

Data Science Workflow at Discovery

Chelsy Xie
May 2nd, 2017



WIKIMEDIA
FOUNDATION

Introduction

Discovery Department

Make the wealth of knowledge and content in the Wikimedia projects easily discoverable.

Project:

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



The screenshot shows the Wikipedia homepage with the following data:

Language	Article Count
English	5 381 000+ articles
Español	1 328 000+ artículos
Русский	1 386 000+ статей
Français	1 860 000+ articles
Português	965 000+ artigos
中文	935 000+ 條目
日本語	1 056 000+ 記事
Deutsch	2 051 000+ Artikel
Italiano	1 348 000+ voci
Polski	1 216 000+ haseł

Below the language statistics is a search bar with the text 'R programming' and a dropdown menu set to 'EN'. The search results show 'R (programming language)' with the description 'programming language for statistical computing'.

Data Analysts at Discovery

- 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), then analyze and report the results
- Work with engineers to design & implement event logging schemas

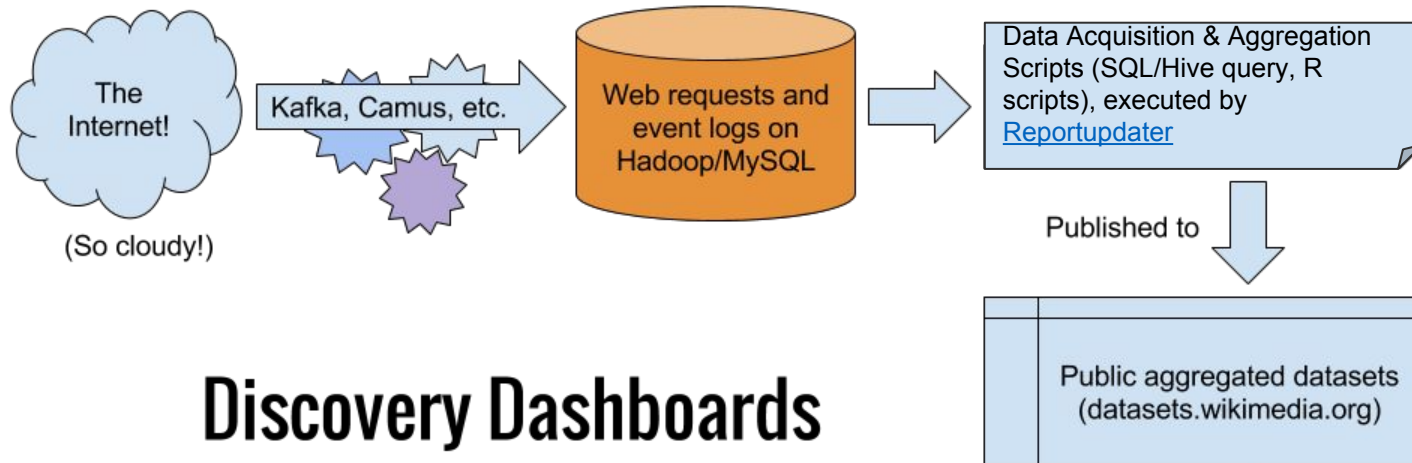
We use R! 🧐

Dashboards

Shiny Dashboards

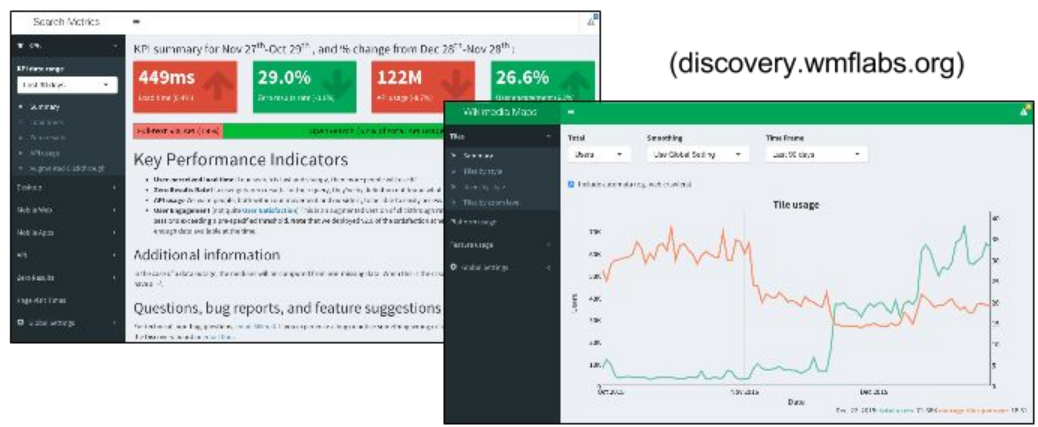
The dashboards contain everything from API usage to direct user interaction, and provide data for internal and external use to see how well we are doing.

- Search Metrics (<https://discovery.wmflabs.org/metrics/>)
- Portal Metrics (<https://discovery.wmflabs.org/portal/>)
- Wikidata Query Service Usage (<https://discovery.wmflabs.org/wdqs/>)
- Wikimedia Maps Metrics (<https://discovery.wmflabs.org/maps/>)
- Externally-referred Traffic (<https://discovery.wmflabs.org/external/>)



Discovery Dashboards

R/Shiny-powered dashboards

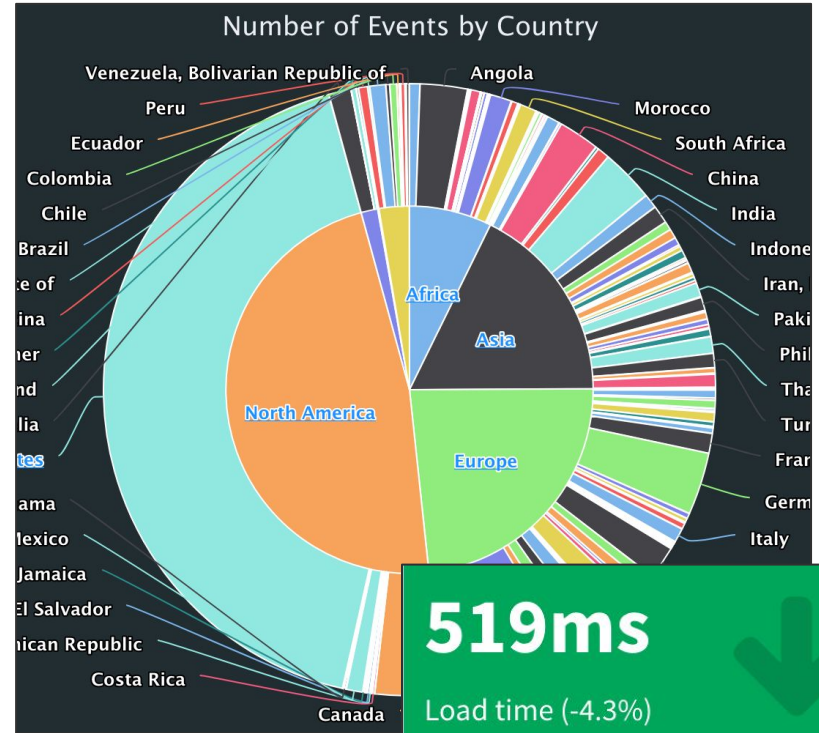


Interactive Graphs

- Time series with dygraph package
- DataTables with DT package
- Pie charts with Highcharter package
- Sparklines with sparkline package
- Example:

<https://discovery.wmflabs.org/portal/>

[#all_country](#)



Shiny is powerfully interactive!

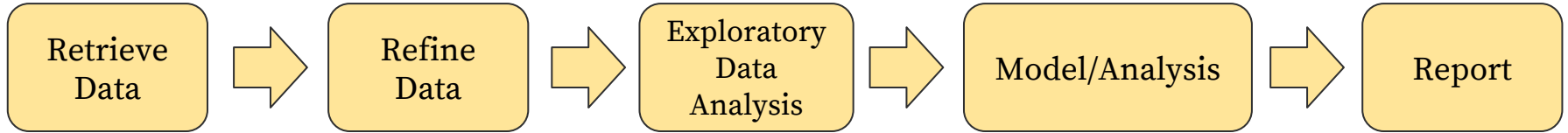
Shiny combines the computational power of R
with the interactivity of the modern web.



WIKIMEDIA
FOUNDATION

Research and Testing

Workflow and Packages



Packages:

readr, wmf*

dplyr, tidyr,
data.table,
lubridate, xts

ggplot2,
ggthemes,
ggally,
cowplot

binom, conting,
BCDA, bsts,
randomForest

RMarkdown,
knitr

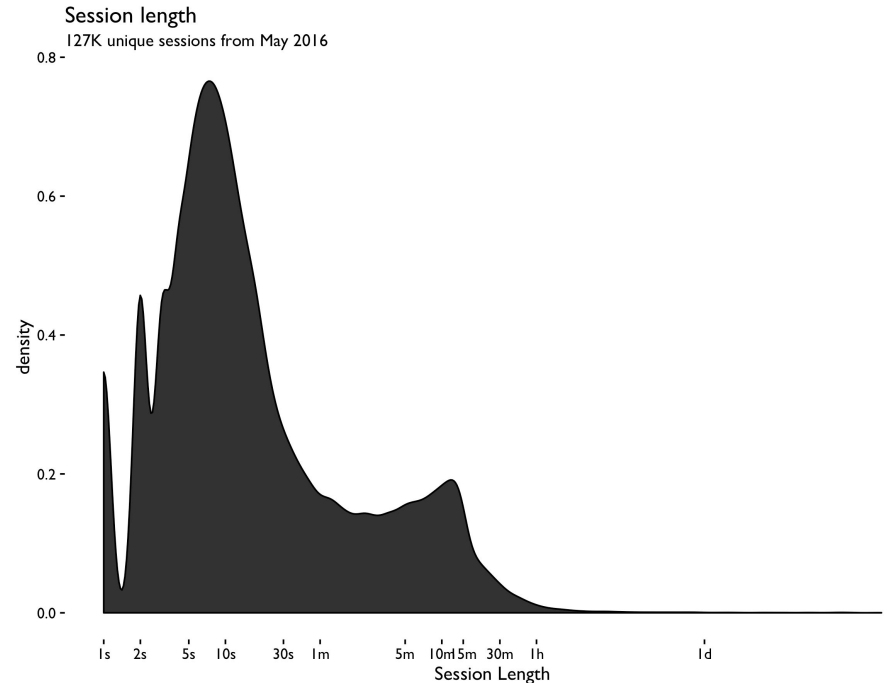


WIKIMEDIA
FOUNDATION

* our open source package for querying Hive & MySQL DBs internally

How long do users stay on Wikipedia.org?

- The most common session length is approximately 10 seconds.
- The majority of the sessions are shorter than 1 minute.
- Around 80% of the English-using visitors' sessions are shorter than 1 minute, and same for United States visitors, while only 45% of the Russian-using visitors' sessions are shorter than that.



Report: <https://git.io/v9tSq>



WIKIMEDIA
FOUNDATION

Should we implement a new search ranking function?

https://wikimedia-research.github.io/Discovery-Search-2ndTest-BM25_jazhth/

Metrics:

- Zero results rate (% of searches w/o results)
- PaulScore*
- Clickthrough rate
- Position of first clicked result
- Dwell-time per visited page
- Scroll
- Query reformulation

Test group has...

- Lower zero results rate
- Worse PaulScore
- Worse clickthrough rate
- Fewer users clicked on the first result first

Conclusion: We are showing test group users worse results 😞



WIKIMEDIA

* click-based measure of relevance, for more details see https://www.mediawiki.org/wiki/Wikimedia_Discovery/Search/Glossary#PaulScore

**Mastering tools for
common tasks
improve the
efficiency of our work.**



WIKIMEDIA
FOUNDATION



THANK YOU



WIKIMEDIA
FOUNDATION

This work is licensed under the Creative Commons Attribution-ShareAlike 4.0 International License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/4.0/> or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.