

Steps on the path to Automated Process Extraction

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

Introduction

- Abram Hindle
 - Masters at UVic under Dr. Daniel German
 - PhD at UWaterloo under Dr. Michael Godfrey and Dr. Ric Holt
- Research Focus
 - Software Evolution w.r.t Change Repositories
 - * Characterization of Behaviour / Process Extraction

Chronological Overview

- Temporal Queries and Patterns
 - Release Mining
 - Architecture Visualization over Time
 - Release Behaviour Discovery
 - Metrics of Changes
 - Topic Trends
 - Fourier Analysis
 - Process Extraction

Motivation

- Software Engineering is to Software Evolution as a radio is to a clock radio
- Reason about time and the past, and potentially the future
- I care about extracting what happened

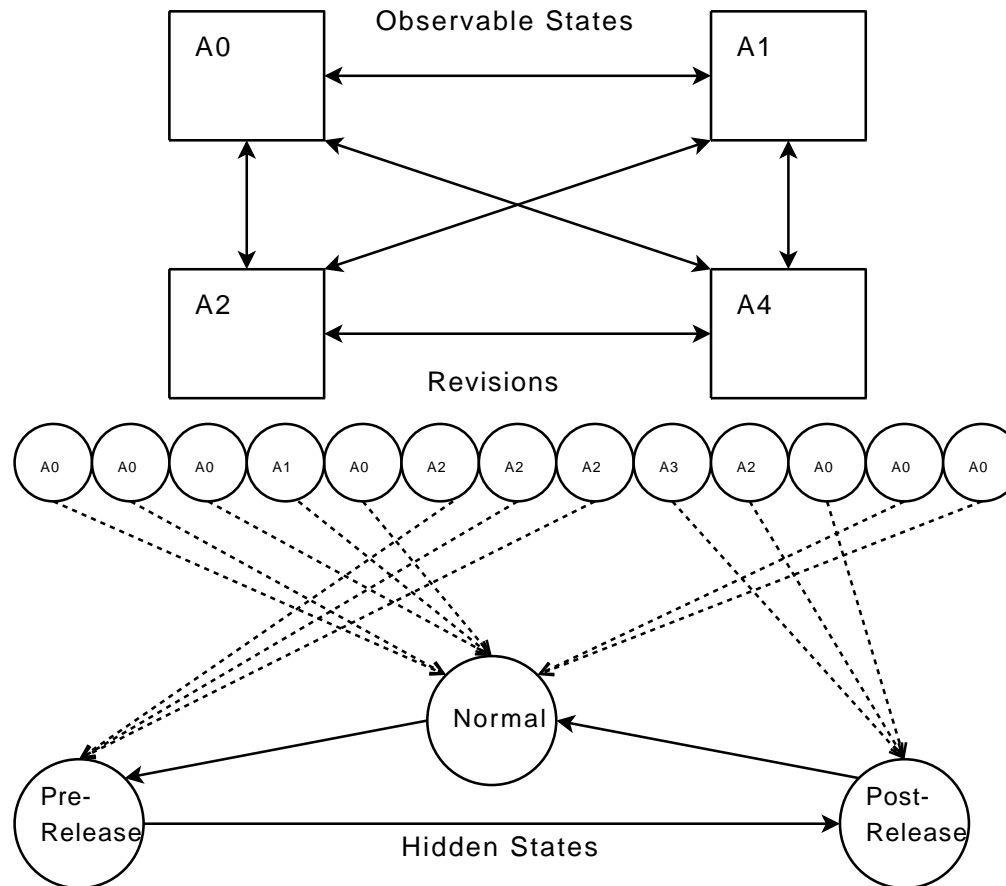
SCQL

- Source Control Query Language
 - Masters work with Dr. Daniel German [HG05]
 - A first order and temporal logic based query language for revisions
 - Similar to Myrung Kim's description inference work. [KNG07]

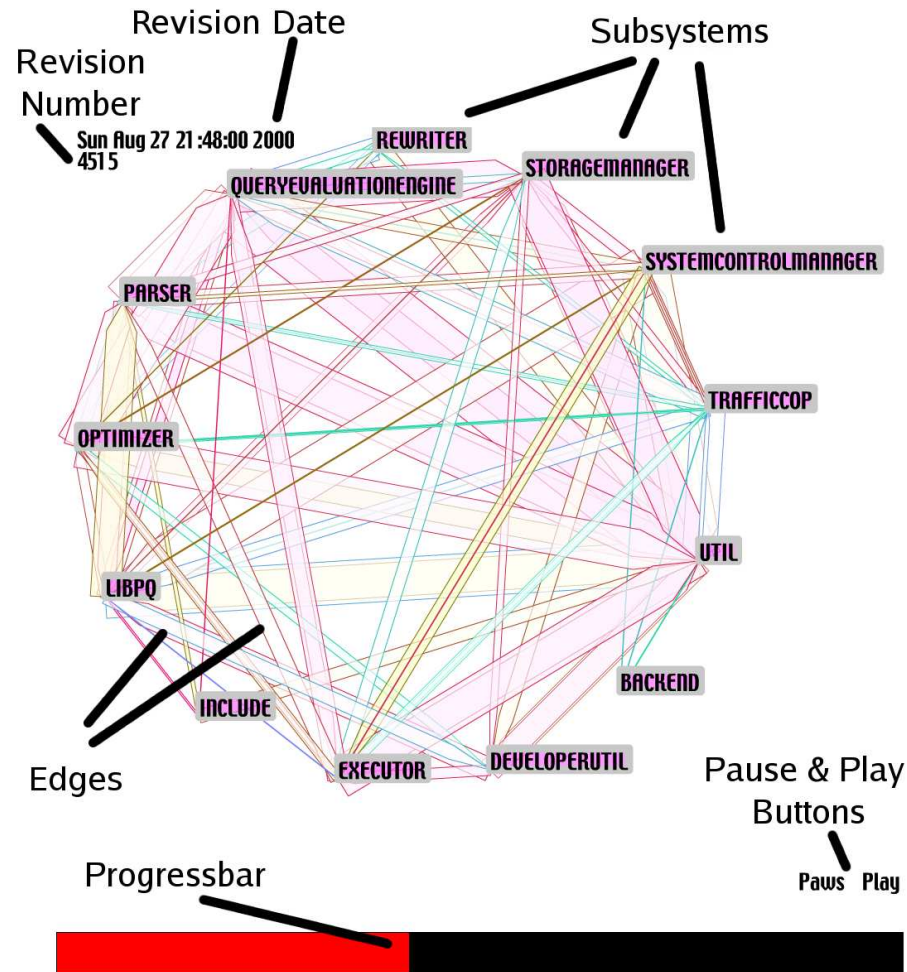
Release Discovery

- Try to find release revisions via
 - Machine Learning,
 - Markov Models,
 - Hidden Markov Models
 - and Markov Decision Processes
- Poor results (low precision and recall 40/40)

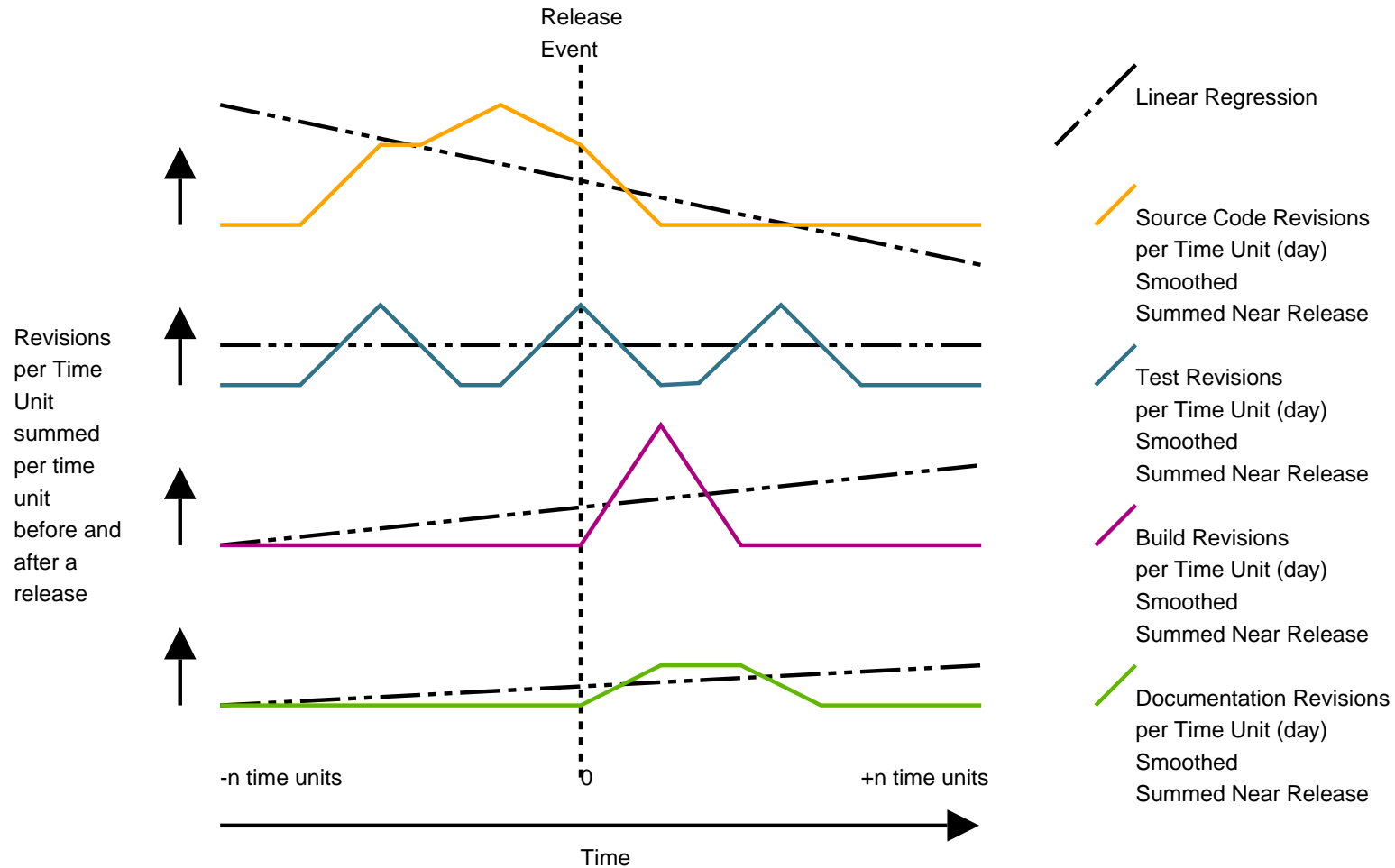
Release Discovery



YARN



Release Patterns



Indentation Metrics

Get the Diff

```
> void square( int * arr, int n ) {
>     int i = 0;
>     for ( i = 0 ; i < n ; i++ ) {
>         arr[ i ] *= arr[ i ];
>     }
> }
```

Measure the Indentation

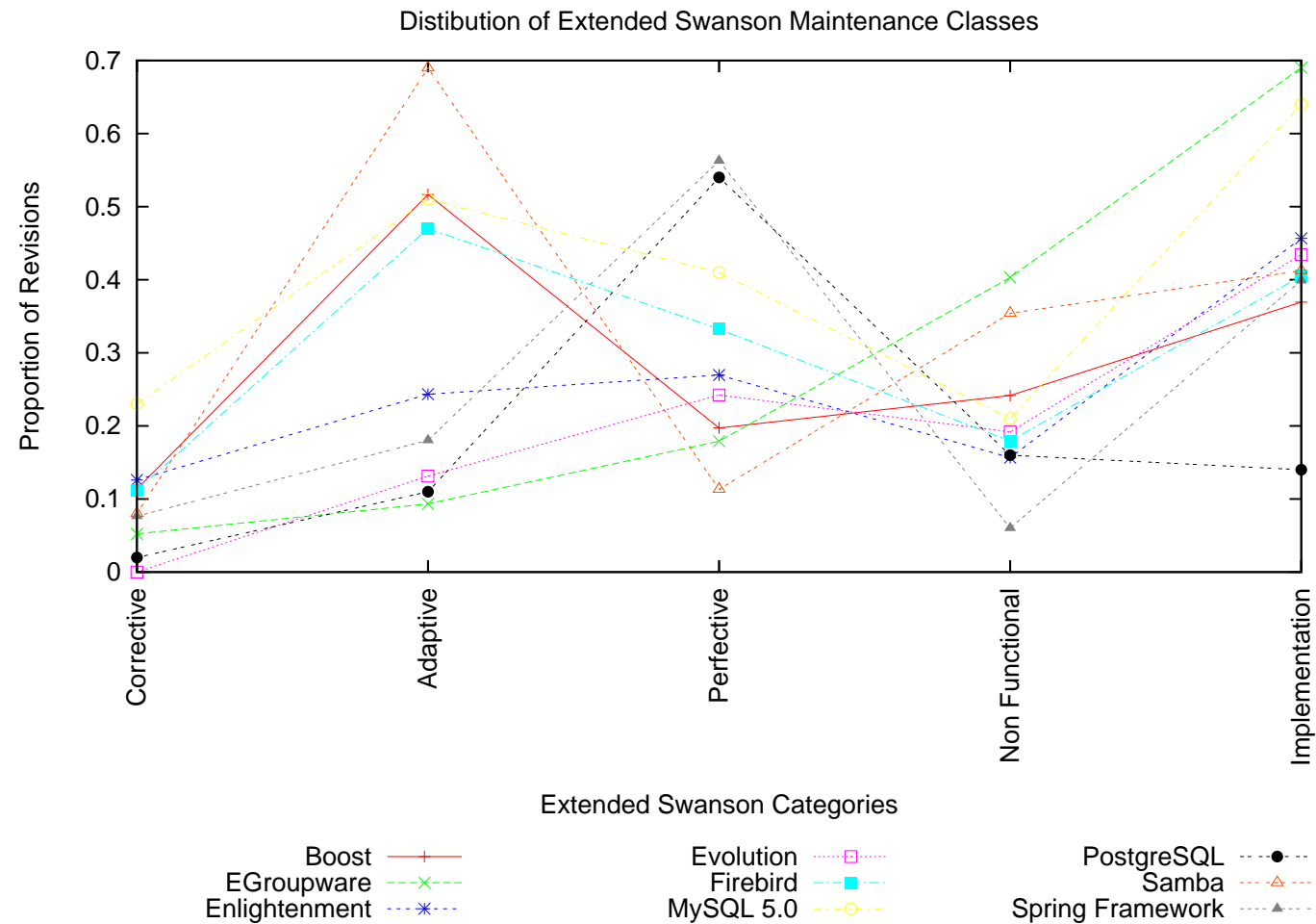
Raw
Indentation
Logical
Indentation

0	4	4	8	4	0
0	1	1	2	1	0

Produce Summary Statistics

Metric	Raw	Logical
LOC	6.000	6.000
AVG	3.330	0.833
MED	4.000	1.000
STD	2.750	0.687
VAR	9.070	0.567
SUM	20.000	5.000
MCC	2.000	2.000
HVOL	152.000	152.000
HDIFF	15.000	15.000
HEFFORT	2127.000	2127.000

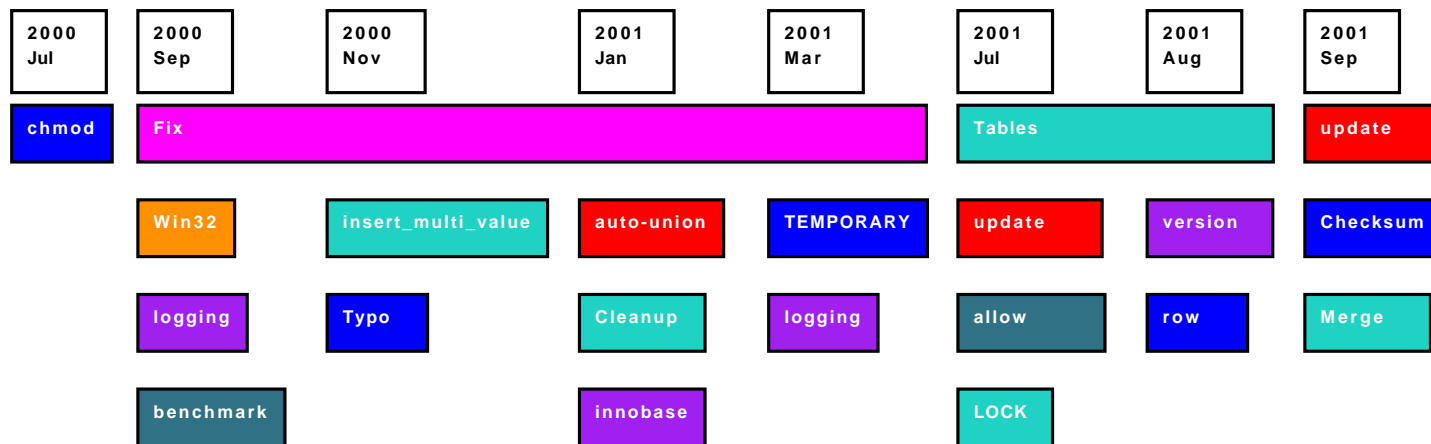
Large Changes



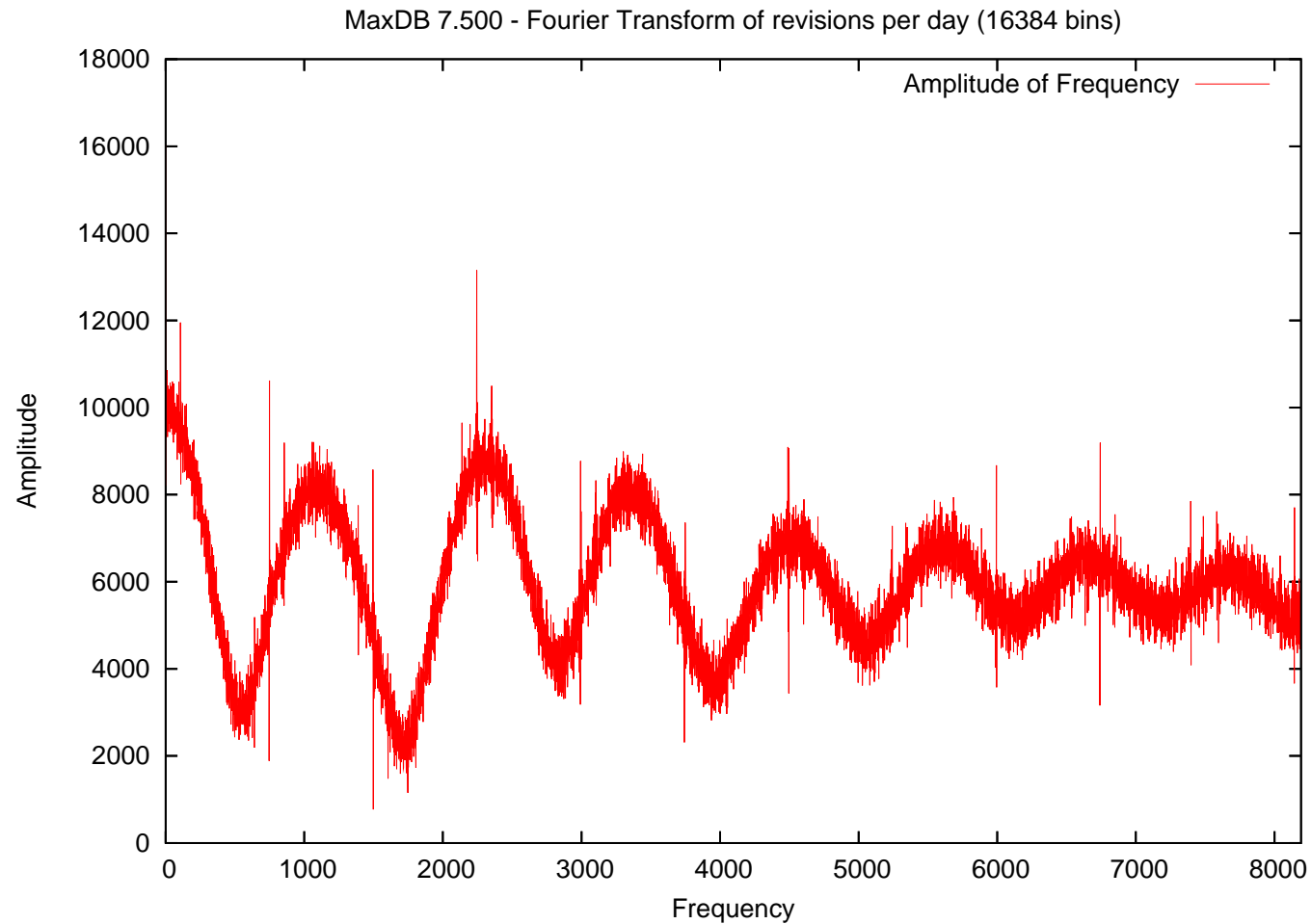
The Present and Future

- That's what had been accepted and submitted (or work in progress)
- The future is extending this work for Process Extraction

Topic Trends

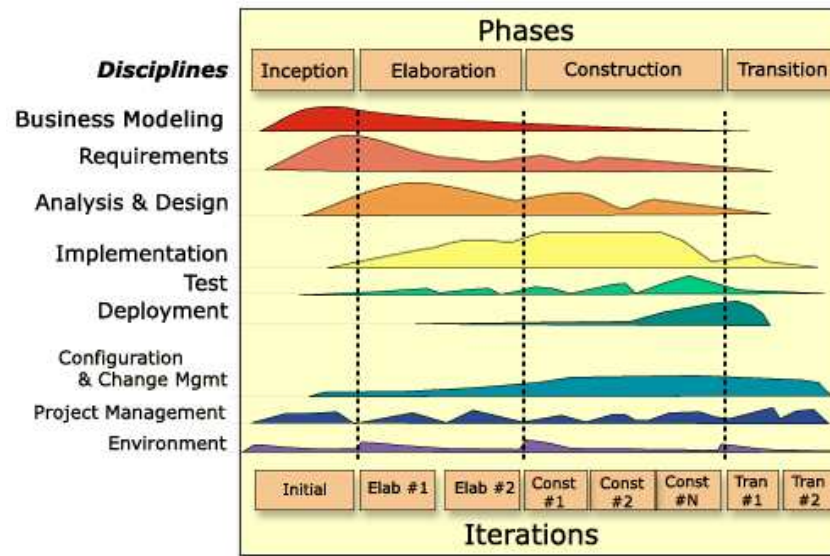


Fourier Transform



Process Extraction

- Combine past techniques and tools
- Partition time by iteration/sub iteration



— (c) 2003 Rational Software Corporation

What have we learned so far

- Statistical measures have told us that projects show a great deal of individuality
 - Results difficult to interpret
- We get better information through manual study rather than automated means
- Repository Behaviour is probably different than the development process

Shock! Controversy!

- Empirical Software Engineering is a sham!
 - We learn little
 - Results are not repeatable across different systems
- True empirical verification of SE might not be feasible beyond a certain size of program.
 - Empirical studies like these don't scale
 - * but descriptive case studies provide lots of wisdom

Conclusions

- Work so far has focussed on resolving facts from source control repositories
- Characterize the behaviour of actors
 - Extract and characterize the software process
- Goal: Automated and manual methods of reliably and accurately describing the software process of a project

References

- [HG05] Abram Hindle and Daniel M. German. Scql: a formal model and a query language for source control repositories. In *MSR '05: Proceedings of the 2005 international workshop on Mining software repositories*, pages 1–5, New York, NY, USA, 2005. ACM.
- [KNG07] Miryung Kim, David Notkin, and Dan Grossman. Automatic inference of structural changes for matching across program versions. In *ICSE '07: Proceedings of the 29th International Conference on Software Engineering*, pages 333–343, Washington, DC, USA, 2007. IEEE Computer Society.