# Dashboarding & A/B Testing


## Oliver Keyes and Mikhail Popov

Analysis Team formerly known as The Swifties

Wikimedia Foundation / Discovery Department

# Dashboards

These contain everything from API usage to direct user interactions, and provide data for internal and external use to see how well we're doing.

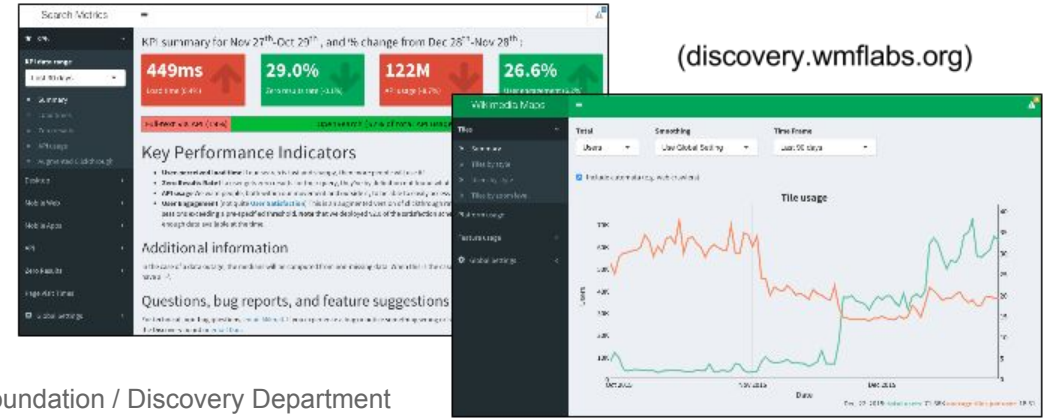
- Search Metrics (http://discovery.wmflabs.org/metrics/)
- Portal Metrics (http://discovery.wmflabs.org/portal/)
- WDQS Metrics (http://discovery.wmflabs.org/wdqs/)
- Maps Metrics (http://discovery.wmflabs.org/maps/)
- External Traffic Referals (http://discovery.wmflabs.org/external/)

# Live demonstration!

(Because those always work when you need them to, right?)


# Discovery Dashboards

R/Shiny-powered dashboards




(discovery.wmflabs.org)

# Data Acquisition & Aggregation

- Scripts on stat1002 (written in any language you'd like!) launched on a schedule.
- Gives us access to EventLogging data, MediaWiki data, search and requests data.
- Can be as computationally intensive as you want (unless you break Erik Z's scripts).
- Can operate on any schedule.

# Our Dashboards

- Built on the “Shiny” web application framework in R - so you can compute and rework values in the dashboard = familiar to analysts.

- Reads data over http from datasets.wikimedia.org, computes on them, generates pretty (interactive) visualisations in JavaScript and HTML.

- Markdown documentation

# Features

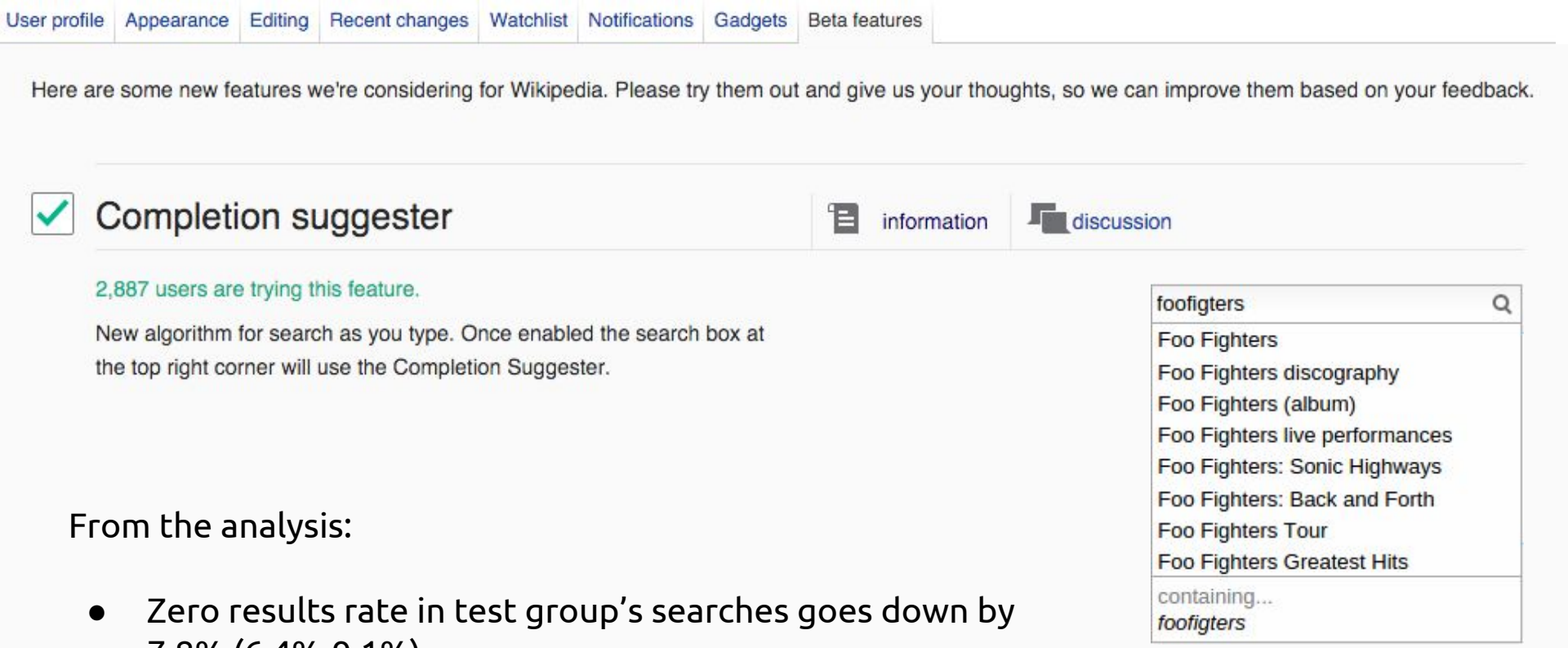
- Reactive contexts (e.g. options to smooth data)
- Easy from an Ops perspective
- Annotations
- Built-in color schemes and themes
- Support for custom CSS/JS (e.g. linking to individual visualizations)
- Support for advanced data manipulation, data tables, CSV export, search

# Ten Dash Commandments

- When writing your collection scripts, *never export percentages*.
- Remember that public means public, so store summary statistics.
- You need a beta version of the dashboards. The darndest things happen.
- Not everything needs to be dashboarded.
- Documentation/descriptions can be written in Markdown.
- Maintenance is a thing.

# A/B Testing

- Data-approach to making changes and adding features (vs “gut feeling”)
- Compare change to status quo
- Measure potential impact:
  - Click-through rate (“conversion rate”)
  - % of searches where 0 results were returned
  - Time spent

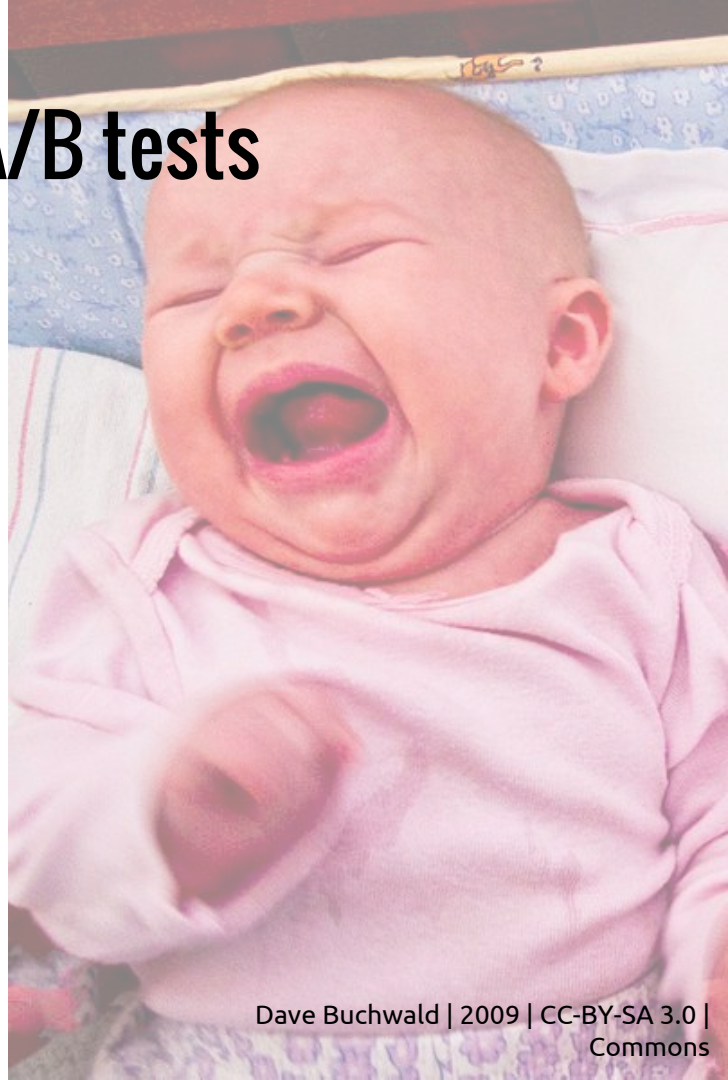



From the analysis:

- Zero results rate in test group's searches goes down by 7.8% (6.4%-9.1%).
- Test group were 1.1-1.2 times more likely to get search results than the control group.

# Lessons learned from our first few A/B tests

- It was an AB/CD test rather than A/B/C/D test.
- We used a traditional hypothesis testing method with sample size of 13.4 million.
  - These tests are usually used in studies with sample sizes of 20-600.
- *p*-value was < 0.001 which is “HIGHLY significant.’
  - p<0.05 is usually “statistically significant.”
- Cohen's w (a measure of effect size) was 0.001
  - 0.1 is considered small in stats literature.



Dave Buchwald | 2009 | CC-BY-SA 3.0 | Commons

# Traditional statistics vs. Bayesian statistics

- Invented for use with paper and pencil.
- Large enough sample size will make tiniest effects “statistically significant”
- Primarily concerned with “p-value”

**Confidence intervals are hard:**

“If we repeat this experiment 100 times, 95 of the 100 intervals we compute will contain the true value. This 4%-6% interval may be one of those 95, or maybe not.”

- Often computationally intensive.
- Can even deal with tiny sample sizes
- Allows for prior knowledge

**Credible intervals are easy:**

“There is a 95% probability the difference between group A and group B is between 4% and 6%.”


BAYESIAN

# Questions?