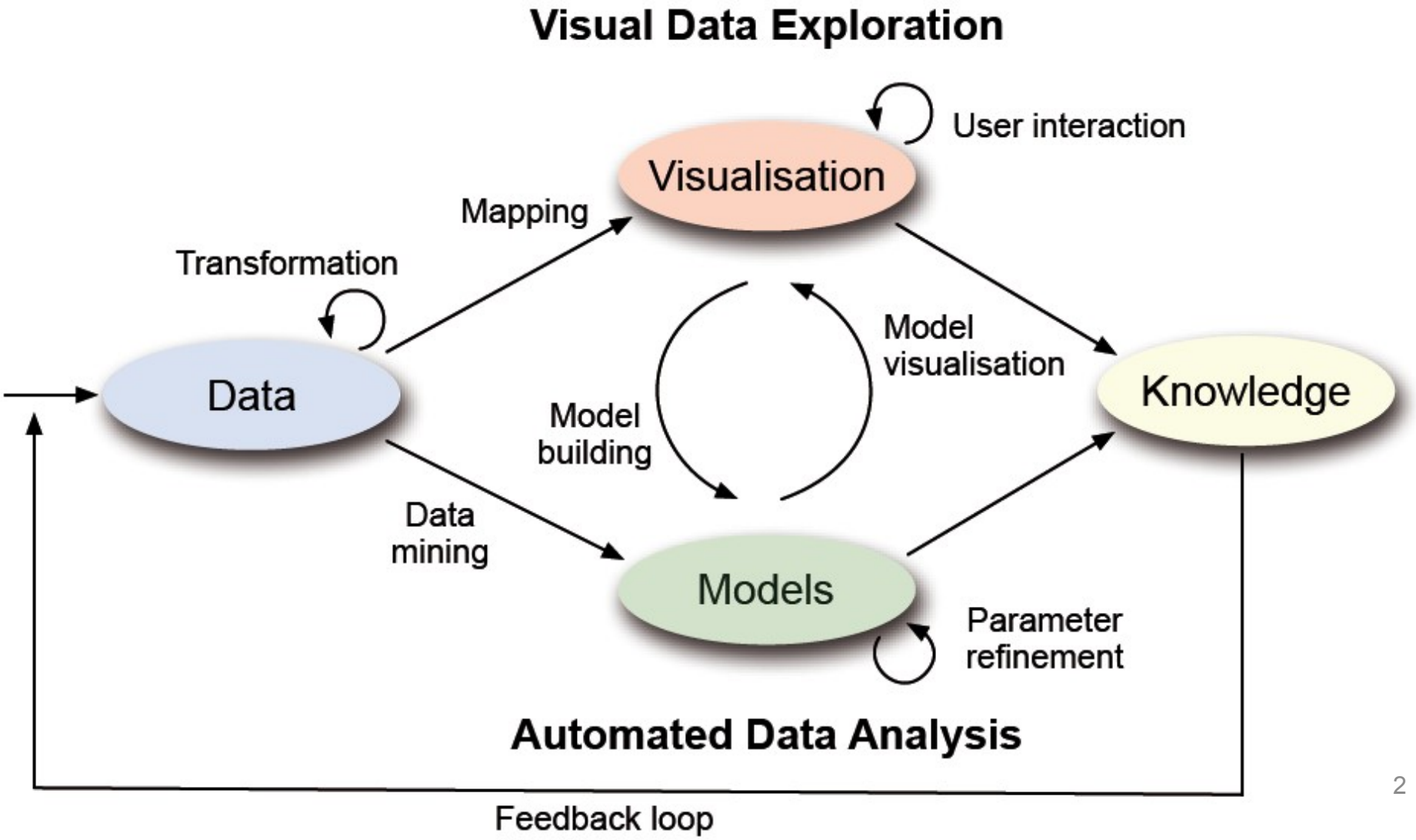
Visual Analysis Tools

Jean-Daniel Fekete
AVIZ/INRIA
Jean-Daniel.Fekete@inria.fr
http://www.aviz.fr/~fekete

The Visual Analytics Process

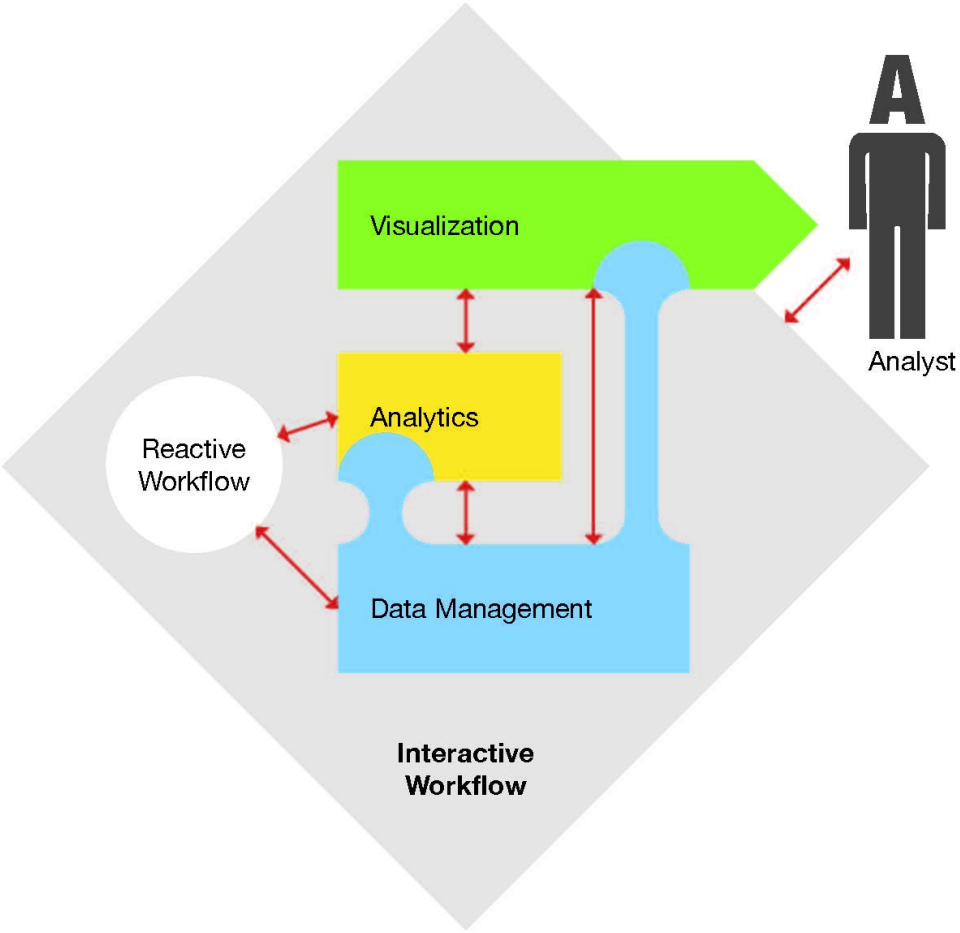
- D. A. Keim, J. Kohlhammer, G. Ellis and F. Mansmann. Mastering The Information Age - Solving Problems with Visual Analytics. Eurographics, 2010.



Analysis Tools for Visual Analytics

Three Layers:

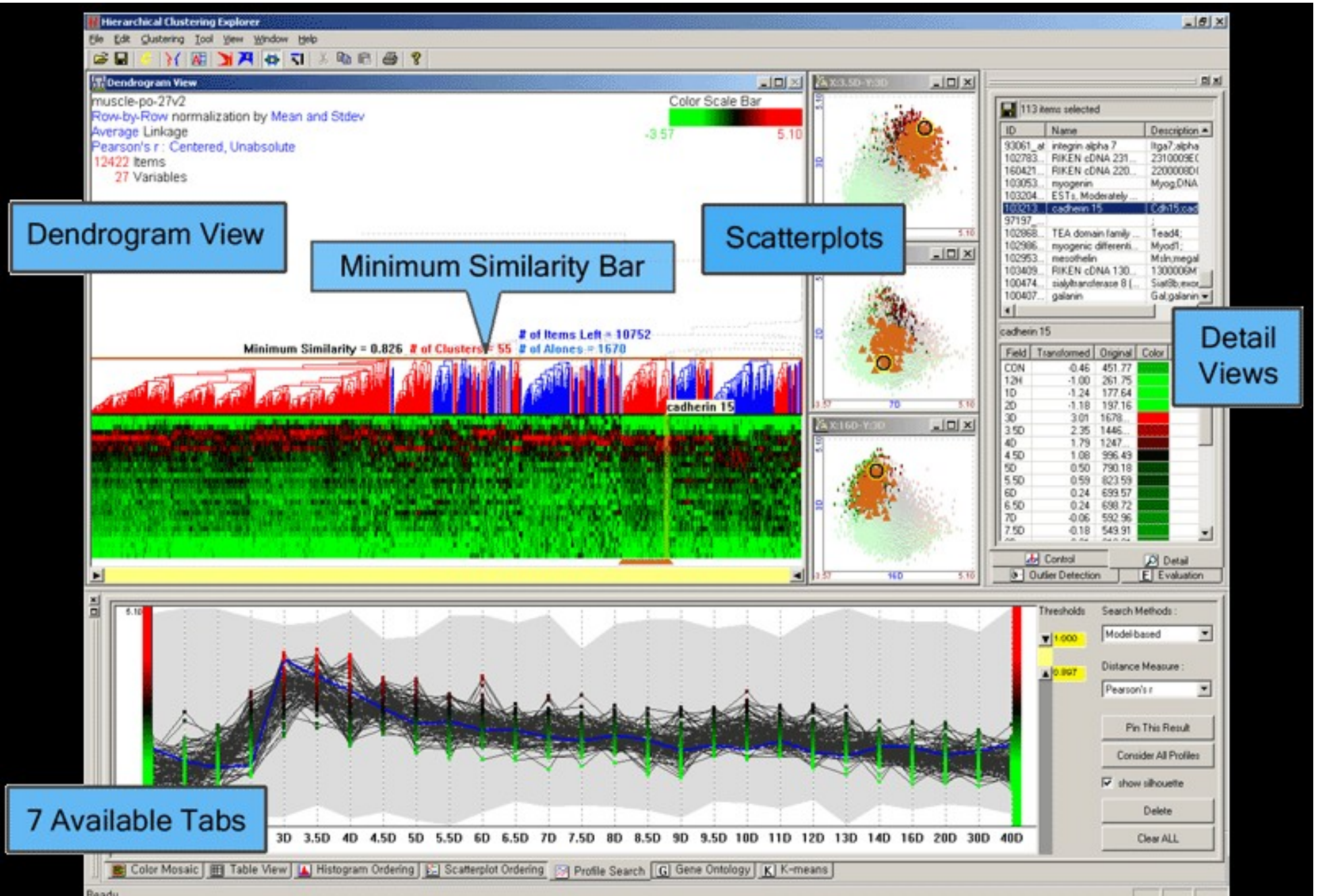
- Data Management
- Analytics
- Visualization



Examples

- Hierarchical Clustering Explorer
- WikiReactive
- HAL Deduplication Framework

Hierarchical Clustering Explorer (Seoh & Shneiderman)

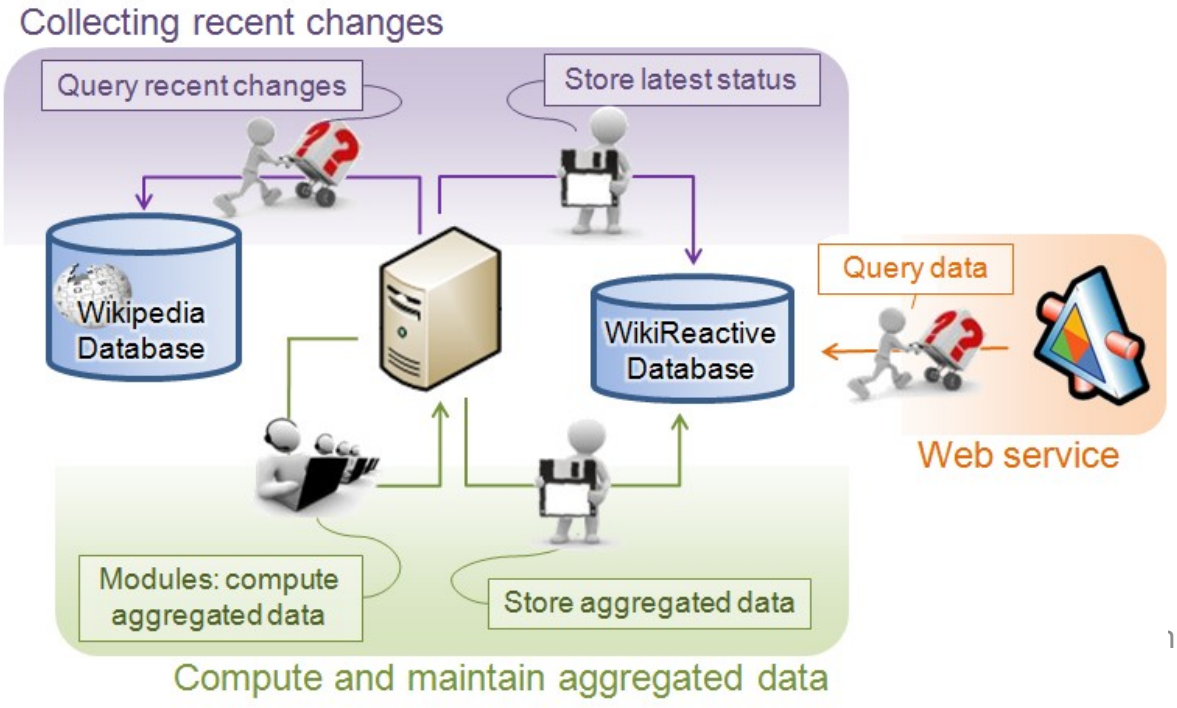


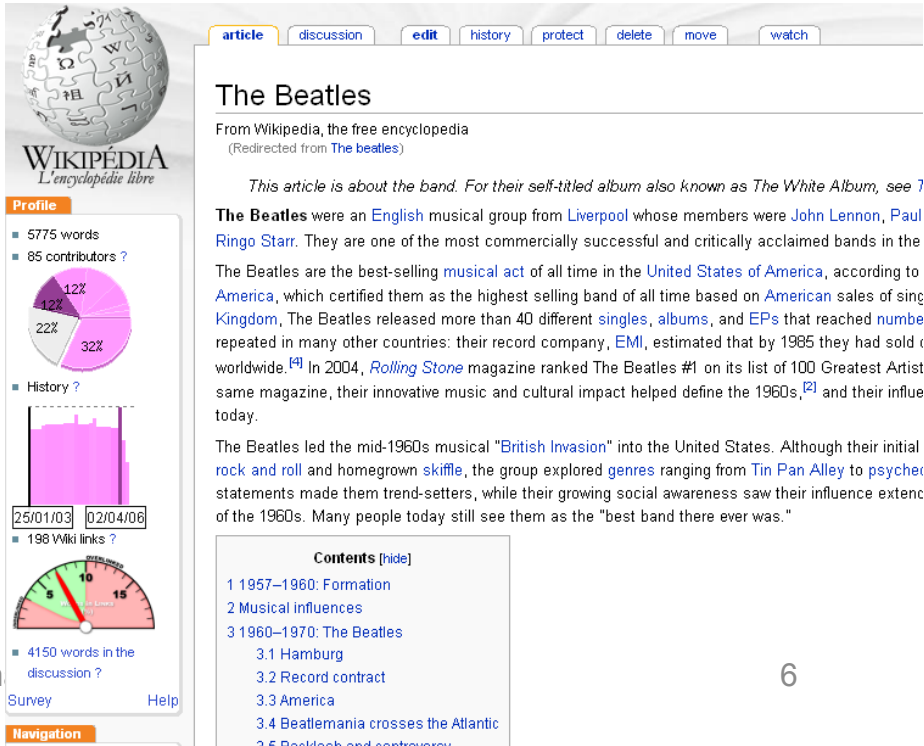
http://www.cs.umd.edu/hcil/hce/

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

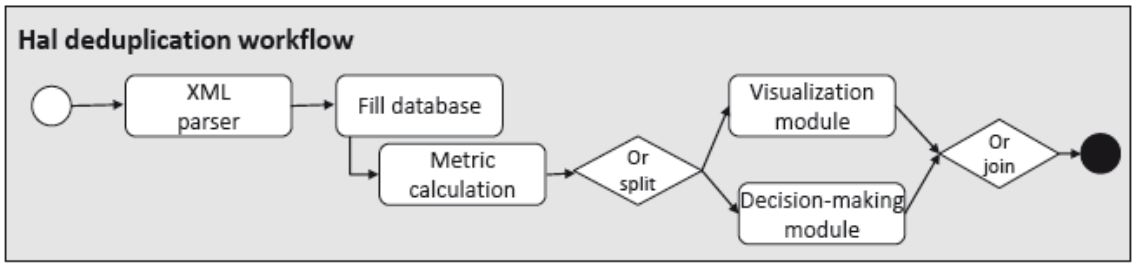
- Collect wikipedia changes and computes derived information
  - Diffs, user contributions, user per character



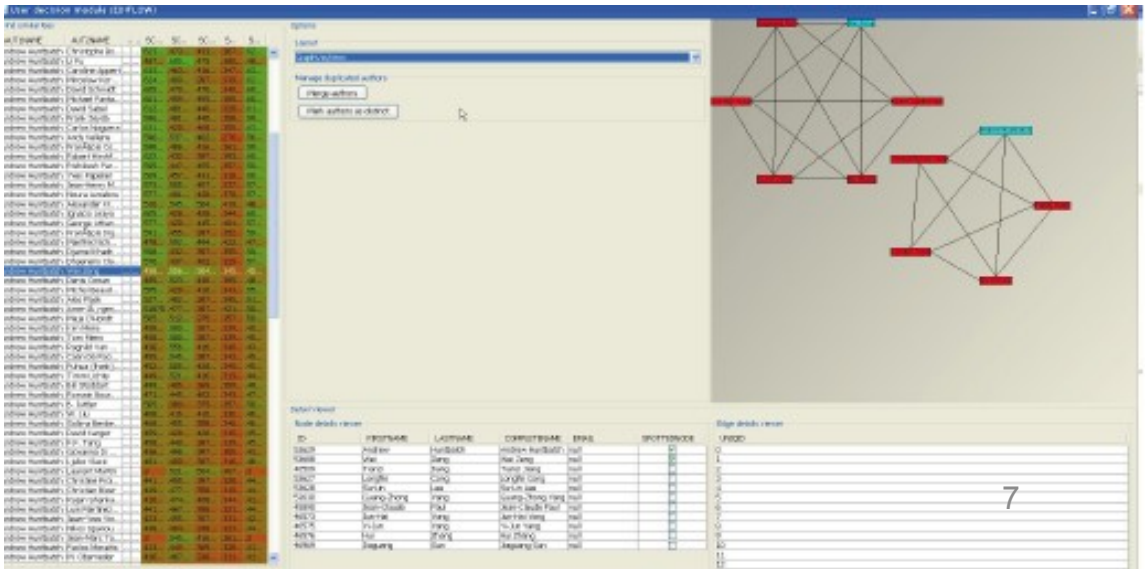


HAL Deduplication framework

- For each article author added to the HAL database



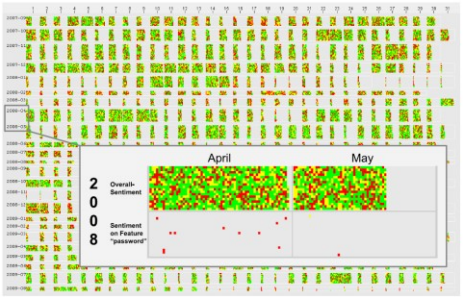
- Computes similarity with all other authors
- Resolve simple case (< or > threshold)
- Show an interface for the other cases

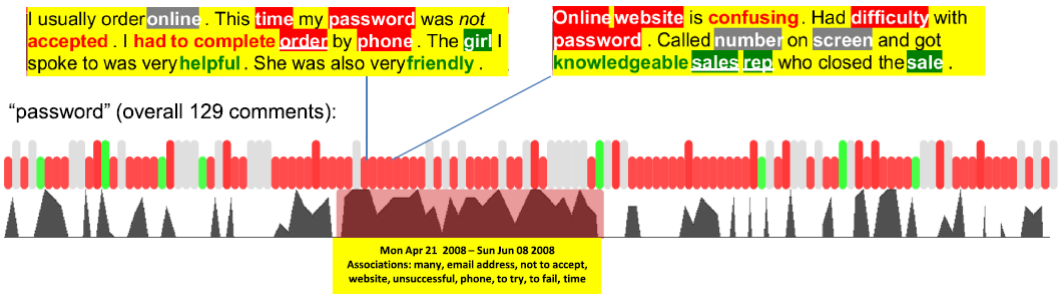


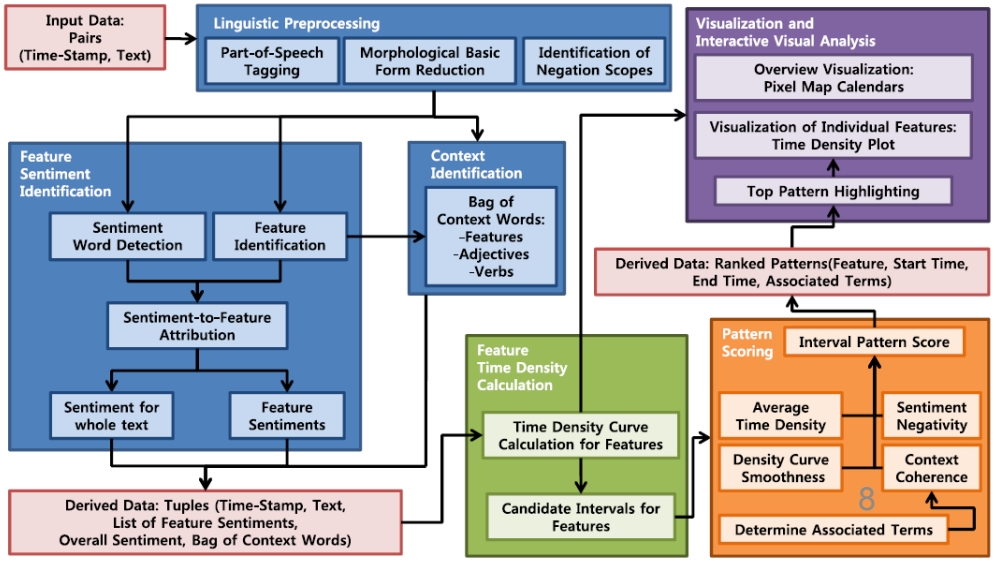
Real-Time Sentiment Analysis

- 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







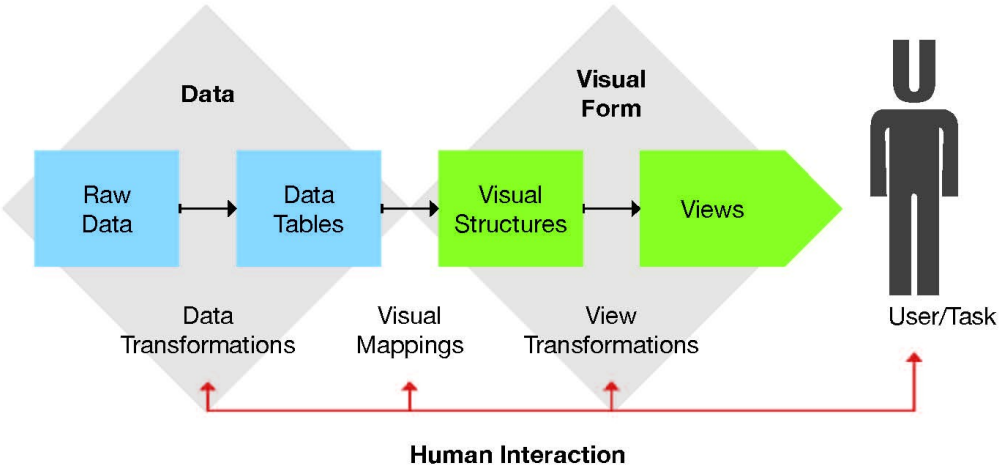
Problem: Bounding Time and Quality

- 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?

Visualization Layer

Reference model
followed by most tools

- Prefuse (Java toolkit)
- Tulip (C++ toolkit)
- D3 (JavaScript toolkit)
- Tableau (Application)
- Spotfire (Application)



Requirements of Visualization

- Number of pixels
- Display Speed
- Human Perception/Cognition

  Main constraints:
  - Animation (<100ms)
  - Interaction (<1s)

    Otherwise,
  - progressive updates (progress bar)

Overcoming the Limitations

- Number of pixels
- Display Speed
- Human Cognition
  - Won't change
  - Cannot wait 3 days for a computed estimate

Analytics

- Analytics is the discovery and communication of meaningful patterns in data. Especially valuable in areas rich with recorded information, analytics relies on the simultaneous application of statistics, computer programming and operations research to quantify performance. Analytics often favors data visualization to communicate insight.

  [Wikipedia]

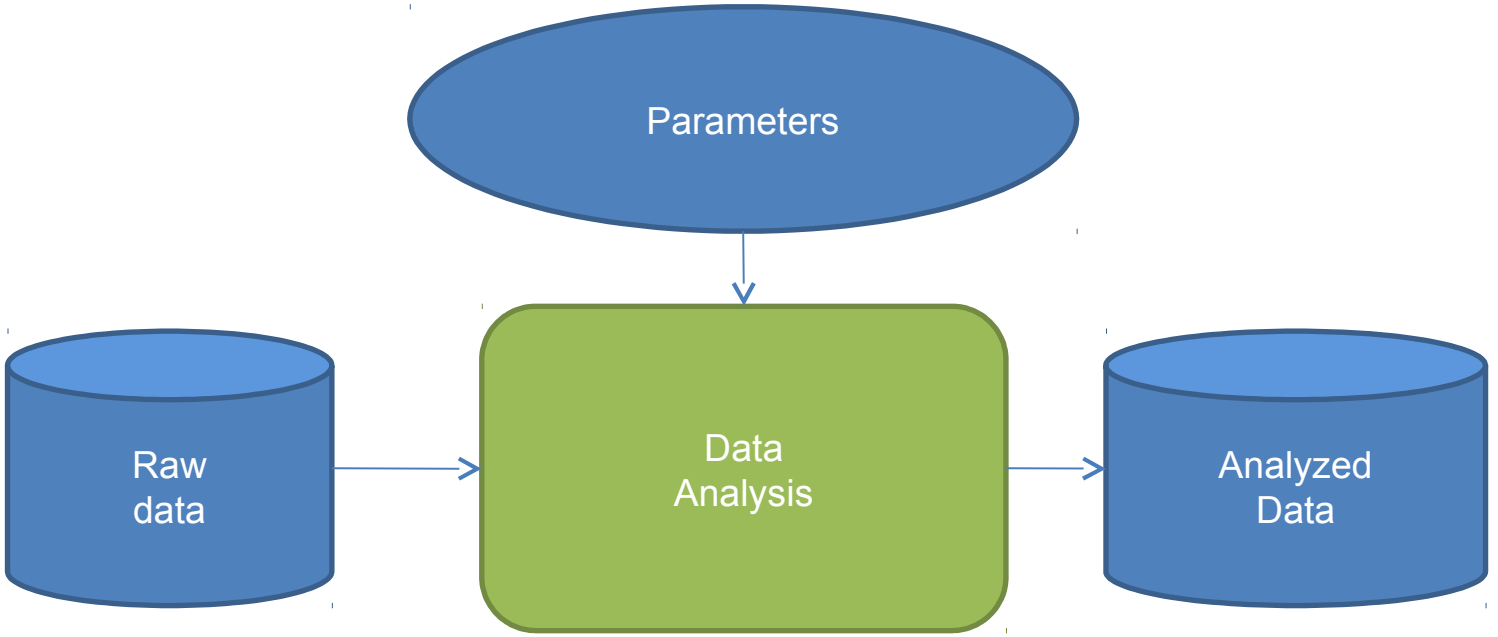
Analytics

- 87 systems listed in Wikipedia
  - http://en.wikipedia.org/wiki/List_of_numerical_analysis_software
- Numerical software packages
  - R, Matlab, Python
- Language Oriented Toolkits
  - LAPACK, LINPACK (Fortran, C),
- Libraries
  - Math, NLP, Video, etc.
-

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

Data Analysis Reference Model



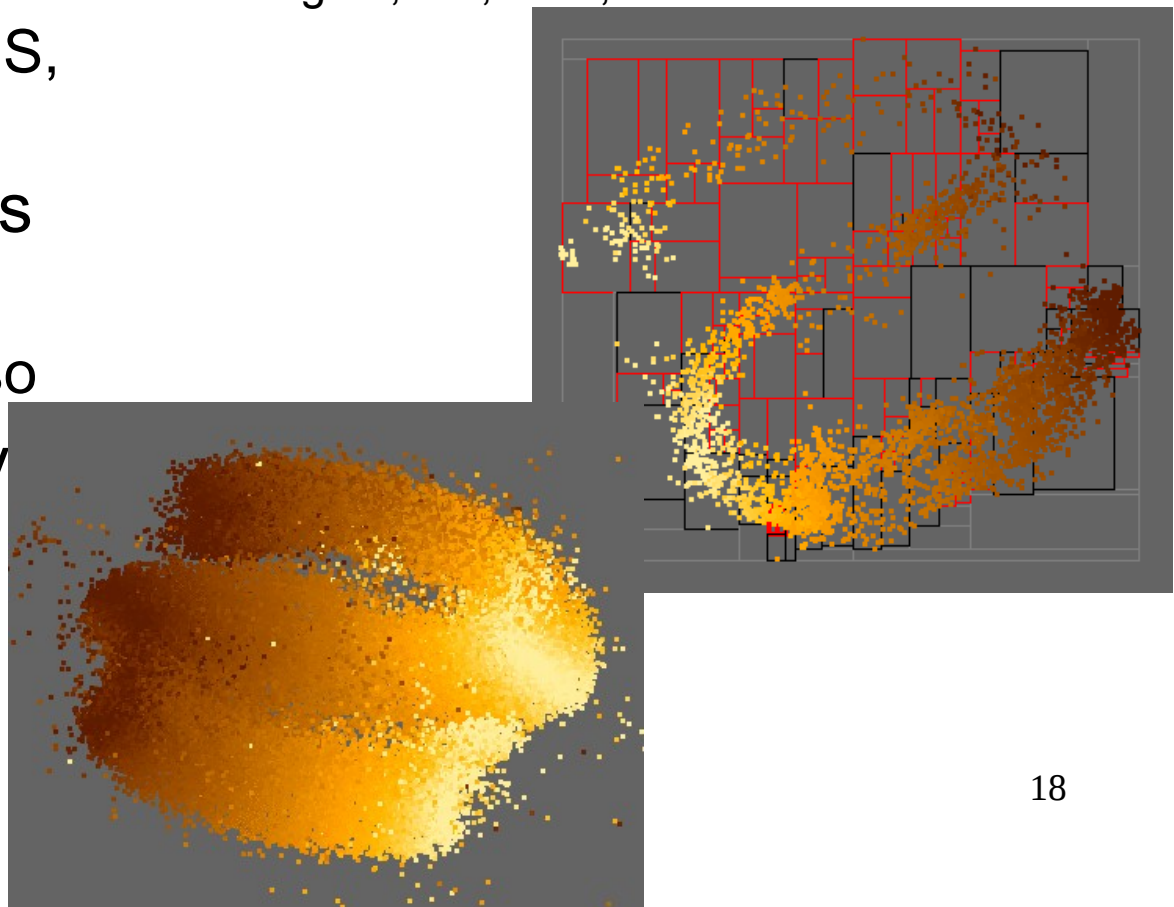
Analytical Strategies

- 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.

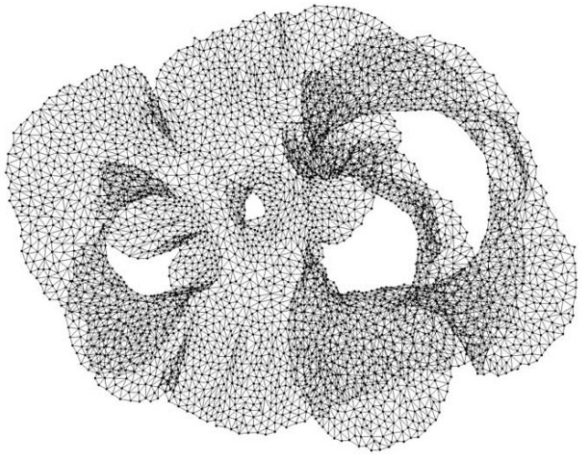


Analytical Strategies: Multiresolution

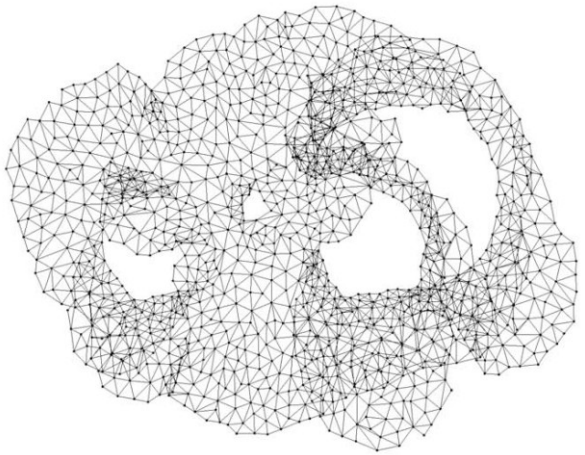
- Some algorithms can start with low resolution and increase it dynamically
- Graph Drawing, Image Transforms, etc.
- Let them pass the results when they are

- Allow them to be steered

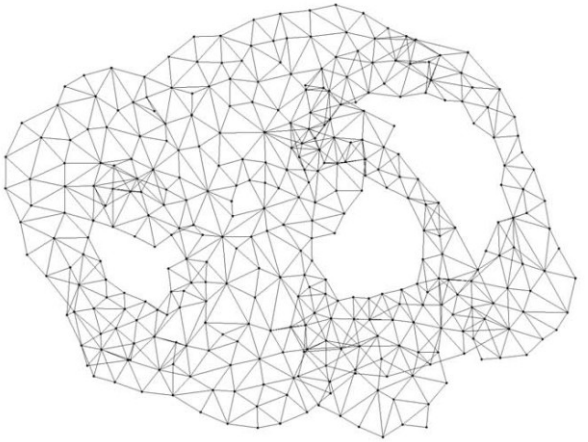
Emden R. Gansner, Yehuda Koren, Stephen C. North, "Topological Fisheye Views for Visualizing Large Graphs," IEEE Transactions on Visualization and Computer Graphics, pp. 457-468, July/August, 2005



4394-node approximation



1223-node approximation



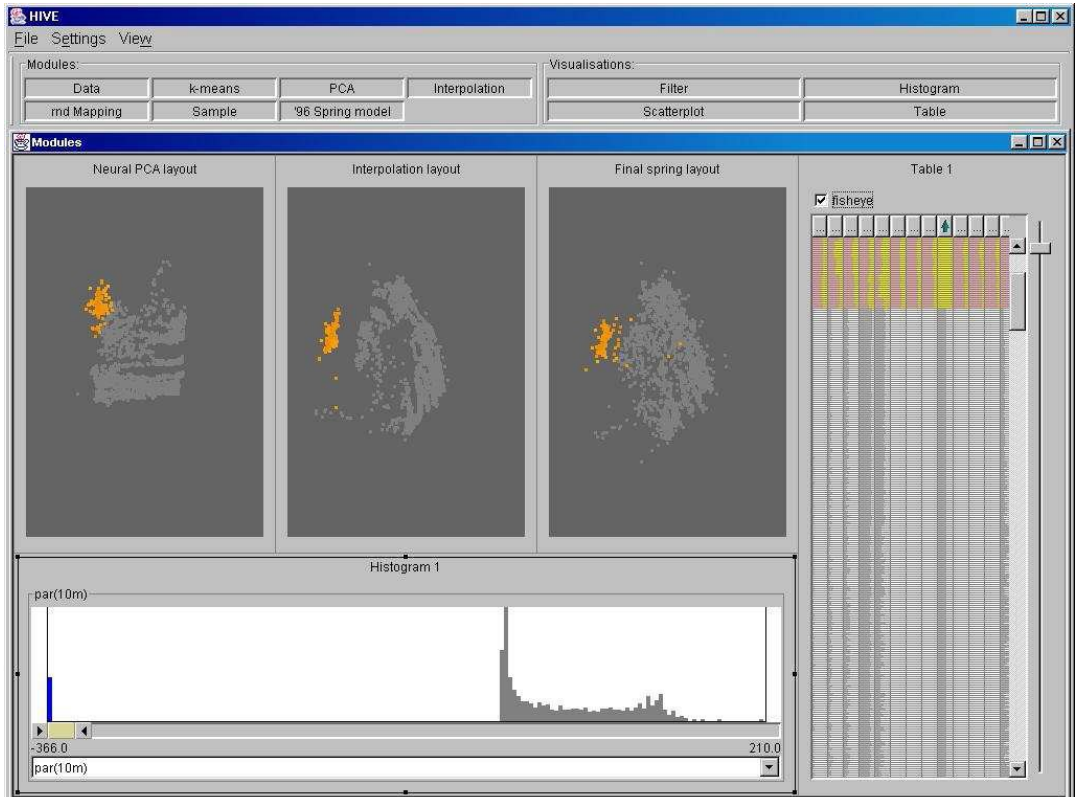
341-node approximation

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

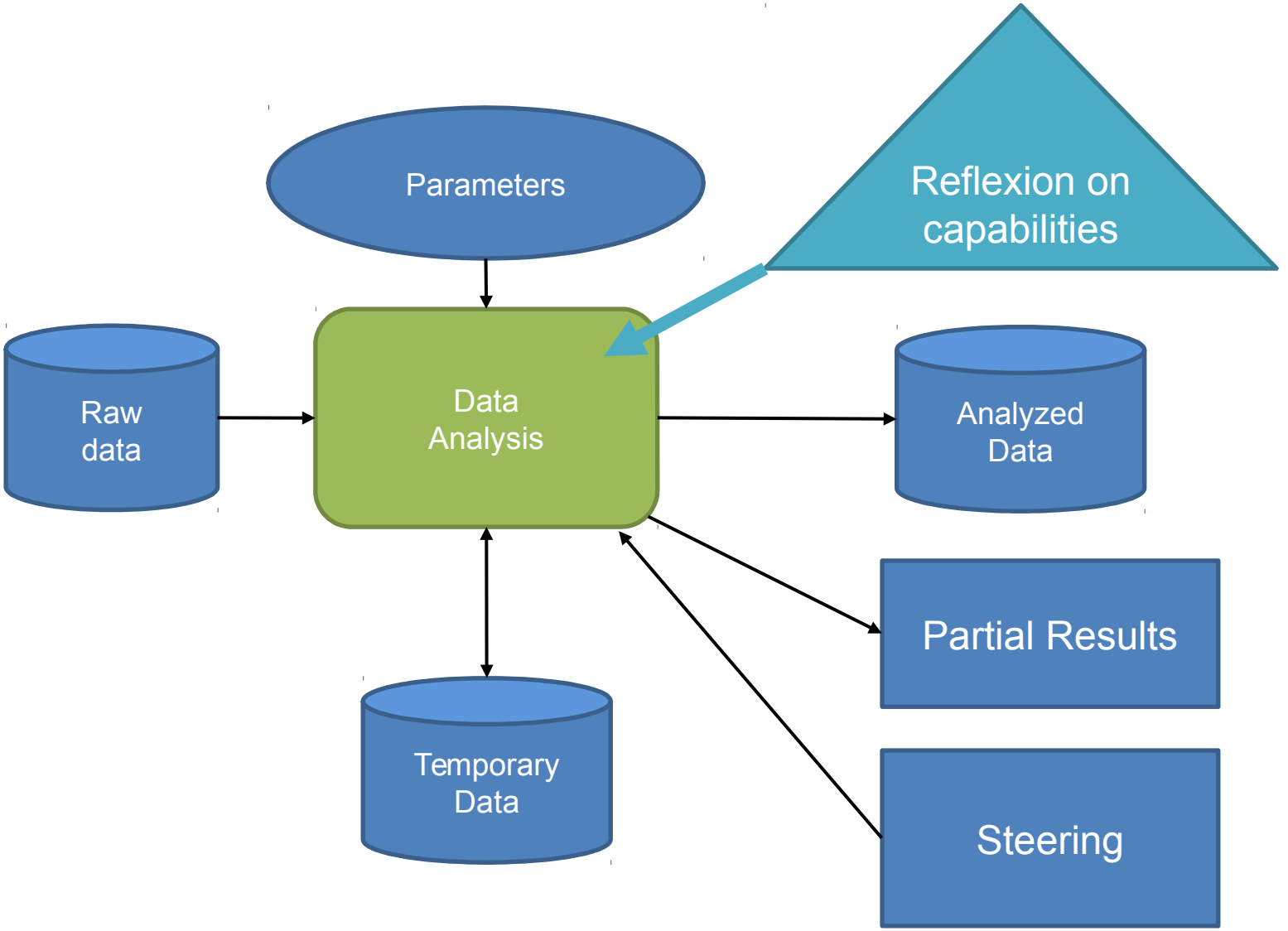


Analytical Strategies: Incremental update

- 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

Data Analysis Improved Ref Model



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?

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

Examples

- R
- Python
- Ipython notebook

R

- http://www.r-project.org/
- Old and mature free system for statistical computation
- Extremely popular and powerful
- Based on the scripting language S
- Rich package system (> 1000 packages)
  - Native libraries (C,Fortran,C++)
- Rich IDE (Rstudio)
- Can run as a server for other systems

Python

- Scripting language
  - Originally aimed at gluing between libraries

  Provides a powerful module system
- Extended with lots of analytical modules
  - Numpy, scipy for general mathematical computations
  - Pandas for data analysis (à la R)
  - Others for specific analyses
    - OpenCV for video, many graph manipulations
  - Connect to all the standard databases

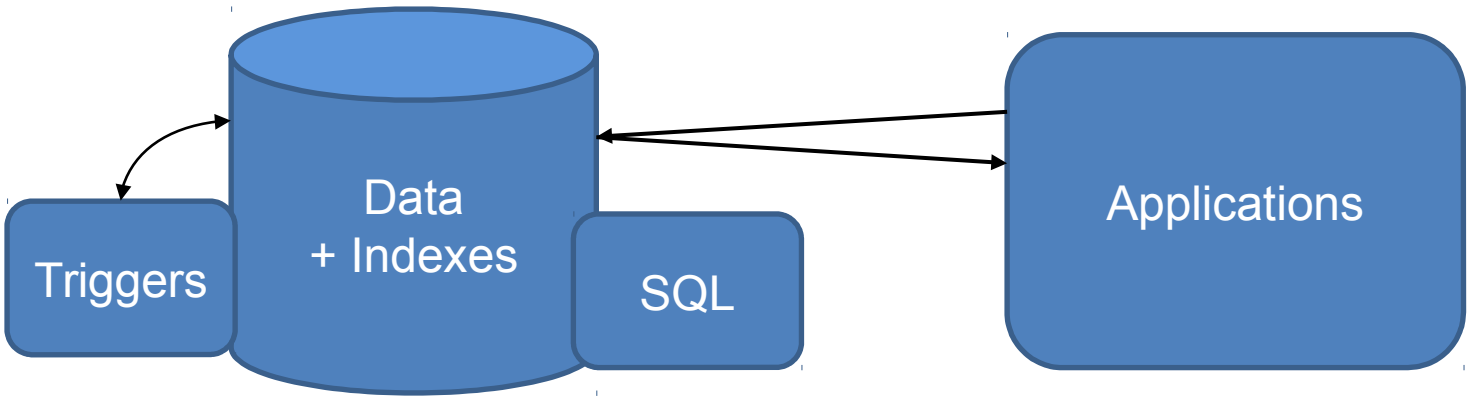
Ipython notebook

- Extension of python
  - Online web-based environment
  - Perform analysis, document then, save them for reuse
  - Format for printing, sharing, documenting
  - Can distribute work on parallel machines
- For demos, see
  - https://github.com/ipython/ipython/wiki/A-gallery-of-interesting-IPython-Notebooks#statistics-machine-learning-and-data-science
- Manages “provenance”

Provenance

- Data provenance documents the inputs, entities, systems, and processes that influence data of interest, in effect providing a historical record of the data and its origins. The generated evidence supports essential forensic activities such as data-dependency analysis, error/compromise detection and recovery, and auditing and compliance analysis. [http://siis.cse.psu.edu/provenance.html]
- Being able to trace back where the data came from, how it has been processed, and why to:
  - Re-apply the process
  - Document the process
  - Check for mistakes or biases
  - Search alternate methods or continuations

Data Management Reference Model



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 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

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

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)

Examples

- MySQL
- MonetDB
- ElasticSearch

MySQL and EdiFlow

- MySQL is a standard SQL DB
- Scalable through replication
- Notification through Triggers
- But notification to applications needs extensions
  - Implement a TCP/IP service for notification
- Works but painful and not portable

MonetDB

- Experimental/Stable High Performance DB
  - https://www.monetdb.org/Home
- SQL
- Scalable by replication
- Column oriented
- With extensions
  - SAMPLE
  - Multidimensional arrays
- Provide a connector with R
  - http://monetr.r-forge.r-project.org/

Coupling R and MonetDB

- Hannes Mühleisen and Thomas Lumley. 2013. Best of both worlds: relational databases and statistics. In Proceedings of the 25th International Conference on Scientific and Statistical Database Management (SSDBM), Alex Szalay, Tamas Budavari, Magdalena Balazinska, Alexandra Meliou, and Ahmet Sacan (Eds.). ACM, New York, NY, USA

```
c<-dbConnect(MonetDB.R(), "monetdb://localhost/db1")

mf<-monet.frame(c,"t1")

mean(subset(mf,c1>42)$c2)
```

- The operation is mainly performed on the database and returned to the R system
- Great when MonetDB knows the operations
- Otherwise, the table columns should be transferred from DB to R

ElasticSearch

- NoSQL database (JSON)
- Scalable horizontally and vertically
  - Replication and decomposition in small units
- Powerful indexing
  - Text with NLP (à la Google)
  - Others with multiple aggregation operators
  - Number / geographical aggregation and binning
- Can delegate some complex operations to the database, retrieving aggregated results

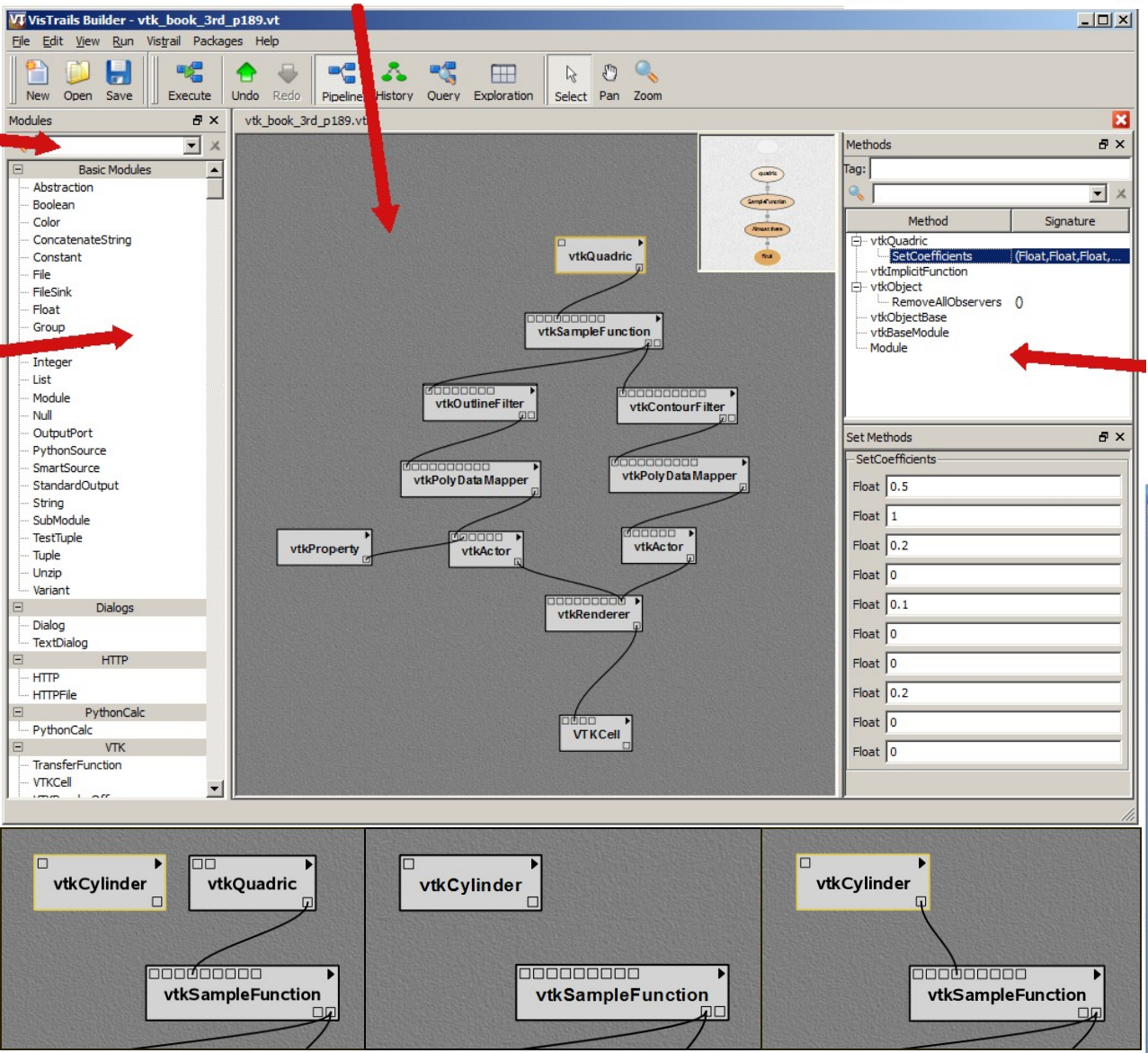
Scientific Workflow Systems

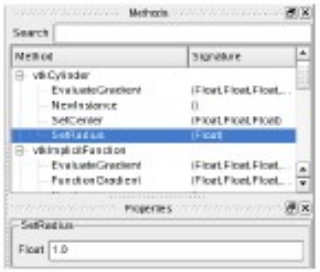
- Combining data management + computation + visualization
- Lots of ad-hoc Scientific Workflow Systems (e.g. Kepler)
- With (Sci) Visualization: VisTrails!
- Impressive system
  - Exploration + data provenance

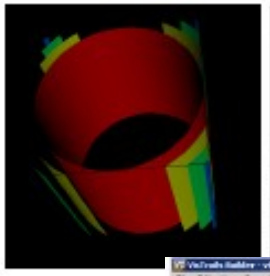
Carlos E. Scheidegger, Huy T. Vo, David Koop, Juliana Freire, and Claudio T. Silva. 2008. Querying and re-using workflows with VsTrails. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data* (SIGMOD '08). ACM, New York, NY, USA, 1251-1254.

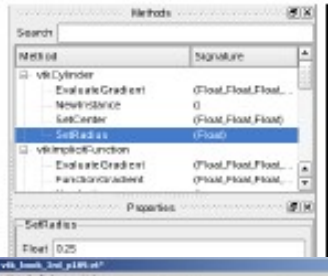
www.vistrails.org

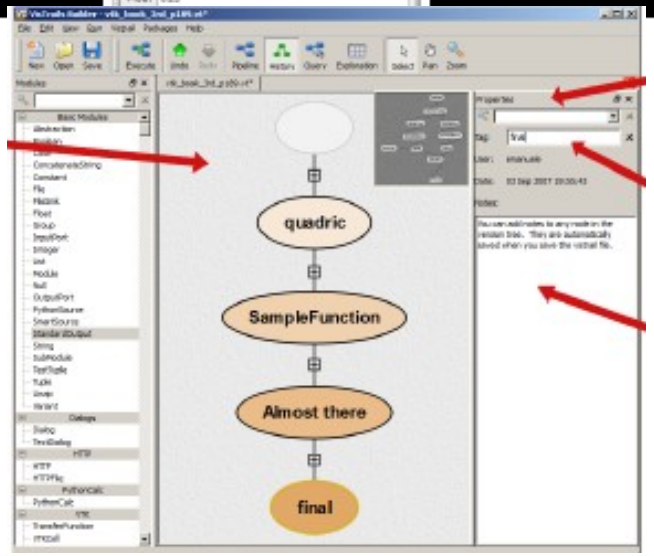
VisTrails

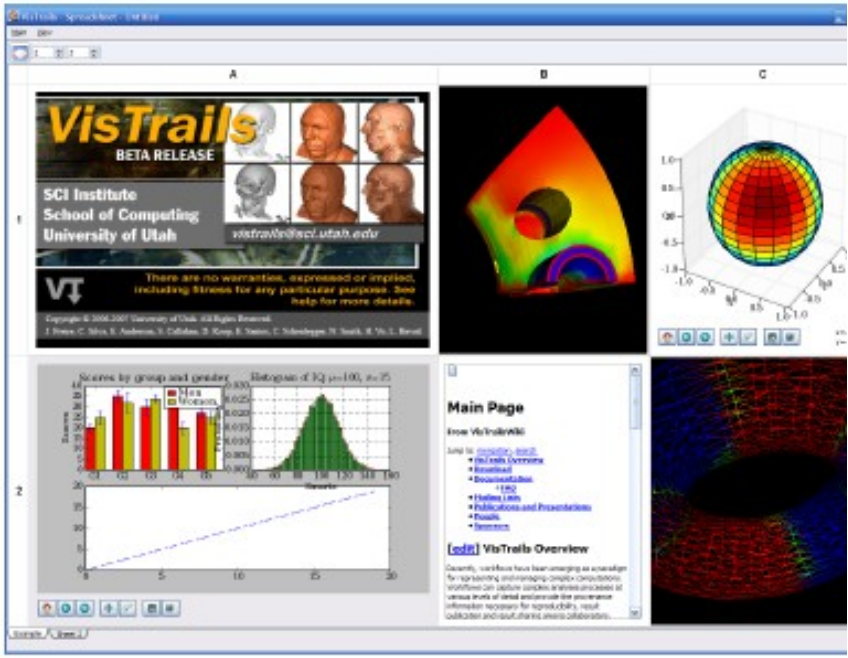












Wed. November 8th

VA Tools

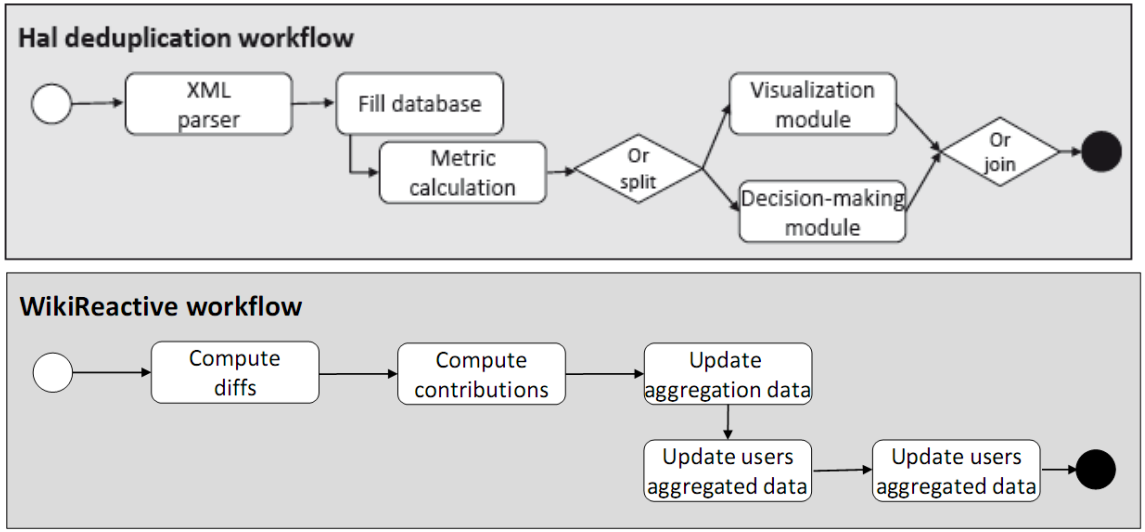
Workflow Systems

- 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.

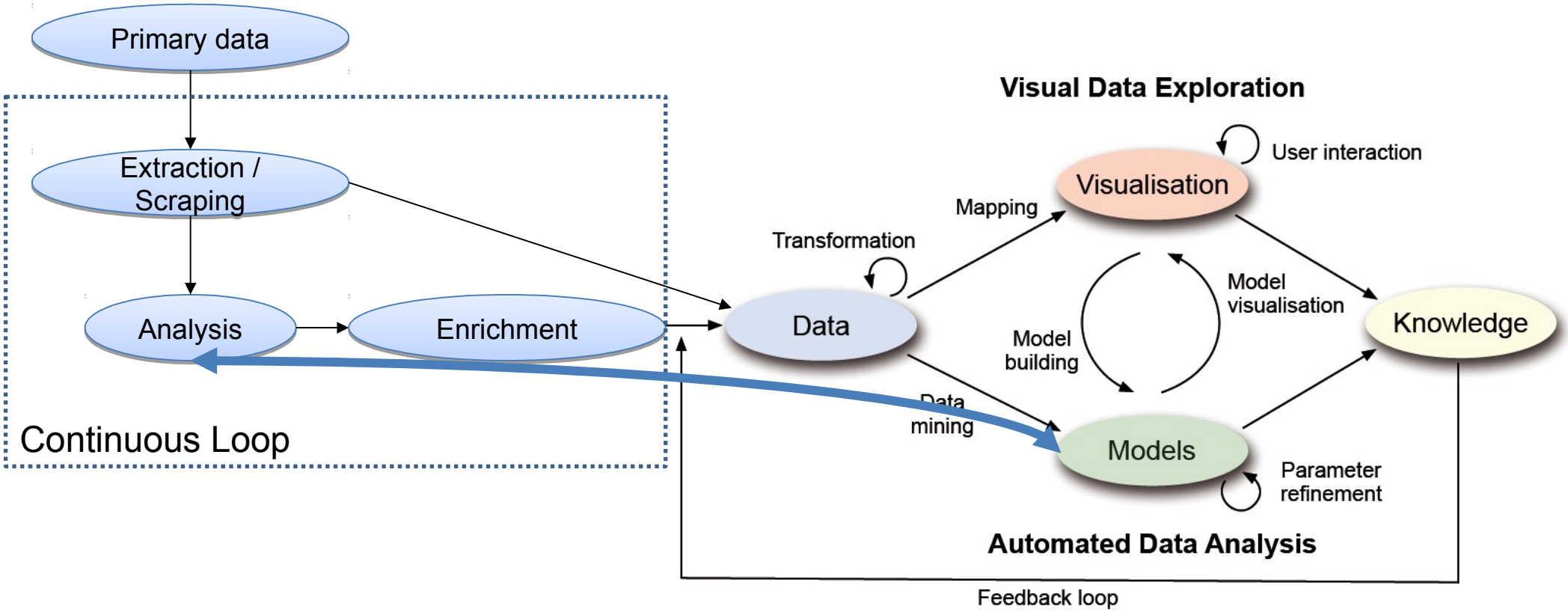
Workflow for the Continuous Loop

- 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



The Visual Analytics Process Extended



Take Away Message

- Off-the-shelf analytics tools are not quite ready for Visual Analytics
  - They don't consider the human in the loop
- Possible to adapt the tools
  - Avoid gratuitous data copy
  - Split computations in smaller chunks to get fast results
- Maintain Data Provenance
  - Not well supported but improving
- Many Existing Vis/Ana/DB Systems are flexible enough to be adapted for your applications