Visual Analytics

Jan Géryk

PV056 Strojové učení a dobývání znalostí, 13. 5. 2014

Témata přednášky

1. Úvod
2. Motivace
3. Historie
4. VA proces
5. Součásti VA
6. Shrnutí a výzvy

Úvod

- Velké objemy dat
    - fenomén informačního přetížení
- Nedostatečná efektivita zpracování
- Standardní analytické nástroje selhávají
- Efektivnější způsoby analýzy
- Samotné uložení dat není problém
- Data ukládána bez pročištění
    - poškozená, nepřesná, chybějící
    - kvalita zdroje dat

Motivace

- Možnosti jak data sbírat a ukládat roste rychleji, než schopnost je analyzovat
- To může vést ke “ztracení“ v datech:
    - irelevantní, špatně zpracovaná, nevhodně prezentovaná data
- Zbytečné plýtvání zdrojů
- Správné informace ve správnou dobu
- Výběr vhodných metod
    - spolehlivé a přínosné informace

Motivace

- Změnit nevýhodu velkého množství dat ve výhodu
- Zprůhlednění celého analytického procesu
- Visual Analytics
    - vizualizace informací a interakce s daty
    - interaktivní rozhraní
    - lidský faktor a strojové zpracování
    - analytik stále řídí celý proces
    - multidisciplinární

Historie

- Automatizované techniky analýzy vznikly nezávisle na vizualizačních a interakčních
- Změna několika klíčových myšlenek vedla ke vzniku Visual Analytics
- Zejména posun od “confirmatory data analysis“ k “exploratory data analysis“
- Definoval John W. Tukey v knize Exploratory data analysis (1977)

Historie

- První zmínka o Visual Analytics v roce 2004
- Termín použit v širším kontextu
- Kombinuje poznatky z několika výzkumných oblastí
- Charakteristiky VA aplikací se objevili už v devadesátých letech (systém CoCo)
    - vylepšení návrhu čipů
    - návrhář monitoruje a řídí průběh
    - rozhraní zobrazuje indikátory výkonnosti čipu a citlivosti

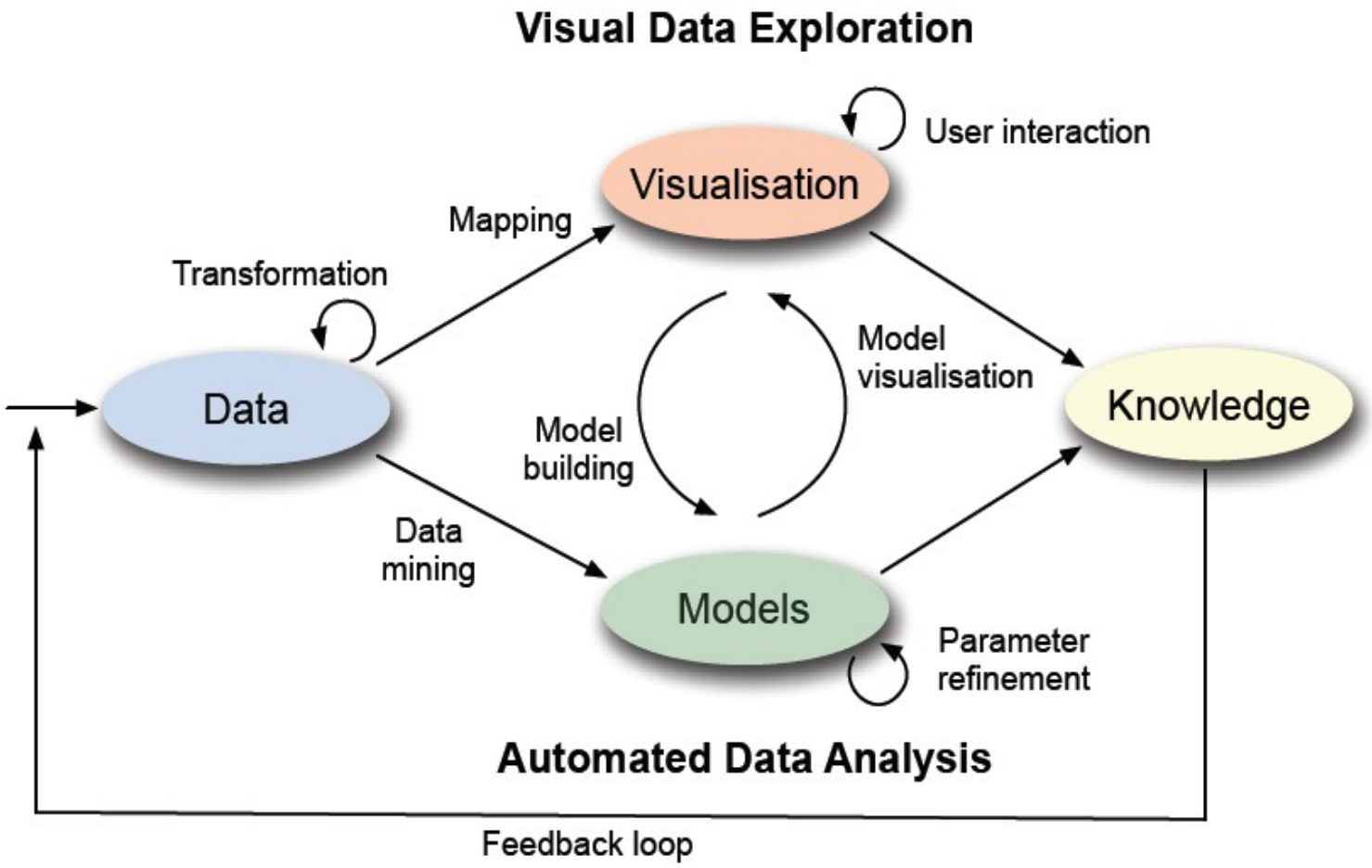
Visual Analytics

- Definice není jednoduchá
    - Multidisciplinární
    - Vizualizace, lidský faktor, analýza dat
- První: "The science of analytical reasoning facilitated by interactive visual interfaces"
    - P. C. Wong and J. Thomas. Visual analytics, 2004
- Lepší: "Visual analytics combines automated analysis techniques with interactive visualisations for an effective understanding, reasoning and decision making on the basis of very large and complex datasets"

Visual Analytics

- Iterativní proces
  - získání dat a předzpracování dat
  - reprezentace znalostí a interakce
  - vyvozování a rozhodování
- Kombinace automatizované analýzy
  - data mining, statistika
- A schopností analytika
  - chápání, vyvozování

Visual Analytics – proces



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

Visual Analytics – proces

- Velmi důležitý krok je předzpracování dat
  - Transformace dat do vhodného formátu
  - Pročištění dat
  - Normalizace dat
- Volba mezi vizualizační a automatizovanou metodou analýzy
- Střídání vizualizačních a analytických metod
- Neustálé zlepšování na základě verifikace předchozích (mezi)výsledků

Visual Analytics - automatizované

- Data mining metody
- Výstupem je model
- Přehlednější úprava parametrů metod
- Výběr jiných analytických metod
- Vizualizace modelu umožní jednodušší vyhodnocení výsledku

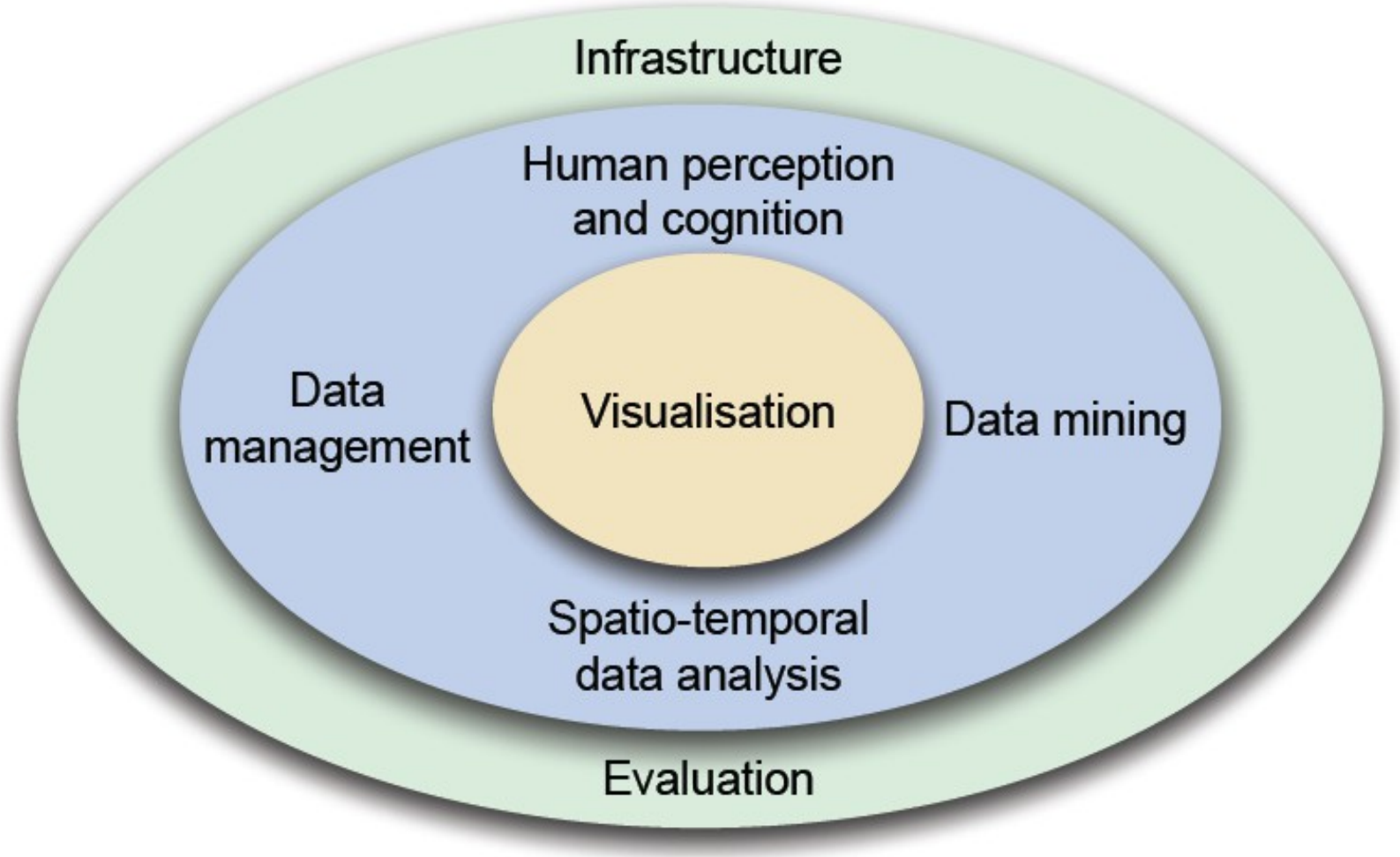
Visual Analytics – proces

- Zpětná vazba nám umožňuje vylepšovat proces analýzy
- Postupné vylepšování modelu umožňuje dříve odhalit problémy
    - Chyby ve zdrojových datech, předzpracování nebo analýze
- Kvalitnější a důvěryhodnější výsledky
- Využití vizualizace
    - Zaměření na různé části nebo typy dat
    - Vygenerované hypotézy je potřeba ověřit analytickými metodami

Visual Analytics – proces

- Poznatky získané při vizualizaci jsou užitečné při dalším směrování analýzy
- Jak vhodně prezentovat zkoumaná data
    - Shneiderman, 1966, “Overview first, zoom/filter, details on demand“
- Nevhodné v kontextu VA
    - Můžeme přijít o důležité informace
- Rozšíření: “Analyse first, show the important, zoom/filter and analyse further, details on demand“
    - analýza s ohledem na požadovaný cíl

Visual Analytics – stavební kameny



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

Visual Analytics – stavební kameny

- Integruje několik vědních disciplín
- Vizualizace je základním stavebním kamenem celého systému
- Slouží k zobrazení dat a výsledků analýz
- Zpřehlednění procesů v ostatních oblastech
    - Data management, data mining

Visual Analytics – Vizualizace

- Stále poměrně nová vědní disciplína
  - Rozvoj v posledních 20 letech
- Scientific visualization
  - Data lze mapovat do 2D/3D
  - Objemy, povrchy, toky, ...
  - Biologie, meterologie, ...
- Information visualization
  - Vizualizace abstraktních dat
  - Stovky dimenzí
  - Business data, sociální sítě, ...

Visual Analytics – Vizualizace

- Prezentace
  - Výběr vhodné techniky závisí na typu uživatele a aplikace
- Confirmatory analysis
  - Vstupem je hypotéza o datech, kterou ověřujeme (potvrdíme/vyvrátíme)
- Exploratory analysis
  - Nemáme danou hypotézu
  - Hledáme potenciálně užitečné informace
  - Důležitá je možnost interakce

Visual Analytics – Data management

- Efektivní a kvalitní správa dat
    - Dobře navržená databáze
- Typicky poskytuje data k analýze
- Integrace heterogenních dat
- Hledání efektivních reprezentací různých druhů dat
- Čištění dat
    - Chybějící data, nepřesná data
- Nové zdroje dat
    - Streamovaná data, sensorové sítě

Visual Analytics – Data mining

- Automatizované metody pro extrakci informací
- Učení s učitelem
  - Algoritmy se aplikují na množiny trénovacích dat
  - Výsledkem jsou modely
  - Klasifikace předtím neviděných dat
  - Rozhodovací stromy, support vector machine, neuronové sítě, ...

Visual Analytics – Data mining

- Učení bez učitele
  - Odhalení struktury dat bez jakékoliv předchozí znalosti
  - Klastrování, asociační pravidla
  - Nastavení parametrů metod analytikem

Visual Analytics – Visual Data mining

- Interaktivní vizualizace
    - Přehlednější nastavení parametrů
- Rozhraní umožňující vizuální prezentaci zkoumaných dat
- Prezentace dat způsobem, který umožní analytikovi lépe pochopit data
- Prezentace výsledků analýz

Visual Analytics – Prostorová a časová analýza

- Prostorová data
  - Dají se vynést do grafu nebo zobrazit na mapě
    - Geografická měření, GPS data
  - Hledání vztahů a zajímavých vzorů
    - Vhodné nástroje, např. efektivní datové struktury
- Časová data
  - Hodnoty se mění v čase
  - Hledání vzorů, trendů a korelací

Visual Analytics – Prostorová a časová analýza

- Prostorová a časová data sebou nesou jisté obtíže
  - umožnit změnu měřítka mapy
  - trend vývoje v určitý den nebo za celý rok
  - často nekompletní a interpolovaná data naměřená v různých časech
  - složité topologické vztahy mezi objekty

Visual Analytics – Vnímání a poznávání

- Reprezentuje lidskou stránku
- Vizuální vnímání je prostředek, kterým člověk interpretuje své okolí
- Poznávání je schopnost tyto informace pochopit a vyvodit závěry
- Poznatky z těchto oblastí jsou důležité hlavně při návrhu uživatelských rozhraní
- Také při návrhu multimodálních interakčních technik
    - Interakce člověka s počítačem užitím více vstupních a výstupních zařízení

Visual Analytics – Infrastruktura

- Efektivní propojení všech procesů, funkcí a služeb VA aplikace
- Rozdílné technologie využíváné v jednotlivých oblastech
- Velká interaktivita klade další požadavky na kvalitu infrastruktury
- Většina takovýchto aplikací je vyvíjena na míru
- Často využívají in-memory databázy místo klasického DBMS

Visual Analytics – Aplikace

- Fyzika a astronomie
    - Vizualizace toků, dynamika tekutin, ...
- Business
    - Finanční trhy, ...
- Monitorování životního prostředí
    - Počasí, data ze satelitů, ...
- Bezpečnost
- Biologie a medicína
- ...

Visual Analytics – Vyhodnocení

- Vyvíjí se velké množství nových technik, metod, modelů a teorií
- Je potřeba vyhodnotit efektivitu, přínos a kvalitu
- Dobré vyhodnocení může odhalit potencialní problémy
- Výzkum a vývoj je díky velkému množství specifických oblastí roztříštěn
    - komplikuje použití jednotných evaluačních metod

Shrnutí - výzvy

- Visual Analytics se musí zabývat čtyřmi důležitými aspekty
- Data: velké množství různorodých typů dat různé kvality
- Uživatelé: vyhovět uživatelským požadavkům a zjednodušit a zpřehlednit analýzu
- Design: robustní a efektivní návrh systému
- Technologie: využití moderních a efektivních technologií

Shrnutí - Data

- Velké množství dat
- Ukládání, získávání a přenos
    - Distribuované databáze, cloudy
- Náročnost zpracování
- Škálovatelnost vizualizací
- In-memory úložiště
    - Lépe vyhovuje požadavkům
    - efektivnější
- Různorodá data
    - Nekvalitní a nekompletní

Shrnutí – Data

- Složitost integrace dat z více zdrojů
- Potřeba transformovat data do jednotného formátu
- Nové typy dat
    - streamovaná data
- Potřeba zpracovat data v reálném čase

Shrnutí – Uživatelé

- Uživatel by měl mít přehled o analýze
- Odkud se data berou, jaké operace byly s daty provedeny
- Chápání nedostatků v datech a výsledcích
    - Omezení chybné interpretace
- Většina DM metod je neintuitivních
    - vyžadují odbornou znalost
    - Vhodná reprezentace dat

Shrnutí – Design

- Aplikace moderních teoretických a praktických znalostí
- Hodně technologií a pro daný problém je potřeba zvolit správné techniky
    - Analytické metody, typ vizualizace
- Použití, evaluace daného řešení
- Potřeba unifikovaného modelu
    - Rychlejší a spolehlivější návrh a implementace

Shrnutí – Technologie

- Analytická část bývá výrazně delší než klasické transakce
    - Je potřeba ukládat mezivýsledky
- Analytik by měl mít neustále přehled o průběhu analýzy a možnost řídit její běh
- Analytik může požadovat data za jeden den nebo za celý rok
- Vývoj webových vizualizačních nástrojů
    - dostupnost

Literatura

- Daniel A. Keim and Florian Mansmann and Jörn Schneidewind and Hartmut Ziegler and Jim Thomas, Visual Analytics: Scope and Challenges, 2008
- David J. Kasik and David Ebert and Guy Lebanon and Haesun Park and William M. Pottenger, Data transformations and representations for computation and visualization, 2009
- VisMaster: http://www.vismaster.eu/book/