Visual Analytics (VA)

Jan Géryk

PV056 Strojové učení a dobývání znalostí, 15. 5. 2012

Osnova přednášky

1. Úvod
2. Motivace
3. Historie
4. Popis
5. Proces
6. Základní součásti
7. Shrnutí

Úvod

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

Motivace

- Samotné uložení dat není problém
- Data ukládána bez pročištění (surová)
  - Poškozená, nepřesná, chybějící
  - Kvalita zdroje dat
- Možnosti jak data sbírat a ukládat roste rychleji, než schopnost je analyzovat
- To může vést ke “ztracení“ v datech:
  - Špatně zpracovaná, nevhodně prezentovaná, irelevantní

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
- VA

Historie

- První zmínka o VA v roce 2004
- Termín použit v širším kontextu
- Oblast kombinující několik dalších oblastí
- Charakteristiky VA aplikací se už objevily např. v systému CoCo (devadesátá léta)

Historie - CoCo

- CoCo: vylepšení návrhu čipů
- Použity numerické optimalizační algoritmy
    - Spousta nevýhod
- Velké zefektivnění při spolupráci se zkušeným návrhářem čipů
    - Návrhář monitoruje a řídí průběh
    - Rozhraní (Cockpit) zobrazující indikátory výkonnosti čipu a citlivosti; rady (AI systém)

Popis

- 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
- Přesnější: "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"

Popis

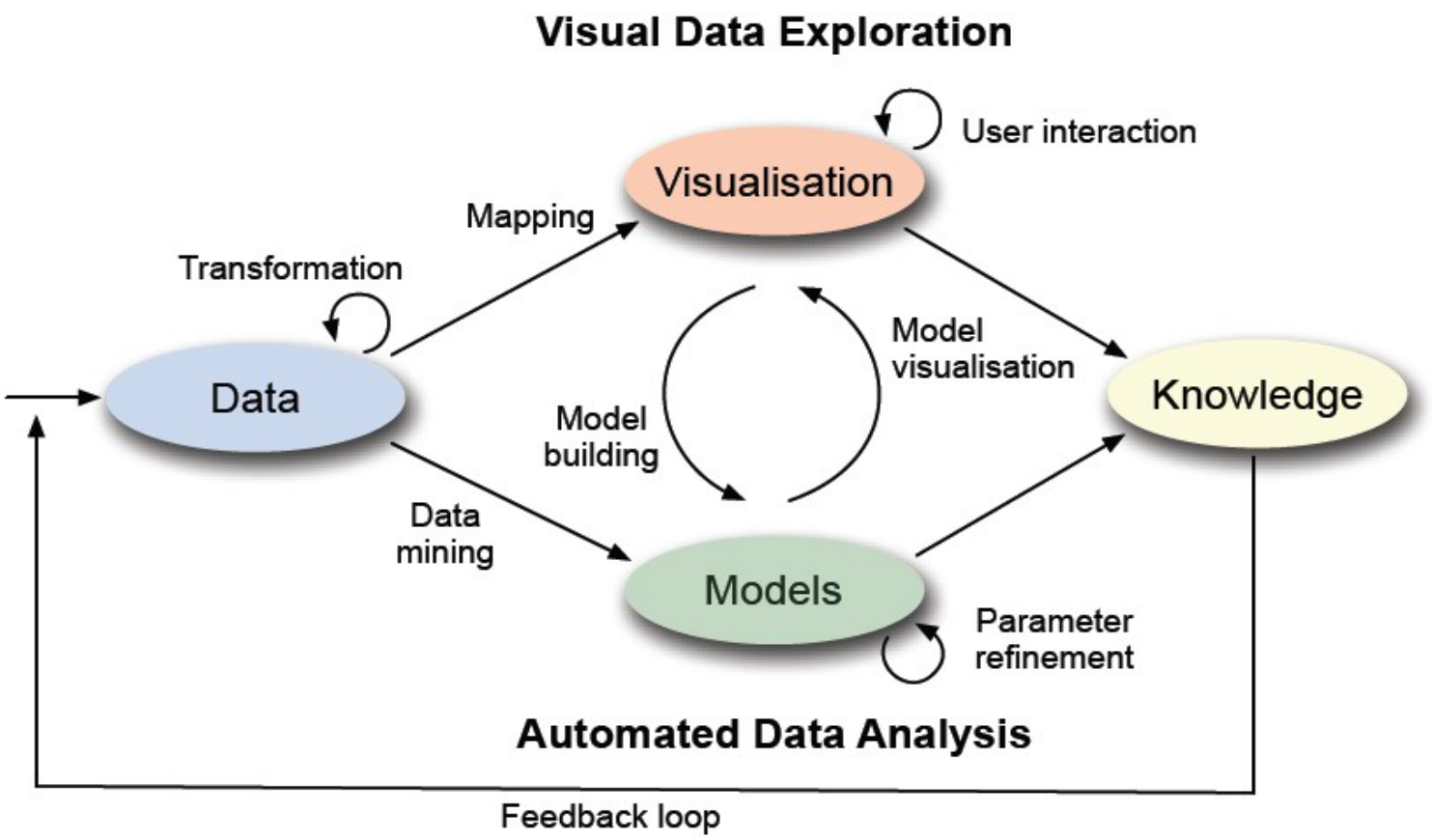
- Zprůhlednění celého analytického procesu
- Vizualizace informací a interakce s daty
  - Lepší přehled a jednodušší rozhodování
- Semi-automatický proces
  - Lidský faktor a strojové zpracování
- Analytik stále řídí celý proces
- Vizualizační problémy, jejichž řešení nezahrnuje metody automatické analýzy dat, nespadají do oblasti VA

Popis

- Iterativní proces
  - Získání dat, předzpracování dat, reprezentace informací, interakce, vyvozování
- Automatická analýza dat
  - KDD, statistika, matematika
- Schopnosti analytika
  - Chápání, vyvozování
- Vytvoření užitečné vizualizace není triviální
  - Několik způsobů jak data reprezentovat
  - Výběr správných metod

Proces

- Přehled fází (ovály) a přechodů (šipky)



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

Proces

- První důležitý krok je předzpracování dat
  - Transformace dat do vhodného formátu
  - Pročištění dat, normalizace
- Volba mezi vizuální a automatickou metodou analýzy
- Střídání vizualizačních a analytických metod
- Neustálé zlepšování na základě verifikace předchozích (mezi)výsledků

Proces

- Zpětná vazba představuje fakt, že se jedná o iterativní 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
  - Interakcí analytika s vizualizacemi a modely
- Poznatky získané při vizualizaci jsou užitečné při dalším směrování analýzy
- Jak vhodně prezentovat zkoumaná data
  - Shneiderman, 1966
  - „Overview first, zoom/filter, details on demand“

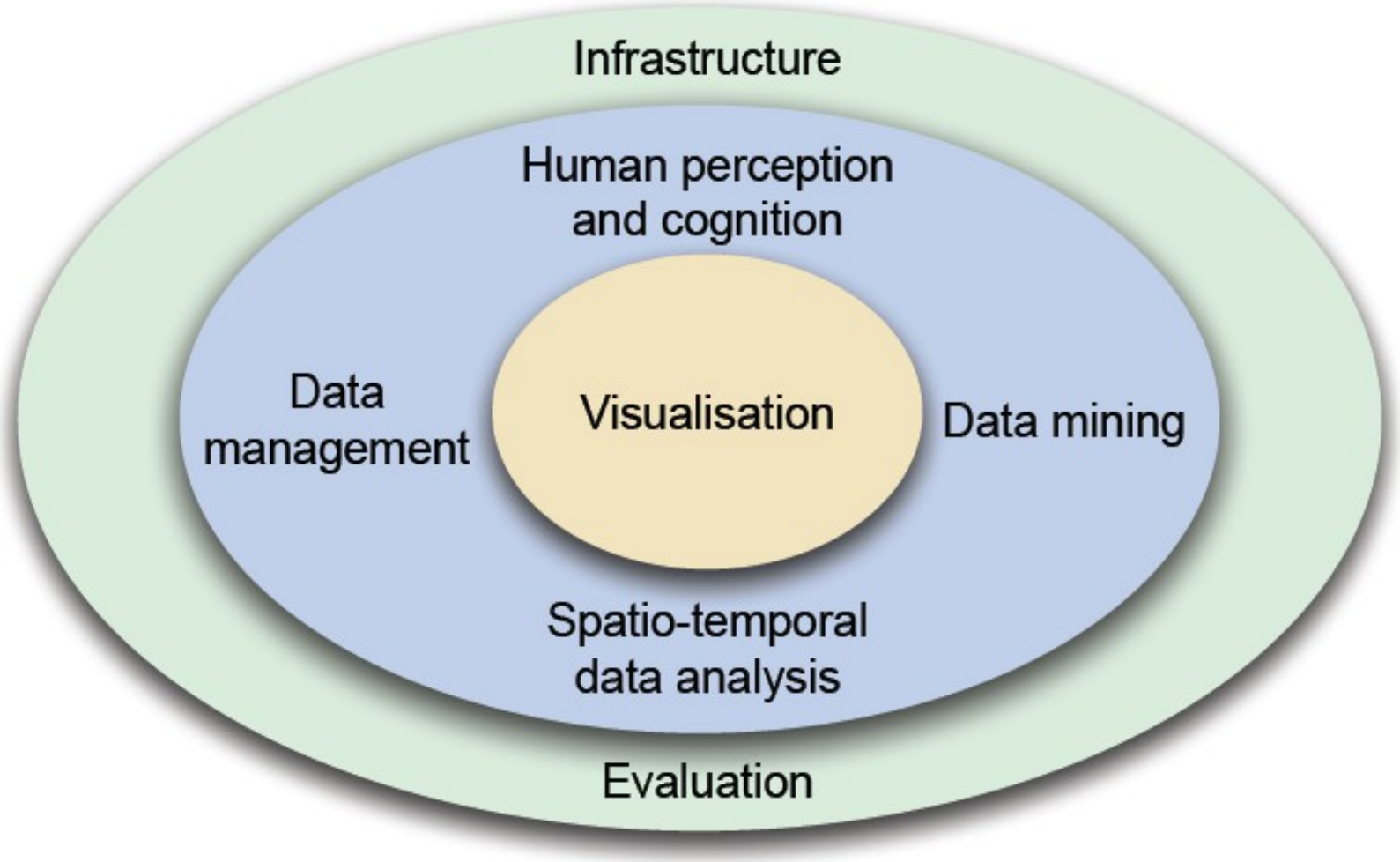
Proces – Vizualizace

- 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 vyplivnout je na obrazovku
  - Je důležitá analýza s ohledem na požadovaný cíl

Proces - Automatické metody

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

Základní součásti



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

Základní součásti

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

Vizualizace

- Poměrně nová vědní disciplína
  - Rozvoj v posledních 20 letech
- Definice podle Colin Ware: grafická (visuální) reprezentace myšlenek a dat
- Tři hlavní přístupy k analýze dat:
  - Prezentace
  - Confirmatory analysis (deduktivní přístup)
  - Exploratory analysis (induktivní přístup)

Vizualizace

- Prezentace
  - Výběr vhodné techniky uživatelem
- Confirmatory analysis
  - Jako vstup máme hypotézu o datech
  - Ověřujeme hypotézu
  - Pomocí vizualizace potvrdíme/vyvrátíme
- Exploratory analysis
  - Není hypotéza
  - Hledáme potenciálně užitečné informace
  - Interaktivita a vizualizace

Vizualizace vědeckých dat

- Dva druhy vizualizací:
  - Vizualizace vědeckých dat
  - Vizualizace informací
- Senzory, simulace, laboratorní testy
- Vizualizace toků, vykreslování objemů
- Lze jednoduše mapovat do 2D/3D prostředí

Vizualizace informací

- Metody pro vizualizaci abstraktních dat
  - Business data, demografie, sociální sítě
- Velké objemy dat, stovky dimenzí
- Různé datové typy
  - Numerická, textová data, grafika, zvuk, video
- Data nelze snadno mapovat do 2D/3D
- Standardní grafové 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

Data mining

- Automatické 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

Data mining

- Učení bez učitele
- Odhalení struktury dat bez jakékoliv předchozí znalosti
- Klastrování
  - Seskupování instancí do tříd na základě společných vlastností
  - Identifikace odchylek v zašumněných datech (předzpracování)
- Asociační pravidla

Visual data mining

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

Prostorová a časová analýza

- Prostorová data
- Data, která se dají vynést do grafu nebo zobrazit na mapě
  - Geografická měření
  - GPS data
- Hledání vztahů a zajímavých vzorů
- Využití efektivních datových struktur
- Podobnostní funkce

Prostorová a časová analýza

- Časová data
- Hodnoty se mění v čase
- Hledání vzorů, trendů a korelací v čase
- Prostorová a časová data sebou nesou jisté obtíže
    - Umožnit změnu měřítka mapy
    - Trend vývoje v určitý den nebo za celý rok
    - Data jsou často nekompletní, interpolovaná a naměřená v různých časech
    - Složité topologické vztahy mezi objekty

Aspekty 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 při návrhu uživatelských rozhraní
- Také při návrhu multimodálních interakčních technik
    - Interakce člověka s počítačem užitím více vstupních a výstupních zařízení

Infrastruktura

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

Aplikace

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

Evaluace

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

Shrnutí

- VA se tedy zabývá čtyřmi oblastmi
- Data: velké množství různorodých typů dat s různou kvalitou
- Uživatelé: vyhovět uživatelským požadavkům a zjednodušit a zpřehlednit analýzu
- Design: kvalitní návrh systému
- Technologie: využití moderních a efektivních technologií

Shrnutí – Data

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

Shrnutí – Data

- Složitost integrace dat z více zdrojů
- Potřeba transformovat data do jednotného formátu
- Nové typy dat
- Nové zdroje dat: streamovaná data
    - Velké dávky nebo neustále
    - Analýza finančních toků
- Potřeba zpracovat data v reálném čase

Shrnutí – Uživatelé

- Uživatel by měl mít přehled o průběhu
- Odkud se data berou
- Jaké operace s daty byly provedeny
  - Čištění, analýza, vizualizace
- Chápání nedostatků v datech a výsledcích
  - Omezení chybné interpretace výsledků
- Většina DM metod je neintuitivních a vyžadují odbornou znalost
  - Vhodná úroveň abstrakce, reprezentace dat na obrazovce

Shrnutí – Design

- Aplikace moderních teoretických a praktických znalostí
- Hodně technologií a pro daný problém je potřeba zvolit správné techniky
  - Analytické metody, typ vizualizace
- Unifikovaný model
  - Rychlejší a spolehlivější návrh a implementace

Shrnutí – Technologie

- Je potřeba ukládat mezivýsledky
- Analytik má neustále přehled o průběhu analýzy a řídí její průběh
- Analytik může požadovat data za jeden den, stejně jako za celý rok

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/