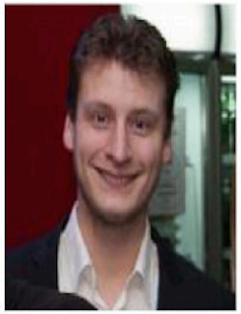
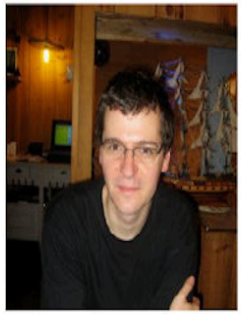
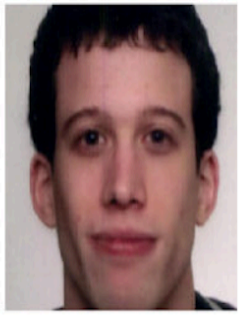
Alexander Serebrenik
Eindhoven University of Technology, NL

Human Aspects of Software Engineering

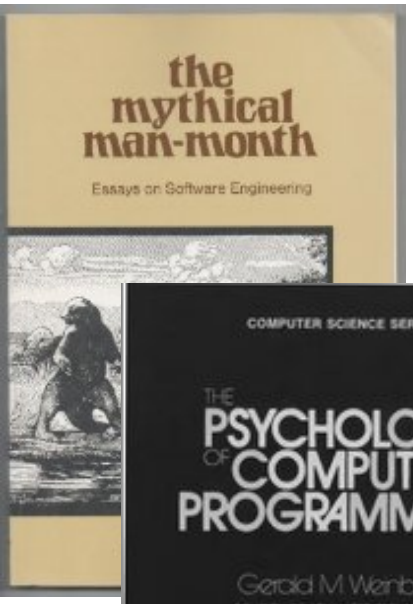




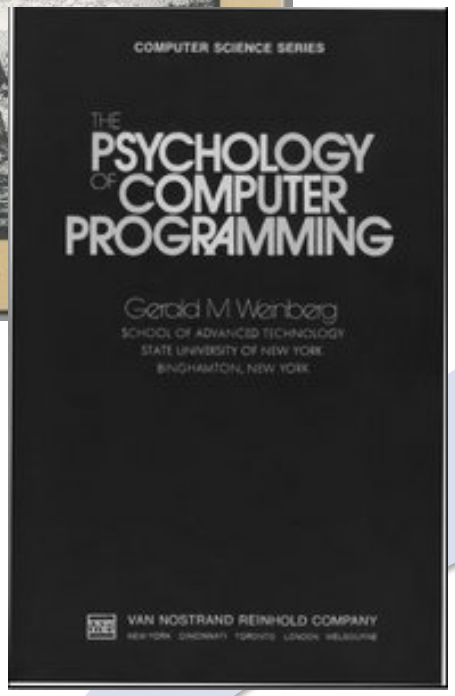





And you!





1970s

1998

MSR

SocialCom,
SocInfo, SBP,
CHASE

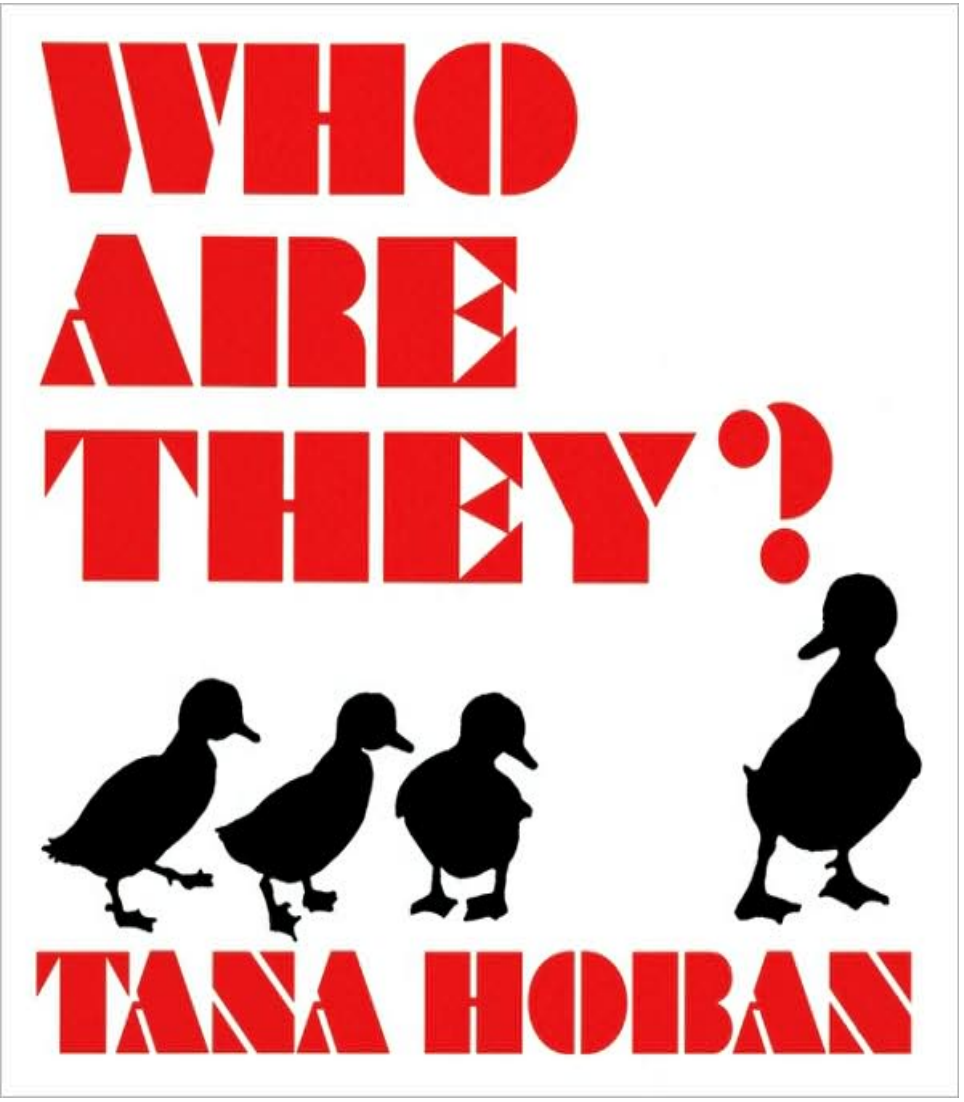
2008

books,
conferences

2014

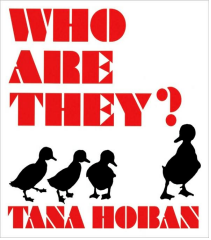
open source

MS Windows post-release fault prediction	**Precision** predicted & correct / predicted	**Recall** predicted & correct / correct
Code churn	78.6%	79.9%
Code complexity	79.3%	66.0%
Code coverage	83.8%	**54.5%**
Code dependencies	74.4%	69.9%
Organizational structure	**86.2%**	**84.0%**
Socio-technical network	76.9%	70.5%


WHO
ARE
THEY?
TANA HOBAN


what
they do
in the
dark
AMANDA COE


WHO ARE THEY?
TANA HOBAN
MEN
> 90% in WordPress & Drupal
> 95% in FLOSS surveys
> 87% in GNOME
> 70% in software-related jobs (NSF)



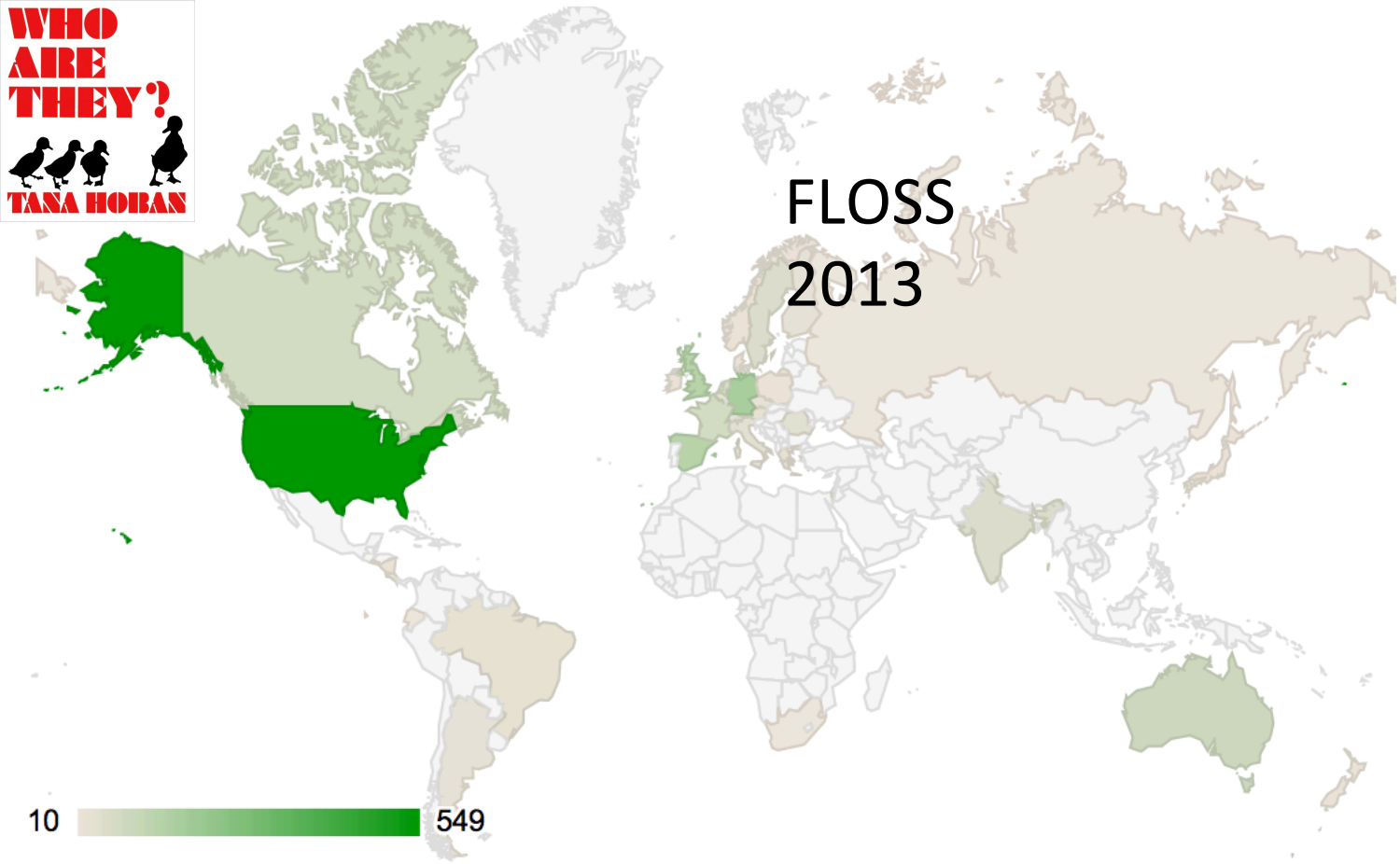
median

stack overflow

FLOSS 2013

YOUNG(?)

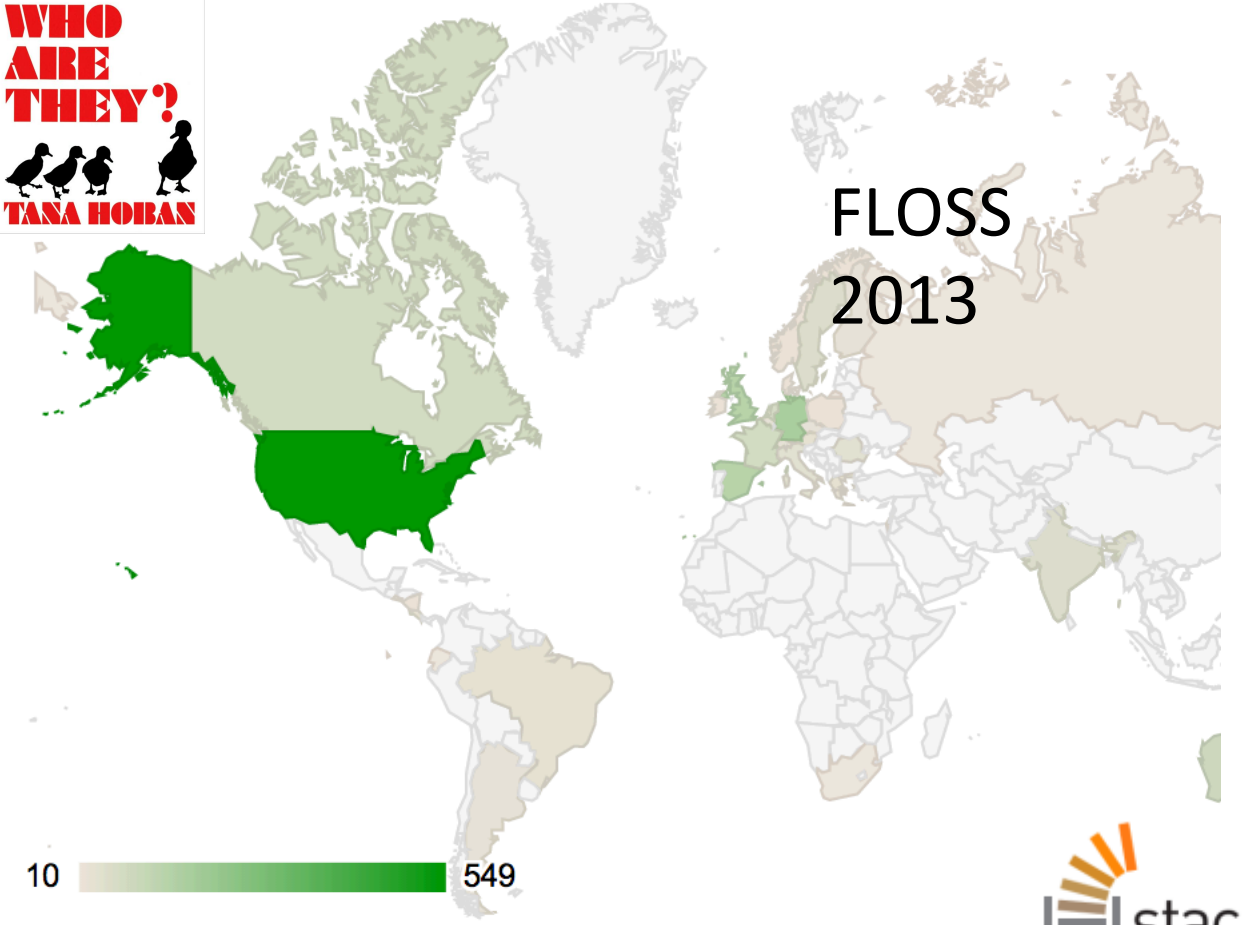
Is Programming Knowledge Related to Age? An Exploration of Stack Overflow. Morrison, P., Murphy-Hill, E. MSR 2013. FLOSS 2013: A survey dataset about free software contributors: challenges for curating, sharing and combining. Robles, G., Arjona-Reina, L., Vasilescu, B., Serebrenik, A., Gonzalez-Barahona, J.M. MSR 2014


WHO ARE THEY?
TANA HOBAN
FLOSS
2013
10
549

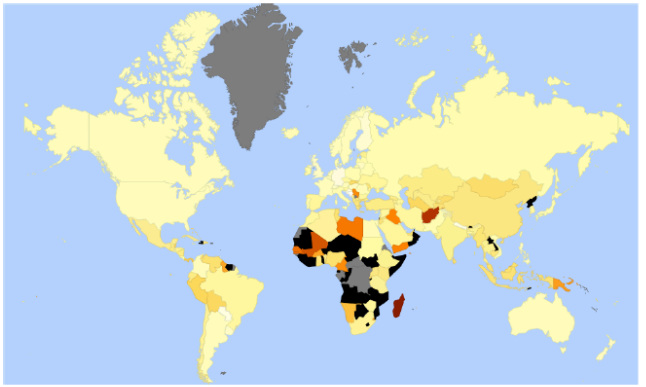
WHO ARE THEY?

TANA HOBAN

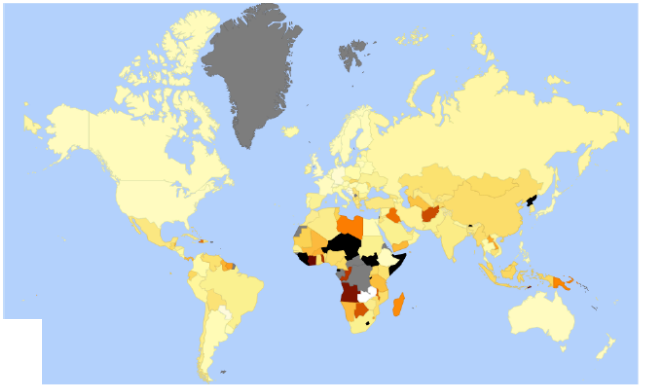
FLOSS 2013



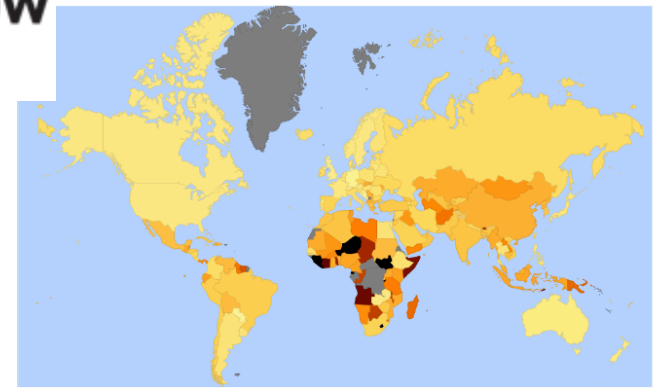
stack overflow



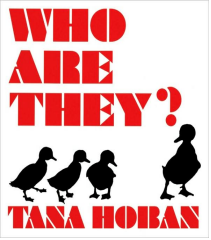
(d) Until June 2010

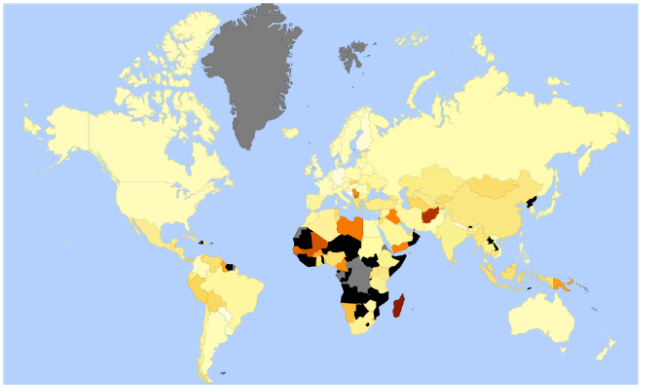


(g) Until December 2011

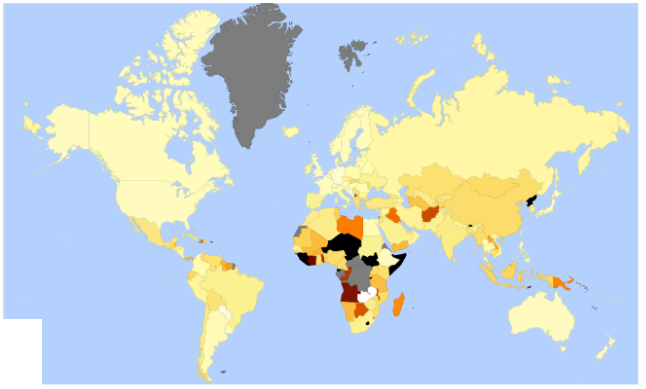


(j) Until June 2013

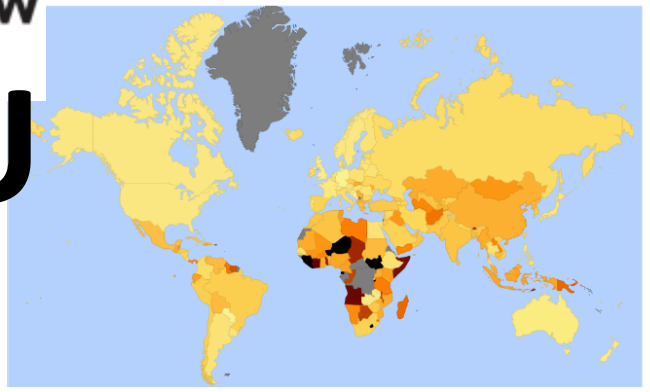




(d) Until June 2010



(g) Until December 2011



(j) Until June 2013

stack overflow

Europe,US,CA,AU
Brazil/Argentina


what
they do
in the
dark
AMANDA COE


stack
overflow
what they do in the dark
AMANDA COE


.cpp


.po


/test/


.jpg


/library/

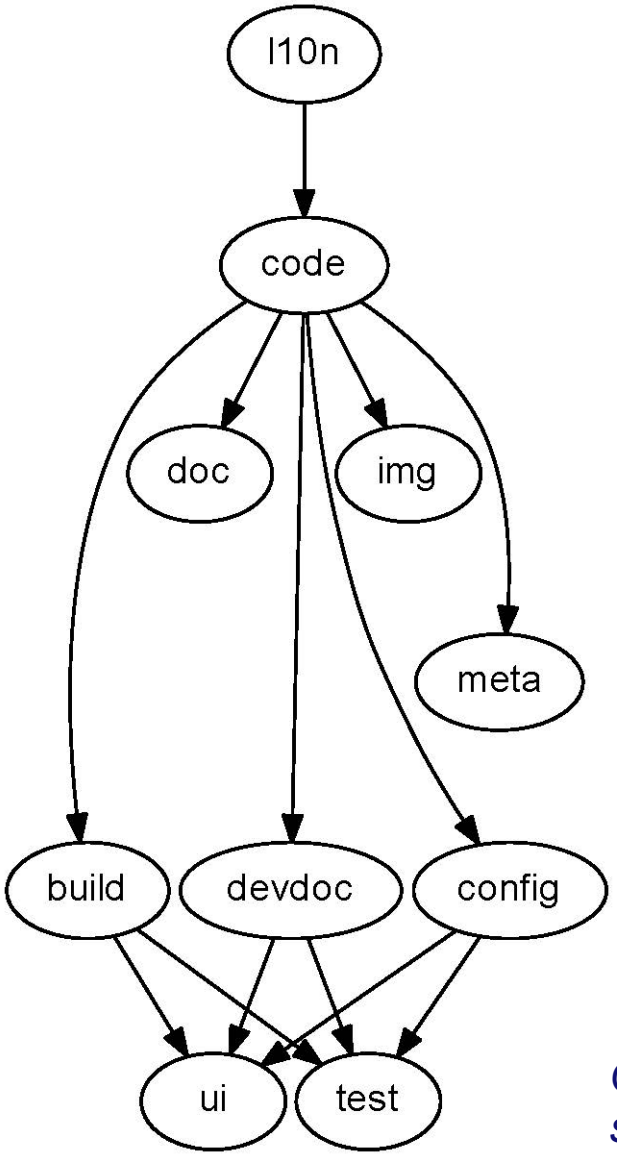

.doc


makefile


.sql

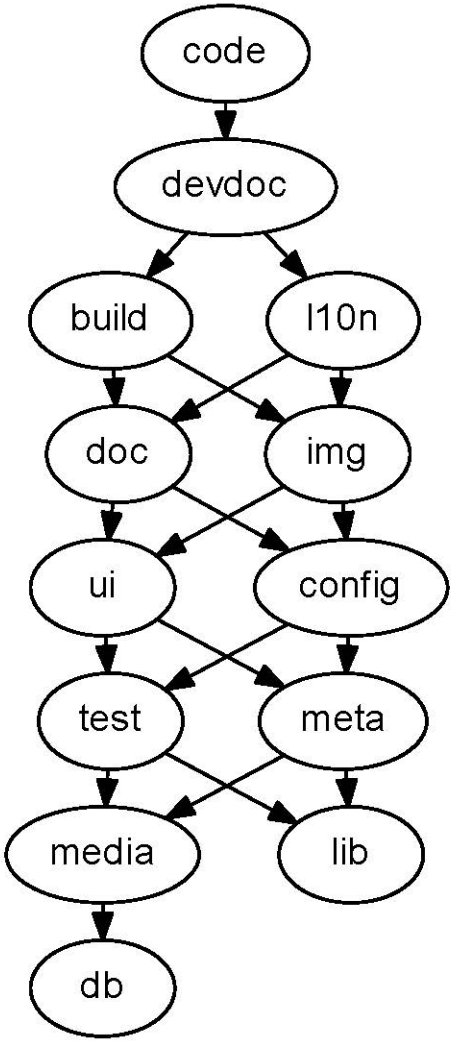

.conf

Occasional contributors



GNOME

Frequent contributors



On the variation and specialization of workload---A case study of the Gnome ecosystem community Vasilescu, B., Serebrenik, A., Goeminne, M., Mens, T. Empirical Sw Engg

For which of those activities would Joe Average be more active than Jane Average?

For which activities no differences would be observed?


.cpp


.po


/test/


.jpg


/library/


.doc


makefile


.sql


.conf


.cpp


.po


makefile


.sql

No differences in *preferences* for activities, differences in the *amount* of commits

stack overflow

sample

WordPress

Drupal

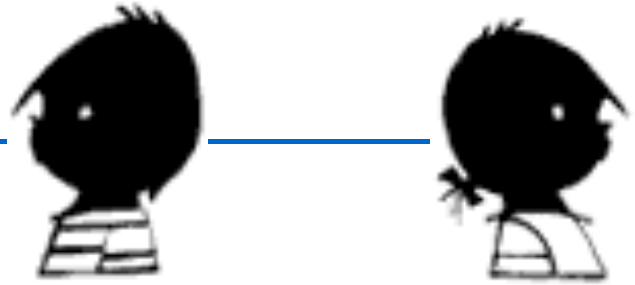
Engage for longer

Ask more questions

No diff in #answers

Women can contribute to SO but choose not to!

Gender, representation and online participation: A quantitative study, Vasilescu, B., Capiluppi, A., and Serebrenik, A., Interacting with Computers



stack overflow

sample

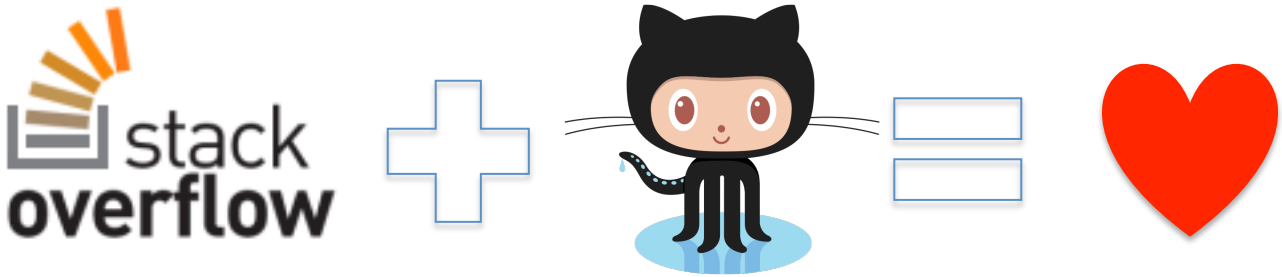
WordPress

Drupal

No significant differences in #questions, #answers, length of engagement

Disengagement of women is activity/ platform specific!


stack
overflow
what they do in the dark
AMANDA COE



- More active committers ask more on SO
- More active askers commit more on GitHub

StackOverflow and GitHub: Associations between software development and crowdsourced knowledge, Vasilescu, B., Filkov, V. and Serebrenik, A., In Social Computing, 2013, IEEE.

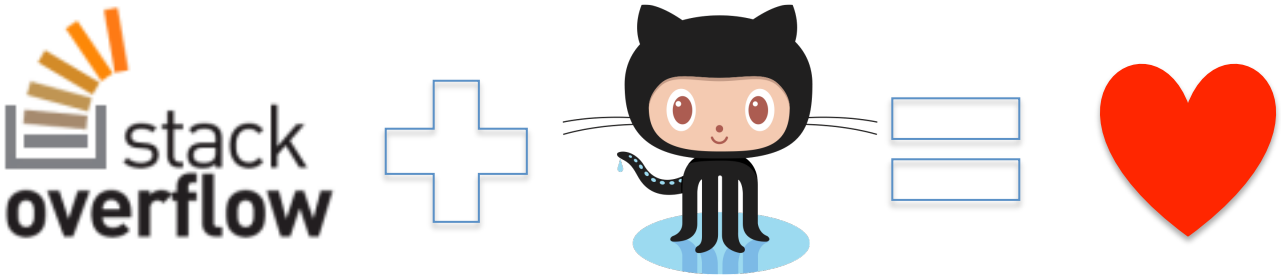


TABLE II: Mutual influence of StackOverflow activities and GitHub committing, for different committers (from least active $Q1$, to most active $Q4$).

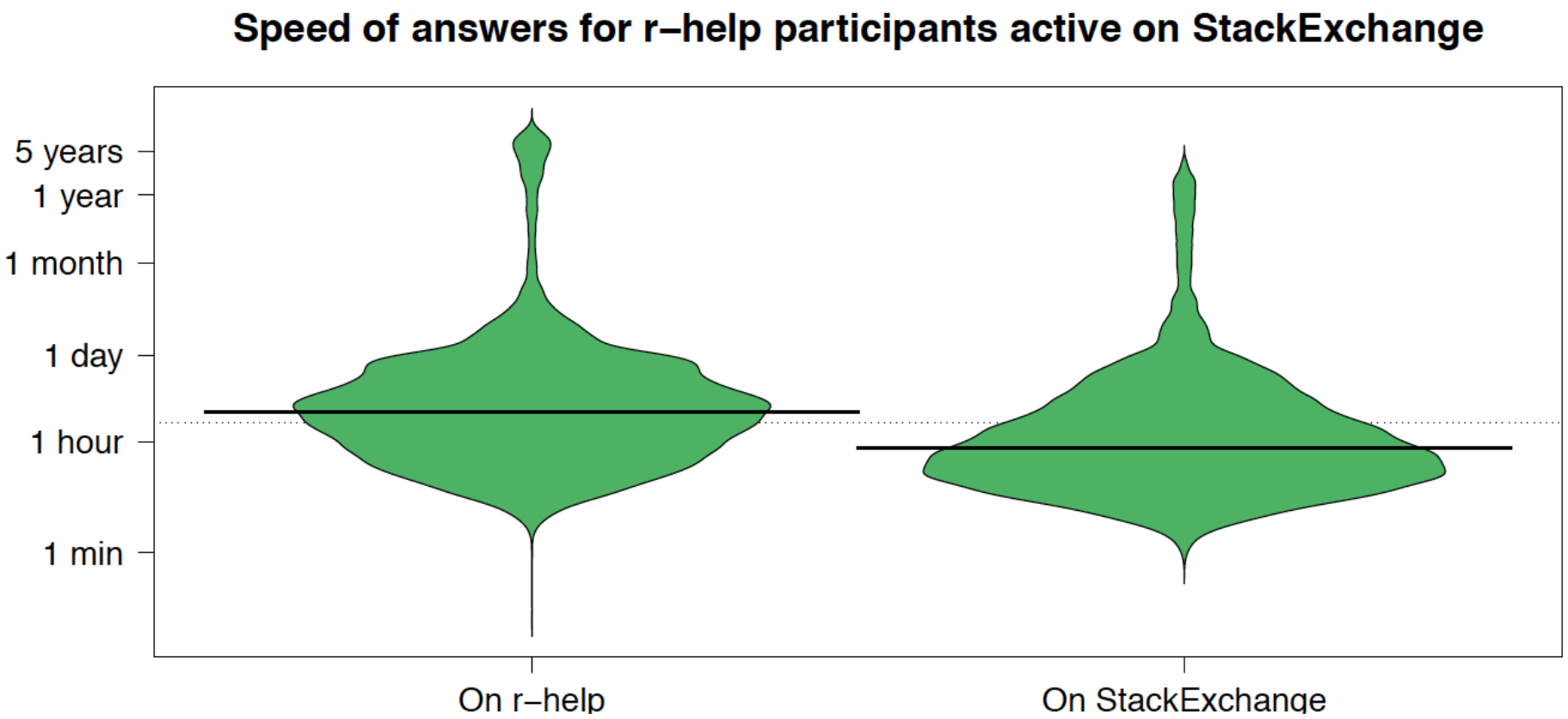
Q	Influence of			
	asking on committing	committing on asking	answering on committing	committing on answering
$Q1$	none	none	none	none
$Q2$	none	inconclusive	inconclusive	inconclusive
$Q3$	accelerates	accelerates	accelerates	accelerates
$Q4$	accelerates	accelerates	accelerates	accelerates

StackOverflow and GitHub: Associations between software development and crowdsourced knowledge, Vasilescu, B., Filkov, V. and Serebrenik, A., In Social Computing, 2013, IEEE.

- Contributors to both communities (more likely developers than non-developers) are more active than those who focus on just one.

How Social Q&A sites are changing knowledge sharing in open source software communities, Vasilescu, B., Serebrenik, A., Devanbu, P.T. and Filkov, V., In CSCW 2014, ACM.

stack overflow + R help = ♥



How Social Q&A sites are changing knowledge sharing in open source software communities, Vasilescu, B., Serebrenik, A., Devanbu, P.T. and Filkov, V., In CSCW 2014, ACM.

Older and Wiser... Up to a Point

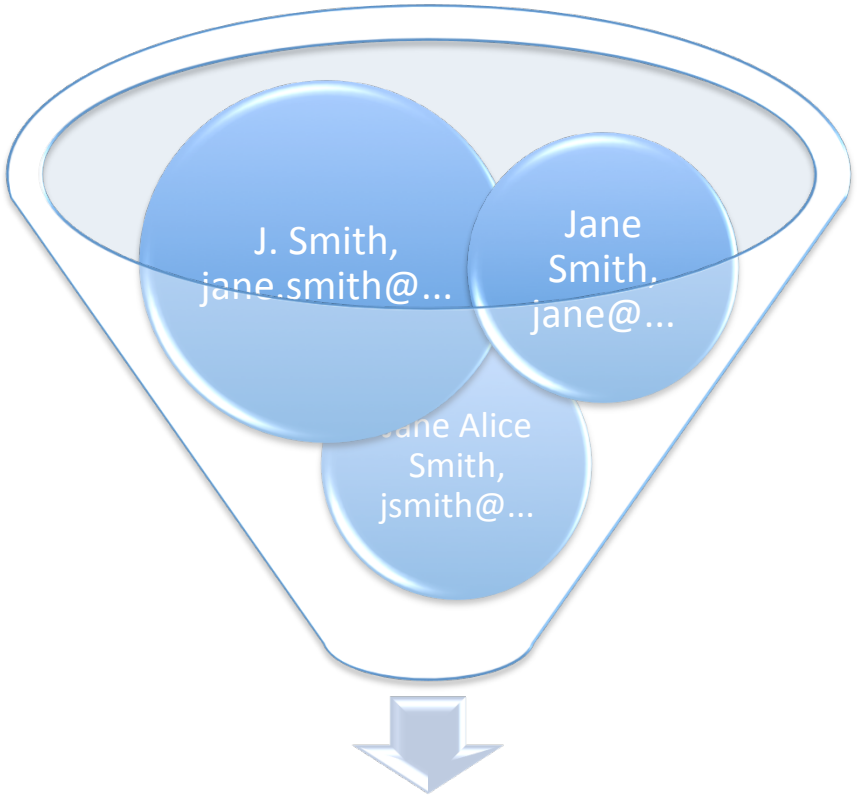
By Philip E. Ross

Posted 3 May 2013 | 11:21 GMT

Tools and Techniques


Robin Williams,
robinw@gmail.com
Euphegenia Doubtfire,
euphegenia@hotmail.com
WHO ARE THEY?
TANA HOBAN

Ordering	Rajesh Sola	Sola Rajesh
Spelling: misspelling, diacritics, punctuation	Rene Engelhard	Fene Engelhard
	Démurget	Demurget
	J. A. M. Carneiro	J A M Carneiro
Middle initials, patronyms, nicknames, additional surnames, incomplete names	Daniel M. Mueth	Daniel Mueth
	Alexander Alexandrov Shopov	Alexander Shopov
	Carlos Garnacho Parro	Carlos Garnacho
	Jacob “Ulysses” Berkman	Jacob Berkman
	A S Alam	Amanpreet Singh Alam
Name variants: transliteration, diminutives	Γιωργοσ	Georgios
	Mike Gratton	Michael Gratton
Software-specific: usernames, projects, tooling artefacts	mrhappypants	Aaron Brown
	Arturo Tena/libole2	Arturo Tena
	(16:06) Alex Roberts	Alex Roberts
Mix	Any combination of those	



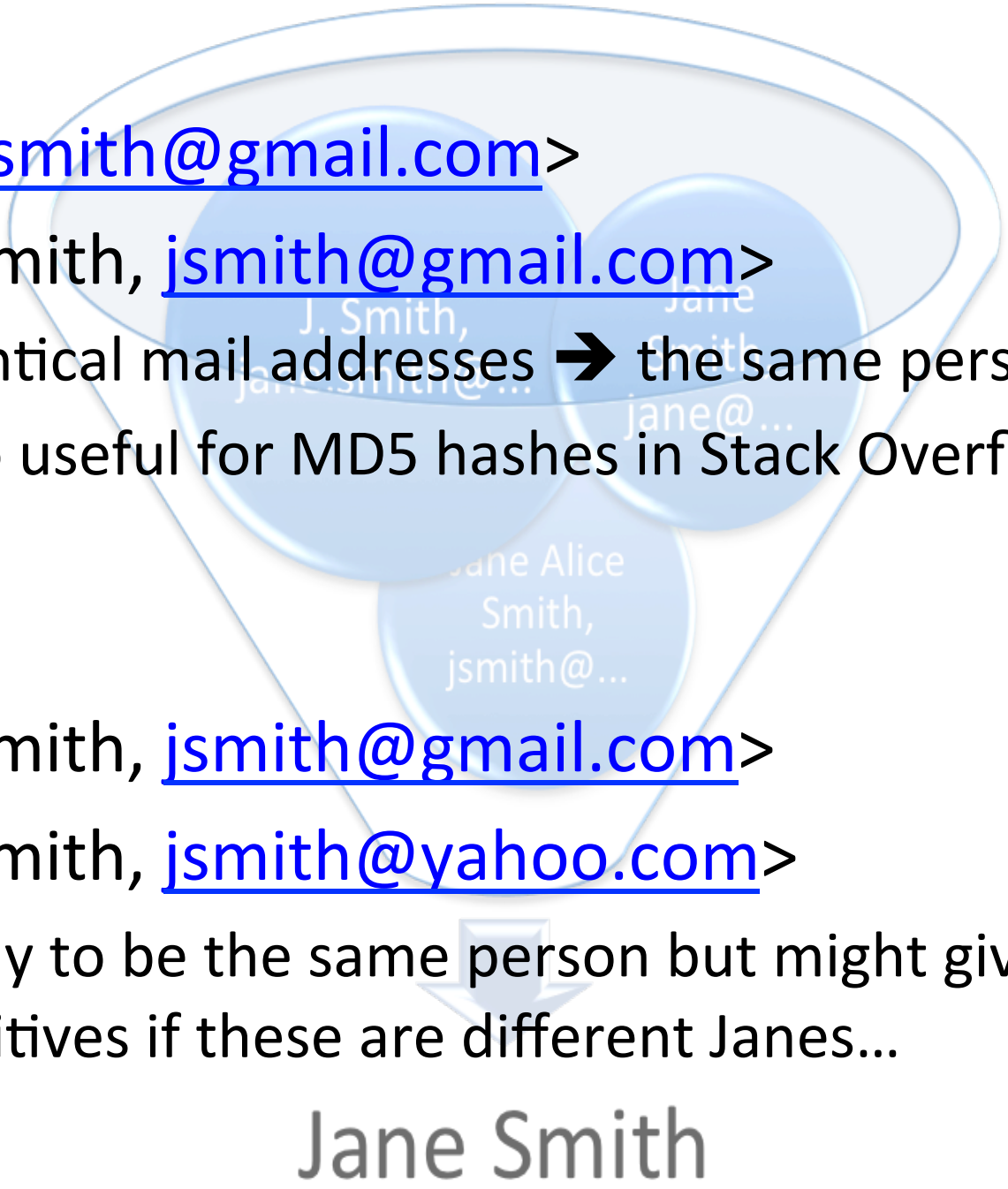
Jane Smith

identity merging/
identity reconciliation/
developer matching

<Jane, jsmith@gmail.com>

<Jane Smith, jsmith@gmail.com>

- identical mail addresses ➔ the same person
- also useful for MD5 hashes in Stack Overflow



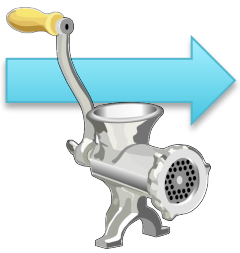
<Jane Smith, jsmith@gmail.com>

<Jane Smith, jsmith@yahoo.com>

- likely to be the same person but might give false positives if these are different Janes…

Latent Semantic Analysis

<John Doe, johnd@domainA>
<John Joseph Doe, johnd@domainA>



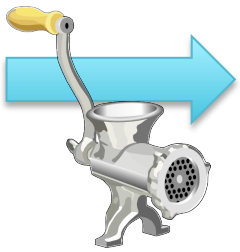
johnd@domainA:
{john, johnd, joseph, doe}

Document-term matrix

	johnd@...	j.doe@...		
john	1	..	..	..
johnd	1	..	..	..
joseph	1	..	..	..
jdoe	?	..	..	..
doe	1	..	..	..

Latent Semantic Analysis

<John Doe, johnd@domainA>
<John Joseph Doe, johnd@domainA>



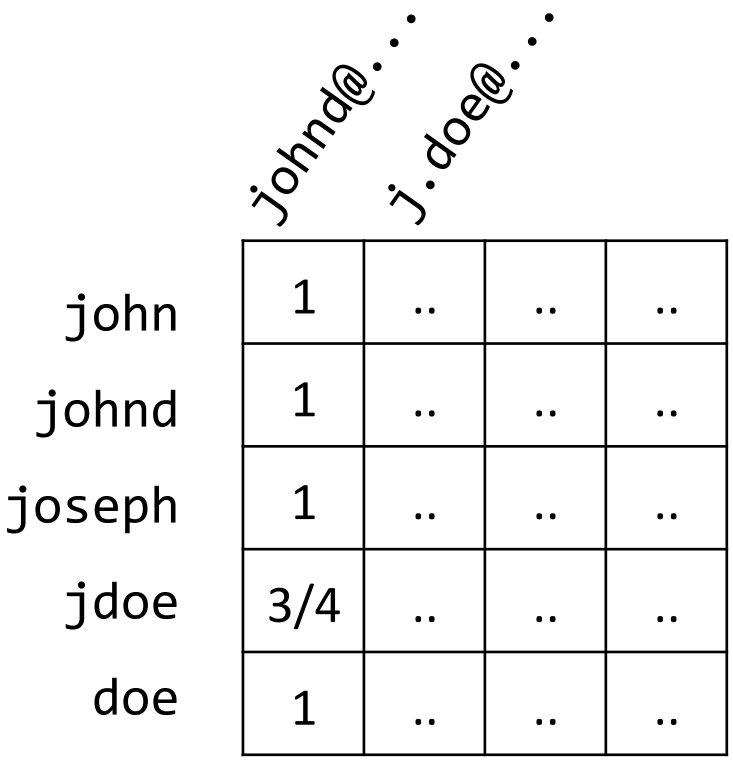
johnd@domainA:
{john, johnd, joseph, doe}

Document-term matrix

	johnd@...	j.doe@...		
john	1	..	..	..
johnd	1	..	..	..
joseph	1	..	..	..
jdoe	3/4	..	..	..
doe	1	..	..	..

max similarity(jdoe,
{john, johnd, joseph, doe})
= similarity(jdoe, doe)
= 1 – Levenshtein(jdoe, doe) /
max(length(jdoe), length(doe))
= 1 – 1/4 = 3/4

Latent Semantic Analysis



Inverse document frequency

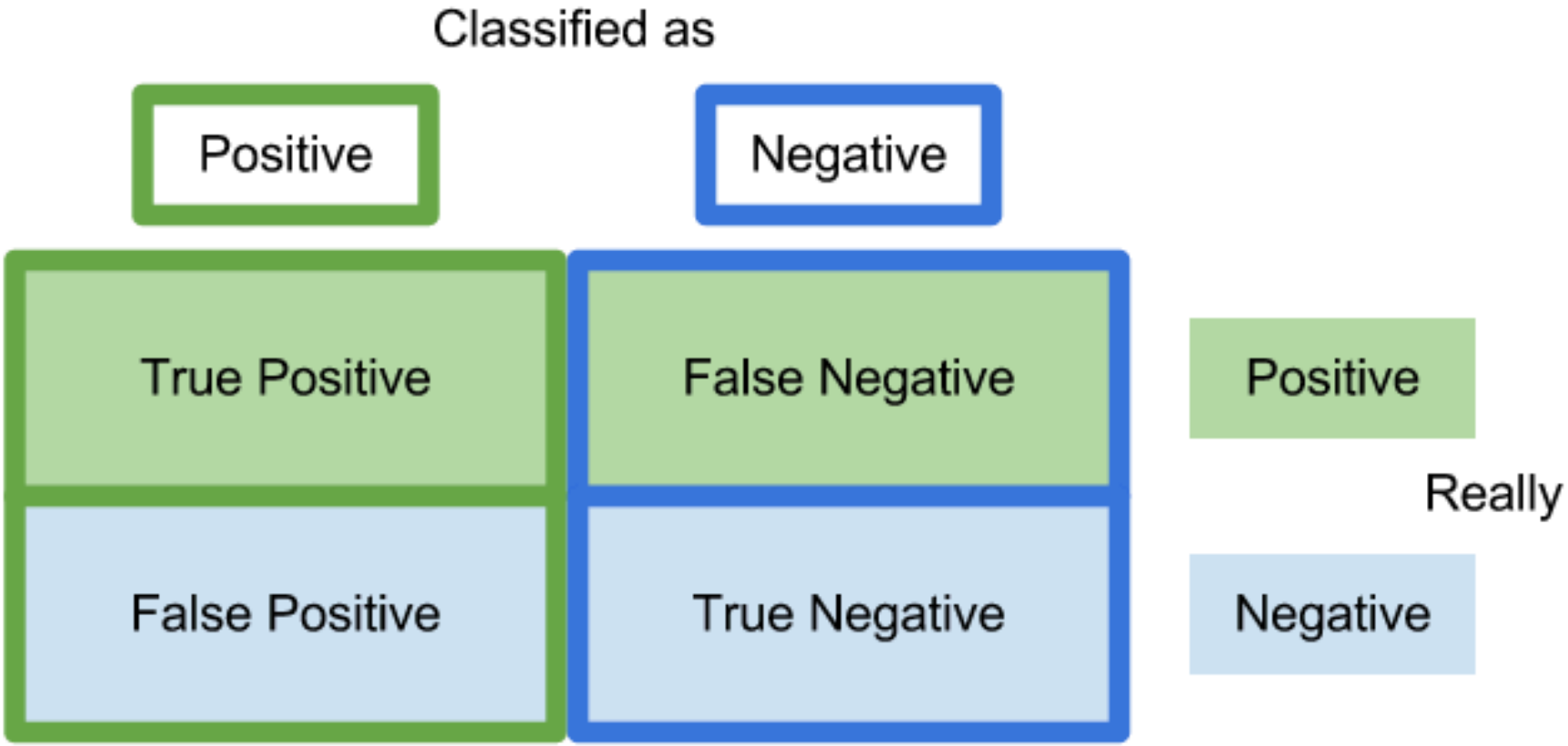
(Singular value decomposition)

Rank (noise) reduction

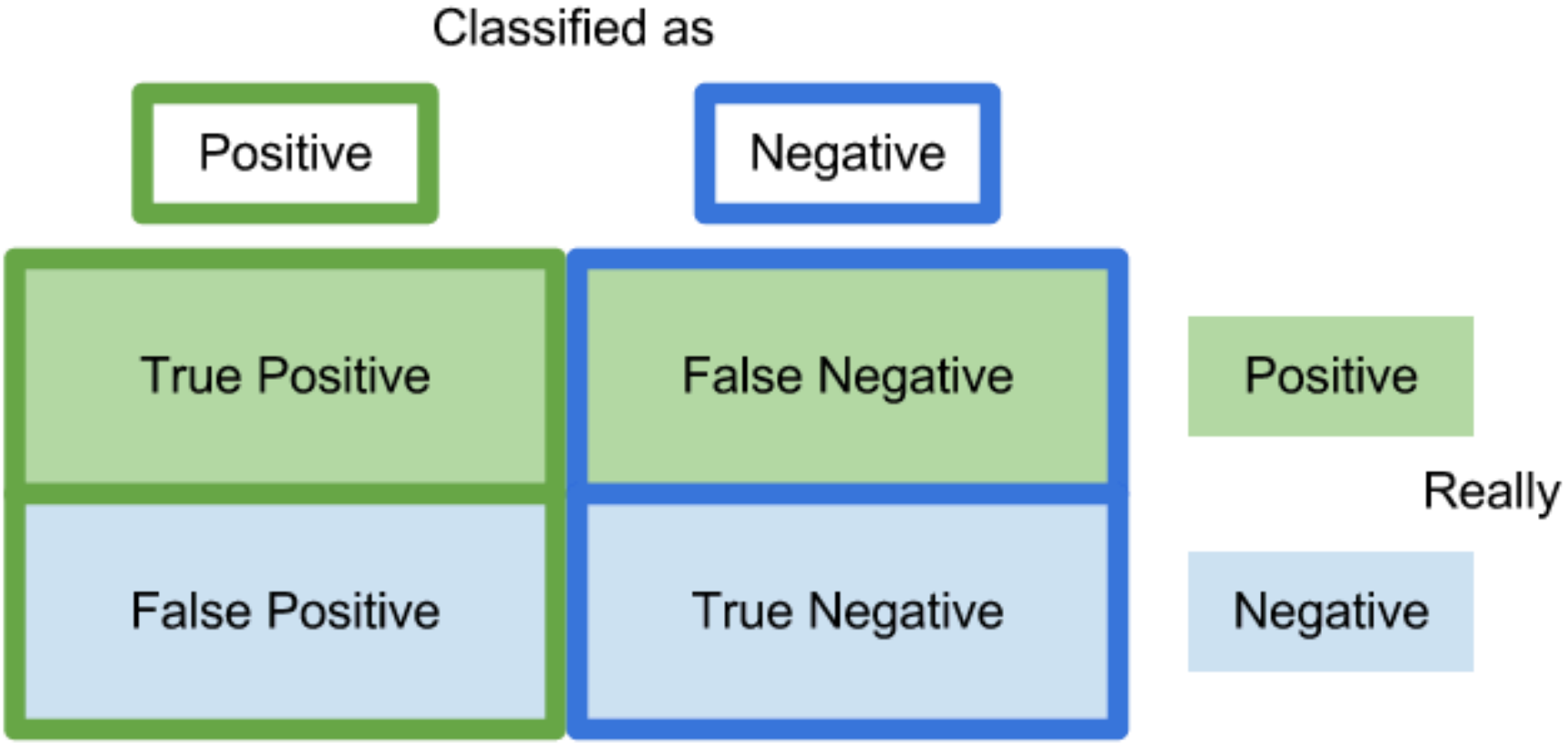
Cosine between documents

Merge similar documents

Which technique is better?



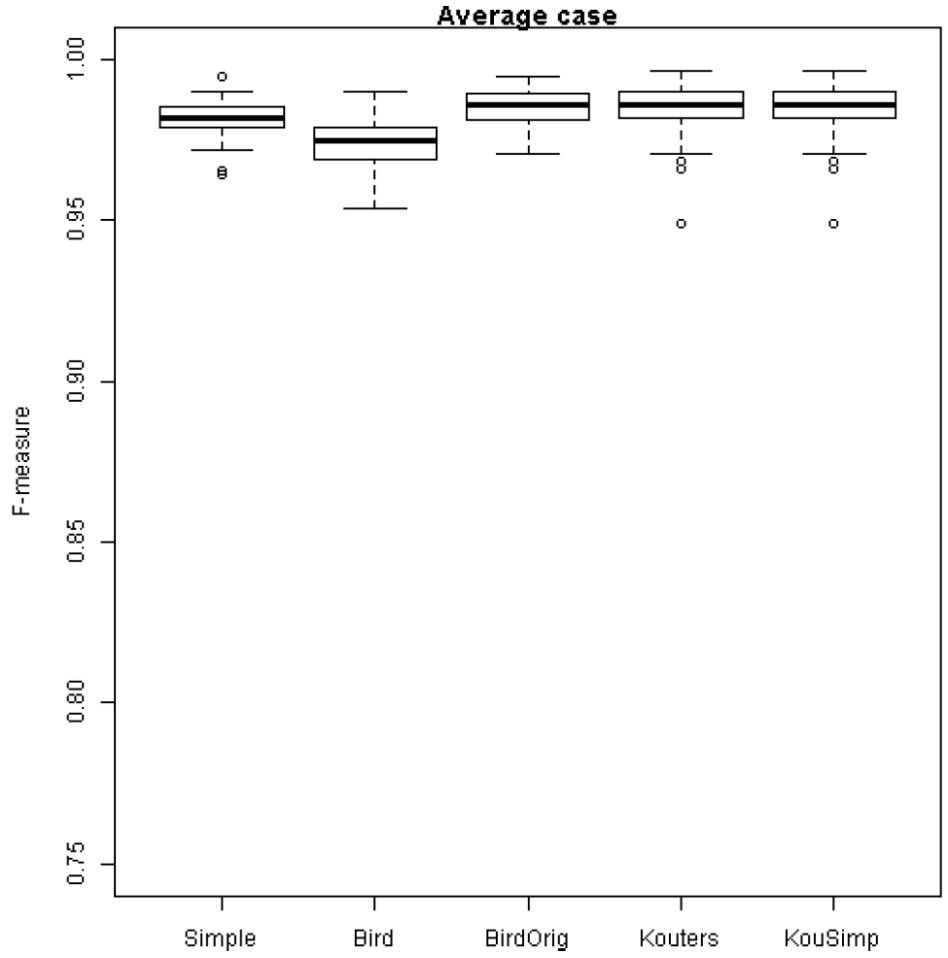
http://uberpython.wordpress.com/2012/01/01/precision-recall-sensitivity-and-specificity/

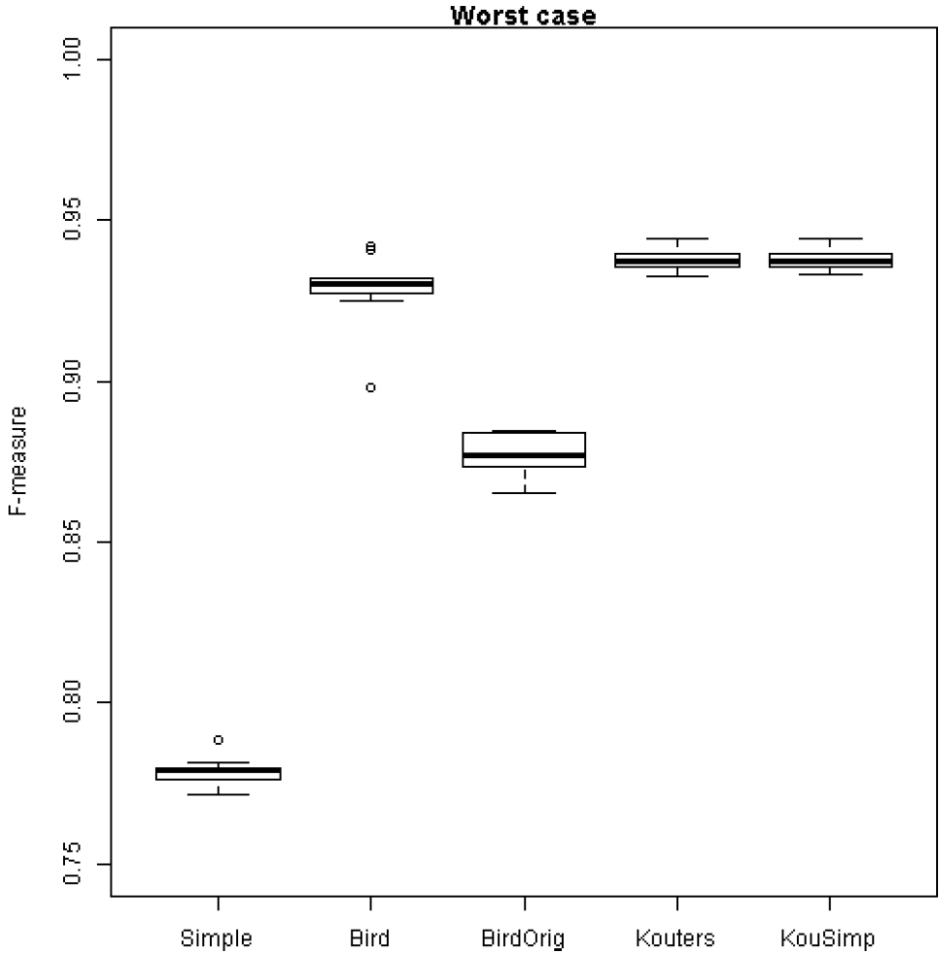


$$\text{Precision} = \frac{tp}{tp + fp} \qquad \text{Recall} = \frac{tp}{tp + fn}$$

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

http://uberpython.wordpress.com/2012/01/01/precision-recall-sensitivity-and-specificity/





Most contributors have only one “identity”: all techniques are good enough for the average case.

Advanced techniques are better in presence of noise.

Choose your weapons wisely!


What is your gender?

Sara Chipps less info

bio

visit

stat

4,168
reputation

● 5 ● 35 ● 83

Sara Chipps less info

bio

visits

stats

4,168
reputation

5 35 83

Andrea Ambu less inf

bio

visi

Andrea Smith less

1
reputation

Sara Chipps less info

bio

visits

stats

4,168
reputation

5 35 83

Andrea Ambu less info

bio

visits

stats

website	andreaa.com
location	Italy
age	25
member for	5 years, 1 months
seen	Oct 20 at 11:01
profile views	1,232

Andrea Smith less in

1
reputation

Sara Chipps less info

bio

visits

stats

4,168
reputation

5 35 83

Andrea Ambu less info

bio

visits

website	andreaa.com
location	Italy
age	25
member for	5 years, 1 months
seen	Oct 20 at 11:01
profile views	1,232

Andrea Smith less in

1
reputation

Name + Location = Gender

vsushkov less info

bio
website vsushkov.com
location Taganrog, Russia
age 23

visits
member for 3 years, 3 months
seen 15 hours ago

stats
profile views 188

1,678
reputation

●1 ●5 ●15

Lonzo less info

bio

visits

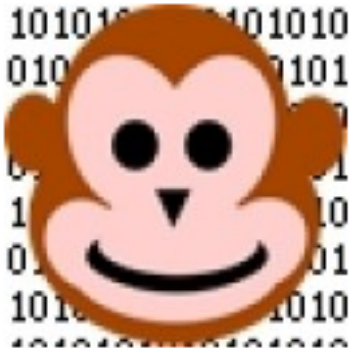
stats

1,177
reputation

●3 ●12 ●18

Lonzo ⇒ Alonzo

w35l3y less info



bio

visits

stats

1,908
reputation

●9 ●27

w35l3y ⇒ wesley

Name +
Location =
Gender

Heuristics:
title + first h1

Ben Kamens

is lead dev at **Khan Academy**, and has been a proud part of **Fog Creek**

<title>Ben Kamens</title>
…
<h1>We’re willing to be embarrassed about what we haven’t done…</h1>

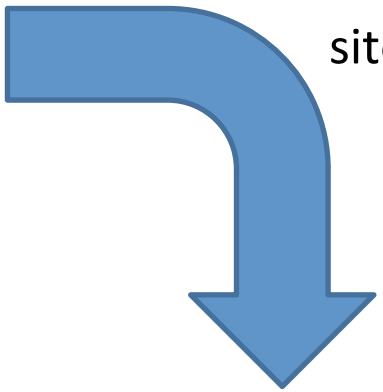
Ben Kamens We're willing to be embarrassed about what we haven't done…

Stanford Named Entity Tagger

<PERSON>Ben Kamens</PERSON> We're willing to be embarrassed about what we haven't done…

Quality of gender resolution: Survey

Self-identification	As inferred			Total
	M	F	?	
M	**60**	3	*43*	106
F	2	**5**	*4*	11



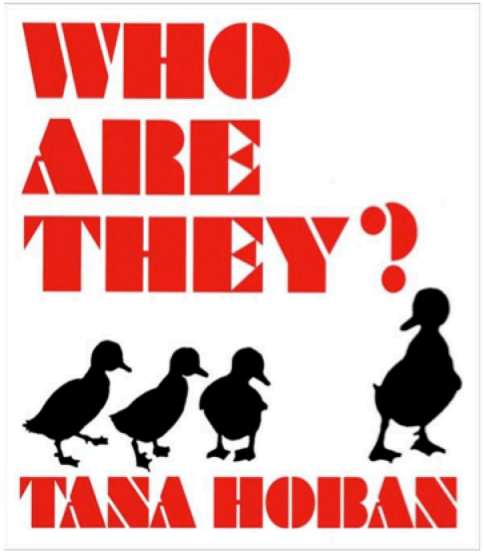
+ avatars, other social media sites (manually)

Self-identification	As inferred			Total
	M	F	?	
M	**90**	3	*13*	106
F	2	**9**	*0*	11

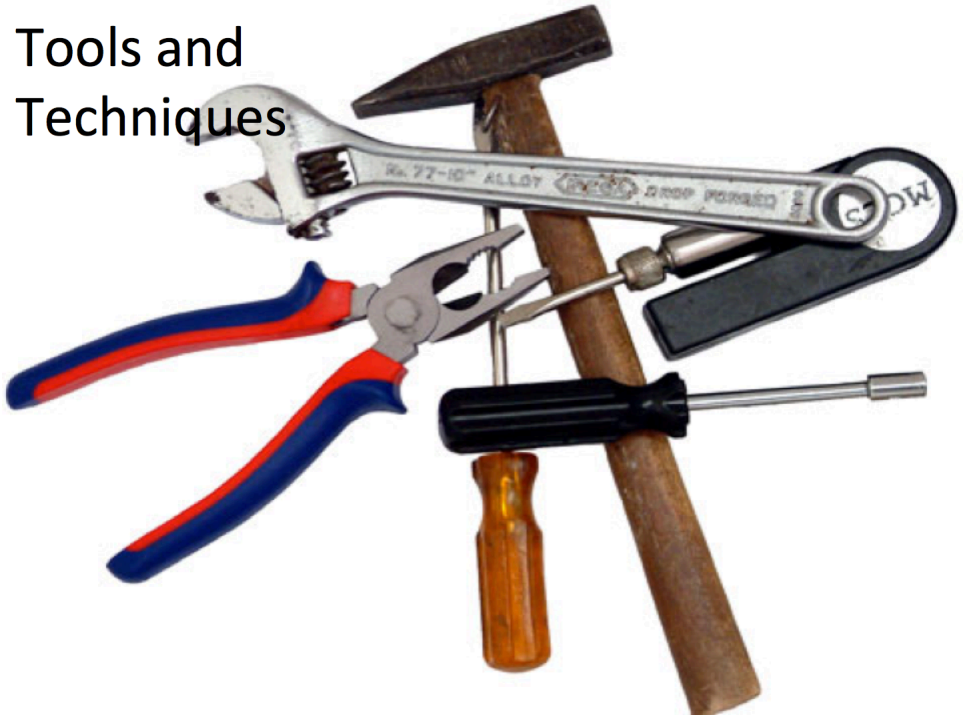

SUMMARY

Alexander Serebrenik
Eindhoven University of Technology, NL

Human Aspects of Software Engineering



Tools and Techniques



SANER 2015
22nd International Conference on Software Analysis, Evolution and Reengineering

- Abstracts: November 7, 2015
- Papers: November 14, 2015