Octocat, here and elsewhere, by GitHub https://octodex.github.com

Lessons in Social Coding

Software Analytics in the Age of GitHub

Bogdan Vasilescu

ISR, School of Computer Science
Carnegie Mellon University

@b_vasilescu http://bvasiles.github.io

Social Web

+

Software Engineering

Social Software Engineering

THE EVOLUTION OF THE "SOCIAL PROGRAMMER"

github SOCIAL CODING

ashley williams
ashleygwilliams

npm, inc
ridgewood, queens, NYC
http://ashleygwilliams.github.io/
Joined on Oct 31, 2011

776 Followers **38** Starred **15** Following

Organizations

Contributions | Repositories | Public activity

Follow

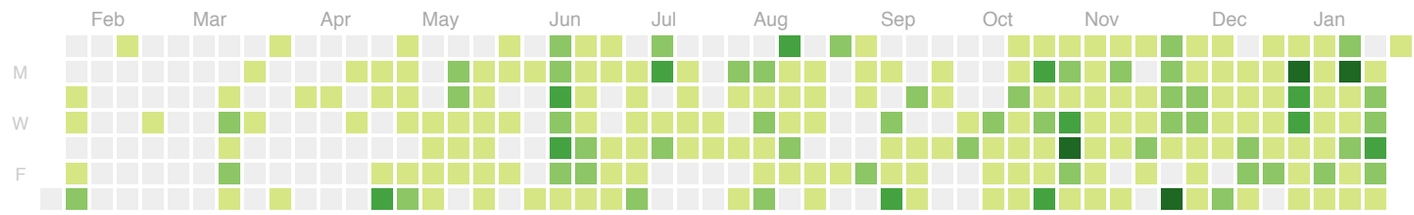
Popular repositories

- **breakfast-repo** — a collection of videos, recordings, and podcast… — 208 ★
- **x86-kernel** — a simple x86 kernel, extended with Rust — 48 ★
- **ashleygwilliams.github.io** — hi, i'm ashley. nice to meet you. — 37 ★
- **jsconf-2015-deck** — deck for jsconf2015 talk, "if you wish to learn e… — 32 ★
- **ratpack** — sinatra boilerplate using activerecord, sqlite, a… — 32 ★

Repositories contributed to

- npm/**docs** — The place where all the npm docs live. — 44 ★
- mozilla/**publish.webmaker.org** — The teach.org publishing service for goggles a… — 2 ★
- npm/**marky-markdown** — npm's markdown parser — 104 ★
- artisan-tattoo/**assistant-frontend** — ember client for assistant-API — 5 ★
- npm/**npm-camp** — a community conference for all things npm — 1 ★

Public contributions



Summary of pull requests, issues opened, and commits. Learn how we count contributions.

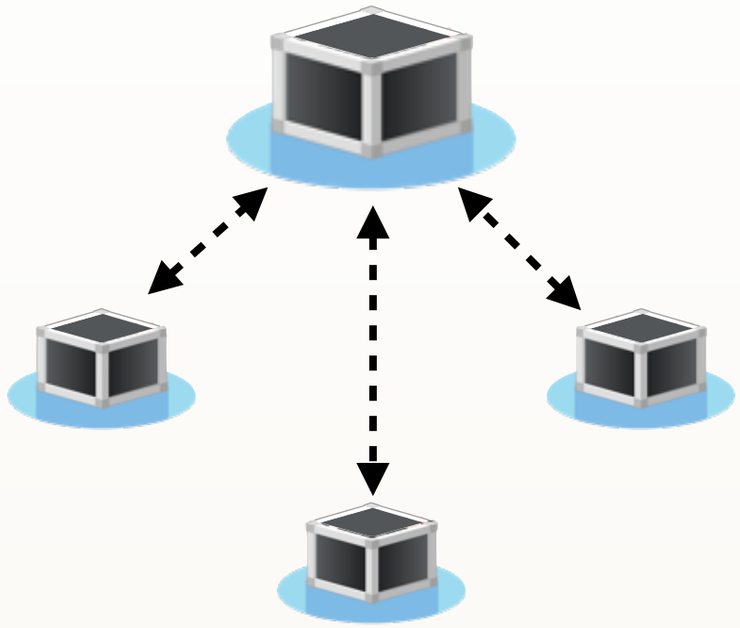
Less More

Contributions in the last year	Longest streak	Current streak
1,886 total	37 days	7 days
Jan 24, 2015 – Jan 24, 2016	October 7 – November 12	January 18 – January 24

https://github.com/ashleygwilliams

- Programming in a socially networked world: the evolution of the social programmer
  C Treude, F Figueira Filho, B Cleary, MA Storey.
  FutureCSD-CSCW 2012
- Social coding in GitHub: transparency and collaboration in an open software repository
  L Dabbish, C Stuart, J Tsay, J Herbsleb.
  CSCW 2012
- Social networking meets software development: Perspectives from GitHub, MSDN, Stack Exchange, and TopCoder
  A Begel, J Bosch, MA Storey.
  IEEE Software 2013

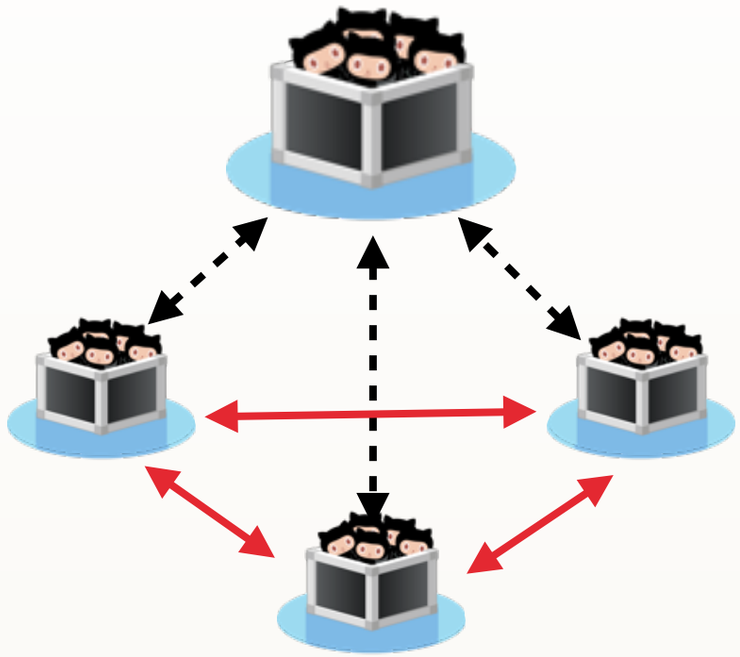
“SOCIAL CODING”: CODE IS MEANT TO BE SHARED



“SOCIAL CODING”: CODE IS MEANT TO BE SHARED

github SOCIAL CODING

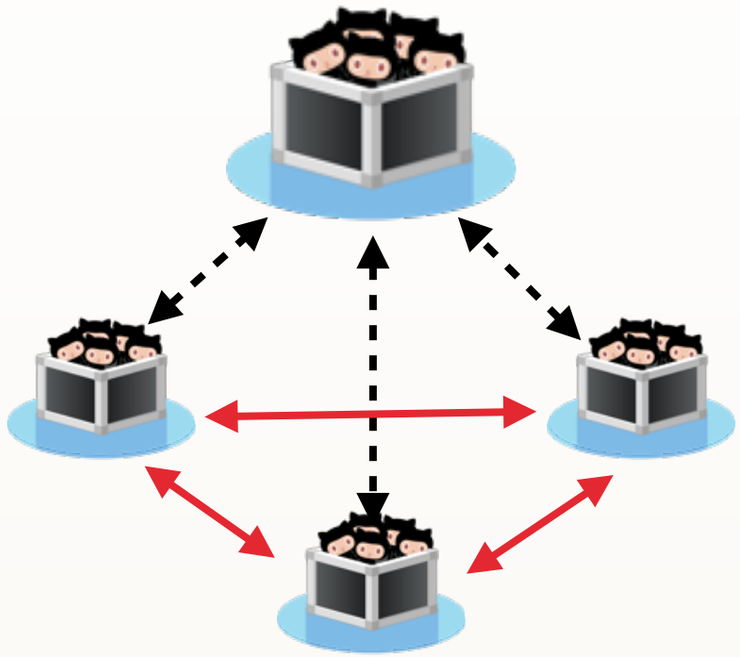
GIT



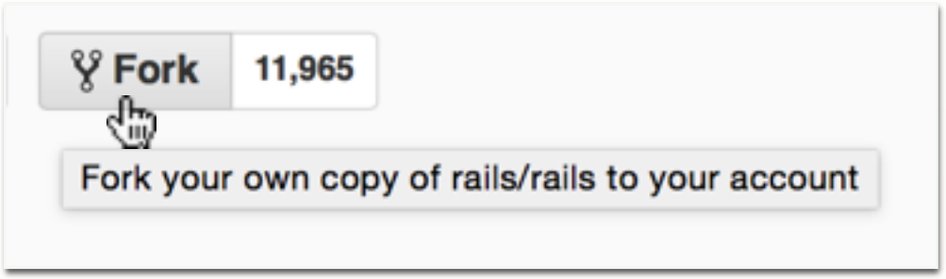
“SOCIAL CODING”: CODE IS MEANT TO BE SHARED

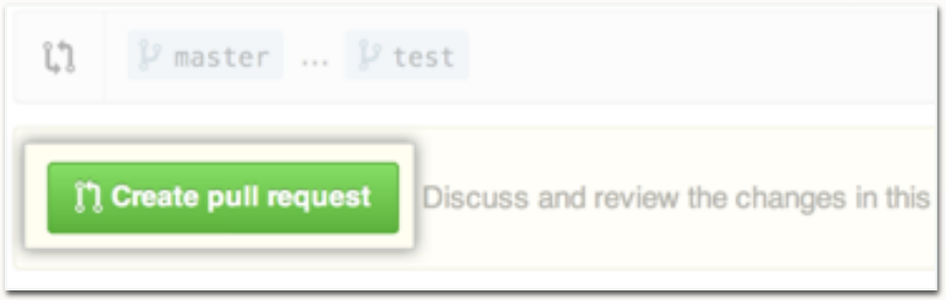
github SOCIAL CODING

GIT



GITHUB UI

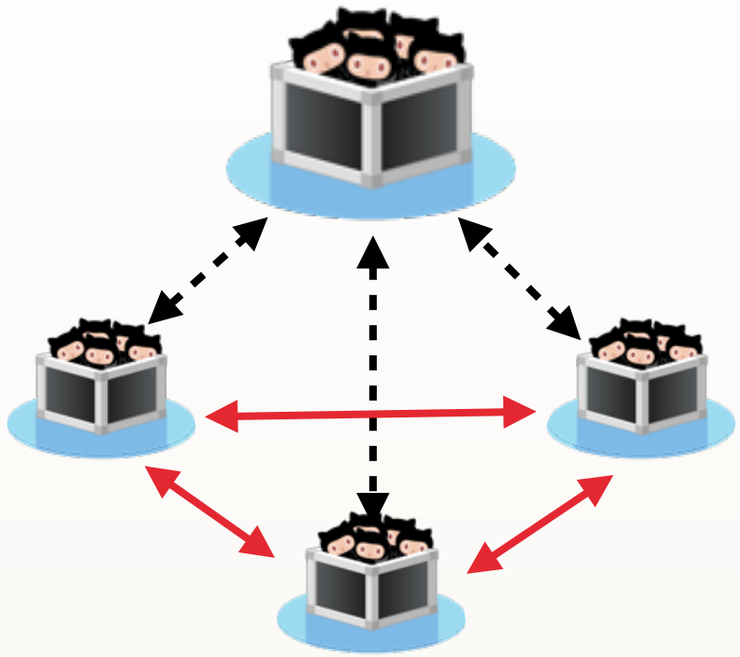




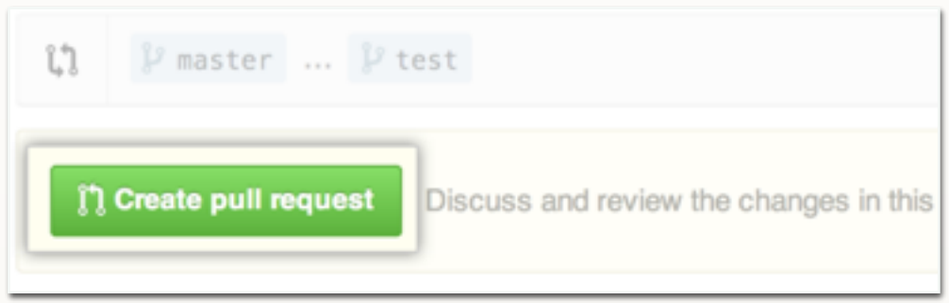
“SOCIAL CODING”: CODE IS MEANT TO BE SHARED

github SOCIAL CODING

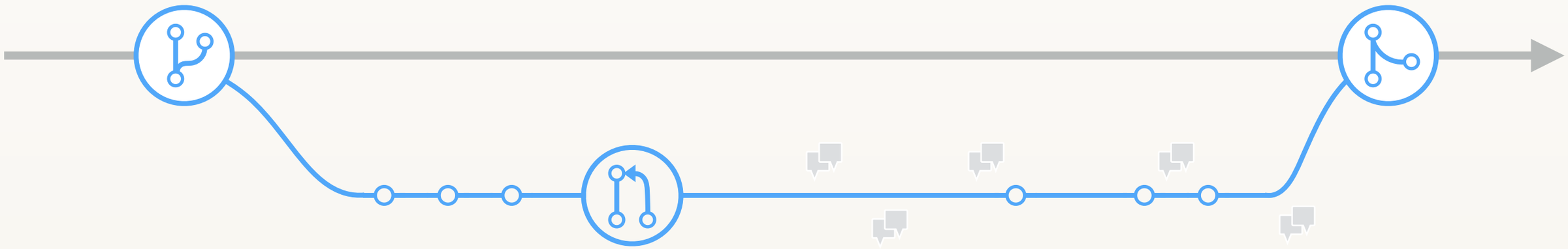
GIT



GITHUB UI



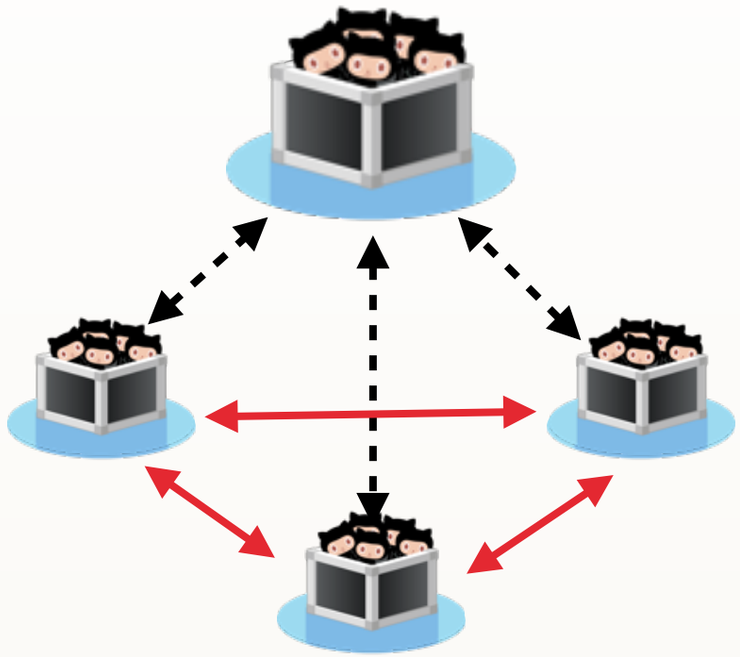
THE “PULL REQUEST” MODEL



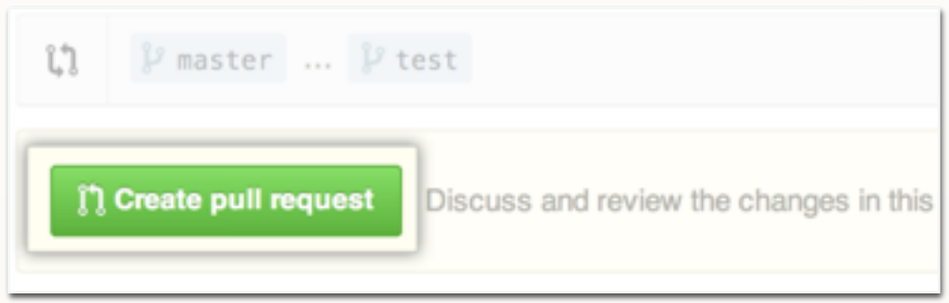
“SOCIAL CODING”: CODE IS MEANT TO BE SHARED

github SOCIAL CODING

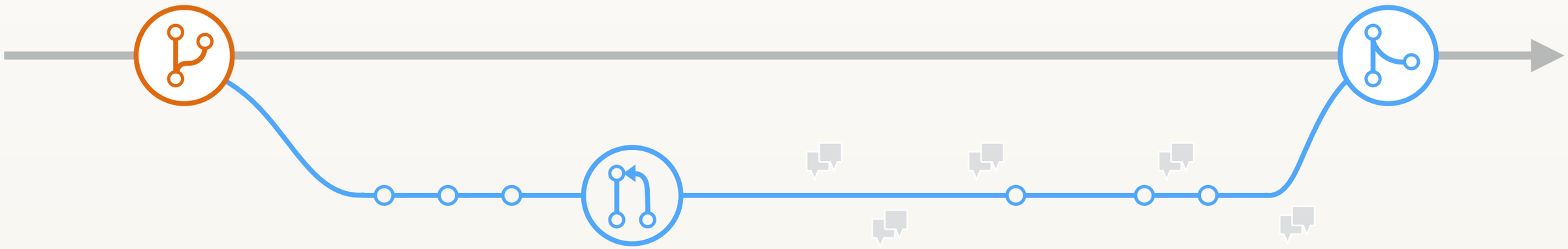
GIT



GITHUB UI



THE “PULL REQUEST” MODEL

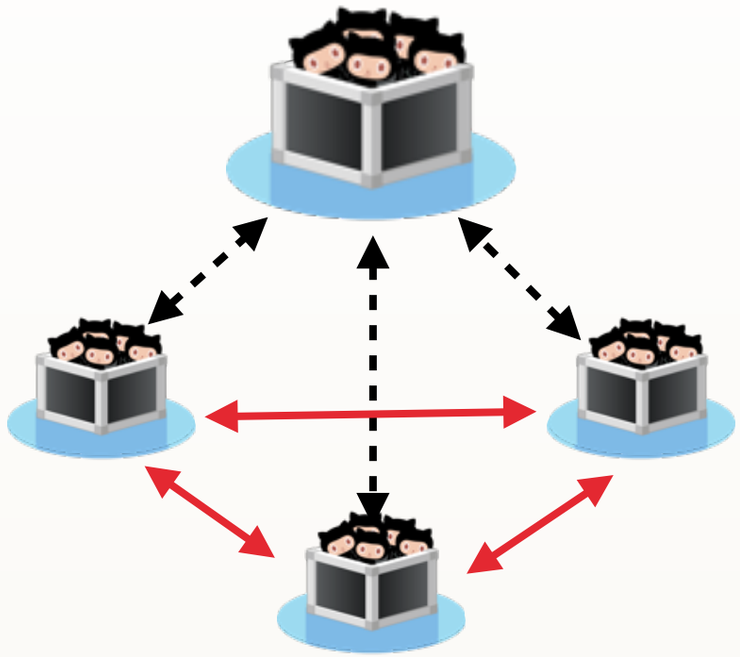


Create
a branch

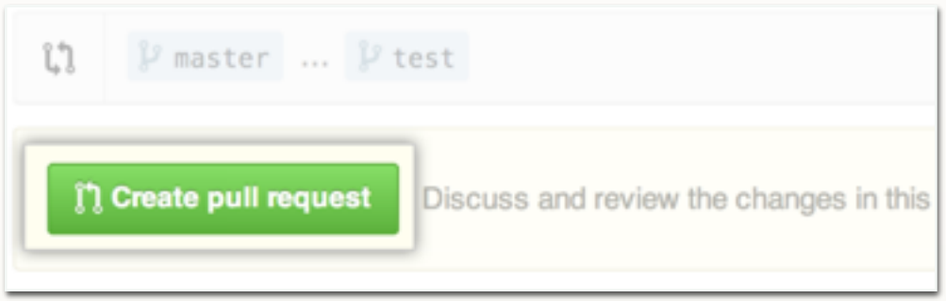
“SOCIAL CODING”: CODE IS MEANT TO BE SHARED

github SOCIAL CODING

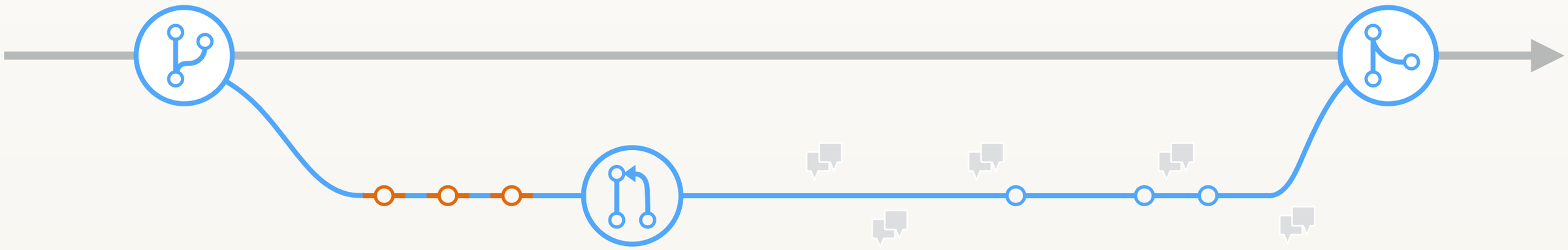
GIT



GITHUB UI



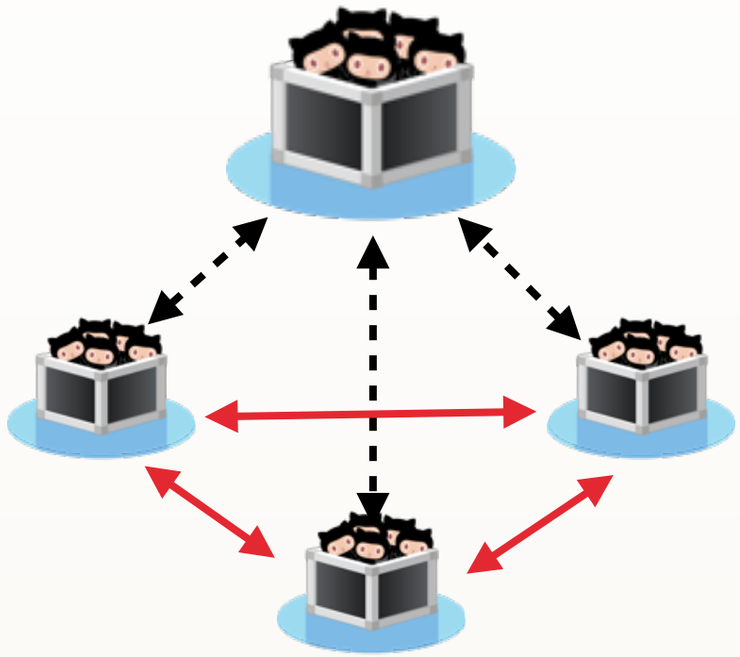
THE “PULL REQUEST” MODEL



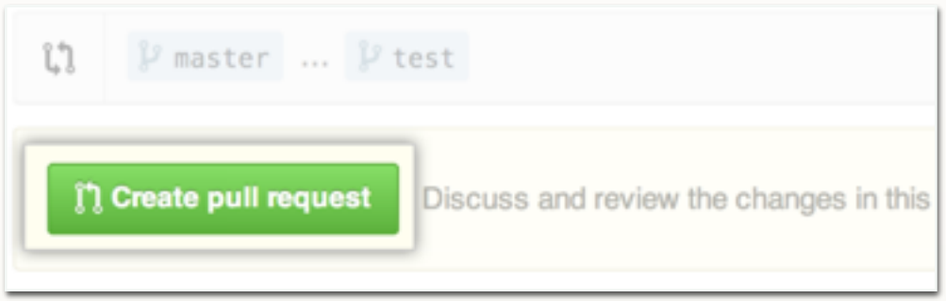
“SOCIAL CODING”: CODE IS MEANT TO BE SHARED

github SOCIAL CODING

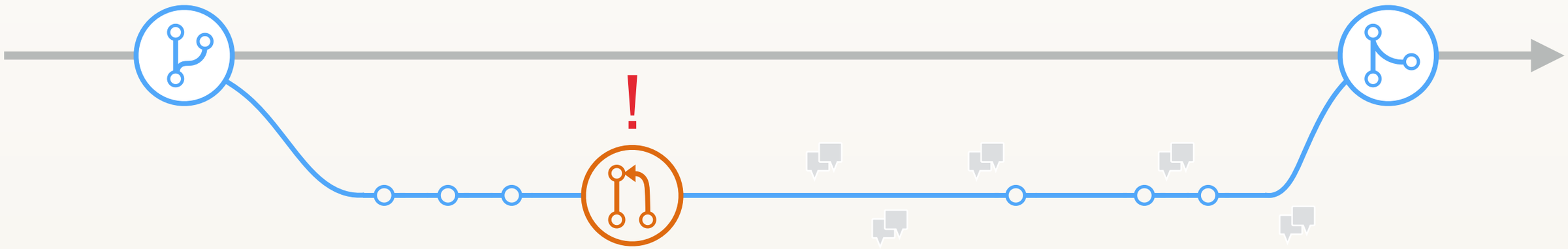
GIT



GITHUB UI



THE “PULL REQUEST” MODEL

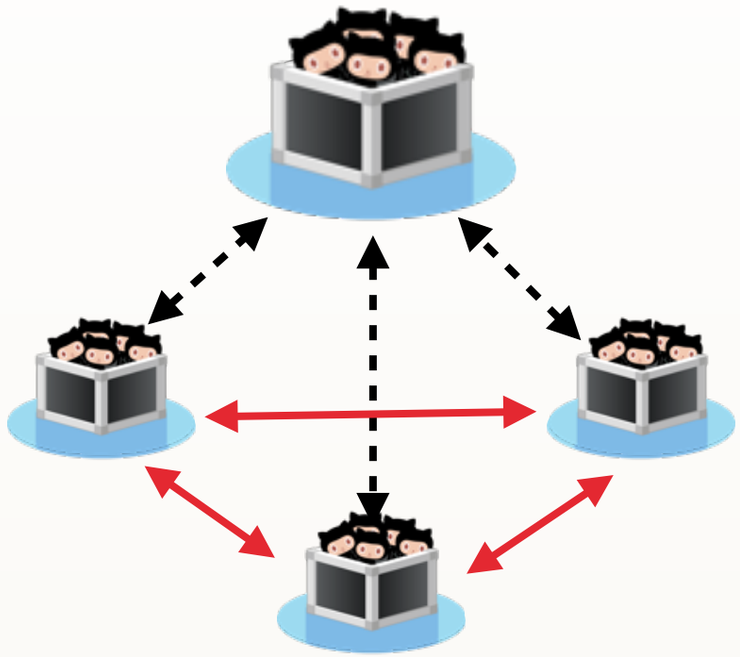


Open a
pull request

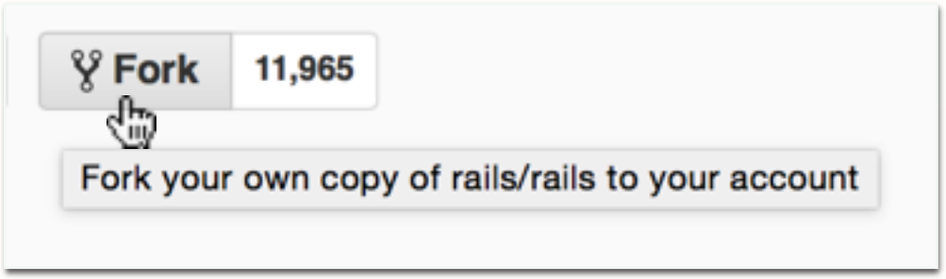
“SOCIAL CODING”: CODE IS MEANT TO BE SHARED

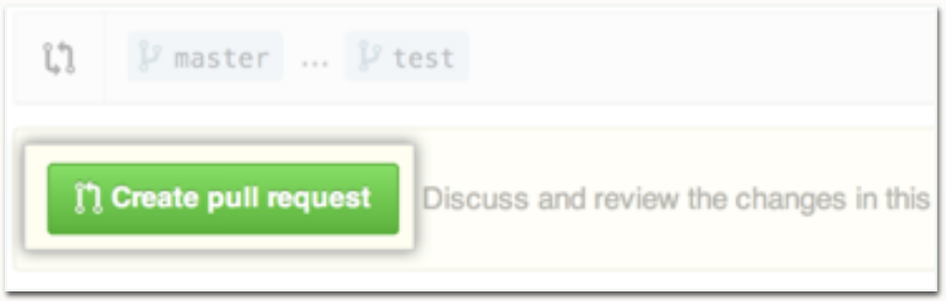
github SOCIAL CODING

GIT

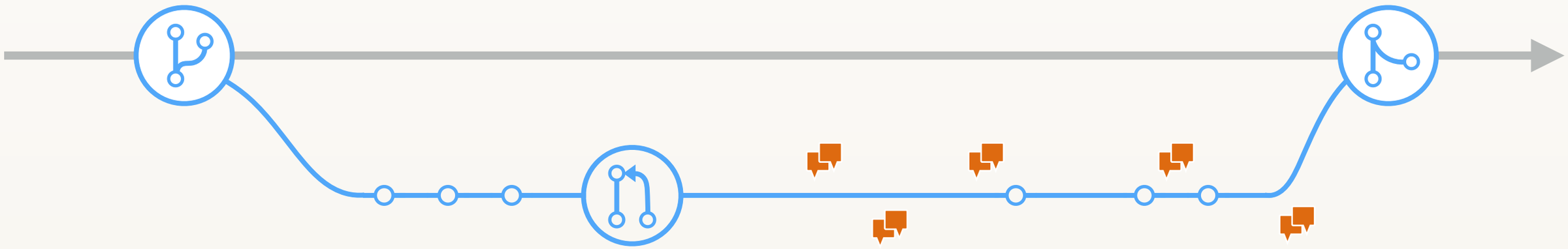


GITHUB UI





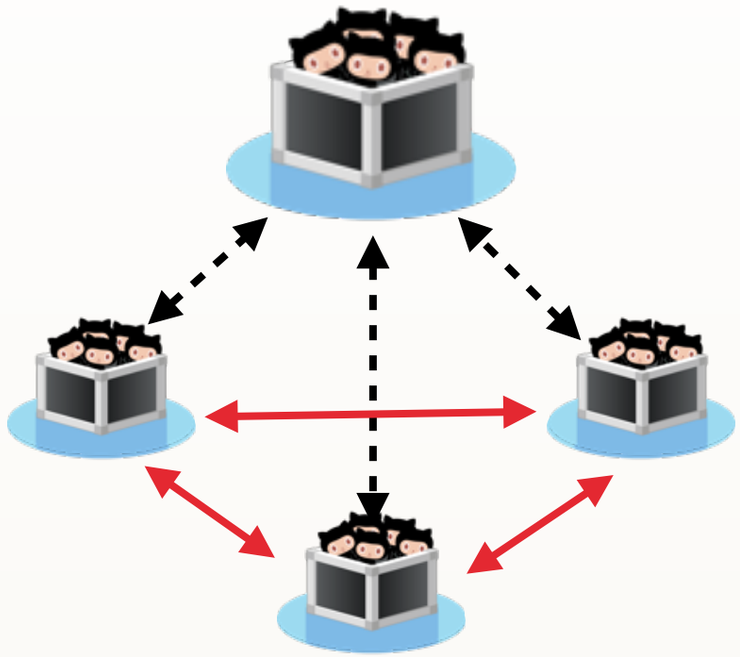
THE “PULL REQUEST” MODEL



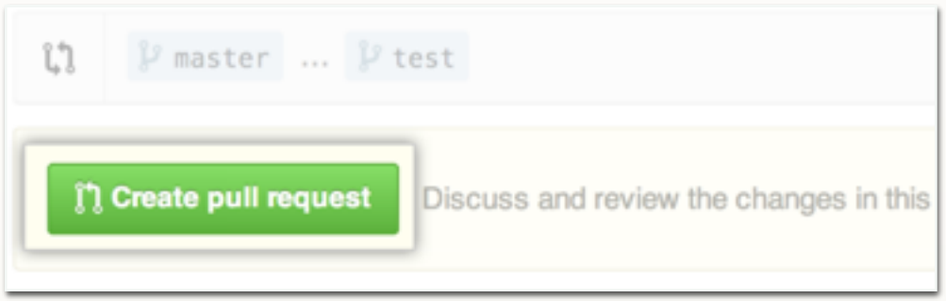
“SOCIAL CODING”: CODE IS MEANT TO BE SHARED

github SOCIAL CODING

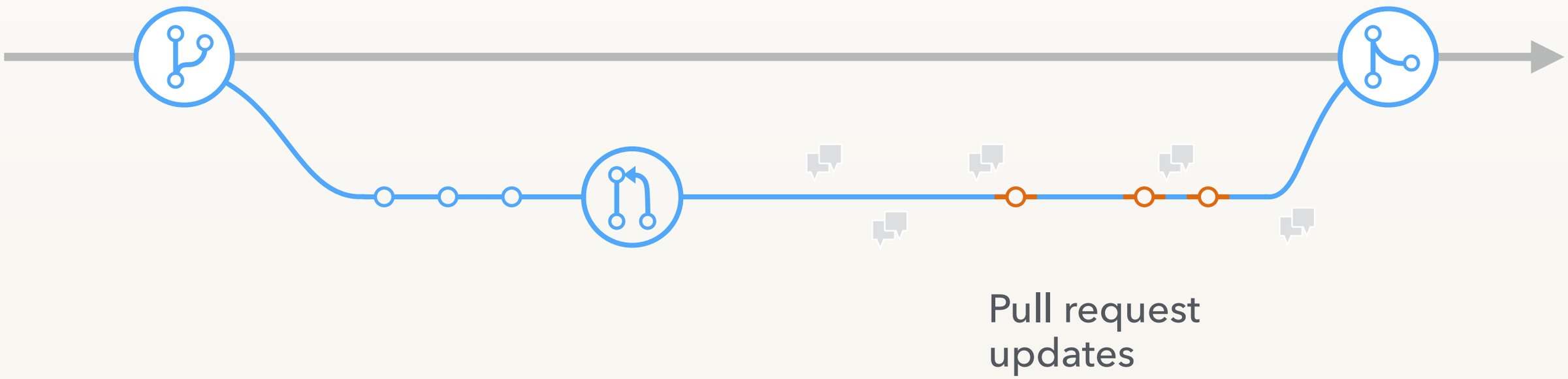
GIT



GITHUB UI



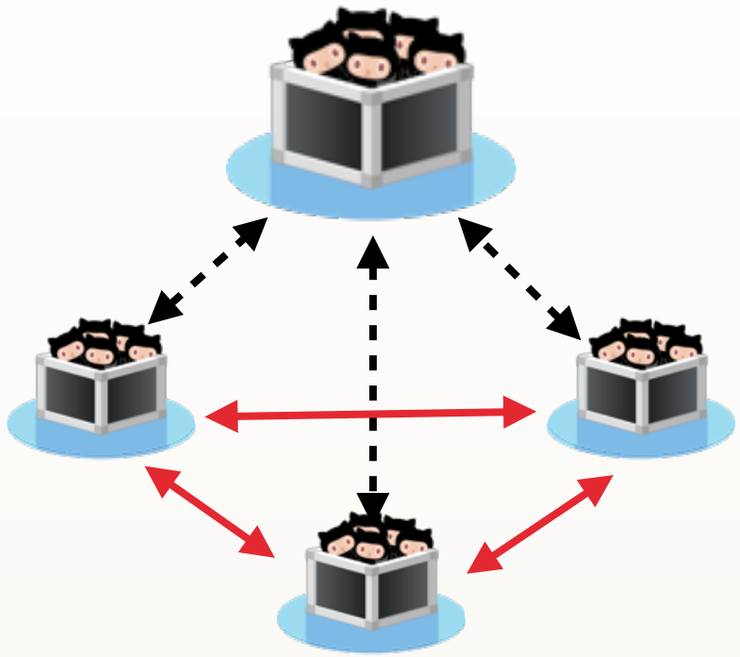
THE “PULL REQUEST” MODEL



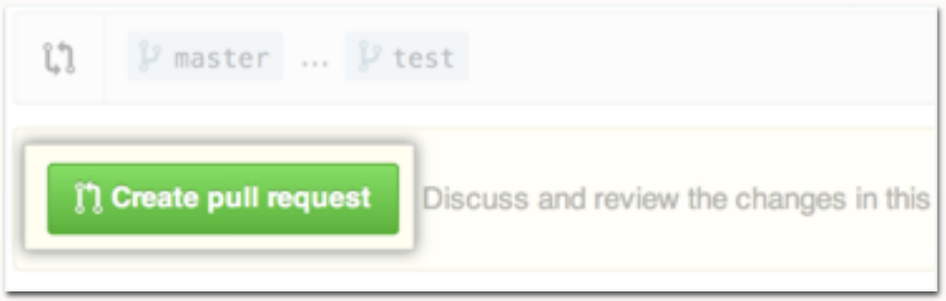
“SOCIAL CODING”: CODE IS MEANT TO BE SHARED

github SOCIAL CODING

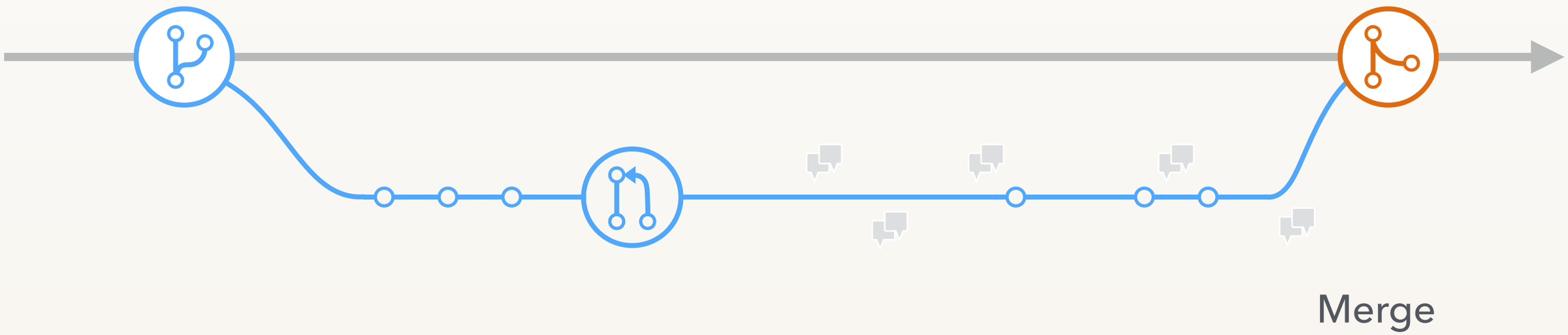
GIT



GITHUB UI



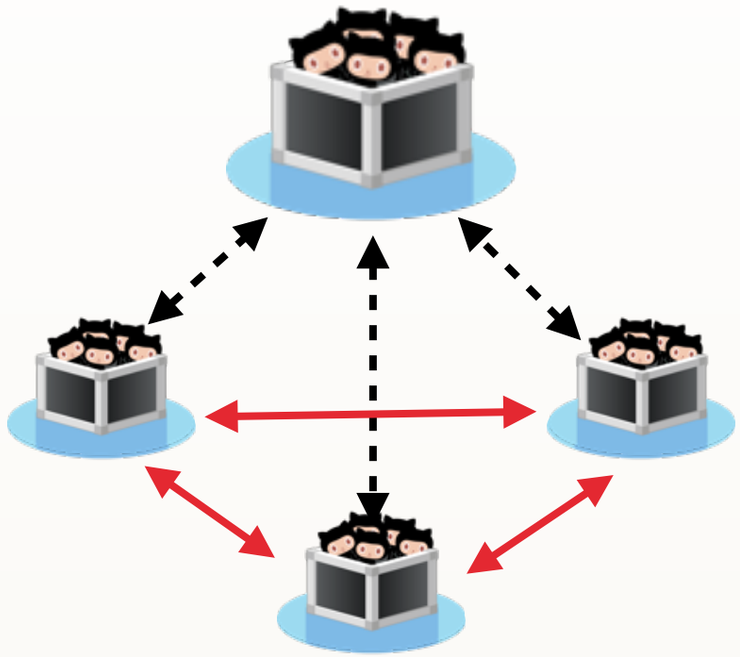
THE “PULL REQUEST” MODEL



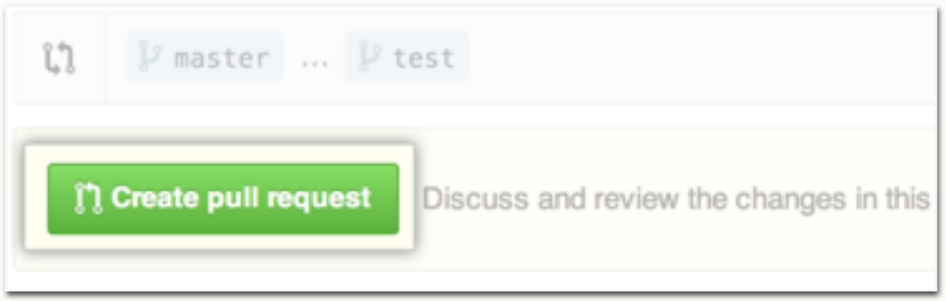
“SOCIAL CODING”: CODE IS MEANT TO BE SHARED

github SOCIAL CODING

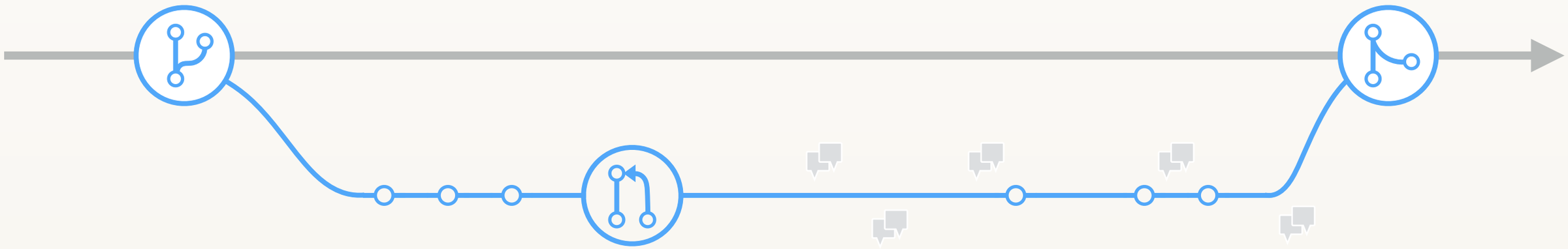
GIT



GITHUB UI



THE “PULL REQUEST” MODEL



Unified development, testing, code review, integration → DEVOPS

Lowest ever barrier to entry for newcomers

Democratic, open, social process

SOFTWARE DEVELOPMENT IS CHANGING

OPEN-SOURCE IS GROWING

Companies:

- 78% run OSS
- 66% build on top of OSS

BlackDuck 2015 survey https://goo.gl/Ltaqqs

SOFTWARE DEVELOPMENT IS CHANGING

OPEN-SOURCE IS GROWING

Companies:

- 78% run OSS
- 66% build on top of OSS

BlackDuck 2015 survey https://goo.gl/Ltaqqs

SOCIAL CODING IS GROWING

12 million people

31 million repositories

- GitHub stats from: https://github.com/about

SOFTWARE DEVELOPMENT IS CHANGING

BlackDuck 2015 survey https://goo.gl/Ltaqqs

OPEN-SOURCE IS GROWING

Companies:

- 78% run OSS
- 66% build on top of OSS

SOCIAL CODING IS GROWING

12 million people

31 million repositories

18.5 million software dev's

- GitHub stats from: https://github.com/about
- World estimates from: http://goo.gl/Htnni9

SOFTWARE DEVELOPMENT IS CHANGING

BlackDuck 2015 survey https://goo.gl/Ltaqqs

OPEN-SOURCE IS GROWING

open source™

Companies:

- 78% run OSS
- 66% build on top of OSS

SOCIAL CODING IS GROWING

12 million people

31 million repositories

18.5 million software dev's

15,000+ people

- GitHub stats from: https://github.com/about
- World estimates from: http://goo.gl/Htnni9

SOFTWARE DEVELOPMENT IS CHANGING

OPEN-SOURCE IS GROWING

Companies:

- 78% run OSS
- 66% build on top of OSS

SOCIAL CODING IS GROWING

12 million people

31 million repositories

18.5 million software dev's

15,000+ people

CULTURE CHANGE

"it's just so uncool not sharing the code in the age of social coding"

- GitHub stats from: https://github.com/about
- World estimates from: http://goo.gl/Htnni9

SOFTWARE DEVELOPMENT IS CHANGING

OPEN-SOURCE IS GROWING

Companies:

- 78% run OSS
- 66% build on top of OSS

SOCIAL CODING IS GROWING

12 million people

31 million repositories

18.5 million software dev's

15,000+ people

CULTURE CHANGE

"it's just so uncool not sharing the code in the age of social coding"

HIRING

- **$100+** /hour:
  - owns popular OSS products;
  - stack**overflow** score > 20K; …
- **$50+** /hour:
  - active OSS contributor;
  - stack**overflow** score > 5K; …

- GitHub stats from: https://github.com/about
- World estimates from: http://goo.gl/Htnni9
- **How Much Do You Cost?** Yegor Bugayenko http://goo.gl/N0mL3F
- Activity traces and signals in software developer recruitment and hiring J Marlow, L Dabbish. *CSCW 2013*

SOFTWARE DEVELOPMENT IS CHANGING

OPEN-SOURCE IS GROWING

Companies:

- 78% run OSS
- 66% build on top of OSS

SOCIAL CODING IS GROWING

12 million people

31 million repositories

18.5 million software dev's

15,000+ people

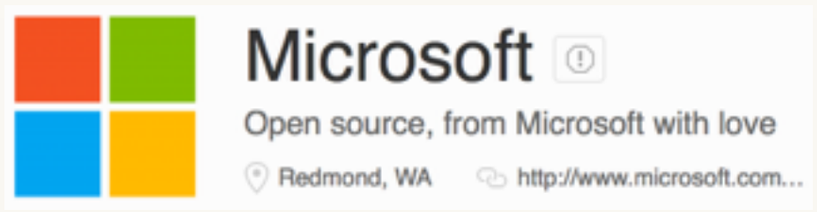
CULTURE CHANGE

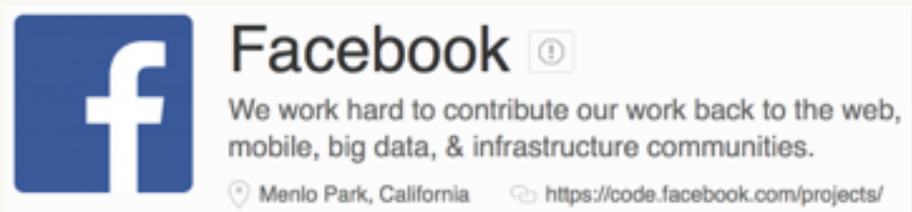
"it's just so uncool not sharing the code in the age of social coding"

HIRING

- **$100+** /hour:
  - owns popular OSS products;
  - stack**overflow** score > 20K; …
- **$50+** /hour:
  - active OSS contributor;
  - stack**overflow** score > 5K; …

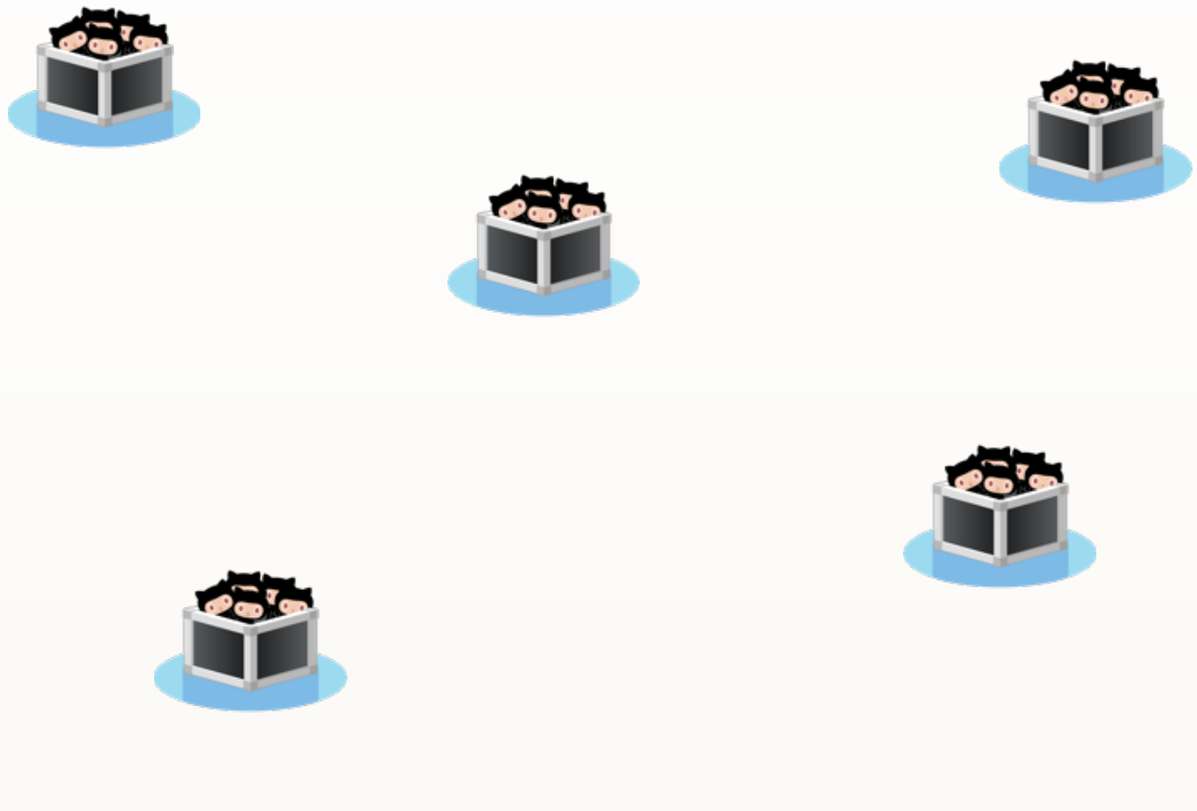
INDUSTRIAL INVOLVEMENT & ADOPTION



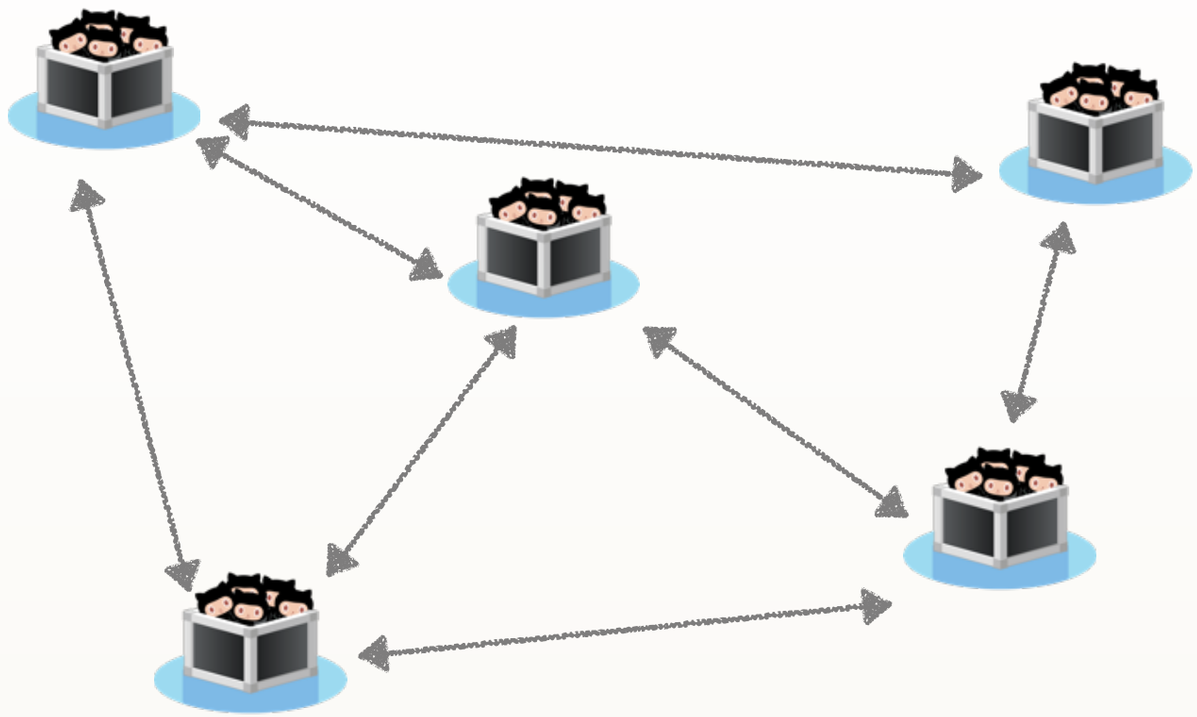


- GitHub stats from: https://github.com/about
- World estimates from: http://goo.gl/Htnni9
- Open source-style collaborative development practices in commercial projects using GitHub E Kalliamvakou, D Damian, K Blincoe, L Singer, DM German. *ICSE 2015*
- How Much Do You Cost? Yegor Bugayenko http://goo.gl/N0mL3F
- Activity traces and signals in software developer recruitment and hiring J Marlow, L Dabbish. *CSCW 2013*

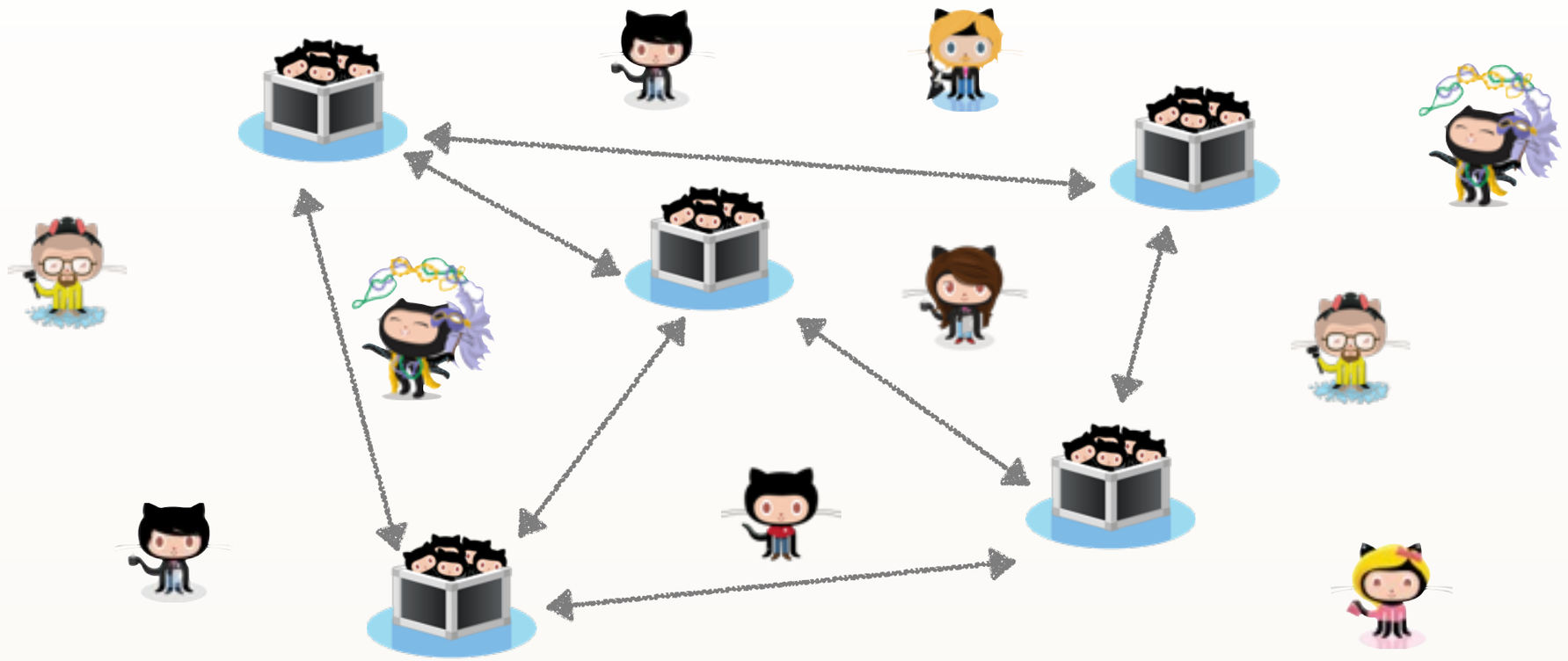
LARGE, DIVERSE, COMPLEX ECOSYSTEM



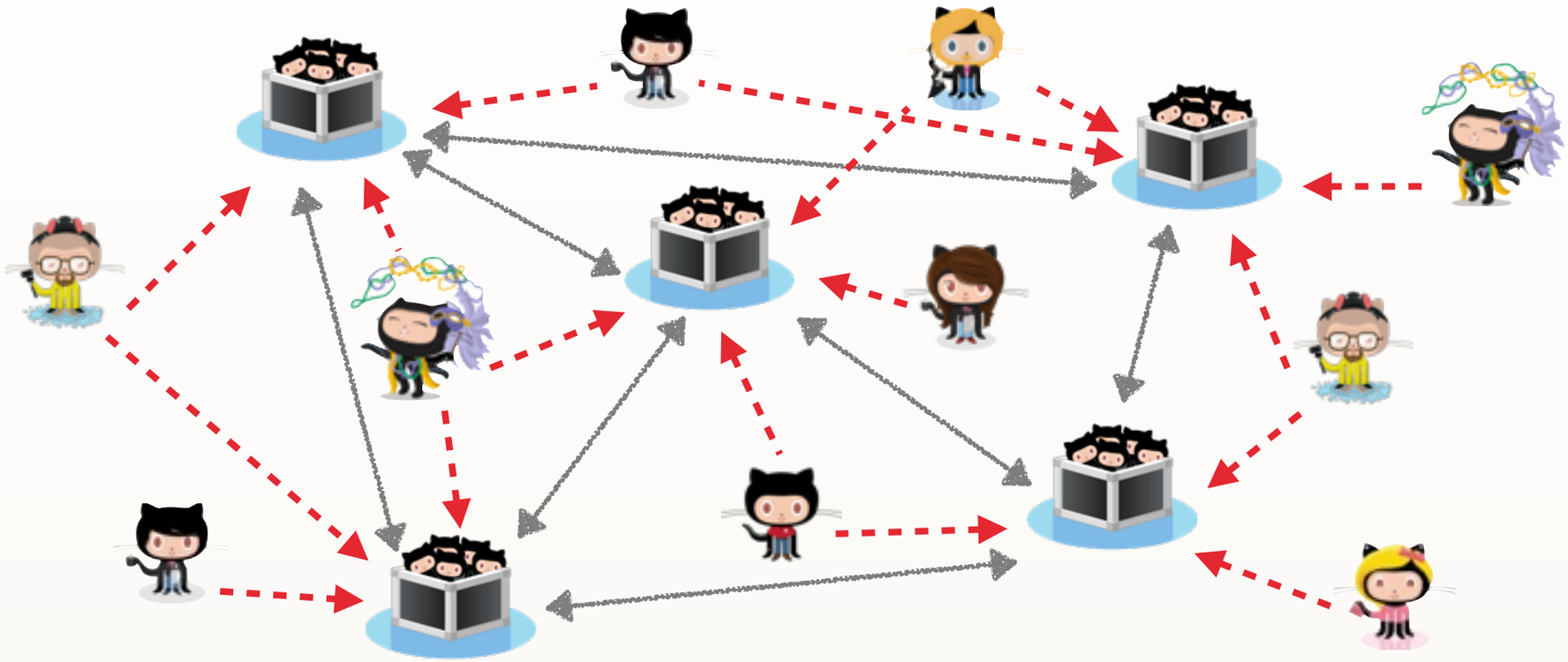
LARGE, DIVERSE, COMPLEX ECOSYSTEM



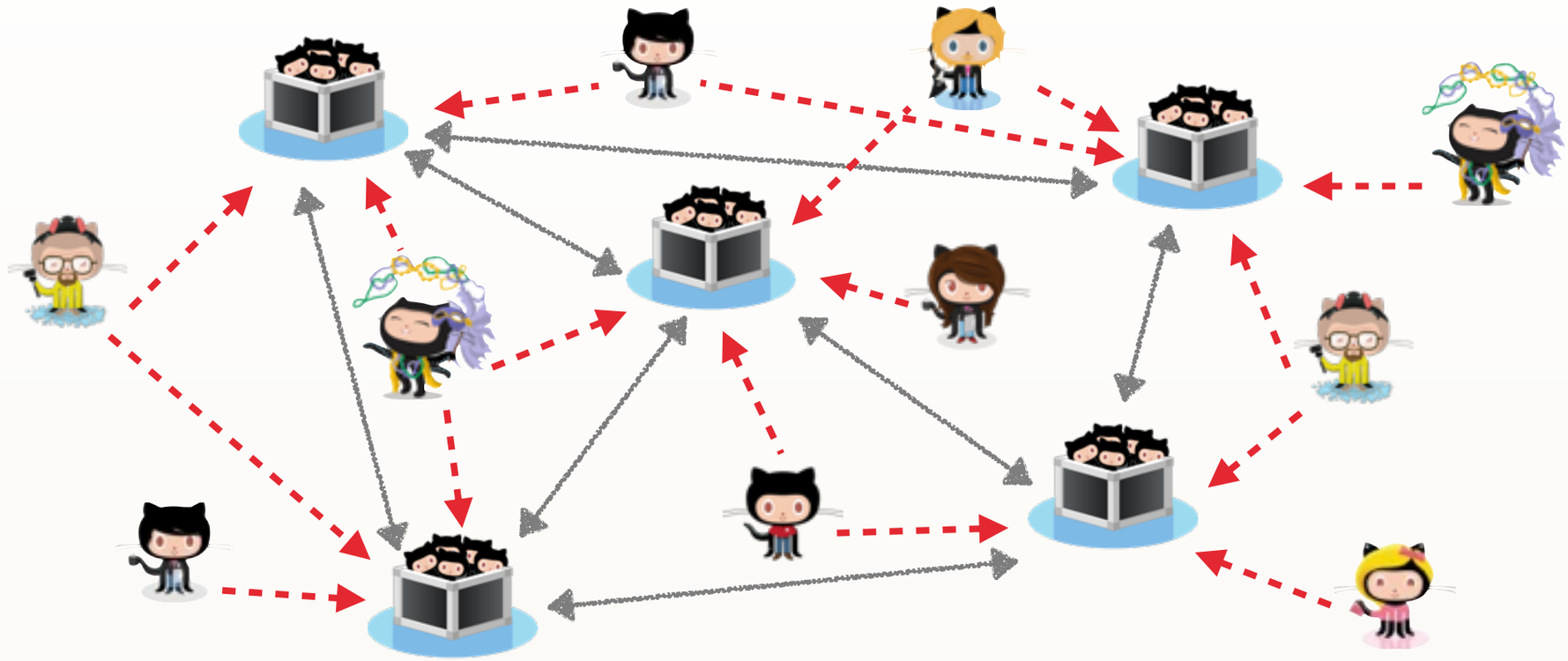
LARGE, DIVERSE, COMPLEX ECOSYSTEM



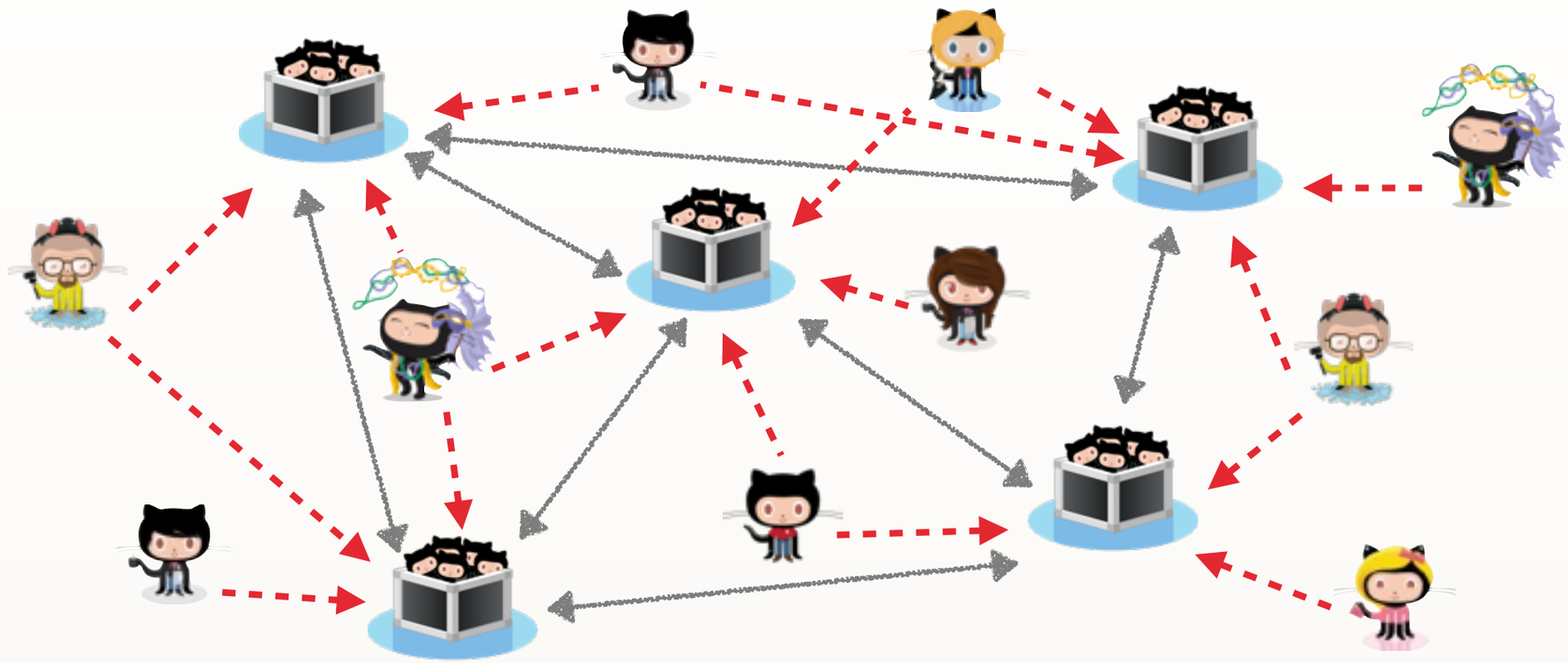
LARGE, DIVERSE, COMPLEX ECOSYSTEM



WE DON'T YET UNDERSTAND THE EFFECTS



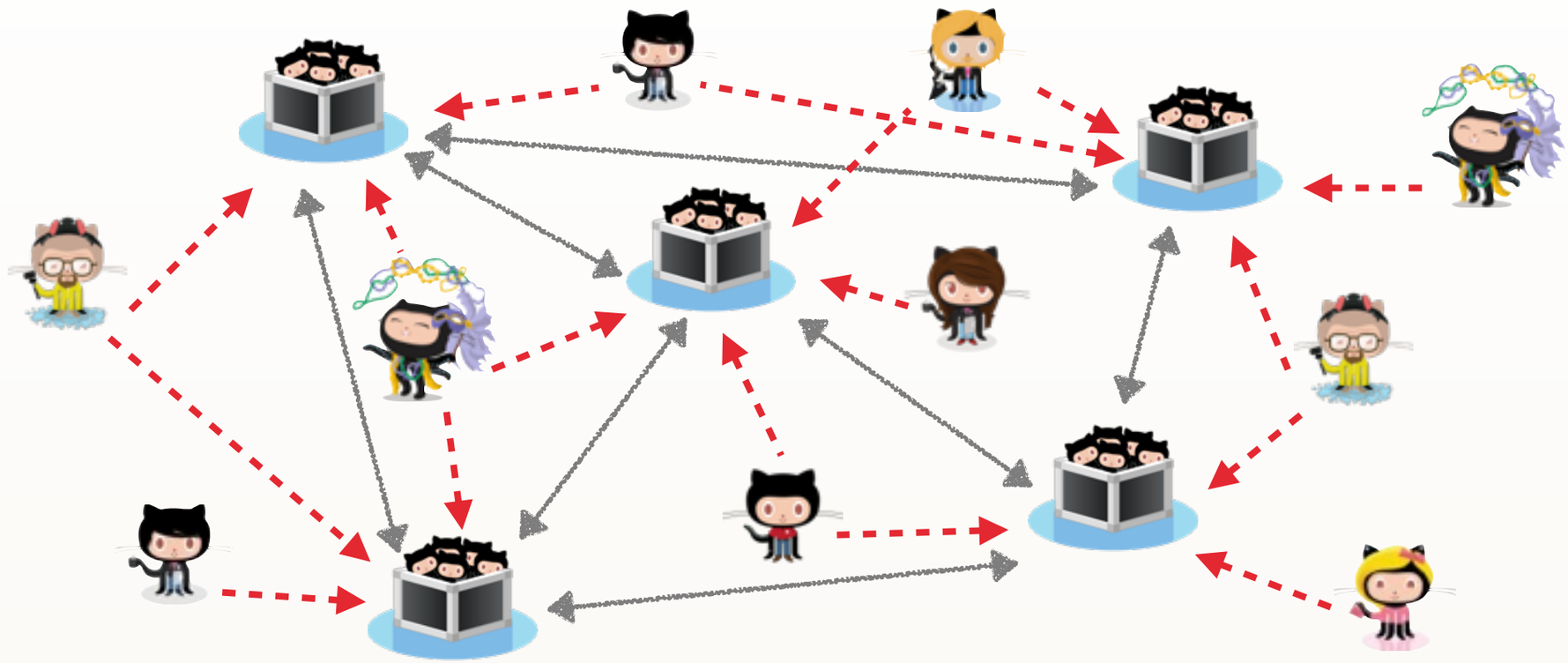
WE DON'T YET UNDERSTAND THE EFFECTS



INDIVIDUAL PRODUCTIVITY?

- Signaling
- Distraction
- Audience pressure

WE DON'T YET UNDERSTAND THE EFFECTS



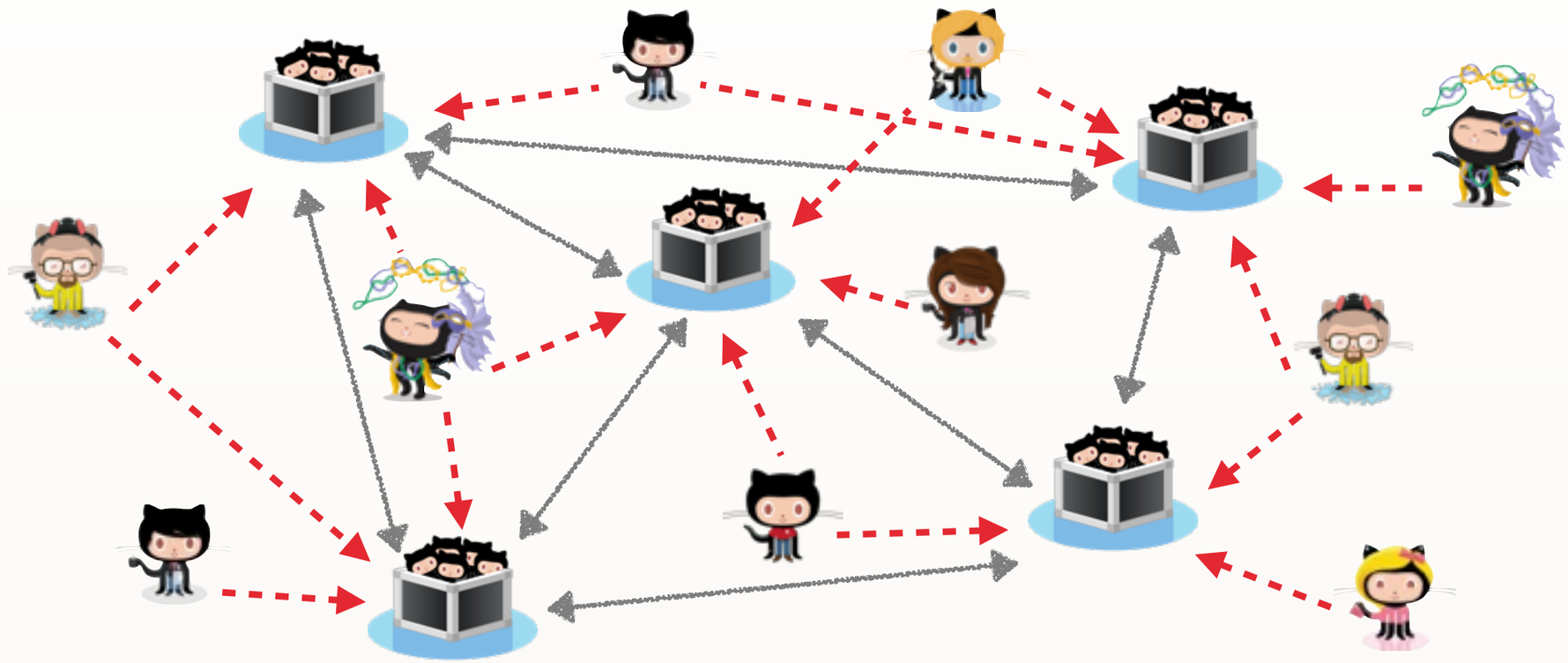
INDIVIDUAL PRODUCTIVITY?

- Signaling
- Distraction
- Audience pressure

TEAM EFFECTIVENESS?

- Teams: large, distributed, diverse
- New technology for process automation

WE DON'T YET UNDERSTAND THE EFFECTS



INDIVIDUAL PRODUCTIVITY?

- Signaling
- Distraction
- Audience pressure

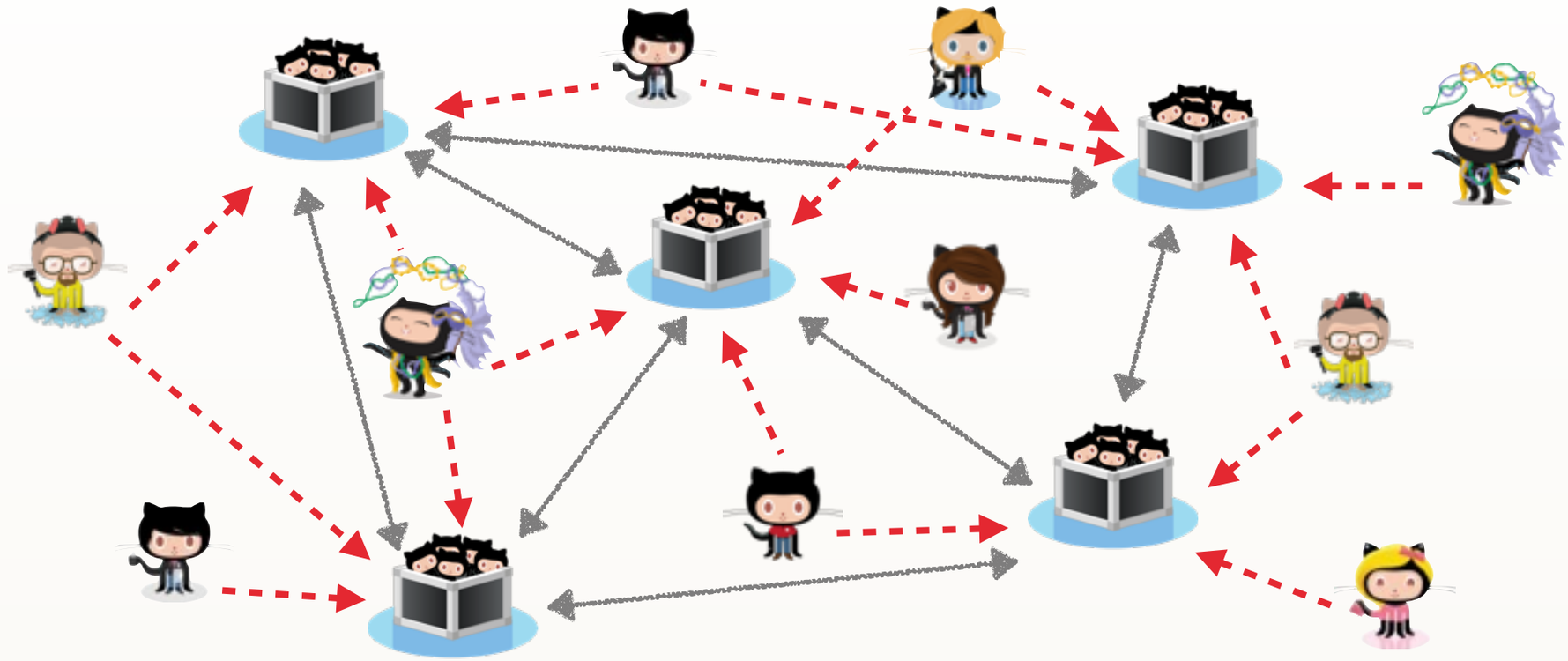
TEAM EFFECTIVENESS?

- Teams: large, distributed, diverse
- New technology for process automation

SOFTWARE QUALITY?

- More contributors
- Faster pace
- DEVOPS

EMPIRICAL STUDIES



EXPERIMENTS

Best way to control for confounds

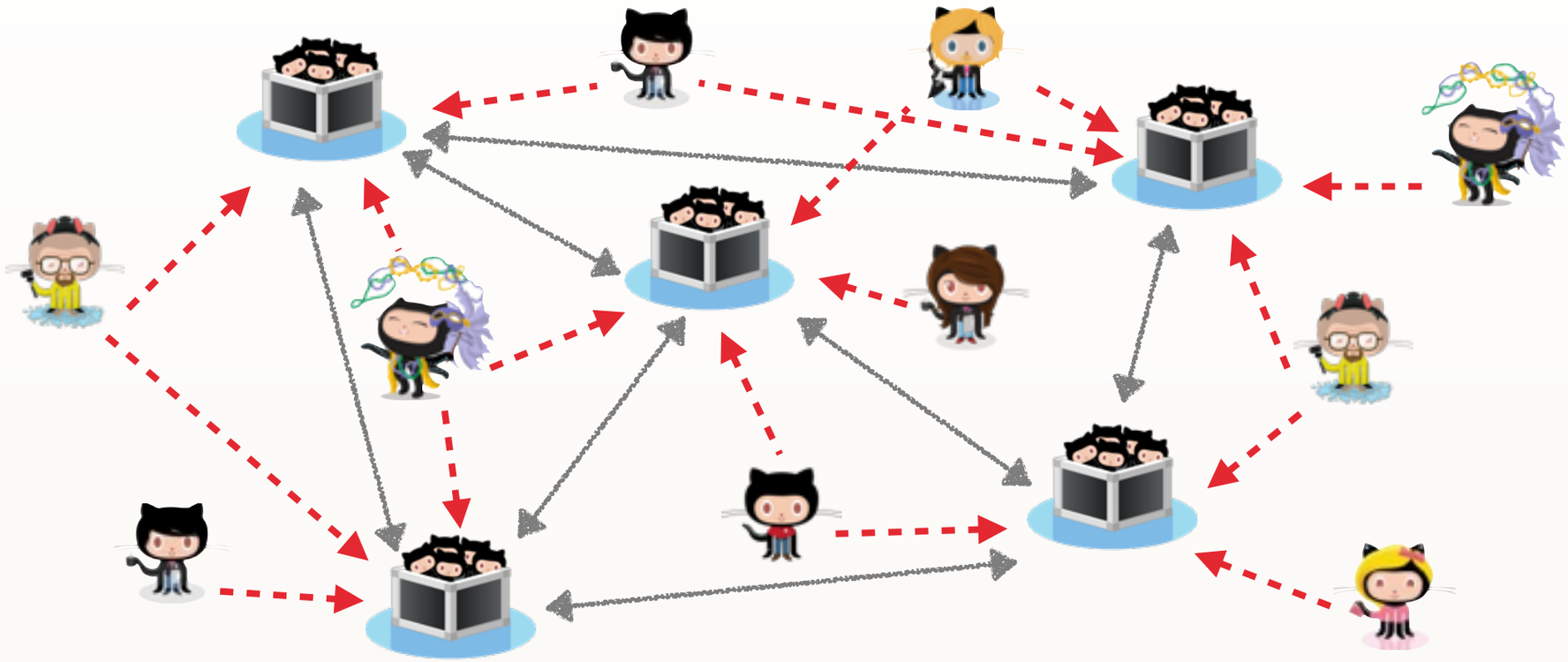
- Small sample size
- Threats to ecological validity
- Relatively expensive

QUASI-EXPERIMENTS

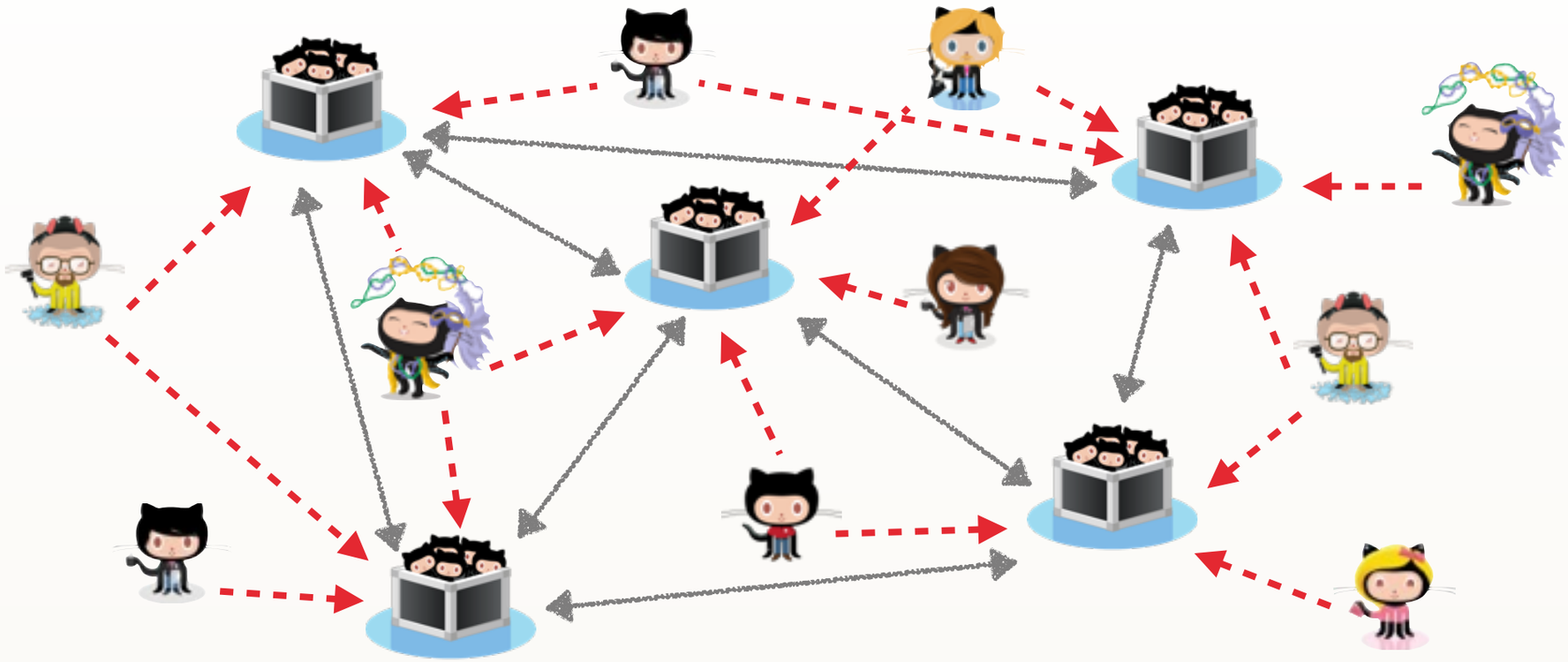
Everything is archived and can be mined

- Large samples
- “Real” data
- More generalizable
- Relatively cheap

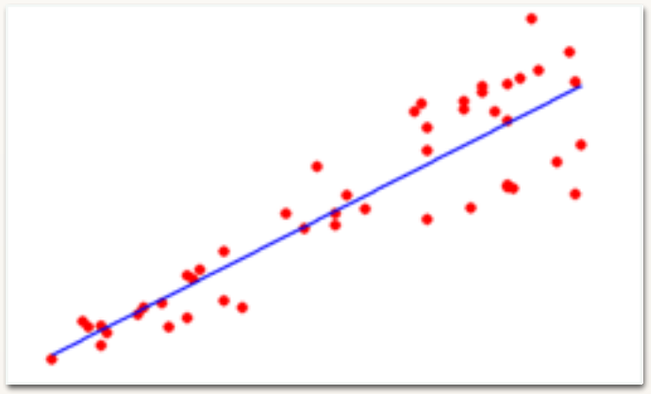
QUASI-EXPERIMENTS



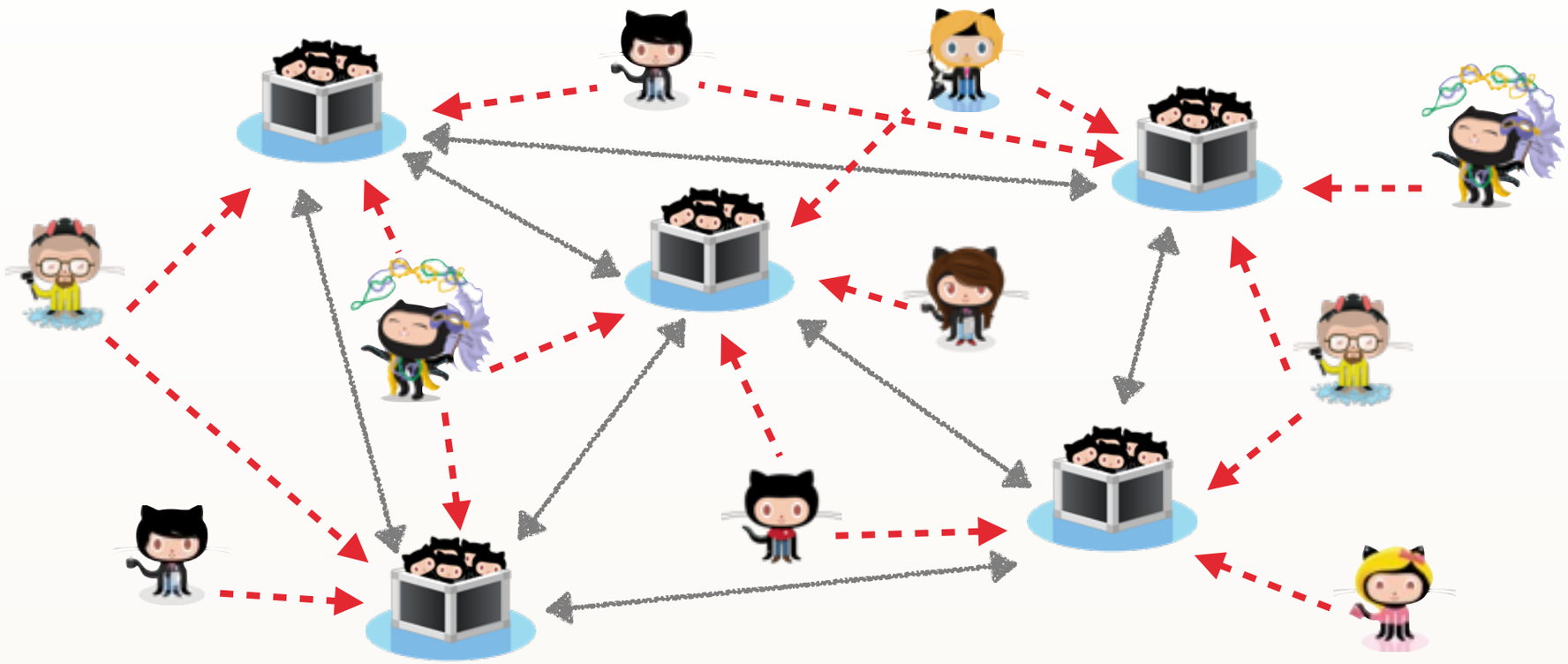
QUASI-EXPERIMENTS



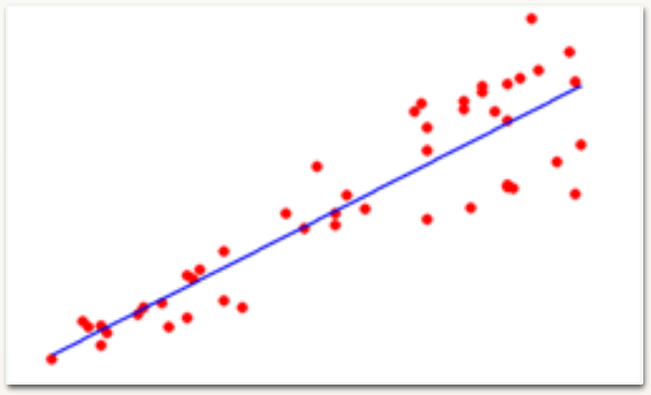
DATA ANALYSIS (STATISTICS) → TRENDS



QUASI-EXPERIMENTS



DATA ANALYSIS (STATISTICS) → TRENDS



- Analyze This! 145 Questions for Data Scientists in Software Engineering
  A. Begel, T. Zimmermann. *ICSE 2014*

DATA-DRIVEN vs. INTUITION-BASED

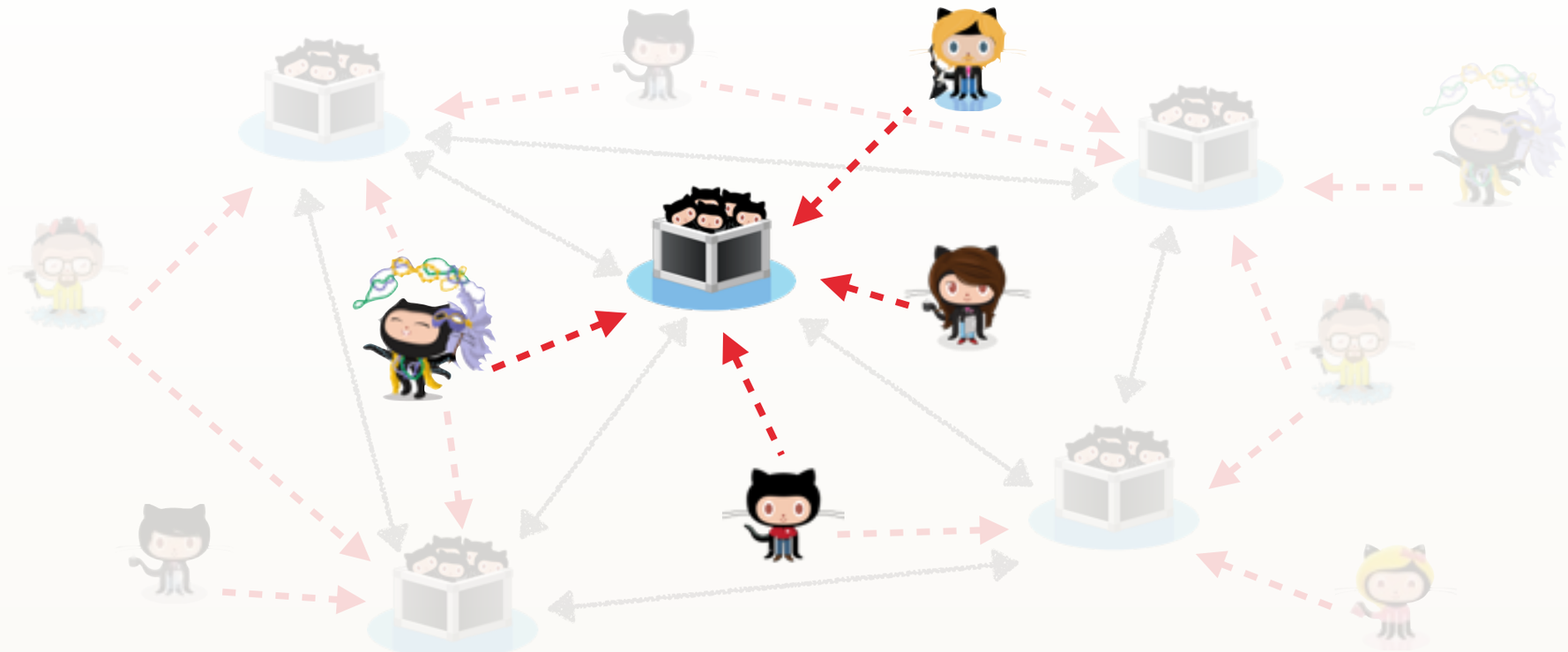
decision making

DATA SCIENTIST:

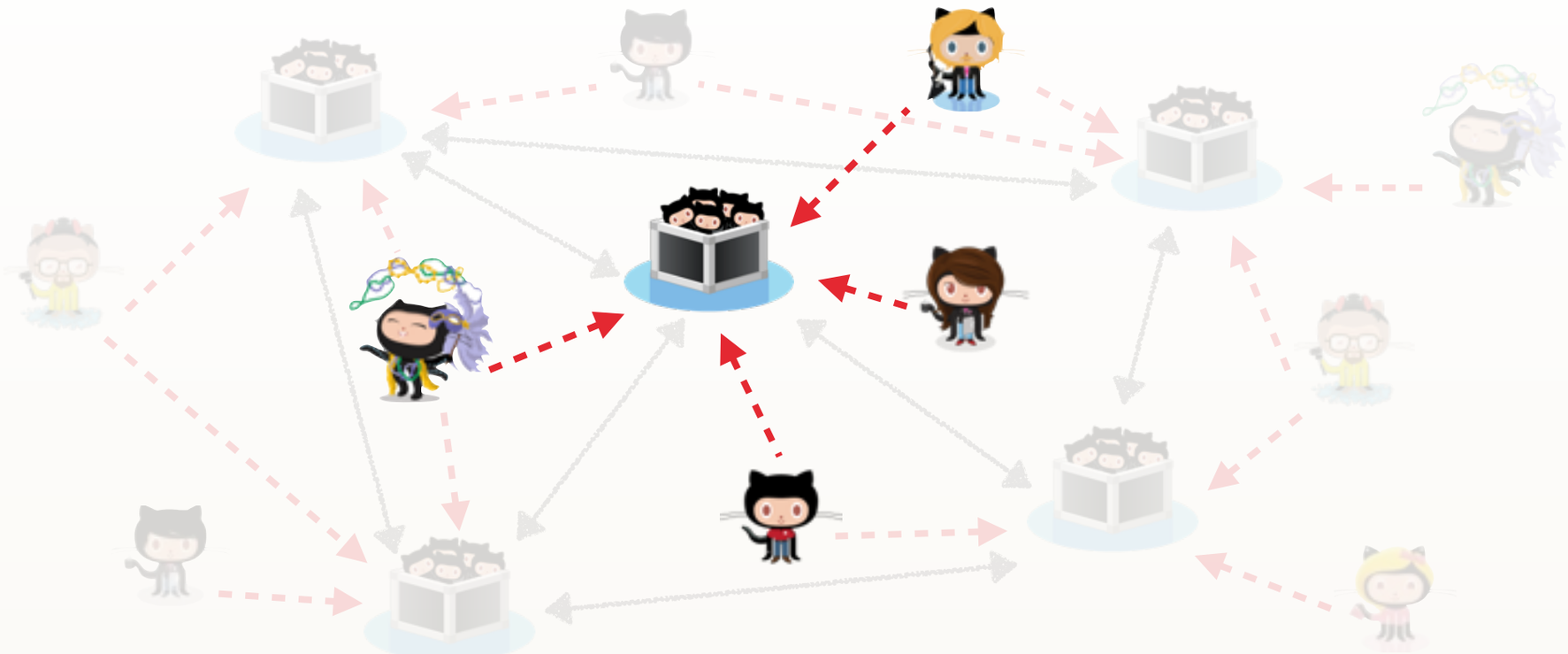
standard on software teams

- The Emerging Role of Data Scientists on Software Development Teams
  M. Kim, T. Zimmermann, R. DeLine, A. Begel. *ICSE 2016*

EXAMPLE: PULL REQUEST EVALUATION TIME



EXAMPLE: PULL REQUEST EVALUATION TIME

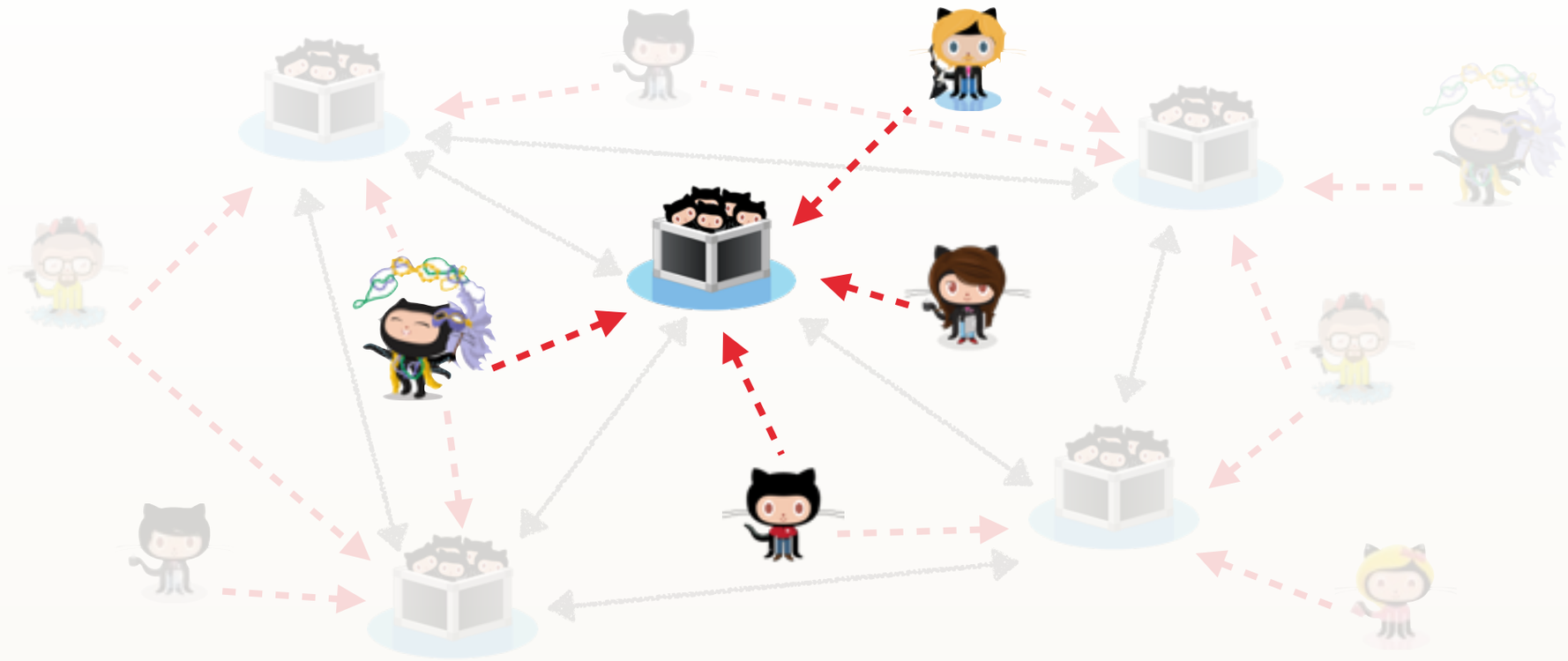


Hypothesis:

Only technical attributes matter:

- Size
- Complexity
- Tests

EXAMPLE: PULL REQUEST EVALUATION TIME



Hypothesis:

Only technical attributes matter:

- Size
- Complexity
- Tests

SOCIAL CODING!

- Submitter is core developer
- Number of followers
- Strength of social connection

… all stronger predictors than including tests

- Influence of social and technical factors for evaluating contribution in GitHub
  J. Tsay, L. Dabbish, J. Herbsleb. *ICSE 2014*
- Wait for it: Determinants of pull request evaluation latency on GitHub
  Y Yu, H Wang, V Filkov, P Devanbu, B Vasilescu. *MSR 2015*

EXPERIMENTAL RISK: BIG DATA TO THE RESCUE

12 million people

31 million repos

EXPERIMENTAL RISK: BIG DATA TO THE RESCUE

12 million people

31 million repos

1 FALSE POSITIVES

	Reject Null Hyp.	Accept Null Hyp.
Null Hyp. TRUE	1	

EXPERIMENTAL RISK: BIG DATA TO THE RESCUE

12 million people

31 million repos

1. FALSE POSITIVES
2. FALSE NEGATIVES

	Reject Null Hyp.	Accept Null Hyp.
Null Hyp. TRUE	1	
Null Hyp. FALSE		2

EXPERIMENTAL RISK: BIG DATA TO THE RESCUE

12 million people

31 million repos

1. FALSE POSITIVES
2. FALSE NEGATIVES
3. CONFOUNDS

	Reject Null Hyp.	Accept Null Hyp.
Null Hyp. TRUE	1	
Null Hyp. FALSE		2

EXPERIMENTAL RISK: BIG DATA TO THE RESCUE

12 million people

31 million repos

1. FALSE POSITIVES
2. FALSE NEGATIVES
3. CONFOUNDS

	Reject Null Hyp.	Accept Null Hyp.
Null Hyp. TRUE	1	
Null Hyp. FALSE		2

HUGE SAMPLE SIZES:

- More stringent a priori about significance level
  → reduce False Positives

EXPERIMENTAL RISK: BIG DATA TO THE RESCUE

12 million people

31 million repos

1. FALSE POSITIVES
2. FALSE NEGATIVES
3. CONFOUNDS

	Reject Null Hyp.	Accept Null Hyp.
Null Hyp. TRUE	1	
Null Hyp. FALSE		2

HUGE SAMPLE SIZES:

- More stringent a priori about significance level
  → reduce False Positives

- Detect even small effects
  → reduce False Negatives

EXPERIMENTAL RISK: BIG DATA TO THE RESCUE

12 million people

31 million repos

1. FALSE POSITIVES
2. FALSE NEGATIVES
3. CONFOUNDS

	Reject Null Hyp.	Accept Null Hyp.
Null Hyp. TRUE	1	
Null Hyp. FALSE		2

HUGE SAMPLE SIZES:

- More stringent a priori about significance level
  → reduce False Positives

- Detect even small effects
  → reduce False Negatives

- Handle more degrees of freedom
  → control for Confounds

EXPERIMENTAL RISK: BIG DATA TO THE RESCUE

12 million people

31 million repos

1. FALSE POSITIVES
2. FALSE NEGATIVES
3. CONFOUNDS

	Reject Null Hyp.	Accept Null Hyp.
Null Hyp. TRUE	1	
Null Hyp. FALSE		2

HUGE SAMPLE SIZES:

- More stringent a priori about significance level
  → reduce False Positives

- Detect even small effects
  → reduce False Negatives

- Handle more degrees of freedom
  → control for Confounds

SEPARATE SIGNAL FROM NOISE:

- Quantify effect size

EXPERIMENTAL RISK: BIG DATA TO THE RESCUE

12 million people

31 million repos

1. FALSE POSITIVES
2. FALSE NEGATIVES
3. CONFOUNDS

	Reject Null Hyp.	Accept Null Hyp.
Null Hyp. TRUE	1	
Null Hyp. FALSE		2

HUGE SAMPLE SIZES:

- More stringent a priori about significance level
  → reduce False Positives
- Detect even small effects
  → reduce False Negatives
- Handle more degrees of freedom
  → control for Confounds

SEPARATE SIGNAL FROM NOISE:

- Quantify effect size
- Mix research methods
  - Quantitative: stats, data mining, …
  - Qualitative: case studies, user surveys, grounded theory, …

EXPERIMENTAL RISK: BIG DATA TO THE RESCUE

12 million people

31 million repos

1. FALSE POSITIVES
2. FALSE NEGATIVES
3. CONFOUNDS

	Reject Null Hyp.	Accept Null Hyp.
Null Hyp. TRUE	1	
Null Hyp. FALSE		2

HUGE SAMPLE SIZES:

- More stringent a priori about significance level
  → reduce False Positives
- Detect even small effects
  → reduce False Negatives
- Handle more degrees of freedom
  → control for Confounds

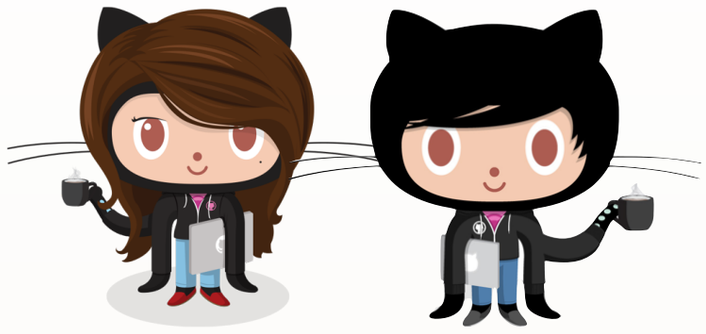
SEPARATE SIGNAL FROM NOISE:

- Quantify effect size
- Mix research methods
  - Quantitative: stats, data mining, …
  - Qualitative: case studies, user surveys, grounded theory, …

VALIDATE DATA FIRST!

- Spot-checking

TODAY



1

TEAM DIVERSITY

[CHI 2015]

2

MULTITASKING ACROSS PROJECTS

[ICSE 2016]

3

CONTINUOUS INTEGRATION

[ESEC/FSE 2015]

DIVERSITY IS RECOGNIZED AS VALUABLE

DIVERSITY IS RECOGNIZED AS VALUABLE

"Driver of internal innovation and business growth" [Forbes]

DIVERSITY IS RECOGNIZED AS VALUABLE

"Driver of internal innovation and business growth" [Forbes]

Companies with diverse executive boards have higher earnings and returns on equity [McKinsey]

DIVERSITY IS RECOGNIZED AS VALUABLE

"Driver of internal innovation and business growth" [Forbes]

Companies with diverse executive boards have higher earnings and returns on equity [McKinsey]

POLL: WHY WOULD WE WANT DIVERSITY?

DIVERSITY IS RECOGNIZED AS VALUABLE

"Driver of internal innovation and business growth" [Forbes]

Companies with diverse executive boards have higher earnings and returns on equity [McKinsey]

BENEFITS:

- access to different networks
- broader views
- creativity
- adaptability
- problem solving

...

→ INFORMATION PROCESSING THEORY

- Salancik, G. R., and Pfeffer, J. A social information processing approach to job attitudes and task design. Admin. Sci. Quart. 23, 2 (1978), 224–253

DIVERSITY IN SOFTWARE TEAMS?

vs.

1. HIGHER RISK OF:

- communication breakdown
- conflict
- confusion
- stress
- discrimination

...

DIVERSITY IN SOFTWARE TEAMS?

vs.

1. HIGHER RISK OF:

- communication breakdown
- conflict
- confusion
- stress
- discrimination

…

→ SIMILARITY ATTRACTION THEORY

→ SOCIAL IDENTITY, SOCIAL CATEGORIZATION THEORY

- Byrne, D. E. The attraction paradigm. Personality and psychopathology. Academic Press, 1971
- Tajfel, H. Social psychology of intergroup relations. Annu. Rev. Psychol. 33, 1 (1982), 1–39

DIVERSITY IN SOFTWARE TEAMS?

vs.

2. OPEN SOURCE / GITHUB ARE MERITOCRACIES

[CHASE 2015]

DIVERSITY IN SOFTWARE TEAMS?

vs.

2. OPEN SOURCE / GITHUB ARE MERITOCRACIES

"More about the contributions to the code than the `characteristics' of the person"

"Any demographic identity is irrelevant"

"Code sees no color or gender"

[CHASE 2015]

DIVERSITY IN SOFTWARE TEAMS?

3. PERCEPTION: OPEN-SOURCE IS UNFRIENDLY TO NEWCOMERS & WOMEN

"I have used a fake GitHub handle (my normal GitHub handle is my first name, which is a distinctly female name) so that people would assume I was male" [CHASE 2015]

DIVERSITY IN SOFTWARE TEAMS?

3. PERCEPTION: OPEN-SOURCE IS UNFRIENDLY TO NEWCOMERS & WOMEN

"I have used a fake GitHub handle (my normal GitHub handle is my first name, which is a distinctly female name) so that people would assume I was male" [CHASE 2015]

GENDER REPRESENTATION

stackoverflow 5.8%

github SOCIAL CODING ~5%

open source 10.9%

Google 18%

16.6%

- FLOSS 2013: A survey dataset about free software contributors: challenges for curating, sharing, and combining G Robles, L Arjona-Reina, B Vasilescu, A Serebrenik, JM Gonzalez-Barahona. *MSR 2014*
- Google Diversity (2015) www.google.com/diversity/index.html#chart
- Inside Microsoft (2015) https://goo.gl/nT4YiI

- Exploring the data on gender and GitHub repo ownership Alyssa Frazee. http://alyssafrazee.com/gender-and-github-code.html
- Stack Overflow 2015 Developer Survey (26,086 people from 157 countries) http://stackoverflow.com/research/developer-survey-2015#profile-gender

DIVERSITY IN SOFTWARE TEAMS?

3. PERCEPTION: OPEN-SOURCE IS UNFRIENDLY TO NEWCOMERS & WOMEN

"I have used a fake GitHub handle (my normal GitHub handle is my first name, which is a distinctly female name) so that people would assume I was male" [CHASE 2015]

GENDER REPRES

Does diversity create added value in GitHub teams?

stackoverfl

5.8%

gle

16.6%

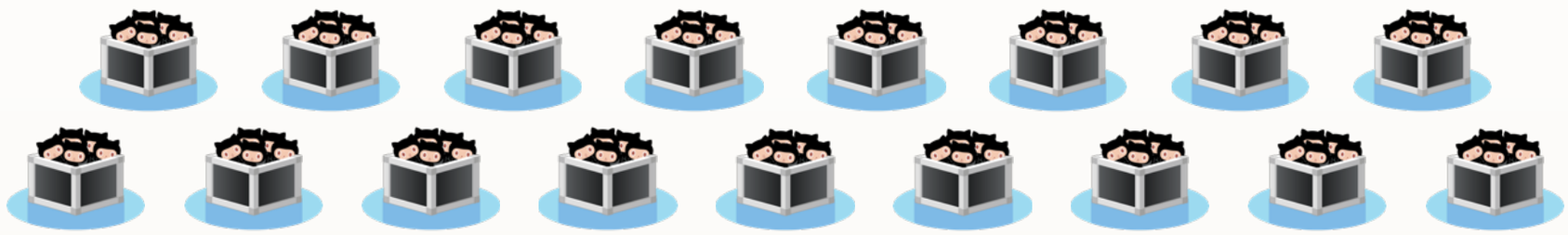
- FLOSS 2013: A survey dataset about
  for curating, sharing, and combining
  A Serebrenik, JM Gonzalez-Barahona
- Google Diversity (2015) www.google.com/diversity/index.html#chart
- Inside Microsoft (2015) https://goo.gl/nT4YiI

o repo ownership
gender-and-github-code.html

- Stack Overflow 2015 Developer Survey (26,086 people from 157 countries)
  http://stackoverflow.com/research/developer-survey-2015#profile-gender

NATURAL EXPERIMENT

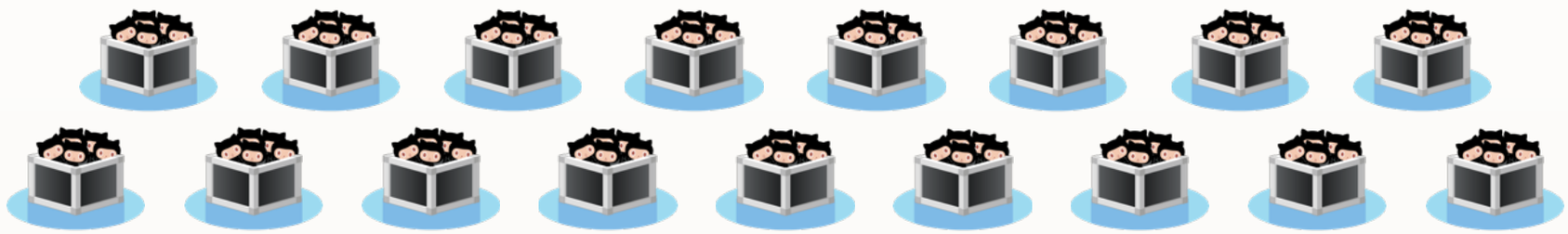
1. Mine data from many **collaborative projects**



[CHI 2015]

NATURAL EXPERIMENT

1. Mine data from many **collaborative projects**

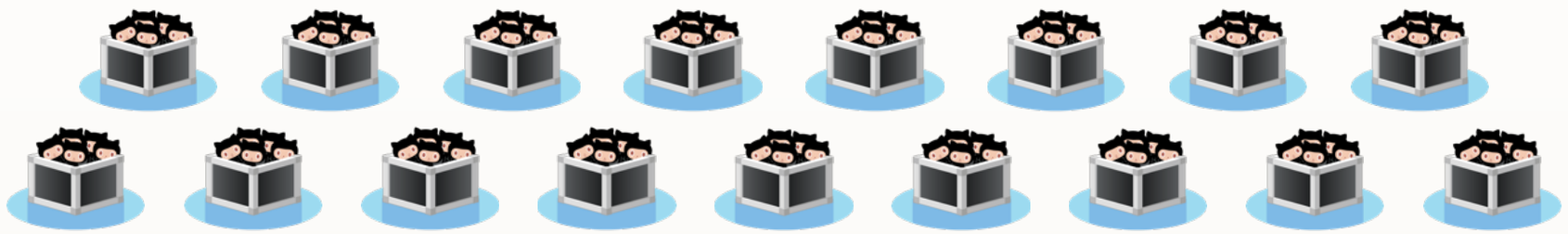


2. Compare **outputs produced per unit time**
in more/less diverse teams

[CHI 2015]

NATURAL EXPERIMENT

1. Mine data from many **collaborative projects**



2. Compare **outputs produced per unit time**
in more/less diverse teams

Gender

Tenure

[CHI 2015]

Team boundaries?

Demographics not salient?

- Demographic diversity and employee attitudes: An empirical examination of relational demography within work units. Riordan, C. M., and Shore, L. M.. *J. Appl. Psychol. 82, 3 (1997),*

User survey

4,500 invitations, 816 responses

- Demographic diversity and employee attitudes: An empirical examination of relational demography within work units. Riordan, C. M., and Shore, L. M.. *J. Appl. Psychol. 82, 3 (1997),*

Team boundaries?

Demographics not salient?

User survey

4,500 invitations, 816 responses

What constitutes a team?

Which differences do people recognize among team members?

Does diversity matter?

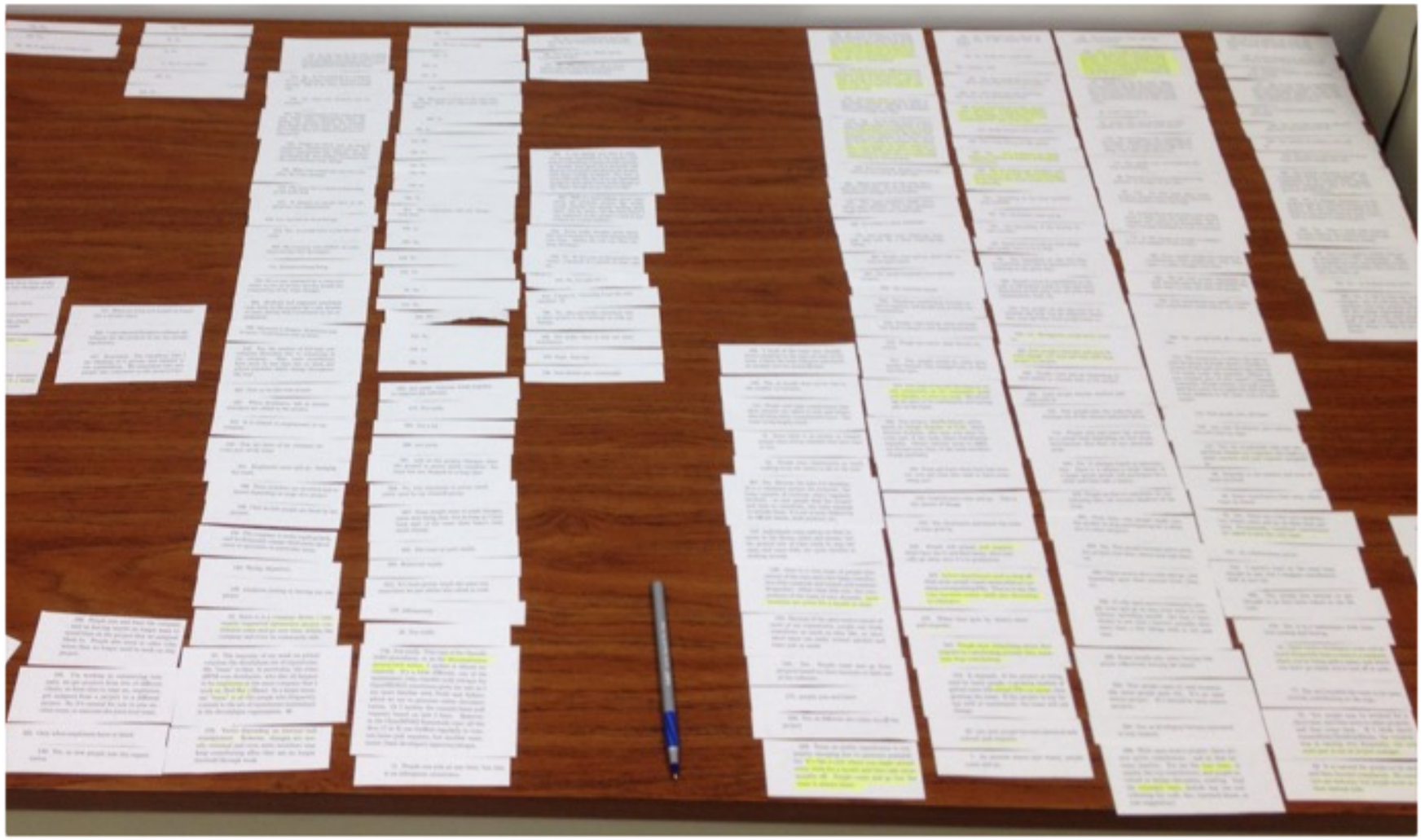
- Demographic diversity and employee attitudes: An empirical examination of relational demography within work units. Riordan, C. M., and Shore, L. M.. *J. Appl. Psychol. 82, 3 (1997),*

[CHASE 2015]

Team boundaries?

Demographics not salient?

Open card sorting



[CHASE 2015]

Team boundaries?

Demographics not salient?

User survey

4,500 invitations, 816 responses

What constitutes a team?

The team is everyone

Which differences do people recognize among team members?

Gender is surprisingly salient

Does diversity matter?

Split opinions

[CHASE 2015]

CHALLENGES

1. EXP. DESIGN **2. DATA MINING** 3. STATISTICAL ANALYSIS

Gender not explicit

Multiple aliases

GENDER TOOL

Gender not explicit

Multiple aliases

Bogdan Vasilescu
bvasiles

University of California
Davis, CA
http://bvasiles.github.io
Joined on Jul 3, 2012

Contributions | Repositories | Publi

Popular repositories

bvasiles.github.io
My website

diversity
A data set for social diversity studies of GitHu…

flask_assets_tutorial
Maxime Bouroumeau-Fuseau's tutorial on flas…

ghtorrent.org
The GHtorrent project website

ght_unmasking_aliases

Contributions

Apr May Jun Jul Aug Sep

http://github.com/tue-mdse/genderComputer
http://github.com/tue-mdse/countryNameManager

- Gender, representation and online participation: A quantitative study.
Vasilescu, B., Capiluppi, A., and Serebrenik, A. *Interacting with Computers 2014*

GENDER TOOL

Gender not explicit

Multiple aliases

Bogdan Vasilescu

bvasiles

University of California
Davis, CA
http://bvasiles.github.io
Joined on Jul 3, 2012

Contributions | Repositories | Publi

Popular repositories

- **bvasiles.github.io** — My website
- **diversity** — A data set for social diversity studies of GitHu…
- **flask_assets_tutorial** — Maxime Bouroumeau-Fuseau's tutorial on flas…
- **ghtorrent.org** — The GHtorrent project website
- **ght_unmasking_aliases**

Contributions

Apr May Jun Jul Aug Sep

http://github.com/tue-mdse/genderComputer
http://github.com/tue-mdse/countryNameManager

- Gender, representation and online participation: A quantitative study. Vasilescu, B., Capiluppi, A., and Serebrenik, A. *Interacting with Computers 2014*

GENDER TOOL

Gender not explicit

Multiple aliases

Bogdan Vasilescu
bvasiles

University of California
Davis, CA
http://bvasiles.github.io
Joined on Jul 3, 2012

Contributions | Repositories | Publi

Popular repositories

- **bvasiles.github.io** — My website
- **diversity** — A data set for social diversity studies of GitHu…
- **flask_assets_tutorial** — Maxime Bouroumeau-Fuseau's tutorial on flas…
- **ghtorrent.org** — The GHtorrent project website
- **ght_unmasking_aliases**

Contributions

Apr May Jun Jul Aug Sep

Bing Maps + Heuristics

USA

http://github.com/tue-mdse/genderComputer
http://github.com/tue-mdse/countryNameManager

- Gender, representation and online participation: A quantitative study.
Vasilescu, B., Capiluppi, A., and Serebrenik, A. *Interacting with Computers 2014*

GENDER TOOL

Gender not explicit

Multiple aliases

Bogdan Vasilescu
bvasiles

University of California
Davis, CA
http://bvasiles.github.io
Joined on Jul 3, 2012

Contributions | Repositories | Publi

Popular repositories

- **bvasiles.github.io** — My website
- **diversity** — A data set for social diversity studies of GitHu…
- **flask_assets_tutorial** — Maxime Bouroumeau-Fuseau's tutorial on flas…
- **ghtorrent.org** — The GHtorrent project website
- **ght_unmasking_aliases**

Contributions

Apr May Jun Jul Aug Sep

Bing Maps + Heuristics

Bogdan + USA

Name frequency tables for 30 countries

male

http://github.com/tue-mdse/genderComputer
http://github.com/tue-mdse/countryNameManager

- Gender, representation and online participation: A quantitative study. Vasilescu, B., Capiluppi, A., and Serebrenik, A. *Interacting with Computers 2014*

GENDER TOOL

Gender not explicit

Multiple aliases

Bogdan Vasilescu

bvasiles

University of California
Davis, CA
http://bvasiles.github.io
Joined on Jul 3, 2012

Contributions | Repositories | Publi

Popular repositories

- **bvasiles.github.io** — My website
- **diversity** — A data set for social diversity studies of GitHu…
- **flask_assets_tutorial** — Maxime Bouroumeau-Fuseau's tutorial on flas…
- **ghtorrent.org** — The GHtorrent project website
- **ght_unmasking_aliases**

Contributions

Apr May Jun Jul Aug Sep

Bing Maps + Heuristics

Bogdan + USA

Name frequency tables for 30 countries

male

Location matters!

- Andrea (Italy) → male
- Andrea (USA) → female

http://github.com/tue-mdse/genderComputer
http://github.com/tue-mdse/countryNameManager

- Gender, representation and online participation: A quantitative study. Vasilescu, B., Capiluppi, A., and Serebrenik, A. *Interacting with Computers 2014*

CHALLENGES

DEALIASING TOOL

Gender not explicit

Multiple aliases

INTUITION:

Laurent Gautier - laurent@cbs.dtu.dk

Laurent Gautier - s010592@student.dtu.dk

Laurent - lgautier@gmail.com

- lgautier@altern.org

- Mining email social networks. Bird, C., et al. *MSR 2006*

DEALIASING TOOL

Gender not explicit

Multiple aliases

INTUITION:

- first name

Laurent Gautier - laurent@cbs.dtu.dk

Laurent Gautier - s010592@student.dtu.dk

Laurent - lgautier@gmail.com

- lgautier@altern.org

- Mining email social networks. Bird, C., et al. *MSR 2006*

DEALIASING TOOL

Gender not explicit

Multiple aliases

INTUITION:

- first name
- email prefix

Laurent Gautier - laurent@cbs.dtu.dk

Laurent Gautier - s010592@student.dtu.dk

Laurent - lgautier@gmail.com

- lgautier@altern.org

- Mining email social networks. Bird, C., et al. *MSR 2006*

CHALLENGES

DEALIASING TOOL

Gender not explicit

Multiple aliases

INTUITION:

- first name
- email prefix
- first initial + last name

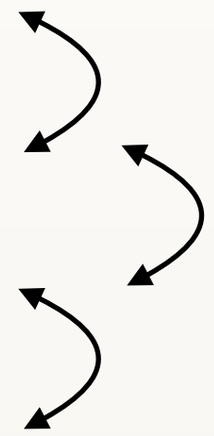
...

Laurent Gautier - laurent@cbs.dtu.dk

Laurent Gautier - s010592@student.dtu.dk

Laurent - lgautier@gmail.com

- lgautier@altern.org



- Mining email social networks. Bird, C., et al. *MSR 2006*

REGRESSION

Outputs produced / unit time
(#Commits/quarter)

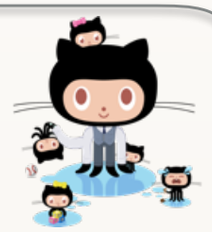
response

REGRESSION

Outputs produced / unit time
(#Commits/quarter)

response

Gender diversity
(Blau)



Tenure diversity
(CV)

main predictors

REGRESSION

Outputs produced / unit time
(#Commits/quarter)

response

Gender diversity
(Blau)

Tenure diversity
(CV)

main predictors

Total commits

Project size

controls

REGRESSION

Outputs produced / unit time
(#Commits/quarter)

response

Gender diversity
(Blau)

Tenure diversity
(CV)

main predictors

Total commits

Project size

Team size

Experience

Human resources

controls

REGRESSION

Outputs produced / unit time
(#Commits/quarter)

response

Gender diversity
(Blau)

Tenure diversity
(CV)

main predictors

Total commits

Project size

Team size

Experience

Human resources

Project age

Time

Evolution of GitHub & time passing

controls

REGRESSION

Outputs produced / unit time
(#Commits/quarter)

response

Gender diversity
(Blau)

Tenure diversity
(CV)

main predictors

Total commits

Project size

Team size

Experience

Human resources

Project age

Time

Evolution of GitHub & time passing

Comments

Forks

Popularity
Distributed development

controls

CHALLENGES

Project	Created on	Project age	Total #commits	#Forks	Time	#Commits	#Comments	Team size	Gender diversity	Commit tenure diversity	Turnover
A	2011-02-15	12	557	51	Q2	47	26	9	0.25	0.47	0.67
					Q5	19	12	10	0.00	0.93	0.75
					Q6	7	13	12	0.25	0.54	0.67
					Q7	56	53	20	0.00	0.56	0.87
B	2010-09-21	11	2075	578	Q4	71	169	83	0.03	0.66	0.87
					Q5	116	219	93	0.05	0.73	0.56
					Q6	186	367	119	0.06	0.80	0.86
					Q7	129	453	114	0.08	0.85	0.82

CHALLENGES

Different projects ...

Project	Created on	Project age	Total #commits	#Forks	Time	#Commits	#Comments	Team size	Gender diversity	Commit tenure diversity	Turnover
A	2011-02-15	12	557	51	Q2	47	26	9	0.25	0.47	0.67
					Q5	19	12	10	0.00	0.93	0.75
					Q6	7	13	12	0.25	0.54	0.67
					Q7	56	53	20	0.00	0.56	0.87
B	2010-09-21	11	2075	578	Q4	71	169	83	0.03	0.66	0.87
					Q5	116	219	93	0.05	0.73	0.56
					Q6	186	367	119	0.06	0.80	0.86
					Q7	129	453	114	0.08	0.85	0.82

Different projects ...

... observed over time

Project	Created on	Project age	Total #commits	#Forks	Time	#Commits	#Comments	Team size	Gender diversity	Commit tenure diversity	Turnover
A	2011-02-15	12	557	51	Q2	47	26	9	0.25	0.47	0.67
					Q5	19	12	10	0.00	0.93	0.75
					Q6	7	13	12	0.25	0.54	0.67
					Q7	56	53	20	0.00	0.56	0.87
B	2010-09-21	11	2075	578	Q4	71	169	83	0.03	0.66	0.87
					Q5	116	219	93	0.05	0.73	0.56
					Q6	186	367	119	0.06	0.80	0.86
					Q7	129	453	114	0.08	0.85	0.82

CHALLENGES 1. EXP. DESIGN 2. DATA MINING 3. STATISTICAL ANALYSIS

Different projects ...

... observed over time

Outputs produced

Project	Created on	Project age	Total #commits	#Forks	Time	#Commits	#Comments	Team size	Gender diversity	Commit tenure diversity	Turnover
A	2011-02-15	12	557	51	Q2	47	26	9	0.25	0.47	0.67
					Q5	19	12	10	0.00	0.93	0.75
					Q6	7	13	12	0.25	0.54	0.67
					Q7	56	53	20	0.00	0.56	0.87
B	2010-09-21	11	2075	578	Q4	71	169	83	0.03	0.66	0.87
					Q5	116	219	93	0.05	0.73	0.56
					Q6	186	367	119	0.06	0.80	0.86
					Q7	129	453	114	0.08	0.85	0.82

Different projects ... | ... observed over time | Outputs produced | Diversity measures

Project	Created on	Project age	Total #commits	#Forks	Time	#Commits	#Comments	Team size	Gender diversity	Commit tenure diversity	Turnover
A	2011-02-15	12	557	51	Q2	47	26	9	0.25	0.47	0.67
					Q5	19	12	10	0.00	0.93	0.75
					Q6	7	13	12	0.25	0.54	0.67
					Q7	56	53	20	0.00	0.56	0.87
B	2010-09-21	11	2075	578	Q4	71	169	83	0.03	0.66	0.87
					Q5	116	219	93	0.05	0.73	0.56
					Q6	186	367	119	0.06	0.80	0.86
					Q7	129	453	114	0.08	0.85	0.82

CHALLENGES 1. EXP. DESIGN 2. DATA MINING 3. STATISTICAL ANALYSIS

Different projects ... | ... observed over time | Outputs produced | Diversity measures

Project	Created on	Project age	Total #commits	#Forks	Time	#Commits	#Comments	Team size	Gender diversity	Commit tenure diversity	Turnover
A	2011-02-15	12	557	51	Q2	47	26	9	0.25	0.47	0.67
					Q5	19	12	10	0.00	0.93	0.75
					Q6	7	13	12	0.25	0.54	0.67
					Q7	56	53	20	0.00	0.56	0.87
B	2010-09-21	11	2075	578	Q4	71	169	83	0.03	0.66	0.87
					Q5	116	219	93	0.05	0.73	0.56
					Q6	186	367	119	0.06	0.80	0.86
					Q7	129	453	114	0.08	0.85	0.82

LINEAR MIXED-EFFECTS REGRESSION

Longitudinal data

Nesting: projects

Random effects: project, time

Random slope: team size | project

RESULTS

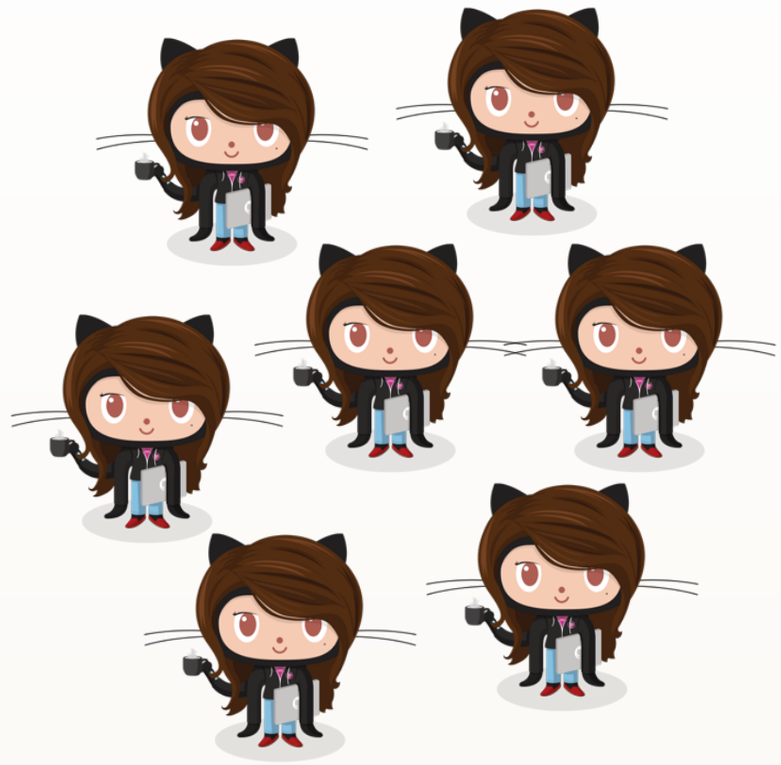
Higher productivity

vs.

RESULTS

Higher productivity

vs.



Other confounds held fixed, **higher team diversity (gender & tenure)** is associated with **increased code production** (commits per quarter),

But small effects!

RESULTS

Higher productivity

vs.

Other confounds held fixed, **higher team diversity (gender & tenure)** is associated with **increased code production** (commits per quarter),

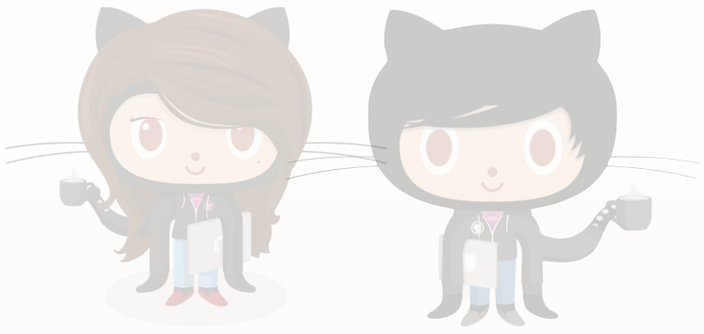
But small effects!

ONGOING / FUTURE WORK:

- Diversity effects beyond code production (e.g., **team cohesiveness** & **code quality**)
- Why are social coding platforms so **exclusive**?

Gamification?

TODAY



1

TEAM DIVERSITY

[CHI 2015]

2

MULTITASKING ACROSS PROJECTS

[ICSE 2016]

3

CONTINUOUS INTEGRATION

[ESEC/FSE 2015]

WORKING ON MULTIPLE PROJECTS IN PARALLEL

REASONS:

- Dependencies
- Downtime
- Being “stuck” in one project
- Request from other dev’s
- Personal interest
- Signaling
- …

- Working for free? Motivations of participating in open source projects
  A. Hars and S. Ou. *HICSS, 2001*
- The open source software development phenomenon: An analysis based on social network theory
  G. Madey, V. Freeh, and R. Tynan. *AMCIS, 2002*
- Activity traces and signals in software developer recruitment and hiring
  J Marlow, L Dabbish. *CSCW 2013*

WORKING ON MULTIPLE PROJECTS IN PARALLEL

REASONS:

- Dependencies
- Downtime
- Being “stuck” in one project
- Request from other dev’s
- Personal interest
- Signaling
- …

PROS:

- Fill downtime
- Cross-fertilisation

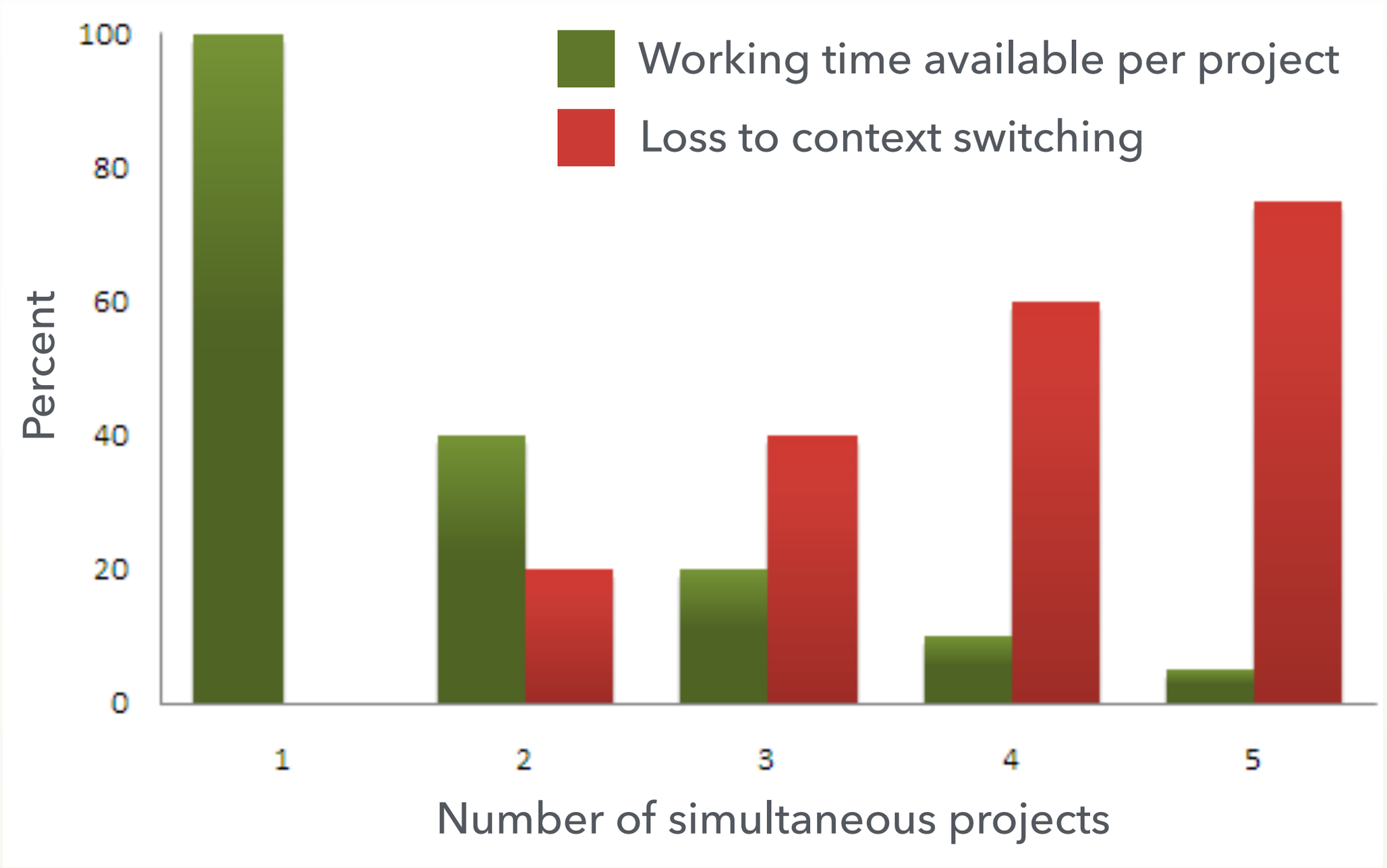
CONS:

- Distraction
- Cognitive switching cost - storing state

- Memory for goals: An activation-based model
  E. M. Altmann and J. G. Trafton.
  Cognitive Science, 26(1):39–83, 2002
- What makes interruptions disruptive? A process-model account of the effects of the problem state bottleneck on task interruption and resumption
  J. P. Borst, N. A. Taatgen, and H. van Rijn. *CHI 2015*

SWITCHING PROJECTS IS EXPENSIVE

ANECDOTAL RULE OF THUMB [G. Weinberg, 1992-7]

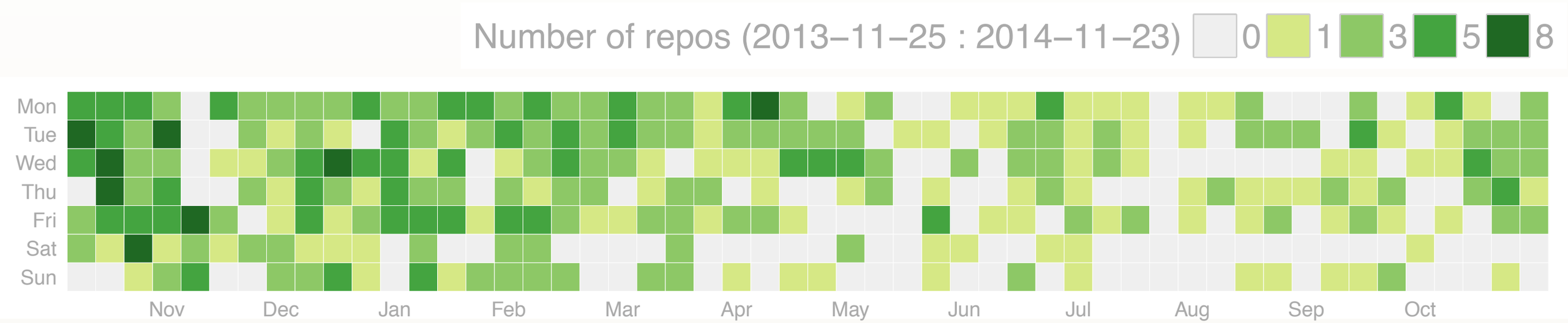


From: http://blog.codinghorror.com/the-multi-tasking-myth/

- Quality Software Management, 1: Systems Thinking. G. Weinberg. 1992. Dorset House Publishing

GITHUB DEV'S MULTITASK ACROSS PROJECTS OFTEN

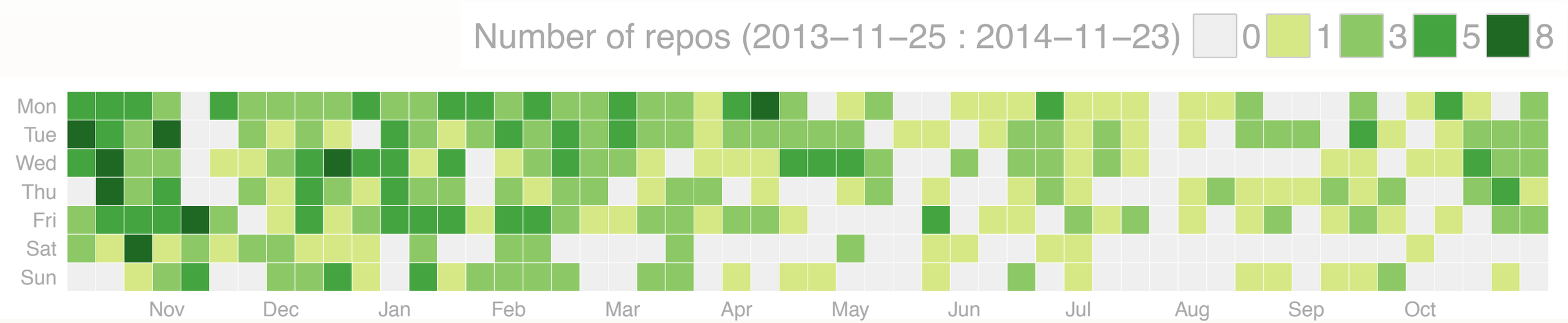
EXAMPLE BEHAVIOR:



[ICSE 2016]

GITHUB DEV'S MULTITASK ACROSS PROJECTS OFTEN

EXAMPLE BEHAVIOR:



PEOPLE WHO MULTITASK:

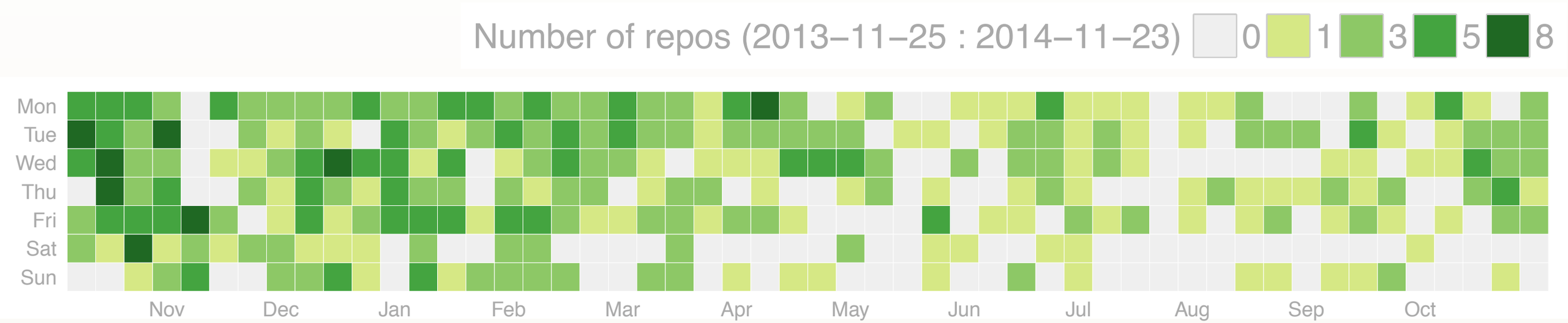
- Feel more productive
- Believe they contribute more code

User survey (128 responses)

[ICSE 2016]

GITHUB DEV'S MULTITASK ACROSS PROJECTS OFTEN

EXAMPLE BEHAVIOR:



PEOPLE WHO MULTITASK:

- Feel more productive
- Believe they contribute more code

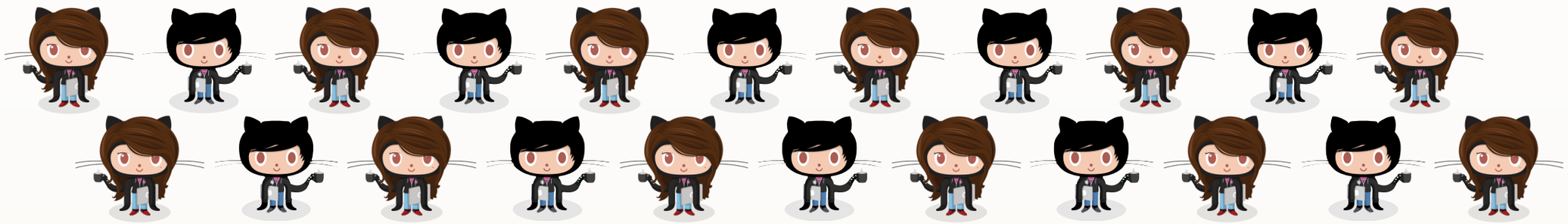
User survey (128 responses)

Is there a limit to multitasking?

[ICSE 2016]

NATURAL EXPERIMENT

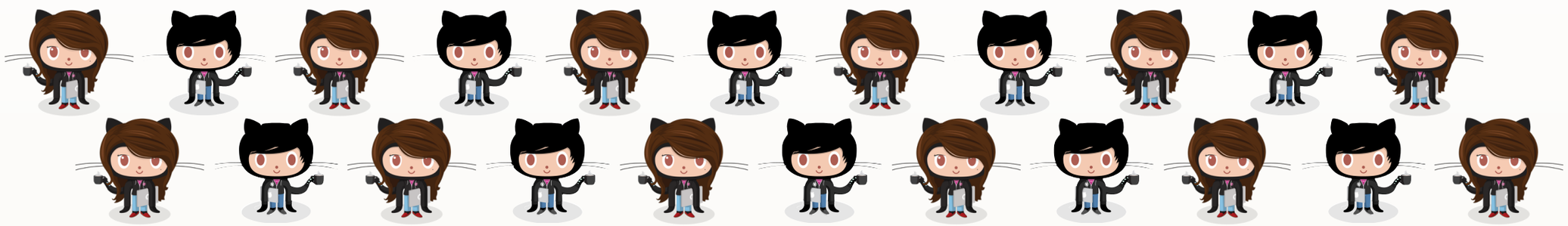
1. Mine data on ~1200 **prolific developers**



[ICSE 2016]

NATURAL EXPERIMENT

1. Mine data on ~1200 **prolific developers**



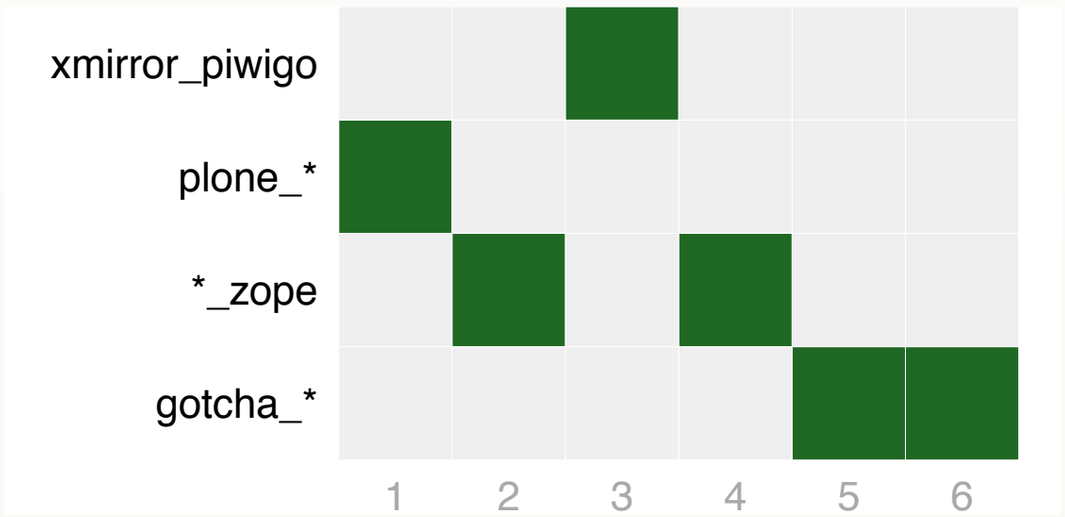
2. Compare **outputs produced per unit time**
(LOC added / week)
in different multitasking & project switching conditions

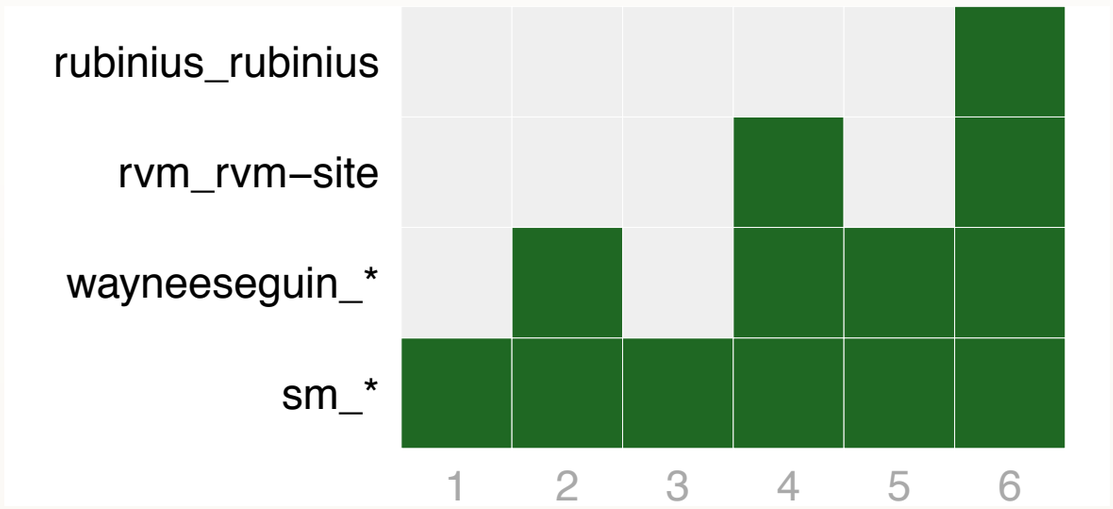
[ICSE 2016]

MULTITASKING DIMENSIONS

1. PROJECTS PER DAY

Working sequentially vs. Within-day multitasking

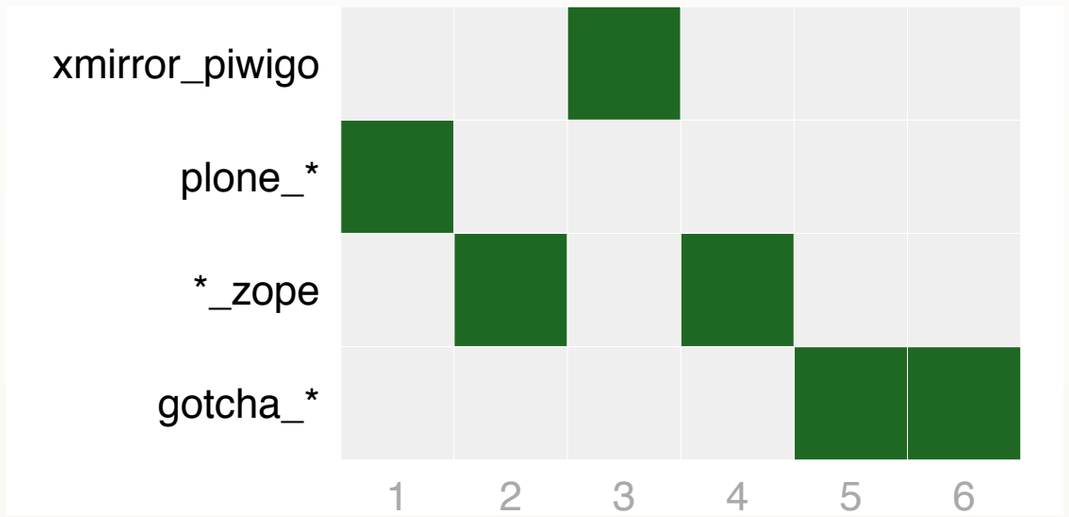


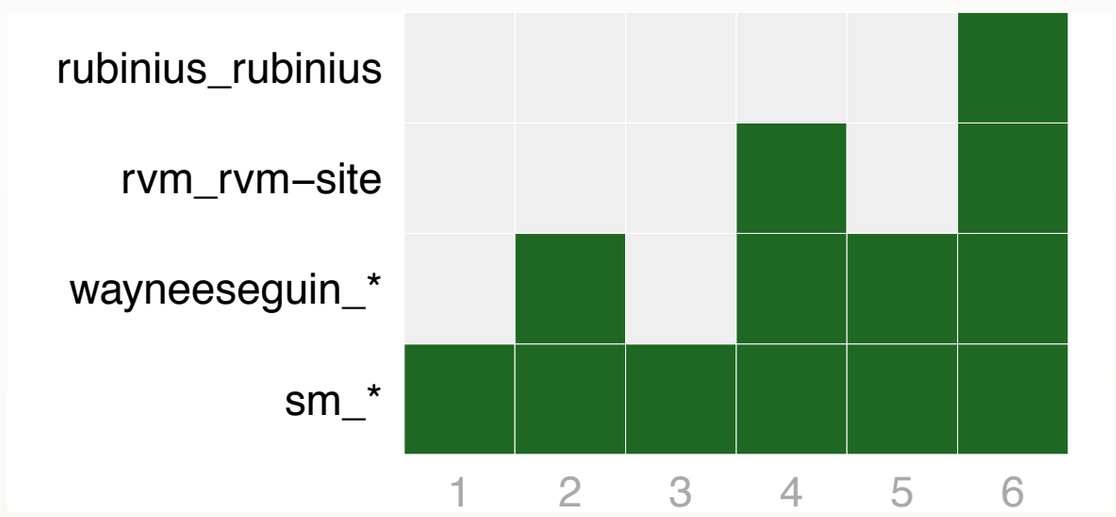


MULTITASKING DIMENSIONS

1. PROJECTS PER DAY

Working sequentially vs. Within-day multitasking





$$\text{AvgProjectsPerDay} = 1$$

$$\text{AvgProjectsPerDay} = 2.2$$

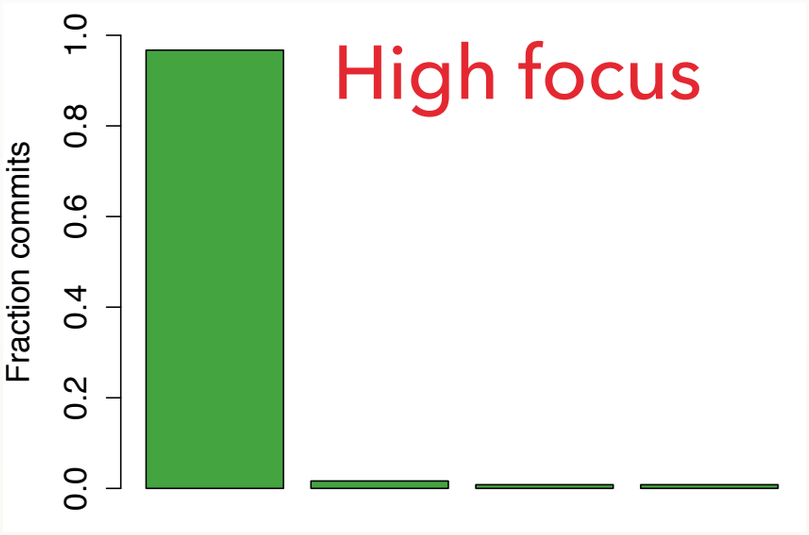
MULTITASKING DIMENSIONS

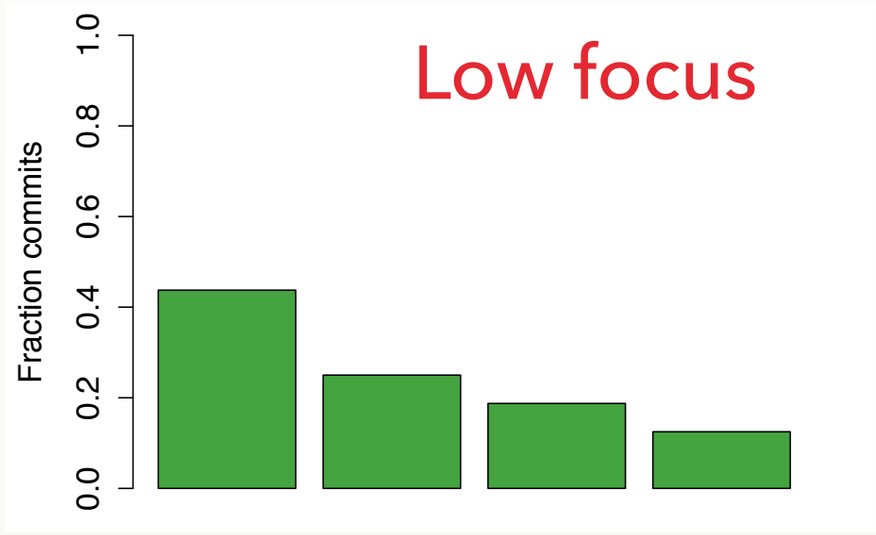
2. WEEKLY FOCUS

Working mostly on one project

vs.

Contributing evenly to all projects





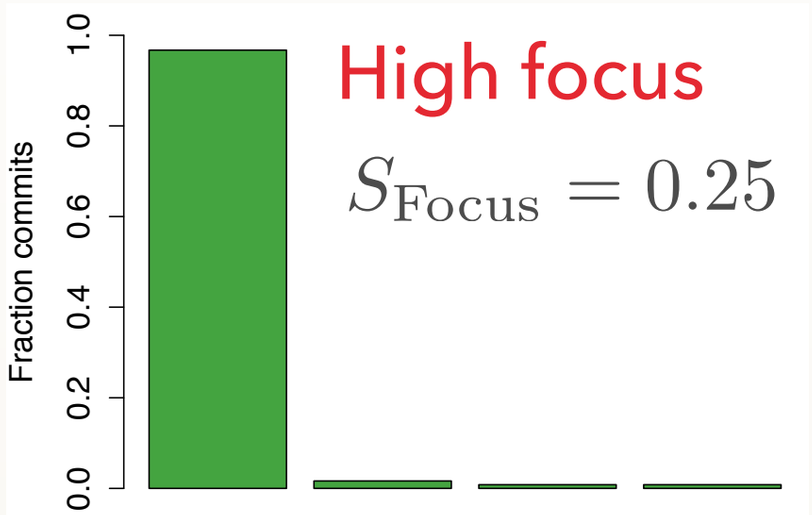
MULTITASKING DIMENSIONS

2. WEEKLY FOCUS

Working mostly on one project

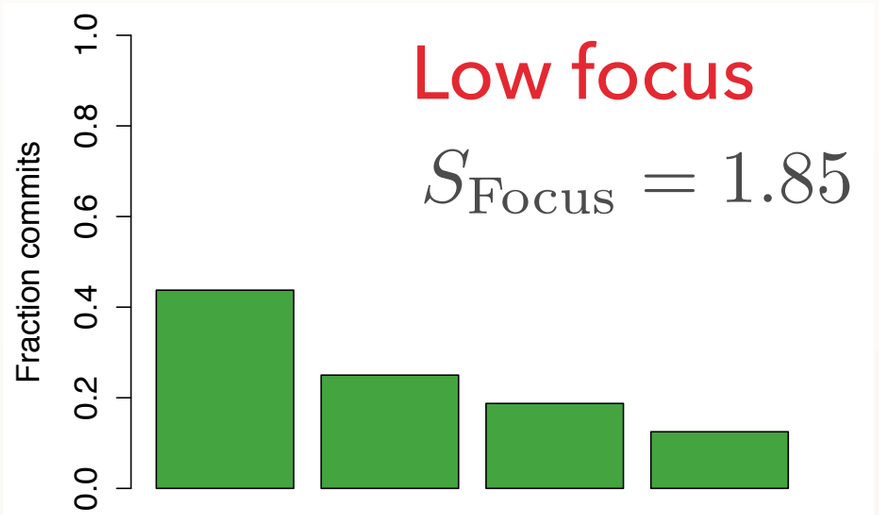
vs.

Contributing evenly to all projects



High focus

$S_{\mathrm{Focus}} = 0.25$



Low focus

$S_{\mathrm{Focus}} = 1.85$

Shannon entropy:

$$S_{\mathrm{Focus}} = -\sum_{i=1}^{N} p_i \log_2 p_i$$

N ← Projects this week

Fraction commits in project i

MULTITASKING DIMENSIONS

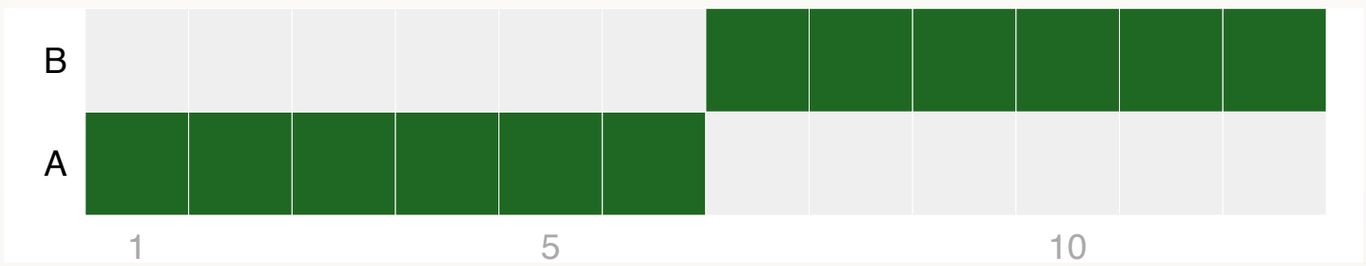
3. DAY-TO-DAY FOCUS

Repetitive day-to-day working style

vs.

Changing focus one day to next

$$\text{AvgProjectsPerDay} = 1$$
$$S_{\text{Focus}} = 1$$



$$\text{AvgProjectsPerDay} = 1$$
$$S_{\text{Focus}} = 1$$

MULTITASKING DIMENSIONS

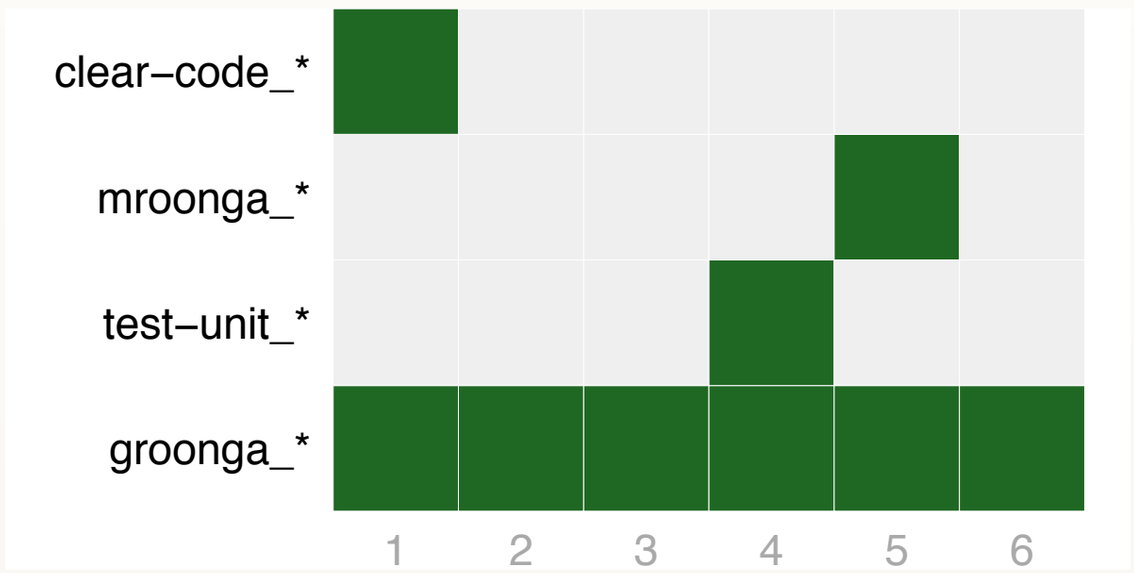
3. DAY-TO-DAY FOCUS

Repetitive day-to-day working style

vs.

Changing focus one day to next

Focus shifting networks



MULTITASKING DIMENSIONS

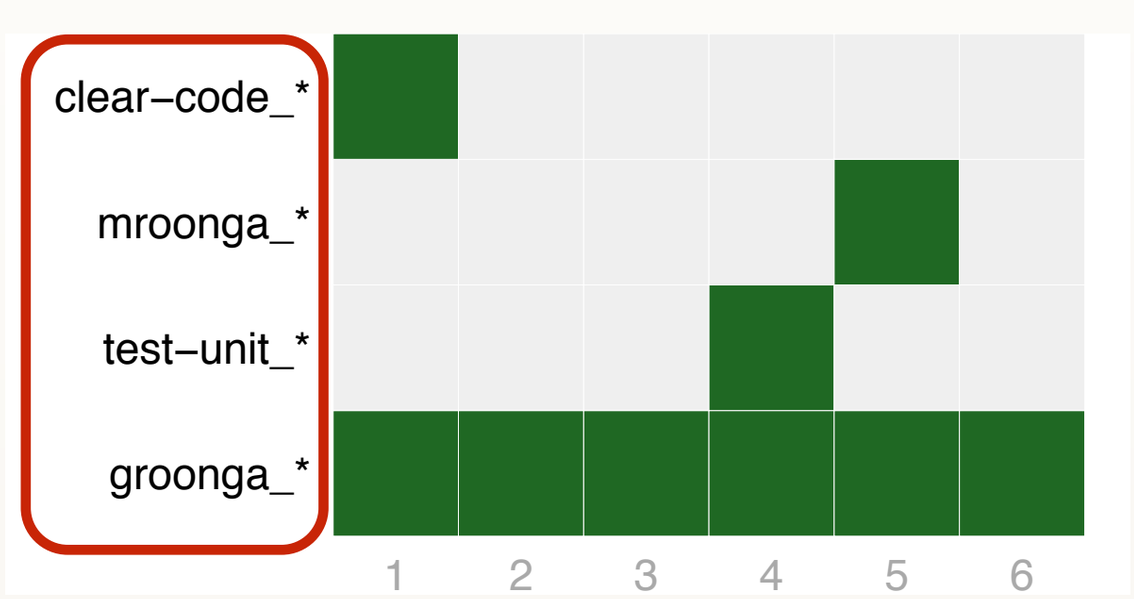
3. DAY-TO-DAY FOCUS

Repetitive day-to-day working style

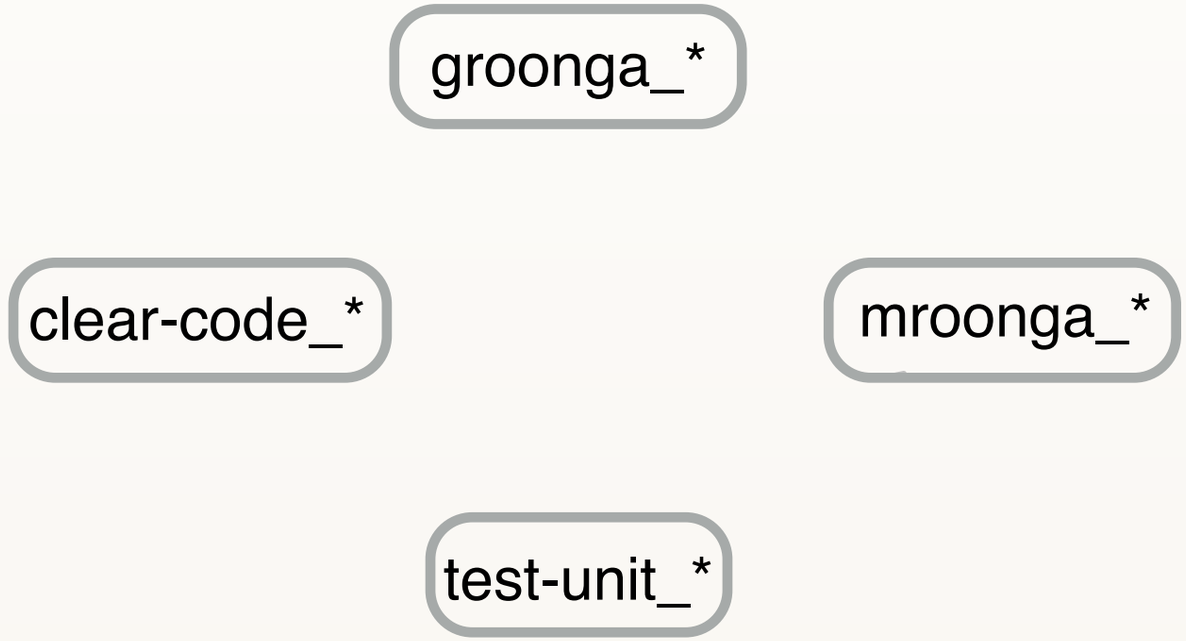
vs.

Changing focus one day to next

Focus shifting networks



nodes



MULTITASKING DIMENSIONS

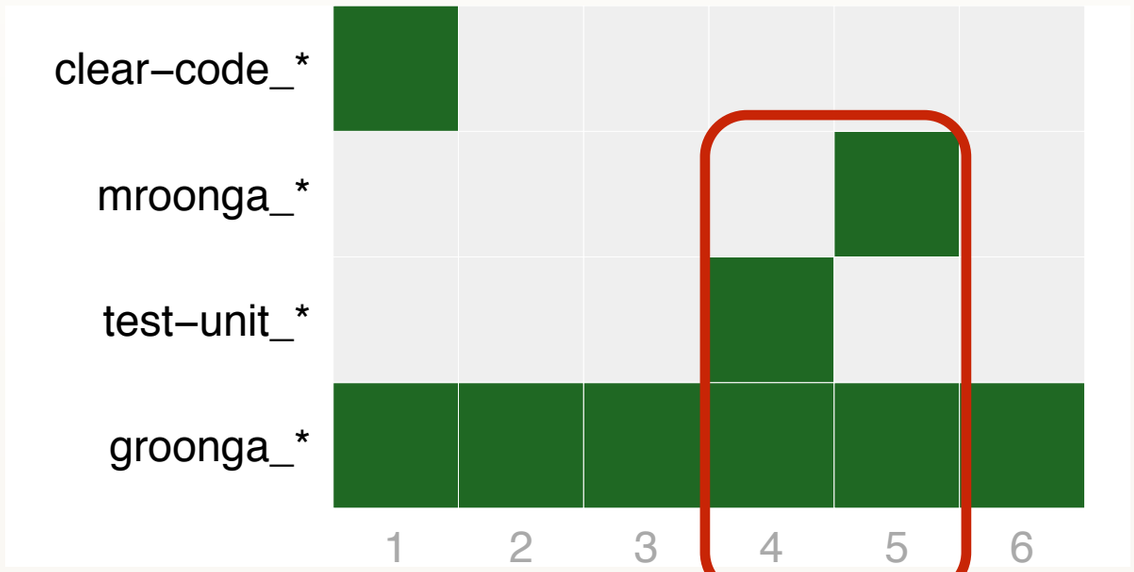
3. DAY-TO-DAY FOCUS

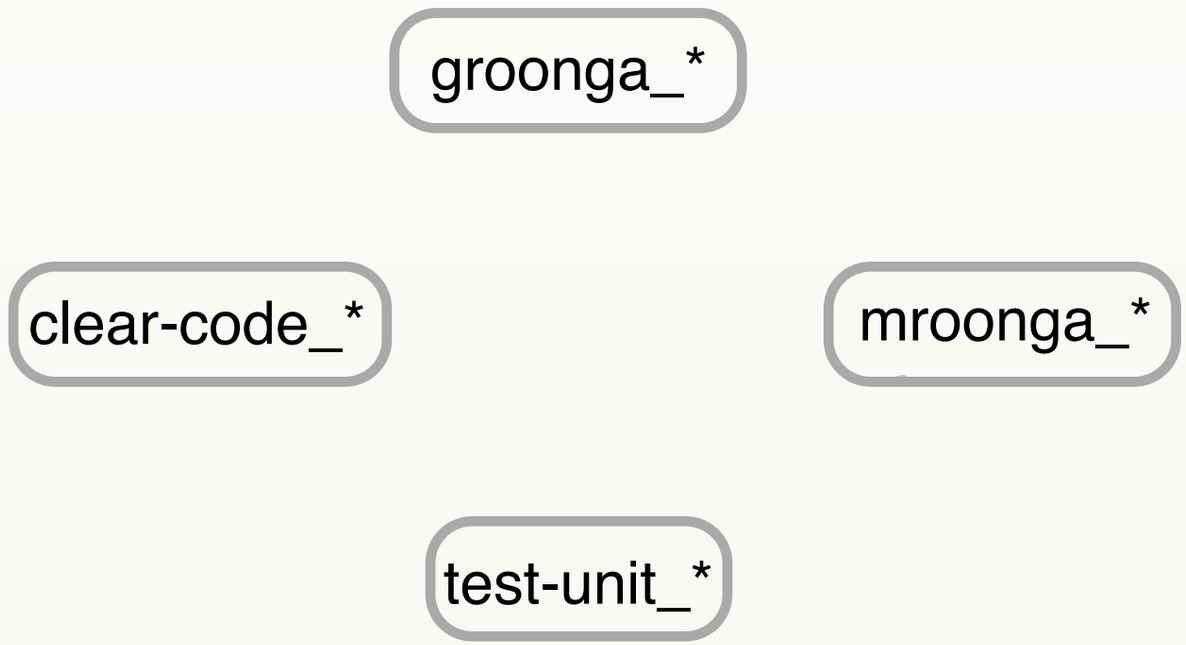
Repetitive day-to-day working style

vs.

Changing focus one day to next

Focus shifting networks





MULTITASKING DIMENSIONS

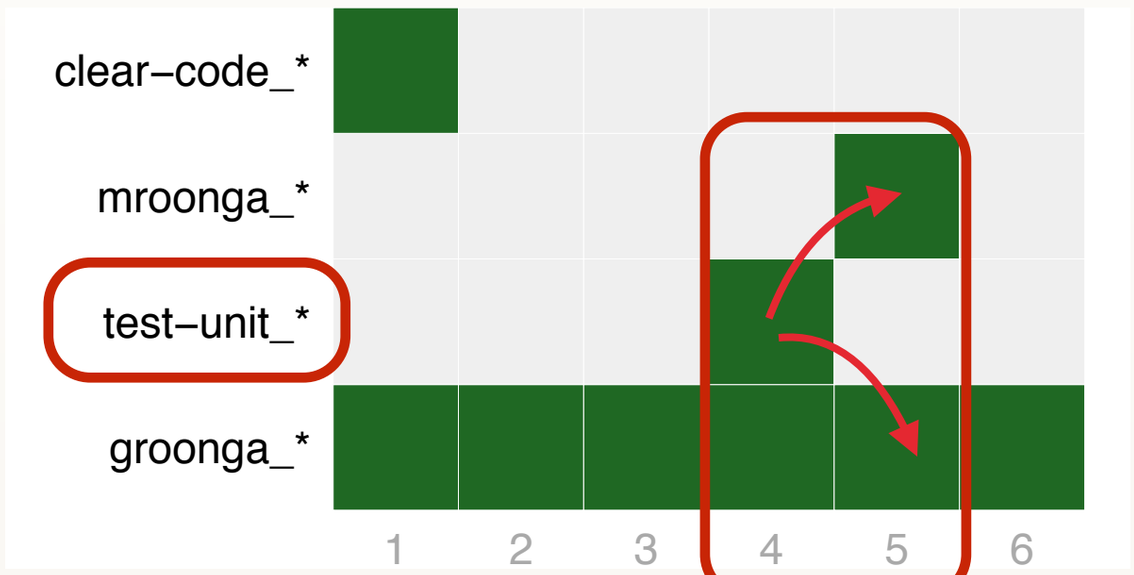
3. DAY-TO-DAY FOCUS

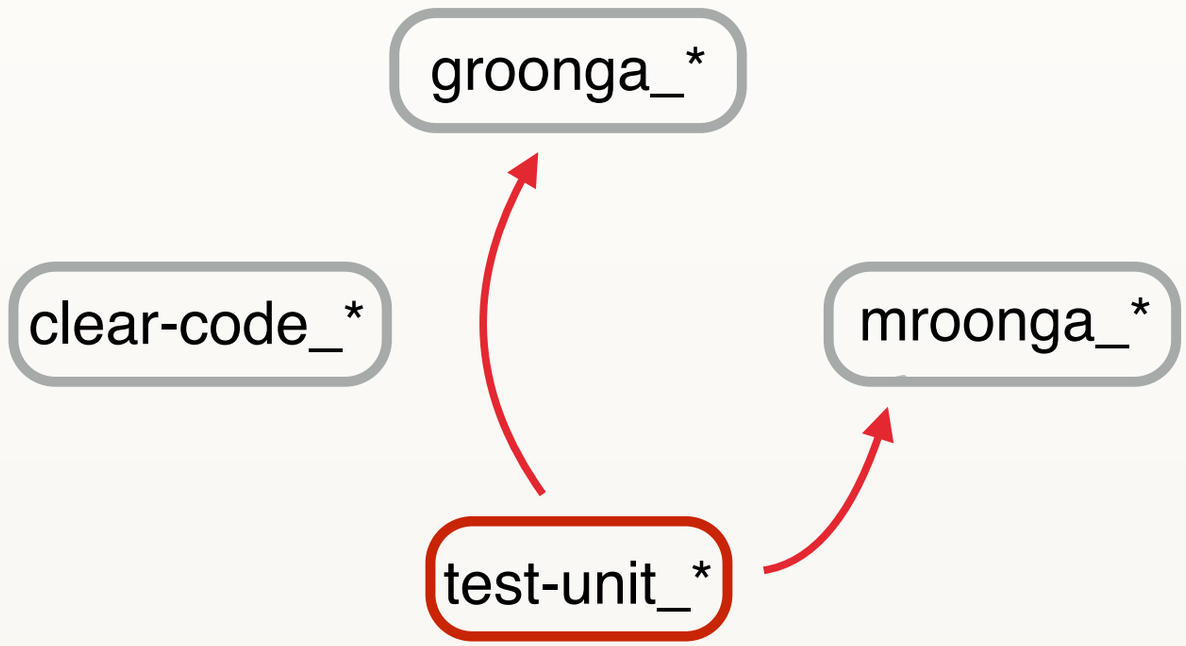
Repetitive day-to-day working style

vs.

Changing focus one day to next

Focus shifting networks





MULTITASKING DIMENSIONS

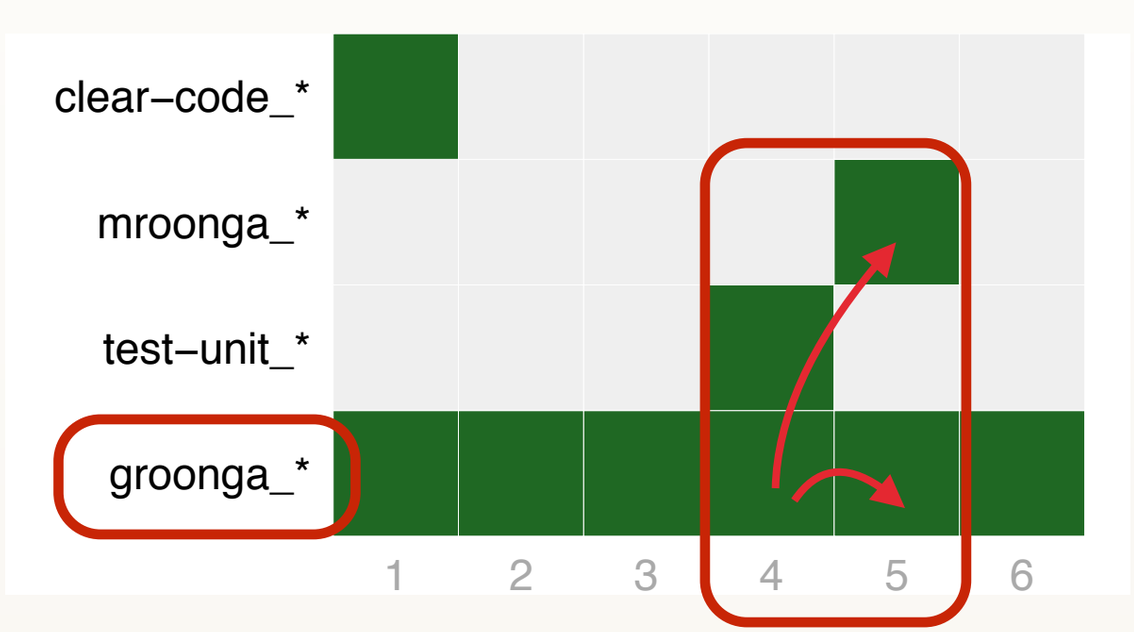
3. DAY-TO-DAY FOCUS

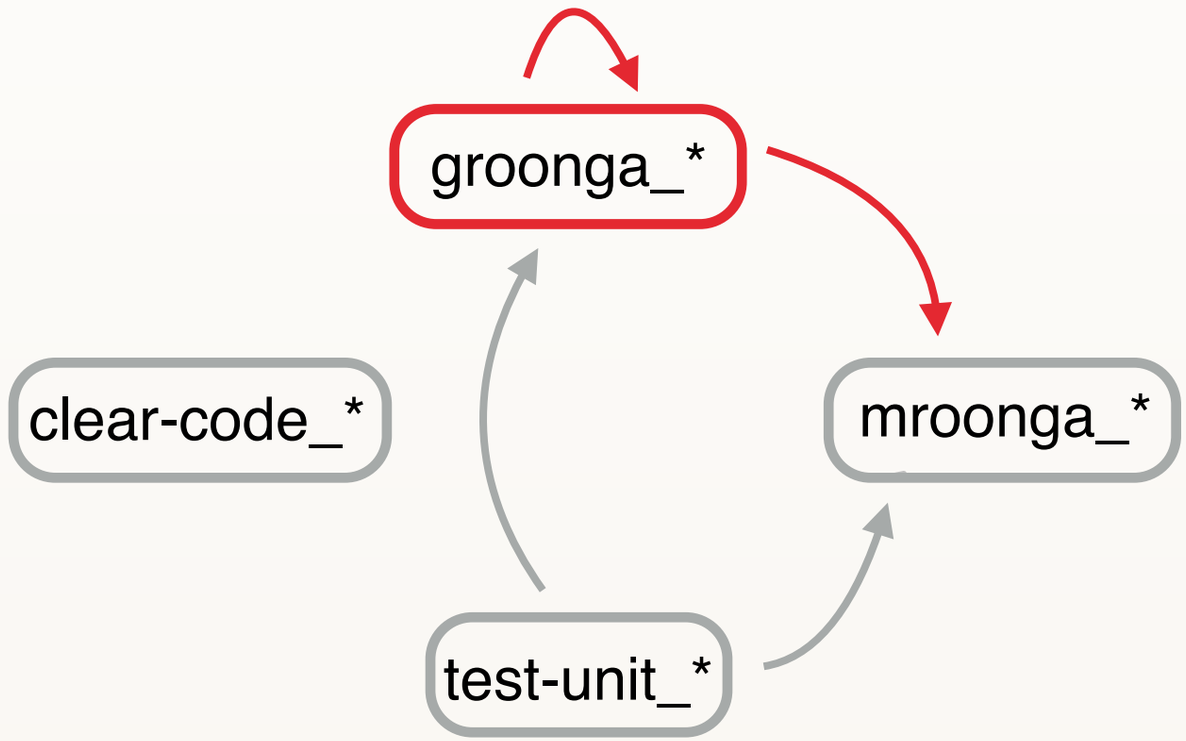
Repetitive day-to-day working style

vs.

Changing focus one day to next

Focus shifting networks





MULTITASKING DIMENSIONS

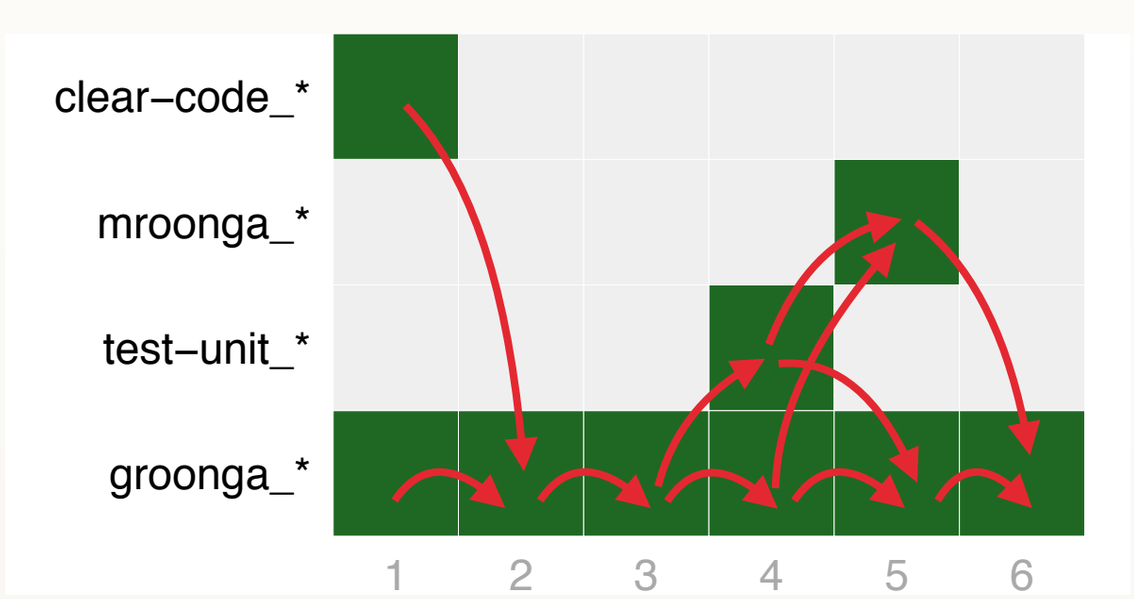
3. DAY-TO-DAY FOCUS

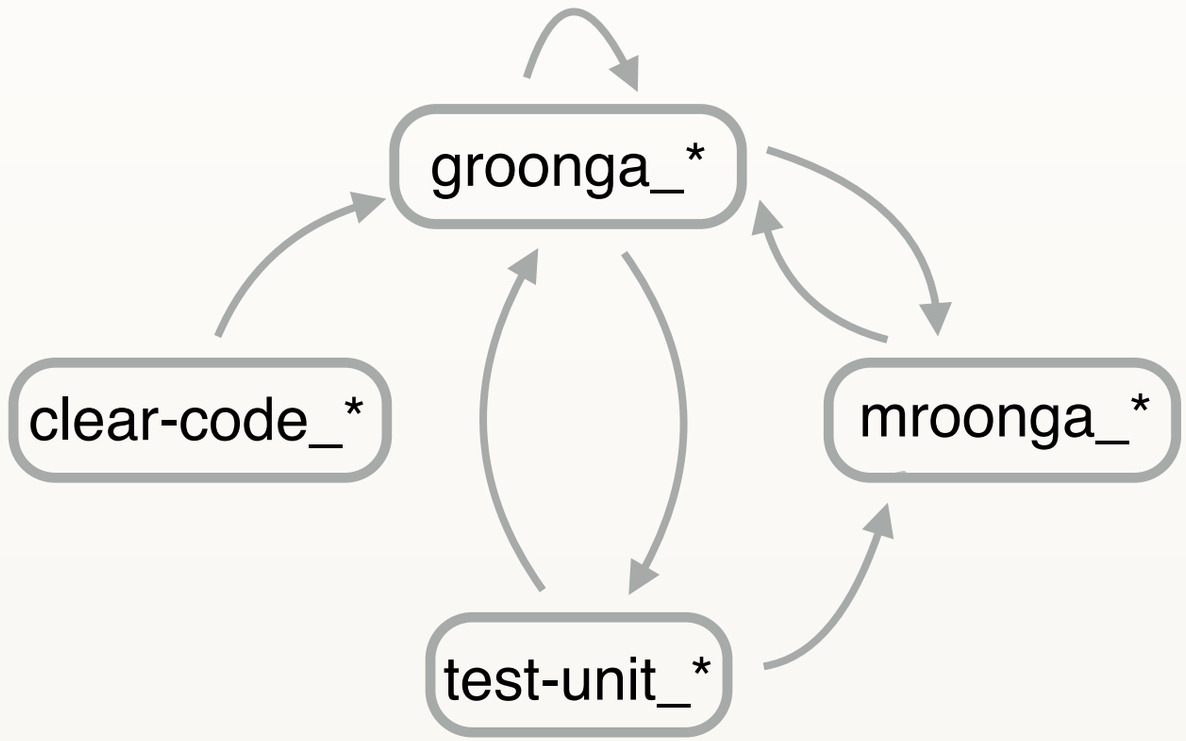
Repetitive day-to-day working style

vs.

Changing focus one day to next

Focus shifting networks





MULTITASKING DIMENSIONS

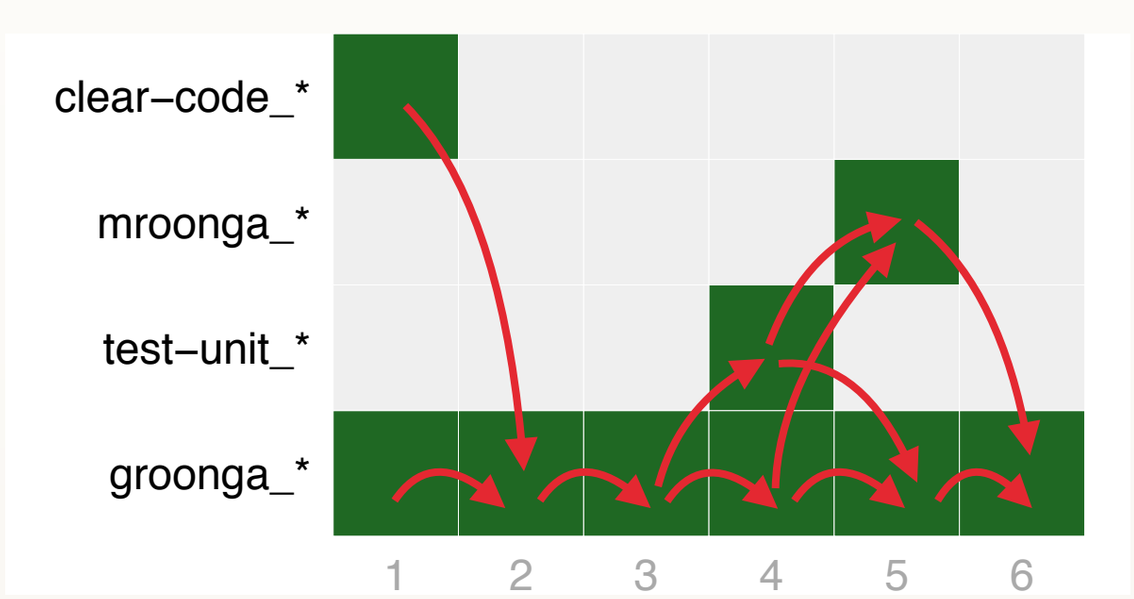
3. DAY-TO-DAY FOCUS

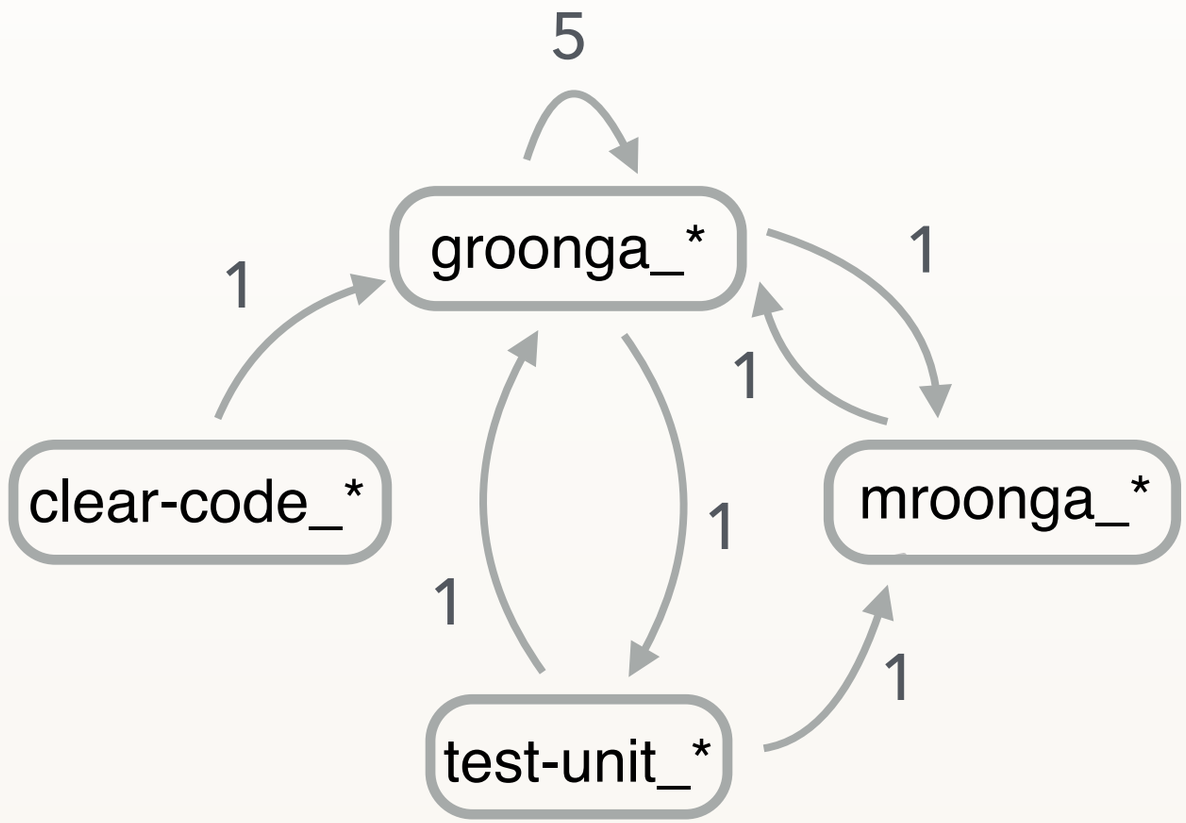
Repetitive day-to-day working style

vs.

Changing focus one day to next

Focus shifting networks





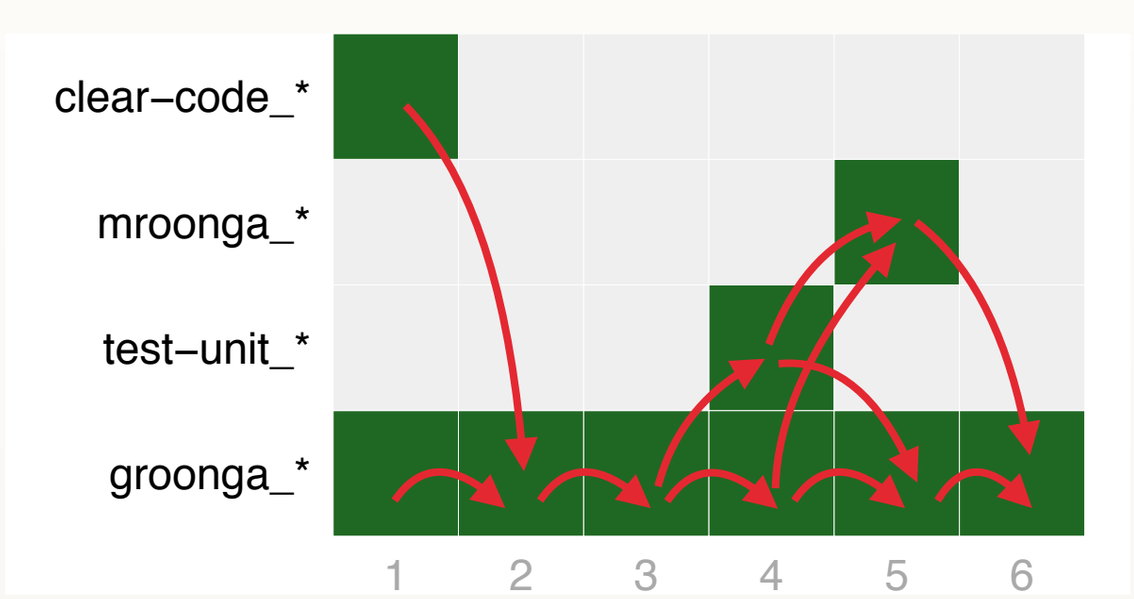
MULTITASKING DIMENSIONS

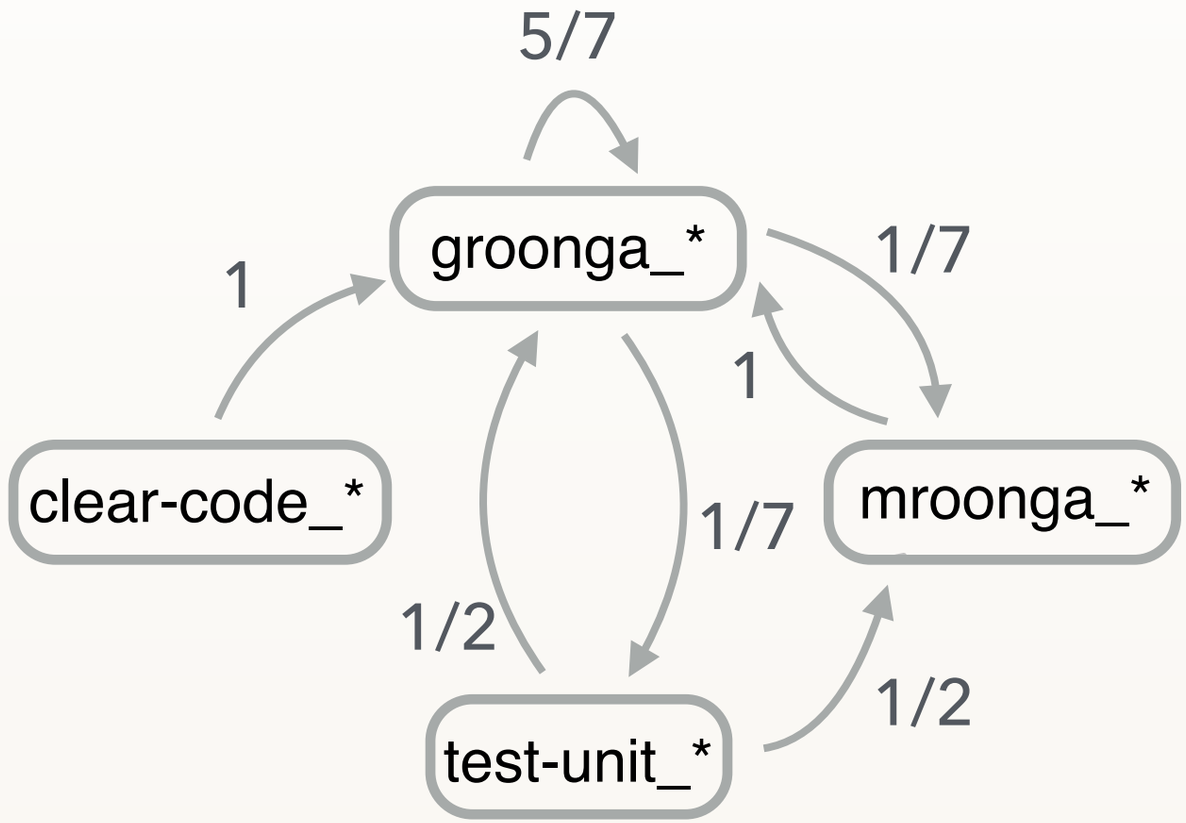
Repetitive day-to-day working style

vs.

Changing focus one day to next

Focus shifting networks





MULTITASKING DIMENSIONS

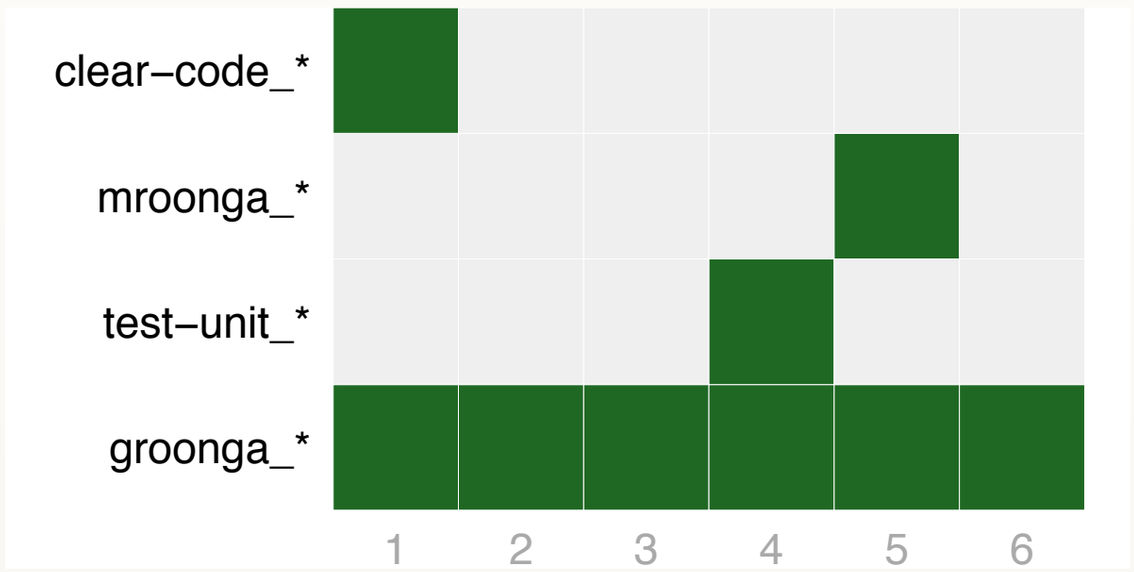
3. DAY-TO-DAY FOCUS

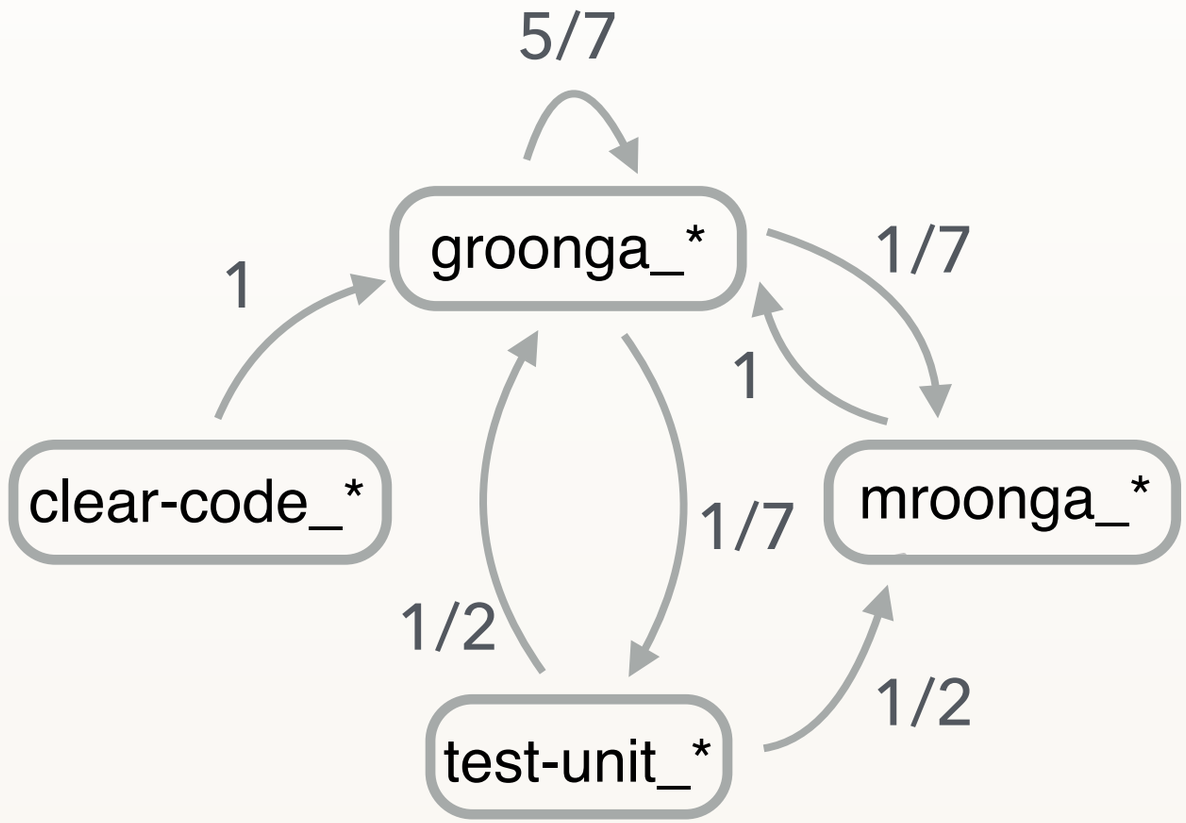
Repetitive day-to-day working style

vs.

Changing focus one day to next

Focus shifting networks





$$S_{\mathrm{Switch}} = -\sum_{i=1}^{N}\left[p_i \sum_{j\in\pi_i} p(j|i)\log_2 p(j|i)\right]$$

Markov entropy

MULTITASKING DIMENSIONS

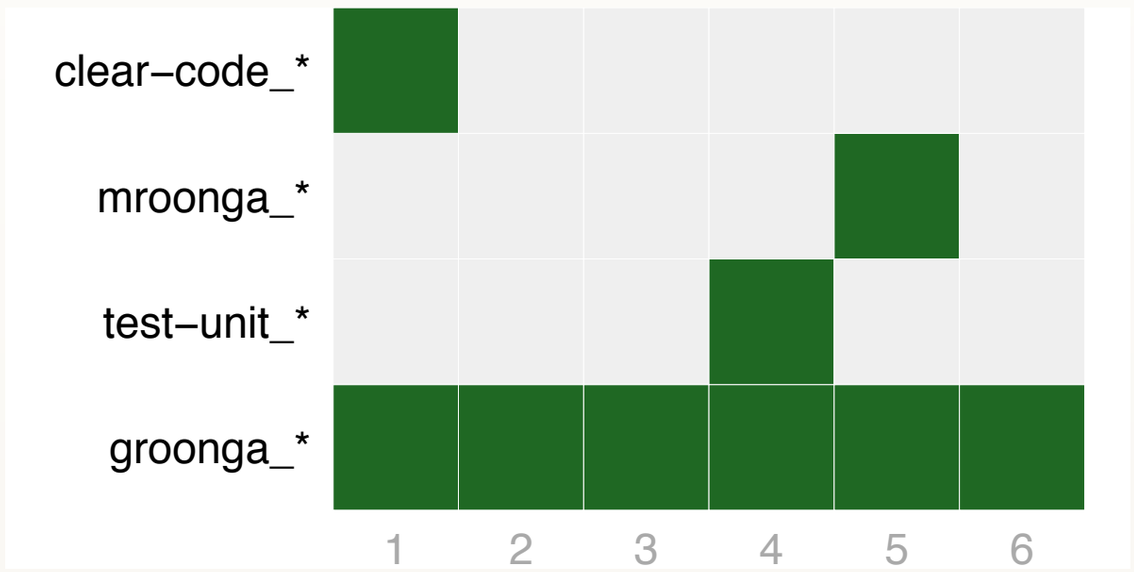
3. DAY-TO-DAY FOCUS

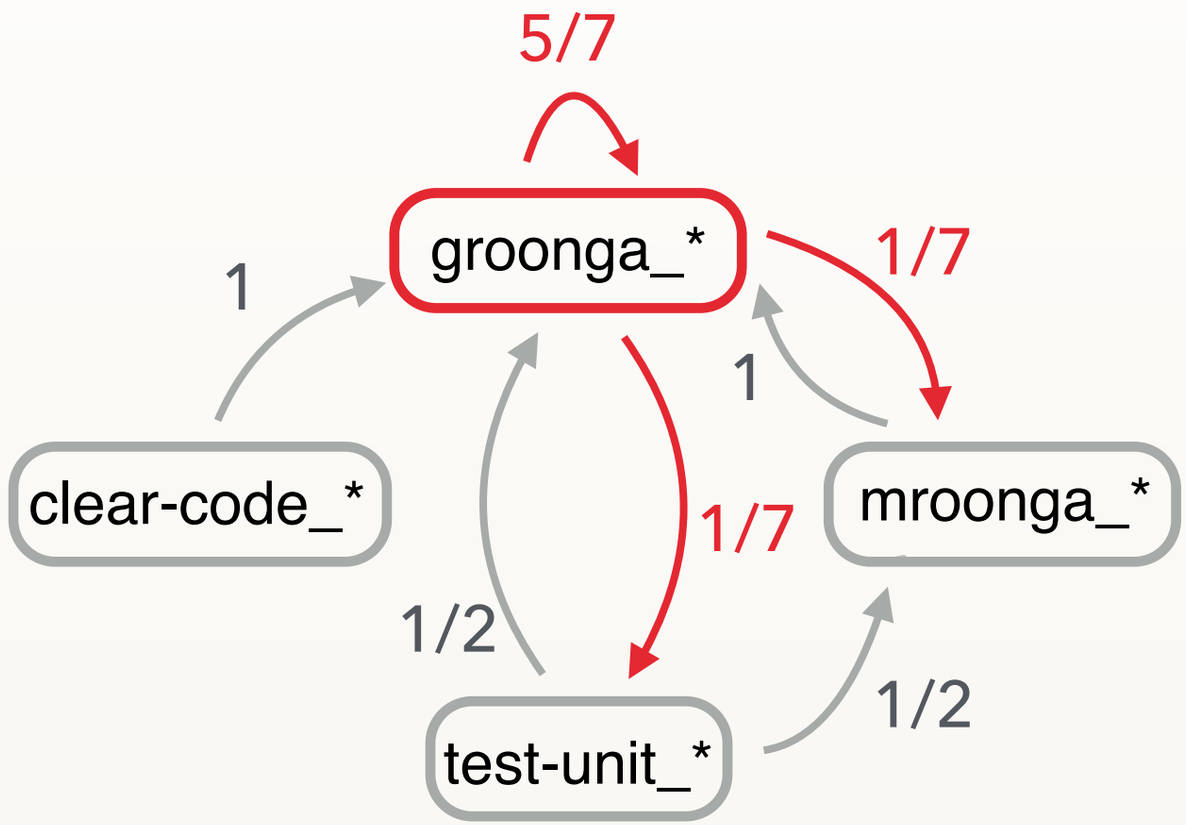
Repetitive day-to-day working style

vs.

Changing focus one day to next

Focus shifting networks





$$S_{\mathrm{Switch}} = -\sum_{i=1}^{N}\left[p_i \sum_{j\in\pi_i} p(j|i)\log_2 p(j|i)\right]$$

How predictable is my behavior tomorrow if today I work on project X?

MULTITASKING DIMENSIONS

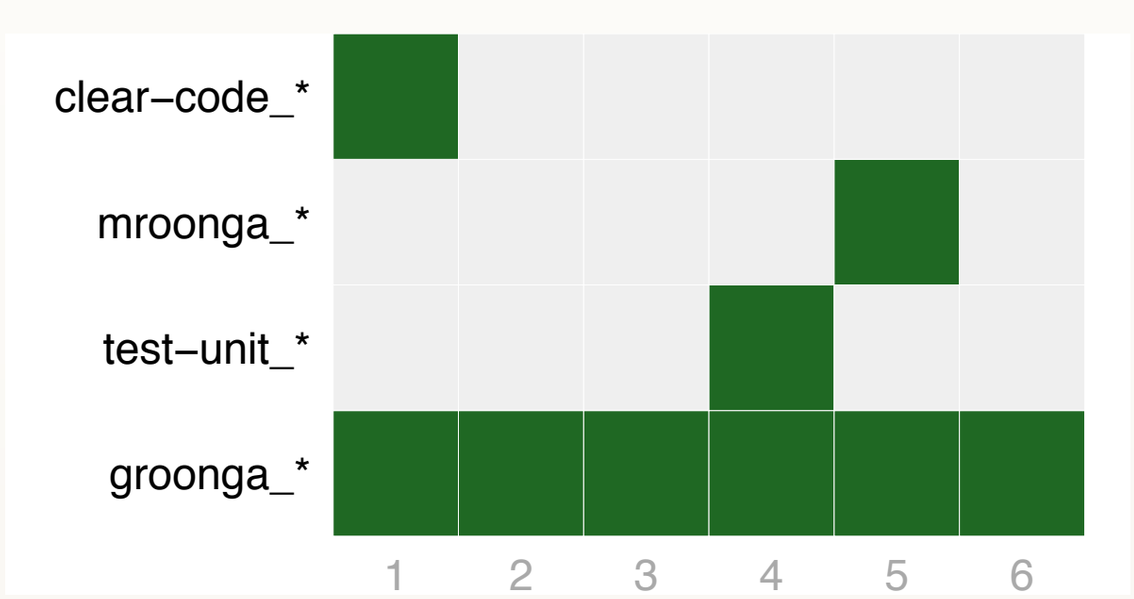
3. DAY-TO-DAY FOCUS

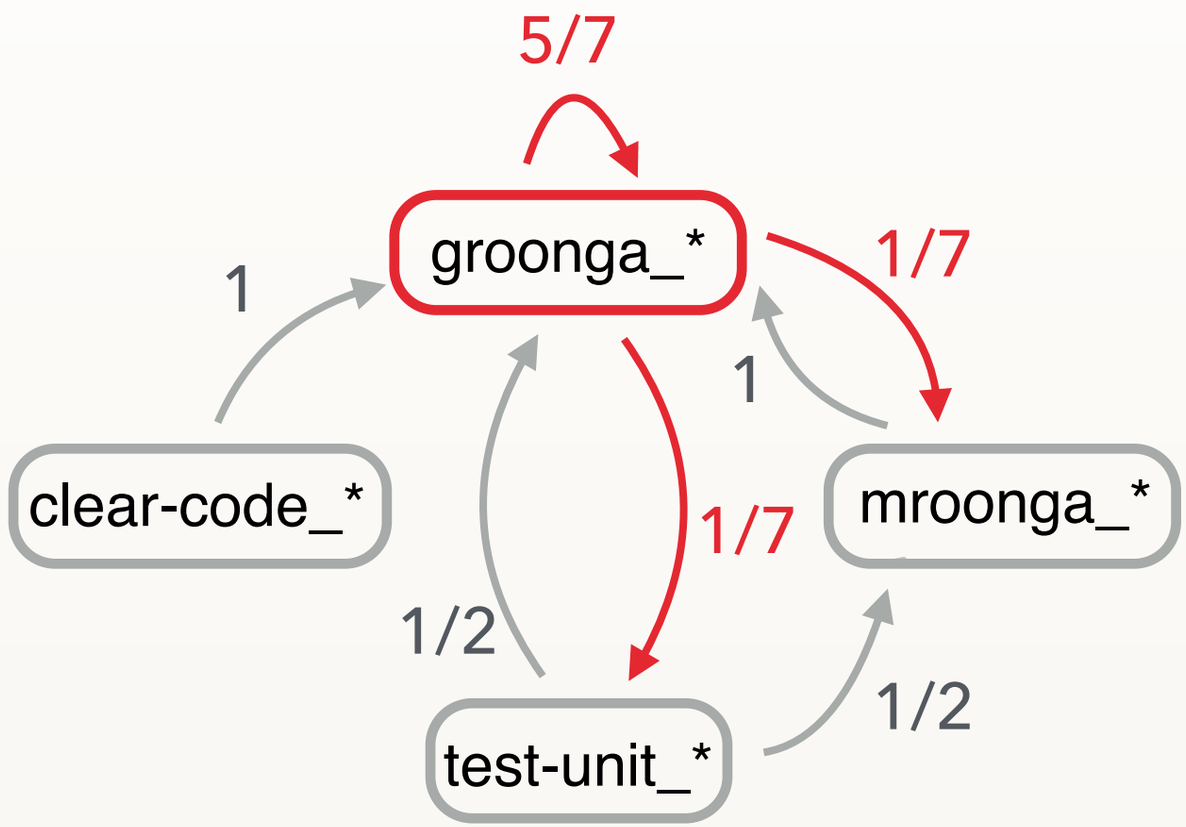
Repetitive day-to-day working style

vs.

Changing focus one day to next

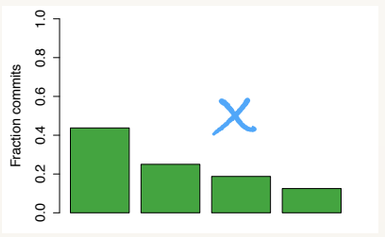
Focus shifting networks





$$S_{\text{Switch}} = -\sum_{i=1}^{N} \left[p_i \sum_{j \in \pi_i} p(j|i) \log_2 p(j|i) \right]$$

How important is project X relative to my other projects?



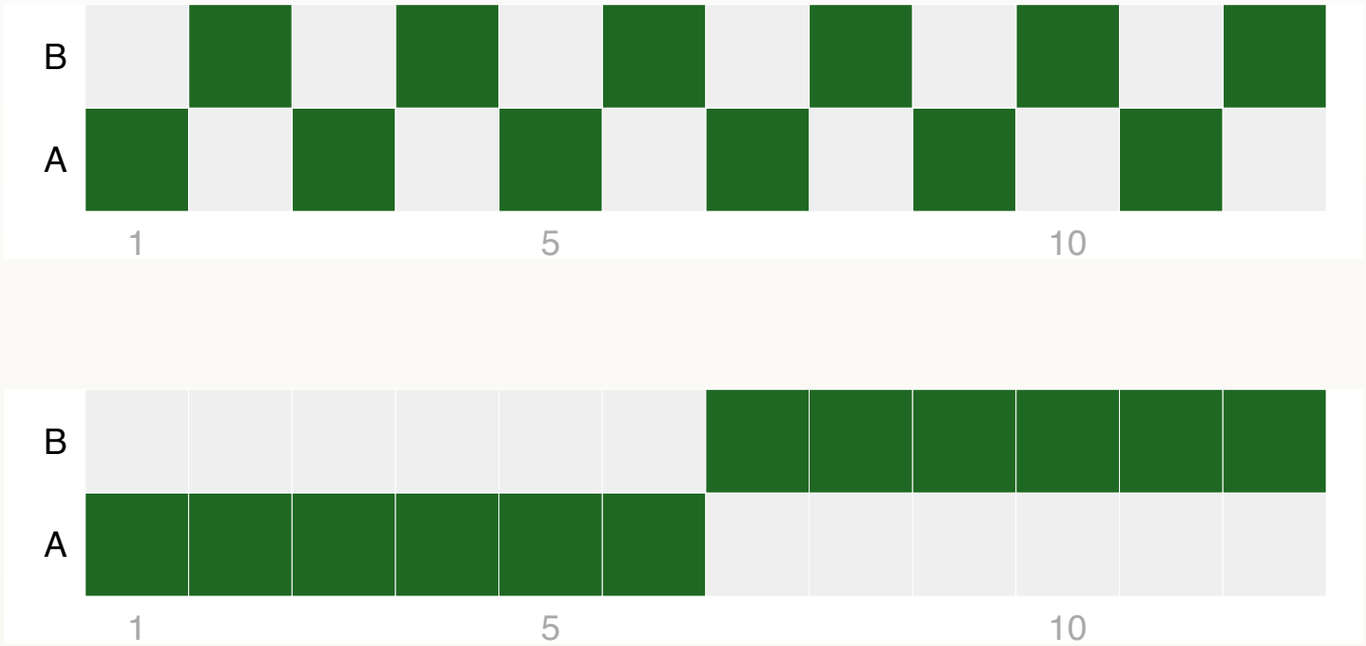
MULTITASKING DIMENSIONS

3. DAY-TO-DAY FOCUS

Repetitive day-to-day working style

vs.

Changing focus one day to next



$S_{\text{Switch}} = 0$

Less repetitive day-to-day

$S_{\text{Switch}} = 0.325$

More repetitive day-to-day

MULTITASKING DIMENSIONS

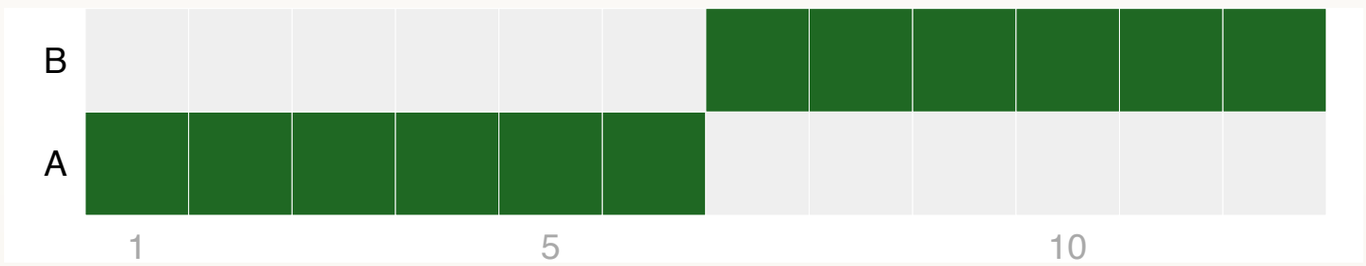
3. DAY-TO-DAY FOCUS

Repetitive day-to-day working style

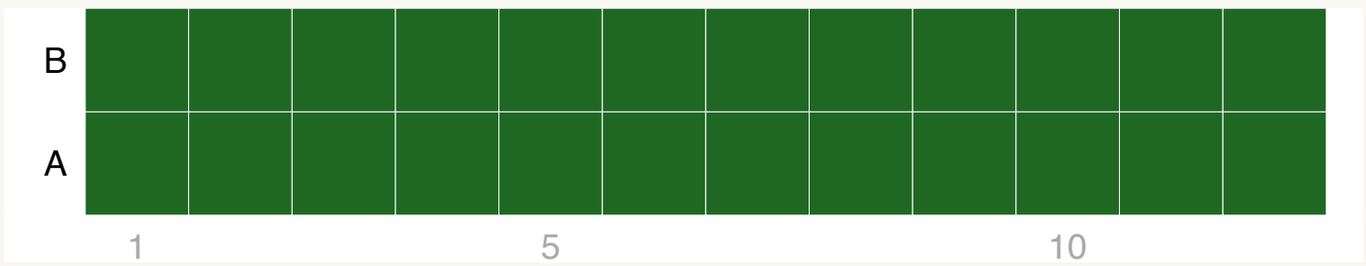
vs.

Changing focus one day to next

$S_{\text{Switch}} = 0$



$S_{\text{Switch}} = 0.325$



$S_{\text{Switch}} = 2$

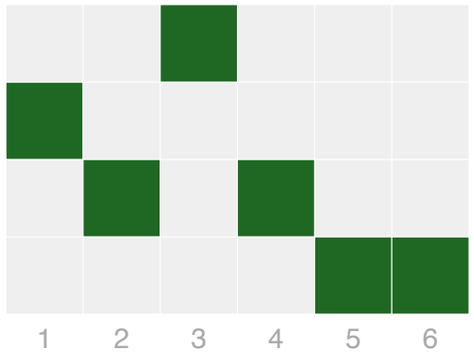
Less repetitive day-to-day

More repetitive day-to-day

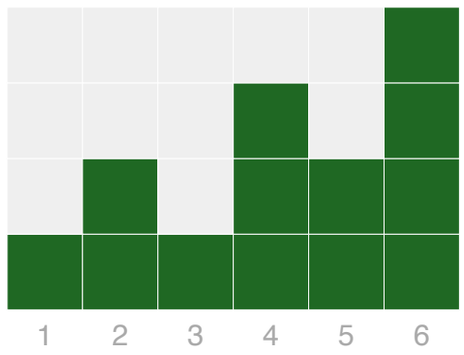
RESULTS – LINEAR MIXED EFFECTS REGRESSION

Higher productivity

Projects per day



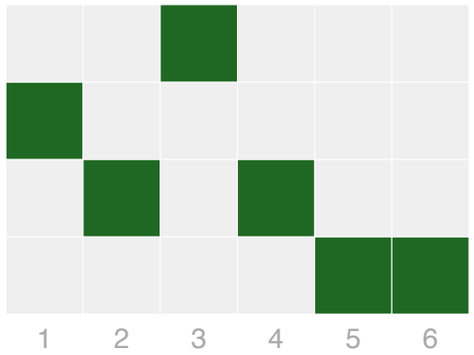
vs.



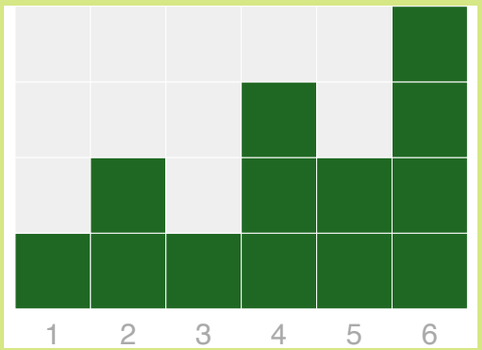
RESULTS – LINEAR MIXED EFFECTS REGRESSION

Higher productivity

Projects per day



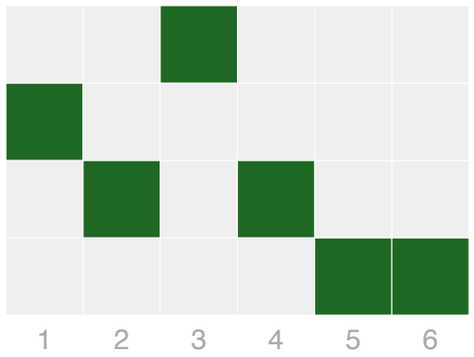
vs.



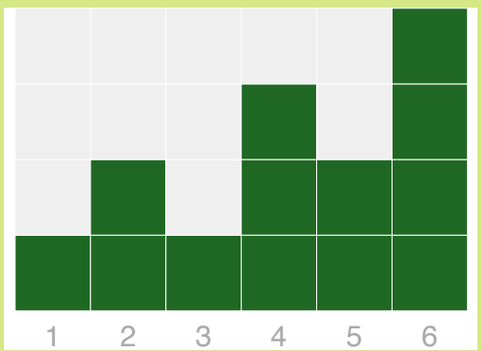
RESULTS – LINEAR MIXED EFFECTS REGRESSION

Higher productivity

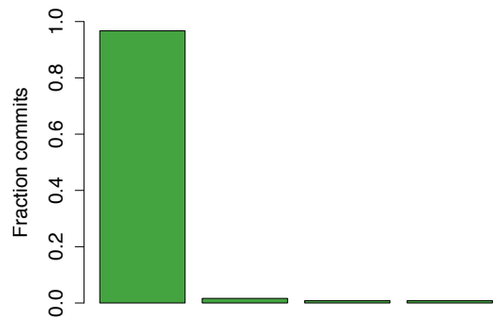
Projects per day



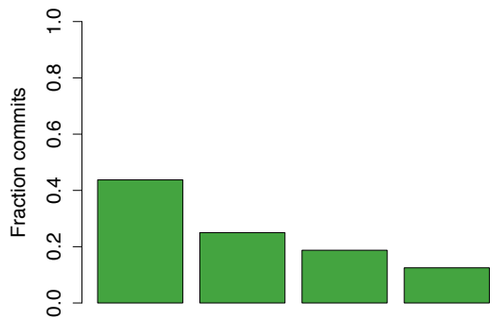
vs.



Weekly focus



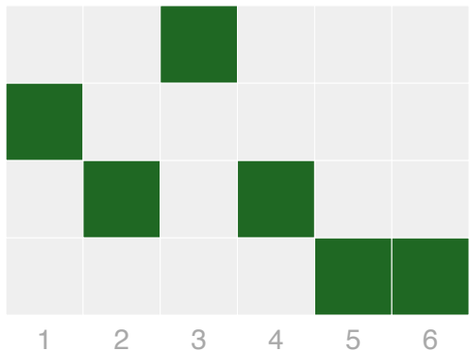
vs.



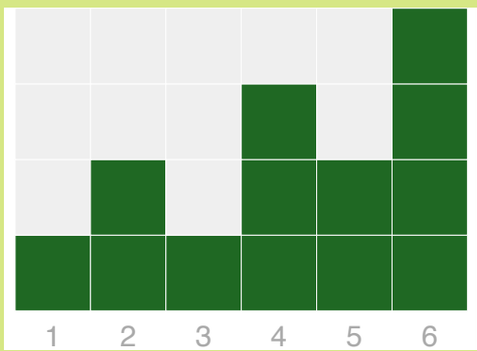
RESULTS – LINEAR MIXED EFFECTS REGRESSION

Higher productivity

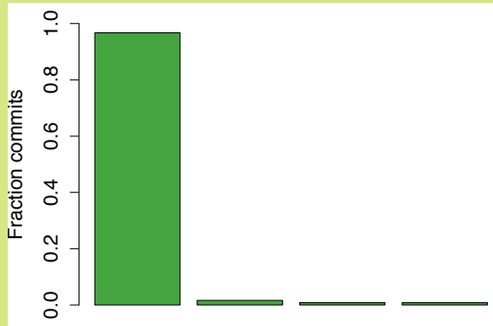
Projects per day



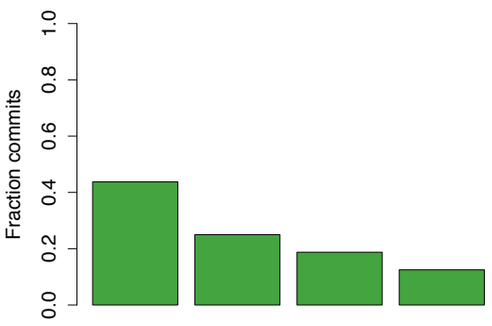
vs.



Weekly focus



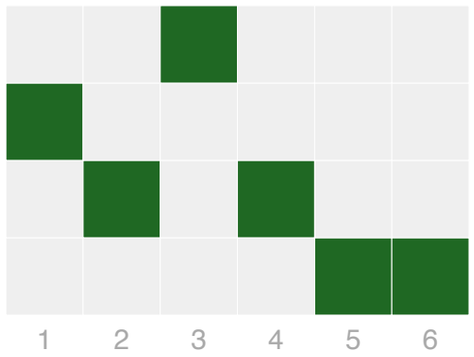
vs.



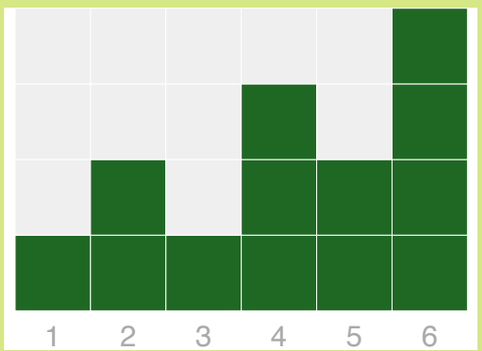
RESULTS – LINEAR MIXED EFFECTS REGRESSION

Higher productivity

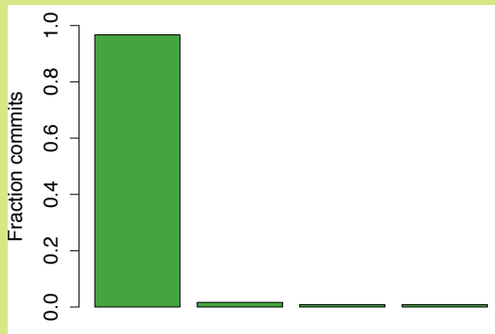
Projects per day



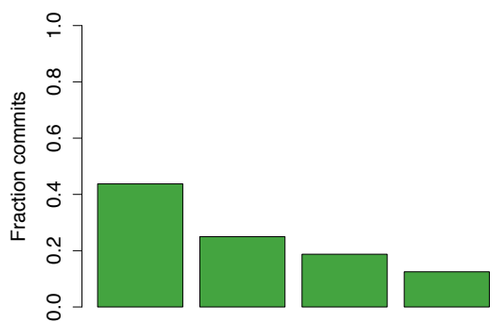
vs.



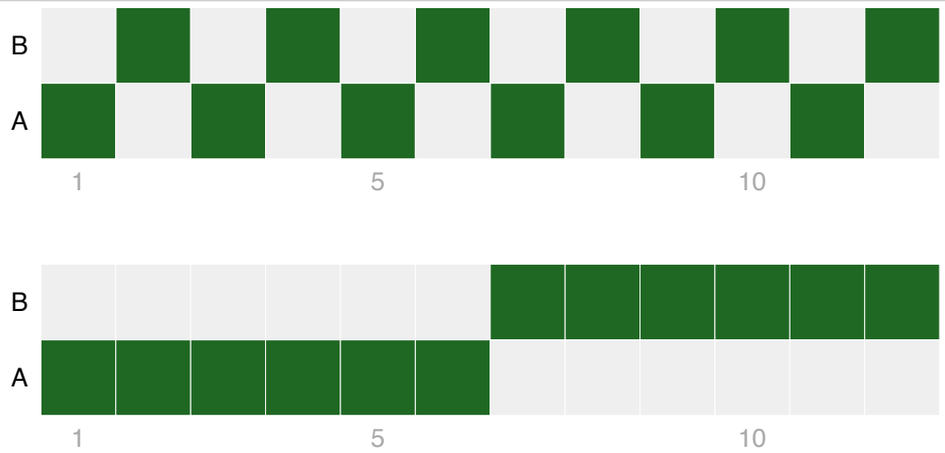
Weekly focus



vs.



Day-to-day focus (repeatability)

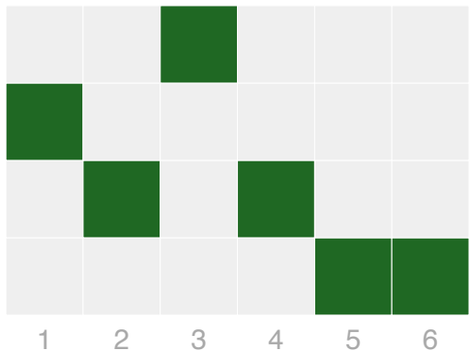


vs.

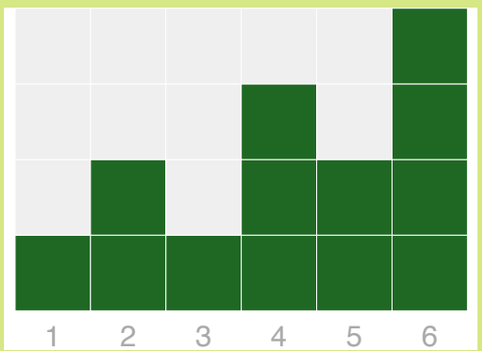
RESULTS – LINEAR MIXED EFFECTS REGRESSION

Higher productivity

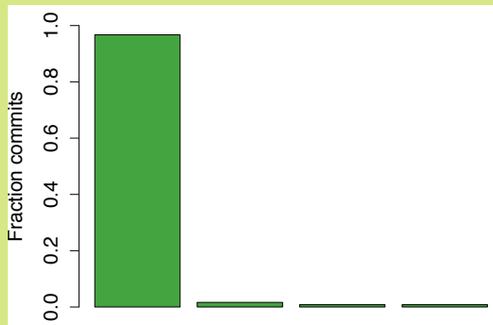
Projects per day



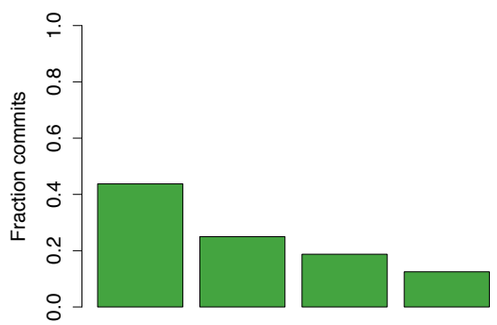
vs.



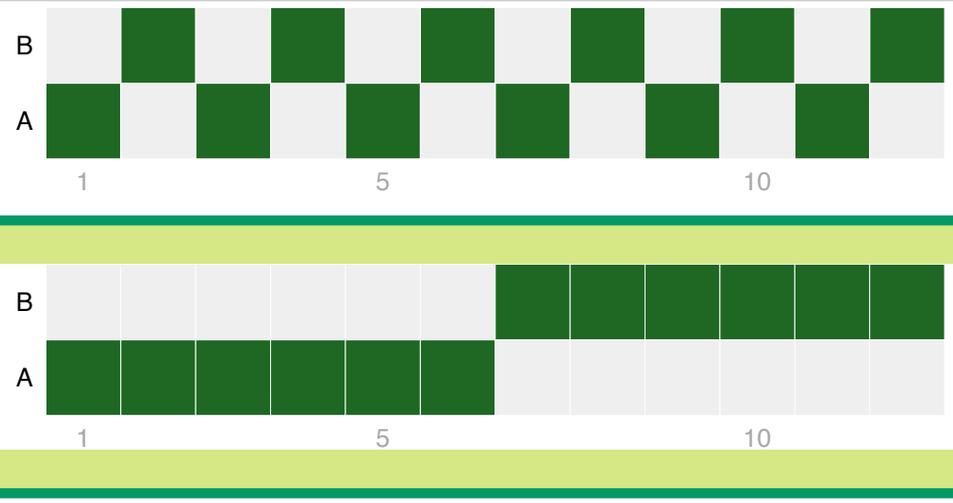
Weekly focus



vs.



Day-to-day focus (repeatability)

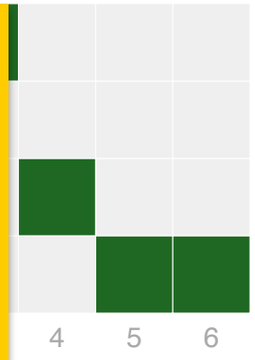


vs.

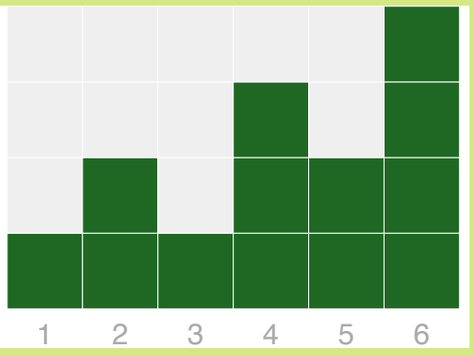
RESULTS – LINEAR MIXED EFFECTS REGRESSION

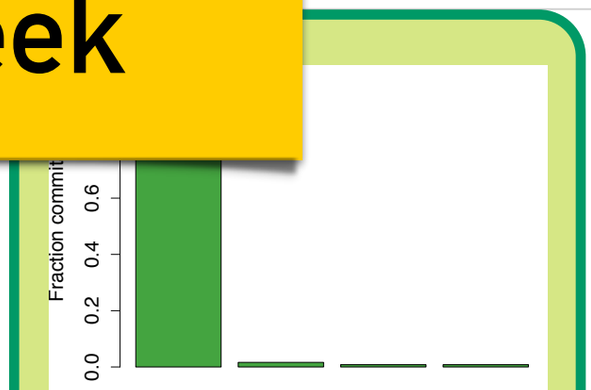
Higher productivity

Interaction effects:
No scheduling is productive beyond 5 projects/week

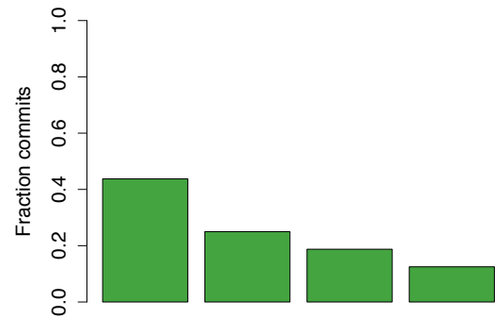


vs.



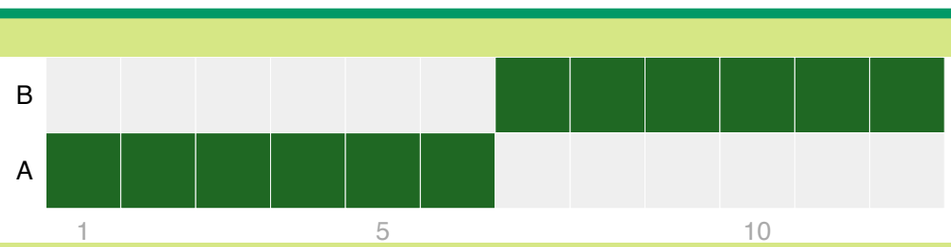


vs.

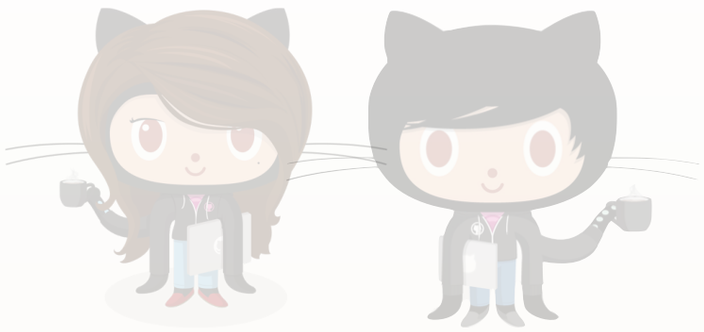


Day-to-day focus (repeatability)

vs.



TODAY



1

TEAM DIVERSITY

[CHI 2015]

2

MULTITASKING ACROSS PROJECTS

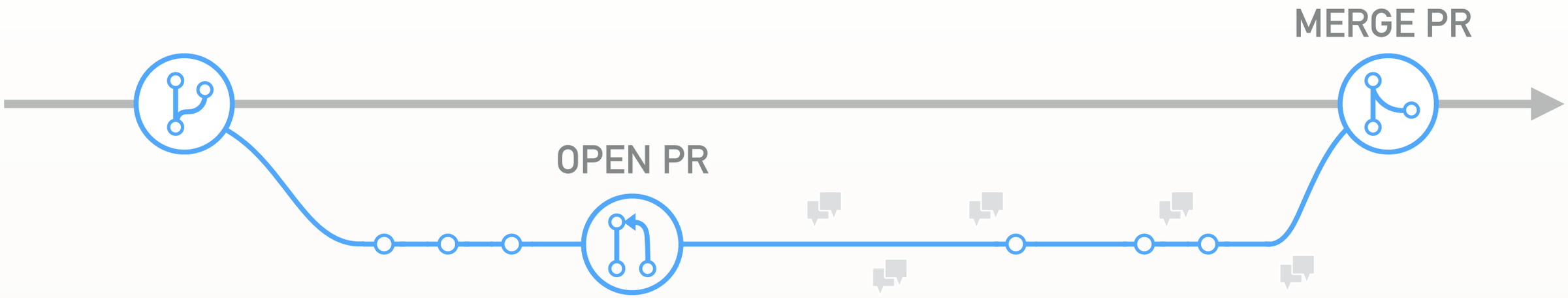
[ICSE 2016]

3

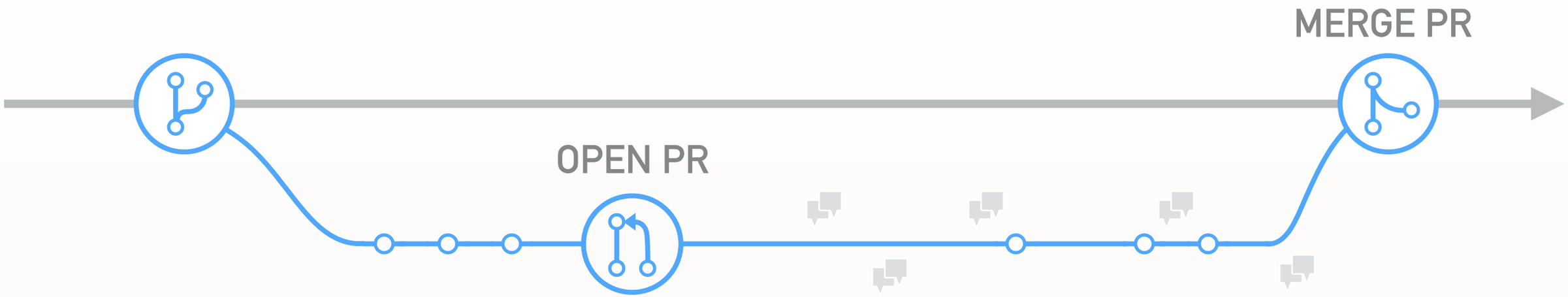
CONTINUOUS INTEGRATION

[ESEC/FSE 2015]

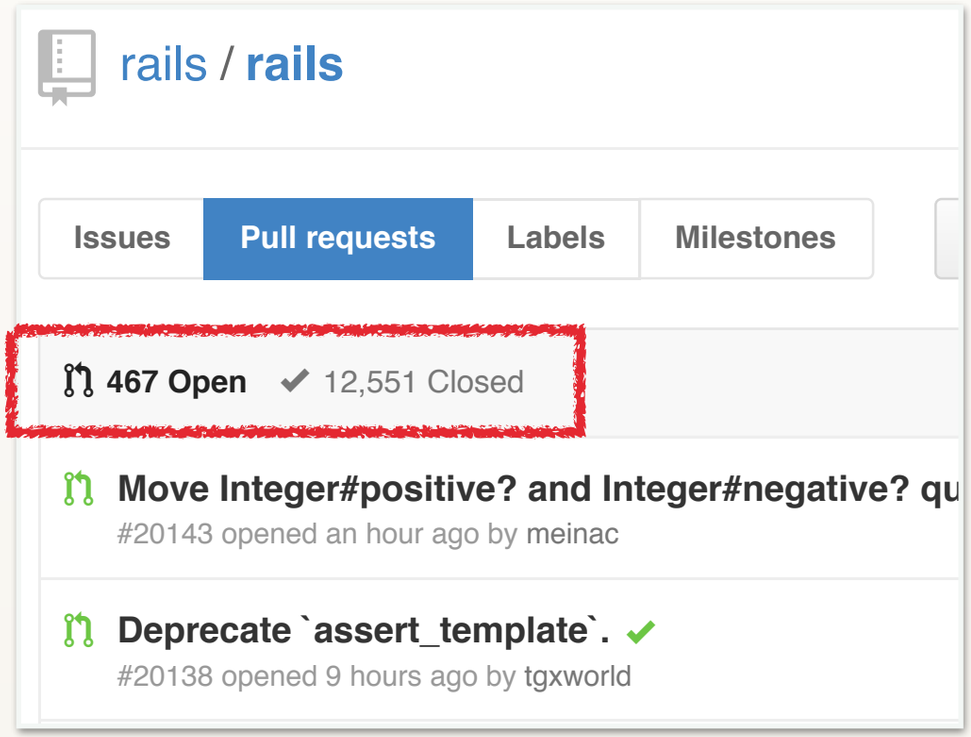
PULL REQUESTS REQUIRE REVIEW



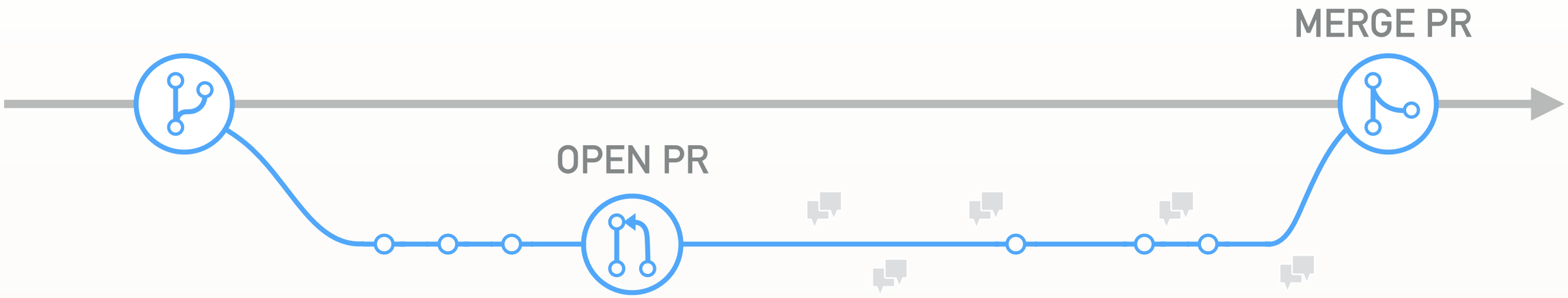
PULL REQUESTS REQUIRE REVIEW



Ruby on Rails



PULL REQUESTS REQUIRE REVIEW



Ruby on Rails

rails / **rails**

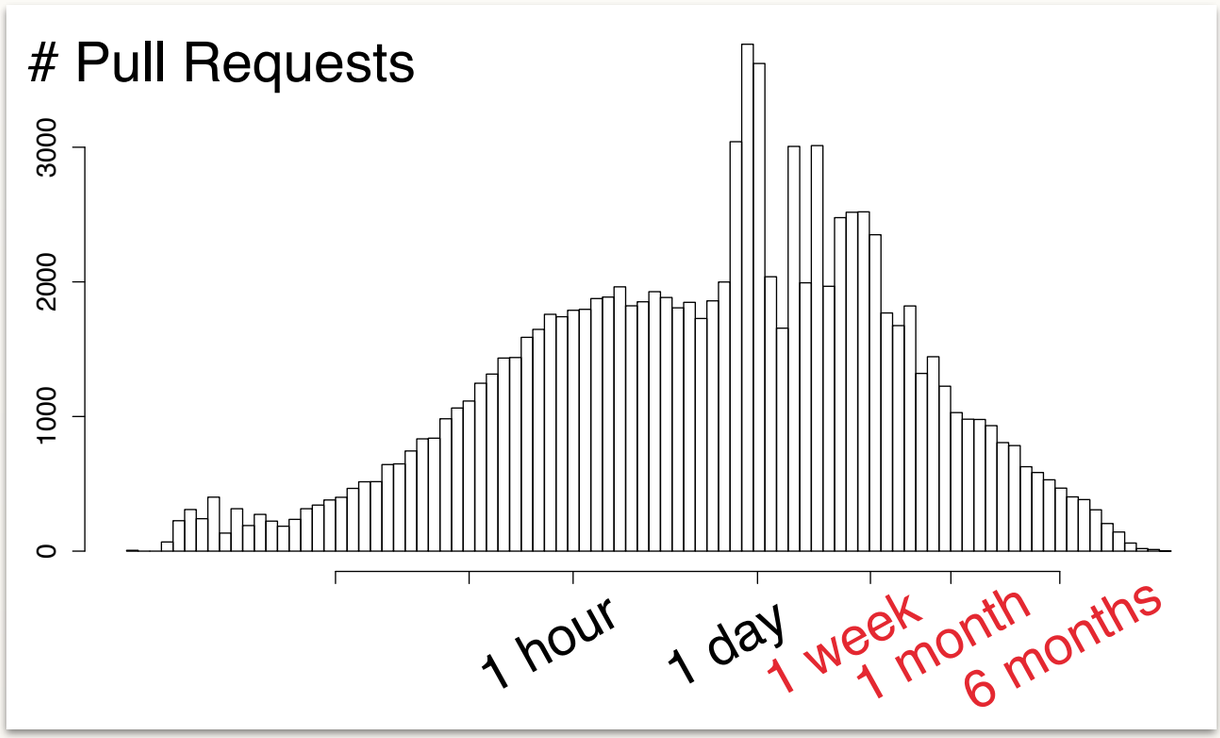
Issues | **Pull requests** | Labels | Milestones

467 Open ✔ 12,551 Closed

Move Integer#positive? and Integer#negative? qu
#20143 opened an hour ago by meinac

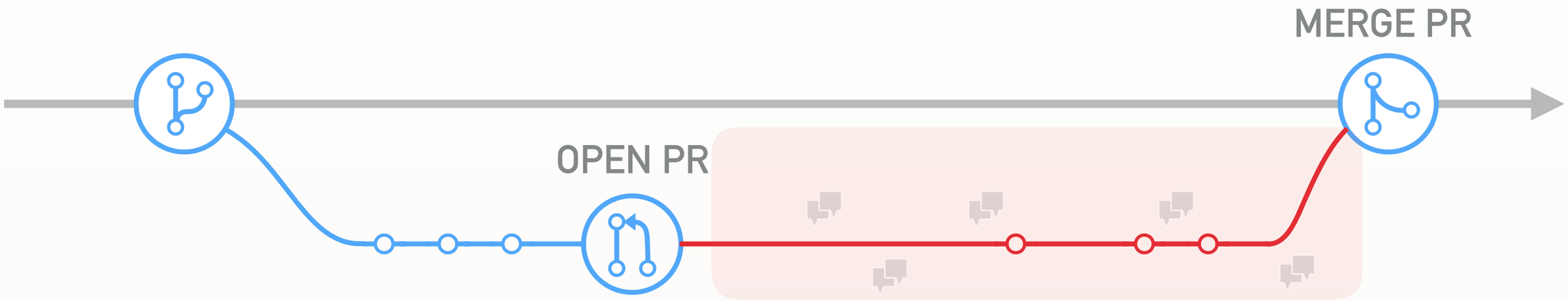
Deprecate `assert_template`. ✔
#20138 opened 9 hours ago by tgxworld

Large GitHub sample



[MSR 2015]

PROCESS AUTOMATION



Is it good? Should I merge? ✗ / ✓

Ruby on Rails

rails / **rails**

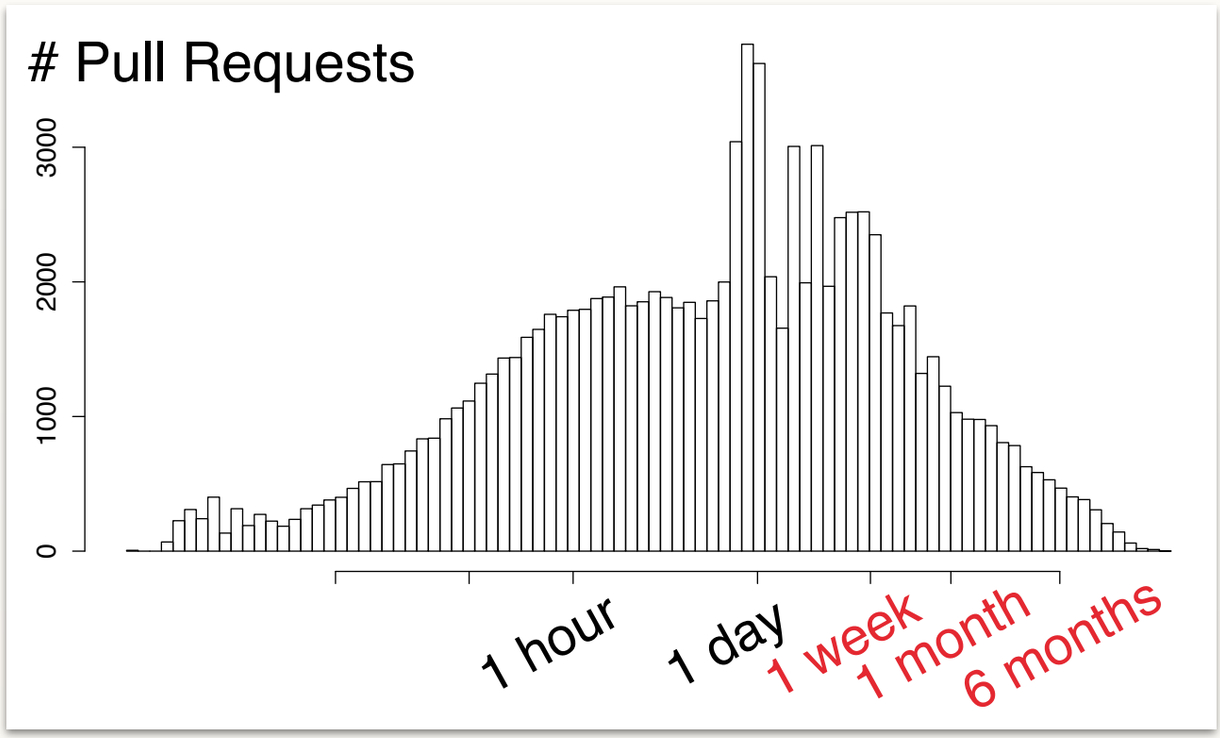
Issues | **Pull requests** | Labels | Milestones

467 Open ✔ 12,551 Closed

Move Integer#positive? and Integer#negative? qu
#20143 opened an hour ago by meinac

Deprecate `assert_template`. ✔
#20138 opened 9 hours ago by tgxworld

Large GitHub sample

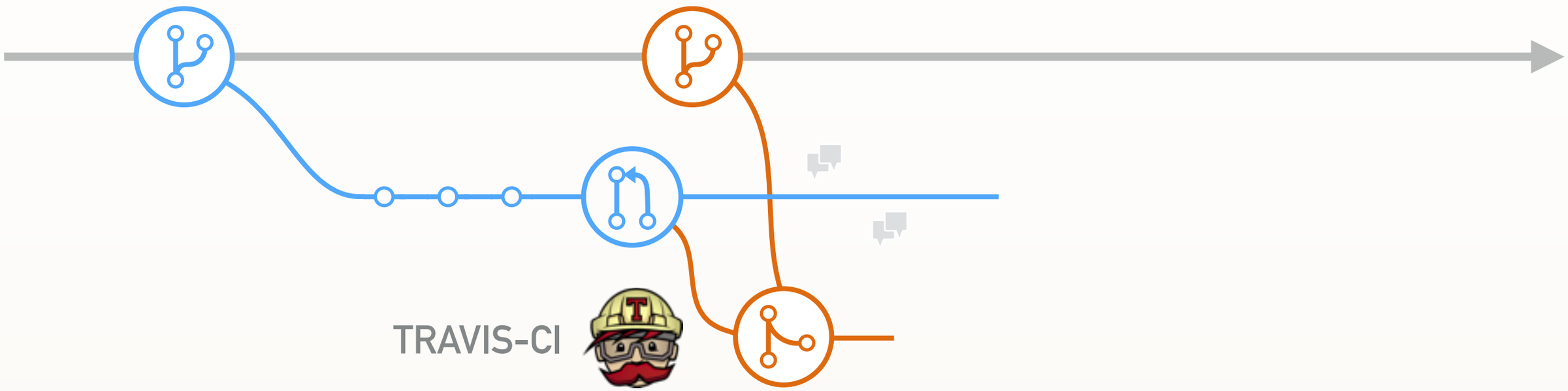


[MSR 2015]

CI PULL REQUEST PROCESS

[ICSME 2014]

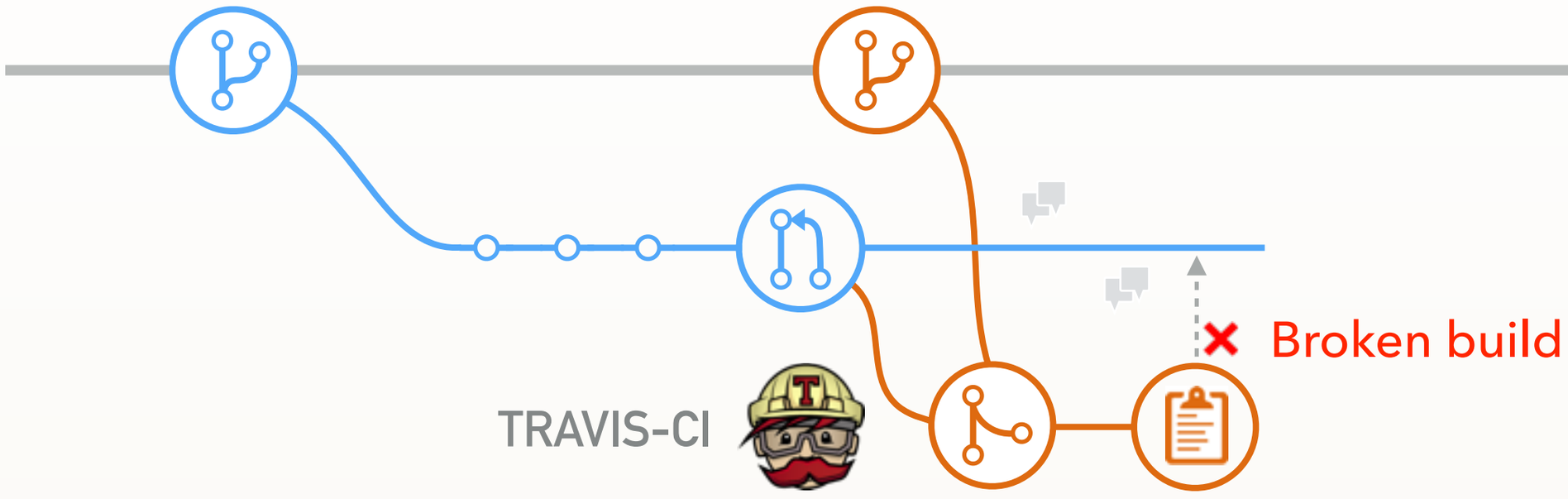
CI PULL REQUEST PROCESS



PR auto
merged into
testing branch

[ICSME 2014]

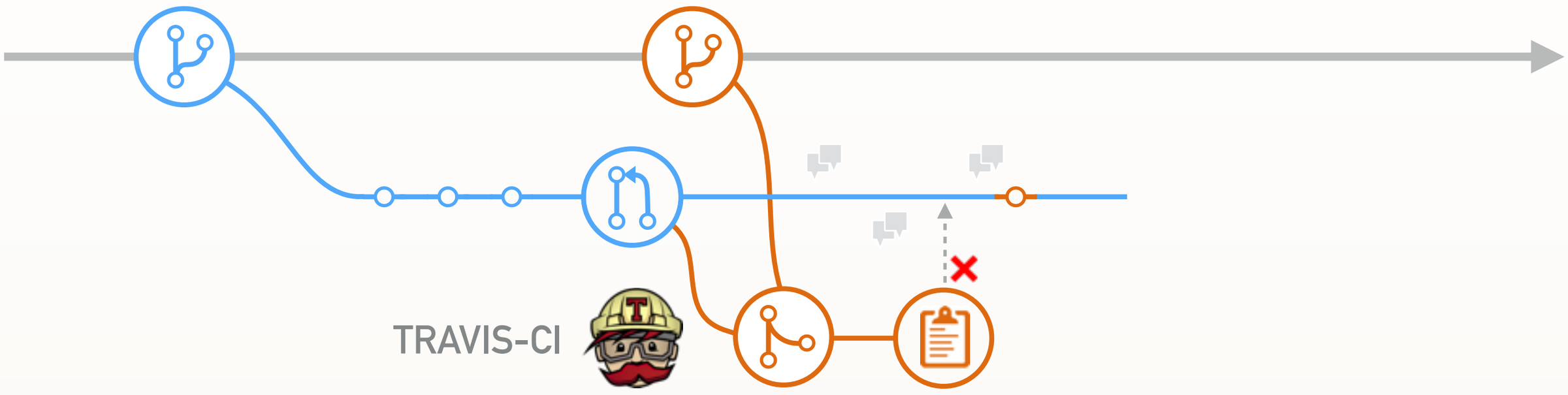
CI PULL REQUEST PROCESS



CI “Build”:
Compile, unit test, integration test, quality analysis, etc.

[ICSME 2014]

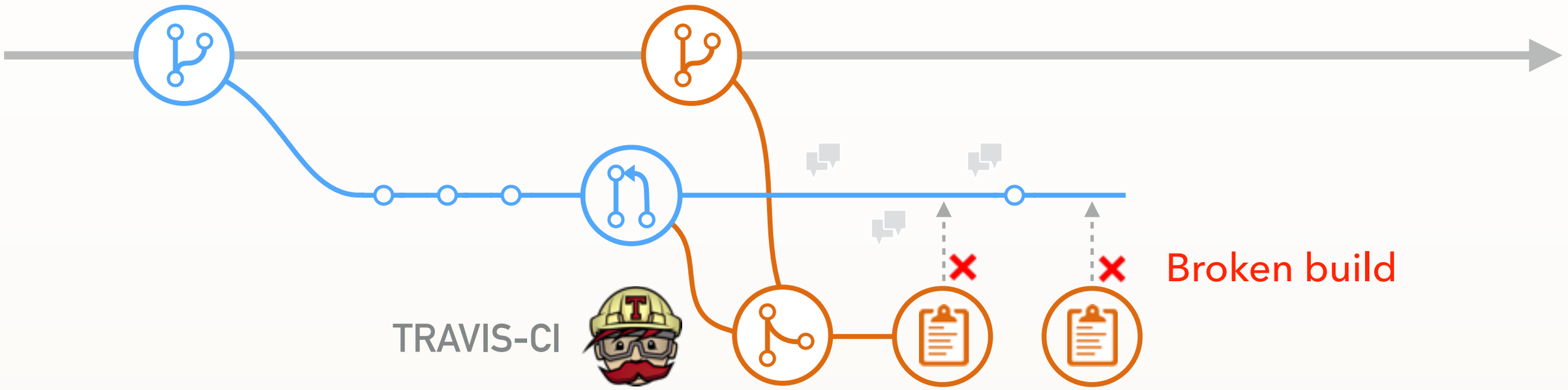
CI PULL REQUEST PROCESS



PR is updated
in response
to failure

[ICSME 2014]

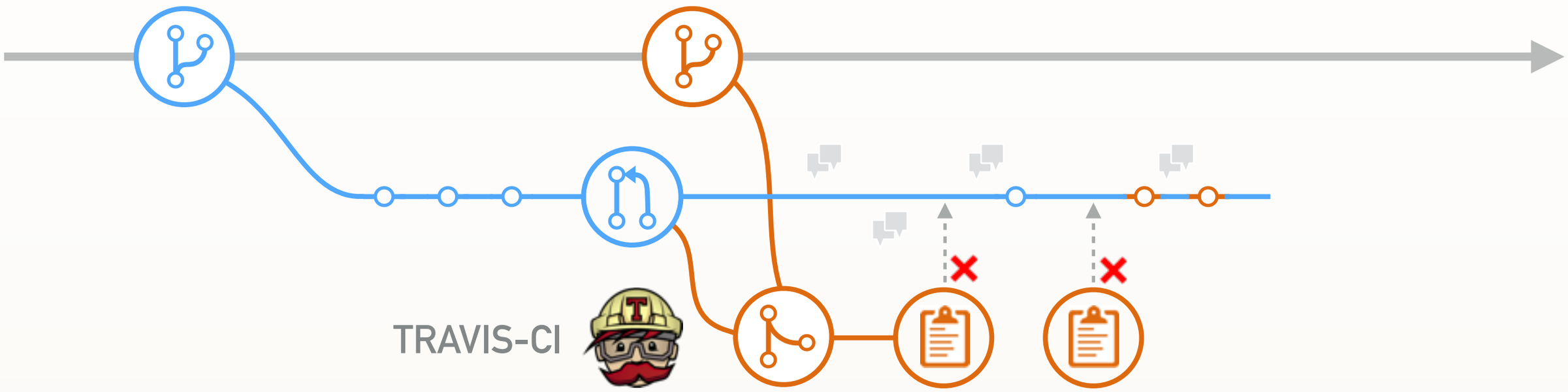
CI PULL REQUEST PROCESS



New build
after update

[ICSME 2014]

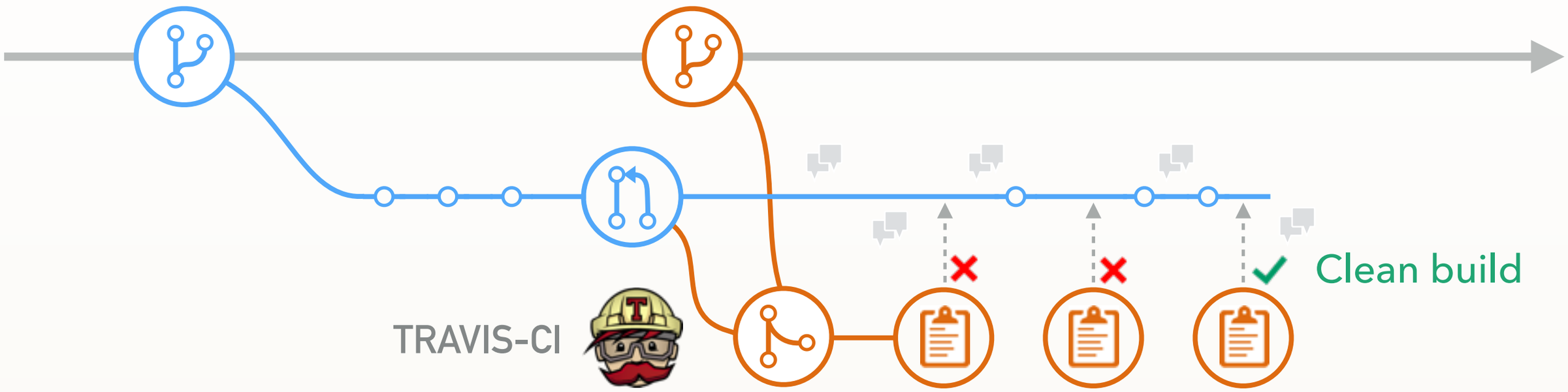
CI PULL REQUEST PROCESS



More
updates

[ICSME 2014]

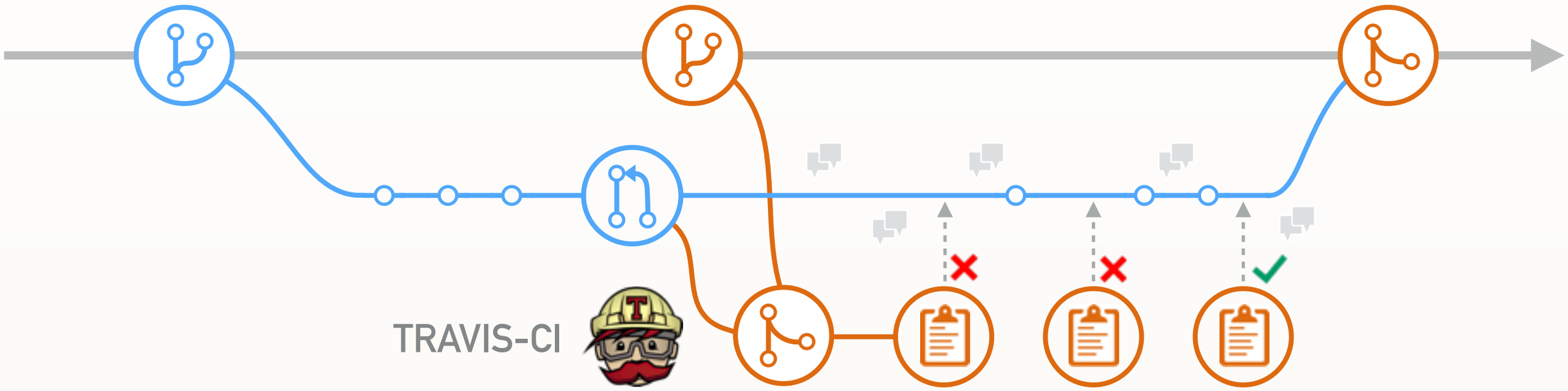
CI PULL REQUEST PROCESS



Tests
finally
pass

[ICSME 2014]

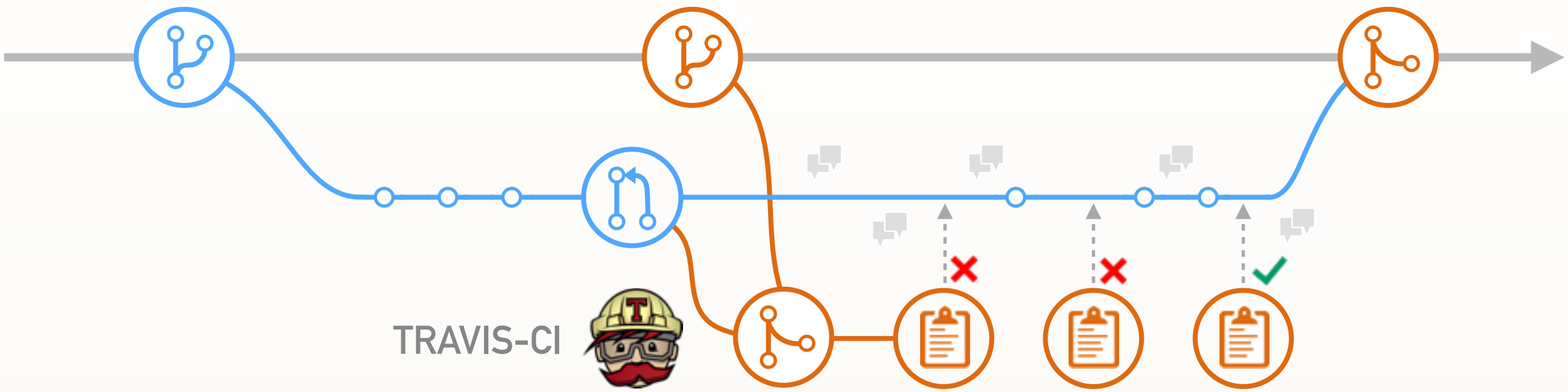
CI PULL REQUEST PROCESS



Merge

[ICSME 2014]

CI PULL REQUEST PROCESS

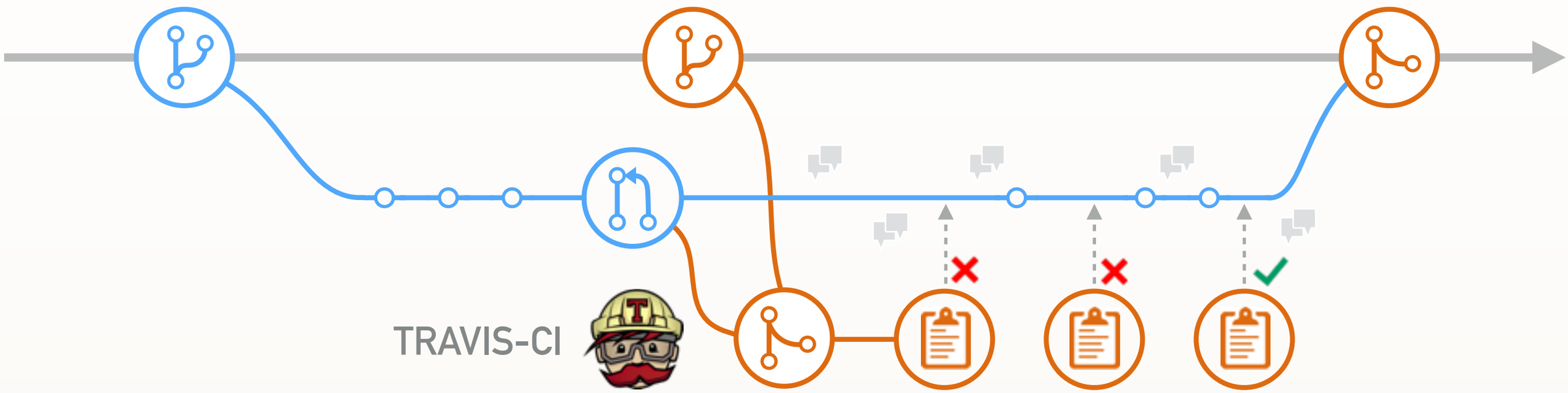


https://www.flickr.com/photos/javierdiazb/14052486641

CI AS GATEKEEPER:

- Integrated in PR process
- Tighter feedback loop
- Find integration errors & regression failures early

CI PULL REQUEST PROCESS



https://www.flickr.com/photos/javierdiazb/14052486641

CI AS GATEKEEPER:

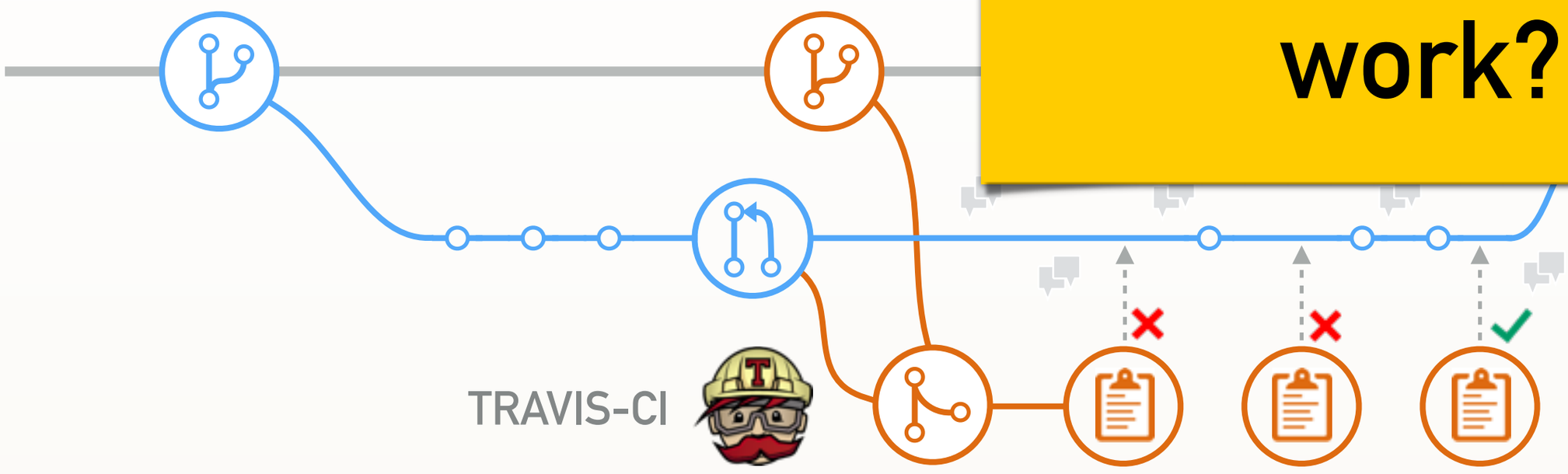
- Integrated in PR process
- Tighter feedback loop
- Find integration errors & regression failures early

http://goo.gl/ermLno

CI AS VALET:

- Automate more of the process
- More time to focus on other things

CI PULL REQUEST PROCESS



How well does it work?

https://www.flickr.com/photos/javierdiazb/14052486641

CI AS GATEKEEPER:

- Integrated in PR process
- Tighter feedback loop
- Find integration errors & regression failures early

http://goo.gl/ermLno

CI AS VALET:

- Automate more of the process
- More time to focus on other things

NATURAL EXPERIMENT

1. Mine data from projects that
adopted Travis-CI

NATURAL EXPERIMENT

1. Mine data from projects that **adopted Travis-CI**

2. Compare **before vs. after**

Pull request throughput

- How many pull requests are closed per month?

Defect rate

- How many bugs are reported per month?

NOT ALL BUGS CREATED EQUAL

Bugs vs. feature requests

STM32L1 get_cpuid() hard faults when using a Cat. 1 or Cat. 2 STM32L1 #3692

New issue

Closed **DipSwitch** opened this issue 12 days ago · 2 comments

DipSwitch commented 12 days ago

From the STM32L1 Reference Manual (31.2 Unique device ID registers (96 bits)):

```
Base address: 0x1FF80050 for Cat.1 and Cat.2 devices and 0x1FF800D0 for Cat.3, Cat.4, Cat.5 and Ca
```

Three solutions possible for this problem:

- Compile time: Via the linkerscript for the device (this I would prefer since this is the cleanest solution in my opinion)

```
MEMORY
{
    rom (rx)        : ORIGIN = 0x08000000, LENGTH = 128K
    ram (rw)        : ORIGIN = 0x20000000, LENGTH = 32K

    cpuid (r)       : ORIGIN = 0x1FF80050, LENGTH = 12
}

_cpuid_address = ORIGIN(cpuid);

INCLUDE cortexm_base.ld
```

Labels

arm

bug

Milestone

Release 2015.09

Assignee

thomaseichinger

Notifications

Subscribe

You're not receiving notifications from this thread.

4 participants

NOT ALL BUGS CREATED EQUAL

Bugs vs. feature requests

STM32L1 get_cpuid() hard faults when using a Cat. 1 or Cat. 2 STM32L1 #3692

New issue

Closed **DipSwitch** opened this issue 12 days ago · 2 comments

DipSwitch commented 12 days ago

From the STM32L1 Reference Manual (31.2 Unique device ID registers (96 bits)):

```
Base address: 0x1FF80050 for Cat.1 and Cat.2 devices and 0x1FF800D0 for Cat.3, Cat.4, Cat.5 and Ca
```

Three solutions possible for this problem:

- Compile time: Via the linkerscript for the device (this I would prefer since this is the cleanest solution in my opinion)

```
MEMORY
{
    rom (rx)        : ORIGIN = 0x08000000, LENGTH = 128K
    ram (rw)        : ORIGIN = 0x20000000, LENGTH = 32K

    cpuid (r)       : ORIGIN = 0x1FF80050, LENGTH = 12
}

_cpuid_address = ORIGIN(cpuid);

INCLUDE cortexm_base.ld
```

Labels

arm

bug

Labels

arm

bug

4 participants

SOCIO-TECHNICAL PROCESS!

Bug reporter matters

Early vs. late discovery

Core developers (early)

Users (late)

SOCIO-TECHNICAL PROCESS!

Bug reporter matters

Early vs. late discovery

Core developers (early)

Users (late)

Other confounds

Project size

Team size

Project test suite size

Project popularity

Issue tracker activity

Project age

CHALLENGES

1. DATA MINING

2. STATISTICAL ANALYSIS

Defect rate
(#Bugs/month)

~

Travis-CI
(T/F)

+

Project age

+

Issue tracker activity

+

Project source code size

+

Project test code size

+

Project popularity

controls

Defect rate (#Bugs/month) ~  Travis-CI (T/F) +

Project age +

Issue tracker activity +

Project source code size +

Project test code size +

Project popularity

controls

ZERO-INFLATED NEGATIVE BINOMIAL REGRESSION

NEGATIVE BINOMIAL

Over-dispersed count data (variance > mean)

ZERO INFLATED

Excess zeros. No bugs reported:

- because high quality?
- because nobody reporting?

- P. D. Allison and R. P. Waterman. Fixed–effects negative binomial regression models. Sociological Methodology, 32(1):247–265, 2002.
- D. Lambert. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics, 34(1):1–14, 1992.

RESULTS

WITH TRAVIS-CI:

- Code grows faster
- Dev's find more defects
- Users don't experience quality changes

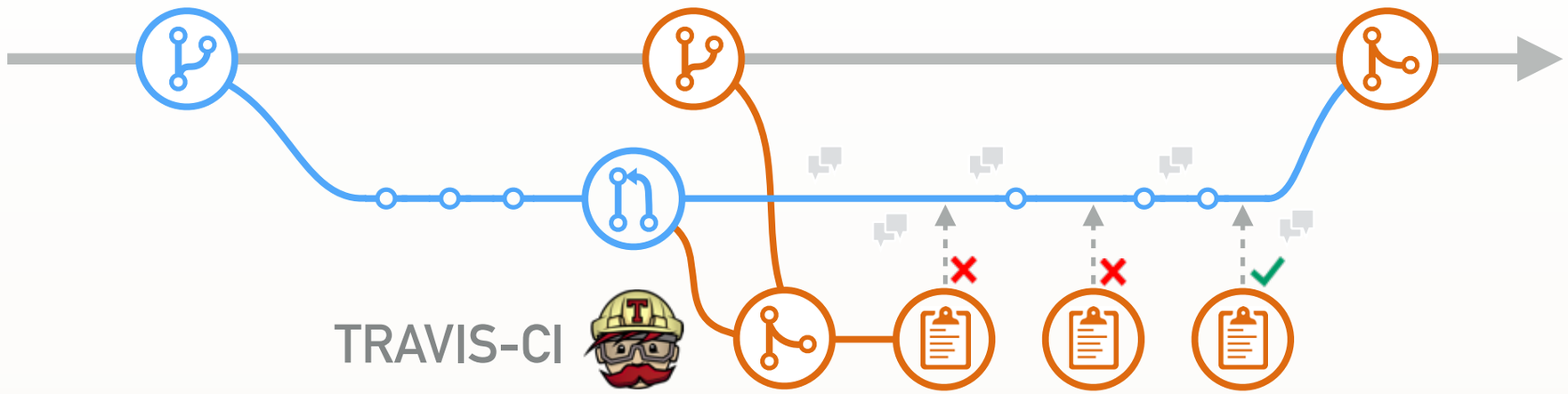
PR throughput
(#PRs/month)

+ 20..40%

Defect rate
(#Bugs/month)

+ 48% (core dev's)
None (users)

ONGOING & FUTURE WORK



>200,000 PROJECTS

- Where and why do CI failures occur?

  Many can be foreseen and prevented

- Do CI failures "predict" eventual defects?

  Yes - focus code review / testing

- How do people learn to program?

  Failures and fixes both logged

- How does the onboarding process change?

  Machine vs. human response
  Fear of losing face?
  Enforce project norms

SUMMARY: PERCEPTION → EVIDENCE

Project maintainers

PERCEPTION: CI REQUIRES BIG INVESTMENT

SUMMARY: PERCEPTION → EVIDENCE

Project maintainers

PERCEPTION: CI REQUIRES BIG INVESTMENT

Teams using CI handle more PRs & find more defects.

FSE '15a

SUMMARY: PERCEPTION → EVIDENCE

Project maintainers

PERCEPTION: CI REQUIRES BIG INVESTMENT

Teams using CI handle more PRs & find more defects.

FSE '15a

PERCEPTION: OPEN-SOURCE IS HOSTILE TO WOMEN

More diverse teams are more productive.

CHI '15

SUMMARY: PERCEPTION → EVIDENCE

Project maintainers

PERCEPTION: CI REQUIRES BIG INVESTMENT

Teams using CI handle more PRs & find more defects.

FSE '15a

PERCEPTION: OPEN-SOURCE IS HOSTILE TO WOMEN

More diverse teams are more productive.

CHI '15

Individual developers

PERCEPTION: MULTITASKING IS EXPENSIVE BUT NOBODY KNOWS WHEN TO STOP

> 5 projects/week always counterproductive

ICSE '16

SUMMARY: PERCEPTION → EVIDENCE

Project maintainers

PERCEPTION: CI REQUIRES BIG INVESTMENT

Teams using CI handle more PRs & find more defects.

FSE '15a

PERCEPTION: OPEN-SOURCE IS HOSTILE TO WOMEN

More diverse teams are more productive.

CHI '15

Individual developers

PERCEPTION: MULTITASKING IS EXPENSIVE BUT NOBODY KNOWS WHEN TO STOP

> 5 projects/week always counterproductive

ICSE '16

PERCEPTION: EXPERIENCE MATTERS THE MOST

Not in first 6 months: social environment more important

FSE '15b

SUMMARY: PERCEPTION → EVIDENCE

Project maintainers

PERCEPTION: CI REQUIRES BIG INVESTMENT

Teams using CI handle more PRs & find more defects.

FSE '15a

PERCEPTION: OPEN-SOURCE IS HOSTILE TO WOMEN

More diverse teams are more productive.

CHI '15

Individual developers

PERCEPTION: MULTITASKING IS EXPENSIVE BUT NOBODY KNOWS WHEN TO STOP

> 5 projects/week always counterproductive

ICSE '16

PERCEPTION: EXPERIENCE MATTERS THE MOST

Not in first 6 months: social environment more important

FSE '15b

Community designers

PERCEPTION: GAMIFICATION IS A GOOD IDEA

Incentivize participation

CSCW '14

But, quicker disengagement

IWC '14

ANALYTICS: NEXT STEPS

Project maintainers

CI BUILD FAILURES

Why do they happen?

Can we automatically prevent them?

Teams

DIVERSITY

Which aspects of team diversity are most important for:

- productivity?
- code quality?
- cohesiveness?
- architecture?

Community designers

DESIGN

Why are social coding platforms so seemingly exclusive?

Individual developers

MULTITASKING

Are there “risky” habits that lead to buggier code?

ANALYTICS + MACHINE LEARNING + NLP + . . .

12 million people

31 million repos

SOON: All the code that will ever be written has already been written.

ANALYTICS + MACHINE LEARNING + NLP + . . .

12 million people

31 million repos

SOON: All the code that will ever be written has already been written.

SOFTWARE DEVELOPMENT BECOMES A SEARCH PROBLEM

- Code snippets
- CI scripts
- Refactoring
- Porting
- Documentation
- Q&A

ANALYTICS + MACHINE LEARNING + NLP + . . .

12 million people

31 million repos

SOON: All the code that will ever be written has already been written.

SOFTWARE DEVELOPMENT BECOMES A SEARCH PROBLEM

- Code snippets
- CI scripts
- Refactoring
- Porting
- Documentation
- Q&A

```
1  using System;
2
3  namespace Demo
4  {
5      class Program00
6      {
7          static void Method01(string[] args)
8          {
9              string pth = "c:\\file.txt";
10             ///how to read file pth line by line
11         }
12     }
13 }
```

stackoverflow

C# Reading a File Line By Line

54

27

I am trying to read some text files, where each line needs to be processed. At the moment I am just using a StreamReader, and then reading each line individually.

I am wondering whether there is a more efficient way (in terms of LoC and readability) to do this using LINQ without compromising operational efficiency. The examples I have seen involve loading the whole file into memory, and then processing it. In this case however I don't believe that would be very efficient. In the first example the files can get up to about 50k, and in the second example, not all lines of the file need to be read (sizes are typically < 10k).

You could argue that nowadays it doesn't really matter for these small files, however I believe that sort of the approach leads to inefficient code.

Thanks for your time!

First example:

```
// open file
using(var file = System.IO.File.OpenText(_LstFilename))
{
    // read file
    while (!file.EndOfStream)
    {
        String line = file.ReadLine();

        // ignore empty lines
        if (line.Length > 0)
        {
            // create addon
            T addon = new T();
            addon.Load(line, _BaseDir);

            // add to collection
            collection.Add(addon);
```

Bing Code Search: http://codesnippet.research.microsoft.com

ANALYTICS + MACHINE LEARNING + NLP + . . .

12 million people

31 million repos

SOON: All the code that will ever be written has already been written.

```
k = [[1, 2], [4], [5, 6, 2], [1, 2], [3], [4]]
///How to remove duplicates from a list of lists?
```

SOFTWARE DEVELOPMENT BECOMES A SEARCH PROBLEM

- Code snippets
- CI scripts
- Refactoring
- Porting
- Documentation
- Q&A

I noticed **you use iterators a lot**. Here's how you can do it with iterators:

The iterator-based solution is faster, but **pull request reviewers tend to prefer this set-based version**:

Don't forget the NULL check! **It's a common bug**.

ANALYTICS + MACHINE LEARNING + NLP + …

12 million people

31 million repos

SOON: All the code that will ever be written has already been written.

SOFTWARE DEVELOPMENT BECOMES A SEARCH PROBLEM

- Code snippets
- CI scripts
- Refactoring
- Porting
- Documentation
- Q&A

103 lines (95 sloc) | 2.61 KB — .travis.yml

```
# After changing this file, check it on:
#   http://lint.travis-ci.org/
language: python

# Run jobs on container-based infrastructure, can be overridden per job
sudo: false

# Travis whitelists the installable packages, additions can be requested
#   https://github.com/travis-ci/apt-package-whitelist
addons:
  apt:
    packages: &common_packages
      - gfortran
      - libatlas-dev
      - libatlas-base-dev
      # Speedup builds, particularly when USE_CHROOT=1
      - eatmydata

cache:
  directories:
    - $HOME/.cache/pip

env:
  global:
    - WHEELHOUSE_UPLOADER_USERNAME=travis.numpy
```

Don’t forget to test against Python 2.6. **Similar code breaks Python 2.6 builds often**.

“BIG CODE”

MICHELANGELO:

“Every block of stone has a statue inside it; it is the task of the sculptor to discover it.”

“BIG CODE”

MICHELANGELO:

“Every block of stone has a statue inside it; it is the task of the sculptor to discover it.”

Almost any software engineering question has an answer inside a big code archive.
It is the task of the data scientist to discover it.

ACKNOWLEDGEMENTS












Netherlands Organisation for Scientific Research

NSF

AFOSR
AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

UC DAVIS
UNIVERSITY OF CALIFORNIA

TU/e Technische Universiteit Eindhoven University of Technology

Baishakhi Ray · Alexander Serebrenik · Vladimir Filkov · Prem Devanbu · Cindy Rubio Gonzalez · Casey Casalnuovo · Daryl Posnett · Yue Yu · Qi Xuan · Mark van den Brand · Kelly Blincoe · Daniela Damian

SOFTWARE DEVELOPMENT IS CHANGING

OPEN-SOURCE IS GROWING

Companies:
- 78% run OSS
- 66% build on top of OSS

SOCIAL CODING IS GROWING

12 million people · 31 million repositories · 18.5 million software dev's · 15,000+ people

CULTURE CHANGE

"it's just so uncool not sharing the code in the age of social coding"

HIRING

- **$100+** /hour:
  - owns popular OSS products;
  - stack**overflow** score > 20K; …
- **$50+** /hour:
  - active OSS contributor;
  - stack**overflow** score > 5K; …

INDUSTRIAL INVOLVEMENT & ADOPTION

Microsoft — Open source, from Microsoft with love — Redmond, WA — http://www.microsoft.com…

Google — https://developers.google.com/

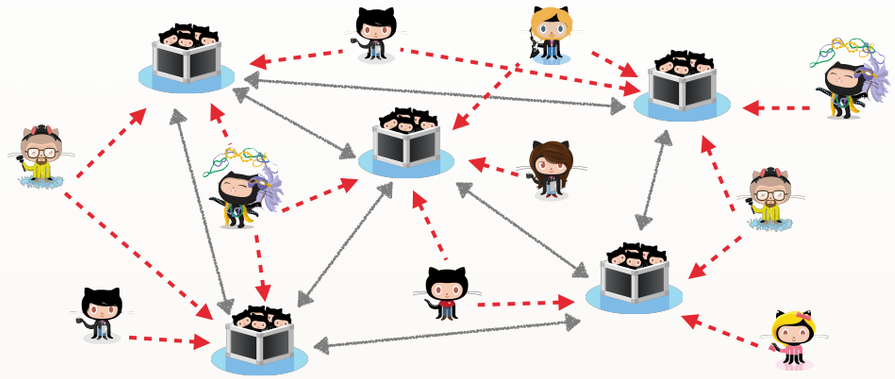
Facebook — We work hard to contribute our work back to the web, mobile, big data, & infrastructure communities. — Menlo Park, California — https://code.facebook.com/projects/

- GitHub stats from: https://github.com/about
- World estimates from: http://goo.gl/Htnni9
- Open source-style collaborative development practices in commercial projects using GitHub E Kalliamvakou, D Damian, K Blincoe, L Singer, DM German. *ICSE 2015*
- How Much Do You Cost? Yegor Bugayenko http://goo.gl/N0mL3F
- Activity traces and signals in software developer recruitment and hiring J Marlow, L Dabbish. *CSCW 2013*

BlackDuck 2015 survey https://goo.gl/LLaqqs

https://news.ycombinator.com/item?id=2982987

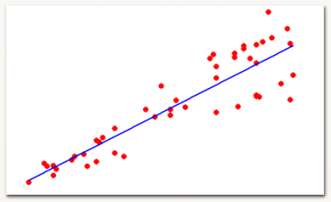
SOFTWARE ANALYTICS TO THE RESCUE



EVERYTHING IS ARCHIVED!

- Source code
- People involved
- Bug reports
- Communication
- …

DATA ANALYSIS (STATISTICS) → TRENDS



DATA-DRIVEN vs. INTUITION-BASED

decision making

DATA SCIENTIST:

standard on software teams

- Analyze This! 145 Questions for Data Scientists in Software Engineering A. Begel, T. Zimmermann. *ICSE 2014*
- The Emerging Role of Data Scientists on Software Development Teams M. Kim, T. Zimmermann, R. DeLine, A. Begel. *ICSE 2016*

EXPERIMENTAL RISK: BIG DATA TO THE RESCUE

12 million people · 31 million repos

1. **FALSE POSITIVES**
2. **FALSE NEGATIVES**
3. **CONFOUNDS**

	Reject Null Hyp.	Accept Null Hyp.
Null Hyp. TRUE	1	
Null Hyp. FALSE		2

HUGE SAMPLE SIZES:

- More stringent a priori about significance level → reduce False Positives
- Detect even small effects → reduce False Negatives
- Handle more degrees of freedom → control for Confounds

SEPARATE SIGNAL FROM NOISE:

- Quantify effect size
- Mix research methods
  - Quantitative: stats, data mining
  - Qualitative: case studies, user surveys, grounded theory

VALIDATE DATA FIRST!

- Spot-checking

SUMMARY: PERCEPTION → EVIDENCE

Project maintainers

PERCEPTION: CI REQUIRES BIG INVESTMENT

Teams using CI handle more PRs & find more defects. — FSE '15a

PERCEPTION: OPEN-SOURCE IS HOSTILE TO WOMEN

More diverse teams are more productive. — CHI '15

Individual developers

PERCEPTION: MULTITASKING IS EXPENSIVE BUT NOBODY KNOWS WHEN TO STOP

>4-5 projects/week always counterproductive — ICSE '16

PERCEPTION: EXPERIENCE MATTERS THE MOST

Not in first 6 months: social environment more important — FSE '15b

Community designers

PERCEPTION: GAMIFICATION IS A GOOD IDEA

Incentivize participation — CSCW '14

But, quicker disengagement — IWC '14