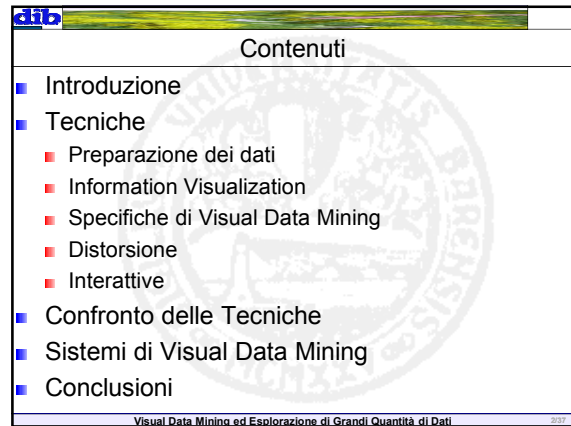




**Visual Data Mining
ed esplorazione di grandi quantità di dati**

Paolo Buono
buono@di.uniba.it

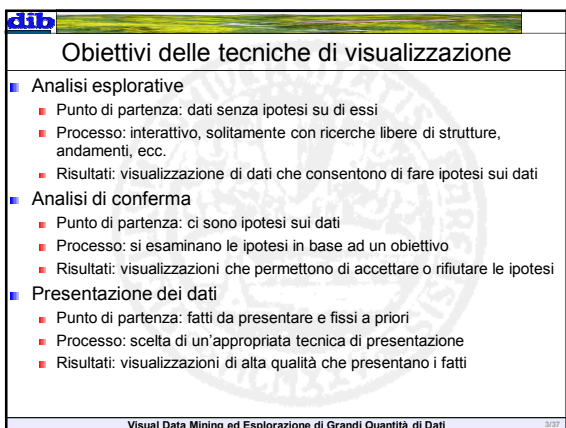
Dipartimento di Informatica
Università degli Studi di Bari



Contenuti

- Introduzione
- Tecniche
 - Preparazione dei dati
 - Information Visualization
 - Specifiche di Visual Data Mining
 - Distorsione
 - Interattive
- Confronto delle Tecniche
- Sistemi di Visual Data Mining
- Conclusioni

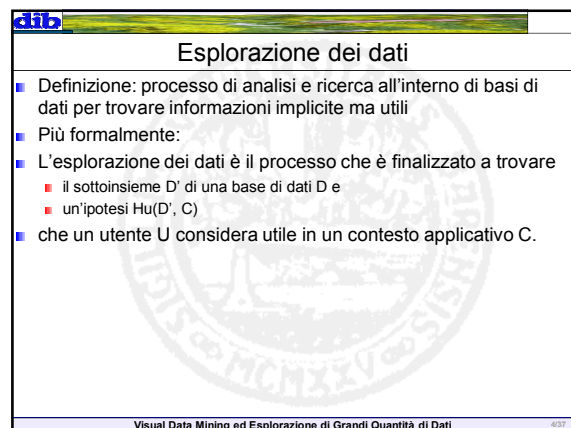
Visual Data Mining ed Esplorazione di Grandi Quantità di Dati



Obiettivi delle tecniche di visualizzazione

- Analisi esplorative
 - Punto di partenza: dati senza ipotesi su di essi
 - Processo: interattivo, solitamente con ricerche libere di strutture, andamenti, ecc.
 - Risultati: visualizzazione di dati che consentono di fare ipotesi sui dati
- Analisi di conferma
 - Punto di partenza: ci sono ipotesi sui dati
 - Processo: si esaminano le ipotesi in base ad un obiettivo
 - Risultati: visualizzazioni che permettono di accettare o rifiutare le ipotesi
- Presentazione dei dati
 - Punto di partenza: fatti da presentare e fissi a priori
 - Processo: scelta di un'appropriata tecnica di presentazione
 - Risultati: visualizzazioni di alta qualità che presentano i fatti

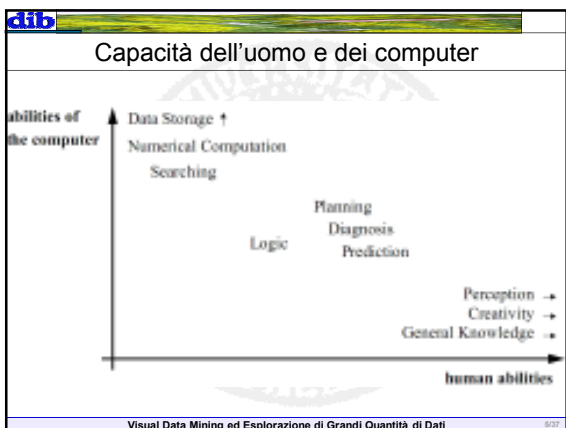
Visual Data Mining ed Esplorazione di Grandi Quantità di Dati



Esplorazione dei dati

- Definizione: processo di analisi e ricerca all'interno di basi di dati per trovare informazioni implicite ma utili
- Più formalmente:
- L'esplorazione dei dati è il processo che è finalizzato a trovare
 - il sottoinsieme D' di una base di dati D e
 - un'ipotesi $Hu(D', C)$
- che un utente U considera utile in un contesto applicativo C .

Visual Data Mining ed Esplorazione di Grandi Quantità di Dati



Capacità dell'uomo e dei computer

Abilities of the computer ↑

Data Storage ↑

Numerical Computation

Searching

Logic

Planning

Diagnosis

Prediction

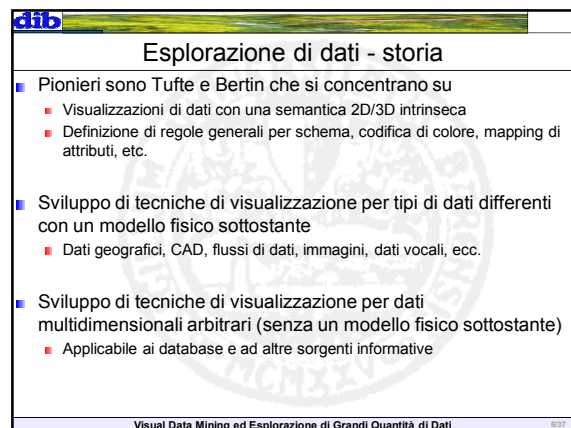
Perception →

Creativity →

General Knowledge →

human abilities →

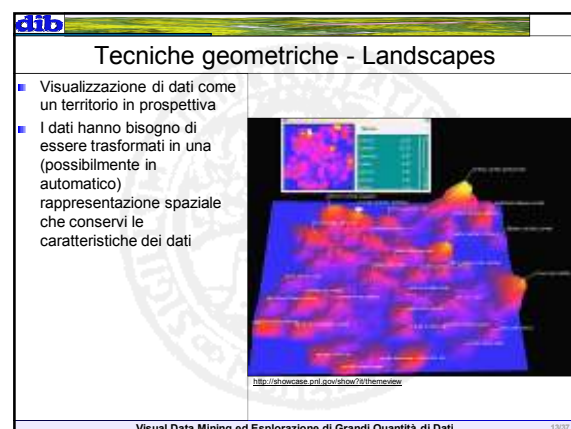
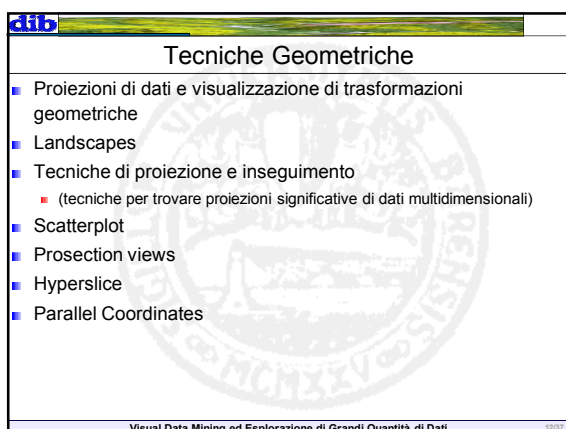
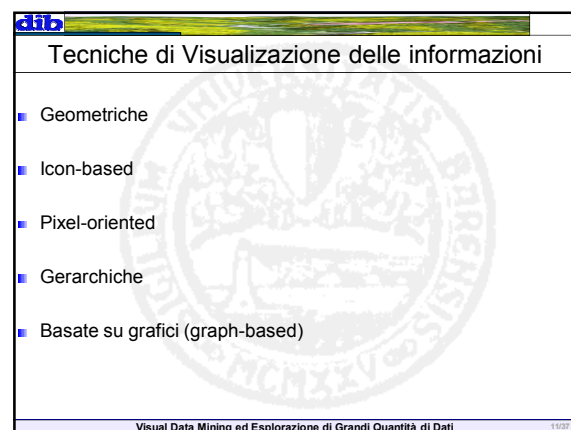
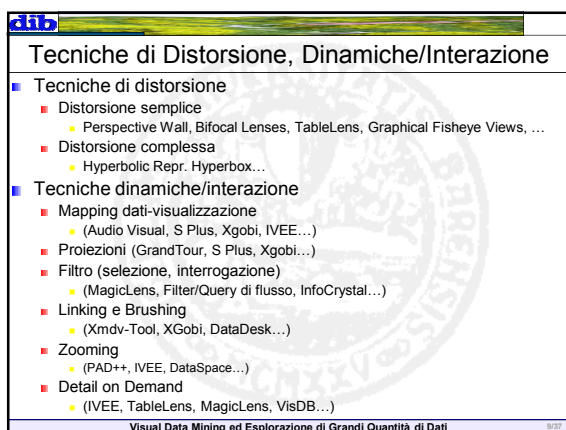
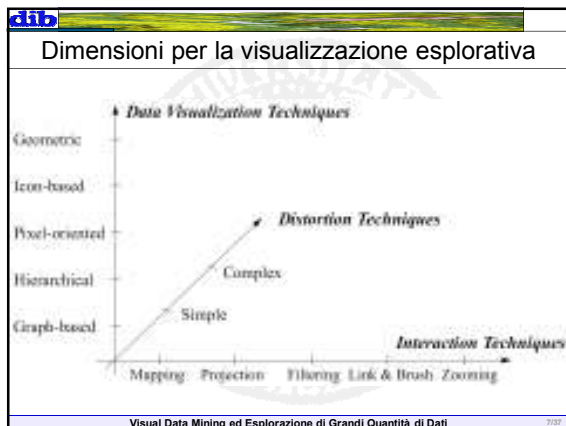
Visual Data Mining ed Esplorazione di Grandi Quantità di Dati

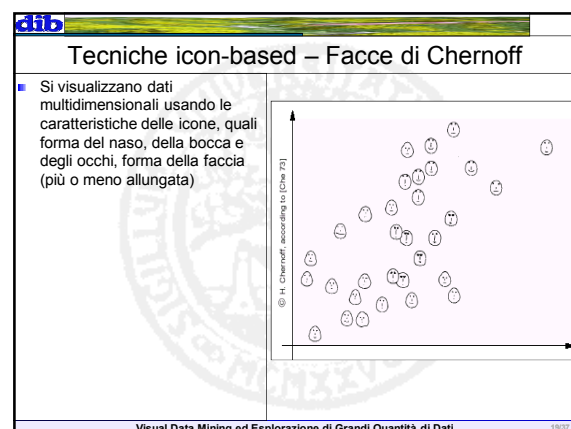
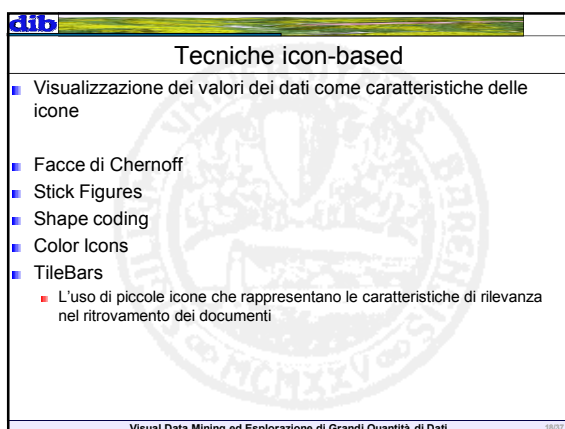
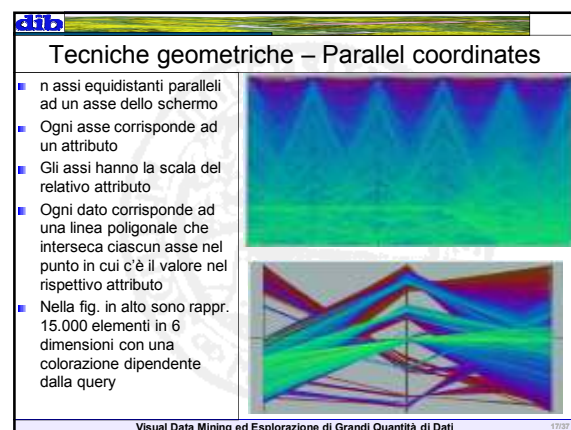
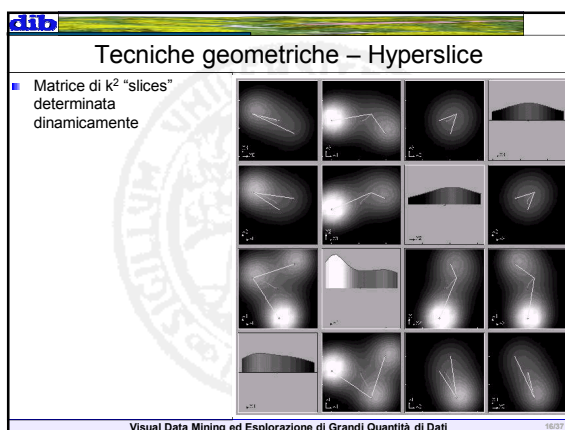
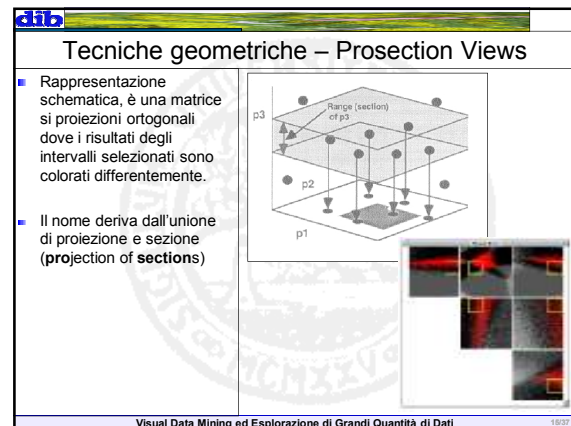
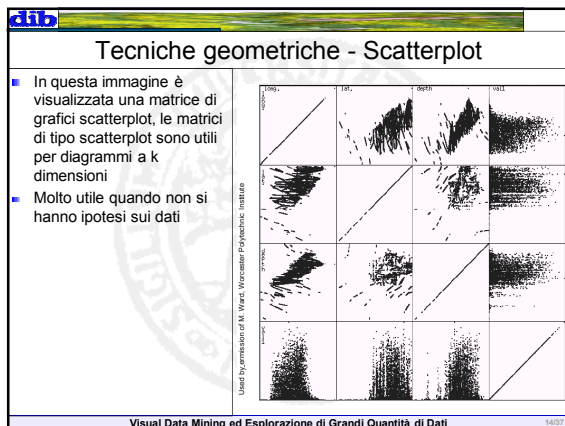


Esplorazione di dati - storia

- Pionieri sono Tufte e Bertin che si concentrano su
 - Visualizzazioni di dati con una semantica 2D/3D intrinseca
 - Definizione di regole generali per schema, codifica di colore, mapping di attributi, etc.
- Sviluppo di tecniche di visualizzazione per tipi di dati differenti con un modello fisico sottostante
 - Dati geografici, CAD, flussi di dati, immagini, dati vocali, ecc.
- Sviluppo di tecniche di visualizzazione per dati multidimensionali arbitrari (senza un modello fisico sottostante)
 - Applicabile ai database e ad altre sorgenti informative

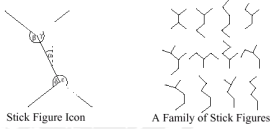
Visual Data Mining ed Esplorazione di Grandi Quantità di Dati





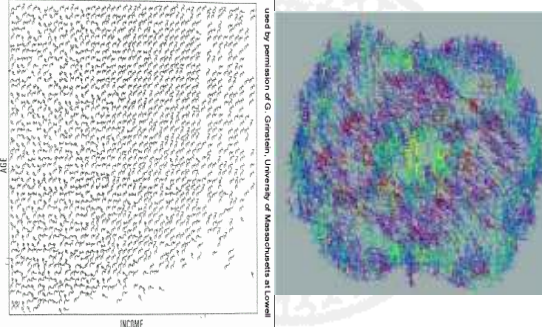
Tecniche icon-based – Stick figures

- Visualizzazione di dati multidimensionali usando le caratteristiche delle icone
- Si effettua il mapping di due attributi di un dato sugli assi da visualizzare e i restanti attributi sono codificati con delle barrette angolate
- La texture (la disposizione di tutte le stick figures lungo l'immagine risultante) mostra certe caratteristiche dei dati



Visual Data Mining ed Esplorazione di Grandi Quantità di Dati

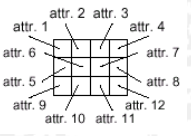
Tecniche icon-based – Stick figures



Visual Data Mining ed Esplorazione di Grandi Quantità di Dati

Tecniche icon-based – Shaped coding

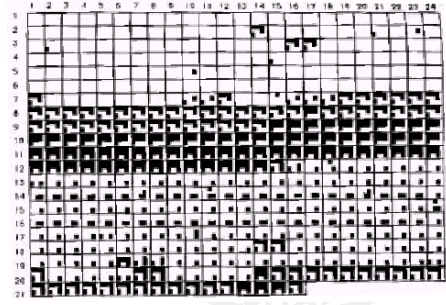
- I dati sono visualizzati usando piccoli insiemi di campi
- Ciascun campo rappresenta un valore di un attributo
- Ecco come possono essere codificati:



- L'insieme di dati è trattato riga per riga secondo un certo criterio (per esempio il tempo per i dati di serie temporali)

Visual Data Mining ed Esplorazione di Grandi Quantità di Dati

Tecniche icon-based – Shaped coding

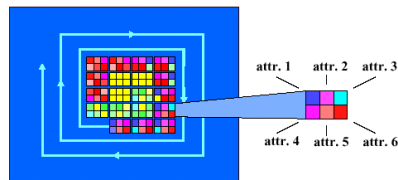


time series of NASA earth observation data

Visual Data Mining ed Esplorazione di Grandi Quantità di Dati

Tecniche icon-based – Color Icons

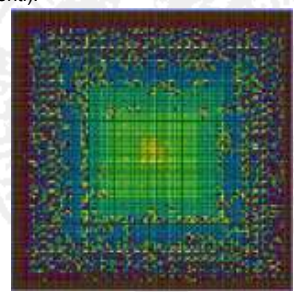
- I dati sono visualizzati usando icone colorate (color icons)
- Le color icons sono vettori di campi colorati che rappresentano un valore di un attributo, possono essere visti come un avanzamento rispetto alla codifica shaped coding
- I criteri con cui sono sistemati i colori dipendono dalla particolare query



Visual Data Mining ed Esplorazione di Grandi Quantità di Dati

Tecniche icon-based – Color Icons

- Ecco un esempio di dati casuali contenenti diversi cluster (raggruppamenti):



Visual Data Mining ed Esplorazione di Grandi Quantità di Dati

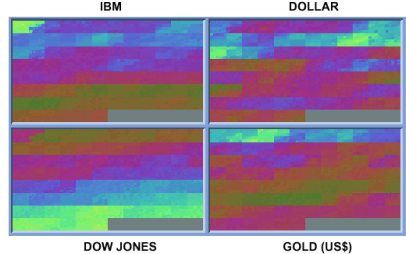
Tecniche pixel-oriented

- Consentono di esplorare i dati e di effettuare analisi su quantità di dati multidimensionali molto grandi
- Sfruttano le abilità percettive umane fornendo un'adeguata presentazione dell'informazione
- L'uso di ciascun pixel visualizza un valore di un dato, questo permette di visualizzare circa 1.3 milioni di valori in un dato istante di tempo
- L'idea è di mettere in relazione ciascun dato con un pixel colorato e di organizzare i dati adeguatamente, si ottiene così la visualizzazione orientata al pixel

Visual Data Mining ed Esplorazione di Grandi Quantità di Dati 39/37

Tecniche pixel-oriented

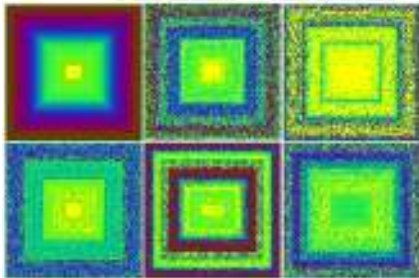
- Ecco un esempio di tecnica pixel-oriented applicata a dati finanziari



Visual Data Mining ed Esplorazione di Grandi Quantità di Dati 39/37

Tecniche pixel-oriented

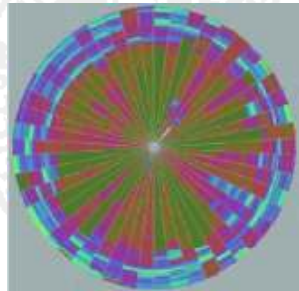
- Esempio di tecnica pixel-oriented dipendente dal risultato di diverse interrogazioni sulla base di dati



Visual Data Mining ed Esplorazione di Grandi Quantità di Dati 39/37

Tecniche pixel-oriented

- Esempio di tecnica pixel-oriented che usa la tecnica dei circle segments



Visual Data Mining ed Esplorazione di Grandi Quantità di Dati 39/37

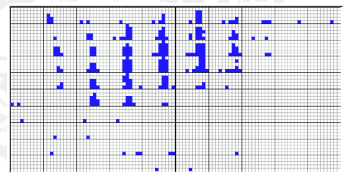
Tecniche gerarchiche

- I dati sono visualizzati secondo partizioni gerarchiche nello spazio
- Dimensional stacking
- Mondri all'interno di mondi
- Treemap
- Cone Trees
- InfoCube

Visual Data Mining ed Esplorazione di Grandi Quantità di Dati 39/37

Tecniche gerarchiche – dimensional stacking

- Lo spazio n-dimensionale è partizionato in uno spazio bidimensionale
- I sottospazi sono 'impilati' uno nell'altro
- Gli attributi sono partizionati in classi
- Gli attributi più importanti sono usati nelle classi più esterne
- Sono indicati soprattutto per dati numerosi che hanno una bassa cardinalità:



used by permission of M. Ward, Worcester Polytechnic Institute

Visual Data Mining ed Esplorazione di Grandi Quantità di Dati 39/37

