Application of Data Visualization

Lunteren Conference

Landelijk Netwerk Mathematische Besliskunde (LNMB)

January 15 2015, Lunteren

dr.ir. Danny Holten
Lead Visualization Scientist & Co-Founder
danny.holten@synerscope.com
+31 (0)6 5202 1711
SynerScope B.V.

Overview

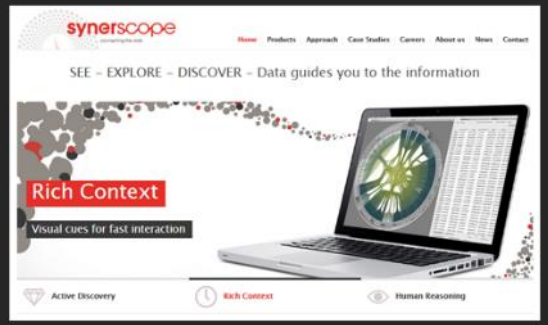
- Background SynerScope & Danny
- “Big Data”
- Making sense of Big Data
  - Analytics & visualization → visual analytics
- Offering a solution
  - Technology, users, domains
- Demo, questions, ...

Danny

- MSc, PhD Comp. Sci., Eindhoven University of Technology (TU/e)
  - "Visualization of Graphs and Trees for Program Understanding"
  - Help SW engineers understand source code using interactive and insightful visual overviews of the SW architecture

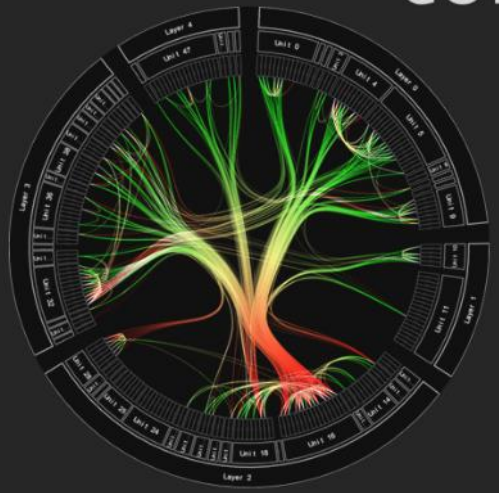
- Information Visualization (InfoVis)
  - "The study of interactive visual representations of abstract data to reinforce human cognition"

- Now
  - Lead Visualization Scientist & co-founder at SynerScope B.V.

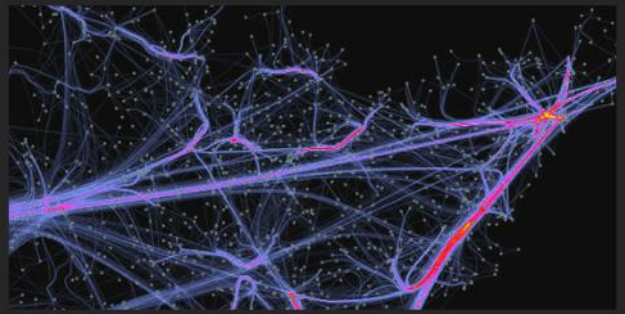


synerscope

connecting the dots



computer
analysis
visiulization
imaging
user
perception
geometry
big
visualization
programming
medical
OpenGL
science
C++
infovis
graphics
analytics
data
hardware
CUDA
SynerScope
scivis
vision
interaction
computational
infographics
learning
machine
studies
WebGL

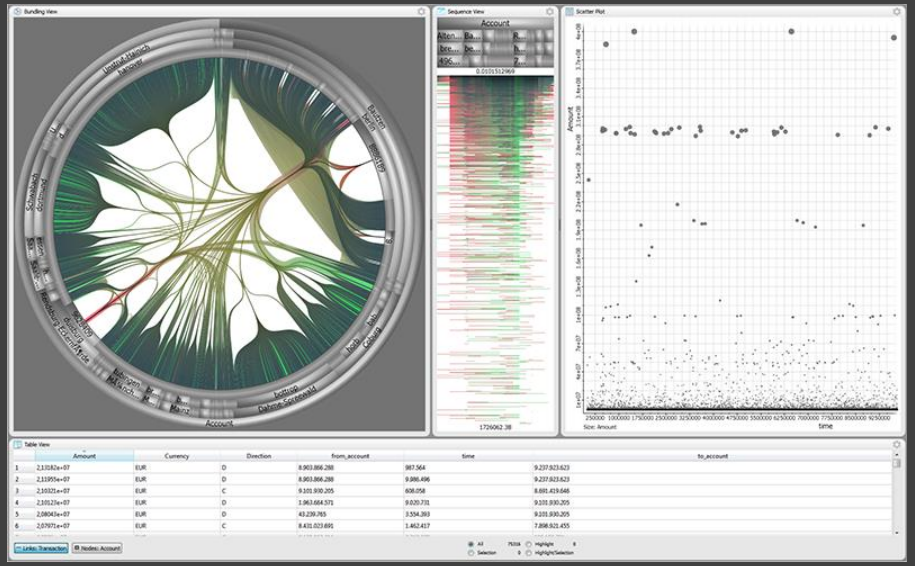


SynerScope B.V.

- Came in contact with Jan-Kees Buenen (CEO) in 2008
  - 1959, MBA Chicago Booth University
- Can we apply this to industrial/commercial data?
  - "Big Data visual analytics software that allows domain experts and analysts to make sense of their data"
  - Initially for source code, now for generic Big Data
- Spin-off of InfoVis PhD research at TU/e
  - Founded in 2011, 16 people, ~~TU/e campus~~ Helvoirt

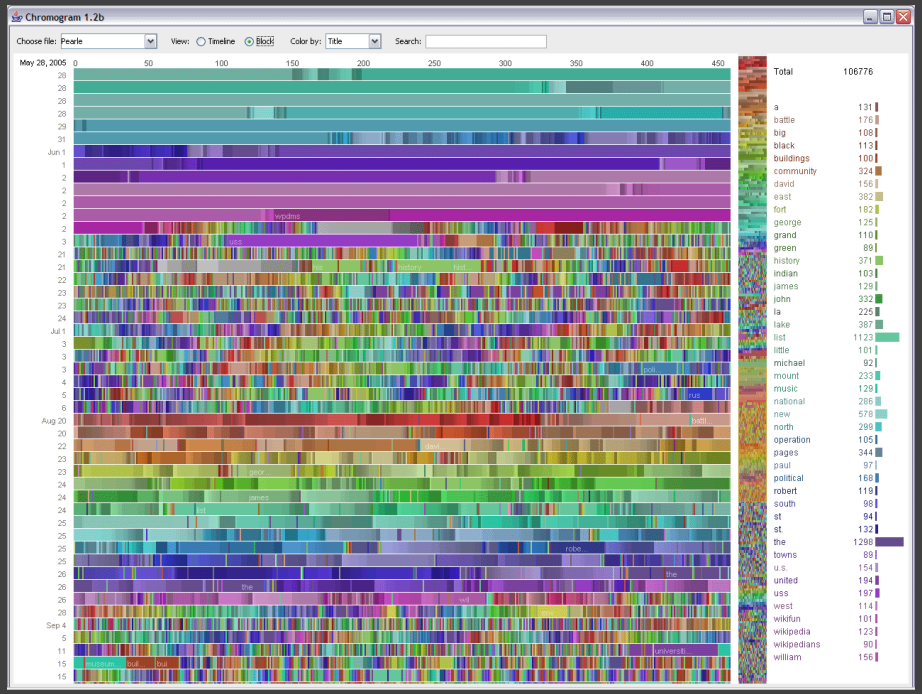
“As a telescope or microscope extend human vision...

...SynerScope extends human analytic capability.”



Big Data (BD)

"All-encompassing term for data so large and complex that it becomes difficult to process using traditional processing applications. Challenges include analysis, capture, curation, search, sharing, storage, transfer, visualization, and privacy violations"

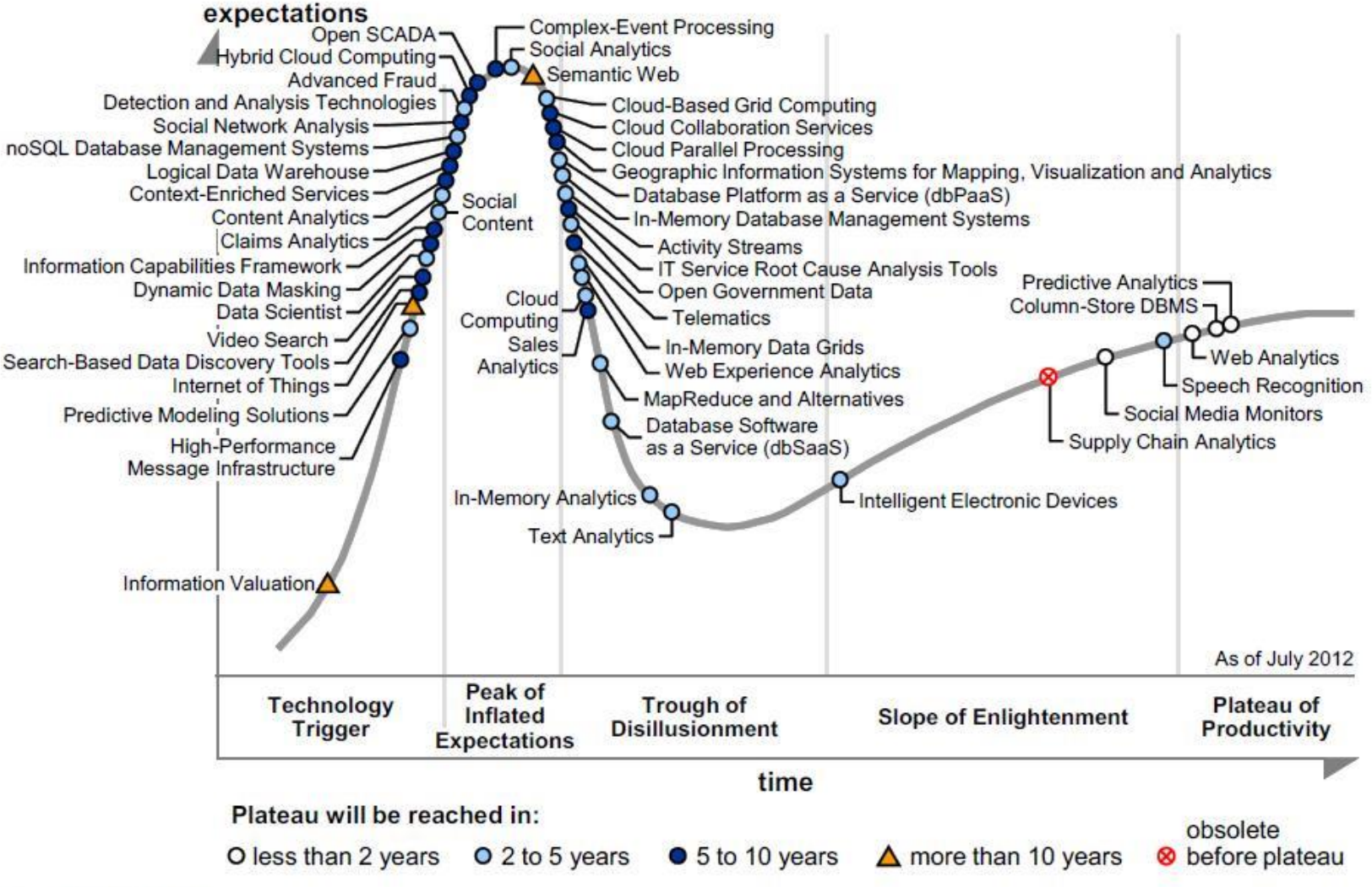


Big Data (BD)

- "The Four V's"
  - Volume, Variety, Velocity, Variability
  - Storage, Multivariate, Streaming, Inconsistent (structure)
- Get value / make sense / get "insight"!
  - Value from large (often already available) data from various sources by combining them and ontaining new, actionable insights
- As an industry

  "Software AG, Oracle, IBM, FICO, Microsoft, SAP, EMC, HP, and Dell spent >$15B on firms specializing in data management and analytics. In 2010, the BD industry was worth >$100B and growing at 10% / year; twice as fast as the software business as a whole" – [The Economist]

Figure 1. Hype Cycle for Big Data, 2012



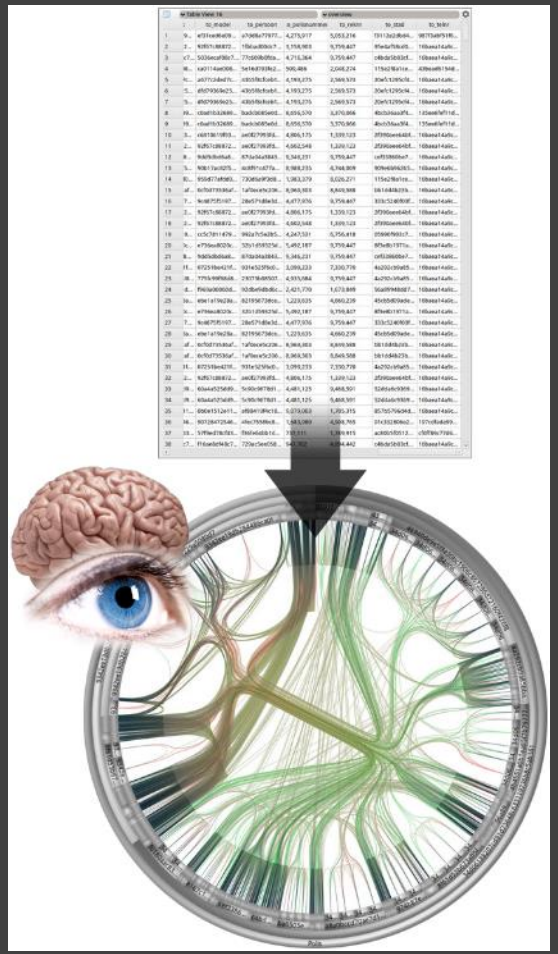
Source: Gartner (July 2012)

Analytics

- Statistical / mathematical techniques
  - Simple summary statistics, regression, model fitting, PCA/MDS, ...
- Machine learning / AI
  - Clustering, classification, prediction, ...
  - Decision trees/forests, Bayesian networks, neural networks, SVMs, ...
- Rule-based engines
  - Flagging, pre-filtering, aggregation, sampling, ...
- Available as libraries / code for “easy” use
- So... Problem(s) solved?

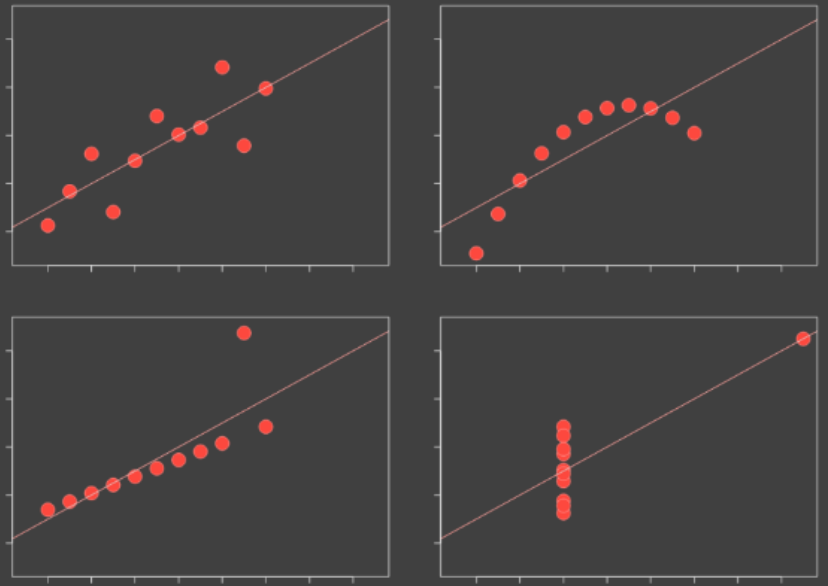
Analytics

- Not a complete solution
  - Data scientist (“data janitor”) needed
  - Users are domain experts, (business) analysts
  - Domain knowledge hard to capture
    → keep “human in the loop”!
  - What to look/code for? “get insights/overview”
- Enter Visualization…
  - Analyze interactively and explore
    → Active Discovery (AD), not just presentation!
  - Visual system evolved to detect anomalies/trends
    → excels at fuzzy pattern recognition in noisy data



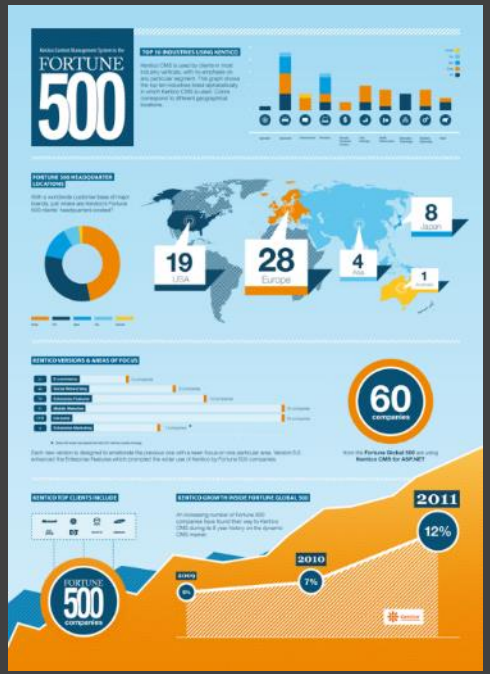
Visualization (cont.)

- Fast interactive AD / ideation cycle
  - hypothesize, test, repeat (offload IT)
- "Confirm the expected, detect the unexpected"
  - Discover new rules, automate
- "Overview first, zoom and filter, then details-on-demand"
  - Use Gestalt principles and pre-attentive processing
- Complementary to data mining & machine learning
  - Combine both into visual analytics

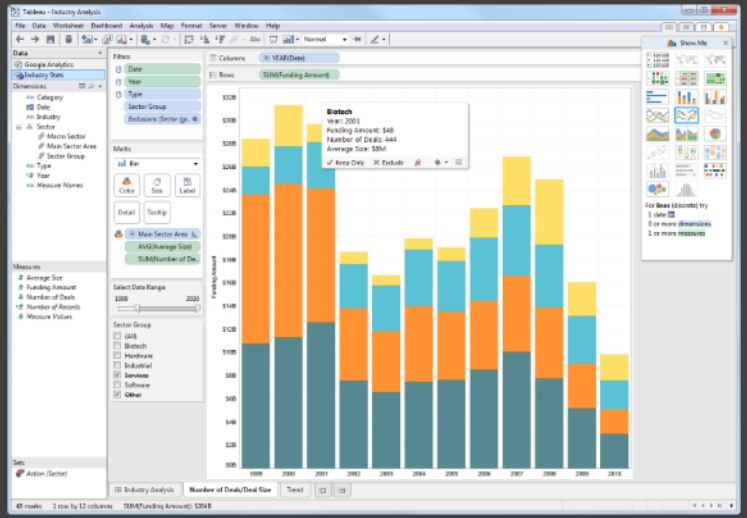


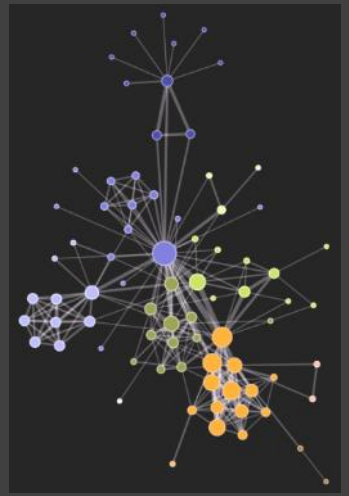
InfoVis != dashboards, infographics

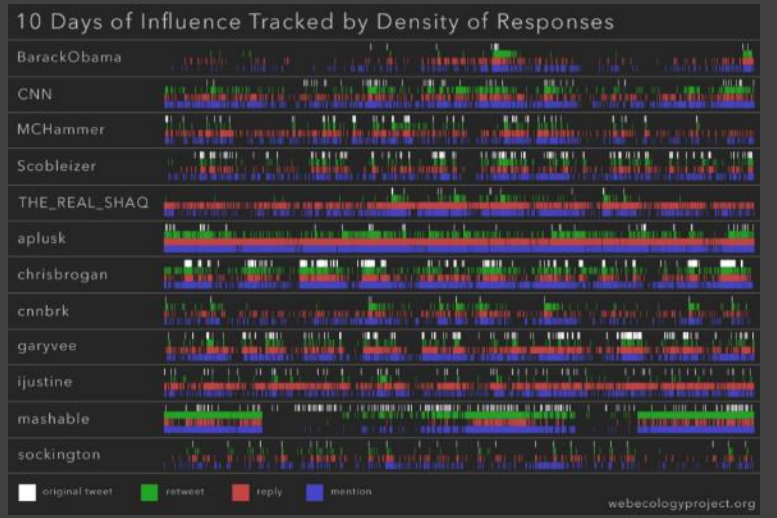
- Aggregated KPIs (monitoring)
- Limited drill-down, interaction
- Discrete and chunked
  → who, what, when, why, how?

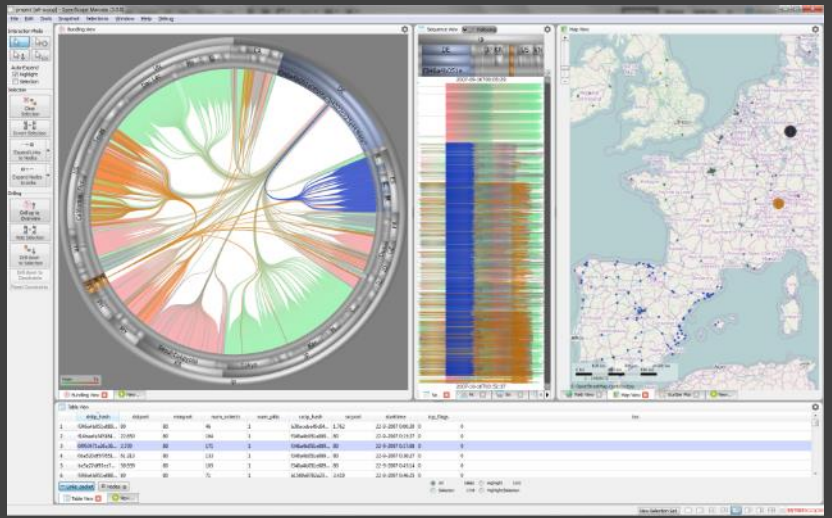


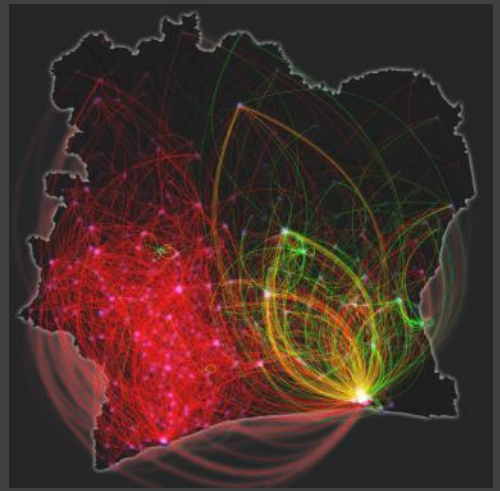
- One-off, manual
- Non-interactive, no analysis
- Storytelling

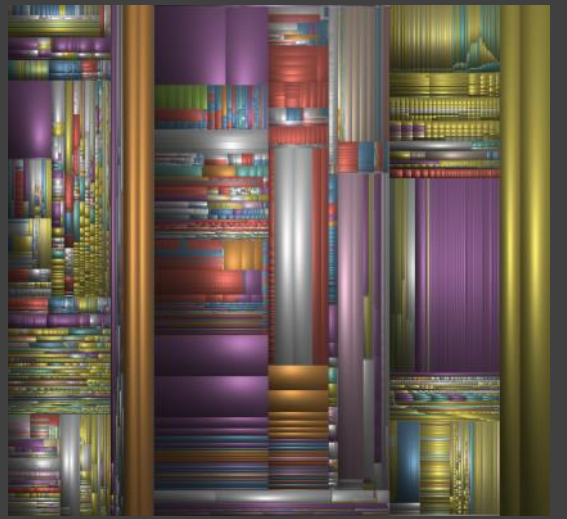





10 Days of Influence Tracked by Density of Responses
BarackObama
CNN
MCHammer
Scobleizer
THE_REAL_SHAQ
aplusk
chrisbrogan
cnnbrk
garyvee
ijustine
mashable
sockington
original tweet
retweet
reply
mention
webecologyproject.org

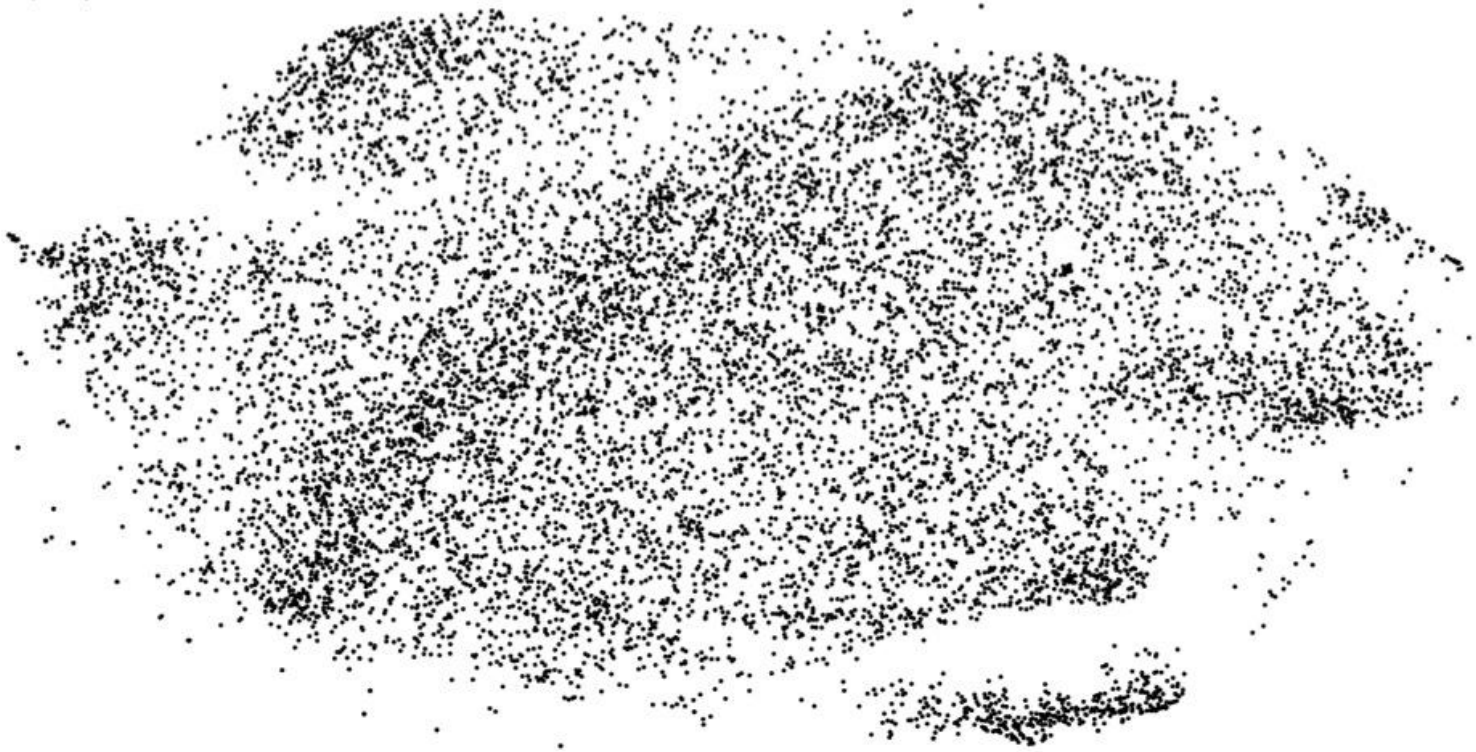


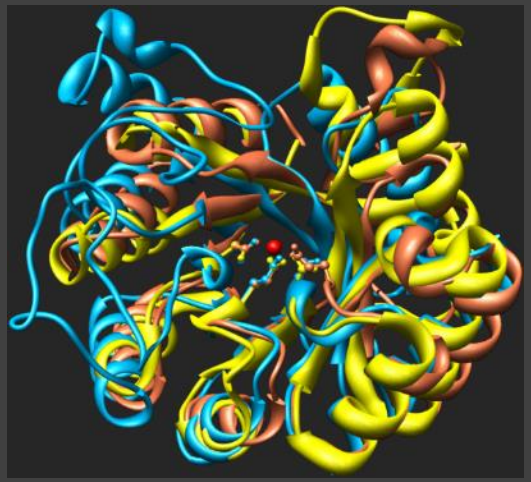


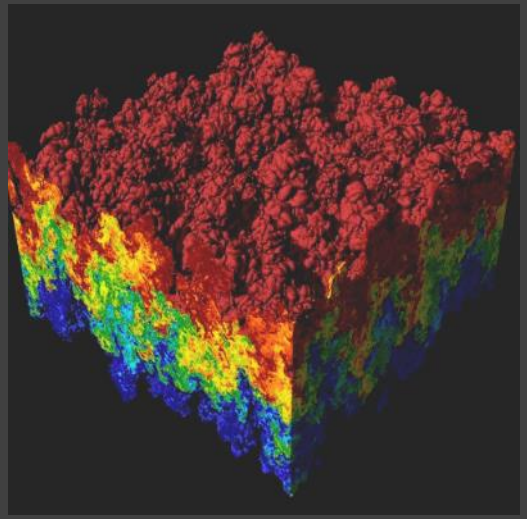


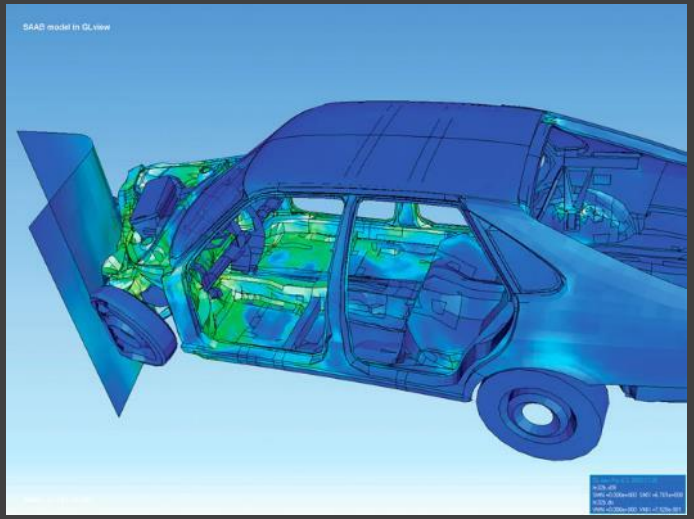
t-SNE
Distributed stochastic Neighbour Embedding

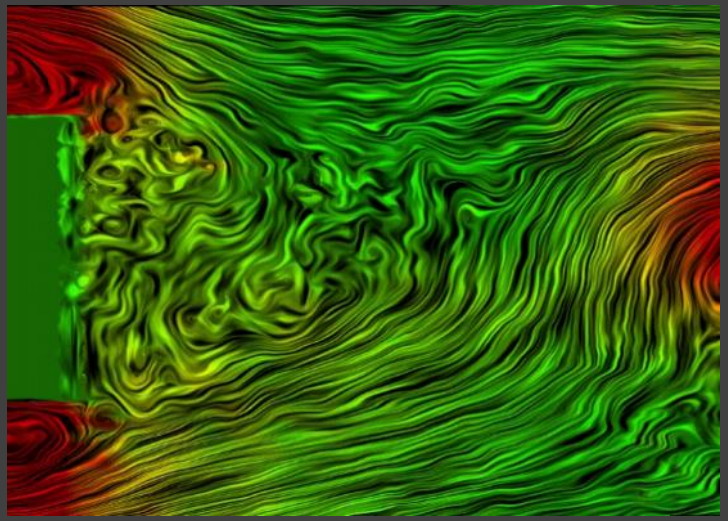
10,557 poems
Perplexity: 50

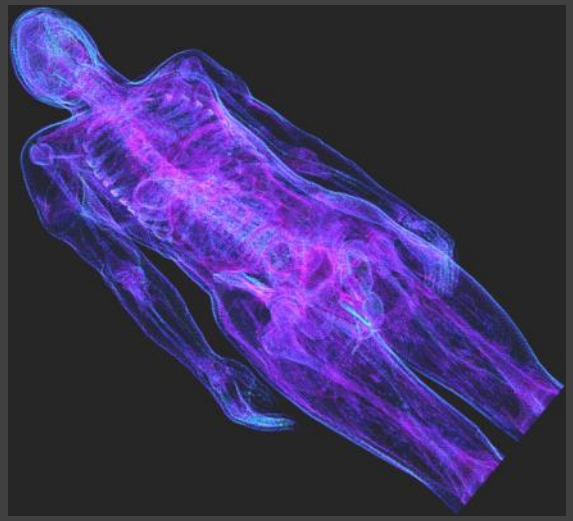


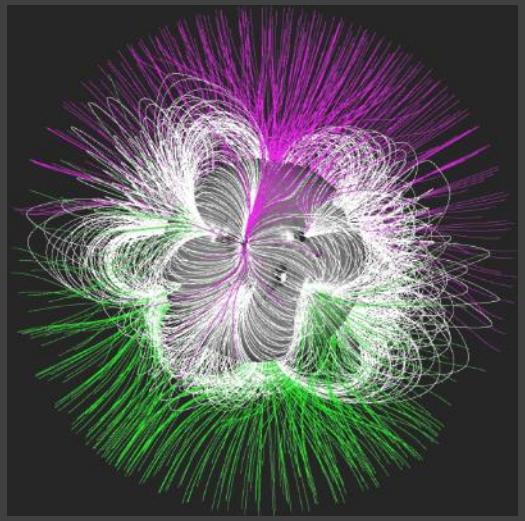





SAAB model in GLview



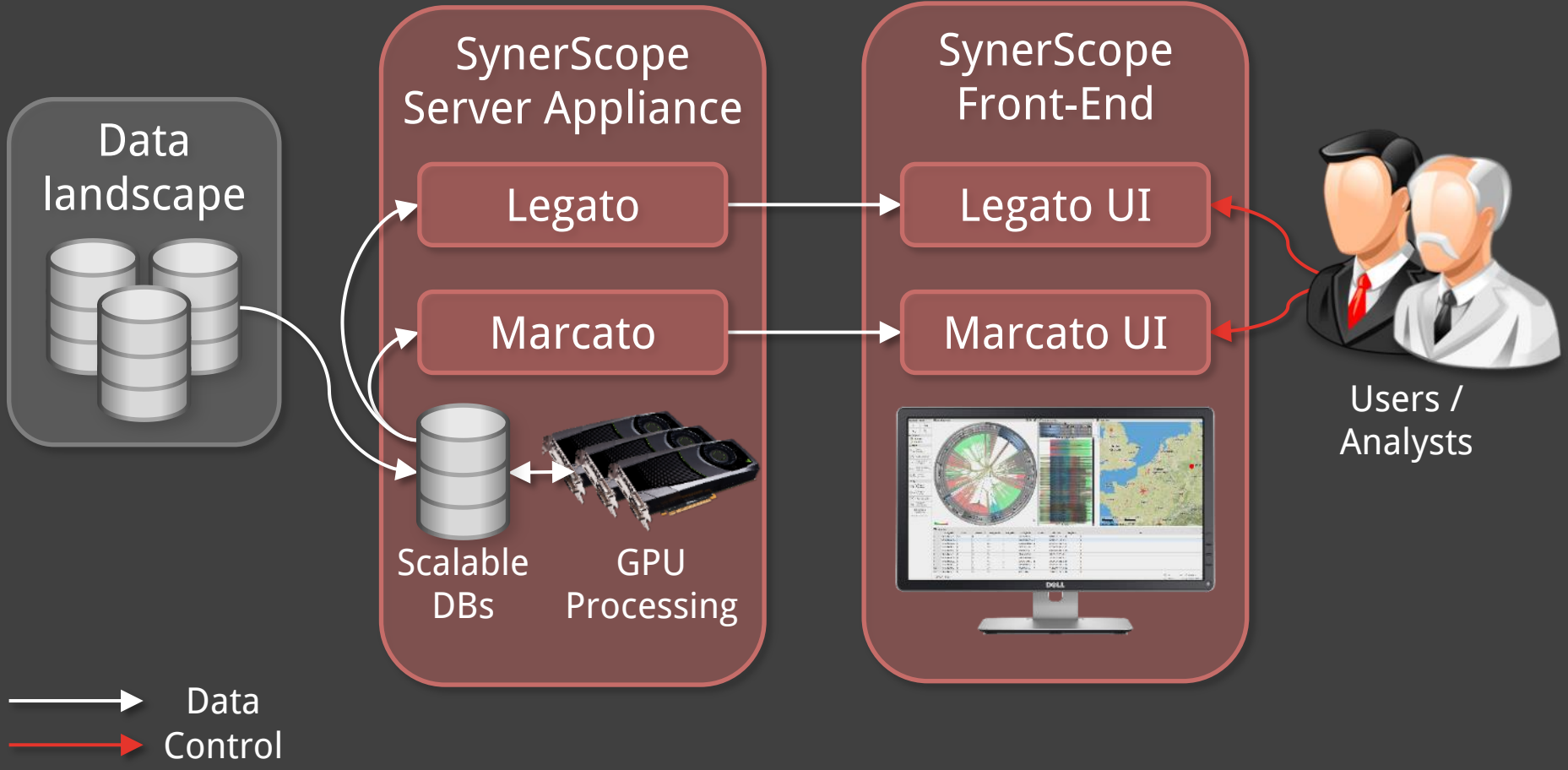




SynerScope Solution

- Connects to multiple DBs, suggests and augments data to analyze
  - Customizable analysis environment → complete solution
- Real-time interaction, highly scalable (millions of nodes/links)
  - On commodity hardware through use of GPUs/CUDA (NVIDIA GRID)
- Built-in visual analytics
  - Data augmentation, cleansing, sentiment analysis, ...
- On-premise hardware appliance
  - Secure server-based storage, analysis, and remote visualization (thin clients) for enterprise scalability

Pipeline


Data landscape
SynerScope Server Appliance
Legato
Marcato
Scalable DBs
GPU Processing
SynerScope Front-End
Legato UI
Marcato UI
Users / Analysts
Data
Control

Legato and Marcato

SynerScope **Legato**

"World map of data landscape"

- Overview of tables, columns, and values
- Data processed using workflow-engine
  - Shows value distributions and data quality/completeness
- Shows suitability for Marcato analysis based on data availability, quality, and completeness
- Data provenance tracking for sensitive data

Legato and Marcato

SynerScope **Marcato**

Visual analytics environment

- Domain experts and analysts explore and analyze their data
  - "Zero programming" if desired
- Multiple Coordinate View (MCV) framework spanning multiple screens
- Extensive collection of predefined visualizations and view-configurable UI
- Add new visualizations using HTML5 (Canvas/SVG/WebGL)

Insurance Claim Analysis

- Find networks of interconnected people and organizations/companies
- Detect deviating connectivity
  → possible fraud ring(s)
- Explore claim-handling process
  → discover trends, patterns, combinations, anomalies
- Quickly switch task perspectives with minimal IT burden
  → domain expert in control

But also...

- Financial transactions
- Communication analysis (SNA, e-mail, social media)
- Software dependencies
- Citation analysis
- Case-law networks
- E-commerce data (buy together)
- Drug interactions
- Any kind (?) of multivariate data

Demo