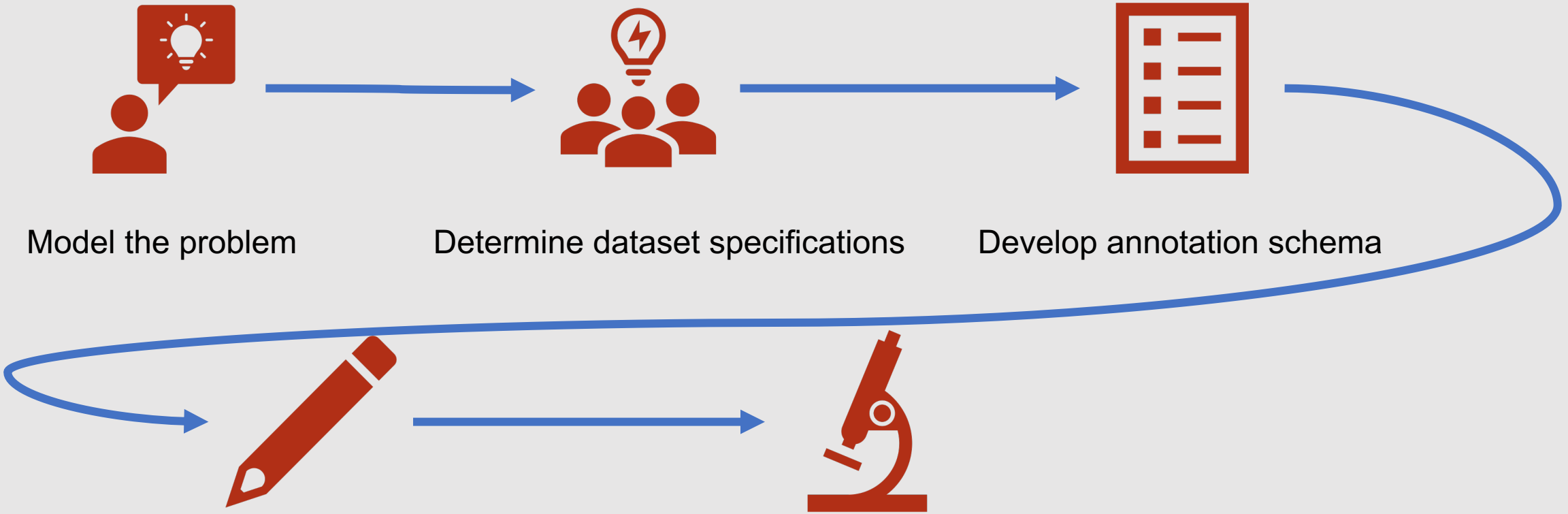
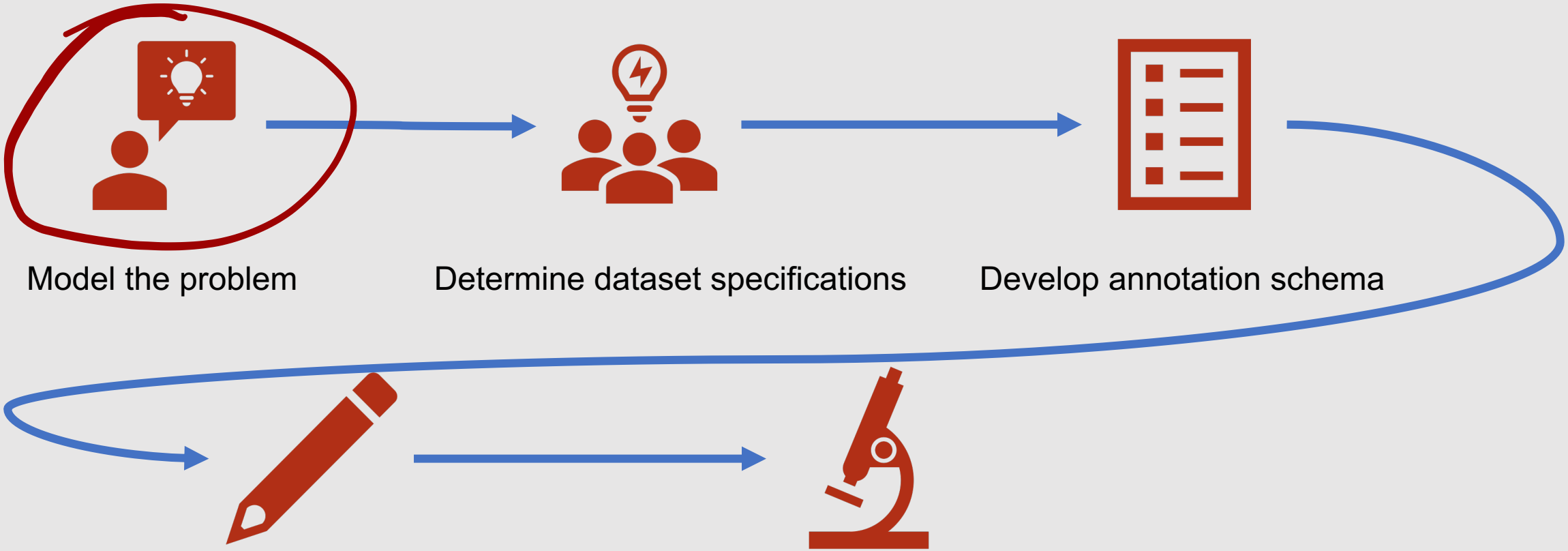
Modeling the Problem

Natalie Parde
UIC CS 521

Typical Data Collection Pipeline



Typical Data Collection Pipeline



Modeling the Problem

- Define a clear annotation goal for your task!
- Answer key questions:
  - What are you trying to do?
  - How are you trying to do it?
  - Which resources best fit your needs?
- Building a dataset is an iterative process …your answers may change as you progress

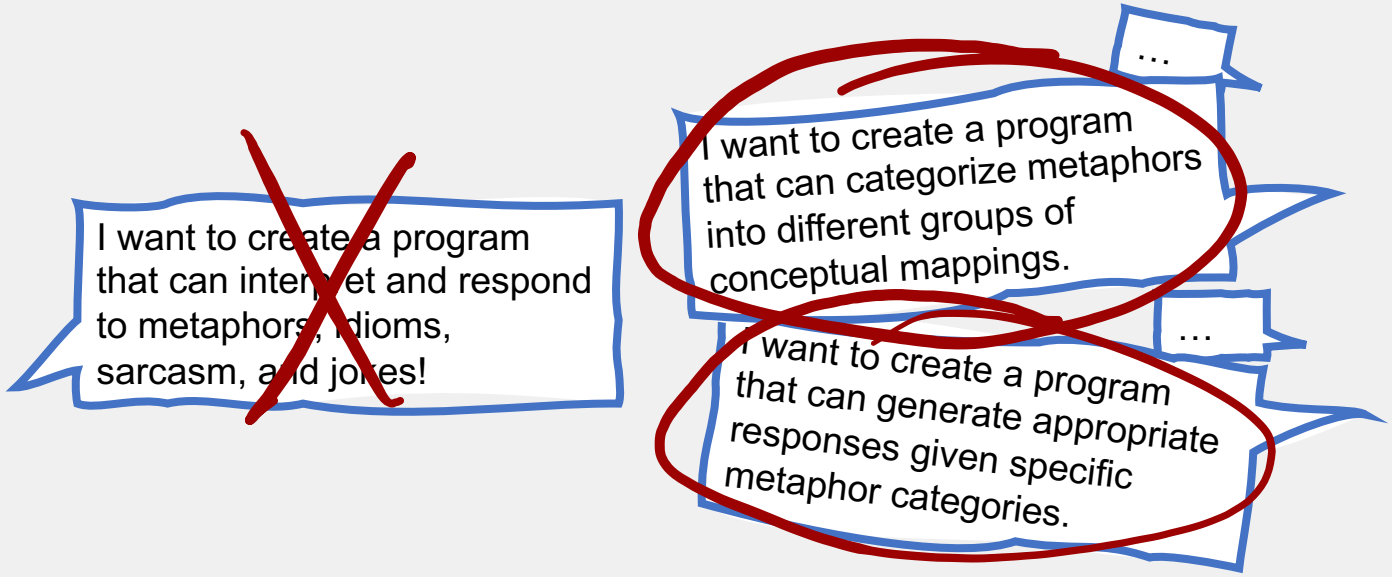
Goal Definition

Write a statement of purpose

Expand upon how you will achieve it

Statement of Purpose

- 1-2 sentence summary describing intended line of research
- Break the task into manageable segments, if necessary
  - It may be unrealistic to collect annotations for everything of interest all at once



Sample Summaries for Well-Known Corpora

Corpus	Summary
Penn Discourse TreeBank	Labels discourse relations between eventualities and propositions in newswires, for learning about discourse in natural language
MPQA Opinion Corpus	Labels opinions with characteristic traits, for use in evaluating emotional language
TimeBank	Labels times, events, and their relationships in news texts, for use in temporal reasoning

How do we move from statement of purpose to longer task description?

Need to balance informativity vs. correctness

- **High informativity:** Annotations are very useful for your task
- **High correctness:** Annotation task is not difficult for annotators to complete accurately

Often the two are at odds with one another!

- With very precise categories (**high informativity**), annotators may easily miss the correct label or make labeling errors (**low correctness**)
- With limited categories (**high correctness**), labels may be less useful for the task of interest (**low informativity**)

These factors are closely related to project scope.

- Two main types of scope in this context:
  - **Scope of the annotation task**
    - How far-reaching is the annotation goal?
  - **Scope of the corpus**
    - How much will be covered?

Scope of the Annotation Task

Define	Define different possible categories of relationships
Determine	Determine which will be most relevant to the task
Think	Think of the annotation task in terms of the classification task you are ultimately trying to solve • Note: Having many classes is not only difficult for annotators; it is also more challenging from a model training perspective

Scope of the Corpus

- What will be the data source(s)?
- Is a single source sufficient for developing generalizable methods for your task?
- Will your corpus need to cover multiple text styles (e.g., tweets and news articles) and/or genres (e.g., self-help and computer science)?
  - Do you need different annotation guidelines for different text styles/genres?