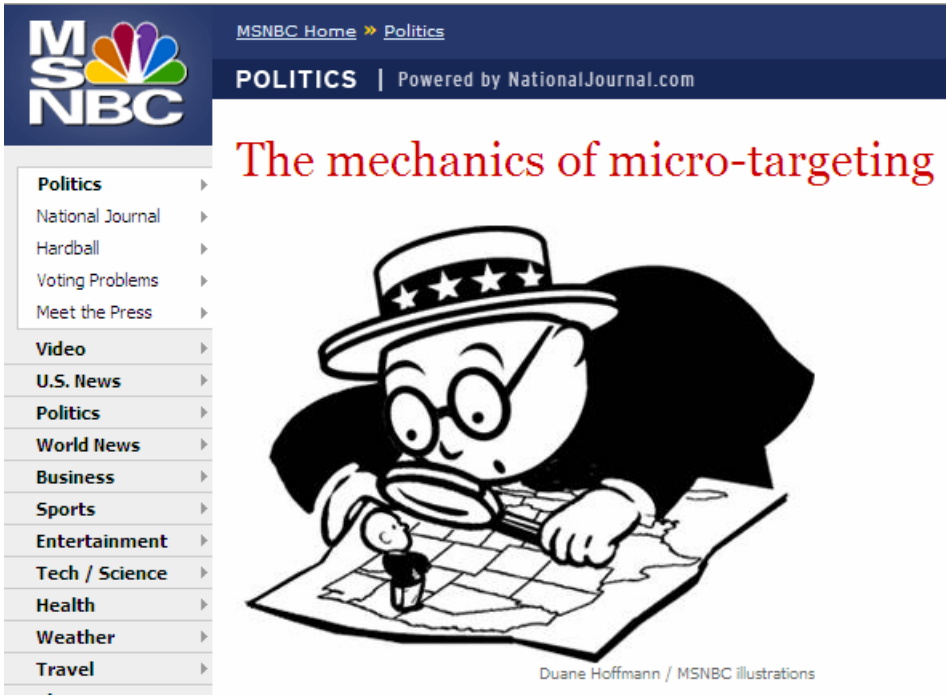
Segmenting Customer Bases in Personalization Applications Using Direct Grouping and Micro-Targeting Approaches

Alexander Tuzhilin
Stern School of Business
New York University
(joint work with Tianyi Jiang)

Personalized Election Campaigns



Targeted messages to specific subgroups of voters, e.g. middle-income Democratic blue-collar gun owners in rural Michigan (T. Curry, MSNBC, 10/2006)

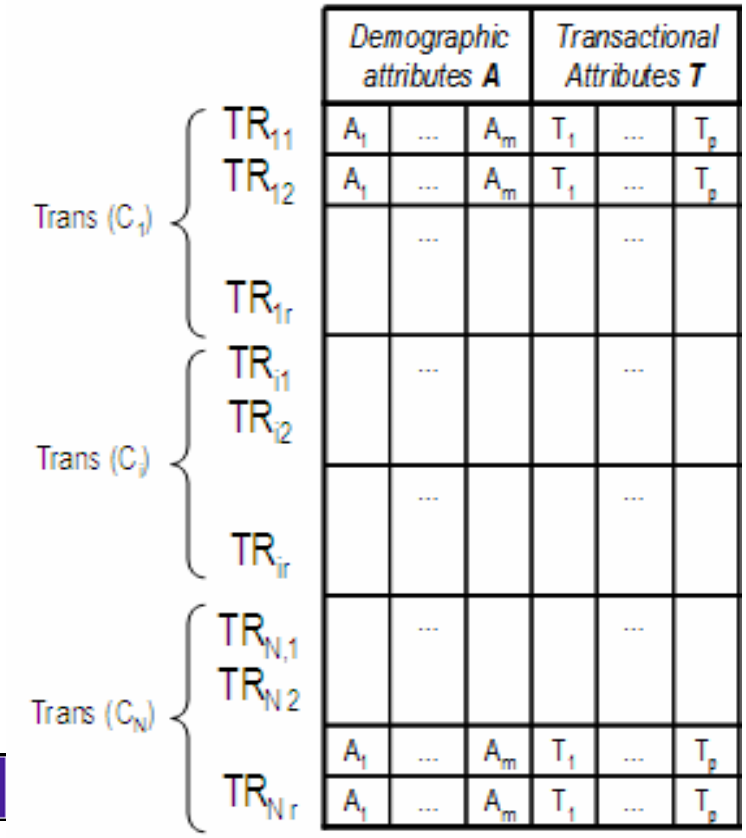
Question: how to do this?

Research Questions:

- How to effectively partition the customer base into non-overlapping segments to achieve better targeting of customers?
- Degree of personalization: What is the unit of analysis? How refined should customer models/segments be?
  - Individual (1-to-1) vs. Segments
  - How small the segments should be?
- Is there a general trend to the distribution of "ideal segment" sizes across the customer population?

Problem Formulation

- Group of *N* customers $p = (C_1, ..., C_N)$ and their data (below)
- Build a model $M: Y = f(X_1, \ldots, X_k)$ on that data for customers *p*
- Measure performance *f(p)* of model *M* using *fitness function f.*



Customer C_i is defined by

- demographic attributes $A = \{A_1, \ldots, A_m\}$
- transactional attributes $T = \{T_1, \ldots, T_p\}$
- transactions $Trans(C_i) = \{TR_{i1}, \ldots, TR_{ir}\}$ performed by customer C_i

Problem Formulation: Example

- Model M_i is a **decision tree** built on data p of customers C_1, ..., C_N
  - Y is buy/not buy; X_1, …, X_k are some demographic and transactional variables
- Examples of fitness function f
  - Percentage of correctly classified instances on the out-of-sample data
  - Relative absolute error on out-of-sample data
  - Area under the ROC curve

Optimal Customer Segmentation (OCS) Problem

Partition customer base *C* of *N* customers into into **mutually exclusive collectively exhaustive** groups $P = \{p_1,...,p_k\}$, s.t., the models M_i built on groups p_i would collectively produce the best performance for the fitness function $f(p_i)$ taken over $p_1,...,p_k$, i.e.,

$$\theta = \sum_{i=1}^{k} \alpha_i \times f(p_i)$$

is maximized over all possible partitions *P*, where α_i is a weighting measure of "importance" of segment *i*.

Note: Focus on *prediction* problems

OCS is an Intractable Problem

Proposition: Optimal Customer Segmentation problem is NP-hard

Solution: find *suboptimal tractable* customer segmentation methods

Related Work

- Customer segmentation and clustering in marketing research
- Data mining research on customer segmentation

Related Work: Marketing Research

- clustering techniques
- mixture models
- (generalized) mixture regression models and continuous mixture distributions (Wedel et al. 2000)

Related Work: Data Mining Research

- Shopping basket clustering (Brijs et al. 2001)
- Hierarchical clustering (Jiang et al. 2006)
- Pattern-Based Clustering (Yang et al. 2003)

Suboptimal Solutions to OCS

- Statistics-based segmentation
- Direct Grouping methods
- 1-to-1 (segments of one customer)

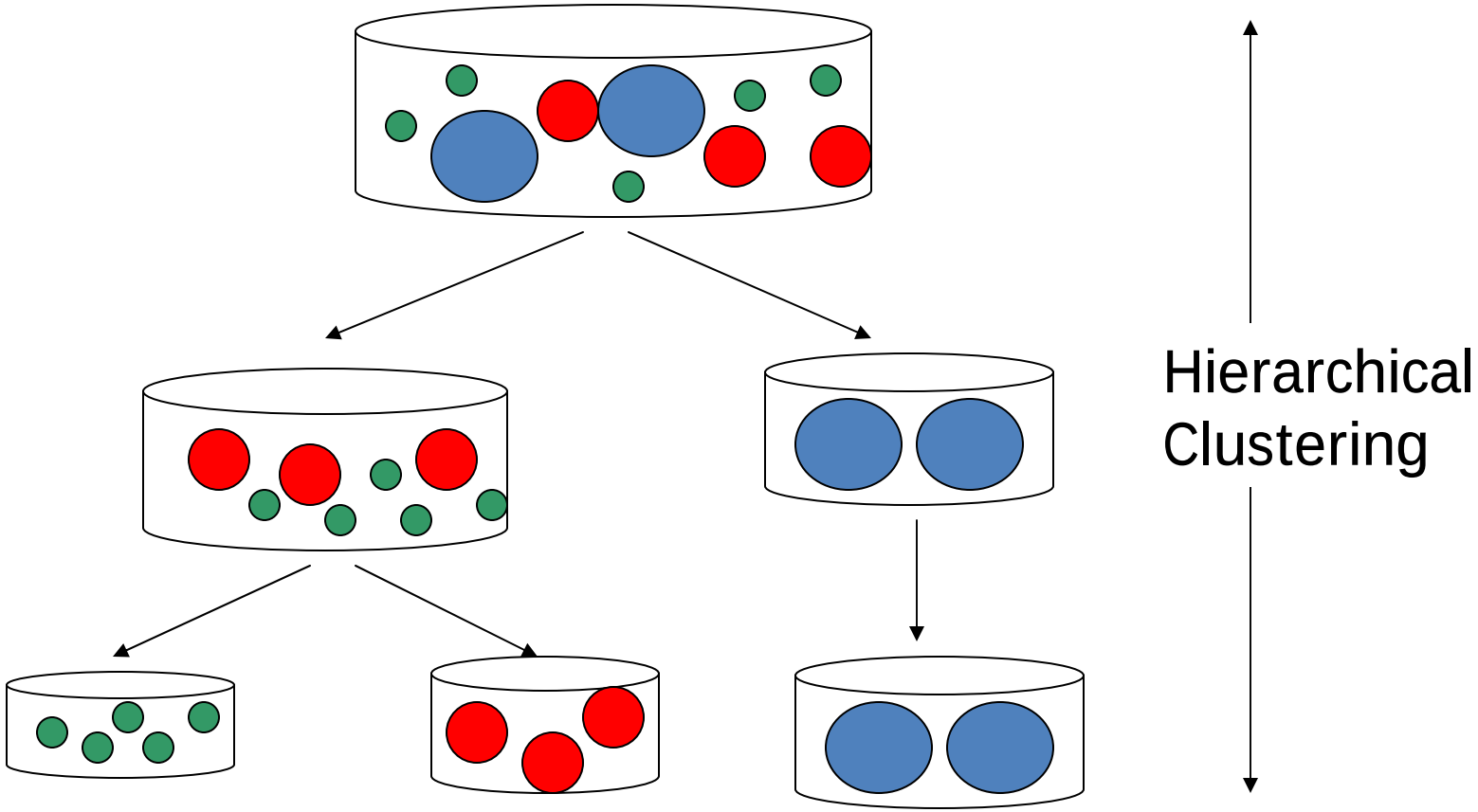
Statistics-Based Segmentation

Group customers by

- computing ***summary statistics*** from customers' demographic and transactional data, e.g.,
  - RFM statistics, avg time spent on a web site, max transaction
- mapping these statistics into points in an *n*-dimensional space
  - one customer – one point
- grouping these points (customers) into clusters (segments)
  - by applying various *clustering* algorithms

Hierarchical Clustering (**HC**) Segmentation

- Dots: points in the R^n space of computed statistics
- Used FarthestFirst method (beats others [JT06])



Other Statistics-Based Methods

Affinity Propagation (AP) [Frey, Dueck, NIPS 2006]:

- like K-means, but uses *exemplars (single customers)* as cluster centers during clustering iterations

Entropy Clustering (EC)

- uses decision trees (C4.5) to form segments based on the labels of dependent variable Y.
- customers grouped it terms of summary statistics to reduce the entropy of the class label

- Experimentally showed that ***HC*** dominates ***AP*** and ***EC***
- Therefore, will focus subsequently on ***HC***

Direct Grouping Methods

Group customers into segments by

- directly combining their (transactional) data
- building predictive model on this data for the group
- measuring the overall fitness score as a linear combination of fitness scores of individual segments

Direct Grouping Methods

Iterative Merge (IM):

- Start with segments containing individual customers
- Iteratively merge two segments Seg_A and Seg_B when
  - predictive model on the *combined* segment $Seg_A \cup Seg_B$ is better than *individual* models on segments Seg_A and Seg_B
  - combining Seg_A with any other existing segments results in a worse performance than the model on $Seg_A \cup Seg_B$.
- Terminate when a fixpoint or a performance threshold is reached

Other Direct Grouping Methods

Iterative Growth (IG):

• Bottom up segmentation approach that adds customers to existing segments one at a time

Iterative Reduction (IR):

• Top down segmentation approach that subtracts customers from existing segments one at a time

• Experimentally showed that ***IM*** dominates ***IG*** and ***IR***
• Therefore, will focus subsequently on ***IM***

1-to-1 Method

Builds customer segments of size 1 (individual models of customers) by learning them *only* from the data pertaining to individual customers

- Needs to have a minimal number of transactions (e.g., 10) per customer

Comparing 1-to-1, Statistics-Based and Direct Grouping Segmentation Methods

Take best-of-breeds in each category

- Statistics-Based: Hierarchical Clustering (HC)
  - with FartherstFirst clustering
- Direct Grouping: Iterative Merge (IM)
- 1-to-1: the standard one

- Do experimental comparison among them
  - Determine which one is better and when

Experimental Settings

Compare these approaches across the following dimensions:

- Types of datasets (ComScore, Nielsen, Synthetic data)
- Types of customers (high vs. low-volume)
- Types of predictive models
  - E.g., classifiers J48 & Naïve Bayes
- Dependent variables
  - 3 variables per dataset
- Performance Measures
  - Root Mean Squared Error – RME
  - Relative Absolute Error – RAE
  - Correctly Classified Instances - CCI

Data Sets: Panel Data

DataSet	Customer Type	% of Total Population	Families	Total Transactions	Average Transactions Per Household
ComScore	High	2.2%	1,000	137,157	92
ComScore	Low	2.2%	1,000	24,344	10
Nielsen	High	32%	500	28,985	136
Nielsen	Low	32%	500	5,007	46
Syn-High	High	100%	1,024	102,400	100
Syn-Low	Low	100%	1,024	10,240	10

Statistical Significance

We apply the **Mann-Whitney rank test** to compare any two *performance distributions*.

The null hypothesis for comparing distributions generated by methods A and B for a performance measure is:

(I) H_0: The distribution of a performance measure generated by method A ***is not*** different from the distribution of the performance measure generated by method B.

H_1+: The distribution of a performance measure generated by method A ***is*** different from the distribution of the performance measure generated by method B in the ***positive*** direction.

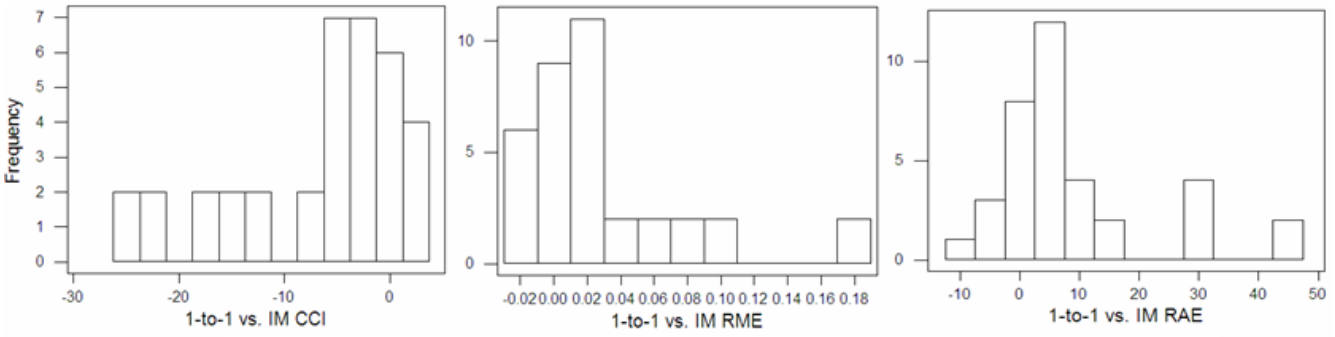
H_1-: The distribution of a performance measure generated by method A ***is*** different from the distribution of the performance measure generated by method B in the ***negative*** direction.

Empirical Results: 1-to-1 vs. Direct Grouping vs. Statistics-Based Methods

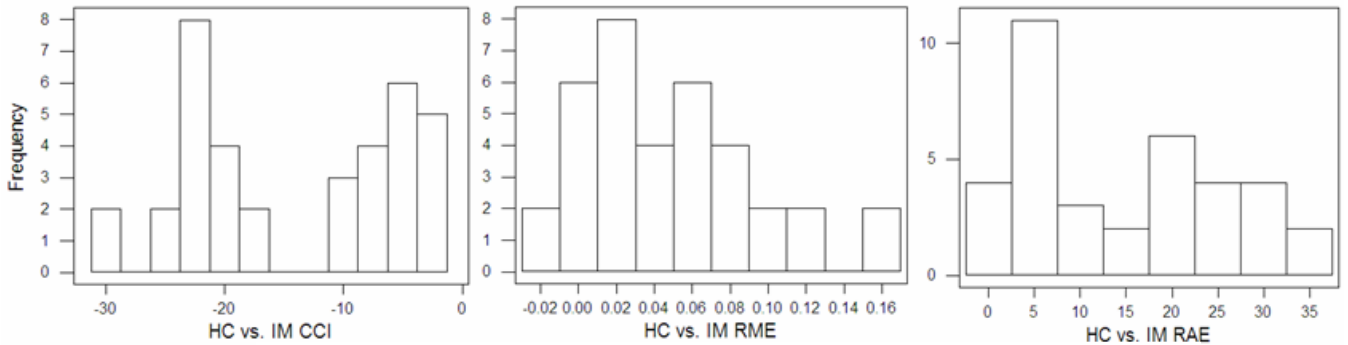
Method	HC		IM	
	H_1+	H_1-	H_1+	H_1-
1-to-1	76	29	21	86
HC	-	-	3	104

- 108 significance tests per methods comparison pair
- Numbers in columns H_1+ and H_1- indicate the number of statistical tests that *reject* hypothesis H_0.

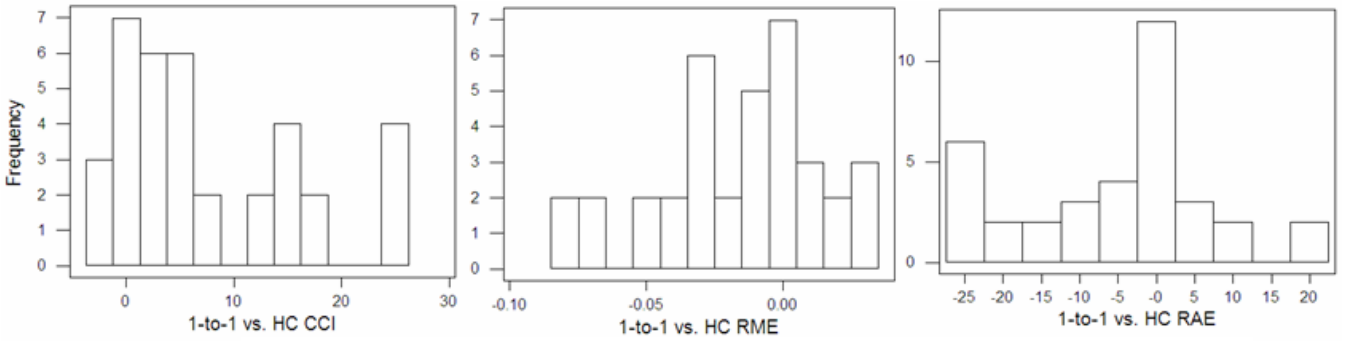
Median Difference Distributions



Median Difference Distributions of 1-to-1 vs. IM



Median Difference Distributions of HC vs. IM



Median Difference Distributions of 1-to-1 vs. HC

Empirical Results

Comparing Direct Grouping, Statistics-Based, & 1-to-1 Methods Among *High-Volume* Customers

Method	HC		IM	
	H_1+	H_1-	H_1+	H_1-
1-to-1	43	11	19	34
HC	-	-	1	53

Total number of tests is 54

Empirical Results

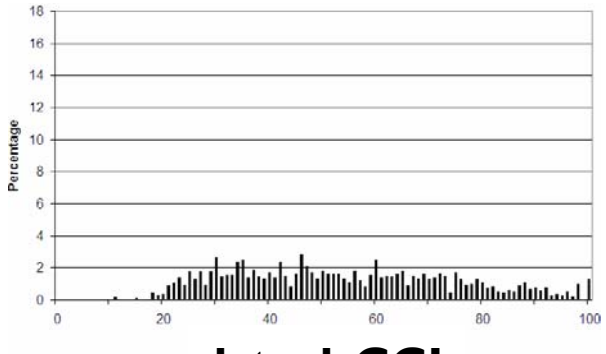
Comparing Direct Grouping, Statistics-Based, & 1-to-1 Methods Among *Low-Volume* Customers

Method	HC		IM	
	H_1+	H_1-	H_1+	H_1-
1-to-1	33	18	2	52
HC	-	-	2	51

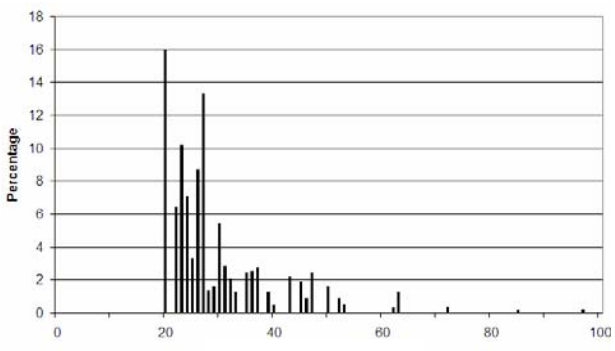
Total number of tests is 54

Empirical Results: Sample CCI scores

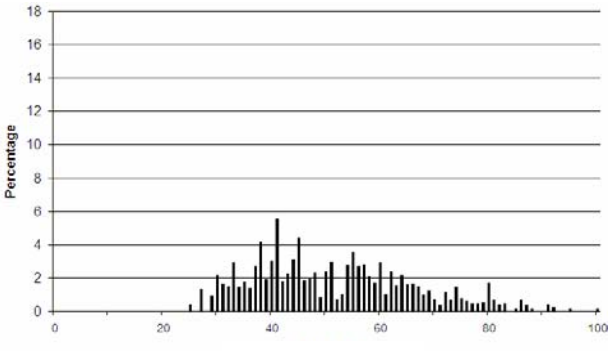
(for Day-of-the-Week predictions across High-Volume ComScore Customers)



1-to-1 CCI



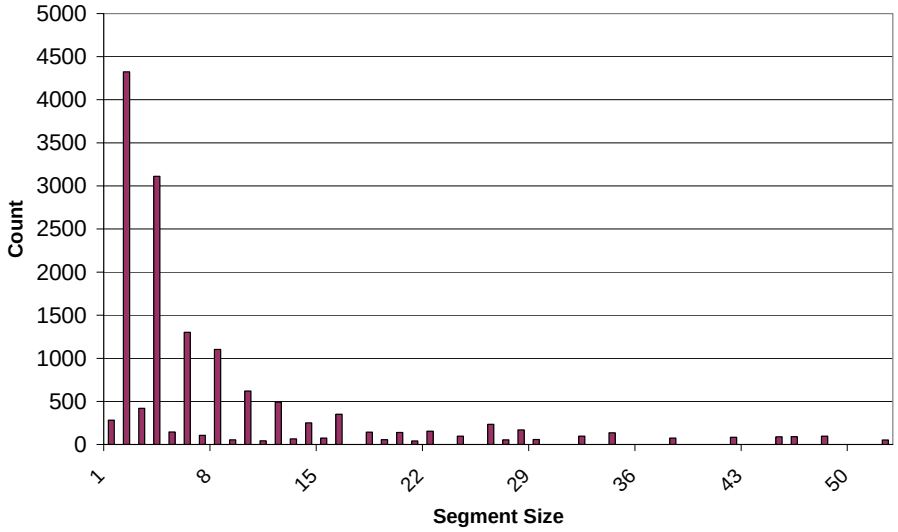
HC CCI



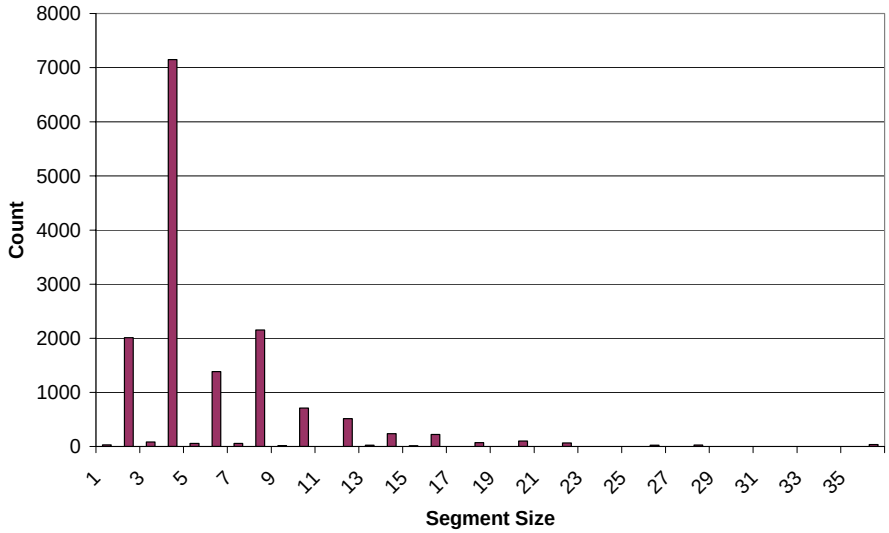
IM CCI

Empirical Results

Segment Size Distribution Generated by IM



High-Volume Datasets



Low-Volume Datasets

Power Law Distribution

Many natural phenomena obeys Power law

(where P_n is the frequency of occurrence of a segment of size n): $\log(P_n) \approx -a \log(n)$

For segment size distributions generated by IM, starting from the highest count:

Customer Type	Starting Segment Size	*a*	*p-value*
High	2	0.828	<0.001
Low	4	1.67	<0.01

Research Implications

- Direct grouping produces better segmentation results than traditional clustering-based and 1-to-1 methods
- Segment size distributions from direct grouping approaches follows the Power Law distribution
- The above results add support for the micro-segmentation approaches to personalization

Micro Targeting Extensions

Product Types × *Customer Matrix*

(√ stands for purchasing product type by customer)

	Customer 1	Customer 2	…	Customer N
Product $Type_1$		√		
Product $Type_2$	√		√	
…	…	…	…	…
Product $Type_L$	√			√

Micro Targeting Method

Iterative Merge Products (IM_Prod):

• Start with segments containing individual customer's specific product type transaction data

• Bootstrap operation to merge small segments based on K-nearest neighbors of customer's product type and demographic summary statistics vectors

• Run ***IM*** but with the unit of analysis being (customers, product types) segments

Empirical Comparisons of Different Approaches

Comparing Three Segmentation Approaches:

- Statistics based (e.g. HC, AP)
- Direct grouping based (e.g. IM)
- Micro Targeting based (IM_Prod)

Across five dimensions of different

- Types of datasets (ComScore, Nielsen, Synthetic data)
- Types of customers (high vs. low-volume)
- Types of predictive models (classifiers J48 & Naïve Bayes)
- Dependent variables (3 variables per dataset)
- Performance Measures
  - Root Mean Squared Error – RME
  - Relative Absolute Error – RAE
  - Correctly Classified Instances - CCI

Empirical Results: Compare All Methods

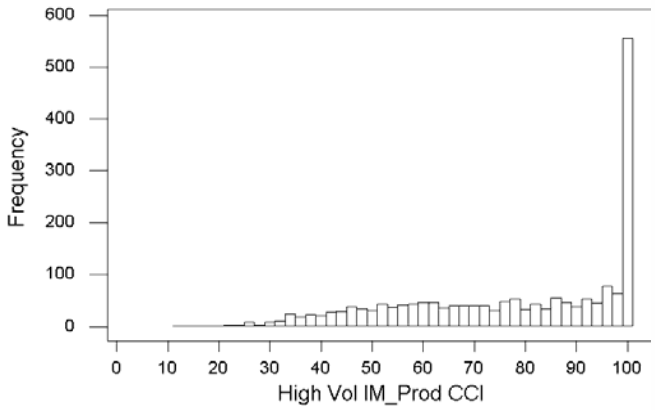
Methods	***HC***		***IM***		***IM_Prod***	
	H+	H-	H+	H-	H+	H-
AP	66	18	12	57	0	**108**
HC	-	-	6	90	0	**108**
IM_Prod	**108**	**0**	**96**	**0**	-	-

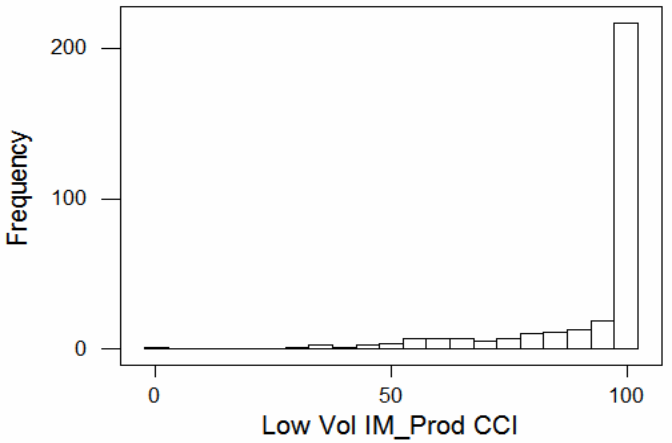
Total number of tests is 108

IM vs. IM_Prod: CCI Score Distributions

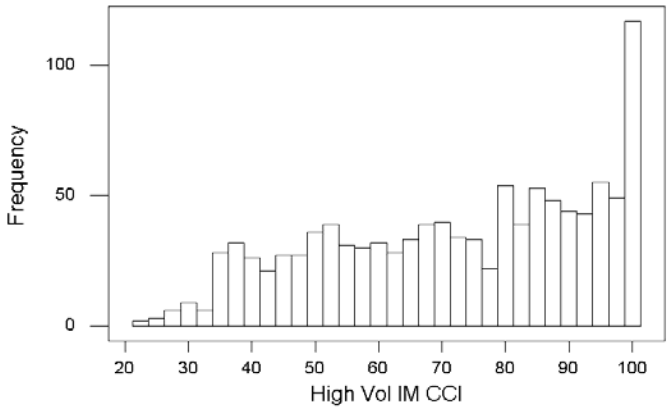
("Day of the Week" prediction for High & Low-Volume ComScore Customers)

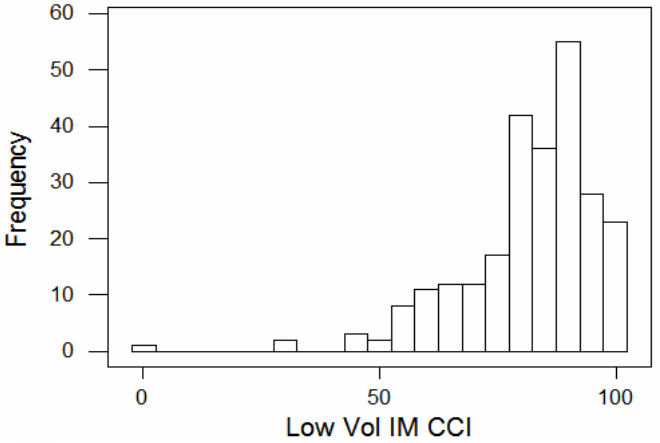
IM_Prod





IM



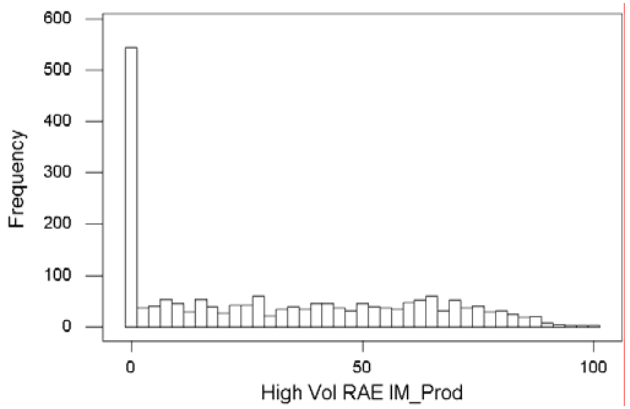


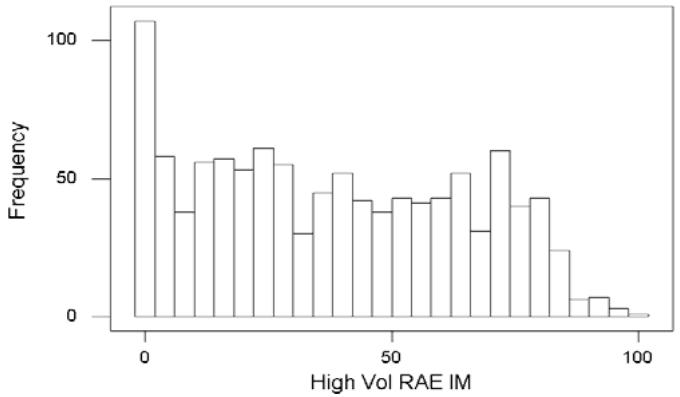
High-Volume Datasets **Low-Volume Datasets**

IM vs. IM_Prod: Error Distributions

("Day of the Week" prediction for High & Low-Volume ComScore Customers)

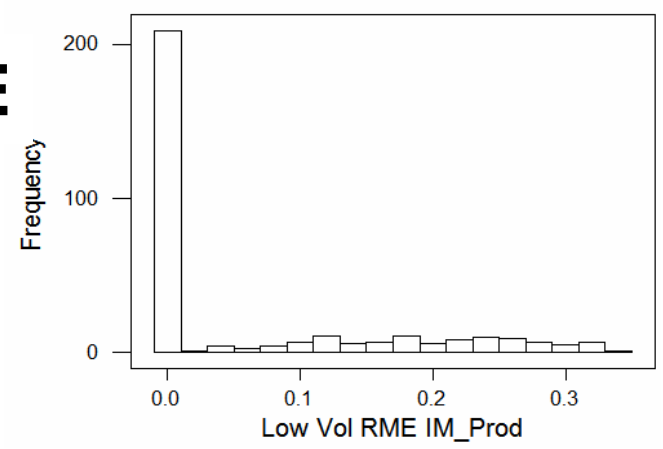
RAE

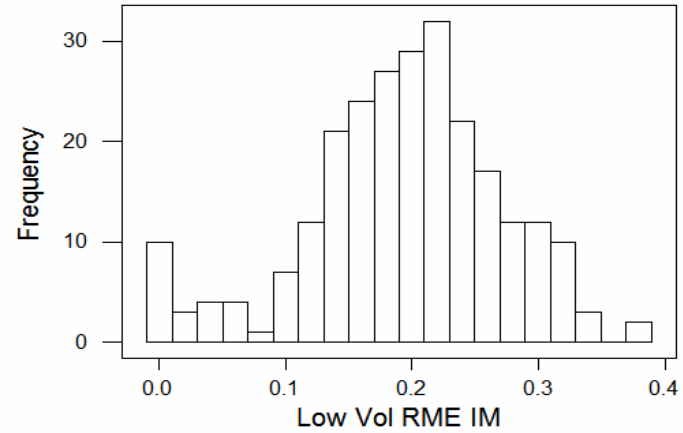




High Volume

RME



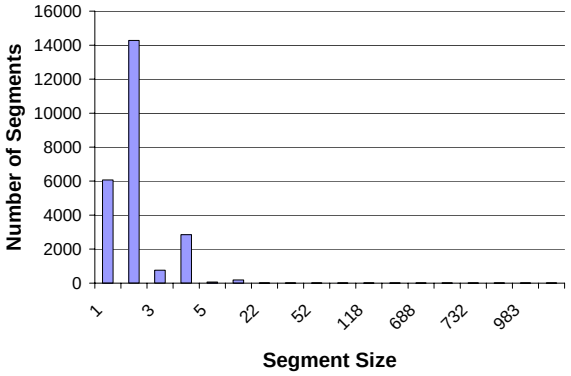


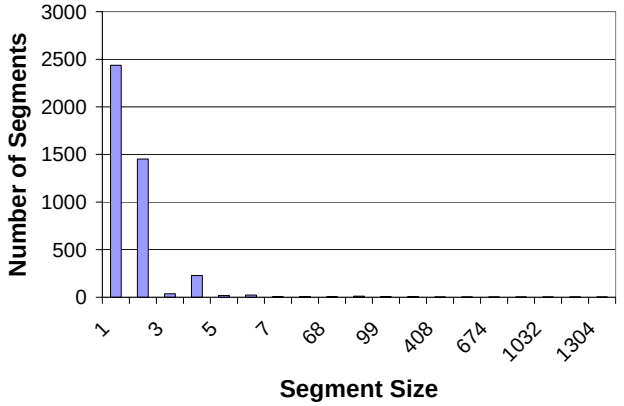
Low Volume

IM_Prod ***IM***

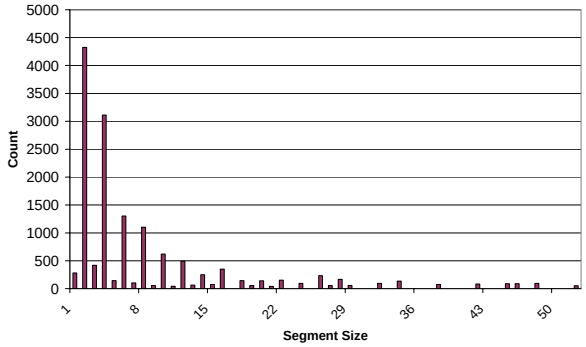
IM vs. IM_Prod: Segment Size Distributions

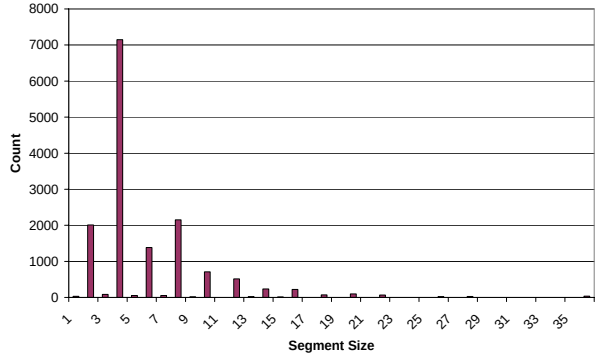
IM_Prod





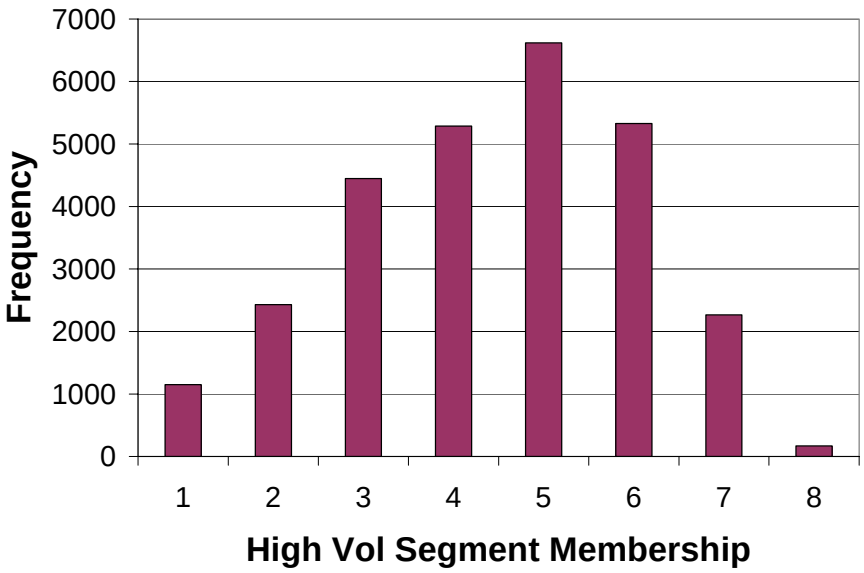
IM

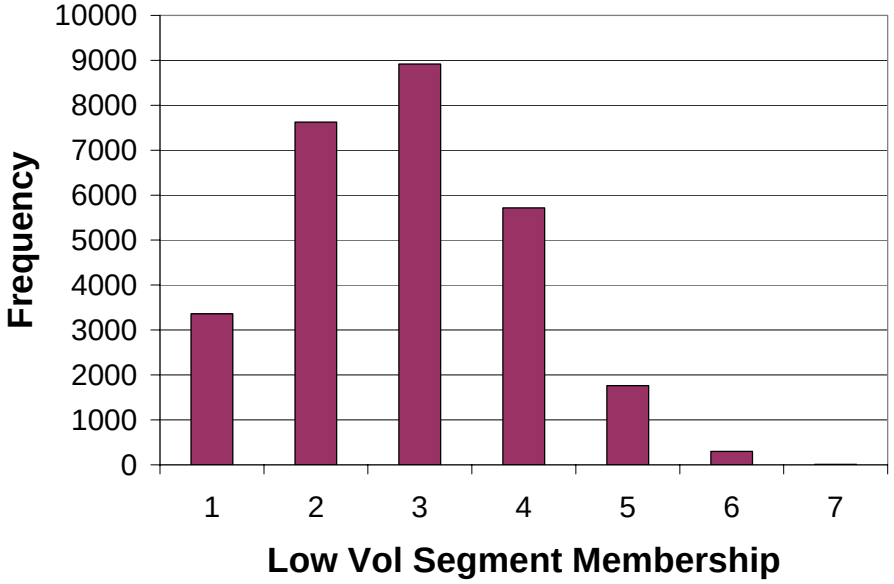




High-Volume Datasets **Low-Volume Datasets**

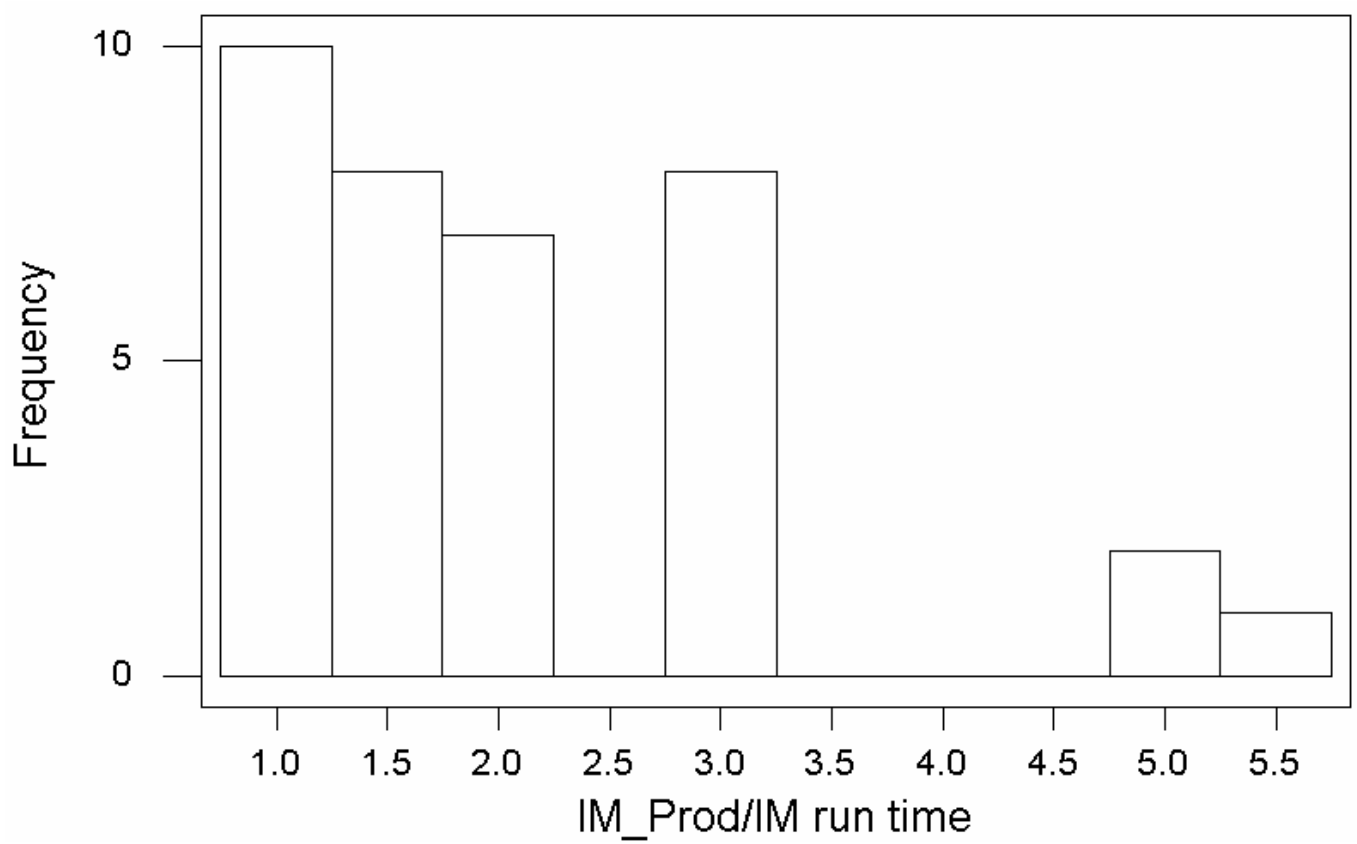
Customer Segment Membership Count Distribution





High-Volume Datasets **Low-Volume Datasets**

IM_Prod vs IM Computational Performance



Conclusions

- Partition customers based on ***micro targeting*** results in formation of "better" customer segmentations than traditional clustering based and fitness-based direct grouping approaches

- Micro targeting produces smaller segments than Direct Grouping methods

- The above results add support for micro-segmentation (partition based on both customer and product types) approaches to personalization

Future Research

- Need to gain more insight into the OCS problem:
  - Does the "Optimal" partition of customer base follow Power Law distribution?

- Investigate scalability and generalizability issues of our approach against different types of very large real world datasets and handling incremental or time series data

- Improve method not just based on predictive accuracy, but also in terms of the standard marketing oriented performance measures such as customer value, profitability and other economics based performance measures