



dstillery

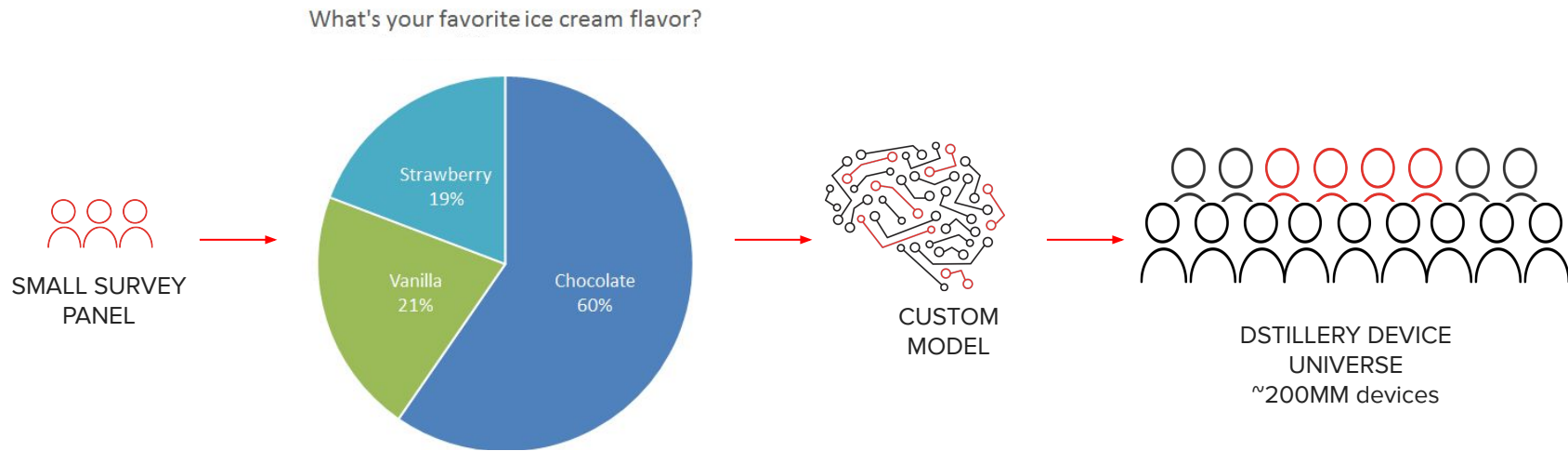
# ONLINE LEARNING OF WEBSITE EMBEDDINGS

for Accurate Prediction of User  
Behavior

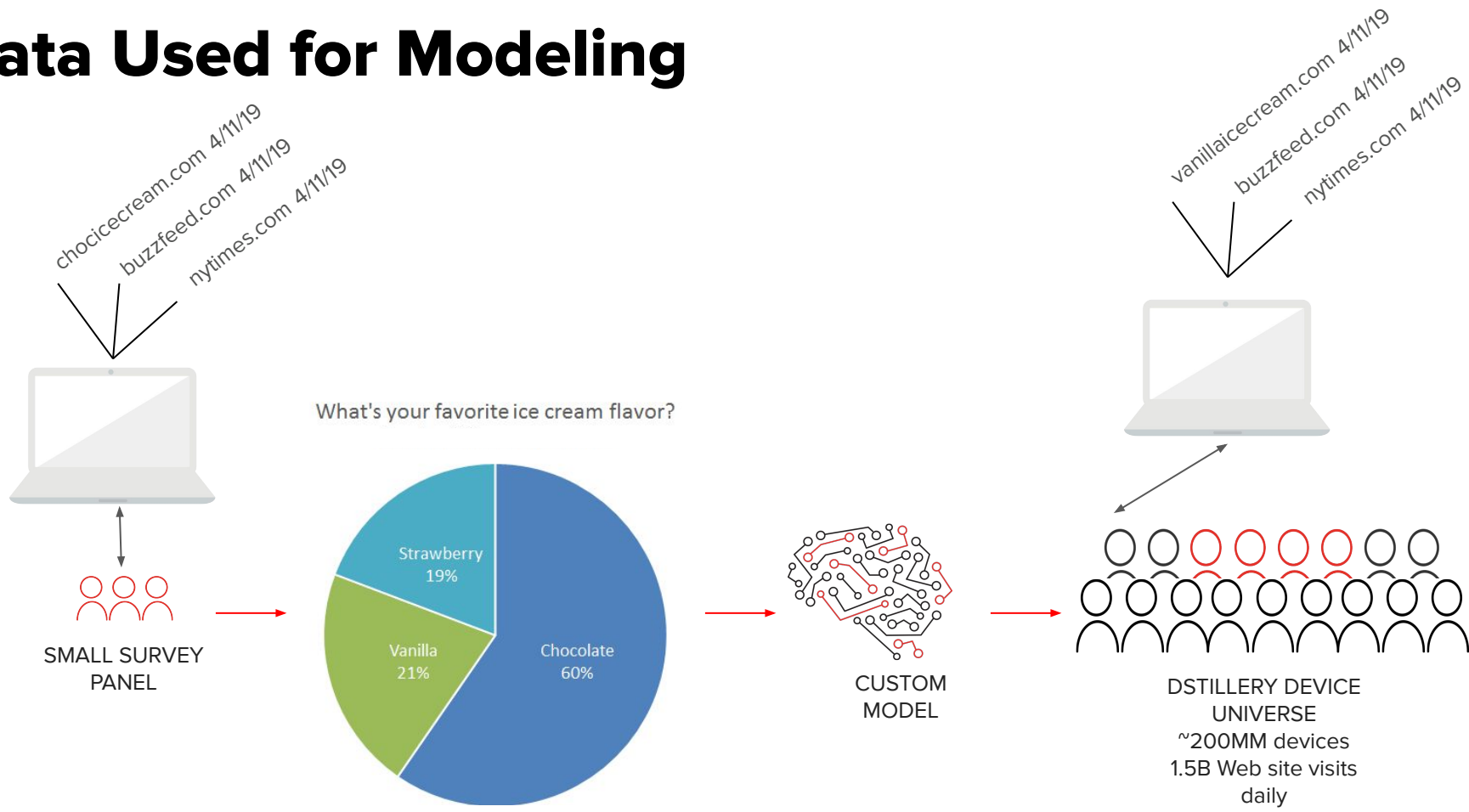
**Even when Data are Scarce**

Amelia White, Director of Data Science Research  
Nov 13, 2019

# Expanding Digital Survey Data



# Data Used for Modeling



[illegible]

# Need for a Reduced Dimensional Feature Space

[illegible]

**REDUCED  
DIMENSIONAL  
FEATURE SPACE**

# Taking Ideas from Natural Language Processing

- Similar data
- Sentences of words

Up to the 1980s, most natural language processing systems were based on complex sets of hand-crafted rules for processing with the introduction of **machine learning** algorithms for language processing. This lessening of the dominance of **Chomskyan** theories of linguistics (e.g. **transformational grammar**)

- Sequences of web sites visited

www.eatright.org www.therabreath.com www.colgate.com www.naturallyella.com  
picysouthernkitchen.com www.modwedding.com www.greenlakejewelry.com www.ingaddiction.com  
www.dietitian.com www.vt.edu www.gimmesomeoven.com www.

- High dimensional categorical features

# Need for a Reduced Dimensional Feature Space

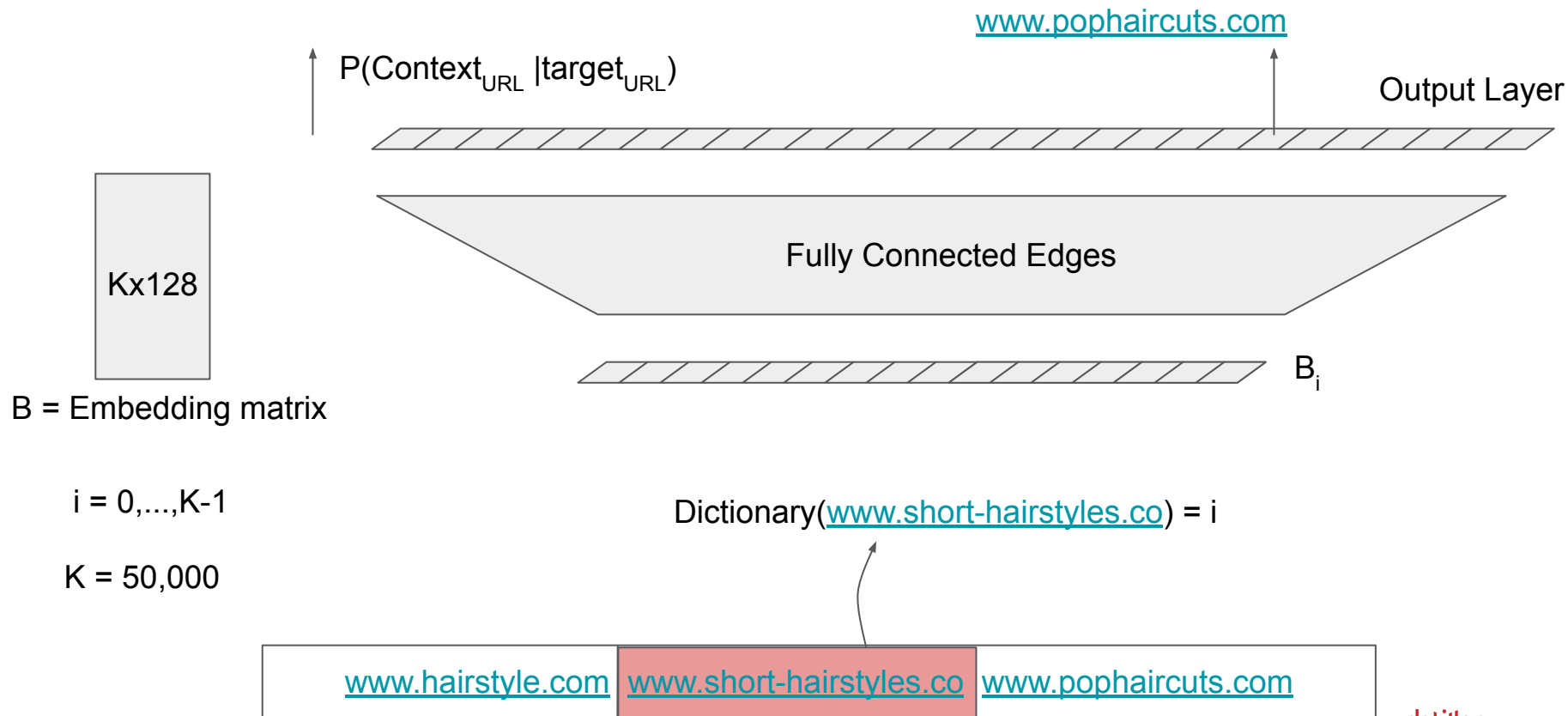
[illegible]

# Need for a Reduced Dimensional Feature Space

[illegible]

Dim 1	Dim 2	...	...	Dim 128
0.67534775	0.43516183	0.00371811	0.58666643	0.13523544
0.10092871	0.02024611	0.19984572	0.64574799	0.54387123
0.98997418	0.26743623	0.34408102	0.71425389	0.02140623
0.0845312	0.53884732	0.37128581	0.65845385	0.35794342
0.6666365	0.53685554	0.34811952	0.28316887	0.48017634
0.79902046	0.9786974	0.14587728	0.59378527	0.53994022
0.71107443	0.16079175	0.78204965	0.45080368	0.29320381
0.64726402	0.09479171	0.09246093	0.87526155	0.51668014

# Website Embeddings V1: word2vec

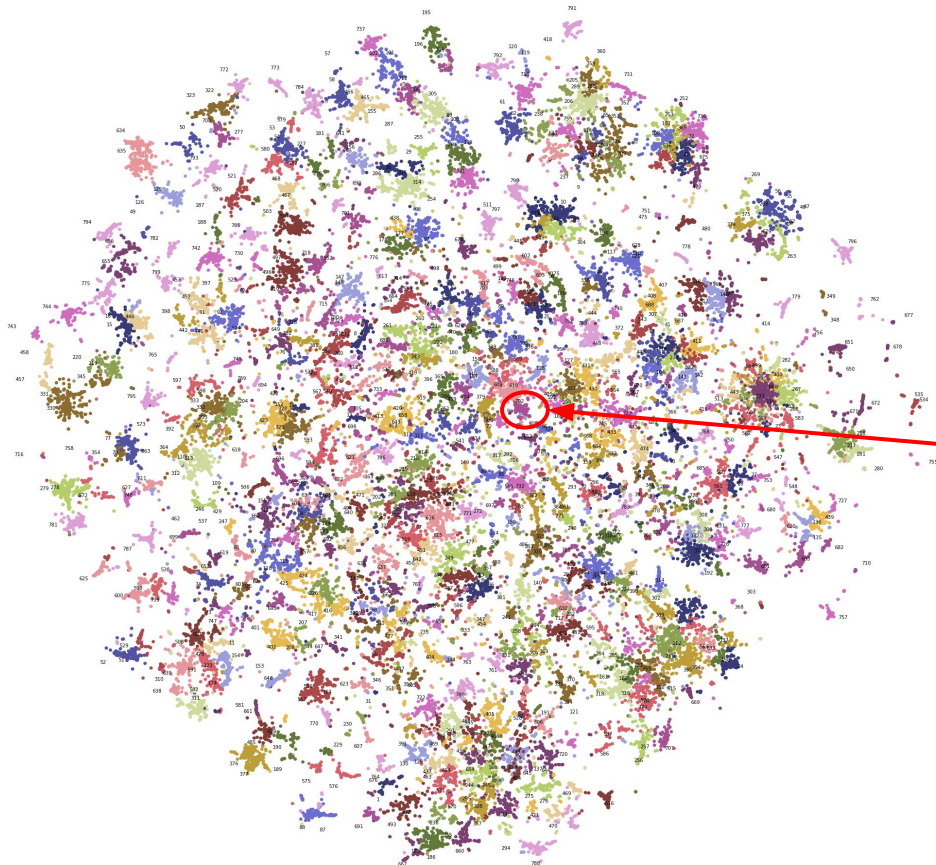


# Training Word2vec

- Trained word2vec with the browsing history of all devices seen in a 2 week time period:
- Browsing history of 430,648,822 devices
- Sequence of 15,077,897,800 site visits

www.blackhairinformation.com www.petcarerx.com www.news-medical.net www.pocket-lint.com www.davidlebovitz.com www.huffpost.com  
www.lassy.com www.colgate.com www.pophaircuts.com www.recapo.com www.skinnymom.com www.att.net www.utahvalley.com www.firstrespon  
olgate.com www.cookingclassy.com www.mobilityawarenessmonth.com www.gurl.com www.att.net www.cookingclassy.com www.inventorspa  
ressywomen.com www.sallysbakingaddiction.com www.rmhealthy.com www.fitbottomedgirls.com www.recaplet.com www.appliancesconnec  
t.all-greatquotes.com www.medtronicdiabetes.com www.recaplet.com www.julianbakery.com www.spendwithpennies.com www.att.net www  
www.jamiecooksitup.net www.legacy.com www.fitbottomedgirls.com www.elvisyorkshireterrier.com www.visual.ly www.bottomlinehealth  
b-hairstyle.com www.short-hairstyles.co www.pophaircuts.com www.legacy.com www.thehealthsite.com www.carmonkeys.com www.bottor  
com www.thediabetessite.com www.good4utah.com www.usana.com www.topsecretrecipes.com www.att.net www.homeremedyshop.com www.lu  
www.petmd.com www.whfoods.com www.ohsweetbasil.com www.healthyandnaturalworld.com www.ohsweetbasil.com break  
estademexico.com www.100x100fan.mx www.jornadameridiana.com www.eonline.com www.tvnotas.com.mx abcnews.go.com www.record.com  
tas.com.mx www.denunciasmx.com www.tvnotas.com.mx www.eluniversal.com www.entretengo.com www.cezy.org www.yucatan.com.mx www.l  
co.com www.tvnotas.com.mx www.classifiedads.com www.starmedios.com www.record.com.mx www.eldeforma.com www.grupopalomooficial  
eforma.com www.quieroavisos.com www.tvnotas.com.mx www.classifiedads.com www.tvnotas.com.mx www.eluniversal.com.mx www.eonline  
notas.com.mx www.launion.com.mx www.tvnotas.com.mx www.eldiariony.com www.prensa.com www.elgrafico.mx www.infobae.com www.seci  
.com.mx www.eluniversal.com.mx www.tvnotas.com.mx www.buenamusic.com www.tvnotas.com.mx www.eldiariony.com break  
m.com www.themagicalslowcooker.com www.minq.com break

# Visualizing Embeddings



Website	Cluster #
www.boardingarea.com	512
www.thepointsguy.com	512
www.taxifarefinder.com	512
www.theflightdeal.com	512
www.uberestimate.com	512
www.sleepinginairports.net	512
www.frugaltravelguy.com	512
www.airchina.us	512
www.cathaypacific.com	512
www.travelskills.com	512
www.travelsort.com	512
www.skyteam.com	512
www.seatmaestro.com	512
www.flyertalk.com	512
www.expertflyer.com	512
www.singaporeair.com	512
www.estimatefares.com	512

# BEYOND WORD2VEC:

- Embedding millions of URLs, with a manageable number of parameters
- Online learning of embeddings

**EMBEDDING  
MORE URLS WITH  
FEWER  
PARAMETERS**

# Hash Embeddings

## Hash Embeddings for Efficient Word Representations

**Dan Svenstrup**

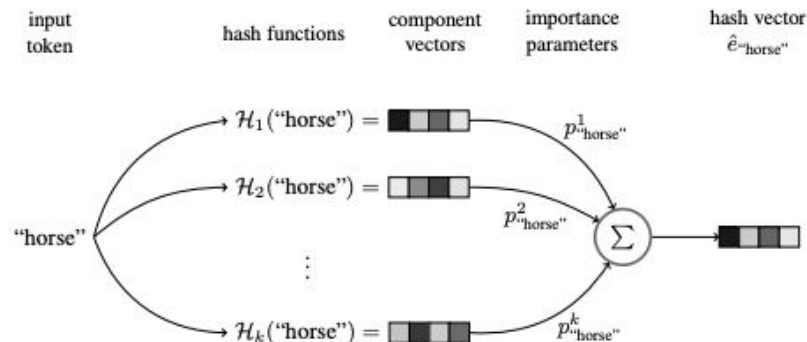
Department for Applied Mathematics and Computer Science  
Technical University of Denmark (DTU)  
2800 Lyngby, Denmark  
dsve@dtu.dk

**Jonas Meinertz Hansen**

FindZebra  
Copenhagen, Denmark  
jonas@findzebra.com

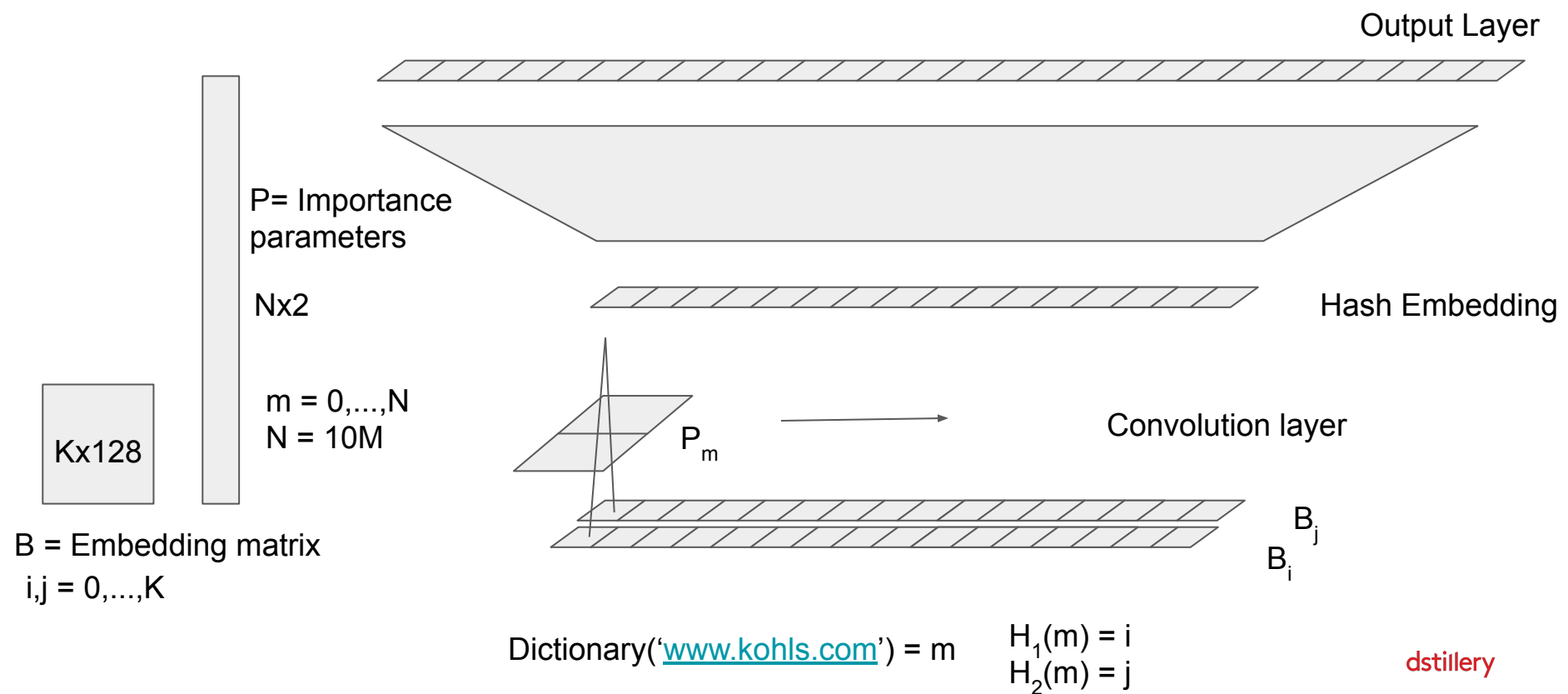
**Ole Winther**

Department for Applied Mathematics and Computer Science  
Technical University of Denmark (DTU)  
2800 Lyngby, Denmark  
olwi@dtu.dk

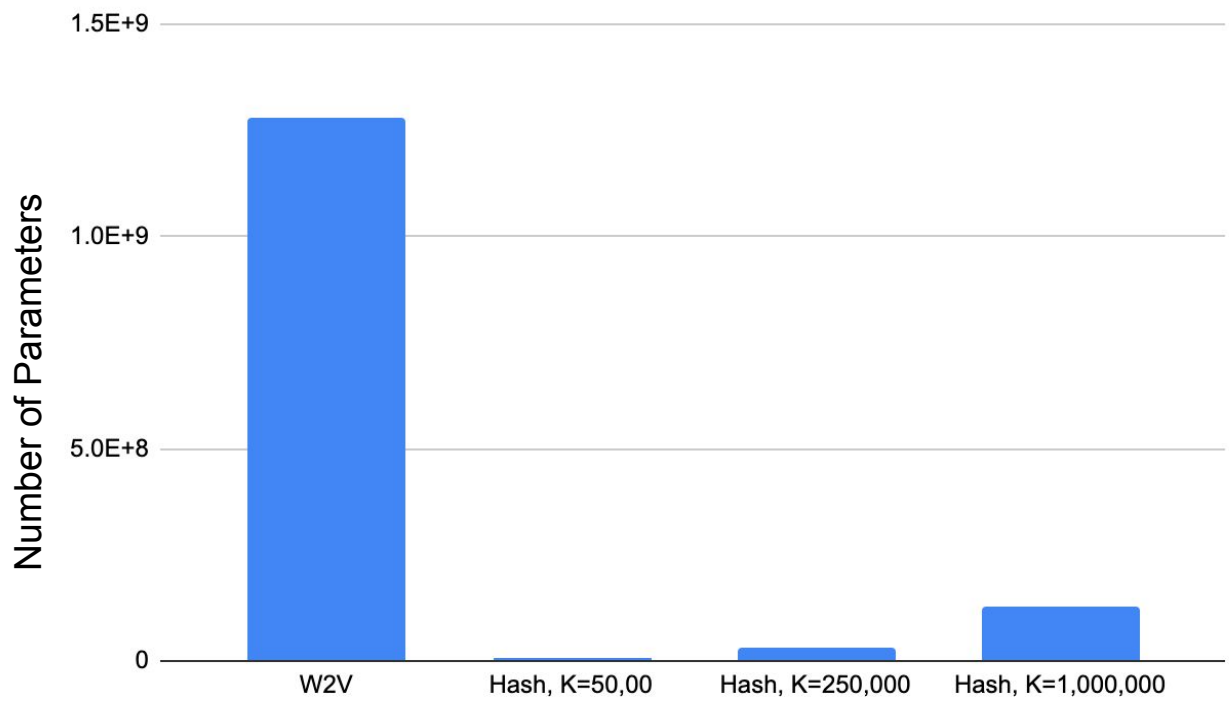


31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

# Website Embeddings V2: Hash embeddings

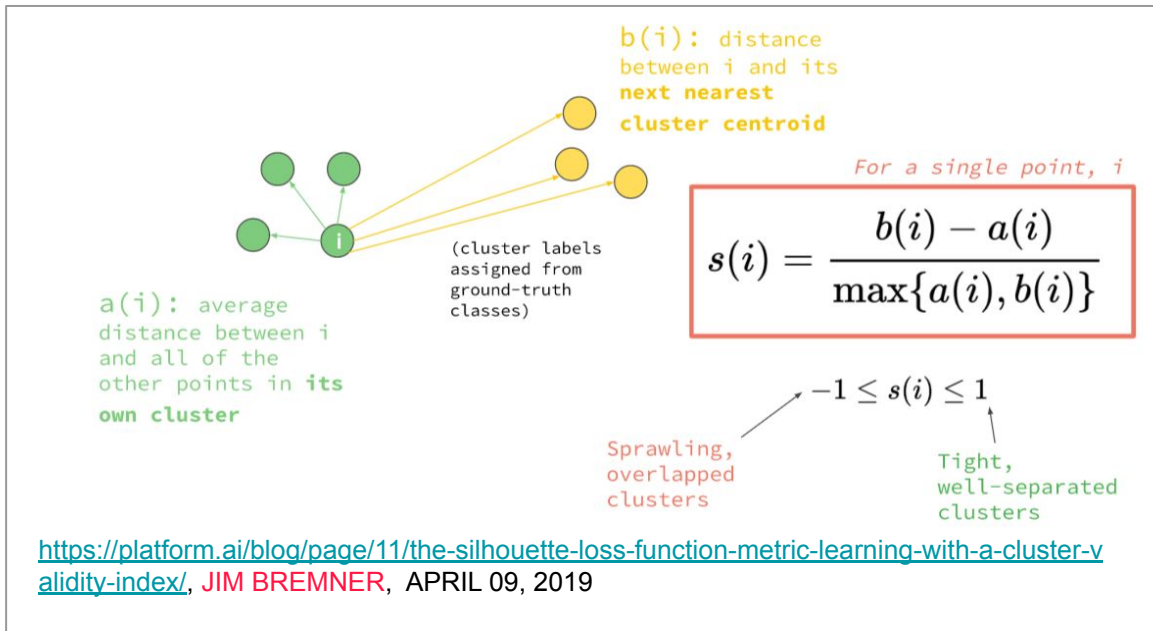


# Hash Embedding Requires Fewer Parameters

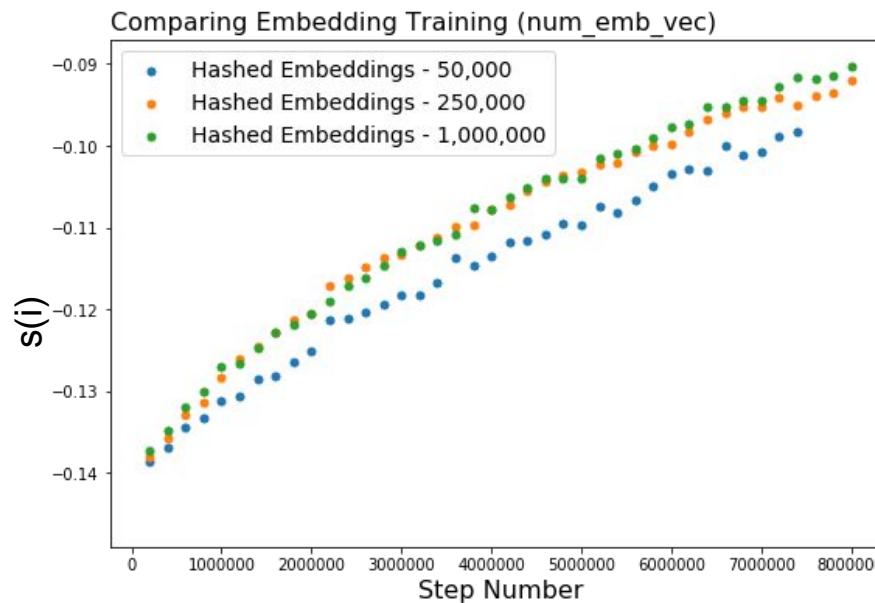
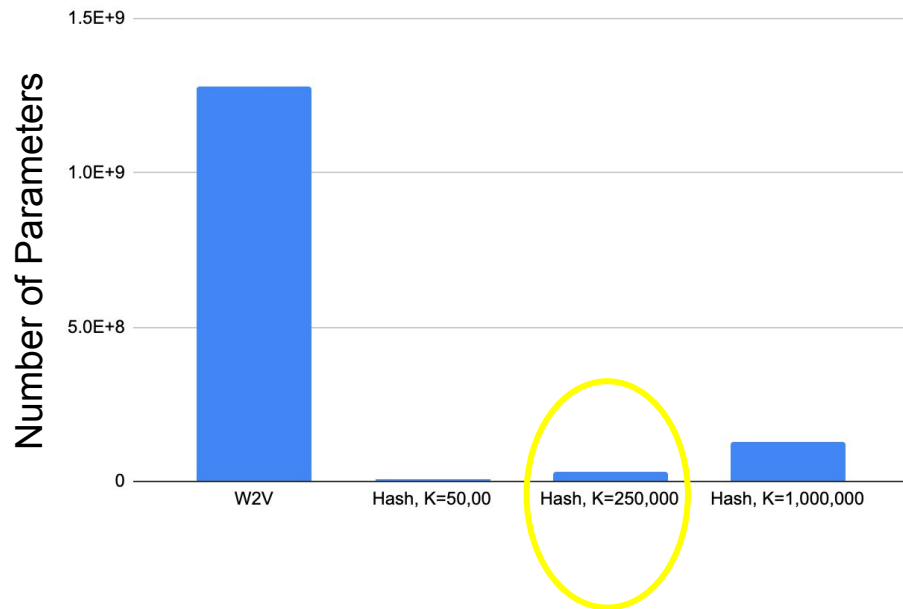


# Measuring Embedding Quality for Parameter Selection

- Selected a 'ground truth' clustering, made from a known high quality embedding
- Used the silhouette score to measure how well test embeddings converged to the ground truth clustering as the network trained

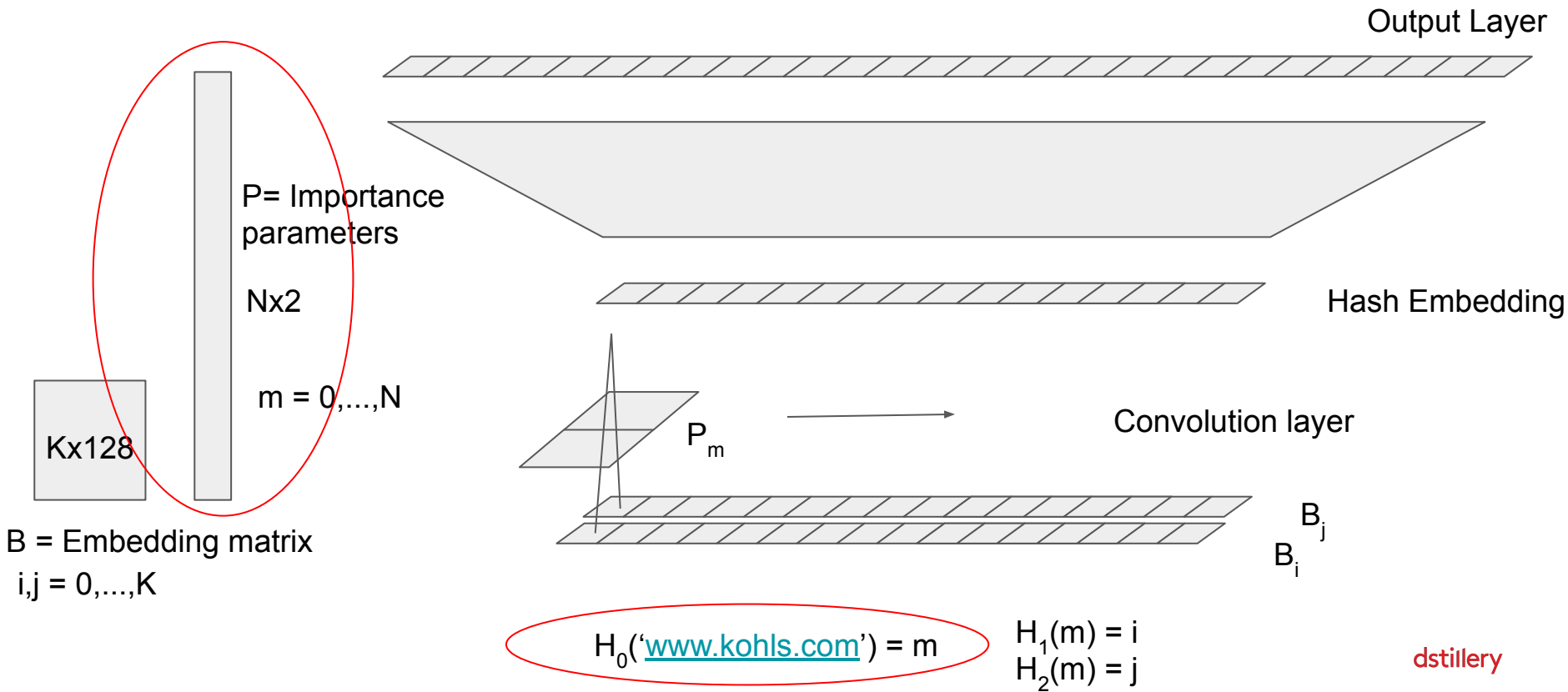


# Good Performance with 100x Fewer Parameters

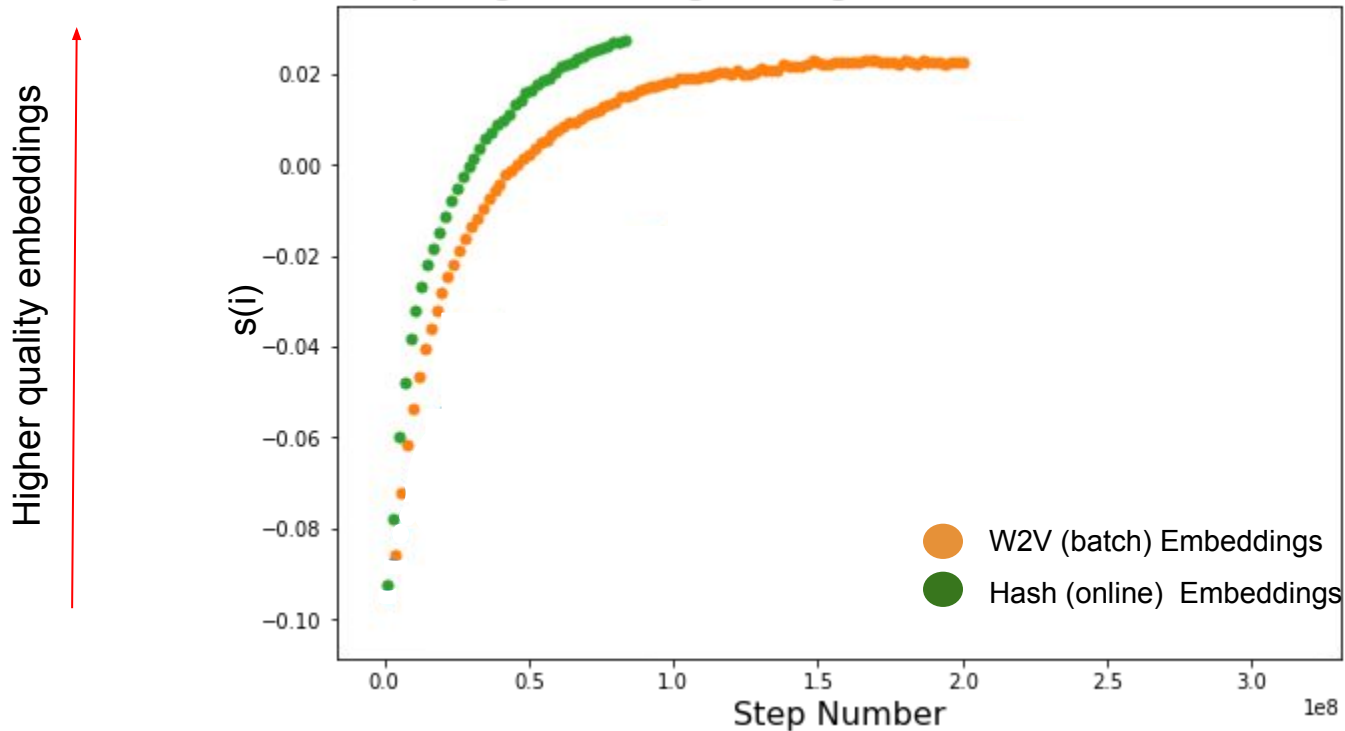


# **ONLINE LEARNING OF EMBEDDINGS**

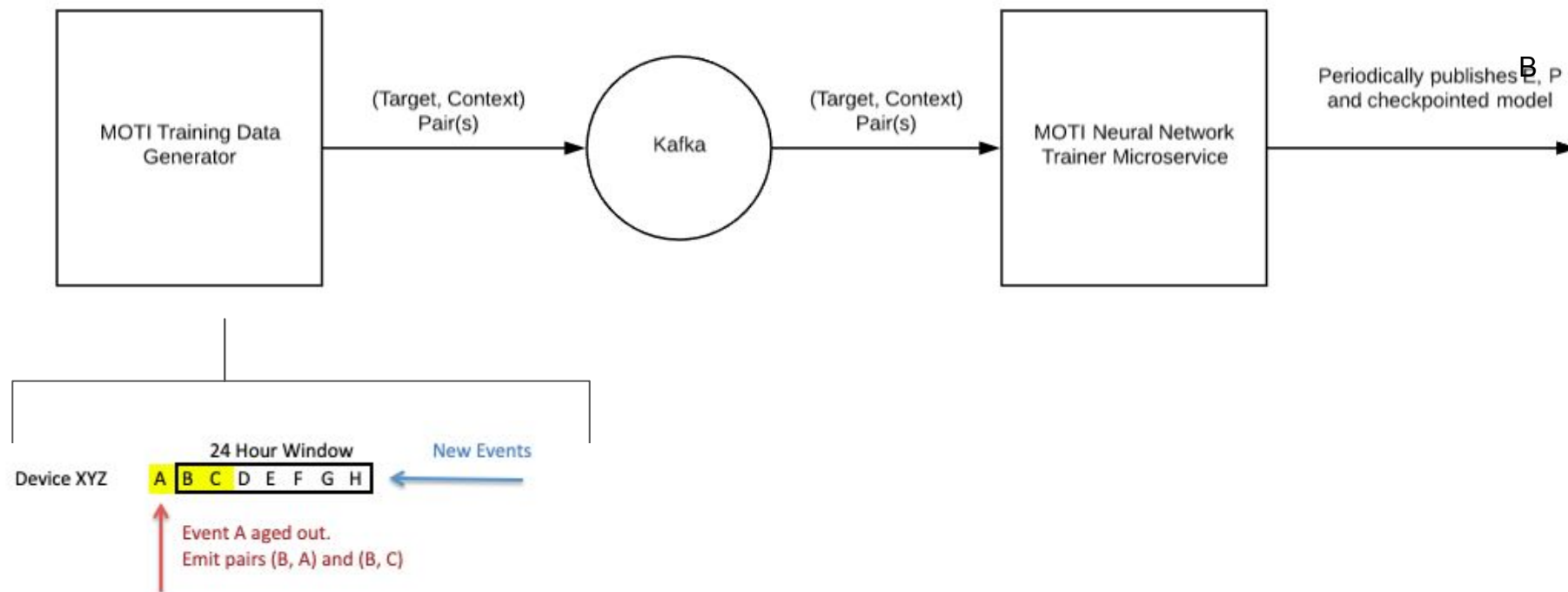
# Website Embeddings V3: Online Learning of Hash Embeddings



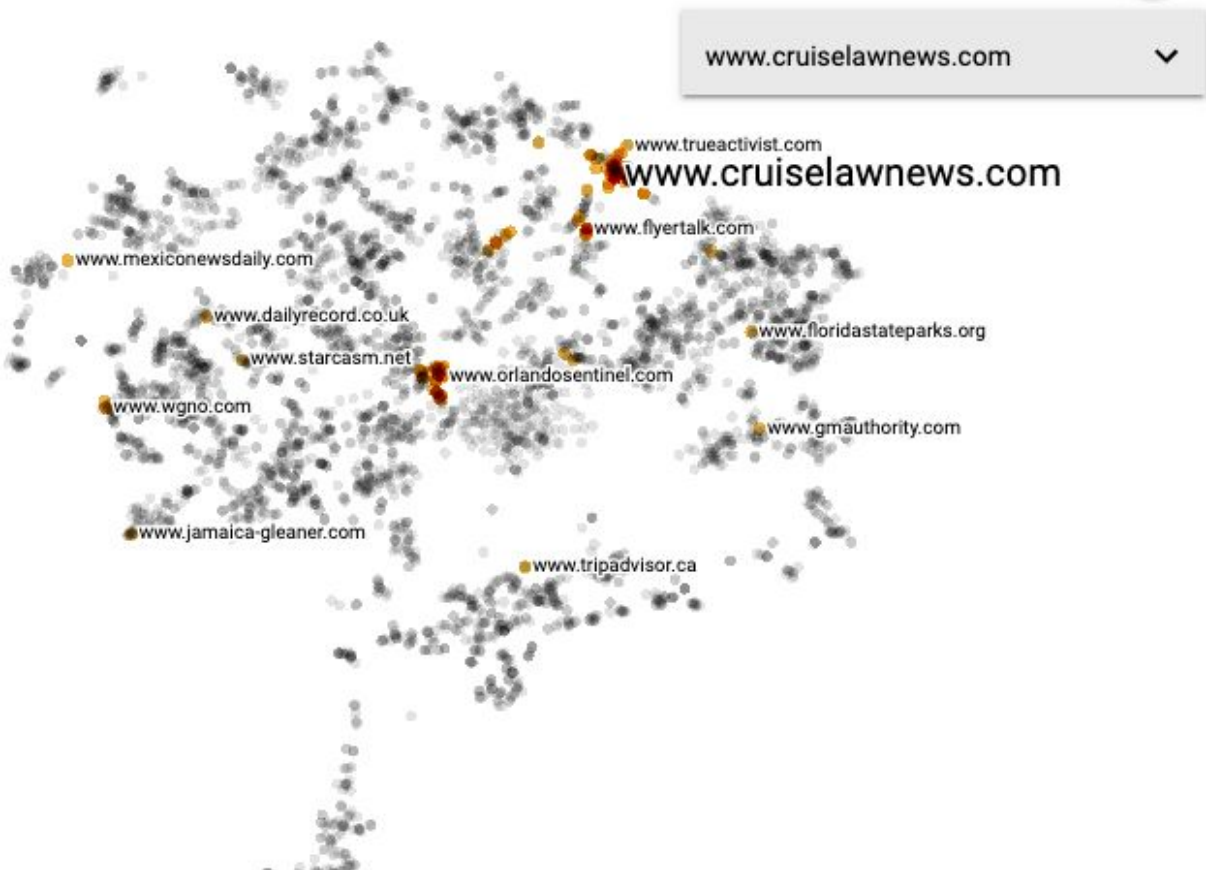
# Online Learning Optimizes Faster than Batch Learning



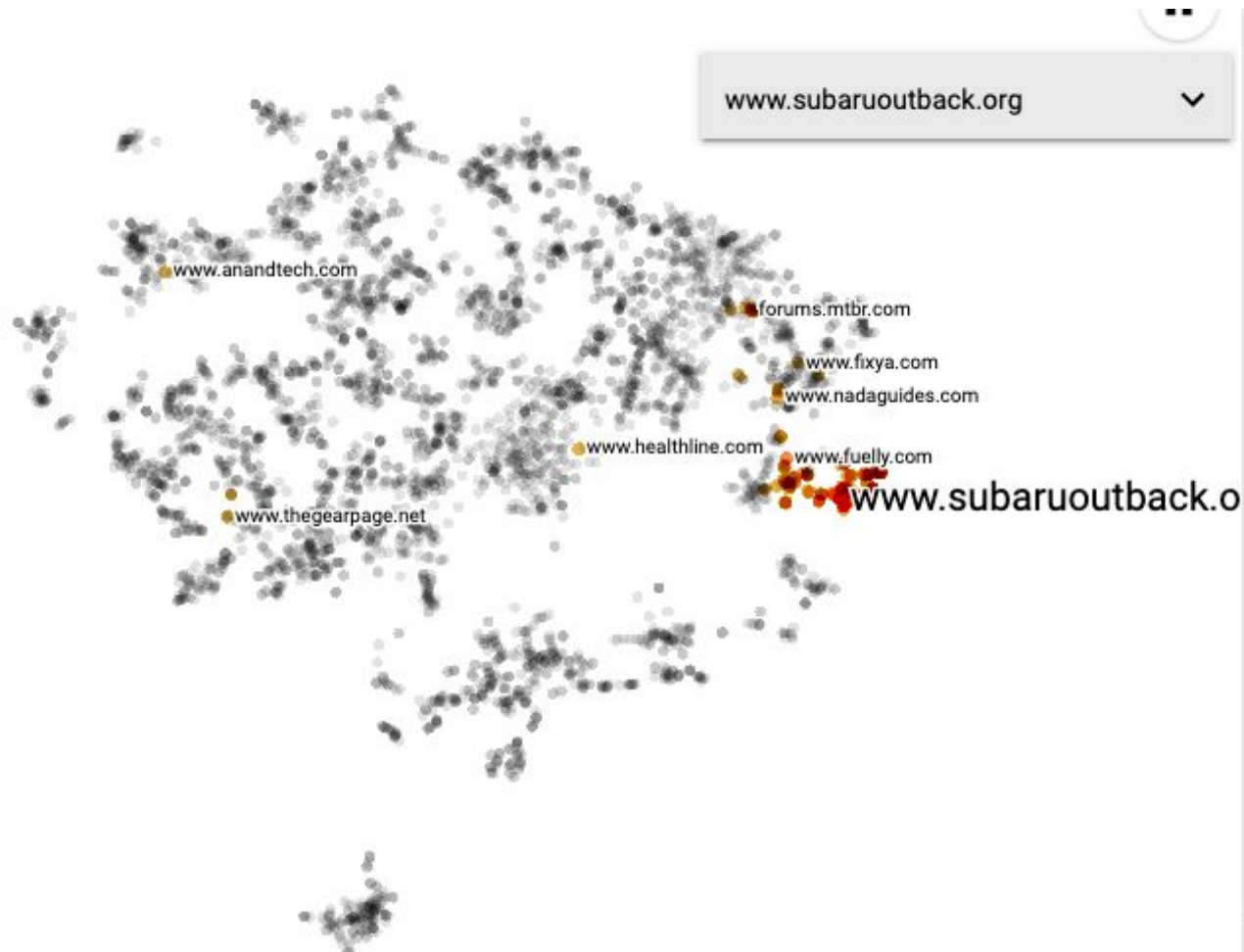
# Training the Online Embeddings



# Distance in Embedding Space



Search	.*	by	
neighbors	?		100
distance	COSINE	EUCLIDEAN	
Nearest points in the original space:			
www.royalcaribbeanblog.com			0.248
www.cruisecritic.com			0.292
www.cruisehive.com			0.333
www.ncl.com			0.391
www.travelpulse.com			0.426
www.caribjournal.com			0.445
www.carnival.com			0.482
www.destinationtips.com			0.538
www.thetravel.com			0.546
www.flyertalk.com			0.548



Search

.\*

by

neighbors

?



100

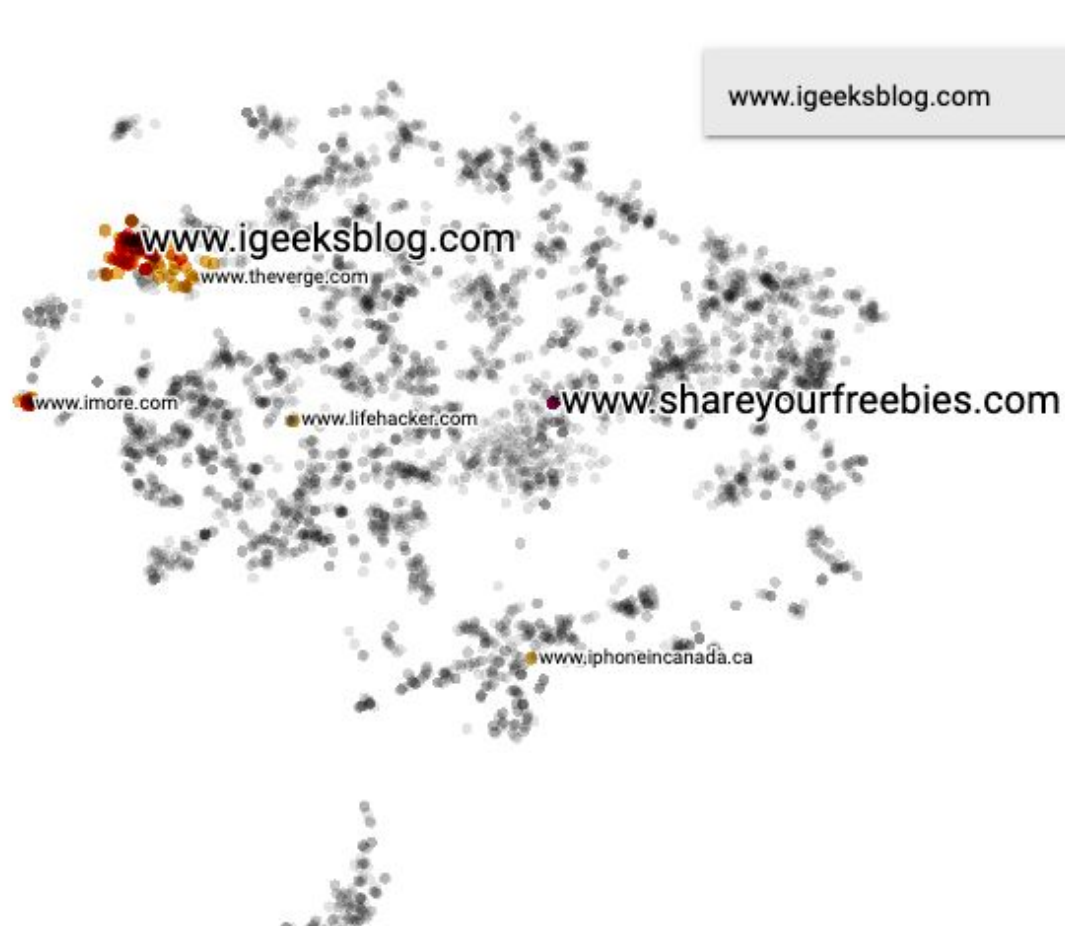
distance

COSINE

EUCLIDEAN

Nearest points in the original space:

<a href="#">www.1010tires.com</a>	0.378
<a href="#">www.toyotanation.com</a>	0.385
<a href="#">www.michelinman.com</a>	0.391
<a href="#">www.acurazine.com</a>	0.404
<a href="#">www.carcomplaints.com</a>	0.407
<a href="#">www.toyota-4runner.org</a>	0.426
<a href="#">www.700r4transmissionhq.com</a>	0.438
<a href="#">www.advanceautoparts.com</a>	0.446
<a href="#">www.ih8mud.com</a>	0.447
<a href="#">www.autoanything.com</a>	0.450



Search	.*	by	
neighbors	<input type="range" value="100"/>		100
distance	COSINE	EUCLIDEAN	
Nearest points in the original space:			
www.appletoolbox.com			0.103
www.idownloadblog.com			0.115
www.iphonelife.com			0.122
www.macworld.com			0.167
www.macworld.co.uk			0.191
www.cultofmac.com			0.200
www.gadgethacks.com			0.205
www.ikream.com			0.241
www.gottabemobile.com			0.243
www.techzilla.com			0.293

www.hypebeast.com



Search



by



neighbors



100

distance

COSINE

EUCLIDEAN

Nearest points in the original space:

<a href="#">www.solecollector.com</a>	0.186
<a href="#">www.sneakernews.com</a>	0.229
<a href="#">www.nicekicks.com</a>	0.236
<a href="#">www.kicksonfire.com</a>	0.272
<a href="#">www.complex.com</a>	0.359
<a href="#">www.hotnewhiphop.com</a>	0.370
<a href="#">www.thefader.com</a>	0.424
<a href="#">www.sneakerbardetroit.com</a>	0.471
<a href="#">www.vogue.com</a>	0.473
<a href="#">www.aa.com</a>	0.484

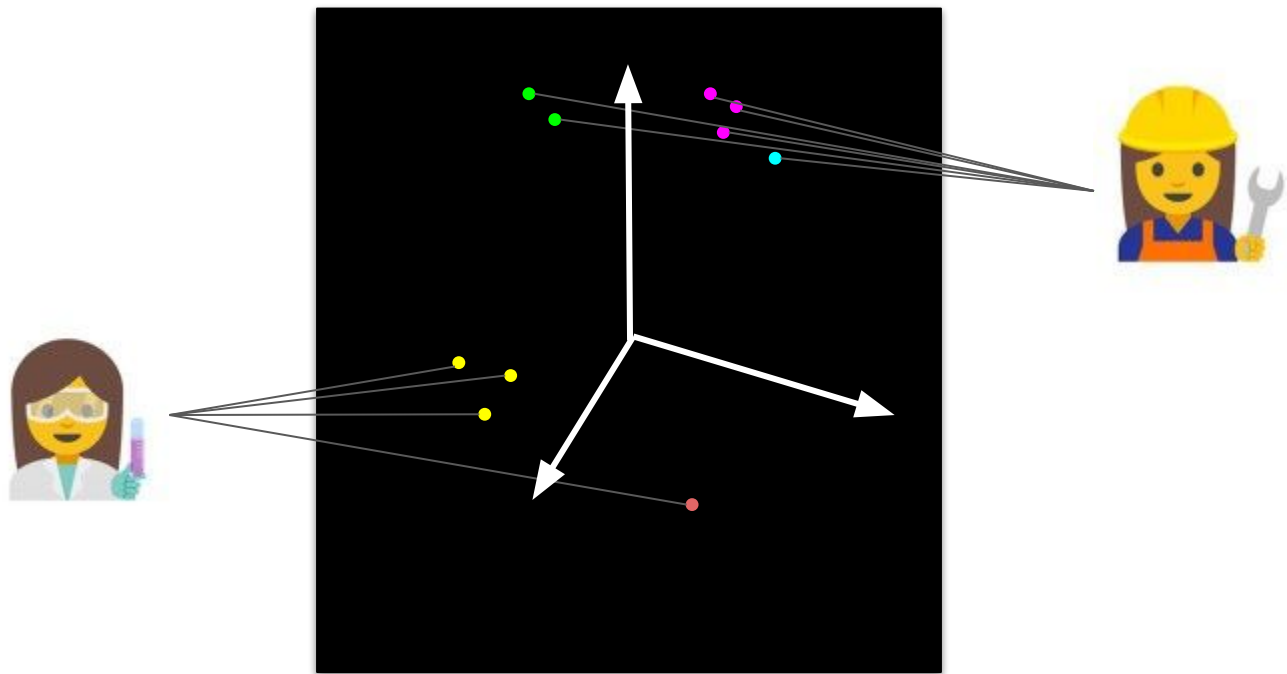
# **MODELING USERS IN EMBEDDING SPACE**

# Need for a Reduced Dimensional Feature Space

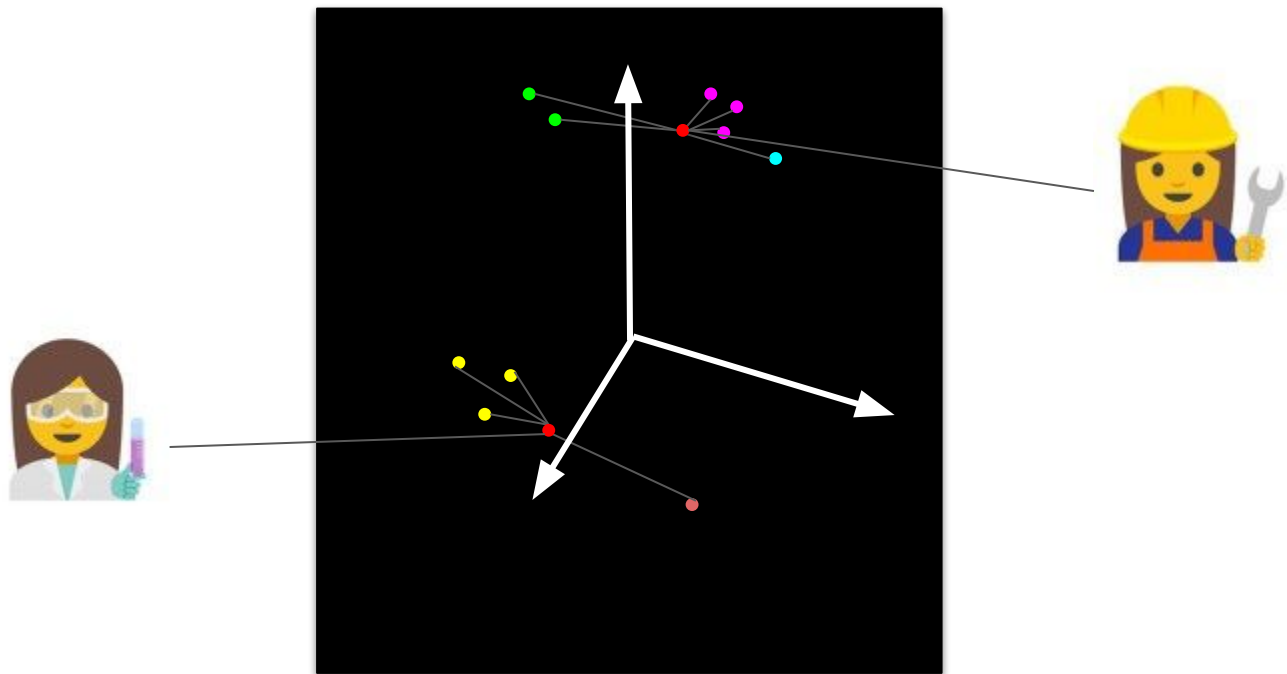
[illegible]

Dim 1	Dim 2	...	...	Dim 128
0.67534775	0.43516183	0.00371811	0.58666643	0.13523544
0.10092871	0.02024611	0.19984572	0.64574799	0.54387123
0.98997418	0.26743623	0.34408102	0.71425389	0.02140623
0.0845318	0.53884738	0.37188681	0.65845885	0.35794342
0.66666667	0.53884738	0.37188681	0.65845885	0.48017634
0.79902046	0.9786974	0.14587728	0.59378527	0.53994022
0.71107443	0.16079175	0.78204965	0.45080368	0.29320381
0.64726402	0.09479171	0.09246093	0.87526155	0.51668014

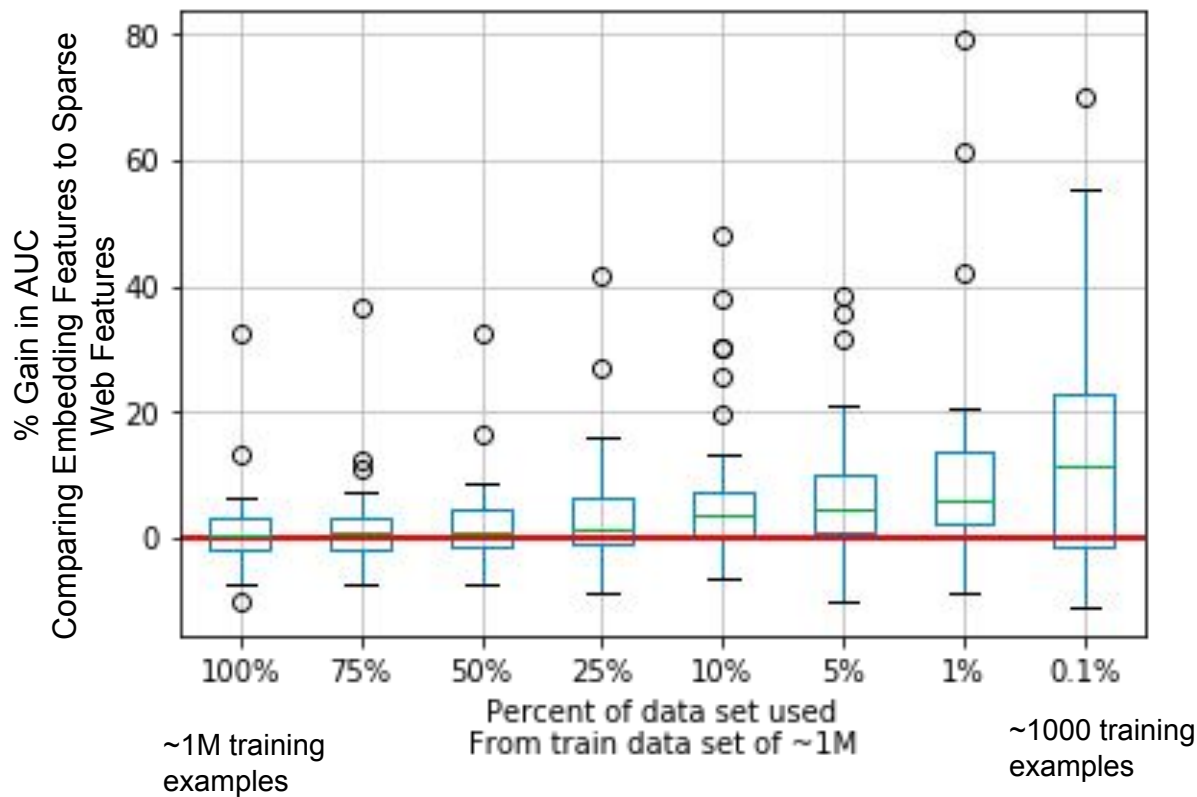
# From URL Embeddings to Models



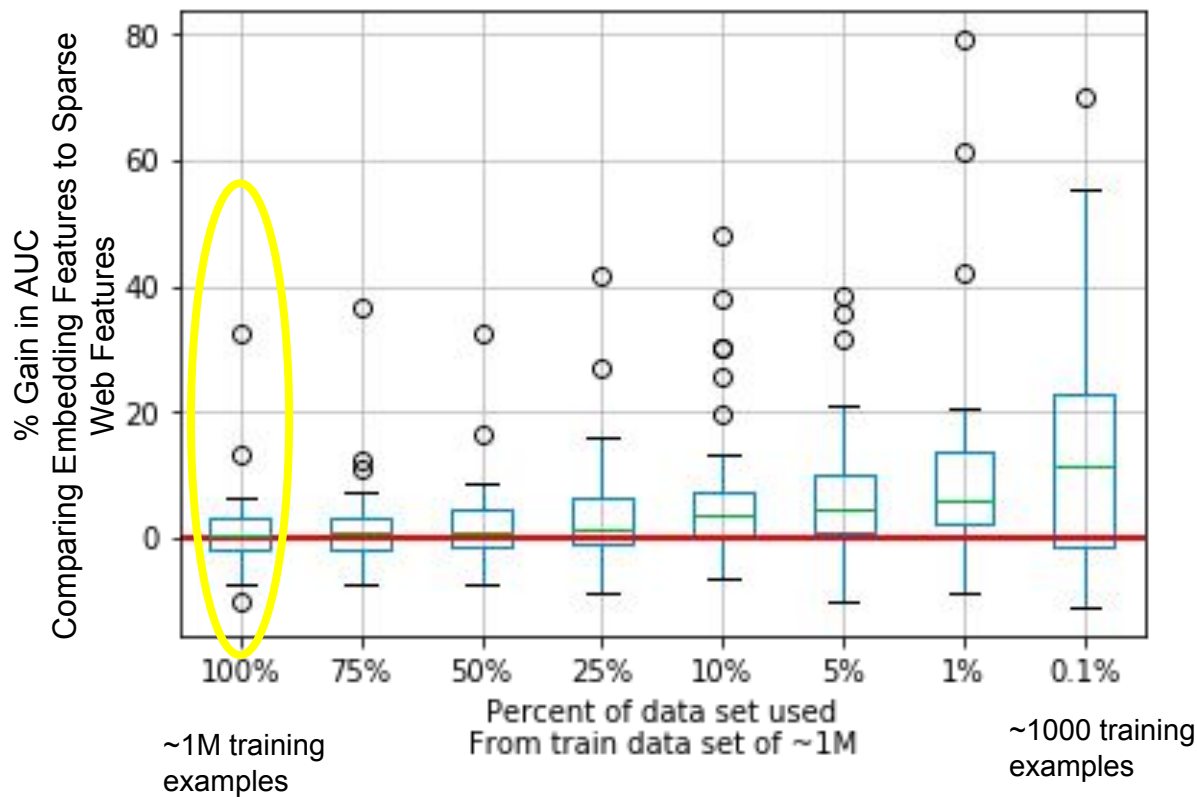
# From URL Embeddings to Models



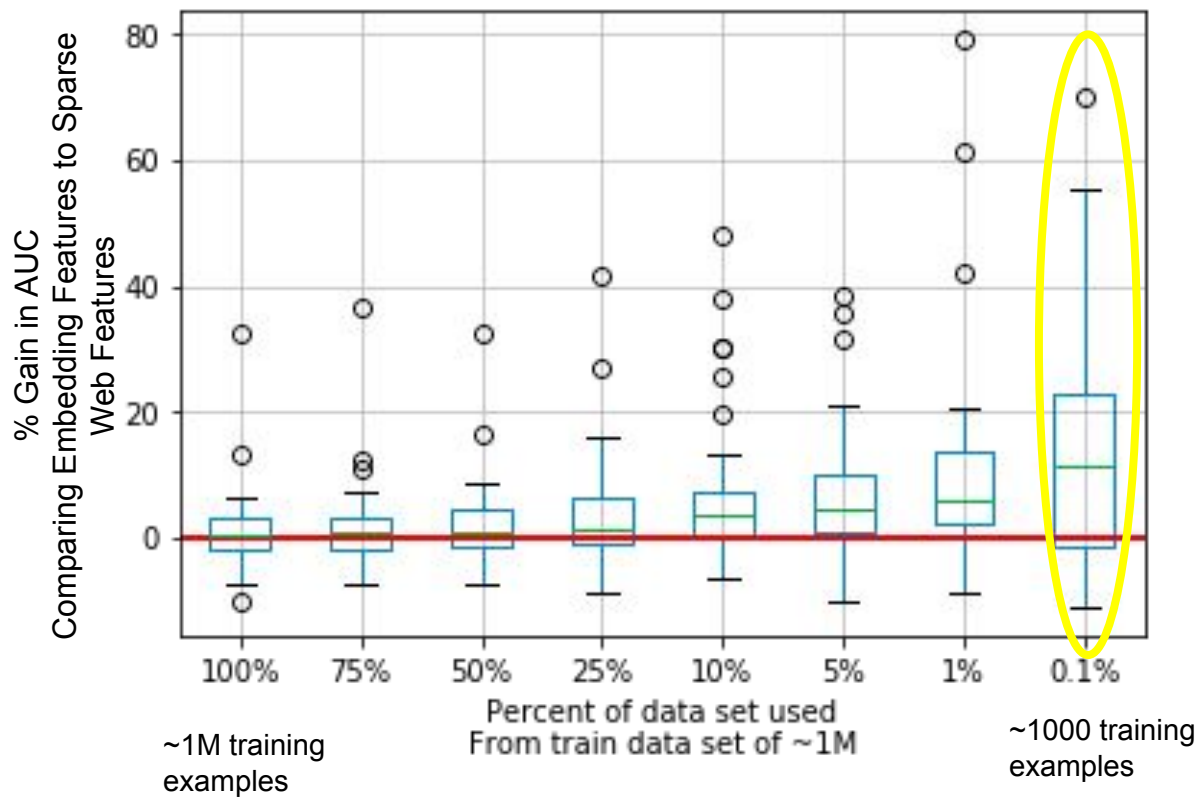
# Embedding Features Outperform Sparse Web Features For Small Data Sets



# Embedding Features Outperform Sparse Web Features For Small Data Sets

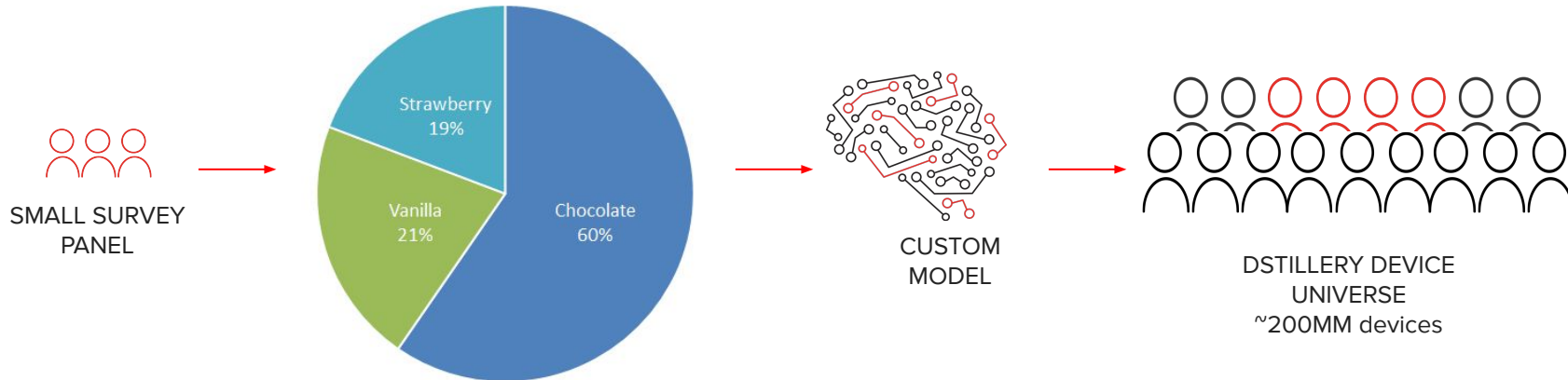


# Embedding Features Outperform Sparse Web Features For Small Data Sets



# MODELING SURVEY DATA

What's your favorite ice cream flavor?



# Case Study: Predicting Ad Influence for Ice Cream Brand

- The Problem:

- A survey company models which people are likely to be influenced by an advertisement for an ice cream brand
- 5.5K survey respondents
- 500 high scoring respondents

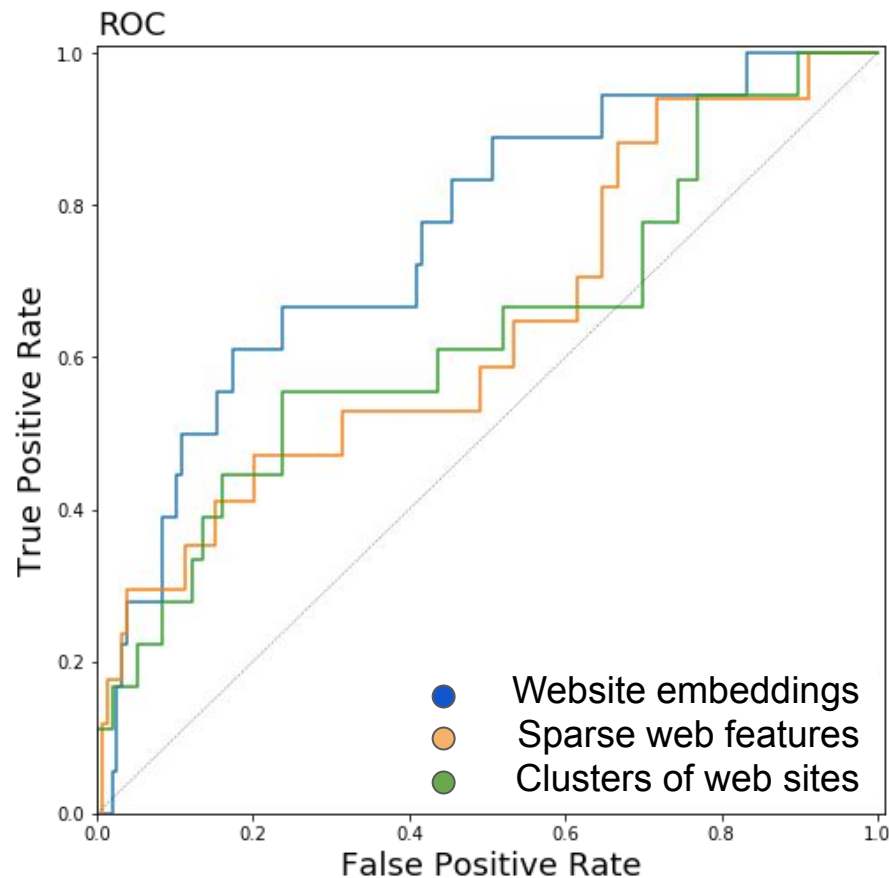
- Our Goal:

- Predicting the high scoring respondents
- Produce audience of devices that are predicted to be influenceable by ad for ice cream brand



# Case Study: Predicting Ad Influence for Ice Cream Brand

- Test AUC on predicting high scoring respondents:
  - Raw web behavior: 64.1
  - Summarized web behavior: 63.5
  - Cookie Embeddings: 75.8





dstillery

# THANK YOU

Presented by Amelia White.

[awhite@dstillery.com](mailto:awhite@dstillery.com)

Contributors:

**Christopher Jenness**

Melinda Han Williams

MLE team:

Wickus Martin

Roger Cost

Justin Moynihan

Patrick McCarthy