Discovery, big data analytics, and Bayesian A/B testing at Wikipedia

Mikhail Popov
Data Scientist // Discovery // Wikimedia Foundation

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)
- collaborates with users around the world
- develops MediaWiki software used by many organizations and companies

I'M THE TOP 7TH WEBSITE!

Wikipedia

Wikimania

wikivoyage

Wikibooks

Commons

Community

Wikisource

Wikiquote

WIKIMEDIA

Wikinews

Wikispecies

Wiktionary

MediaWiki

Wikiversity

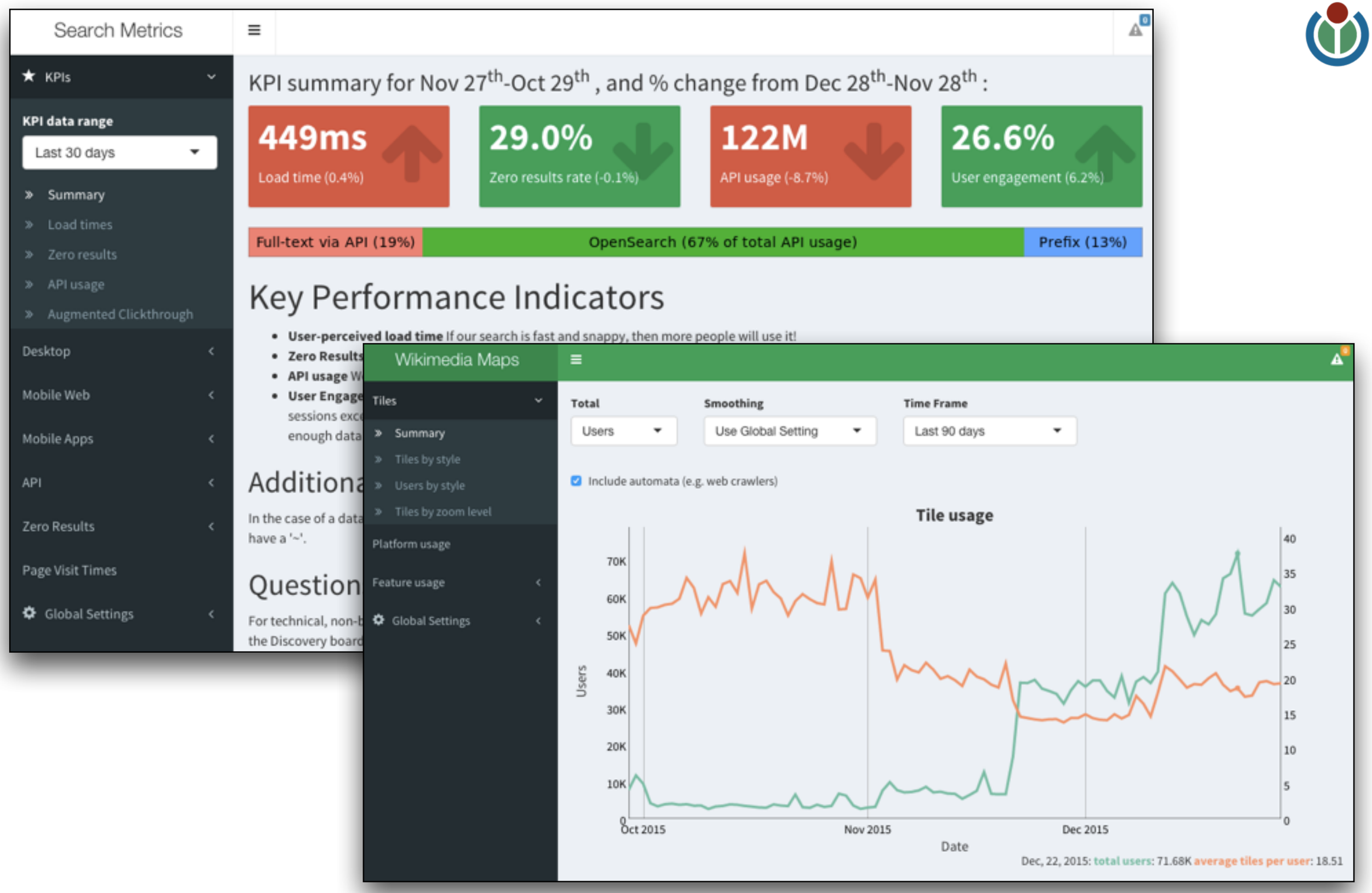
Incubator

WIKIDATA

Discovery Department

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

- **Projects**:
  - Search feature and APIs
  - wikipedia.org portal
  - Maps, in collaboration with OpenStreetMap
  - Wikidata Query Service



Analysis as support: data retrieval & visualization via dashboarding.

Data & Technologies

Web Requests (Search)

- *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

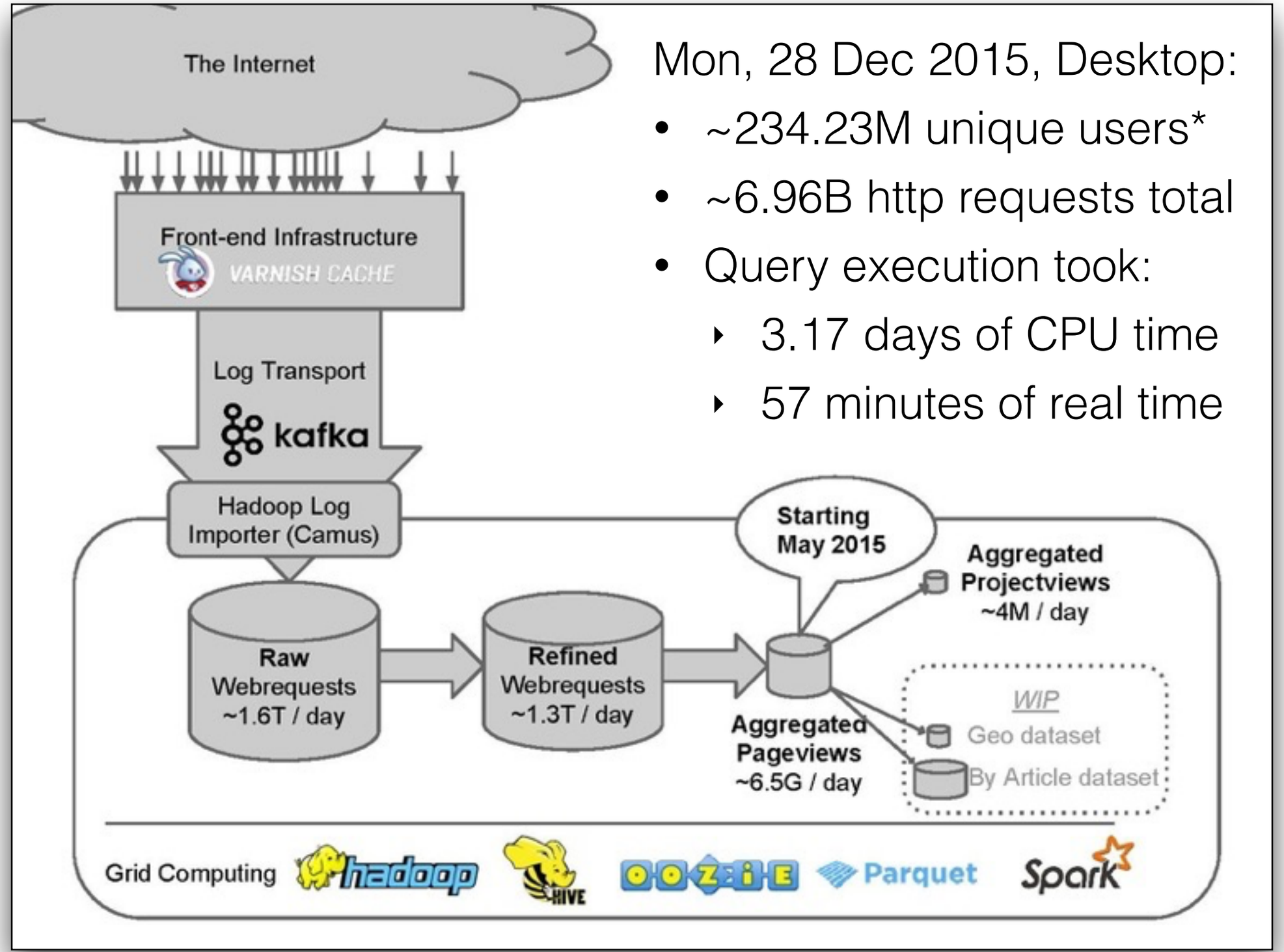
Event Logging (User Actions)

- 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 0.1% of anonymous users recorded
- Mon, 28 Dec 2015, Desktop:

action	total events	average per user	total users
click-result	2.5K	1.01	2.5K
impression-results	27.1K	4.72	5.8K
session-start	54.6K	8.58	6.4K
submit-form	3.9K	1.01	3.9K



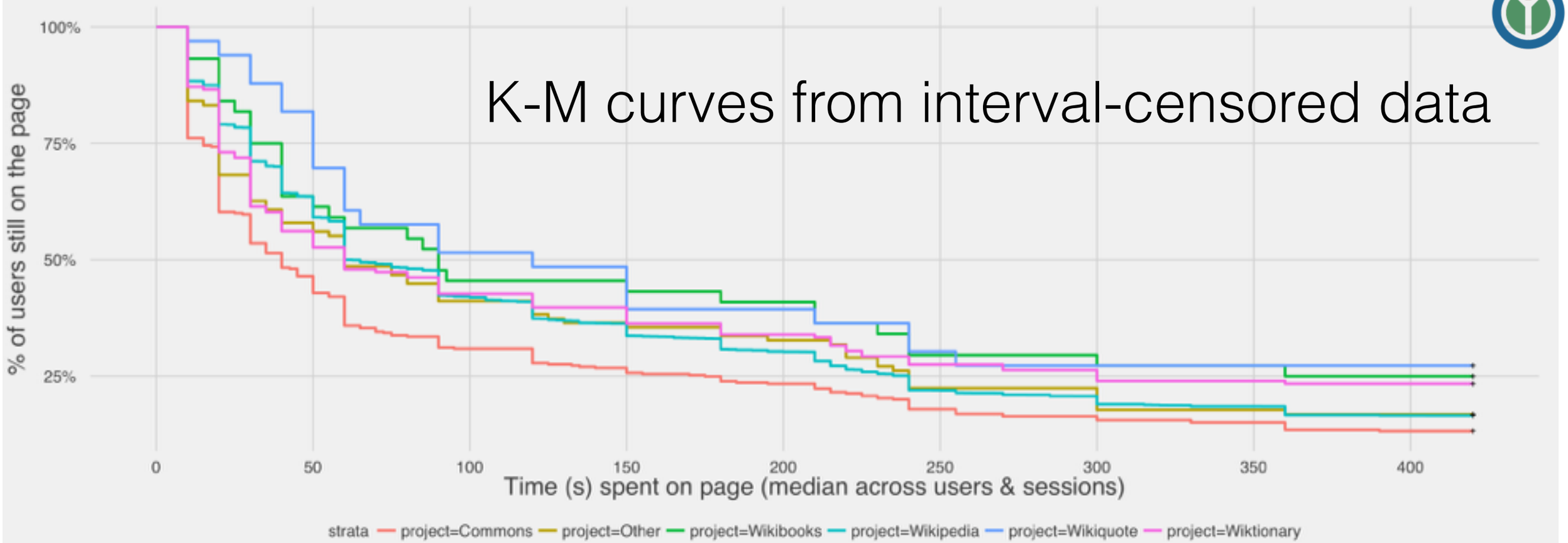
Mon, 28 Dec 2015, Desktop:

- ~234.23M unique users*
- ~6.96B http requests total
- Query execution took:
  - 3.17 days of CPU time
  - 57 minutes of real time

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*)

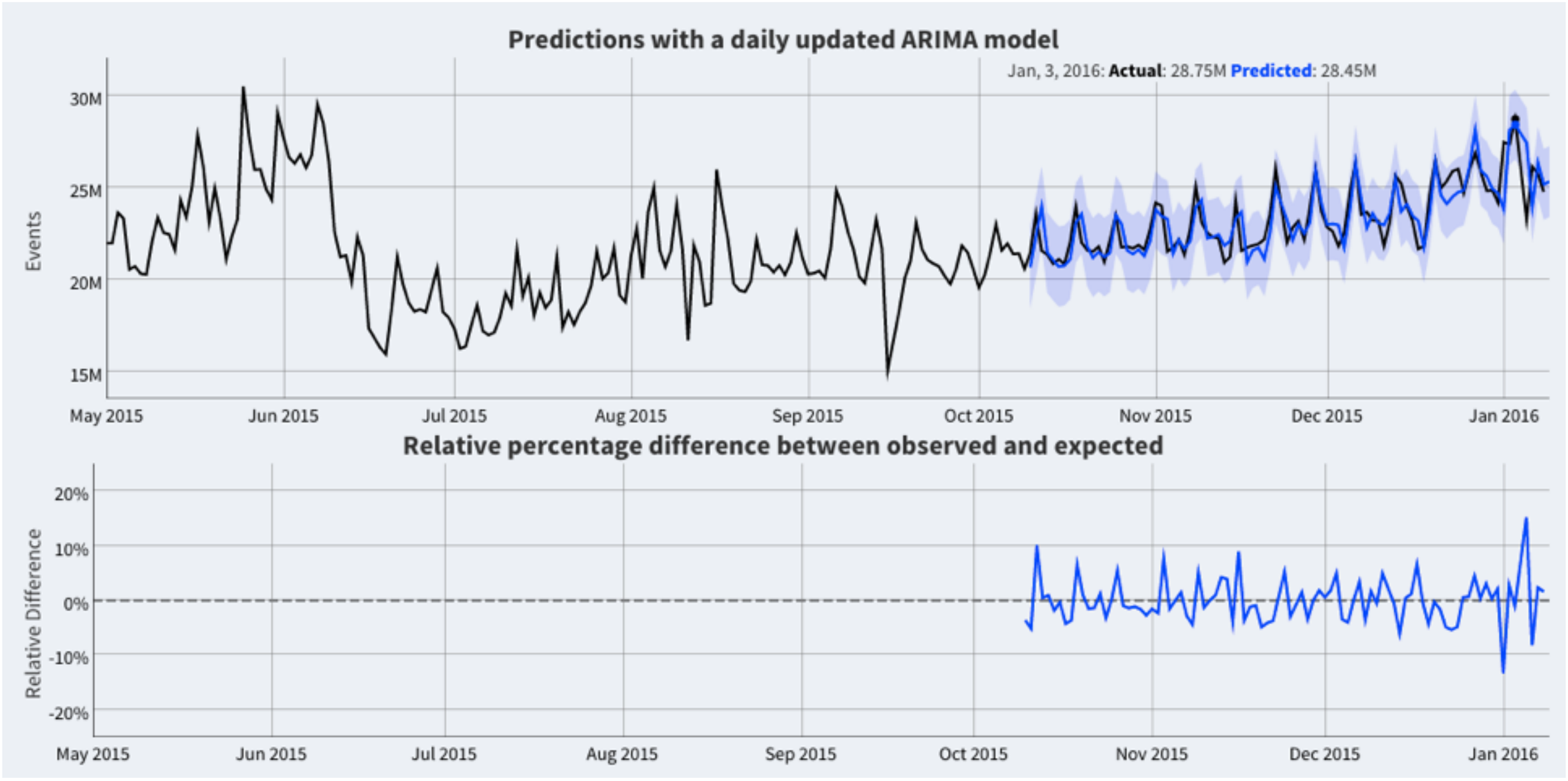
- Test group's zero results rate goes down by 7.8% (95% Credible Interval: 6.4–9.1%)
- Test group is 1.1-1.2 times more likely to get results
- Deployed as a Beta feature



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



Live: `http://discovery-experimental.wmflabs.org/forecast/`

Codebase: `https://github.com/bearloga/wmf-delphi`

Thank you!

Got additional questions?

Work email:
`mpopov@wikimedia.org`

Personal email:
`mikhail@mpopov.com`

Twitter: @bearloga

Jobs at WMF: `https://wikimediafoundation.org/wiki/Work_with_us`