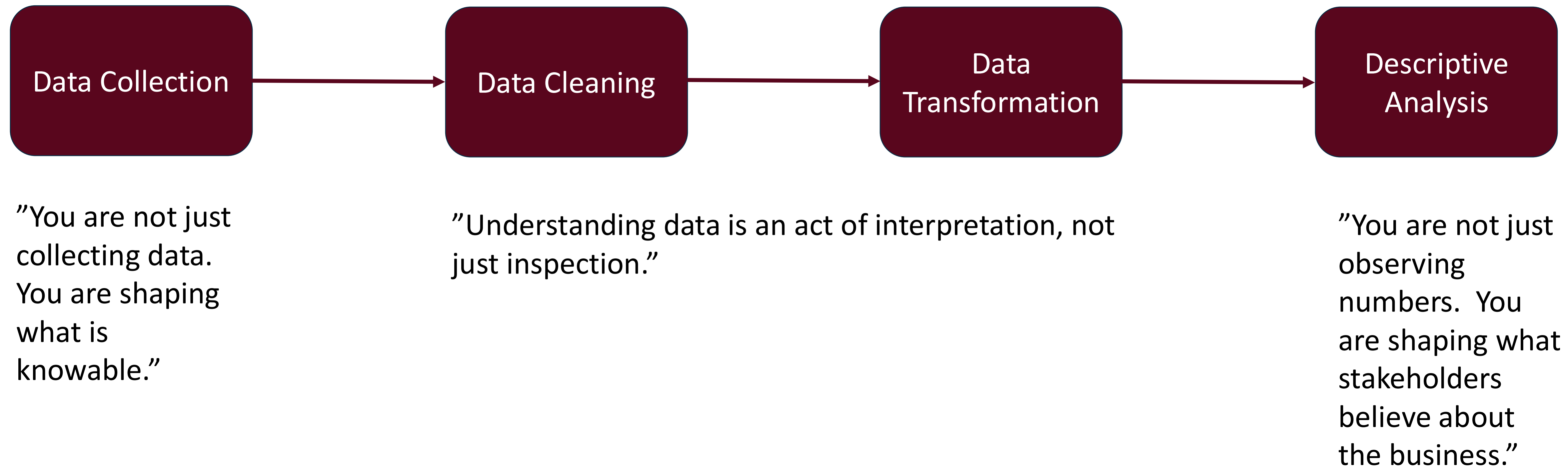


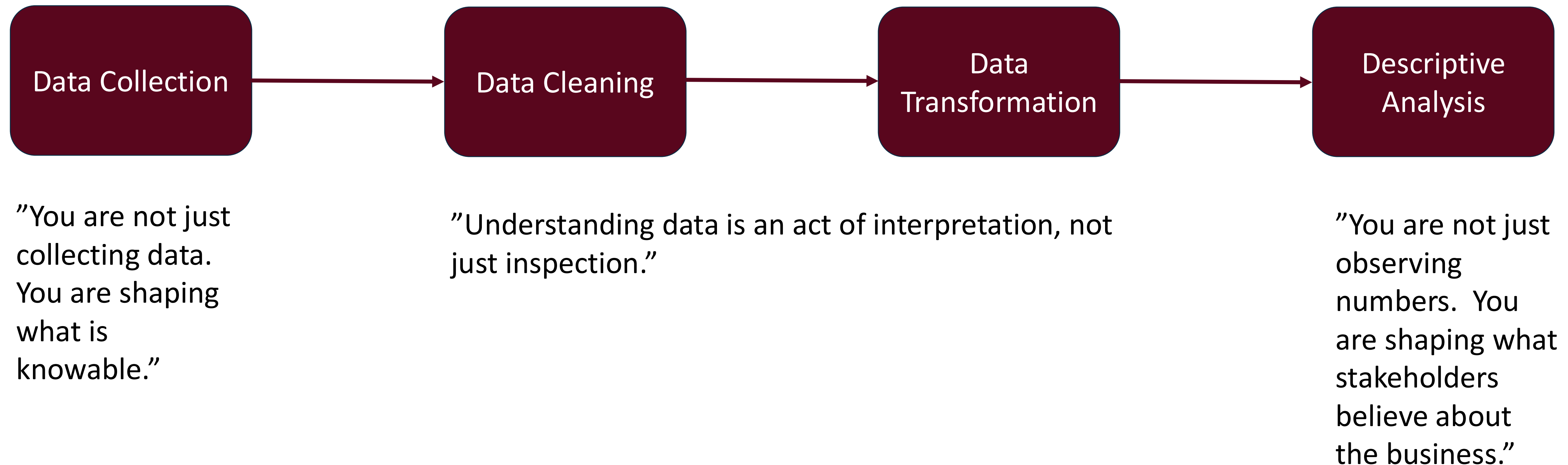
Recap – Descriptive Analysis

Work Flow



Program Throughlines

Communication, Testing & Security



Data Collection Framework



Stage	Key Questions About the Data Being Collected
1. Intent	What are we trying to answer or achieve? What will the data be used for?
2. Source	Where does the data live? Who owns it? How stable is it?
3. Structure	What is the data's shape? Structured, semi-structured, unstructured?
4. Semantics	What do the values mean? Are they standardized? Interpretable?
5. Access	How do we get it? What are the technical, legal, and ethical constraints?
6. Timing	How often is the data needed? Is freshness important?
7. Quality	How reliable is it? How do we know if it's incomplete or inconsistent?
8. Impact	How does this collection influence downstream analysis or decision-making?

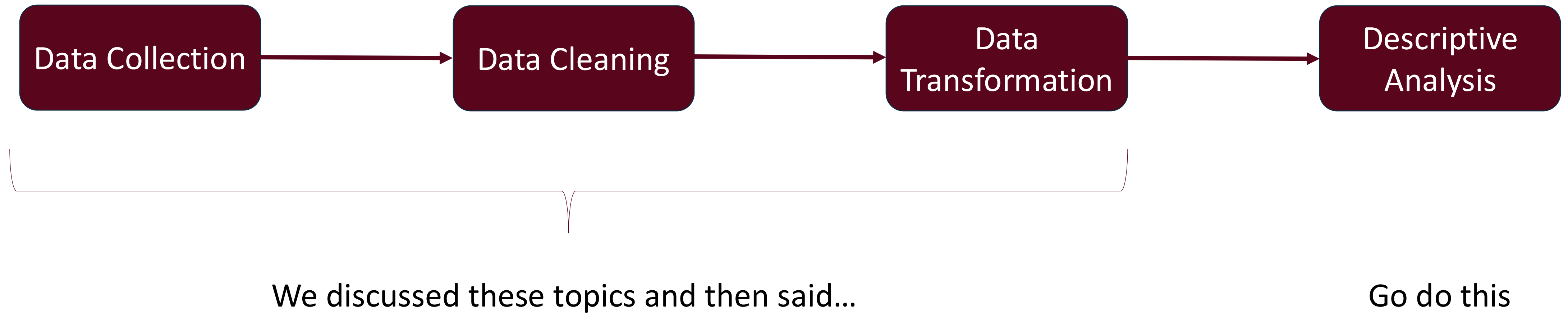
Data Cleaning & Transformations



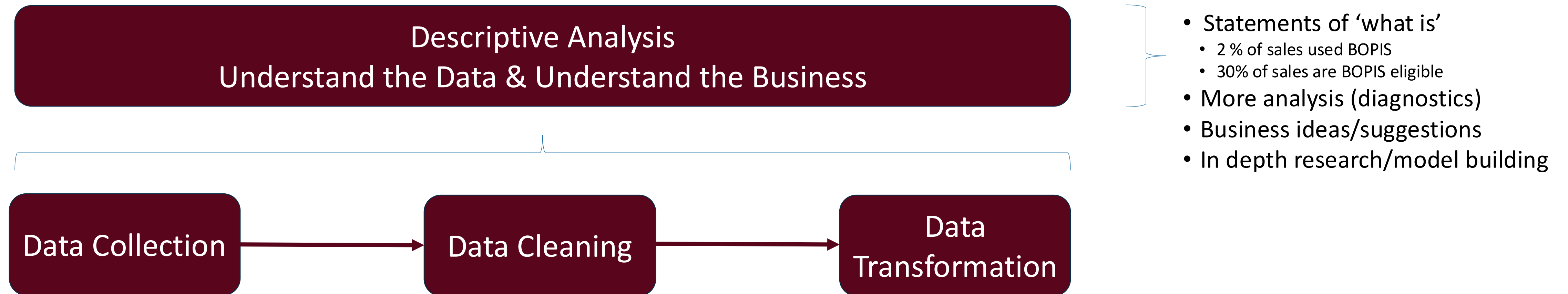
Cleaning	
• Missing values	<ul style="list-style-type: none">• Ignore rows with missing values• Use substitution methods – mean, mean of similar rows, regressed values, etc.
• Duplicate rows	<ul style="list-style-type: none">• Validate legitimacy• Delete redundant rows
• Outliers/Noise	<ul style="list-style-type: none">• Validate• Bin• Smoothing• Cluster
• Inconsistent/ Conflicting Data	<ul style="list-style-type: none">• Correct the dataset

Transformations	
• Integration	<ul style="list-style-type: none">• Join (make wider)• Append (merge)• Edit metadata
• Generalization of data	<ul style="list-style-type: none">• Hierarchy climbing (low level attributes (e.g., street name) is replaced with a higher level attribute (e.g., city))
• Normalization/ Scaling	<ul style="list-style-type: none">• Z-score method• Min-Max method• Decimal scaling
• Aggregation	<ul style="list-style-type: none">• Applying summary or aggregation operators to data (daily lab procedures to calculate yearly lab procedures)
• Dimensionality reduction to eliminate insignificant variables	<ul style="list-style-type: none">• Feature selection• Attribute sampling• Heuristic methods
• Data compression	<ul style="list-style-type: none">• Reduce data size with methods such as Principal Component Analysis, etc.
• Feature engineering	<ul style="list-style-type: none">• Add new variables ('features' in ML terminology) derived from existing data (e.g, first name from full name)

The Classroom Approach



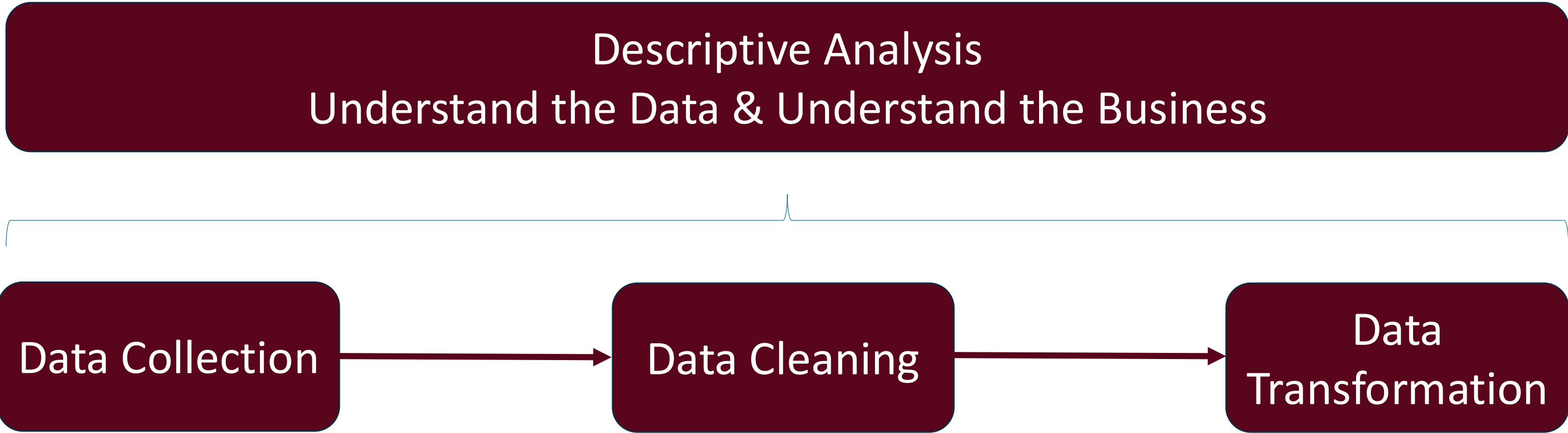
The Real-World Approach



The Real-World Approach: Starting with Questions

Channel Performance & Attribution

- Analyze Journey patterns
 - How many journeys have 1, 2, 3, etc. touchpoints?
 - What is the average # of touchpoints per journey?
 - Median touchpoints?
 - Most common touchpoint?
 - Conversion rate by touchpoint?
- Most common journey path
 - What are the top 10 journey paths (need full path)
 - How many single vs multi touchpoint journeys are there?
- Channel sequence patterns
 - Number of journeys by first touch channel?
 - Number of journeys by last touch channel?
 - Journey duration analysis
 - Avg, median & max duration
- Revenue attribution by different models
 - Matrix channel by first touch revenue, first touch %, last touch revenue, last touch %
 - Is the channel over or under credited based on last touch? (last – first touch %)



1. What data is needed to answer each question? What data is available?
2. What metrics will be used for each questions?
3. What are the operational definitions for those metrics? (e.g., What defines a product as BOPIS eligible: the purchase is online and there is at least one store that has inventory for the product being purchased)

Sample Questions

Channel Performance & Attribution

- Analyze Journey patterns
 - How many journeys have 1, 2, 3, etc. touchpoints?
 - What is the average # of touchpoints per journey?
 - Median touchpoints?
 - Most common touchpoint?
 - Conversion rate by touchpoint?
- Most common journey path
 - What are the top 10 journey paths (need full path)
 - How many single vs multi touchpoint journeys are there?
- Channel sequence patterns
 - Number of journeys by first touch channel?
 - Number of journeys by last touch channel?
 - Journey duration analysis
 - Avg, median & max duration
- Revenue attribution by different models
 - Matrix channel by first touch revenue, first touch %, last touch revenue, last touch %
 - Is the channel over or under credited based on last touch? (last – first touch %)
- Online to offline channel impact
 - How many customer journeys started online and ended in a store?
 - Associated revenue?
 - Percentage of all journeys?
 - Average touchpoints?
 - Which online channels drove store business?
- Device patterns
 - What are the top 10 device patterns in a customer journey?
 - What percentage of journeys experience device switching?
 - How is revenue associated with devices?

Operational Efficiency

- Current Inventory State
 - SKU
 - How many skus do we have?
 - # units?
 - retail value?
 - cost value?
 - How is inventory distributed between stores and warehouses?
 - How often do we have stockouts?
 - Do stockouts occur in stores or warehouses?
 - How many days of supply do we have?
 - Days supply < 7, >30
 - What is our inventory turnover rate?
- Fulfillment patterns
 - What is the distribution of order fulfillment (ship, in-store, BOPIS)?
 - Channel performance – mobile, online, store
 - # orders?
 - Revenue?
 - average order value?
 - Shipping cost?
 - Avg, median, max, # orders with free shipping
 - Returns?
 - Return rate, total returns by channel
- Current Cross Channel Gaps
 - For each channel, what fulfillment methods are used (counts and \$\$)
 - BOPIS
 - What % of orders use BOPIS?
 - Avg # items using BOPIS?
 - How many products are available in at least on store?
 - What is the potential # orders that could use BOPIS?
 - Store Inventory coverage
 - Avg stores per product?
 - Number of products in all stores?
 - Product in only one store?

Customer Acquisition and Value

- Acquisition Cost by Channel
 - What is the true customer acquisition cost by channel?
 - Which channel has the highest acquisition cost?
 - The lowest?
 - Are there channels where acquisition cost exceed profit?
- Lifetime Value
 - Which channel has been providing the best lifetime value?
 - How do lifetime values differ between single and omnichannel?
 - How many single channel customers?
 - Multichannel?
 - What is the omnichannel adoption by customer segment?
 - Are omni channel customers worth more than single channel?