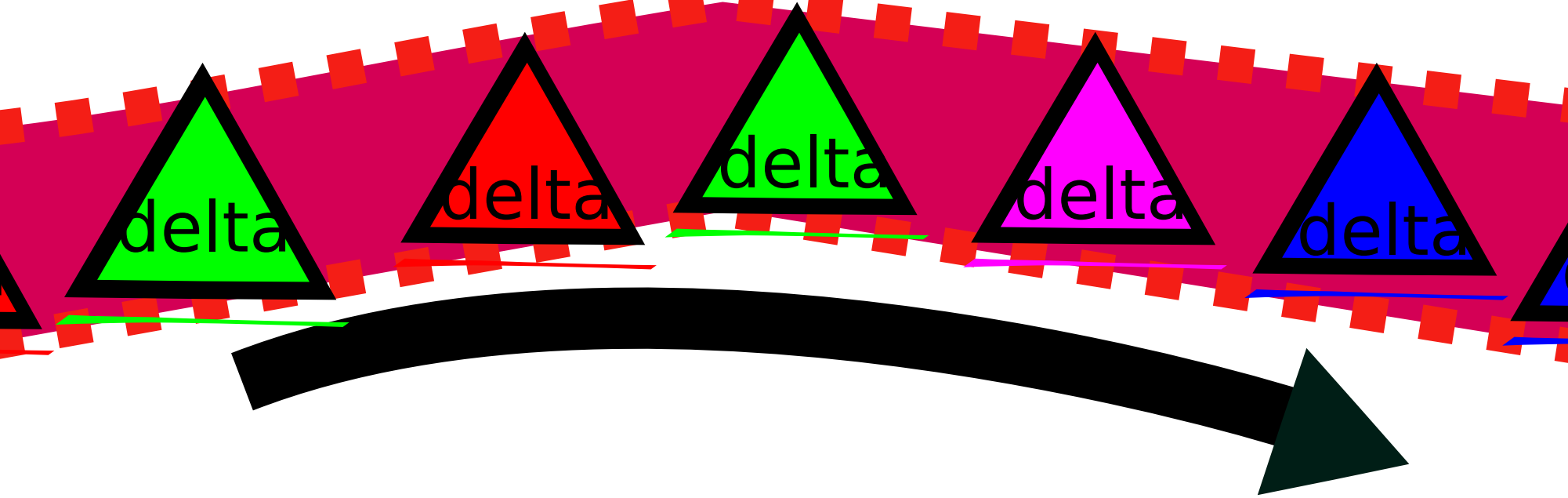
What's Hot and What's Not: Windowed Developer Topic Analysis

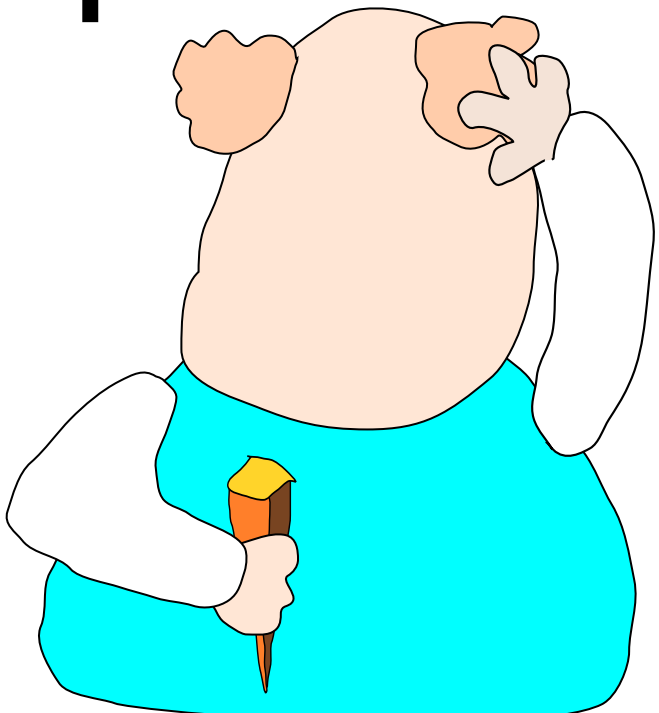
Abram Hindle, Michael W. Godfrey, Richard C. Holt

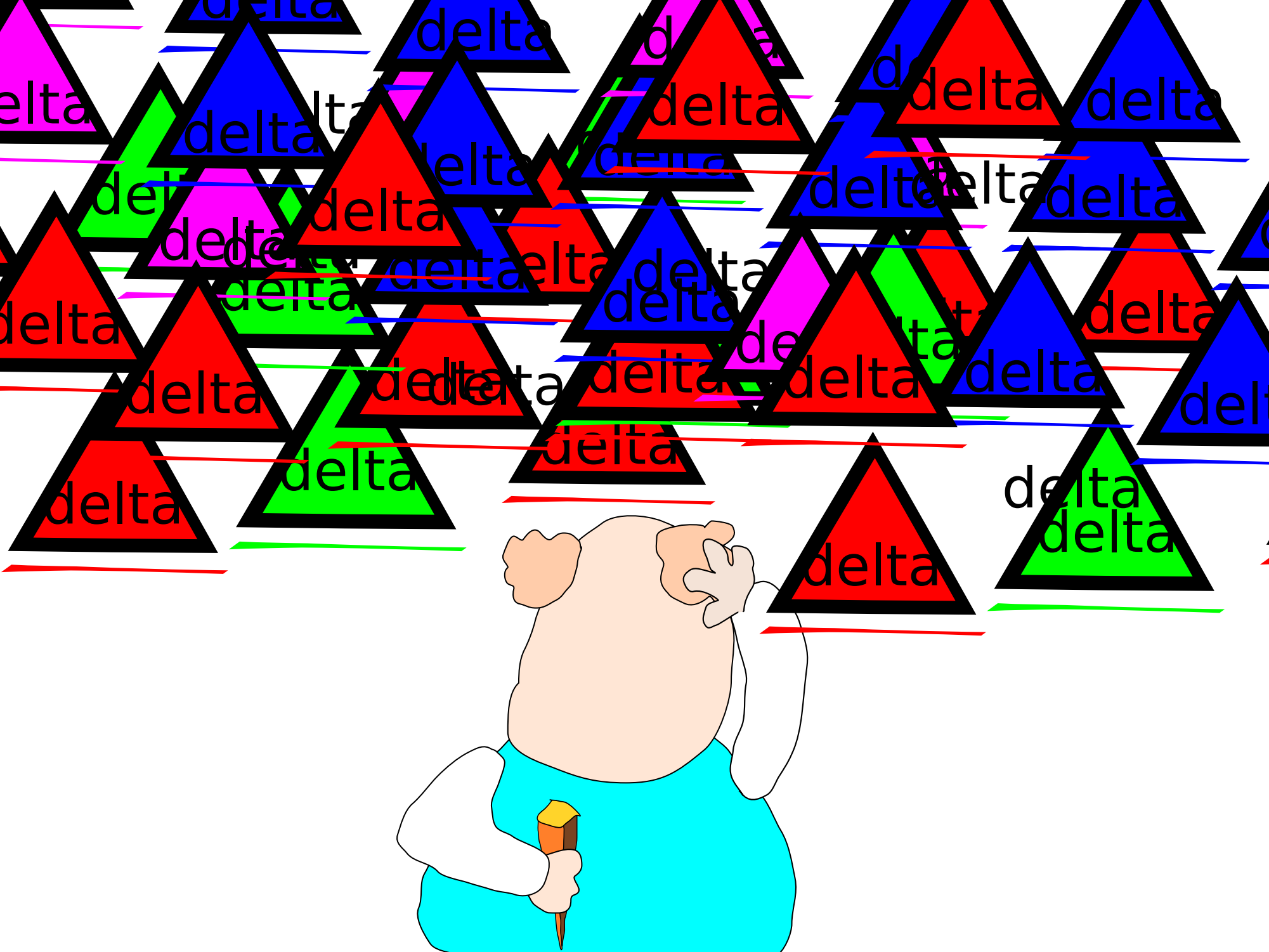
Software Architecture Group
David R. Cheriton School of Computer Science
University of Waterloo
Canada
http://swag.uwaterloo.ca/

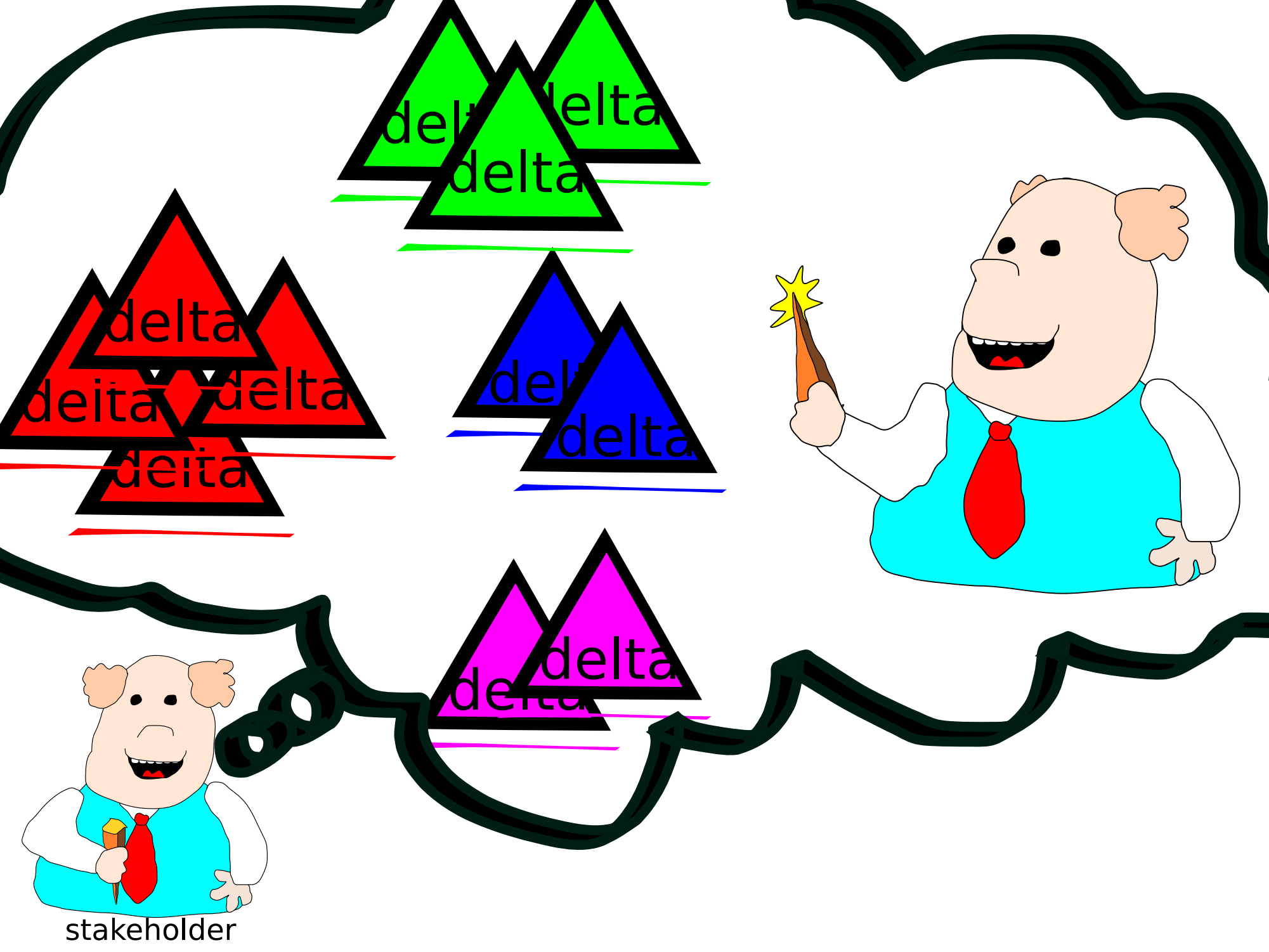
{ahindle,migod,holt}@cs.uwaterloo.ca



Development History

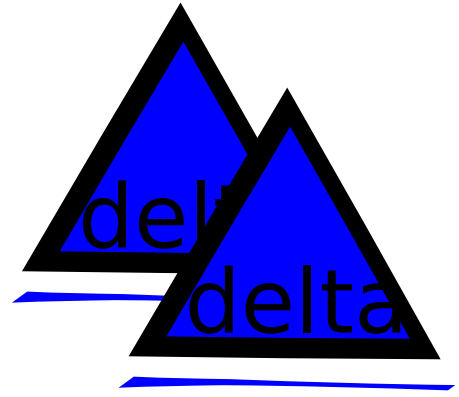



delta


delta
delta
delta
delta
delta
delta
delta
delta
delta
stakeholder

Bug Fixes

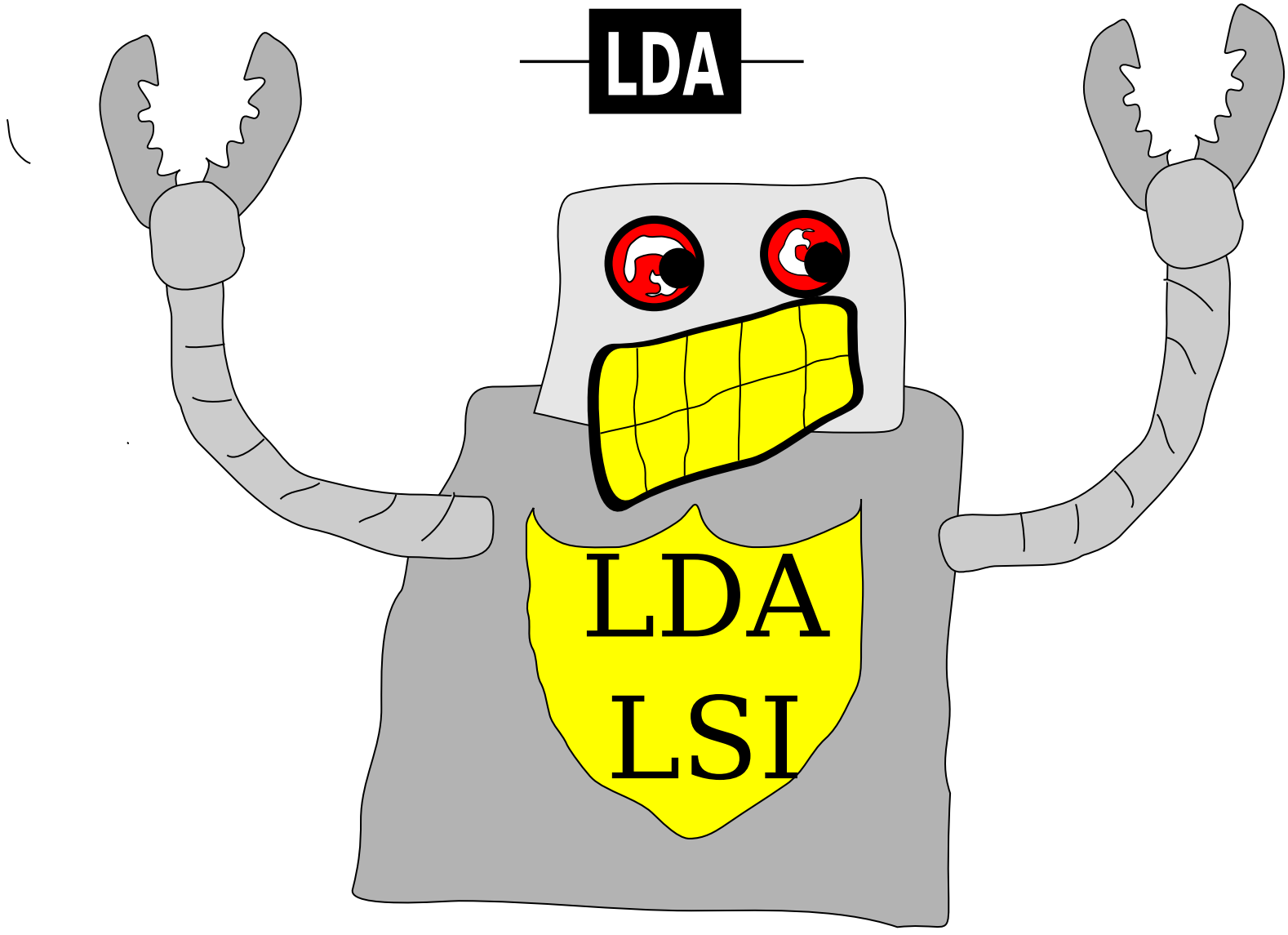
Database
Support



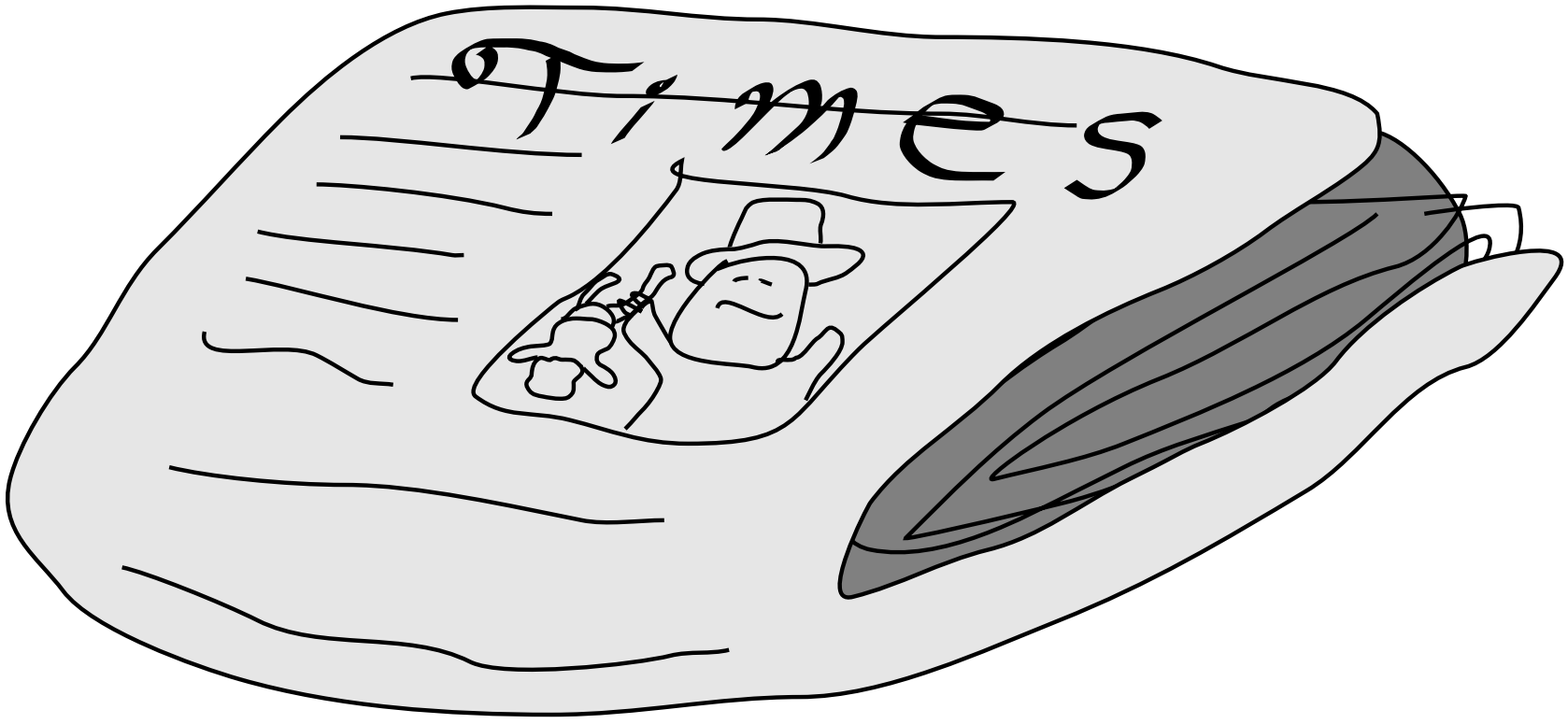
Performance
Regression

Portability

Our blackbox

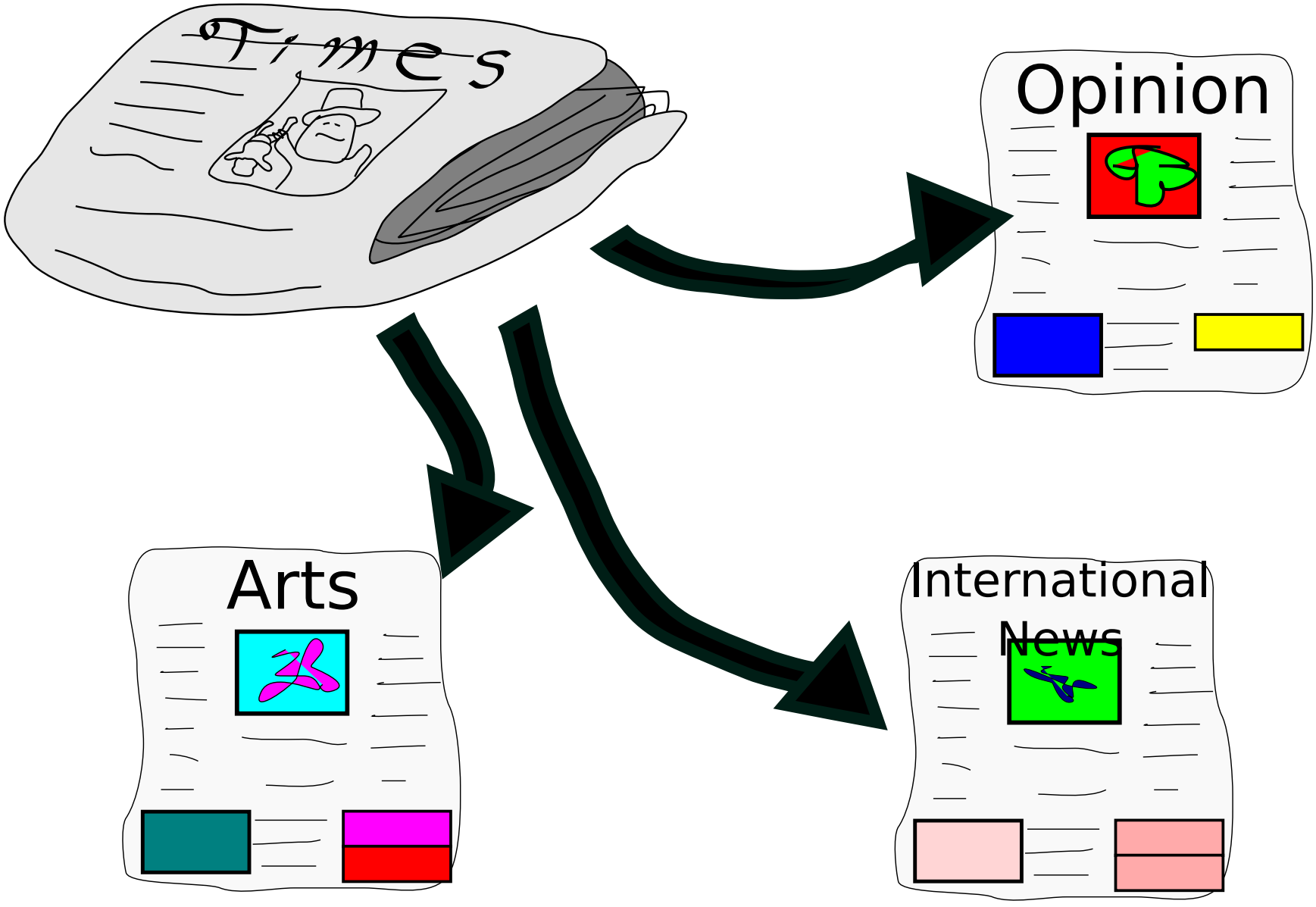


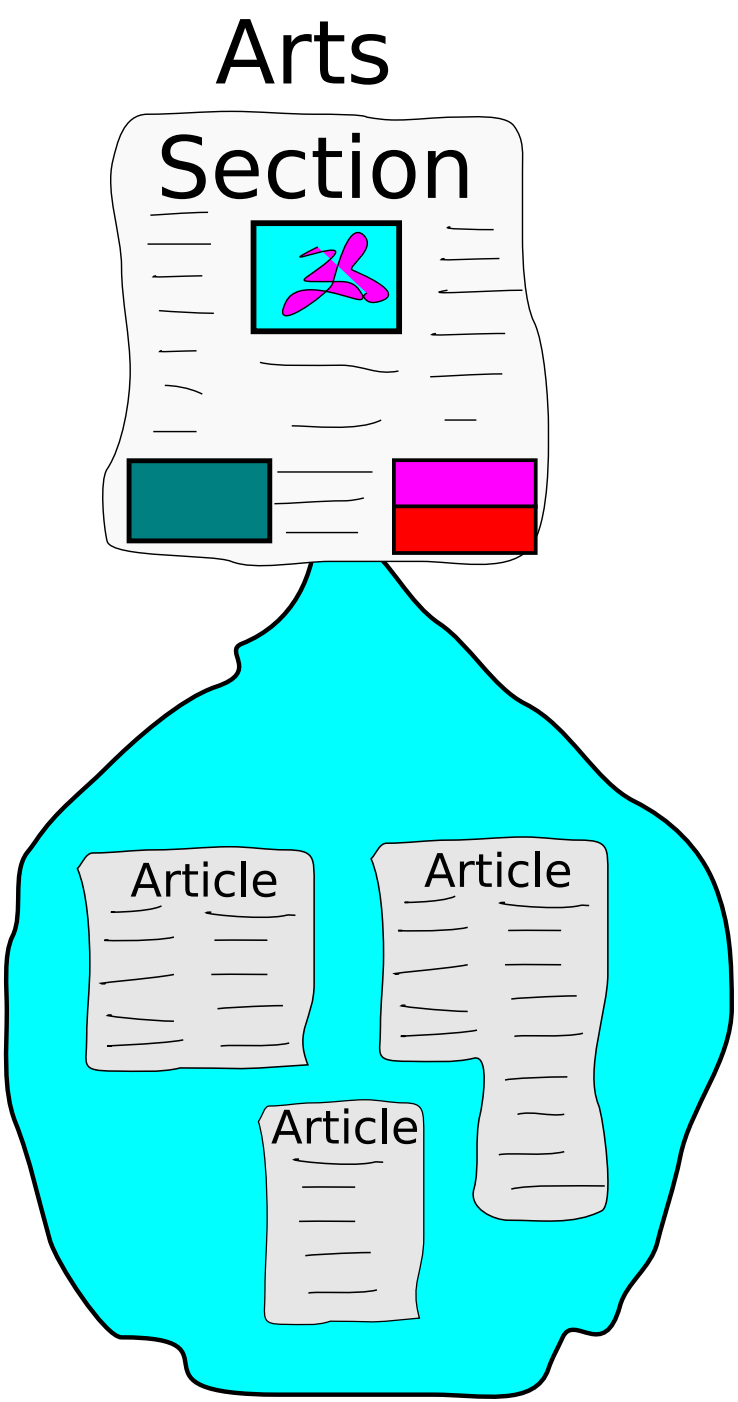
Example

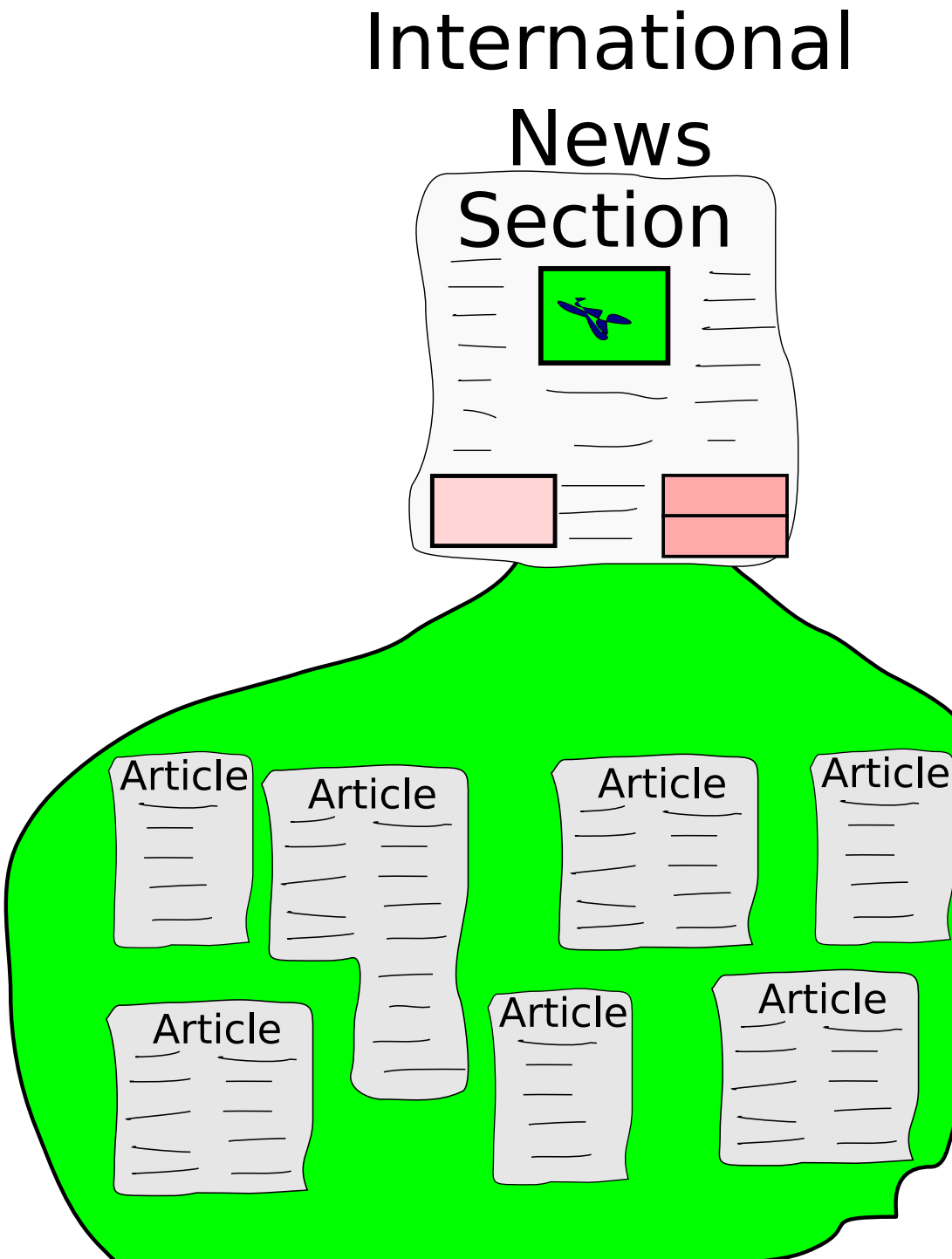


[Blei]

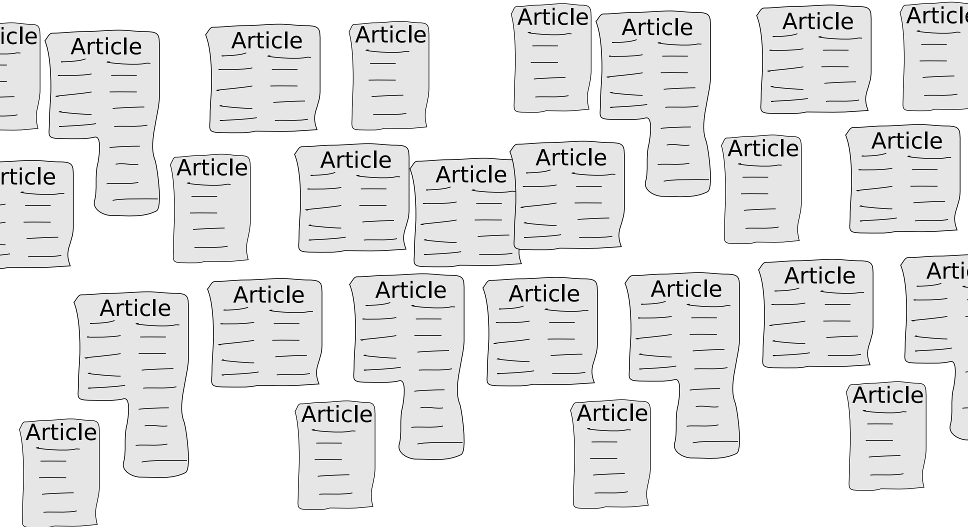
apologies to those with prior LDA/LSI experience

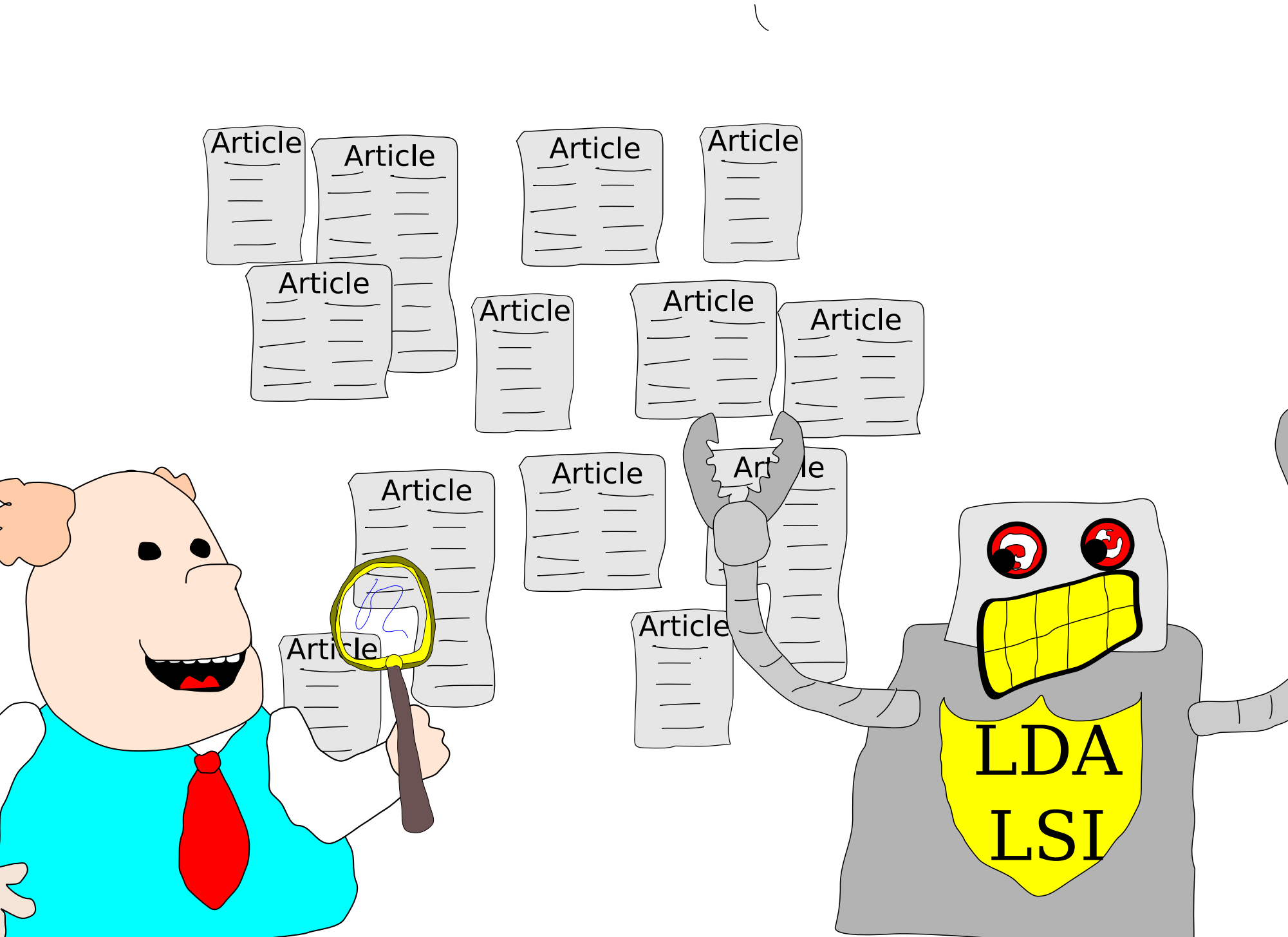

Times
Opinion
Arts
International
News

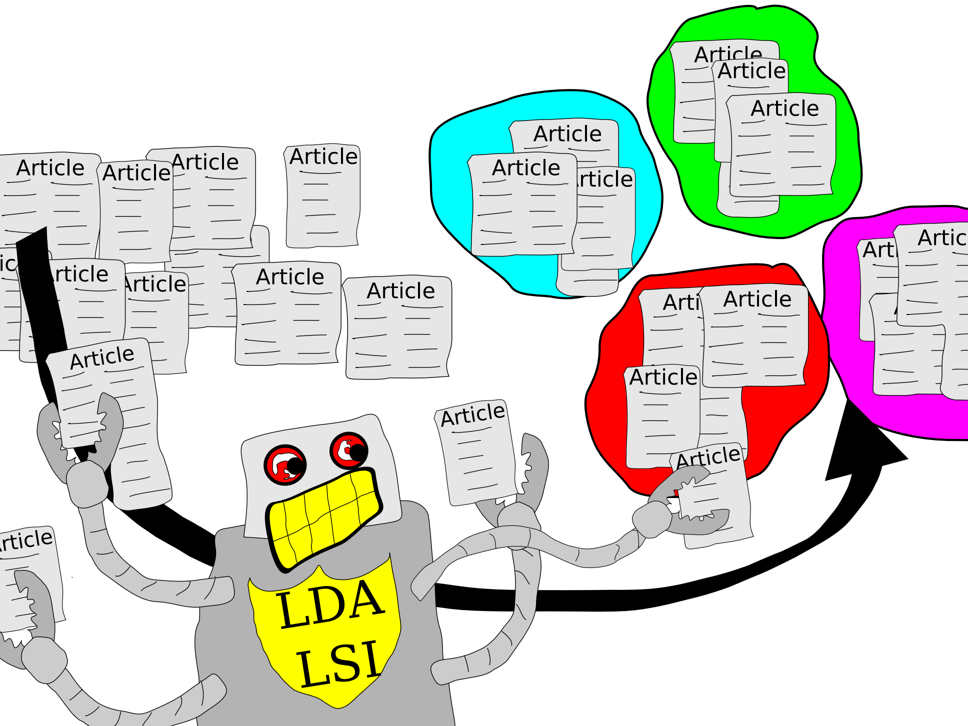

Arts
Section
Article
Article
Article

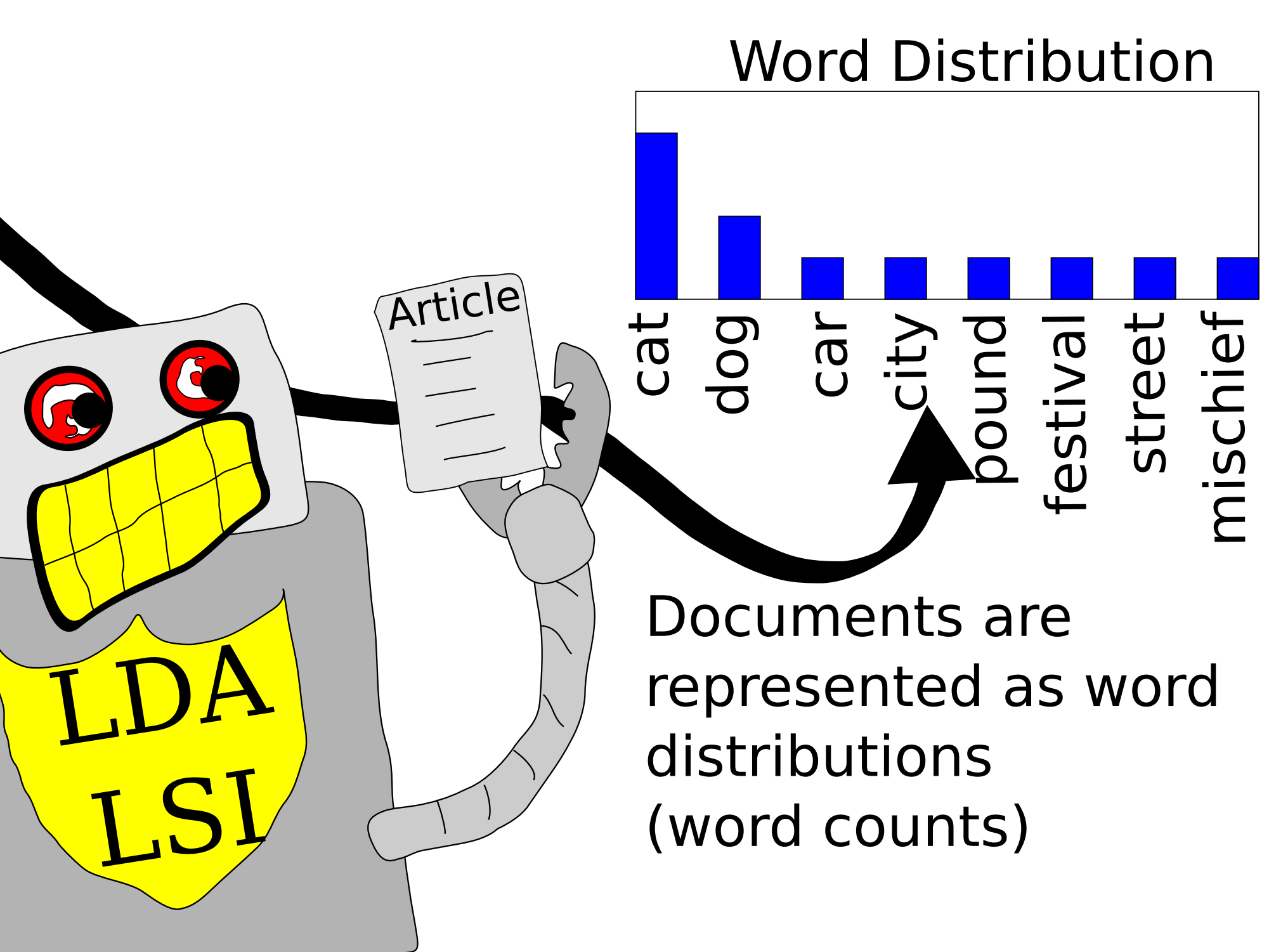

International
News
Section
Article
Article
Article
Article
Article
Article
Article

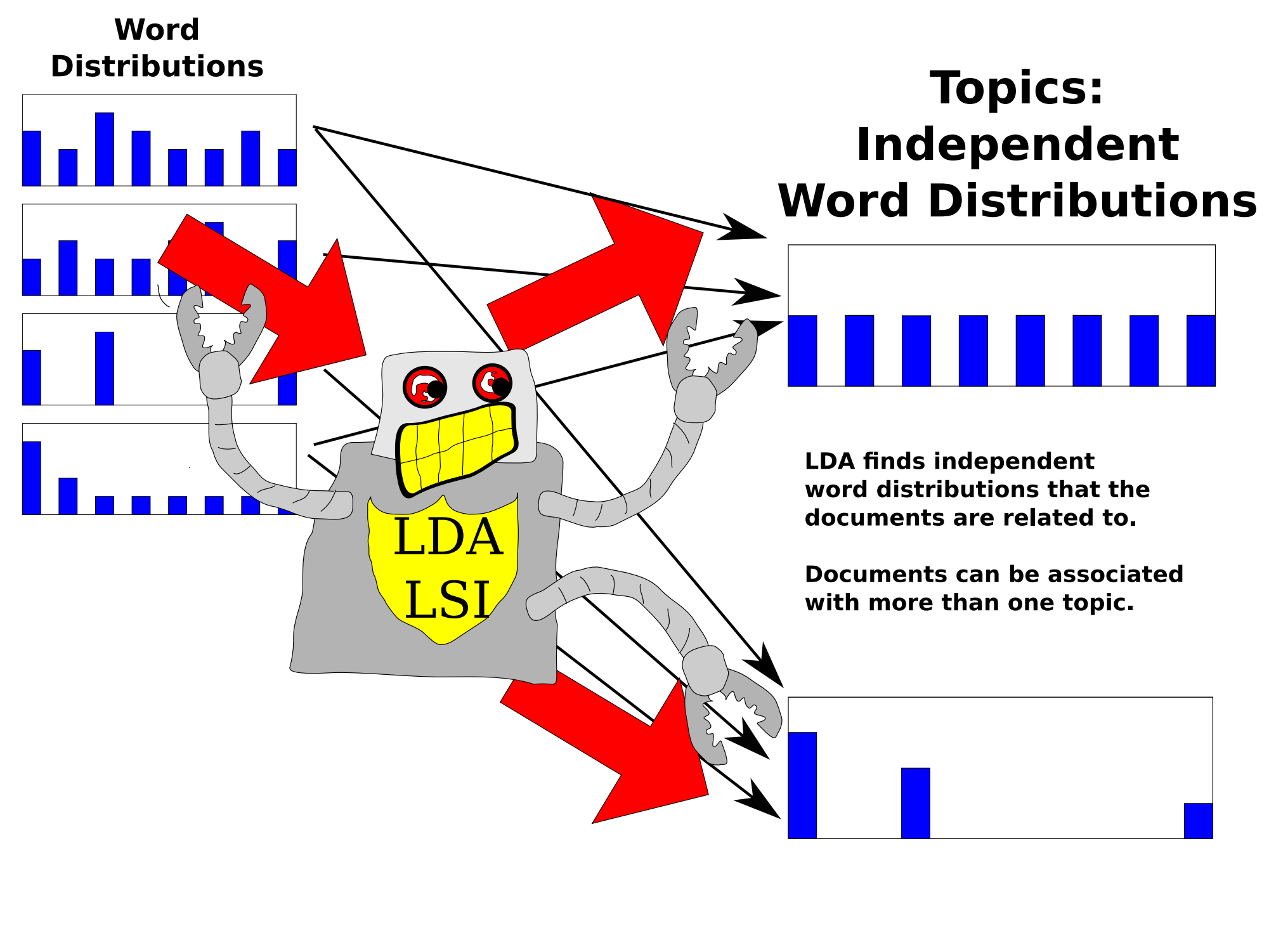
What if we didn't know what section the articles were in?

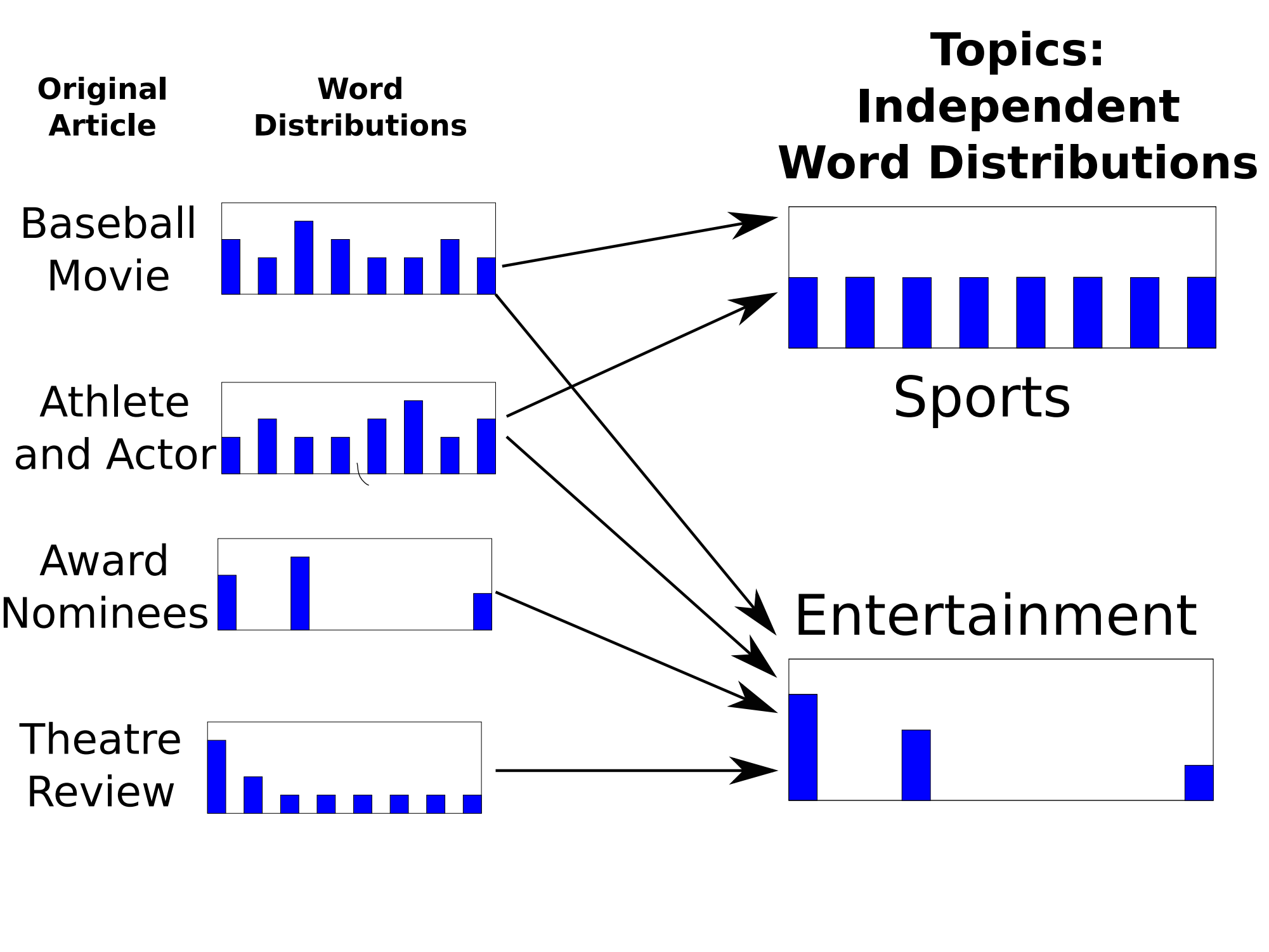



Article
Article
Article
Article
Article
Article
Article
Article
Article
Article
Article
Article
Article
LDA
LSI

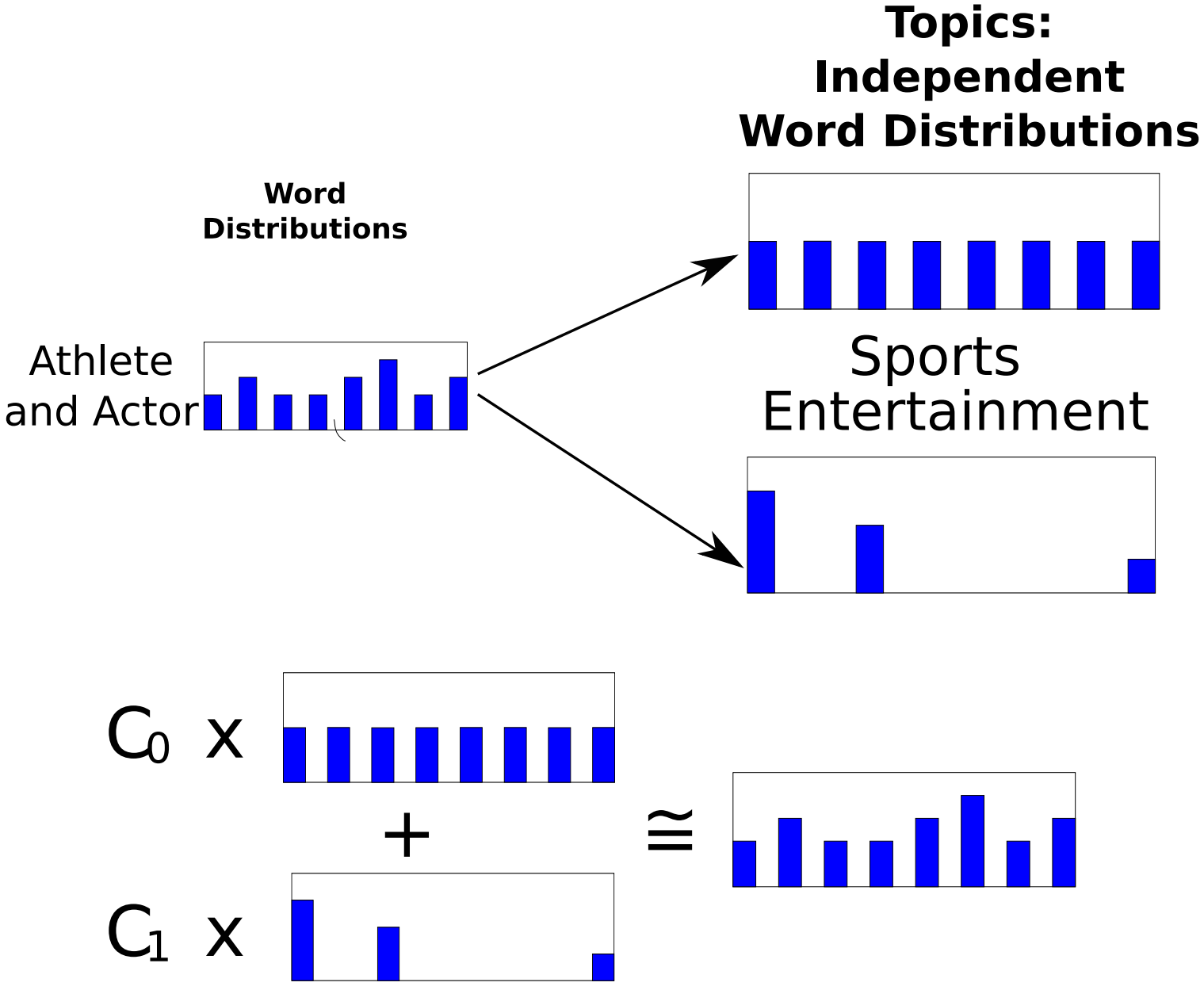

Article
Article
Article
Article
Article
Article
Article
Article
Article
Article
Article
Article
Article
Article
Article
Article
Article
LDA
LSI

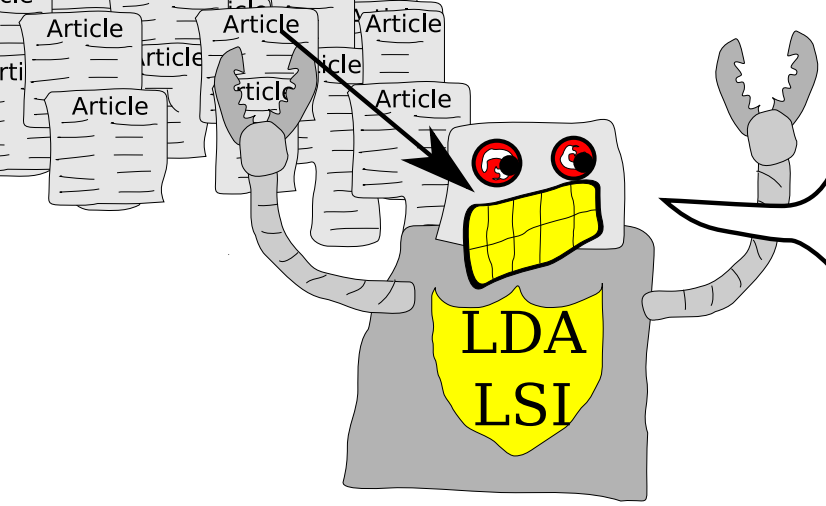

Word Distribution
cat
dog
car
city
pound
festival
street
mischief
Article
LDA
LSI
Documents are represented as word distributions (word counts)


Word Distributions
Topics: Independent Word Distributions
LDA
LSI
LDA finds independent word distributions that the documents are related to.
Documents can be associated with more than one topic.


Original Article
Word Distributions
Topics: Independent Word Distributions
Baseball Movie
Athlete and Actor
Award Nominees
Theatre Review
Sports
Entertainment

Documents are represented as a linear combination of independent topics





Here are two topics. I don't know what they are about!

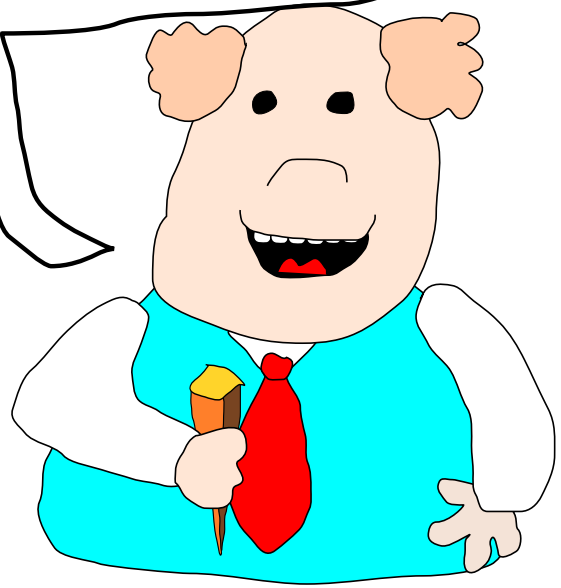
Topic 1

* play
* game
* inning
* player
* quarter
* opponent
* ...

Topic 2

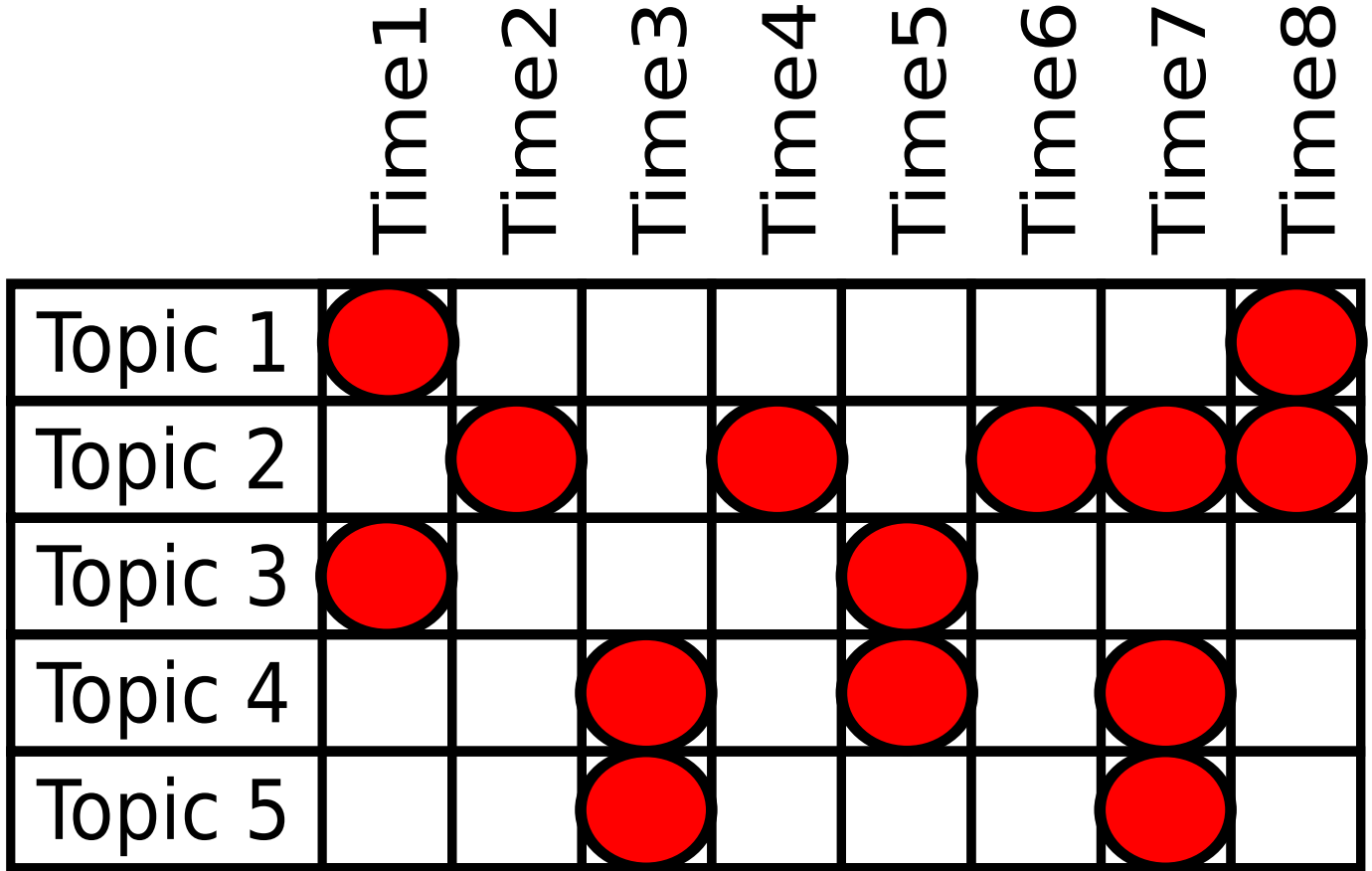
* gambling
* play
* night life
* comedy
* movie
* theatre
* ...

These word lists look look like: **Sports** and **Entertainment** !

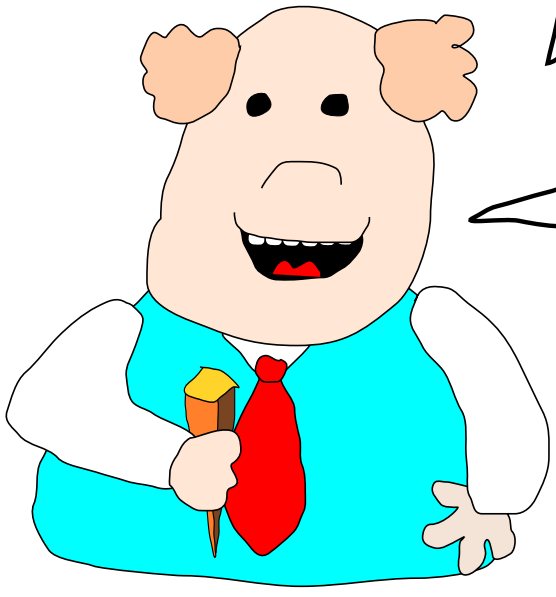


Previous authors

* Linstead et al. & Lukins et al.
* analyzed the entire history
* extracted N topics across the entire project's life time.



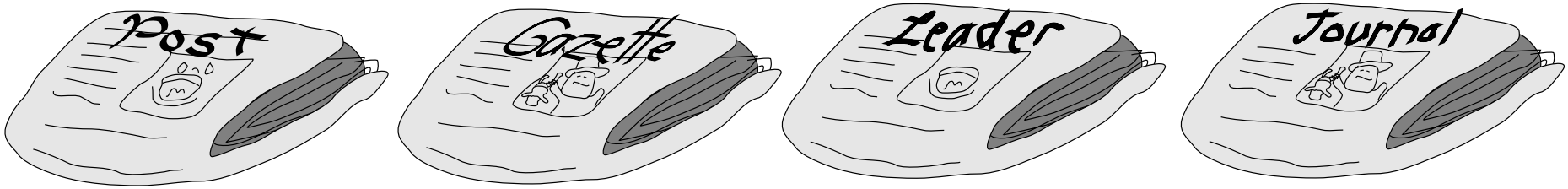
Topic analysis (LDA) can group articles around topics. Can we apply this to software?



Can we find topics, that developers were concerned with, from change-logs

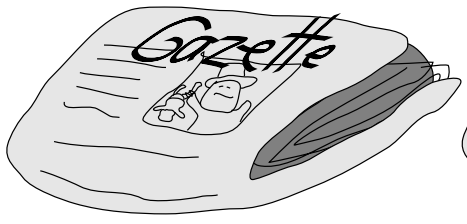
Common topics dominate!
If we did this to newspapers we'd expect to extract the sections, not necessarily local news trends!

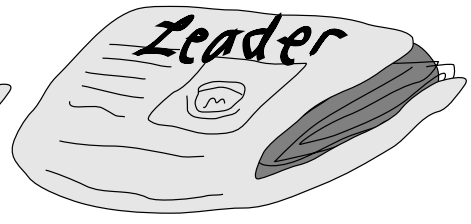
	Time1	Time2	Time3	Time4	Time5	Time6	Time7	Time8
News	●							●
Opinion		●		●		●	●	●
Sports	●				●			
Life			●		●		●	
Auto			●				●	

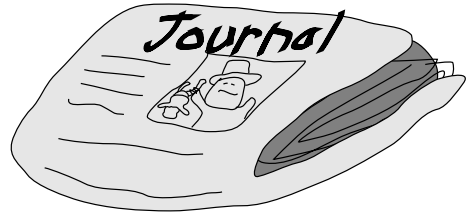


Perhaps more local topics are more interesting?

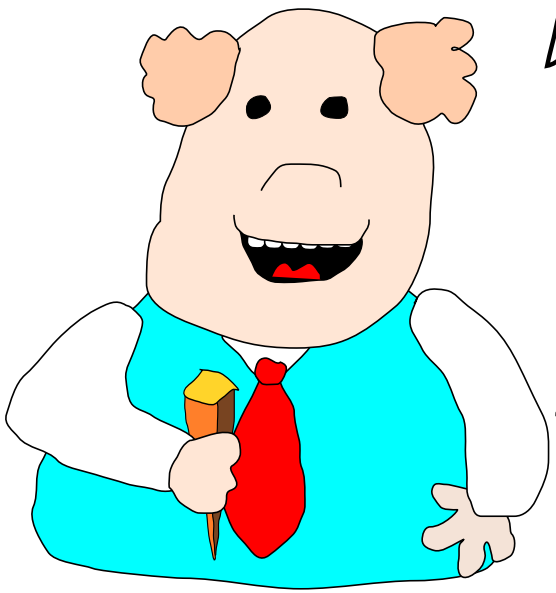
	Time1	Time2	Time3	Time4	Time5	Time6	Time7	Time8
Border Tension								●
Playoffs		●		●		●	●	●
Celebrity Death	●							
Copyright Case			●		●		●	
Local Crime							●	



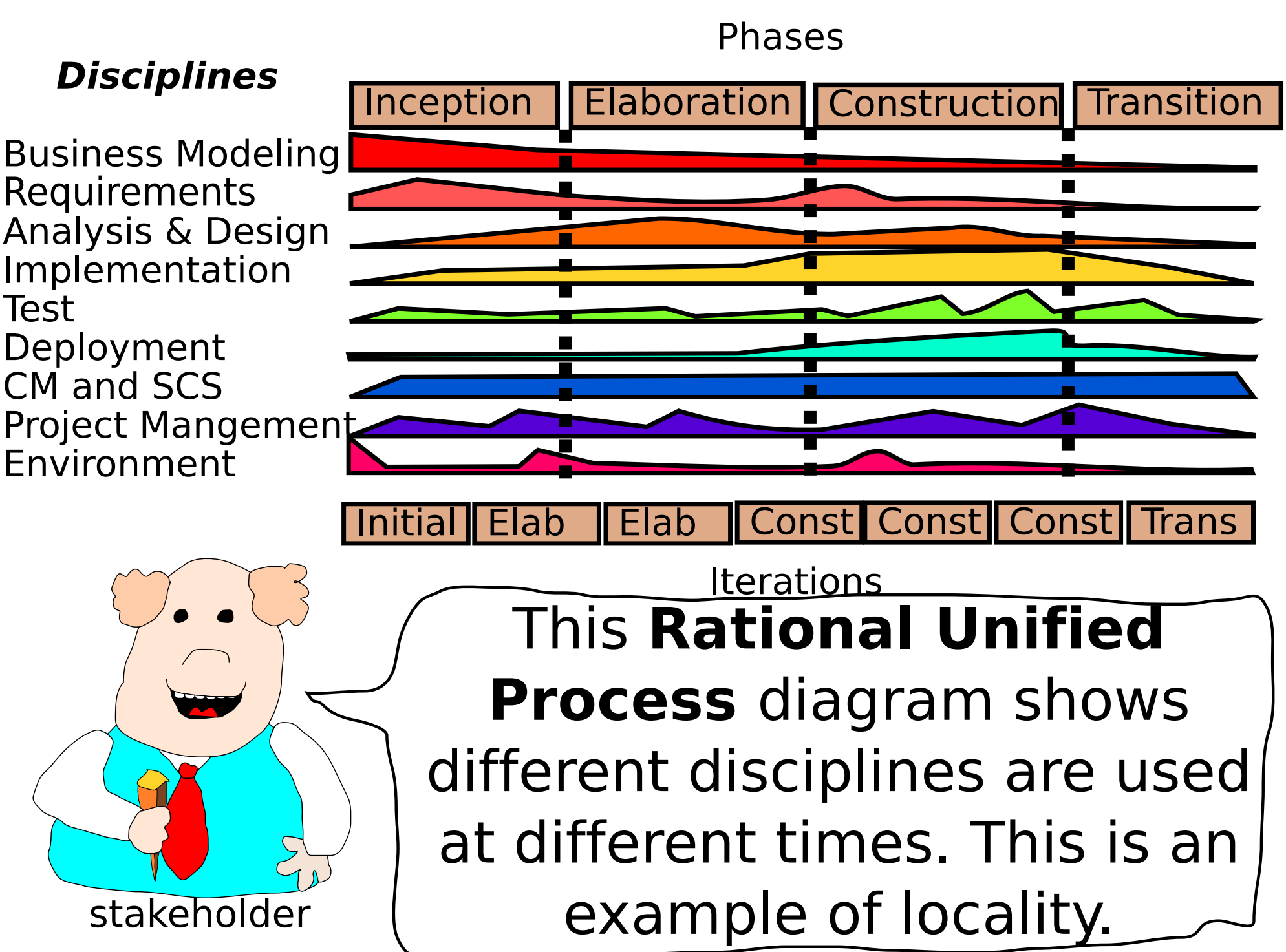




Local topic analysis is relevant to software because we care about what just occurred and often deal with iterations of development

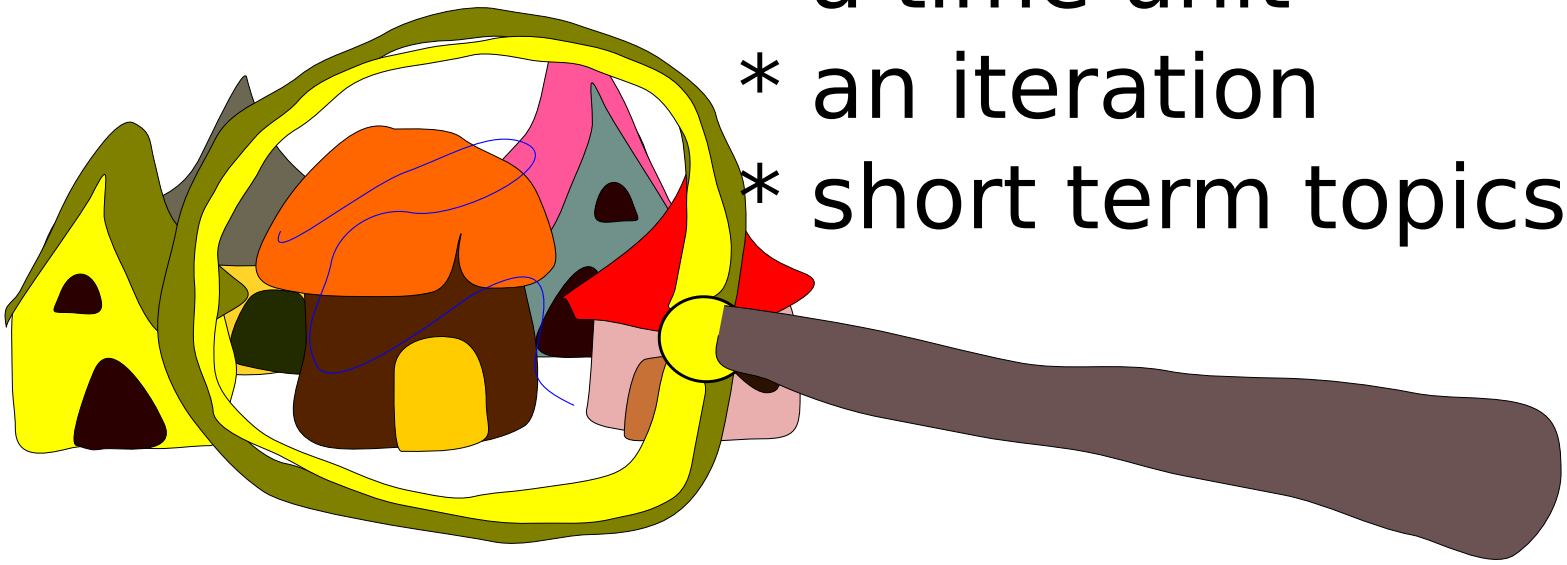


One iteration might have totally different develop-ment topics than another


Phases
Disciplines
Inception
Elaboration
Construction
Transition
Business Modeling
Requirements
Analysis & Design
Implementation
Test
Deployment
CM and SCS
Project Mangement
Environment
Initial
Elab
Elab
Const
Const
Const
Trans
Iterations
stakeholder
This Rational Unified Process diagram shows different disciplines are used at different times. This is an example of locality.

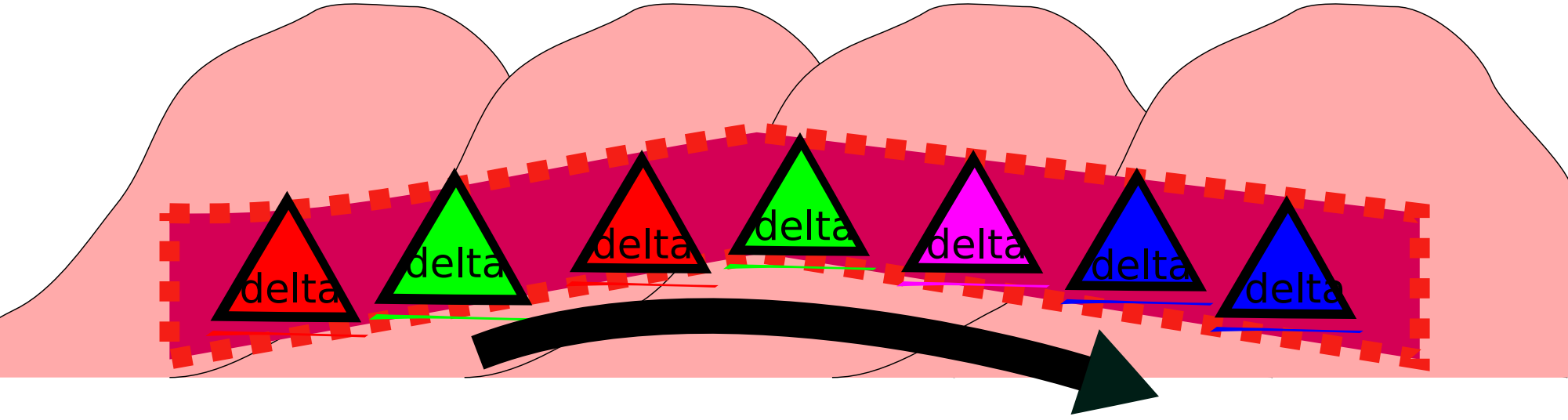
Local analysis

* a time unit
* an iteration
* short term topics



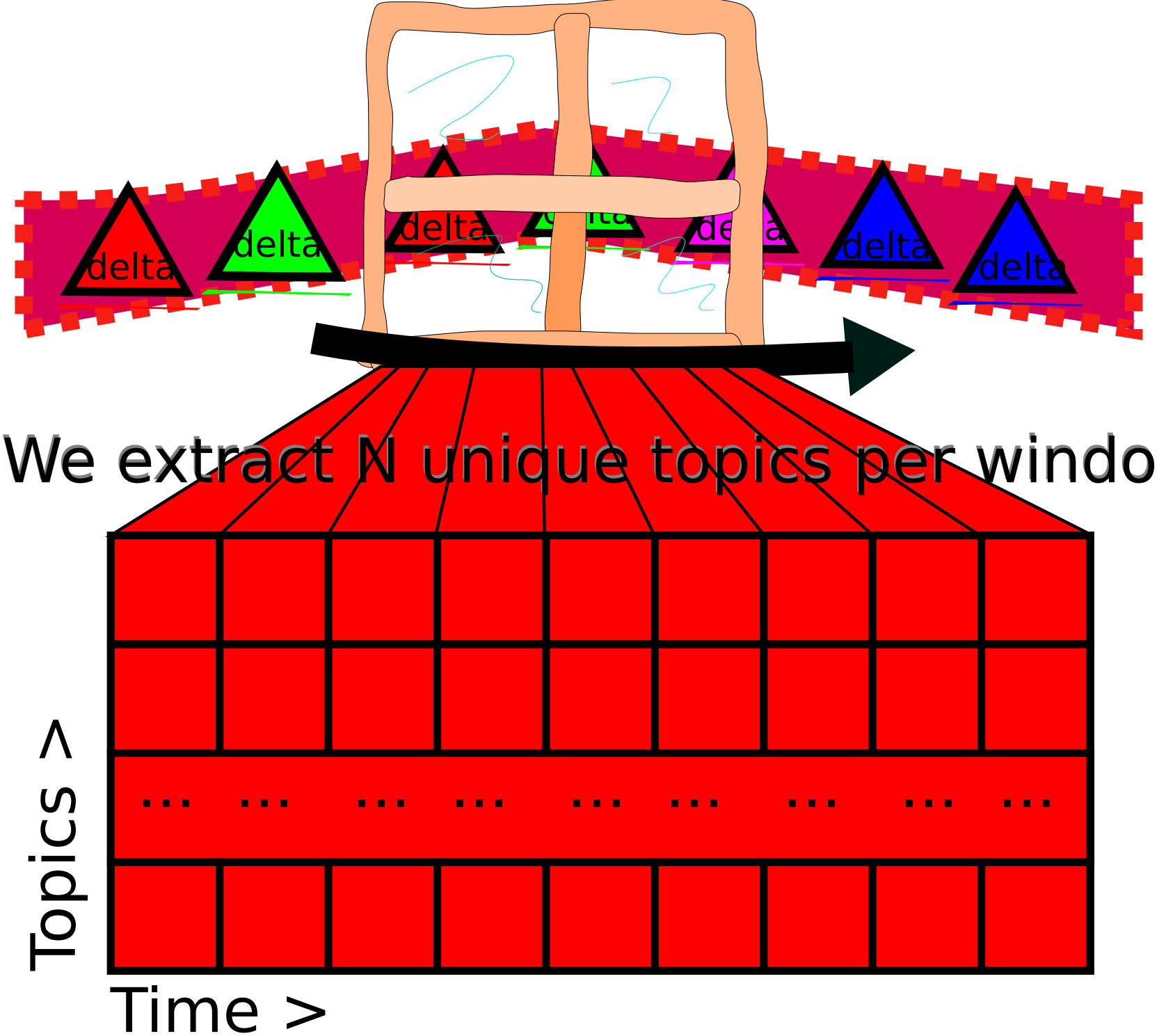
Global Analysis

* relate local topics globally
* Recurring or dominant topics

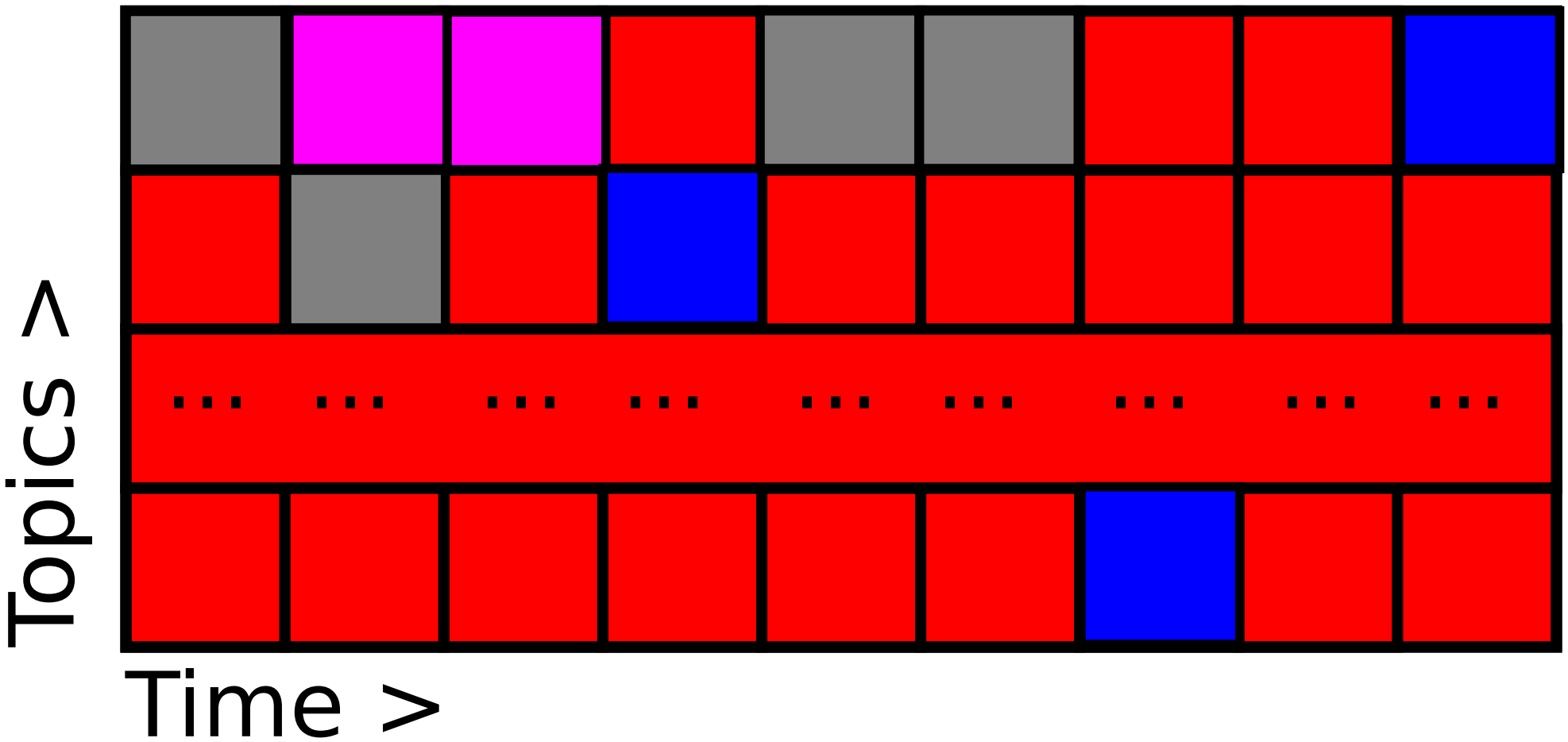


Development History

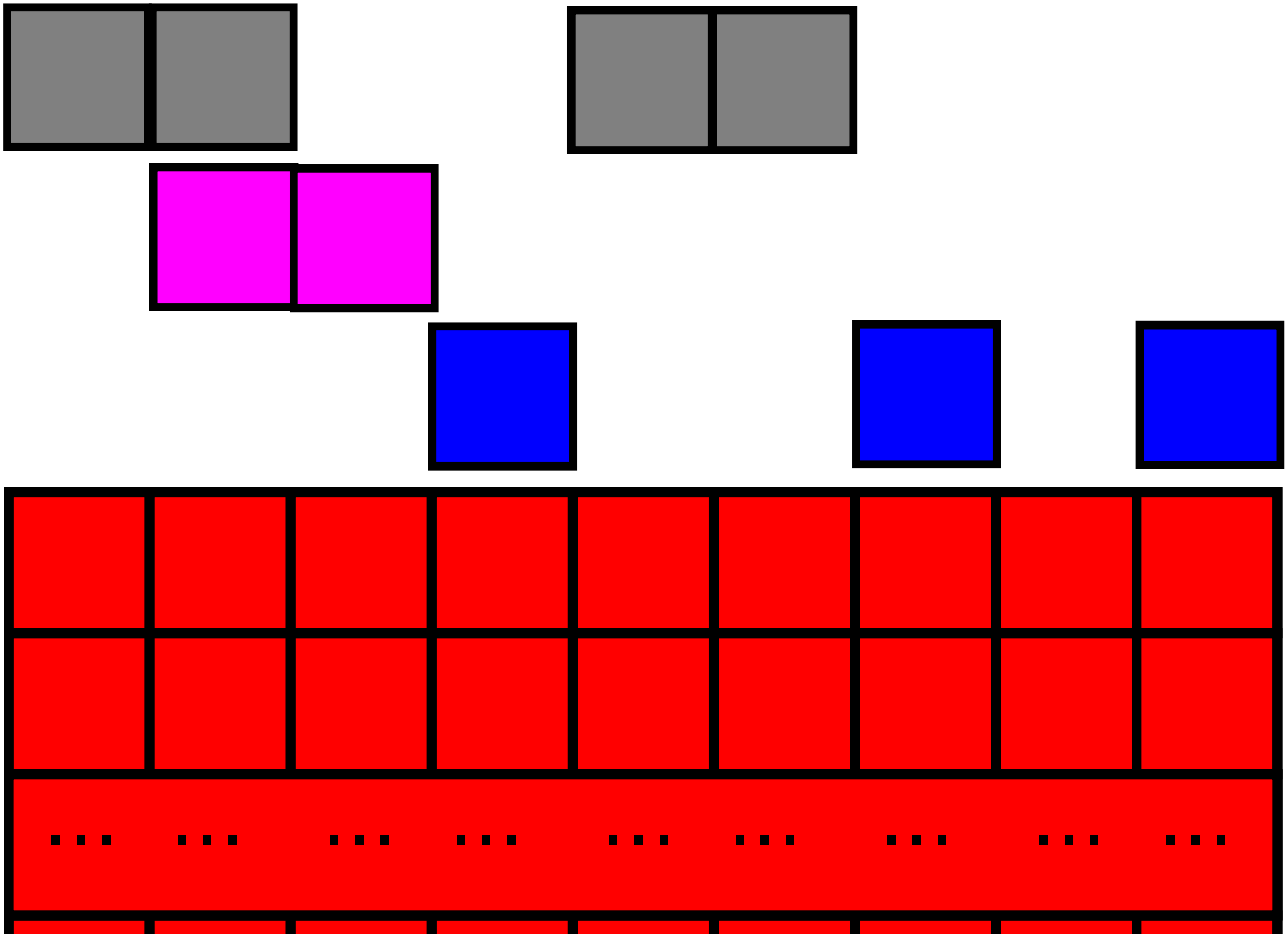
Apply a sliding window


delta
delta
delta
delta
delta
We extract N unique topics per window!
Topics >
Time >

We noticed that some topics are quite similar, they might not appear just locally, they might reoccur globally!

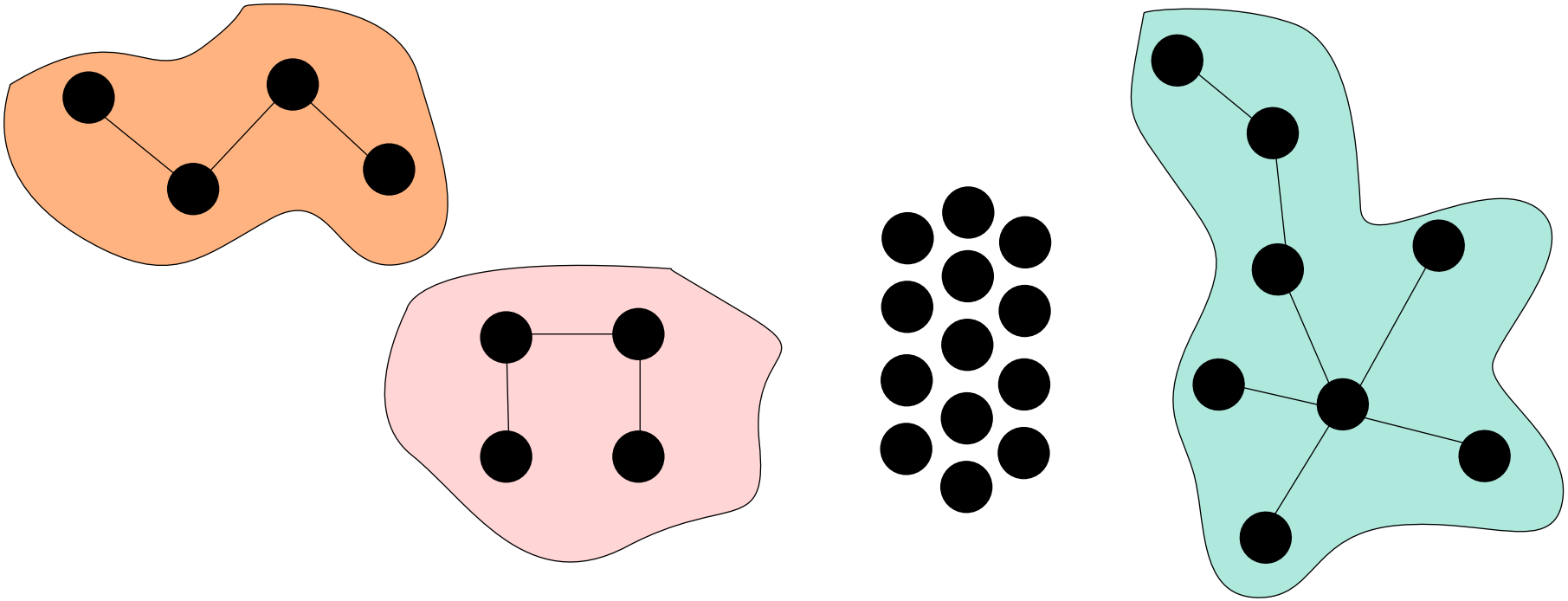


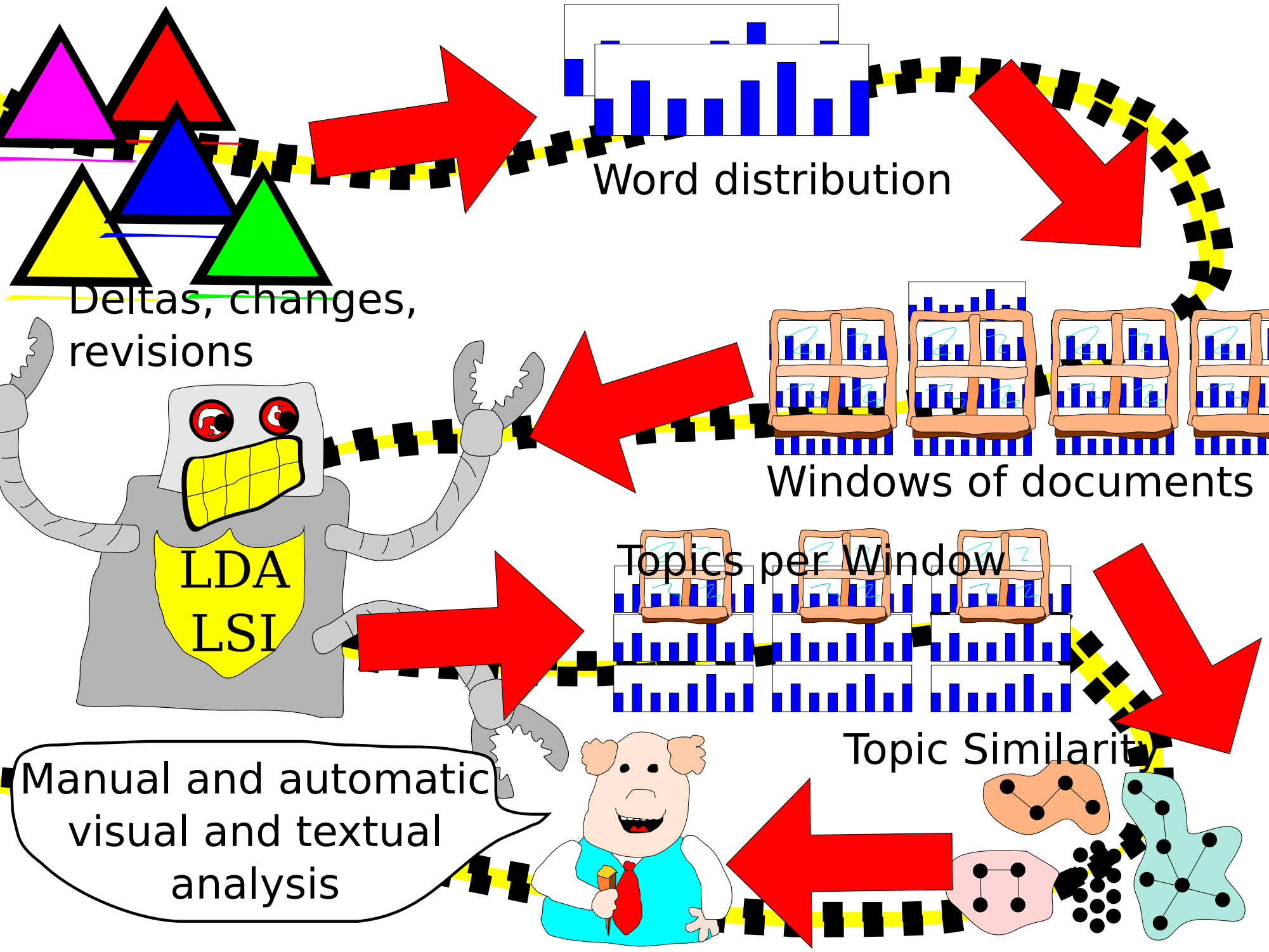
We could group the similar topics together



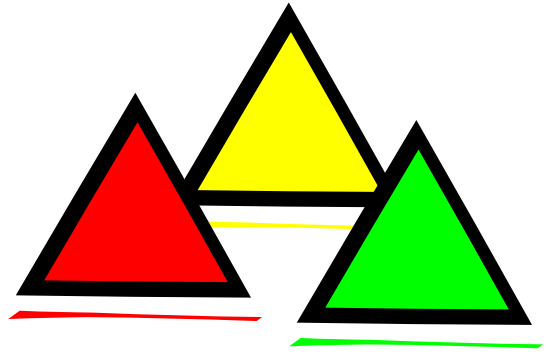
Our method of similarity:

* Compare top 10 most frequent words per topic
* Those that match beyond a threshold are immediately similar
* Otherwise 2 topics are similar if they are transitively related

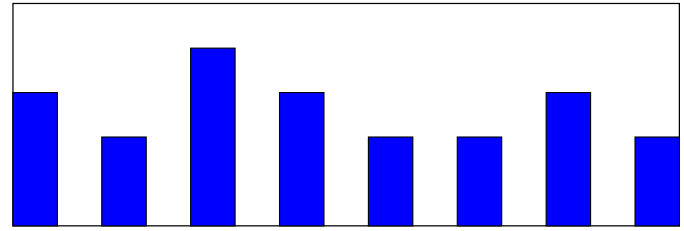



Word distribution
Deltas, changes, revisions
Windows of documents
LDA
LSI
Topics per Window
Topic Similarity
Manual and automatic visual and textual analysis

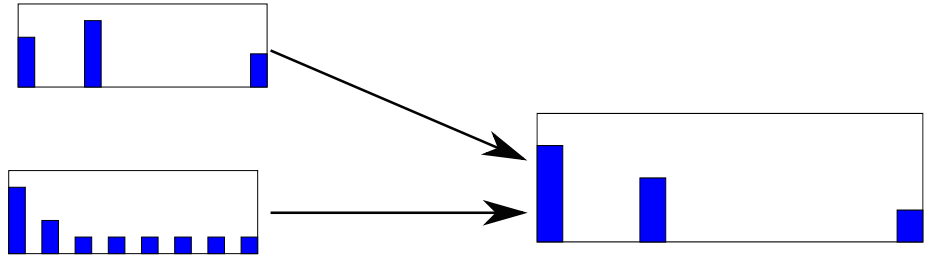
Message



Word Distribution

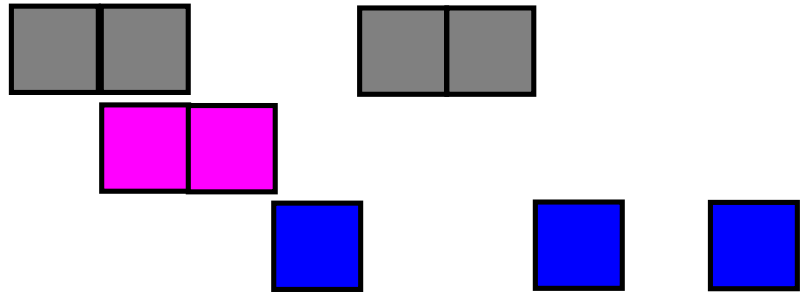


Topic

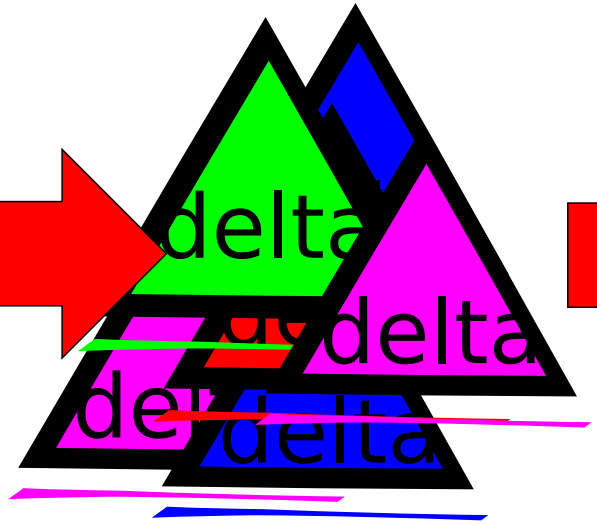


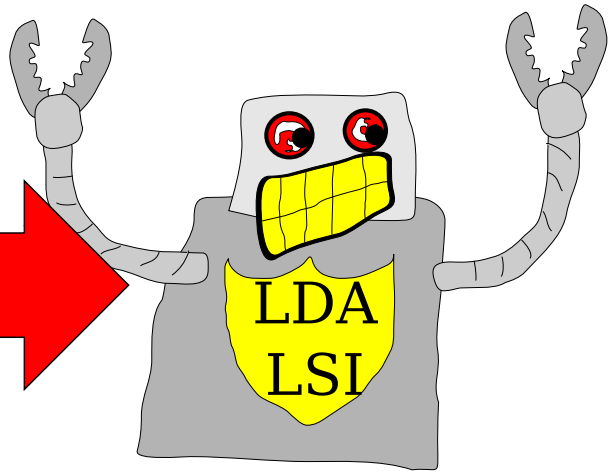
Top 10 Words:
* **perforce**
* **bug #**
* **POSIX**
* **Opteron**
* **...**

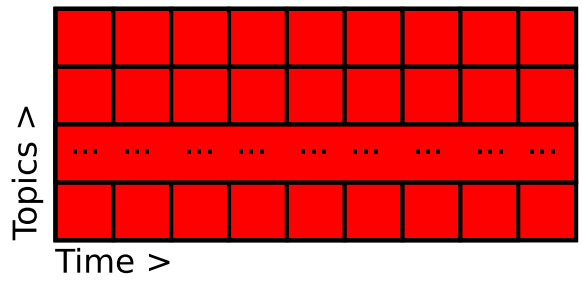
Trend



Initial Case Study

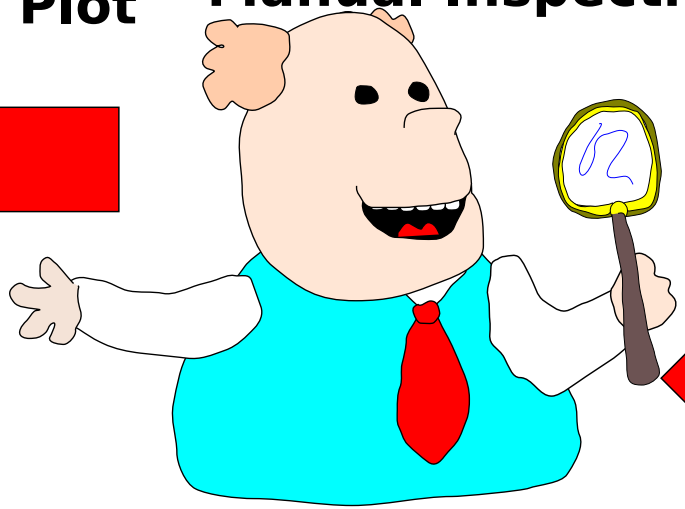




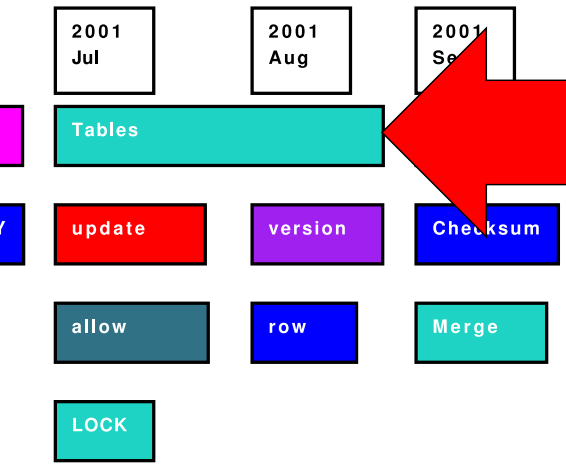


Extracted Topics

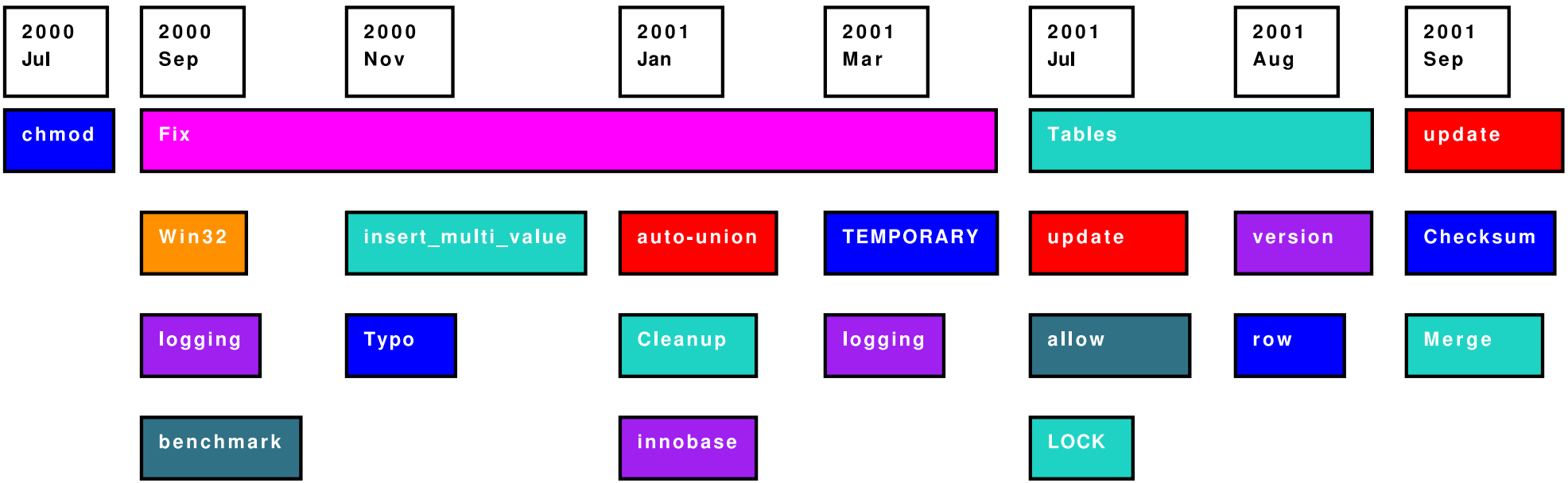
Manual Inspection



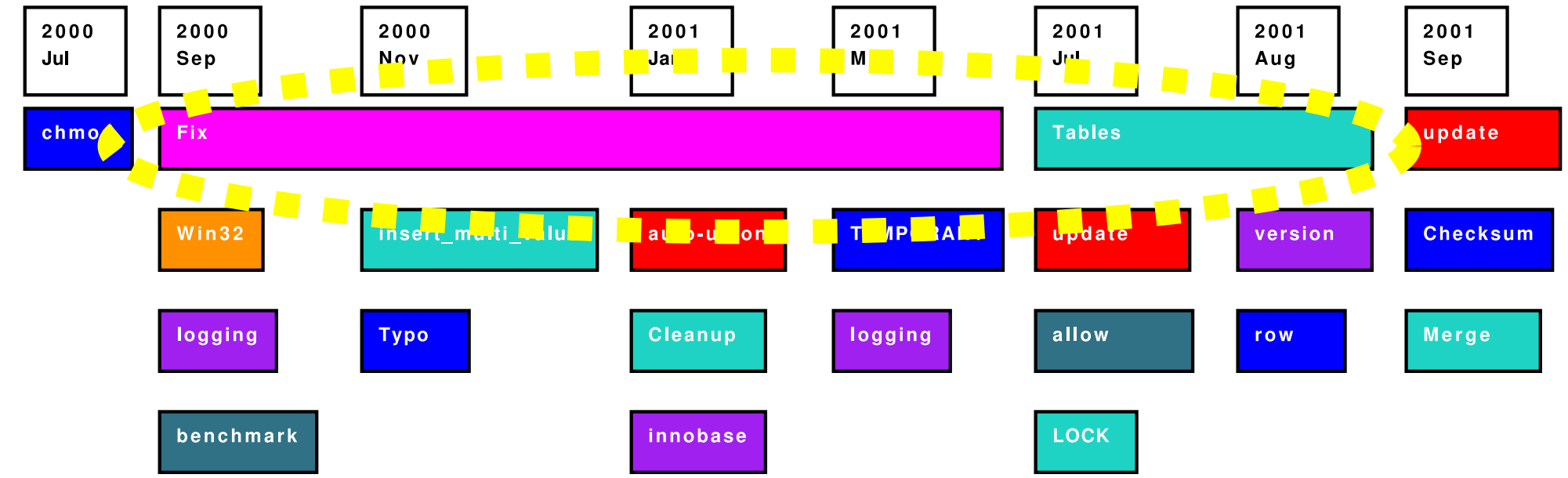
Manual Generated Plot

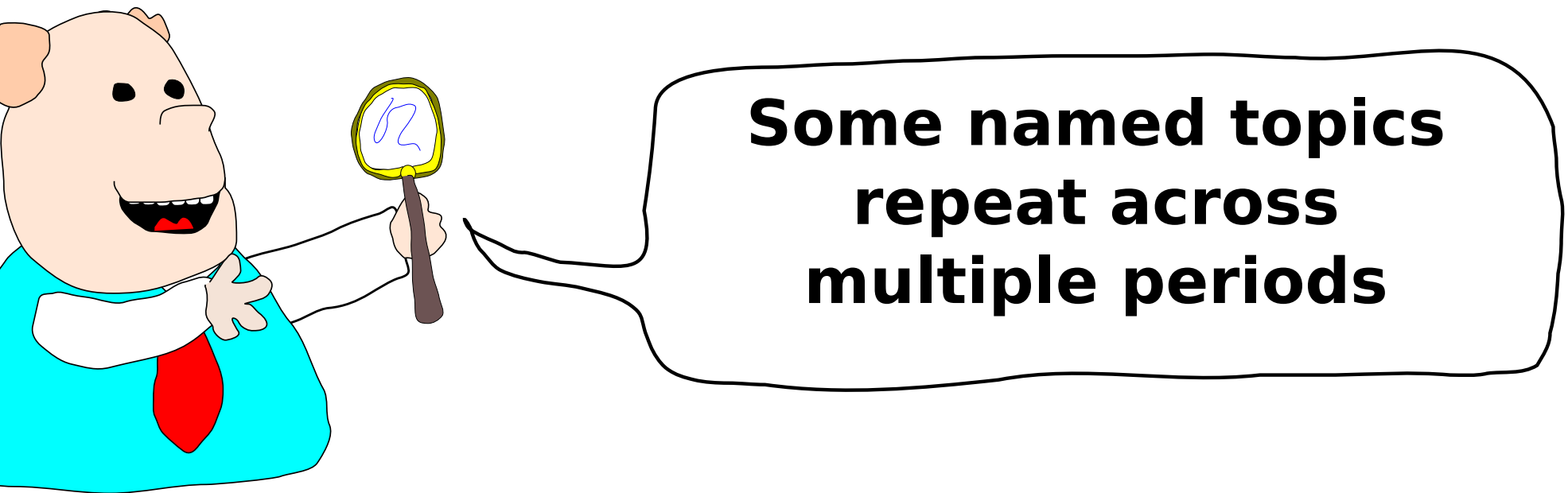


MySQL 3.23 Case Study

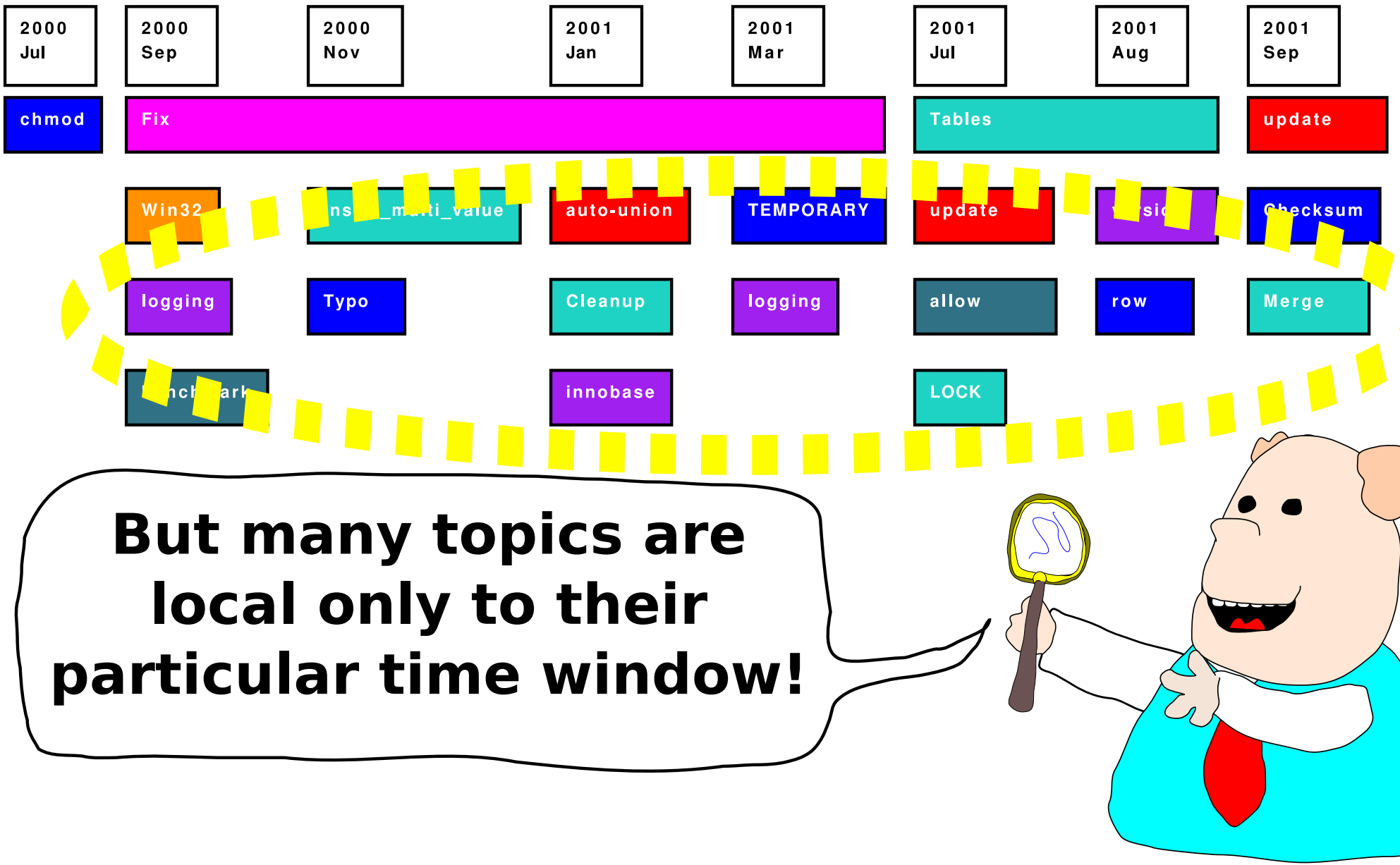


MySQL 3.23 Case Study

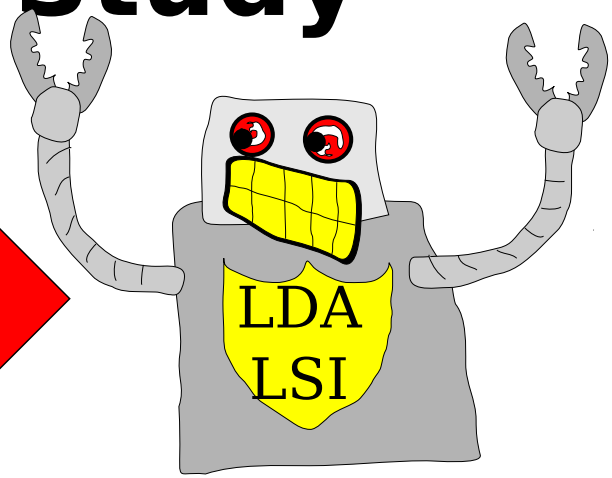




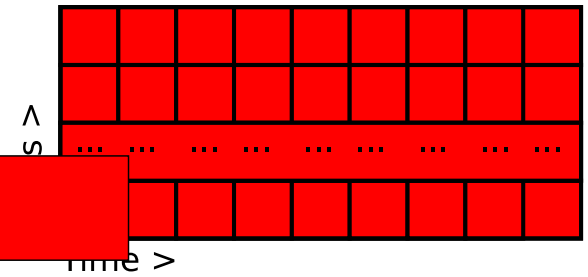
MySQL 3.23 Case Study



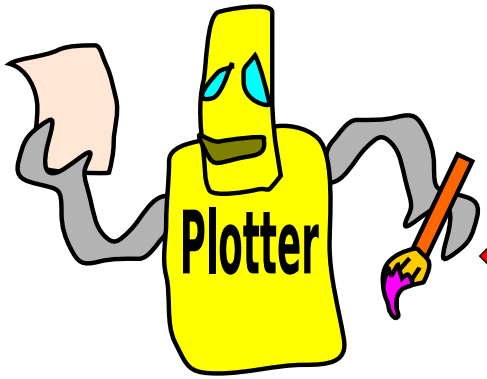
MaxDB 7.500 Case Study



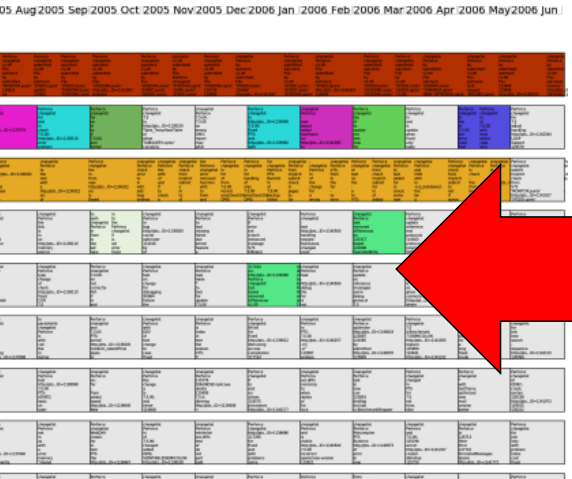
Extracted Topics



Automated Analysis

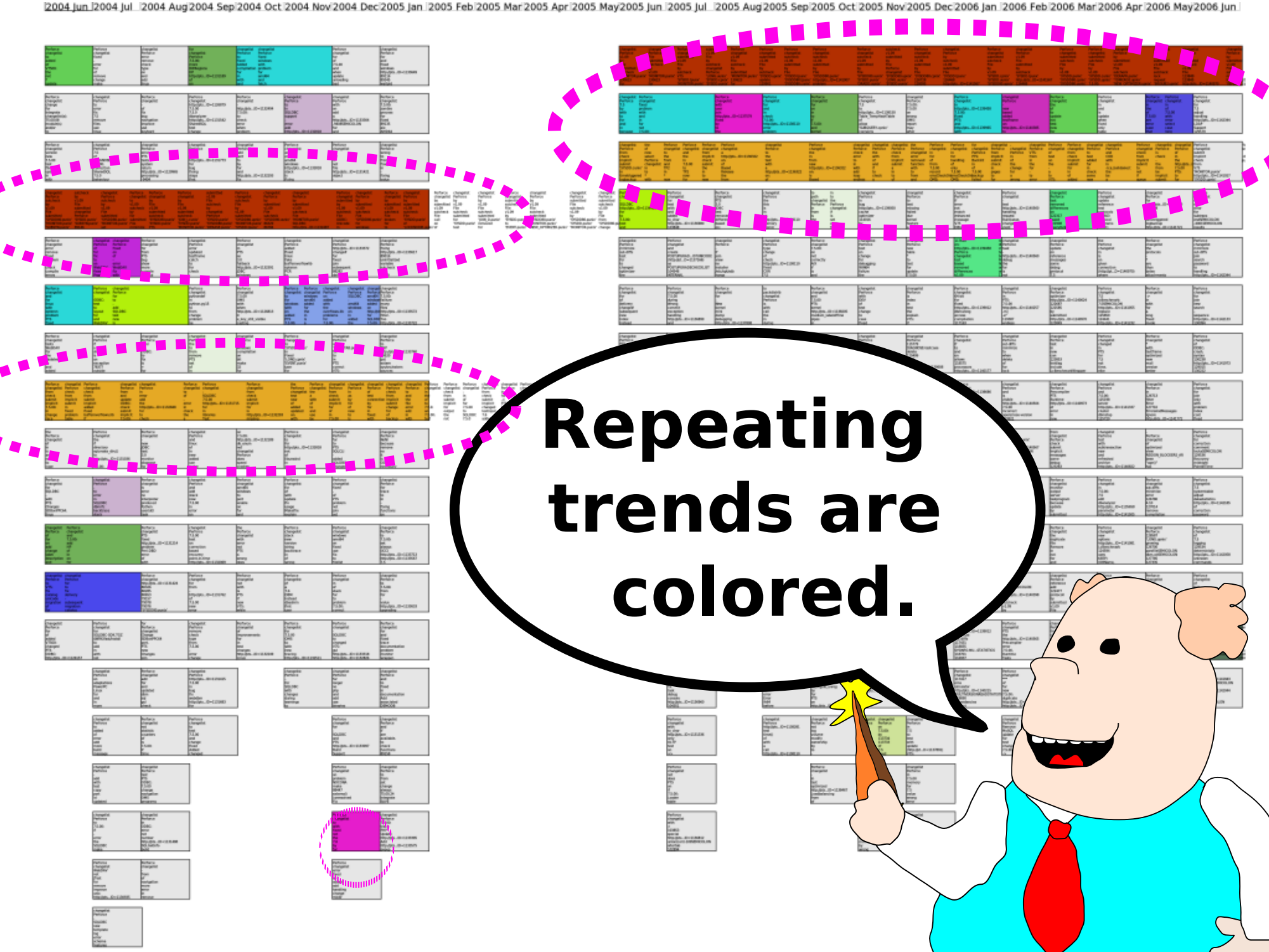


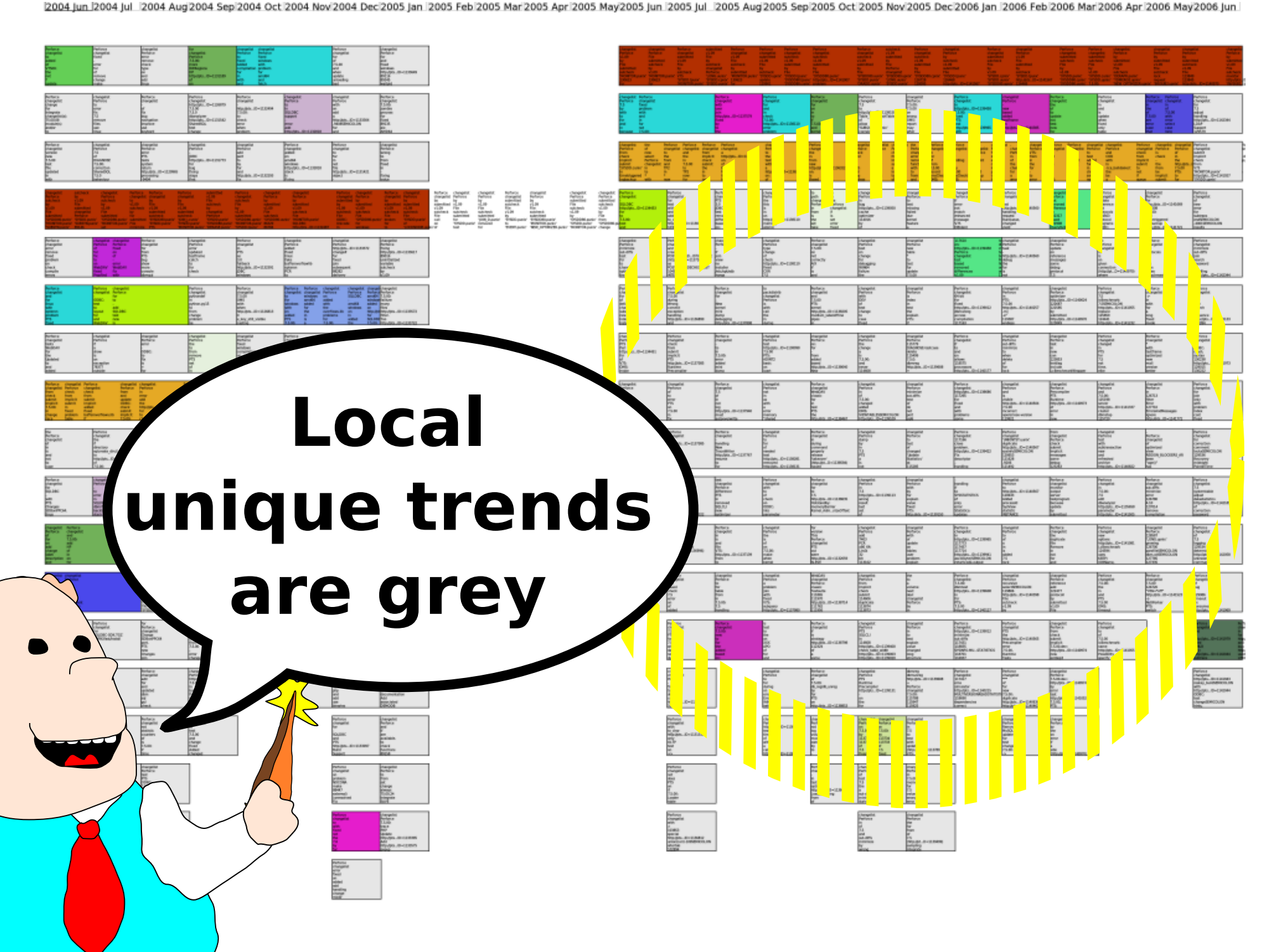
Auto-Generated Plot

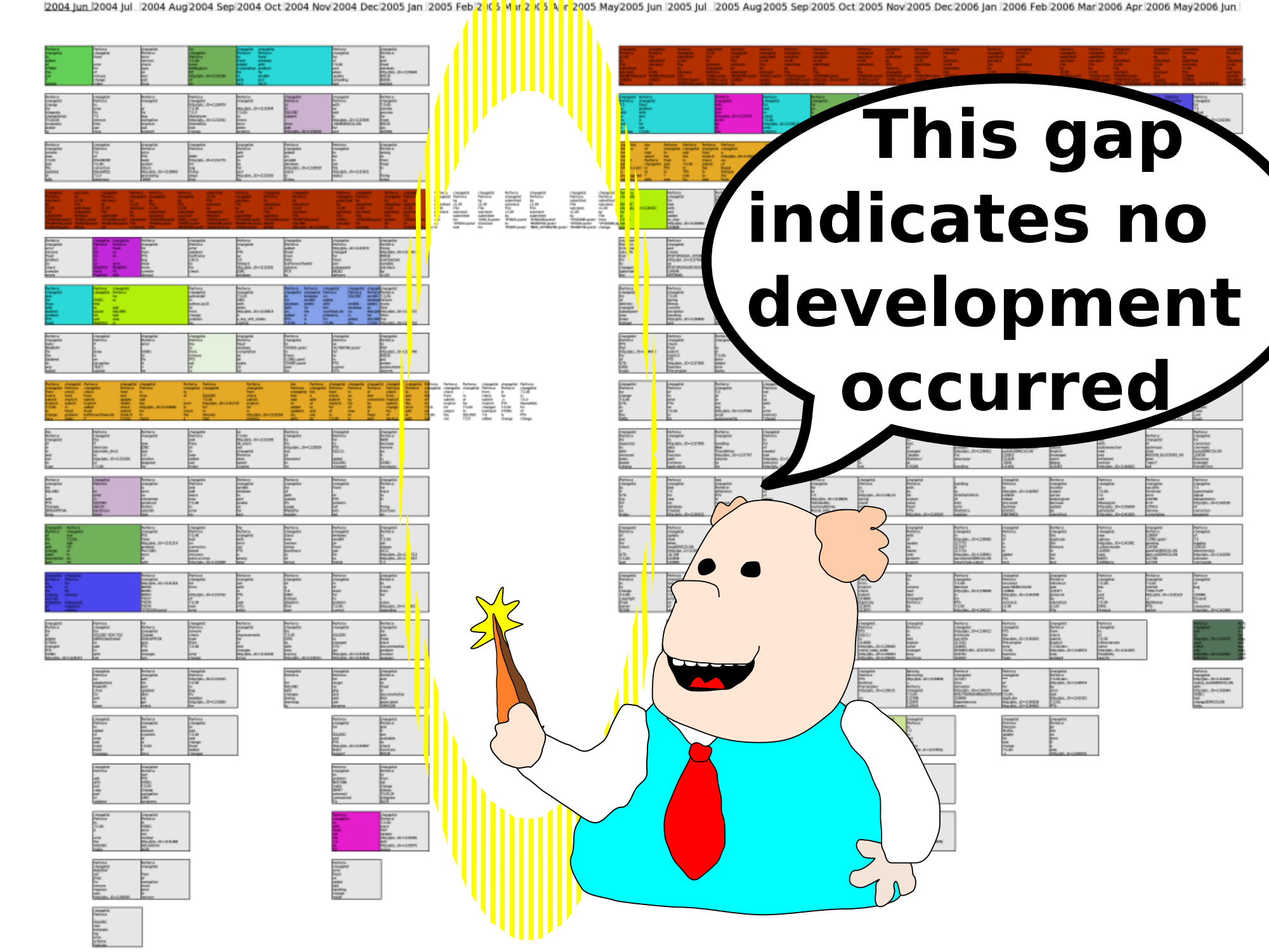


Many Documents
Few Documents
2004 June
2004 July
2004 August
2004 September
2004 October
2004 November
2004 December
2005 January
2005 February
2005 March
2005 April
2005 May
2005 June
2005 July
2005 August
2005 September
2005 October
2005 November
2005 December
2006 January
2006 February
2006 March
2006 April
2006 May
2006 June
2006 July
Topic 1
Topic 10
Topic 20


2004 Jun
2004 Jul
2004 Aug
2004 Sep
2004 Oct
2004 Nov
2004 Dec
2005 Jan
2005 Feb
2005 Mar
2005 Apr
2005 May
2005 Jun
2005 Jul
2005 Aug
2005 Sep
2005 Oct
2005 Nov
2005 Dec
2006 Jan
2006 Feb
2006 Mar
2006 Apr
2006 May
2006 Jun


2004 Jun 2004 Jul 2004 Aug 2004 Sep 2004 Oct 2004 Nov 2004 Dec 2005 Jan 2005 Feb 2005 Mar 2005 Apr 2005 May 2005 Jun 2005 Jul 2005 Aug 2005 Sep 2005 Oct 2005 Nov 2005 Dec 2006 Jan 2006 Feb 2006 Mar 2006 Apr 2006 May 2006 Jun
Repeating trends are colored.


2004 Jun 2004 Jul 2004 Aug 2004 Sep 2004 Oct 2004 Nov 2004 Dec 2005 Jan 2005 Feb 2005 Mar 2005 Apr 2005 May 2005 Jun 2005 Jul 2005 Aug 2005 Sep 2005 Oct 2005 Nov 2005 Dec 2006 Jan 2006 Feb 2006 Mar 2006 Apr 2006 May 2006 Jun
Local unique trends are grey


2004 Jun 2004 Jul 2004 Aug 2004 Sep 2004 Oct 2004 Nov 2004 Dec 2005 Jan 2005 Feb 2005 May 2005 Jun 2005 Jul 2005 Aug 2005 Sep 2005 Oct 2005 Nov 2005 Dec 2006 Jan 2006 Feb 2006 Mar 2006 Apr 2006 May 2006 Jun
This gap indicates no development occurred

changelist Perforce sutcheck by v1.09 File submitted 'SYSDD80.punix' 'MONITOR.punix' 'SUBSYN.punix'	sutcheck by v1.09 File submitted changelist Perforce 'SYSDD.punix' 'SUBSYN.punix' 89146	changelist Perforce sutcheck by v1.09 File submitted 'SYSDD80.punix' 'MONITOR.punix' set	Perforce changelist to v1.09 sutcheck by for submitted File minimize	Perforce changelist by File v1.09 submitted sutcheck 'SYSDD80.punix' 'SYSDD.punix' PTS	Perforce changelist by File v1.09 submitted sutcheck 'JOIN_J.cunix' 'SYSDD.punix' 'MONITOR.punix'	submitted v1.09 File sutcheck by changelist Perforce 'SYSDD.punix' 'SYSDD80.punix' 'DERASE.punix'	Perforce changelist sutcheck File by v1.09 submitted 'SYSDD80.upnix' 'MONITOR.punix' 'SYSDD.tpunx'	changelist Perforce by submitted v1.09 File sutcheck 'SYSDD80.uunix' 86329 86784	changelist Perforce by submitted v1.09 sutcheck File 'MONITOR.punix' SQLDBC http://pts...ID=1132487	Perforce changelist submitted by v1.09 sutcheck File ODBC: Unicode in	changelis Perforce by submitte v1.09 sutcheck File 'SYSDD8 for windows

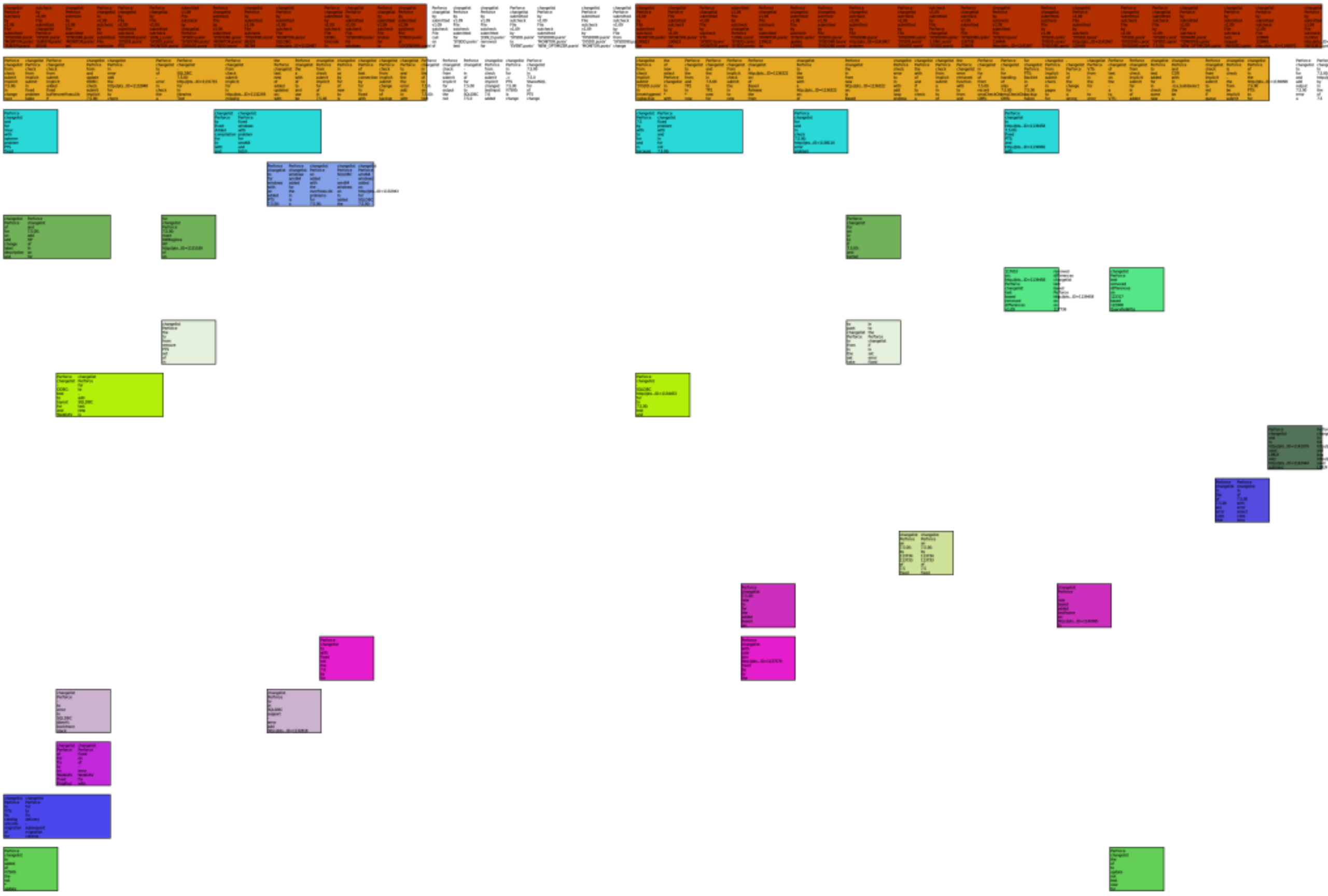
Perforce changelist error remove fixed problem on check compile errors	changelist Perforce of for fix to on WebDAV fixed Propfind	changelist Perforce fixed on of - error WebDAV fix with	Perforce changelist for from PTS bug show more console dbmgui	changelist Perforce error updated testframe catch to for check -	Perforce changelist in PTS as 3.0 fallback http://pts...ID=1132391 JDBC windows	changelist Perforce added fixed linux fake bufferoverflowlib opteron PCR to	changelist Perforce http://pts...ID=1133372 change# for fixed join subsequent 88282 delivery

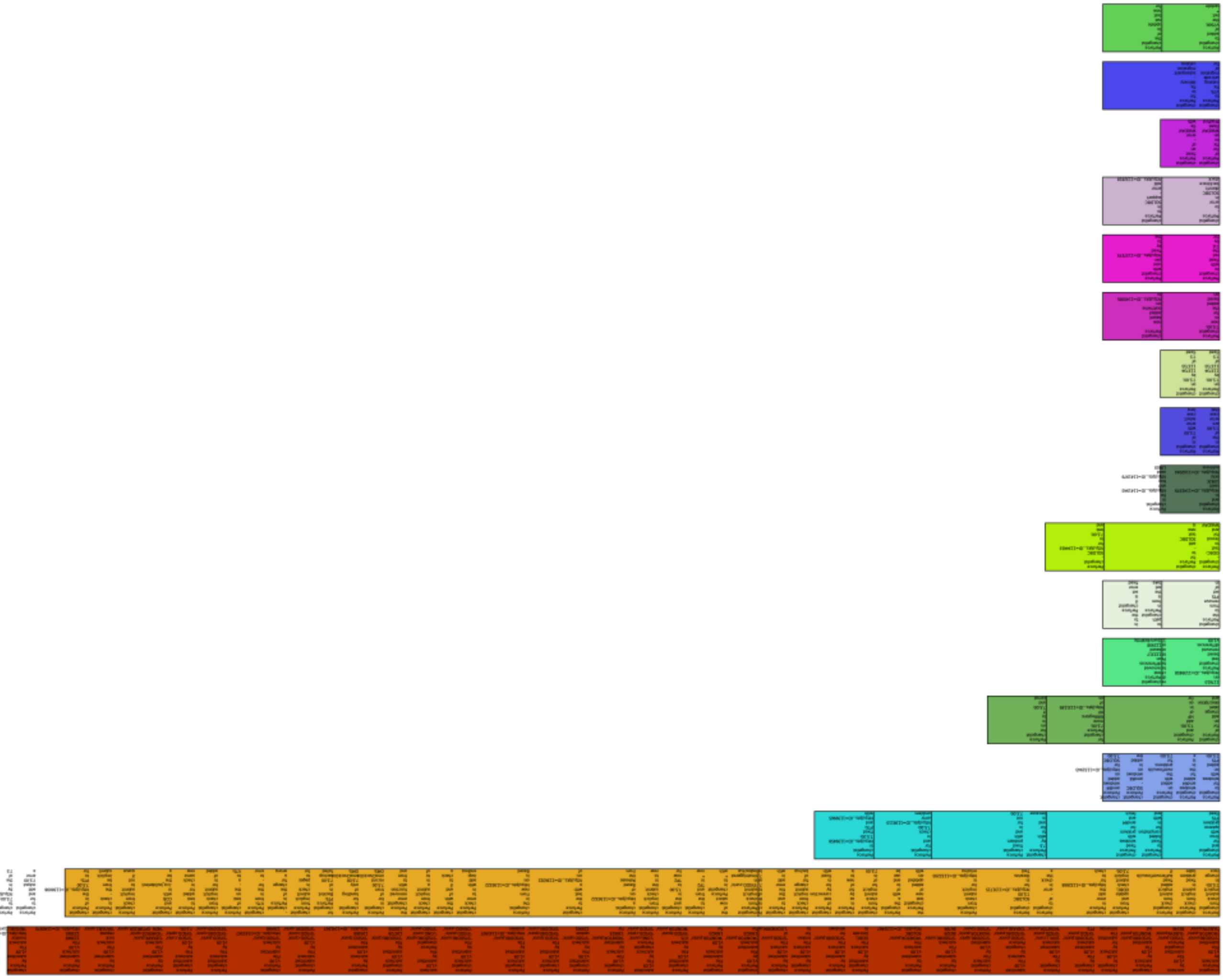
Perforce changelist and for linux with opteron problem PTS fixed	Perforce changelist - ODBC: test to layout for and WebDAV	changelist Perforce for to - add SQLDBC test new is	Perforce changelist pythondef in python-py15 to from change problem be	Perforce changelist 7.5.00 OMS with when http://pts...ID=1126813 0 e_key_still_visible casting	Perforce changelist to for windows with on added PTS 7.5.00:	Perforce changelist windows amd64 added for the in is a	changelist Perforce on added with the overflowu.lib problems for 7.5.00:	changelist Perforce SQLDBC - amd64 windows on in added the	c P a w a o h f S 7.

Perforce changelist tests WebDAV for the Updated to and added	changelist Perforce if a show is an exception 78377 outside	changelist Perforce error - ODBC: to fix is + for	changelist Perforce the to from remove PTS set of in	changelist Perforce fixed windows compilation on bit make 32 for	Perforce changelist to 'SYSDD.upnix' in Fixed 'LONG.cpnix' 'EVENT.punix' user for	changelist Perforce on 'ALTERTAB.punix' for to in PTS correct V75:					

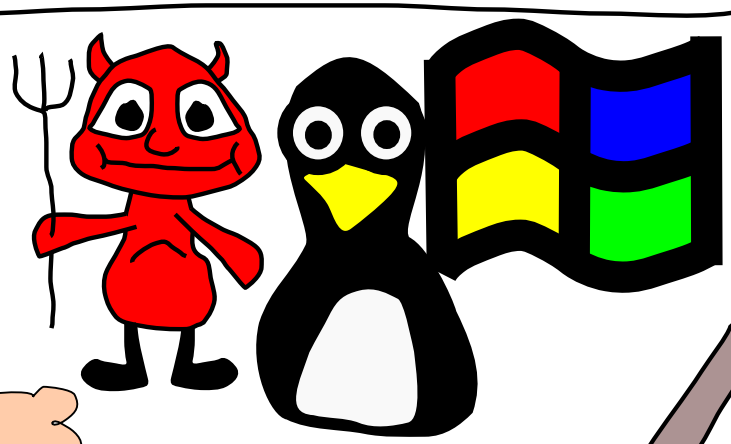
Perforce changelist from check submit implicit 7.5.00: in change race	changelist Perforce check from implicit submit in fixed problem table	Perforce changelist check from submit implicit added fixed bufferoverflowu.lib if	changelist Perforce from and update ODBC: check submit implicit 7.5.00:	changelist Perforce in error add the http://pts...ID=1132848 for to check	Perforce changelist to of - error in check the a	Perforce changelist - SQLDBC 7.5.00 http://pts...ID=1131715 for in libraries Test	Perforce changelist from check submit implicit for is http://pts...ID=1132293 missing	the Perforce changelist test new of added updated on: with	Perforce changelist the a with of to and use be	changelist Perforce from check submit implicit for of in 7.5.00	changelist Perforce in as by for of new to is	c P o t c b f i w

2004 Jun 2004 Jul 2004 Aug 2004 Sep 2004 Oct 2004 Nov 2004 Dec 2005 Jan 2005 Feb 2005 Mar 2005 Apr 2005 May 2005 Jun 2005 Jul 2005 Aug 2005 Sep 2005 Oct 2005 Nov 2005 Dec 2006 Jan 2006 Feb 2006 Mar 2006 Apr 2006 May 2006 Jun

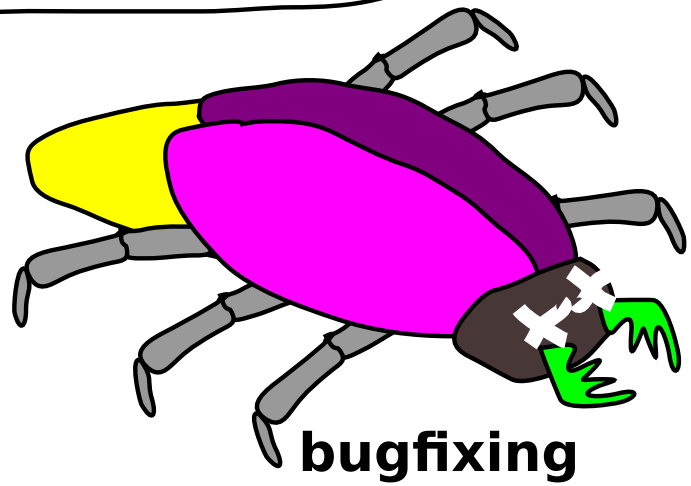




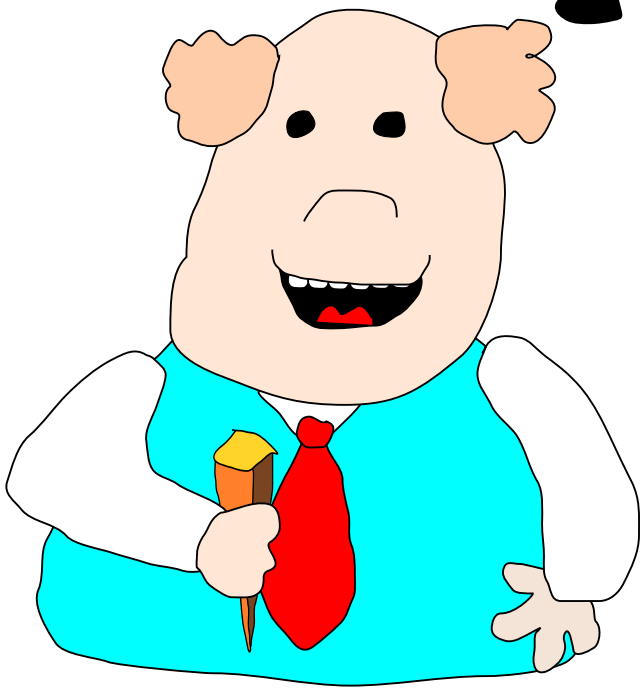
We can name topics based on certain kinds of words

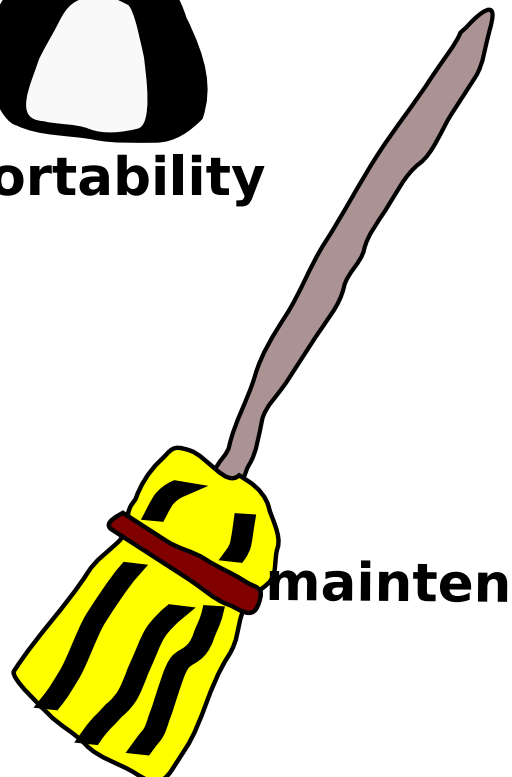


portability



bugfixing



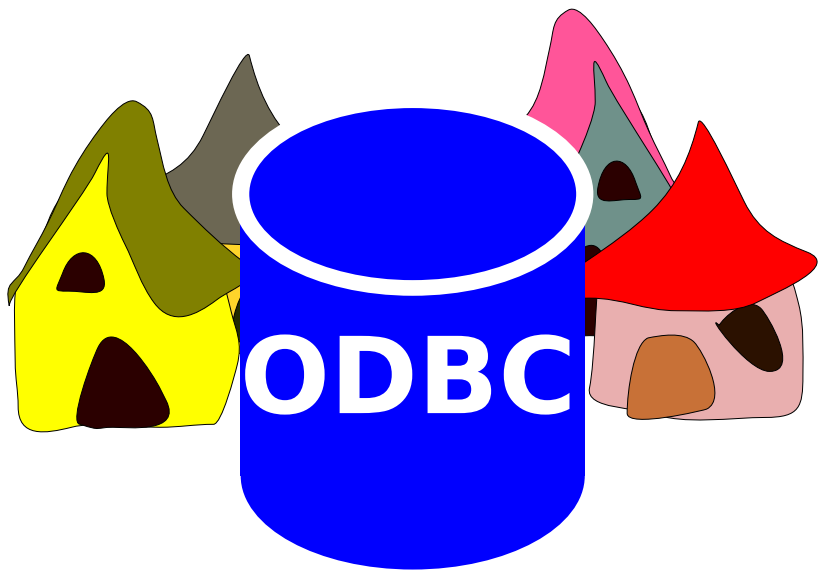


maintenance

'ilities

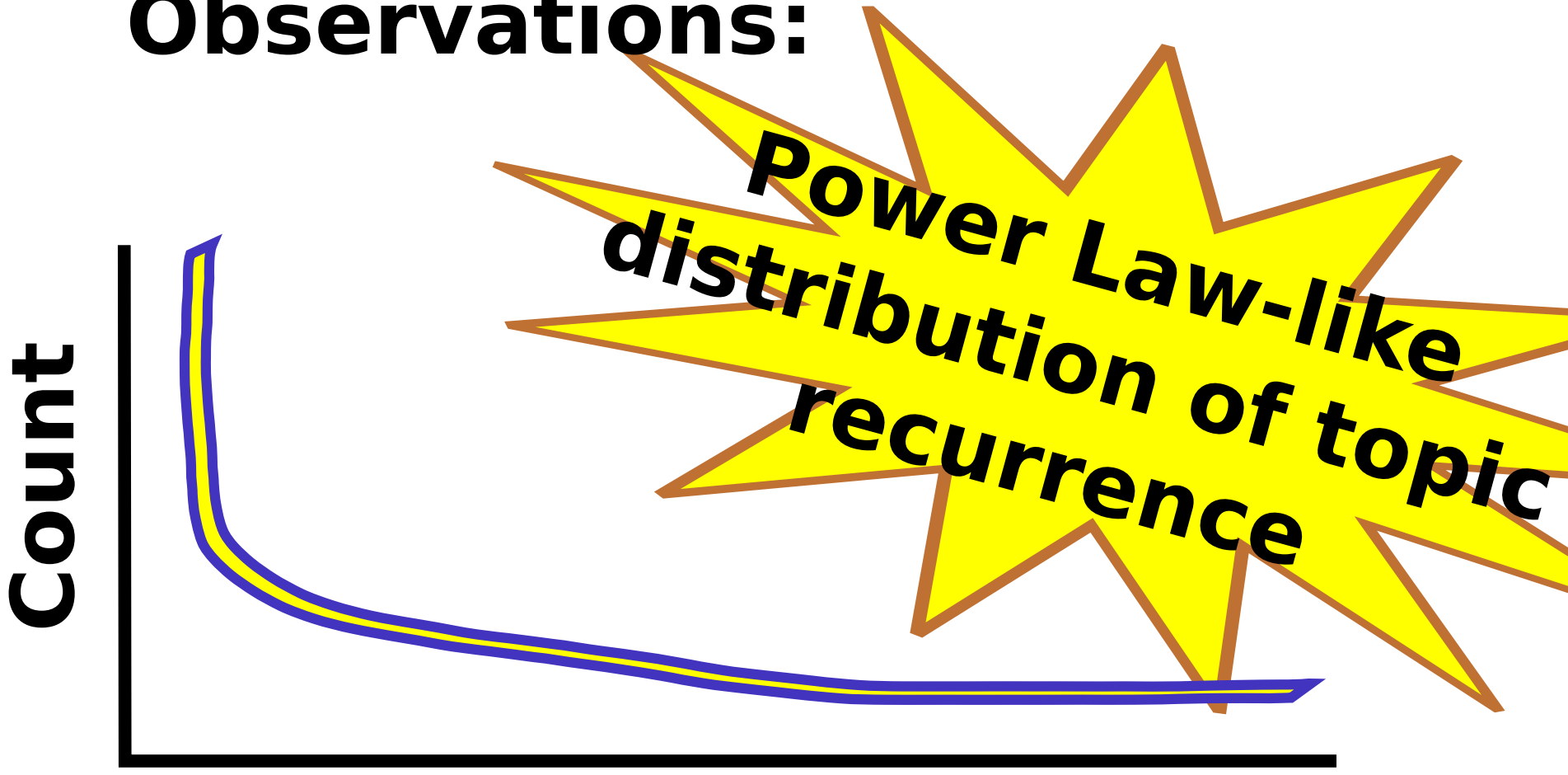
Observations

* some important local topics were not found with global topic analysis

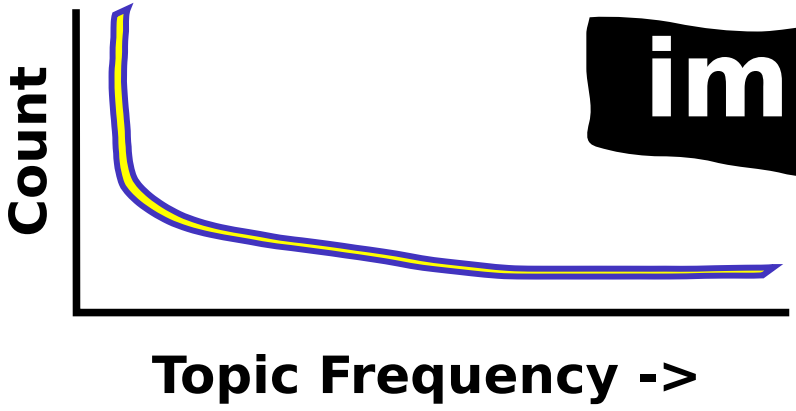


vs

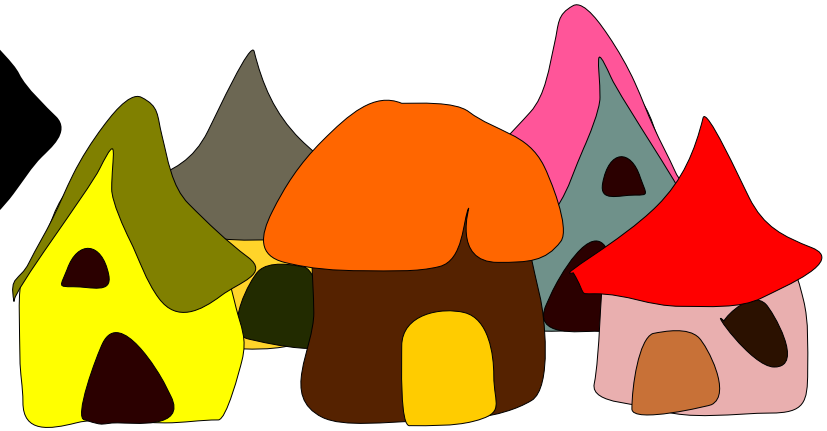
Observations:



Observations:

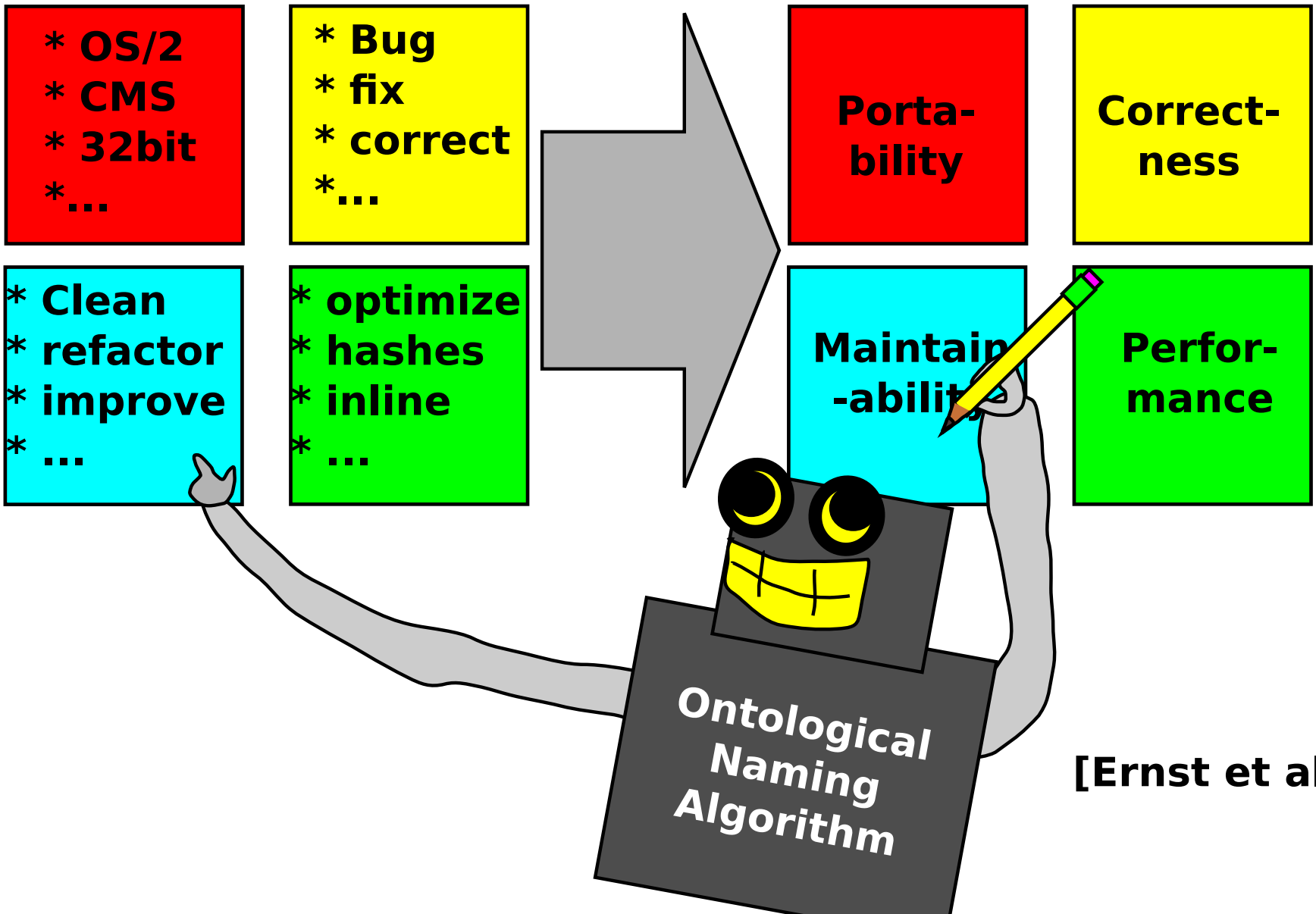


implies



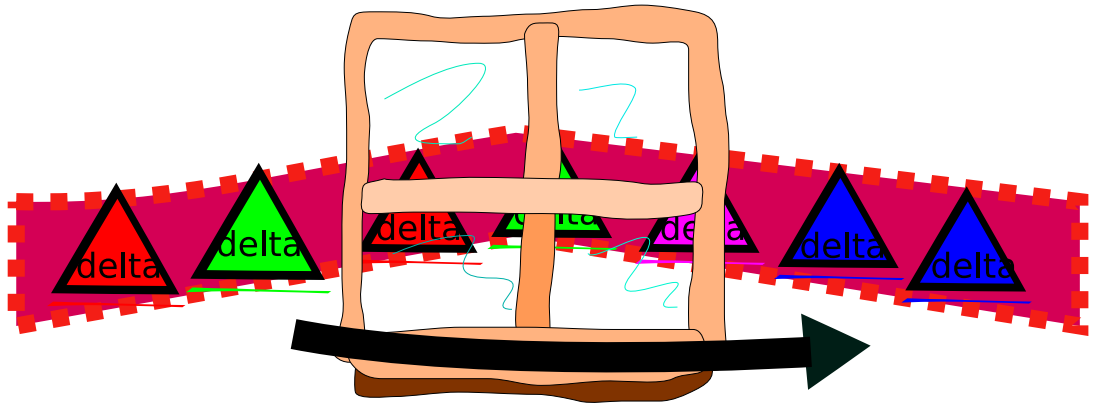
with some

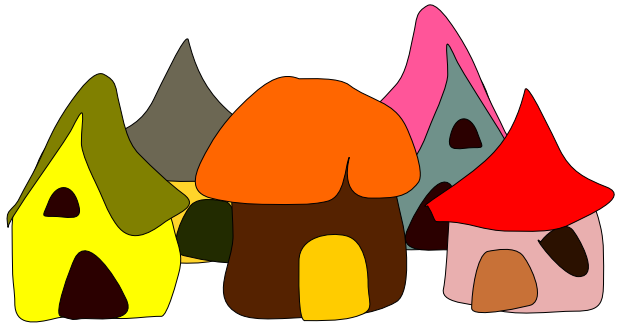
Future Work



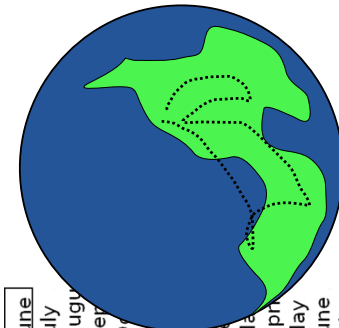
[Ernst et al.]

Conclusions





vs



vs