Practical Data Science & Analytics

JOSE ABELENDA
LYFT

THOUGHT LEADERS IN DATA SCIENCE & BIG DATA ANALYTICS

UC BERKELEY

FEBRUARY 1, 2016

Intro.

Jose Abelenda

- Director of marketing analytics as Lyft
- Director of marketing analytics at Hotwire
- Previously senior data scientist at PayPal
- Also worked at JP Morgan Chase, Washington Mutual, Providian Fin.
- Berkeley alumni.

What it is all about

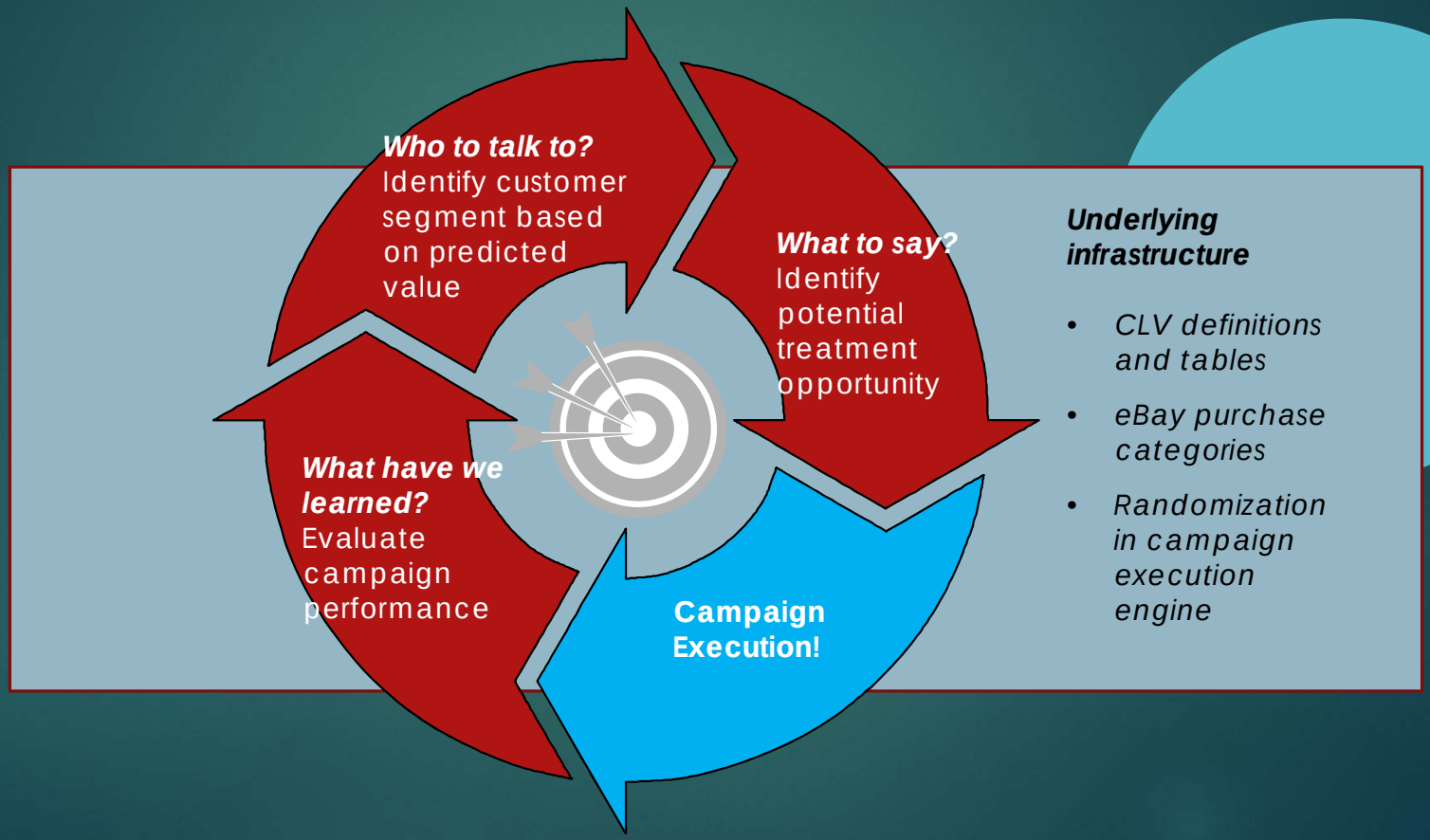
Context:

- Customer lifecycle – Customer engagement.

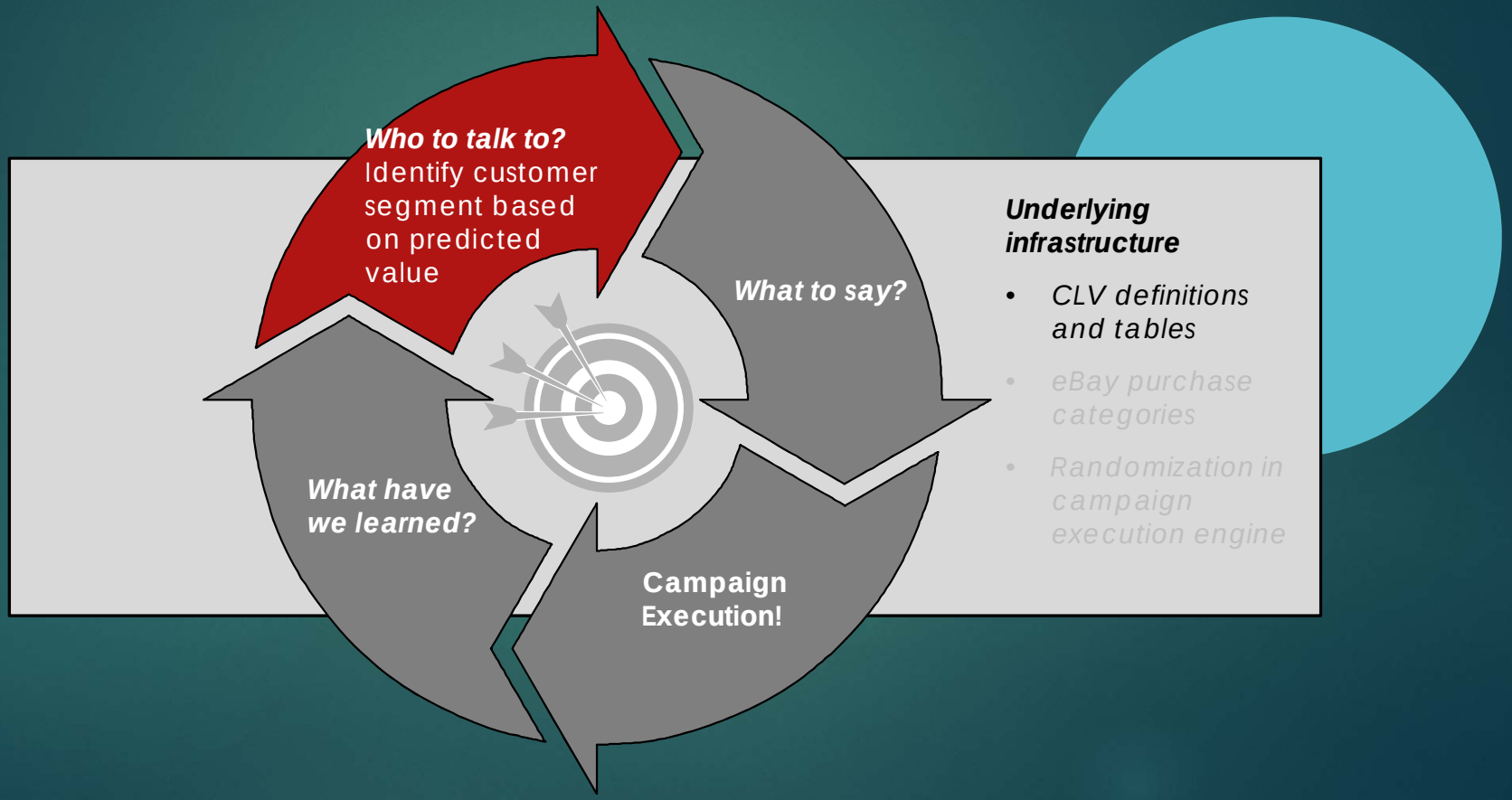
Topics:

- Supervised learning
- Unsupervised learning
- Causal inference

Customer Lifecycle



Who should we talk to?



Who to talk to: which of my customers drive the most value

- 80/20 Rule
- Traditional approaches CLV and RFM do remarkably well, but are backward looking
- PayPal's business need:
  - Drive customer engagement as early as possible in the lifecycle.
  - Empower/Evaluate the effectiveness of customer acquisition efforts
- Solution: build a model to predict first year cumulative customer value after 30 days from signup

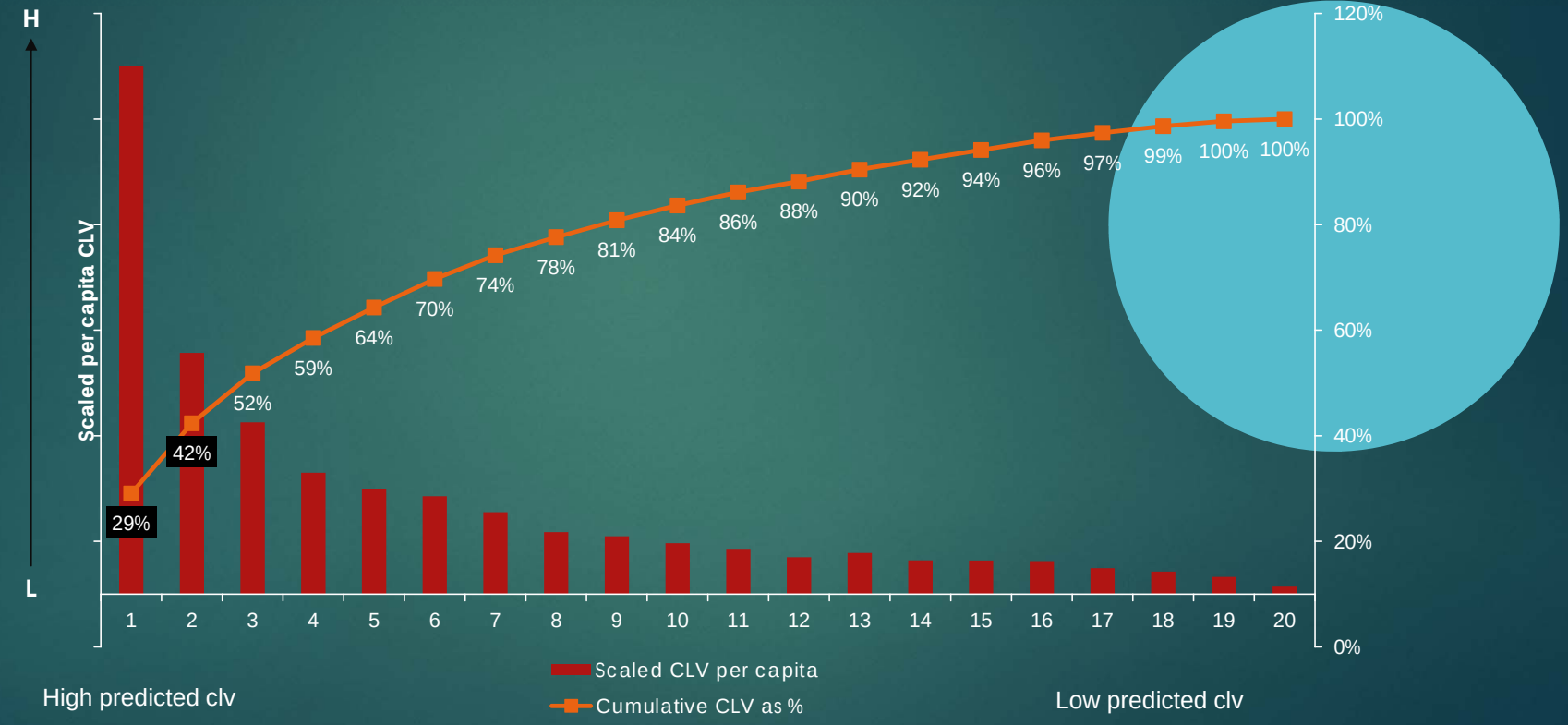
Key underlying infrastructure

- Customer CLV tables and definitions
  - Every source and sink of cash clearly itemized and tracked at an individual level and on a monthly basis going back several years
- Customer fact DataMart
  - All relevant customer events tracked over time
- Transaction fact DataMart
  - Complete description of transaction attributes

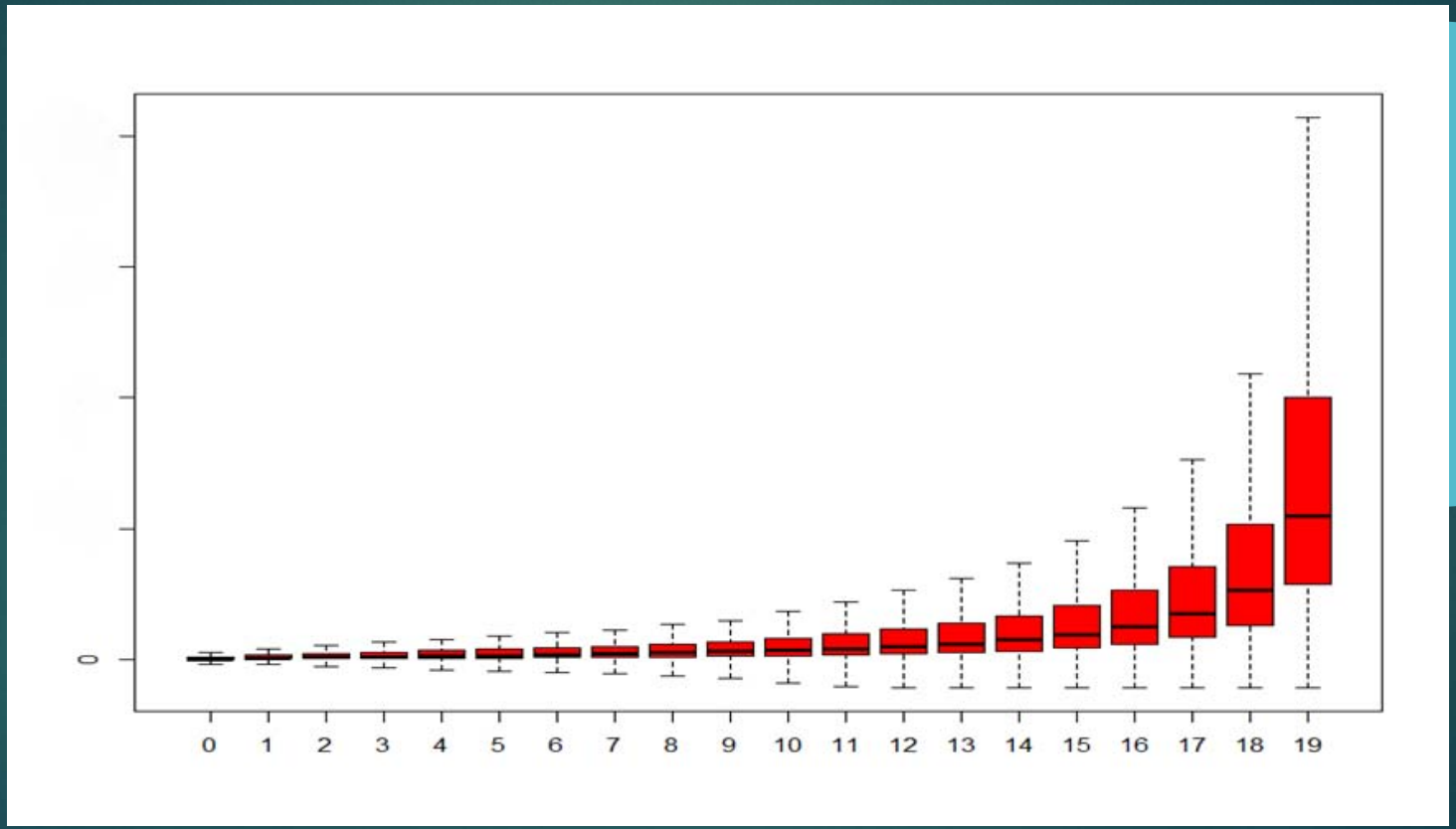
Which tools /techniques did we use

- Hadoop i.e. Hive for feature generation
- R running on a large RAM machine
  - Regression
  - CART – ‘rpart’ R package
  - Random forest – ‘randomForest’ R package
- All free open source software / somewhat expensive hardware

Actual Y1 CLV by predicted CLV deciles



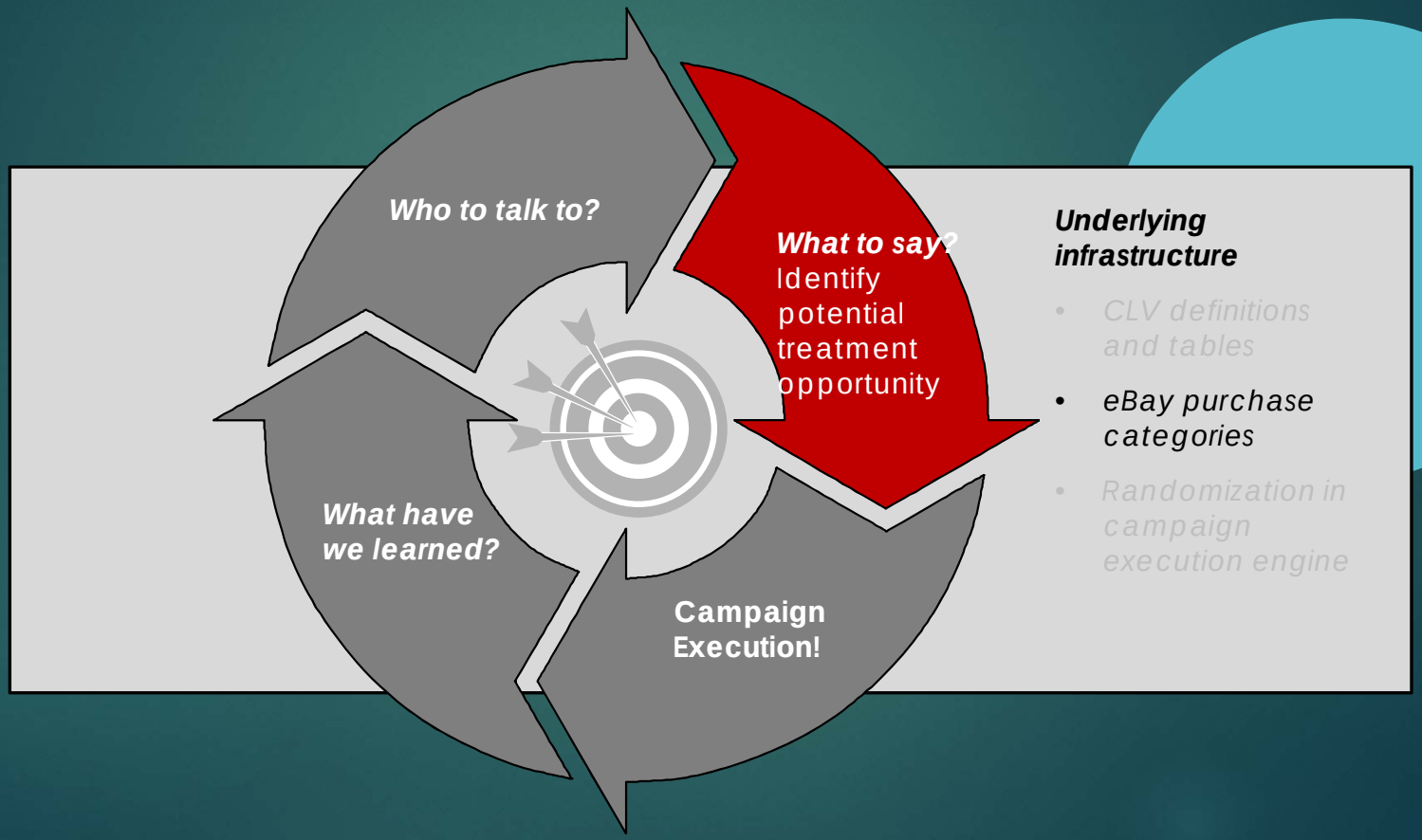
Predicted vs. actual



Take aways

- Focus on a genuine business need:
  - Drive customer engagement as early as possible in the lifecycle.
  - Empower/Evaluate the effectiveness of customer acquisition efforts
- Great data infrastructure leads to reliable and scalable results with shorter development cycles
- Off the shelf algorithms can do a great job if you have the right data

What should we say?



What should we say: which content/offers will be most relevant

- Promote well know merchants, provide rich offers
- Building recommendation engine / expensive, time consuming, requires enterprise that can consume thousands of recommendations
- PayPal's business need:
  - Promote merchants/brands/products which are most relevant. Leverage its rich ecosystem
  - Empower the consumer engagement team with actionable insights about customer tastes and preferences
- Solution: build a rich yet interpretable segmentation based on customer past purchases and implied preferences

Key underlying infrastructure

- eBay /PP purchase categories.
  - 40 high level categories classifying all items sold and purchased on eBay or any PP merchant.
  - 2nd and 3rd level of granularity for each high level category.
- Customer fact DataMart
  - All relevant customer events tracked over time
- Transaction fact DataMart
  - Complete description of transaction attributes

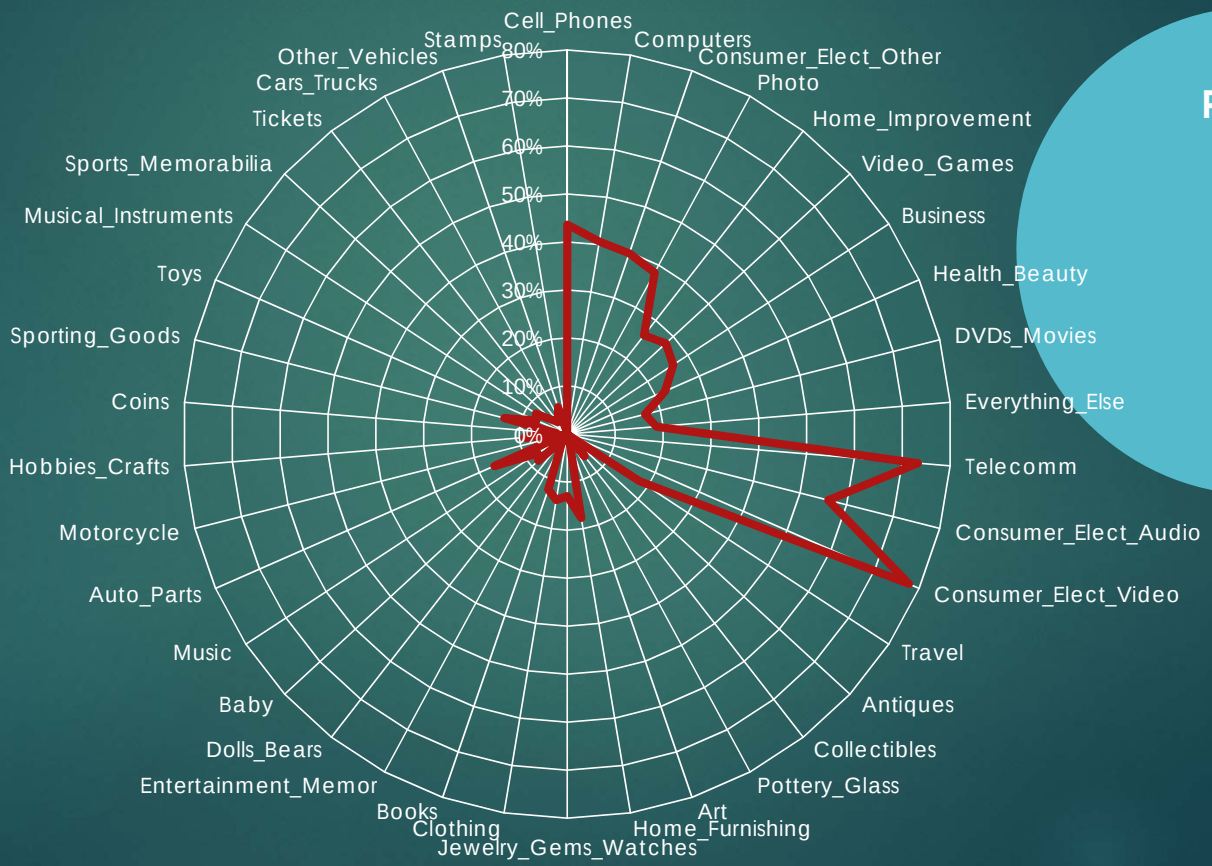
Which tools /techniques did we use

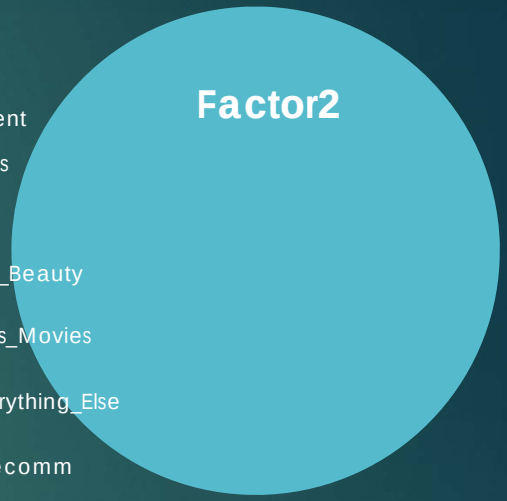
- Teradata SQL for feature generation
- SAS running on enterprise server
  - Factor Analysis
- High quality enterprise software
- Expensive enterprise hardware

Segmentation output

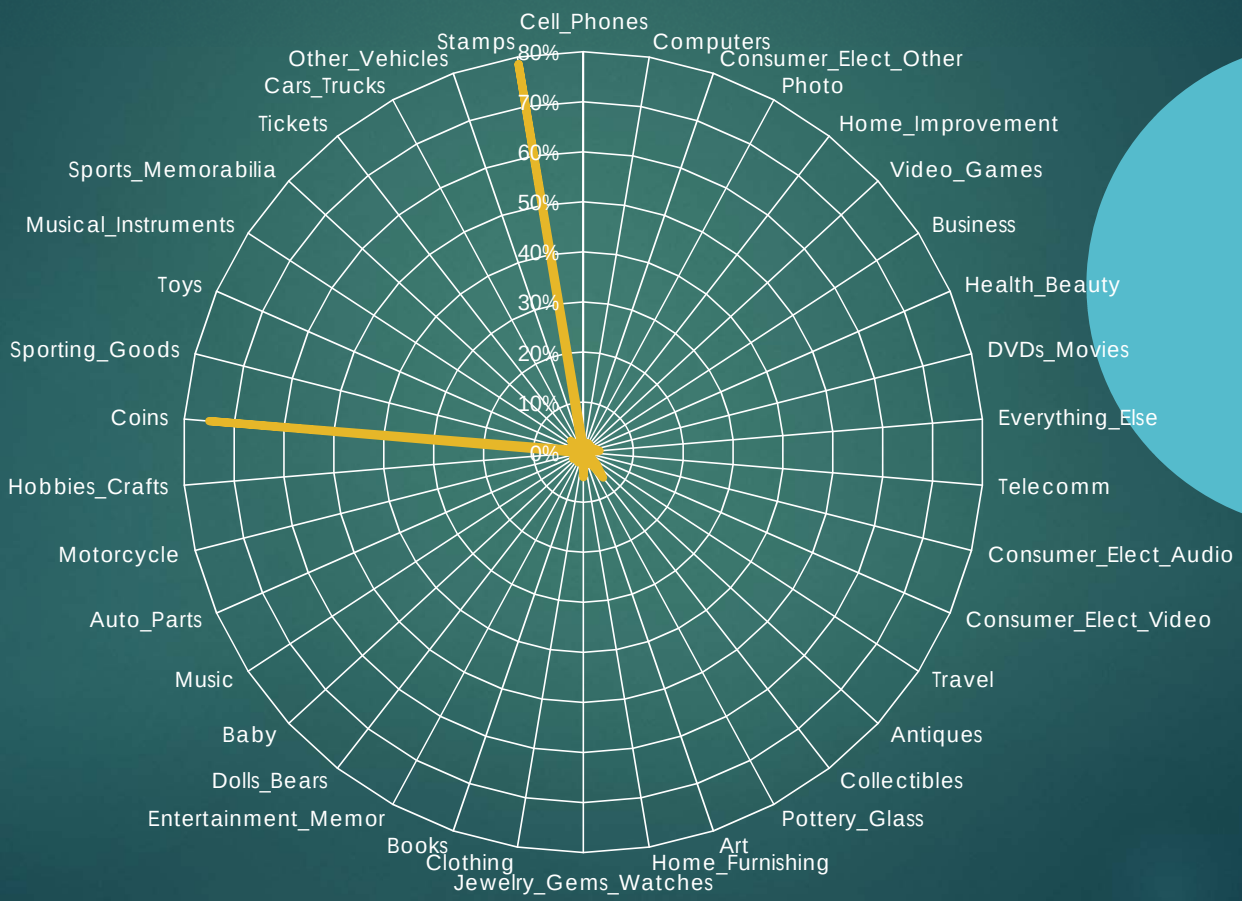
- 18 Main clusters
- ~ 60 sub-segments
- Capturing several dimensions of consumer behavior:
  - Pure buyer vs. Buyer/Seller
  - eBay positive bias, neutral, negative bias
  - Multi category purchaser vs. single category behavior
  - Association between different verticals
  - Frequency of purchase
  - Big ticket items

Some cluster display rich multi category behavior



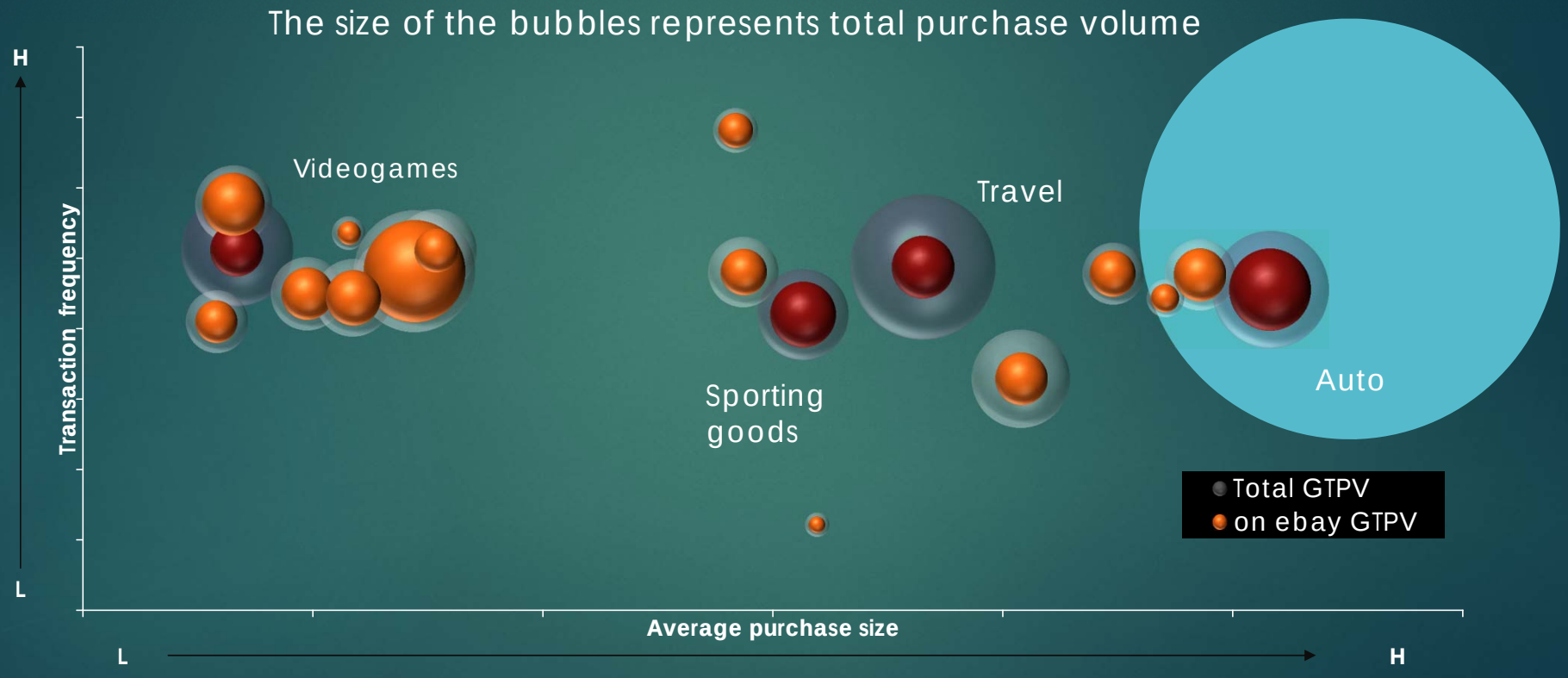


Other clusters display very narrow skew



Factor11

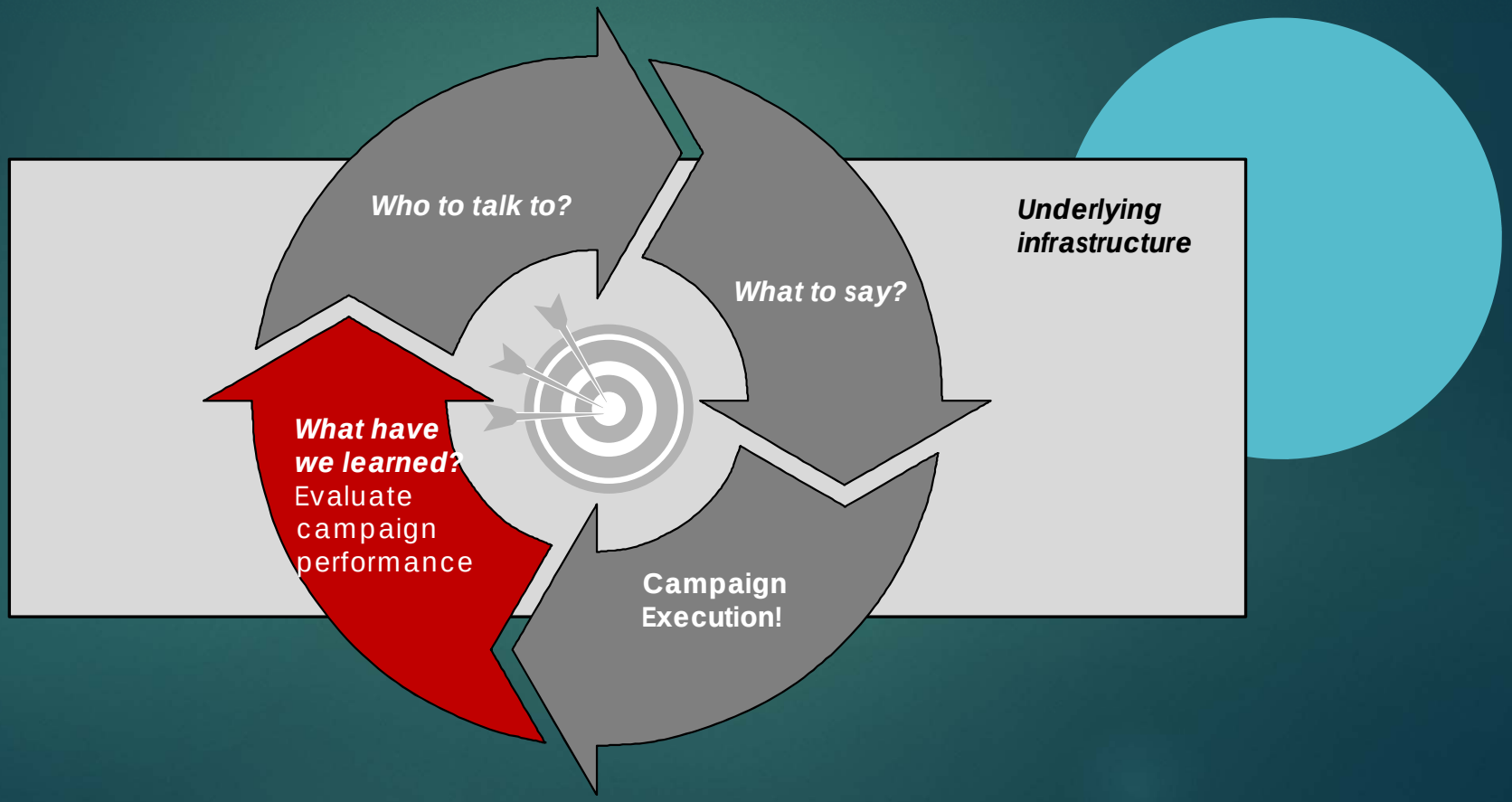
Visualizing cluster characteristics



Take-aways

- Focus on a genuine business need:
  - Promote merchants/brands/products which are most relevant.
  - Empower the consumer engagement team with actionable insights about customer tastes and preferences
  - Help marketing professionals visualize and understand complex multivariate relationships
- Great data infrastructure leads to reliable and scalable results with shorter development cycles
- Off the shelf algorithms can do a great job if you have the right data

What have we learned?



What have we learned: how do we analyze campaign performance

- Pre vs. post. Actual vs. forecast performance
- Test and control, a very good starting point
- PayPal's business need:
  - Understand the drivers of campaign performance
    - Are we driving more customers to purchase, or simply getting more purchases from active customers? Are we pushing larger basket sizes?
  - Understand campaign outcome on different customer groups: high value/ engaged vs. medium or low engagement customers
- Solution: Build program level randomized holdout as well as E.D. for individual campaigns. Use regression techniques to read treatment effects on multiple metrics: activation, avg txn number, etc

Key underlying infrastructure

- Ability to generate and persist a program level random holdout as well as experimental designs for individual campaigns
  - Well staffed and strongly motivated analytics team
  - Enterprise committed to continuous measurement and improvement.
- Customer fact DataMart
  - All relevant customer events tracked over time
- Transaction fact DataMart
  - Complete description of transaction attributes

Which tools /techniques did we use

- Teradata SQL to generate customer list and campaign files.
- SAS to generate stratified random samples
  - Proc Survey Select
- R and Stata
  - Instrumental Variables and 2SLS
  - Logistic regression, Selection models and Zero inflation models

- High quality enterprise software as well as open source free software
- Expensive enterprise hardware, med/low cost server and pcs.

Does the campaign have greater effect on engaged or marginal customers?

- Engaged customers are more likely to open our emails.
- So are opportunistic shoppers
- We cannot simply compare openers to non-openers or individuals opening say 3+ emails to rest as those comparisons are biased!
- Use random assignment to test and holdout as an instrument and run 2SLS

What makes a good instrument

- Let Z_i be our instrument and D_i be our variable of interest
  - Cov(Zi, $error_i$) = 0
  - Cov(Zi, Di) <> 0

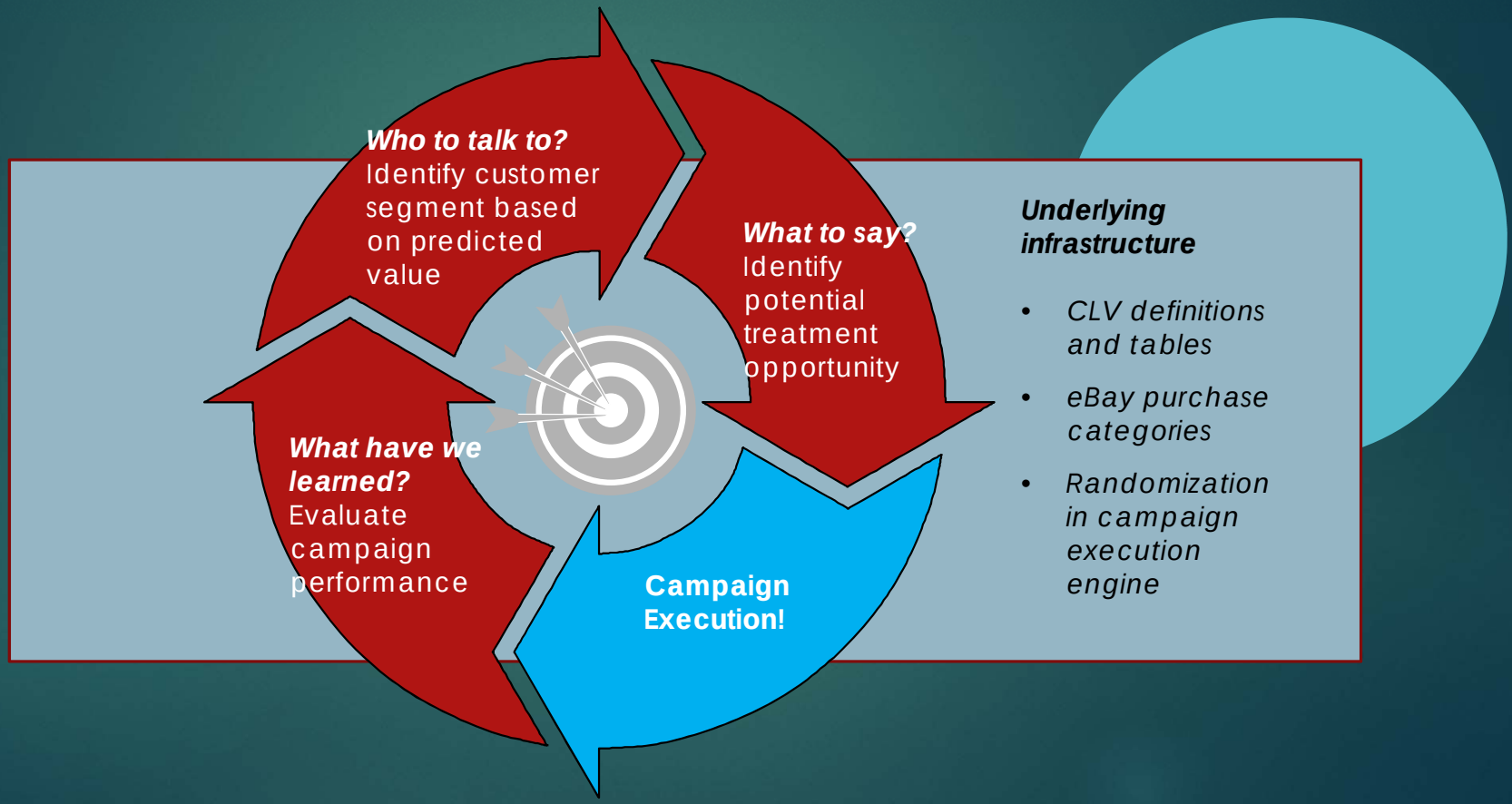
Use 2SLS to get an unbiased estimate of treatment effect

- Use our instrument, i.e. the randomly assigned treatment/ holdout group as well as our covariate set to fit *num_open* i.e. the number of emails opened.
  - $num\_open_i = \pi_0 + \pi_1 treat_i + X\pi + error_i$
- In the second stage we use the our fitted estimate of *num_open* and regress on transactions per active
  - $Txn\_act_i = \beta_0 + \beta_1 \hat{num\_open}_i + x\beta + error_i$

Take-aways

- Focus on a genuine business need:
  - Understand the drivers of campaign performance
    - Are we driving more customers to purchase, or simply getting more purchases from active customers? Are we pushing larger basket sizes?
  - Understand campaign outcome on different customer groups: high value/ engaged vs. medium or low engagement customers
  - There is a concave relationship between engagement and the lift from marketing programs
    - Medium engagement customers generate most of the value

One more time



Q1: Estimate elasticity to price

- One of the fundamental problems we are trying to solve is how to price a ride. How should the fair price of a ride be determined. Currently the price structure has two main components; a base price calculated on a per hour and per mile basis, and a surcharge which is generated dynamically based on supply demand imbalances. The computation of the surcharge is 'efficient' but is very much a local optimization and dependent on the overall price level. The base price, in particular, is a product of legacy pricing decisions made in the past, and of responses to competitive pressure from other ride-share services. What we would like to do is estimate price response functions by city that help us understand how demand is affected by changes in base price. This analysis should also help us understand the degree to which we are not simply providing an alternative to other ride-share options or traditional cab companies but transportation options in general ,including public transportation and driving your own vehicle.

Q2: Model the ride ecosystem

- Ride-sharing is essentially a two sided network comprising passengers demanding rides and drivers supplying those rides. More passenger i.e. greater demand leads to more drivers joining the network, and in turn more drivers means greater availability and better standard of service which then draws in more demand and so on. It is in theory a virtuous cycle but it is open to many hiccups. In particular large excesses of either supply or demand can lead to very negative experiences for both would be passengers and drivers and can lead to individuals abandoning the network. One way to model this micro market is to think of a production function whose output is rides and whose inputs are passenger demand for rides and available driver hours supplied. For any fixed number of rides there are isoquants representing different combination of demand and supply pairs that will produce that level of rides. However only a subset of those combinations will be optimal from the point of view of positive experience for both side of the network and for its continued growth. We want to be able to model more rigorously such production functions and compute optimal operating parameters. In particular we want to be able to do so at a city level and incorporate city specific characteristics that might make different operating parameters optimal or perhaps unfeasible in different environments.

Q3: Taxonomy of rides

- A very successful and much practice method of classifying consumers and understanding consumer behavior is based on grouping consumer by the type of goods they purchase. This method relies on having a meaningful taxonomy of the different product being sold and consumed. For a retailer this is a potentially cumbersome but rather straightforward task as most products produced and sold fall in well defined categories. What our company produces are rides, which are not however so easily classified. What we would like to accomplish as a foundational step is create a meaningful taxonomy of the different type of rides that passengers take. In particular we believe that tagging destination and point of origin as well as the time of day when the ride occurred, and perhaps its duration or route, would provide a meaningful basis for a taxonomy.

Q4: Identify the persuadables

- Our company, like many other organizations, relies to some extent on monetary incentives as levers to stimulate both demand from would be passengers, and availability from would be drivers. Given the sums of money that are at times involved it is imperative that these expenditures are as efficient as possible. We have executed and continue to field a number of test to measure the incremental impact of those initiatives. However the outcome of these test is generally somewhat binary i.e. the promotion generate positive value/ is break even or instead it generates losses. What we would like to do is to use this data to go a step further and understand which individuals are more likely to display 'incremental' behavior because of the promotion i.e. identity the 'persuadable' and focus our incentives on these individuals. This would also allow us to exclude from promotions individuals where promotional incentive would most likely lead just to cannibalized revenue.

Q5: Scalable test design and inference

- As most enterprises we rely on experimental designs to optimize the features of our mobile app and our website, as well as the array of market incentive we provide to our customers. These activities requires both an effective design of potentially complex experiments as well as the ability to perform valid statistical inference on the results. We would like to work with subject matter experts to ensure that both the design of our experiments as well as the procedures used to read them are statistically sound and optimal from the point of view of maximizing readable signal from the available samples of mobile/website events and customer-promotion interactions.