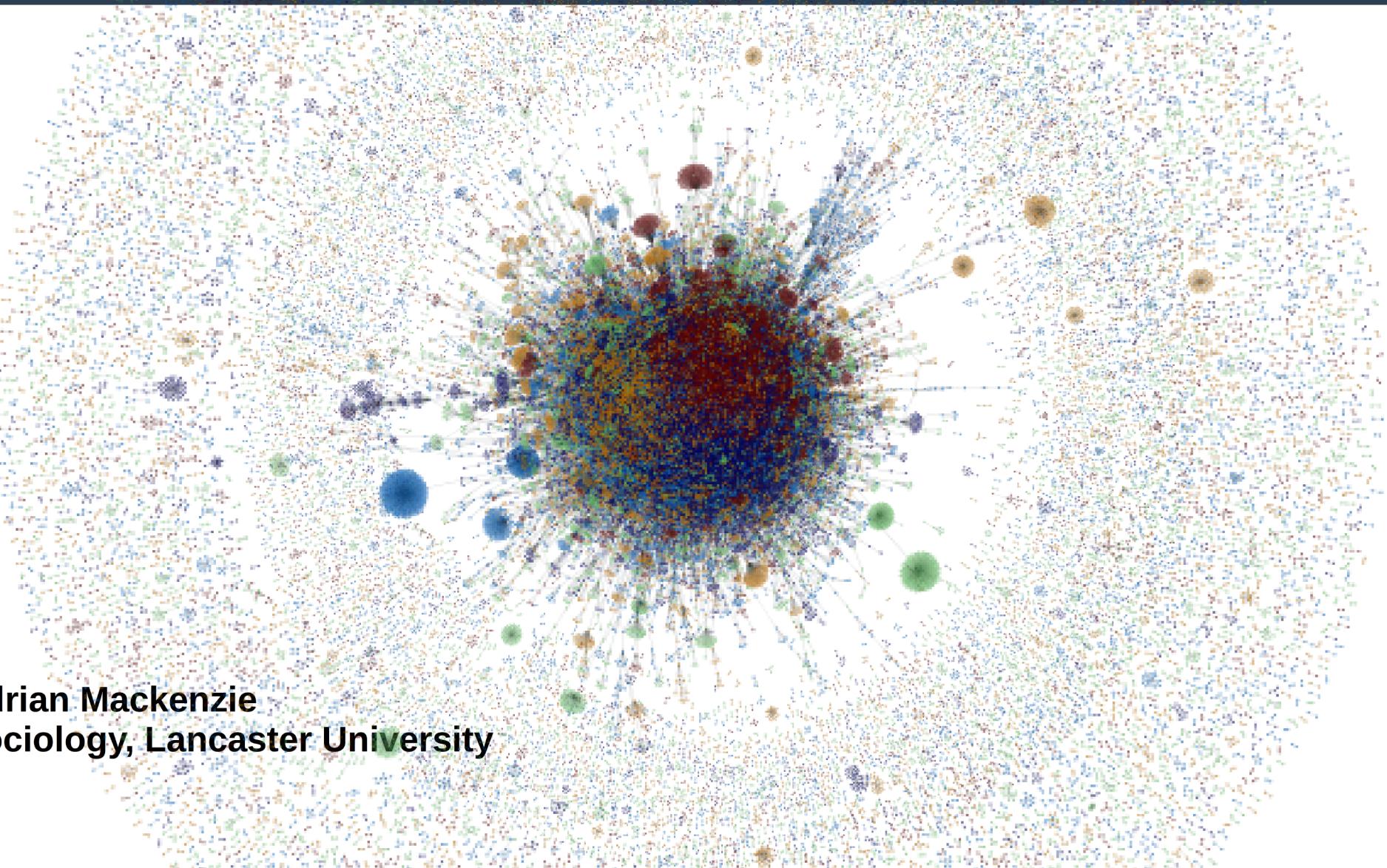


# Many names and large numbers: methods for imitative fluxes in the data-as-raw-material imaginary?



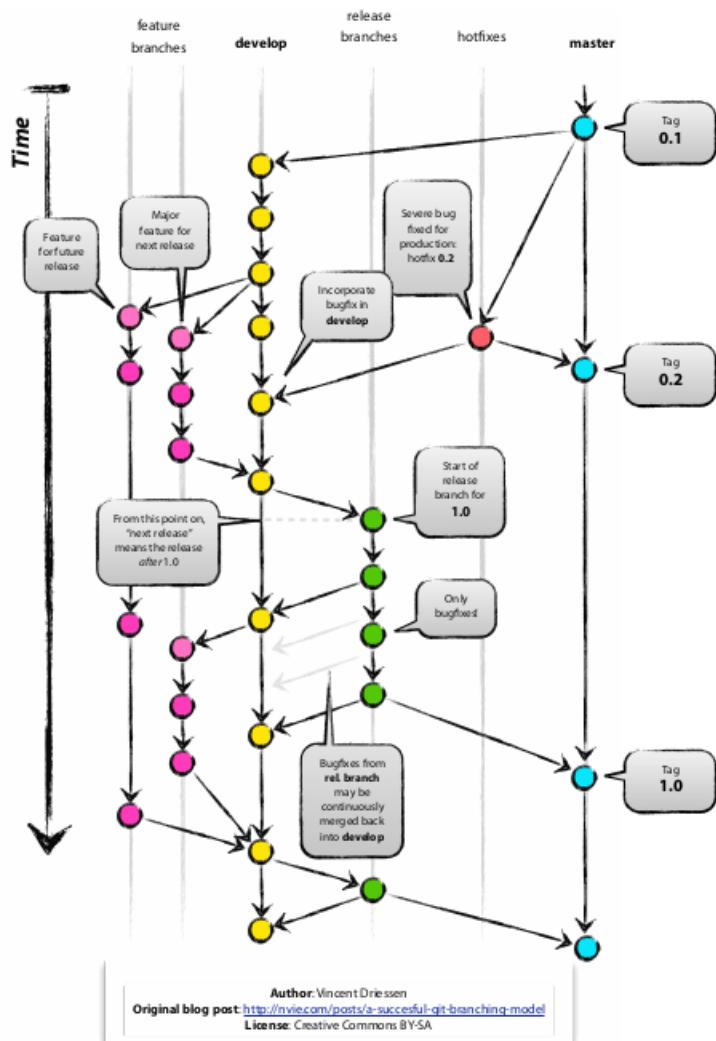
Adrian Mackenzie  
Sociology, Lancaster University

# 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?

File Edit View Help

fixed: save\_issue now checks if issue is associated with a pull request in a clean way  
Merge branch 'master' of http://github.com/metacommunities/metacommunities  
Merge branch 'stu/history-git'  
feature added: history\_git uploads a wide table of activity to big query  
Merge branch 'stu/history-git'  
fix: 'create\_db not found'  
feature added: history\_git uploads a wide table of activity to big query  
fix: 'create\_db not found'  
analysis history debug remotes/origin/analysis history debug Merge branch 'master' of https://github.com/m...  
Merge branch 'stu/history-git'  
refactored history git, so module is now a folder rather than a single file  
Merge branch 'stu/crossref-tags'  
crossref so w/ github: will now upload to bigquery  
Merge branch 'stu/licenses'  
abandoning searching for licenses as github have restricted the ability to search code  
brickwall: cant scrape  
problem: api code search requires repo or user - part 2  
problem: api code search requires repo or user  
Merge branch 'stu/upload-stackoverflow'  
remotes/origin/stu/upload-stackoverflow I can upload to BQ and Cloud SQL, all my code works  
changed names of tag tables to be more consistent  
added tag\_summary table  
mysql now creates tags rather than python code  
refactored import and dump into a single loop  
Merge branch 'master' into stu/upload-stackoverflow  
mysql dumps are now gzipped  
updating queries  
looking at early repos to see if githubbers make github  
Merge branch 'master' of https://github.com/metacommunities/metacommunities  
added list of licenses to be searched for  
Merge branch 'stu/identify-repo-licenses'  
recon on license searching  
Merge branch 'stu/classify-orgs'  
files are more tidy

stuples <nope>  
rian39 <a.mackenzie@lancaster.ac.uk>  
stuples <nope>  
Richard Mills <r.mills@lancs.ac.uk>  
rian39 <a.mackenzie@lancaster.ac.uk>  
rian39 <a.mackenzie@lancaster.ac.uk>  
stuples <nope>  
stuples <nope>  
stuples <nope>  
stuples <nope>  
stuples <nope>  
stuples <nope>

2014-04-16 10:53:00  
2014-04-14 17:14:36  
2014-04-14 16:15:33  
2014-04-14 15:15:47  
2014-04-14 10:45:41  
2014-04-14 10:44:11  
2014-04-14 15:15:47  
2014-04-14 10:44:11  
2014-03-03 08:30:45  
2014-03-28 15:38:56  
2014-03-21 20:52:33  
2014-03-28 15:37:50  
2014-03-28 15:37:08  
2014-03-28 12:45:25  
2014-03-28 12:42:45  
2014-03-27 15:46:08  
2014-03-26 10:13:56  
2014-03-26 10:13:31  
2014-03-28 12:38:23  
2014-03-26 16:13:28  
2014-03-21 21:05:51  
2014-03-21 21:03:27  
2014-03-21 19:22:52  
2014-03-20 16:46:12  
2014-03-20 15:23:49  
2014-03-20 15:19:12  
2014-03-26 11:46:40  
2014-04-03 08:30:19  
2014-03-26 10:03:51  
2014-03-22 12:57:28  
2014-03-22 12:49:43  
2014-03-21 19:11:25  
2014-03-22 12:48:38  
2014-03-22 12:36:04

SHA1 ID: a9afa6f5ab574db5e5bf86c24fccf86110a633b1 ← → Row 16 / 484

Find ↓ ↑ commit containing: Exact All fields

Search

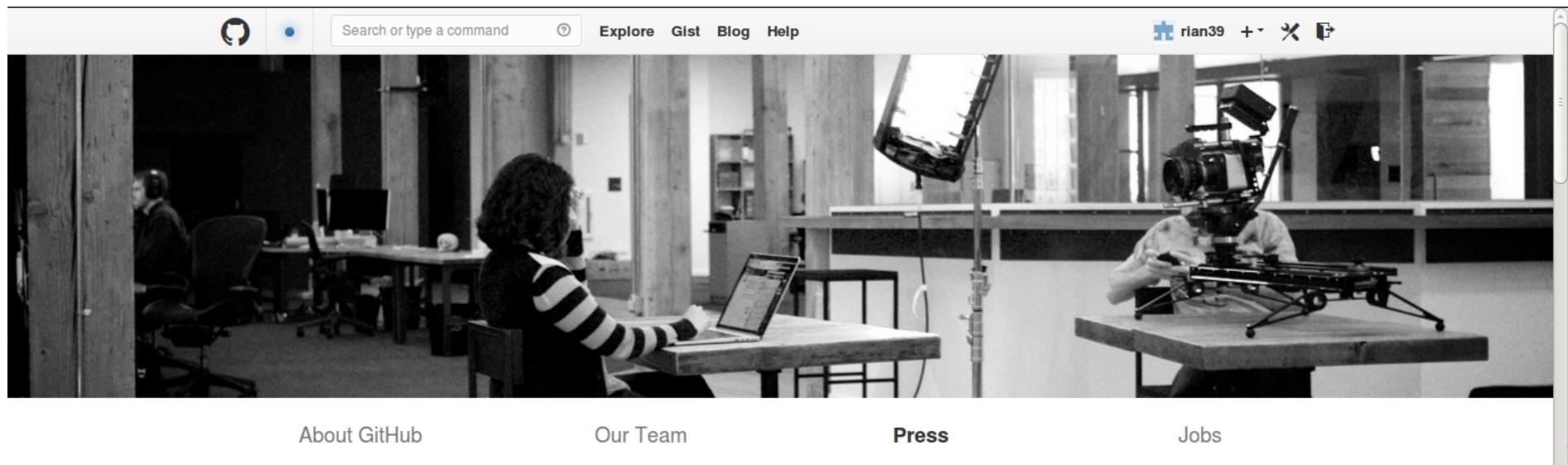
Diff Old version New version Lines of context: 3

Author: rian39 <a.mackenzie@lancaster.ac.uk> 2016-01-05 13:51:44  
Committer: rian39 <a.mackenzie@lancaster.ac.uk> 2016-01-05 13:51:44  
Parent: 16ec79efab44a8cf36ef46ecd9b4a794cb711614 (Merge branch 'publ...  
Child: 73e2f880086e3ae15dc84de15c726dc44fb98438 (fixing fork plots)  
Branches: analysis/actor\_copying, analysis/repo\_census, master, publication/material\_cultures\_london2016, publication/ss...  
remotes/origin/analysis/actor\_copying, remotes/origin/analysis/repo\_census, remotes/origin/master, remotes/origin/publication/material\_cultures\_london2016,

Comments

publications/material\_cultures\_2016/.~lock.material\_culture\_refs.csv#  
publications/material\_cultures\_2016/mackenzie\_material\_culture\_feb2016.md  
publications/material\_cultures\_2016/mackenzie\_material\_culture\_feb2016.odp  
publications/material\_cultures\_2016/mackenzie\_material\_cultures.md  
publications/material\_cultures\_2016/material\_culture\_refs.csv  
publications/material\_cultures\_2016/title.txt

# November 2013: Github founder resigns after sexual harrassment claims



The screenshot shows the GitHub homepage. At the top, there's a search bar with placeholder text "Search or type a command". Below it is a navigation menu with links for "Explore", "Gist", "Blog", and "Help". On the right side of the header, there's a user profile icon for "rian39" and some other account-related icons. The main content area features a large black and white photograph of several people working at desks in an office environment. Below the photo, there are four navigation links: "About GitHub", "Our Team", "Press" (which is underlined in orange), and "Jobs". In the "Press" section, there's a text block that reads: "There are **6.1M** people collaborating right now across **13.2M** repositories on GitHub. Developers from all around the world are building amazing things together. Their story is our story." A small arrow points from this text block down towards the bottom right corner of the page. At the very bottom left, there are two links: "GitHub: Innovations That Mattered in 2013" and "GitHub's Tom Preston-Werner: How We Went Mainstream". On the bottom right, there's a contact email "press@github.com" and a logo for "GitHub Fast Shoot".

There are **6.1M** people collaborating right now across **13.2M** repositories on GitHub. Developers from all around the world are building amazing things together. Their story is our story.

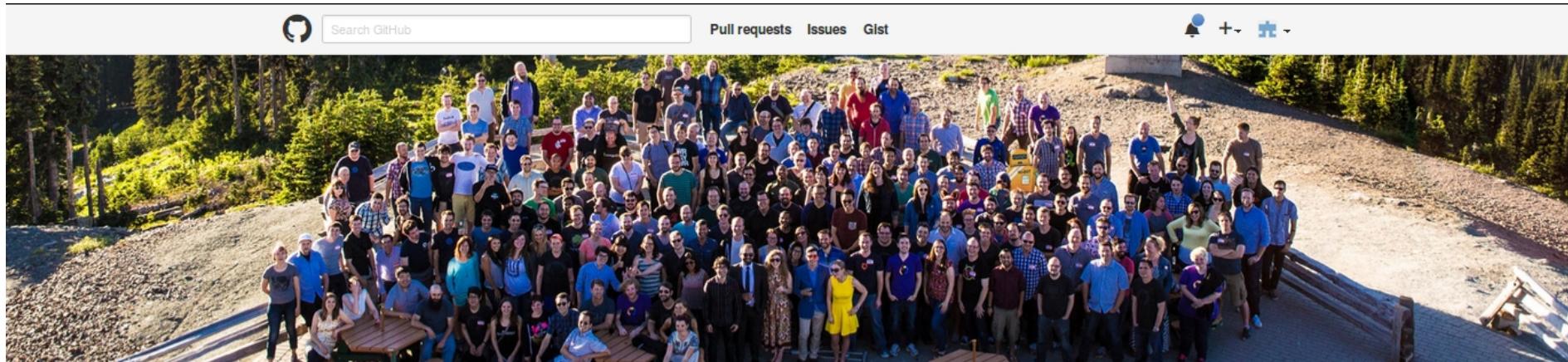
GitHub: Innovations That Mattered in 2013  
GovLoop, December 12, 2013

GitHub's Tom Preston-Werner: How We Went Mainstream  
ReadWriteWeb November 18, 2013

press@github.com

'Their story is our story'

# November 2013

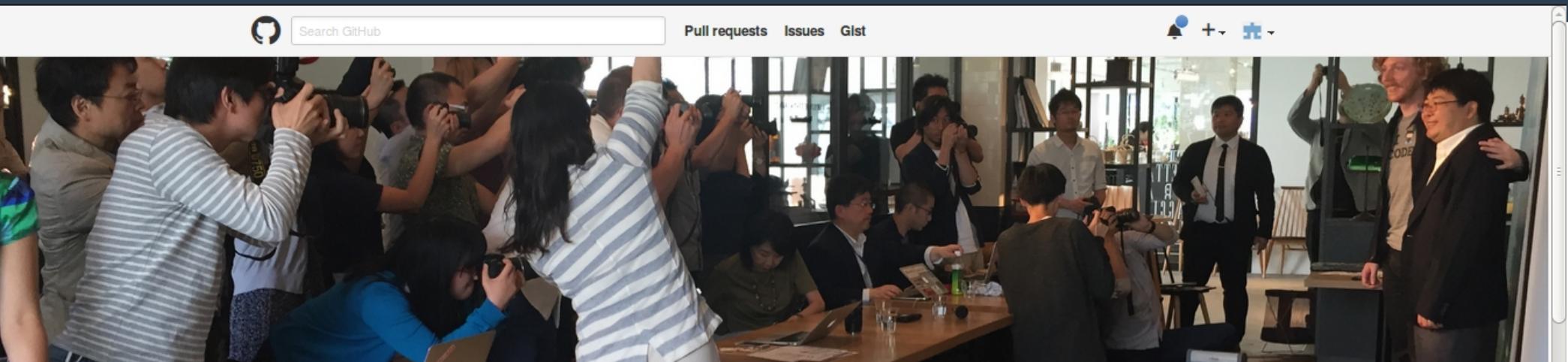
[About GitHub](#)[Our Team](#)[Press](#)[Jobs](#)

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](#)

# November 2015: China denial of service attack



About GitHub

Our Team

**Press**

Jobs

There are **11.8M** people collaborating right now across **29.9M** repositories on GitHub. Developers from all around the world are building amazing things together. Their story is our story.

**Awards**

2014

CNBC: Disruptor 50

MIT Tech Review: 50 Smartest Companies

Fast Company: Most Innovative Companies

 Founded in February, 2008

 HQ in San Francisco

 \$100M series A from a16z

 ...

# 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

```
{  
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. "payload": {  
15. "action": "started"  
16. },  
17. "public": true,  
18. "created_at": "2014-05-23T08:40:56Z",  
19. "org": {  
20. "id": 5497318,  
21. "login": "OpenSensorsIO",  
22. "gravatar_id": "1e0218942846ec8ef59f5d679dbca78",  
23. "url": "https://api.github.com/orgs/OpenSensorsIO",  
24. "avatar_url": "https://avatars.githubusercontent.com/  
25. }  
},  
},
```

Actor information



Repository Information

Azondi



Azondi is a data digestion service. ↗

# 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**

**Ilya Grigorik**

briandoll@github.com

igrigorik@google.com

@briandoll

@igrigorik

# The data has always/already been collected by someone

igrigorik / githubarchive.org

Watch 54 Star 1,061 Fork 73

Code Issues 4 Pull requests 0 Pulse

Graphs

Contributors

Commits

Code frequency

Punch card

Network

Members

Mar 11, 2012 – Nov 25, 2015

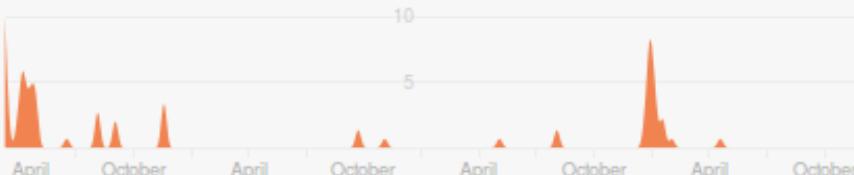
Contributions: Commits

Contributions to master, excluding merge commits



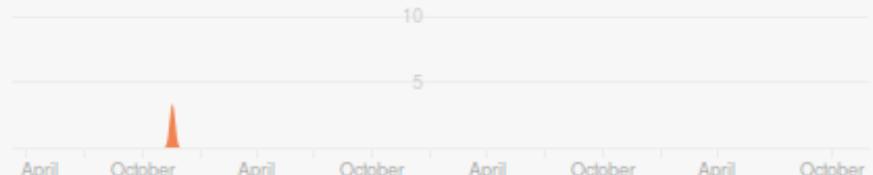
igrigorik

79 commits / 9,808 ++ / 9,555 --



klangner

5 commits / 205 ++ / 91 --



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

The screenshot shows the GitHub Archive website. At the top, there's a navigation bar with links for "Geogic Analytics", "Google+ Platform", "Reinvigorate", "StatHat", "Twitter Button", and "Unofficial GitHub Buttons". Below the header, there's a main content area with a "git" icon and the text "GitHub Archive". It includes social sharing buttons for GitHub, Star (653), Google+, and Twitter. A large text block explains the project's purpose: "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." Below this, there's a section for "daily top new & watched repository reports" with a "Sign up here" link and a GitHub logo with "819 Subscribers". Another section details "18 event types" and provides wget commands for different time periods. A note at the bottom states "Note: timeline data is available starting February 12, 2011." Finally, there's a Ruby example for processing JSON events.

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

git GitHub Archive

Star 653

Google+ Twitter

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	<code>wget http://data.githubarchive.org/2012-04-11-15.json.gz</code>
Activity for April 11, 2012	<code>wget http://data.githubarchive.org/2012-04-11-{0..23}.json.gz</code>
Activity for April 2012	<code>wget http://data.githubarchive.org/2012-04-{01..31}-{0..23}.json.gz</code>

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 'yaml'
4
5 gz = open('http://data.githubarchive.org/2012-03-11-12.json.gz')
```

COMPOSE QUERY

Query History  
Job History

## 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

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

githubarchive	
└─ github	
└─ actor_frequency	
└─ actor_location	
└─ twitter_forks	
└─ github_explore	
└─ github_proper	
└─ stack_overflow	
└─ githubarchive:github	

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

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

Screenshot of a GitHub blog post titled "The GitHub Data Challenge".

The post was published on May 1, 2012, by briandoll under the Watercooler category.

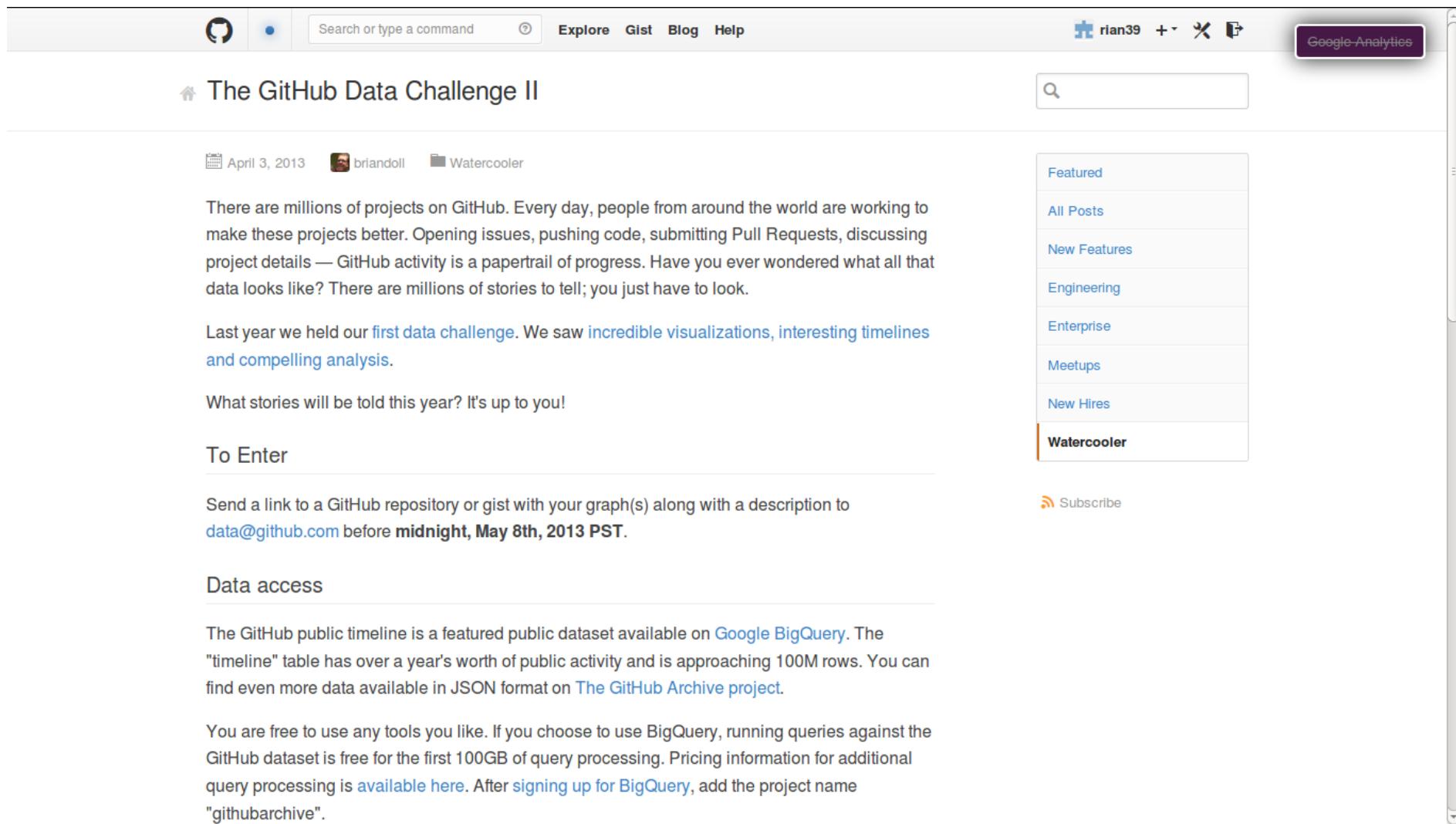
The text of the post reads: "The GitHub public timeline is now easy to query and analyze. With hundreds of thousands of events in the timeline every day, there are countless stories to tell."



The sidebar on the right lists categories: Featured, All Posts, New Features, Engineering, Enterprise, Meetups, New Hires, and Watercooler. The Watercooler category is highlighted with a border.

A link at the bottom of the page is: <https://camo.githubusercontent.com/5e6eb0b00d714eb5b8ec84254205c61c2a97c68d/687474703a2f2f6f63746f6465782e6769746875622e636f6d2f696d616765732f77616c646f6361742e6a7067>

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



The screenshot shows a GitHub page titled "The GitHub Data Challenge II". The page features a search bar at the top right and navigation links for Explore, Gist, Blog, and Help. A sidebar on the right lists categories like Featured, All Posts, New Features, Engineering, Enterprise, Meetups, New Hires, and Watercooler, with "Watercooler" highlighted. The main content discusses GitHub activity and past challenges, and provides instructions for entering the current challenge by sending a link to data@github.com by May 8th, 2013 PST. It also mentions the availability of GitHub data on Google BigQuery and in JSON format.

Search or type a command

Explore Gist Blog Help

rian39 + × ↗ Google Analytics

## The GitHub Data Challenge II

April 3, 2013 briandoll Watercooler

There are millions of projects on GitHub. Every day, people from around the world are working to make these projects better. Opening issues, pushing code, submitting Pull Requests, discussing project details — GitHub activity is a papertrail of progress. Have you ever wondered what all that data looks like? There are millions of stories to tell; you just have to look.

Last year we held our [first data challenge](#). We saw [incredible visualizations](#), [interesting timelines](#) and [compelling analysis](#).

What stories will be told this year? It's up to you!

### To Enter

Send a link to a GitHub repository or gist with your graph(s) along with a description to [data@github.com](mailto:data@github.com) before **midnight, May 8th, 2013 PST**.

### Data access

The GitHub public timeline is a featured public dataset available on [Google BigQuery](#). The "timeline" table has over a year's worth of public activity and is approaching 100M rows. You can find even more data available in JSON format on [The GitHub Archive project](#).

You are free to use any tools you like. If you choose to use BigQuery, running queries against the GitHub dataset is free for the first 100GB of query processing. Pricing information for additional query processing is [available here](#). After signing up for BigQuery, add the project name "githubarchive".

Featured

All Posts

New Features

Engineering

Enterprise

Meetups

New Hires

**Watercooler**

Subscribe

# Typical stories: “all individuals are similar”

The screenshot shows a web page titled "THE OPEN SOURCE REPORT CARD". At the top, there is a teal header with the text "Dear recruiters: While you read this, make sure that you remember that GitHub is not your CV, and that these stats only provide a *biased and one-sided view*. This is just a toy. Don't take it too seriously!" followed by a button labeled "OK. I promise!". In the top right corner of the header, there is a small box containing "SayHelloAds" and "Google Analytics". Below the header, the main content area has a white background. It features a large input field with the placeholder "Enter a GitHub username". Above this input field, there is a descriptive text: "Enter a GitHub username to see a dynamically generated progress report for their open source contributions". Further down, there is a section titled "HOW IT WORKS" with a detailed explanation of how the service uses GitHub data to generate reports. At the bottom of the page, there is a footer with some additional text and links.

Dear recruiters:  
While you read this, make sure that you remember that GitHub is not your CV, 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!

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.’



# Story: 'coding is hard work'



GEEKSTA

Sections

Projects

Geeklog

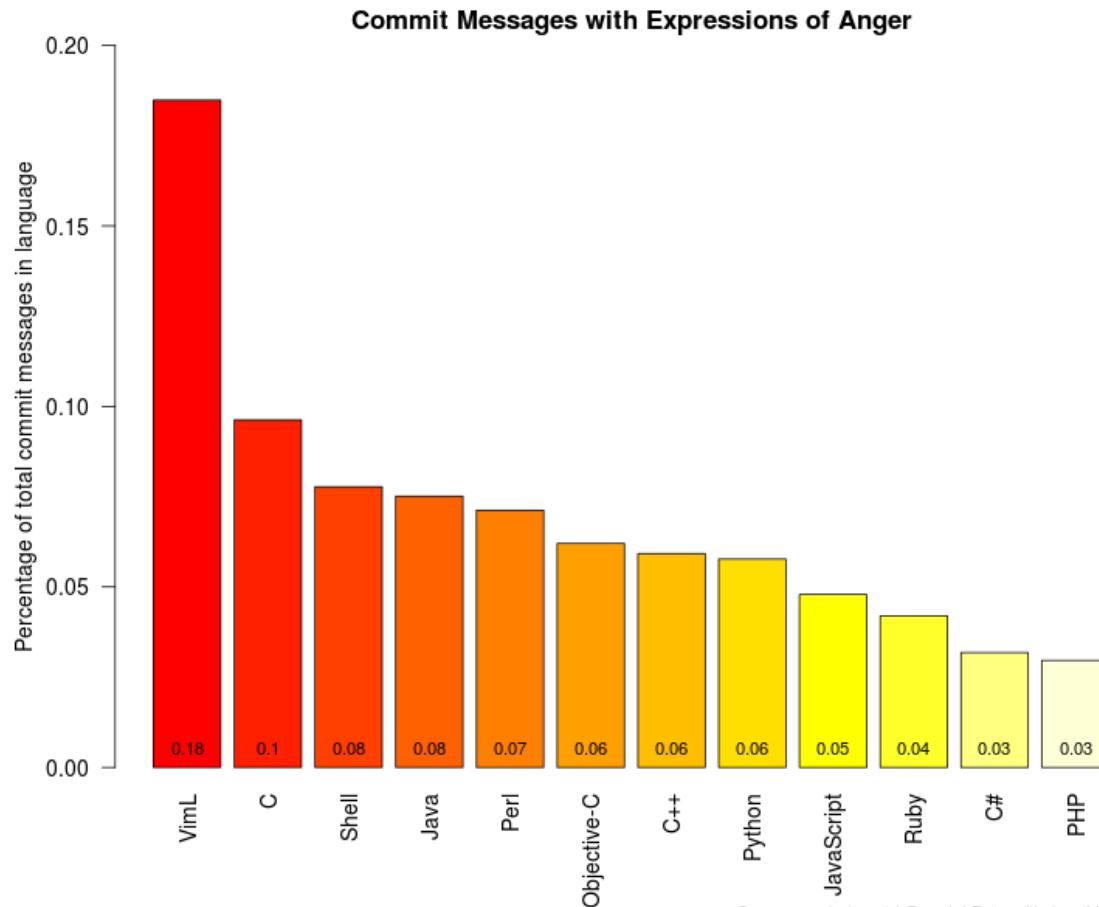
Tools

Shirts



Regular Expression

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



Source: geeksta.net / @yaph | Data: githubarchive.org

# Story: 'we need live numbers'

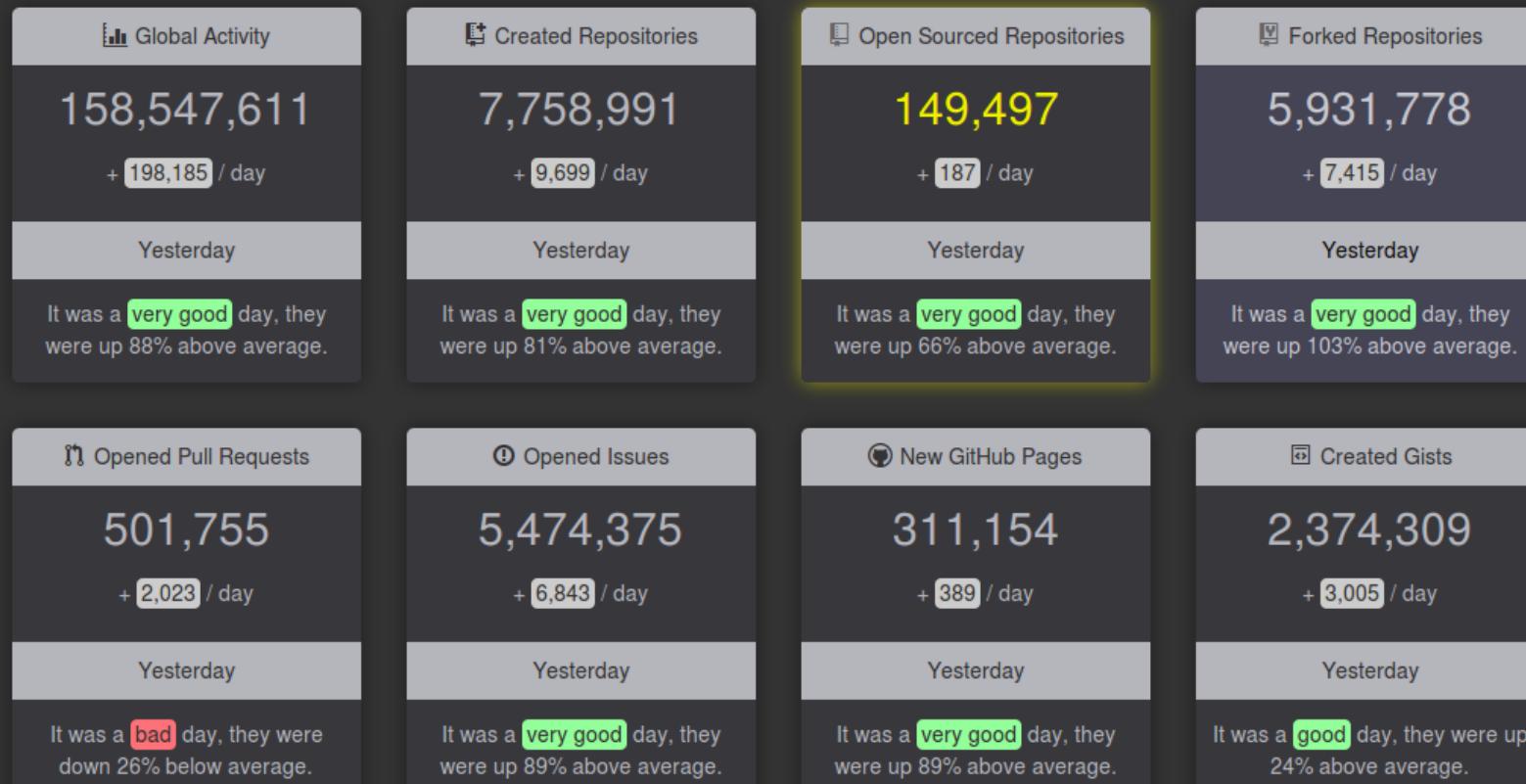
Octoboard

Global activity Repository Public Fork Pull Request Issue GH Pages Google Analytics

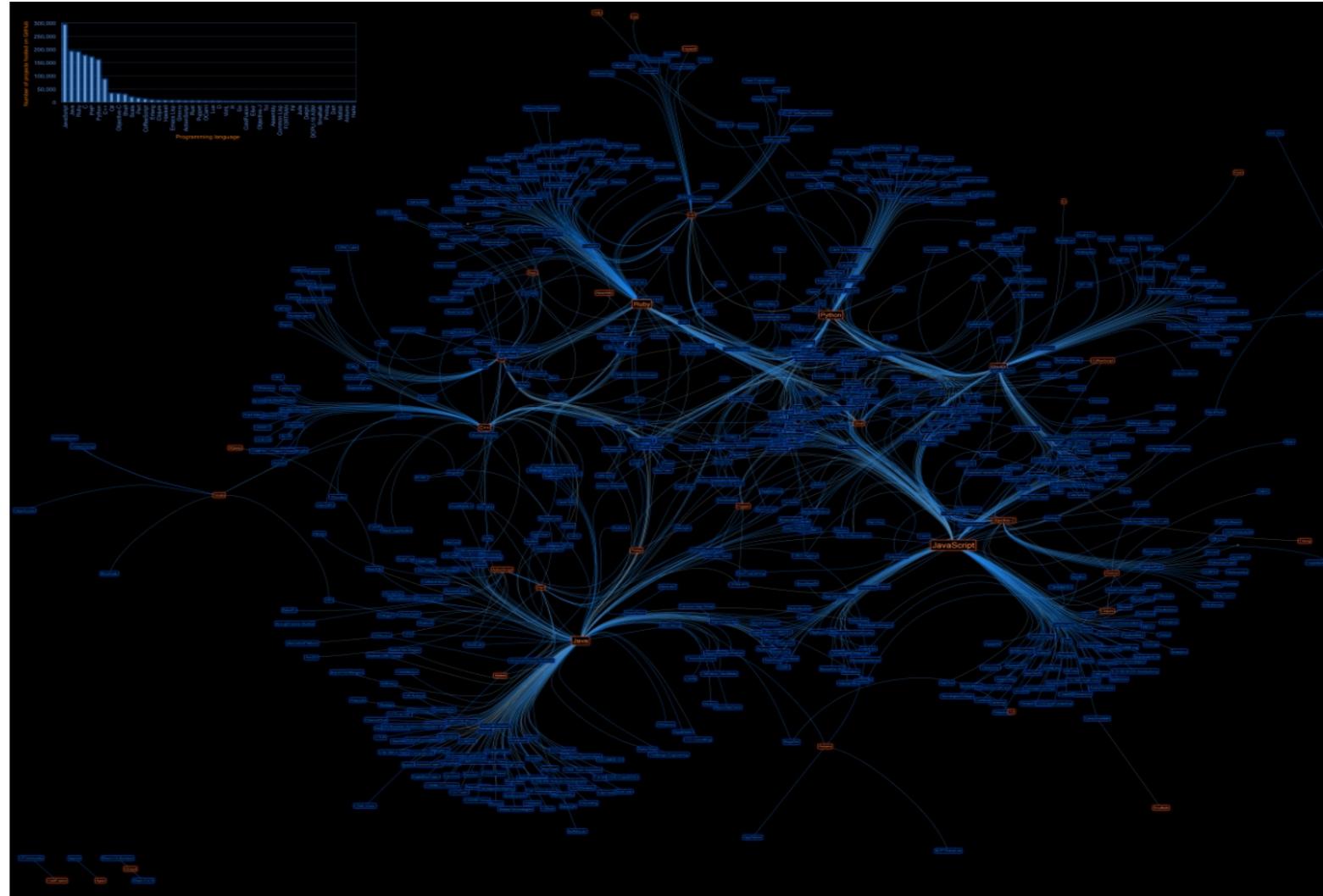
## GitHub activity dashboard.

Octoboard is based on GitHub Archive : each day, it scans new GitHub events archives and computes a few stats, with a 15 days history. You can see some general data on this page, or use menu for more information about language and history.

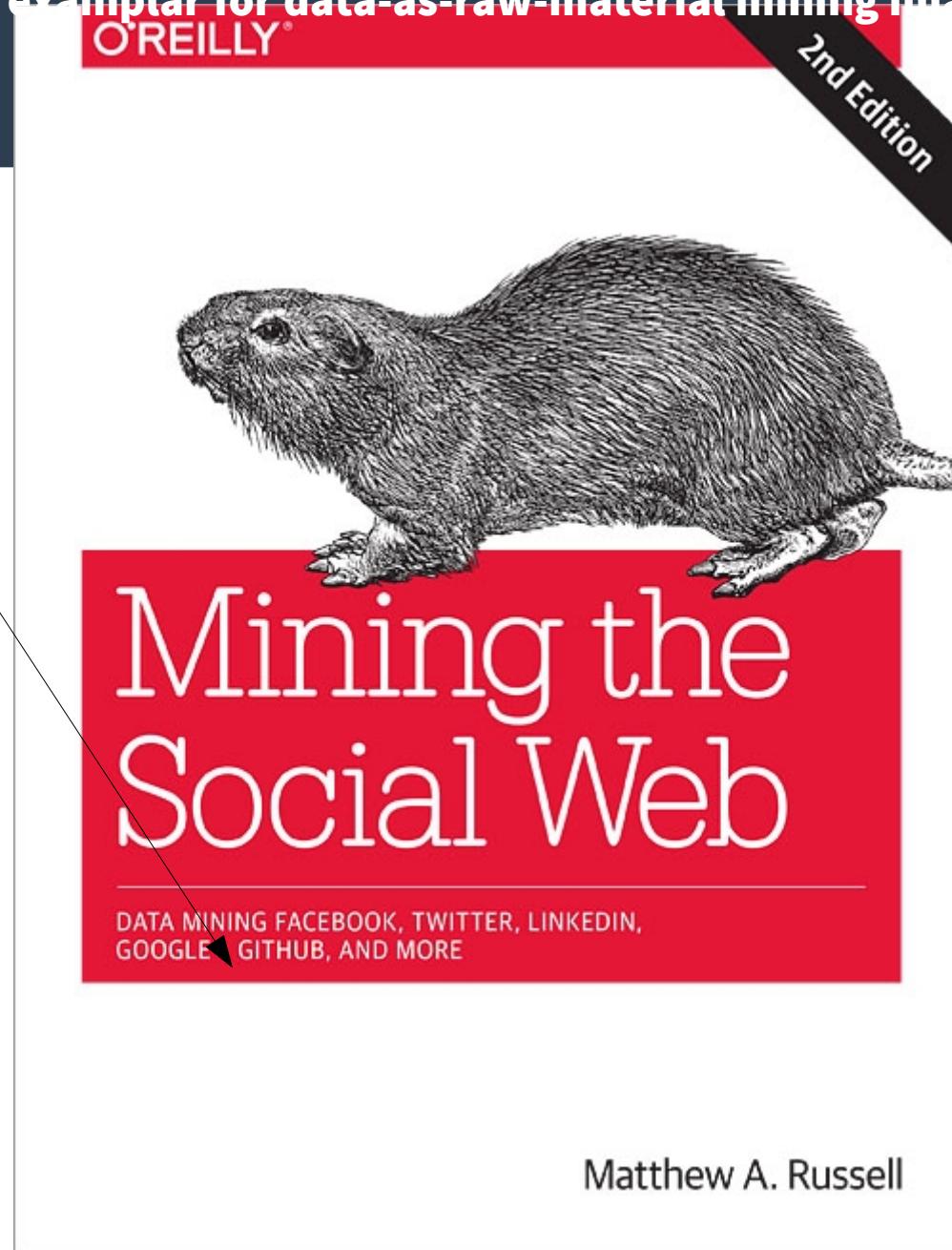
Since march 11th, 2012 :



# Story: 'Code is a social network'



**Story: 'Github data is an exemplar 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) Python 2

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 data

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

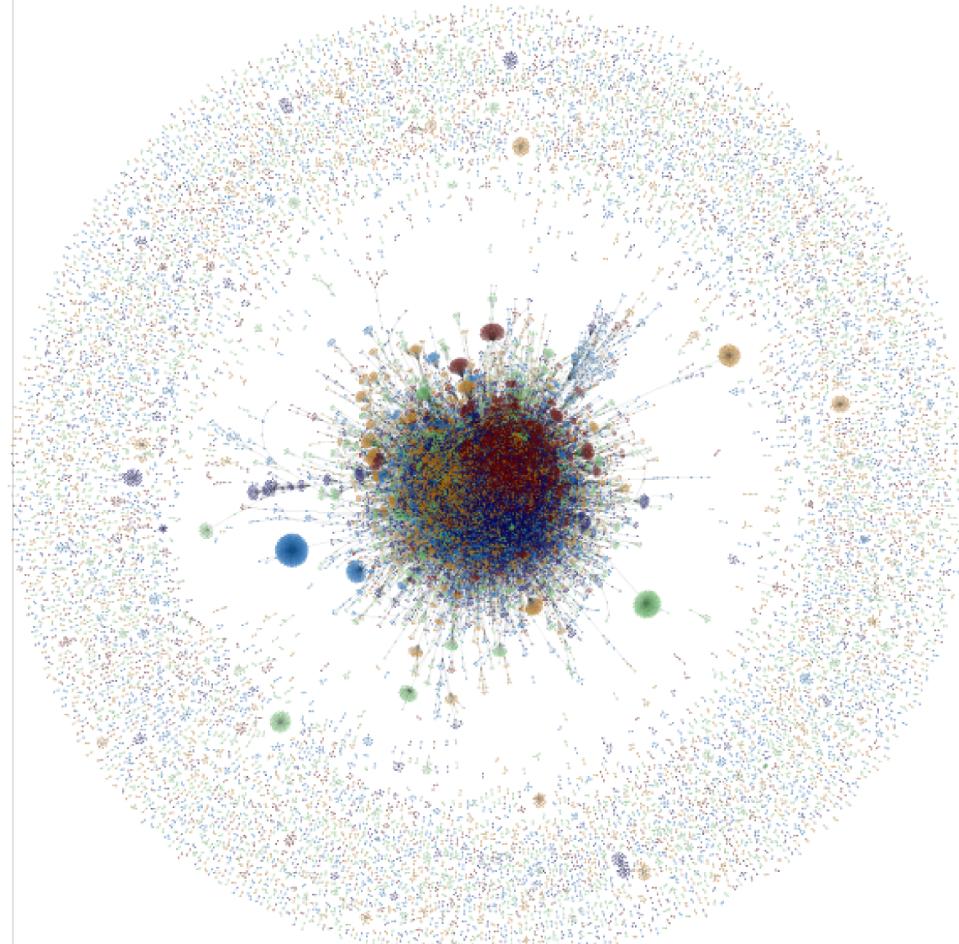
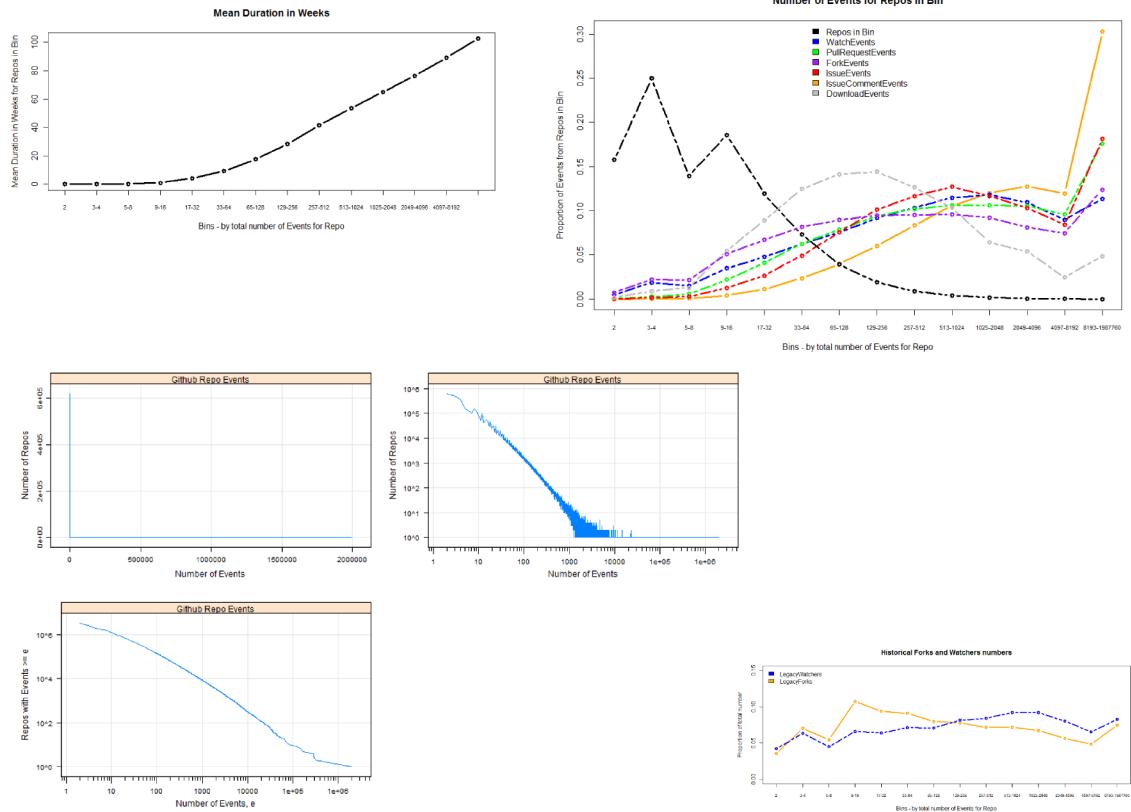
  

```
In [17]: [[i, lda_model.print_topic(i)] for i in range(15)]
```

```
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())
```

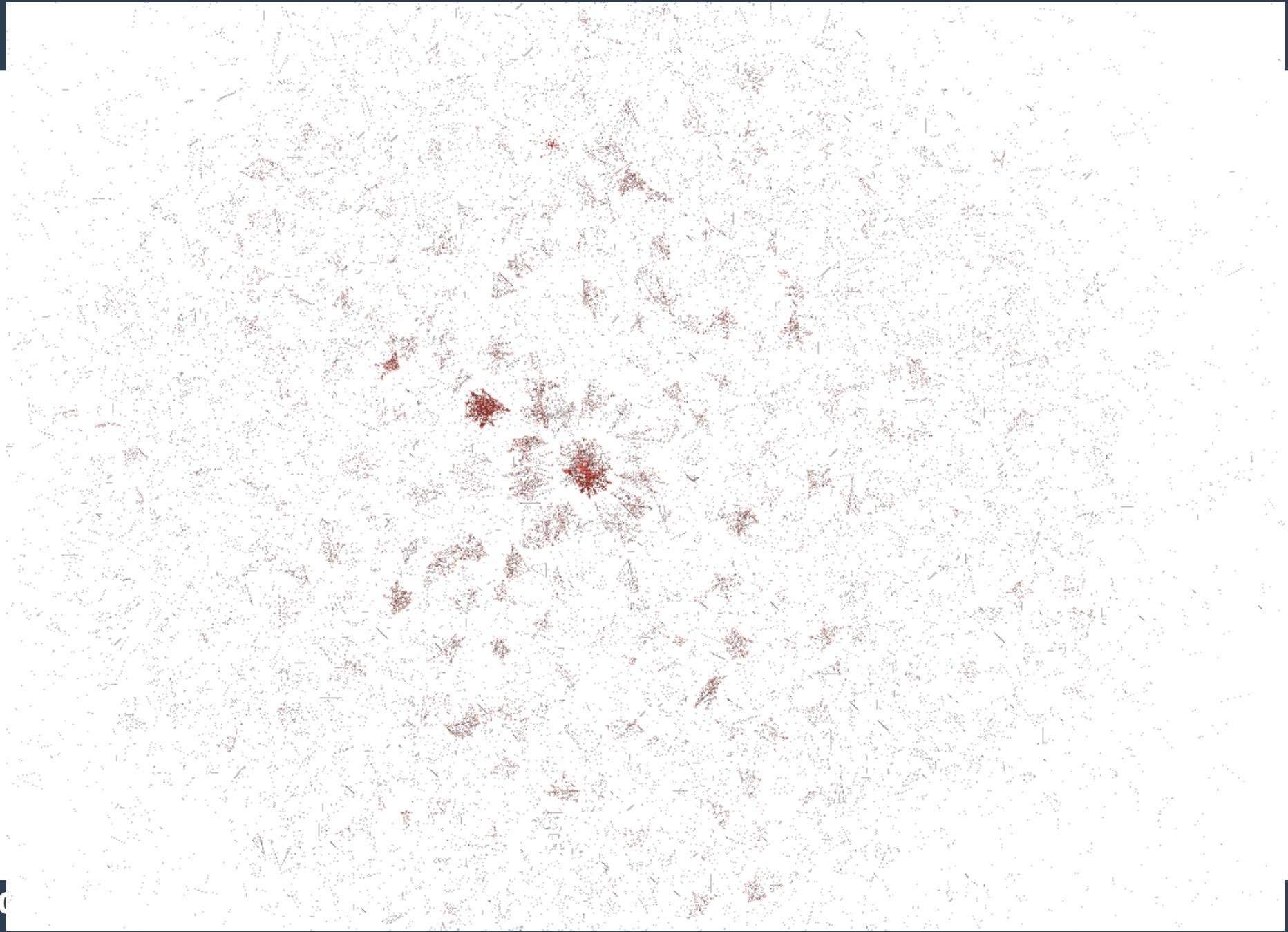
# Statistician says 'no' 'Digital sociologist' says 'yes'



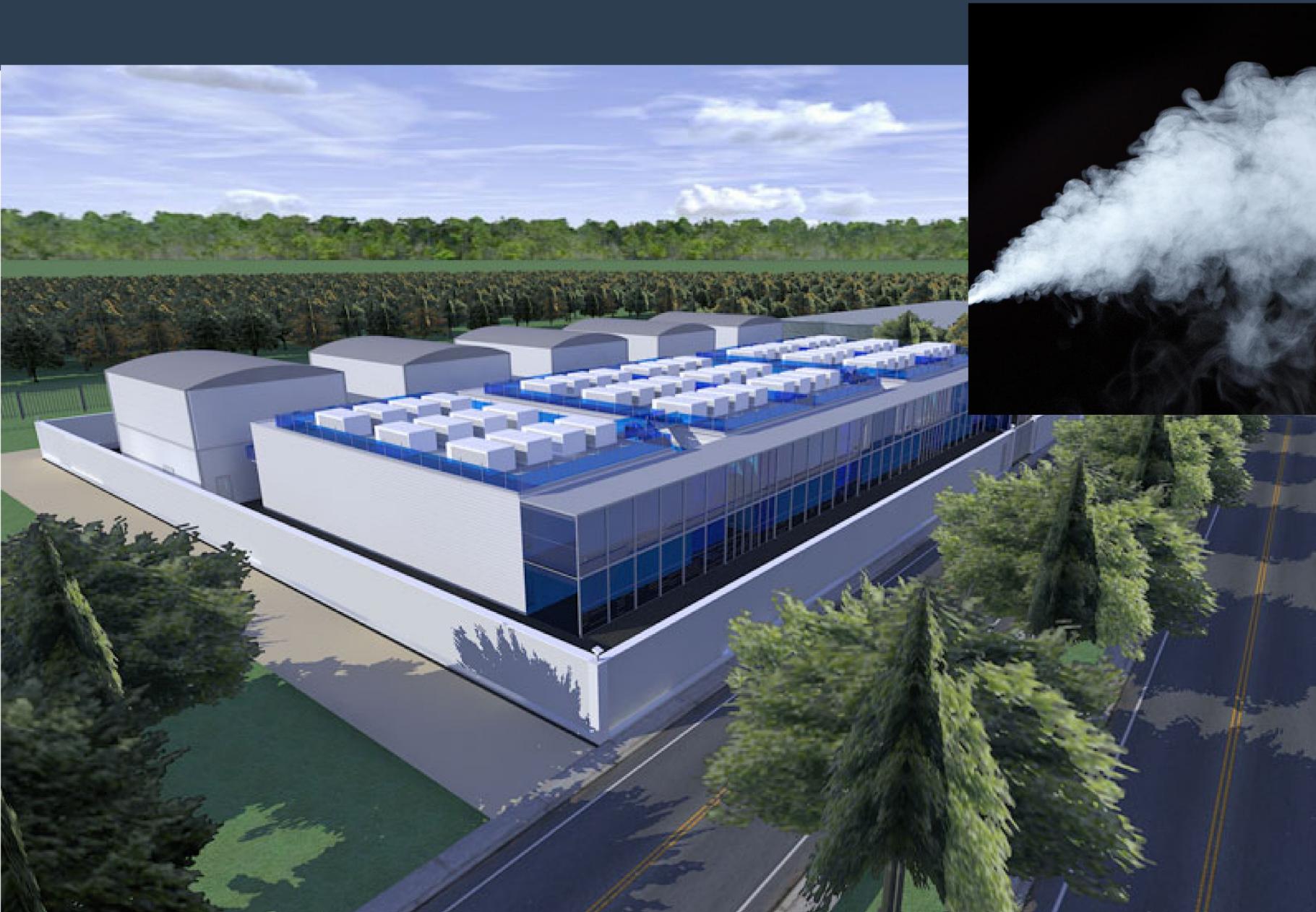
statistics

digital sociology?

# Statistician says 'yes'



# Steam billows from one query ...



Projects

Billing

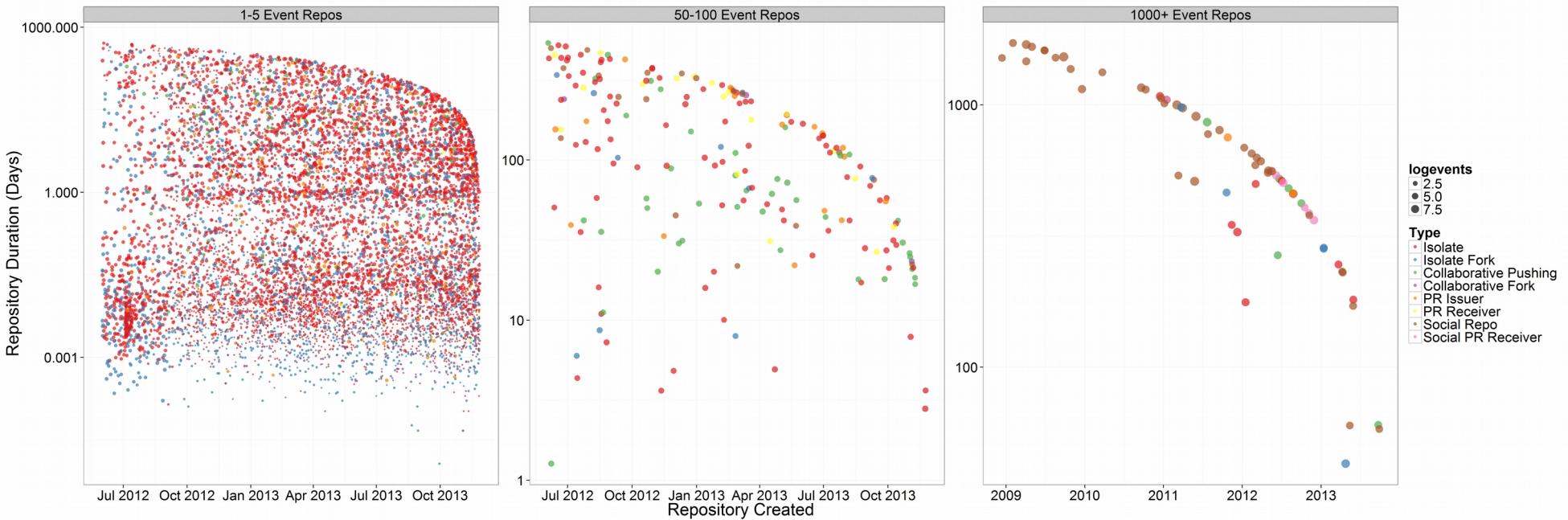
Account settings

	Oct 1, 2013	Starting balance		10.54
	Sep 1, 2013 - Sep 30, 2013		\$10.54	(\$2,265.66)
<a href="#">Documents (2)</a>				
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)
<a href="#">Documents (2)</a>				
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)
<a href="#">Documents (2)</a>				
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
<a href="#">Documents (2)</a>				

[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**

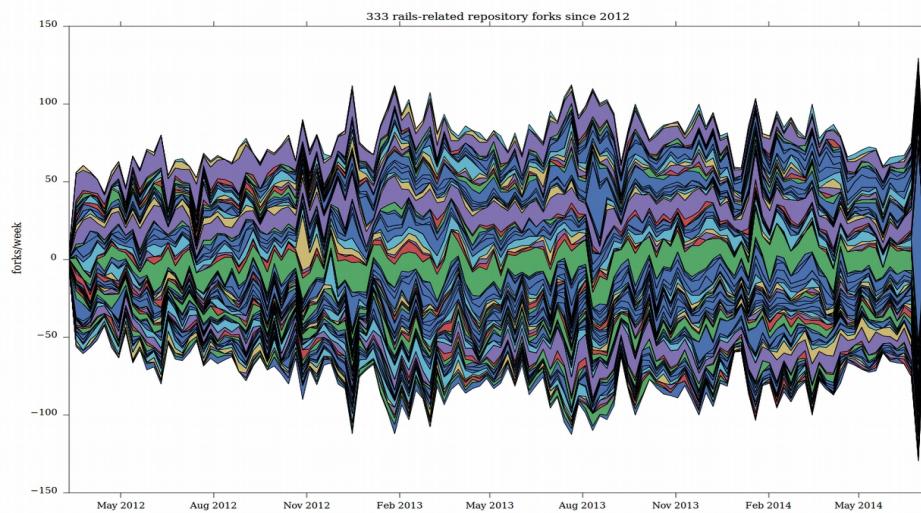
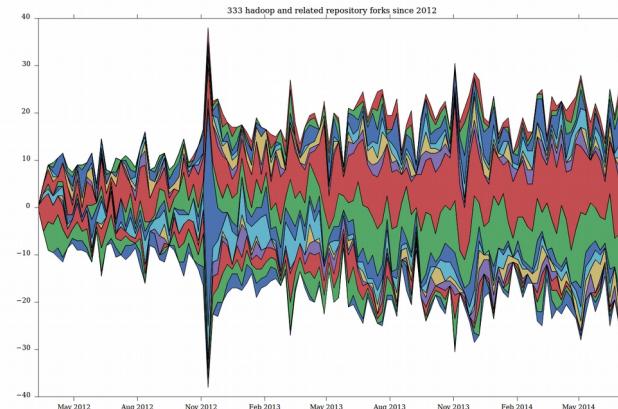
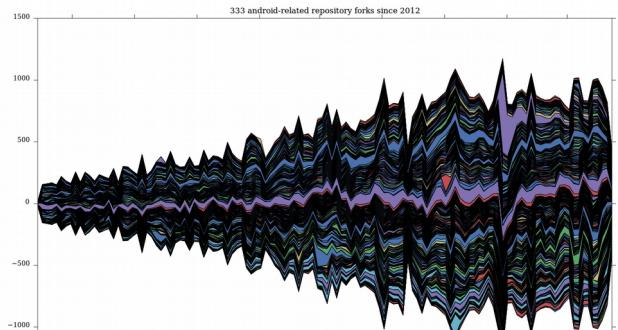
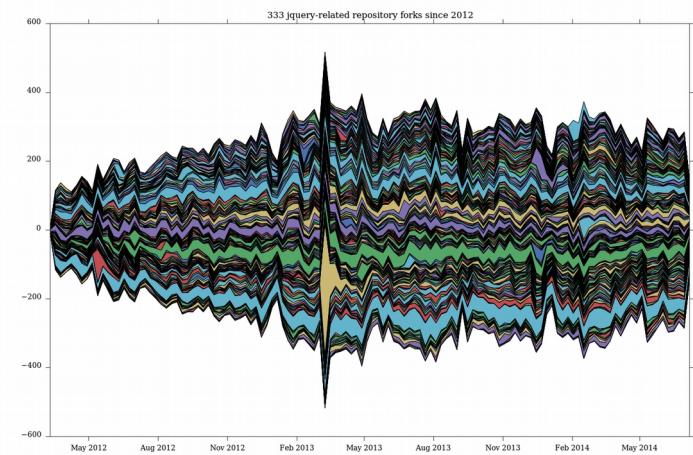
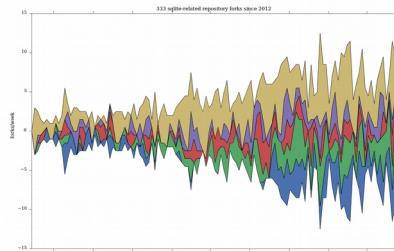
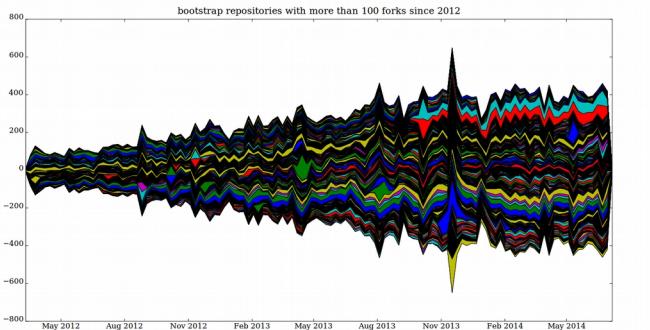
## Early repositories ( c. 2008)

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 Oksiane/Websockets  
23 chenxiruanhai/XScrollView  
24 jhonM17/footer-fixed-bootstrap  
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 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