Automated Topic Naming to Support Cross-project Analysis of Software Maintenance Activities

Abram Hindle
Dept. of Computer Science
University of California, Davis
Davis, CA, USA
abram@softwareprocess.es

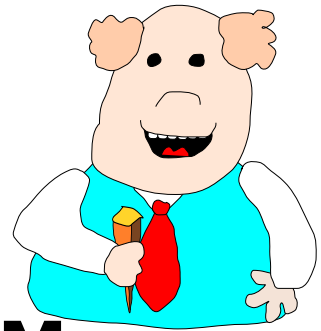
Neil A. Ernst
Dept. of Computer Science
University of Toronto
Toronto, Ontario, CANADA
nernst@cs.toronto.edu

Michael W. Godfrey
David Cheriton School of
Computer Science
University of Waterloo
Waterloo, Ontario, CANADA
migod@uwaterloo.ca

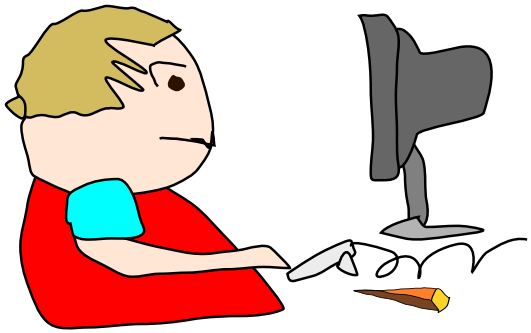
John Mylopoulos
Dept. Information Eng. and
Computer Science
University of Trento
Trento, ITALY
jm@disi.unitn.it

US NSF SHF Medium 0964703

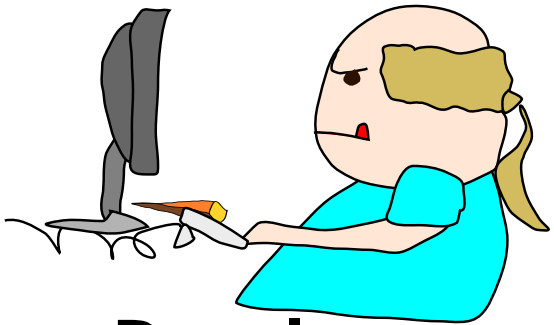
Who Cares About Quality?



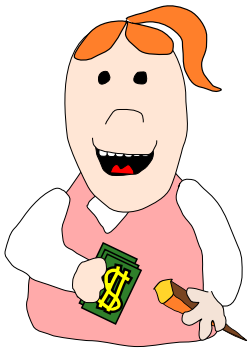
Managers



New Developers



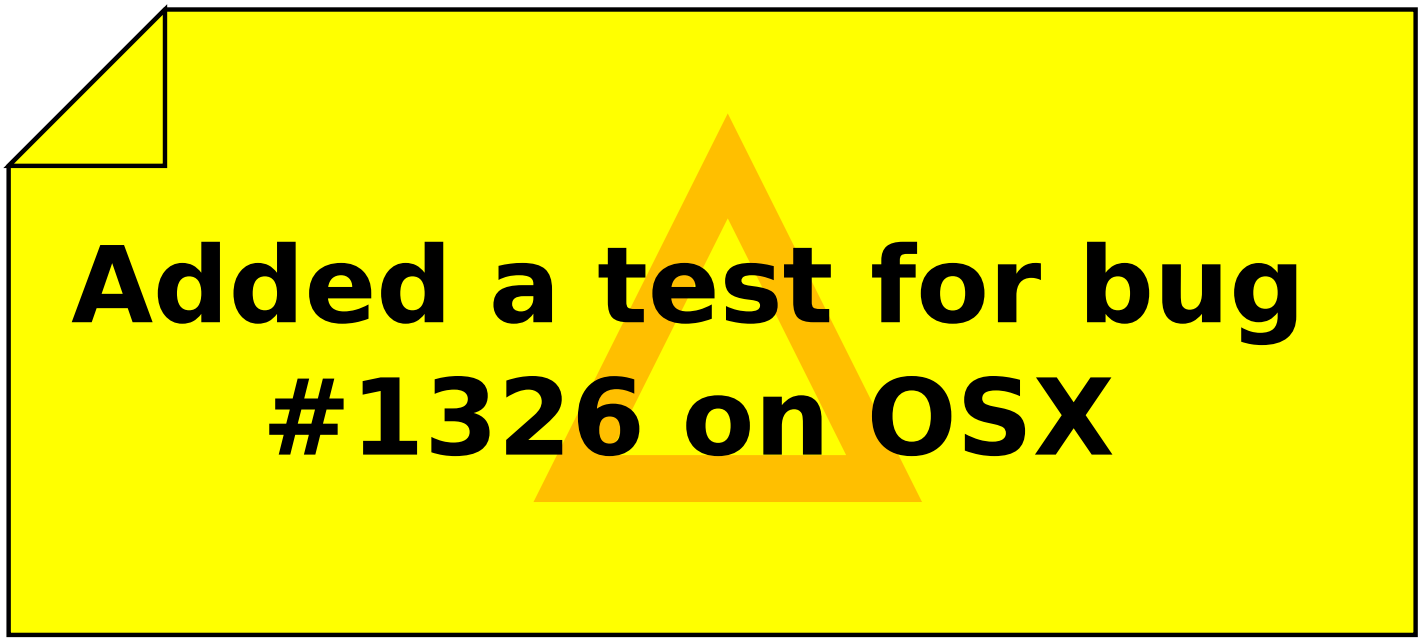
Developers



Investors

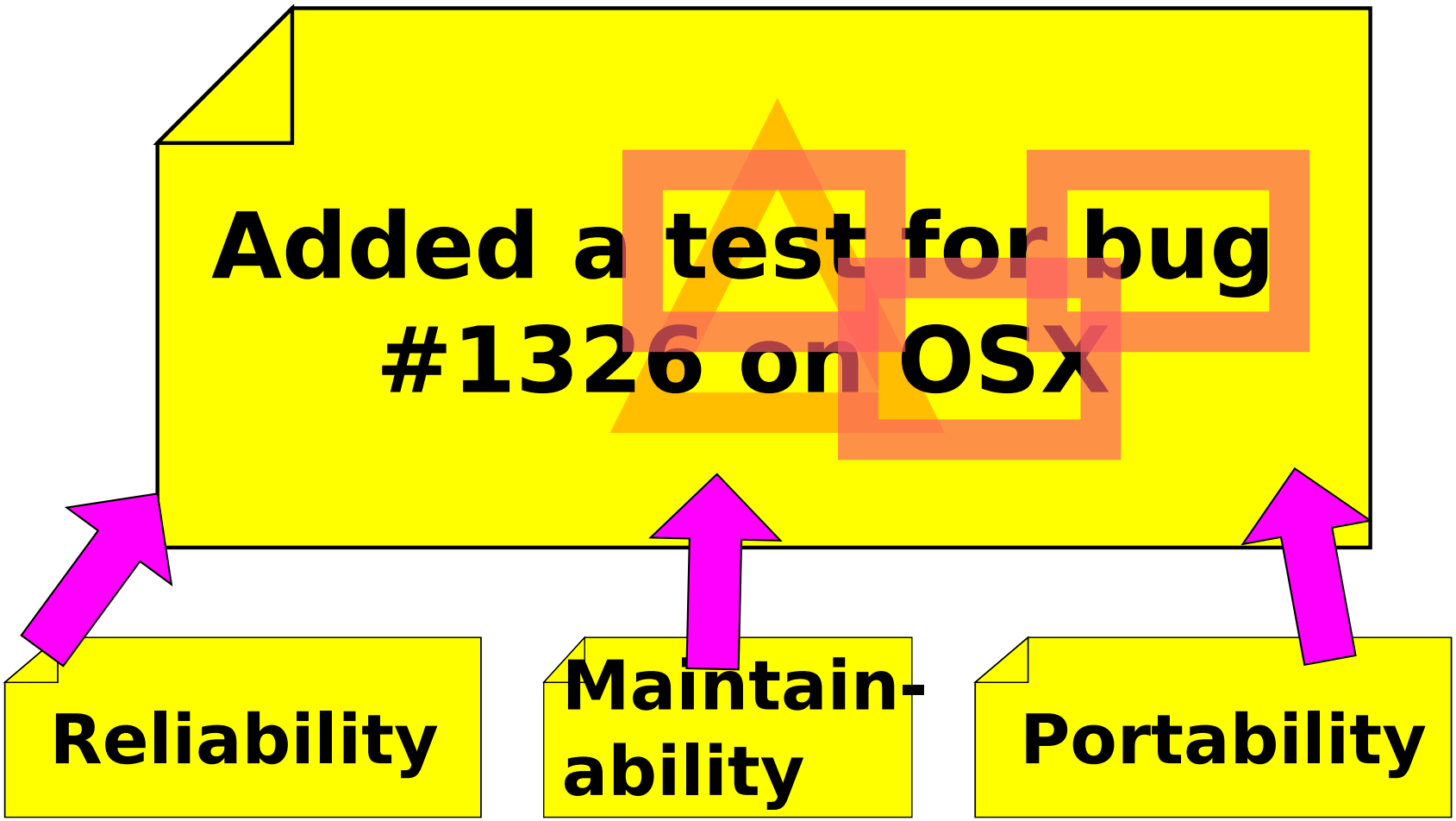
Customers

What is this commit about?

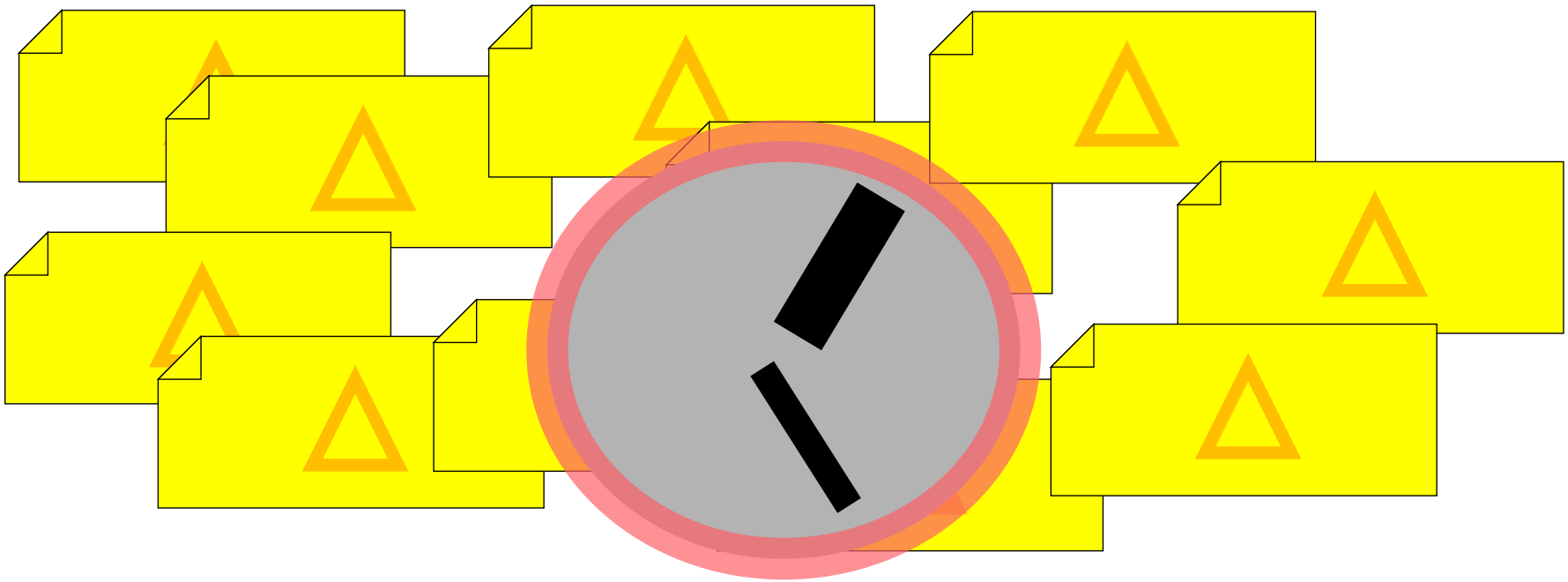


What is this commit about?

What is this commit about?



But we have many commits..

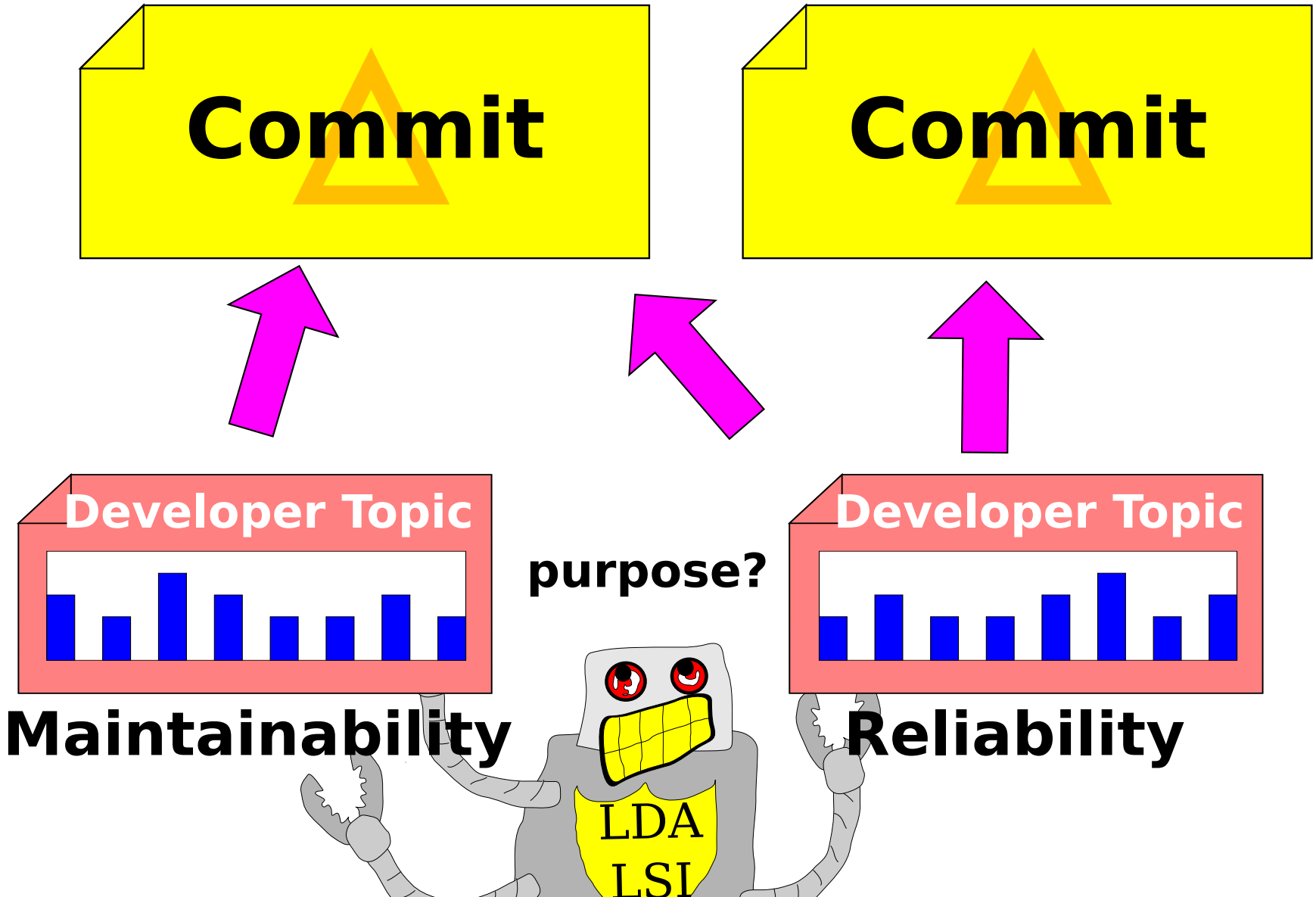


Reliability

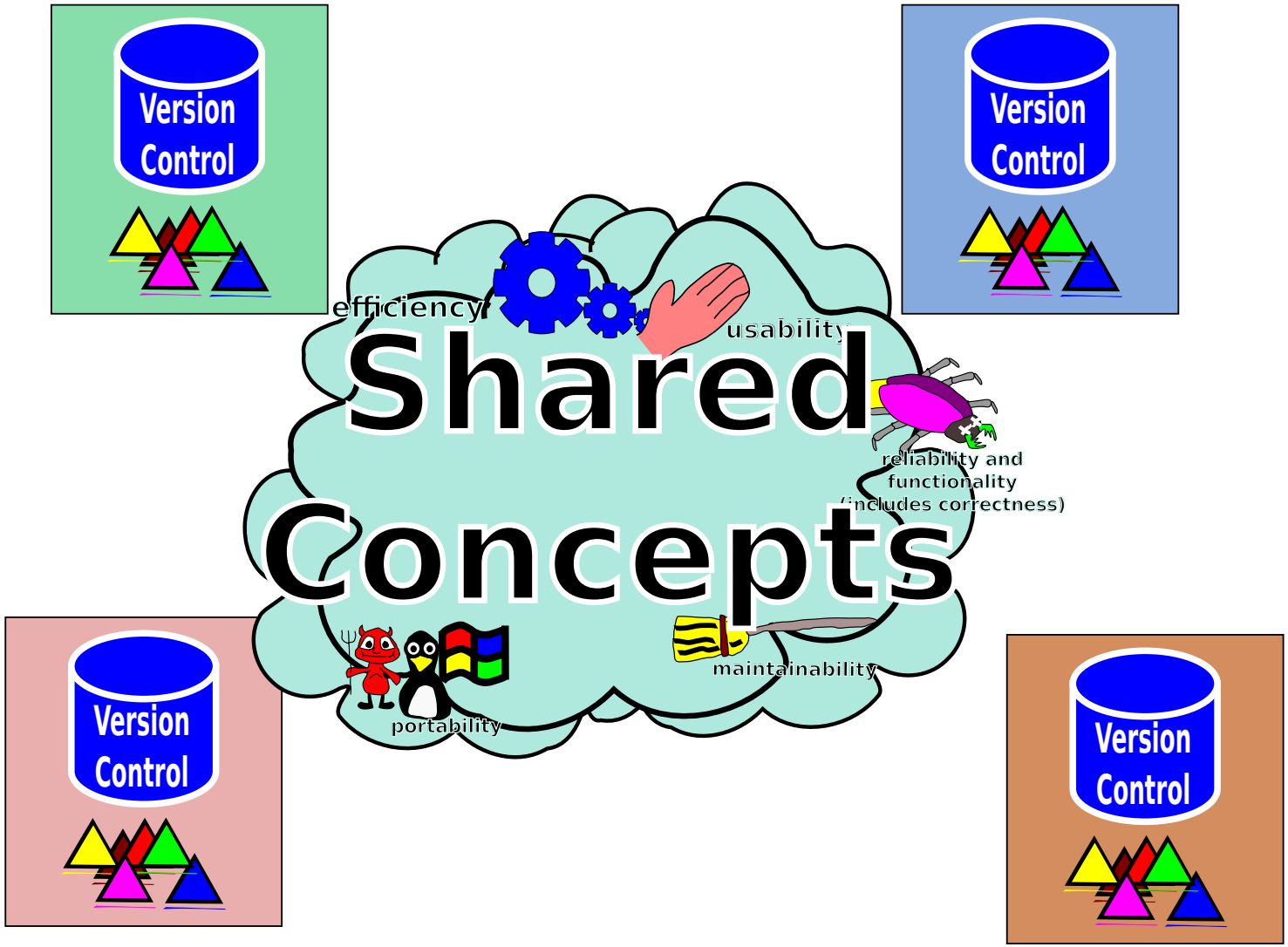
Maintain-ability

Portability

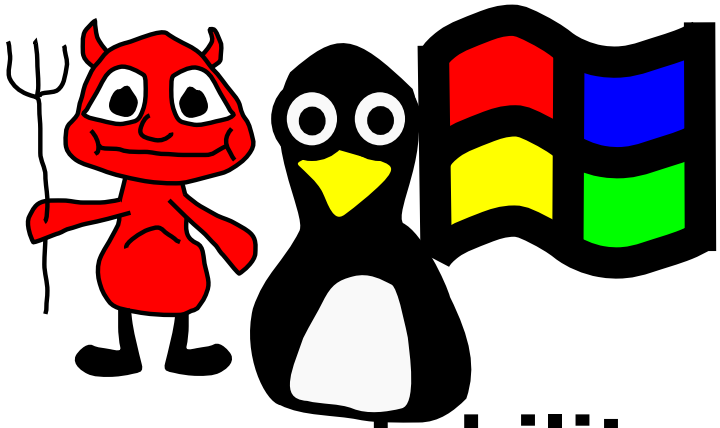
Developer Topics



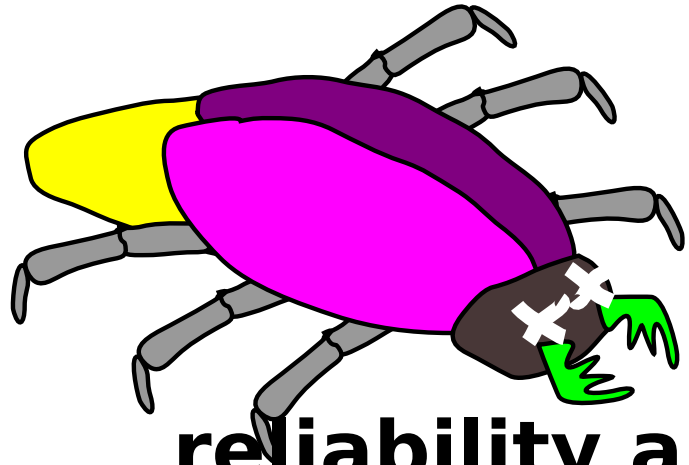
Cross Project Relevance



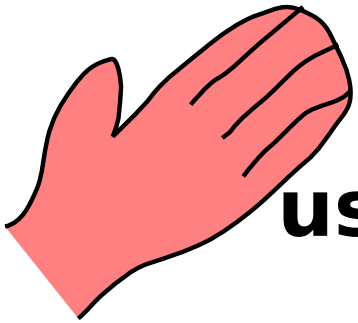
Quality-related Non Functional Requirements (NFRs)



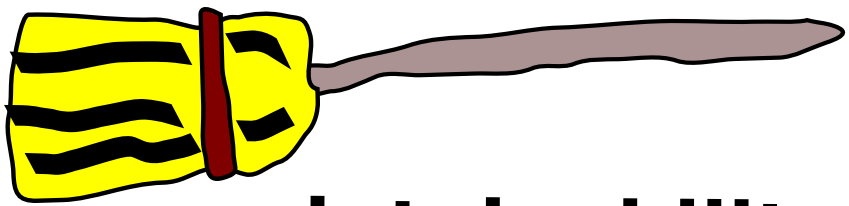
portability



reliability and functionality (includes correctness)

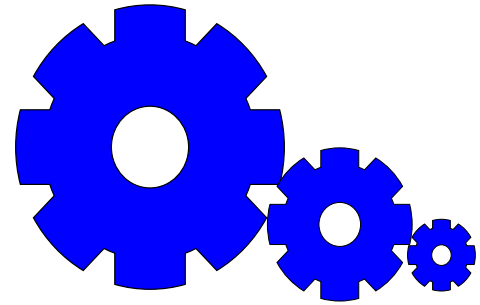


usability



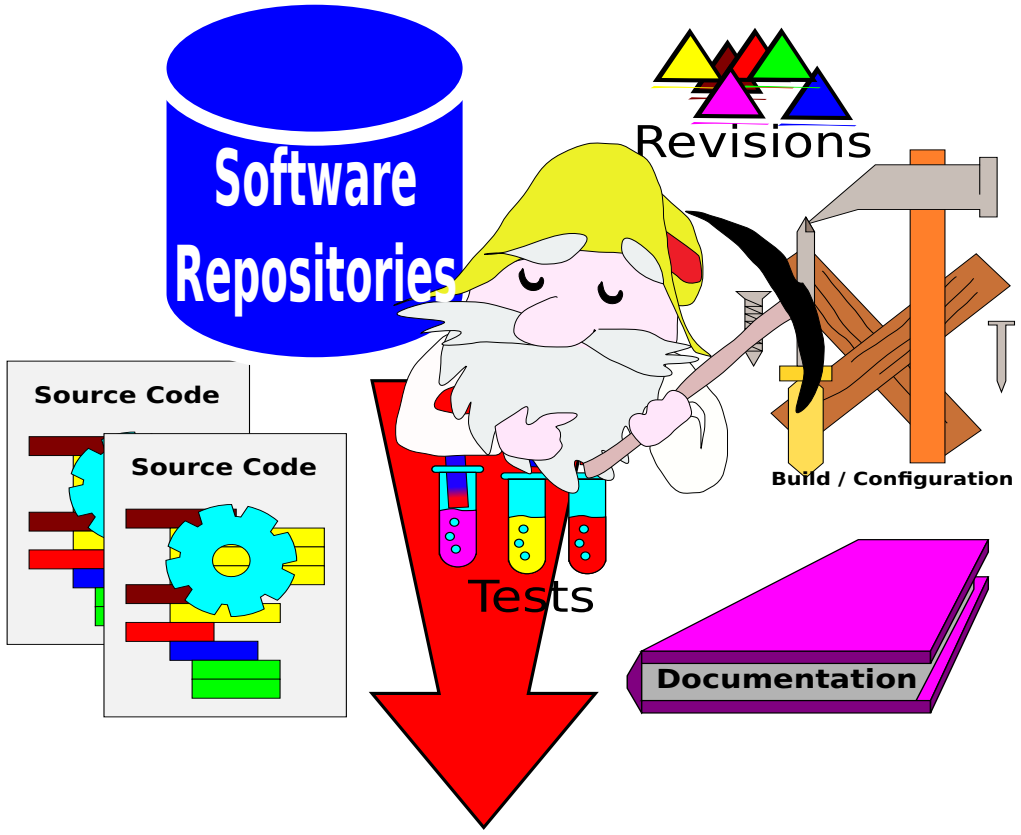
maintainability

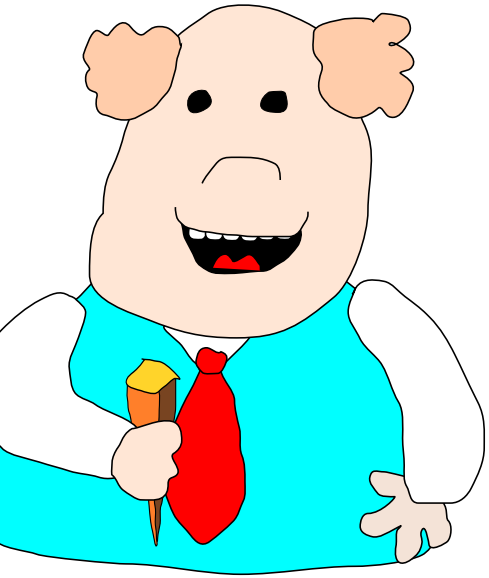
efficiency



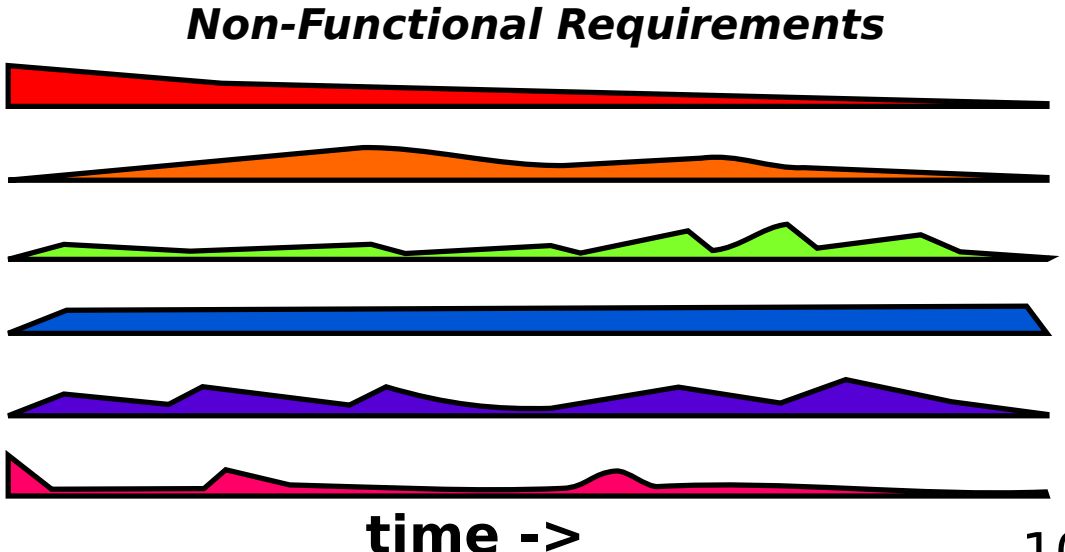
[iso9126]
[cleland-huang03]
[ernst10]

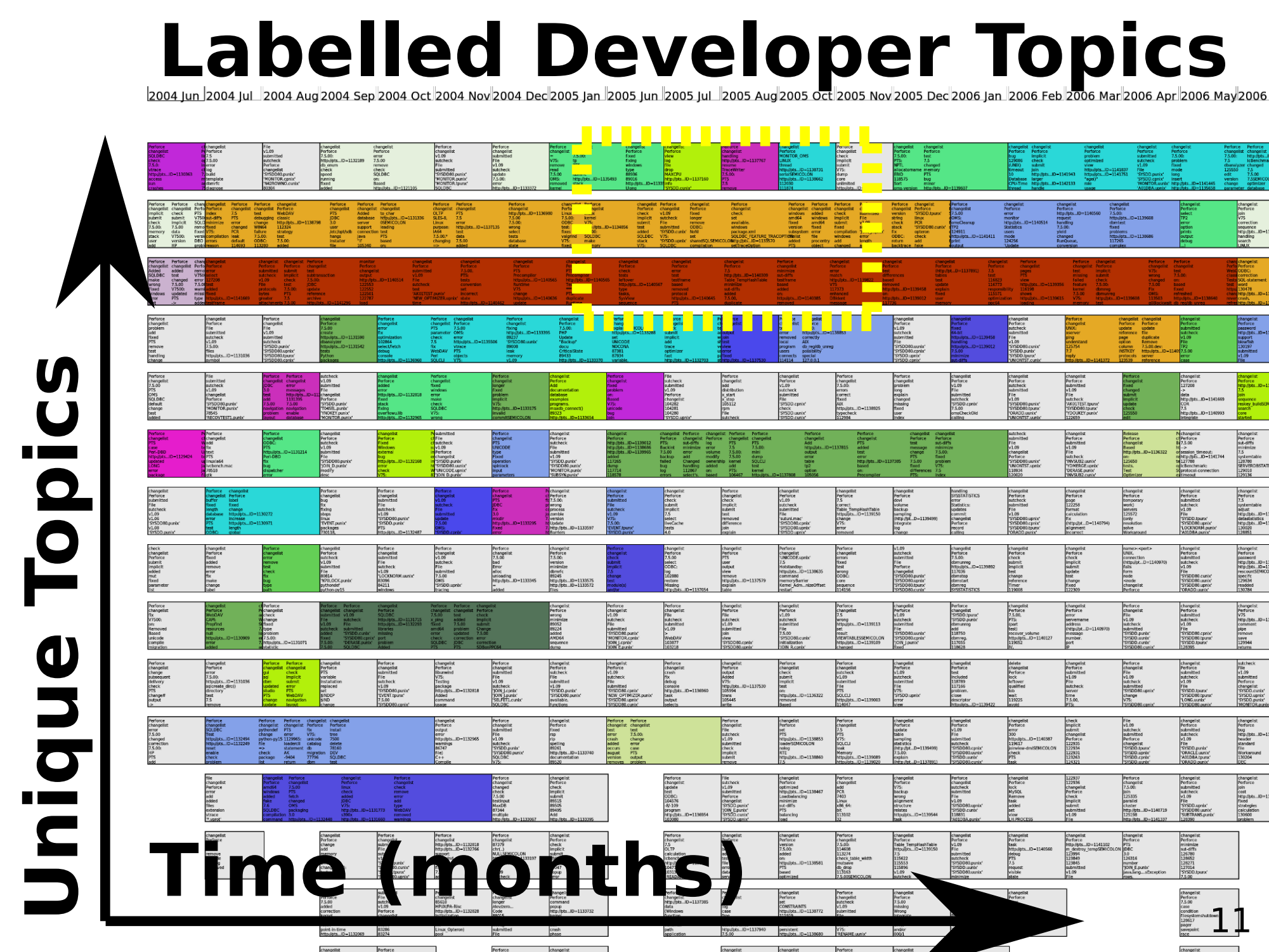
Can't we just summarize quality related efforts within this project?



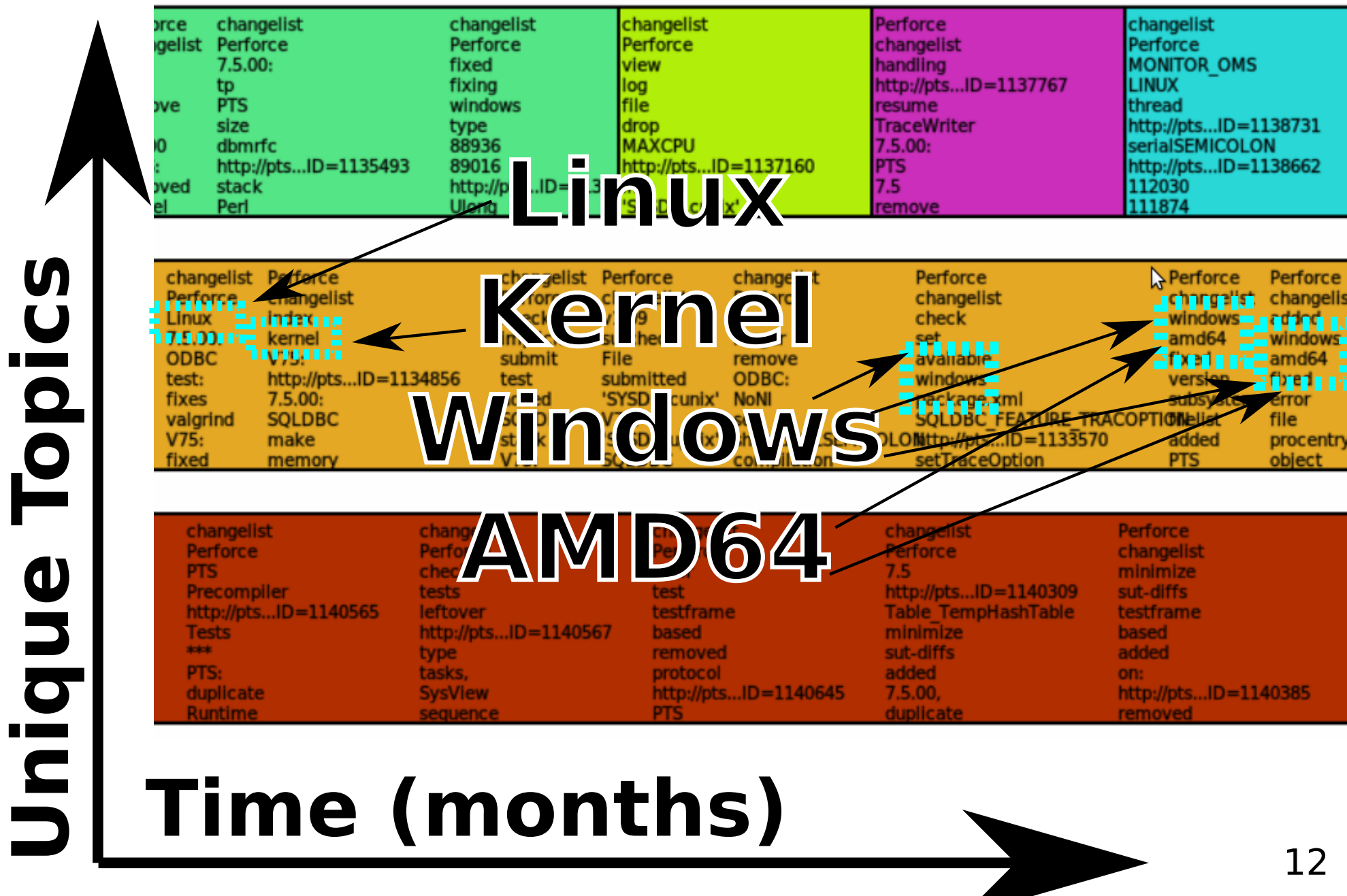


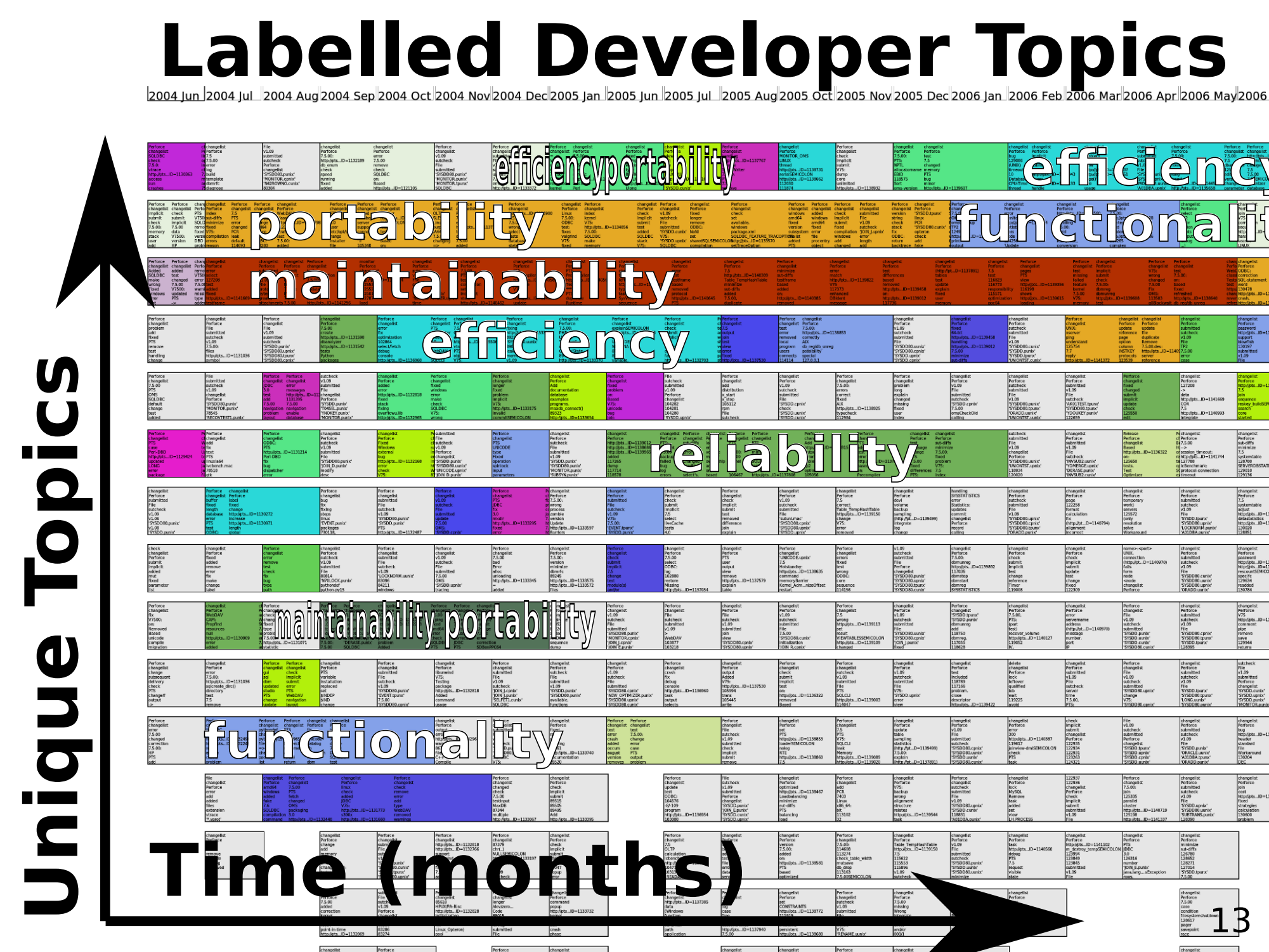
Maintainability
Functionality
Portability
Efficiency
Usability
Reliability



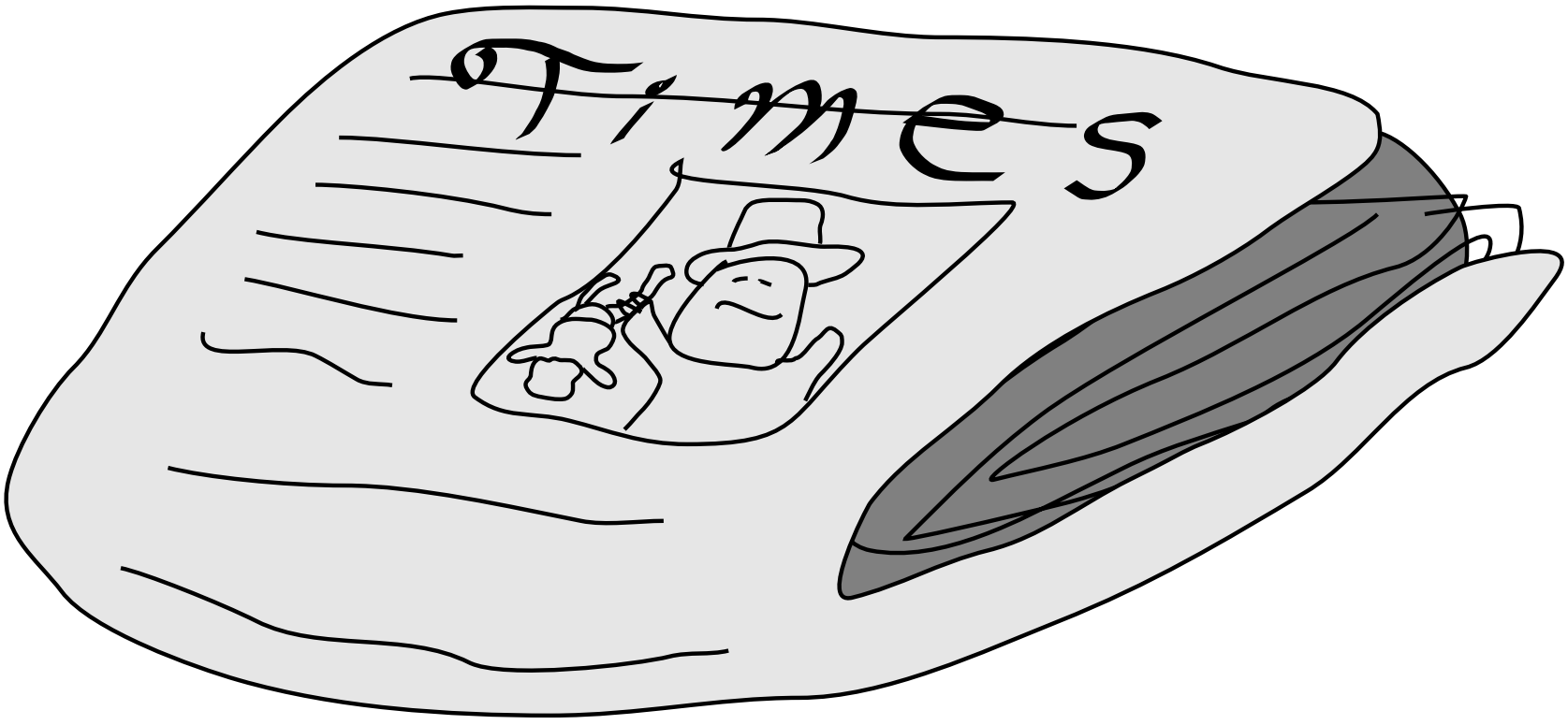

Labelled Developer Topics
2004 Jun
2004 Jul
2004 Aug
2004 Sep
2004 Oct
2004 Nov
2004 Dec
2005 Jan
2005 Jun
2005 Jul
2005 Aug
2005 Oct
2005 Nov
2005 Dec
2006 Jan
2006 Feb
2006 Mar
2006 Apr
2006 May
2006
Unique Topics
Time (months)
11

Labelled Developer Topics



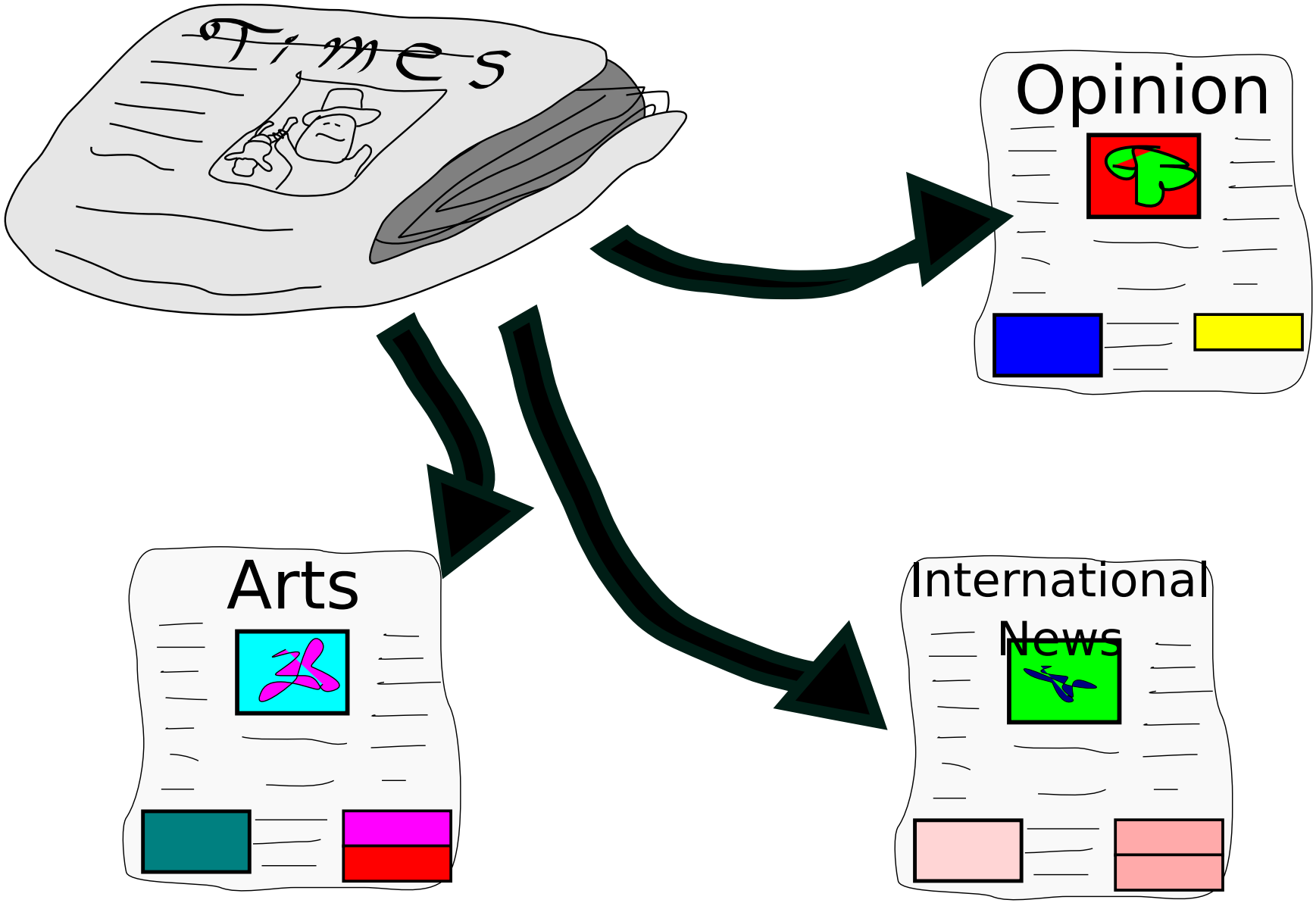

Labelled Developer Topics
2004 Jun
2004 Jul
2004 Aug
2004 Sep
2004 Oct
2004 Nov
2004 Dec
2005 Jan
2005 Jun
2005 Jul
2005 Aug
2005 Oct
2005 Nov
2005 Dec
2006 Jan
2006 Feb
2006 Mar
2006 Apr
2006 May
2006
efficiencyportability
efficiency
portability
functionalit
maintainability
efficiency
reliability
maintainability portability
functionality
Unique Topics
Time (months)

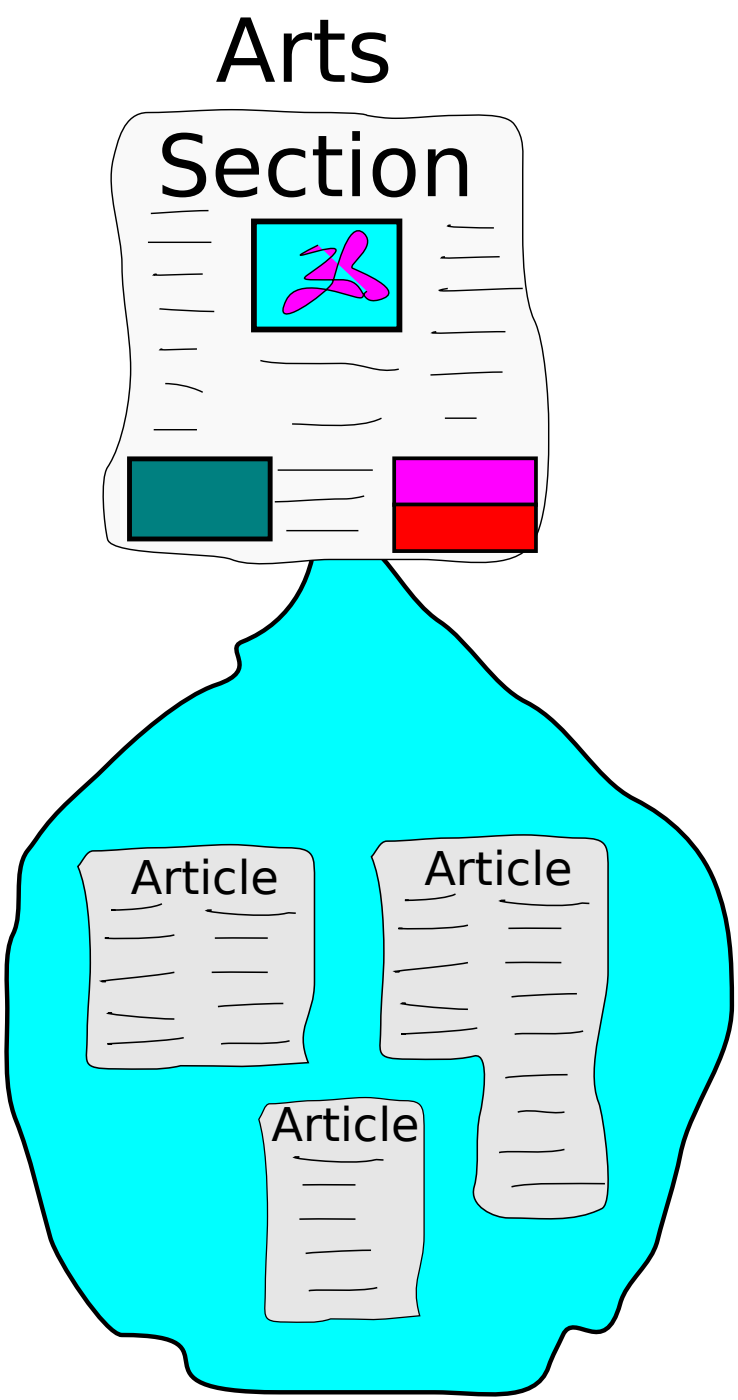
Example

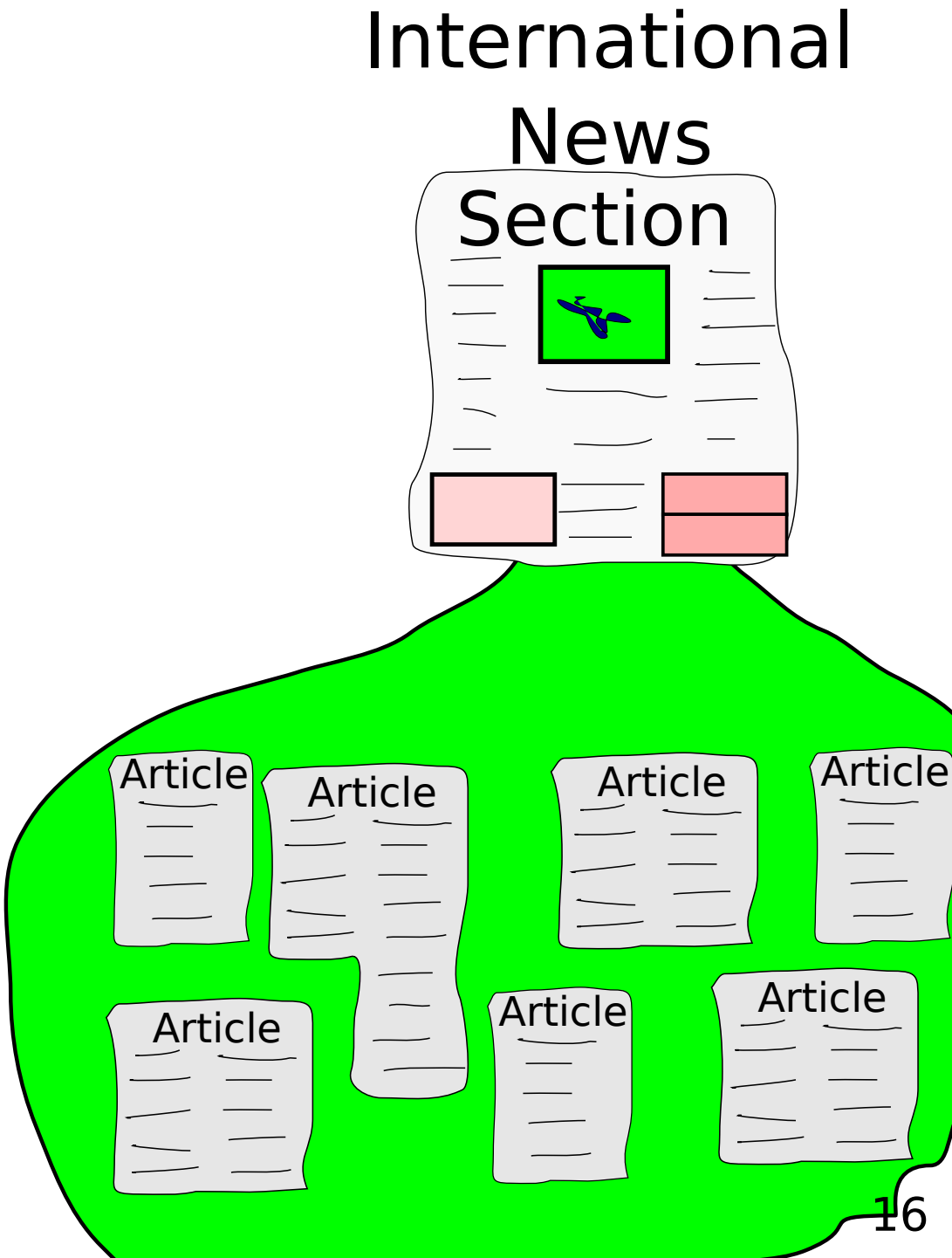


[Blei]

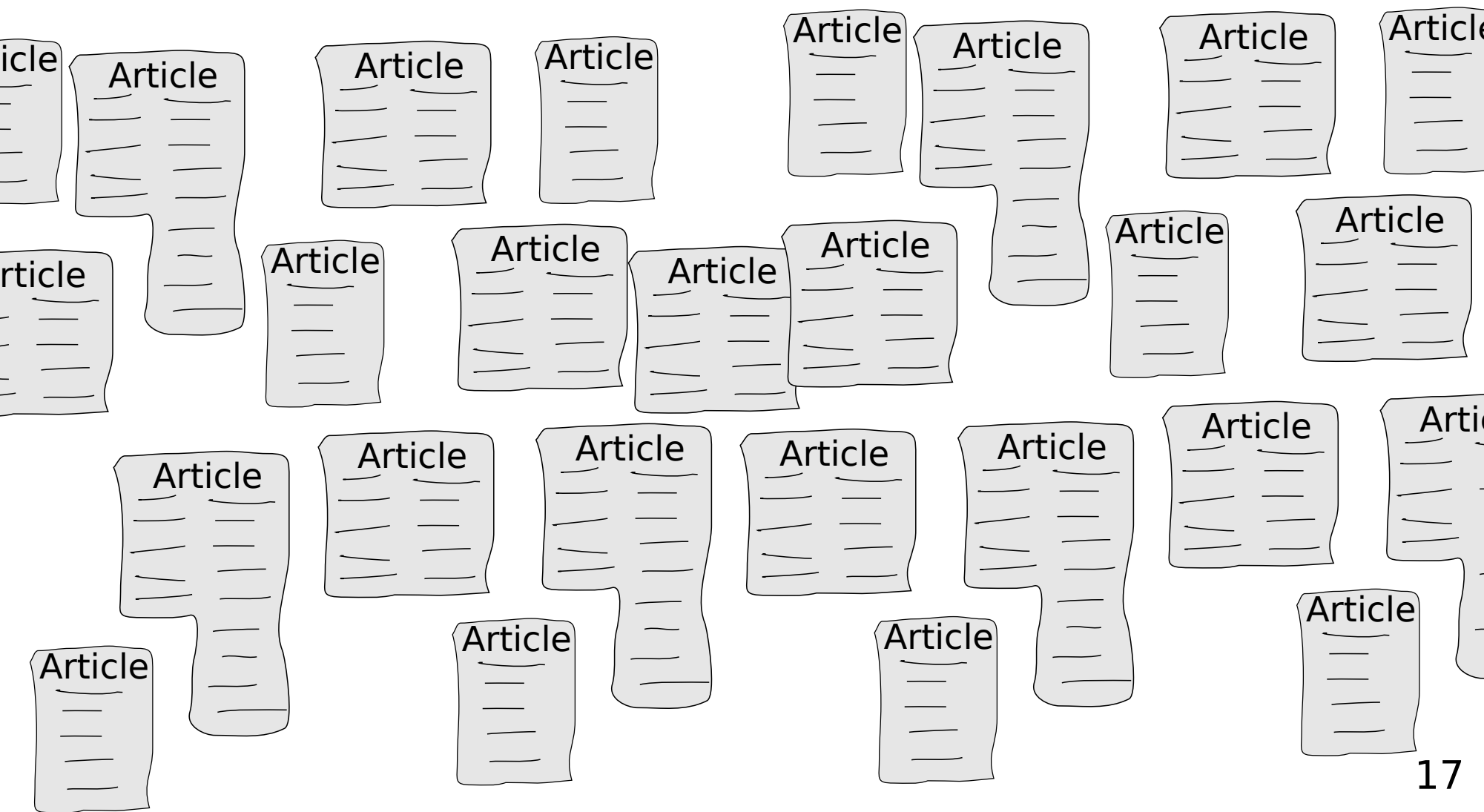
apologies to those with prior LDA/LSI experience

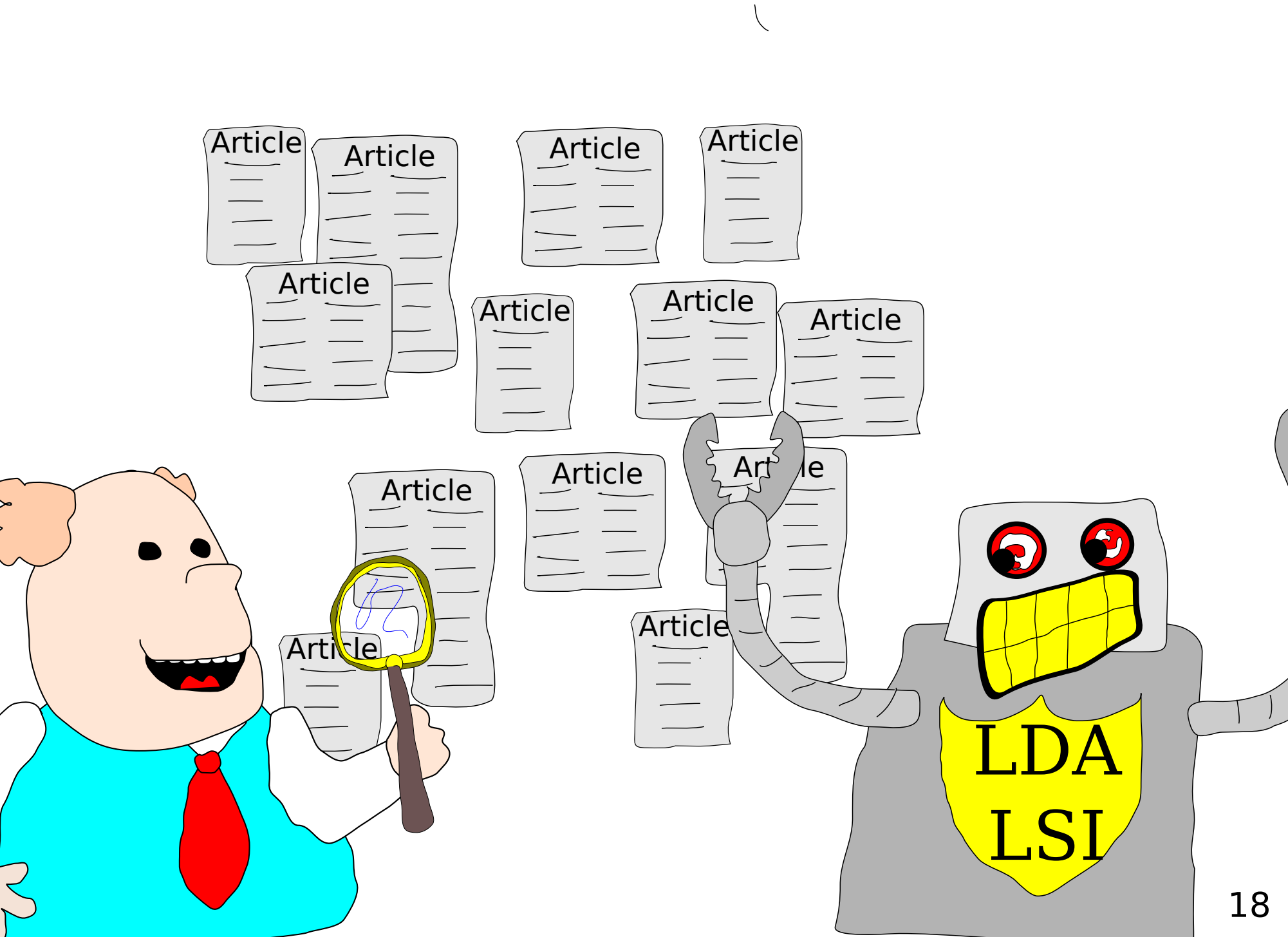

Times
Opinion
Arts
International
News

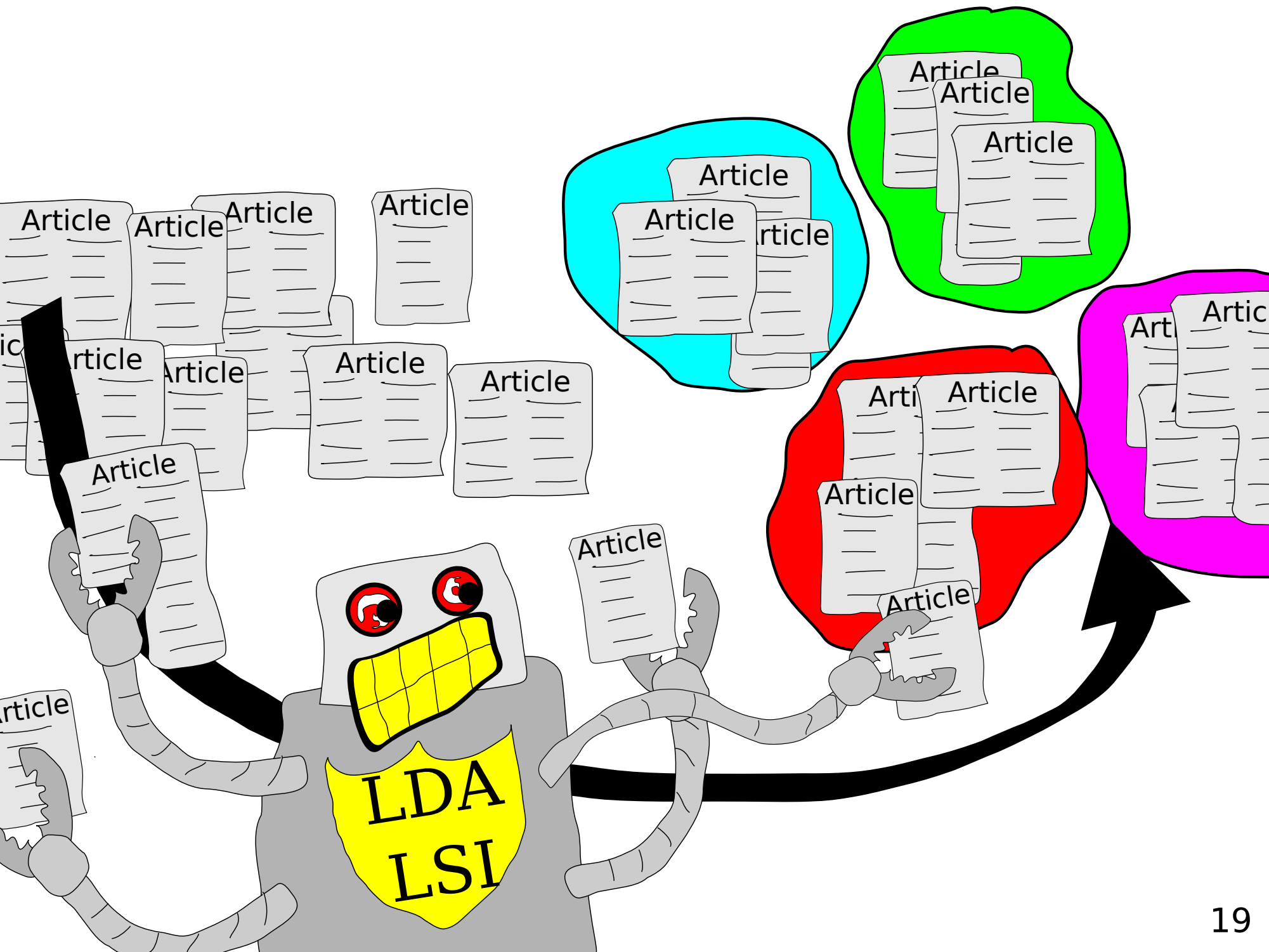

Arts
Section
Article
Article
Article

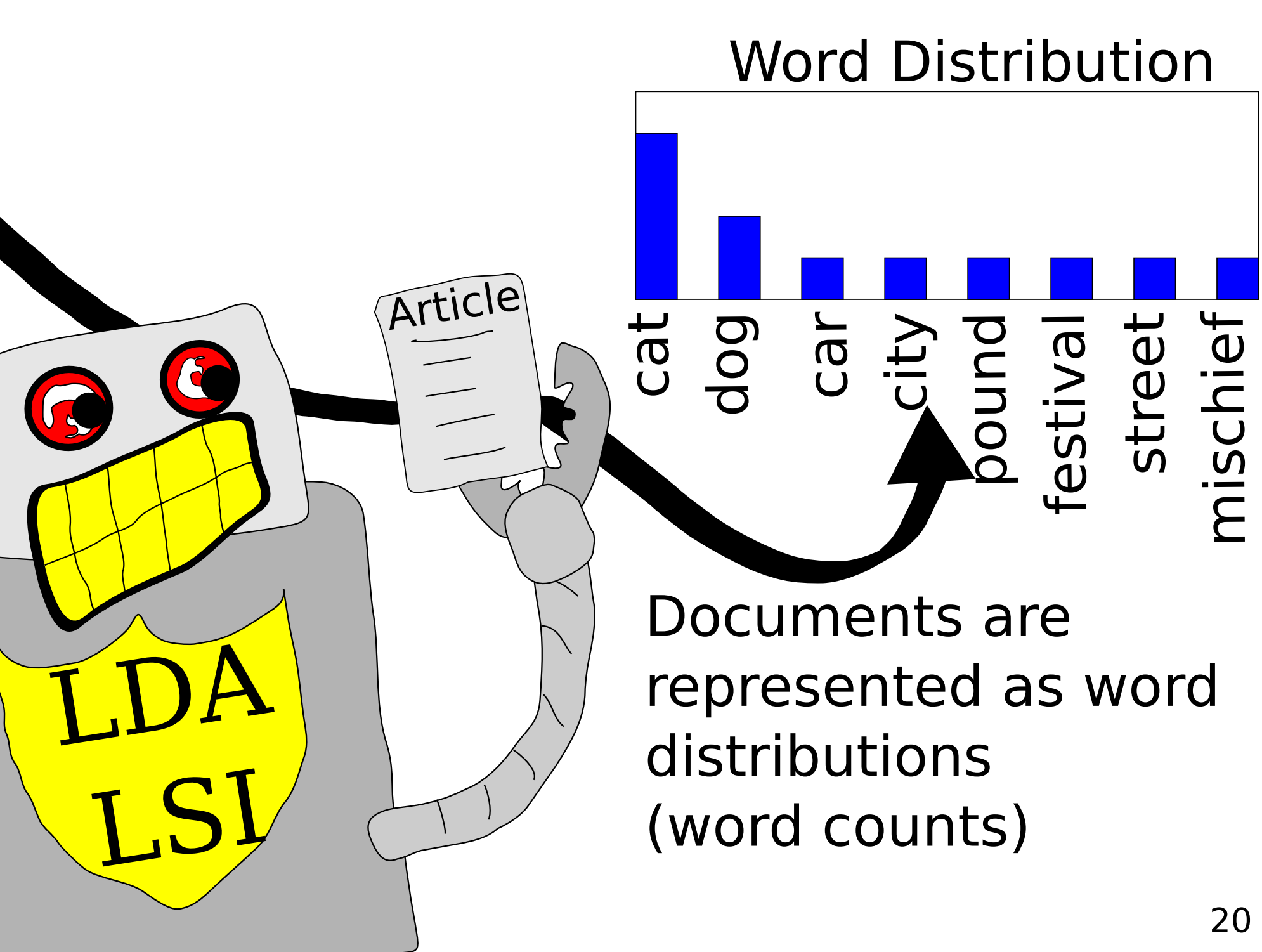

International
News
Section
Article
Article
Article
Article
Article
Article
Article

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



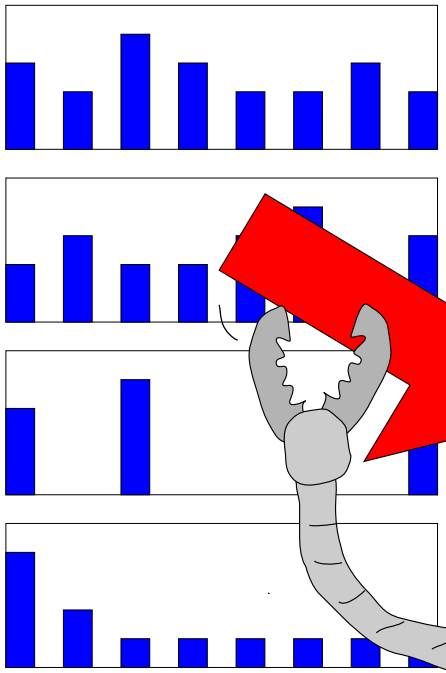

Article
Article
Article
Article
Article
Article
Article
Article
Article
Article
Article
Article
Article
LDA
LSI

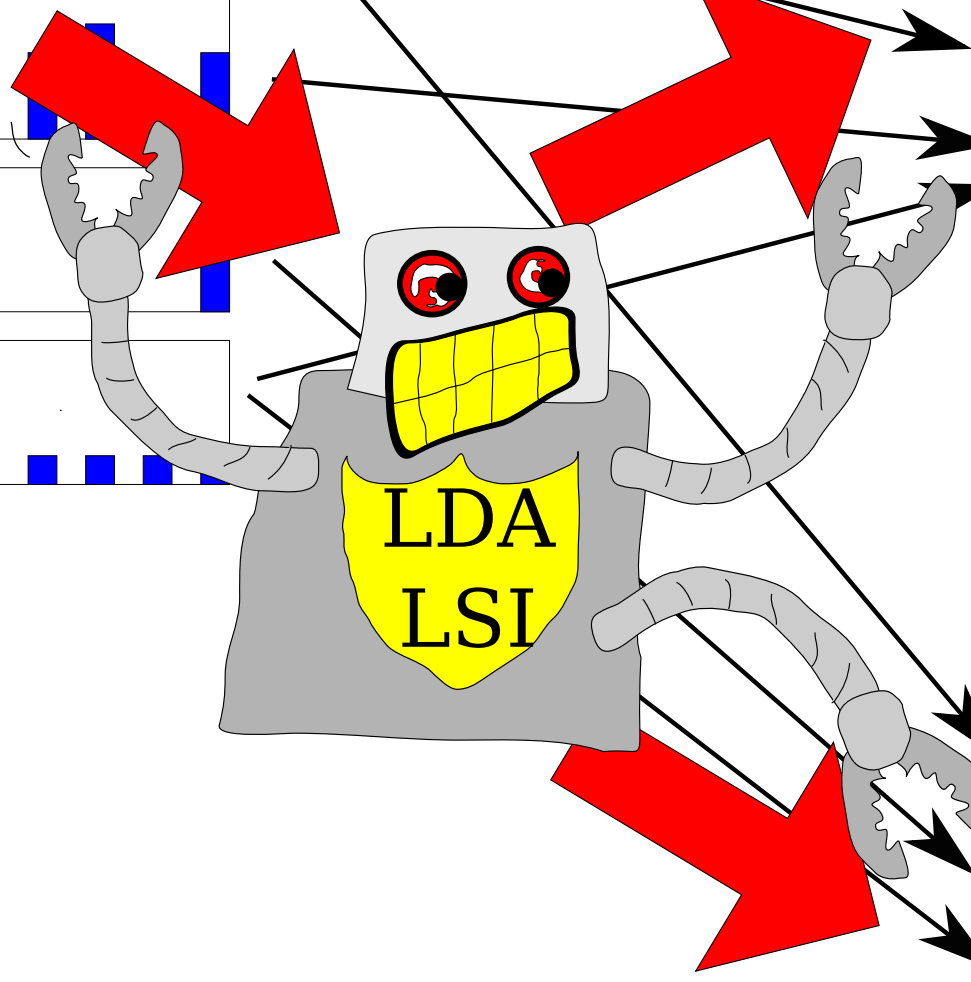

Article
LDA
LSI



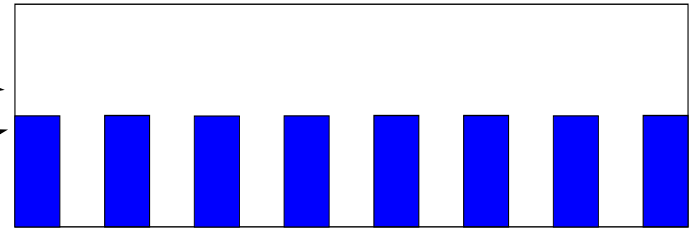
Documents are represented as word distributions (word counts)

Word Distributions



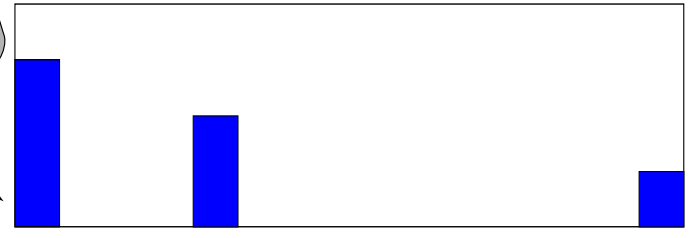


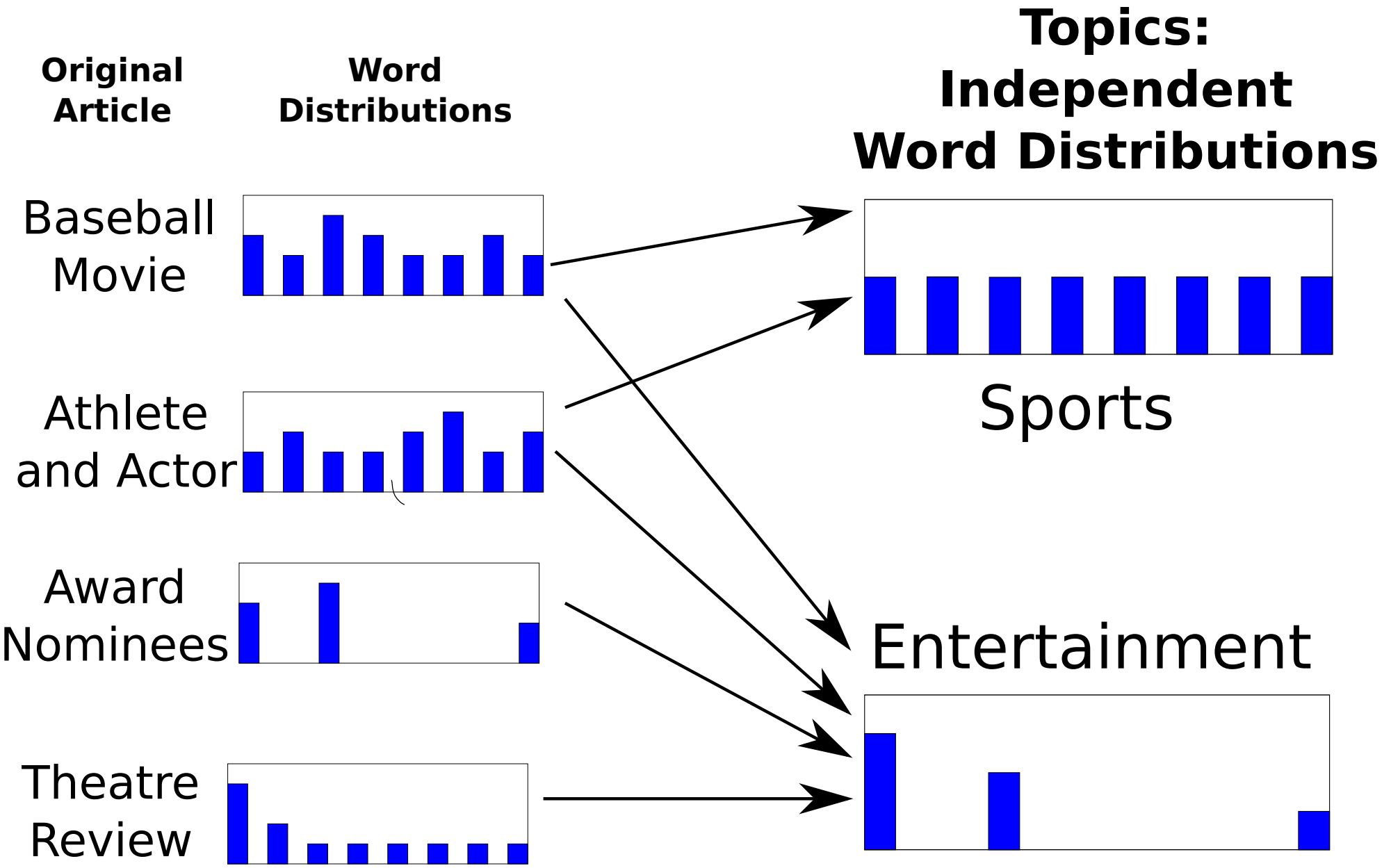
Topics: Independent Word Distributions



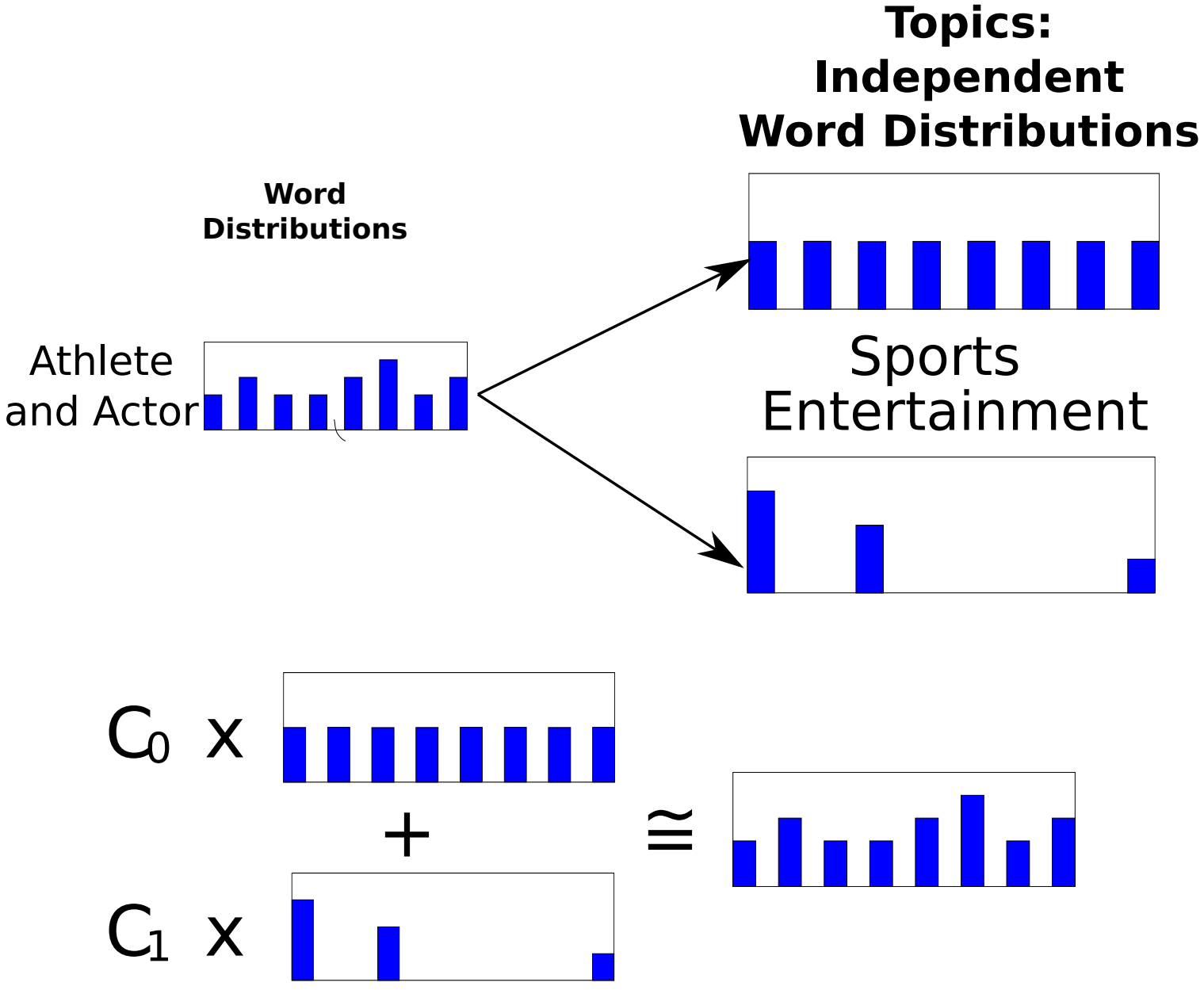
LDA finds independent word distributions that the documents are related to.

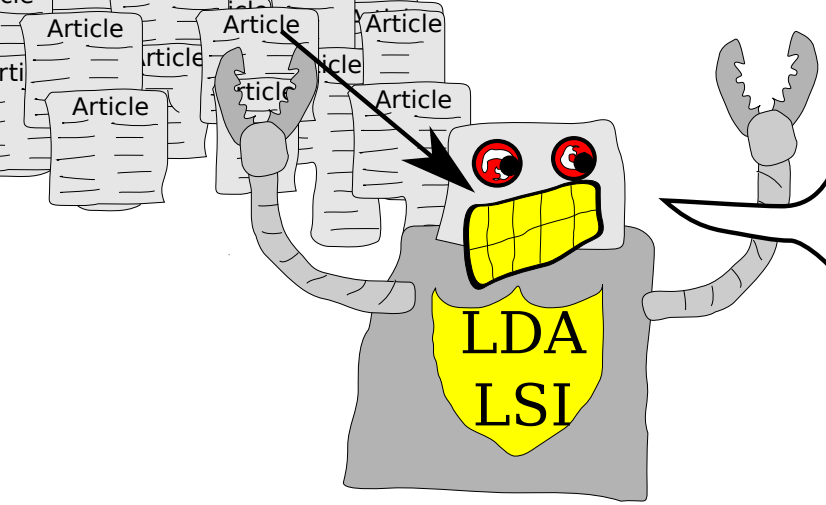
Documents can be associated with more than one topic.




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



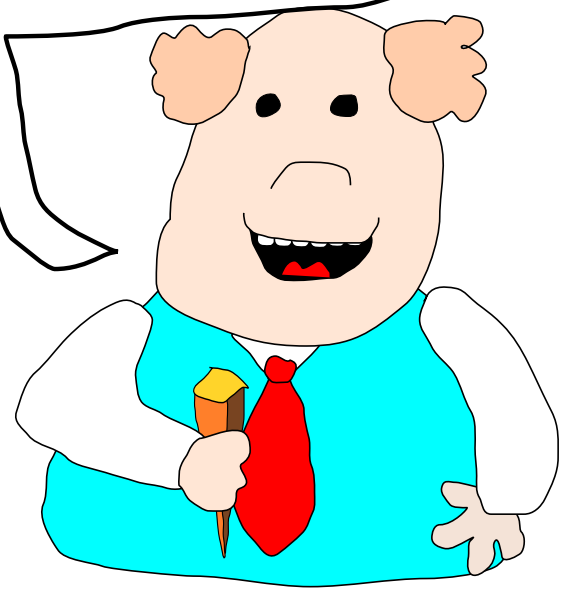


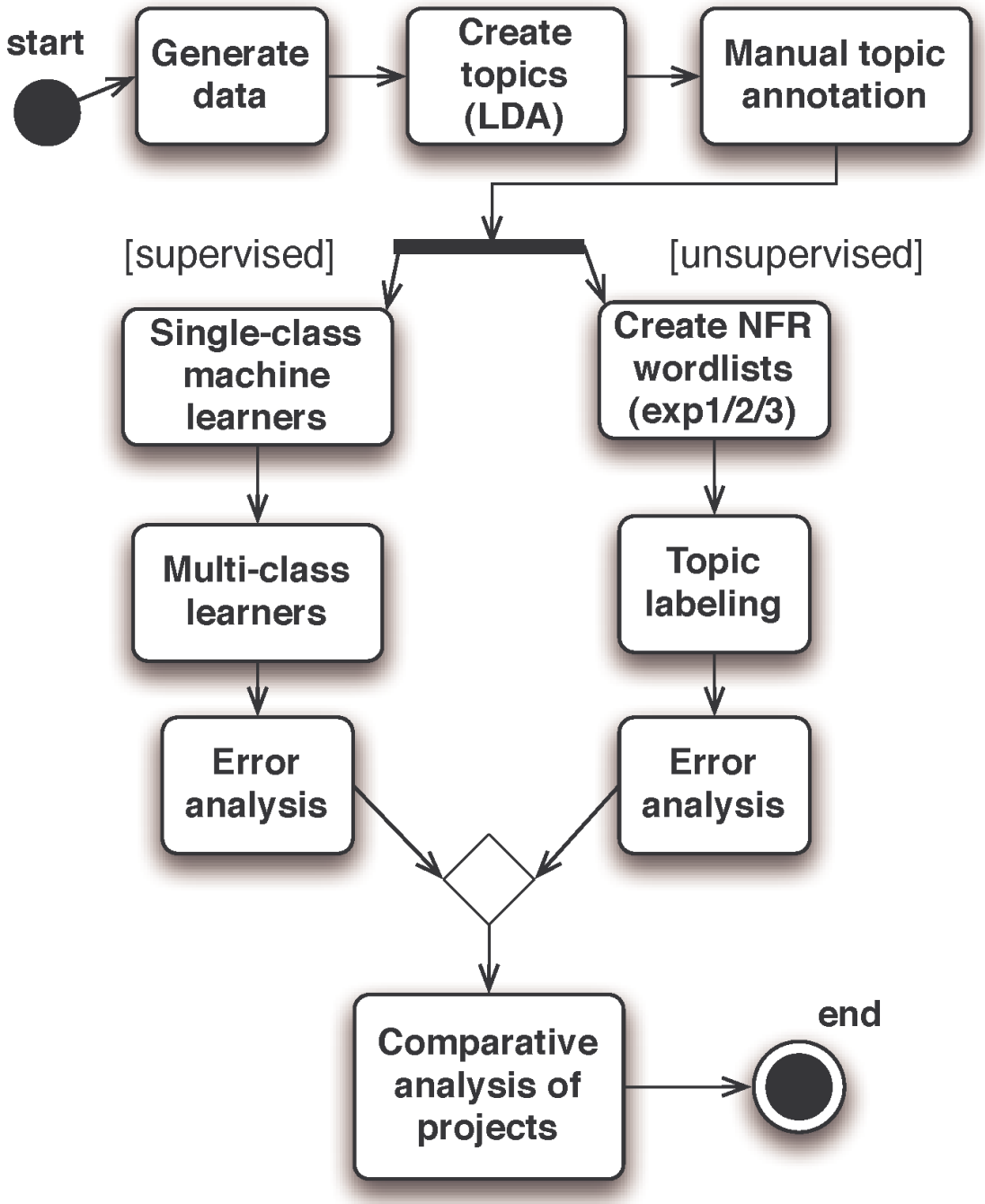
Here are two topics. I don't know what they are about!

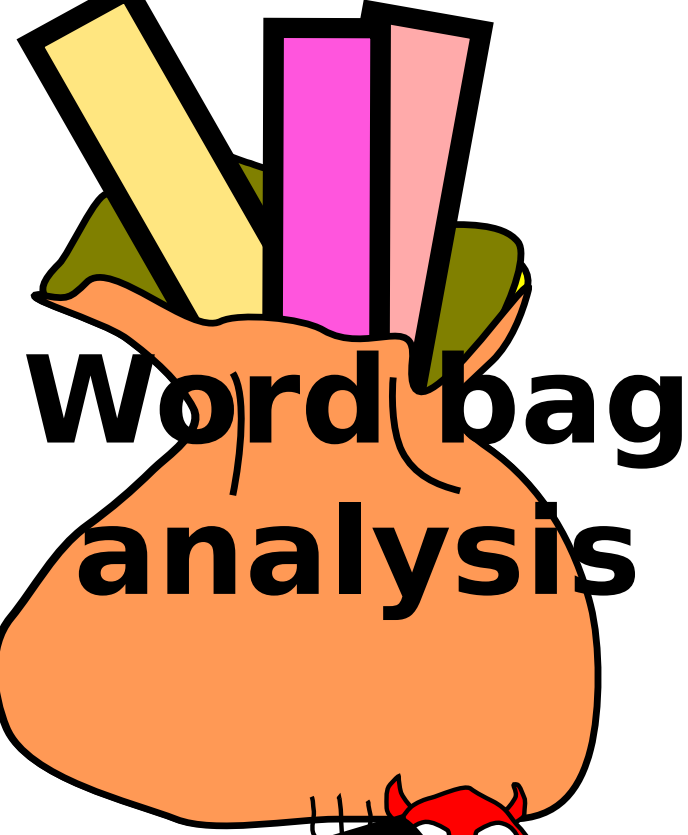
Topic 1
* play
* game
* inning
* player
* quarter
* opponent
* ...

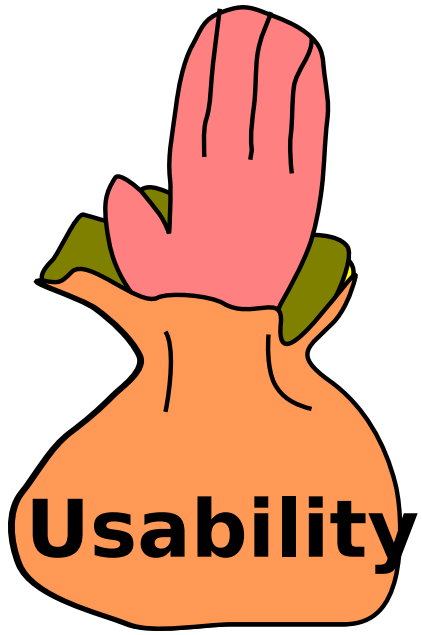
Topic 2
* gambling
* play
* night life
* comedy
* movie
* theatre
* ...

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




start
Generate data
Create topics (LDA)
Manual topic annotation
[supervised]
[unsupervised]
Single-class machine learners
Create NFR wordlists (exp1/2/3)
Multi-class learners
Topic labeling
Error analysis
Error analysis
Comparative analysis of projects
end


Word bag
analysis


Usability


Maintainability


Portability

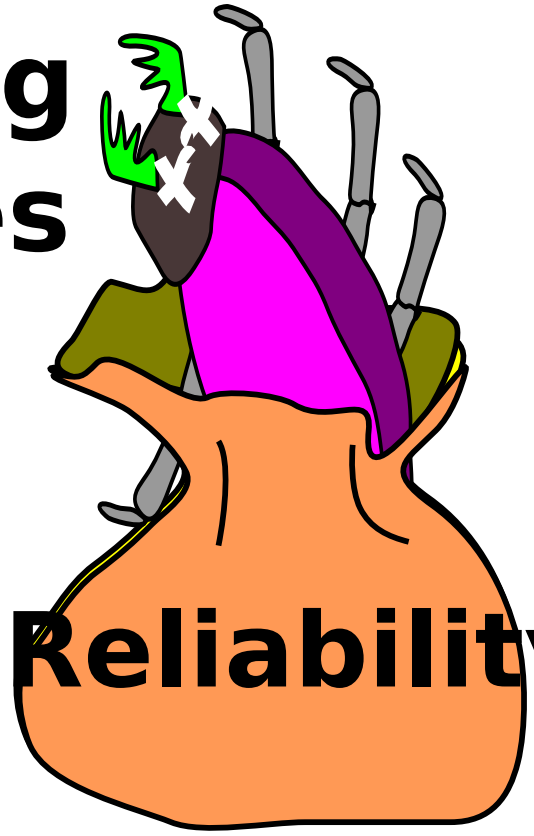

Reliability


Efficiency

Word Bag Examples

Portability

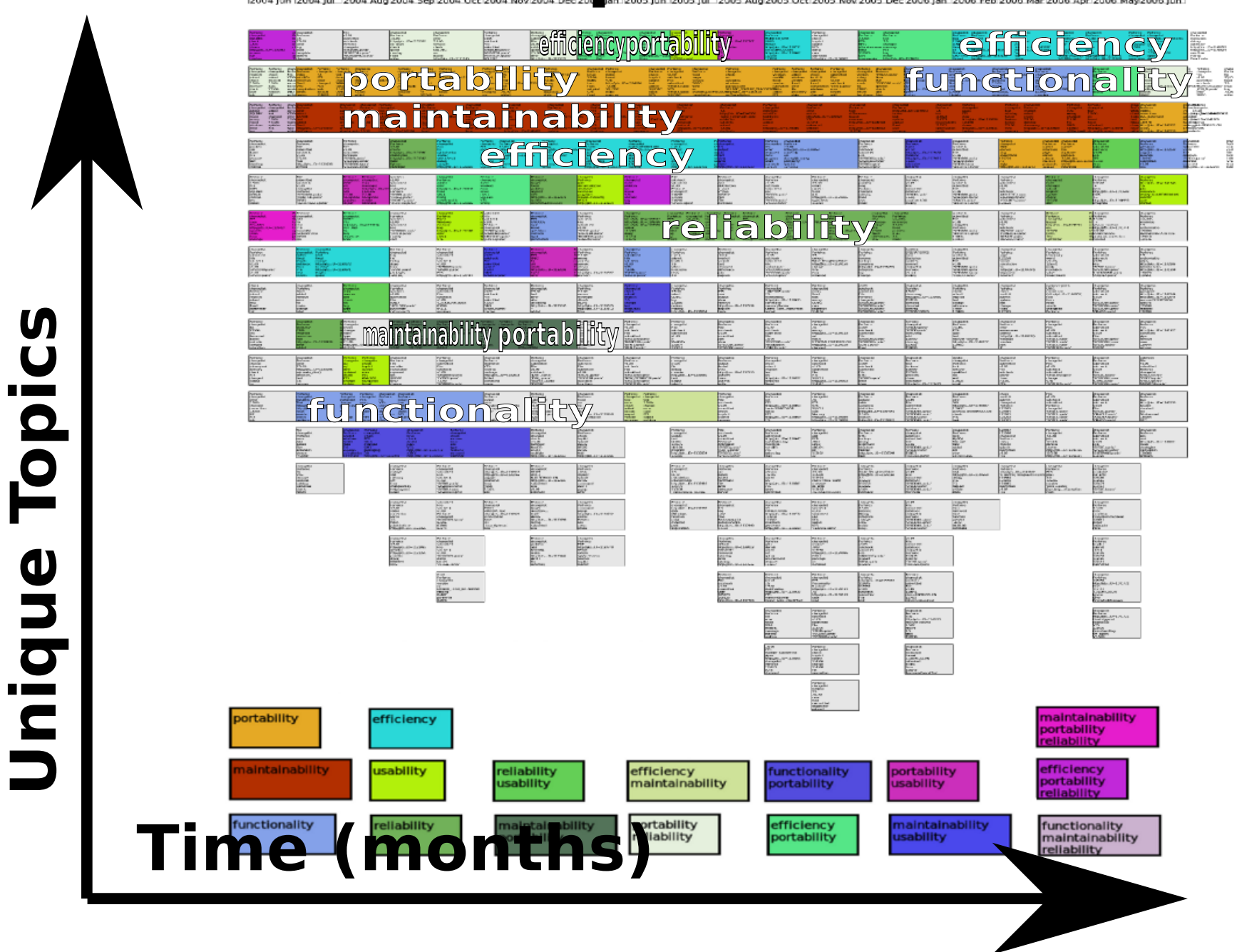
portability
transferability
interoperability
documentation
internationalization
i18n
...



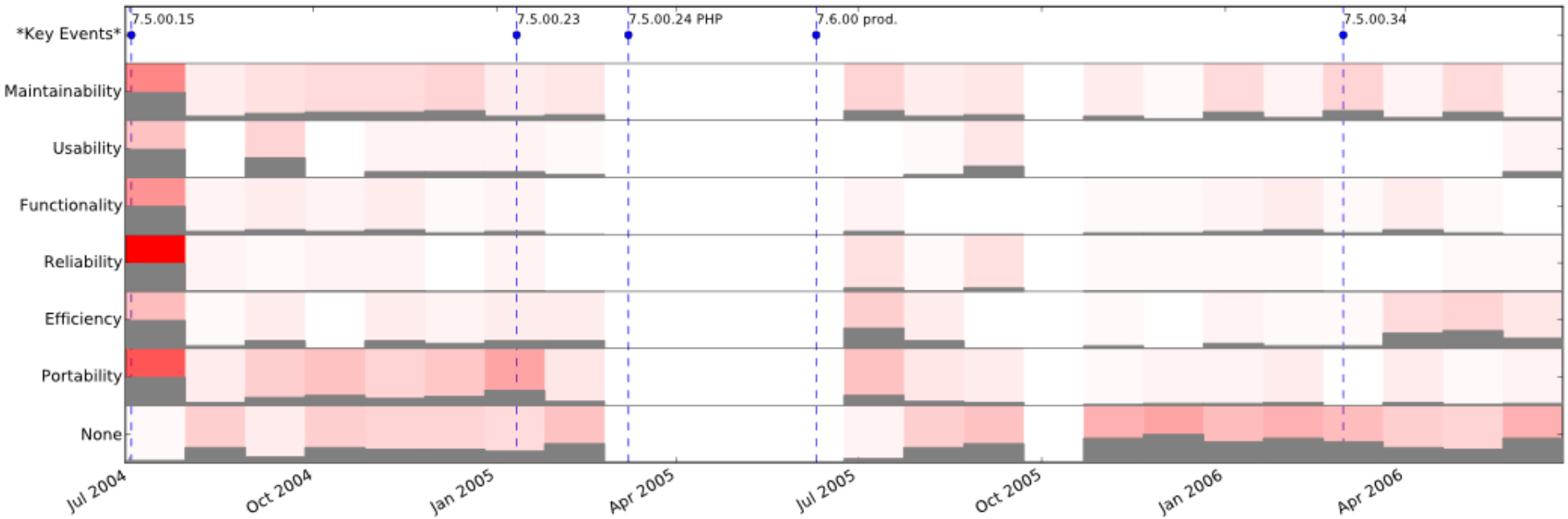
Reliability

reliability
failure
error
redundancy
fails
bug
...

Labelled Topics of MaxDB 7.500



MaxDB 7.500 Timeline

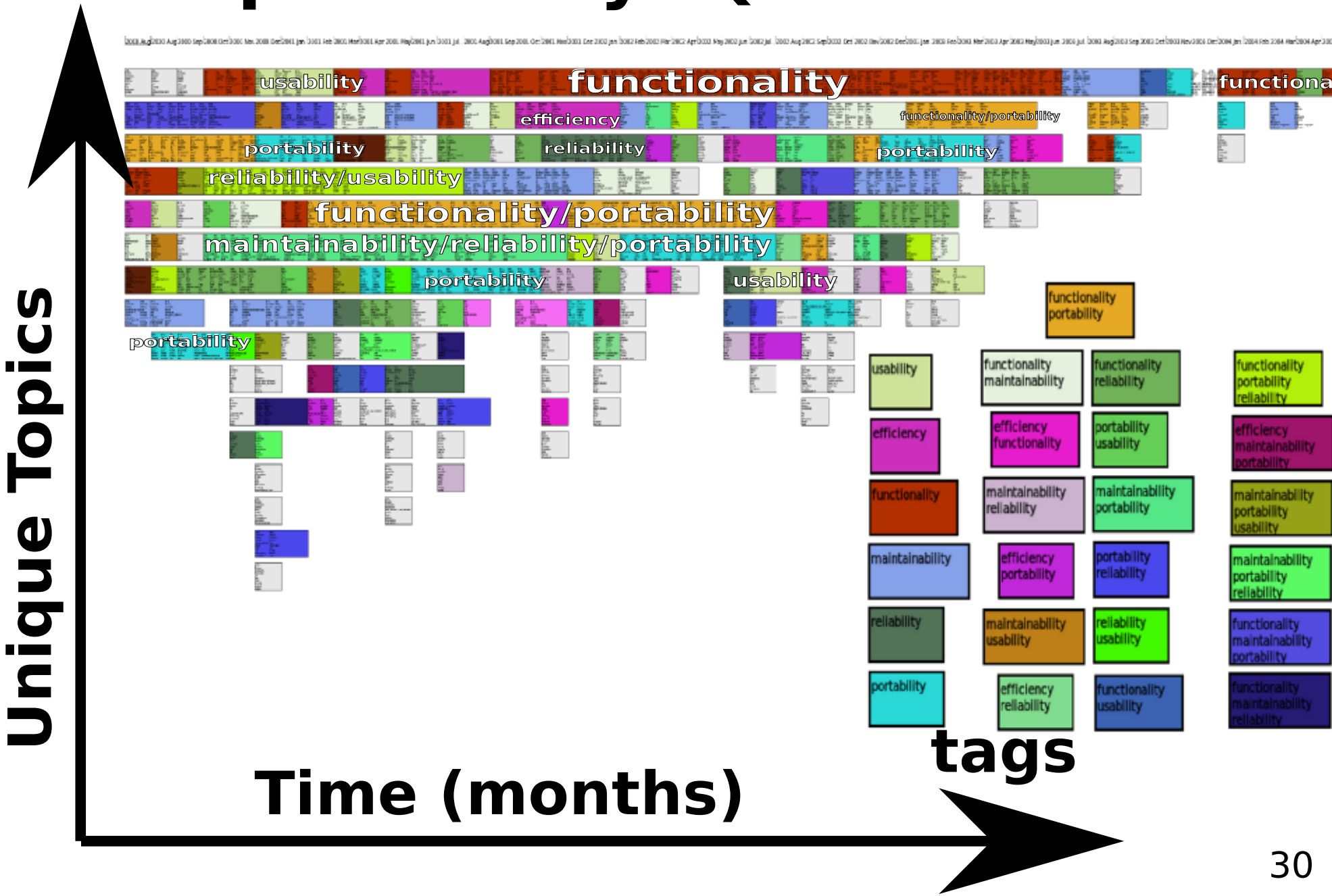


Maintainability
Portability

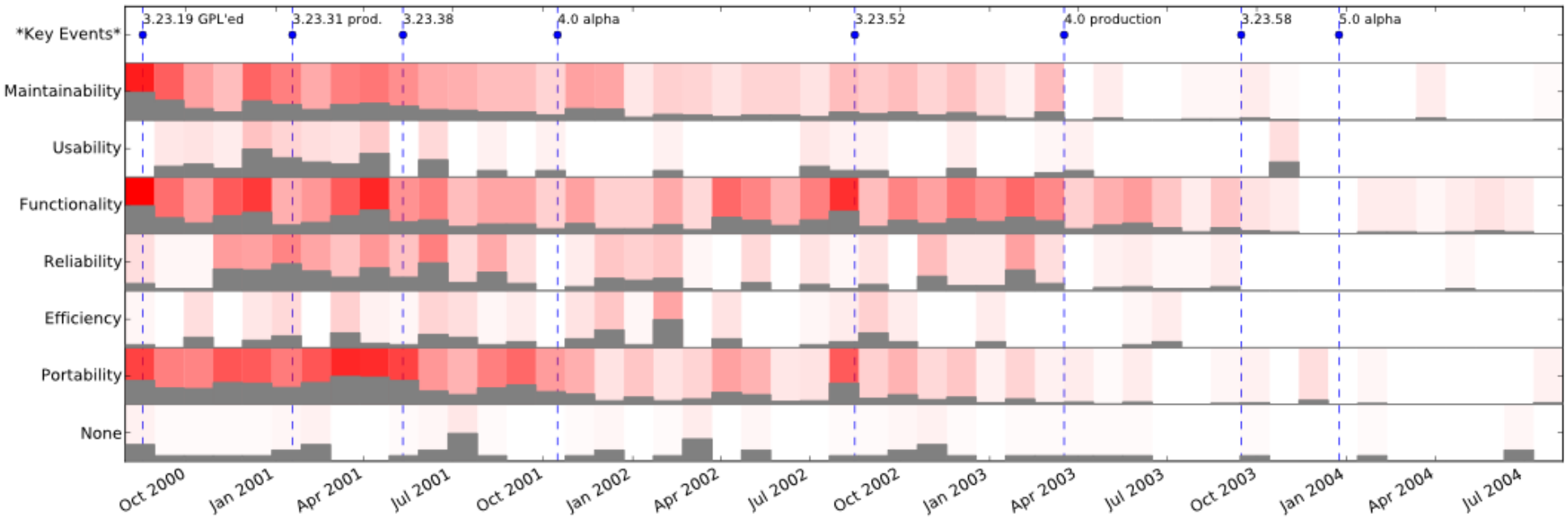
Maintainability
Portability
Reliability
Effeciency

Maintainability
Effeciency

Topics of MySQL 3.23



MySQL 3.23 Timeline

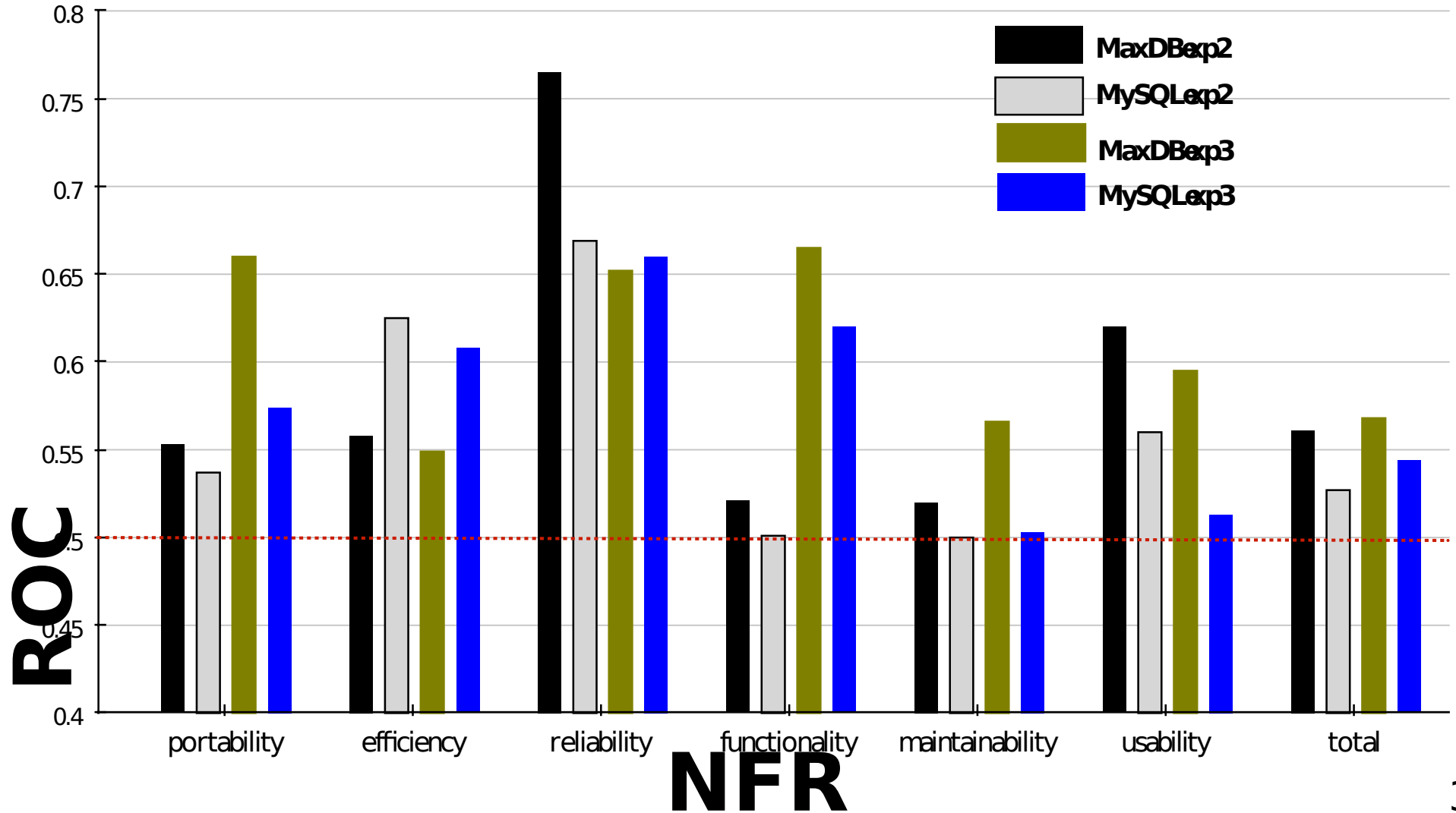


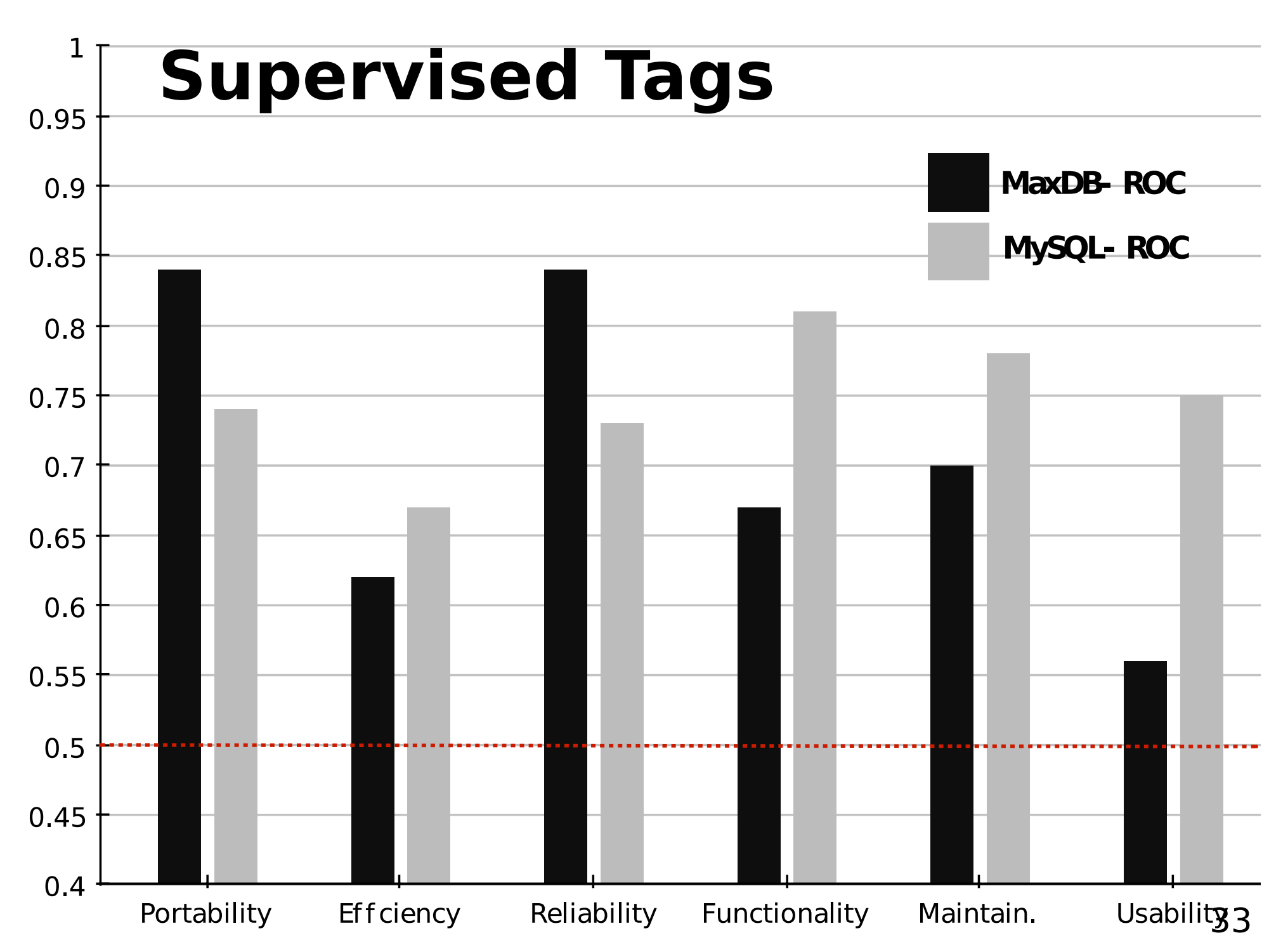
Maintainability
Functionality
Reliability
Portability

Maintainability
Functionality
Efficiency
Portability

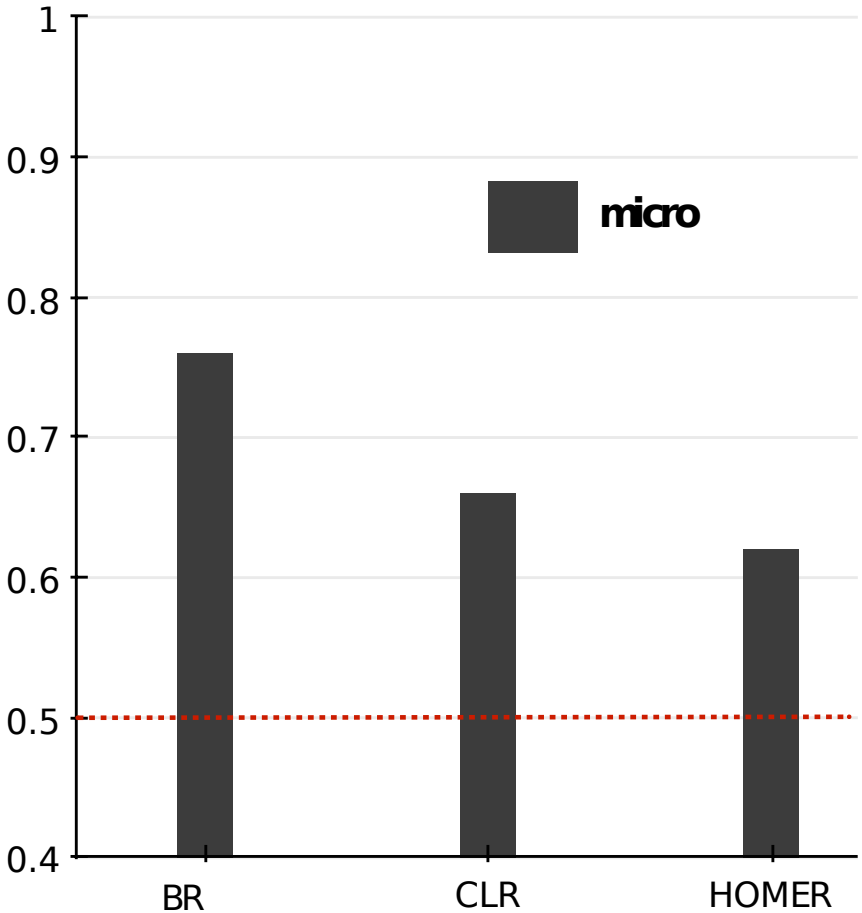
Functionality

ROC Values of Semi-Supervised Word Bags

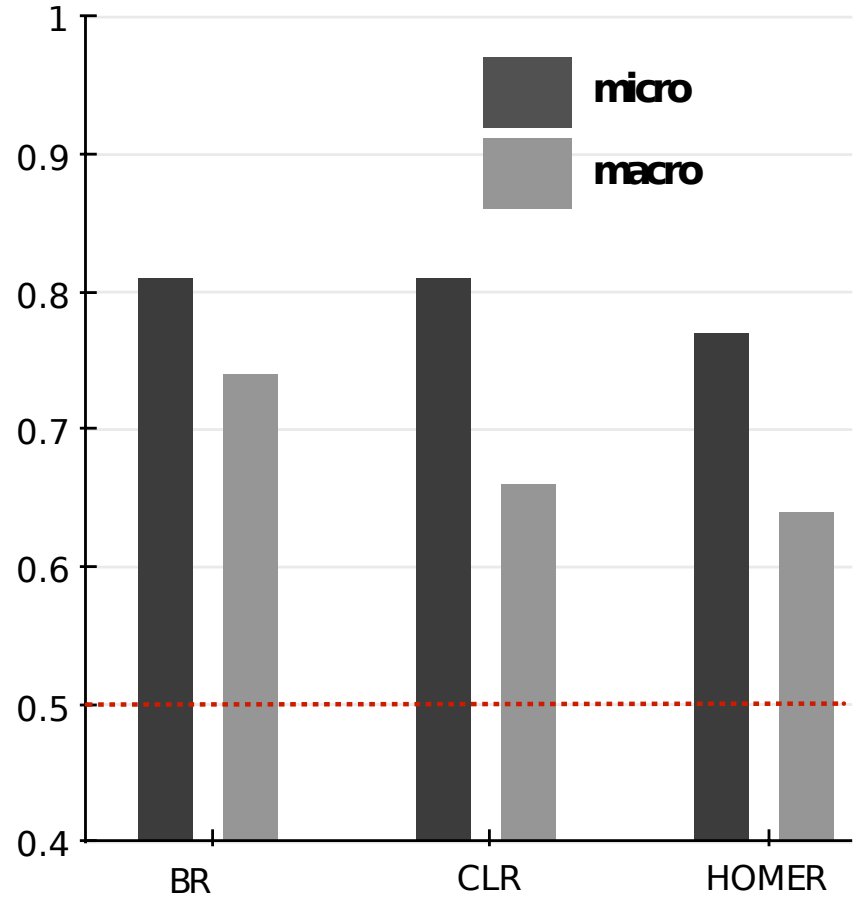



Supervised Tags
MaxDB- ROC
MySQL- ROC
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0.4
Portability
Effciency
Reliability
Functionality
Maintain.
Usability

Supervised Multitag Classifiers: MySQL and MaxDB



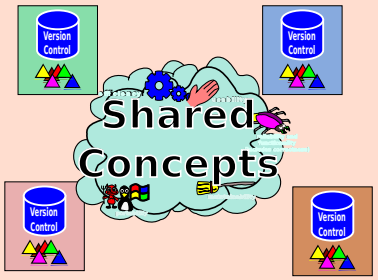
MySQL Classifiers

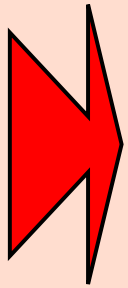


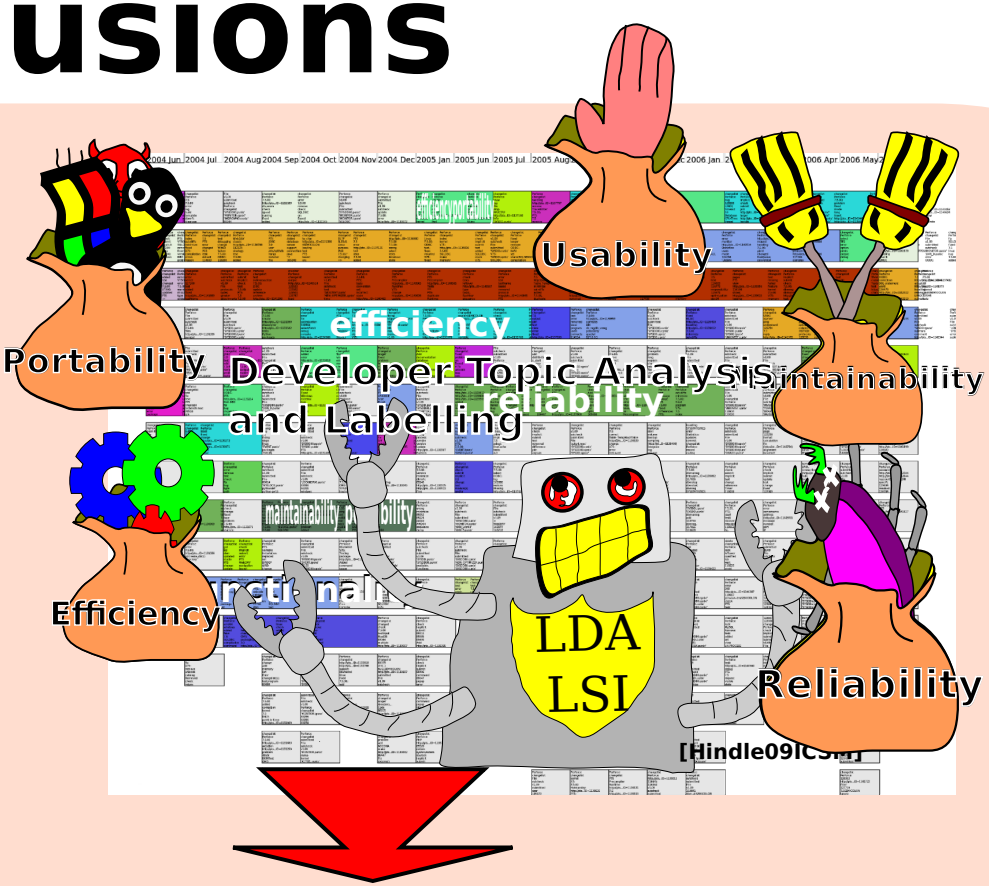
MaxDB Classifiers

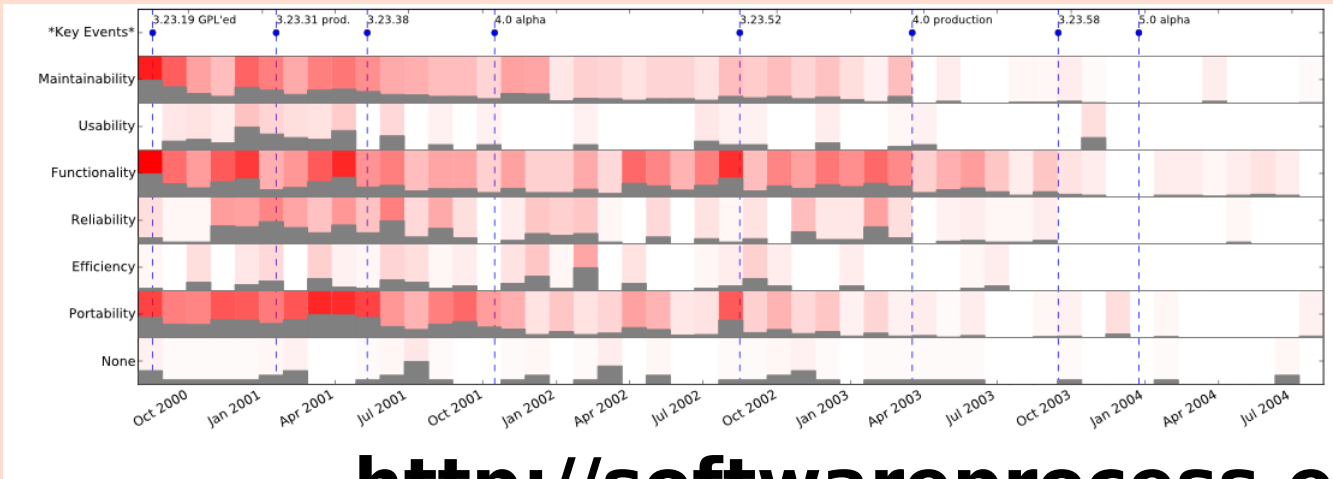
Conclusions

Source Control



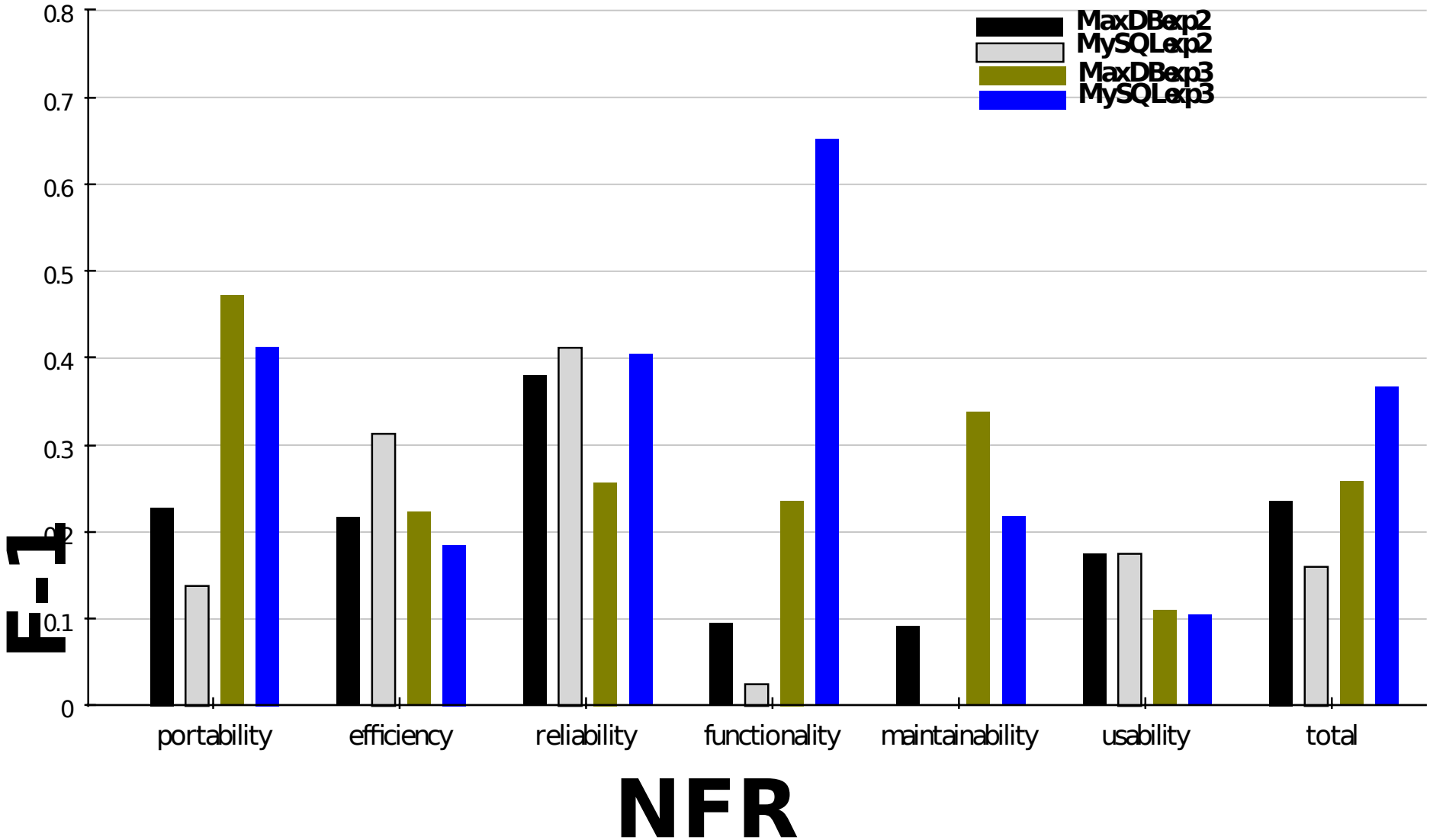






http://softwareprocess.es/name/

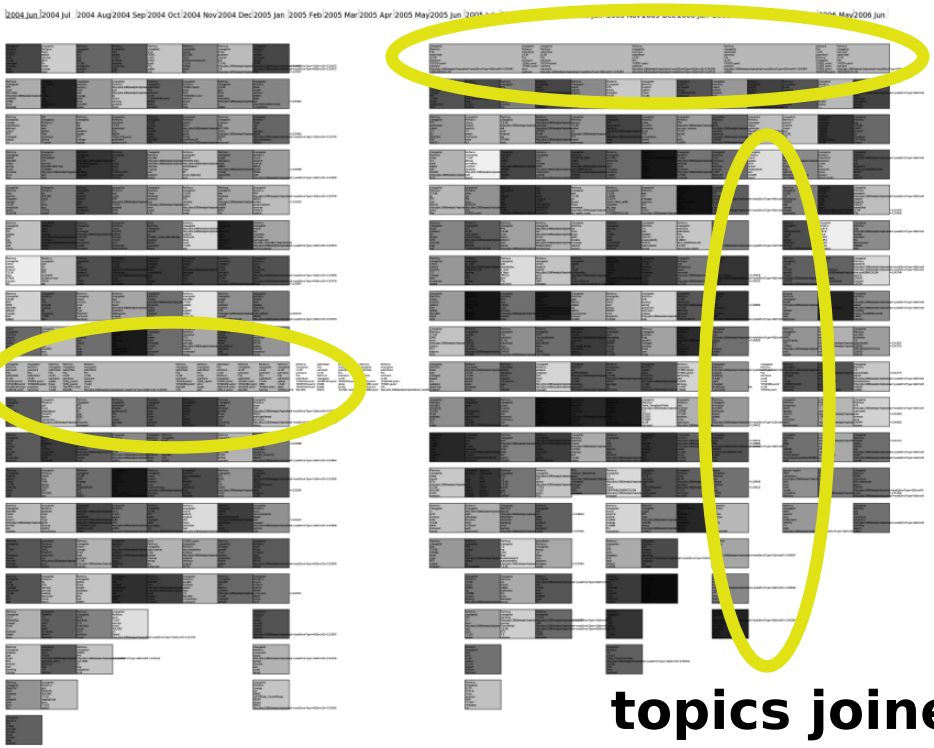
F-1 Measure of Semi-Supervised Word Bags



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

Annotation: Stop Words

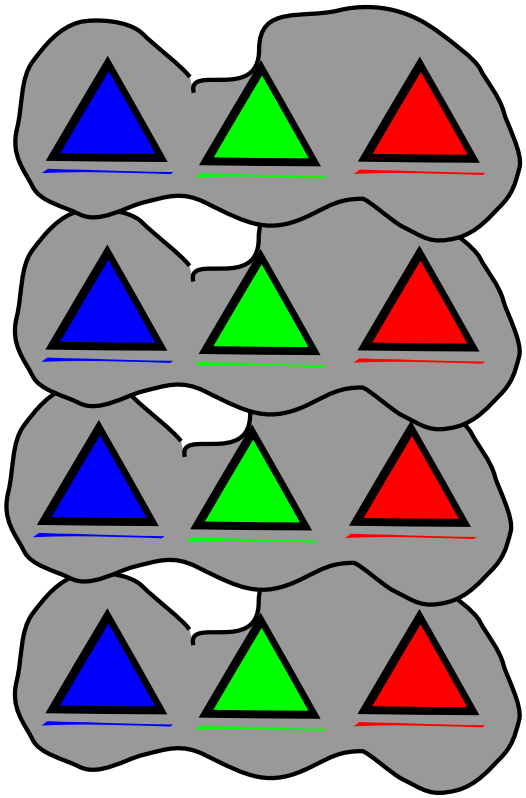
MaxDB 7.500 Case Study



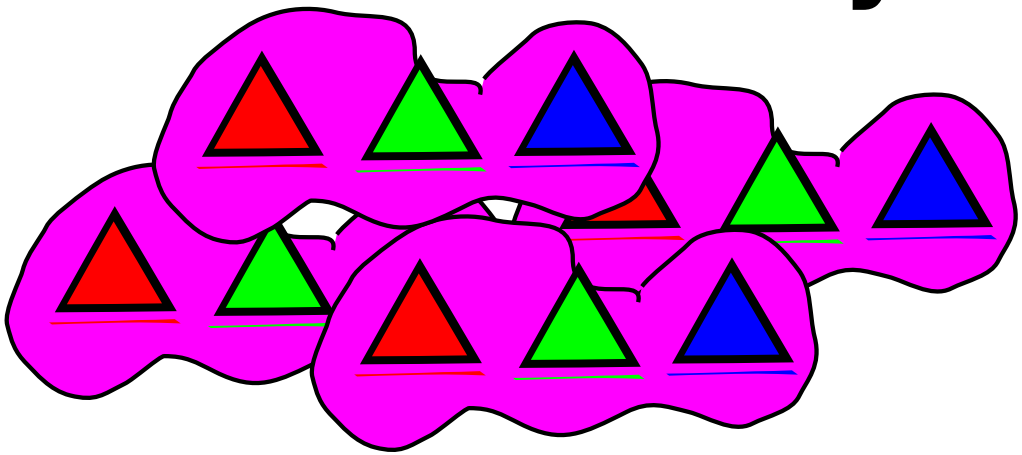
topics joined
due to similarity

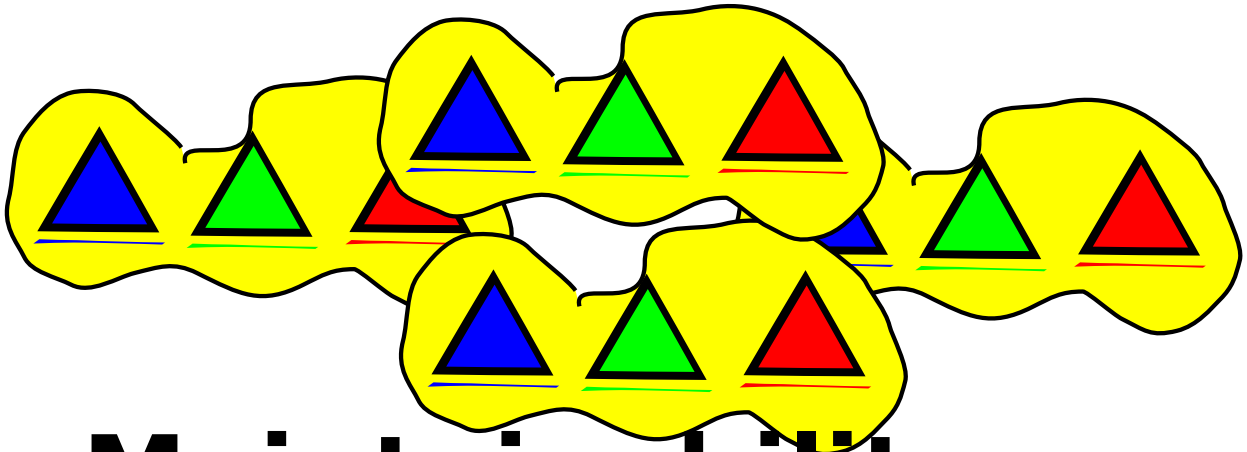
2 long trends
instead of one

Annotation: Training Sets



Maintainability+





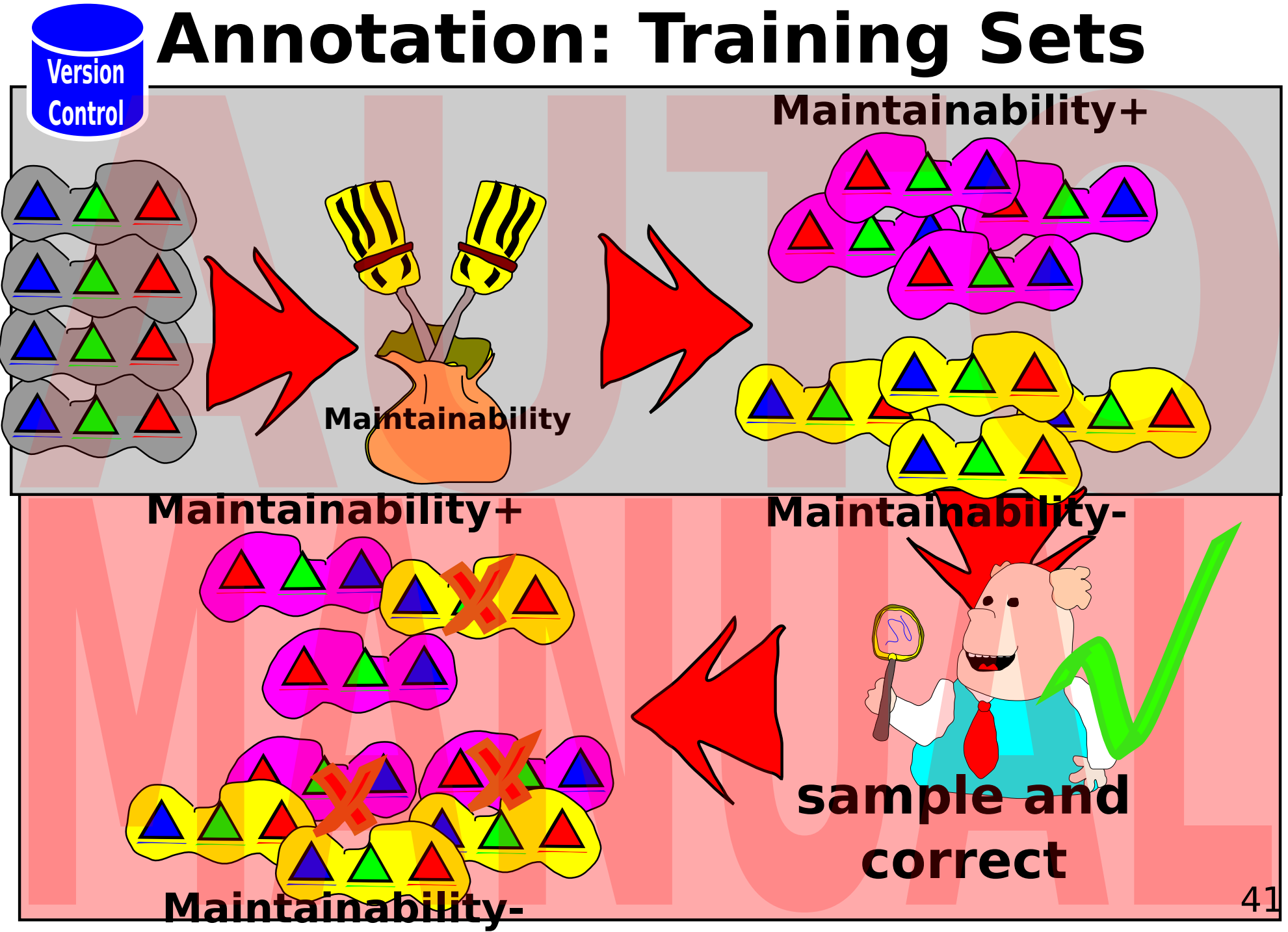
Maintainability-

Annotation: Stop Words

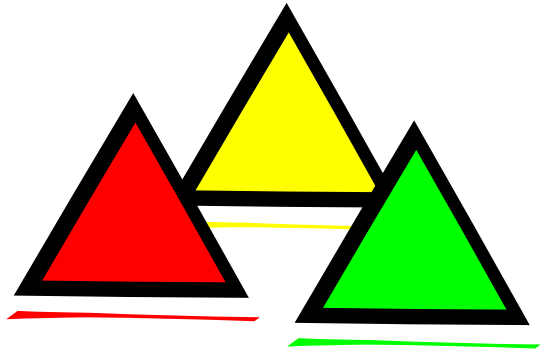
Used in topic analysis or to reduce # of features for learners.

perh
clearly
between
done ther
who becaus
haven't move
in asking exam
nevertheless doe
sensible our some
elsewhere upon ask
beforehand ie found
anywhere it contain
everywhere detai
need associ
specifyin
con
fo

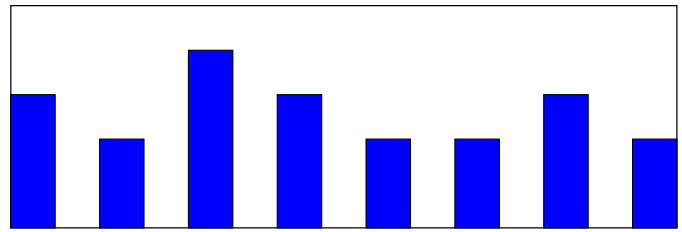
Annotation: Training Sets



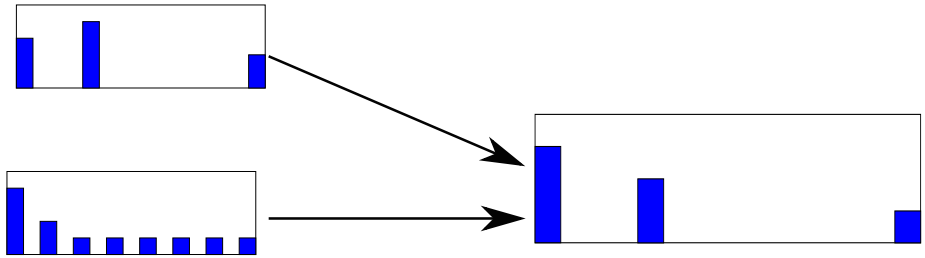
Message



Word Distribution



Topic



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

Trend