[m]clusters

AUDIENCES FIRST

Copenhagen, Denmark

GroupM Mentor: Kristjan Brødreskift

Capstone Project By

Will Fein and Gerard Woytash

January - August, 2018

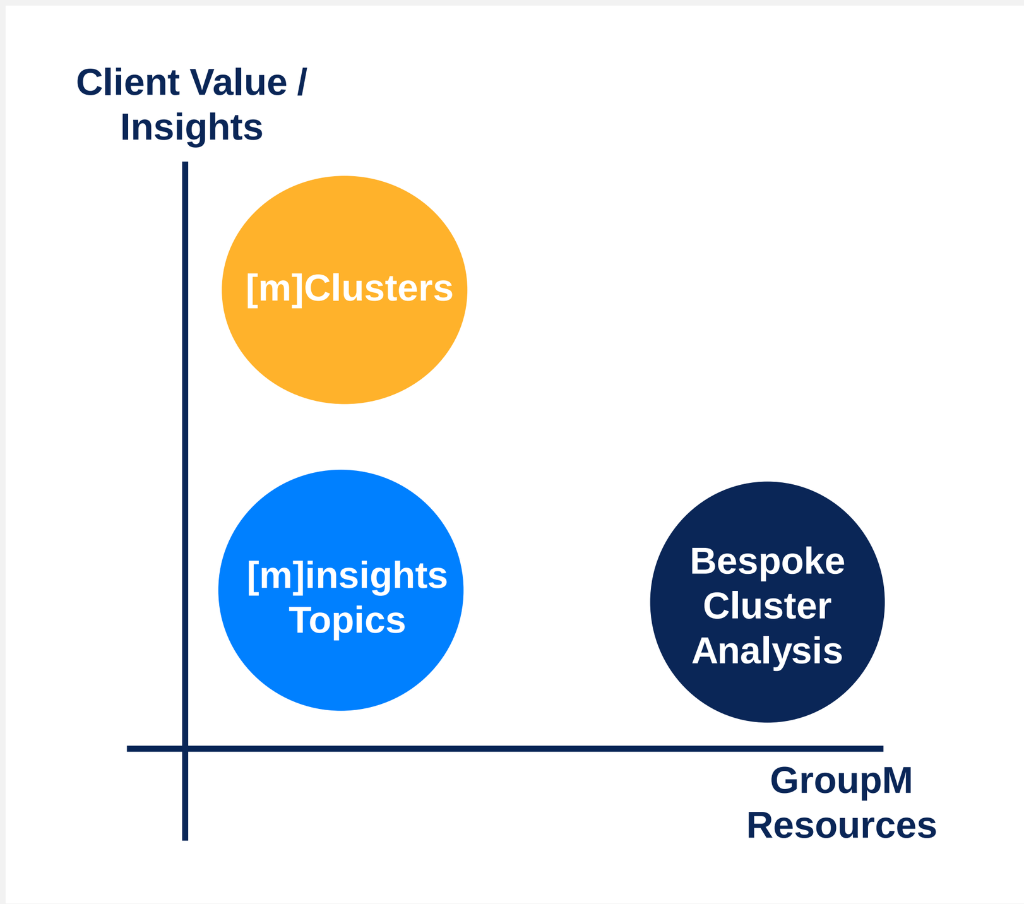
MIT Advisor: Karen Zheng

Who is GroupM?

GroupM is a leading global Media Agency. Advertising agencies make ads, but Media Agencies *place* them, and as online advertising and personalization increasingly dominate the marketplace, ad placement becomes more and more important. GroupM is focused on showing the right ads to the right people at the right time, and to this end has become a leader in data-driven solutions.

What are [m]Clusters?

[m]Clusters are segments of the population that are defined by particular online behaviors. GroupM clients can score their website's visitors against these segments, and can also target these segments for future advertising.



GroupM's Proprietary Data

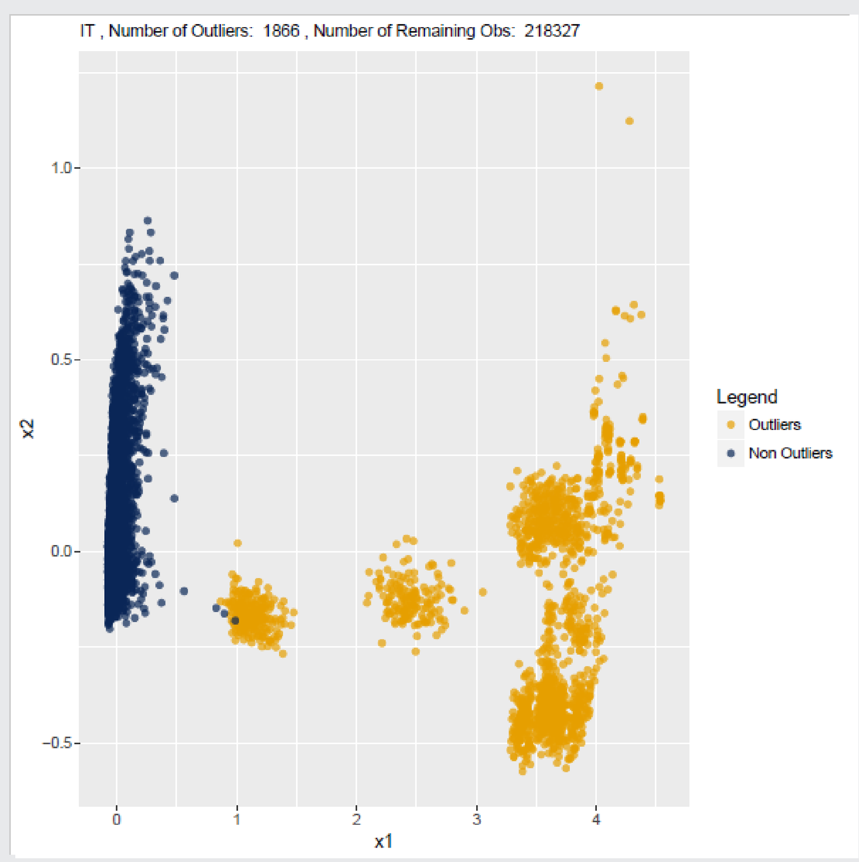
GroupM gave us access to their extraordinary and proprietary user interest data, among the most comprehensive in the world. As GroupM participates in display ad auctions, they record site visits for nearly all online users and nearly all webpages. These webpages are classified by a semantic engine, and then the counts of visits are converted to binary tables specifying whether a given user is "interested in" a given webpage type. It was these behavior datasets – and no additional demographic data – that we used to make our behavior-based segmentation.

visitor_id	Accessories	Adult_Education	Adventure_Travel	Advertising	Africa	Agriculture
ID_103[redacted]	0	1	0	0	0	0
ID_103[redacted]	1	0	0	0	0	0
ID_103[redacted]	1	1	0	1	0	0

PROCESS

Automated Data Cleaning

Our datasets had *bad users*, which we can either attribute to online bots or to failures of the semantic engine. Including these *bad users* hurt our clustering results. To make our entire process replicable, however, we couldn't leave behind any steps that centered around our ability to find and judge outliers. Our automated data cleaning notebook uses drastic dimensionality reduction with MCA and DBScan clustering to automatically detect outliers.



Dimensionality Reduction

High-dimensional data and binary data are both poorly suited to clustering analysis. The first suffers from the curse of dimensionality, and the latter makes distance calculations difficult. So, we used PCA to compress the user interest data and convert binary interests into continuous features. The curse of dimensionality – the notion that distance measurements converge in high dimensions – is not just a theoretical problem. Its business consequence is too many users placed in a 'leftover cluster.'

Ask Me About…

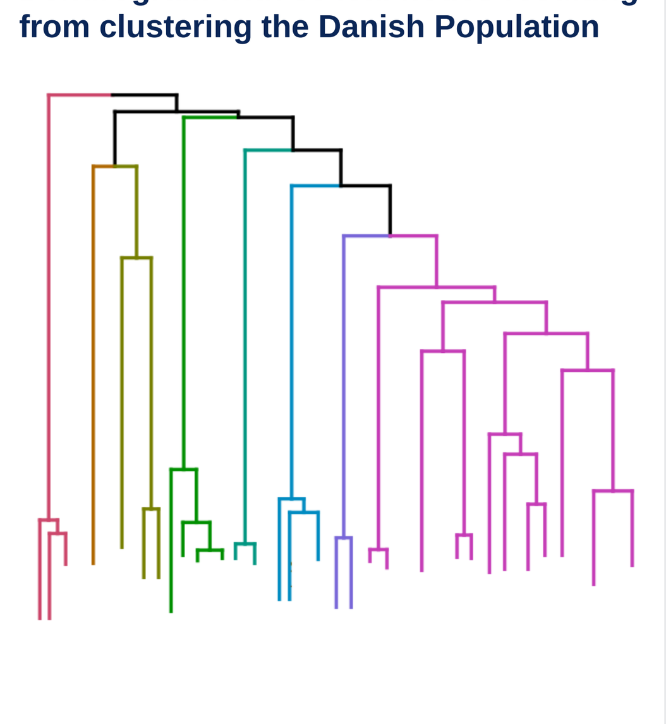
Things we'd love to chat about that didn't make this poster:

- Statistics for evaluating clustering algorithms
- Synthesizing user interests datasets from recoverable truth
- Dimensionality reduction with binary data
- Finding pre-determined clusters with supervised learning
- Ad buying and levels of personalization
- GDPR and demographic user data
- Copenhagen, Denmark or the Nordic Region!

Clustering

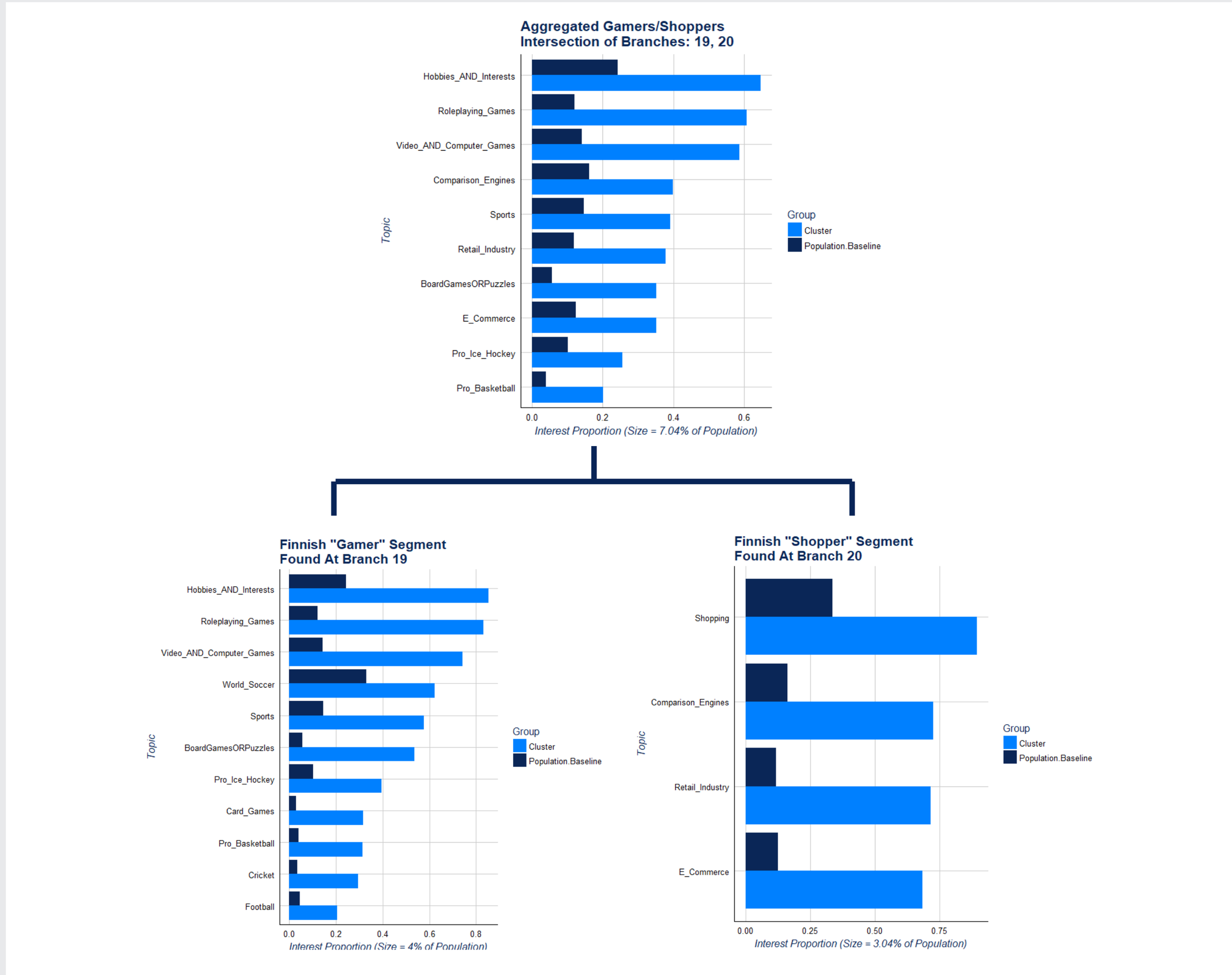
The next was to perform agglomerative hierarchical clustering with linkage determined by Ward's Method. We kept the first 30 branches of these trees.

Dendrogram with 30 leaf nodes resulting from clustering the Danish Population



Tree Pruning

"Choosing K" is always a difficult step in clustering, and we included business experts for this step of the process. In this business context, it's a question of aggregation. Which splits of the tree separated distinguishable audiences, and which split a single audience into two? The importance of this question centers on our ability to sell the segments to GroupM client managers, and on their ability to sell the segments to their clients. So, we asked local experts in each country to help us choose the level of aggregation for their market.



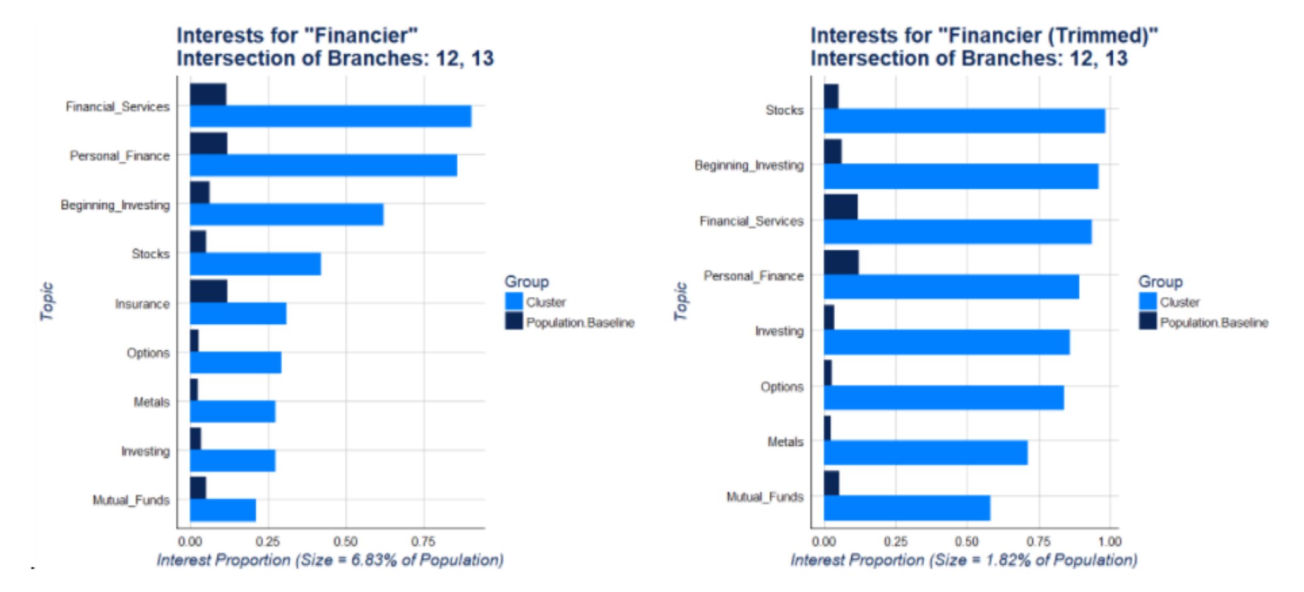
INTEGRATION & RESULTS

Final Segments

Audience Segment	Denmark	Sweden	Norway	Finland
The Attentive Parent	✓		✓	
The Car Aficionado	✓	✓	✓	✓
The Car Buyer	✓	✓		✓
The Do-It-Yourselfer		✓	✓	✓
The Driven Professional			✓	✓
The Engaged Citizen	✓	✓	✓	✓
The Fashionista	✓	✓		
The Financier	✓	✓	✓	✓
The Foodie	✓	✓	✓	✓
The Gamer	✓	✓	✓	✓
The Health Enthusiast		✓	✓	
The Interior Decorator		✓	✓	✓
The Mainstream Media Consumer	✓	✓		✓
The New Parent	✓	✓		
The News Junkie		✓	✓	✓
The Office Computer	✓	✓	✓	✓
The Online Shopper	✓			✓
The Real Estate Buyer	✓	✓		✓
The Sports Fan	✓	✓	✓	✓
The Student	✓	✓	✓	✓
The Stylish Parent	✓	✓	✓	✓
The Traveler	✓		✓	✓

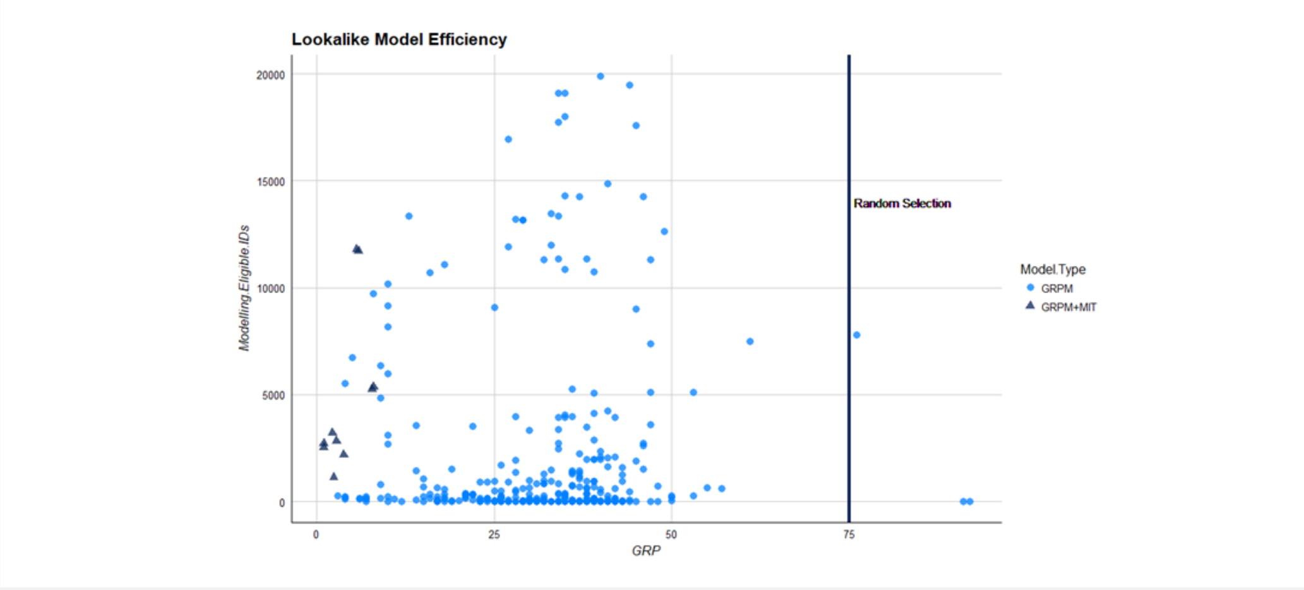
Trimming Clusters

When it came time to upload our "segments," we had to define each segment in the GroupM platform by passing in a list of user ids. We chose these user ids carefully, building algorithms that essentially trim the clusters and amplify the signals.



Lookalike Results

One metric of important to GroupM is the ability of their bidding and insights engine to identify the right users from the population to match a segment. They evaluate this with a statistic called GRP, which measures the efficiency of the lookalike models. It can be thought of as similar to recall, except that smaller GRP numbers mean that the models are more successful. MIT segments outperformed nearly all benchmark models.



Buying Trial Results

We activated our segmentation for three different clients, who each selected certain segments to target. These selections were based on segments "over-indexed" for their current customers, and the results shown below demonstrate the monetary value of [m]Clusters.

Client Industry	Buying Strategy	Click Through Rate	Conversion Rate	Cost Per Conversion	Impressions
Automotive	Broad Strategy	1%	NA	1	>345,000
	[m]Cluster Activation	1.25%	NA	0.94	>16,000
Telecom	Broad Strategy	1%	1%	1	>398,000
	[m]Cluster Activation	1.36%	2.07%	0.77	>202,000
Amusement Park	Broad Strategy	1%	0.01%	1	>24,000
	[m]Cluster Activation	1.01%	9.84%	0.09	>2,400

Neural Networks for Cluster Scoring and latent Sparse Dimensions

Neural networks allowed us to find sparse latent dimensions that could explain the hierarchical clustering tree in the original feature space and define the segments with few, shared dimensions. They also allowed to score our segments on how easy they were to distinguish from the population and from other segments. Our results were almost uniformly excellent (precision and recall > 0.9), and where they were not they were informative. In Finland, it seems our Interior Decorator and Real Estate Buyer segments may contain many similar individuals.