Visual analytics

Masarykova univerzita, fakulta informatiky
Juraj Jurčo, 173001@mail.muni.cz

Motivácia

- Užívatelia, aplikácie a zariadenia zbierajú enormné množstvo dát
- Zakiaľ množstvo týchto dát rapídne rastie, možnosti spracovávať a analyzovať tieto dáta stúpajú oveľa pomalšie

Motivácia

- V roku 2002 ~ 22EB dát
- V roku 2006 ~ 161EB dát
- V roku 2009 ~ 800EB (nárast o 62% oproti roku 2008)
- Vízia do roku 2020 ~ 35ZB = $35x2^{70}$
- 70% všetkých týchto dát je produkovaných jednotlivcami
- 30% produkujú firmy
- 25% všetkých dát tvoria obrázky [9]

Čo je vizuálna analýza?

- Thomas a Cook ju v knihe *Illuminating the Path*[2] definovali ako: “vedu ktorá uľahčuje analytické rozhodovanie pomocou interaktívnych vizualizácií”
- Pomáha lepšie si predstaviť štruktúru dát
- Kombinuje výpočetnú silu počítačov a ľudské schopnosti chápania, dávania do súvislostí a vyvodzovania záverov

Čo nie je vizuálna analýza

- Veľká grafová štruktúra bez popisu
- Diagramy bez legendy
- Obrázky, ktoré nemajú žiadnu výpovednú hodnotu
- Obrázky bez významovej interpretácie

Prečo analyzovať?

- Porovnanie
- Preskúmanie vzťahov
- Predpoveď
- Testovanie hypotéz
- Vytváranie pojmov a teórií
- Skúmanie
- Kontrola
- Vysvetlenie

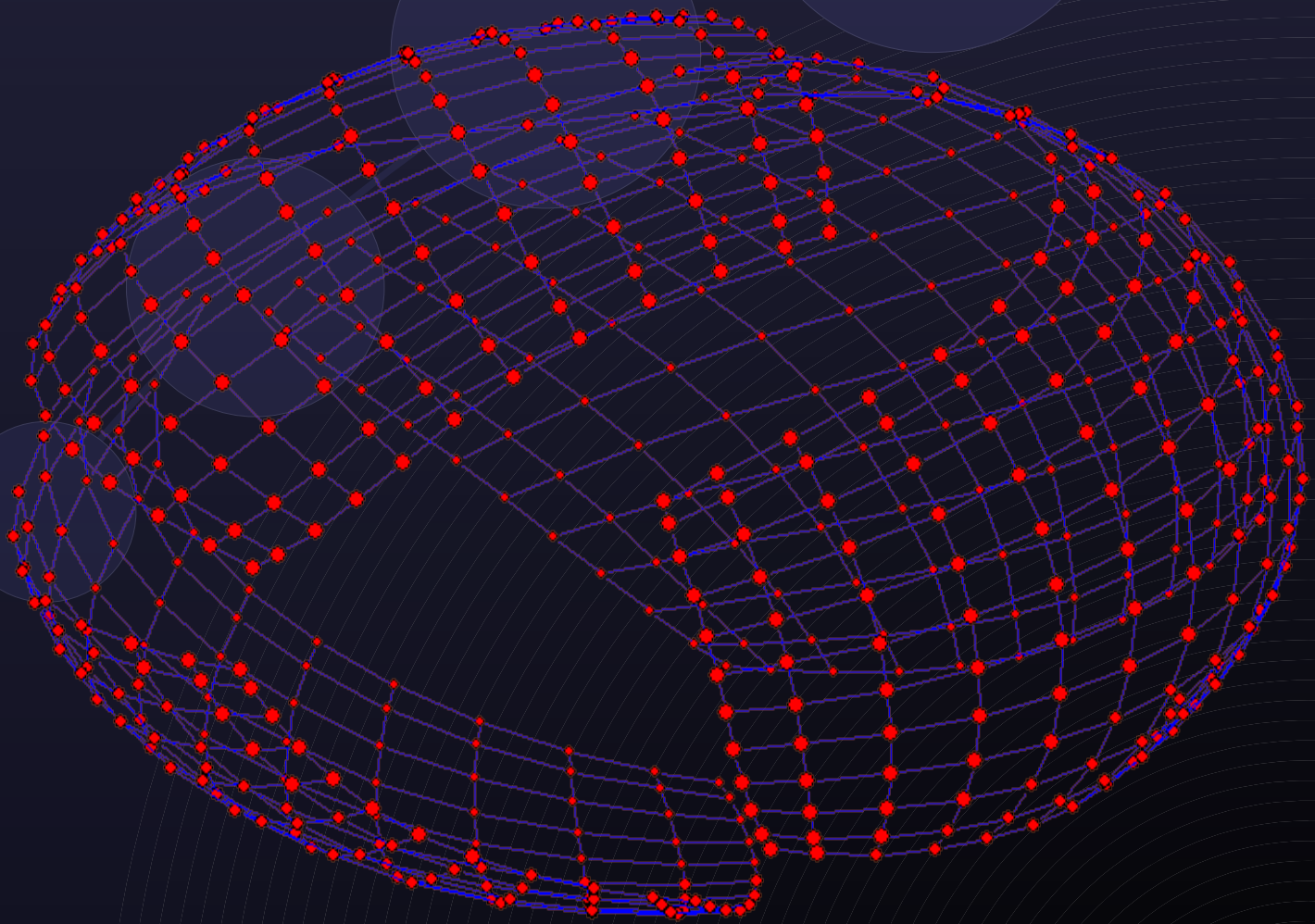
Vizuálna analýza

- Je iteratívny proces, ktorý zahŕňa:
  - Zbieranie dát
  - Spracovanie dát
  - Reprezentovanie znalostí
  - Interakcia
  - Rozhodovanie

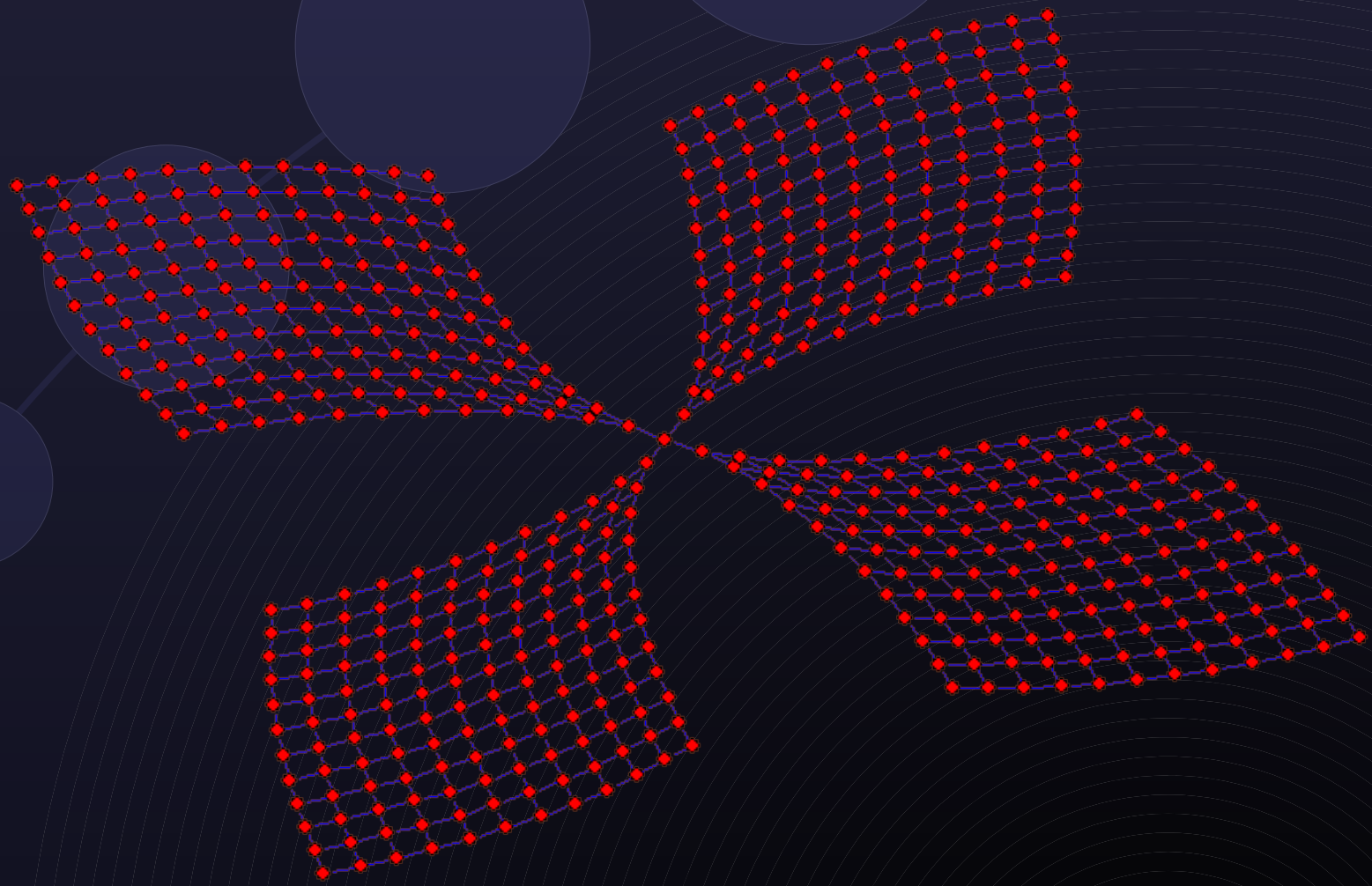
Príklad - dáta

```
*vertices    577
     1 "96"                          0.8466    0.4138    0.7969
     2 "129"                         0.6201    0.5566    0.0623
     3 "85"                          0.1176    0.2683    0.4126
     4 "34"                          0.2916    0.5293    0.9077
     5 "111"                         0.3779    0.7583    0.1404
     6 "46"                          0.5084    0.1362    0.7958
     7 "129"                         0.3966    0.0528    0.5924
     8 "72"                          0.5845    0.8913    0.7233
     9 "138"                         0.5389    0.6703    0.9238
    10 "36"                          0.8604    0.2249    0.3788
    11 "70"                          0.5489    0.2934    0.9155
    12 "15"                          0.9400    0.6278    0.5281
    13 "92"                          0.8779    0.2391    0.5945
    14 "113"                         0.5302    0.6782    0.0785
    15 "58"                          0.5311    0.2204    0.8732
    16 "127"                         0.5535    0.4433    0.0488
    17 "10"                          0.1236    0.7587    0.5340
    18 "122"                         0.7407    0.4509    0.8922
    19 "87"                          0.8451    0.2151    0.6418
    20 "114"                         0.5052    0.0328    0.4340
    21 "81"                          0.1233    0.4571    0.2484
    22 "112"                         0.4988    0.6446    0.0645
    23 "47"                          0.6066    0.7815    0.8437
    24 "40"                          0.6671    0.7251    0.8608
    25 "143"                         0.5239    0.9274    0.3344
```

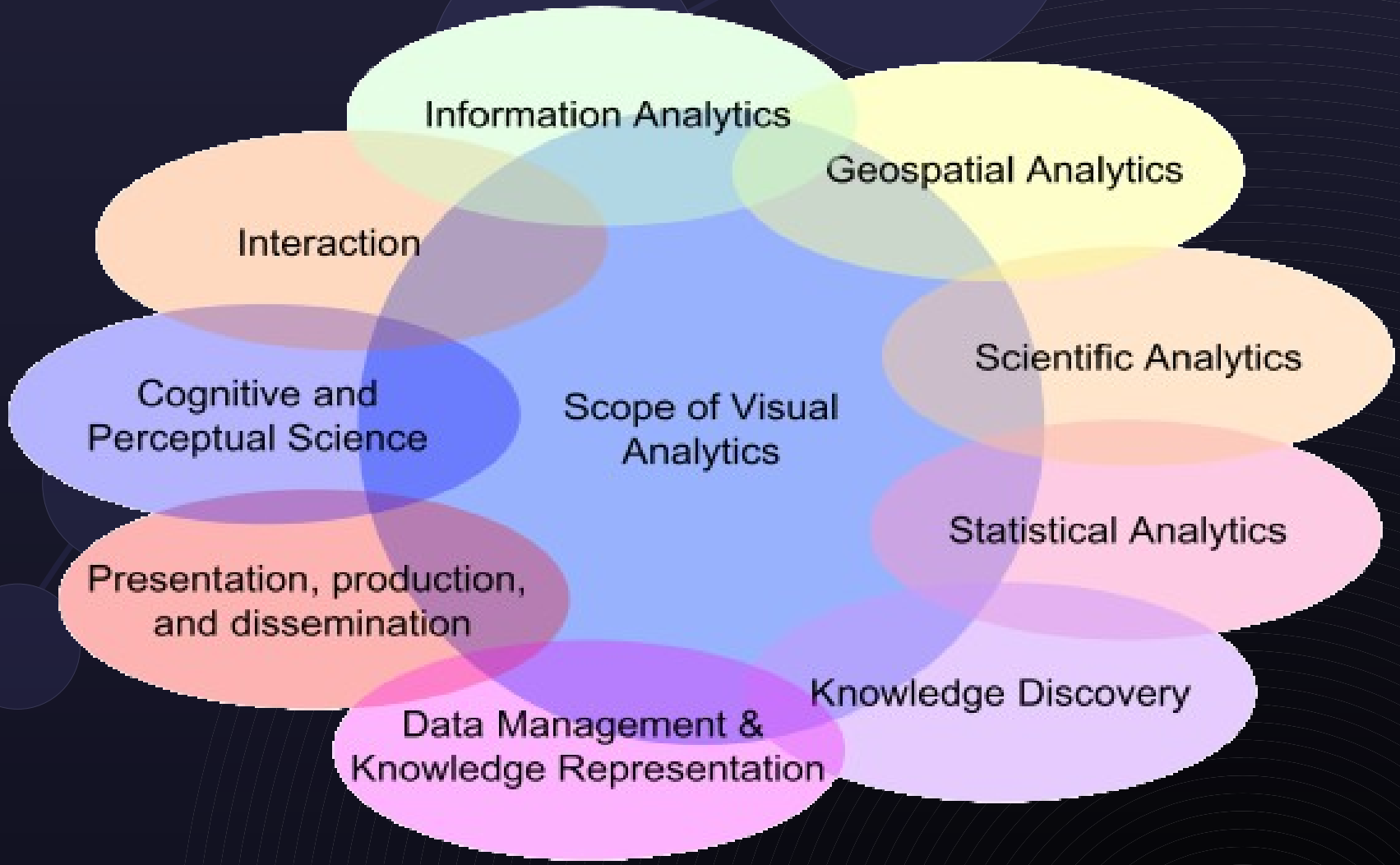
Príklad - vizualizácia



Príklad - vizualizácia

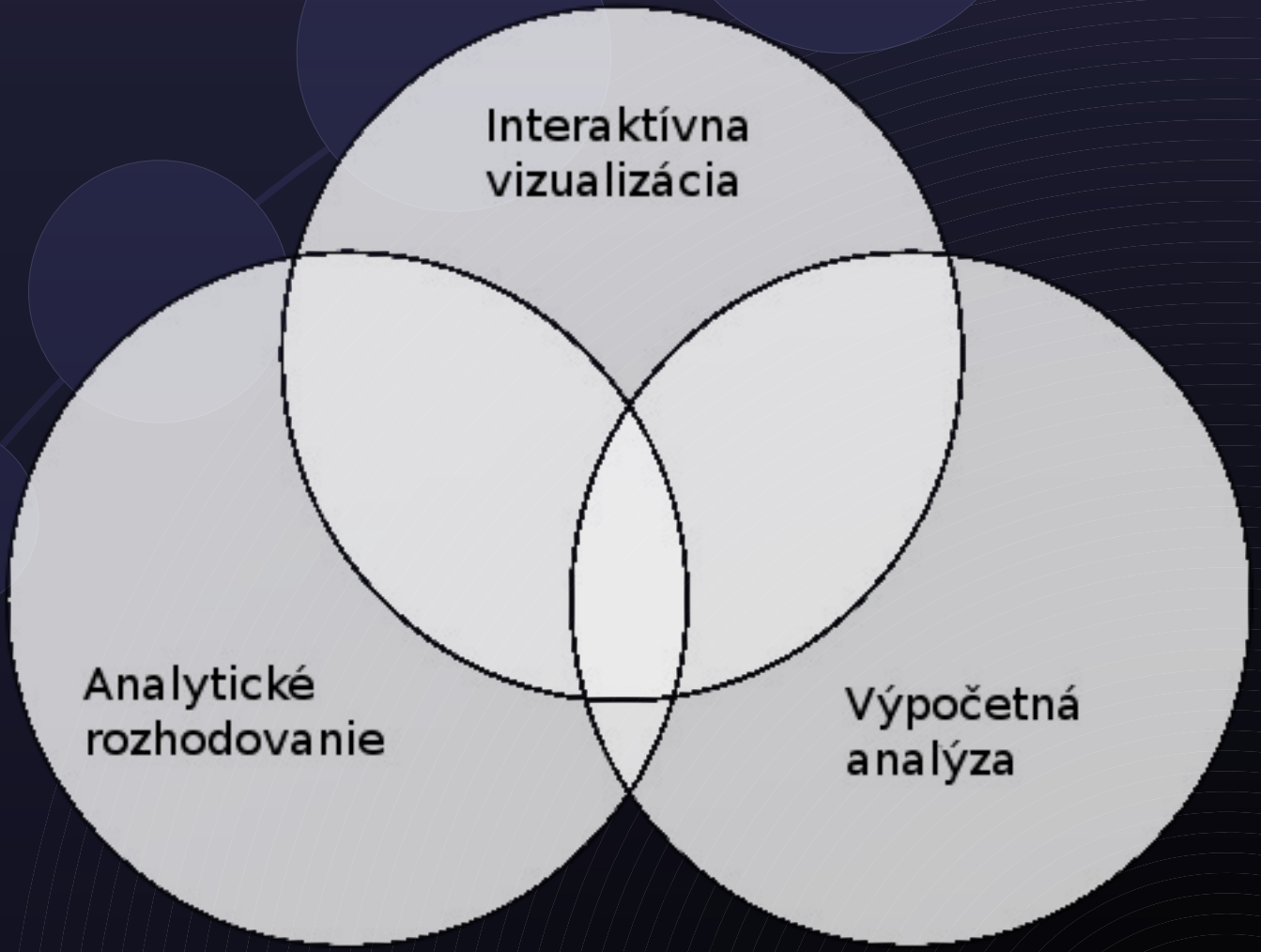


Rozsah pôsobnosti vizuálnej analýzy



Keim at al., 2007 [1]

Hlavné komponenty

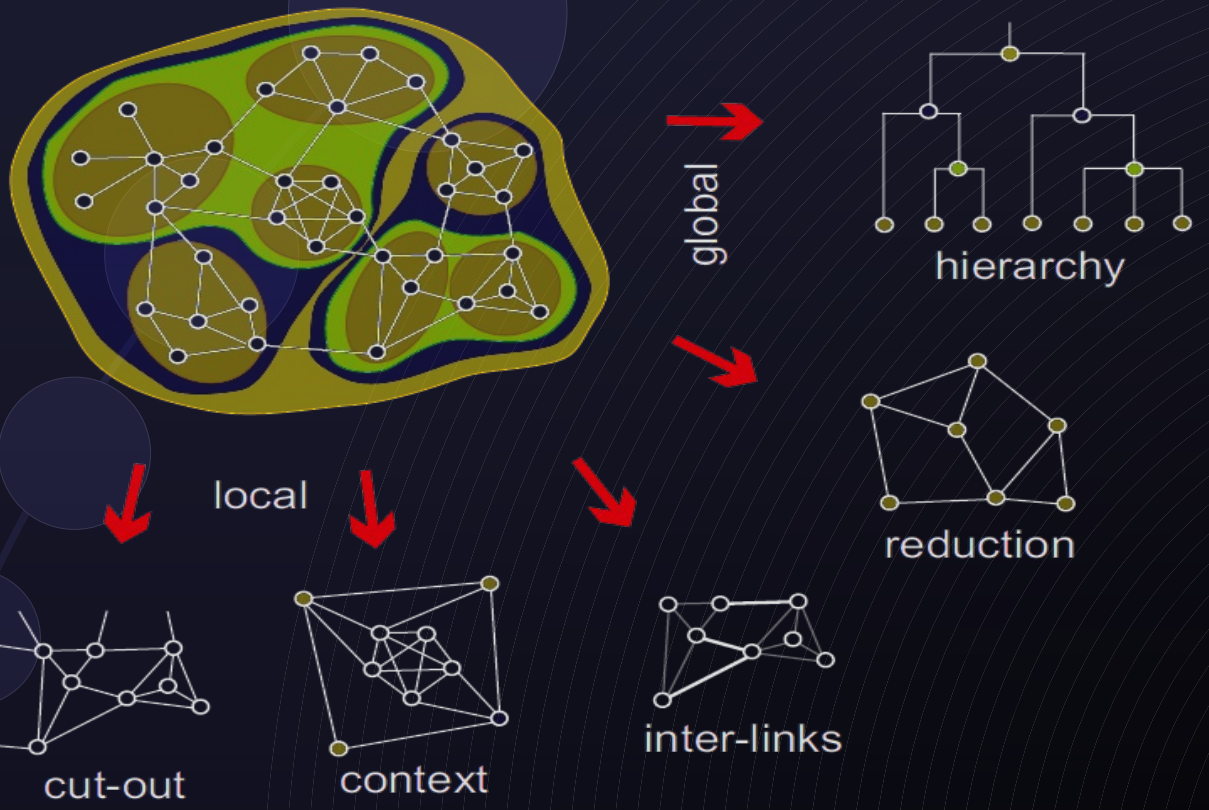


Metódy vizualizácií

- Grafy
- Diagramy
- Stromy
- Geograficko-priestorová
- (Farebná abeceda)

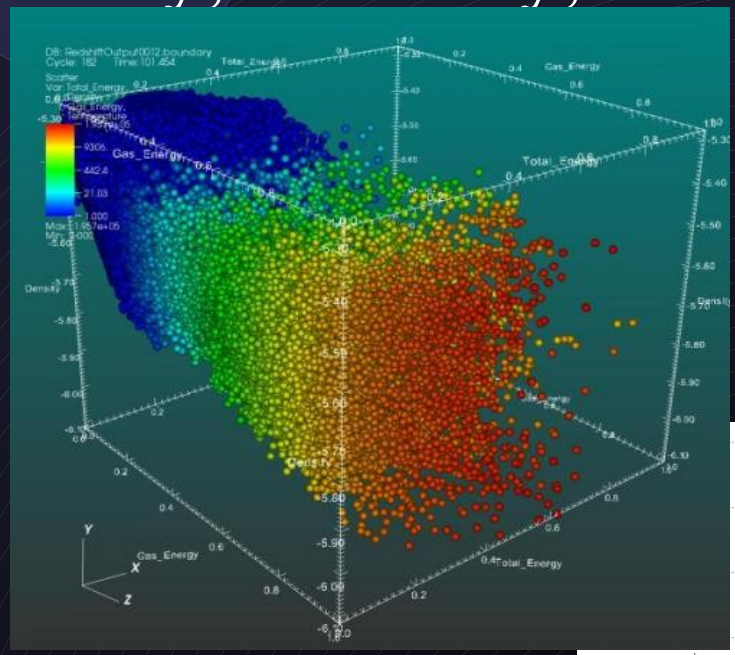
Grafy

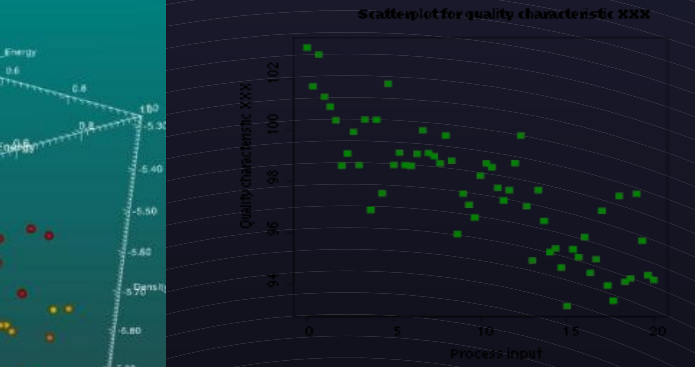
- Veľkosť a farba vrcholu
- Extrakcia podčastí grafu
- Hrúbka a farba hrany

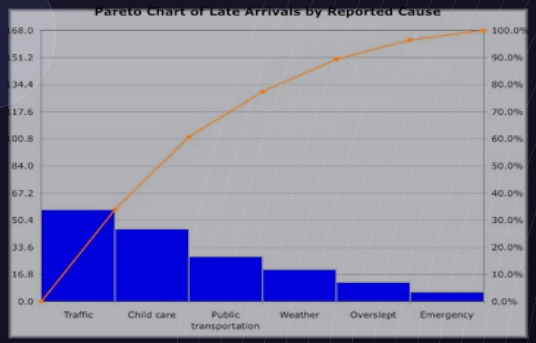


Diagramy

- Histogram, paličkový, koláčový, časová os, krivka, tok
- Bodkový
- Paretov graf
- Sviečkový

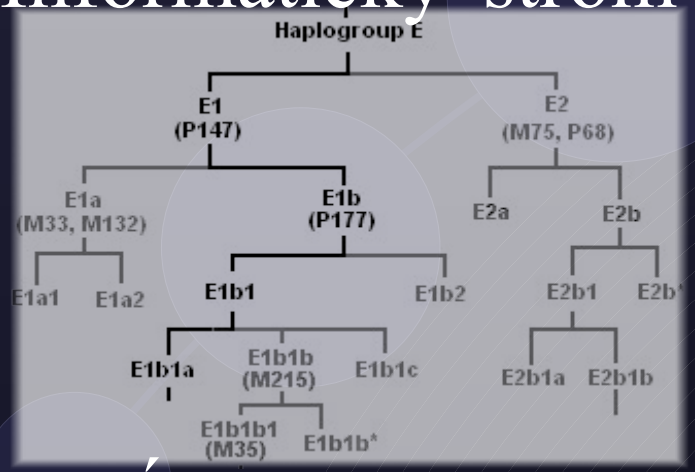




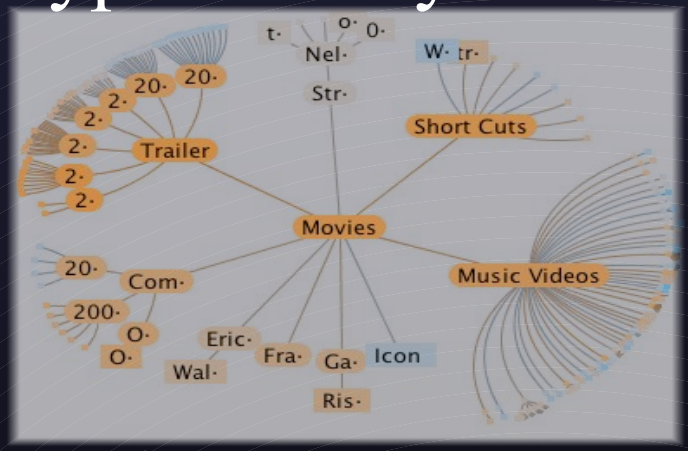


Stromy I.

- 'Informatický' strom :-)

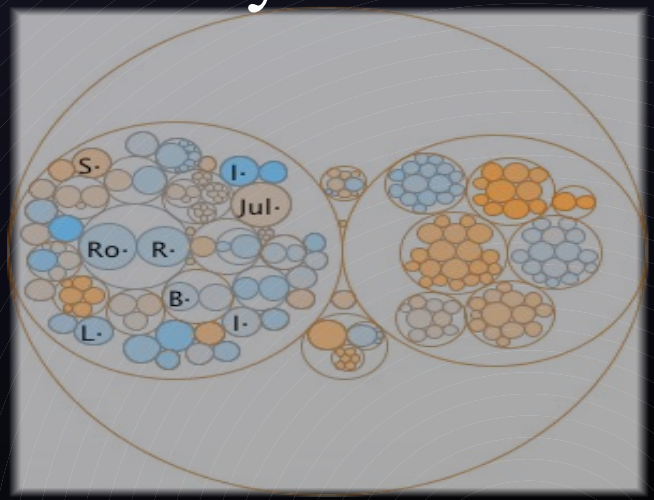


- Hyperbolický strom



- Obdĺžnikový strom

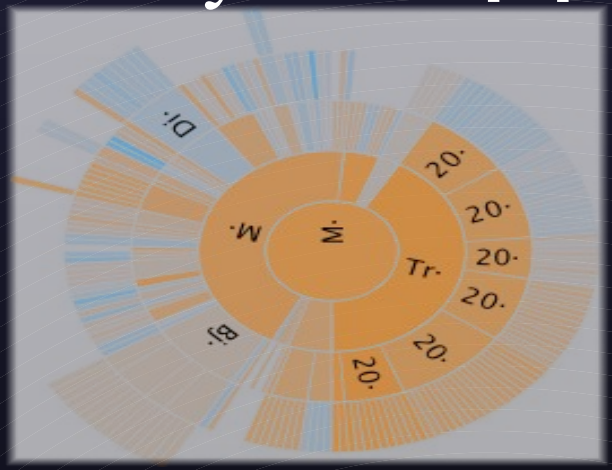
- Kruhový strom



Stromy II.

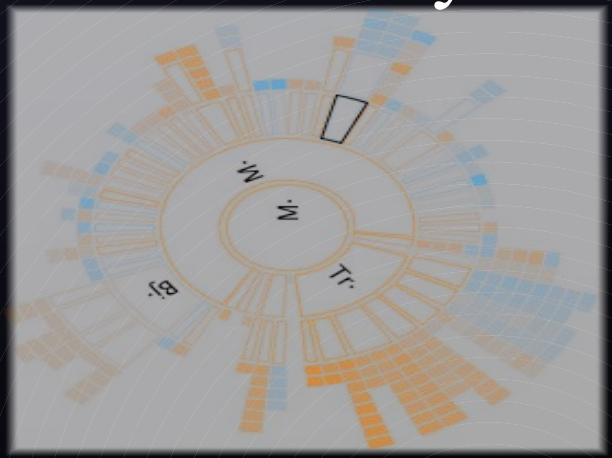
- Cencúľový strom (icicle)

- Slnečný strom [10]



- Strom 'ľadový lúč' (iceray)

- Strom 'slnečný lúč' (sunray)



Geograficko-priestorová vizualizácia

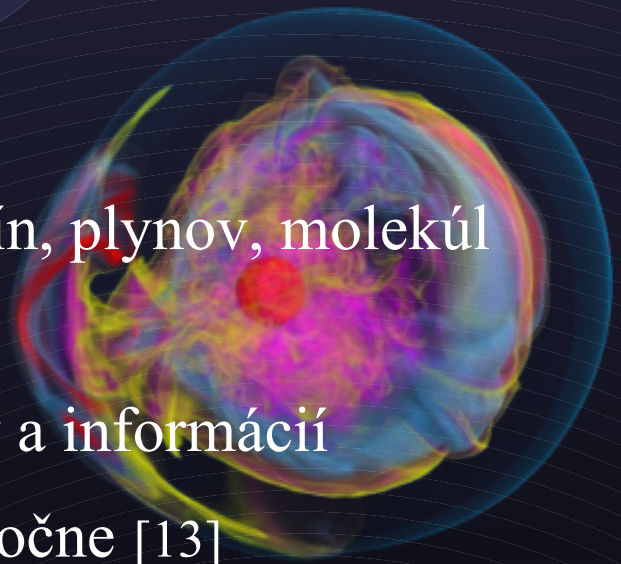
- Geografické dáta sú zakreslené do mapy
- Definovanie vzťahov a anomálií
- Príklady:
  - Správanie sa antarktických ľadovcov[3]
  - Epidémie a zdravotná starostlivosť
  - Bezpečnosť: evakuácia ľudí pri výbuchu bomby[4]
  - Požiare - http://theivac.org/content/pie-fire-video
- http://geoanalytics.net

Farebná abeceda

- Každé písmeno abecedy má svoju farbu
- Iný pohľad na text
- Viac o farebnej abecede: http://www.christianfaur.com/color/
  - Ofarbenie online
  - Font

10 najväčších výziev

1. Fyzika a astronómia
   - Vizualizácia toku, dynamika kvapalín, plynov, molekúl
   - Terabajty dát obsahujúcich šum
   - Objavovanie nových javov, vzťahov a informácií
   - CERN – LHC produkuje 15PB dát ročne [13]



2. Firmy
   - Sklady, komodity, cenné papiere, meny, burza
   - Analýza minulých a súčastných stavov, predpovede

10 najväčších výziev

3. Sledovanie životného prostredia
   - Terabajty dát zozberianých po svete alebo zo satelitov
   - Analýza minulých stavov a predpoveď do budúcnosti
4. Katastrofy a správa krízových situácií
   - Vizuálna analýza môže upresniť postup pri katastrofách (povodne, hurikány, požiare, výbuch sopky, tsunami...)
   - Zistenie rozsahu škôd, identifikácia cieľov, stanovenie priorít a efektívna koordinácia rôznych špecialistov v zasiahnutých oblastiach

10 najväčších výziev

5. Bezpečnosť

- Informácie o teroristoch a incidentoch
- Spájanie týchto informácií a vyhľadávanie súvislostí
- VisAware – kto, kde, kedy.

6. Softvérová analýza

- Analýza zdrojových kódov softvéru a jeho závislostí
- Debugovanie, správa, optimalizácia, reštrukturalizácia

10 najväčších výziev

7. Biológia, medicína a zdravie

- Genetika – ľudský genóm približne 3mld. Nukleotidov
- Proteomika, metabolické dráhy (foldit - http://fold.it)

8. Inžinierstvo

- Optimalizácia toku – zobrazenie odporu vetra/vody
- Nárazové testy automobilov – automobil zložený zo státisícov kociek
- Vizuálna analýza môže pomôcť návrhárom pochopiť deformáciu pri náraze a identifikovať kľúčové body kde je nevyhnutná optimalizácia

10 najväčších výziev

9. Osobný informačný manažment

- Efektívna analýza osobnej e-mailovej komunikácie

10. Doprava

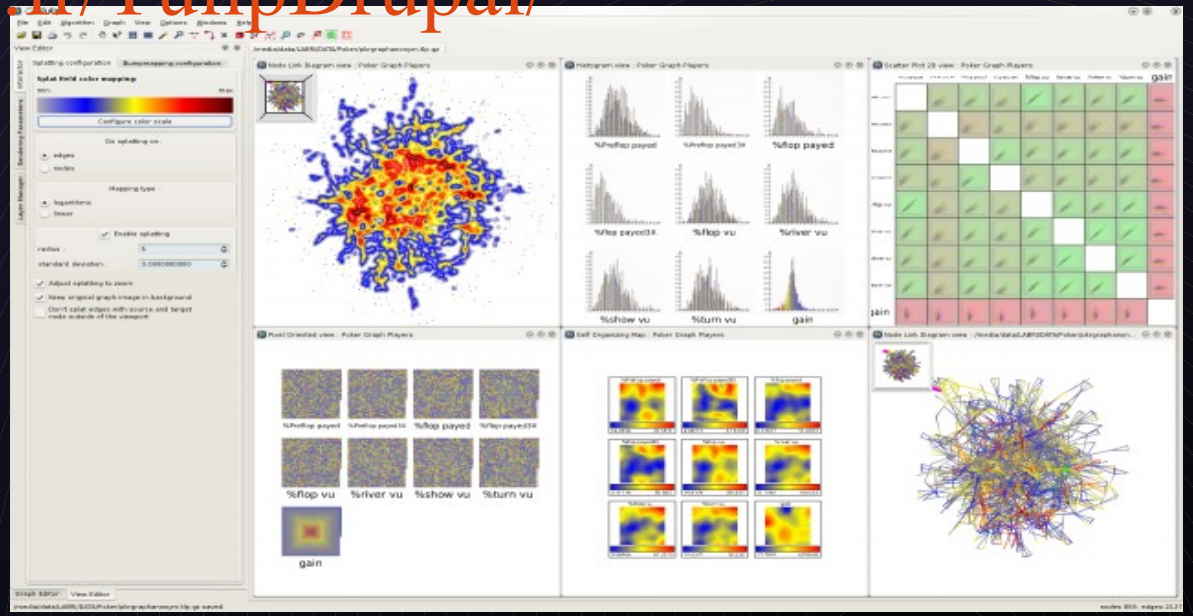
- Množstvo senzorov – detekcia váhy vozidla, vyťaženia cesty, kamery, GPS, textové správy o nehodách, informácie o počasí
- Analýza aktuálneho stavu dopravnej siete
- Algoritmy na detekciu zmien v toku
- Predpoveď dopravnej situácie

Tulip

LaBRI

INRIA

- Analýza a vizualizácia relačných dát
- Napísaný v c++, framework umožňuje rozširovanie (deb, exe na sourceforge, LGPL lic.)
- http://tulip.labri.fr/TulipDrupal/



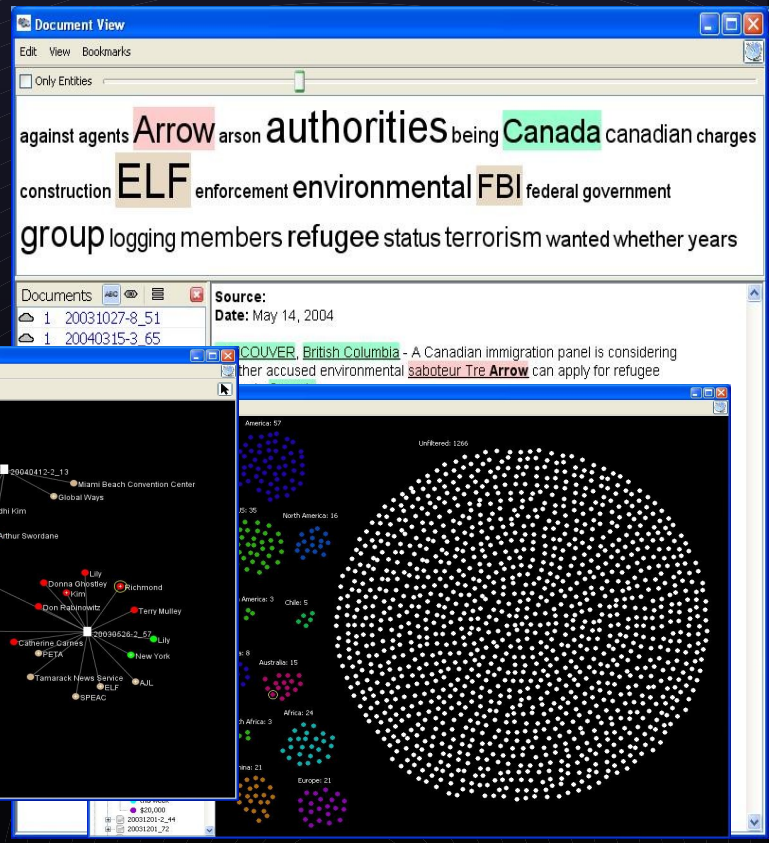
Pajek

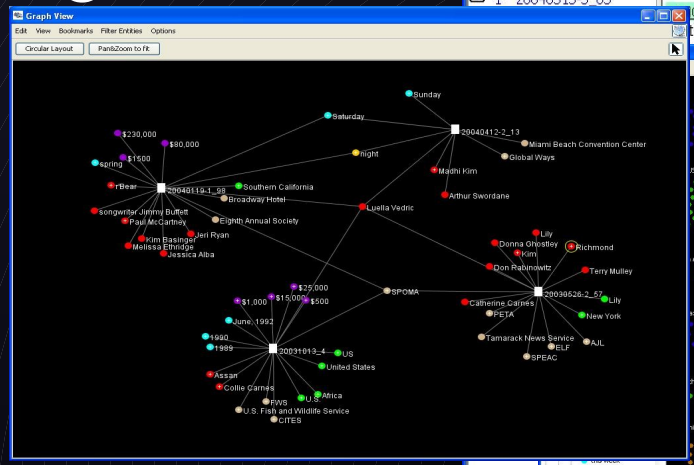
- Program na analýzu grafových štruktúr
- Rozdelenie siete na menšie časti
- Úpravy siete
- Pre nekomerčné použitie zdarma

jigsaw

- Analýza neštrukturovaných textových dokumentov[11]
- Ako sú dokumenty medzi sebou prepojené
- Zameraný na jednotlivé dokumenty a ich obsah
  - Ľudia, miesta, organizácie...

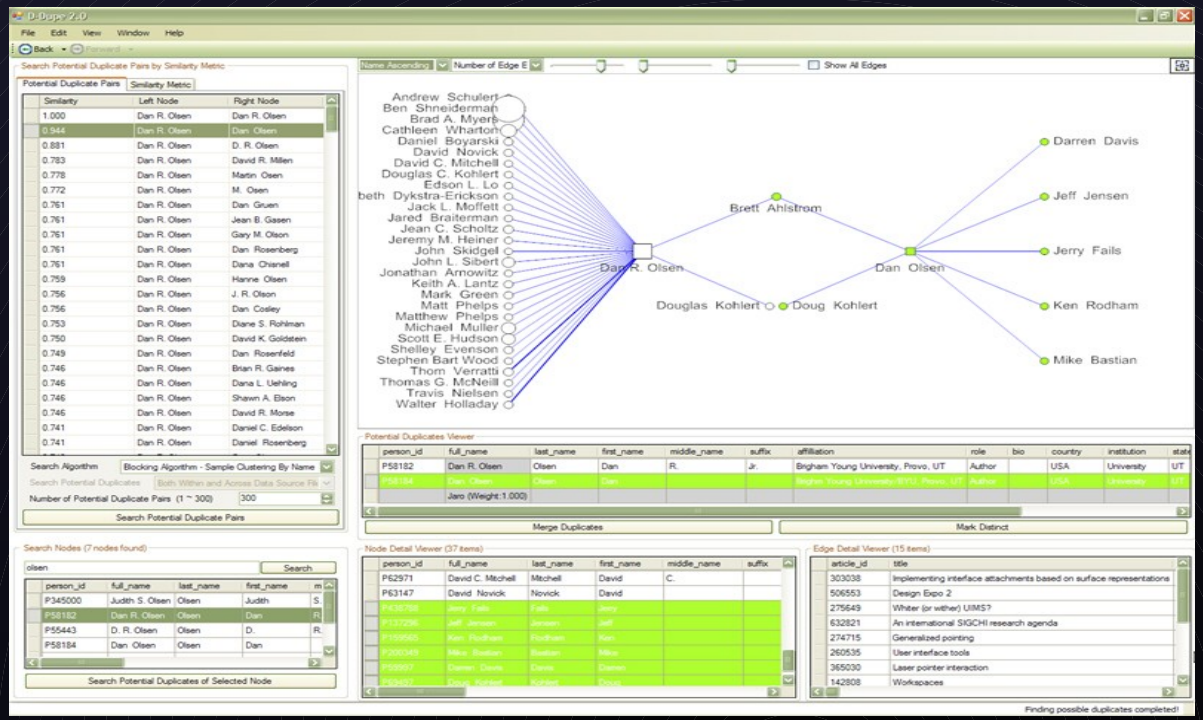
pozn. platené z grantov





D-Dupe

- Hľadanie potencionálnych duplicitných uzlov v sieťach. Napr. bibliografické zdroje. [12]
- Pre nekomerčné použitie zdarma
- Len pre Windows

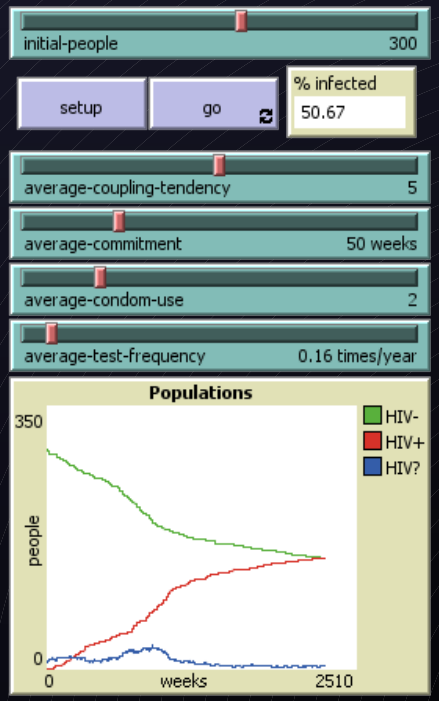


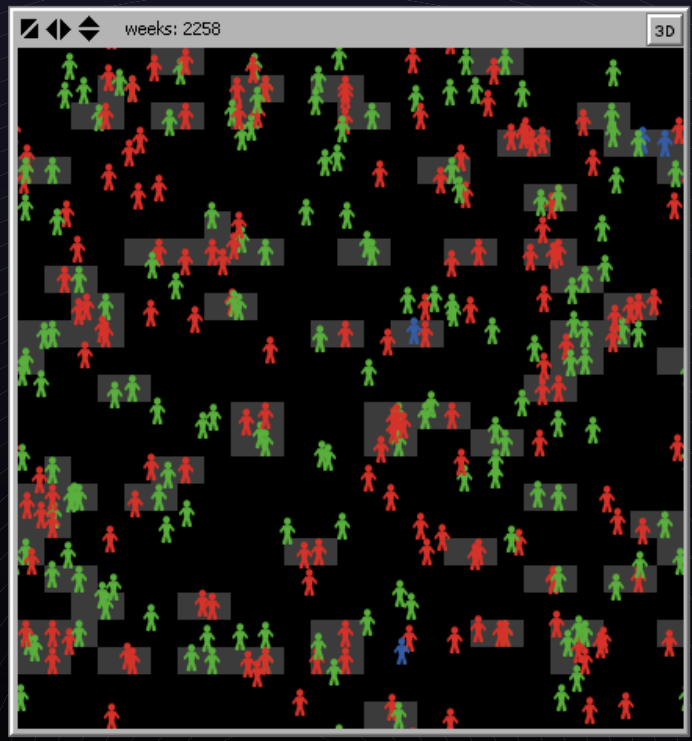
IN-SPIRE

- IN-SPIRE - http://in-spire.pnl.gov/
  - http://www.youtube.com/watch?v=7bIRmJlhjbQ
- Zameraný na veľké kolekcie dokumentov
- Zhlukovanie kolekcií podľa témy [11]
- Stránka momentálne nedostupná

Ďalšie softvérové nástroje

- Štatistické programy
  - Statistica, SPSS
- Maple
- Mathematica
- Simulačné programy
  - NetLogo, Stella





Ďakujem za pozornosť :-)

Literatúra

[1] Keim, Daniel a., Florian Mansmann, and Jim Thomas. "Visual analytics." ACM SIGKDD Explorations Newsletter 11, no. 2 (2010): 5. http://portal.acm.org/citation.cfm?doid=1809400.1809403.

[2] Thomas, J.J., Cook, K.A.: Illuminating the Path. IEEE Computer Society Press, Los Alamitos (2005)

[3] Turdukulov, Ulanbek, Connie Blok. "Visual analytics to explore iceberg movement." Geo-Information Science 2008: 1-4. http://geoanalytics.net/GeoVis08/a21.pdf.

[4] Andrienko, Gennady, Natalia Andrienko. "Geospatial Visual Analytics : GeoPKDD project," September 2008. http://geoanalytics.net/GeoVisualAnalytics08/s13.pdf.

[5] W.J. Frawley, G. Piatetsky-Shapiro, and C.J. Matheus. Knowledge discovery in databases: an overview. In G. Piatetsky-Shapiro and W.J. Frawley, editors, Knowledge Discovery in Databases. AAAI / MIT Press, 1991.

[6] Lipo, Carl; O'Brien, Michael J., eds (2006). Mapping Our Ancestors: Phylogenetic Approaches in Anthropology and Prehistory. Piscataway: Transaction Publishers.

[7] Robinson, Peter M.W.; O'Hara, Robert J. (1996). "Cladistic analysis of an Old Norse manuscript tradition". Research in Humanities Computing 4: 115–137. http://rjohara.net/cv/1996-rhc.

Literatúra

[8] Jerison, Harry J. (2003), "On Theory in Comparative Psychology", in Sternberg, Robert J.; Kaufman, James C., The evolution of intelligence, Mahwah, NJ: Lawrence Erlbaum Associates, Inc., p. 254 .

[9] IDC Go-to-Market Services: The Digital Universe Decade - Are you Ready?. URL: <http://www.emc.com/collateral/demos/microsites/idc-digital-universe/iview.htm> [cit. 2010-06-27].

[10] Visualization of large tree structures.URL:<http://www.randelshofer.ch/treeviz/> [cit. 2010-06-27]

[11] Jigsaw: Visualization for Investigative Analysis. URL: <http://www.cc.gatech.edu/gvu/ii/jigsaw/vacviews-jigsaw.pdf> [cit. 2010-06-27]

[12] D-Dupe: A Novel Tool for Interactive Data Deduplication and Integration. URL:<http://www.cs.umd.edu/projects/linqs/ddupe/index.htm> [cit. 2010-06-27]

[13] Worldwide LHC Computing Grid. URL: <http://public.web.cern.ch/public/en/lhc/Computing-en.html> [cit. 2010-06-27]

[14]Batagelj, Vladimir; Mrvar, Andrej. Pajek [počítačový program]. Ver. 1.26. [Slovinsko], 1998 [cit. 2010-02-06]. Dostupné z <http://vlado.fmf.uni-lj.si/pub/networks/pajek/pajek126.exe>. Zdarma pre nekomerčné použitie.

Literatúra

[15]Batagelj, Vladimir; Mrvar, Andrej. Pajek - Program for Analysis and Visualization of Large Networks: Reference Manual, 5 Máj 2009, posledná aktualizácia 7 Január 2010. URL: < http://pajek.imfm.si/lib/exe/fetch.phpid=download&cache=cache&media=dl:pajekman126.pdf>.

[16]Boldiš, Petr. Bibliografické citace dokumentů podle CSN ISO 690 a CSN ISO 690-2: Část 2 – Modely a příklady citací u jednotlivých typů dokumentů. Verze 3.0 (2004). c 1999–2004, posledná aktualizácia 11. 11. 2004. URL: <http://www.boldis.cz/citace/citace2.pdf>.

[17] Gregorovič, Tomáš. Extrakce informací ze sociálních médií, diplomová práca, 8 Február 2010. URL: <https://is.muni.cz/th/139855/fi_m/dp.pdf>.