Feature Engineering for Bot Detection

Raja Iqbal

CEO | Chief Data Scientist

datasciencedojo

data science for everyone

Agenda

- Introduction
    - Growth of internet traffic
    - Good bots vs. bad bots
- Some common bot types
- Feature engineering for detecting bots
- Question

Acknowledgement

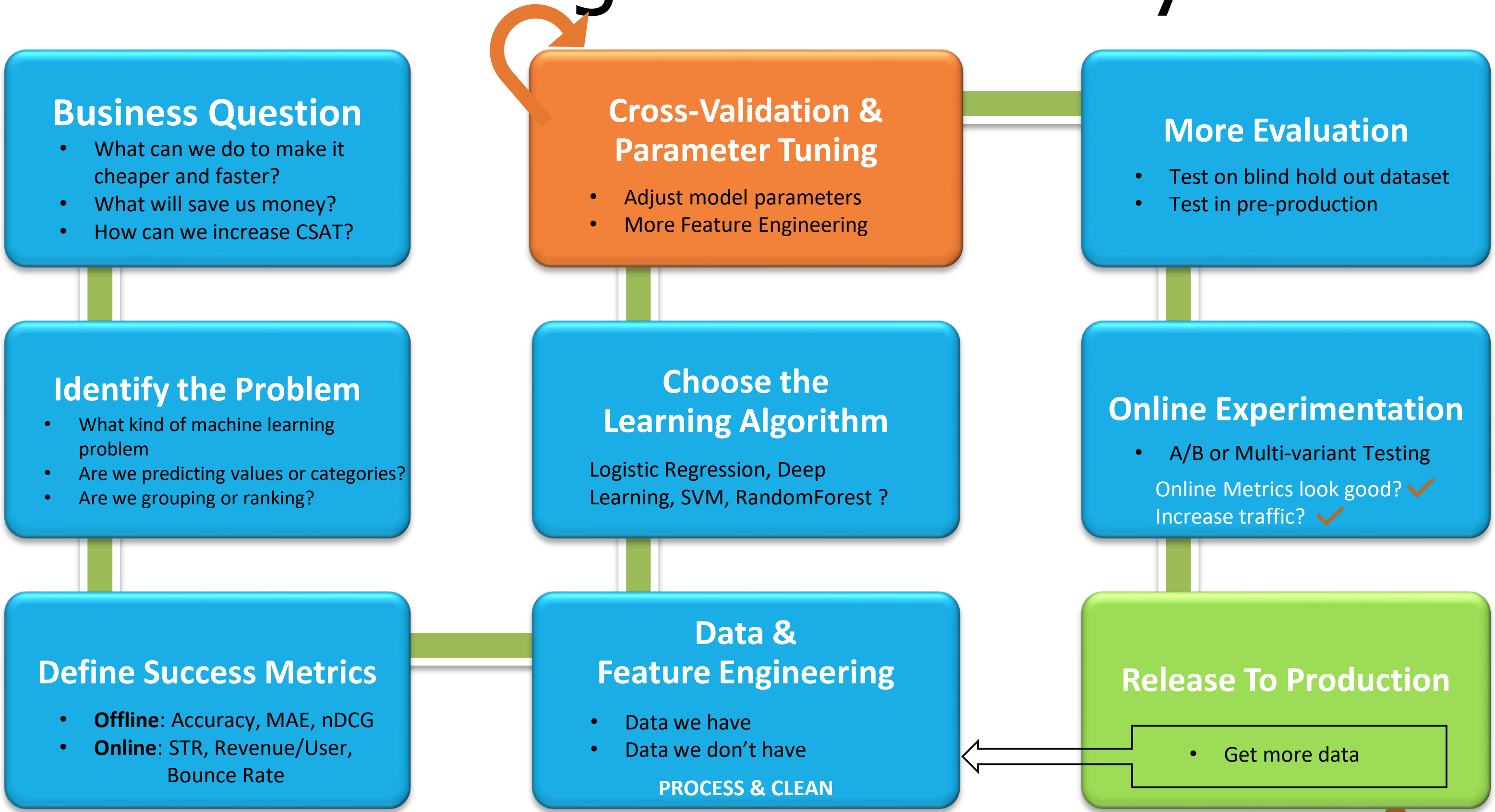
The discussion is based on the following paper:

Classification of Automated Web Traffic,

Greg Buehrer, Jack W. Stokes, Kumar Chellapilla and John C. Platt

Why this talk?

Machine Learning Model Life Cycle



Business Question

- What can we do to make it cheaper and faster?
- What will save us money?
- How can we increase CSAT?

Identify the Problem

- What kind of machine learning problem
- Are we predicting values or categories?
- Are we grouping or ranking?

Define Success Metrics

- **Offline**: Accuracy, MAE, nDCG
- **Online**: STR, Revenue/User, Bounce Rate

Data & Feature Engineering

- Data we have
- Data we don't have

PROCESS & CLEAN

Choose the Learning Algorithm

Logistic Regression, Deep Learning, SVM, RandomForest ?

Cross-Validation & Parameter Tuning

- Adjust model parameters
- More Feature Engineering

More Evaluation

- Test on blind hold out dataset
- Test in pre-production

Online Experimentation

- A/B or Multi-variant Testing

Online Metrics look good? ✓
Increase traffic? ✓

Release To Production

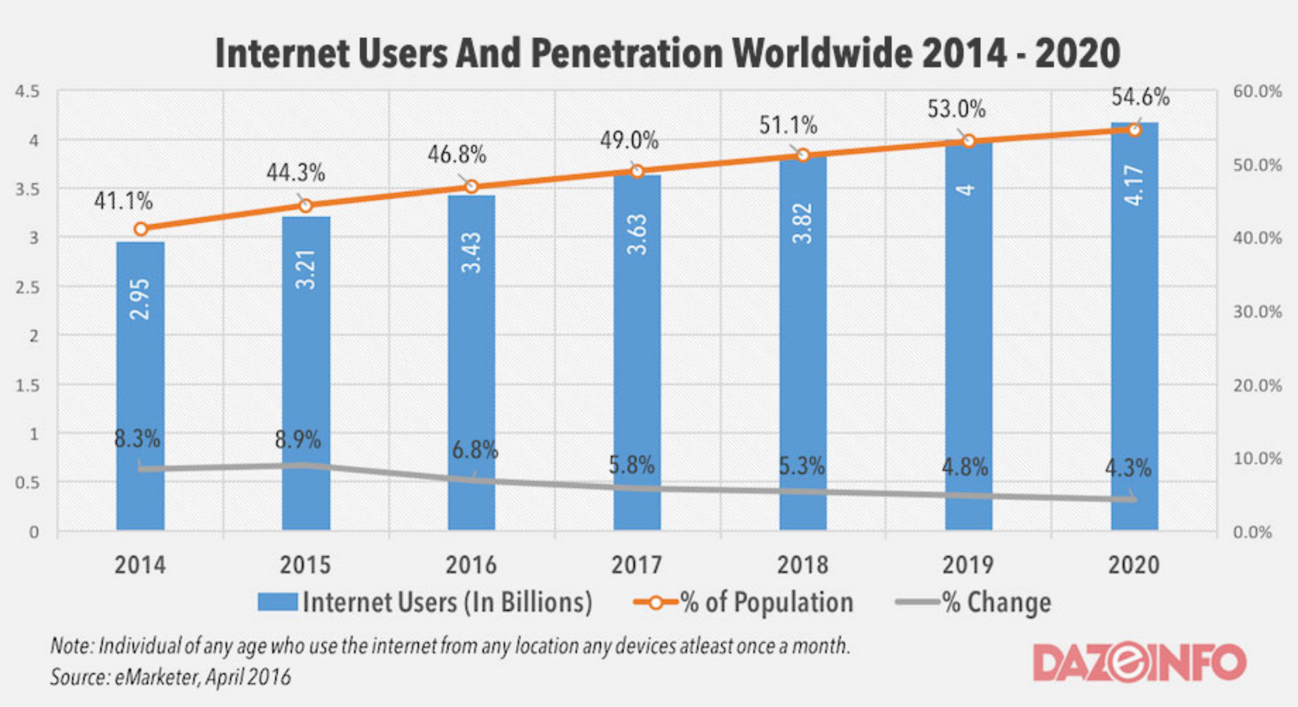
- Get more data

datasciencedojo
data science for everyone

Feature Engineering is important

- All good models start with good feature engineering.

Internet Usage and Growth



Note: Individual of any age who use the internet from any location any devices atleast once a month.
Source: eMarketer, April 2016

What is a (internet) bot?

- A software application that runs automated tasks over the internet.
- Used for simple, repetitive tasks to be performed faster than humans.

Example: Indexing and crawling

- A Web **crawler**, is an Internet bot that systematically browses the World Wide Web, typically for the purpose of Web **indexing**.
- Web search engines and some other sites use Web crawling or spidering software to update their index (or other web content)

Example: Chatbot

- A computer program which conducts a conversation via auditory or textual methods.
- Designed to simulate how a human would behave as a conversational partner

More bots…

Good

- Spider bot
- Trading bot
- Data bot
- FeedFetcher bot

Bad

- Email bot
- Bandit bot
- Transfer bot
- Zombie bot
- AdFraud bot

Scope of this talk

- We will only discuss bots specific to web

Why are web bots a ‘problem’?

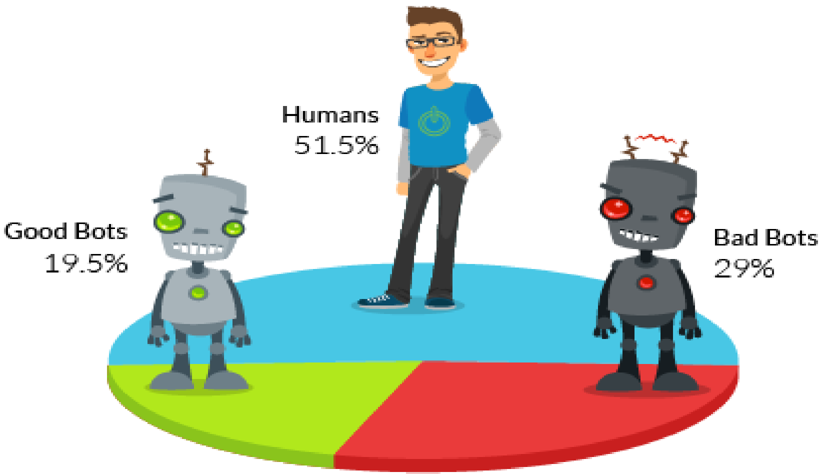
- Reduced QoS (Quality of Service)
- Machine learning models learn behavior that does not represent actual customers
- Click frauds and incorrect metrics calculation

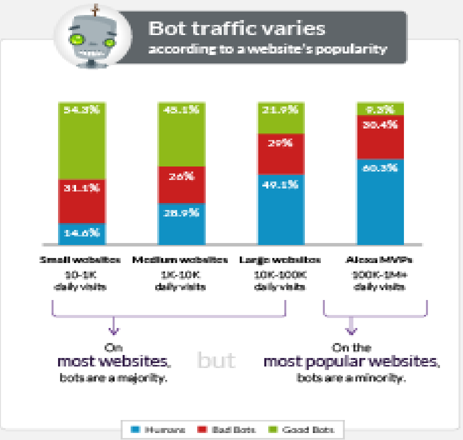
Traffic Trends 2015

Global Bot Traffic Report 2015

For the first time, humans are more active than bots, accounting for 51.5% of all website traffic.

This is a breakdown of online traffic in 2015:





On **most websites,** bots are a majority. but On the **most popular websites,** bots are a minority.

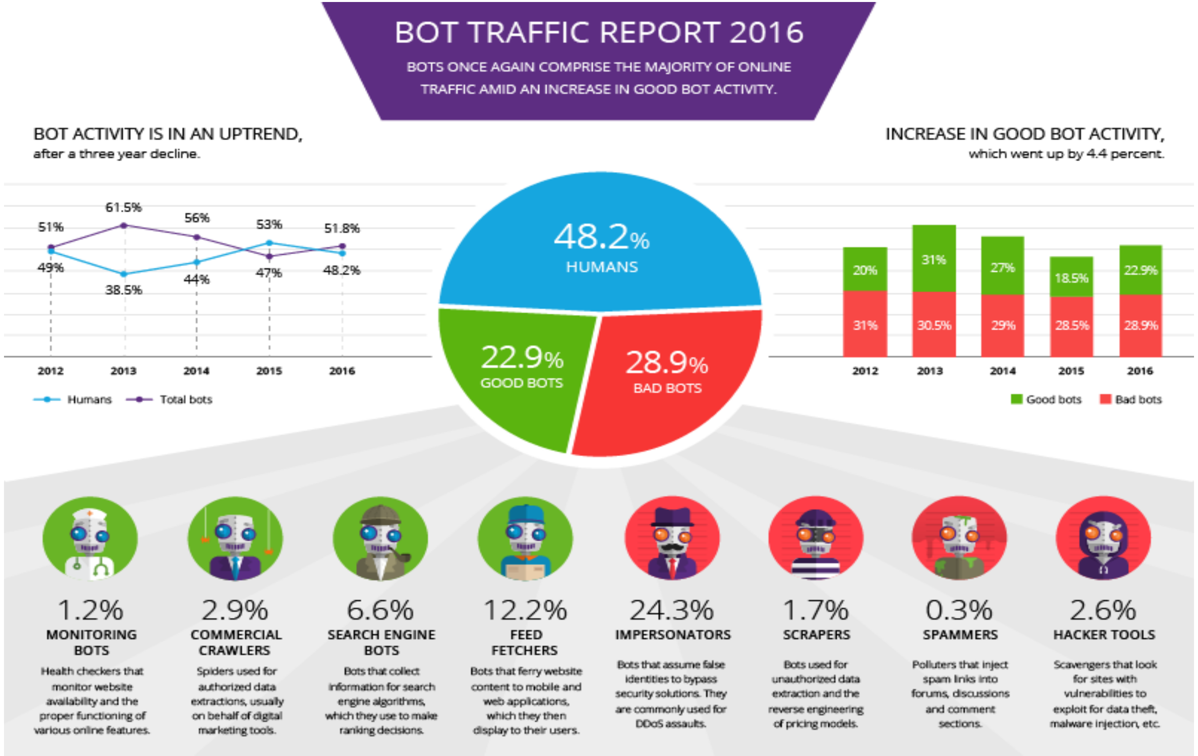
As a rule of thumb

When a website attracts more humans, the relative amount of good bots decline, while bad bot traffic stays the same.

source IMPERVA INCAPSULA

datasciencedojo
data science for everyone

Traffic Trends 2016



BOT TRAFFIC REPORT 2016

BOTS ONCE AGAIN COMPRISE THE MAJORITY OF ONLINE TRAFFIC AMID AN INCREASE IN GOOD BOT ACTIVITY.

BOT ACTIVITY IS IN AN UPTREND, after a three year decline.

INCREASE IN GOOD BOT ACTIVITY, which went up by 4.4 percent.

1.2% MONITORING BOTS

Health checkers that monitor website availability and the proper functioning of various online features.

2.9% COMMERCIAL CRAWLERS

Spiders used for authorized data extractions, usually on behalf of digital marketing tools.

6.6% SEARCH ENGINE BOTS

Bots that collect information for search engine algorithms, which they use to make ranking decisions.

12.2% FEED FETCHERS

Bots that ferry website content to mobile and web applications, which they then display to their users.

24.3% IMPERSONATORS

Bots that assume false identities to bypass security solutions. They are commonly used for DDoS assaults.

1.7% SCRAPERS

Bots used for unauthorized data extraction and the reverse engineering of pricing models.

0.3% SPAMMERS

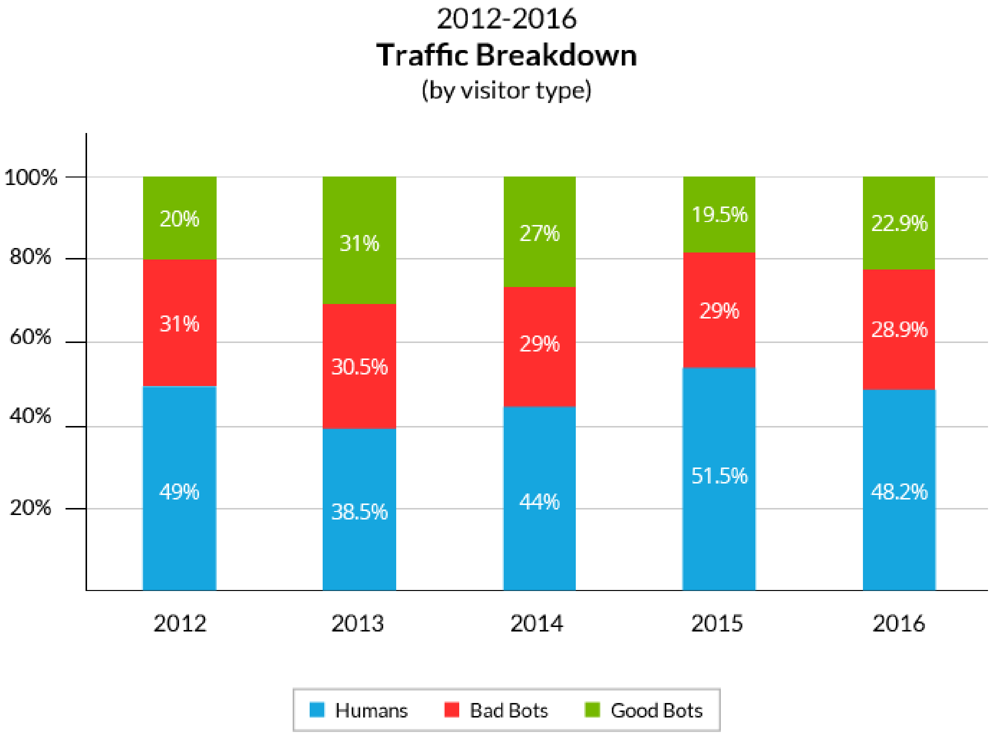
Polluters that inject spam links into forums, discussions and comment sections.

2.6% HACKER TOOLS

Scavengers that look for sites with vulnerabilities to exploit for data theft, malware injection, etc.

source IMPERVA INCAPSULA

Traffic Breakdown



source IMPERVA INCAPSULA

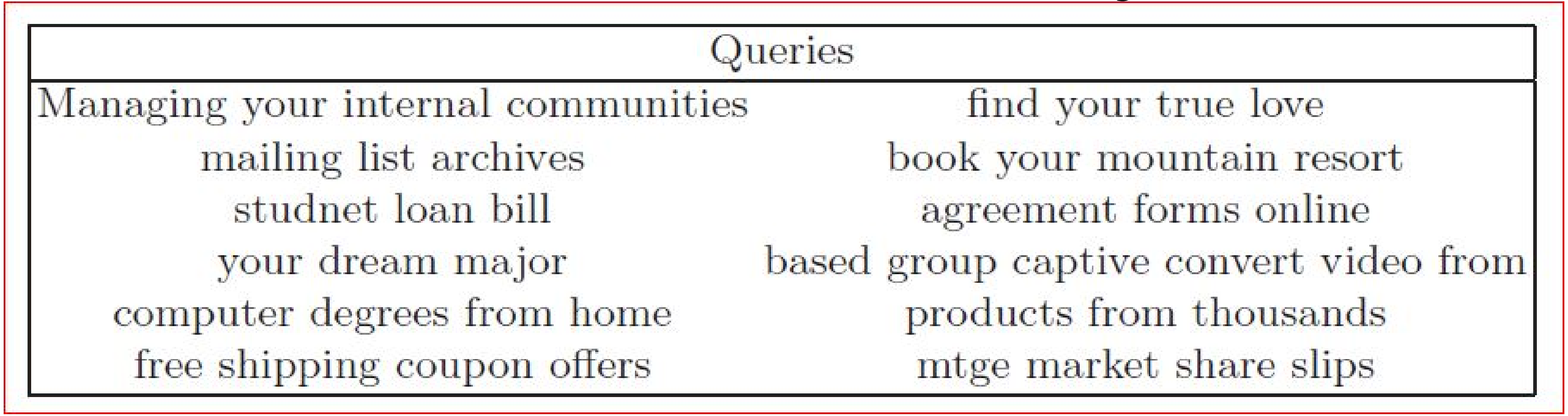
What are some typical bots?

Typical web bots

- Spam bot
- Finance bot
- URL bot
- Real Estate Bot
- Stock Bot
- Simple query bot(originating from many cities)

Spam Bot

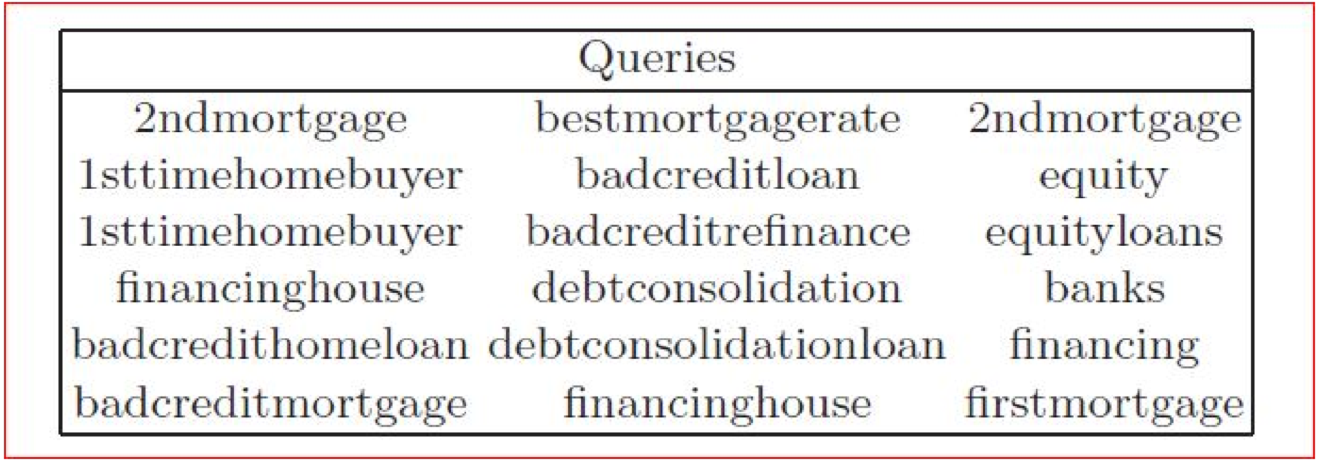
- Scans the index for top spam words
- Queries often but clicks rarely



Queries	
Managing your internal communities	find your true love
mailing list archives	book your mountain resort
studnet loan bill	agreement forms online
your dream major	based group captive convert video from
computer degrees from home	products from thousands
free shipping coupon offers	mtge market share slips

Finance Bot

- Ascertains which websites are most correlated with these finance terms

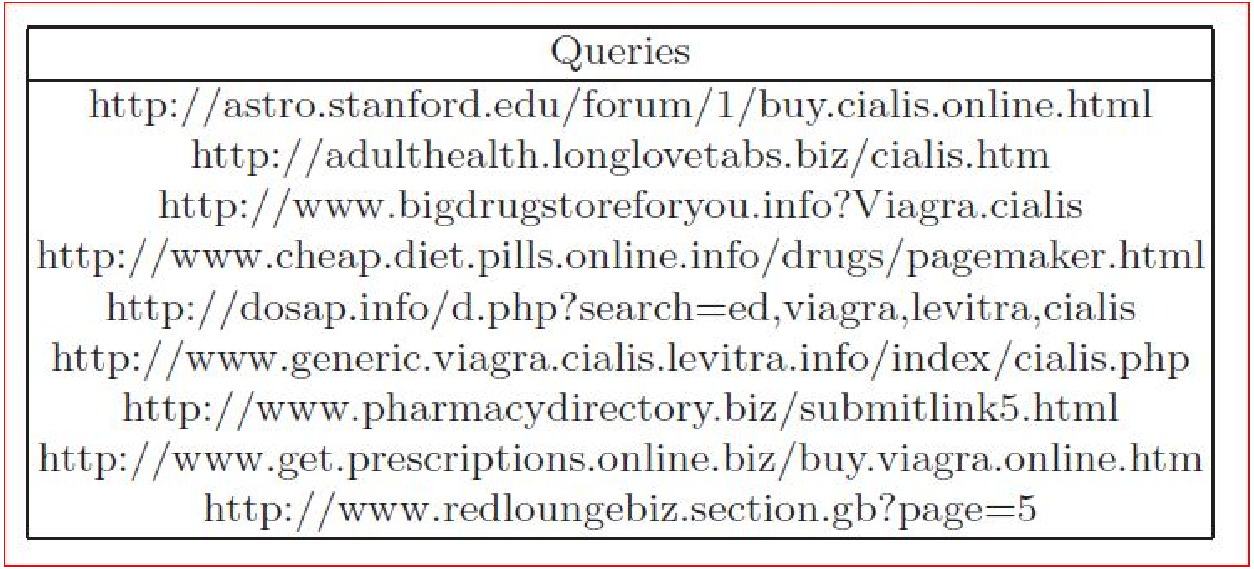


Queries		
2ndmortgage	bestmortgagerate	2ndmortgage
1sttimehomebuyer	badcreditloan	equity
1sttimehomebuyer	badcreditrefinance	equityloans
financinghouse	debtconsolidation	banks
badcredithomeloan	debtconsolidationloan	financing
badcreditmortgage	financinghouse	firstmortgage

URL bot

- Websites owned by spammers or legitimate domains hacked by hackers
- Presumably the bot is attempting to boost its search engine rank

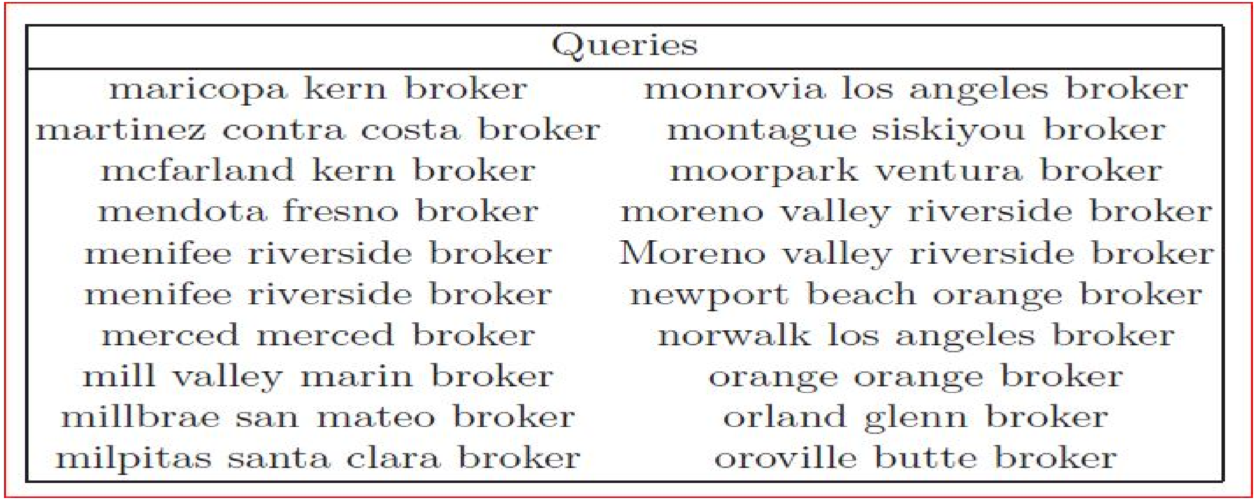
URL Bot



Queries
http://astro.stanford.edu/forum/1/buy.cialis.online.html
http://adulthealth.longlovetabs.biz/cialis.htm
http://www.bigdrugstoreforyou.info?Viagra.cialis
http://www.cheap.diet.pills.online.info/drugs/pagemaker.html
http://dosap.info/d.php?search=ed,viagra,levitra,cialis
http://www.generic.viagra.cialis.levitra.info/index/cialis.php
http://www.pharmacydirectory.biz/submitlink5.html
http://www.get.prescriptions.online.biz/buy.viagra.online.htm
http://www.redloungebiz.section.gb?page=5

Real estate bot

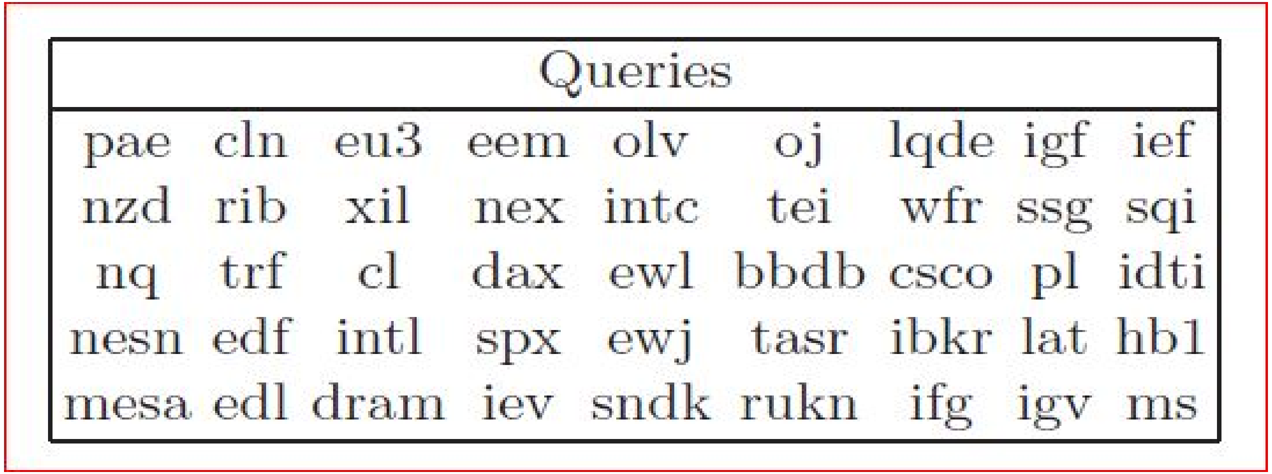
- Attempting to find the top ten broker results for mortgage broker keywords



Queries	
maricopa kern broker	monrovia los angeles broker
martinez contra costa broker	montague siskiyou broker
mcfarland kern broker	moorpark ventura broker
mendota fresno broker	moreno valley riverside broker
menifee riverside broker	Moreno valley riverside broker
menifee riverside broker	newport beach orange broker
merced merced broker	norwalk los angeles broker
mill valley marin broker	orange orange broker
millbrae san mateo broker	orland glenn broker
milpitas santa clara broker	oroville butte broker

Stock bot

- Searching for financial news related to particular companies

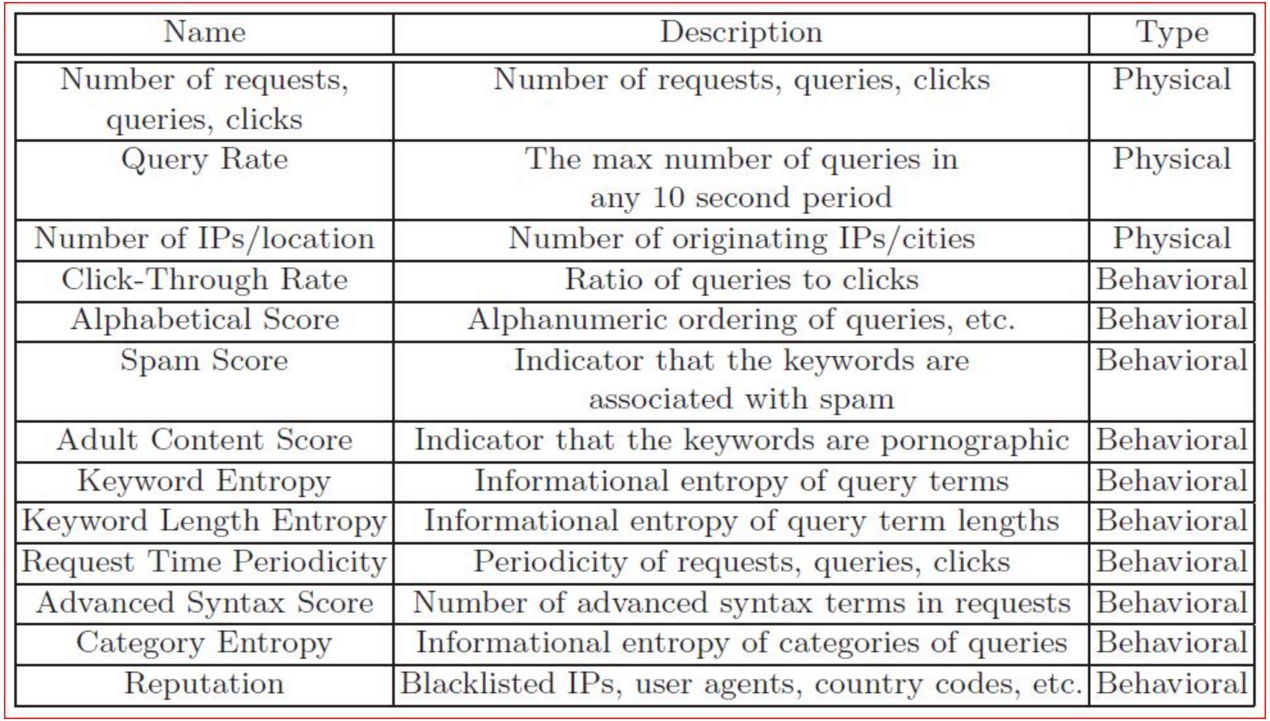


Queries								
pae	cln	eu3	eem	olv	oj	lqde	igf	ief
nzd	rib	xil	nex	intc	tei	wfr	ssg	sqi
nq	trf	cl	dax	ewl	bbdb	csco	pl	idti
nesn	edf	intl	spx	ewj	tasr	ibkr	lat	hb1
mesa	edl	dram	iev	sndk	rukn	ifg	igv	ms

Feature Engineering

- We generally classify these features into two groups
  - Physical model of a human
  - Behavioral patterns of bots

Quantitative Analysis



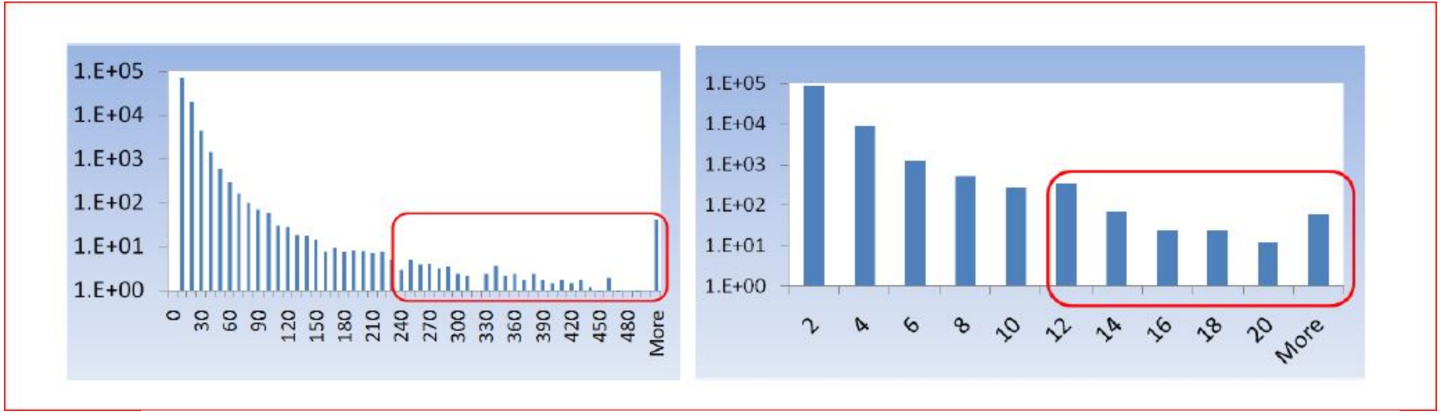
Name	Description	Type
Number of requests, queries, clicks	Number of requests, queries, clicks	Physical
Query Rate	The max number of queries in any 10 second period	Physical
Number of IPs/location	Number of originating IPs/cities	Physical
Click-Through Rate	Ratio of queries to clicks	Behavioral
Alphabetical Score	Alphanumeric ordering of queries, etc.	Behavioral
Spam Score	Indicator that the keywords are associated with spam	Behavioral
Adult Content Score	Indicator that the keywords are pornographic	Behavioral
Keyword Entropy	Informational entropy of query terms	Behavioral
Keyword Length Entropy	Informational entropy of query term lengths	Behavioral
Request Time Periodicity	Periodicity of requests, queries, clicks	Behavioral
Advanced Syntax Score	Number of advanced syntax terms in requests	Behavioral
Category Entropy	Informational entropy of categories of queries	Behavioral
Reputation	Blacklisted IPs, user agents, country codes, etc.	Behavioral

Physical Features

- Number of Queries, Clicks, Page Views etc.
- Query Rate
- Number of IP Addresses / Locations

Physical: Count of Queries and Clicks

- A user can submit 100 queries a day, but it occurs with an unnatural probability

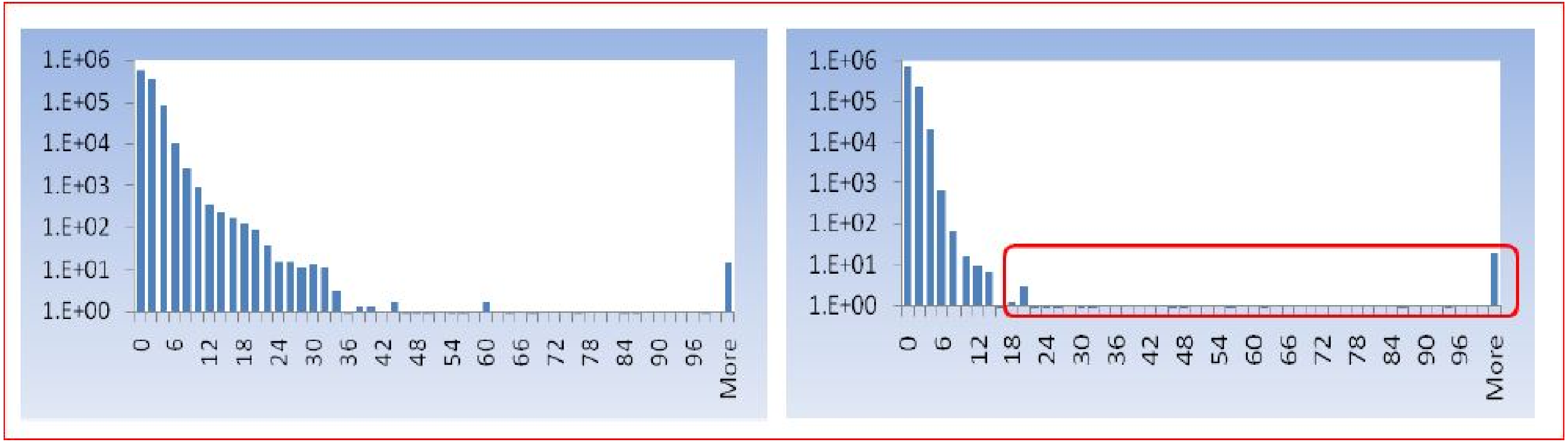


Number of requests (left), and maximum requests in any 10 second interval (right).

Physical: No. of IP Addresses / Locations

- A human cannot be in two distant places at the same time(or in a short interval)
- What if a user’s cookie is compromised and used to make queries from different geographies?

Physical: No. of IP Addresses / Locations



Distinct IP address (all four octets) (left), and distinct IP address (first two octets)

Behavioral Features

- Click-through Rate
- Alphabetical ordering queries
- Spam score

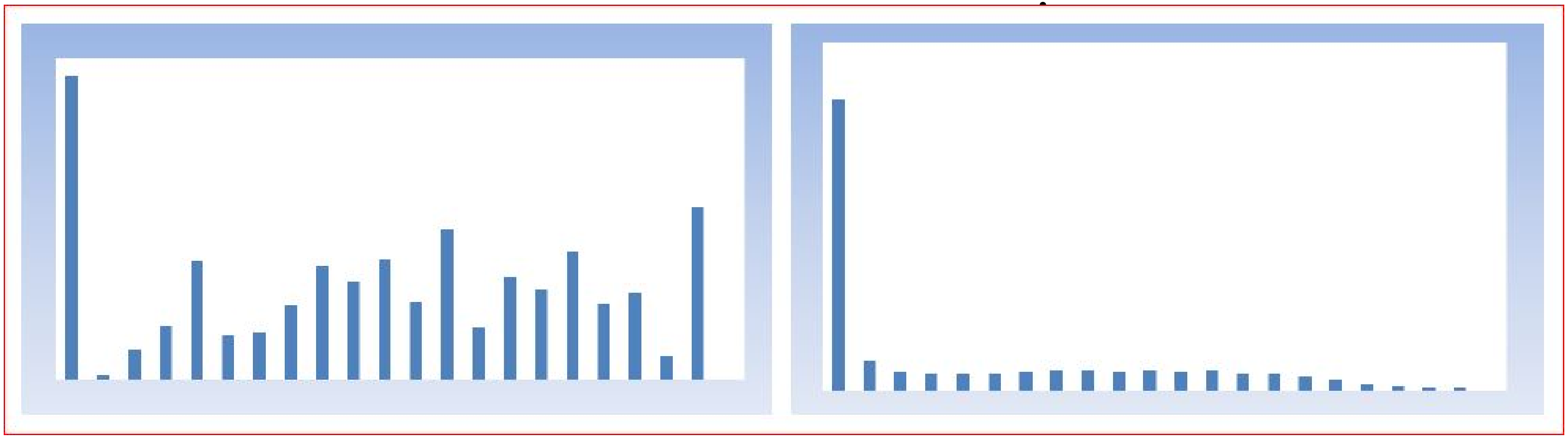
Behavioral: Click-through Rate

- A bot that clicks on no links
- A bot the clicks on every link
- A bot that clicks on targeted links

Behavioral: Click-through Rate

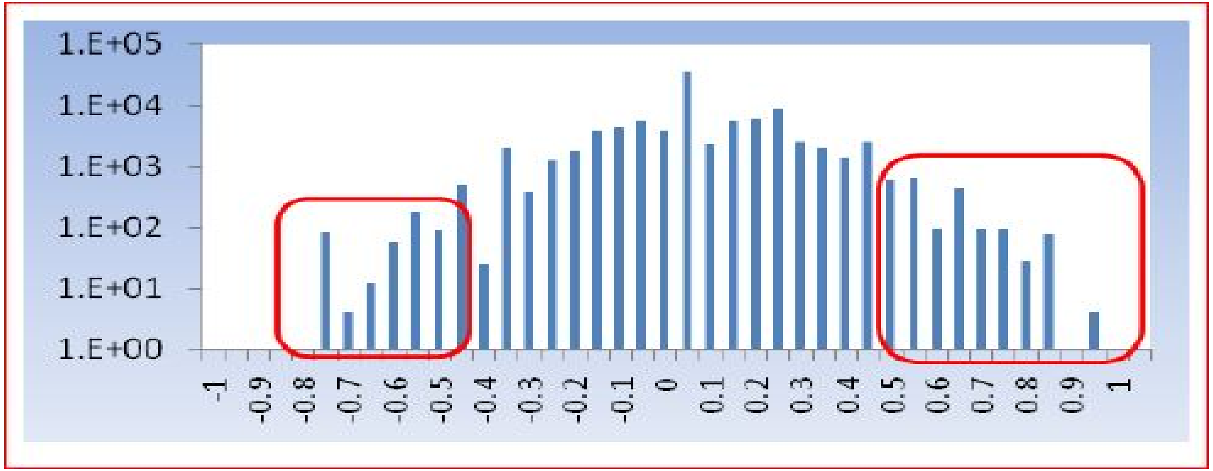
- Click through rates for all users

- Those users with ten times as many



Behavioral: Alphabetical ordering

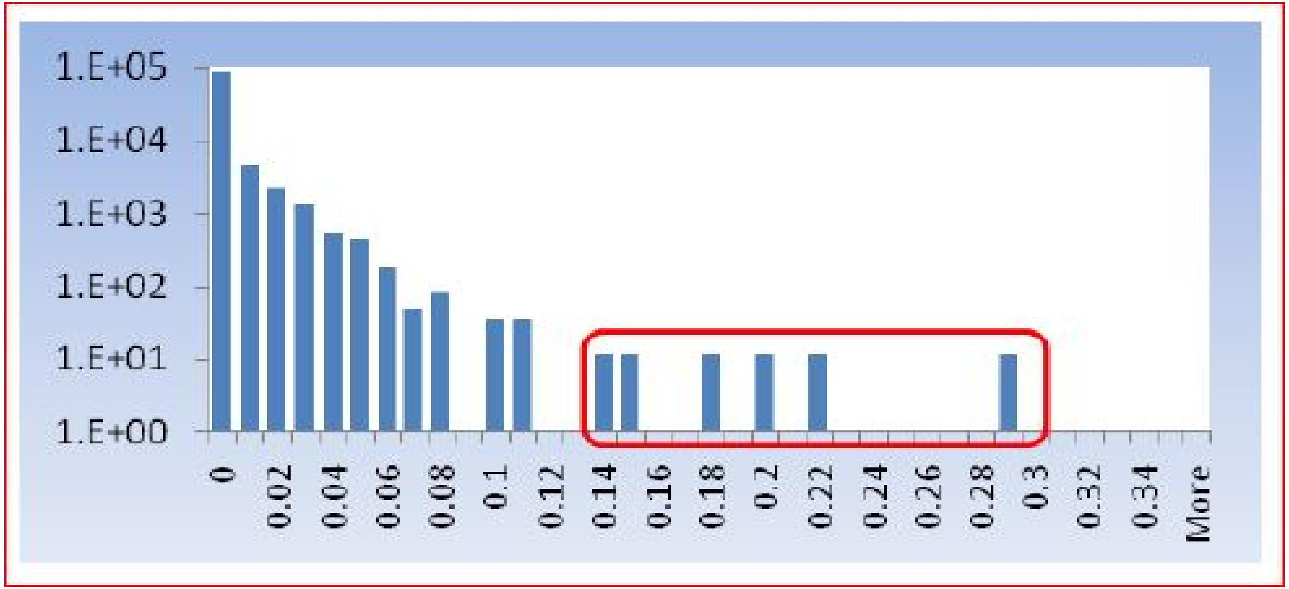
- Order the queries chronologically for each pair



datasciencedojo
data science for everyone

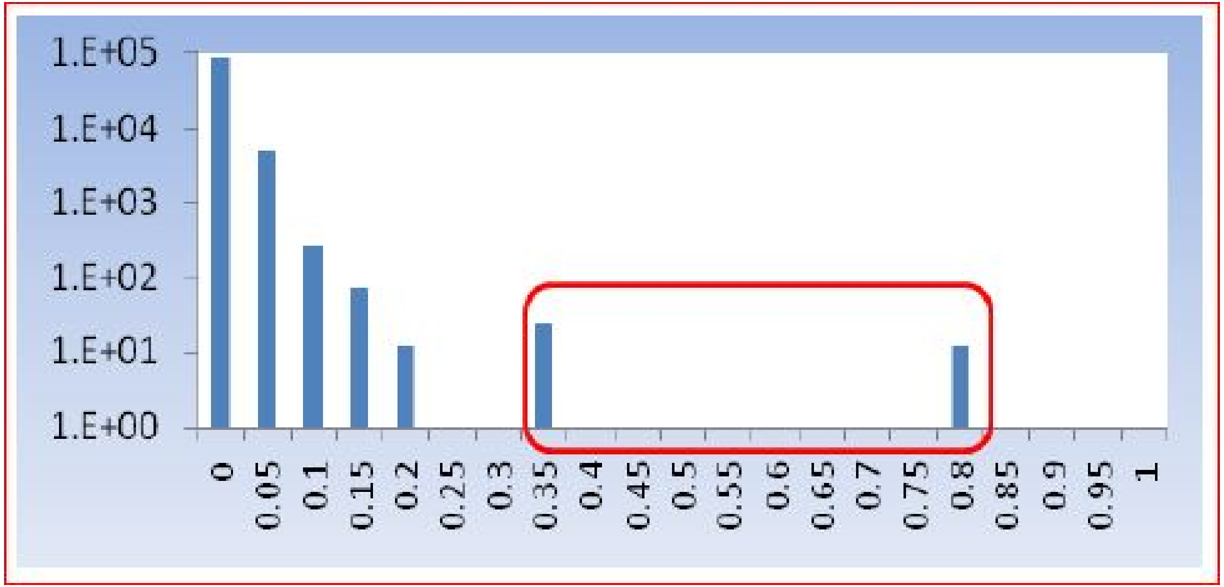
Behavioral: Spam score

- By using a bag of <spam words, weights)



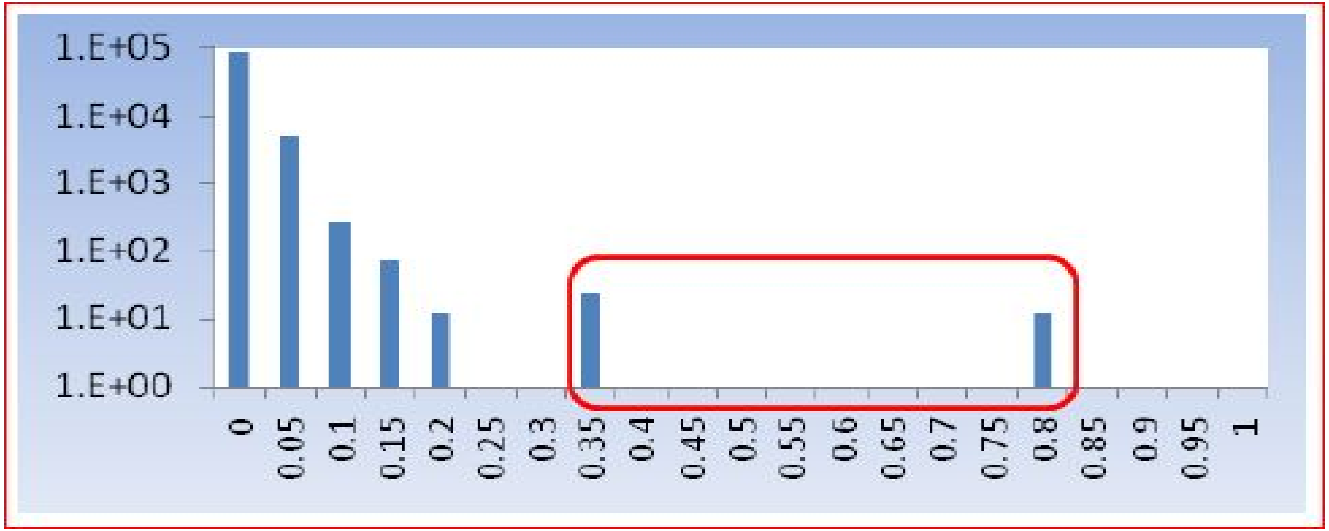
Behavioral: Adult content scores

- By using a bag of <adult word, weight)



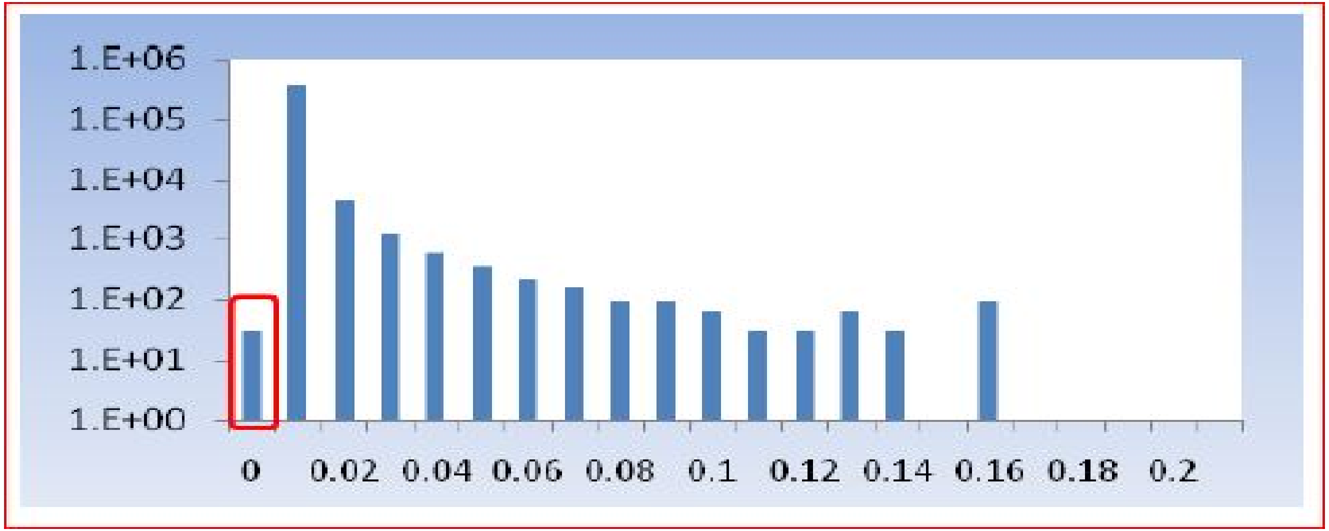
Behavioral: Query keyword entropy

- Map of <word, count> pairs for each userID



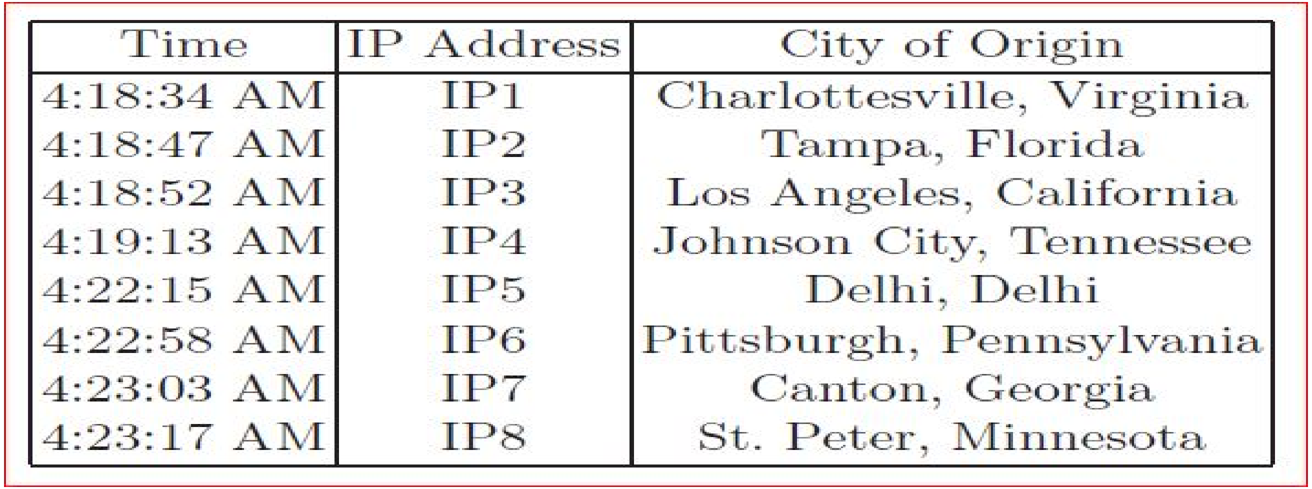
Behavior: Query length entropy

- If the word lengths are roughly the same



Behavior: Varying Geography

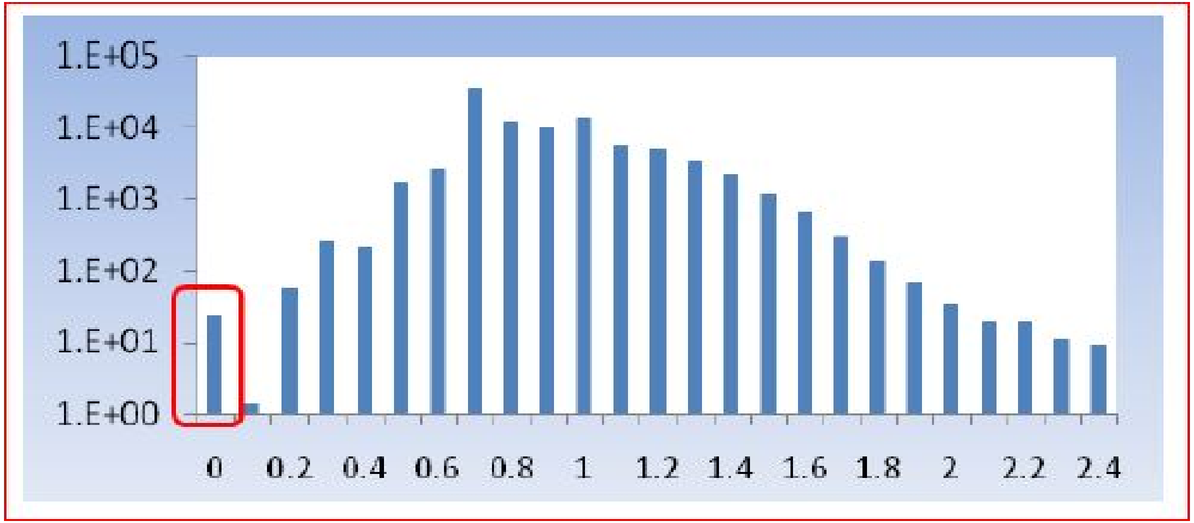
- Attempting to automate traffic through anonymous browsing tools



Time	IP Address	City of Origin
4:18:34 AM	IP1	Charlottesville, Virginia
4:18:47 AM	IP2	Tampa, Florida
4:18:52 AM	IP3	Los Angeles, California
4:19:13 AM	IP4	Johnson City, Tennessee
4:22:15 AM	IP5	Delhi, Delhi
4:22:58 AM	IP6	Pittsburgh, Pennsylvania
4:23:03 AM	IP7	Canton, Georgia
4:23:17 AM	IP8	St. Peter, Minnesota

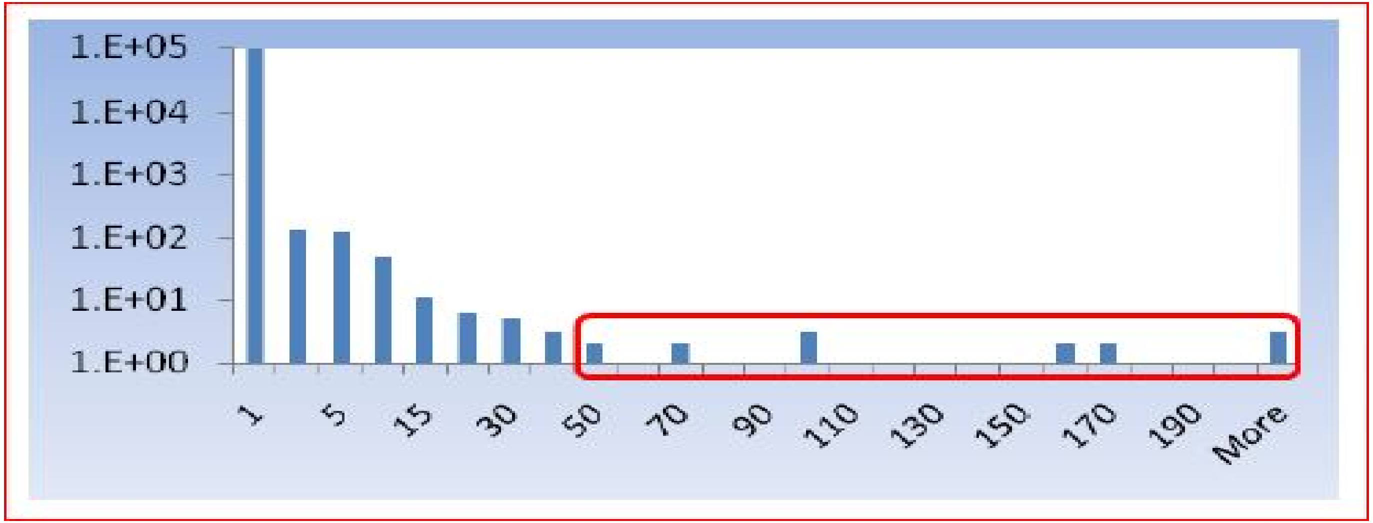
Behavior: Query Time Periodicity

- Capture requests at regular interval say 15 minutes



Behavior: Advanced query syntax

- Keep a total count of all advanced terms for each user throughout the day



Behavior: Category entropy

- Capturing the number of distinct categories associated with a userID
- Assigning category hierarchy to each query

Reputation and trends

- Black listed ip–addresses and user agents

Questions?

Questions?