Measuring and mining evolution of software projects

Alexander Serebrenik

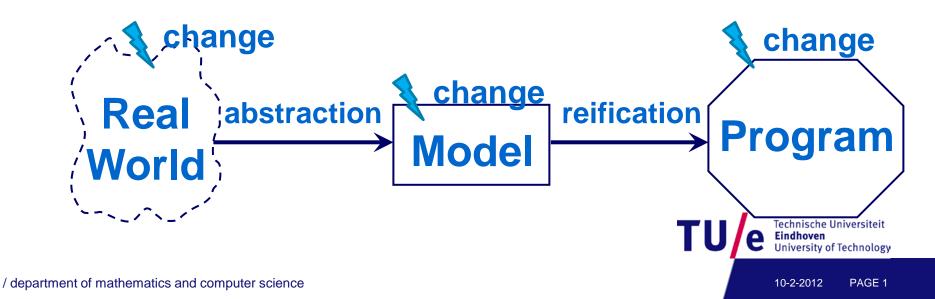
Technische Universiteit **Eindhoven** University of Technology

Where innovation starts

TU

Maintenance: 75-95% costs

- Software is
 - crucial for modern society
 - more complex than any other human artifact
 - subject to change
 - GNOME, 10 years, 1000 developers, 2.5 millions changes
 - Mozilla, 6 years, > 100 developers, > 1 million changes



Evolution: one change a top of another

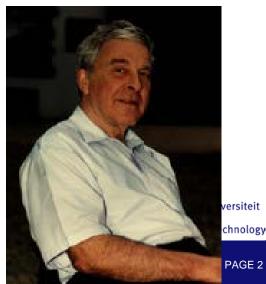
Evolution is staged process of progressive change over time in the properties, attributes, characteristics, behaviour of some material or abstract, natural or artificial, entity or



system

Charles Darwin

Meir Manny Lehman



/ department of mathematic

Why do we want to study software evolution?

- Software = the weakest link (often)
- Evolution "in general" makes things more complex
 - Science:
 - What is the nature of software evolution?
 - Psychology, sociology and organization theory, economics, law...
 - Engineering:
 - Where did the things go wrong?
 - incorrect, too complex, out of sync with other artefacts
 - Where can/will the things go wrong?
 - prediction, week spot identification, ...
 - What can we do to prevent the things from going wrong?

Where did the things go wrong?

• What was wrong?

- incorrect, too complex, out of sync with other artefacts
- Metrics:
 - "You can't control what you can't measure" (*DeMarco*)

• When did it happen?

- last month, before the release, when fixing a bug
- Repository mining



Measuring With Metrics



/ department of mathematics and computer science

10-2-2012 Page 5

Software metrics: Examples

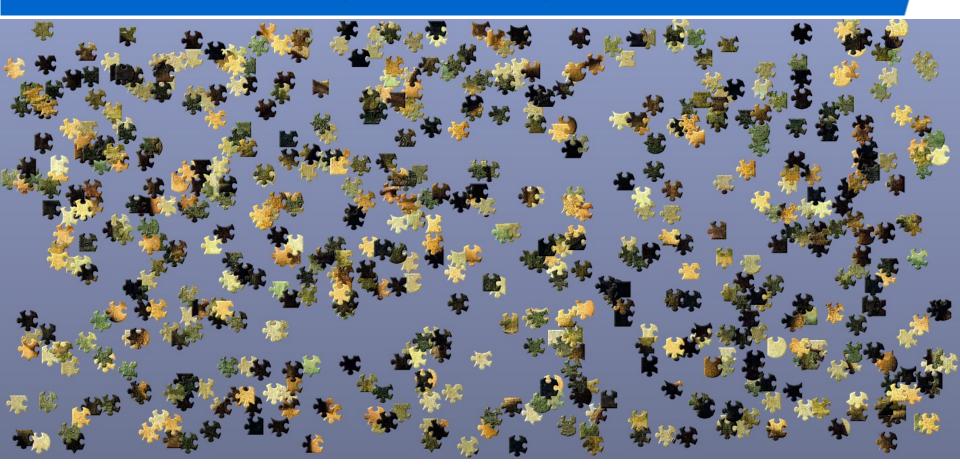
- Size
- Complexity
- Inheritance
- Comments
- Churn

. . .

- Coupling/cohesion
- Abstractness

TU/e Technische Universiteit Eindhoven University of Technology

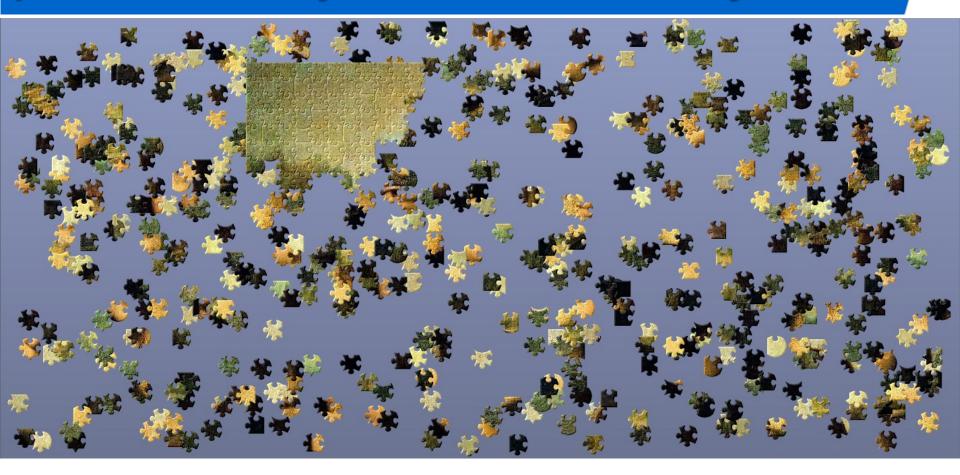
Metrics are usually computed at a *low* level: classes, methods, ...





10-2-2012 PAGE 7

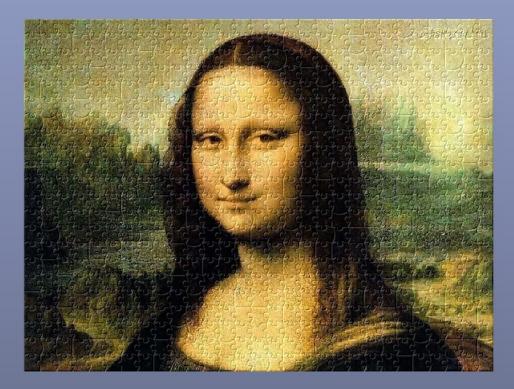
Multitude of data values obscures a general picture of the system maintainability





10-2-2012 PAGE 8

That we are actually interested in!





10-2-2012 PAGE 9

/W&I / MDSE

We need aggregation techniques to get the whole picture



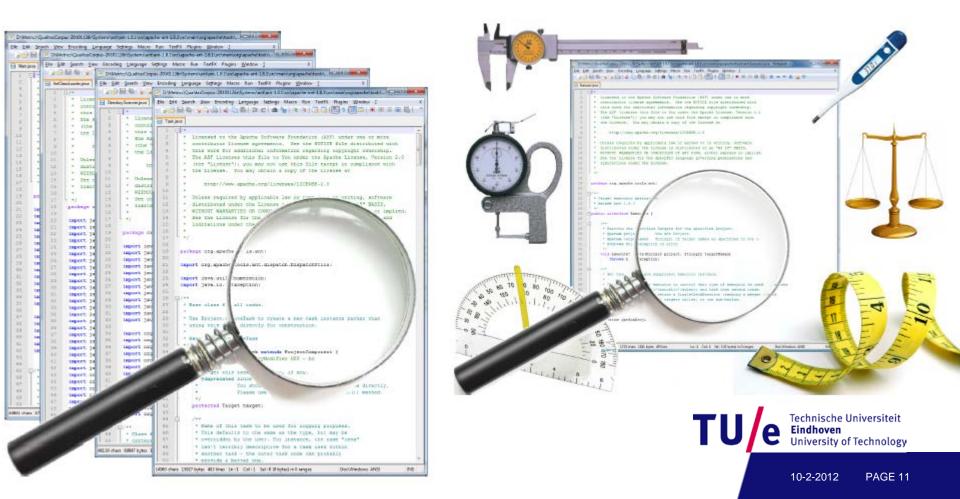
/ department of mathematics and computer science

10-2-2012 Page 10

Two kinds of aggregation

Same metrics, different artifacts

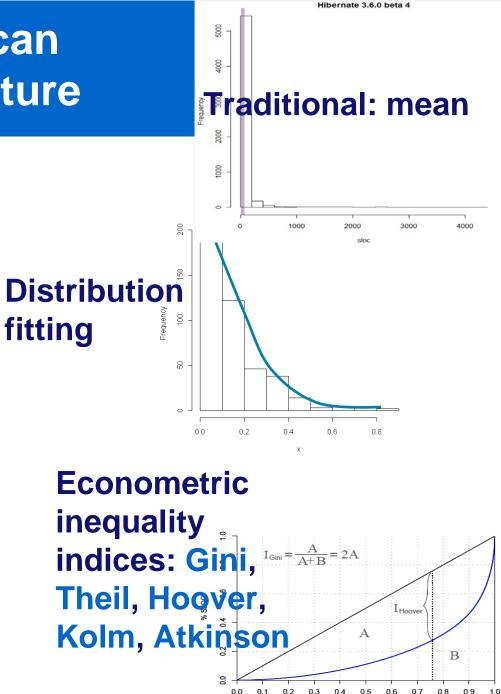
Same artifact, different metrics



Various techniques can be found in the literature

Same metrics, different artifacts

Bit Bit Zept. Ver. Franker Lengers Strag. Here fun Terff Hige: Enter 1 Weiner
Health Des ge man gen finde jungers freque text he fully finder junke I is an intervention is funder was an is an intervention in the funder was an is an intervention is funder was an is an intervention is funder was an is an intervention is funder was an is an isotrom was an intervention is funder was an isotrom was an intervention in the funder was an isotrom was an intervention in the funder was an isotrom was an intervention in the funder was an isotrom was an intervention in the funder was an isotrom was an intervention in the funder was an isotrom was an intervention in the funder was an isotrom was an intervention in the funder was an isotrom was an intervention in the funder was an isotrom was an intervention in the funder was an isotrom was an isotrom was an intervention in the funder was an isotrom was an intervention in the funder was an isotrom was an isotrom was an isotrom was an intervention in the funder was an isotrom was an intervention in the funder was an isotrom was an intervention in the funder was an isotrom was an intervention in the funder was an isotrom was an intervention in the funder was an isotrom was an intervention in the funder was an isotrom was an intervention in the funder was an isotrom was an intervention in the funder was an isotrom was an intervention in the funder was an intervention inthe funder was an intervention intervention inthe funder was an i
Image: State in the state i
Image: Section of the first part of the formation of the section
Image: Second
1 1
1 1
<pre>1</pre>
3 1
3 1
11 1 </th
13 13 14 15 15 15 16 17 17 18 16 16 17 17 18 <
13 1 </th
13 11 13 14 <
13 1 1 + NUTE 10 + 1 13 1 + NUTE 10 + 1 + 1 14 1 + Nute 1 + 1 + 1 14 + Nute 1 + Nute 1 + 1 15 1 + Nute 1 + Nute 1 16 1 Nute Nute 1 + Nute 17 + 1 + Nute 1 + Nute 18 Nute 1 + Nute 1 + Nute 18 Nute 1 + Nute 1 + Nute 18 Nute 10 + Nute 1 + Nute 18 Nute Nute 10 + Nute Nute 18 <
13 13 14 5 15 14 <t< th=""></t<>
13 0 1 </th
13 po 10 - 11 - Ber th 11 - - Ber th 11 - - Ber th 11 - - - Ber th -
23 13 package 10 • instructional works the Lioune (• * Shill, 24 13 10 • instructional works the Lioune (• * Shill, 27 13 10 • instructional works the Lioune (• * Shill, 28 10 • instructional works the Lioune (• * shill, 28 10 • instructional works the Lioune (• * shill, 28 10 • instructional works the Lioune (• instructional works the lioune (28 10 • instructional works the lioune (• instructional works the lioune (28 10 • instructional works the lioune (• instructional works the lioune (29 10 • instructional works the lioune (• instructional works the lioune (29 10 • instructional works the lioune (• instructional works the lioune (29 10 • instructional works the lioune (• instructional works the lioune (29 10 • instructional works the lioune (• instructional works the lioune (29 10 • instructional works the lioune (• instructional works the lioune (29 10 • instructional works the lioune (• instructional works the lioune (
22 10 10 10 11 11 10 <
10 11 1000000000000000000000000000000000000
1 1 1 1 1 1 21 1 1 1 1 1 23 1 1 1 1 24 1 1 1 1 25 1 1 1 1 26 1 1 1 1 27 1 1 1 1 28 1 1 1 29 1 1 1 21 1 1 1 25 1 1 1 26 1 1 1
25 to 12 theorem 1 in theorem 1
27 in 29 manet 20 ii ingert 20 20 sether transfer 10.000
27 th in an import is 24 import int in import cog, spann Cools. art. mapaten huspatenting
ter 24 tegent per 23 tegent per 24 tegent per 24
12 in import 14 import 147 ID import 1476 and Inderstoton)
The second se
The second part of the second pa
and a support ju 27 import jaw 1
at in the second s
and an insert re at a state of a state of the state of th
and all import 3m import ord and a man to the start
A in the second of a state of a state of a second of a
11 import in 33 import ord
10 Q/2 17 Impart pr 11 Impart
A 10 lagest of 1 have a supervised and
at the import of a directly.
A 14 TABLET C. A 11 METSON
postected Tappet target:
AND the II and III and III and III
11 E /**
 Class e 40 This defaults to the same as the type, but may be
and + contains +1 * overstidden he the inter. For instance, the anse "news"
ALLO that SHAT here at a last terribly descriptive for a task west vitin
warring weriges a 40 * another task - the doted task code can probably *
1990 data 1997 bete 43 line 1x1 Col1 Still Bistepini ange DetWindows AVS 16



Mean

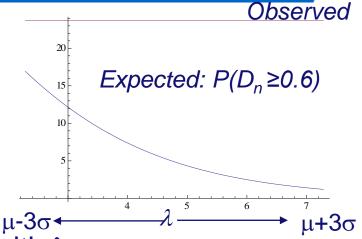
- 3 OS systems: ArgoUML, Adempiere, Mogwai
- Correlation SLOC/bugs
 - Kendall's τ : ranked, does not assume normality
 - SLOC/class
 Bugs: "By No
 Ke
 Mogwai
 0.197
 < 0.01
- Inconsistent results:
- Expresses central tendency, unreliable for skewed distributions

[Vasilescu, Serebrenik, v.d. Brand WETSoM'11 ACM]

e Technische Universiteit Eindhoven University of Technology

Distribution fitting

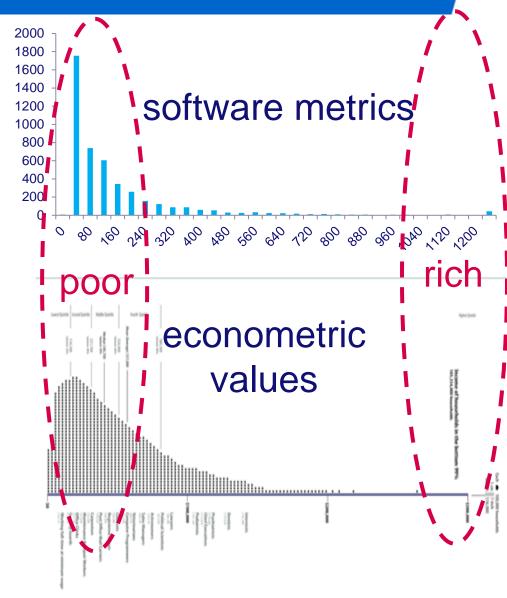
- Useful: estimation of excess values
- Case study
 - 21 Java OSS
 - D_n: abstractness/instability balance
- Distribution
 - Similar to exponential, parameterized with $\boldsymbol{\lambda}$
 - λ of diff benchmarks are normally distributed [Serebrenik, Roubtsov, v.d. Brand ICPC'09 IEEE]
- Fitting: involves guessing the distribution family
 - Controversial for SLOC: log-normal or double Pareto?
 - Even more error-prone for more complex metrics
 - Avoid, if possible...



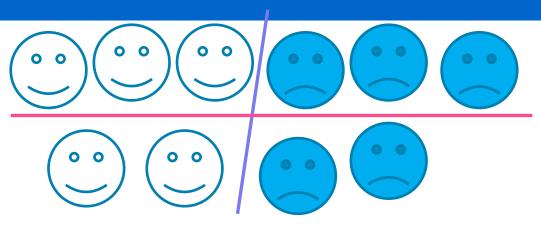
Econometric inequality indices

Measures of inequality

- wealth, expenditure, income
- SLOC, #Classes, function points
- Calculation is metricsindependent
- Some support explanation of inequality:
 - region, education, gender
 - prog. lang., domain, maintainer



Explanation of inequality: Intuition

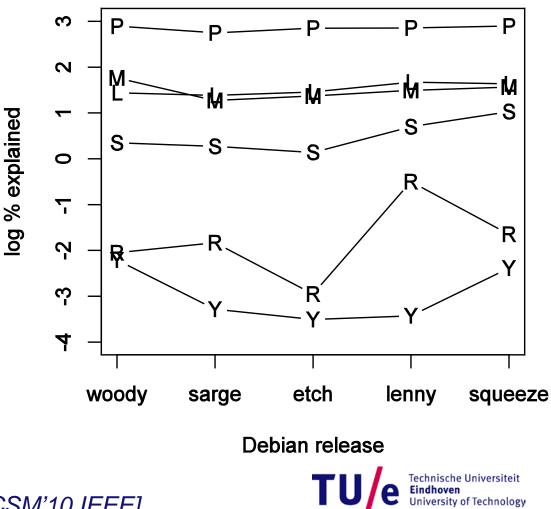


- Why are some people wealthier than others?
- Why are some files larger/more complex than others?
- Partition individuals in groups
 - Partition = explanation
 - Inequality within the groups and between the groups
 - Better explanation: more inequality between the groups
 - Lila is better than red



Explanations: which one is better?

- Theil index
- Debian Linux
 - 1469062 files
 - Metrics: SLOC per file
- Partitions (lenny)
 - Package (17.5%)
 - Maintainer (5.3%)
 - Language (4.5%)
 - Section (2%)
 - Repository (0.6%)
 - PrioritY (0.03%)



[Serebrenik, v.d. Brand ICSM'10 IEEE]

Inequality indices: Which one is better?

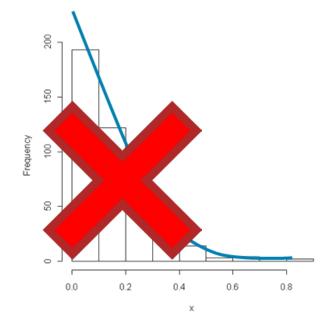
- Theil, Gini, Hoover, Atkinson, Kolm, ...
 - Theil, Gini, Hoover and Atkinson agree
 - Any can be chosen from the correlation point of view
 - Advantages: decomposable (Theil), fixed range (Gini, Hoover), applicable to negative values (Gini, Hoover)
 - Kolm and mean agree
 - for SLOC, not for more advanced metrics
 - Kolm is a better alternative

[Vasilescu, Serebrenik, v.d. Brand ICSM'11 IEEE]

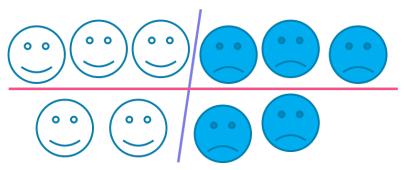


Aggregation techniques: Summary









Econometric inequality indices



/ department of mathematics and computer science

10-2-2012 Page 19

Mining Repositories



/ department of mathematics and computer science

10-2-2012 Page 20



Test #14352 fails sometimes



know how to fix it!

The error should be somewhere here... What does this code do?

sual Studio

SUBVERSION

Foundation

/W&I/MDSE

10-2-2012 PAGE 21

Technische Universiteit

University of Technology



How can the repositories serve you?

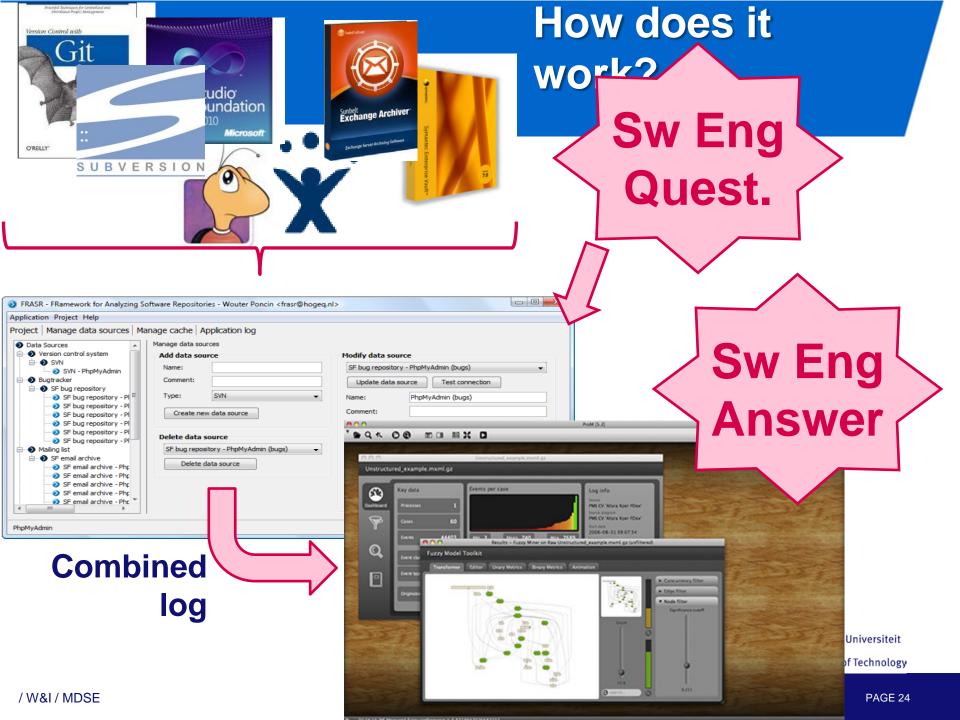
- Is the documentation up-to-date?
- How fast are the bugs resolved?
- Who is responsible for
 - Bugs
 - Overtly complex code
 - Code guidelines violations?
- What parts are covered by tests?





10-2-2012

PAGE 23



Existing work...

Existing repository mining works often limited to

- one repository or one software engineering question
- we need a generic framework

Process mining

		_
Business processes	Software processes	
One data source	Multiple data sources	
"Natural case":association of different eventsclaim id, person id, vehicle id	Many different options files, developers, topics 	
Explicit events	Implicit events is the mail relevant to the bug report? 	
Unique data representation	Different representations in different data sources	niversiteit Technology
nt of mathematics and computer science	10-2-2012	Page 25

Our studies so far

Open-Source software:

- developer roles
- use of Bugzilla (intended vs. actual) [Poncin, Serebrenik, v.d. Brand CSMR 2011 IEEE]

Student capstone projects

- adherence to guidelines
- quality of the development process
- developer roles

[Poncin, Serebrenik, v.d. Brand OOPSLA Comp. 2011 ACM]



10-2-2012

PAGE 26

Case study 1: Developer roles

- **Question:** Classify developers according to their roles
- Classification of Nakakoji et al. IWPSE 2002: 8 roles
 - Core member involved for a relatively long period and made significant contributions to the development and evolution of the system
- Case study
 - **aMSN**: instant messaging application
 - 38 million downloads, 20th most popular at SourceForge
 - February 26, 2002 July 9, 2010
 - 7 bug repositories: 3137 bug reports
 - 3 mail archives: 34947 messages
 - Subversion: 12062 commits



10-2-2012

PAGE 27

Case study 1: Results ProM Dotted Chart visualization

	3 25.5.2004	25.5.2005 2	5.5.2006 19:0 [25.5.2007 2:0:0	24.5.2008 2:0:0	24.5.2009 2:0:0	ź
							(((()-0 -0)))
Philippe Valembois - Phil/lephilo	usophe					• •	
erome Gagnon-Voyer/germi∎atd							
lenrique Cesar Ulbr tor hatredma	m •((((((()))))•)•••	() ••• •					
/ivia Nikolaidou/vivia		• • •		(** •• • • •()••	• • • • •		
Boris Faure/billiob	•						
arel Demeyer/scapor	-				•	•	
			• 🕞				
ander Hoentjen/tjikkun	• • • • • • • • • • • • • • • • • • • •			() () () () () () () () () () () () () (
'om Hennigan/tomhennigan				•••••••			
(arol Kr®ka/kkrizkæ©	() () • • • • • • • • • • • • • • • • •		• •				
rederic Diaz/gadget_boy		•••••••••	•				
larry Vennik/thaven	• • • • • • • •				•		
'homas Geirhovd/tg90nor		• • • •					
Mark/koenijn	• • • • • • • • • • • • • • • • • • • •						
,	File added Bug opened	Mail Bug other	File mo	dified			
evelopers					TU	Eindhoven	Universiteit of Technology
and the second							5105 00

Timo

10-2-2012

PAGE 29

/ Department of Mathematics and Computer Science

Case study 1: Results

Core developers (examples)

10-2-2012

PAGE 30

	2003 25.5.2004	25.5.2005 2 ⁻ P-0	25.5.2006 2'9'0	25.5.2007 2:0:0	24.5.2008 2:0:0	24.5.2009 2:0:0	24.5 2:0:0
Alaoui Youness/kakarot							
Philippe Valembois - Phil/lep	hilousophe				••••	• • •	•••
Jerome Gagnon-Voyer/germi					•		•
Henrique Cesar Ulbr en hatre	dman •(((())))		_				
Vivia Nikolaidou/vivia				•••••			
EProblem Arieh Schneier/lio lion	in the orig		ificatio	o <mark>hoora</mark>			
Karel Demeyer/scapor				•		•	
Tom Jenkins/bluetit			• 🕒			•	
Sander Hoentjen/tjikkun	•					•	
Tom Hennigan/tomhennigan				(()C ()C (()		••• ••••	
Karol Kr©ka/kkrizka© Frederic Diaz/gadget boy			Pe	eripher	al deve	elopers	
Harry Vennik/thaven							
Thomas Geirhovd/tg90nor					• ••		
Mark/koenijn	•	9- 0					
	File added Bug opened	Mail Bug other	File m	odified	Bu	g report	er
					Τυ/	e Technische Univers Eindhoven University of Techn	

/ Department of Mathematics and Computer Science

Case study 1: Classification

Role	#devel	#developers		
Bug reporter	1443			
Bug fixer	3	Bugs are usually fixed		
Peripheral developer	29	by peripheral		
Active developer	6	developers		
Core member	7			
Project leader	3	Only ticket-		
Other	234	commented or		
Total	1725	mail-reply		

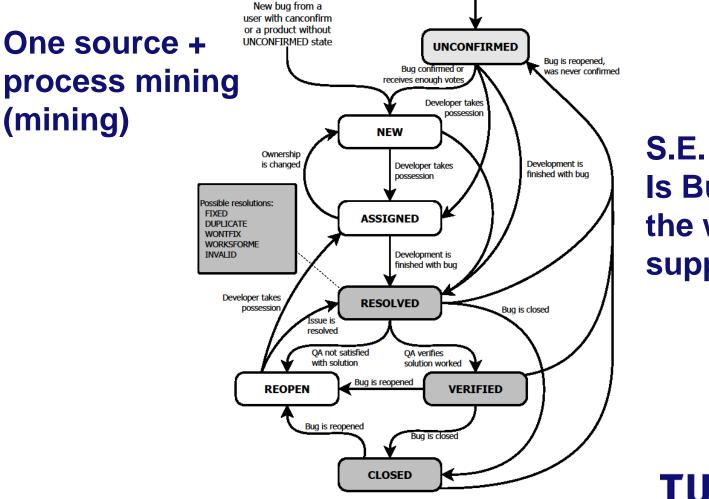
/ Department of Mathematics and Computer Science

10-2-2012 PAGE 31

Technische Universiteit **Eindhoven** University of Technology

TU

Case study 2: Bug life cycle in Bugzilla Theory according to the Bugzilla Guide



S.E. question: Is Bugzilla used the way it is supposed to be?

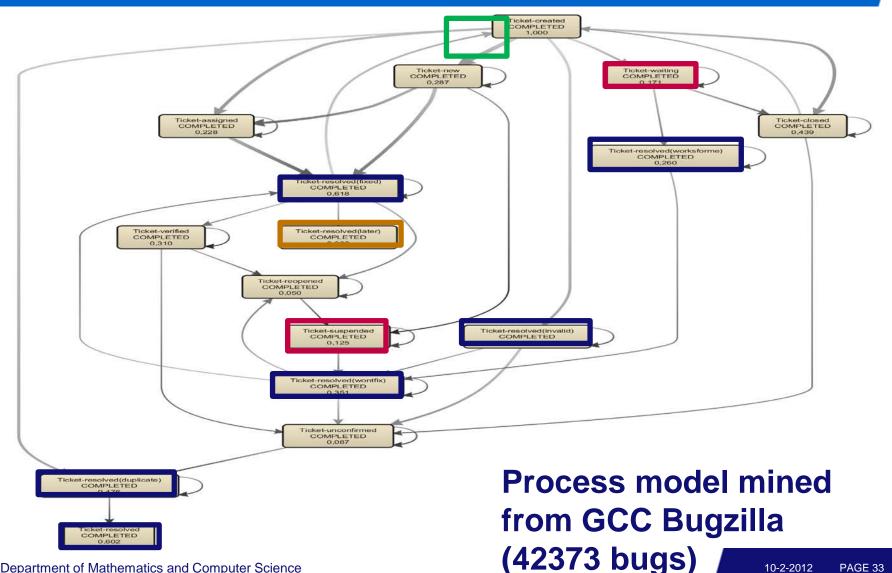
10-2-2012 PAGE 32

Technische Universiteit

University of Technology

Eindhoven

Case study 2: Bug life cycle in Bugzilla **Practice vs. Theory**



/ Department of Mathematics and Computer Science

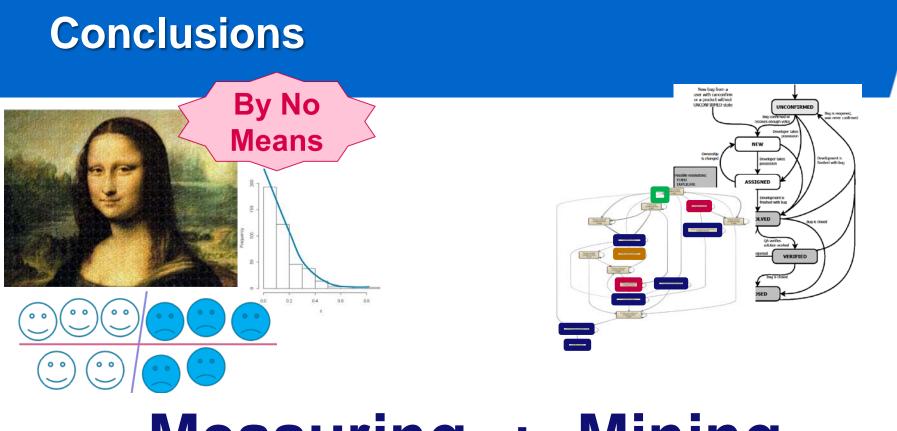


Information is available in software repositories

- Just waiting to be mined
- Numerous opportunities and chances







Measuring + Mining Interested? Join us!



/ department of mathematics and computer science

10-2-2012 Page 35