

Predictability a Predicament?

Claudia Perlich
@claudia_perlich

This presentation is for illustrative and discussion purposes only. This presentation and any accompanying information is high-level in nature and may not be complete. Rather, such information is meant as a summary discussion of potentially complex concepts, systems and/or processes. Moreover, Two Sigma may have market views or opinions (e.g., in respect of investment theses or particular securities discussed herein) that materially differ from those discussed, and may have a significant financial interest in (or against) one or more of such positions or theses. This document does not contain certain material information about the systems, processes and/or techniques employed by Two Sigma and/or the products and strategies managed by Two Sigma, including important disclosures and risk factors associated with an investment in connection therewith.

The information contained herein is not intended to provide, and should not be relied upon for investment, accounting, legal or tax advice. The information provided does not purport to advise you personally concerning the nature, potential, value or suitability of any particular sector, geographic region, security, portfolio of securities, transaction, investment strategy or other matter. You are encouraged to make an independent investigation as to the matters discussed, including by consulting your own tax, legal, accounting and other advisors.

Certain information contained herein is based on data obtained from third parties and, although believed to be reliable, has not been independently verified by anyone at or affiliated with Two Sigma Investments, LP (“TSI”) or Two Sigma Advisers, LP (“TSA,” and together with TSI, the “Firm” or “Two Sigma”), and its accuracy or completeness cannot be guaranteed.

This presentation may make reference to the use of simulations. Simulations have inherent limitations, including that they are designed with the benefit of hindsight and do not represent actual trading. In addition, certain factors will likely cause the simulations to generate better performance than live trading of the associated strategy. These factors include, but are not limited to, the following: (i) the simulations benefit from the use of the most recent models, research and optimizer settings used by Two Sigma, which have been historically-tested and in most cases have been fit to a period similar to the one shown and, therefore, benefit from portfolio settings that have been found to be optimal for such time period; and (ii) the lack of operating expenses and other expenses which, if included, would degrade performance in a compounded fashion over time. Actual returns will be different than those of the simulations. In addition, the interpretation of simulated performance results is an inherently subjective process, requires significant interpretation by portfolio management personnel, and is ultimately based upon the knowledge, expertise and subjective beliefs of portfolio management personnel about the workings of the strategies, techniques and markets.

This presentation does not constitute an offer to sell or the solicitation of any offer to buy any security or other interest, which may only be made at the time a qualified offeree receives a confidential offering memorandum describing the offering, the applicable subscription agreement and certain related materials (the “Offering Document(s)”). This presentation and any associated information is qualified in its entirety by the such Offering Document(s). In the case of any inconsistency between the descriptions or terms in this presentation and those contained in the applicable Offering Documents (as may be amended from time to time), the applicable Offering Document shall control.

This presentation is the property of Two Sigma, should be used solely for discussion purposes, and any unauthorized distribution of this information without the express written consent of Two Sigma is prohibited.

The information is presented as of the date hereof unless otherwise indicated. Any statements set forth herein (or made in connection with this presentation) regarding future events constitute only subjective views or beliefs, should not be relied upon, and are subject to change due to a variety of factors, many of which cannot be predicted or quantified and are beyond the control of Two Sigma, its affiliates, and agents. Future actual results could differ materially from any discussed herein (or made in connection with this presentation), and no assurances are given that these statements are now, or will prove to be, accurate or complete, in whole or in any part.

Some of the images, copyrights or trademarks used herein may qualify for fair usage in the context described below. Some of these images or logos may be protected by copyright and/or trademark. Such copyright or trademark is most likely owned by the company or corporation that produced the image or logo. It is believed that the use of a limited number of these images and/or logos purely for identification and visual association qualifies as fair use under international copyright laws. Use of such image, copyright or trademark does not imply any association with such organization (or endorsement of such organization) by Two Sigma, nor vice versa.

The recipient is reminded that past performance is not an indicator or guarantee of future results. There is no guarantee that the concepts, methodologies or strategies discussed (or any vehicle employing such strategies) will be able to achieve its investment objective or that any potential investor will make a profit, or will not sustain losses, including a total loss of their investment.



How 'predictable' is a predictive modeling task? (given the data)

Pizza for
Dinner?



Random

We can't choose our lives, but
WE CAN DECIDE
what to do with the joys
or griefs we're given.

Paulo Coelho

SYMPHONY OF LOVE
Photo by Unsplash

Sexual
Orientation



Deterministic

2σ

For now let's try to predict who is male ...



Predicting Probability (Male) in Facebook



```
00000d41ed774823fca142945ec915c0,1,,,,,,,,,en_GB,,
00000dee02d70cf8c0d79f96b6d1c59d,0,,,,,,,,,en_US,,
00000f232abfe25a80156fe069395460,0,1992,20,2,2,,,,,,,,,19,-5
00000f4ba0cff946b1c0e3b051287ede,0,1993,19,2,,,,,,,,,en_US,310,8
0000130571654e3afaa62f4e9d2e4f63,0,,,2,2,,,,,,,,,en_US,193,7
00001544469ae9b408869a463a1dd77a,1,1984,28,2,,,,,,,,,en_US,,-4
```

40% Men

Data:

Facebook public dataset with 200K anonymized users, their demographics and their likes

Methodology:

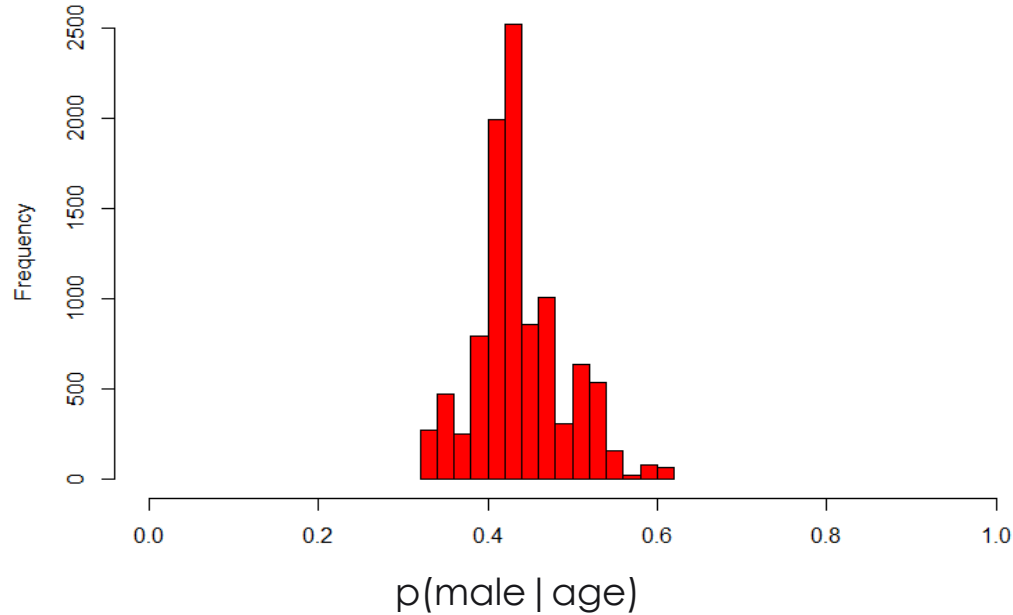
Logistic regression on sparse representation

```
00001544469ae9b408869a463a1dd77a 100198443380917
00001544469ae9b408869a463a1dd77a 100248613695
00001544469ae9b408869a463a1dd77a 10050726267
00001544469ae9b408869a463a1dd77a 101021248409
00001544469ae9b408869a463a1dd77a 101054236602446
00001544469ae9b408869a463a1dd77a 101425333232551
00001544469ae9b408869a463a1dd77a 10148199466
00001544469ae9b408869a463a1dd77a 10150154095435553
00001544469ae9b408869a463a1dd77a 10161539667
00001544469ae9b408869a463a1dd77a 101844593108
00001544469ae9b408869a463a1dd77a 101936079845392
00001544469ae9b408869a463a1dd77a 101987301816
00001544469ae9b408869a463a1dd77a 102038567018
00001544469ae9b408869a463a1dd77a 102040023230884
00001544469ae9b408869a463a1dd77a 10212595263
00001544469ae9b408869a463a1dd77a 102168219824412
00001544469ae9b408869a463a1dd77a 1022548733
```

Take 1: Predict Gender Based on age ...



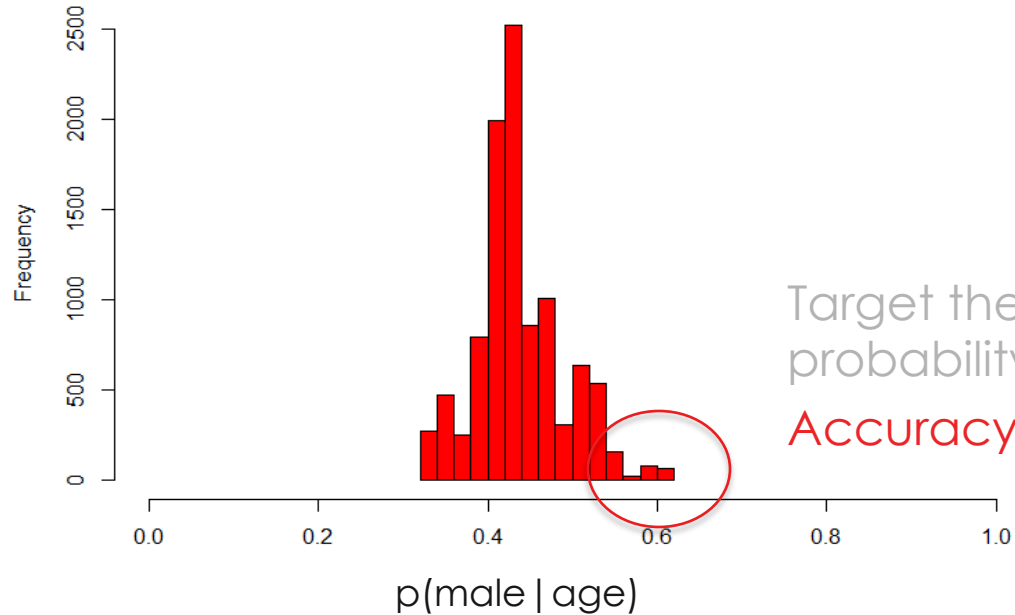
Gender based on age: very little signal ...



Overall Accuracy: 60%

AUC: 58%

Gender based on age: very little signal ...



Target the 1% with highest probability:

Accuracy: 75%

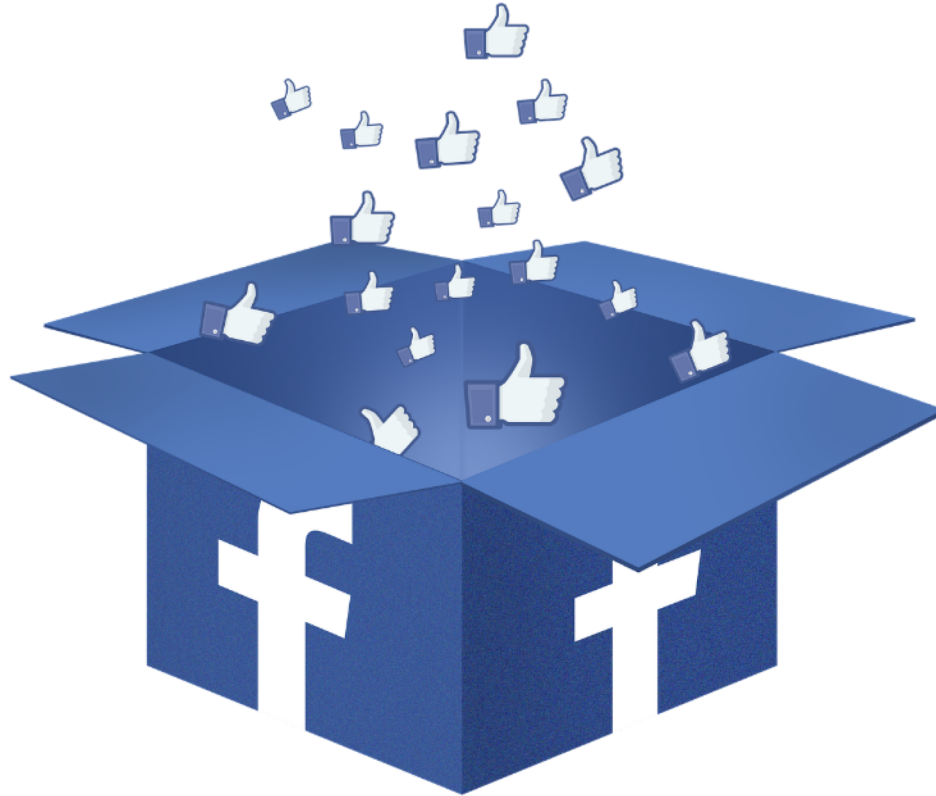
Overall Accuracy: 60%

AUC: 58%

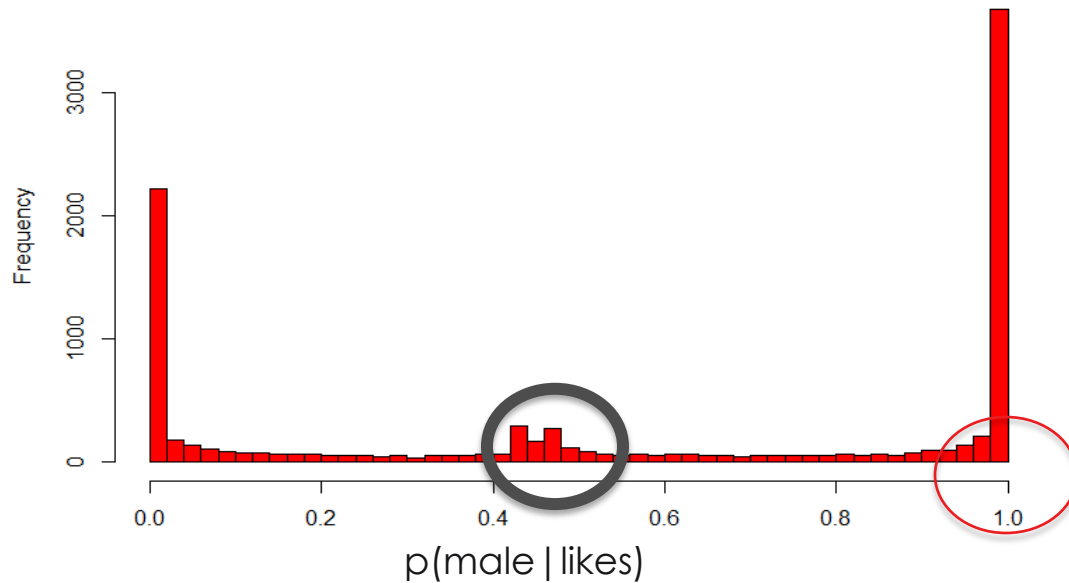
From small to bigger data ...



Take 2: Gender based on all your likes



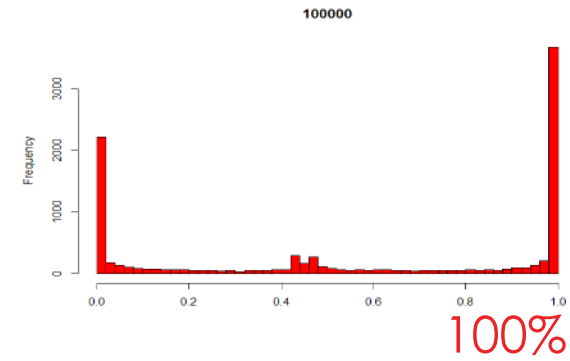
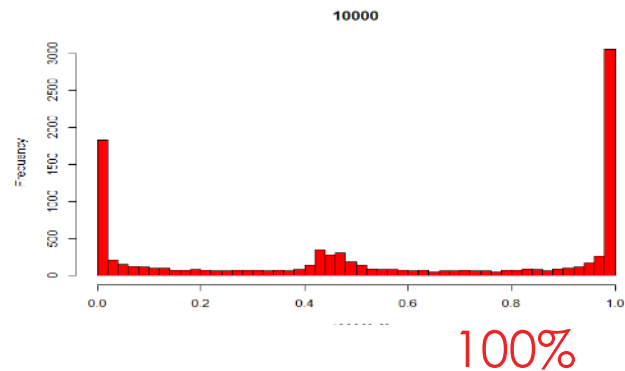
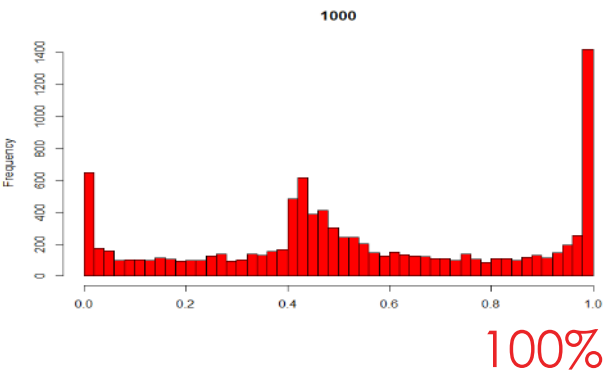
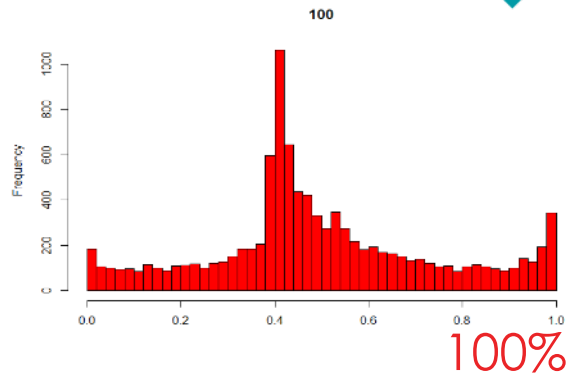
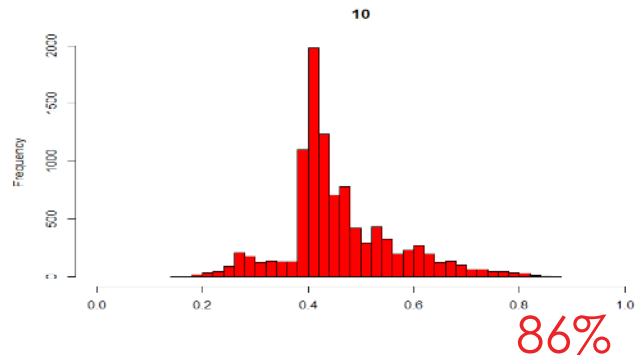
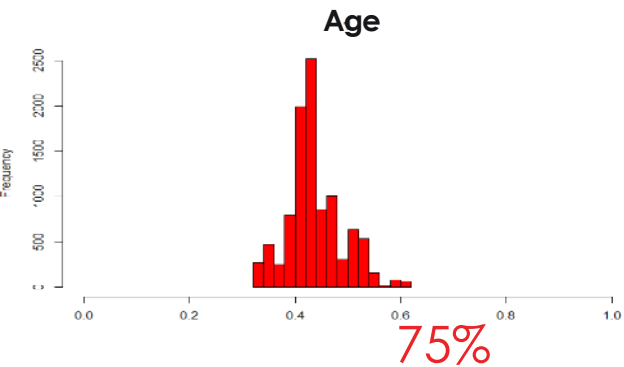
Predict gender based on all likes: a lot of signal ...



Target top 1%
Accuracy: 100%

Overall Accuracy: 83%

Progression: from age to all 'likes'



But what happens if your problem is a mixture of both?



Random



Deterministic

Models tend to focus where the signal is ...



Witness a spike in human predictability ...



Death of
free will?

Predicting digital events based on browsing histories



SGD Model (Model ID: 1728493)

Marketer	Microsoft (Empowering Media SF) [6000]
Offer	Microsoft Azure Evergreen [7599]
Offer Level	40: Free Trial [19945]
Model Created At	2016-03-06 12:36:39
Model Event End Time	2016-03-06 17:40:00
# Positive	217
# Negative	2,170
# Features	362,059
Parent Model	1728493
HDFS Path	hdfs://hdfs2/user/sgd/models/19945/OL19945_FM2_BS20160306123636.model

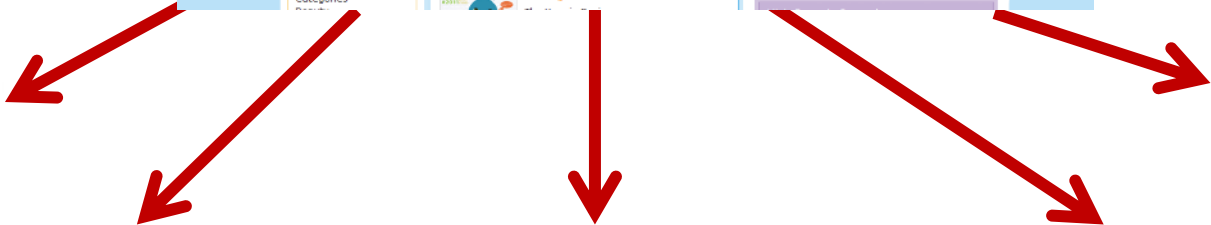
Logistic Regression
 Stochastic gradient descent
 Hashing
 Streaming
 L1 & L2 Penalties

Top Features with Highest Scores

Show 10 entries

Search:

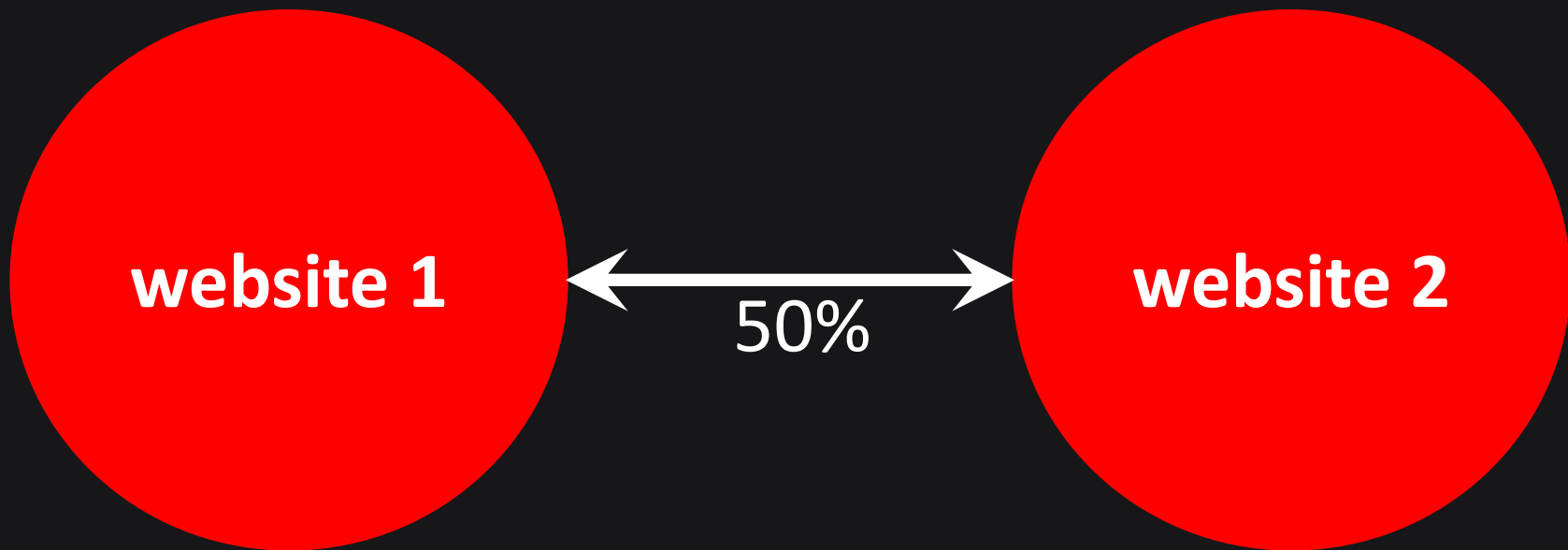
Feature	Coef
[46303095]	3.117
blogs.msdn.com [33892607]	3.010
www.why-recycle.com www.technet.com [46550201]	2.932
www.lz95.org [35757707]	2.779
www.dorisandjillycook.com www.dotnetfunda.com www.edenprairiecenter.com [35375642]	2.508
www.onebox.com www.avanquest.com [35325508]	2.467
www.office.com [46380052]	2.410
[46305049]	2.297
www.verbasoftware.com [46446087]	2.294
www.rosebyrne.org www.socialapocalypse.com www.milfmovies.biz www.milwaukee-wisconsin-wi.com www.msexchange.org [35355270]	2.288

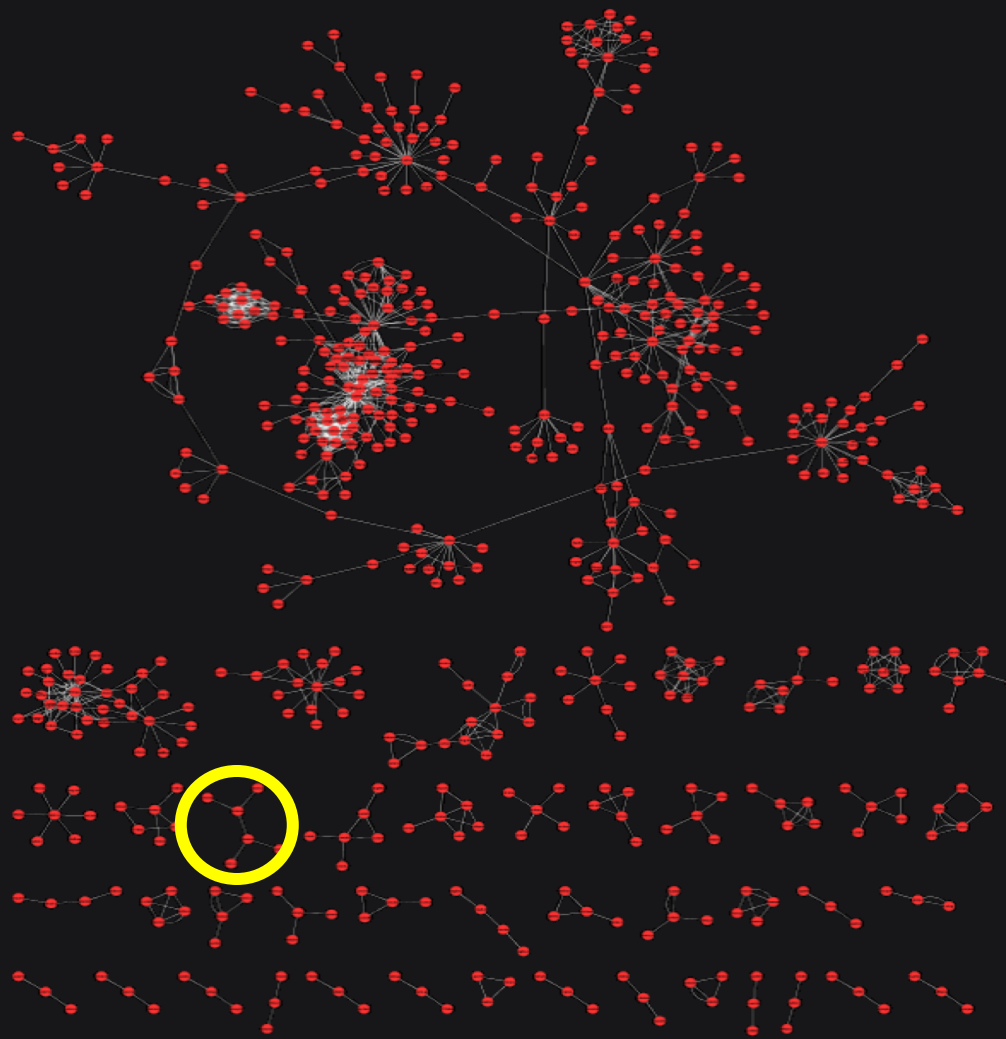


URL's that are very predictive for more than 10 brands

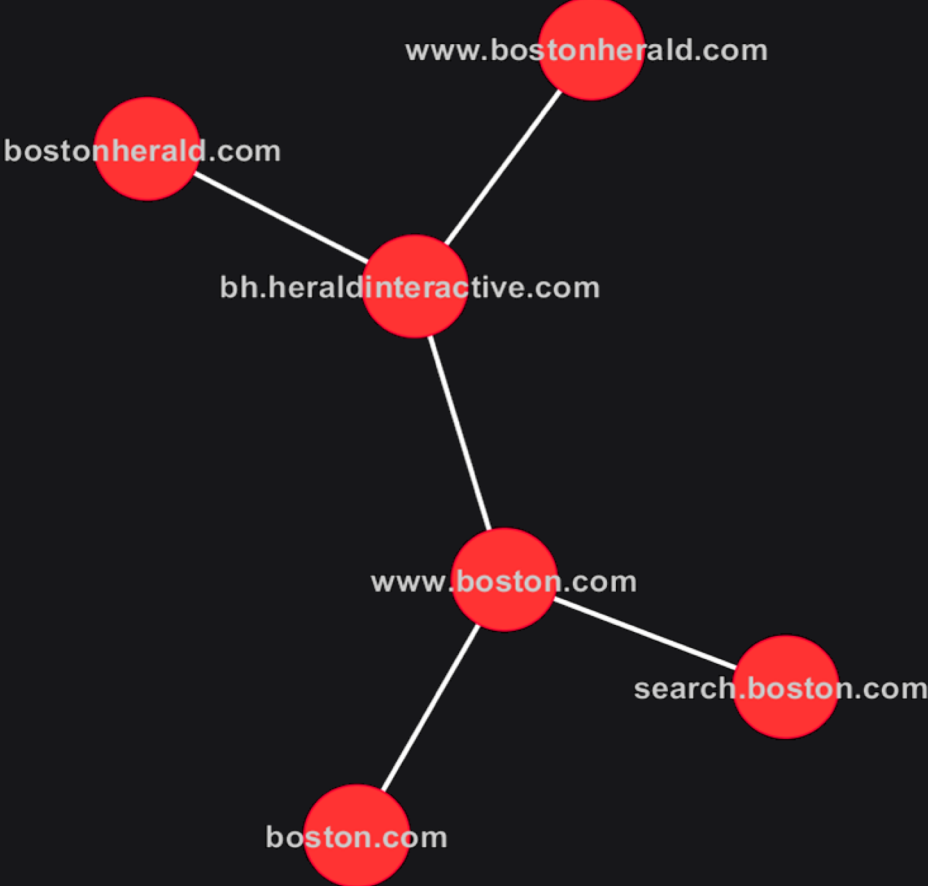
www.womenshealthbase.com
www.filmannex.com
www.ffog.net
www.drugsnews.org
www.menshealthbase.com
www.dailyfreshies.com
www.hark.com
www.gossipcenter.com
www.articletrunk.com
www.411answers.com
www.dailyrx.com
www.all-allergies.com
www.knowvehicles.com
www.chinaflix.com
www.parentingnewsstories.com
www.wrestlingnewz.com
www.gourmandia.com

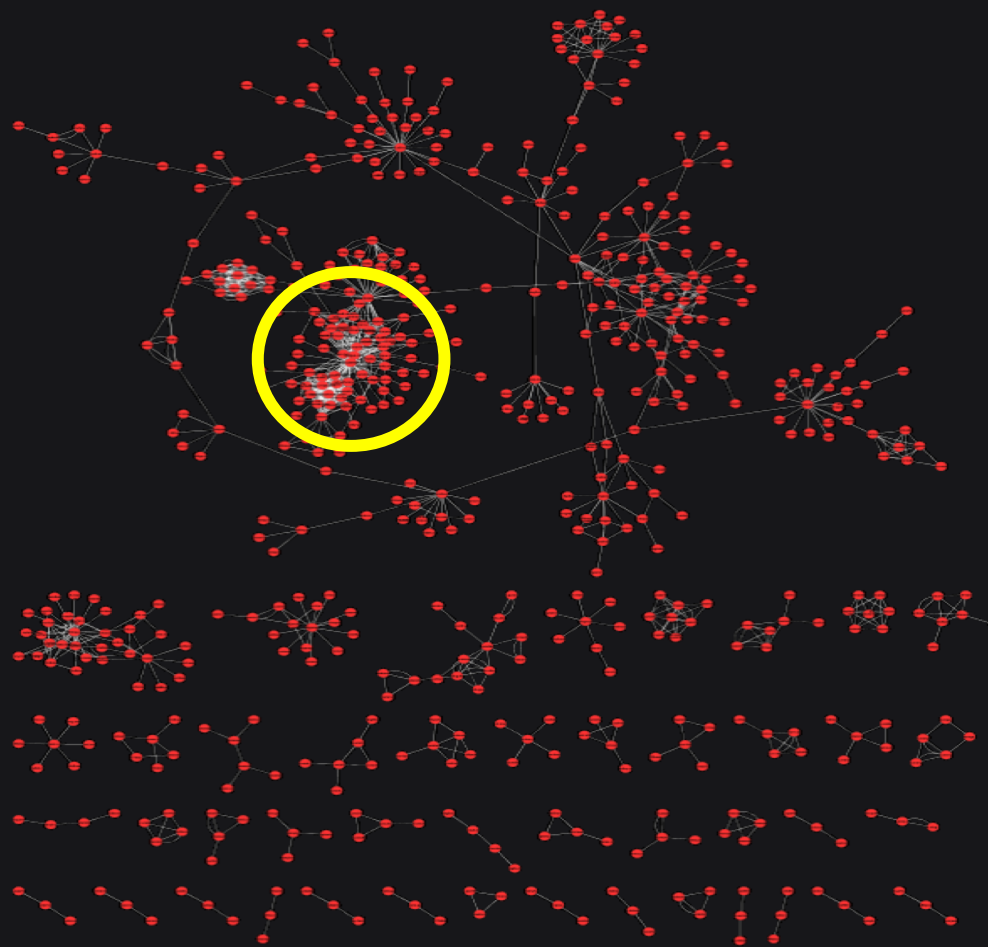
Traffic overlap of cookies from Bid Request





Boston Herald





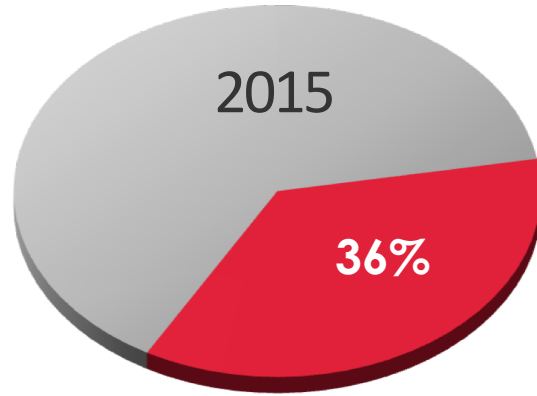


But what has that to do with predictability of brands?

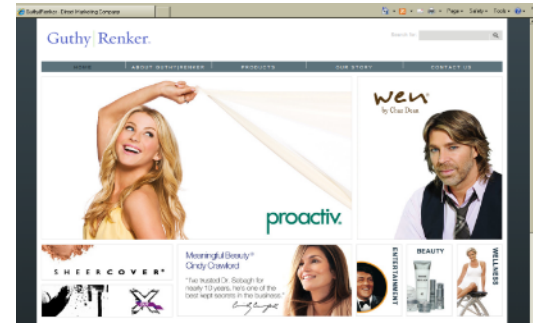
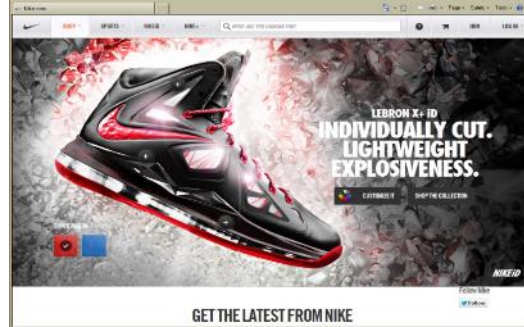
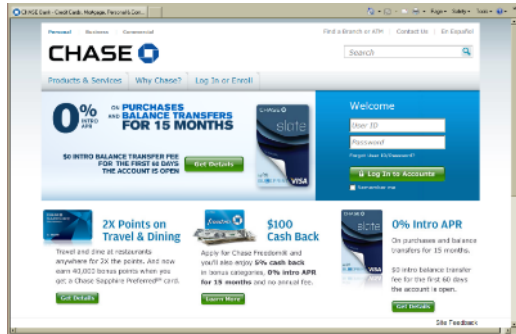


Death of
free will?

Two populations surfing the we: Bot vs. Human

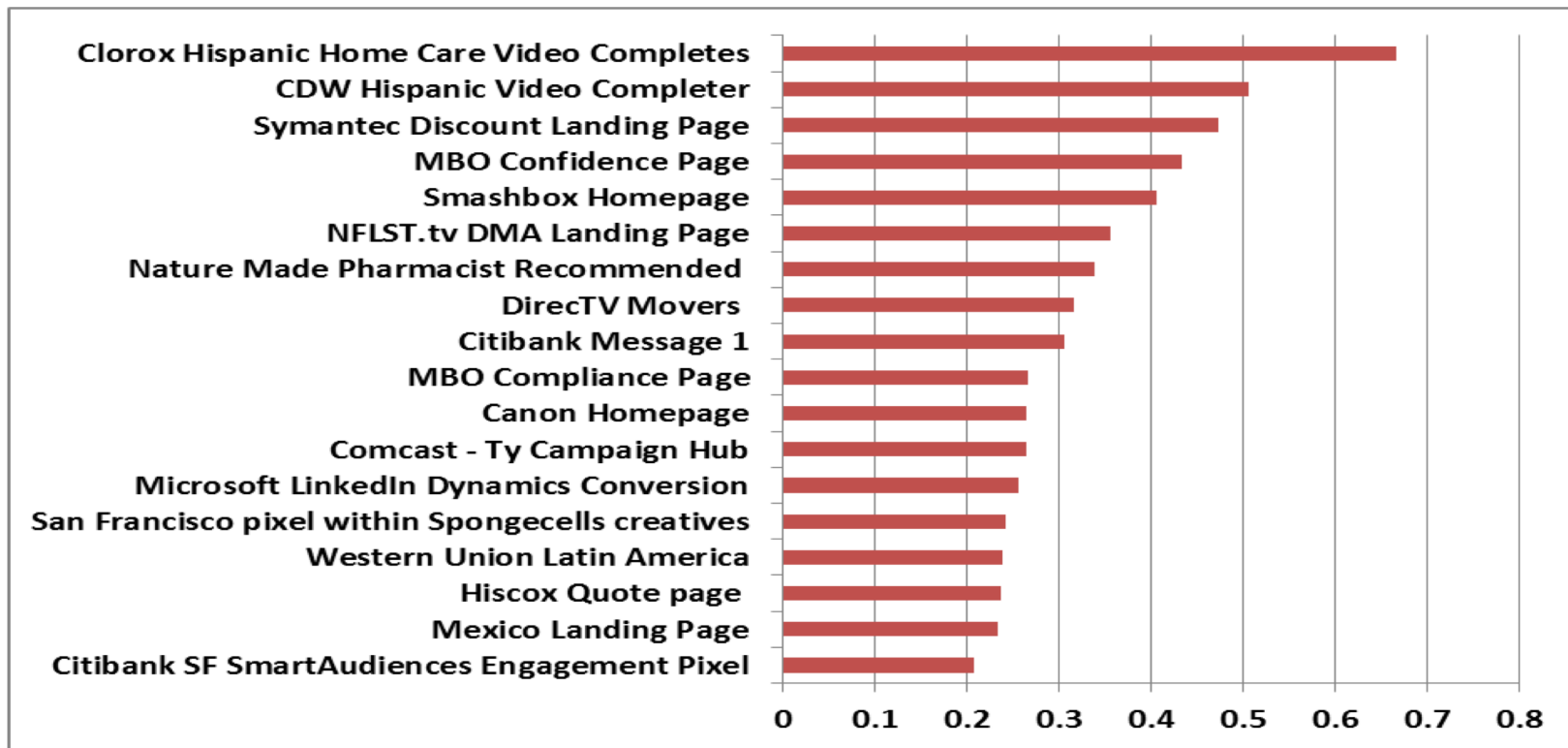


Bots are executing conversion events



- 'Cookie Stuffing' increases the value of the ad for retargeting
- Messing up Web analytics ...
- Messes up my models because a bot is easier to predict than a human

Percent bot traffic on conversion metrics

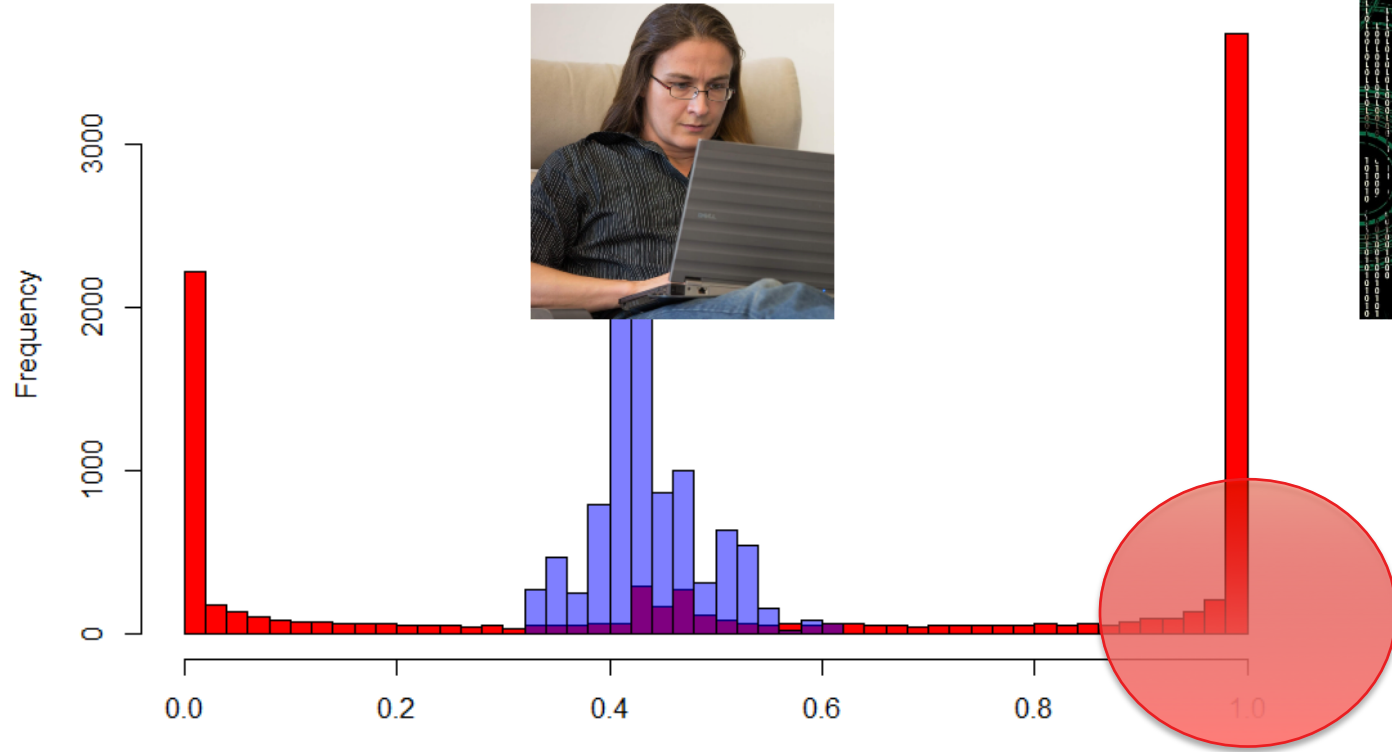




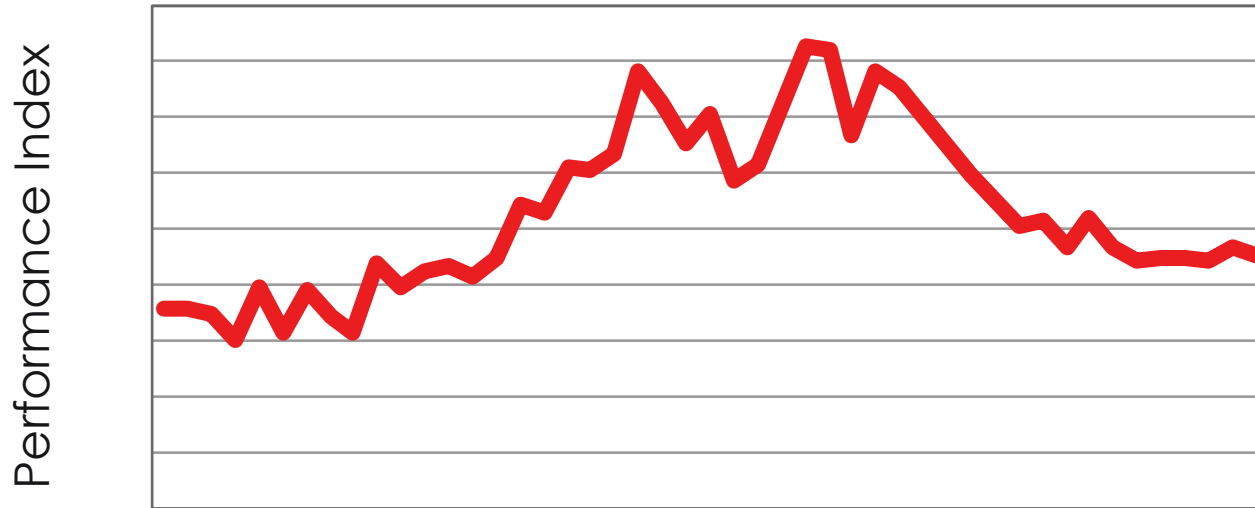
Random

Deterministic

Bot activity has more signal – higher predictions



Eliminate labels generated by bots



3 more weeks in spring 2012

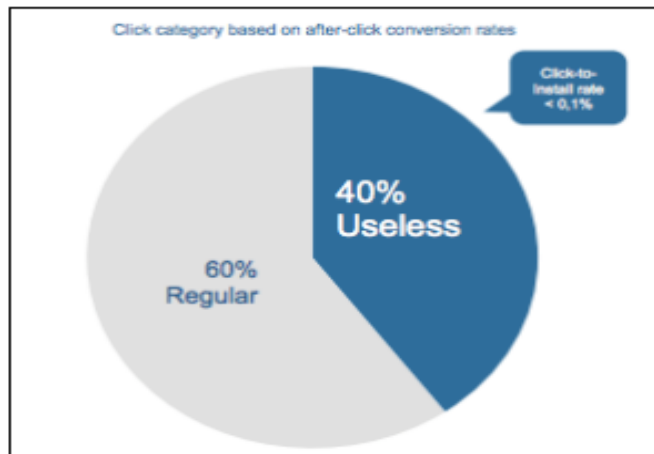
What about Clicks?

“Measure of consumers interest in the product”



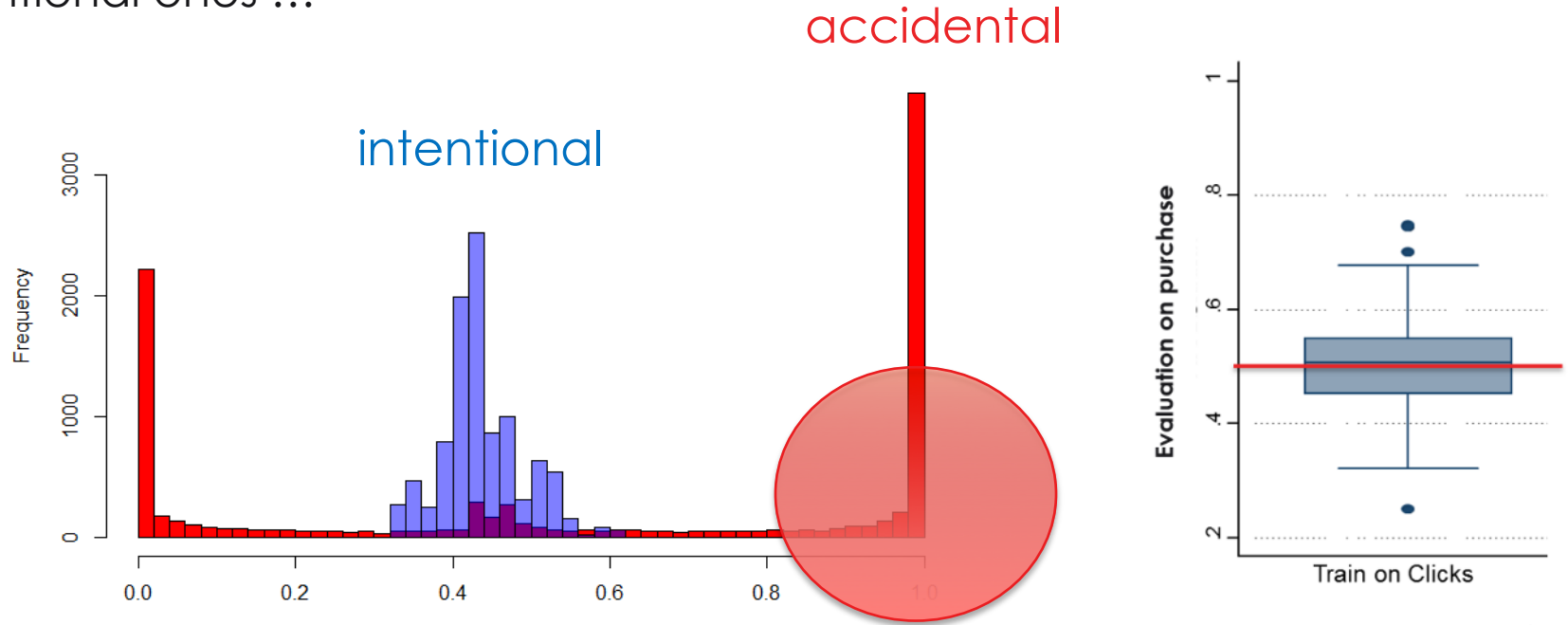
High percentage of Accidental Clicks on Mobile Ads

A study by the Trademob mobile app marketing company showed that 40% of mobile ad clicks are either accidental or fraudulent.

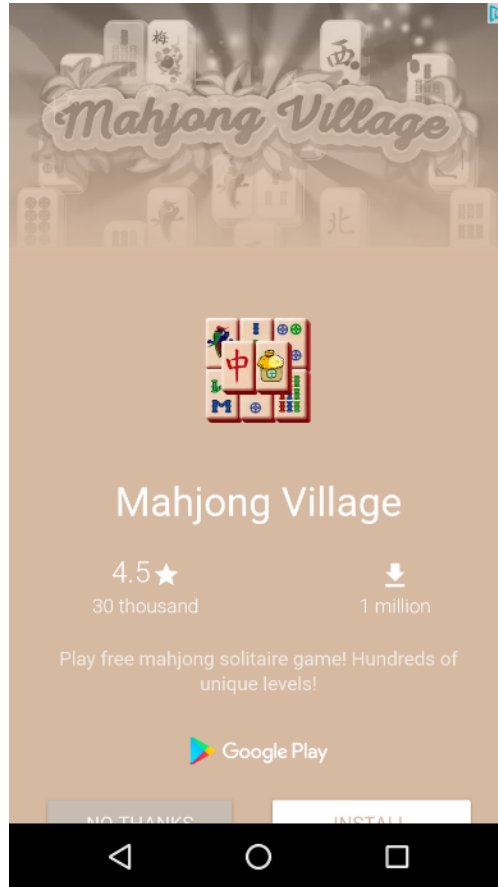


Accidental clicks are more predictable than intentional

Model learns to predict the accidental much more easily than the intentional ones ...

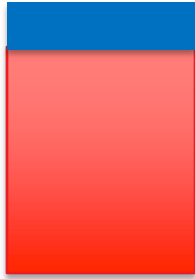


'Optimizing Clicks ...'



Old days of the Click Metric ...

Strategy 1:
target women



intentional

<

accidental

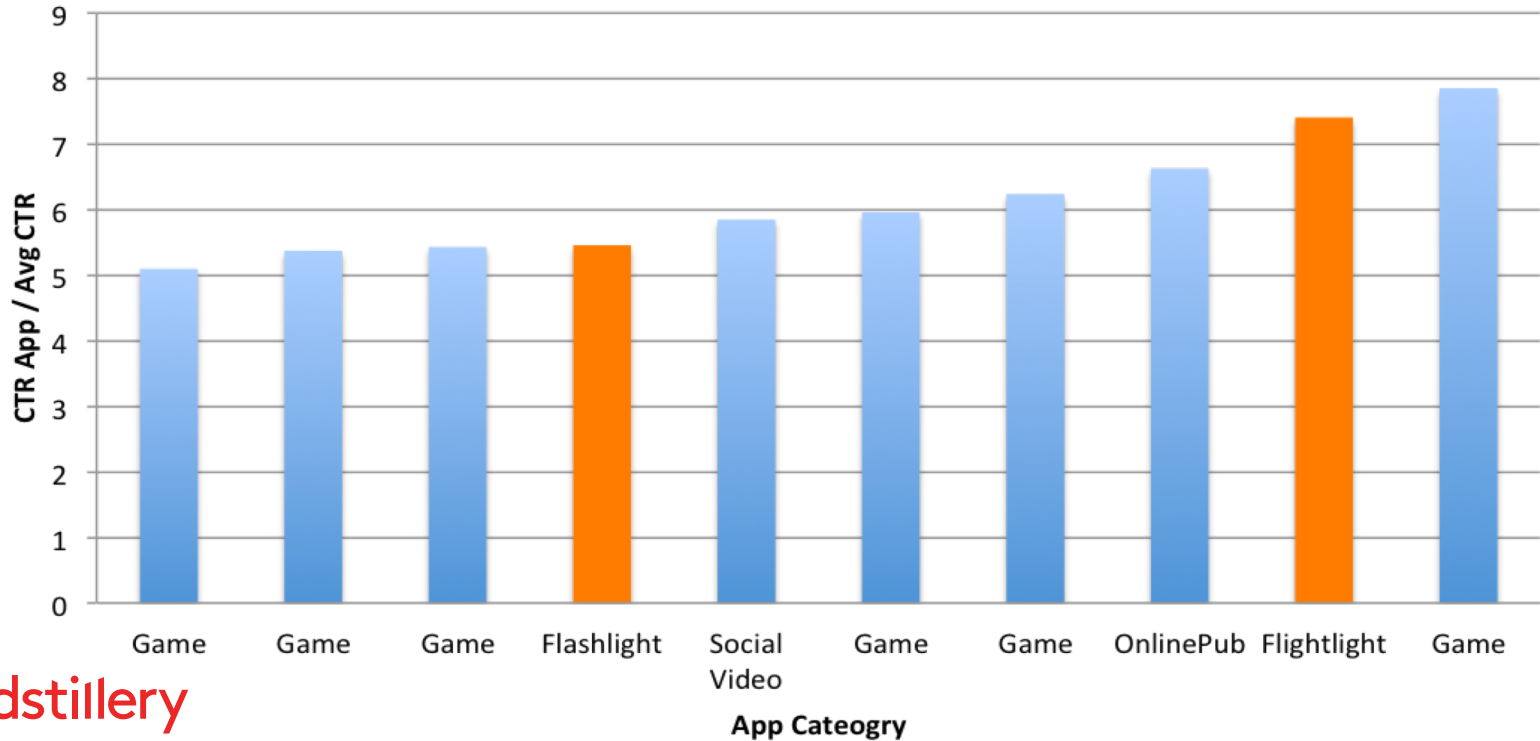


Strategy 2:
target men



Model learned to predict accidents: People fumbling in the dark ...

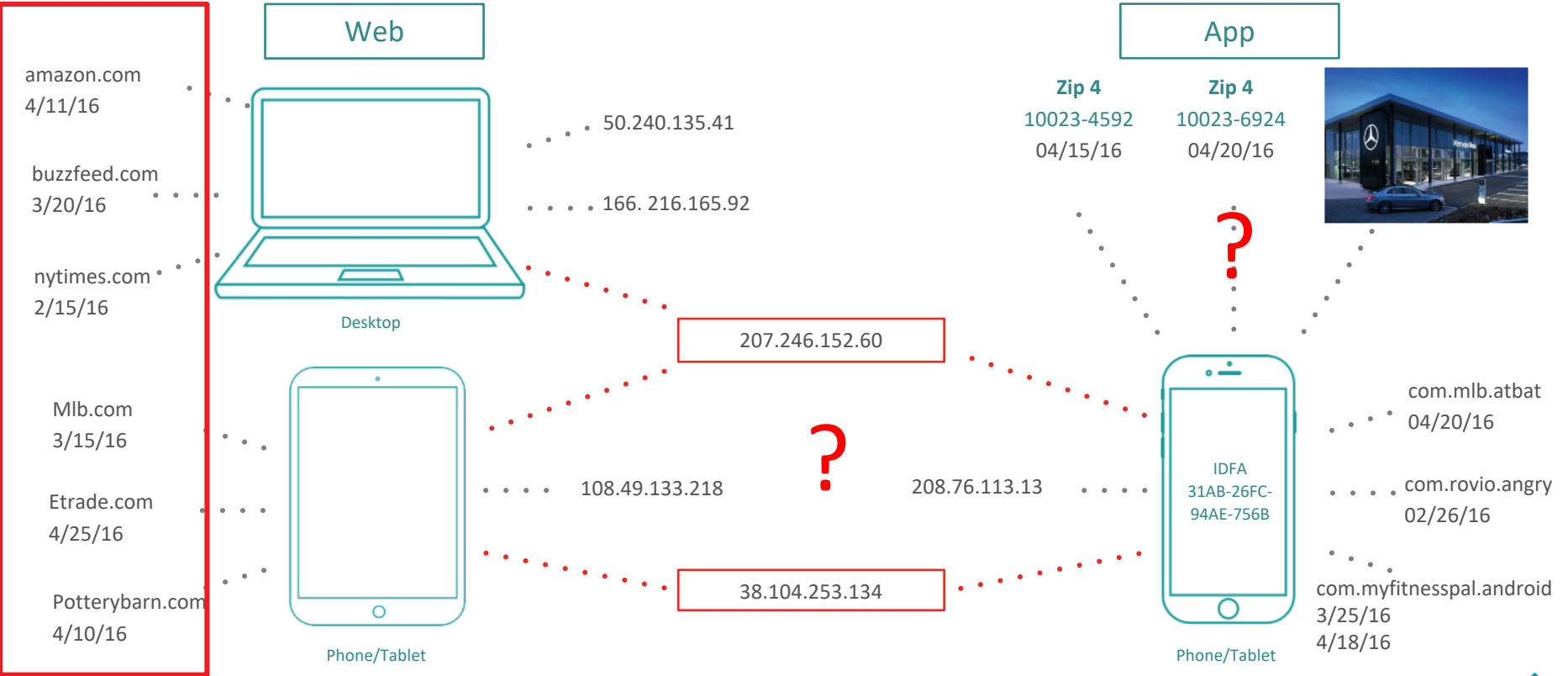
Top 10 Apps by CTR



Finding good audiences for luxury cars? Predict dealership visits?



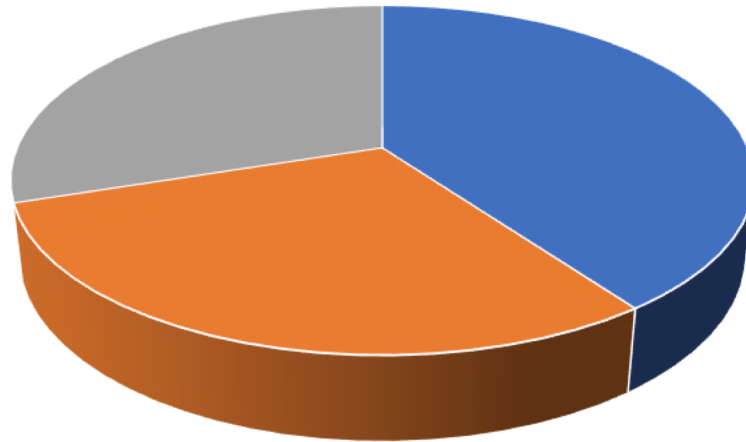
Browsing history of your neighbor who hacked your WIFI?



Potentially three populations in the location prediction

People who are
somewhere close

**People who are indeed at
dealership and their history**



People who hacked into your WIFI

Identify people who will go to Mercedes dealerships



URL	Logistic Parameter
www.mercedesremotestart.com	2.97
www.girardtoyota.com	2.92
www.bluelineprop.com	2.91
www.toyotaoflaredo.com	2.70
www.pikespeakacura.com	2.44
www.motorcarsacura.com	2.39
www.hobsonrealtors.com	2.35
www.fayettevillencmls.com	2.35
www.coast2coastmixtapes.com	2.28
www.fwproperties.com	2.26
www.dealeron.com	2.25
www.smartfitnesscoach.com	2.23
www.pearsonlongman.com	2.21
www.pattersonauto.com	2.14
www.acuraofmodesto.com	2.11
www.f2it.co.uk	2.10
www.keyeshyundai.com	2.09
www.stevenscreekhyundai.com	2.06
www.columbusrealtors.com	2.05
www.whitmanathletics.org	2.05
www.napletonautosales.com	2.02
www.dealerelite.net	2.00
www.valuinsight.com	1.99
www.travel-shows.com	1.98
www.gorealestateagents.com	1.98

'In the market' signal



URL	Logistic Parameter
www.mercedesremotestart.com	2.97
www.girardtoyota.com	2.92
www.bluelineprop.com	2.91
www.toyotaoflaredo.com	2.70
www.pikespeakacura.com	2.44
www.motorcarsacura.com	2.39
www.hobsonrealtors.com	2.35
www.fayettevillencmls.com	2.35
www.coast2coastmixtapes.com	2.28
www.fwproperties.com	2.26
www.dealeron.com	2.25
www.smartfitnesscoach.com	2.23
www.pearsonlongman.com	2.21
www.pattersonauto.com	2.14
www.acuraofmodesto.com	2.11
www.f2it.co.uk	2.10
www.keyeshyundai.com	2.09
www.stevenscreekhyundai.com	2.06
www.columbusrealtors.com	2.05
www.whitmanathletics.org	2.05
www.napletonautosales.com	2.02
www.dealerelite.net	2.00
www.valuinsight.com	1.99
www.travel-shows.com	1.98
www.gorealestateagents.com	1.98

Real Estate



URL	Logistic Parameter
www.mercedesremotestart.com	2.97
www.girardtoyota.com	2.92
www.bluelineprop.com	2.91
www.toyotaoflaredo.com	2.70
www.pikespeakacura.com	2.44
www.motorcarsacura.com	2.39
www.hobsonrealtors.com	2.35
www.fayettevillencmls.com	2.35
www.coast2coastmixtapes.com	2.28
www.fwproperties.com	2.26
www.dealeron.com	2.25
www.smartfitnesscoach.com	2.23
www.pearsonlongman.com	2.21
www.pattersonauto.com	2.14
www.acuraofmodesto.com	2.11
www.f2it.co.uk	2.10
www.keyeshyundai.com	2.09
www.stevenscreekhyundai.com	2.06
www.columbusrealtors.com	2.05
www.whitmanathletics.org	2.05
www.napletonautosales.com	2.02
www.dealerelite.net	2.00
www.valuinsight.com	1.99
www.travel-shows.com	1.98
www.gorealestateagents.com	1.98

Fitness ...



URL	Logistic Parameter
www.mercedesremotestart.com	2.97
www.girardtoyota.com	2.92
www.bluelineprop.com	2.91
www.toyotaoflaredo.com	2.70
www.pikespeakacura.com	2.44
www.motorcarsacura.com	2.39
www.hobsonrealtors.com	2.35
www.fayettevillencmls.com	2.35
www.coast2coastmixtapes.com	2.28
www.fwproperties.com	2.26
www.dealeron.com	2.25
www.smartfitnesscoach.com	2.23
www.pearsonlongman.com	2.21
www.pattersonauto.com	2.14
www.acuraofmodesto.com	2.11
www.f2it.co.uk	2.10
www.keyeshyundai.com	2.09
www.stevenscreekhyundai.com	2.06
www.columbusrealtors.com	2.05
www.whitmanathletics.org	2.05
www.napletonautosales.com	2.02
www.dealerelite.net	2.00
www.valuinsight.com	1.99
www.travel-shows.com	1.98
www.gorealestateagents.com	1.98

How much randomness can a model absorb?

We will randomly switch the gender value for increasing percentages

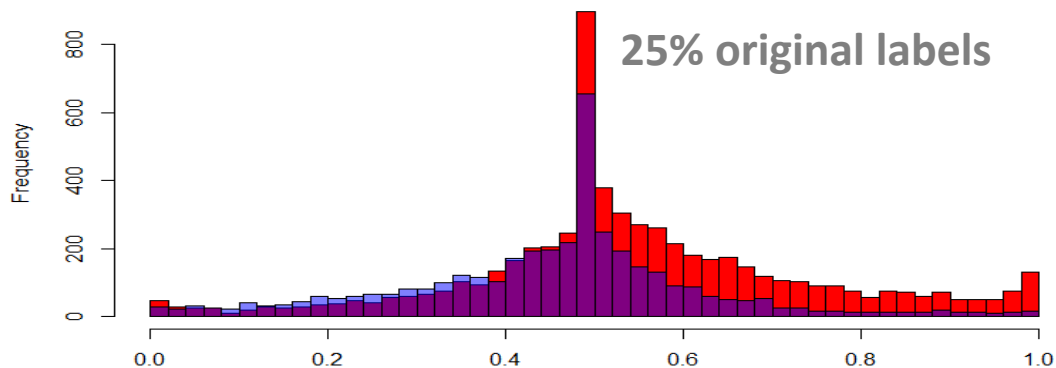
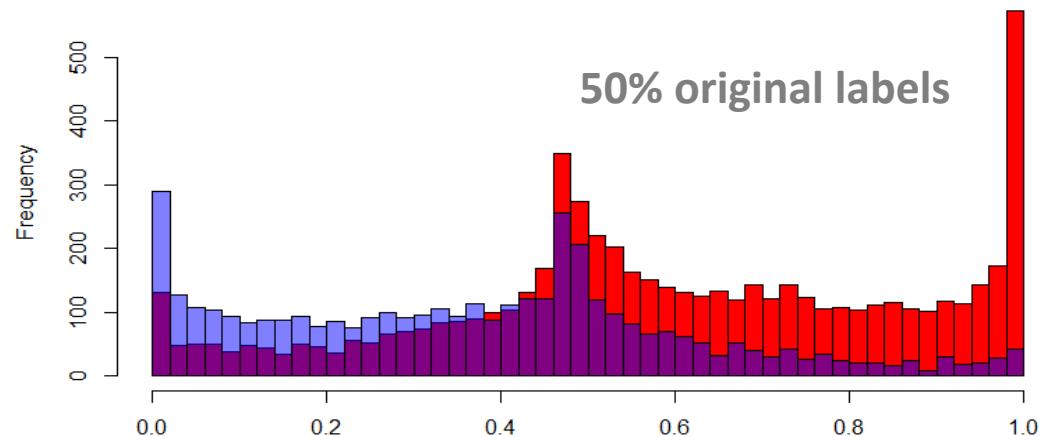
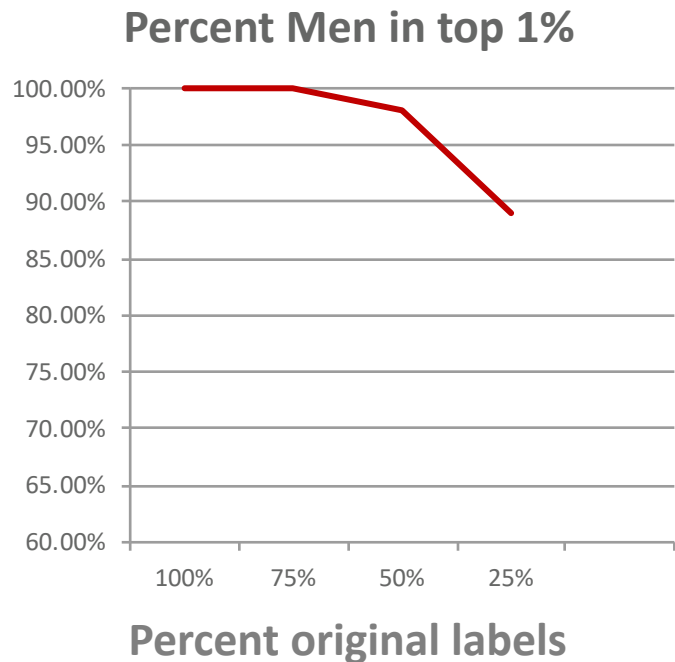
```
0000d41ed774823fca142945ec915c0,1,,,,,,,,,en_GB,,  
0000dee02d70cf8c0d79f96b6d1c59d,0,,,,,,,,,en_US,,  
0000f232abfe25a80156fe069395460,0,1992,20,2,2,,,,,,,,,19,-5  
0000f4ba0cff946b1c0e3b051287ede,0,1993,19,2,,,,,,,,,en_US,310,8  
0000130571654e3afaa62f4e9d2e4f63,0,,2,2,,,,,,,,,en_US,193,7  
00001544469ae9b408869a463a1dd77a,1,1984,28,2,,,,,,,,,en_US,,-4
```



```
0000d41ed774823fca142945ec915c0,0,,,,,,,,,en_GB,,  
0000dee02d70cf8c0d79f96b6d1c59d,0,,,,,,,,,en_US,,  
0000f232abfe25a80156fe069395460,1,1992,20,2,2,,,,,,,,,19,-5  
0000f4ba0cff946b1c0e3b051287ede,0,1993,19,2,,,,,,,,,en_US,310,8  
0000130571654e3afaa62f4e9d2e4f63,1,,2,2,,,,,,,,,en_US,193,7  
00001544469ae9b408869a463a1dd77a,1,1984,28,2,,,,,,,,,en_US,,-4
```

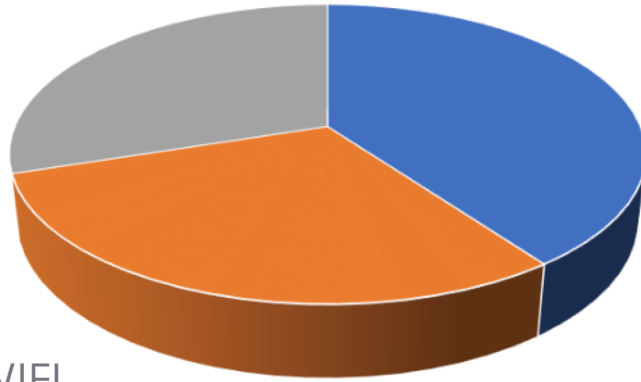


Performance is fairly stable even under significant noise



Predict who go to dealerships

People who are
somewhere close



Hacked WIFI

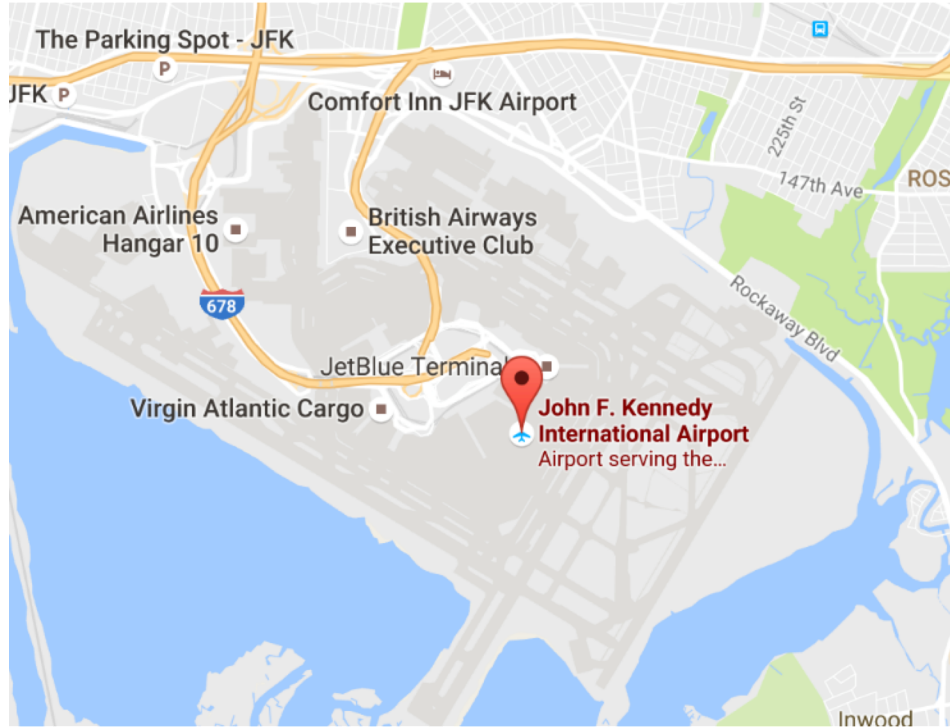
**People who
want a car ...**



Where do we find frequent travelers?



What do you think indicates people going to JFK?



URL

www.iglesiaelfaroinc.org
www.jumpseatnews.com
www.bluelineprop.com
www.ktxdtv.com
www.southjefffootball.org
www.unitedafa.org
www.parliamenthouse.com
www.yunghova.com
www.interlinetravel.com
www.aclin.org
www.swissport.com
www.gcsanc.com
www.swacu.org
www.airlinepilotcentral.com
www.homotrophy.com
www.beggfuneralhome.net
www.tvathletics.org
www.2shopper.com
www.nextmagazine.com
www.dailyjocks.com
www.pullzone.com
www.diamondoffshore.com
www.myerspolaris.com
www.ryandeyer.com
www.okllo.com
www.ifihadtochoose.com
www.ivoirmixdj.com

Logistic Parameter

2.38
2.25
2.21
2.14
2.1
2.09
2.07
2.06
2.03
2.03
2.03
2.01
2.01
1.97
1.97
1.94
1.92
1.91
1.91
1.87
1.87
1.86
1.86
1.86
1.84
1.83
1.83



URL	Logistic Parameter
www.iglesiaelfaroinc.org	2.38
www.jumpseatnews.com	2.25
www.bluelineprop.com	2.21
www.ktxdtv.com	2.14
www.southjefffootball.org	2.1
www.unitedafa.org	2.09
www.parliamenthouse.com	2.07
www.yunghova.com	2.06
www.interlinetravel.com	2.03
www.aclin.org	2.03
www.swissport.com	2.03
www.gcsanc.com	2.01
www.swacu.org	2.01
www.airlinepilotcentral.com	1.97
www.homotrophy.com	1.97
www.beggfuneralhome.net	1.94
www.tvathletics.org	1.92
www.2shopper.com	1.91
www.nextmagazine.com	1.91
www.dailyjocks.com	1.87
www.pullzone.com	1.87
www.diamondoffshore.com	1.86
www.myerspolaris.com	1.86
www.ryandeyer.com	1.86
www.okllo.com	1.84
www.ifihadtochoose.com	1.83
www.ivoirmixdj.com	1.83



URL	Logistic Parameter
www.iglesiaelfaroinc.org	2.38
www.jumpseatnews.com	2.25
www.bluelineprop.com	2.21
www.ktxdtv.com	2.14
www.southjefffootball.org	2.1
www.unitedafa.org	2.09
www.parliamenthouse.com	2.07
www.yunghova.com	2.06
www.interlinetravel.com	2.03
www.aclin.org	2.03
www.swissport.com	2.03
www.gcsanc.com	2.01
www.swacu.org	2.01
www.airlinepilotcentral.com	1.97
www.homotrophy.com	1.97
www.beggfuneralhome.net	1.94
www.tvathletics.org	1.92
www.2shopper.com	1.91
www.nextmagazine.com	1.91
www.dailyjocks.com	1.87
www.pullzone.com	1.87
www.diamondoffshore.com	1.86
www.myerspolaris.com	1.86
www.ryandeyer.com	1.86
www.okllo.com	1.84
www.ifihadtochoose.com	1.83
www.ivoirmixdj.com	1.83



URL	Logistic Parameter
www.iglesiaelfaroinc.org	2.38
www.jumpseatnews.com	2.25
www.bluelineprop.com	2.21
www.ktxdtv.com	2.14
www.southjefffootball.org	2.1
www.unitedafa.org	2.09
www.parliamenthouse.com	2.07
www.yunghova.com	2.06
www.interlinetravel.com	2.03
www.aclin.org	2.03
www.swissport.com	2.03
www.gcsanc.com	2.01
www.swacu.org	2.01
www.airlinepilotcentral.com	1.97
www.homotrophy.com	1.97
www.beggfuneralhome.net	1.94
www.tvathletics.org	1.92
www.2shopper.com	1.91
www.nextmagazine.com	1.91
www.dailyjocks.com	1.87
www.pullzone.com	1.87
www.diamondoffshore.com	1.86
www.myerspolaris.com	1.86
www.ryandeyer.com	1.86
www.okllo.com	1.84
www.ifihadtochoose.com	1.83
www.ivoirmixdj.com	1.83

Predict who goes to JFK?

People who work there ...



An ornate, gold-colored picture frame with intricate scrollwork and floral patterns. The frame is rectangular and has a wide, decorative border. The center of the frame is a solid black rectangle.

Big Picture

Beware of (unintentional) discrimination based on predictability



LinkedIn Jobs: Targeted Email Delivery is Here

 Parker Barrile February 28, 2011



LinkedIn just rolled out an exciting new feature that makes it even easier to reach qualified candidates automatically. Members who view jobs on LinkedIn will now receive targeted job recommendations automatically via email. Powered by LinkedIn's 'Jobs You May Be Interested In' system, the emails include jobs that are personalized for each recipient based on his or her professional profile.

Ad related to latanya sweeney ⓘ

[Latanya Sweeney Truth](#)
www.instantcheckmate.com/

Looking for **Latanya Sweeney**? Check **Latanya Sweeney's Arrests**.

Ads by Google

[Latanya Sweeney, Arrested?](#)

1) Enter Name and State. 2) Access Full Background Checks Instantly.
www.instantcheckmate.com/

[Latanya Sweeney](#)

Public Records Found For: **Latanya Sweeney**. View Now.
www.publicrecords.com/

[La Tanya](#)

Search for La Tanya Look Up Fast Results now!
www.ask.com/La+Tanya

Thank You!

Claudia Perlich

Senior Data Scientist
@claudia_Perlich

www.twosigma.com/careers