



**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



The screenshot displays the KNIME platform interface with a workflow and several data visualization windows.

**Workflow:** The workflow starts with a **File Reader** node (Read NCI), followed by a **Row Filter** node (Outlier), a **Domain Calculator** node (Node 47), and a **Color Manager** node (Color by class). The workflow then branches into two parallel paths. The left path consists of **MetaNode 1:1** (Node 35), **MoSS** (Node 59), and **Column Filter** (Node 94). The right path consists of **MetaNode 1:1** (Node 35), **Parallel Coordinates**, **Interactive Table** (Node 87), and another **Interactive Table** (Node 88). Both paths converge at a **Neighbor** node, which then feeds into a **Neighborhood Visualizer** (Node 82).

**Scatter Plot (#49) - Scatterplot:** A scatter plot showing data points colored by class. The X-axis ranges from 100.0 to 940.0, and the Y-axis ranges from 27.0 to 361.0.

**File Reader Node Description:** 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.

**Interactive Table (#87) - Table View (58 x 7):**

Key	sm:Smiles	Support in Focus (rel)	Support in Complement (rel)	Atom Num.	Bar
0	<chem>Nc1ccc2c(c1)oc(=O)c2</chem>	14.42%	0.02%	19	20
1	<chem>Nc1ccc2c(c1)oc(=O)c2</chem>	14.42%	0.03%	18	19

**Interactive Table (#88) - Table View (51 x 2):**

Key	sm:Smiles	Class
624486	<chem>Nc1ccc2c(c1)oc(=O)c2</chem>	Cl
624487	<chem>Nc1ccc2c(c1)oc(=O)c2</chem>	Cl
633772	<chem>Nc1ccc2c(c1)oc(=O)c2</chem>	Cl

**Neighborhood Visualizer (#82) - Interactive Neighborgrams:** A visualization showing neighborgrams for different classes. The X-axis is labeled 'ClusterQuality' and the Y-axis is labeled 'Key'. The visualization shows three columns: 'Unary FP', 'Atompar FP', and 'VolSurf'. The data points are colored by class, and the visualization shows the distribution of neighborgrams for each class.



# 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 loads and integrates data from diverse data sources:

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

## File Reader



Excel import

## Database Connector



Node 0:1:8

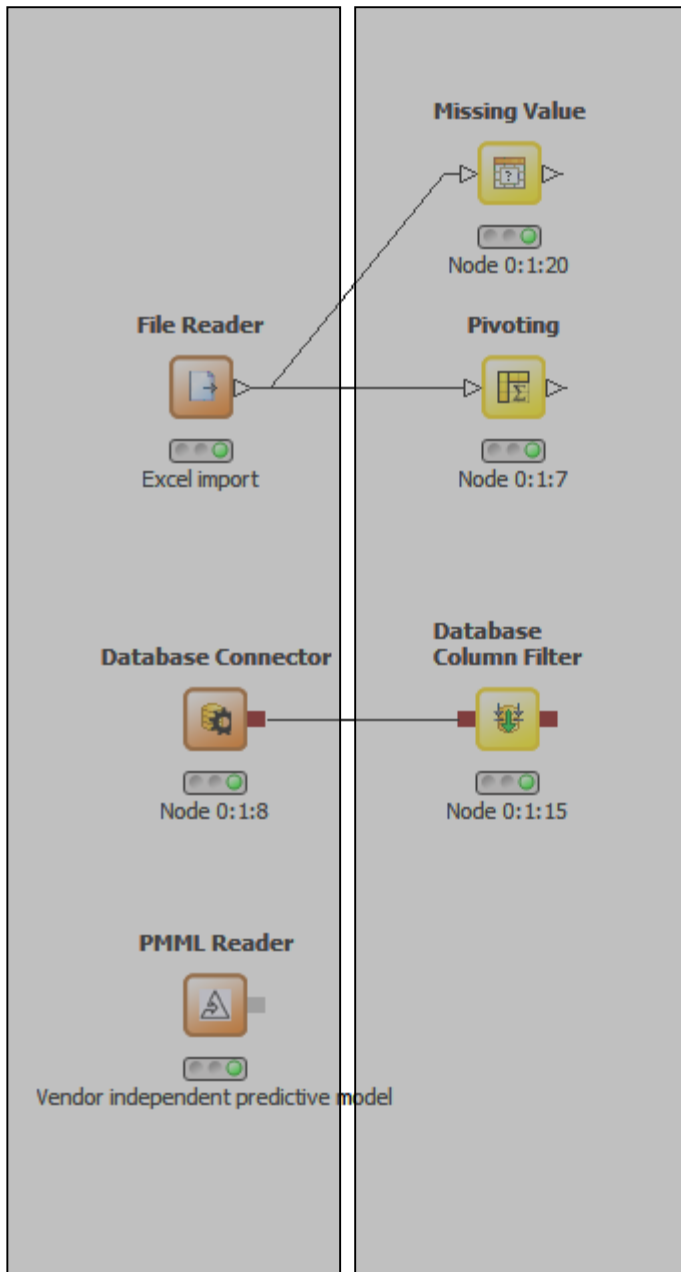
## PMML Reader

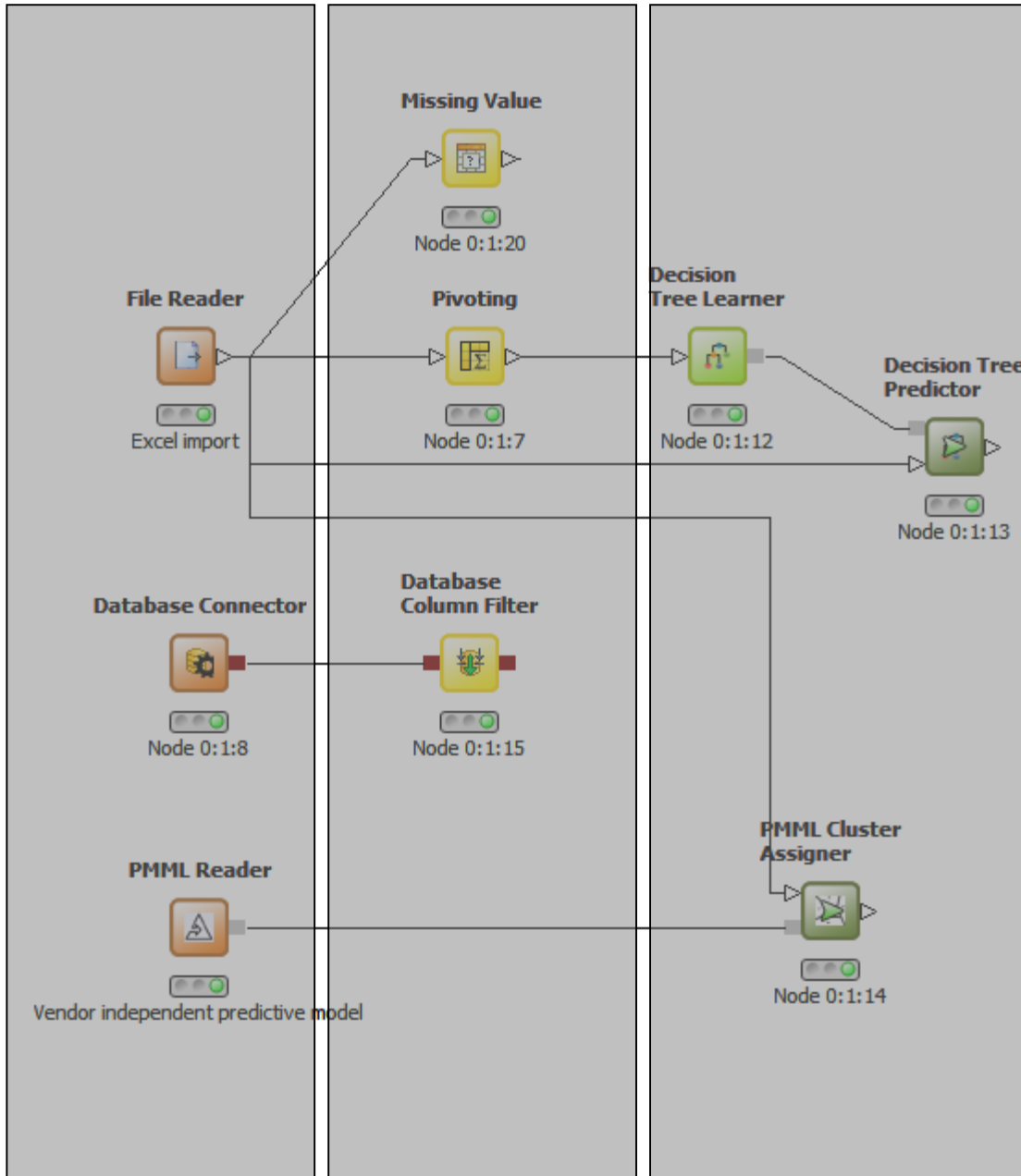


Vendor independent predictive model

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

- Data preprocessing
- Data fusion
- Data transformation

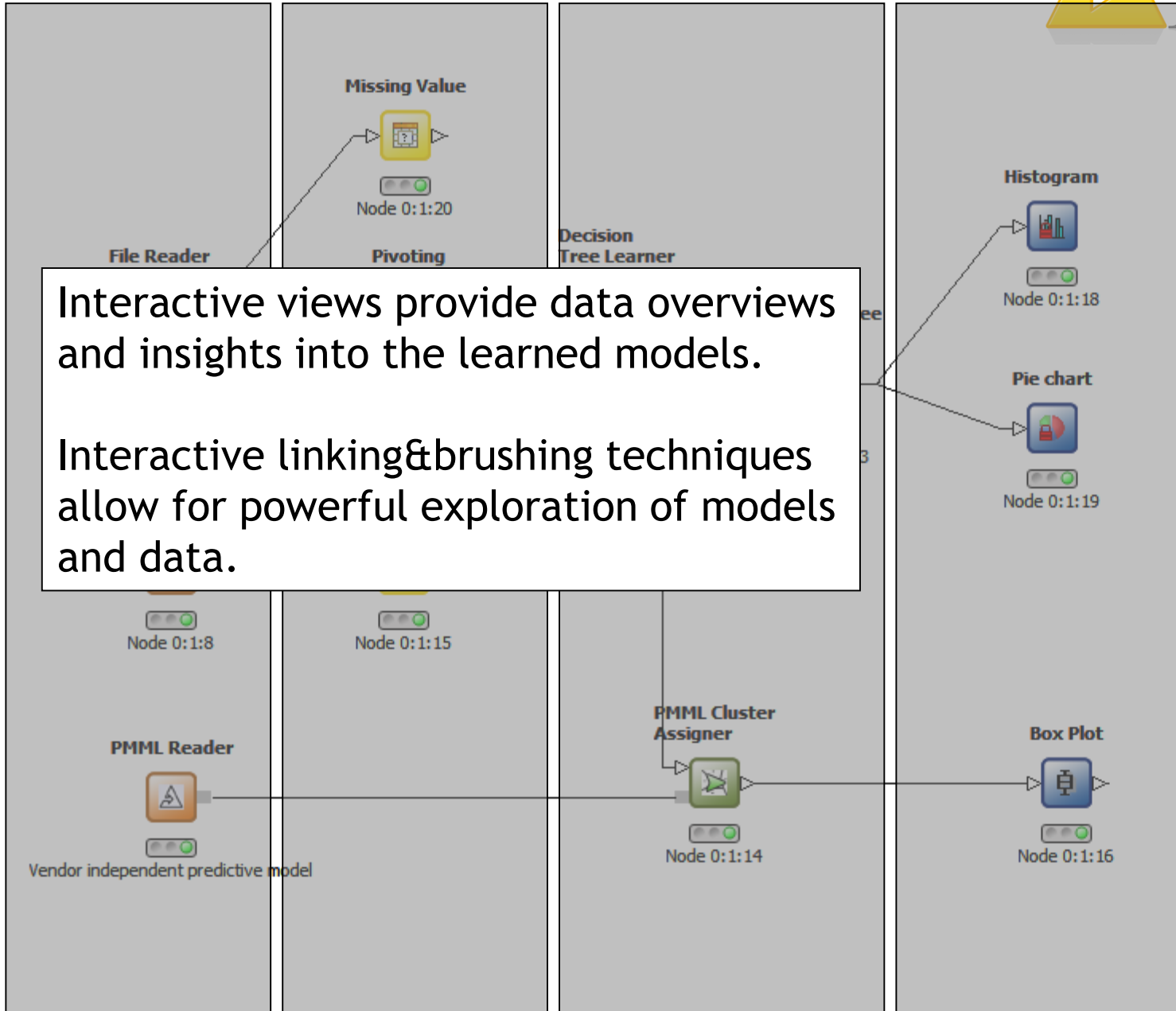




In addition to standard data mining techniques, KNIME adds cutting edge data analysis algorithms. (...thanks to its academic roots)

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

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





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

PMML Reader



Vendor independent predictive model

Assigner



Node 0:1:14

Box Plot



Node 0:1:16

R Snippet (Local)



Node 0:1:21

Logistic



Node 0:1:22

Weka Predictor



Node 0:1:23

# 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

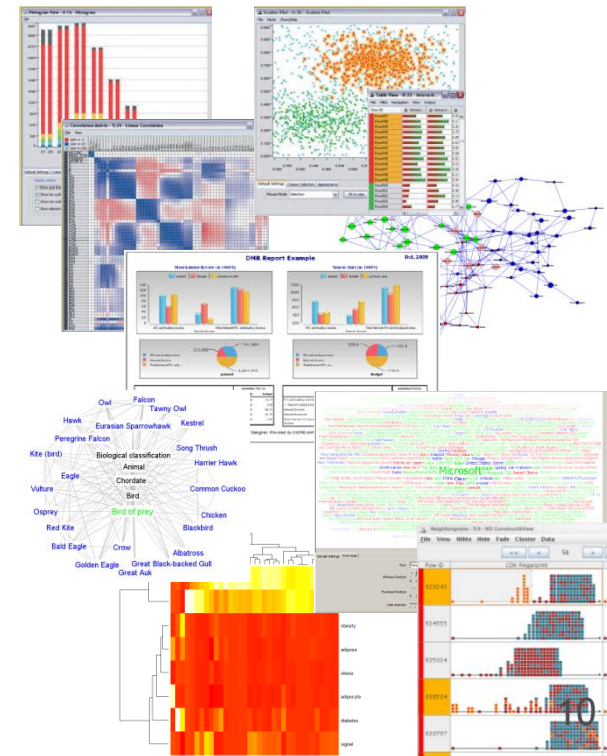
PMML Support

Social Media Analysis

Business Intelligence

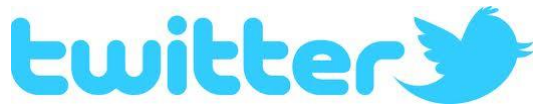
Community Nodes

3rd Party Nodes



# Social Media Data

Water Water Everywhere, and not a drop to drink



# 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



**Slashdot**  Library

stories

recent

popular

ask slashdot

book reviews

games

idle

yro

hardware

linux

management

mobile

science

security

storage

**Shmoocoon 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.

Sparrowsrevolution writes with this excerpt from a Forbes piece recounting a scary demo at the just-ended [Shmoocoon](#):

"[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 which

[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.

Post

## Basic Numbers:

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

Comments

## Case Study Example: Slashdot

- Very rich data sources about customers !
- We want to establish:
  - How users feel about the discussed topic
  - Whether it matters how users feel
  - A more general abstraction of the results

Sentiment Analysis

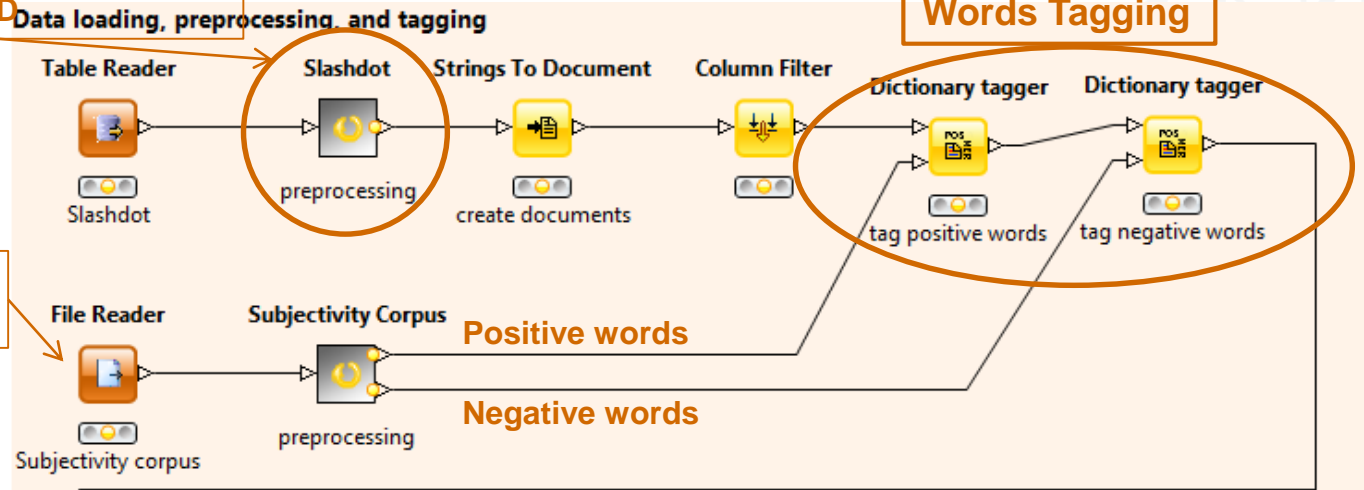
Network Analytics

Clustering

# Sentiment Analysis



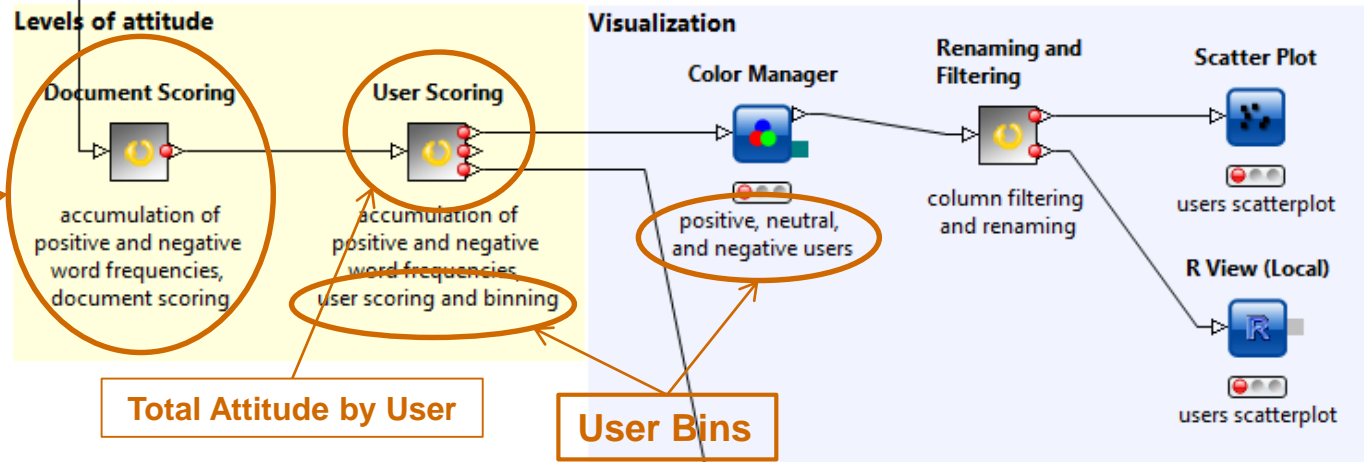
Remove anonymous users,  
group by PostID



**Words Tagging**

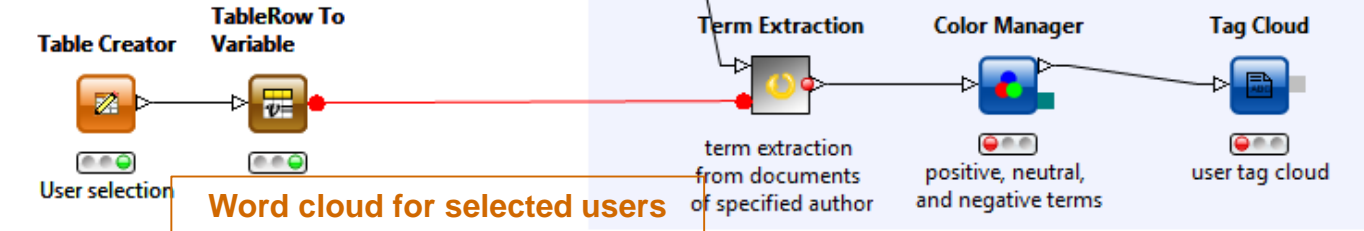
**MPQA  
Corpus**

**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

burners  
 bus copier bum-bailey  
 400-year buses **coward** **drive** conversion buggler  
**able** cad coxcomb drivers gob-kissing dregs control briefly  
 access calf **craven** dvd **golden** krempf gleeking dread-bolted contempting breath  
 action campaigns **creative** dwarf google lagerlout **misery** **kiss** giggling **doubts** **contempt** brain  
 addition canker **criminal** earth gormless land monkeys pink middle jkct fungus **doubtful** consequences borders  
 advertisement central criminals edge greens lawyer **monster** pisshead responses pillock mid-day **jerk** frothy **disease** conception boggish  
 agency character **criticism** edit grimy **leech** monument plaintiff rest sporen response pick meta-stupid **irony** **foul** directors completely **bloody**  
 churlish criticizes effluvia grotty lemon **moral** plonker **revulsion** spread uncensored spell request phone mer iraq foot-licking dimension **complaints** **blood**  
**embarrassment** guard lends moulitsas ponce rhythm squalid unknowing zukunft unabbreviated speech representation **personality** mention **intelligent** **foolish**  
 guardsman lepers mutant posters **ridiculous** **stale** **unnecessary** zâfâ'niga on-line youth twit species remind people medium intellect **fool** dewberry blinking  
 naff **prison** rock-hard statement unpleasantness ã,ã sd1901020223 **copyright** measure worm trite spd **reform** penintentiary media institution foal  
**profane** rocks statements **unprofitable** **advice** **sour** **national** **advertising** **zaw** mass **wide** translation spattering **recognition** **pathetic** fly  
 leprous **nasty** professional root stefan untrue association **spot** **marketing** weight trans-stupid source recipient past instance  
 profoundest sd4511428251 **stench** users attention uwg **spots** **nickel** **robbery** grannie **weed** tosser **sorrow** reality paper  
 guy nausea **protection** sd4750433211 stock utilization **bad** world **weasel** tiny-fisted **sore** **real** own meatslapper **insight**  
**emptiness** half-wit life neutrons public self strip violators copiers **stupid** **little** german **weary** **threatening** sod readers opposite measures  
 nincompoop publicity self-important **strong** virtual country **worthless** **campaign** **law** filmwirtschaft **waste** themselves snake read opinion **meaningless**  
 cinema equivalent hawk link novella puke-drooling **senile** **stupidity** **void** due zkm example wallpaper tantrums smell reach oops **meaningful** insensate  
 evacuation history lives noxious punishment **serious** **success** volker economy **vulgar** sun slogan **rat** onion meaning inputs flirt-gill detentions blazing  
 circle crook-pated evidence hold local **objection** putrescent sex suck **vomit** **suitable** site rape ones maximum initiative flesh design competition bite  
 alias city curdled exchange holds loose **obnoxious** queasy **shame** **suicidal** singularity rancid one-handed matters **infringements** flat **desert** birth  
**appeal** clack-dish dank exchanges **hot** louisiana ocean question sheep rabidly oik material **inexperienced** flap-mouthed derailing compatibility  
 article clapper-clawed dankish experience human macon offal rabbit ogre markos industry **filthy** **depraved** company beliefs  
 artless clearing-up data eyes hurricane maggot offers mandatory **indignation** **filth** dense companies beasts  
 available clotpole day **fantasy** idea malformity ignoramus film dehydrated-rock-hard communicates beak  
 bafelsh cockered **deception** fate **ignoble** **fiend** **degenerate** comment base-court  
 bag cockup deer federation **deficient** com barf



# 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



**Slashdot**  Library

stories

recent

popular

ask slashdot

book reviews

games

idle

yro

cloud

hardware

linux

management

mobile

science


security

storage

---

**Shmoocoon 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.



Sparrowsrevolution writes with this excerpt from a Forbes piece recounting a scary demo at the just-ended [Shmoocoon](#):

"[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 which

[Reply to This](#) [Parent](#)

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

---

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

by Joice640k (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 actually

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

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

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

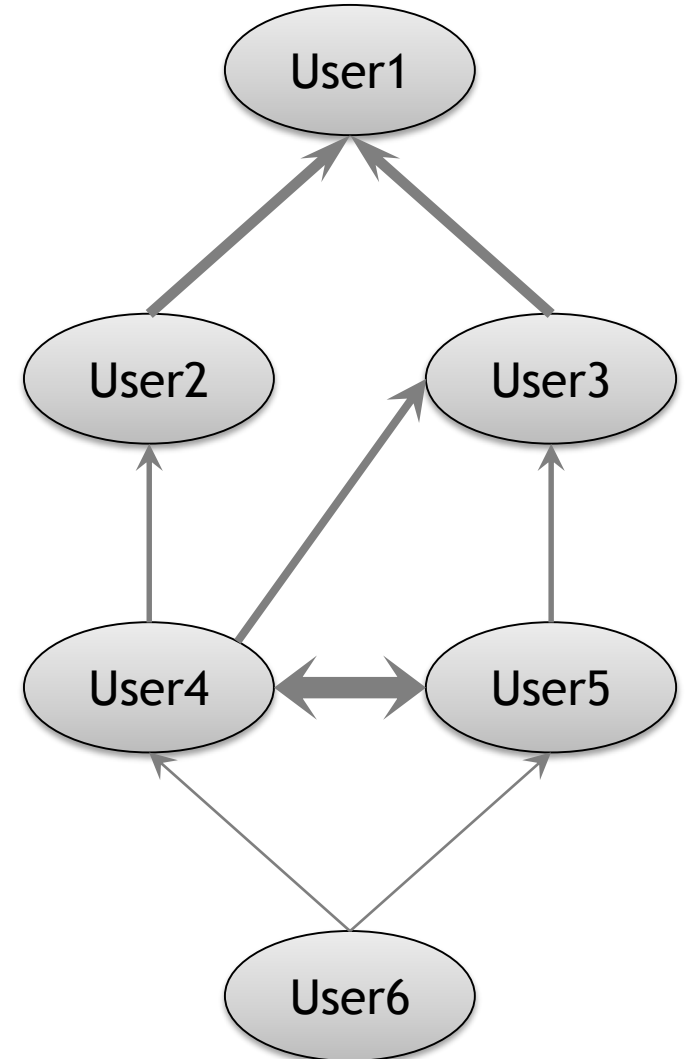
**Re:** The CVV used here, I believe, isn't the one printed on the back of the card. I believe

**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

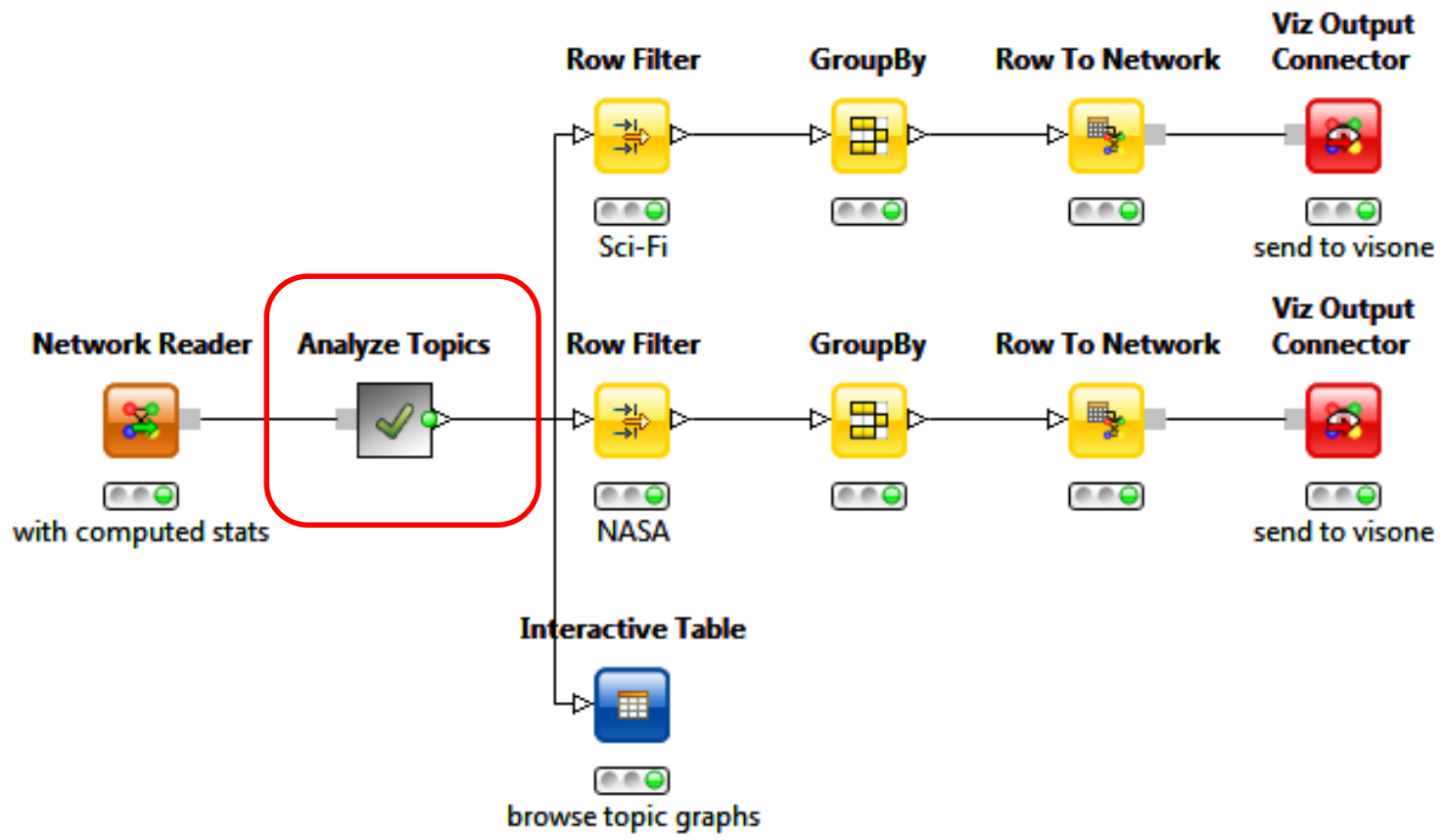
**Re:** Paget's firm has been working on a more sophisticated fix: a credit-card-shaped protective

**1 reply beneath your current threshold.**





# Topic Graphs







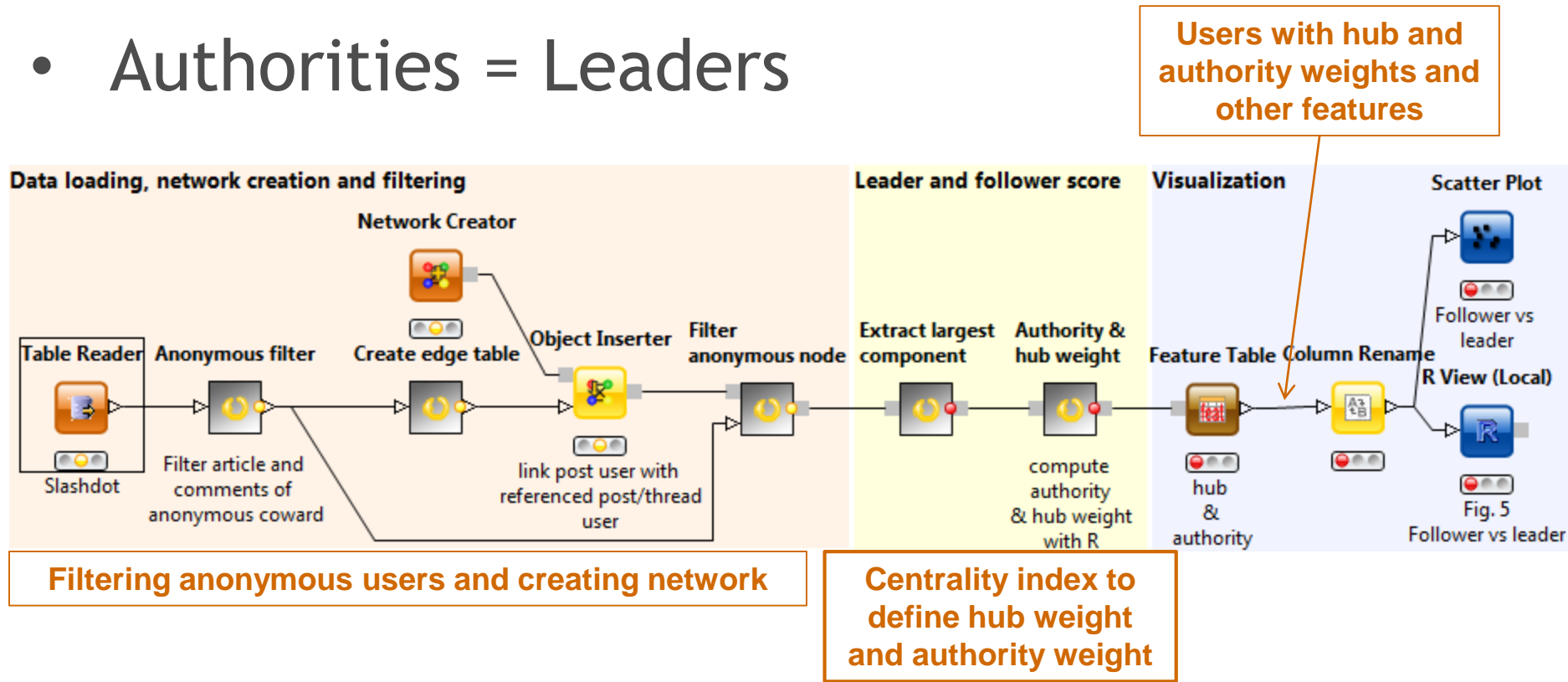






# Hubs & Authorities

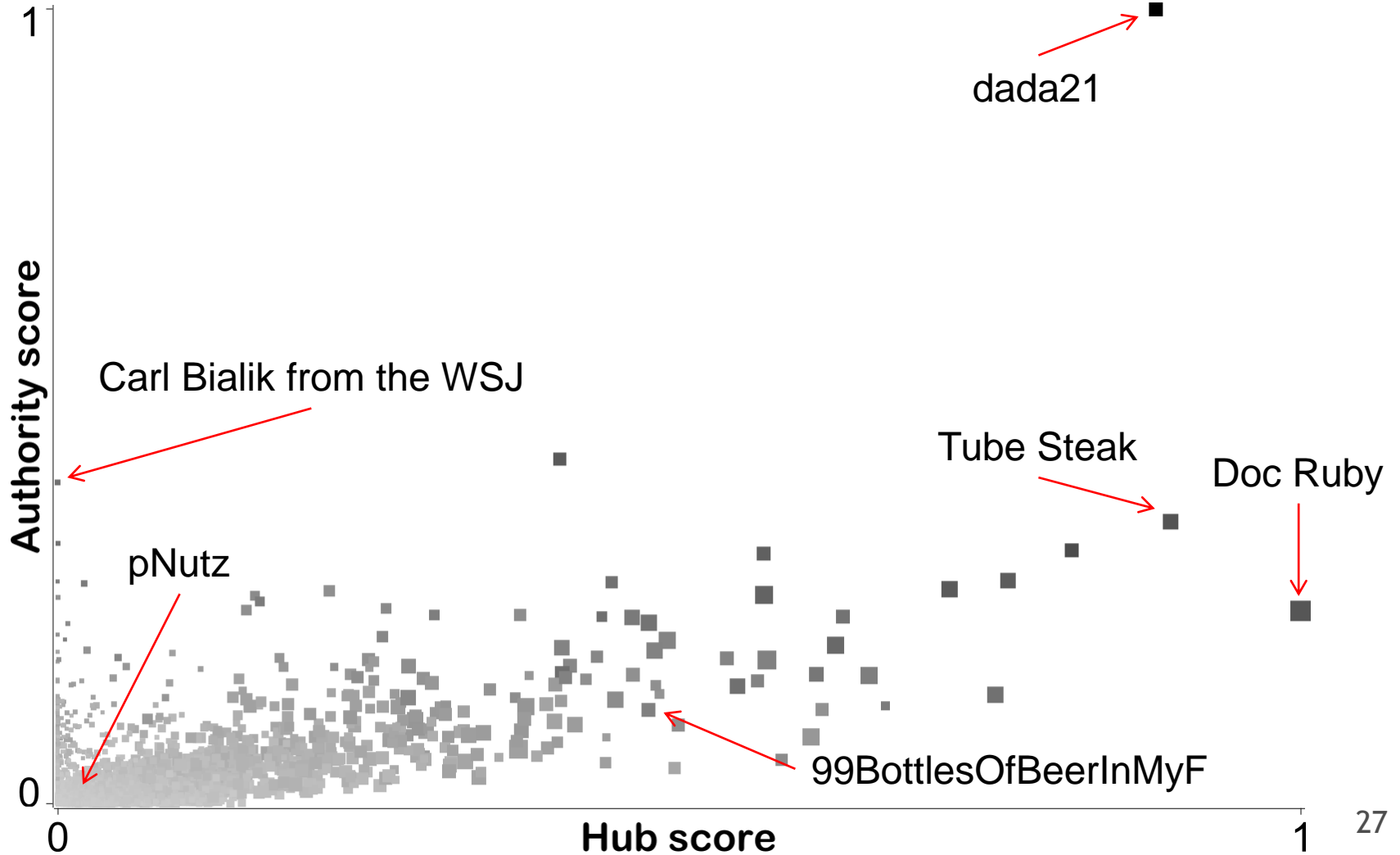
- Hubs = Followers
- Authorities = Leaders



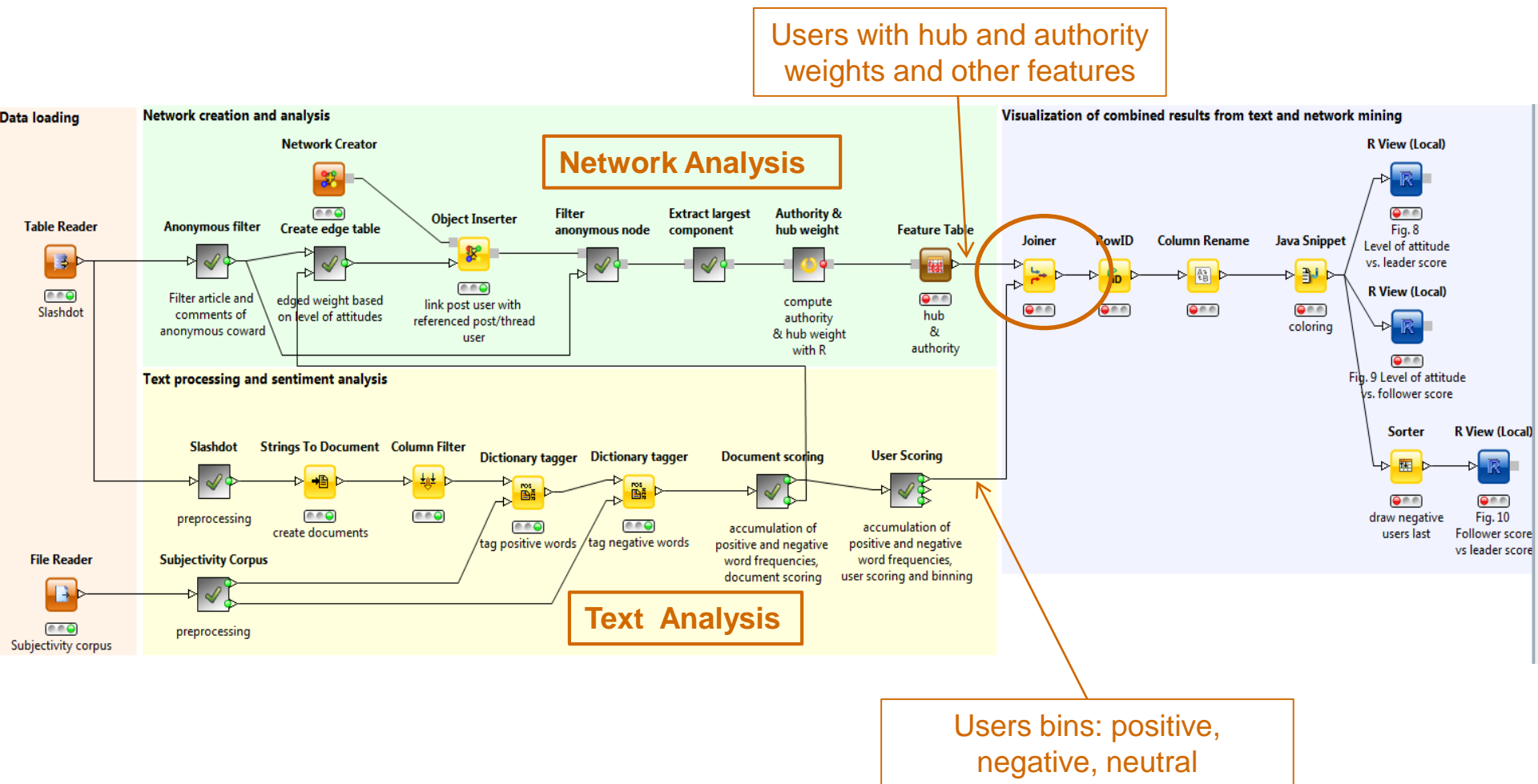


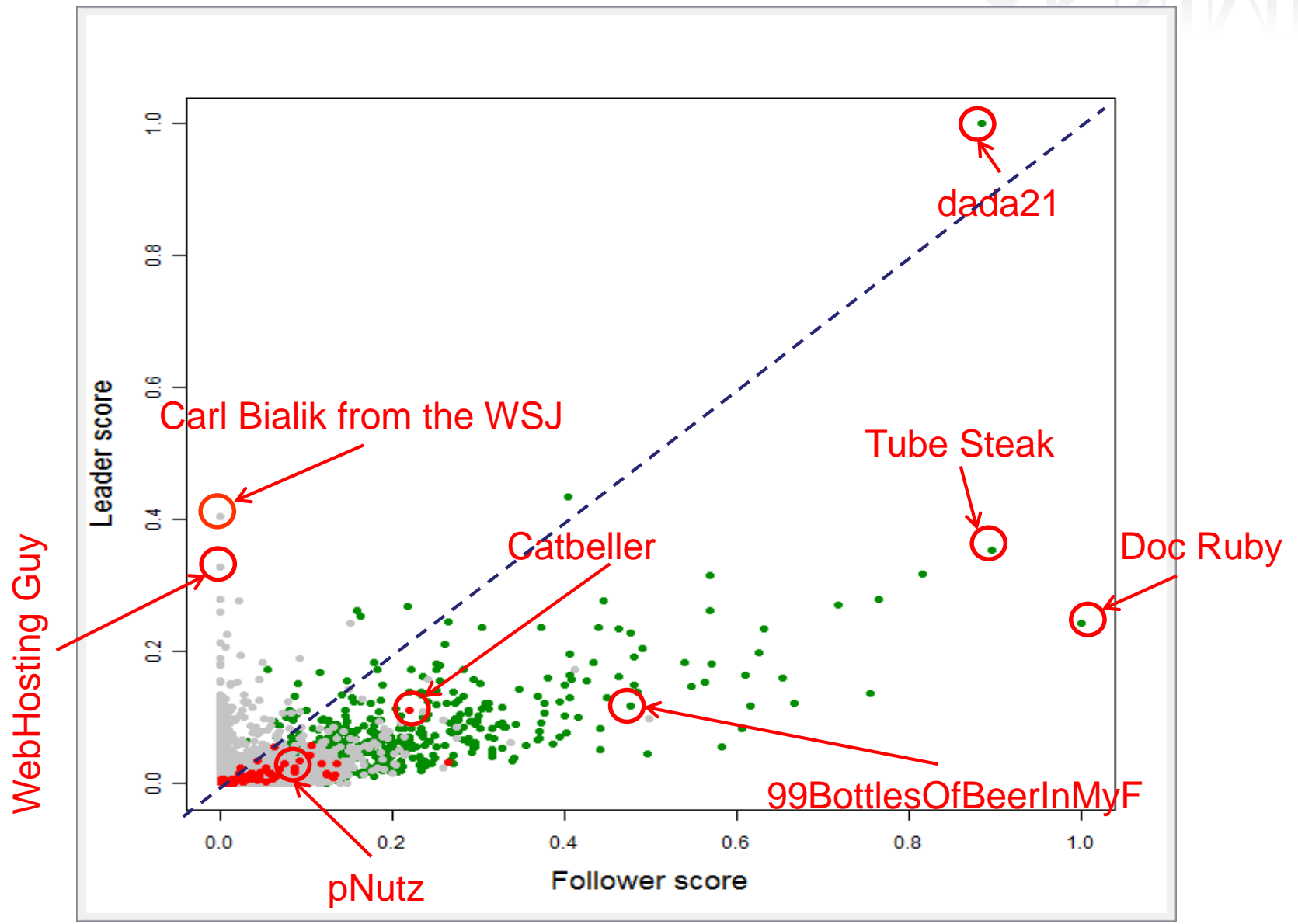


# Hubs & Authorities



# KNIME: Bringing it all together

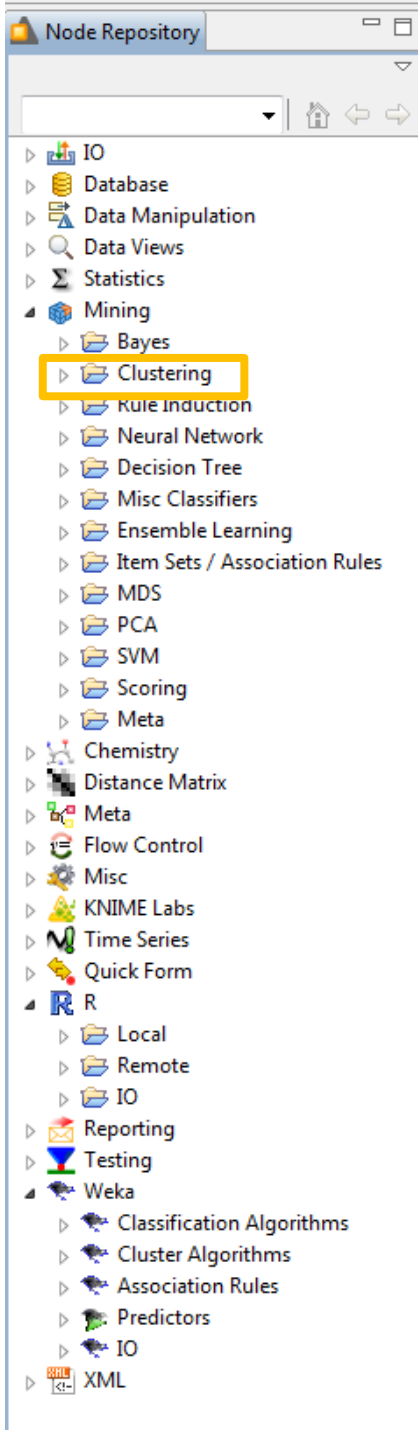




## What we have found ...

- 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?



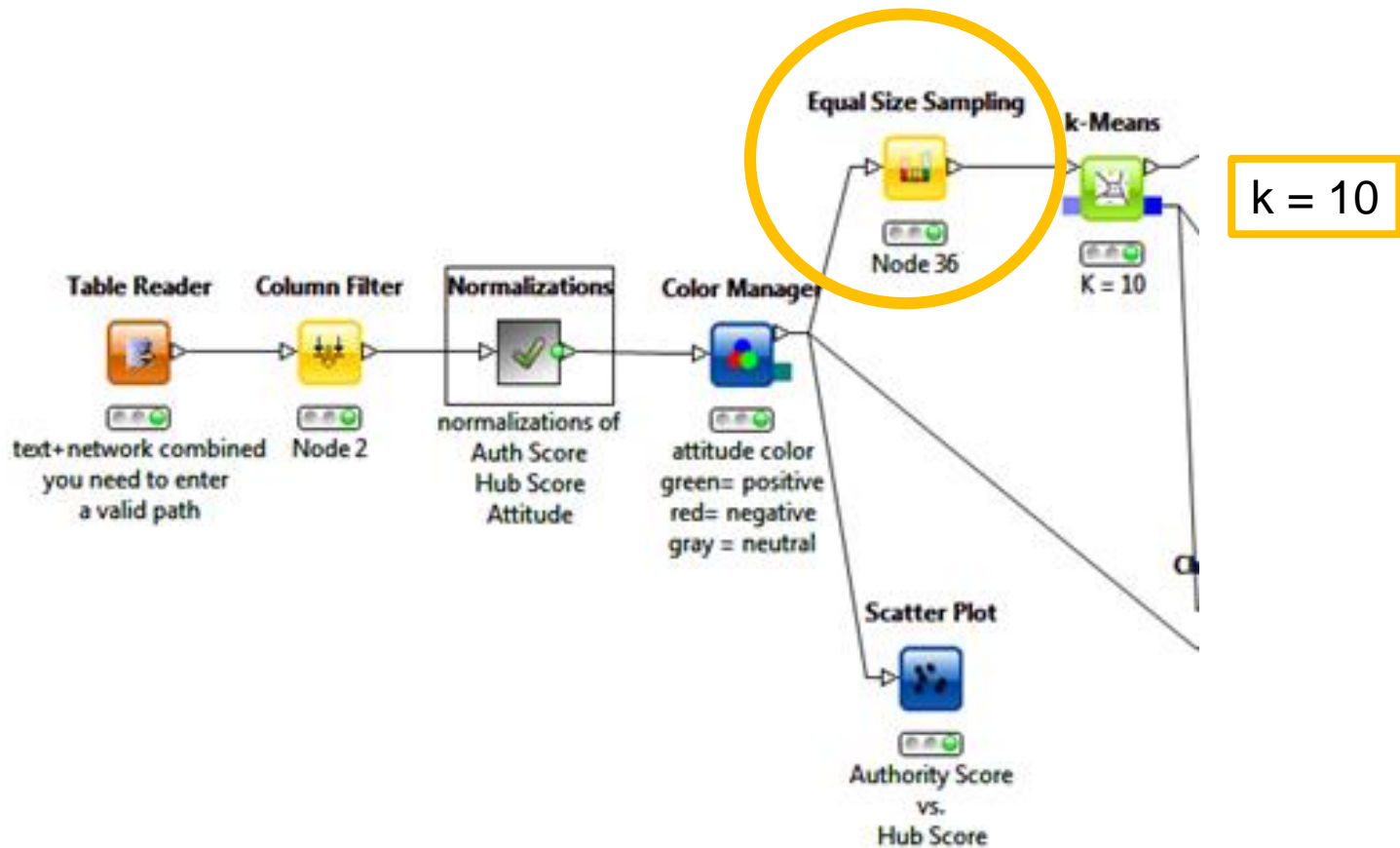
## Why Clustering?

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



k-Means algorithm

# Re-sampling the Training Set



# The k-Means Clusters

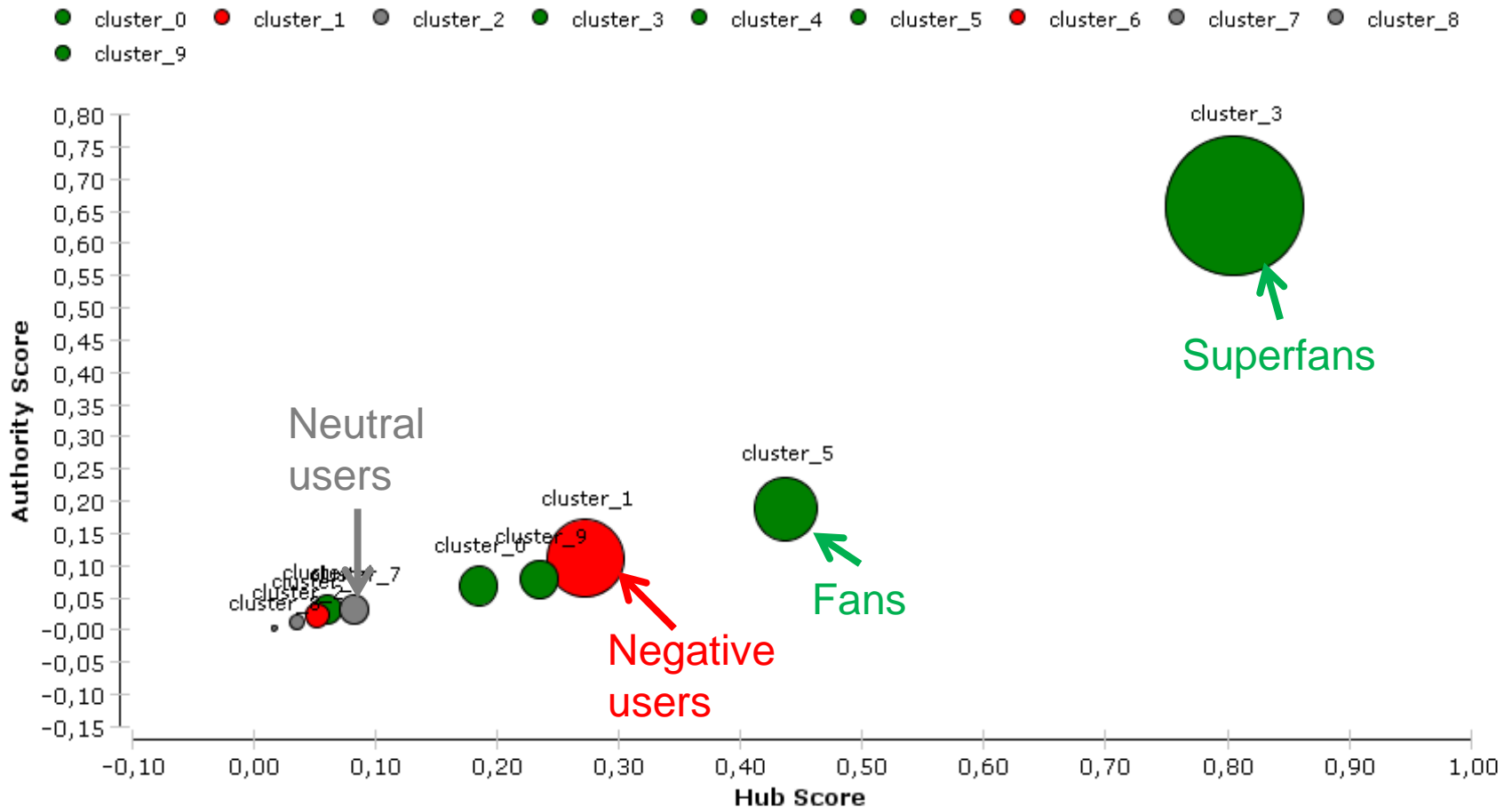
## Leaders, Followers, Positive and Negative Thinkers

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

## Leaders vs. Followers

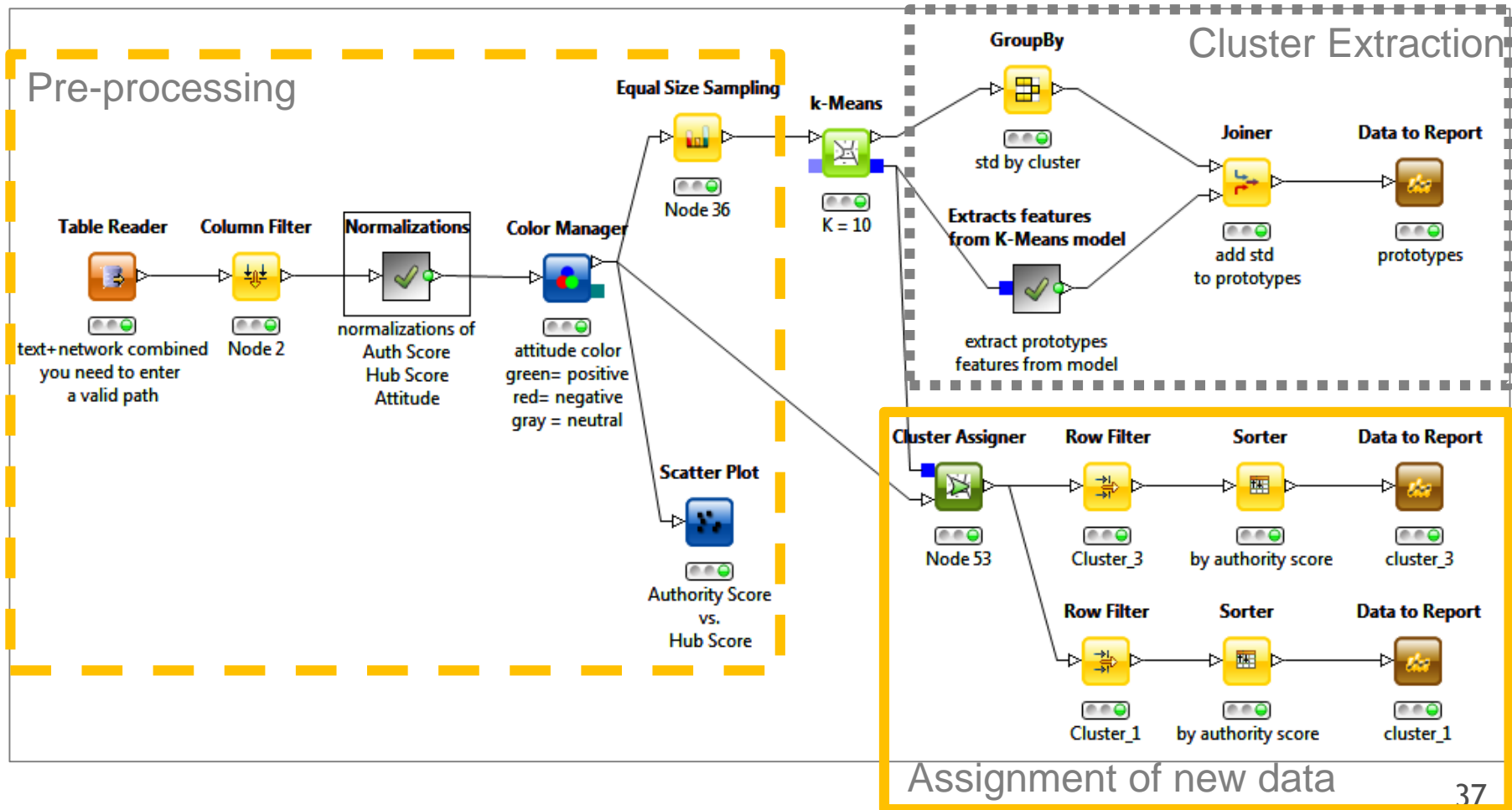




## Additional Discoveries

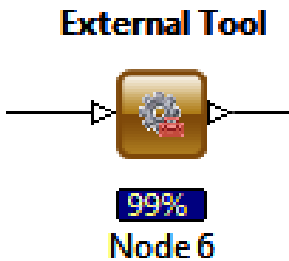
- There are only very few real leaders!  
Authority and hub scores identify active participants rather than leaders.
- Superfans can be found in cluster\_3
- Negative and (sigh!) active users are collected in cluster\_1.
- Neutral users are usually inactive (cluster\_2, cluster\_7, and cluster\_8)
- 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

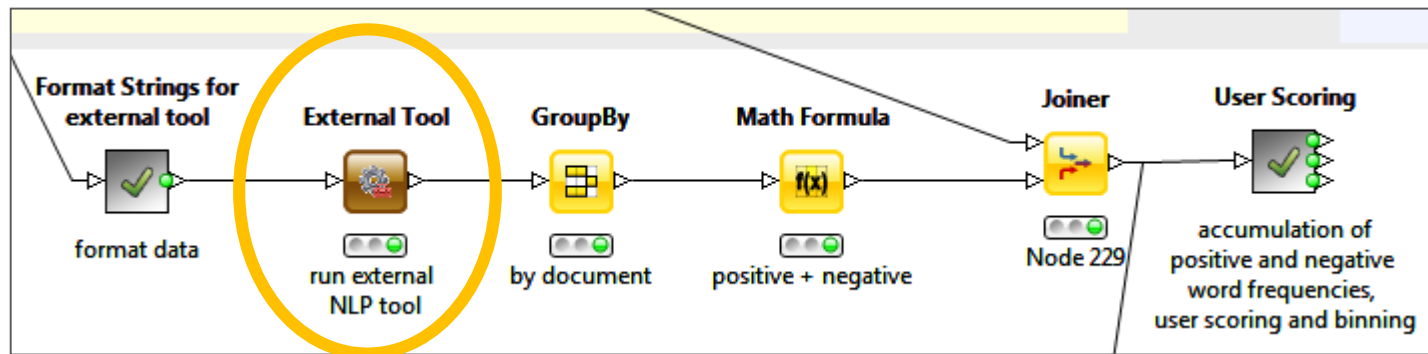
The „External Tool“ node executes **any** external program from command line

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

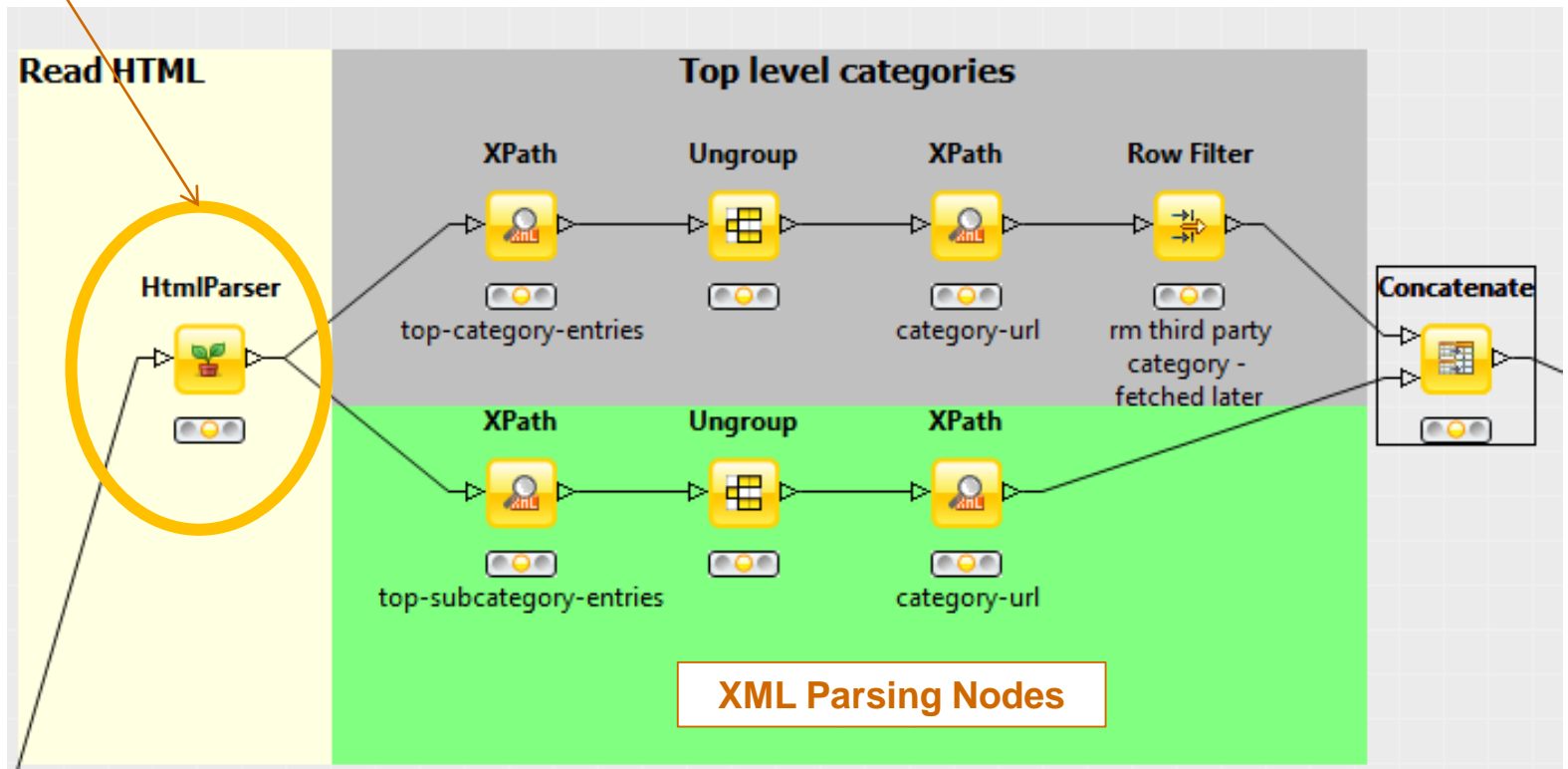
SentiStrength v2.2 (still interactive)



**External Tool and  
Generic Web Service Client**

# Web Crawling Workflow

Community Web Crawler Node



## 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](http://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](http://www.knime.com)