machine learning conference

Identifying Miscategorized Products

M. Hidayath Ansari hidayata@
Applied Scientist II | Browse

Subhadeep Chakraborty sdeep@
Sr SDE | Browse

AMAZON CONFIDENTIAL

Customer Experience – Shopping Funnel

Customer Experience – Shopping Funnel

Customer Experience – Shopping Funnel

Humidifier
Sleeves

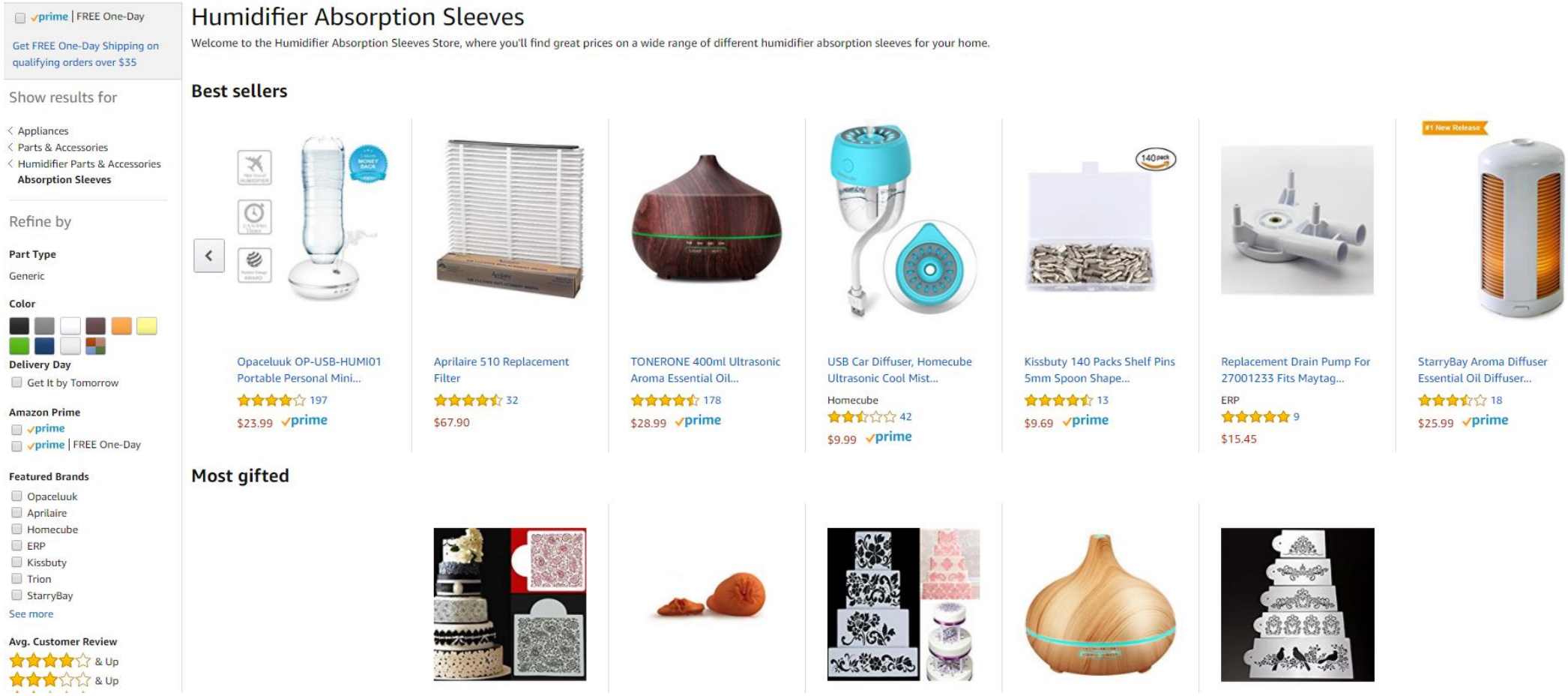
Customer Experience – Shopping Funnel

Humidifier
Sleeves

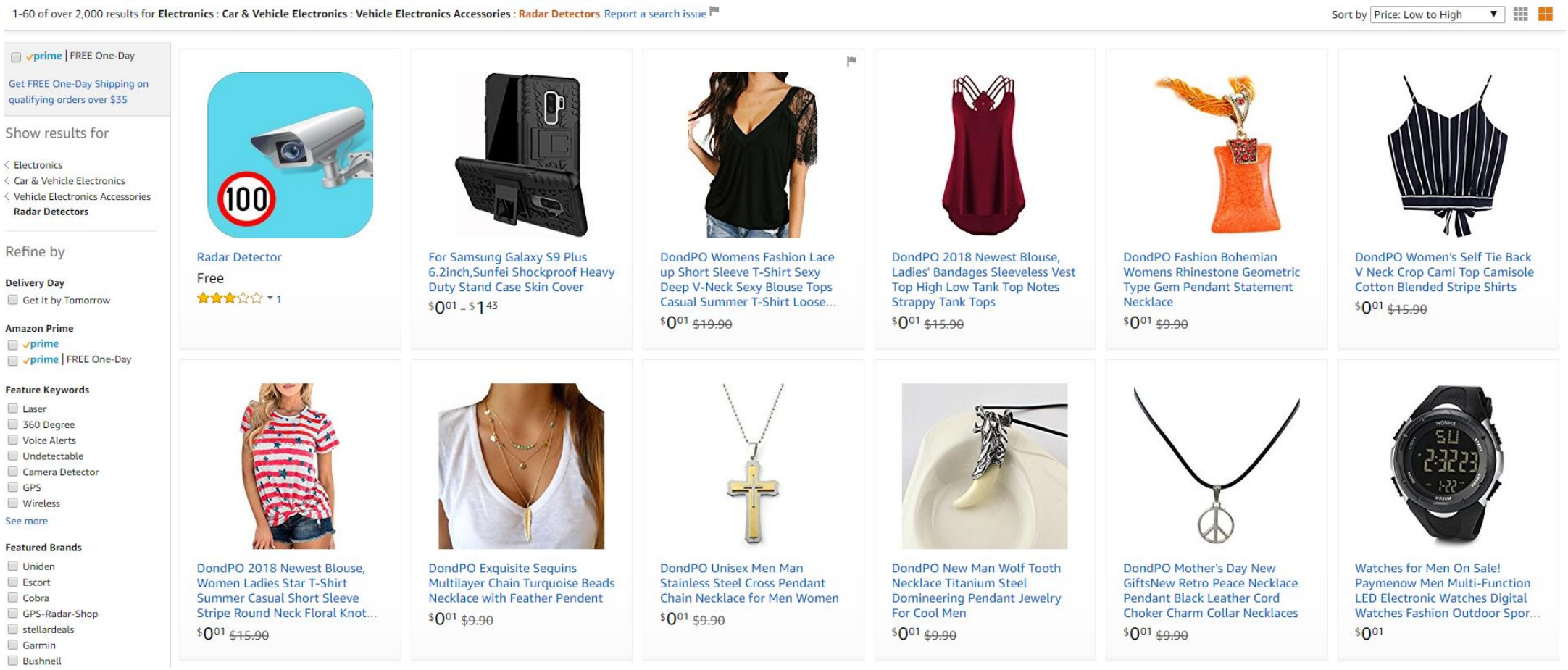
Customer Experience – Shopping Funnel

Humidifier
Sleeves

Customer Experience – Category Landing Page



Customer Experience: Sort by Price



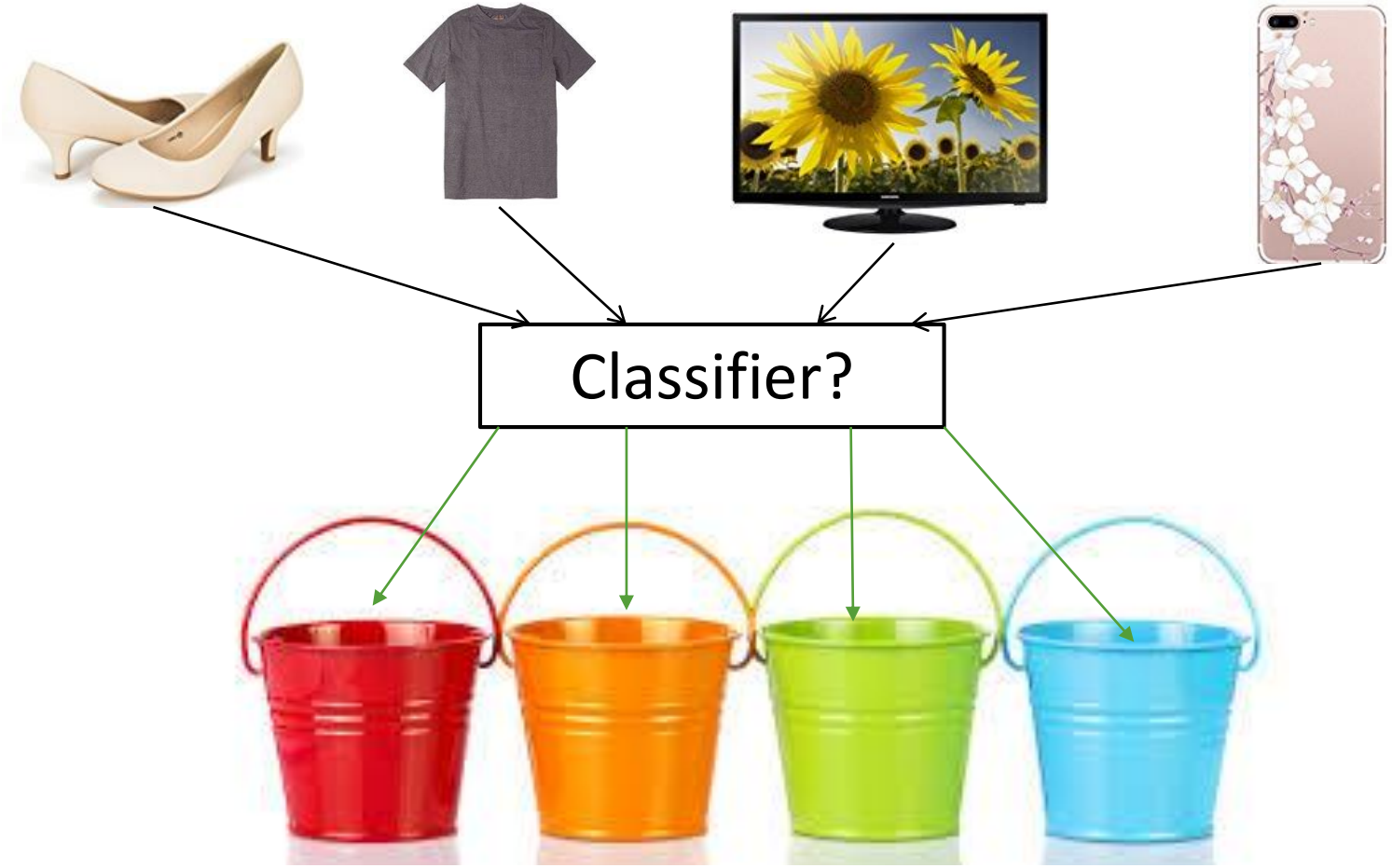
"Radar Detectors"

Product Discovery

Amazon Browse: Enabling customers' discovery experience

Without having to weed through irrelevant products

Finding Product Miscategorizations



BUT …

Product Miscategorizations as Outliers

Product Miscategorizations as Outliers

“Which of these doesn’t belong?”

AMAZON CONFIDENTIAL

Product Miscategorizations as Outliers

"Which of these doesn't belong?"

Product Miscategorizations as Outliers

"Which of these doesn't belong?"



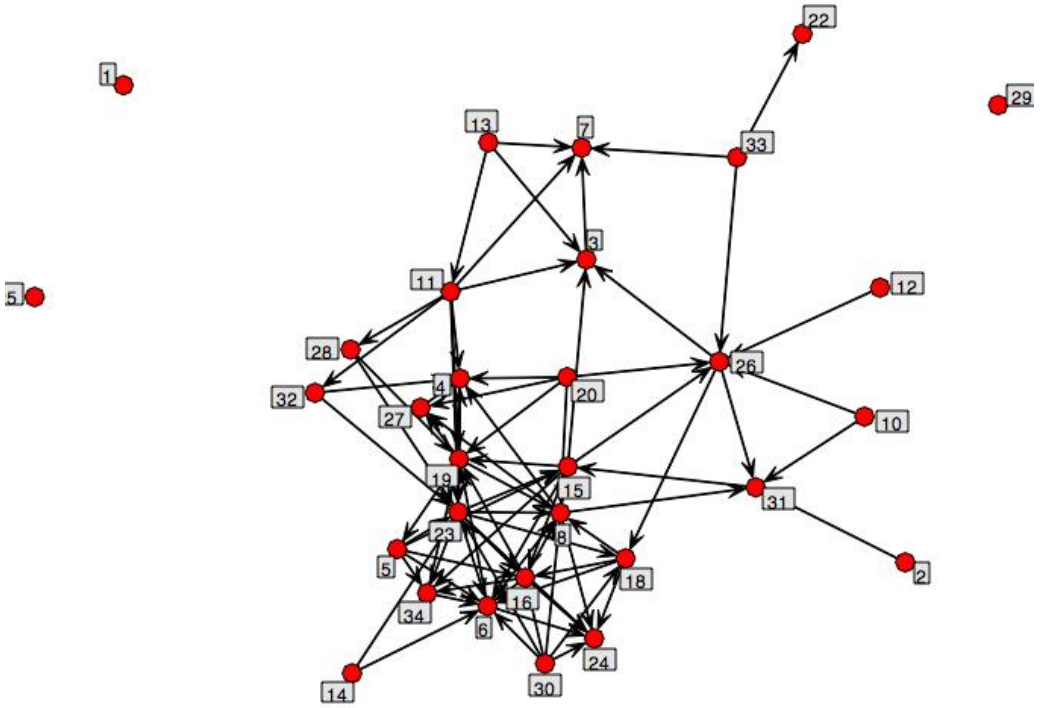




Unsupervised/Lightly supervised methods

PageRank

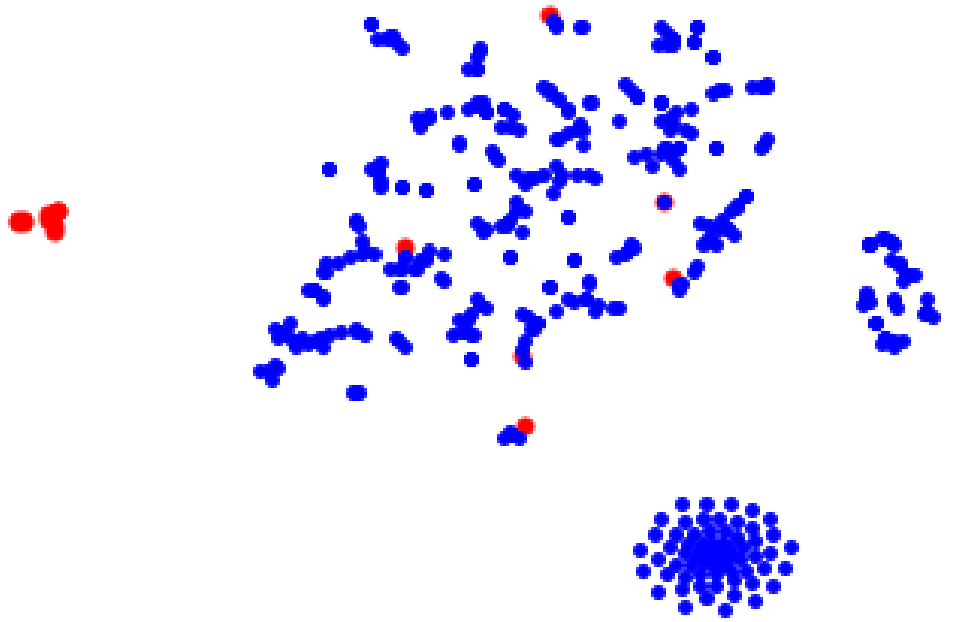
- Finds the most “central”/”important nodes
- Conversely …



Outrank – Example with ASINs

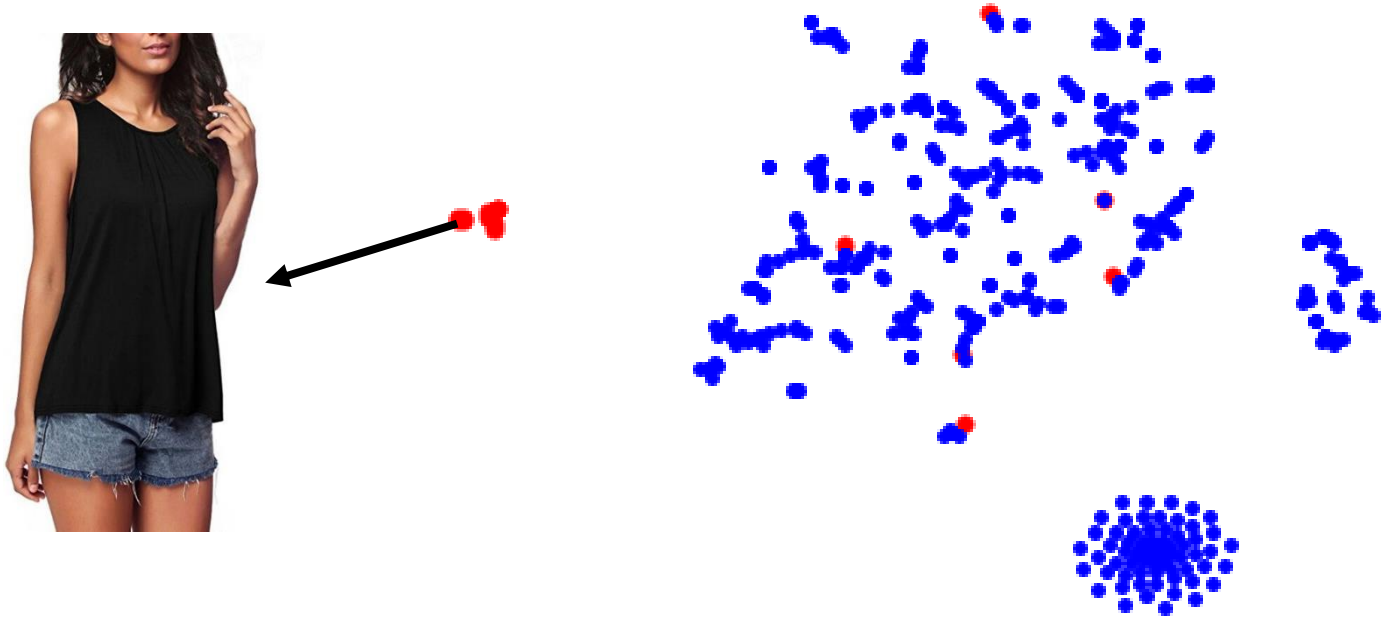
Dishwashers (dimensionally reduced)

Outrank – Example with ASINs



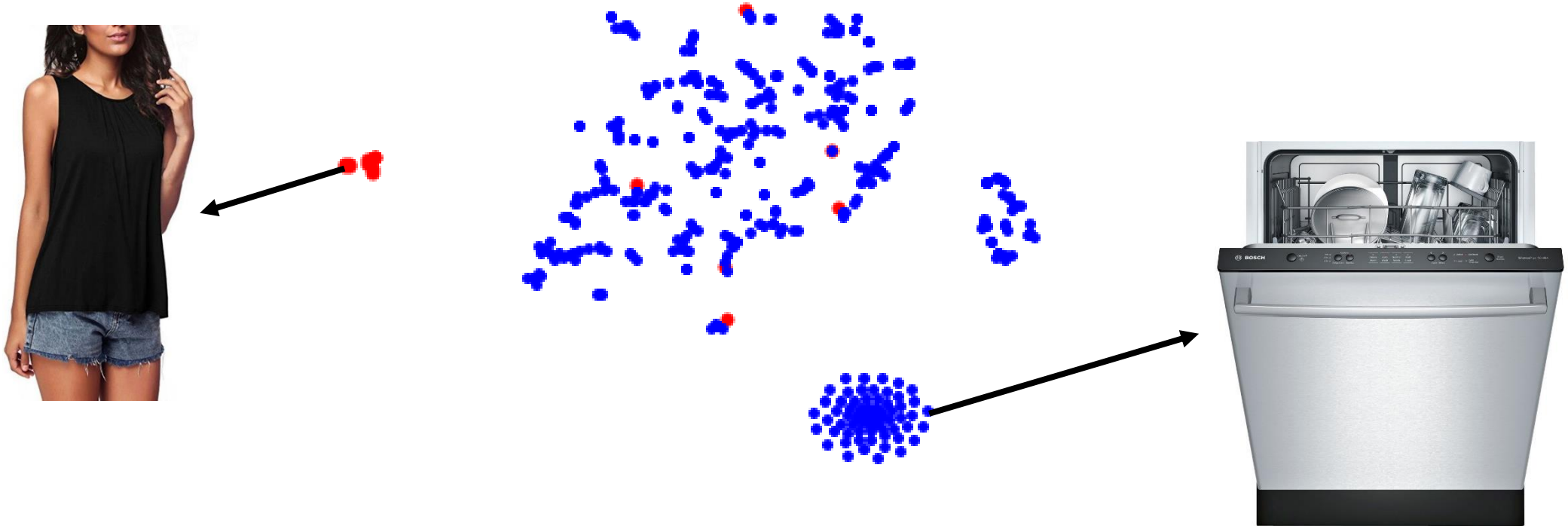
Dishwashers (dimensionally reduced)

Outrank – Example with ASINs



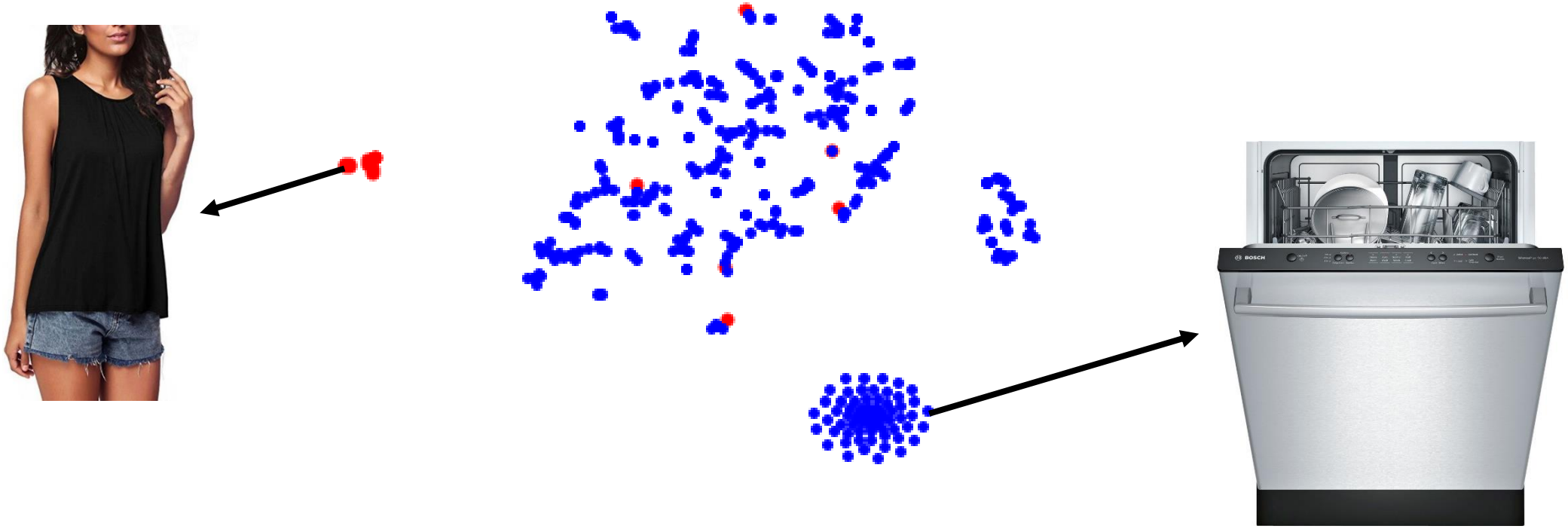
Dishwashers (dimensionally reduced)

Outrank – Example with ASINs



Dishwashers (dimensionally reduced)

Outrank – Example with ASINs



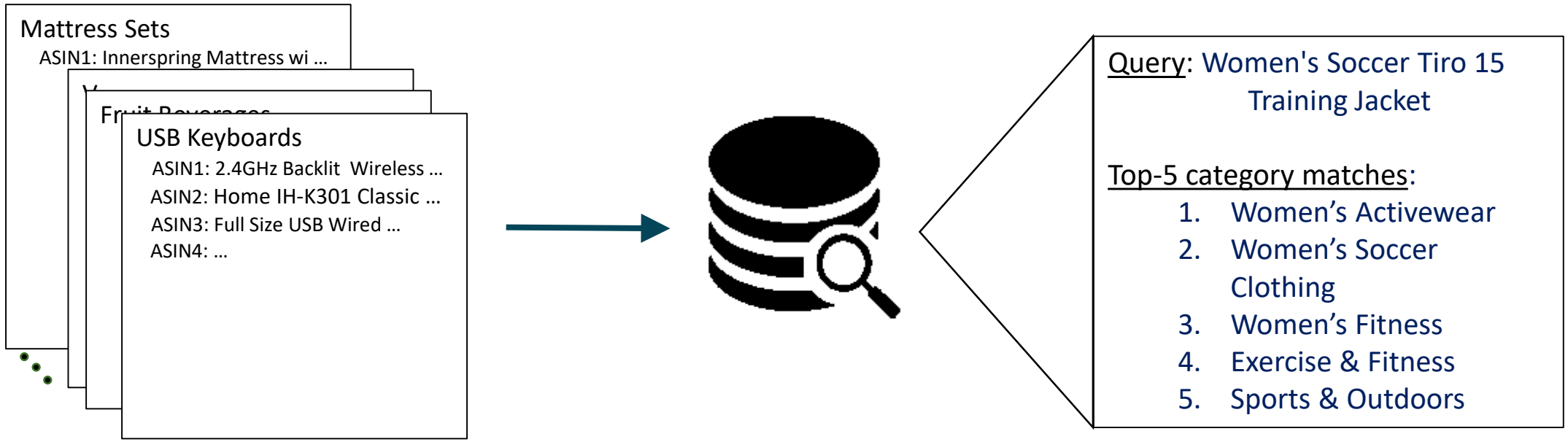
Still, Outrank has a high false positive rate

Dishwashers (dimensionally reduced)

Moonesinghe, H. D. K., and Pang-Ning Tan. "Outrank: a graph-based outlier detection framework using random walk." *International Journal on Artificial Intelligence Tools* 17.01 (2008): 19-36.

Miscategorizations – Using Search

Idea: Use a search index for product categories to find miscategorized products



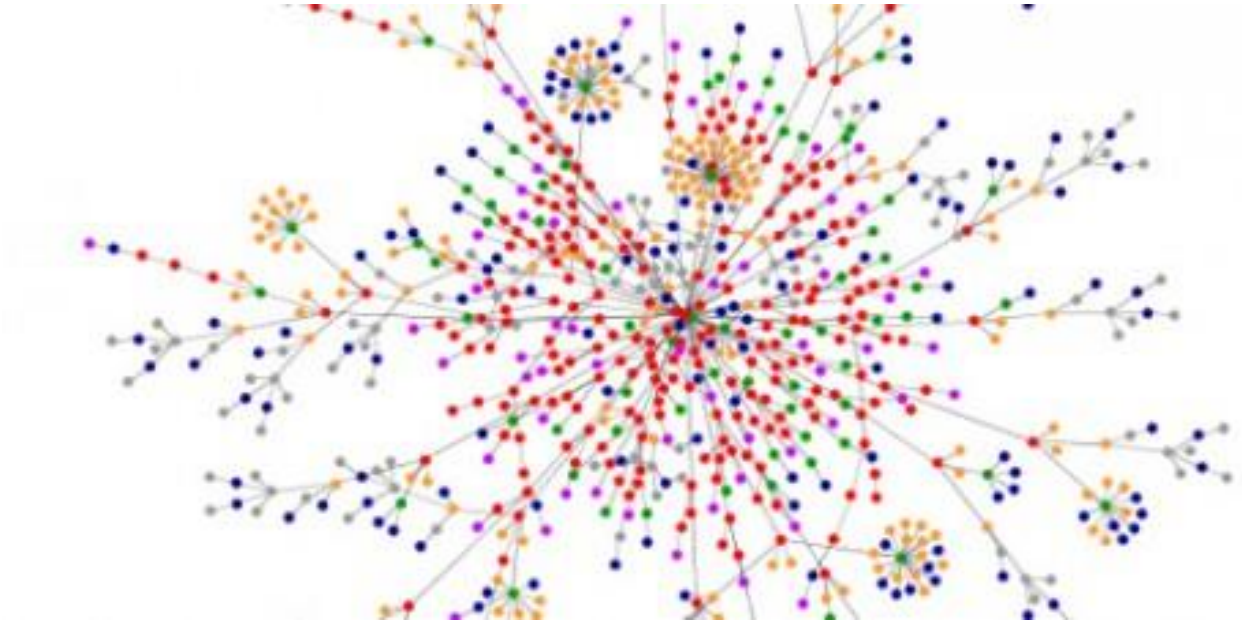
Singular outliers – Results & Impact

Marketplace	Assignments removed	Audited miscat Precision	Irrelevancy Reduction	OPS impact through irrelevancy reduction
US	2,319,105	91.8%	24% → 9%	$15,925,972
UK	2,168,596	97.2%	20% → 14%	$4,886,300

Results for bestseller ASINs

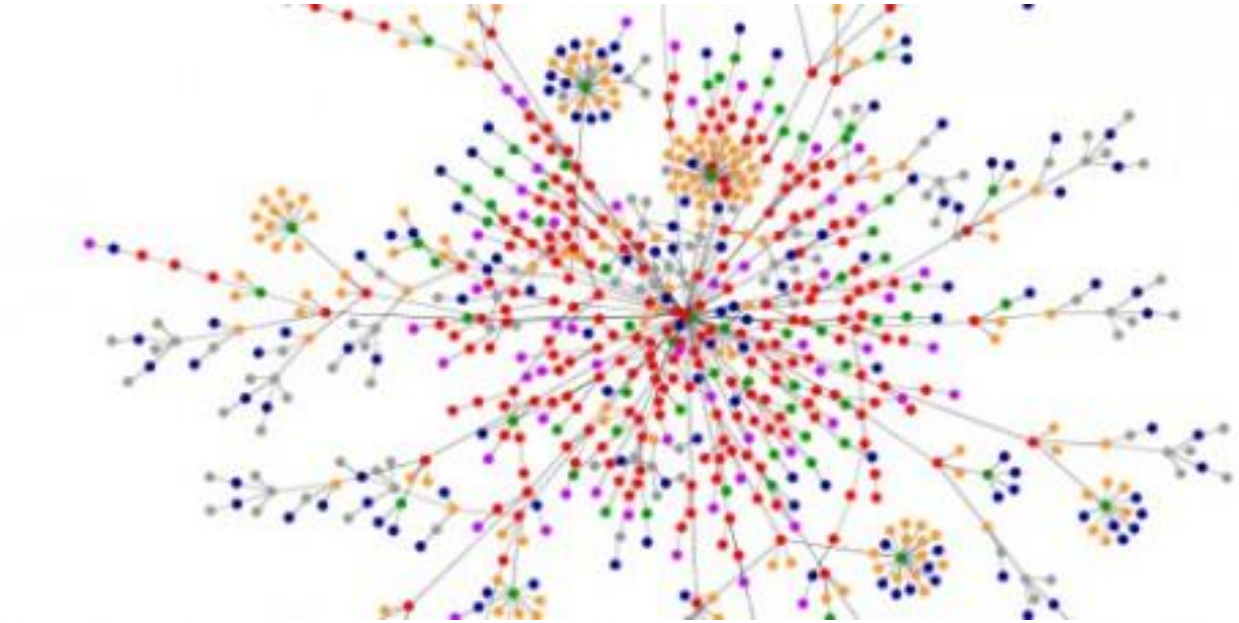
AMAZON CONFIDENTIAL

Clump outliers – Using expert input



Clump outliers – Using expert input

1. Select “exemplars”
2. Get judgments
3. Propagate judgments



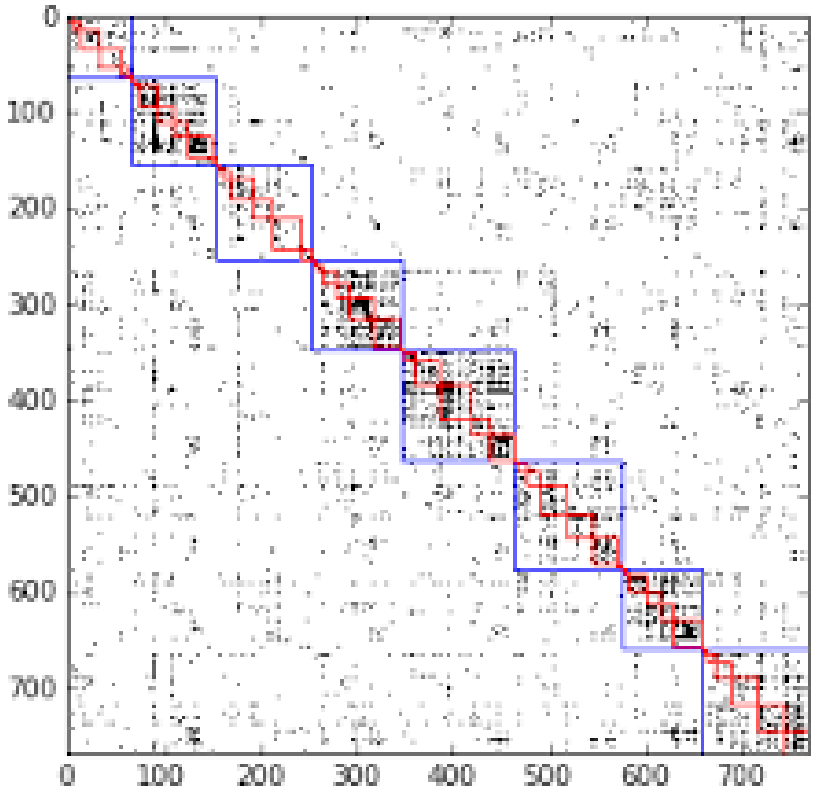
Clump outliers – Affinity Matrix

Constructing a graph adjacency/ASIN affinity matrix:

1. Vectorize ASINs using averaged GloVe
2. Find the 250 nearest neighbors of each ASIN
3. Construct matrix using (Euclidean) distances to nearest neighbors
4. Apply Gaussian kernel to each non-zero entry:

$$A_{ij} = \exp\left(\frac{-dist(\boldsymbol{x}_i, \boldsymbol{x}_j)}{2\sigma^2}\right)$$

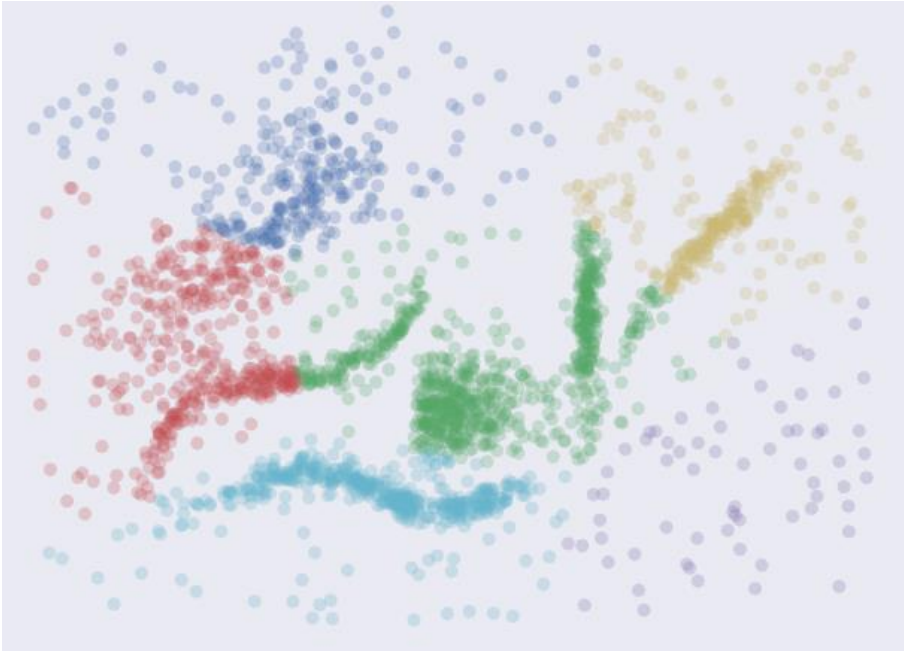
5. Symmetrize matrix



Clump outliers – Exemplar Selection

Choosing a set of “representative” ASINs:

1. Spectral clustering using affinity matrix
2. Compute the “degree vector”
3. Select ASIN with highest “degree” in each cluster

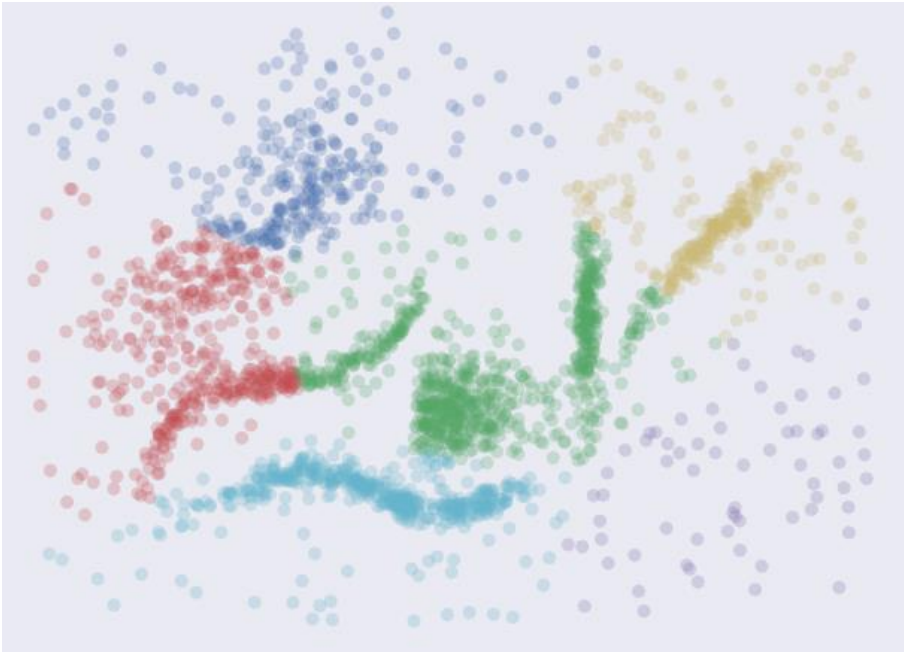


AMAZON CONFIDENTIAL

Clump outliers – Exemplar Selection

Choosing a set of “representative” ASINs:

1. Spectral clustering using affinity matrix
2. Compute the “degree vector”
3. Select ASIN with highest “degree” in each cluster



Obtain judgments for selected ASINs

AMAZON CONFIDENTIAL

Clump outliers – Judgment Propagation

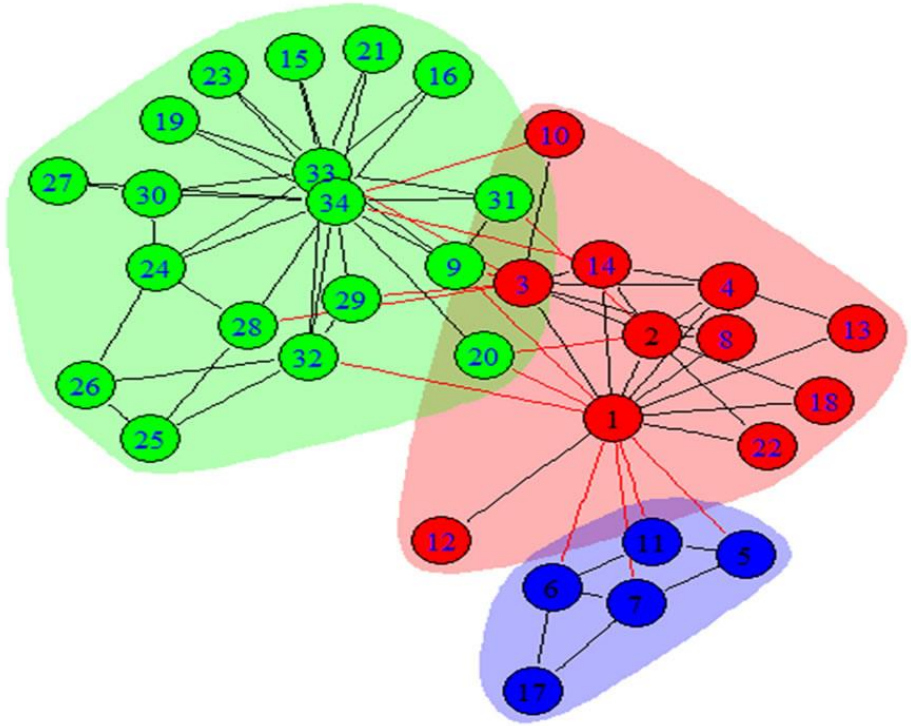
Propagating expert judgments:

1. Construct $S = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$

2. Iterate $F(t+1) = \alpha SF(t) + (1-\alpha)Y$ until convergence

   (for α between 0 and 1)

3. Obtain labels for ASINs from the limit of $F(t)$ by applying thresholds



Clump outliers – Results

Dataset	Total ASINs	Negative exemplars	Negative Propagations	Miscat Precision	Miscat Recall
TVs	19,345	20/50	5,117	98%	77%
Storage Benches	2,123	19/20	1,647	100% (est.)	-
Radar Detectors	15,225	13/19	9,427	100% (est.)	-
Cooktops	22,266	40/50	19,263	100% (est.)	-
Dart Cases & Wallets	28,992	98/100	27,151	100% (est.)	-
50 UK categories	1,748,936	2,372/5000	818,176	95.8% (audit)	-

AMAZON CONFIDENTIAL

Towards a Pristine Catalog

Questions?