Software Process Recovery using Recovered Unified Process Views

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

What are we going to do?

Theory

Business Modeling
Requirements
Analysis & Design
Implementation
Test
Deployment
CM and SCS
Project Mangement
Environment

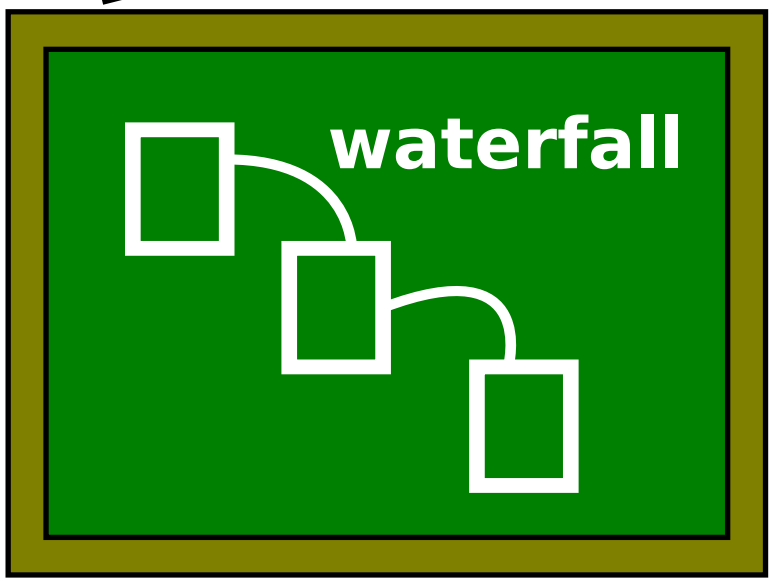
Practice

Process

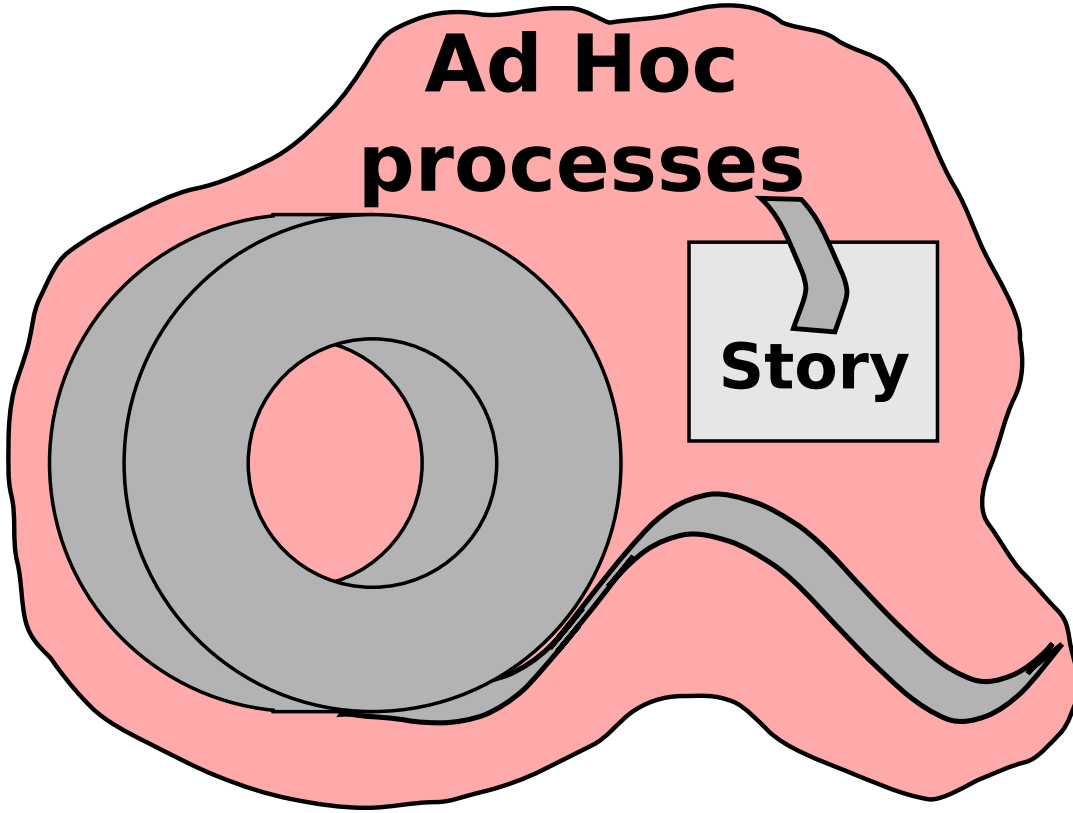
prescribed processes

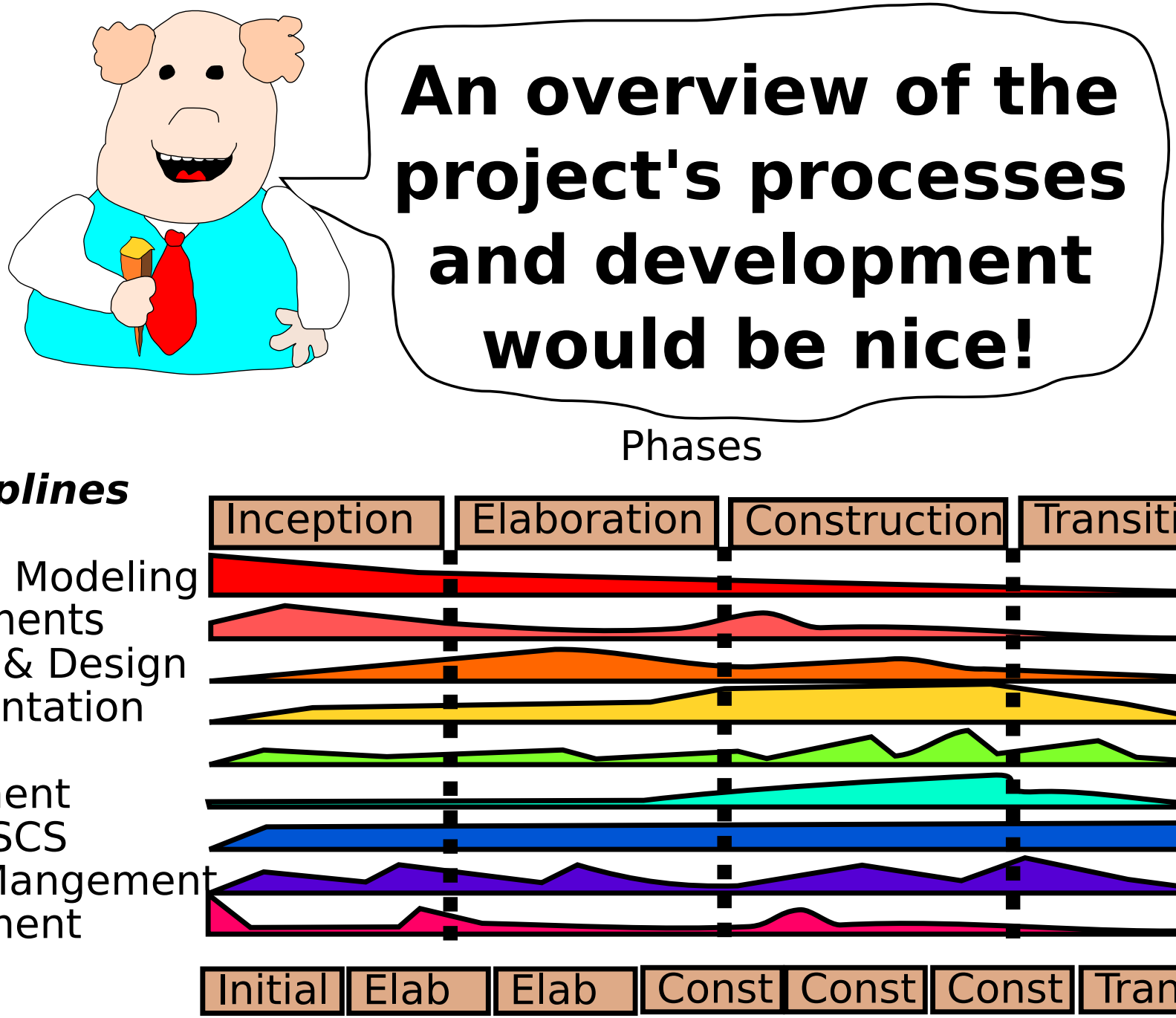
Rx

- Test First
- Scrums
- Story Cards



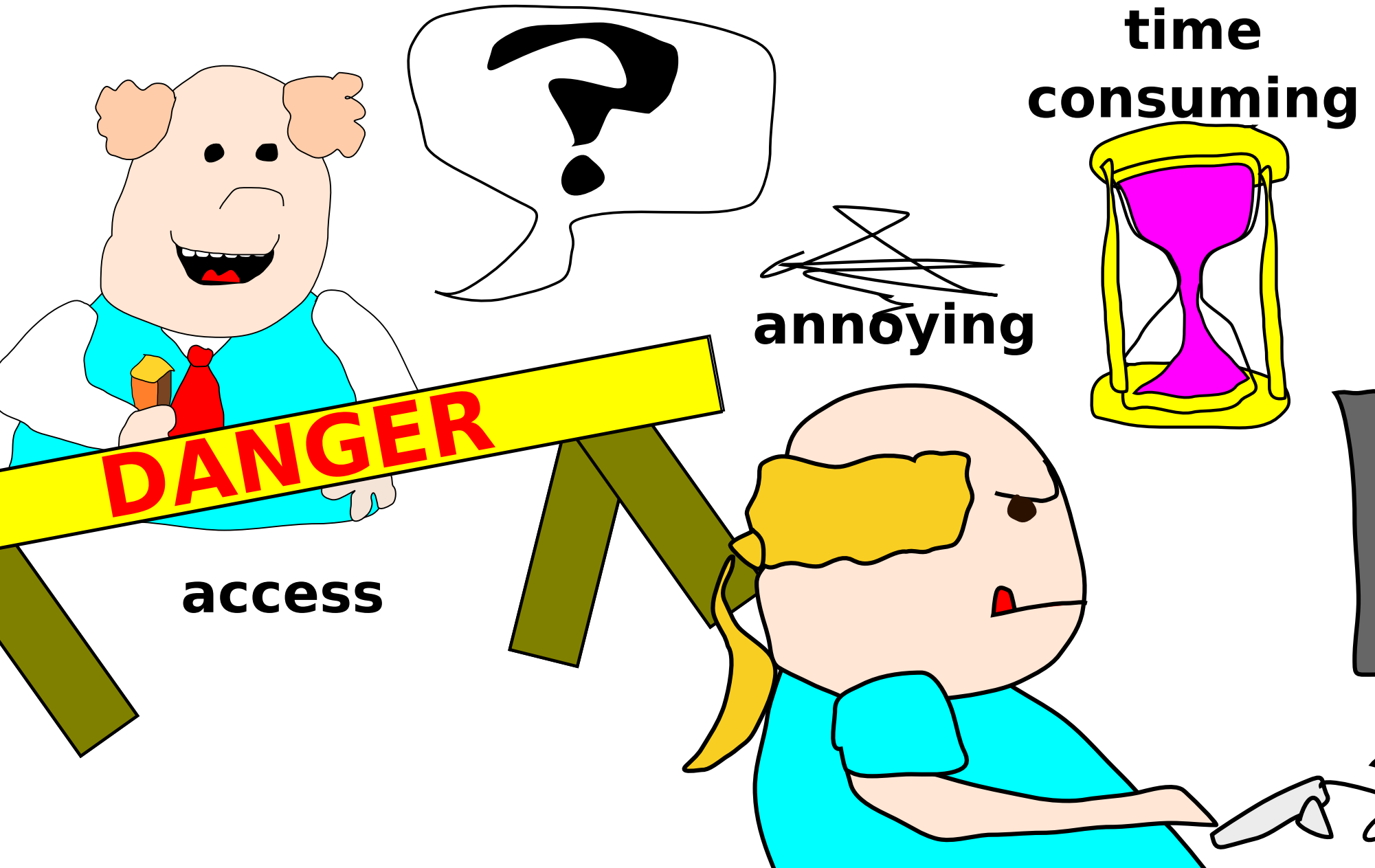
Formal Processes



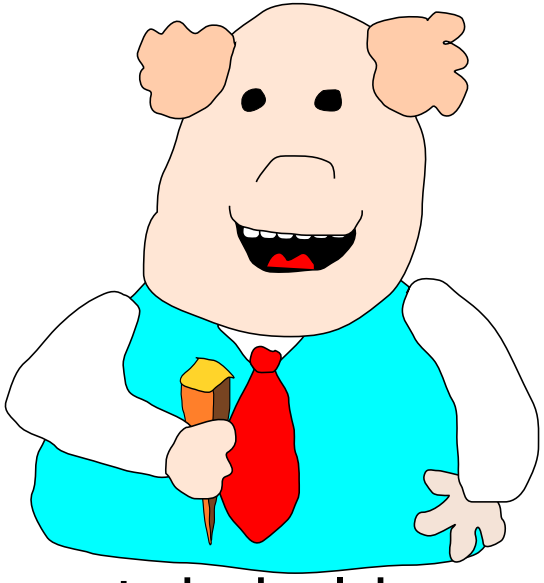


Example UP Process

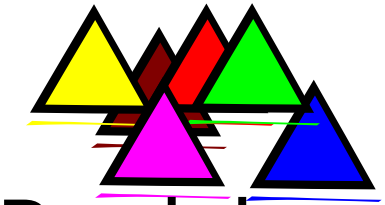
How to get an overview: Interviews



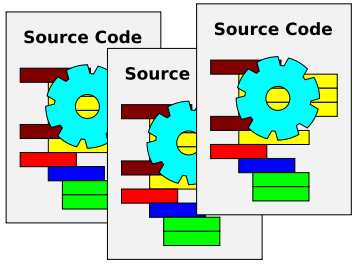
How to get an overview: Mining Software Repositories



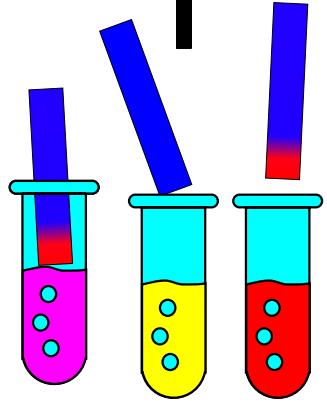
stakeholder



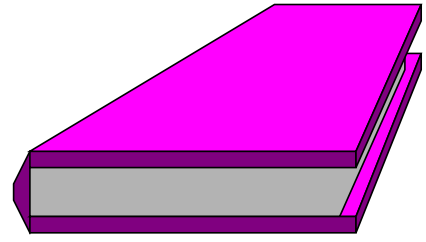
Revisions



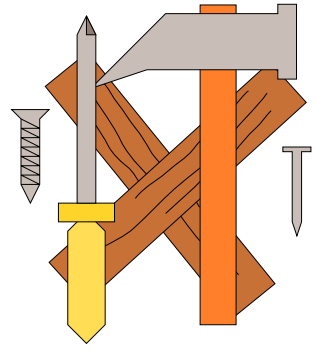
Source



Tests

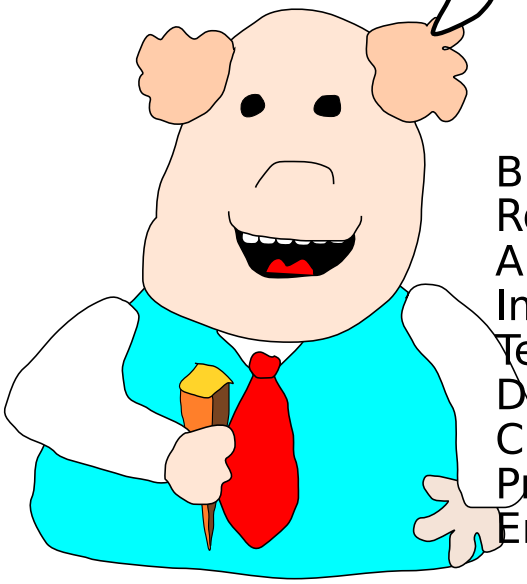


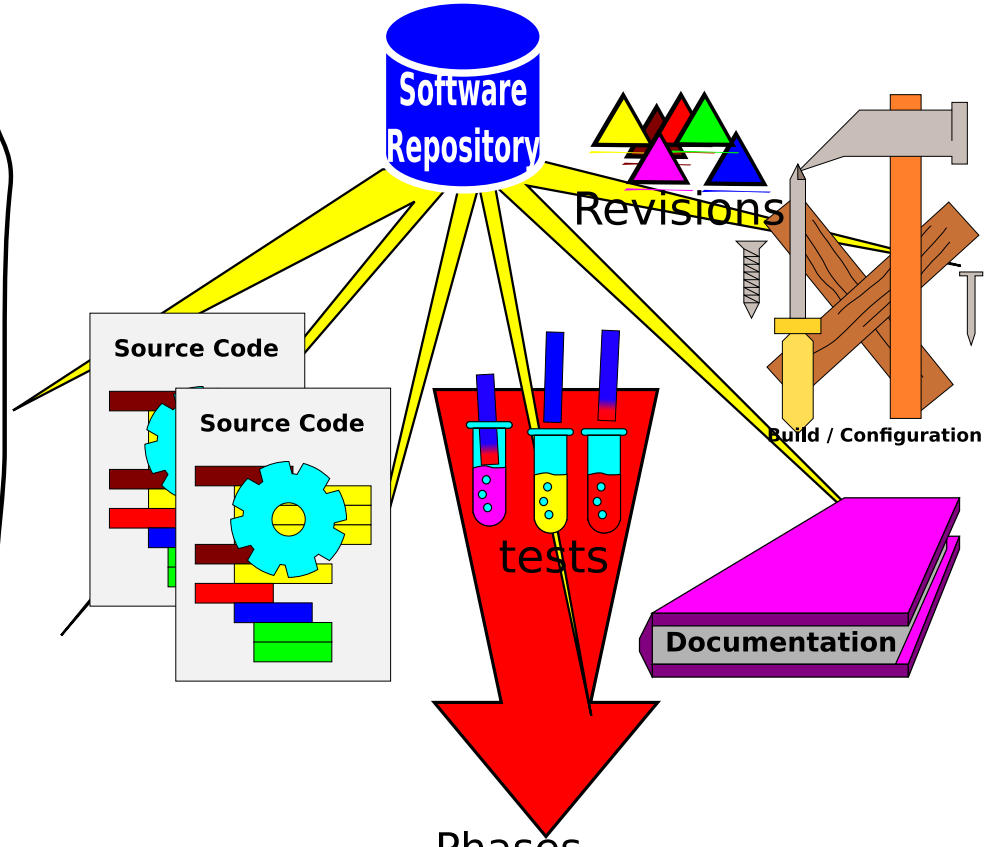
Docs



Build

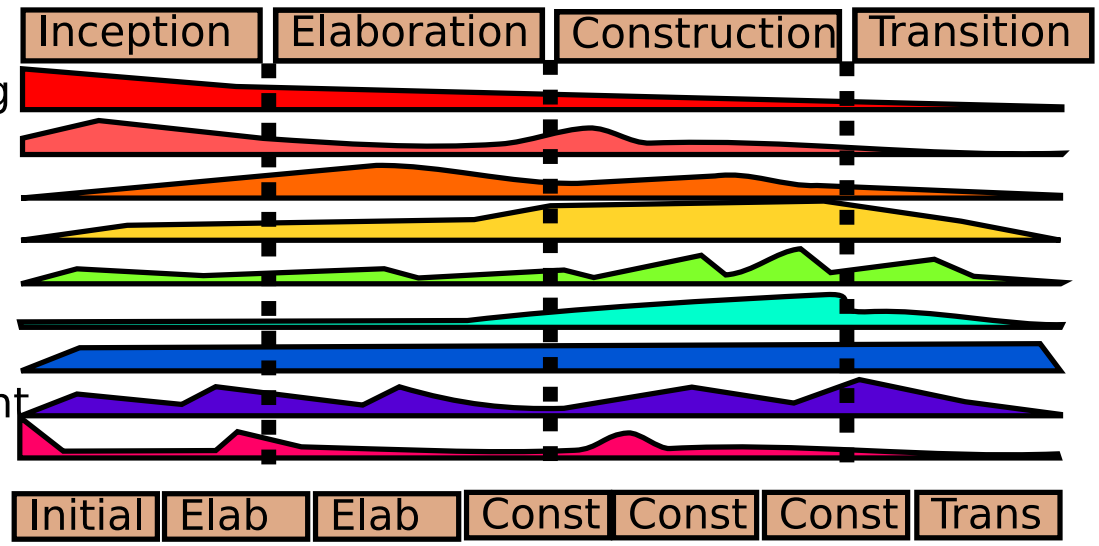
Can't we just summarize what is going on within this project?





Disciplines

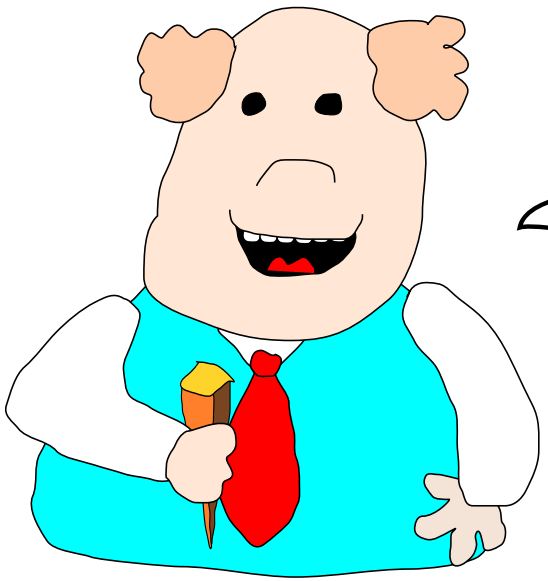
Business Modeling
Requirements
Analysis & Design
Implementation
Test
Deployment
CM and SCS
Project Mangement
Environment



Proposed Process

Recovered Process

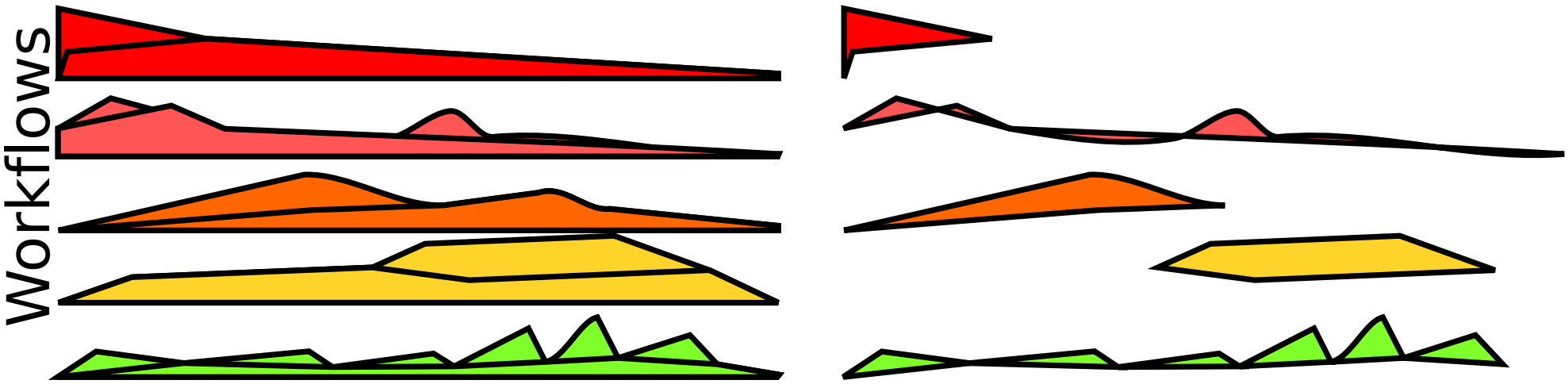
Workflows

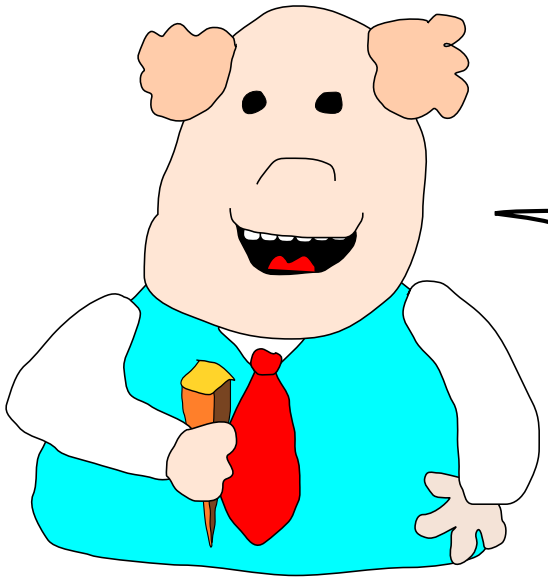


Is my proposed process actually being used?

Proposed and Recovered Process Overlayed

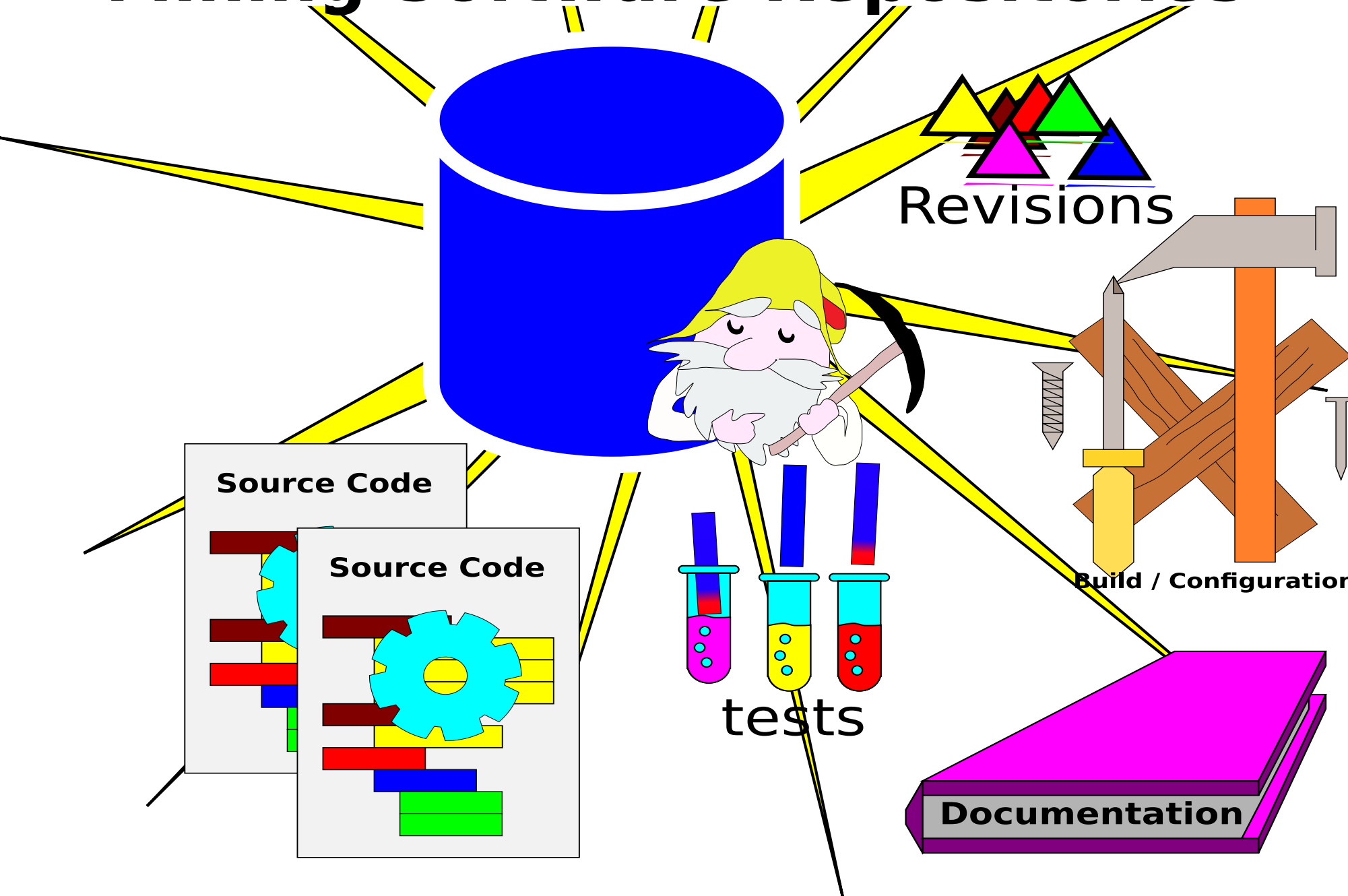
Differences between Proposed and Recovered



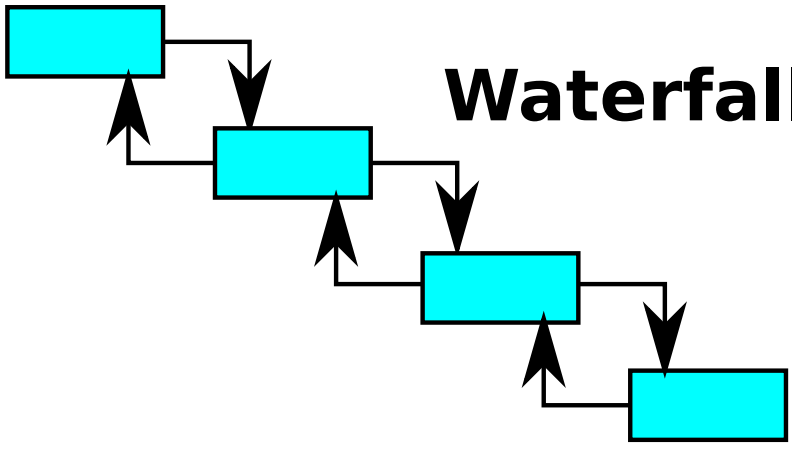


I can compare and contrast the observed process versus the expected process!

Mining Software Repositories



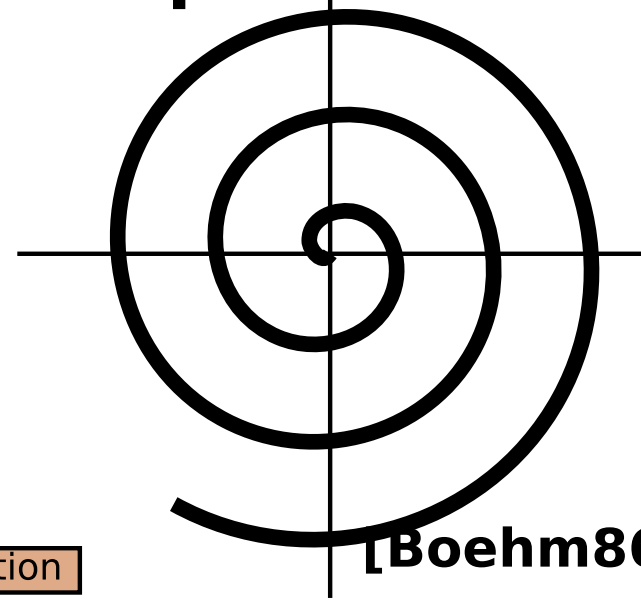
Software Development Processes



Waterfall

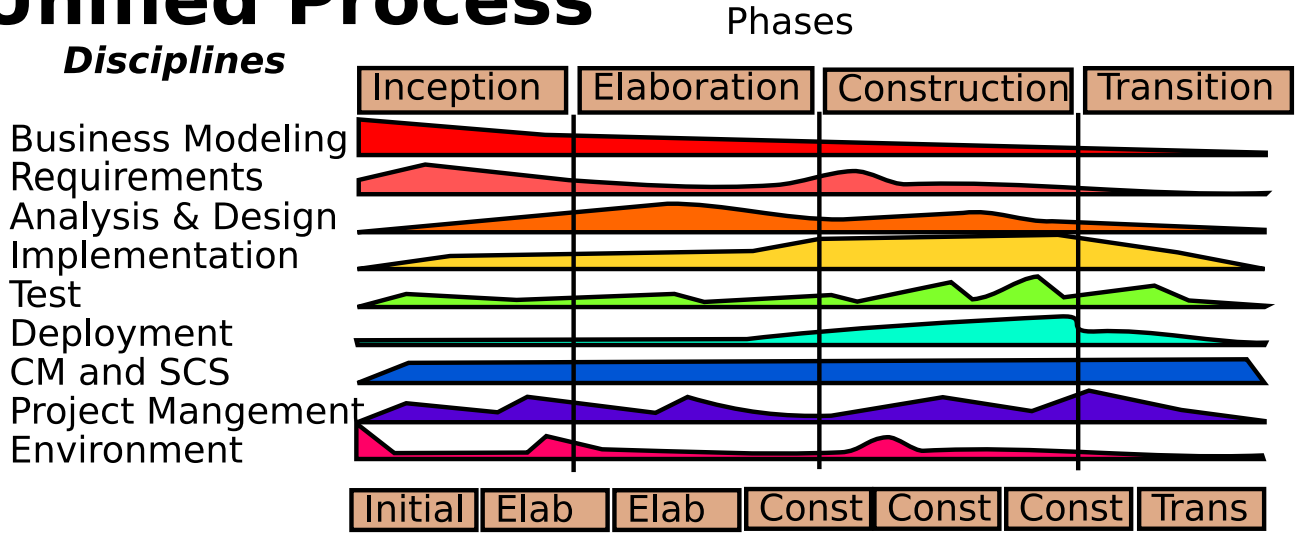
[Royce87]

Spiral



[Boehm86]

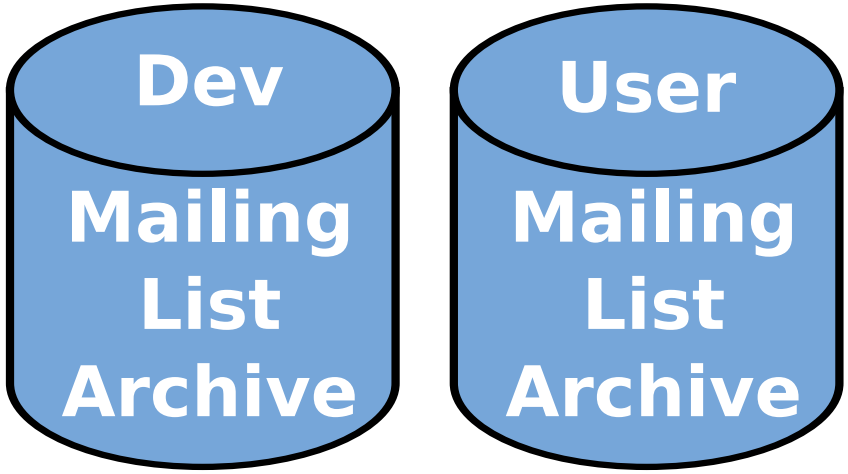
Unified Process

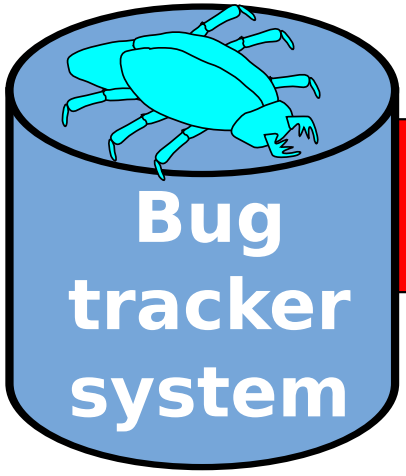


[Jacobson99]

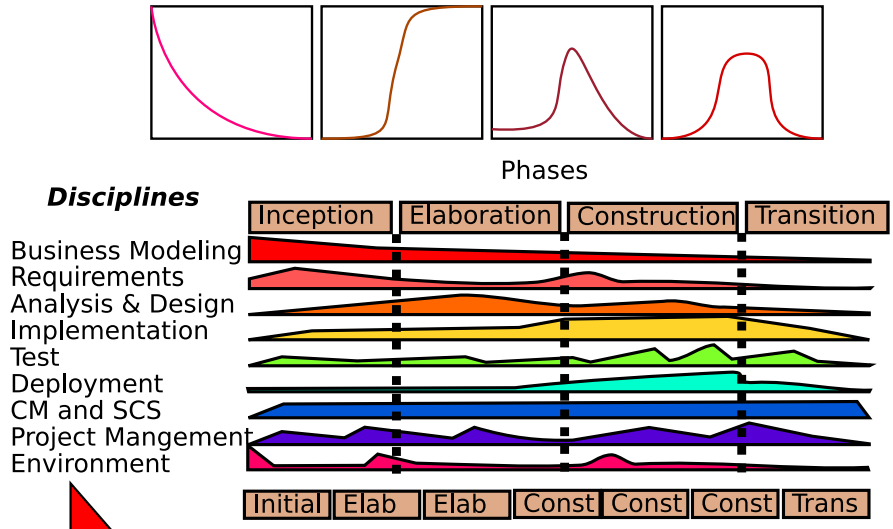
* **CMM**
* **SDLC**

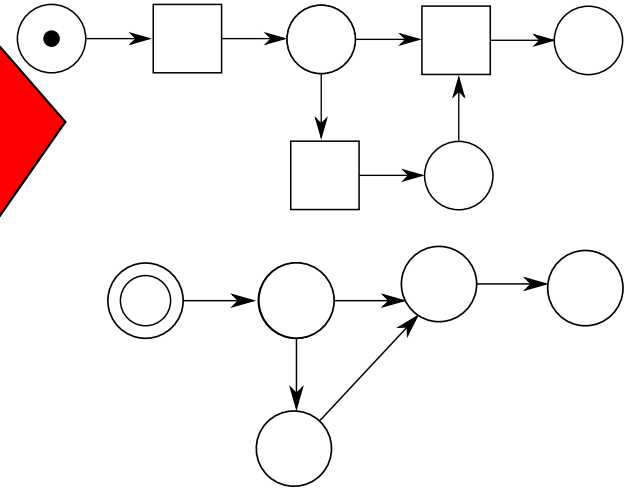
Process Recovery



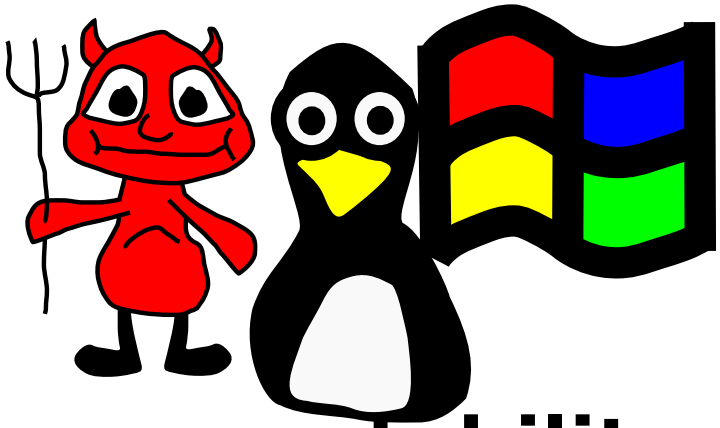


after the fact

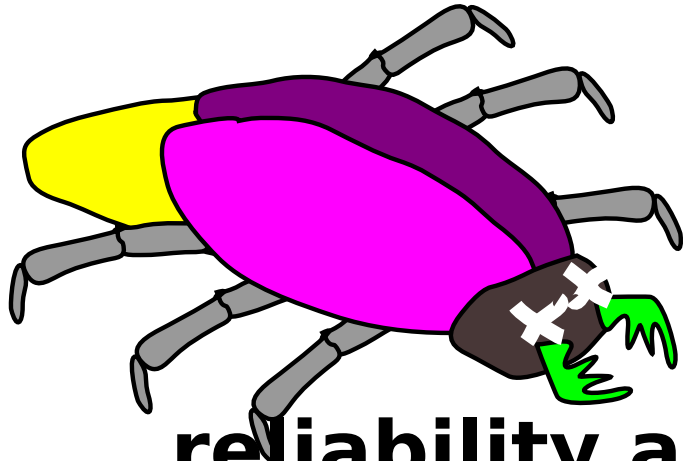




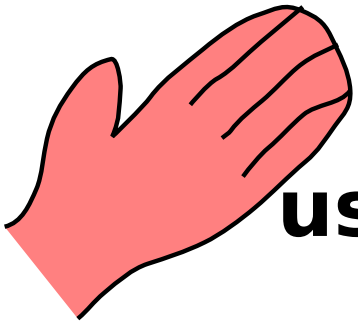
Quality Related
Non functional requirements



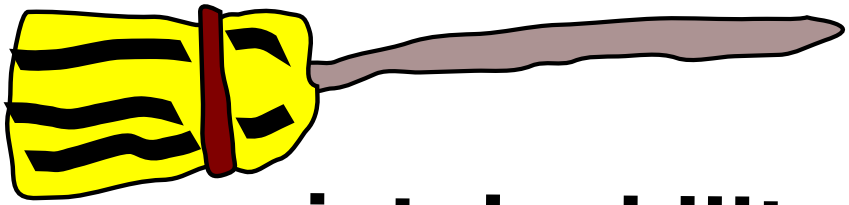
portability



reliability and
functionality
(includes correctness)

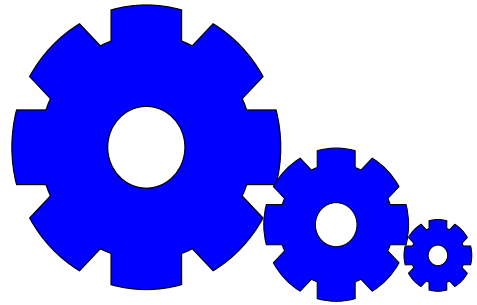


usability



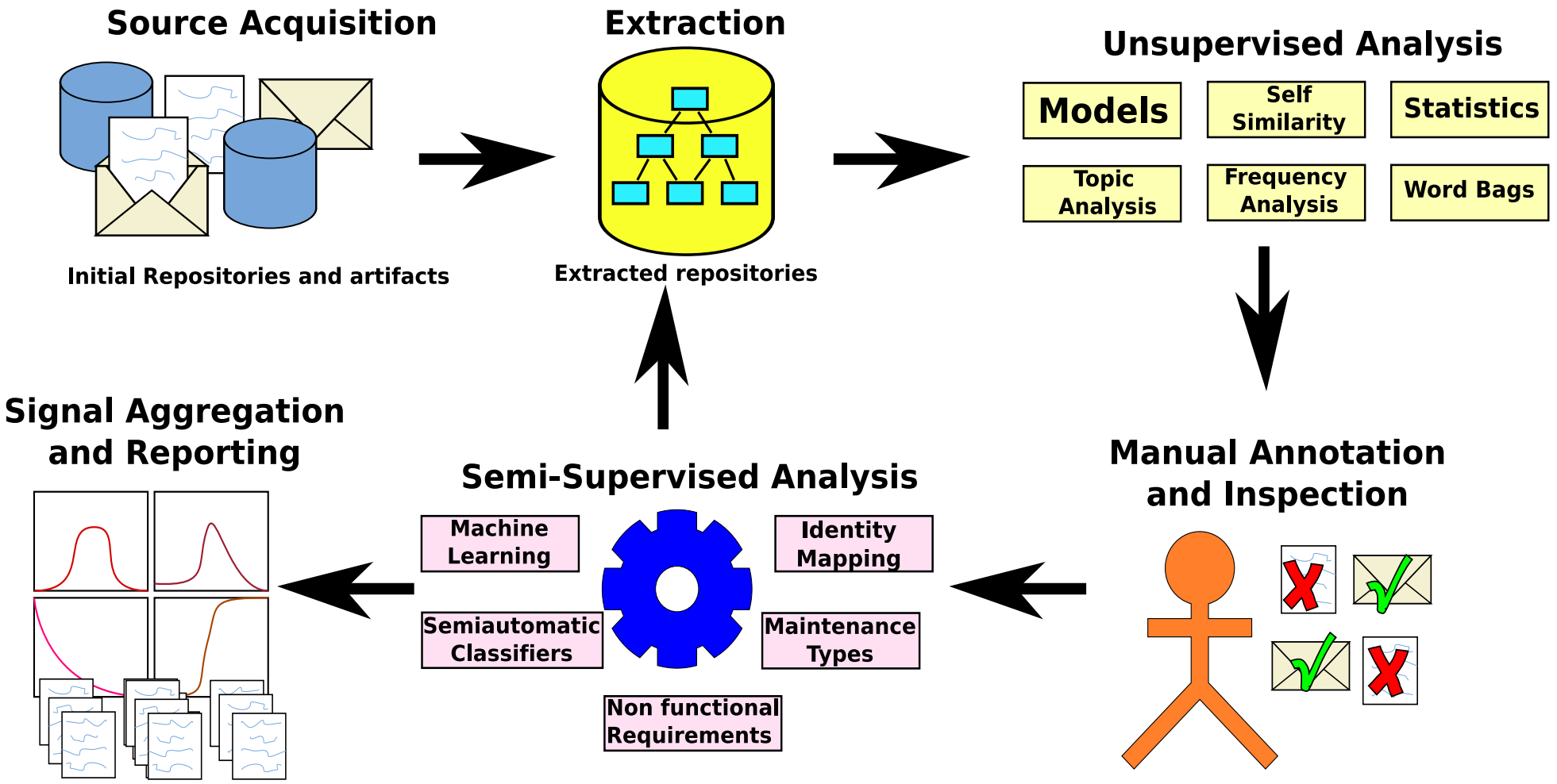
maintainability

efficiency

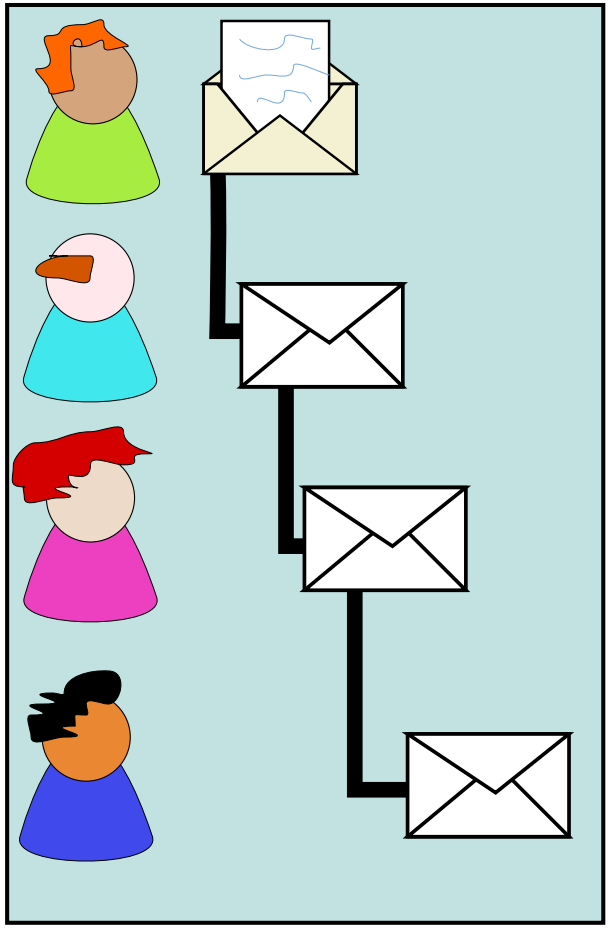


[cleland-huang03]
[ernst10]

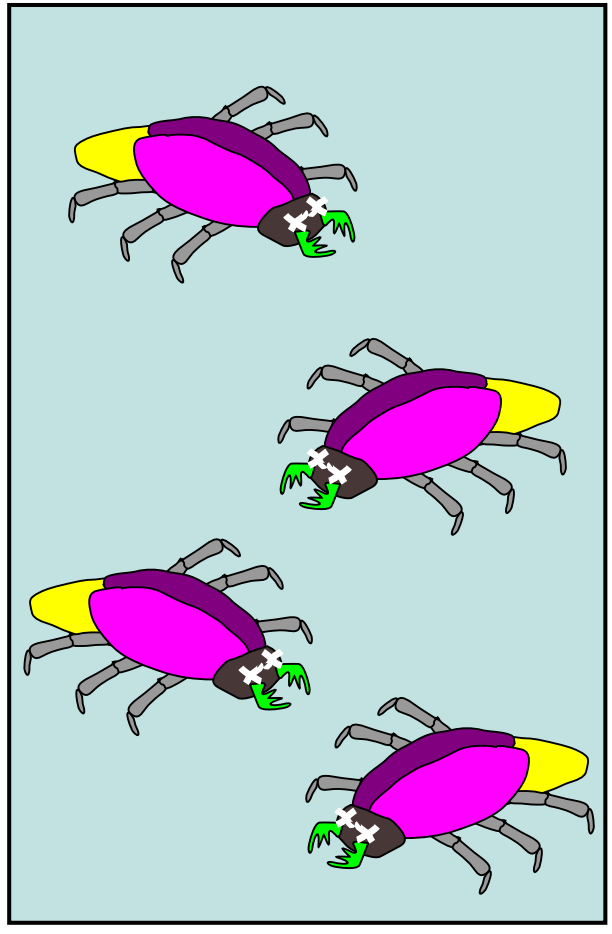
Methodology:
Recovered Unified Process Views



Source Acquisition



discussions

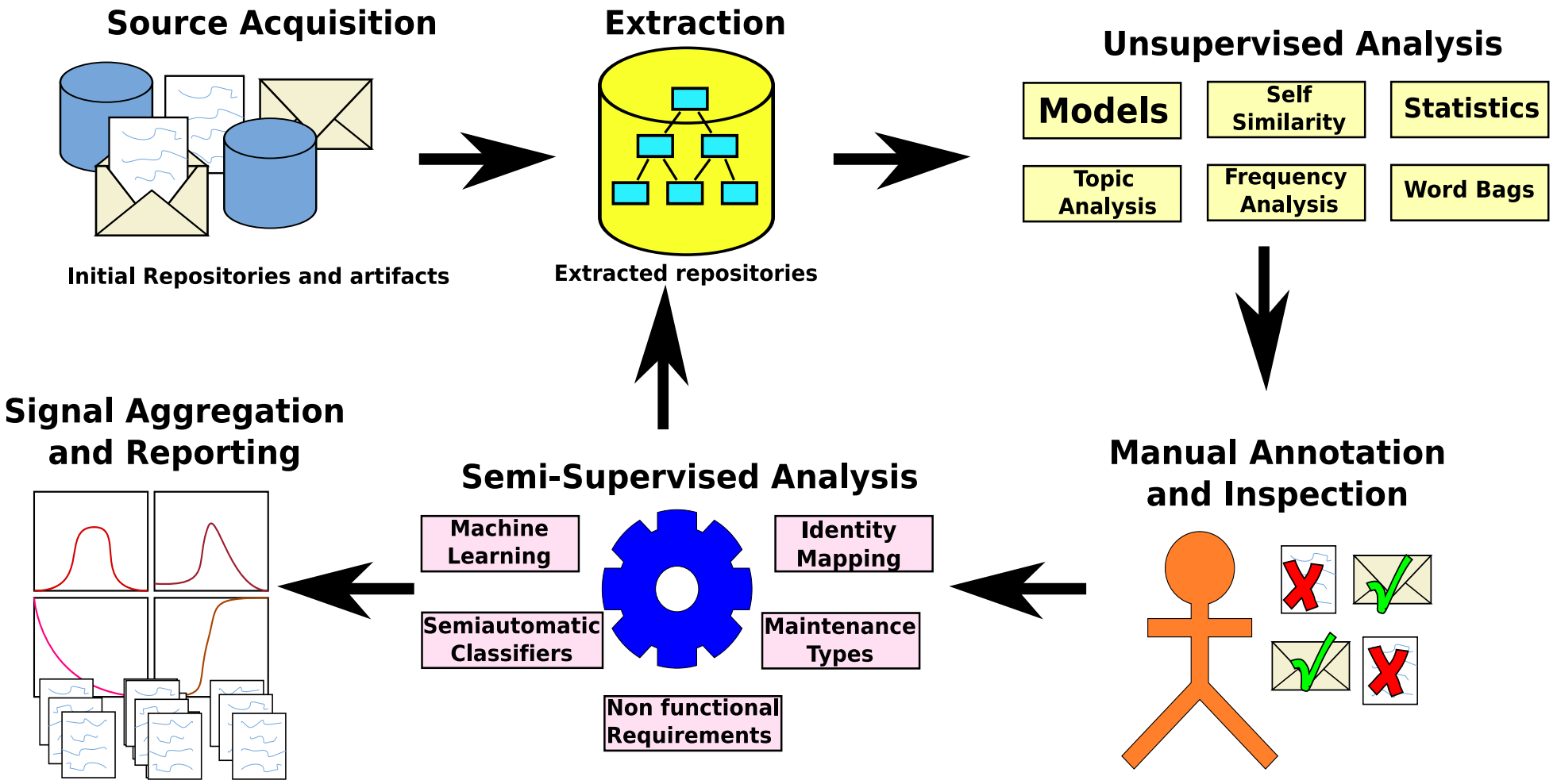


bugs

source

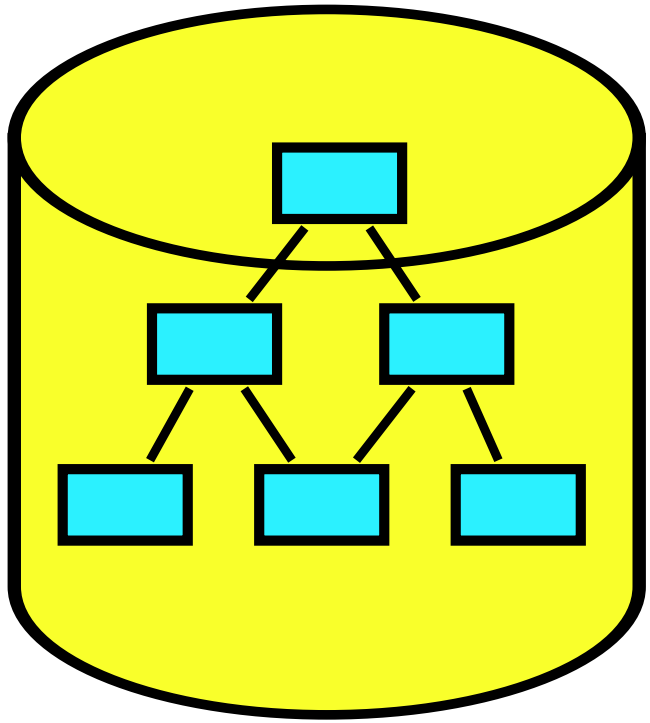
Initial Repositories and artifacts

Methodology:
Recovered Unified Process Views



on

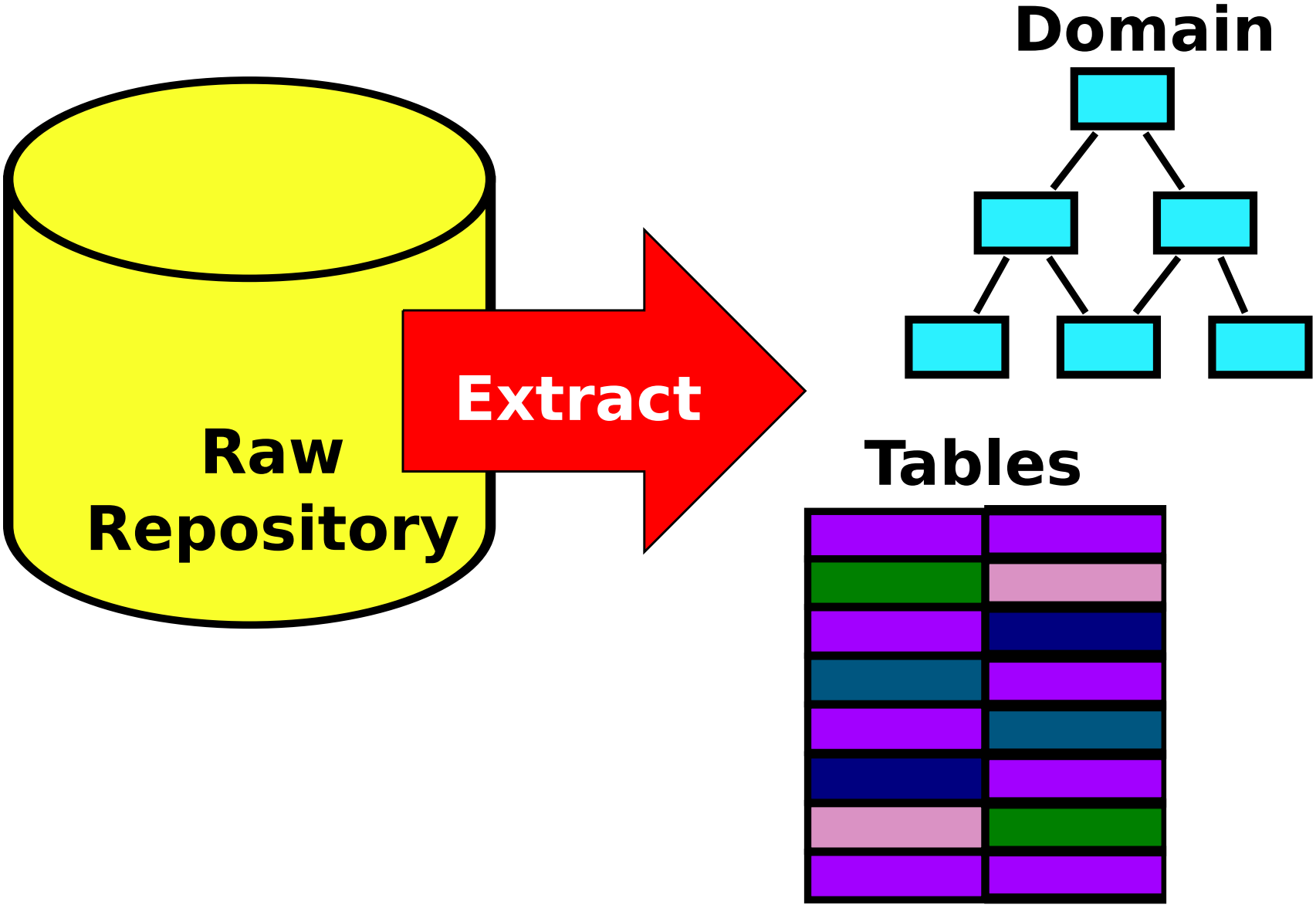
Extraction



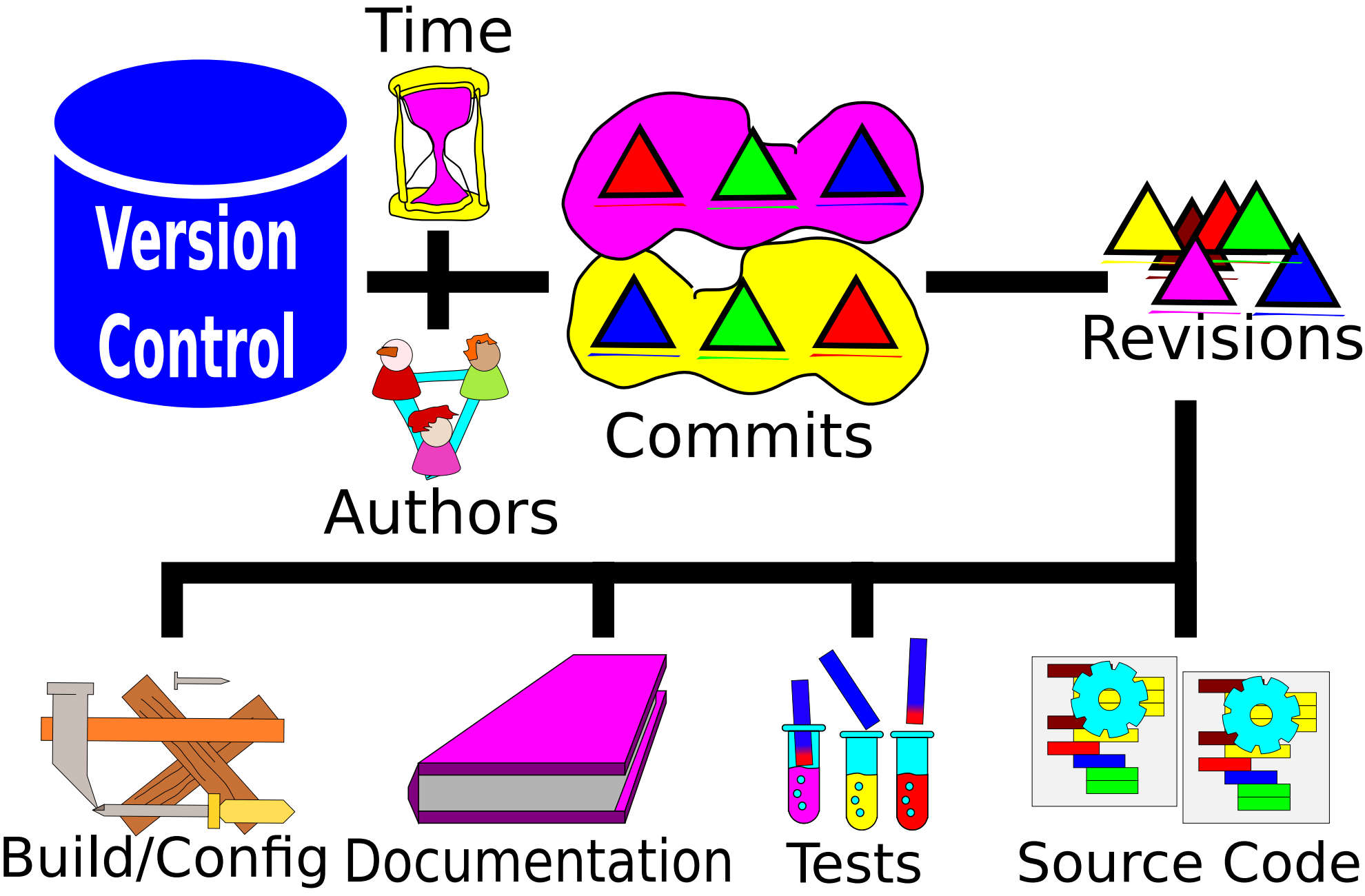
nacts

Extracted repositories

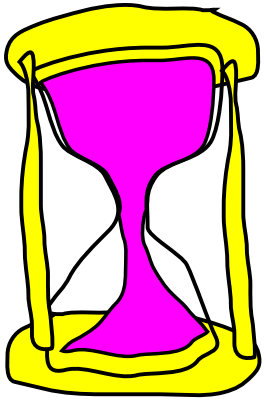
Extraction

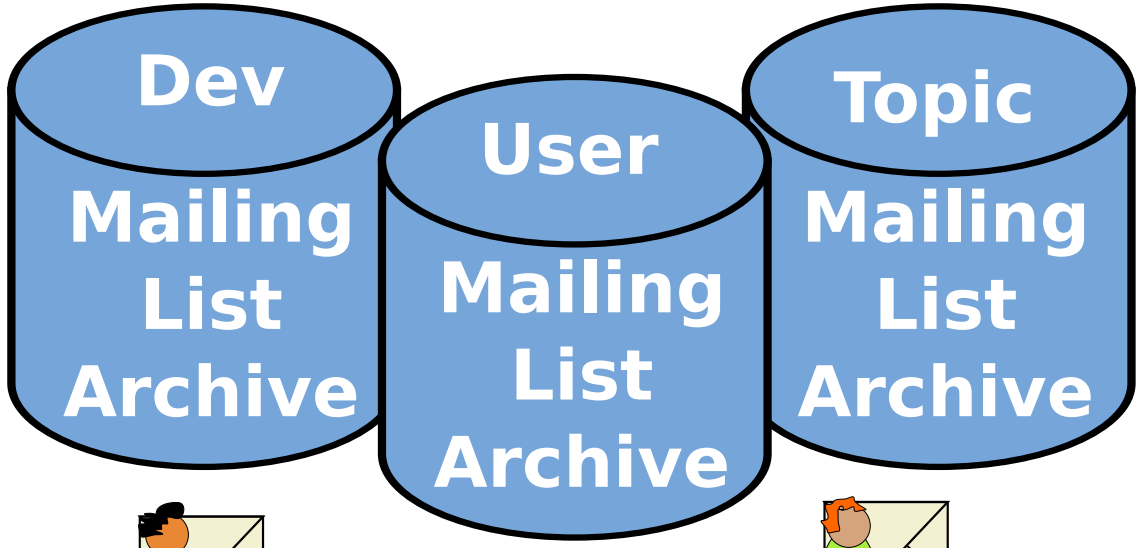


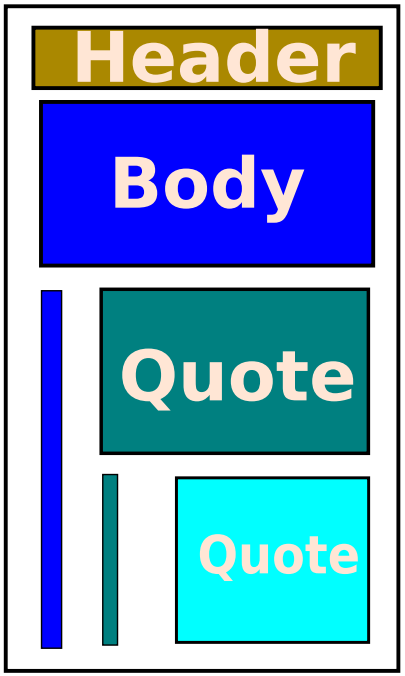
Extraction: Version Control



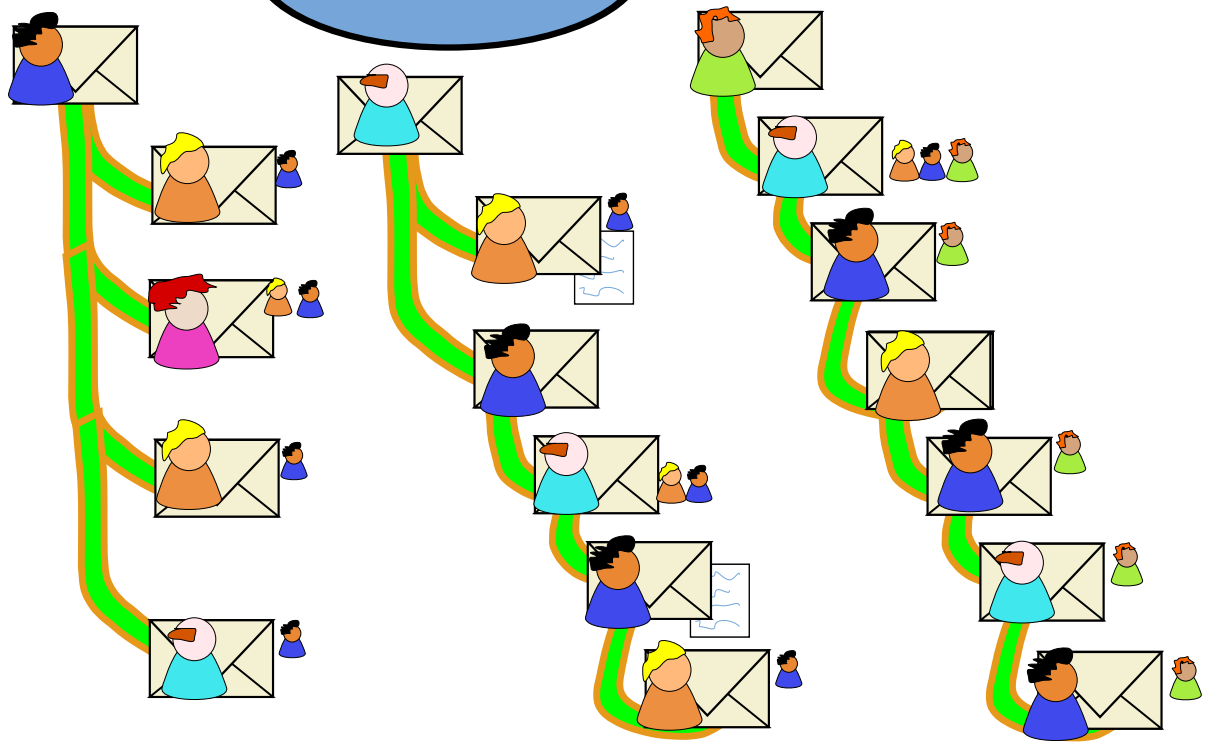
Extraction: Mailing list archives



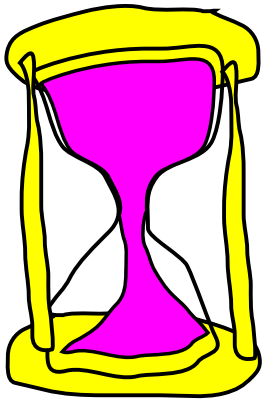


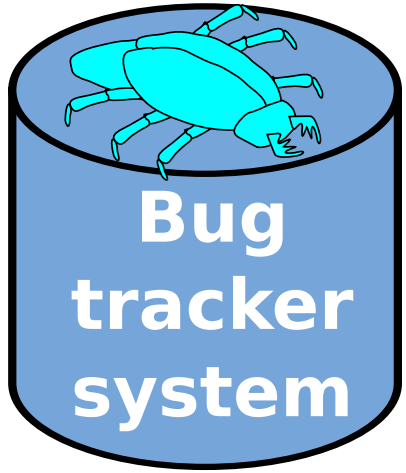


Email



Extraction: Bug trackers

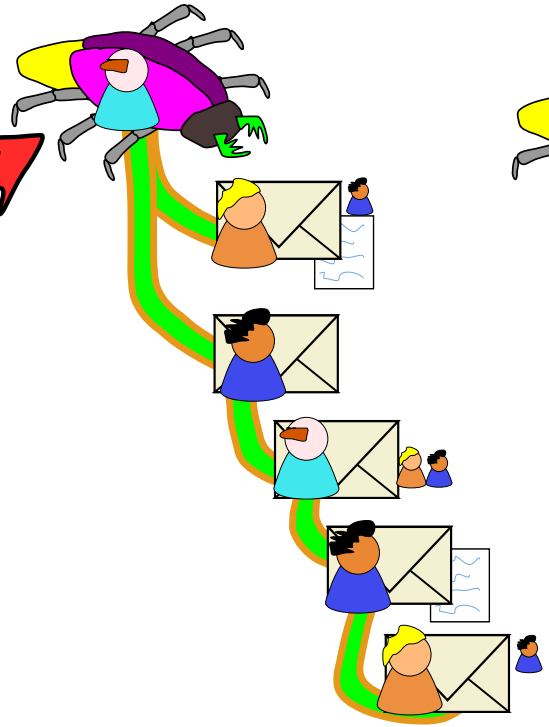


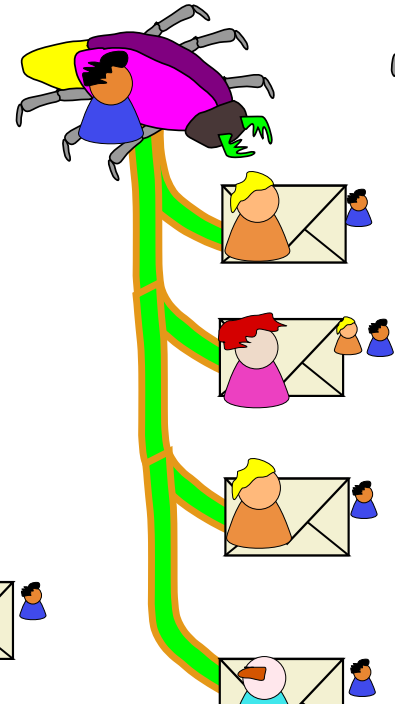


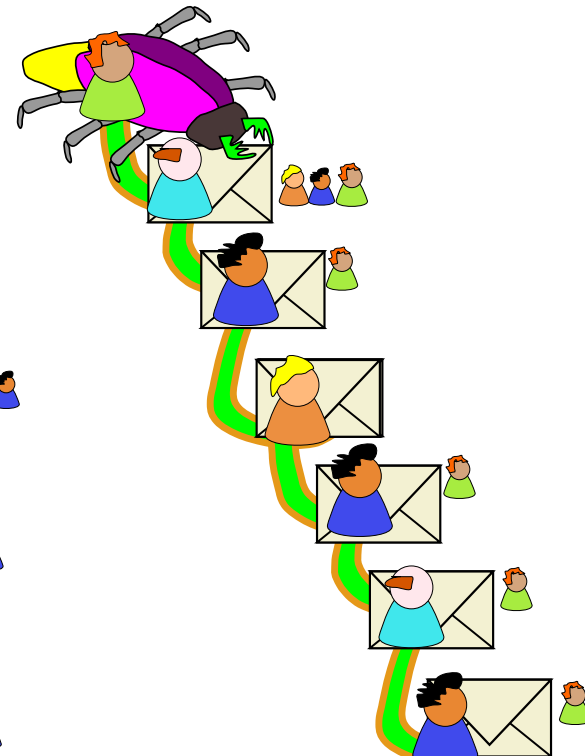
bug id

Cricket, Jim #57981

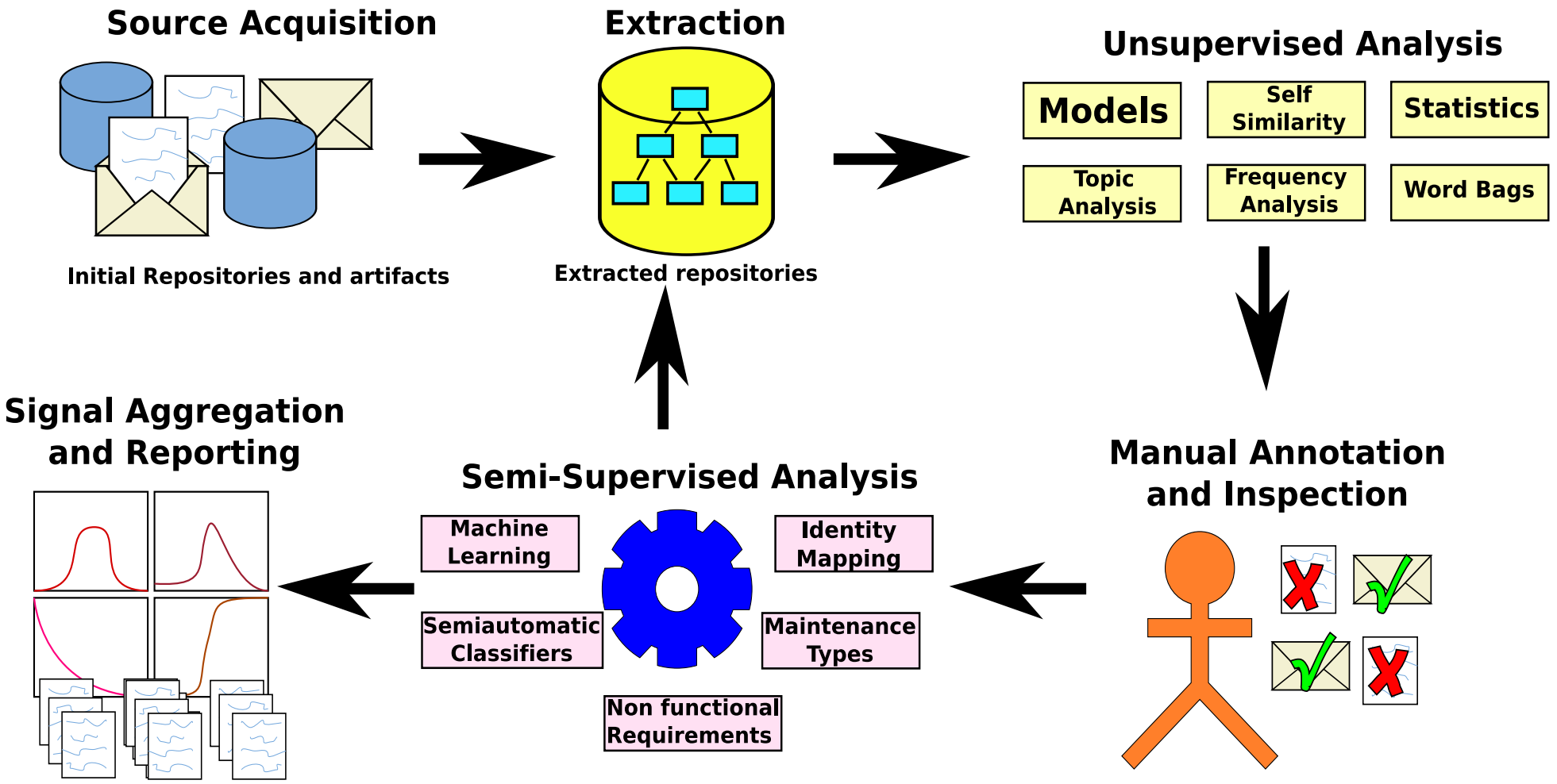
Summary
Severity
Description
...







Methodology:
Recovered Unified Process Views



Unsupervised Analysis

Models

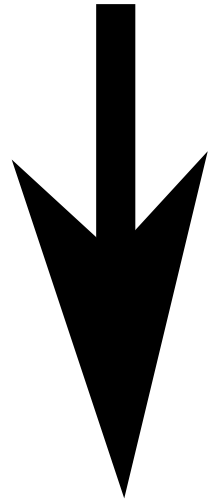
Self Similarity

Statistics

Topic Analysis

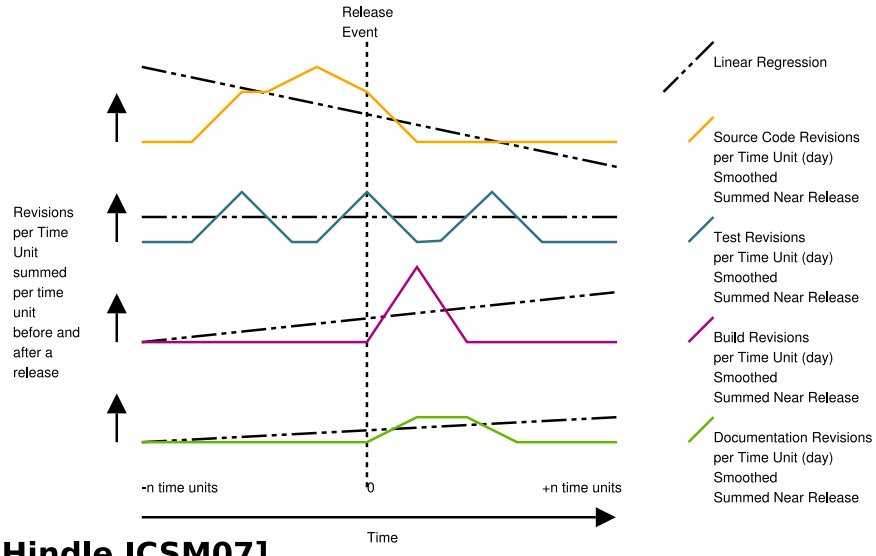
Frequency Analysis

Word Bags



Unsupervised Analysis

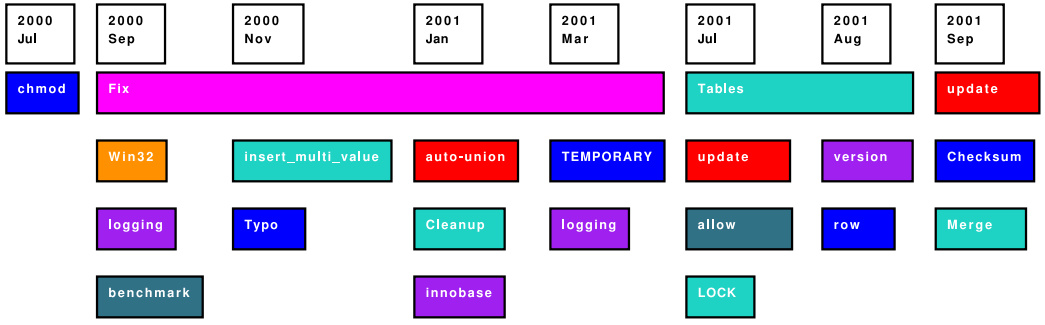
Partition Revisions



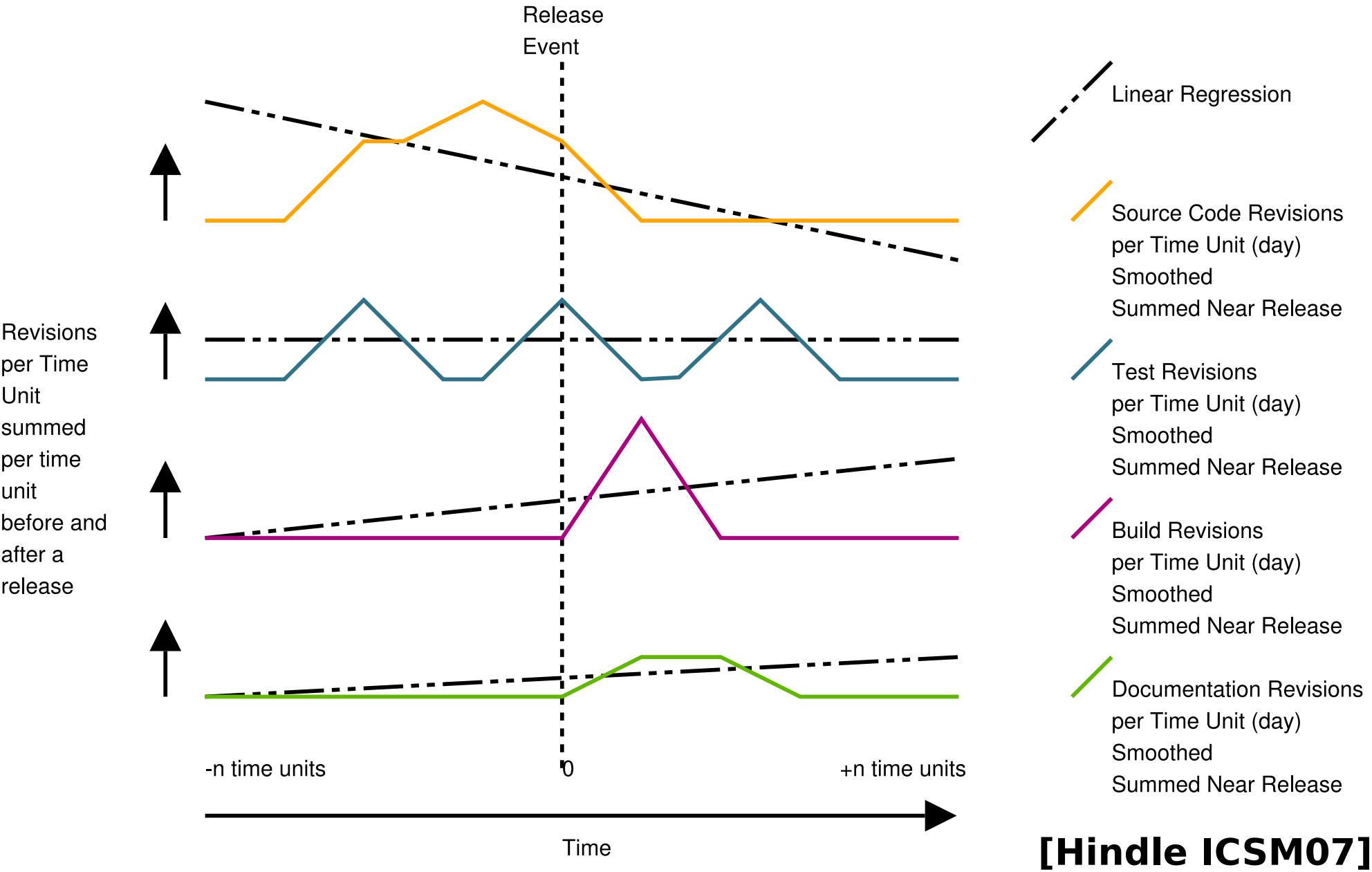
[Hindle ICSM07]

Word Bag Analysis

Topic Analysis

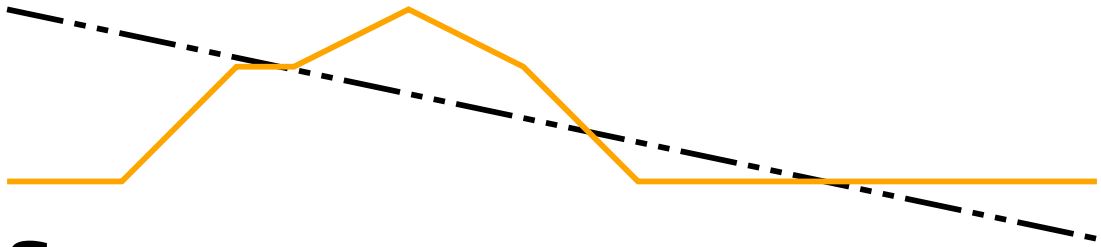


Unsupervised Analysis: STBD



[Hindle ICSM07]

Unsupervised Analysis: Source



*.hs

*.ml

*.tcl

*.php

*.lisp

*.c

*.C

*.pl

*.rb

*.java

*.cpp

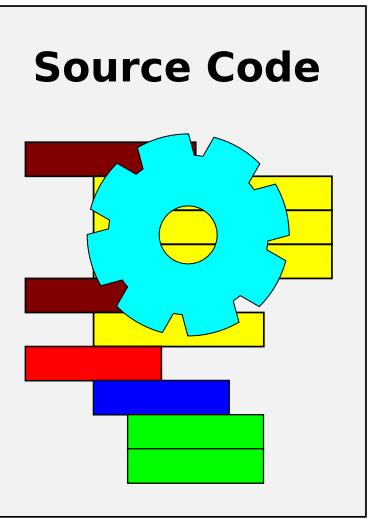
*.py

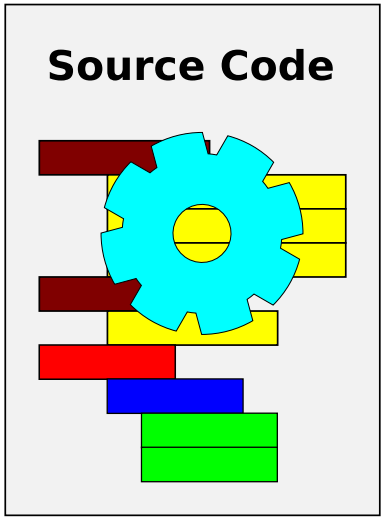
*.scm

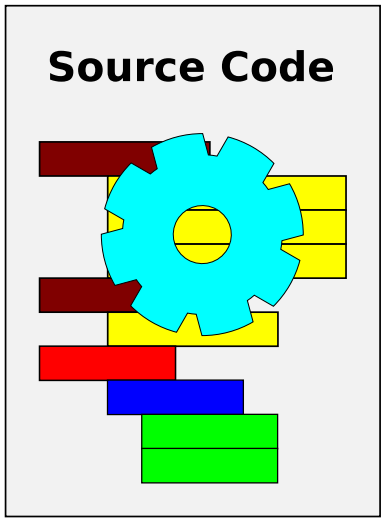
*.sql

*.pl

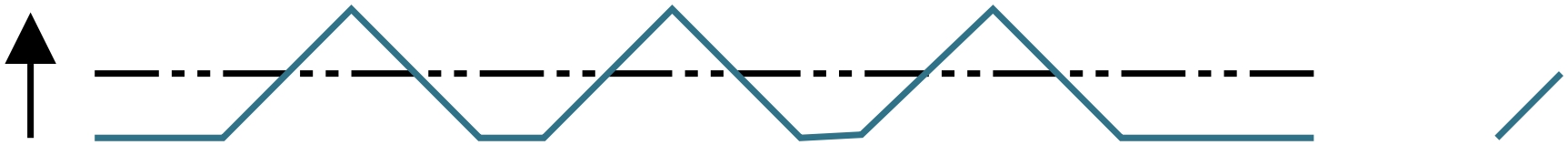
*.pm







Unsupervised Analysis: Testing



Tests

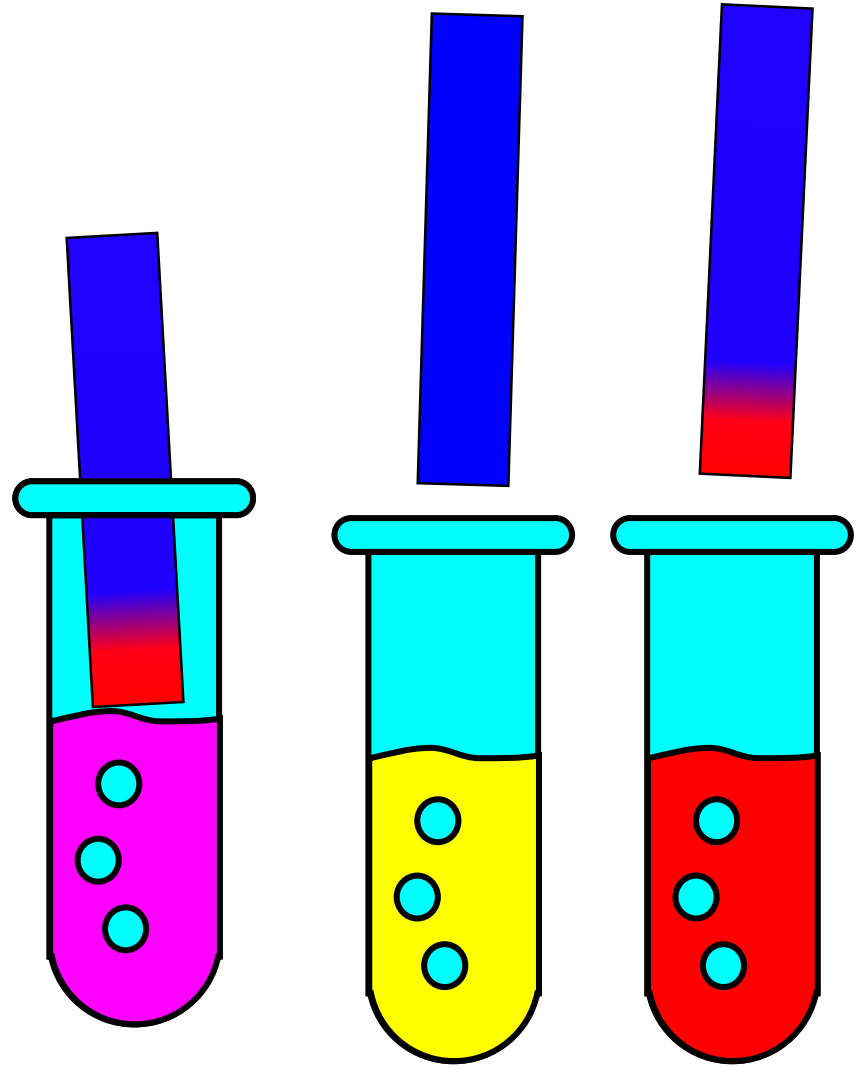
*.t

test

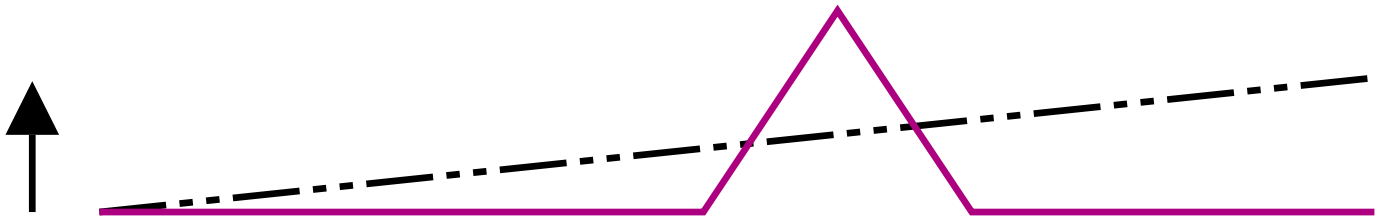
unit tests

*.test

Test



Unsupervised Analysis: Build files



Build

Makefile.*

Makefile

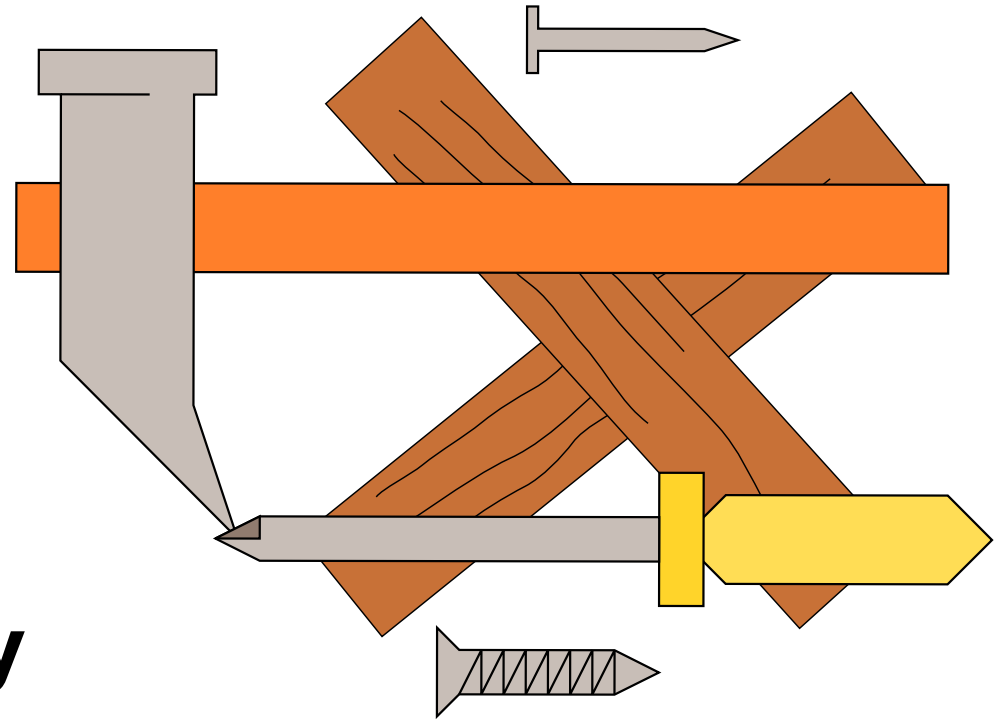
configure

configure.*

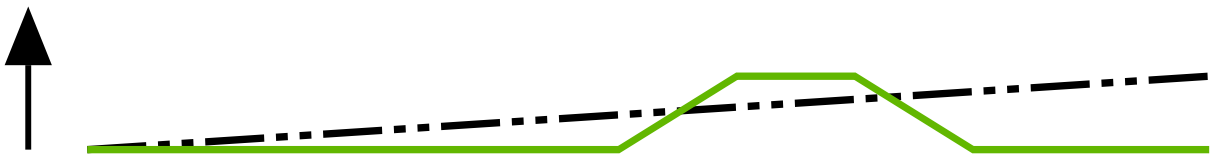
build.xml

setup.hs

setup.py



Unsupervised Analysis: Documentation



Documentation

FILES

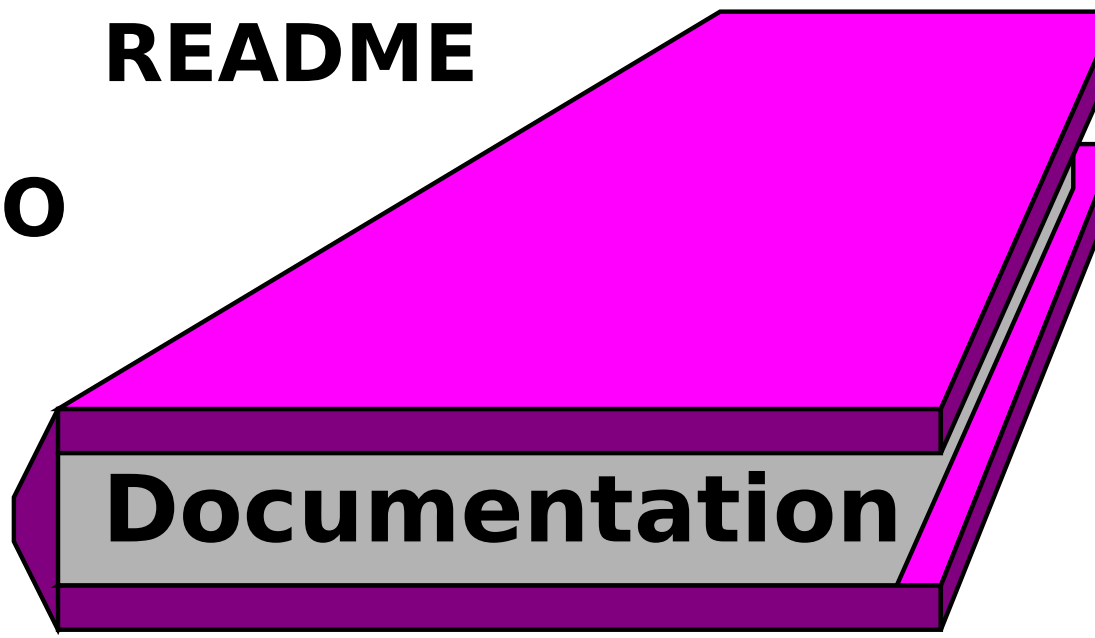
*.tex *.doxygen

doc/ INSTALL

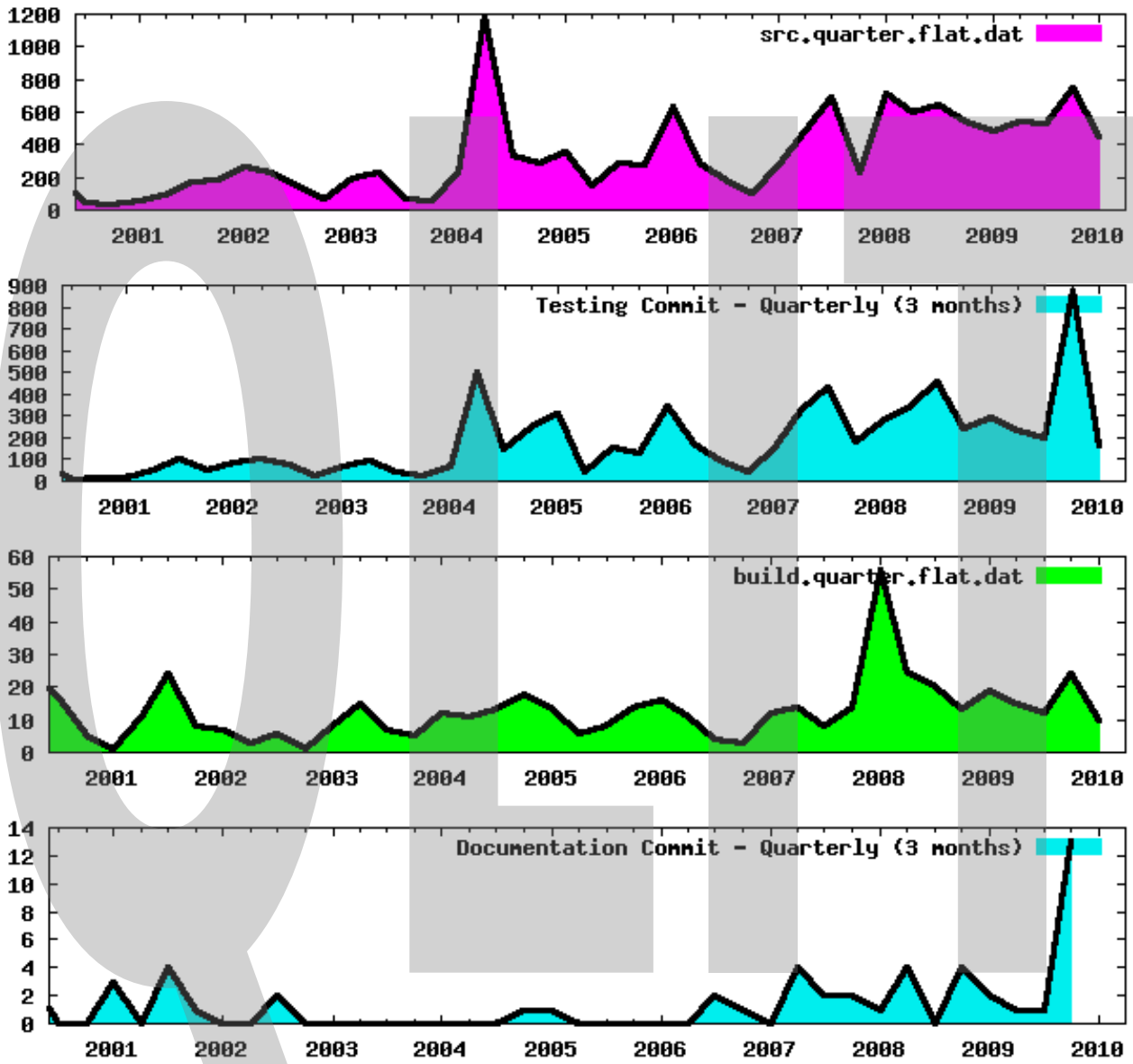
AUTHORS README

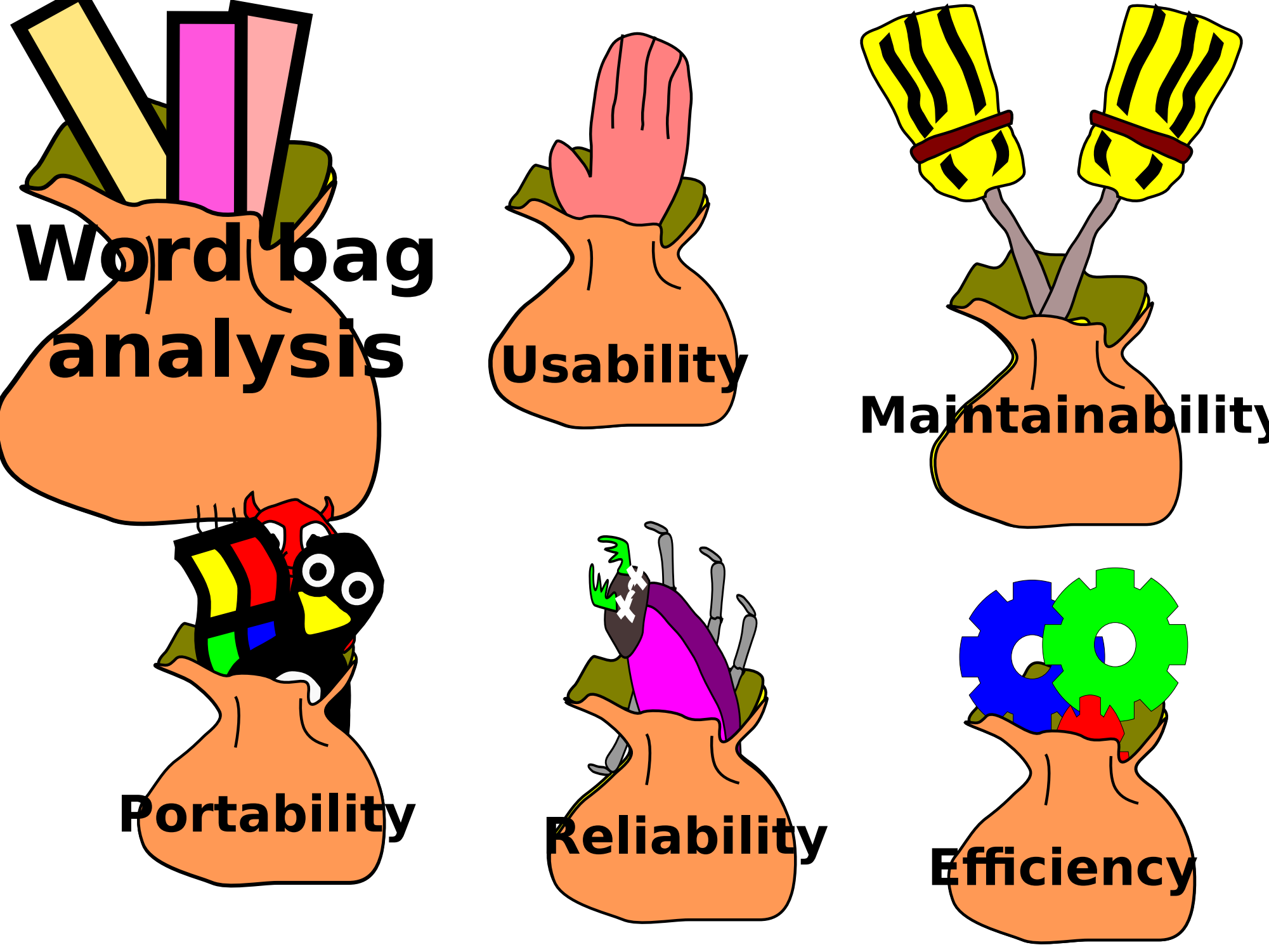
*.txt TODO

maybe: *.html
*.png
*.svg



Unsupervised Analysis: STBD applied

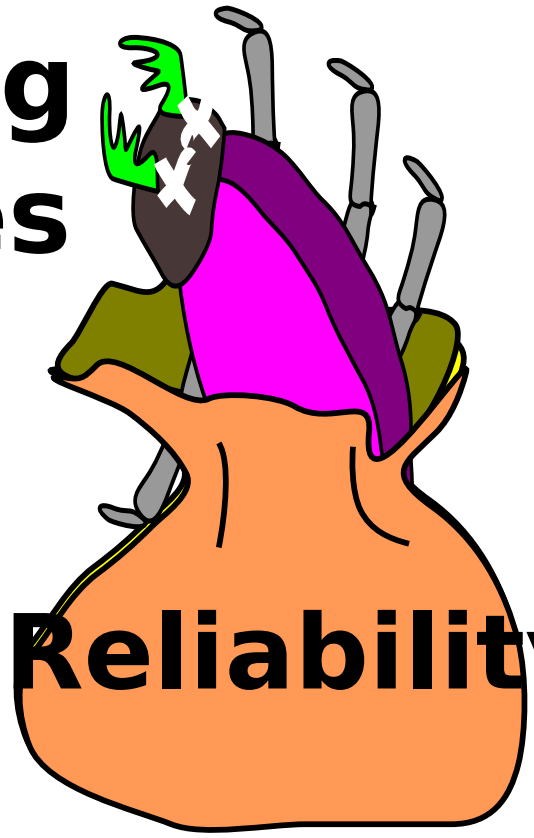



Word bag analysis
Usability
Maintainability
Portability
Reliability
Efficiency

Word Bag Examples

Portability

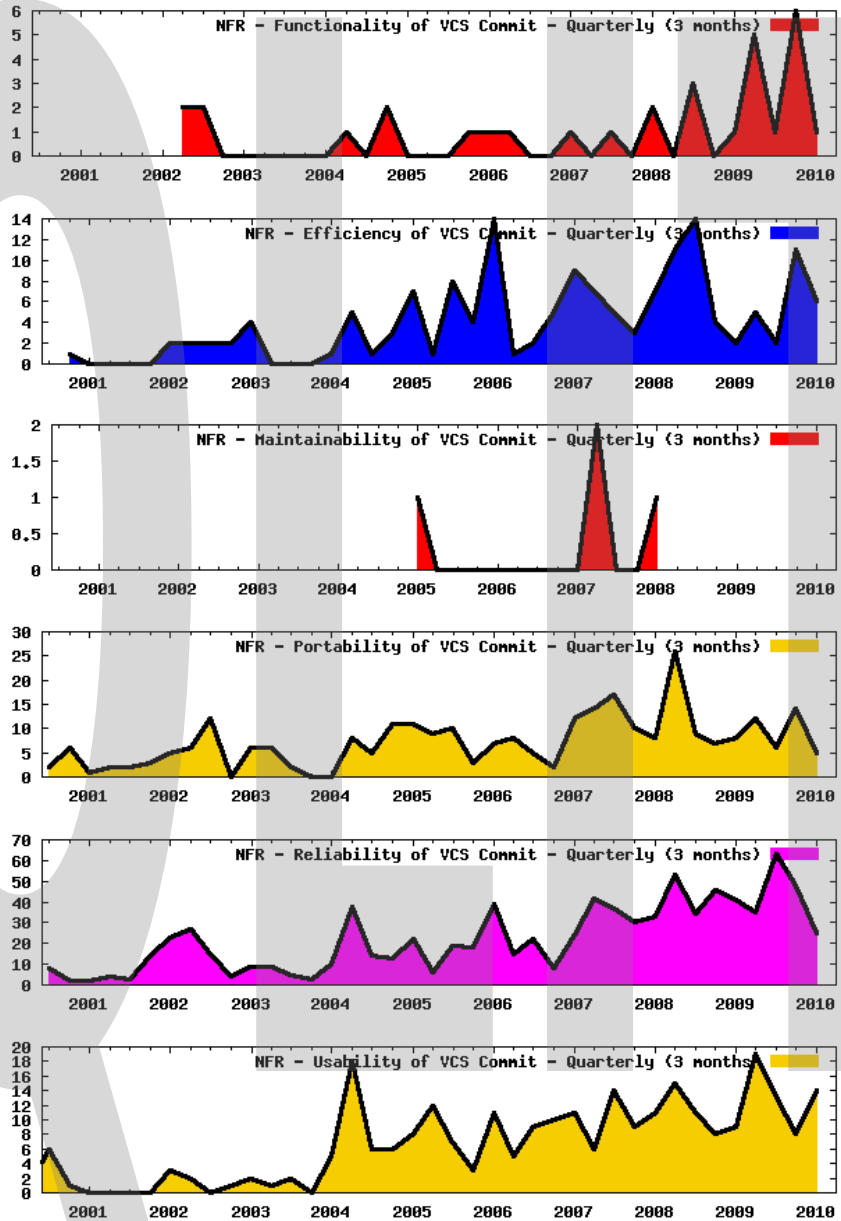
portability
transferability
interoperability
documentation
internationalization
i18n
...



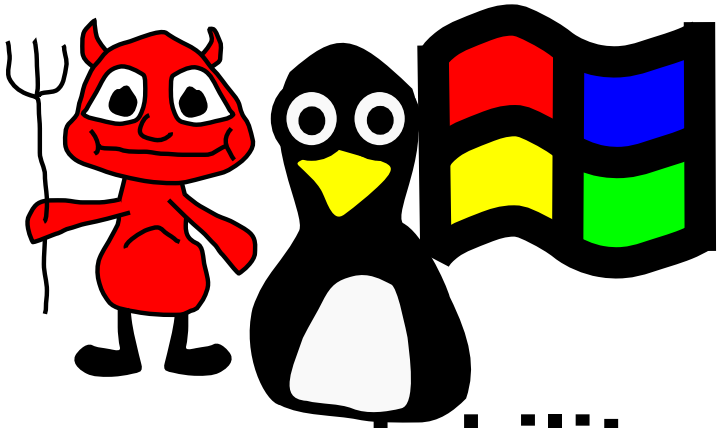
Reliability

reliability
failure
error
redundancy
fails
bug
...

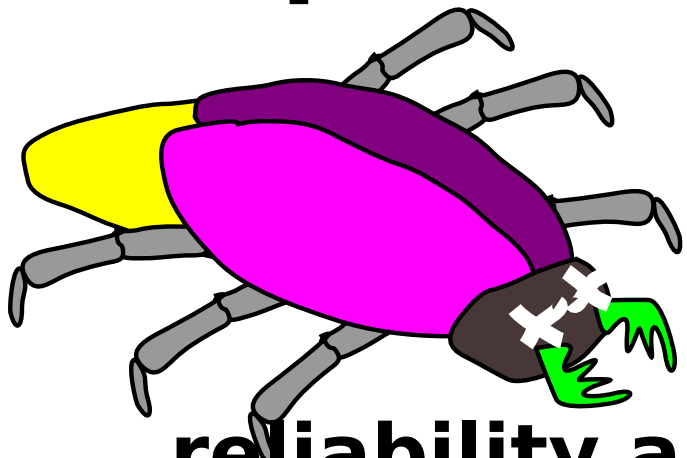
Unsupervised Analysis: Word Bag Applied



Label topics by Non functional requirements



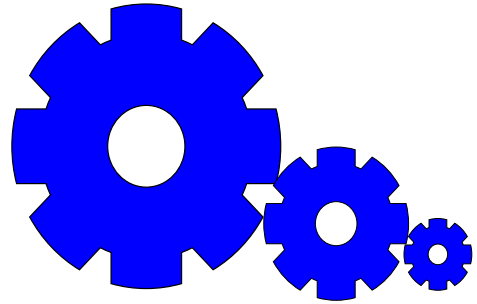
portability

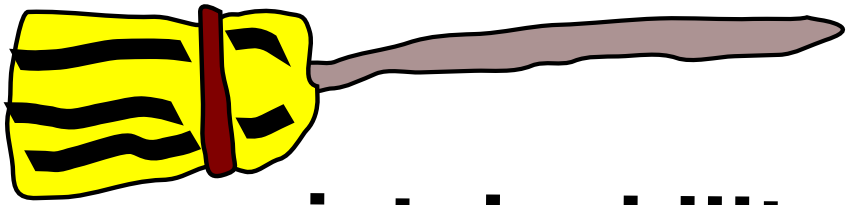


reliability and functionality (includes correctness)

usability

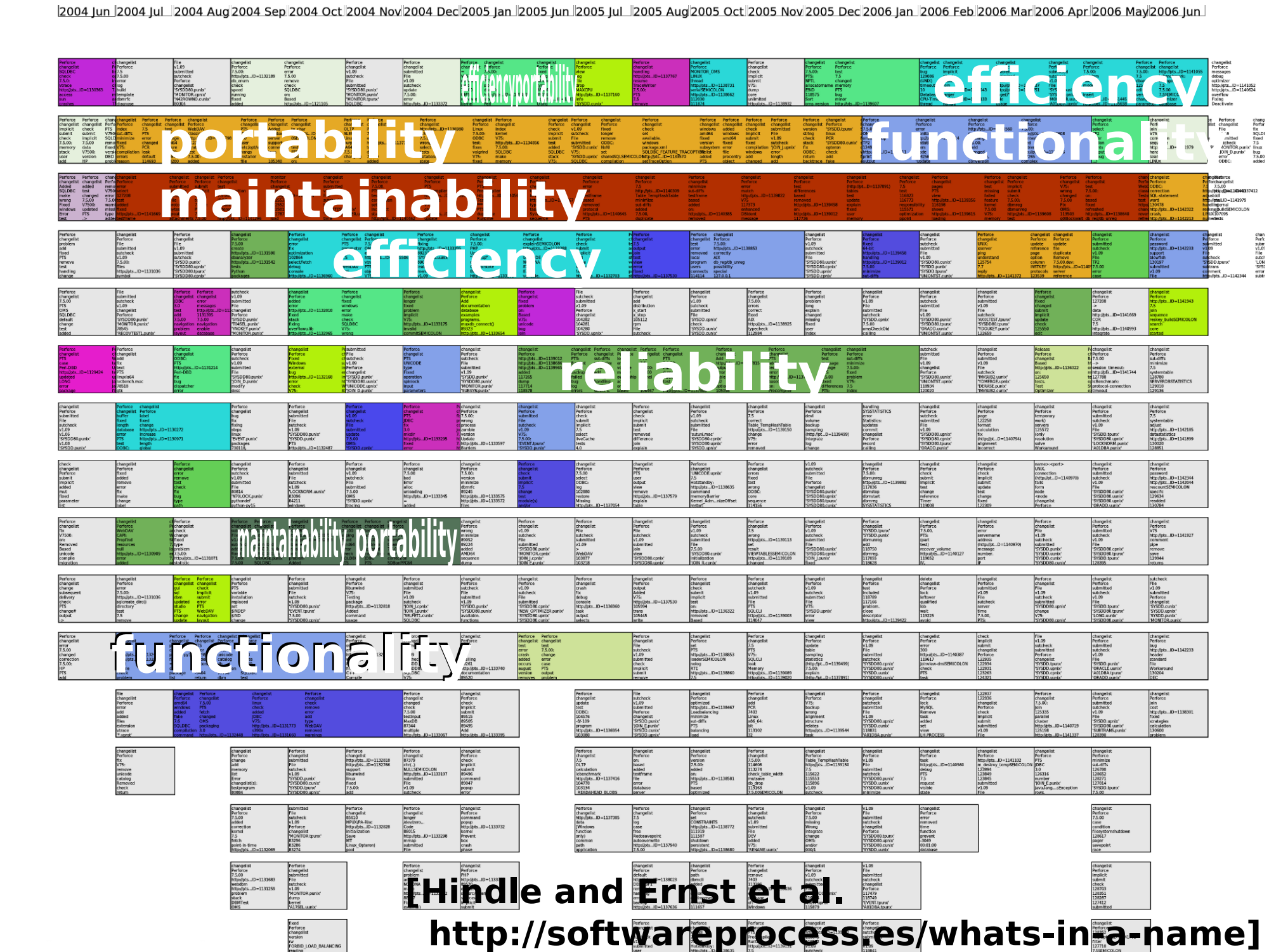
efficiency



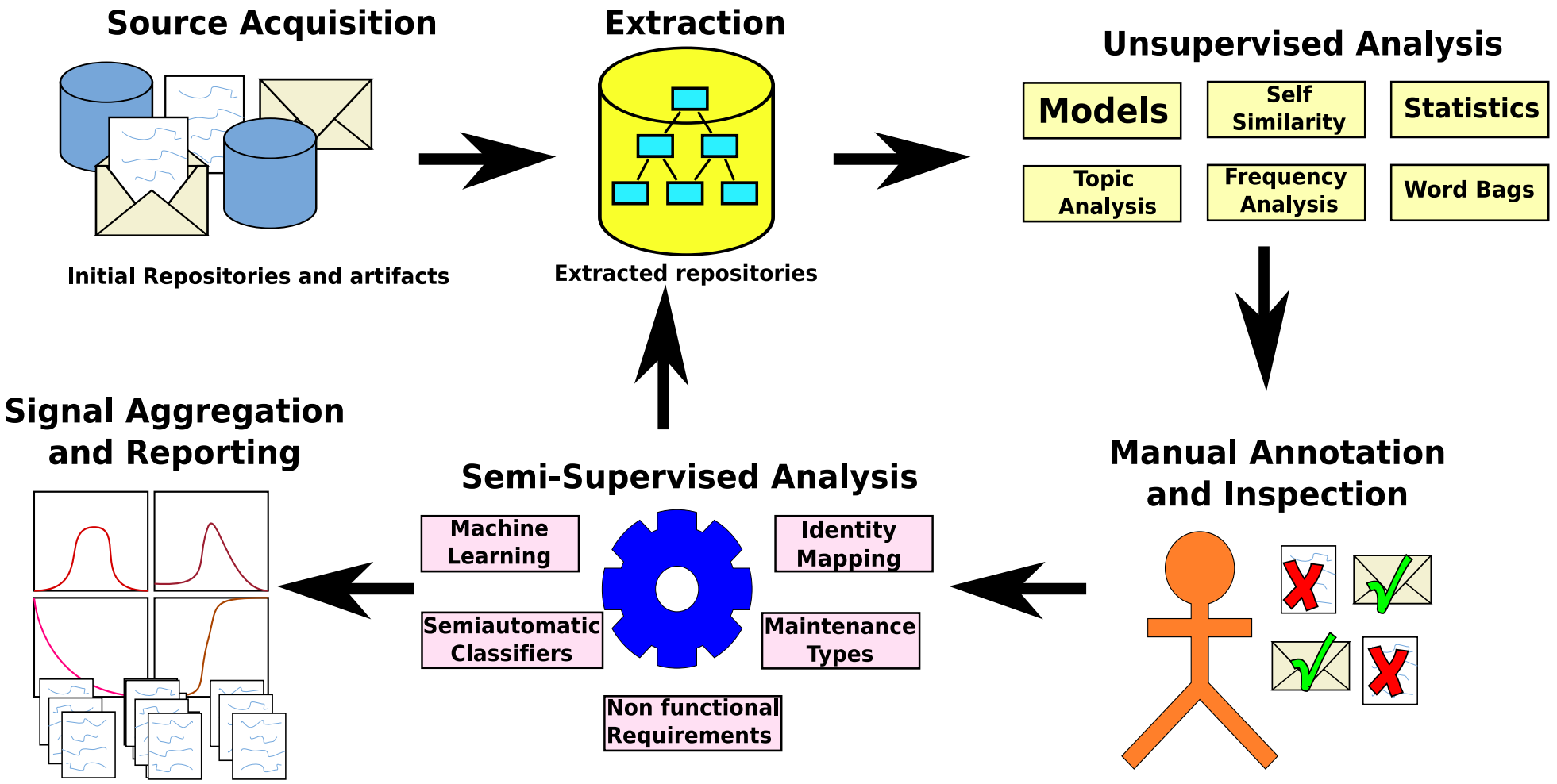


maintainability

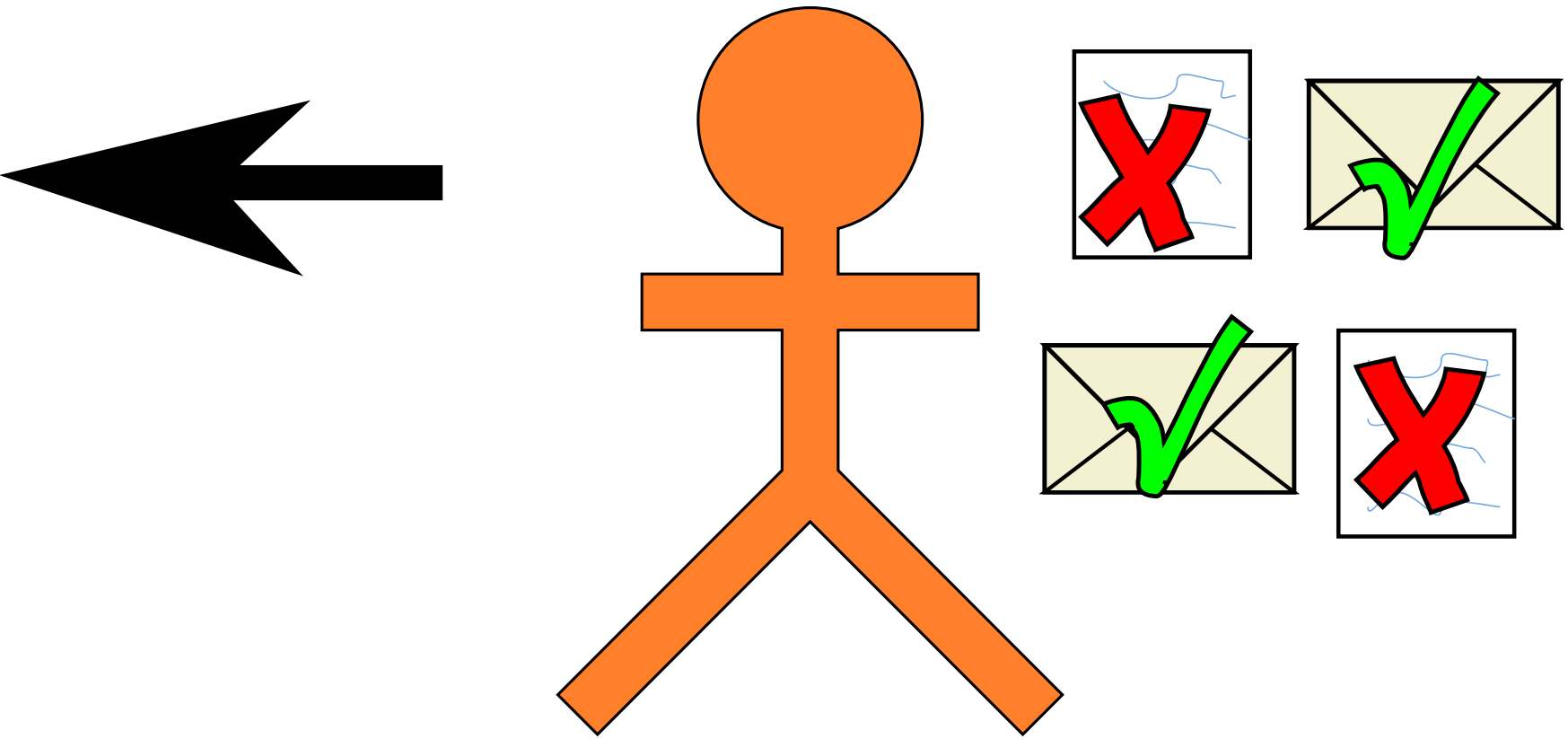
http://softwareprocess.es/whats-in-a-name/


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 Jun
efficiency
efficiency
portability
functionality
maintainability
efficiency
reliability
maintainability
portability
functionality
[Hindle and Ernst et al.
http://softwareprocess.es/whats-in-a-name]

Methodology:
Recovered Unified Process Views



Manual Annotation and Inspection

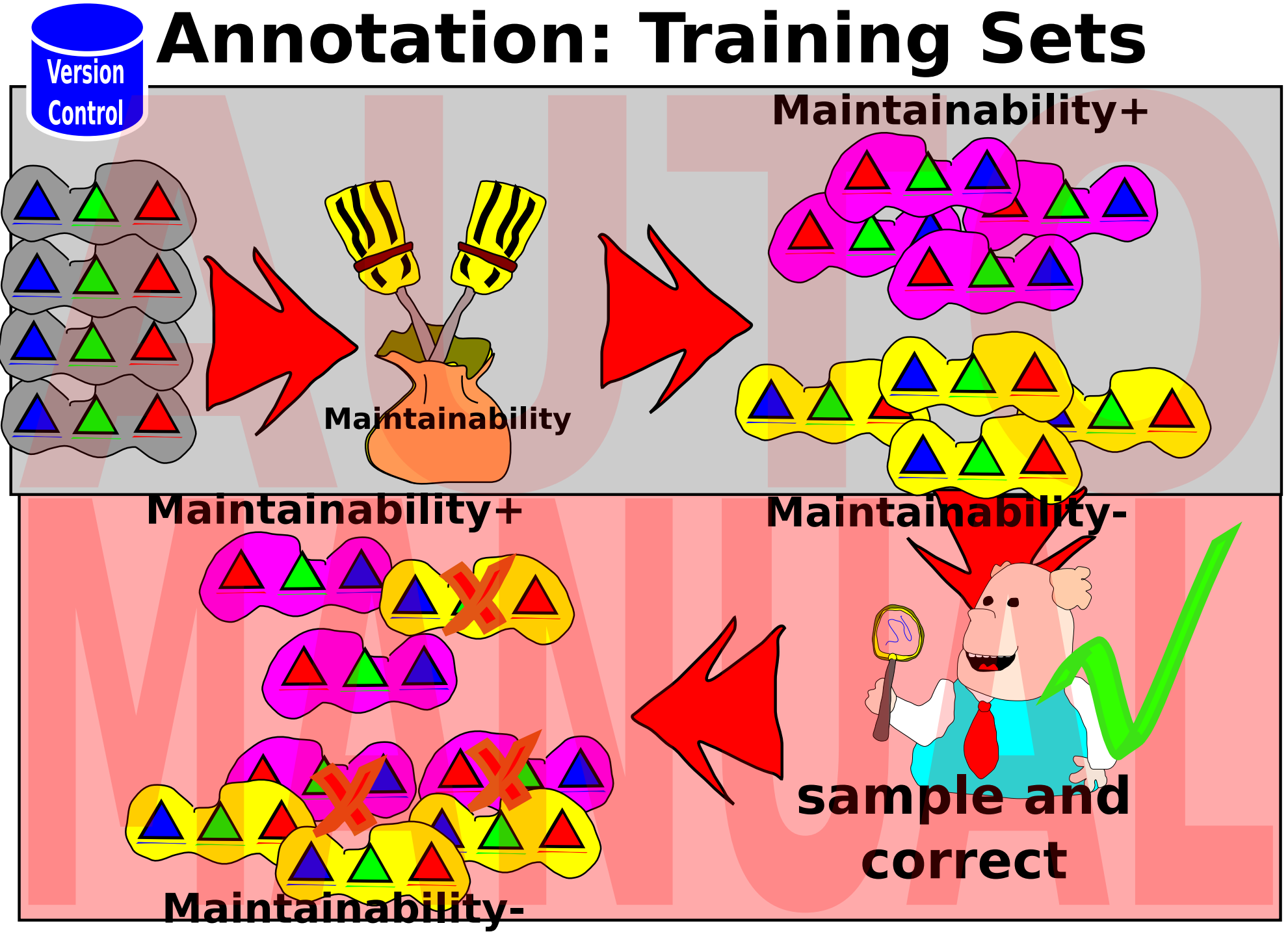


Annotation: Stop Words

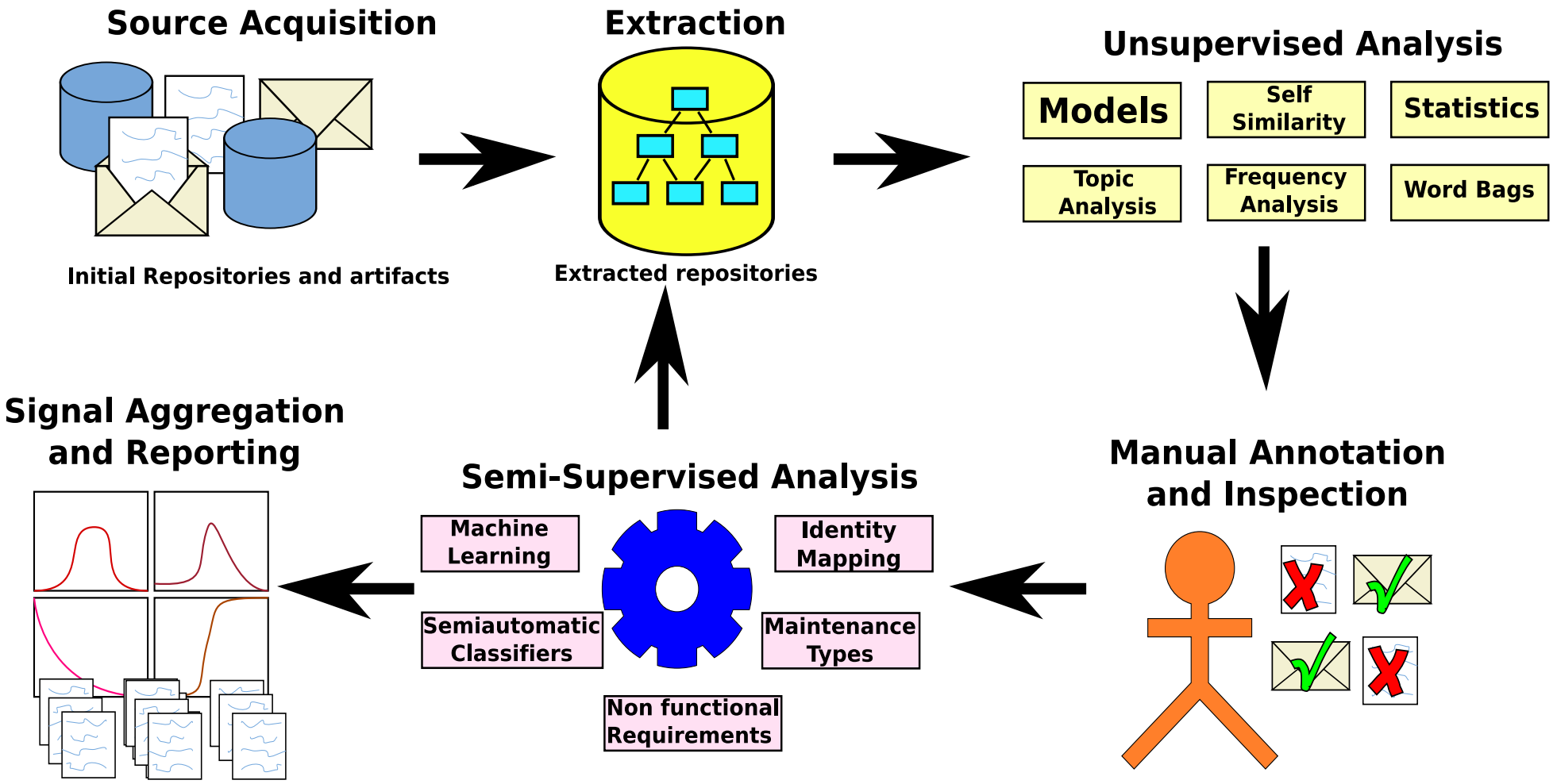
Used in topic analysis or to reduce # of features for learners.

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

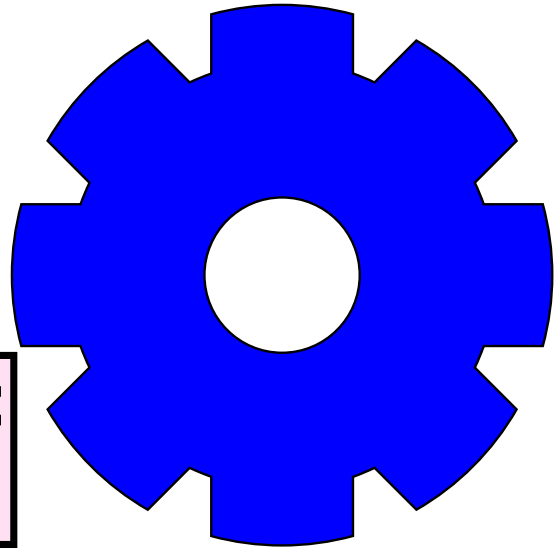
Annotation: Training Sets



Methodology:
Recovered Unified Process Views



Semi-Supervised Analysis



Machine Learning

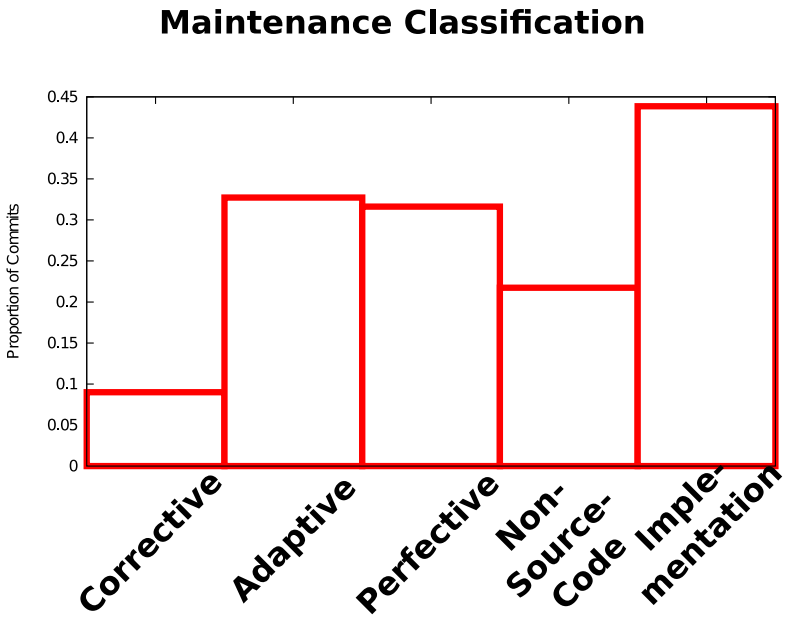
Identity Mapping

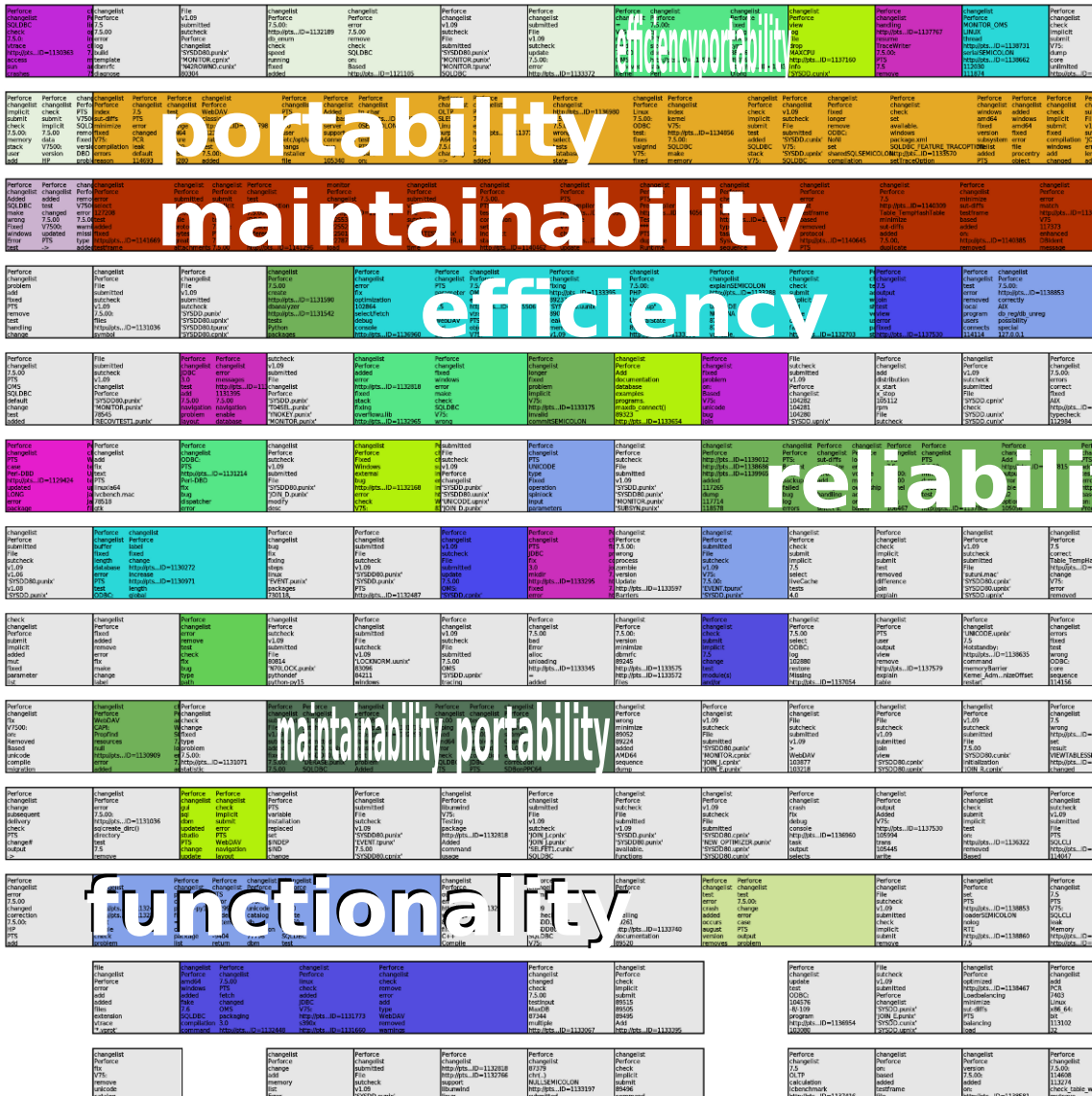
Semiautomatic Classifiers

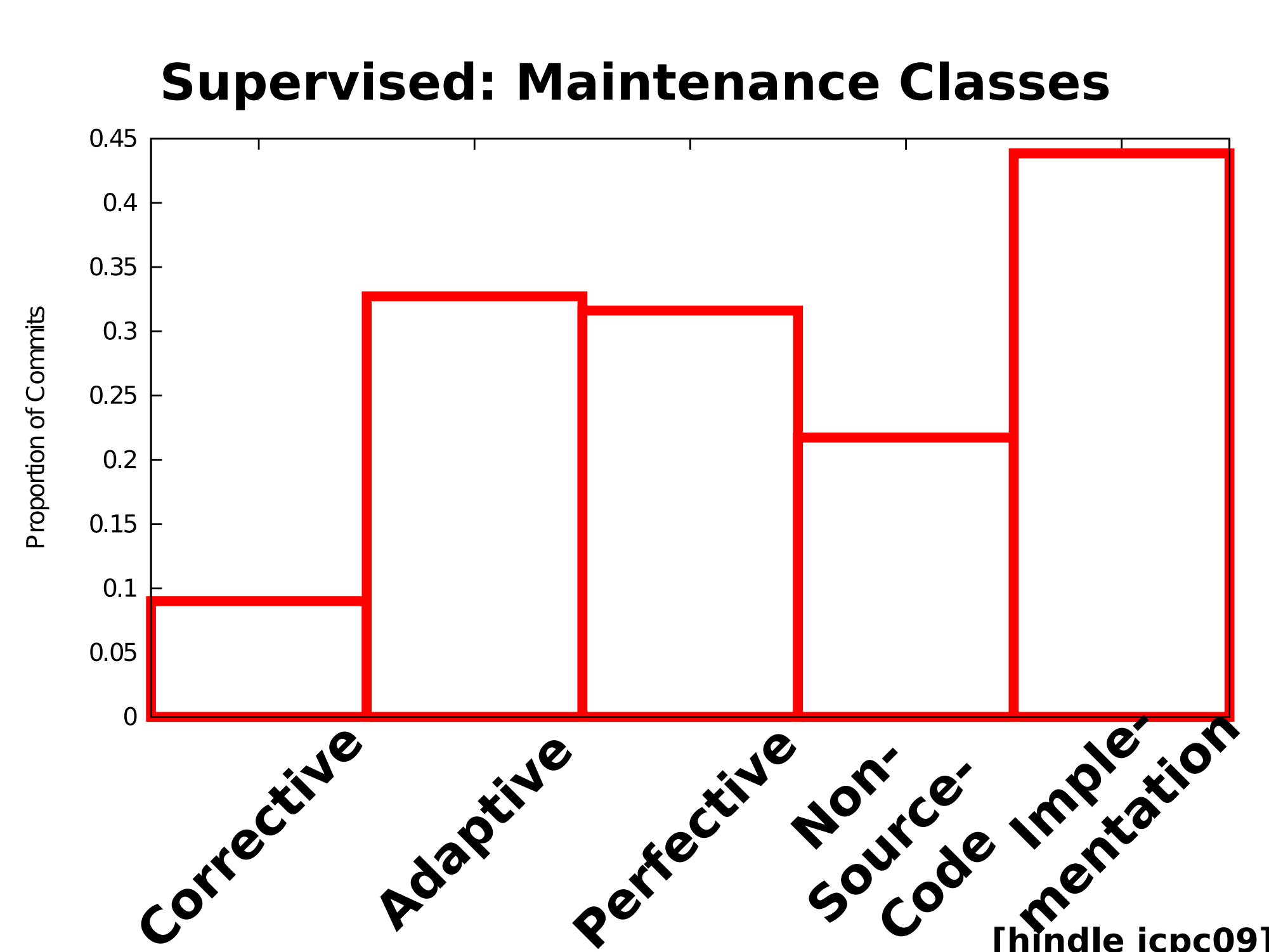
Maintenance Types

Non functional Requirements

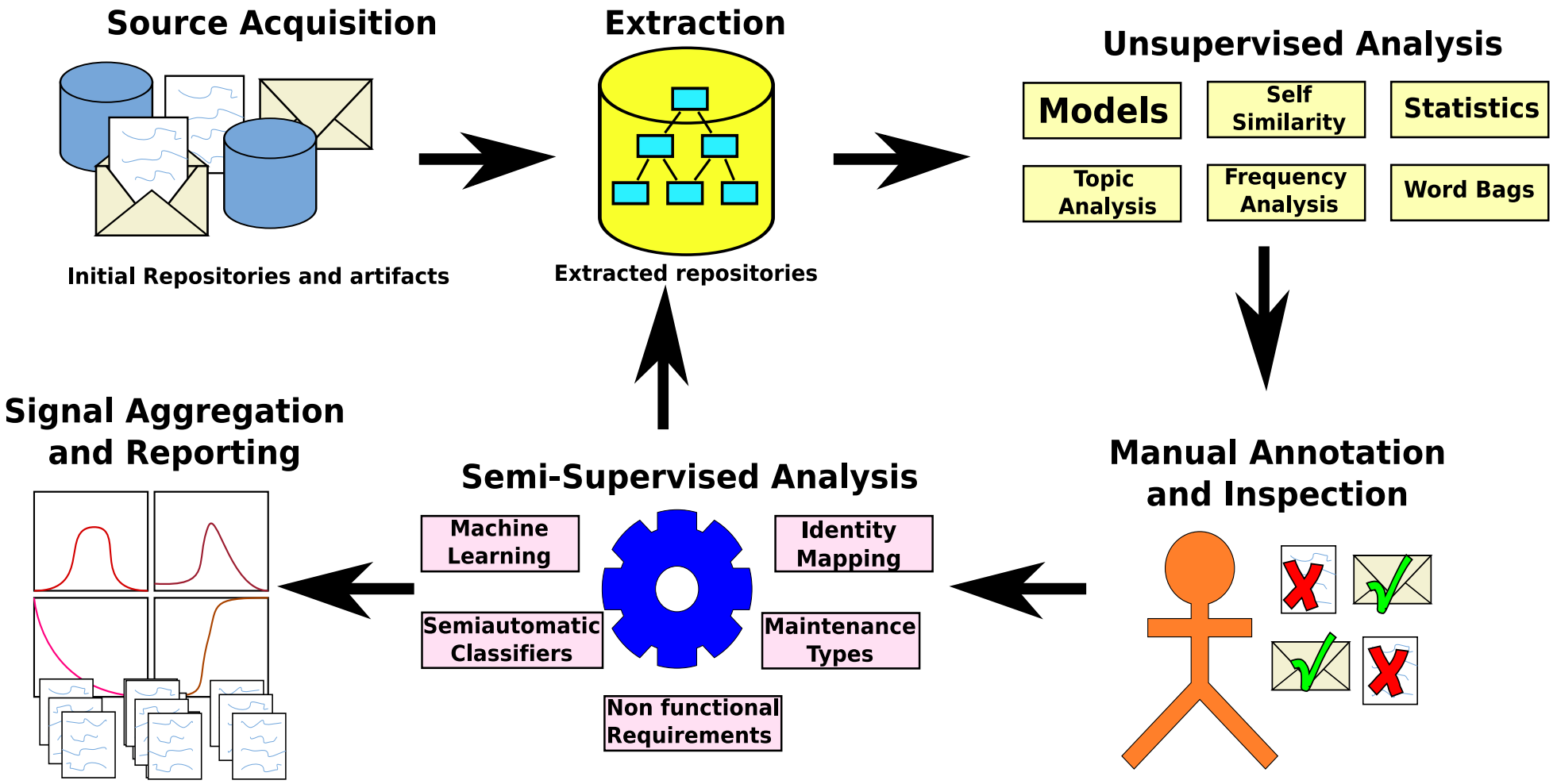
Supervised Analysis



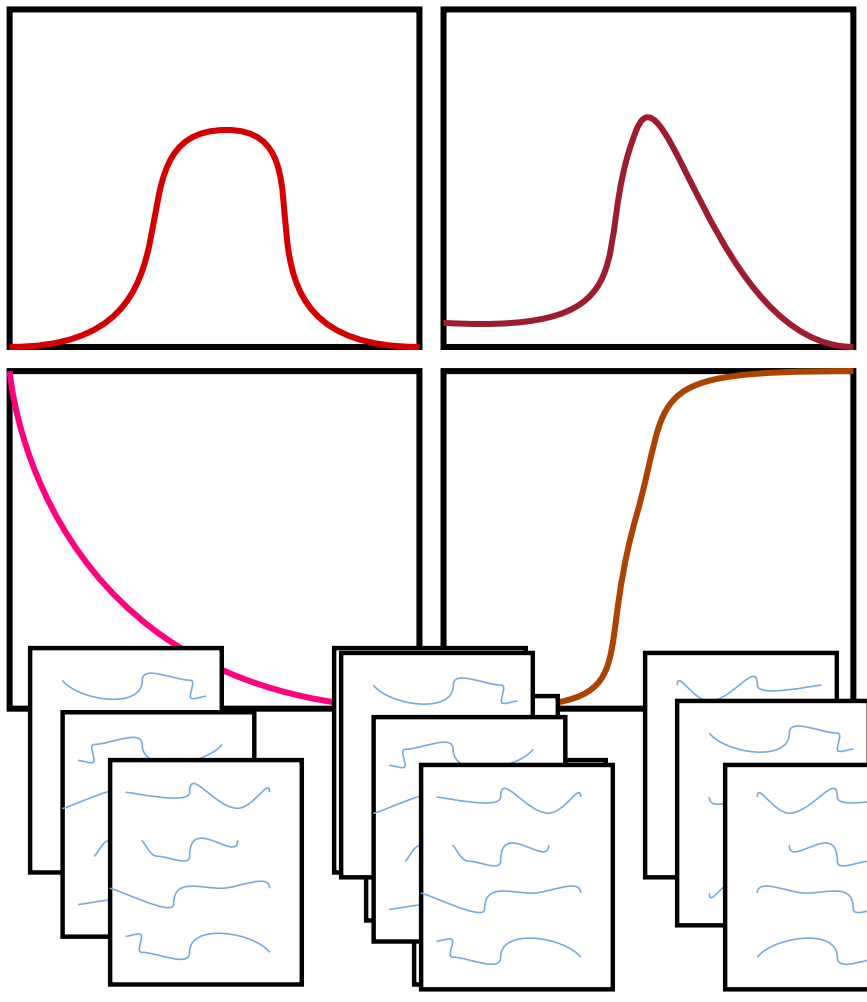



Supervised: Maintenance Classes
Proportion of Commits
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Corrective
Adaptive
Perfective
Non-Source-Code
Imple-mentation
[hindle icpc09]

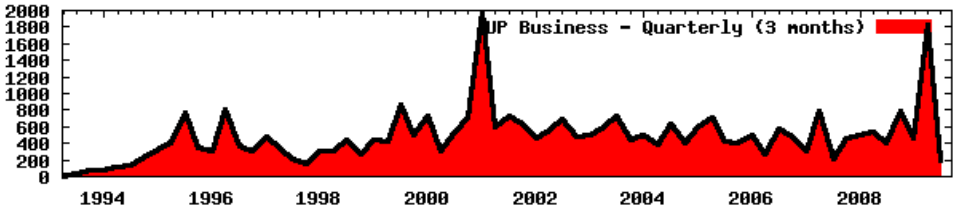
Methodology:
Recovered Unified Process Views

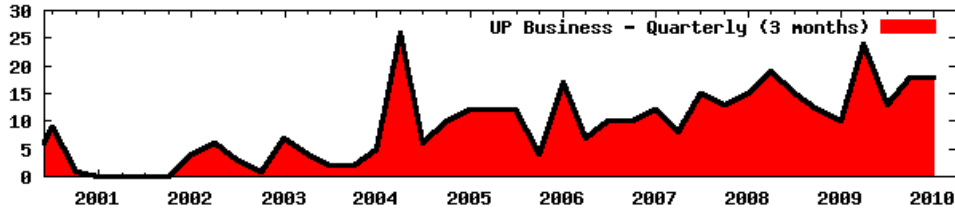


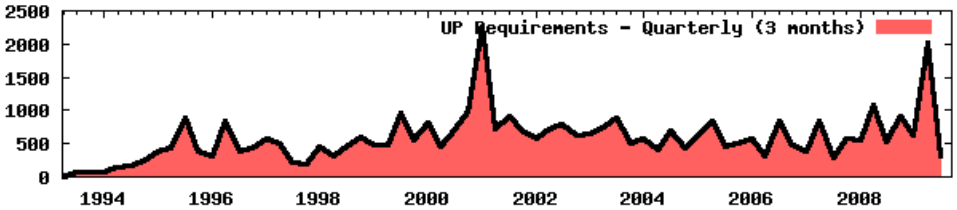
Signal Aggregation and Reporting

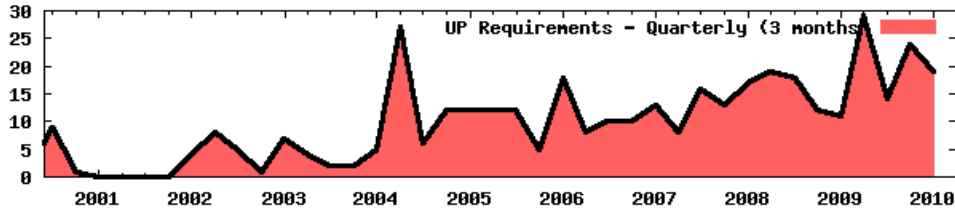


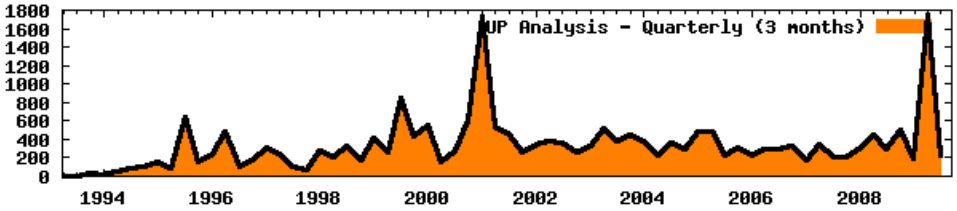
Reporting

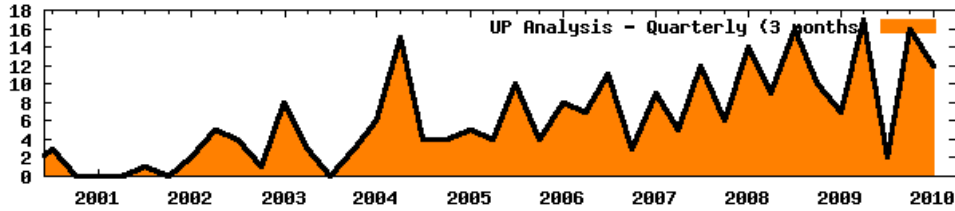

UP Business - Quarterly (3 months)
2000
1800
1600
1400
1200
1000
800
600
400
200
0
1994
1996
1998
2000
2002
2004
2006
2008

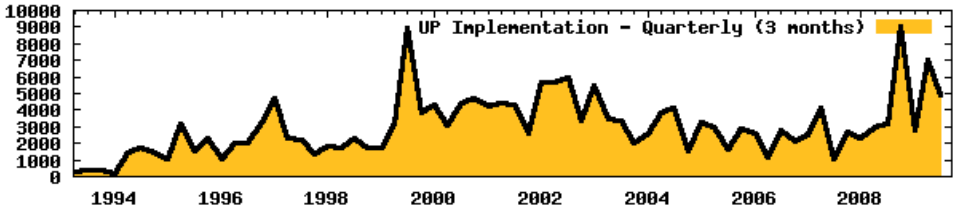

UP Business - Quarterly (3 months)
30
25
20
15
10
5
0
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010

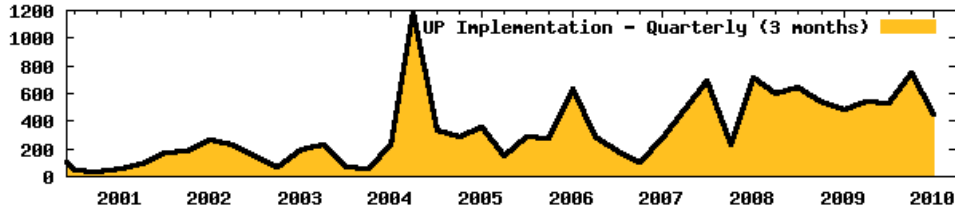

UP Requirements - Quarterly (3 months)
2500
2000
1500
1000
500
0
1994
1996
1998
2000
2002
2004
2006
2008

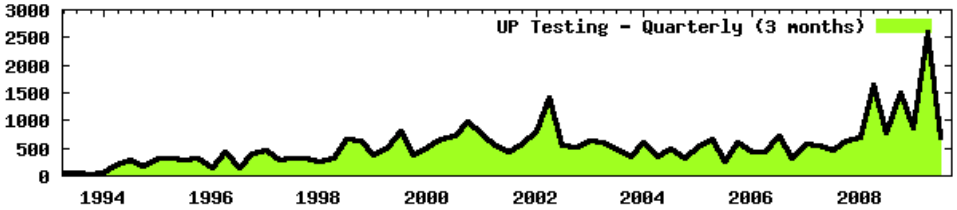

UP Requirements - Quarterly (3 months)
30
25
20
15
10
5
0
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010

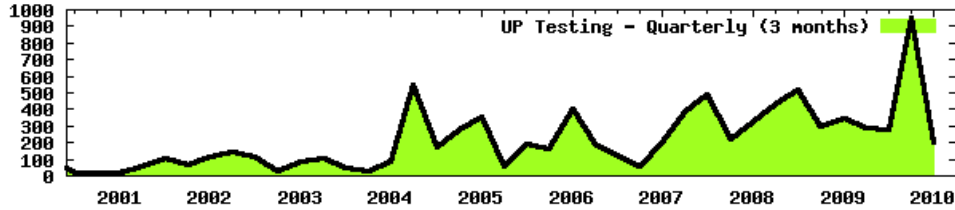

UP Analysis - Quarterly (3 months)
1800
1600
1400
1200
1000
800
600
400
200
0
1994
1996
1998
2000
2002
2004
2006
2008

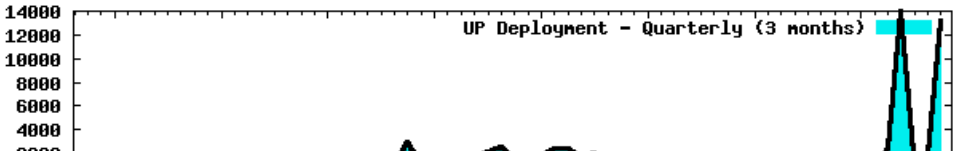

UP Analysis - Quarterly (3 months)
18
16
14
12
10
8
6
4
2
0
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010

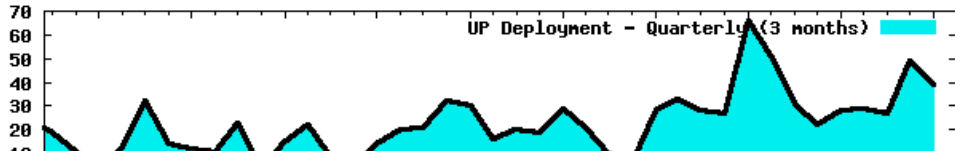

UP Implementation - Quarterly (3 months)
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
1994
1996
1998
2000
2002
2004
2006
2008


UP Implementation - Quarterly (3 months)
1200
1000
800
600
400
200
0
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010

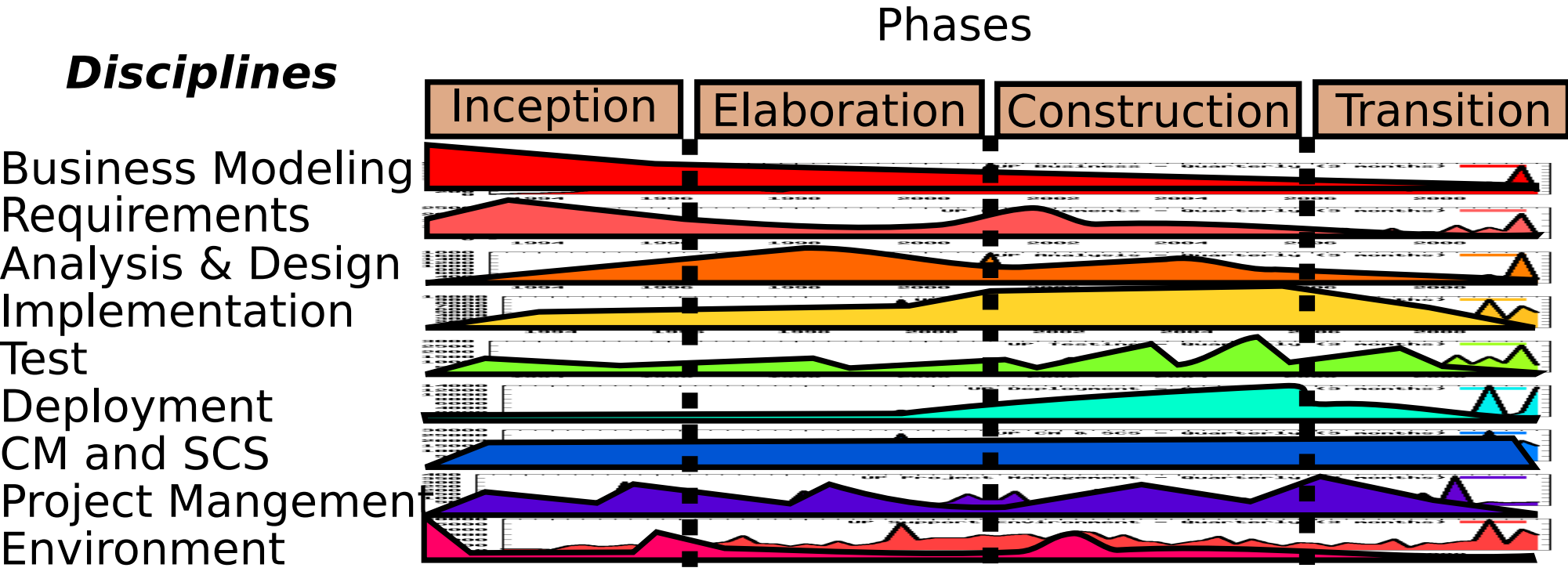

UP Testing - Quarterly (3 months)
3000
2500
2000
1500
1000
500
0
1994
1996
1998
2000
2002
2004
2006
2008


UP Testing - Quarterly (3 months)
1000
900
800
700
600
500
400
300
200
100
0
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010


UP Deployment - Quarterly (3 months)
14000
12000
10000
8000
6000
4000

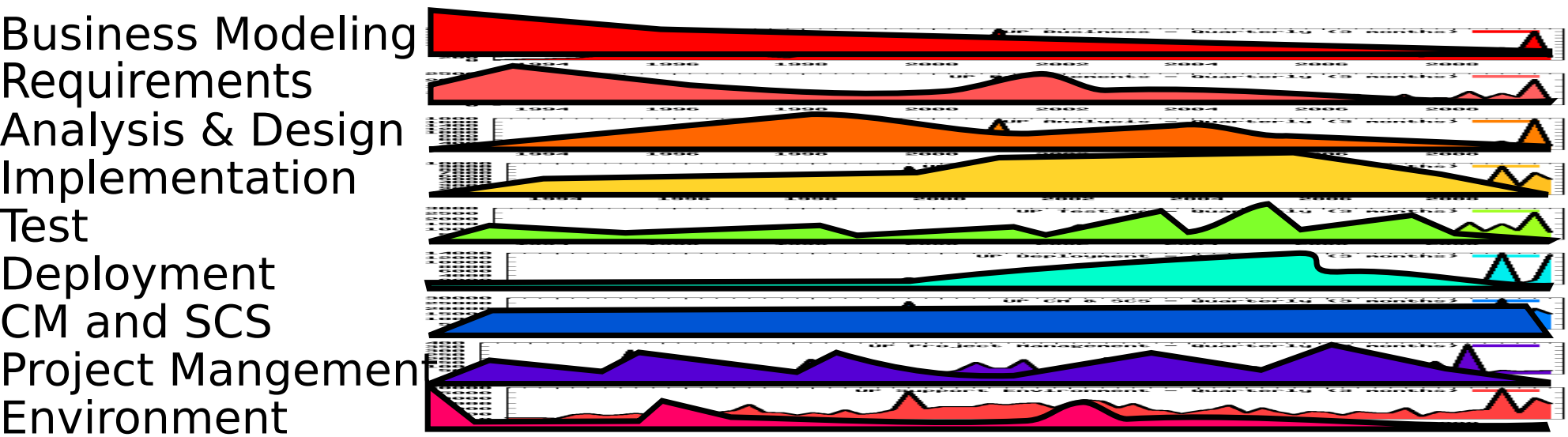

UP Deployment - Quarterly (3 months)
70
60
50
40
30
20

Recovered Unified Process Views



Recovered Unified Process Views

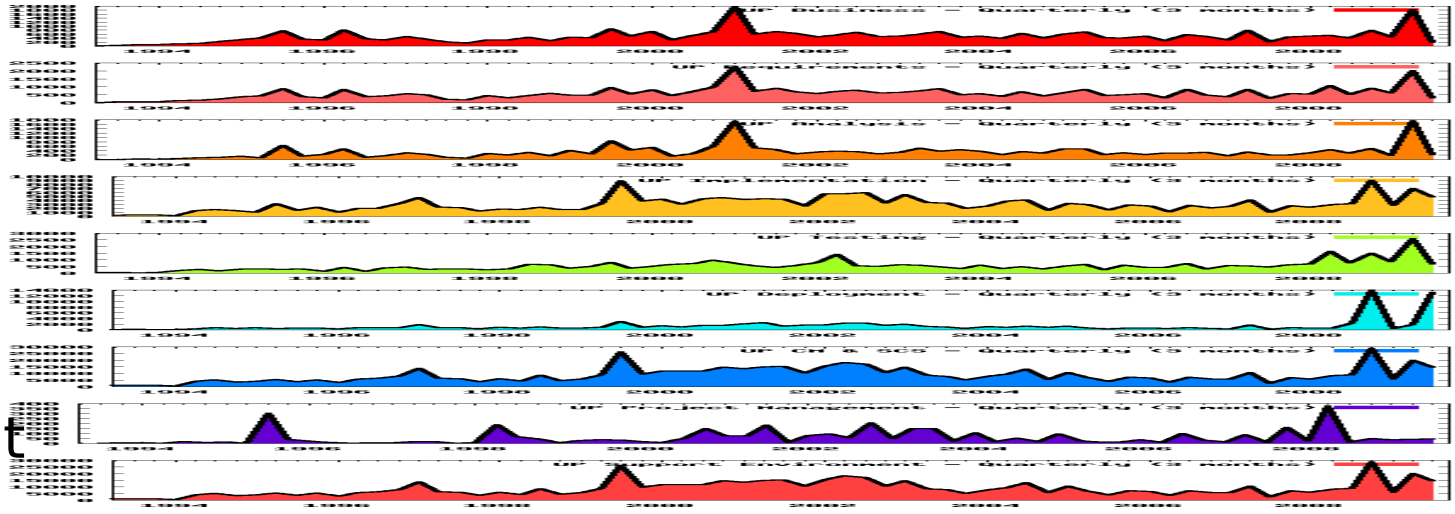
Disciplines



Recovered Unified Process Views

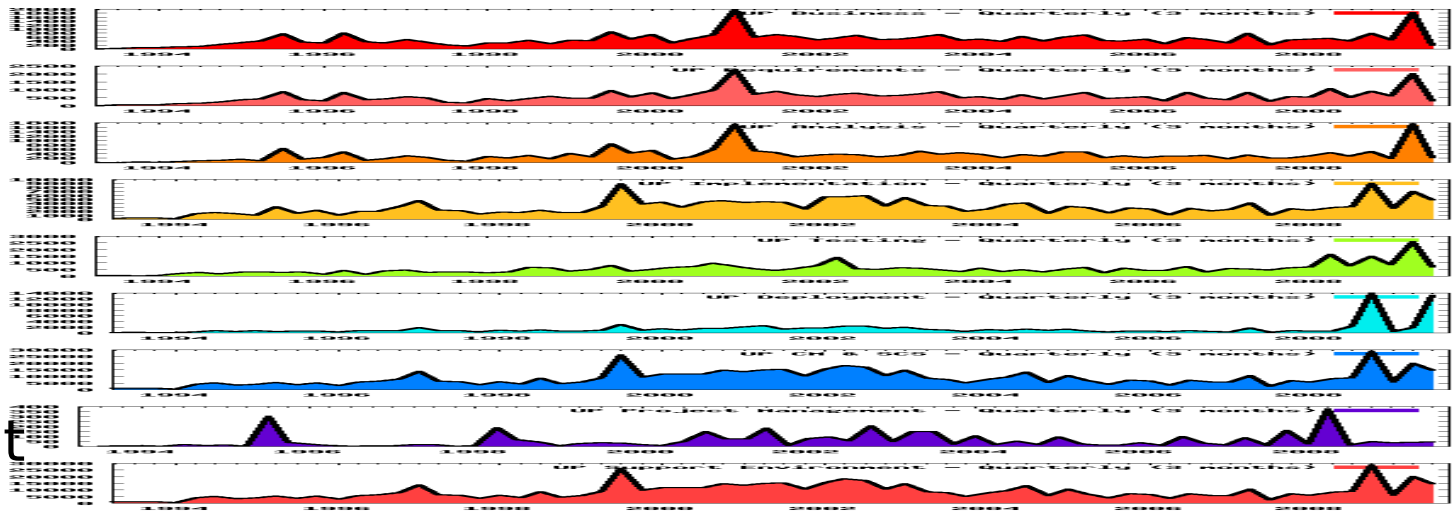
Disciplines

Business Modeling
Requirements
Analysis & Design
Implementation
Test
Deployment
CM and SCS
Project Mangement
Environment



Recovered Unified Process Views

Business Modeling
Requirements
Analysis & Design
Implementation
Test
Deployment
CM and SCS
Project Mangement
Environment



Recovered Unified Process Views

Business Modeling

Requirements

Analysis & Design

Implementation

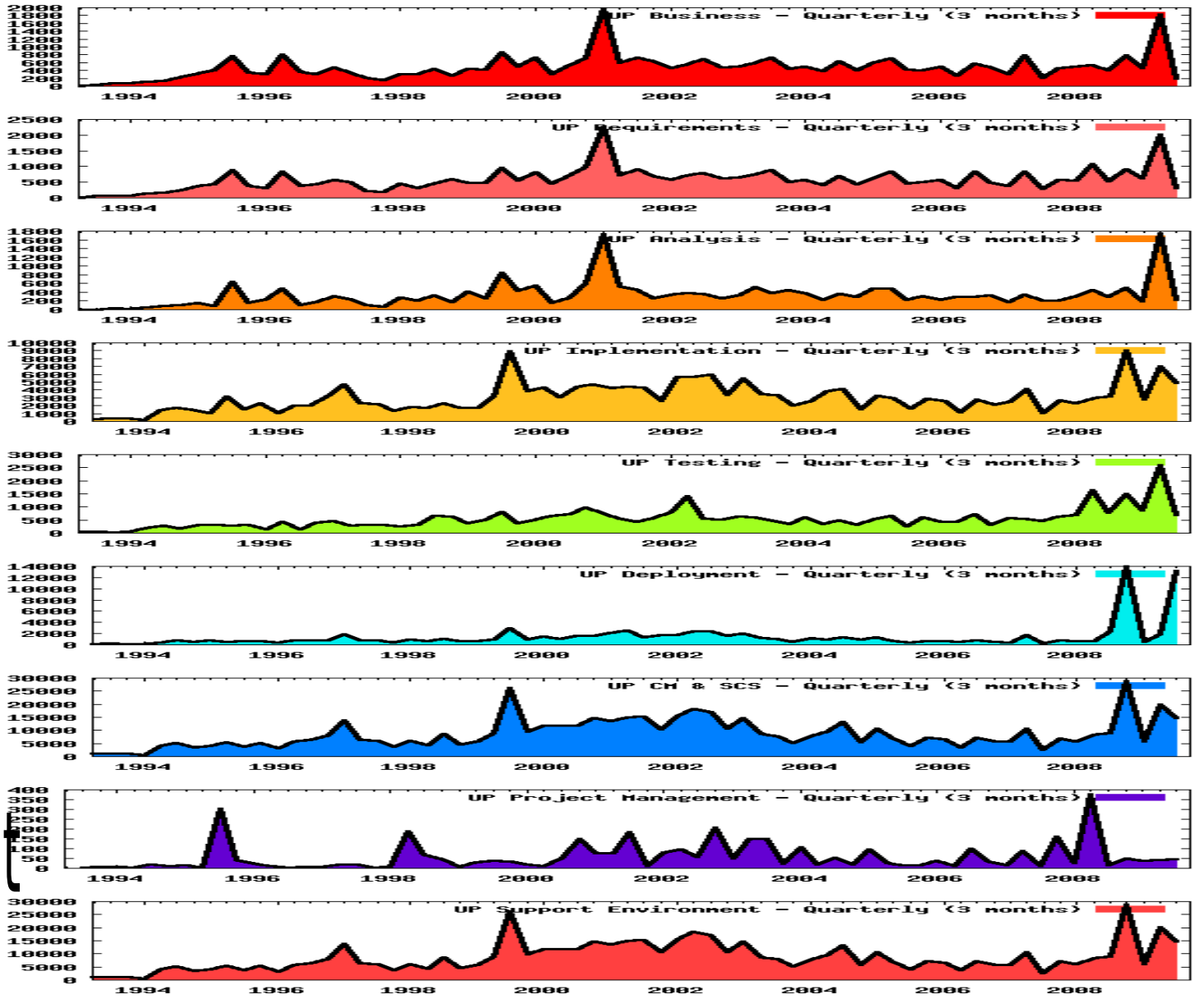
Test

Deployment

CM and SCS

Project Mangement

Environment



Recovered Unified Process Views

Business Modeling

Requirements

Analysis & Design

Implementation

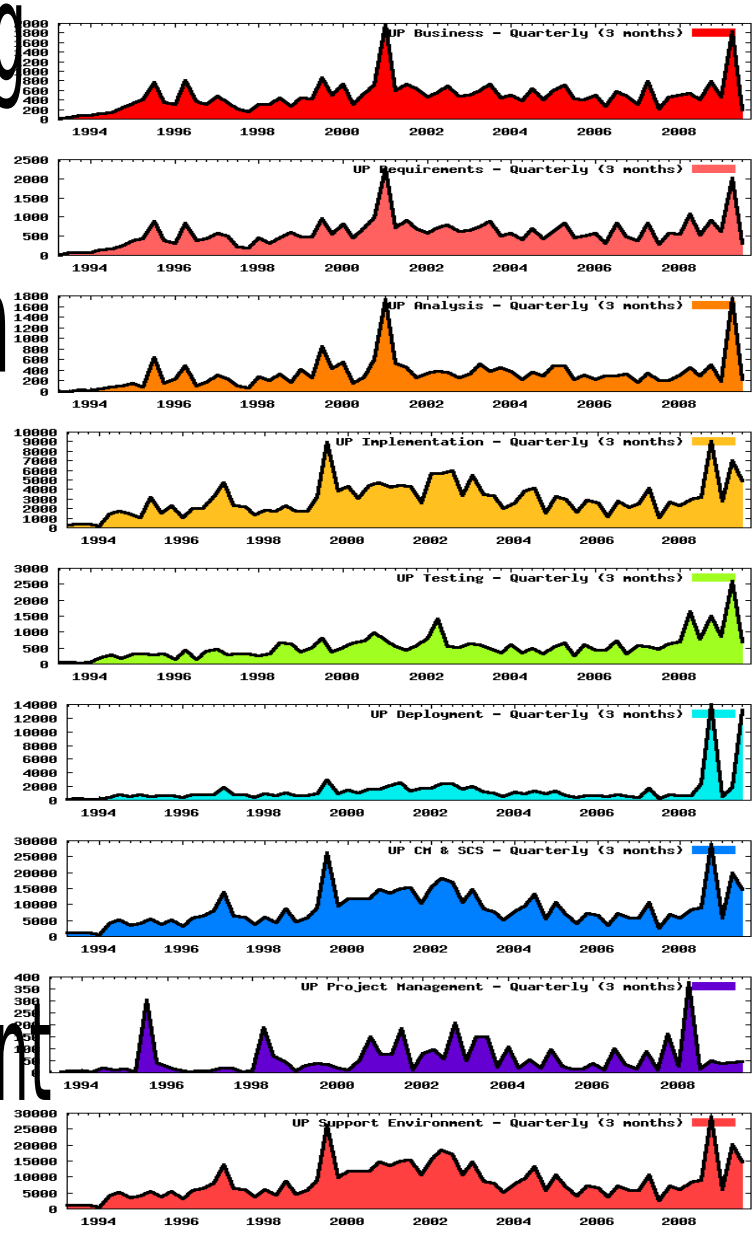
Test

Deployment

CM and SCS

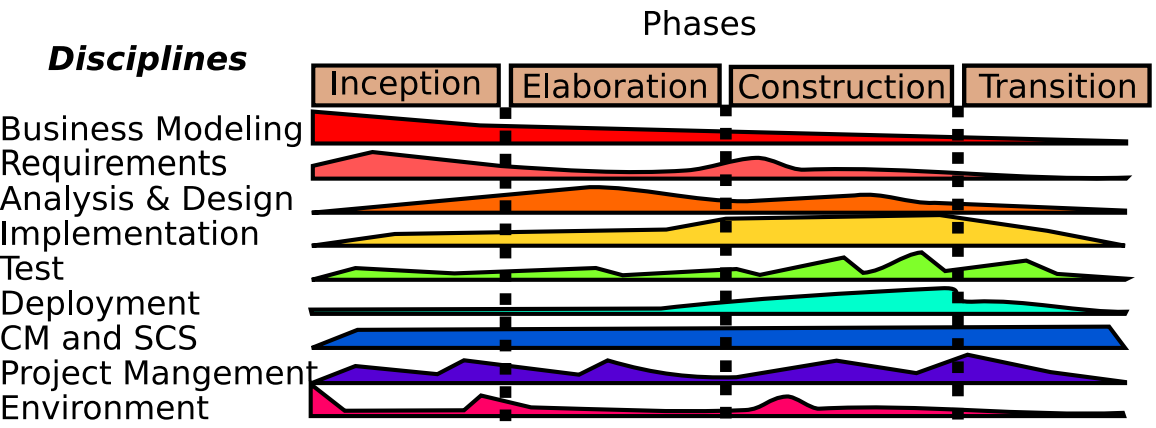
Project Mangement

Environment

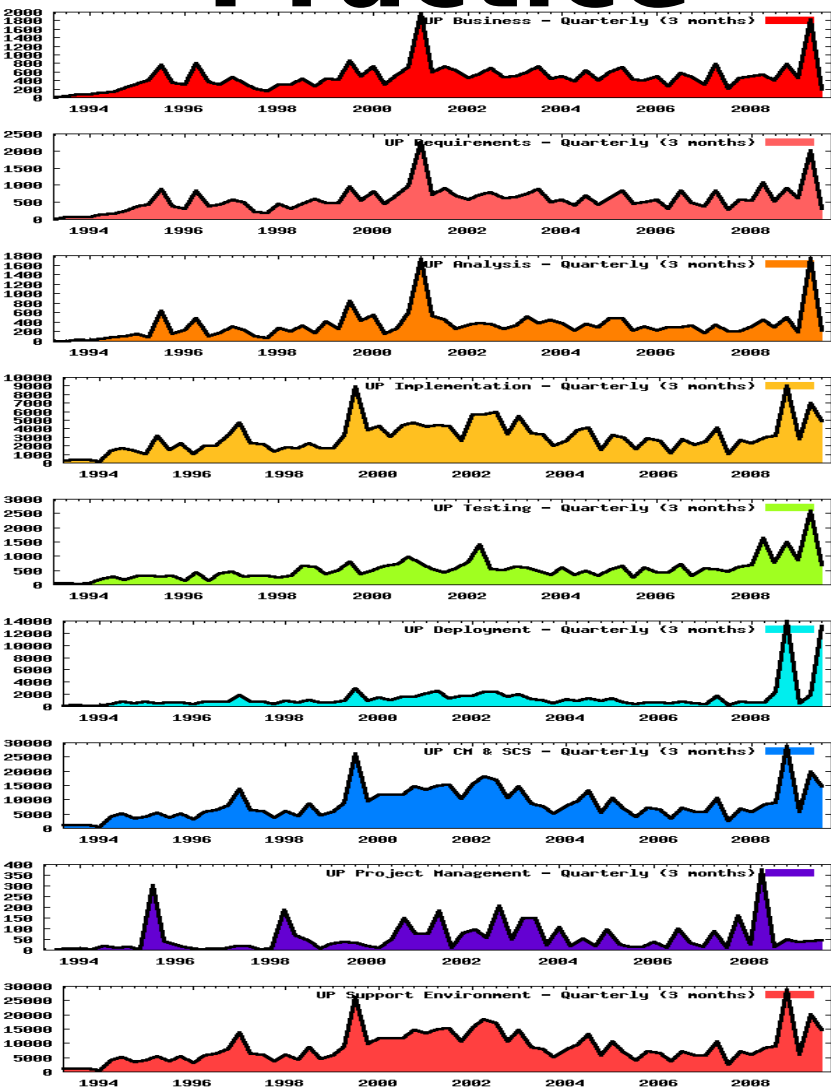


Recovered Unified Process Views

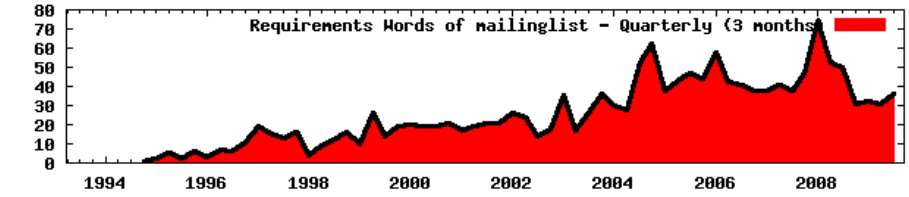
Theory

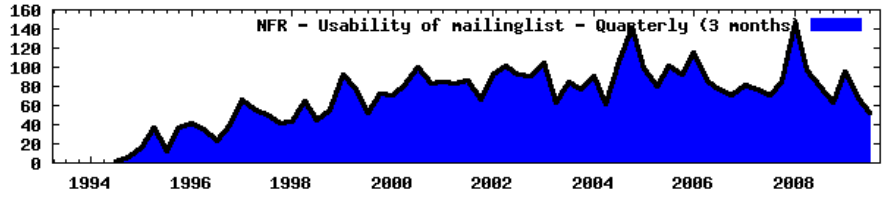


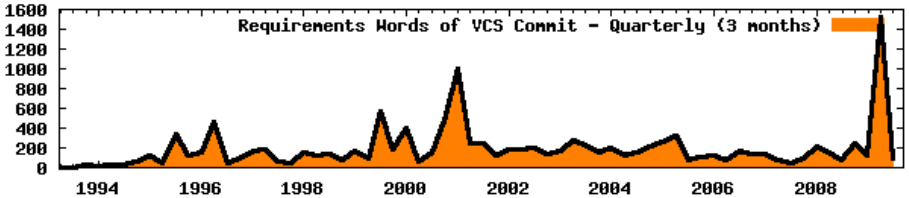
Practice



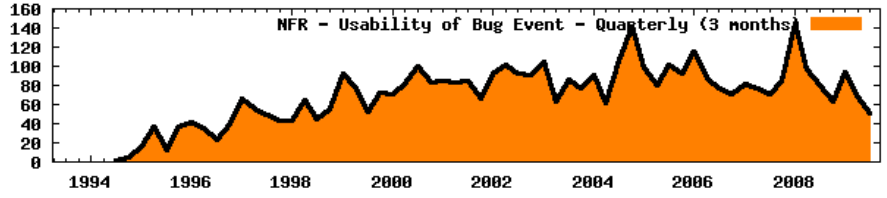
UP Business Modelling Signal

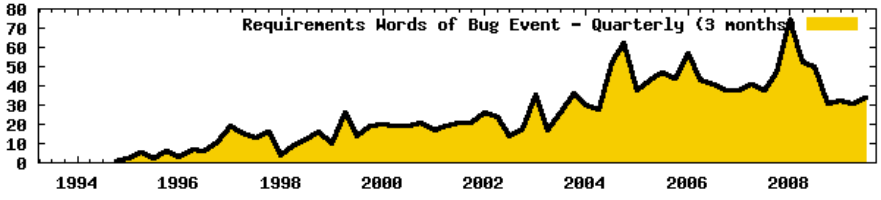


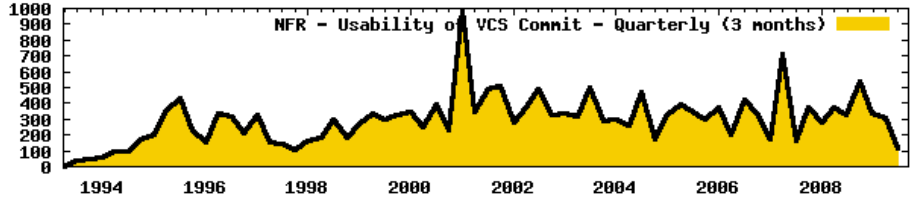




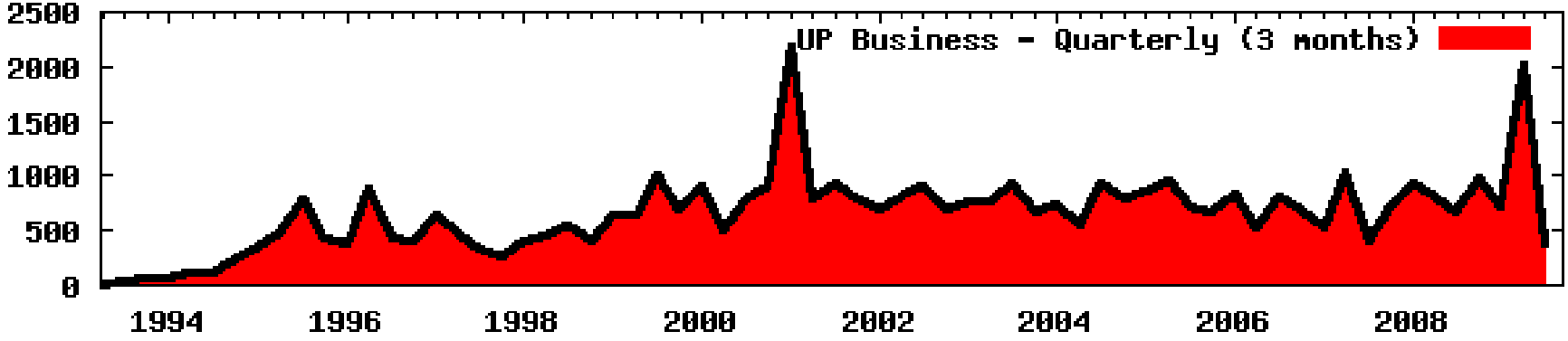
+



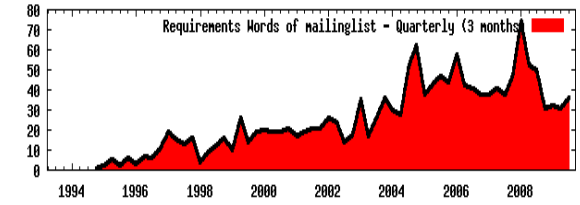


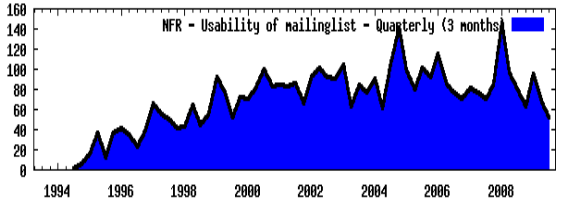


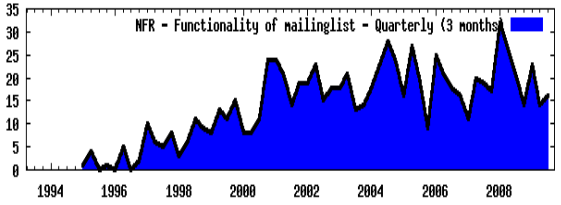
=

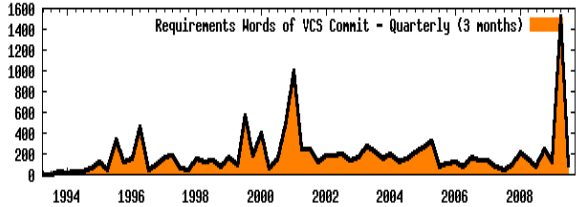


UP Requirements Signal

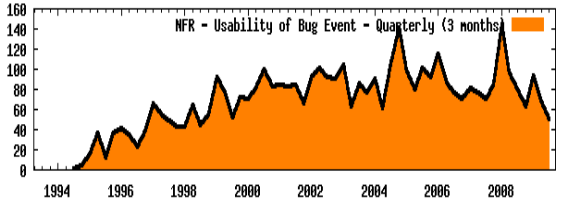




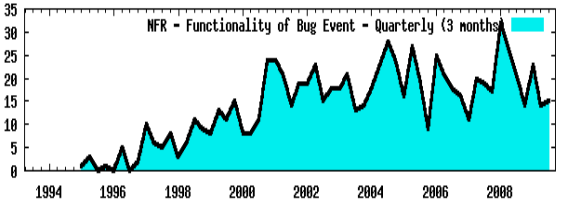


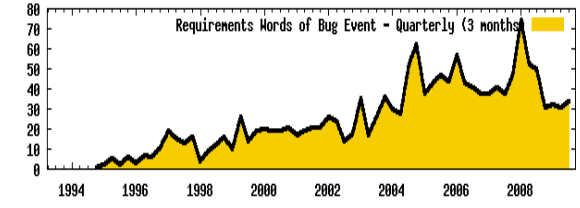


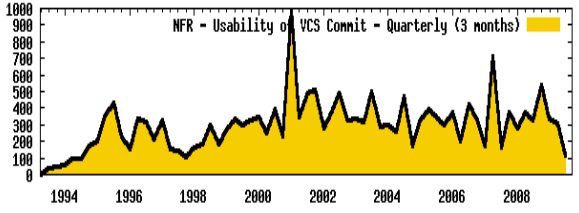
+

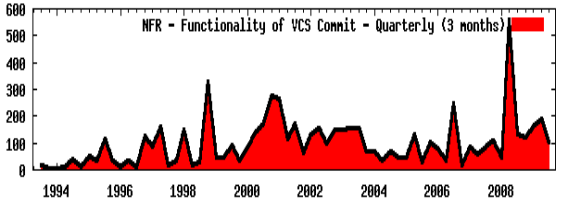


+

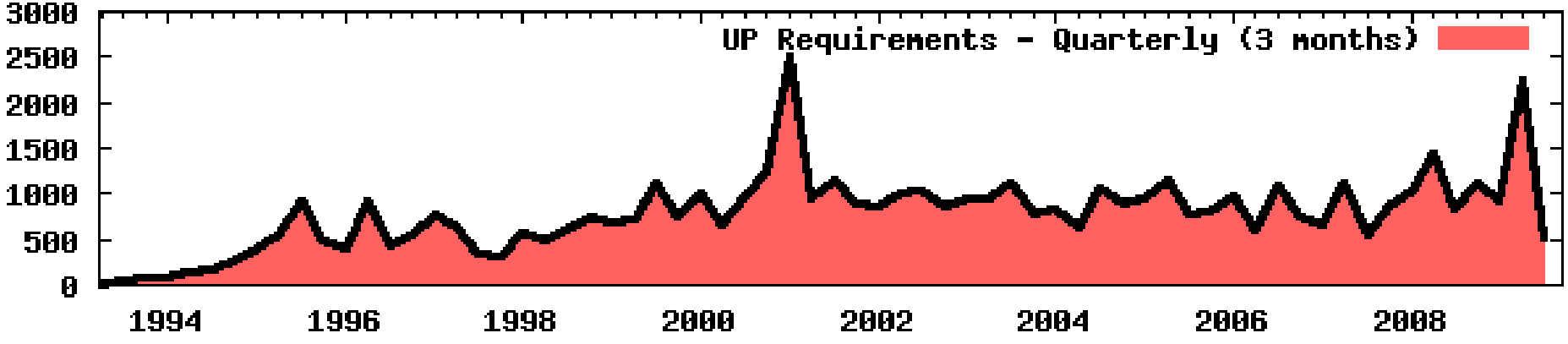




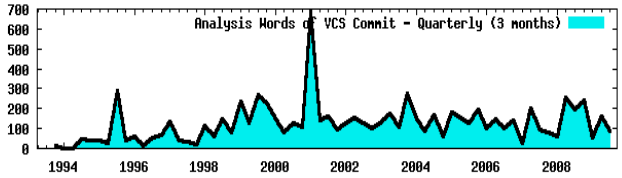


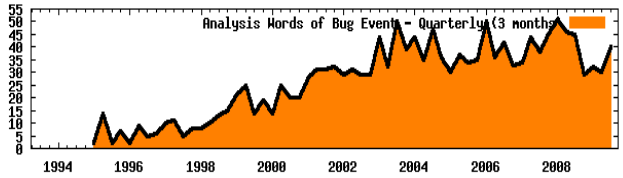


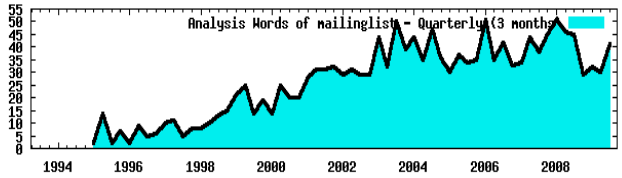
=

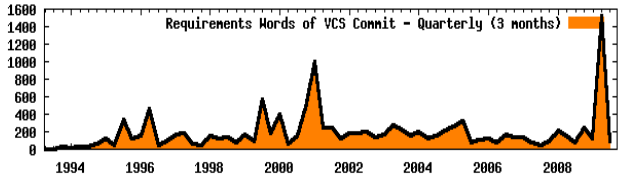


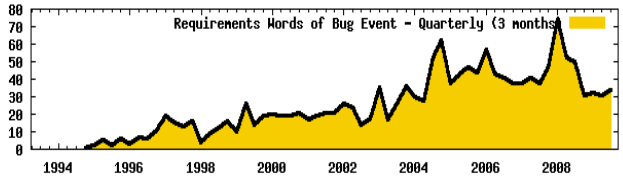
UP Analysis Signal

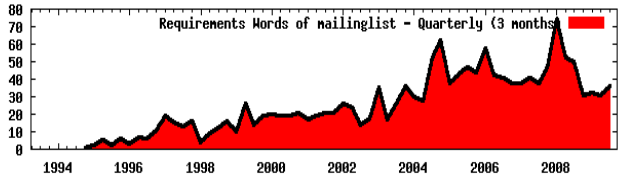


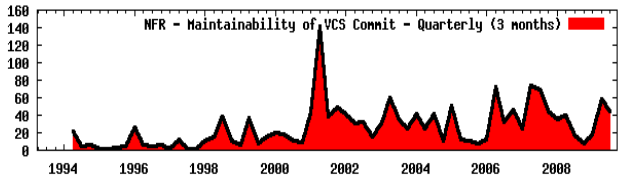


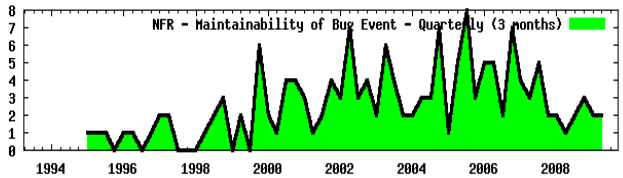


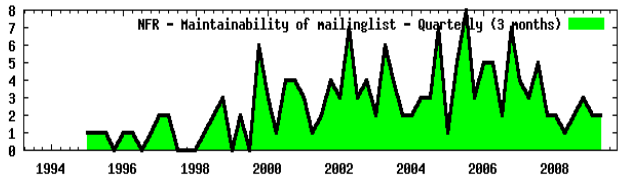




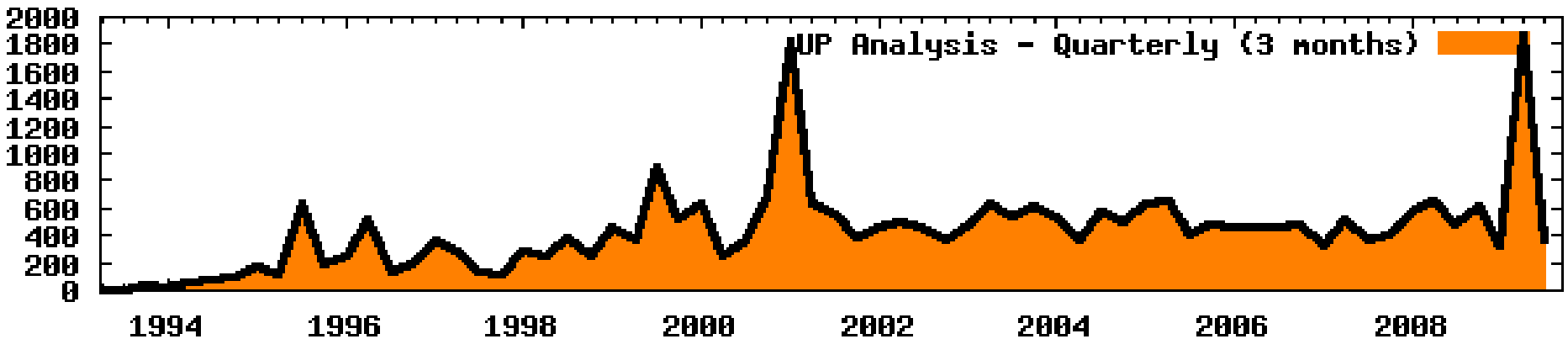




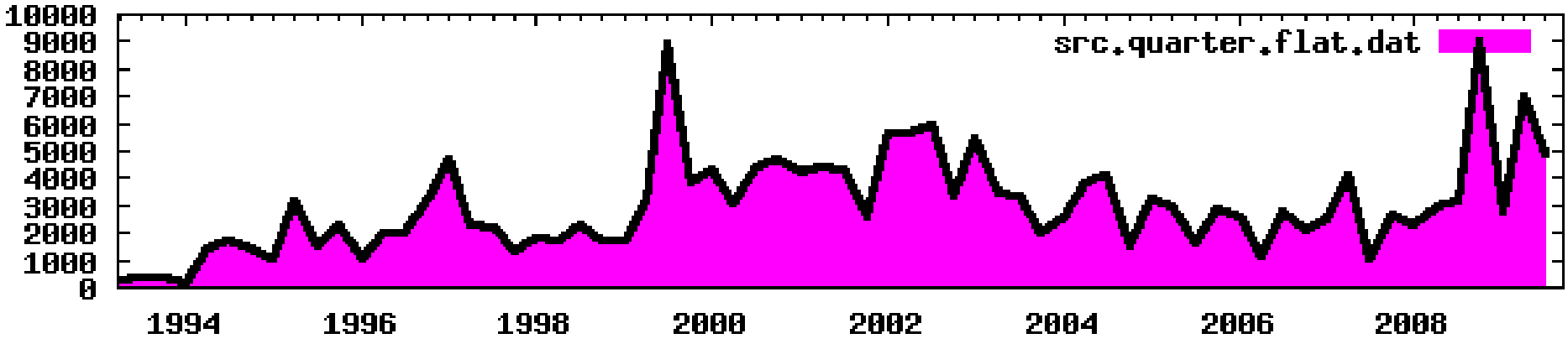




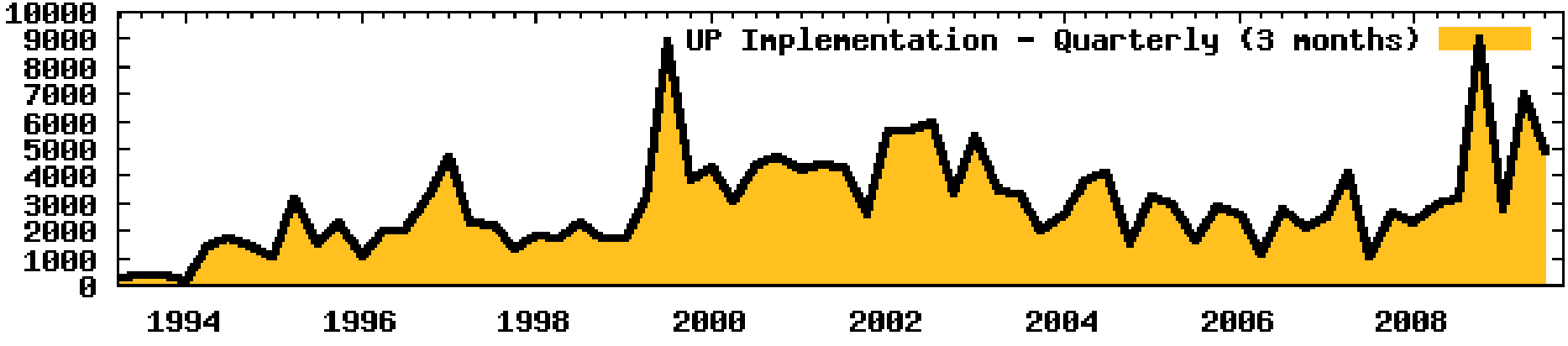
=



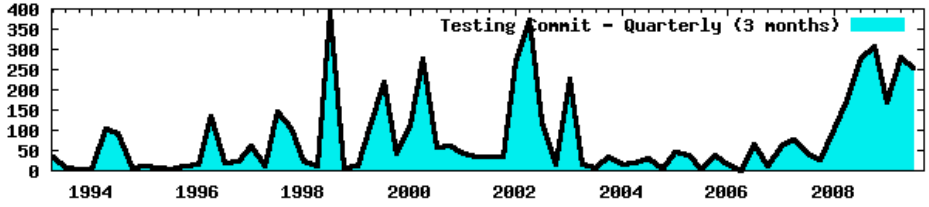
UP Implementation Signal

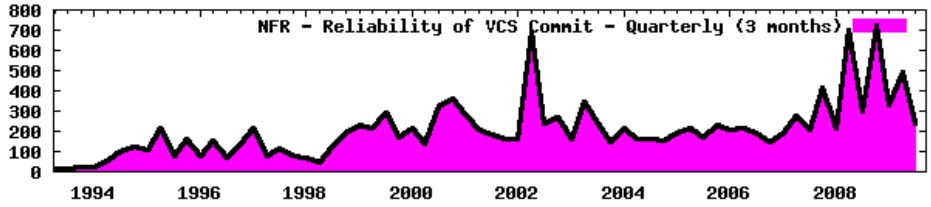


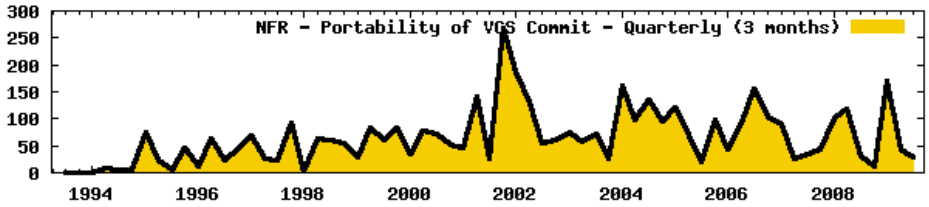
=

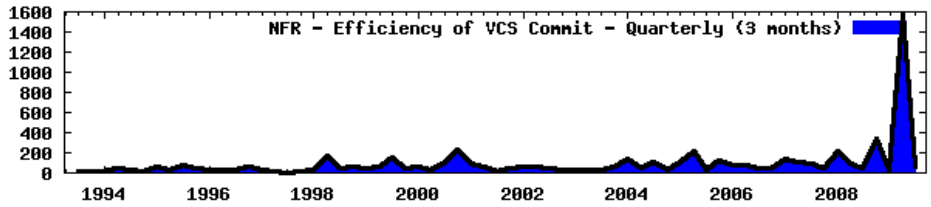


UP Testing Signal

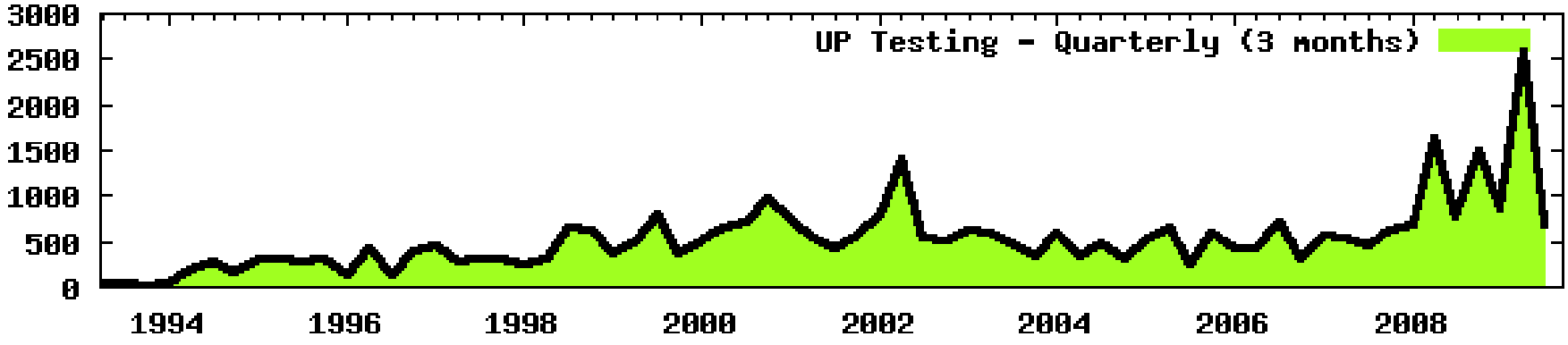




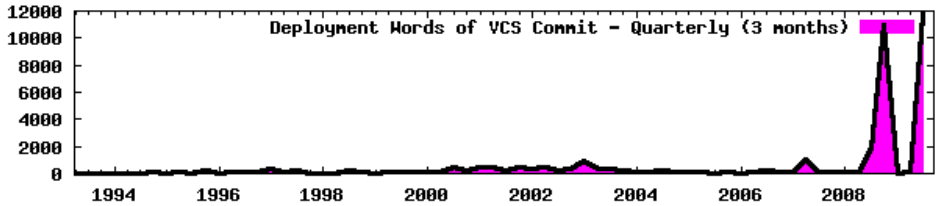


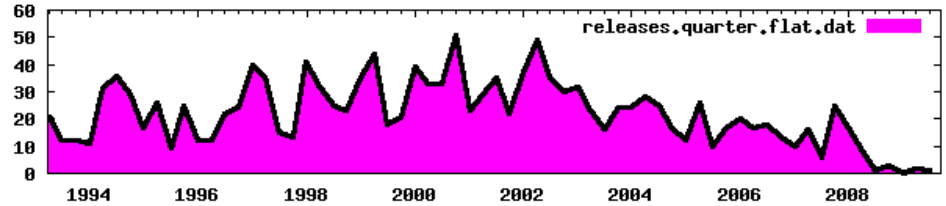


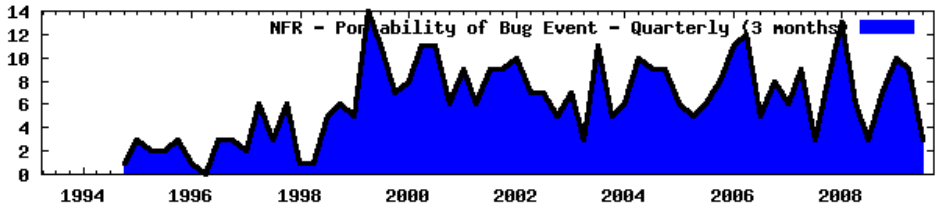
=

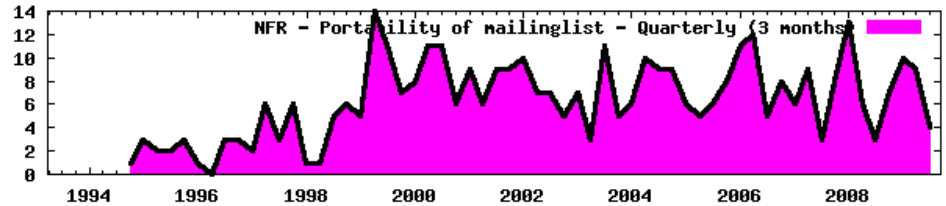


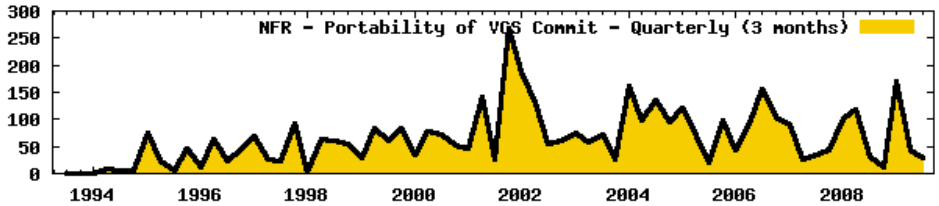
UP Deployment Signal

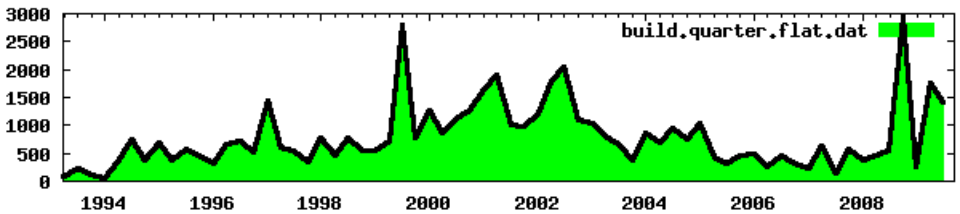


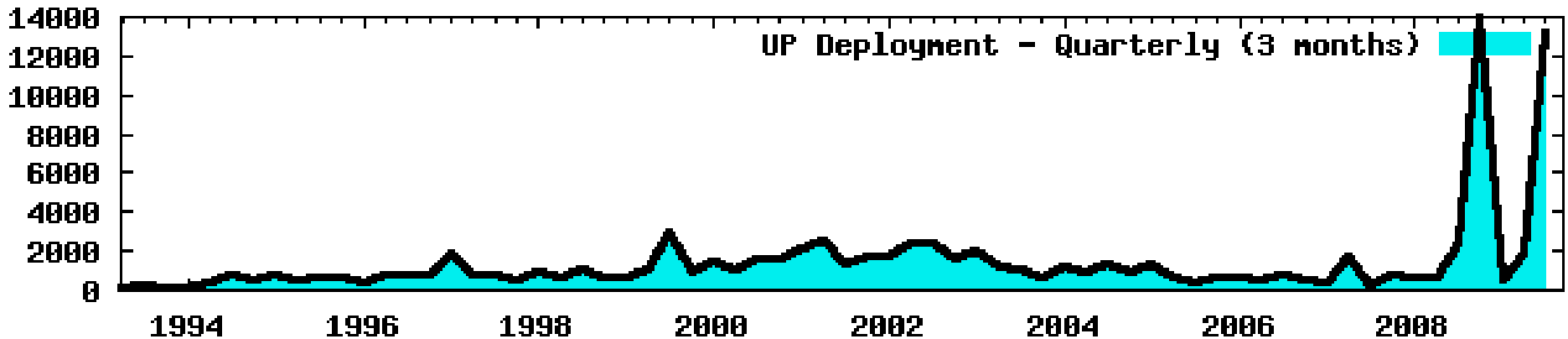




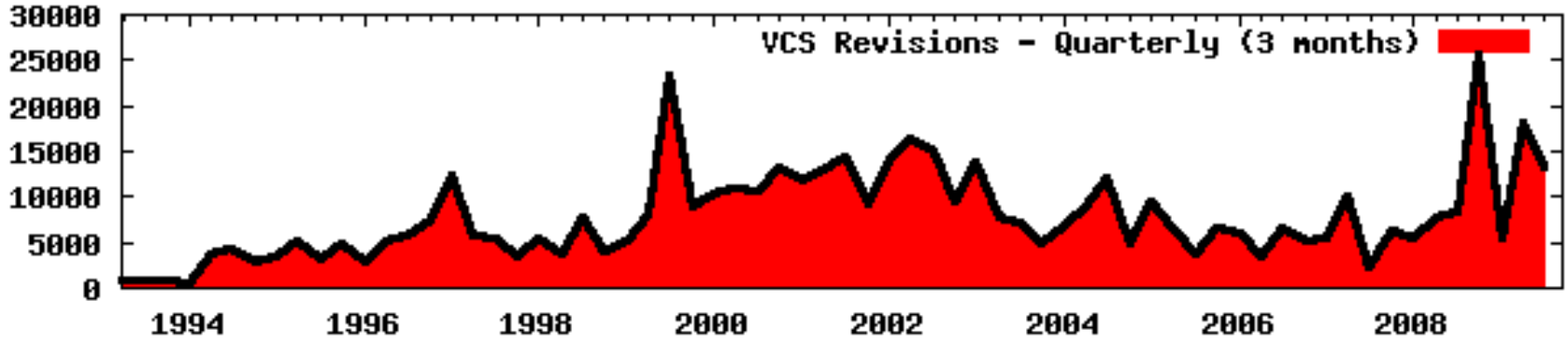


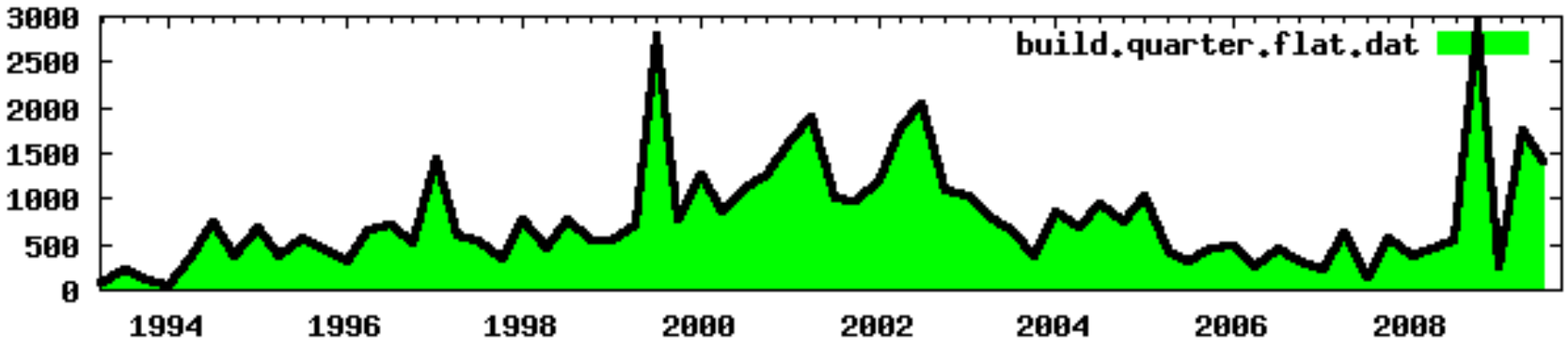




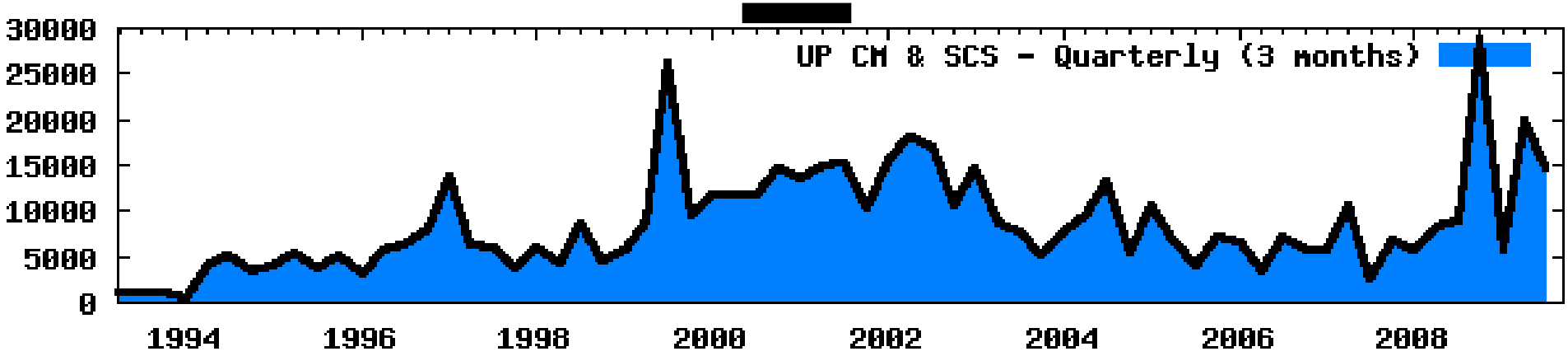


UP Configuration Managment and SCS

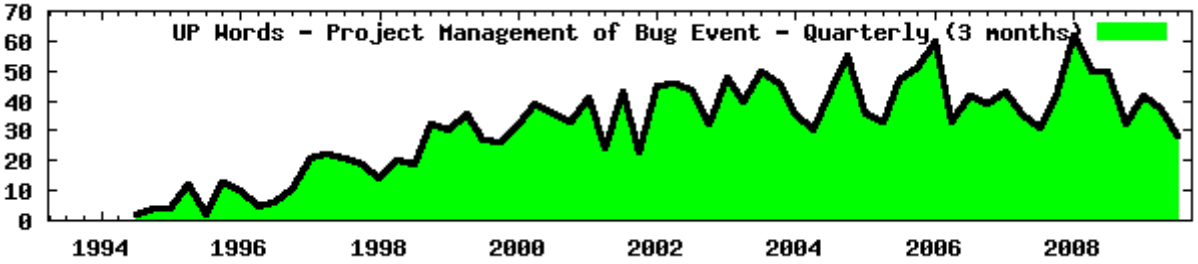


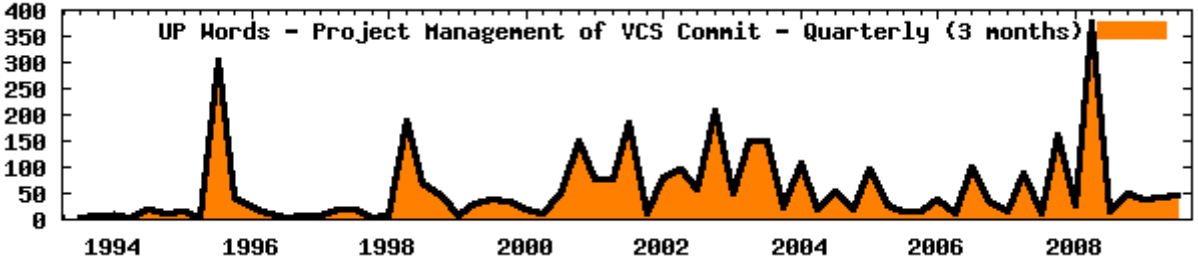


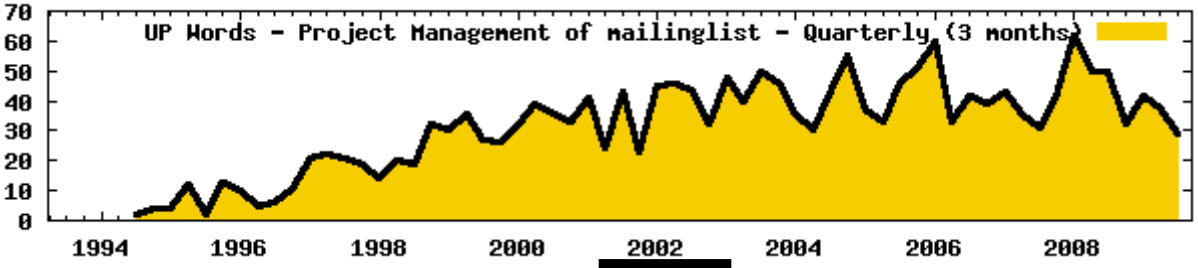
=

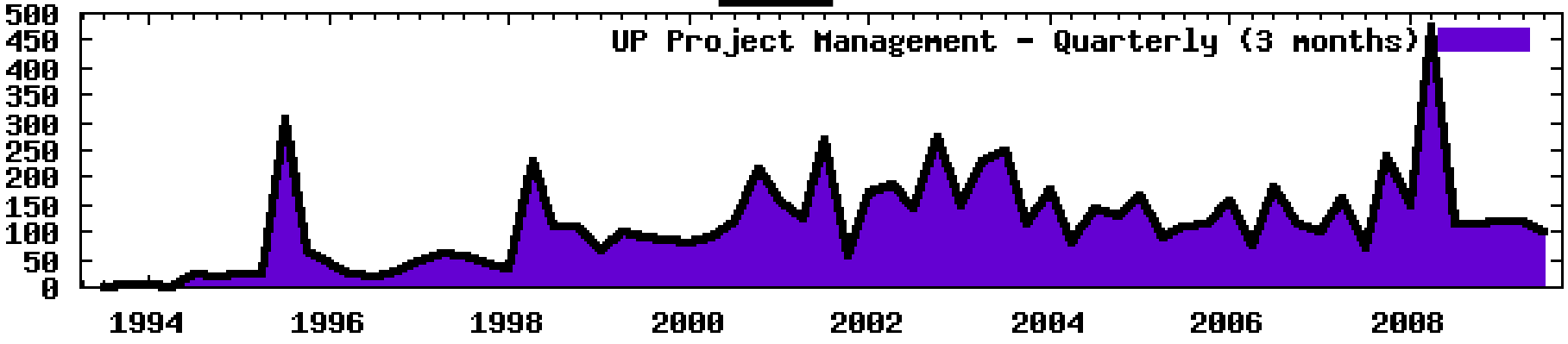


UP Project Management Signal

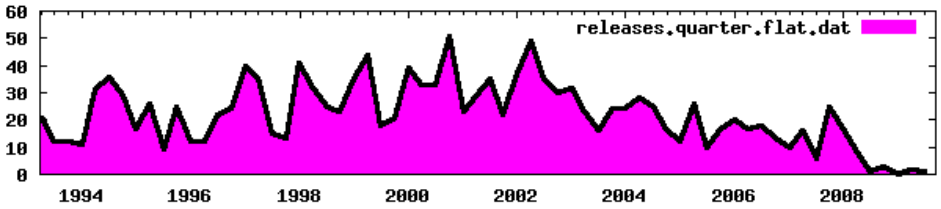


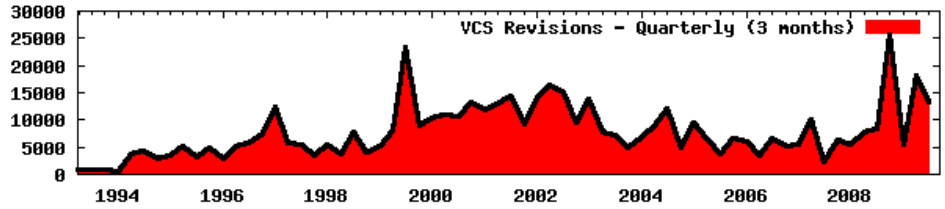


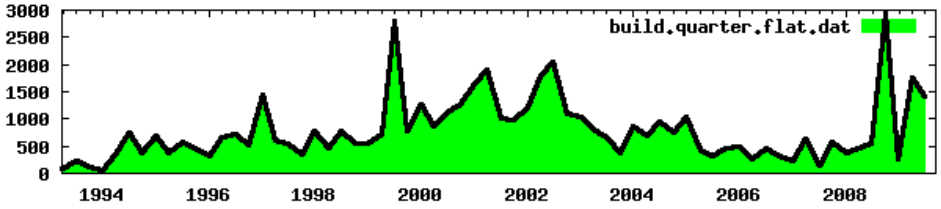


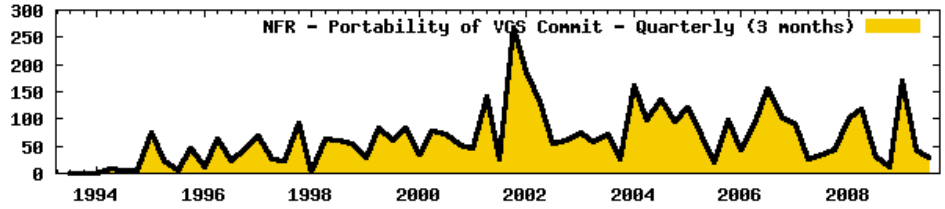


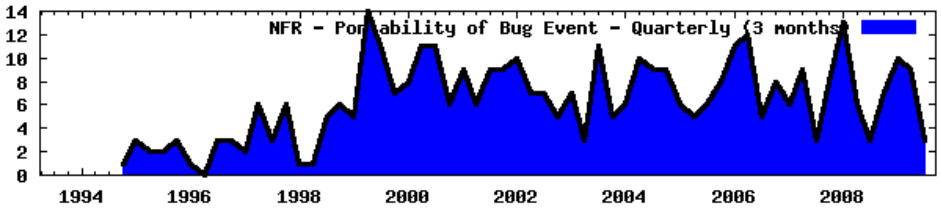
UP Environment Signal

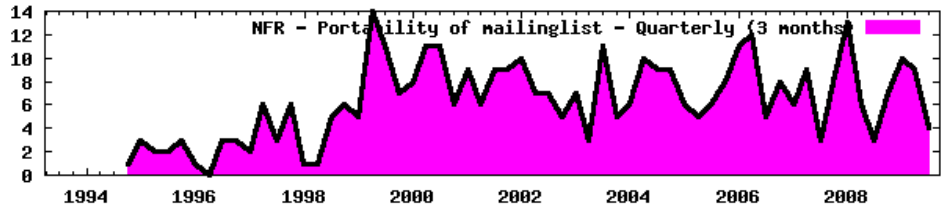




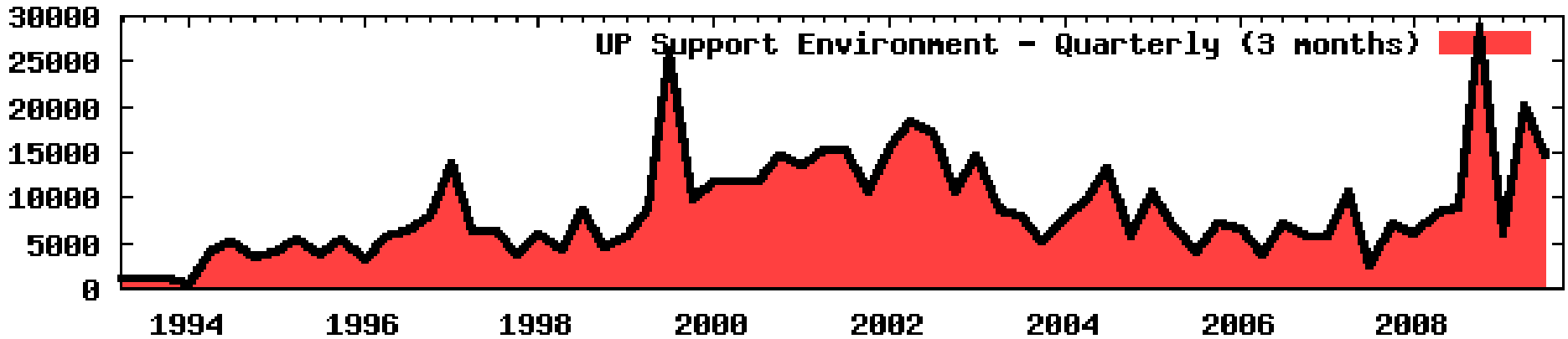




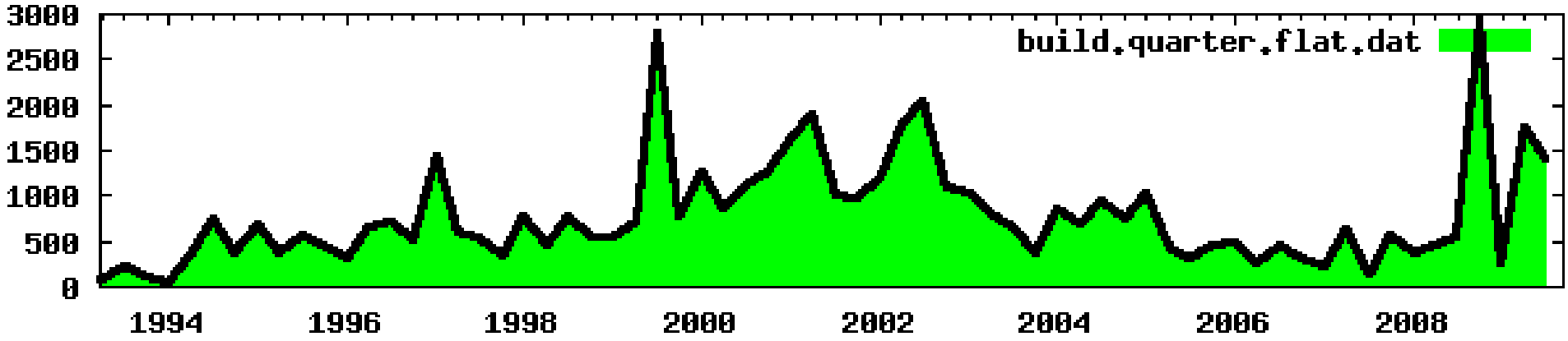




=



Process Heavy Signals: Build Commits



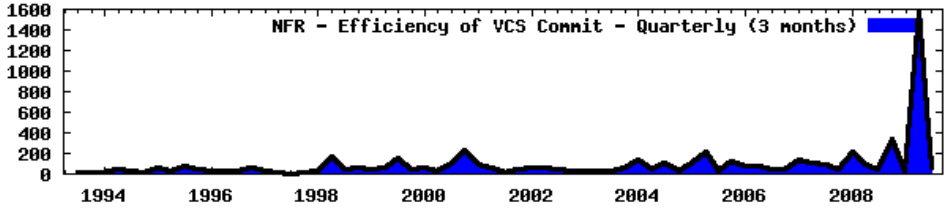
Related to

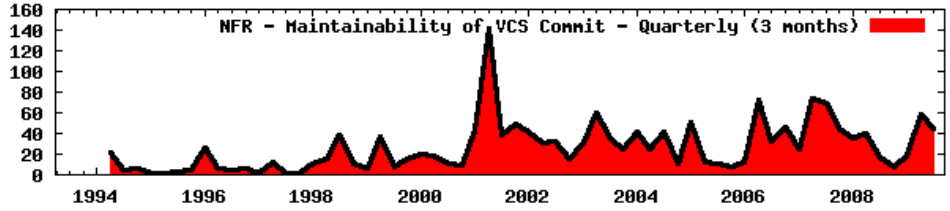
- portability

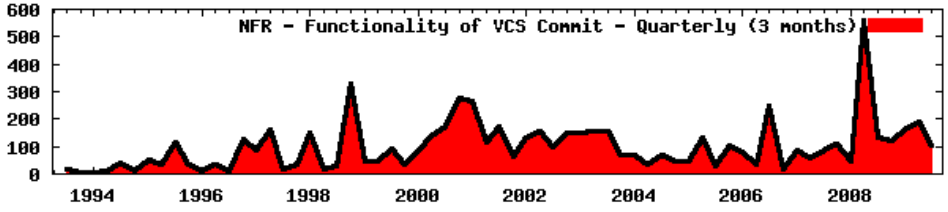
- change in modularity

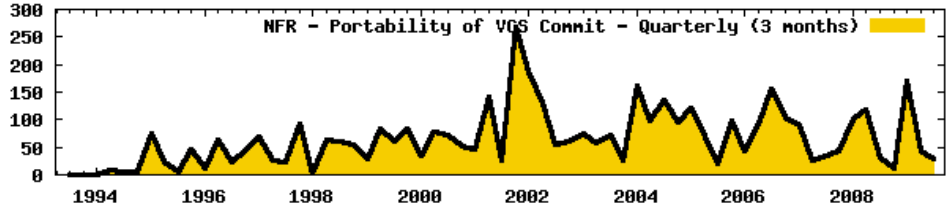
- feature addition and removal

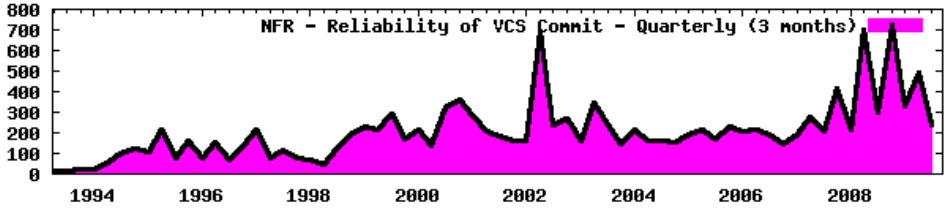
Process Heavy Signals:
Non Functional Requirements

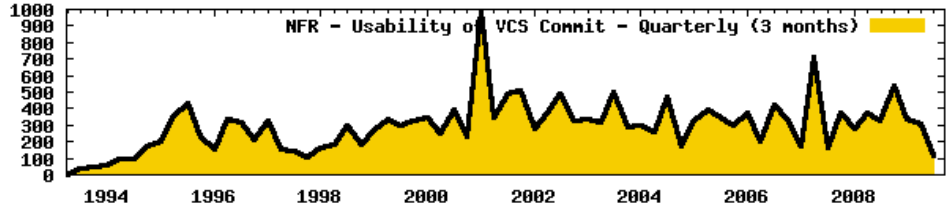






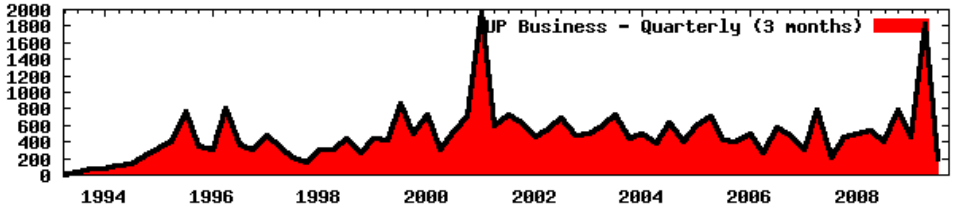


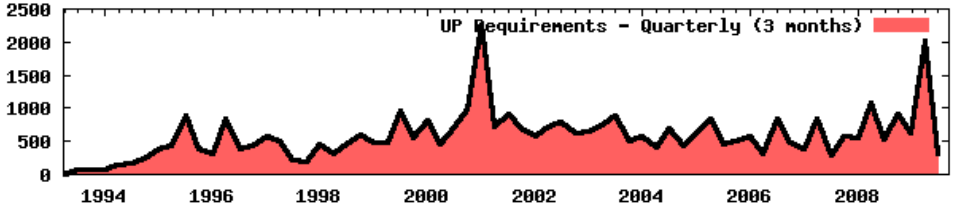


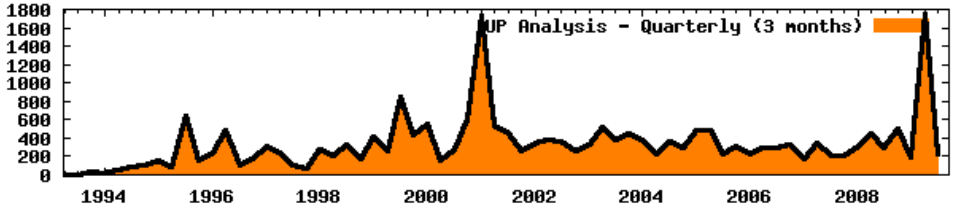


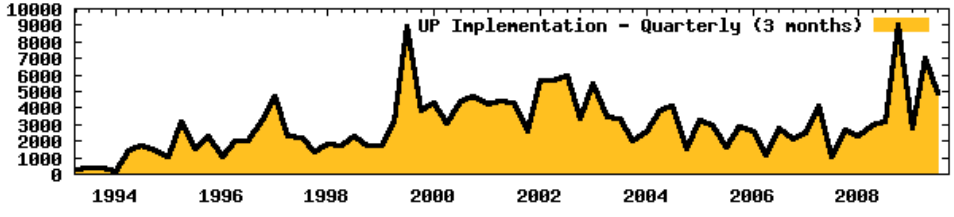
UP Observability

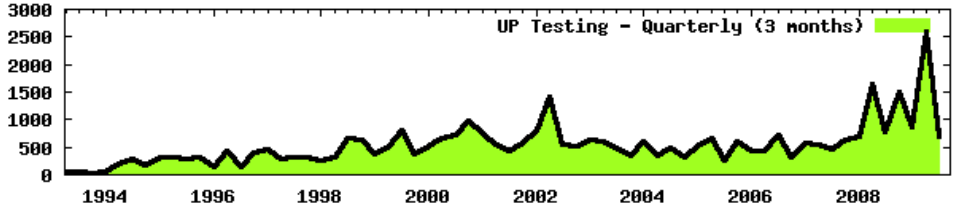
FreeBSD Case Study

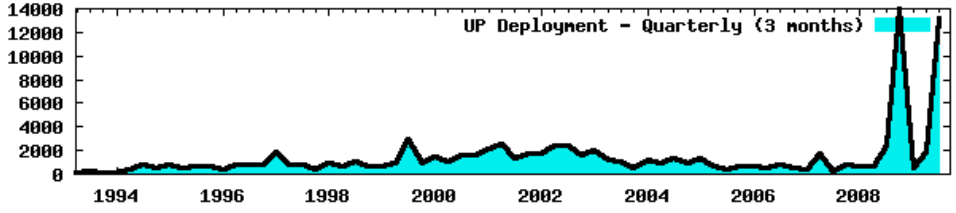


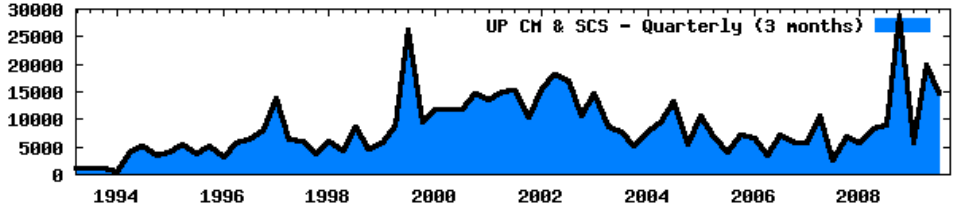


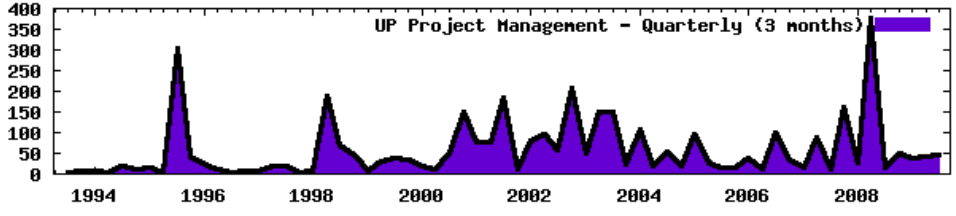


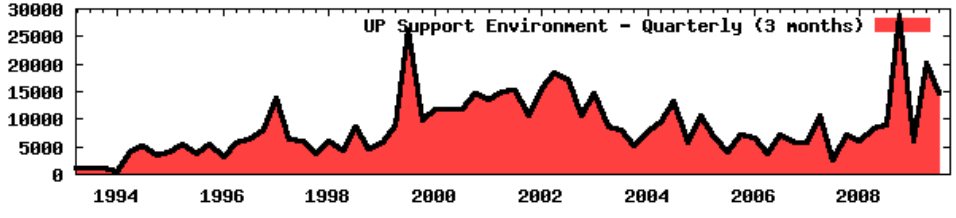




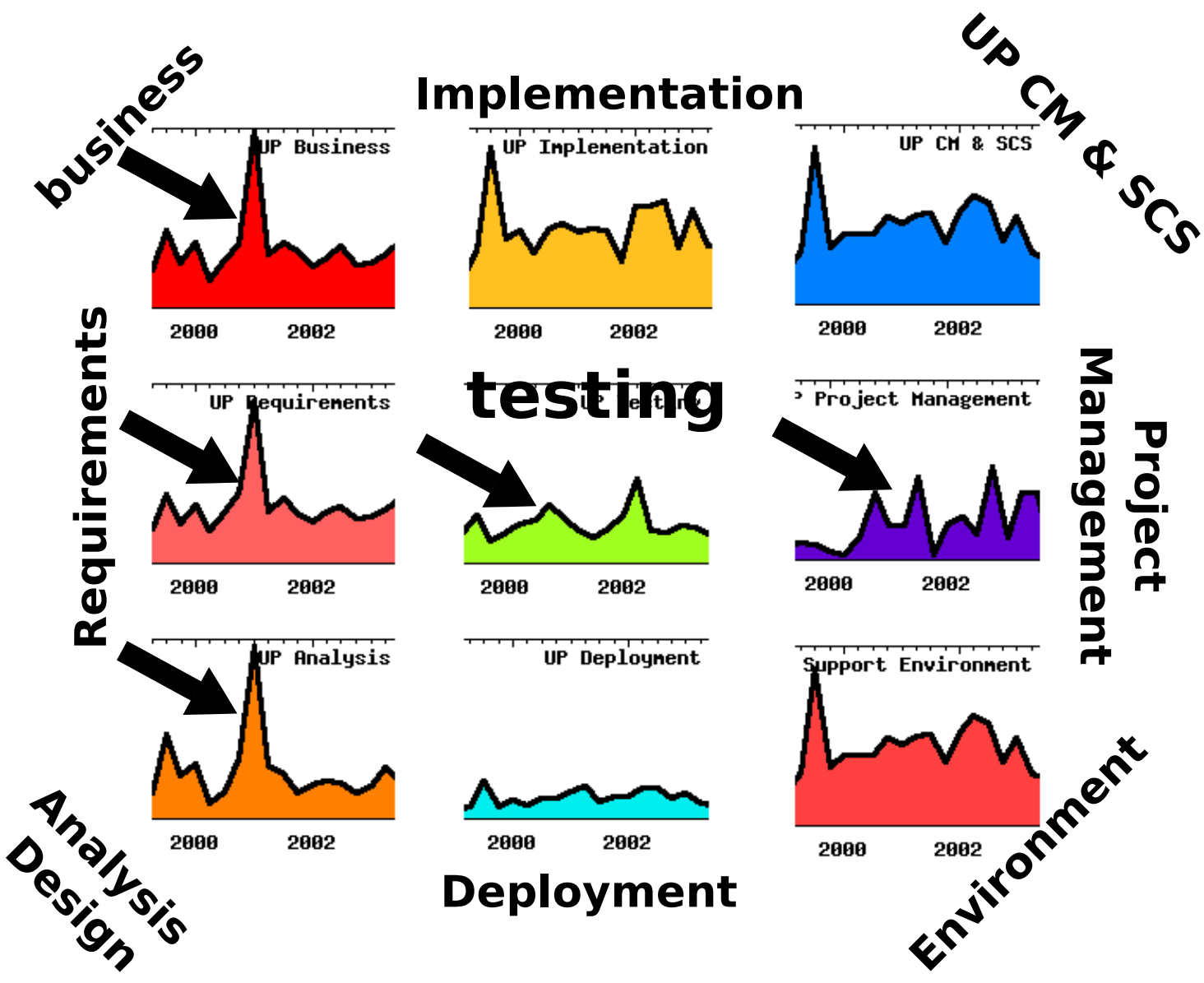




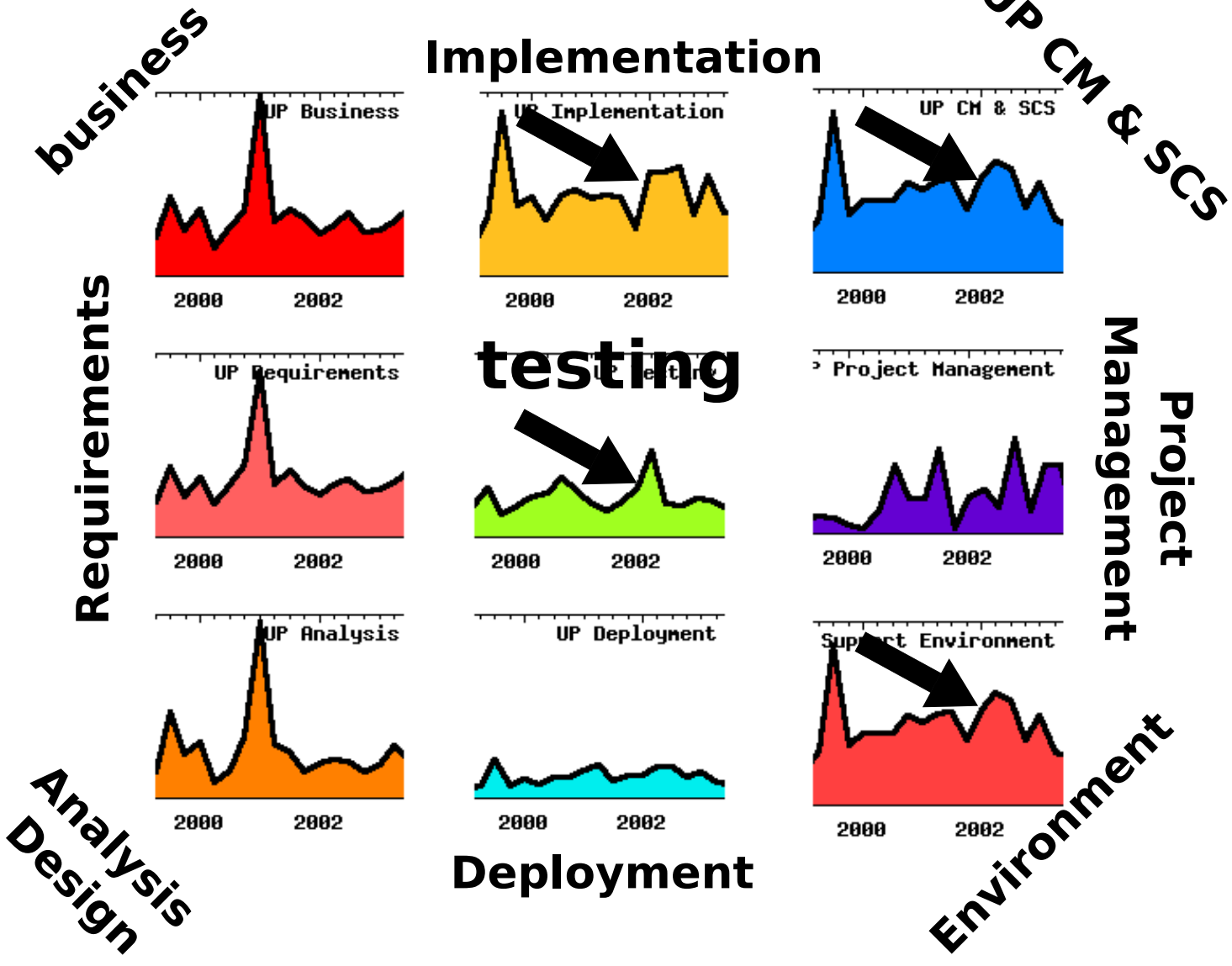




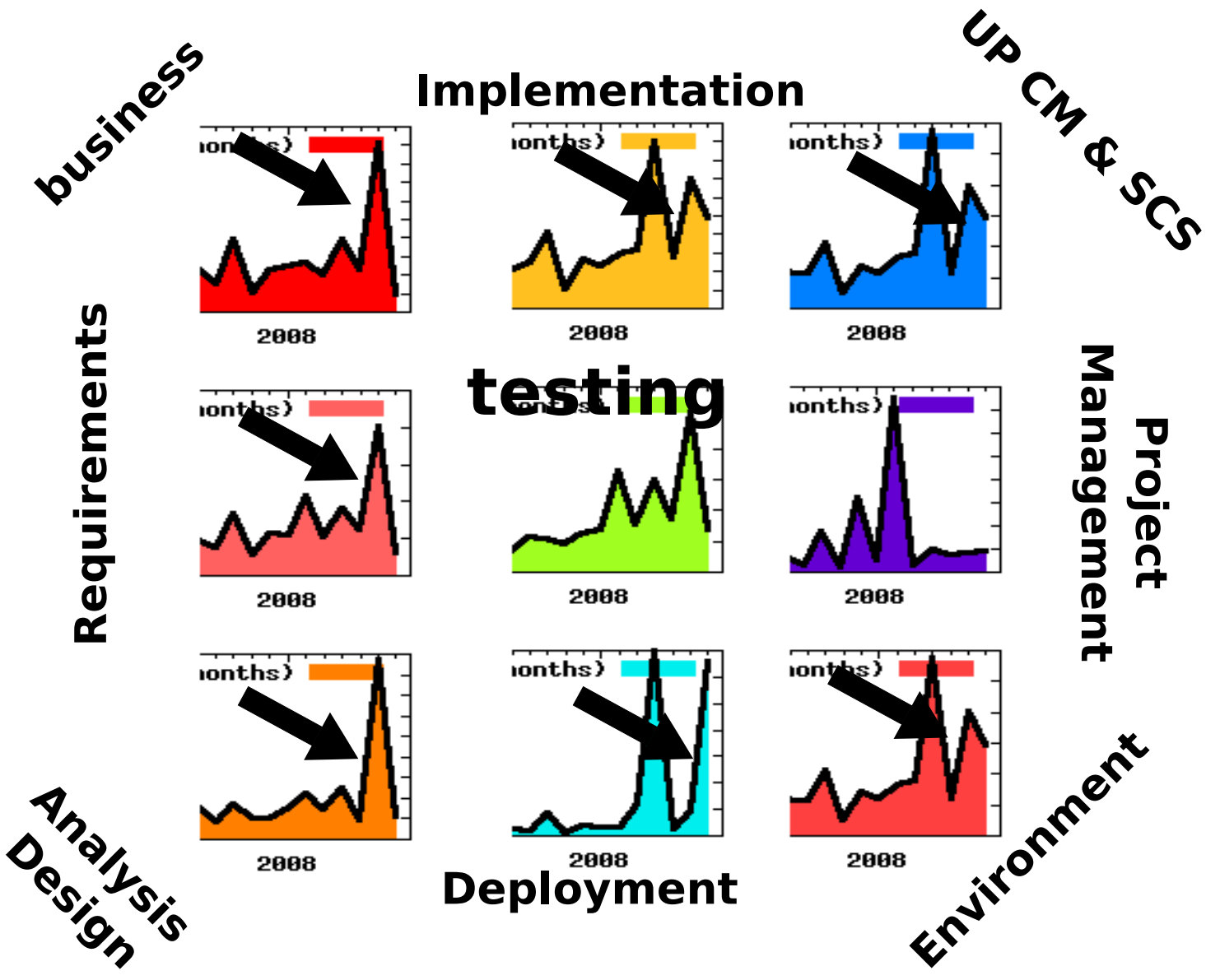
FreeBSD Case Study: 2001



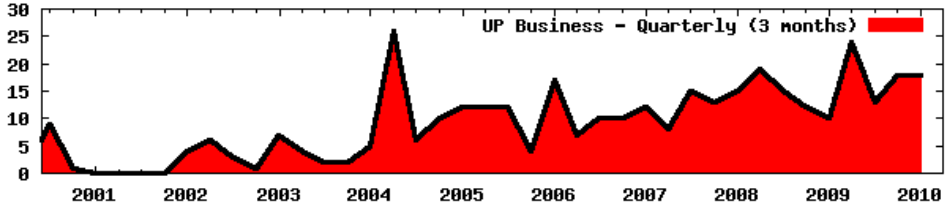
FreeBSD Case Study: 2002

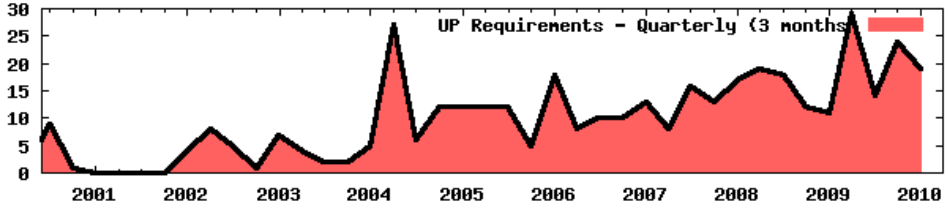


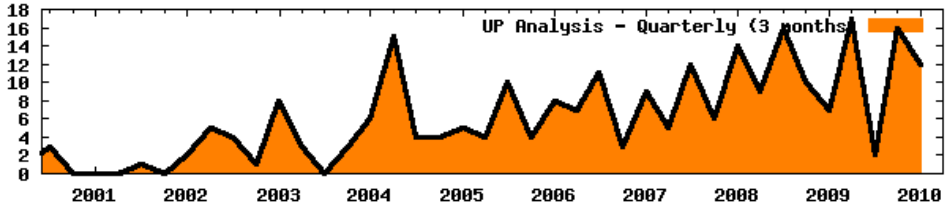
FreeBSD Case Study: 2009

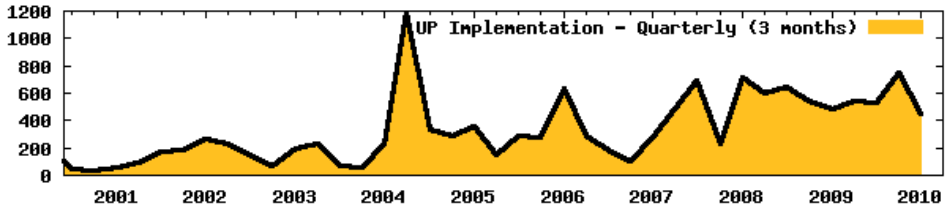


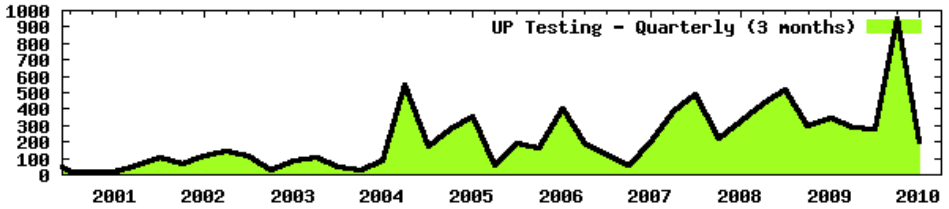
SQLite Case Study

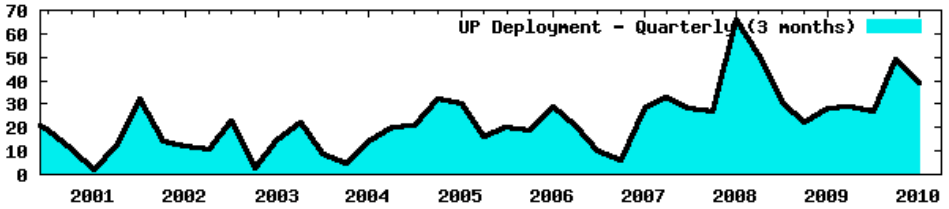


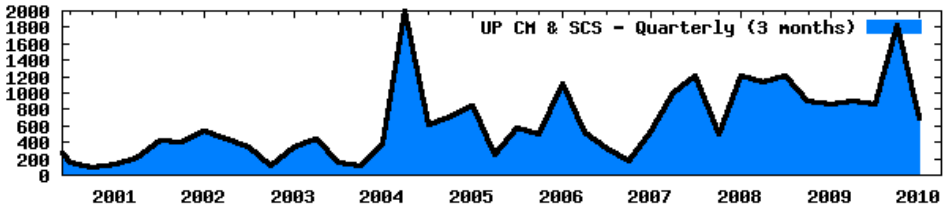


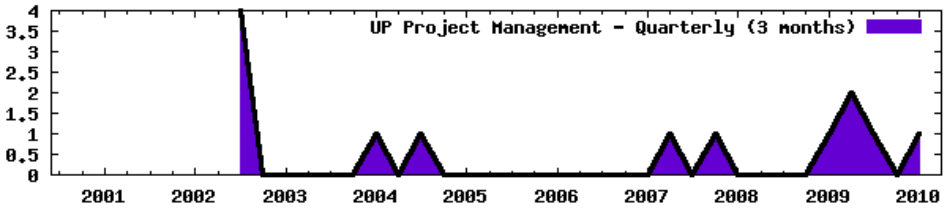


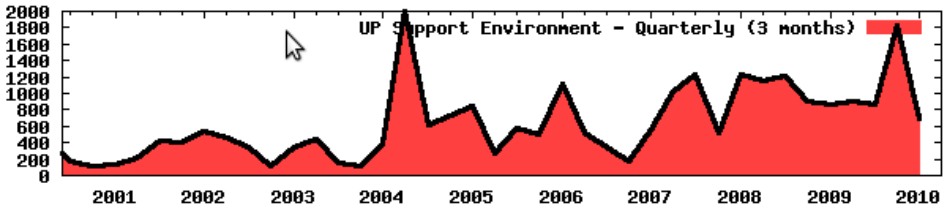




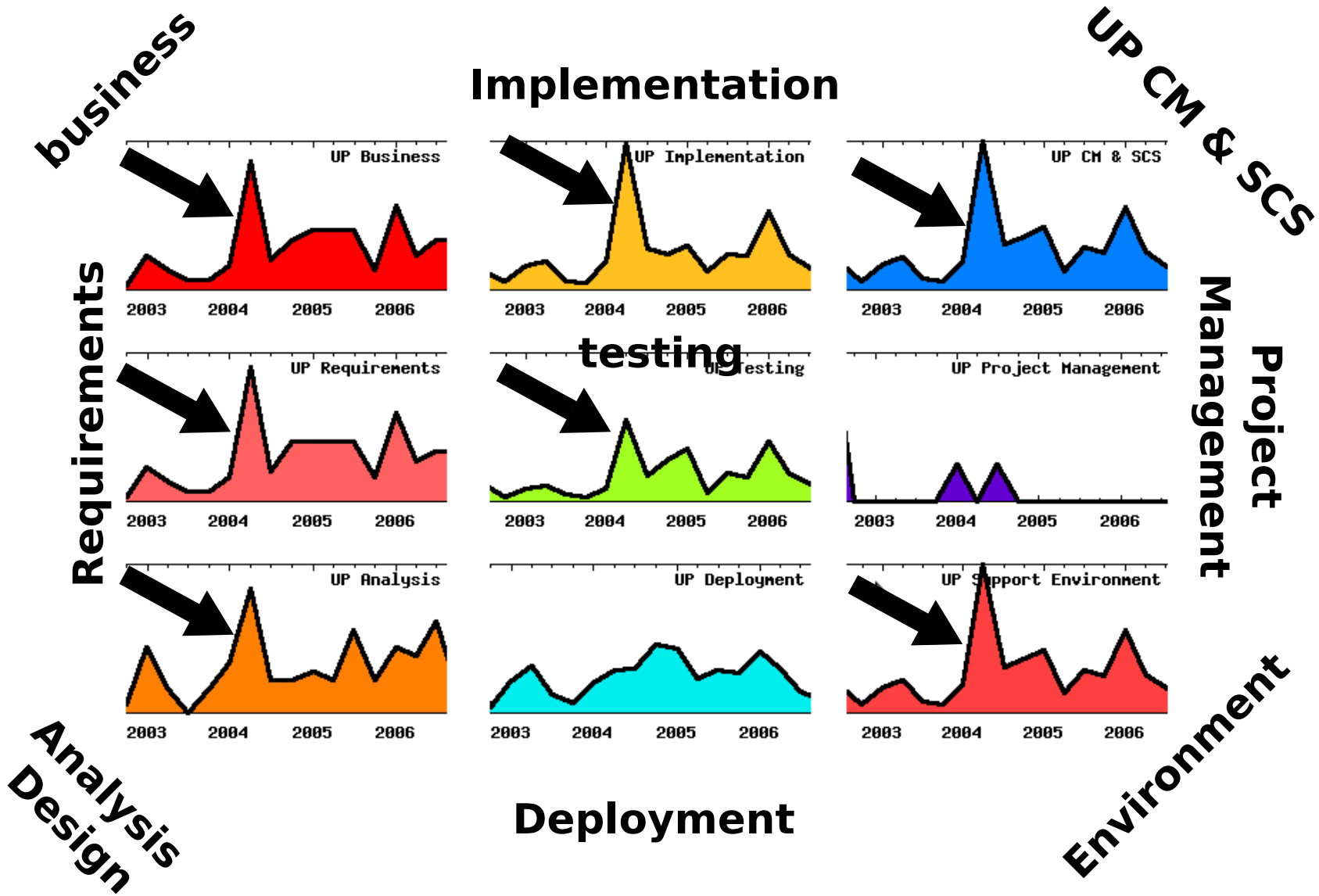




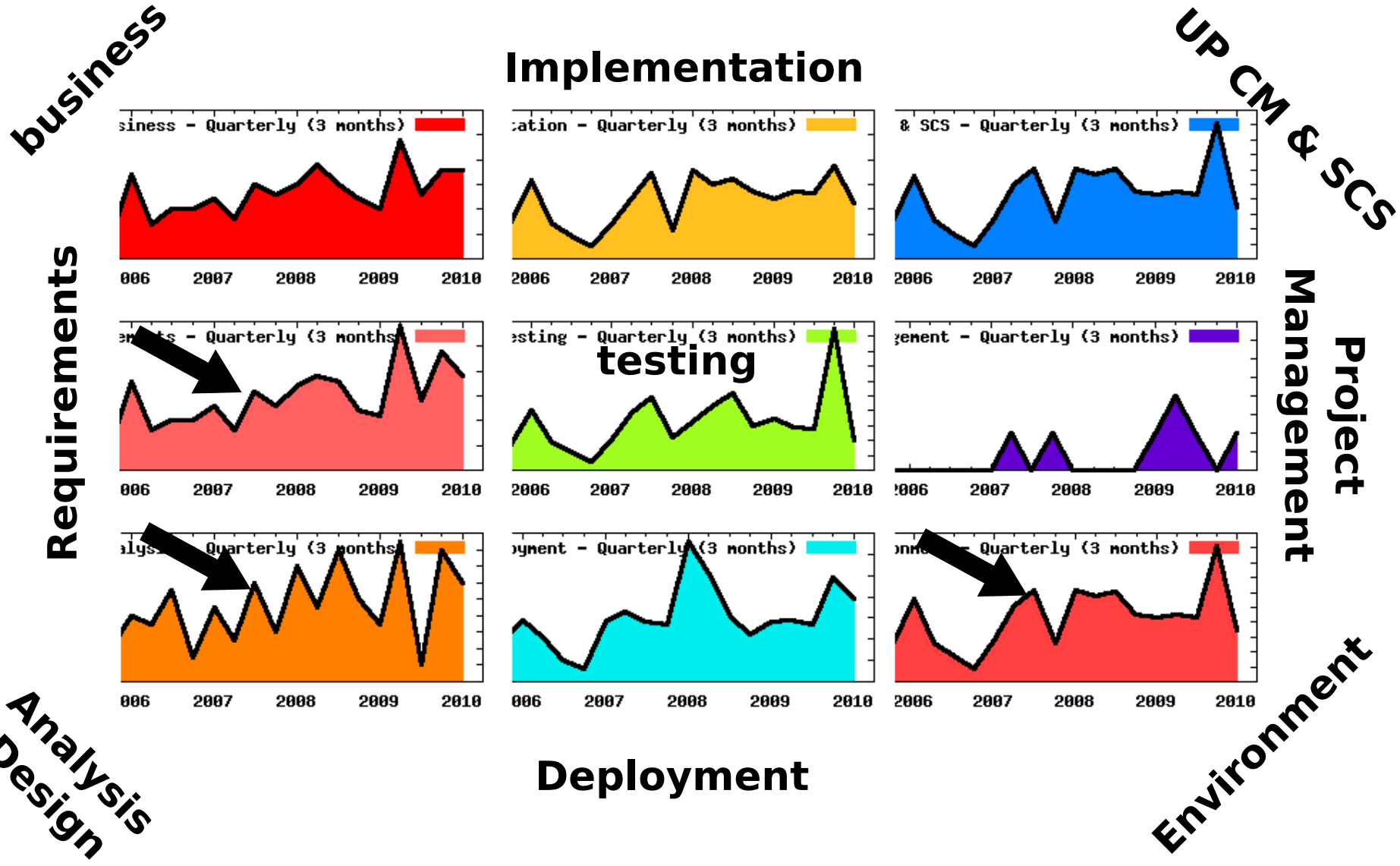




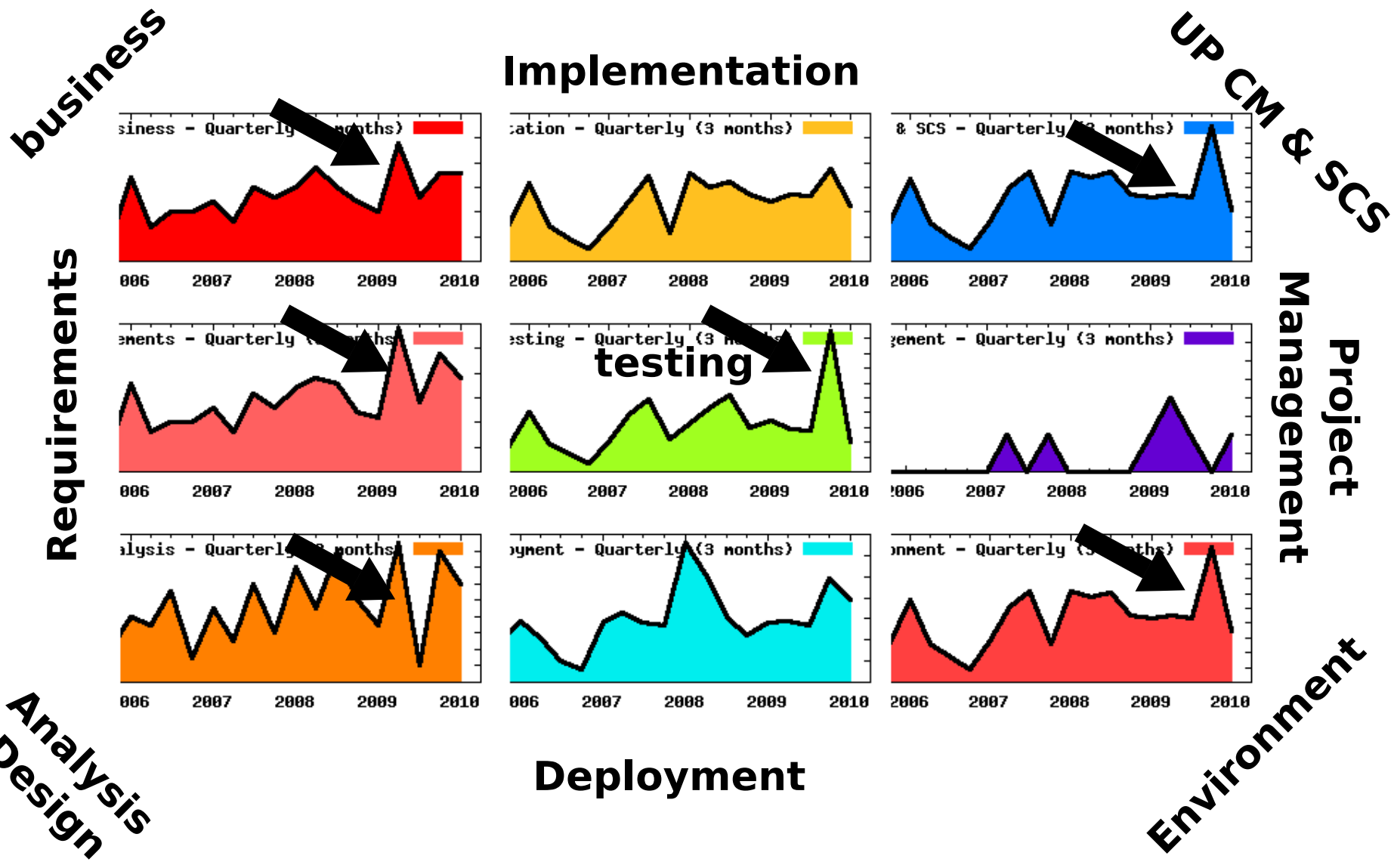
SQLite Case Study: 2004



SQLite Case Study: 2007-2008

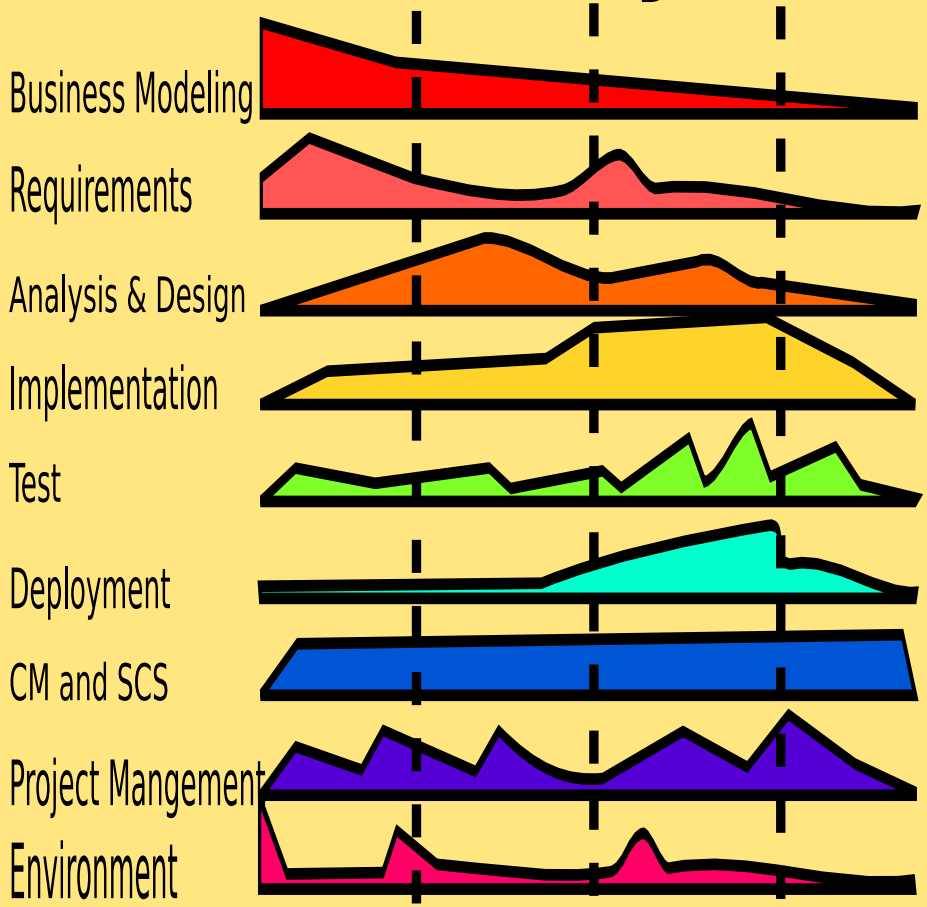


SQLite Case Study: 2009

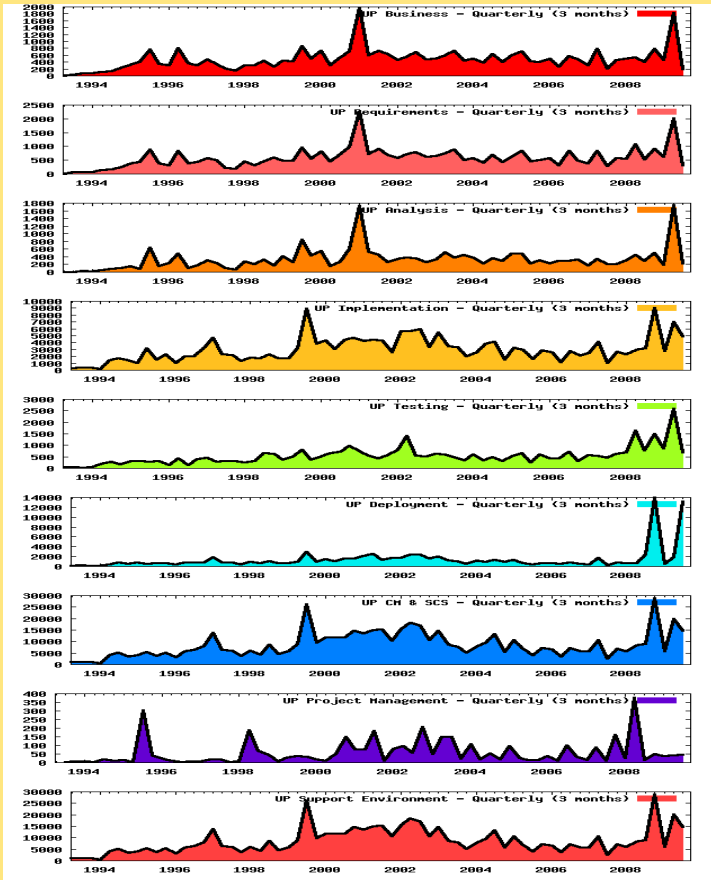


What have we done?

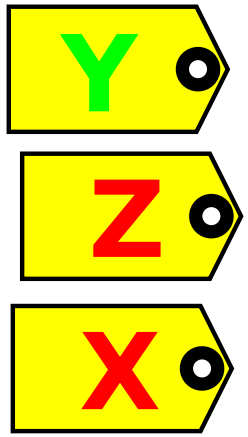
Theory



Practice

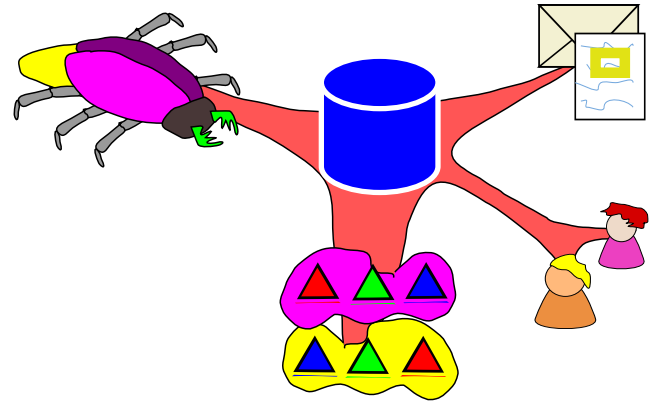


Looking Forward
What can other tools do to help?

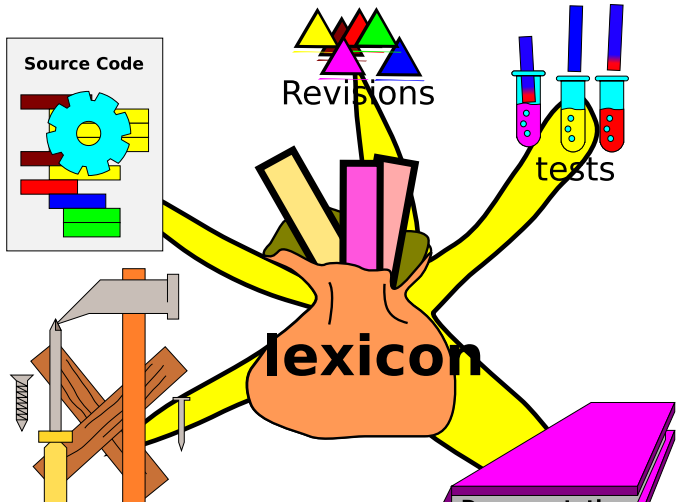


Tagging

Keep all artifacts in repositories



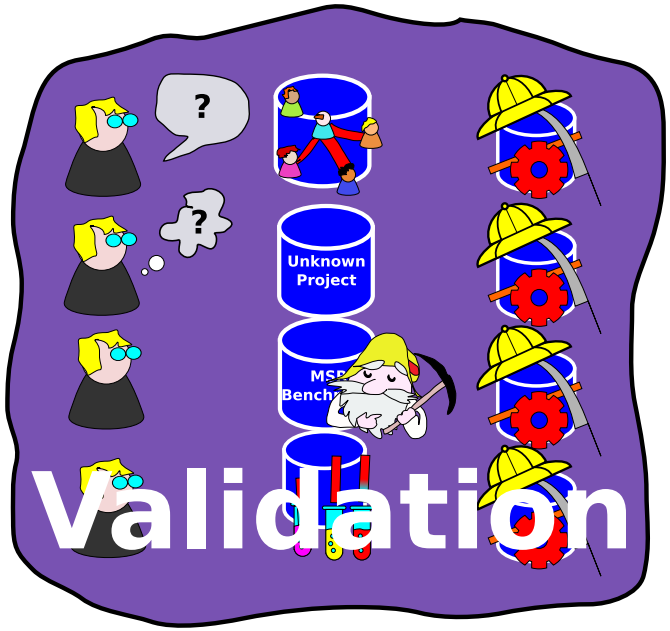
Traceability



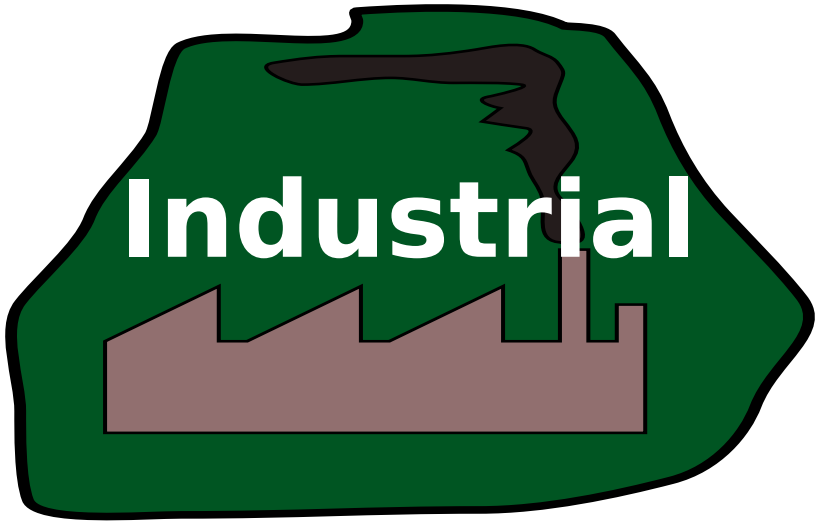
Common Lexicon and Vocabulary

Future Work


People
and
teams

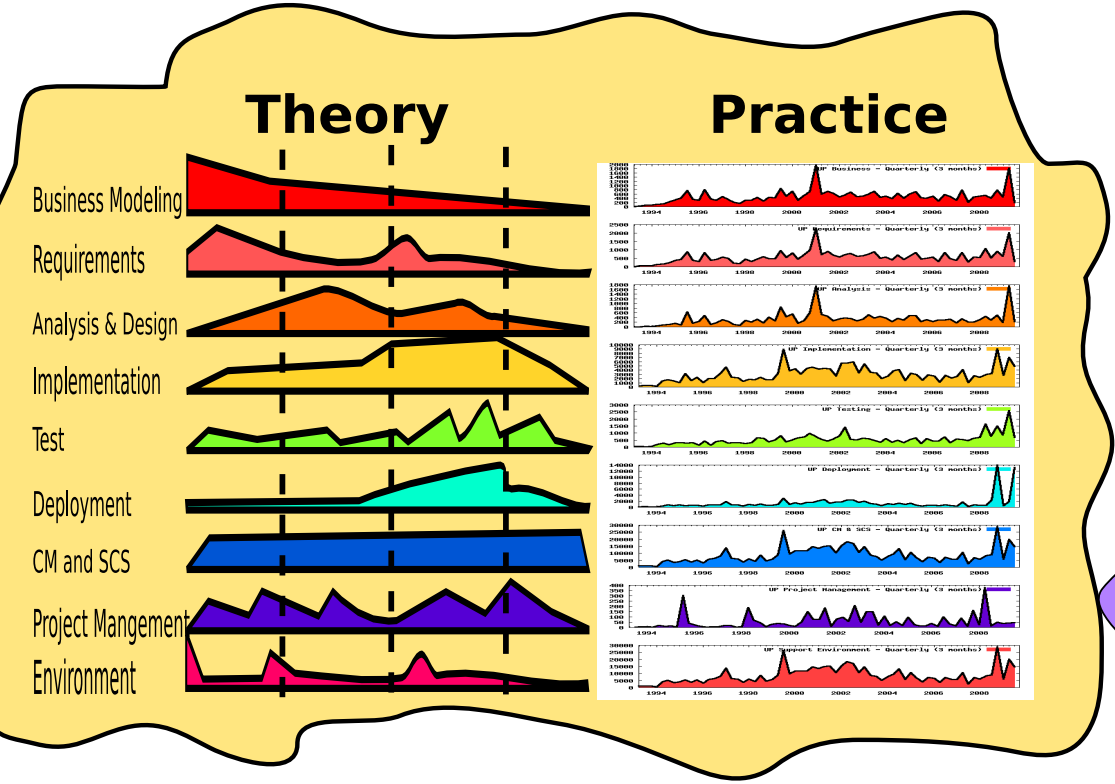

?
?
Unknown
Project
Validation

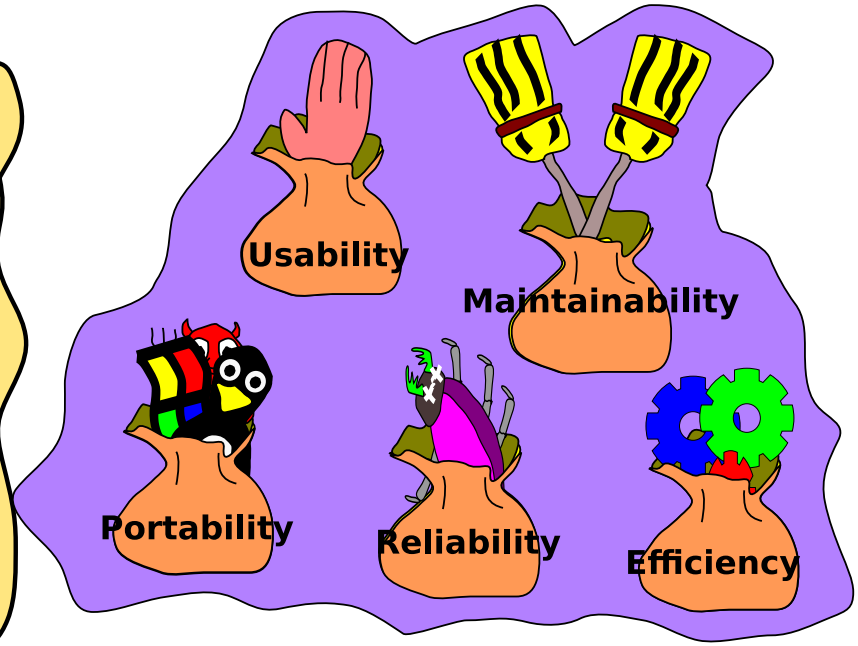

Accuracy


Industrial


Iteration
Identification

Conclusions





Related publications
- RUPV: ICSM10
- NFR topic labels: http://softwareprocess.es/name
- Developer Topics: ICSM09
- Release Patterns: MSR07 & ICSM07
- Maintenance Categories: ICPC09 & MSR08

http://softwareprocess.es/RUPV/