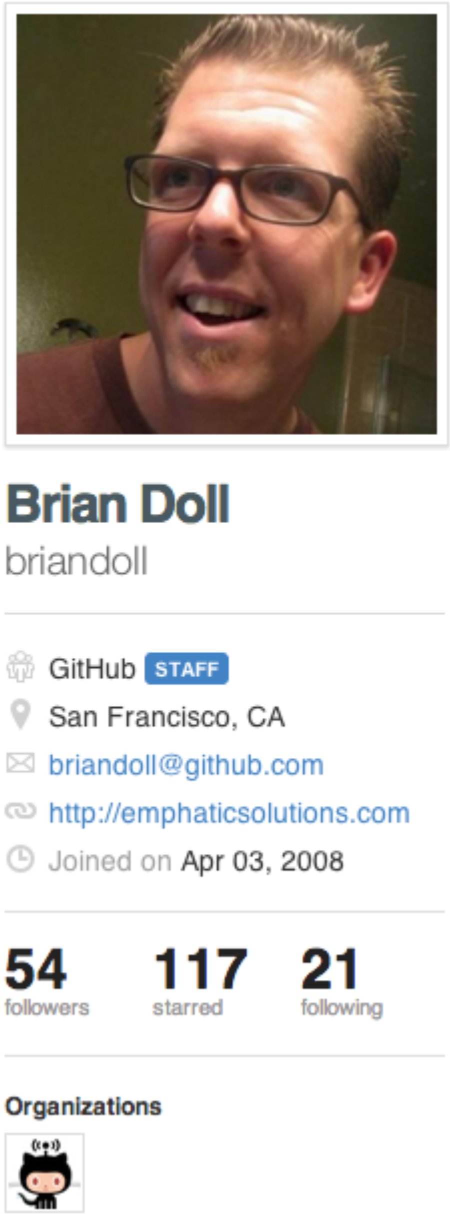
Analyzing Millions of GitHub Commits

what makes developers happy, angry, and everything in between?

Brian Doll briandoll@github.com @briandoll
Ilya Grigorik igrigorik@google.com @igrigorik



Brian Doll

briandoll

GitHub STAFF

San Francisco, CA

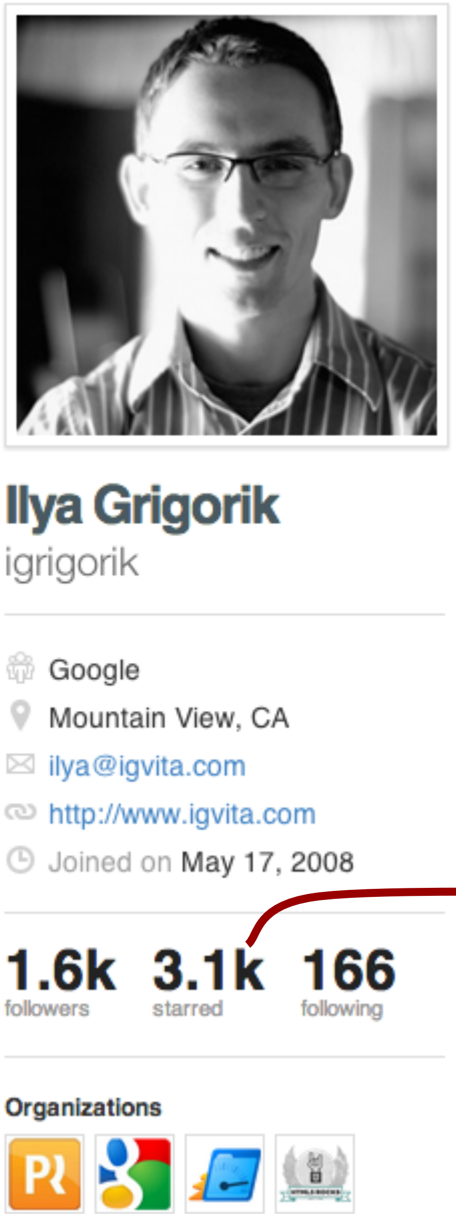
briandoll@github.com

http://emphaticsolutions.com

Joined on Apr 03, 2008

54	117	21
followers	starred	following

Organizations



Ilya Grigorik

igrigorik

Google

Mountain View, CA

ilya@igvita.com

http://www.igvita.com

Joined on May 17, 2008

1.6k	3.1k	166
followers	starred	following

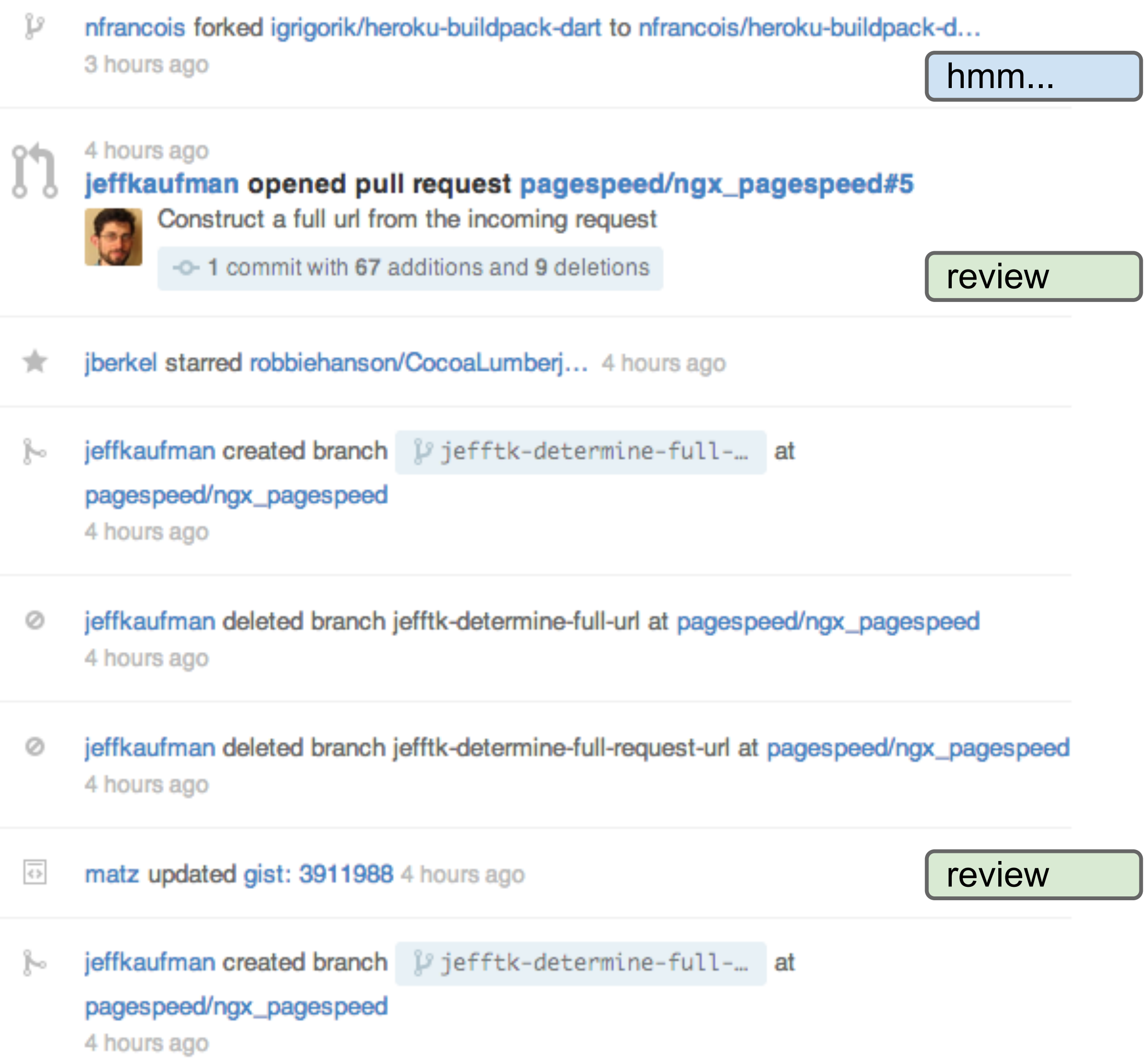
Organizations

<facepalm>

"Keeping up with ***3000+*** *open-source projects is not easy... If only there was a better way!"*

Ilya, circa early 2012

(Ilya's) Burning questions...



nfrancois forked igrigorik/heroku-buildpack-dart to nfrancois/heroku-buildpack-d...
3 hours ago

4 hours ago
jeffkaufman opened pull request pagespeed/ngx_pagespeed#5
Construct a full url from the incoming request
1 commit with 67 additions and 9 deletions

jberkel starred robbiehanson/CocoaLumberj... 4 hours ago

jeffkaufman created branch `jefftk-determine-full-...` at pagespeed/ngx_pagespeed
4 hours ago

jeffkaufman deleted branch jefftk-determine-full-url at pagespeed/ngx_pagespeed
4 hours ago

jeffkaufman deleted branch jefftk-determine-full-request-url at pagespeed/ngx_pagespeed
4 hours ago

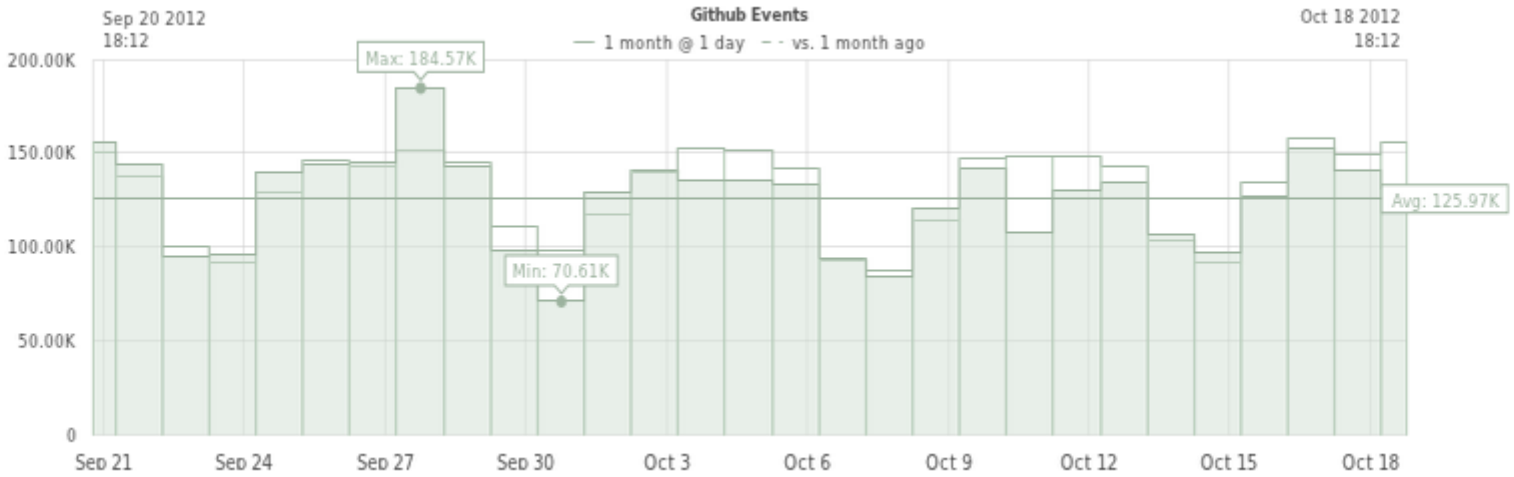
matz updated gist: 3911988 4 hours ago

jeffkaufman created branch `jefftk-determine-full-...` at pagespeed/ngx_pagespeed
4 hours ago

- **What were the hot new projects today?**
  - In Ruby land...
  - In JavaScript land...
  - Globally?
- **Did anyone commit something interesting or controversial?**
- **For the people I follow, which projects did they follow or contribute to?**
- **What are the emerging projects, or languages?**
- **...**

@briandoll @igrigorik

GitHub is *kinda a big deal* in open-source...

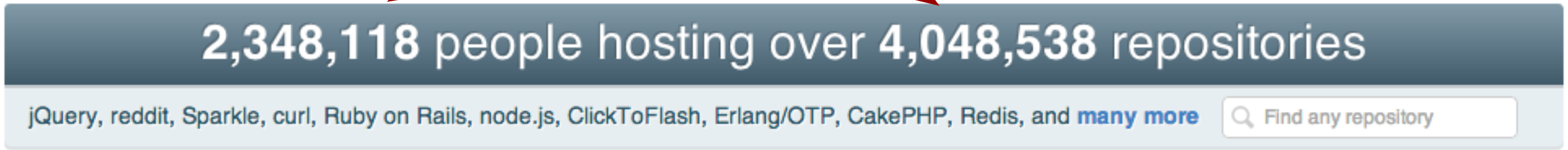


Activity stats:

- **Max:** 184,570 events / day
- **Avg:** 125,970 events/day

- **1~2** events / second!

BigNumber (tm)



2,348,118 people hosting over **4,048,538** repositories

jQuery, reddit, Sparkle, curl, Ruby on Rails, node.js, ClickToFlash, Erlang/OTP, CakePHP, Redis, and **many more**

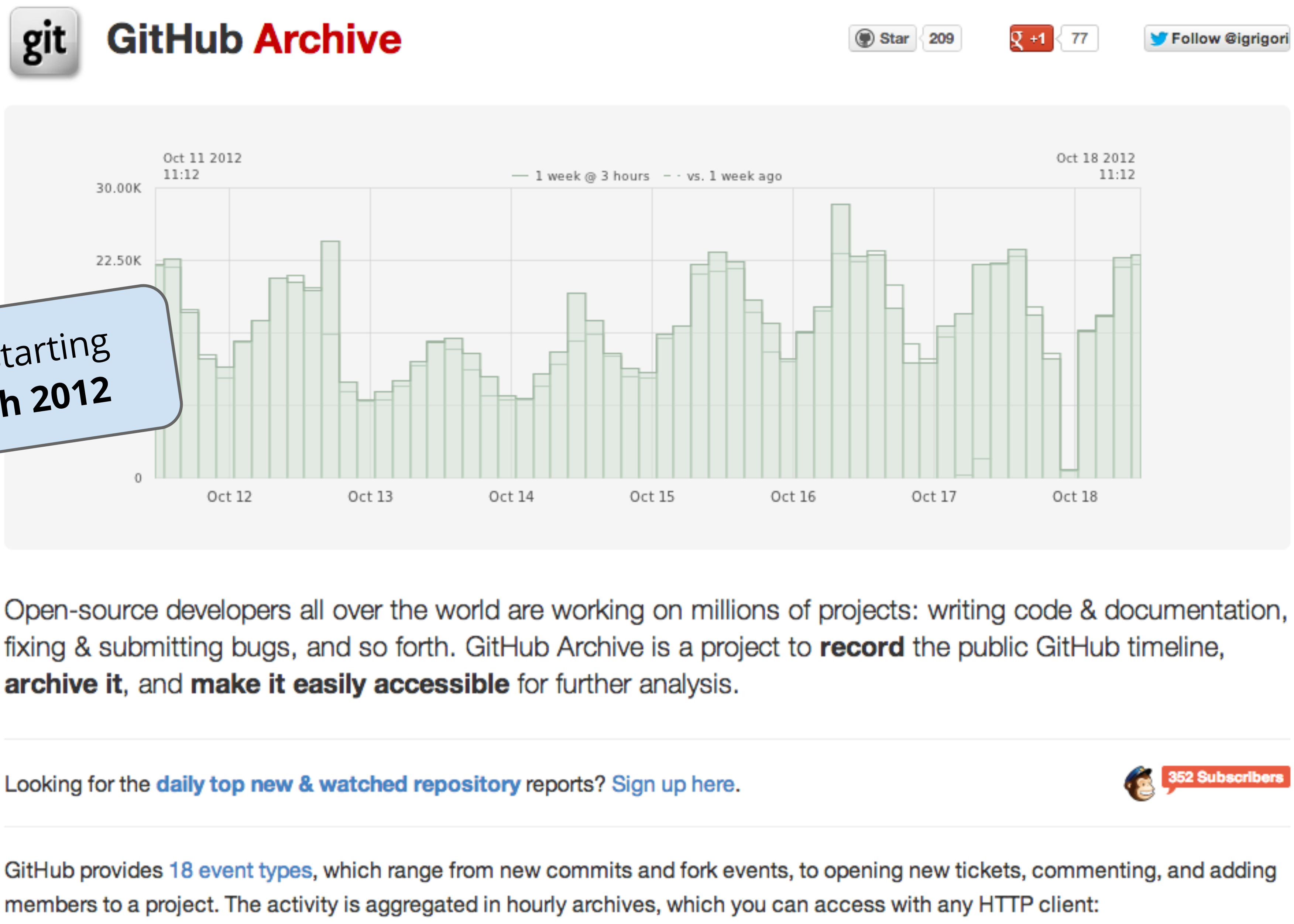
facebook Microsoft vmware redhat LinkedIn mozilla

The ***"aha"*** moment:

> *It's not my timeline, it's the **global timeline** that **contains the answers**.*

Now if only we had access to the GitHub archive...

(one weekend later...)



GitHub Archive

Open-source developers all over the world are working on millions of projects: writing code & documentation, fixing & submitting bugs, and so forth. GitHub Archive is a project to **record** the public GitHub timeline, **archive it**, and **make it easily accessible** for further analysis.

Looking for the **daily top new & watched repository** reports? Sign up here.

GitHub provides 18 event types, which range from new commits and fork events, to opening new tickets, commenting, and adding members to a project. The activity is aggregated in hourly archives, which you can access with any HTTP client:

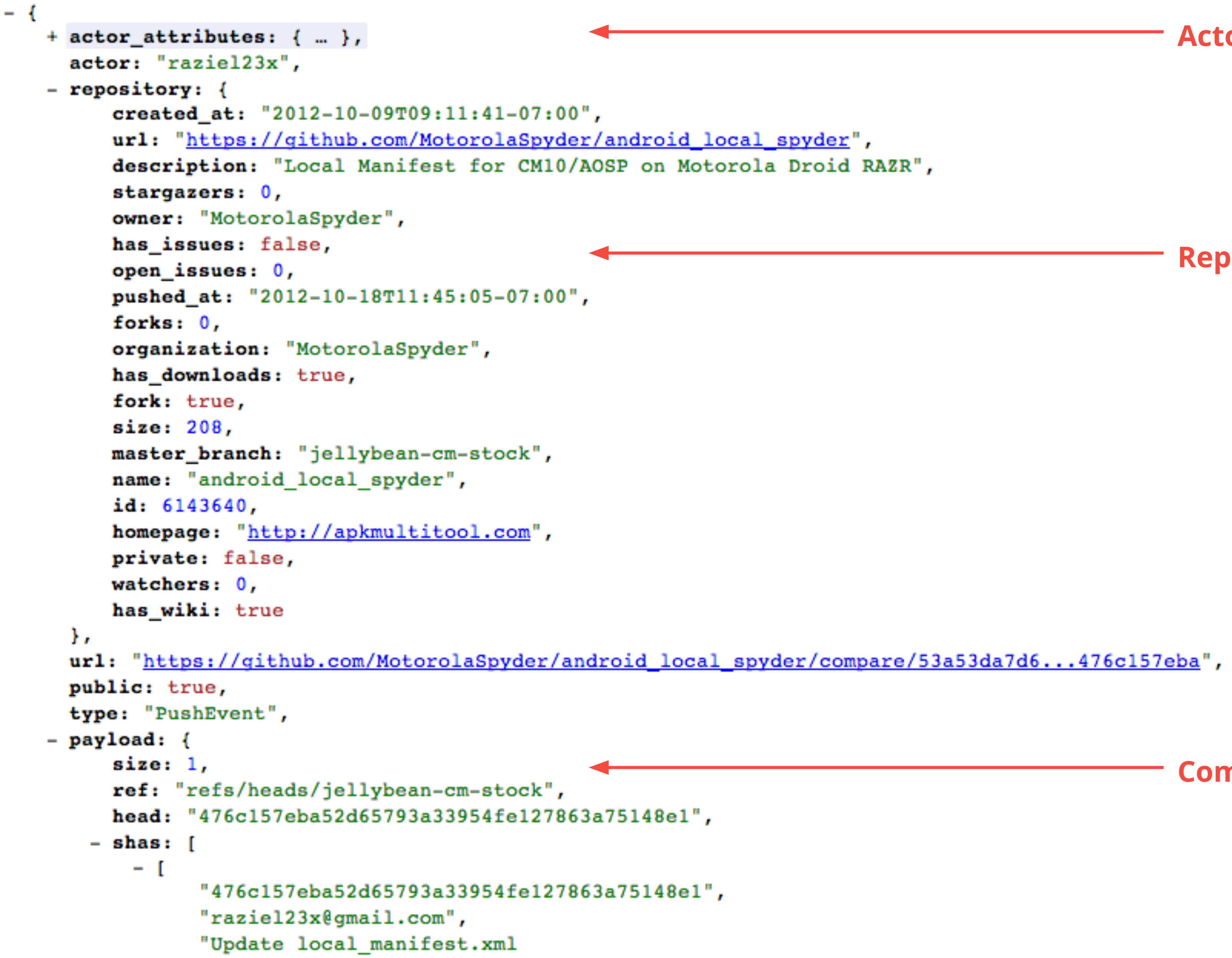
Data starting
March 2012

http://www.githubarchive.org collector code @ https://github.com/igrigorik/githubarchive.org/

Anatomy of an event

- CommitCommentEvent
- CreateEvent
- DeleteEvent
- DownloadEvent
- FollowEvent
- ForkEvent
- ForkApplyEvent
- GistEvent
- GollumEvent
- IssueCommentEvent
- IssuesEvent
- MemberEvent
- PublicEvent
- PullRequestEvent
- PullRequestReviewCommentEvent
- PushEvent
- TeamAddEvent
- WatchEvent

18 event types. JSON payload, meta-data rich.



```
- {
  + actor_attributes: { … },
    actor: "raziel23x",
  - repository: {
        created_at: "2012-10-09T09:11:41-07:00",
        url: "https://github.com/MotorolaSpyder/android_local_spyder",
        description: "Local Manifest for CM10/AOSP on Motorola Droid RAZR",
        stargazers: 0,
        owner: "MotorolaSpyder",
        has_issues: false,
        open_issues: 0,
        pushed_at: "2012-10-18T11:45:05-07:00",
        forks: 0,
        organization: "MotorolaSpyder",
        has_downloads: true,
        fork: true,
        size: 208,
        master_branch: "jellybean-cm-stock",
        name: "android_local_spyder",
        id: 6143640,
        homepage: "http://apkmultitool.com",
        private: false,
        watchers: 0,
        has_wiki: true
    },
    url: "https://github.com/MotorolaSpyder/android_local_spyder/compare/53a53da7d6...476c157eba",
    public: true,
    type: "PushEvent",
  - payload: {
        size: 1,
        ref: "refs/heads/jellybean-cm-stock",
        head: "476c157eba52d65793a33954fe127863a75148e1",
      - shas: [
          - [
                "476c157eba52d65793a33954fe127863a75148e1",
                "raziel23x@gmail.com",
                "Update local_manifest.xml
```

Actor information

Repository information

Commit data

GZIP archive(s)

Query	Command
Activity for April 11, 2012 at 3PM PST	wget http://data.githubarchive.org/2012-04-11-15.json.gz
Activity for April 11, 2012	wget http://data.githubarchive.org/2012-04-11-{0..23}.json.gz
Activity for April 2012	wget http://data.githubarchive.org/2012-04-**{01..31}**-**{0..23}**.json.gz

- Raw JSON data
- Hourly archives
- Easy access
- Uploaded every hour

+ Tool agnostic

- Lots of work

- Non-interactive

- Hard to analyze large ranges

Dremel, err... BigQuery

Publication Data

Venue

Proc. of the 36th Int'l Conf on Very Large Data Bases (2010), pp. 330-339

Publication Year

2010

*"Dremel is a scalable, **interactive ad-hoc query system for analysis of read-only nested data**. By combining multi-level execution trees and columnar data layout, it is **capable of running aggregation queries over trillion-row tables in seconds**. The system scales to thousands of CPUs and petabytes of data, and has thousands of users at Google."*

developers.google.com/**bigquery**

GitHub Archive =

JSON data

Meta-data rich

BigQuery =

Interactive ad-hoc analysis

Trillion-row tables

Table scan friendly (no indexes)

Column storage for efficient access

...

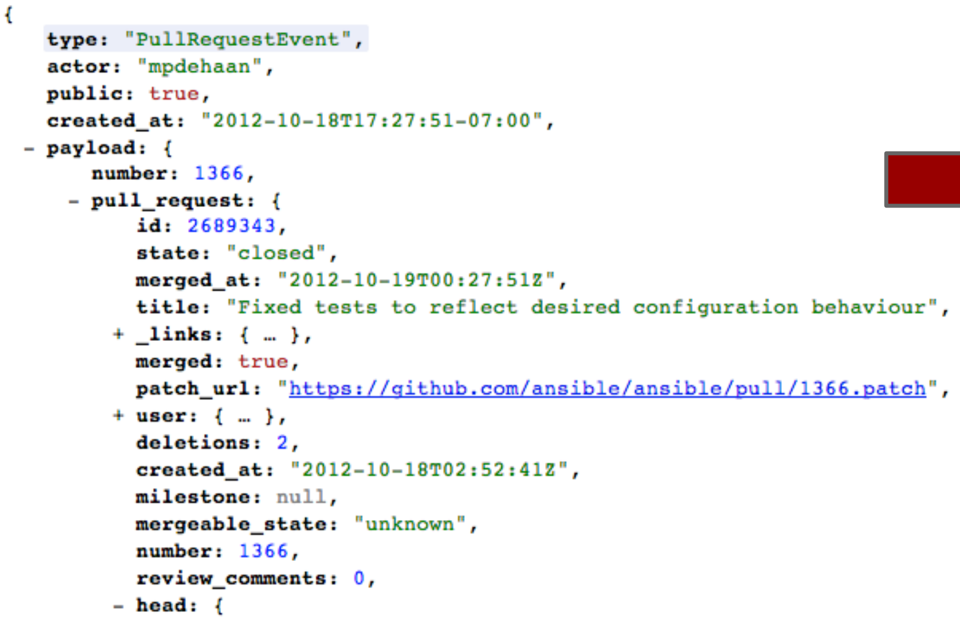
BigQuery + GitHub = Profit *

* still working on the profit part

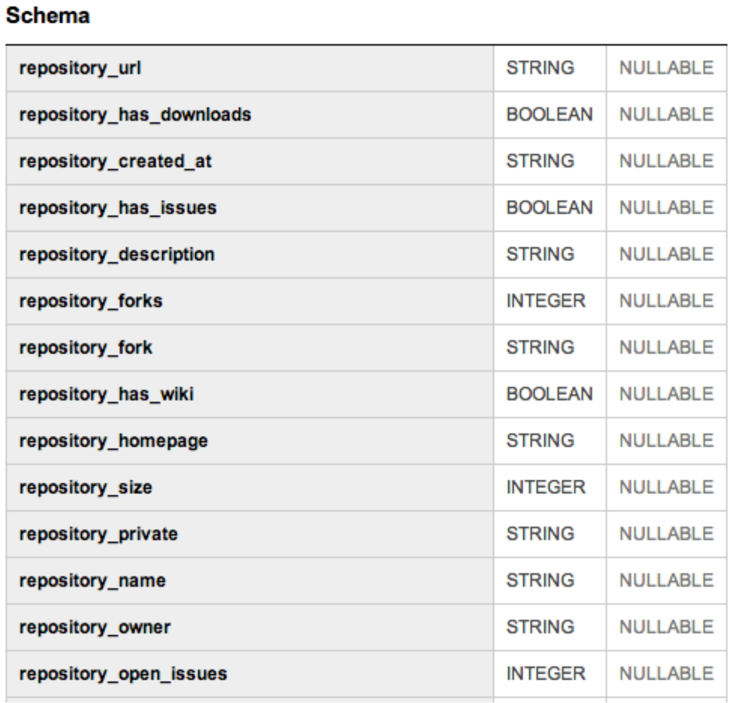
Data import in 3 commands - *automation ftw!*

1

```
$ wget http://data.githubarchive.org/2012-04-11-15.json.gz

$ ruby flatten.rb 2012-04-11-15.json.gz > flat.csv.gz
```



```
{
   type: "PullRequestEvent",
   actor: "mpdehaan",
   public: true,
   created_at: "2012-10-18T17:27:51-07:00",
 - payload: {
       number: 1366,
     - pull_request: {
           id: 2689343,
           state: "closed",
           merged_at: "2012-10-19T00:27:51Z",
           title: "Fixed tests to reflect desired configuration behaviour",
         + _links: { … },
           merged: true,
           patch_url: "https://github.com/ansible/ansible/pull/1366.patch",
         + user: { … },
           deletions: 2,
           created_at: "2012-10-18T02:52:41Z",
           milestone: null,
           mergeable_state: "unknown",
           number: 1366,
           review_comments: 0,
         - head: {
```



Schema

repository_url	STRING	NULLABLE
repository_has_downloads	BOOLEAN	NULLABLE
repository_created_at	STRING	NULLABLE
repository_has_issues	BOOLEAN	NULLABLE
repository_description	STRING	NULLABLE
repository_forks	INTEGER	NULLABLE
repository_fork	STRING	NULLABLE
repository_has_wiki	BOOLEAN	NULLABLE
repository_homepage	STRING	NULLABLE
repository_size	INTEGER	NULLABLE
repository_private	STRING	NULLABLE
repository_name	STRING	NULLABLE
repository_owner	STRING	NULLABLE
repository_open_issues	INTEGER	NULLABLE

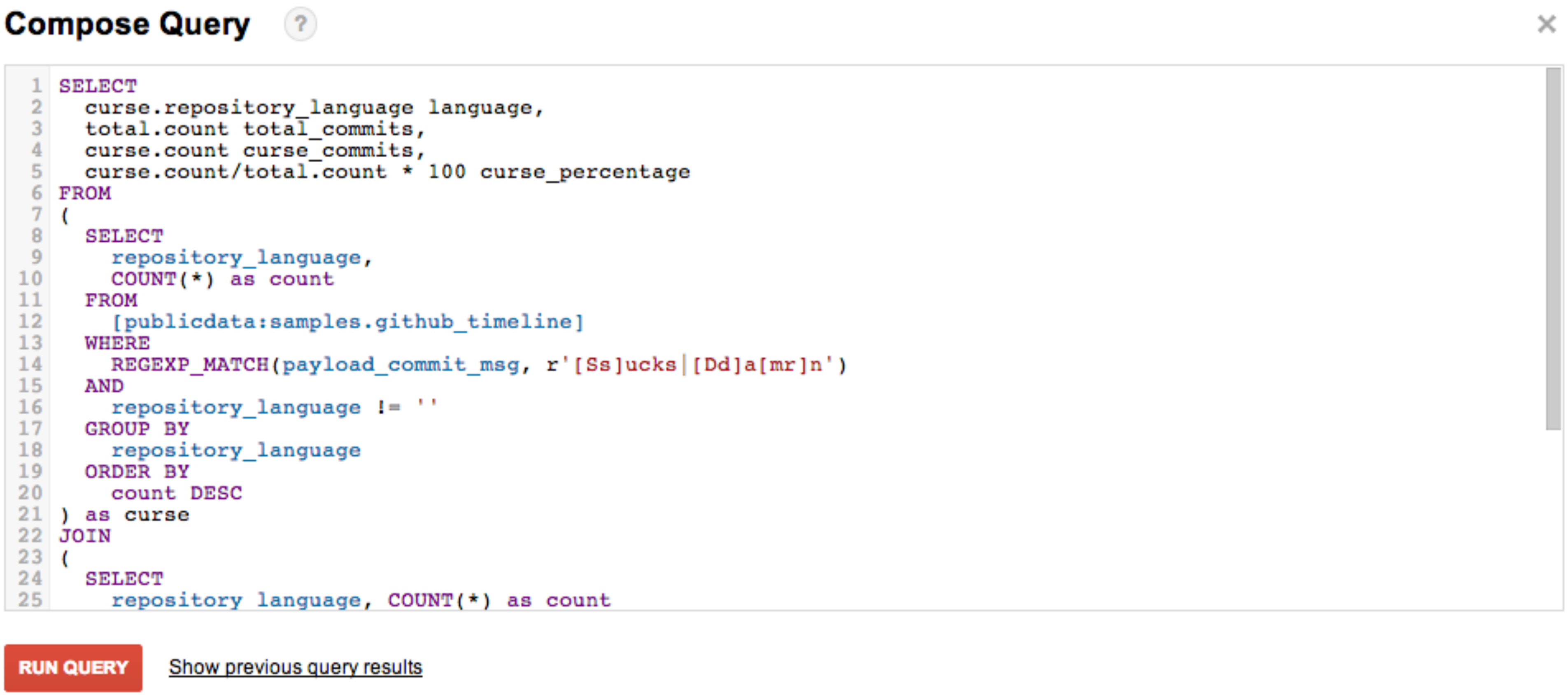
2

```
$ bq load github.timeline flat.csv.gz
```

Hourly cron-job to import flattened CSV **

@briandoll @igrigorik

A RegExp against entire table? Why not...



```sql
SELECT
  curse.repository_language language,
  total.count total_commits,
  curse.count curse_commits,
  curse.count/total.count * 100 curse_percentage
FROM
(
  SELECT
    repository_language,
    COUNT(*) as count
  FROM
    [publicdata:samples.github_timeline]
  WHERE
    REGEXP_MATCH(payload_commit_msg, r'[Ss]ucks|[Dd]a[mr]n')
  AND
    repository_language != ''
  GROUP BY
    repository_language
  ORDER BY
    count DESC
) as curse
JOIN
(
  SELECT
    repository_language, COUNT(*) as count
```

Speaking of interactive, ad-hoc analysis..

- BigQuery <3 table scans
- What's an index? **Table scans are no slower** than any other query...

https://gist.github.com/671fe0d3cb5e669a4fd6

Not your ....'s SQL language

Timestamp Functions

- FORMAT_UTC_USEC
- PARSE_UTC_USEC
- UTC_USEC_TO_DAY
- ...

Aggregate Functions

- AVG, COUNT
- STDDEV, VARIANCE
- **QUANTILES**
- **TOP, ...**

String Functions

- CONTAINS
- SUBSTR
- CONCAT, RPAD, LPAD
- ...

SQL bread and butter

- JOIN
- HAVING
- GROUP BY
- ORDER BY
- ...

Nested Record Functions

- WITHIN
- FLATTEN
- Scoped aggregation...

Other Functions

- CASE
- IF
- HASH
- **... and many others**

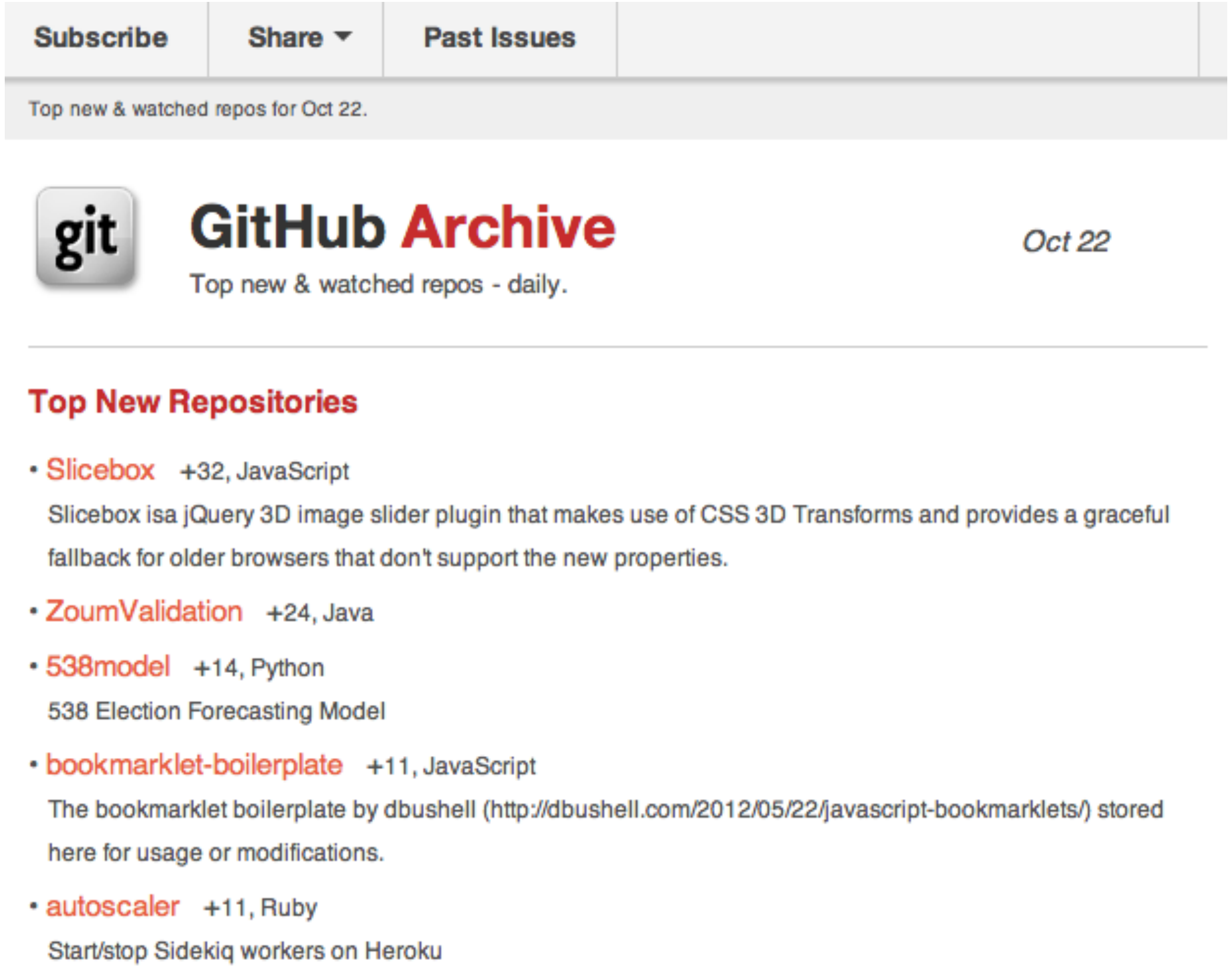
https://developers.google.com/bigquery/docs/query-reference

GitHub Daily (email) reports!

Speaking of scratching an itch...

https://www.githubarchive.org/

GitHub Daily: GitHub + BigQuery + MailChimp



Top new & watched repos for Oct 22.

GitHub Archive

Top new & watched repos - daily.

Oct 22

Top New Repositories

- Slicebox +32, JavaScript
  Slicebox isa jQuery 3D image slider plugin that makes use of CSS 3D Transforms and provides a graceful fallback for older browsers that don't support the new properties.
- ZoumValidation +24, Java
- 538model +14, Python
  538 Election Forecasting Model
- bookmarklet-boilerplate +11, JavaScript
  The bookmarklet boilerplate by dbushell (http://dbushell.com/2012/05/22/javascript-bookmarklets/) stored here for usage or modifications.
- autoscaler +11, Ruby
  Start/stop Sidekiq workers on Heroku

1. Cronjob
   a. Run query via ***bq***
   b. Export JSON
   c. Render HTML template
   d. Email via MailChimp

2. ~30 line of code

http://www.githubarchive.org/

GitHub Daily = GitHub Archive + BigQuery + MailChimp

```
SELECT repository_name, repository_language, repository_description, COUNT(repository_name) as cnt,
repository_url
FROM github.timeline
WHERE type="WatchEvent"
    AND PARSE_UTC_USEC(created_at) >= PARSE_UTC_USEC("#{yesterday} 20:00:00")
        AND repository_url IN (

             SELECT repository_url
             FROM github.timeline
             WHERE type="CreateEvent"
                     AND PARSE_UTC_USEC(repository_created_at) >= PARSE_UTC_USEC('#{yesterday} 20:00:00')
                     AND repository_fork = "false"
                     AND payload_ref_type = "repository"
             GROUP BY repository_url

         )
GROUP BY repository_name, repository_language, repository_description, repository_url
HAVING cnt >= 5
ORDER BY cnt DESC
LIMIT 25
```

1

http://www.githubarchive.org/ - https://gist.github.com/f8742314320e0a4b1a89

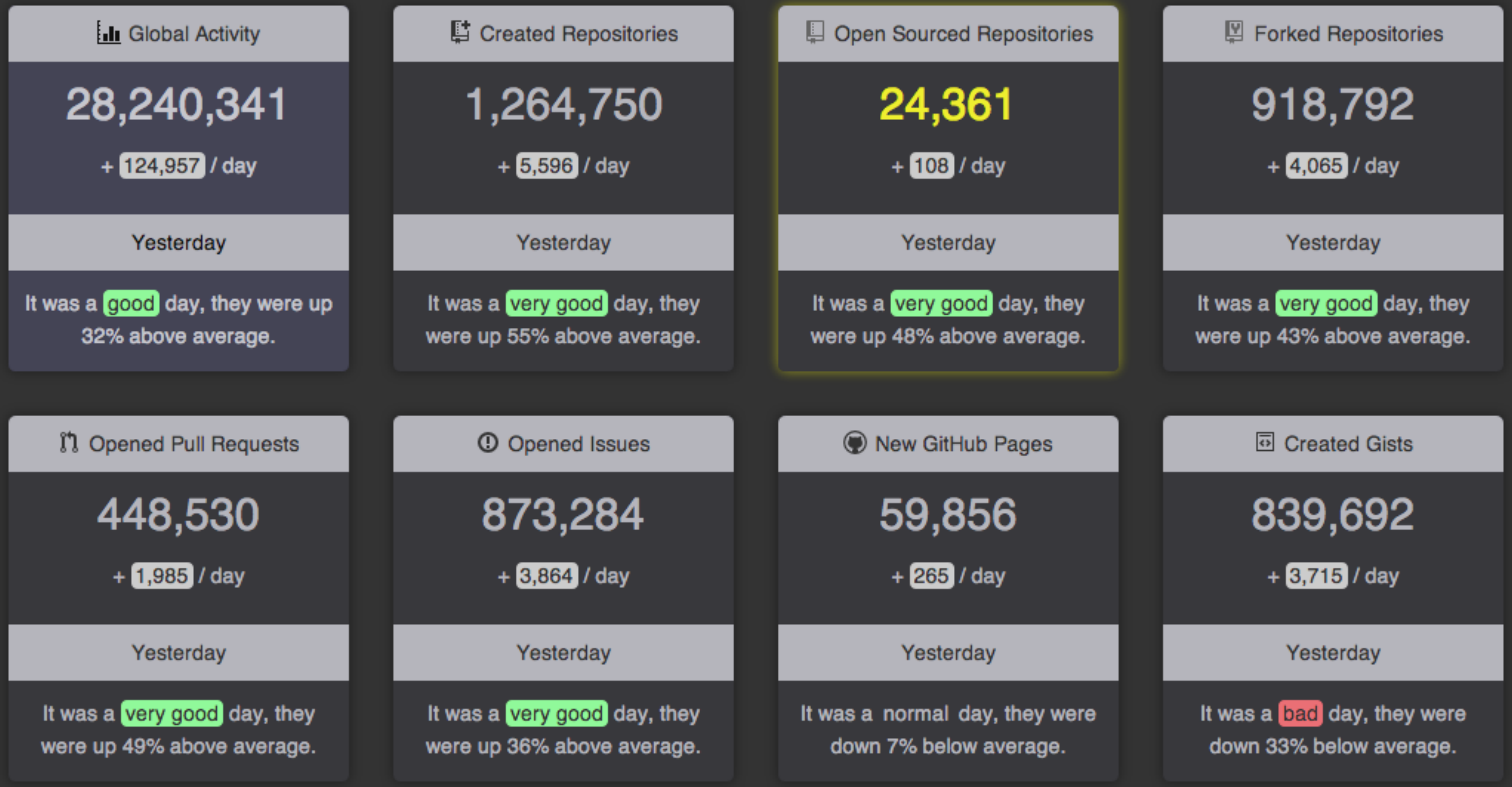
@briandoll @igrigorik

GitHub Data Challenge

Analyze with BigQuery, submit your entries...

https://github.com/blog/1112-data-at-github

octoboard.com - stats since March 11, 2012



Global Activity

28,240,341

+ 124,957 / day

Yesterday

It was a good day, they were up 32% above average.

Created Repositories

1,264,750

+ 5,596 / day

Yesterday

It was a very good day, they were up 55% above average.

Open Sourced Repositories

24,361

+ 108 / day

Yesterday

It was a very good day, they were up 48% above average.

Forked Repositories

918,792

+ 4,065 / day

Yesterday

It was a very good day, they were up 43% above average.

Opened Pull Requests

448,530

+ 1,985 / day

Yesterday

It was a very good day, they were up 49% above average.

Opened Issues

873,284

+ 3,864 / day

Yesterday

It was a very good day, they were up 36% above average.

New GitHub Pages

59,856

+ 265 / day

Yesterday

It was a normal day, they were down 7% below average.

Created Gists

839,692

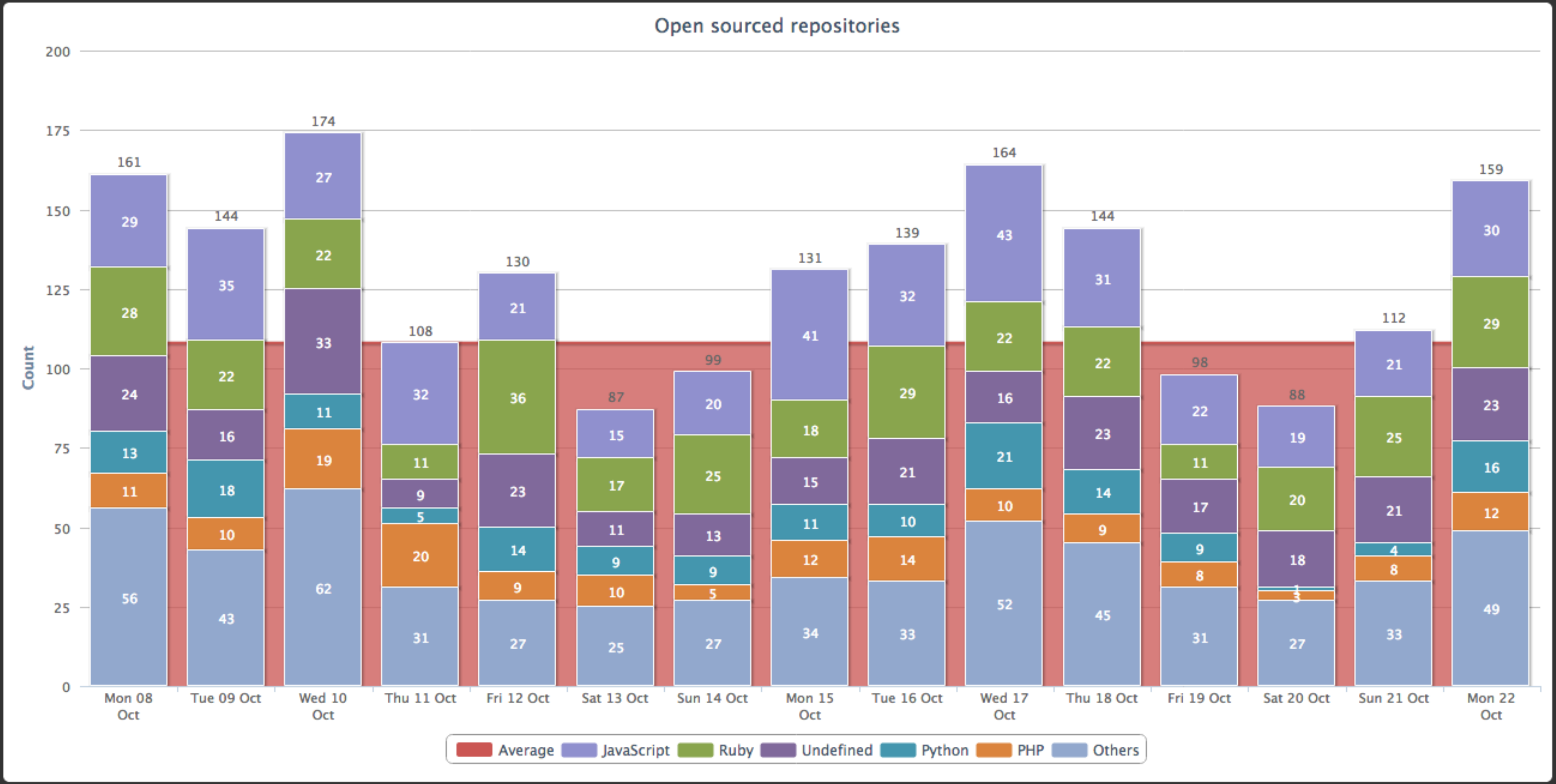
+ 3,715 / day

Yesterday

It was a bad day, they were down 33% below average.

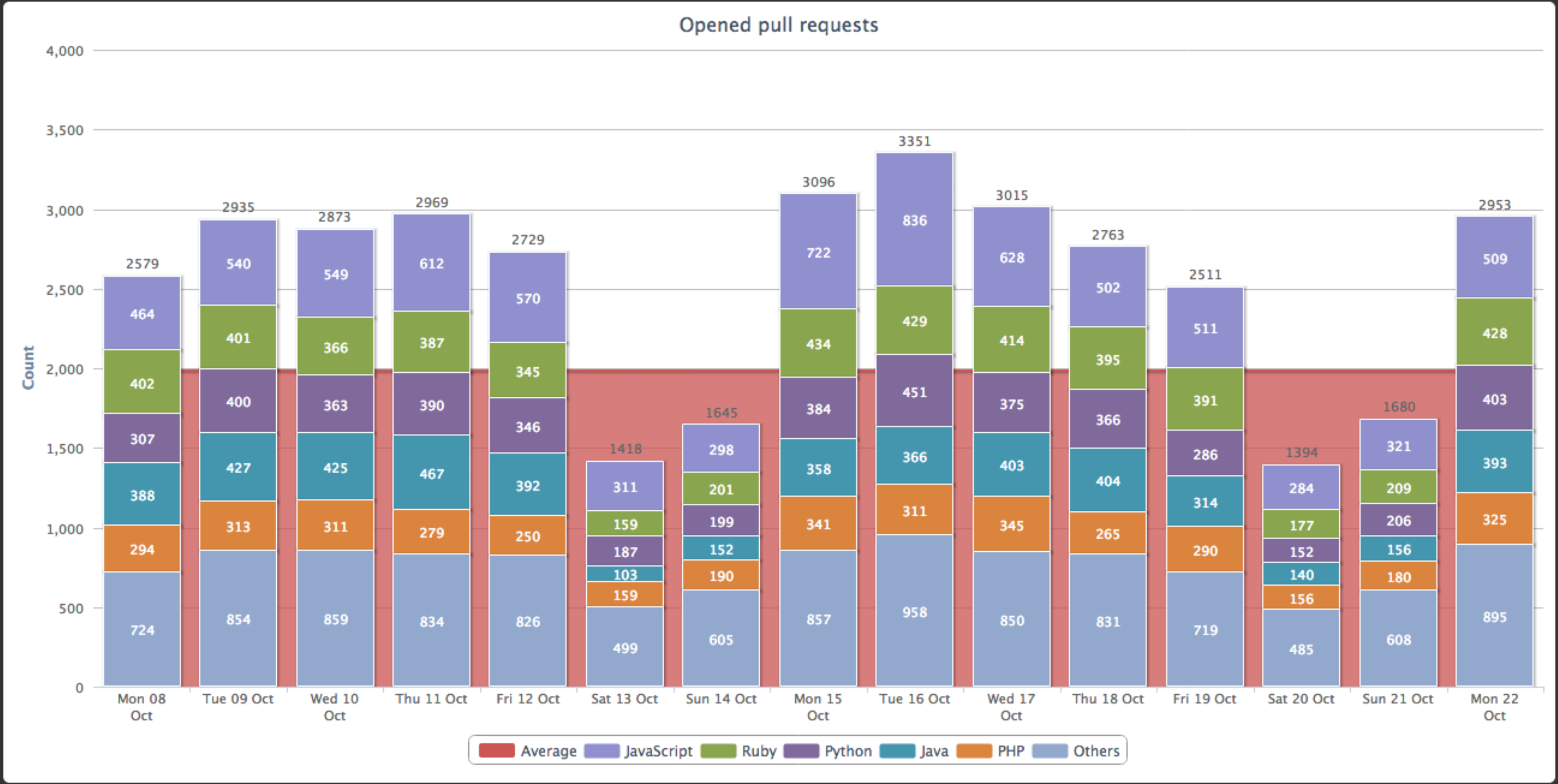
Denis Roussel https://github.com/KuiKui/Octoboard

~108 private repositories released to the public / day



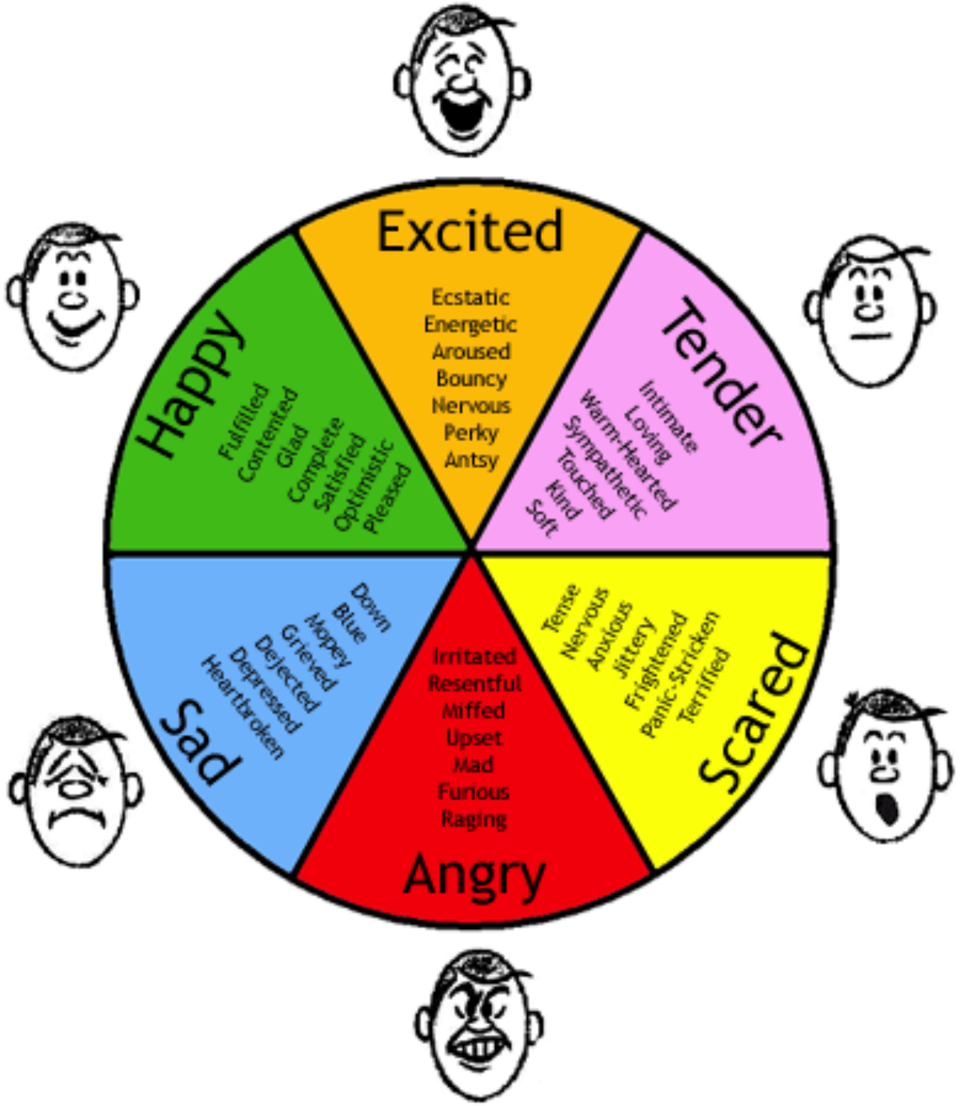
Active JavaScript and Ruby communities on GitHub.

~2000 Pull requests / day - which languages?



2x the activity on weekdays than on weekends! Saturday's are the slowest.

Emotional impact of programming languages...



Ramiro Gomez
https://github.com/yaph

http://geeksta.net/geeklog/exploring-expressions-emotions-github-commit-messages/

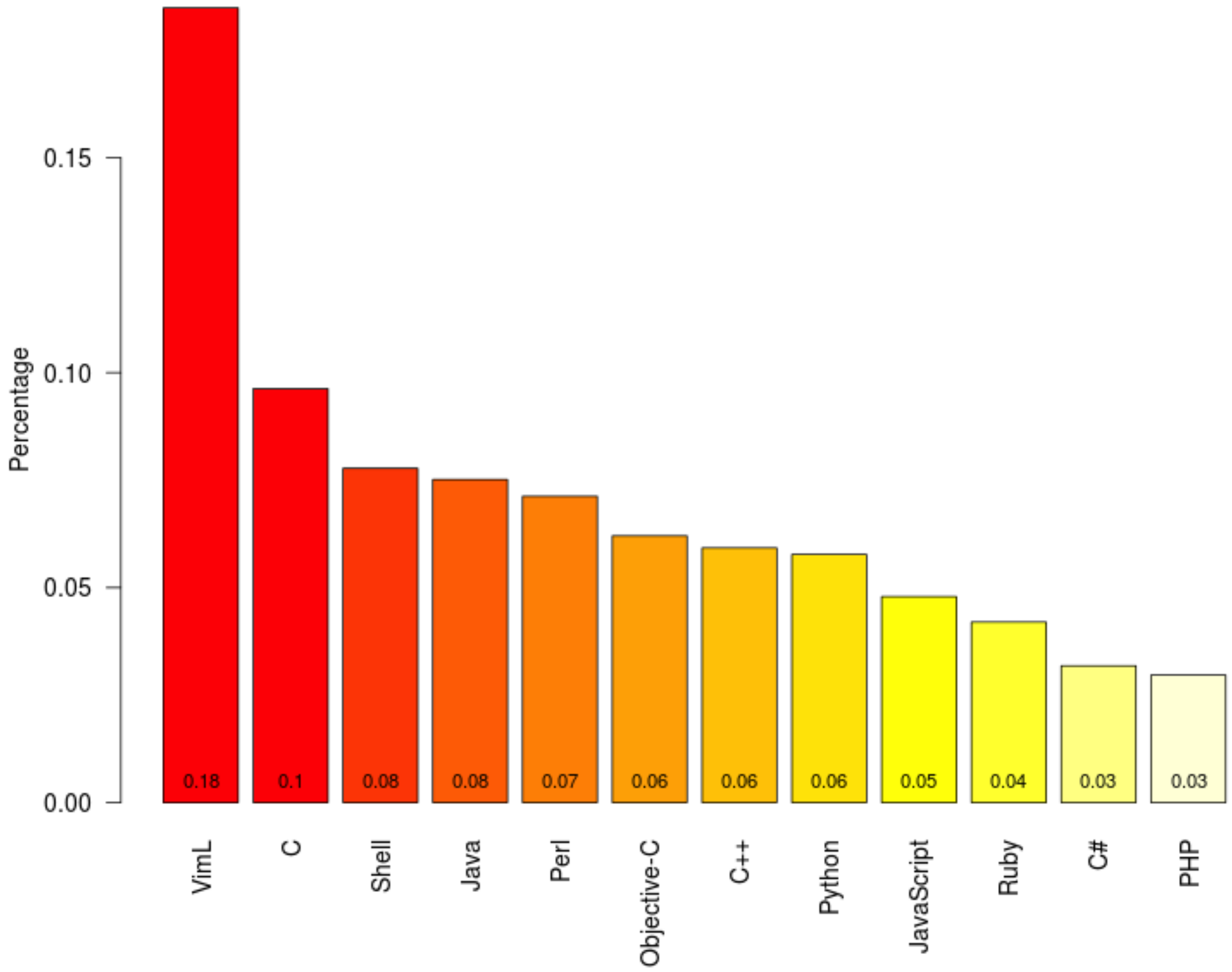
Emotional impact ... example query for "joy"

```
SELECT repository_language, COUNT(*) as cntlang
  FROM [githubarchive:github.timeline]
  WHERE repository_language != ''
    AND payload_commit_msg != ''
    AND PARSE_UTC_USEC(created_at) < PARSE_UTC_USEC('2012-05-09 00:00:00')

    AND REGEXP_MATCH(payload_commit_msg,
          r'(?i)\b(yes|yay|hallelujah|hurray|bingo|amused|cheerful|excited|glad|proud)\b')

GROUP BY repository_language
ORDER BY cntlang DESC
```

Table-scans for the win!

https://github.com/yaph/gh-emotional-commits

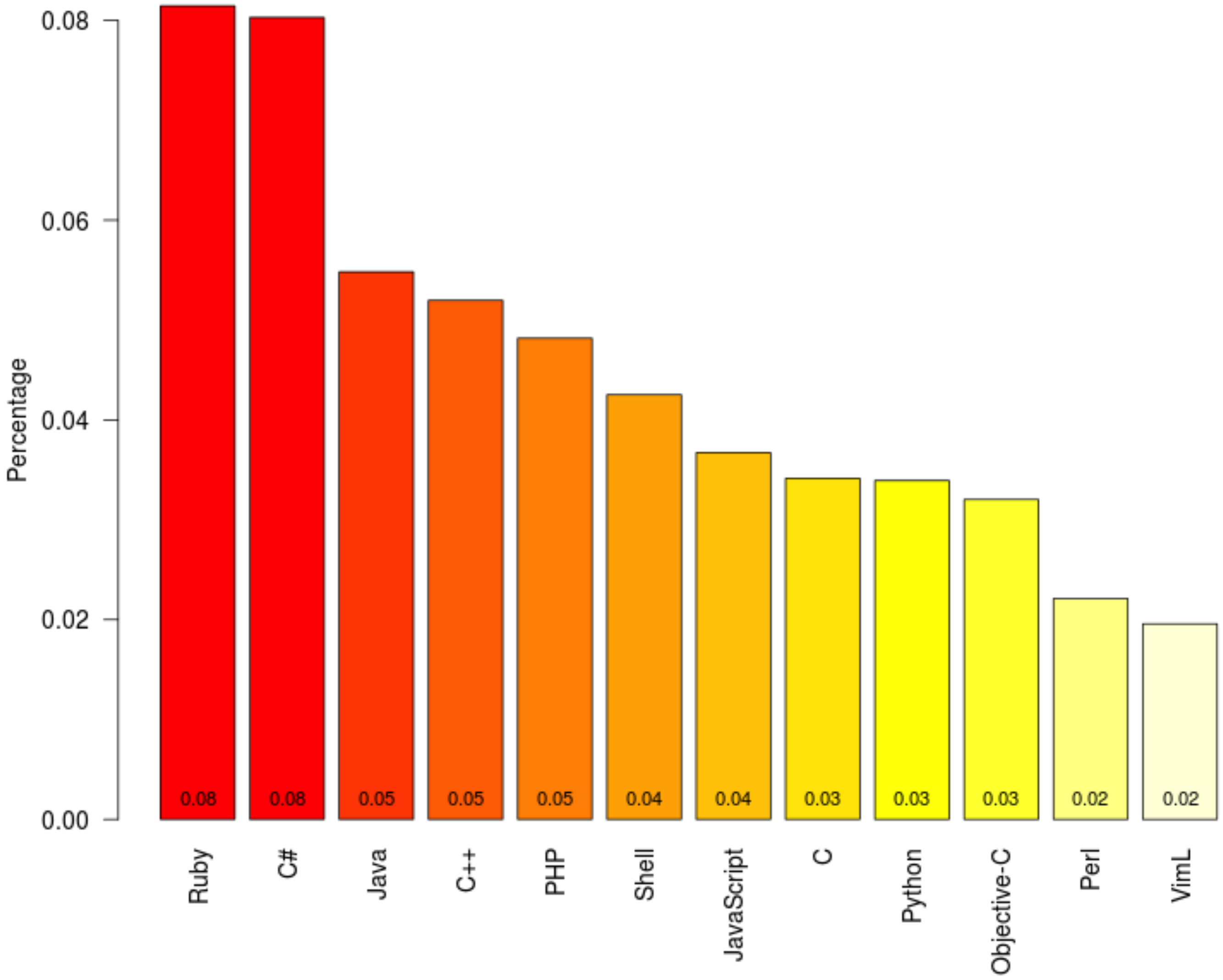
Emotional impact: anger



- VimL takes the top spot
- **C** makes more people angry than **Java**? Interesting!
- Python makes more people angry than Ruby... *But we all knew that! :-)*

http://geeksta.net/geeklog/exploring-expressions-emotions-github-commit-messages/

Emotional impact: amusement



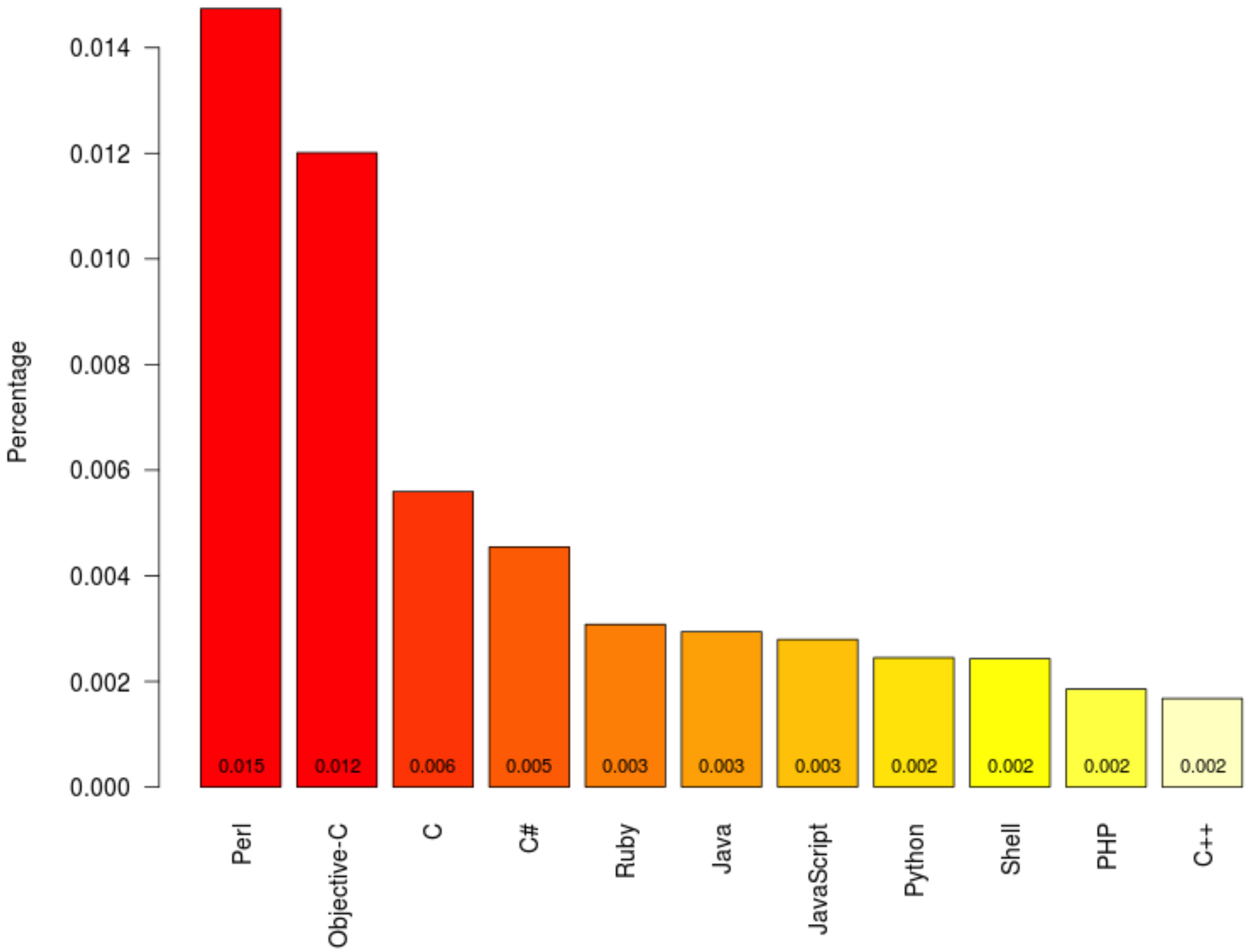
- **Ruby** takes #1
- What's so amusing about **C#???** :)

Regexp:

(?i)\b(ha(ha)+|he(he)+|lol|rofl|lmfao|lulz|lolz|rotfl|lawl|hilarious)\b

http://geeksta.net/geeklog/exploring-expressions-emotions-github-commit-messages/

Emotional impact: surprise



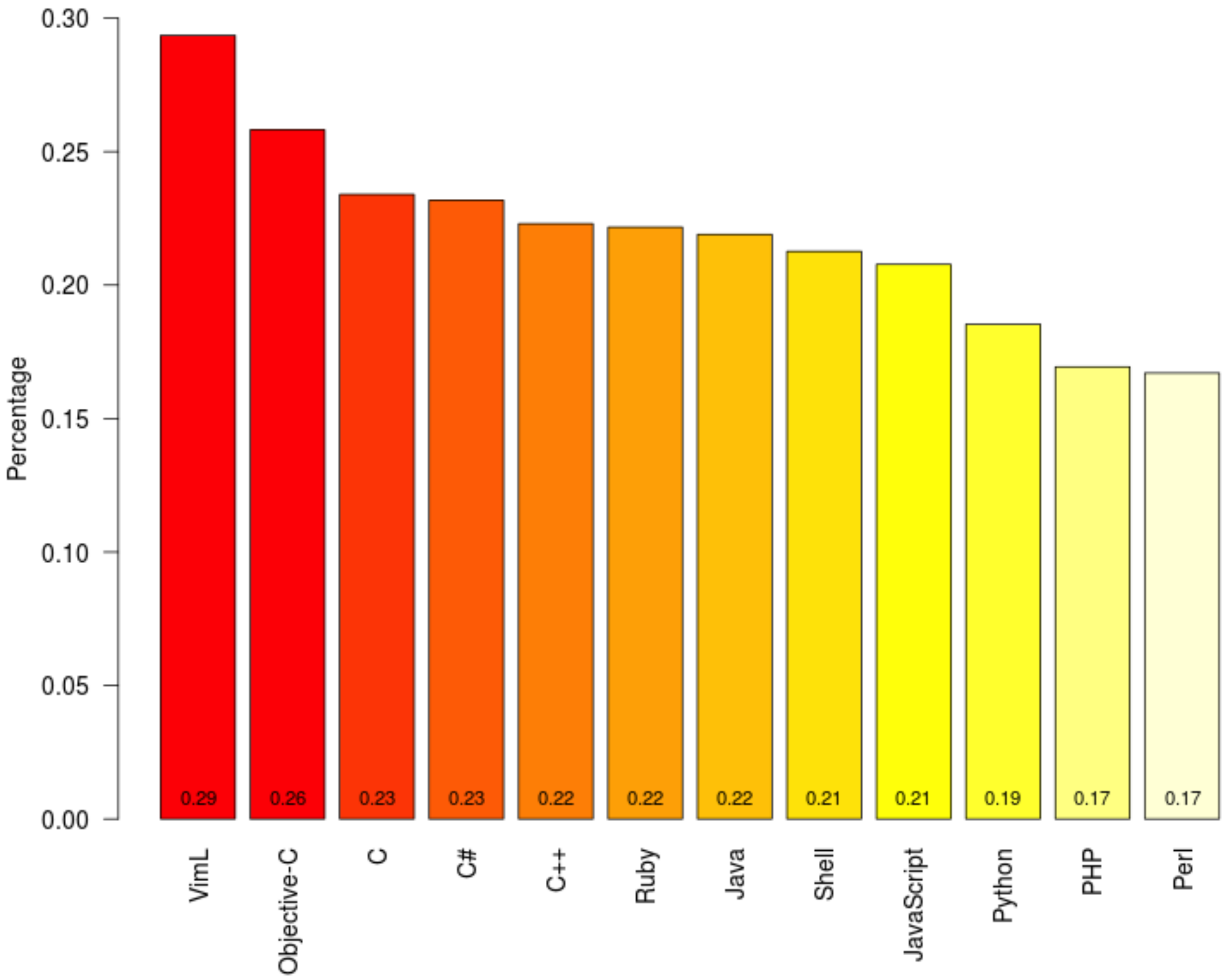
- **Perl, of course…**
- Or, if it has a ***/C/*** as part of the name

Regexp:

(?i)\b (yikes|gosh|baffled|stumped|surprised|shocked)\b

http://geeksta.net/geeklog/exploring-expressions-emotions-github-commit-messages/

Emotional impact: swear word inducing...



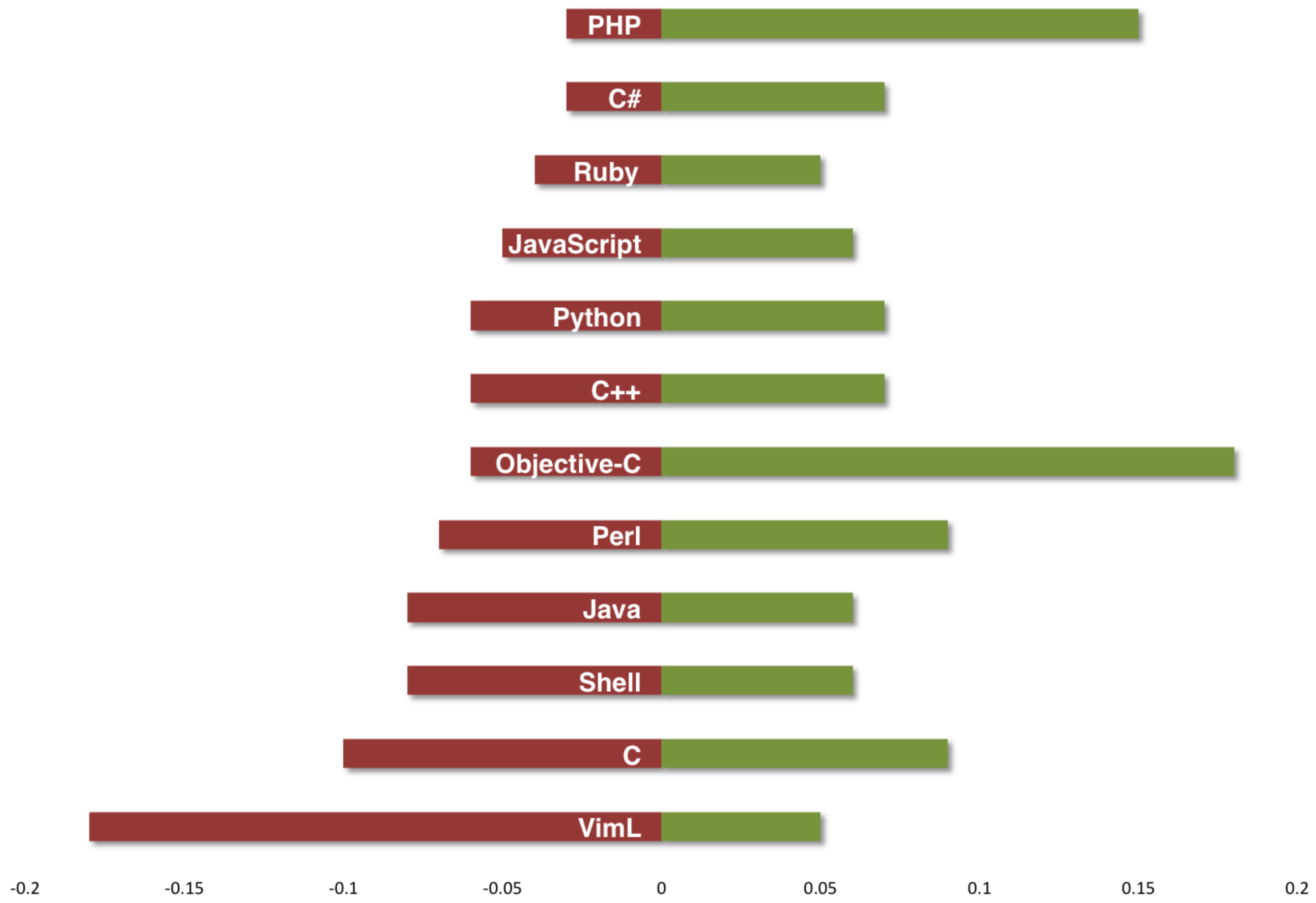
- If it has a ***/C/*** as part of the name, **it'll make you swear.**

Regexp:

(snip)   :-)

http://geeksta.net/geeklog/exploring-expressions-emotions-github-commit-messages/

Emotional impact: Anger vs. Joy

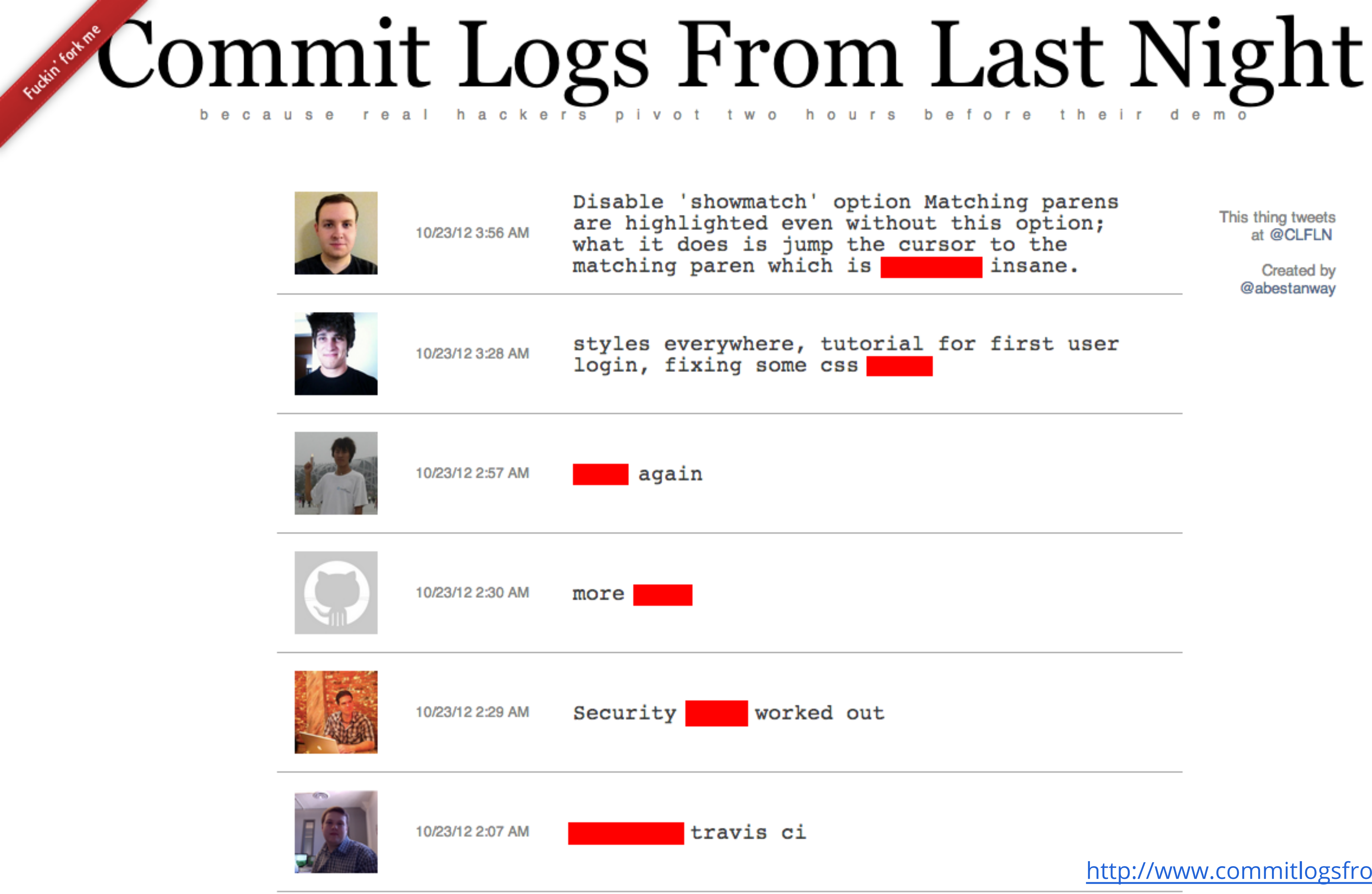


How do they stack up?

- **PHP, Objective-C** and **C#** are net positive

- **Java, Shell** and **C** are fairly even while **VimL** is just bad news

@briandoll @igrigorik



Fuckin' fork me

Commit Logs From Last Night

because real hackers pivot two hours before their demo

10/23/12 3:56 AM	Disable 'showmatch' option Matching parens are highlighted even without this option; what it does is jump the cursor to the matching paren which is [redacted] insane.
10/23/12 3:28 AM	styles everywhere, tutorial for first user login, fixing some css [redacted]
10/23/12 2:57 AM	[redacted] again
10/23/12 2:30 AM	more [redacted]
10/23/12 2:29 AM	Security [redacted] worked out
10/23/12 2:07 AM	[redacted] travis ci

This thing tweets at @CLFLN

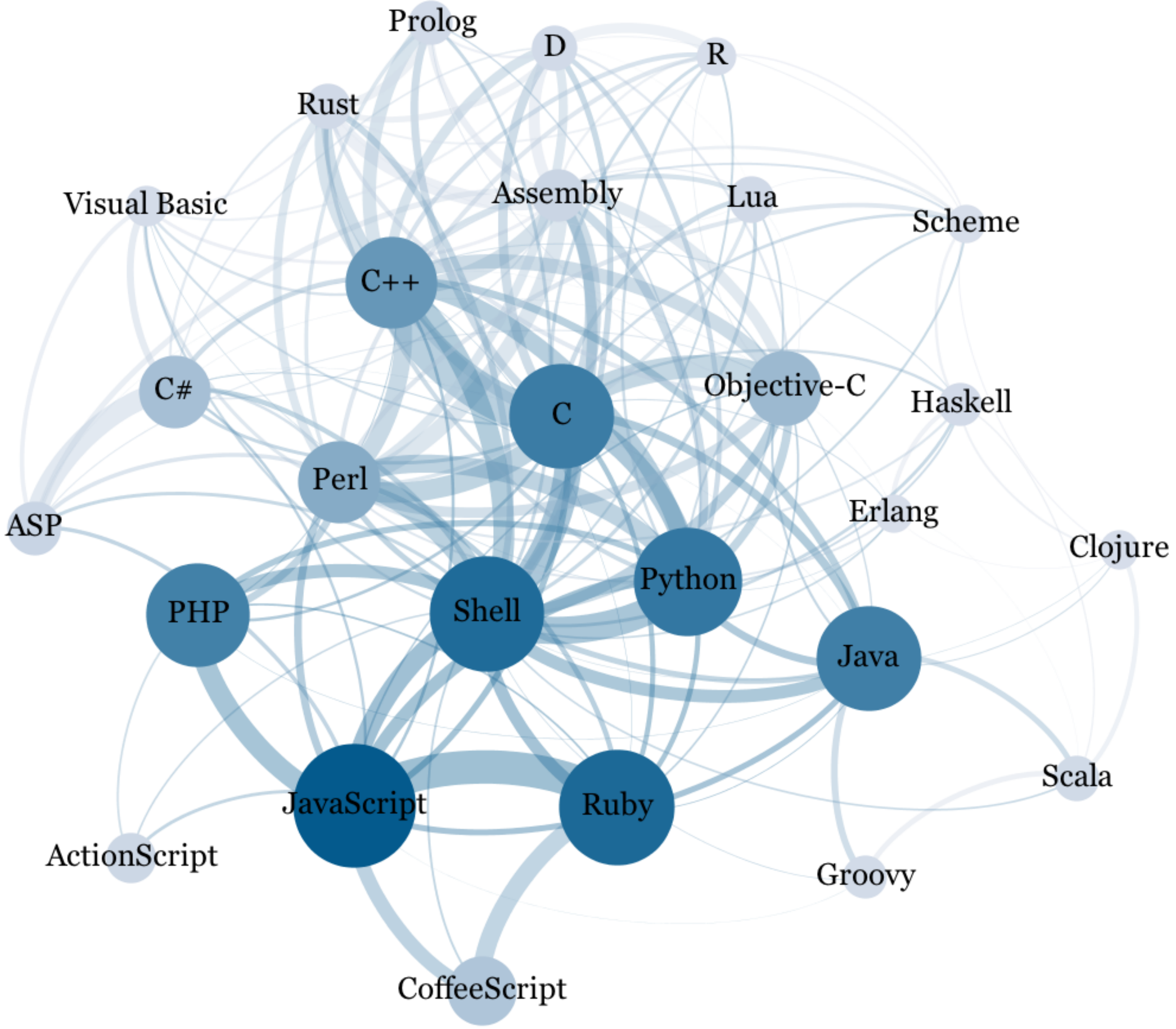
Created by @abestanway

http://www.commitlogsfromlastnight.com/

Programming language associations

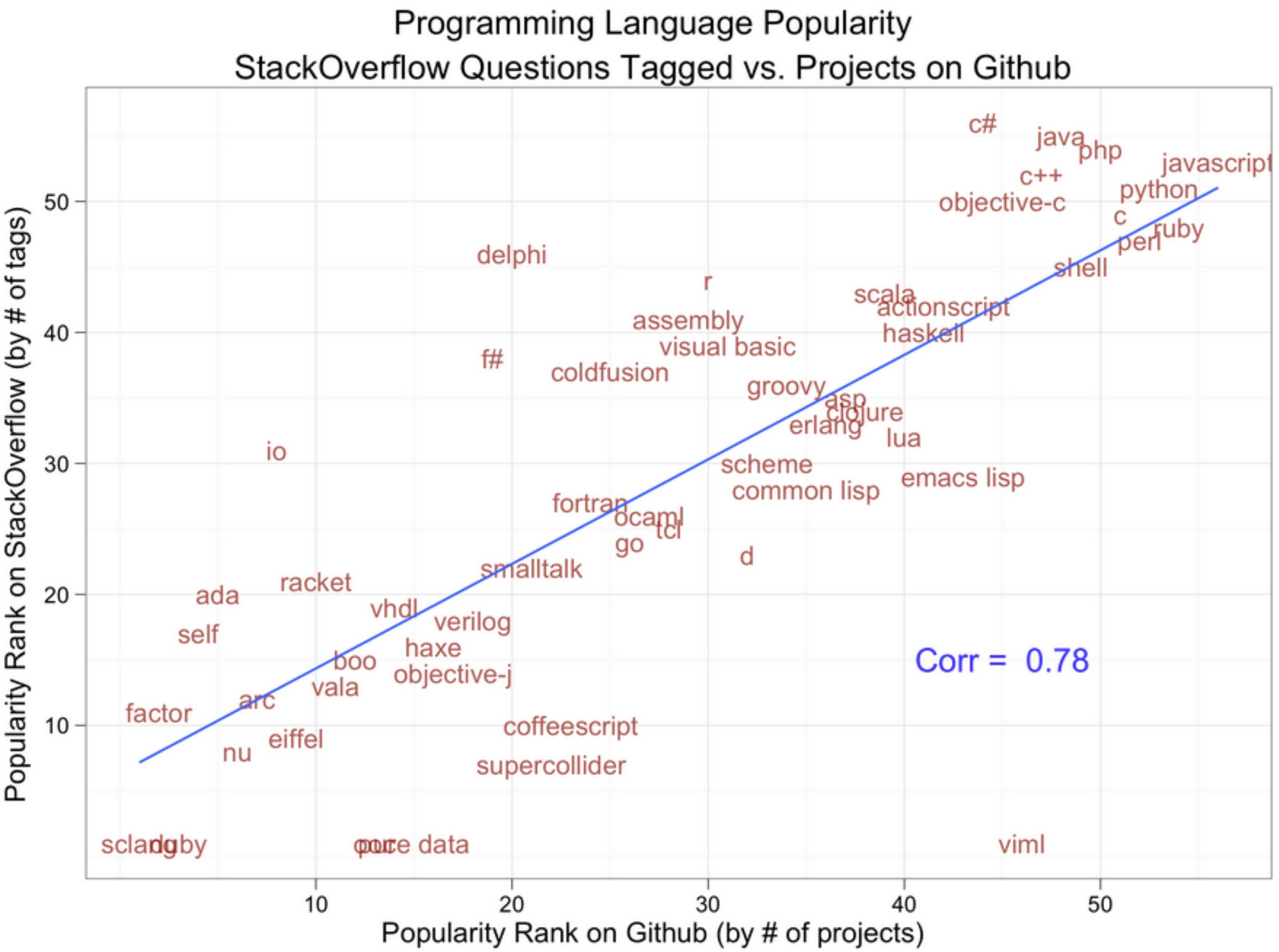
A **Ruby** programmer is ***very likely to know*** **JavaScript**, while a **Perl** programmer is not.

Java is a popular language, but stands primarily alone.



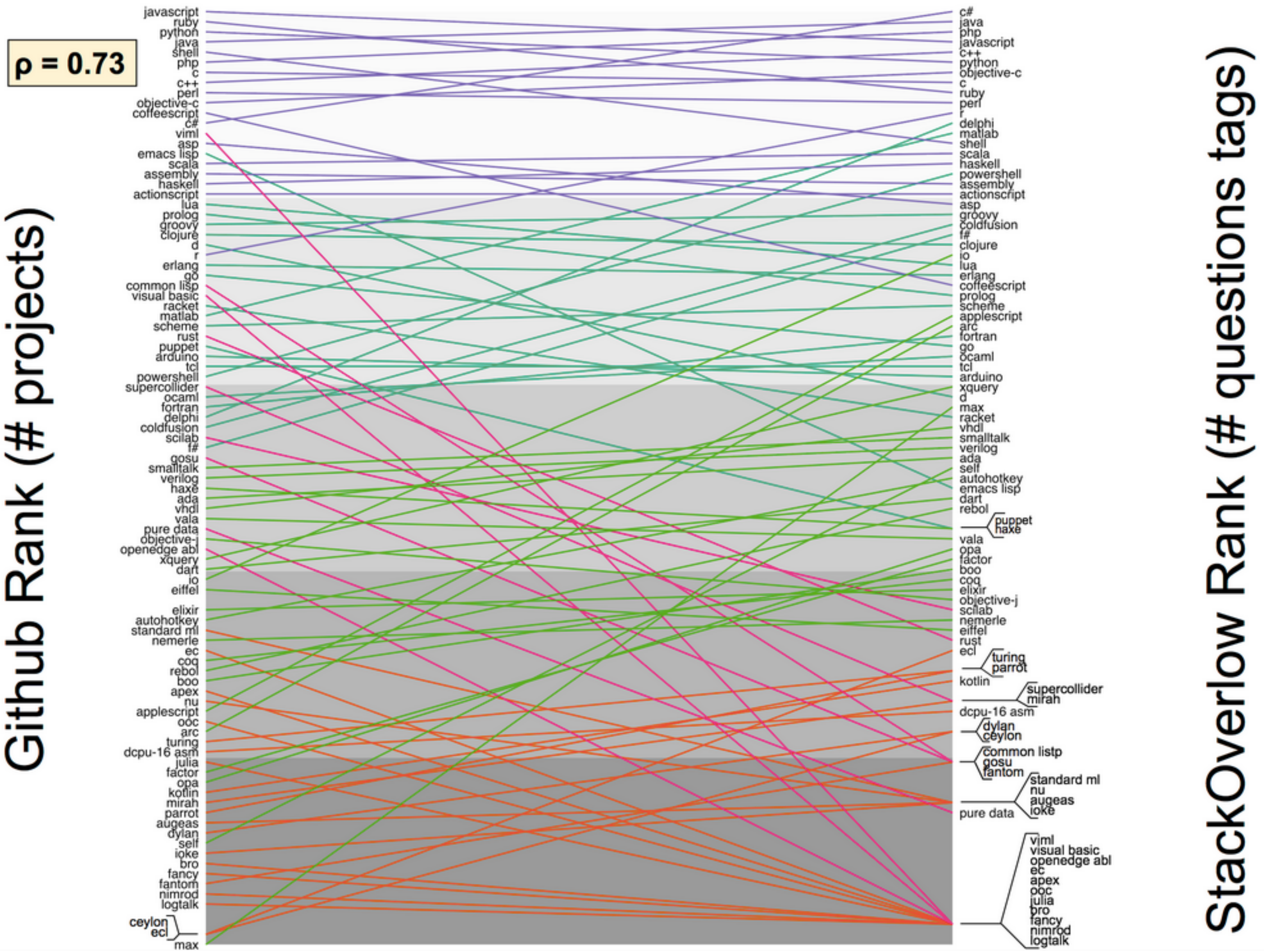
https://github.com/mjwillson/ProgLangVisualise

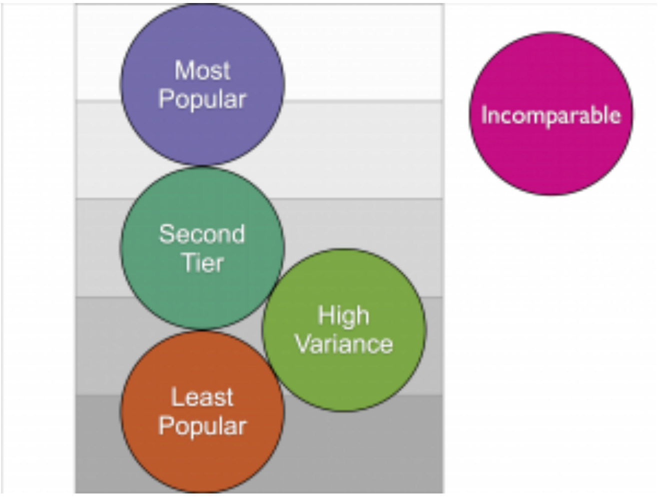
@briandoll @igrigorik



http://www.drewconway.com/zia/?p=2892

@briandoll @igrigorik

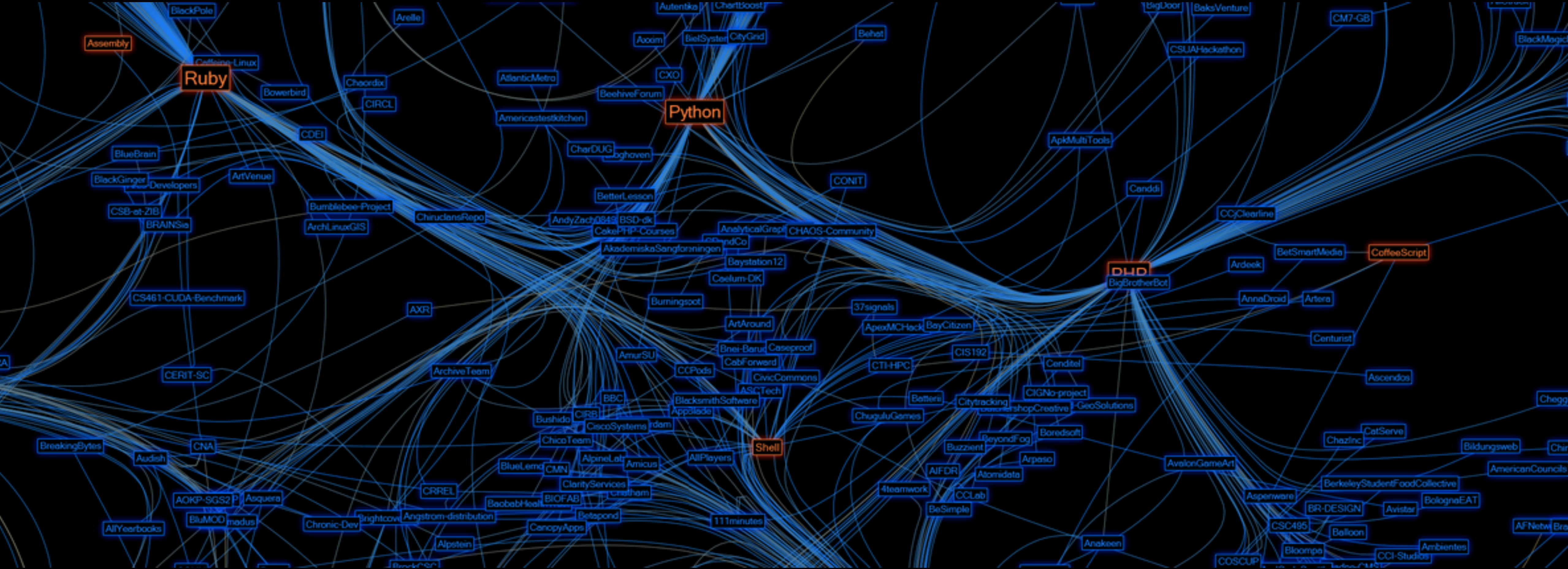




There is a lot of existing **VimL, common lisp** and **visual basic** code, but everyone is afraid to ask questions about them?

http://www.drewconway.com/zia/?p=2892

Repository activity by language

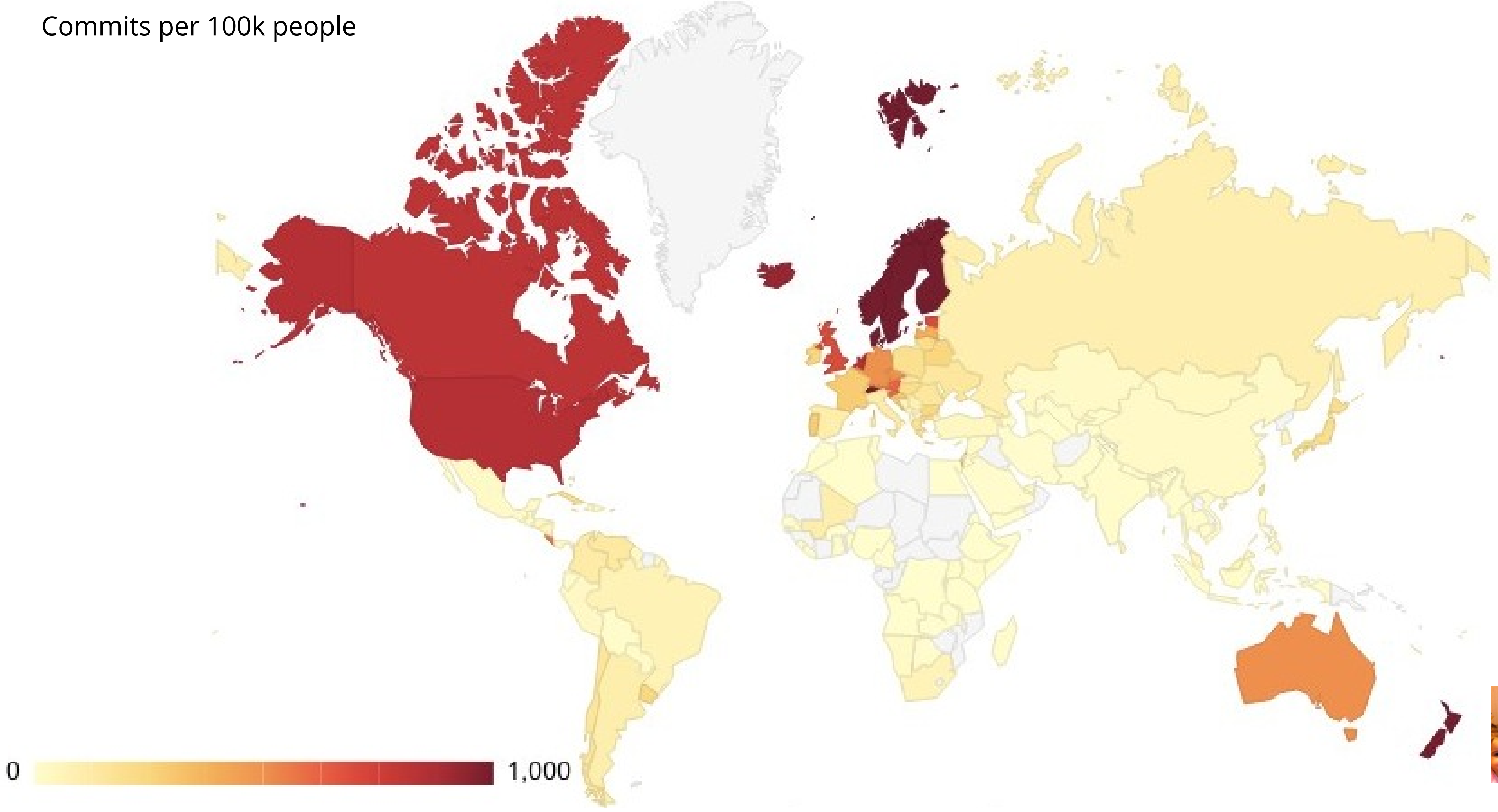


Mapping organizations with 250+ projects on GitHub to their respective programming languages

http://zoom.it/kCsU

GitHub activity by country

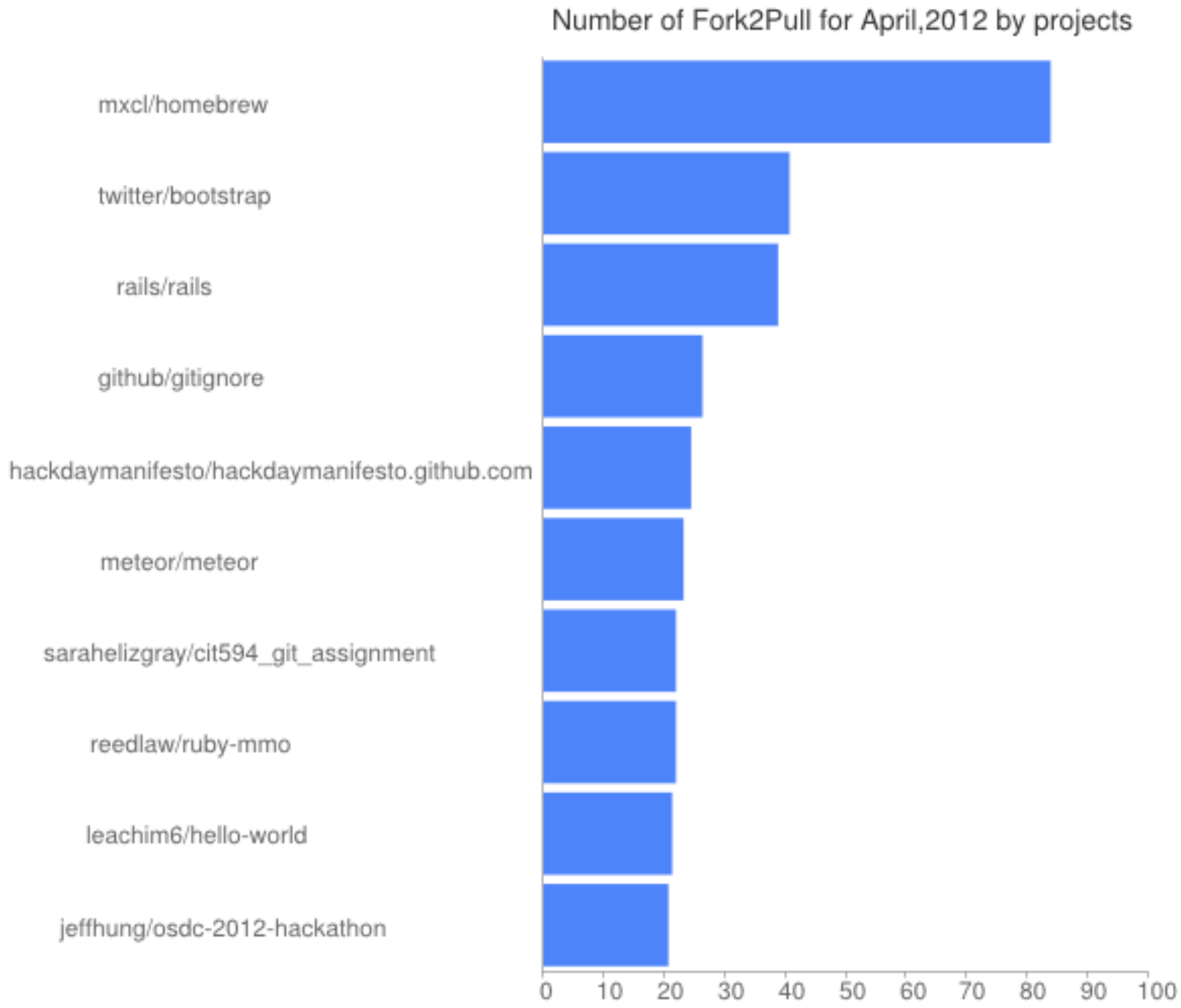
Commits per 100k people



http://bl.ocks.org/2727882

@briandoll @igrigorik

Projects using the fork to pull paradigm...



1. homebrew
2. bootstrap
3. rails
4. gitignore
5. ...

Jean-Noël Avila

jnavila

France

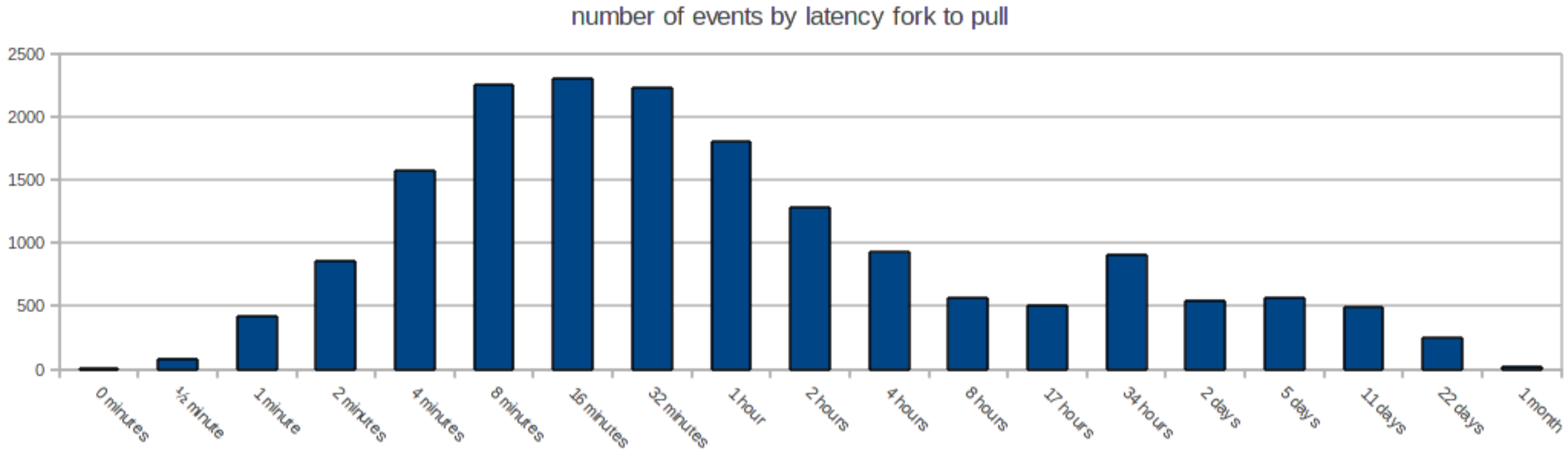
jn.avila@free.fr

http://aviblog.free.fr

Joined on Nov 20, 2009

https://gist.github.com/2623537

Pull request latency!



- 50%+ pull requests come in within **1 hour** of the fork
- 80%+ pull requests come in within **1 day** of the fork

1/2 minute? Spelling mistakes, etc!

https://gist.github.com/2623537

@briandoll @igrigorik

Pull request latency: the query...

```
SELECT
 COUNT(DISTINCT ForkTable.url) AS f2p_number,
 FLOOR(LOG2((PARSE_UTC_USEC(PullTable.created_at)-PARSE_UTC_USEC(ForkTable.created_at))/30000000)) AS f2p_interval_log_2_minute        4
FROM
 (SELECT
   url,
   repository_url,
   repository_language,
   MIN(created_at) AS created_at                                         1
  FROM
   [githubarchive:github.timeline]
  WHERE type='ForkEvent'
  AND PARSE_UTC_USEC(created_at) >= PARSE_UTC_USEC('2012-04-01 00:00:00')
  AND PARSE_UTC_USEC(created_at) < PARSE_UTC_USEC('2012-05-01 00:00:00')
  GROUP BY
   repository_language,
   repository_url,
   url)
 AS ForkTable
 INNER JOIN

 (SELECT
   ... )          2

 AS PullTable
 ON
  ForkTable.repository_url=PullTable.repository_url AND                          3
  ForkTable.url=PullTable.payload_pull_request_head_repo_html_url
 WHERE PARSE_UTC_USEC(PullTable.created_at)>PARSE_UTC_USEC(ForkTable.created_at)
 GROUP BY
  f2p_interval_log_2_minute
ORDER BY
 f2p_interval_log_2_minute ASC
```

https://gist.github.com/2623537#file_fork2_pull_request_by_latency.sql

Creating a Shared Understanding of Testing Culture on a Social Coding Site

Leibniz Universität Hannover & Universidade Federal do Rio Grande do Norte

Does the eye of the public make for better and well tested code?

Just by watching how other, more senior project members behave, they learn what a good commit looks like.

Infrastructure with low barriers seems to be very important in getting developers to test their contributions.

Because contribution has become so easy, project owners reported seeing what they called ***drive-by commits***.

http://blog.leif.me/2012/09/github-testing/

Research: Analysis of OSS development using DNA sequencing tools

by Aron Lindberg and Tim Henderson at Case Western Reserve University

What is the "social DNA"
of successful open source projects?

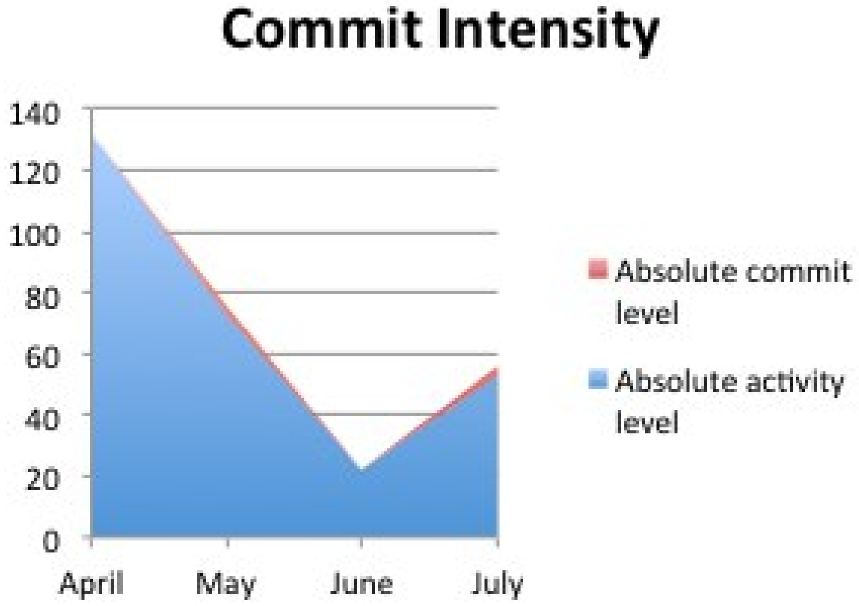
Research: Analysis of OSS development using DNA sequencing tools

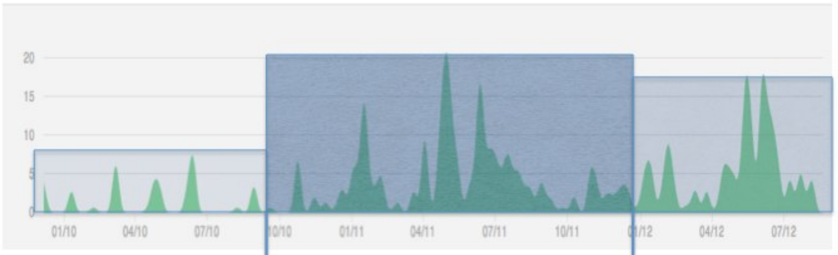
by Aron Lindberg and Tim Henderson at Case Western Reserve University

Overall activity levels are tightly coupled with commit levels

Success breeds success; i.e. communities that are growing or declining are likely to continue the trajectory that they have started (An object in motion...)

Don't ignore those who commit infrequently or only report bugs: growing a leadership pipeline through quickly establishing a broad base of developers supports long-term success





e.g.:
- Start-up phase
- Low level of variation
- Low commit intensity
- Community increases slowly

e.g.:
- Intensive development phase
- High level of variation
- High commit intensity
- Rapid increases in community

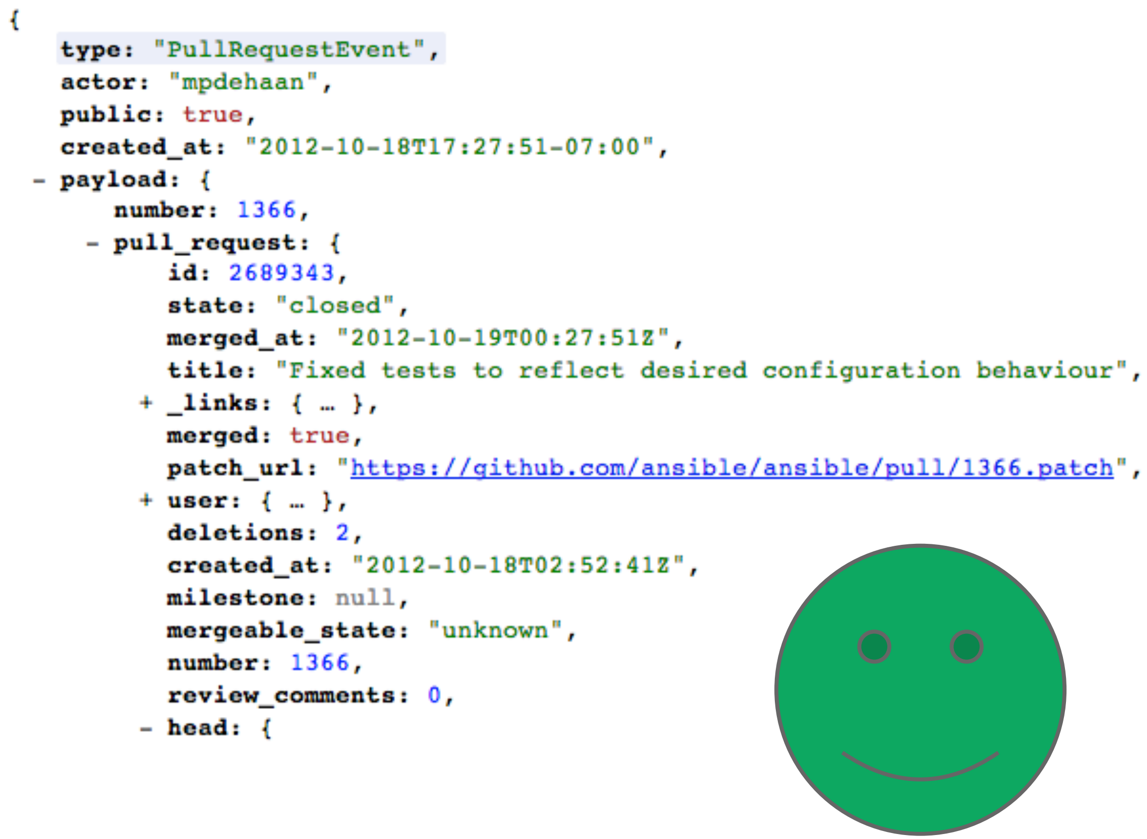
e.g.:
- Maintenance phase
- Low level of variation
- Low commit intensity
- Increases in community level out

NEW!!!

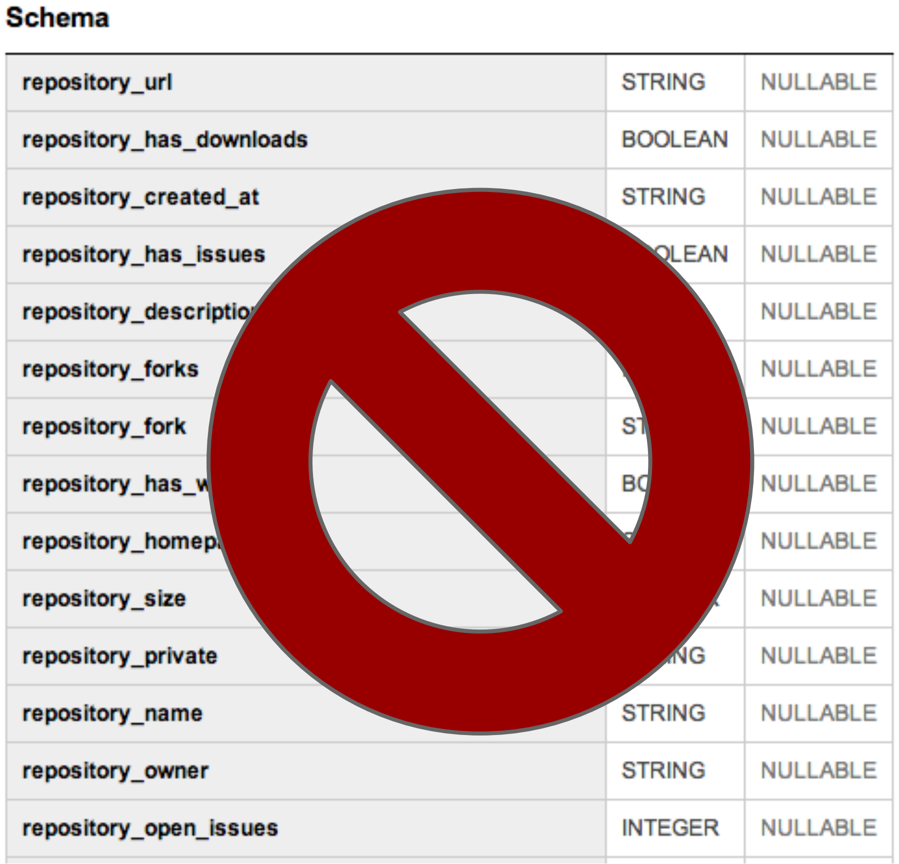
Moar & Better Data!

Import in progress...

SELECT expr1 WITHIN RECORD, expr2 WITHIN node_name...



```
{
    type: "PullRequestEvent",
    actor: "mpdehaan",
    public: true,
    created_at: "2012-10-18T17:27:51-07:00",
  - payload: {
        number: 1366,
      - pull_request: {
            id: 2689343,
            state: "closed",
            merged_at: "2012-10-19T00:27:51Z",
            title: "Fixed tests to reflect desired configuration behaviour",
          + _links: { … },
            merged: true,
            patch_url: "https://github.com/ansible/ansible/pull/1366.patch",
          + user: { … },
            deletions: 2,
            created_at: "2012-10-18T02:52:41Z",
            milestone: null,
            mergeable_state: "unknown",
            number: 1366,
            review_comments: 0,
          - head: {
```



Schema

repository_url	STRING	NULLABLE
repository_has_downloads	BOOLEAN	NULLABLE
repository_created_at	STRING	NULLABLE
repository_has_issues	OLEAN	NULLABLE
repository_descriptio		NULLABLE
repository_forks		NULLABLE
repository_fork	ST	NULLABLE
repository_has_w	BO	NULLABLE
repository_homep		NULLABLE
repository_size		NULLABLE
repository_private	NG	NULLABLE
repository_name	STRING	NULLABLE
repository_owner	STRING	NULLABLE
repository_open_issues	INTEGER	NULLABLE

Support for nested (JSON) data in BigQuery! *New import in process...*

http://googledevelopers.blogspot.com/2012/10/got-big-json-bigquery-expands-data.html

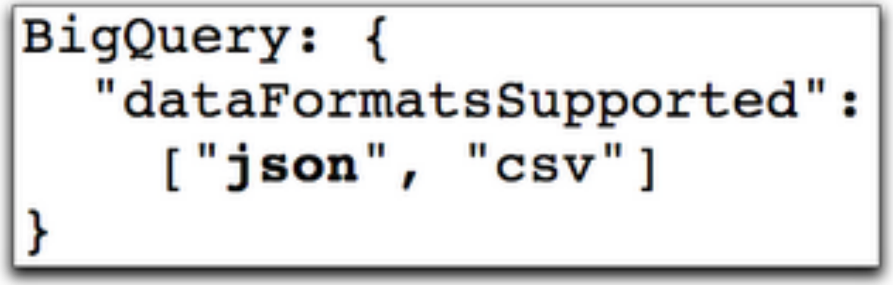
@briandoll @igrigorik

Kudos to GitHub....

Github Archive data now goes back to **Feb 12, 2011**

- ○ **Feb 12, 2011 - Now!**

- **Raw JSON data for 2011:**
  - ○ `wget http://data.githubarchive.org/201{1,2}-{01.12}-{01..31}-{0..23}.json.gz`



```
BigQuery: {
   "dataFormatsSupported":
      ["json", "csv"]
}
```

SELECT questions FROM audience

Brian Doll	briandoll@github.com	@briandoll
Ilya Grigorik	igrigorik@google.com	@igrigorik