Data Collection for Business Analytics

About Data: Introduction and Overview

Session 1 @ CBA Batch 12

Oct 2019

Sudhir Voleti

ISB

Why care about DC?

- DC is about Data + Collection.

- What challenges might you face in collecting data?
- Knowing *what* data to collect
- → Hunting for data sources
- → mining raw data
- → assessing data quality
- → Processing and transforming the raw data
- → judging business relevance
- → budgeting cost and time
- → Estimating data value
- Etc.

The Age of Data

"If *Land* was the primary raw material of the agricultural age,

and *Iron* that of the industrial age,

then *Data* is the primary raw material of the information age."

Session Outline

- A Motivating Example
  - Data, value and valuations - the Uber example.

- Preliminaries
  - Anatomy of a business, Nature of Analytics

- Data and Measurement
  - Measurement and the Theory of Scales
  - Data Types and Data Dichotomies

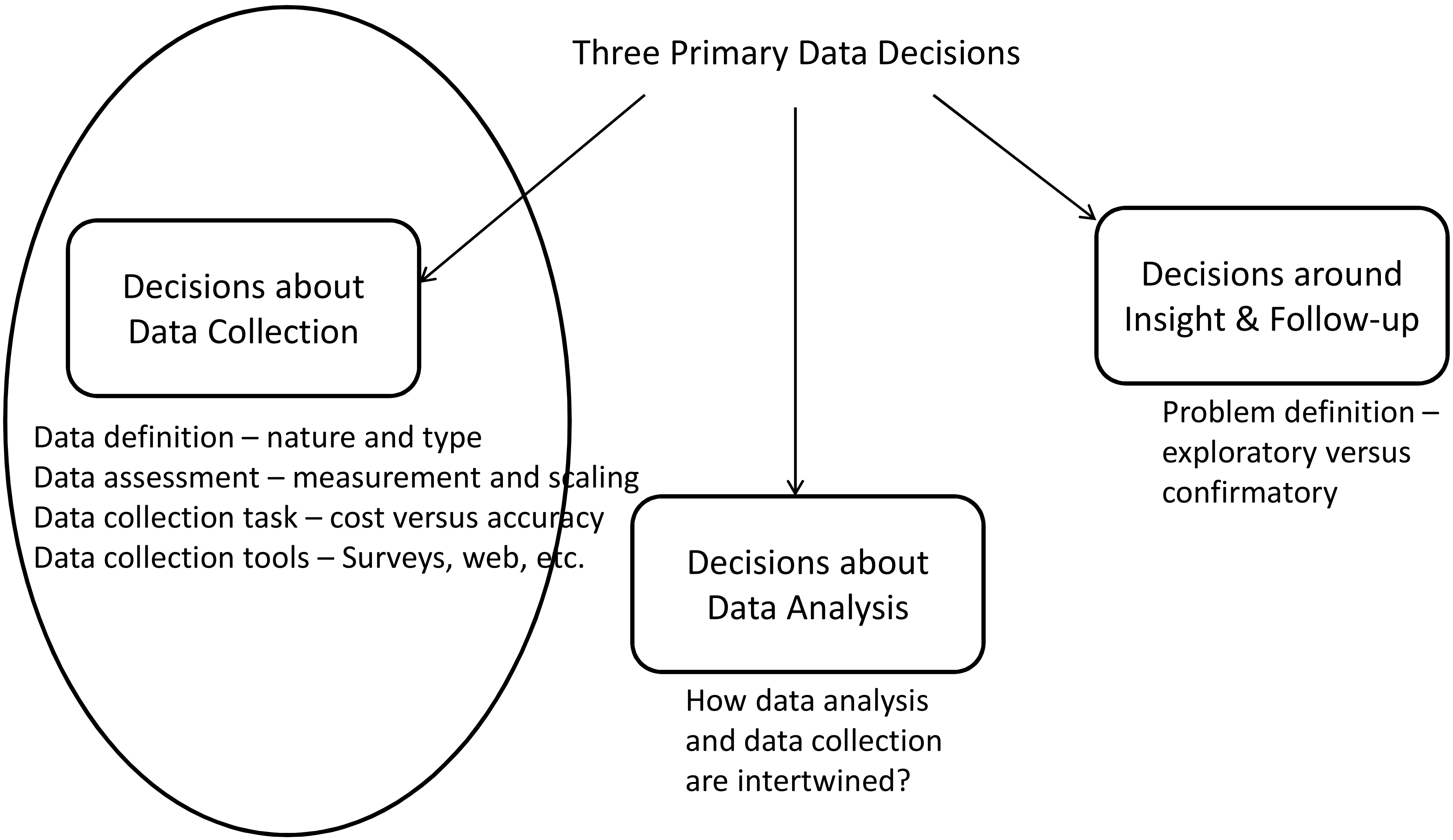
- Basic structure of [Traditional] Survey Research
  - Perceptual Mapping using Survey data and shinyapps

- Session Wrap-up

Some Preliminaries

DC’s Intended Scope

Decisions about data must be made.



Three Primary Data Decisions

Decisions about Data Collection

Data definition – nature and type
Data assessment – measurement and scaling
Data collection task – cost versus accuracy
Data collection tools – Surveys, web, etc.

Decisions about Data Analysis

How data analysis and data collection are intertwined?

Decisions around Insight & Follow-up

Problem definition – exploratory versus confirmatory

About me…

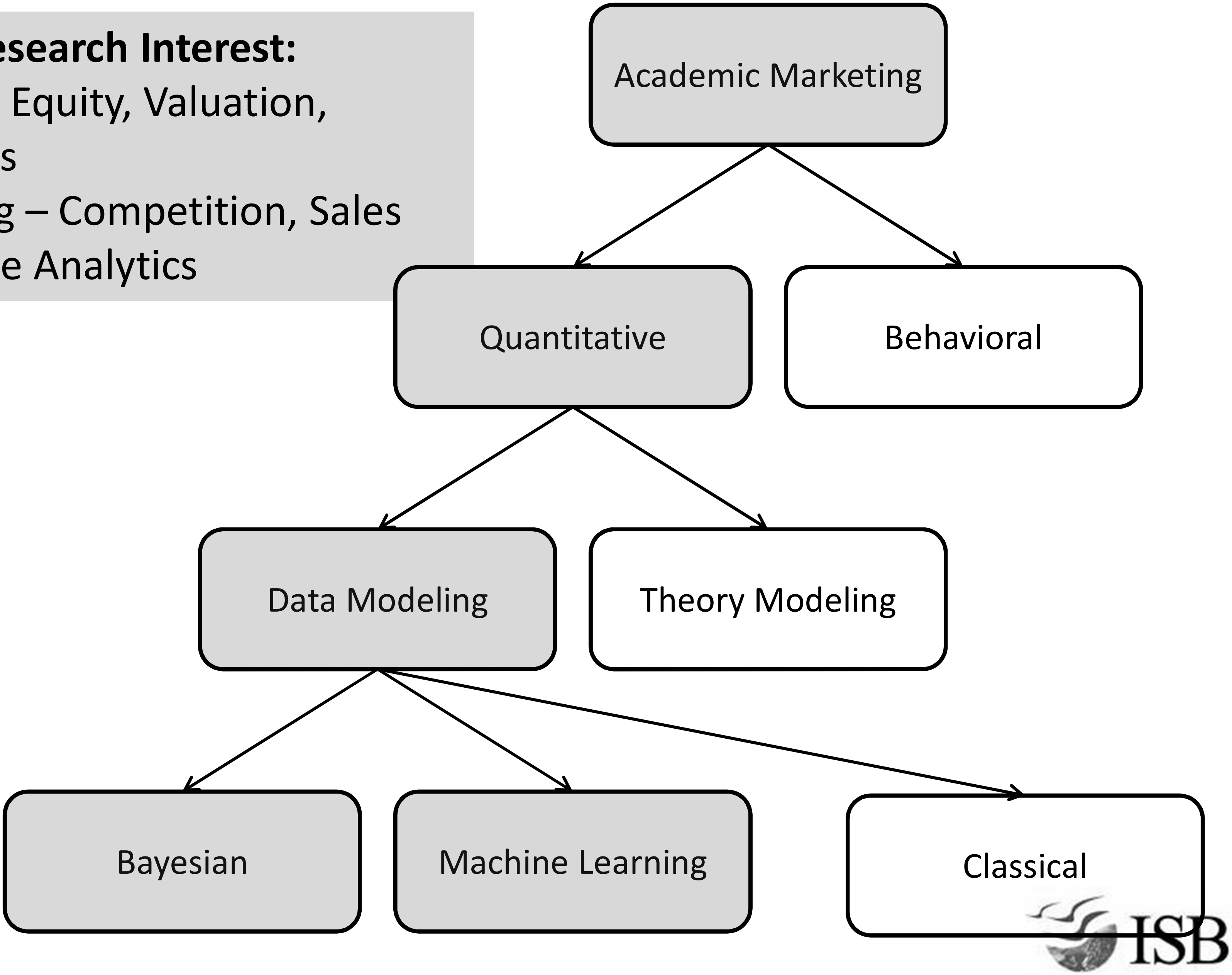
- **Academic Credentials:**
  - PhD in Marketing – Univ of Rochester (2009)
  - MS in Applied Statistics – Univ of Rochester (2006)
  - PGDM – IIM Calcutta (2001)
  - B.E. – BIT Mesra (1998)

- **Industry Experience:**
  - Software Programmer with Cognizant 1998-99
  - Management Consultant with Accenture 2001-02
  - Data Analyst – Daymon Consumer Insights Division 2006-08
  - Academic Faculty with ISB – 2009 onwards
  - Been involved in a Tech Startup – Modak Analytics – 2012

About my Research…

Topics of Research Interest:

1. Brands – Equity, Valuation, Dynamics
2. Modeling – Competition, Sales
3. Predictive Analytics



Announcements

- DC will be more of a *training workshop* than a regular lecture based course.
  - Syllabus outline was tentative, there may be a few changes to it.

- I’ll assume you:
  - [1] will install the requisite [open-source] software,
  - [2] have your own Github pages,
  - [3] have no prior exposure to DC.

- *Primers* will be conducted, as required.

- Assignments will be there and a final exam.

- Qs, feedback etc are welcome.

Course Preliminaries

- Is your R and Rstudio installed and ready to go?

- Installed Python 3.x and Spyder (or Jupyter) as well?
  - I prefer Py’s Anaconda distribution

- Downloaded and ready with materials for today’s session?

- Some of what follows may seem terribly basic to some veterans of R and/or Py. So be it.

- Ready to start?

Conceptual Preliminaries: Basic Concepts

- This is a course on *Business Analytics*.
- Q1. What is a ‘*business*’?
- Q2. What is the nature of ‘Analytics’?
  - Art? Craft? Science? Magic?
- Q3. What are the implications of the answers to the above?

Prelimir

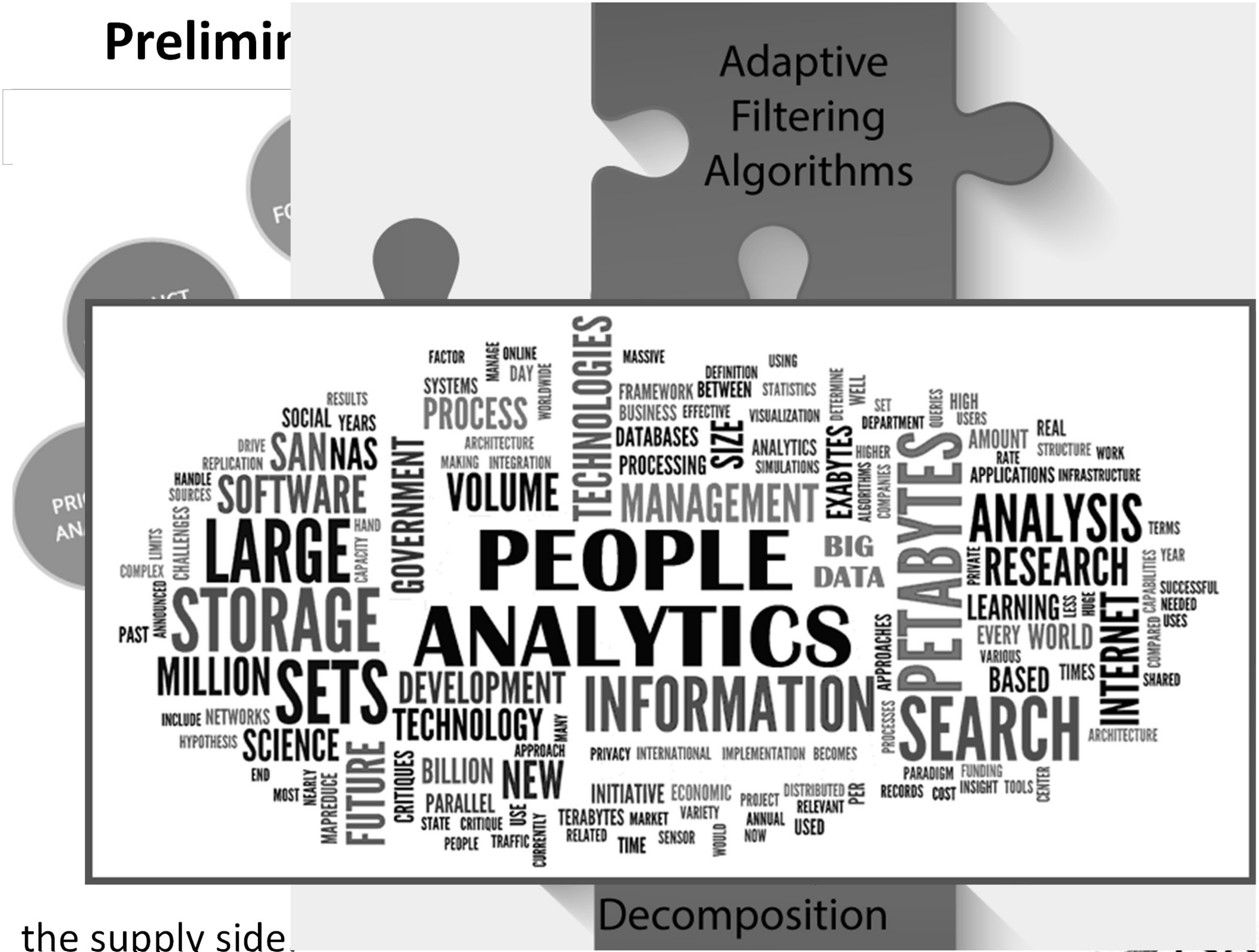
-

-

-

-

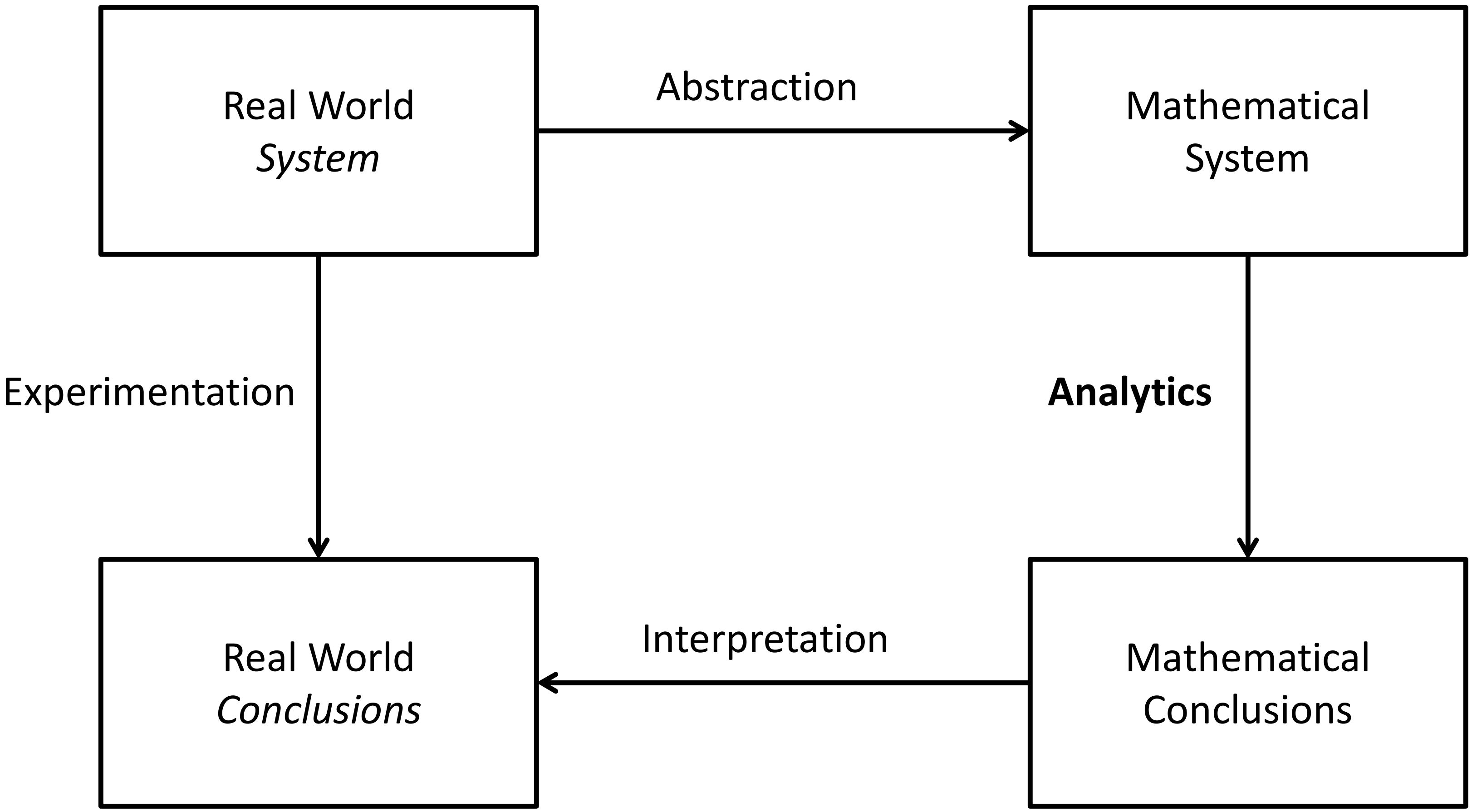
the supply side.


Adaptive Filtering Algorithms
Decomposition
FO
PRI AN
FACTOR ONLINE MANAGE DAY WORLDWIDE MASSIVE USING DEFINITION
SYSTEMS PROCESS TECHNOLOGIES FRAMEWORK BETWEEN STATISTICS DETERMINE WELL SET DEPARTMENT QUERIES HIGH USERS
RESULTS SOCIAL YEARS BUSINESS EFFECTIVE VISUALIZATION
DRIVE SAN NAS ARCHITECTURE DATABASES ANALYTICS HIGHER AMOUNT RATE REAL STRUCTURE WORK
REPLICATION MAKING INTEGRATION PROCESSING SIZE SIMULATIONS ALGORITHMS COMPANIES APPLICATIONS INFRASTRUCTURE
HANDLE SOURCES SOFTWARE GOVERNMENT VOLUME MANAGEMENT EXABYTES PETABYTES ANALYSIS
LIMITS CHALLENGES LARGE HAND CAPACITY PEOPLE BIG DATA PRIVATE RESEARCH TERMS YEAR
COMPLEX STORAGE ANALYTICS LEARNING LESS HUGE INTERNET CAPABILITIES SUCCESSFUL NEEDED USES
PAST ANNOUNCED EVERY WORLD VARIOUS COMPARED
MILLION SETS DEVELOPMENT INFORMATION APPROACHES BASED TIMES SHARED
INCLUDE NETWORKS TECHNOLOGY MANY PROCESSES SEARCH ARCHITECTURE
HYPOTHESIS SCIENCE APPROACH PRIVACY INTERNATIONAL IMPLEMENTATION BECOMES
END NEARLY FUTURE CRITIQUES BILLION NEW INITIATIVE ECONOMIC PARADIGM FUNDING
MOST MAPREDUCE PARALLEL USE PROJECT DISTRIBUTED PER RECORDS COST INSIGHT TOOLS CENTER
STATE CRITIQUE TERABYTES MARKET VARIETY ANNUAL RELEVANT
PEOPLE TRAFFIC CURRENTLY RELATED SENSOR WOULD NOW USED TIME

Preliminaries: The Anatomy of Analytics

Problem definition is critical here

Domain Knowledge is critical here



Real World *System*

Abstraction

Mathematical System

Experimentation

Analytics

Real World *Conclusions*

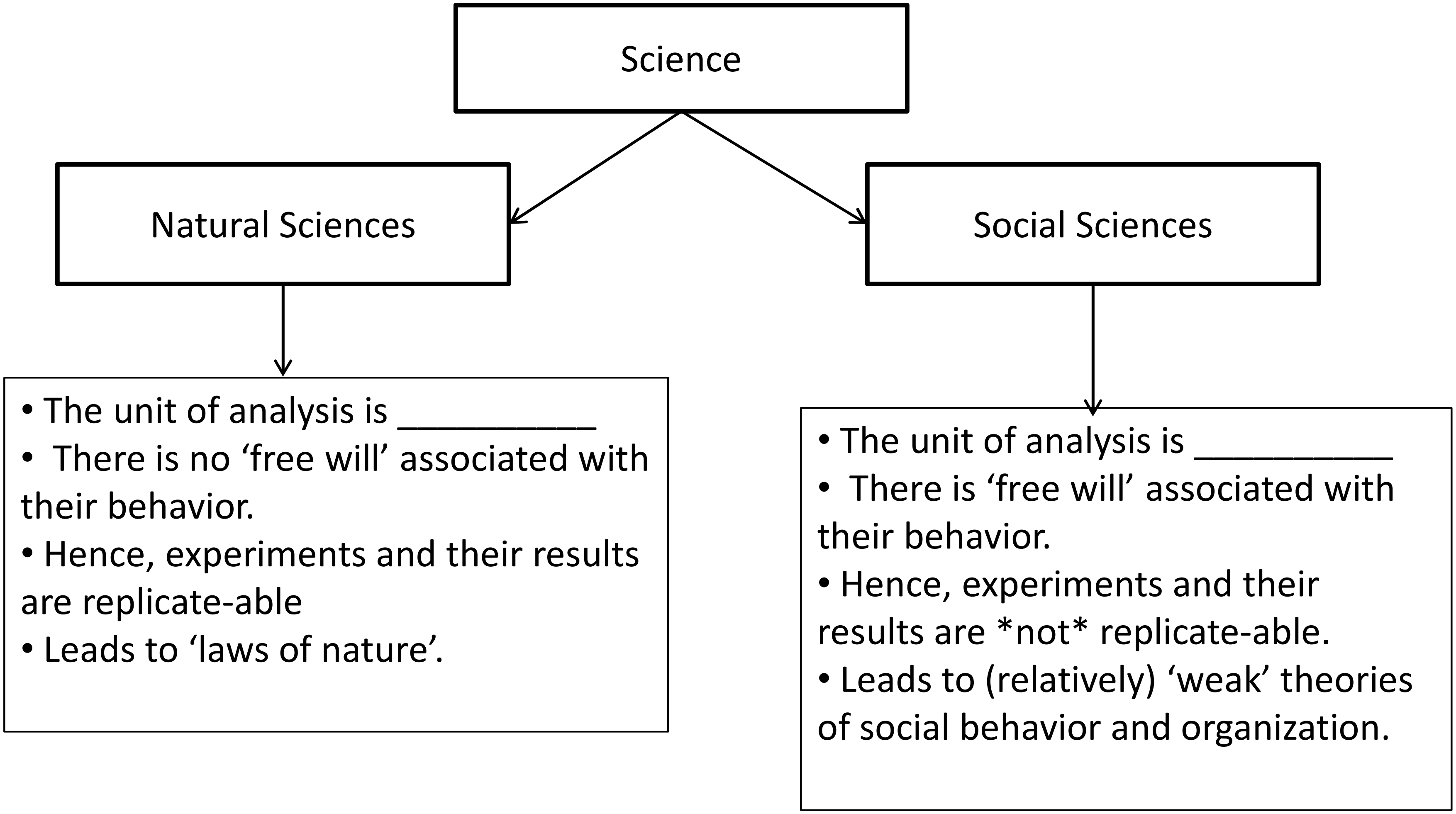
Interpretation

Mathematical Conclusions

Translation and Implementation are key here

Inference capabilities are key here

Preliminaries: Is ‘Analytics’ Scientific?



Science

Natural Sciences

Social Sciences

- The unit of analysis is ___________
- There is no ‘free will’ associated with their behavior.
- Hence, experiments and their results are replicate-able
- Leads to ‘laws of nature’.

- The unit of analysis is ___________
- There is ‘free will’ associated with their behavior.
- Hence, experiments and their results are *not* replicate-able.
- Leads to (relatively) ‘weak’ theories of social behavior and organization.

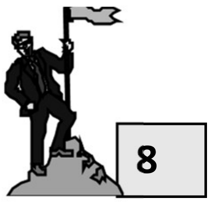
Bottomline: There’s only so much **precision** in our **measurements** and our results that we can expect.

Theory of Scales

Four Feature Types

- There are 4 types of Data based on the quality of information contained and corresponding to these are 4 primary scales.
- **Nominal**
  - Merely labels. No further information can be gleaned.
  - Example: “Coke” and “Pepsi”.
- **Ordinal**
  - Conveys only upto preference information. Direction alone.
  - Example: “I prefer Coke to Pepsi”.
- **Interval**
  - Conveys relative magnitude information, in addition to preference.
  - Example: “I rate Coke a 7 and Pepsi a 4 on a scale of 10”.
- **Ratio**
  - Conveys information on an absolute scale.
  - Example: “I paid Rs 11 for Coke and Rs 12 for Pepsi”.

Primary Scales of Measurement

Scale					
Nominal	Numbers Assigned to Runners				Finish
Ordinal	Rank Order of Winners	Third place	Second place	First place	Finish
Interval	Performance Rating on a 0 to 10 Scale	8.2	9.1	9.6	
Ratio	Time to Finish, in Seconds	15.2	14.1	13.4	

Types of Scales: Examples of Common Analysis

NOMINAL	ORDINAL	INTERVAL	RATIO
Mode	Mode	Mode	Mode
Frequencies	Median	Median	Median
Percentages	Frequencies	Mean	Mean
	Percentages	Frequencies	Frequencies
	Some Statistical Analysis	Percentages	Percentages
		Variance	Variance
		Standard Deviation	Standard Deviation
		Most Statistical Analysis	Ratio of numbers
			All Statistical Analysis

Q-Quickfire Question

- Mr Fernando measures favorability of the Airtel brand on a 1-5 scale (higher means more favorable). Jai gives Airtel a 2 whereas Aditi gives it a 4.

- Which of the following statements hold true.

- (A) Airtel is twice as much favored by Aditi as Jai.
- (B) The difference between Jai's and Aditi's ratings is 2 points.
- (C) Jai is not favorably inclined towards Airtel. Aditi is.
- (D) On a 1-9 scale, Jai would have given 4 & Aditi would have given 6.
- (E) Can't say. It depends.

Q-Quickfire Question

- Mr Fernando measures Airtel usage time in minutes/day. Jai reports an average of 20 minutes whereas Aditi reports an average of 40 minutes.
- Which of the following statements hold true.
- (A) Airtel is used twice as much by Aditi as by Jai.
- (B) The difference between Jai’s and Aditi’s avg usage is 20 minutes.
- (C) Aditi uses Airtel more than Jai on any given day.
- (D) Aditi’s Airtel bill is higher than Jai’s.
- (E) Can’t say. It depends.

Q - Quickfire Question

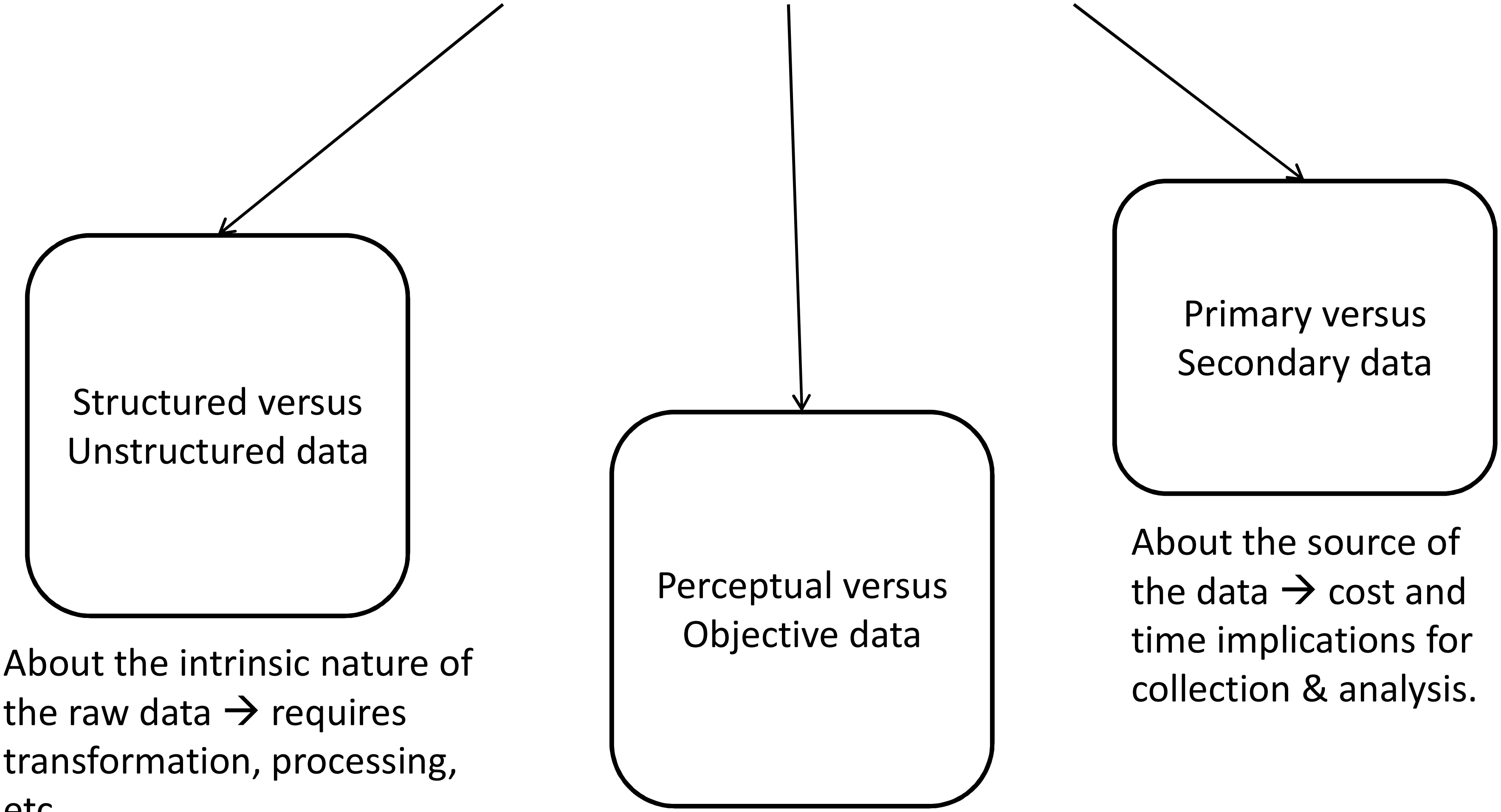
- Which of the following data are (i) Nominal, (ii) Ordinal, (iii) Interval, and (iv) Ratio. Choose the most informative description for each of the items below.

- (A) Passport numbers.
- (B) Quality rankings.
- (C) Social class categorization (‘lower’, ‘middle’, ‘upper’ class).
- (D) Market share.
- (E) Store formats (‘Food stores’, ‘drug stores’, ‘mass merchandisers’, ‘online stores’ etc.).
- (F) BSE sensex levels

Measurement Basics:

Data Types and Data Dichotomies

3 Basic Data Dichotomies



Structured versus Unstructured data

About the intrinsic nature of the raw data → requires transformation, processing, etc.

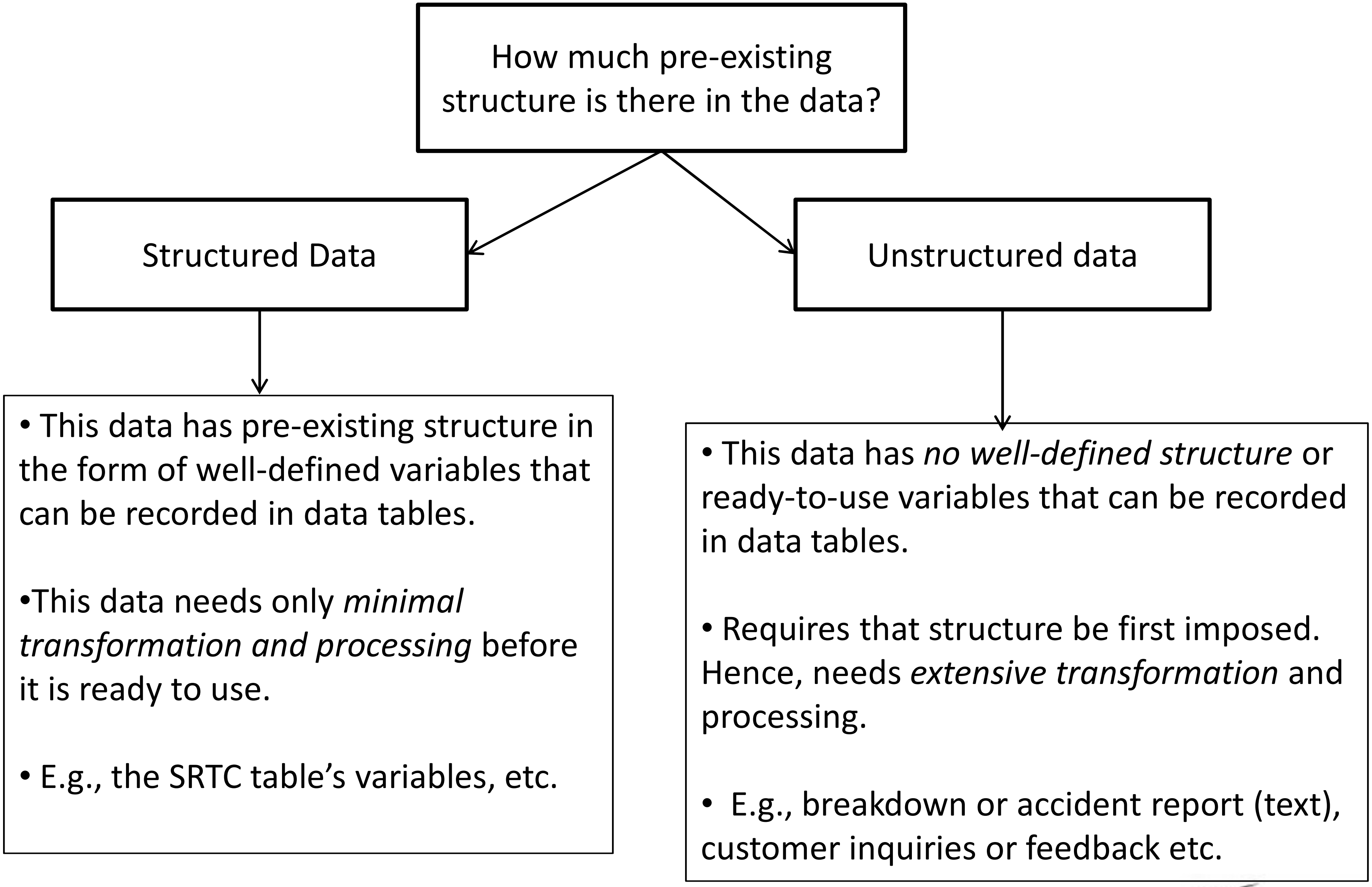
Perceptual versus Objective data

About whether data collected is subjective or objective → implications for measurement and for analytics

Primary versus Secondary data

About the source of the data → cost and time implications for collection & analysis.

The Structured Vs Unstructured Data Dichotomy



How much pre-existing structure is there in the data?

Structured Data

- This data has pre-existing structure in the form of well-defined variables that can be recorded in data tables.
- This data needs only *minimal transformation and processing* before it is ready to use.
- E.g., the SRTC table’s variables, etc.

Unstructured data

- This data has *no well-defined structure* or ready-to-use variables that can be recorded in data tables.
- Requires that structure be first imposed. Hence, needs *extensive transformation* and processing.
- E.g., breakdown or accident report (text), customer inquiries or feedback etc.

Quick Q on Structured vs Unstructured Data

- Which of the following data are Structured data - i.e., can directly be used as variables in a dataset? Why or why not?

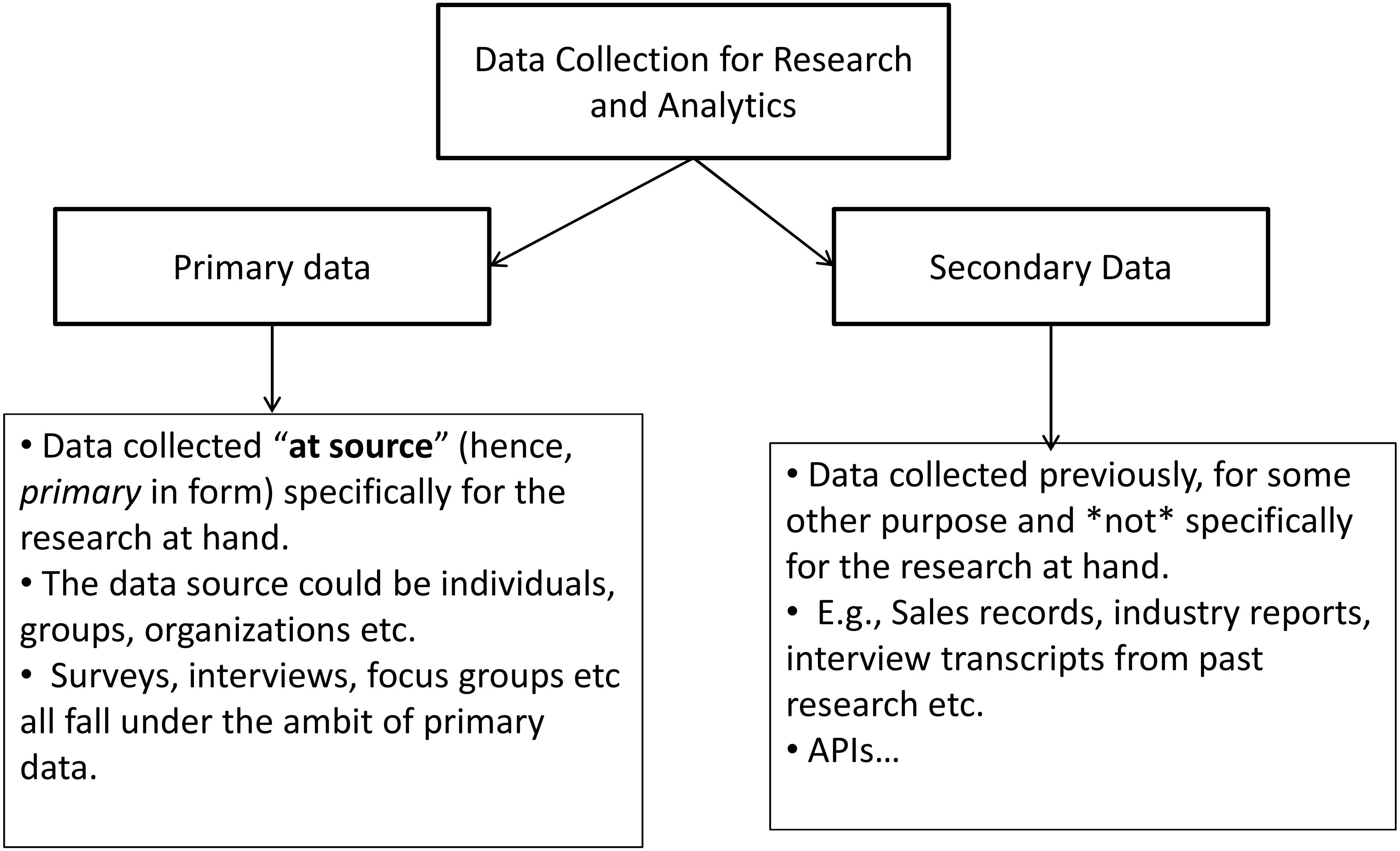
- (a) Aadhaar fingerprints
- (b) PAN number
- (c) Address on the ration card
- (d) Jan dhan account number
- (e) Scheduled versus actual departure of APSRTC buses
- (f) date of birth on school certificate
- (g) photo on the passport

Perceptual versus Objective data

- **Perceptual Data:**
- Subjective data - about which two people can reasonably disagree.
- E.g., I give Virat Kolhli a 8/10, you give him a 7/10.
- Usually about people's perceptions of quality, service, performance, etc.
- Usually compared to some reference or prior expectations.

- **Objective data:**
- Facts that are independent of subjective perception.
- E.g., Virat's strike rate is 83.3.
- Usually about events measured in physical attributes, space, mass, time etc.

The Primary Vs Secondary Data Dichotomy



Data Collection for Research and Analytics

Primary data

Secondary Data

- Data collected “**at source**” (hence, *primary* in form) specifically for the research at hand.
- The data source could be individuals, groups, organizations etc.
- Surveys, interviews, focus groups etc all fall under the ambit of primary data.

- Data collected previously, for some other purpose and *not* specifically for the research at hand.
- E.g., Sales records, industry reports, interview transcripts from past research etc.
- APIs…

Basic Structure of Survey Research:

A Conceptual Primer

Survey Principles: Introduction to Survey Research

- What is survey research?
- What are its main *components*?

- The **conjunction** of a certain kind of ________________ with a certain approach to _____________ constitutes Survey Research as a distinct Mktg Research (MKTR) tool.

- Why care about surveys?
- What are surveys *best* at?
- Surveys *sample* respondents with an intention to *project* responses onto the larger population.

Survey Principles: Survey Research as a Descriptive Tool

- What do survey results look like typically?

  - A **percentage figure** ("35% of home PC owners are dissatisfied with their ISP.")
  - A **frequency count** ("On the average, a household buys toothpaste once in 3 months.")
  - A **cross-tabulation** ("47% of car-owners have our product whereas only 12% of bike-owners do.")

- Could these descriptive estimates relate to *Market size estimation*? How?

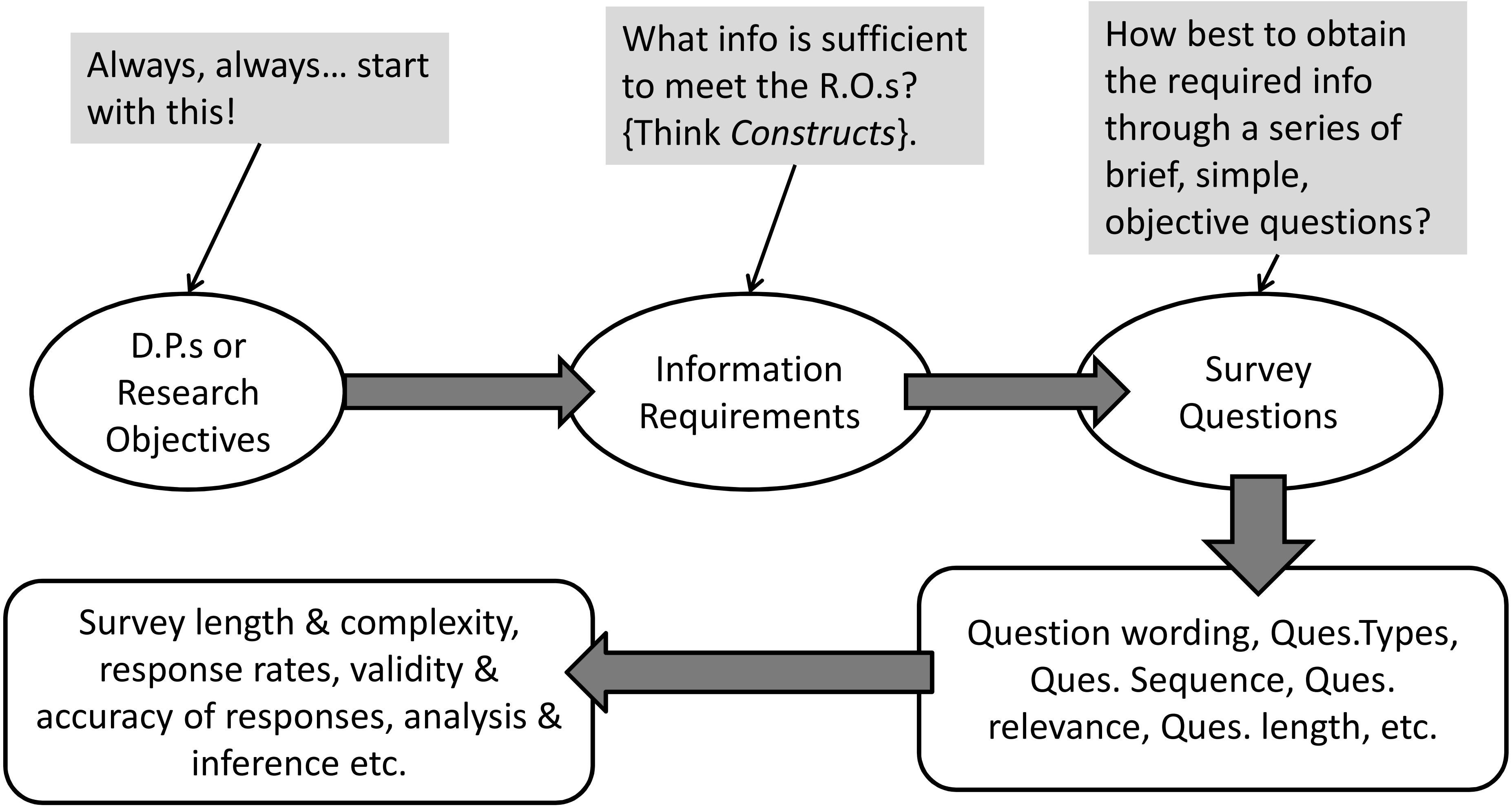
- If you want *prediction* not *description*, would you still use surveys? Why (not)?

Survey Principles: Q1 – Matching Tools to Situations

- Decide if Survey research best fits the following situations? Why?

- (i) “If I lower prices by 10%, by how much will my sales rise?”
- (ii) “How many of our mobile customers want a more flexible rate plan?”
- (iii) “If the competitor launches a new rate plan, how many of our current customers will switch?”
- (iv) “Does our advertising connect at all with 16-24 yr olds?”
- (v) “What value added services can we sell along with our core product?”

Survey Design - Basic Principles



Always, always… start with this!

What info is sufficient to meet the R.O.s? {Think *Constructs*}.

How best to obtain the required info through a series of brief, simple, objective questions?

D.P.s or Research Objectives

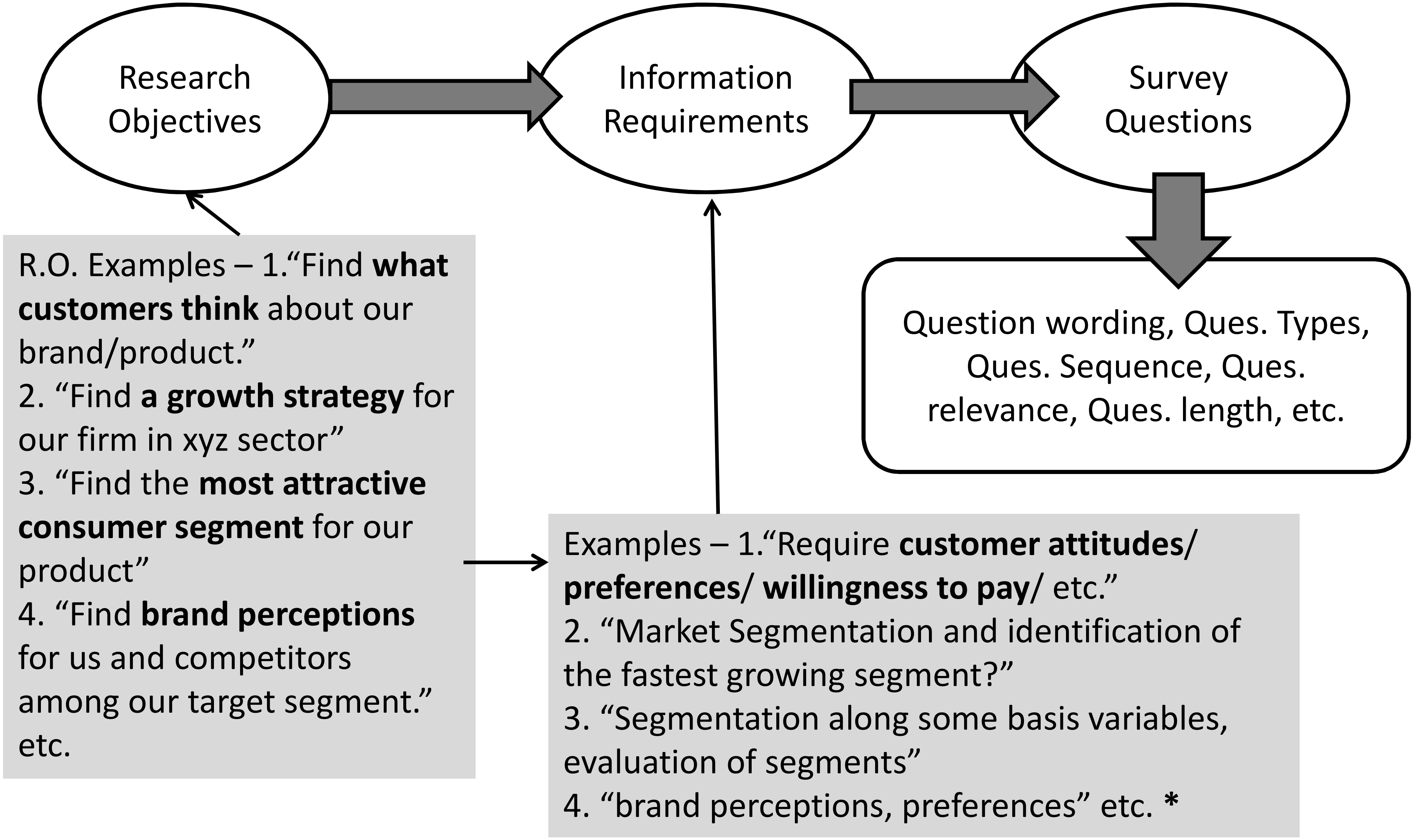
Information Requirements

Survey Questions

Question wording, Ques.Types, Ques. Sequence, Ques. relevance, Ques. length, etc.

Survey length & complexity, response rates, validity & accuracy of responses, analysis & inference etc.

Survey Design - Basic Principles



Research Objectives

Information Requirements

Survey Questions

Question wording, Ques. Types, Ques. Sequence, Ques. relevance, Ques. length, etc.

R.O. Examples – 1.“Find **what customers think** about our brand/product.”
2. “Find **a growth strategy** for our firm in xyz sector”
3. “Find the **most attractive consumer segment** for our product”
4. “Find **brand perceptions** for us and competitors among our target segment.” etc.

Examples – 1.“Require **customer attitudes/ preferences/ willingness to pay**/ etc.”
2. “Market Segmentation and identification of the fastest growing segment?”
3. “Segmentation along some basis variables, evaluation of segments”
4. “brand perceptions, preferences” etc. *****

ISB

Survey Types & Use: Some Global Examples

- The *Net-Promoter Score* (NPS) - 1-dimensional summary response with a lot of diagnostic & predictive power.

Only 1 Q asked "How likely are you to recommend us?" on a 1-10 scale.
NPS = #Promoters - #Detractors

- Mobile-Surveys are the next frontier. When might they be more effective than websurveys?*

Capture customers' reactions in-situ rather than retrospectively, tailored to location and context.

Survey Types & Uses: Some Global Examples

- Active data collection on purchases - the Infoscout example.

Infoscout actively incentivizes consumers to scan & report shopping receipts of *every* purchase they make --> enabling attitudinal measurements alongside behavioral ones.

- [Actively] Tracking word-of-mouth dynamics - the Keller Fay example.

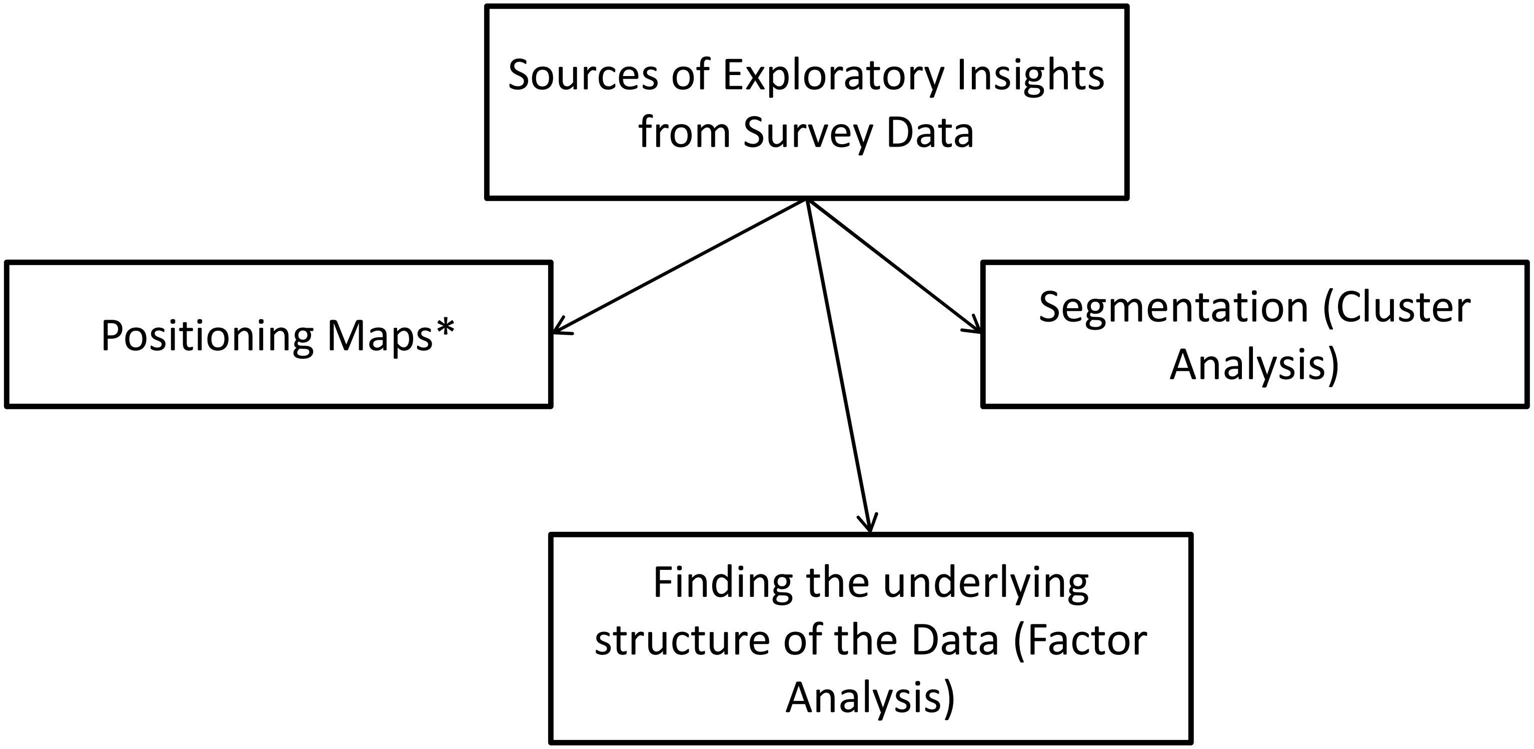
Incentivizes people to report whenever brands are mentioned in casual conversations via an app.

Points to Ponder:

Q: What kind of analytics is possible from the postmodern era survey DC?

Q: What kind of Qs can be asked?

What kind of (Exploratory) Insights can Survey Data yield?



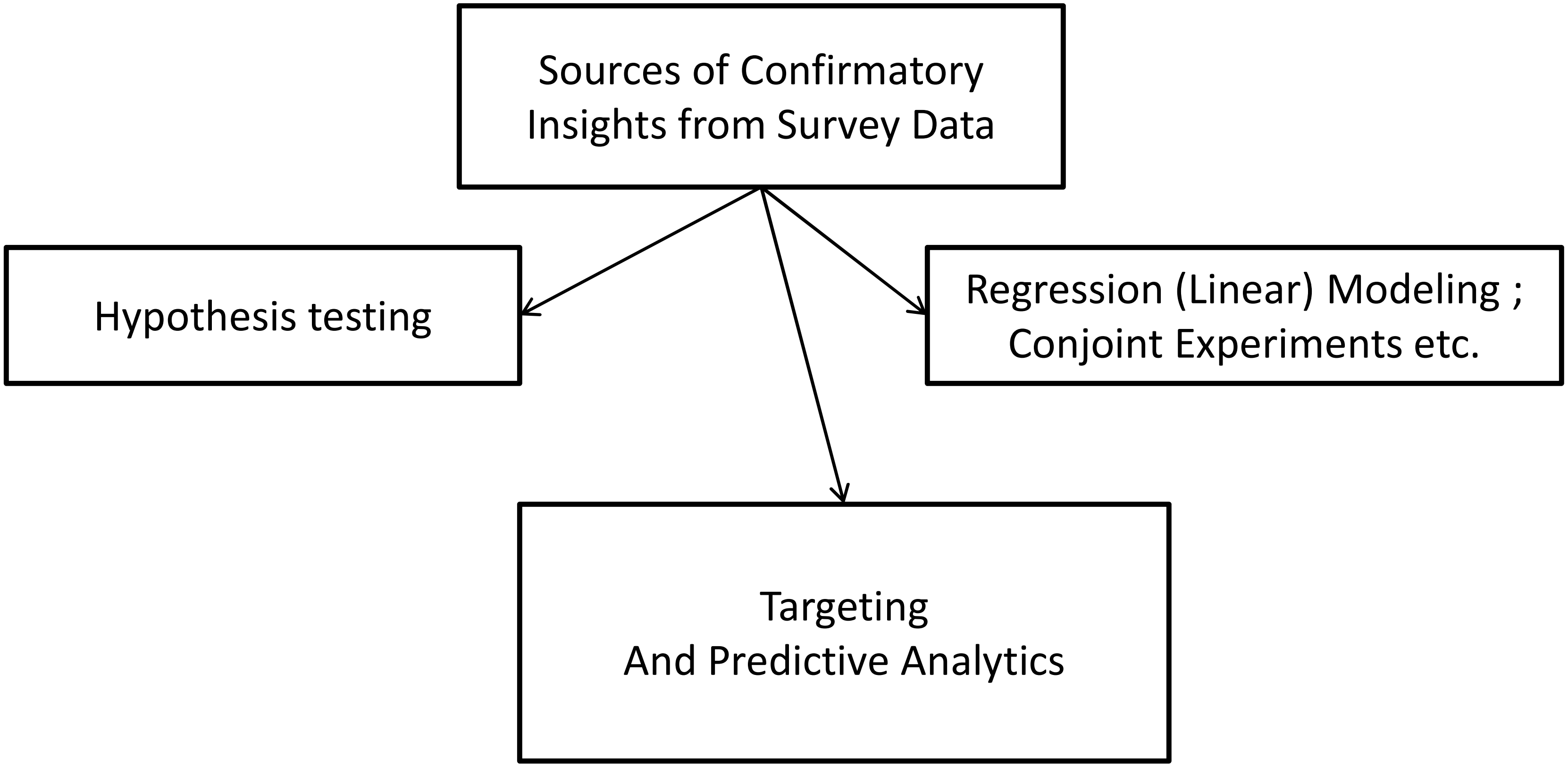
Sources of Exploratory Insights from Survey Data

Positioning Maps*

Segmentation (Cluster Analysis)

Finding the underlying structure of the Data (Factor Analysis)

What kind of Confirmatory Insights can Survey Data yield?



Sources of Confirmatory Insights from Survey Data

Hypothesis testing

Regression (Linear) Modeling ; Conjoint Experiments etc.

Targeting And Predictive Analytics

Survey Design: Basic Principles - Recap

- Is problem formulation a pre-requisite for survey design?

  Yes. Else, the object of interest wouldn't be known → which precludes all descriptive work

- What problem / research types would you use surveys for?

  Confirmatory. Descriptive. With well-defined DPs and ROs…

- What is the survey method's primary strength?

  Population level projections of response profiles

- Its primary weaknesses?

  Cost, time, complexity, inflexibility

Next up: A use-case involving primary DC via surveys, some math and some visualization for business insight. And shinyapps.

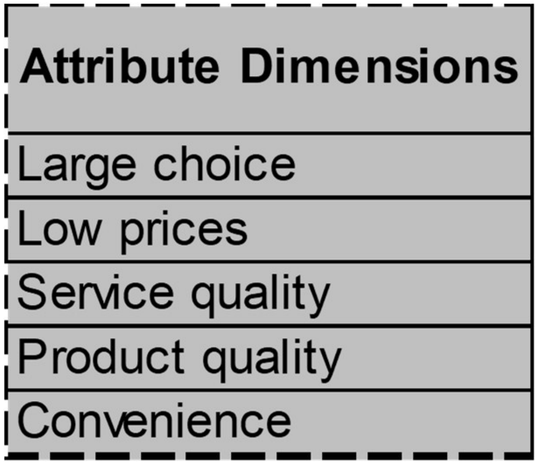
Perceptual Mapping
Using
Survey Data

[Officestar example via *shinyapps]*

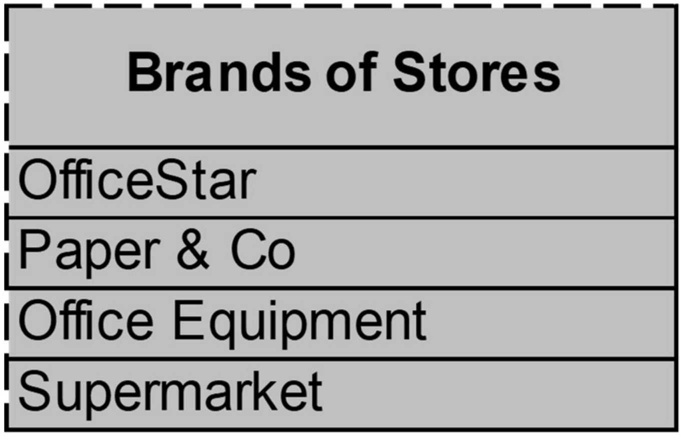
Officestar: About the Problem at Hand

- **D.P. 1:** How do our customers view our department store against our competitor stores [in the Office Supplies Business on 5 key dimensions.]?

- Say, the firm (Office Star) has 3 other competitors and has identified 5 dimensions it believes are key.



Attribute Dimensions
Large choice
Low prices
Service quality
Product quality
Convenience

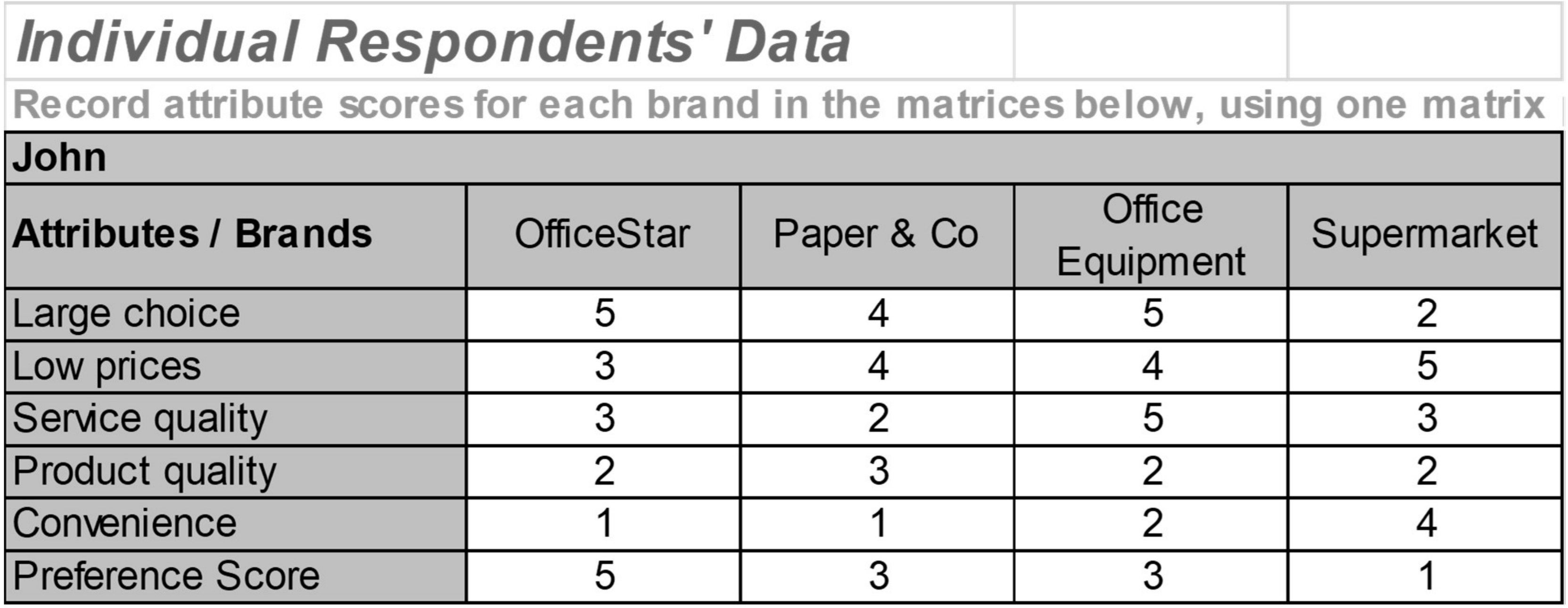


Brands of Stores
OfficeStar
Paper & Co
Office Equipment
Supermarket

- The challenge now is to collect data on how customers evaluate *each* brand on *each* attribute.

Officestar: Data Collection

- A survey is administered to target segment customers and perception data is collected in a matrix format thus:



Individual Respondents' Data

Record attribute scores for each brand in the matrices below, using one matrix

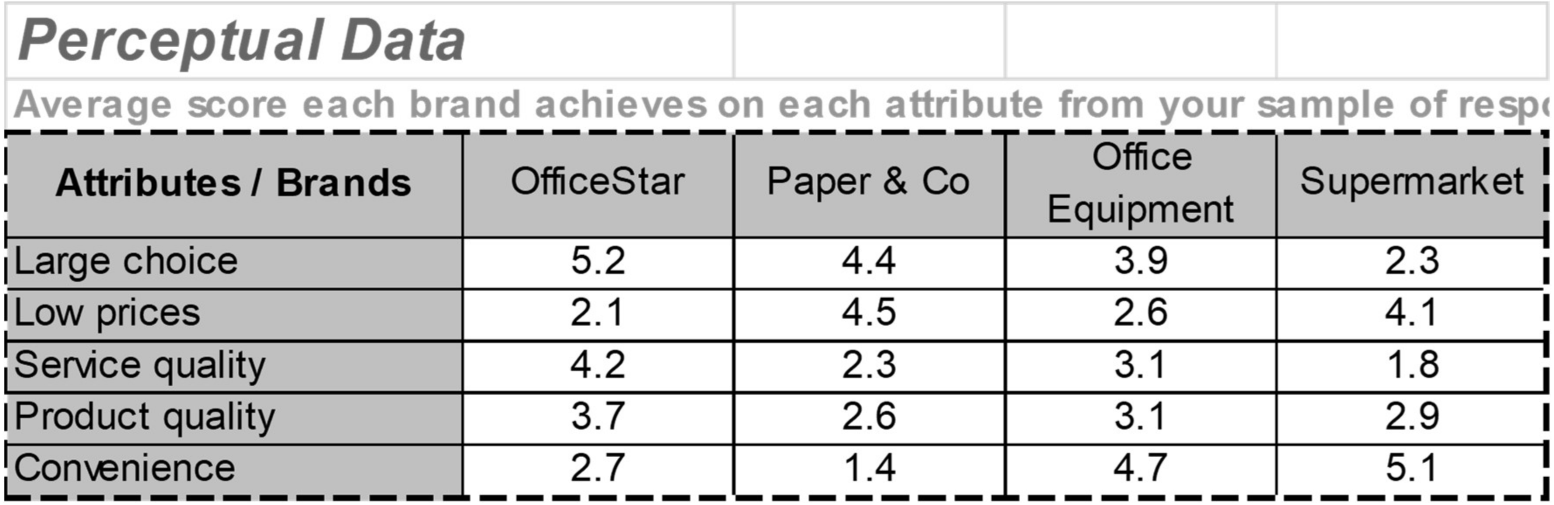
John

Attributes / Brands	OfficeStar	Paper & Co	Office Equipment	Supermarket
Large choice	5	4	5	2
Low prices	3	4	4	5
Service quality	3	2	5	3
Product quality	2	3	2	2
Convenience	1	1	2	4
Preference Score	5	3	3	1

Note that each respondent answers 6 x 4 = 24 Qs.

Officestar: Run the Analysis

- Find the **average rating** each brand gets on each attribute across respondents and tabulate it. Thus the resulting table could like this:



Perceptual Data

Average score each brand achieves on each attribute from your sample of resp

Attributes / Brands	OfficeStar	Paper & Co	Office Equipment	Supermarket
Large choice	5.2	4.4	3.9	2.3
Low prices	2.1	4.5	2.6	4.1
Service quality	4.2	2.3	3.1	1.8
Product quality	3.7	2.6	3.1	2.9
Convenience	2.7	1.4	4.7	5.1

- The above table is the prescribed format for **the *R shinyapp*.**

Preliminaries: Intro to R Shinyapps

- Consider a code-averse colleague.
- Now consider a typical R function's structure:

```
                 my.func <- function(inp1, inp2, ...){   ui.R


                  [some pre-processing, exception-checking etc]     global.R*


   server.R       result1 <- [processing block 1]
                  result2 <- [processing block 2]
                          ...

                 outp <- list(result1, result2, ...)   }  # my.func ends
           ui.R
```

- Q: How to get your code-averse colleague to engage with your work?
- An important consideration is *interactivity.*

Opening and Using the JSM App

- Run code for the jsm-shinyapp from Rstudio and examine its layout.

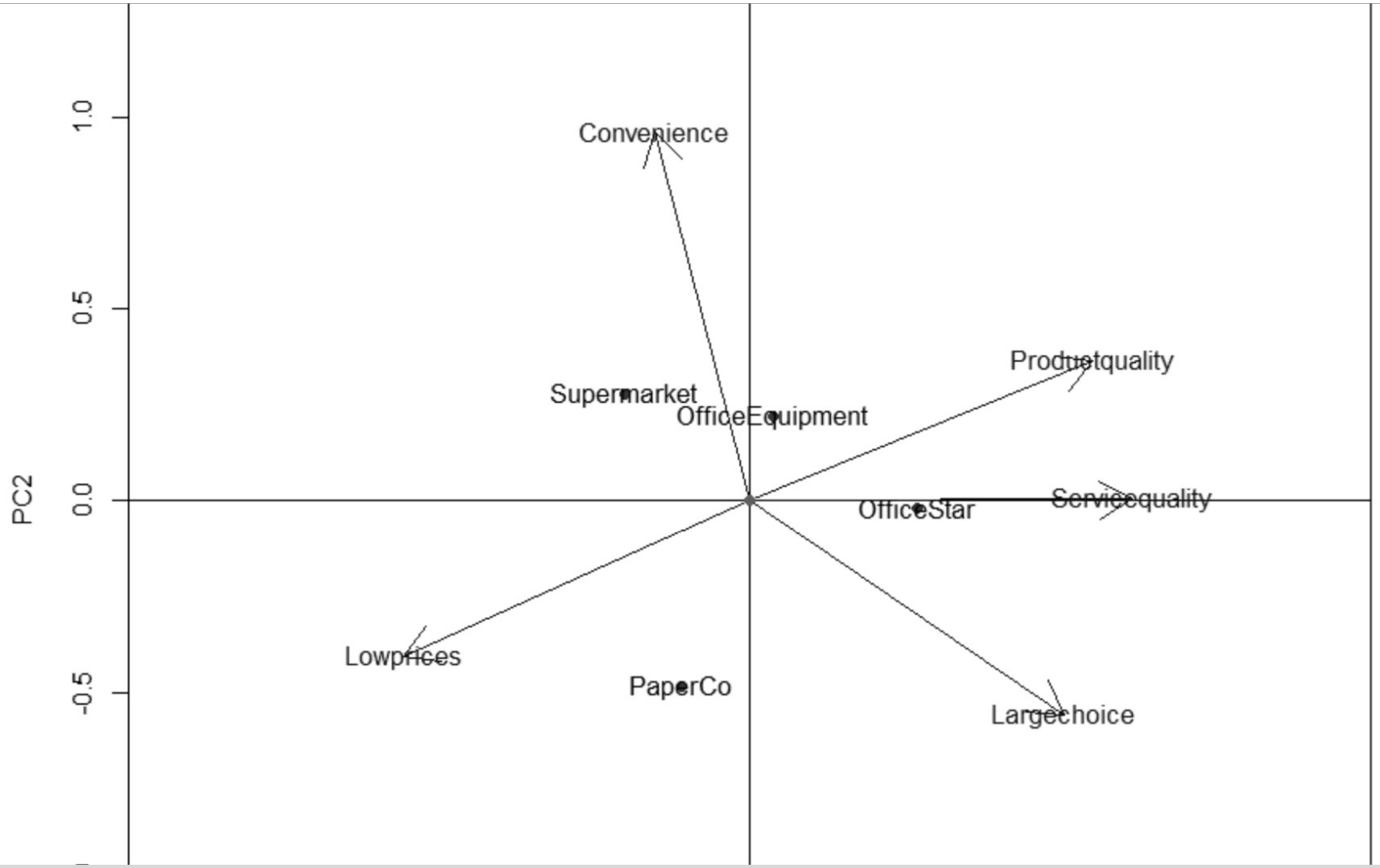
- What input fields do you see?

- What output tabs do you see?

- Now read in the datasets ‘officestar perceptual.csv’ and preference.csv

- Next, we walk through the output on how to read JSMs and what business analytic insights are available from there.

Officestar: Perceptual Map in R



Some Qs to think about:

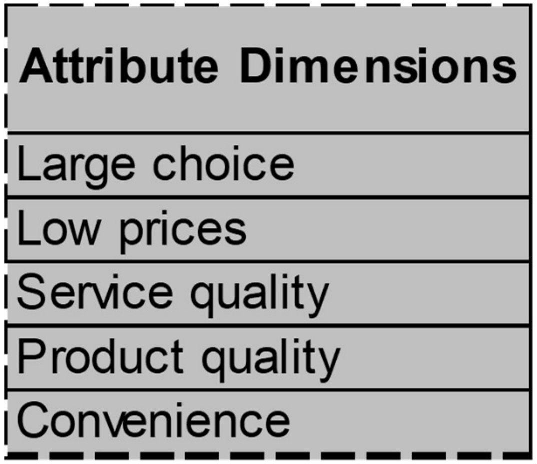
[1]. Which firm is perceived to be highest on (a) Service quality (b) Convenience (c) Low Prices (d) Product Quality?

[2]. Between which two attributes do you see the most white-space inviting potential entry?

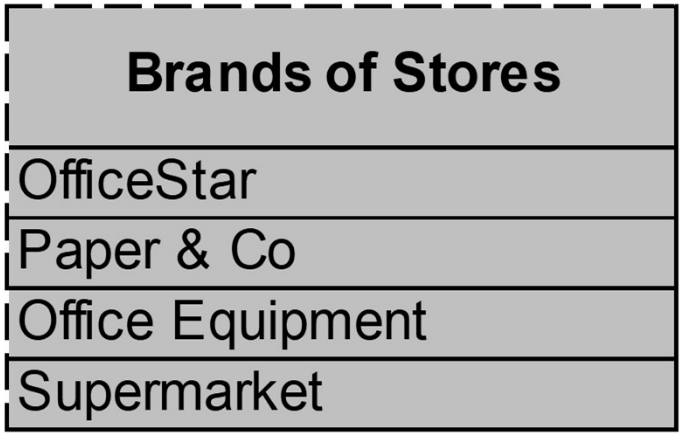
Officestar: The Problem at Hand continues

- **D.P. 2: Which stores do customers prefer** among our store and our competitor stores in the Office Supplies Business on 5 key dimensions.

- The two R.O.s are related but different. The former deals with perceptions, the second with preference.
  - Sure enough, there's a perceptual map to address the first question and a preference map to answer the second.



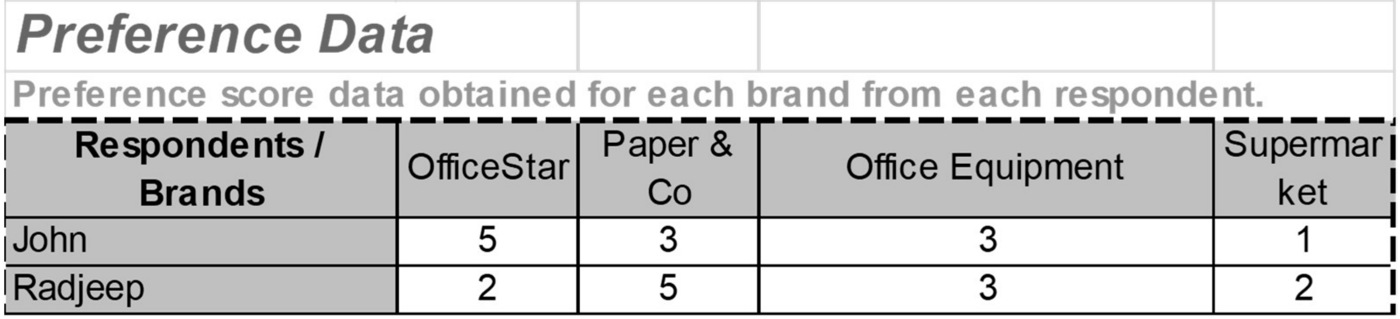
Attribute Dimensions
Large choice
Low prices
Service quality
Product quality
Convenience



Brands of Stores
OfficeStar
Paper & Co
Office Equipment
Supermarket

Officestar: Run the Analysis (JSM app)

- Preference data when entered into MEXL look something like this:



Preference Data

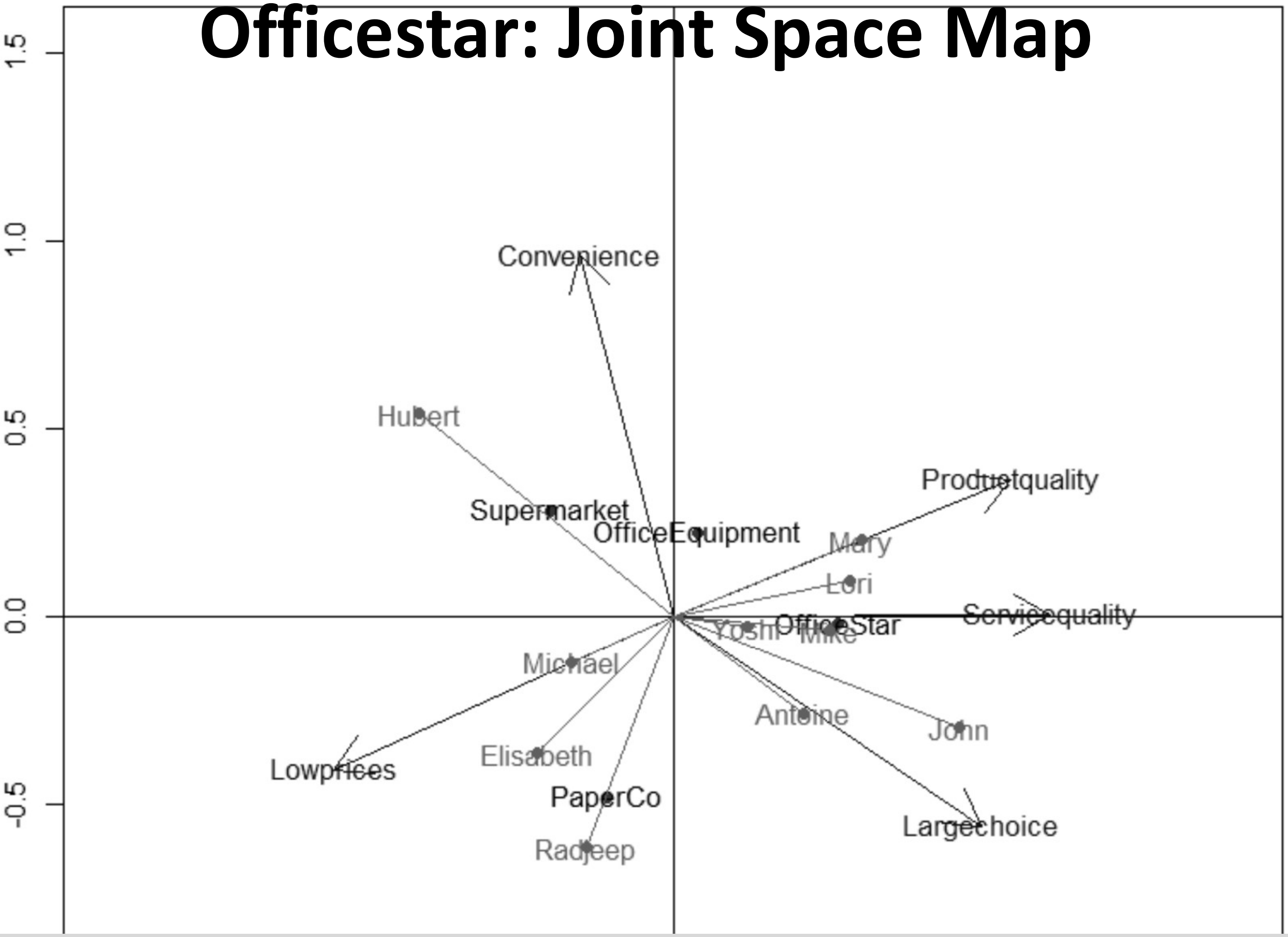
Preference score data obtained for each brand from each respondent.

Respondents / Brands	OfficeStar	Paper & Co	Office Equipment	Supermarket
John	5	3	3	1
Radjeep	2	5	3	2

- We use the same input into R also.
- We now **overlay** respondent preference vectors onto the perceptual map we saw earlier.
  - The result is called a *Joint space-map* or **JSM**.

> - While a perceptual map allows us to ask if there is a gap in the market, …
> - … the JSM allows us to see if there is “a market in the gap”. - Gary Lilien

Officestar: Joint Space Map



Quick – Check [1]. Who seems to most value (a) low prices, (b) convenience and (c) product quality AND service quality?

[2]. On what attributes should Officestar compete with the other firms, based on the JSM?

[3]. Assuming the sample genuinely represents the population, what might be the market share of Officestar?

Officestar Exercise: Notes on SWOT Insights

- JSMs, with caveats*, are a neat way reveal a full SWOT analysis for a focal brand.

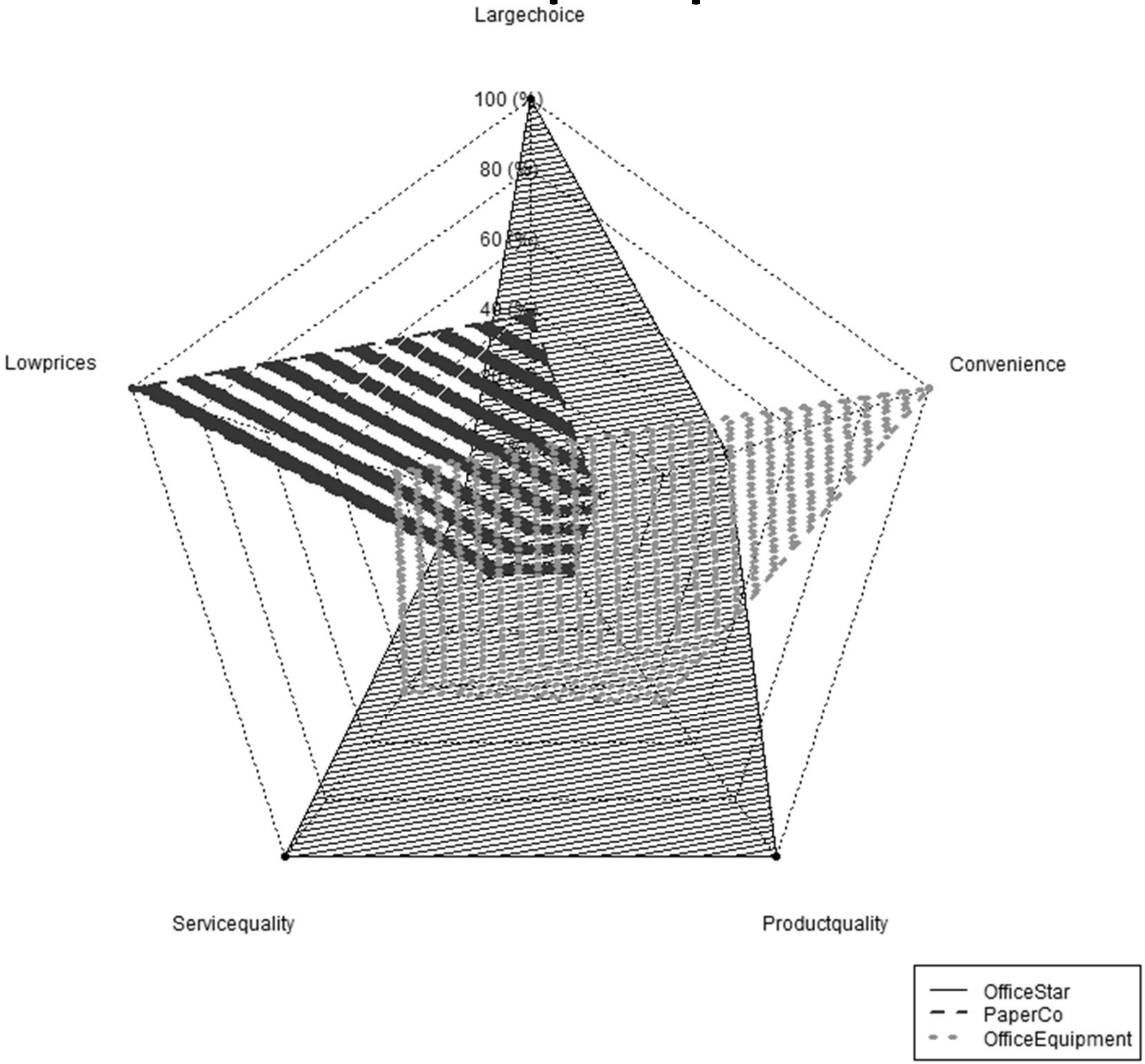
- JSMs reveal not only preferences, but could reveal preference-share leading to estimates of market share.

- A brand’s ‘S’ & ‘W’ (strengths and weaknesses) around attributes are revealed.

- The ‘O’ in SWOT (opportunities) in terms of white spaces are revealed.

- Conversely, the ‘T’ in SWOT (Threats) in terms of potential entry are also visible.

Officestar example: Spider Charts



Session Wrap-up

- The Intro and Overview session is over.

- Next session on, we will delve into *code*, in workshop mode.
  - Effort is to make available code on LMS at the earliest
  - Ensure you have the required modules and packages downloaded & ready

- If you're new or unfamiliar with R and/or Py, ensure you're able to replicate all classwork problems at home.

- Recall the shinyapp we ran?
  - Aim was to illustrate downstream use of survey data
  - We'll have a small workshop on how to build shiny apps

Thank You

Q & A

ISB