

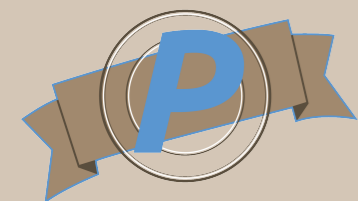
A Sea Change is Coming to Patent Analytics – Brought to You by Big Data

Anthony Trippe

Managing Director – Patinformatics, LLC

International Conference for the Information
Community (ICIC)

Vienna, Austria – 15 October 2013



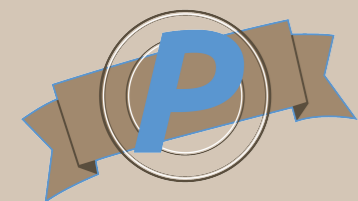
When I Submitted the Abstract

- Using IBM's Many Eyes for Generating Valuable Patent Analytics Insights
- I promise to show some examples of interesting visualizations using Many Eyes but...
- I predict a Sea Change in patent analysis, which based on the sophistication of this audience seems like a more interesting topic for discussion

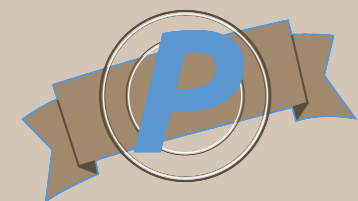


When I Submitted the Abstract

- The abstract actually describes the What to Look for When Using Public PAIR talk I gave at the PIUG Annual Conference
- Some of that material is covered in the workshop for Wednesday but you can also find portions of it at the following links:
- <http://www.patinformatics.com/presentations/>
- <http://www.patinformatics.com/blog/what-to-look-for-when-using-us-public-pair-an-infographic/>



So What is this Sea Change You Speak of?



Big Data and the Field of Data Science



The Transformation has Already Started

- Patent information professionals have been working with big data for years, decades even, but didn't use a catchy phrase to describe what they were doing
- The universe of available patent documents, worldwide, is well over 80 million
- The small molecule universe is well over 70 million substances
- Biggest change in over a decade for analytics



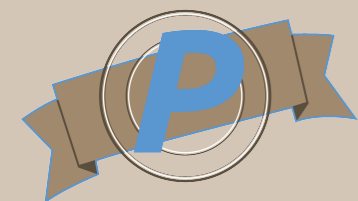
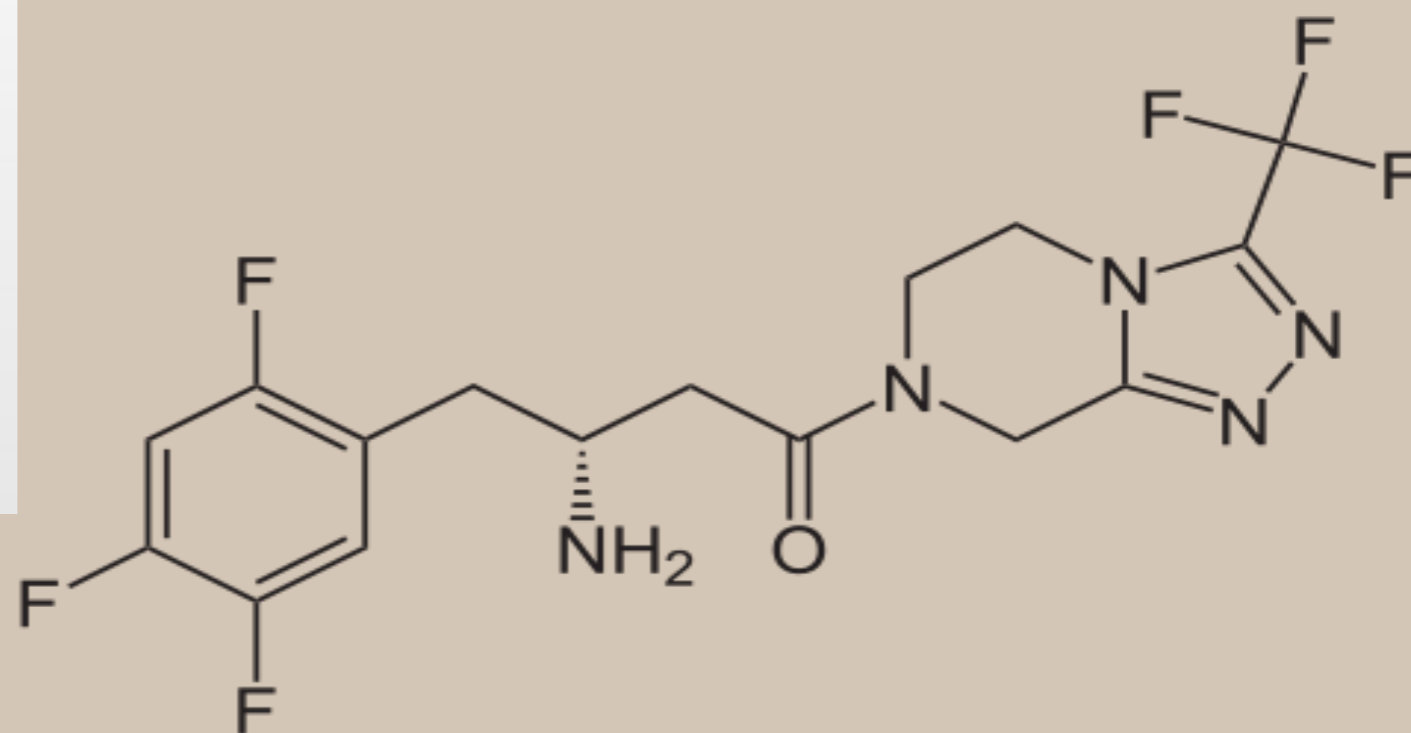
New STN Platform – Built for Big Data



- Hadoop implements a computational paradigm named MapReduce, where the application is divided into many small fragments of work, each of which may be executed or re-executed on any node in the cluster. In addition, it provides a distributed file system that stores data on the compute nodes, providing very high aggregate bandwidth across the cluster. It enables applications to work with thousands of computation-independent computers and petabytes of data.



Hadoop Makes the Following Possible



How Many of You Have Heard of R?

The screenshot shows the STN database interface. The top navigation bar includes 'STN Projects' and 'Working On - Januvia'. The main search area is titled 'CAPLUS/REGISTRY' and contains a search bar. Below the search bar, there are several search results listed, each with a chemical structure and a description. The results are: 1. Acetic acid, 2-[4-[[[(2S)-2-pyrrolidinylcarbonyl]amino]methyl]phenoxy]-, ethyl ester; 2. Cyclopentanecarbonitrile, 2-[[[(2S)-2-amino-4-[3,4-dihydro-1-(2-hydroxyethyl)-6,7-dimethoxy-2(1H)-isoquinolinyl]-1,4-dioxobutyl]-, (1R)-]; 3. Cyclopentanecarbonitrile, 2-[[[(2S)-2-amino-4-[3,4-dihydro-1-(2-hydroxyethyl)-6,7-dimethoxy-2(1H)-isoquinolinyl]-1,4-dioxobutyl]-, (1R)-]; 4. Cyclopentanecarbonitrile, 2-[[[(2S)-2-amino-4-[3,4-dihydro-1-(2-hydroxyethyl)-6,7-dimethoxy-2(1H)-isoquinolinyl]-1,4-dioxobutyl]-, (1R)-].

The screenshot shows the RStudio interface. The top navigation bar includes 'File Edit Code View Project Workspace Plots Tools Help'. The main workspace contains a script titled 'diamondPricing.R' with the following code:

```
1 library(ggplot2)
2 source("plots/formatPlot.R")
3
4 view(diamonds)
5 summary(diamonds)
6
7 summary(diamonds$price)
8 aveSize <- round(mean(diamonds$carat), 4)
9 clarity <- levels(diamonds$clarity)
10
11 p <- qplot(carat, price,
12            data=diamonds, color=clarity,
13            xlab="Carat", ylab="Price",
14            main="Diamond Pricing")
15
```

The console shows the output of the code, including the summary of the diamonds data and the execution of the qplot function. The plot shows a scatter plot of Price vs. Carat, with points colored by clarity. The plot is titled 'Diamond Pricing' and has a legend for Clarity with categories: I1, SI2, SI1, VS2, VS1, VVS2, VVS1, and IF.

It's Coming!

How Many of You are Using OpenRefine?

Extensions: **Freebase** ▾

ows records Show: 5 10 25 50 rows « first < previous 1 - 10 next > last »

Publication Num	Title	Title - DWPI	Citing Patents	Count of Citing	Citing Pat 1st As	Citing References Details - Patents
US8442695B2	Auto-adaptable energy management apparatus	Auto-adaptable energy management apparatus for managing e.g. energy production for residential customers through energy management user interface of network device, has processor creating personalized schedule	US20120130513A1 US8375118B2	2	VERIZON PATENT & LICENSING INC VERIZON PATENT &	<ul style="list-style-type: none"> Facet ▶ Text filter Edit cells ▶ Edit column ▶ Transpose ▶ Sort... View ▶ Reconcile ▶
US8428782B2	Energy management system and method	Temperature control apparatus for home energy management system has processor which deploys current set-point stored within memory which stores temperature set-	US20120179547A1 US20120023225A1 US20120047092A1 US20120063334A1 US20120116597A1 US20120130513A1		GEN ELECTRIC GEN ELECTRIC VERIZON PATENT & LICENSING INC NAT UNIV TSINGHUA ALLURE ENERGY INC NEST LABS INC VERIZON PATENT & LICENSING INC	<ul style="list-style-type: none"> US20120179547A1,US, ,7 (Pre-search),2011-01-06,GEN ELECTRIC US20120023225A1,US, ,7 (Pre-search),2009-07-20, US20120047092A1,US, ,7 (Pre-search),2010-08-17, US20120063334A1,US, ,7 (Pre-search),2010-10-21,GEN ELECTRIC US20120116597A1,US, ,7 (Pre-search),2010-11-09,GEN ELECTRIC US20120130513A1,US, ,7 (Pre-search),2010-11-18,VERIZON PATENT & LICENSING INC US20120166010A1,US, ,7 (Pre-search),2010-12-28,NAT UNIV TSINGHUA US8024073B2,US, ,0 (Examiner),2009-08-21,ALLURE ENERGY INC US8280536B1,US, ,1



How Many of You are Using OpenRefine?

- Clean up data fields
- Split multi-value cells
- Split into several columns
- Count string length
- Determine how many times an item occurs in a cell

• Many, many more uses for manipulating data

Patinformatics, LLC[®]

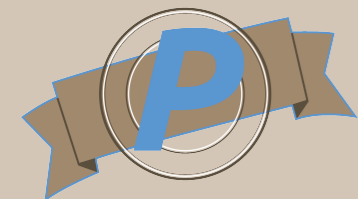
Data Driven Decisions

Patent Strategy and Analytics Services

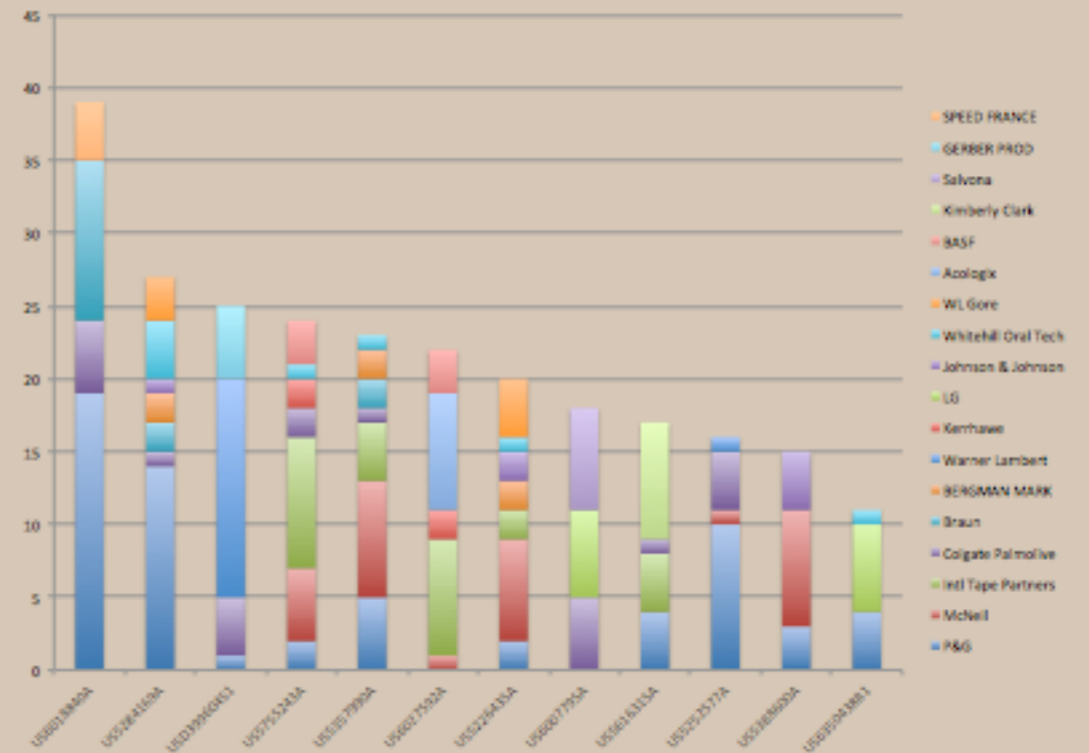
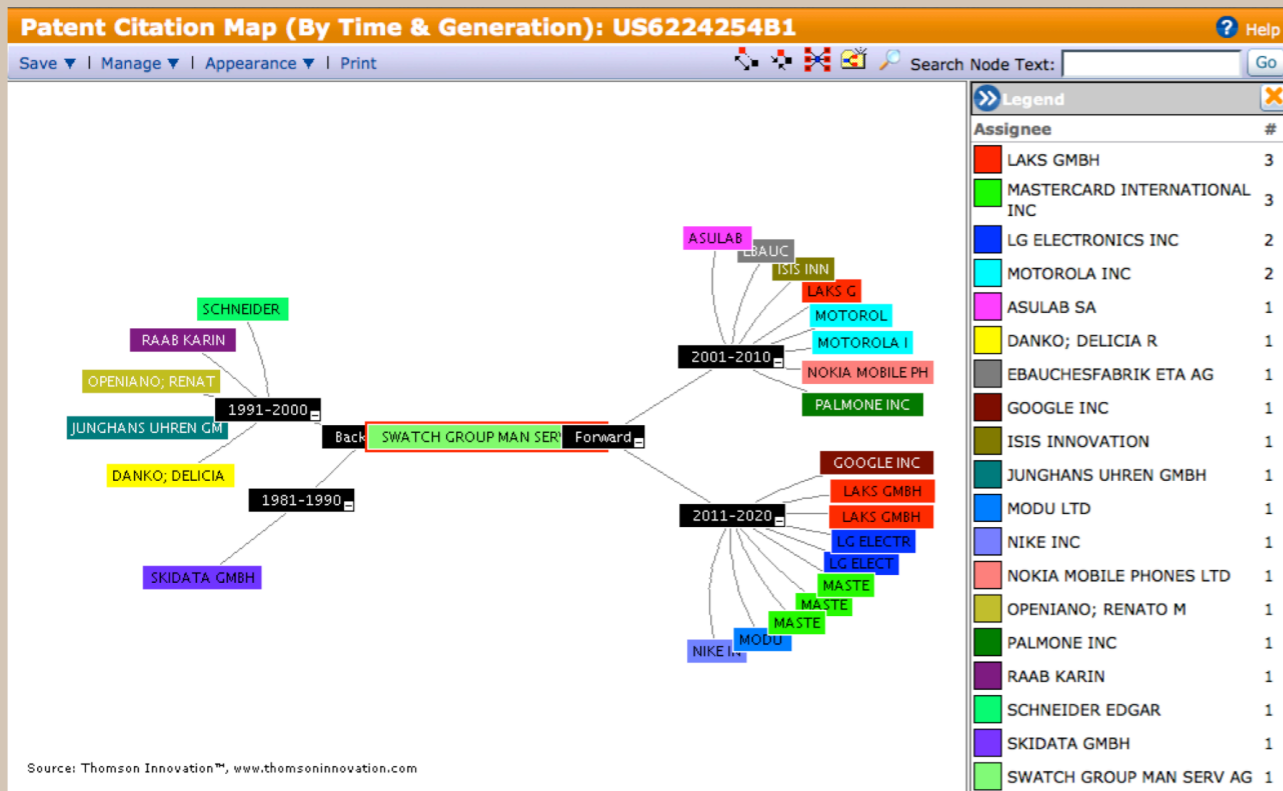
REPRESENTATIVE EXAMPLES OF RE- APPLIED TECHNIQUES



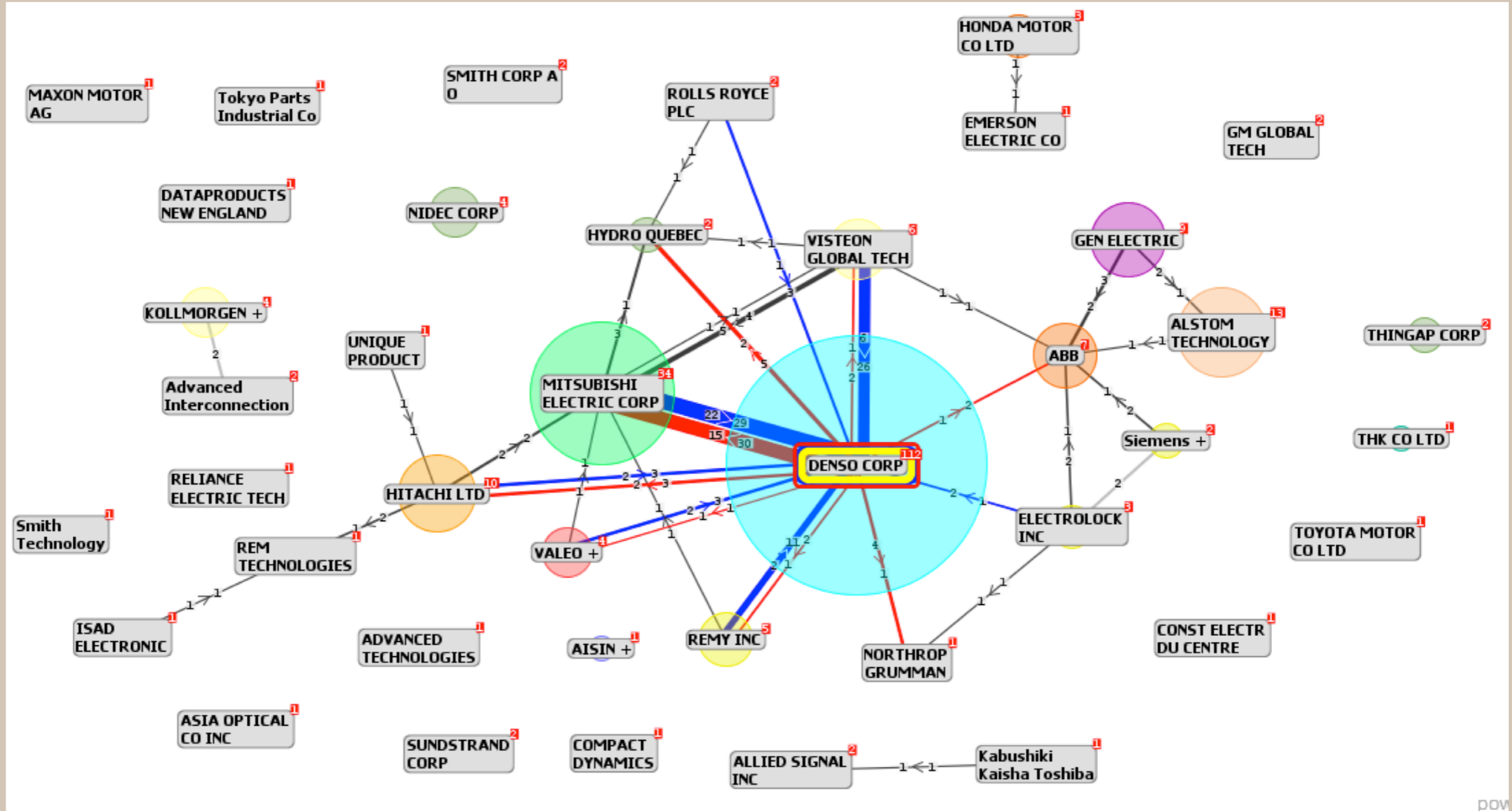
Big Data is Concerned with the Same Things We Are



Traditional Citation Analysis



Next Generation Citation Analysis



pow



Tools to Help with Network Analysis

- TouchGraph – <http://www.touchgraph.com/navigator>
- Cytoscape – <http://www.cytoscape.org>
- Sci2 – <https://sci2.cns.iu.edu/user/index.php>
- Also available in commercial tools such as Orbit.com, Relecura and Intellixir



Tools to Visualize and Analyze Text Data

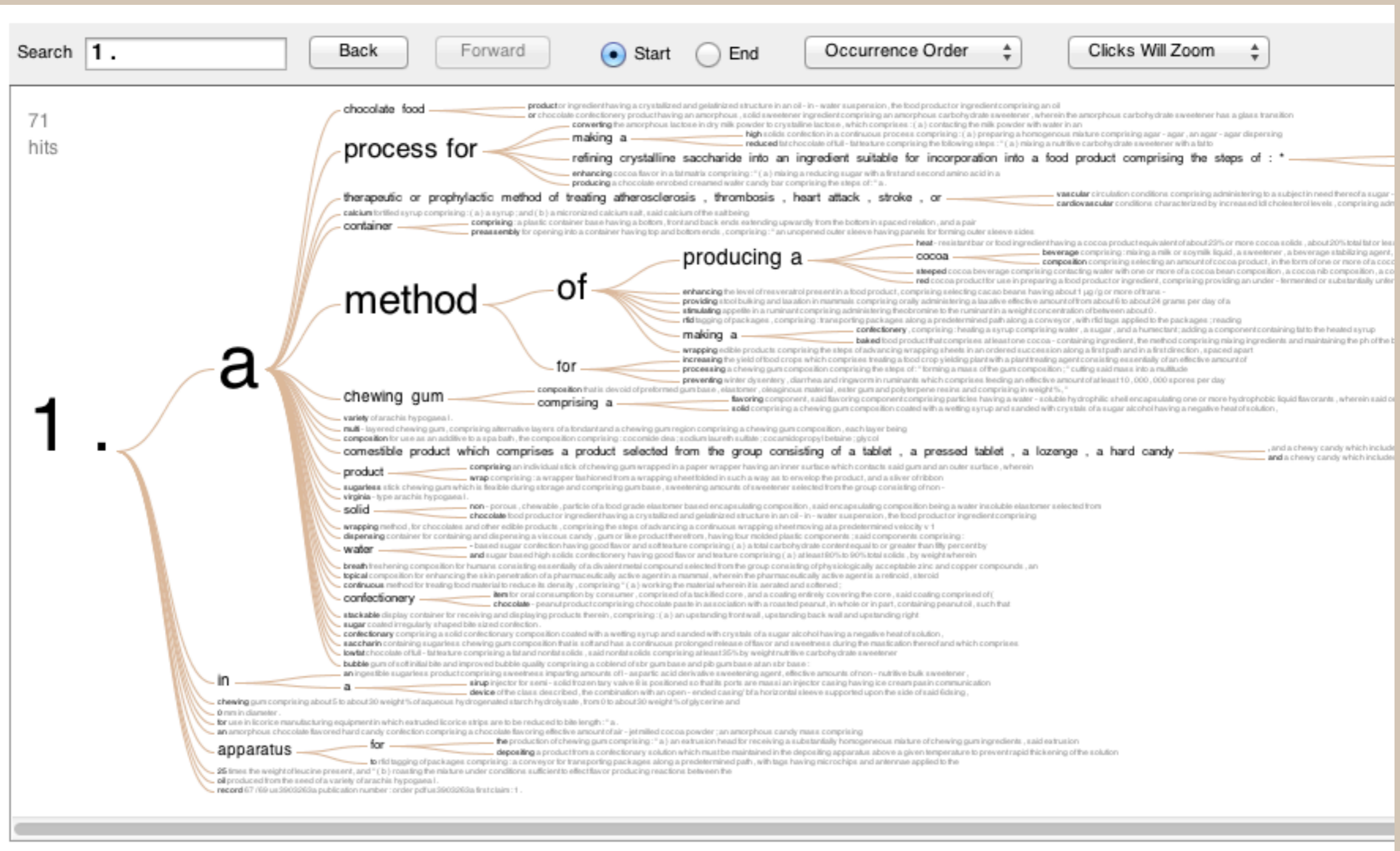
- Many Eyes – http://www-958.ibm.com/software/data/cognos/manyeyes/page/Visualization_Options.html
- Jigsaw – <http://www.cc.gatech.edu/gvu/ii/jigsaw/>
- Sci2 – <https://sci2.cns.iu.edu/user/index.php>
- Weka - <http://www.cs.waikato.ac.nz/ml/weka/>
- R - <http://tm.r-forge.r-project.org>
- Commercial patent tools also available as well



Network Diagrams – Family Trees



Word Tree – Claims Analysis



Word Tree – Claims Analysis



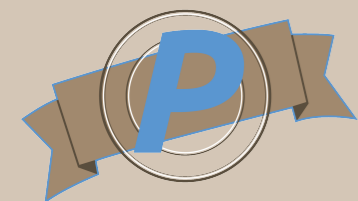
Tag Clouds – Hedge & Synonym Discovery

1 word 2 word compare

Search:

Showing 64 Out of 5009

chocolate baking chocolate bars **chocolate beverage** chocolate bittersweet **chocolate cake** chocolate candies chocolate caramel
chocolate chocolate chocolate chocolate-coated chocolate chocolate-flavored chocolate claimed chocolate coating **chocolate comprising**
chocolate confectionery chocolate dark chocolate drink chocolate drinks chocolate flavor chocolate flavouring chocolate fondant
chocolate food chocolate foods chocolate icing chocolate industry chocolate juice **chocolate liquor**
chocolate maker chocolate method **chocolate milk** chocolate paste chocolate phase chocolate preferably chocolate preferred chocolate preparation chocolate processes
chocolate product chocolate products chocolate pudding **chocolate scent** chocolate semi-sweet
chocolate semisweet chocolate syrup chocolate tend chocolate useprods **chocolate white** chocolate-coated bars chocolate-coated candies
chocolate-coated crystalline chocolate-coated foods chocolate-coated product chocolate-flavored bars **chocolate-flavored beverage** chocolate-flavored candies
chocolate-flavored confectionery chocolate-flavored drinks chocolate-flavored foods **chocolate-flavored product** chocolate-flavoured hard
chocolate-like product chocolate-like products chocolate-peanut prod **chocolates 2** chocolates produced chocolatey flavor



Conclusions

- The advent of Big Data and Data Science is creating an environment for growth that has not been seen in more than a decade
- Data structures and algorithms for dealing with very large data collections will be directly applicable to the analysis of scientific literature
- Looking for methods outside of our areas of expertise can provide new means for providing insight and value to our own data and analysis

