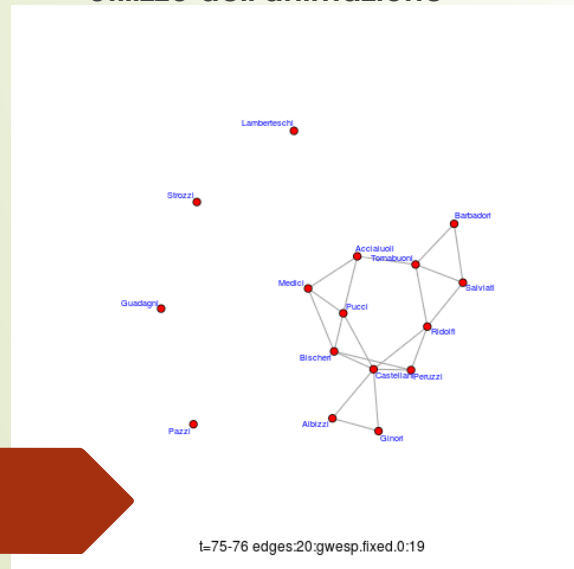


Utilizzo dell'animazione

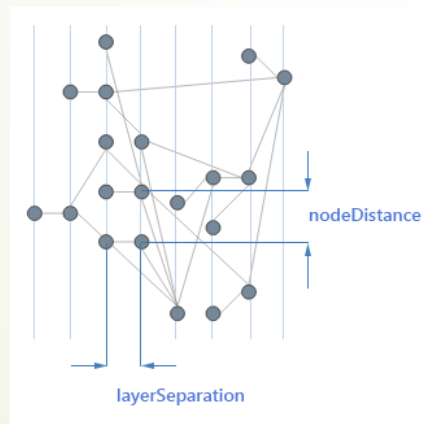


Layout basati su vincoli

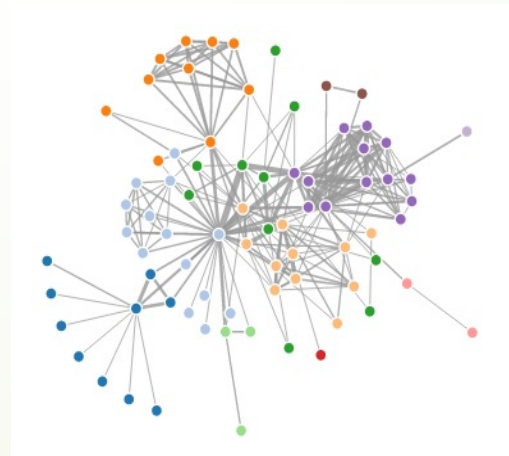
Visualizzazione statica con dimensione temporale



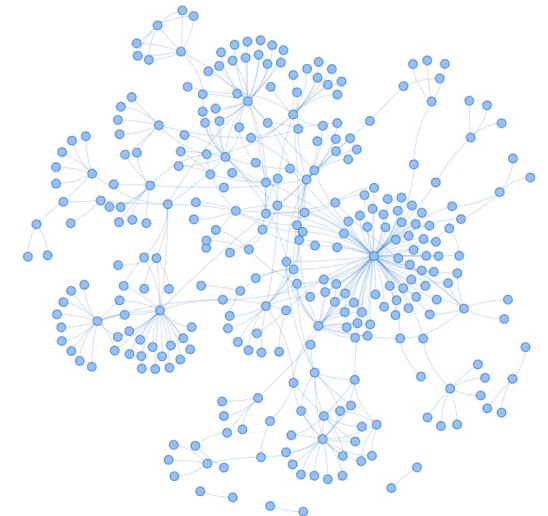
t=75-76 edges:20:gwsp.fixed.0:19



Layout a strati



Approcci multiscala



Layout basati sulle forze

User Interaction [1]

DATA MANIPULATION

Influisce sulla selezione dei dati da visualizzare o può modificare i valori dei dati

DATA SELECTION

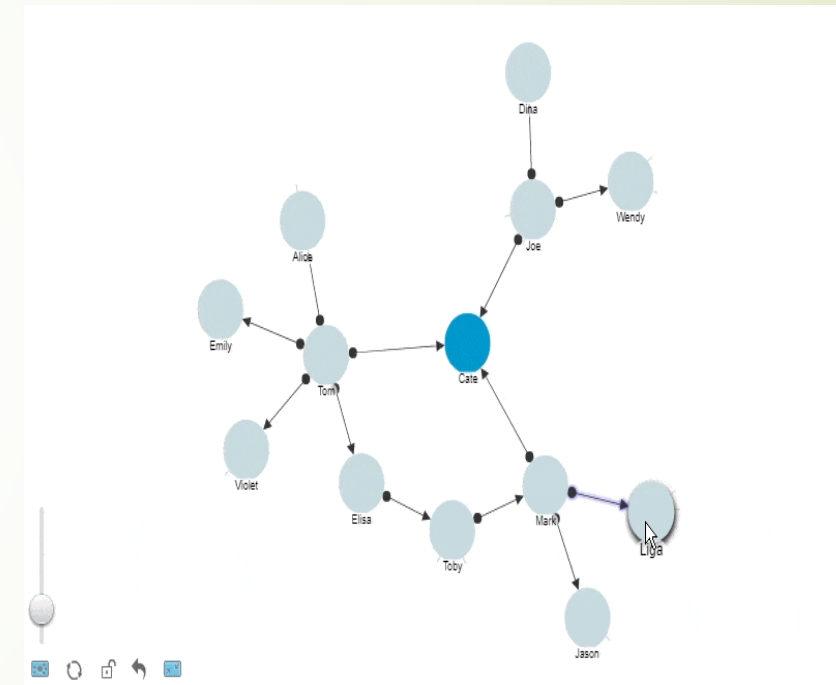
La selezione dei dati può seguire 3 percorsi:

1. Top-Down: parte dall'intero grafo e quindi vincola la parte del set di dati da visualizzare filtrando secondo criteri o selezionando manualmente. Questo approccio permette di avere prima una panoramica del grafo completo, per poi concentrarsi solo sulle parti interessanti
2. Bottom-up: parte da un nodo selezionato e mostra successivamente più nodi/archi su richiesta.
3. Middle-out: combina l'approccio top-down e bottom-up.

MODIFICA AI VALORI DEI DATI

L'utente può modificare i valori dei dati su un livello o creare/modificare aggregazioni di grafi

1. Graph editing: l'utente può eliminare o aggiungere in modo interattivo nodi o archi direttamente dall'interfaccia
2. Interactive graph aggregation: utilizzata per semplificare il grafo. Gli algoritmi di aggregazione dei grafi producono riepiloghi piccoli e comprensibili e possono evidenziare le comunità nella rete, il che facilita notevolmente la sua interpretazione.

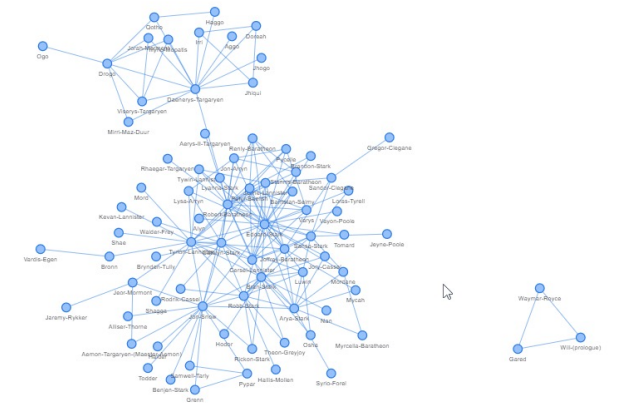
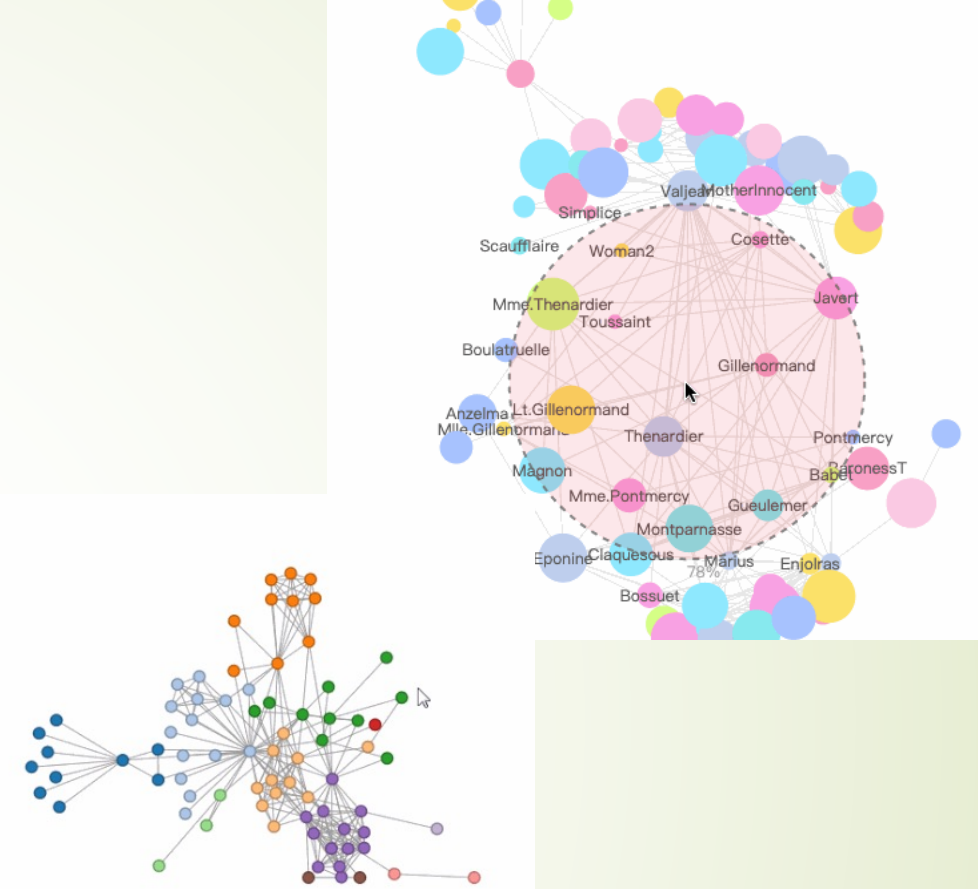


VIEW INTERACTION

Riguarda la regolazione del tipo di presentazione visiva e dei suoi parametri

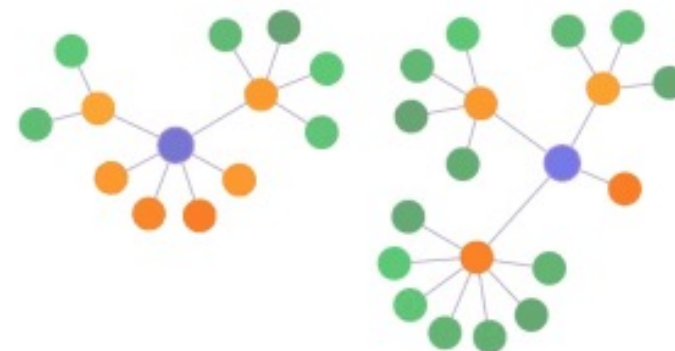
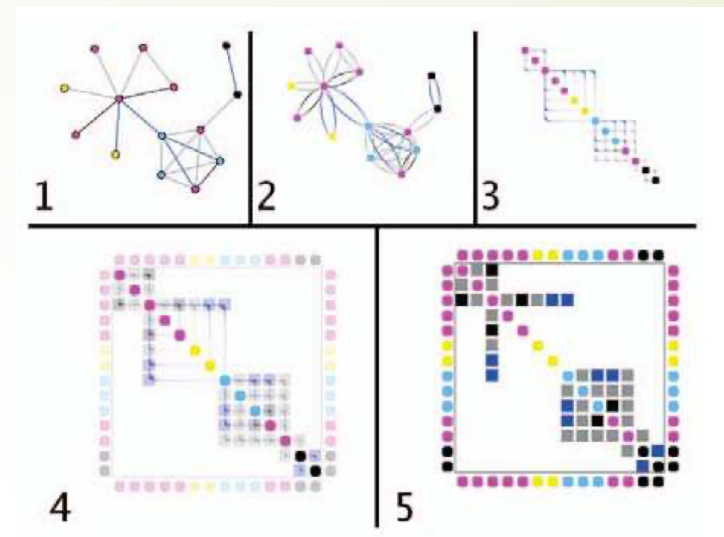
1. CAMBIAMENTO DEI PARAMETRI

- **Highlighting:** evidenziare nodi o parti del grafo particolarmente interessanti
- **Linking and brushing:** si utilizzano più viste coordinate per mostrare i dati da diverse prospettive. In queste viste, la modifica a una visualizzazione viene trasferita automaticamente alle altre viste
- **Panning:** consente di navigare lungo gli archi di un nodo selezionato nel grafo e quindi esplorare la struttura del grafo. Può essere combinato con lo zoom sull'arco e la distorsione dei nodi
- **Zoom:** aumenta il livello di dettaglio effettuando drill-down a livelli inferiori di aggregazione
- **Distorsione:** assegna più spazio agli elementi in aree focalizzate e quindi migliorano la leggibilità dei dati di interesse. La selezione interattiva di aree di messa a fuoco aiuta a esplorare le diverse parti dei dati in modo più dettagliato
 - Grafiche fish-eye: risolvono la sovrapposizione degli archi spostandoli in un'area più ampia. Questa tecnica è particolarmente utile nei grafi geografici, dove non si vuole il riposizionamento di nodi e archi
 - Filtering: spostare la posizione di nodi e archi



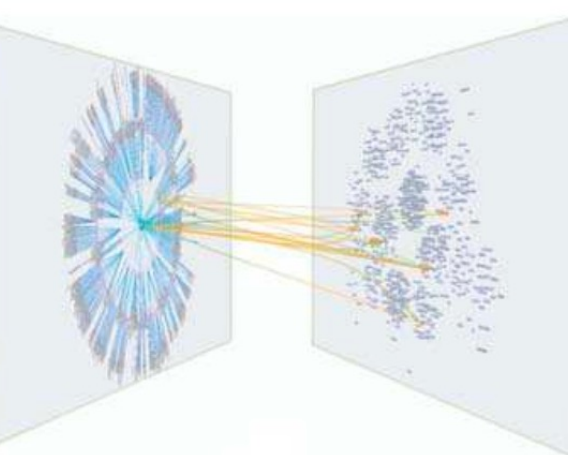
CAMBIAMENTO DEL TIPO DI VISUALIZZAZIONE

La modifica del layout influisce sulla posizione degli elementi di dati sullo schermo. Può essere eseguito modificando il layout automaticamente o spostando manualmente i nodi. È inoltre possibile modificare il tipo di rappresentazione dei dati, questa modifica può interessare l'intera visualizzazione dei dati o solo una parte e permette di ottenere nuove informazioni sui dati.



Analisi dei grafi [1]

- ▶ L'analisi algoritmica dei grafi, è utile durante tutte le fasi del processo di analisi visiva del grafo. Le tecniche pertinenti consentono, ad esempio, di ridurre un grafo di grandi dimensioni a un grafo più piccolo prima della visualizzazione, di cercare strutture di grafi specifiche di interesse o di trovare somiglianze e dissomiglianze per generare viste di grafi comparativi.
- ▶ Nella maggior parte delle attività dell'utente, l'analisi delle relazioni tra le entità nel grafo e la valutazione della struttura del grafo globale giocano un ruolo chiave. Queste attività possono essere efficacemente supportate da una combinazione di analisi algoritmiche di grafi e visualizzazione interattiva. I metodi algoritmici consentono, ad esempio:
 - di calcolare le proprietà del nodo/arco
 - Identificare nodi importanti: nelle reti, alcuni nodi svolgono un ruolo specifico a causa della loro posizione all'interno della rete. La codifica a colori di nodi o bordi in base a valori di metrica o la visualizzazione di metriche e reti in più viste collegate (come elenchi, grafi a dispersione o coordinate parallele) vengono utilizzate a questo riguardo. Offrono la possibilità di scegliere interattivamente le metriche di interesse e di filtrare/evidenziare i nodi in base a queste metriche.
 - Analisi delle connessioni tra i nodi: oltre a concentrarsi sui singoli nodi, è possibile analizzare le relazioni tra due nodi, tipicamente mediante il calcolo e l'evidenziazione dei percorsi più brevi tra le entità. Di solito, tale analisi è combinata con la selezione interattiva di due entità di interesse
 - Analisi della struttura del grafo: In molte applicazioni, tipi specifici di sottostrutture svolgono un ruolo importante. Ad esempio, nei social network, le clique identificano comunità altamente connesse. Il tipo di struttura può essere scelto interattivamente dall'utente per supportare vari compiti analitici.
 - di identificare i cluster nei grafi, i cui risultati sono visualizzati in modo interattivo.



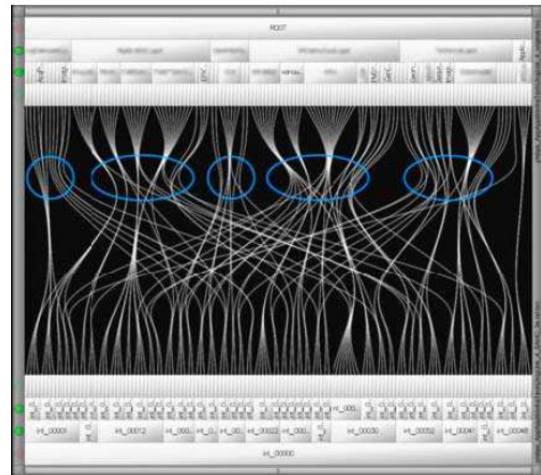
(a) One-to-one graph matching

Confronto tra grafi [1]

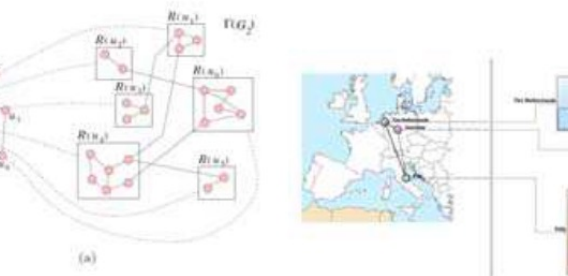
- Un compito analitico particolarmente importante è l'esame delle somiglianze e delle differenze tra grafi multipli, con particolare attenzione agli aspetti strutturali. Tale differenza può essere identificata dalle etichette identiche dei nodi in entrambi i grafi o da algoritmi di corrispondenza dei grafi. Dopo la corrispondenza, viene utilizzata la visualizzazione per esplorare le differenze.

- Confronto dei nodi uno-a-uno:** il compito più comune nel confronto dei grafi è l'abbinamento di singoli nodi da un grafo ai singoli nodi del secondo grafo. Mostra entrambi i grafi su piani separati e disegna i collegamenti corrispondenti tra i nodi corrispondenti. Per il confronto delle gerarchie, si disegnano le due gerarchie in parti opposte della visualizzazione e sul collegamento dei loro nodi foglia. In entrambi i casi, la visibilità dei link corrispondenti può essere aumentata mediante l'edge bundling.

- Confronto da uno a molti** nodi di due grafi: il confronto da uno a molti nodi riguarda la corrispondenza di un nodo in un grafo con molti nodi in un altro grafo. Si visualizzano queste connessioni uno-a-molti con una bassa sovrapposizione di collegamenti.



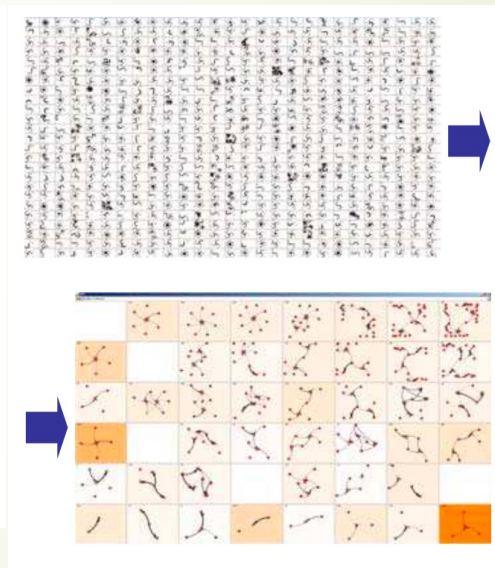
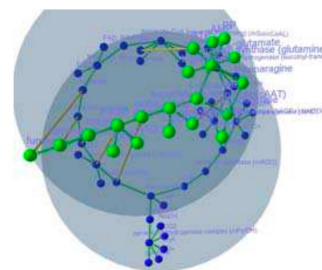
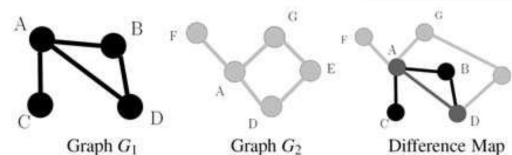
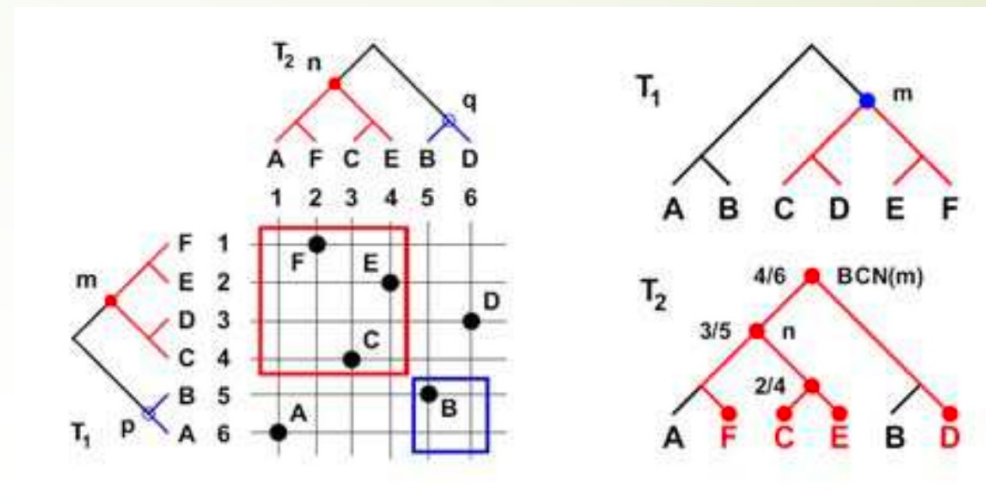
(b) One-to-one hierarchy matching



(c) One-to-many graph matching

- Differenze strutturali tra due grafi:** quando si analizzano le differenze strutturali tra due grafi, gli analisti sono spesso interessati a identificare quali collegamenti o parti del grafo corrispondono o differiscono dall'altro. Per l'analisi degli alberi, può essere utilizzata l'analisi e l'evidenziazione delle differenze strutturali tra due alberi. Per i grafi generali, si possono utilizzare sia viste di grafi multi-livello, sia la sovrapposizione di due reti con l'evidenziazione di parti strutturali comuni, oppure si utilizza l'aggregazione dei grafi e il filtraggio per rivelare le differenze strutturali.

Somiglianza strutturale tra grafi multipli: è spesso basato sulla loro descrizione da parte di diverse proprietà del grafo come la dimensione del grafo, la densità, la connessione ecc. Queste proprietà possono essere utilizzate per l'esplorazione di grandi insiemi di grafi, o per la determinazione della somiglianza strutturale tra grafi. La somiglianza del grafo può servire come input per il raggruppamento di grafi (raggruppamento di grafi simili). Il clustering aiuta a ottenere una panoramica dei tipi di grafi in database di grafi di grandi dimensioni.





Sfide future

- Problemi di scalabilità nel disegno di grafi
- Tipi di grafi per la rappresentazione di grafi dinamici e composti
- Edge visualization
- Espansione dei sistemi di analisi visiva con maggiore controllo da parte degli utenti
- Integrazione di vari tipi di dati nell'analisi visiva
- Affrontare nuove attività analitiche per l'esame di somiglianze e differenza tra grafi
- Sistemi di Visual Analytics adattabili
- Valutazione delle tecniche di Visual Analytics
- Tassonomie e benchmark



BIBLIOGRAFIA

- [1] Von Landesberger, Tatiana, et al. "Visual analysis of large graphs: state-of-the-art and future research challenges." *Computer graphics forum*. Vol. 30. No. 6. Oxford, UK: Blackwell Publishing Ltd, 2011.
- [2] Beck, Fabian, Michael Burch, and Stephan Diehl. "Towards an aesthetic dimensions framework for dynamic graph visualisations." *2009 13th international conference information visualisation*. IEEE, 2009
- [3] Shneiderman, Ben, and Aleks Aris. "Network visualization by semantic substrates." *IEEE transactions on visualization and computer graphics* 12.5 (2006): 733-740.
- [4] Ma, Kwan-Liu, and Chris W. Muedler. "Large-scale graph visualization and analytics." *Computer* 46.7 (2013): 39-46.
- [5] Niggemann, Oliver. *Visual data mining of graph based data*. Diss. Paderborn, Univ., Diss., 2001, 2001.



Visual Clustering

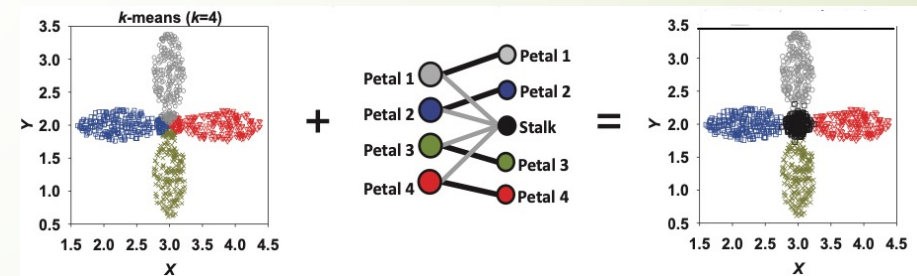
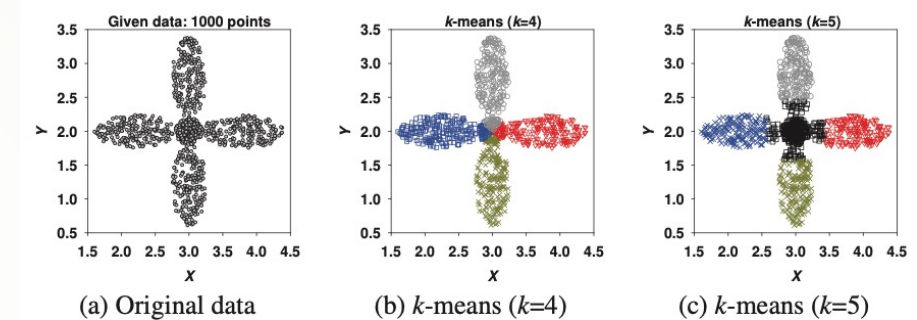


CLUSTERING

- ▶ Durante l'analisi dei dati, un'attività ampiamente utilizzata consiste nel trovare gruppi di oggetti dell'insieme di dati che condividono caratteristiche simili. In tal modo, gli utenti acquisiscono informazioni dettagliate sui propri dati, li comprendono e riducono persino la natura ad alta dimensionalità. Questi gruppi concettuali sono comunemente chiamati **cluster** [1]
- ▶ Poiché gli strumenti di clustering mancano di input specifici del dominio e dell'utente, i risultati non sono sempre pertinenti o convenienti per l'utente finale
- ▶ Il raggruppamento è un compito intrinsecamente soggettivo poiché dipende dall'interpretazione, dalla necessità e dall'interesse dell'utente
- ▶ I dati del mondo reale possono avere diversi raggruppamenti plausibili, e un raggruppamento non supervisionato, non ha modo di stabilire un raggruppamento adatto alle esigenze dell'utente poiché ciò richiede una conoscenza del dominio esterno
- ▶ Vi è una crescente necessità di metodi che coinvolgono gli utenti finali direttamente nel processo di clustering per adattarlo a specifici domini applicativi e consentirgli di adattarsi continuamente alle loro esigenze

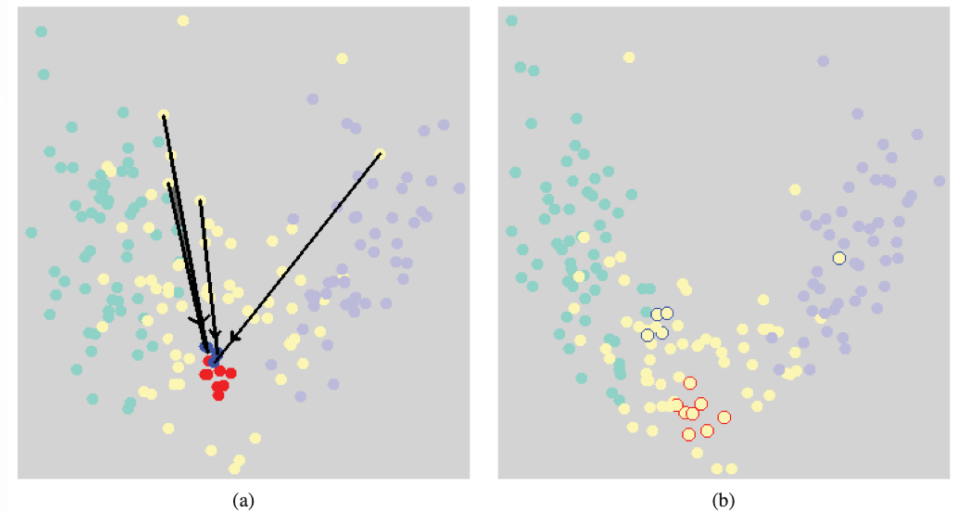
ESEMPIO [5]

- utilizziamo un set di dati sintetico composto da 1000 punti bidimensionali. Il set di dati è composto da quattro petali e uno stelo ciascuno contenente 200 punti.
- Quando l'utente applica un semplice clustering k-mean, con un'impostazione di quattro clusters (cioè $k = 4$), il fiore viene diviso in quattro parti dove i petali sono effettivamente in cluster diversi, ma ciascuno dei petali occupa anche un quarto delle punte del gambo del fiore.
- Quando viene utilizzata un'impostazione di cinque cluster, l'utente ottiene il clustering. È evidente che i cinque clusters generati da k-mean non sono in grado di differenziare nettamente il gambo dai petali.
- L'utente può fornire un input all'algorithmo per quanto riguarda il risultato atteso. Leggendo da sinistra a destra, vediamo che l'utente si aspetta che i quattro cluster vengano suddivisi (sparpagliati) in cinque cluster. Leggendo da destra a sinistra, vediamo che il gambo dovrebbe raccogliere punti da tutti i cluster attuali, ma c'è una corrispondenza uno a uno tra i petali desiderati e i petali originali. I risultati di un tale raggruppamento forniscono petali e stelo ben separati, a differenza del risultato fornito da semplici k-means con $k = 5$



INTERAZIONE CON IL RISULTATO

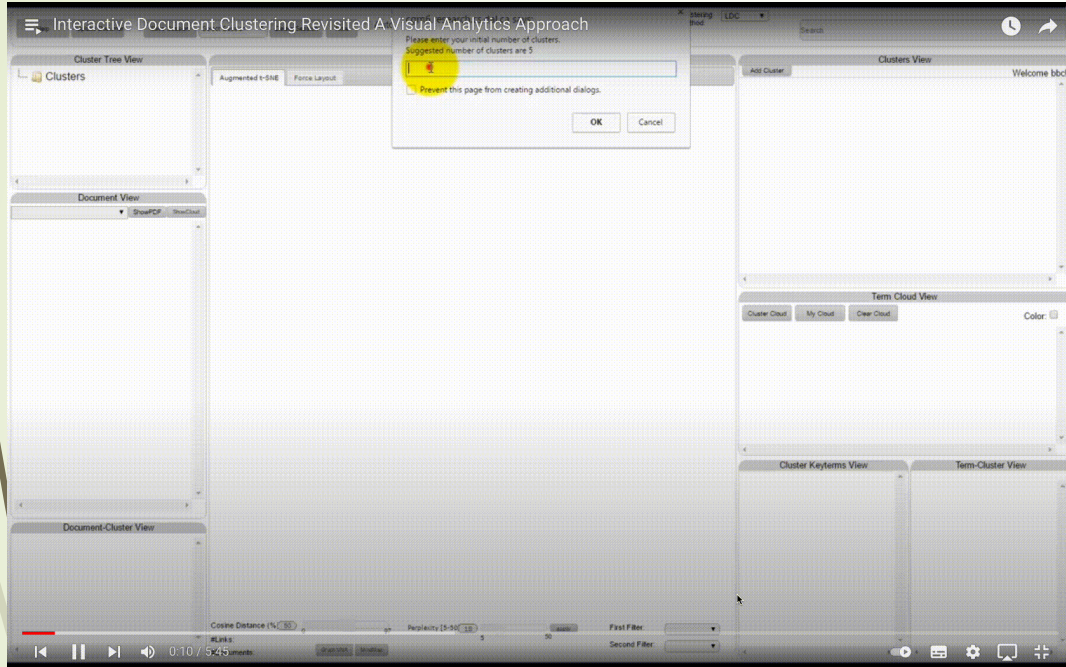
- ▶ la funzionalità per consentire all'utente di perfezionare in modo iterativo e dinamico i risultati del clustering attraverso operazioni significative e della massima importanza per ottenere un processo di clustering più basato sull'utente.
- ▶ Il tipo più complesso di interazione prevede che la macchina esegua il clustering (iniziale), presenti i risultati all'utente tramite la visualizzazione e solo successivamente dia all'utente la possibilità di interagire con i risultati del clustering [3].
- ▶ L'utente può:
 - spostare manualmente i punti dati tra i cluster
 - interagire con le visualizzazioni
 - rimuovere i membri del cluster, dividere i cluster in cluster più piccoli, unire i cluster e ridistribuire i membri del cluster
 - selezionare una parte del set di dati
 - interagire con i risultati del clustering fornendo feedback sotto forma di selezione, scarto e messa a punto dei candidati del cluster



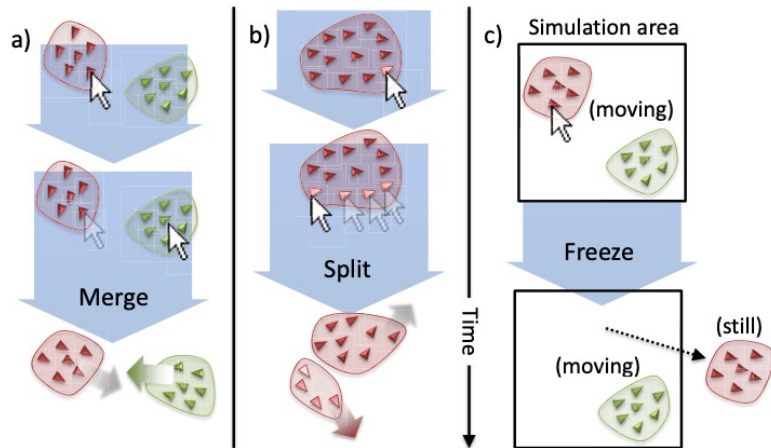
INTERAZIONE CON I PARAMETRI

- ▶ Una volta eseguito e visualizzato il clustering iniziale, gli utenti devono essere in grado di rieseguire il clustering utilizzando parametri diversi.
- ▶ i singoli utenti devono decidere quali parametri regolare per migliorare i risultati. ad esempio, regolare il numero di cluster o i parametri della soglia di somiglianza
- ▶ gli utenti possono:
 - eseguire la riduzione delle dimensioni
 - aggiornare in modo interattivo le opzioni di pre-elaborazione per i dati e il clustering
 - applicare diversi algoritmi di clustering
 - confrontare visivamente gli effetti di queste modifiche su coordinate parallele, grafici a dispersione e visualizzazioni di etichette di cluster
- ▶ la visualizzazione utilizza contemporaneamente grafici a dispersione, coordinate parallele, pesi dei singoli parametri dimensionali e granularità diverse. L'evidenziazione di un elemento di dati in un fotogramma fa sì che lo stesso elemento venga evidenziato in altre viste, consentendo un facile confronto tra diverse impostazioni. L'utente può anche riorganizzare questi frame avvicinando i dati con forme simili.

OPERAZIONI DI INTERAZIONE [4]



- **Confrontare somiglianza**
- **Modificare il numero di cluster (o altri parametri)**
- **Rimuovere i cluster**
- **Aggiungere cluster:** Molte applicazioni consentono all'utente di aggiungere cluster, prima del processo di raggruppamento o dopo aver esaminato i risultati.
- **Correzione dell'errore**
- **Manipolare features**
- **Dividere e unire cluster**





BIBLIOGRAFIA

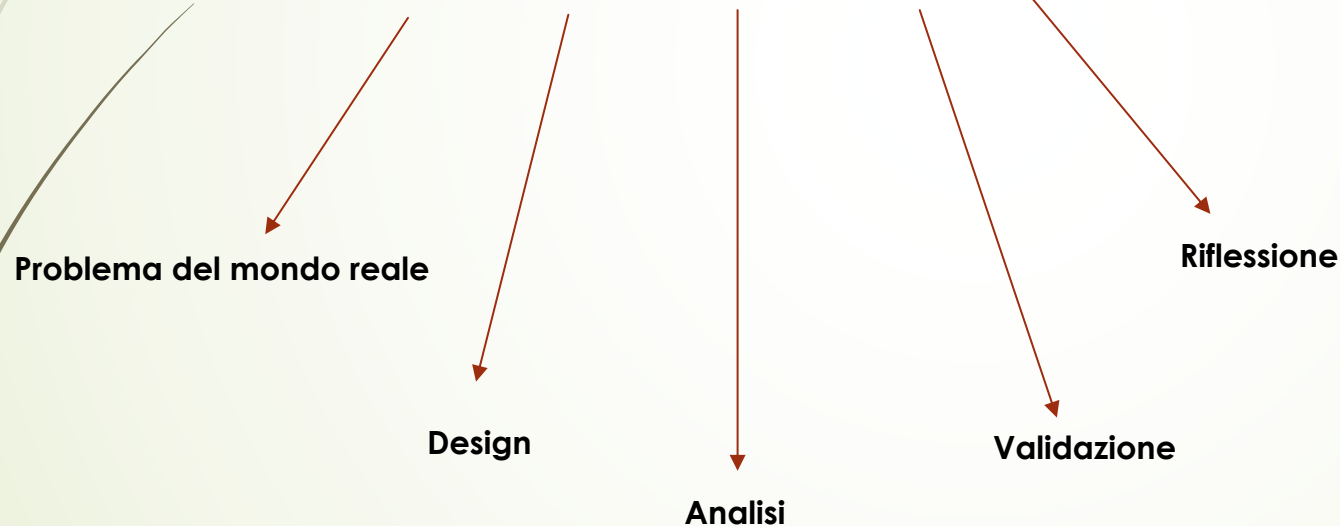
- [1] Bae, Juhee, et al. "Interactive clustering: A comprehensive review." *ACM Computing Surveys (CSUR)* 53.1 (2020): 1-39.
- [2] Brown, Eli T., et al. "Dis-function: Learning distance functions interactively." *2012 IEEE conference on visual analytics science and technology (VAST)*. IEEE, 2012.
- [3] Cao, Nan, et al. "Dicon: Interactive visual analysis of multidimensional clusters." *IEEE transactions on visualization and computer graphics* 17.12 (2011): 2581-2590.
- [4] Erra, Ugo, Bernardino Frola, and Vittorio Scarano. "An interactive bio-inspired approach to clustering and visualizing datasets." *2011 15th International Conference on Information Visualisation*. IEEE, 2011.
- [5] Hossain, M. Shahriar, et al. "Scatter/gather clustering: Flexibly incorporating user feedback to steer clustering results." *IEEE Transactions on Visualization and Computer Graphics* 18.12 (2012): 2829-2838.
- [6] Lee, Hanseung, et al. "iVisClustering: An interactive visual document clustering via topic modeling." *Computer graphics forum*. Vol. 31. No. 3pt3. Oxford, UK: Blackwell Publishing Ltd, 2012.



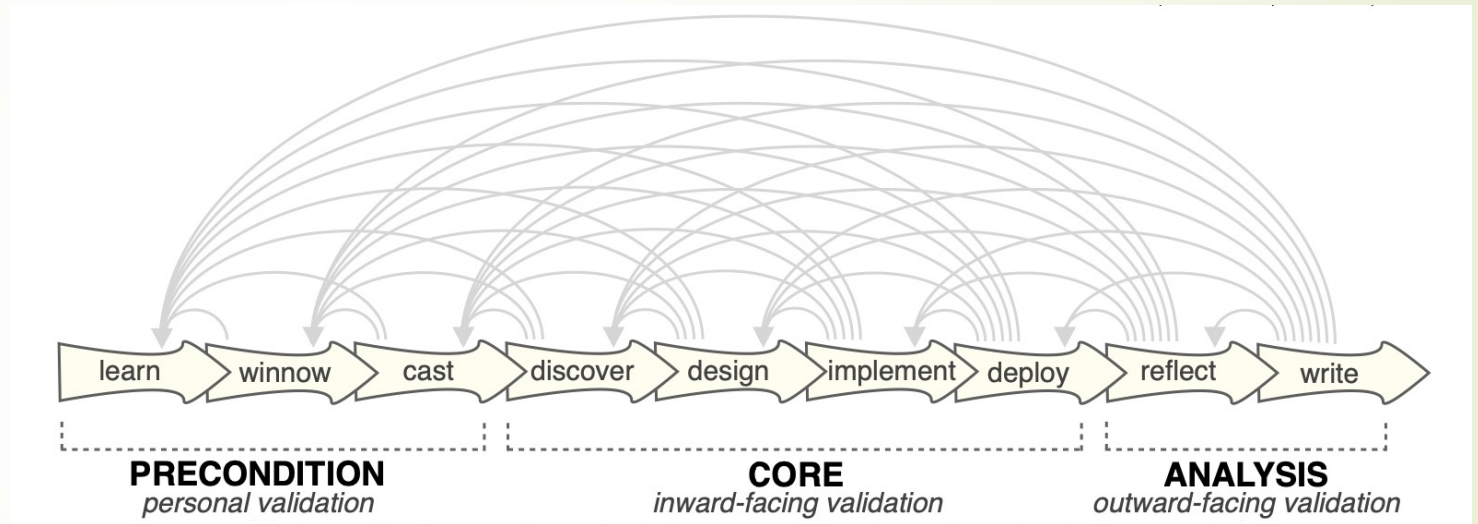
Tecniche di valutazione

STUDIO DI PROGETTAZIONE DELLA VISUALIZZAZIONE [4]

- Uno studio di progettazione è un progetto in cui i ricercatori della visualizzazione analizzano uno specifico problema del mondo reale affrontato dagli esperti di dominio, progettano un sistema di visualizzazione che supporti la risoluzione di questo problema, convalidano il progetto e riflettono sulle lezioni apprese al fine di perfezionare le linee guida di progettazione della visualizzazione.
- Uno studio di progettazione implica:



1. una fase preliminare che descrive cosa deve essere fatto prima di iniziare uno studio di progettazione;
2. una fase centrale che presenta le fasi principali della conduzione di uno studio progettuale;
3. una fase di analisi che illustra il ragionamento analitico alla fine.





PRECONDITION

1. **LEARN**: una preconditione cruciale per condurre uno studio di progettazione efficace è una solida conoscenza della letteratura sulla visualizzazione, comprese le tecniche di codifica e interazione visiva, linee guida di progettazione e metodi di valutazione.
2. **WINNOW**: L'obiettivo di questa fase è identificare le collaborazioni più promettenti.
3. **CAST**: identificare i ruoli dei collaboratori

CORE

1. **DISCOVER**: A questo livello, un designer di visualizzazione deve conoscere le attività e i dati di un determinato dominio. L'output di questo livello è spesso un insieme dettagliato di domande poste o azioni eseguite da utenti target per una raccolta eterogenea dei dati
2. **DESIGN**: Dopo aver raggiunto una comprensione condivisa di un problema con gli esperti di dominio nella fase di individuazione, il ricercatore della visualizzazione può iniziare a progettare una soluzione di visualizzazione. Questa fase consiste nel mappare problemi e dati nel vocabolario del dominio in una descrizione più astratta e generica. L'output di questo livello è una descrizione delle operazioni e dei tipi di dati che sono l'input richiesto per prendere decisioni di codifica visiva al livello successivo. L'altro aspetto di questa fase consiste nel trasformare i dati grezzi nei tipi di dati che possono essere affrontati dalle tecniche di visualizzazione. L'obiettivo è trovare il giusto tipo di dati in modo che una sua rappresentazione visiva affronti il problema che spesso richiede la trasformazione di dati grezzi in un tipo derivato di una forma diversa
3. **IMPLEMENT**: creare un algoritmo per eseguire automaticamente la codifica visiva e le interazioni
4. **DEPLOY**: distribuzione di uno strumento e la raccolta di feedback sul suo utilizzo in natura. L'obiettivo principale nella convalida di un sistema distribuito è scoprire se gli esperti di dominio sono effettivamente aiutati dalla nuova soluzione.

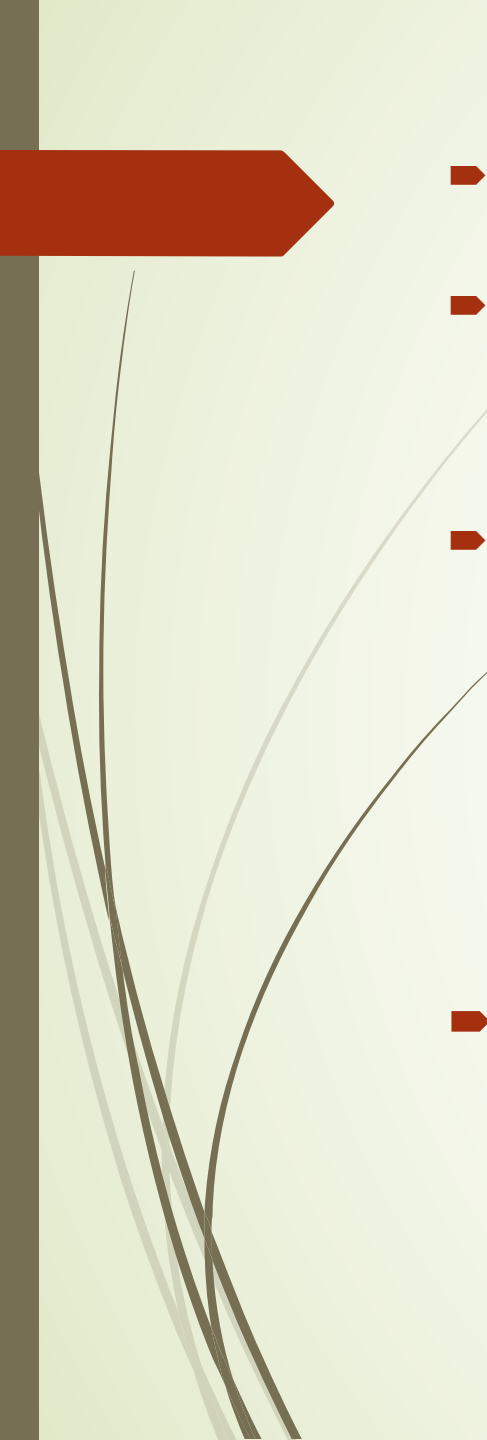
ANALYSIS

1. REFLECT: CONFIRM, REFINE, REJECT, PROPOSE GUIDELINES

È particolarmente informativo per migliorare le linee guida di progettazione attualmente disponibili: sulla base di nuove scoperte, le linee guida proposte in precedenza possono essere confermate, fornendo ulteriori prove della loro utilità; raffinato o ampliato con nuove intuizioni; rifiutati quando vengono applicati ma non funzionano; o potrebbero essere proposte nuove linee guida.

- **Problema caratterizzato erroneamente** → Intervistare e osservare il pubblico di destinazione per verificare la caratterizzazione
- **Le operazioni e i tipi di dati scelti non risolvono il problema** → Far provare lo strumento a un membro della comunità di utenti target per raccogliere prove che lo strumento sia effettivamente utile
→ Osservare e documentare in che modo il pubblico di destinazione utilizza il sistema
- **Il design scelto non è efficace nel comunicare l'astrazione desiderata alla persona che utilizza il sistema** → Misurazione quantitativa dei risultati forniti dal sistema. Ad esempio, il numero di incroci e piegature degli archi
- **L'algorithm non è ottimale in termini di tempo o prestazioni di memoria** → Analizzare la complessità computazionale dell'algorithm
- **L'algorithm potrebbe non essere corretto** (non soddisfa le specifiche per la codifica visiva o il design dell'interazione) → Presentare immagini o video create dall'algorithm riguardanti il suo utilizzo, con cui il lettore può vedere direttamente che gli obiettivi di correttezza dell'algorithm sono stati raggiunti

2. WRITE: Scrivere un documento di studio di progettazione

- 
- ▶ Gli strumenti di VDM hanno la particolarità di rappresentare i dati in una forma visiva con cui gli utenti possono interagire per ottenere informazioni.
 - ▶ Un elevato livello di qualità delle informazioni, è essenziale per gli utenti. Questo può essere ottenuto attraverso un'interazione di alta qualità con il sistema e una visualizzazione di alta qualità dei dati;
 - ▶ La misurazione della qualità d'uso, implica la misurazione di aspetti quali:
 - Efficacia
 - Efficienza
 - Soddisfazione degli utentinel raggiungimento degli obiettivi specificati in uno specifico contesto di utilizzo;
 - ▶ L'analisi di questi sistemi viene fatta su 3 livelli [7]:
 1. Visualizzazione
 2. Interazione
 3. Informazione



Qualità d'uso

- ▶ La qualità d'uso di uno strumento di VDM è definita come la totalità delle caratteristiche dello strumento che riflettono la sua capacità di soddisfare le esigenze degli utenti.
- ▶ Le caratteristiche principali di uno strumento di VDM che influenzano la soddisfazione dell'utente sono:
 - Visualizzazione dei dati
 - Interazione con il sistema
 - Informazioni ottenute



Qualità di visualizzazione

- È la capacità del sistema di trasformare i dati di input e di renderli accessibili. I problemi da analizzare sono:
 - 1. Impostazioni iniziali**
si riferiscono ai requisiti relativi al formato dei dati di input, al grado di astrazione dei dati e all'impostazione dei parametri per la visualizzazione
 - 2. Visualizzazione dei dati**
possibilità di visualizzare la struttura dei dati, la variazione, il contenuto e il confronto tra dati
 - 3. Attività di esplorazione**
possibilità di visualizzare una panoramica dei dati, dettagli, applicazione di filtri, dettagli su richiesta e relazioni
 - 4. Funzioni di reporting**
funzioni di sistema che consentono all'utente di trasferire i risultati all'esterno dell'applicazione per vari scopi

Qualità di interazione

- Riguarda le valutazioni su quanto gli utenti considerano il sistema facile da usare e da apprendere, accurato (affidabile), efficace ed efficiente.
- Le tecniche di interazione sono classificate in 5 gruppi:
 1. **Facilità d'uso**
capacità del sistema di essere facilmente controllabile dall'utente e di fornire all'utente libertà di azione
 2. **Apprendimento**
facilità e rapidità con cui gli utenti imparano ad usare il sistema per poter eseguire le attività desiderate
 3. **Precisione** (affidabilità)
frequenza e gravità di errori o guasti del sistema
 4. **Efficienza**
misura il grado con cui gli utenti ritengono che il sistema li aiuti nel loro lavoro (personalizzare le azioni frequenti, migliorare le prestazioni lavorative, ricevere risposte rapide alle domande)
 5. **Supportabilità**
possibilità di accesso degli utenti alla documentazione e al supporto quando necessario



Qualità di informazione

- ▶ Valutare la misura in cui gli utenti sono soddisfatti delle informazioni ottenute
- ▶ Le 4 caratteristiche delle informazioni che gli utenti potrebbero richiedere sono:
 1. **Ricchezza di informazioni**
significa completezza, utilità, e interesse. Inoltre, le informazioni ottenute, devono corrispondere alle esigenze e alle aspettative degli utenti
 2. **Accuratezza delle informazioni**
riguarda il grado in cui le informazioni sono precise, corrette e coerenti con le conoscenze degli utenti
 3. **Chiarezza delle informazioni**
le informazioni sono presentate in modo chiaro e comprensibile e consentono interpretazioni e deduzioni
 4. **Novità dell'informazione**
riflette la caratteristica dell'informazione di essere nuova e aggiornata



► Altre valutazioni che possono essere fatte su un sistema di VDM riguardano [5]:

1. Adattabilità

capacità del sistema di adattarsi alle esigenze dell'utente senza alcun intervento esplicito da parte dell'utente o la sua capacità di reagire in base al contesto e alle esigenze e preferenze dell'utente.

2. Curabilità

capacità dell'utente di correggere una situazione non desiderata.

3. Gestione degli errori

possibilità di evitare o ridurre gli errori e di correggerli quando si verificano. Ad esempio, avere la possibilità di proporre un altro metodo di analisi, nel caso in cui quello scelto non vada a buon fine, senza avere un crash del sistema.

4. Feedback

avere informazioni che mostrano all'utente che le operazioni sono in corso, il rapporto sullo stato di avanzamento delle operazioni e una risposta che informi l'utente dell'azione compiuta e dei possibili risultati

5. Guida per l'utente

avere modi disponibili per consigliare, orientare, informare, istruire e guidare gli utenti durante le loro interazioni con il computer.

6. Molteplicità dei risultati

capacità del sistema di fornire diversi tipi di visualizzazione per lo stesso insieme di dati



BIBLIOGRAFIA

- [1] Dorina. "Evaluating multidimensional visualization techniques in data mining tasks." (2008).
- [2] Badjio, Edwige Fangseu. "Quality evaluation of visual data mining tools." *IADIS International Conference Applied Computing 2005*. 2005. Endert, Alex, et al. "The human is the loop: new directions for visual analytics." *Journal of intelligent information systems* 43.3 (2014): 411-435.
- [3] Munzner, Tamara. "A nested model for visualization design and validation." *IEEE transactions on visualization and computer graphics* 15.6 (2009): 921-928.
- [4] Sedlmair, Michael, Miriah Meyer, and Tamara Munzner. "Design study methodology: Reflections from the trenches and the stacks." *IEEE transactions on visualization and computer graphics* 18.12 (2012): 2431-2440.
- [5] Badjio, Edwige P. Fangseu, and François Poulet. "Usability of Visual Data Mining Tools." *ICEIS* (5). 2004.
- [6] Badjio, Edwige P. Fangseu, and François Poulet. "Visual Data Mining Tools: Quality Metrics Definition and Application." *ICEIS*. 2005.
- [7] Marghescu, Dorina, Mikko Rajanen, and Barbro Back. "Evaluating the quality of use of visual data-mining tools." *Proc. 11th Europ. Conf. IT Evaluation*. 2004.