

# Modeling the Evolution of Topics in Source Code Histories



**Stephen W.  
Thomas**



**Bram  
Adams**



**Ahmed E.  
Hassan**



**Dorothea  
Blostein**



**SOFTWARE ANALYSIS  
& INTELLIGENCE LAB**

**Microsoft®**

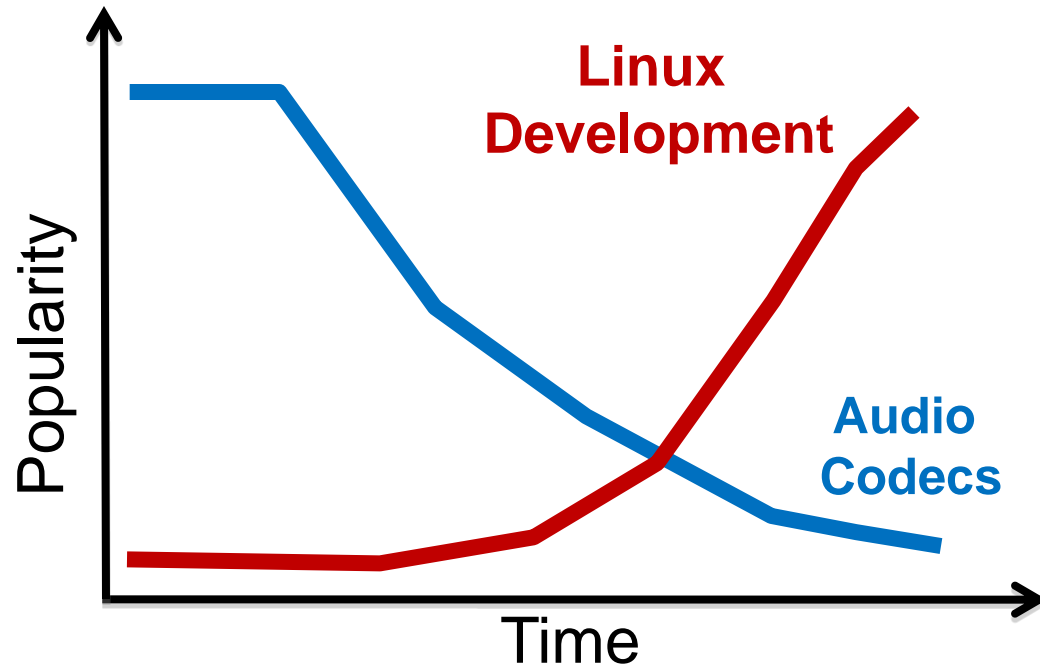
**skype™**



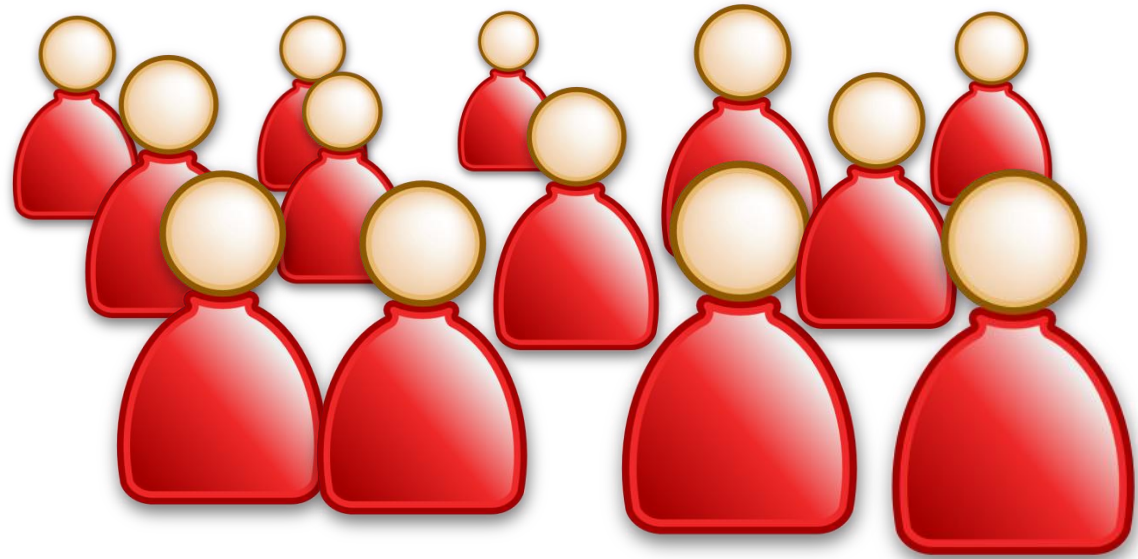
*What have the Skype developers been interested in?*



Microsoft manager

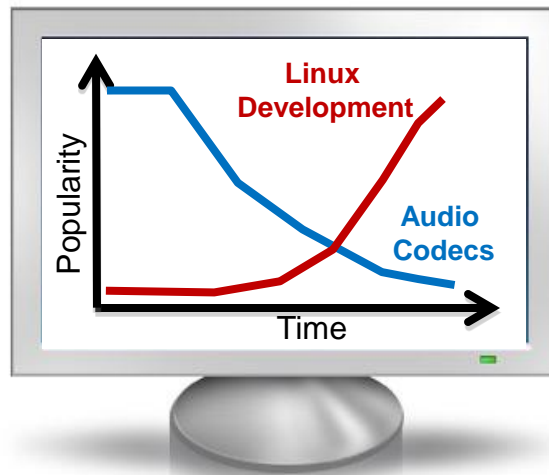


# What are developers working on?



Option 1: Speak with every developer

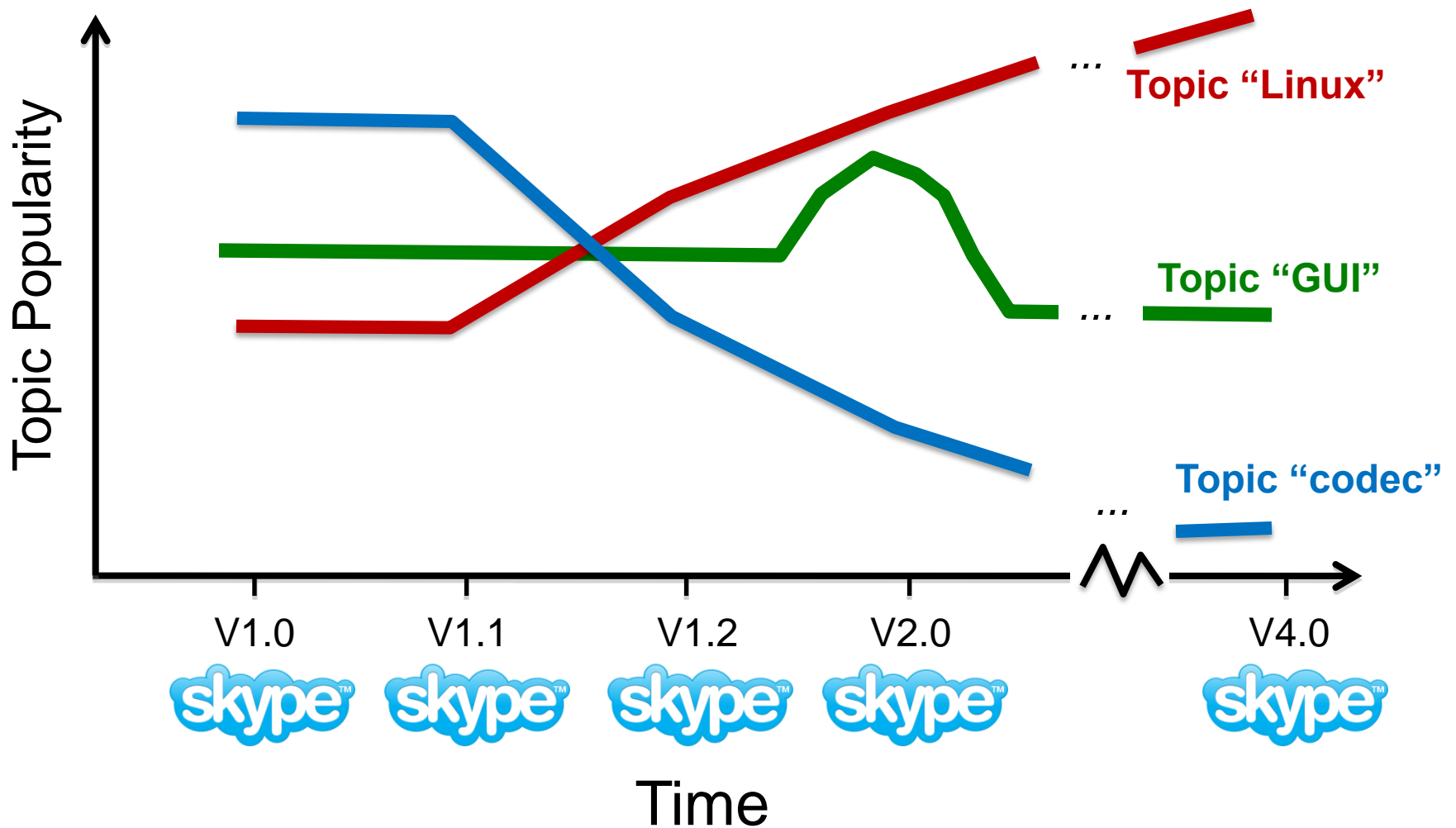
---



Option 2: Use automated tool

# Tool: Topic Evolution Models

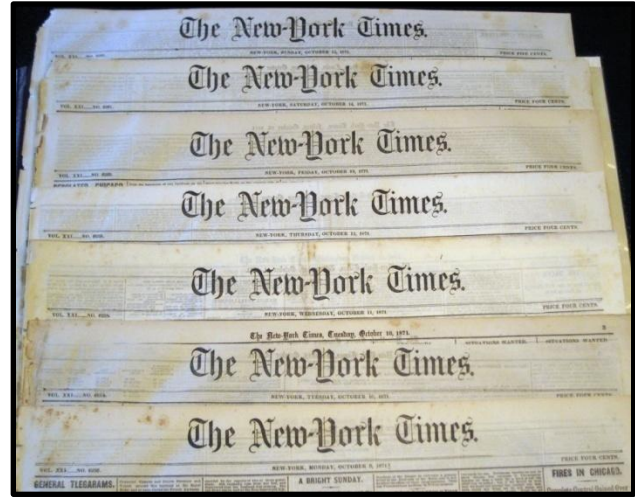
Applied to Source Code Histories



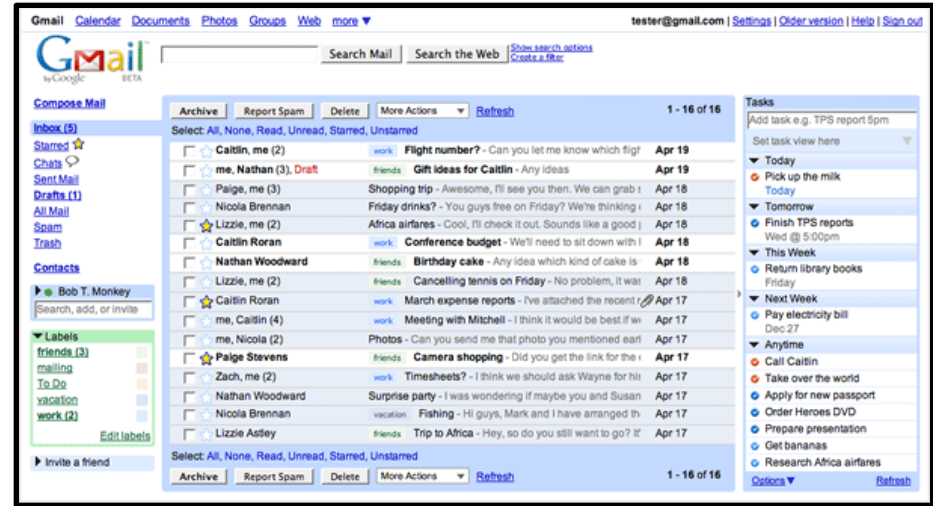
# Success in Other Domains



Conference Proceedings



Newspaper Articles

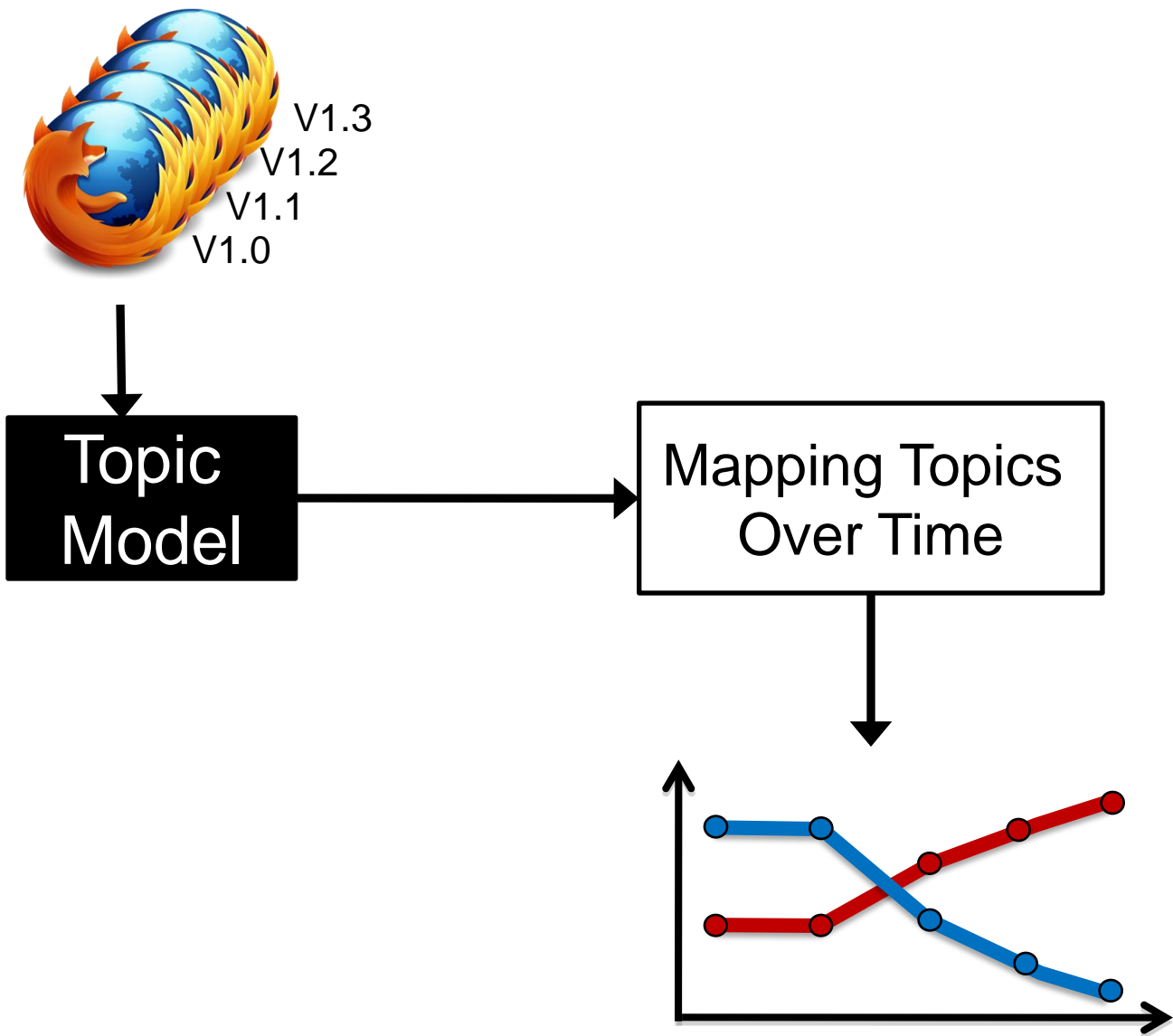


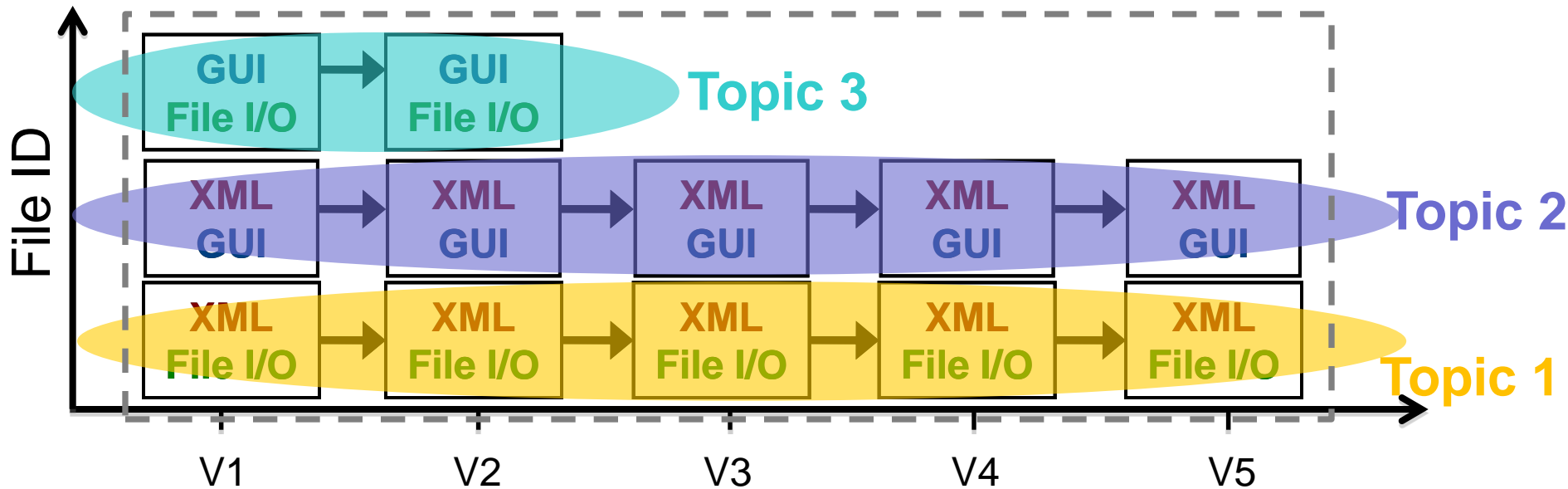
Email Archives

# Topic Evolution on Source Code



# Background: The Hall Model





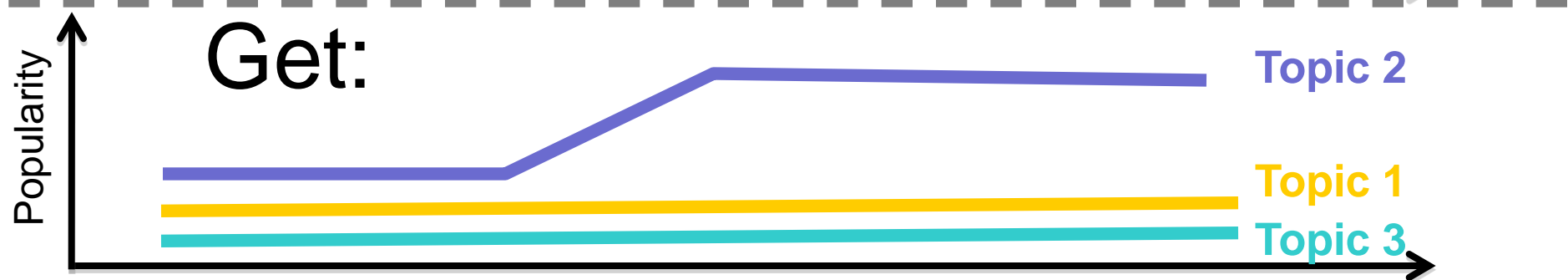
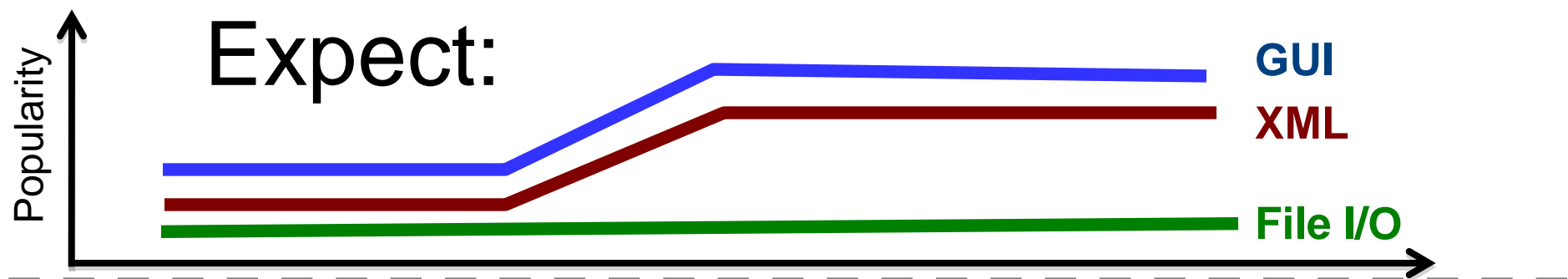
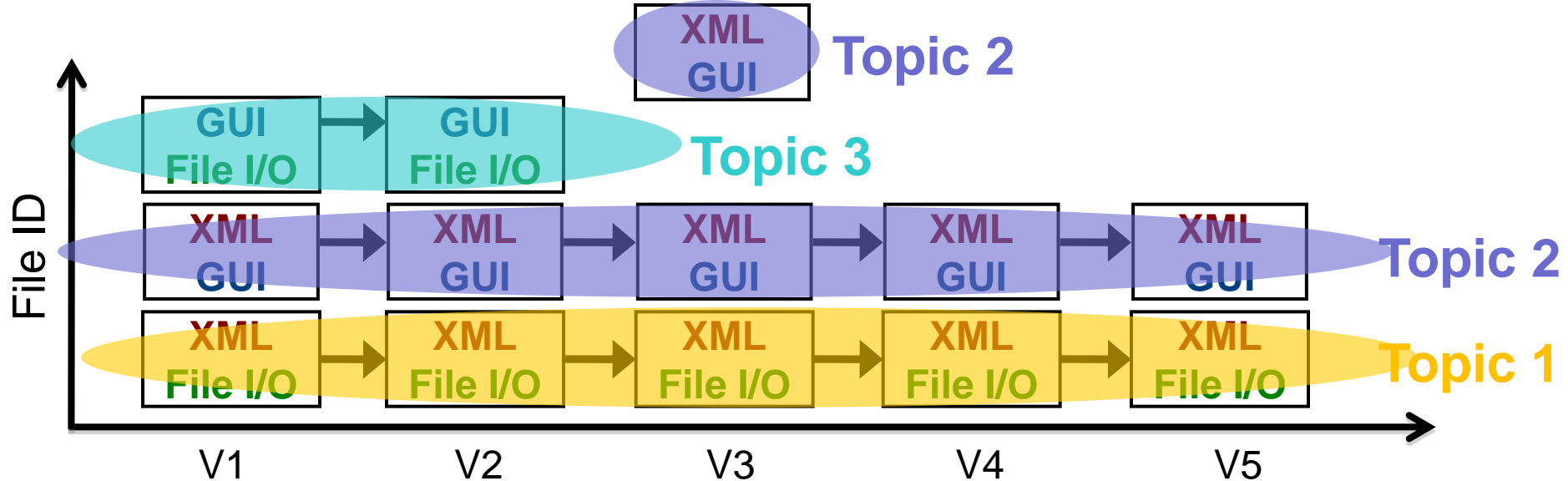
**Expect:**

- Topic 1: XML**
- Topic 2: GUI**
- Topic 3: File I/O**

**Get:**

- Topic 1: XML+ File I/O**
- Topic 2: XML + GUI**
- Topic 3: GUI+ File I/O**

**Problem: Topics are muddled, not distinct**



**Problem: Evolutions not sensitive or accurate**

# Problems due to duplication

Found in Source Code Histories



Topics are muddled, not distinct



Evolutions are not accurate

# Real-World Duplication

63% files don't change

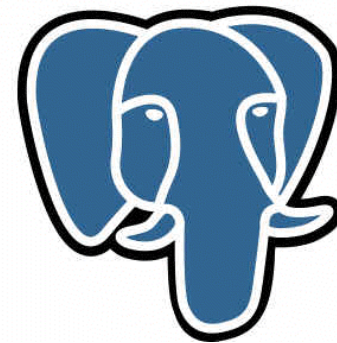
99.8% words don't change



84% files don't change

99.8% words don't change

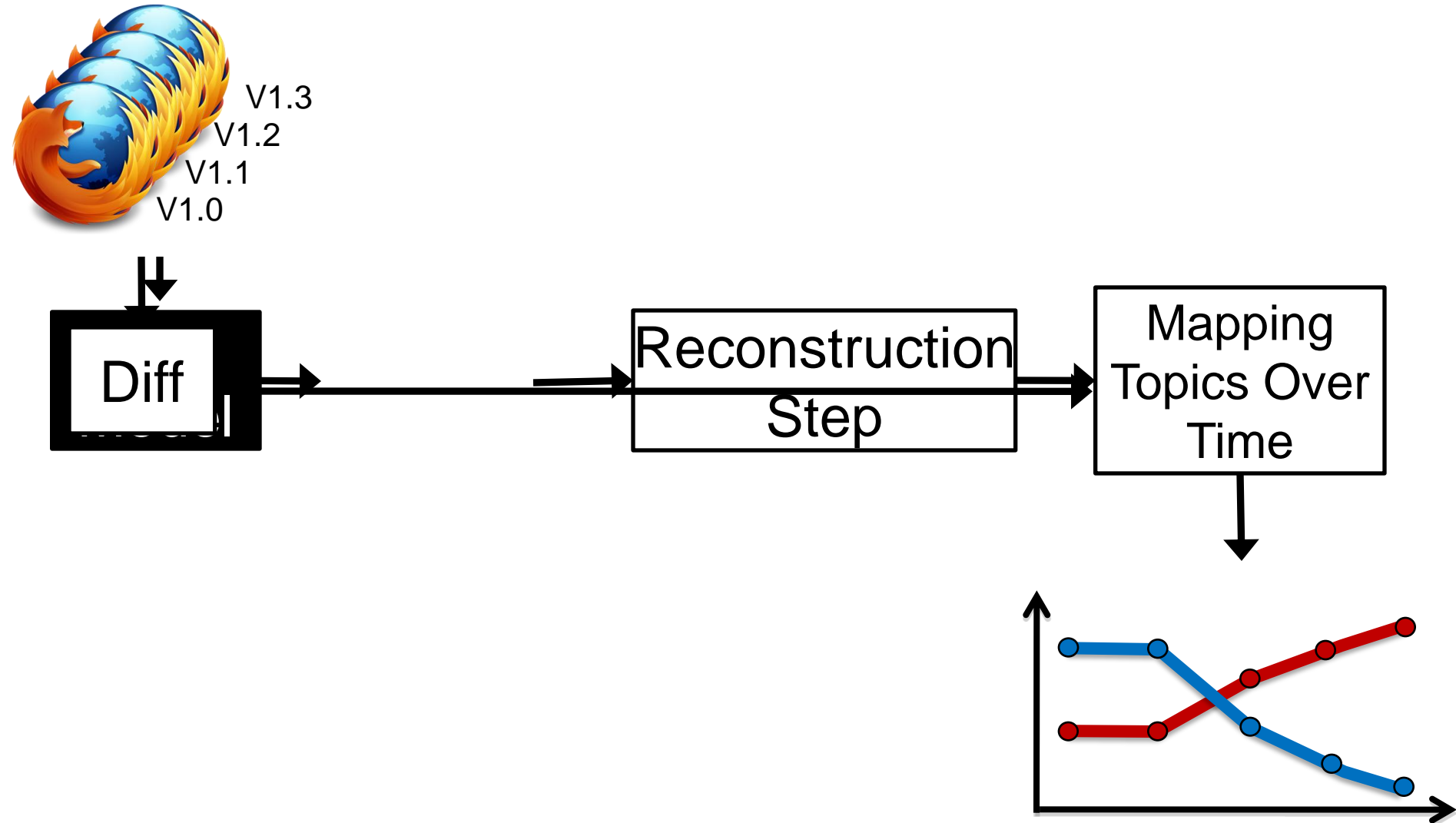
PostgreSQL





# The Diff Model

# The Diff Model



# Diff Step

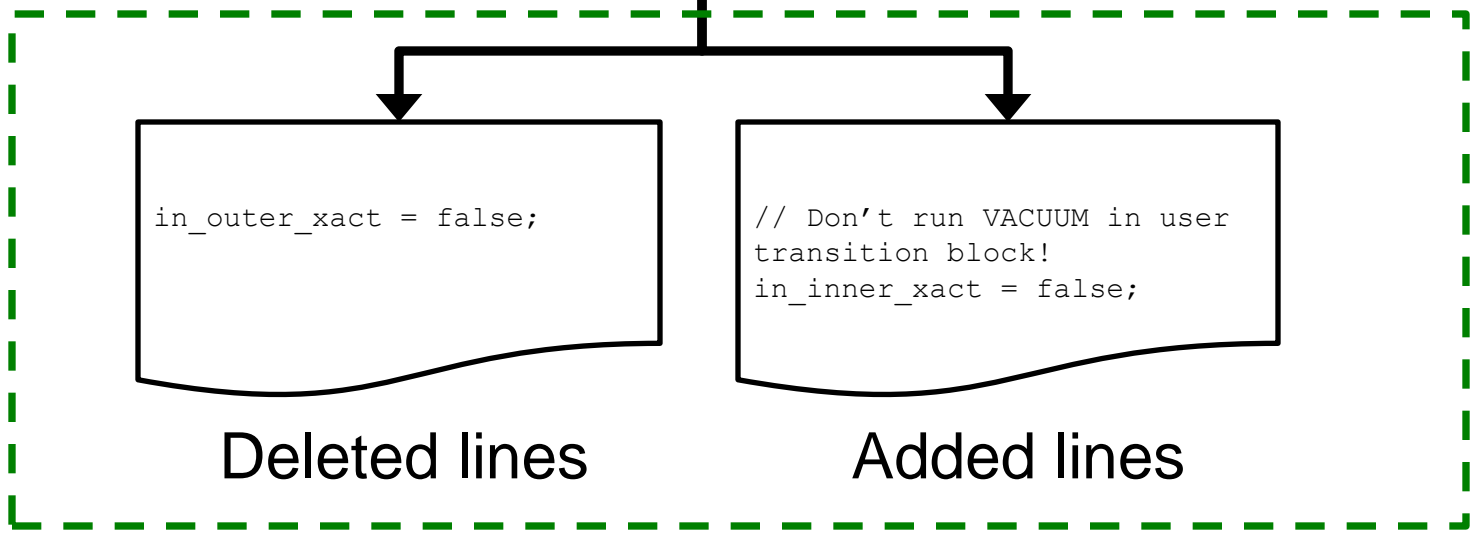
```
...  
if (vacstmt->options & VACOPT_VACUUM){  
  PreventTransactionChain(isTopLevel, stmttype);  
  in_outer_xact = false;  
}  
...
```

Version 5.3.7

```
...  
// Don't run VACUUM in user transition block!  
if (vacstmt->options & VACOPT_VACUUM){  
  PreventTransactionChain(isTopLevel, stmttype);  
  in_inner_xact = false;  
}  
...
```

Version 5.3.8

**Diff**



```
in_outer_xact = false;
```

Deleted lines

```
// Don't run VACUUM in user  
transition block!  
in_inner_xact = false;
```

Added lines

# Reconstructing Topic Memberships

Topic Model

Topic Model

Topic Model

Infer

GUI (90%)  
XML (10%)

GUI (100%)  
XML (0%)

GUI (20%)  
XML (80%)

GUI (77%)  
XML (23%)

First Version  
(1000 lines)

Deleted Lines  
(200 lines)

Added Lines  
(150 lines)

Second Version  
(950 lines)

$$(1000 * 90\%) - (200 * 100\%) + (150 * 20\%) = 730$$

$$(1000 * 10\%) - (200 * 0\%) + (150 * 80\%) = 220$$

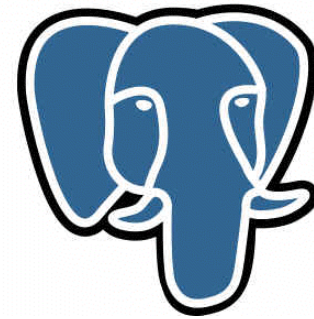
# Case Studies



Drawing Application Framework  
(Java)

13 releases (5.2.0 – 7.5.1)  
613 files  
84K SLOC

PostgreSQL



Database Management System  
(C)

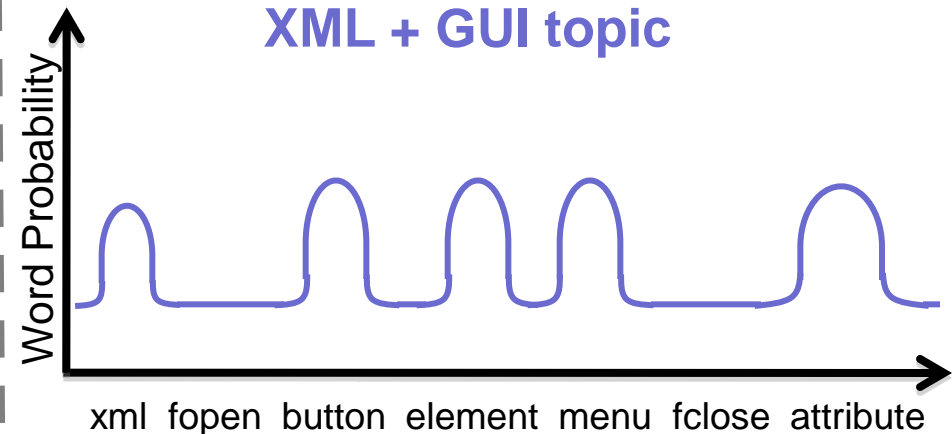
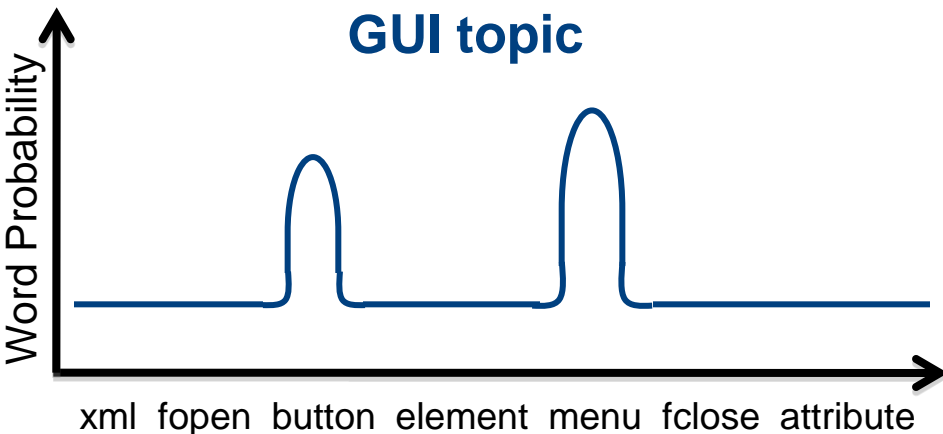
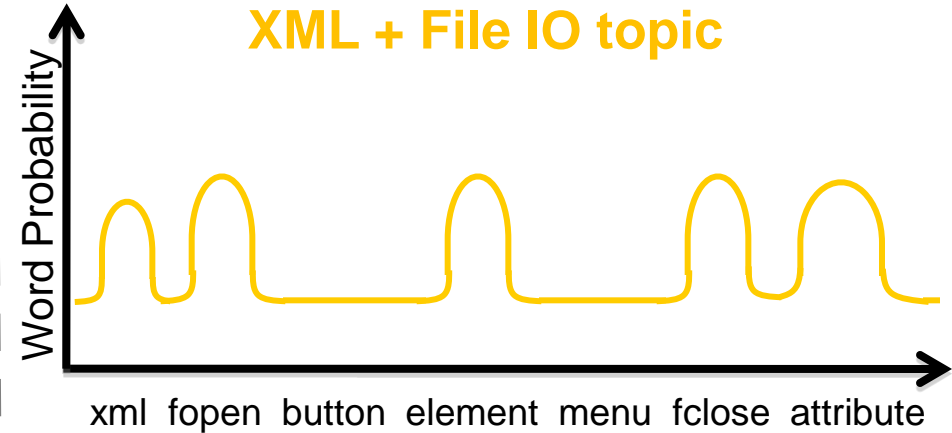
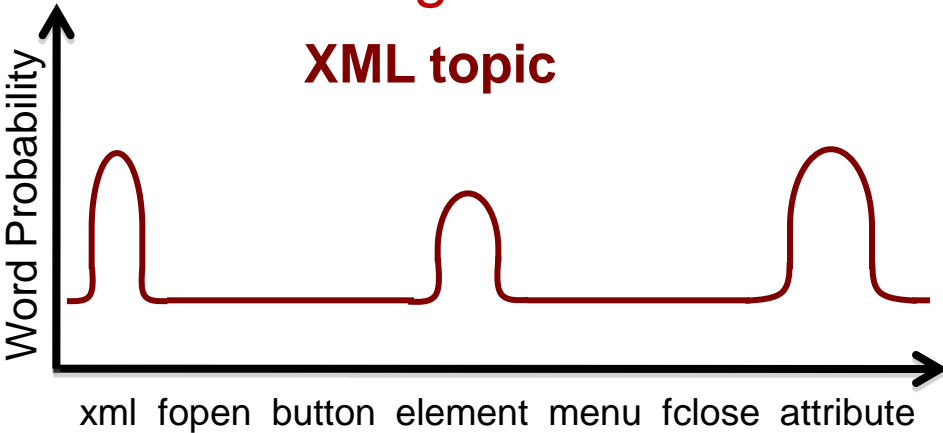
46 releases (7.0.0 – 8.3.5)  
844 files  
501K SLOC

*I bet the Diff model  
discovers topics that  
are more **distinct!***



# Measuring Distinctness

With KL-Divergence

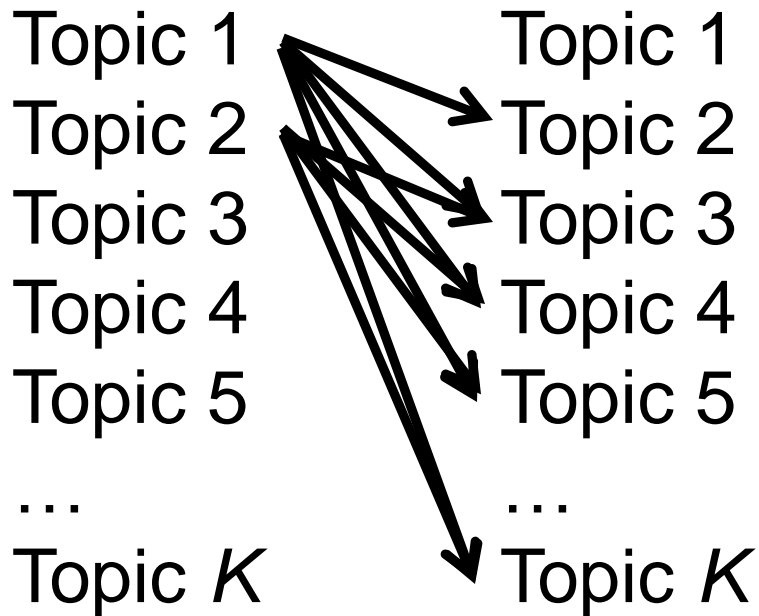


**High KL divergence**  
**High distinctness**

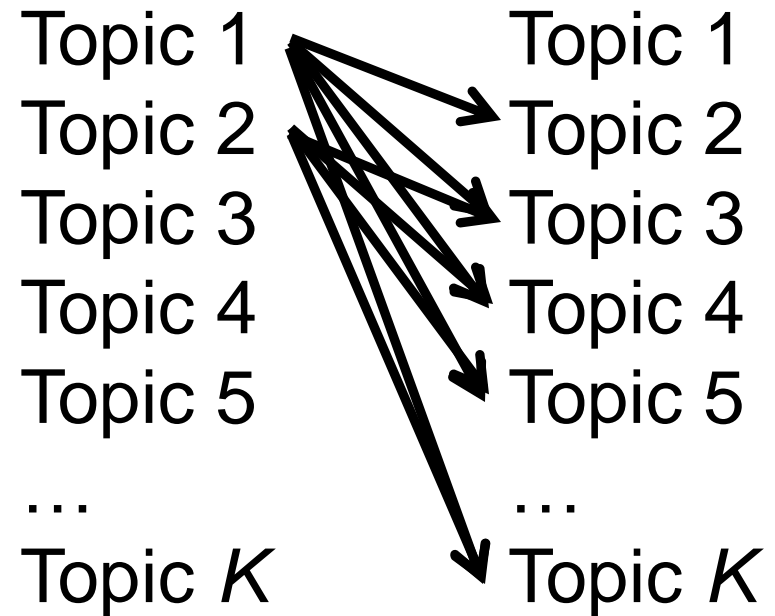
**Low KL divergence**  
**Low distinctness**

# Average Topic Distinctness

## Hall Topics



## Diff Topics



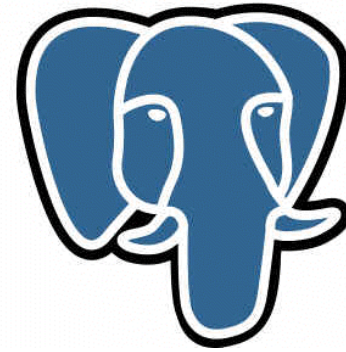
# Diff makes more distinct topics



**JHotDraw**

+32%

PostgreSQL



+38%

# Problems due to duplication

Found in Source Code Histories



Topics are muddled, not distinct



Diff makes more distinct topics



Evolutions are not accurate

*I bet the Diff model  
discovers more **accurate**  
topic evolutions*



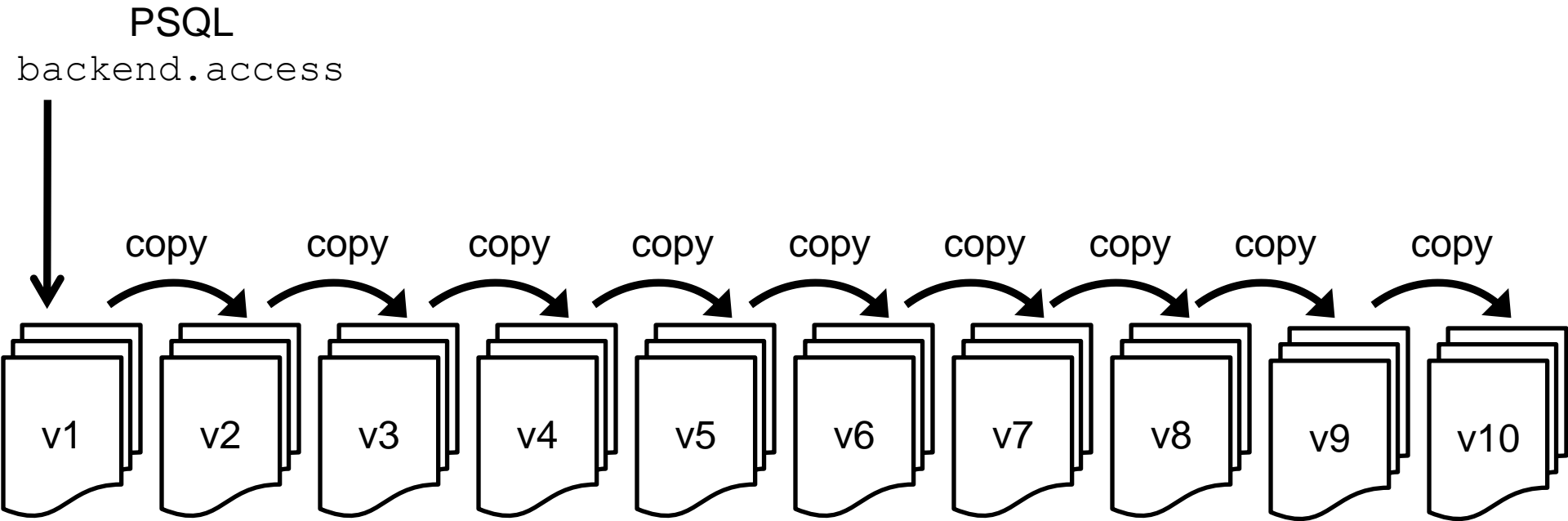
# Measuring Accuracy



No oracle dataset

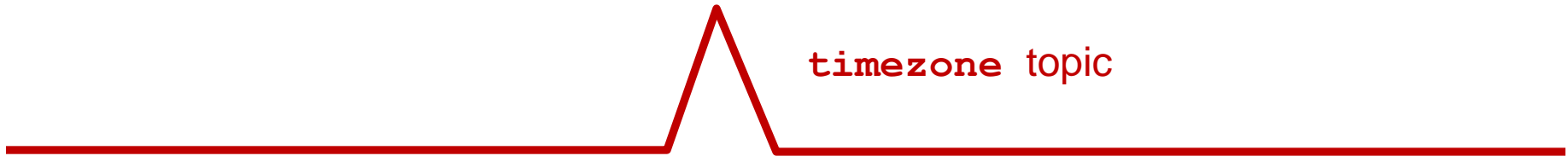
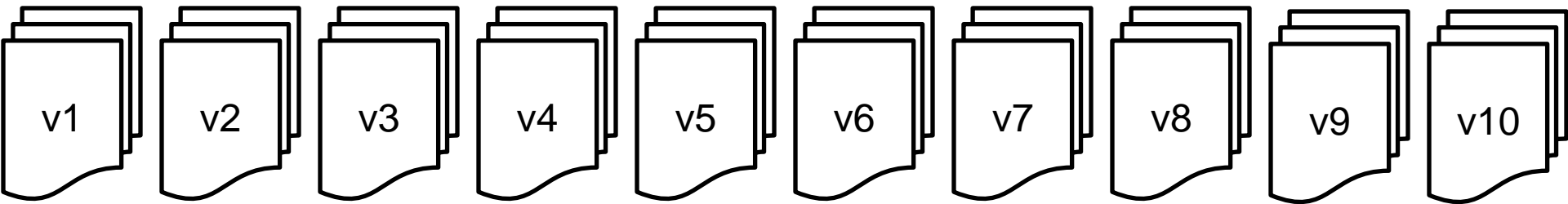
1. Create simulated scenario by hand  
*Truth known*
2. Manually investigate evolutions in JHotDraw and PSQL  
*Truth learned*

# Simulated Project

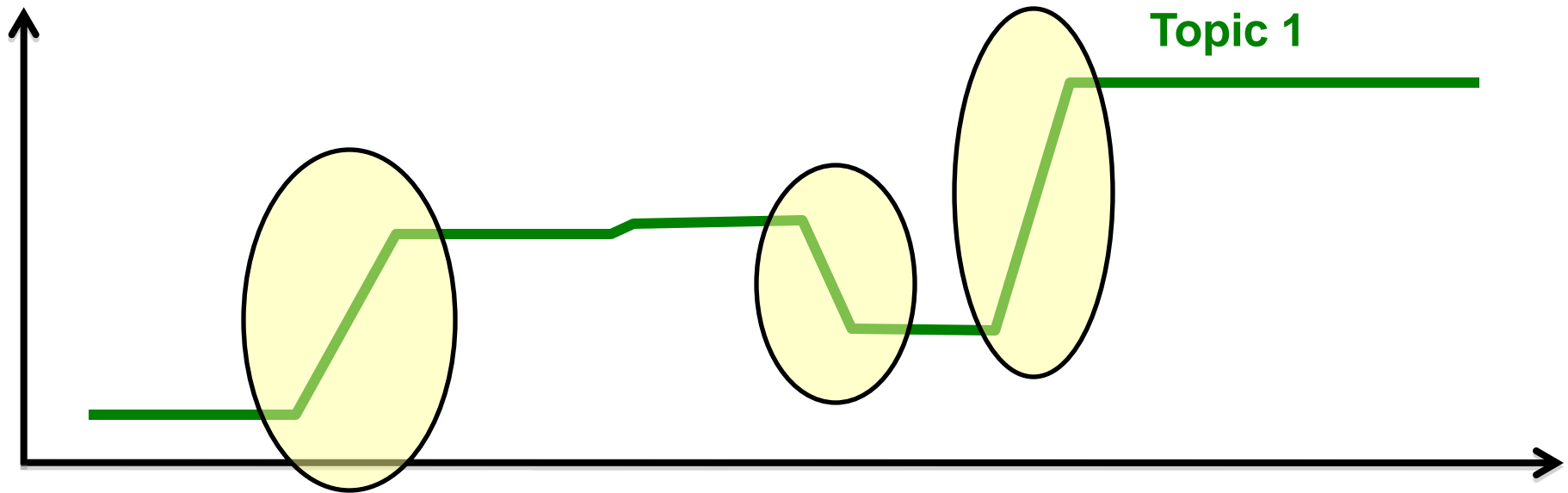


# Simulated Scenario 1

3 files from PSQL  
timezone



# Manual Investigation



1 .Select change events

2. Validate against project documentation  
(commit logs, release notes, etc.)

# Diff makes more accurate topics

*Simulated Project*

**+25%**  
precision

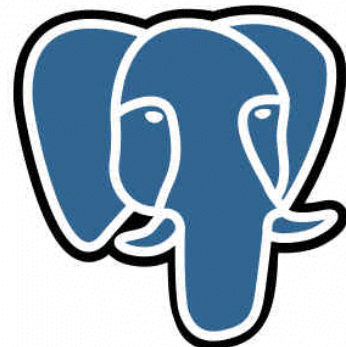
**+100%**  
recall



**JHotDraw**

**+33%**  
precision

PostgreSQL



**+47%**  
precision

# Problems due to duplication

Found in Source Code Histories



Topics are muddled, not distinct



Diff makes more distinct topics



Evolutions are not accurate

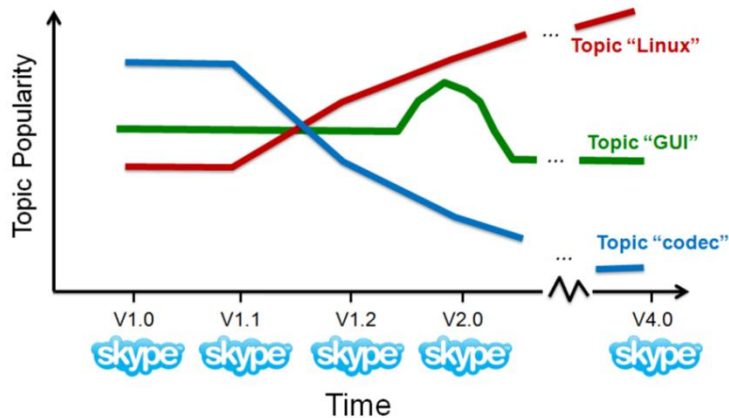


Diff makes more accurate evolutions

# Summary

## Tool: Topic Evolution Models

Applied to Source Code Histories



[5]

## Topic Evolution on Source Code



[7]



## The Diff Model

[13]

## Problems due to duplication

Found in Source Code Histories



Topics are muddled, not distinct



Diff makes more distinct topics



Evolutions are not accurate



Diff makes more accurate evolutions

[29]

[30]