

# Mining Software Repositories with Topic Models

Stephen W. Thomas (sthomas@cs.queensu.ca)

Queen's University, Canada



## 1. Problems to Address

What is the **design rationale** behind this code?

What high-level **topics** are developers interested in, and how have they been **changing**?

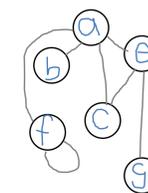


Mining software repositories can help:

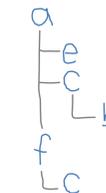
- Developer thoughts and intentions are encoded in **source code identifiers and comments**, and
- discussions about code design is contained in **email archives**.

Both of these repositories are *unstructured*.

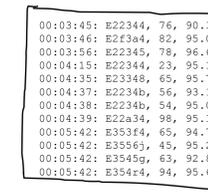
## 2. Challenges with Unstructured Data



Static call graphs

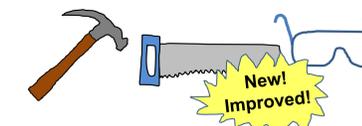


Stack traces

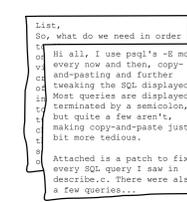


Execution logs

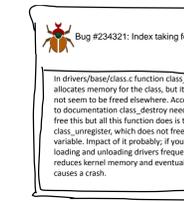
Need better methodologies and techniques to use (exploit!) unstructured data.



Code comments and identifiers



Email archives



Bug reports

Unstructured data.

Challenges:

- Natural language and noise
- Huge volumes of data, evolving over time
- No explicit links between data

## 3. Existing (Advanced) Approach to Exploit Unstructured Data

Use statistical **topic models**, from NLP, to structure and analyze the unstructured data.

**BONUS!** Topic models scale and are meta-model agnostic (i.e., work on any text documents).

**WARNING!** Topics models to designed for use on source code (Thomas et al. 2011).

**Latent Dirichlet Allocation (LDA)** Unsupervised; automatic; discovers *topics* from any corpus

The X-Files is an **American** science fiction **television series** and a part of The X-Files franchise, created by screenwriter Chris Carter. The program originally **aired** from September 10, 1993 to May 19, 2002. The **show** was a hit for the Fox ...

The Indianapolis Colts are a professional **football team** based in Indiana. The **quarterback**, Peyton Manning, has the lowest **tackle** percentage in league at only 0.3 **tackles** per game. He has the **league MVP** award 75 times and can lift a car with ...

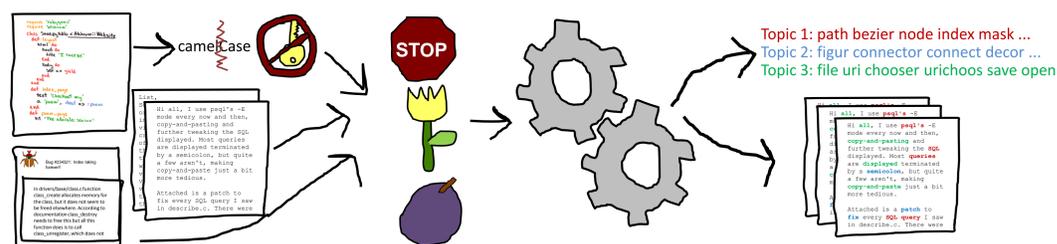
Married... with Children is a sitcom about a dysfunctional family living in Chicago that **aired** for 11 seasons. The **show**, notable for being the first prime time **television series** to air on Fox, ran from April 5, 1987, to ...

Rioja is made from **grapes** grown not only in the Autonomous Community of La Rioja, but also in parts of Navarre and the Basque province ... prohibited the passing of carts through streets near **wine cellars**, in case...

- Topic 1: **american, television, show, series, aired, ...**  
 Topic 2: **football, quarterback, touchdown, tackle, ...**  
 Topic 3: **wine, grape, taste, cellar, spain, glass, ...**

Traditional applications of LDA Clustering; IR; concept location

Applying to code, emails, and bugs split identifiers; remove programming keywords; stop; stem; prune



## 4. Software Engineering Applications

**Concept location and concept evolution** (In progress)

*Method:* Use LDA to discover topics and their evolutions

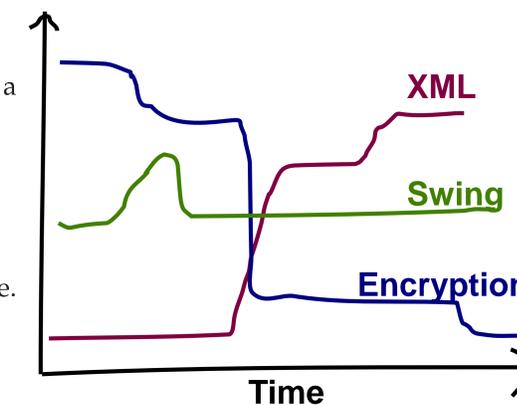
*Benefits:* Allows practitioners to monitor and reason about code at a higher level:

- Code refactoring and understanding
- Project monitoring
- Acquisition support

*Contribution:* Prior work did not consider characteristics of source code.

Our *Diff model* accounts for source code characteristics, resulting in:

- Better topics
- More sensitive and accurate topic evolutions



**Source-email traceability link recovery** (In progress)

*Method:* Link engine based on keywords (Bacchelli et al. 2010) and topics

*Benefits:* Allows developers to locate design discussions of code

- Understand design rationale
- Recover the flow of knowledge

*Contribution:* Prior work was based on keywords only. Our work also considers higher-level concepts, resulting in better traceability links.

For more information: <http://cs.queensu.ca/~sthomas>