Building a Real-time, Big Data Analytics Platform with Solr

Trey Grainger

Manager, Search Technology Development

@ careerbuilder®

Overview

- Intro to Analytics with Solr
- Real-world examples & Faceting deep dive
- Solr enhancements we're contributing:
  - Distributed Pivot Faceting
  - Pivoted Percentile/Stats Faceting
- Data Analytics with Solr… the next frontier

My Background

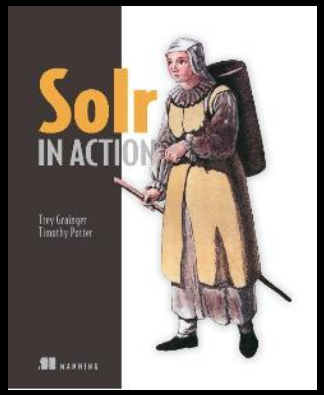
Trey Grainger

Manager, Search Technology Development

@ **CareerBuilder**.com

Relevant Background

- Search & Recommendations
- High-volume, Distributed Systems
- NLP, Relevancy Tuning, User Group Testing, & Machine Learning



Other Projects

- Co-author: *Solr in Action*
- Founder and Chief Engineer @ Celiaccess.com

About Search @ careerbuilder®

- Over 2.5 million new jobs each month
- Over 50 million actively searchable resumes
- ~300 globally distributed search servers (in the U.S., Europe, & Asia)
- Thousands of unique, dynamically generated indexes
- Over ½ Billion actively searchable documents
- Over 1 million searches an hour

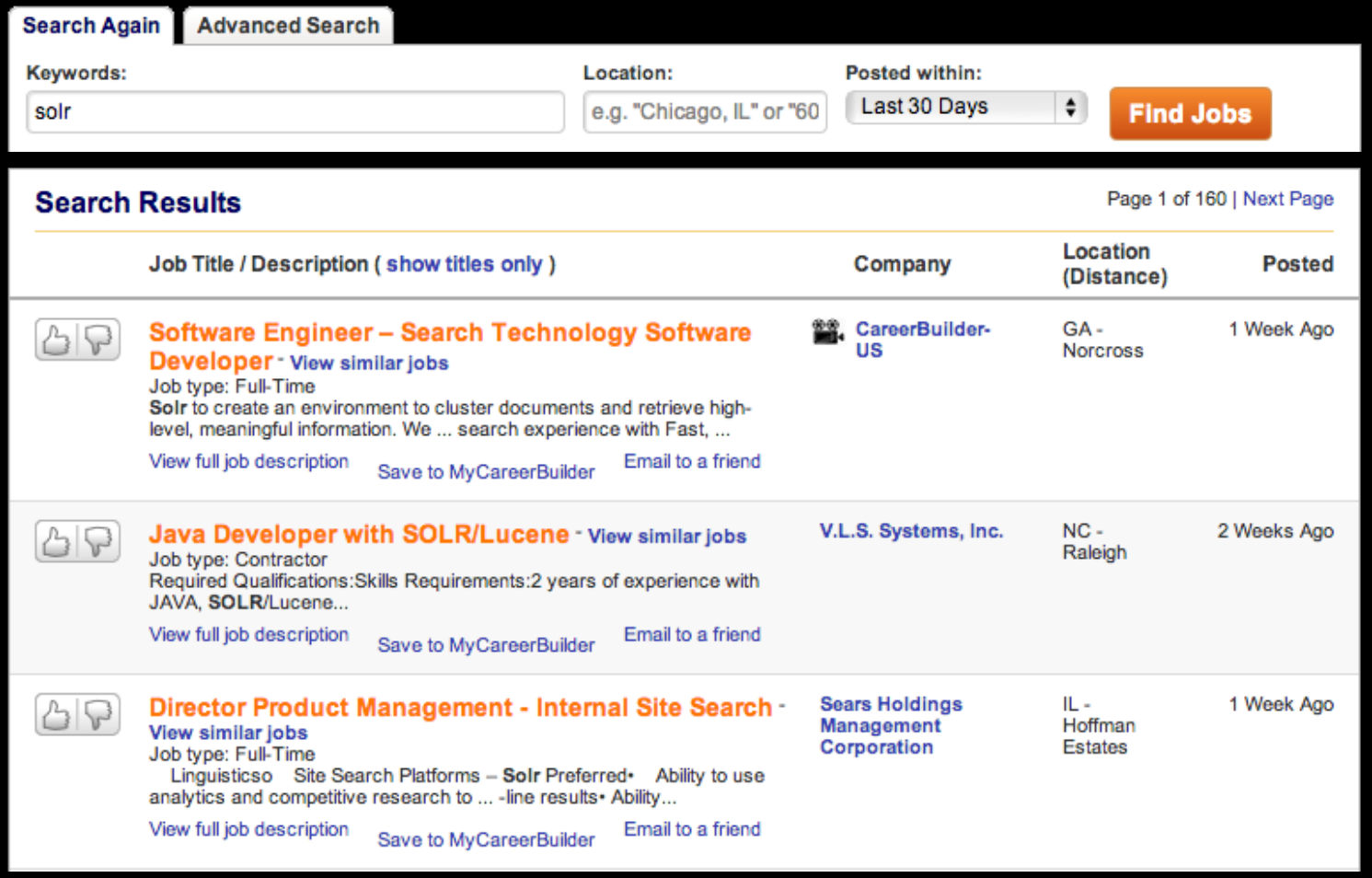
Our Search Platform

- Generic Search API wrapping Solr + our domain stack
- Goal: Abstract away search into a simple API so that any engineer can build search-based product with no prior search background
- 3 Supported Methods (with rich syntax):
  - AddDocument
  - DeleteDocument
  - Search

  **users pass along their own dynamically-defined schemas on each call*
- Building out as Restful API so search can be used "from anywhere"

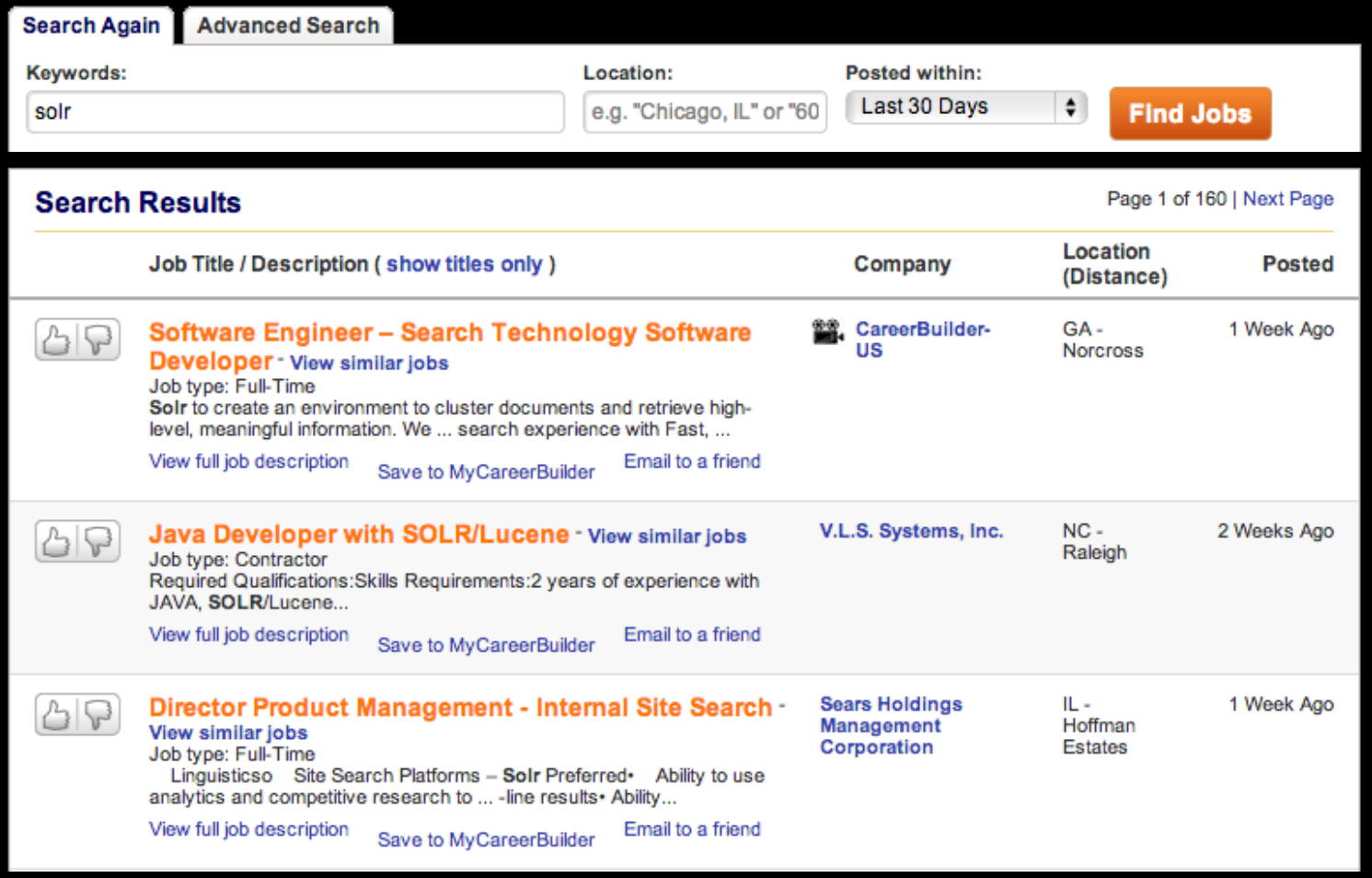
Paradigm shift?

- Originally, search engines were designed to return a ranked list of relevant documents



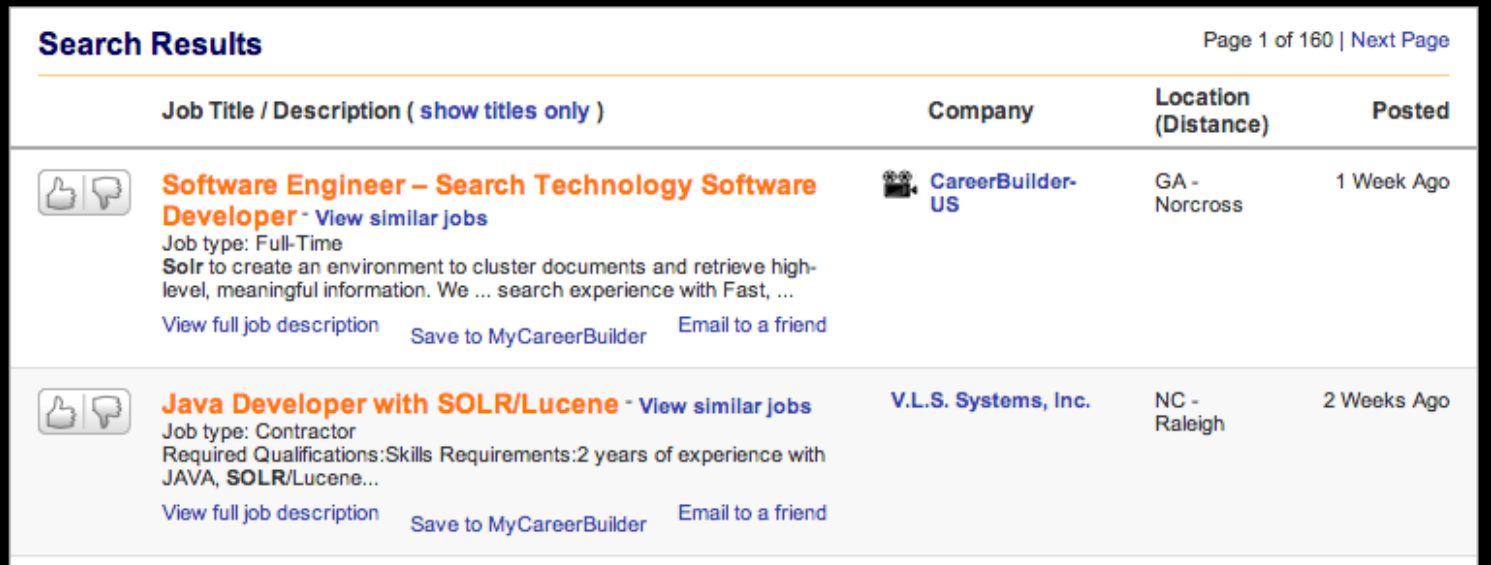
Paradigm shift?

- Then, features like faceting were added to augment the search experience with aggregate information…

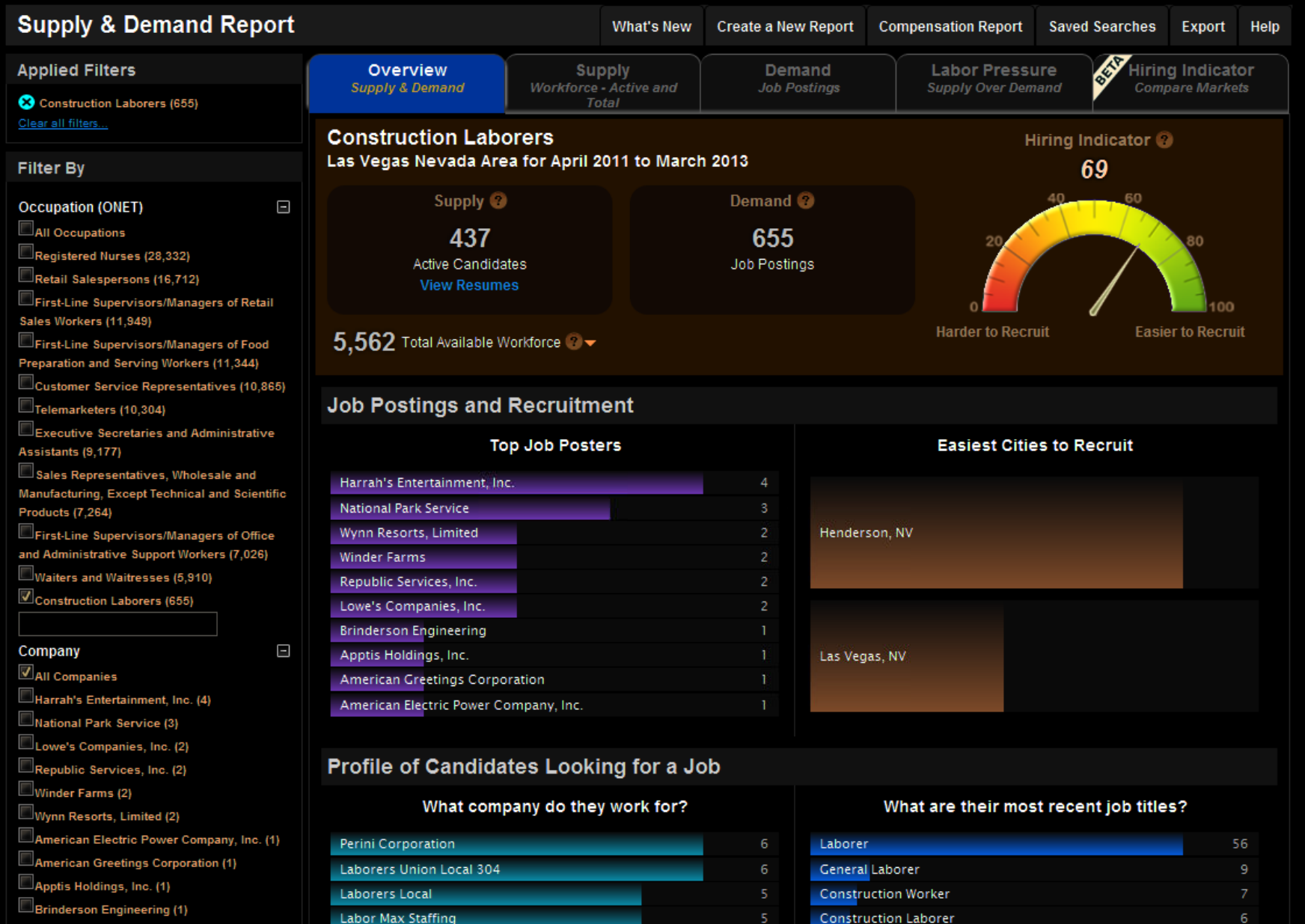


Paradigm shift?

- But… what kinds of products could you build if you only cared about the aggregate calculations?



Workforce Supply & Demand



Faceting Overview

/solr/select/?q=...&facet=true

//Field Faceting
&facet.field=city

```
"facet_fields":{
   "city":[
   "new york, ny",2337,
   "los angeles, ca",1693,
   "chicago, il",1535,
   ... ]}
```

//Query Faceting:
&facet.query={!frange key="0 to 10 km" l=0 u=10 incll=false}geodist()
&facet.query={!frange key="10 to 25 km" l=10 u=25 incll=false}geodist()
&facet.query={!frange key="25 to 50 km" l=25 u=50 incll=false}geodist()
&facet.query={!frange key="50+" l=50 incll=false}geodist()
&sfield=location
&pt=37.7770,-122.4200

```
"facet_queries":{
   "0 to 10 km":1187,
   "10 to 25 km":462,
   "25 to 50 km":794,
   "50+":105296
},
```

//Range Faceting
&facet.range=years_experience
&facet.range.start=0
&facet.range.end=10
&facet.range.gap=1
&facet.range.other=after

```
"facet_ranges":{
    "years_experience":{
     "counts":[
      "0",1010035,
      "1",343831,
      ...
      "9",121090
     ], ...
     "after":59462}}
```

Multi-select Faceting with Tags / Excludes

Faceting with No Filters Selected:

▶State	▶Price Range	▶City
☐ Georgia (6)	☐ < $5 (6)	☐ Atlanta, GA (6)
☐ California (4)	☐ $5 - $10 (5)	☐ New York, NY (4)
☐ New York (4)	☐ $10 - $20 (3)	☐ Austin, TX (3)
☐ Texas (3)	☐ $20 - $50 (6)	☐ San Francisco (3)
☐ Illinois (2)	☐ $50+ (0)	☐ Chicago (2)
...		...

```
&facet=true
&facet.field=state
&facet.field=price_range
&facet.field=city
```

Faceting with Filter of state:California selected:

▶State	▶Price Range	▶City
☑ **California (4)**	☐ < $5 (0)	☐ San Francisco (3)
	☐ $5 - $10 (2)	☐ Los Angeles (1)
	☐ $10 - $20 (0)	
	☐ $20 - $50 (2)	
	☐ $50+ (0)	

```
&facet=true
&facet.field=state
&facet.field=price_range
&facet.field=city
&fq=state:California
```

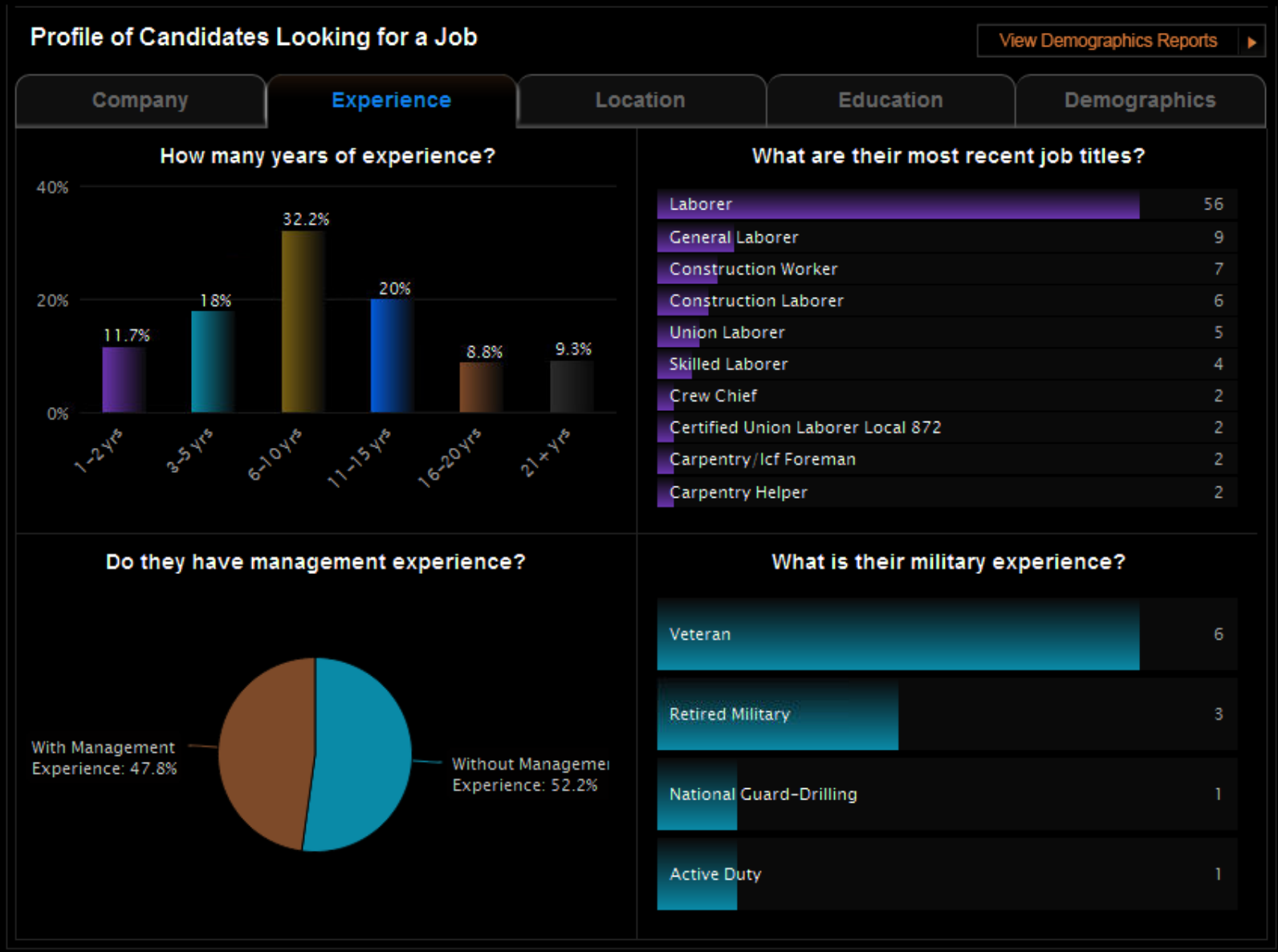
Multi-Select Faceting with state:California selected:

▶State	▶Price Range	▶City
☐ Georgia (6)	☐ < $5 (0)	☐ San Francisco (3)
☑ **California (4)**	☐ $5 - $10 (2)	☐ Los Angeles (1)
☐ New York (4)	☐ $10 - $20 (0)	
☐ Texas (3)	☐ $20 - $50 (2)	
☐ Illinois (2)	☐ $50+ (0)	
...		

```
&facet=true
&facet.field={!ex="stateTag"}state
&facet.field=price_range
&facet.field=city
&fq={!tag="stateTag"}state:California
```

*Slide Taken from Ch 8 of *Solr in Action*

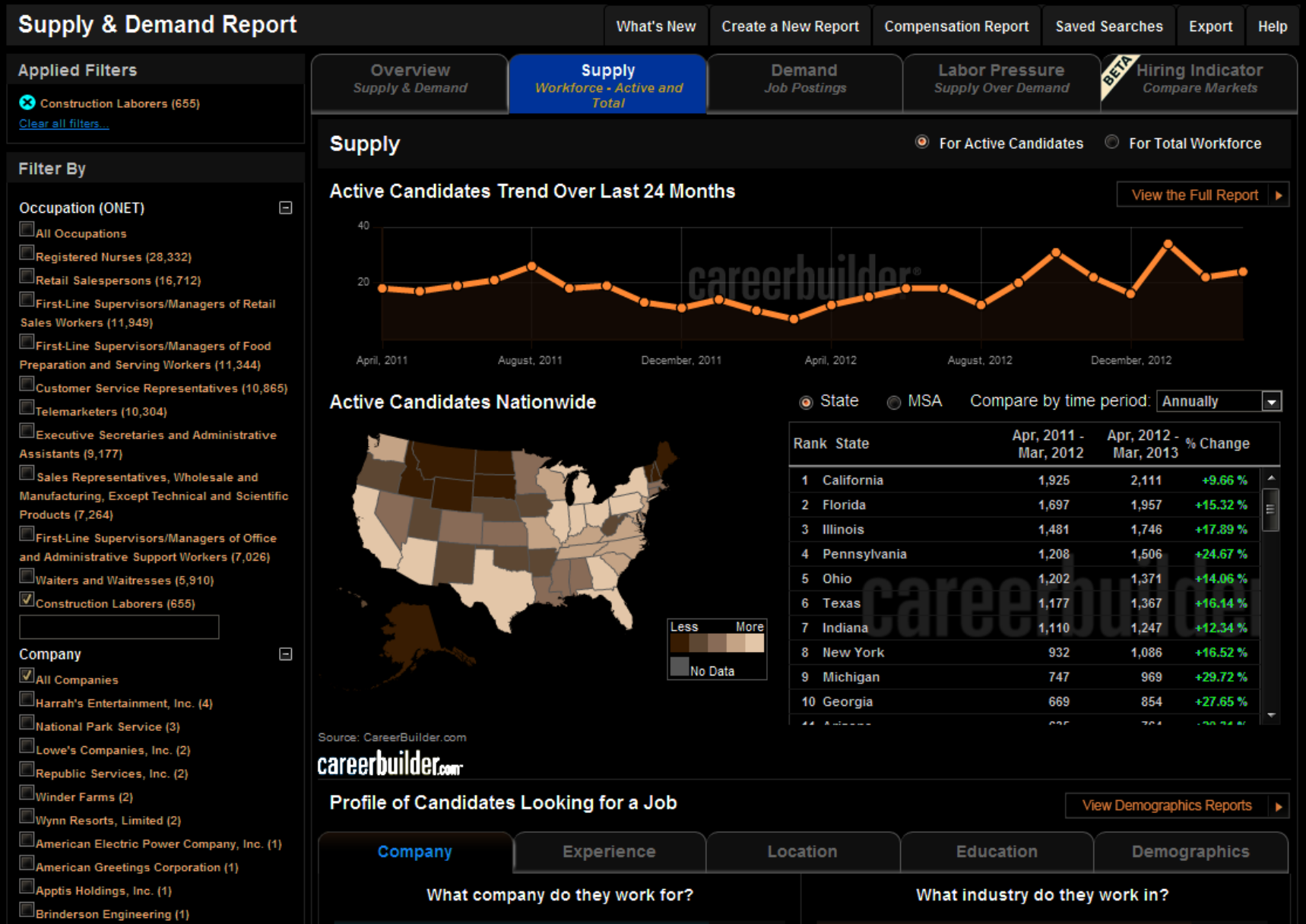
Supply of Candidates



Why Solr for Analytics?

- Allows “ad-hoc” querying of data by keywords
- Is good at on-the-fly aggregate calculations (facets + stats + functions + grouping)
- Solr is horizontally scalable, and thus able to handle billions of documents
- Insanely Fast queries, encouraging user exploration

Supply of Candidates



Demand for Jobs

Overview *Supply & Demand* | Supply *Workforce - Active and Total* | **Demand** *Job Postings* | Labor Pressure *Supply Over Demand* | BETA Hiring Indicator *Compare Markets*

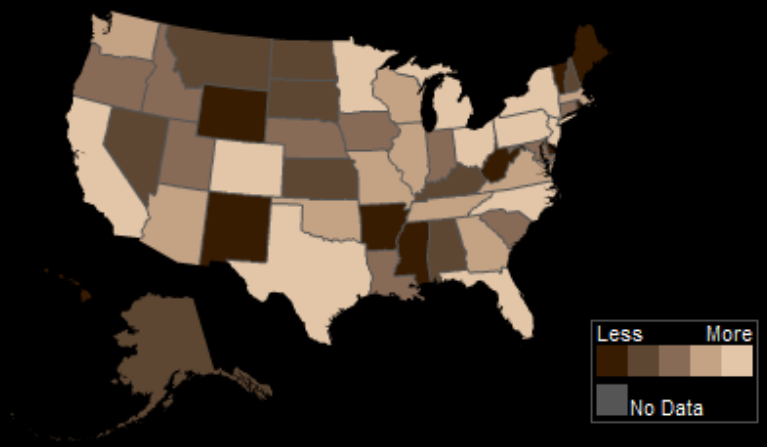
Demand

Job Posting Trend Over Last 24 Months

What is this? View the Full Report

Demand Geography Details

State MSA Compare by time period: Annually



Rank	State	Apr, 2011 - Mar, 2012	Apr, 2012 - Mar, 2013	% Change
1	Texas	5,343	8,009	+49.90 %
2	California	4,743	6,466	+36.33 %
3	Florida	3,965	5,785	+45.90 %
4	Pennsylvania	4,816	4,616	-4.15 %
5	Ohio	4,022	4,517	+12.31 %
6	New York	3,880	4,434	+14.28 %
7	Michigan	2,984	3,637	+21.88 %
8	Colorado	1,961	3,407	+73.74 %
9	North Carolina	2,129	2,790	+31.05 %
10	New Jersey	2,344	2,764	+17.92 %

Sources: CareerBuilder.com, Wanted Technologies

careerbuilder.com WANTED TECHNOLOGIES

Supply over Demand (Labor Pressure)

Overview — Supply & Demand | Supply — Workforce - Active and Total | Demand — Job Postings | **Labor Pressure — Supply Over Demand** | BETA Hiring Indicator — Compare Markets

Labor Pressure

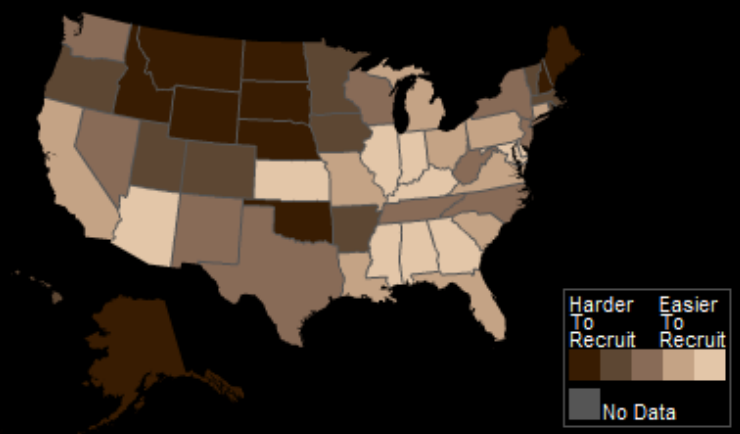
(•) For Active Candidates ( ) For Total Workforce

Trend Over the Last 24 Months

View the Full Report ▸

Active Labor Pressure Geography Details

(•) State ( ) MSA Compare by time period: Annually



Rank	State	Apr, 2011 - Mar, 2012	Apr, 2012 - Mar, 2013	% Change
1	Illinois	1.05	0.92	-12.38 %
2	Delaware	0.94	0.88	-6.38 %
3	Indiana	0.94	0.84	-10.64 %
4	Kentucky	0.59	0.52	-11.86 %
5	Georgia	0.54	0.49	-9.26 %
6	Mississippi	0.56	0.46	-17.86 %
7	Kansas	0.54	0.45	-16.67 %
8	Maryland	0.41	0.44	+7.32 %
9	Arizona	0.51	0.40	-21.57 %
10	Alabama	0.45	0.39	-13.33 %

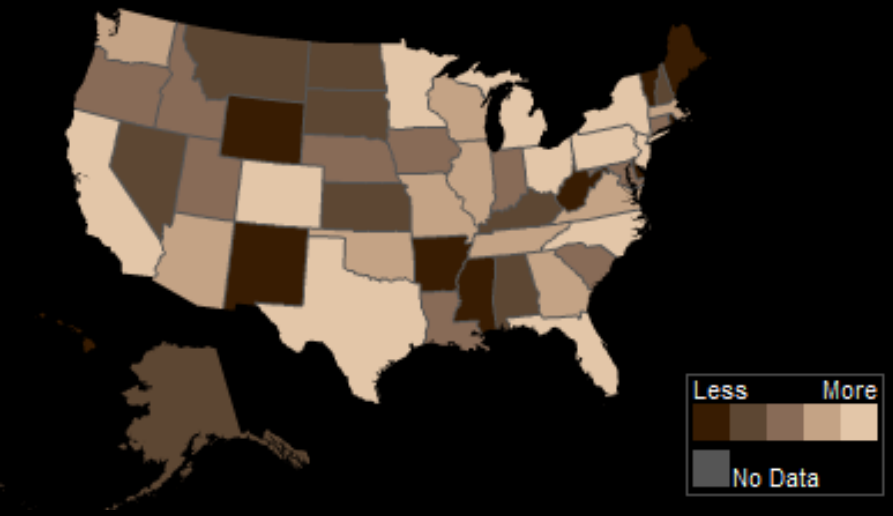
Sources: CareerBuilder.com, Wanted Technologies

careerbuilder.com WANTED TECHNOLOGIES

Wait, how’d you do that?

Building Blocks…

/solr/select/?q=…&facet=true&facet.field=month*



/solr/select/q=...&facet=true&facet.field=state

/solr/select/?q=…&facet=true&
facet.field=military_experience

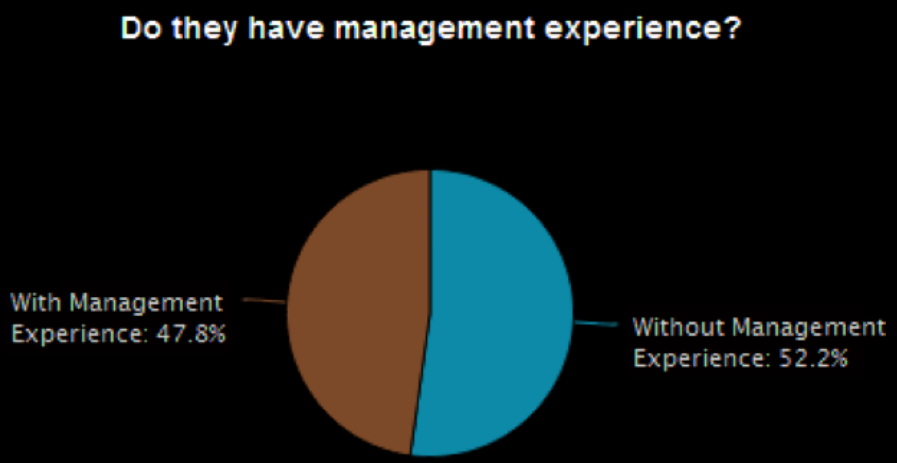
*string field in format 201305

Building Blocks…

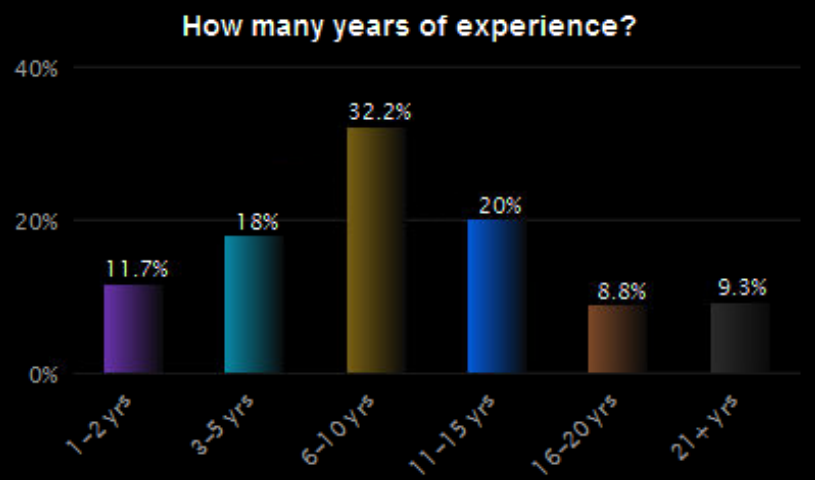
/solr/select/?
q="construction worker"&
fq=city:"las vegas, nv"&
facet=true&
facet.field=company

/solr/select/?
q="construction worker"&
fq=city:"las vegas, nv"&
facet=true&
facet.field=lastjobtitle

Building Blocks…



/solr/select/?q=...&facet=true&
facet.field=management_experience



/solr/select/? q=...&
facet=true&facet.field=experience_ranges

That’s pretty basic… what else can you do?

Radius Faceting

Applied Filters

Food Service Managers (2,253)

Clear all filters...

Filter By

Occupation (ONET)

- [ ] All Occupations
- [ ] Registered Nurses (29,372)
- [ ] Computer Systems Analysts (18,682)
- [ ] Web Developers (17,673)
- [ ] Truck Drivers, Heavy and Tractor-Trailer (17,147)
- [ ] Retail Salespersons (15,279)
- [ ] Accountants (12,949)
- [ ] Sales Managers (12,329)
- [ ] Computer Support Specialists (11,751)
- [ ] Customer Service Representatives (11,595)
- [ ] Executive Secretaries and Administrative Assistants (11,382)
- [x] Food Service Managers (2,253)

Company

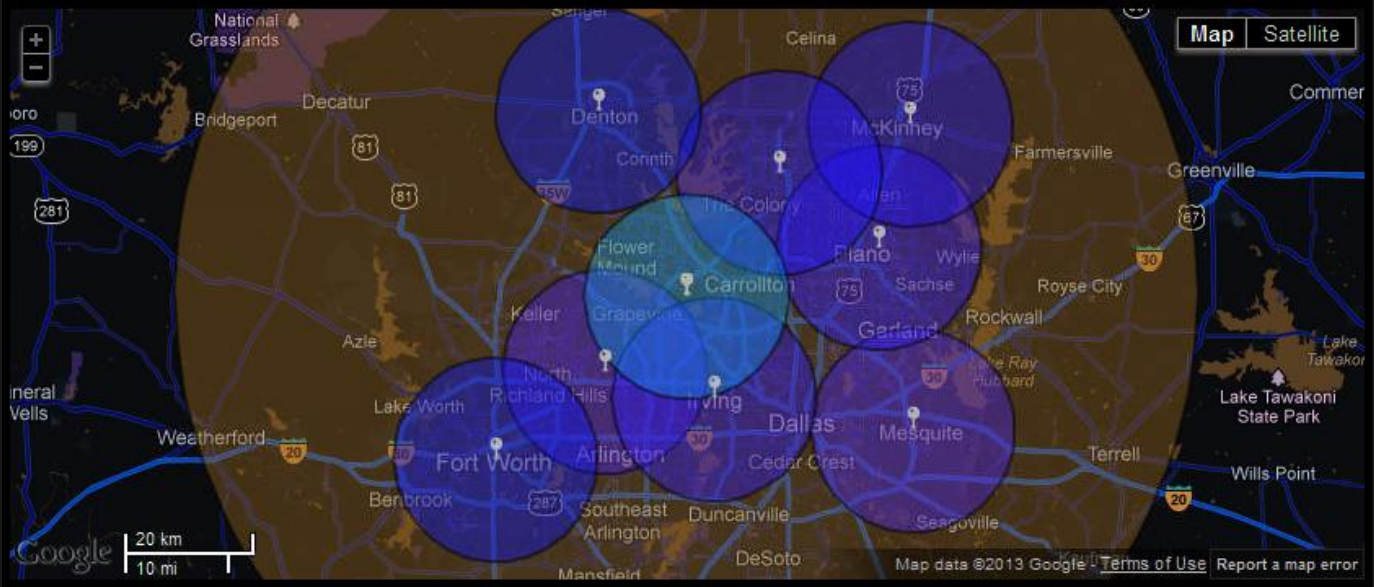
- [x] All Companies
- [ ] Aramark Corporation (73)
- [ ] Compass Group Llc (37)
- [ ] Compass Group International Llc (28)
- [ ] Compass Group Usa, Inc. (28)
- [ ] Compass Group Usa Investments Llp (17)
- [ ] Omni Hotels Corporation (15)
- [ ] Premier Solutions Inc (13)
- [ ] Panda Express, Inc. (12)
- [ ] Culinaire International, Inc. (11)
- [ ] Panda Restaurant Group, Inc. (11)

Industry (NAICS)

Staffing

Overview *Supply & Demand* | Supply *Workforce - Active and Total* | Demand *Job Postings* | Labor Pressure *Supply Over Demand* | BETA Hiring Indicator *Compare Markets*

Nearby Cities with Hiring Indicator Data



Indicator	City	State	Distance (mi)	Active Supply	Demand	Compensation	Apps/Job
72	Mesquite	TX	25.72	168	235	38k 60k	Average
72	Bedford	TX	10.86	112	189	38k 60k	Below Average
70	Irving	TX	10.4	295	500	38k 60k	Average
68	Plano	TX	19.36	372	660	38k 60k	Above Average
64	Mckinney	TX	27.53	184	265	38k 60k	Below Average
63	Frisco	TX	15.06	133	311	38k 60k	Average
55	Fort Worth	TX	24.59	1,354	1,838	38k 60k	Above Average
49	Denton	TX	20.01	182	371	38k 60k	Below Average
18	Dallas	TX	--	1,612	12,445	38k 60k	Above Average
0	Westlake	TX	11.79	0	122	38k 60k	Insufficient Data

Harder to Recruit — Easier to Recruit — Insufficient Comparable Data

Hiring Comparison per Market

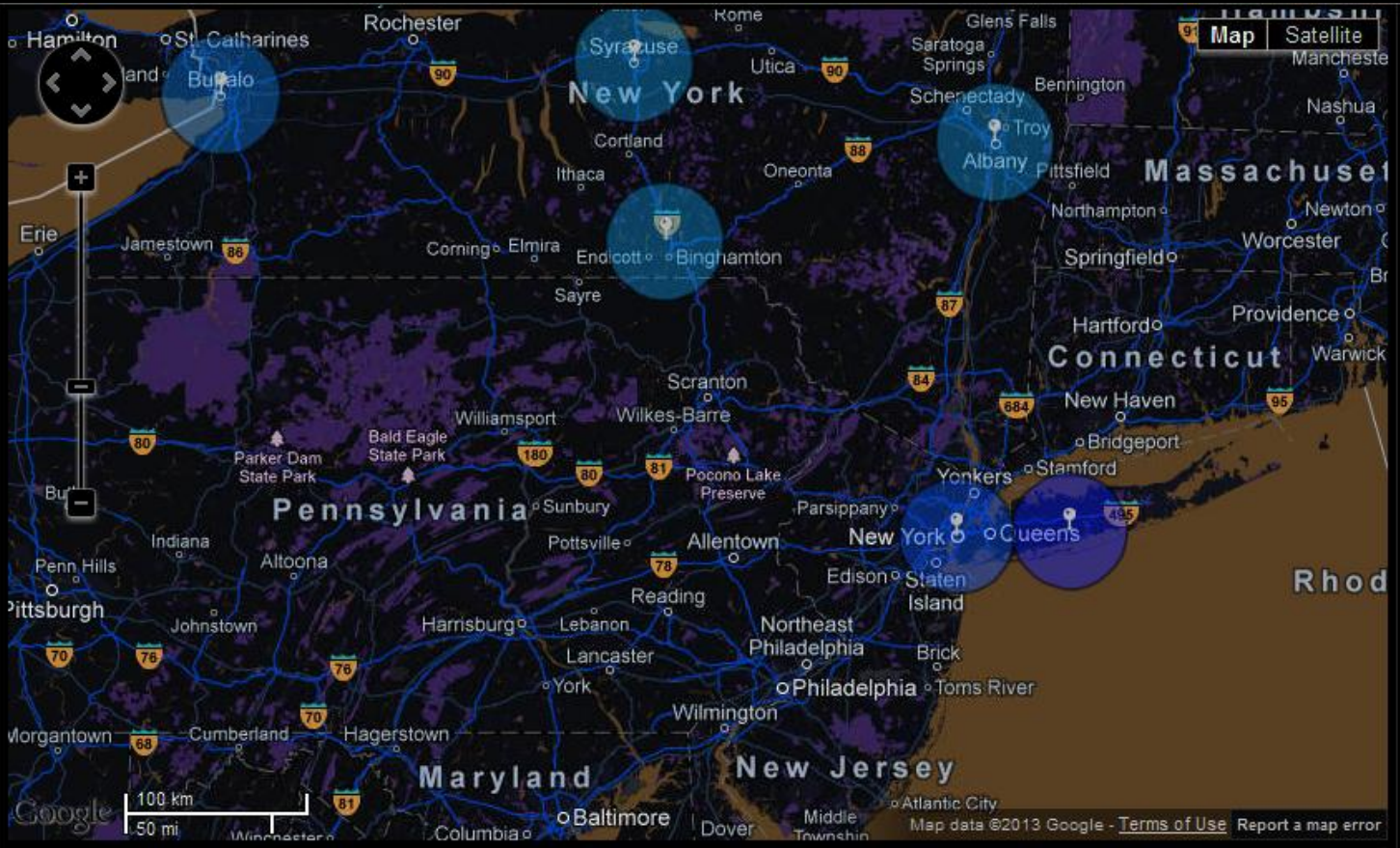
- [x] Registered Nurses (53,112)
- [ ] Medical and Health Services Managers (4,475)
- [ ] Licensed Practical and Licensed Vocational Nurses (3,466)
- [ ] Nursing Aides, Orderlies, and Attendants (1,411)
- [ ] Home Health Aides (985)
- [ ] Physician Assistants (977)
- [ ] Medical Assistants (510)
- [ ] Dietitians and Nutritionists (417)
- [ ] Nursing Instructors and Teachers, Postsecondary (367)
- [ ] Nurse Practitioners (366)
- [x] Advanced Practice Psychiatric Nurses (90)
- [x] Acute Care Nurses (18)
- [x] Critical Care Nurses (14)

Company

- [x] All Companies
- [ ] North Shore- Long Island Jewish Health System (5,194)
- [ ] Lenox Hill Hospital (1,277)
- [ ] Newyork- Presbyterian Hospital (1,060)
- [ ] Visiting Nurse Service Of New York (628)
- [ ] Onward Healthcare (578)
- [ ] Catholic Health System (356)
- [ ] Us Department Of Veterans Affairs (348)
- [ ] Bayada Nurses, Inc (311)
- [ ] Health Quest Management Corporation Vii (296)
- [ ] Lourdes Hospital (294)

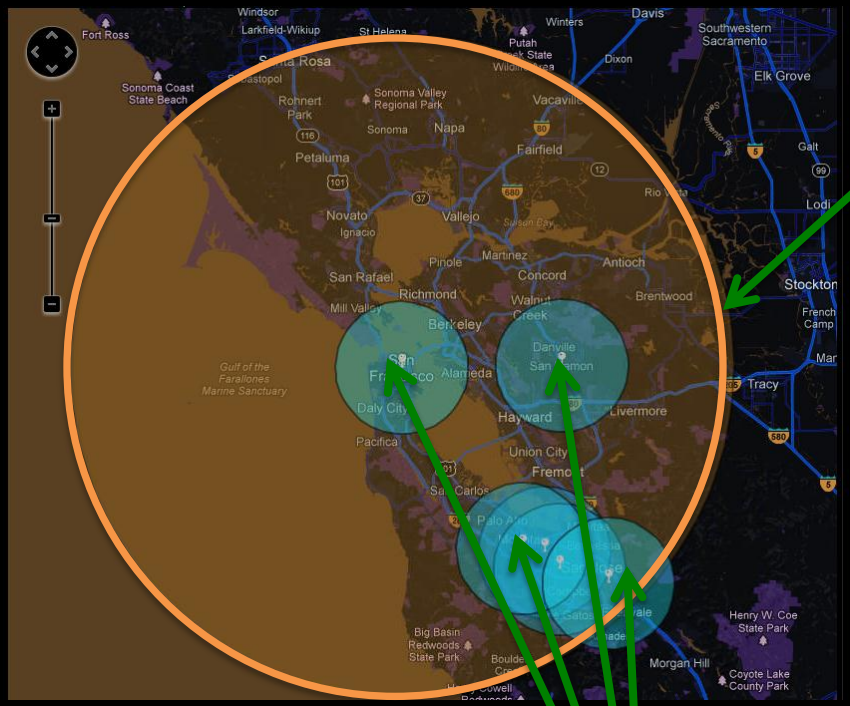
Industry (NAICS)

Staffing



Indicator	City	State	Active Supply	Demand	Compensation	Apps/Job
47	Bay Shore	NY	70	613	55k 91k	Below Average
31	New York	NY	366	22,167	54k 89k	Average
22	Buffalo	NY	43	4,537	47k 67k	Weak
19	Binghamton	NY	37	1,431	42k 65k	Below Average
17	Syracuse	NY	59	3,642	46k 65k	Weak
16	Albany	NY	63	5,599	42k 69k	Weak

Geo-spatial Analytics



Query 1:

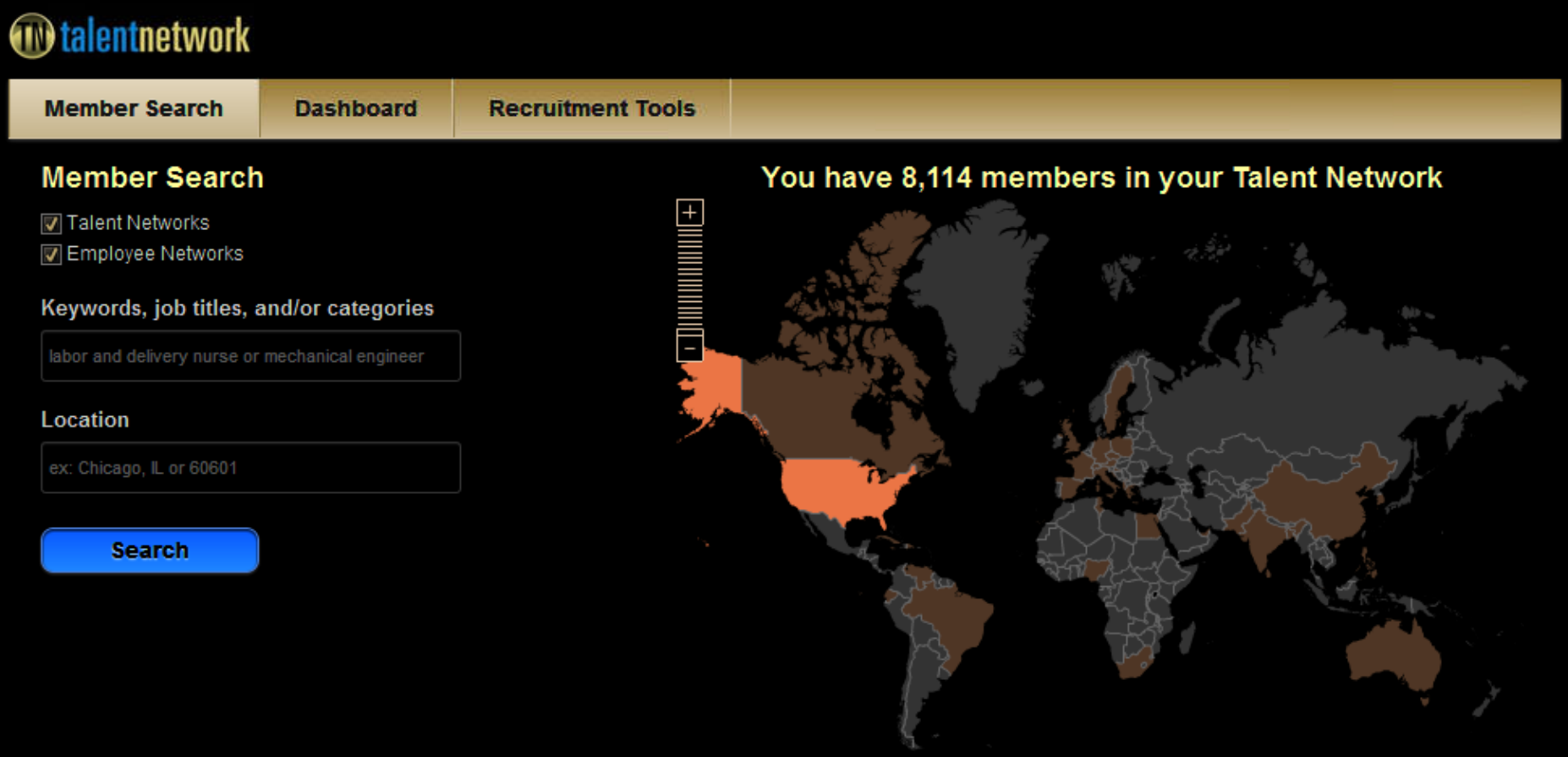
/solr/select/?...
fq={!geofilt sfield=latlong pt=37.777,-122.420 d=80}
&facet=true&facet.field=city&

```
"facet_fields":{
   "city":[
    "san francisco, ca",11713,
    "san jose, ca",3071,
    "oakland, ca",1482,
    "palo alto, ca",1318,
    "santa clara, ca",1212,
    "mountain view, ca",1045,
    "sunnyvale, ca",1004,
    "fremont, ca",726,
    "redwood city, ca",633,
    "berkeley, ca",599]}
```

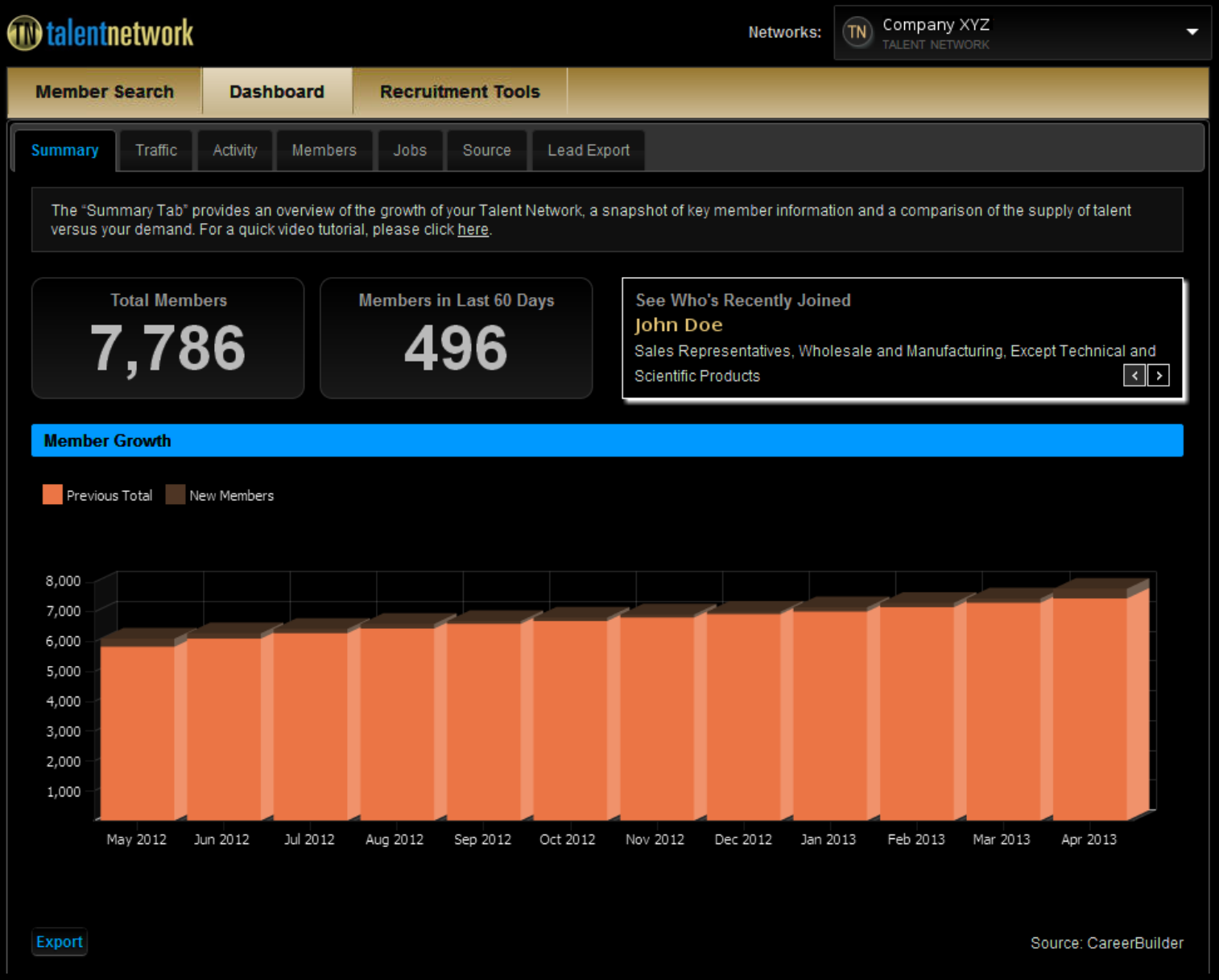
Query 2:

```
/solr/select/?...
&facet=true&facet.field=city&
fq=( _query_:"{!geofilt sfield=latlong pt=37.7770,-122.4200 d=20} "  //san francisco
      OR _query_:"{!geofilt sfield=latlong pt=37.338,-121.886 d=20} " //san jose
       ...
      OR _query_:"{!geofilt sfield=latlong pt=37.870,--122.271 d=20} " //berkeley
     )
```

Customer-specific Analytics



Customer-specific Analytics



Customer-specific Analytics

talentnetwork

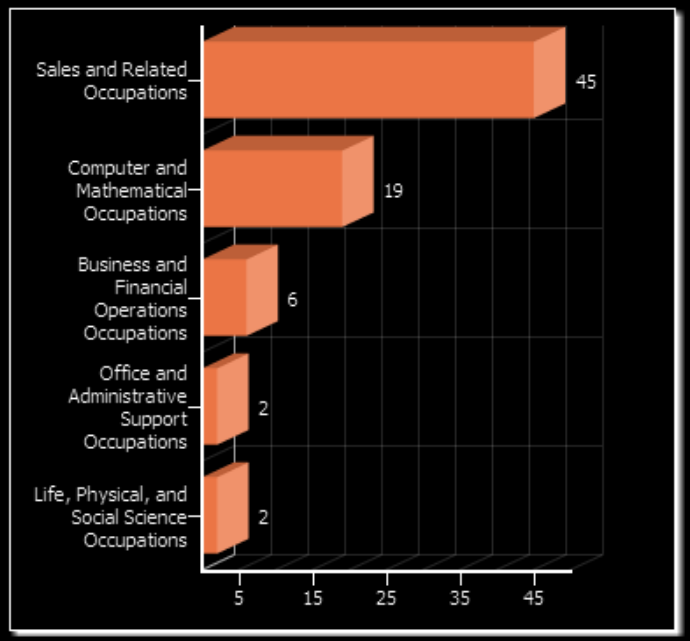
Networks: TN Company XYZ TALENT NETWORK

Member Search | Dashboard | Recruitment Tools

Summary | Traffic | Activity | Members | Jobs | Source | Lead Export

The "Jobs Tab" allows you to analyze the types and locations of jobs in your Talent Network. For a quick video tutorial, please click here.

Jobs By Category

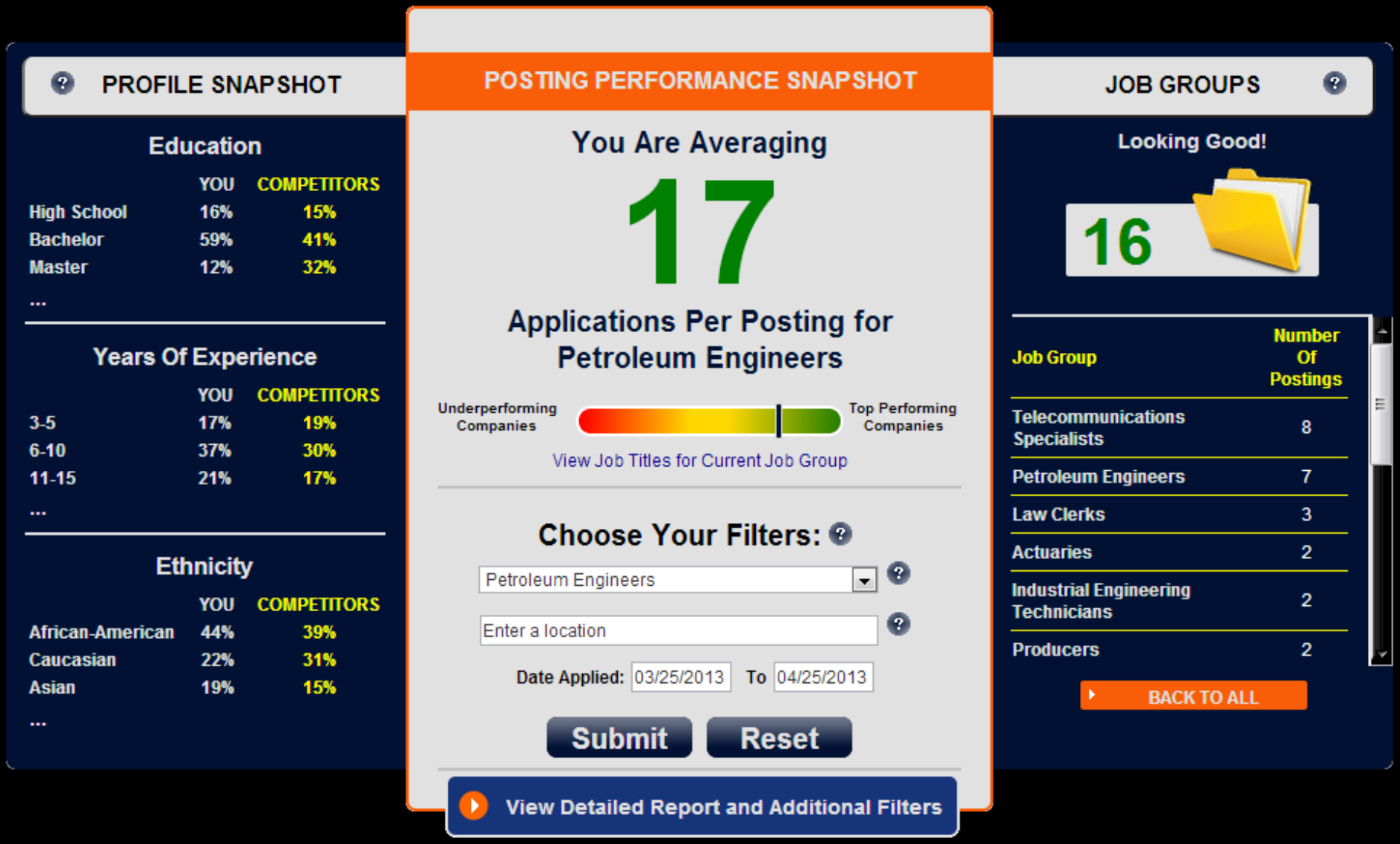


Jobs by Category	Total Active Jobs
Sales and Related Occupations	45
Computer and Mathematical Occupations	19
Business and Financial Operations Occupations	6
Office and Administrative Support Occupations	2
Life, Physical, and Social Science Occupations	2
Arts, Design, Entertainment, Sports, and Media Occupations	1

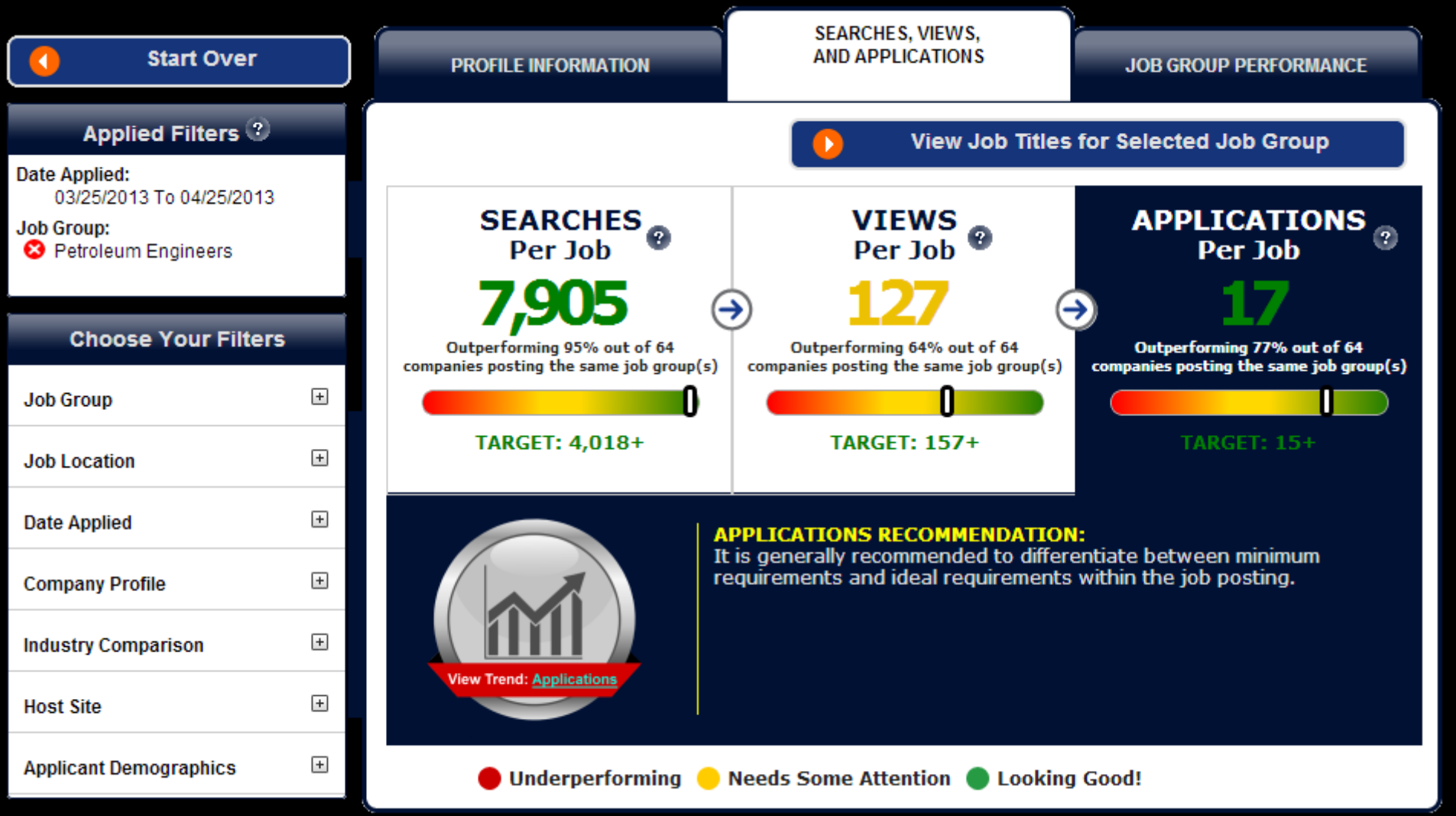
Export

Source: CareerBuilder

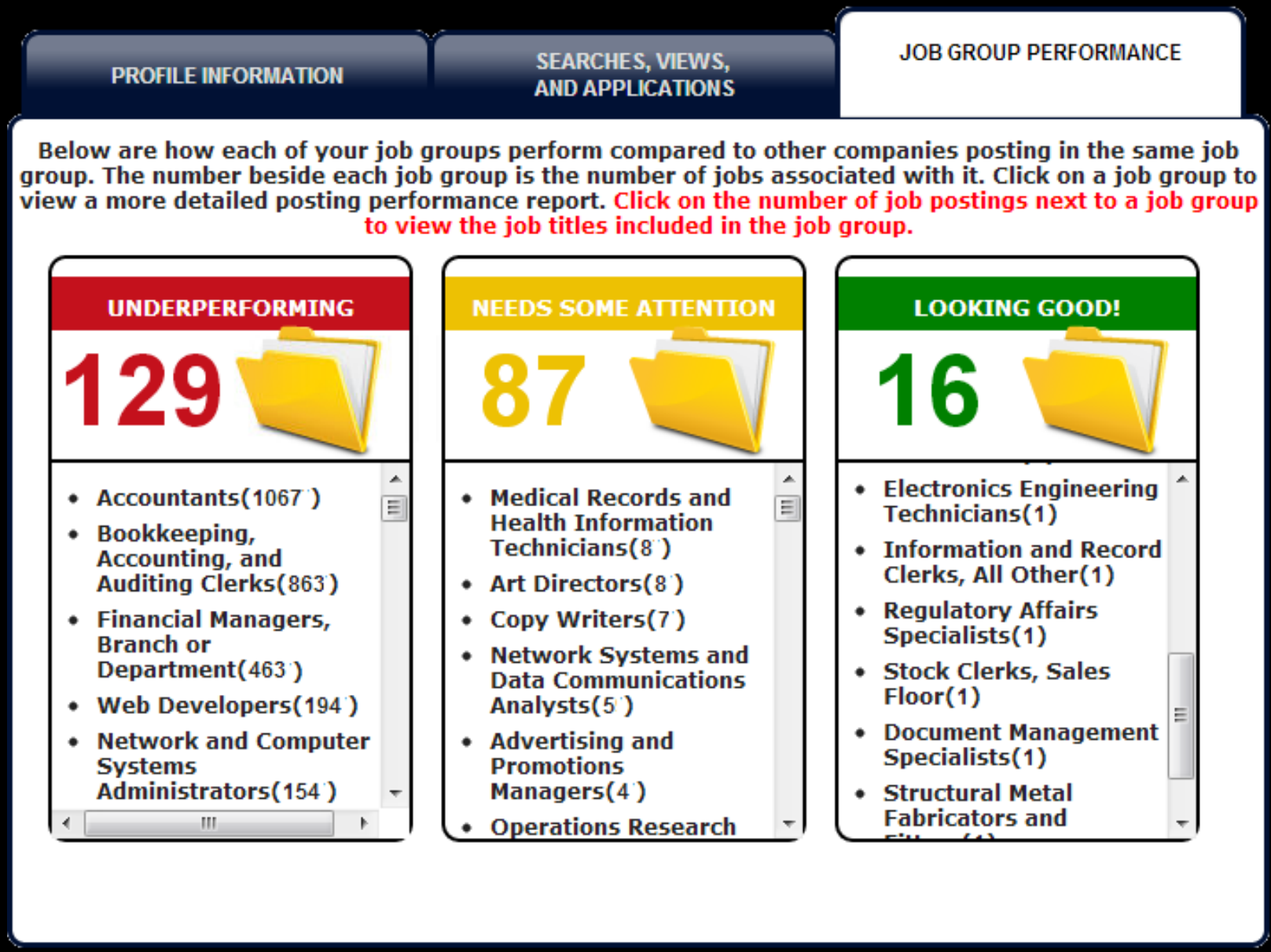
Customer's Performance vs. Competition



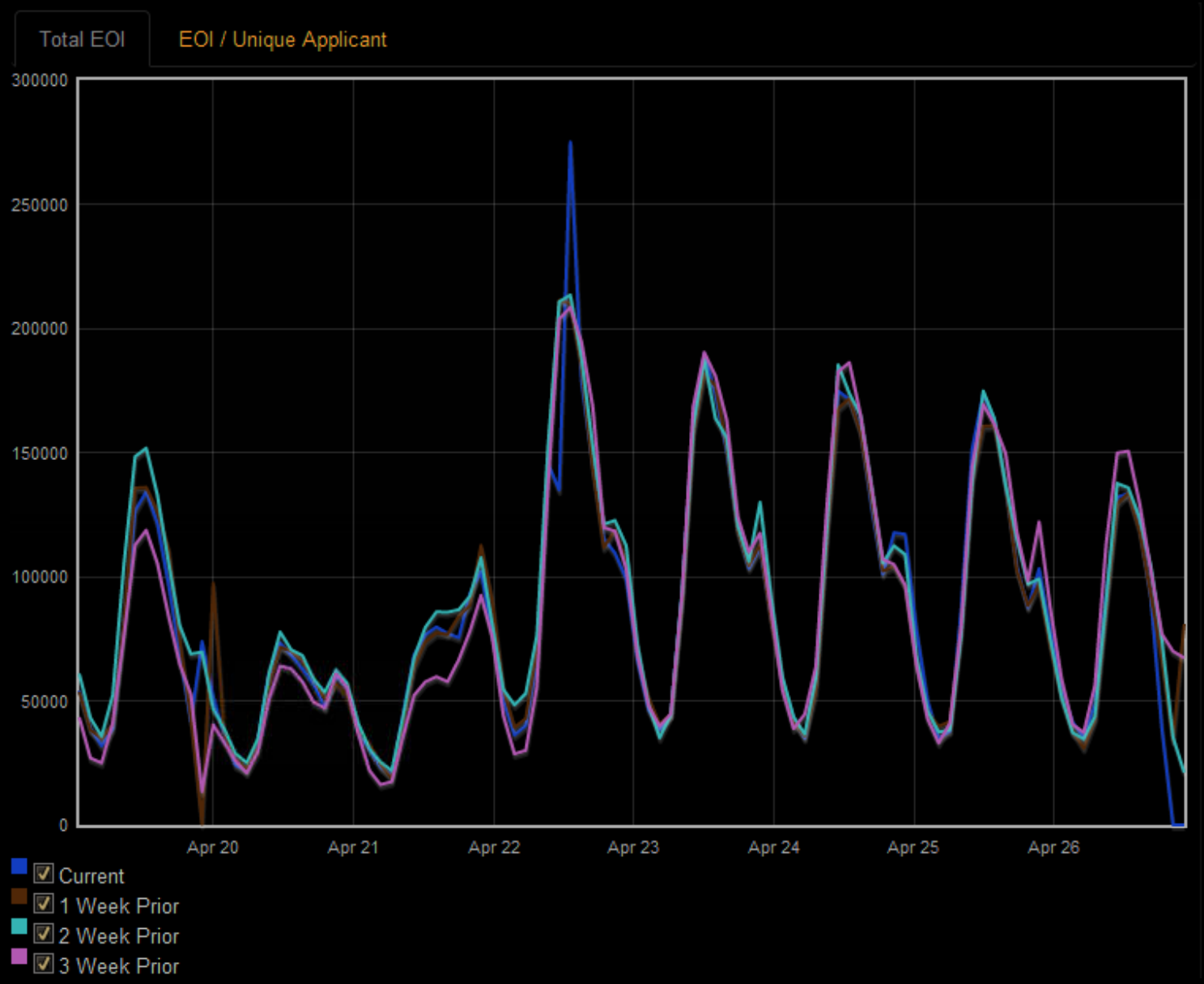
Customer's Performance vs. Competition



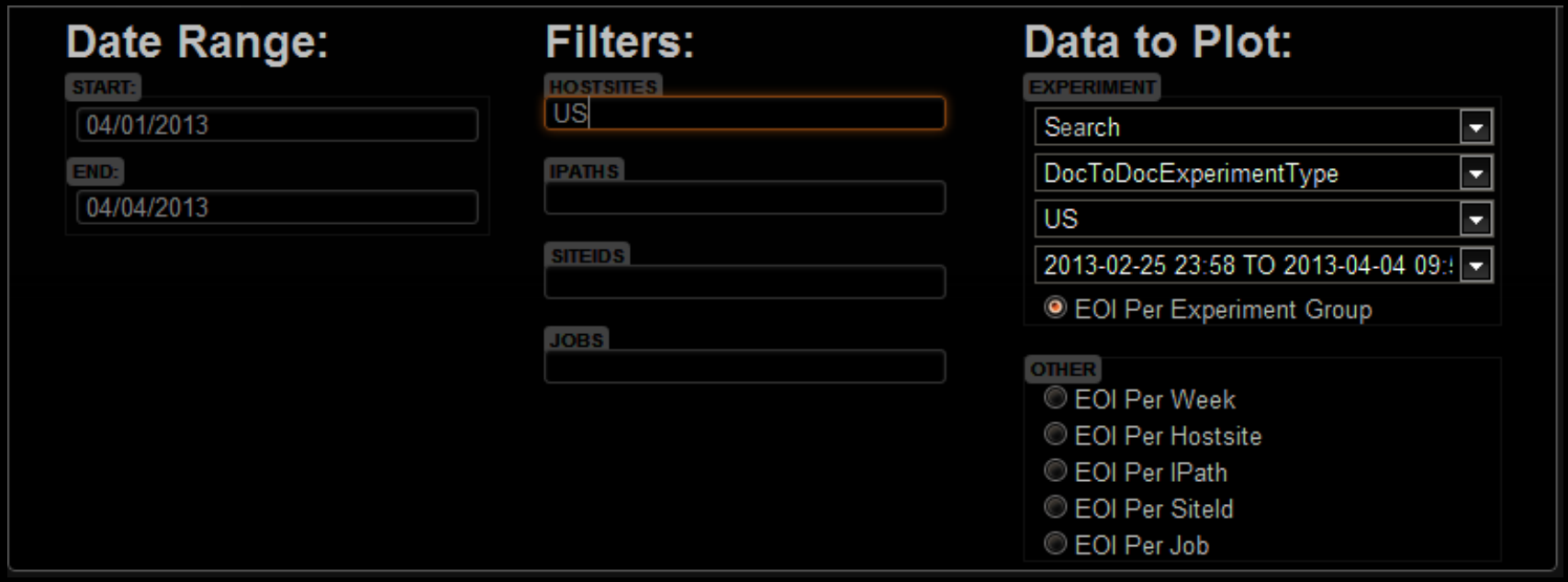
Customer's Performance vs. Competition



Event Stream Analysis



A/B Testing


Date Range:
START:
04/01/2013
END:
04/04/2013
Filters:
HOSTSITES
US
IPATHS
SITEIDS
JOBS
Data to Plot:
EXPERIMENT
Search
DocToDocExperimentType
US
2013-02-25 23:58 TO 2013-04-04 09:
EOI Per Experiment Group
OTHER
EOI Per Week
EOI Per Hostsite
EOI Per IPath
EOI Per SiteId
EOI Per Job

A/B Testing


Date Range:
Filters:
Data to Plot:
START:
04/01/2013
END:
04/04/2013
Total EOI
EOI / Unique Applicant
25000
20000
15000
10000
5000
0
Apr 1 12:00
Apr 2 0:00
Apr 2 12:00
Apr 3 0:00
Apr 3 12:00
Apr 4 0:00
Apr 4 12:00
Apr 5
Group 0: MultipleONetsNoTfldf (25% traffic)
Group 1: MultipleONetsWithTfldf (25% traffic)
Group 2: Off (25% traffic)
Group 3: TfldfOnly (25% traffic)

A/B Testing

1. Hash all users based upon user identifier in application stack
2. Flow user event streams into Solr with user identifier
3. Implement custom hashing function in Solr
4. Facet based upon custom function
5. Profit!

A/B Testing – Adding Function to Solr

```
//Custom Solr Plugin
public class ExperimentGroupFunctionParser extends ValueSourceParser {
    public ValueSource parse(FunctionQParser fqp) throws SyntaxError{
        String experimentName = fqp.parseArg();
        ValueSource uniqueID = fqp.parseValueSource();
        int numberOfGroups = fqp.parseInt();
        double ratioPerGroup = fqp.parseDouble();
        return new ExperimentGroupFunction (experimentName, uniqueID, numberOfGroups, ratioPerGroup);
    }
}

public class ExperimentGroupFunction extends ValueSource {
…
   @Override
    public FunctionValues getValues(Map context, AtomicReaderContext readerContext) throws IOException {
       final FunctionValues vals = uniqueIdFieldValueSource.getValues(context, readerContext);
            return new IntDocValues(this) {
                @Override
                public int intVal(int doc) {
                    //returns an deterministically hashed integer indicating which test group a user is in
                   return GetExperimentGroup(experimentName, vals.strVal(doc), numGroups, ratioPerGroup);
                }
    }
…
}

//SolrConfig.xml:
<valueSourceParser name=”experimentgroup"
    class="com.careerbuilder.solr.functions. ExperimentGroupFunctionParser" />
```

A/B Testing – Faceting on Function

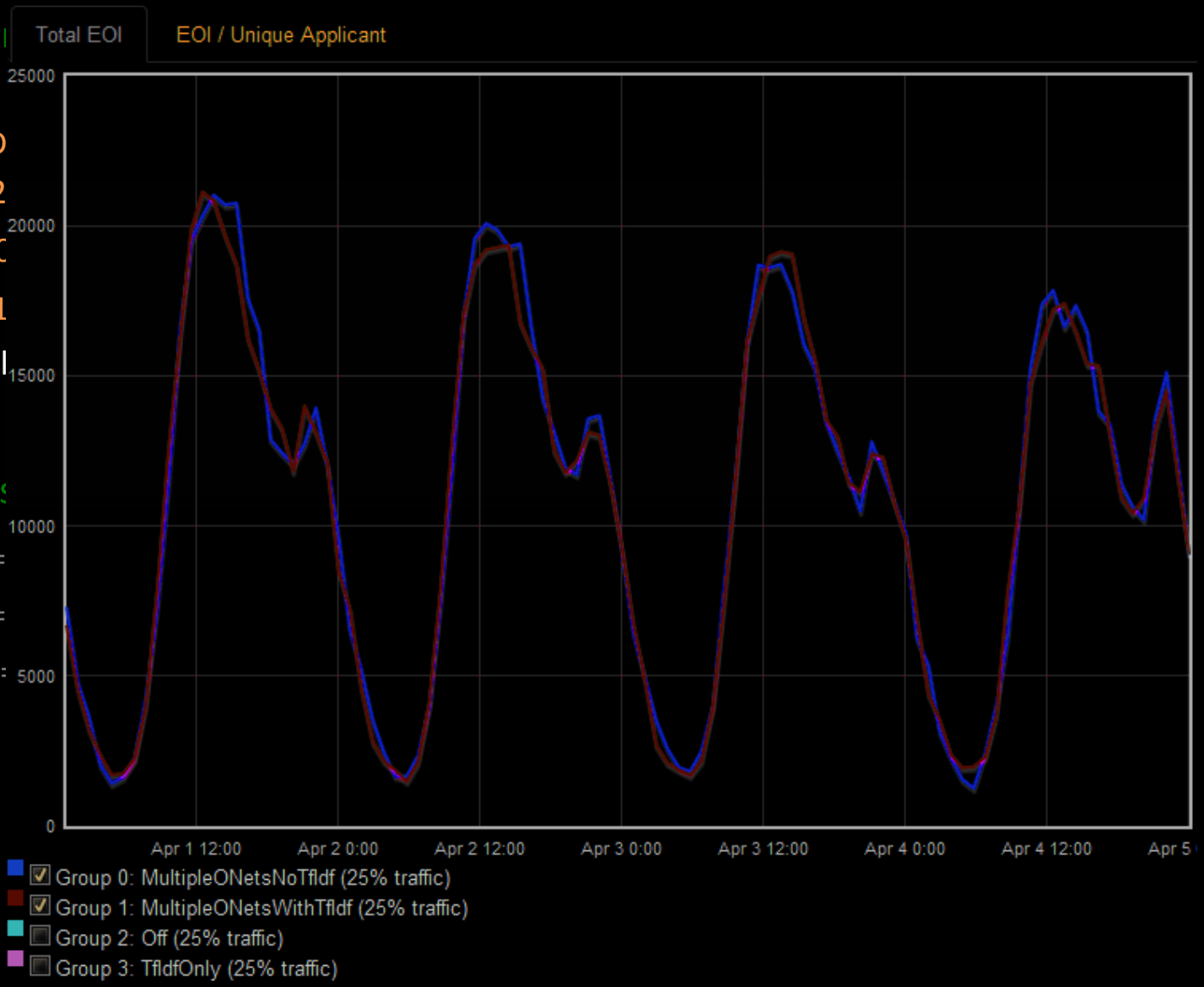
```
//Group 1 Time Series Facet Query:
/solr/select?
facet.range=eventDate&
facet.range.start=2013-04-01T01:00:00.000Z&
facet.range.end=2013-04-07T01:00:00.000Z&
facet.range.gap=+1Hour&
fq={!frange l=1 u=1 key="Group 1"}experimentgroup("EXPERIMENT_NAME", userIdField, 4, 0.25)

//repeat for groups 2-4
...fq={!frange l=2 u=2 key="Group 2"}experimentgroup("EXPERIMENT_NAME", userIdField, 4, 0.25)
...fq={!frange l=3 u=3 key="Group 3"}experimentgroup("EXPERIMENT_NAME", userIdField, 4,
0.25)...fq={!frange l=4 u=4 key="Group 4"}experimentgroup("EXPERIMENT_NAME", userIdField, 4,
0.25)
```

A/B Testing – Faceting on Function

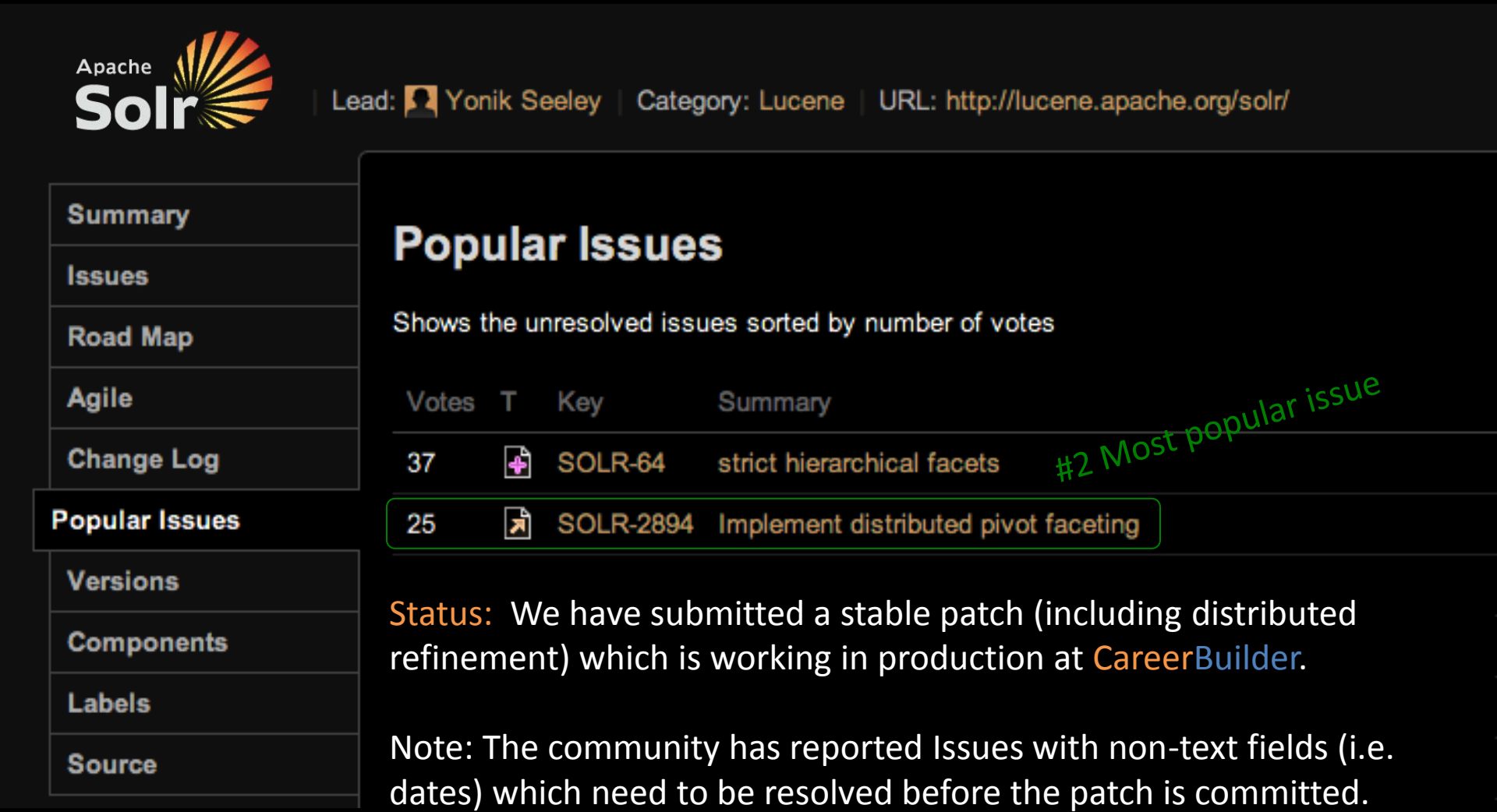
```
//Group 1 Time Se
/solr/select?
facet.range=eventD
facet.range.start=2
facet.range.end=2(
facet.range.gap=+1
fq={!frange l=1 u=1 l

//repeat for groups
...fq={!frange l=2 u=
...fq={!frange l=3 u=
0.25)...fq={!frange l
0.25)
```



Solr Patches in progress

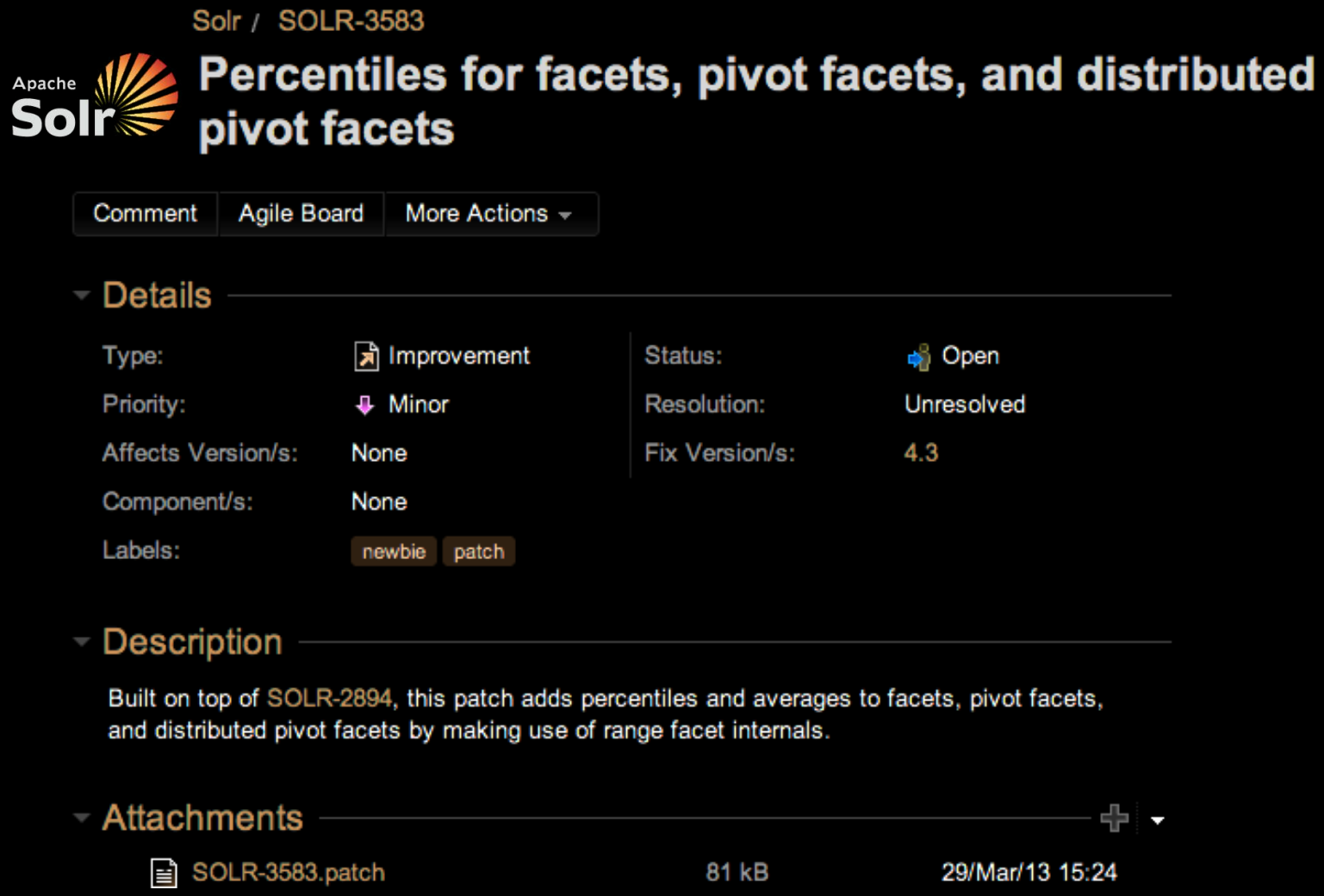
SOLR-2894: “Distributed Pivot Faceting”



Status: We have submitted a stable patch (including distributed refinement) which is working in production at CareerBuilder.

Note: The community has reported Issues with non-text fields (i.e. dates) which need to be resolved before the patch is committed.

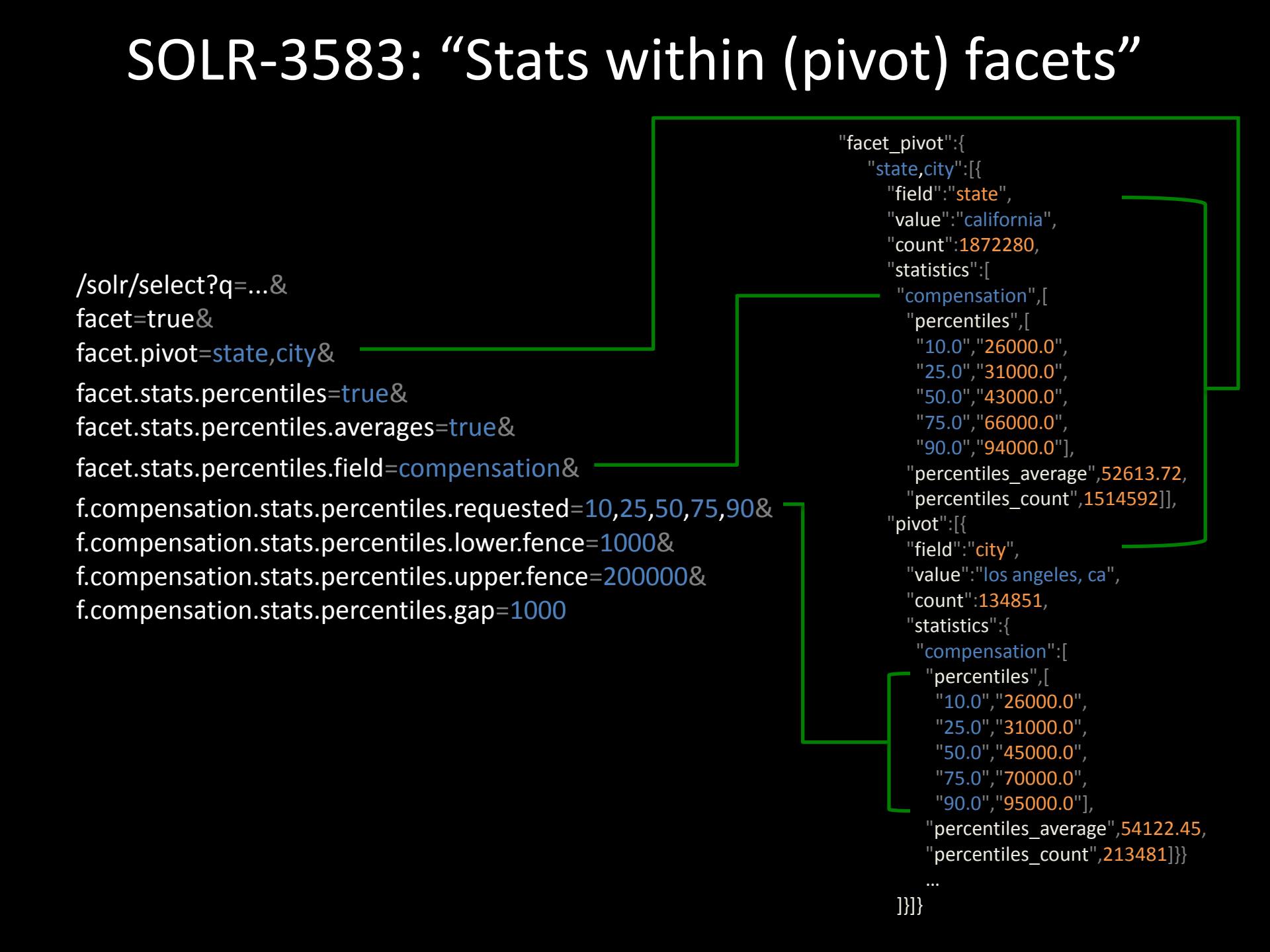
SOLR-3583: “Stats within (pivot) facets”



Status: We have submitted an early patch (built on top of SOLR-2894, distributed pivot facets), which is in production at CareerBuilder.

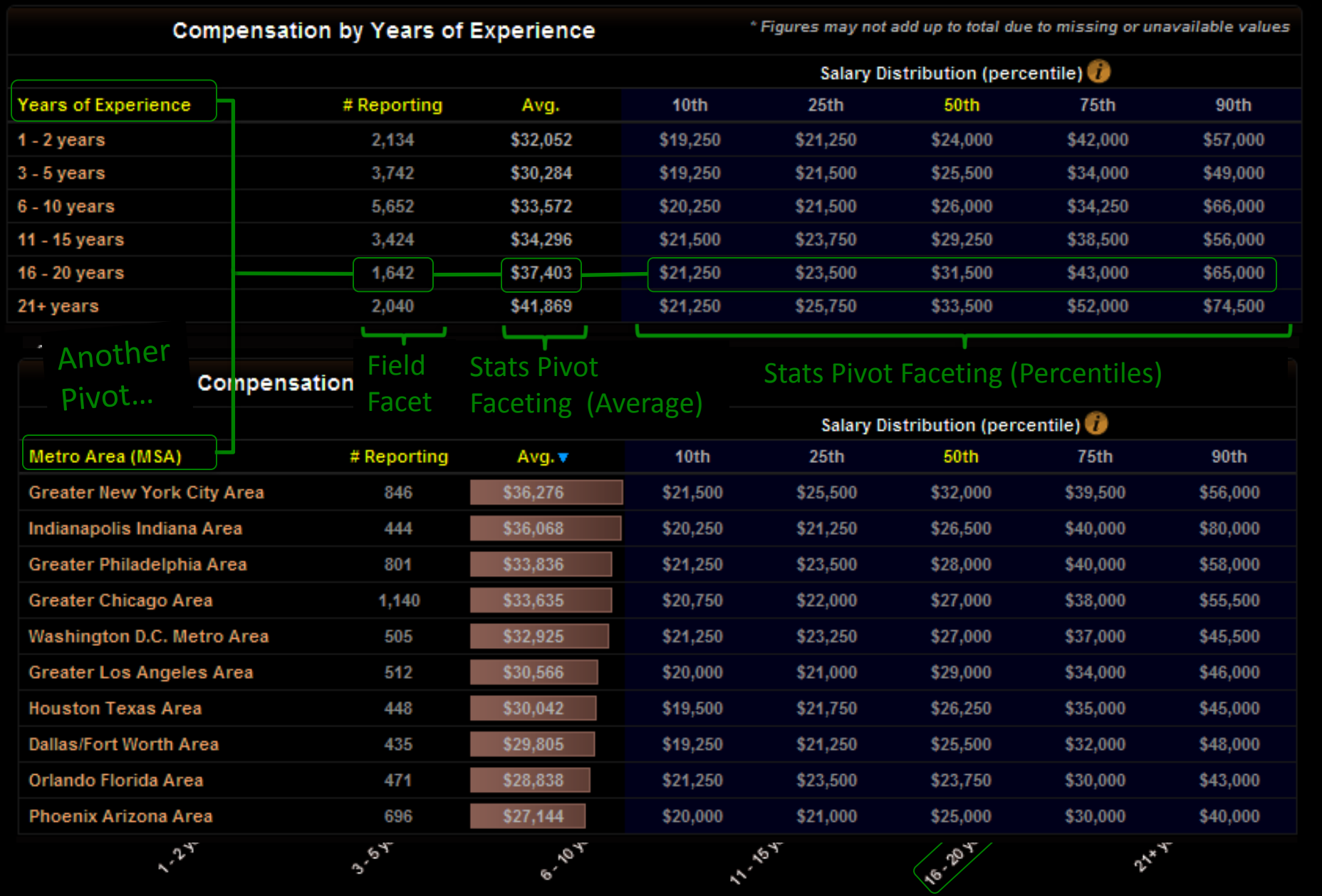
SOLR-3583: “Stats within (pivot) facets”

```
/solr/select?q=...&
facet=true&
facet.pivot=state,city&
facet.stats.percentiles=true&
facet.stats.percentiles.averages=true&
facet.stats.percentiles.field=compensation&
f.compensation.stats.percentiles.requested=10,25,50,75,90&
f.compensation.stats.percentiles.lower.fence=1000&
f.compensation.stats.percentiles.upper.fence=200000&
f.compensation.stats.percentiles.gap=1000
```



```
"facet_pivot":{
  "state,city":[{
    "field":"state",
    "value":"california",
    "count":1872280,
    "statistics":[
     "compensation",[
      "percentiles",[
       "10.0","26000.0",
       "25.0","31000.0",
       "50.0","43000.0",
       "75.0","66000.0",
       "90.0","94000.0"],
      "percentiles_average",52613.72,
      "percentiles_count",1514592]],
    "pivot":[{
      "field":"city",
      "value":"los angeles, ca",
      "count":134851,
      "statistics":{
       "compensation":[
        "percentiles",[
         "10.0","26000.0",
         "25.0","31000.0",
         "50.0","45000.0",
         "75.0","70000.0",
         "90.0","95000.0"],
        "percentiles_average",54122.45,
        "percentiles_count",213481]}}
        ...
    ]}]}
```

Real-world Use Case



Data Analytics with Solr… the next frontier…

Solr Analytics Wish List…

- Faceting Architecture Redesign:
  - "All" Facets should be pivot facets (default pivot depth = 1)
  - Each "Pivot" could be a field, query, or range
  - Meta information (like the Stats patch) could be nested in each pivot
  - Backwards compatibility with current facet response format for a while…
- Grouping
  - Grouping should also support multiple levels (Pivot Grouping)
  - If "Pivot Grouping" were supported… what about faceting within each of the pivot groups?
- Remaining caches should be converted to "per-segment" to enable NRT
- These kinds of changes would enable Solr to return rich Data Analytics calculations in a single search call where many calls are required today.

Recap

Billions of documents

+

Ad-hoc querying by keyword

+

Horizontal Scalability

+

Sub-second query responses

+

Facet on “anything” – field, function, etc.

+

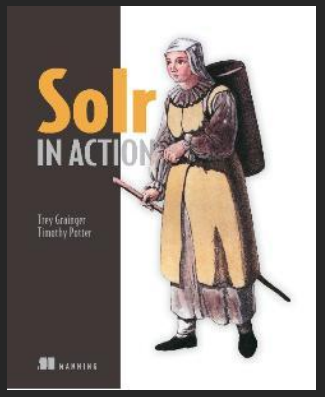
User friendly visualizations/UX

Building a real-time, Big Data Analytics Platform with Solr

Contact Info

- **Trey Grainger**

trey.grainger@careerbuilder.com
@treygrainger



Other presentations:

http://www.treygrainger.com

http://solrinaction.com

Yes, we are hiring @CareerBuilder. Come talk with me if you are interested…