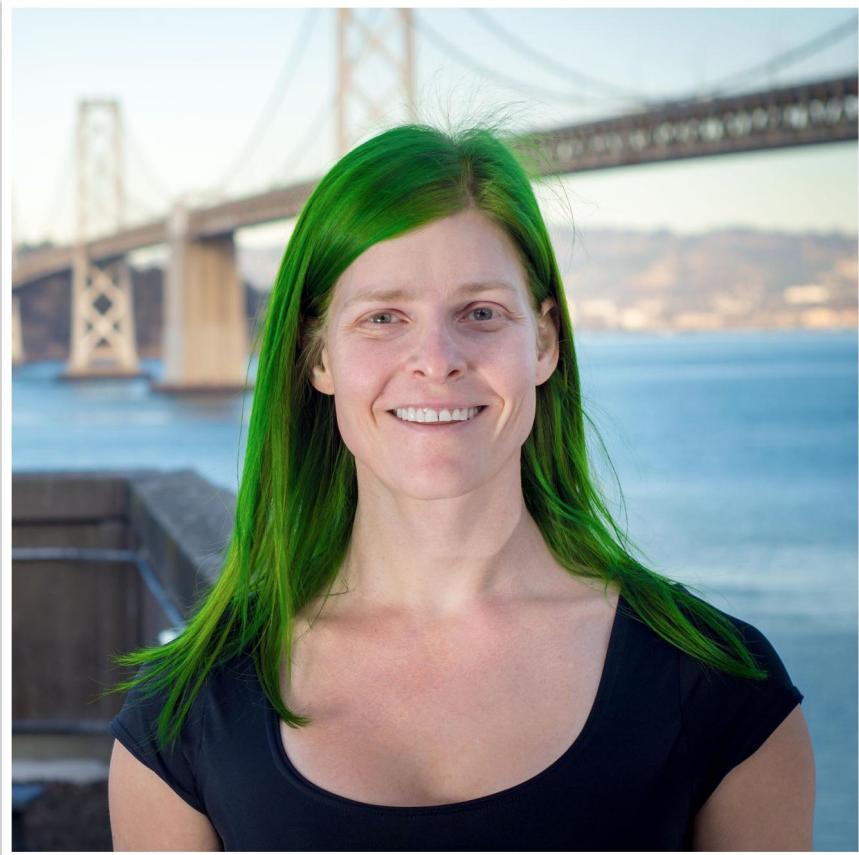




# Exploring Open Data with BigQuery



Google Cloud Platform



# Jen Tong

Developer Advocate  
Google Cloud Platform

[@MimmingCodes](https://twitter.com/MimmingCodes)  
[little418.com](http://little418.com)



# Agenda

- Origin story
- Count stuff
- How it works
- Some cool open data
- Do something useful



# Google Research Publications

The image shows two side-by-side screenshots of Mac OS X desktop environments. Both screens feature a PDF viewer window with a toolbar at the top and a sidebar on the left.

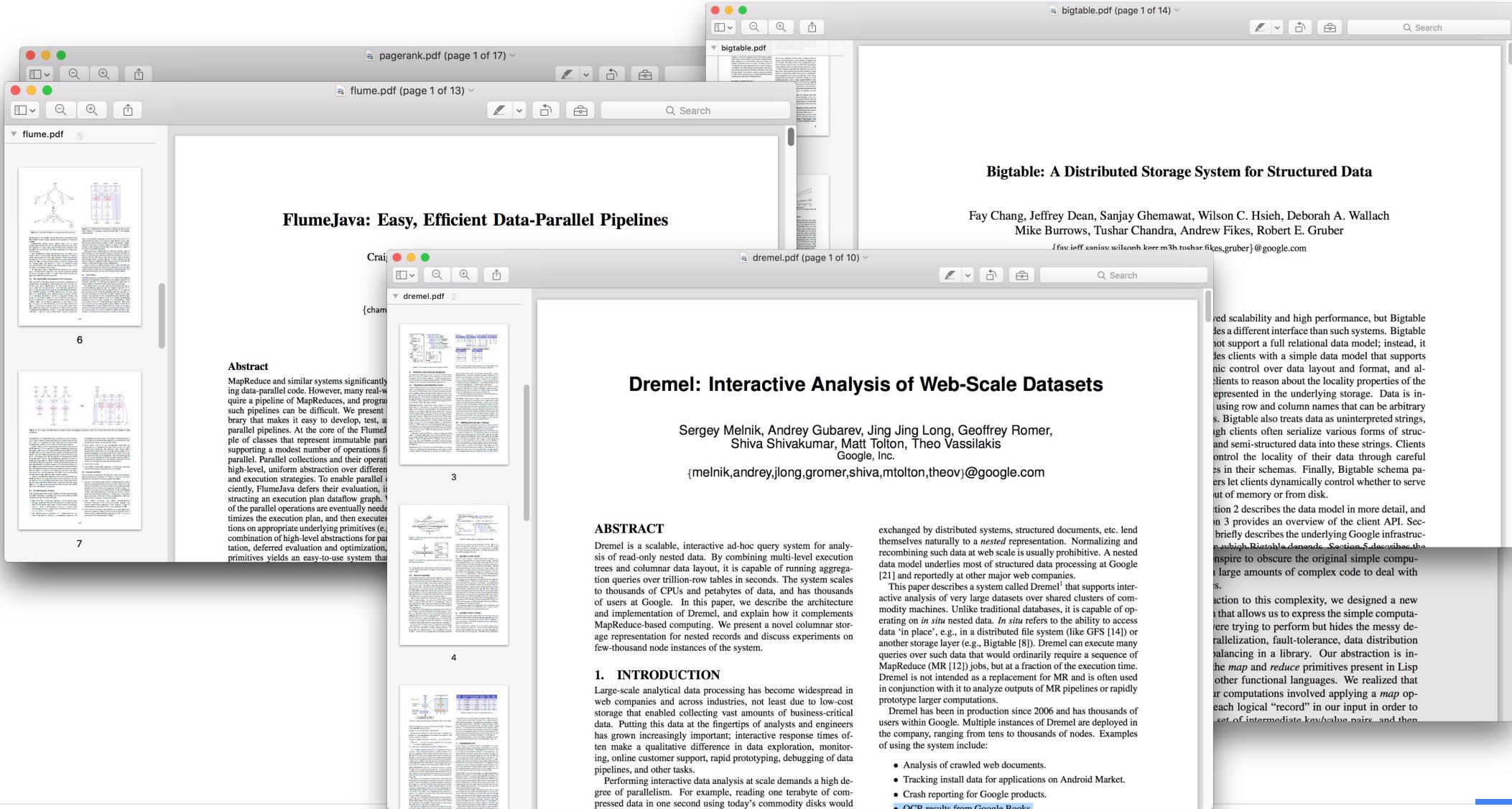
**Left Screen (pagerank.pdf):**

- Title:** The PageRank Citation Ranking: Bringing Order to the Web
- Date:** January 29, 1998
- Abstract:** The importance of a Web page is an inherently subjective matter, which readers interests, knowledge and attitudes. But there is still much that can be about the relative importance of Web pages. This paper describes PageRank rating Web pages objectively and mechanically, effectively measuring the human attention devoted to them.
- Introduction and Motivation:** The World Wide Web creates many new challenges for information retrieval. It is heterogeneous. Current estimates are that there are over 150 million web pages.

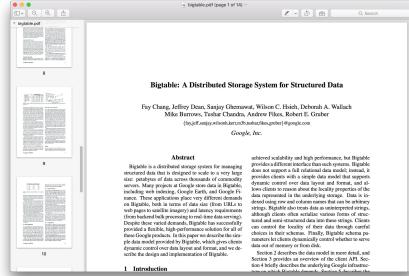
**Right Screen (mapreduce.pdf):**

- Title:** MapReduce: Simplified Data Processing on Large Clusters
- Authors:** Jeffrey Dean and Sanjay Ghemawat
- Emails:** jeff@google.com, sanjay@google.com
- Organization:** Google, Inc.
- Abstract:** MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a *map* function that processes a key/value pair to generate a set of intermediate key/value pairs, and a *reduce* function that merges all intermediate values associated with the same intermediate key. Many real world tasks are expressible in this model, as shown in the paper.
- Text:** Programs written in this functional style are automatically parallelized and executed on a large cluster of commodity machines. The run-time system takes care of the details of partitioning the input data, scheduling the program's execution across a set of machines, handling machine failures, and managing the required inter-machine communication. This allows programmers without any experience with parallel and distributed systems to easily utilize the resources of a large distributed system.
- Given Day:** given day, etc. Most such computations are conceptually straightforward. However, the input data is usually large and the computations have to be distributed across hundreds or thousands of machines in order to finish in a reasonable amount of time. The issues of how to parallelize the computation, distribute the data, and handle failures conspire to obscure the original simple computation with large amounts of complex code to deal with these issues.
- Reaction:** As a reaction to this complexity, we designed a new abstraction that allows us to express the simple computations we were trying to perform but hides the messy details of parallelization, fault-tolerance, data distribution and load balancing in a library. Our abstraction is inspired by the *map* and *reduce* primitives present in Lisp and many other functional languages. We realized that most of our computations involved applying a *map* operation to each logical "record" in our input in order to compute a set of intermediate key/value pairs, and then

# Google Research Publications

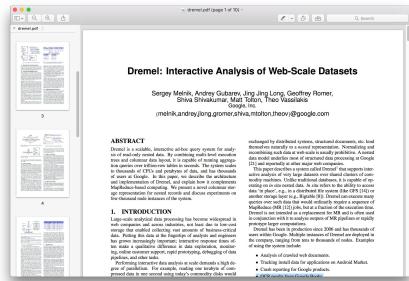


# Open Source Implementations



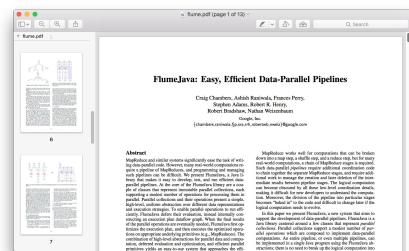
Bigtable →

APACHE  
**HBASE**



Flume →

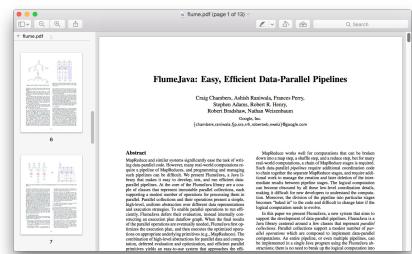
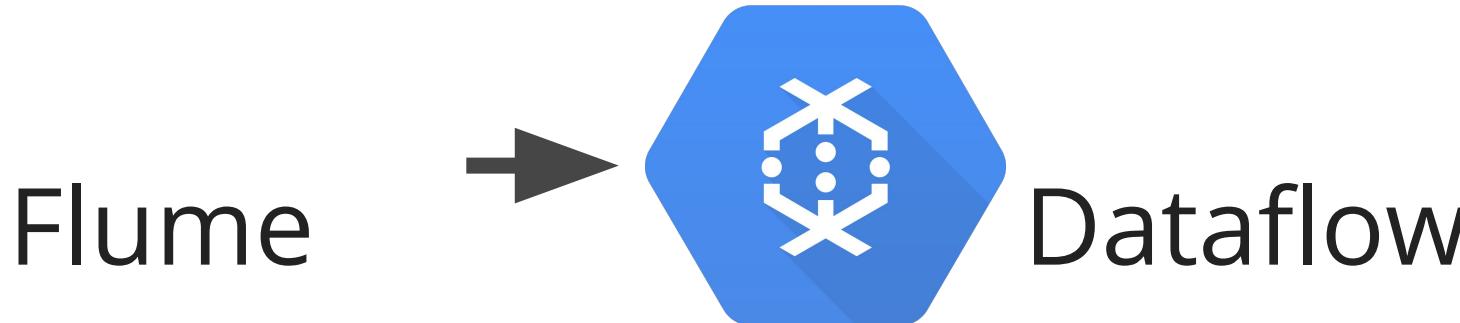
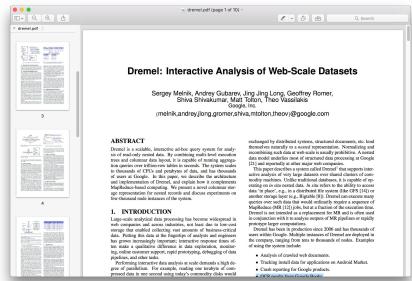
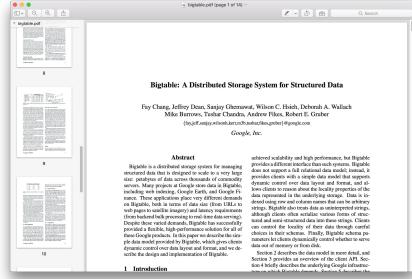
CRUNCH



Dremel →

APACHE  
**DRILL**

# Managed Cloud Versions





Google BigQuery

# Let's count some stuff

# Words in Shakespeare

```
SELECT count(word)  
FROM publicdata:samples.shakespeare
```

# Wikipedia hits over 1 hour

```
SELECT sum(requests) as total  
FROM [fh-bigquery:wikipedia.pagecounts_20150511_05]
```

# Wikipedia hits over 1 month

```
SELECT sum(requests) as total  
FROM [fh-bigquery:wikipedia.pagecounts_201505]
```

# Several years of Wikipedia data

```
SELECT sum(requests) as total  
FROM  
[fh-bigquery:wikipedia.pagecounts_201105],  
[fh-bigquery:wikipedia.pagecounts_201106],  
[fh-bigquery:wikipedia.pagecounts_201107],  
...
```

# Several years of Wikipedia data

```
SELECT
    SUM(requests) AS total
FROM
    TABLE_QUERY(
        [fh-bigquery:wikipedia],
        'REGEXP_MATCH(
            table_id,
            r"pagecounts_2015[0-9]{2}$")')
```

# How about a RegExp

```
SELECT
    SUM(requests) AS total
FROM
    TABLE_QUERY(
        [fh-bigquery:wikipedia],
        'REGEXP_MATCH(
            table_id,
            r"pagecounts_2015[0-9]{2}$")')
WHERE
    (REGEXP_MATCH(title, '.*[dD]inosaur.*'))
```

# How did it do that?

o\_O

# Qualities of a good RDBMS



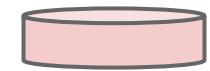
# Qualities of a good RDBMS

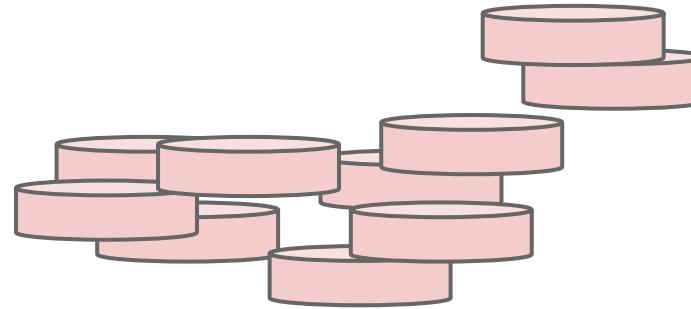
- Inserts & locking
- Indexing
- Cache
- Query planning

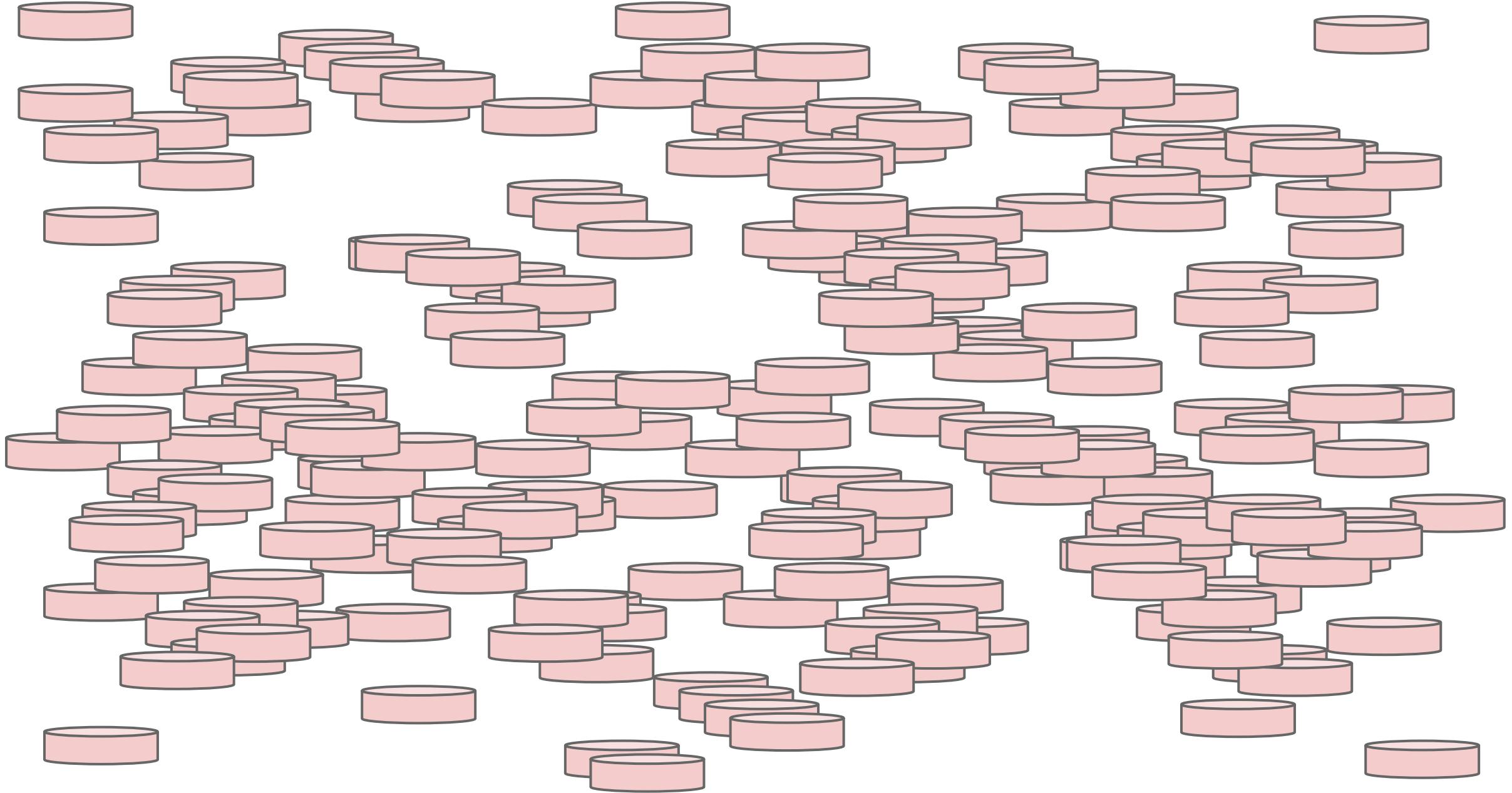


# Qualities of a good RDBMS

- 
- Inserts & locking
  - Indexing
  - Caching
  - Query planning





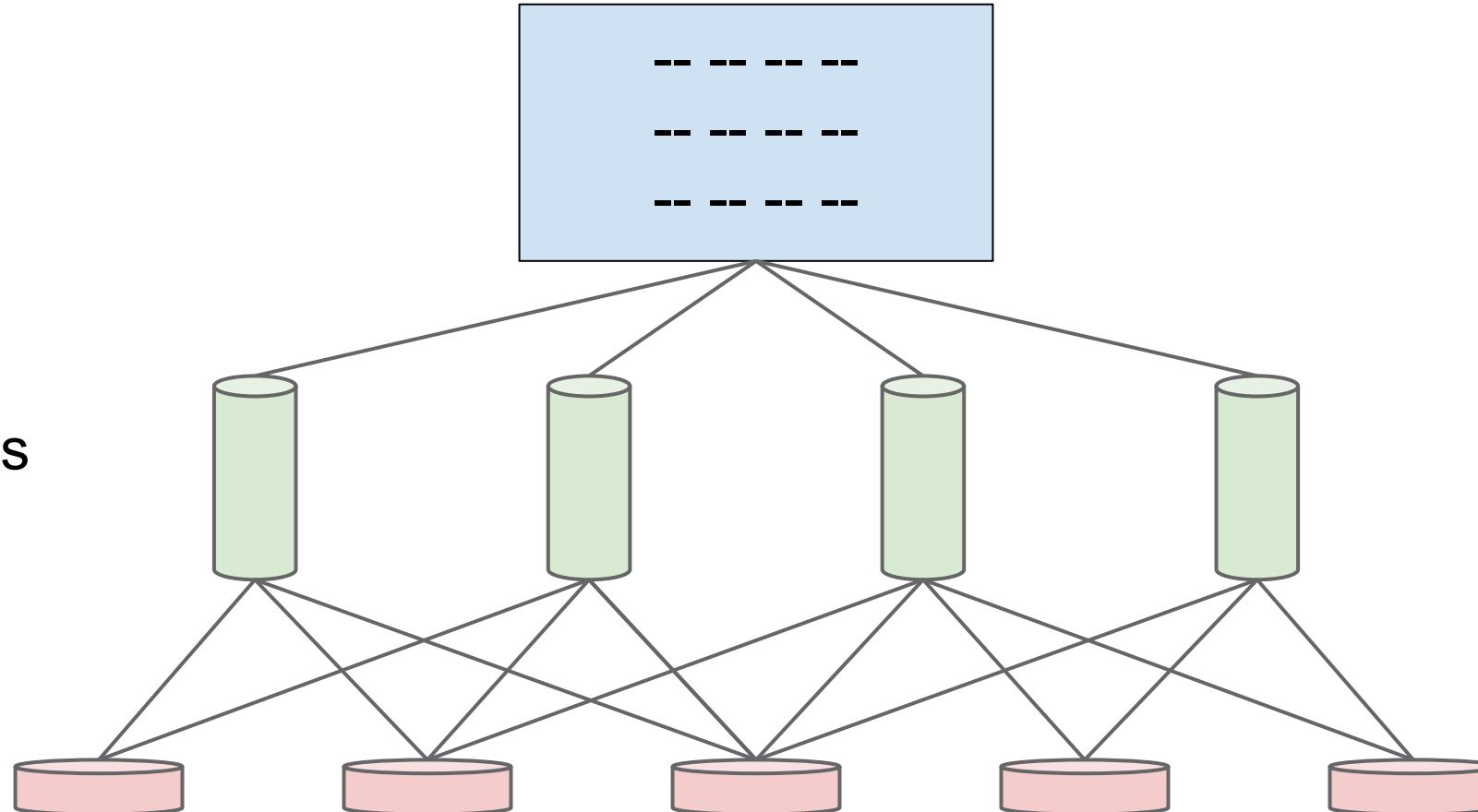


# Storing data

Table

Columns

Disks

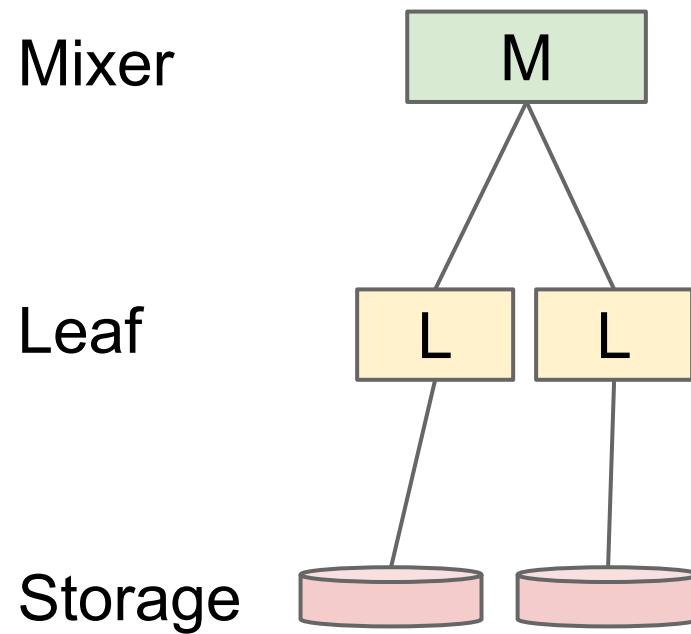


# Reading data: Life of a BigQuery

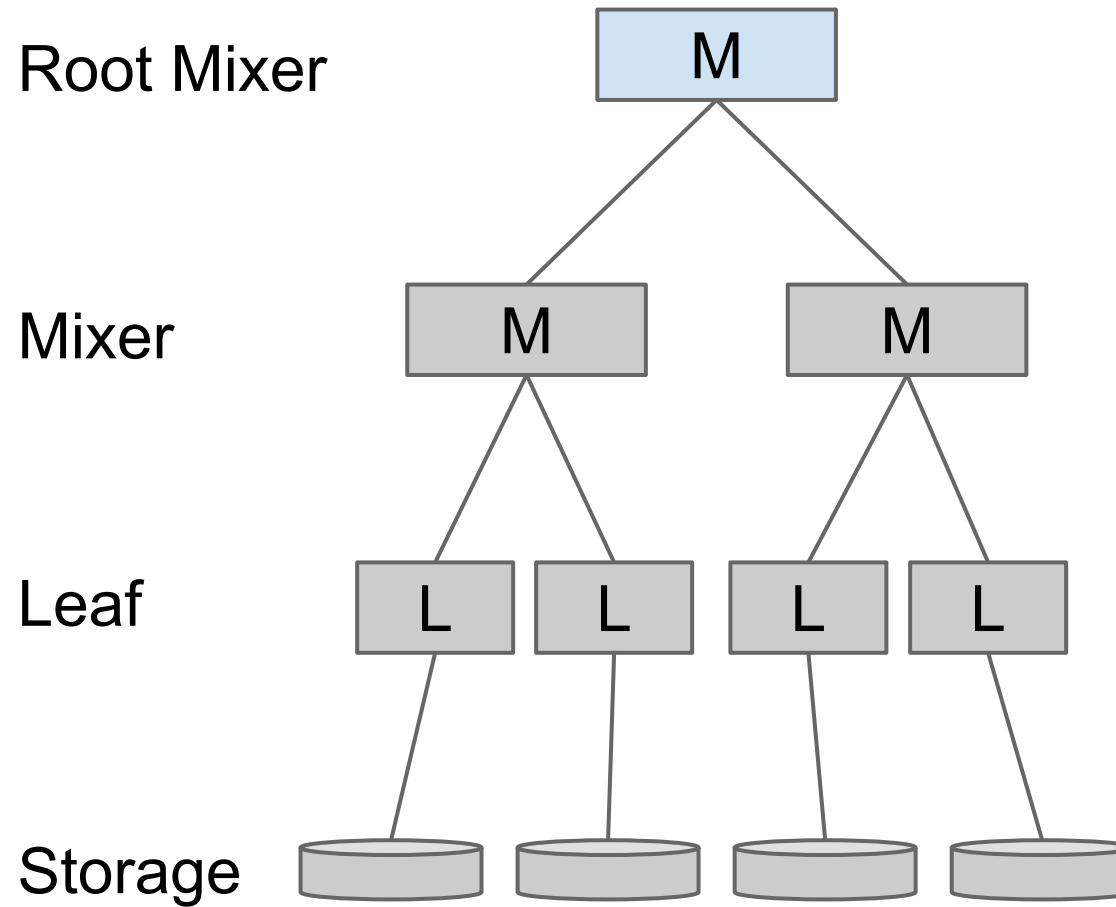
```
SELECT sum(requests) as sum
FROM (
    SELECT requests, title
    FROM [fh-bigquery:wikipedia.
pagecounts_201501]
    WHERE
        (REGEXP_MATCH(title, '[Jj]en.+'))
)
```



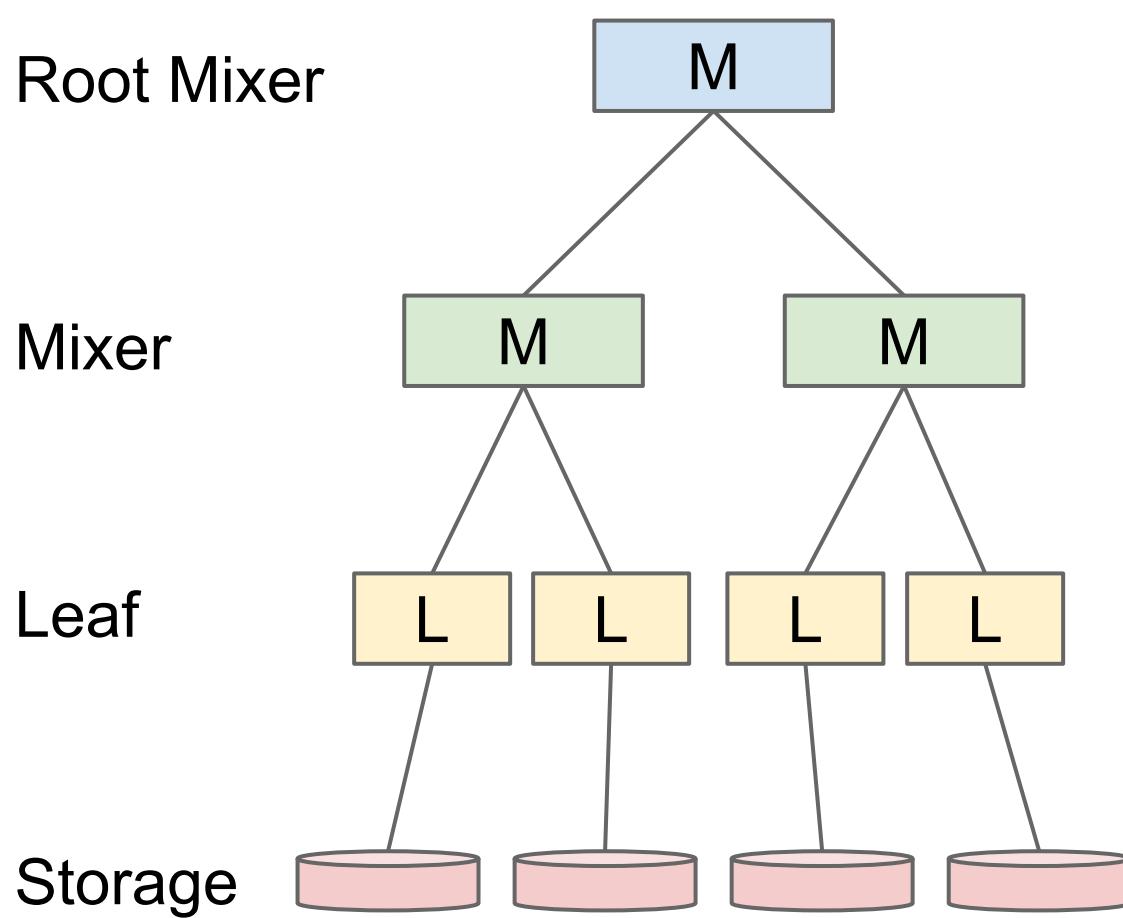
# Life of a BigQuery



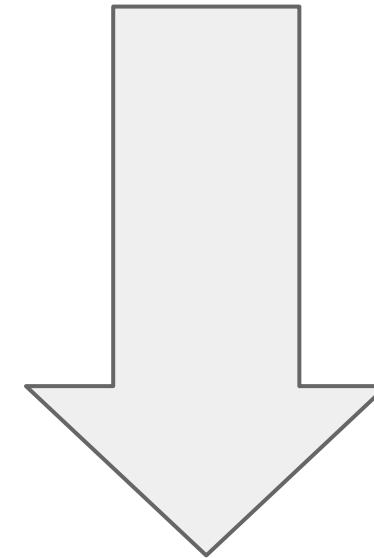
# Life of a BigQuery



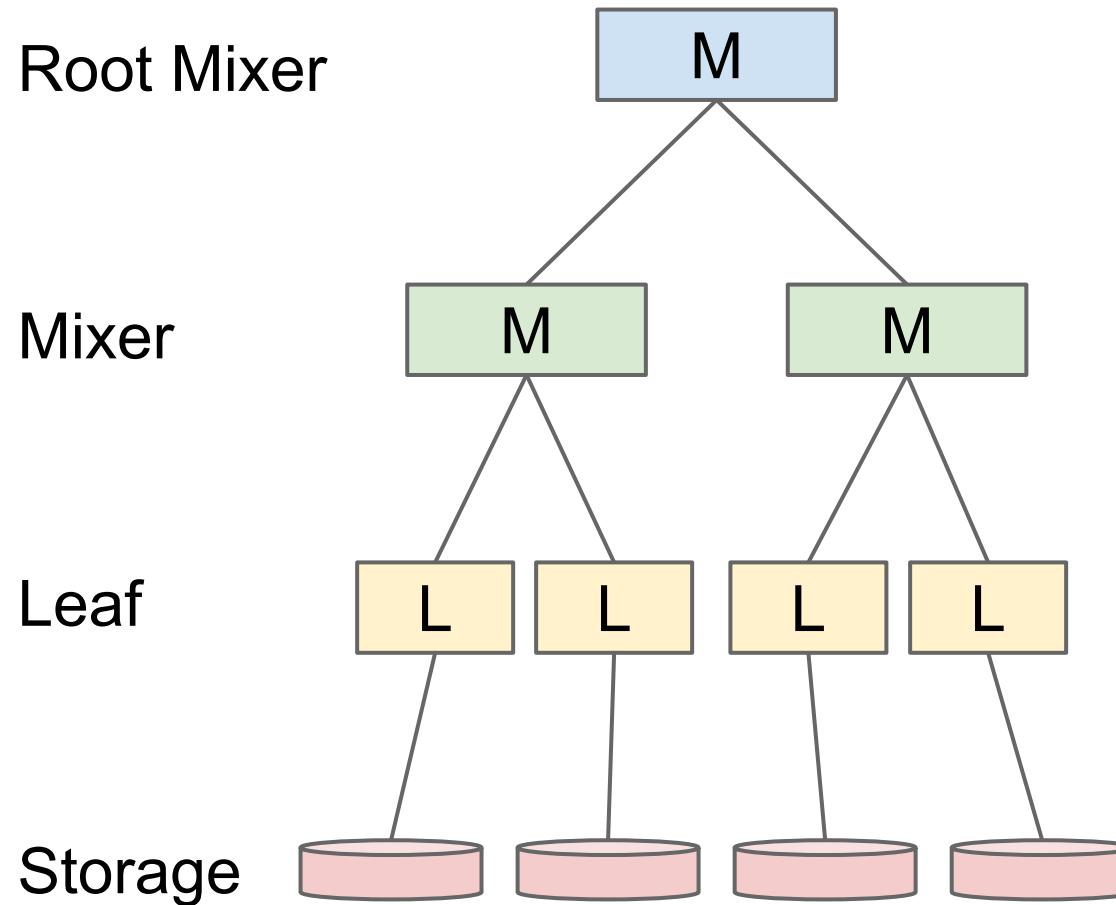
# Life of a BigQuery



Query

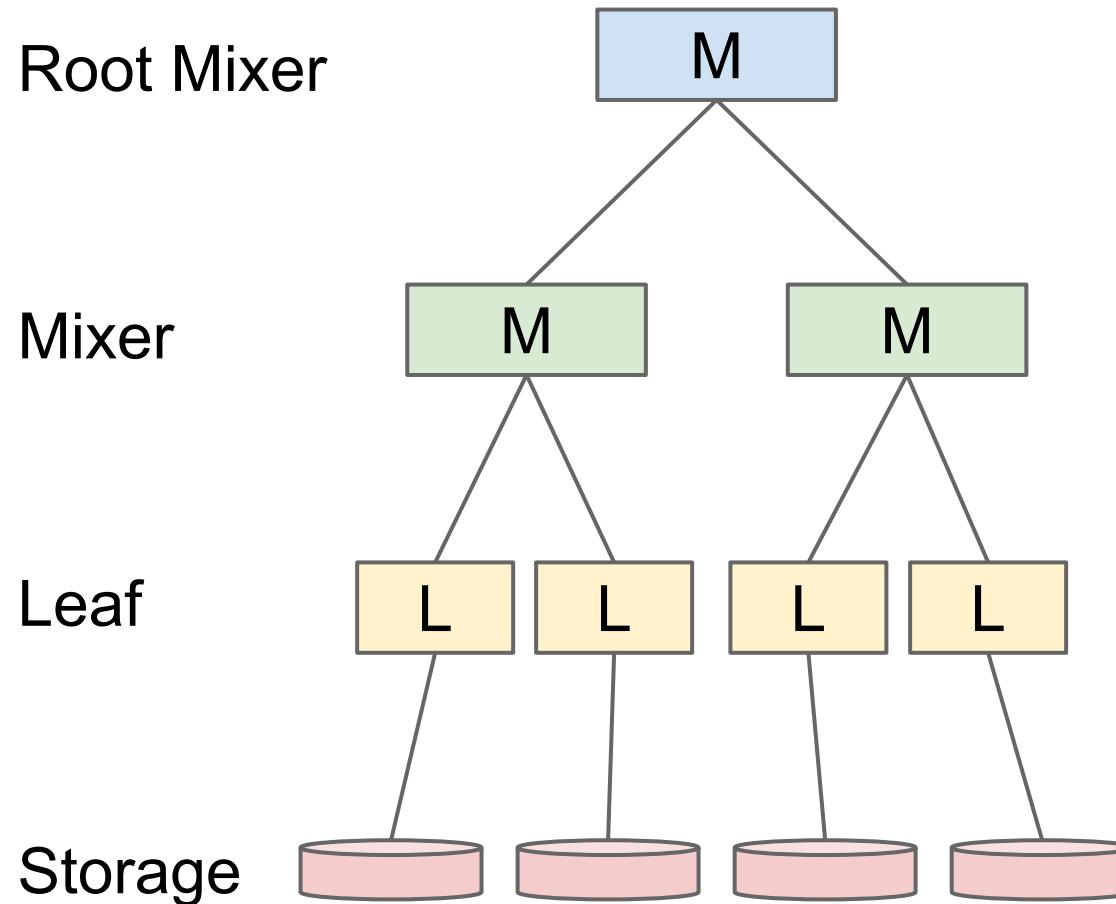


# Life of a BigQuery



`SELECT requests, title`

# Life of a BigQuery

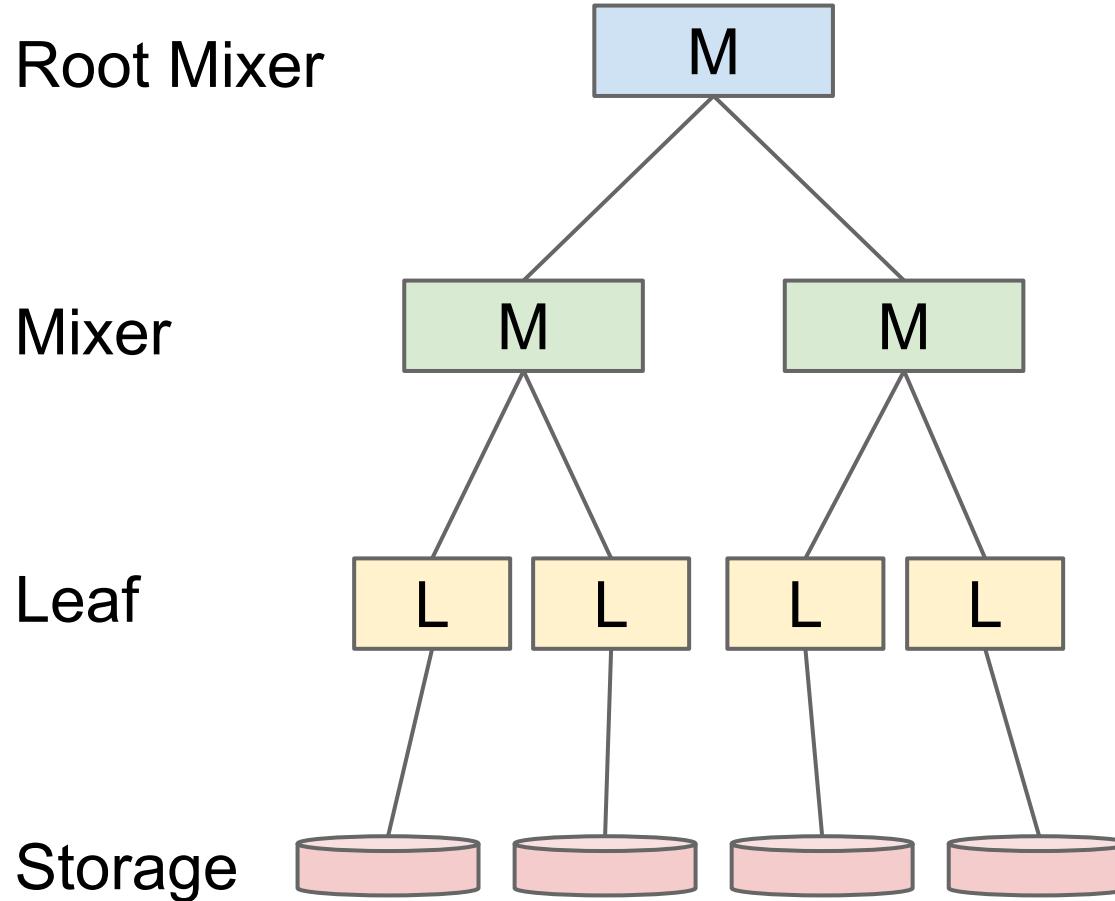


WHERE  
(REGEXP\_MATCH(title, '[Jj]en.+'))

5.4 Bil

SELECT requests, title

# Life of a BigQuery



`SELECT sum(requests)`

`5.8 Mil`

`WHERE`  
`(REGEXP_MATCH(title, '[Jj]en.+'))`

`5.4 Bil`

`SELECT requests, title`

# Life of a BigQuery

Root Mixer

M

Mixer

M

M

Leaf

L

L

L

L

Storage



SELECT sum(requests)



SELECT sum(requests)

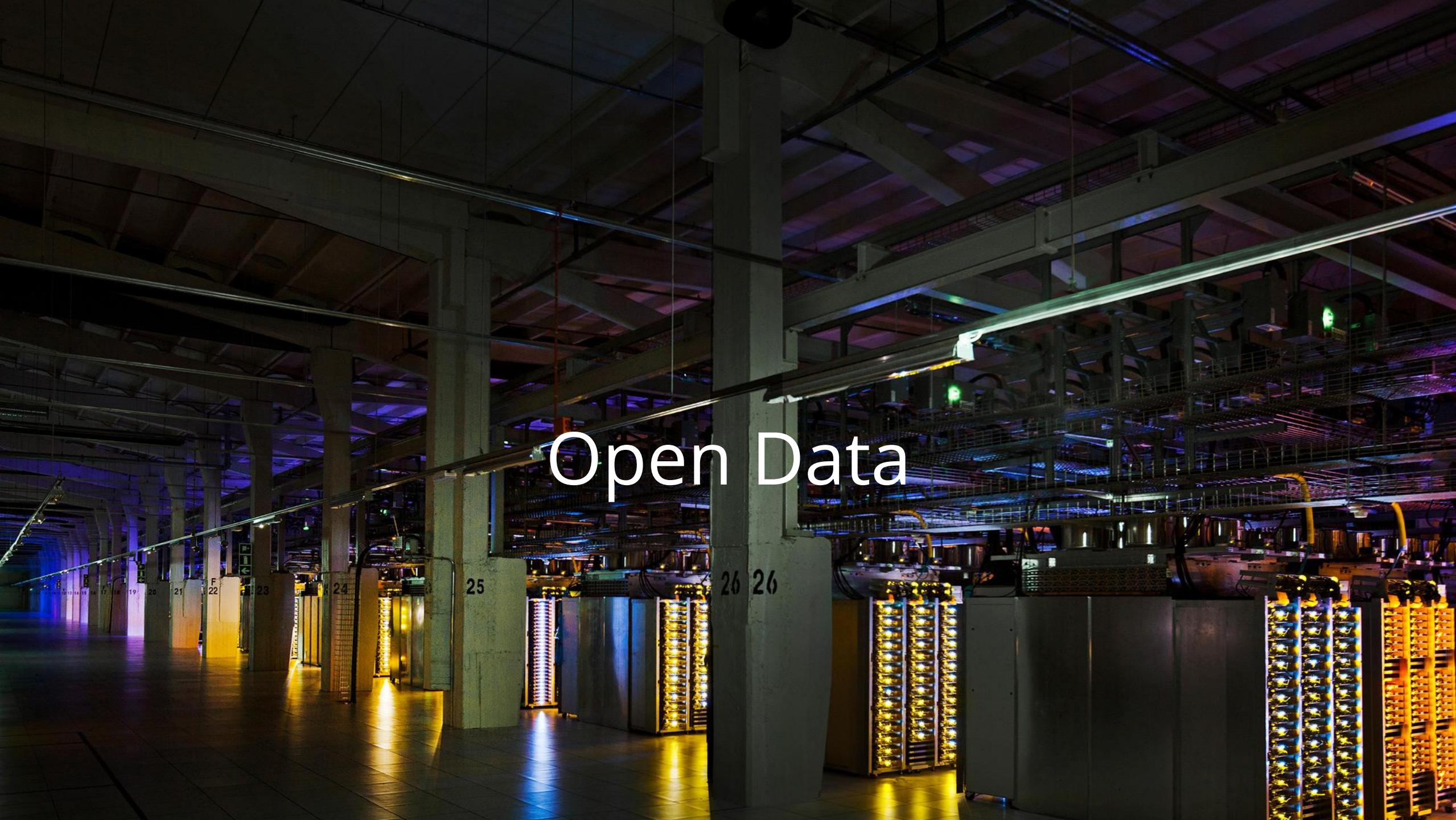


WHERE  
(REGEXP\_MATCH(title, '[Jj]en.+'))



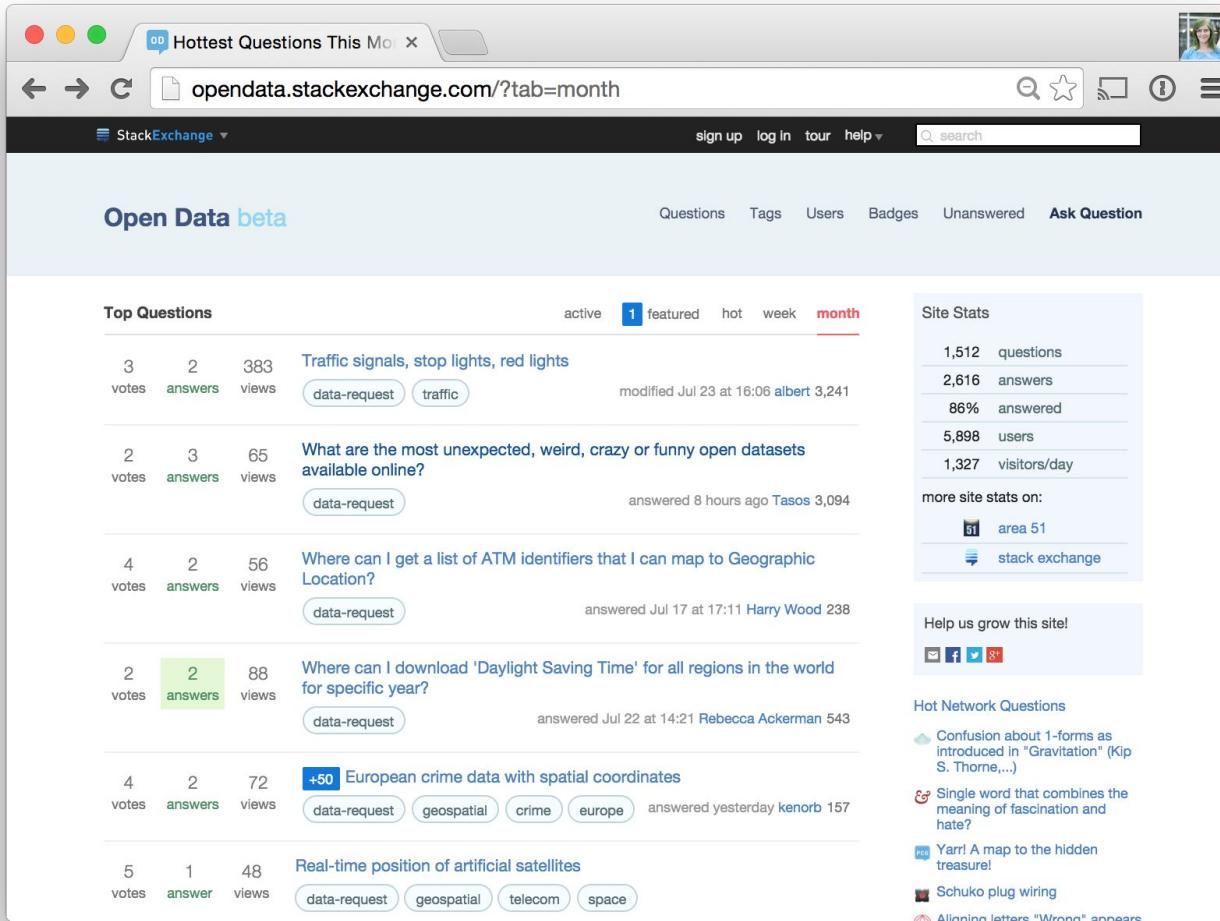
SELECT requests, title





# Open Data

# Finding Open Data



The screenshot shows a web browser window displaying the Open Data beta site on Stack Exchange. The URL in the address bar is [opendata.stackexchange.com/?tab=month](https://opendata.stackexchange.com/?tab=month). The page features a header with navigation links for sign up, log in, tour, help, and search. Below the header, there's a navigation bar with links for Questions, Tags, Users, Badges, Unanswered, and Ask Question.

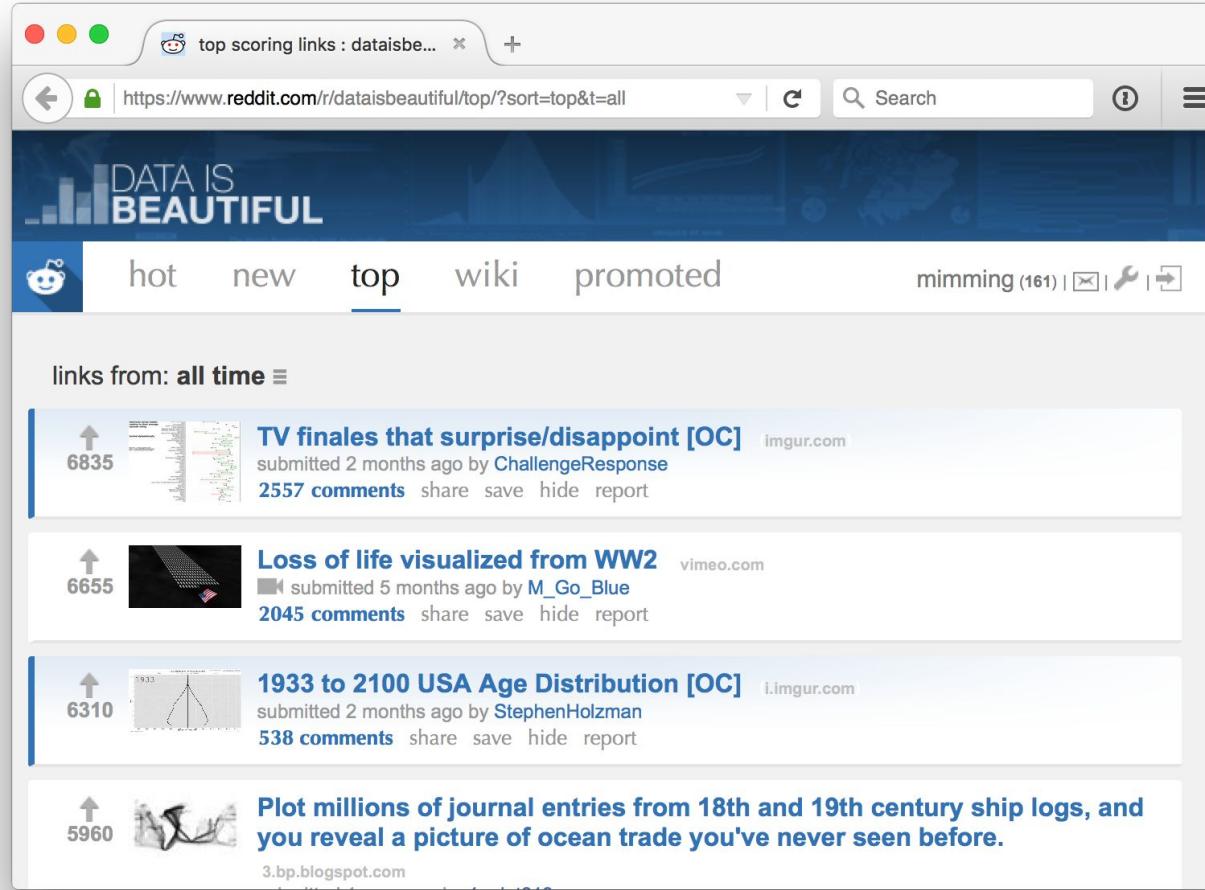
The main content area is titled "Top Questions" and displays a list of six questions sorted by month. Each question card includes the number of votes, answers, and views, along with a brief description and a "data-request" button. The questions are:

- 3 votes, 2 answers, 383 views: Traffic signals, stop lights, red lights (modified Jul 23 at 16:06 by albert 3,241)
- 2 votes, 3 answers, 65 views: What are the most unexpected, weird, crazy or funny open datasets available online? (answered 8 hours ago by Tasos 3,094)
- 4 votes, 2 answers, 56 views: Where can I get a list of ATM identifiers that I can map to Geographic Location? (answered Jul 17 at 17:11 by Harry Wood 238)
- 2 votes, 2 answers, 88 views: Where can I download 'Daylight Saving Time' for all regions in the world for specific year? (answered Jul 22 at 14:21 by Rebecca Ackerman 543)
- 4 votes, 2 answers, 72 views: +50 European crime data with spatial coordinates (answered yesterday by kenorb 157)
- 5 votes, 1 answer, 48 views: Real-time position of artificial satellites (data-request)

On the right side of the page, there's a "Site Stats" section showing various metrics like 1,512 questions and 2,616 answers. Below that is a "more site stats on:" section with links to Area 51 and Stack Exchange. There's also a "Help us grow this site!" section with social media sharing icons. At the bottom, there's a "Hot Network Questions" section with links to other related questions.

[opendata.stackexchange.com](https://opendata.stackexchange.com)

# Finding Open Data



[reddit.com/r/dataisbeautiful](https://www.reddit.com/r/dataisbeautiful)

# Finding Open Data

The screenshot shows a Mac OS X browser window displaying the subreddit <https://www.reddit.com/r/bigquery/wiki/datasets>. The page content is as follows:

**Datasets publicly available on Google BigQuery**

- Post more at <http://www.reddit.com/r/bigquery>
- Find and share interesting queries at <http://bigqueri.es/categories>
- Ask questions at <http://stackoverflow.com/questions/tagged/google-bigquery>
- Get started now (5 minutes, no credit card needed).

**Sample tables**

- Samples described: <https://cloud.google.com/bigquery/docs/sample-tables>

**GDELT Worldwide news and events (340GB and growing every 15 minutes)**

- GDELT announcement
- GDELT v2 announcement
- Top words queries
- All events: <https://bigquery.cloud.google.com/table/gdelt-bq:full.events>

**GDELT American Television Global Knowledge Graph dataset released: (>28 GB)**

- >740,000 broadcasts codified
- <https://blog.gdeltproject.org/announcing-the-american-television-global-knowledge-graph-dataset-released/>

**Datasets publicly available on Google BigQuery**

- Sample tables
- GDELT Worldwide news and events (340GB and growing every 15 minutes)
- GDELT American Television Global Knowledge Graph dataset released: (>28 GB)
- Worldwide Weather 1929-2015 (23 GB)
- Mexico
- Wikipedia (380GB per month)
- GitHubArchive (87.2 GB per year, and growing every day)
- Genomics (3.4 TB + 9.8 TB + ...)
- Cancer Genomics (>400 GB)
- HttpArchive (42 GB per run)
- Firebase (142 GB)
- New York (130 GB+)
- Eclipse Developer Tools
- Soccer
- Measurement Lab
- Airplanes
- Reddit (546 GB of comments, and growing)
- From Datadives
- GeoIP Geolocation
- Hacker News (4 GB)
- Austin
- Open Library (35 GB)
- Iowa liquor sales (879MB)
- Deezer music playlists (~1GB)

**bigquery**

**subscribe** 1,309 readers

~5 users here now

- BigQuery documentation
- What is BigQuery [video]
- StackOverflow support
- Interesting queries
- Open datasets
- Related subreddits:
  - /r/dataflow
  - /r/bigdata
  - /r/datamining
  - /r/datasets
  - /r/AppEngine
  - /r/firebase
  - /r/googlecloud
  - Multi-reddit
- Recommended posts:
  - Getting started
  - NYC taxi trips queries
  - Reddit comments queries
  - GDELT video, taxi trips video
  - Top posts

Moderator: [@felipehoffa](#)

created by [fnoffa](#) a community for 2 years

**WIKI TOOLS**

recent wiki revisions

[reddit.com/r/bigquery/wiki/datasets](https://www.reddit.com/r/bigquery/wiki/datasets)

# Time to explore

# GSOD

The screenshot shows a web browser window for the National Centers for Environmental Information (NCEI) at [www.ncdc.noaa.gov](http://www.ncdc.noaa.gov). The page features the NOAA logo and the text "Formerly the National Climatic Data Center (NCDC)... [more about NCEI](#)". The navigation menu includes Home, Climate Information, Data Access, Customer Support, Contact, About, and a Search bar. A sidebar on the left lists ways to assist users, such as searching for data at a location or finding specific datasets. The main content area displays a map titled "June 2015 Regional Climate Impacts and Outlooks" showing precipitation anomalies and climate impacts across the Eastern United States. The map includes labels like "Percent of Normal Precipitation (%)" and "June 1 - August 31, 2012". A small footer at the bottom of the content area shows page numbers 1 through 5.

# Find nearby weather data

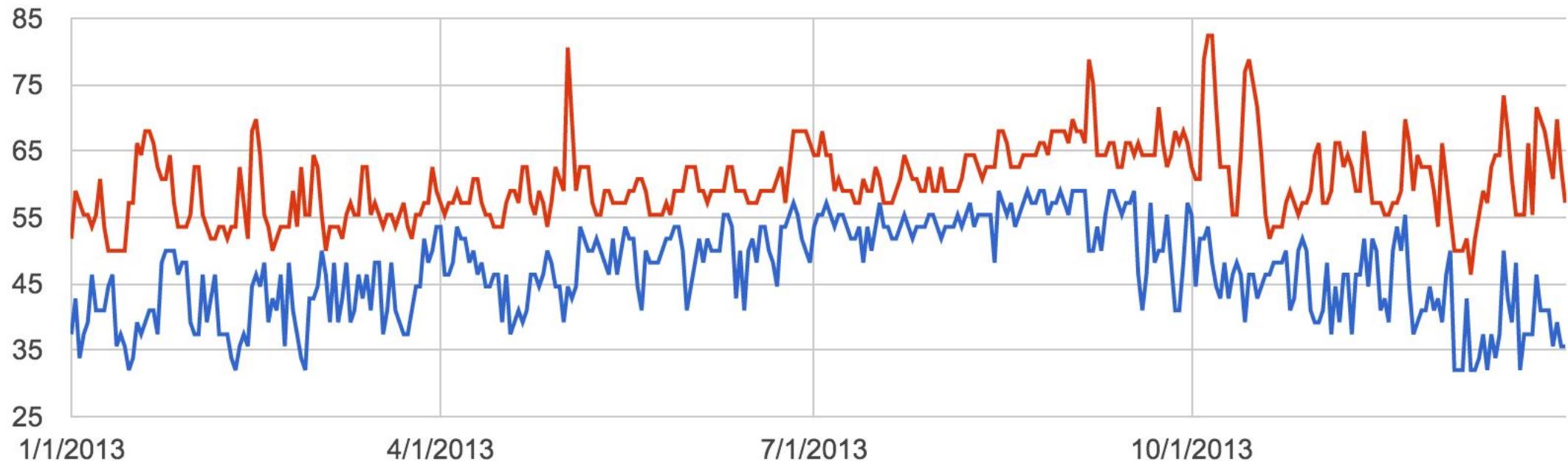
```
select name
from [fh-bigquery:weather_gsod.stations]
where
  state == 'GA' and
  usaf in(select stn from (SELECT count(stn) as cnt, stn
    FROM [fh-bigquery:weather_gsod.gsod2015]
    where stn <> '999999'
    group by stn order by cnt desc))
group by name
order by name ASC;
```

# Weather in Atlanta

```
SELECT DATE(year+mo+da) day, min, max
FROM [fh-bigquery:weather_gsod.gsod2015]
WHERE stn IN (
    SELECT usaf FROM [fh-bigquery:weather_gsod.stations]
    WHERE name = 'ATLANTA MUNICIPAL')
AND max < 200
ORDER BY day;
```

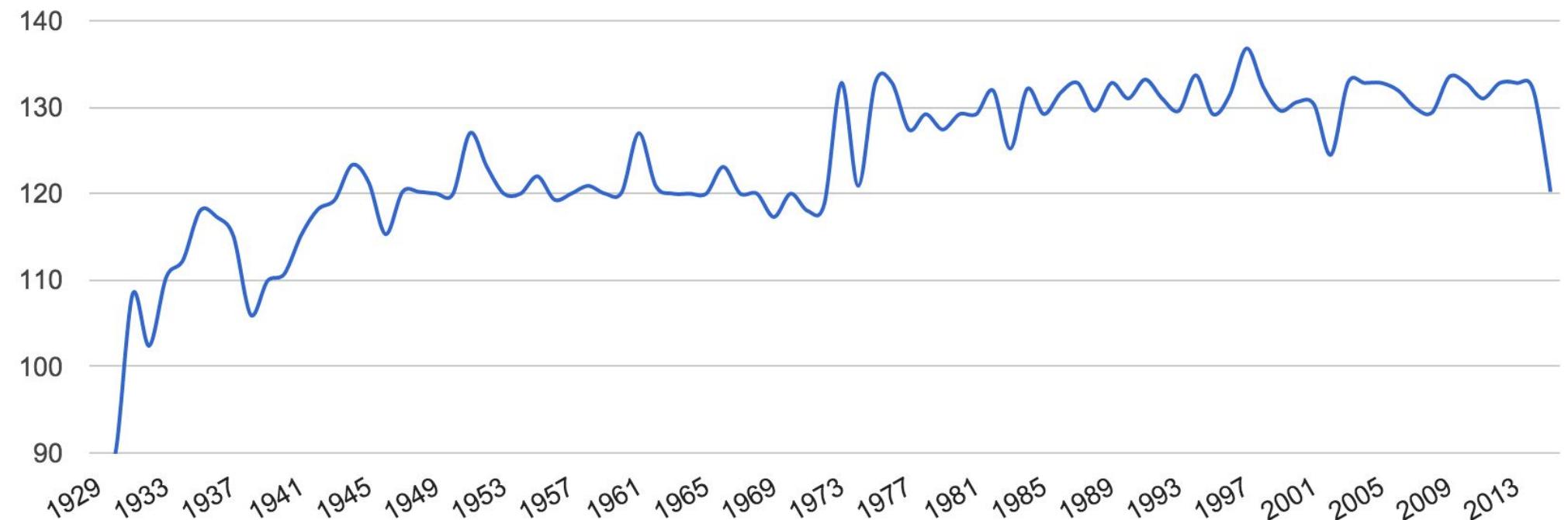
# Weather in Half Moon Bay

```
SELECT DATE(year+mo+da) day, min, max  
FROM [fh-bigquery:weather_gsod.gsod2013]  
WHERE stn IN (  
    SELECT usaf FROM [fh-bigquery:weather_gsod.stations]  
    WHERE name = 'HALF MOON BAY AIRPOR')  
AND max < 200  
ORDER BY day;
```



# Global high temperatures

```
SELECT year, max(max) as max
FROM
  TABLE_QUERY(
    [fh-bigquery:weather_gsod],
    'table_id CONTAINS "gsod"')
where max < 200
group by year order by year asc
```



# GDELT

The screenshot shows a web browser window for "The GDELT Project" at [gdeltpoint.org](http://gdeltpoint.org). The page features a large globe visualization showing global connectivity and event density. A network of colored lines (green, yellow, blue) radiates from various points on the globe, particularly concentrated over Asia and North America. Overlaid on the globe is a semi-transparent white box containing text and a legend. The text inside the box reads:

Nov 02, 2013  
Number of violent attacks   Number of Protests

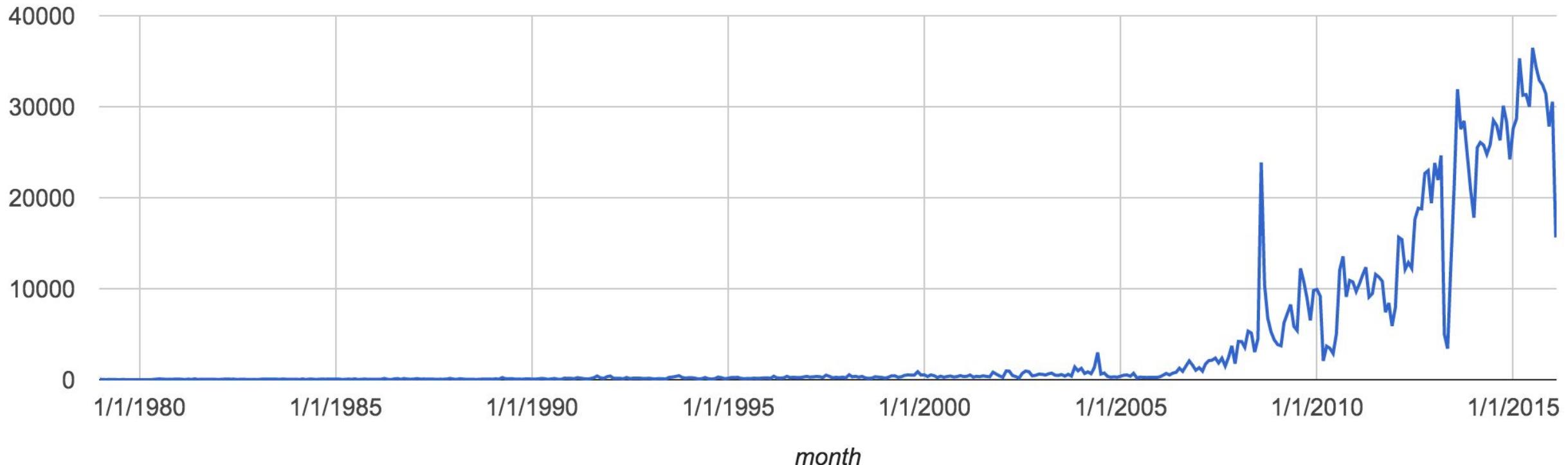
A Global Database of Society  
Supported by Google Ideas,  
the GDELT Project monitors the

Visualization credit GDELT Project.

The main menu bar includes links for Blog, Data, Solutions, and About. The top right corner of the browser window shows a user profile picture and standard browser controls.

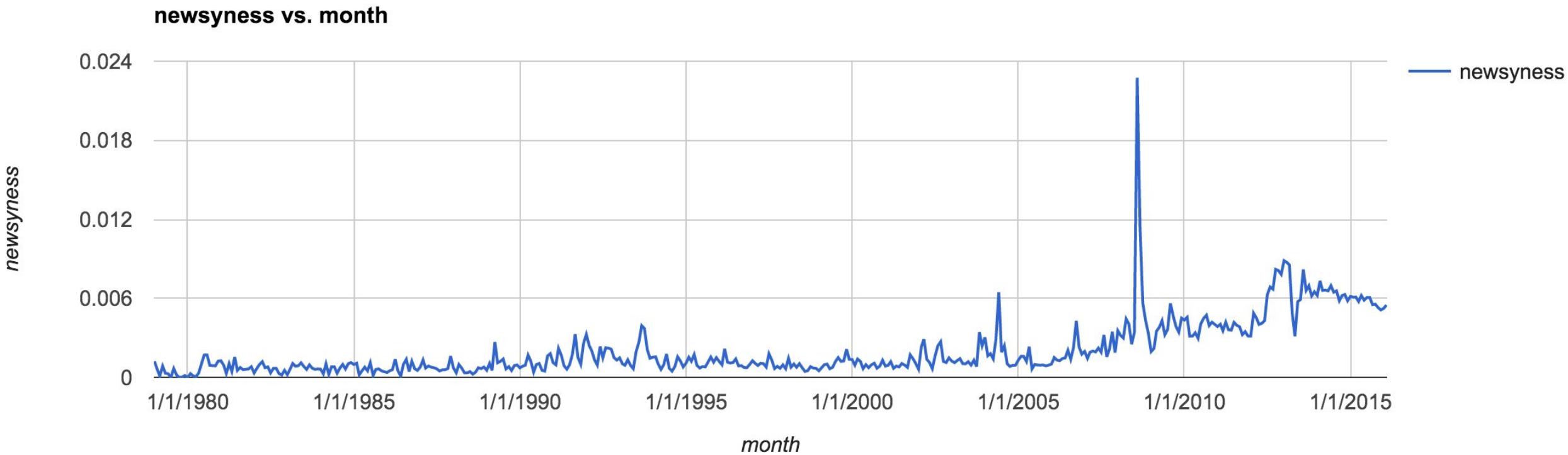
# Stories per month - GA

```
SELECT DATE(STRING(MonthYear) + '01') month,  
       SUM(ActionGeo_ADM1Code='USGA') GA  
  FROM [gdelt-bq:full.events]  
 WHERE MonthYear > 0  
 GROUP BY 1 ORDER BY 1
```



# Stories per month, normalized

```
SELECT DATE(STRING(MonthYear) + '01') month,  
       SUM(ActionGeo_ADM1Code='USGA') / COUNT(*) newsyness  
FROM [gdelt-bq:full.events]  
WHERE MonthYear > 0  
GROUP BY 1 ORDER BY 1
```



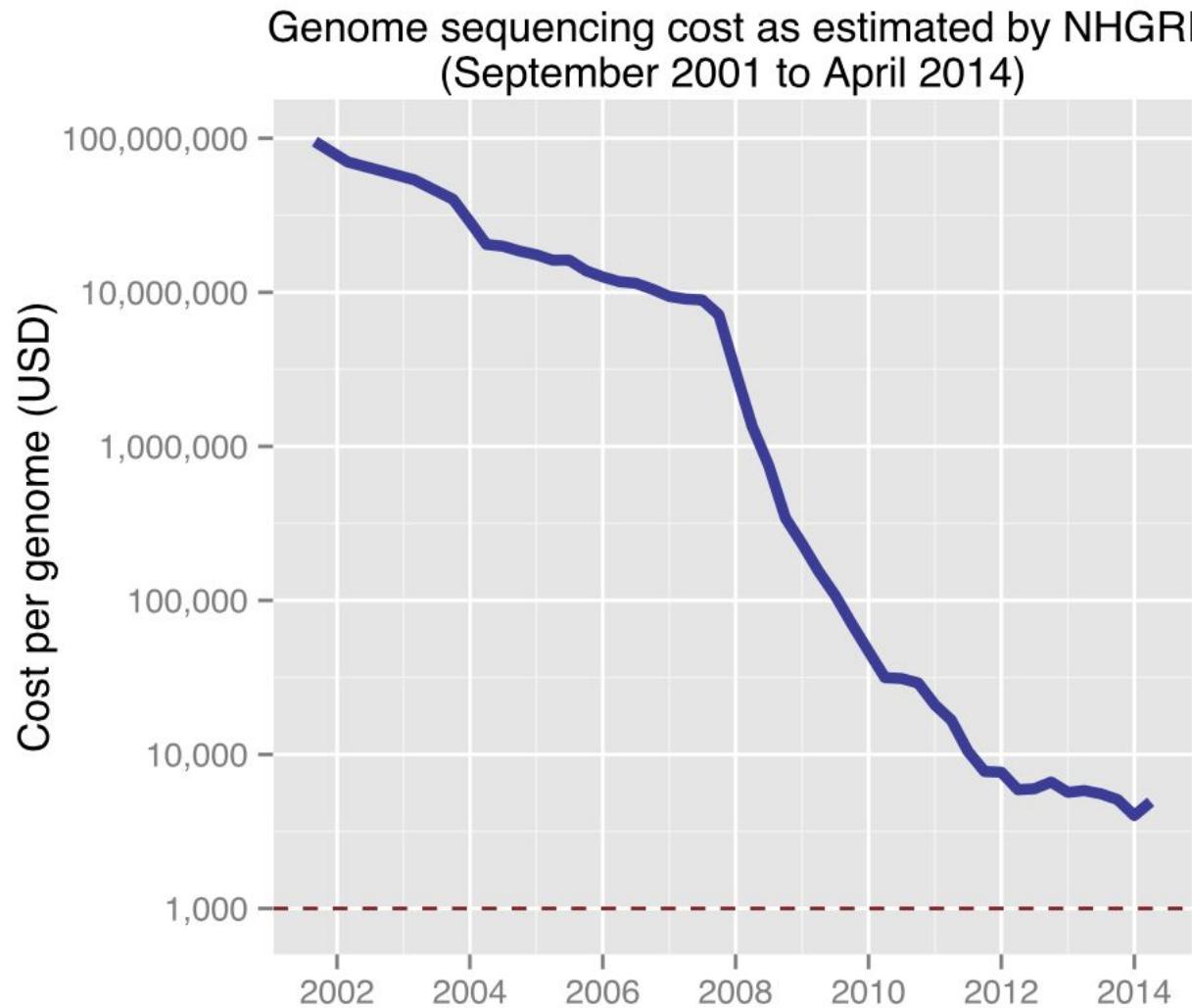
# Genomics



<https://developers.google.com/genomics/>

AAAGAGAGGGCCACCAAGGTTGGCGGCCAACCGGGGTAACCTAGGGGGCTAAAACCCCCAACTTGTAACCCCAAAAAA  
AAAGGAACCCCCTAGCTCTAGAGCTGCCAGCCCTCCGGCGGAAGGGGCCTGGTGGAGCTAGCCAGAGACCCCTGGCCCC  
TTACAGCTAGGGACTTAGCCAGCTCTAGCTAACCTGGGTTTCCCCCAAAAAATTGGTGGACGGTTAACCAAGGGGCCCC  
CGGGGGAAAAATTCTGGGAAGGCCTTTGGACGGCAGTTAACCAAAACCAAAATTAAATTCCGGAATTAACCAACCCAATTAA  
TTTAACCCCCCGCCAACCTCCACAGCCTCTCCGGAAAAGGGGCCGGCAATTCCCCGGCCCCAGGTAACTACCCGGAAG  
GCCAGAATTGGAGTTCTCCACCCCTCCAGGGCCGGCTAGTTAGAAAGAGAGCCACCCCTTTCCAACCCGGTTAAAAAG  
AGCTCCAGCTCTGGCTAGAGAGTTCCCCCCCCGGTTGGTCCCCAAGGTTGGCCTTGGCCCCAATTCCAACCGGGCCTCCGGC  
TCCAGCCTTTCCCCGGCCCAACCCCTTAGAAAGCCCTAAACCCCTTAACCTCTGGCCGGCCCTTCCAACCGGTTGGGAA  
CCAACCAATTAAACCTGGTTGGTTAGGGGAACTCTAGCCGGTGGCTAGAGCCCTGGGCTCCAGAGTTAGAGCTCCGGG  
GGCTAGCTGGTTCTCCCTCTGGGCTCCTGGGGGGAACCTTCCCTCCGGAGAGCCCTAGGGCTCTCCAAAGAAAA  
GGAACCAAAGCCAGAGGGCGGGAGAGCTAAGTCTCTAGGAACCCCAAGGCCTTAAGGGCTGGTCCCCGGCACCCCCCTC  
TGGCCAACGTCTCTTTAACGTAACCTCCCTGGTTGGTTAATTGGGCCAAGGGCTTGGTCCAACCTCTCC  
GGGAACTAGGGGCTCTAGCCCTCTAGCCTGGGAGTTCTACCTGGTCTTTCCACCTGTAGACAGAGCTCTGGAAC  
TAGCCAACCGGCTTAGGAACCACGGAGCCTTGGTAGCTGGAACTTCCCTAGCCTTGGCCCTACGTTGGTTAAAG  
CCAGCTTGGCCCCCCCCCCCCGGGCCCTTGGGGGACCTGTAGGGAGGGAACCTTGGGGGGCCCAAGGCCGG  
AGGTTGGGCCGGTTGGATTGGTTAATTGGTTGGCCGGAACCTTCCAACCGGCCCCAAACGGAAAGAGCCAACCG  
GCCAACCTGGCCACCCAGAAAAACCTGGAGCCCCCAGCTTGTAAAGAAACTCTCTCCCCGGGTTCCACGGTTAACCTGGT  
TTCCGGAGTTAACGTTAACGCCGGAAACCGGAAGGGACTCTGGGGGAGAATTGGAAAATTCCCCAGCTACCTCTAGCT  
TTCCTGGCCCCGGCCAACTAGAAACAGCTCCCTGGTTGGGGGACCTTAAGGCCGGCTTAAGGAGCCGGCT  
CTTTGGGGAGGGAGAAAGCTCCAGGGCTCTCAACCAATTGTAATTAAAAGGGGCCCCAAACGGCCAATTCTTAACGG  
TTAAAGGGACCCCCAGCCAGCTCCCTAGGGTTCCGGGTTCTCCAGAGGGCAGGTTTCCCCAGCCGTCTCCCTACGGCTA  
GCCCAAAACTCTGGTTAACGCCAGGGCTAGGGTTAACCAAGGGCCGGGTTAAAAAATTGGCCCTAGACAGGGTTCTCTAG

# Cost to sequence a genome



# 1000 Genomes

The screenshot shows the homepage of the 1000 Genomes website ([www.1000genomes.org](http://www.1000genomes.org)) displayed in a web browser. The page features a dark header with the project's name in large yellow text and a subtitle "A Deep Catalog of Human Genetic Variation". Below the header is a decorative background image of numerous human chromosomes. A navigation bar contains links for Home, About, Data, Analysis, Participants, Contact, Browser, Wiki, and FTP search, along with a search bar. The main content area is titled "LATEST ANNOUNCEMENTS" and includes a date ("WEDNESDAY SEPTEMBER 30, 2015") and a bold heading "A global reference for human genetic variation". It describes the Phase 3 publication and provides download links for the dataset. To the right, there are sections for "NAVIGATION" (linking to Frequently Asked Questions) and "LINKS" (with icons for megaphone, chromosomes, and document, leading to All Project Announcements, Sample and Project Information, and Media Archive).

## 1000 Genomes

A Deep Catalog of Human Genetic Variation

Home   About   Data   Analysis   Participants   Contact   Browser   Wiki   FTP search

Search

### LATEST ANNOUNCEMENTS

WEDNESDAY SEPTEMBER 30, 2015

**A global reference for human genetic variation**

The Phase 3 publication, [A global reference for human genetic variation](#) and the Phase 3 Structural variation publication, [An integrated map of structural variation in 2,504 human genomes](#) are now available from [Nature](#) alongside a [celebration of 25 years of the Human Genome Project](#)

The variants from the Phase 3 analysis are available in [ftp/release/20130502/](#) and extended information about the SV dataset can be found in [ftp/phase3/integrated\\_sv\\_map/](#).

Both these papers are open access and should be free for everyone to read and download.

If you have any questions about the data these papers are based on or how to access it please email [info@1000genomes.org](mailto:info@1000genomes.org)

<http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/>

### NAVIGATION

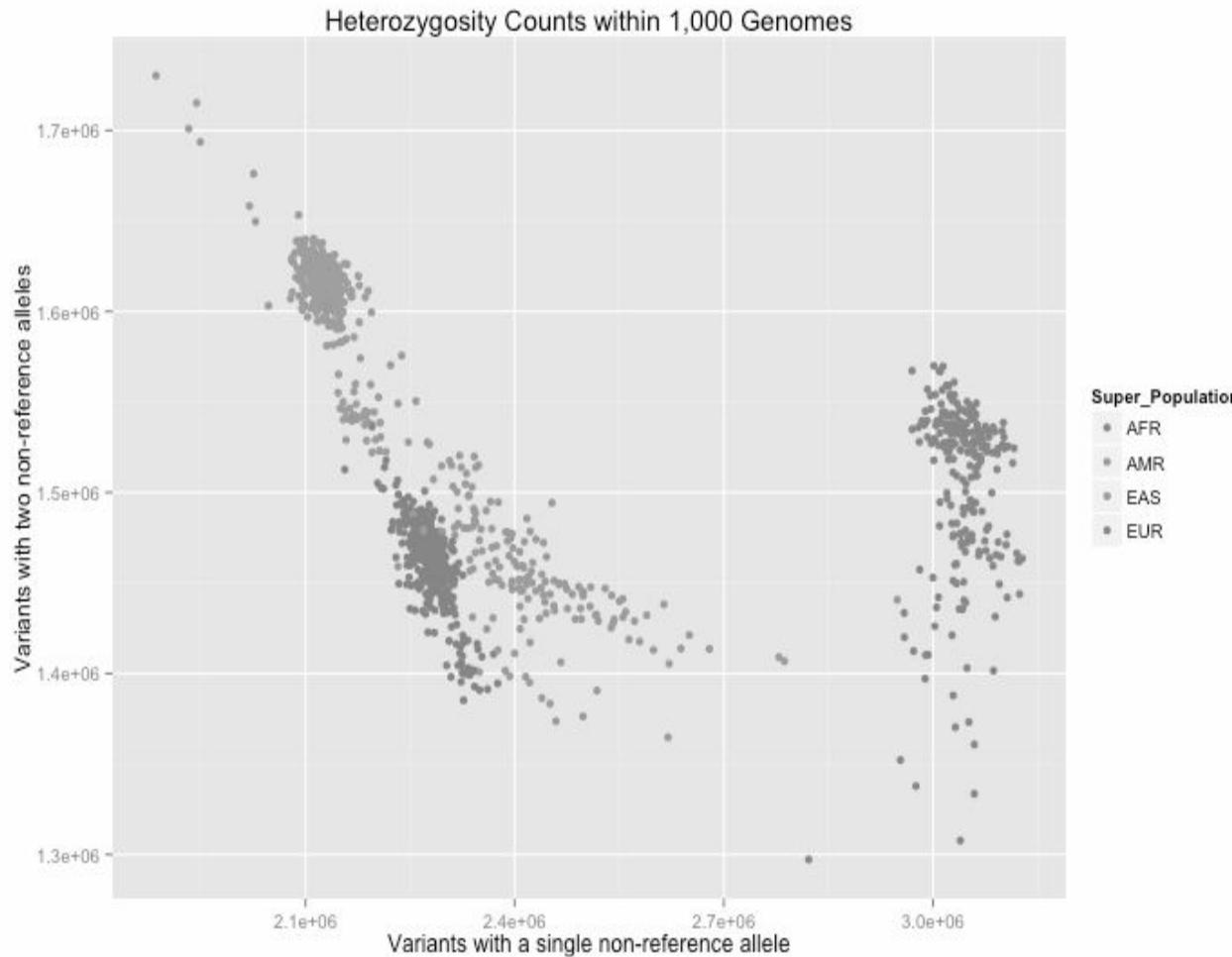
- Frequently Asked Questions

### LINKS

- All Project Announcements
- Sample and Project Information
- Media Archive

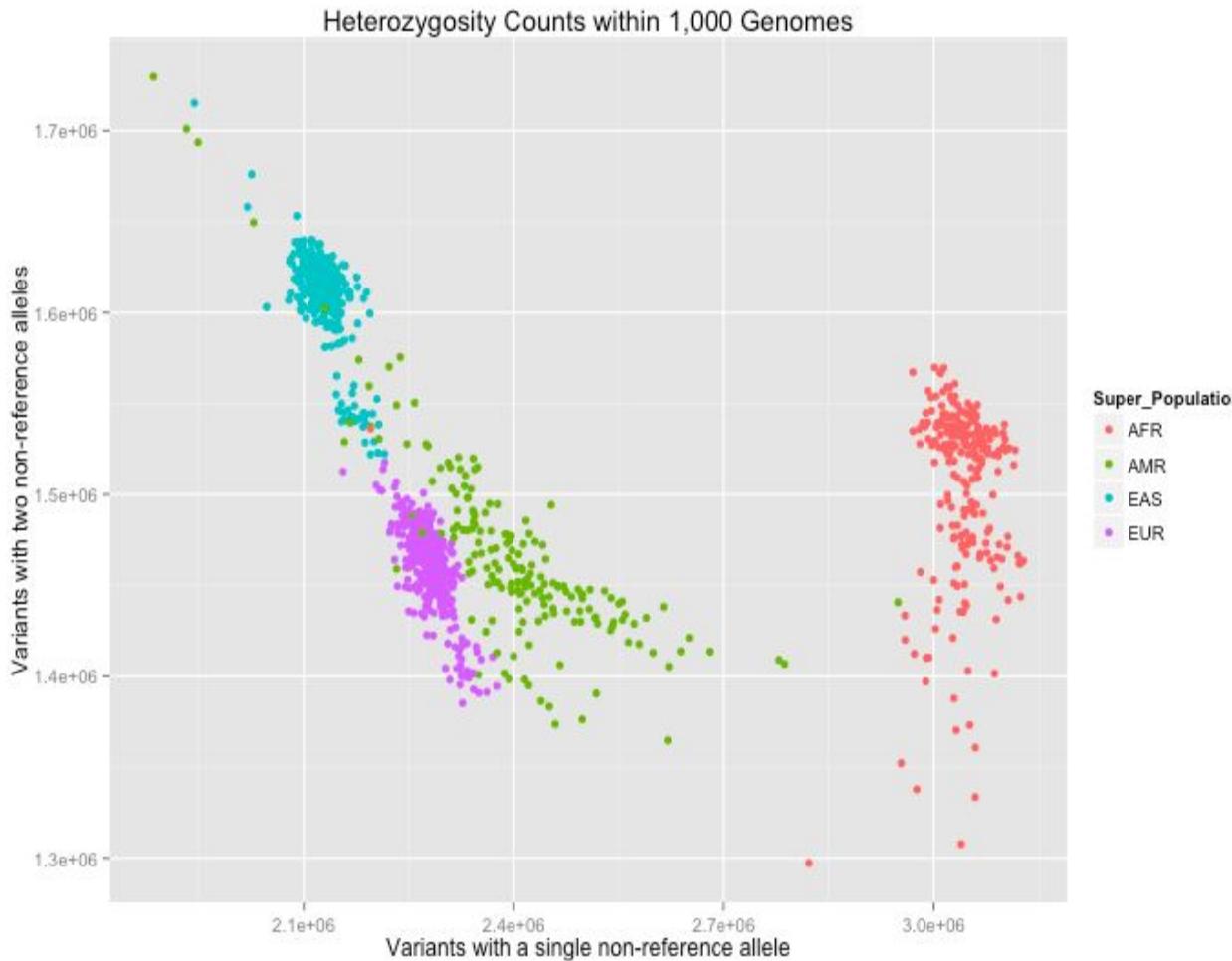
# Genomics

```
SELECT Sample, SUM(single), SUM(double),
FROM (
  SELECT call.call_set_name AS Sample,
    SOME(call.genotype > 0) AND NOT EVERY(call.
genotype > 0) WITHIN call AS single,
    EVERY(call.genotype > 0) WITHIN call AS double,
  FROM [genomics-public-data:1000_genomes.variants]
  OMIT RECORD IF reference_name IN ("X","Y","MT"))
GROUP BY Sample ORDER BY Sample
```



# Genomics

```
SELECT Sample, SUM(single), SUM(double),
FROM (
  SELECT call.call_set_name AS Sample,
    SOME(call.genotype > 0) AND NOT EVERY(call.
genotype > 0) WITHIN call AS single,
    EVERY(call.genotype > 0) WITHIN call AS double,
  FROM [genomics-public-data:1000_genomes.variants]
  OMIT RECORD IF reference_name IN ("X","Y","MT"))
GROUP BY Sample ORDER BY Sample
```



# Something useful:

Use Wikipedia data to pick a movie

1. Wikipedia edits
2. ???
3. Movie recommendation

# Follow the edits

Wikipedia page for **Hackers (film)**. The page summary states: "Hackers is a 1995 American teen science fiction thriller crime film directed by Iain Softley and starring Jonny Lee Miller, Angelina Jolie, Renoly Santiago, Matthew Lillard, Jay Winters, Lorraine Bracco and Fisher Stevens. The film follows the exploits of a group of gifted high school hackers and their involvement in a corporate extortion conspiracy. Made in the 1990s when the Internet was unfamiliar to the general public, it reflects the ideals laid out in the Hacker Manifesto quoted in the film, "This our world now... the world of the electron and the switch [...] We exist without skin color, without nationality, without religious bias... and you call us criminals. [...] Yes, I am a criminal. My crime is that of curiosity." Hackers has achieved cult classic status.<sup>[4]</sup>" Below the summary is a poster for the movie.



Same  
editor

Wikipedia page for **Transformers (film)**. The page summary states: "This article is about the 2007 live action film. For the 1986 animated film, see [The Transformers: The Movie](#)". Below the summary is a poster for the movie.

Wikipedia page for **WarGames**. The page summary states: "This article is about the 1983 film. For the 2001 film, see [War Game \(film\)](#). For other uses, see [War Game \(disambiguation\)](#)". Below the summary is a poster for the movie.

# Pick a great movie

```
select title, id, count(id) as edits
from [publicdata:samples.wikipedia]
where
    title contains 'Hackers'
    and title contains '(film)'
    and wp_namespace = 0
group by title, id
order by edits
limit 10
```

# Find edits in common

```
select title, id, count(id) as edits
from [publicdata:samples.wikipedia]
where contributor_id in (
  select contributor_id
  from [publicdata:samples.wikipedia]
  where
    id=264176
    and contributor_id is not null
    and is_bot is null
    and wp_namespace = 0
    and title CONTAINS '(film)'
  group by contributor_id)
  and wp_namespace = 0
  and id != 264176
  and title CONTAINS '(film)'
group each by title, id
order by edits desc
limit 100
```

# Discover the most broadly popular films

```
select id from (
  select id, count(id) as edits
  from [publicdata:samples.wikipedia]
  where
    wp_namespace = 0
    and title CONTAINS '(film)'
  group each by id
  order by edits desc
  limit 20)
```

# Edits in common, minus broadly popular

```
select title, id, count(id) as edits
from [publicdata:samples.wikipedia]
where contributor_id in (
    select contributor_id
    from [publicdata:samples.wikipedia]
    where
        id=264176
        and contributor_id is not null
        and is_bot is null
        and wp_namespace = 0
        and title CONTAINS '(film)'
    group by contributor_id)
and wp_namespace = 0
and id != 264176
and title CONTAINS '(film)'
and id not in (
```

```
select id from (
    select id, count(id) as edits
    from [publicdata:samples.wikipedia]
    where
        wp_namespace = 0
        and title CONTAINS '(film)'
    group each by id
    order by edits desc
    limit 20
)
group each by title, id
order by edits desc
limit 100
```

# What we talked about

- Origin story
- Count stuff
- How it works
- Some cool open data
- Practical applications



# The end

- Try BigQuery
  - [bigquery.cloud.google.com](https://bigquery.cloud.google.com)
- Me: [@MimmingCodes](https://twitter.com/MimmingCodes)
- Slides: [mimming.com/presos/](https://mimming.com/presos/)
- Queries: [github.com/mimming/snippets](https://github.com/mimming/snippets)





*ISB teams with Google for  
NCI Cancer Genomics Cloud project*

Research Blog: Facilitating

← → C googleresearch.blogspot.com/2014/07/facilitating-genomics-research-with.html

Google Research Blog

The latest news from Research at Google

## Facilitating Genomics Research with Google Cloud Platform

Posted: Wednesday, July 30, 2014

g+1 249

Twitter Facebook

Posted by Paul C. Boutros, Ontario Institute for Cancer Research, Josh Stuart, UC Santa Cruz, Adam Margolin, Oregon Health & Science University; Nicole Deflaux and Jonathan Bingham, Google Cloud Platform and Google Genomics

The understanding of the origin and progression of cancer remains in its infancy. However, due to rapid advances in the ability to accurately read and identify (i.e. sequence) the DNA of cancerous cells, the knowledge in this field is growing rapidly. Several [comprehensive sequencing studies](#) have shown that alterations of single base pairs within the DNA, known as [Single Nucleotide Variants](#) (SNVs), or duplications, deletions and rearrangements of larger segments of the genome, known as [Structural Variations](#) (SVs), are the [primary causes of cancer](#) and can

Defn  $T_n(C_{ij}) = \sum_{e \in E} C_{ij} \delta_{e \in P(e)}$   
 $\sum_{e \in E} D_e \log_2 \frac{1}{D_e} < \sum_{e \in E} nF(e)$  before hashing w/  
 $\hat{R}_n(F) = \frac{1}{n} \sum_{e \in E} \left[ \sum_{i=1}^n \delta_{e \in P_i(e)} \right] S$  if the  
 $S = \int_{-\infty}^{\infty} p(x) dx$  in the continuous case. In the  
(long tail queries)  
 $R_n(F) = \frac{1}{n} \sum_{e \in E} \left[ \sum_{i=1}^n \delta_{e \in P_i(e)} \right] S$  for the  
binning approach.

Research at Google

google.com/+ResearchatGoogle