Active Data

http://activedata.allizom.org

Kyle Lahnakoski
Engineering Productivity
(formerly Auto-Tools Team)
…but still the “A Team”

Objective

- Data driven decision making
- Defining metrics and tracking progress
- Comparing current state to past performance

Data Warehouse

> *a copy of transaction data specifically structured for query and analysis*

-- Ralph Kimball

> *If you have to wait minutes or hours for a question to be answered, you simply can’t iterate on hypotheses and investigate in a meaningful way.*

-- Jeffrey Wang
(just someone on internet*)

* https://amplitude.com/blog/2015/08/25/scaling-analytics-at-amplitude/

Data Warehouse

- Fast data access
- Reduce effort to get data
- Offset query load from transactional systems
- Standardize data
- Comprehensive single source
- Share data

Data Warehouse

- Fast data access
- Reduce effort to get data
- Offset query load from transactiona
- Standardize data
- Comprehensive single source
- Share data

- Supporting ad hoc reporting and charts
- Allow quick exploration and discovery
- Spark new types of analysis

Data Warehouse

- Fast data access
- Reduce effort to get data
- Offset query load from transactional
- Standardize data
- Comprehensive single source
- Share data

- Eliminate log parsing
- Reduce time to find needles in haystack
- No database schema to declare
- No indexing or caching for speed

Data Warehouse

- Fast data access
- Reduce effort to get data
- Offset query load from transactional systems
- Standardize data
- Comprehensive single source
- Share data

- Report queries often require expensive joins
- Long time series demand lots of data

Data Warehouse

- Fast data access
- Reduce effort to get data
- Offset query load from transactiona
- Standardize data
- Comprehensive single source
- Share data

- Standard naming convention
- Common dimensions and lookup tables
- Identical format

Data Warehouse

- Fast data access
- Reduce effort to get data
- Offset query load from transactional systems
- Standardize data
- Comprehensive single source
- Share data

Data from multiple systems in single location

Data Warehouse

- Fast data access
- Reduce effort to get data
- Offset query load from transactional systems
- Standardize data
- Comprehensive single so
- Share data

- Enable others for increased mindshare
- Do not block because data is inaccessible
- No web service setup
- No production services

Data Warehouse

Active Data

- Fast data access
- Reduce effort to get data
- Offset query load from transactional systems
- Standardize data
- Comprehensive single source
- Share data

Active Data

- Fast data access
- Reduce effort to get data
- Offset query load from transactional
- Standardize data
- Comprehensive single source
- Share data

- 3 billion test results, query response in under a minute
- Speed limited by response volume (of course)
- Eg “B*yte count of structured logs from August*” less than 3 seconds

Active Data

- Fast data access
- Reduce effort to get data
- Offset query load from transactional s
- Standardize data
- Comprehensive single source
- Share data

- Query language to request data
- Summarize with aggregates
- Focus on particular features with filters
- Pull the raw records

Active Data

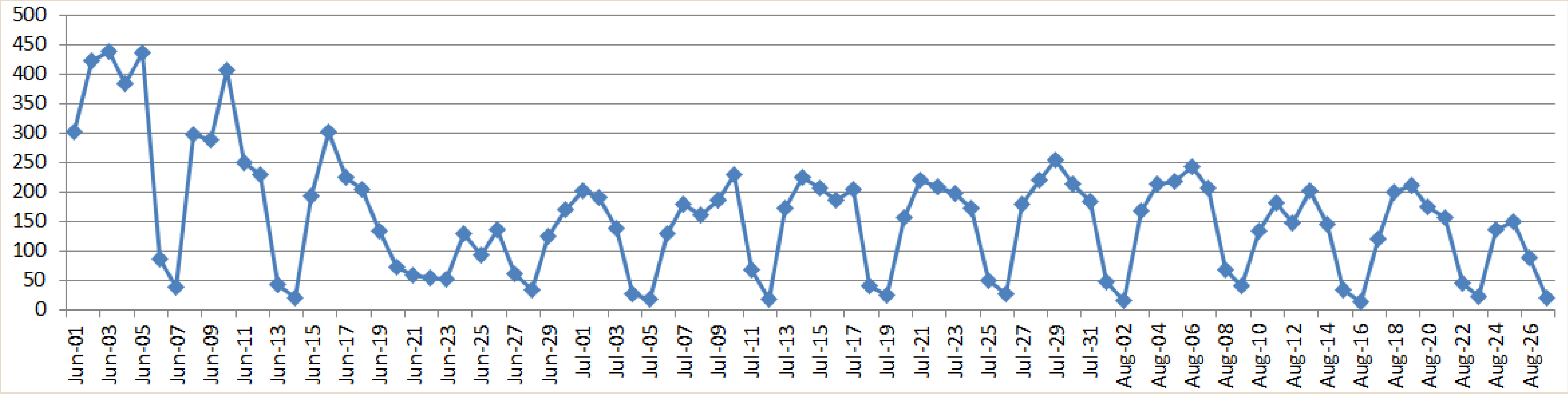
- Fast data access
- Reduce effort to get data
- Offset query load from transactional systems
- Standardize data
- Comprehensive single source
- Share data

- Use `repo` as a cache of hg.mozilla.org
- MoDevMetrics (ActiveData precursor) holds historical bug data for trending dashboards

Active

- Fast data access
- Reduce effort to get data
- Offset query load from transactional systems
- Standardize data
- Comprehensive single source
- Share data

New changesets per day over past 3 months (4sec)



- Use `repo` as a cache of hg.mozilla.org
- BugzillaES (ActiveData precursor)

Active Data

- Fast data access
- Reduce effort to get data
- Offset query load from transactional
- Standardize data
- Comprehensive single source
- Share data

- Uses pulsetranslator for normalized build properties
- hg.mozilla.org changeset metadata added to all results

Active Data

- Fast data access
- Reduce effort to get data
- Offset query load from transactional s
- Standardize data
- Comprehensive single source
- Share data

- Unit test structured logs
- Mercural repo
- Buildbot properties
- Orangefactor
- Talos performance metrics
- Bugzilla?
- Treeherder?

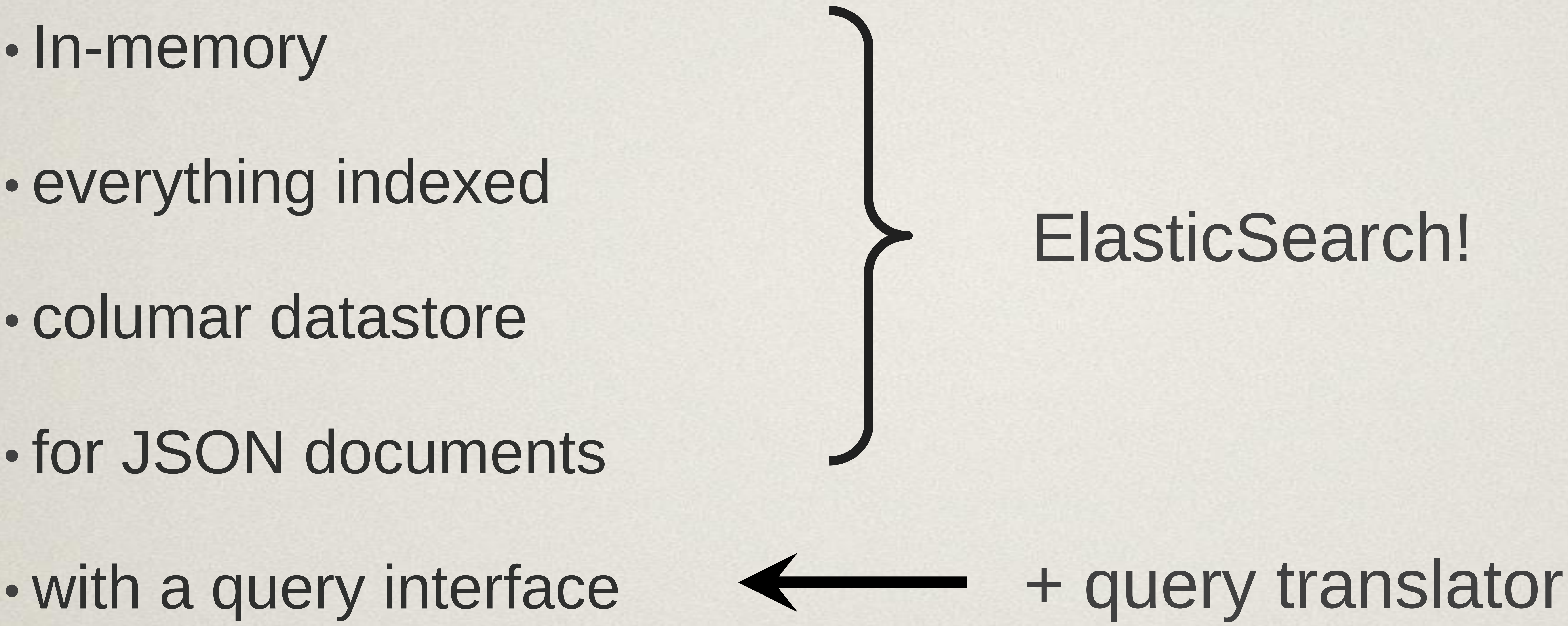
Active Data

- Fast data access
- Reduce effort to get data
- Offset query load from transactional systems
- Standardize data
- Comprehensive single source
- Share data

```
curl http://activedata.allizom.org/query -X POST -d
"{\"from\":\"unittest\"}"
```

Active Data

What is it?



- In-memory
- everything indexed
- columar datastore
- for JSON documents
- with a query interface

ElasticSearch!

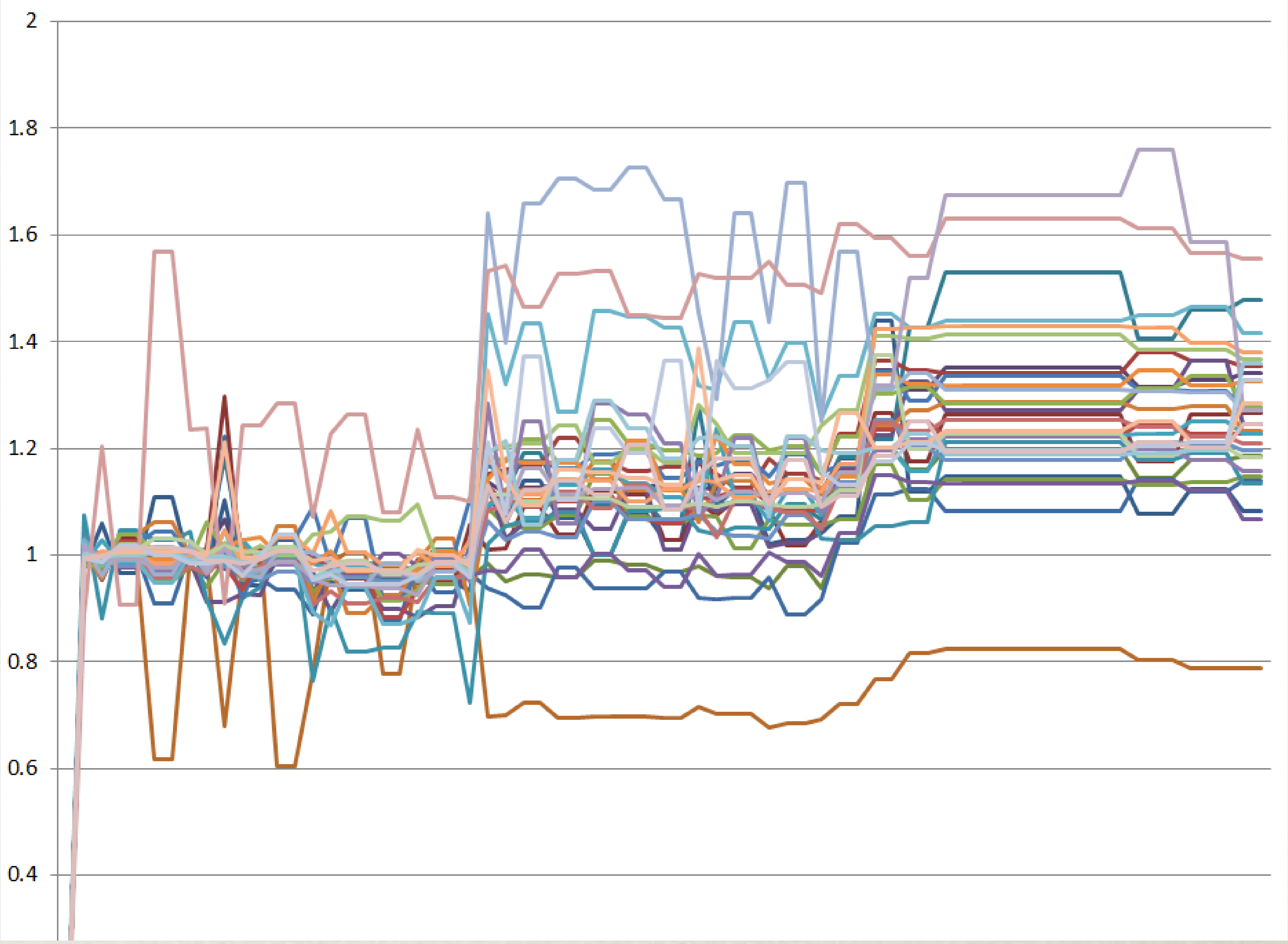
+ query translator

Contents

- Unit test results (2 months, 2 billion test results)
- Mercural repo
- Buildbot properties
- Orange Factor
- Talos performance metrics
- Bugzilla?
- Treeherder?

Examples

- Identify change in test times

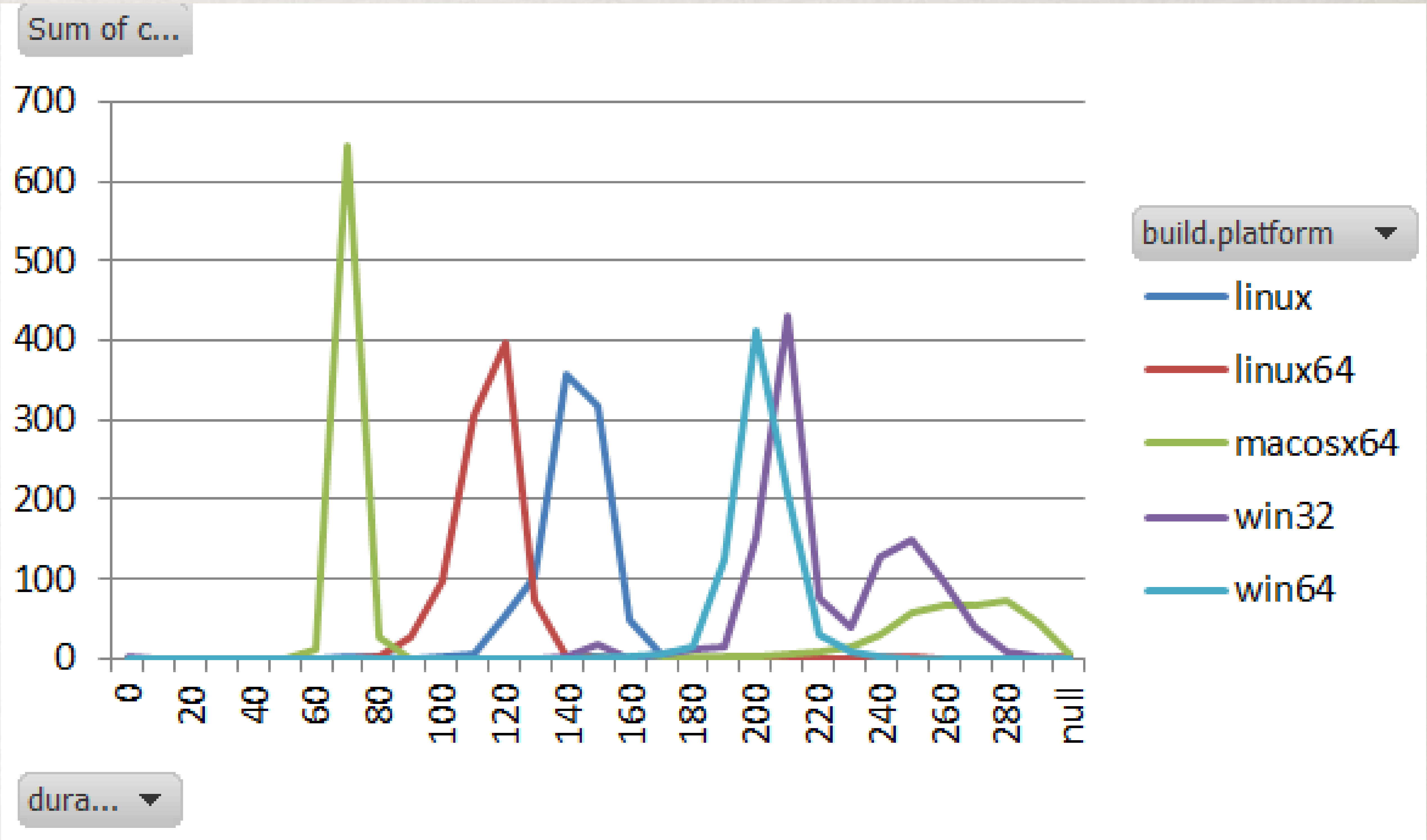


* ActiveData does not include visualization a this time

Examples

- Identify change in test times

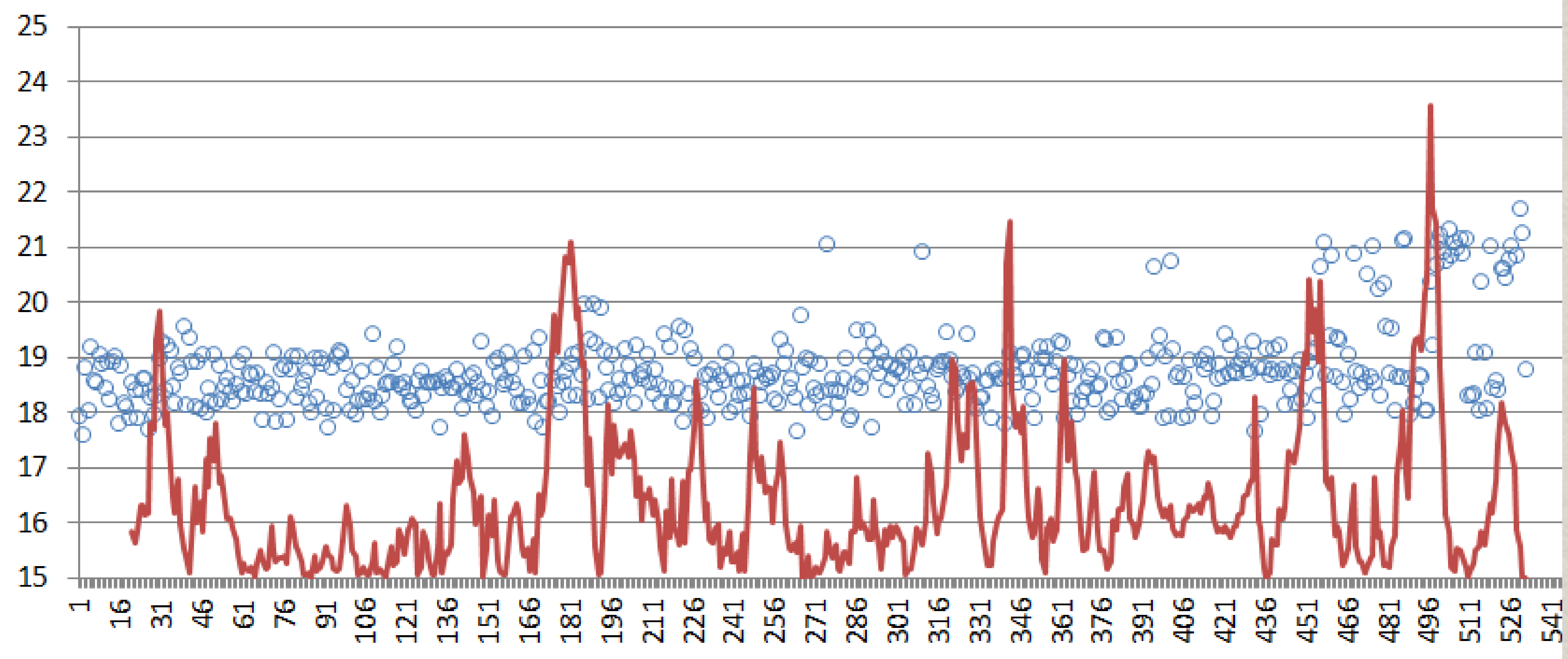
- Test-time distributions



* ActiveData does not include visualization a this time

Examples

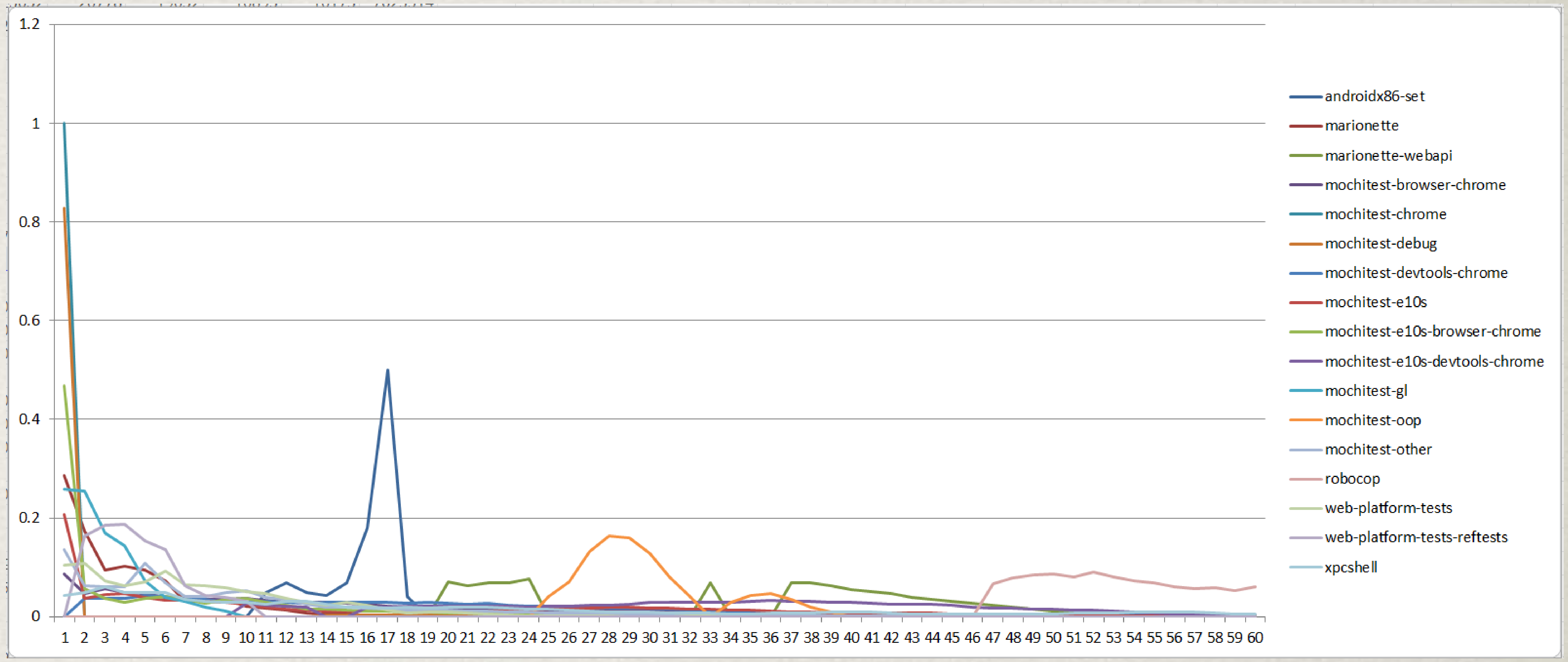
- Identify change in test times
- Test time distributions
- Visualize perf regression



* ActiveData does not include visualization a this time

Examples

- Identify change in test times
- Test time distributions
- Visualize perf regression
- Fail rate by time into suite

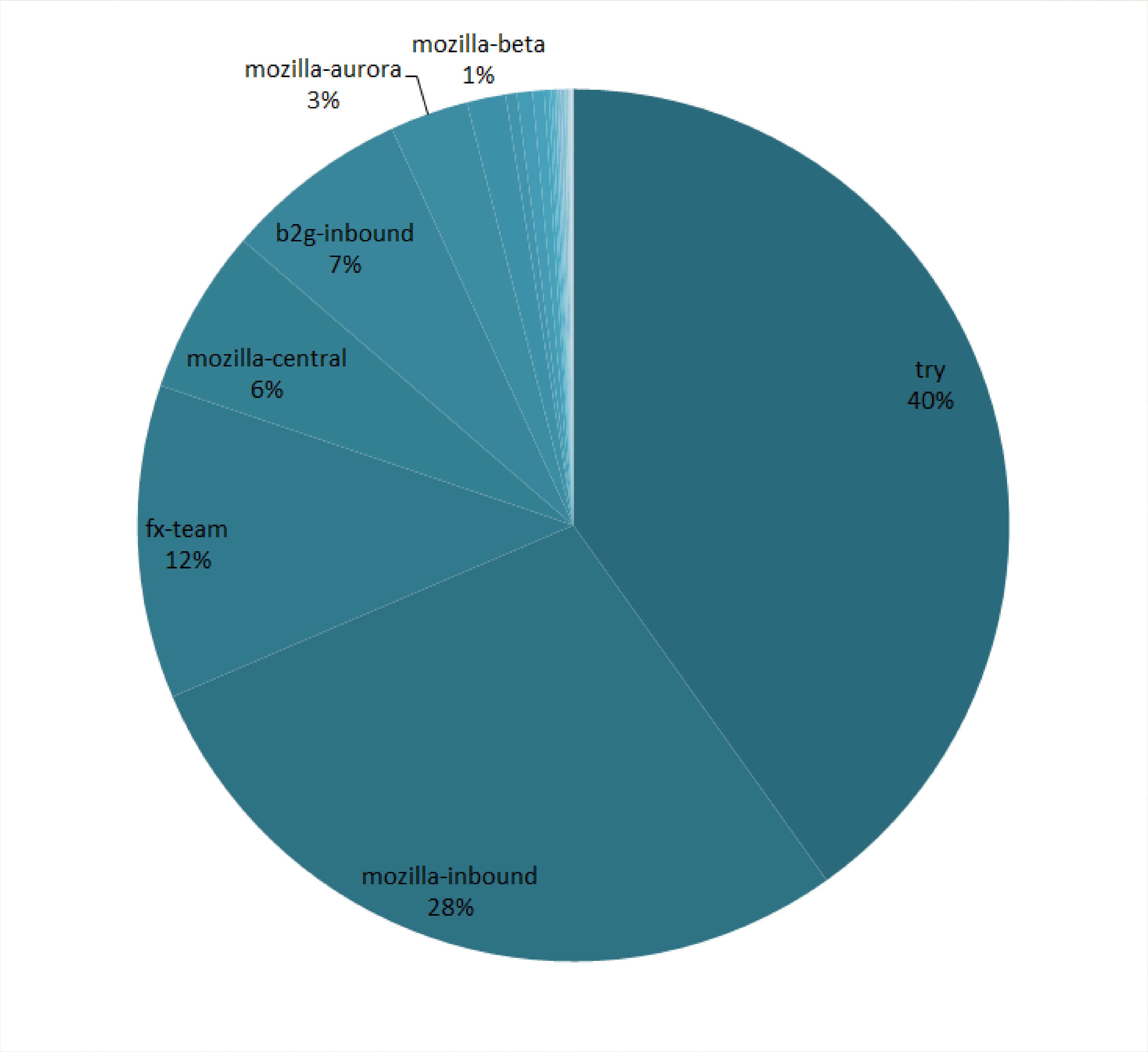


* ActiveData does not include visualization a this time

mozilla

Examples

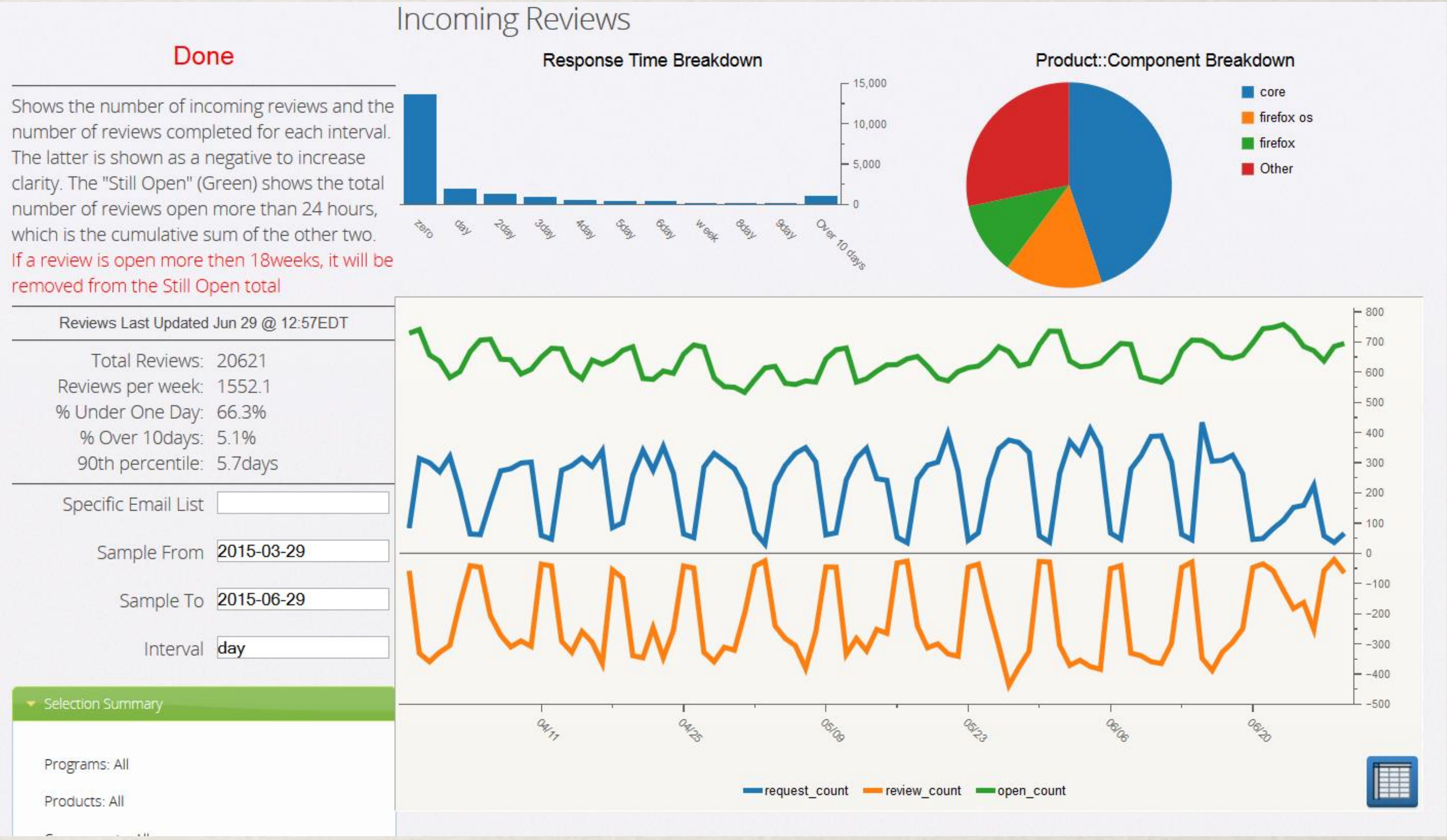
- Identify change in test times
- Test time distributions
- Visualize perf regression
- Fail rate by time into suite
- CPU time by Branch



* ActiveData does not include visualization a this time

Examples

- Identify change in test times
- Test time distributions
- Visualize perf regression
- Fail rate by time into suite
- CPU time by branch
- Reviews over time*



Incoming Reviews

Done

Shows the number of incoming reviews and the number of reviews completed for each interval. The latter is shown as a negative to increase clarity. The "Still Open" (Green) shows the total number of reviews open more than 24 hours, which is the cumulative sum of the other two. If a review is open more then 18weeks, it will be removed from the Still Open total

Reviews Last Updated Jun 29 @ 12:57EDT

Total Reviews:	20621
Reviews per week:	1552.1
% Under One Day:	66.3%
% Over 10days:	5.1%
90th percentile:	5.7days
Specific Email List	
Sample From	2015-03-29
Sample To	2015-06-29
Interval	day

Selection Summary

Programs: All

Products: All

* From MoDevMetrics, the ActiveData precursor

Limitations

- Query language is still limited, complex analysis must be done on client
- Not designed for complex relations: Only transactional data; data with little or no lifecycle; can be modeled well.
- Data is dumb: Must make the effort to explore the results and avoid misinterpreting the data.

More Details

Query Tool	http://activedata.allizom.org/tools/query.html
Service Endpoint	http://activedata.allizom.org/query
Wiki	https://wiki.mozilla.org/Auto-tools/Projects/ActiveData
Code	https://github.com/klahnakoski/ActiveData

Active Data

http://activedata.allizom.org

Kyle Lahnakoski
klahnakoski@mozilla.com