Data Science Workflow at Discovery

Chelsy Xie
May 2nd, 2017
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
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/](https://discovery.wmflabs.org/metrics/))
Data Acquisition & Aggregation Scripts (SQL/Hive query, R scripts), executed by Reportupdater.

Published to

Public aggregated datasets (datasets.wikimedia.org)

Discovery Dashboards

R/Shiny-powered dashboards

(discovery.wmflabs.org)
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.
Research and Testing
Workflow and Packages

Retrieve Data → Refine Data → Exploratory Data Analysis → Model/Analysis → Report

Packages:
- readr, wmf*
- dplyr, tidyr, data.table, lubridate, xts
- ggplot2, ggthemes, ggally, cowplot
- binom, conting, BCDA, bsts, randomForest
- RMarkdown, knitr

* 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
Should we implement a new search ranking function?


**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 😞

* 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.
THANK YOU