DATA PREPROCESSING TECHNIQUES

@ AIMS Senegal

27 Oct- 14 Nov, 2025

ABOUT ME

- My name is **Rockefeller**.
- Senior Data Scientist, **Morae**.
- Building and deploying GenAI applications in production

- PhD defense soon to happen (AI applied to Dynamical systems)

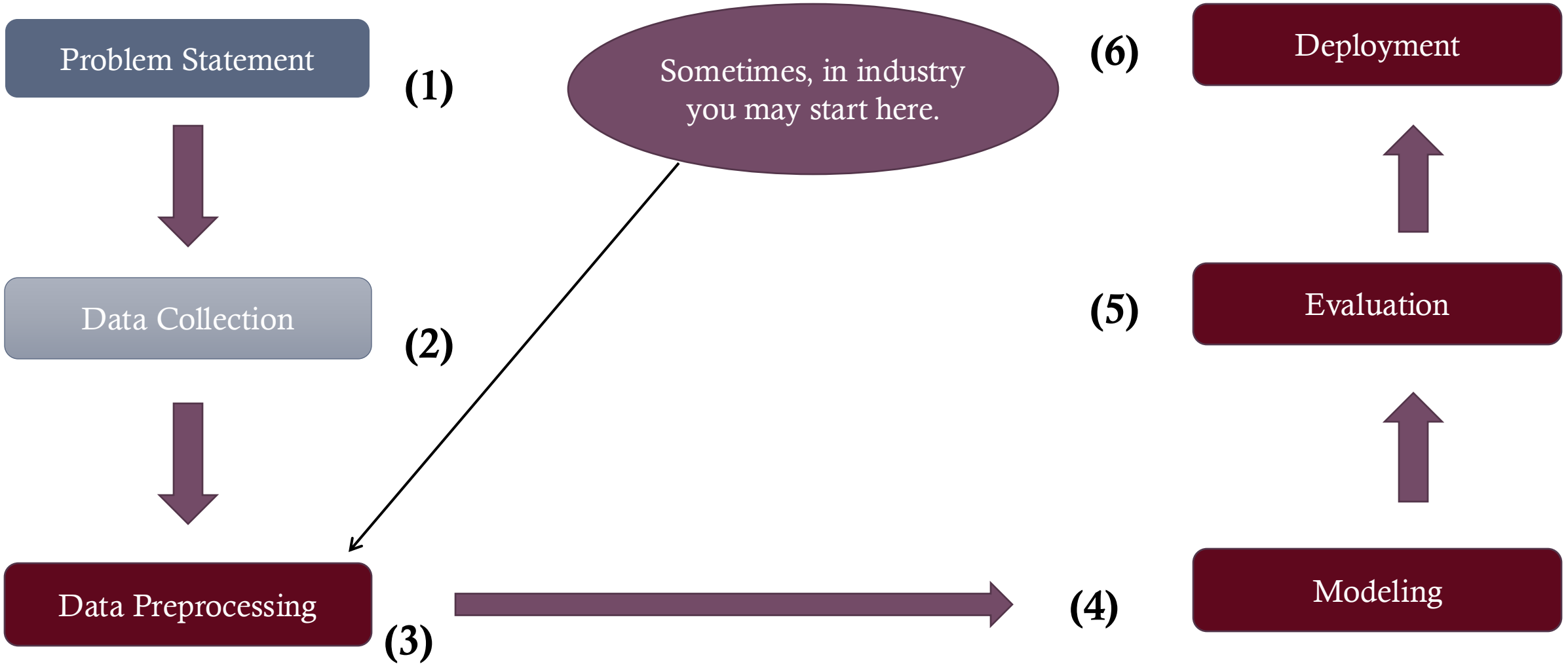
MOTIVATION FOR THIS COURSE

1. Fitting models with raw data is (often) the guarantee of building biased models.

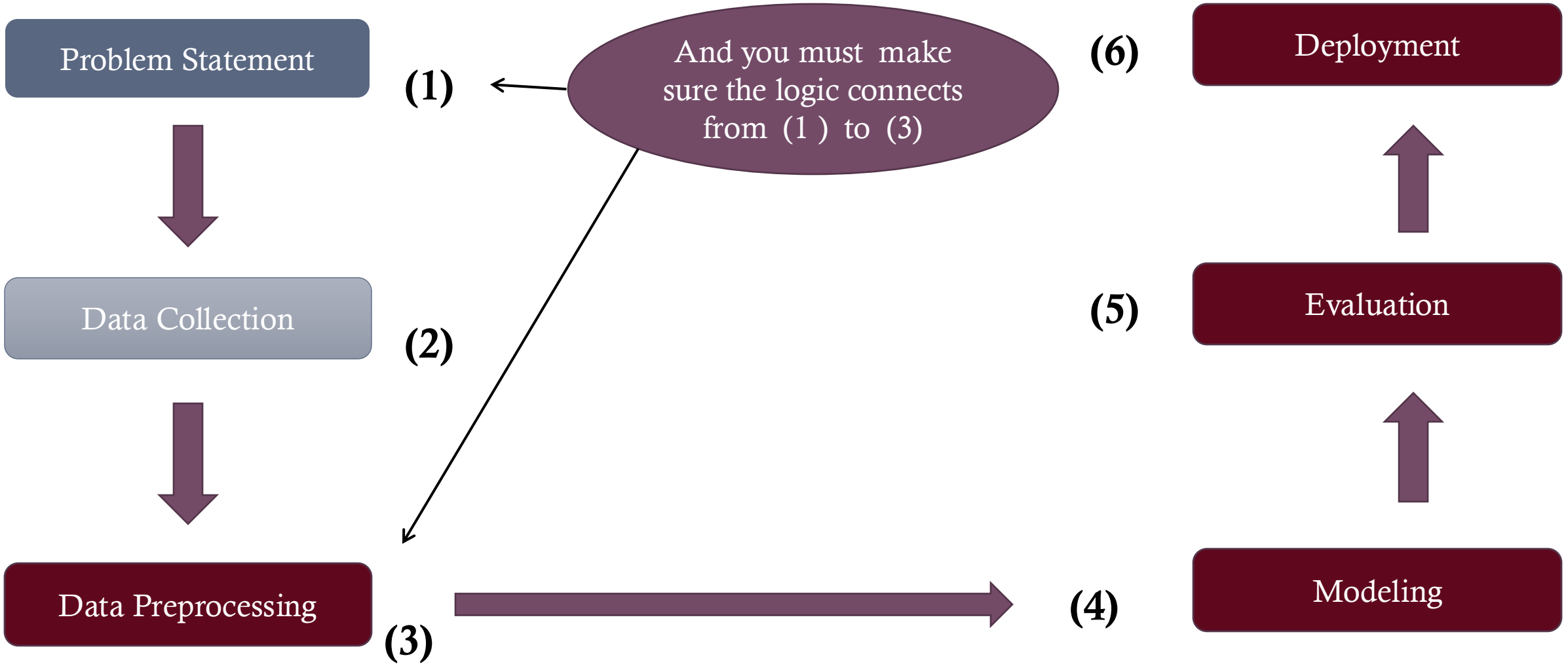
2. Most of the attention is often directed towards building models assuming the data is already **modelled** and **structured.**

TYPICAL DATA SCIENCE PROJECT ROADMAP (**IN ACADEMIA**)

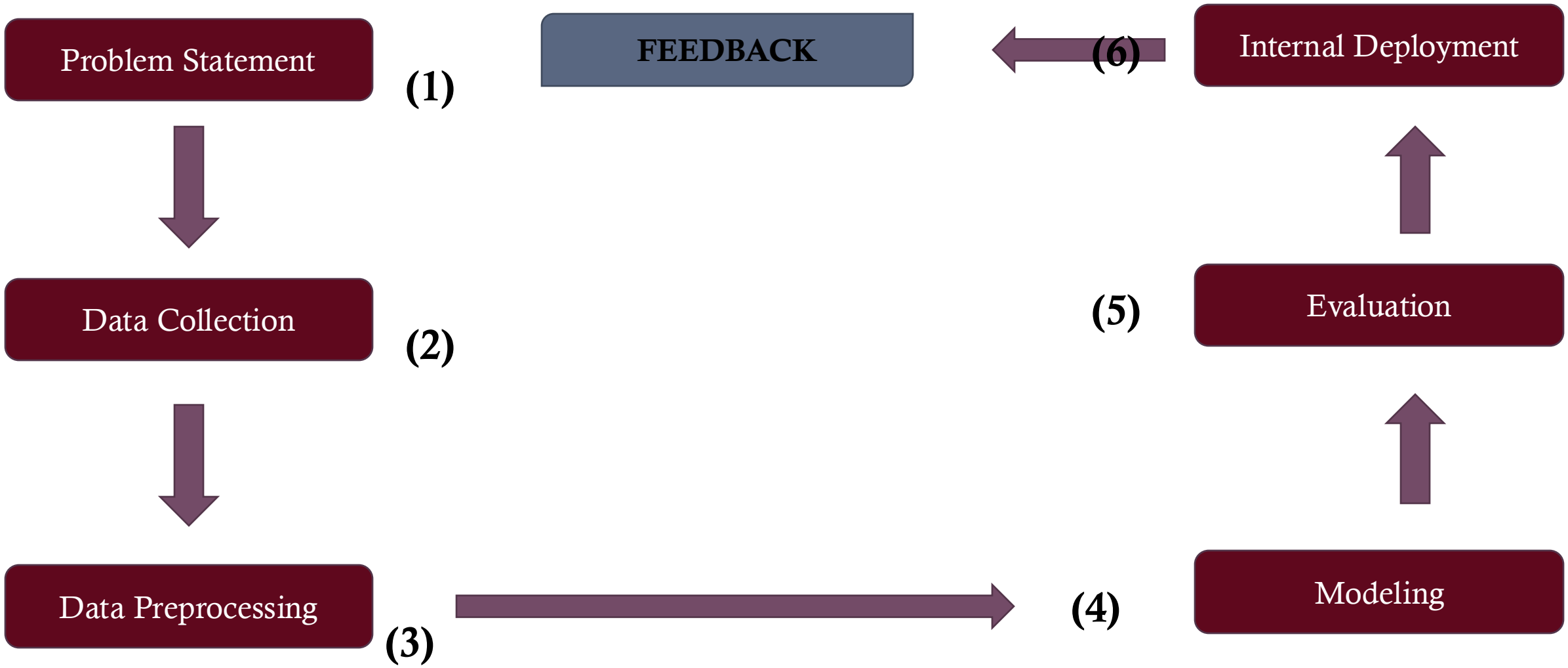
TYPICAL DATA SCIENCE PROJECT ROADMAP (IN INDUSTRY)



TYPICAL DATA SCIENCE PROJECT ROADMAP (IN INDUSTRY)

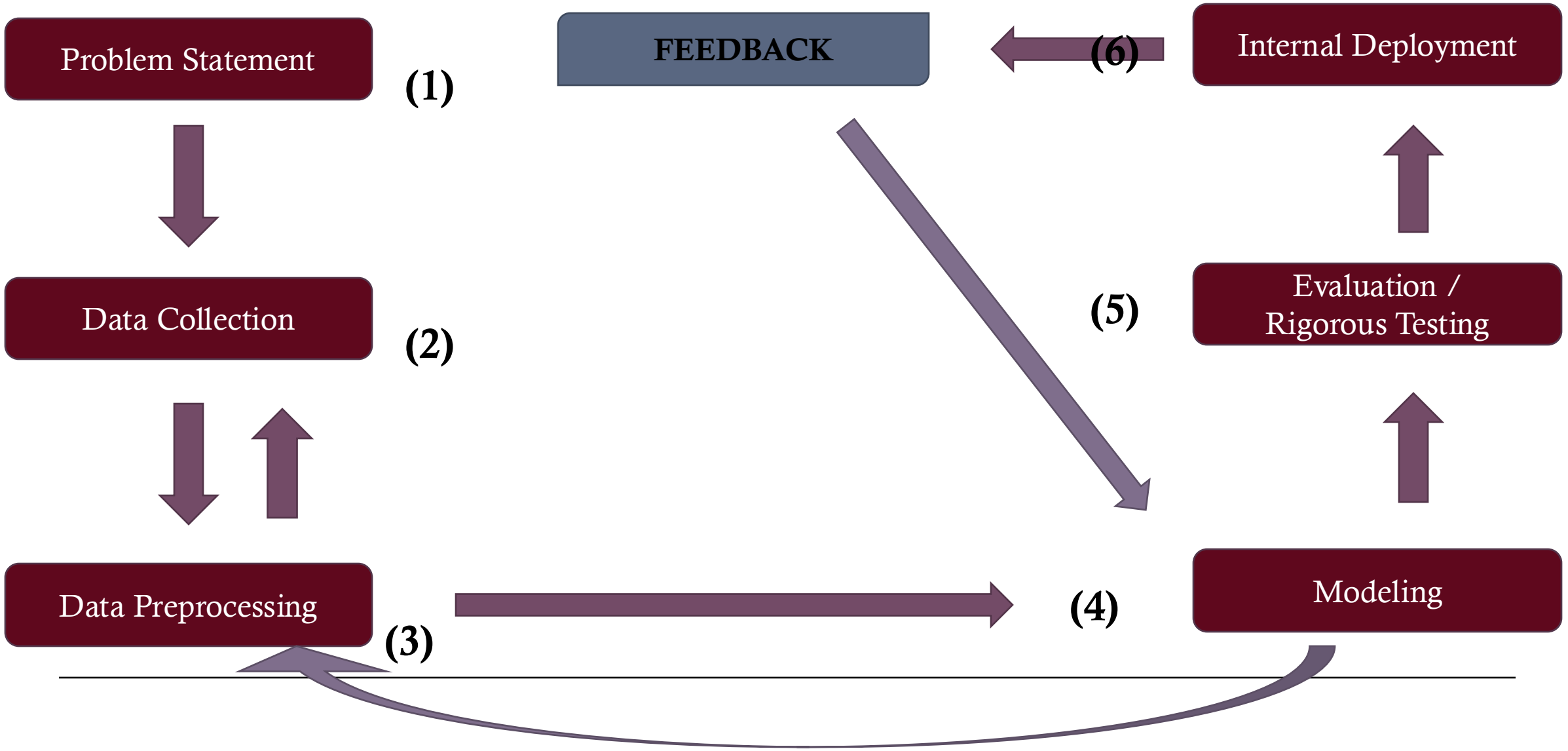


TYPICAL DATA SCIENCE PROJECT ROADMAP



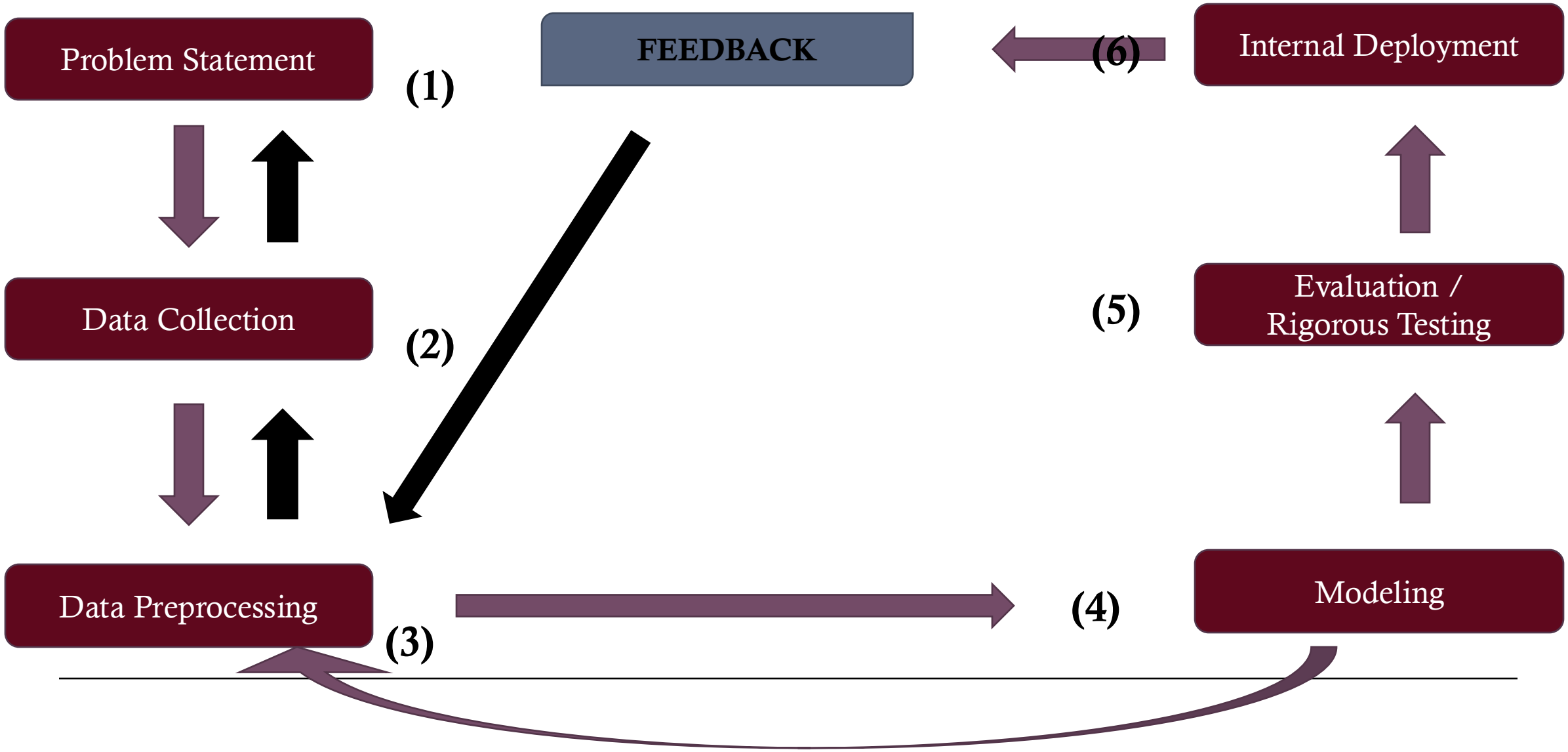
TYPICAL DATA SCIENCE PROJECT ROADMAP

Continuous Development and evaluation



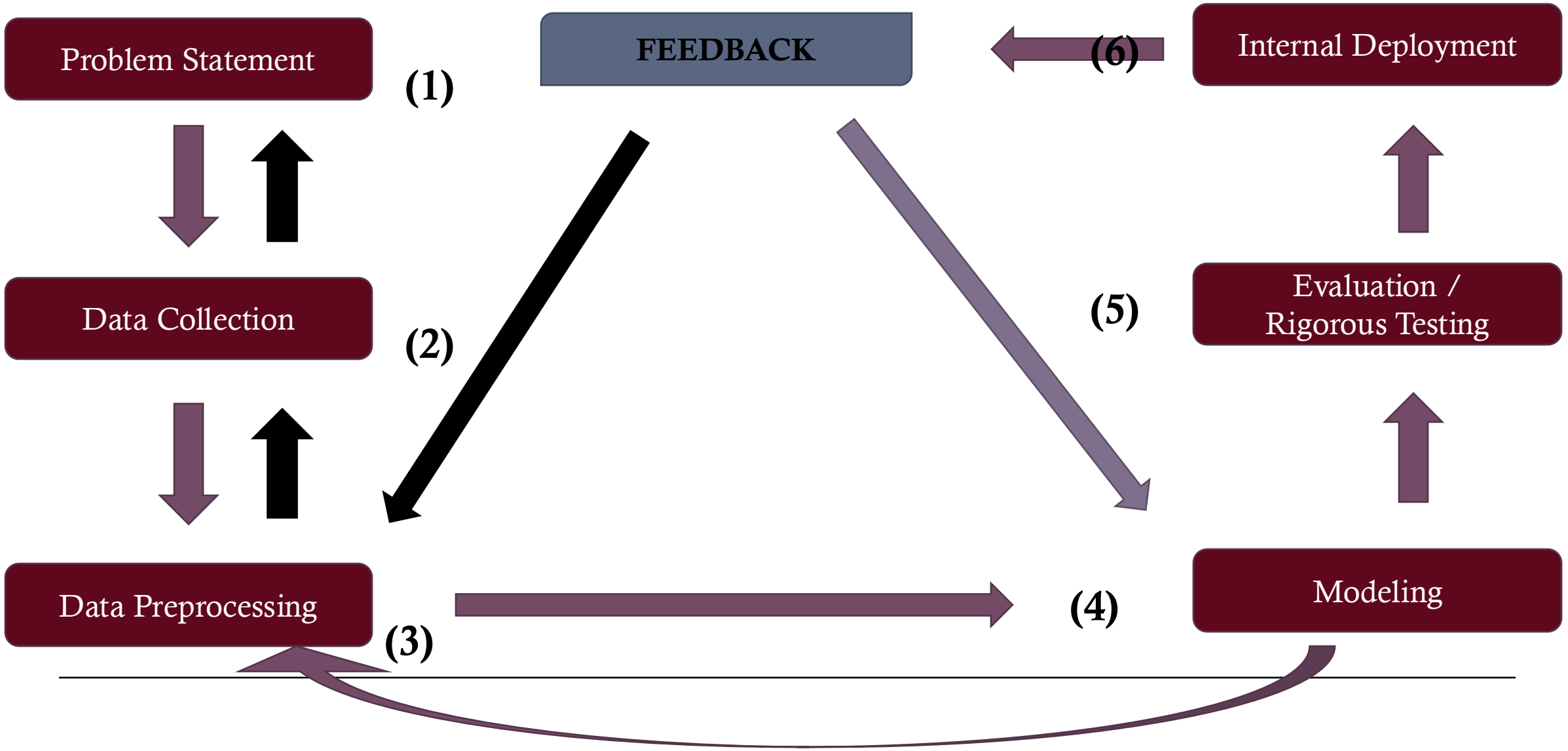
TYPICAL DATA SCIENCE PROJECT ROADMAP

Continuous Development and eveluation

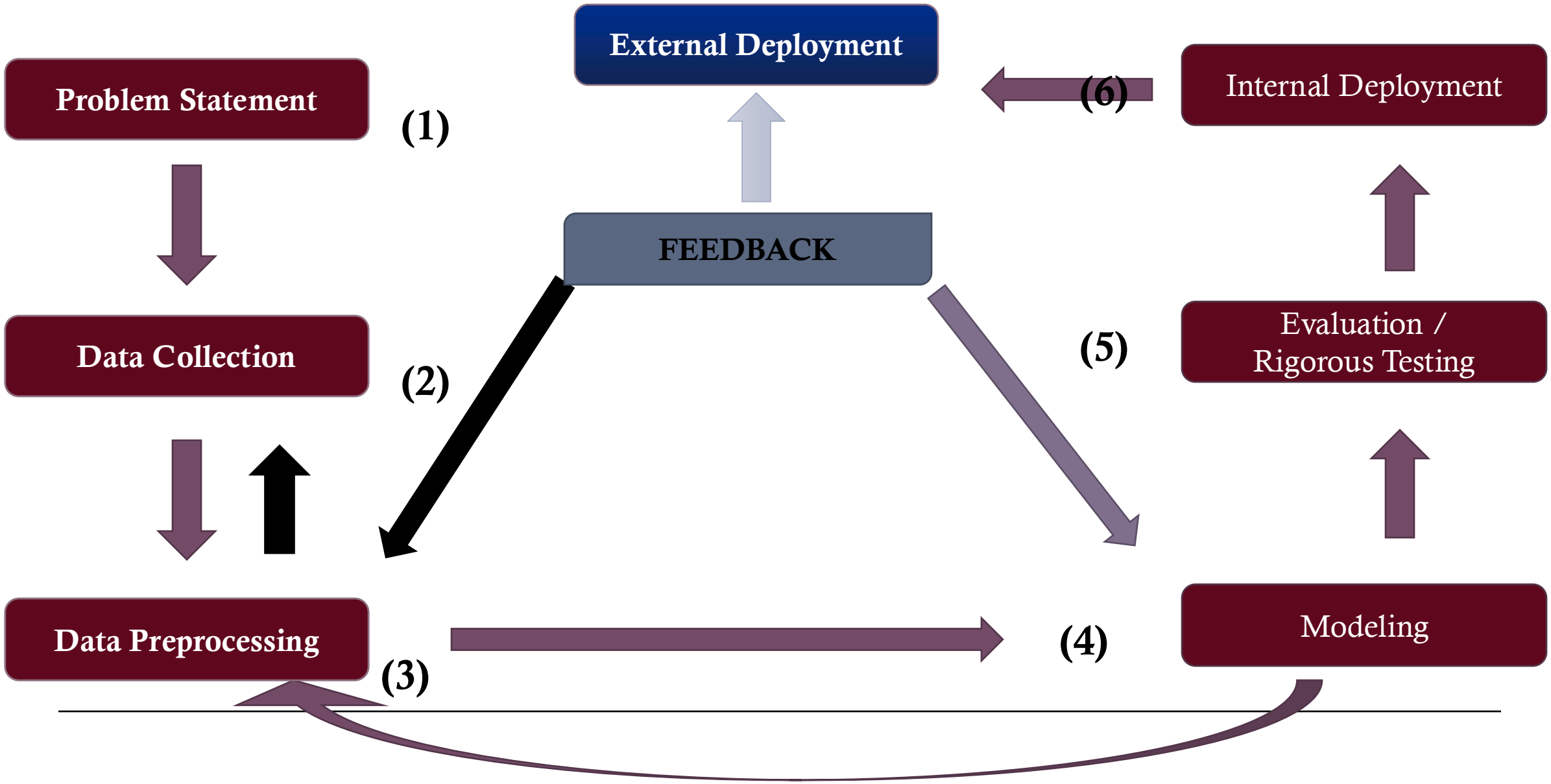


TYPICAL DATA SCIENCE PROJECT ROADMAP

Continuous Development and eveluation

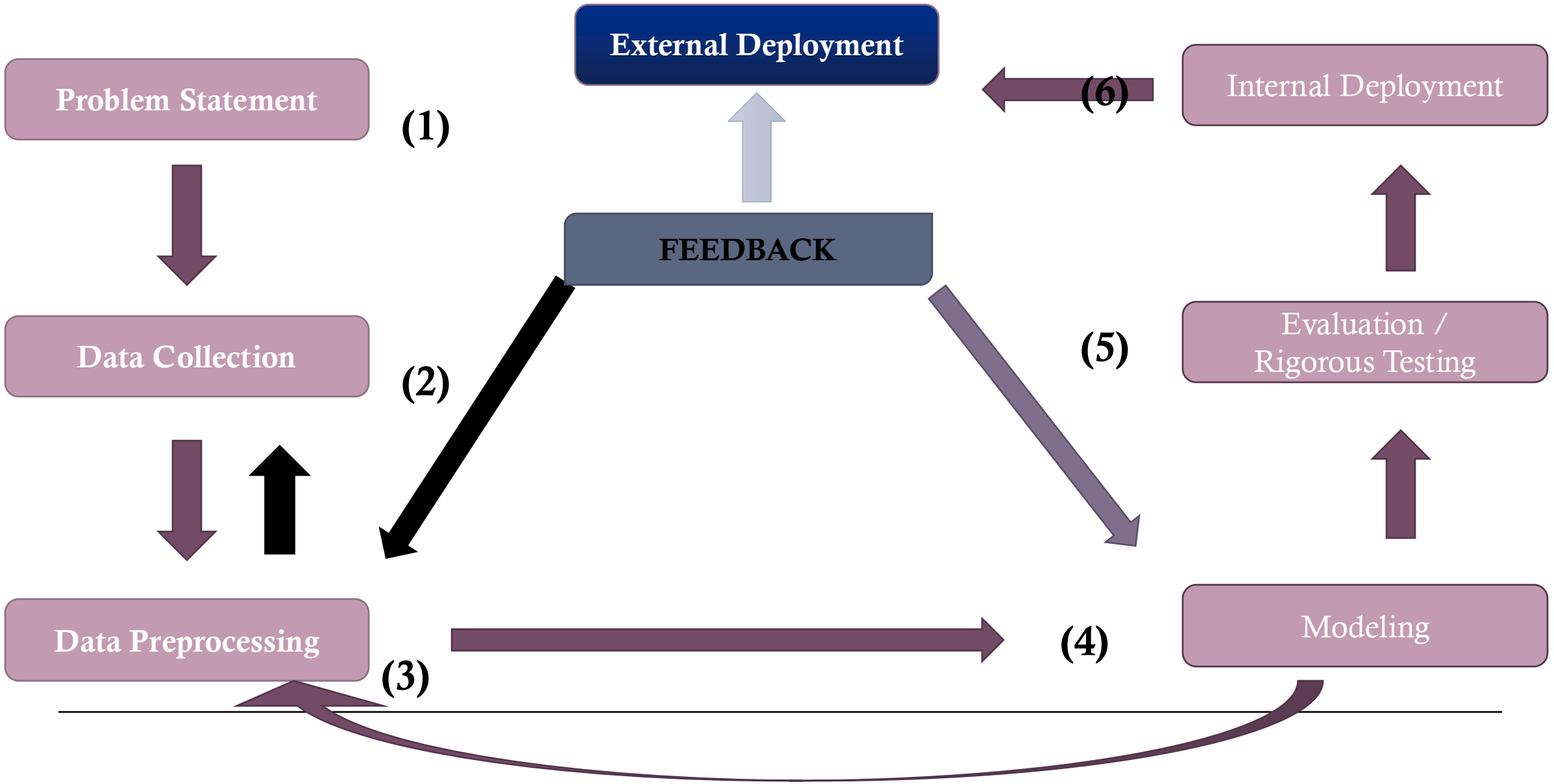


TYPICAL DATA SCIENCE PROJECT ROADMAP



TYPICAL DATA SCIENCE PROJECT ROADMAP

Continuous Development and eveluation



WHAT WE WILL DO.

- Principles of data modeling
- Data Reading methods,
- Data visualization techniques,
- Data cleaning procedures,
- Feature Engineering techniques
- Reporting/story telling on

ON ..

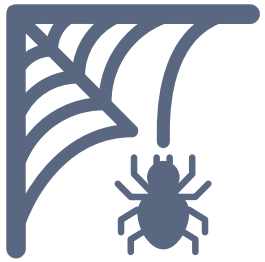
Tabular (IID) data,

Time series

Text data (if time allows)

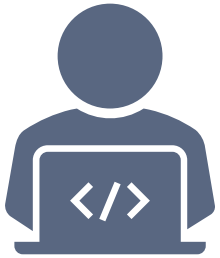
Image data (if time allows)

PART 0: PREREQUISITES



1. Get to know the Tools

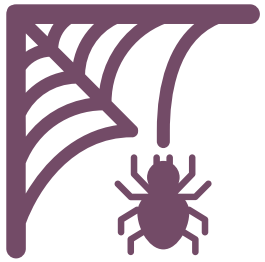
(VS Code , Jupyter Notebook, Colab, bash scripting, Linux, Environment and Packages management



2. Python essentials for this course

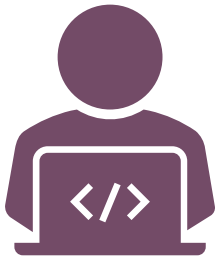
(Functions, classes) **but this is not the focus**

PART 1: DATA MODELING



1. Data modeling

How do you represent what you want to model



2. Python essentials for data modeling

Dataclasses and Pydantic datamodels

PART 1: WORKING WITH TABULAR DATA

1. Working with Series and DataFrames

2. Data Reading Methods

3. Introducing Features and Observations

4. Handling Text Data

5. Grouping the Data

6. Basic Data Explorations

7. Data Organization Methods

8. Customizing Functions

PART 2: WORKING WITH TIME SERIES

1. WORKING WITH TIME DATA

2. BASIC DATA MANIPULATION ON TIME SERIES

3. ADVANCED MANIPULATION ON TIME SERIES

TIPS FOR SUCCESS

1. It is a practical data manipulation course, **IT IS NOT A PROGRAMMING COURSE.**

2. You should focus on :
   - Understanding what the data is about and what to do with it.
   - Asking the right questions to extract value from the data.

TIPS FOR SUCCESS

3. Do not focus on :

- Getting the Python codes right. I will provide the Python codes for you to play with. However, I care about self-documenting code, it's good hygiene!!!

4. Do not write code while I am teaching, RATHER ASK QUESTIONS.

TIPS FOR SUCCESS

1. DO NOT BE SHY !!!!

TIPS FOR SUCCESS

1. DO NOT BE SHY !!!!

2. Ask Questions

TIPS FOR SUCCESS

1. DO NOT BE SHY !!!!

2. Ask Questions

3. Ask Questions again

TIPS FOR SUCCESS

1. DO NOT BE SHY !!!!

2. Ask Questions

3. Ask Questions again

4. Ask Questions again and again

PREPROCESSING DATA IS PART OF MODELING. WHAT IS MODELING ?

WHAT DO WE MEAN BY **MODELING**?

Modeling means building a simplified representation of something: (an object, system, or process, to understand and predict its behavior).

Examples:

- Language Modeling : ChatGPT is a large language model, a simplified representation of how human language flows.
- Disease Modeling: Building a simplified representation of how diseases spread within a population.
- Fluid Modeling: Building a simplified representation of how a liquid flows under different pressures within a given environment

THE MISSING STEP IN (DATA) MODELING

Step 1: Assume an object Step 2: **Model its behavior**

But what happens between Step 1 and Step 2?

How is the object itself *represented* in our data? This representation is not a given; it is a fundamental modeling choice.

SOME EXAMPLES

Fluid dynamics:

Behavior Model: Navier-Stokes Equations (how velocity & pressure evolve).

The Missing Question: Before we solve these equations, **how is the fluid *represented* in our data?** As a continuous field? A set of discrete particles? A mesh of finite volumes?

Financial markets

Behavior Model: Black-Scholes Equation (how prices evolve).

The Missing Question: Before applying the equation, how is the market represented? As a price series? An order book? A network of transactions?

WHY AND HOW?

These scientific models are brilliant at answering **'how' and 'why' questions about behavior**:

- *How* will this rocket move when the engines fire?
- *Why* does a storm system evolve so unpredictably?
- *How* will air flow over a new airplane wing?
- *What* is a fair price for this stock option given market volatility?

However**, Someone has already correctly defined and structured the fundamental entities involved.** So, there is missing learning step here!

THE MISSING STEP IN (DATA) MODELING

Step 1: Assume an object Step 2: **Model its behavior**

- The choice of representation enables the behavior model.
- The representation defines what "behavior" we can even see. This part of the course is about …

Step 0: Data-Centric Representation.

THE "MISSING STEP": REPRESENTATION

This is precisely where data modeling comes in. Before we can apply a complex behavioral model, we must first answer the fundamental question: '**What is this thing, structurally?'**

- Before modeling disease spread (behavior), we must define what a patient is(**structure: id, age, vaccination_status, location**).
- Before modeling trader behavior in a market, we must define what a trader is(**structure: name, stall_id, goods_sold, years_experience**).
- Before predicting crop yield (behavior), we must define what a farm is(structure: **location, size, soil_type, crop_variety**).

Data modeling provides the essential structural foundation upon which all behavioral analysis is built. It's the blueprint that defines our objects so our equations and algorithms have something meaningful to act upon."

OBJECTS ARE OFTEN MULTI-DIMENSIONAL

A Farmer in Northern Ghana:

- **Biographical**: Name, Age, Village
- **Financial**: Annual revenue, Access to credit, Mobile money usage
- **Agricultural**: Crops grown, Farm size, Irrigation access, Years of experience
- **Social**: Cooperative membership, Family size, Education level

A Maize Crop in Nigeria:

- **Environmental**: Soil pH, Rainfall amount, Average temperature
- **Agricultural**: Seed variety, Planting date, Fertilizer type used
- **Economic**: Market price, Transportation cost, Yield per hectare

The same object can be represented in dozens of different ways depending on what we are trying to understand.

MODELING IS CHOOSING RELEVANT DIMENSIONS

We cannot capture every possible attribute. Effective data modeling is the art and science of selecting the *most relevant* dimensions for our specific problem.

Contrasting Examples:

To predict a student's exam performance in Accra:

- **Relevant:** Past grades, Hours studied, School attendance, Access to textbooks
- **Probably Irrelevant:** Shoe size, Favorite music genre, Mother's maiden name

To predict mobile money adoption in Abidjan:

- **Relevant:** Age, Smartphone ownership, Trust in digital systems, Proximity to agent
- **Probably Irrelevant:** Blood type, Height, Favorite football team

THE ROLE OF DOMAIN KNOWLEDGE

Agricultural Context:

- *Common Sense:* Rainfall affects crop growth
- *Domain Knowledge (from a Senegalese farmer):* "The timing of the first rains matters more than the total amount. Also, the type of soil determines which millet variety will thrive."

Healthcare Context:

- *Common Sense:* Symptoms indicate disease
- *Domain Knowledge (from a Nigerian doctor):* "For malaria, we need travel history to endemic regions, previous infection history, and access to bed nets - not just current fever temperature."

Financial Context:

- *Common Sense: Income determines loan repayment*
- *Domain Knowledge (from a Ghanaian loan officer):* "For small traders, consistency of daily sales and membership in a market association are better predictors than total monthly income."

Key Message: Domain experts help us separate signal from noise in our attribute selection.

WHAT HAPPENS WHEN WE CHOOSE WRONG?

Selecting the wrong dimensions leads to models that don't work in the real world.

Real West African Consequences:

The Overly Simple Model:

Problem: Trying to predict crop yields using only "farm size"

Reality: A 1-hectare farm with irrigation and good soil can outproduce a 5-hectare farm with poorsoil and no water

Result: Poor predictions, wasted resources

The Kitchen Sink Model:

Problem: Collecting 100+ attributes because "more data is better"

Reality: Most attributes are irrelevant, making analysis complicated and expensive

Result: "Curse of dimensionality" - the model finds false patterns in the noise

The Culturally Blind Model:

Problem: Using Western attributes for African contexts

Example: Modeling "formal employment status" instead of capturing the diverse income streams common in West African informal economies

Result: The model misses the actual economic reality

MODELING MOBILE MONEY USAGE ACROSS TOUNOUNGA MARKET

```
user_1 = ['Amina', 'Ghana', 23, 'MTN Momo', '0501234567', 450.50]

user_2 = {'name': 'Chijioke', 'country': 'Nigeria', 'age': 'thirty-five', 'provider': 'OPay'}

user_3 = ['Kwame', 31, 'Vodafone Cash', 200.00, '233201234567']
```

- What do you observe?
- What's missing?
- Wrong types?
- Unstructured data = confusion.

THE SOLUTION: DATA MODELS

A Data Model defines:

- What fields exist.
- Their types.
- Validation rules (e.g., age > 0).

Benefits:

- Consistency and validation.
- Faster preprocessing.
- Reliable insights.

COURSE ROADMAP

- Small 5 mins test everyday
- 2 Quiz per week
- 1 assignment per week