elasticsearch
beyond full-text search

#gotoaar #elasticsearch

Alexander Reelsen
@spinscale
alexander.reelsen@elasticsearch.com

About me

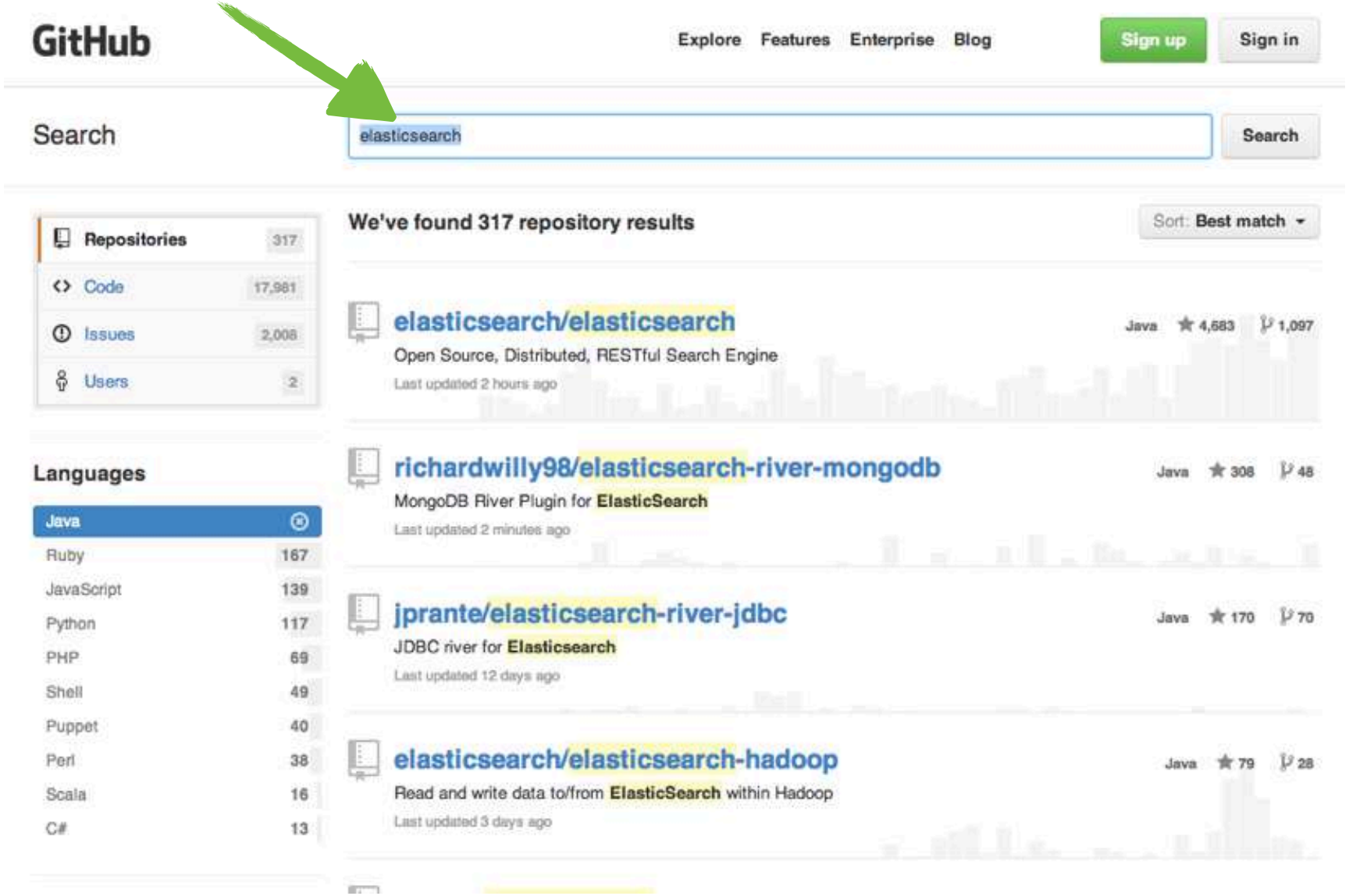
- Elasticsearch core developer

  Features, bug fixing, package maintenance, documentation, blog posts
- Development support
- Production support
- Trainings
- Conferences & talks

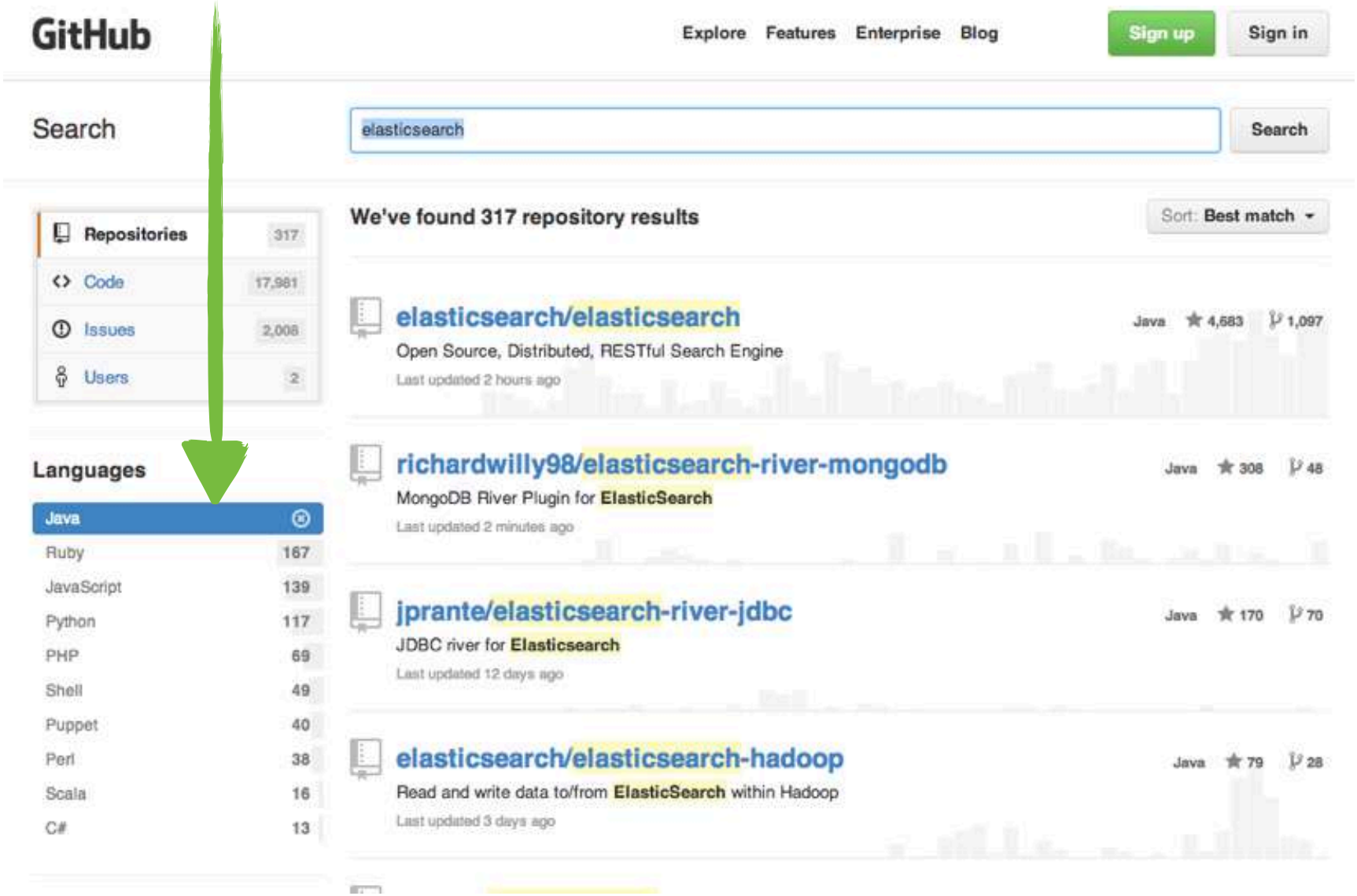
- Interests: Java, JavaScript, web apps

Beyond full-text search?

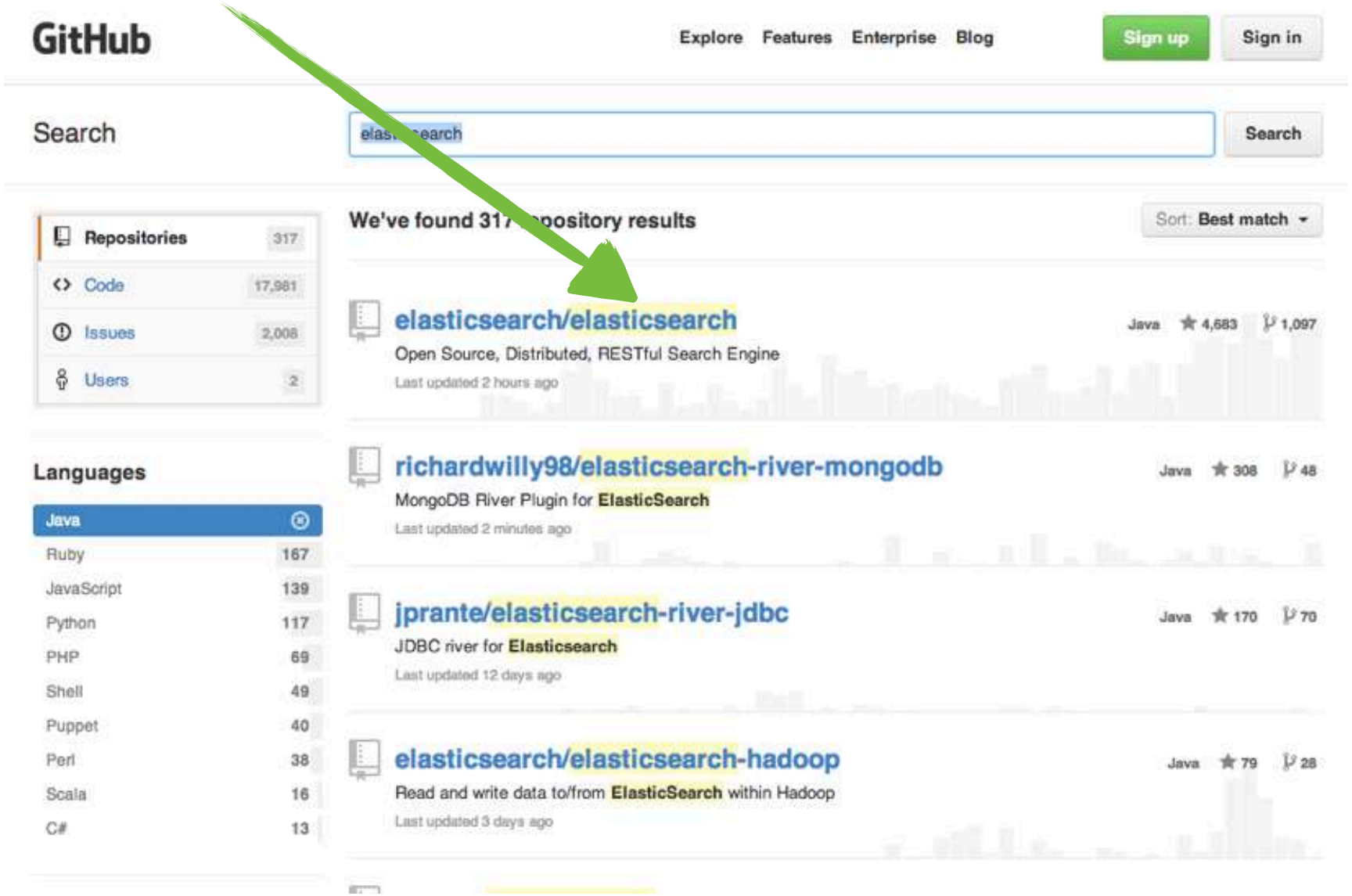
Unstructured search



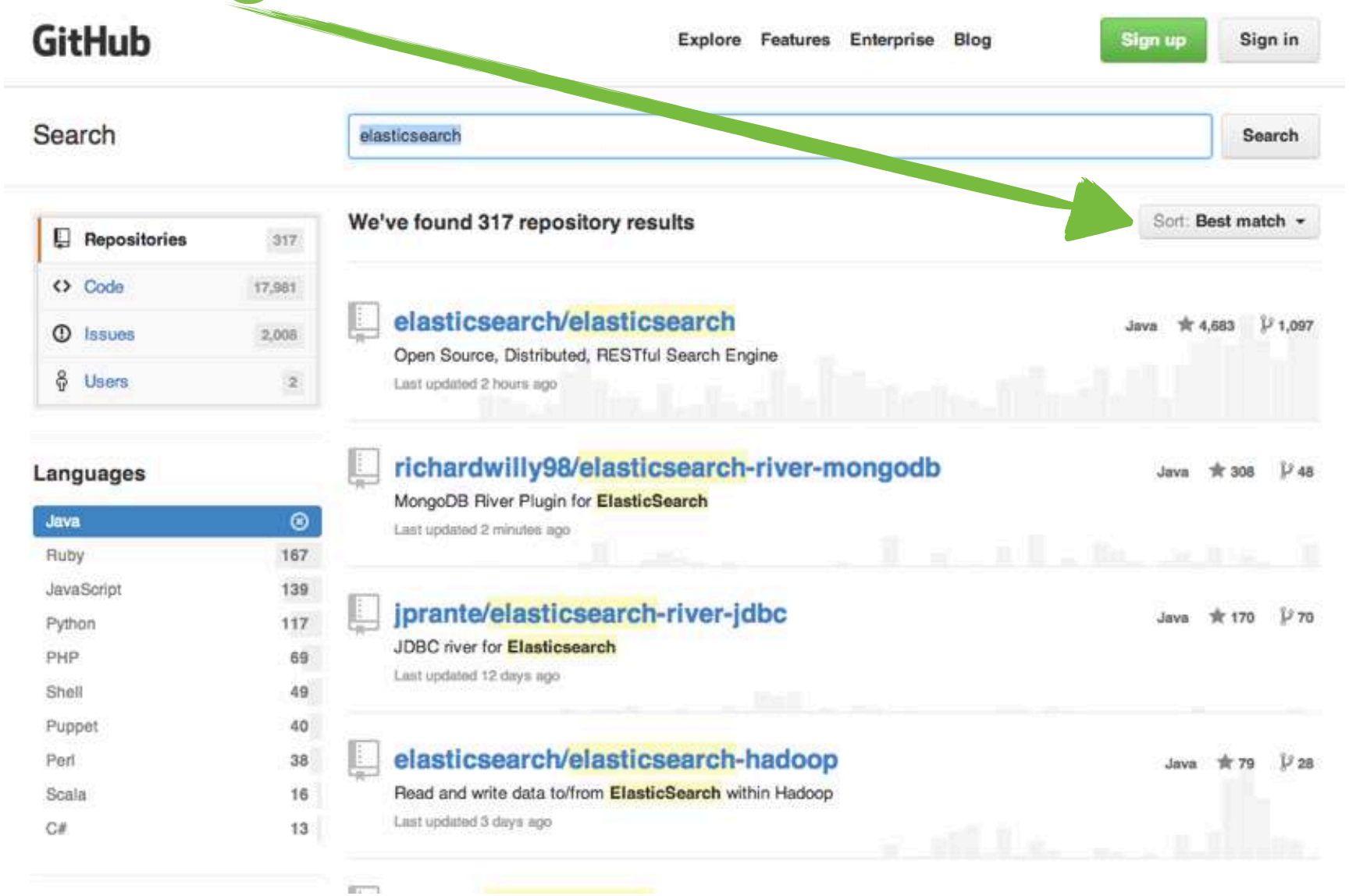
Structured search



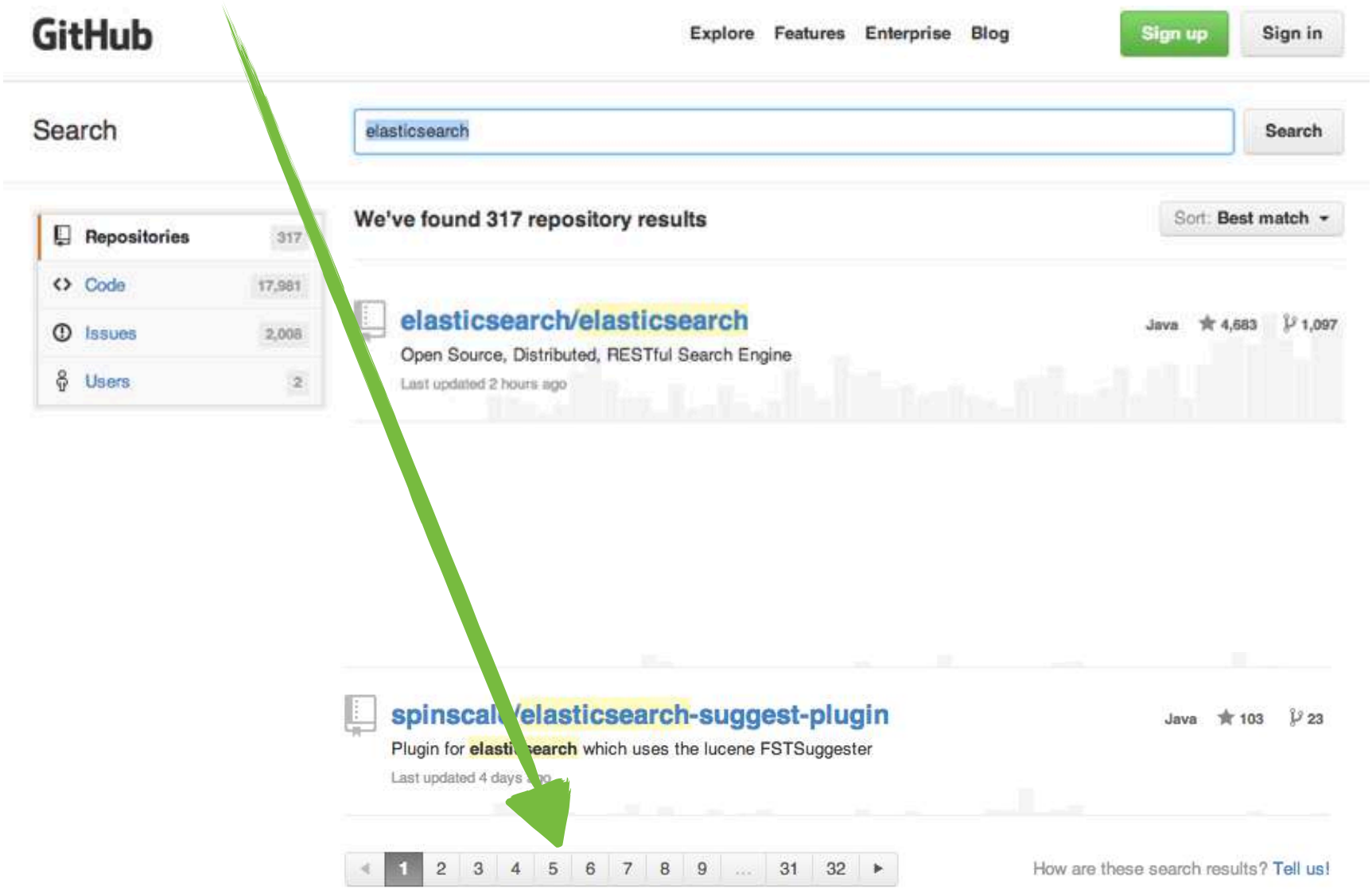
Enrichment



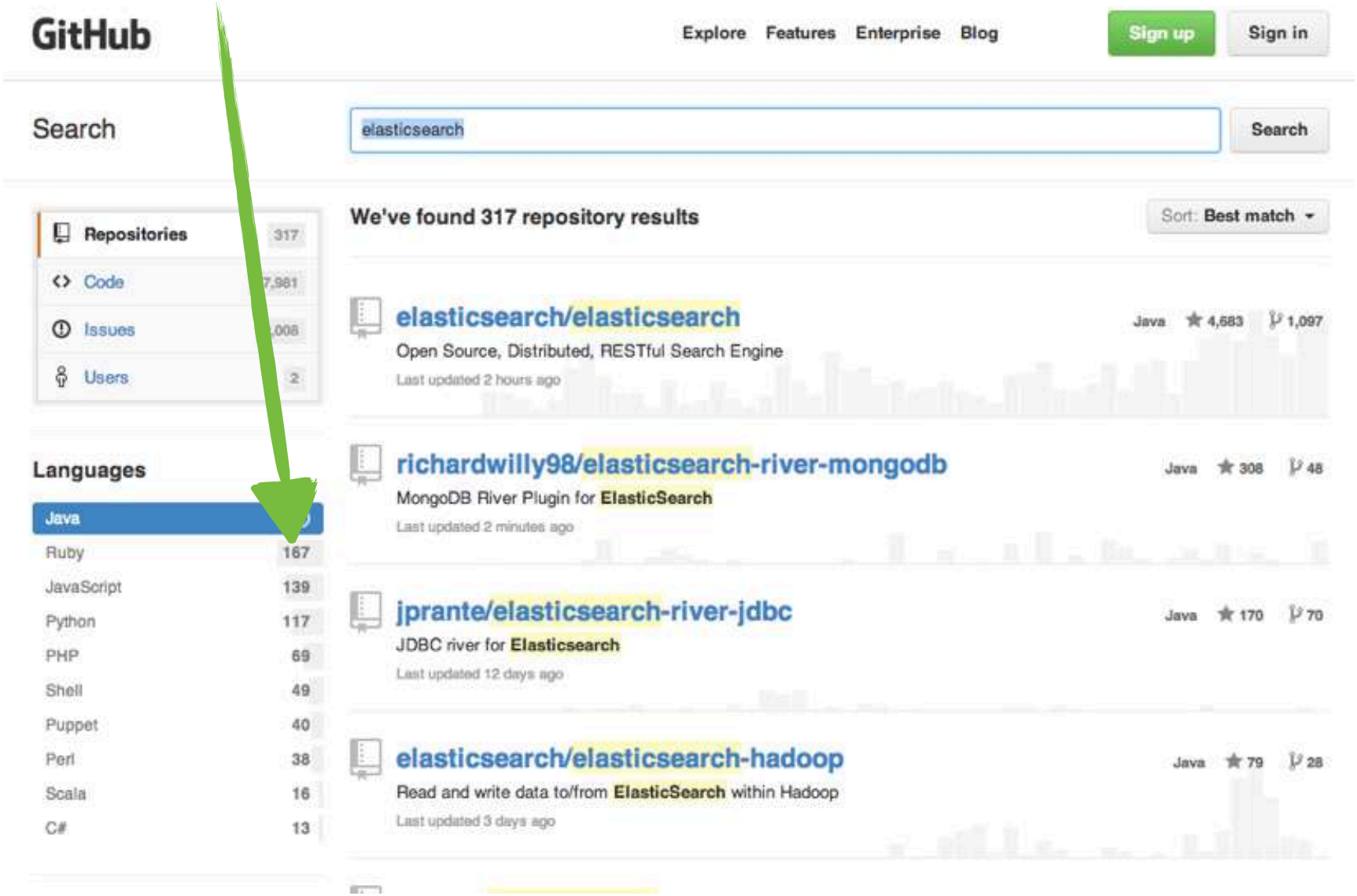
Sorting



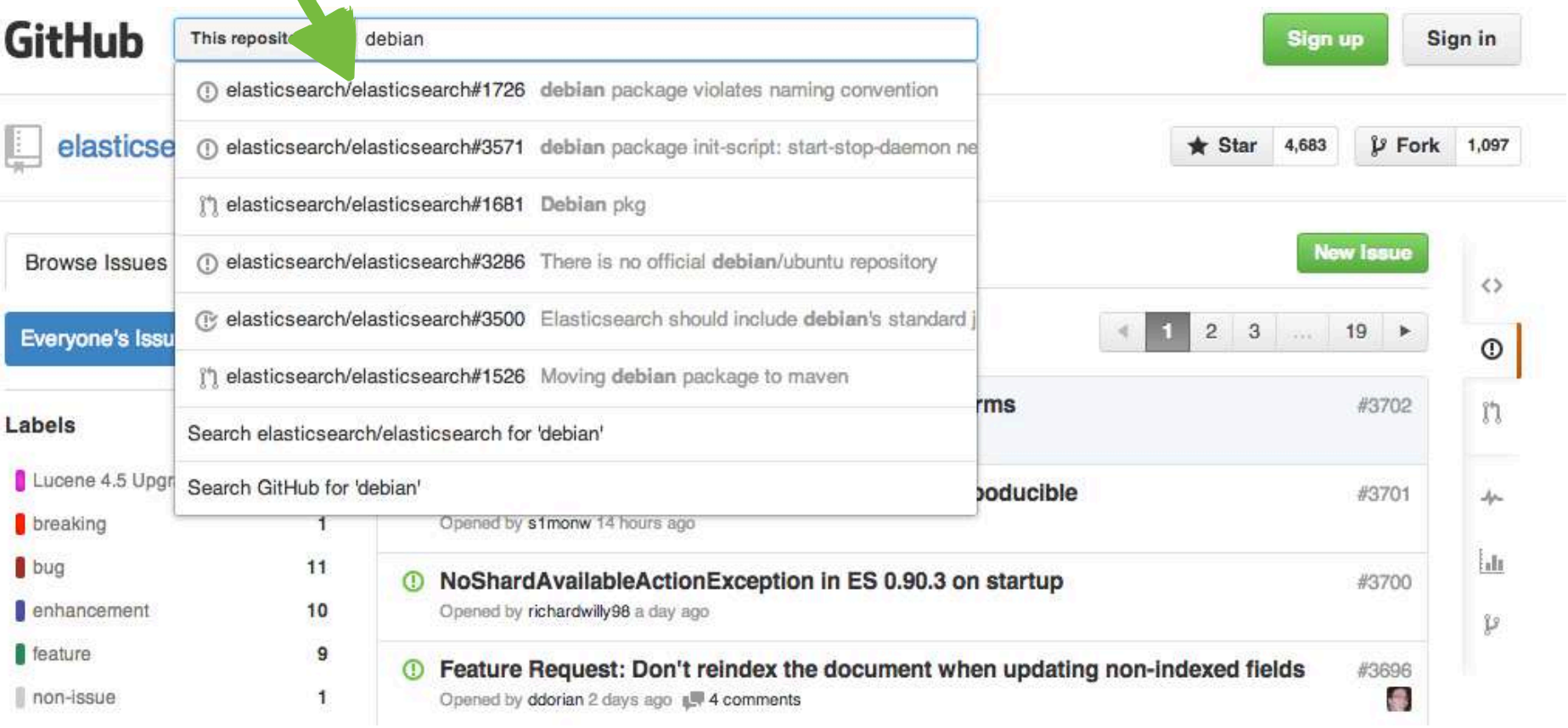
Pagination



Aggregation



Suggestions



Introduction

Elasticsearch in 10 seconds

- Schema-free, REST & JSON based distributed document store
- Open source: Apache License 2.0
- Zero configuration

- Used by github, mozilla, soundcloud, stack overflow, foursquare, fog creek, stumbleupon

Zero configuration

```
$ wget https://download.elasticsearch.org/...
$ tar -xf elasticsearch-0.90.5.tar.gz
$ ./elasticsearch-0.90.5/bin/elasticsearch -f
... [INFO ][node][Ghost Maker] {0.90.5}[5645]: initializing ...
```

Index & search data

```
curl -X PUT localhost:9200/products/product/1 -d '
{
  "created_at" : "2013/09/05 15:45:10",
  "name" : "Macbook Air",
  "price" : {
    "net" : 1699,
    "tax" : 322.81,
  }
}'
```

```
curl -X GET 'localhost:9200/products/product/_search?q=macbook'
```

Distributed

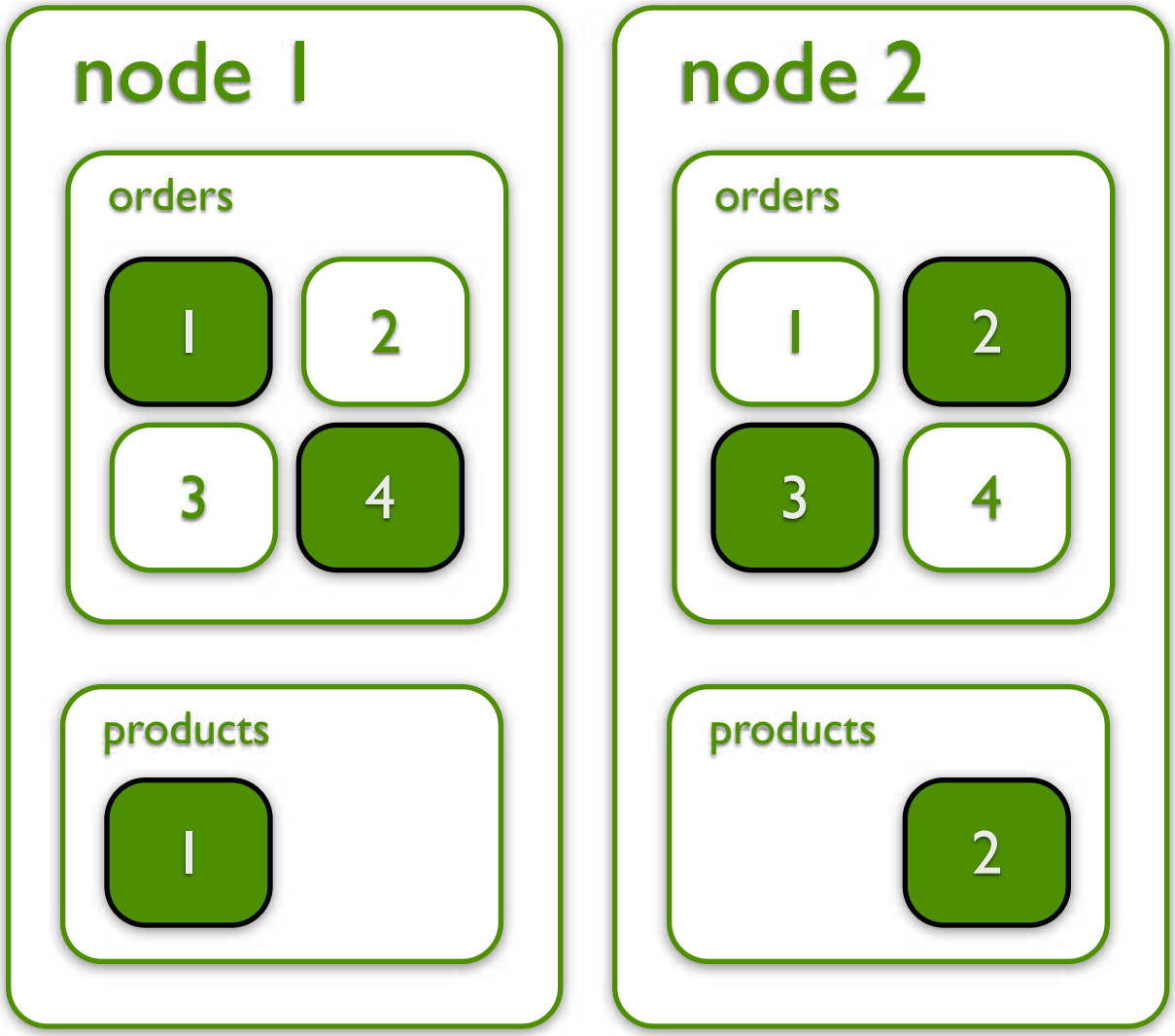
- Replication: Data duplication
  - Read scalability
  - Removing SPOF
- Sharding: Data partitioning
  - Split logical data over several machines
  - Write scalability
  - Control data flows

Distributed

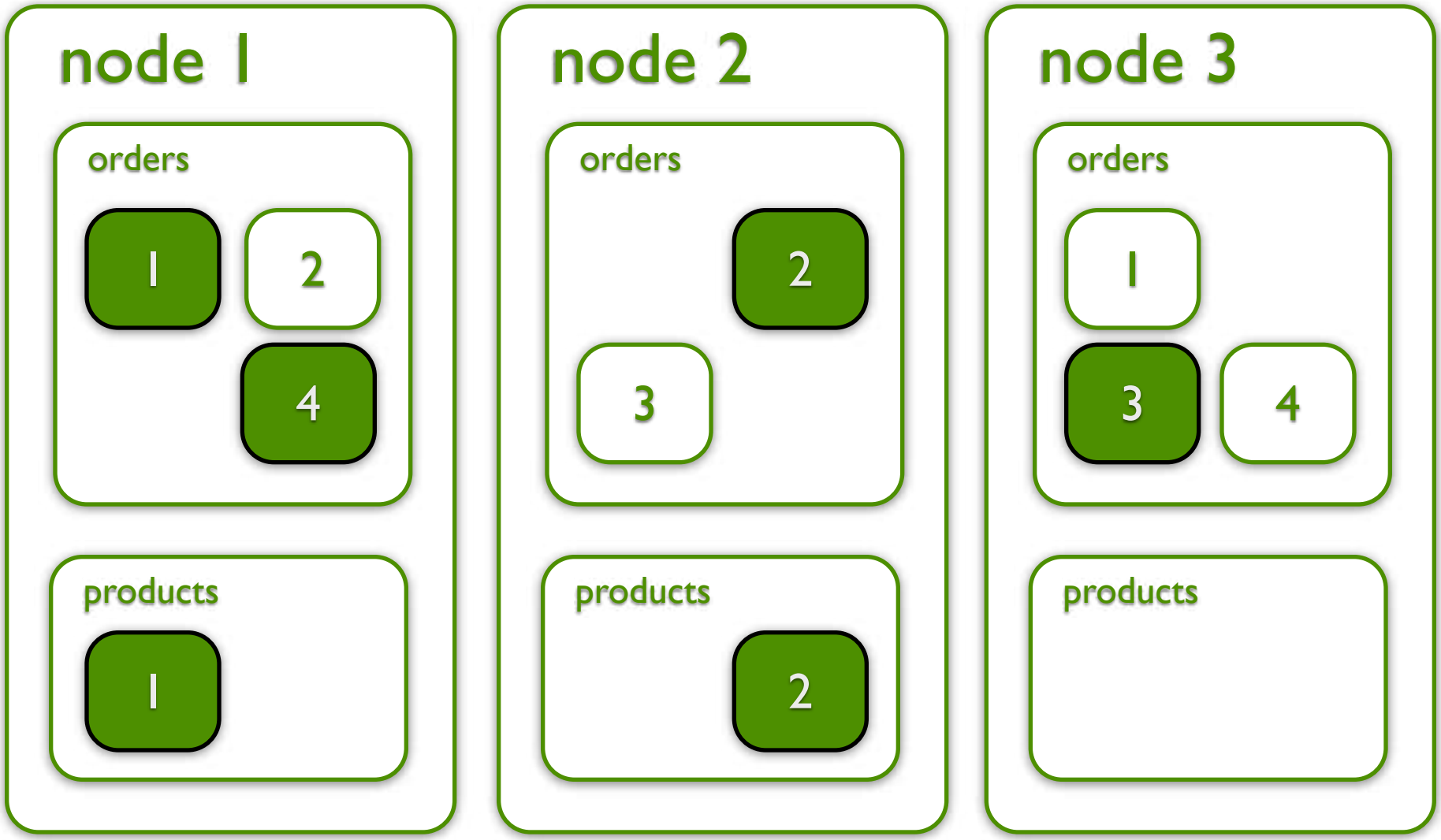

```
curl -X PUT localhost:9200/orders -d '{
  "settings.index.number_of_shards" : 4
  "settings.index.number_of_replicas" : 1
}'
```

```
curl -X PUT localhost:9200/products -d '{
  "settings.index.number_of_shards" : 2
  "settings.index.number_of_replicas" : 0
}'
```

Distributed



Distributed



Ecosystem

- Plugins
- Clients for many languages

  Ruby, Python, PHP, Perl

  Javascript, Scala, Clojure
- Kibana & Logstash
- Hadoop integration

From data to information

What is data?

- Whatever provides value for your business

- Domain data

  Internal: Orders, products

  External: Social media streams, email

- Application data

  Log files

  Metrics

Asking questions to your data

- How many orders were created?
- How many orders were created in the last month?
- How many orders were created every day in the last month?
- What is the average revenue per shopping cart?
- What is the average shopping cart size per order (EUR or #items)? Per hour?

Order as JSON

```
curl -X PUT localhost:9200/orders/order/1 -d '
{
  "created_at" : "2013/09/05 15:45:10",
  "items" : [
    ...
  ]
  "total" : 245.37
}'
```

Asking questions to your data

- How many orders were created? count
- How many orders were created in the last month?

```
curl -X GET http://localhost:9200/orders/order/_count
```

- What is the average revenue per shopping cart?
- What is the average shopping cart size per order (EUR or #items)? Per hour?

Asking questions to your data

- How many orders were created?
- How many orders were created in the last month?
- How many orders were created every day in the last month?


```
curl -X GET http://localhost:9200/orders/order/_count -d '{
  "range": {
    "created_at": {
      "gte": "2013/09/01",
      "lt":  "2013/10/01"
    }
  }
}'
```

Asking questions to your data

- How many orders were created?
- How many orders were created in the last month?
- How many orders were created every day in the last month?
- What is the average revenue per shopping cart?
- What is the average shopping cart size per order (EUR or #items)? Per hour?

Asking questions to your data

```
curl -X GET http://localhost:9200/orders/order/_search -d '{
  "facets": {
    "created": {
      "date_histogram" : {
        "field" : "created_at",
        "interval" : "1d"
      },
      "facet_filter" : {
        "range": {
          "created_at": {
            "gte": "2013/09/01",
            "lt" : "2013/10/01"
          }
        }
      }
    }
  }
}'
```

count/day

filter

Asking questions to your data

- How many orders were created?
- How many orders were created in the last month?
- How many orders were created every day in the last month?
- What is the average revenue per shopping cart?
- What is the average shopping cart size per order (EUR or #items)? Per hour?

filter

scripting

stats

Asking questions to your data

```
curl -X GET http://localhost:9200/orders/order/_search -d '{
  "facets": {
    "avg_revenue": {
      "facet_filter" : {
        "range": {
          "created_at": {
            "gte": "2013/09/01",
            "lt" : "2013/10/01"
          }
        }
      },
      "statistical" : {
        "script" : "doc[\u0027total\u0027].value * 0.1 + 2"
      }
    }
  }
}'
```

filter

scripting

stats

Asking questions to your data

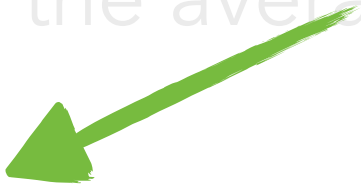
- How many orders were created?
- How many orders were created in the last month?
- How many orders were created every day in the last month?
- What is the average revenue per shopping cart?
- What is the average shopping cart size per order (EUR or #items)? Per hour?

filter

scripting

stats

per hour

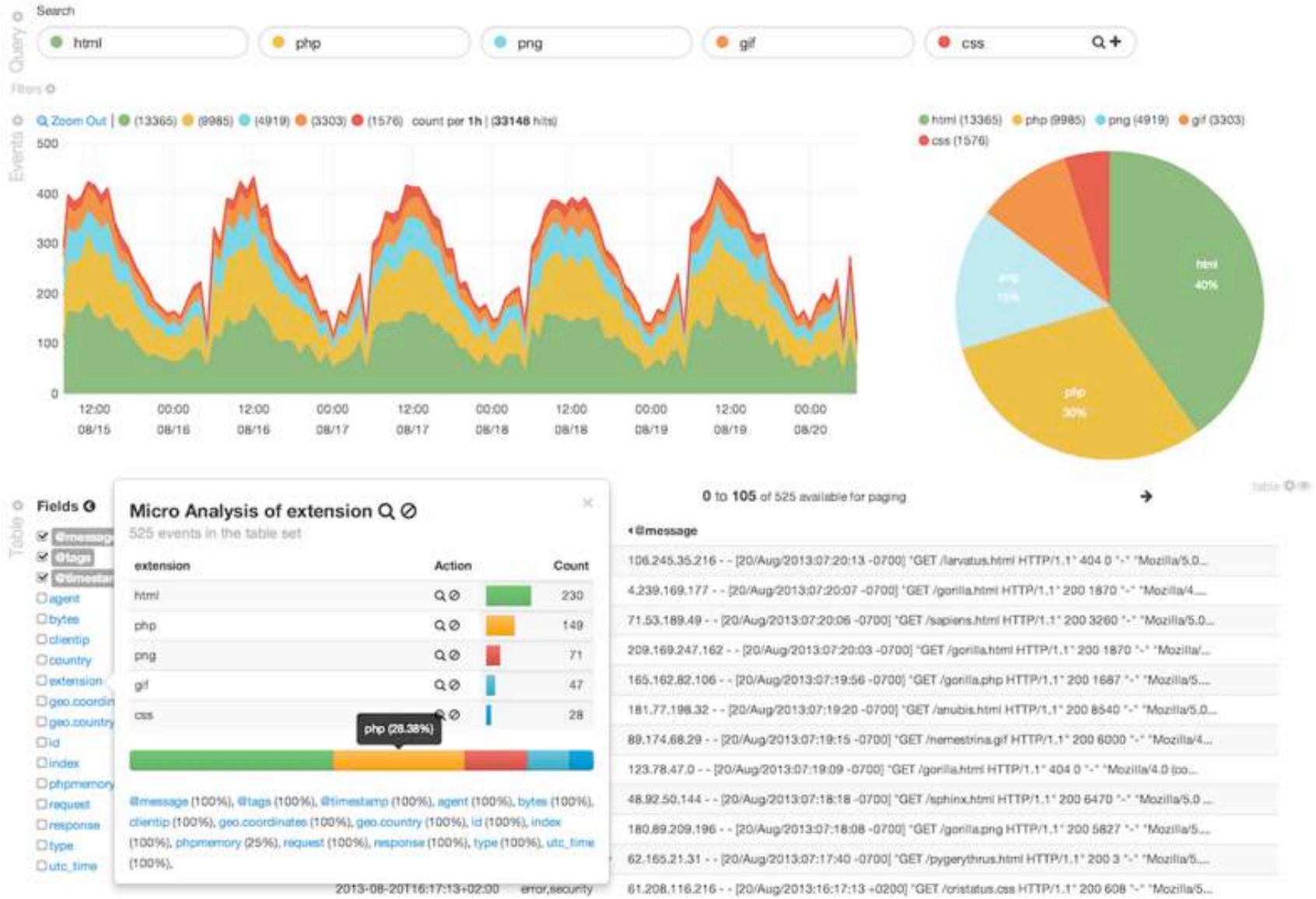


From data to visualization

From numbers to simplicity

- JSON is not a management compatible notation
- Writing your own visulization app for all the different data is tedious
- Enter Kibana!

Kibana

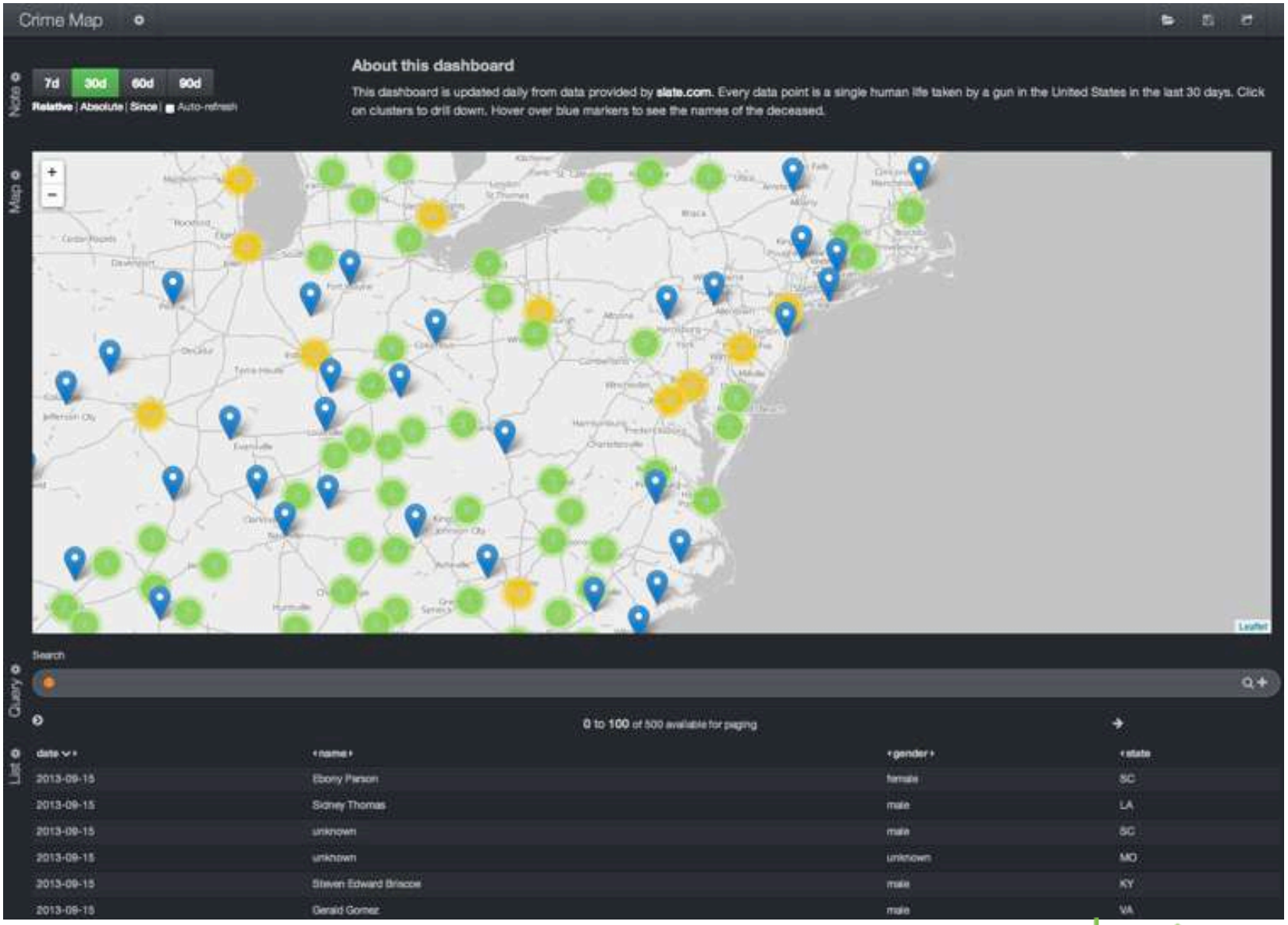


elasticsearch.

Kibana

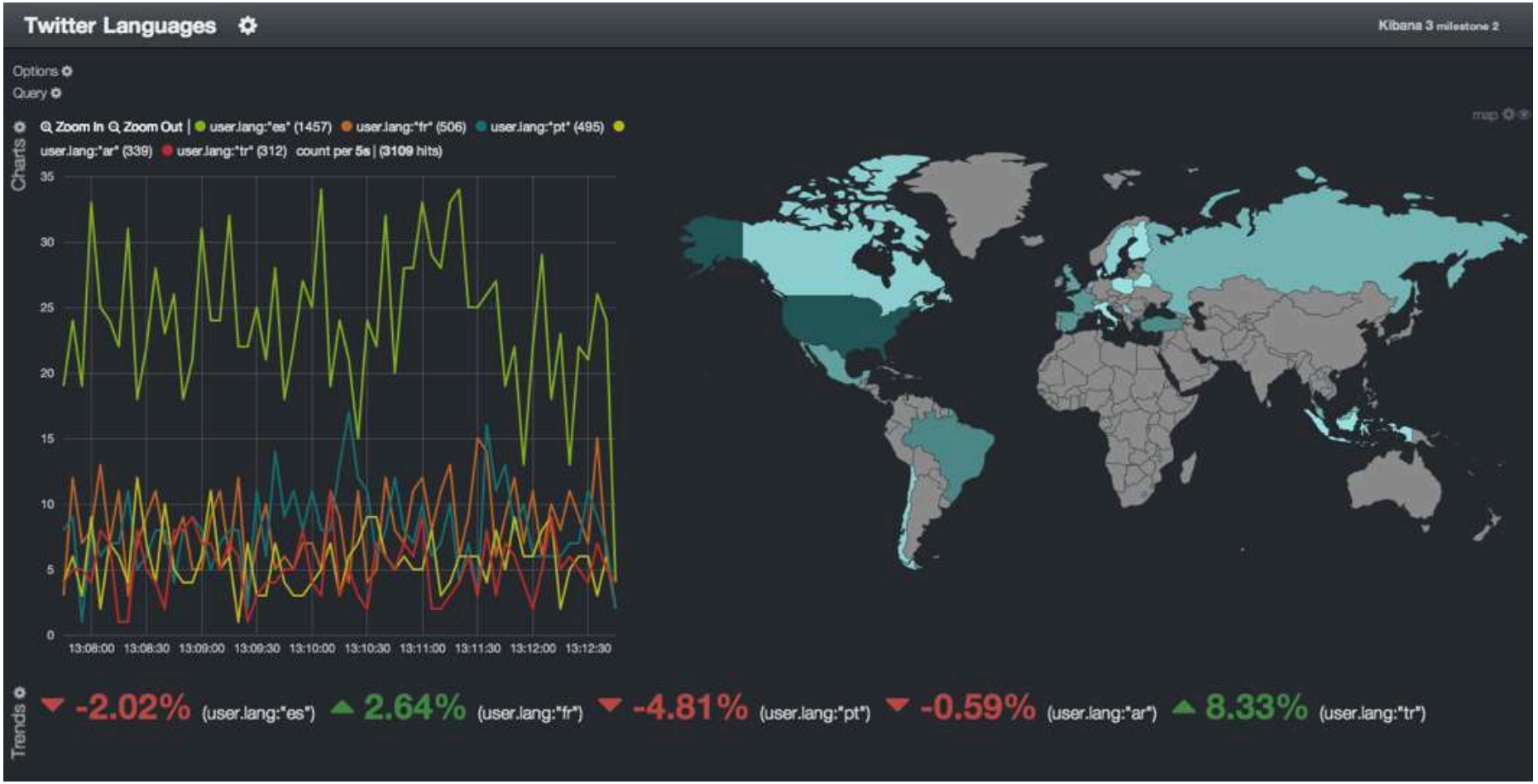
elasticsearch.

Kibana



elasticsearch.

Kibana



elasticsearch.

From data
to notification

Houston, we have a problem!

- The average response time of your payment API just increased over 2 seconds over the last 15 minutes
- A credit card fraud detection kicks in
- Visits are exploding after the television commercial
- The “win-a-car” voucher has reached its usage limit
- Memory usage exceeds physical memory

Meet the metrics library!

- Measure inside your application

- Gauges, Timers, Counters, Meters, Histograms
- Healthchecks

- Report to elasticsearch

Meet the metrics library!

```
MetricRegistry metrics = new MetricRegistry();
```

```
Meter requestsMeter = metrics.meter("incoming-http-requests");

// in your app code
requestsMeter.mark(1);
```

```
Timer responses = metrics.timer("responses"));

Timer.Context context = responses.time();
try {
    // etc;
    return "OK";
} finally {
    context.stop();
}
```

Metrics elasticsearch reporter

- Reports from your application into elasticsearch
- Uses HTTP, no elasticsearch dependency
- Realtime notification via percolation

  Sent an email, a pager alert or a MQ message

Percolation

- Normal: Index documents, run queries
- Percolator: Register queries, run against documents

- Use-case: Price agent, contextual ads, classification before indexing (geo, tag, categorization), metrics

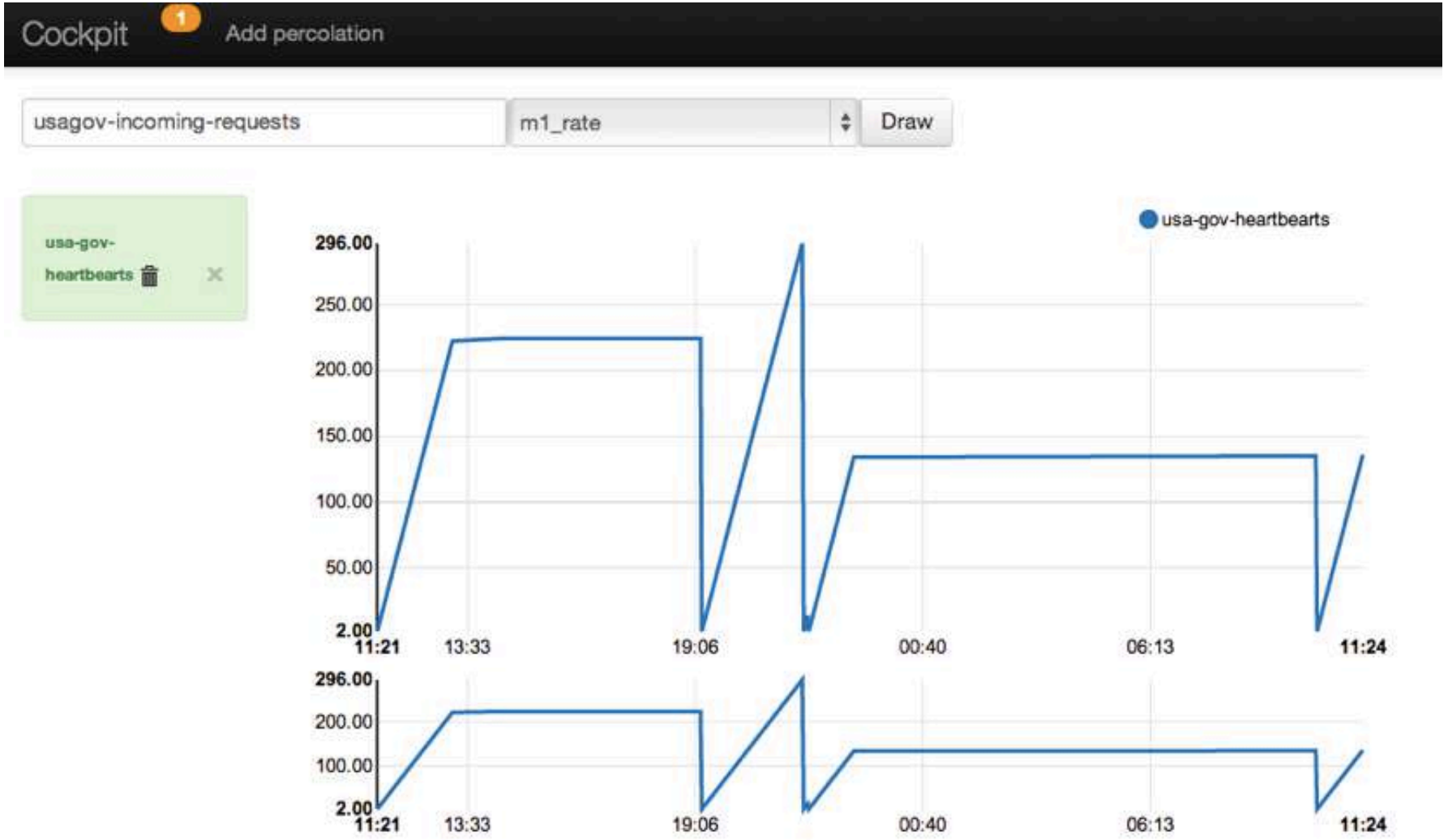
Percolation support

```
ElasticsearchReporter reporter =
    ElasticsearchReporter.forRegistry(registry)
        .percolateNotifier(new PagerNotifier())
        .percolateMetrics(".*")
        .build();
reporter.start(60, TimeUnit.SECONDS);
```

```
public class PagerNotifier implements Notifier {

  @Override
  public void notify(JsonMetric metric, String id) {
    // send pager duty here
  }
}
```

Cockpit - Sample App



elasticsearch.

From data to insight

Know it all!

- Long term data required (index everything!)
- Visualization is a great start
- Deep insight into your data required

  Know your data

  Know your data format

  Concrete questions with lots of dimensions

Aggegrations

- aka: composable facets
- Take the output of a facet operation
- Use it as an input of another facet operation

- Remember: What is the average shopping cart value per order per hour?

Aggegrations

```
curl -X GET 'http://localhost:9200/orders/order/_search' -d '{
"aggs" : {
    "avg_shopping_cart_per_hour" : {
      "filter" : {
        "range": {
          "created_at": {
            "gte": "2013/09/01",
            "lt" : "2013/10/01"
          }
        }
      },
      "date_histogram" : {
        "field" : "created_at",
        "interval" : "1h"
      },
      "aggregations" : {
        "avg" : { "avg" : { "field" : "total" } }
      }
} } }'
```

Aggegrations

```
curl -X GET 'http://localhost:9200/orders/order/_search' -d '{
"aggs" : {
    "avg_shopping_cart_per_hour" : {
      "filter" : {
        "range": {
          "created_at": {
            "gte": "2013/09/01",
            "lt" : "2013/10/01"
          }
        }
      },
      "histogram" : {
        "script" : "doc[\u0027created_at\u0027].date.hourOfDay",
      },
      "aggregations" : {
        "avg" : { "avg" : { "field" : "total" } }
      }
} } }'
```

Ask complex questions

- Product pageviews

  Sum of page views per price range including price statistics (min/max/avg/sum/count)

- Geo location

  Physical store: Home of buyers per weekday combined with money spent

- Protip: Reduce memory consumption using probalistic data structures, losing precision

roundup

Roundup

Insight

Visualization

Notification

Thanks for listening!

We're hiring
http://www.elasticsearch.com/about/jobs

#gotoaar #elasticsearch

Alexander Reelsen
@spinscale
alexander.reelsen@elasticsearch.com

roadmap

Roadmap

- Elasticsearch 1.0

  Distributed percolator (already in master)

  Aggregations

  Snapshot/Restore

links

Links

- Elasticsearch

  http://www.elasticsearch.org
- Logstash

  http://logstash.net
- Kibana

  http://three.kibana.org
- elasticsearch-metrics-reporter

  https://github.com/elasticsearch/metrics-elasticsearch-reporter-java

Links

- Clients

  http://www.elasticsearch.org/blog/unleash-the-clients-ruby-python-php-perl/
- Metrics

  http://metrics.codahale.com/
- Aggregations

  https://github.com/elasticsearch/elasticsearch/issues/3300
- Elasticsearch Hadoop integration

  https://github.com/elasticsearch/elasticsearch-hadoop

Links

- Talk on probalistic data structures

  http://www.infoq.com/presentations/scalability-data-mining

- Icons

  http://www.doublejdesign.co.uk/

  http://www.iconarchive.com/