

Discovery, Big Data, web analytics, and Bayesian A/B testing at Wikipedia



Mikhail Popov (@bearloga)

Data Scientist // Discovery // Wikimedia Foundation

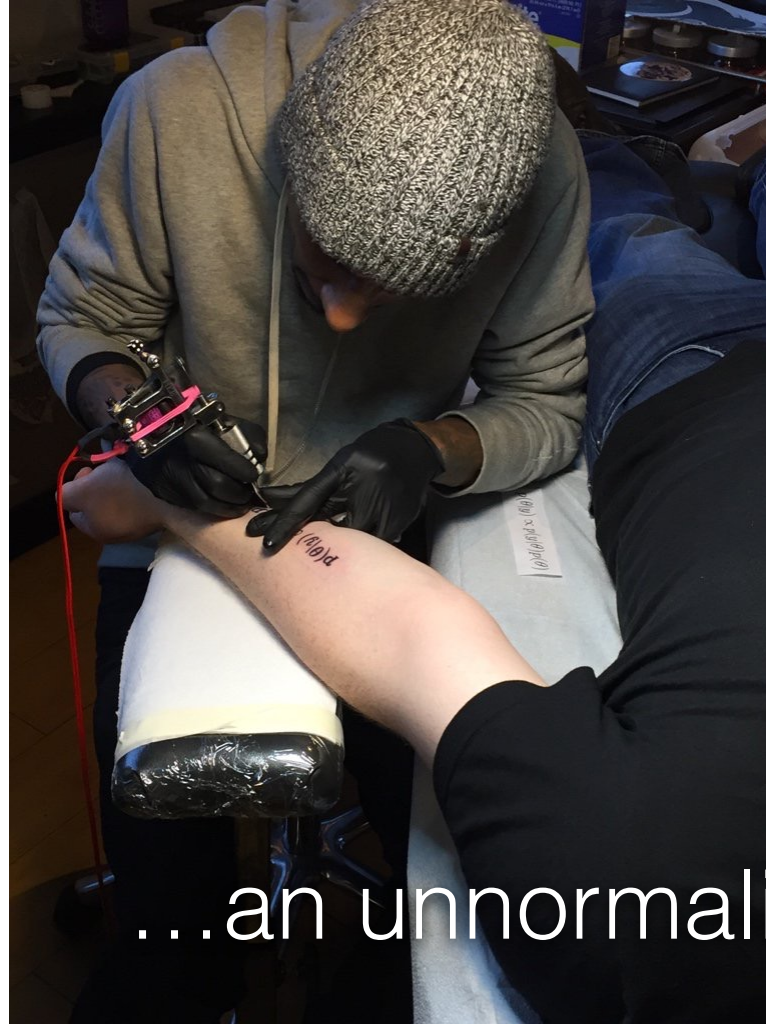


I was/am...

- 🍕 Bachelor in Mathematics (CSUF NSM, 2012)
- 🍕 Master in Statistical Practice (CMU, 2013)
- 🍕 Statistician for a neuropsychology research program at University of Pittsburgh / UPMC for 2 years
- 🍕 Data Scientist / Data Analyst / Quantitative Analyst / Statistician / why do we have so many names??? 🤔
at Wikimedia Foundation
- 🍕 Tattoo designer-typesetter (Wait, what?)

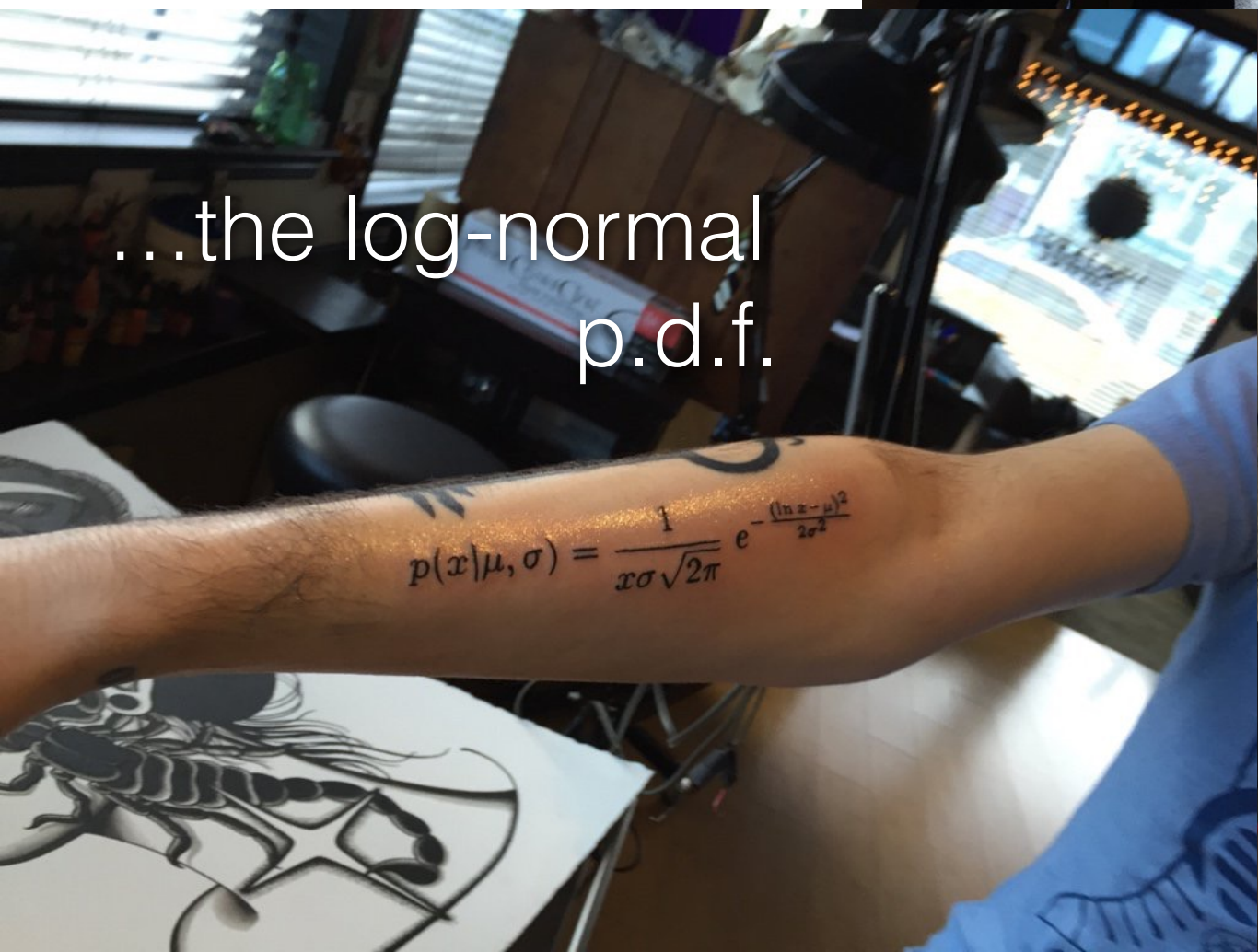
9 January 2016:

Three data science dorks walk into a tattoo parlor in Berkeley, holding LaTeX-typeset formulas for...



...an unnormalized posterior density!

...the log-normal
p.d.f.



...the Poisson
p.m.f.





The Wikimedia Foundation (WMF)

- 🍏 A nonprofit charitable organization.
- 🍏 Committed to creating a world in which every single human being can freely share in the sum of all knowledge (e.g. Wikipedia, Wikipedia Zero).
- 🍏 Collaborates with users around the world.
- 🍏 Develops MediaWiki software used by many organizations and companies (e.g. Intel, Wikia, NASA).
- 🍏 Fights for people's privacy (see *Wikimedia Foundation, et al. v. National Security Agency, et al.* – https://en.wikipedia.org/wiki/Wikimedia_Foundation_v._NSA)

I'M THE TOP
6TH WEBSITE!

Wikimania



Wikipedia

Wikibooks

Commons

Community

Wikisource

Wikiquote

Wikinews

Wikispecies

Wiktionary

MediaWiki

Wikiversity

Incubator





Discovery Department

Building the anonymous path of discovery to a trusted and relevant source of knowledge.






Projects:

- 🔍 Search feature & APIs
- 🔍 Wikipedia Portal (wikipedia.org)
- 🔍 Maps, in collaboration with OpenStreetMap
- 🔍 Wikidata Query Service

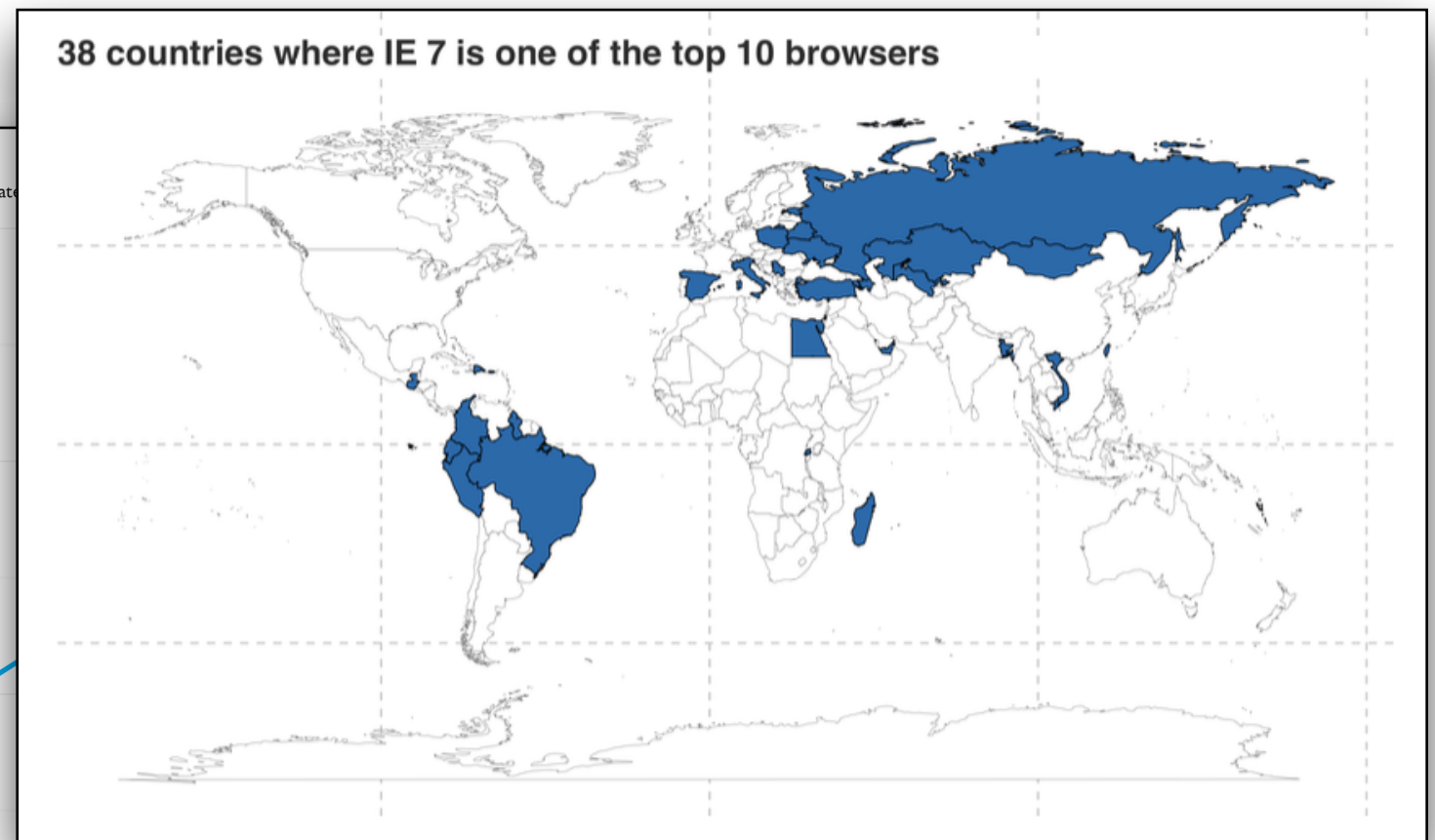
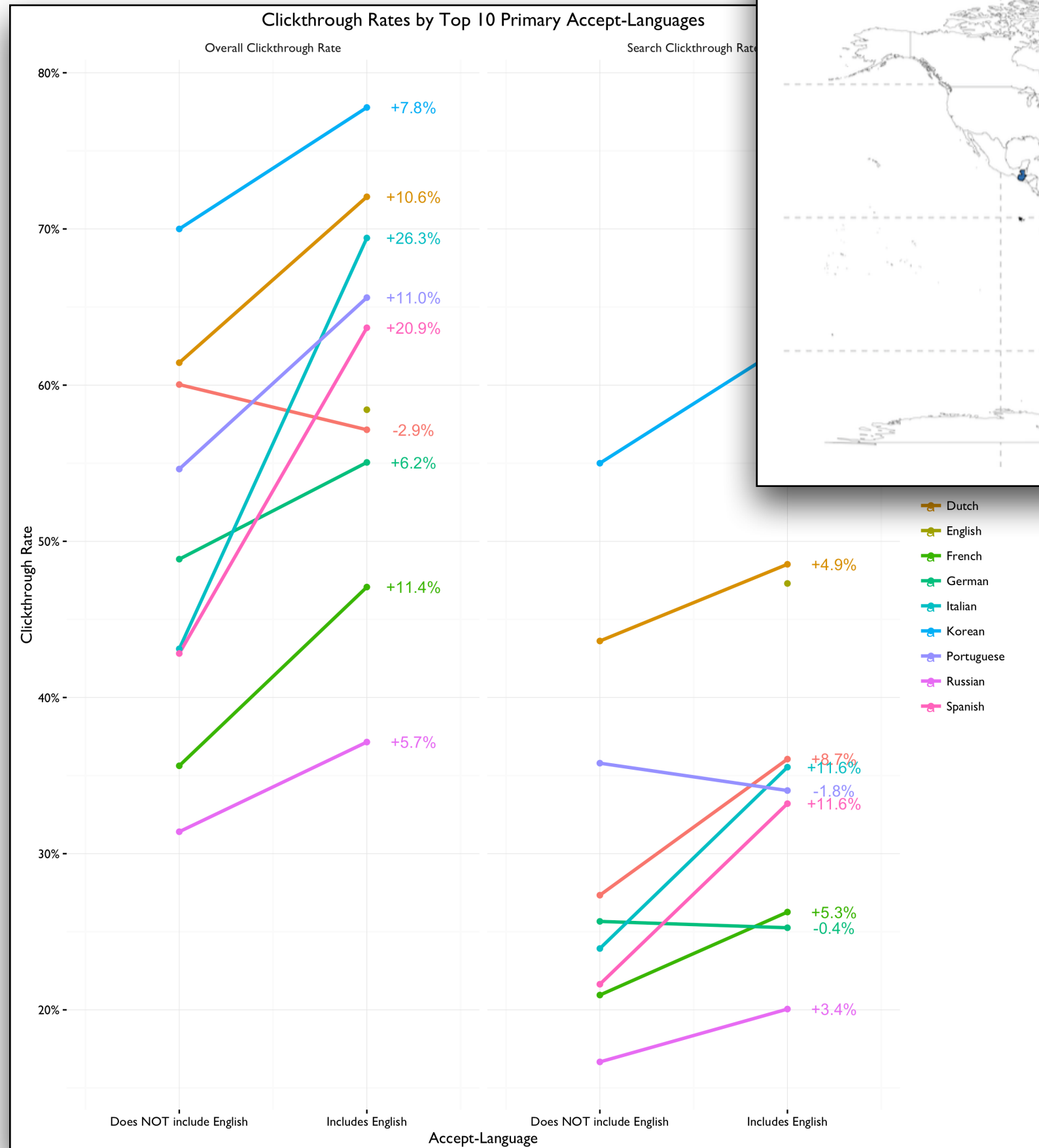
The screenshot shows the Wikipedia search interface. At the top, there are language links for Italiano, Portuguese, and Chinese. Below these, a search bar contains the text 'Beyoncé' and a dropdown menu set to 'EN'. To the right of the search bar is a magnifying glass icon. Below the search bar, the search results are displayed. The first result is 'Beyoncé', described as 'American singer, actress and song writer, American singer, actress and song writer'. The second result is 'Beyoncé discography', described as 'discography'. The third result is 'Beyoncé (album)', described as 'fifth studio album by American singer and songwriter Beyoncé Knowles'. To the left of the search results, there are language links for Français, Hō-ló-oē, and Eest. To the right, there are language links for Binisaya, Eest, Eλληνι, and Bahasa Minang.



Discovery's Analysts

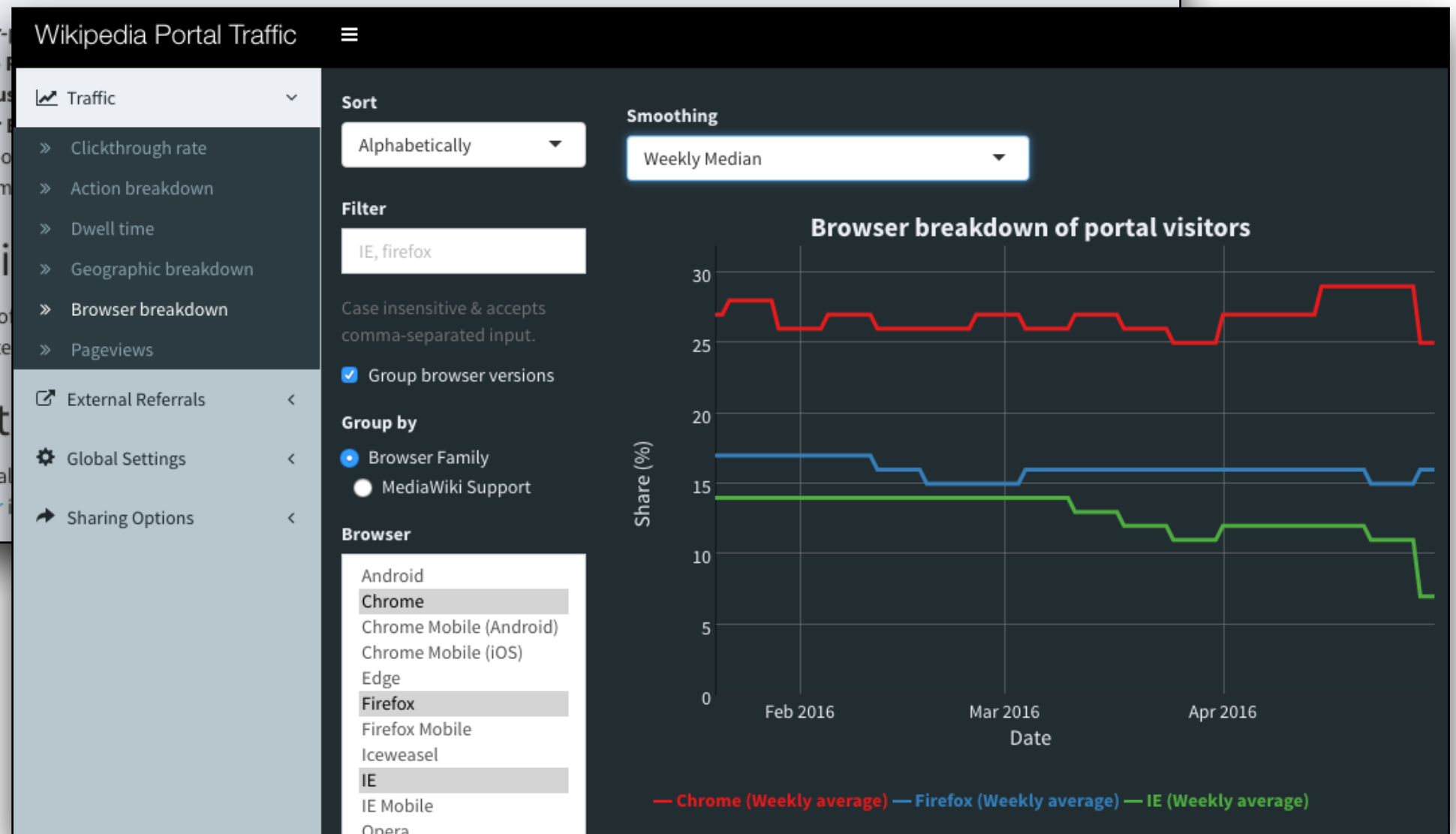
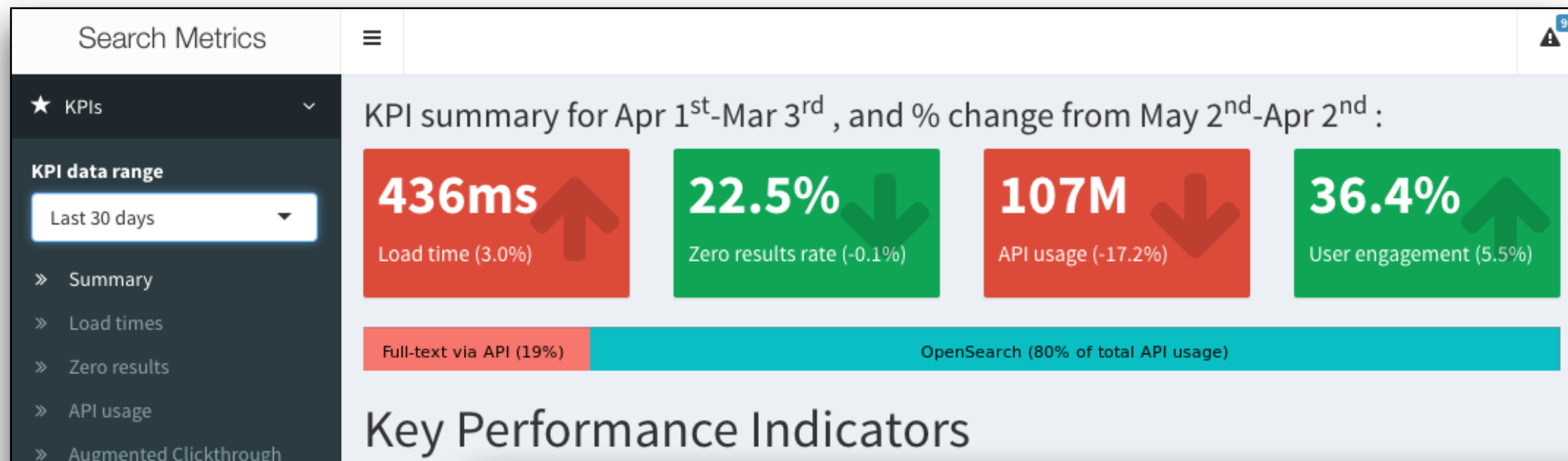
-  Integrated within department instead of separate dep't
-  Provide ad-hoc analyses and reports as needed
-  Build and maintain dashboards for tracking *key performance indicators* (KPIs) and other metrics
-  Consult with teams in design of experiments – “A/B tests” – e.g. narrowing down affected population
-  Work with engineers to design event logging schemas

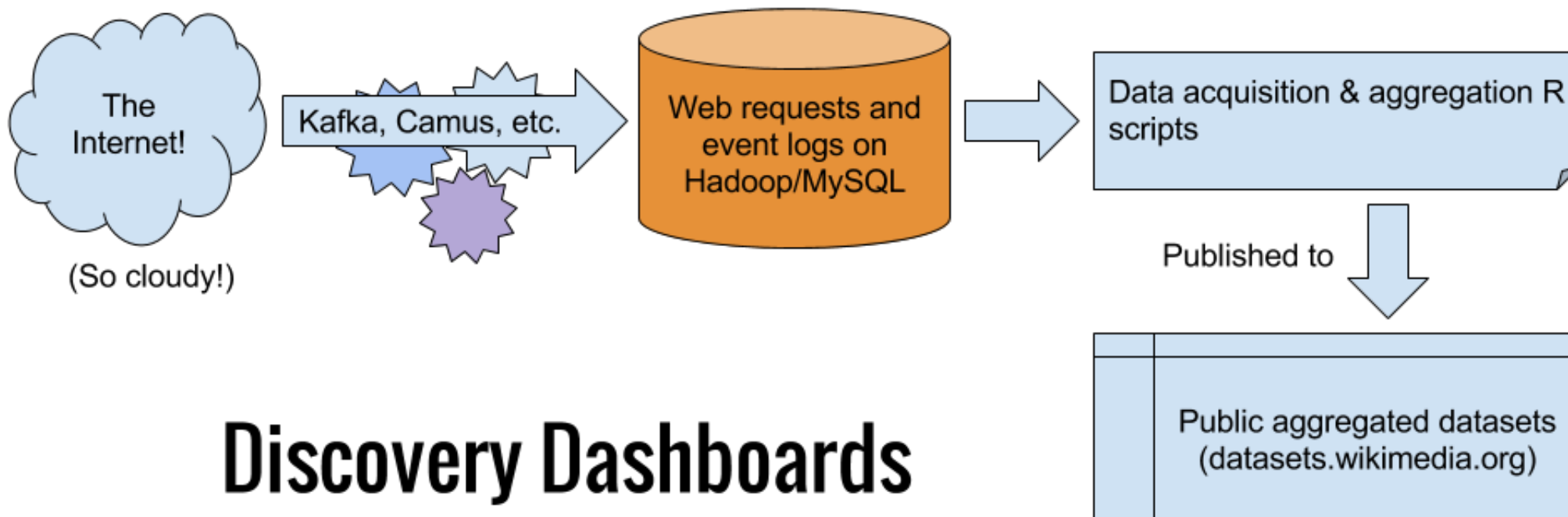
Ad-hoc analyses include...



↑ Global browser usage and JavaScript support.

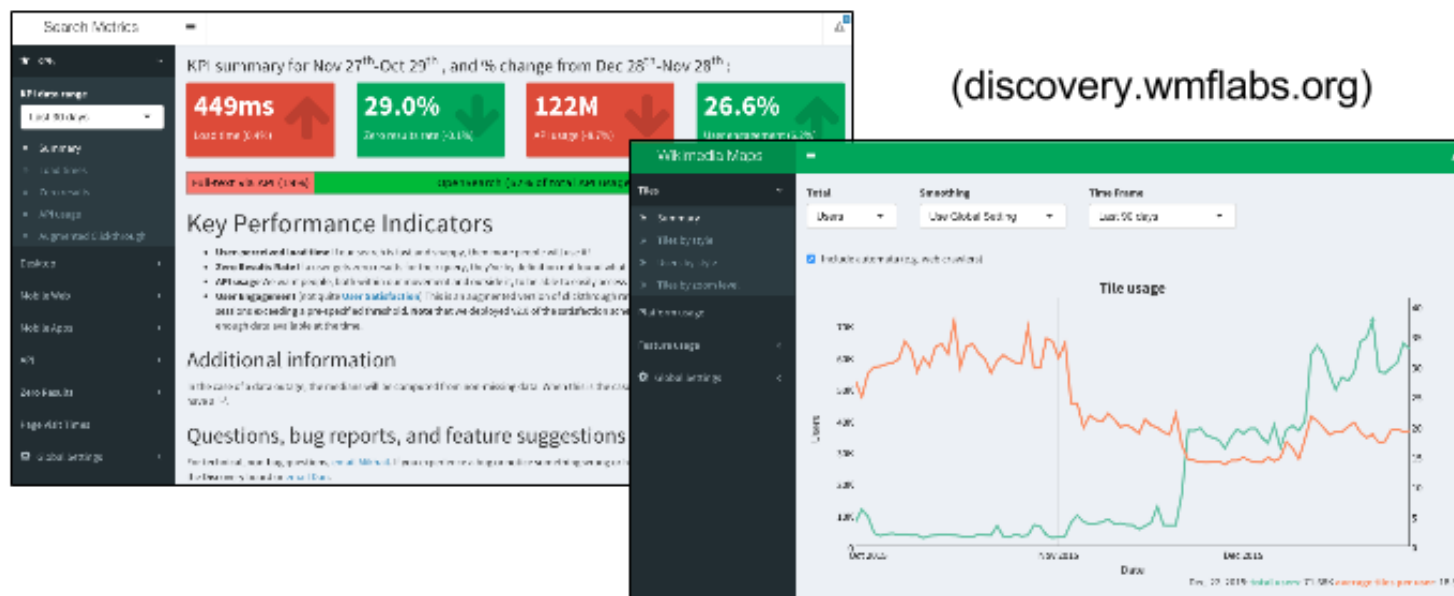
← Engagement with wikipedia.org based on whether your list of preferred languages includes English or not.





Discovery Dashboards

R/Shiny-powered dashboards



Database-to-Dashboard Pipeline



Data & Technologies

Web & Search Requests

- ✓ *MapReduce* with Hadoop Distributed File System (HDFS)
- ✓ Kafka (log buffer) → HDFS via LinkedIn's Camus pipeline
- ✓ Includes IP addresses, request referrers, user agents, queries
- ✓ Retrieve and aggregate data with HiveQL and User-defined functions (UDFs) written in Java

Event Logging (User Actions)

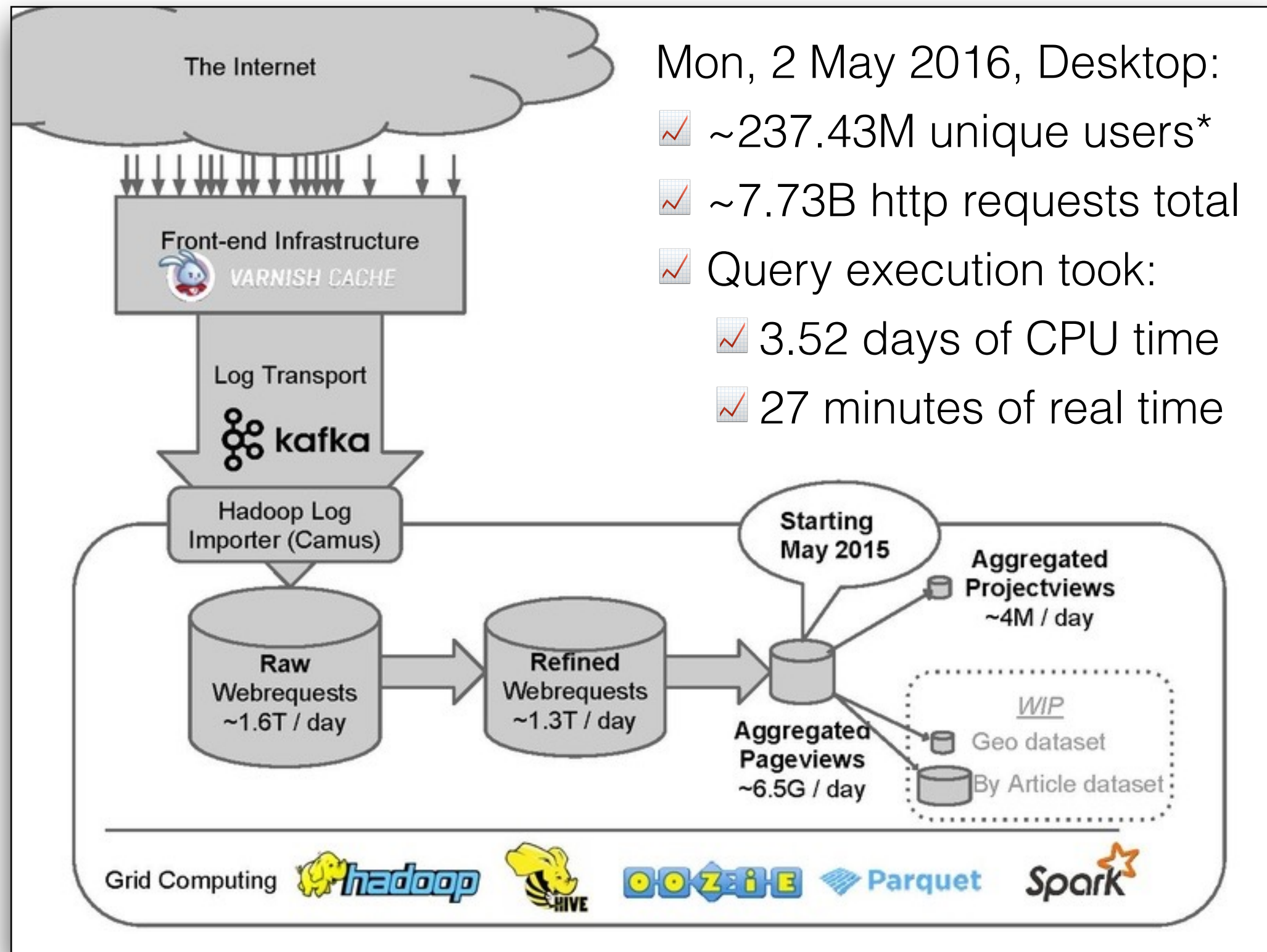
- ✓ JavaScript event triggers (e.g. clickthroughs, edits, pings)
- ✓ Kafka → MariaDB (MySQL)
- ✓ Varying degrees of privacy (anonymized → containing PII) & sampling (all vs. 1 in n users)
- ✓ Retrieve and aggregate using Structured Query Language



Event Logging Volume

- Only 0.5% of sessions (users) get selected for anonymous tracking
- Monday, 2 May 2016, on Desktop: 23K sessions (users) recorded

| event | total events | average per user |
|-----------------------|--------------|------------------|
| search results page | 226K | 9.93 |
| opened a result | 2.8K | 1.37 |
| “I’m alive!” check-in | 14K | 8.04 |
| clicked | 22K | 1.73 |





My team only uses R

Development: devtools, testthat, roxygen2, Rcpp, RStudio + GitHub

Data Manipulation & Analysis

- 🔧 Data wrangling: readr, dplyr, tidyr, data.table, lubridate, xts
- 🔧 Web analytics: urltools, rgeolocate, iptools, uaparser
- 🔧 Bayesian analysis of categorical data: BCDA

Data Visualization & Reporting

- ✍️ ggplot2 + ggthemes + ggally
- ✍️ RMarkdown + knitr for reproducible reporting
- ✍️ shiny + shinydashboard + dygraphs (for time series)
- ✍️ All code (but not data) is public at github.com/wikimedia-research

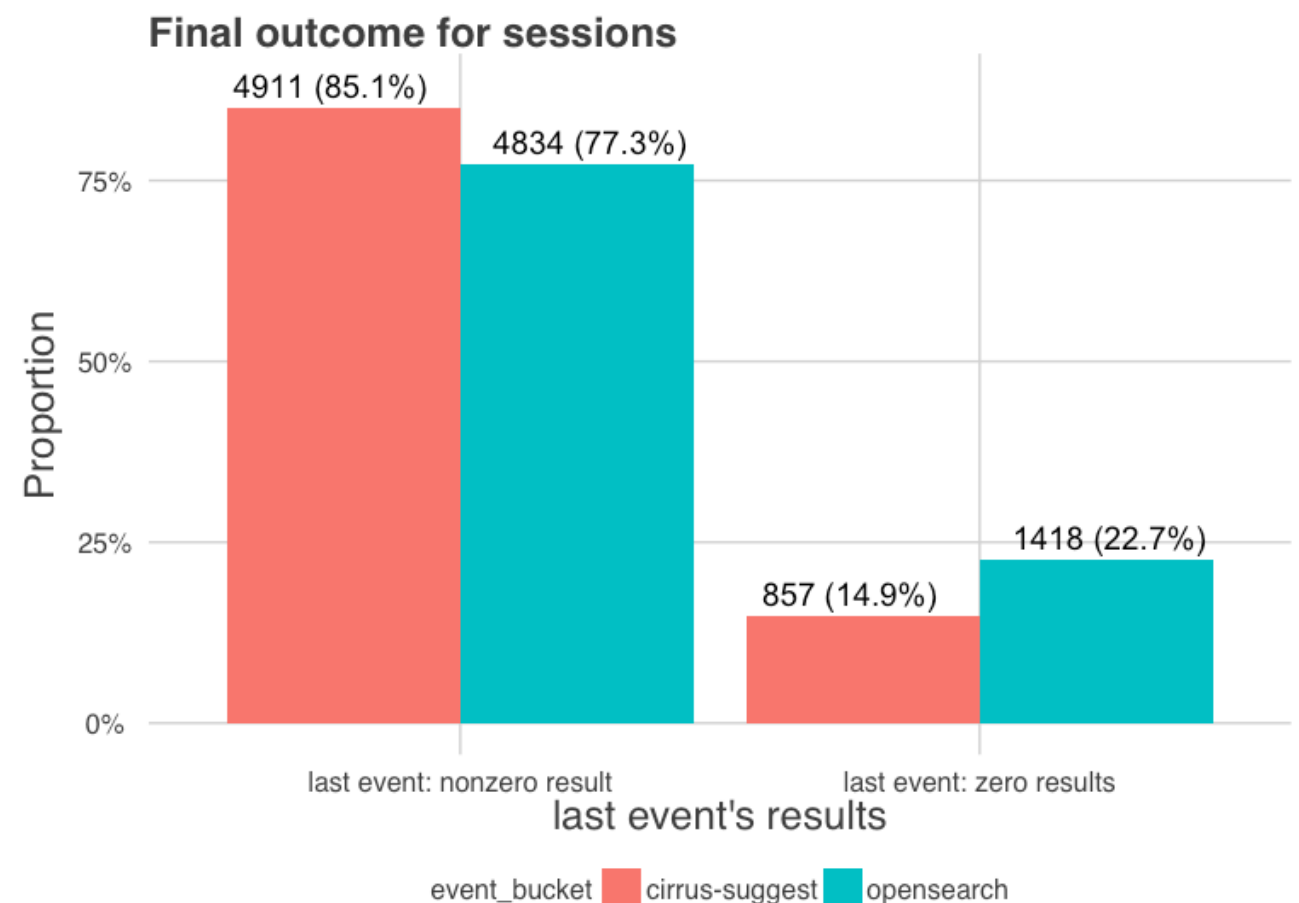


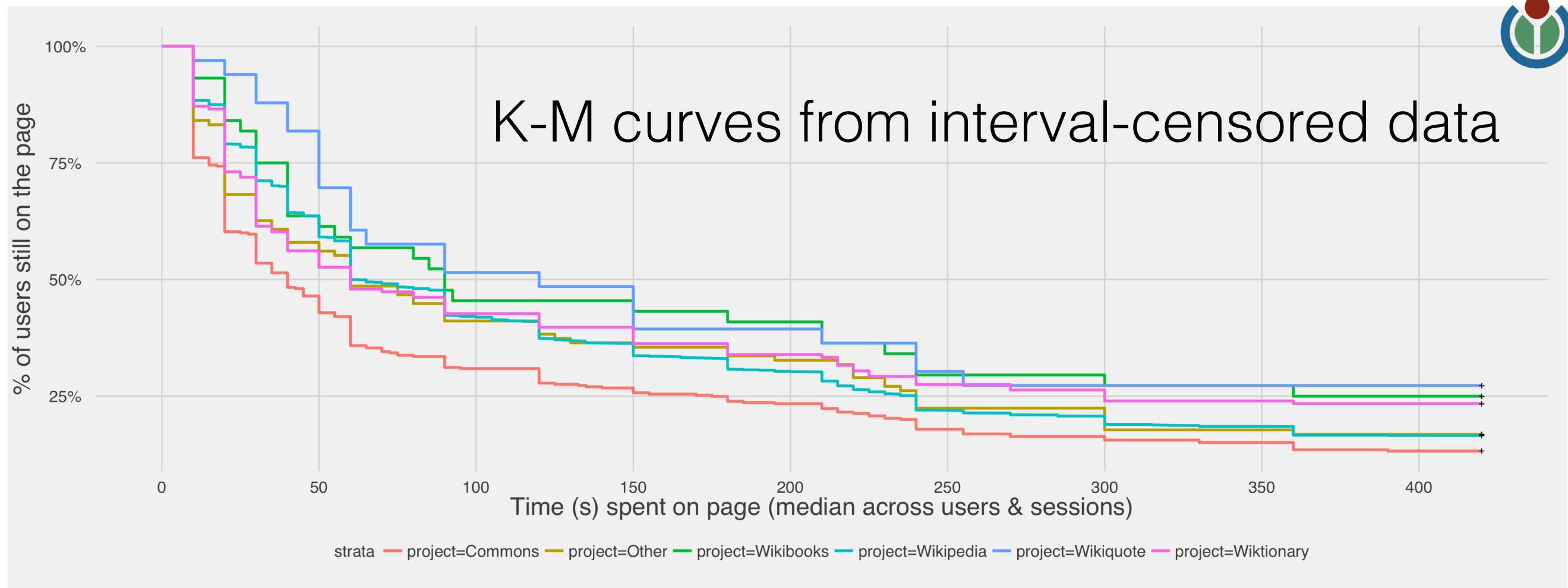
Testing Completion Suggester's Impact on Zero Results Rate



- 🔍 Tolerance for typos and stop words
- 🔍 Fewer typos suggested (e.g. searching for *Airton Senna* will properly suggest *Ayrton Senna*)

- 🔍 Used Bayesian methodology (Beta-Binomial model) to compute difference, relative risk, and credible intervals.
- 🔍 Deployed as opt-in beta → deployed to production!





Human-Computer Interaction Research: User Satisfaction

Scenario 1: Find out Barack Obama's birth date.

Scenario 2: Learn about Black Lives Matter movement.

Scenario 3: I search for Lin-Manuel Miranda's *Hamilton*, stay there for 5s, and then decide I want to read about Aaron Burr.

Scenario 4: I search a vague term with no specific task in mind.



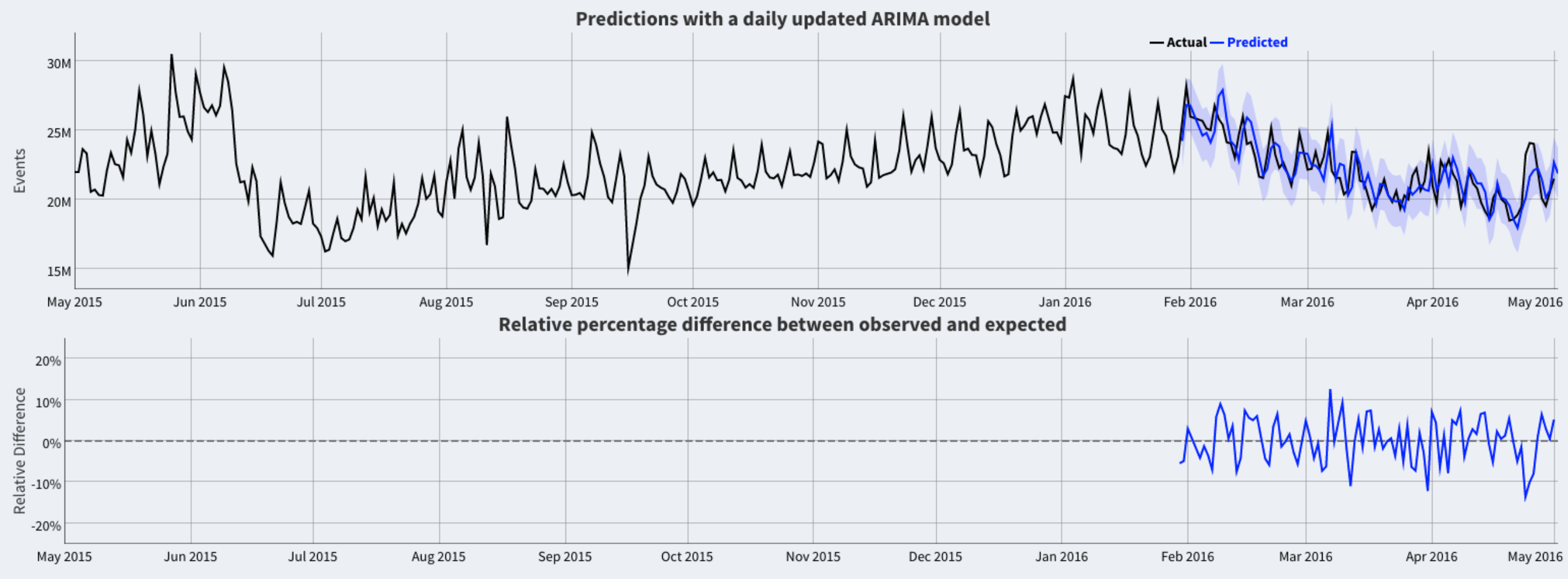
Research: daily forecasts to detect anomalies & abuse of service

5.19% less requests than expected

Relative % difference for 'yesterday' (2016-05-01)

~21.9M (20.07M-23.73M)

Expected requests and 80% Confidence Interval for 'today' (2016-05-02)



Prototype using ARIMA: <http://discovery-experimental.wmflabs.org/forecast/>

Future work will probably focus on Bayesian Structural Time Series Modeling



Thank you!

Got additional questions?

Work email:

`mpopov@wikimedia.org`

Personal email:

`mikhail@mpopov.com`

Twitter: @bearloga

We're hiring data analysts/scientists!

See the jobs page at WMF:

https://wikimediafoundation.org/wiki/Work_with_us

