NC STATE UNIVERSITY

Exploratory Data Analysis for Clinical Research

February 4th, 2026

Emily Griffith
John Slankas

NC STATE UNIVERSITY

Outline

- Welcome and introduction
- Foundations
- The EDA Lifecycle
- Interpretation and next steps
- Wrap up and Q&A

NC STATE UNIVERSITY

Welcome!

Who we are:

- Emily Griffith
  Professor of the Practice, Department of Statistics
  Director of Consulting, Data Science and AI Academy, NC State

- John Slankas
  Senior Research Scholar, Laboratory for Analytic Sciences, NC State

We manage a subaward from NC TraCS to help provide support for clinical research using data science techniques.

E,J

E

NC STATE UNIVERSITY

A quick note

- We will discuss
  - EDA requirements for different types of analytical questions
  - The four dimensions of exploration
  - Data quality issues and limitations
  - Determining appropriate next steps after an EDA
- We will provide
  - A synthetic dataset of respiratory patient records from OMOP CDM
  - A notebook with workshop exercises
    - We will post solutions to these exercises on our website!

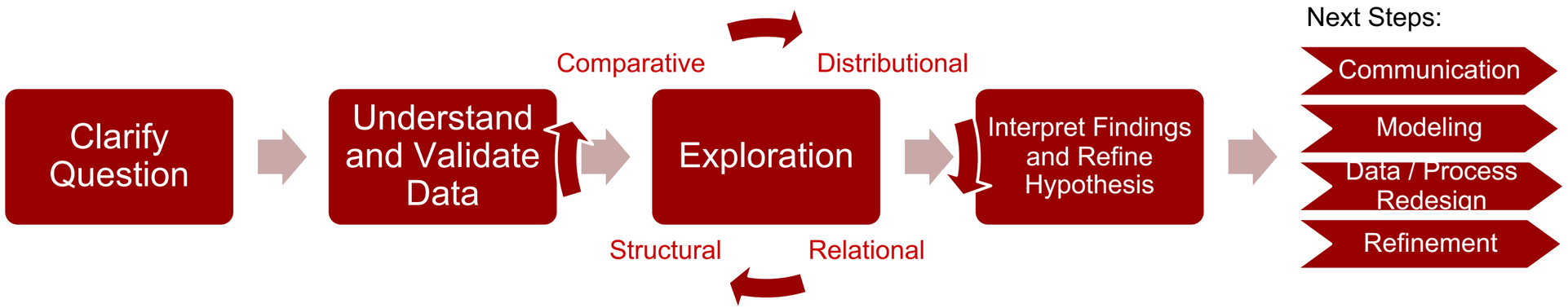
NC STATE UNIVERSITY

Extras for you!

- We have a website with:
  - a sample data set
  - sample exercises
  - solutions videos

NC STATE UNIVERSITY

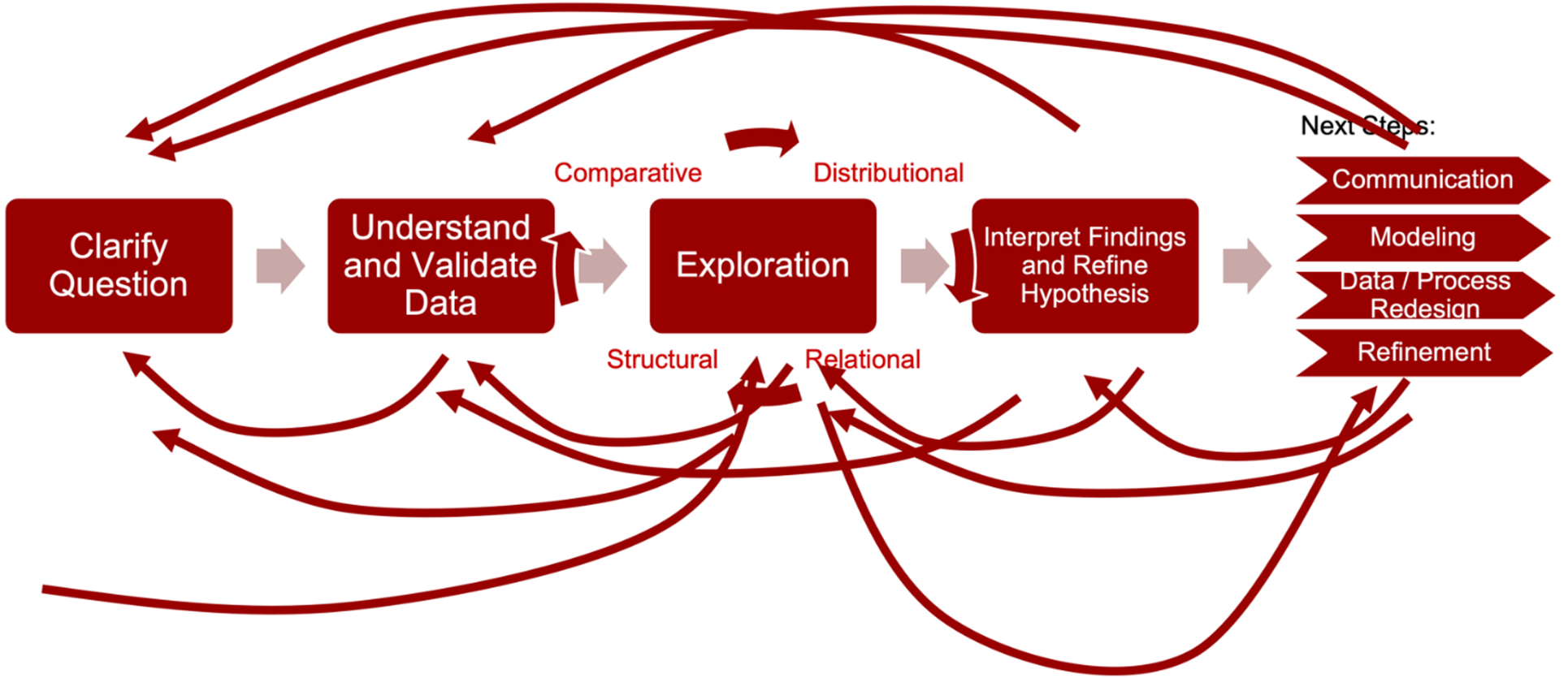
Exploratory Data Analysis Lifecycle



J

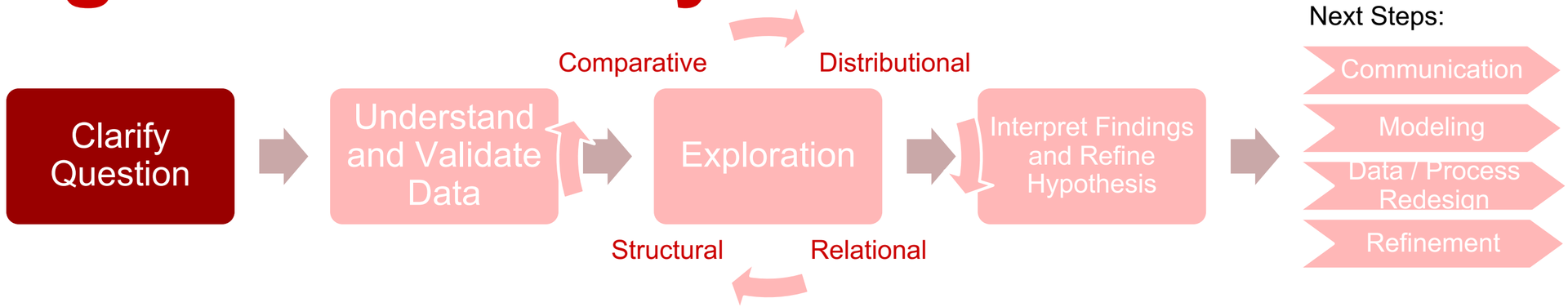
NC STATE UNIVERSITY

The "Reality" of an EDA "Lifecycle"



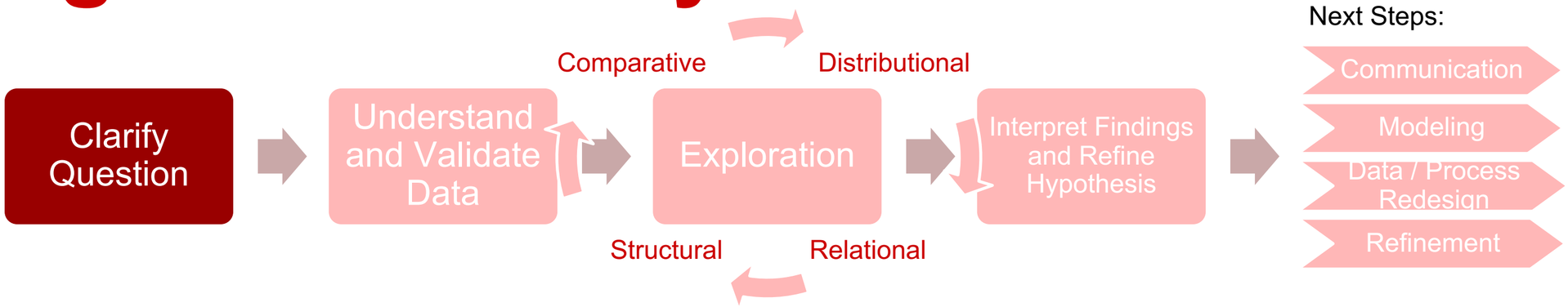
J

Stage One: Clarify Question



Descriptive: What happened?
Exploratory: What patterns exist?
Inferential: Does this generalize beyond the sample?
Predictive: What is likely to happen next?
Prescriptive: What action should we take?
Causal: What would happen if we intervened?
Mechanistic: What underlying process produces the behavior?

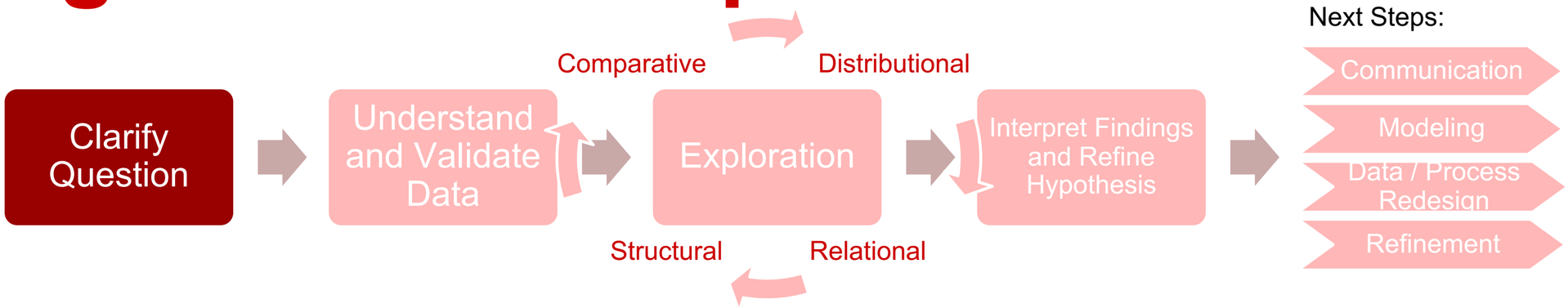
Stage One: Clarify Question



Descriptive: What happened?
Exploratory: What patterns exist?
Inferential: Does this generalize beyond the sample?
Predictive: What is likely to happen next?
Prescriptive: What action should we take?
Causal: What would happen if we intervened?
Mechanistic: What underlying process produces the behavior?

What counts as relevant evidence?
What assumptions are present?
What can be concluded?

Stage One: Descriptive Questions

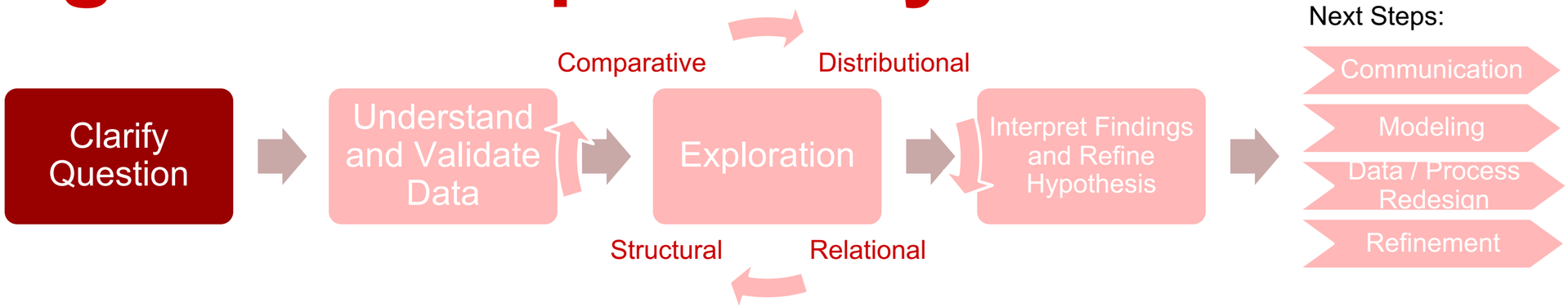


Descriptive: What happened?

- Summarize events, conditions, or behavior as recorded in the data.
- Provide clear, accurate reporting
- No claims of generality or forecasting.

- What is the distribution of respiratory conditions (e.g., pneumonia, cough, dyspnea) across all patient visits?
- What are the mean, median, and range of vital signs (oxygen saturation, respiratory rate, heart rate) for suspected COVID-19 patients?
- How many inpatient vs. outpatient visits occurred, and what is the average length of stay for inpatient visits?

Stage One: Exploratory Questions

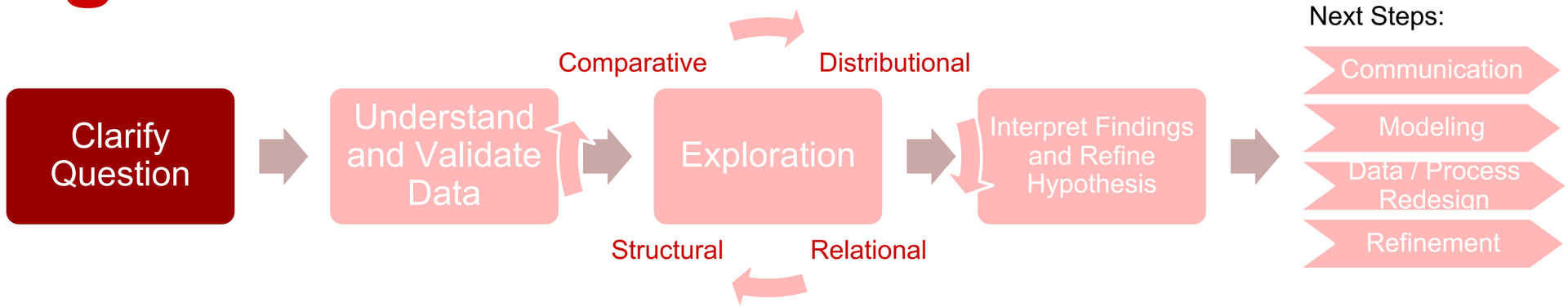


Exploratory: What patterns and relationships exist?

- Search for structure, trends, anomalies, or relationships
- Gain insights or guide further analysis.
- Goal: hypothesis generation rather than confirmation.

- Is there a relationship between body temperature and oxygen saturation levels in suspected COVID-19 patients?
- Do patients with multiple respiratory conditions (e.g., pneumonia AND dyspnea) show different vital sign patterns than those with single conditions?

Stage One: Inferential Questions

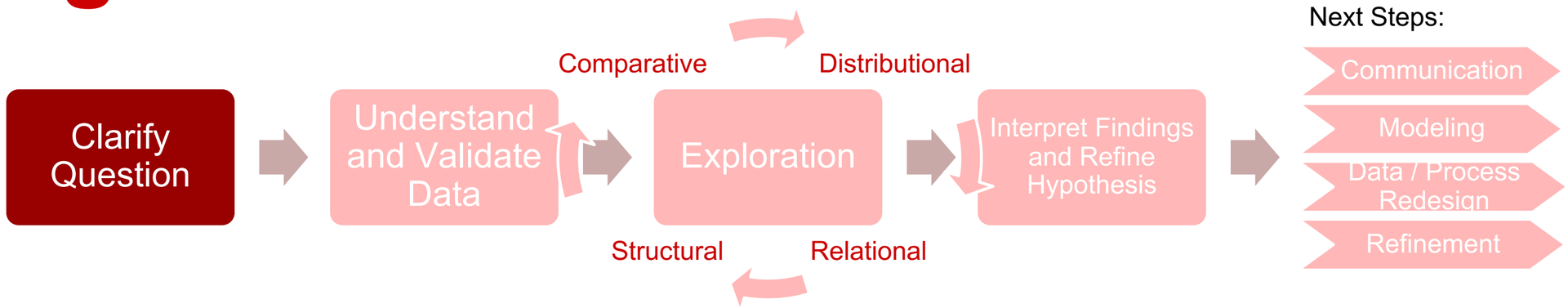


Inferential: Does this generalize beyond the sample?

- Representative of the population, not just the sample
- Requires statistical assumptions and careful design

- Can we infer that older patients have significantly higher rates of inpatient admission for respiratory conditions?
- Is the mean respiratory rate significantly different between patients with pneumonia vs. those without?

Stage One: Predictive Questions

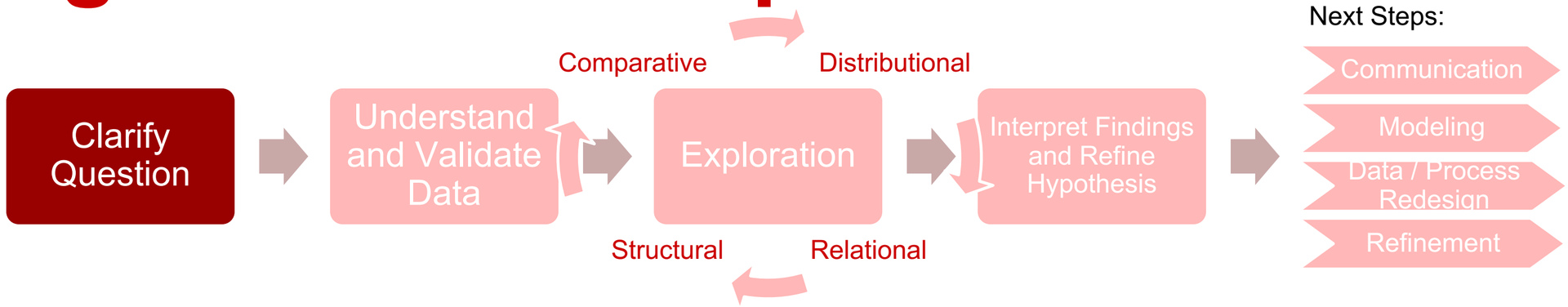


Predictive: What is likely to happen next?

- Estimate unknown or future outcomes using existing data
- Typically evaluated using an accuracy-based metric

- Based on vital signs at admission, can we predict which patients will require inpatient care?
- Based on historical visit patterns, can we predict which patients are likely to have recurring respiratory visits?

Stage One: Prescriptive Questions

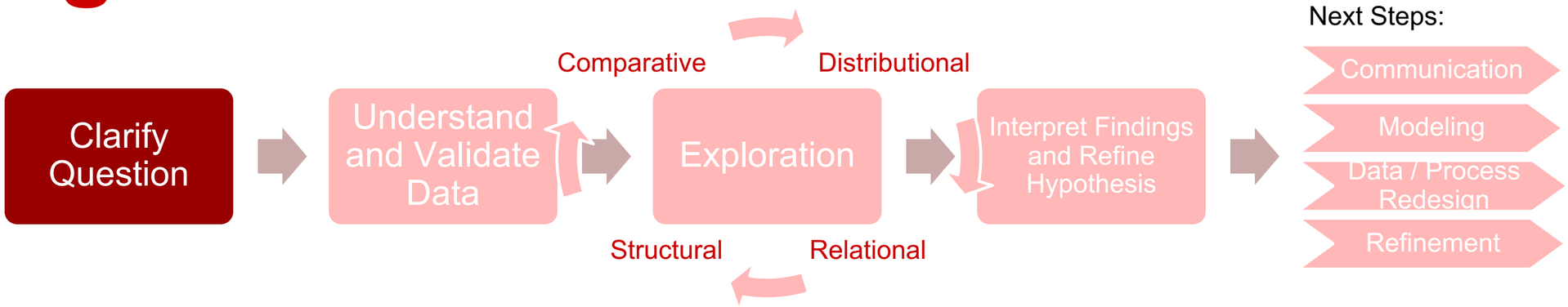


Prescriptive: What action should we take?

- Select the best action given
  - Costs
  - Benefits
  - Constraints
  - Risk tolerances

- Which combination of vital sign thresholds should trigger escalation to inpatient care for respiratory patients?
- Based on patient profiles, which demographic groups should be prioritized for flu vaccination outreach?

Stage One: Causal Questions

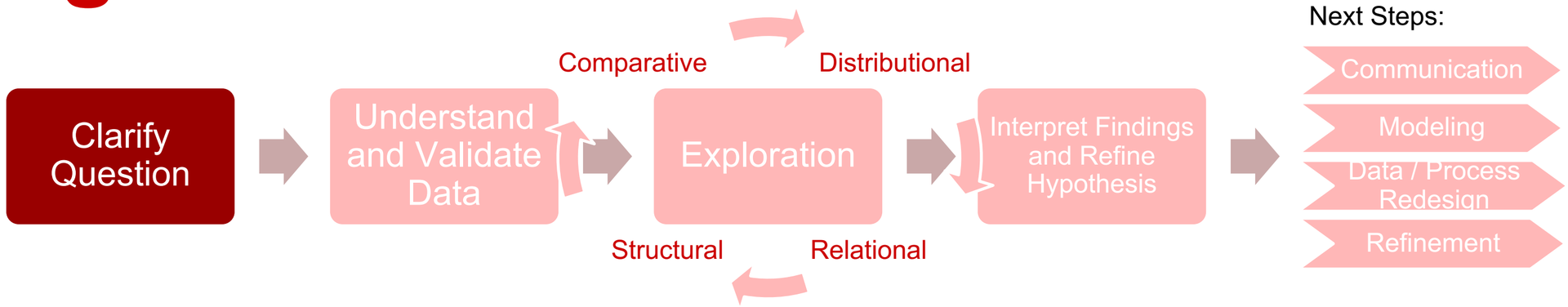


Causal: What would happen if we intervened?

- Estimate the effect of a deliberate change/treatment
- Data must support causal inference

- Would earlier flu vaccination reduce the incidence of severe respiratory conditions requiring hospitalization?
- Would applying treatment *x* for patients with abnormal respiratory rates at first presentation prevent disease progression?

Stage One: Mechanistic Questions

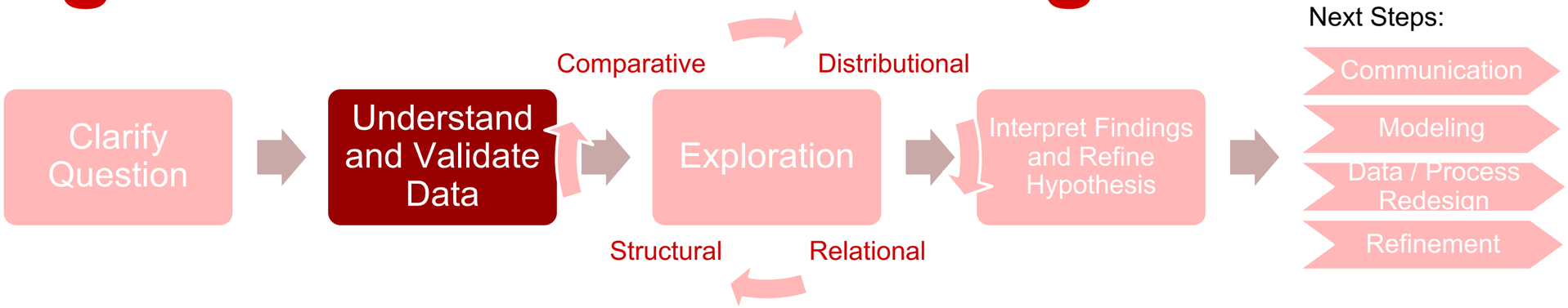


Mechanistic: What process produces this behavior?

- Seek to understand how and why a system behaves the way it does

- How do hospital admission criteria and triage protocols explain the observed patterns of inpatient vs. outpatient classification for similar presenting symptoms?
- What standard clinical workflows or care pathways explain why certain respiratory conditions consistently receive more complete vital sign documentation than others?

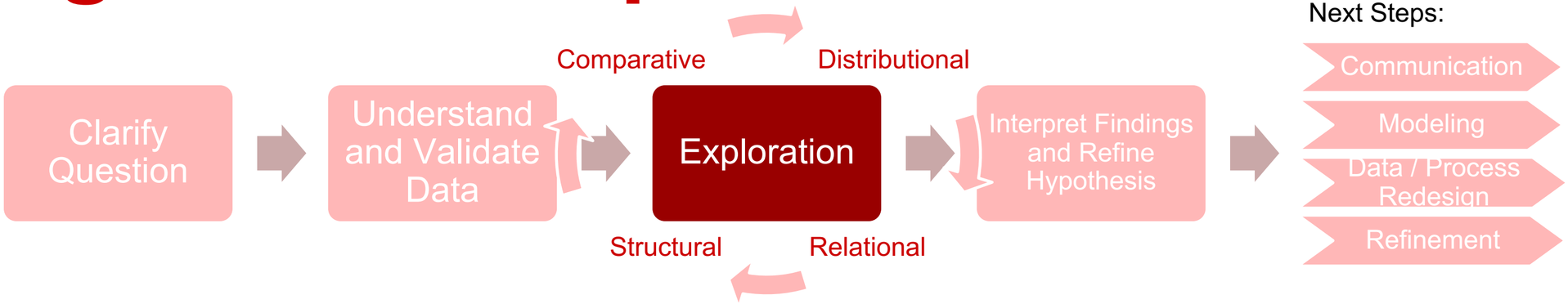
Stage Two: Understanding Data



Key Questions:

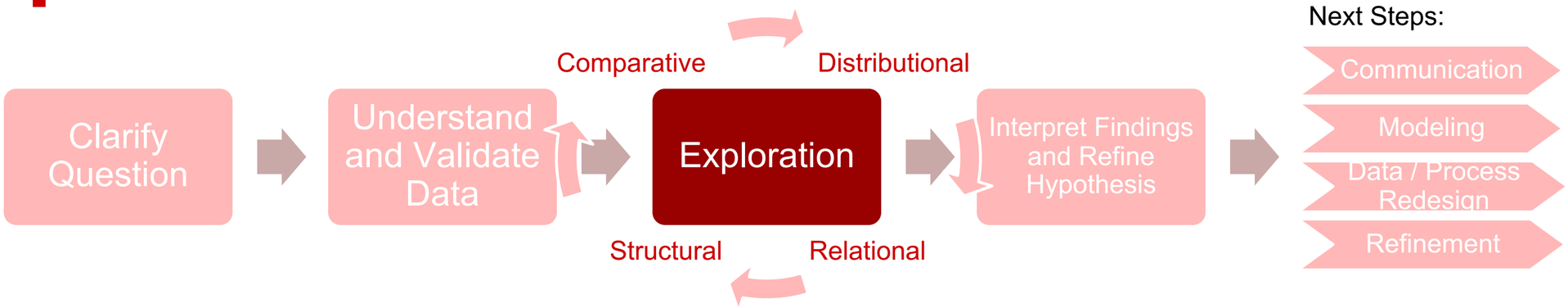
- What kinds of data are we actually working with?
- What does a single “row” represent in this dataset?
- How are clinical concepts encoded and standardized?
- What structure exists across time and hierarchy?
- What data quality or representation issues should we expect?

Stage Three: Exploration



- Exploration asks: *What does the data want to tell us?*
- Each exploration lens reveals **different kinds of evidence**
- We will examine the data across **four complementary dimensions**
- Not all questions require all four dimensions
- Exploration may **confirm assumptions—or challenge them**

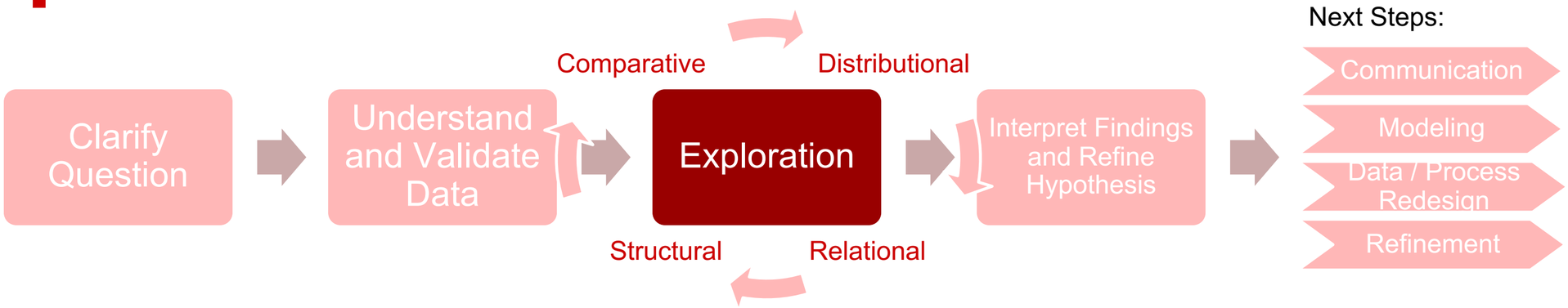
Exploration: Distributional / Univariate



- Examine ranges, central tendency, and variability
- Identify skewness, multimodality, and extreme values
- Detect implausible or clinically impossible measurements
- Understand measurement frequency and missingness

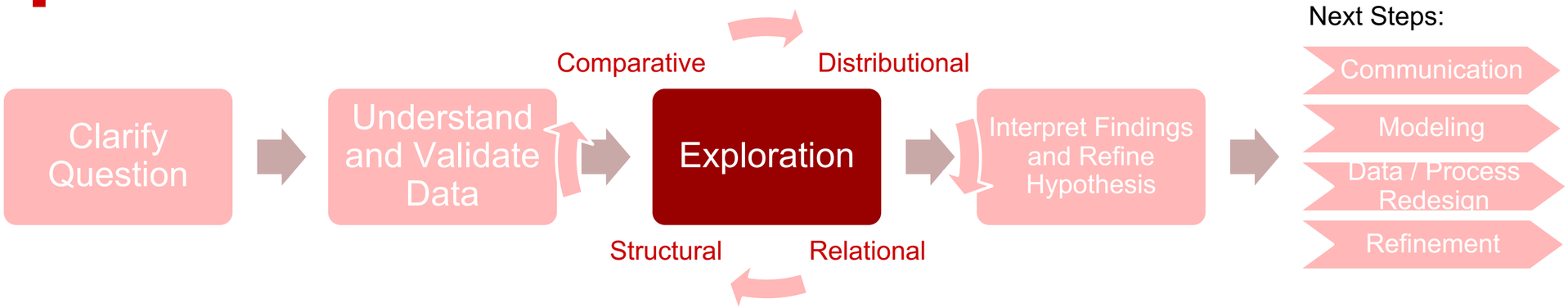
Distributional EDA often reveals data quality issues

Exploration: Relational / Bivariate



- Explore associations, correlations, and dependencies
- Identify nonlinear or threshold-based relationships
- Look for relationships that suggest hypotheses
- Distinguish signal from spurious correlation

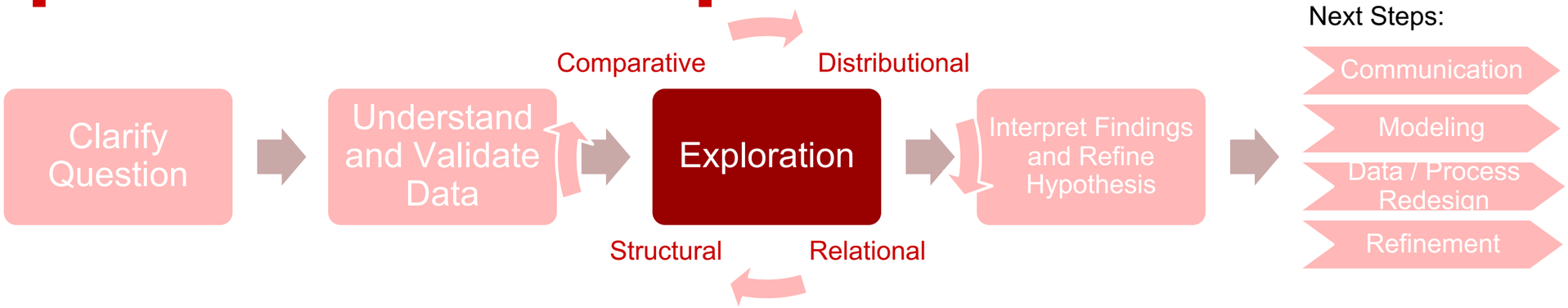
Exploration: Structural



Structural: Temporal, Hierarchical, Sequential, Geographical

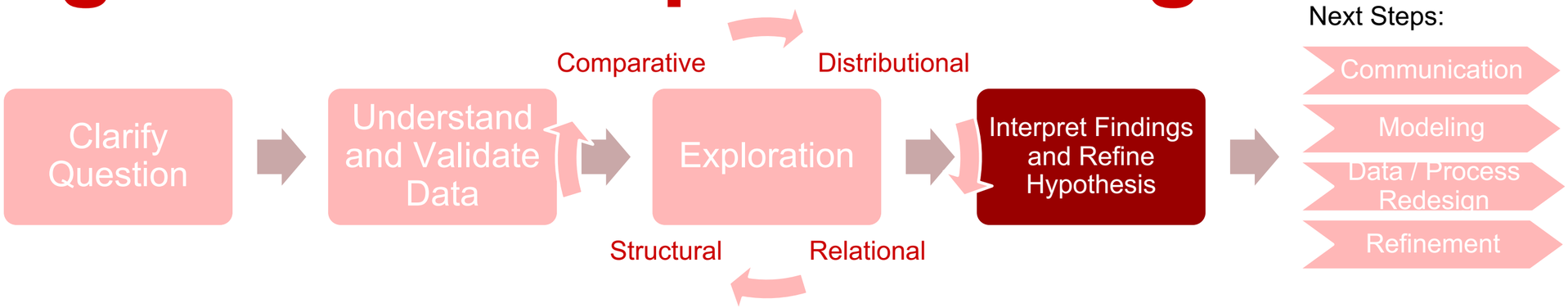
- Identify repeated measurements within patients
- Examine temporal ordering of observations
- Understand nesting (observations → visits → patients)
- Detect bursts, gaps, and monitoring patterns

Exploration: Comparative



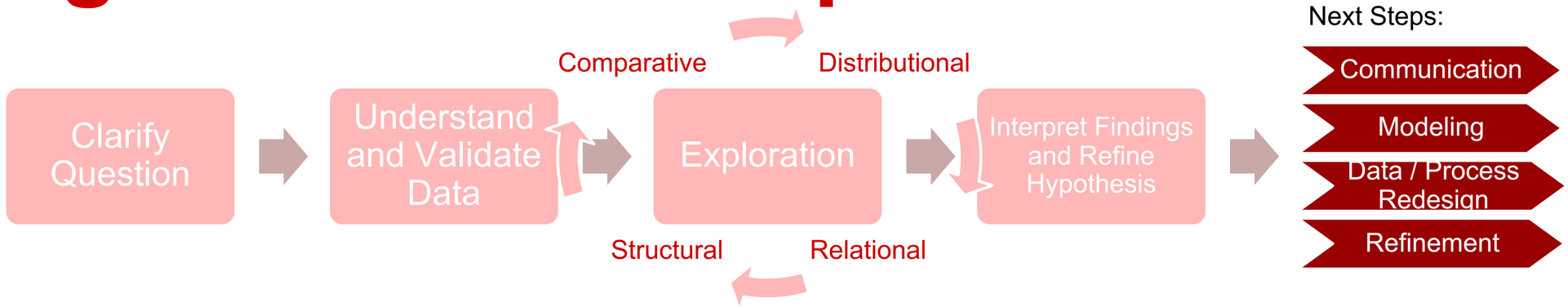
- Compare distributions across groups
- Identify systematic differences between populations
- Contrast stable vs unstable patients
- Reveal inequities or operational differences

Stage Four: Interpret Findings



- What patterns are robust vs fragile?
- Which findings align with clinical expectations?
- What surprised us?
- What assumptions might be violated?
- Which findings deserve follow-up analysis?

Stage Five: Next Steps



- Refine or reframe analytical questions
- Identify data gaps and collection needs
- Decide whether inferential or predictive modeling is appropriate
- Communicate findings to stakeholders
- Determine whether process or data redesign is needed

NC STATE UNIVERSITY

1-Minute EDA Checklist

Before you plot

- ☐ What does one row represent (patient, visit, observation)?
- ☐ Are there repeated measurements or nested structure?
- ☐ How were these data generated in clinical practice?

While exploring

- ☐ Have I examined distributions before relationships?
- ☐ Am I respecting time, hierarchy, and sequencing?
- ☐ Am I comparing groups that are meaningfully comparable?

Before concluding

- ☐ Are patterns robust or driven by a few cases?
- ☐ Could workflows or monitoring explain what I see?
- ☐ What assumptions am I implicitly making?

Before moving on

- ☐ What questions can the data answer well?
- ☐ What questions require more data or stronger design?
- ☐ What is the *next* responsible analytical step?

NC STATE UNIVERSITY

References and Suggested Resources

- *Communicating with Data: The Art of Writing for Data Science* by D Nolan and S Stoudt
- *Effective Data Storytelling: How to Drive Change with Data, Narrative, and Visuals* by Brent Dykes
- *What is the question?* by J Leek and R Peng in *Science* **347**, 1314-1315 (2015). DOI: 10.1126/science.aaa6146
- *Exploratory Data Analysis* by J Tukey (a classic!)

NC STATE UNIVERSITY

We welcome your feedback!

- We’d appreciate your feedback
- Thanks to NC TraCS for supporting this work
- Thank you for attending!