Exploring Open Data with BigQuery

Jen Tong

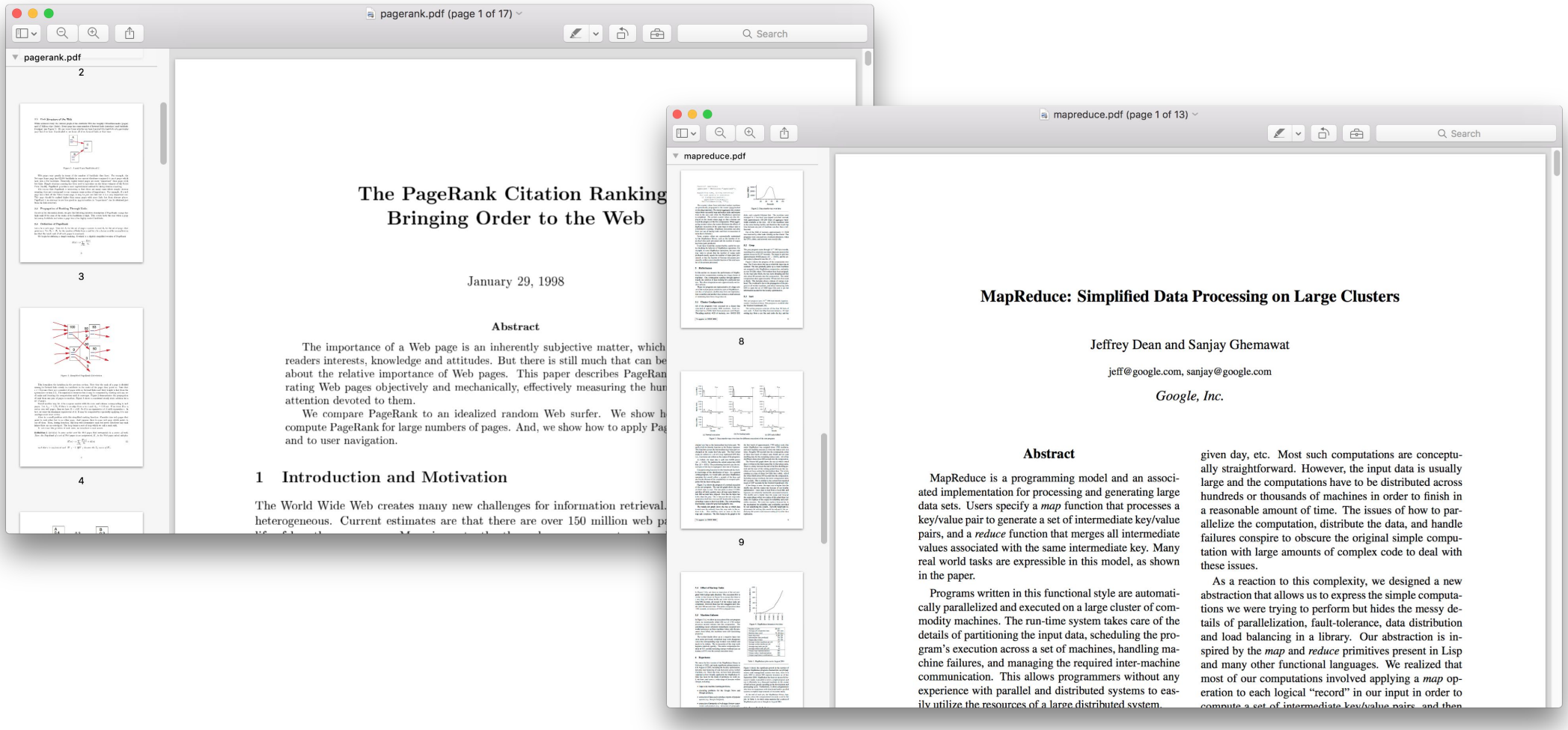
Developer Advocate
Google Cloud Platform

@MimmingCodes
little418.com

Agenda

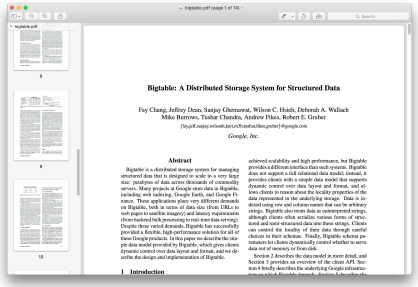
- Origin story
- Count stuff
- How it works
- Some cool open data
- Do something useful

Google Research Publications

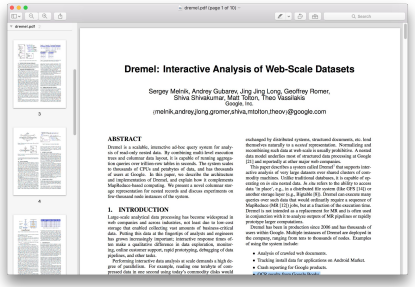


Google Research Publications

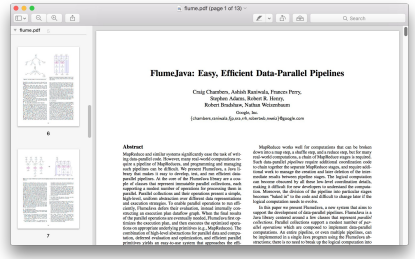
Open Source Implementations


Bigtable: A Distributed Storage System for Structured Data

Bigtable → APACHE HBASE

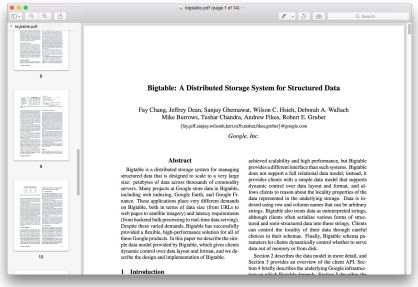

Dremel: Interactive Analysis of Web-Scale Datasets

Flume → CRUNCH

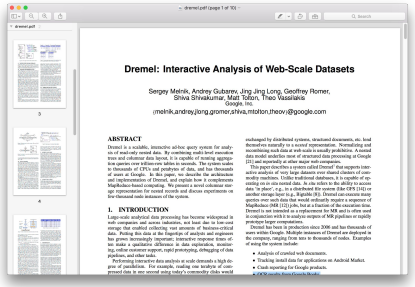

FlumeJava: Easy, Efficient Data-Parallel Pipelines

Dremel → APACHE DRILL

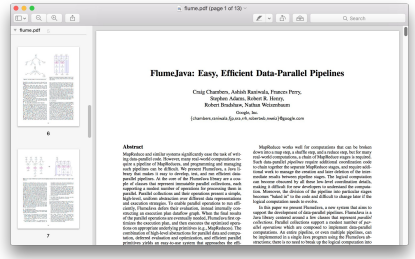
Managed Cloud Versions


Bigtable: A Distributed Storage System for Structured Data

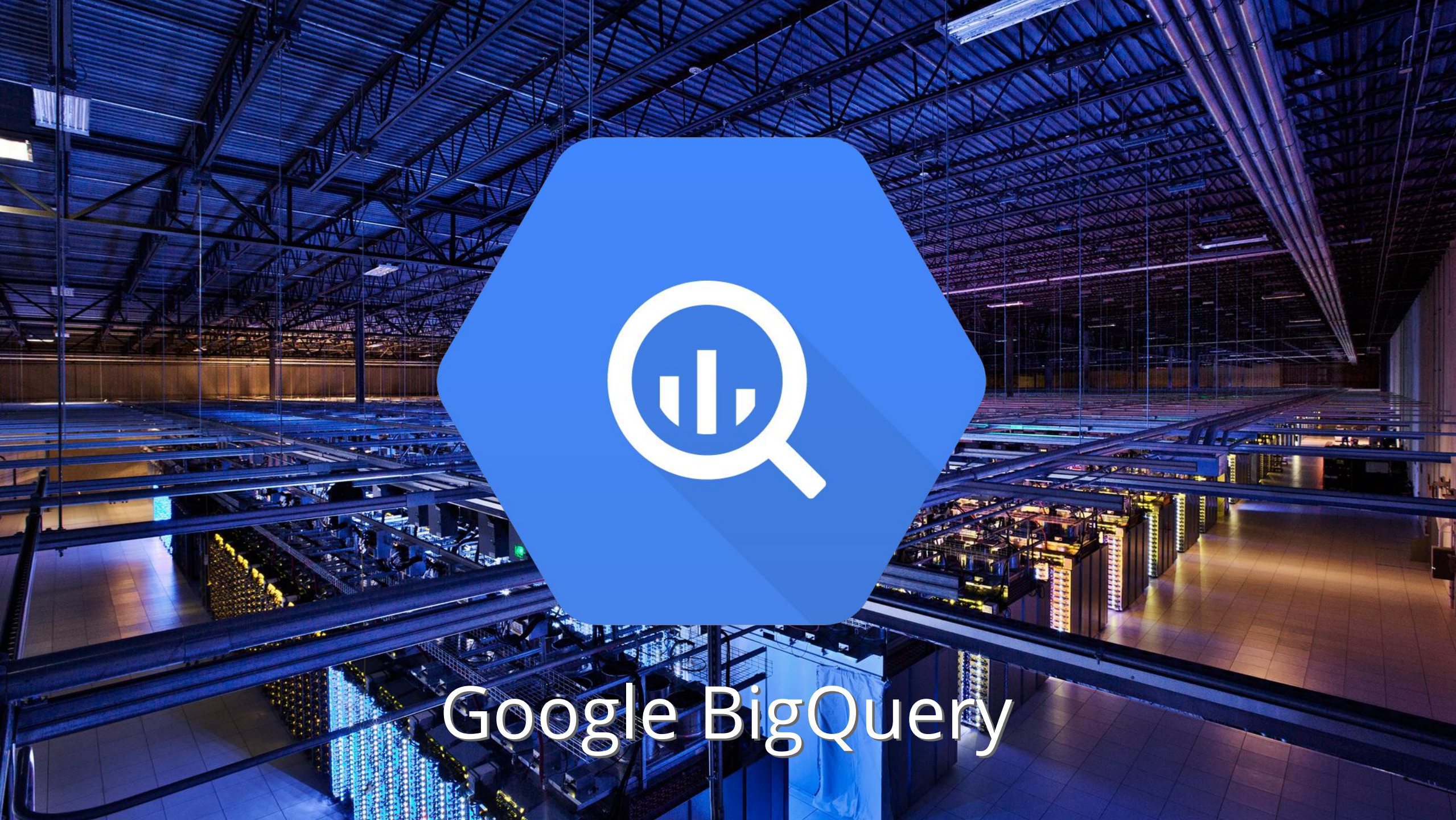
Bigtable → Bigtable


Dremel: Interactive Analysis of Web-Scale Datasets

Flume → Dataflow


FlumeJava: Easy, Efficient Data-Parallel Pipelines

Dremel → BigQuery



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?

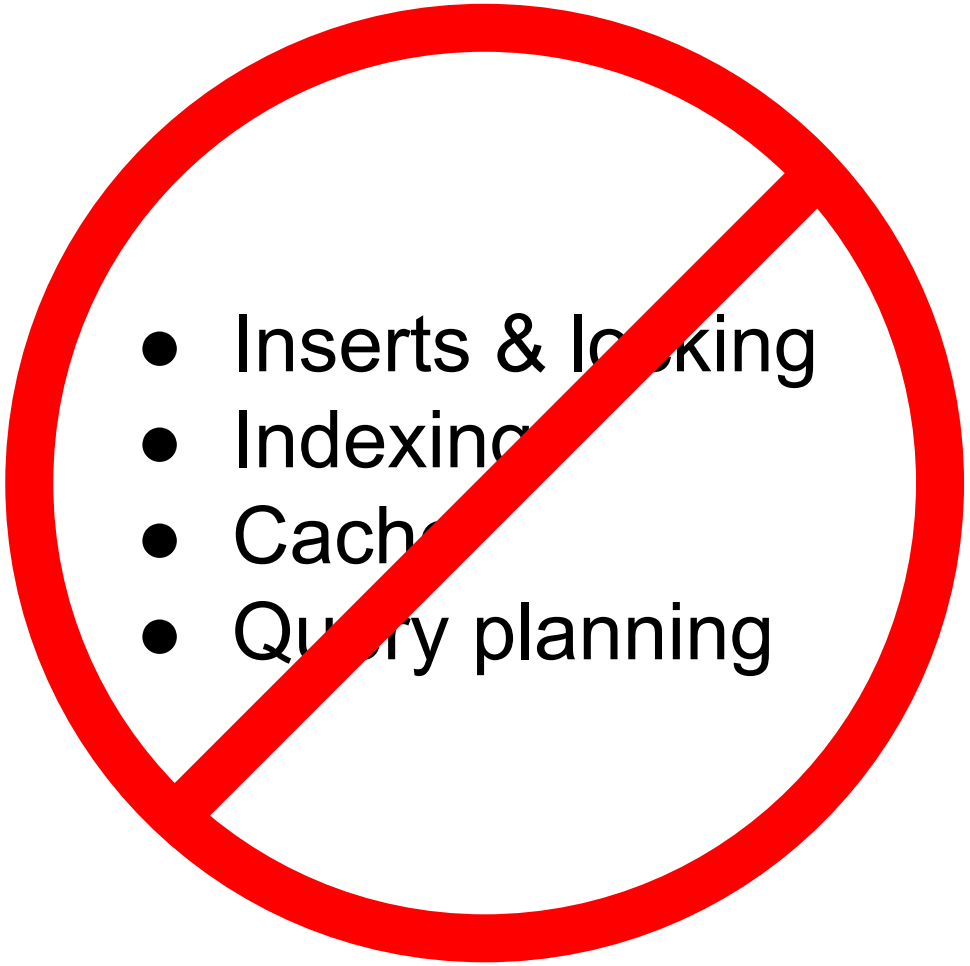
o_O

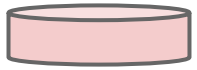
Qualities of a good RDBMS

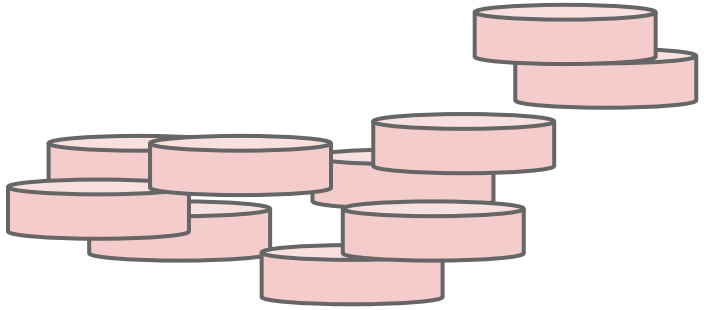
Qualities of a good RDBMS

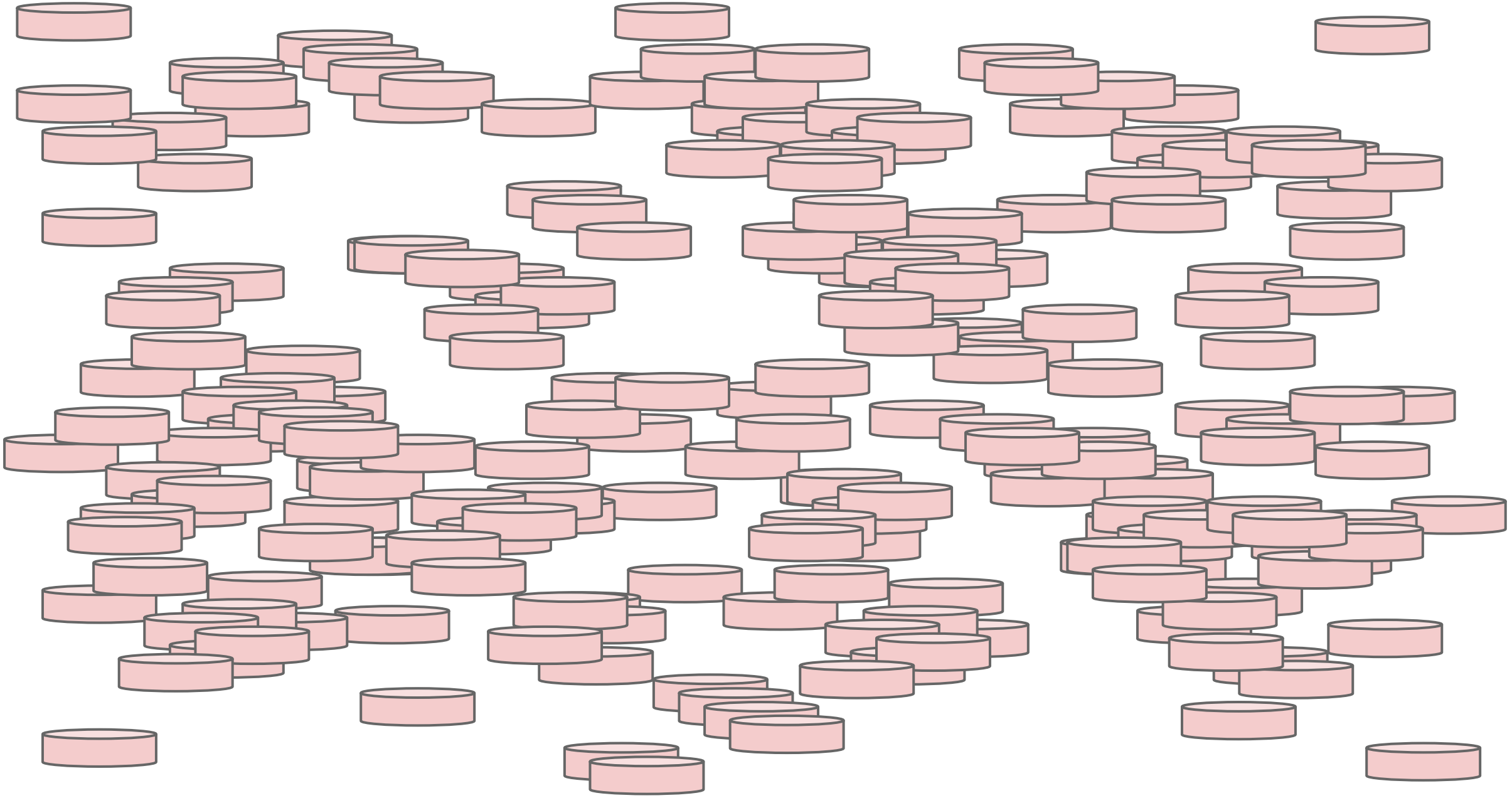
- Inserts & locking
- Indexing
- Cache
- Query planning

Qualities of a good RDBMS

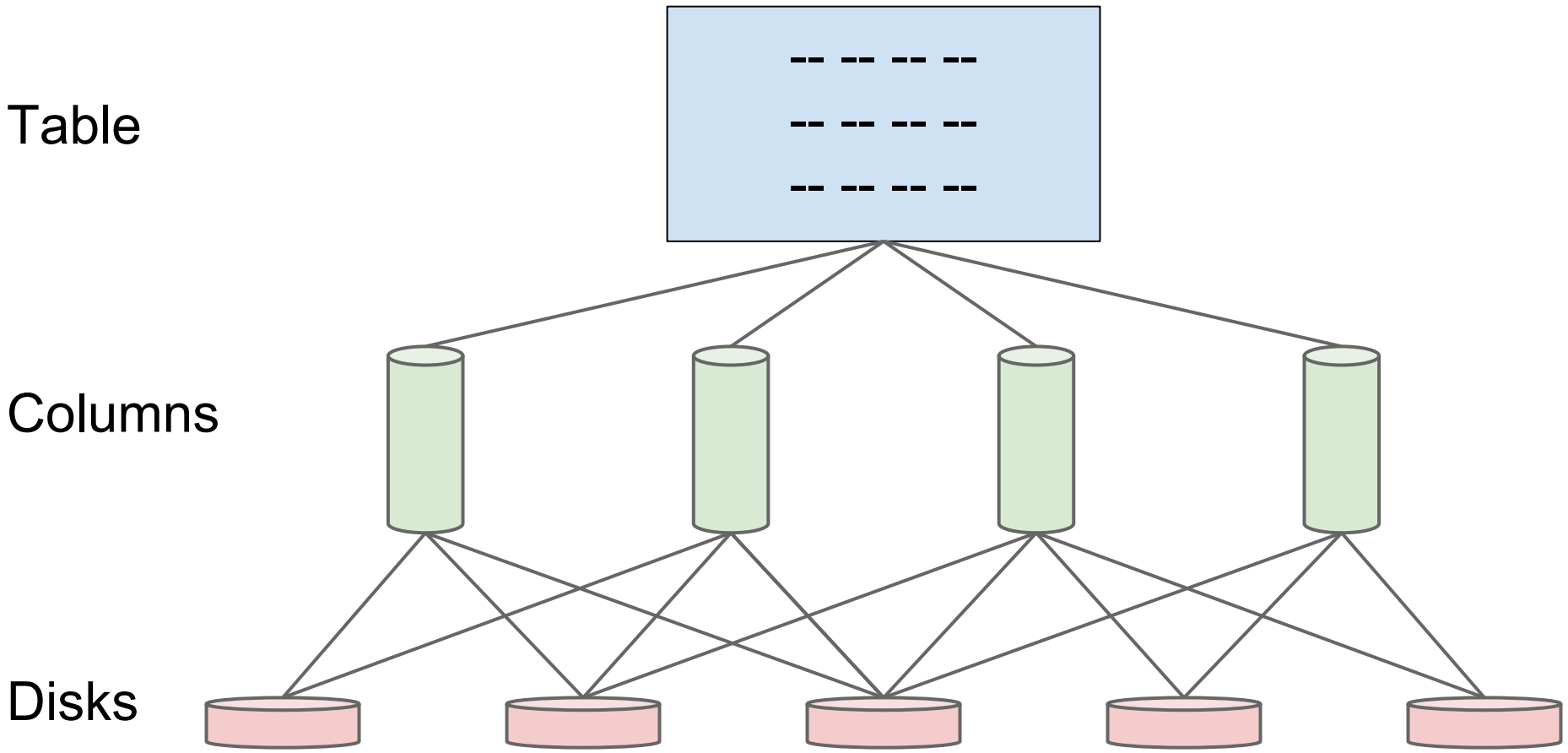








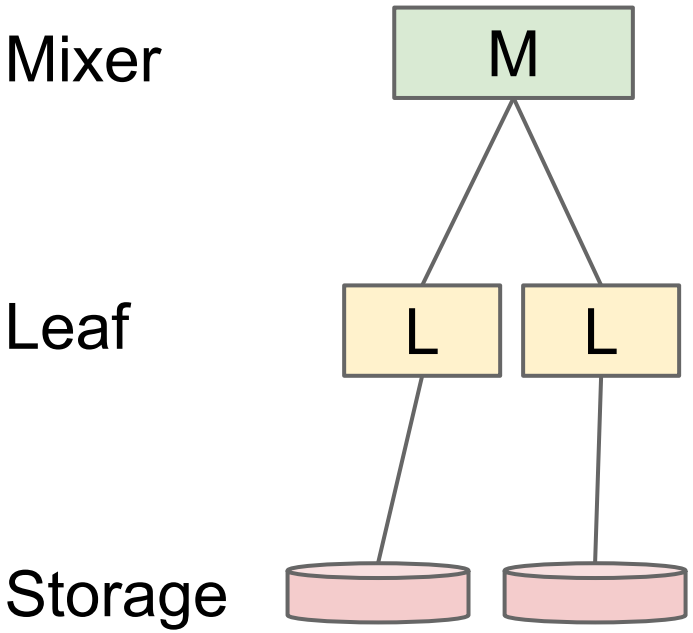
Storing data



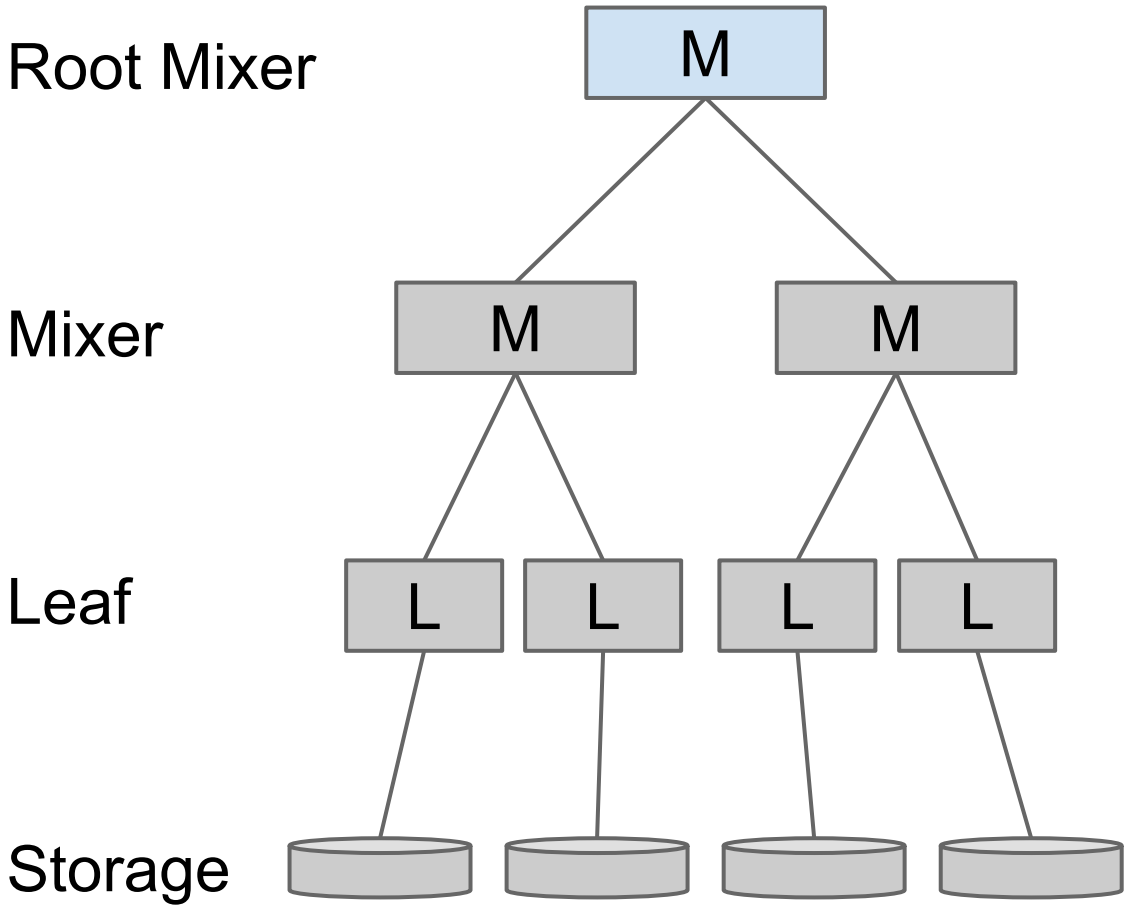
Reading data: Life of a BigQuery

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

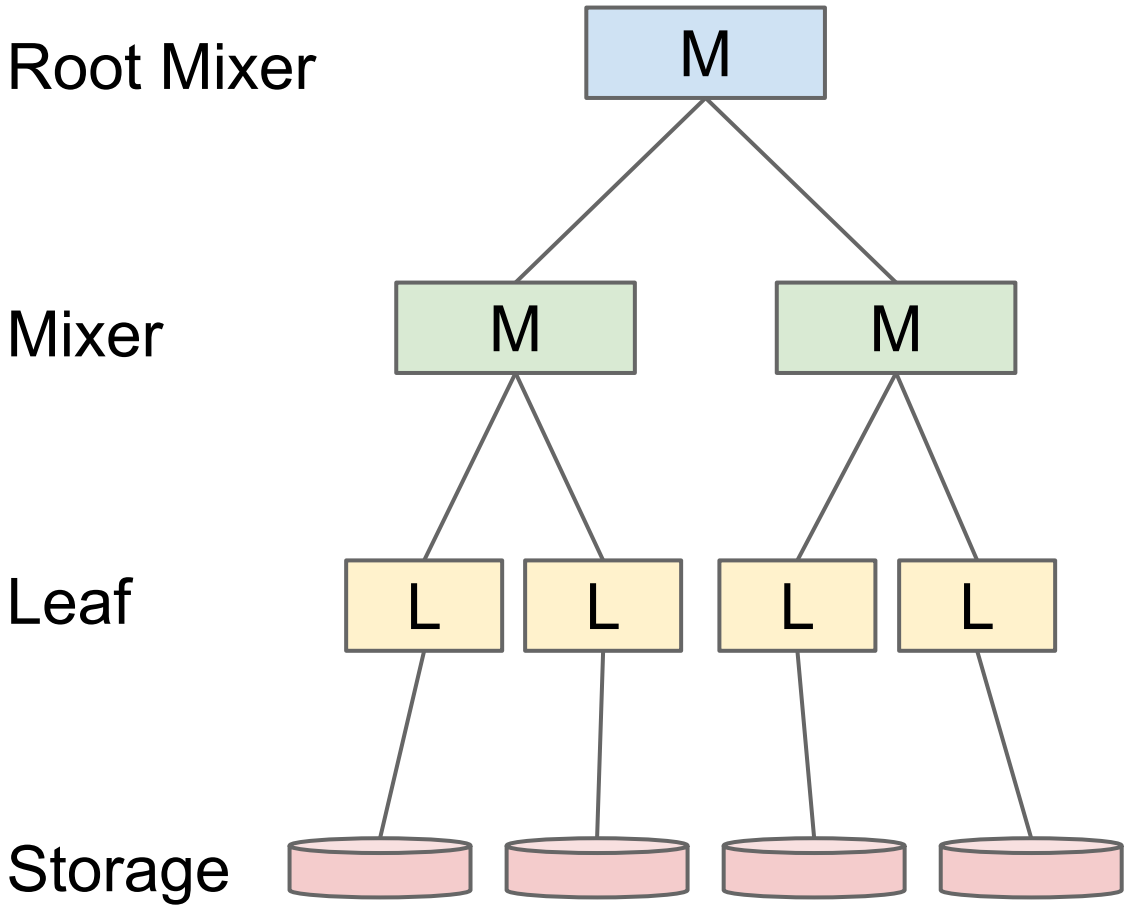
Life of a BigQuery



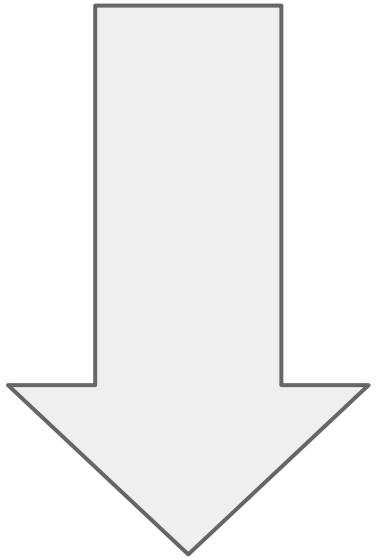
Life of a BigQuery



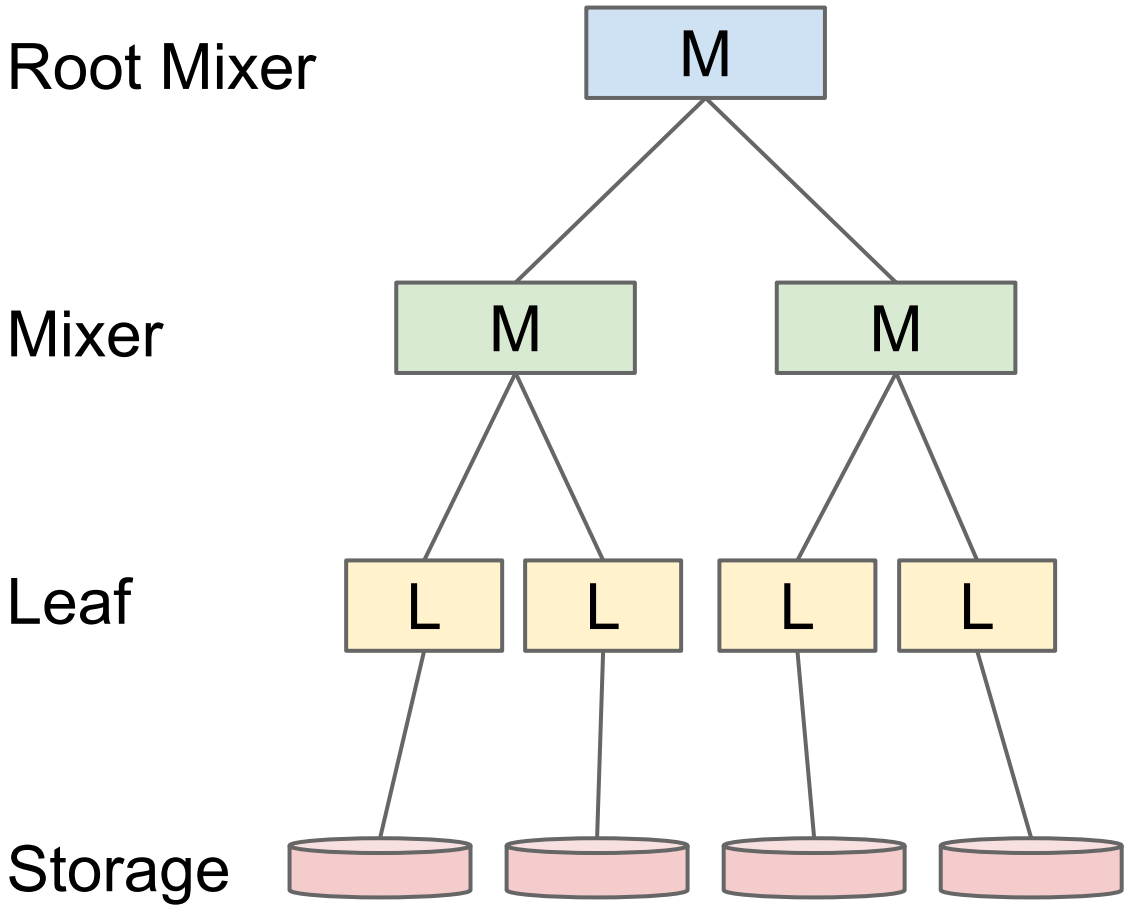
Life of a BigQuery



Query

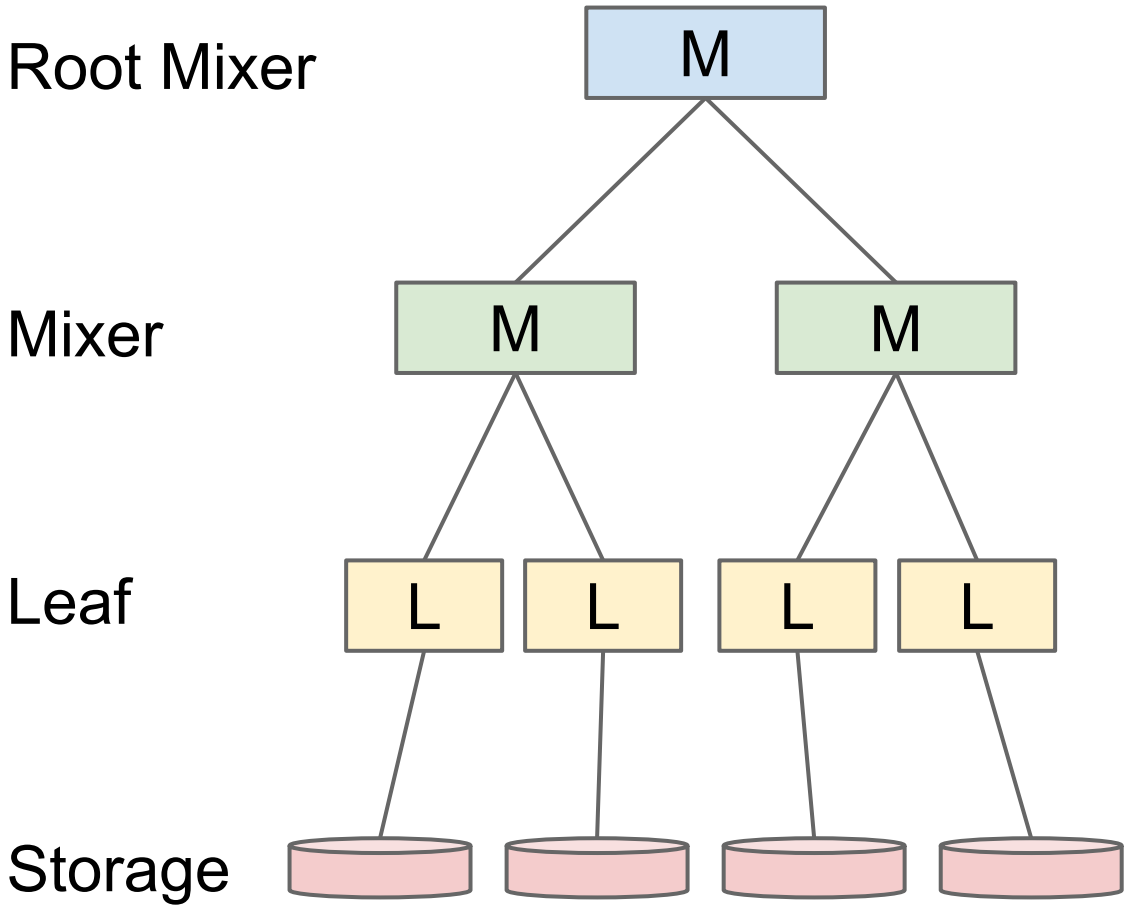


Life of a BigQuery




```
SELECT requests, title
```

Life of a BigQuery

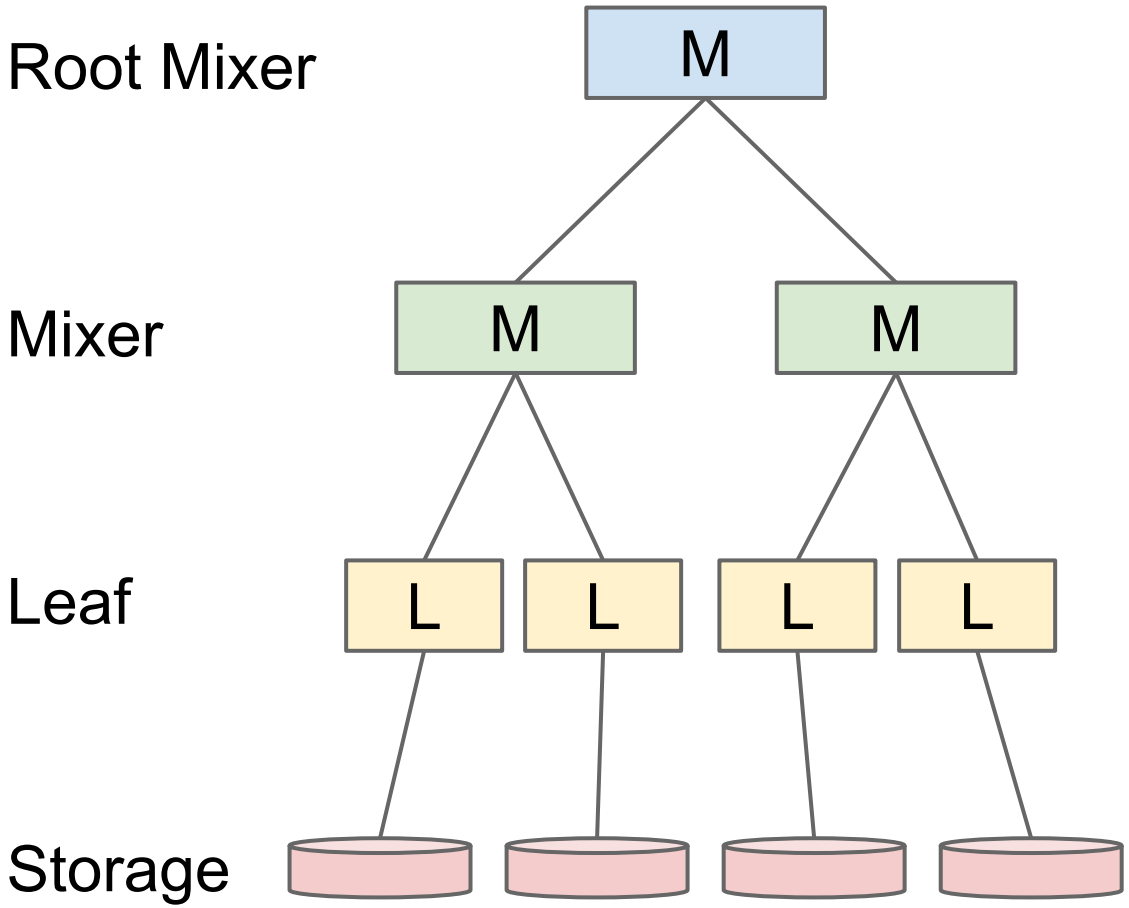



```
WHERE
(REGEXP_MATCH(title, '[Jj]en.+'))
```

5.4 Bil

```
SELECT requests, title
```

Life of a BigQuery




```
SELECT sum(requests)
```

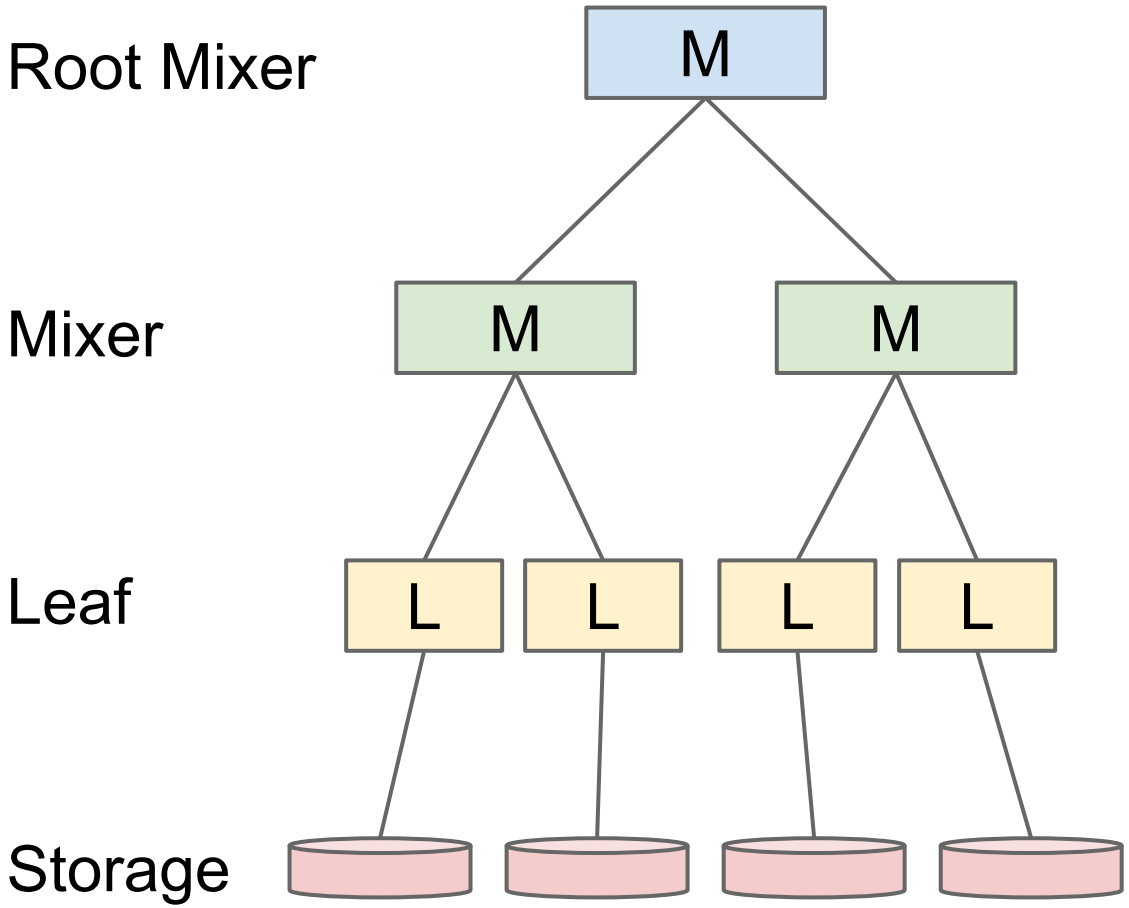
5.8 Mil

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

5.4 Bil

```
SELECT requests, title
```

Life of a BigQuery

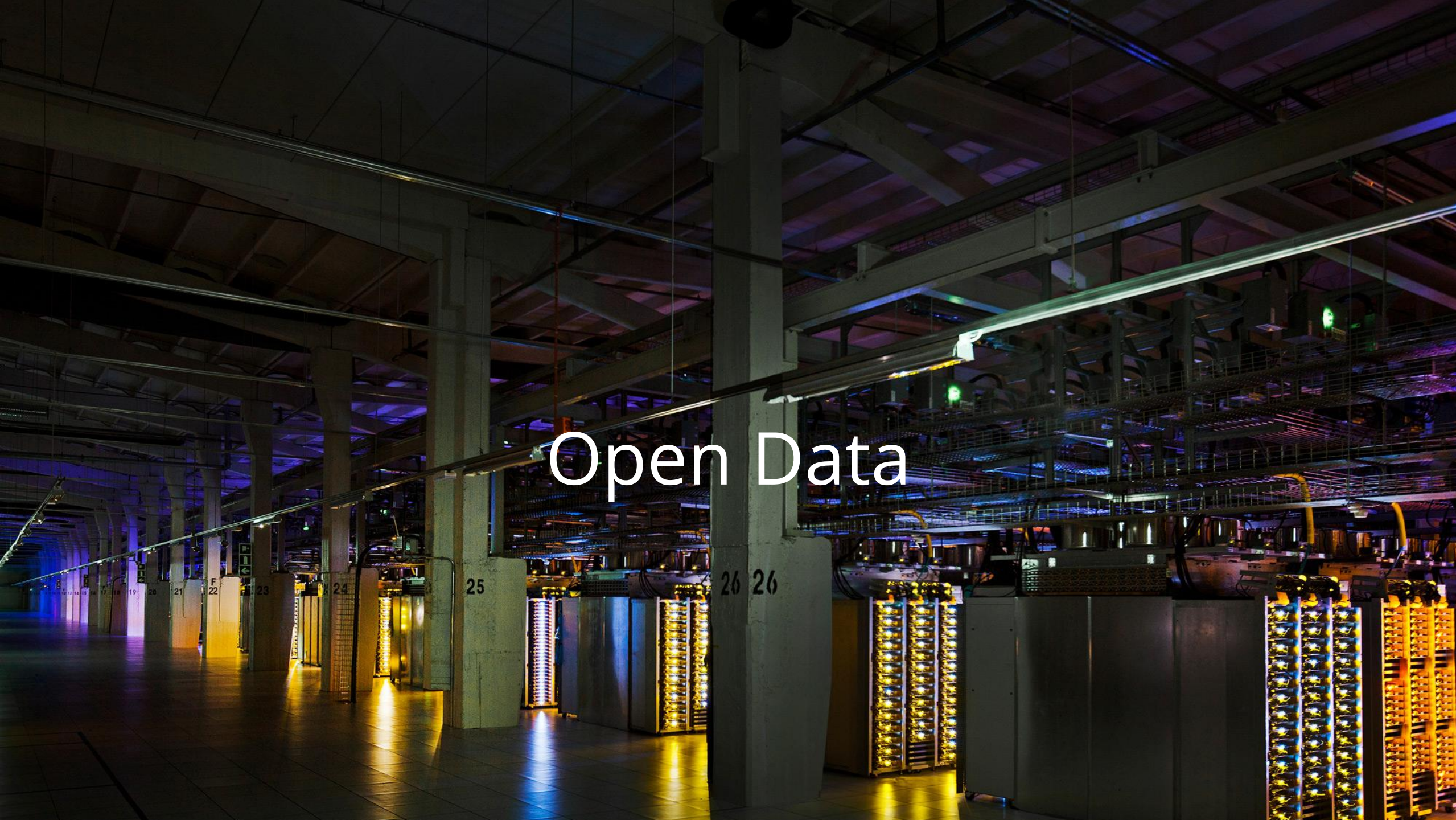



```
SELECT sum(requests)
```

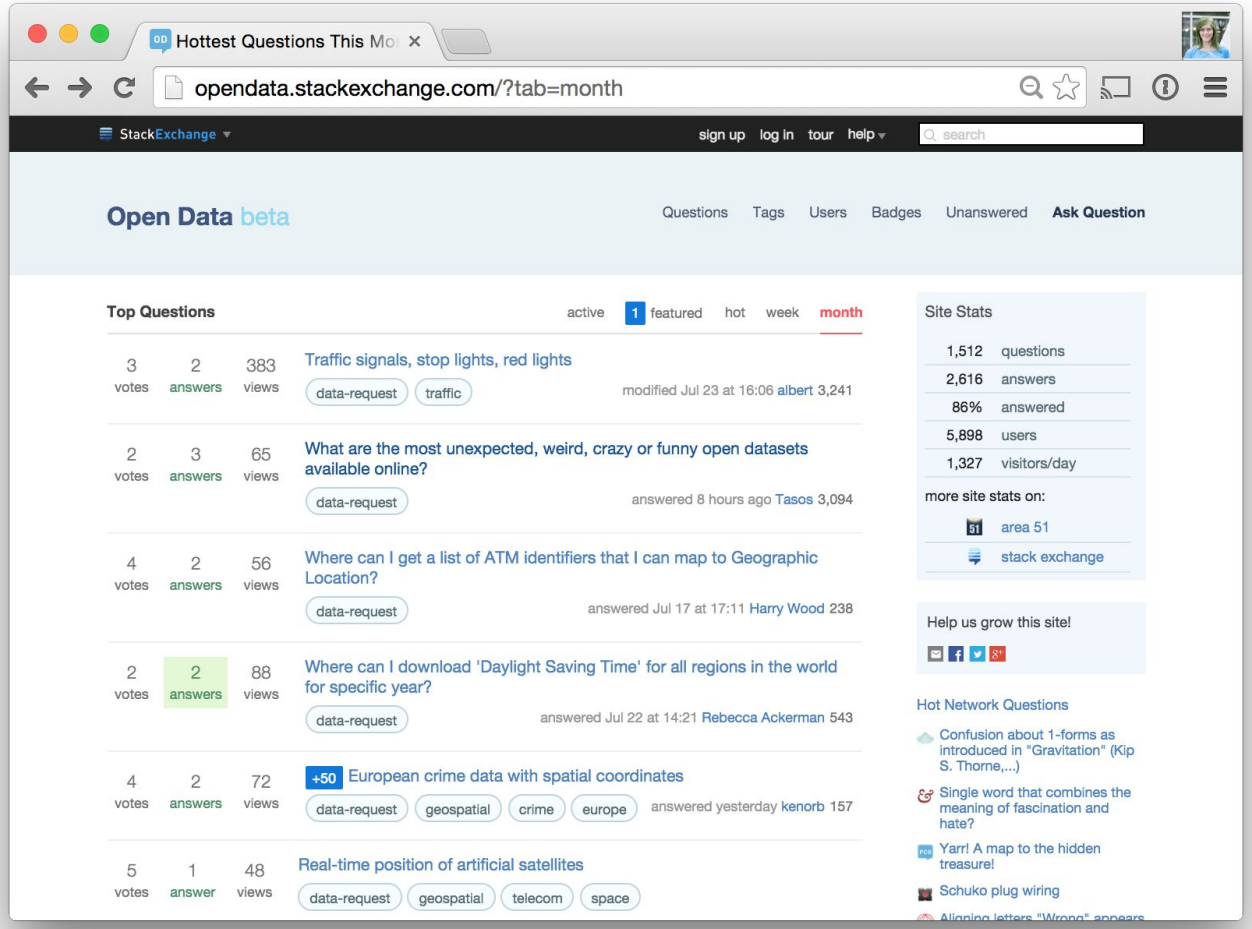
```
SELECT sum(requests)
```


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


```
SELECT requests, title
```

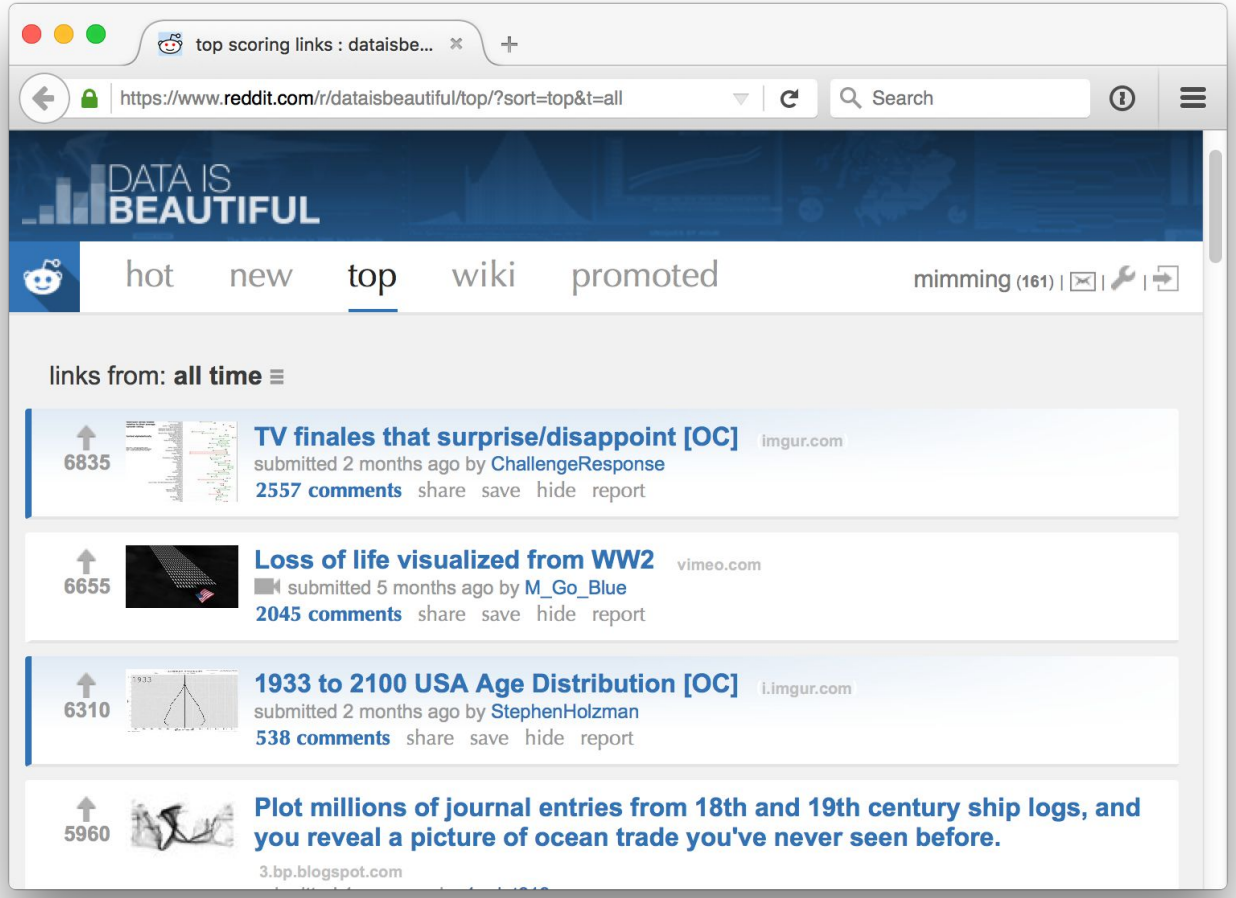

Open Data

Finding Open Data



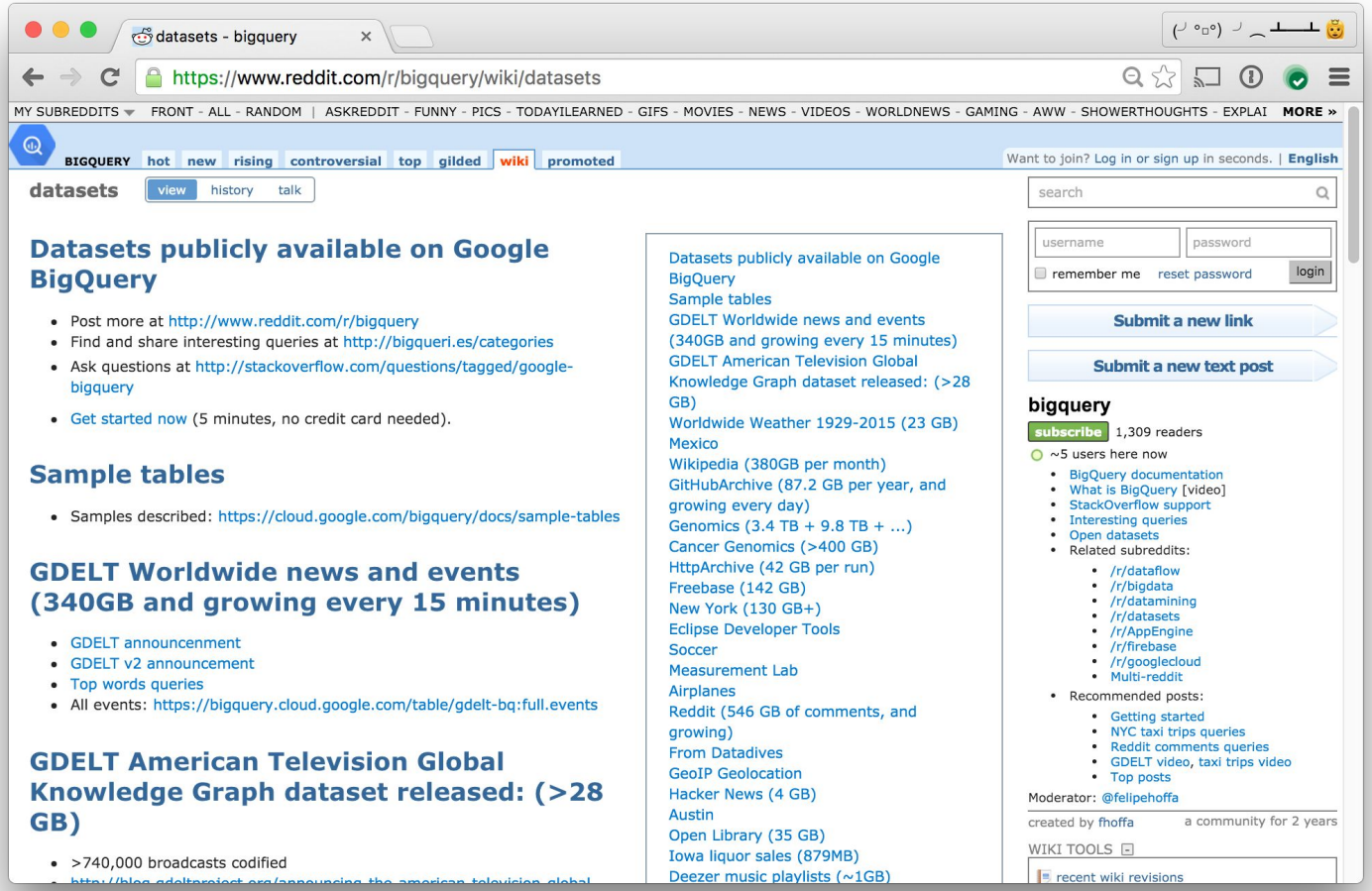
opendata.stackexchange.com

Finding Open Data



reddit.com/r/dataisbeautiful

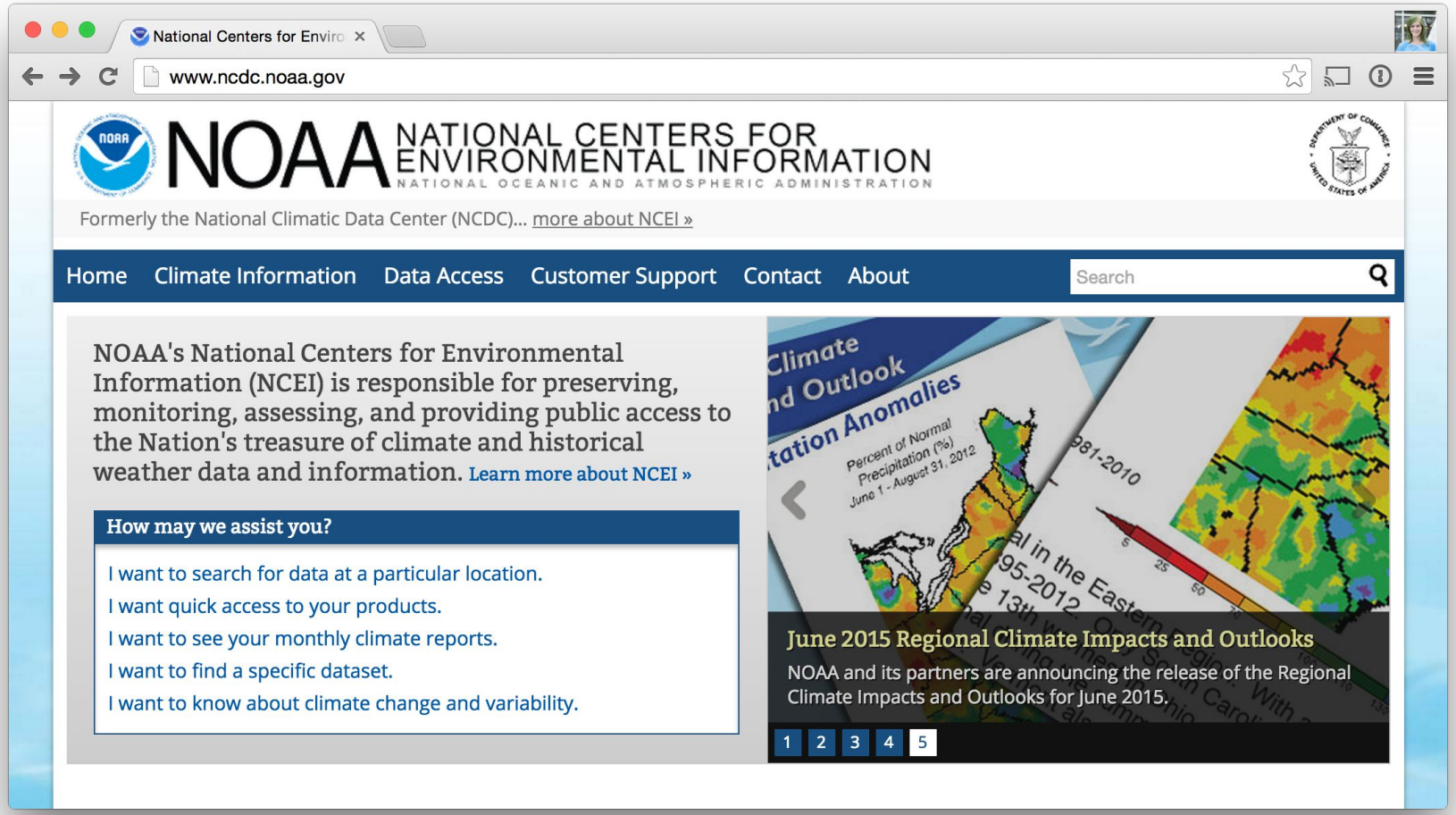
Finding Open Data



reddit.com/r/bigquery/wiki/datasets

Time to explore

GSOD


National Centers for Enviro
www.ncdc.noaa.gov
NOAA NATIONAL CENTERS FOR ENVIRONMENTAL INFORMATION
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
Formerly the National Climatic Data Center (NCDC)... more about NCEI »
Home
Climate Information
Data Access
Customer Support
Contact
About
Search
NOAA's National Centers for Environmental Information (NCEI) is responsible for preserving, monitoring, assessing, and providing public access to the Nation's treasure of climate and historical weather data and information. Learn more about NCEI »
How may we assist you?
I want to search for data at a particular location.
I want quick access to your products.
I want to see your monthly climate reports.
I want to find a specific dataset.
I want to know about climate change and variability.
June 2015 Regional Climate Impacts and Outlooks
NOAA and its partners are announcing the release of the Regional Climate Impacts and Outlooks for June 2015.
1 2 3 4 5

Find nearby weather data

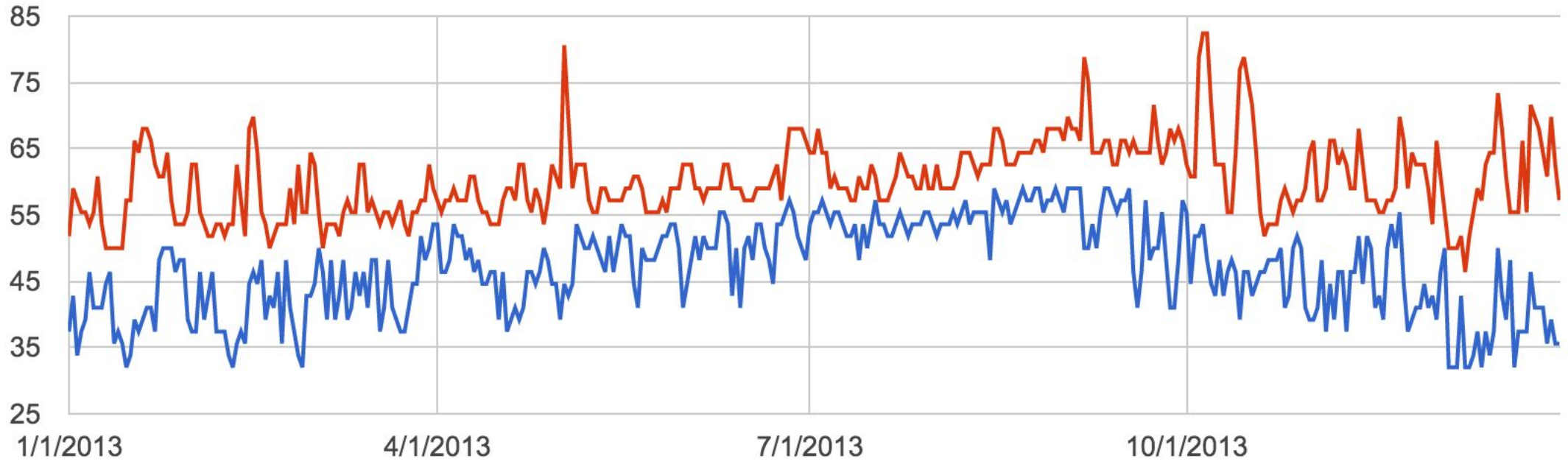
```
select name
from [fh-bigquery:weather_gsod.stations]
where
  state == 'GA' and
  usaf in(select stn from (SELECT count(stn) as cnt, stn
    FROM [fh-bigquery:weather_gsod.gsod2015]
    where stn <> '999999'
    group by stn order by cnt desc))
group by name
order by name ASC;
```

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 MUNICICPAL')
AND max < 200
ORDER BY day;
```

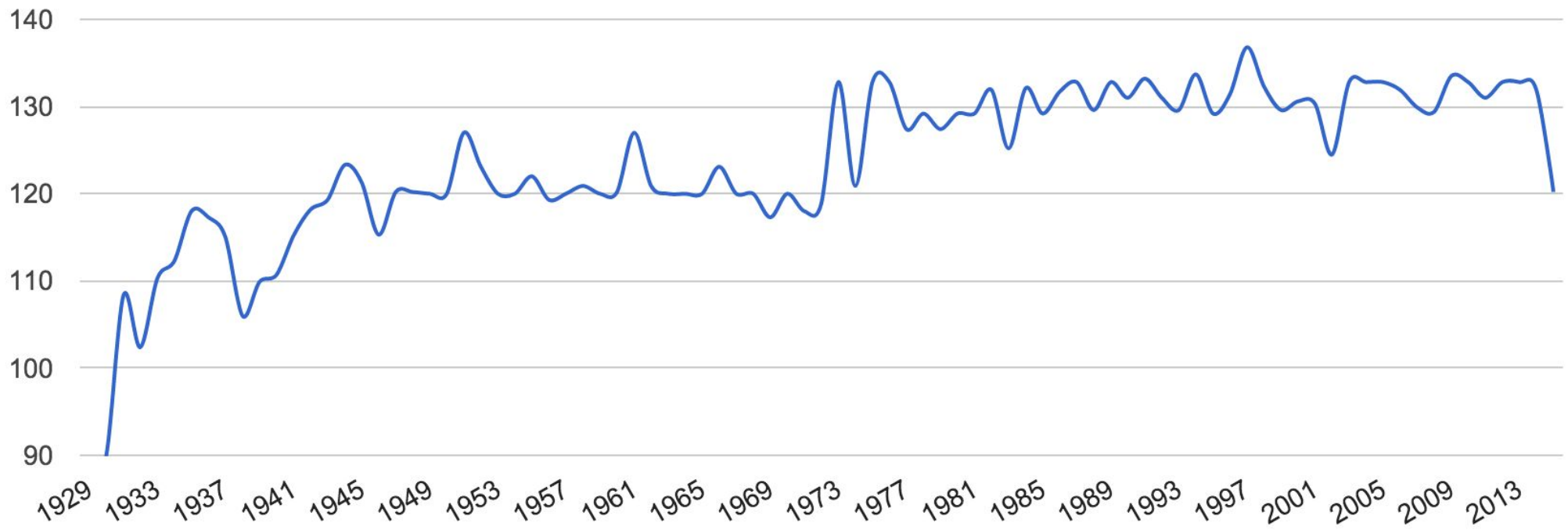
Weather in Half Moon Bay

```
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
```

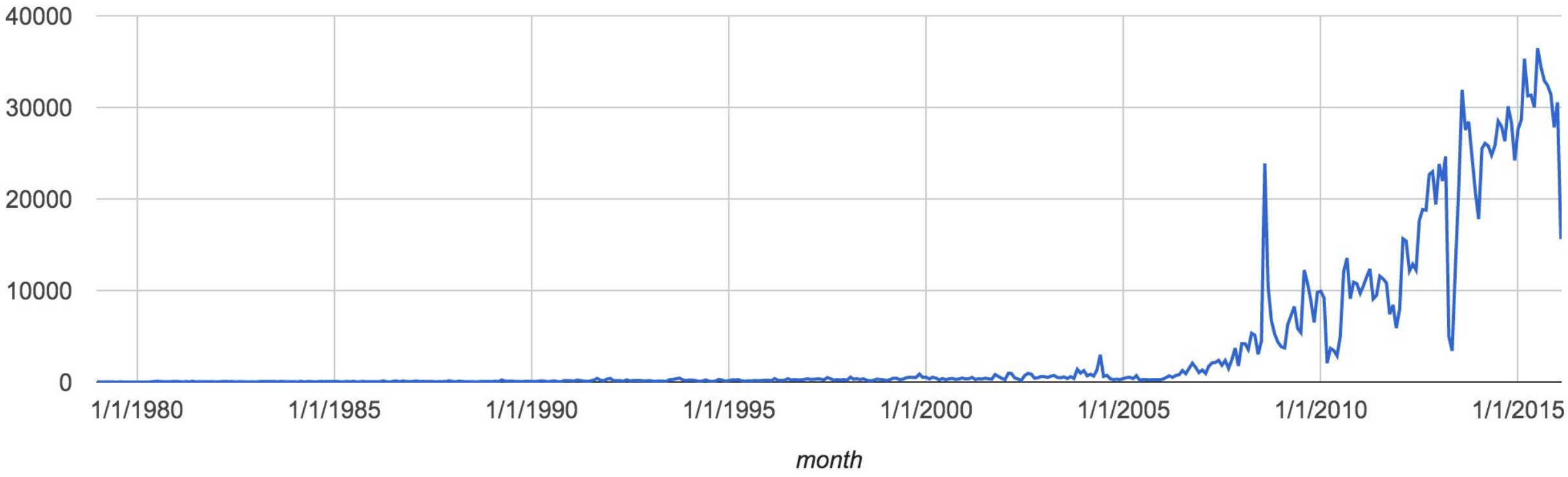


GDELT


The GDELT Project
gdeltproject.org
The GDELT Project
Blog
Data
Solutions
About
Intro
Watching
Computing
Downloading
Blogging
Starting
Watching Our World Unfold
Visualization credit GDELT Project.
Nov 02, 2013
Number of violent attacks
Number of Protests
A Global Database of Society
Supported by Google Ideas, the GDELT Project monitors the

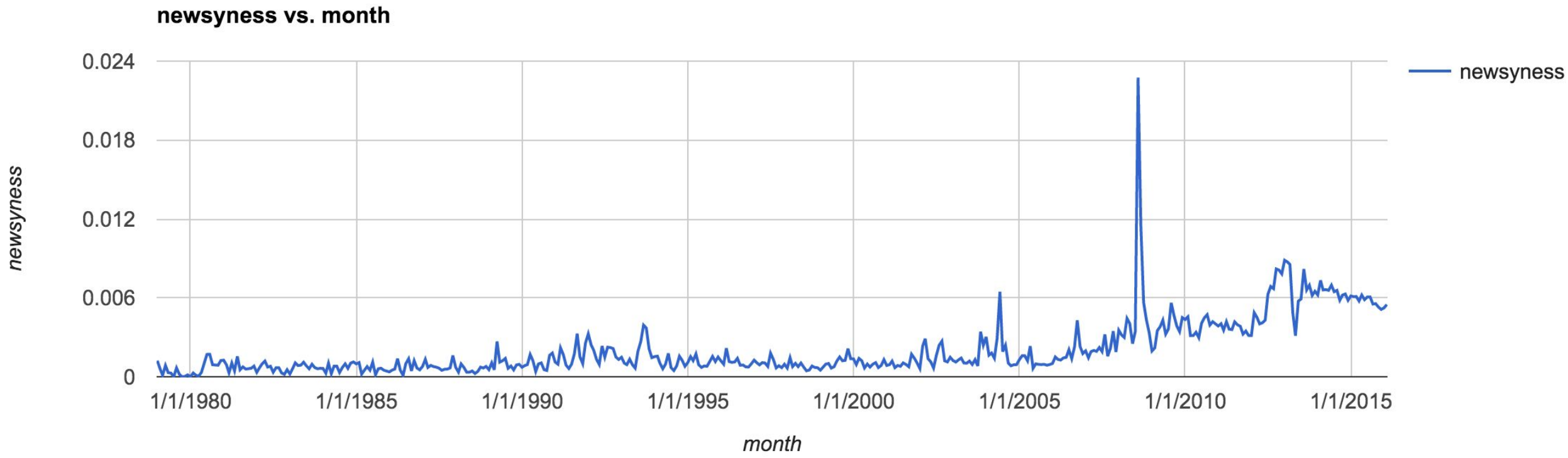
Stories per month - GA

```
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
```



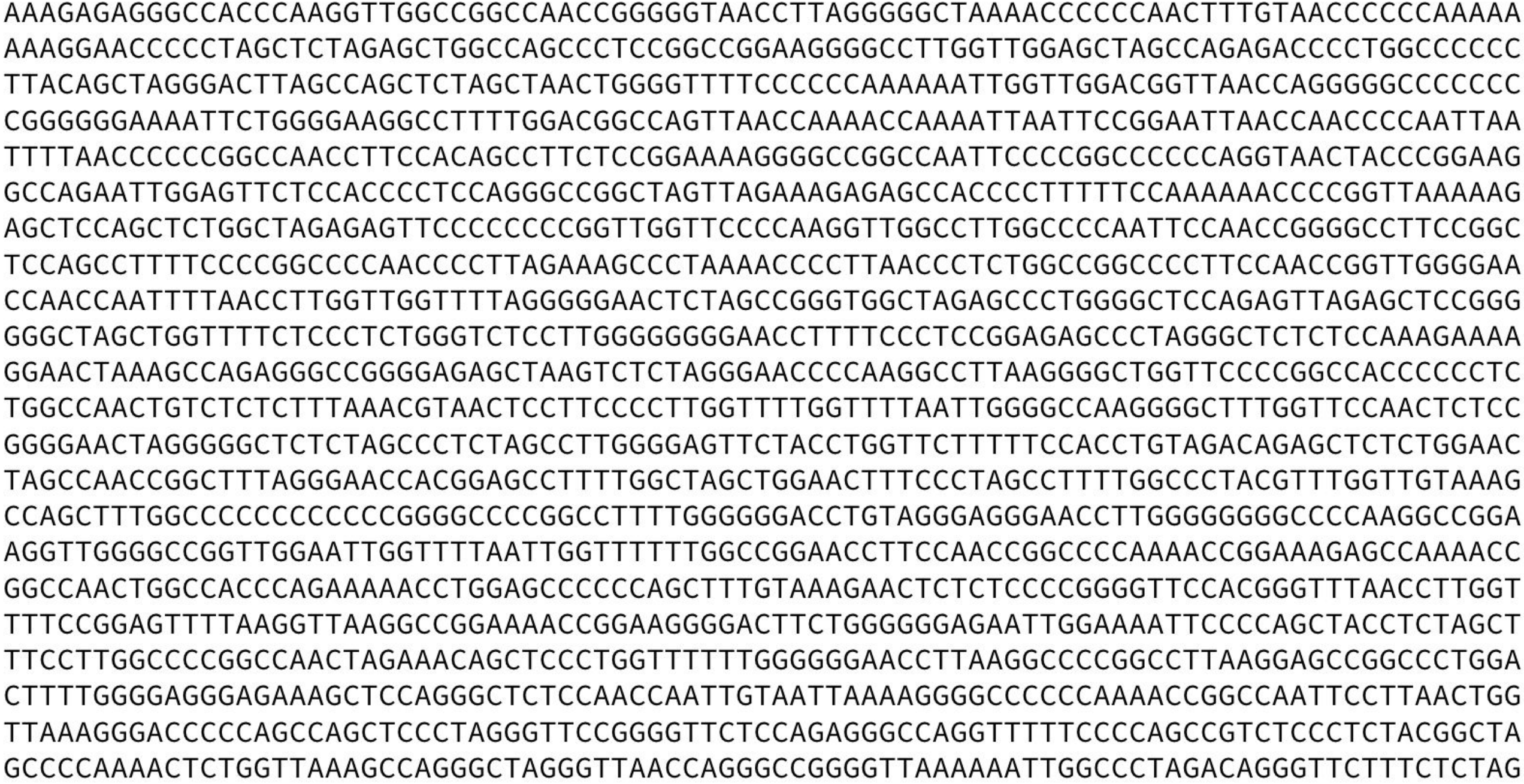
Stories per month, normalized

```
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
```

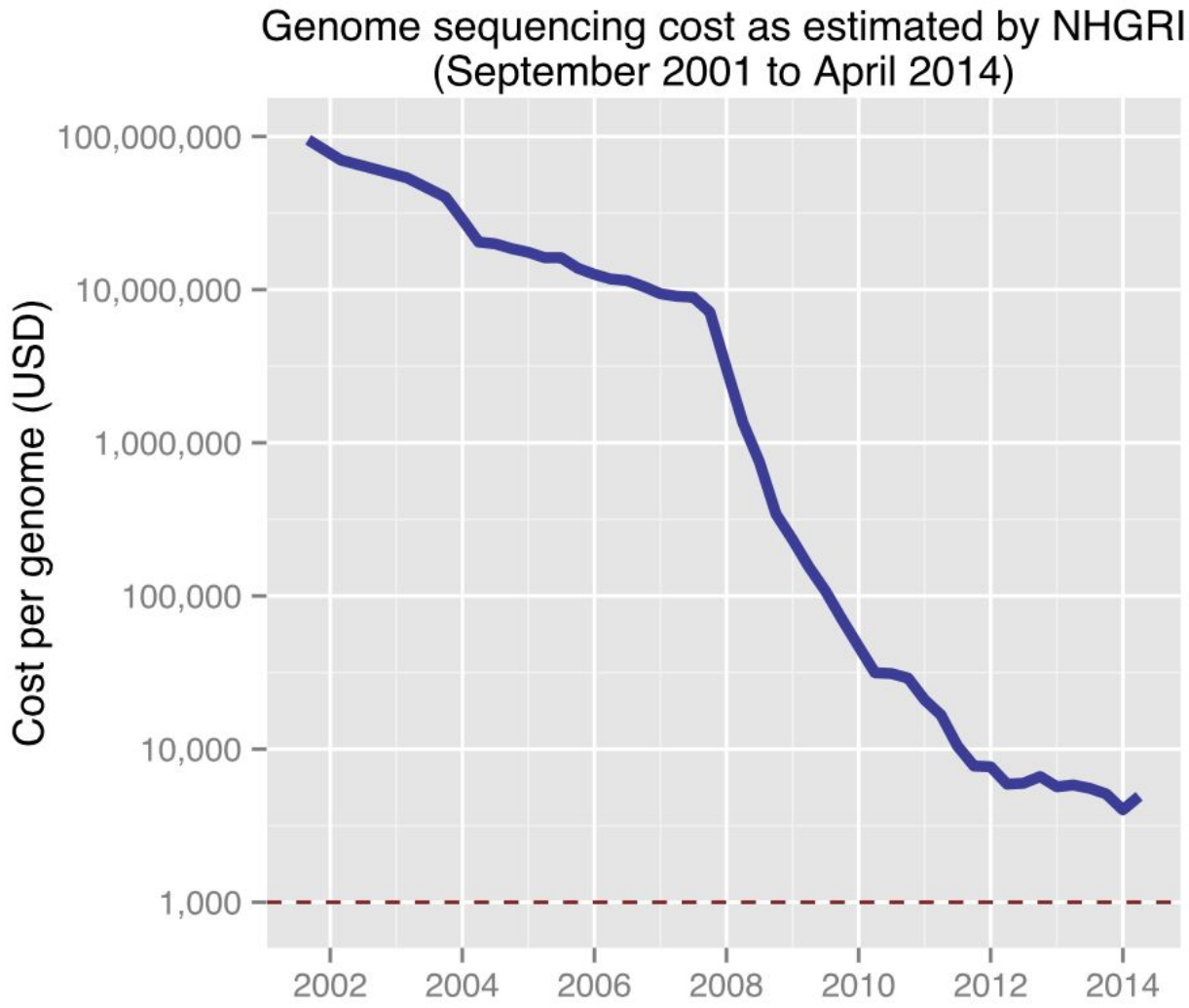


Genomics

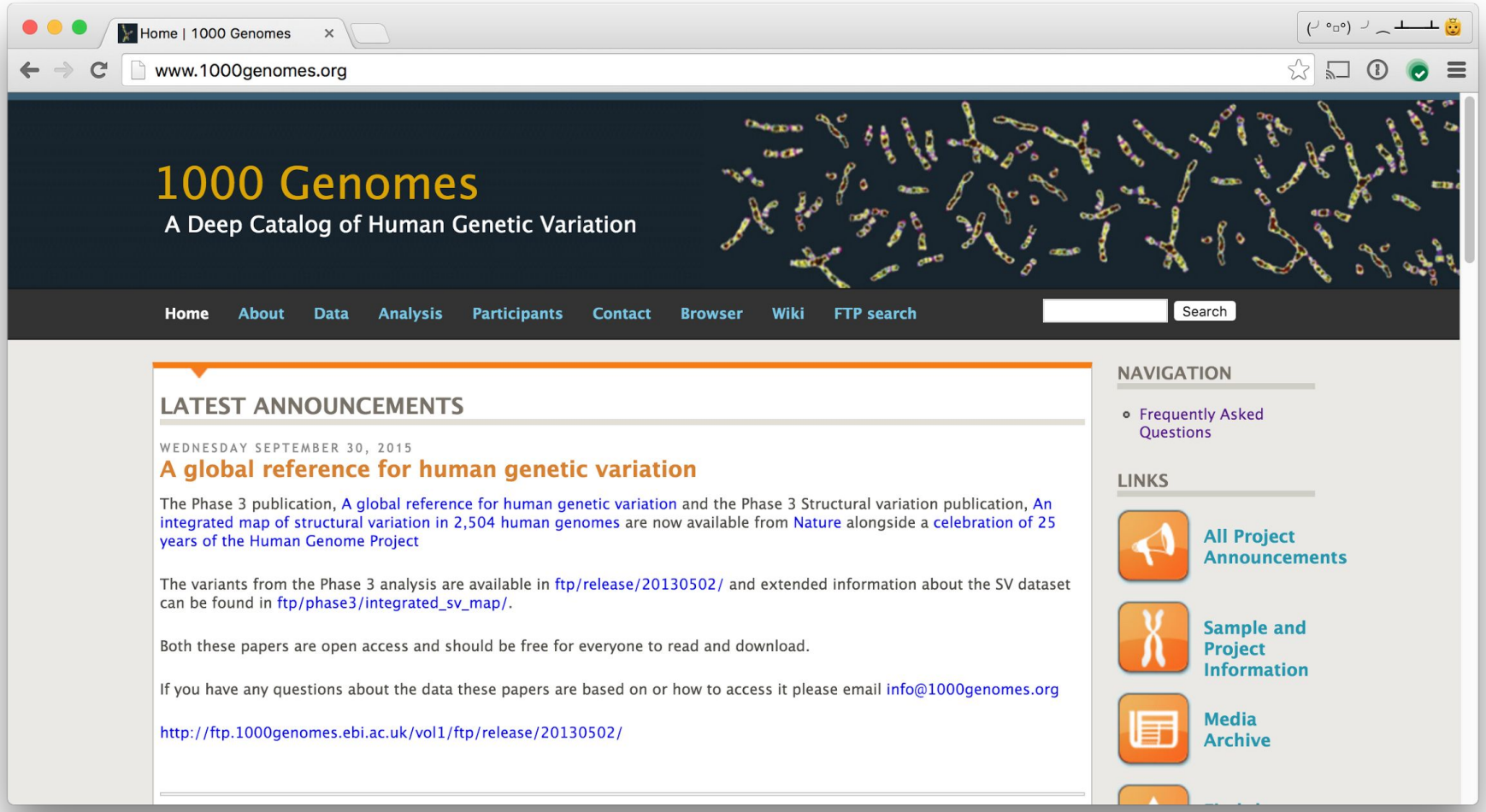
https://developers.google.com/genomics/


AAAGAGAGGGCCACCCAAGGTTGGCCGGCCAACCGGGGGTAACCTTAGGGGGCTAAAACCCCCCAACTTTGTAACCCCCCAAAAA
AAAGGAACCCCCTAGCTCTAGAGCTGGCCAGCCCTCCGGCCGGAAGGGGCCTTGGTTGGAGCTAGCCAGAGACCCCTGGCCCCCC
TTACAGCTAGGGACTTAGCCAGCTCTAGCTAACTGGGGTTTTCCCCCCAAAAAATTGGTTGGACGGTTAACCAGGGGGCCCCCCC
CGGGGGGAAAATTCTGGGGAAGGCCTTTTGGACGGCCAGTTAACCAAAACCAAAATTAATTCCGGAATTAACCAACCCCAATTAA
TTTTAACCCCCCGGCCAACCTTCCACAGCCTTCTCCGGAAAAGGGGCCGGCCAATTCCCCGGCCCCCCAGGTAACTACCCGGAAG
GCCAGAATTGGAGTTCTCCACCCCTCCAGGGCCGGCTAGTTAGAAAGAGAGCCACCCCTTTTTCCAAAAAACCCCGGTTAAAAAG
AGCTCCAGCTCTGGCTAGAGAGTTCCCCCCCCGGTTGGTTCCCCAAGGTTGGCCTTGGCCCCAATTCCAACCGGGGCCTTCCGGC
TCCAGCCTTTTCCCCGGCCCCAACCCCTTAGAAAGCCCTAAAACCCCTTAACCCTCTGGCCGGCCCCTTCCAACCGGTTGGGGAA
CCAACCAATTTTAACCTTGGTTGGTTTTAGGGGGAACTCTAGCCGGGTGGCTAGAGCCCTGGGGCTCCAGAGTTAGAGCTCCGGG
GGGCTAGCTGGTTTTCTCCCTCTGGGTCTCCTTGGGGGGGGAACCTTTTCCCTCCGGAGAGCCCTAGGGCTCTCTCCAAAGAAAA
GGAACTAAAGCCAGAGGGCCGGGGAGAGCTAAGTCTCTAGGGAACCCCAAGGCCTTAAGGGGCTGGTTCCCCGGCCACCCCCCTC
TGGCCAACTGTCTCTCTTTAAACGTAACTCCTTCCCCTTGGTTTTGGTTTTAATTGGGGCCAAGGGGCTTTGGTTCCAACTCTCC
GGGGAACTAGGGGGCTCTCTAGCCCTCTAGCCTTGGGGAGTTCTACCTGGTTCTTTTTCCACCTGTAGACAGAGCTCTCTGGAAC
TAGCCAACCGGCTTTAGGGAACCACGGAGCCTTTTGGCTAGCTGGAACTTTCCCTAGCCTTTTGGCCCTACGTTTGGTTGTAAAG
CCAGCTTTGGCCCCCCCCCCCCGGGGCCCCGGCCTTTTGGGGGGACCTGTAGGGAGGGAACCTTGGGGGGGGCCCCAAGGCCGGA
AGGTTGGGGCCGGTTGGAATTGGTTTTAATTGGTTTTTTGGCCGGAACCTTCCAACCGGCCCCAAAACCGGAAAGAGCCAAAACC
GGCCAACTGGCCACCCAGAAAAACCTGGAGCCCCCCAGCTTTGTAAAGAACTCTCTCCCCGGGGTTCCACGGGTTTAACCTTGGT
TTTCCGGAGTTTTAAGGTTAAGGCCGGAAAACCGGAAGGGGACTTCTGGGGGGAGAATTGGAAAATTCCCCAGCTACCTCTAGCT
TTCCTTGGCCCCGGCCAACTAGAAACAGCTCCCTGGTTTTTTGGGGGGAACCTTAAGGCCCCGGCCTTAAGGAGCCGGCCCTGGA
CTTTTGGGGAGGGAGAAAGCTCCAGGGCTCTCCAACCAATTGTAATTAAAAGGGGCCCCCCAAAACCGGCCAATTCCTTAACTGG
TTAAAGGGACCCCCAGCCAGCTCCCTAGGGTTCCGGGGTTCTCCAGAGGGCCAGGTTTTTCCCCAGCCGTCTCCCTCTACGGCTA
GCCCCAAAACTCTGGTTAAAGCCAGGGCTAGGGTTAACCAGGGCCGGGGTTAAAAAATTGGCCCTAGACAGGGTTCTTTCTCTAG

Cost to sequence a genome

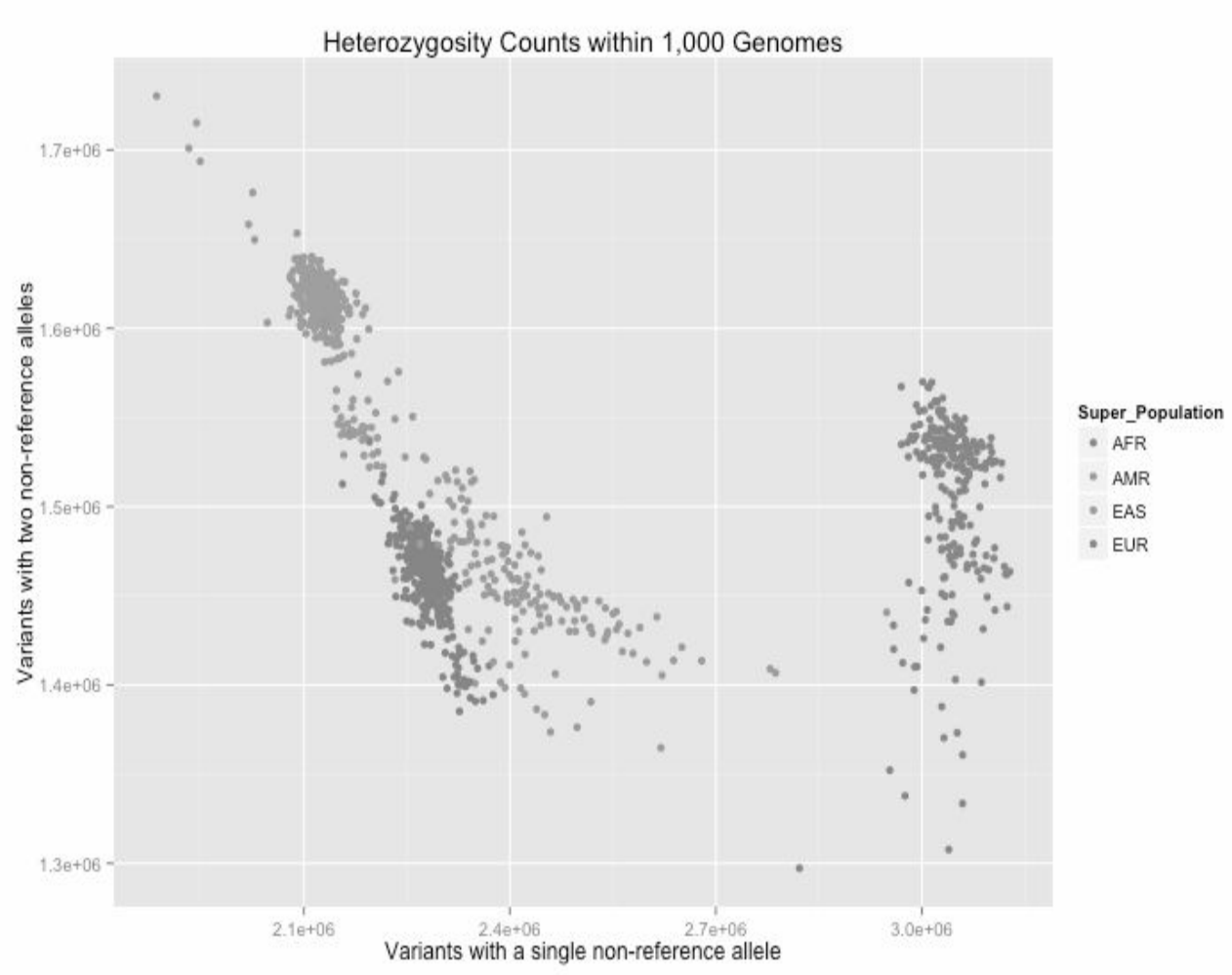


1000 Genomes



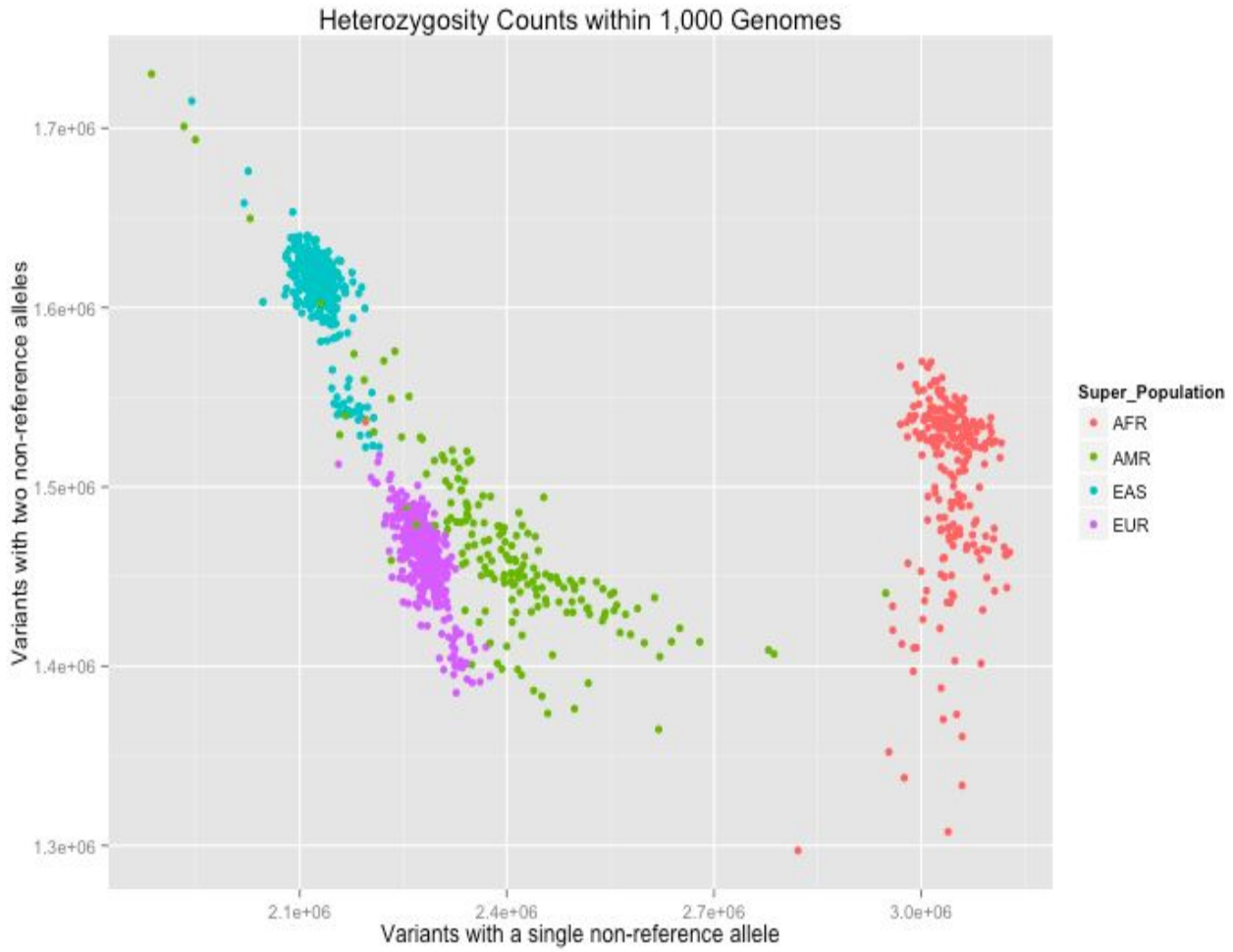
Genomics

```
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
```



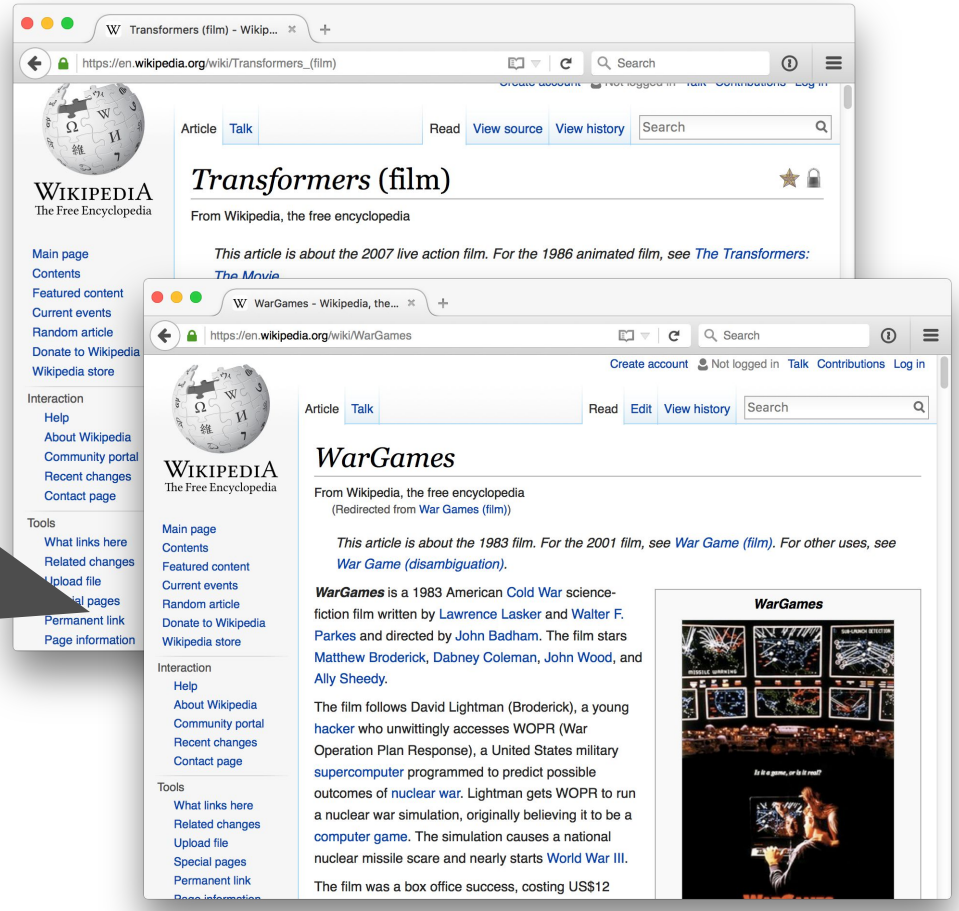
Something useful:

Use Wikipedia data to pick a movie

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

Follow the edits

Same editor



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
```

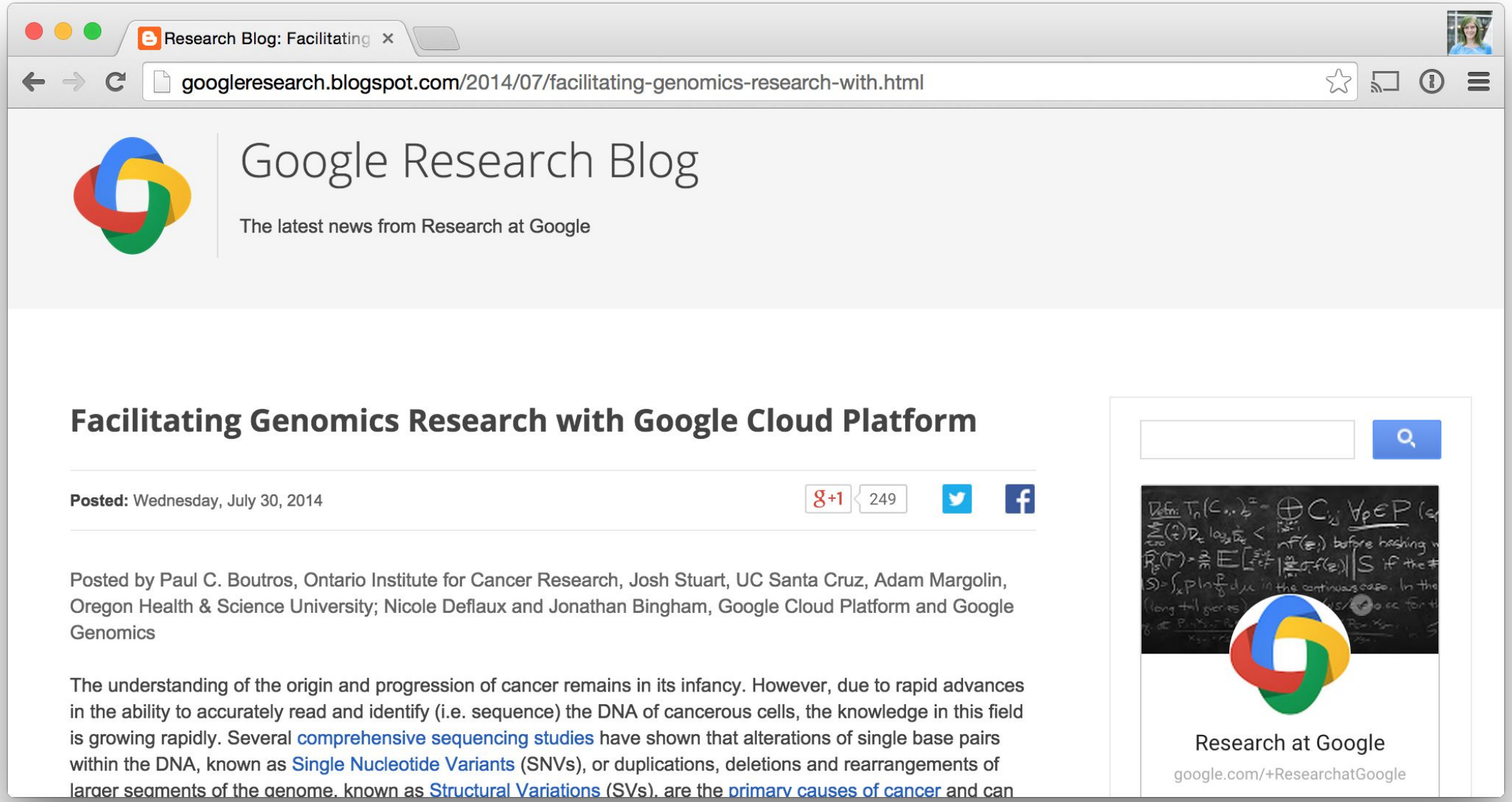
What we talked about

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

The end

- Try BigQuery
  - bigquery.cloud.google.com
- Me: @MimmingCodes
- Slides: mimming.com/presos/
- Queries: github.com/mimming/snippets

ISB teams with Google for NCI Cancer Genomics Cloud project


Research Blog: Facilitating
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
+1 249
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
Research at Google
google.com/+ResearchatGoogle