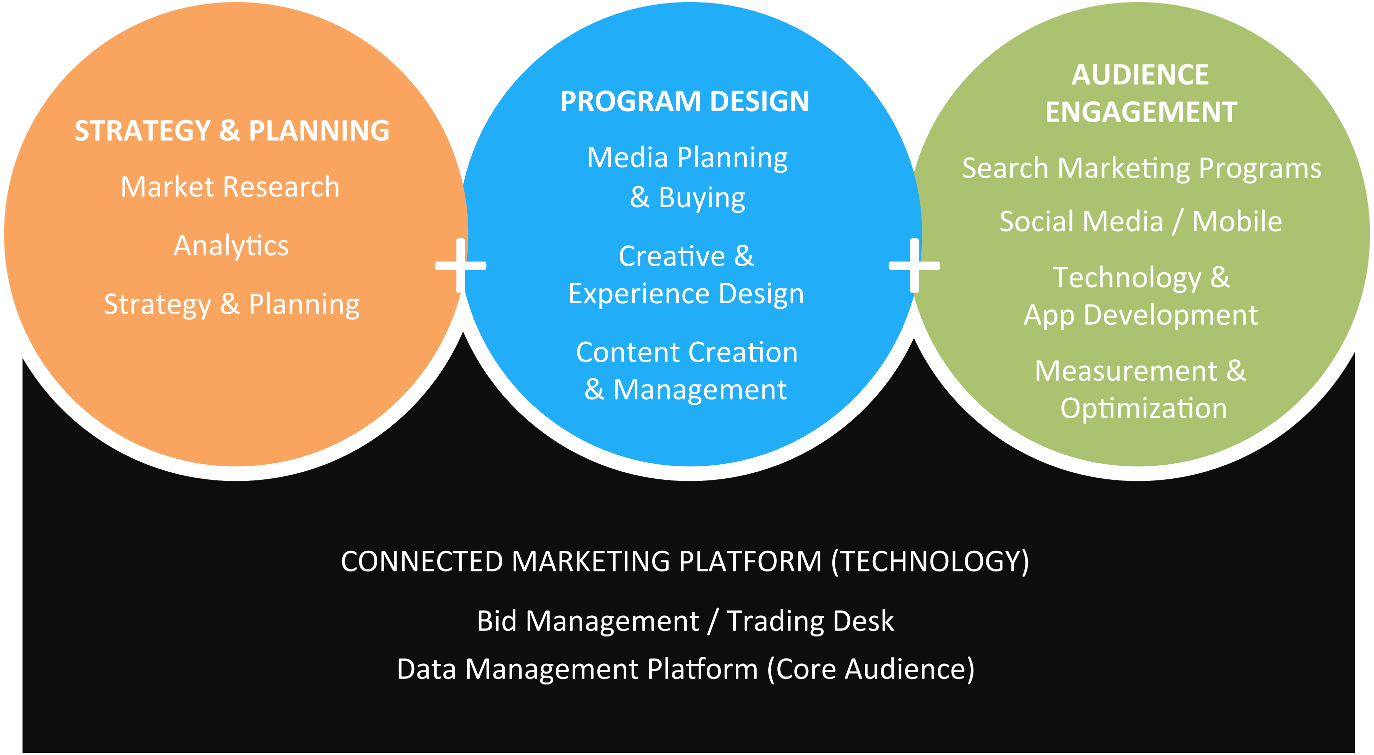
Real-time Interactive Big Data Analysis Using In-Memory Computing

Mike Joyce – Manager Software Engineer, iCrossing
Shawn Nguyen – Lead Software Engineer, iCrossing

icrossing



STRATEGY & PLANNING

Market Research

Analytics

Strategy & Planning

+

PROGRAM DESIGN

Media Planning & Buying

Creative & Experience Design

Content Creation & Management

+

AUDIENCE ENGAGEMENT

Search Marketing Programs

Social Media / Mobile

Technology & App Development

Measurement & Optimization

CONNECTED MARKETING PLATFORM (TECHNOLOGY)

Bid Management / Trading Desk

Data Management Platform (Core Audience)

DIGITAL AGENCY INSIDE A

CONTENT EMPIRE

Leveraging audience insights:

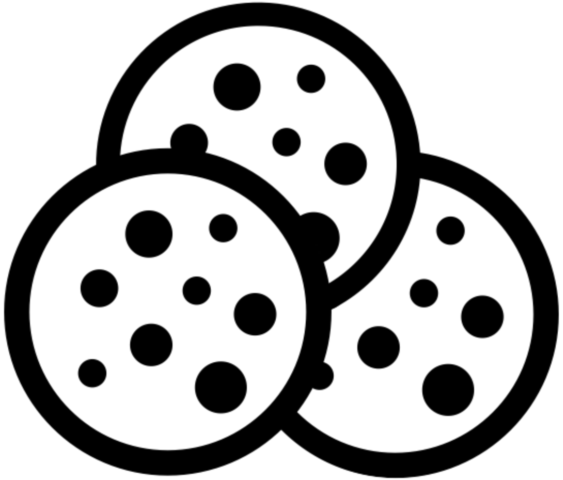
ELLE HouseBeautiful Esquire COSMOPOLITAN

- 20+ brands
- 30+ TV networks
- 50+ newspapers
- 300+ magazines

Big Data - Cookies!

- Subscribers
- Visitors
- International
- Multiple devices

300+ million
unique cookies



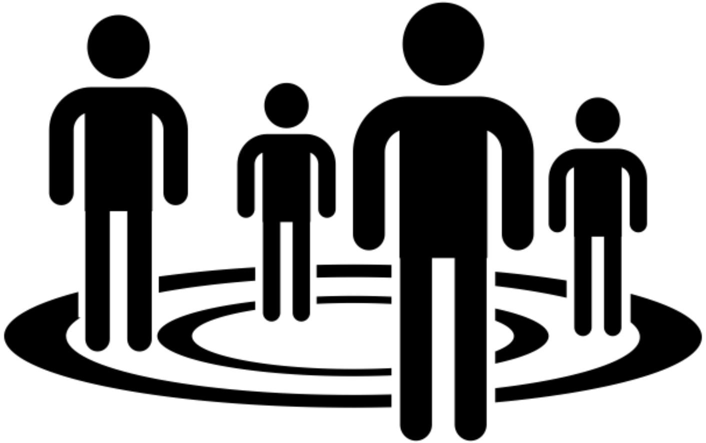
Created by Nathan Thomson
from the Noun Project

DMP Audience Data

Attributes

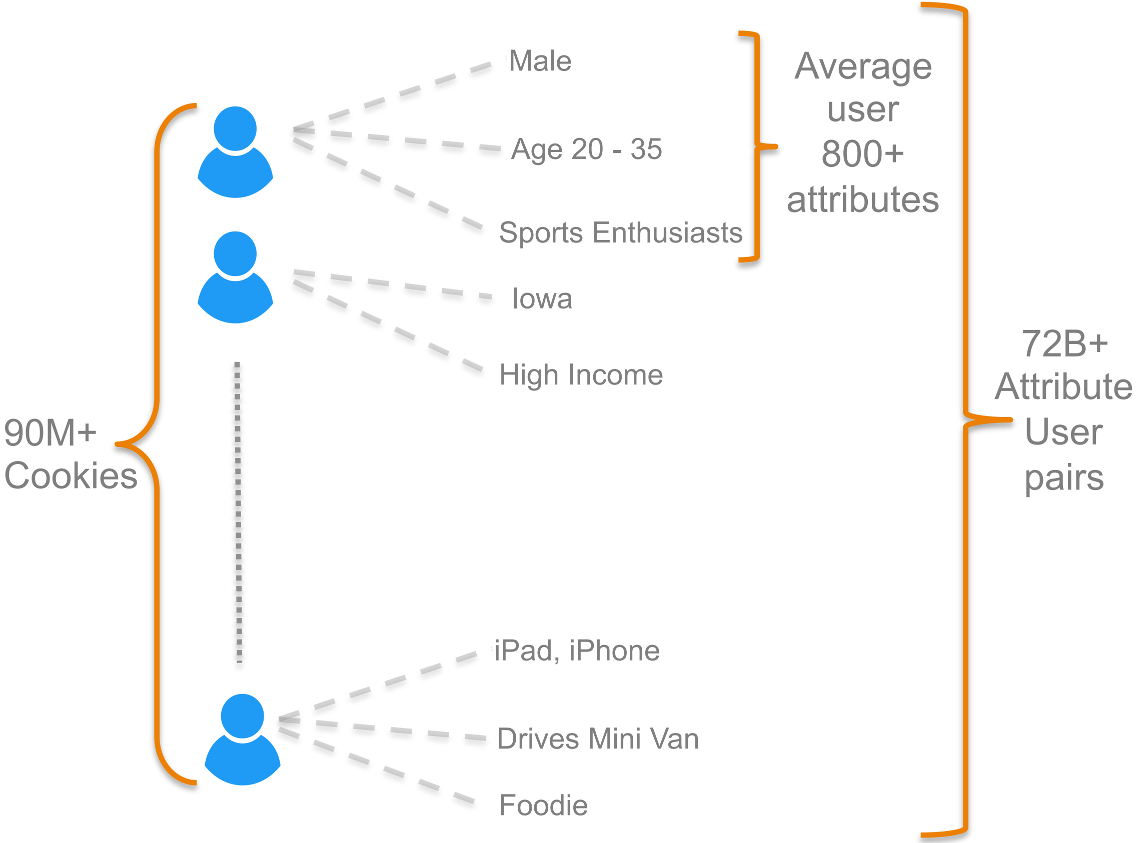
- Geographic
- Demographic
- Behavioral
- Psychographic

11,000+ Unique Attributes



Created by Creative Stall
from the Noun Project

Cookies + Audience Attributes = Super Big Data!



Male

Age 20 - 35

Sports Enthusiasts

Average user 800+ attributes

Iowa

High Income

90M+ Cookies

72B+ Attribute User pairs

iPad, iPhone

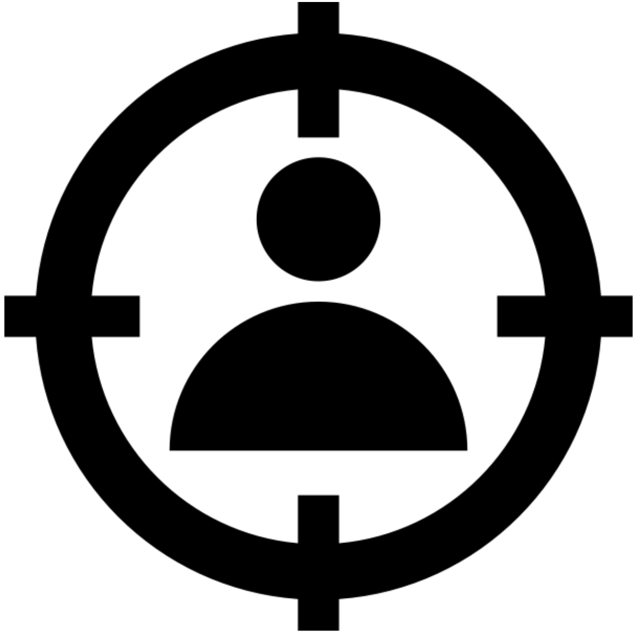
Drives Mini Van

Foodie

Audiences – Targeting vs Discovering

- Who you are targeting

- How do you connect with them?
- What describes them?

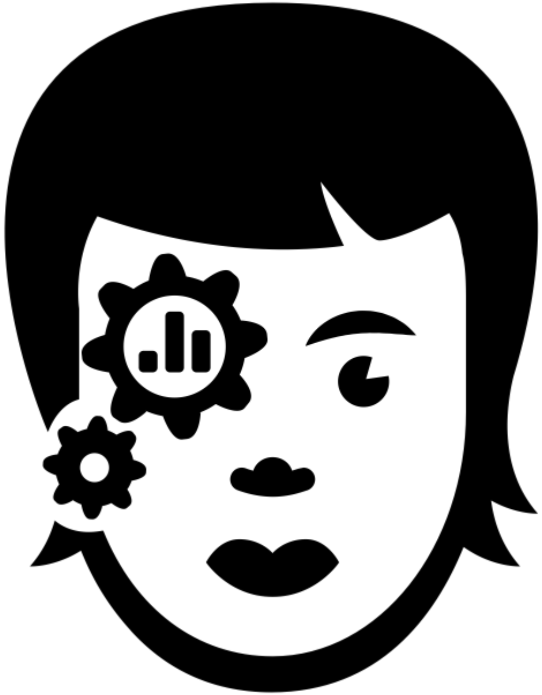


Created by Creative Stall
from the Noun Project

Data Scientists

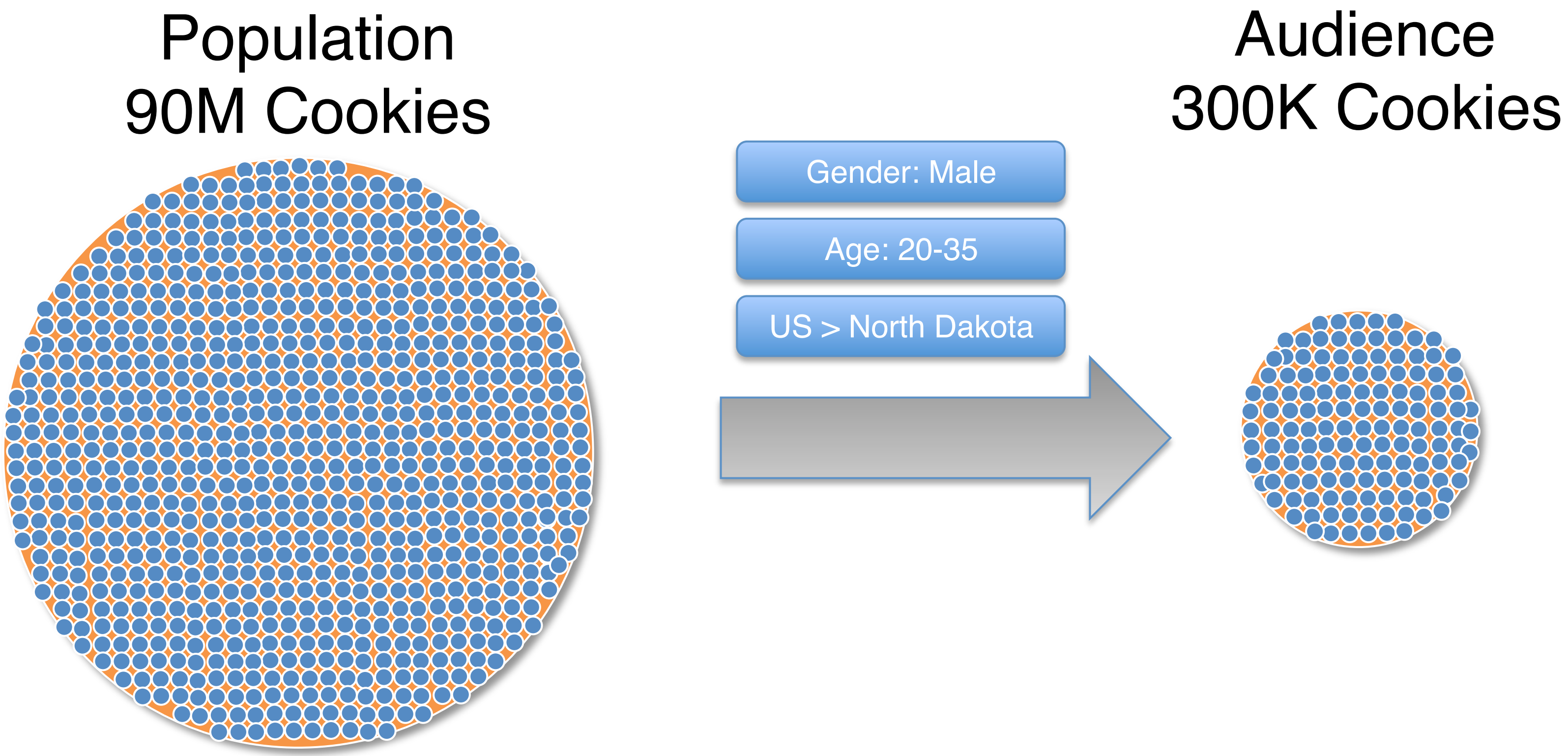
<u>Discovering Audience Attributes</u>

1. Define an audience using attributes
2. Identify all attributes of cookies in audience
3. Calculate highly indexing attributes



Created by Thibault Geffroy
from the Noun Project

1) Define the Audience



Population
90M Cookies

Gender: Male

Age: 20-35

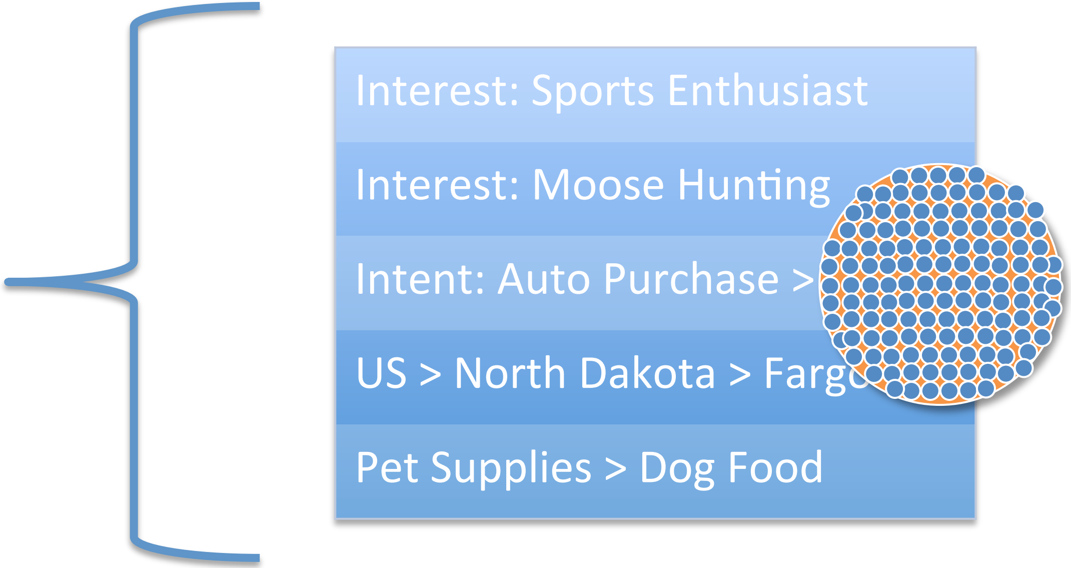
US > North Dakota

Audience
300K Cookies

2) Audience Attributes

Attributes of Cookies in Audience

Audience 300K Cookies



Interest: Sports Enthusiast

Interest: Moose Hunting

Intent: Auto Purchase >

US > North Dakota > Fargo

Pet Supplies > Dog Food

3) Index the Attributes

Attributes of
Cookies in Audience

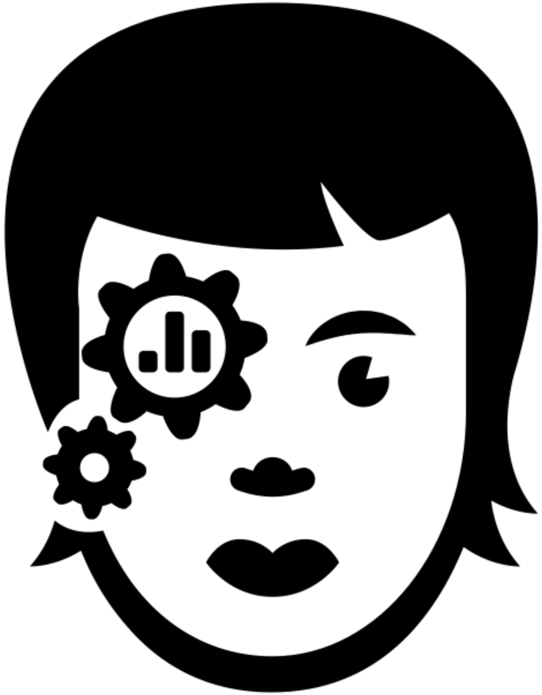
- Interest: Sports Enthusiast
- Interest: Moose Hunting
- Intent: Auto Purchase > Truck
- US > North Dakota > Fargo
- Pet Supplies > Dog Food

Attribute	Audience	Population
Interest: Sports Enthusia		
Interest: Moose Hunting		
Intent: Auto Purchase >		
US > North Dakota > Far		
Pet Supplies > Dog Food	6%	9%

Data Scientists

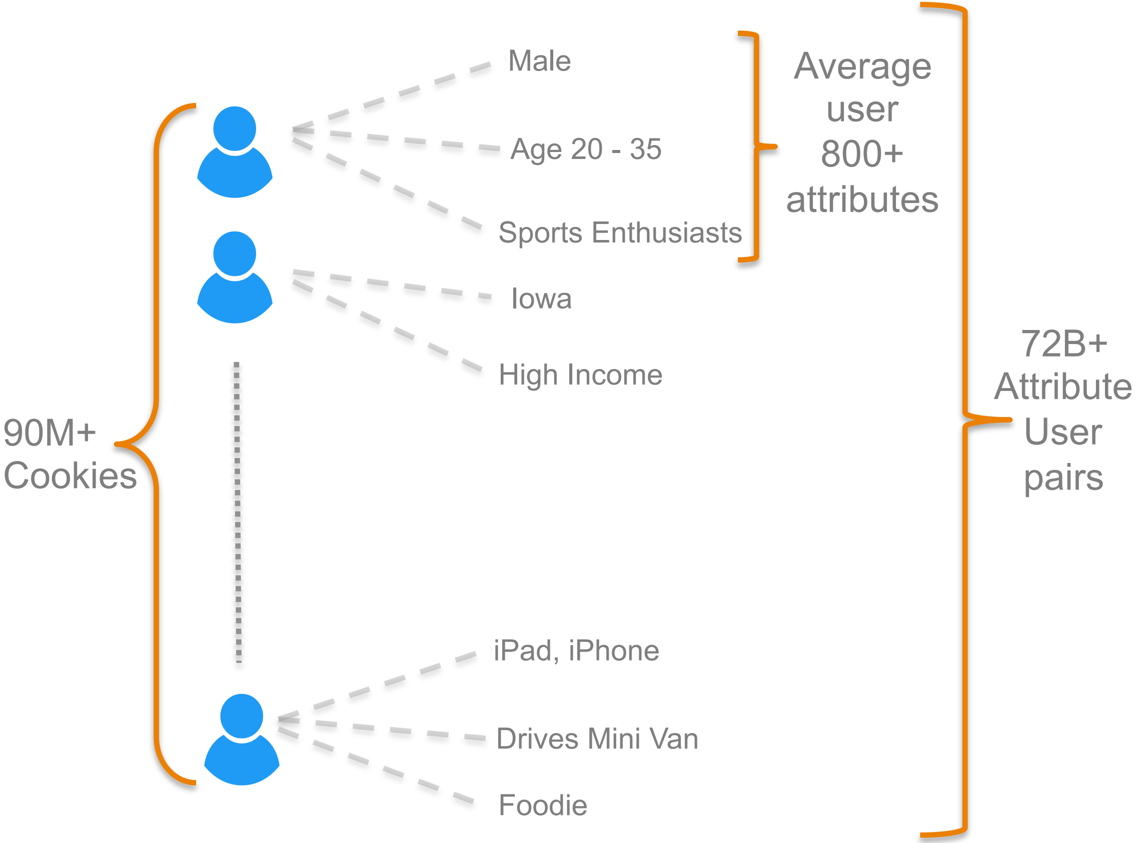
Development Ask

1. Make it accessible to "normals"
2. Exportable visualizations & calculations
3. Reduce query time from 1 hr to 1 sec



Created by Thibault Geffroy from the Noun Project

Why is this Hard?

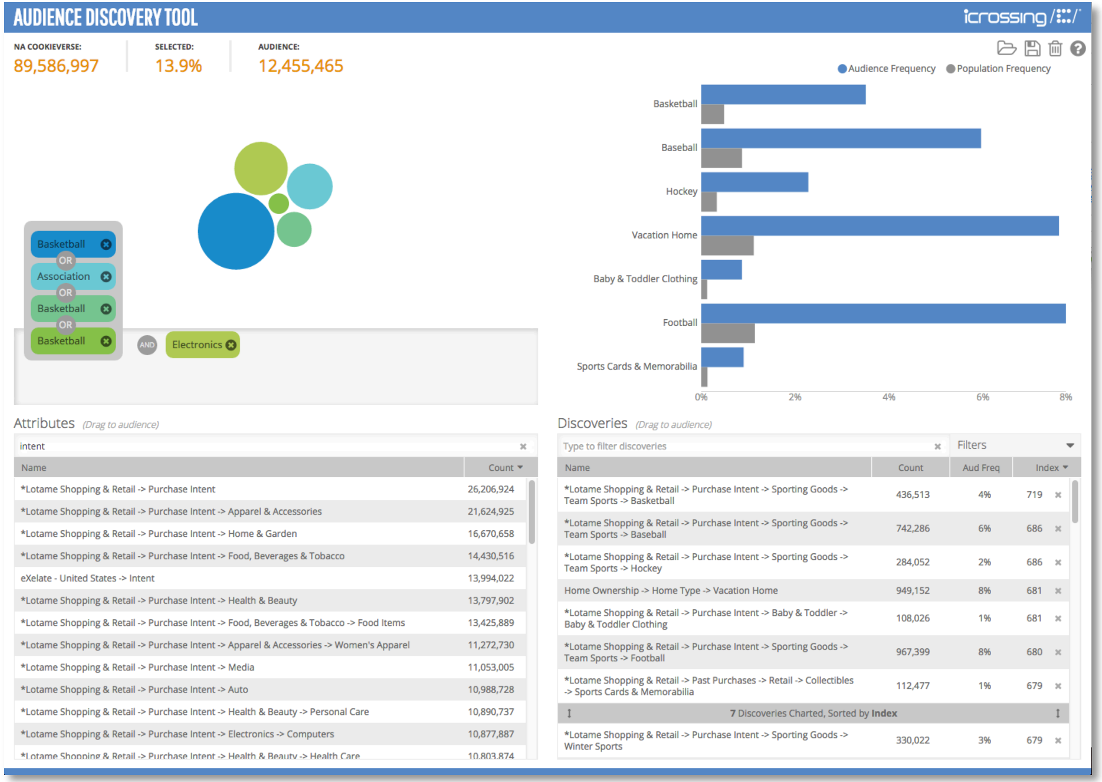


<u>Algorithm</u>

1. Check every cookie if it satisfies audience criteria
2. Collect all attributes for every audience cookie
3. Calculate percentages & index

Within 1 sec !!!!!!

The Answer – Audience Discovery Tool



NA COOKIEVERSE: 89,586,997

SELECTED: 13.9%

AUDIENCE: 12,455,465

Attributes *(Drag to audience)*

intent

Name	Count
*Lotame Shopping & Retail -> Purchase Intent	26,206,924
*Lotame Shopping & Retail -> Purchase Intent -> Apparel & Accessories	21,624,925
*Lotame Shopping & Retail -> Purchase Intent -> Home & Garden	16,670,658
*Lotame Shopping & Retail -> Purchase Intent -> Food, Beverages & Tobacco	14,430,516
eXelate - United States -> Intent	13,994,022
*Lotame Shopping & Retail -> Purchase Intent -> Health & Beauty	13,797,902
*Lotame Shopping & Retail -> Purchase Intent -> Food, Beverages & Tobacco -> Food Items	13,425,889
*Lotame Shopping & Retail -> Purchase Intent -> Apparel & Accessories -> Women's Apparel	11,272,730
*Lotame Shopping & Retail -> Purchase Intent -> Media	11,053,005
*Lotame Shopping & Retail -> Purchase Intent -> Auto	10,988,728
*Lotame Shopping & Retail -> Purchase Intent -> Health & Beauty -> Personal Care	10,890,737
*Lotame Shopping & Retail -> Purchase Intent -> Electronics -> Computers	10,877,887
*Lotame Shopping & Retail -> Purchase Intent -> Health & Beauty -> Health Care	10,803,874

Discoveries *(Drag to audience)*

Type to filter discoveries

Filters

Name	Count	Aud Freq	Index
*Lotame Shopping & Retail -> Purchase Intent -> Sporting Goods -> Team Sports -> Basketball	436,513	4%	719
*Lotame Shopping & Retail -> Purchase Intent -> Sporting Goods -> Team Sports -> Baseball	742,286	6%	686
*Lotame Shopping & Retail -> Purchase Intent -> Sporting Goods -> Team Sports -> Hockey	284,052	2%	686
Home Ownership -> Home Type -> Vacation Home	949,152	8%	681
*Lotame Shopping & Retail -> Purchase Intent -> Baby & Toddler -> Baby & Toddler Clothing	108,026	1%	681
*Lotame Shopping & Retail -> Purchase Intent -> Sporting Goods -> Team Sports -> Football	967,399	8%	680
*Lotame Shopping & Retail -> Past Purchases -> Retail -> Collectibles -> Sports Cards & Memorabilia	112,477	1%	679
7 Discoveries Charted, Sorted by **Index**			
*Lotame Shopping & Retail -> Purchase Intent -> Sporting Goods -> Winter Sports	330,022	3%	679

- Audience discovery
  - Cookie Attributes
  - Frequency vs Population
- Built for non-technical users
  - Strategy
  - Sales / Account
  - Anyone
- Flexible
  - Research tool
  - In-meeting, iterative discovery
- Approachable
  - Real-time
  - Results in seconds
  - Simple, elegant interface
  - Multiple export formats

“Making science accessible”

Data Processing R& D

Traditional Relational Databases

- Long load time
- Complex queries resulting in long query times
- Rigid data model

Non Traditional Databases

- Lack of complex query feature
- Large memory footprint requirement
- Aggregation query exceeded by many 10x of seconds

Neo4j
the graph database

cassandra

mongoDB
{ name: mongo, type: DB }

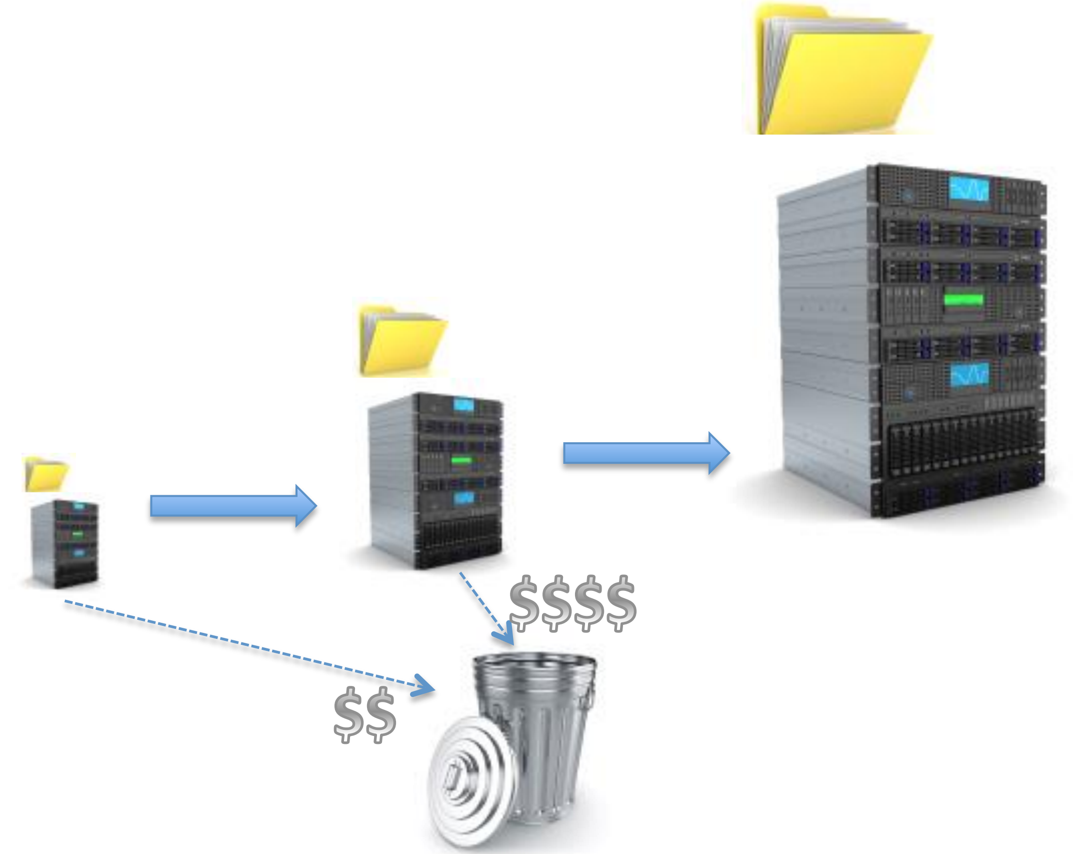
redis

The Low Hanging Fruit

- In memory cache
- Customizable query using Java code
- Relatively low data loading time

Neo4j
the graph database

The Vertical Problem



Distributed Computing Ecosystem

- Not production ready
- Data import fails without explanation
- Aggregation fails to impress

cloudera IMPALA

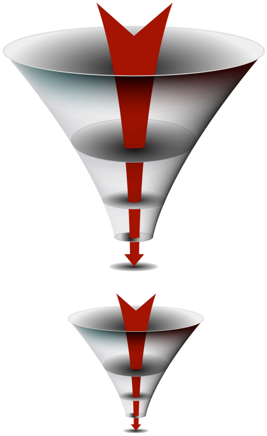
SHARK

Spark
Lightning-Fast Cluster Computing

Back to Basics

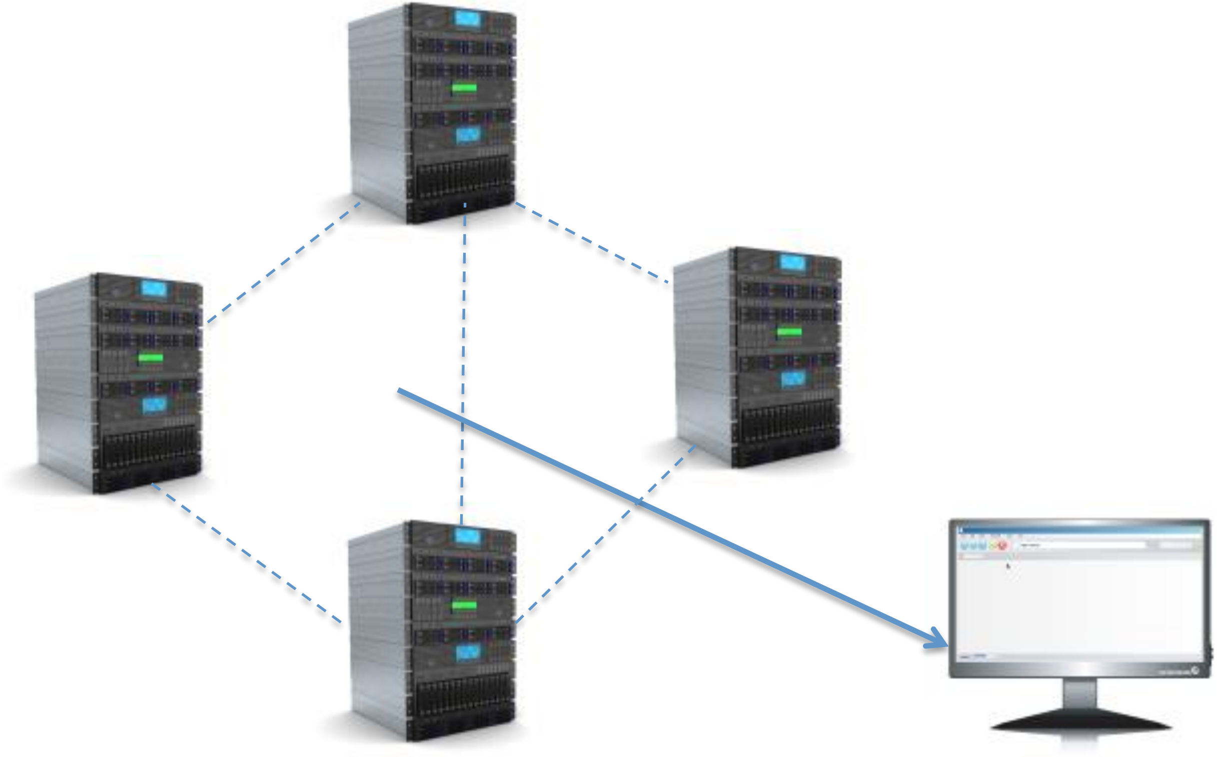
- Pure Java code solution
- Data and logic must exists in same memory space
- Capable of advanced filtering

GridGain



- Distributed computing, low overhead
- Data locality
- Minimal code migration

The Distributed Solution



The Challenges

- Tedious manual data distribution
- Gar building and deployment issues
- Development challenges

What We Learned

- Indexed data requiring minor calculations -- databases (relational & noSQL) great

- Large non-indexed data -- the data & processing need to live in the same (memory) space