Visualization and text mining of patent and non-patent data

Anton Heijs

Treparel Information Solutions
Delft, The Netherlands
http://www.treparel.com/

ICIC conference , Nice , France , 2008

Outline

Introduction

Mining + visualization = visual analytics

- More data - less insight
- Overview needed

Text mining approaches controlling over/under-fit

Rule based : difficult to control under fit

Natural language processing : difficult to control over fit

Machine learning : controls over-fit and under-fit

Document clustering : finding patterns unsupervised

Unsupervised text mining : clustering

- Input is a vector representation of all documents
- Separate the data in its natural groups based on a similarity metric
- Output is a similarity matrix

Document classification : finding patterns supervised

Supervised text mining : classification

- Input is a vector representation of all documents
- Determine a hyper plane which separates all the data
  - Binary classification : classify to 1 class
  - Multi class classification : classify to multiple classes
  - Compound classification : combine multiple binary classifiers
- Output is a matrix with classification scores

Text analytics

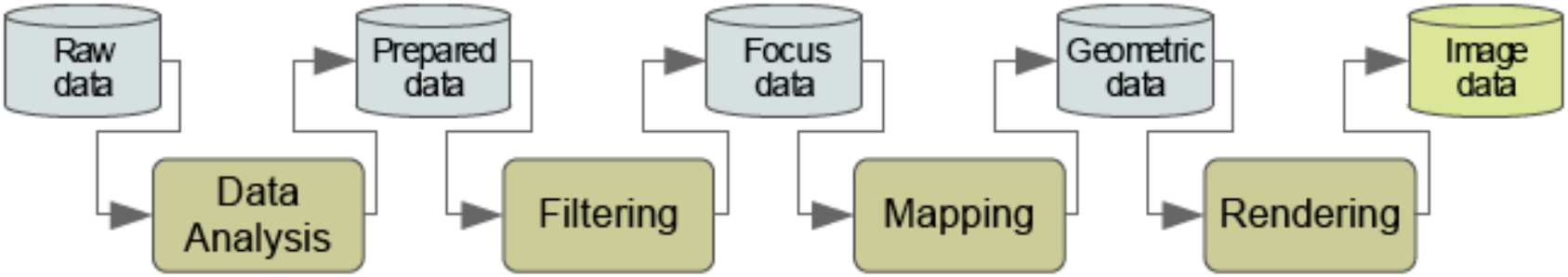
Text analytics is text mining and visualization

- Combined use of mining and visualization for analysis
- Uses text mining to determine trends in the data
- Use visualization to make complex data understandable

Applications on patent and non-patent literature

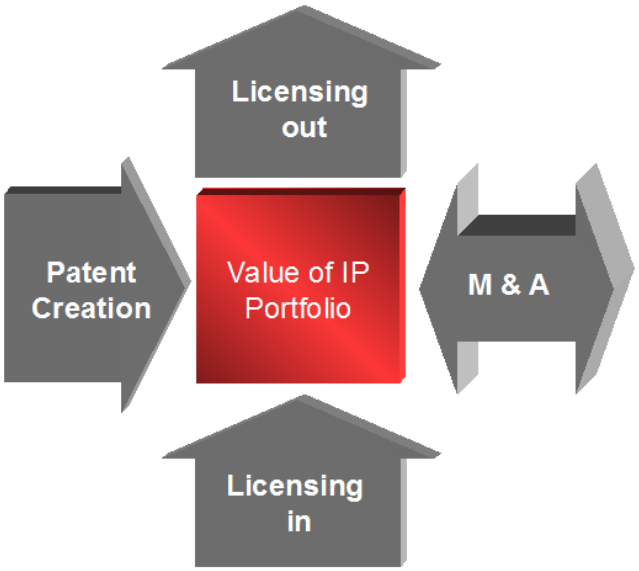
Text mining and visualization steps

- management and preprocessing data
- text mining (filtering the data in the pipeline)
- mapping filtered data to geometry
- rendering and interaction

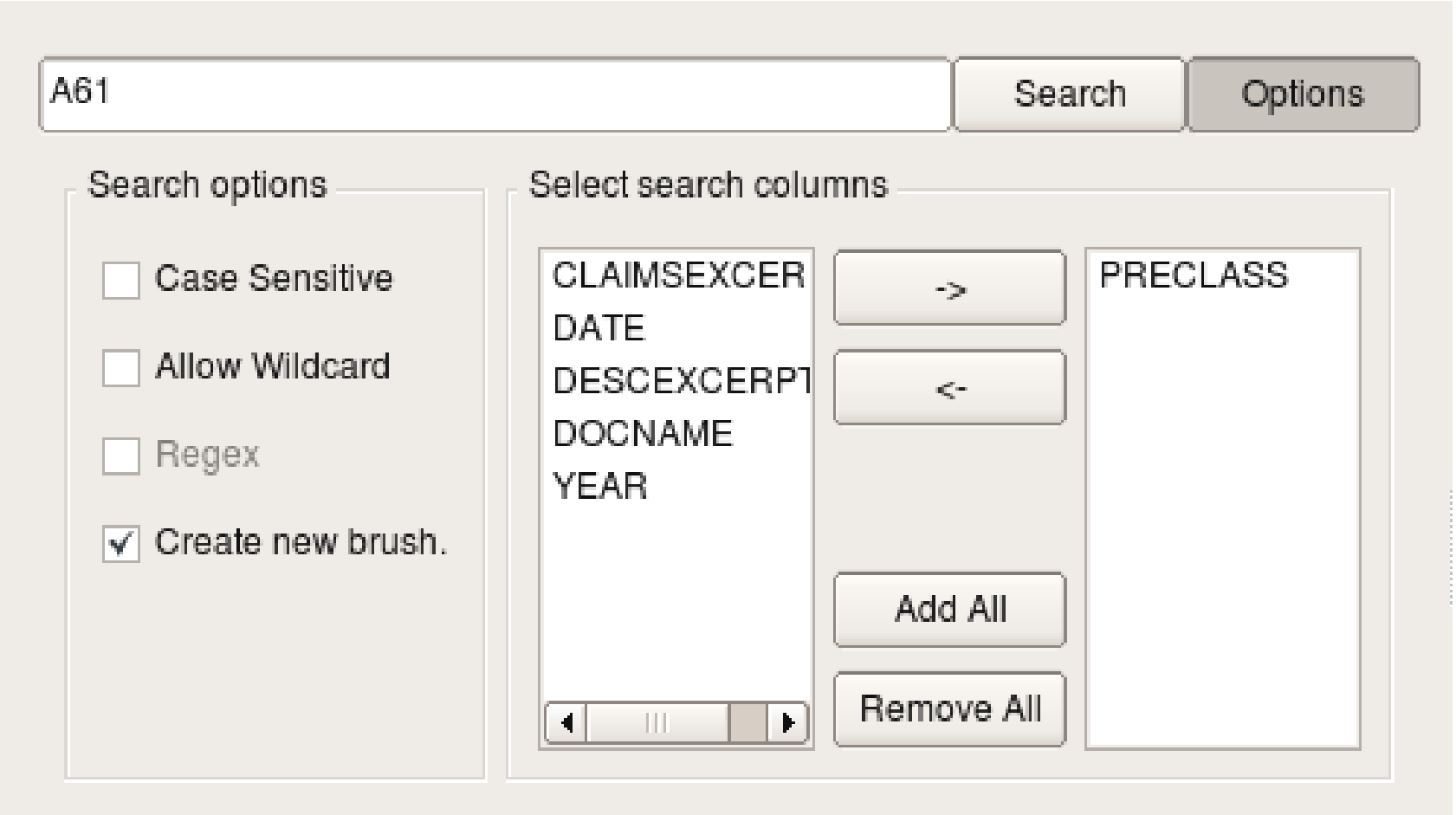


Patent analytics applications

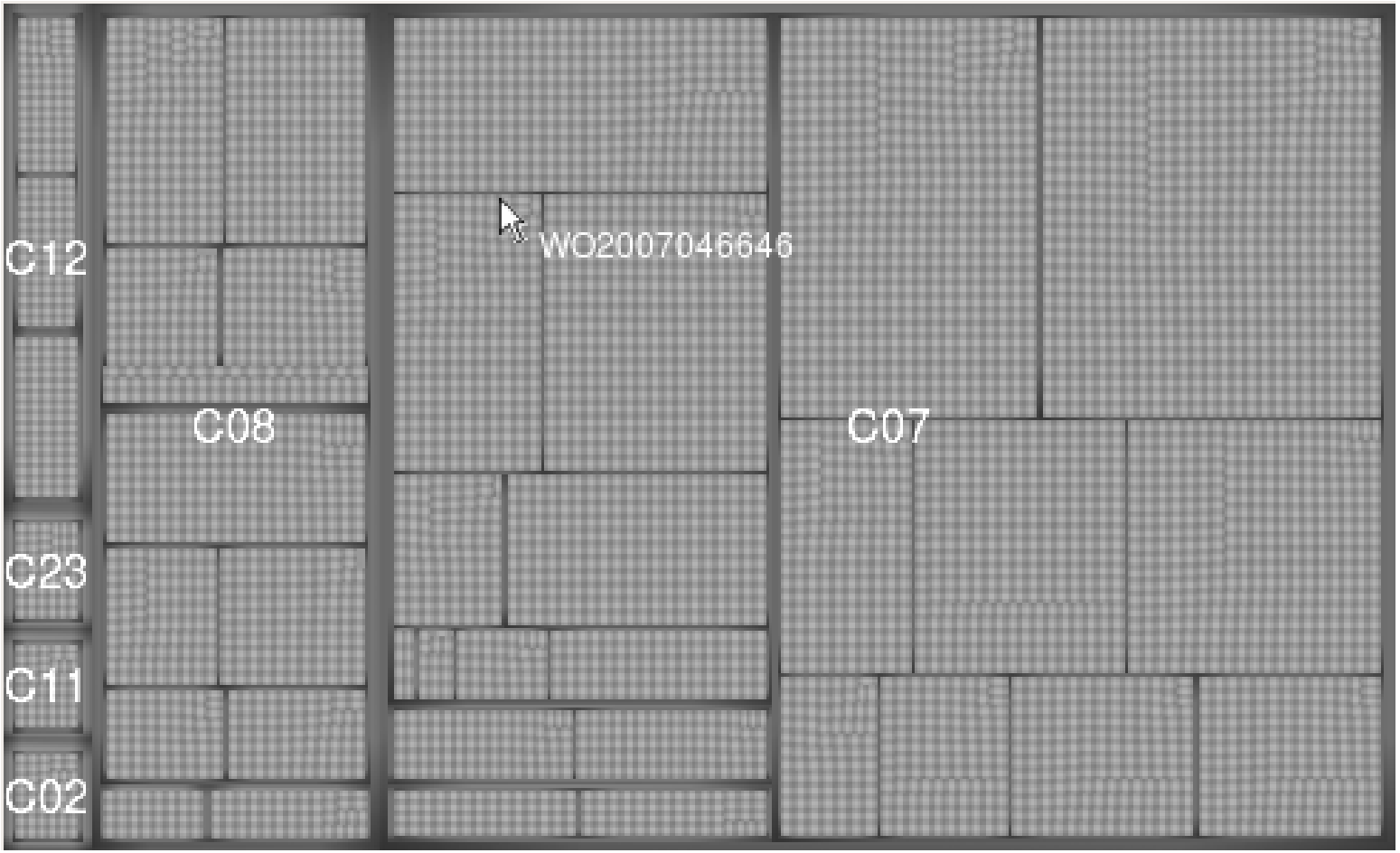
- Patent search using classification and clustering algorithms
- Patent landscaping, spotting new opportunities
- Patent ranking, utilization optimization and valuation



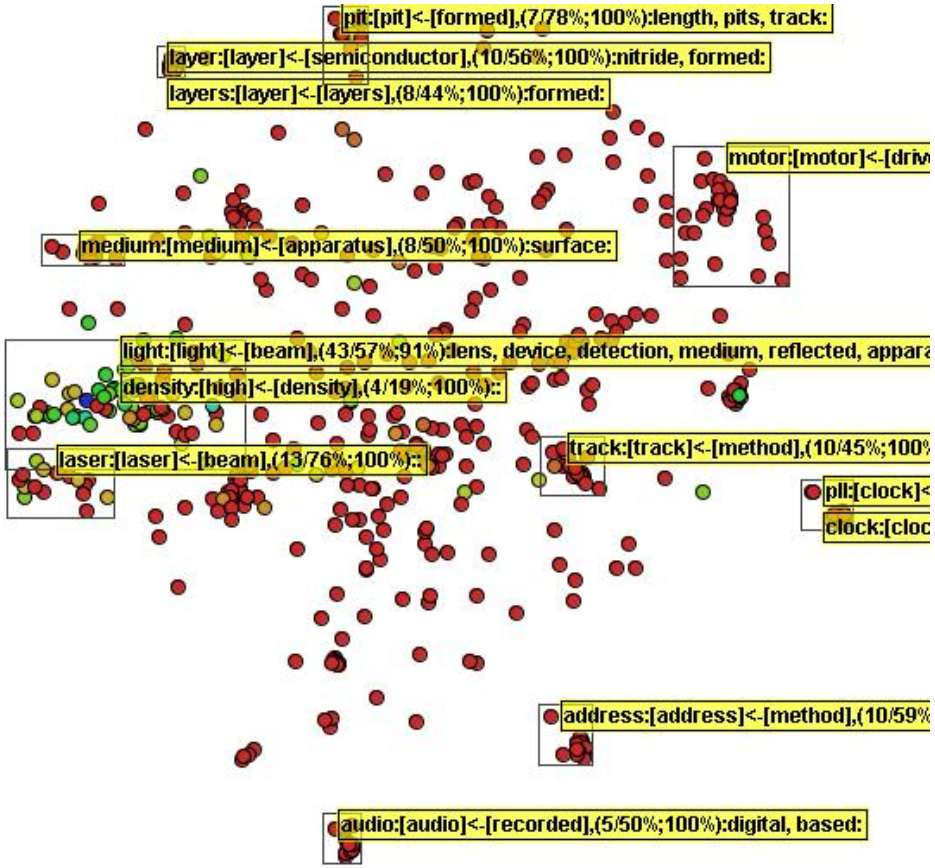
Search by querie by class



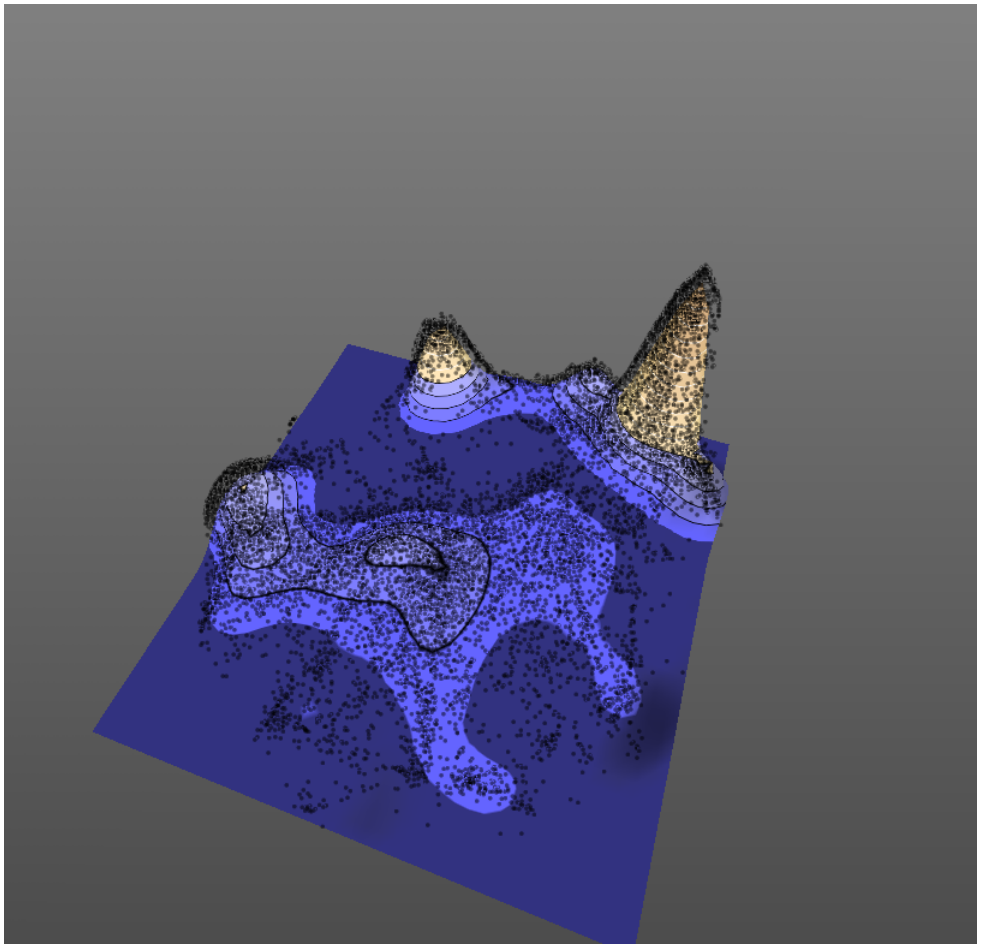
Tree map visualization of the patent classification



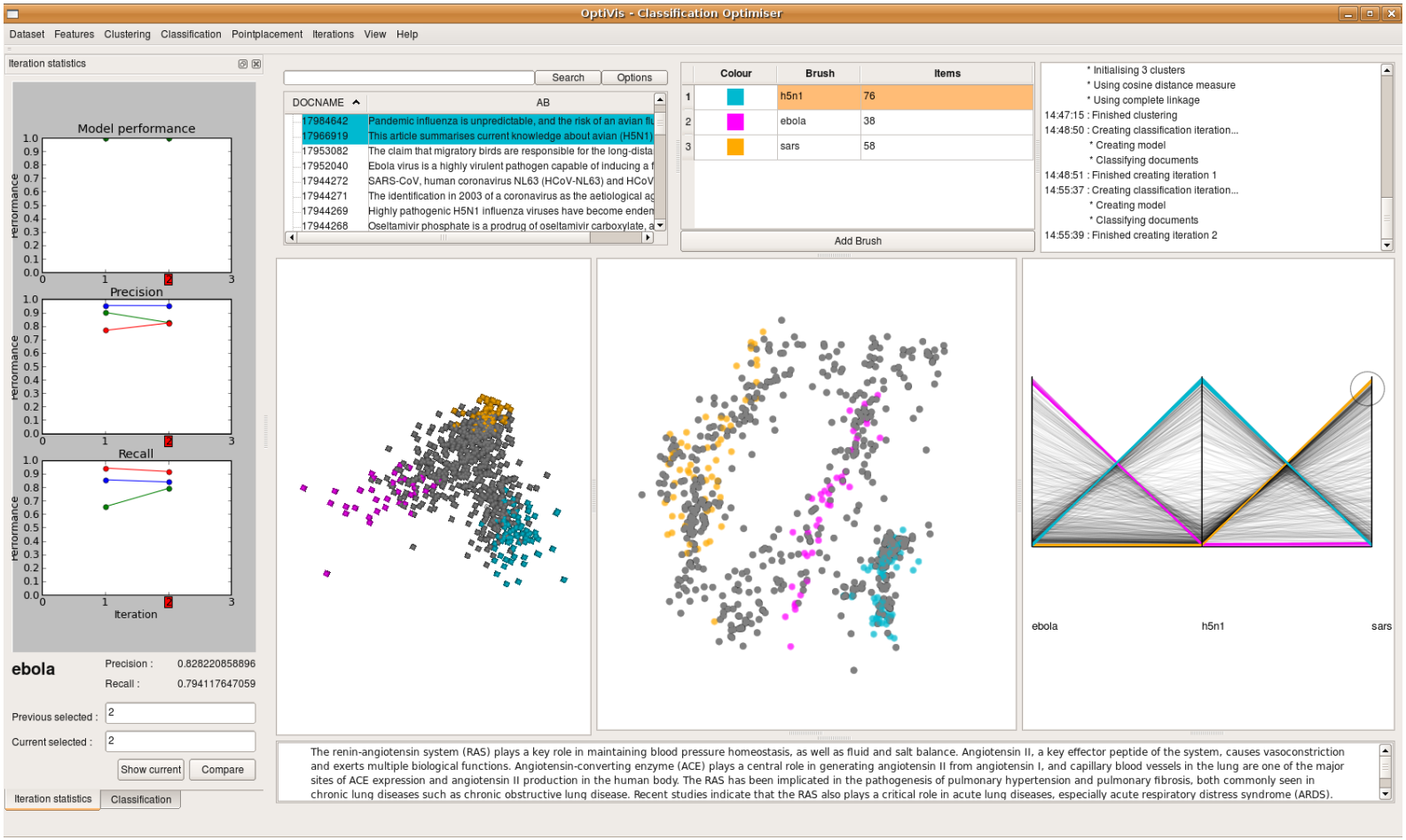
Clustering of patents of 7 classes in Optical Recording



Clustering of large patent sets

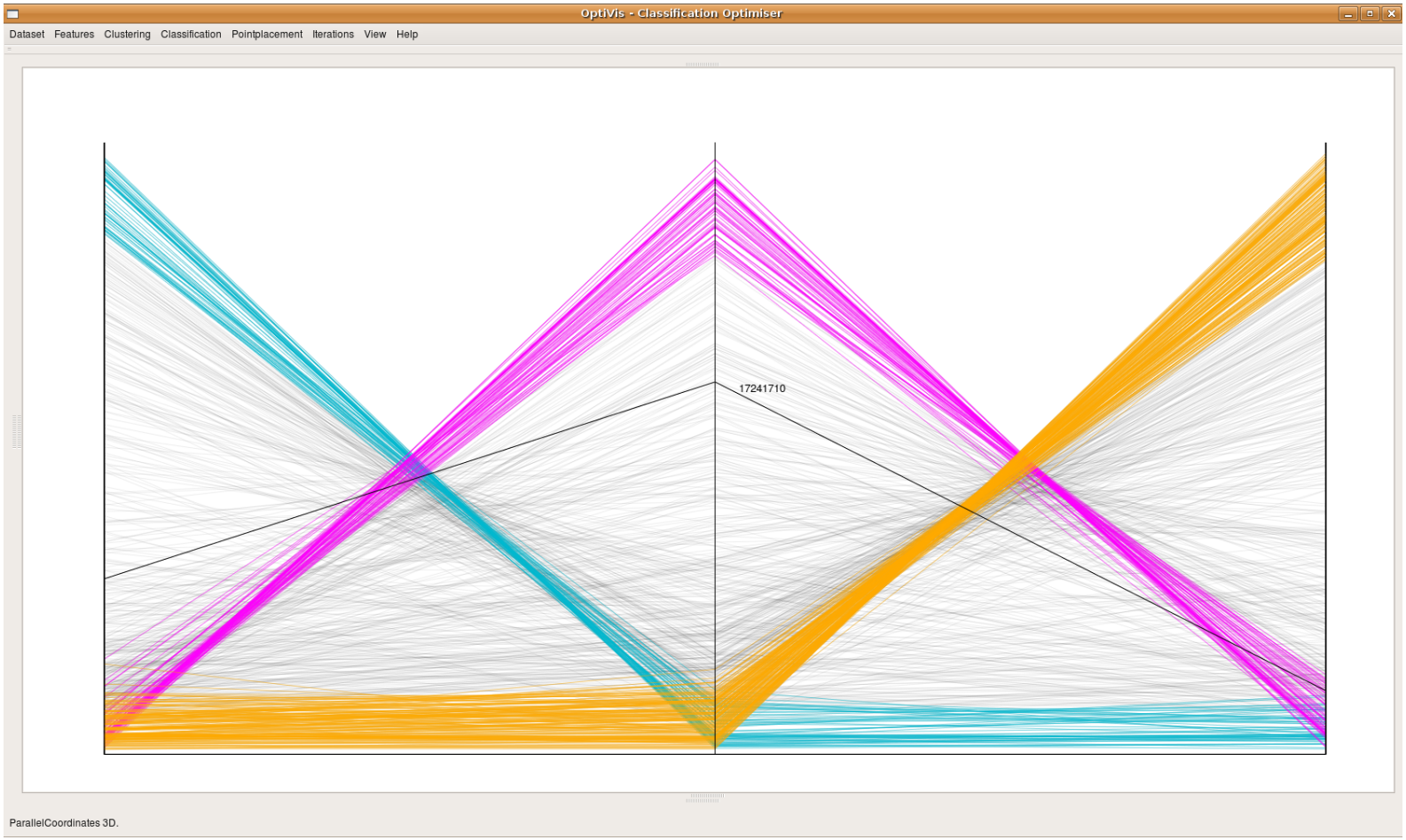


Building a classifier

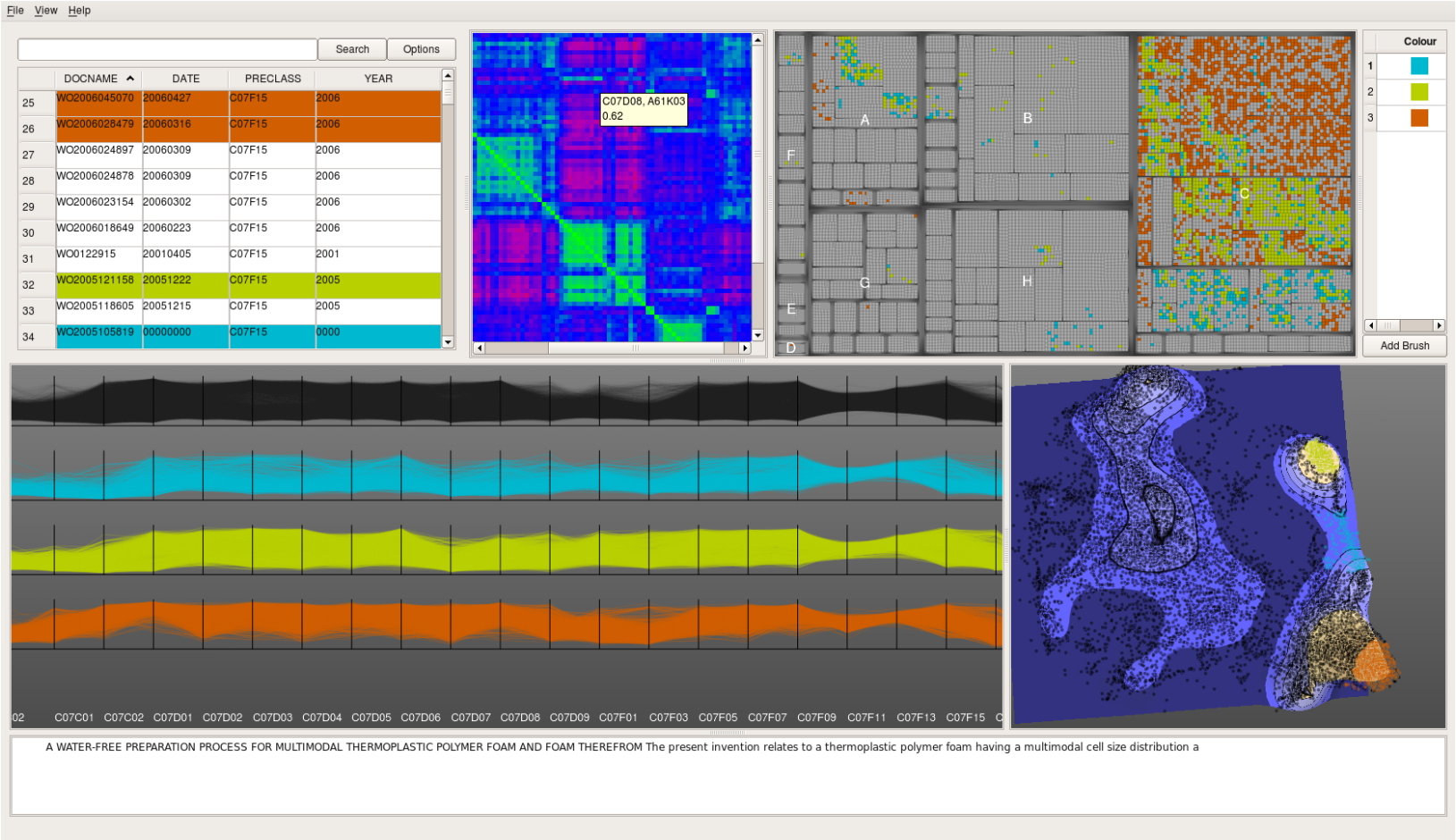


Testing a classifier

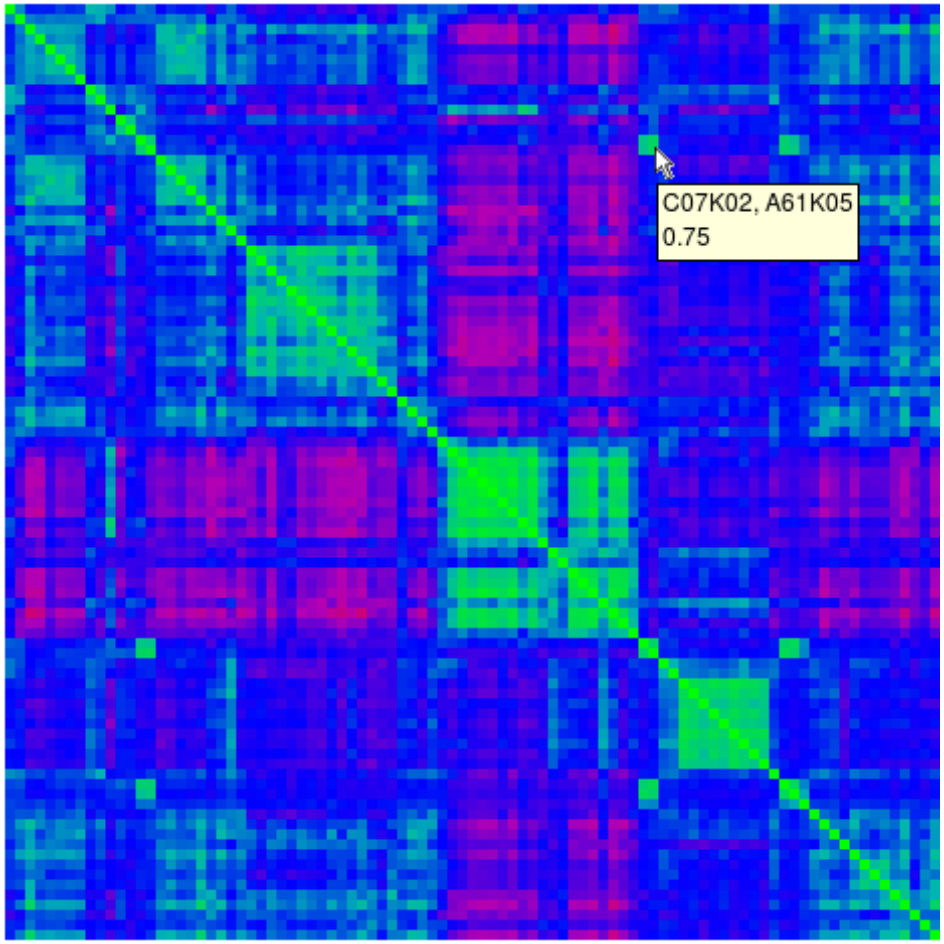
Visualization of the classifier performance



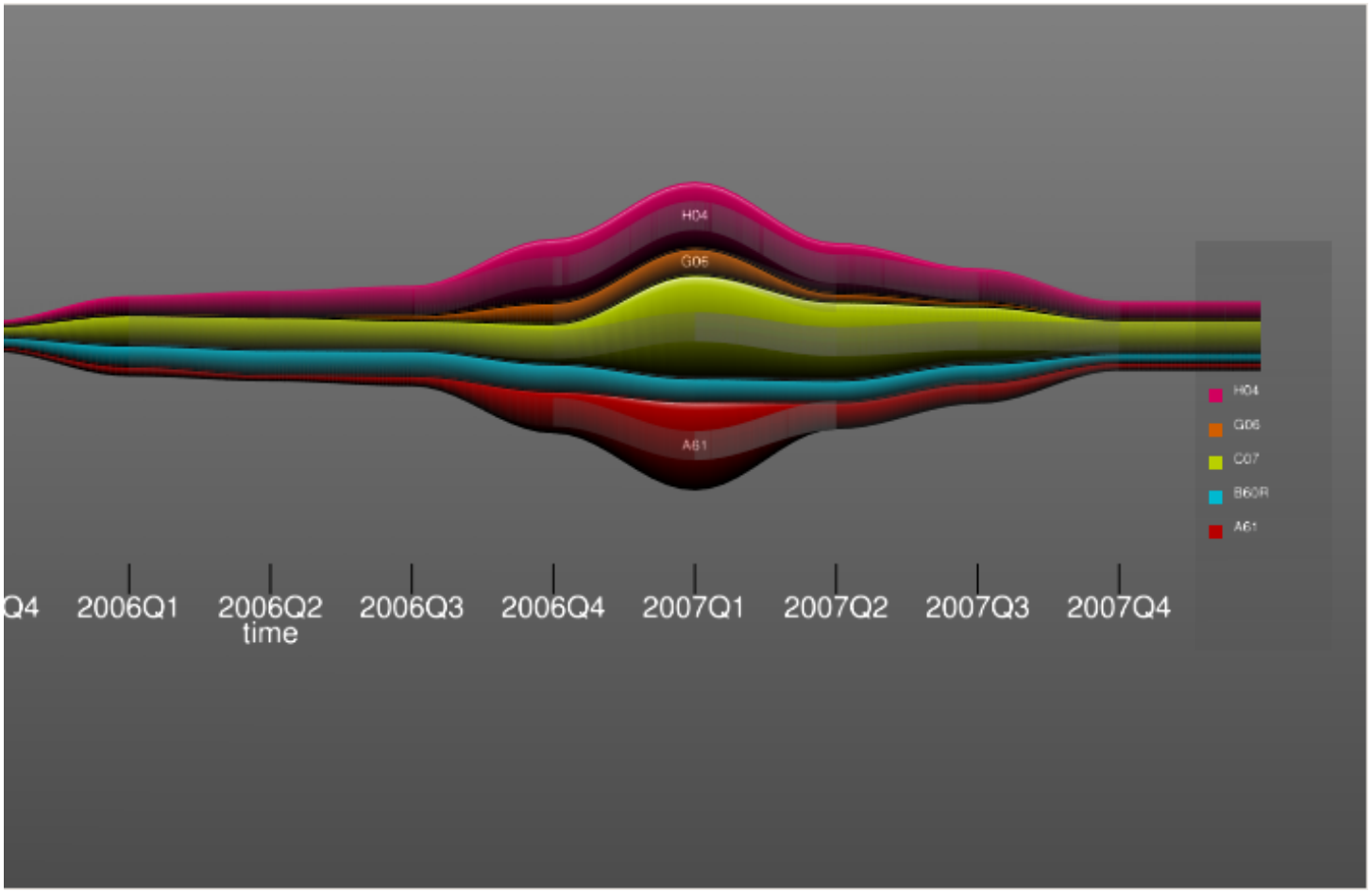
Patent analytics application combining mining and visualization



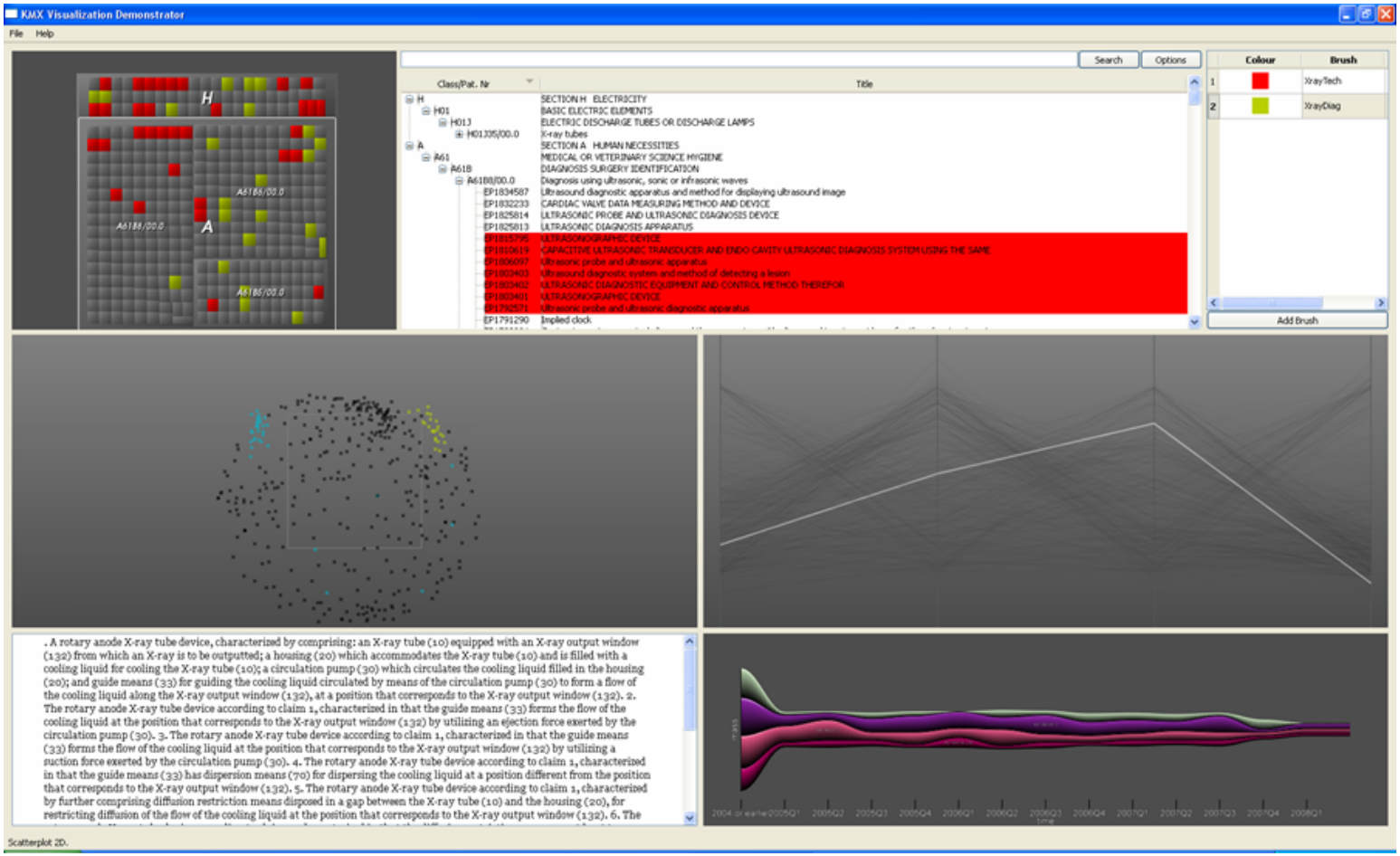
Correlation between patents



Patents over time



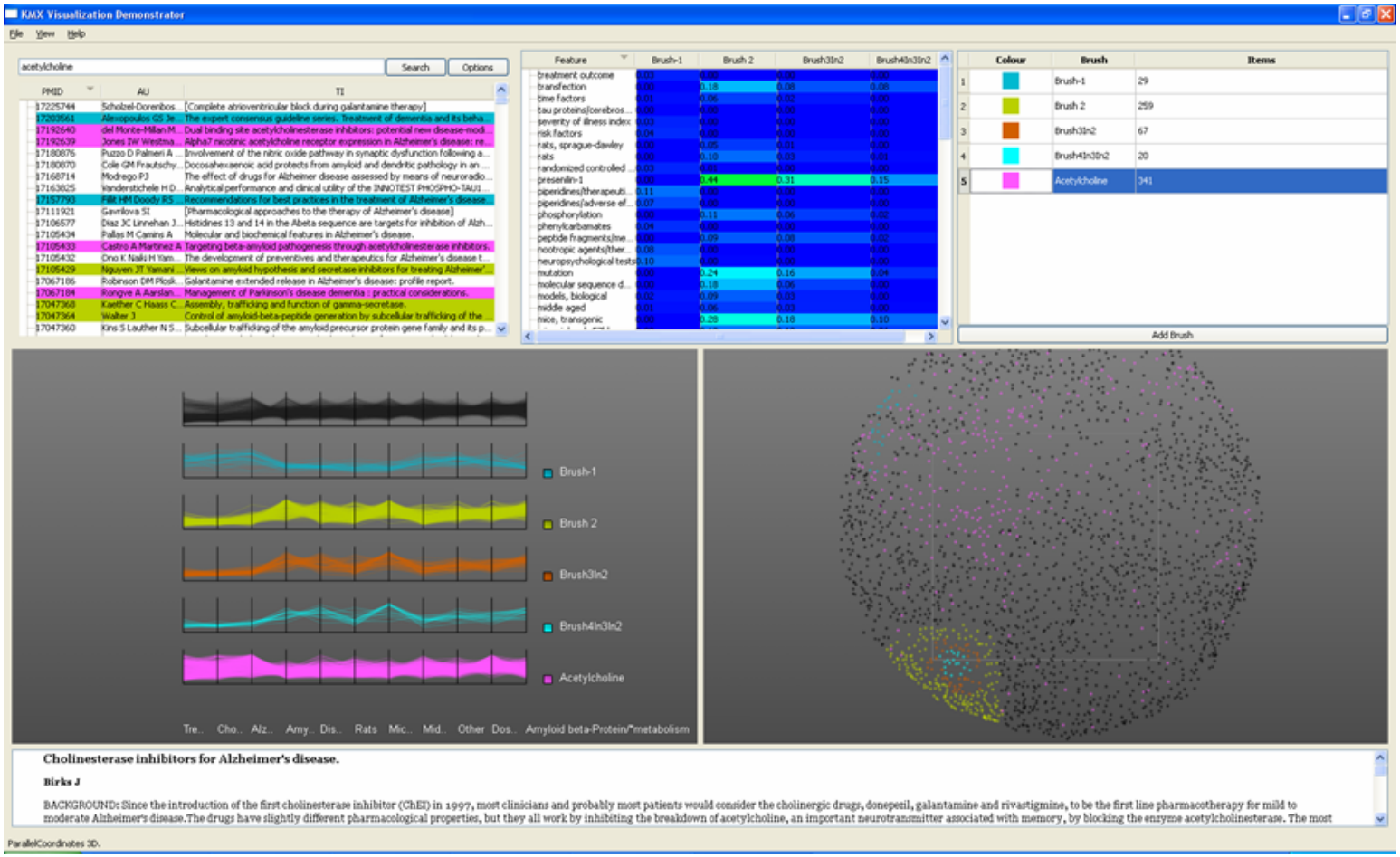
Patents visualization hierarchically, supervised, unsupervised and over time



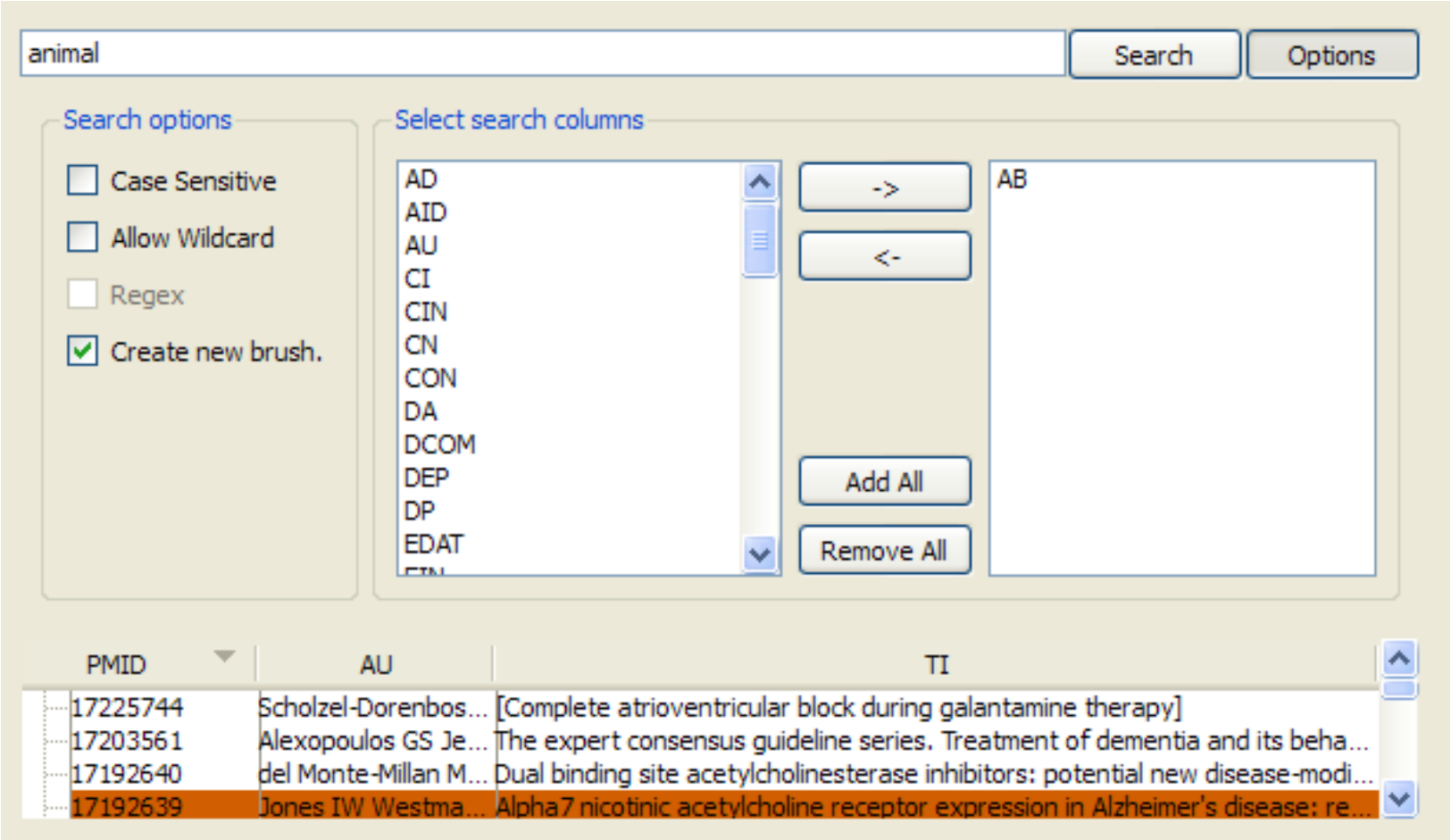
Text mining and visualizations applications on Medline

- Search for facts and relationships (using classification and clustering)
- Disambiguation (using classification)
- Named entity recognition (using classification)
- Concept and relationships discovery
- Knowledge discovery by integrating text mining and visualization with
  - Data mining : mining table data from experiments
  - Image mining : micro array analysis
  - Graph mining : path way analysis

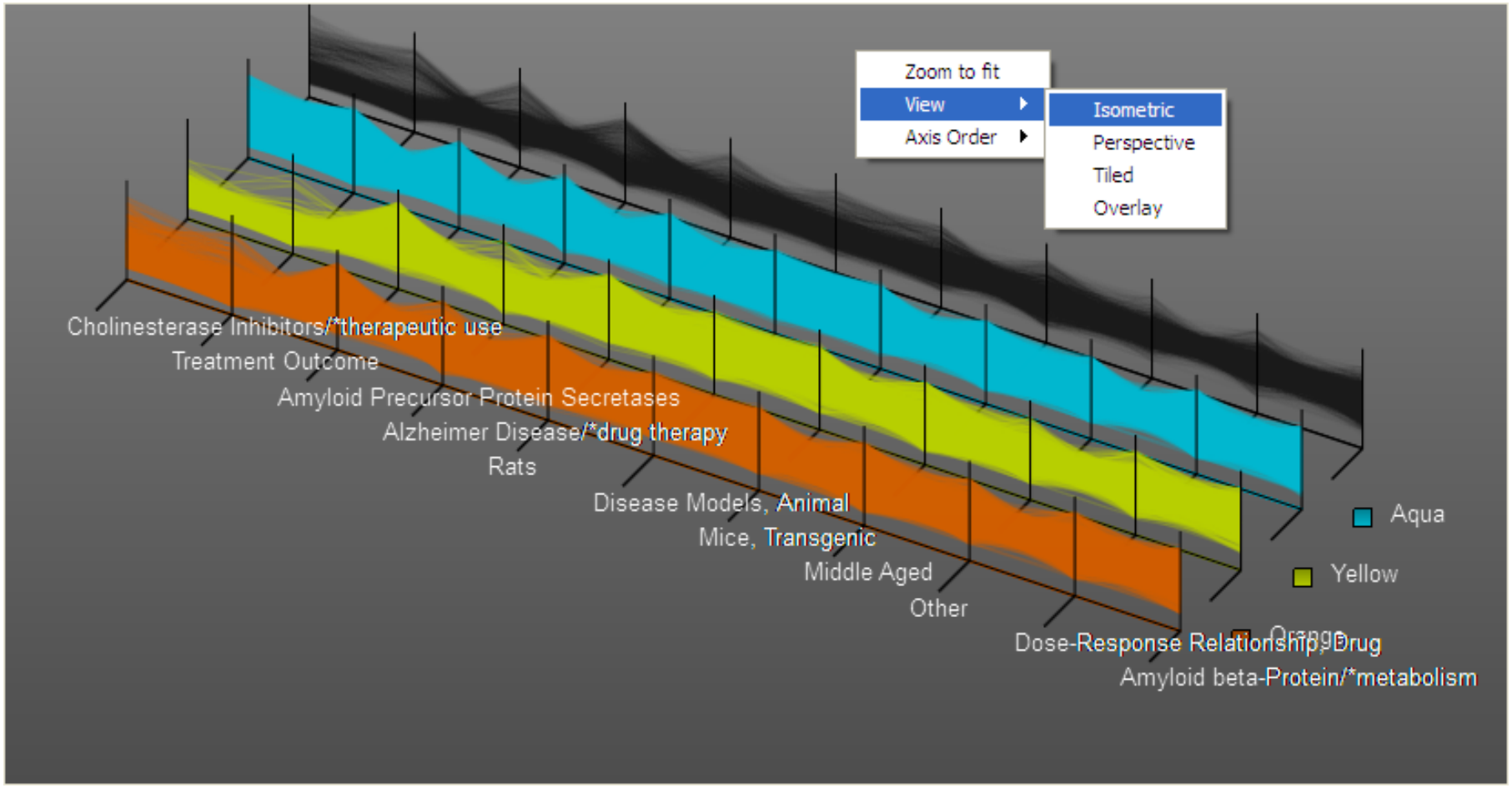
Mining and visualization Medline



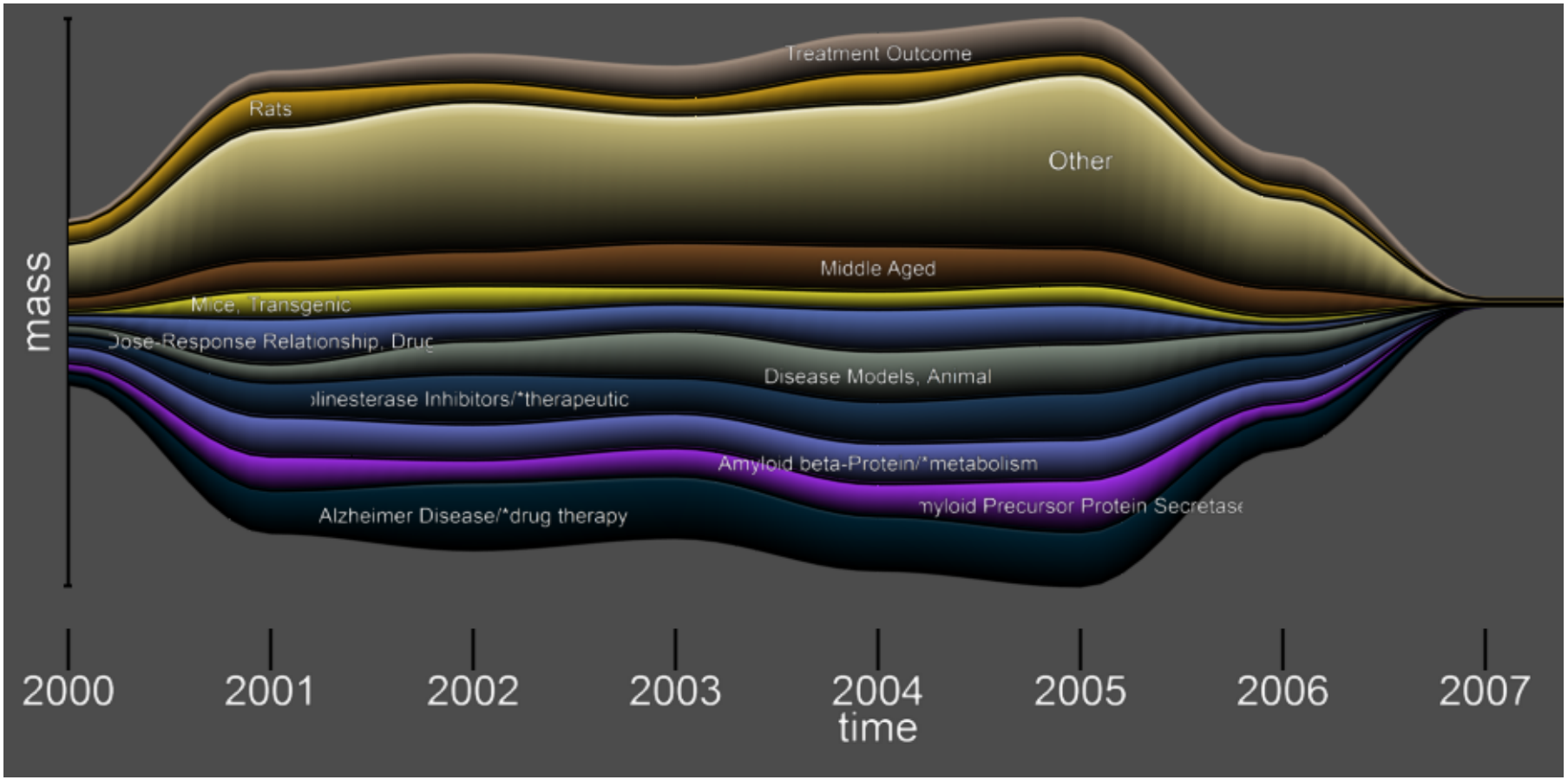
Search Medline by querie



Classification of Medline documents to multiple topics



Visualization of trends in Medline documents over time



Conclusions

- Combination of text mining and visualization algorithms enables more application solutions
- Multiple application solutions are a combination of a small set of text mining algorithms
- Multiple application solutions require multiple coupled view visualizations to provide complete

Future development

- Patent analytics will include patent ranking and valuation
- Text mining will integrate with data mining , for instance for patent portfolio management
- Stand alone applications and applications with work flow support will advance

Trends Patterns Relationships

http://www.treparel.com/

Enabling you to learn more and see more