The Relevance of Application Domains in Empirical Findings

Andrea Capiluppi
Brunel University London (UK)

Nemitari Ajienka
Edge Hill University (UK)

What this talk is about

1) Software ecosystems
2) Application domains

Contention/Conclusion

1) Different ways to define a software ecosystem
2) Ecosystems based on domains
3) OO metrics are sensitive to domain-based ecosystems
4) Software health related to domains

THE APACHE™ SOFTWARE FOUNDATION

ECLIPSE™ FOUNDATION

GNOME™ FOUNDATION

Software ecosystem

A set of actors functioning as a unit and interacting with a shared market for software and services, together with the relationships among them

Domains

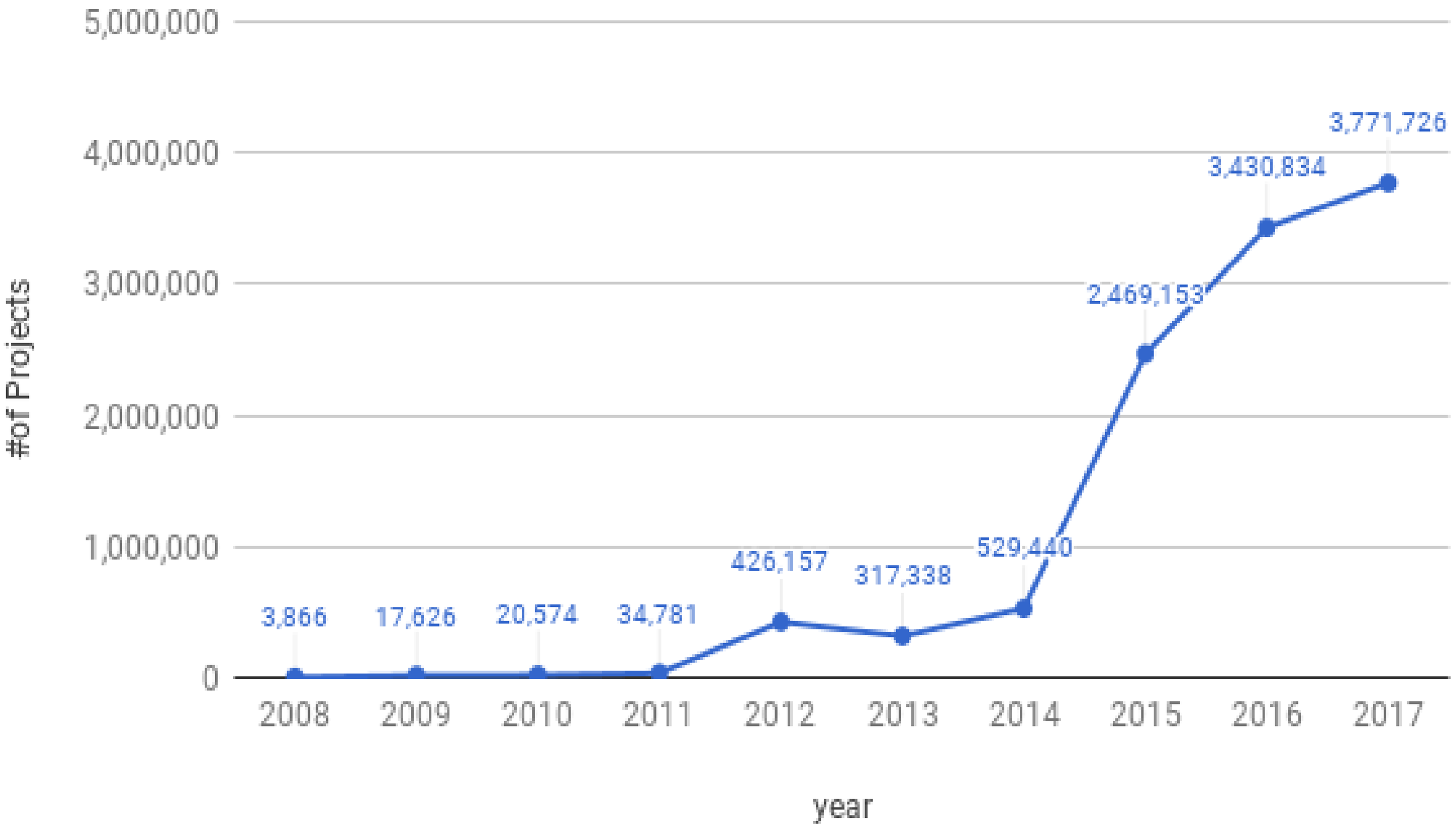
Thank you reusable cliparts

Domains

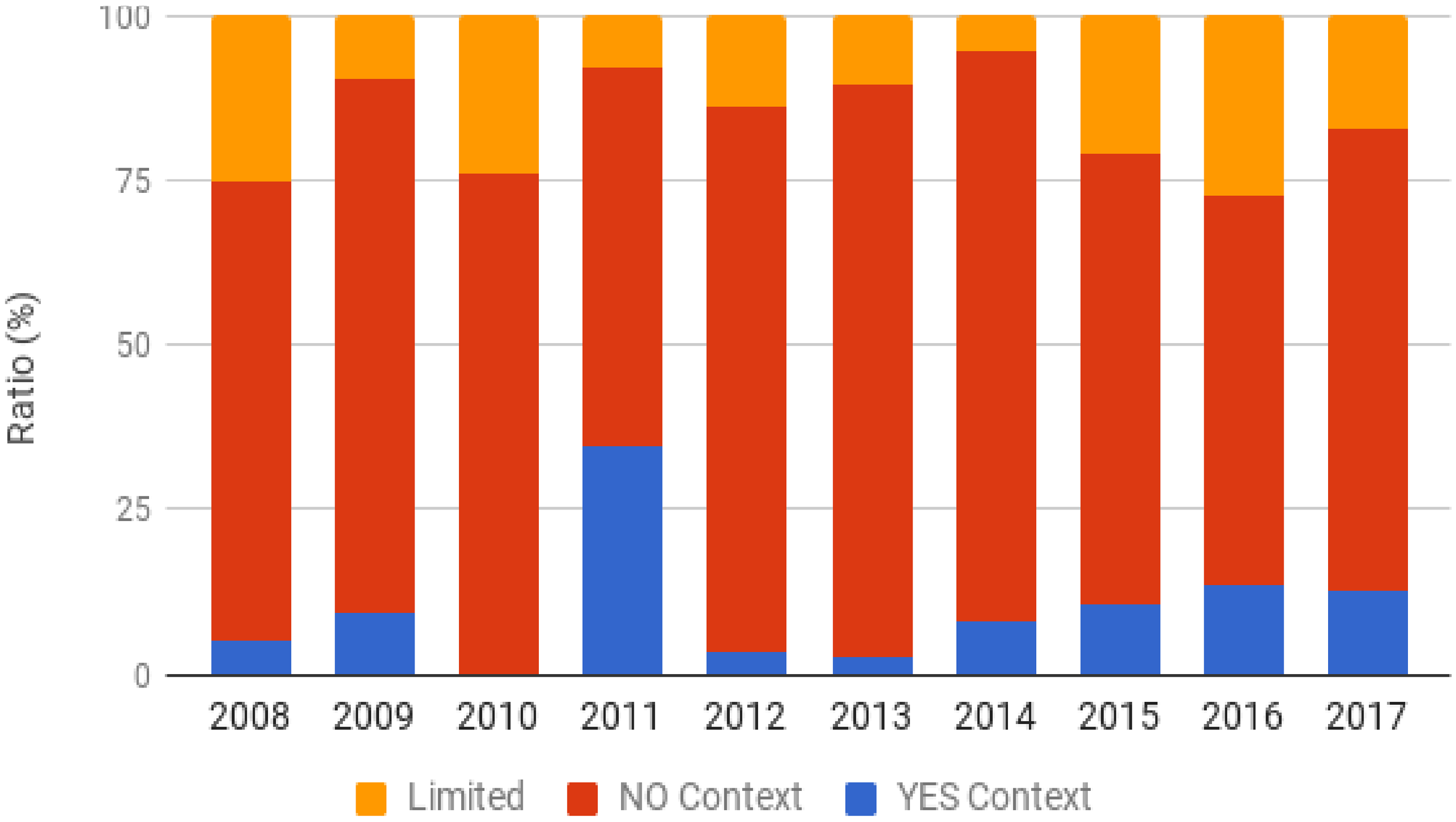
Thank you reusable cliparts

Background of the study

Cumulative number of OSS projects used by papers at MSR



Did they mention context? Or domain?

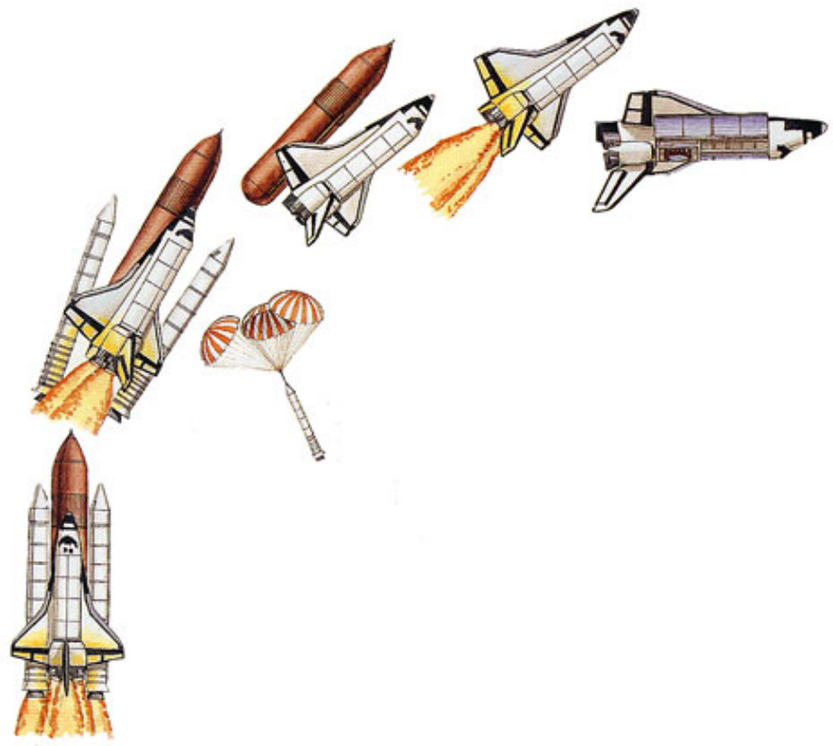


How

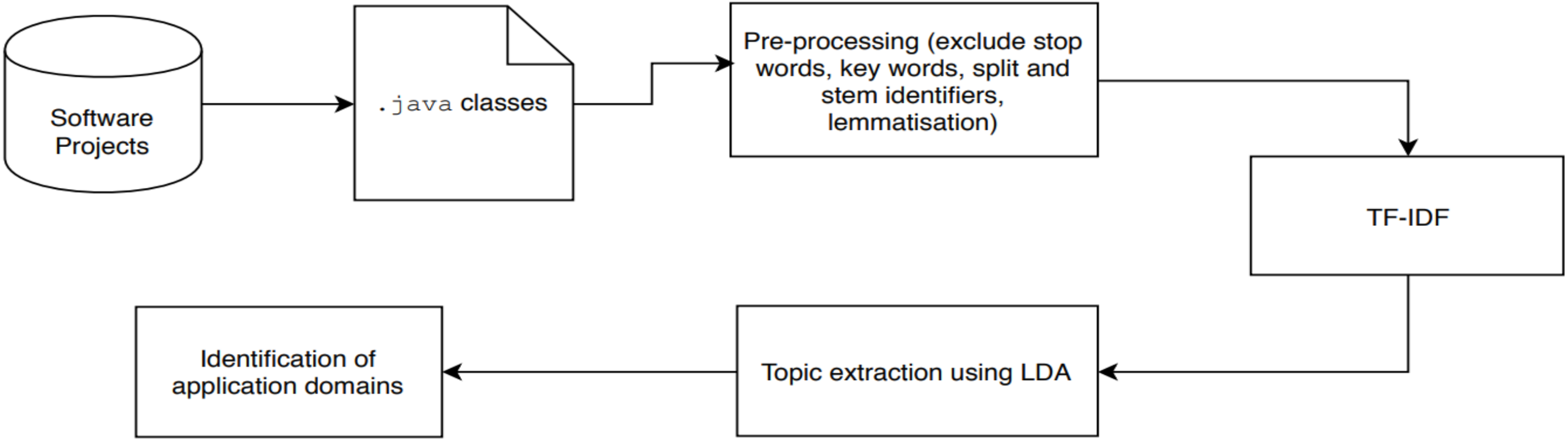
Step ONE
give *any* software system a unique domain

Step TWO
retrieve, cluster and analyse by domain

Step THREE
obtain results, check for differences



How


Software Projects
.java classes
Pre-processing (exclude stop words, key words, split and stem identifiers, lemmatisation)
TF-IDF
Topic extraction using LDA
Identification of application domains

From Code to Application Domain

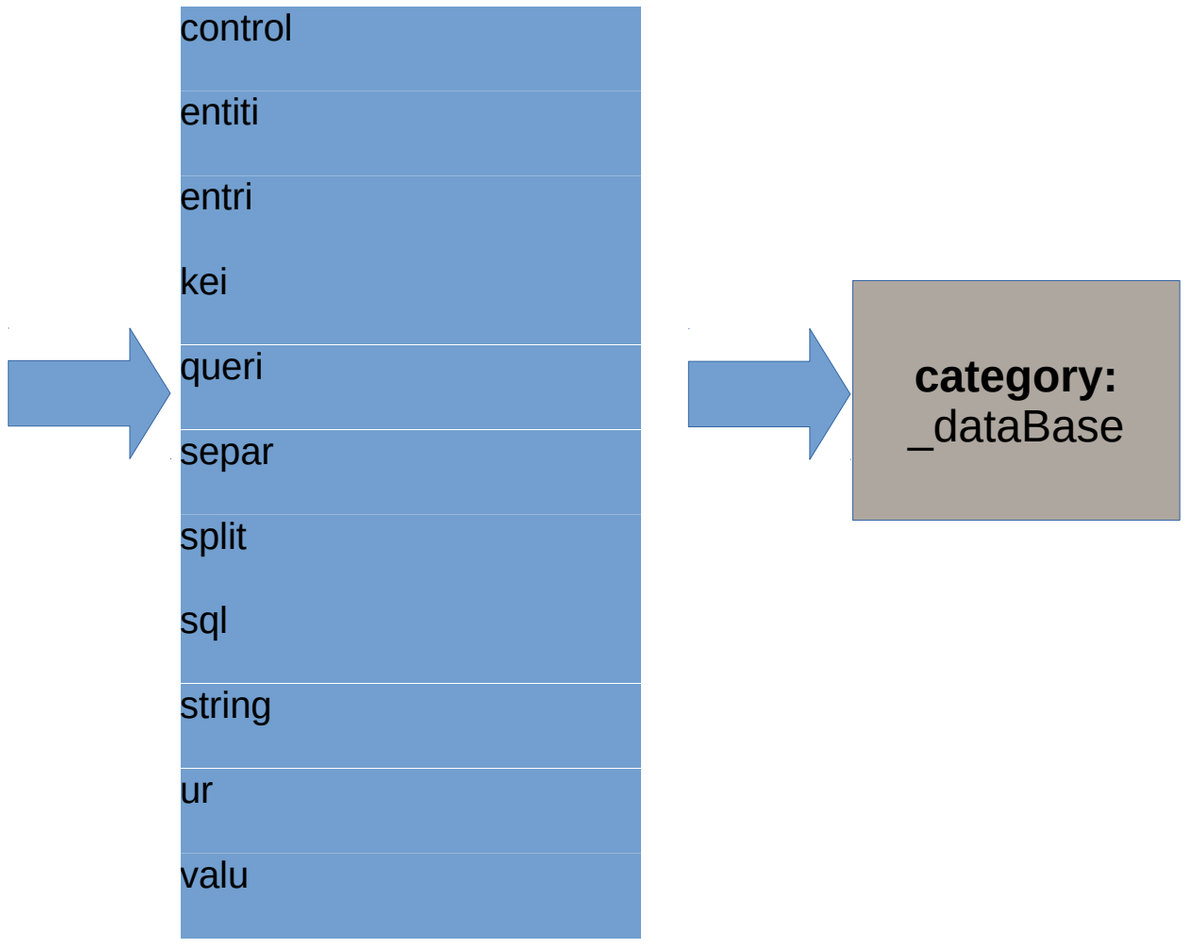
```
package tmacsoftware.ursql;

public class UrSQLEntry
{

    private String key;
    private String value;
    private String firstKey;
    private String firstValue;
    private UrSQLEntity entity;

    public UrSQLEntry()
    {
    }

    public UrSQLEntry(String query)
    {
        String[] split = query.split
(UrSQLController.KEY_VALUE_SEPARATOR);
        this.key = split[0];
        this.value = split[1];
        this.firstKey = this.key;
        this.firstValue = this.value;
    }
```



TF-IDF via Python

exampl print quot quot println string format win quot propos oper choic propos oper choic main string arg **TRAVEL** except **TRAVEL** handl ph **TRAVEL** handl ph add custom john mari ann bob ted luci ph add destin london op op op ph add destin berlin op op op op ph add destin madrid op op op ph add destin rome op op op ph add destin berlin **TRAVEL** except ex println ex messag println ph destin op berlin london madrid ph add propos pro london ph add propos pro berlin ph add propos pro madrid ph add propos pro rome ph add propos pro pragu **TRAVEL** except ex println ex messag list string list ph set **USER** pro john mari ann bob tom linda println list linda tom list ph set **USER** pro ann bob println list list ph set **USER** pro ann ted luci println list list ph set oper pro op op op op op println list op op list ph set oper pro op op op println list ph add quot pro op ph add quot pro op ph add quot pro op ph add quot pro op **TRAVEL** except ex println ex messag list quot quot ph quot pro println quot oper op ph make choic pro john op ph make choic pro ann op ph make choic pro bob op ph make choic pro mari op ph make choic pro ted op **TRAVEL** except ex println ex messag quot ph win quot pro print quot win quot propos pro oper op choic ph add quot pro op ph add quot pro op ph make choic pro ann op ph make choic pro luci op ph make choic pro ted op ph win quot pro print quot win quot propos pro oper op choic sort map string integ quot dest ph total quot destin println quot dest quot dest quot dest london madrid sort map integ list string opr quot ph oper number quot println opr quot opr quot opr quot op op op op sort map string **USER** dest ph number **USER** destin println **USER** dest **USER** dest **USER** dest berlin london madrid

Latent Dirichlet Allocation (LDA) via Python

Topic 0: 0.003*"**stream**" + 0.003*"**bodi**" + 0.003*"**header**" + 0.003*"**content**" + 0.003*"**id**" + 0.002*"**benchmark**" + 0.002*"**type**" + 0.002*"**ssl**" + 0.002*"**socket**" + 0.002*"**stori**"

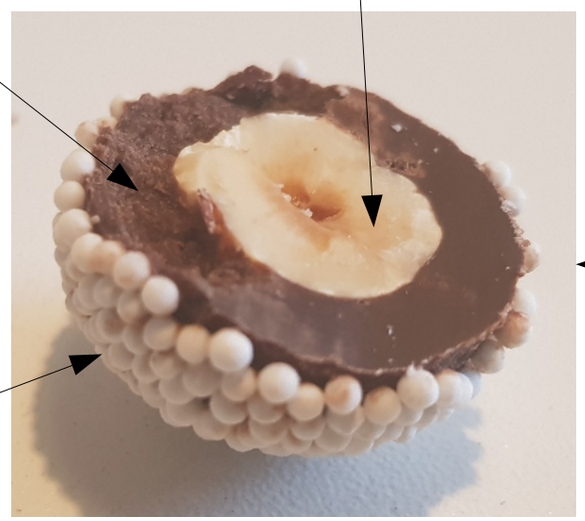
Topic 1: 0.002*"**entiti**" + 0.002*"**url**" + 0.002*"**proxi**" + 0.002*"**slack**" + 0.002*"**event**" + 0.001*"**frame**" + 0.001*"**filter**" + 0.001*"**client**" + 0.001*"**equal**" + 0.001*"**session**"

Topic 2: 0.005*"**cooki**" + 0.004*"**header**" + 0.004*"**interceptor**" + 0.003*"**chain**" + 0.003*"**url**" + 0.002*"**bodi**" + 0.002*"**certif**" + 0.002*"**content**" + 0.002*"**client**" + 0.002*"**timeout**"

Topic 3: 0.005*"**cach**" + 0.004*"**socket**" + 0.004*"**connect**" + 0.004*"**bodi**" + 0.003*"**rout**" + 0.003*"**server**" + 0.003*"**web**" + 0.003*"**header**" + 0.003*"**client**" + 0.003*"**url**"

Topic 4: 0.006*"**event**" + 0.006*"**socket**" + 0.005*"**certif**" + 0.005*"**address**" + 0.005*"**cach**" + 0.004*"**file**" + 0.003*"**deleg**" + 0.003*"**connect**" + 0.003*"**server**" + 0.003*"**inet**"

How



Which categories/topics/domains?

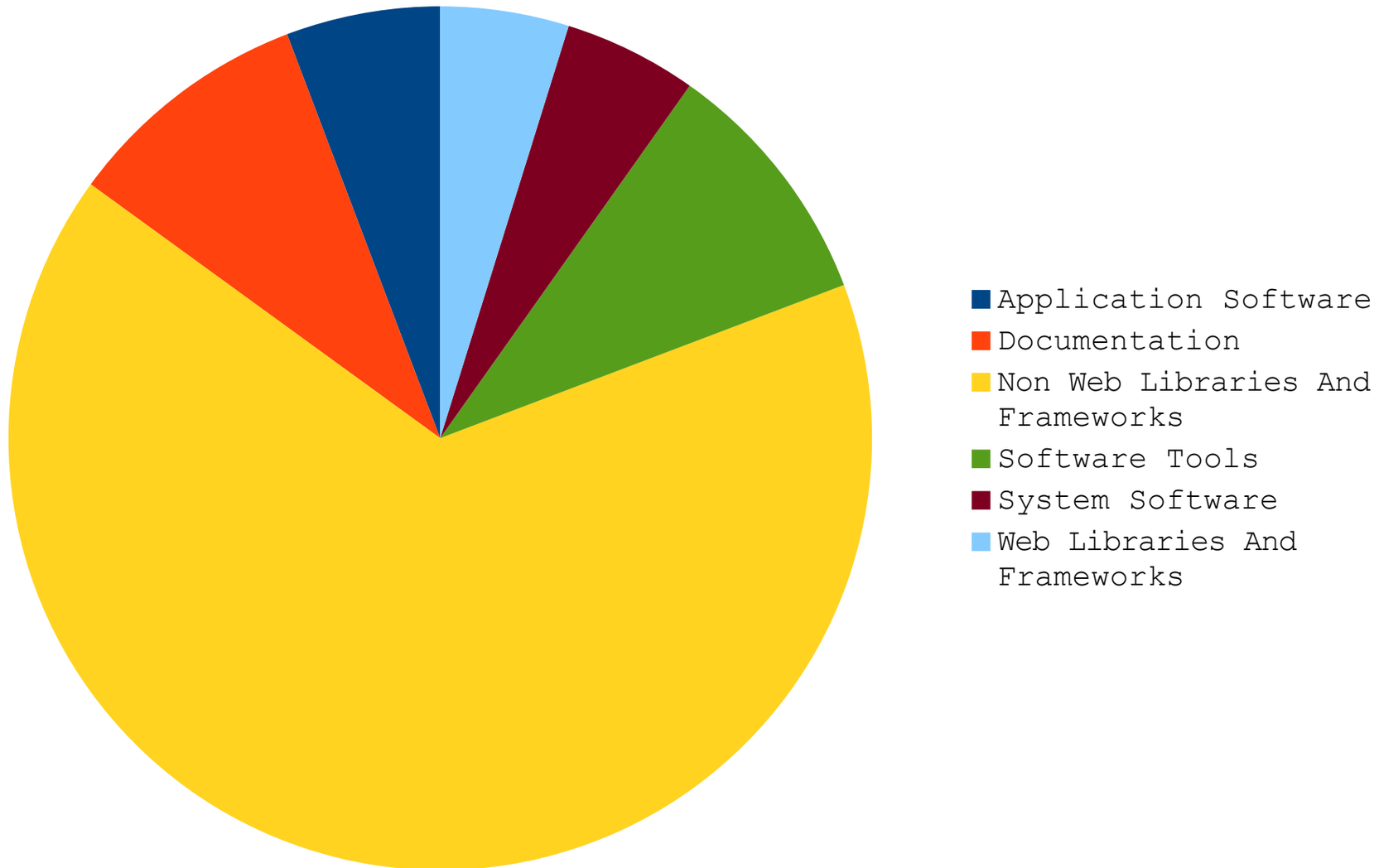
Preassigned?

- Application Software
- Documentation
- Non Web Libraries And Frameworks
- Software Tools
- System Software
- Web Libraries And Frameworks

* H. Borges, A. Hora, M. T. Valente, Understanding the factors that impact the popularity of GitHub repositories, in: 2016 IEEE International Conference on Software Maintenance and Evolution (ICSME), IEEE, 2016, pp. 334–344.

* H. Borges, M. T. Valente, Whats in a GitHub star? Understanding repository starring practices in a social coding platform, Journal of Systems and Software 146 (2018) 112–129.

Preassigned?



SourceForge.net domains

1:Communications

2:Database

3:Desktop Environment

4:Education

5:Formats and Protocols

6:Games/Entertainment

7:Internet

8:Mobile

9:Multimedia

10:Office/Business

11:Other/Nonlisted Topic

12:Printing

13:Religion and Philosophy

14:Scientific/Engineering

15:Security

16:Social sciences

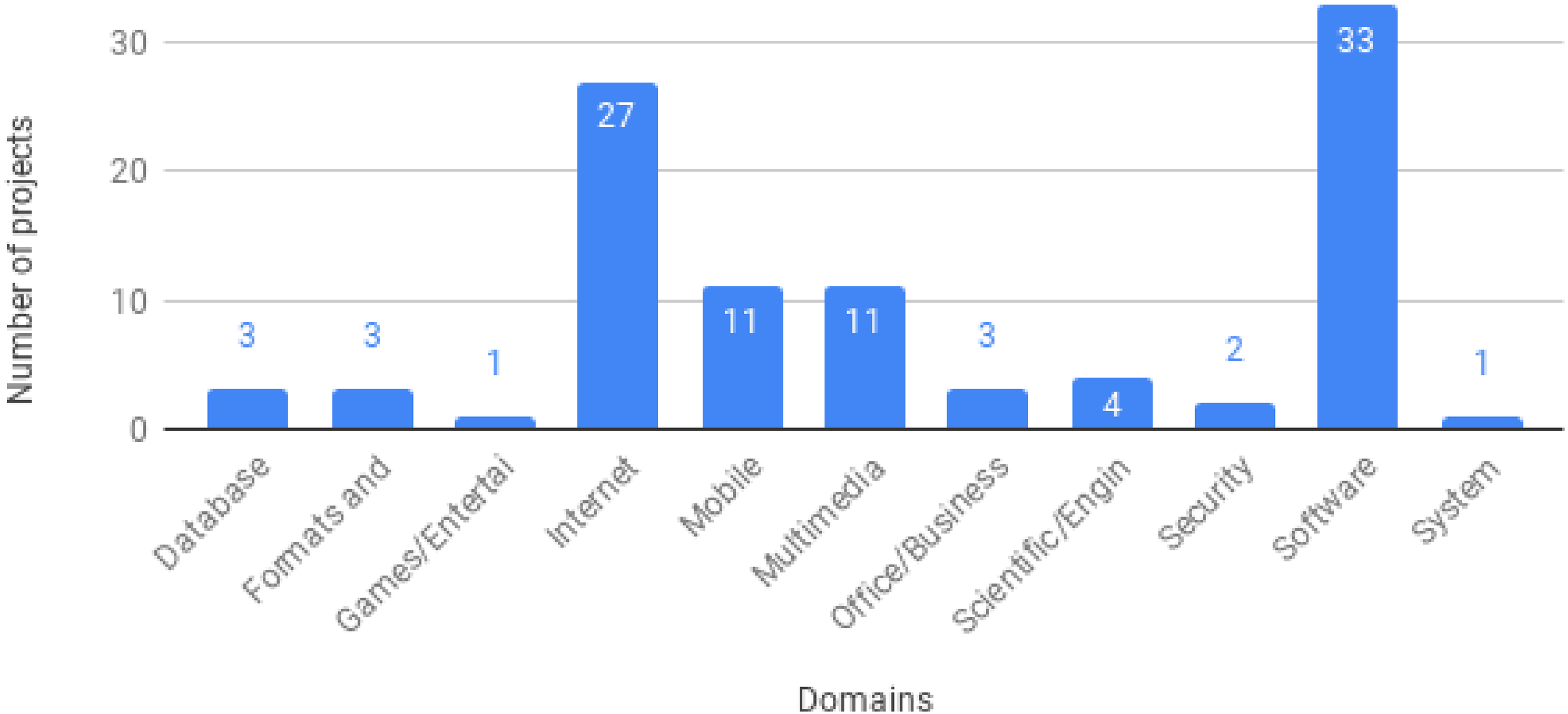
17:Software Development

18:System

19:Terminals

20:Text Editors

Triangulation (100 GitHub projects)



Structural attributes chosen

- NOC
- DIT
- CBO
- RFC
- WMC
- LCOM
- NIM
- IFANIN
- NIV

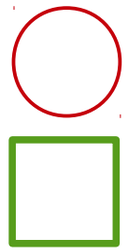
Domains make a difference

	Mobile	Multimedia	Software Devel
Internet	✓	✓	✓
Mobile	x	*NOC*	*NOC*
Multimedia		x	*IFANIN*
Software Devel			x

Only for the NOC attribute we cannot reject the hypothesis "***the NOC values of Mobile and Multimedia come from the same distribution***"

Past evidence?

- CBO & RFC
- RFC & LCOM



Giancarlo Succi, Witold Pedrycz, Snezana Djokic, Paolo Zuliani, and Barbara Russo. 2005. An empirical exploration of the distributions of the Chidamber and Kemerer object-oriented metrics suite. Empirical Software Engineering 10, 1 (2005), 81–104.

Past evidence?

Internet

	IFANIN	CBO	NOC	NIM	NIV	WMC	RFC	DIT
CBO	-i							
NOC	-i	i						
NIM	s	M	i					
NIV	i	M	i	M				
WMC	i	M	i	AP	M			
RFC	-i	M	i	M	s	M		
DIT	-s	i	-i	i	-i	i	L	
LCOM	-i	s	i	M	M	M	M	s

Mobile

	IFANIN	CBO	NOC	NIM	NIV	WMC	RFC	DIT
CBO	i							
NOC	-i	i						
NIM	s	M	s					
NIV	i	M	i	XL				
WMC	i	M	s	AP	XL			
RFC	-i	s	s	L	M	L		
DIT	-s	s	s	s	s	s	I	
LCOM	-i	M	s	L	L	L	M	s

Multimedia

	IFANIN	CBO	NOC	NIM	NIV	WMC	RFC	DIT
CBO	i							
NOC	i	i						
NIM	i	s	i					
NIV	i	M	i	M				
WMC	-i	L	i	s	i			
RFC	-i	L	i	s	i	AP		
DIT	-M	s	i	s	i	i	s	
LCOM	i	s	i	M	L	i	i	s

Software Development

	IFANIN	CBO	NOC	NIM	NIV	WMC	RFC	DIT
CBO	i							
NOC	-i	i						
NIM	i	M	i					
NIV	i	i	i	i				
WMC	i	M	i	AP	i			
RFC	-i	s	i	s	i	s		
DIT	-s	i	-i	i	i	i	M	
LCOM	-i	M	i	M	i	M	s	s

Software Health?

Internet

	IFANIN	CBO	NOC	NIM	NIV	WMC	RFC	DIT
CBO	-i							
NOC	-i	i						
NIM	s	M	i					
NIV	i	M	i	M				
WMC	i	M	i	AP	M			
RFC	-i	M	i	M	s	M		
DIT	-s	i	-i	i	-i	i	L	
LCOM	-i	s	i	M	M	M	M	s

Mobile

	IFANIN	CBO	NOC	NIM	NIV	WMC	RFC	DIT
CBO	i							
NOC	-i	i						
NIM	s	M	s					
NIV	i	M	i	XL				
WMC	i	M	s	AP	XL			
RFC	-i	s	s	L	M	L		
DIT	-s	s	s	s	s	s	L	
LCOM	-i	M	s	L	L	L	M	s

Multimedia

	IFANIN	CBO	NOC	NIM	NIV	WMC	RFC	DIT
CBO	i							
NOC	i	i						
NIM	i	s	i					
NIV	i	M	i	M				
WMC	-i	L	i	s	i			
RFC	-i	L	i	s	i	AP		
DIT	-M	s	i	s	i	i	s	
LCOM	i	s	i	M	L	i	i	s

Software Development

	IFANIN	CBO	NOC	NIM	NIV	WMC	RFC	DIT
CBO	i							
NOC	-i	i						
NIM	i	M	i					
NIV	i	i	i	i				
WMC	i	M	i	AP	i			
RFC	-i	s	i	s	i	s		
DIT	-s	i	-i	i	i	i	M	
LCOM	-i	M	i	M	i	M	s	s

RFC v WMC

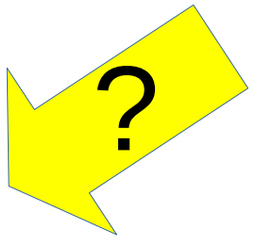
Future work

An oracle of domains

1:Communications
2:Database
3:Desktop Environment
4:Education
5:Formats and Protocols
6:Games/Entertainment
7:Internet
8:Mobile
9:Multimedia
10:Office/Business
11:Other/Nonlisted Topic
12:Printing
13:Religion and Philosophy
14:Scientific/Engineering
15:Security
16:Social sciences
17:Software Development
18:System
19:Terminals
20:Text Editors

Main domain	Secondary domain	Projects	Parsed (via SVN)
Office/Business		3,896	341
Communications		4,103	408
Desktop Environment		535	49
SW Development		14,702	569
SW Development	Database	1,975	254
SW Development	Text Editors	1,119	140
SW Development	Formats & Protocols	1,642	357
Games/Entertainment		4,360	538
Home/Education	Printing	176	27
Home/Education	Religion & Philosophy	98	13
Home/Education	Social sciences	190	26
Internet		9,046	594
Mobile		260	50
Multimedia		3,294	411
Other/Nonlisted Topic		1,042	42
Scientific/Engineering		5,742	855
Security		1,111	128
Terminals		127	21

Oracle and assignment to domain clusters



Conclusion

1) Different ways to define a software ecosystem
2) Ecosystems based on domains
3) OO metrics are sensitive to domain-based ecosystems
4) Software health dependent on domain

Thanks for having me

Questions?