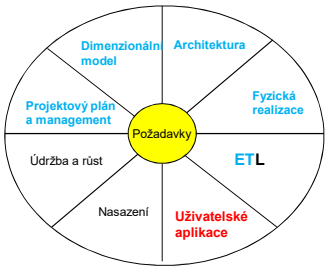
Databázové systémy a metody zpracování dat

Vizualizace velkých dat – BIG DATA

zdroj: http://www.consultadd.com/services-2/big-data-training

11.přednáška

Architektura



Úvod

- Velké objemy dat -fenomén informačního přetížení
- Trpí efektivita zpracování
- Standardní analytické nástroje selhávají
- Potřeba inteligentnějších a efektivnějších nástrojů a metod podporujících analytický proces

Analýza velkých dat

- Data arrives as sequence of items at high speed, forever.
- Can't store them all.
- Can't go back; or too slow
- Data Stream Axioms
  - 1 Only one; t-th item available at time t only
  - 2 Small processing time per item
  - 3 Small memory, certainly sublinear in stream length; sketches or summaries
  - 4 Able to provide answers at any time
    - (Ricard Gavalda, Sep 2, 2015 Summer School on Data Sciences for Big Data Porto)
- Platí i pro analytické vizualizace?

Motivace

- Samotné uložení dat dosud nebyl problém
- Data ukládána bez pročištění
  - Poškozená, nepřesná, chybějící
  - Kvalita zdroje dat
- Možnost, jak data sbírat a ukládat, roste rychleji než schopnost je analyzovat
- To může vést ke špatné interpretaci dat:
  - Špatně zpracovaná, nevhodně prezentovaná, irelevantní

Popis

- Analytik stále řídí celý proces
- Vizualizační problémy, jejichž řešení nezahrnuje metody automatické analýzy dat, nespadají do oblasti VA (Visual Analytics)
- Iterativní proces
  - Získání dat, předzpracování dat, reprezentace informací, interakce, vyvozování

Popis

- Automatická analýza dat
  - KDD, statistické metody
- Schopnosti analytika
- Vytvoření užitečné vizualizace není triviální
  - Spousta způsobů jak data prezentovat
  - Výběr správných metod

Motivace

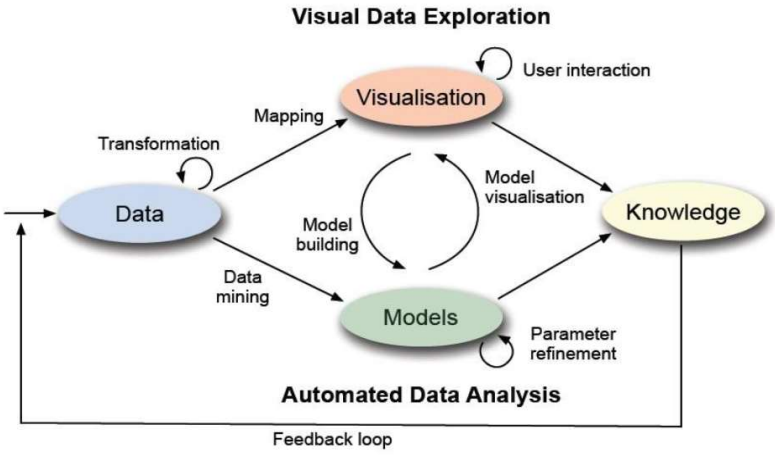
- Zbytečné plýtvání zdrojů
- V mnoha oblastech jsou správné informace získané ve správnou dobu rozhodující
- Výběr vhodných metod
  - Ať už analytických nebo jiných
  - Spolehlivé a přínosné informace
- Změnit nevýhodu velkého množství dat ve výhodu
- Visual Analytics

Popis

- Definice není jednoduchá
  - –Multidisciplinární
  - –Vizualizace, lidský faktor, analýza dat
- Definice: "The science of analytical reasoning facilitated by interactive visual interfaces„

  (P. C. Wong and J. Thomas. Visual analytics, 2004)
- Zprůhlednění celého analytického procesu
- Semi-automatický proces
  - –Lidský faktor a strojové zpracování

Proces



zdroj: http://www.vismaster.eu/book/

Popis

- První důležitý krok je předzpracování dat
  - Transformace dat do vhodného formátu
  - Pročištění dat, normalizace
- Volba mezi vizualizační nebo automatickou metodou analýzy
- Střídání vizualizačních a automatických metod
- Postupné zlepšování výsledků na základě verifikace předchozích mezivýsledků

Proces

- Postupné vylepšování modelu umožňuje dříve odhalit problémy
  - Chyby v předzpracování
  - Chyby ve zdrojových datech
  - Nevhodný postup analýzy
- Kvalitnější a důvěryhodnější výsledky

Proces

- Znalosti mohou být získány:
  - Vizualizací
  - Analytickými metodami
- Poznatky získané při vizualizaci jsou užitečné při dalším směrování analýzy
- Jak prezentovat zkoumaná data
  - „Overview first, zoom and filter, details on demand“
  - Visual Information-Seeking Mantra, Schneiderman

Proces

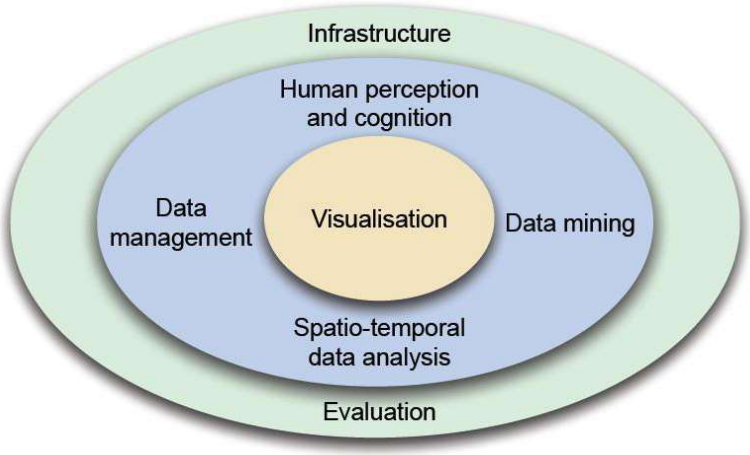
- Tento přístup však není vhodný v kontextu VA
  - V masivních objemech dat je obtížné vytvořit přehled
  - Mohli bychom přijít o důležité informace
- Rozšíření: „Analyse first, show the important, zoom/filter and analyse further, details on demand“
- Nelze jen shromáždit data a zobrazit je
- Větší důraz na analýzu s ohledem na požadovaný cíl

Proces - Automatické metody

- DM metody
- Výstupem je model
- Možnost interakce s daty
- Přehlednější úprava parametrů metod
- Výběr jiných metod
- Vizualizace modelu umožní jednodušší vyhodnocení výsledků

zdroj: http://www.vismaster.eu/book/

Základní součásti



Základní součásti

- Integruje několik vědních disciplín
- Vizualizace je základním stavebním kamenem celého systému
- Slouží k zobrazení
  - Dat
  - Výsledků analýz
- Zpřehlednění procesů v ostatních oblastech

Analýza dat

- Dva hlavní přístupy k analýze dat:
  - Konfirmační analýza
  - Explorační analýza
- Konfirmační analýza
  - Jako vstup máme hypotézu o datech, kterou ověřujeme
- Explorační analýza
  - Není přímo daná hypotéza, ale hledáme potenciálně užitečné informace a vztahy v datech
  - Důležité jsou interaktivita a vizualizace

Analýza dat

- Metody pro vizualizaci abstraktních dat
  - Business data, sociální sítě
- Velké objemy vícedimenzionálních dat
- Různé datové typy
  - Numerická, textová data, grafika, zvuk, video
- Data nelze snadno mapovat do 2D/3D
- Standardní vizualizační techniky nejsou efektivní

Správa dat

- Efektivní a kvalitní správa dat
  - Dobře navržená databáze
- Poskytuje data k analýze
- Efektivní reprezentace různých druhů dat
- Integrace heterogenních dat
- Čištění dat
  - Chybějící data, nepřesná data
- Nové zdroje dat
  - Streamovaná data, senzorové sítě

Efektivní propojení

- Efektivní propojení všech procesů, funkcí a služeb
- Rozdílné technologie využívané v jednotlivých oblastech
- Interaktivita klade vysoké požadavky na kvalitu infrastruktury
- Většina VA systémů je vyvíjena na míru
- Často využívají in-memory databáze místo klasických DBMS

Evaluace

- Vyvíjí se velké množství nových technik a metod
- Je potřeba vyhodnotit efektivitu, přínos a kvalitu
- Dobré vyhodnocení může odhalit potenciální problémy
- Výzkum a vývoj je díky velkému množství specifických oblastí roztříštěn

–Komplikuje použití jednotných evaluačních metod

Analýza časových dat

Hodnoty se mění v čase

- Hledání vzorů, trendů a korelací v čase
- Časová data sebou nesou specifické obtíže
  - Trend vývoje v určitý den nebo za celý rok
  - Data jsou často nekompletní, interpolovaná a naměřená v různých časech

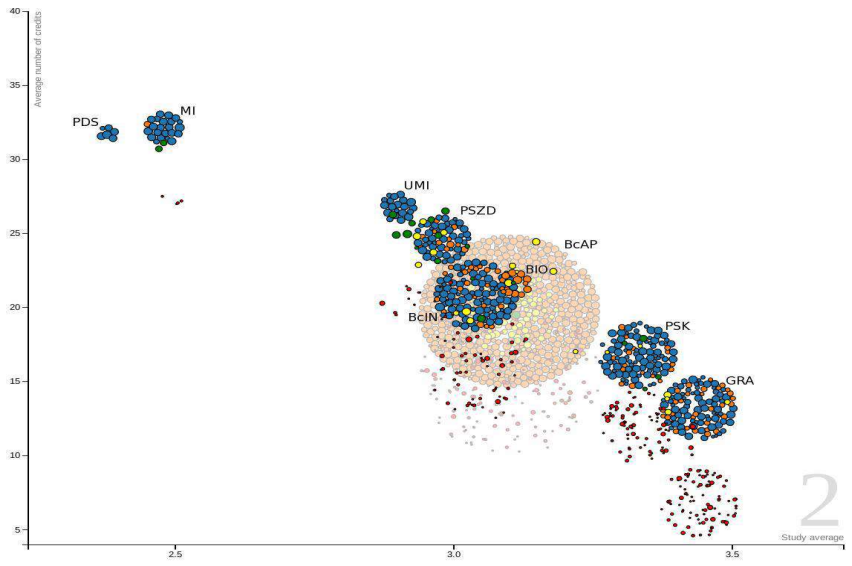
VA (Visual Analytics) nástroj

- Využívá několik metod založených na animovaných grafech
  - Motion Charts
- Původně navržený pro analýzu dat z informačního systému univerzity
- Hledání a ověřování hypotéz o datech
- Je rozšířen základní koncept MC metod
  - Zobrazení více dimenzí
  - Zobrazení více bodů
  - Zobrazení více animací

MC (Motion Charts)

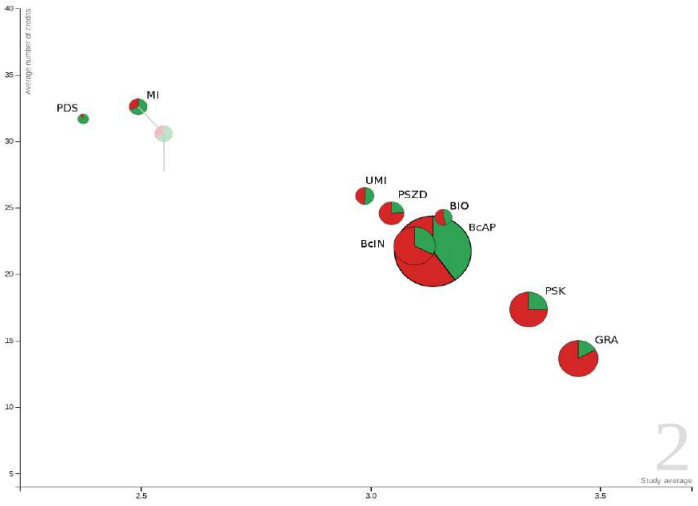
- Animované grafy:
  - –Zobrazují několik ukazatelů (dimenzí) v čase
- Jednodušší identifikace vzorů a trendů v datech
- Mapování dimenzí
  - –Důležitá část analýzy
  - –Neexistuje optimální metoda
  - –Ovlivněno charakteristikou dat a zkoumanou hypotézou
- Primárně navržené pro prezentaci dat

Skupiny studentů dle oboru

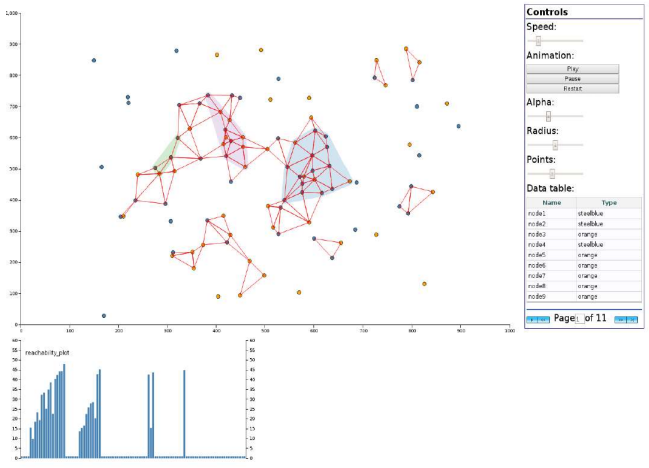


- Studenti zapsaní do bakalářského studia
  - roky 2008 až 2014
- Velké entity
  - Reprezentují konkrétní obor studia
  - Velikost odpovídá počtu studentů
- Malé entity
  - Reprezentují konkrétní studenty
  - Velikost odpovídá počtu získaných kreditů
  - Barva odpovídá stavu studia

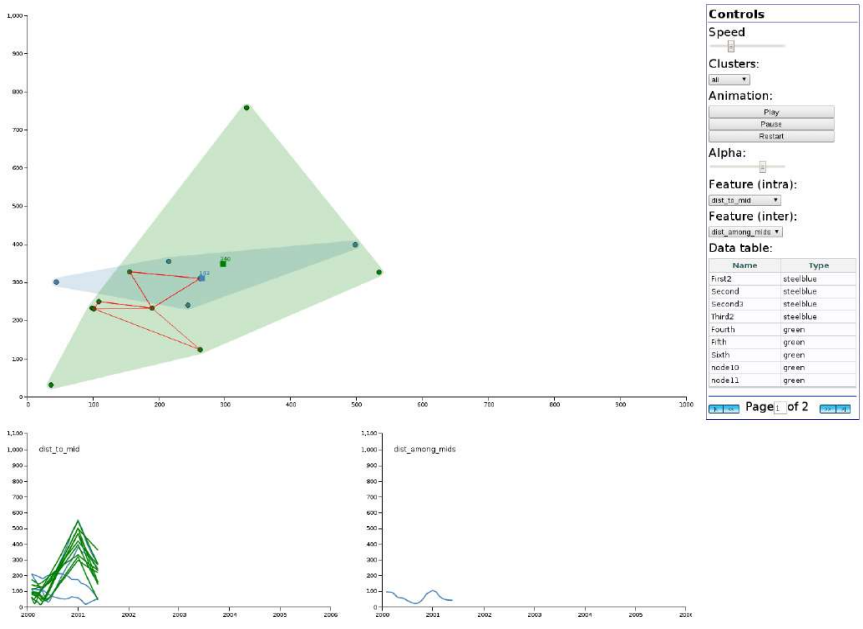
- animace vyjadřují
  - Průběh studia
  - Přerušení studia
  - Změna oboru studia
  - Změna programu studia
- Číslo semestru představuje časovou složku
- Vážený průměr je mapován na X osu
- Průměrný počet kreditů je mapován na Y osu
- Velikost vyjadřuje počet získaných kreditů



Přidání dimenze pomocí pie chart



- Hledají se shluky entit
- Ordering points to identify the clustering structure (OPTICS)
  - Density-based clustering
- Nepotřebuje počet shluků
- Vstupní parametry
  - Prohledávaná vzdálenost
  - Počet entit nutných k vytvoření shluku
- Graf dosažitelnosti
- Minimální kostra grafu



- Zkoumají se charakteristiky
  - Entit v rámci shluků
  - Shluků vzájemně
- Spočítané charakteristiky se zobrazují v grafech pod hlavním oknem
- Minimální kostra grafu pro každý shluk

- Mohou nástroje vizuální analýzy pomoci při zpracování velkých dat?
- Příklady použití:
  - Vizualizace četností
  - Vizualizace agregací
  - Vizualizace průměru
  - Existující algoritmy lineární složitosti – téměř nepoužitelné - příliš pomalé…

Příklad

Příklad: přibližný výpočet průměru

epsilon-delta Aproximace E(X)

|Aproximace E(X) – E(X)| < epsilon s pravděpodobností delta

Je možno ukázat (viz např. Gavalda, Summer school Porto), že stačí vzít vzorek velikosti

1/(2*epsilon^2)*ln(2/delta)

VIZUALIZACE MEDICÍNSKÝCH DAT

Nástroje pro vizualizaci dat

- Cíl – snížit informační zatížení
- Inteligentní abstrakce
- Vizualizace zajímavých příznaků
- Zobrazení složitých vztahů mezi daty

Vizualizace lékařských dat

- Využití obrazové informace – MRI, PET, CT
- Grafické zobrazení dat a informací jiného než obrazového charakteru (jednorozměrné signály, numerická data, apod.)

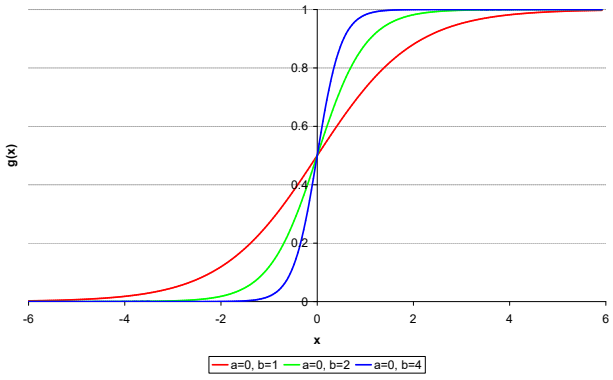
Analýza a úprava jednotlivých atributů

- Zpráva o stavu proměnných
  - typ (spojitá X diskrétní)
  - rozsah definičního oboru (počet použitých hodnot)
  - rozsah a frekvence výskytů (histogram)
  - typ rozdělení a jeho statistické charakteristiky
- Upozornit na
  - osamělé mimořádné hodnoty (outliers)
  - téměř konstantní atributy (možné vynechat)
  - nevyplněná datová pole
  - znečištění dat
    - data neodpovídají deklarovanému formátu
    - hodnoty neodpovídají deklarované množině

Analýza a úprava jednotlivých atributů

Příklady úprav

- Náhrada chybějících údajů - provádí se tak, aby zůstala zachována hodnota směrodatné odchylky uvažovaného atributu
- Úprava rozsahu hodnot atributů pomocí logistické transformace (velmi důležité v každé metodě, která počítá se vzdáleností objektů - např. CBR, nejbližší sousedi, shlukování)



$$g(x) = (1 + e^{-a - bx})^{-1}$$

Analýza a úprava jednotlivých atributů

- Monotónní atributy – představují obvykle jednoznačnou identifikaci pro uvažované objekty, např. pořadové číslo měření, číslo bankovního účtu. Rostou bez omezení a při tom jejich přímá hodnota jako taková nemá pro vytvoření modelu význam.
- Řady – tvořené hodnotami veličin, které jsou pravidelně měřeny a zaznamenávány (např. EKG, burzovní koeficienty). Vždy jsou vztaženy k jediné monotónní veličině, která slouží jako index.
  - často jako index slouží čas -> časová řada
  - Prostředky k analýze:
    - Fourierova analýza
    - Vlnková (wavelet) transformace umožňuje získání časově-frekvenčního popisu signálu

Úpravy a analýza dat ve stavovém prostoru

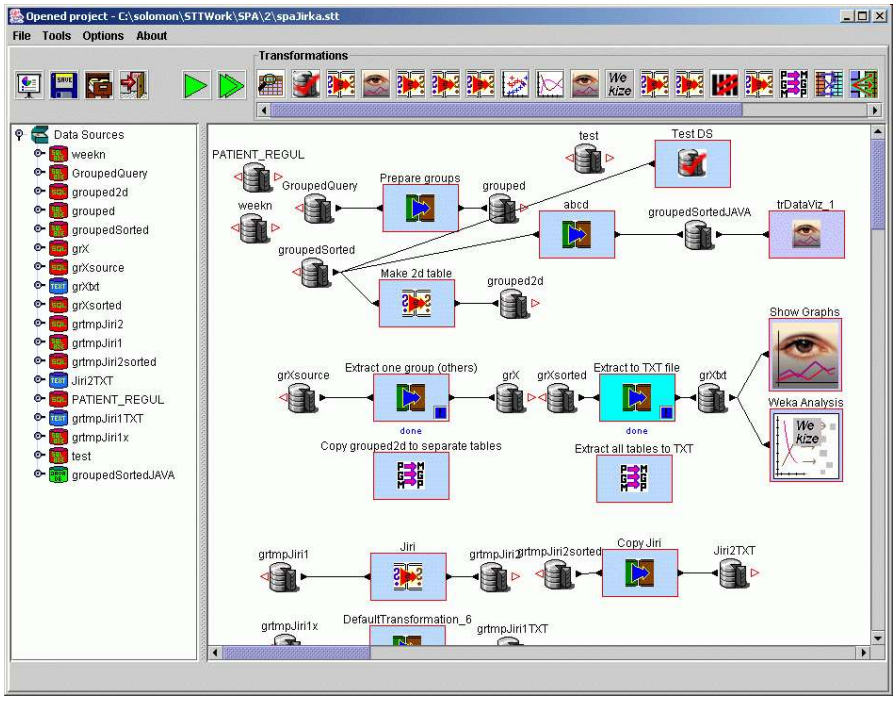
Příklady úprav

- Snížení dimenze
  - vynecháním
    - konstantních atributů
    - atributů řídce obsazených
    - atributů s duplicitní informací (rok narození X věk, apod.)
  - sloučením
    - atributů řídce obsazených – z několika řídce obsazených atributů je možné zřetězením vytvořit jeden nový (PVP - present value pattern)
- Zvýšení dimenze
  - obohacení doplněním údajů z jiných zdrojů (např. meteorologická měření, demografické údaje, apod.)
  - rozšíření
    - přidání odvozených atributů (např. pohlaví z rodného čísla, apod.)
    - „otočení“ dat (reverse pivoting) - nový atribut a_{n+1} přebírá údaj z objektu následujícího. Pro každý objekt *i* platí $a_{n+1}(i) = a_n(i+1)$.

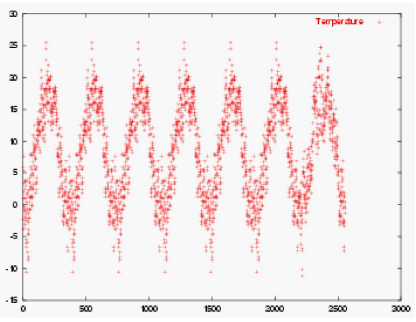
Úpravy a analýza dat ve stavovém prostoru

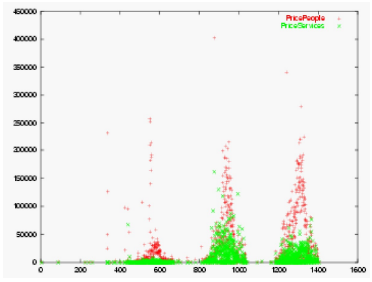
Příklady úprav

- Agregace dat - použití metod datových skladů. údaje o více objektech obsažené na několika řádcích jsou vztaženy k jedinému obecnějšímu objektu (tvoří tedy v novém souboru jedinou řádku).
- Vizualizace – např. umístění datového souboru ve stavovém prostoru úlohy, přirozené shluky, nepravidelné deformace,...
- Statistické přístupy snižování dimenze
  - podmíněná entropie
  - CHAID (Chi-square Automatic Interaction Detector)
  - hledání hlavních komponent (návrh vhodné lin. kombinace)
- Využití neuronových sítí - Řídce propojená autoasociativní neuronová síť (Sparcely Connected Autoasociative Neural Net: SCANN)

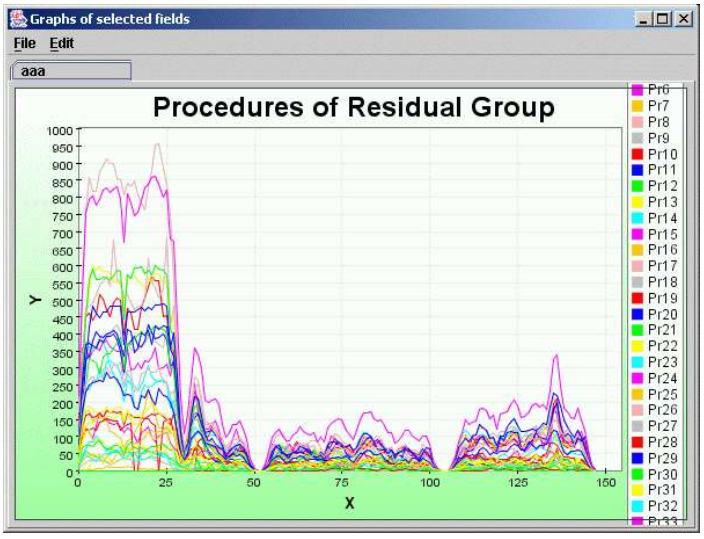


Sumatra TT – rozptylový diagram

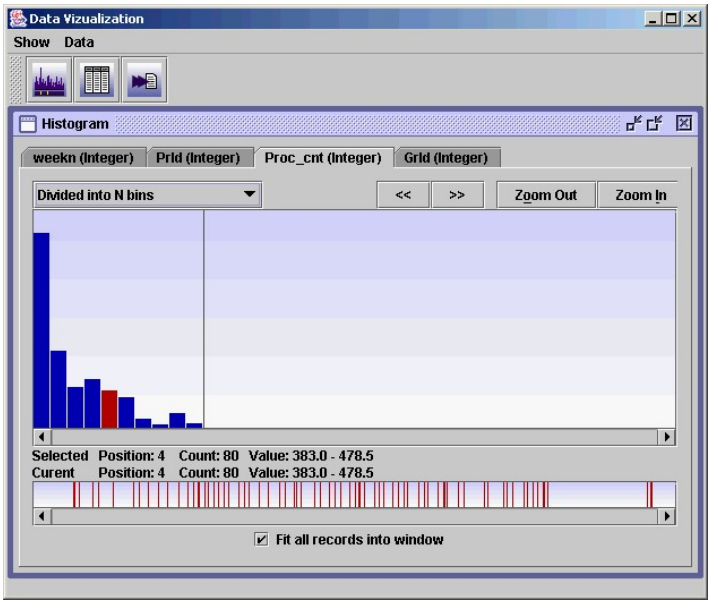




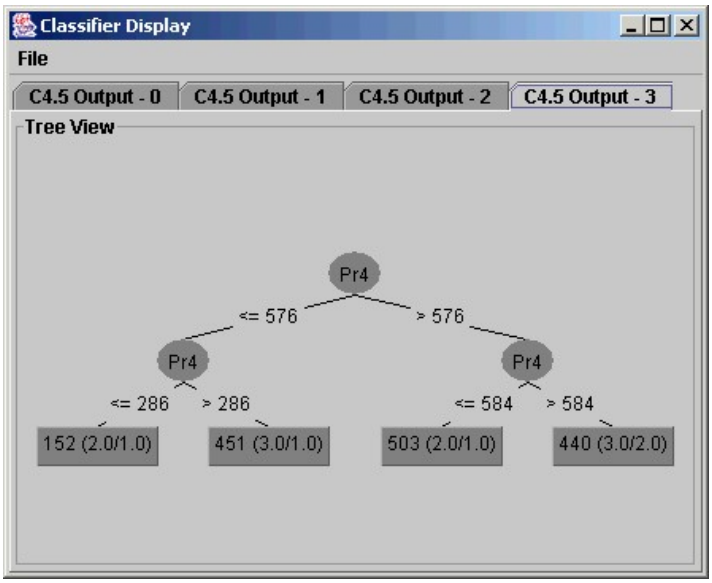
Sumatra TT – grafy vybraných proměnných



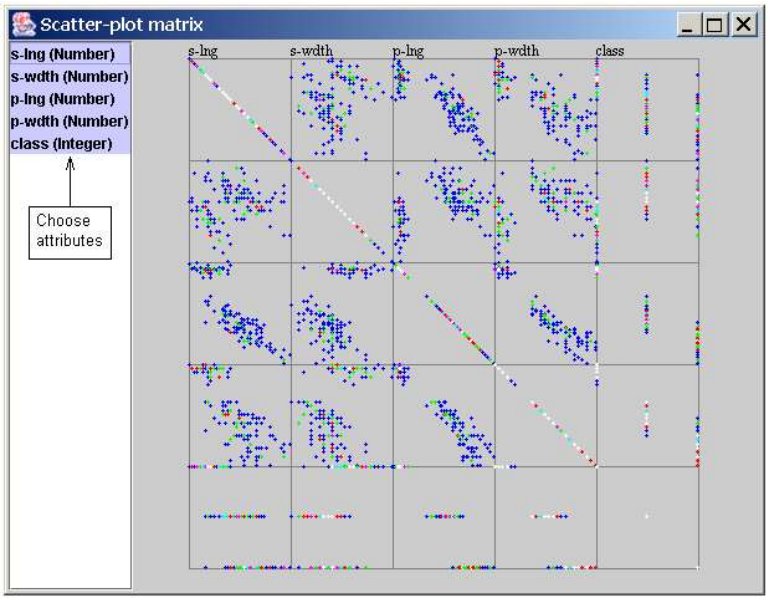
Sumatra TT – histogram



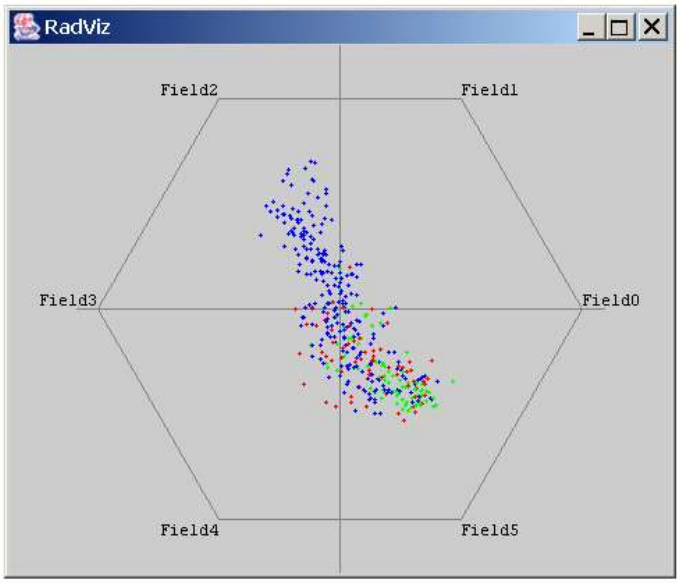
Sumatra TT – rozhodovací strom



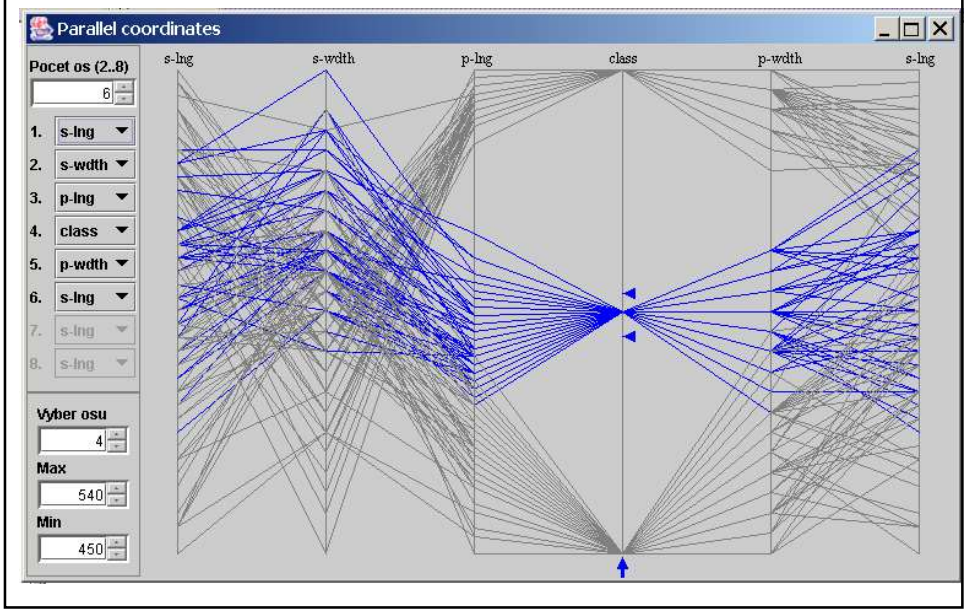
Sumatra TT – rozptylová matice



Sumatra TT



Sumatra TT – vztahy mezi atributy



Další možnosti zobrazení

- Frekvenční spektrum
- Výkonové spektrum
- Spektrální kulisy
- Mapy

- Interaktivní zobrazení – mapa + průběh signálu

- Tyto možnosti budou prezentovány v konkrétních aplikacích.

Závěr

- medicína – velké objemy dat
- větší počet přístrojů přímo propojených s počítači - více vstupních dat pro vyhodnocování
- efektivní vyhodnocování velkého objemu dat
- často neznámé explicitní relace mezi daty - obtížná interpretace - nástroje dobývání znalostí
- integrace s vizualizačními nástroji – podpora rychlejšího porozumění složitým, velkým a dynamicky rostoucím souborům dat