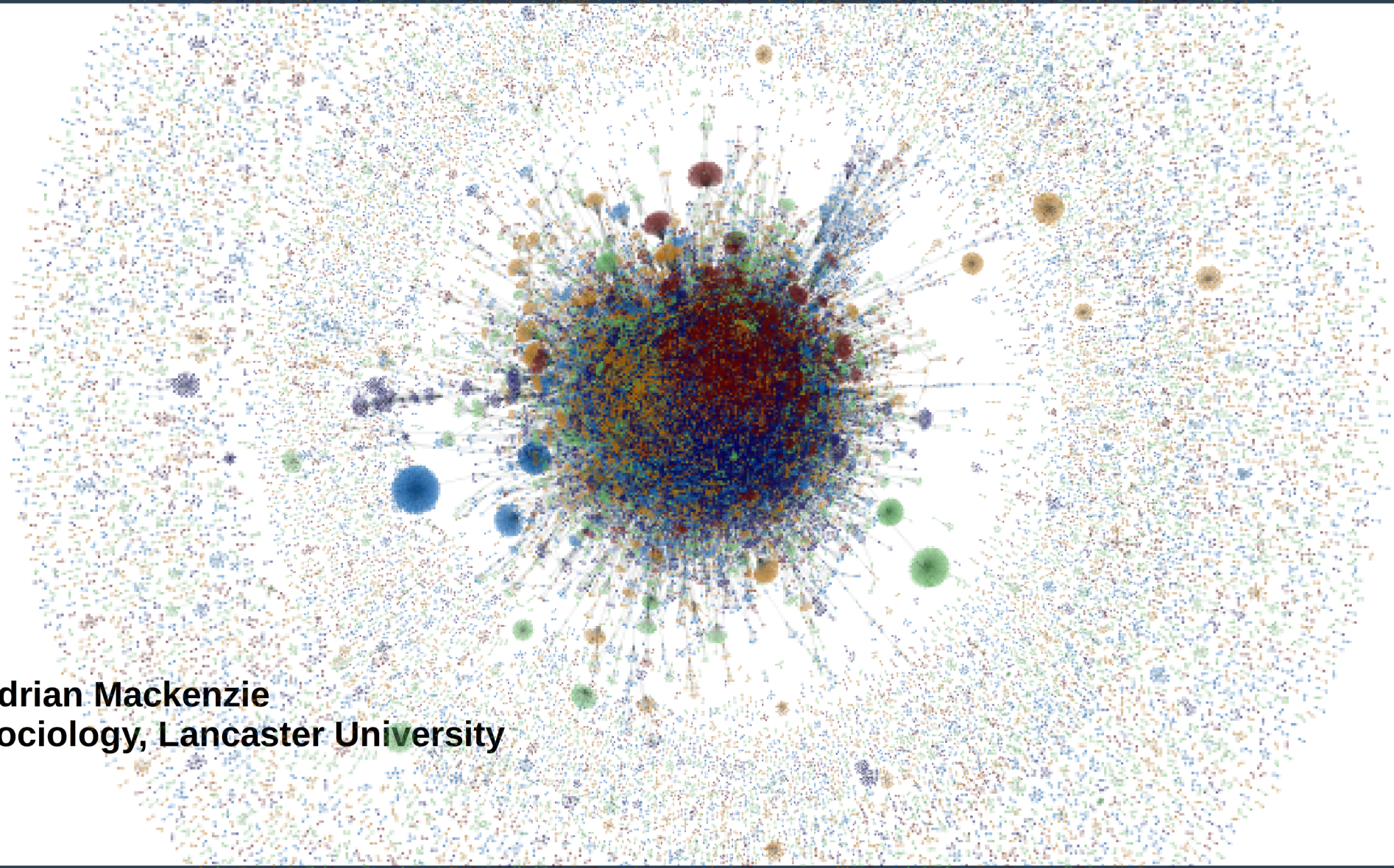
Many names and large numbers: methods for imitative fluxes in the data-as-raw-material imaginary?



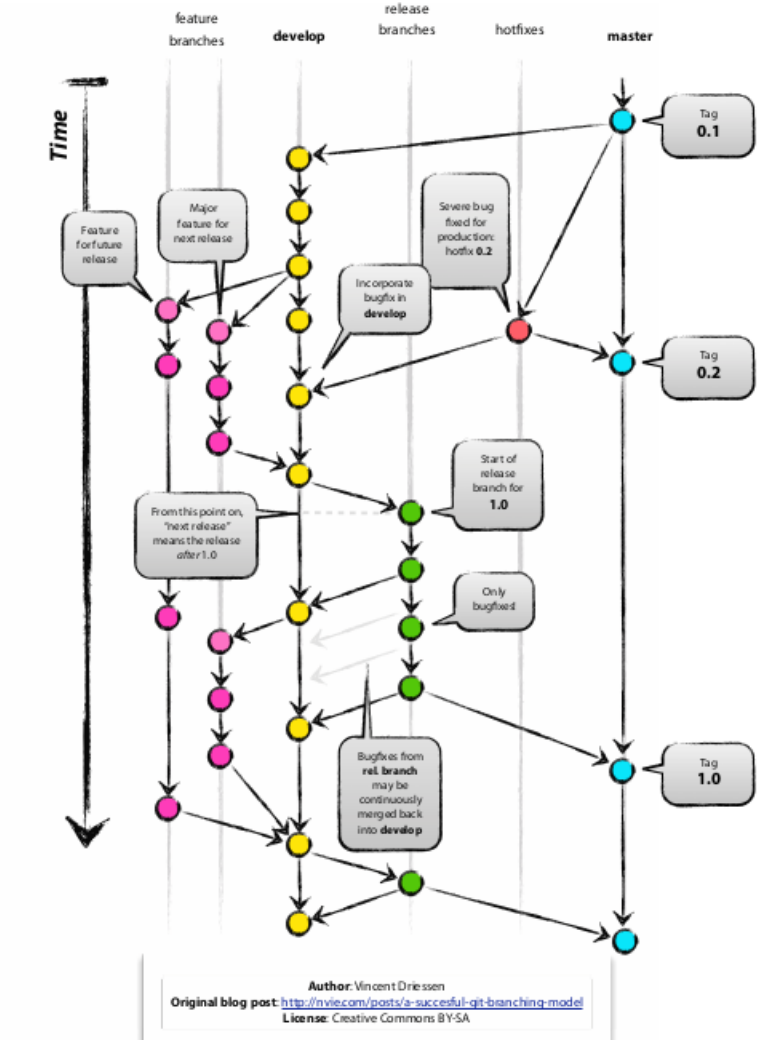
Adrian Mackenzie
Sociology, Lancaster University

Overview: Github case

- **Github as *social media platform* that hosts ~20 million code repositories;**
  - Compare earlier accounts of coding (Kelty – recursive publics; Coleman – coding freedom)
  - Github code as symptom of and infrastructure for burgeoning post-social attachments;
- **Github as example of *large number (N=ALL)***
  - the contemporary political economies of counting and graphing the many or the *N = ALL*
  - obstacles to and potentials for number-based ethnographic sensibilities;
  - how to re-count and make large numbers
  - What work numbers makes us do;

Part A: Post-social attachments in code

Github: post-social attachments in collective coding



"Since these markets are exteriorized and concentrated on screen, traders not only participate in these markets, they relate to them as a complex 'other' with which they are strongly, even obsessively, engaged. The term 'postsocial relationships' refers to new kinds of bonds such as those constructed between humans and objects."

Knorr-Cetina, "Traders' Engagement with Markets. A Postsocial Relationship, " 2002, 162

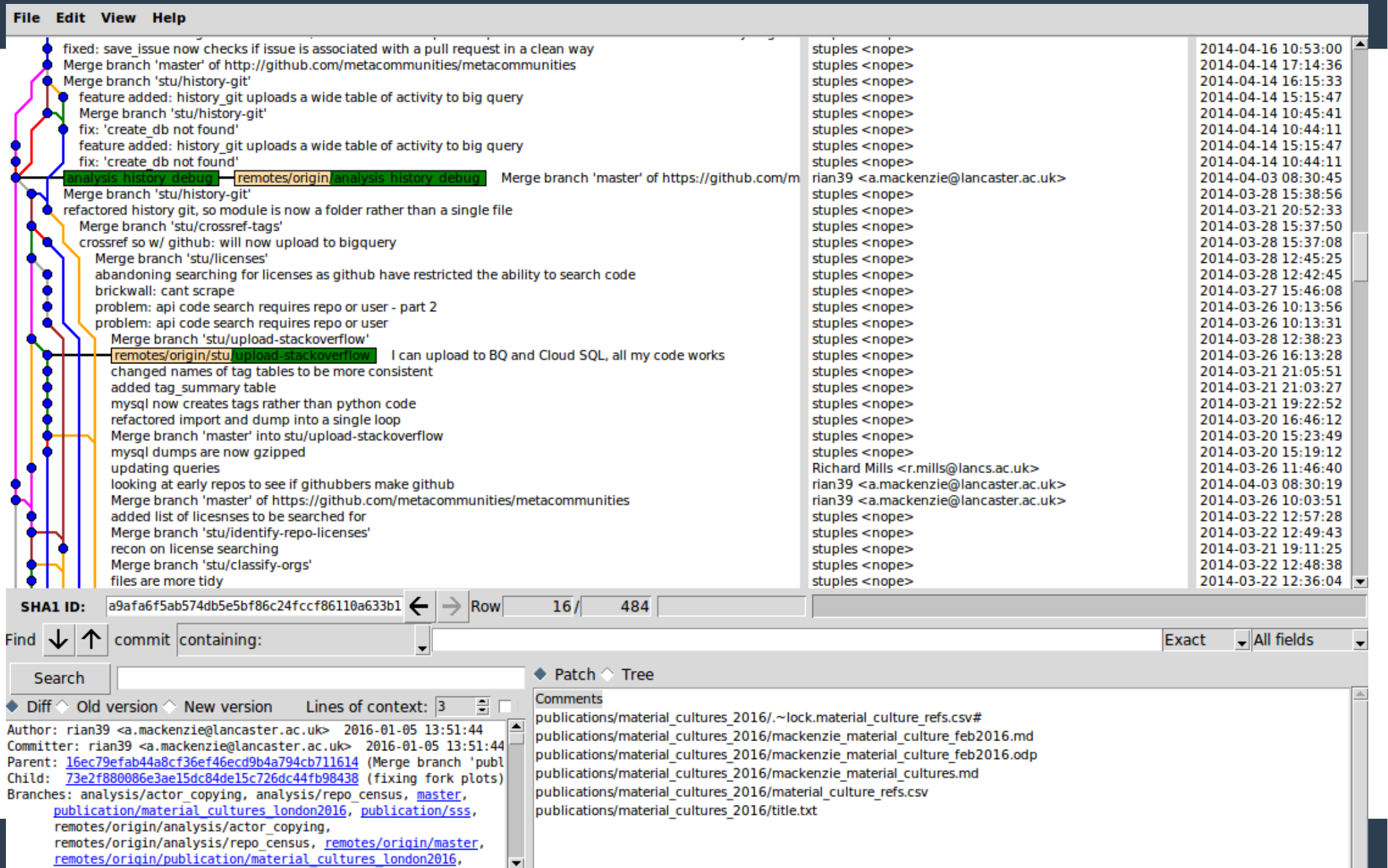
"Since these repositories are exteriorized and concentrated on screen, coders not only participate in these repositories, they relate to them as a complex 'other' with which they are strongly, even obsessively, engaged. The term 'postsocial relationships' refers to new kinds of bonds such as those constructed between humans and objects."

Mackenzie, "Coder's Engagement with Code Repositories',

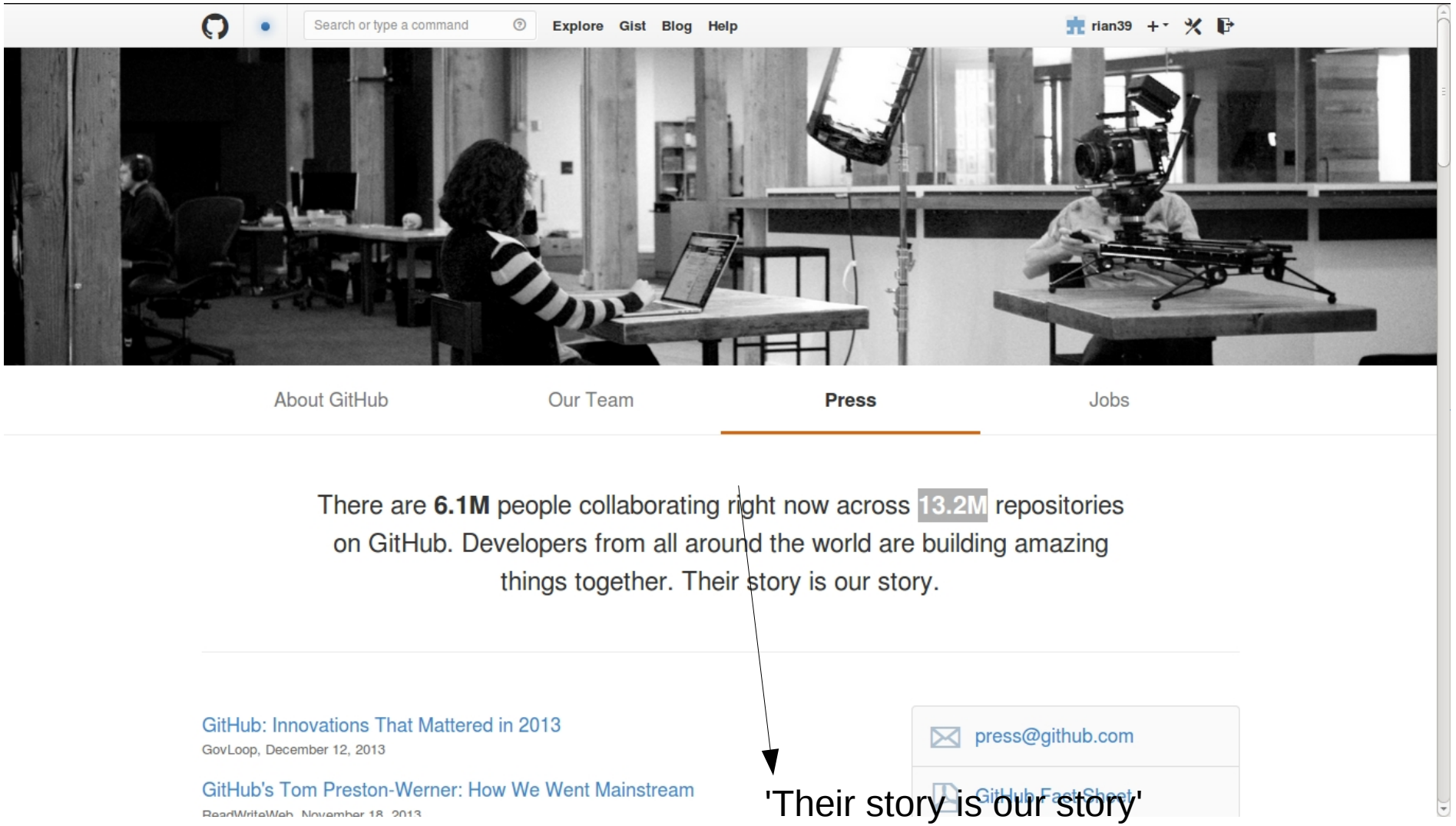
~

~

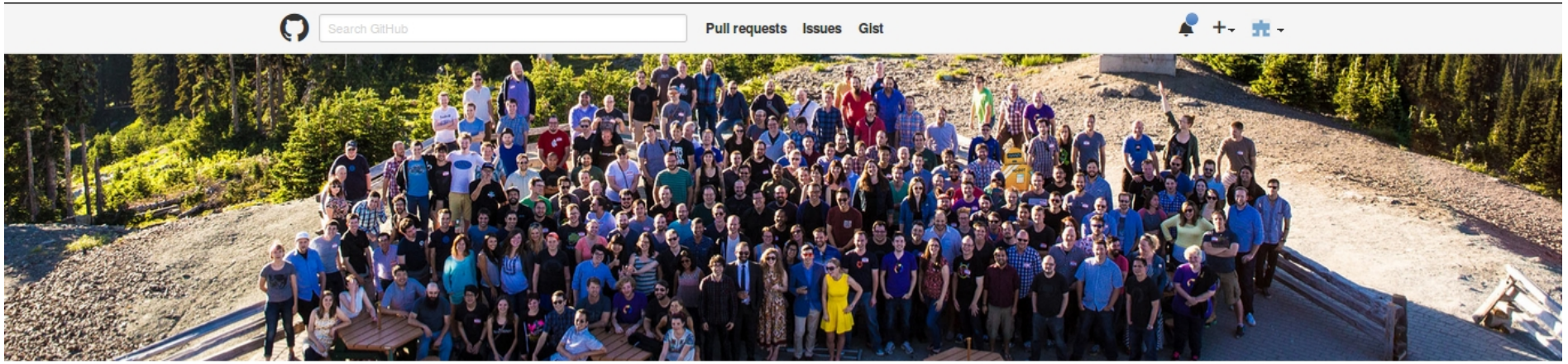
Ethnography also becoming post-social?



November 2013: Github founder resigns after sexual harrassment claims



November 2013



About GitHub | Our Team | Press | Jobs

GitHub is how people build software. With a community of more than 11 million people, developers can discover, use, and contribute to over 29 million projects using a powerful collaborative development workflow.

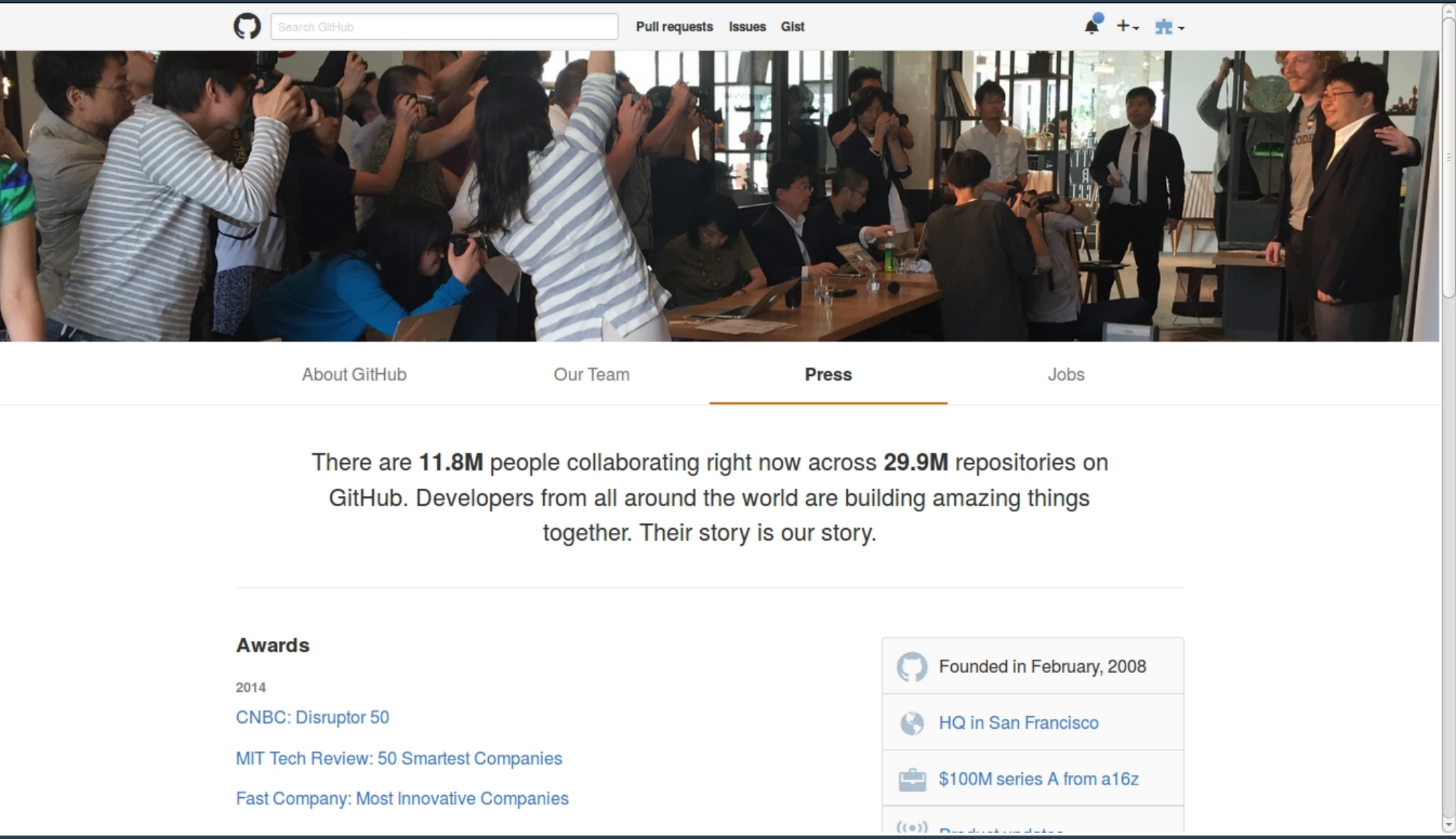
Whether using GitHub.com or your own instance of GitHub Enterprise, you can integrate GitHub with third party tools, from project management to continuous deployment, to build software in the way that works best for you.

Follow us on Twitter

© 2015 GitHub, Inc. Terms Privacy Security Contact Help

Status API Training Shop Blog About Pricing

November 2015: China denial of service attack



Part B: Data-as-raw-material imaginary*

* imaginary: unstable cluster of statements/articulations whose relations concern sense of whole + sensible forms projected as densities, boundaries, or diagrammatic connection

Data-as-raw-material imaginary?

Introduction

The ESRC and Google are pleased to announce the Google Data Analytics Social Science Research Call.

Data is the new raw material of the 21st century. It allows citizens to hold governments to account, drives improvements in public services by informing choice, and provides a feedstock for innovation and growth. As open-source data and data integration grows, it is a key time to better understand how it maps onto and possibly significantly strengthens, the ability of academics to understand society.

Background

The ESRC spends between £15-20 million per year on the collection and curation of data, providing access to a wide variety of data resources and developing the methodological tools and techniques for building and exploiting data resources. This has helped to create a world-leading data infrastructure. Much of this funding has been focused on the creation and support of large scale surveys and other aggregate statistics provided by other major agencies like the ONS. However, as the volume of less structured online data resources become increasingly available the ESRC has recognised the potential of capturing and connecting huge data flows in new ways to support conceptual, methodological and empirical advances in social and economic research.

We have been identifying key partners through which to develop the potential of publicly available online data. As a first step we are collaborating with Google to identify demand from the social science community in working with online data and to develop a suite of demonstrator projects which highlight its true value for better understanding society.

The problem(s) of numbers

In any practical going-on with numbers, what matters is that they can be *made* to work, and *making them work* is a politics. Yet this is a politics that completely evades conventional foundationist analysis (Verran, 2001: 88)

Data as raw material?

http://api.github.com/events

Event information

Actor information

Repository Information


```
{
1.   "id": "2111998059",
2.   "type": "WatchEvent",
3.   "actor": {
4.     "id": 1459103,
5.     "login": "mmemetea",
6.     "gravatar_id": "4532d1e4885f579ca7d9aa8748418817",
7.     "url": "https://api.github.com/users/mmemetea",
8.     "avatar_url": "https://avatars.githubusercontent.com/u/1459103?"
9.   },
10.  "repo": {
11.    "id": 14802742,
12.    "name": "OpenSensorsIO/azondi",
13.    "url": "https://api.github.com/repos/OpenSensorsIO/
14.  },
15.  "payload": {
16.    "action": "started"
17.  },
18.  "public": true,
19.  "created_at": "2014-05-23T08:40:56Z",
20.  "org": {
21.    "id": 5497318,
22.    "login": "OpenSensorsIO",
23.    "gravatar_id": "1e0218942846ec8ef59f5d679dbca7
24.    "url": "https://api.github.com/orgs/OpenSensorsIO",
25.    "avatar_url": "https://avatars.githubusercontent.com
  }
},
```


Azondi

Azondi is a data digestion service. A

3 Perspectives on data and its analysis

1) Latour & Venturini, 2012 'The Whole is Smaller than the Part: How **Digital Navigation** May Modify Social Theory.'

   – **traceability** frees social theory from macro-micro split

2) Marres, Noortje, 2013, 'The redistribution of methods: on **intervention** in digital social research, broadly conceived'

   - **device-specific research** must look at platform-data format **and** what overflows it

3) Gill, Ros, and Andy Pratt. 2008. 'In the Social Factory?: **Immaterial Labour**, Precariousness and Cultural Work'.

   - **symbolic-analytical work** is changing (including ours?)

GithubArchive.org collects it all

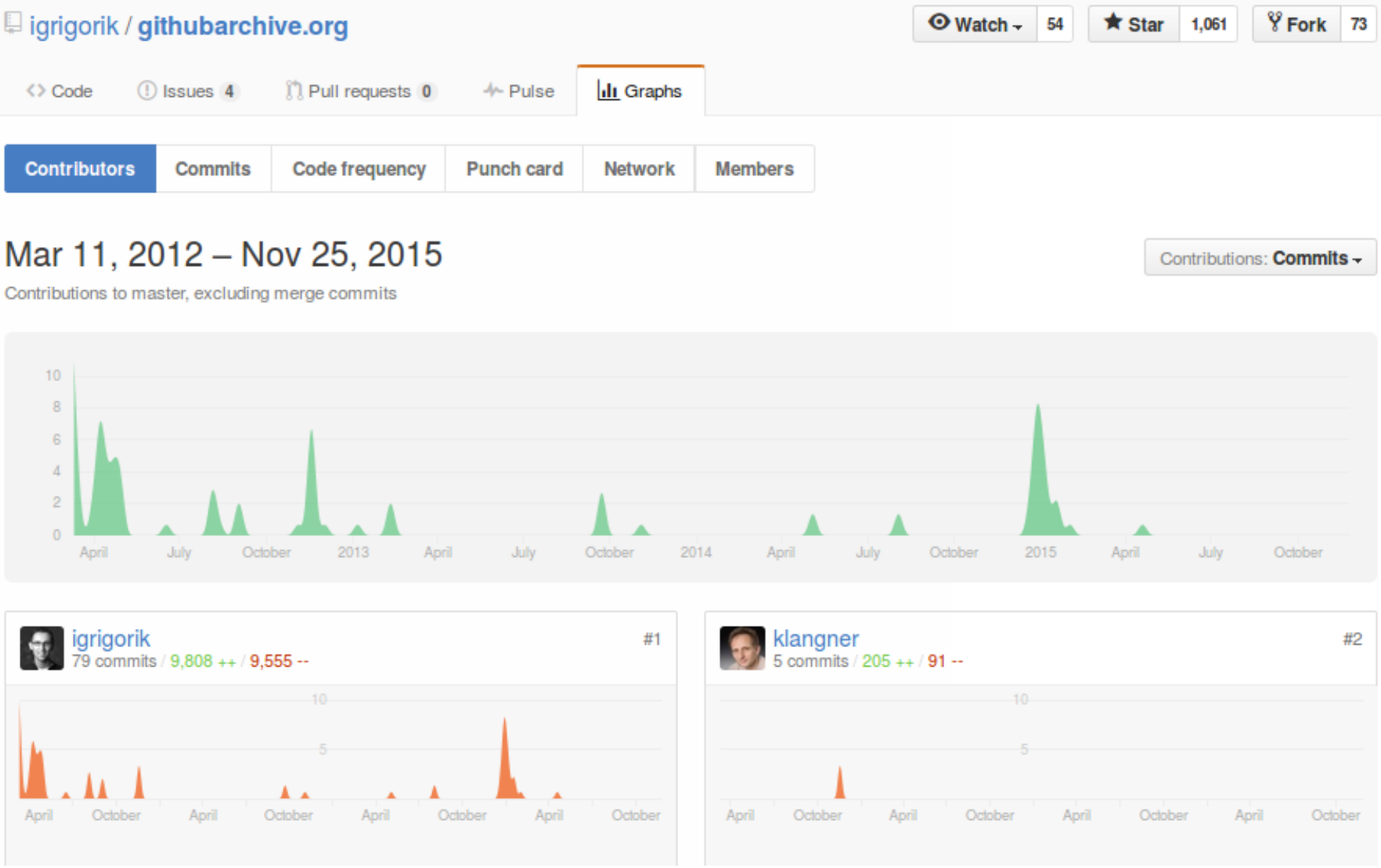
+

Analyzing Millions of GitHub Commits

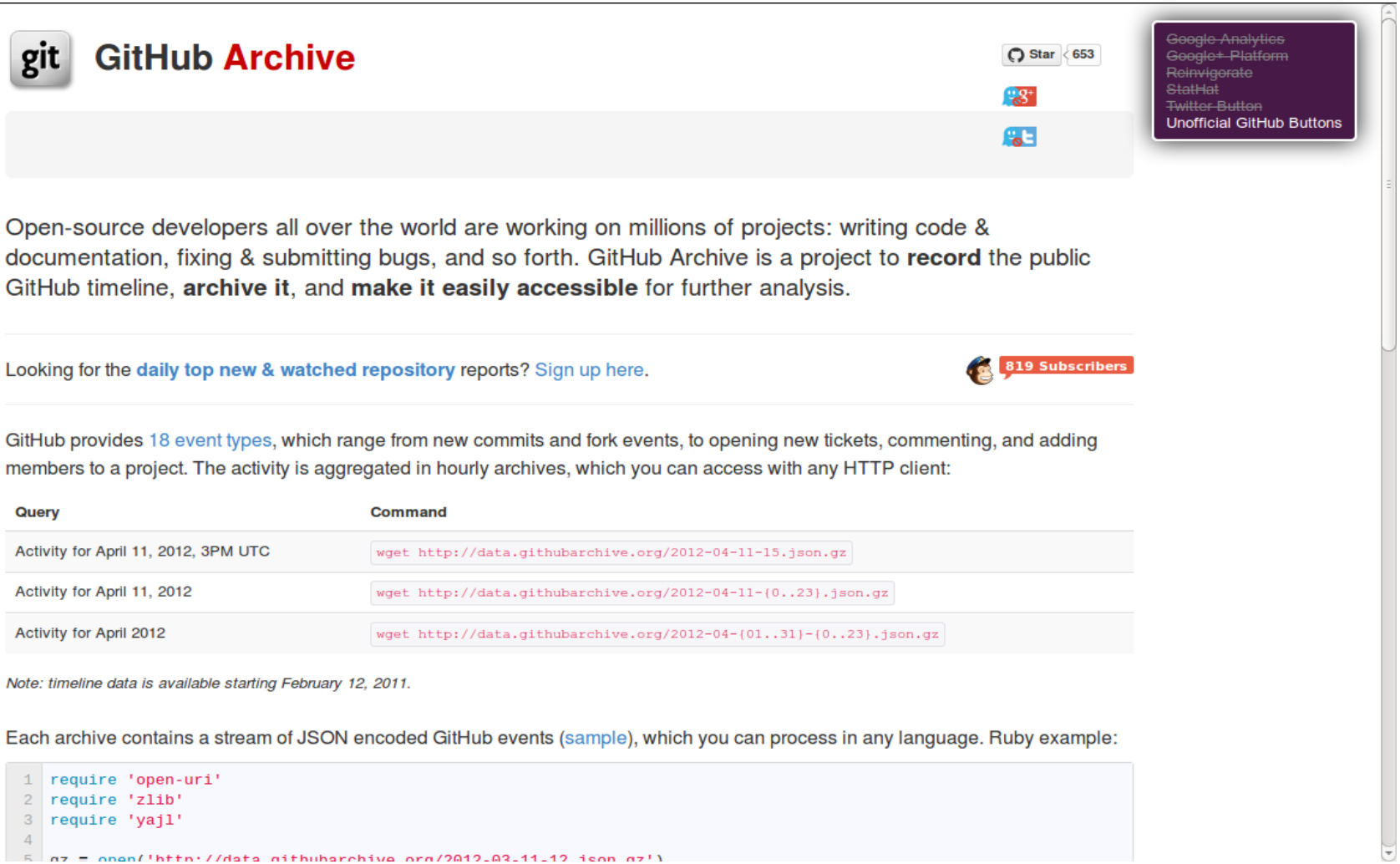
what makes developers happy, angry, and everything in between?

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

The data has always/already been collected by someone



Public 'events: ~400 million (1 March 2011 – today)



GitHub Archive

Star 653

Google Analytics
Google+ Platform
Reinvigorate
StatHat
Twitter Button
Unofficial GitHub Buttons

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.

819 Subscribers

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:

Query	Command
Activity for April 11, 2012, 3PM UTC	`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`

Note: timeline data is available starting February 12, 2011.

Each archive contains a stream of JSON encoded GitHub events (sample), which you can process in any language. Ruby example:

```
1 require 'open-uri'
2 require 'zlib'
3 require 'yajl'
4
5 gz = open('http://data.githubarchive.org/2012-03-11-12.json.gz')
```

Google bigquery

COMPOSE QUERY

Query History

Job History

githubarchive

- github
  - actor_frequency
  - actor_location
  - twitter_forks
- github_explore
- github_proper
- stack_overflow
- githubarchive:github

Table Details: timeline

Schema | Details | Query Table

Description

Describe this table...

Table Info

Table ID	githubarchive:github.timeline
Table Size	112 GB
Number of Rows	185,116,106
Creation Time	7:49am, 29 Apr 2012
Last Modified	9:22am, 31 Mar 2014

Preview

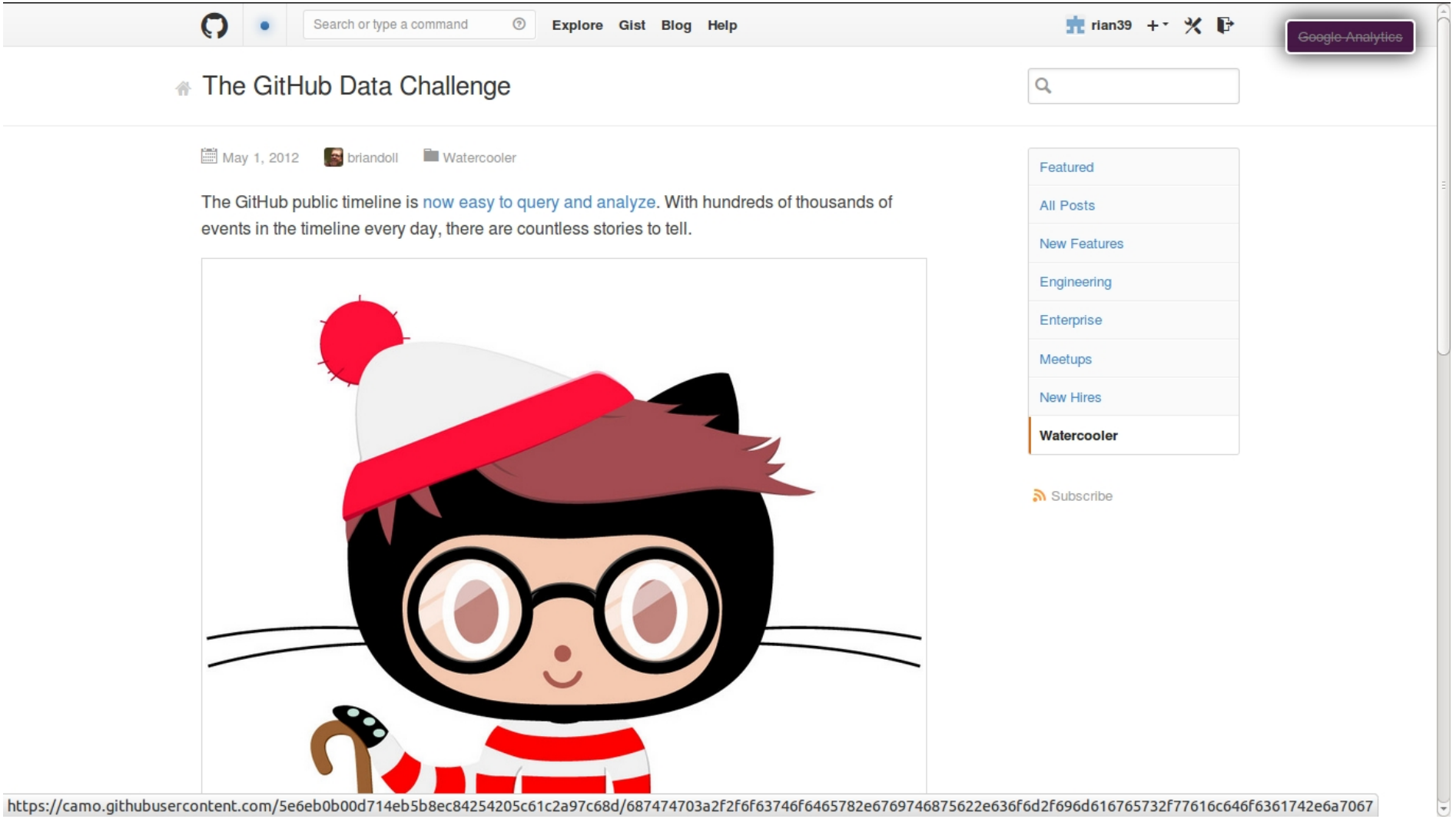
GoogleBigQuery: 'capturing and connecting of the huge data flows.' (ESRC Call)

Query Results 11:19am, 31 Mar 2014

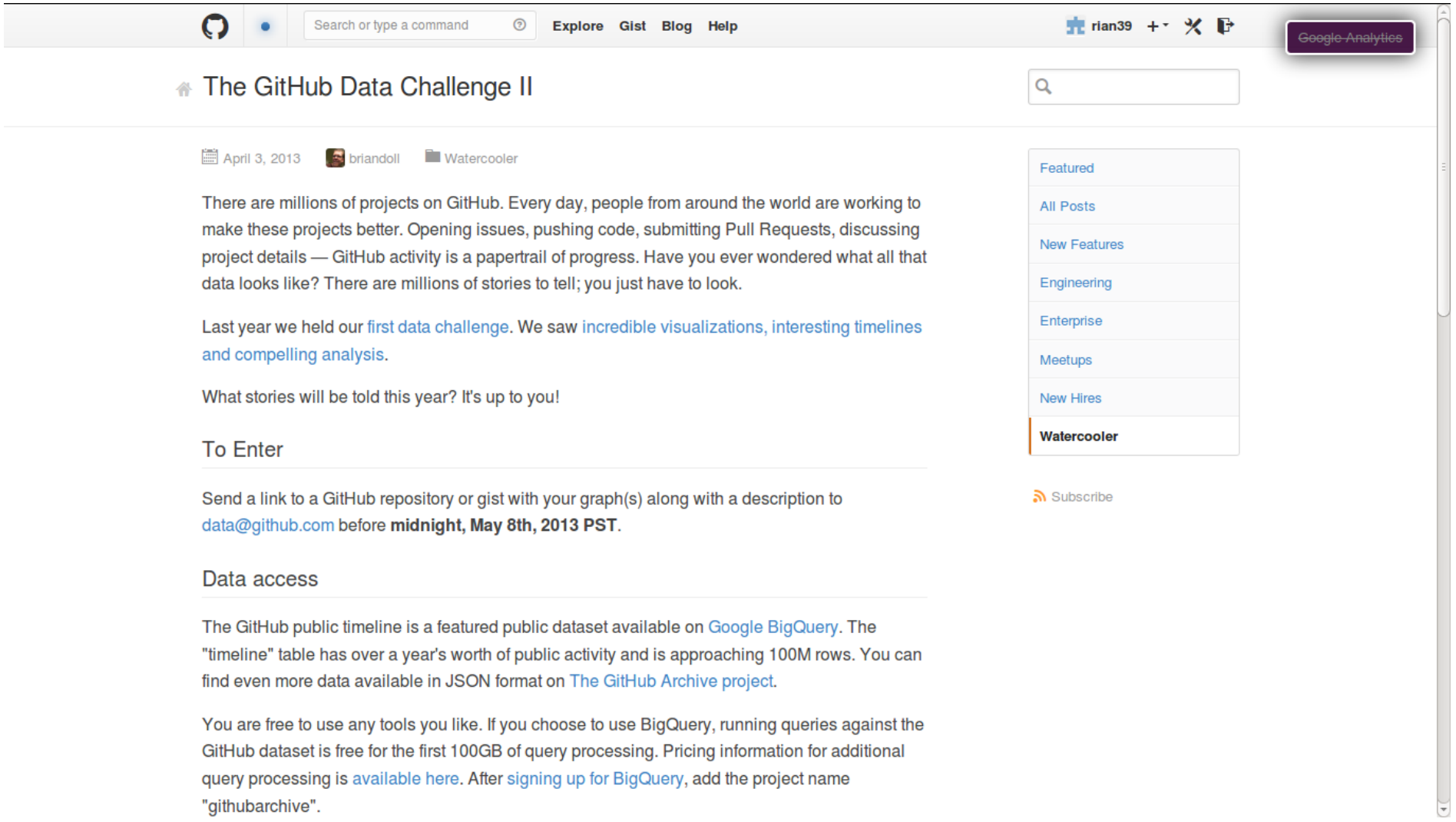
Row	created_at	type	actor	repository_name	url
1	2014-03-15 09:04:43	PushEvent	vvakame	DefinitelyTyped	https://github.com/borisyankov/DefinitelyTyped/compare/3d3832de3e...93a2313f62
2	2014-03-15 09:04:42	PushEvent	gizmomogwai	plist	https://github.com/gizmomogwai/plist/compare/12eb82d283...83ad8b5f26
3	2014-03-15 09:04:48	IssueCommentEvent	ben-lin	node.inflection	https://github.com/dreamerslab/node.inflection/issues/16#issuecomment-37721104
4	2014-03-15 09:04:48	WatchEvent	cbmd	mongofill	https://github.com/koubas/mongofill
5	2014-03-15 09:04:48	WatchEvent	svett	molokai	https://github.com/tomasr/molokai
6	2014-03-15 09:04:47	GollumEvent	swordray	wiki	https://github.com/ruby-china/wiki/wiki/RubyGems
7	2014-03-15 09:04:46	PushEvent	elfet	purephp	https://github.com/elfet/purephp/compare/dc24b70f00...c94c524e76
8	2014-03-15 09:04:46	PullRequestEvent	nikkypx	rets_data	https://github.com/arcticleo/rets_data/pull/2
9	2014-03-15 09:04:46	WatchEvent	mjaneczek	google-authenticator	https://github.com/jaredonline/google-authenticator
10	2014-03-15 09:04:53	PushEvent	micmath	Rye	https://github.com/micmath/Rye/compare/69c02b1aea...7885ea9e36
11	2014-03-15 09:04:53	PushEvent	fitret	daigon	https://github.com/fitret/daigon/compare/8897fa9404...77e5f48276
12	2014-03-15 09:04:53	WatchEvent	joeyates	spork	https://github.com/sporkrb/spork
13	2014-03-15 09:04:53	CreateEvent	grcnva	mcs-chatboard	https://github.com/grcnva/mcs-chatboard
14	2014-03-15 09:04:51	PushEvent	albatrossen	bungeebouncer	https://github.com/albatrossen/bungeebouncer/compare/57a3771590...7794ffb531
15	2014-03-15 09:04:51	PushEvent	catalinstanciu	ChessEngine-Xboard-cpp	https://github.com/catalinstanciu/ChessEngine-Xboard-cpp/compare/9bedce1744...8b98d2ff45

First < Prev Rows 1-15 of 50 Next > Last

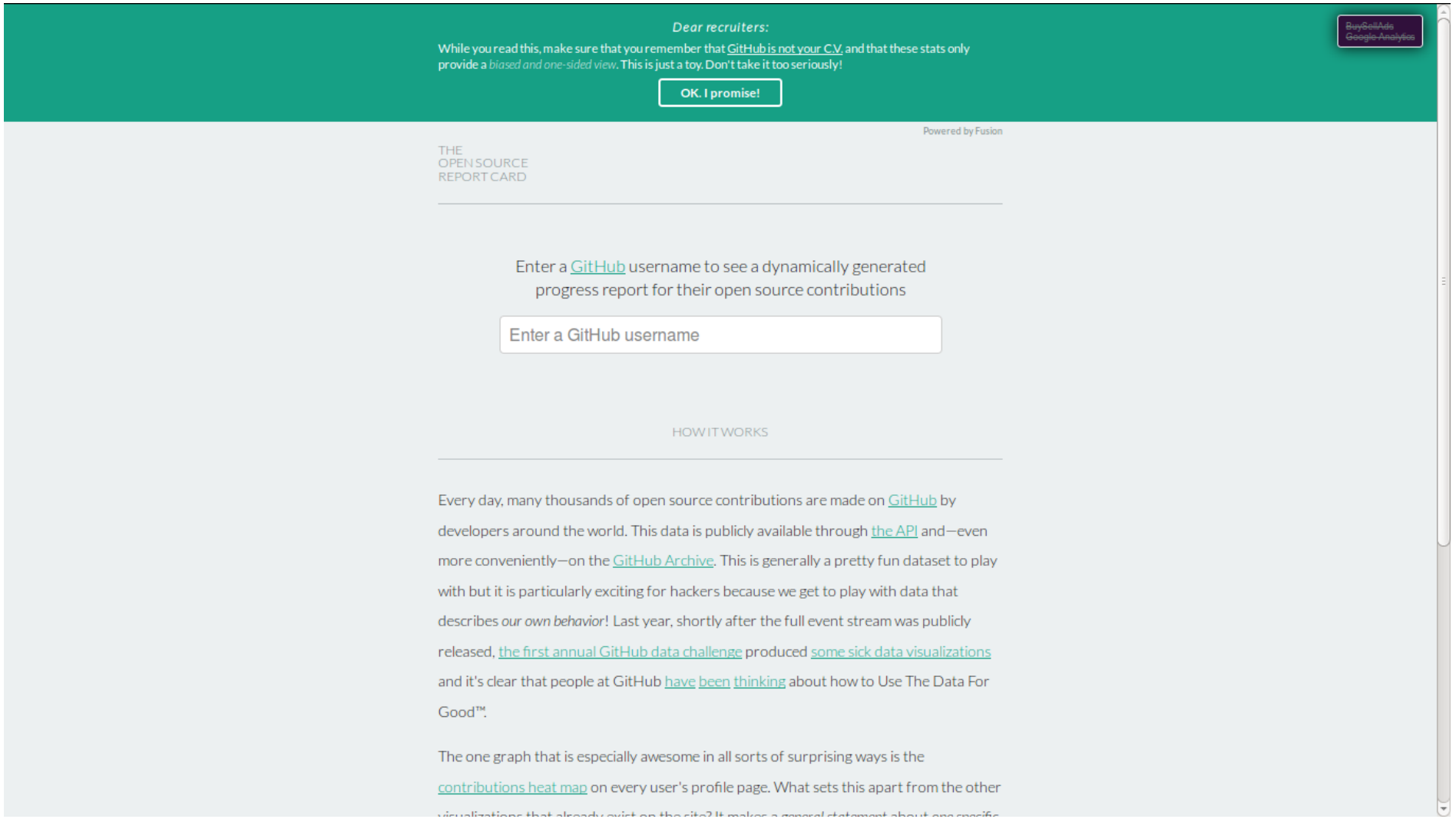
'Countless stories to tell': other people are working with the data



'What stories will be told this year? Its up to you'



Typical stories: “all individuals are similar”



Dear recruiters:

While you read this, make sure that you remember that GitHub is not your C.V. and that these stats only provide a *biased and one-sided view*. This is just a toy. Don't take it too seriously!

OK. I promise!

BuySellAds
Google Analytics

Powered by Fusion

THE
OPEN SOURCE
REPORT CARD

Enter a GitHub username to see a dynamically generated progress report for their open source contributions

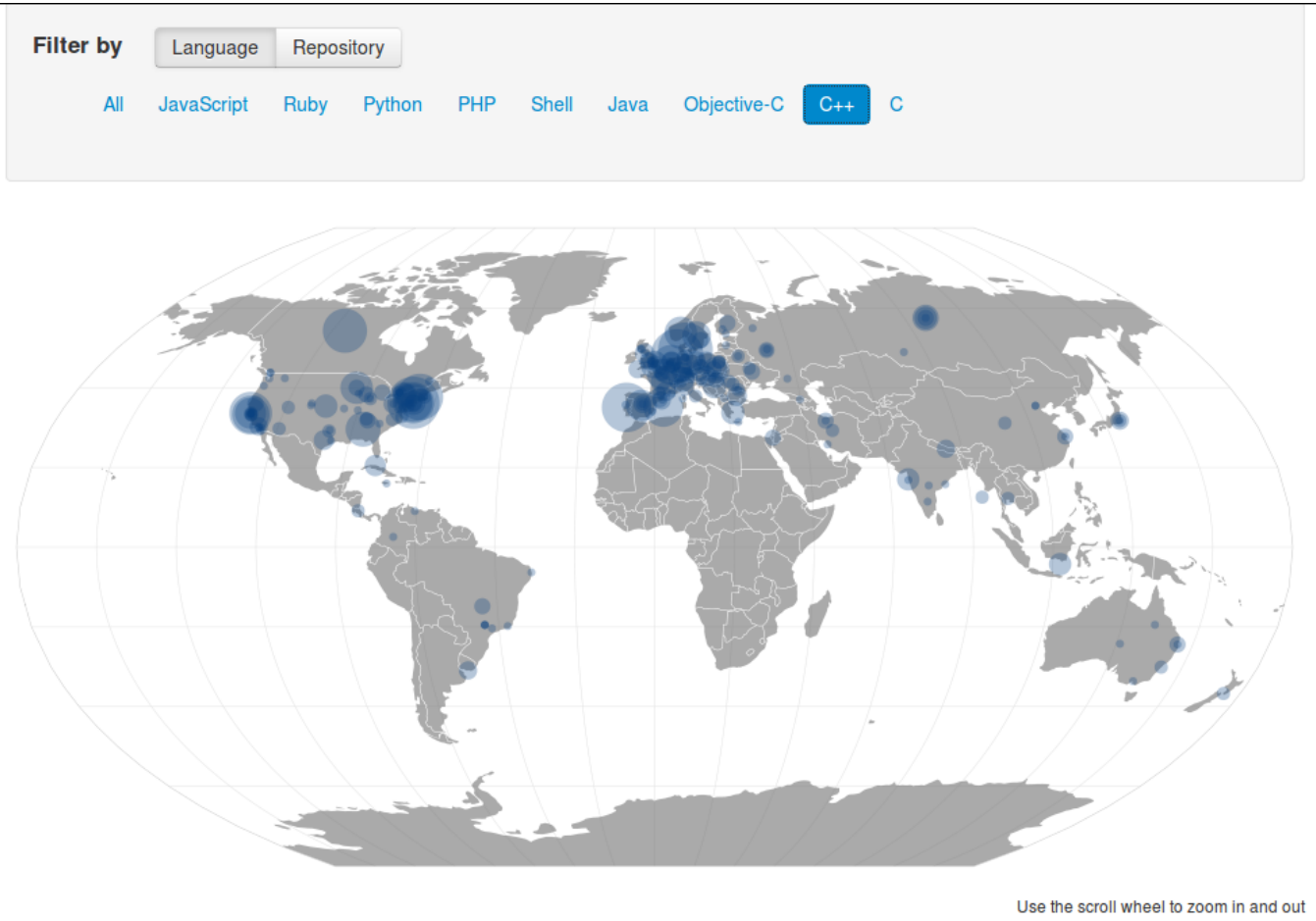
Enter a GitHub username

HOW IT WORKS

Every day, many thousands of open source contributions are made on GitHub by developers around the world. This data is publicly available through the API and—even more conveniently—on the GitHub Archive. This is generally a pretty fun dataset to play with but it is particularly exciting for hackers because we get to play with data that describes *our own behavior*! Last year, shortly after the full event stream was publicly released, the first annual GitHub data challenge produced some sick data visualizations and it's clear that people at GitHub have been thinking about how to Use The Data For Good™.

The one graph that is especially awesome in all sorts of surprising ways is the contributions heat map on every user's profile page. What sets this apart from the other visualizations that already exist on the site? It makes a general statement about one specific

Stories: “Repositories live in familiar places: SF, NY, LON, etc.'



GitHub's public timeline summarizes the open source work of millions of developers the world over.

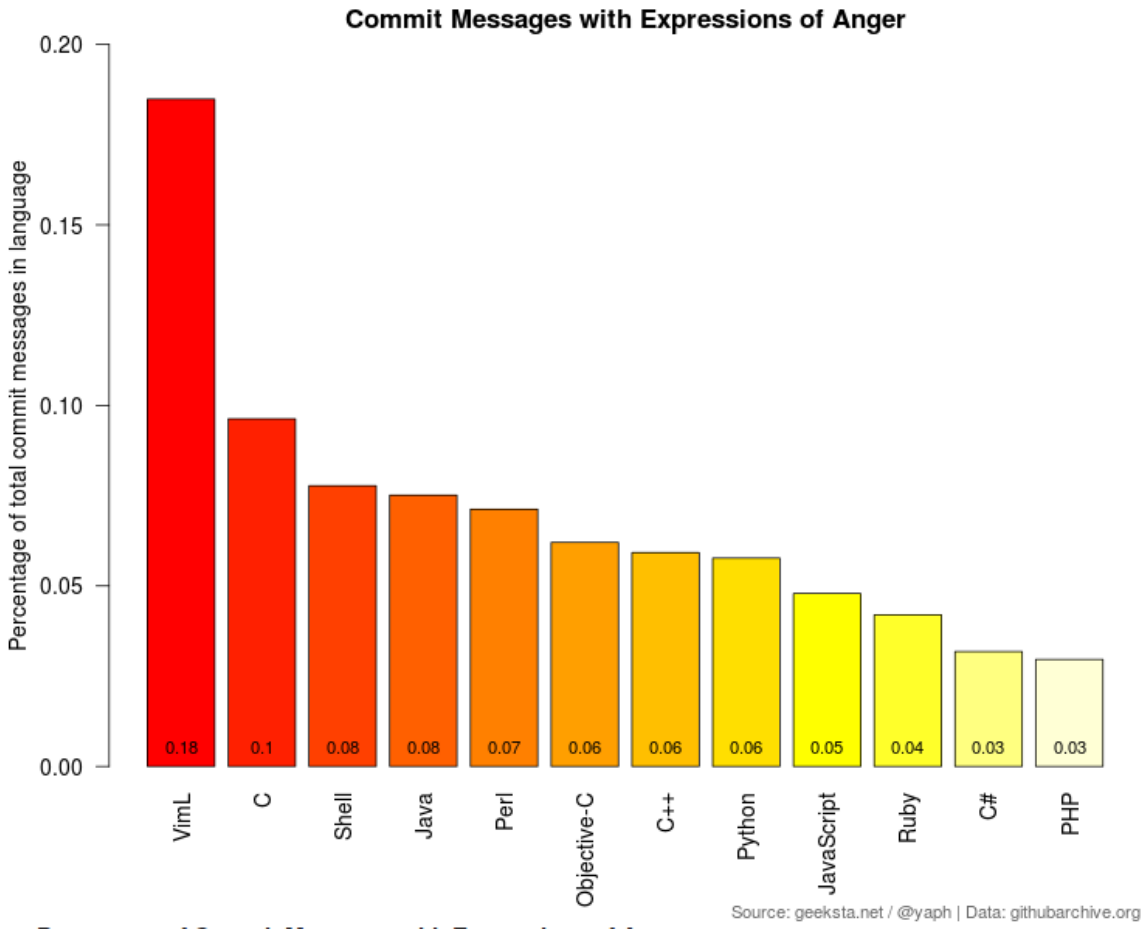
This data set contains contributions to the top 200 GitHub repositories during the first four months of

Story: 'coding is hard work'

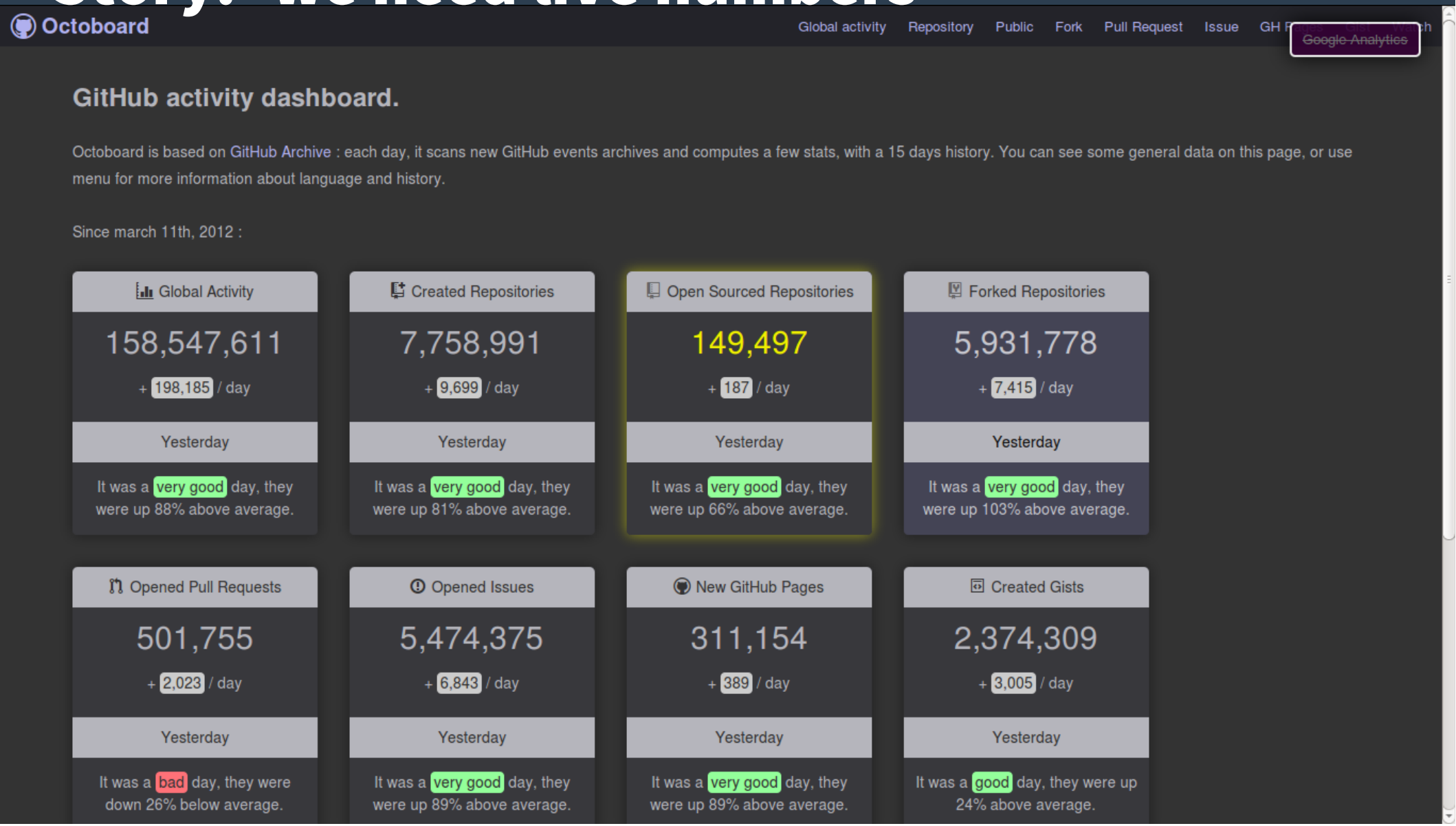
GEEKSTA Sections Projects Geeklog Tools Shirts +

Regular Expression

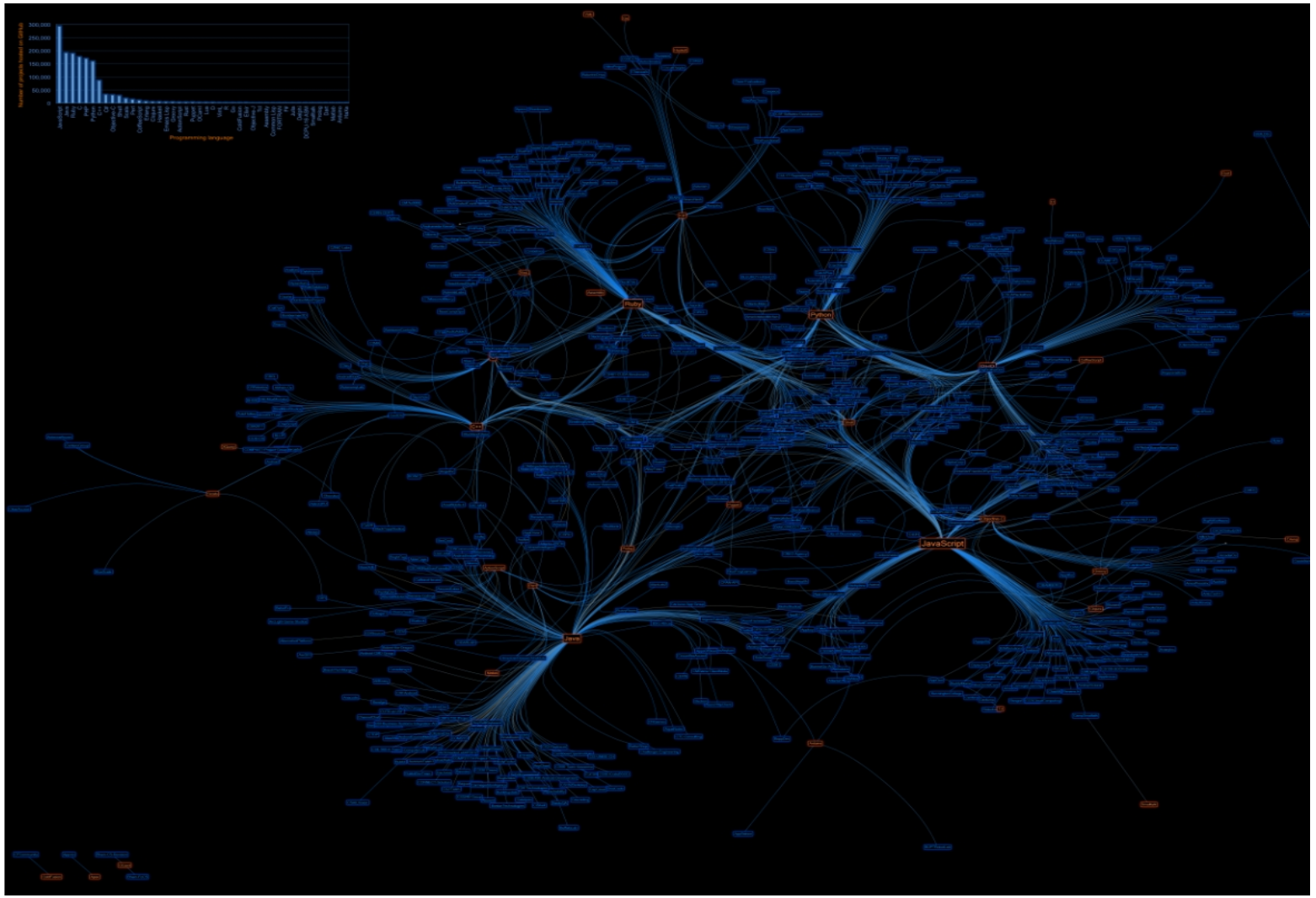
```
(?i)\b(a+rgh|angry|annoyed|annoying|appalled|bitter|cranky|hate|hating|mad)\b'
```



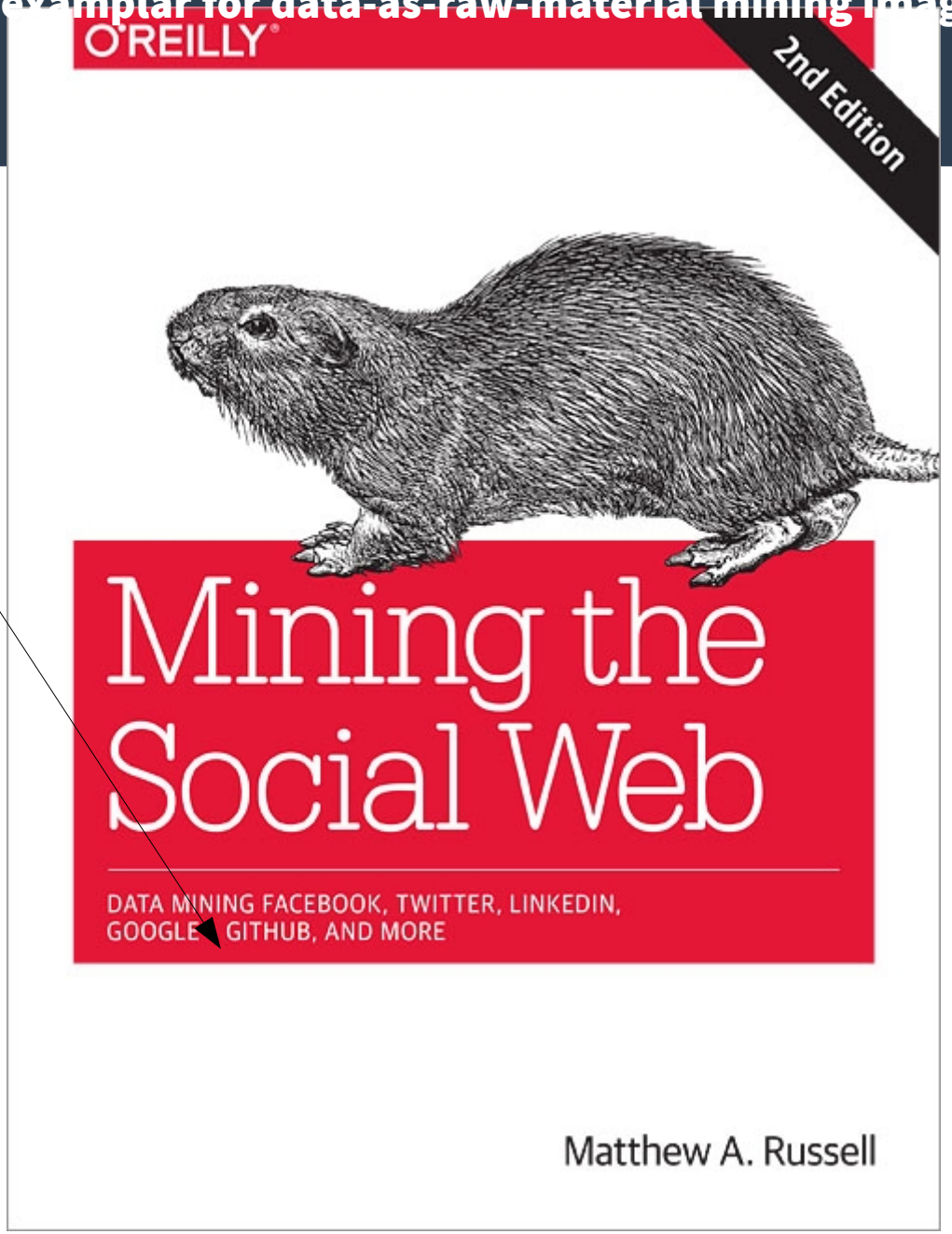
Story: 'we need live numbers'



Story: 'Code is a social network'



Story: 'Github data is an examplar for data-as-raw-material mining imaginary'



Part C: re-counting 'capital numbers'

Statistician to digital sociologist: “All your questions are too difficult”

jupyter Github_project_topics Last Checkpoint: an hour ago (autosaved)

File Edit View Insert Cell Kernel Help Python 2

Code Cell Toolbar: None

In [1]:
```
import google_bigquery_access as gbq
import pandas as pd
import datetime as dt
import matplotlib.pyplot as plt
import gensim as gs
```

Ways of characterising what Github projects are actually about

How do we know what repositories are about? Can we know even know whether a repository has any software in it?

In [2]:
```
query = """select repository_name, repository_description, repository_language
from [publicdata:samples.github_timeline]
limit 5000;"""

repo_df = gbq.query_table(query, 5000)
```

```
executing query:
select repository_name, repository_description, repository_language
from [publicdata:samples.github_timeline]
limit 5000;
has a rows attribute
5000 of   5000 (5000, 3)
```

In [13]:
```
repo_df.repository_description = repo_df.repository_description.fillna(' ')
stoplist = set('for from but an on is or that a of the and to in with this that be using -'.split())
```

“All questions are too difficult” Use big

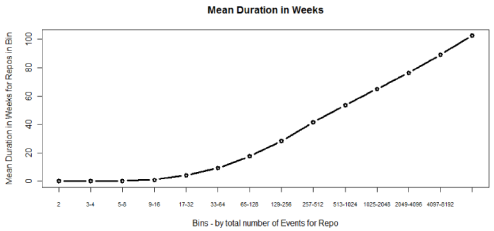
In [15]:
```
lda_model = gs.models.ldamodel.LdaModel(corpus=git_corpus, id2word=dictionary, num_topics=15,update_every=0, passes=50)
```

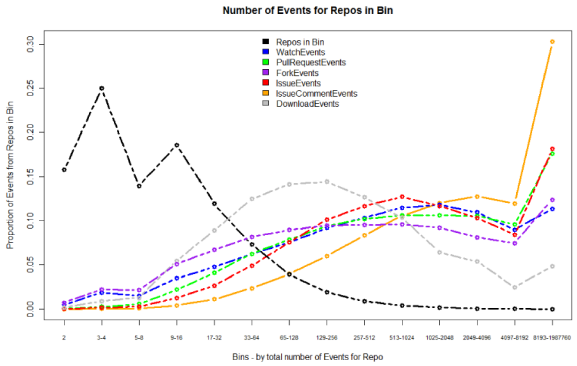
In [17]:
```
[[i, lda_model.print_topic(i)] for i in range(15)]
```

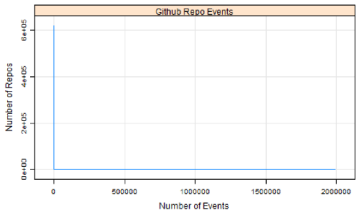
Out[17]:
```
[[0,
  u'0.080*plugin + 0.052*interface + 0.043*mirror + 0.038*email + 0.036*adding + 0.035*git-svn + 0.035*trac + 0.035*email
2trac, + 0.032*jquery + 0.016*website'],
 [1,
  u'0.019*your + 0.016*test + 0.013*extension + 0.012*password + 0.010*ios + 0.010*node + 0.010*like + 0.009*estimation +
 0.009*realistic + 0.009*strength'],
 [2,
  u'0.026*server + 0.021*data + 0.020*use + 0.014*easy + 0.013*api + 0.011*engine + 0.011*project + 0.011*source + 0.011*
open + 0.010*node.js'],
 [3,
  u'0.034*javascript + 0.028*web + 0.023*application + 0.022*framework + 0.021*python + 0.017*testing + 0.014*into + 0.01
1*open + 0.011*node.js + 0.010*xml'],
 [4,
  u'0.021*as + 0.017*blog + 0.015*rails + 0.014*library + 0.013*your + 0.012*personal + 0.010*java + 0.010*tool + 0.009*s
imple + 0.009*data'],
 [5,
  u'0.031*framework + 0.016*web + 0.015*building + 0.014*management + 0.014*your + 0.011*php + 0.011*html5 + 0.011*work +
 0.010*applications + 0.010*simple,'],
 [6,
  u'0.032*see + 0.030*here + 0.030*nothing + 0.015*source + 0.012*by + 0.011*project + 0.008*advanced + 0.008*not + 0.007
*test + 0.007*open'],
 [7,
  u'0.021*code + 0.020*framework + 0.016*server + 0.015*salesforce + 0.015*application + 0.014*repository + 0.012*web + 0
.012*project + 0.011*scripts + 0.011*collection'],
 [8,
  u'0.036*app + 0.029*ruby + 0.027*rails + 0.016*tutorial + 0.016*application + 0.011*web + 0.010*written + 0.010*at + 0.
009*example + 0.009*management'],
 [9,
  u'0.033*my + 0.019*toolkit + 0.013*ruby + 0.013*python + 0.013*base + 0.010*vim + 0.010*js + 0.010*android + 0.008*twit
ter + 0.008*(cyanogenmod)'],
```

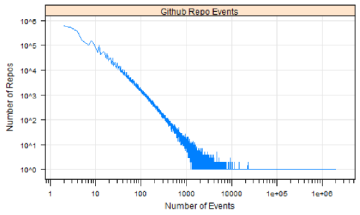
In [13]:
```
repo_df.repository_description = repo_df.repository_description.fillna(' ')
stoplist = set('for from but an on is or that a of the and to in with this that be using -'.split())
```

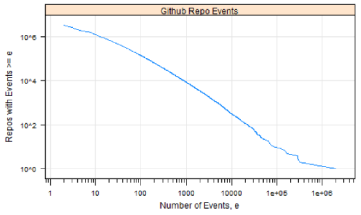
Statistician says 'no'
'Digital sociologist' says 'yes'

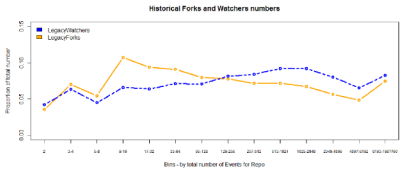




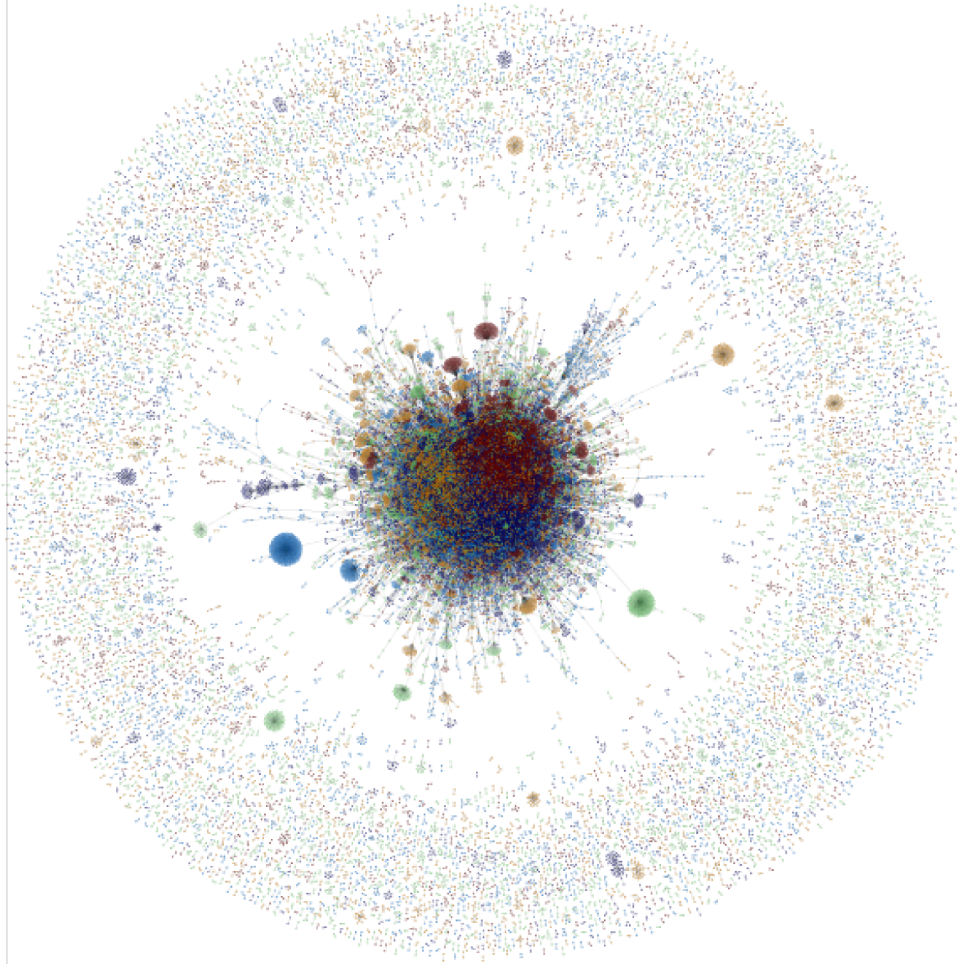






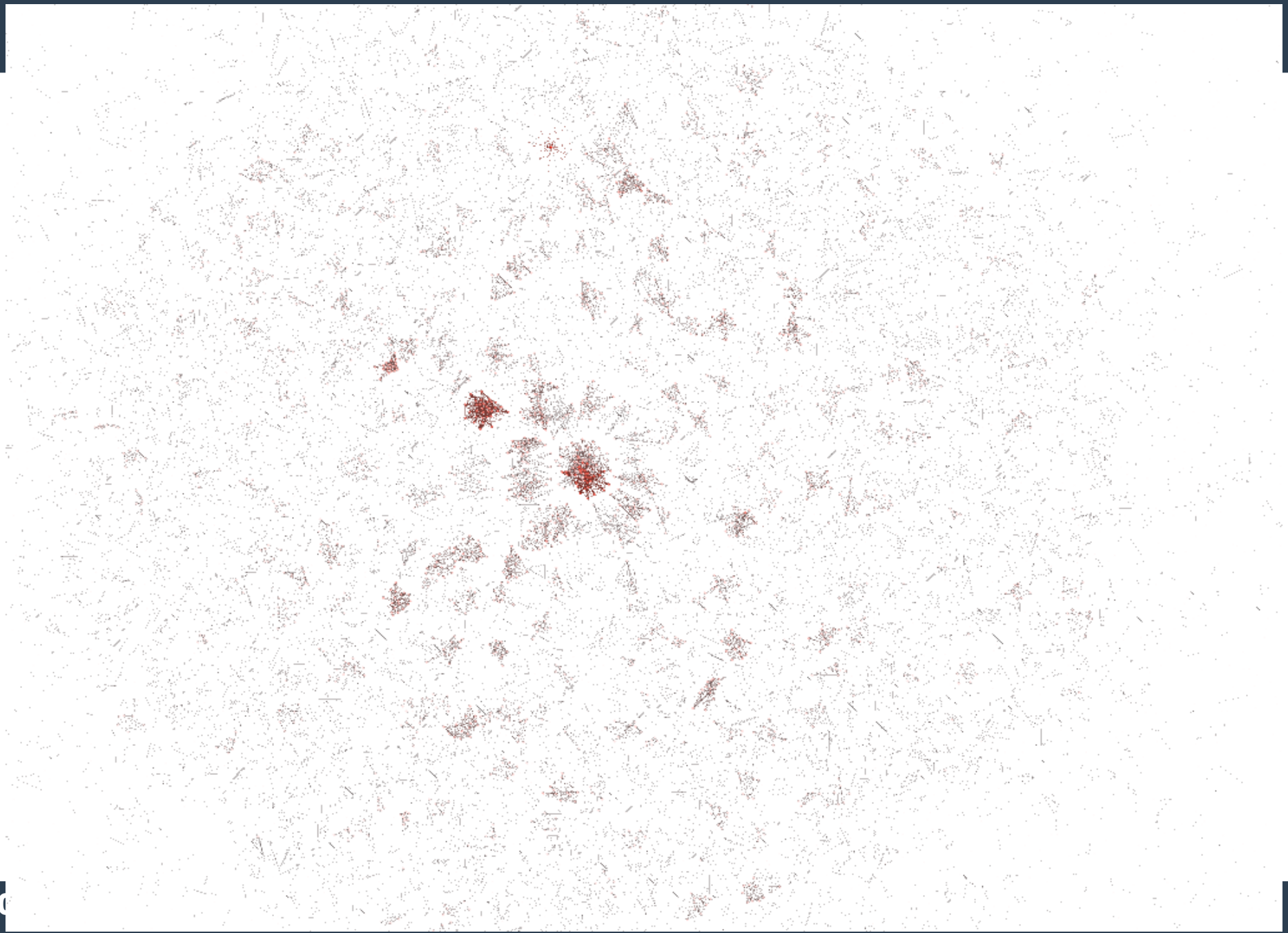


statistics



digital sociology?

Statistician says 'yes'



Steam billows from one query ...

Google Developers Console

+Adrian

Projects

Billing

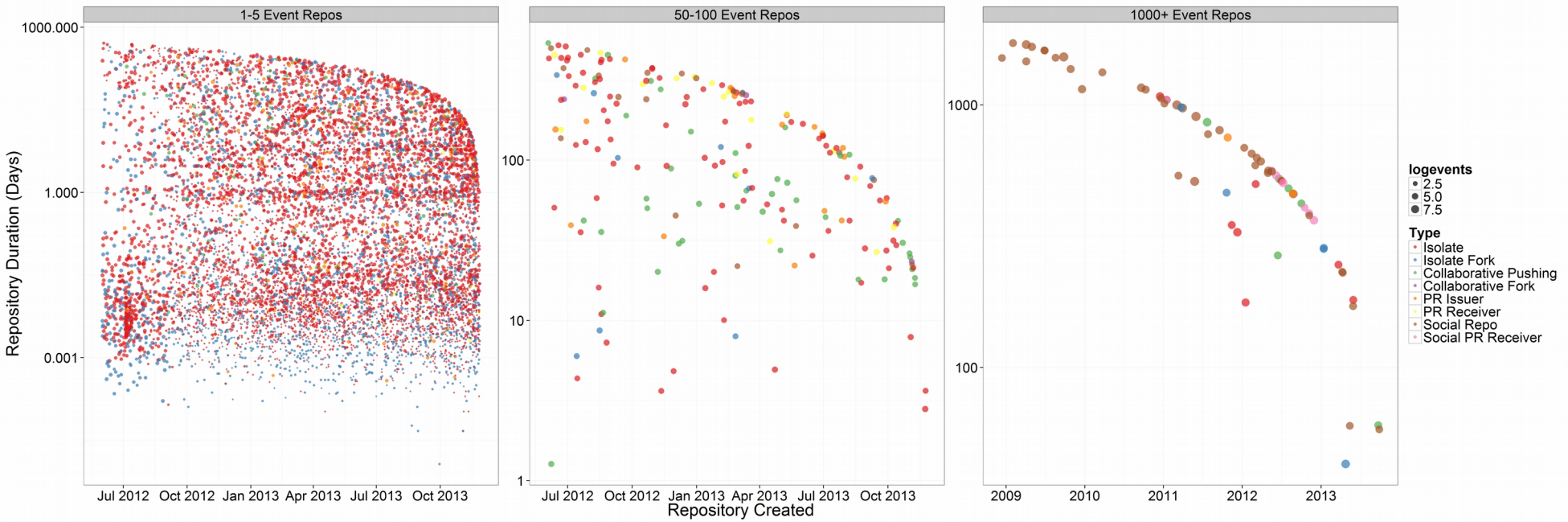
Account settings

Oct 1, 2013	Starting balance			10.54
Sep 1, 2013 - Sep 30, 2013		**$10.54**	**($2,265.66)**	**$10.54**
▸ Documents (2)				
Sep 30, 2013	VAT	1.97		10.54
Sep 1, 2013 - Sep 30, 2013	BigQuery Analysis: 172.789 GB (Project:237471208995)	2.49		8.57
Sep 1, 2013 - Sep 30, 2013	BigQuery Storage: 77.839 GB-month (Project:237471208995)	6.08		6.08
Sep 10, 2013	Automatic payment: Visa ...4260		(2,265.66)	
Sep 9, 2013	Automatic payment declined: Visa ...4260 for $2,265.66. No reason provided by your financial institution			2,265.66
Sep 4, 2013	Automatic payment declined: Visa ...4260 for $2,265.66. No reason provided by your financial institution			2,265.66
Sep 3, 2013	Automatic payment declined: Visa ...4260 for $2,265.66. No reason provided by your financial institution			2,265.66
Sep 1, 2013	Starting balance			2,265.66
Aug 1, 2013 - Aug 31, 2013		**$2,265.66**	**($121.73)**	**$2,265.66**
▸ Documents (2)				
Aug 31, 2013	VAT	423.66		2,265.66
Aug 1, 2013 - Aug 31, 2013	BigQuery Analysis: 53815.969 GB (Project:237471208995)	1,835.96		1,842.00
Aug 1, 2013 - Aug 31, 2013	BigQuery Storage: 77.351 GB-month (Project:237471208995)	6.04		6.04
Aug 3, 2013	Automatic payment: Visa ...4260		(121.73)	
Aug 1, 2013	Starting balance			121.73
Jul 1, 2013 - Jul 31, 2013		**$121.73**	**($22.56)**	**$121.73**
▸ Documents (2)				
Jul 31, 2013	VAT	22.76		121.73
Jul 1, 2013 - Jul 31, 2013	BigQuery Analysis: 2818.778 GB (Project:237471208995)	92.93		98.97
Jul 1, 2013 - Jul 31, 2013	BigQuery Storage: 77.285 GB-month (Project:237471208995)	6.04		6.04
Jul 3, 2013	Automatic payment: Visa ...4260		(22.56)	
Jul 1, 2013	Starting balance			22.56
Jun 4, 2013 - Jun 30, 2013		**$4.22**	**$0.00**	**$22.56**
▸ Documents (2)				

Return to original console Send feedback Privacy & Terms

Big bill for big query: $USD1835 + $USD 420 VAT = $**2200**

“Give me something chunky”



The millions repositories on Github are mostly ephemeral **imitations**

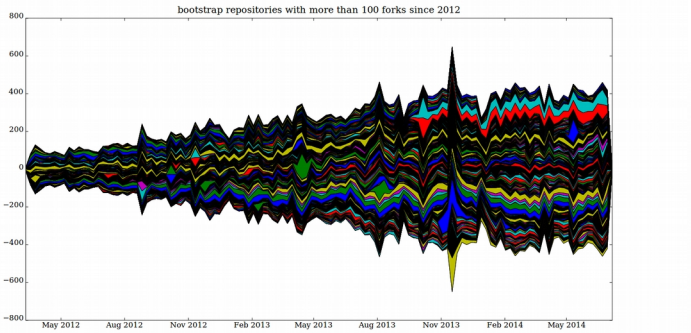
Early repositories (c. 2008)

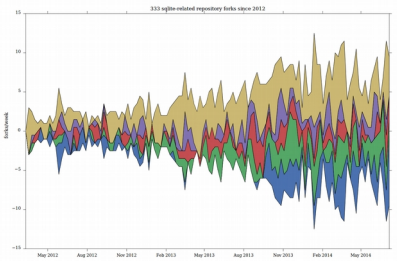
1 mojombo/grit
2 wycats/merb-core
3 rubinius/rubinius
4 mojombo/god
5 vanpelt/jsawesome
6 wycats/jspec
7 defunkt/exception_logger
8 defunkt/ambition
9 labria/restful-authentication
10 technoweenie/restful-authentication
11 technoweenie/attachment_fu
12 topfunky/bong
13 anotherjesse/s3
14 anotherjesse/taboo
15 mojombo/glowstick
16 wycats/merb-more
17 macournoyer/thin
18 jamesgolick/resource_controller
19 defunkt/cache_fu
20 bmizerany/sinatra
21 rtomayko/sinatra
22 jnewland/gsa-prototype
23 defunkt/mofo
24 schacon/ruby-git
25 mmower/simply_versioned
26 abhay/calais
27 mojombo/chronic
28 al3x/git-wiki

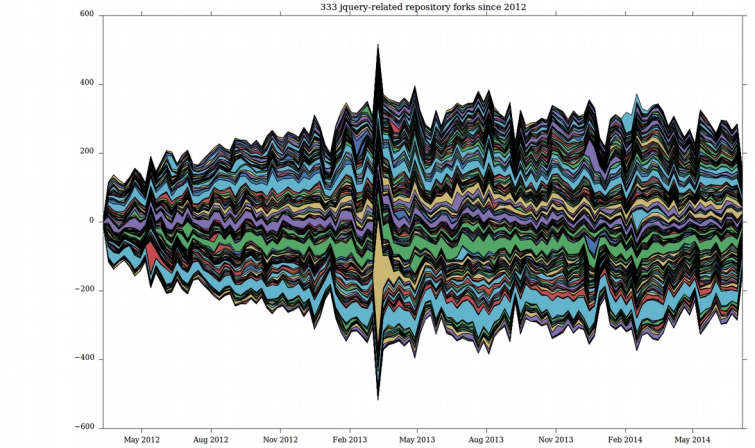
Recent repositories (c. 2014)

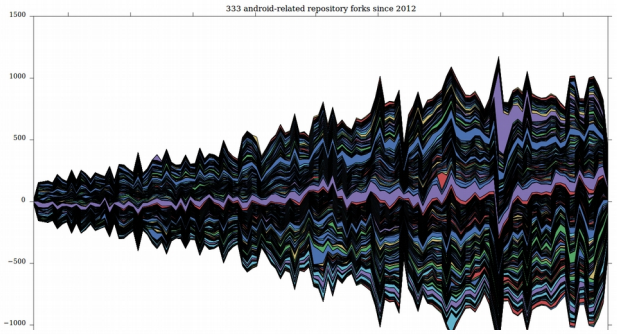
1 2m1tsu3/practice
2 tylerdmace/ledomme
3 yehiaelghaly/xssya
4 istvan-antal/commandjs
5 JohnKrigbjorn/ObjectOne
6 chenx/Ci35_1
7 mohsenbezanj/AI_Project
8 Dineshkarthik/blogengine
9 sapanbhuta/Sapari
10 gwoodroof/chat-gwoodroof
11 tryuichi/Hello-World
12 prateek0020/NepTravelMate
13 discoverfly/discoverfly.github.io
14 evan-007/ng-wikiful
15 sanemat/zipcode-jp
16 donreamey/PJKiller
17 discoverfly/discover
18 jkkorean/MIUI-KK
19 Artofacks1/ionic-app
20 jkkorean/MIUI-JB
21 tycho01/rails-i18n
22 Oksisane/Websockets
23 chenxiruanhai/XScrollView
24 jhonM17/footer-fixed-boostrap
25 McPringle/workshops
26 JosemyD/testing-laravel
27 wngravette/FlogResources
28 Ethico/temp

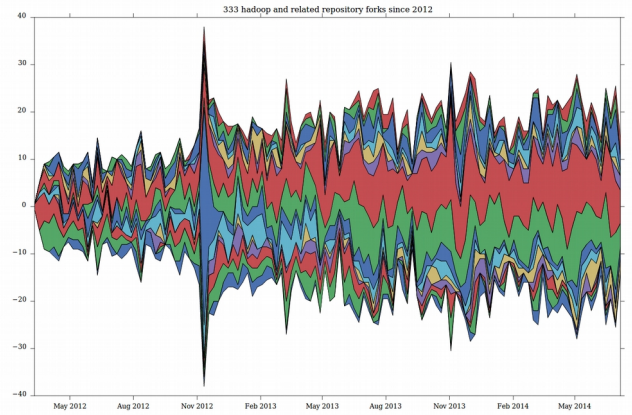
Mapping imitative fluxes: follow the copying of names across repositories

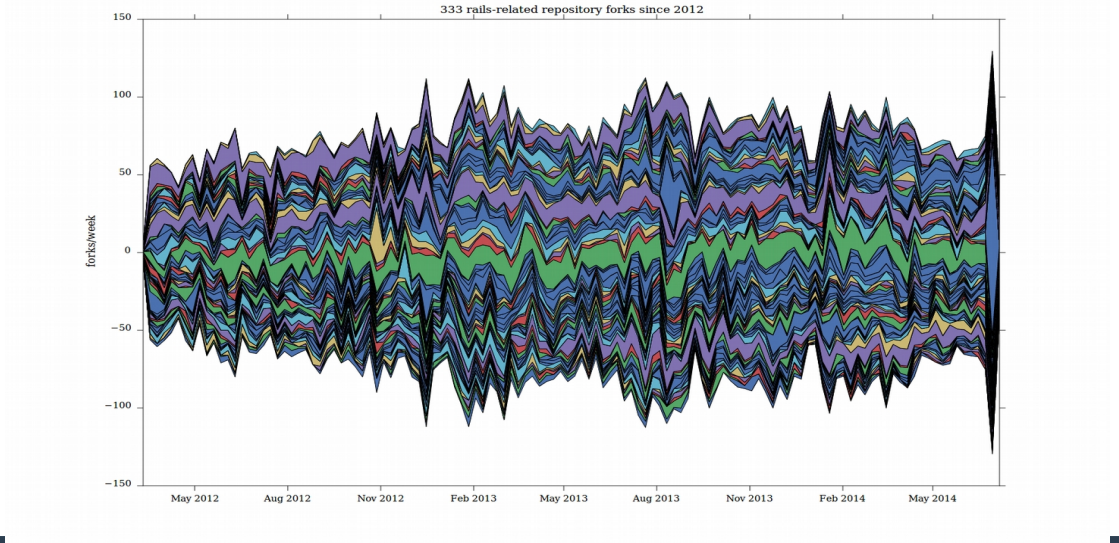












Conclusions? Work in the data-as-raw-material imaginary

- Data-as-raw-material imaginary: *very* disconnected from the liveliness of the data as imitative fluxes and its congealing in large or 'capital' numbers
- What happens to perspectives on data format?
  - **Traceability and digital navigation** (Latour) assumes data is a trace *left behind* not something being intensively worked (e.g. in data challenges and the data-as-raw-material imaginary)
  - **Device-specific research and re-distribution of methods** (Marres) suggests all traces are 'mixtures' that format, attract, and overflow;
    - Definitely resists naïve claims about data power and transparency, but
  - Transformation in **symbolic-analytic work** (Gill & Pratt) associated with data*:*
    - Expand to include academic research feeding into large numbers and data visuality, not just the work that of creatives

Conclusions? Work in the data-as-raw-material imaginary

"What I am suggesting is that there are multiple ways to do the relation unity/plurality; hence there are multiple sorts of numbers."
Verran, 2001, 107

Re-counting large numbers might:

- Begins to differentiate another kind of **contemporary numerality** (alongside ordinals, parameters, percentages, constants, real numbers, differentials, one-many): **capital numbers**
- **Re-scale** them in relation to networks of imitation replete with differences and similarities
- Be done from **within** the data not standing outside it
- Might be politically useful in resisting the **capitalisation** of large number

Acknowledgements

Research funded by ESRC (UK Economic and Social Research Council) under 'Google Data Analytics' programme, 2013-2014

Richard Mills and Stuart Sharples – research associates

Matthew Fuller (Goldsmiths) and Andrew Goffey (Nottingham) – co-researchers