Technology Mining for Technology Intelligence: Mastering Big Data

Irina Efimenko

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

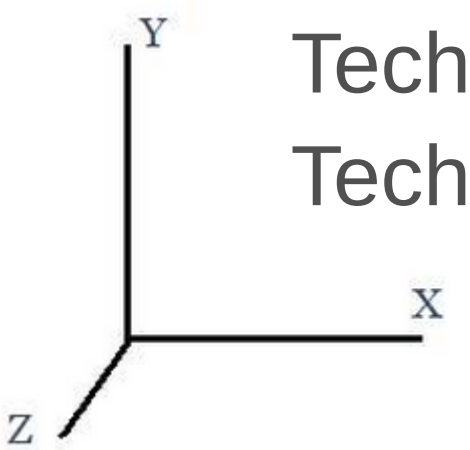
Tech Mining and Technology Intelligence

Technology intelligence indicates the concept and applications that transform data hidden in patents or scientific literature into technical insight for technology strategy-making support. The existing frameworks and applications of technology intelligence mainly focus on obtaining text-based knowledge with text mining components.

A patent time series processing component for technology intelligence by trend identification functionality. H. Chen et al. Neural Computing and Applications, 2014

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

What is Tech Mining?



Technology Mining =
Technology + Mining

Technology: (task / target / application) + input data

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Technology Dimension – I

Future-oriented technology analysis (FTA)
Foresight
Roadmaps
Technology Management
Innovation policy …

Tech Mining?

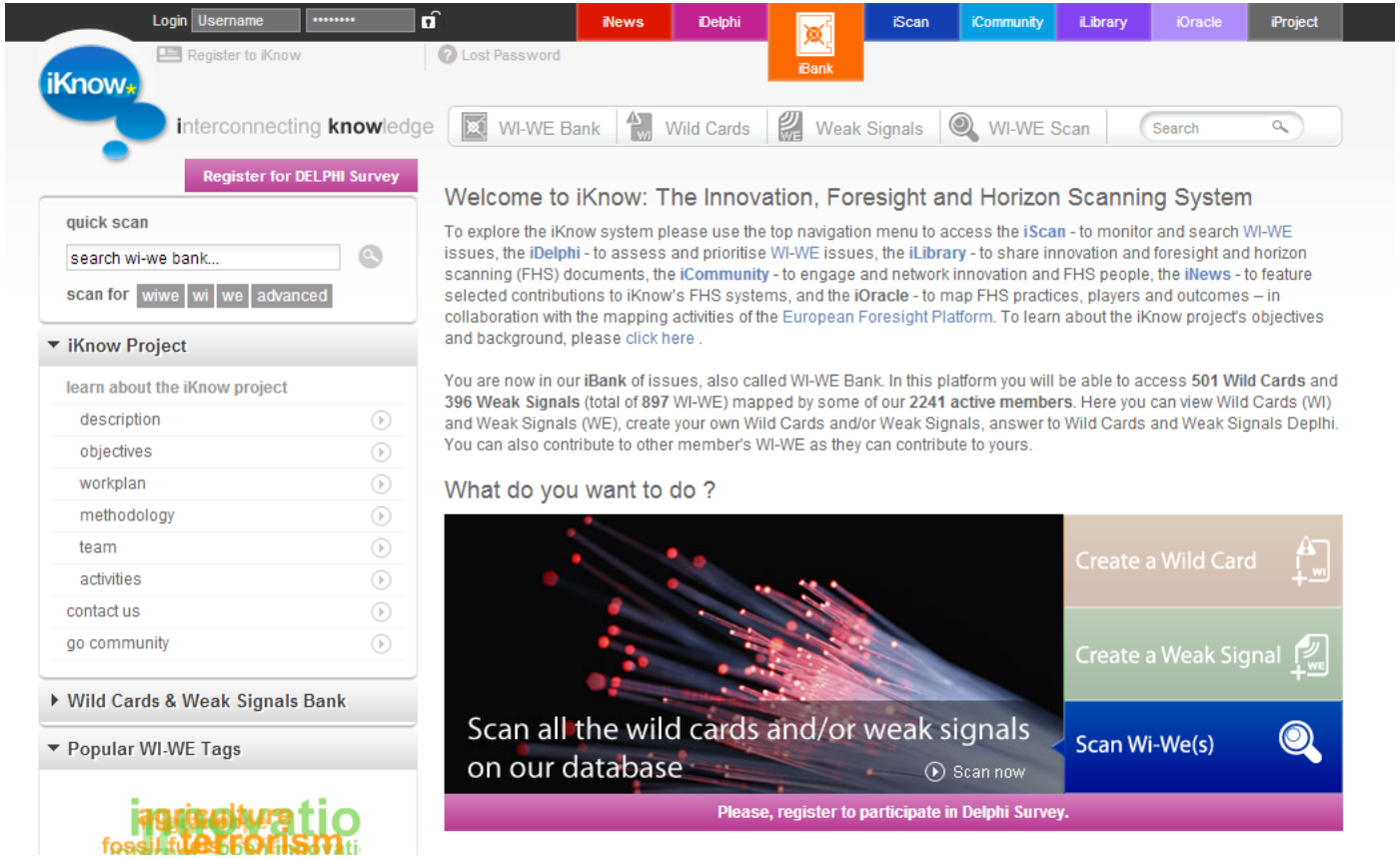
Not necessarily
(“Mining” is absent)

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Technology Dimension – II

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Technology Dimension – III



Futures Diamond Company. iKnow and other projects. Crowdsourcing, etc.

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Technology Dimension – IV

Wikivote! company

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Mining Dimension – I

NLP, statistical data, etc.
Sentiment analysis
Named Entity Recognition (NER)
Fact Extraction
Question Answering (QA)
Clustering

Tech Mining?

Not necessarily
(“Tech” is absent)

Mining Dimension – II

Even for "scientific" data it's not always the case (data + task)

Data – I

Demands of **Format** + **Content**

*"Tech Mining, a special form of "Big Data" analytics, aims to generate **Competitive Technical Intelligence (CTI)** using bibliometric and text-mining software as well as other analytical & visualization applications for analyses of Science, Technology & Innovation (**ST&I**) information resources…"*

GTM2015 Conference

- ✓ Research papers, patents, web scraping (technology news, foresight reports, etc.)
- ✓ Novel data sources (e.g. roadmaps)

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Data – II

Table 1. Six Information Types

Medium \ Message	**1) Technology**	**2) Context**
A) Databases	Research funding, publication & patent abstracts, citations	Business, market, policy, popular opinion
B) Internet	Technical content sites	Company sites, blogs, etc.
C) People	Technical experts	Business experts

https://www.thevantagepoint.com/

Tasks

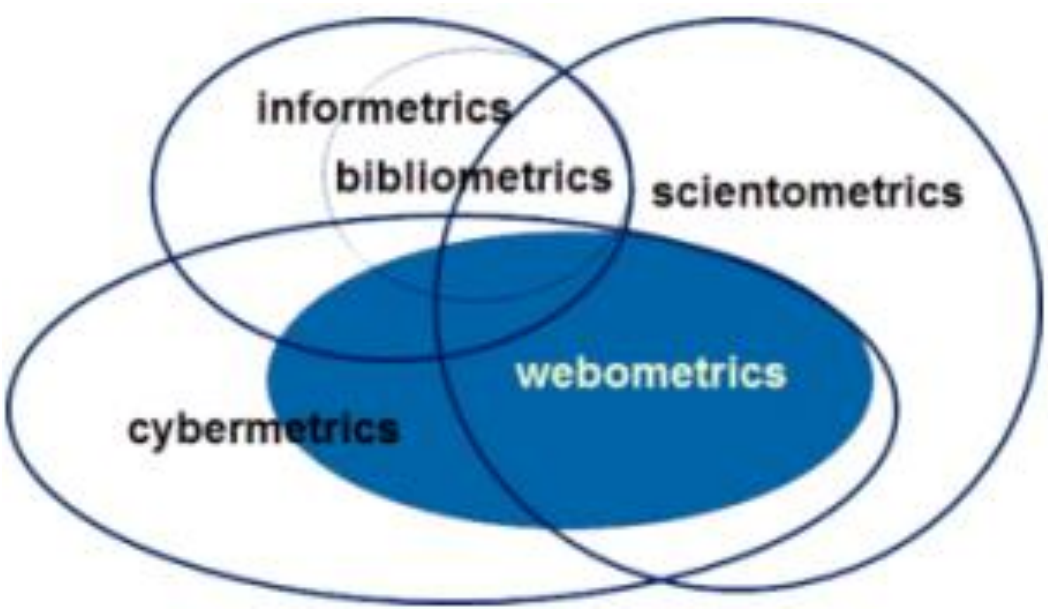
Demands of **Business Task** + **IT Task**

*"Tech Mining, a special form of "Big Data" analytics, aims to generate **Competitive Technical Intelligence (CTI)** using bibliometric and text-mining software as well as other analytical & visualization applications for analyses of Science, Technology & Innovation (**ST&I**) information resources…"*

- ✓ IT: Extraction, BI, Visualization
- ✓ BT: basically, any STI task (however, one can name typical ones)

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Nature of Data: Basics – I



Nature of Data: Basics – II

Webometrics also called **cybermetrics** is a relatively new information discipline that aims to quantitatively measure web phenomena. By web phenomena we mean the number of hyperlinks on a web site, their structure and patterns of use by web users, and their reciprocal links by other websites.

Generally speaking, this data can be mined by search engine websites although several experts in webometrics recommend the use of *data cleansing* heuristics to ensure their reliability.

http://blogs.ubc.ca/

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Nature of Data: Basics – III

Bibliometrics and **Scientometrics** are two closely-related fields that aim to measure scientific publications and science in general. A lot of the research that falls under this topic involves citation analysis, or examining how scholars cite one another in publications.

Author citation data can show a lot about scholar networks and scholarly communication, linkages between scholars, and the development of areas of knowledge over time.

http://blogs.ubc.ca/

Vasily Nalimov

Derek J de Solla Price

Eugene Garfield

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Nature of Data: Basics – IV

Altmetrics (*alternate or alternative metrics*) is a field of web-based metrics that accounts for total author influence. Quantifying this scholarly activity is new and differs from citation metrics. *Altmetrics*, or alternative citation metrics, provides researchers and scholars with new ways to track influence for emerging scholarly communication on the web (number of links, downloads, etc.).

http://blogs.ubc.ca/

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Scientometrics – I

Наукометрия — дисциплина, изучающая эволюцию науки через многочисленные измерения и статистическую обработку научной информации (количество научных статей, опубликованных в данный период времени, цитируемость и т. д.).

Scientometrics is the study of measuring and analyzing science, **technology** and **innovation**. Major research issues include the measurement of impact, reference sets of articles to investigate the impact of journals and institutes, understanding of scientific citations, mapping scientific fields and the production of indicators for use in policy and management contexts. *Wikipedia*

Science → *Technology*

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Not only h-index

THOMSON REUTERS
Web of Science

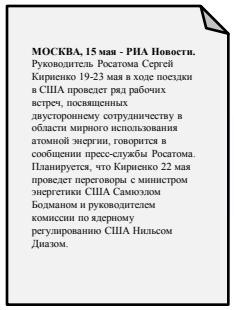
Scopus

Scientometrics – II

- ✓ *Initially: efficiency of X*
- ✓ *For TM (TI): any corresponding task. Mostly for fields and partnering teams*

Tech Mining?
(e.g. opinion mining vs. centers of excellence[1])

Text **Model** **KB**



типы объектов и
типы отношений

РАБОТАТЬ В ОРГАНИЗАЦИИ

ОРГАНИЗАЦИЯ

ПЕРСОНА

Руководитель Росатома Сергей Кириенко 19-23 мая в ходе поездки в США проведет ряд встреч...

a team, a shared facility or an entity that provides leadership, evangelization, best practices, research, support and/or training for a focus area

Beware the Competition!

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Tech Mining?

- ✓ *Monitoring research papers in Medicine for DSS in Diagnostics? No.*
- ✓ *Monitoring research papers in Medicine for early detection of promising treatment methods? Then Yes!*

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Challenge to Scientometrics

19th International Conference on Science and Technology Indicators

"Context Counts: Pathways to Master Big and Little Data"

3 - 5 September 2014

Leiden, The Netherlands

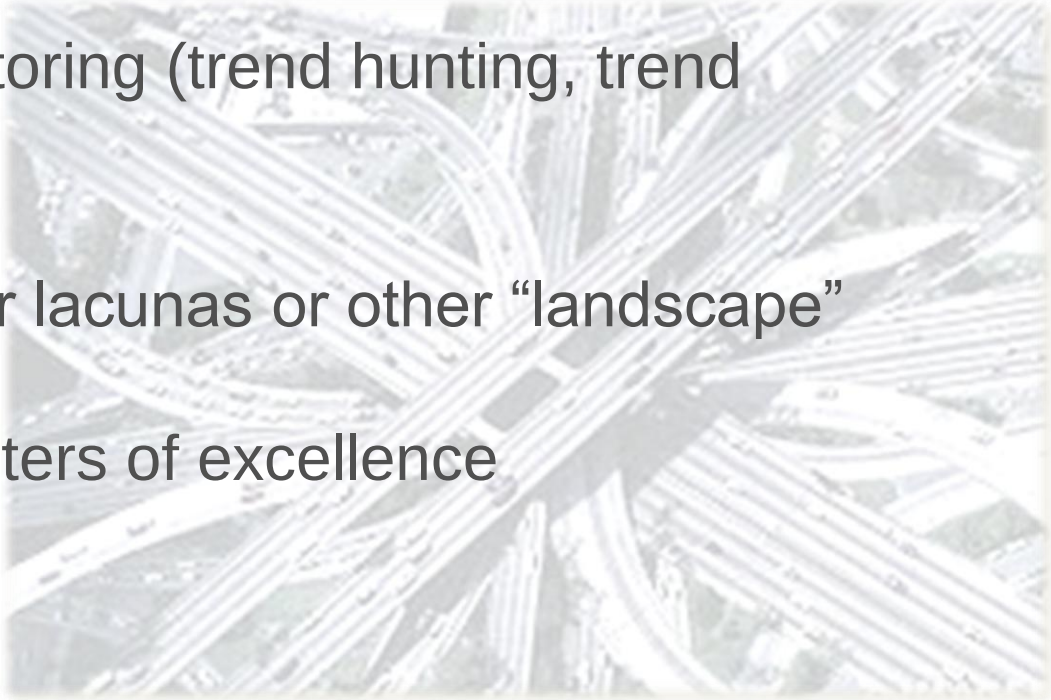
5th Annual

Global TechMining Conference

2015.09.16 • Atlanta, Georgia

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

“Standard” tasks – I

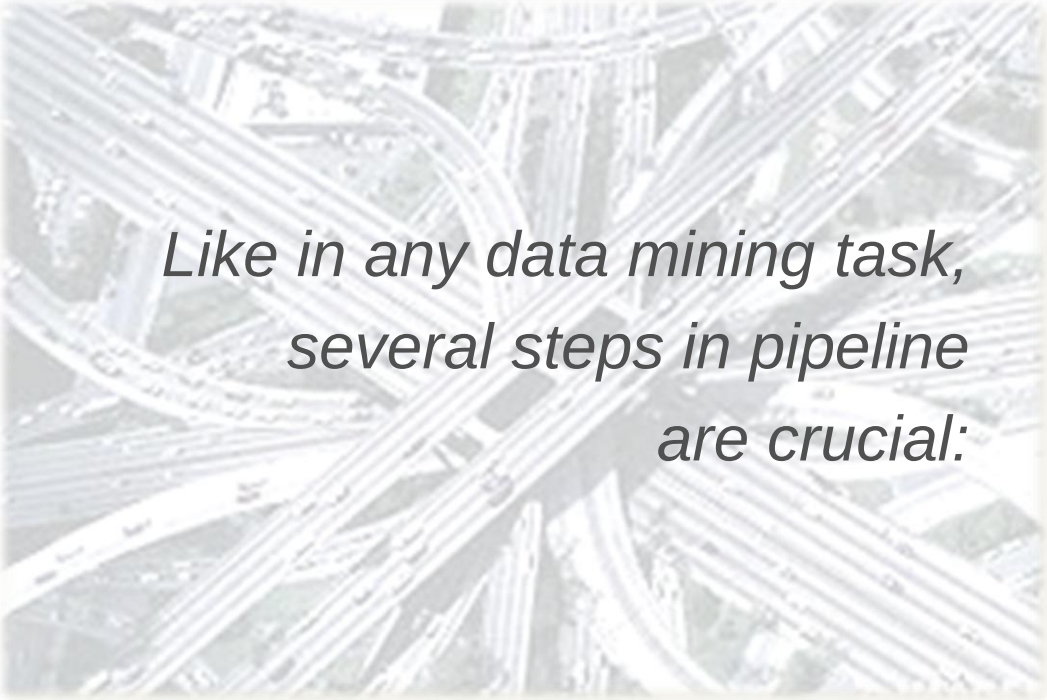


- ✓ Technology trend monitoring (trend hunting, trend watch…)
- ✓ Maps of S&T
- ✓ Patent analysis (e.g. for lacunas or other “landscape” tasks)
- ✓ Partnering teams / Centers of excellence

  - – Bibliometrics
  - – Text mining (does not necessarily mean *Semantic Analysis*). Clustering techniques, etc.
  - – Newcomer: Semantics

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

“Standard” tasks – II



“Extraction”

→

Visualization

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

“Standard” tasks – III

Here’s a 10-step approach to Tech Mining.

1. Spell out the focal intelligence questions and decide how to answer them
2. Get suitable data
3. Search (iterate) & retrieve ~abstract records
4. Import into text mining software [e.g., Thomson Data Analyzer (TDA), VantagePoint]
5. Clean and consolidate the data - Clumping
6. Analyze
7. Visualize
8. Integrate with internet analyses and expert opinion
9. Interpret and summarize findings; communicate those (possibly multiple ways)
10. Standardize and semi-automate where possible

https://www.thevantagepoint.com/

Hunting for an automated mode!

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

“Standard” tasks – IV

Specific tasks, e.g. clumping (is it enough?):

We have devised a topical Term Clumping process (Zhang et al., 2014a) to:

a) consolidate related terms via general purpose and special thesauri
b) clean up terminology (e.g., to combine singular and plural forms of a terms, and to overcome small name discrepancies) via fuzzy matching routines
c) identify more informative terms or remove common terms (depending on the data set) via Term Frequency Inverse Document Frequency (TFIDF) analysis
d) consolidate highly related terms via correlation rules (e.g., co-occurrence of terms in many records indicates possible relationship)
e) cluster topics for further analyses via term clustering approaches (e.g., Principal Components Analysis, Latent Dirichlet Allocation, K-Means, etc.).

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Technology trend monitoring

1) Terms Extraction. Not just n-grams (but meaningful terms)
2) Escape from getting too general or garbage results (“Semantic Technologies” or “Clean Energy”). Not just TF-IDF
3) Time Series for trend extrapolation and growth modeling. Time and Life Cycle Concept
4) Prominent solutions, Disruptive technologies

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Life Cycle Concept – I

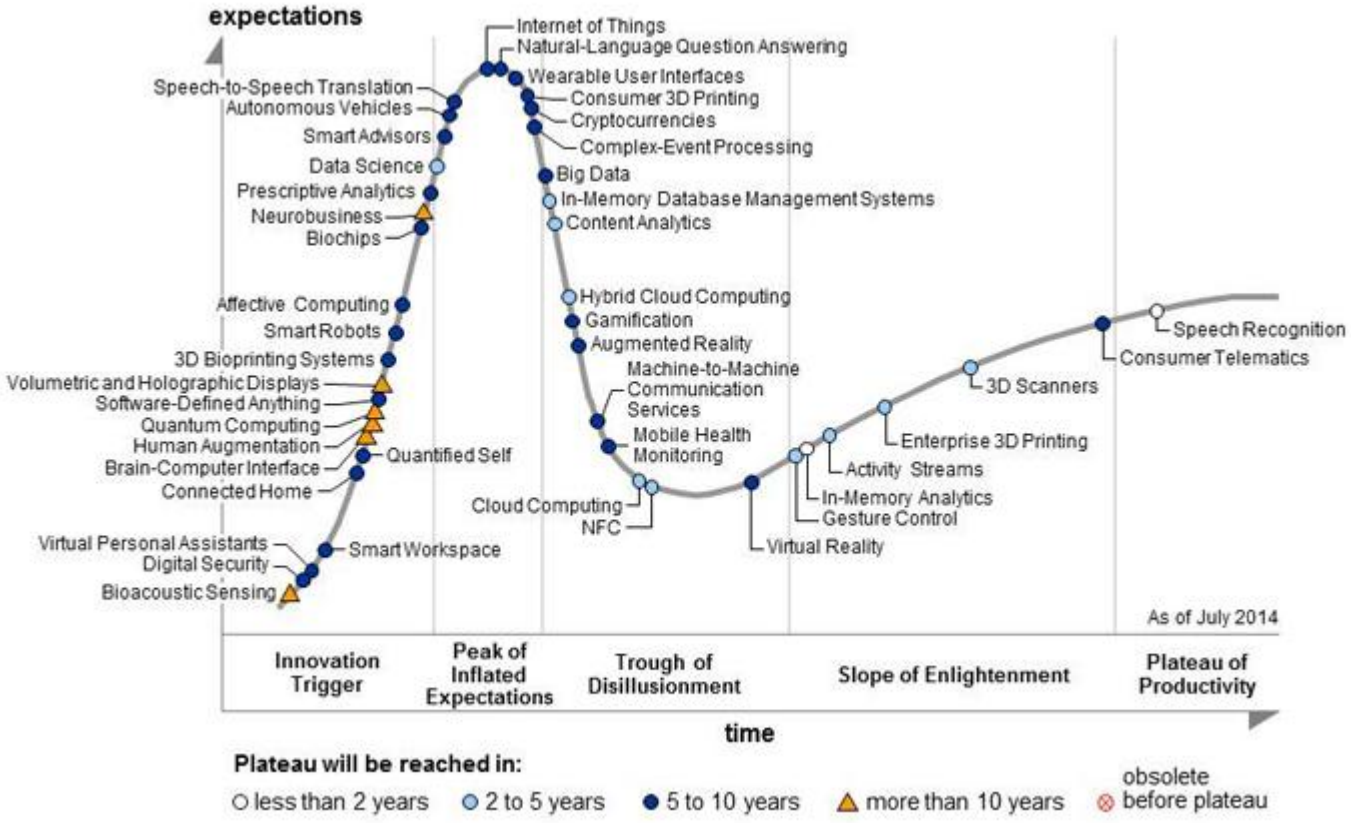
Data types

- Fundamental research [reflected by research project funding (e.g., NSF Award or NIH CRISP database) and publication abstract databases (e.g., Science Citation Index; PubMed)]
- Applied research & development [reflected, for instance, by engineering oriented publication abstract databases (INSPEC, EI Compendex)]
- Invention [consider patent databases such as Derwent World Patent Index and PatStat (from the European Patent Office for academic use]
- Commercial Application [e.g., new products databases, Business Index, marketing data sources]
- Broader Contextual Factors and Implications [e.g., Lexis-Nexis, Factiva, Congressional Record]

TECH MINING of Science & Technology Information Resources for Future-oriented Technology Analyses.
A. Porter, Y. Zhang, 2015

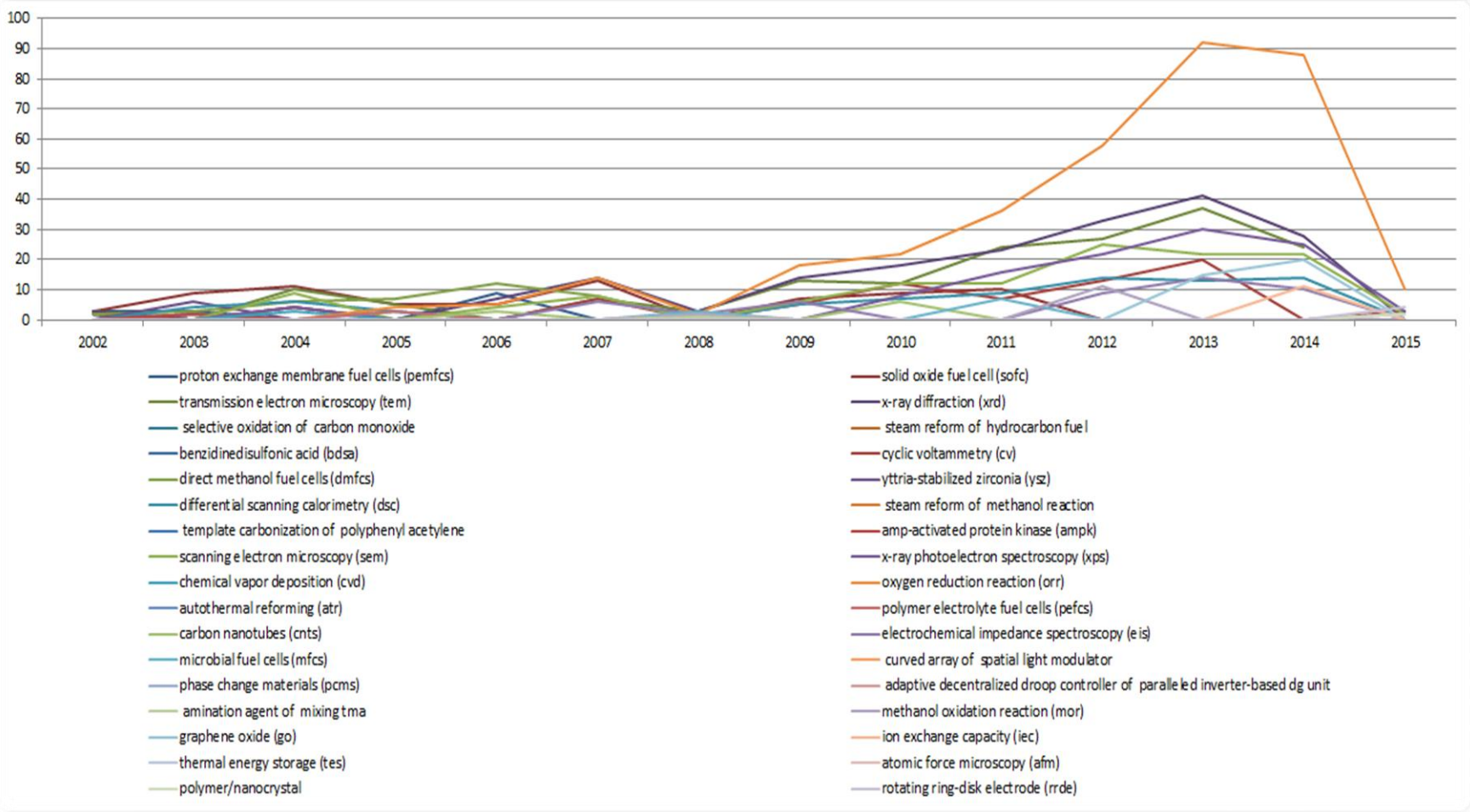
Life Cycle Concept – II

Gartner's Hype Cycle (http://www.gartner.com/)



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Emerging Technologies in Energy



I. Efimenko, V. Khoroshevsky (in press)

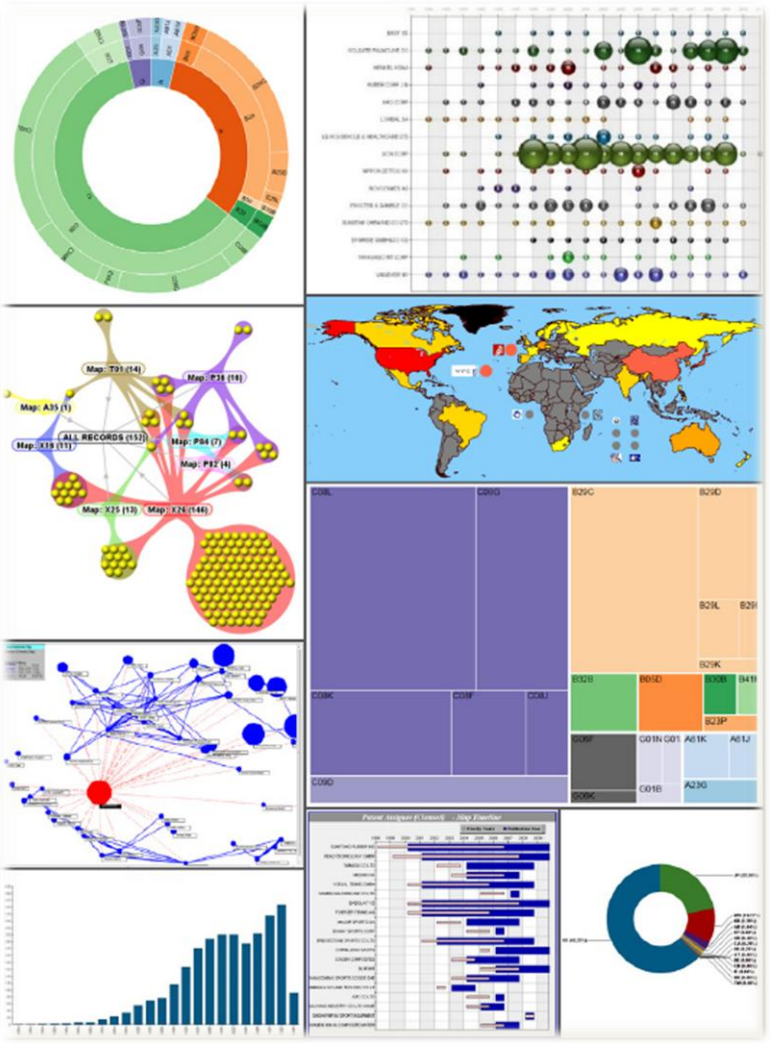
Vantage Point

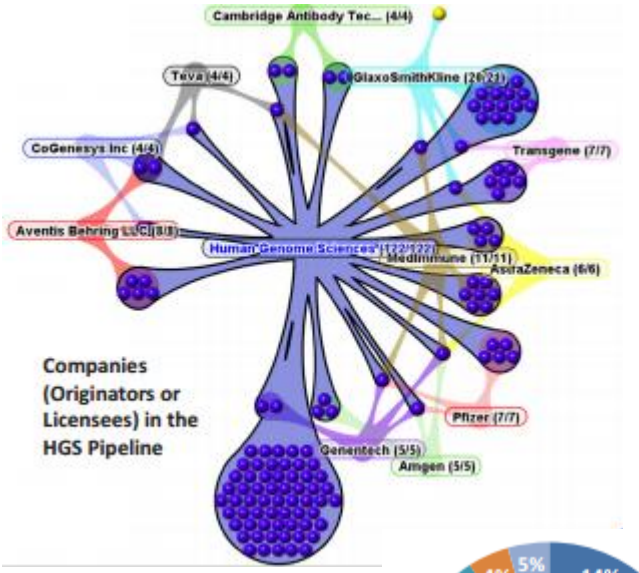
“VantagePoint is a powerful text-mining tool for discovering knowledge in search results from patent and literature databases. VantagePoint helps you rapidly understand and navigate through large search results, giving you a better perspective—a better vantage point—on your information. The perspective provided by VantagePoint enables you to quickly find WHO, WHAT, WHEN and WHERE, helping you clarify relationships and find critical patterns”

- ✓ Semantics - ?
- ✓ “Tech mining evangelists”

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

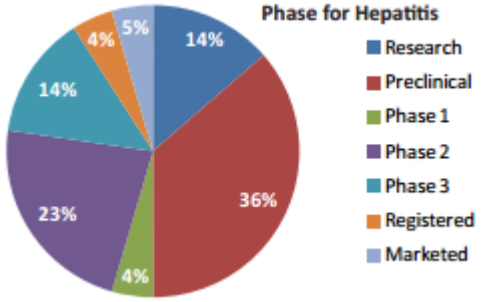
VP Examples: not (just) trends





Companies

Landscape



TM and Related tasks:
BizInt Smart Charts for Patents, BizInt Smart Charts for Drug Pipelines, BizInt Smart Charts Reference Rows

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Maps of S&T – I

Founding fathers: **SciTech Stategies, Inc.**

Bibliometrics and other advanced techniques

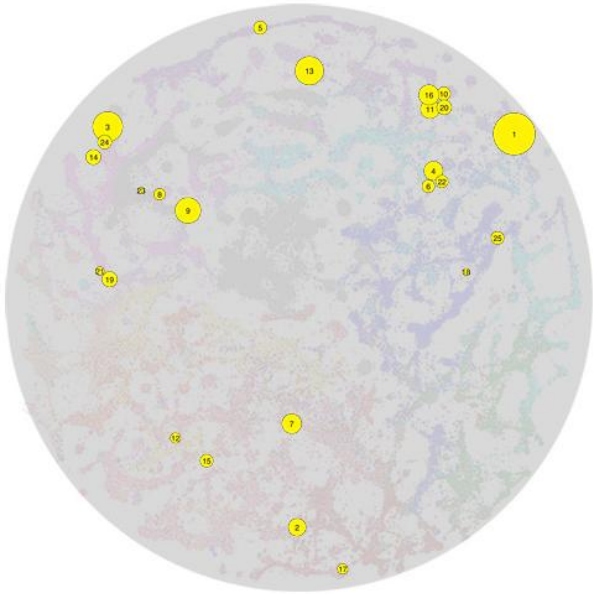
Business value

Emergence =

Novelty (or newness) + Growth

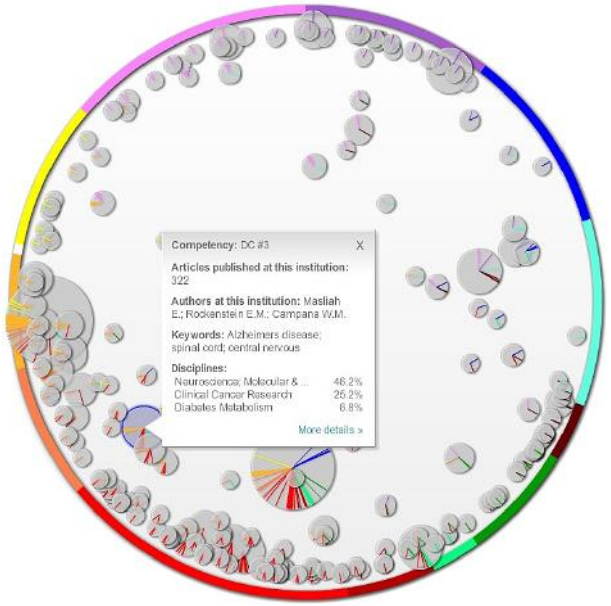
Maps of S&T – II

Dynamic maps for
Emerging Opportunities



Weak signals,
early detection, etc.

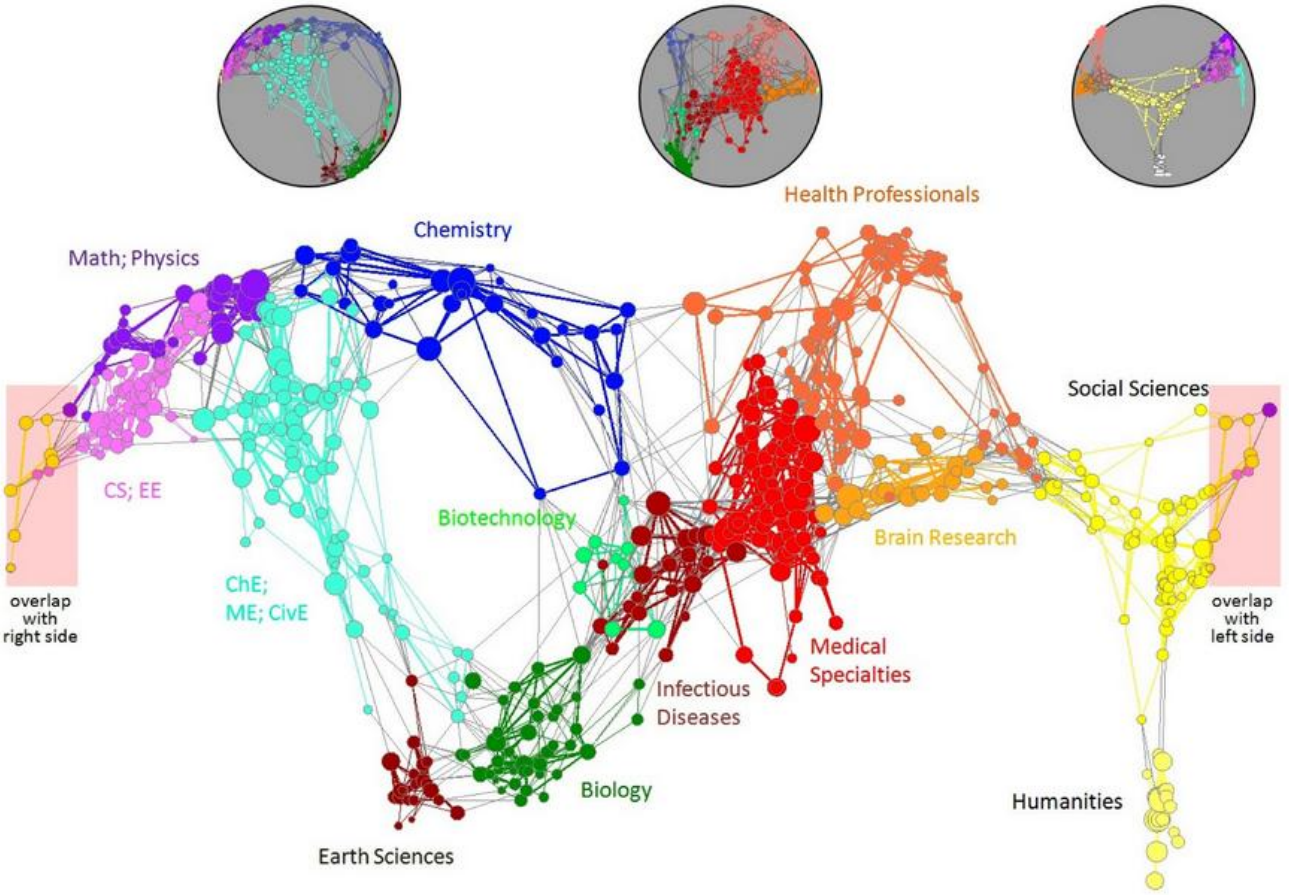
Winning teams



Research Strengths

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

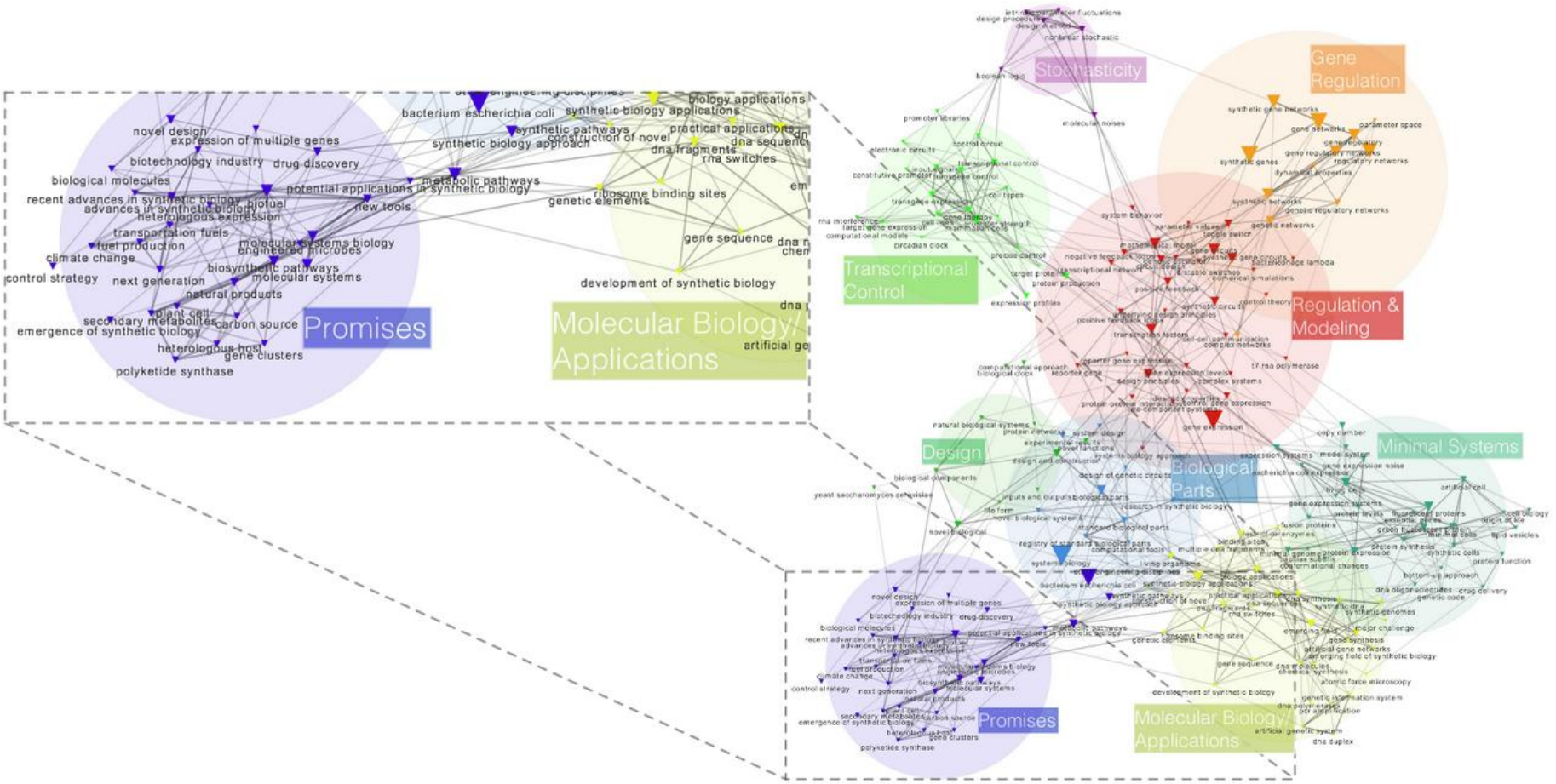
Maps of S&T – III



http://www.soic.indiana.edu/news/story.html?story=Global-Maps-Science

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Maps of S&T – IV



http://digitalmethods-seminar.org/wp-content/uploads/2014/03/2014_SyntheticBio.png

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

New Algorithms (Examples) – I

Example from *Identifying emerging topics in science and technology. H. Small, K. Boyack, R. Klavans. Research Policy, 2014*:

Co-citation model: multi-step process

1) Clusters of cited papers are created for each separate year within the citation database.

2) Current papers from the annual slice are assigned to the clusters of cited references based on their bibliographies.

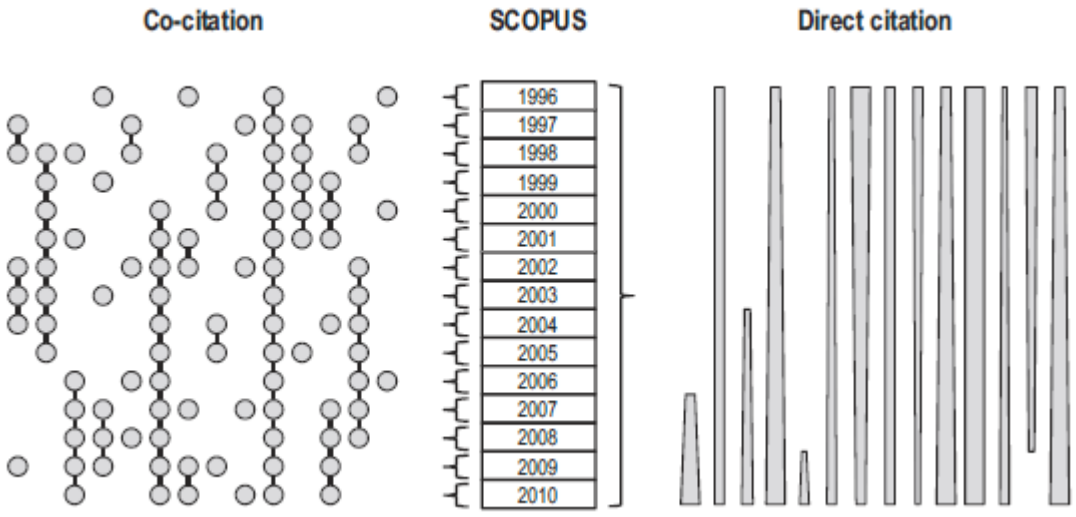
Each cluster thus consists of papers and the group of common cited references that most informed the current work.

3) Finally, clusters from adjacent years are linked using shared reference papers into cluster strings (called threads), which turns a series of static views of the structure of science into a dynamic view.

New Algorithms (Examples) – II

Direct citation model:

Citation links between articles are used to create clusters of articles using the full set of Scopus articles in a single clustering process. The algorithm uses a variant of modularity-based clustering, which attempts to maximize the ratio of links within clusters to links between clusters. To account for the different linkage degrees of different papers (outlinks and inlinks, or citing and cited links), each link is normalized by the number of references in the citing paper…



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

New Algorithms (Examples) – III

Two models are combined to select emergent clusters. The direct citation model is the primary source of information, and the co-citation model is used to augment (or modify) that information.

Emergence Potential (EP) function:

A direct citation cluster is rewarded for having articles in a specific year that are also “new” in the co-citation model.

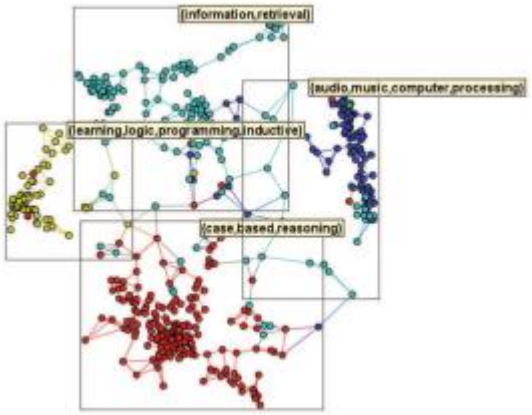
A direct citation cluster is penalized in a specific year for having articles in prior years. With this approach very new direct citation clusters with high growth rates whose papers are also in new co-citation threads are the most highly ranked, and are nominated as the most emergent.

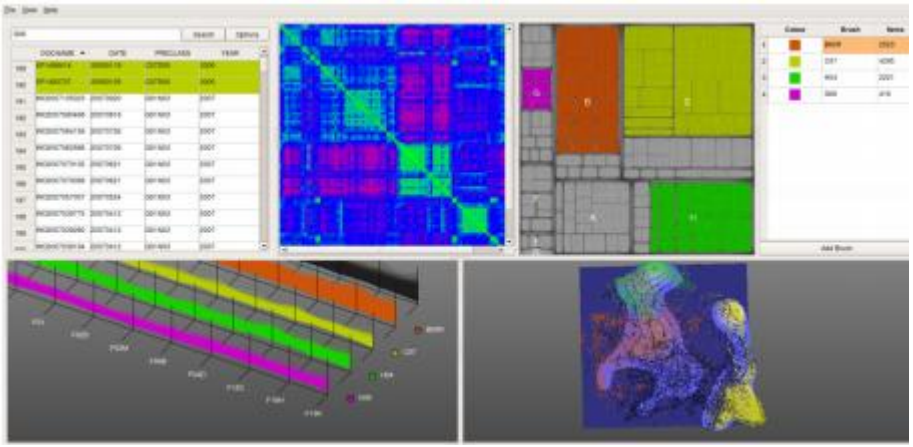
Specifically, the approach used is to count the papers in each direct citation cluster that belong to new threads (one or two years old) in the co-citation model for a given year.

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Patent Analysis – I

Treparel.com





- ✓ Links between patents and papers (policy issues). Not only Tech Mining
- ✓ Large patent data sets (navigation)
- ✓ Top holders, etc.

Patent Analysis – II

Treemapping technique *(Wikipedia)*

Treemaps display hierarchical (tree-structured) data as a set of nested rectangles. Each branch of the tree is given a rectangle, which is then tiled with smaller rectangles representing sub-branches. A leaf node's rectangle has an area proportional to a specified dimension on the data. Often the leaf nodes are colored to show a separate dimension of the data.

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Patent Analysis – III

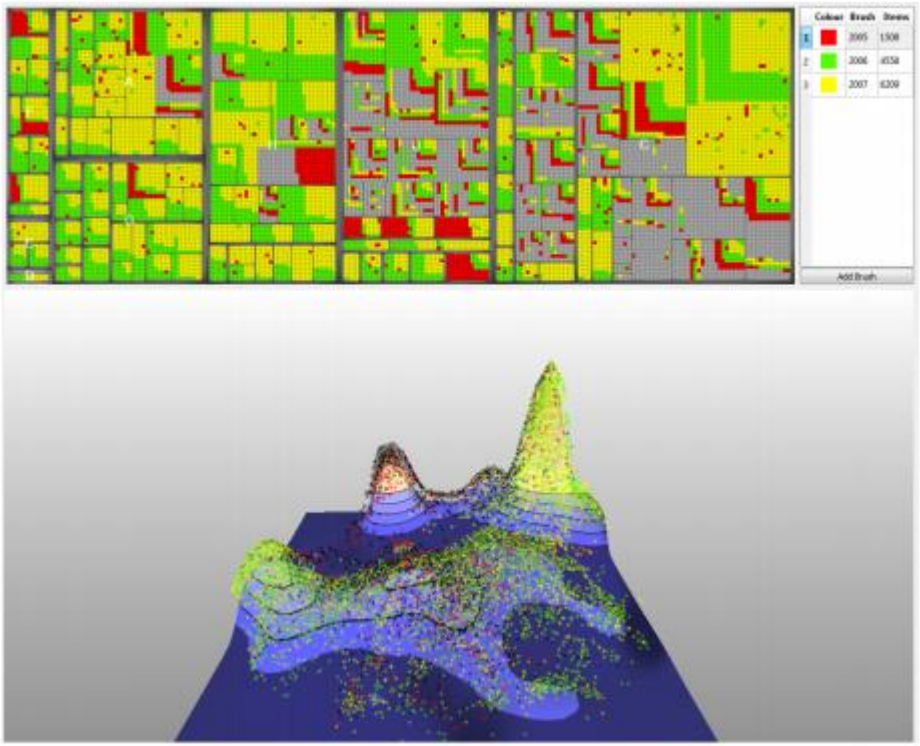
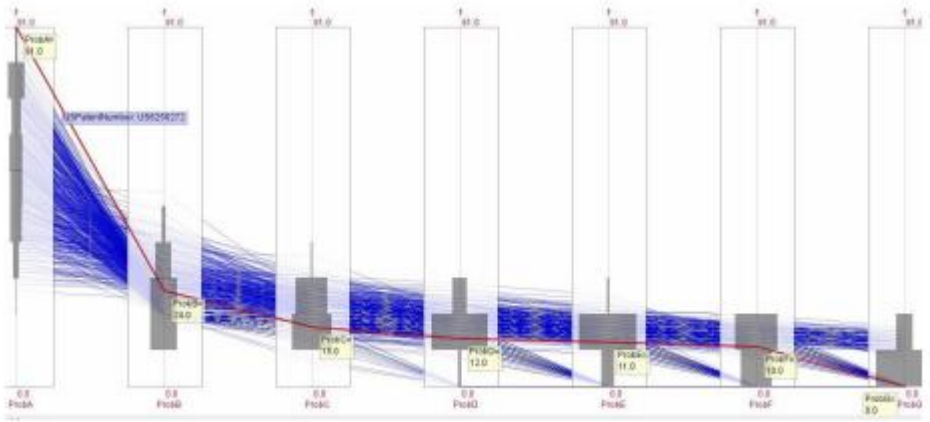


Figure 3: Combined use of two visualizations in KMX (tree map and clustering) to show the patent data hierarchically (tree map) and unsupervised (3D clustering where the height is the density of the patents especially prominent for the an-organic and organic chemistry) and the color is used to display the pattern in the patent data over time.

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Patent Analysis – IV

- ✓ Other tools and techniques
- ✓ Patents + Papers
- ✓ Patent classes analytics, e.g. *Trends of patents over classes*
- ✓ *Treatment techniques*



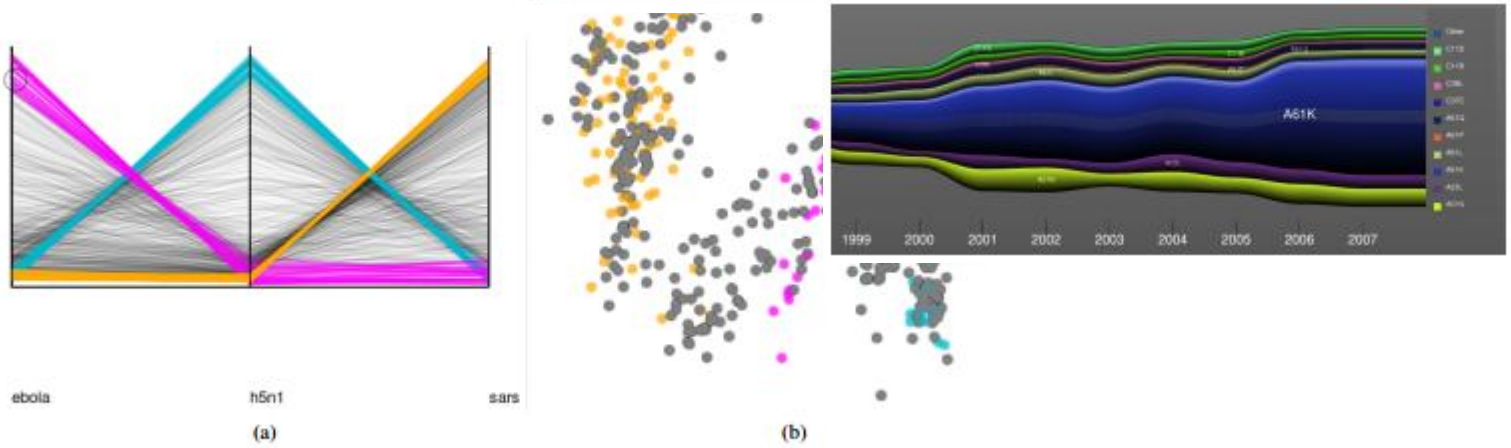


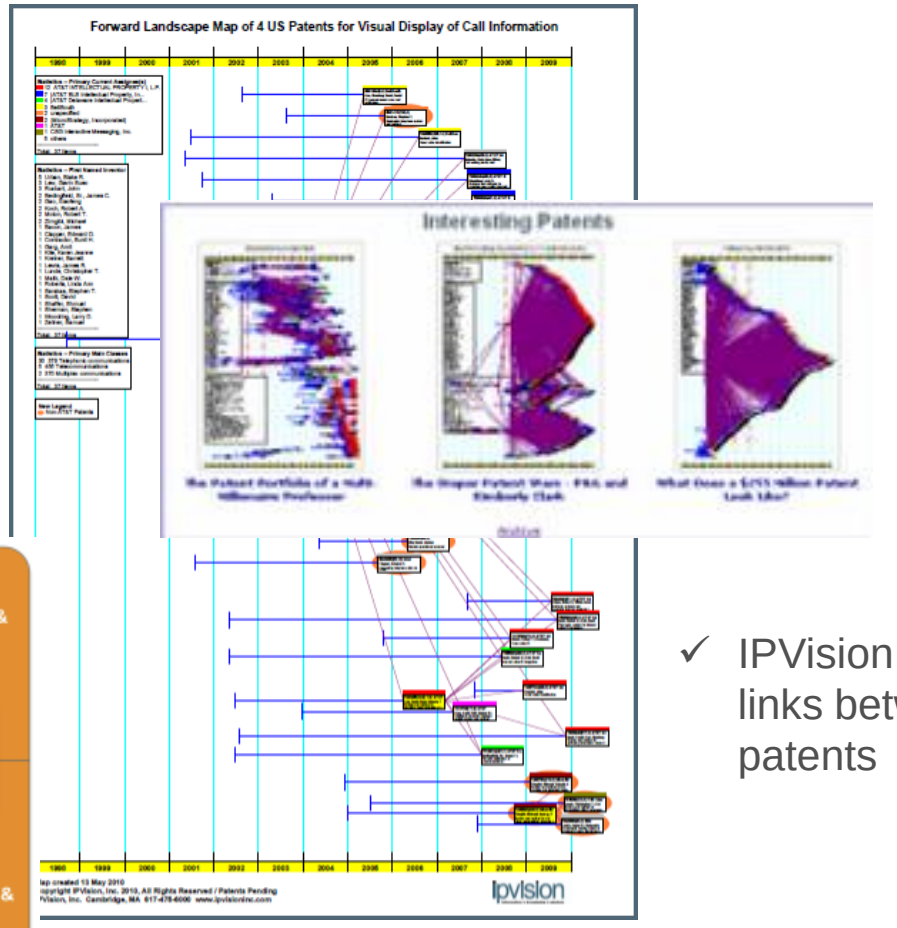
Figure 4: Parellel coordinates visualization and cluster visualization of 3 Medline clusters (Ebola (purple), H5N1 (blue) and SARS (yellow)).

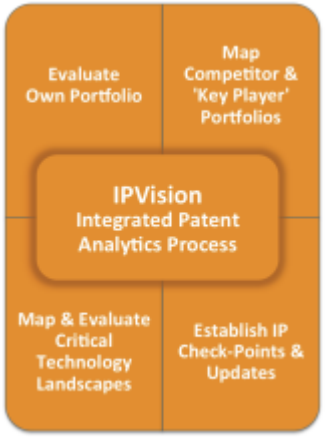
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Patent Analysis – V

IPVision's standard reports:

- ✓ IP Portfolio Evaluations and Assessments
- ✓ Patent and Technology Landscape Analysis
- ✓ Patent and Patent Portfolio Claims Analytics
- ✓ IP Due Diligence Reports
- ✓ IP Portfolio Chain of Title Analysis

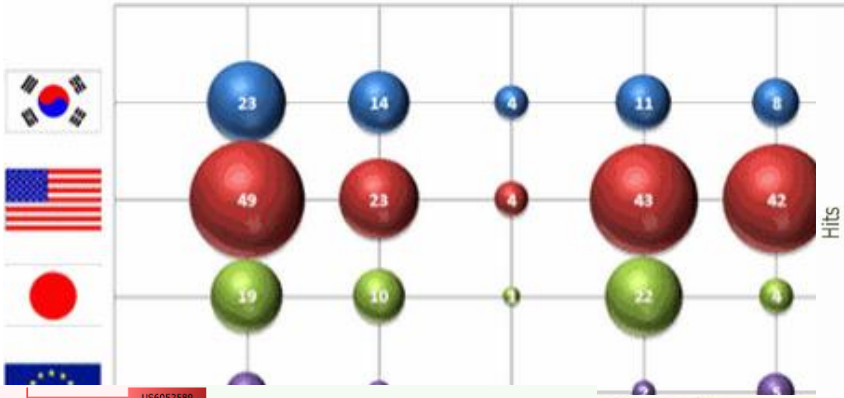




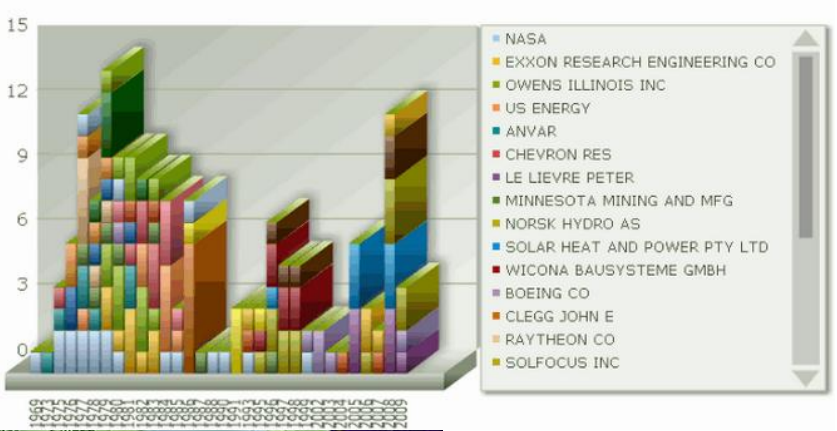
- ✓ IPVision: various links between patents

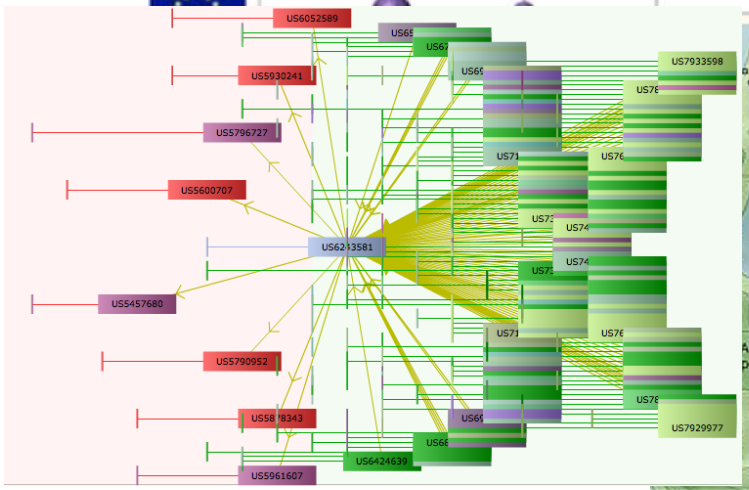
Patent Analysis: Much More

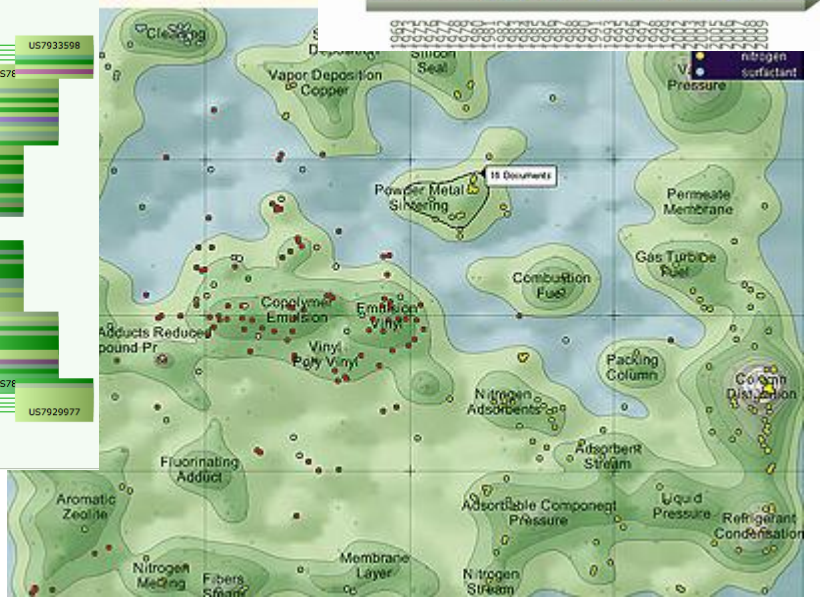
✓ Countries (Displaybank)



✓ Intellogist







✓ ThemeScape: text-mining module in Aureka

✓ Similar tools: WoS

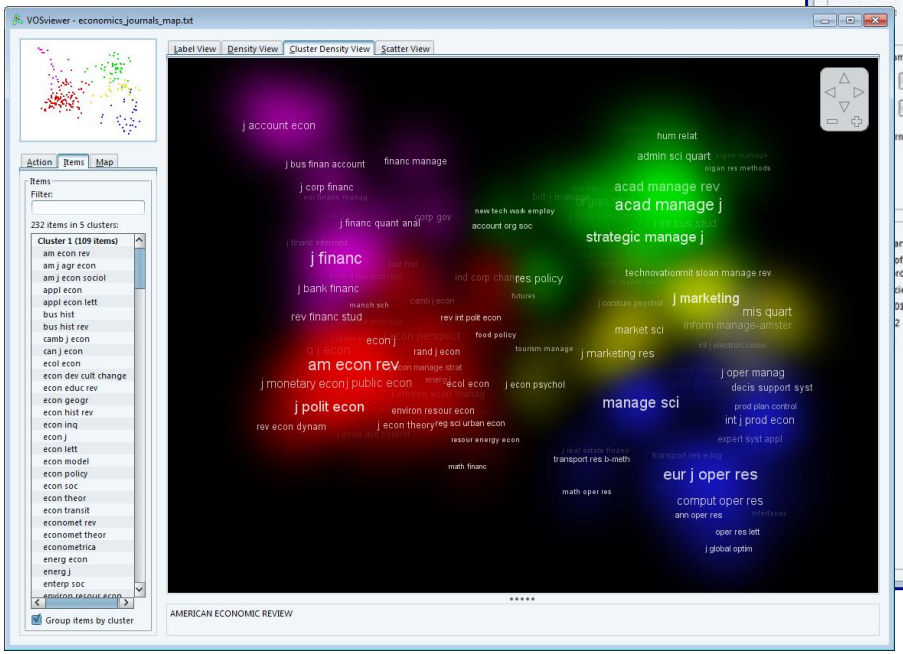
✓ Citations (Sumobrain)

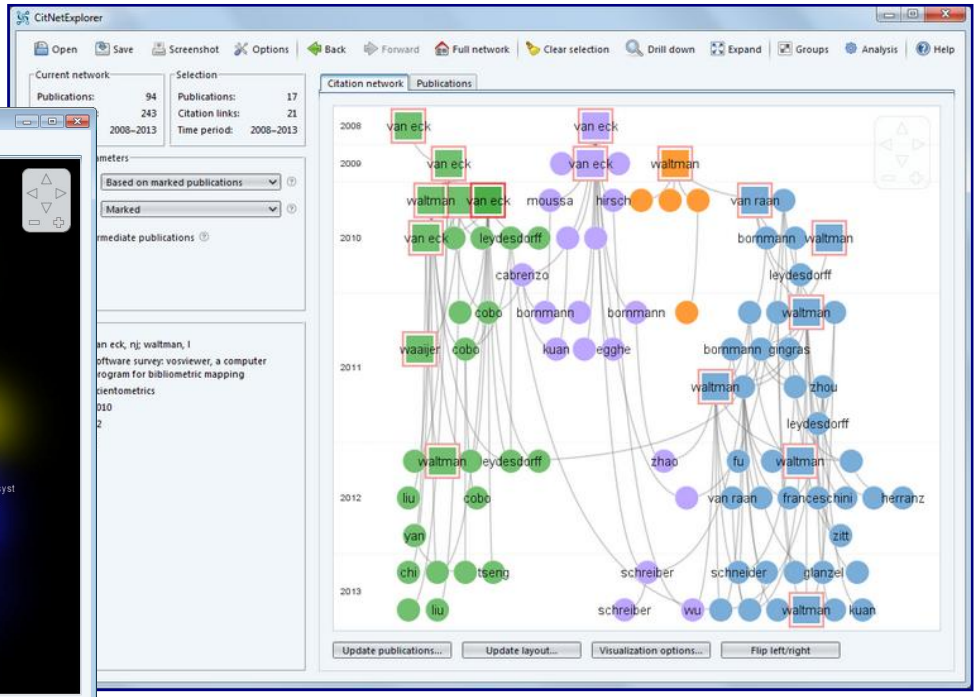
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Other / Related Tasks – I

Leiden University

✓ VosViewer



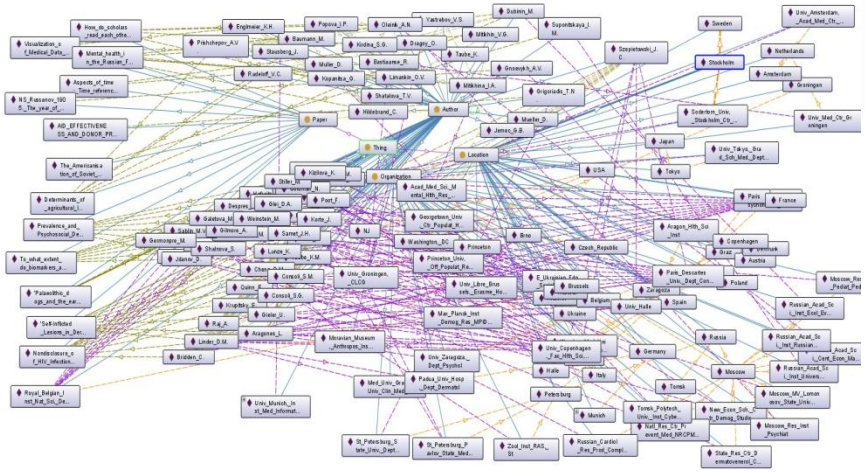


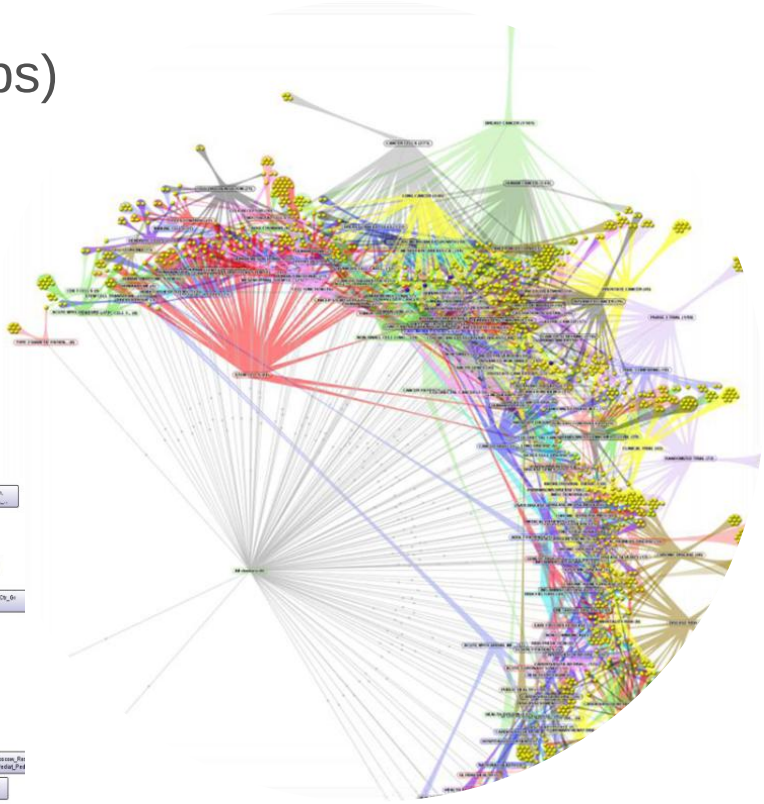
✓ CitNet Explorer

CitNetExplorer is a software tool for visualizing and analyzing citation networks of scientific publications. The tool allows citation networks to be imported directly from the Web of Science database. Citation networks can be explored interactively, for instance by drilling down into a network and by identifying clusters of closely related publications.

Other / Related Tasks – II

- ✓ Research fronts (WoS + Aduna Maps)
- ✓ “Shadow groups”

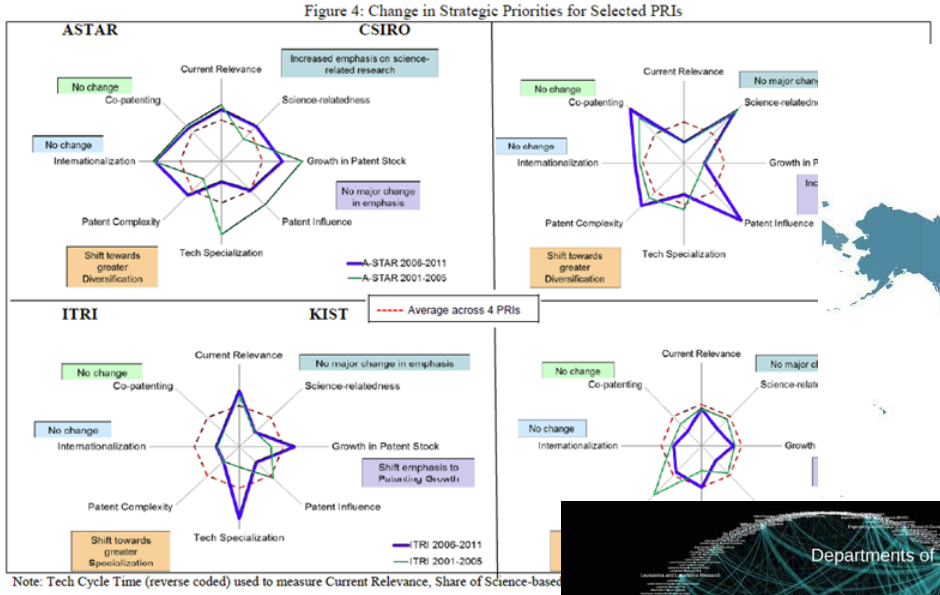




I. Efimenko, V. Khoroshevsky

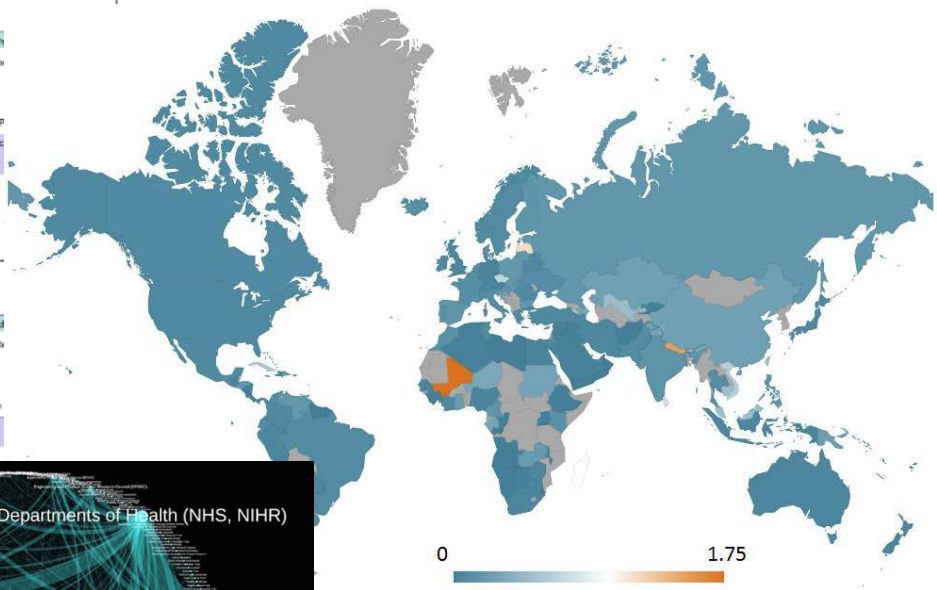
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

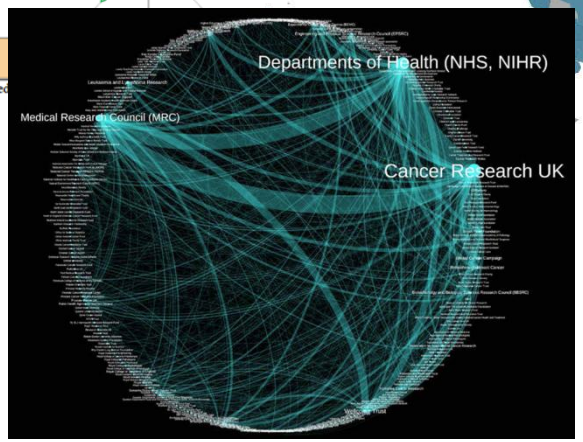
Other / Related Tasks – III



✓ Radars

✓ Global analytics





✓ Major Funders

The STI2014 Conference Proceedings

Other / Related Tasks – IV

- ✓ Indicator Engineering

We encourage anyone with information to share to visit us there, and we wish you a very nice conference.

DATA INDICATOR BUBBLES

cons of this idea? Please contact the office in the foyer.

DID METRICS KILL THE CAT?

According to the famous mearurement problem in quantum mechanics, one cannot do a measurement without severely influencing the measured system.

A somewhat similar problem can be identified in Scientometrics.

In his presentation, Paul Wouters warned that researchers in the scientometrics field should be aware that their measurements influence the scientific world. When asked whether this is a problem for

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Other / Related Tasks – V

✓ “Traditional” tools for new tasks: win-win partnerships

http://www.almostscientific.com/

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

And so…

- Irina Efimenko, Vladimir Khoroshevsky, Ed Noyons. (Map of Science)^2: Fields of S&T in Scientometric and Tech Mining Papers. 2016
- Irina Efimenko, Vladimir Khoroshevsky, Ed Noyons. Anticipating Future Pathways of Science, Technologies & Innovations: (Map of Science)^2 Approach. 2016
- Vladimir Khoroshevsky, Irina Efimenko. From Mining to Meaning: Identification of Meaningful Patterns in Technology Trend Monitoring. 2016
- Irina V. Efimenko, Vladimir F. Khoroshevsky. Peaks, Slopes, Canyons, Plateaus: Identifying Technology Trends throughout the Life Cycle // International Journal of Innovation and Technology Management. 2015 (forthcoming)
- Efimenko I. V., Khoroshevsky V. F. New Technology Trends Watch: An Approach and Case Study // Lecture Notes in Computer Science. 2014. No. 8722. P. 170-177

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

iefimenko@hse.ru
veassi@mail.ru
+7-916-101-4840