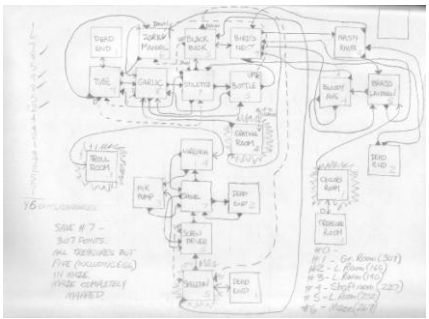
Aviz

informatics mathematics Inria Visual Analytics Project

Infrastructures for Visual Analytics: You are in a maze of twisty little passages, all alike!

Jean-Daniel Fekete
Jean-Daniel.Fekete@inria.fr
www.aviz.fr



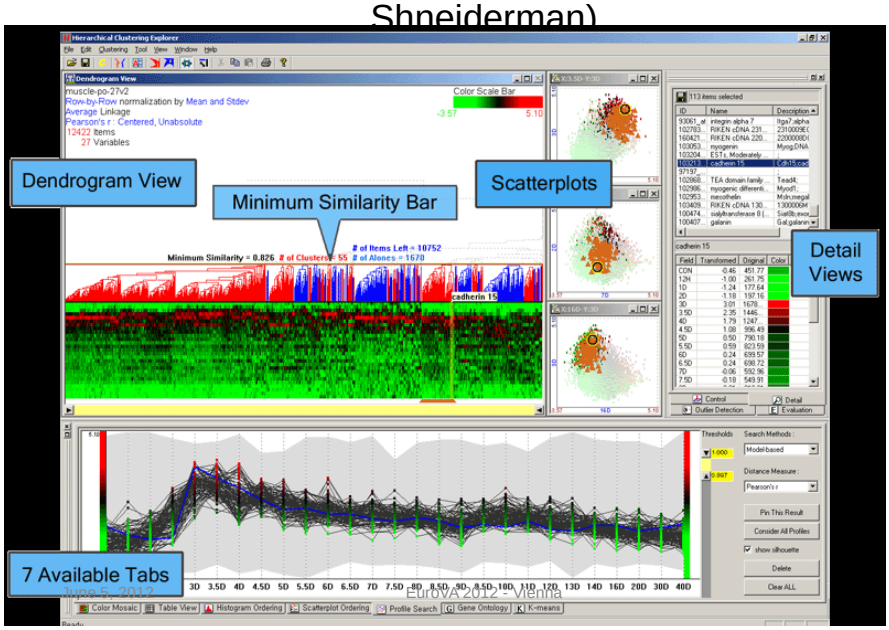
Issues

- VA combines data storage, indexing, analysis, exploration and dissemination through visualization
- When data is big and analyses are complex, interaction suffers from long computations and no guidance:
  - the human analytical process is hampered
- **So far, software infrastructures issues have been addressed in ad-hoc ways by each application**
- **This is not sustainable, even in the short term**

Current Solutions

- An InfoVis Team extends nice interactive visualizations with analytical capabilities and standard storage capabilities
  - Analysis algorithms are sub-optimal
  - Storage is ad-hoc
- A Machine-Learning Team extends nice learning algorithms with visualizations and storage
  - Visualization and interaction is simplistic
  - Storage is ad-hoc
- Same for Database Teams

Hierarchical Clustering Explorer (Seoh & Shneiderman)



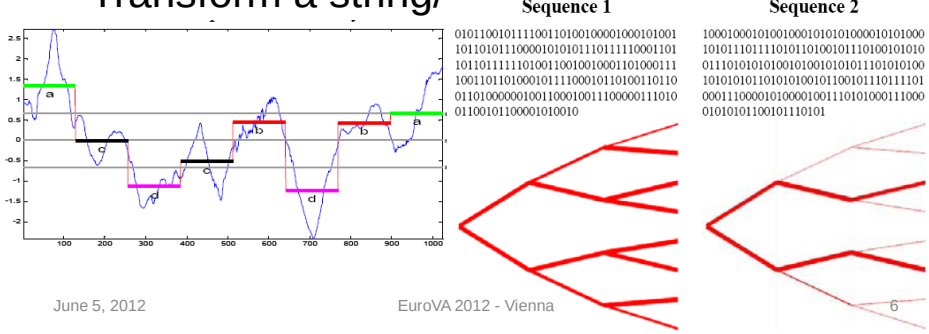
http://www.cs.umd.edu/hcil/hce/

Visualization do the Analysis

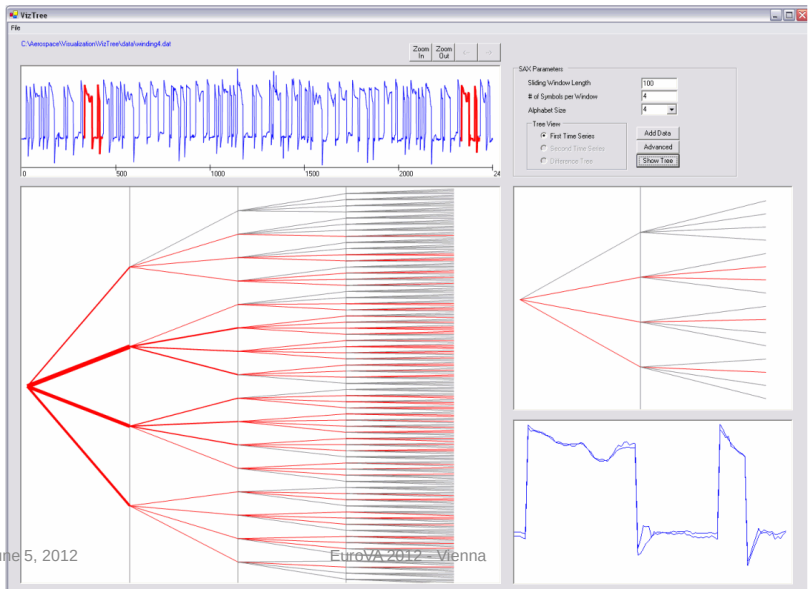
- Very nice visualization
- Sub-optimal algorithms or unreasonable amount of time spent
- Storage-agnostic
- What about:
  - Reuse?
  - Performance?
  - Scalability?

VizTree (Lin et al. 04)

- Lin, J., Keogh, E., Lonardi, S., Lankford, J. P. & Nystrom, D. M. (2004). Visually Mining and Monitoring Massive Time Series. In proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, WA. Aug 22-25.
- Transform a time-series into a string
- Transform a string/token into a tree branch



VizTree Visualization



Machine Learning do the visualization

- Very nice model
- Sub-optimal visualization or unreasonable amount of time spent
- Storage-agnostic
- What about:
  - Reuse?
  - Performance?
  - Scalability?

Improved Solutions: Pairwise Collab.

- Database + Machine Learning
  - Very impressive (e.g. Google with the Cloud)
- Text Analysis and Information Visualization
  - Very promising
- Still, ad-hoc
  - Scalable?
  - Reusable?
  - Interactive?

The Problem

- No Agreed Reference Model for Visual Analytics
- Visual Analytics applications need to re-implement the algorithms, visualizations and interactions
- Complex components cannot be easily re-implemented
- Visualizations and interactions are poorly re-implemented
- Industry cannot sell components

Needs

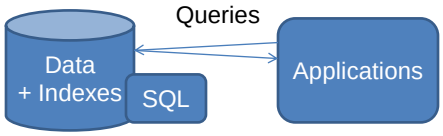
- What are the needed capabilities?
  - Flexible storage / indexing
  - Asynchronous computation
  - Continuous notification of partial results
  - Steering of algorithms to work on “interesting” areas
  - Composition of hybrid algorithms
  - Assessments of the quality of analysis results (Rank algorithms by Feature?)
- How to assemble components?
  - Modularity
  - Separation of concern
  - Abstracting the wealth of hardware configurations

Domains Involved

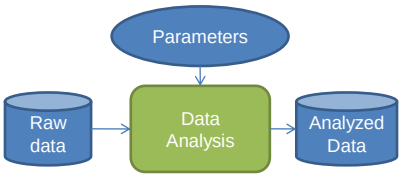
- Data Management / Databases
- Analysis
  - Statistics
  - Machine-Learning
  - Text Mining
  - Image Analysis
  - Video Analysis
  - Graph Mining (e.g. social network analysis)
- Visualization
  - Infovis
  - SciVis
  - GeoVis

Software Reference Models

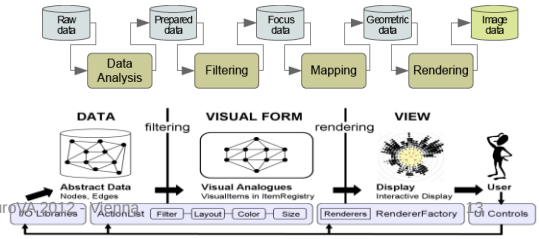
- Databases
  Data Management



- *Analysis*

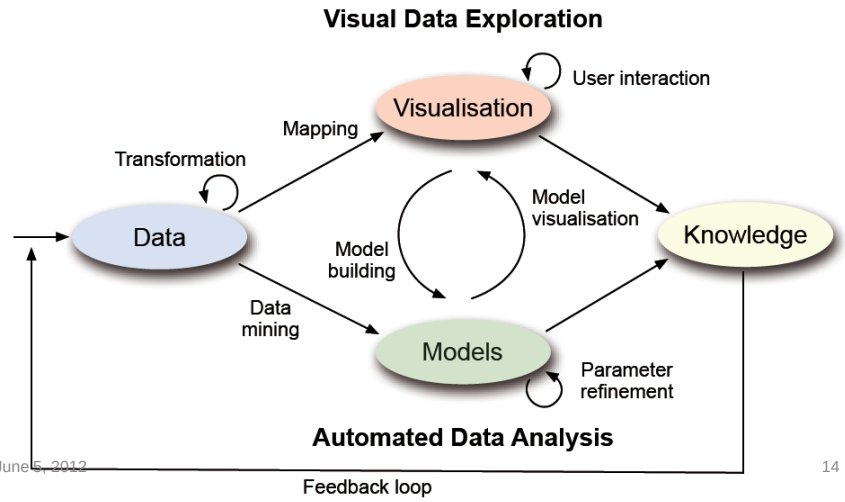


- Visualization

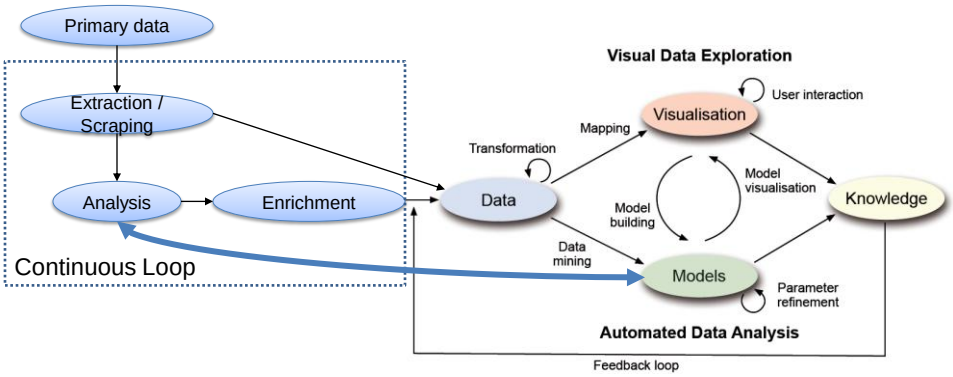


The Visual Analytics Process

- D. A. Keim, J. Kohlhammer, G. Ellis and F. Mansmann. Mastering The Information Age - Solving Problems with Visual Analytics. Eurographics, 2010.



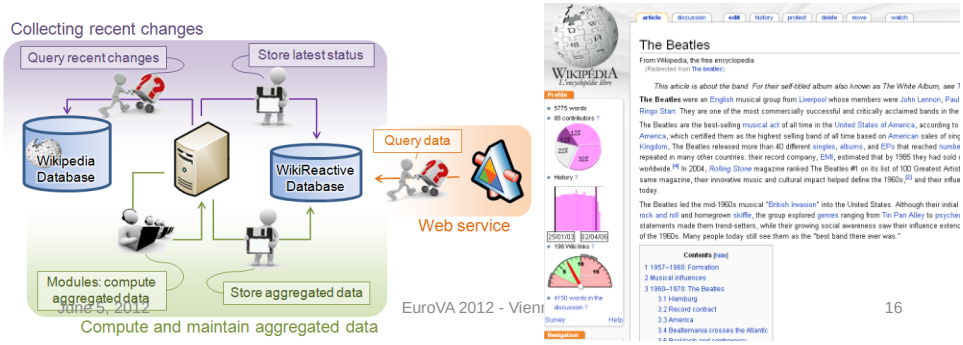
The Visual Analytics Process Extended



WikiReactive

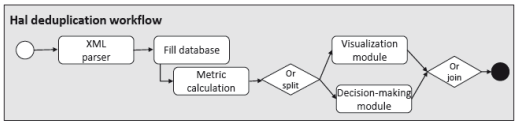
N. Boukhelifa, F. Chevalier and J.D. Fekete Real-time Aggregation of Wikipedia Data for Visual Analytics. In Proceedings of Visual Analytics Science and Technology. VAST '10. 147-154. 2010

- Collect wikipedia changes and computes derived information
  - Diffs, user contributions, user per character

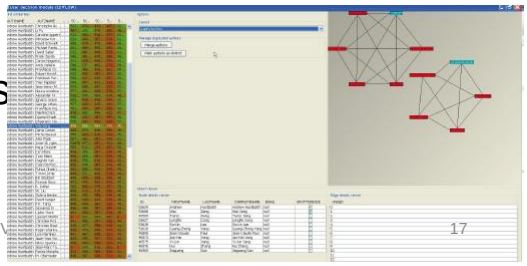


HAL Deduplication framework

- For each article author added to the HAL database

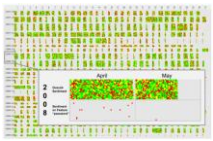


- Computes similarity with all other autors
- Resolve simple case (< or > threshold)
- Show an interface for the other cases

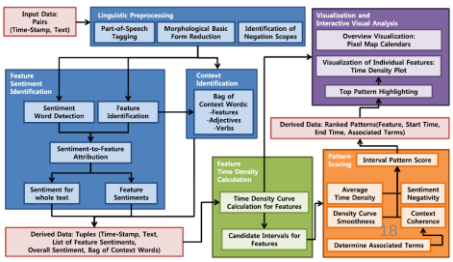


Real-Time Sentiment Analysis

- Christian Rohrdantz, Ming C. Hao, Umeshwar Dayal, Lars-Erik Haug, and Daniel A. Keim. 2012. Feature-Based Visual Sentiment Analysis of Text Document Streams. *ACM Trans. Intell. Syst. Technol.* 3, 2, Article 26 (February 2012), 25 pages.



- For each new document scrapped
- Compute part-of-speech tagging, lemmatization, negation detection, feature extraction, sentiment detection, sentiment-to-feature mapping



Problem: Bounding Time and Quality

- Visualization is User Centric
  - Visualization will only show a small amount of data
  - Visualization need interactive time
  - How can we address the scale in interactive time?
- Analysis is Program Centric
  - Analysis will read data, process it and store its results in the end
  - Analysis will produce unbounded amounts of data in unbounded time
  - How can we get something in a bounded time?
- Databases is Data Centric
  - Databases will store and retrieve unbounded amounts of data in unbounded (but fast) time
  - How can we bound time with a specified level of quality?

Vision

- In the future, Visual Analytics will rely on **components** or **modules**
- The components will interoperate based on a reference model
  - Abstractly defined but implemented by several providers
- Need to avoid
  - “One system does all” (e.g. VTK)
  - Many fragmented/incompatible systems
- Need to go step by step
  - We need a Research Programme

Extending Reference Models

- The Visualization Reference Models
- The Data Management Reference Model
- The Data Analysis Reference Model

- Connecting Them Together

The InfoVis Reference Model Extended

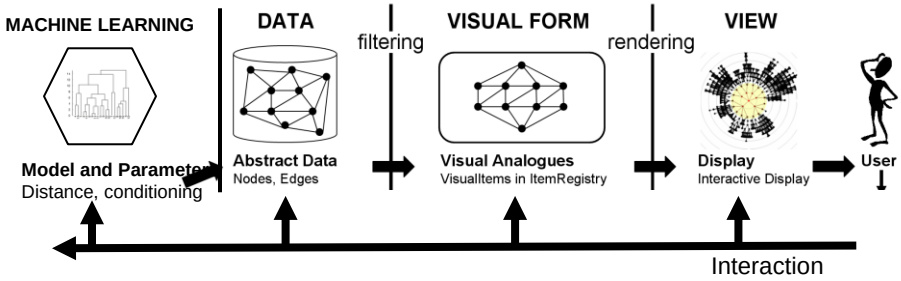
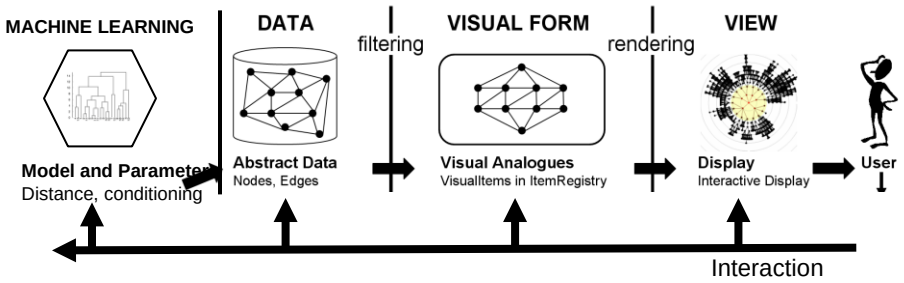


Illustration by J. Heer

Ed H. Chi, John T. Riedl, "An Operator Interaction Framework for Visualization Systems," p. 63, *IEEE InfoVis '98*, 1998.

Visualization and Visual Analytics



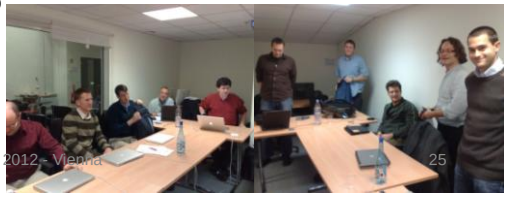
- Extend the pipeline to the left
  - Analytical components need to be integrated and controlled interactively
  - The results of analysis should be stored in the data repository iteratively/progressively
  - Data storage cannot remain “in-memory”

Can Visualization be Componentized?

- Yes
- Done in VTK
- Done in each of the InfoVis Toolkits
- Now, done on top of the Java InfoVis Toolkits
  - The Obvious Abstract Toolkit

Obvious History

- VisMaster WP4 organized a workshop in Paris, Dec 4-6 2008
- Invitation only
  - Already had 3 open workshops on Information Visualization Infrastructures
  - Wanted a “hand on” approach instead of sharing knowledge
  - 2 busy days with Academics and Industrials
- Outcome
  - Practical specifications (code.google.com/p/obvious)
  - High level discussions
  - Commitments to test and conform to it as much as reasonable
  - INRIA just hired an engineer for 2 years to develop and maintain the work: Pierre-Luc Hémery
- 12 participants
  - Baudel, Thomas (ILOG/IBM)
  - Favart, Christophe (BO/SAP)
  - Fekete, Jean-Daniel (INRIA)
  - Fisher, Danyel (Microsoft Research)
  - Heer, Jeffrey (Stanford Univ.)
  - O'Madadhain, Joshua (Google)
  - Piringer, Harald (VRVis)
  - Santucci, Giuseppe (Univ. Roma)
  - Smoot, Mike (UCSD)
  - Theus, Martin (Augsburg Univ.)
  - Weaver, Chris (Univ. of Oklahoma)
  - Wood, Jo (City Univ. London)



Lord of the Toolkits
One Toolkit to Bind Them All!

- Pierre-Luc Hémery hired by INRIA for 2 years to implement it
- Encapsulates the well-understood InfoVis Reference Model for Java Toolkits
- Currently encapsulates:
  - The InfoVis Toolkit (Fekete 04)
  - Prefuse (Heer 05)
  - Improvise (Weaver 05)
  - JDBC as Data Model

http://code.google.com/p/obvious

Vis/InfoVis/GeoVis Unification?

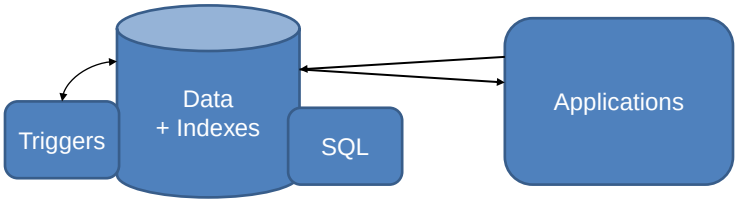
- There is no reason why the pipelines cannot be merged at various levels
  - Data, compositing, view, with brushing&linking
- More research is needed beyond juxtaposition of components
  - Embedding
  - Hybrids
  - Merging?

Missing Parts?

- Scalability
  - Unbounded data can arrive with VA
  - How to avoid flooding the user and the system
  - **Aggregation** becomes mandatory, coupled with a "Budget" model?
    - N. Elmqvist, J.-D. Fekete. Hierarchical Aggregation for Information Visualization: Overview, Techniques and Design Guidelines. *IEEE Transactions on Visualization and Computer Graphics*, 16(3):439-454, 2010.
- Asynchronous Updates of Visualization
  - Data will arrive at any time due to dynamic computation or data collection
  - Analytical queries will take time to complete
  - D. Fisher, I. Popov, S. M. Drucker, and mc schraefel, Trust Me, I'm Partially Right: Incremental Visualization Lets Analysts Explore Large Datasets Faster, in Proceedings of the 2012 Conference on Human Factors in Computing Systems (CHI 2012), ACM Conference on Human Factors in Computing Systems, 5 May 2012

Extending Reference Models

- The Visualization Reference Models
- The Data Management Reference Model



- The Data Analysis Reference Model
- Connecting Them Together

Data Management and Visual Analytics

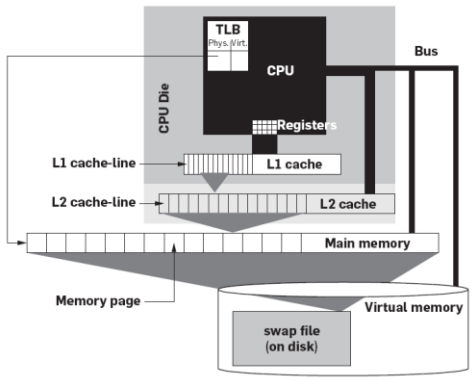
- Several layers of storage semantics
  - Flat files, XML, HFS, SQL Databases, NoSQL, Storage on the Cloud
- Services
  - ACID (Atomicity, Consistency, Isolation, Durability)
  - Persistence
  - Indexing
  - Distribution
  - Typing
  - Notification
  - Interactive Performance
  - Computation

Data Management Services for VA

- Persistence
  - Required
- ACID
  - Required but Atomicity needs extensions
- Indexing
  - Required by Visualization
- Distribution
  - Useful for Data Management, Analysis and Visualization
- Typing
  - Required
- Notification
  - Required by Visualization
- Interactive Performance
  - Required by Visualization
- Computation
  - Required by Analysis and Visualization

Data Management for VA

- Peter A. Boncz, Martin L. Kersten, and Stefan Manegold. 2008. **Breaking the memory wall in MonetDB.** *Commun. ACM* 51, 12 (December 2008), 77-85.
- Reimplementing in-memory databases for Visualization is a waste of time and effort
- In a distributed system, the Database should be seen as the shared memory
- The main memory becomes a cache of the database
- Database people should do it, not Vis people



Experiment: DBMS Caching with Obvious

- One binding of the Obvious data model is written with JDBC:
  - Allows to read tuples on demand from a DBMS table and store them in memory while they are used
  - Keeps a bidirectional link between memory and the DBMS
- What happened when the DBMS table changes?
  - A DB trigger is called
  - The Obvious table is notified that something changed
  - Changed data is read again (eagerly or lazily)
- Tested with Oracle and MySQL
  - Access time for Prefuse and IVTK are about 1ms for 100,000 items, 100 times faster than Oracle or MySQL

Database Issues

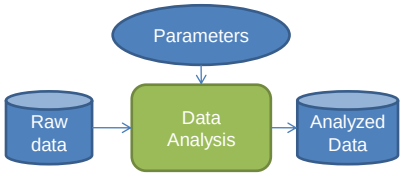
- Analysis frequently add attributes
  - Column oriented vs. Row oriented
- Transactions?
  - Yes
  - But extended (snapshot isolation, long transactions)
- Extended typing
  - Should be able to express the semantics of attributes beyond their representation type
- SQL?
  - Implementation issue but why not for queries
- Notification management
  - Should improve on the standard Trigger mechanism
- Indexing and Aggregation
  - More flexibility is required. Geospatial extensions have been specified, we need other extensions
- Fast bounded interruptible query management
  - Sidirourgos, L. - Kersten, M.L. - Boncz, P.A. SciBORQ: Scientific data management with Bounds On Runtime and Quality 2011 - Proceedings of the biennial Conference on Innovative Data Systems Research 2011
  - The Researcher's Guide to the Data Deluge: Querying a Scientific Database in Just a Few Seconds Proceedings of International Conference on Very Large Data Bases 2011 (VLDB) , p.585–597

What about Cloud and Big Tables?

- Visualizing data in the cloud
  - http://googleblog.blogspot.fr/2008/11/visualizing-data-in-cloud.html
  - Scalability is limited!
- The Cloud is bad for interaction
  - High throughput/high latency
  - Perfect for the continuous loop or large model computation
- More work is needed to steer the computations in the Cloud

Extending Reference Models

- The Visualization Reference Models
- The Data Management Reference Model
- The Data Analysis Reference Model



- Connecting Them Together

Analysis Infrastructures

- Lots of high-quality Analytical components available
- New standards to perform Machine-Learning as a service (DMX or PMML, Google Prediction API)
- However, their reference model (sic) is VERY POOR
- How can we improve it?

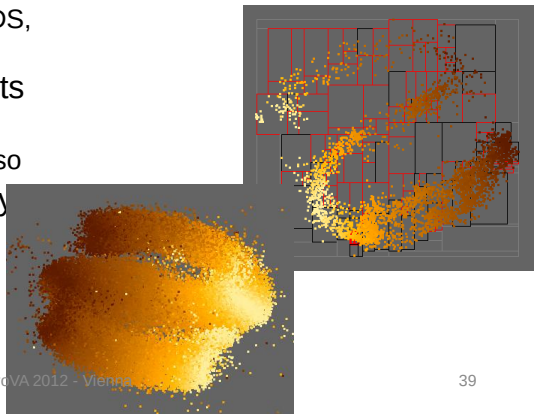
Analytical Strategies

- Pre-computation and storage
  - Ad-hoc methods (run algorithms for a long time)
  - Cloud computing (BigTable + MapReduce)
- Iterative (Steerable) Algorithms
- Multi-resolution progressive algorithms
- Hybrid algorithms
- Incremental update strategies

Analytical Strategies: Iterative (Steerable)

- Lots of algorithms are implemented by iterative refinements
  - Image blurring, Force-based Graph Layout, MDS, TSP, PCA
- Let them pass the results of iteration steps
  - Maybe every second or so
- Some can be steered by the user's viewpoint
  - Let them be dynamically steered

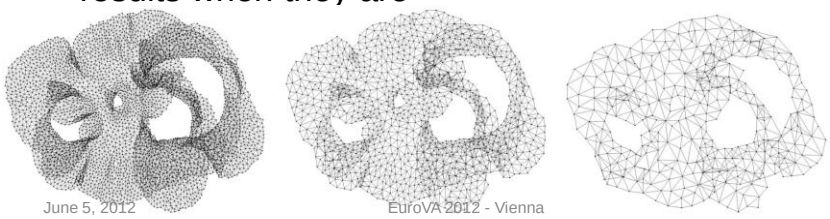
Matt Williams and Tamara Munzner. 2004. Steerable, Progressive Multidimensional Scaling. In *Proceedings of the IEEE Symposium on Information Visualization* (INFOVIS '04). IEEE Computer Society, Washington, DC, USA, 57-64.



Analytical Strategies: Multiresolution

- Some algorithms can start with low resolution and increase it dynamically
- Graph Drawing, Image Transforms, etc.
- Let them pass the results when they are
- Allow them to be steered

Emden R. Gansner, Yehuda Koren, Stephen C. North, "Topological Fisheye Views for Visualizing Large Graphs," IEEE Transactions on Visualization and Computer Graphics, pp. 457-468, July/August, 2005



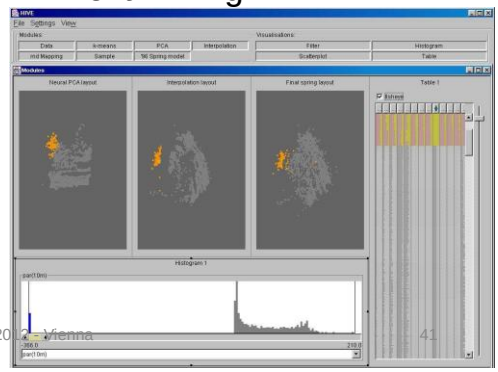
4394-node approximation 1223-node approximation 341-node approximation

Analytical Strategies: Hybrid Algo.

- Clustering a huge dataset?
- HC is quadratic: not possible
- K-Means is linear but requires a good K
- Sample -> HC -> Estimate good K -> k-Means
- Need a good sampling

Ross, G. and Chalmers, M. (2003) A visual workspace for constructing hybrid MDS algorithms and coordinating multiple views. Information Visualization, 2 (4). pp. 247-257.

Does not work well for Text mining

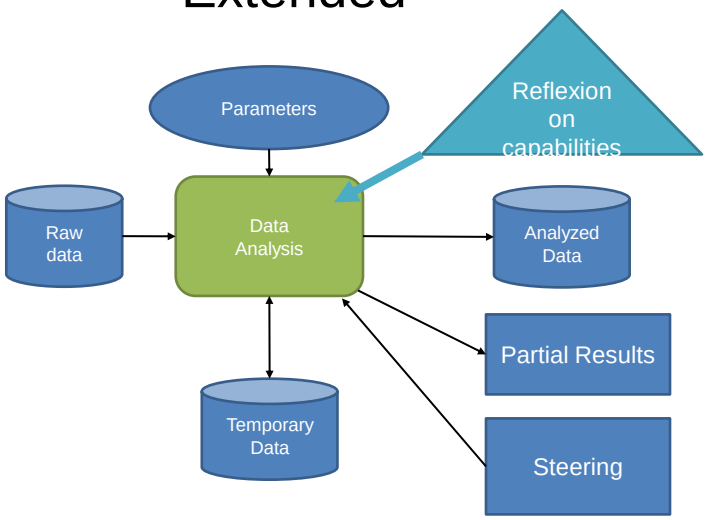


Analytical Strategies: Incremental update

- HC is made in two steps:
  1. Compute (di)similarity matrix
  2. Create clusters
- Step 1 is quadratic
- When items are added or deleted, updating the matrix is linear
- Keep the matrix!

- Same for several algorithms: store temporary computations that are expensive and updatable

Analysis Reference Model Extended



Additional Problem

- Multiple existing analysis environments
  - R, Matlab, Excel, SPSS, SAS, etc.
- People are comfortable in their environment
- Lots of code already exists, sometimes substantial in size and complexity
- If we use them and pass the results between environments, the time is bounded by data transmission
- What should we do?
  - Integrate all the environments? (impractical)
  - Create a new one that will solve everything?
  - Find a way to lower the data transmission time (Data Management Issue)

Analysis: Summary

- Components should be restructured for interaction
- Who will do it?
- Hybrid algorithms can reuse existing components as they are but not the others
- Components need to expose their capabilities to the pipeline
- Expressing the interactive capabilities of components is a research issue
- Multiple environments will exists, how can we lower substantially the data transmission cost?

Extending Reference Models

- The Visualization Reference Models
- The Data Management Reference Model
- The Data Analysis Reference Model

- Connecting Them Together

Building VA Systems

- Coping with the diverse hardware and software solutions
  - Connecting parts from the huge and growing diversity
- We cannot rely on one software solution
  - We need to abstract the solution into a reference model and rely on it
- It can be done
  - VisTrails and Ediflow: workflow systems to connect and run VA dynamically

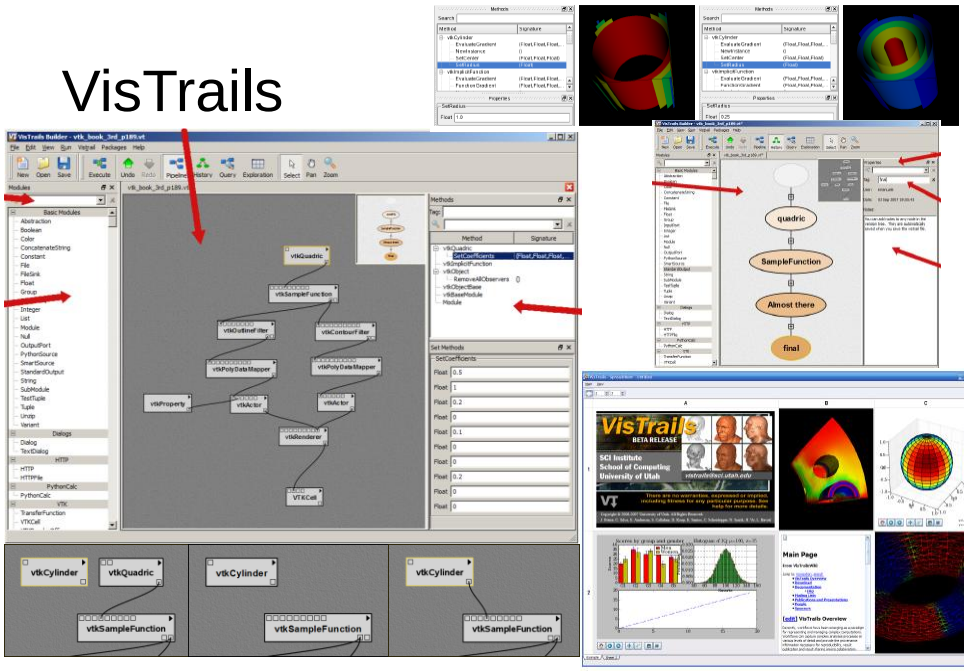
Scientific Workflow Systems

- Combining data management + computation + visualization
- Lots of ad-hoc Scientific Workflow Systems (e.g. Kepler)
- With (Sci) Visualization: VisTrails!
- Impressive system
  - Exploration + data provenance

Carlos E. Scheidegger, Huy T. Vo, David Koop, Juliana Freire, and Claudio T. Silva. 2008. Querying and re-using workflows with VsTrails. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data* (SIGMOD '08). ACM, New York, NY, USA, 1251-1254.

www.vistrails.org

VisTrails



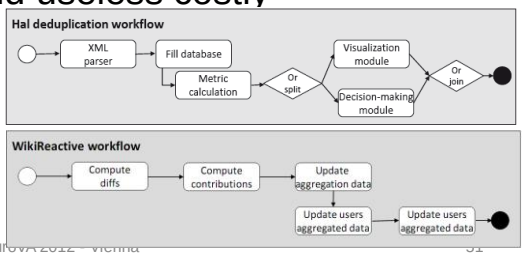
Workflow Systems

- Once the pipeline is componentized, it can be manipulated in a workflow system
- Currently, VisTrails relies on VTK
  - Work underway to work with Java/Jython and Obvious
- More work is needed
  - To add continuous manipulation to VisTrails
  - To hide the complexity to simple users
- Composing complex and powerful applications or prototypes should be made easier!
- Opportunities to separate the work specification from its implementation
  - Run locally, on a Cloud, on an HPC, etc.

Workflow for the Continuous Loop

- V. Benzaken, J.-D. Fekete, P.-L. Hémery, W. Khemiri, I. Manolescu. EdiFlow: data-intensive interactive workflows for visual analytics. International conference on Data Engineering, Apr 2011, Hannover, Germany.

- Specify the workflow, EdiFlow maintains data consistency by running the required modules when the data changes
  - Strategies to avoid useless costly recomputations

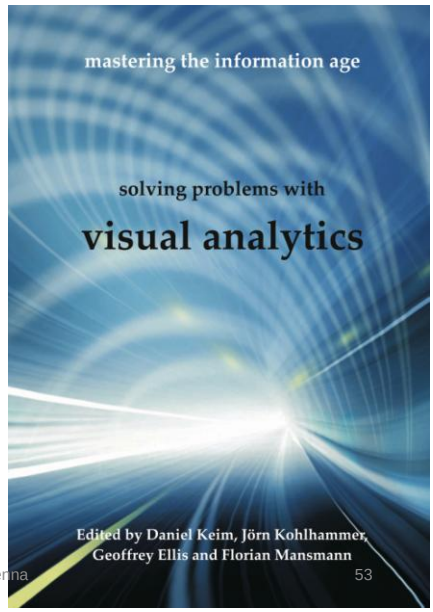


Summary for Infrastructures

- Visual Analytics Architectures are immature
  - They stretch the existing architectures far beyond their initial goals
  - They require complex functionalities and algorithms to be re-implemented over and over again
- We need to involve the specialists of the respective fields to solve the problems
  - Database researchers and practitioners are interested
  - Data Analysis researchers and p. are interested
  - Visualization researchers should meet too!
  - Workflows allow to connect components in a declarative way while maintaining analytic provenance
- Huge benefits in term of Research and Markets

Read the BOOK!

www.vismaster.eu



Contributions

- VisMaster collaboration:
  - Thomas Baudel (IBM/ILOG)
  - Joe Parry (i2)
  - Harald Piringer (VRVis)
- Dagstuhl Seminar on Information Visualization, Visual Data Mining and Machine Learning
- Dagstuhl seminar on Scalable Visual Analytics