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.

1. Problems to Address

2. Challenges with Unstructured Data

Need better methodologies and techniques to use (exploit!) 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!** Topic models to designed for use on source code (Thomas et al. 2011).

**Contribution:**
- Prior work was based on keywords only. Our work also considers higher-level concepts, resulting in better traceability links.
- Allows practitioners to monitor and reason about code at a higher level:
  - Code refactoring and understanding
  - Project monitoring
  - Acquisition support

**Benefits:**
- Better topics
- More sensitive and accurate topic evolutions

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

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

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

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

4. Software Engineering Applications

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

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

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