

Enriching Reverse Engineering with Semantic Clustering

Adrian Kuhn, Tudor Gîrba, Stéphane Ducasse
Software Composition Group
University of Berne, Switzerland

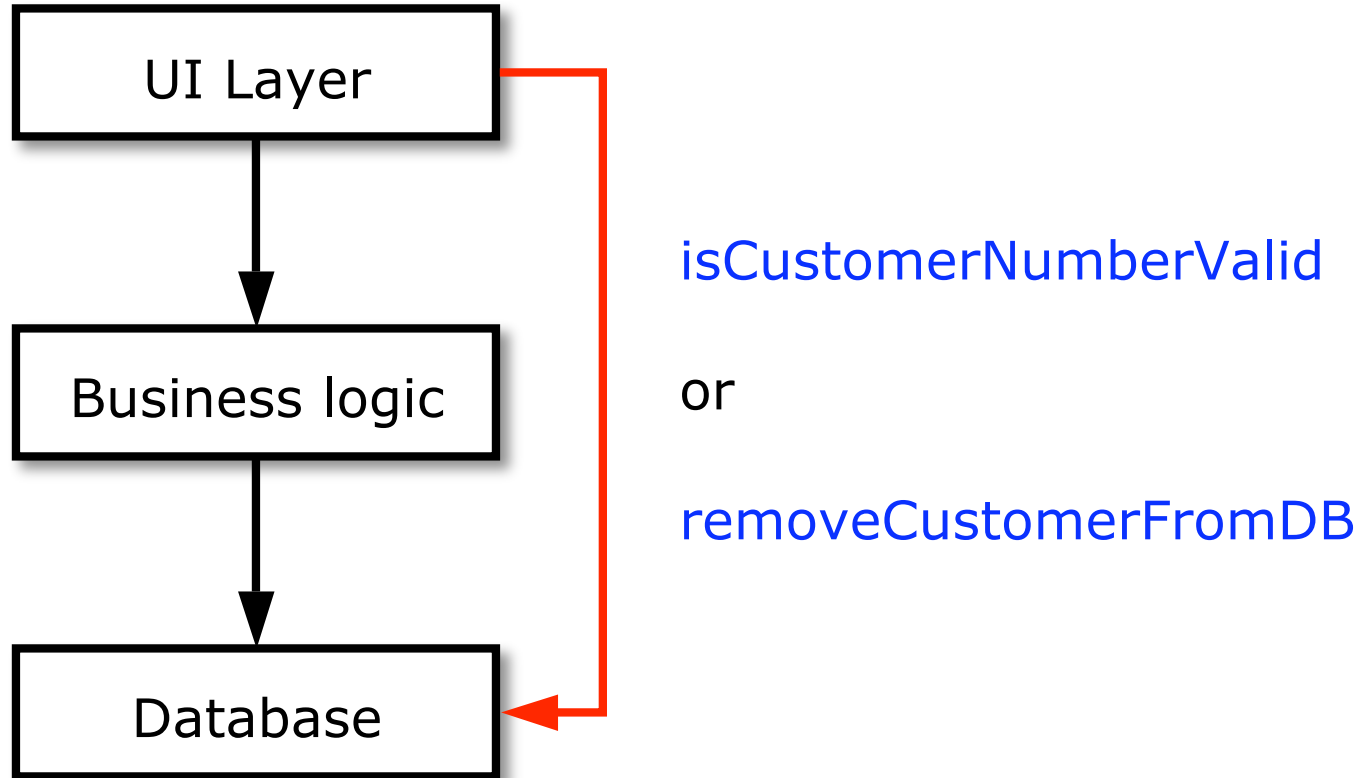
Context: reverse engineering a system must go beyond the structure

Most reverse-engineering approaches are dedicated to the structure.

But to understand a system, we have to know

- both its structure
- and its **domain semantics**

Problem: what is the meaning of that all?



Solution: look at the names!

Assumption:

Developers use **meaningful names**,
which capture the domain knowledge.

Thus:

We analyze the naming information...

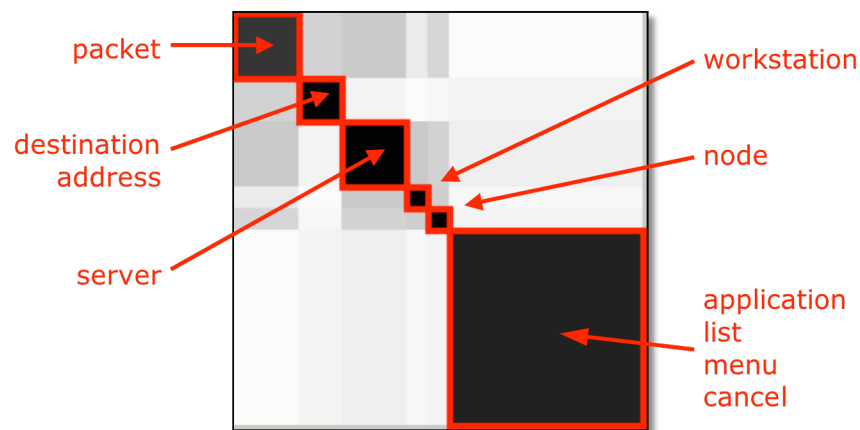
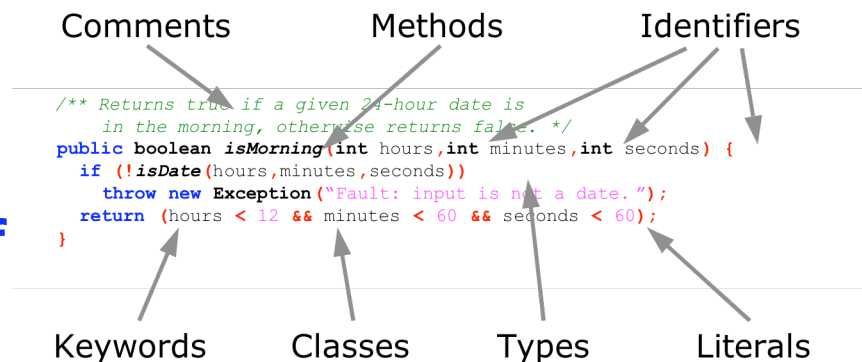
Semantic clustering in a nutshell

We treat source code as text documents.

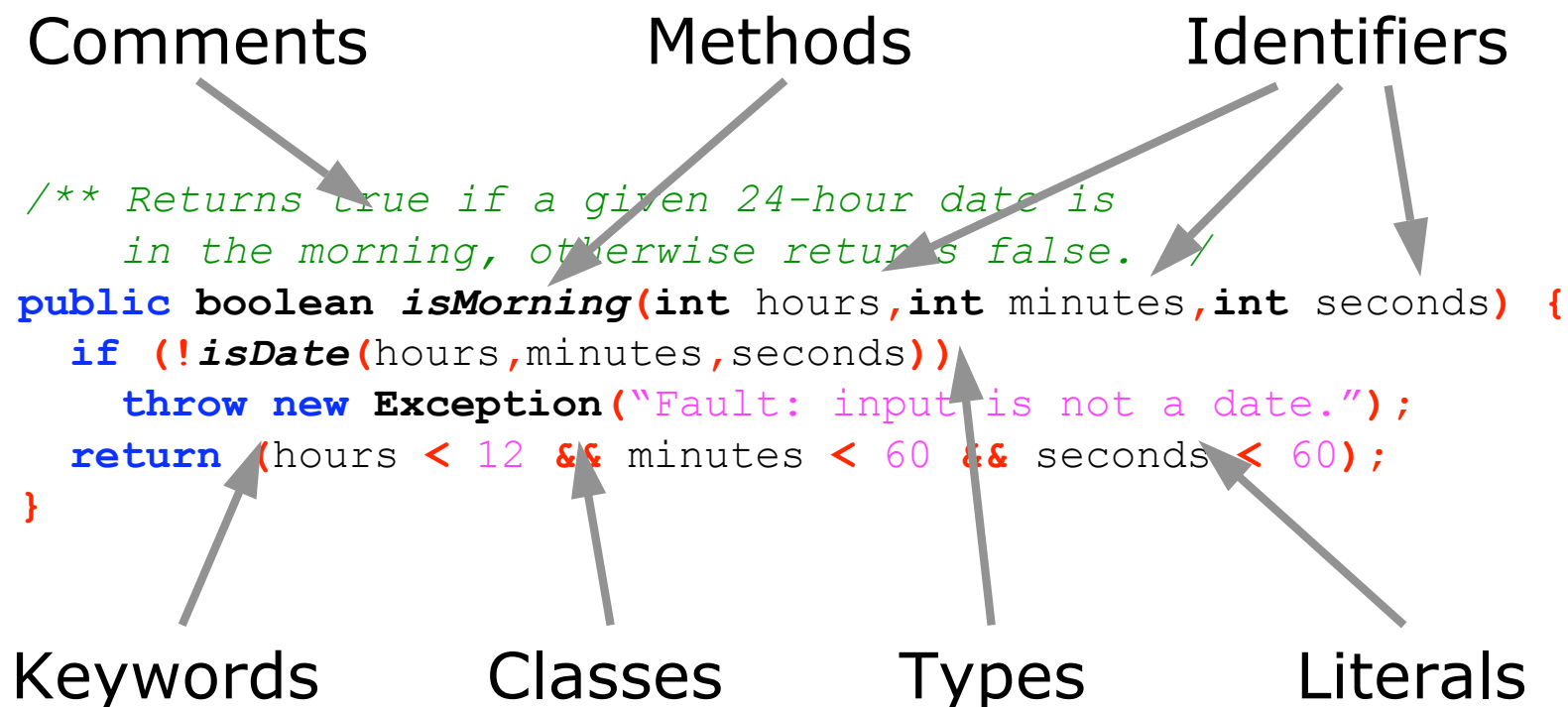
Two documents are similar if they use the same words.

We cluster the documents based on the usage of words.

We use automatically retrieved labels to describe the clusters.



Source code is **natural language text**, and each code item has a name



Thus we apply **Information Retrieval** on the source code

Hapax™

Hapax searches over 10,000 documents containing about 20,000 terms.

The term "hapax legomenon" is Greek and refers to a word that occurs only once in a given body of text. Hapax is a tool built on top of the Moose re-engineering framework.

Anything can be used as **document...**

Documents are not necessarily source files.

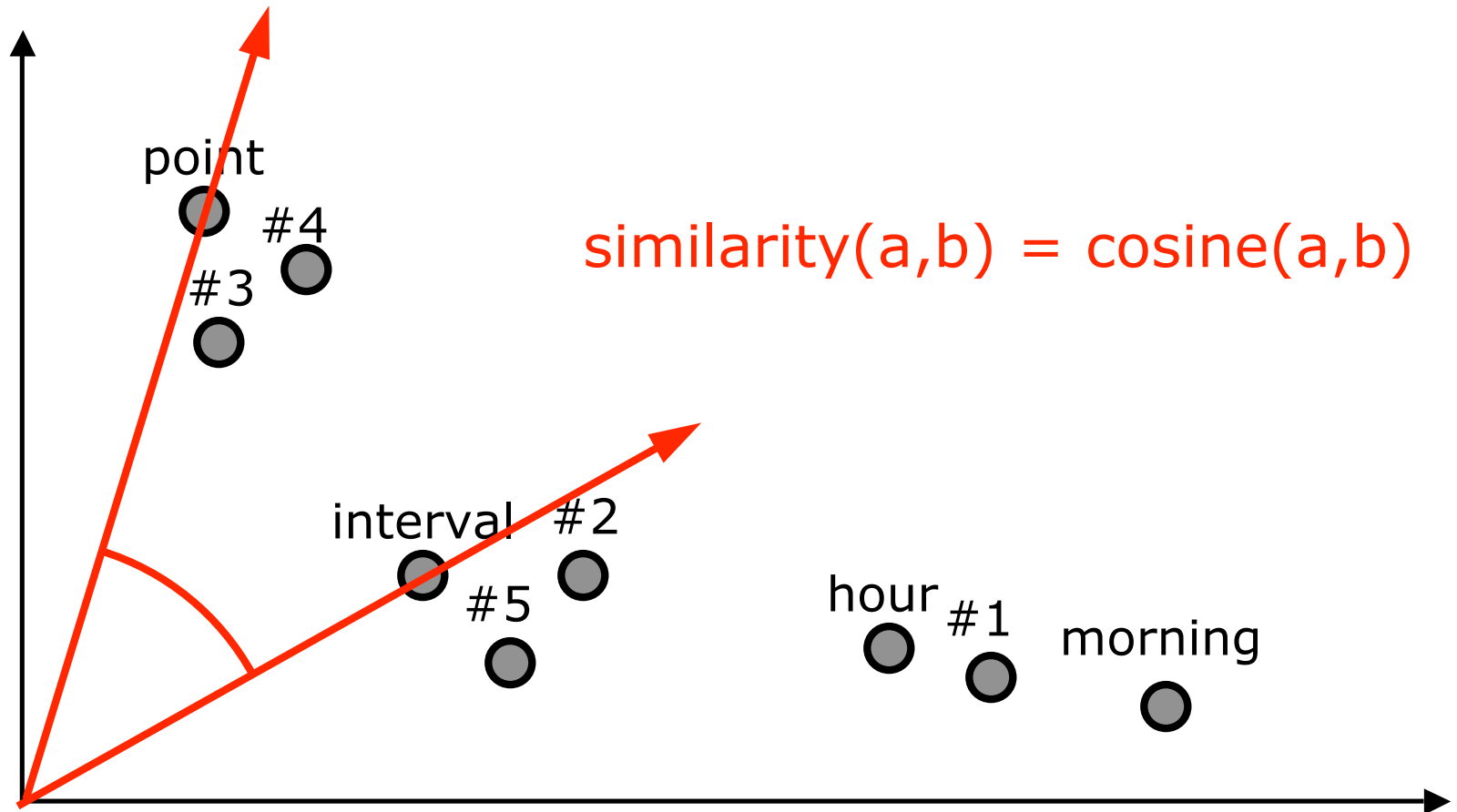
We prefer to split the source at structural levels such as: **package, class, method...**

But any software artifact that has a textual representation is a possible document

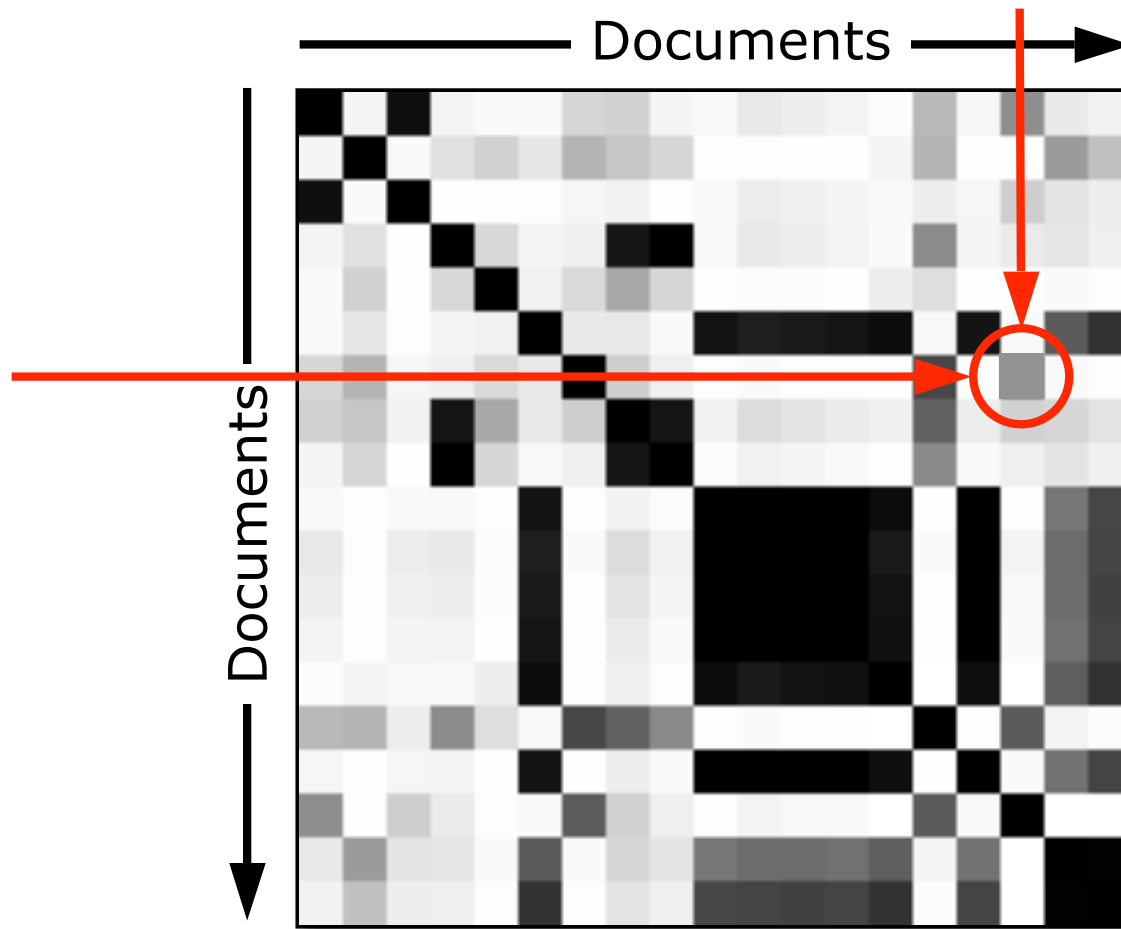
We count the **word frequencies** in each document

	doc 1	doc 2	doc 3	doc 4	doc 5	...
hour	2	3		3	2	
interval		2	1		1	
morning	1				1	
point			4	4		
...						

Latent semantic indexing puts all the documents and terms in one space



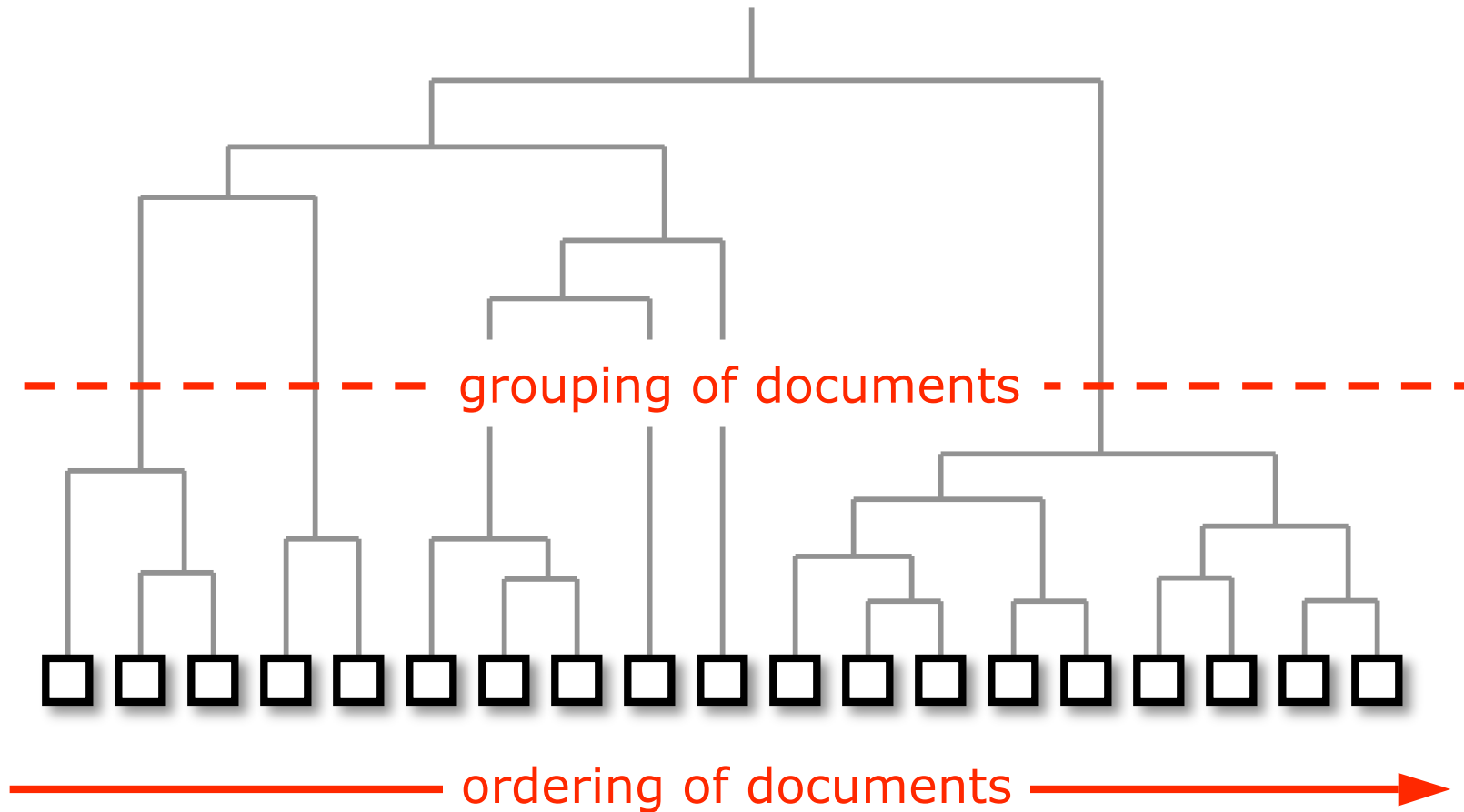
The **correlation matrix** shows all similarities between all documents



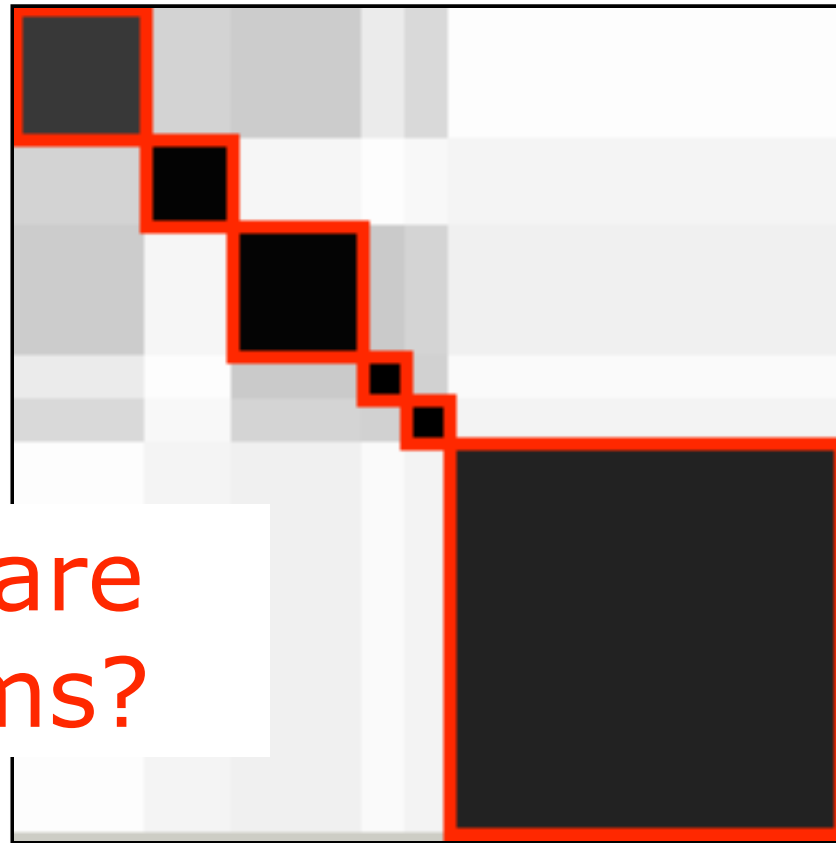
But an **unordered** matrix looks like television tuned to a dead channel...



Clustering yields **both** an ordering and a grouping of documents

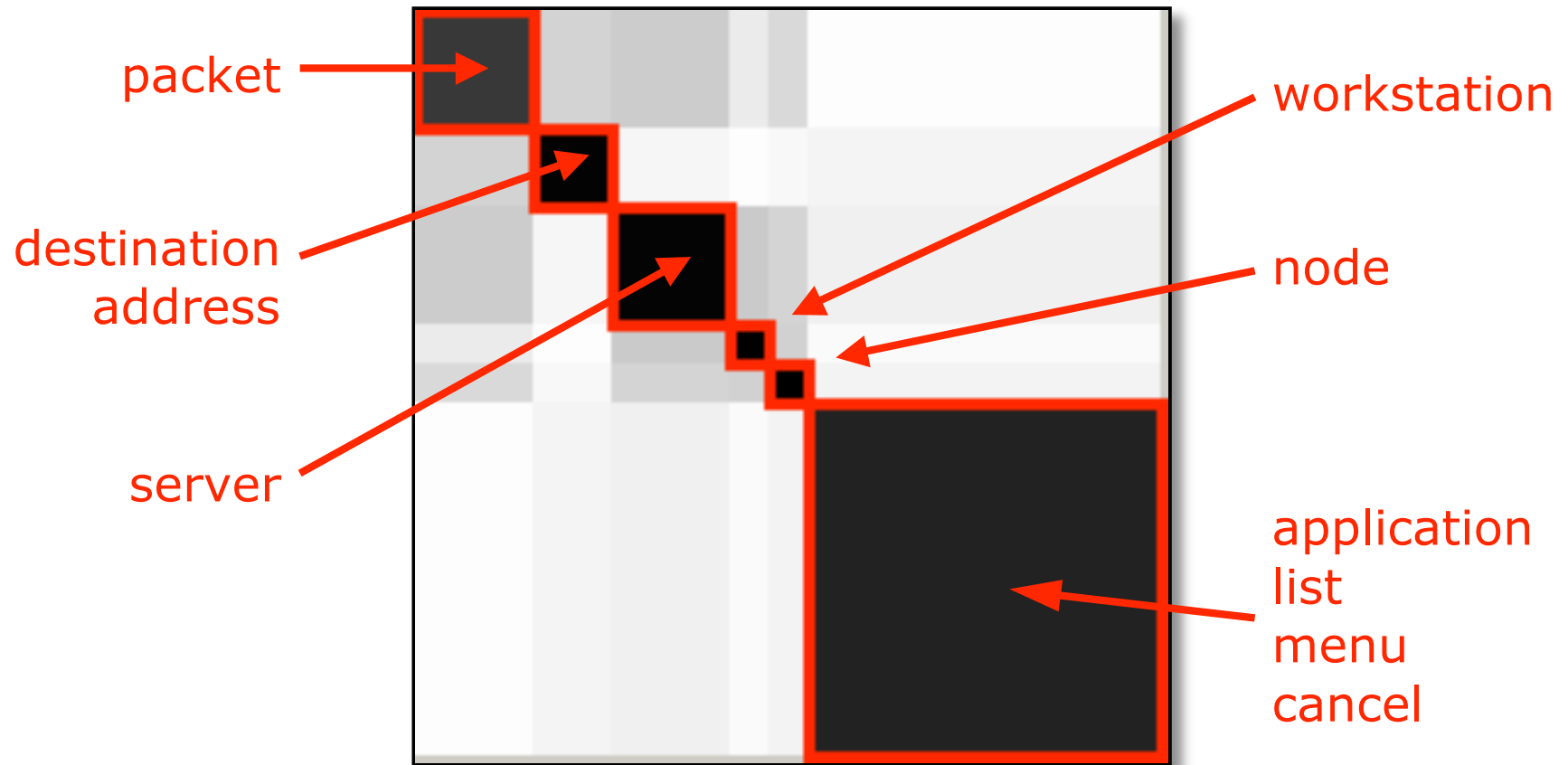


A **cluster** is set of documents which use similar terms

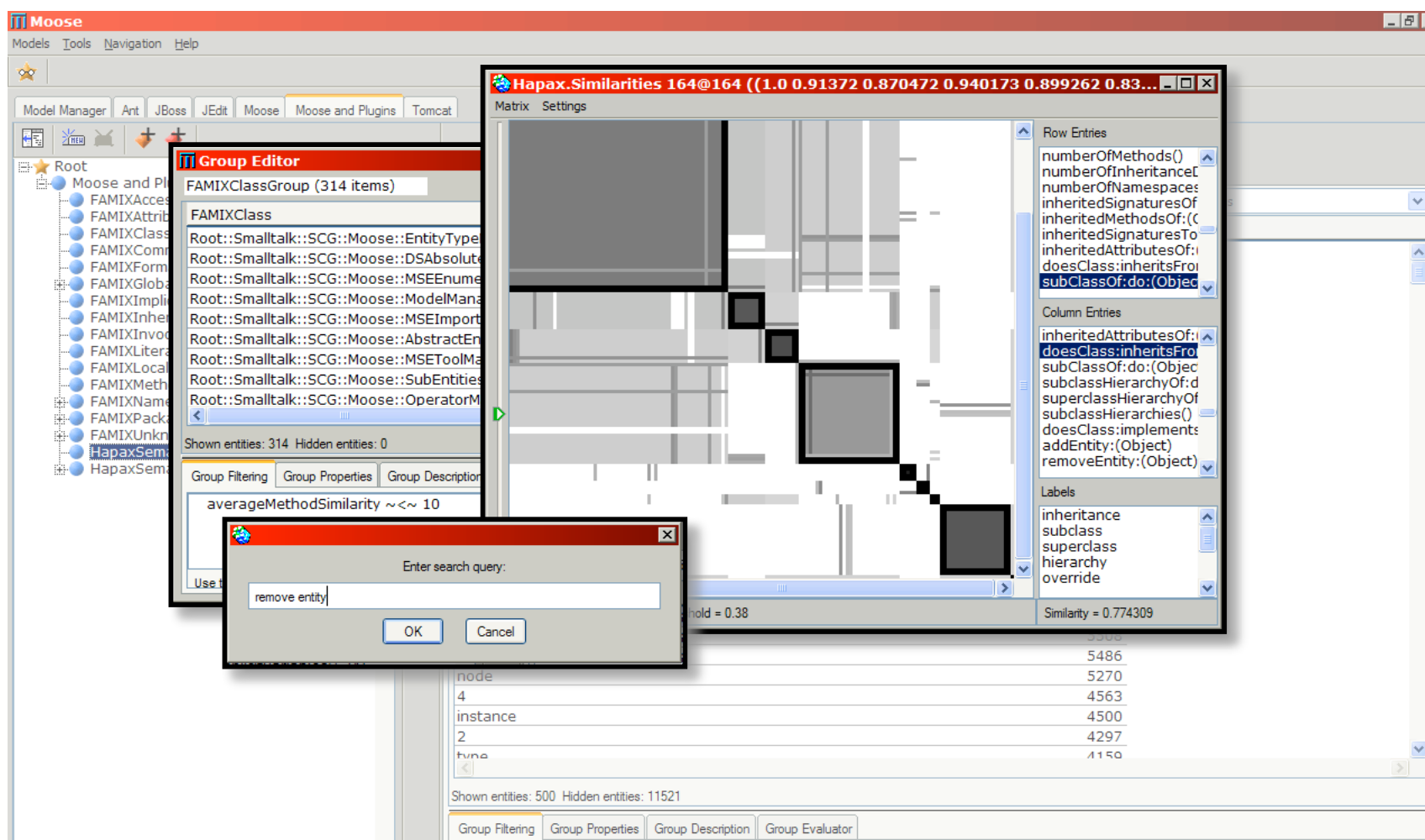


But what are these terms?

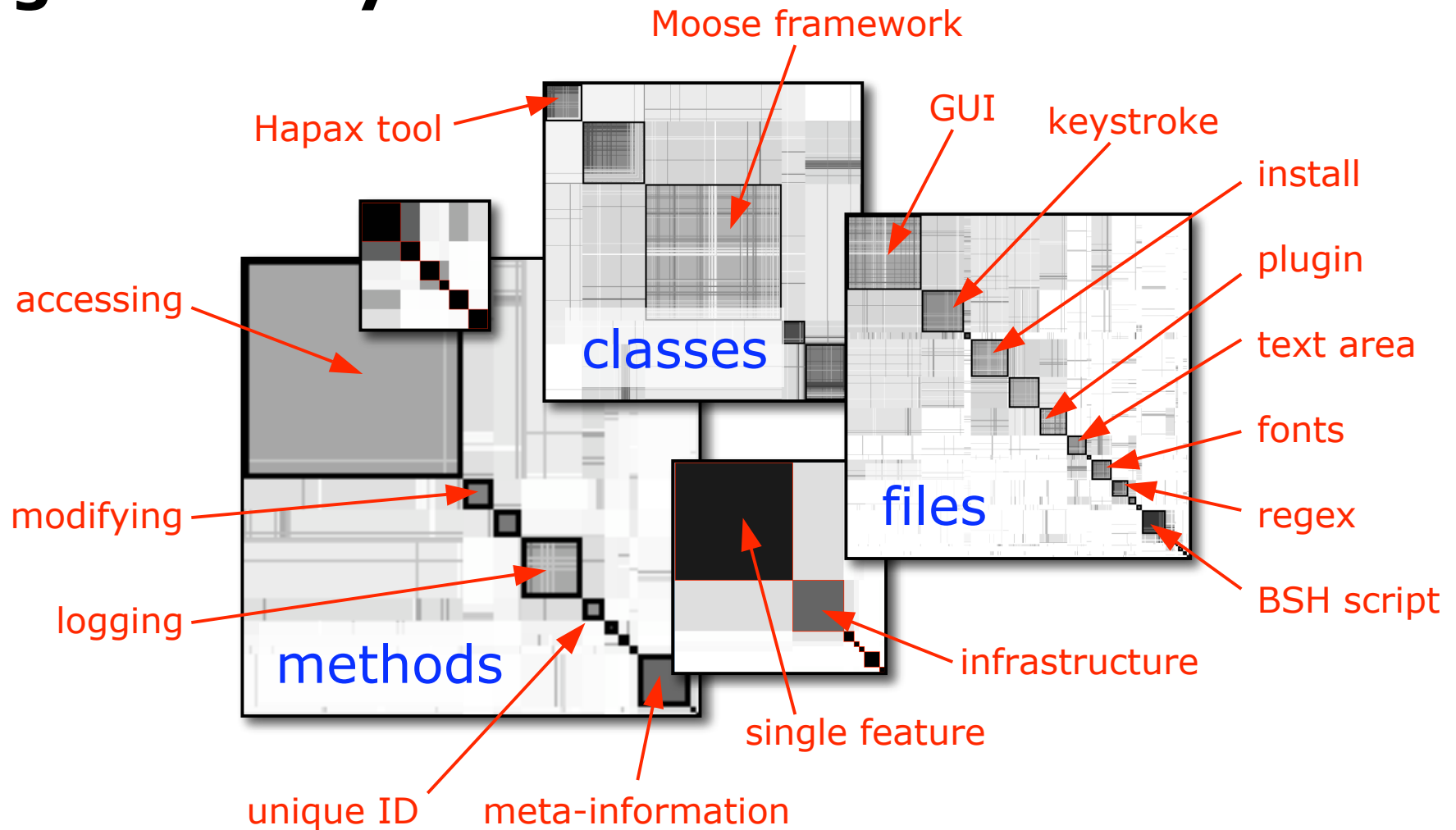
Automatically retrieved **labels** describe the clusters



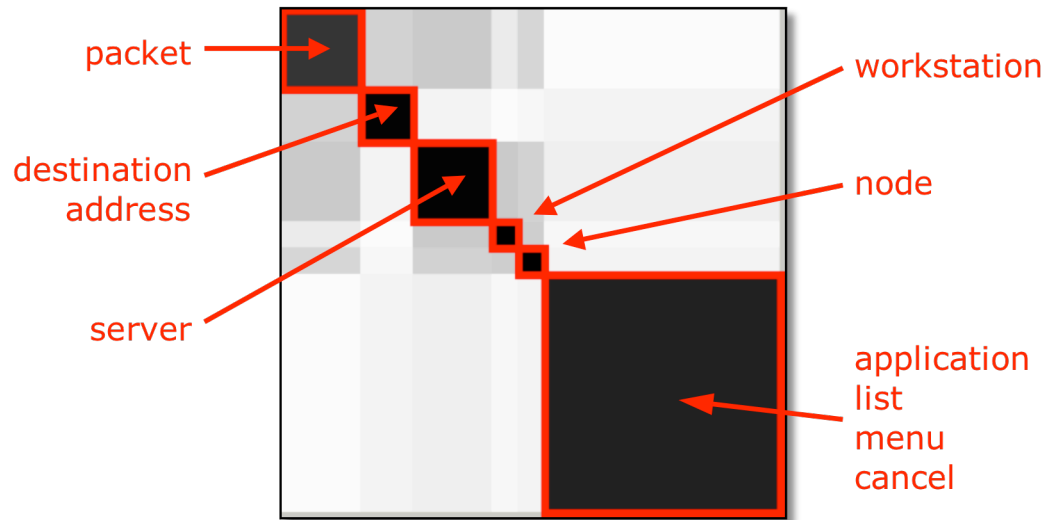
Moose and Hapax: explore structure and semantics with the same tool



Hapax was applied at different levels of granularity...



Conclusion: we should not just look for boxes, we should look for names too!



Questions?