

Search Is the New Big Data

Loren Siebert
DigitalGov Search Team
April 10, 2014

TL;DR

1. Search is Easy
2. Search is Hard
3. Search has many shades of grey

About DigitalGov Search


- Search as a Service for ~1500 gov/mil sites
- Citizens get commercial search results augmented with customer-specific content
- Agencies get powerful and timely analytics

On the Search Side

- Many different document types, from tweets to PDFs
- Some small, some big (~1 Billion documents)


Recommended by U.S. Geological Survey

★ [Natural Hazards Mission Area](#)
www.usgs.gov/natural_hazards
The USGS works with many partners to monitor, assess, and conduct targeted research on a wide range of natural hazards so that policymakers and the public have the understanding they need to enhance preparedness, response and resilience.

Recent tweets for 'earthquakes' by U.S. Geological Survey 

★ [USGS Big Quakes](#) @USGSBigQuakes
Strong **earthquake**, NEAR COAST OF TARAPACA, CHILE, Apr-7 13:43 UTC, on.doi.gov/1fWoldz
about 3 hours ago





★ **Earthquake Information** by U.S. Geological Survey

 [Earthquake Updates](#) [Earthquake data](#)
[Map of latest earthquakes](#) [Mobile and text alerts](#)
[Did You Feel It?](#)

USGS Earthquake
earthquake.usgs.gov
USGS **Earthquake** Hazards Program, responsible for monitoring, reporting, and researching **earthquakes** and **earthquake** hazards

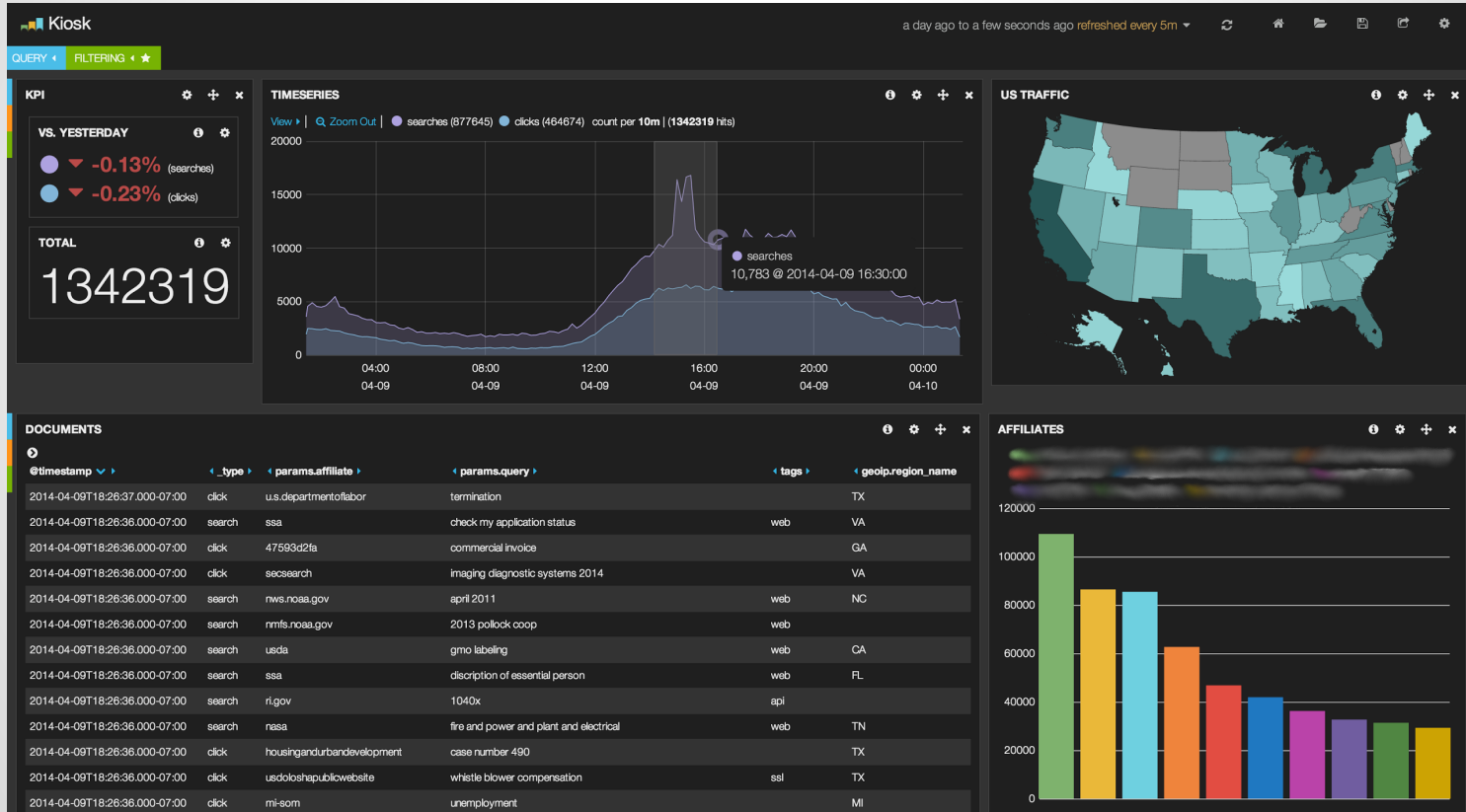
Earthquakes - Earthquake Hazards Program
earthquake.usgs.gov/earthquakes/map
Latest Earthquakesv0.4.4, 2014-01-07. List Clicking the list icon in the top right corner will load the **earthquake** list. Map Clicking the map icon in ...

Videos of 'earthquakes' by U.S. Geological Survey

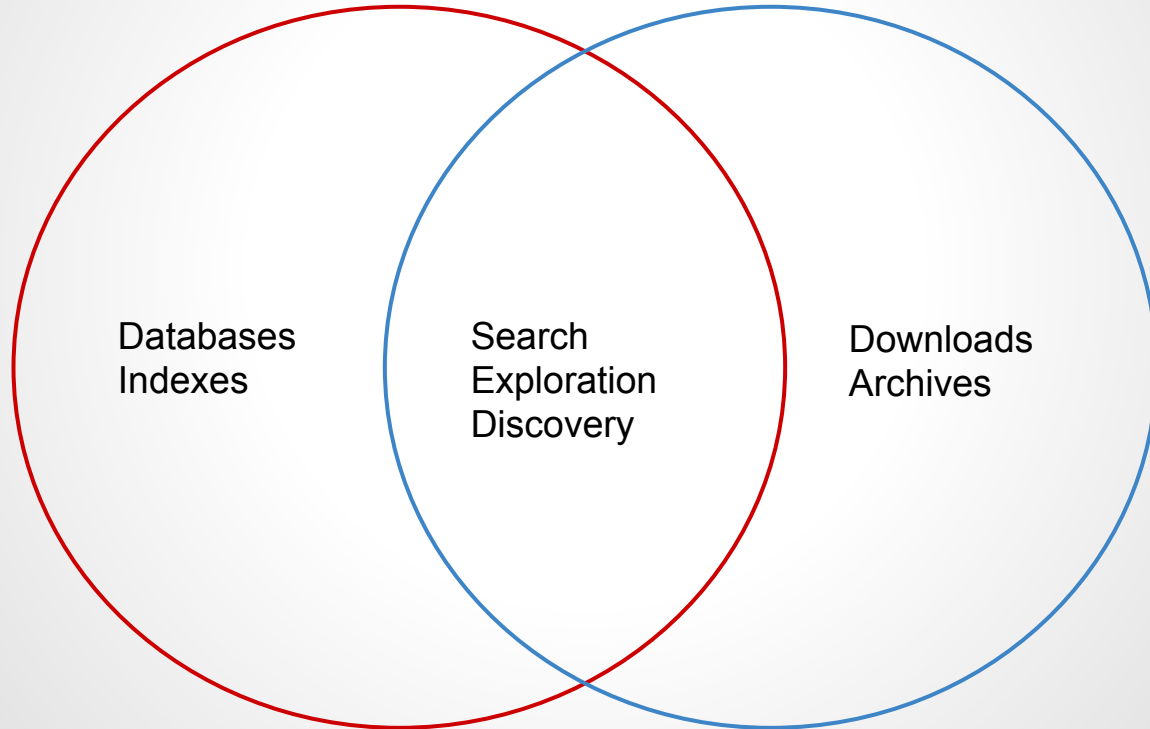
★    

[1964 Quake: The Great Alaska E ...](#) [Magnitude 9.2: The 1964 Great ...](#) [Northridge, CA Earthquake](#) [Shock Waves: 100 Years After](#)
2/27/2014 ... 1/16/2014 1/16/2014 9/17/2013

On the Analytics Side



Transparency: Tech meets Data



Search is Easy

```
PUT /contacts/entry/1
```

```
{  "name": "National Security Agency",  
  "city": "Fort Meade",  
  "state": "MD",  
  "notes": "summer intern job"}
```

```
GET /contacts/entry/_search?q=Agency
```

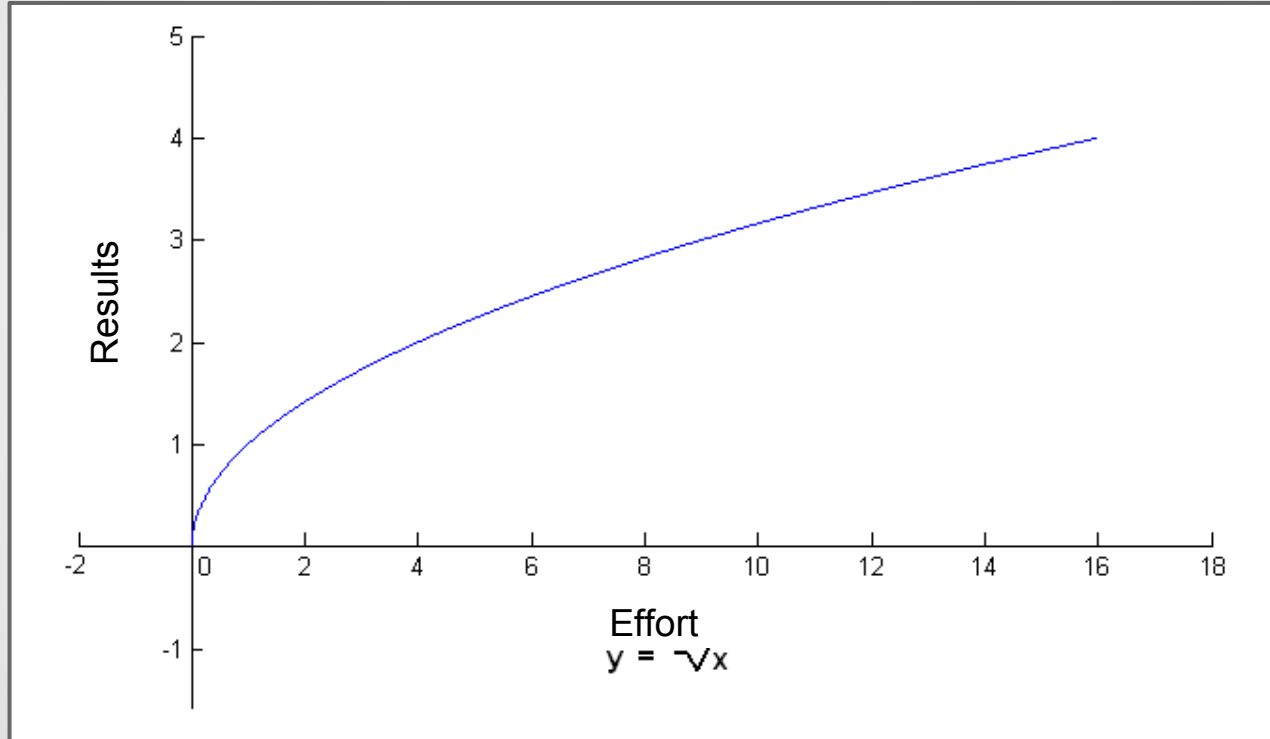
```
"National Security Agency"
```

That worked, but ...

```
{  "name": "National Security Agency",  
  "city": "Fort Meade",  
  "state": "MD",  
  "notes": "summer intern job"}
```

Query term	Hits
Ft. Meade	0
md	0
the National Security Agency	0
National Securite Agency	0
interns	0

Search is Hard



Recall & Relevancy

recall:

fraction of *relevant*
documents that are
retrieved

relevancy:

fraction of *retrieved*
documents that are
relevant

TF-IDF

- The more the term appears in a *document*, the higher the term frequency (TF).
- The more the term appears across the *corpus*, the lower the inverse document frequency (IDF).
- Additional signal can help improve relevancy.

Popular Search Software

Lucene



Solr



Elasticsearch



Sprinkle Search Magic

```
{  "name": "National Security Agency",  
  "city": "Fort Meade",  
  "state": "MD",  
  "notes": "summer intern job"}
```

Query term	Problem	Solution
Ft. Meade	Ft. vs Fort	synonyms
md	case	downcase
the National Security Agency	the	stopwords
National Securite Agency	spelling, accent	fuzziness, folding
interns	word form	stemming

Powerful Query Capabilities

What Agencies in Marylandd have interns?

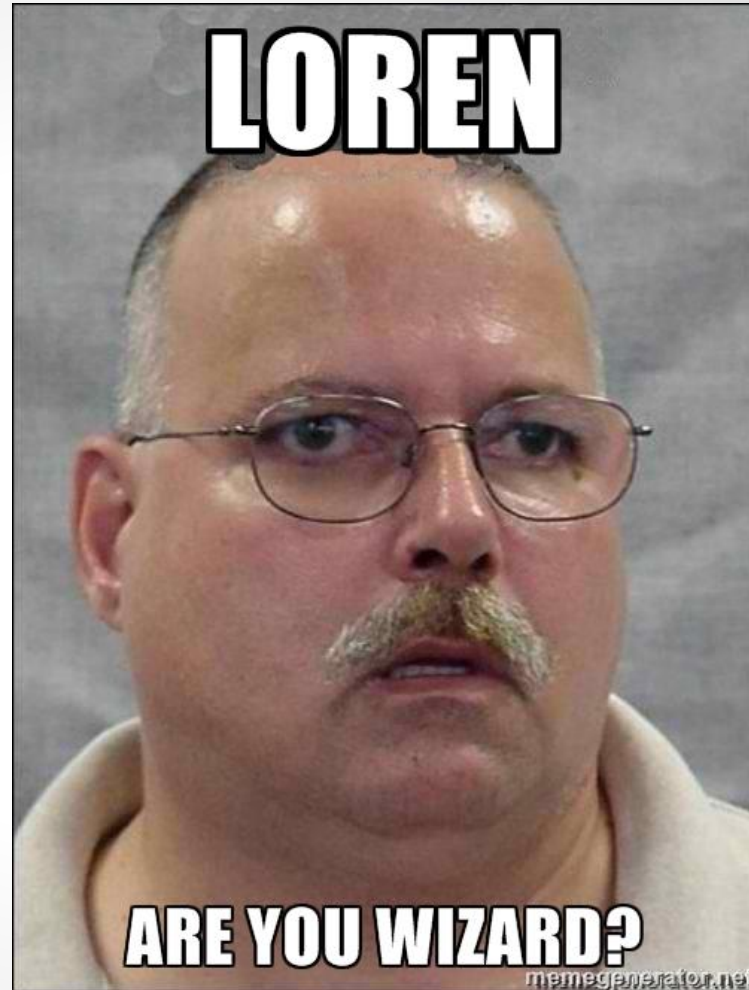


```
GET /contacts/entry/_search?q=What%20%C3%85gencies%20  
in%20Marylandd%20have%20interns%3F
```

"National Security **Agency**"

"summer **intern** job"

Ship it!



Demo Day

Internal Revenue Service



`GET /contacts/entry/_search?q=Internal%20Revenue%20Service`

Demo Day

Internal Revenue Service



`GET /contacts/entry/_search?q=Internal%20Revenue%20Service`

`"summer intern job"`

Demo Day

Agency for International Development



```
GET /contacts/entry/_search?q=Agency%20for%20International%20Development
```

Demo Day

International Development



`GET /contacts/entry/_search?q=International%20Development`

`"summer intern job"`

A Snowball's Chance in English

Raw Term	Stemmed Token
interns, internal, international	intern-
securities, security	secur-
Maine, main	main-
season, seasoning	season-
image, imaging	imag-
physics, physical	physic-
IRS	ir-

Best Practices

A search system is a database with an opinion.
Where does it get that opinion?

- Sensible defaults get you pretty far.
- Analysis chain is where the effort goes.
- Refinement is ongoing.
- Make it easy to reindex.

Search APIs & Data Products

What do you expose?

- Schema?
- Filters?
- Lucene itself?

How much “search magic” is enough?

- Stemming, synonyms, stopwords, etc

Thank you!

<http://search.digitalgov.gov>

202-505-5315 | @DG_Search