Network Analytics
meets
Text Mining
for Social Media Analysis

Dr. Bernd Wiswedel
KNIME.com AG

Agenda

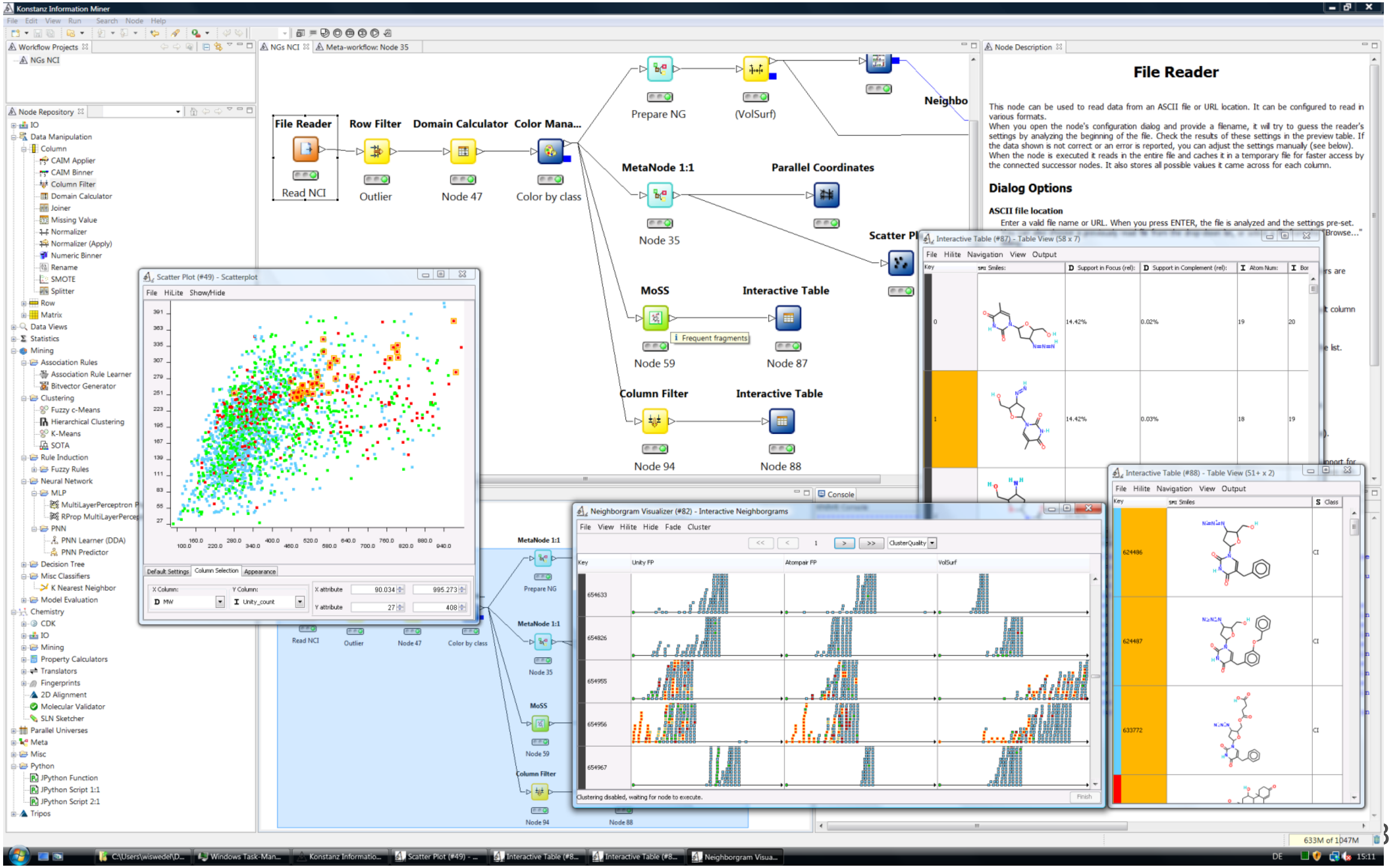
(Brief Overview of KNIME)

Social Media Data – Challenges

Case Study: Slashdot

- Text Mining: Sentiment Analysis
- Network Mining: Topic Graphs
- Combination

The KNIME Platform



Node Description

File Reader

This node can be used to read data from an ASCII file or URL location. It can be configured to read in various formats.
When you open the node's configuration dialog and provide a filename, it will try to guess the reader's settings by analyzing the beginning of the file. Check the results of these settings in the preview table. If the data shown is not correct or an error is reported, you can adjust the settings manually (see below).
When the node is executed it reads in the entire file and caches it in a temporary file for faster access by the connected successor nodes. It also stores all possible values it came across for each column.

Dialog Options

ASCII file location

Enter a valid file name or URL. When you press ENTER, the file is analyzed and the settings pre-set.

A Brief History of KNIME

2004: KNIME development commences

2006: KNIME v1 released

2006: Spin-off in Konstanz, Germany

2006-2007: First commercial partners

2008: KNIME moves to Zurich

2010: Enterprise products released

Status Quo:

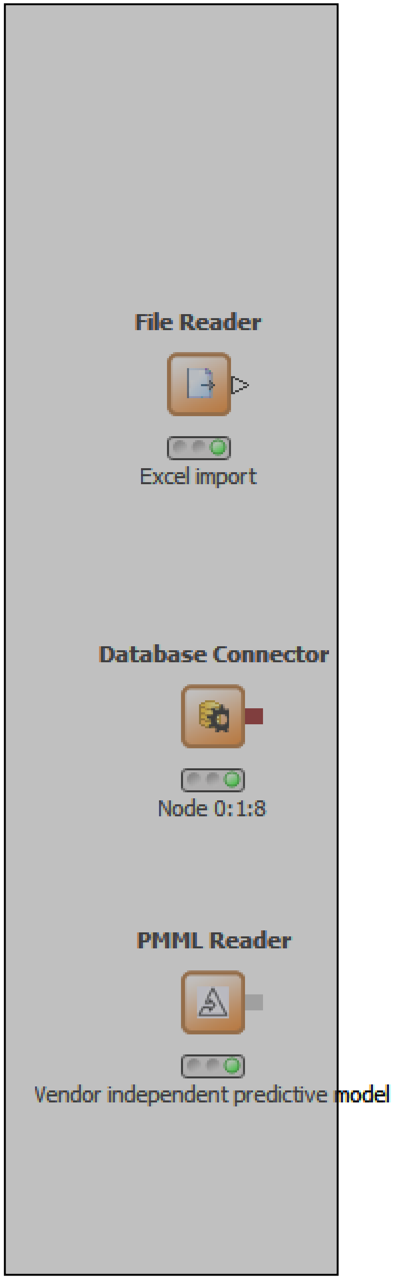
- KNIME used in 30+ countries:
  +3000 Organizations
  ~30% Life Science
  ~70% Business Intelligence, Analytics, Data Mining
  +50 Very Active Community Developers

KNIME 2.8 released in July 2013

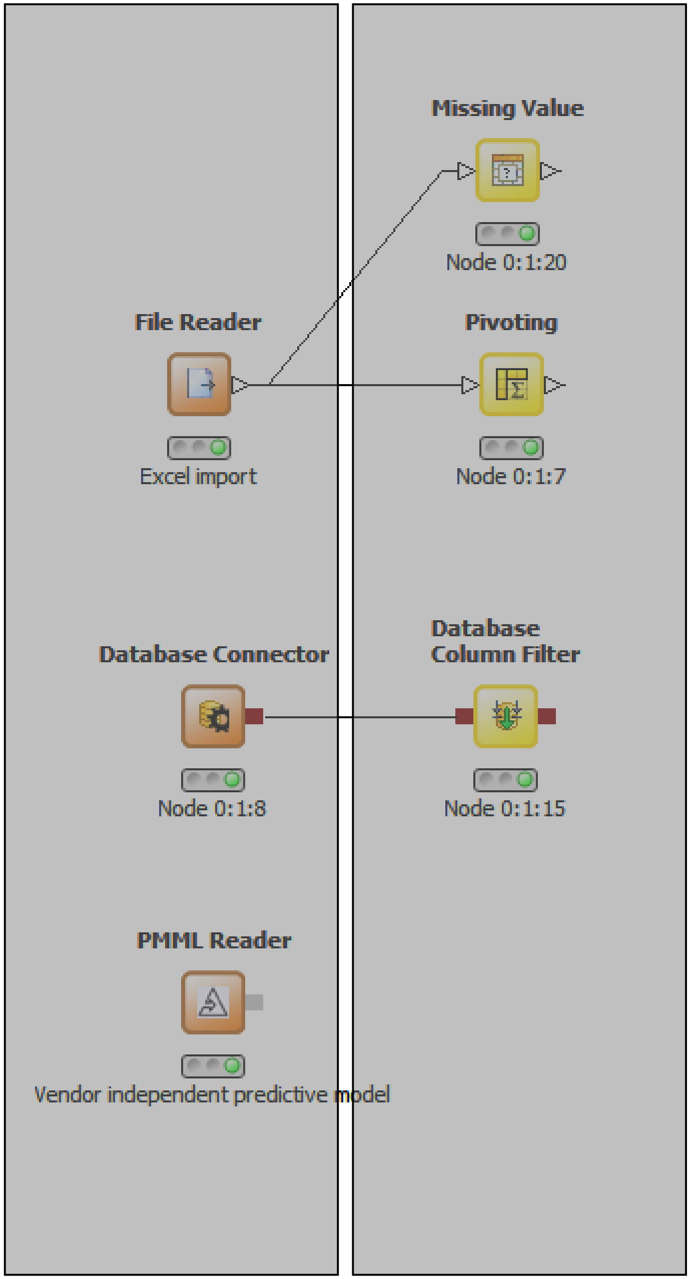
KNIME

KNIME loads and integrates data from diverse data sources:

- Different data bases
- Various file formats (CSV, XML, SDF, etc.)



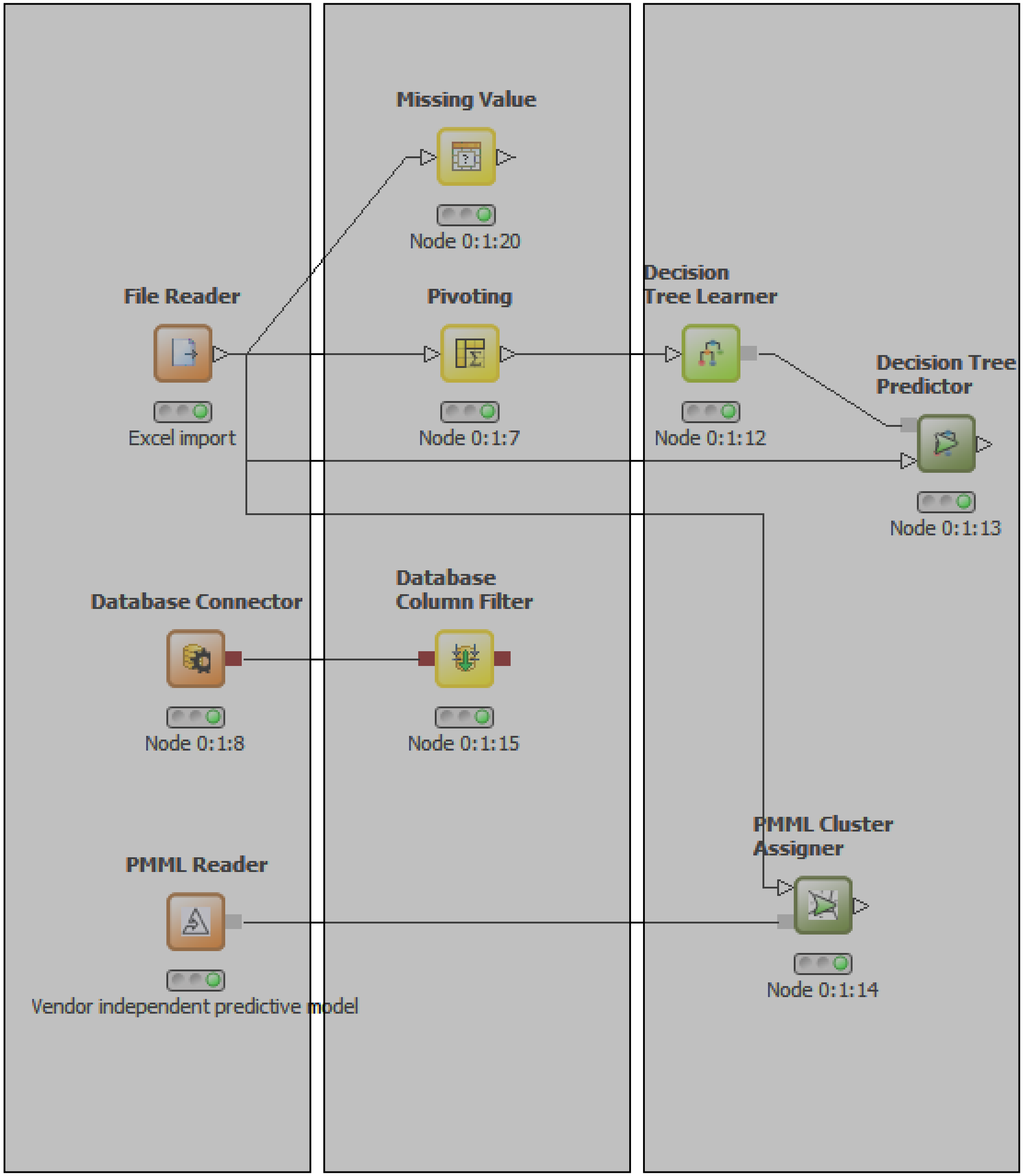
KNIME



KNIME provides huge repository of modules for easy-to-use, modular

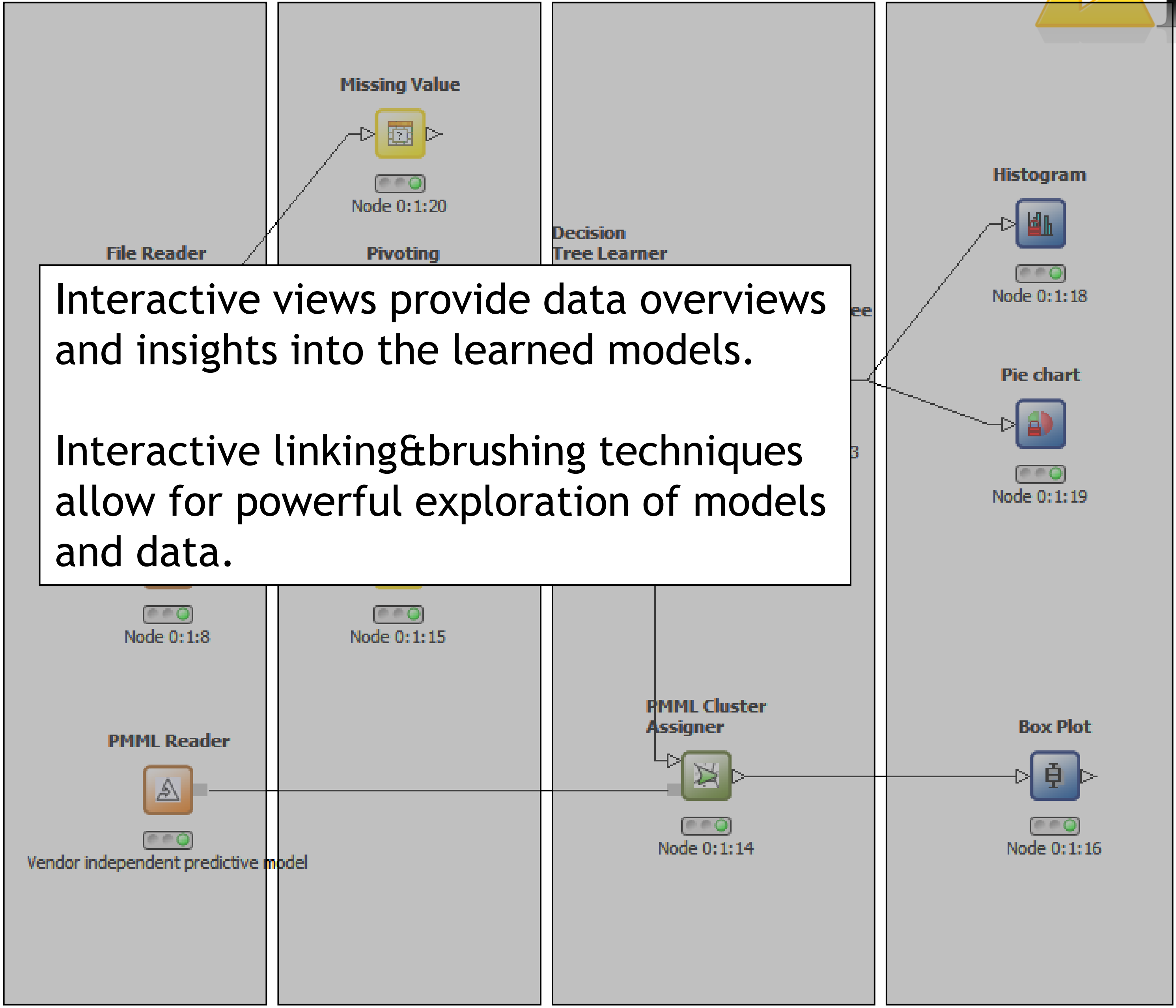
- Data preprocessing
- Data fusion
- Data transformation

KNIME



In addition to standard data mining techniques, KNIME adds cutting edge data analysis algorithms. (…thanks to its academic roots)

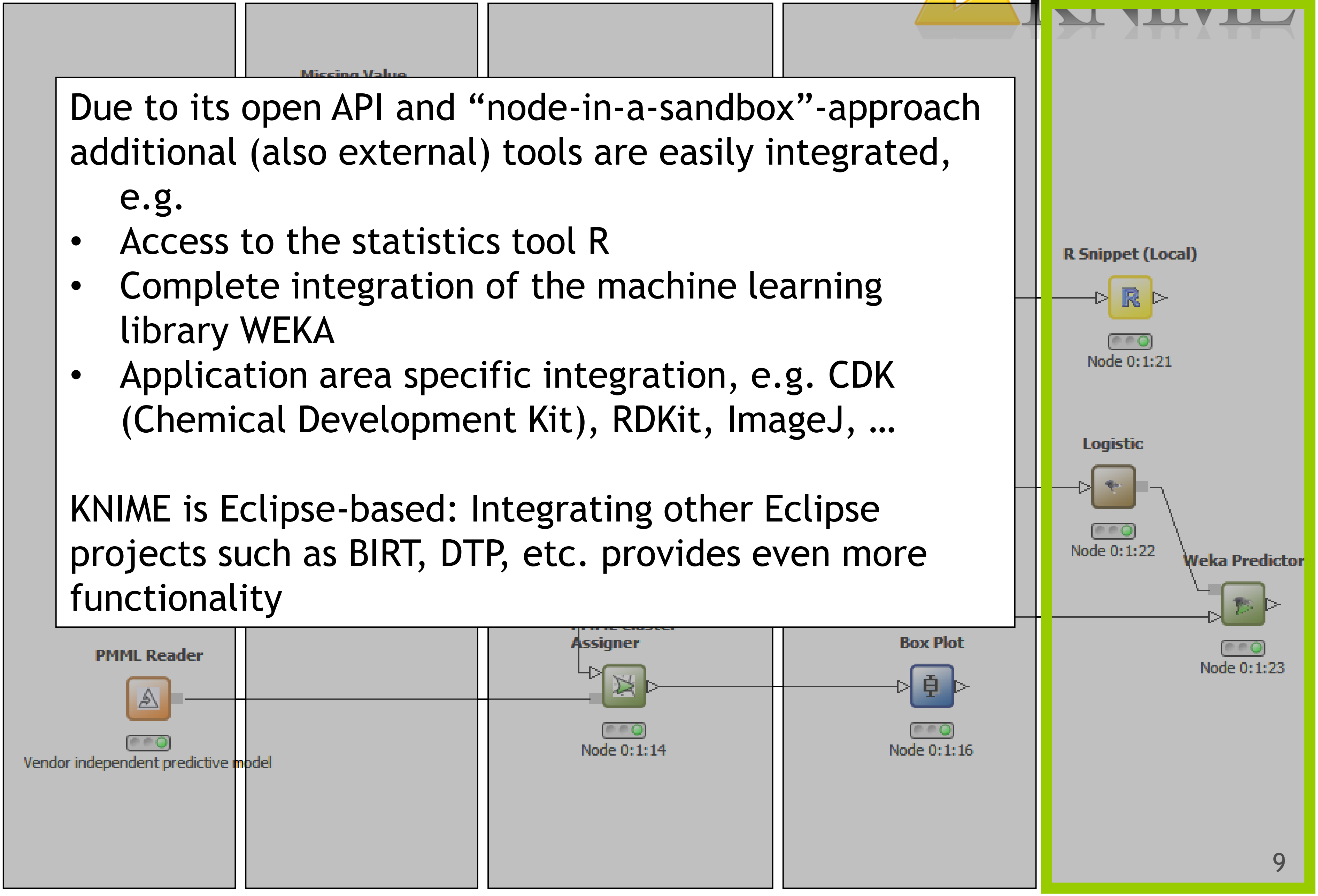
KNIME



Interactive views provide data overviews and insights into the learned models.

Interactive linking&brushing techniques allow for powerful exploration of models and data.

KNIME



Due to its open API and “node-in-a-sandbox”-approach additional (also external) tools are easily integrated, e.g.

- Access to the statistics tool R
- Complete integration of the machine learning library WEKA
- Application area specific integration, e.g. CDK (Chemical Development Kit), RDKit, ImageJ, …

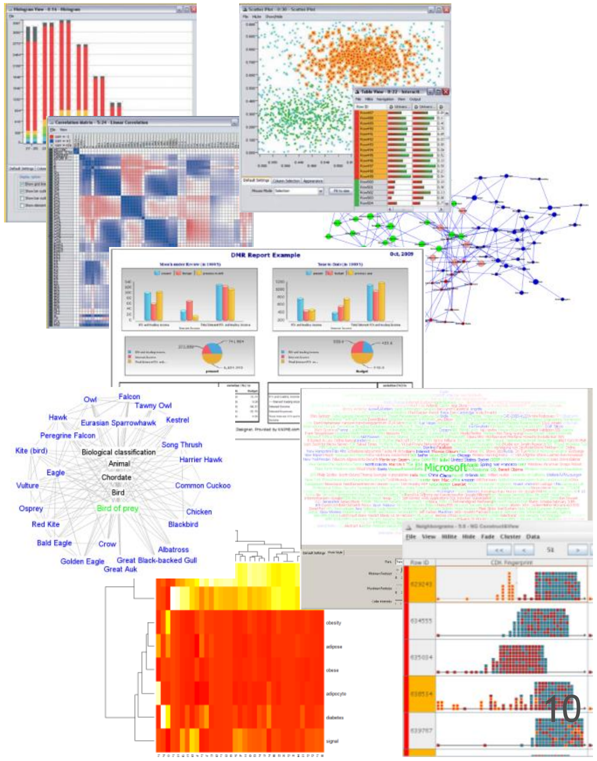
KNIME is Eclipse-based: Integrating other Eclipse projects such as BIRT, DTP, etc. provides even more functionality

KNIME Selected Node Highlights

Over 1000 native and embedded nodes included:

- Statistics
- Data Mining
- Time Series
- Image Processing
- Neighborgrams
- Web Analytics
- Text Mining
- Network Analysis
- WEKA
- R

- Database Support
- ETL
- Text Processing
- Data Generation
- XML Support
- PMML Support
- Social Media Analysis
- Business Intelligence
- Community Nodes
- 3rd Party Nodes



Social Media Data

Water Water Everywhere, and not a drop to drink

facebook

twitter

KNIME tech

skype

WIKIPEDIA
The Free Encyclopedia

Google+

Slashdot
/.

LinkedIn

Social Media Data

Water Water Everywhere, and not a drop to drink

What companies do with it:

- Download and keep
- Topic [Shift] Detection (email content routing, detect market interest shift, clinical studies, query non structured DBs, ...)
- Sentiment Analysis (marketing, polls, elections, ...)
- Connection Analysis (influencers, risk analysis, ...)
- ....

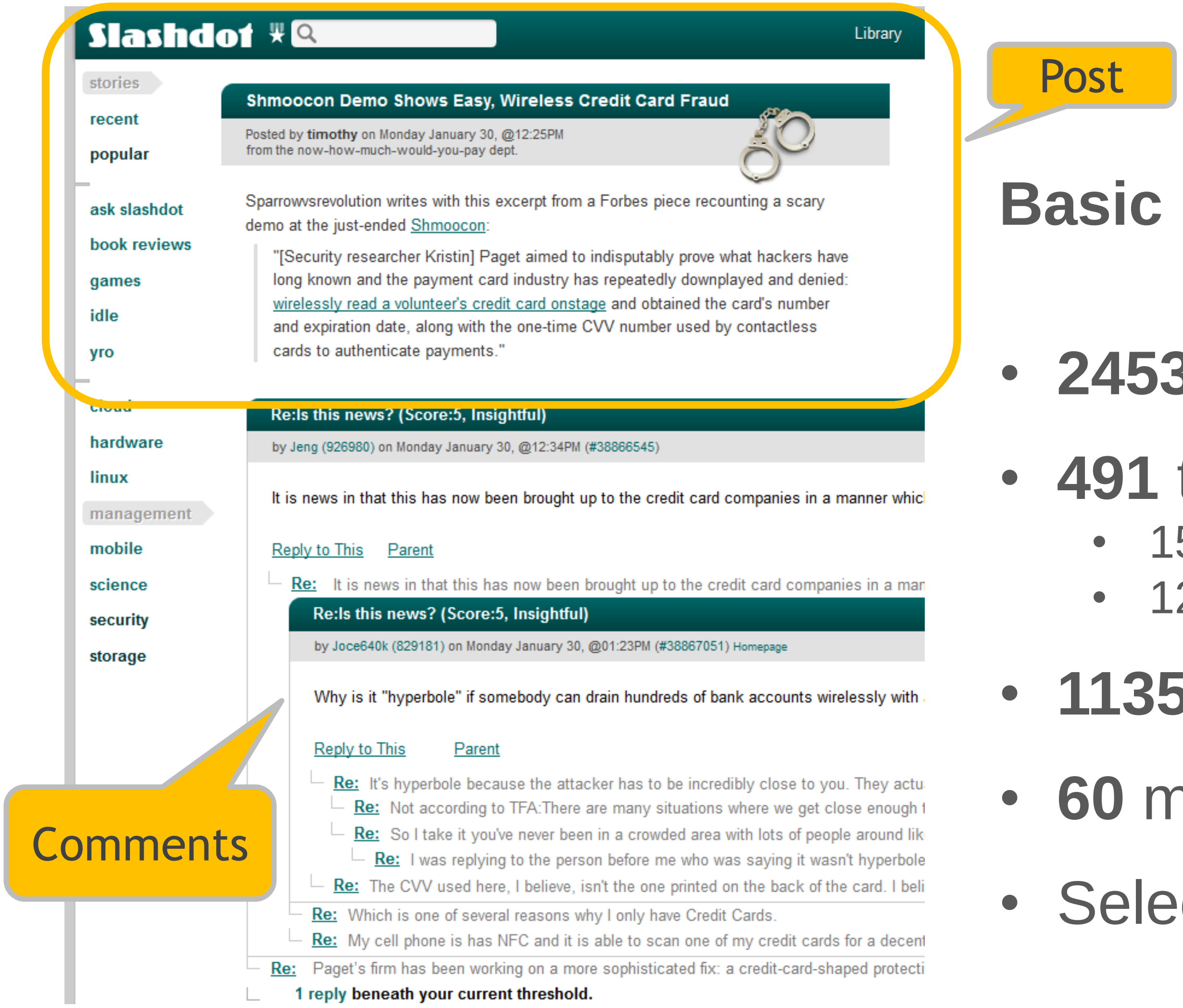
Social Media Data

Water Water Everywhere, and not a drop to drink

The Analysis Tools:

- Web Crawlers
- Visual Exploration
- Topic Detection (Text Mining, NLP, Ontologies)
- Sentiment Score (Text Mining, NLP)
- Influence Score (Network Analytics)
- Find Groups (Predictive Analytics)

Case Study Example: Slashdot Data



Shmoocon Demo Shows Easy, Wireless Credit Card Fraud

Posted by **timothy** on Monday January 30, @12:25PM
from the now-how-much-would-you-pay dept.

Sparrowvsrevolution writes with this excerpt from a Forbes piece recounting a scary demo at the just-ended Shmoocon:

> "[Security researcher Kristin] Paget aimed to indisputably prove what hackers have long known and the payment card industry has repeatedly downplayed and denied: wirelessly read a volunteer's credit card onstage and obtained the card's number and expiration date, along with the one-time CVV number used by contactless cards to authenticate payments."

Re:Is this news? (Score:5, Insightful)

by Jeng (926980) on Monday January 30, @12:34PM (#38866545)

It is news in that this has now been brought up to the credit card companies in a manner whic

Reply to This Parent

Re: It is news in that this has now been brought up to the credit card companies in a man

Re:Is this news? (Score:5, Insightful)

by Joce640k (829181) on Monday January 30, @01:23PM (#38867051) Homepage

Why is it "hyperbole" if somebody can drain hundreds of bank accounts wirelessly with

Reply to This Parent

Re: It's hyperbole because the attacker has to be incredibly close to you. They actu

Re: Not according to TFA:There are many situations where we get close enough t

Re: So I take it you've never been in a crowded area with lots of people around lik

Re: I was replying to the person before me who was saying it wasn't hyperbole

Re: The CVV used here, I believe, isn't the one printed on the back of the card. I beli

Re: Which is one of several reasons why I only have Credit Cards.

Re: My cell phone is has NFC and it is able to scan one of my credit cards for a decent

Re: Paget's firm has been working on a more sophisticated fix: a credit-card-shaped protecti

1 reply beneath your current threshold.

Post

Comments

Basic Numbers:

- **24532** users
- **491** threads with
  - 15 – 843 responses
  - 12 – 507 users
- **113505** posts
- **60** main topics
- Selected Topic: **Politics**

Case Study Example: Slashdot

- Very rich data sources about customers !

- We want to establish:

Sentiment Analysis

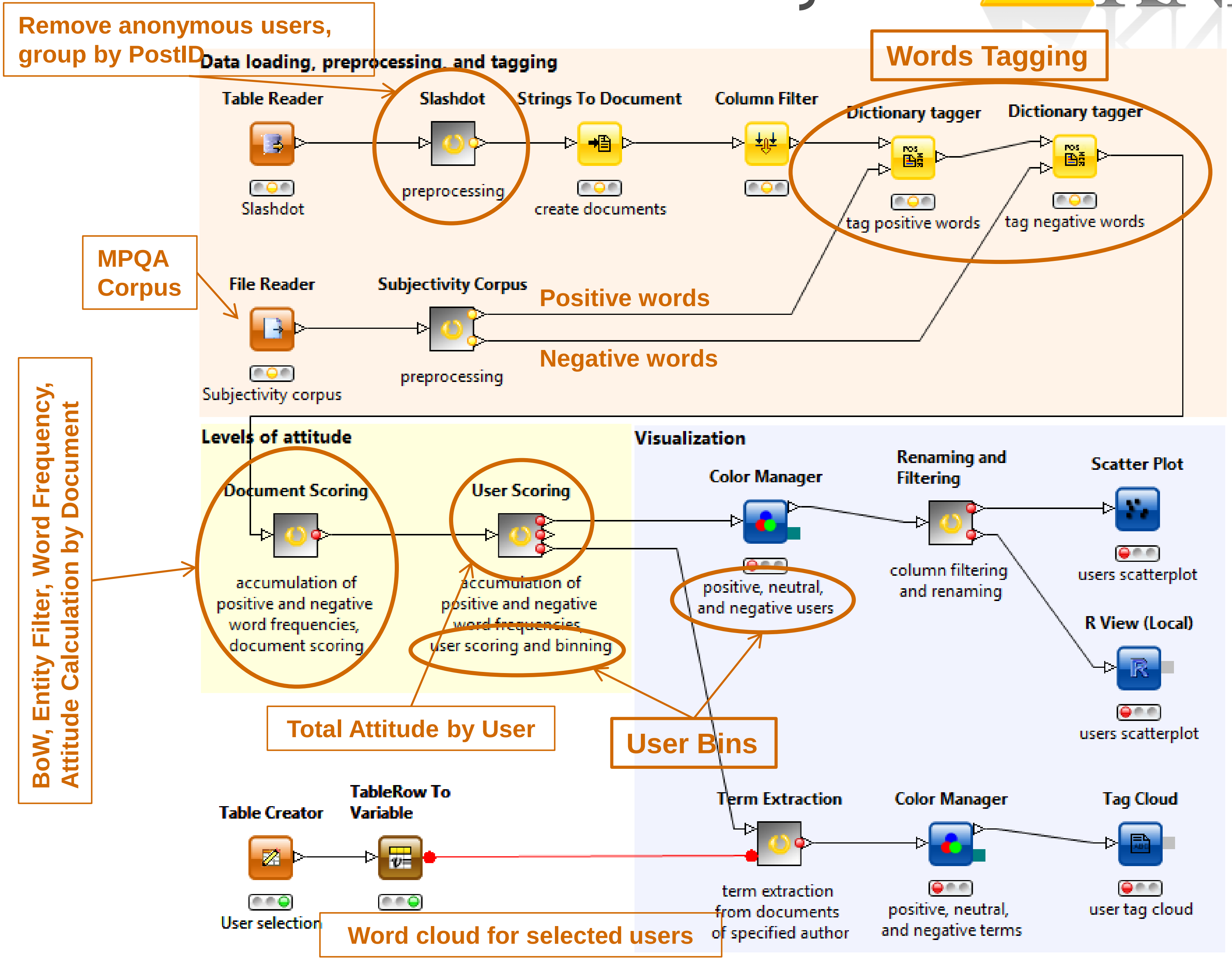
  - How users feel about the discussed topic
  - Whether it matters how users feel

Network Analytics

  - A more general abstraction of the results

Clustering

Sentiment Analysis



Remove anonymous users, group by PostID

Words Tagging

MPQA Corpus

Positive words

Negative words

BoW, Entity Filter, Word Frequency, Attitude Calculation by Document

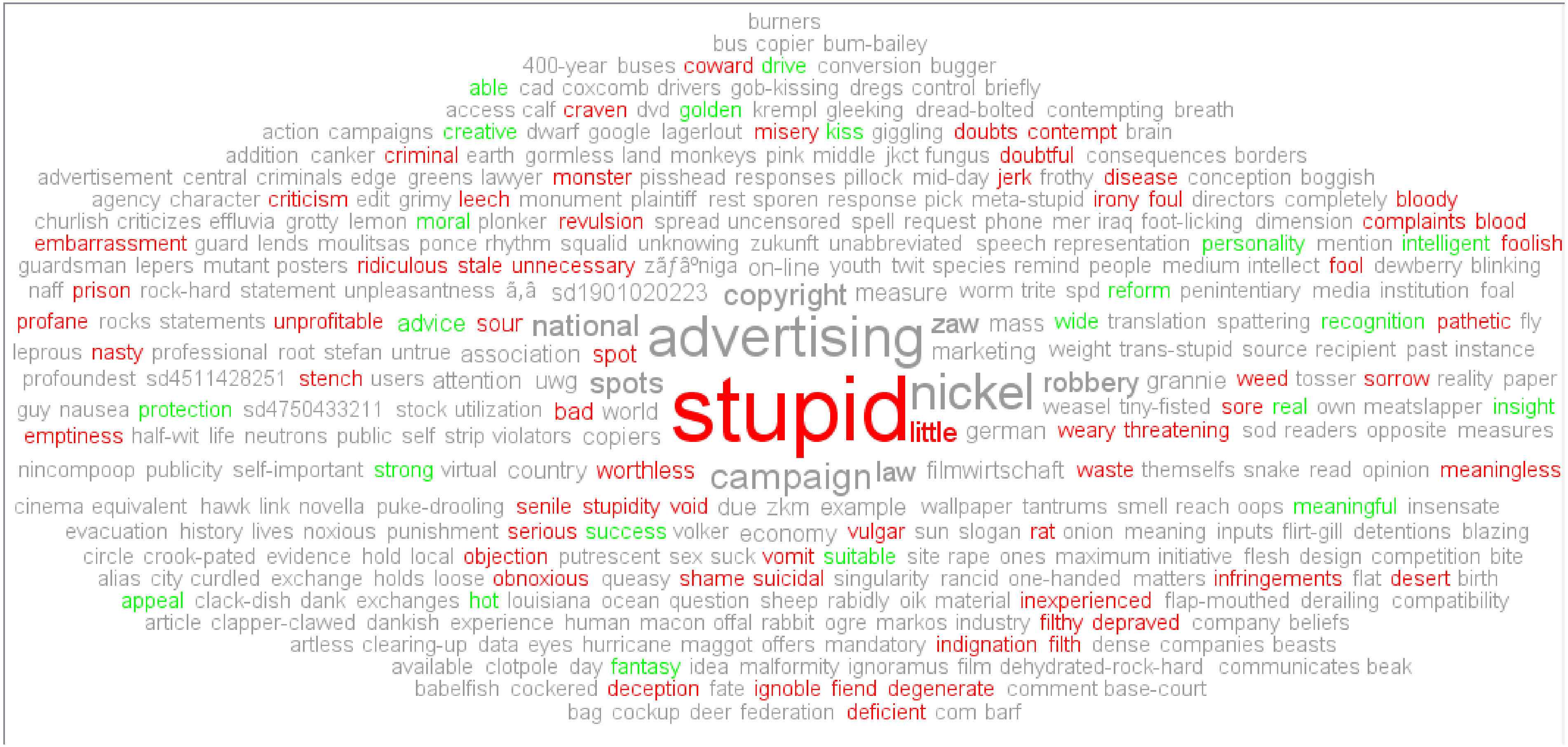
Total Attitude by User

User Bins

Word cloud for selected users

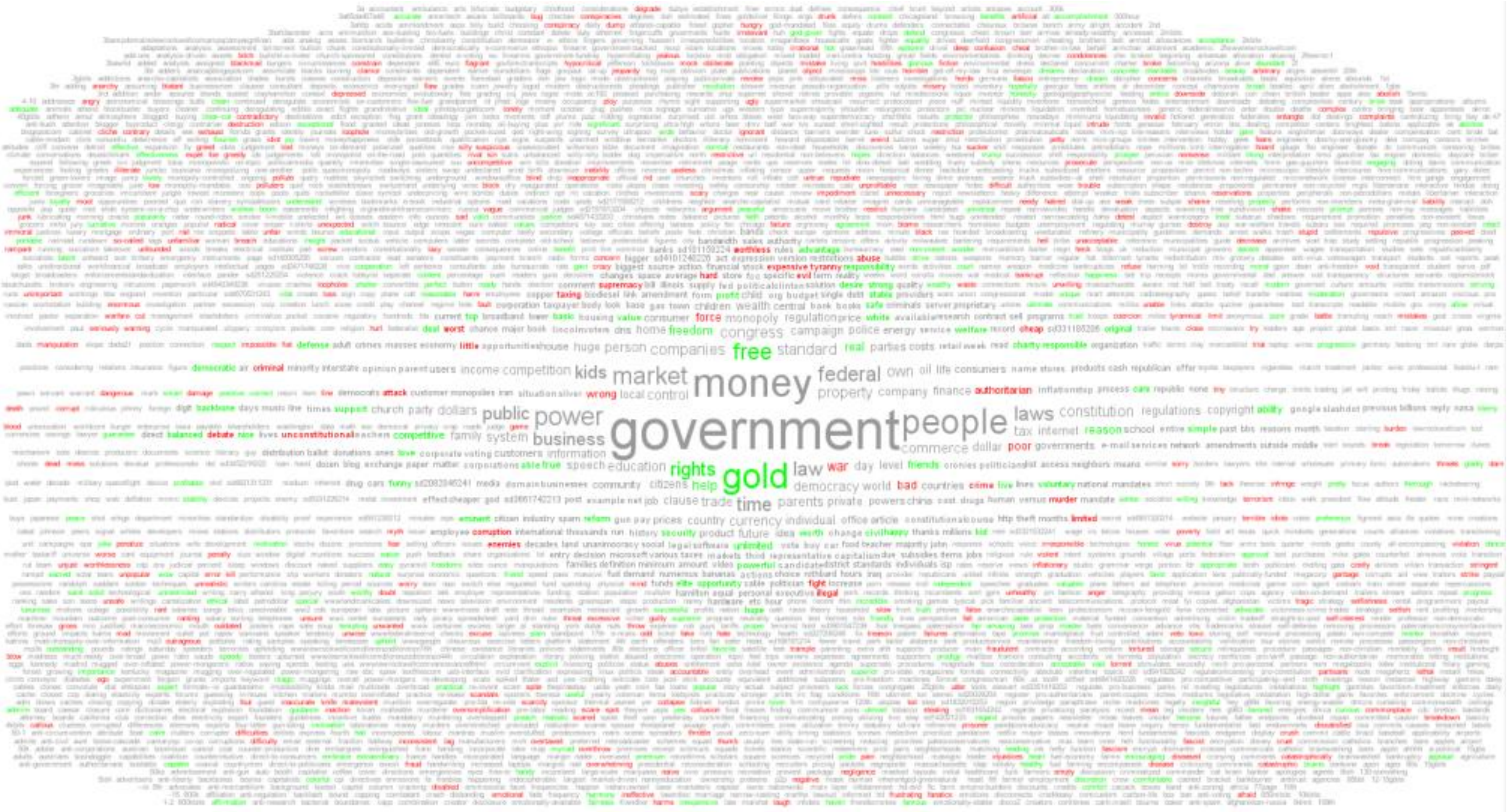
Slashdot – Text Mining

Most Negative User pNutz



Slashdot – Text Mining

Most Positive User dada21



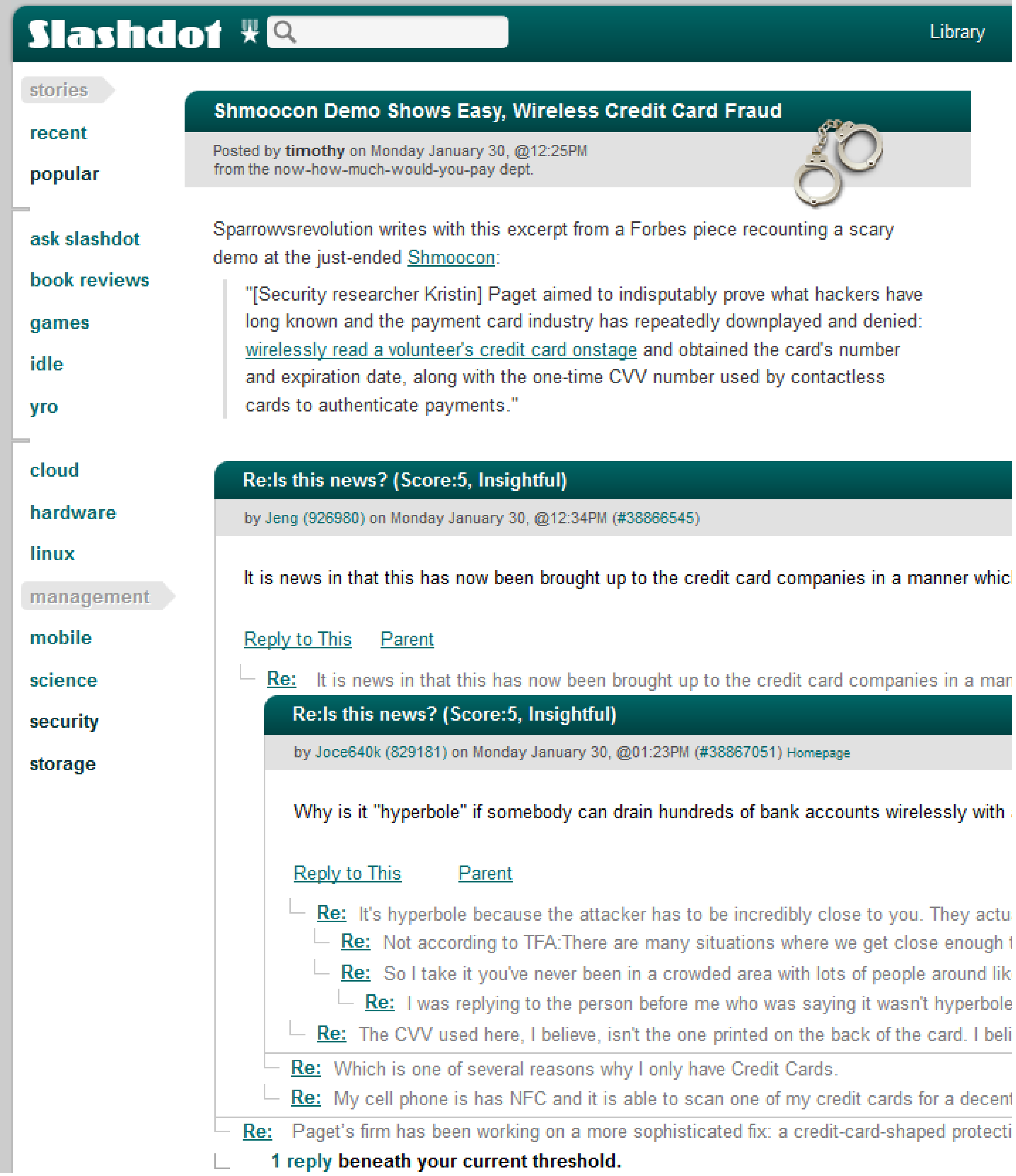
Slashdot – Sentiment Analysis

- **16016** positive users
- **7107** negative users

- Most positive user: dada21 (2838 positive/1725 negative words)
- Most negative user: pNutz (43 positive/109 negative words)

- Which Topics have positive users in common ?
  - Government
  - People
  - Law/s
  - Money
  - Market
  - Parties

Network Creation



Shmoocon Demo Shows Easy, Wireless Credit Card Fraud

Posted by **timothy** on Monday January 30, @12:25PM
from the now-how-much-would-you-pay dept.

Sparrowvsrevolution writes with this excerpt from a Forbes piece recounting a scary demo at the just-ended Shmoocon:

> "[Security researcher Kristin] Paget aimed to indisputably prove what hackers have long known and the payment card industry has repeatedly downplayed and denied: wirelessly read a volunteer's credit card onstage and obtained the card's number and expiration date, along with the one-time CVV number used by contactless cards to authenticate payments."

Re:Is this news? (Score:5, Insightful)

by Jeng (926980) on Monday January 30, @12:34PM (#38866545)

It is news in that this has now been brought up to the credit card companies in a manner whic

Reply to This Parent

Re: It is news in that this has now been brought up to the credit card companies in a mar

Re:Is this news? (Score:5, Insightful)

by Joce640k (829181) on Monday January 30, @01:23PM (#38867051) Homepage

Why is it "hyperbole" if somebody can drain hundreds of bank accounts wirelessly with

Reply to This Parent

Re: It's hyperbole because the attacker has to be incredibly close to you. They actu

Re: Not according to TFA:There are many situations where we get close enough t

Re: So I take it you've never been in a crowded area with lots of people around lik

Re: I was replying to the person before me who was saying it wasn't hyperbole

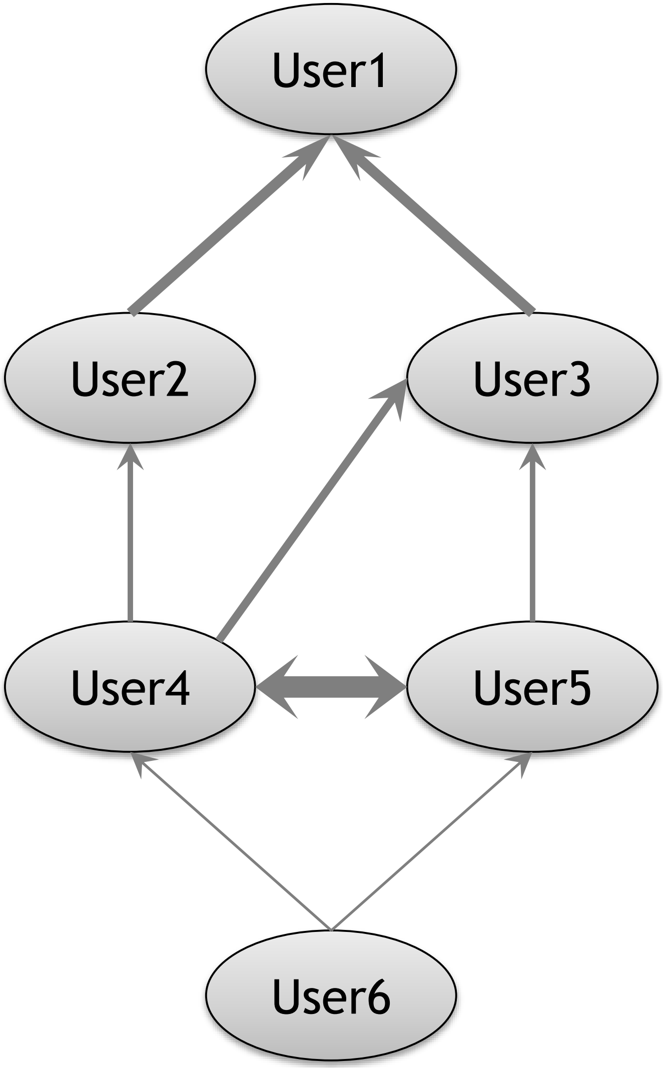
Re: The CVV used here, I believe, isn't the one printed on the back of the card. I beli

Re: Which is one of several reasons why I only have Credit Cards.

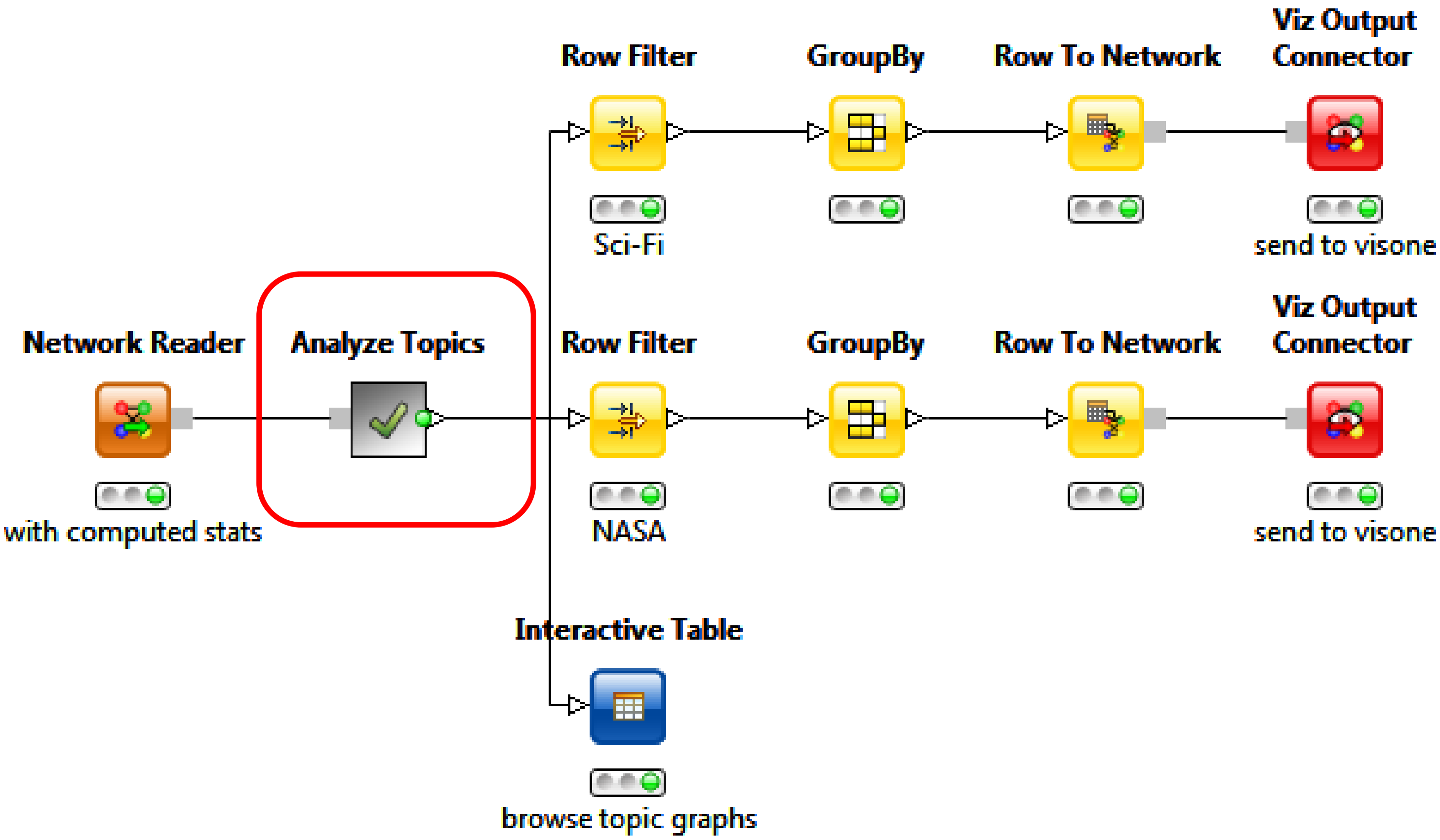
Re: My cell phone is has NFC and it is able to scan one of my credit cards for a decent

Re: Paget's firm has been working on a more sophisticated fix: a credit-card-shaped protecti

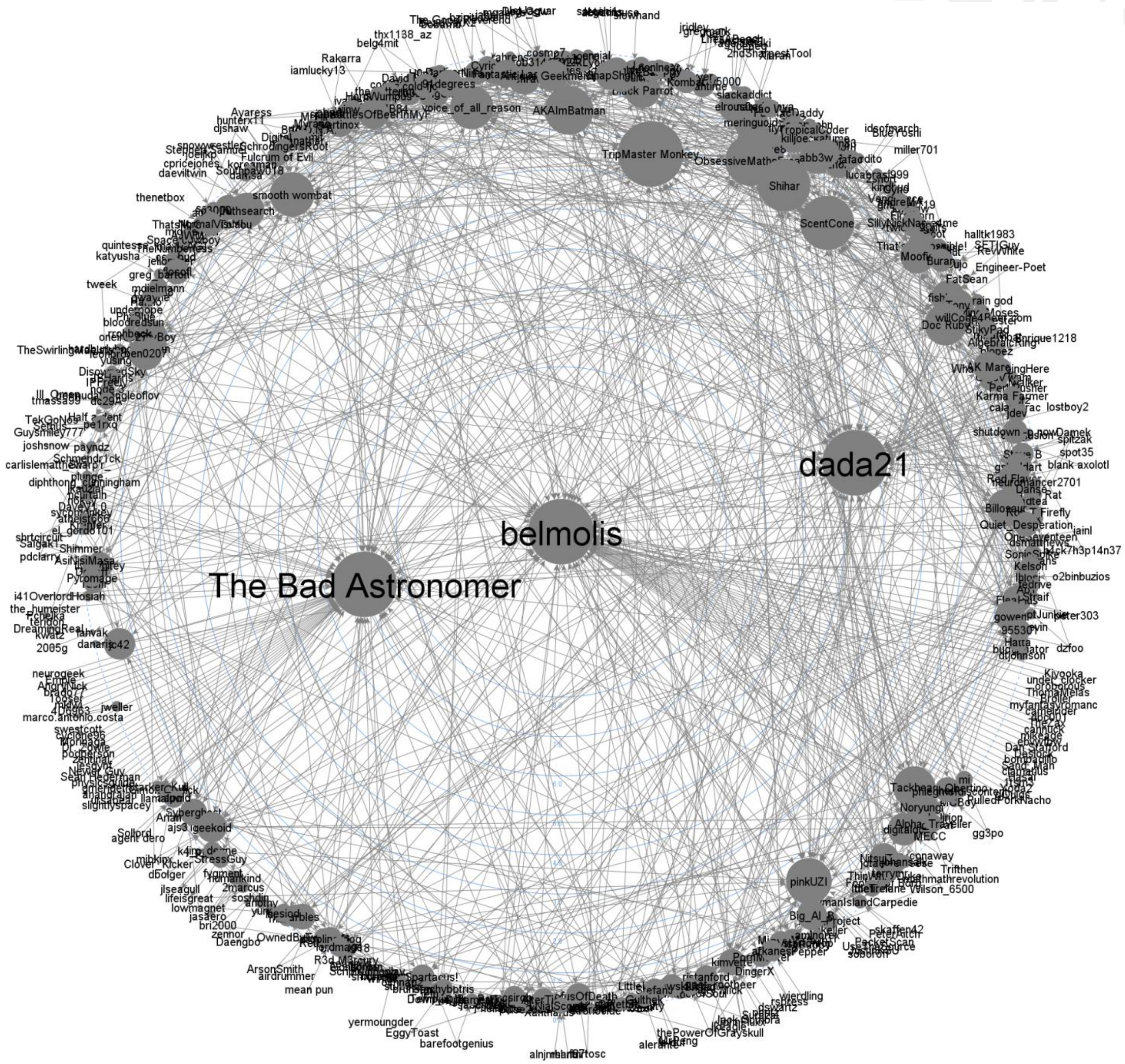
1 reply beneath your current threshold.



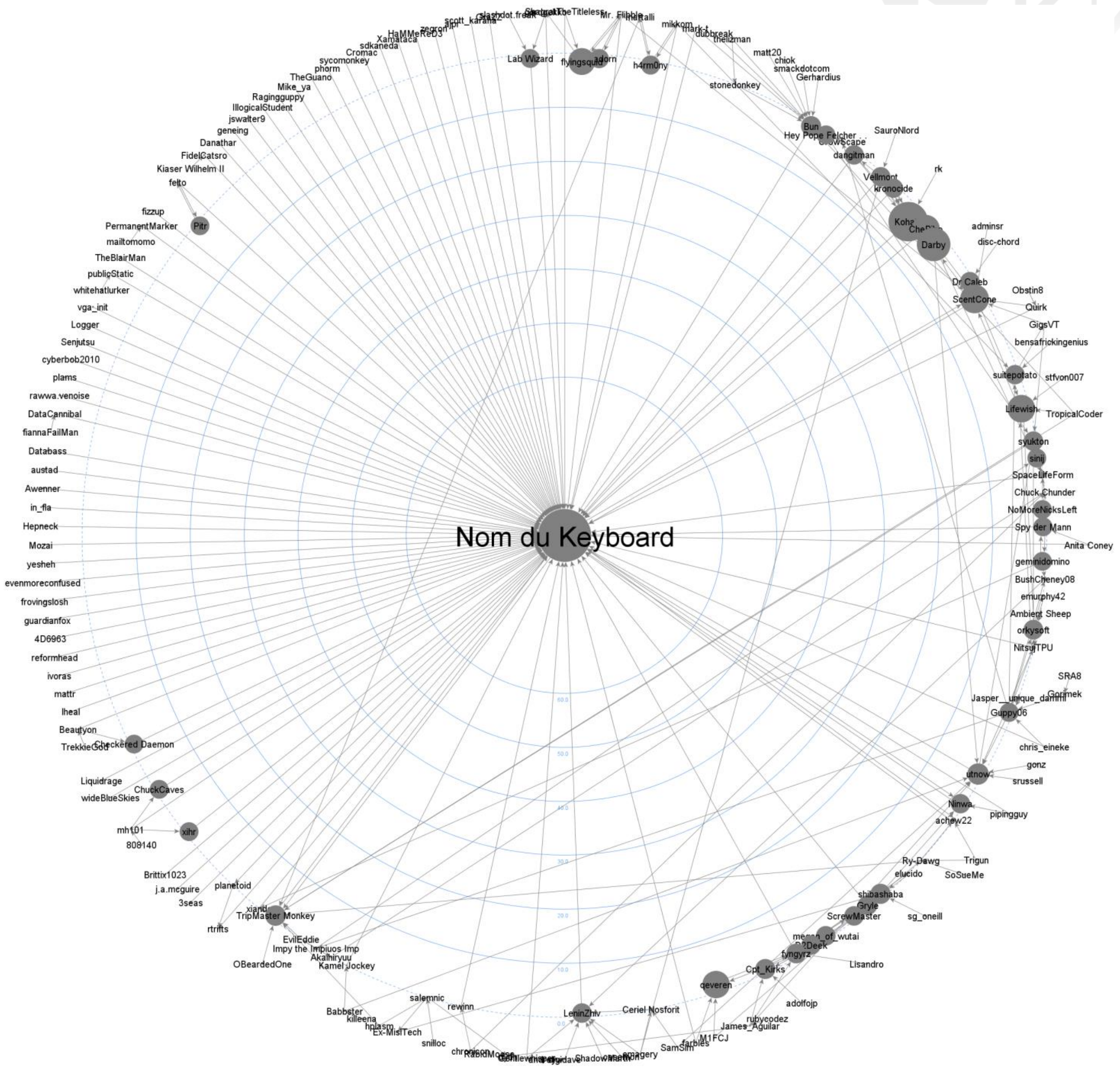
Topic Graphs



Topic Graph: NASA



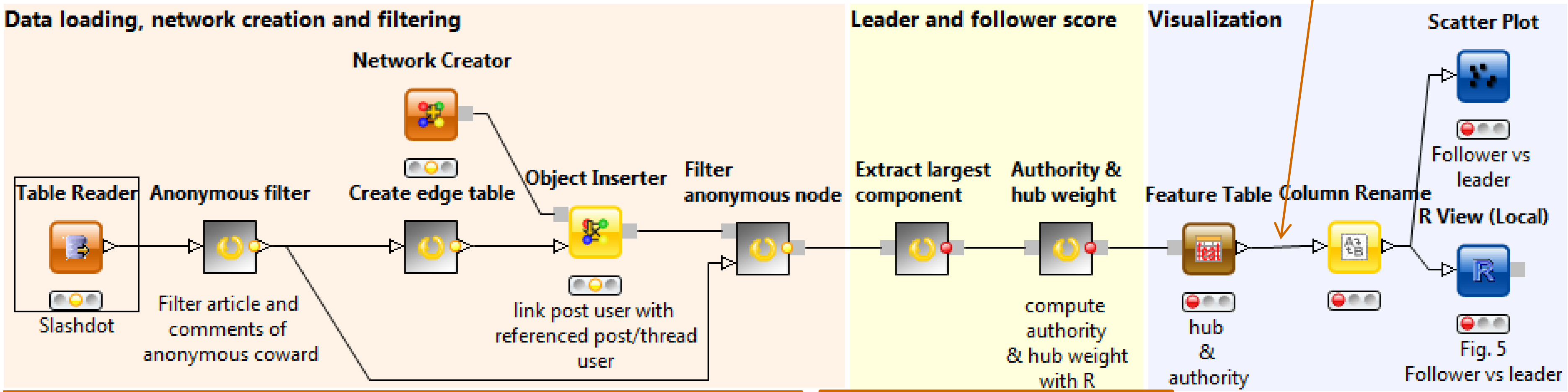
Topic Graph: Sci-Fi



Hubs & Authorities

- Hubs = Followers
- Authorities = Leaders

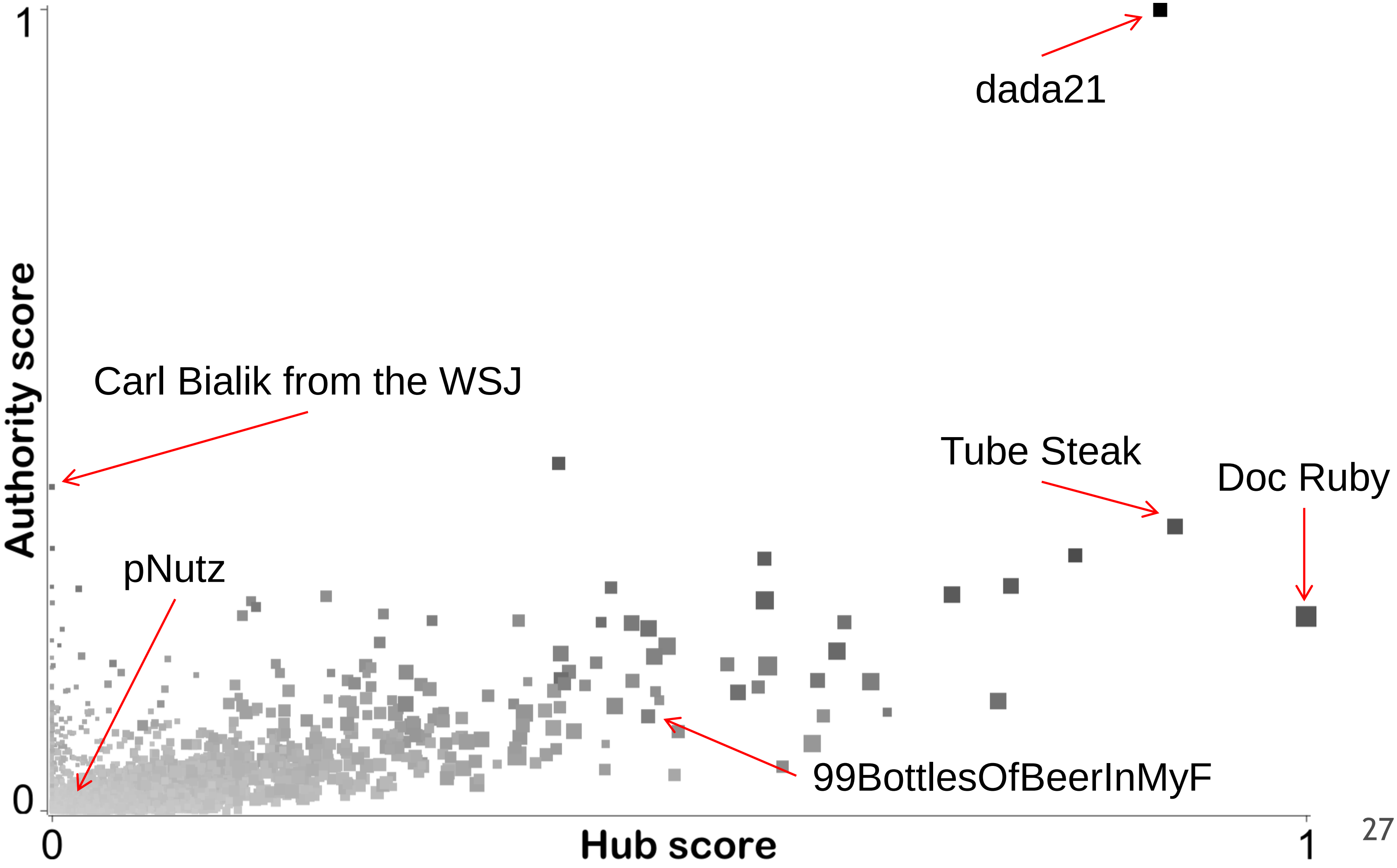
Users with hub and authority weights and other features


Data loading, network creation and filtering
Network Creator
Table Reader
Slashdot
Anonymous filter
Filter article and comments of anonymous coward
Create edge table
Object Inserter
link post user with referenced post/thread user
Filter anonymous node
Leader and follower score
Extract largest component
Authority & hub weight
compute authority & hub weight with R
Visualization
Feature Table
hub & authority
Column Rename
Scatter Plot
Follower vs leader
R View (Local)
Fig. 5 Follower vs leader

Filtering anonymous users and creating network

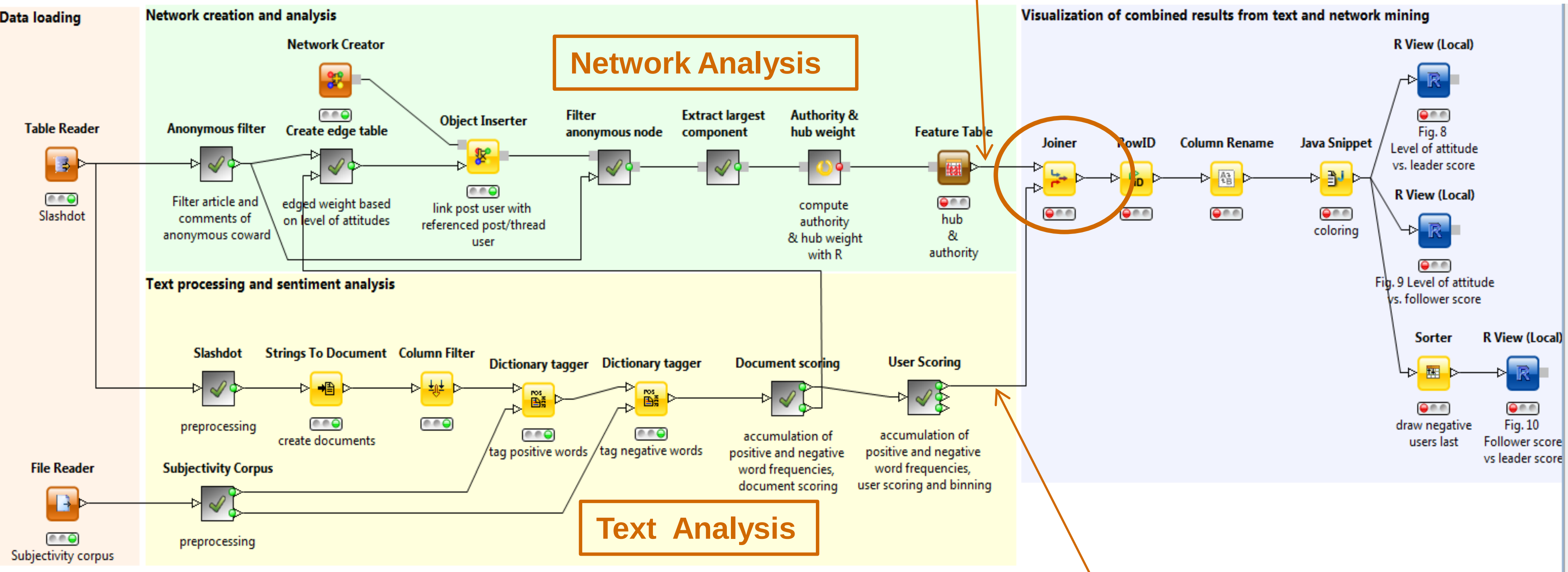
Centrality index to define hub weight and authority weight

Hubs & Authorities



KNIME: Bringing it all together

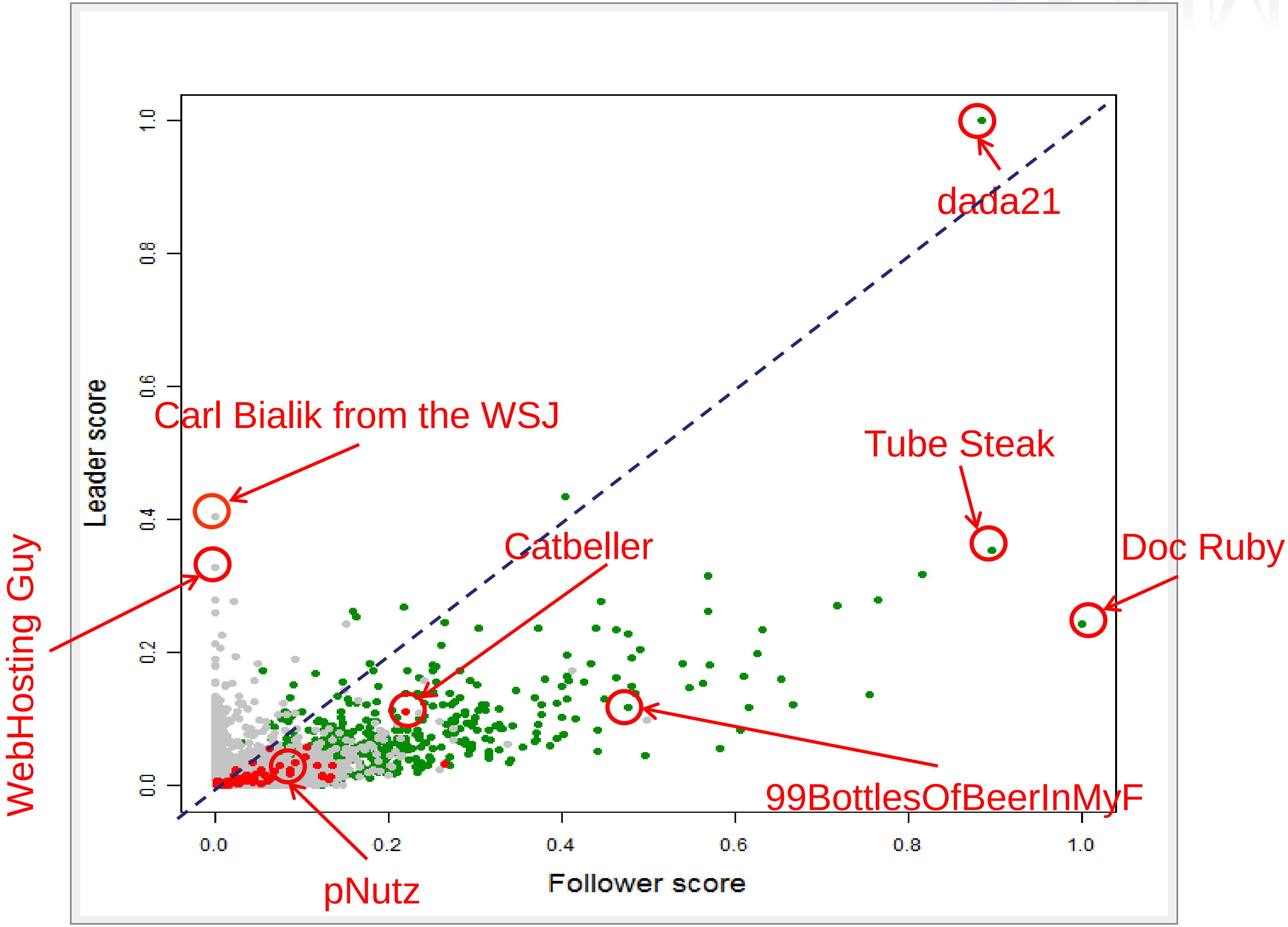
Users with hub and authority weights and other features


Data loading
Table Reader
Slashdot
File Reader
Subjectivity corpus
Network creation and analysis
Network Creator
Anonymous filter
Filter article and comments of anonymous coward
Create edge table
edged weight based on level of attitudes
Object Inserter
link post user with referenced post/thread user
Filter anonymous node
Extract largest component
Authority & hub weight
compute authority & hub weight with R
Feature Table
hub & authority
Text processing and sentiment analysis
Slashdot
preprocessing
Strings To Document
create documents
Column Filter
Dictionary tagger
tag positive words
Dictionary tagger
tag negative words
Document scoring
accumulation of positive and negative word frequencies, document scoring
User Scoring
accumulation of positive and negative word frequencies, user scoring and binning
Subjectivity Corpus
preprocessing
Visualization of combined results from text and network mining
Joiner
owID
Column Rename
Java Snippet
coloring
R View (Local)
Fig. 8 Level of attitude vs. leader score
R View (Local)
Fig. 9 Level of attitude vs. follower score
Sorter
draw negative users last
R View (Local)
Fig. 10 Follower score vs leader score

Network Analysis

Text Analysis

Users bins: positive, negative, neutral


Leader score
0.0
0.2
0.4
0.6
0.8
1.0
Follower score
0.0
0.2
0.4
0.6
0.8
1.0
dada21
Carl Bialik from the WSJ
Tube Steak
Catbeller
Doc Ruby
WebHosting Guy
99BottlesOfBeerInMyF
pNutz

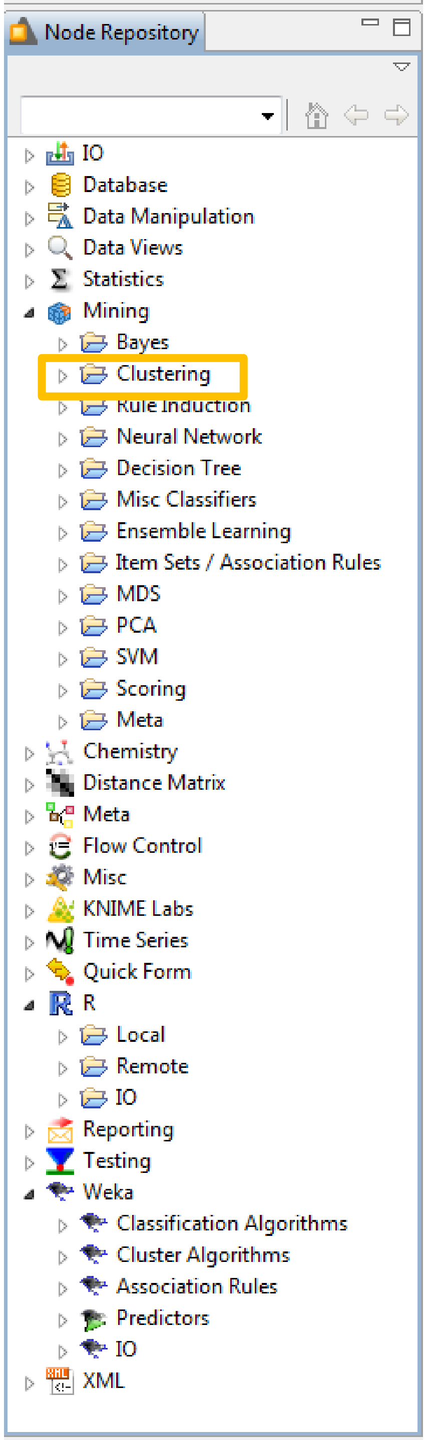
What we have found ...

- The **positive** leaders
- The **neutral** leaders
- The **negative** leaders
- The inactive users

What identifies each group?

How do I identify a new user?

How do I handle each user?



Node Repository

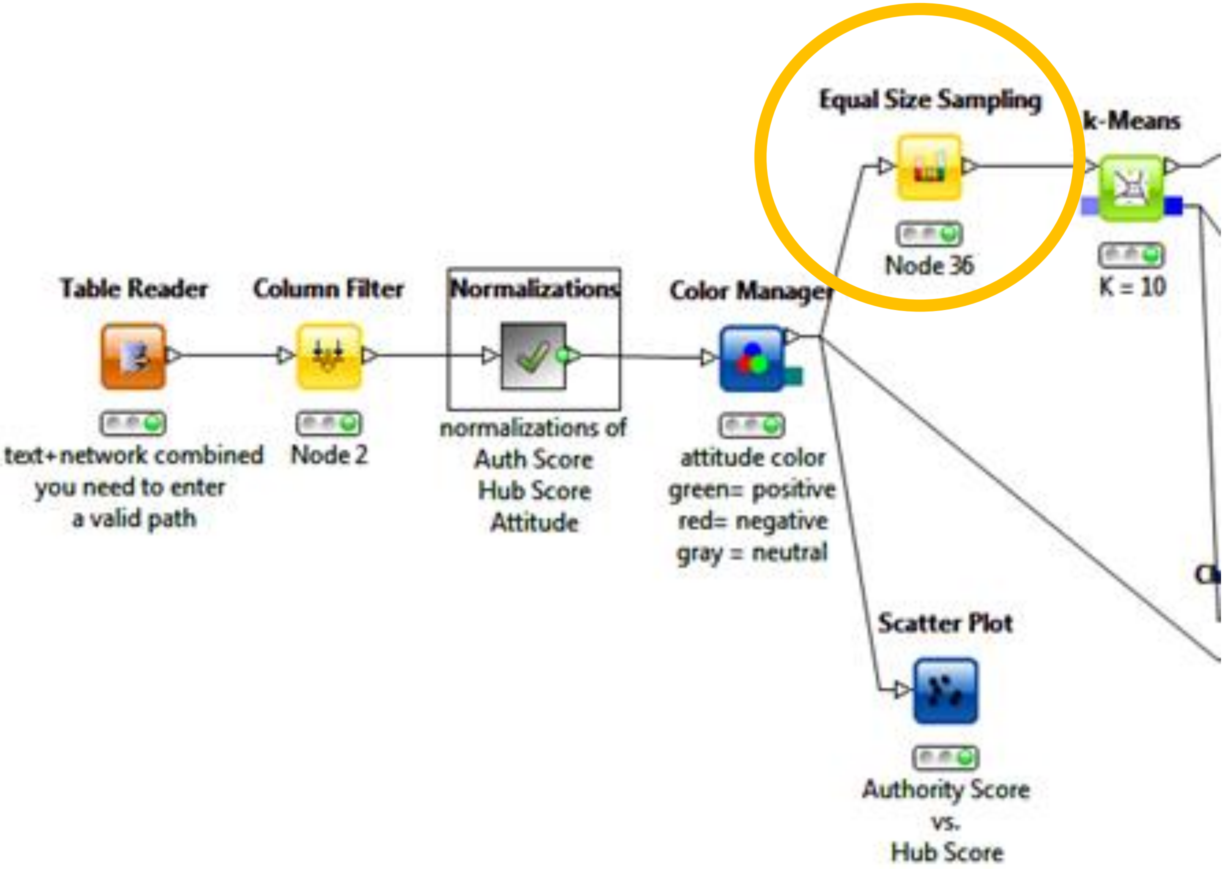
- IO
- Database
- Data Manipulation
- Data Views
- Statistics
- Mining
  - Bayes
  - Clustering
  - Rule Induction
  - Neural Network
  - Decision Tree
  - Misc Classifiers
  - Ensemble Learning
  - Item Sets / Association Rules
  - MDS
  - PCA
  - SVM
  - Scoring
  - Meta
- Chemistry
- Distance Matrix
- Meta
- Flow Control
- Misc
- KNIME Labs
- Time Series
- Quick Form
- R
  - Local
  - Remote
  - IO
- Reporting
- Testing
- Weka
  - Classification Algorithms
  - Cluster Algorithms
  - Association Rules
  - Predictors
  - IO
- XML

Why Clustering?

- No a priori knowledge (not even on a subset of users)
- Prediction and interpretation capabilities required

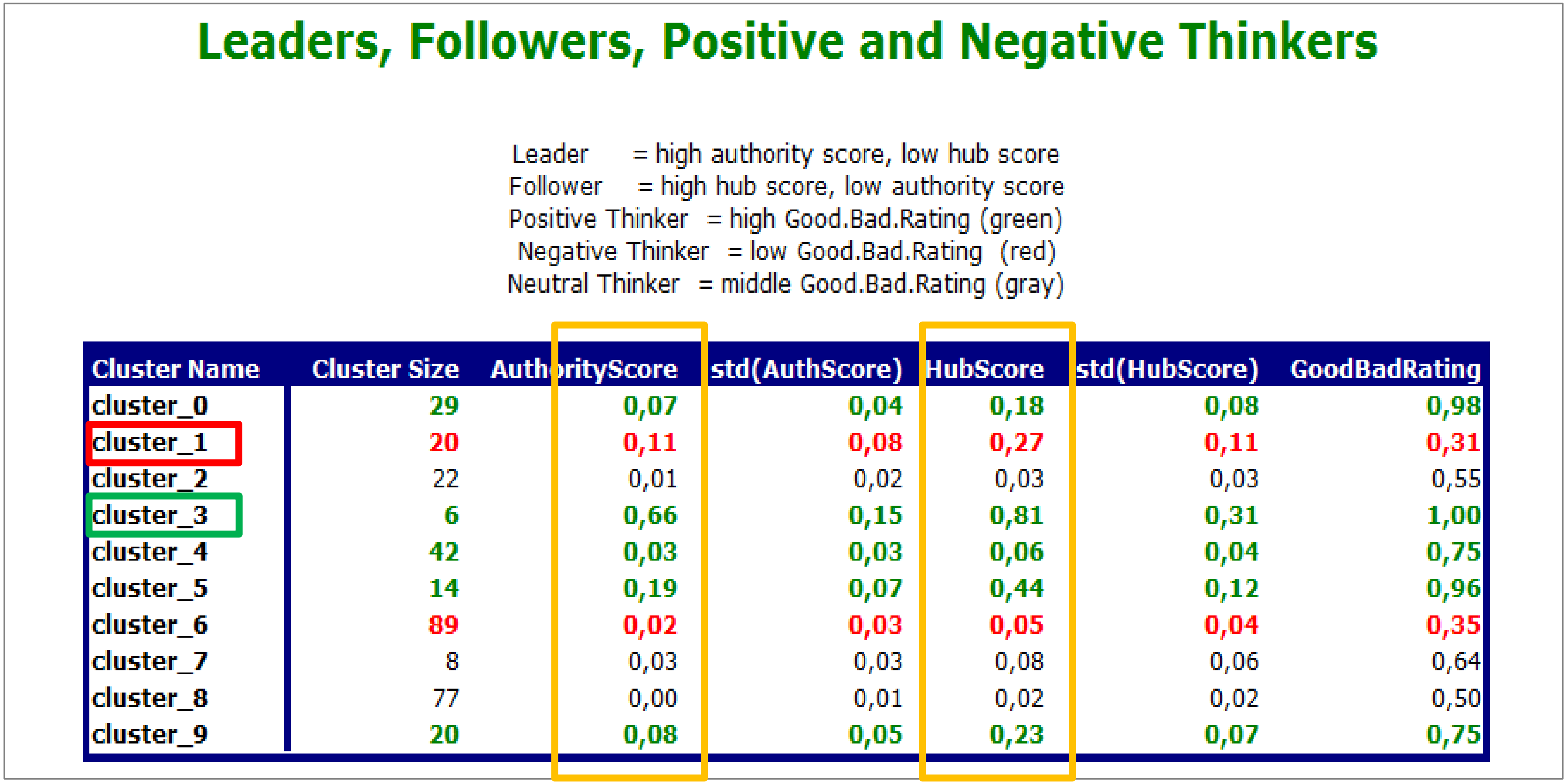
k-Means algorithm

Re-sampling the Training Set



k = 10

The k-Means Clusters

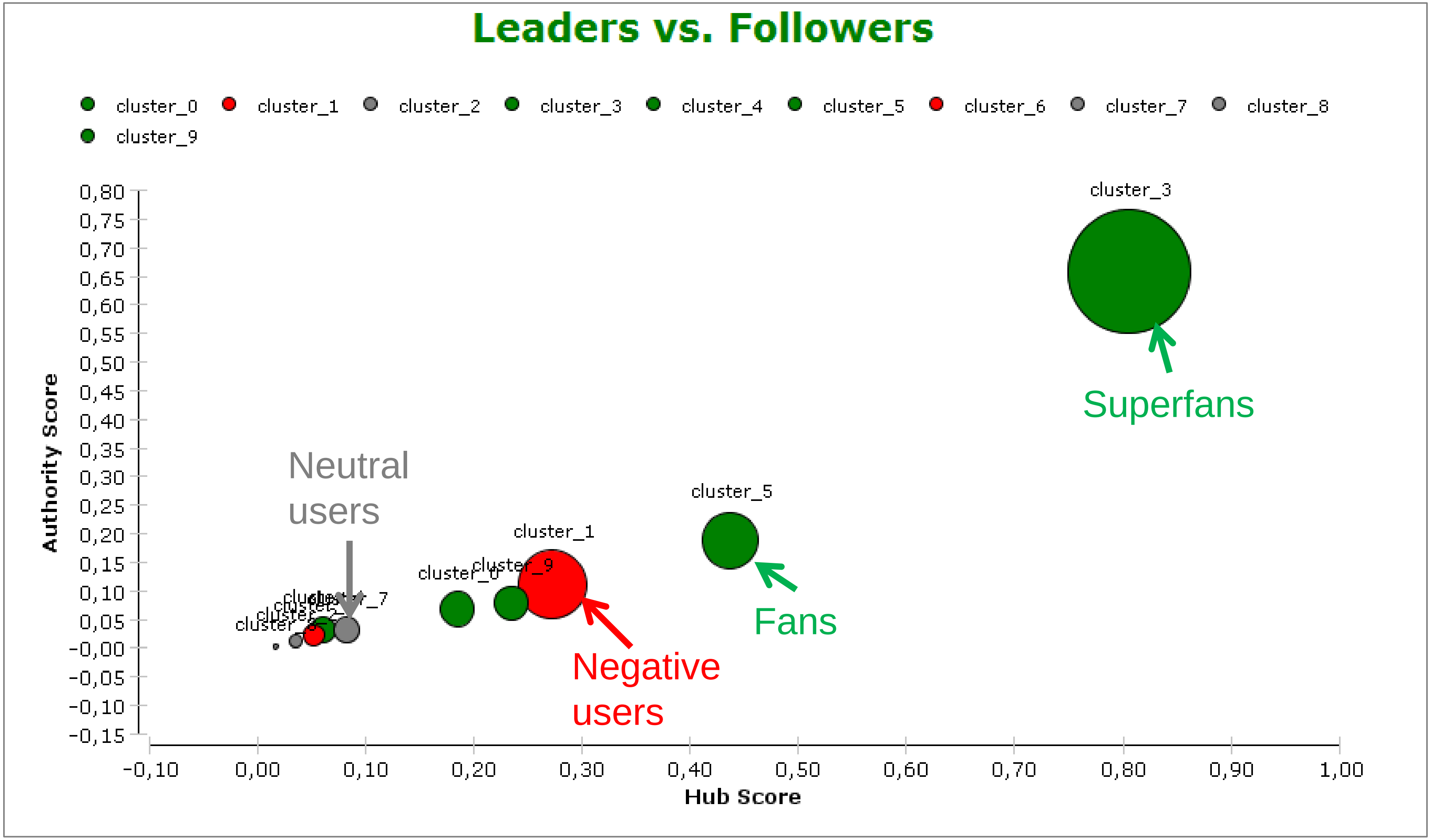


Leaders, Followers, Positive and Negative Thinkers

Leader = high authority score, low hub score
Follower = high hub score, low authority score
Positive Thinker = high Good.Bad.Rating (green)
Negative Thinker = low Good.Bad.Rating (red)
Neutral Thinker = middle Good.Bad.Rating (gray)

Cluster Name	Cluster Size	AuthorityScore	std(AuthScore)	HubScore	std(HubScore)	GoodBadRating
cluster_0	29	0,07	0,04	0,18	0,08	0,98
cluster_1	20	0,11	0,08	0,27	0,11	0,31
cluster_2	22	0,01	0,02	0,03	0,03	0,55
cluster_3	6	0,66	0,15	0,81	0,31	1,00
cluster_4	42	0,03	0,03	0,06	0,04	0,75
cluster_5	14	0,19	0,07	0,44	0,12	0,96
cluster_6	89	0,02	0,03	0,05	0,04	0,35
cluster_7	8	0,03	0,03	0,08	0,06	0,64
cluster_8	77	0,00	0,01	0,02	0,02	0,50
cluster_9	20	0,08	0,05	0,23	0,07	0,75

The k-Means Clusters



Superfans

Fans

Neutral users

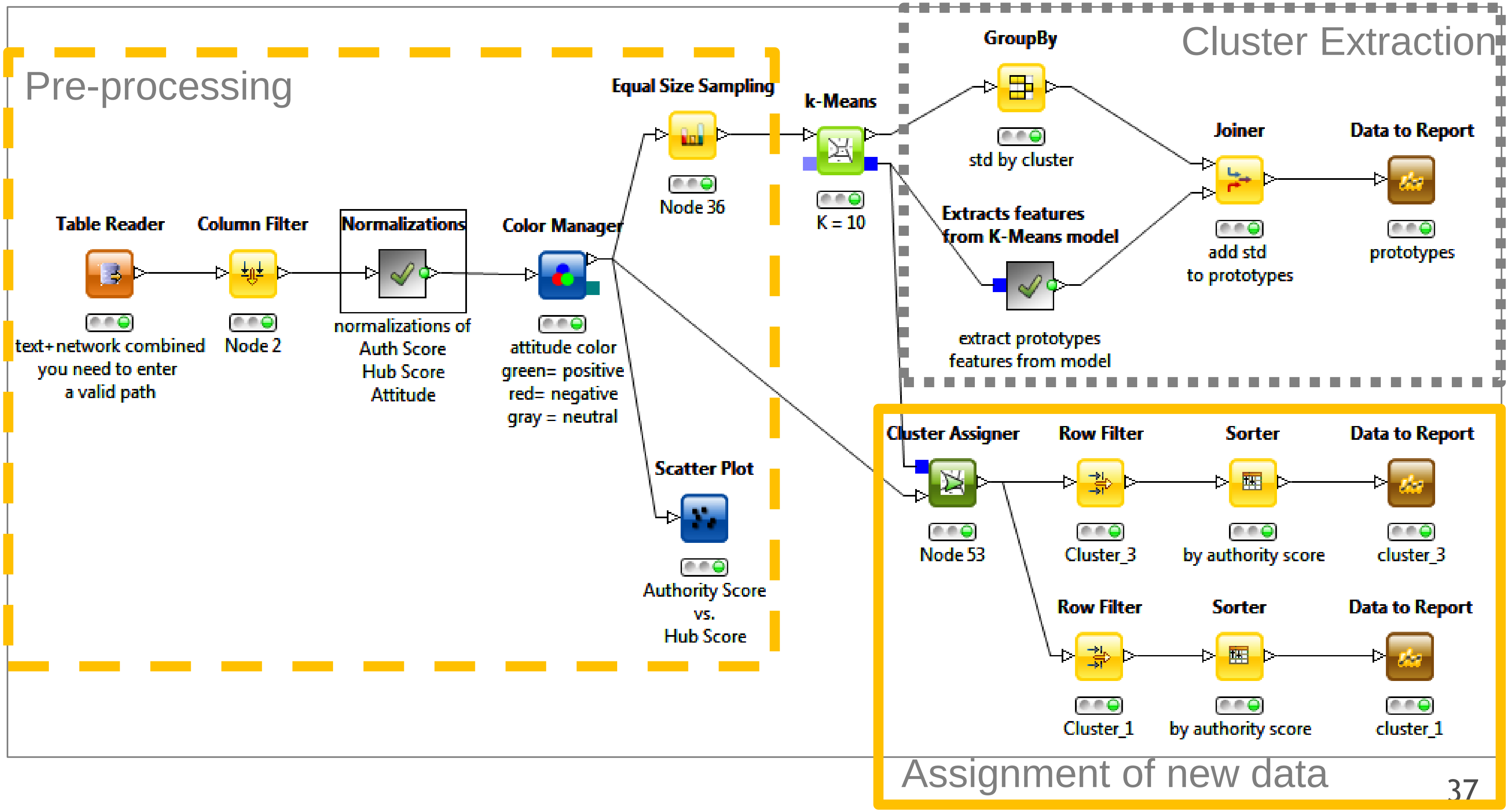
Negative users

Additional Discoveries

- There are only very few real leaders!

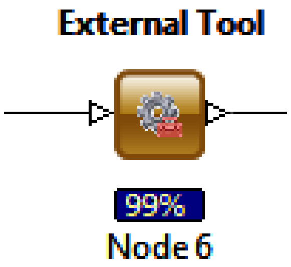
  Authority and hub scores identify active participants rather than leaders.
- Superfans can be found in cluster_3
- Negative and (sigh!) active users are collected in cluster_1.
- Neutral users are usually inactive (cluster_2, cluster_7, and cluster_8)
- Positive users with different degrees of activity are scattered across the remaining clusters.

The operational Workflow



Notes

- **MPQA** Corpus: publicly available Subjectivity Lexicon (http://www.cs.pitt.edu/mpqa/lexicons.html)
- User Characterization is Sum -> **Mean**
- **NLP**: No sentence splitting, no negation identification.
- For a more refined syntax-based sentiment analysis -> „**External Tool**“ node



External Tool Node

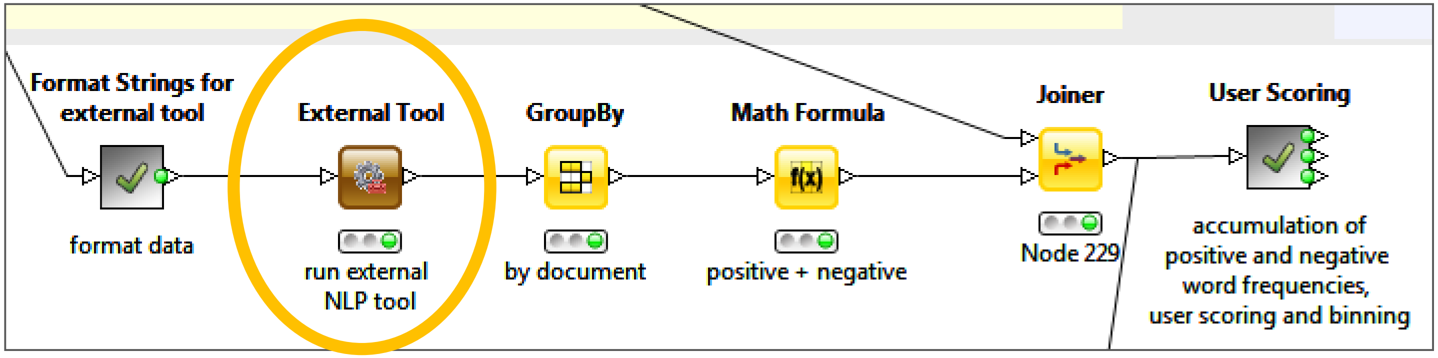
The „External Tool“ node executes **any** external program from command line

1. Writes input data to an input file
2. Calls Tool to run on input file and command line options and to write results to output file
3. Reads output file and presents data at output port

Alternative Sentiment Analysis

Free non-interactive Command Line running Tools for Sentiment Analysis not found

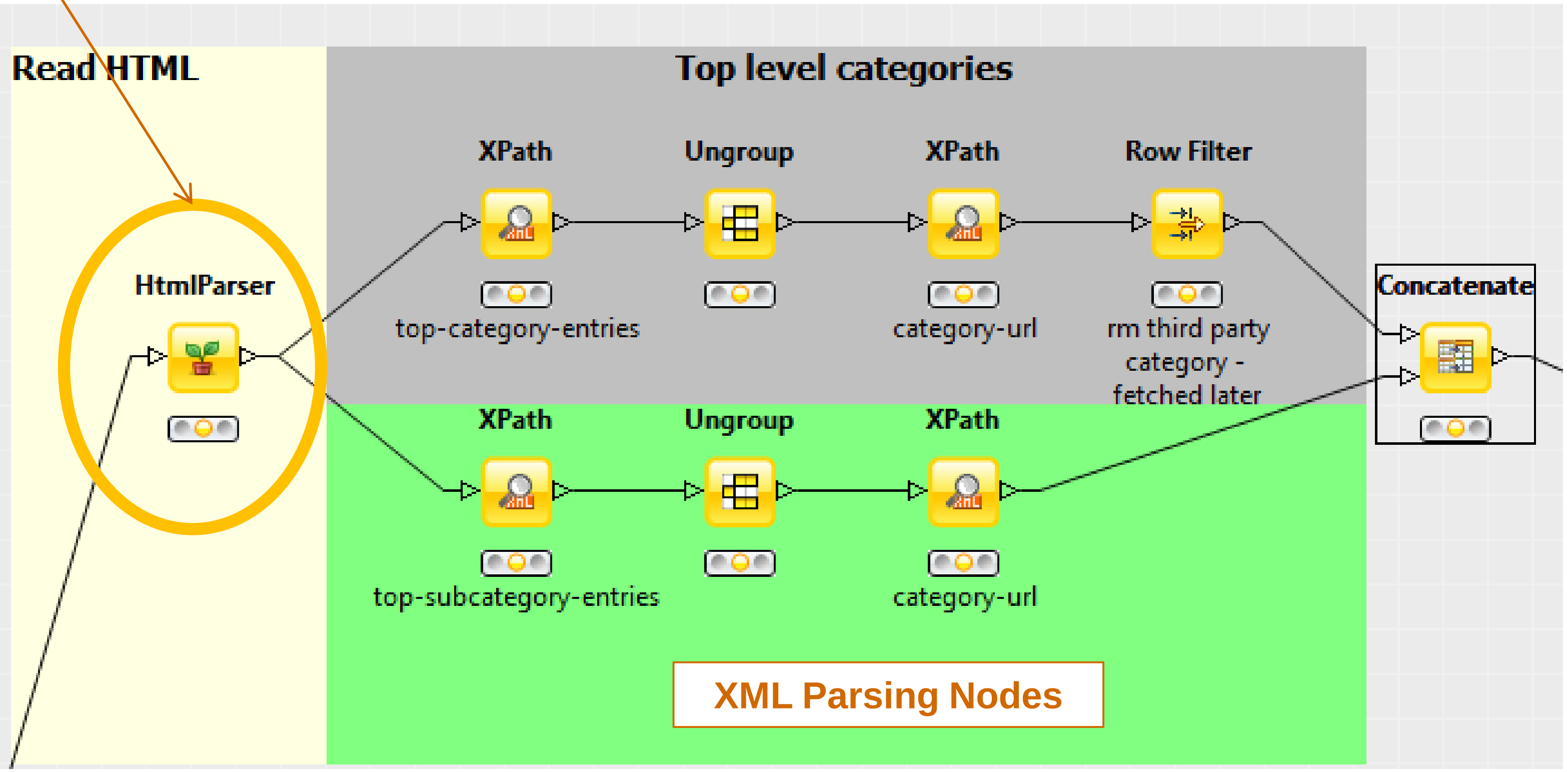
SentiStrength v2.2 (still interactive)



External Tool and Generic Web Service Client

Web Crawling Workflow

Community Web Crawler Node



XML Parsing Nodes

Next Steps

- Integrate topic information
- Integrate user demographic and behavioural information
- Discover [time series] patterns for early detection of negative users and superfans
- Try other techniques, maybe even on manually segmented data, to discover new user segments

Where do I find more?

Whitepaper: www.knime.org/white-papers

Includes Complete Workflows + Data

- text mining
- network mining
- combined analysis

(note the above 3 process huge data and require 16G memory)

- clustering

Open Source Software: KNIME www.knime.com