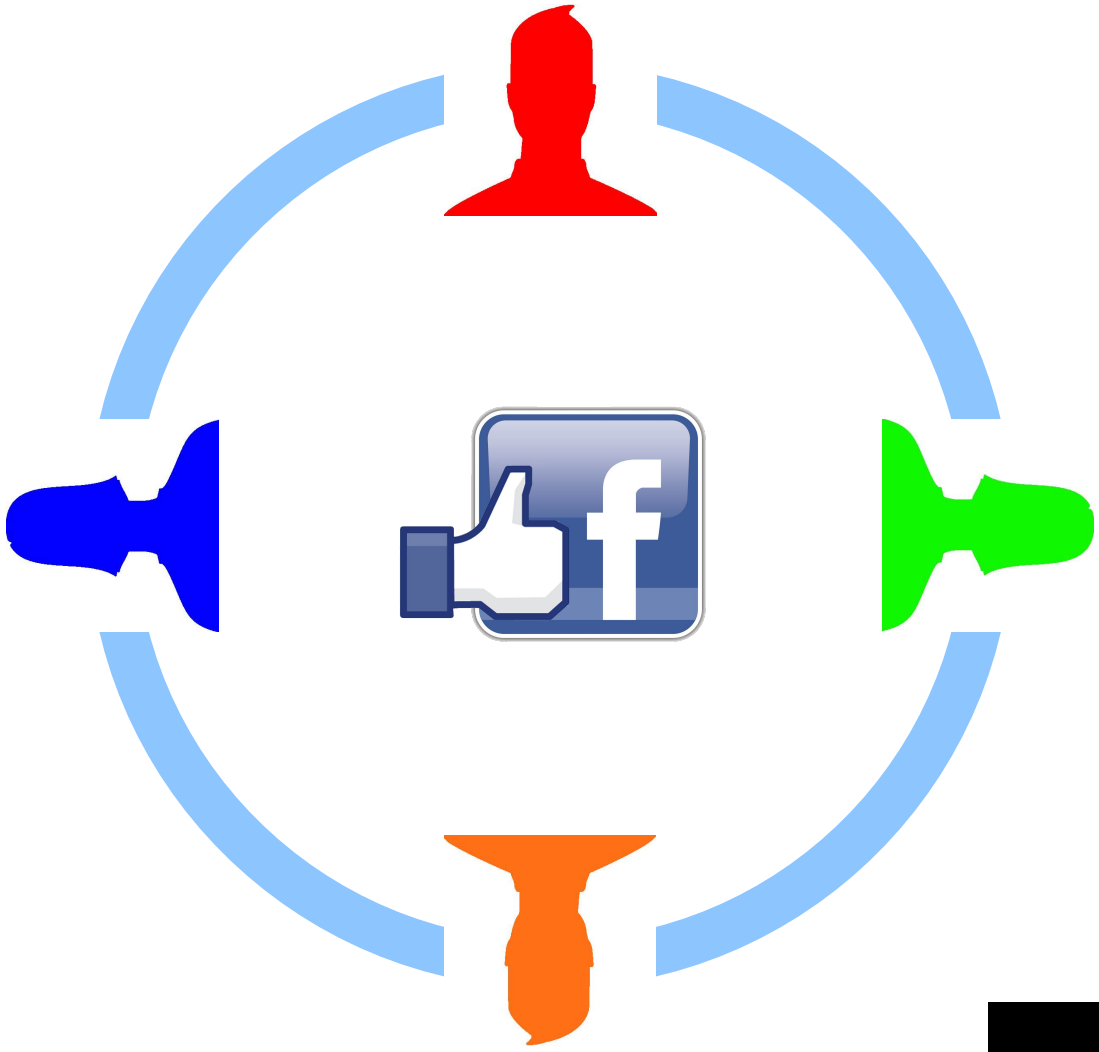
Who R U ?

InfoScout

Classifying Users Based on Their FaceBook Likes



[redacted]

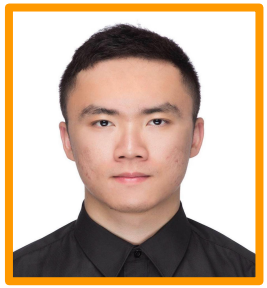
The Team: **The Lion Tamers**

InfoScout

Adi Dick-Charnilas
Research Lead

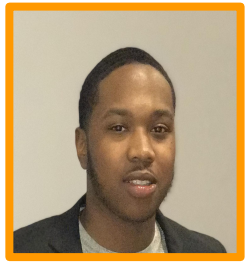
Qiao Yang
Design Lead

Shiqing Feng
Analytics Lead



Zening Wang
Business Analyst Lead

Adam Kirstein
Data lead



Eric Pridgen
Project Manager

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Timeline

InfoScout

Data exploration
Data cleaning, exploratory analysis & defining methodologies
June 5 - June 12

Research Proposal
Define our goals & methodology
June 16

Initial Pitch
Pitch presentation to InfoScout
June 22

Clustering
Refine our methodology and cluster model
June 22-July 26

Finalizing Results
Prepare an UI app, final analysis and presentation
July 26- August 8th

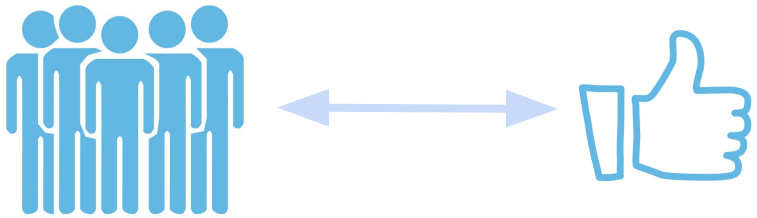
Final Presentation
Present final results and conclusions to InfoScout
August 9th

Project Description

InfoScout

Source Data:

- Facebook like data
- User Demographic data



Objective:

- Segment the user groups according to similar Facebook like history
- Determine relationships between the user clusters and their Facebook likes

Hypothesis:

By clustering the user profiles based on their demographic attributes and their liked Facebook pages, InfoScout can utilize the subsequent results to:

A. Predict consumer behavior
B. Identify consumer sectors likely to be drawn to a particular Facebook page

Cleansing The Data

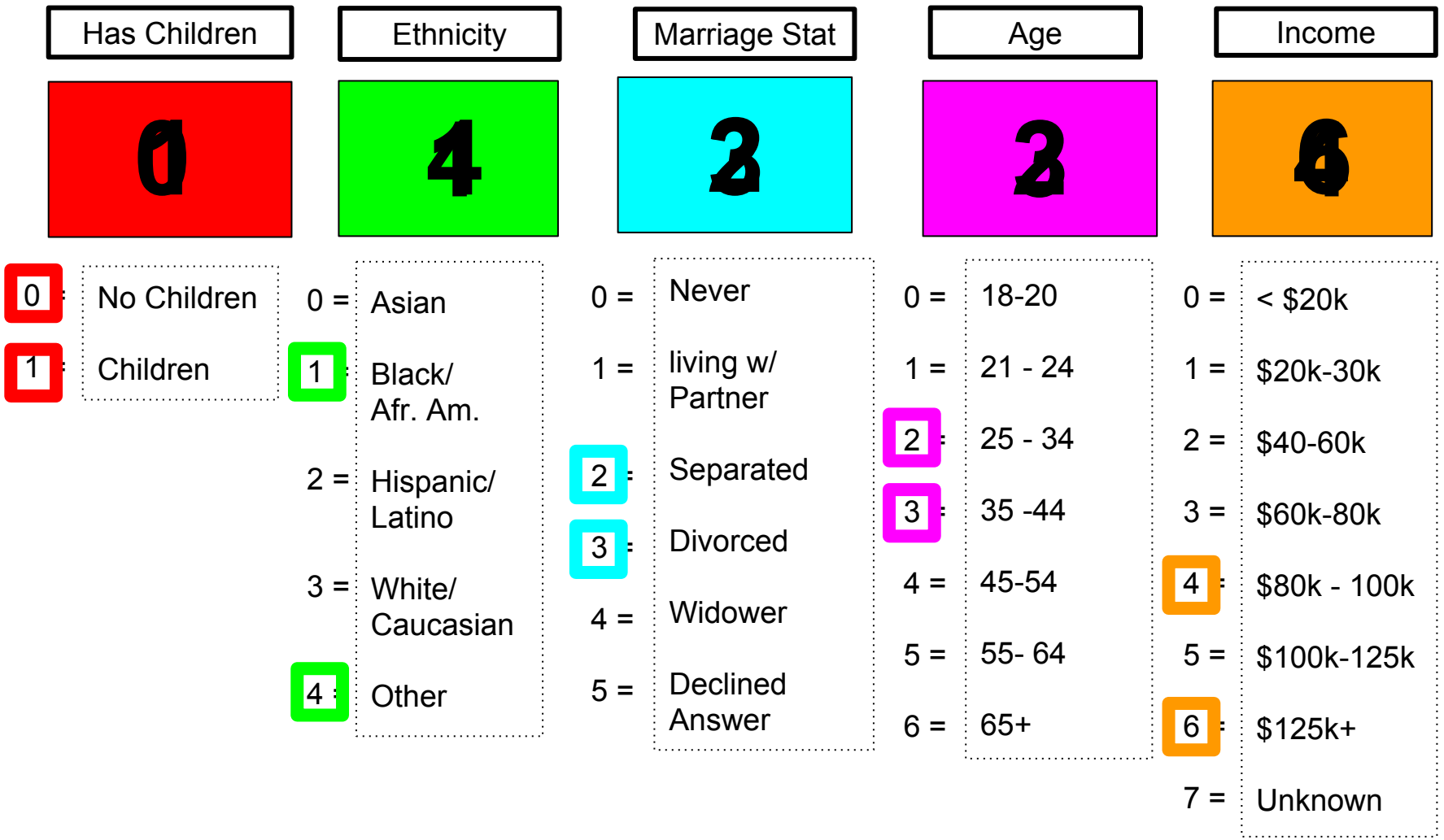
InfoScout

Major duplication problem based on Like I.D. Column

User ID	Category	Name	Like ID	Category List name
123456	Retail	Brand X	123	Furniture
123456	Retail	Brand X	124	Appliance
123456	Retail	Brand X	125	Electronics
123456	Retail	Brand X	126	Furniture
123456	Retail	Brand X	127	Appliance
111111	Retail	Brand X	128	Electronics
44444	Retail	Brand X	129	Furniture
44444	Retail	Brand X	130	Appliance
44444	Retail	Brand X	131	Electronics
111111	Retail	Brand X	132	Appliance
44444	Retail	Brand X	134	Electronics
44444	Retail	Brand X	135	Electronics
44444	Retail	Brand X	136	Furniture

Defining Profile Codes

InfoScout



Methodology

InfoScout

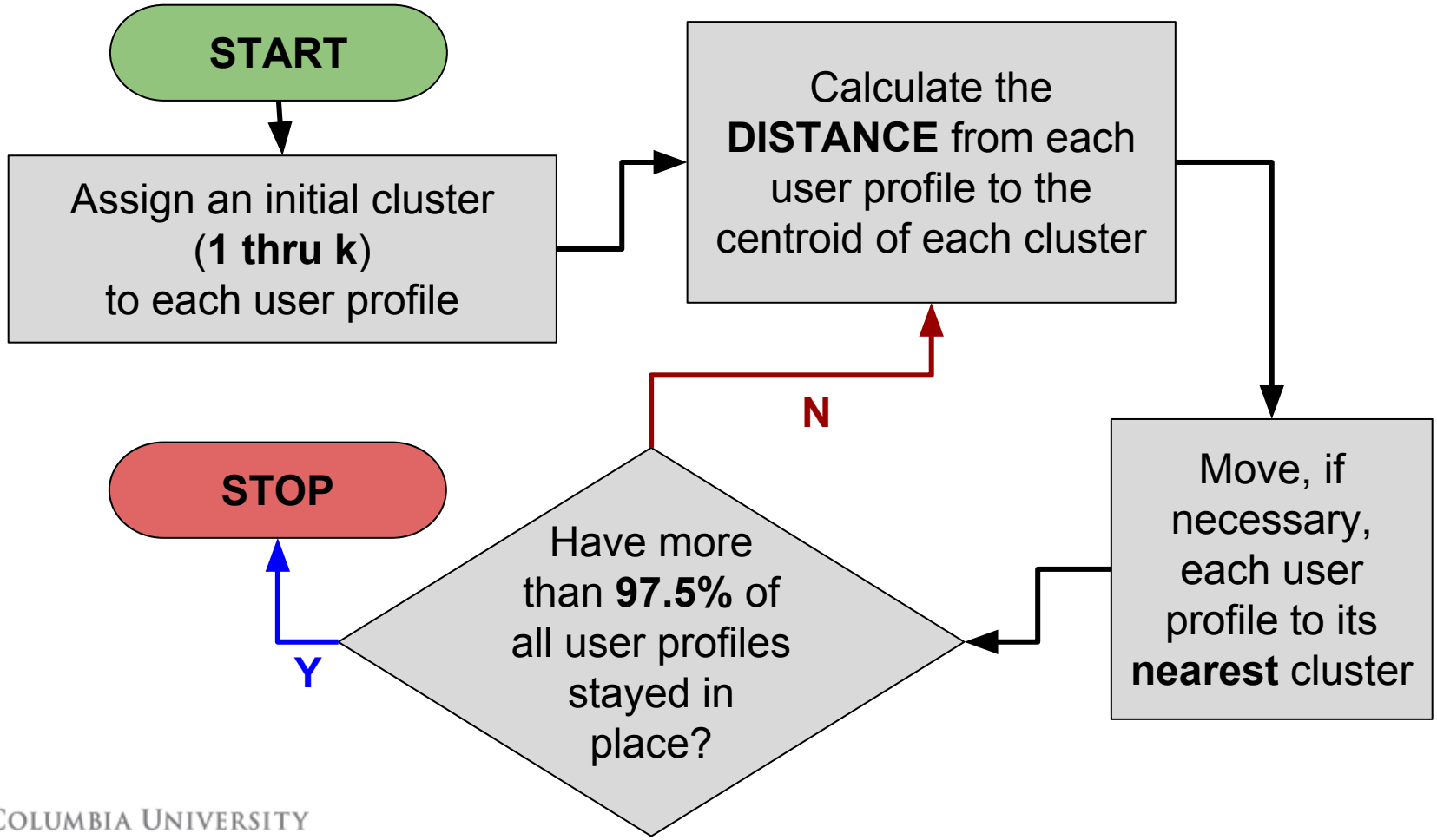
Some Numbers

- 482,055 Total Facebook Likes
- 49,093 Distinct Users
- 3,920 Distinct User Profiles
- 288 Likes by Most Frequent User
- 56,607 Likes by Most Frequent User Profile
  - Profile Code: 13245 - No Children, White/Caucasian, Married, Age 45-54, Income $100k-$125k

Methodology

InfoScout

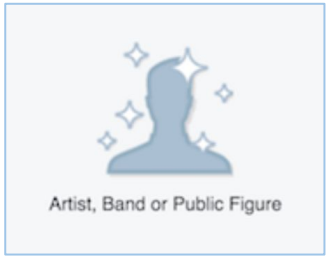
The Steps to Clustering

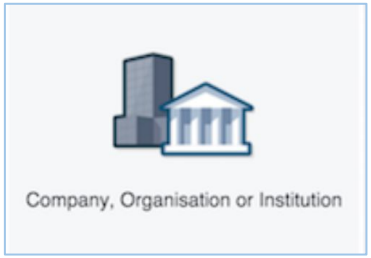


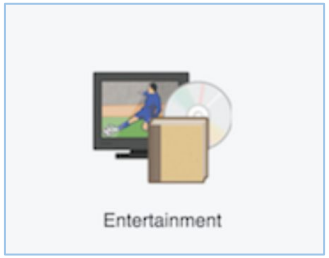
Distance Formula

InfoScout

- We chose to focus on the largest 41 categories in our given Facebook data
- These 41 categories (out of 199 total) accounted for more than 90% of the likes in the data set
- Using the other 158 categories would have added marginal benefit, but added tremendously to the computational runtime





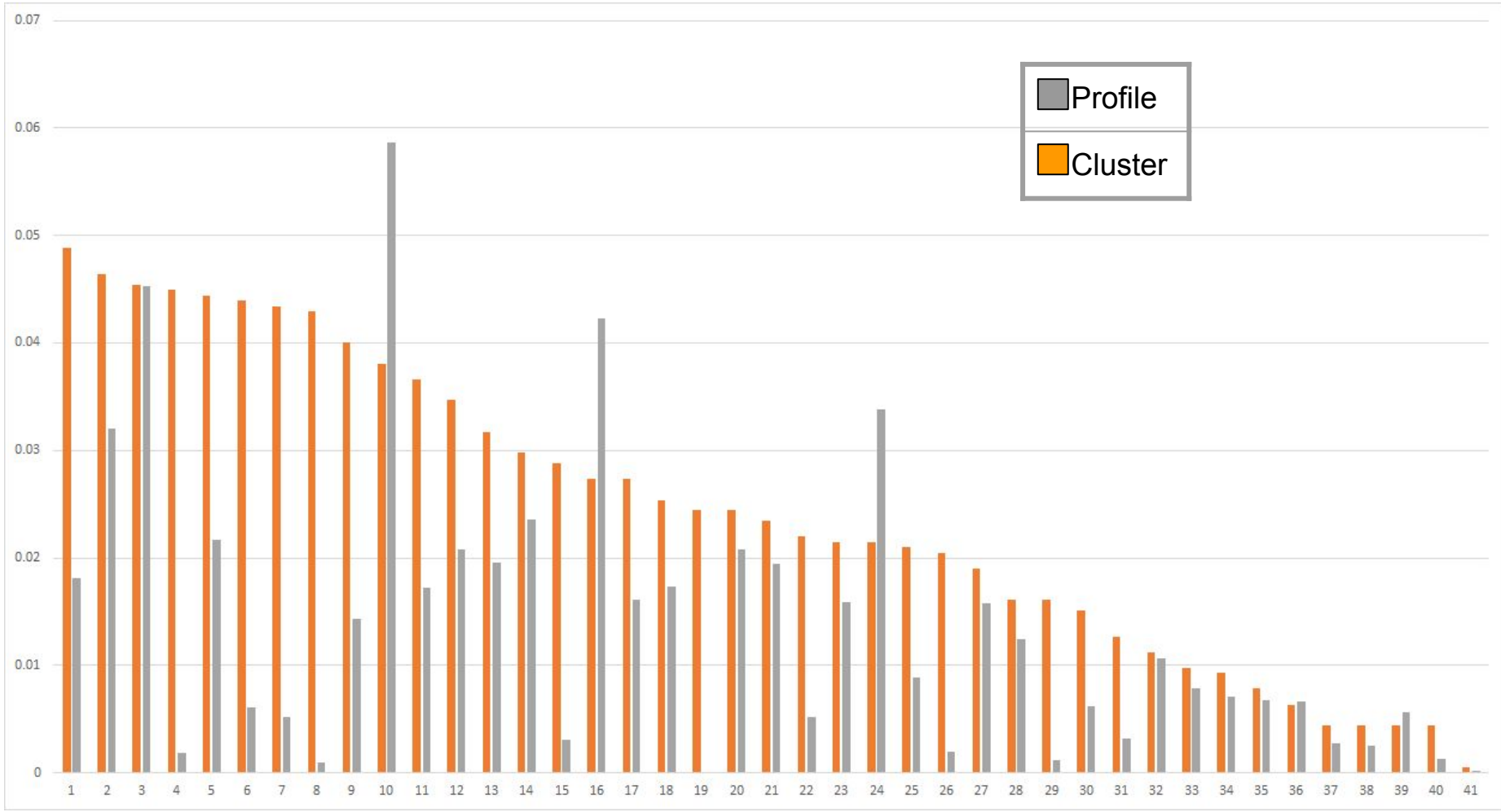


$$D_{a,b} = \sqrt{\sum_{j=1}^{41} \left(p_{a,j} - c_{b,j}\right)^2 \cdot c_{b,j}}$$

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

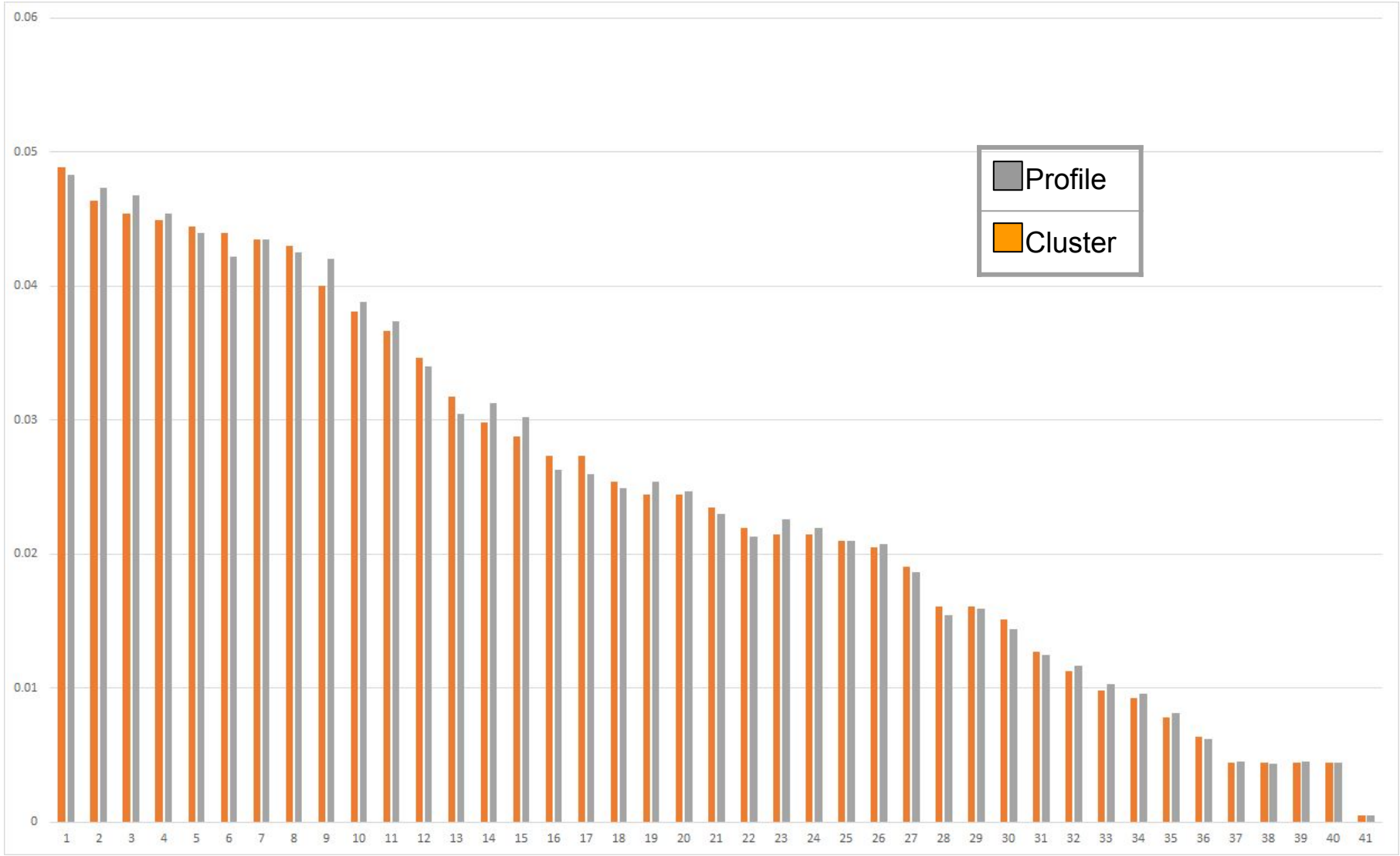
Calculating Distances: Initial Iteration

InfoScout



Calculating Distances: Final Iteration

InfoScout



Why 20 Clusters?

InfoScout

15 Clusters

Cluster	Have Children?	Ethnicity	Marital Status	Age	Income
1	No	White/Caucasian	Never Married	21-24	- $20k
2	Yes	White/Caucasian	Living with Partner	35-44	- $20k
3	Yes	Asian	Never Married	35-44	$80k-100k
4	Yes	White/Caucasian	Living with Partner	25-34	$40k-60k
5	Yes	Black or African American	Never Married	35-44	$40k-60k
6	No	White/Caucasian	Never Married	45-54	$20k-40k
7	Yes	White/Caucasian	Living with Partner	45-54	$100k-125k
8	No	Asian	Never Married	35-44	$60k-80k
9	No	White/Caucasian	Living with Partner	65+	$60k-80k
10	No	White/Caucasian	Living with Partner	55-64	$100k-125k
11	Yes	White/Caucasian	Living with Partner	35-44	$60k-80k
12	No	White/Caucasian	Never Married	25-34	- $20k
13	No	White/Caucasian	Never Married	35-44	$125k +
14	No	White/Caucasian	Never Married	25-34	$40k-60k
15	No	White/Caucasian	Never Married	35-44	$20k-40k

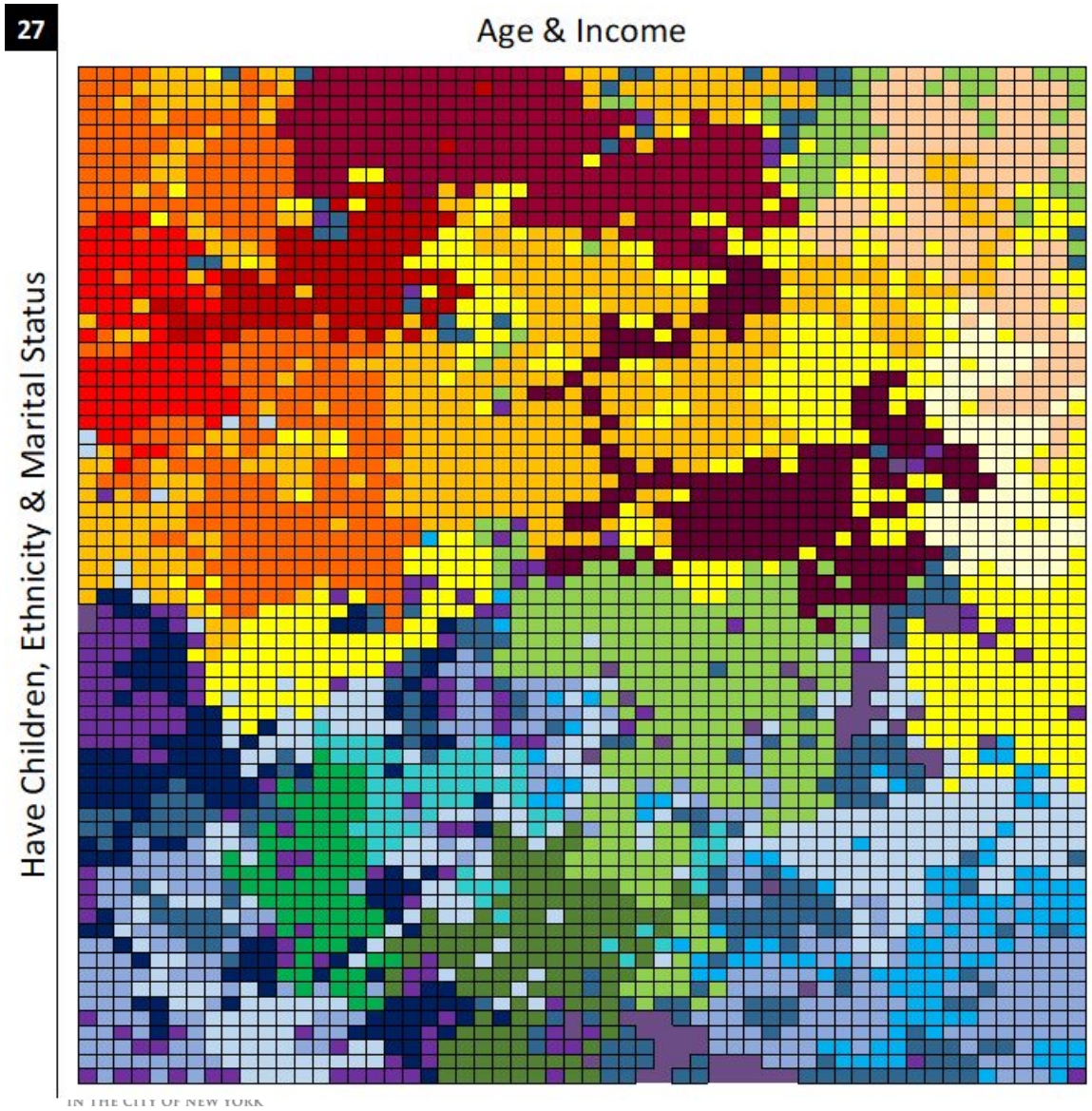
20 Clusters

InfoScout

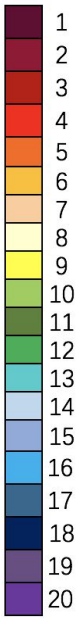
Cluster	Have Children?	Ethnicity	Marital Status	Age	Income
1	No	White/Caucasian	Never Married	35-44	$20k-40k
2	No	White/Caucasian	Never Married	25-34	$100k-125k
3	Yes	White/Caucasian	Living with Partner	55-64	- $20k
4	No	White/Caucasian	Living with Partner	45-54	$20k-40k
5	No	Asian	Married	25-34	$60k-80k
6	No	White/Caucasian	Widower	65+	$60k-80k
7	No	Asian	Never Married	25-34	$40k-60k
8	No	White/Caucasian	Living with Partner	35-44	$60k-80k
9	Yes	White/Caucasian	Never Married	55-64	$60k-80k
10	No	White/Caucasian	Living with Partner	55-64	$40k-60k
11	Yes	Hispanic/Latino	Living with Partner	25-34	$20k-40k
12	Yes	White/Caucasian	Living with Partner	35-44	$125k +
13	Yes	White/Caucasian	Living with Partner	35-44	$80k-100k
14	No	White/Caucasian	Never Married	25-34	- $20k
15	Yes	White/Caucasian	Living with Partner	25-34	$40k-60k
16	Yes	White/Caucasian	Living with Partner	45-54	$125k +
17	No	White/Caucasian	Never Married	21-24	$80k-100k
18	Yes	Black or African American	Never Married	35-44	$40k-60k
19	Yes	White/Caucasian	Living with Partner	45-54	$100k-125k
20	No	White/Caucasian	Living with Partner	65+	$60k-80k

Visualization of Clustering Process

InfoScout



- X-axis: All possible combinations of users' Age & Income
- Y-axis: All possible combinations of other demographic features
- Each color represents a cluster



Final Clusters - A Few Examples

InfoScout

Cluster 7

No Children

Asian

25-34

$40K-$60K

Cluster 13

Have Children

White/Caucasian

35-44

Living with Partner

Cluster 6

No Children

White/Caucasian

65+

$60K-$80K

Final Clusters & Insights

InfoScout

Young and married?

American Airlines

Have Children?

Walt Disney World

65+, Widower?

Making more than $80K?

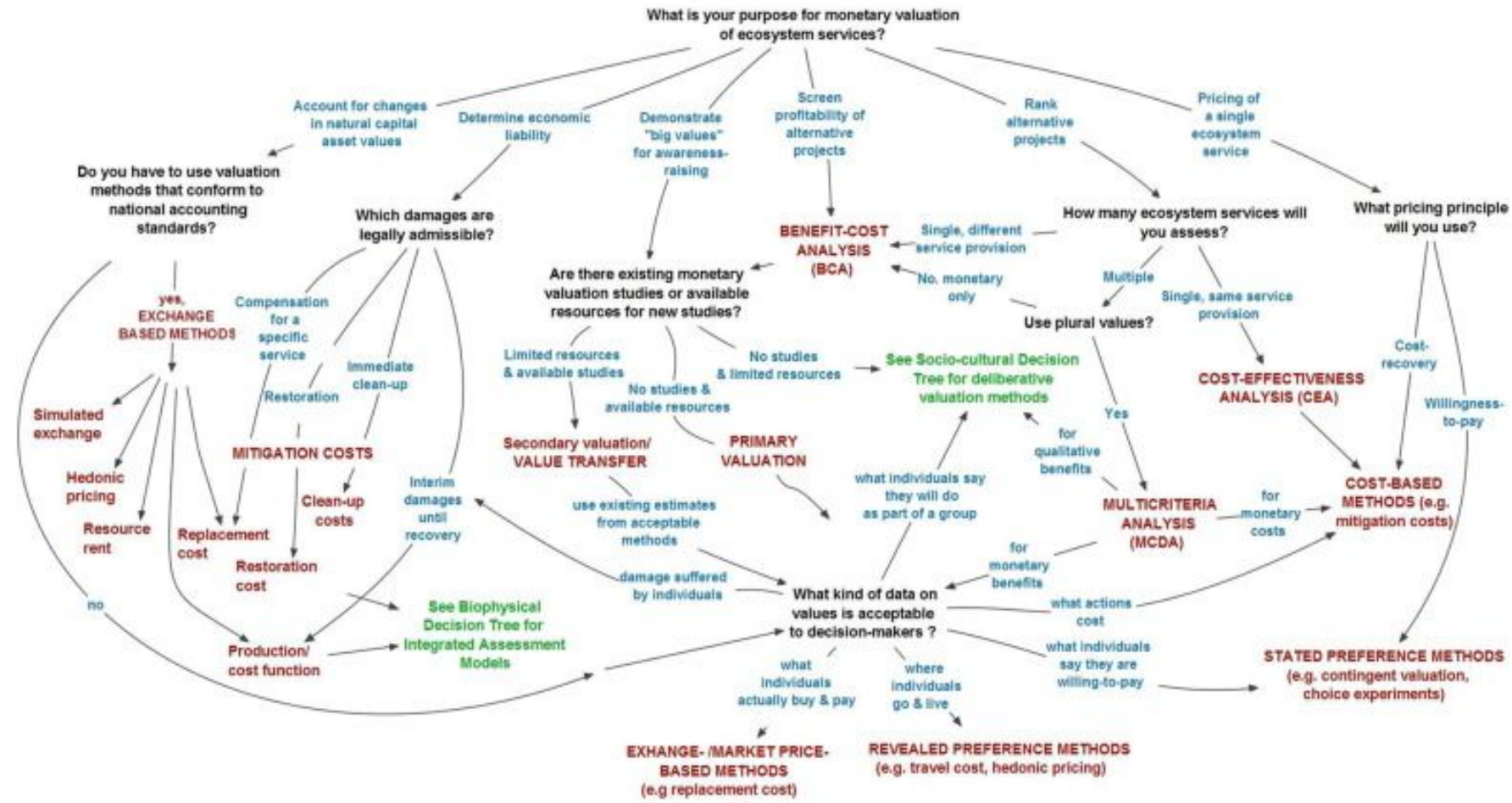
Bath & Body Works

Making less than $40K?

Denny's

How to Access the Results

InfoScout



COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

How Can This Improve?

InfoScout

User Data:

- Gender
- Geographic Location
- Household Size
- Education
- Occupation
- Etc.

Time/ Resources:

- Refine Approach
- Test additional methods/models

Computing Power:

- Compute higher K-levels.
- Use the full range of Facebook Categories.

Data/ Model Integrity:

- Real demographic data could've led to more stable cluster recommendations.
- Model maintenance: Adapting model for the ingestion of new data.

Conclusions

InfoScout

We Have Found:

- Facebook like data combined with user demographics can yield valuable information regarding the users who **like a particular page** and the **pages liked** by a particular user.
- Up to a limit, **more segments** of the user population will provide **more accurate** predictions of consumer behavior.
- **Qualitative analysis** of the clustered data can uncover strong relations within the data that could sometimes be overlooked by an automated process.

Thank You!

InfoScout

Conclusions

InfoScout

We Have Found:

- Facebook Like data can be valuable in yielding new user groups based on consumer demographics and like-behavior.

- K-increases in clusters further refines the insights output from our model.
  - Increases actionability.

Conclusion

InfoScout

Recommendation for Following Steps

- **Expand Data Source**
  - **Customer Information**
    - Collect data from Instagram for analysis of younger consumer group
  - **Consumption Behaviors**
    - Collaborate with online platforms (Amazon, Ebay, etc.) to directly access online shopping data
    - Collaborate with physical retailers (Walmart, Whole Foods, etc.) to directly access offline consumption data
- **Improve Model**

10 Clusters

Cluster	Have Children?	Ethnicity	Marital Status	Age	Income
1	No	White/Caucasian	Never Married	25-34	- $20k
2	No	White/Caucasian	Divorced	35-44	$60k-80k
3	Yes	White/Caucasian	Living with Partner	35-44	- $20k
4	Yes	White/Caucasian	Living with Partner	45-54	$100k-125k
5	No	White/Caucasian	Living with Partner	35-44	$100k-125k
6	No	White/Caucasian	Never Married	25-34	$60k-80k
7	No	White/Caucasian	Never Married	25-34	$40k-60k
8	Yes	White/Caucasian	Living with Partner	35-44	$80k-100k
9	Yes	Asian	Never Married	35-44	$80k-100k
10	No	White/Caucasian	Living with Partner	65+	$60k-80k

InfoScout

Detailed Approach:

Step 1: Assign each feature in our data-frame a value - specifically user attributes (i.e. Has Children, 0 = No, 1 = Yes)

Step 2: Assigned user profile codes to each of our 20 clusters, K= 20

Step 3: Computed the distance of each cluster with respect to the user profile code (Calculated how similar or dissimilar were clusters and profile coders or user demographic data)

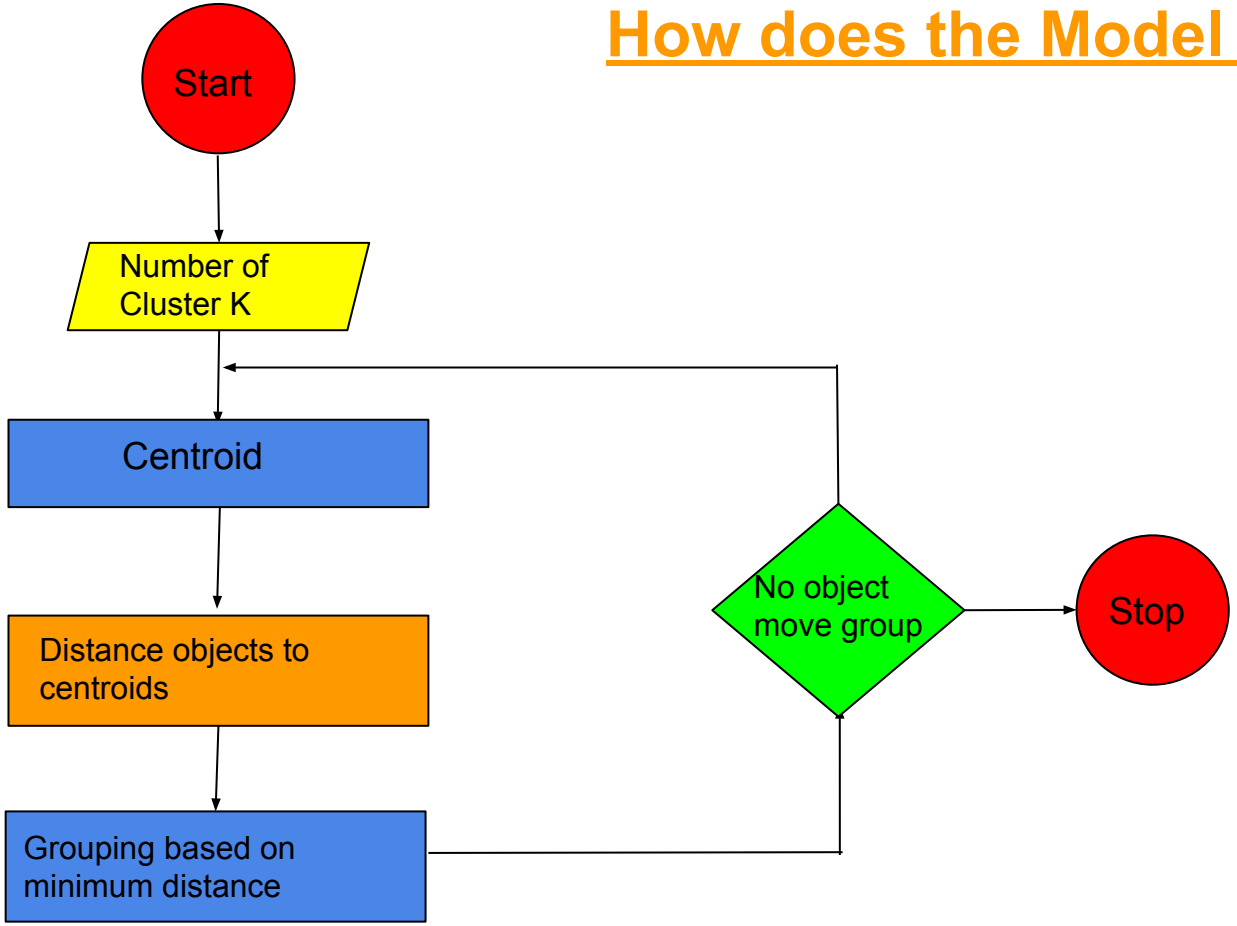
**********DISTANCE FORMULA**********

Step 4: Identified the minimum distance for each cluster node to identify a 20 unique “Cluster Representatives”

Step 5: Used data visualization techniques to provide a comprehensive display of our final output

InfoScout

How does the Model Work?

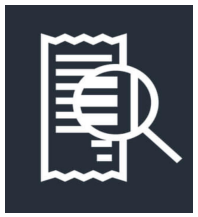


Who Is InfoScout?

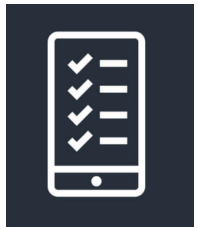
- InfoScout helps brands and retailers grow via next generation **consumer insights**
- Captures receipt transaction information via photograph uploads to one of two of their apps.
- Administers a survey to capture the consumer's motivation for purchase
- Provides this information with their Industry clientele.

Profile the consumer

Capture Receipt photos



Extract the Data



Survey the Consumer

Provide Insights

What's Next? Get Valuable Insights!

InfoScout

Intro - shiny app

Slide Assignments

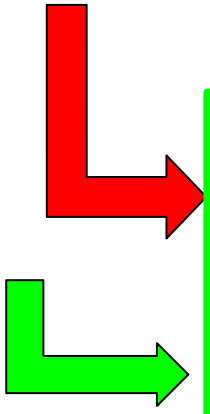
InfoScout

Slide	Owner		Slide	Owner
Cover	Day		- Why k=20?	Shiqing
The Team	Day		- Description of clusters	Adi
Timeline	Qiao		- Visual of clusters (before/after)	Qiao
Project Description (2) - objectives/hypothesis	Zening		- Shiny App - Intro and demo	Adi
Data Cleansing Data Structuring (user profiles)	Adam		Conclusions	
Methodology (3) k-Means (flow charts)	Eric		- Hypothesis proven?	Adam
Distance Formula (bar charts) - Why 41 categories?	Adi		- Limitations (time, computing power, resources, data integrity)	Adam
Results			- What next? - Recommendations	Zening

Reduced row count down to unique occurences of the Category names for each user.

InfoScout

Major duplication problem based on Like I.D. Column



User.ID	Category	Name	like.id	Category.List.Name	Category.List.Item.ID	Age
2155716	Local Business	hhgregg	3.016200e+14	Electronics Store	187937741228885	45-54
2155716	Local Business	hhgregg	3.016200e+14	Appliances	150060378385891	45-54
2155716	Local Business	hhgregg	3.016200e+14	Furniture Store	162845797101278	45-54
1003100	Local Business	hhgregg	1.504221e+14	Appliances	150060378385891	25-34
1003100	Local Business	hhgregg	3.726175e+14	Appliances	150060378385891	25-34
1003100	Local Business	hhgregg	2.021716e+14	Appliances	150060378385891	25-34
1003100	Local Business	hhgregg	1.483026e+14	Appliances	150060378385891	25-34
1003100	Local Business	hhgregg	3.092456e+14	Appliances	150060378385891	25-34
1003100	Local Business	hhgregg	5.370985e+14	Appliances	150060378385891	25-34
1003100	Local Business	hhgregg	3.001964e+14	Appliances	150060378385891	25-34
1003100	Local Business	hhgregg	3.037826e+14	Appliances	150060378385891	25-34
1003100	Local Business	hhgregg	1.504221e+14	Electronics Store	187937741228885	25-34
1003100	Local Business	hhgregg	3.726175e+14	Electronics Store	187937741228885	25-34
1003100	Local Business	hhgregg	2.021716e+14	Electronics Store	187937741228885	25-34
1003100	Local Business	hhgregg	1.483026e+14	Electronics Store	187937741228885	25-34
1003100	Local Business	hhgregg	3.092456e+14	Electronics Store	187937741228885	25-34
1003100	Local Business	hhgregg	5.370985e+14	Electronics Store	187937741228885	25-34
1003100	Local Business	hhgregg	3.001964e+14	Electronics Store	187937741228885	25-34
1003100	Local Business	hhgregg	3.037826e+14	Electronics Store	187937741228885	25-34
1003100	Local Business	hhgregg	1.504221e+14	Furniture Store	162845797101278	25-34
1003100	Local Business	hhgregg	3.726175e+14	Furniture Store	162845797101278	25-34
1003100	Local Business	hhgregg	2.021716e+14	Furniture Store	162845797101278	25-34
1003100	Local Business	hhgregg	1.483026e+14	Furniture Store	162845797101278	25-34
1003100	Local Business	hhgregg	3.092456e+14	Furniture Store	162845797101278	25-34
1003100	Local Business	hhgregg	5.370985e+14	Furniture Store	162845797101278	25-34
1003100	Local Business	hhgregg	3.001964e+14	Furniture Store	162845797101278	25-34
1003100	Local Business	hhgregg	3.037826e+14	Furniture Store	162845797101278	25-34
1003100	Local Business	hhgregg	1.504221e+14	Appliances	150060378385891	25-34
1003100	Local Business	hhgregg	3.726175e+14	Appliances	150060378385891	25-34
1003100	Local Business	hhgregg	2.021716e+14	Appliances	150060378385891	25-34