moz://a

Heavy Users

Identification, Usage Patterns and Attributes

3.1.2019

Kris Thayer
Outreachy Intern

Who I Am

- Outreachy Intern in Product Data Science
- Love math and numbers
- Engineering major in college
- Software developer
- Raising children
- Data scientist

What's Outreachy?

Internships in Free and Open Source Software (FOSS)

- Support people from groups underrepresented in tech
- Projects in:
  - programming
  - user experience
  - documentation
  - illustration
  - graphical design
  - data science
- 46 interns at 18 organizations
- 11 interns at Mozilla
- Software Freedom Conservancy

My Outreachy Project

Heavy Users

- Definition
- Attributes
- Usage Patterns
- Drive decisions about features
- Promotion retention and growth

Prior Work on Heavy Users

Proposed definition of heavy users

2016

- Brendan Colloran, Strategy and Insights team
- Wanted page views
- Proposed session hours
- Decided active ticks
- Defined for a day
- Top active 10% of total active ticks for 28 day window
- 28 days
  - aligns with Mau
  - no undue lag
  - no undue churn and bounciness

Proposed definition of heavy users

2017

- Product Data Science Team
- 90th percentile or above
- Sum of active ticks
- 28 days
- Defined as of a day

Attitudes of Heavy Desktop Users

2018

- Rosanne Scholl , Firefox Strategy & Insights team
- Combined survey and telemetry data
- Identified by hours of use, URI count and search count
- Range of heaviness retained
- Hours of use from subsession hours
- Measurements for 7 days
- Averaged over a week for days with values
- Less sensitive to weekly patterns and volatility

What is a Heavy User?

- Usage patterns beyond a normal, average use
- Online media, websites, apps, digital services
- Dr. Dik Warren Twedt coined “heavy-half” in 1964
- 80/20 rule

Exploratory Data Analysis

Initial exploration

- Determine a cutoff
- Data in main_summary table
- Databricks notebooks
- Week of data in September to avoid summer and major holidays
- 1% sample of client data with sample id 42

Metrics of Interest

- Intensity, time and revenue impact
- URI, subsession hours, active hours and search count
- URI and hours:
  - main_summary table
    - individual pings
    - daily totals
    - weekly averages of daily totals
- Search count:
  - search_clients_daily table
    - daily totals
    - weekly averages of daily totals

Active Hours vs Subsession Hours

Active Hours

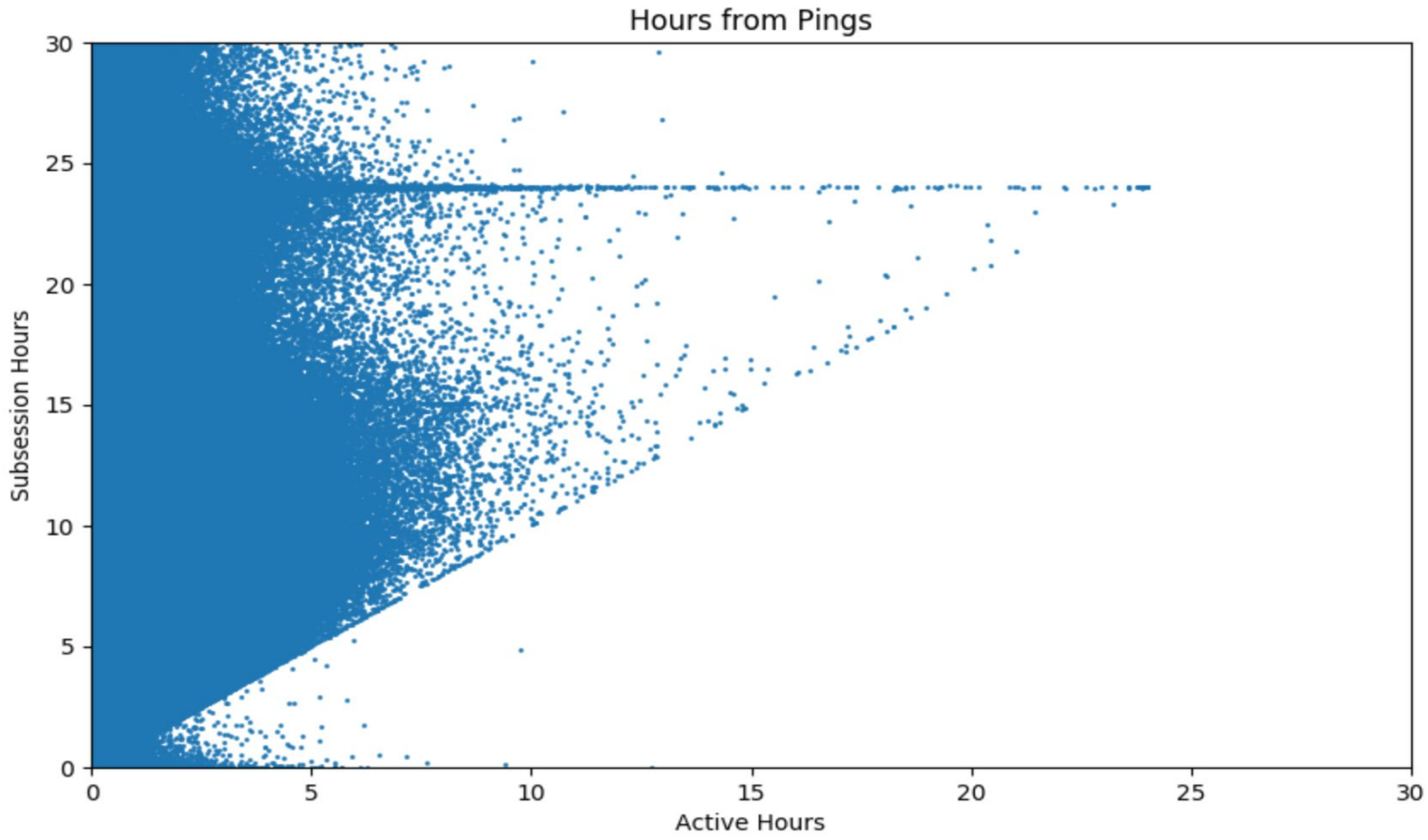
- Active ticks count of 5 second increments
- Converted by active hours
- Active use by browser
- Undercount time watching video, reading

Subsession Hours

- How long client computer running Firefox
- Includes background/idle time
- Subject to measurement error

Correlation

- For 1 week of ping data
- Correlation is 0.004



Summary Stats

All the daily client records for a week

	Active Hours	Subsession Hours
Count	24,422,216	24,422,210
Mean	0.169737	2.579780
Standard Deviation	0.423988	422.106222
Minimum	0.0	0.0
Maximum	71.6972	1,576,182.1558

Outliers

- Active hours > 24 hours: 4 pings
- Subsession hours > 25 hours
  - 198,466 pings
  - 131,870 unique client ids

Determine Cutoff(s)

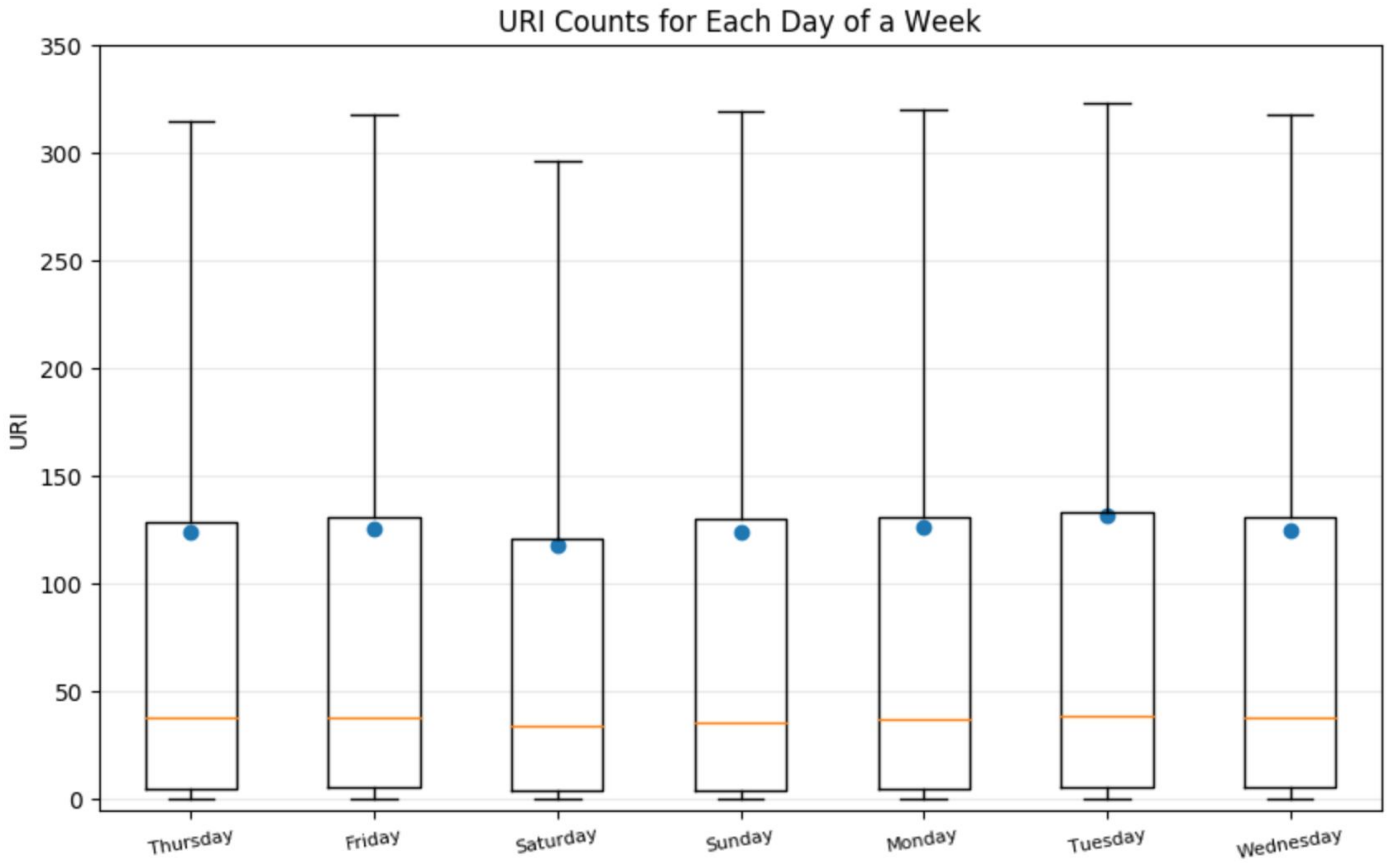
Summary Stats

All daily client records for a week

	Submission Date	Total Daily URI	Total Daily Active Hours	Total Daily Search Count
Count	6,436,229	6,436,229	6,436,229	6,436,229
Mean		125.427	0.642	4.187
Standard Deviation		2041.688	1.359	42.622
Minimum	20180920	0.0	0.0	0.0
Maximum	20180926	4530092	446.719	26937

URI Counts by Day

- Distribution of URI count data for each day of the week
- Mean value shown with blue dot



Segments of Data

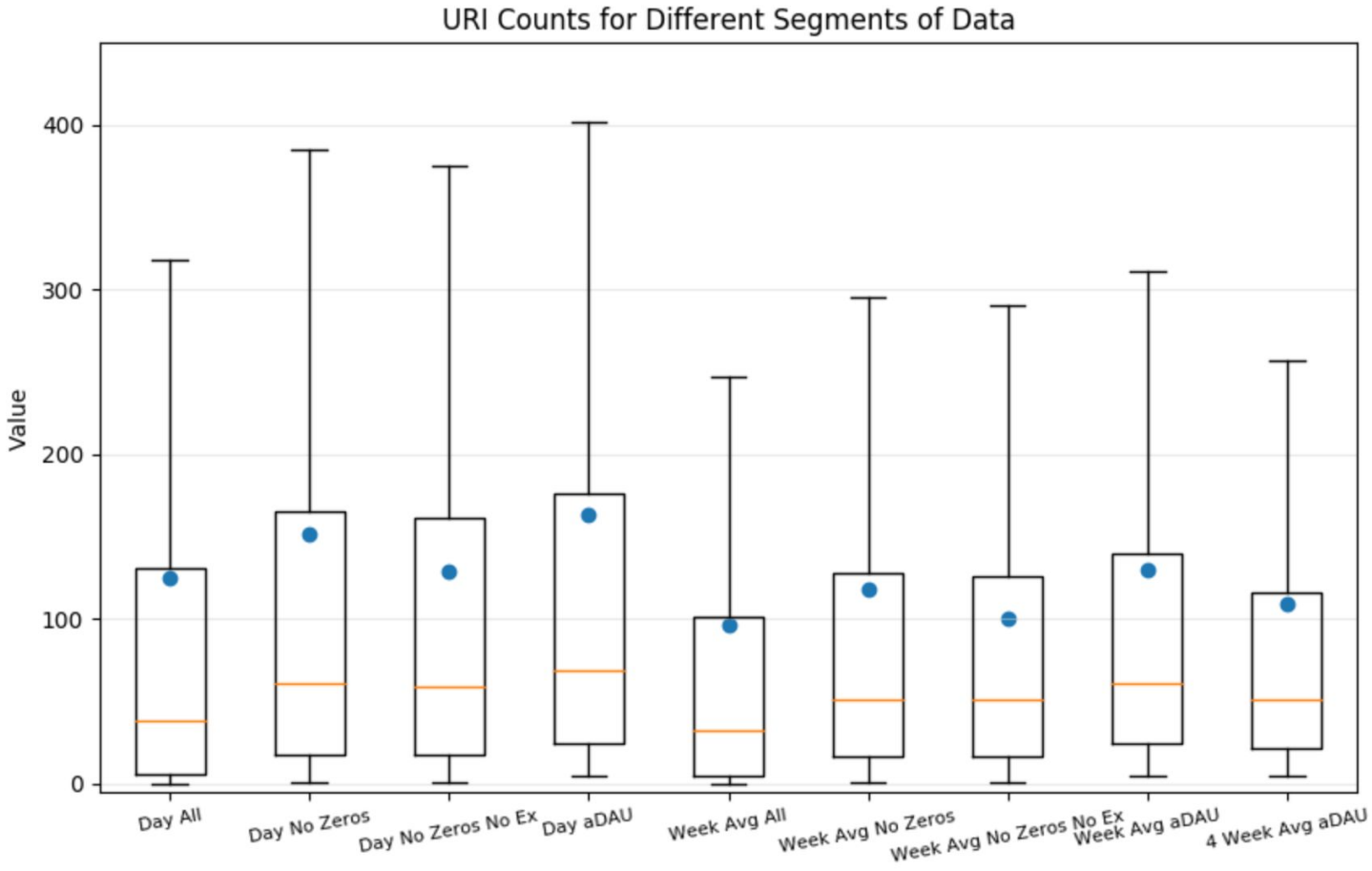
- Compensate for variability
- Reduce effect of outliers
- Reduce effect of inactive days
- 4 segments
  - all the records
  - URI and active hours both > 0
  - URI and active hours both > 0 AND URI and search counts < extreme values
  - aDAU users

Segment Stats

- URI and Active Hours both 0
  - 8.7% of records in 1 week
  - 237,144 distinct clients
- Extreme value above 99.5th percentile
  - URI 1500
  - 0.67% records in 1 week with extreme values
  - 27,411 clients with extreme values
- aDAU records
  - 75% records in 1 week
  - 1,432,289 clients
- Reference - all records
  - 6,436,229 in 1 week
  - 1,851,293 unique clients

URI Counts by Segment

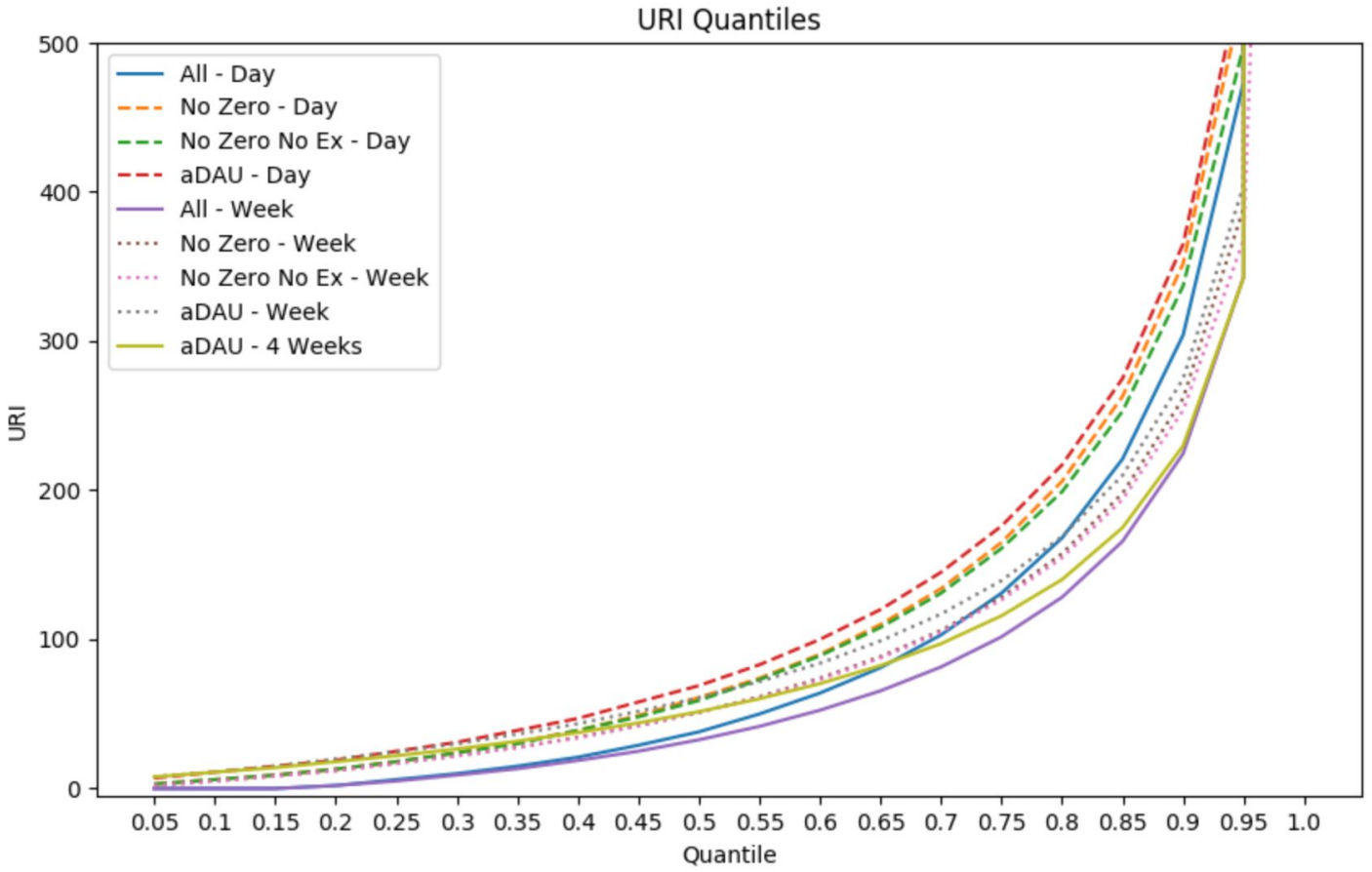
- 4 Segments for 1 day
- 4 Segments for 1 week average
- aDAU for 4 week average
- Search count and active hours follow same pattern



URI Quantiles

For different data segments

- All values increase sharply at 95th percentile
- Search count and active hours follow the same pattern



URI Quantiles for data segments

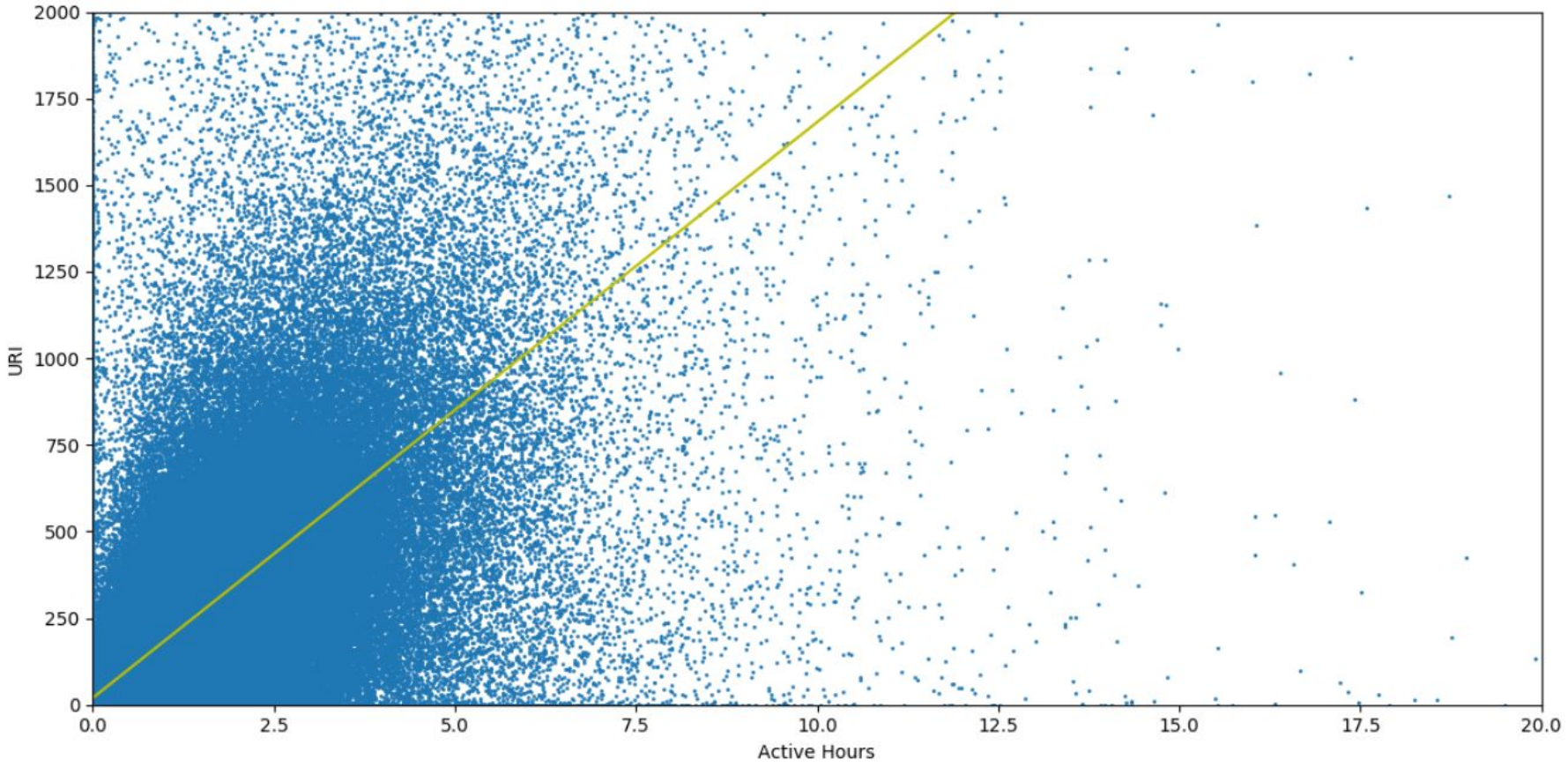
Set of most likely URI heavy user cutoffs

Percentile	Day All	Day No Zero	Day No Zero No Extreme	Day aDAU	Week All	Week No Zero	Week No Zero No Extreme	Week aDAU	4 Week aDU
75	131.00	165.00	161.00	176.00	101.71	128.25	126.43	139.40	115.81
80	168.00	206.00	199.00	217.00	128.25	157.50	154.75	169.00	140.10
90	304.00	352.00	337.00	365.00	224.50	261.60	253.57	274.67	229.75
95	474.00	534.00	498.00	550.00	343.00	389.40	368.00	404.60	342.60

Correlation between Metrics

Daily Totals from 1 Day

- Search count to URI: 0.12
- Active hours to URI: 0.25
- Active hours to search count: 0.35



Chosen Cutoffs

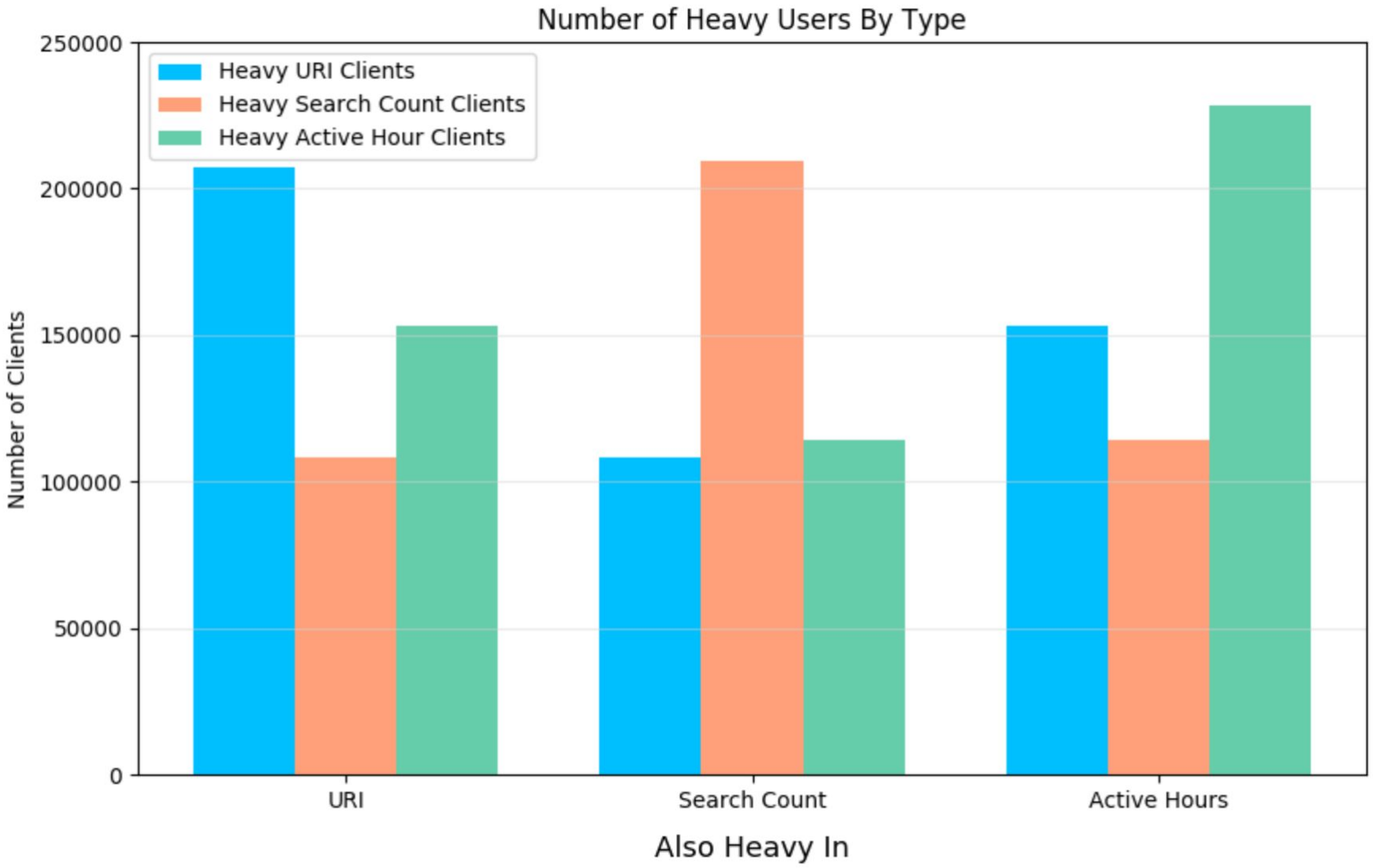
- 3 metrics for cutoffs
- Chose cutoffs from 80th percentile of weekly aDAU
  - weekly average reduce volatility
  - aDAU records ensure active users
  - Specific cutoffs for this date
    - URI 169.0
    - Active Hours 0.93

Percentage of Heavy Users

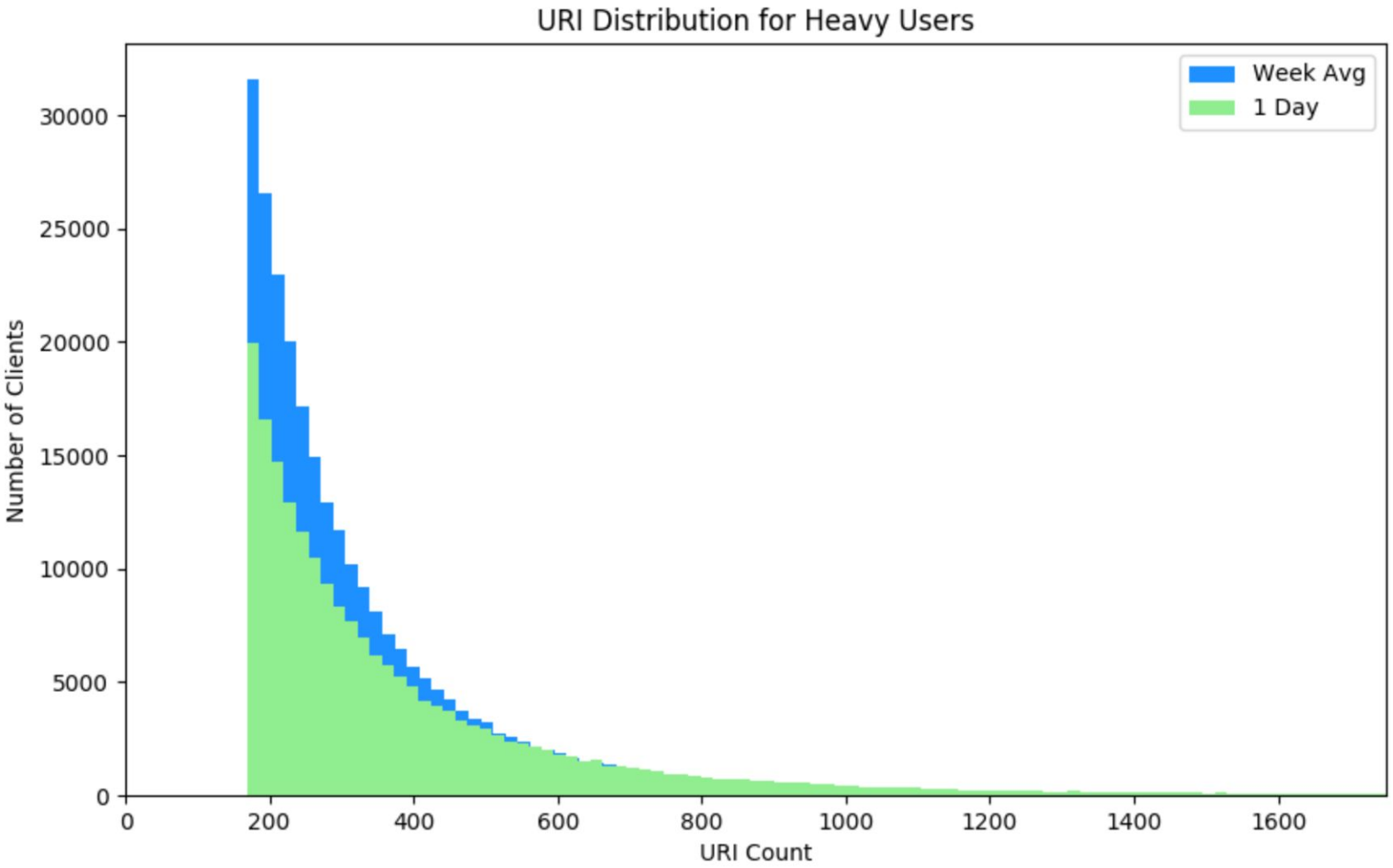
Heavy User Type	1 Day	Weekly Average
URI	19.93%	14.66%
Search Count	20.16%	16.61%
Active Hours	21.94%	15.69%
Any of the 3	34.47%	27.55%
All of the 3	8.54%	5.86%

Number of Heavy Users by Type

- First blue bar - heavy URI users also heavy in URI
- Second blue bar - heavy URI users also heavy in search count
- Third blue bar - heavy URI users also heavy in active hours



Distribution of Heavy Users

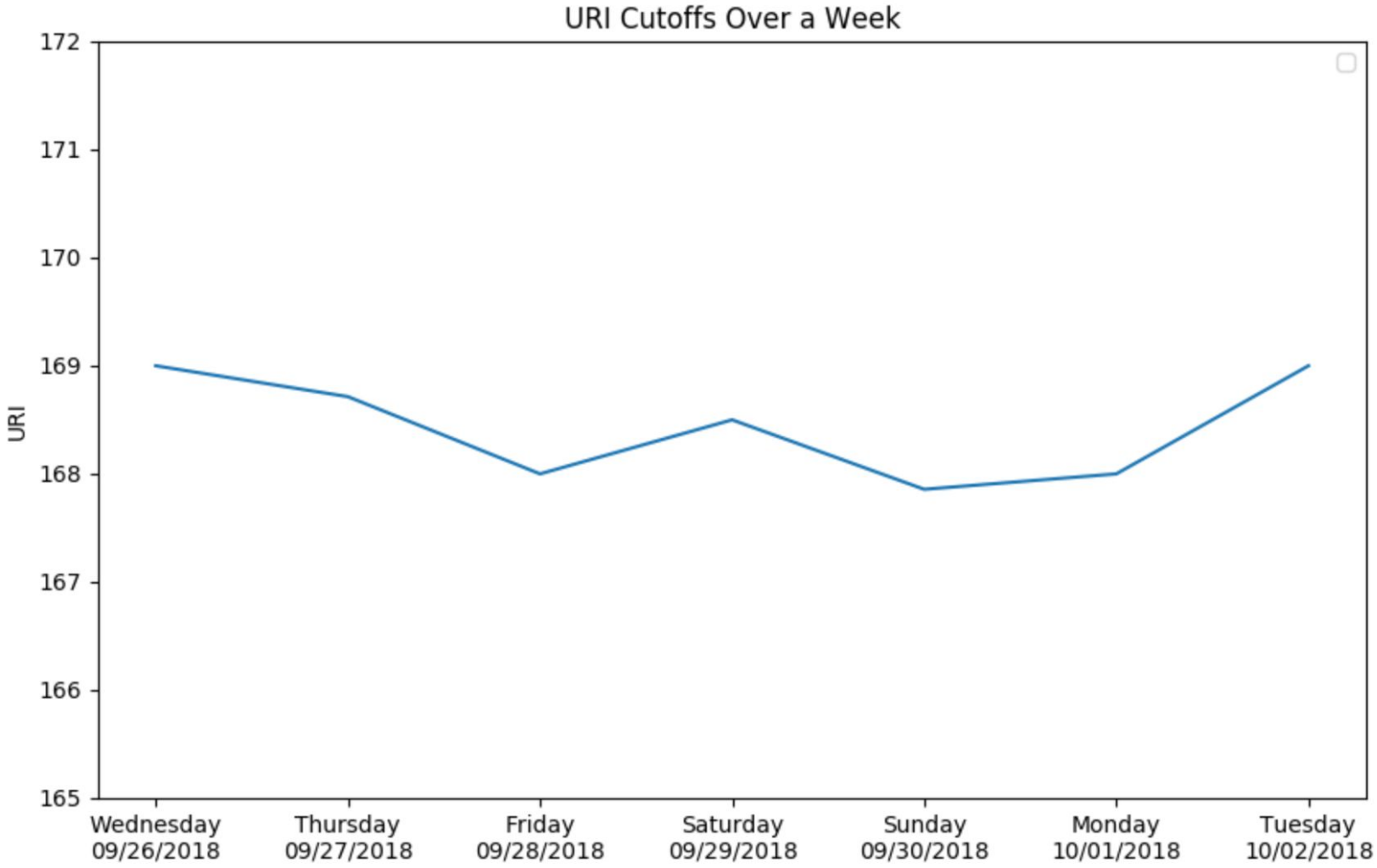


Definition of Heavy User

- On given day D
- For metrics URI, search count and active hours
- The 80th percentile and above of the metric for 1 week ending on day D where the total daily URI count >= 5, of daily values averaged over the week for days that have data

Heavy Users Over Time

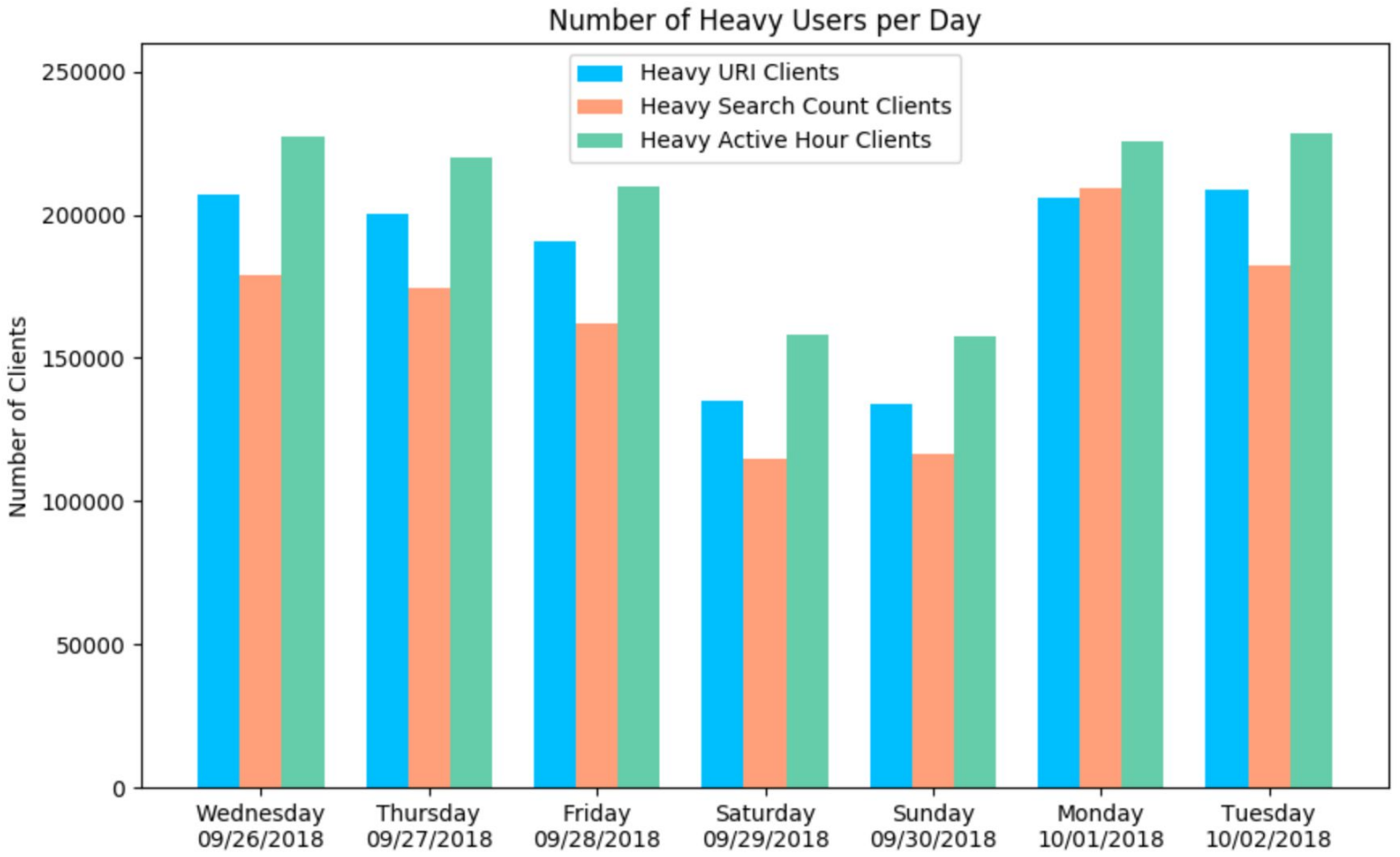
URI Cutoff over a Week



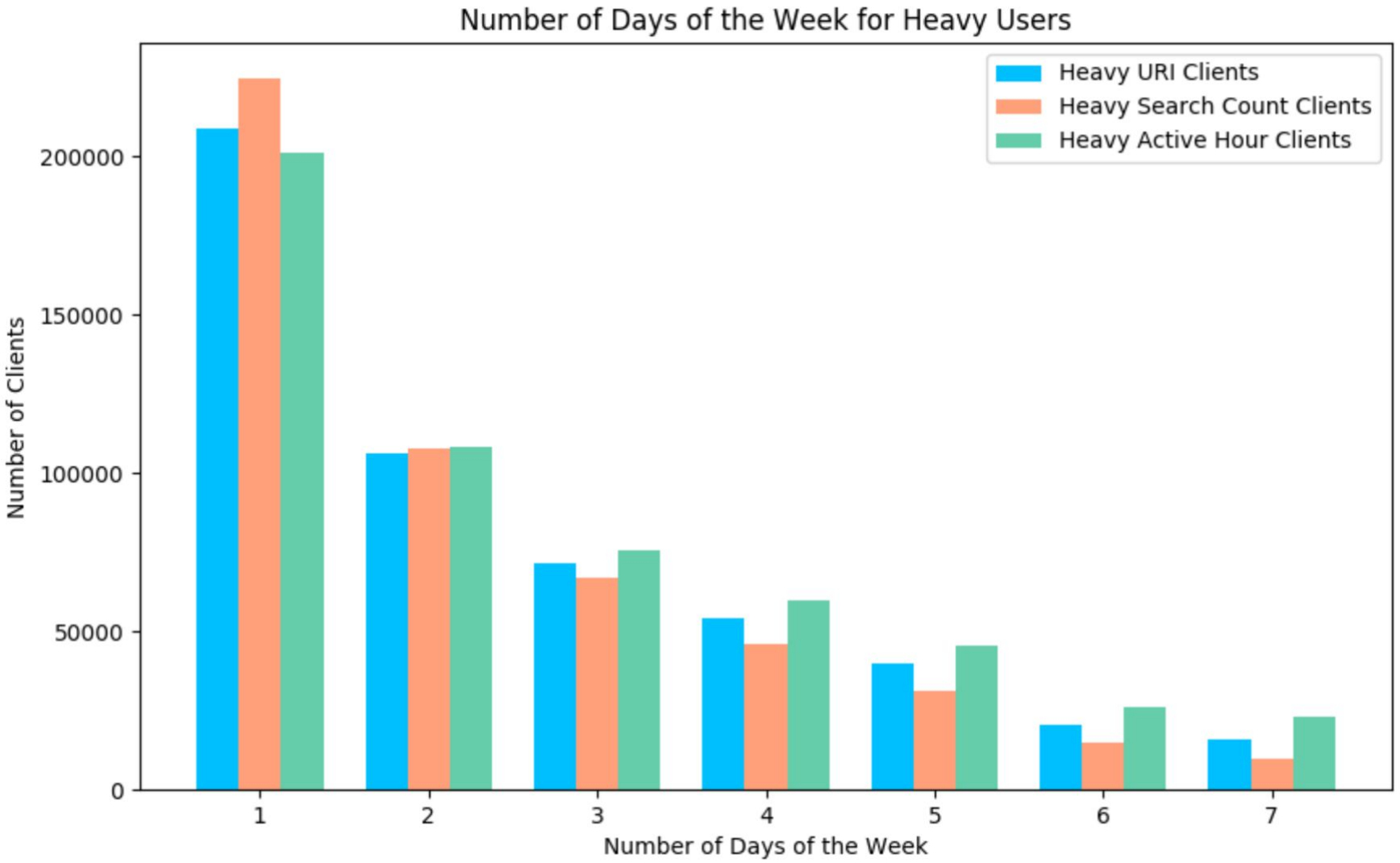
Number of Heavy Users per Day

Variation over Days of the Week

- Number of Heavy Users decreases on the weekend



Number of Days a Week



Contradictory Results?

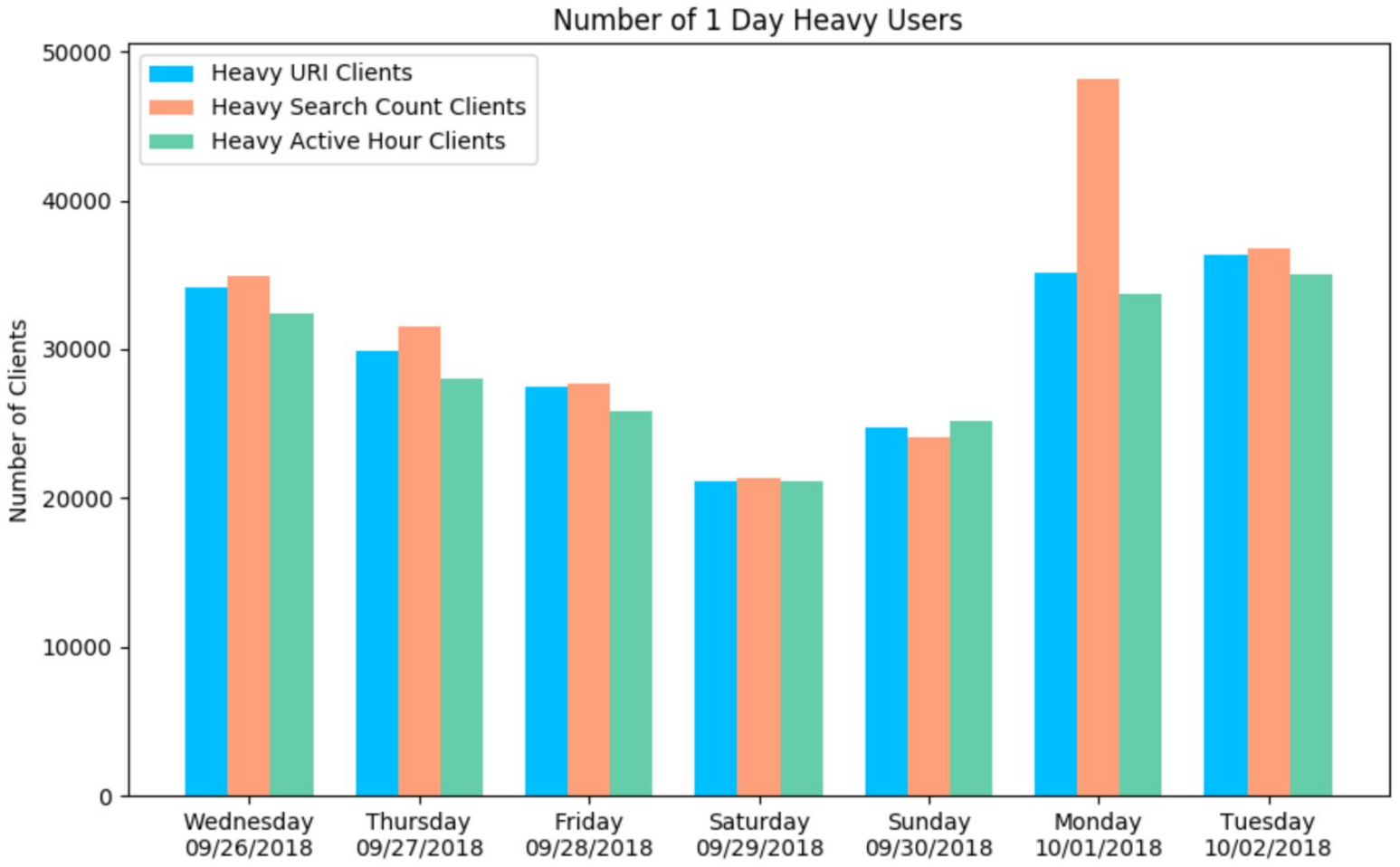
Different Users each Day

	Wednesday	Thursday
Total Heavy Users for the day	207,344	200,380
Heavy Users only on this day	100,924	93,960
Percentage Heavy Users only on this day	48%	47%

Number of 1 Day Heavy Users

Different Users each Day

- Heavy Users only 1 day of the week are distributed throughout the week



Retention Rate

Are they still heavy users in the next 6 weeks

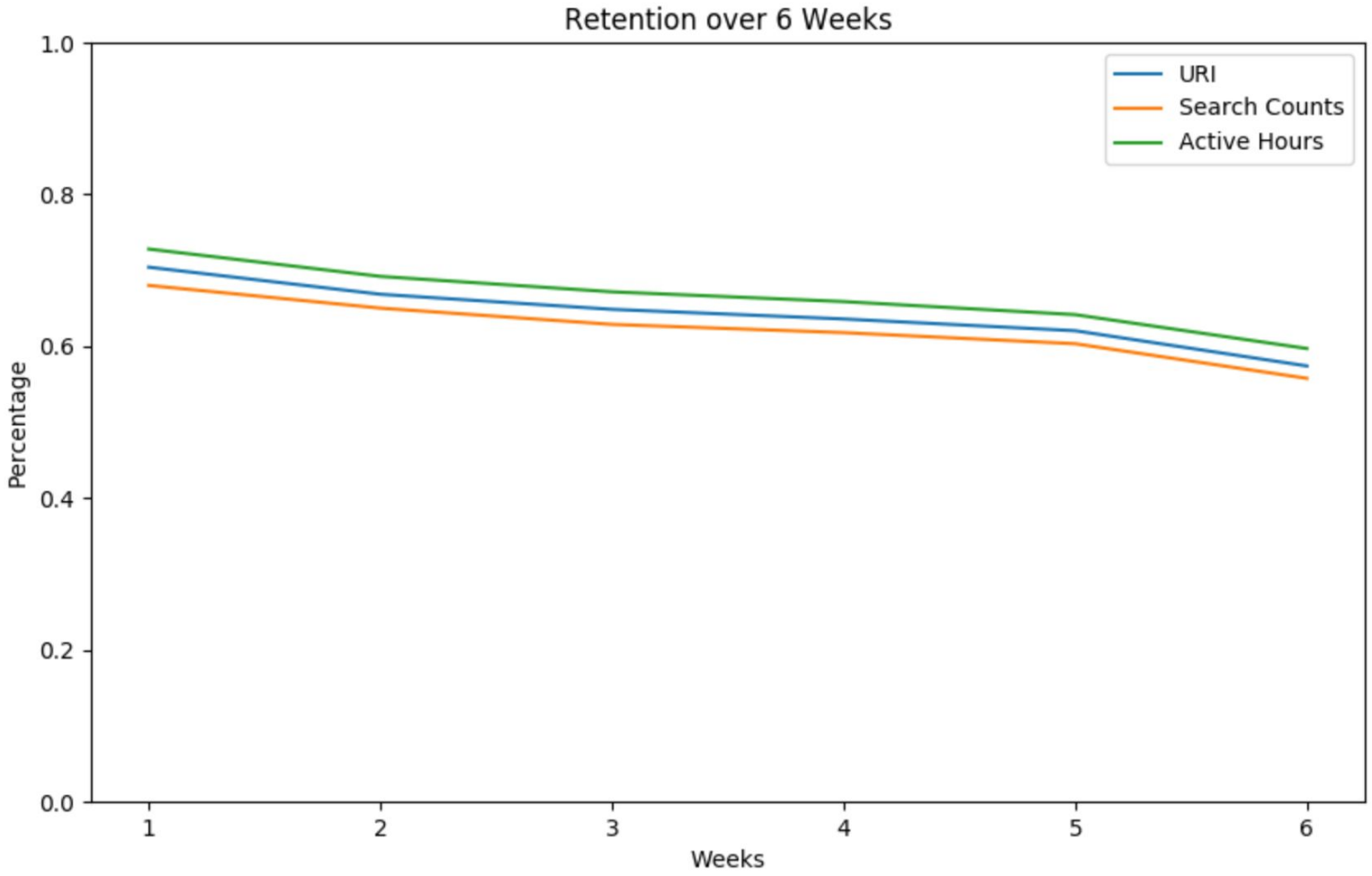
Period	Retention	CI 95% Semi Interval*
1	0.7041	0.00125
2	0.6684	0.00128
3	0.6486	0.00130
4	0.6359	0.00131
5	0.6204	0.00133
6	0.5740	0.00135

*The 95% CI spans the range `retention ± ci_95_semi_interval`

Retention Rate

Are they still heavy users in the next 6 weeks

- Similar for all 3 types of heavy users



Attributes of Heavy Users

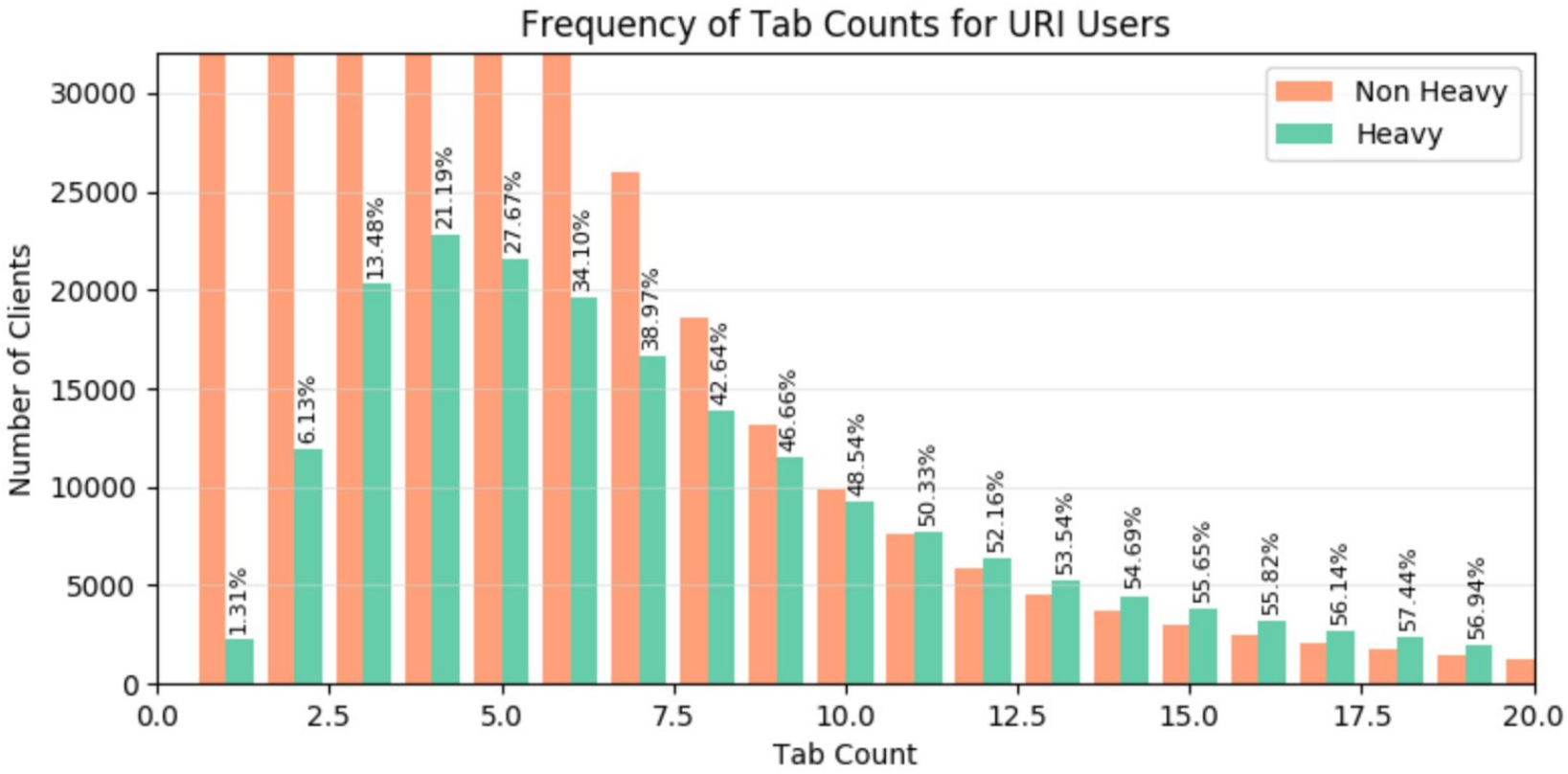
Explanation of Frequency Bar Charts

- One day of data
- Frequencies and percentages of heavy users for:
  - Tab count
  - Window count
  - Active addons count
  - Sessions started on this day
  - Normalized channel
  - OS
  - Is default browser
  - Country
- Group by the attribute and count the number of clients - for heavy users, then non-heavy users
- Each user is either heavy or non-heavy
- Only clients with non-null value for attribute

Tab Counts

Maximum tab count for 1 day

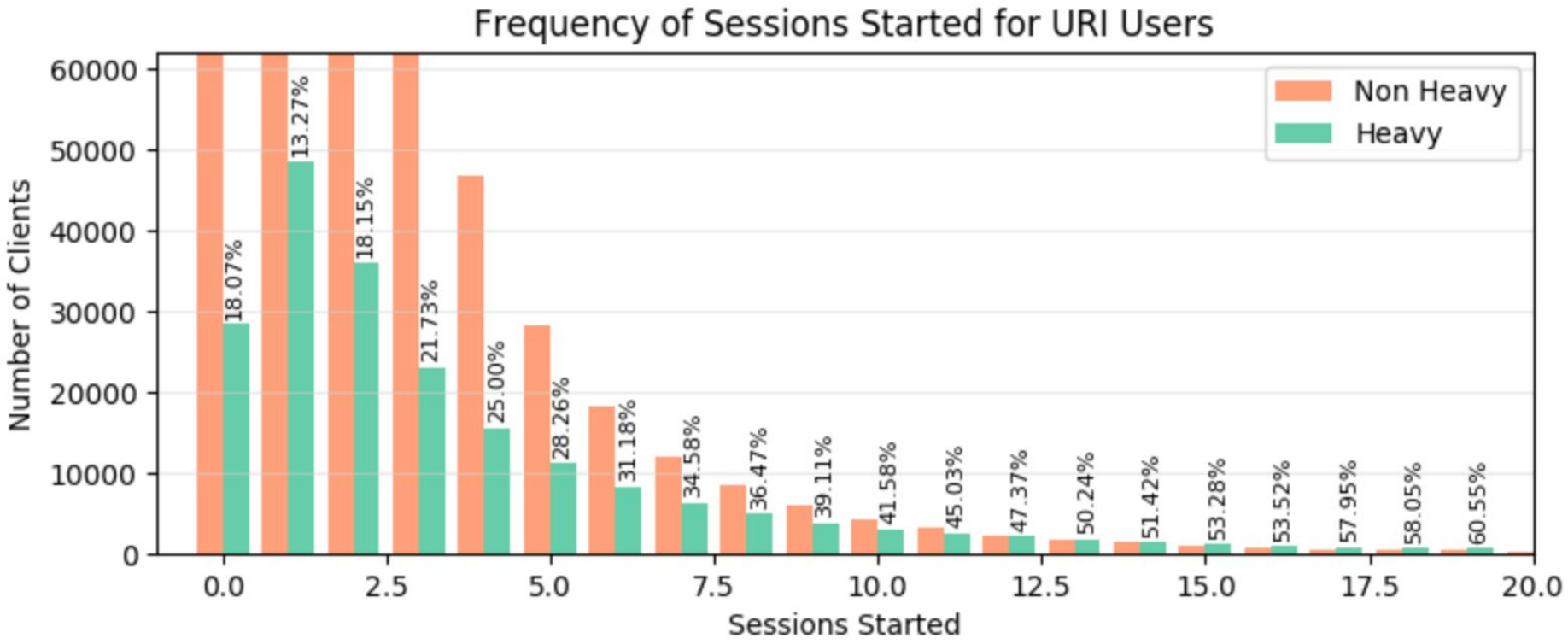
- Data for 1 actual user
- Higher tab counts, higher percentage of heavy users



Sessions Started

Sum of sessions started on this day

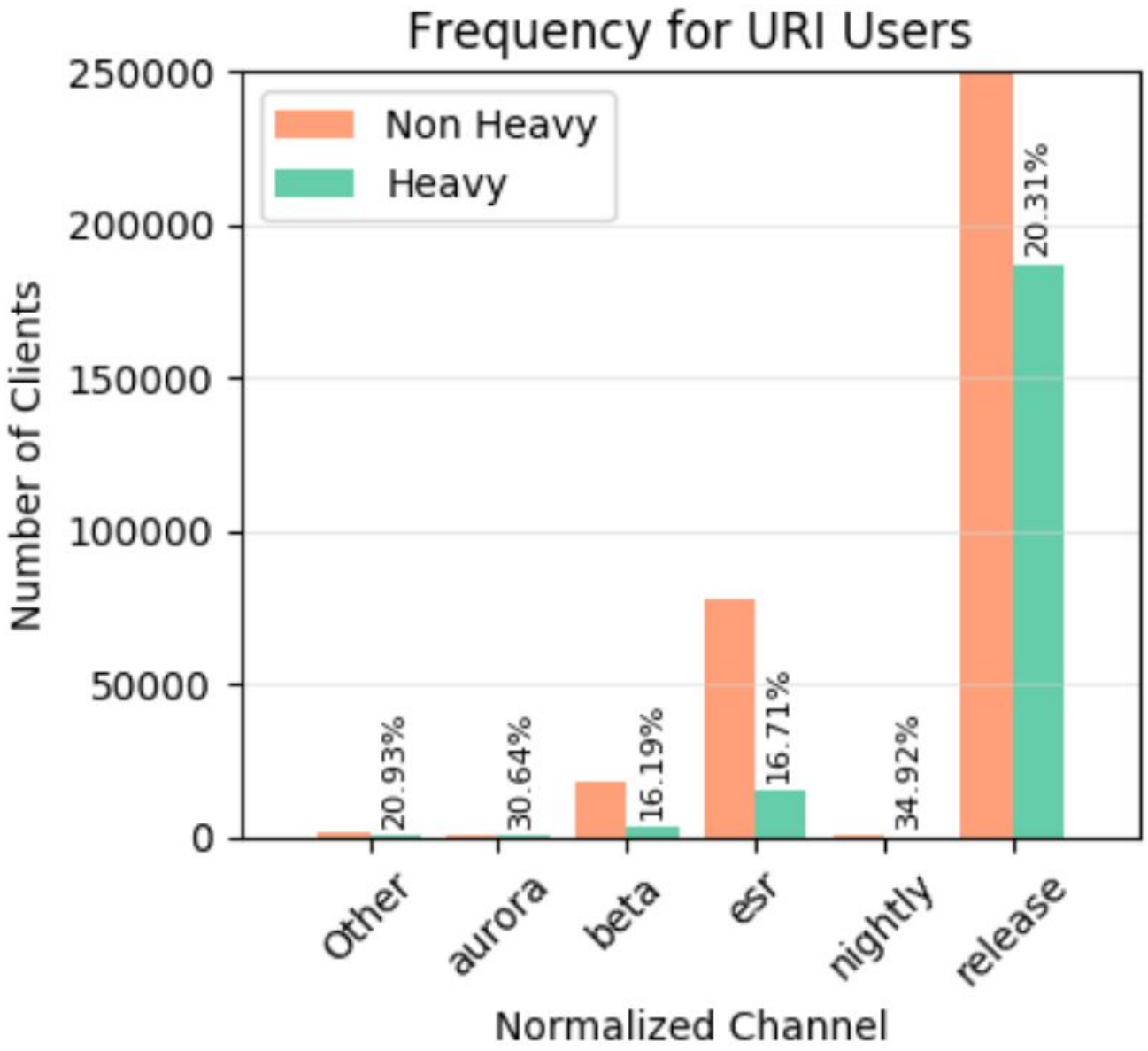
- Higher number of session started, higher percentage of heavy users



Normalized Channel

First normalized channel in pings for a day

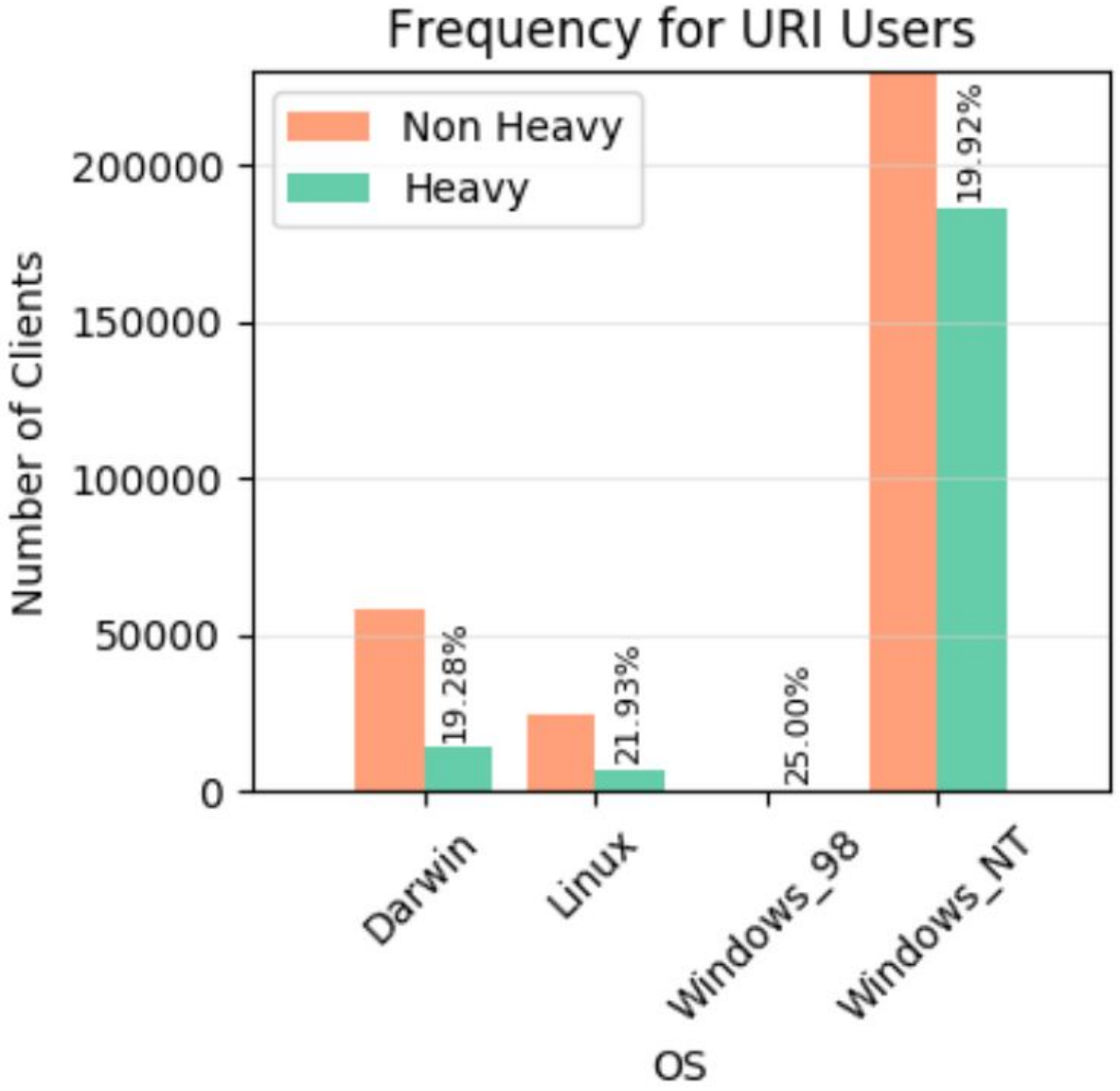
- Percentage high for nightly and aurora
- Percentage low for beta and esr



OS

First OS in pings for a day

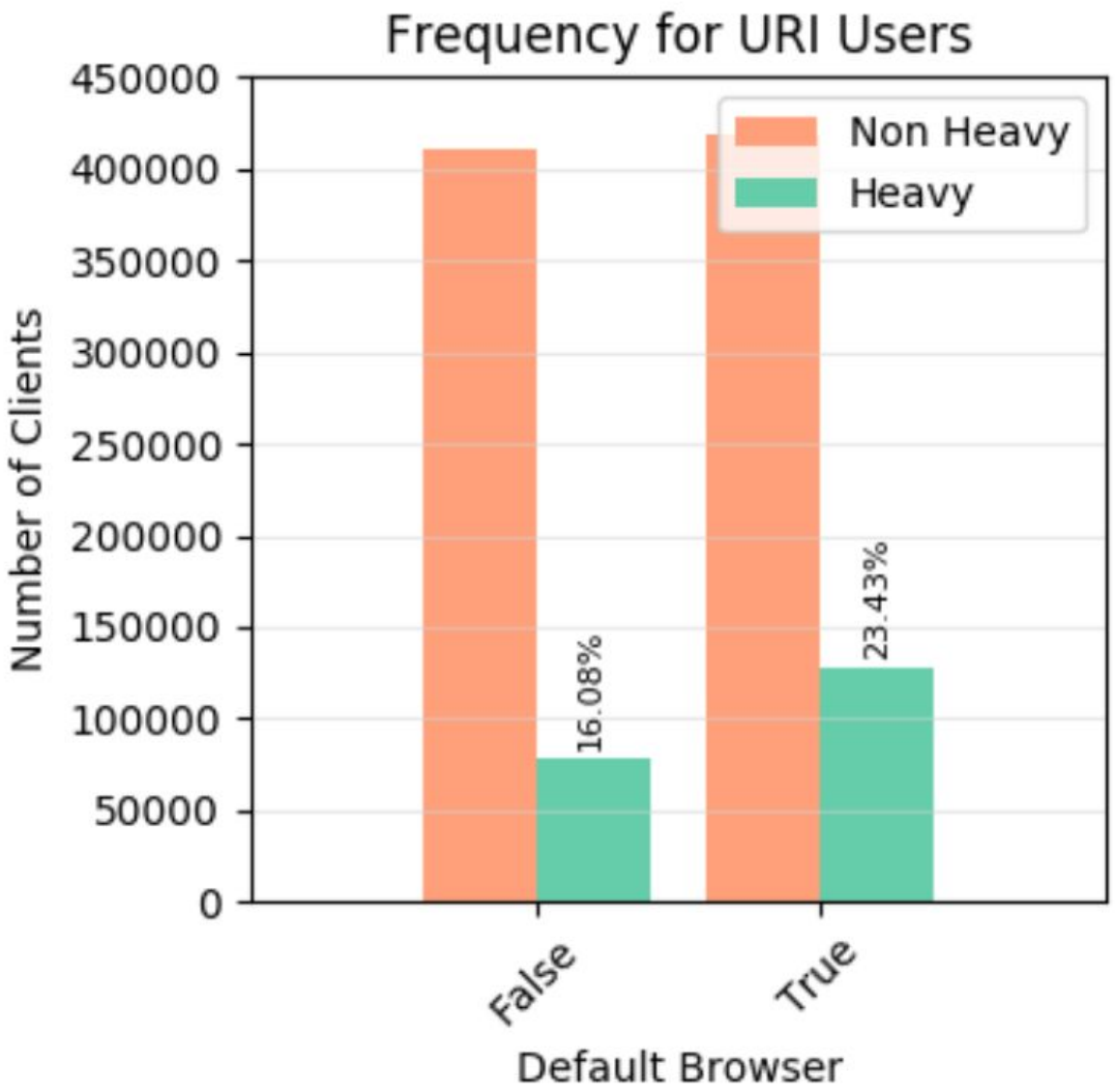
- Heavy users evenly distributed



Default Browser

First Is Default Browser in pings for a day

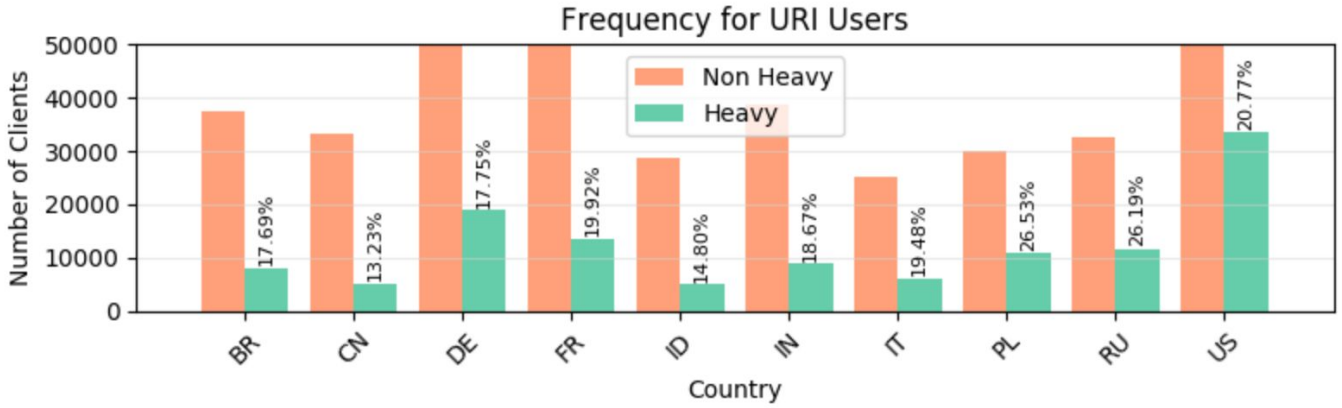
- Higher percentage of clients who have chosen Firefox as their default browser are heavy users

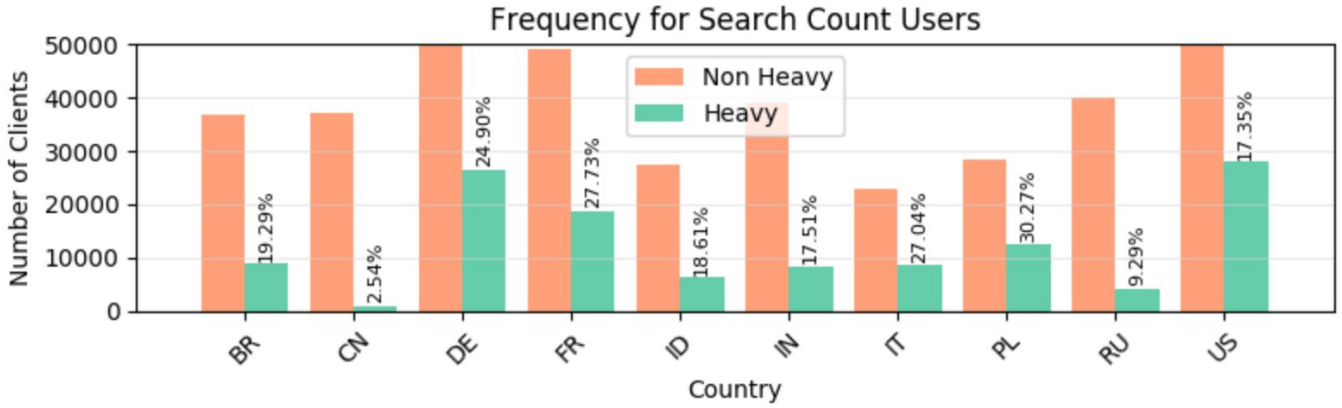


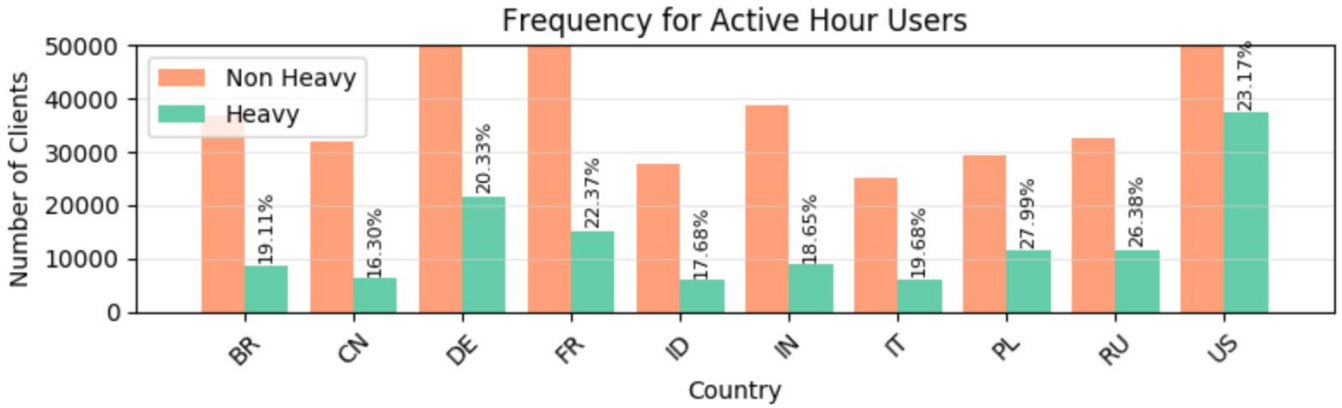
Top 10 Countries

First country in pings for a day

- Patterns different between the heavy user types
- China and Russia - low percentages for search count
- France and Italy - high percentage for search count
- Poland - high percentage in all three categories







Key Findings

- Not one cutoff
  - 3 metrics not closely correlated
  - Cutoff values will change over time
- Averaging over week lowers volatility
- Drawbacks
  - URI counts don’t include private browsing
  - Active Hours don’t include videos and reading
- Each attribute analyzed for each heavy user type

Main Summary vs Clients Daily

Summary Stats - Active Hours

Daily Totals

	Clients Daily Table	Main Summary Table
Count	1,045,624	1,045,624
Mean	0.81745	0.64331
Standard Deviation	16.1003	1.561
Minimum	0.0	0.0
Maximum	14,137.22	436.93

Records with Discrepancies

- 26 records in clients_daily where active hours > maximum from main_summary
- Pings for these clients have 0 active hours.

Unusual Usage Patterns

Zero URI and Zero Active Hours

- 20% pings in 1 day
- 165,413 distinct clients
- Pings totaled over a day
  - 83,621 clients

Zero URI and 24 Active Hours

- One client
  - For 7 days a week
  - For 4 weeks
  - For most days of 12 weeks
- 18 clients in a week with this pattern

Multiple Users, Same Client ID

- Same client_id, submission_date_s3 and profile_subsession_counter
  - Duplicate pings or multiple users?
- One client_id has 1157 pings with same profile_subsession_counter on same day
  - Some Windows, some Linux
- One client_id has 37,964 pings in a week
  - 10 different countries

More than 1000 URI and Zero Active Ticks

- 1,656 distinct client_ids in a week

Thank You