Mining Software Data

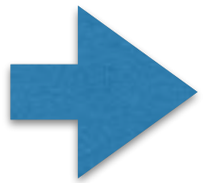
María Gómez

Software Engineering Course — Summer Semester 2017

How Software is built is changing…

- Code centric
- In-lab testing
- Centralized development
- Long product cycle

….



- Data pervasive
- Debugging in the large
- Distributed development
- Continuous release

….

Slide adapted from: https://de.slideshare.net/taoxiease/software-mining-and-software-datasets

Software Data

- Large amount of **artefacts** are generated in the **sw development process**
- Increased amount of data available in software archives through large open source projects

Software Decision Making

Sw developers rely on their **prior experiences** to plan sw projects, fix bugs, prioritise testing, etc.

Mining Software Repositories (MSR)

Let’s mine software data!

What?

Why?

How?

What is *Mining Software Repositories* (MSR)?

”The **MSR** field analyzes rich **data** available in **software repositories** to extract useful and **actionable information** about software projects and systems”. (Source: msrconf.org)

What is Mining Software Repositories (MSR)?

Main goals:

- Gather and exploit data produced by developers (and other sw stakeholders) in the software development process.
- Uses data available in repositories to support development activities (e.g., defect assignment, software validation, evolution and planning).
- Discover hidden patterns and trends.
- Transform static record-keeping repositories into active repositories to guide decision processes.
- Applies data extraction and analysis to make decisions and predictions.

[1] The Road Ahead for Mining Software Repositories. Ahmed E. Hassan.
[2] Effective Mining of Software Repositories. Marco D'Ambros, Romain Robbes.

MSR

- *What types of software data are available to mine?*
- *Which data mining techniques can be used in MSR?*
- *Which software engineering tasks can be assisted with MSR?*

MSR

- ***What types of software data are available to mine?***

- *Which data mining techniques can be used in MSR?*

- *Which software engineering tasks can be assisted with MSR?*

What to mine?

Software repositories refer to artefacts produced and archived during software development processes by developers and other stakeholders.

What to mine?

Different types of **repositories**[1]:

Historical Repositories

Code Repositories

Runtime Repositories

[1] The Road Ahead for Mining Software Repositories. Ahmed E. Hassan.

What to mine?

Historical Repositories

Record information about the evolution and progress of a project

Examples:

- Version control systems (CVS, SVN, Git, Mercurial)
- Bug repositories (Bugzilla, JIRA)
- Mailing lists (e-mails, wiki pages)
- Development collaboration sites (StackOverflow)

What to mine?

Code Repositories

Contain source code of various applications
Developed by several developers

Examples:

- Code bases (SourceForge, GoogleCode)
- Project ecosystems (GitHub)

What to mine?

Runtime Repositories

Contain information about the execution and usage of an application

Examples:

- Crash reports
- Field logs
- Execution traces

What to mine?

Other Repositories

Examples:

- App Stores (Google Play Store, Apple App Store)
  - Contain mobile apps and user feedbacks (reviews, ratings)

What to mine?

Historical Repositories

Runtime Repositories

Cross-link of repositories!

Code Repositories

Other Repositories

Why MSR?

- Better manage software projects
  - Produce higher-quality software systems that are delivered on time and within budget
- Support maintenance of software systems
- Improve software design/reuse
- Learn from past to guide future development

1 MSR Conference: http://2017.msrconf.org/#/home
2 Mining Software Engineering Data. Ahmed E. Hassan & Tao Xie.

Target Audience

- Software practitioners
  - Project Manager
  - Developers
  - Designers
  - Testers
  - Usability engineers
  - Engineers

MSR

- What types of software data are available to mine?
- ***Which software engineering tasks can be assisted with MSR?***
- Which data mining techniques can be used in MSR?

Applications of MSR

- Estimate developer efforts
- Change impact and propagation
- Risk management (trends)
- Fault analysis and prediction
- Test reduction, minimisation and selection
- Continuous quality assurance
- Post-release maintenance

Applications of MSR

- New bug report
  - Estimate fix effort
  - Mark duplicate
    - Suggest experts and fix
- New change
  - Suggest APIs
  - Warn about risky code or bugs
  - Suggest locations to co-change

MSR

- What types of software data are available to mine?
- Which software engineering tasks can be assisted with MSR?
- ***Which data mining techniques can be used in MSR?***

MSR Process

Repositories

EXTRACT → ANALYZE → SYNTHESIZE

Actionable Information

MSR Process

Repositories

EXTRACT → ANALYZE → SYNTHESIZE

Actionable Information

Data Extraction

- Extract data from different repositories
- Selection of input data
  - Processing (e.g., filtering)
- Constraints to help with scalability

MSR Process

Repositories

EXTRACT → ANALYZE → SYNTHESIZE

Actionable Information

Data Analysis

- Process the data
- Link data between repositories
- **Empirical analysis** to the data

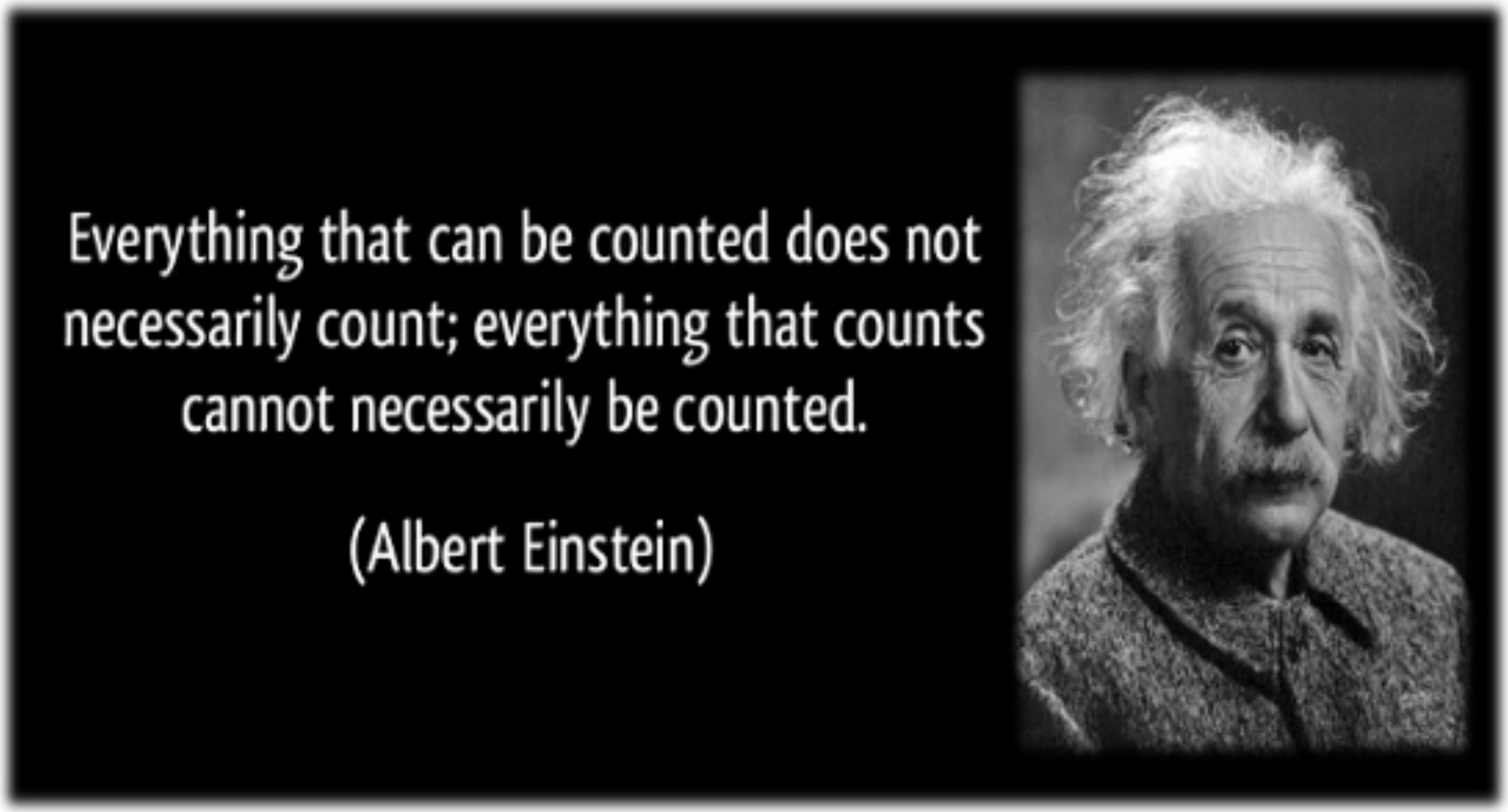
Types of Empirical Analysis

Different types of empirical analysis can be performed in repositories:

- Quantitative vs qualitative
- Regression models
- Grounded theory
- Machine learning/data mining

Types of Empirical Analysis

Quantitative vs qualitative



Types of Empirical Analysis

Quantitative vs qualitative

Quantitative

Data is numerical

Data can be measured

Qualitative

Data non-numerical

Data can be observed

Types of Empirical Analysis

Quantitative vs qualitative

Example quantitative study:

Do performance bugs take more time to fix?

Are performance bugs fixed by more experienced developers?

Example qualitative study:

What are the advantages/disadvantages of shared code ownership from the developers perspective?

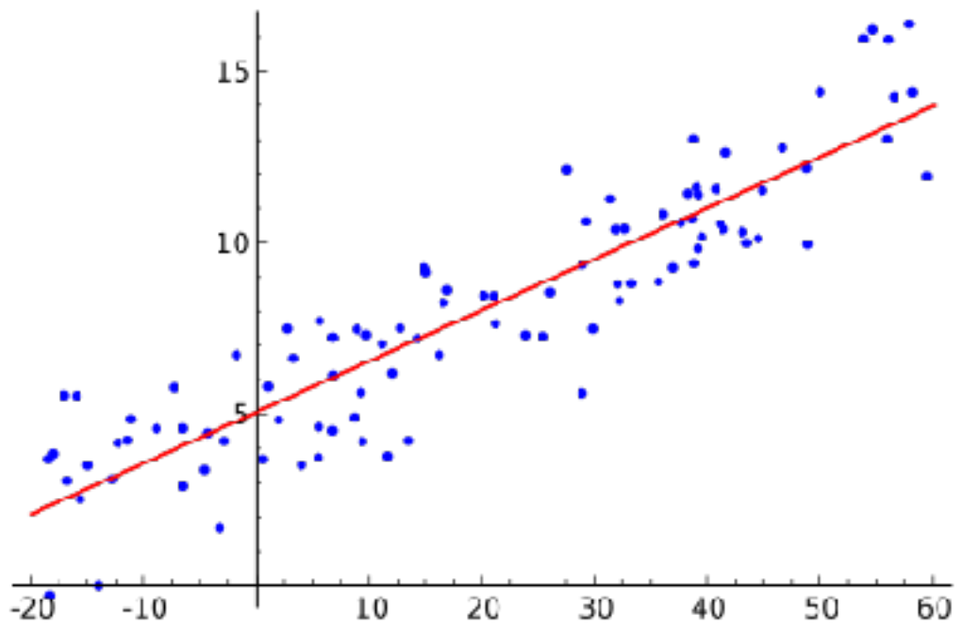
Types of Empirical Analysis

Regression models

- Estimate relationship among variables
- Widely used for prediction and forecasting

Example:

What factors contribute to delays on bug fixing time most?



Types of Empirical Analysis

Grounded theory

- Building theory from data
- Discovery of emerging patterns in data

Types of Empirical Analysis

Grounded theory

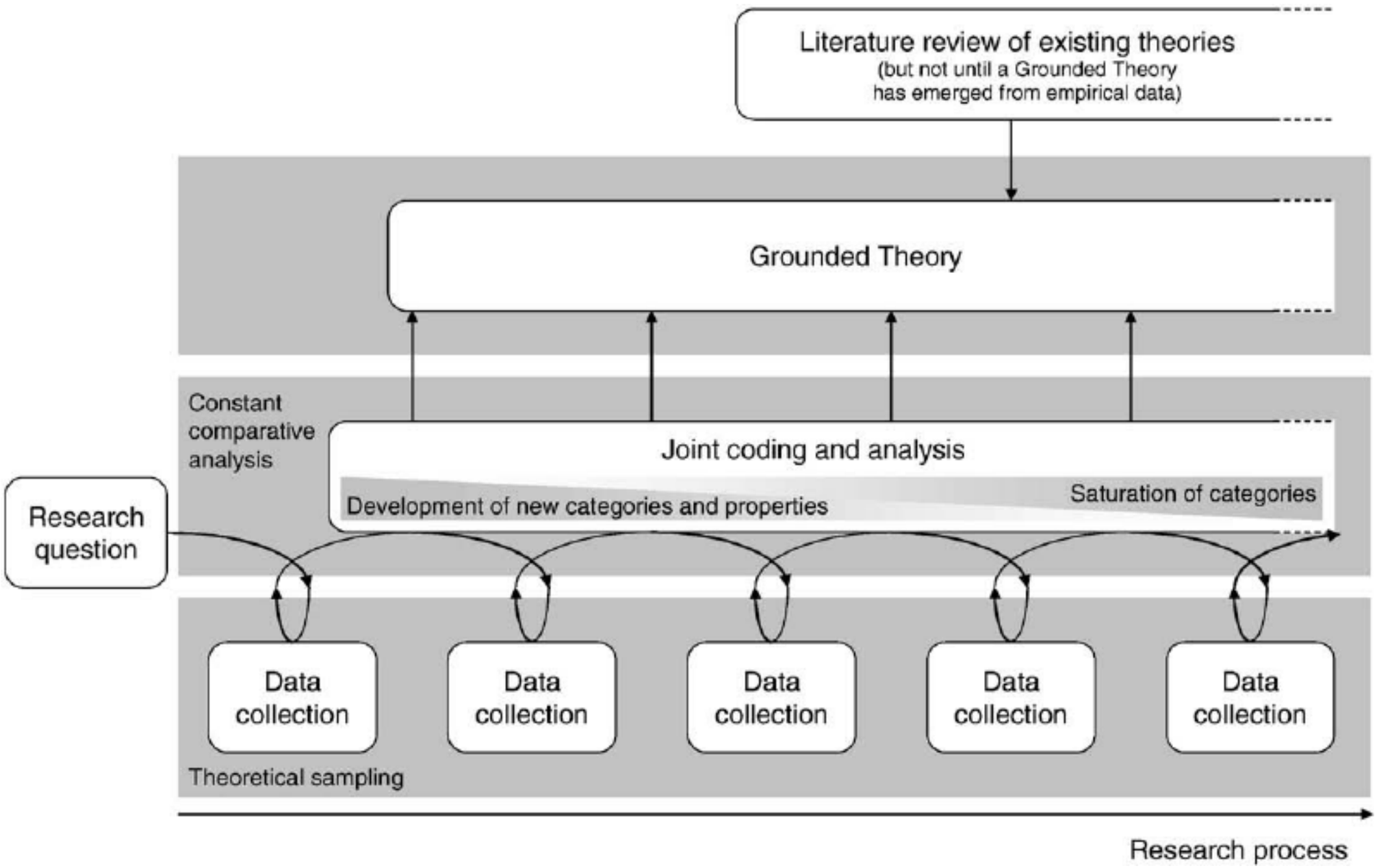


Figure source: https://www.researchgate.net/figure/222301824_fig1_Fig-1-Basic-process-of-the-Grounded-Theory-approach

Types of Empirical Analysis

Machine learning/data mining techniques

- Association Rules and Frequent Patterns
- Classification
- Clustering

Data mining techniques

Association Rules and Frequent Patterns

- Find frequent patterns in a database
- Itemset: set of items
  - Support of itemsets
  - Confidence of rules

Association Rules Example

Transaction	Items
t_1	Bread,Jelly,PeanutButter
t_2	Bread,PeanutButter
t_3	Bread,Milk,PeanutButter
t_4	Beer,Bread
t_5	Beer,Milk

$X \Rightarrow Y$	s	α
Bread ⇒ PeanutButter	60%	75%
PeanutButter ⇒ Bread	60%	100%
Beer ⇒ Bread	20%	50%
PeanutButter ⇒ Jelly	20%	33.3%
Jelly ⇒ PeanutButter	20%	100%
Jelly ⇒ Milk	0%	0%

I = { Beer, Bread, Jelly, Milk, PeanutButter}

Support of {Bread,PeanutButter} is 60%

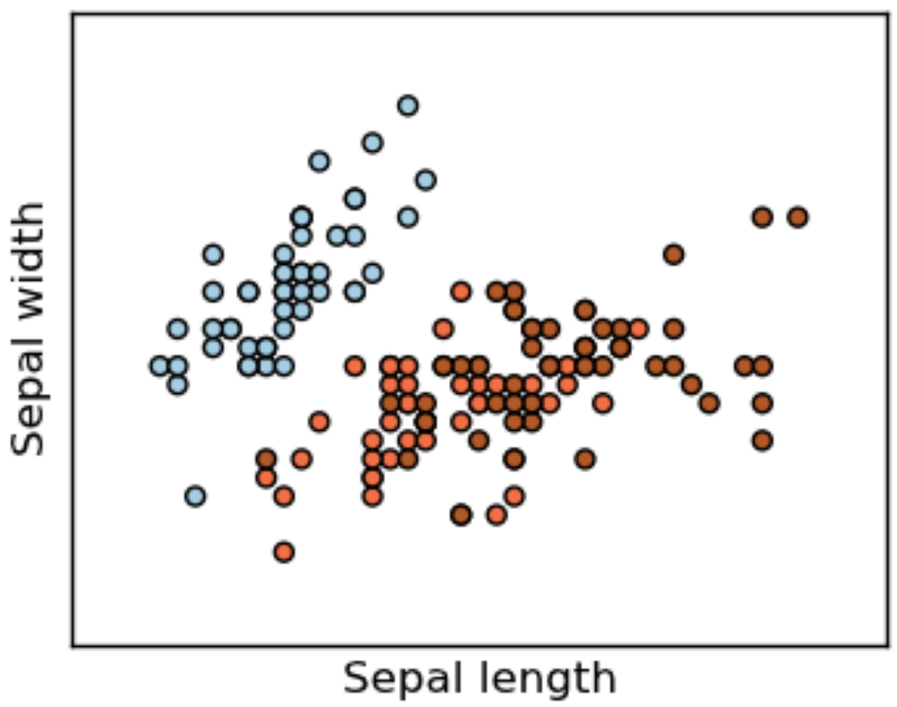
Image source: https://image.slidesharecdn.com/3-150328084211-conversion-gate01/95/31-mining-frequent-patterns-with-association-rulesmca4-4-638.jpg?cb=1427532681

Data mining techniques

Classification

- Supervised learning

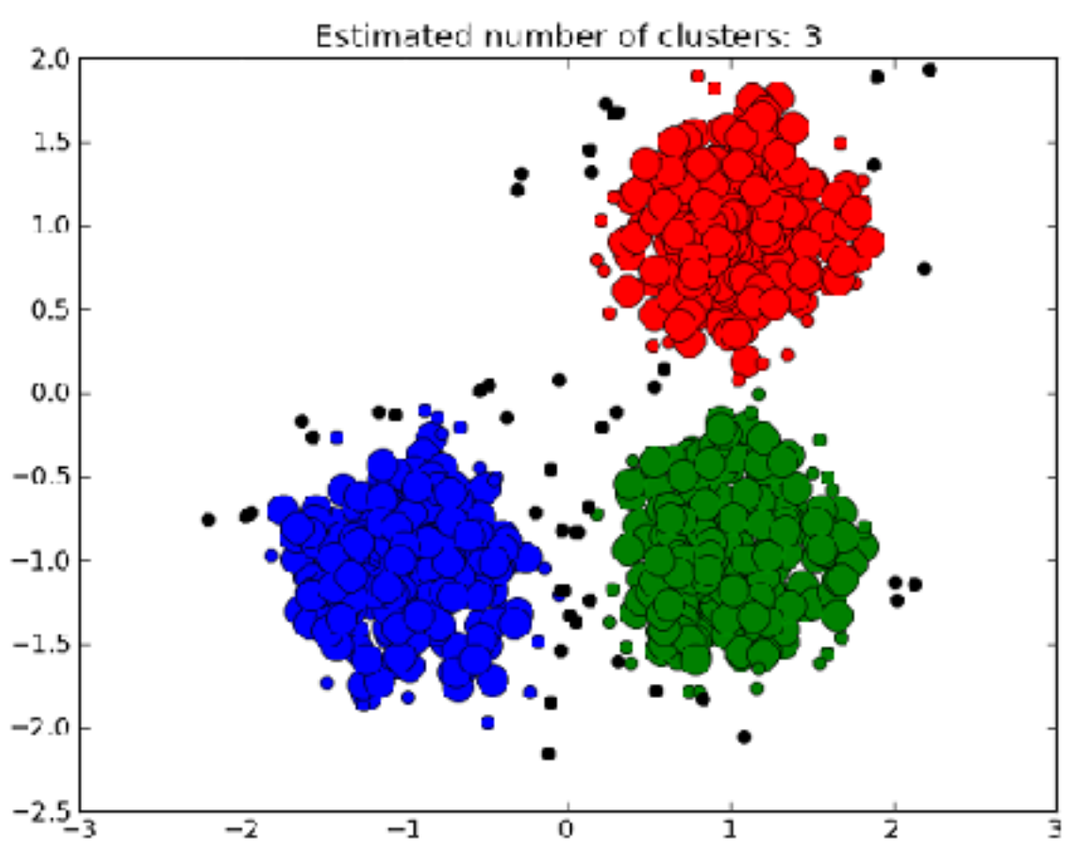
    1. Construct model with labeled objects (training set).
    2. Apply model to unlabelled objects.



Data mining techniques

Clustering

- Unsupervised learning (no predefined classes)
- Group similar data



Analysis Tools

Data mining and analysis tools:

- **R**
  http://www.r-project.org/
  Free software for statistical computing and graphics

- **Weka**
  http://www.cs.waikato.ac.nz/ml/weka/
  Open-source tool containing a collection of machine learning and data mining algorithms.

MSR Process

Repositories

EXTRACT → ANALYZE → SYNTHESIZE

Actionable Information

Data Synthesis

- Report / visualisation of outcome
- Understand the needs of practitioners
- Help practitioners to make decisions
  - Don’t replace them!

Actionable Outputs

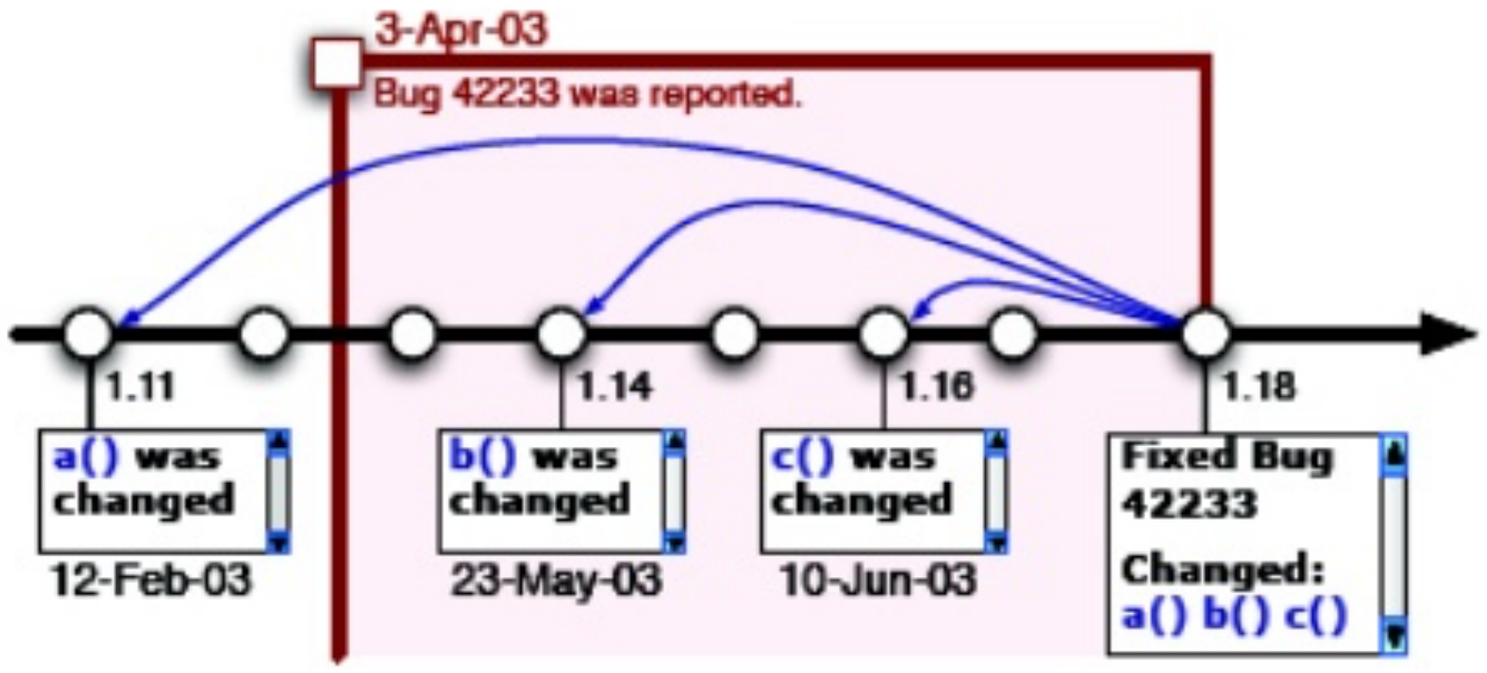
- Developer feedback
- Bug prediction
- Quality assurance
- Architecture analysis
- ……

What can we learn from software data?

MSR Application Examples

Can we predict bugs?

- Link bug fixes to source code changes
- Eclipse/Mozilla repos and bug-trackers
- Correlations found!



When do changes induce fixes? Jacek Sliwerski, Thomas Zimmermann and Andreas Zeller. (MSR' 05)

Can we predict bugs? (2)

Using Imports in Eclipse to Predict Bugs

71% of files that import compiler packages, had to be fixed later on.

```
import org.eclipse.jdt.internal.compiler.lookup.*;
import org.eclipse.jdt.internal.compiler.*;
import org.eclipse.jdt.internal.compiler.ast.*;
import org.eclipse.jdt.internal.compiler.util.*;
...
import org.eclipse.pde.core.*;
import org.eclipse.jface.wizard.*;
import org.eclipse.ui.*;
```

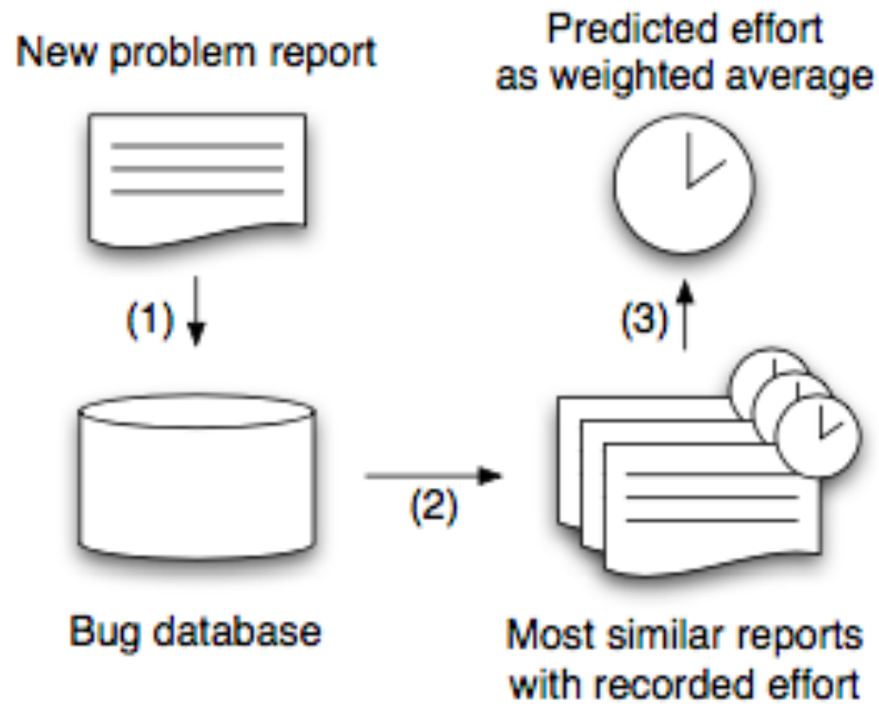
14% of all files that import ui packages, had to be fixed later on.

Schröter, Zimmermann, Zeller. Predicting Component Failures at Design Time. ISESE 2006.

Example source: https://de.slideshare.net/taoxiease/software-mining-and-software-datasets

How Long will it Take to Fix this Bug?

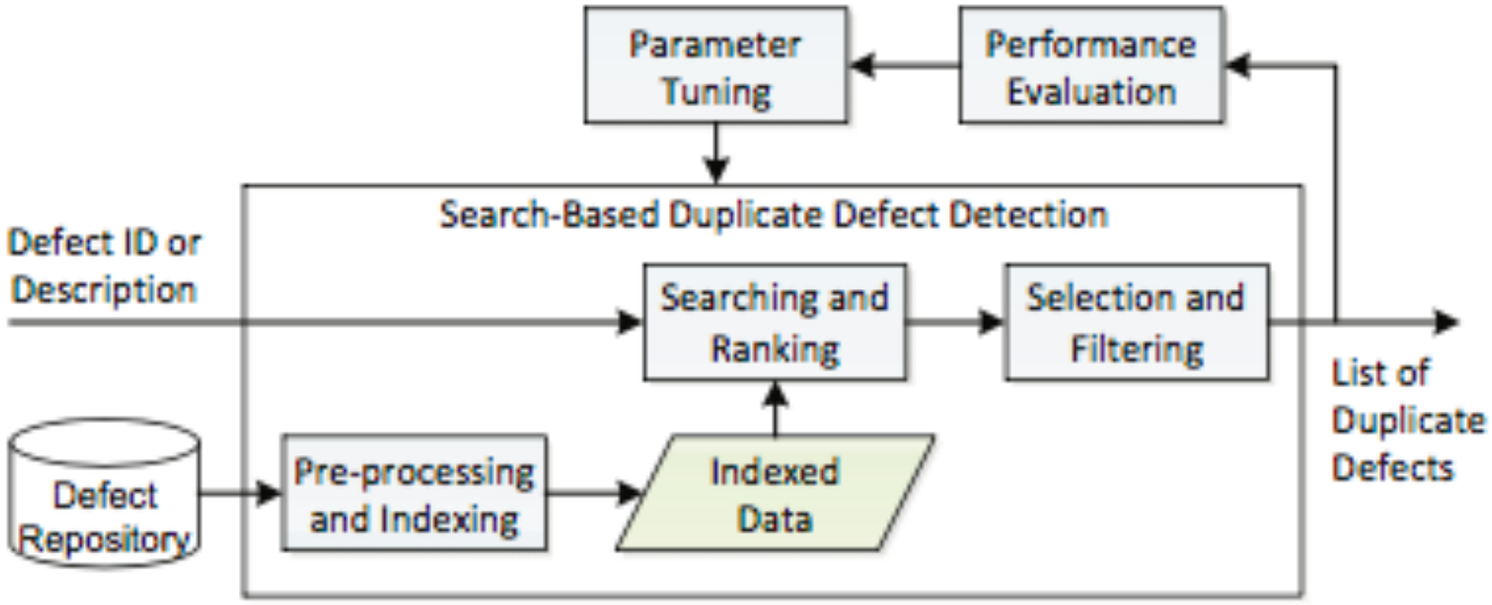
- Predicting effort to fix a bug
- Mine bug databases
- Text similarity to identify reports closely related



How Long will it Take to Fix This Bug? C. WeiB, R. Premraj, T. Zimmermann, A. Zeller. (MSR' 07)

Can we identify duplicate bug reports?

- Mine bug repositories (e.g., Bugzilla, Jira)
- Use information retrieval to find similar reports and rank them.



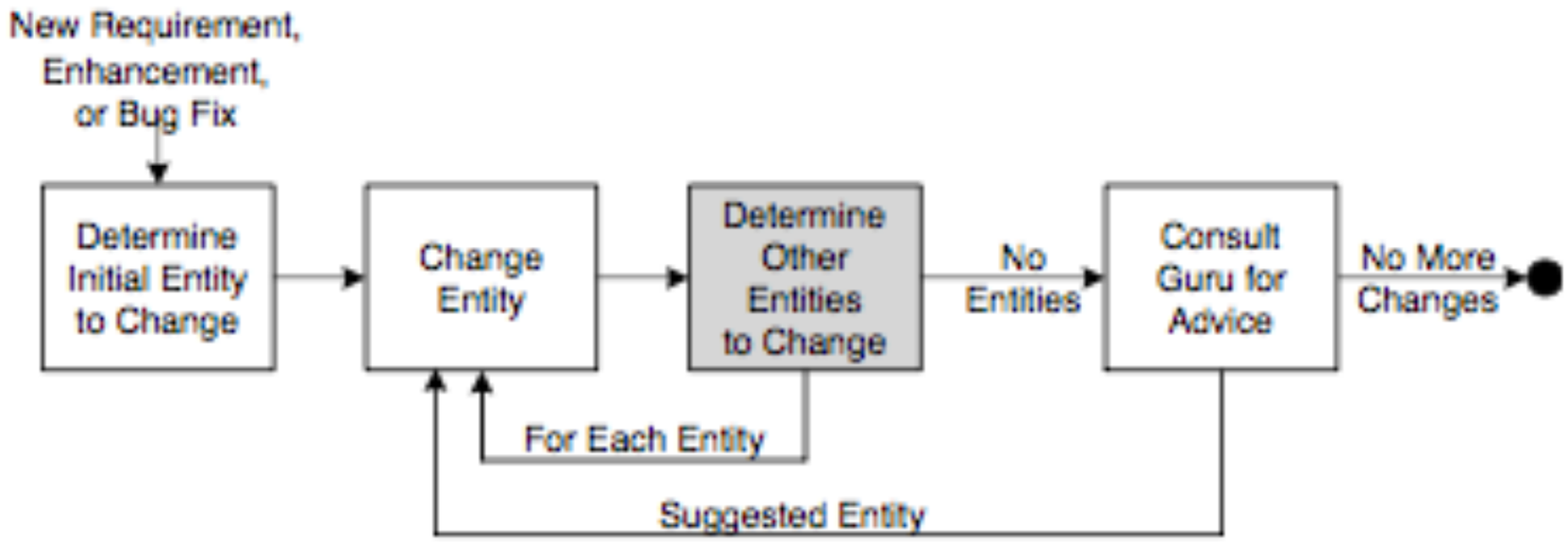
The Duplicate Defect Detection (DDD) Framework

Search-Based Duplicate Defect Detection: An Industrial Experience. Amoui, M., Kaushik, N., Al-Dabbagh, A., Tahvildari, L., Li, S., & Liu, W. (MSR'13)

Change Propagation

How does a change in one source code entity propagate to other entities?

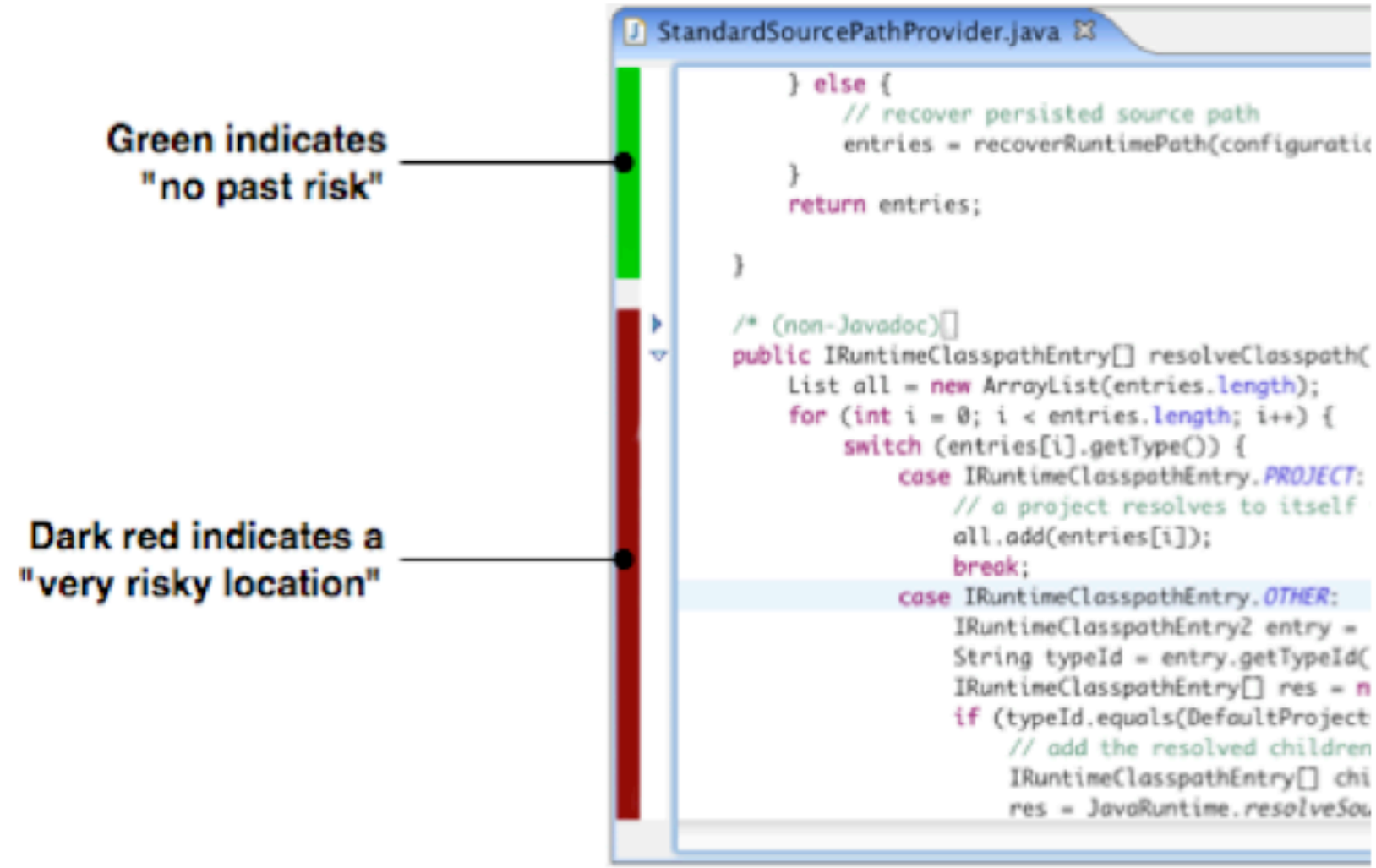
- Predict change propagation
- Mine association rules from change history



Predicting Change Propagation in Software Systems. Ahmed E. Hassan and Richard C. Holt (ICSM '04)

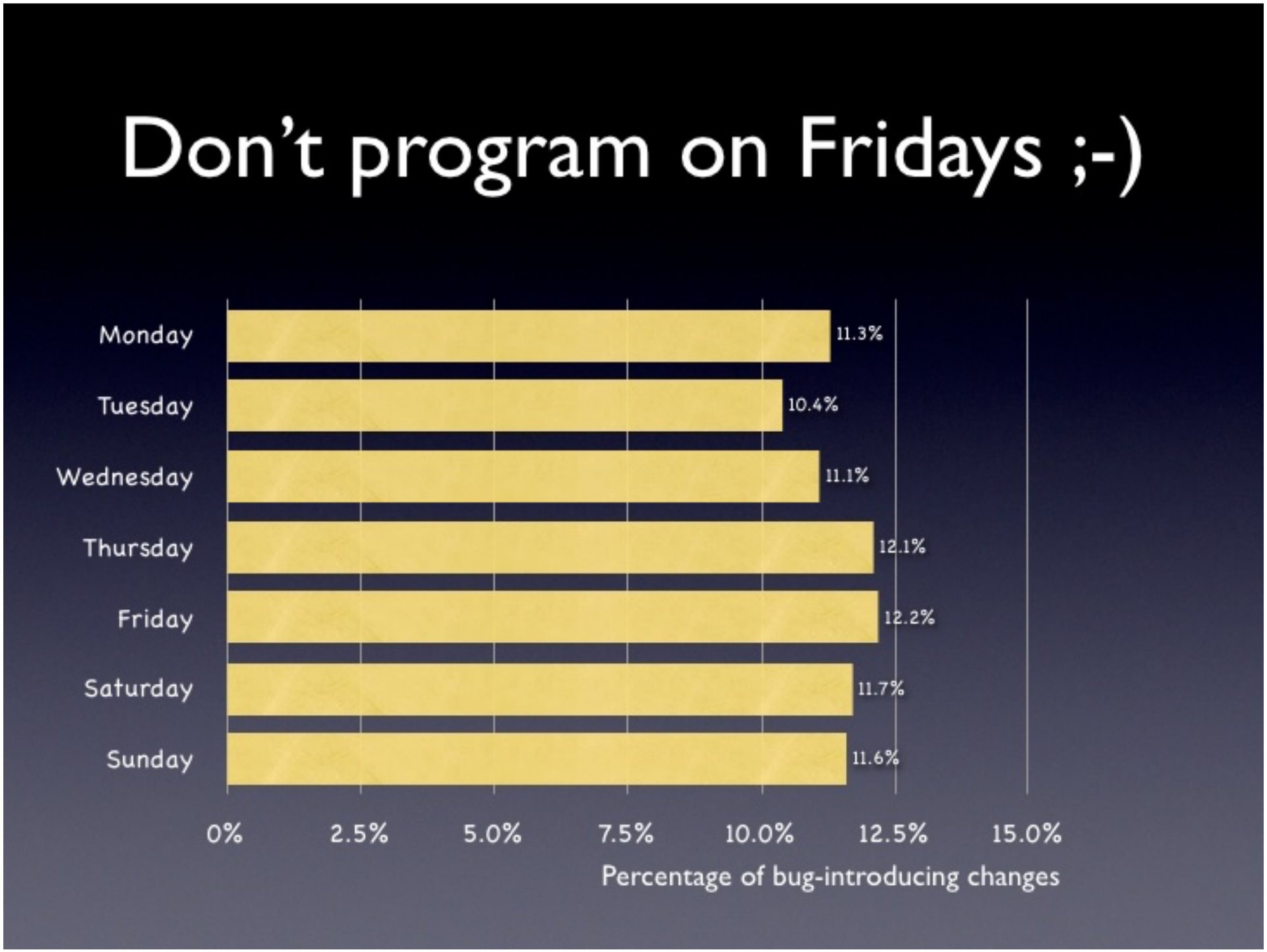
Classify Changes as Buggy or Clean

- Can we warn developers that there is a bug in a change”?
- Identifying bug-introducing changes from bug-fix data



Automatic Identification of Bug-Introducing Changes. Kim, S., Zimmermann, T., Pan, K., & James Jr, E. (ASE’ 06)

Classify Changes as Buggy or Clean



Automatic Identification of Bug-Introducing Changes. Kim, S., Zimmermann, T., Pan, K., & James Jr, E. (ASE’ 06)

Classification of security bug reports

Document Classification: (Non)Security Bug Reports

Term-by-document frequency matrix quantifies a document

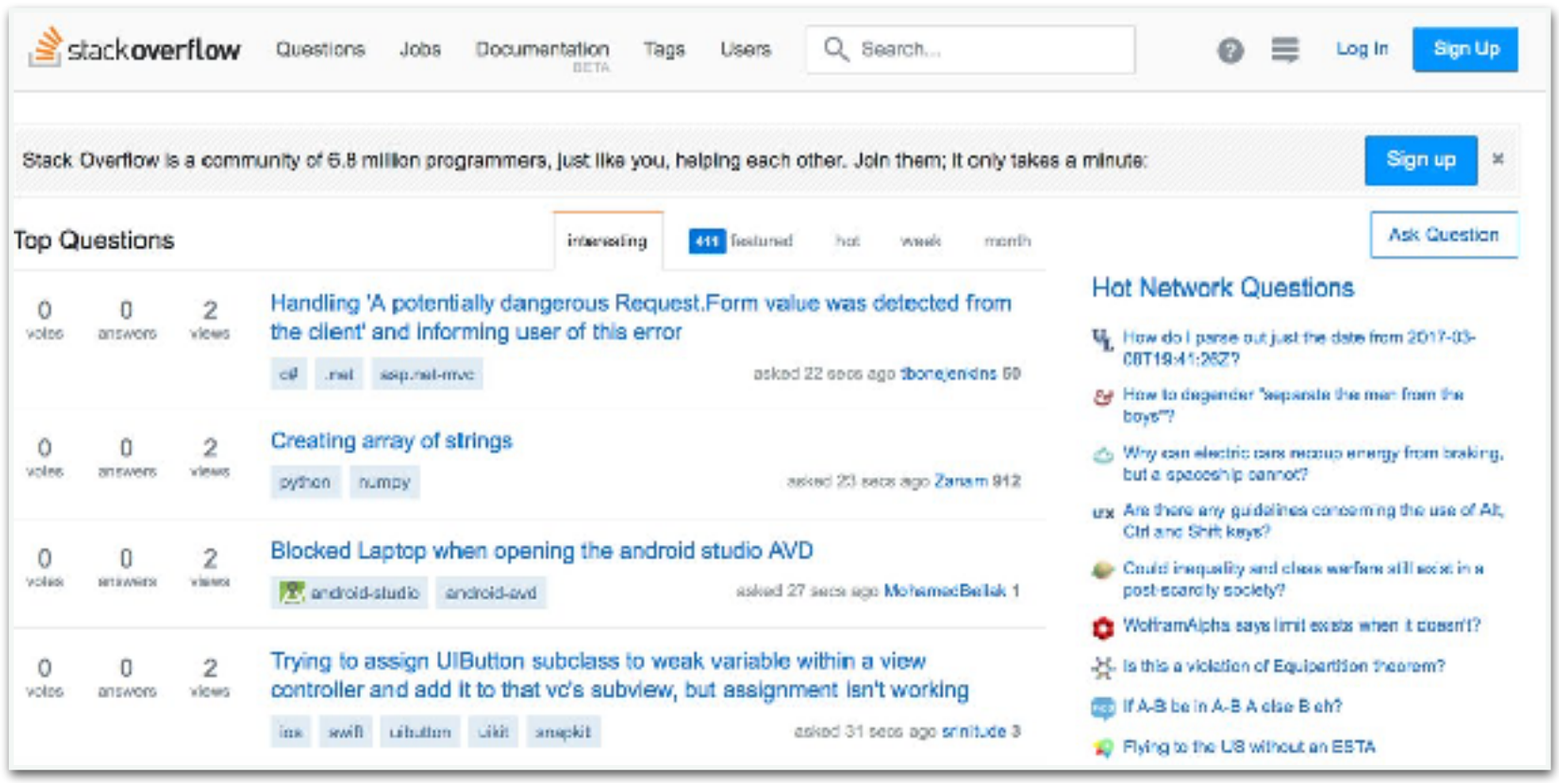
Term *Start List*	Bug Report 1	Bug Report 2	Bug Report 3
Attack	1	0	1
Buffer Overflow	1	0	0
Vulnerability	3	0	0
…			
	Label: Security	**Label: Non-Security**	**Label: ?**

M. Gegick, P. Rotella, T. Xie. Identifying Security Bug Reports via Text Mining: An Industrial Case Study. MSR'10

Example source: https://de.slideshare.net/taoxiease/software-mining-and-software-datasets

Mining questions about software energy consumption

- Mine communities (StackOverflow)
- Use thematic analysis (e.g. LDA, Classifier) to find common themes in questions & answers
- Interpret themes



Mining questions about software energy consumption. Pinto, G., Castor, F., & Liu, Y. D. (MSR' 14)

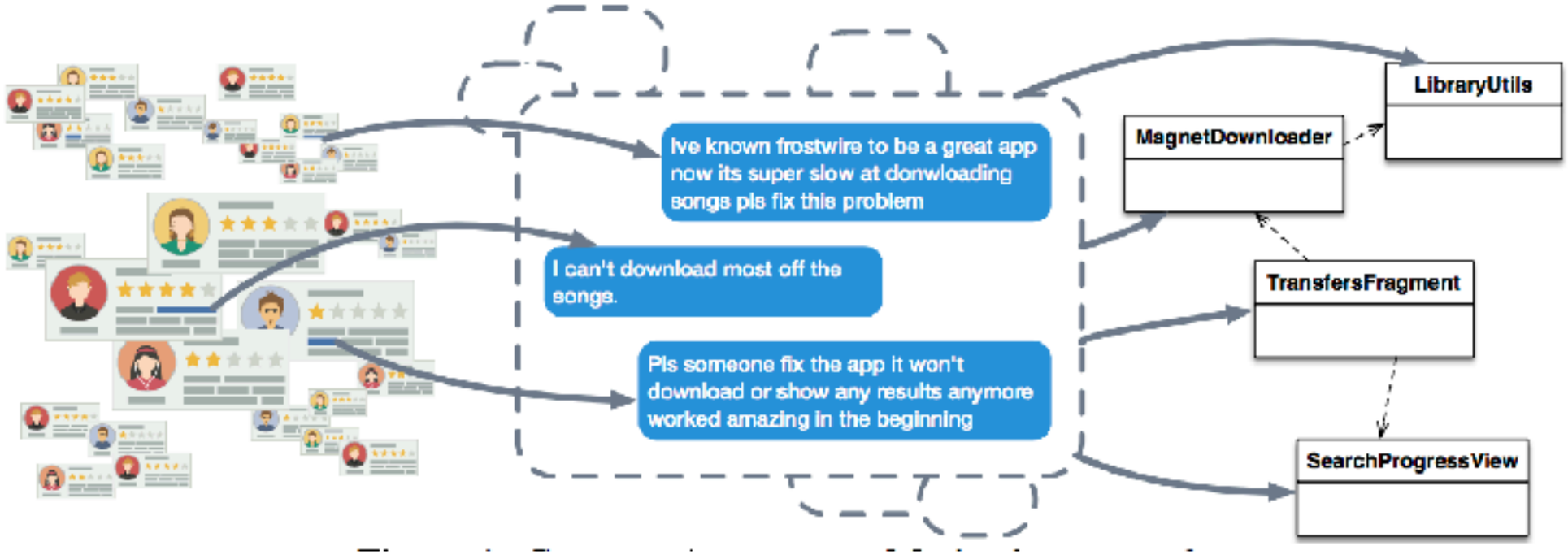
API change and fault proneness impact success

- Relationship between success of Android apps and Android API instability
- Measure success through user ratings in app store
- Measure fault-proneness through number of bugs fixed in the used APIs

API change and fault proneness: a threat to the success of Android apps. M. Linares et al. (FSE'13)

Recommending and Localizing Change Requests for Mobile Apps based on User Reviews

- Automatic classification of user reviews from Google Play store
- Link to the source code entities to be changed
- Recommend developers changes to sw artefacts



Recommending and Localizing Change Requests for Mobile Apps based on User Reviews. F. Palomba et. al. (ICSE'17)

MSR in Practice

Slide extracted from: https://de.slideshare.net/taoxiease/software-mining-and-software-datasets

Tools for Mining Software Repositories

- Available mining tools
  - Libresoft Tools. http://tools.libresoft.es/
  - CVSAnaly. VS/SVN/Git repository log parser
  - MLStats. Mailman and Mboxes parser
  - Bicho. Bugzilla and SF.net tracker parser

MSR Repositories

Data Repositories available online:

- FLOSSmole repository of open source snapshots. flossmole.org/
- Github. http://www.ghtorrent.org
- iBUGS. www.st.cs.uni-saarland.de/ibugs/
- MetricsGrimoire toolset. https://metricsgrimoire.github.io
- PROMISE repository. http://openscience.us/repo/
- Software-artifact Infrastructure Repository. http://sir.unl.edu/portal/index.php
- Ultimate Debian Database. https://wiki.debian.org/UltimateDebianDatabase
- Apache SVN commits. https://github.com/monperrus/apache-svn-commits
- Socorro: Mozilla Crash Stats. https://wiki.mozilla.org/Socorro

References

- The International Conference on Mining Software Repositories. 2017.msrconf.org

- Mining Software Engineering Data. Ahmed E. Hassan & Tao Xie.

- The Road Ahead for Mining Software Repositories. Ahmed E. Hassan

- Software Intelligence: The Future of Mining Software Engineering Data. Ahmed E. Hassan & Tao Xie.

- Effective Mining of Software Repositories. M. D'Ambros & Romain Robbes.