From Knowledge Discovery To Understanding: Making Sense of Large Datasets

Rainer Stuike-Prill, FIZ Karlsruhe
and
Christine McCue, CAS

18 October 2005

cas

FIZ KARLSRUHE

Topics to be presented

- Challenges in today's information environment
- Information overload: separating good from bad
- Technological advances: we must keep up
- Various types of available analysis and visualization (A&V) tools
- Where STN AnaVist fits in

In today's information environment, some challenges are heightened

More than ever, information "means business."

- to be efficient
- to be innovative
- to be cost effective
- to be competitive

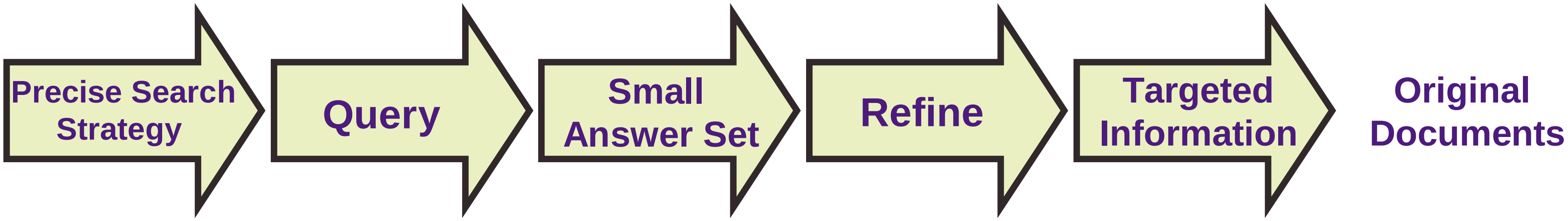
Information keeps growing

- In 2004:
  - more than 865,000 document records were added to the CA/CAplus databases . . .
  - and more than 2.3 million small molecules were added to REGISTRY
  - USPTO issued more than 187,000 patents (fiscal 2004)
  - EPO received over 178,000 patent applications
  - JPO had over 423,000 applications and registrations

Information is not always what you think it is

The data looked good. How could I have known the "scientists" were only 13 years old?

Classic online search and retrieval is most effectively used to answer specific questions



Precise Search Strategy

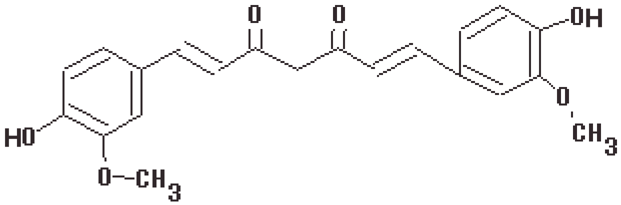
Query

Small Answer Set

Refine

Targeted Information

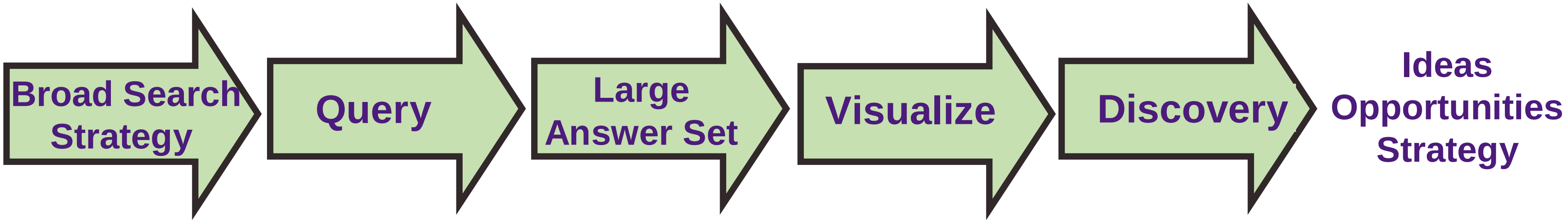
Original Documents

- What is the CAS Registry Number® for DMSO?
- What patents exist for the regeneration of transition metal iron phosphate catalysts?
- Is  a novel compound?

Professional searchers need to assimilate and present sci-tech information effectively

- R&D strategy
- Competitive intelligence
- Corporate intellectual property protection
- Business-critical decisions

Today, search professionals need the "big picture" view delivered by A&V to support strategic decision making



Broad Search Strategy → Query → Large Answer Set → Visualize → Discovery → Ideas Opportunities Strategy

- Who has water purification technology I can potentially license?
- What are the trends in my competitor's area of research?
- What are the potential applications of our patented compounds?

A classical application of A&V software in the US has been national security

ClearResearch from ClearForest . . . will be used to quickly analyze the FBI's entire document repository--over 1 billion existing documents plus up to one thousand new documents per day— *Clearforest press release, April 15, 2003*

Starlight Information Visualization System-developed for the U.S. defense community-uncovers key relationships hidden in large, complex, and dynamic information collections…integrates structured, unstructured, spatial, and multimedia data, offering comparisons of information at multiple levels of abstraction—*Battelle* press release, August 21, 2003

Many more A&V resources have appeared in recent years

ThemeScape

DELPHION

Aureka

anacubis

OMNIVIZ

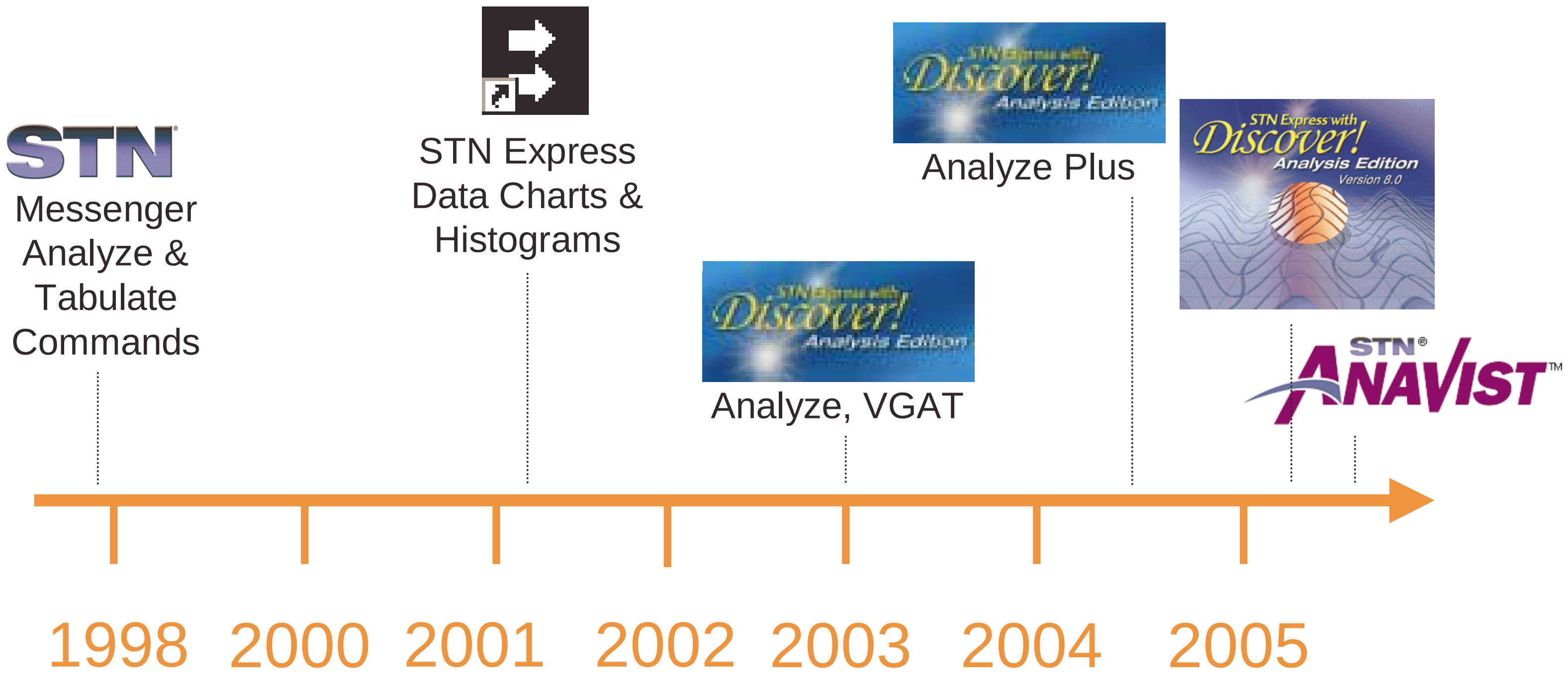
MicroPatent

VantagePoint

A&V features and capabilities vary substantially

- Text clustering
- Various cluster mapping
- Co-coding, co-assignee clustering from structured data
- Co-occurrence matrixes
- Organize documents & knowledge
- Citations, link citations, corporate trees
- Temporal analysis and circle graphs
- Charts and graphs for document taxonomies
- Wizard-driven and Visual C++ association rules generators

STN has also been developing A&V tools



Messenger Analyze & Tabulate Commands

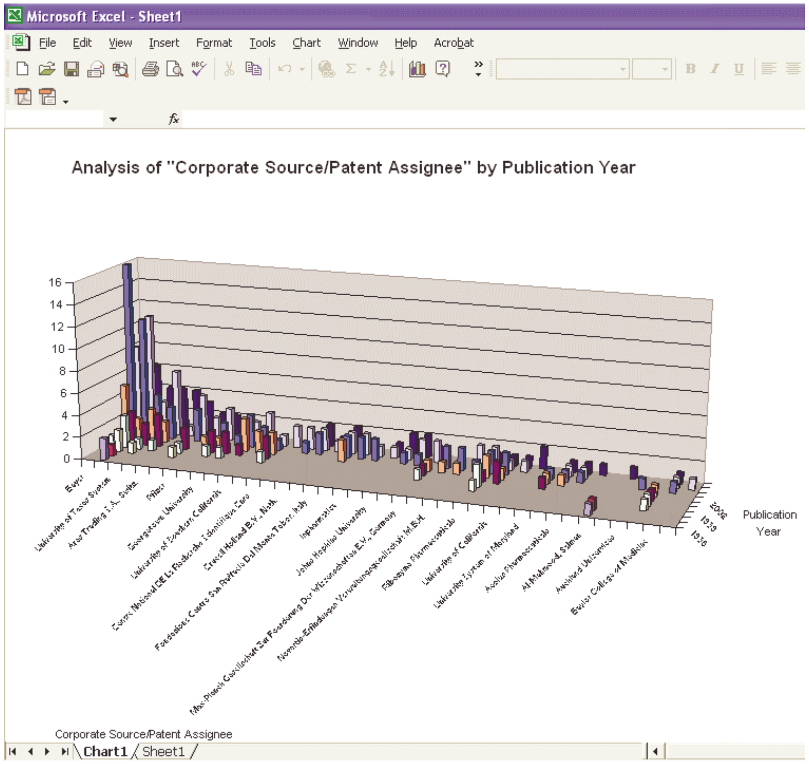
STN Express Data Charts & Histograms

Analyze, VGAT

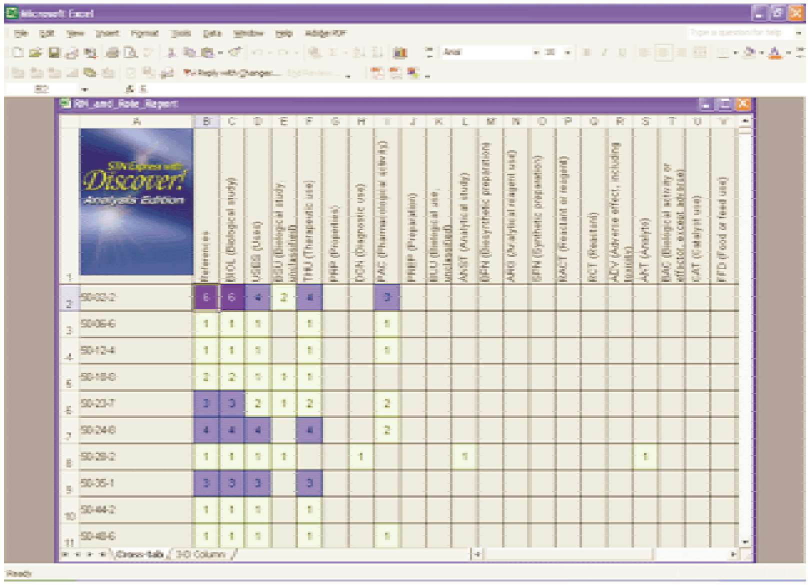
Analyze Plus

1998 2000 2001 2002 2003 2004 2005

Traditional data mining provides rapid analysis of structured data



- Multiple sources analyzed together
- No need to download large files
- Good for statistical analysis



	References	BIOL (Biological study)	USES (Uses)	BSU (Biological study unclassified)	THU (Therapeutic use)	PRP (Properties)	DGN (Diagnostic use)	PAC (Pharmacological activity)	PREP (Preparation)	BUU (Biological use, unclassified)	ANST (Analytical study)	BPN (Biosynthetic preparation)	ARG (Analytical reagent use)	SPN (Synthetic preparation)	RACT (Reactant or reagent)	RCT (Reactant)	ADV (Adverse effect, including toxicity)	ANT (Analyte)	BAC (Biological activity or effector, except adverse)	CAT (Catalyst use)	FFD (Food or feed use)
50-02-2	6	6	4	2	4			3													
50-06-6	1	1	1		1			1													
50-12-4	1	1	1		1			1													
50-18-0	2	2	1	1	1																
50-23-7	3	3	2	1	2			2													
50-24-8	4	4	4		4			2													
50-28-2	1	1	1	1			1				1							1			
50-35-1	3	3	3		3																
50-44-2	1	1	1		1																
50-48-6	1	1	1		1			1													

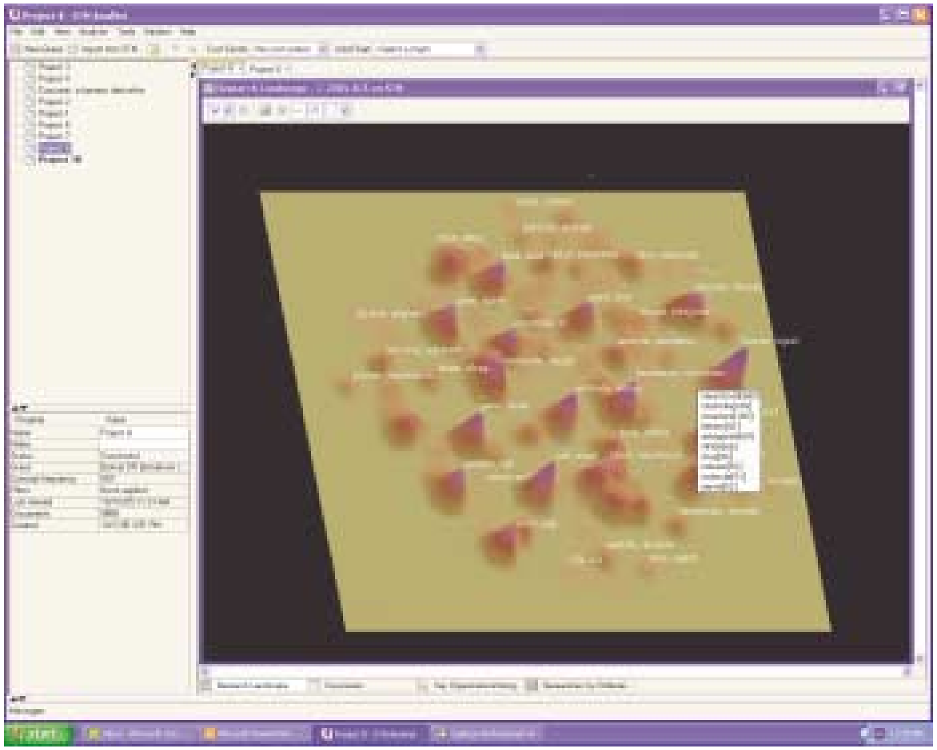
cas®

FIZ KARLSRUHE

Some structured data provide insights into content but can be one-dimensional

- An indexing term may be applied for a concept which is not the main focus of the document
- An indexing term may not be applied for a concept which is still significant within a document but is not new

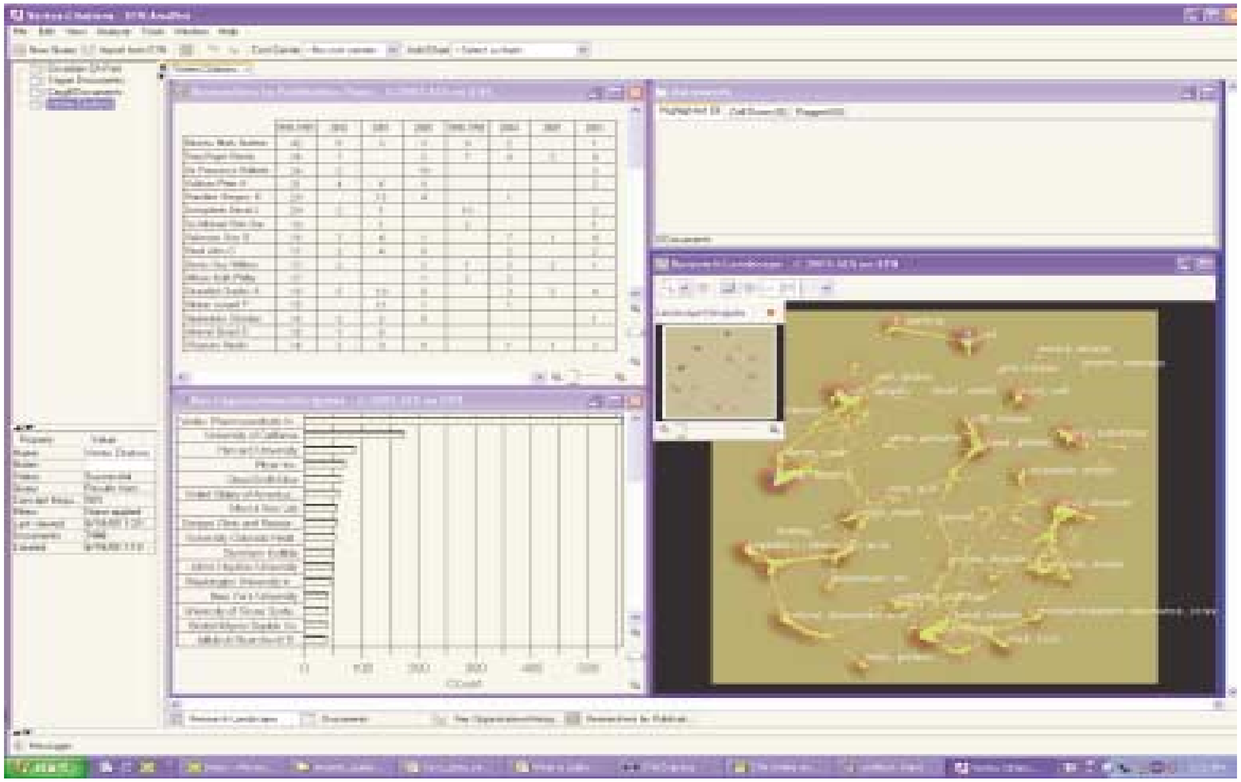
Text mining of unstructured data can offer deeper insight into content



- Allows for mixture of data types
- Shows data relationships
- Provides insight for broad questions

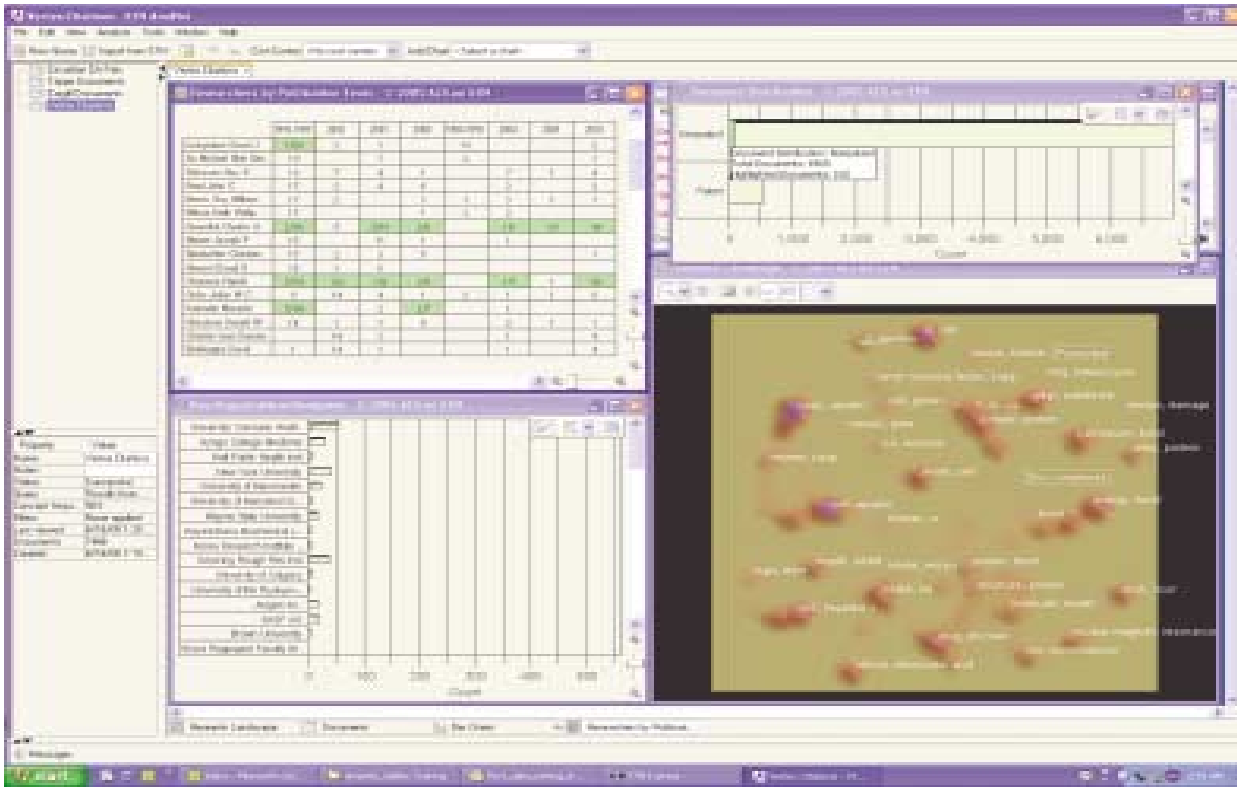
Now there is the ability to combine data and text mining in an A&V tool

STN AnaVist presents data and text mining results upon visualization



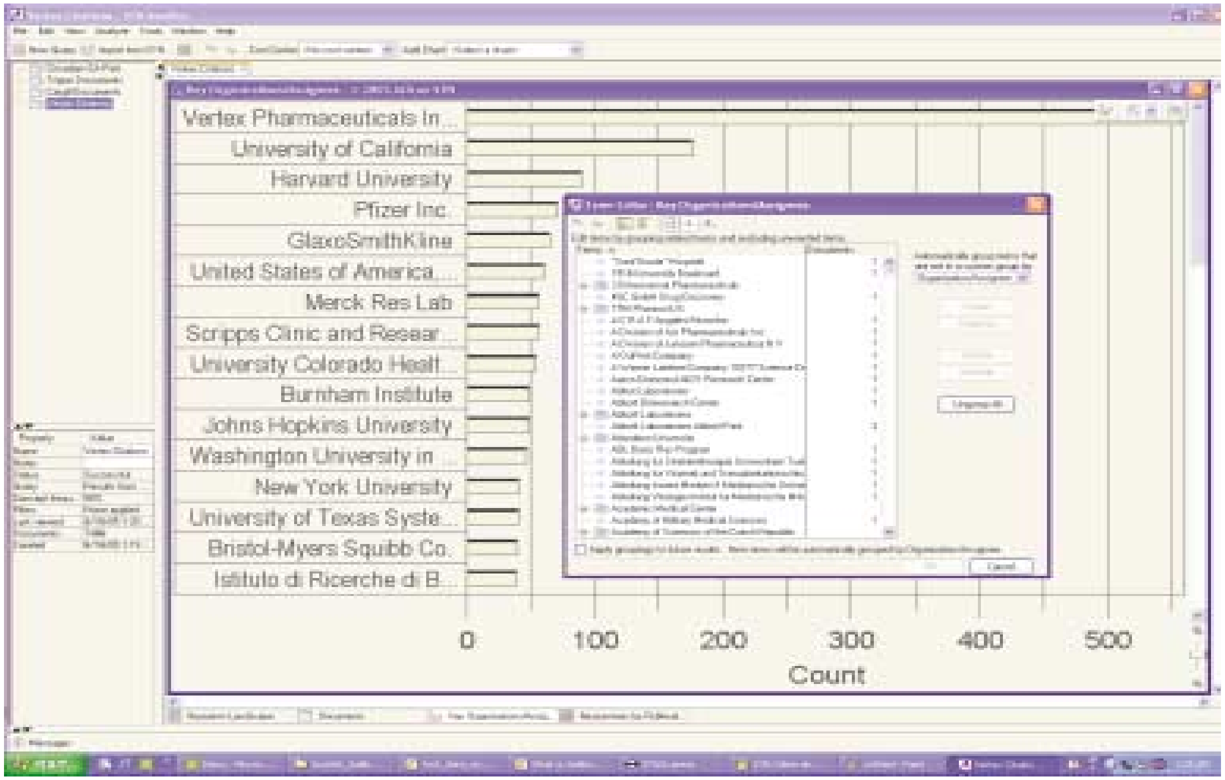
- Analyze multi-database content—patent and non-patent literature
- Visualize research landscape, key organizations, and key researchers

Links among data and charts uncover relationships



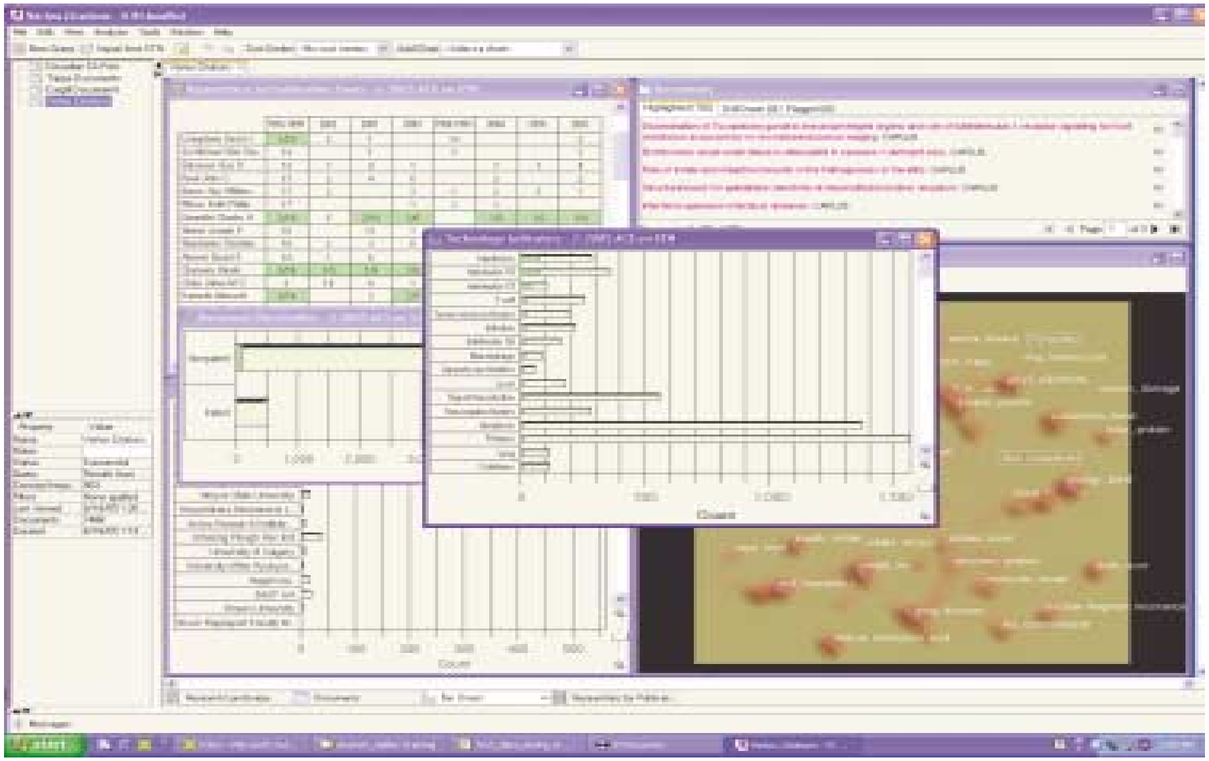
- Interactive charts and maps highlight data relationships
- Related documents are sent to document window for viewing

Data grouping and clean-up are available to produce more meaningful results


Vertex Pharmaceuticals In...
University of California
Harvard University
Pfizer Inc.
GlaxoSmithKline
United States of America...
Merck Res Lab
Scripps Clinic and Resear...
University Colorado Healt...
Burnham Institute
Johns Hopkins University
Washington University in ...
New York University
University of Texas Syste...
Bristol-Myers Squibb Co.
Istituto di Ricerche di B...
0
100
200
300
400
500
Count

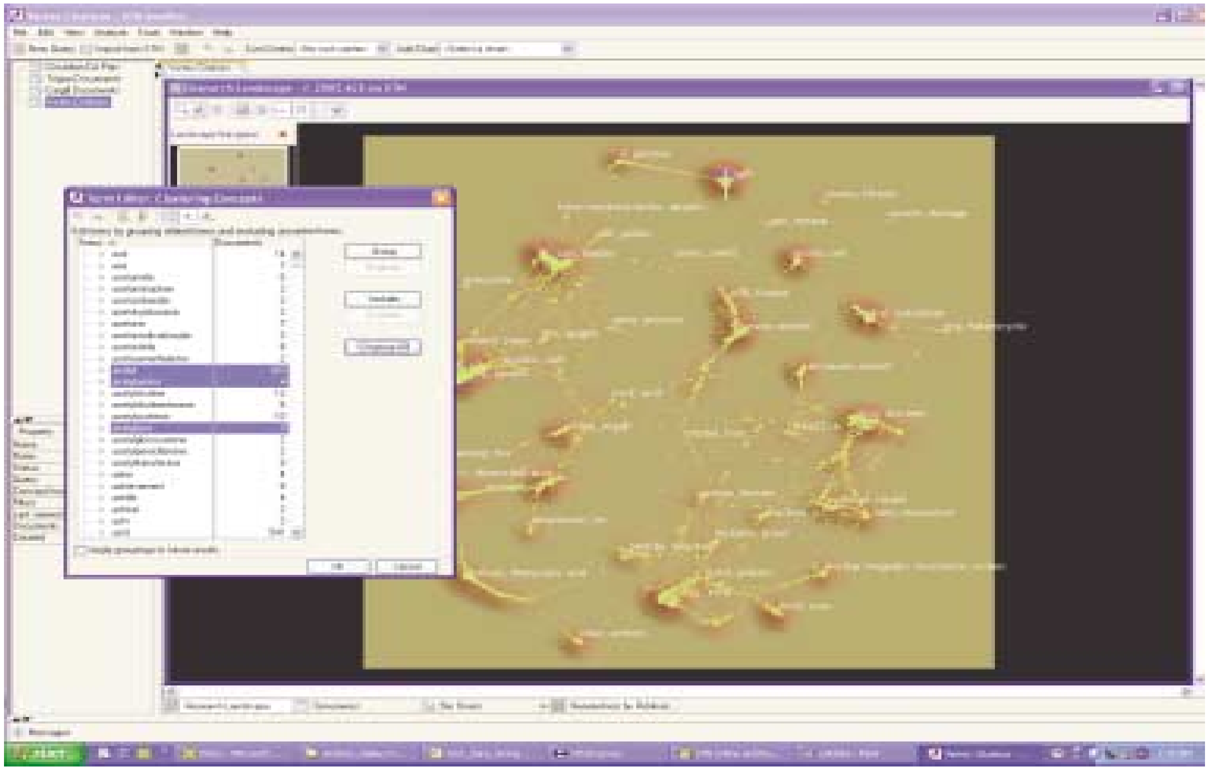
- CAS company name thesaurus
- Save customizations and use in future visualizations

Document-centric research landscape provides perspective that individual data elements do not



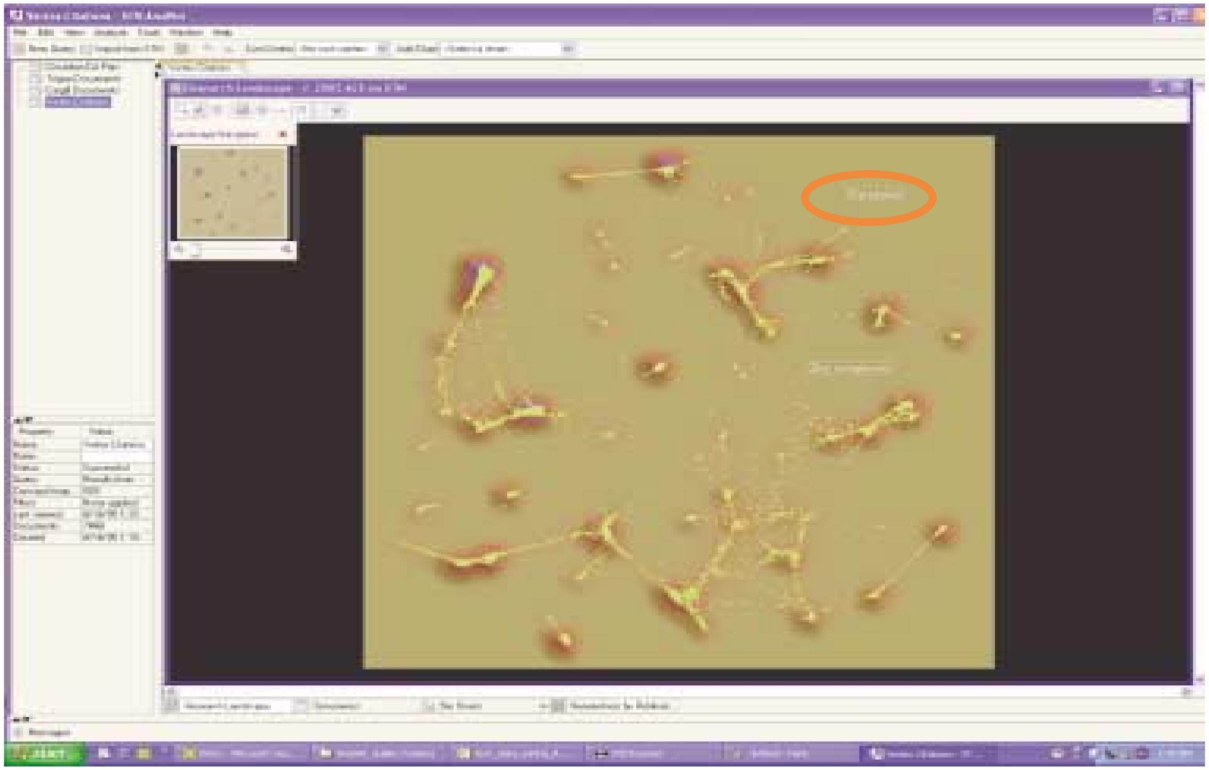
- Concept standardization across data collections based on CAS vocabulary

Excluding non-content bearing words and re-visualizing can improve map results



- Clustering concepts can be edited

Adding user-preferred jargon and annotations helps clarify the cluster map



- Change label to add custom jargon and more illustrative descriptions

Conclusions

- Today's environment requires software that helps users intellectually process large amounts of information
- A&V tools provide a visual representation of data that can uncover new insights
- Quality data combined with sophisticated software provide the best results