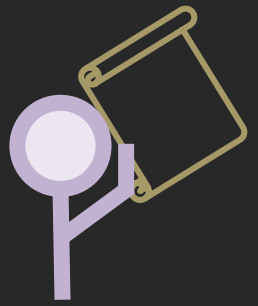
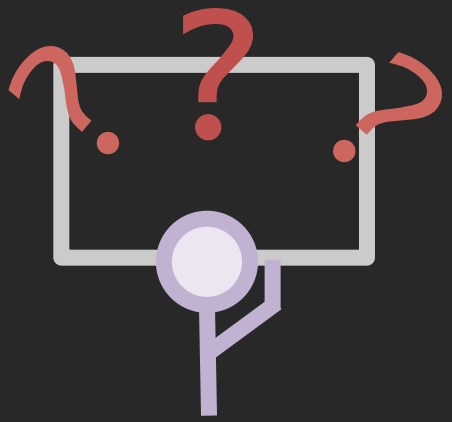
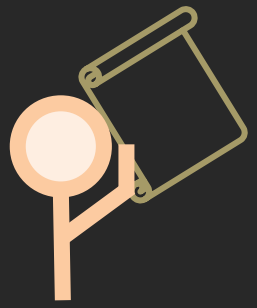
Building Blocks for Exploratory Data Analysis Tools

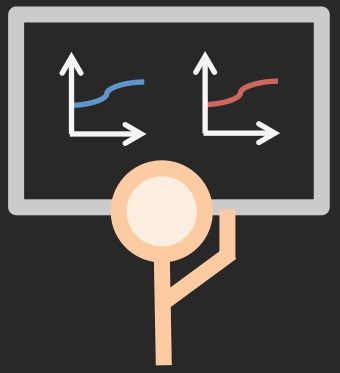
Sara Alspaugh — UC Berkeley
Archana Ganapathi — Splunk Inc.
Marti Hearst — UC Berkeley
Randy Katz — UC Berkeley

Motivation

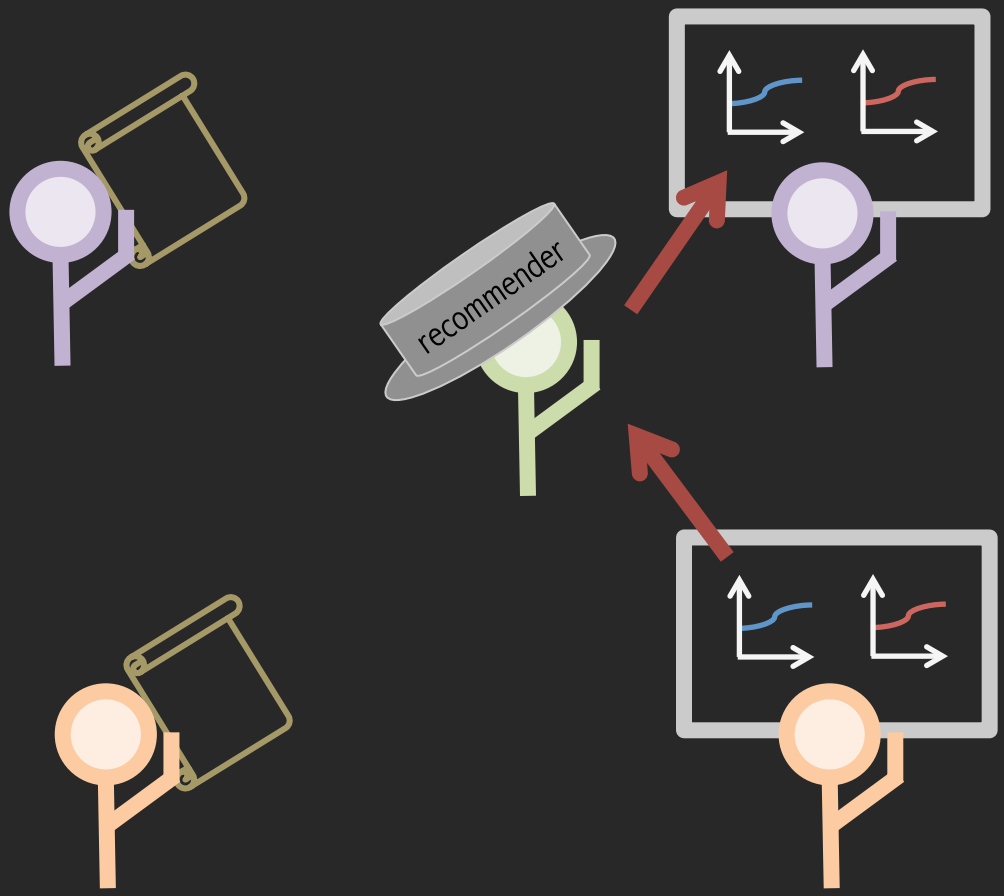








Motivation



Key Idea

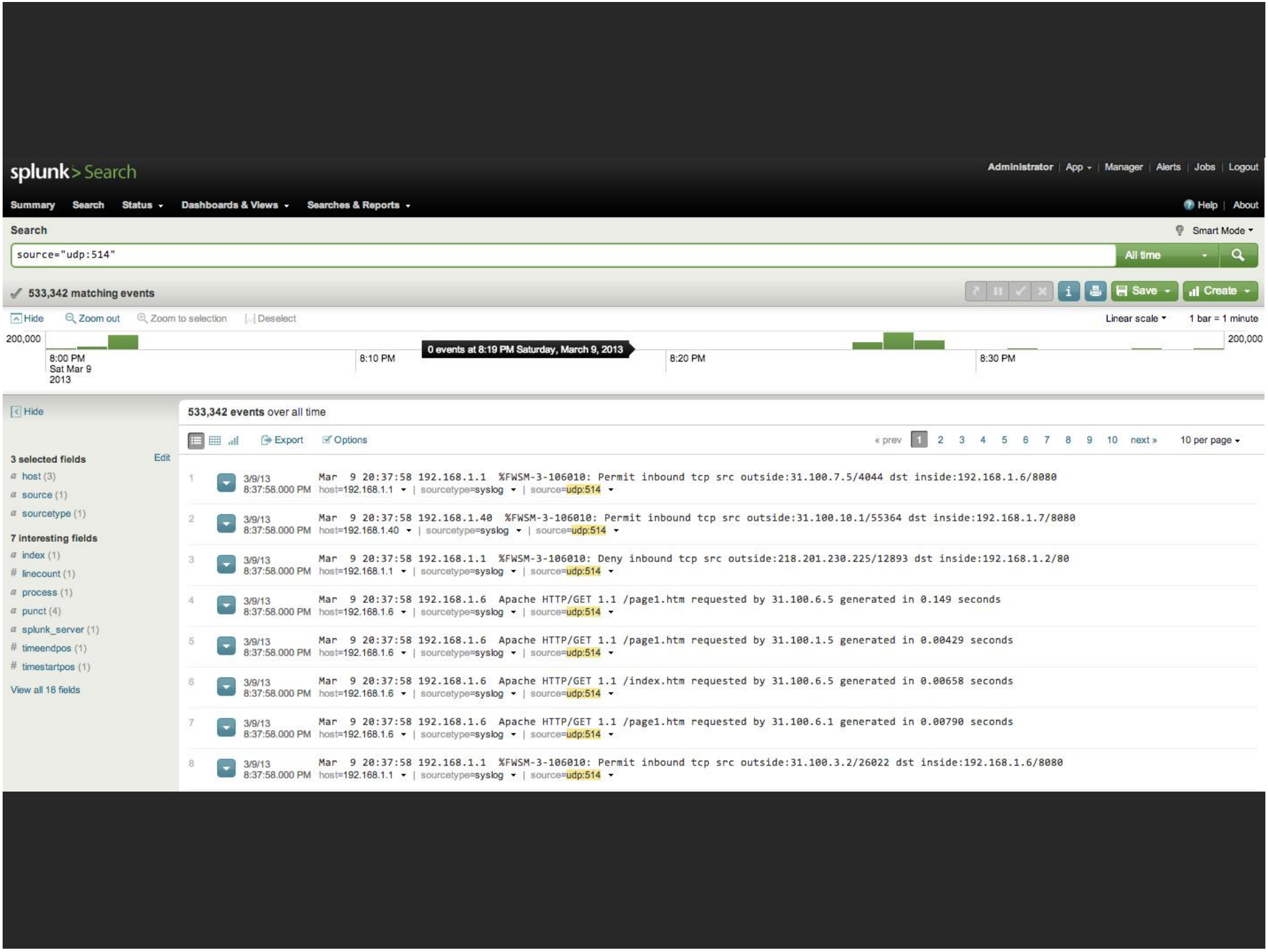
- exploit similarity between data sets to make recommendations
  - but we don’t have analyzed data sets, so use queries (small programs) instead

Approach

1. collect user queries from Splunk
   - but not data sets
2. apply latent semantic analysis
   - test key idea
   - use extension of this for recommending

Splunk

- View a demo here: http://www.eecs.berkeley.edu/~alspaugh/misc/splunk_demo_screencast.mov
  - you might need QuickTime for your the browser
  - be patient, it can take a while to load
- Splunk collects and indexes large amounts of semi-structured time series data
- Data is often log data
  - each time-stamped entry is a row
- Users visualize data via GUI and query language
- Data is processed in stages expressed in queries:
  - `command arguments | command arguments | ...`
- No schema; key-value pairs are extracted at run time
  - think bags of key-value pairs instead of tables


splunk> Search
Administrator | App | Manager | Alerts | Jobs | Logout
Summary Search Status Dashboards & Views Searches & Reports
Help | About
Search
Smart Mode
source="udp:514"
All time
533,342 matching events
Save Create
Hide Zoom out Zoom to selection Deselect
Linear scale 1 bar = 1 minute
200,000
0 events at 8:19 PM Saturday, March 9, 2013
8:00 PM Sat Mar 9 2013
8:10 PM
8:20 PM
8:30 PM
Hide
533,342 events over all time
Export Options
« prev 1 2 3 4 5 6 7 8 9 10 next » 10 per page
3 selected fields Edit
host (3)
source (1)
sourcetype (1)
7 interesting fields
index (1)
linecount (1)
process (1)
punct (4)
splunk_server (1)
timeendpos (1)
timestartpos (1)
View all 18 fields
1 3/9/13 8:37:58.000 PM Mar 9 20:37:58 192.168.1.1 %FWSM-3-106010: Permit inbound tcp src outside:31.100.7.5/4044 dst inside:192.168.1.6/8080
host=192.168.1.1 | sourcetype=syslog | source=udp:514
2 3/9/13 8:37:58.000 PM Mar 9 20:37:58 192.168.1.40 %FWSM-3-106010: Permit inbound tcp src outside:31.100.10.1/55364 dst inside:192.168.1.7/8080
host=192.168.1.40 | sourcetype=syslog | source=udp:514
3 3/9/13 8:37:58.000 PM Mar 9 20:37:58 192.168.1.1 %FWSM-3-106010: Deny inbound tcp src outside:218.201.230.225/12893 dst inside:192.168.1.2/80
host=192.168.1.1 | sourcetype=syslog | source=udp:514
4 3/9/13 8:37:58.000 PM Mar 9 20:37:58 192.168.1.6 Apache HTTP/GET 1.1 /page1.htm requested by 31.100.6.5 generated in 0.149 seconds
host=192.168.1.6 | sourcetype=syslog | source=udp:514
5 3/9/13 8:37:58.000 PM Mar 9 20:37:58 192.168.1.6 Apache HTTP/GET 1.1 /page1.htm requested by 31.100.1.5 generated in 0.00429 seconds
host=192.168.1.6 | sourcetype=syslog | source=udp:514
6 3/9/13 8:37:58.000 PM Mar 9 20:37:58 192.168.1.6 Apache HTTP/GET 1.1 /index.htm requested by 31.100.6.5 generated in 0.00658 seconds
host=192.168.1.6 | sourcetype=syslog | source=udp:514
7 3/9/13 8:37:58.000 PM Mar 9 20:37:58 192.168.1.6 Apache HTTP/GET 1.1 /page1.htm requested by 31.100.6.1 generated in 0.00790 seconds
host=192.168.1.6 | sourcetype=syslog | source=udp:514
8 3/9/13 8:37:58.000 PM Mar 9 20:37:58 192.168.1.1 %FWSM-3-106010: Permit inbound tcp src outside:31.100.3.2/26022 dst inside:192.168.1.6/8080
host=192.168.1.1 | sourcetype=syslog | source=udp:514

splunk> Call Detail Reporting Console

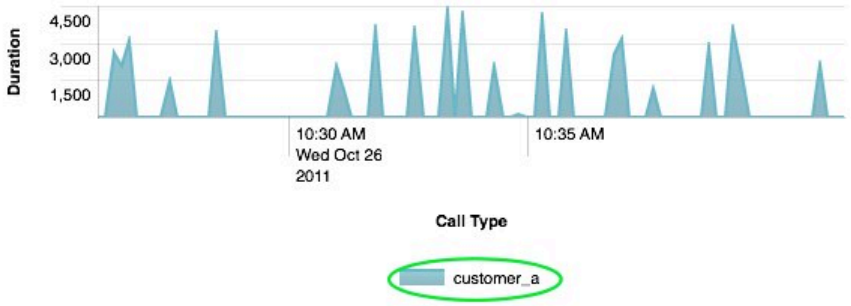
Logged in as customera_user1 | App | Manager | Jobs | Logout

CDR Dashboard | Dashboards | Views | Searches & Reports | Help | About

CDR Dashboard | Actions

Average Call Duration

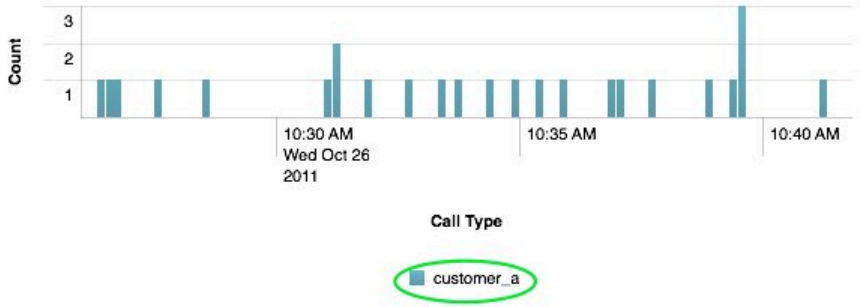
refreshed: **today at 10:41:48 AM.**



View results

Count of ISDN recipients

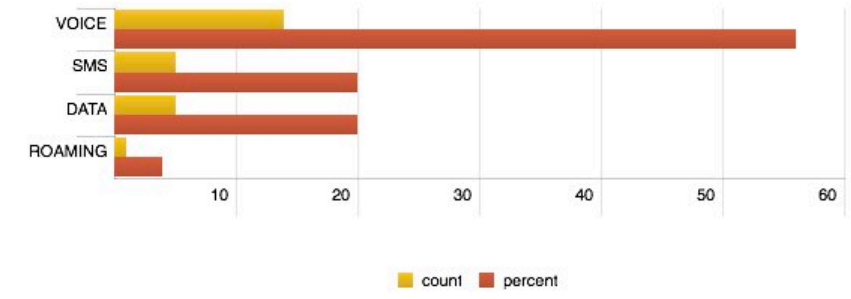
refreshed: **today at 10:41:47 AM.**



View results

Top Mediums

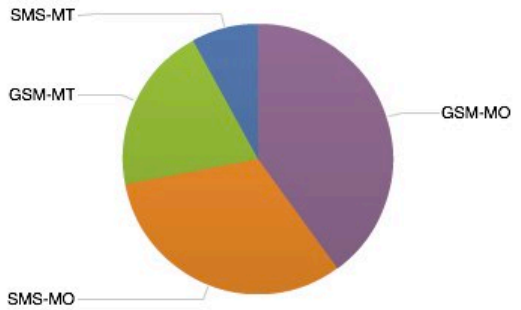
refreshed: **today at 10:41:47 AM.**



View results

Top Types of Calls

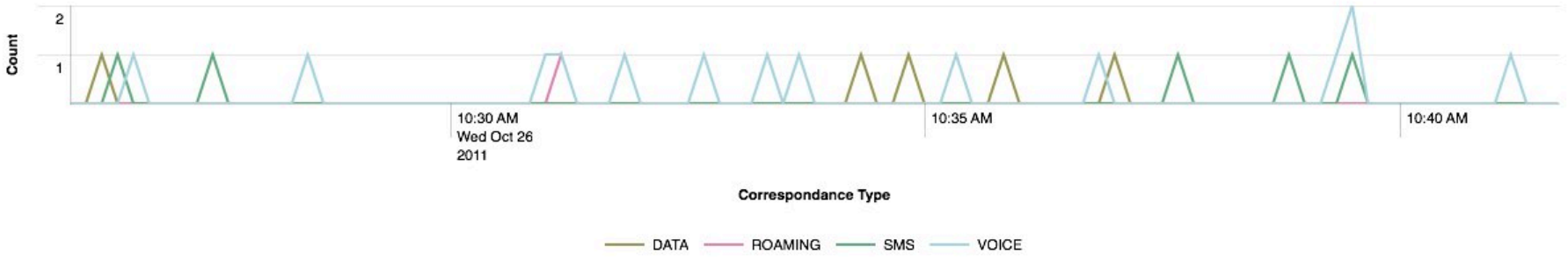
refreshed: **today at 10:41:47 AM.**



View results

Count of ISDN Callers by Correspondance Type

refreshed: **today at 10:41:47 AM.**



View results

Splunk Query Example

```
search source=eqs7day-M1.csv
| eval Description=
      case(Depth<=70, "Shallow",
          Depth>70 AND Depth<=300, "Mid",
          Depth>300, "Deep")
| table Datetime, Region, Depth, Description
```

- commands and operators
- field (i.e., key or column)
- value (i.e., column values)
- pipe to next command

LSA

DOCUMENTS

1. Romeo and Juliet.

2. Juliet: O happy dagger!

3. Romeo died by dagger.

4. “Live free or die”, that’s the New-Hampshire’s motto.

5. Did you know, New-Hampshire is in New-England.

```
d1 : romeo, juliet.
d2 : juliet, happy, dagger
d3 : romeo, die, dagger.
d4 : live, free, die,
        New-Hampshire
d5 : New-Hampshire
```

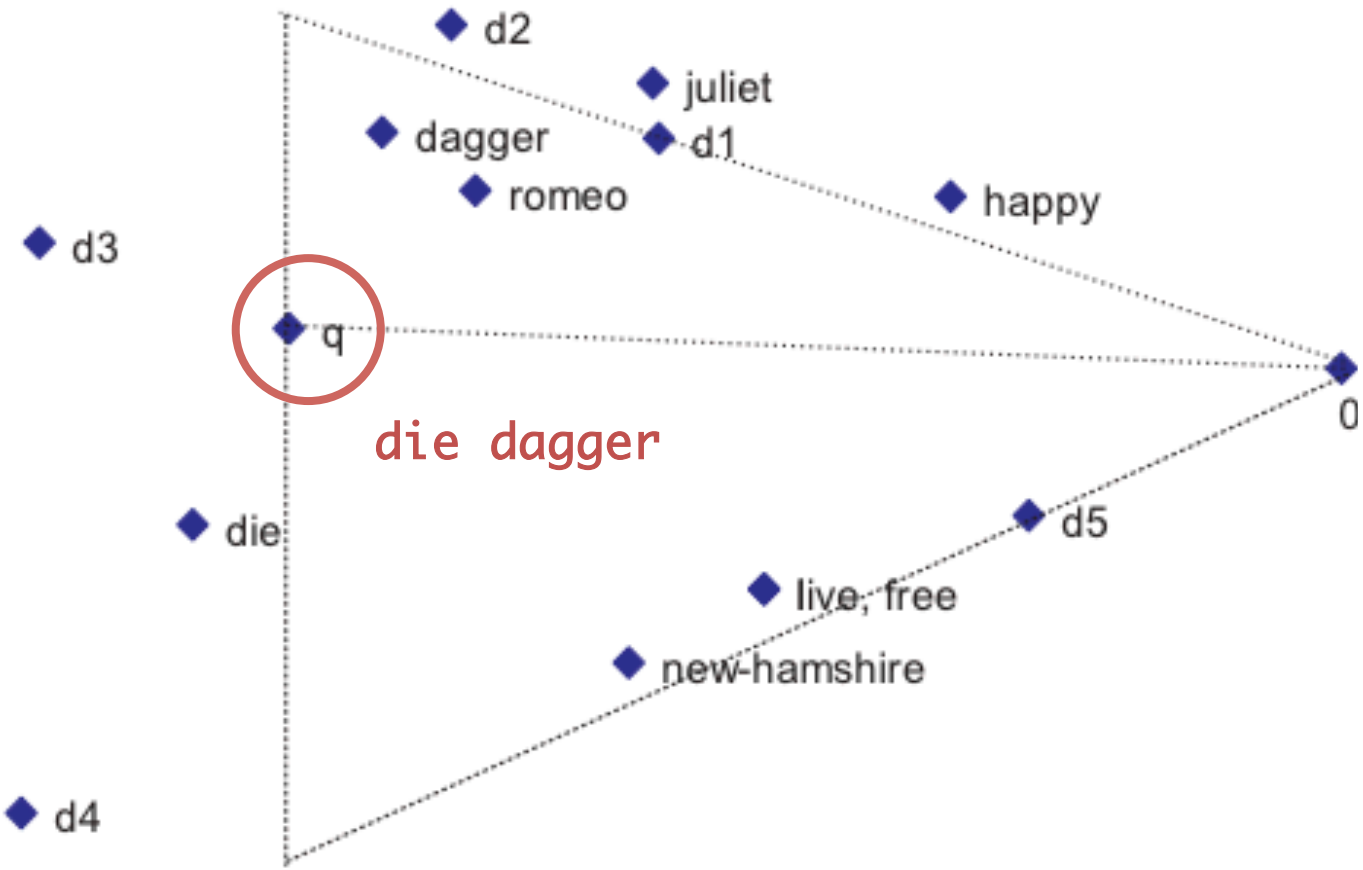
QUERIES

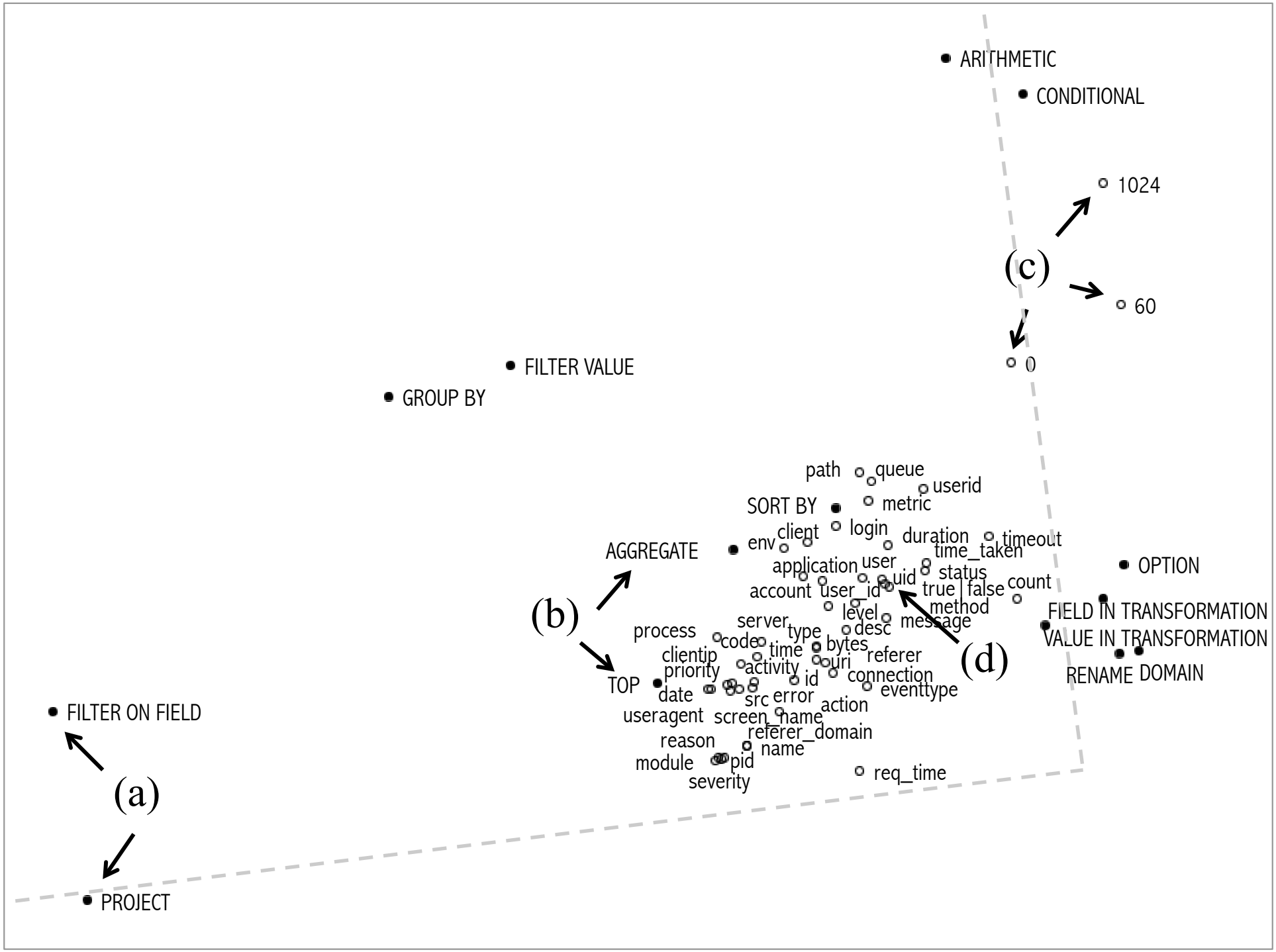
1.
```
search sourcetype=access_combined
| where isnull(task_queue)
| timechart count span=1min
| eval count=count/60
```

2.
```
search host="appspot.com" change_time=*
| eventstats count as Total
| bucket change_time span=log10
| stats count as Count, max(Total) as
    Total by change_time
| eval percentage=Count/Total*100
```

```
bucket     : change_time
search     : sourcetype, host, change_time
eval       : count, total, percentage,
                 60, 100
eventstats : count
stats      : count, total, change_time
timechart  : count, 1min
where      : task_queue
```

Document example from:
Thomo, Alex. Latent Semantic Analysis (Tutorial). www.engr.uvic.ca/~seng474/svd.pdf


d2
juliet
dagger
d1
romeo
happy
d3
q
0
die dagger
die
d5
live, free
new-hamshire
d4


ARITHMETIC
CONDITIONAL
1024
60
0
(c)
FILTER VALUE
GROUP BY
path
queue
userid
SORT BY
metric
client
login
duration
timeout
env
time_taken
AGGREGATE
application
user
status
OPTION
uid
count
account
user_id
true | false
(b)
method
level
message
server
FIELD IN TRANSFORMATION
process
type
desc
code
bytes
VALUE IN TRANSFORMATION
clientip
time
referer
(d)
priority
activity
uri
TOP
id
connection
RENAME DOMAIN
date
src
error
eventtype
action
useragent
screen_name
referer_domain
reason
name
module
pid
req_time
severity
FILTER ON FIELD
(a)
PROJECT

Conclusion

- Evidence for making recommendations based on similarity of data:

  inconclusive but promising

- Possible approaches:
  - recommendation algorithms: LSI, nearest neighbor
  - dimensionality reduction: NMF, t-SNE (hat tip reviewer #3)
  - Bayesian: naïve, hierarchical
  - natural language processing
  - others?

Thank you.

Questions?