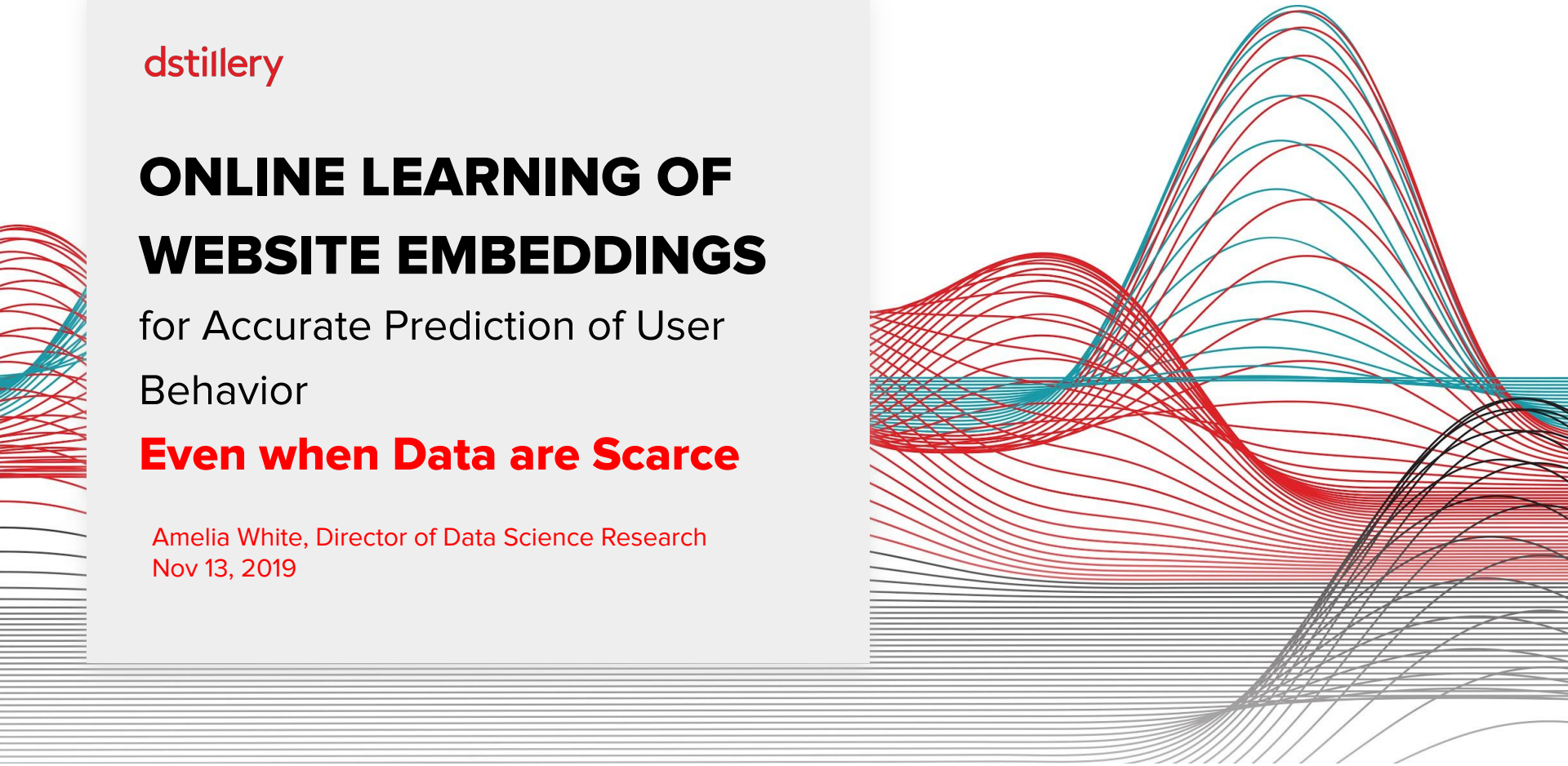
dstillery

ONLINE LEARNING OF WEBSITE EMBEDDINGS

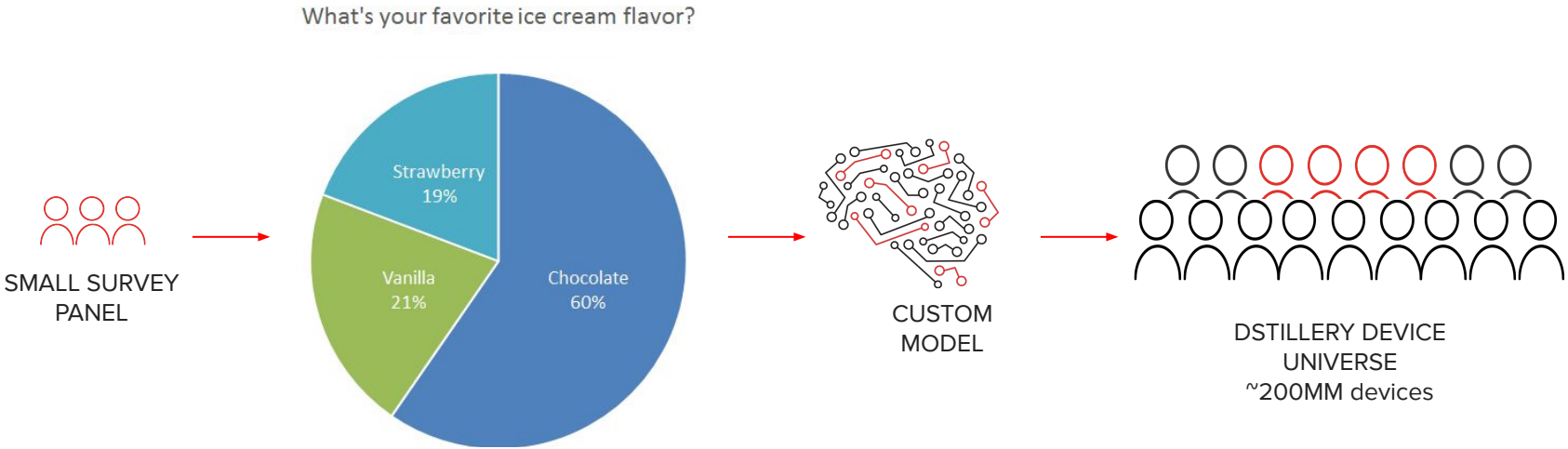
for Accurate Prediction of User Behavior

Even when Data are Scarce

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

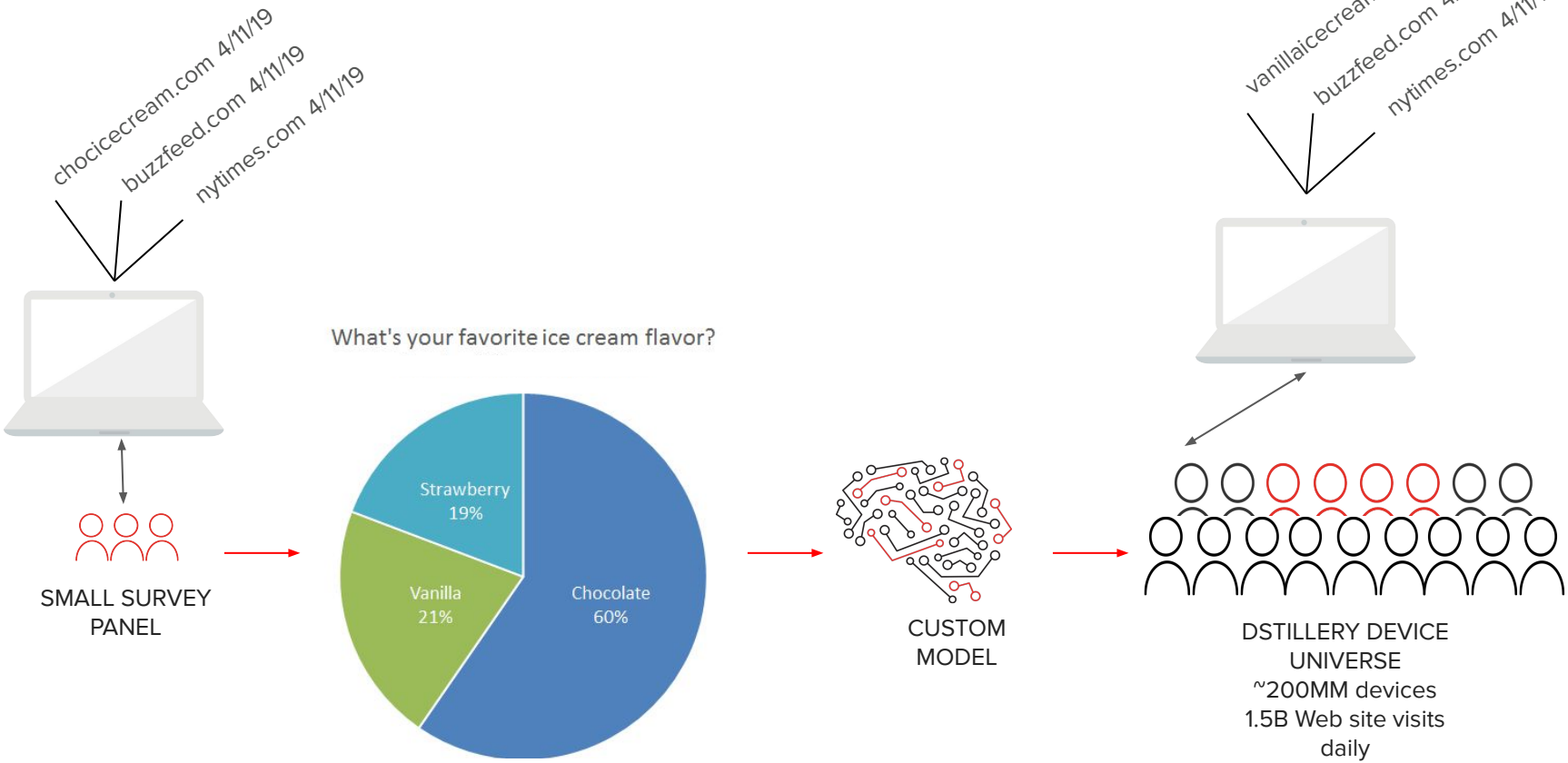


Expanding Digital Survey Data



dstillery

Data Used for Modeling



SMALL SURVEY PANEL

CUSTOM MODEL

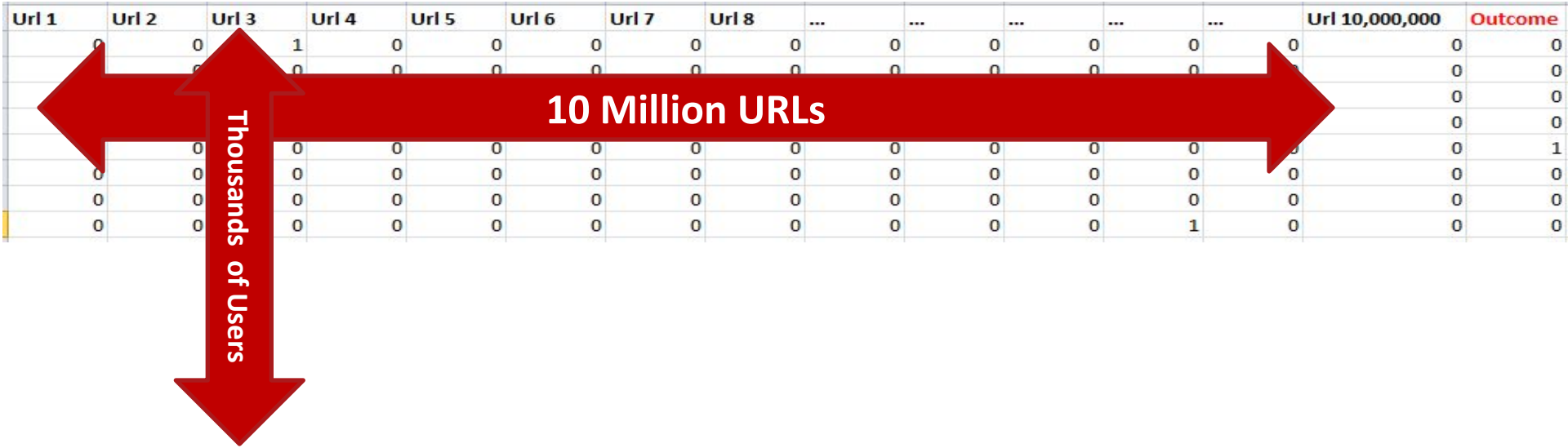
DSTILLERY DEVICE UNIVERSE
~200MM devices
1.5B Web site visits daily

Url 1	Url 2	Url 3	Url 4	Url 5	Url 6	Url 7	Url 8	...	...	...	...	...	Url 10,000,000	Outcome
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
0	[illegible]	[illegible]	0	0	0	0	0	0	0	0	0	0	0	1
[illegible]	0	0	0	0	0	0	0	0	0	0	0	[illegible]	0	0
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	0
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	0
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	0
[illegible]	0	0	0	1	0	0	0	0	0	0	0	[illegible]	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	1	0	0	0	0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	[illegible]	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

10 Million URLs

Millions of Users

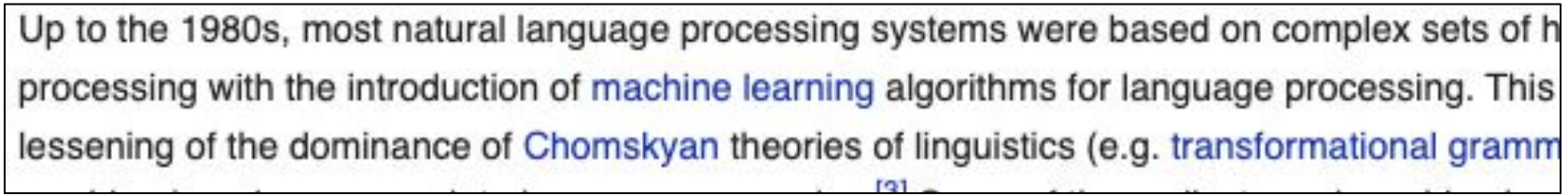
Need for a Reduced Dimensional Feature Space



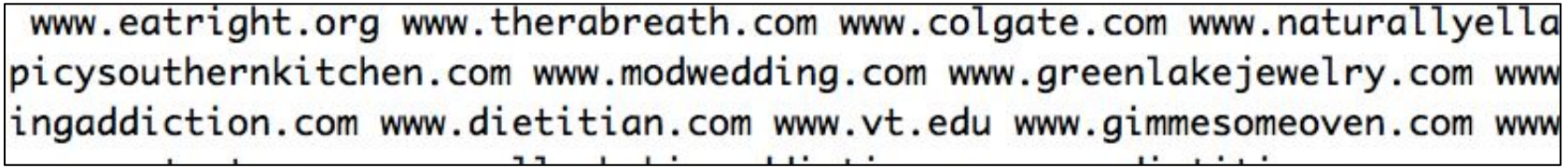
REDUCED DIMENSIONAL FEATURE SPACE

Taking Ideas from Natural Language Processing

- Similar data
- Sentences of words



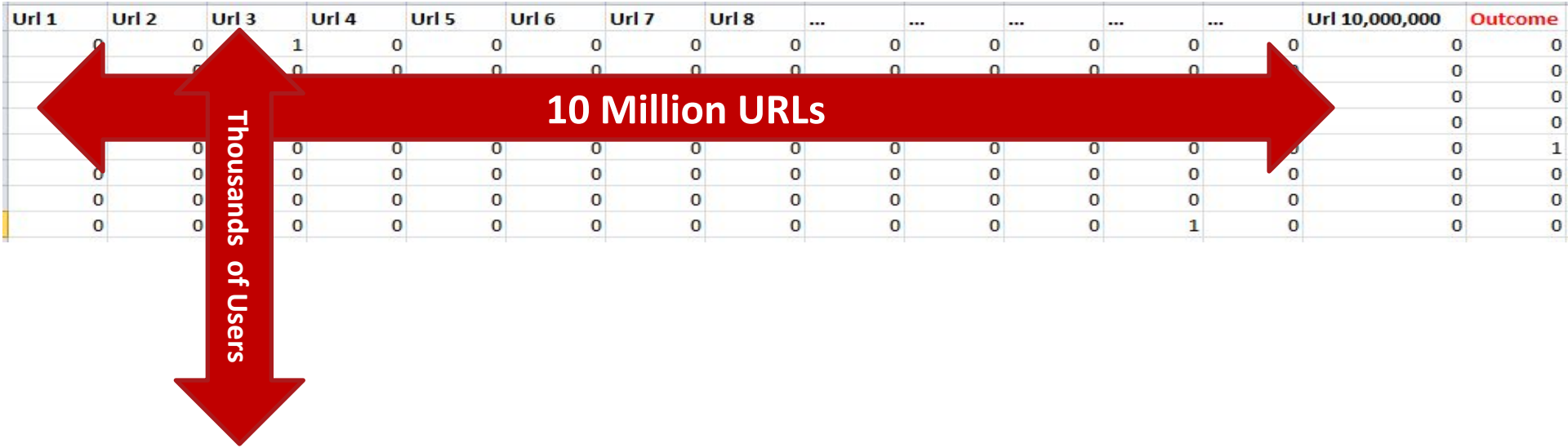
- Sequences of web sites visited



- High dimensional categorical features

dstillery

Need for a Reduced Dimensional Feature Space



Need for a Reduced Dimensional Feature Space

Url 1	Url 2	Url 3	Url 4	Url 5	Url 6	Url 7	Url 8	...	...	...	...	...	Url 10,000,000	Outcome
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
		0											0	0
													0	0
													0	0
	0	0	0	0	0	0	0	0	0	0	0		0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0

10 Million URLs

Thousands of Users

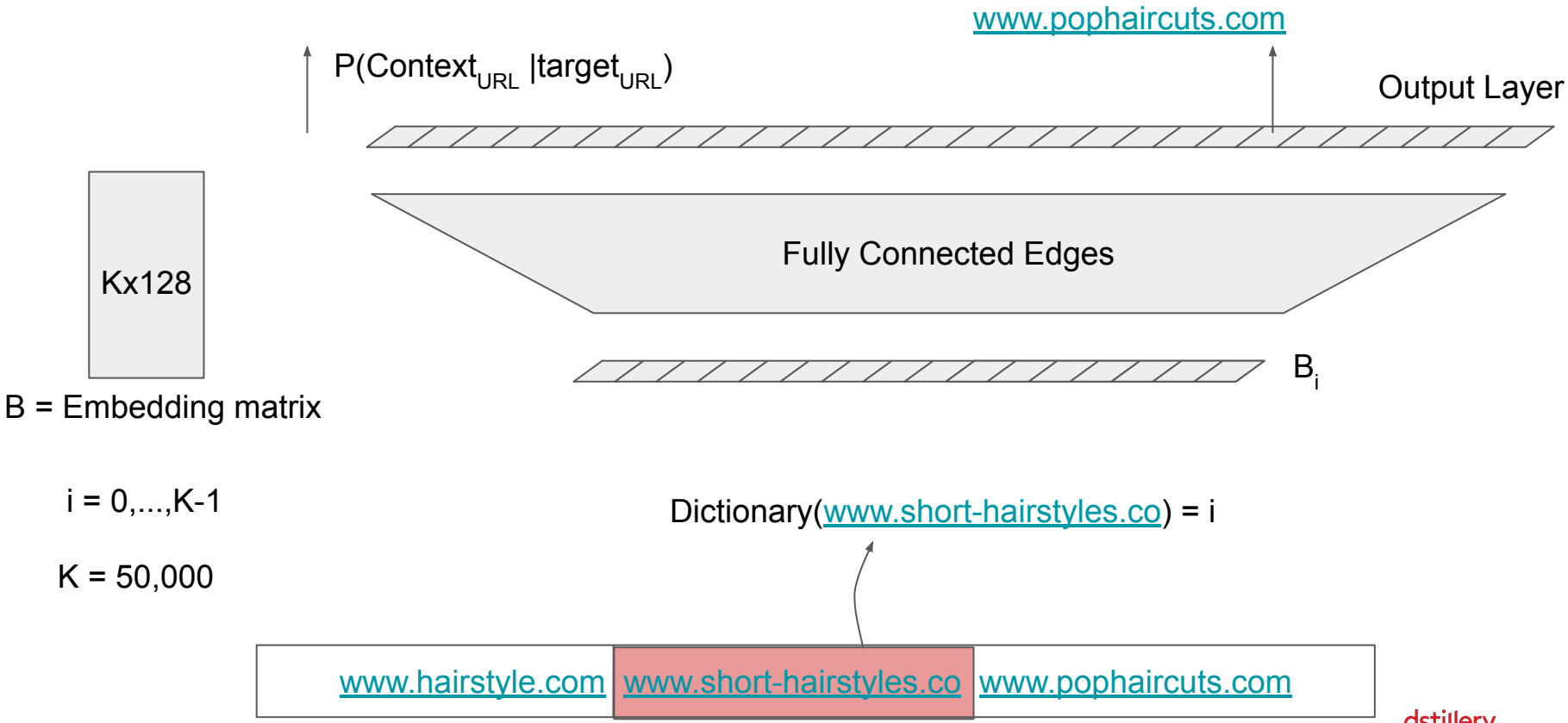
Dim 1	Dim 2	...	...	Dim 128
0.67534775	0.43516183	0.00371811	0.58666643	0.13523544
0.10092871	0.02024611	0.19984572	0.64574799	0.54387123
0.98997418	0.26743623	0.34408102	0.71425389	0.02140623
0.0845				35794342
0.6666				48017634
0.79902046	0.9786974	0.14587728	0.59378527	0.53994022
0.71107443	0.16079175	0.78204965	0.45080368	0.29320381
0.64726402	0.09479171	0.09246093	0.87526155	0.51668014

128 Dimensional Embedding Space

Thousands of Users

dstillery

Website Embeddings V1: word2vec



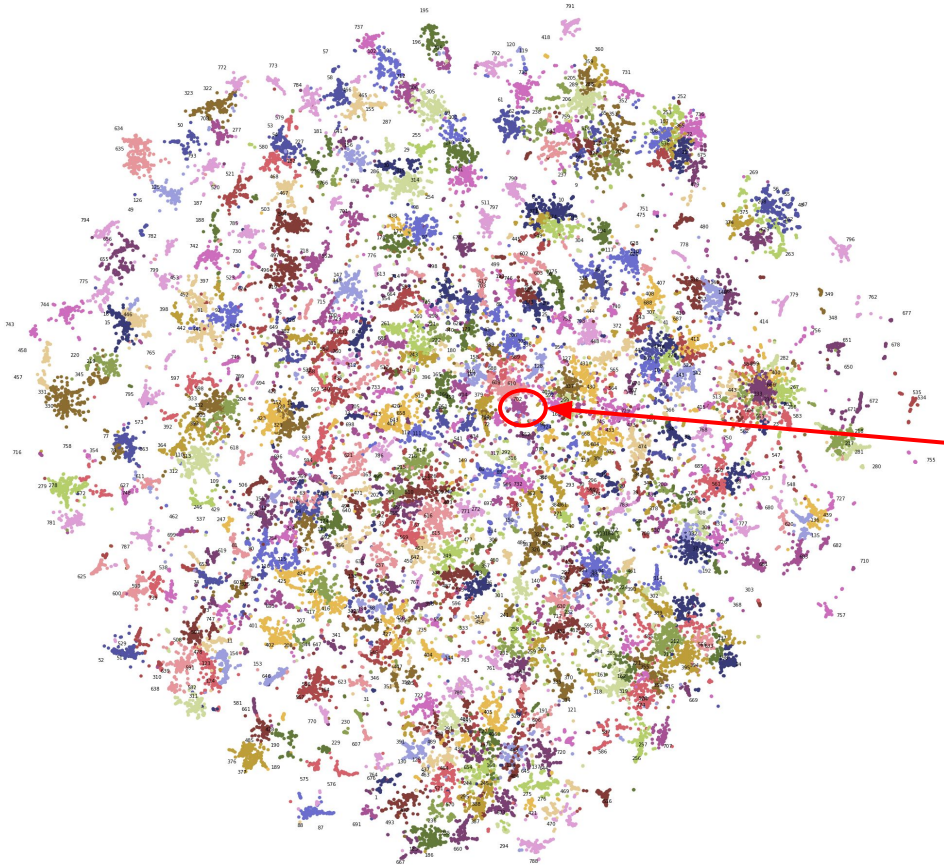
dstillery

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

dstillery

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

dstillery

BEYOND WORD2VEC:

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

EMBEDDING MORE URLS WITH FEWER PARAMETERS

Hash Embeddings

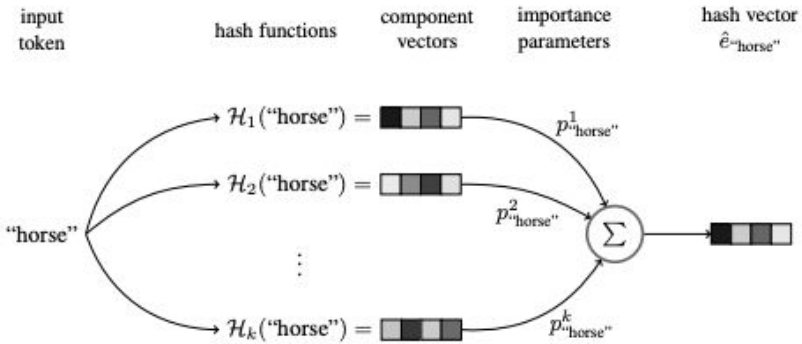
Hash Embeddings for Efficient Word Representations

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

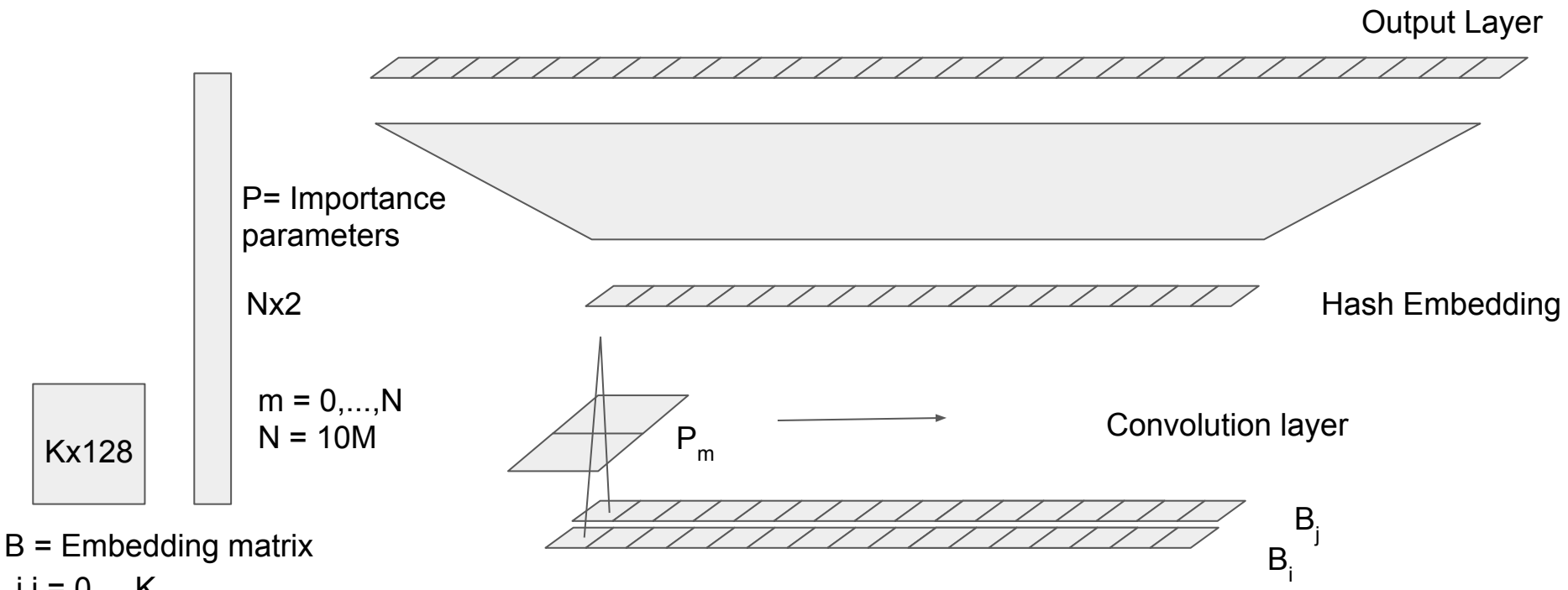
Ole Winther
Department for Applied Mathematics and Computer Science
Technical University of Denmark (DTU)
2800 Lyngby, Denmark
olwi@dtu.dk

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



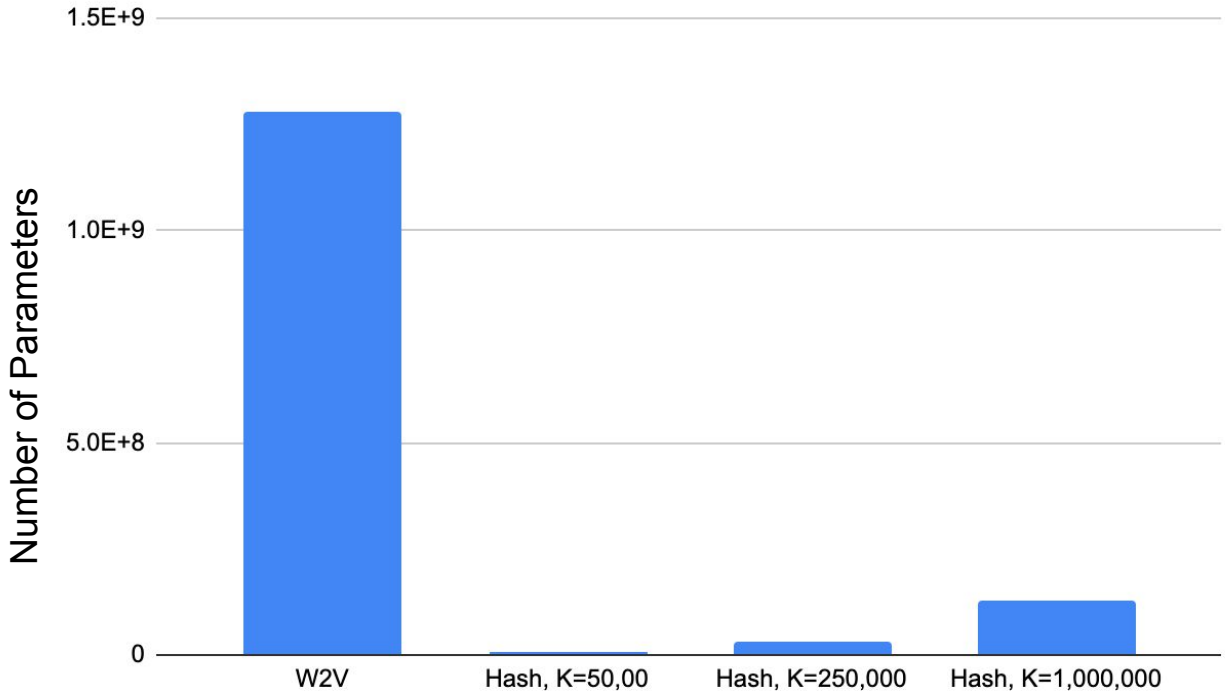
dstillery

Website Embeddings V2: Hash embeddings



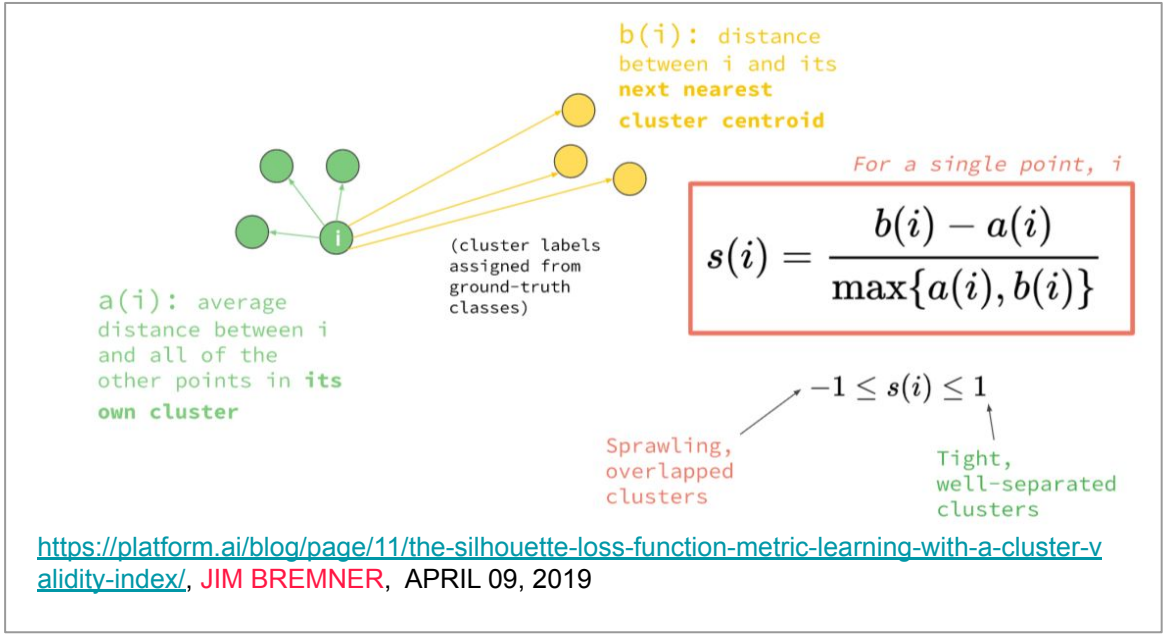
dstillery

Hash Embedding Requires Fewer Parameters



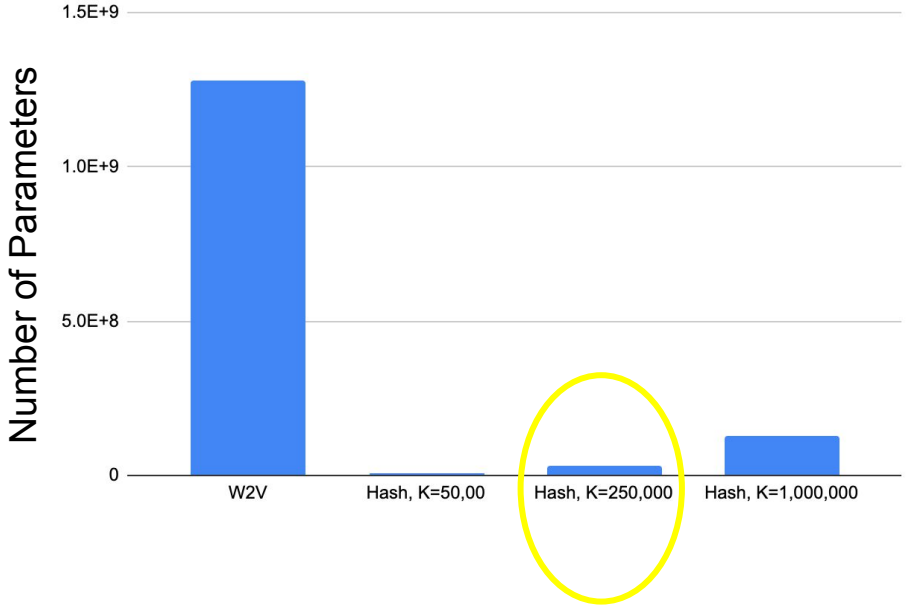
Measuring Embedding Quality for Parameter Selection

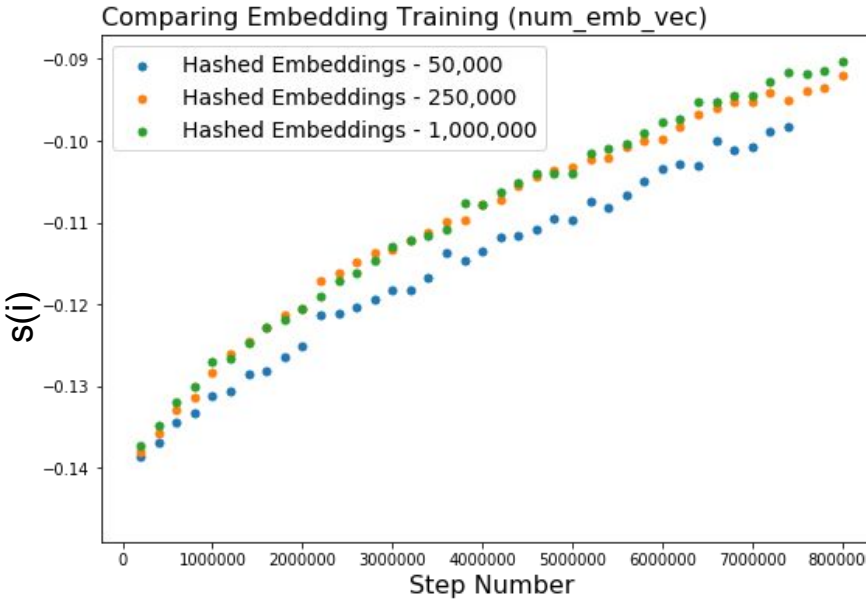
- Selected a ‘ground truth’ clustering, made from a known high quality embedding
- Used the silhouette score to measure how well test embeddings converged to the ground truth clustering as the network trained



https://platform.ai/blog/page/11/the-silhouette-loss-function-metric-learning-with-a-cluster-validity-index/, JIM BREMNER, APRIL 09, 2019

Good Performance with 100x Fewer Parameters

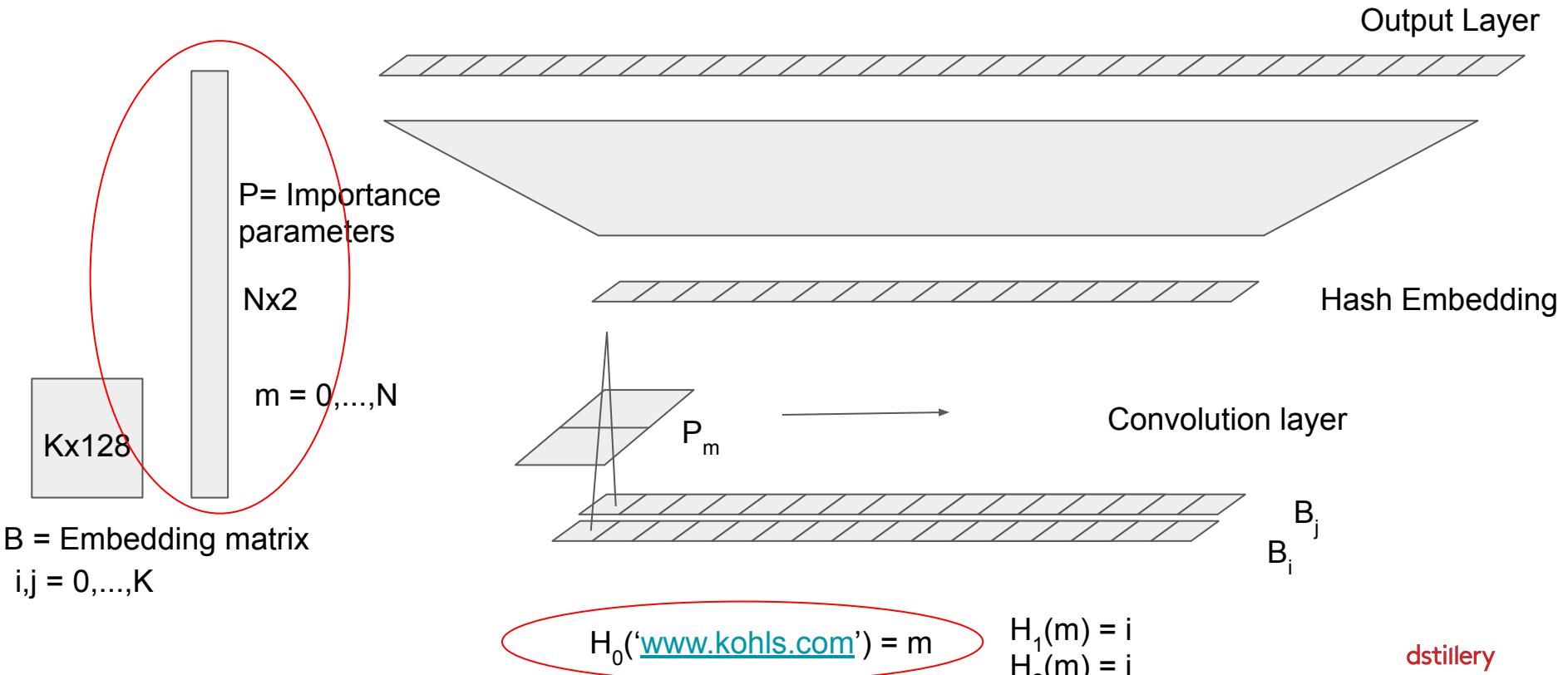




dstillery

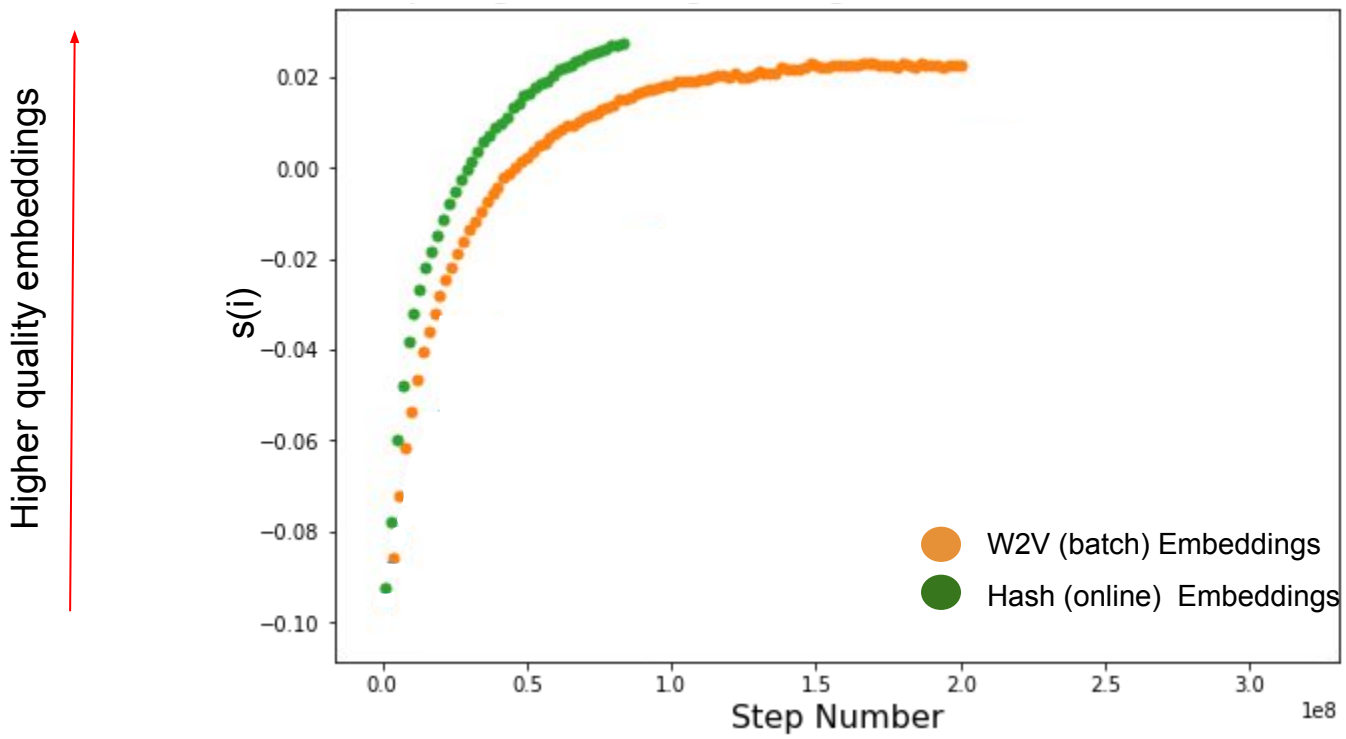
ONLINE LEARNING OF EMBEDDINGS

Website Embeddings V3: Online Learning of Hash Embeddings



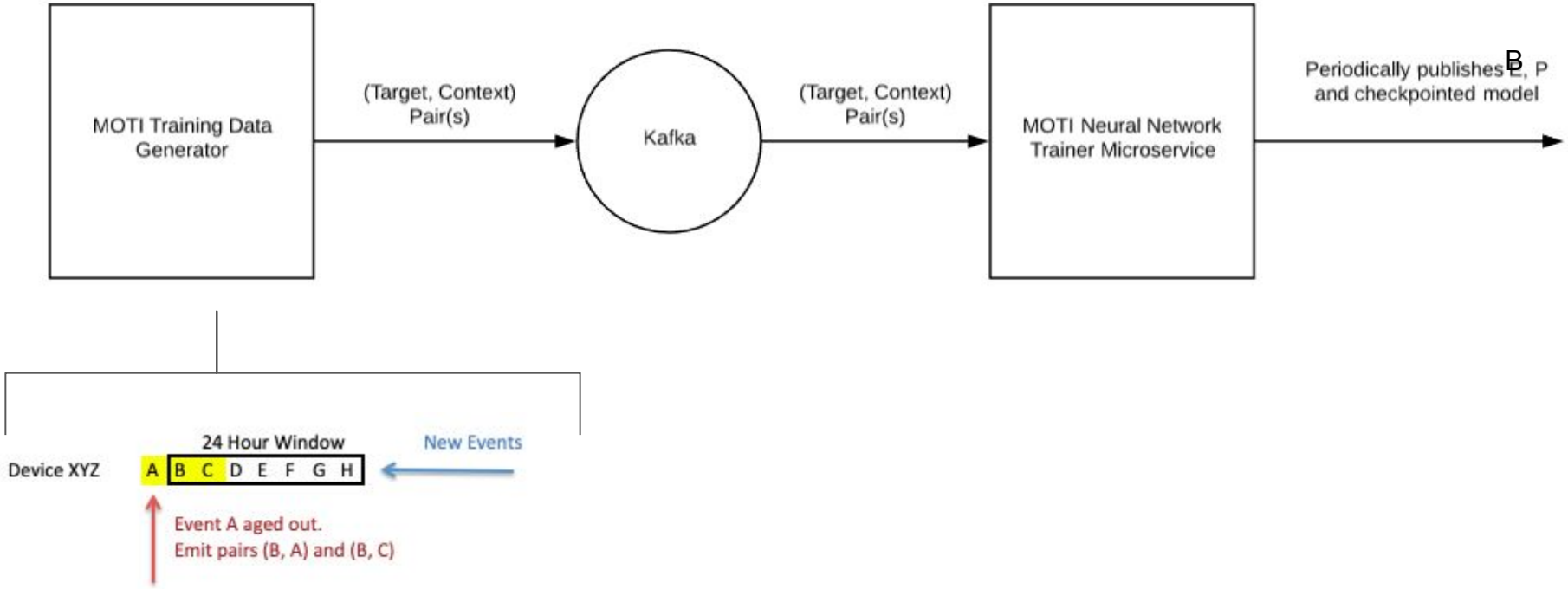
dstillery

Online Learning Optimizes Faster than Batch Learning

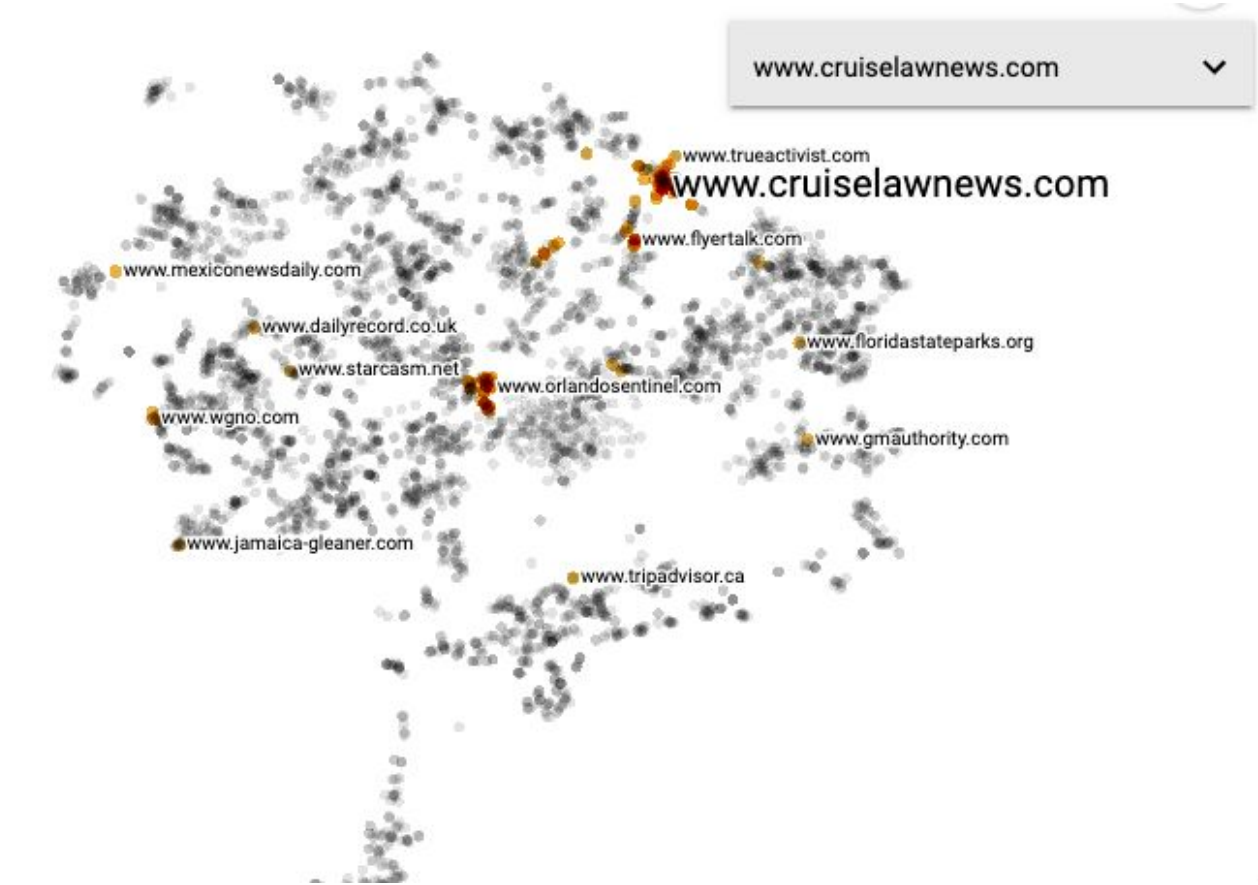


dstillery

Training the Online Embeddings



Distance in Embedding Space



www.cruiselawnews.com

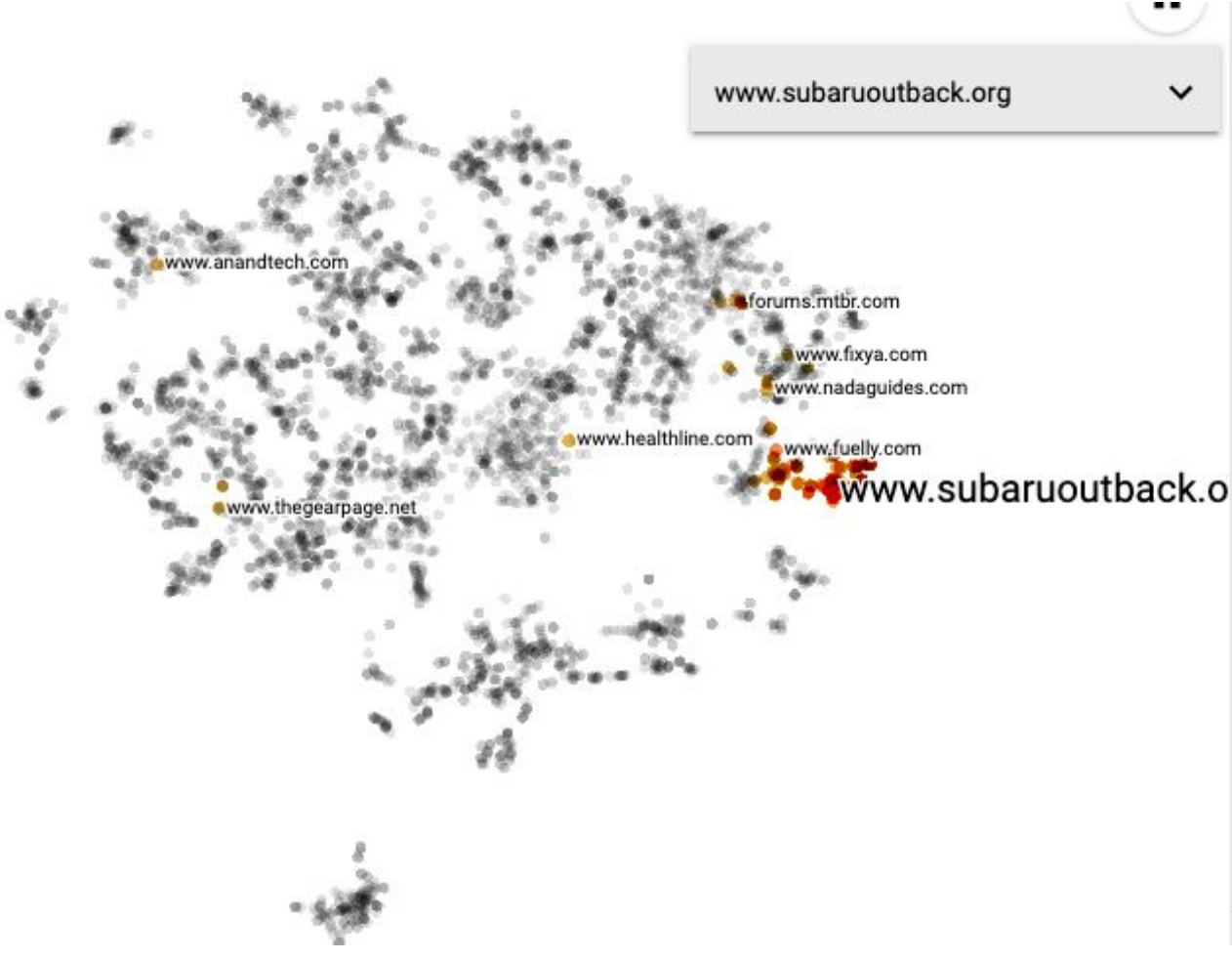
Search .* by

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



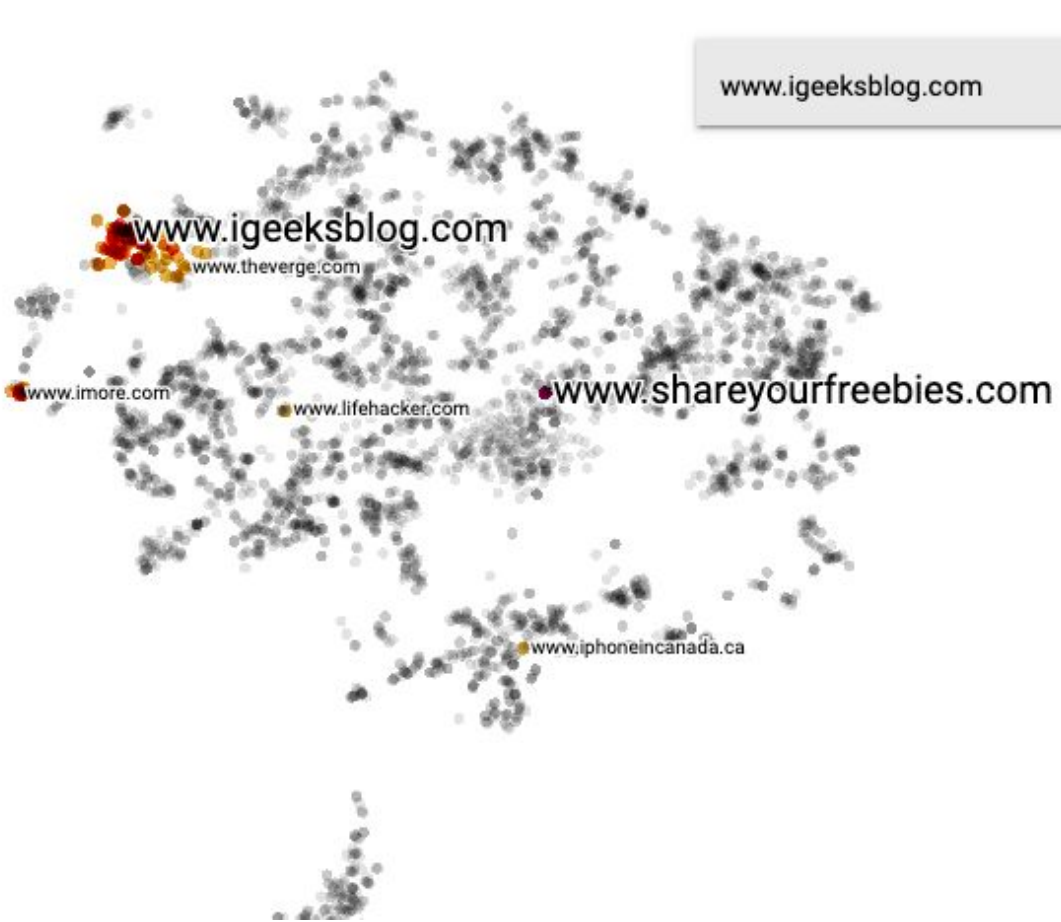
Search .* by

neighbors 100

distance COSINE EUCLIDEAN

Nearest points in the original space:

www.1010tires.com	0.378
www.toyotanation.com	0.385
www.michelinman.com	0.391
www.acurazine.com	0.404
www.carcomplaints.com	0.407
www.toyota-4runner.org	0.426
www.700r4transmissionhq.com	0.438
www.advanceautoparts.com	0.446
www.ih8mud.com	0.447
www.autoanything.com	0.450



www.igeeksblog.com

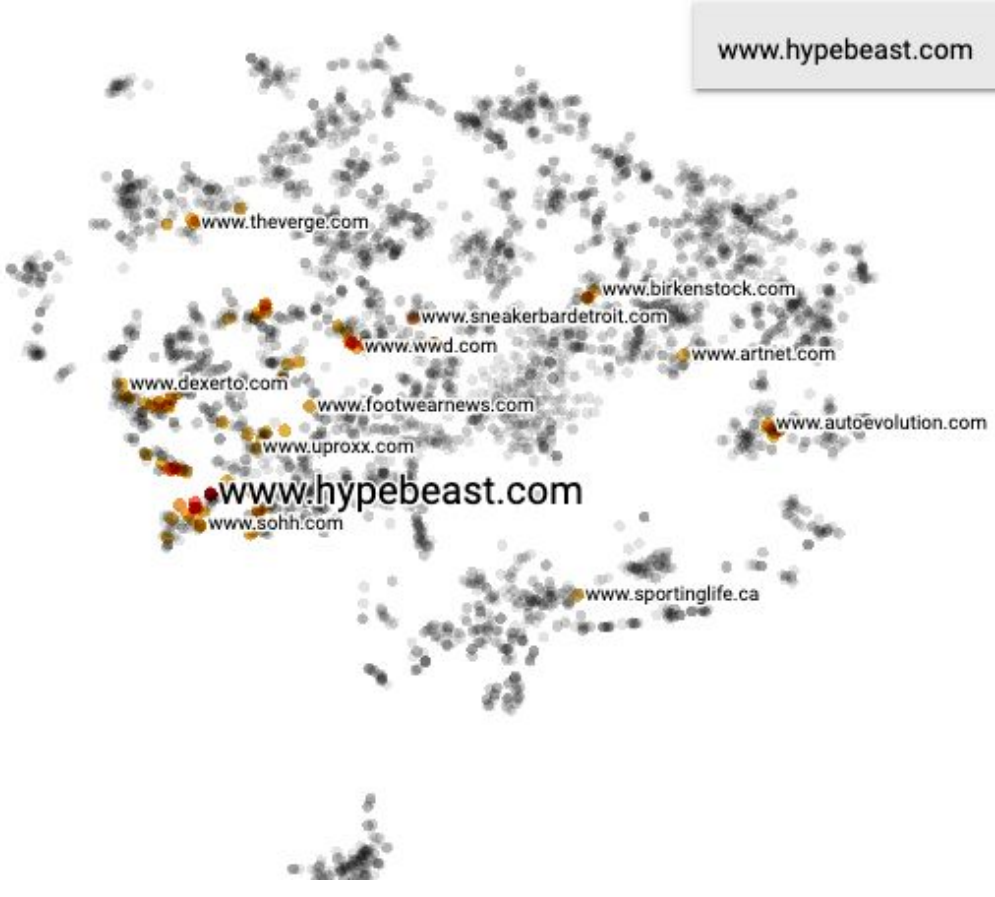
Search .* by

neighbors 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.techzillo.com	0.293



www.hypebeast.com

Search .* by

neighbors 100

distance COSINE EUCLIDEAN

Nearest points in the original space:

www.solecollector.com	0.186
www.sneakernews.com	0.229
www.nicekicks.com	0.236
www.kicksonfire.com	0.272
www.complex.com	0.359
www.hotnewhiphop.com	0.370
www.thefader.com	0.424
www.sneakerbardetroit.com	0.471
www.vogue.com	0.473
www.gq.com	0.484

MODELING USERS IN EMBEDDING SPACE

Need for a Reduced Dimensional Feature Space

Url 1	Url 2	Url 3	Url 4	Url 5	Url 6	Url 7	Url 8	...	...	...	...	...	Url 10,000,000	Outcome
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
		0											0	0
													0	0
													0	0
	0	0	0	0	0	0	0	0	0	0	0		0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0

10 Million URLs

Thousands of Users

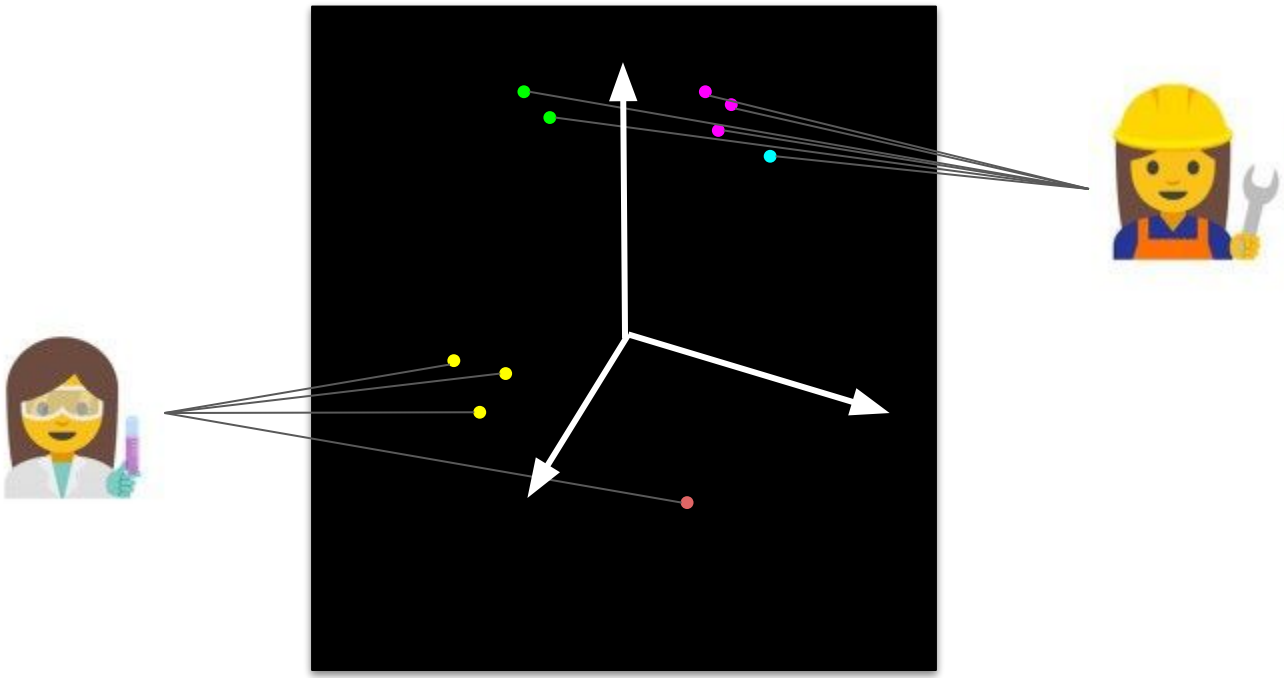
Dim 1	Dim 2	...	...	Dim 128
0.67534775	0.43516183	0.00371811	0.58666643	0.13523544
0.10092871	0.02024611	0.19984572	0.64574799	0.54387123
0.98997418	0.26743623	0.34408102	0.71425389	0.02140623
0.0845				35794342
0.6666				48017634
0.79902046	0.9786974	0.14587728	0.59378527	0.53994022
0.71107443	0.16079175	0.78204965	0.45080368	0.29320381
0.64726402	0.09479171	0.09246093	0.87526155	0.51668014

128 Dimensional Embedding Space

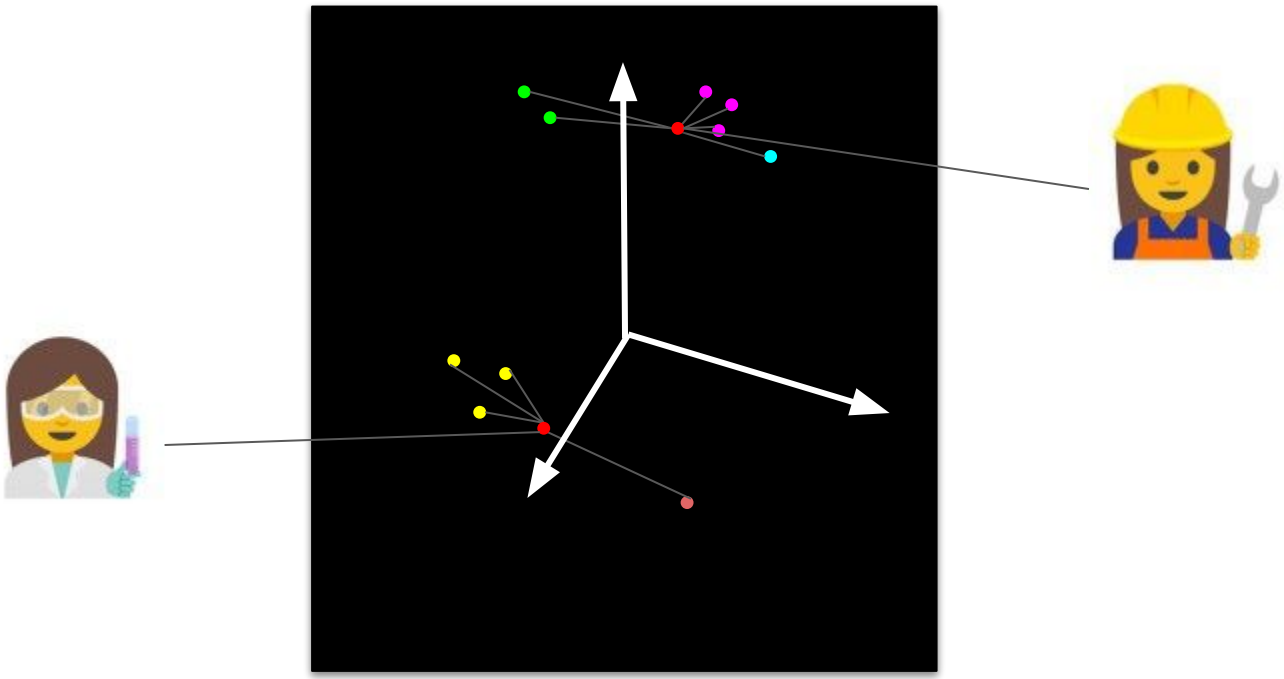
Thousands of Users

dstillery

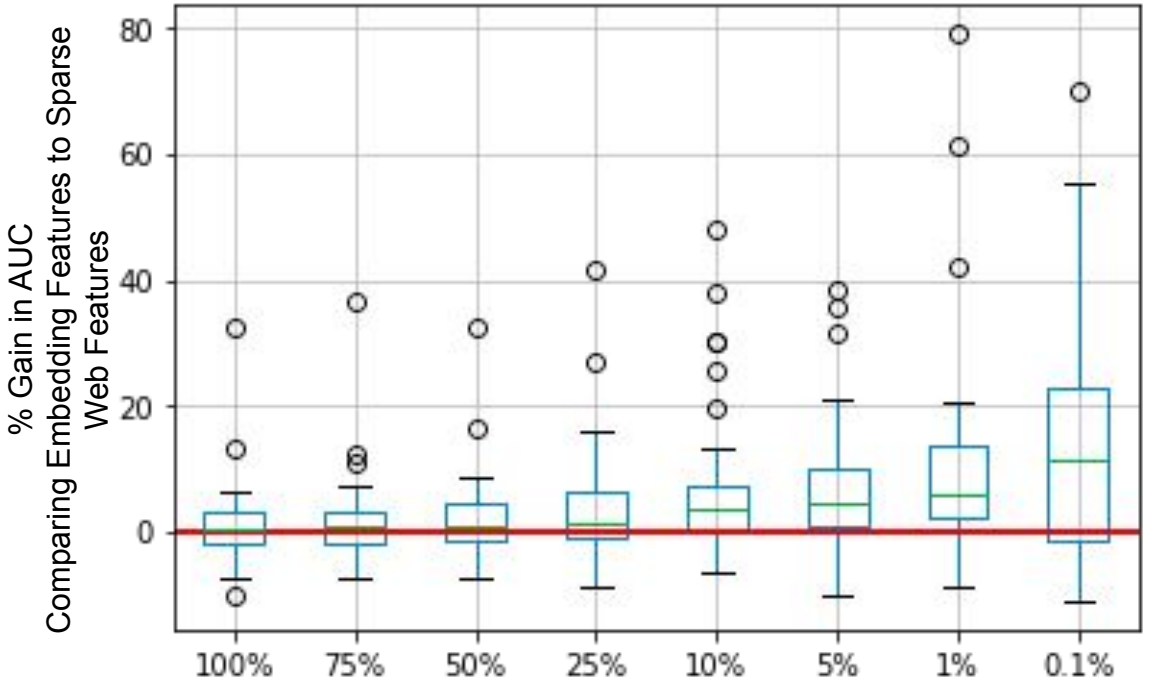
From URL Embeddings to Models



From URL Embeddings to Models

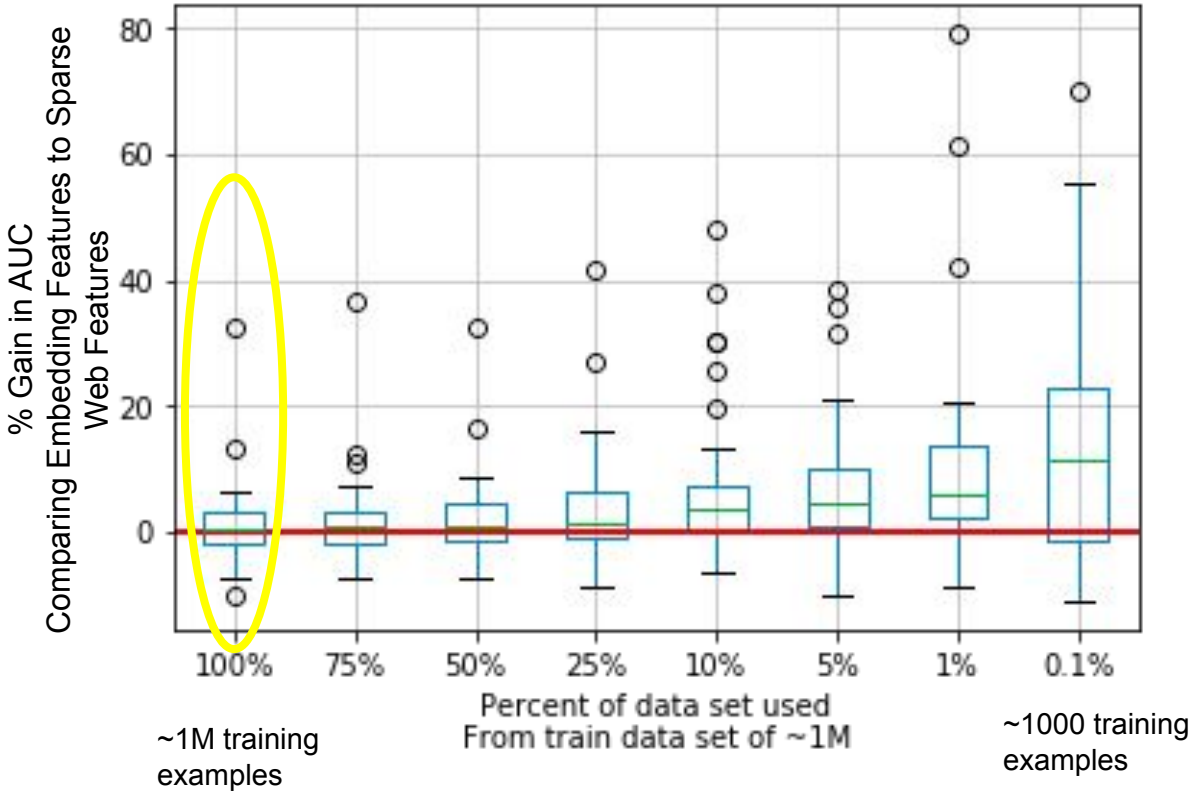


Embedding Features Outperform Sparse Web Features For Small Data Sets



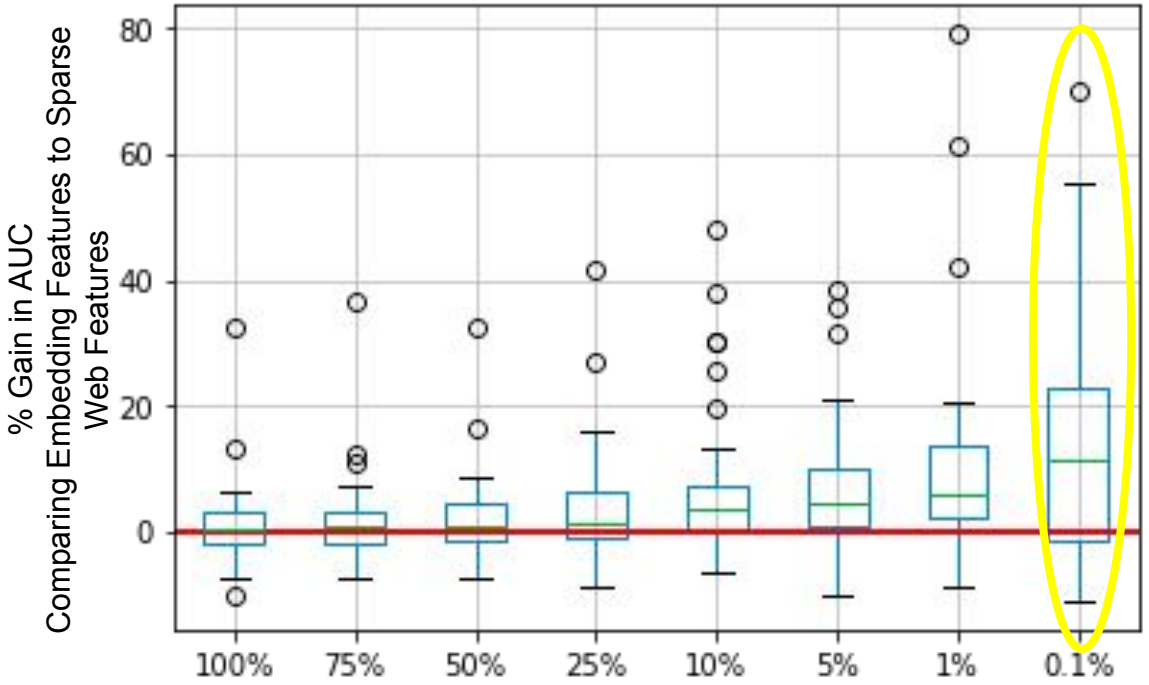
dstillery

Embedding Features Outperform Sparse Web Features For Small Data Sets



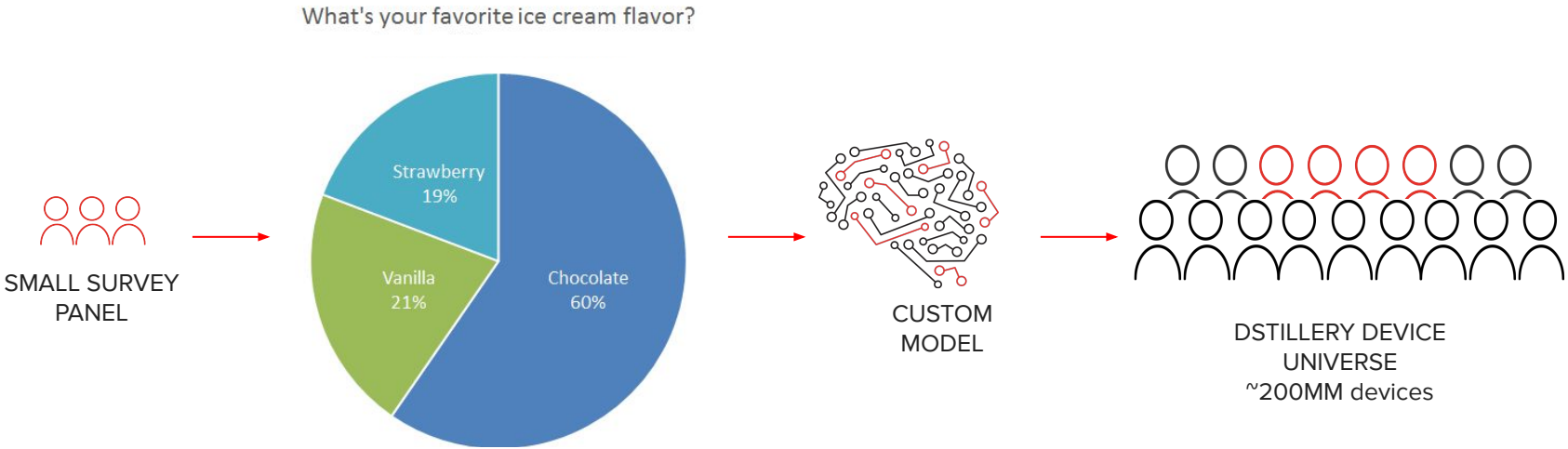
dstillery

Embedding Features Outperform Sparse Web Features For Small Data Sets



dstillery

MODELING SURVEY DATA



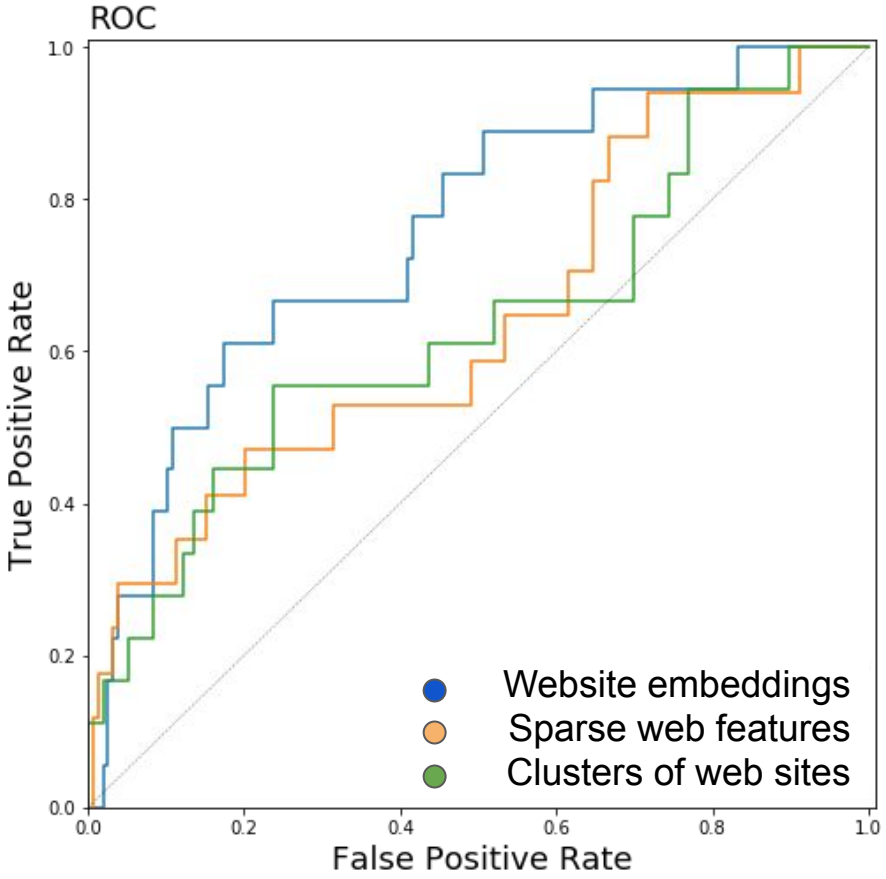
Case Study: Predicting Ad Influence for Ice Cream Brand

- 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

dstillery

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

dstillery

THANK YOU

Presented by Amelia White.
awhite@dstillery.com

Contributors:
Christopher Jenness
Melinda Han Williams

MLE team:
Wickus Martin
Roger Cost
Justin Moynihan
Patrick McCarthy