Promises and perils of identity merging

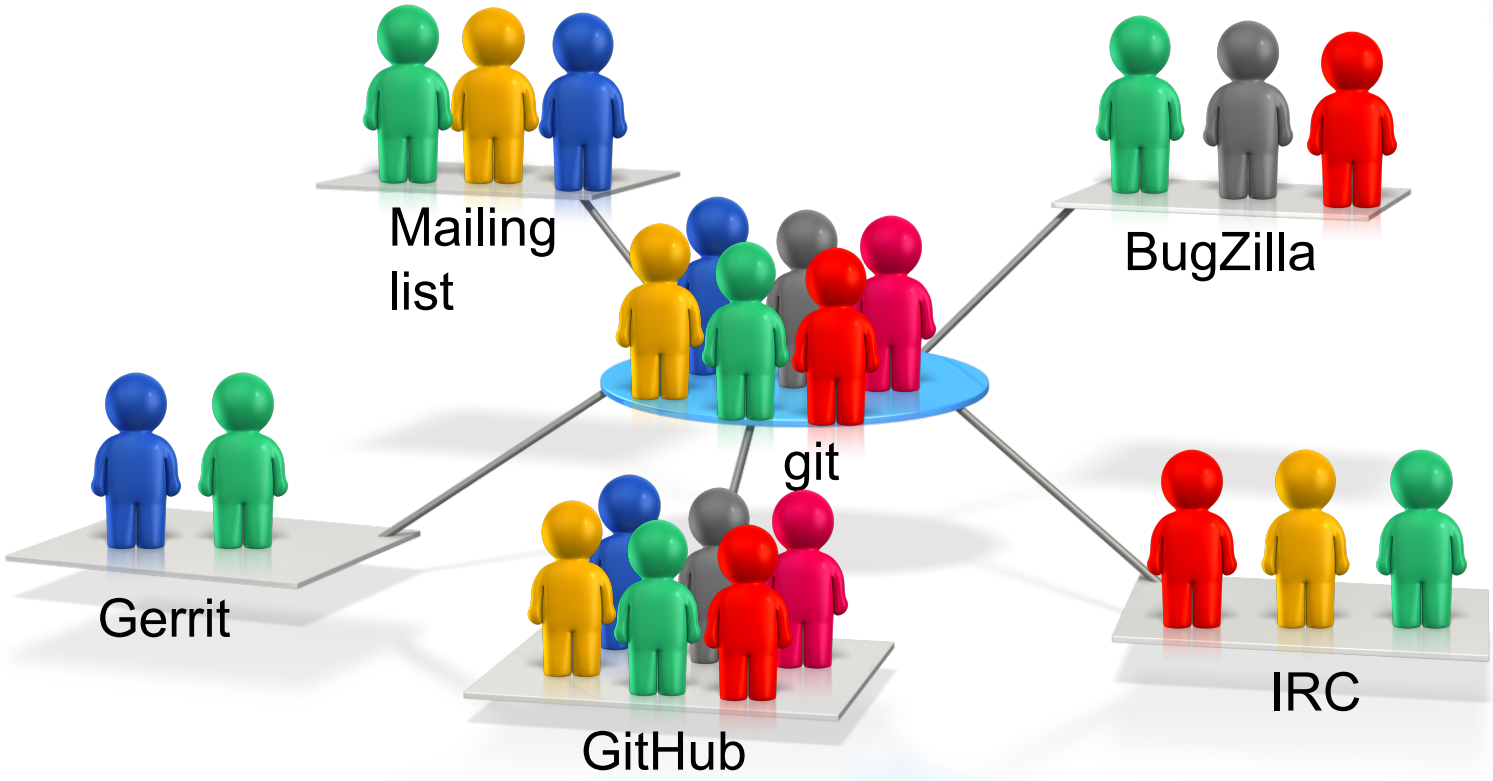
Eleni Constantinou

University of Mons

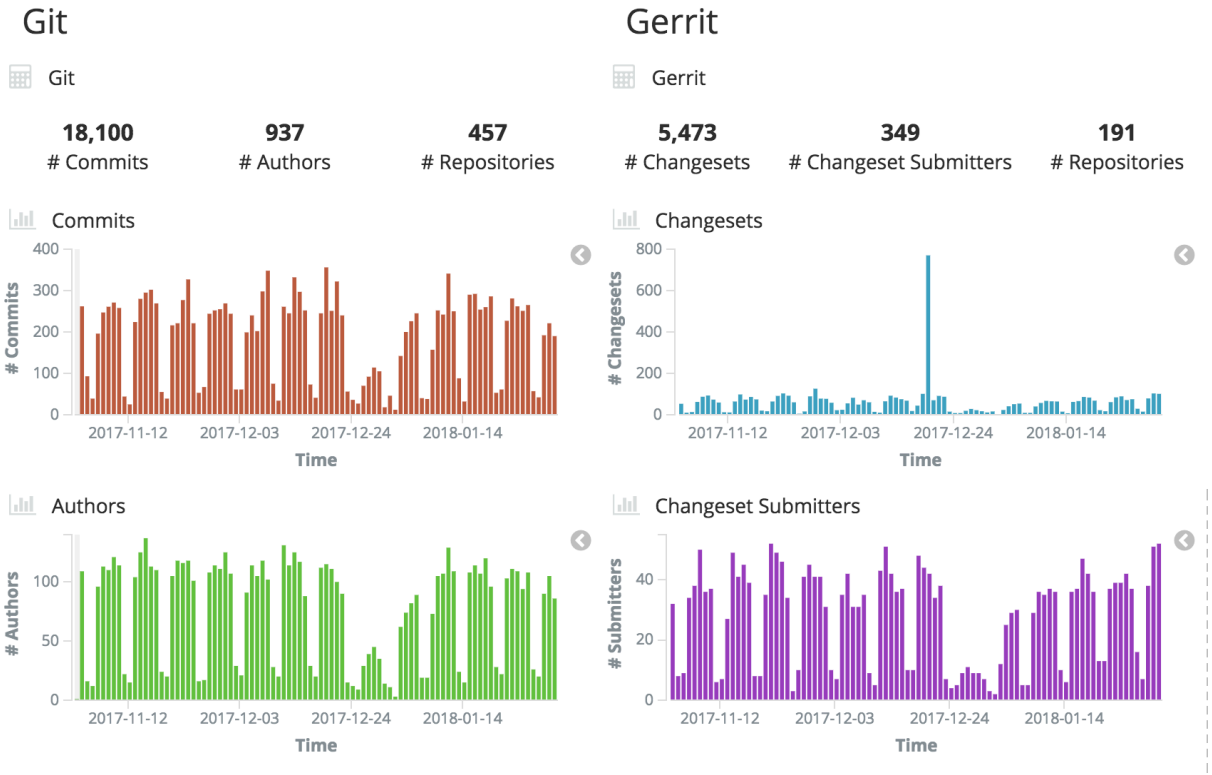
CHAOSS

chaoss.community

Collaboration



Software analytics



CHAOSS

chaoss.community

Software analytics

Bitergia analytics

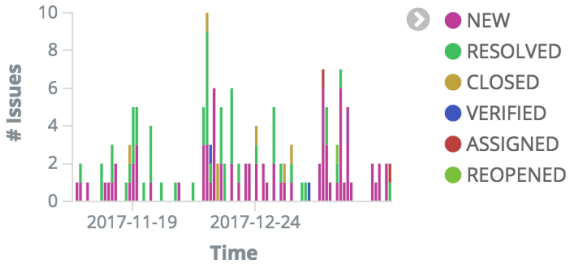
Author	Commits	Projects	Added Lines	Removed Lines	Avg. Files
	372	4	164707	150748	13.898

Bugzilla

Issues

144	2	21
# Issues	# Submitters	# Products

Issues

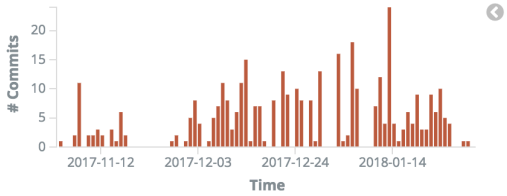


Git

Git

373	1	4
# Commits	# Authors	# Repositories

Commits



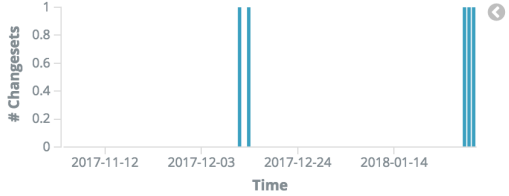
Authors

Gerrit

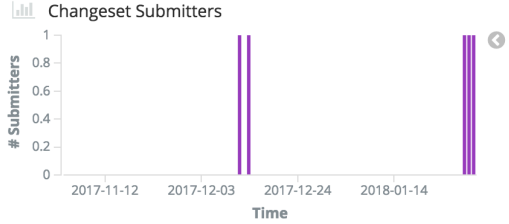
Gerrit

5	1	1
# Changesets	# Changeset Submitters	# Repositories

Changesets



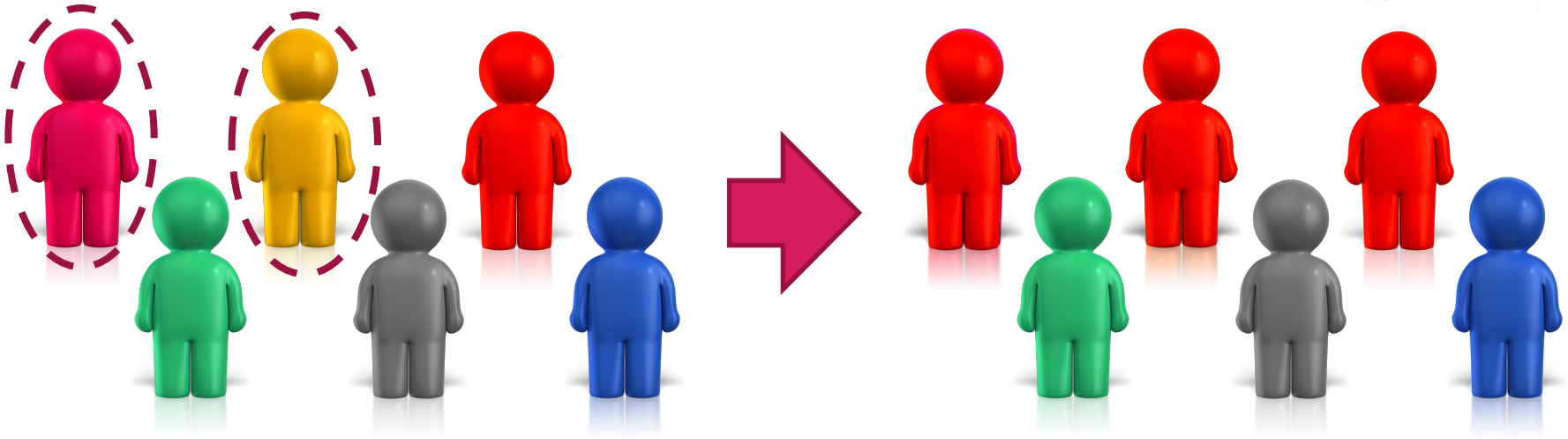
Changeset Submitters



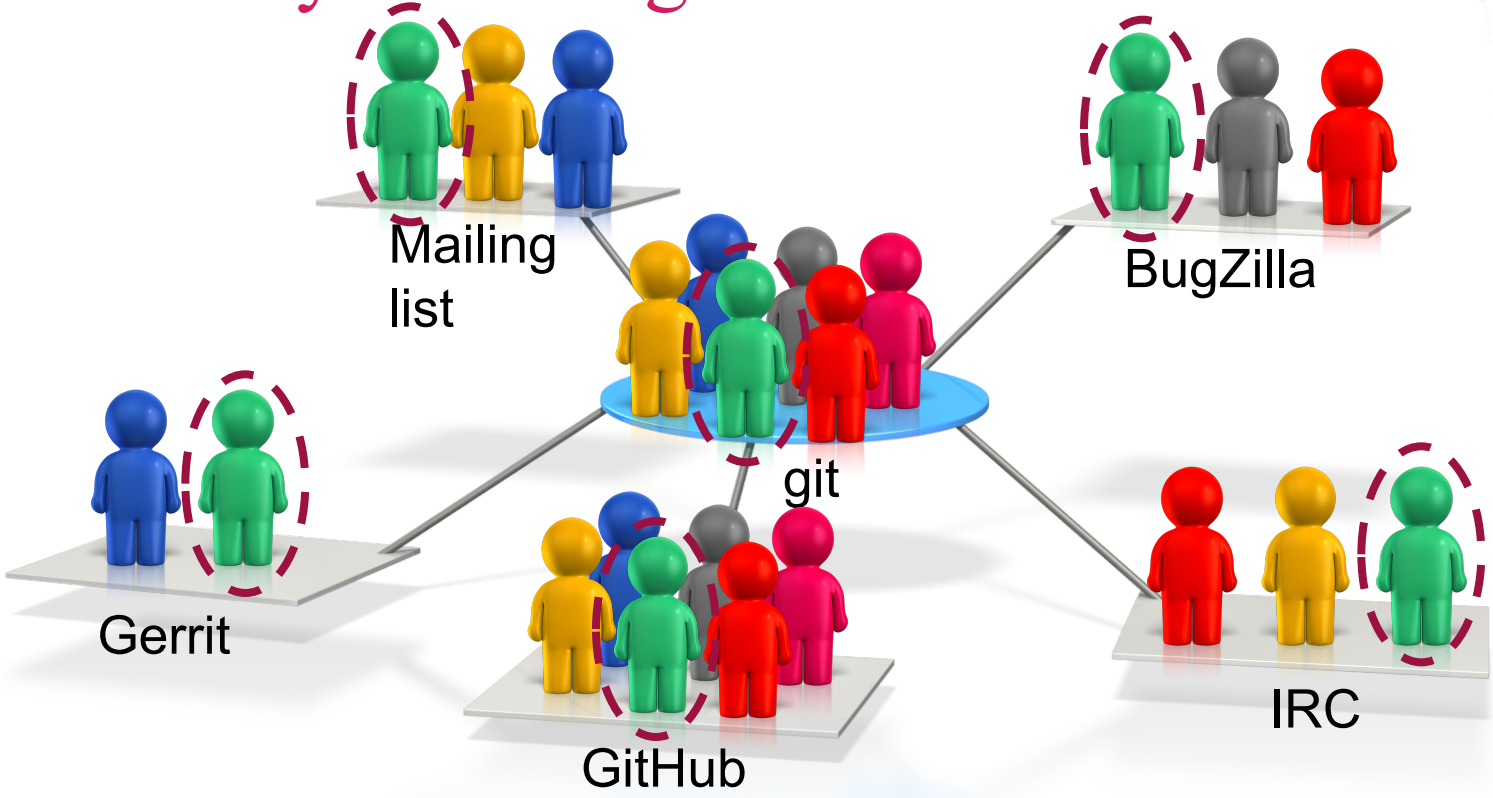
CHAOSS

chaoss.community

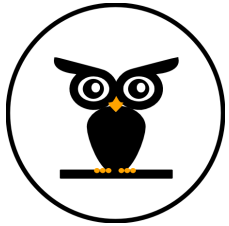
Identity merging



Identity matching



SortingHat



id	name	email	username	source	uuid
0cac4ef	Quan Zhou	quan@bitergia.com	NULL	git	0cac4ef
0ef1c4a	Jesus M. Gonzalez-Barahona	jgbarah@gmail.com	NULL	git	0ef1c4a
11cc034	quan	zhquan7@gmail.com	NULL	git	11cc034
35c0421	Alberto Martín	alberto.martin@bitergia.com	NULL	git	35c0421
37a8187	Alberto Martín	albertinisg@users.noreply.github.com	NULL	git	37a8187
3ca4e85	Daniel Izquierdo Cortazar	dicortazar@gmail.com	NULL	git	3ca4e85
4fcec5a	dpose	dpose@sega.bitergia.net	NULL	git	4fcec5a
5b358fc	dpose	dpose@bitergia.com	NULL	git	5b358fc
692ad15	Andre Klapper	a9016009@gmx.de	NULL	git	692ad15
6dcf98c	Daniel Izquierdo	dizquierdo@bitergia.com	NULL	git	6dcf98c
75fc28e	Santiago Dueñas	sduenas@bitergia.com	NULL	git	75fc28e
7ad0031	Alvaro del Castillo	acs@thelma.cloud	NULL	git	7ad0031
8fac15f	alpgarcia	alpgarcia@gmail.com	NULL	git	8fac15f
9aed245	Alvaro del Castillo	acs@bitergia.com	NULL	git	9aed245

Identity merging

Username:
econst
Full name:
Eleni Constantinou
Email address:
eleni.constantinou@umons.ac.be

GitHub

Username:
econstan
Full name:
Eleni Constantinidou
Email address:
el.const@mydomain.com

GitHub

Identity merging

Username:
econst
Full name:
Eleni Constantinou
Email address:
eleni.constantinou@umons.ac.be

GitHub

Username:
econstan
Full name:
Eleni Constantinidou
Email address:
el.const@mydomain.com

GitHub

Individuals use labels that do not reflect their identity

sigma-lambda-stop

titib

Identity labels might need pre-processing

jon.doe@users.noreply.github.com

j.doe@apple.com

jon.doe@gmail.com

jon.doe@google.com

Some identities do not correspond to humans

CHAOSS

chaoss.community

Individuals use common labels

root

unknown

admin

root@localhost.localdomain

Individuals use very similar labels for their identities

j.wright@apple.com

j.wright@gmail.com

j.wright@amazon.com

CHAOSS

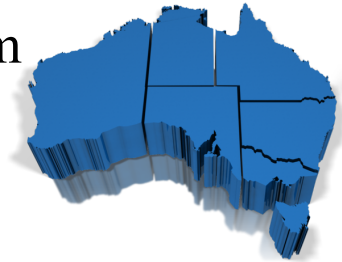
chaoss.community

Individuals use very similar labels for their identities

j.wright@apple.com

j.wright@gmail.com

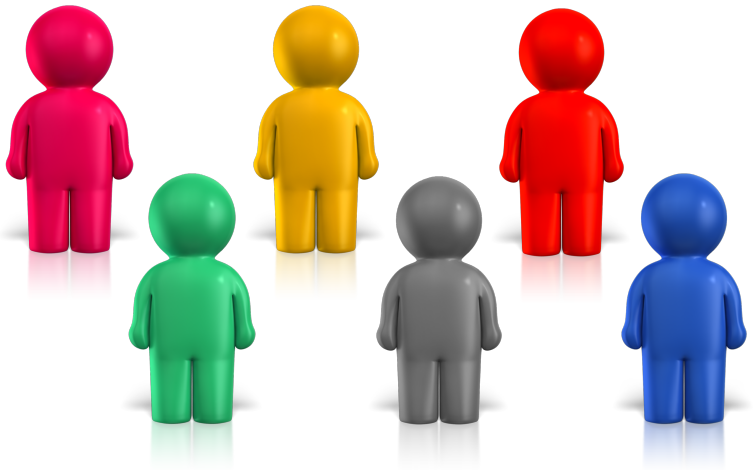
j.wright@amazon.com



CHAOSS

chaoss.community

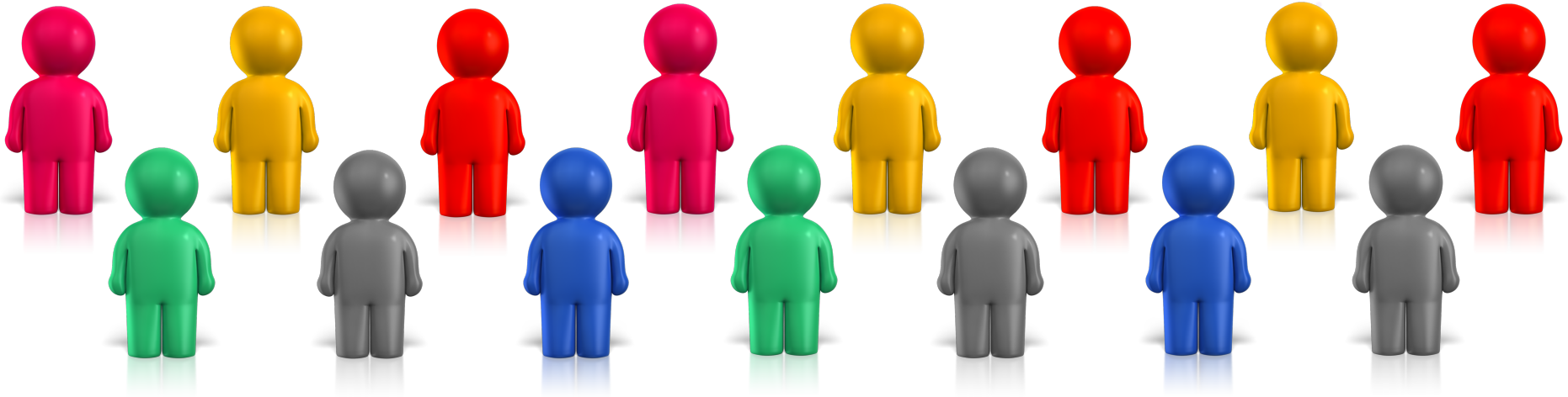
Identity merging accuracy depends on the examined set of identities



CHAOSS

chaoss.community

Identity merging accuracy depends on the examined set of identities



CHAOSS

chaoss.community

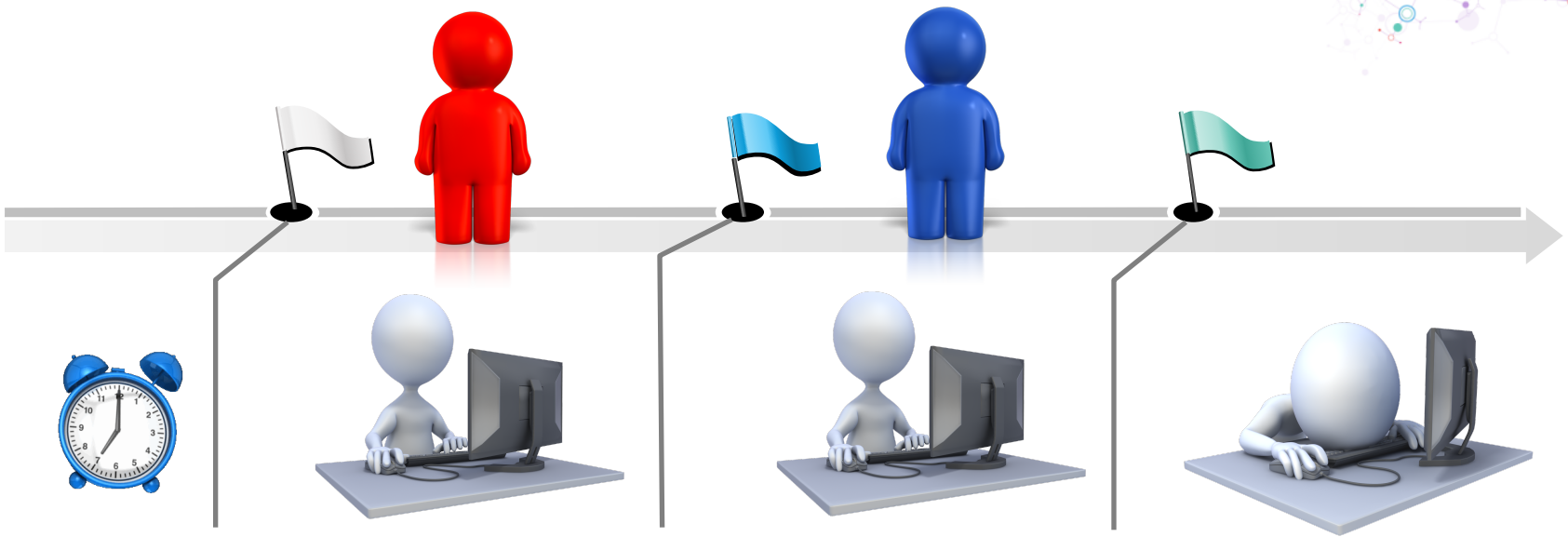
Different platforms provide different identity information

GitHub	Mailing list	StackOverflow
Username: jdoe	**Username:**	**Username:** Jon Doe
Full name: Jon Doe	**Full name:** Jon Doe	**Full name:**
Email address: jon.doe@gmail.com	**Email address:** jon.doe@gmail.com	**Email address:**

Different platforms provide different identity information

GitHub	Mailing list	StackOverflow
Username: jdoe	**Username:**	**Username:** jon
Full name: Jon Doe	**Full name:** Jon Doe	**Full name:**
Email address: jon.doe@gmail.com	**Email address:** jon.doe@gmail.com	**Email address:**

Individuals use multiple identities in a different capacity



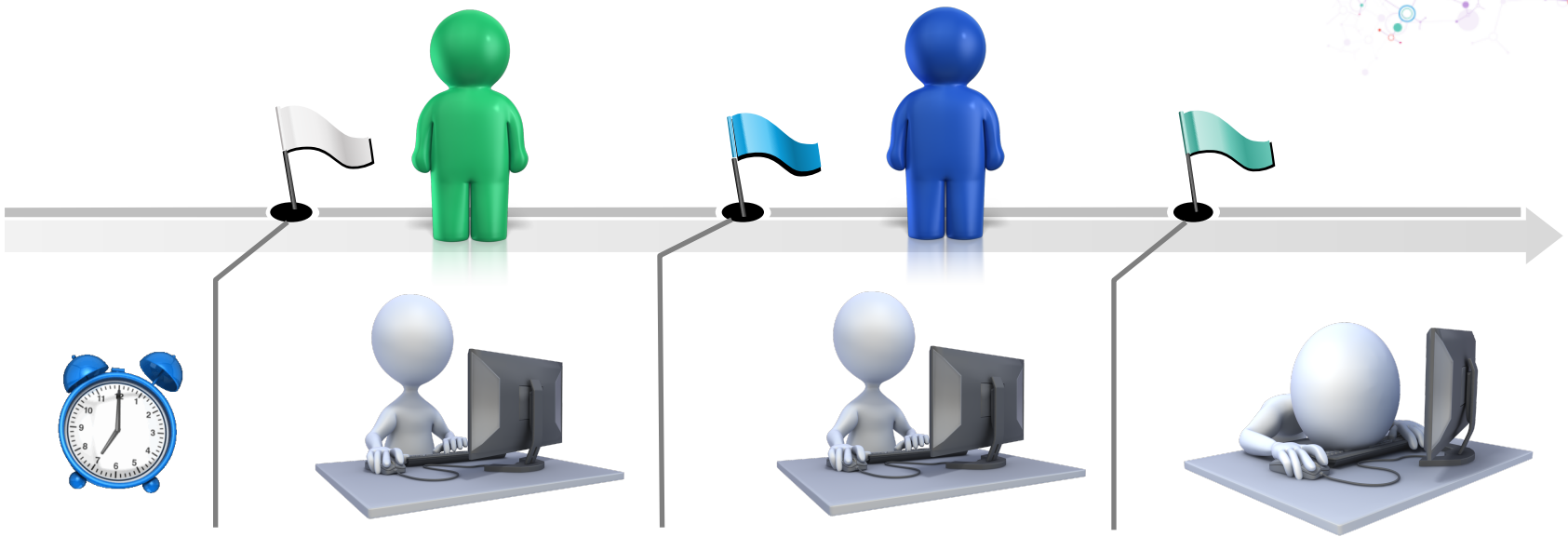
j.doe@apple.com

j.doe@gmail.com

CHAOSS

chaoss.community

Individuals use multiple identities in a different capacity



j.doe@amazon.com

j.doe@gmail.com

CHAOSS

chaoss.community

The label set associated with identities is incomplete

Username:
econst
Full name:
Eleni Constantinou
Email address:
eleni.c@domain1.com

GitHub

Username:
elconst
Full name:

Email address:
eleni.c@domain2.com

GitHub

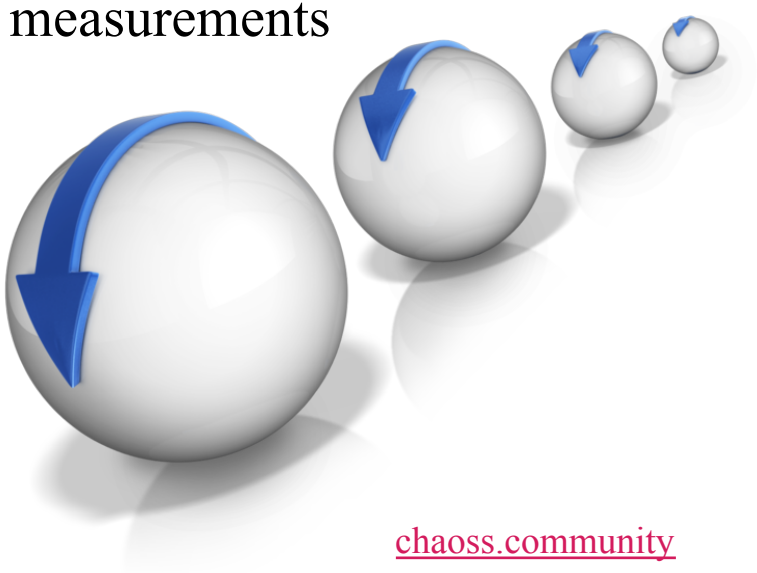
Impact of erroneous merges

Underestimate individuals' activity

Overestimate community joining/leaving measurements

Snowball effect

Cost of fixing erroneous merges



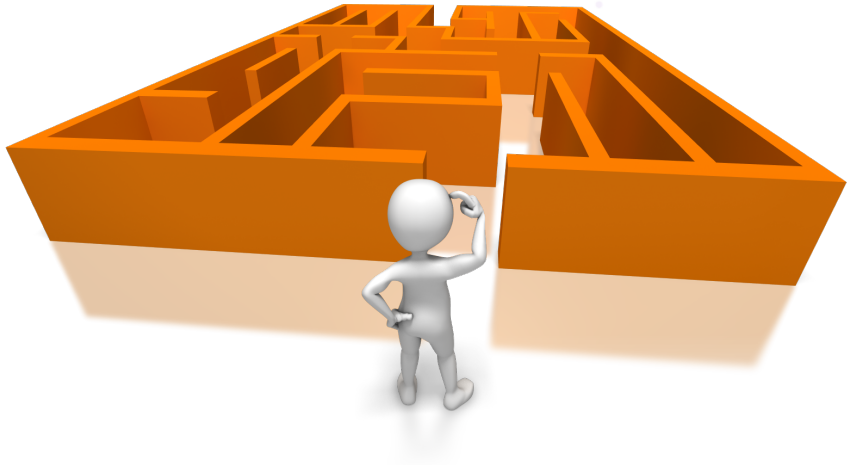
Conclusion

Merging identities is hard

Automation should be limited

Manual inspection

Consider additional information (when available)



Thank you!

CHAOSS

chaoss.community