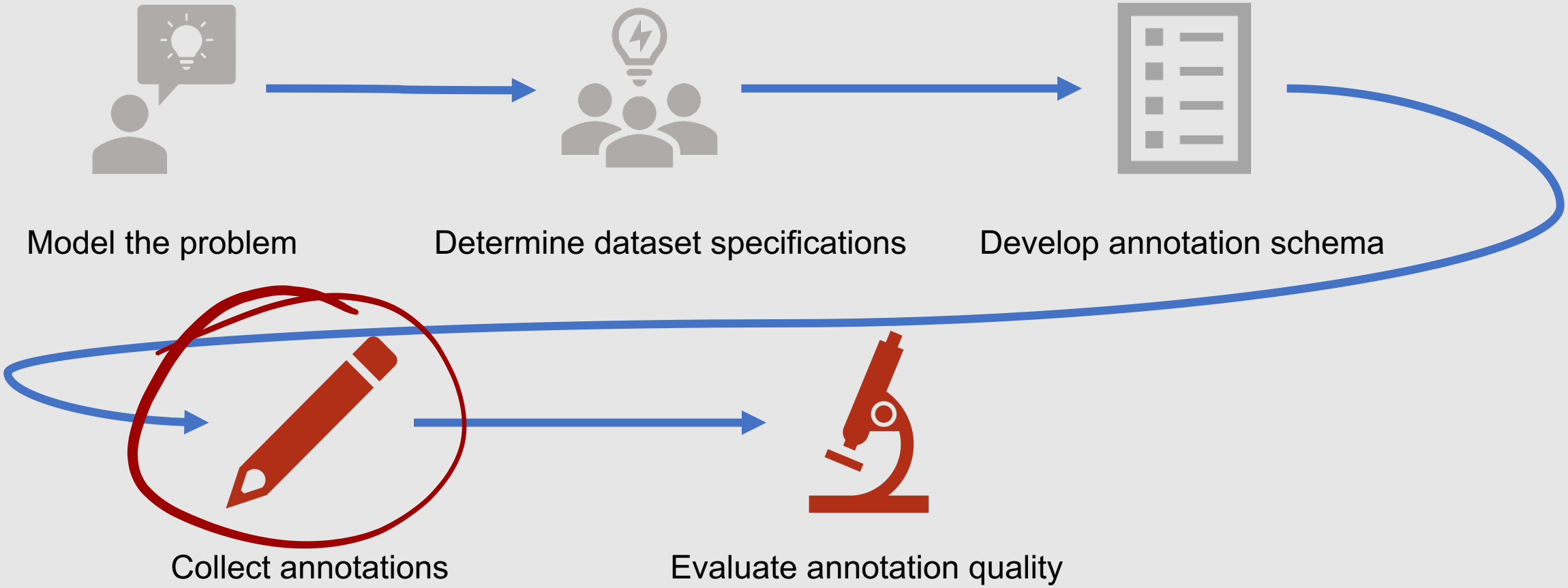
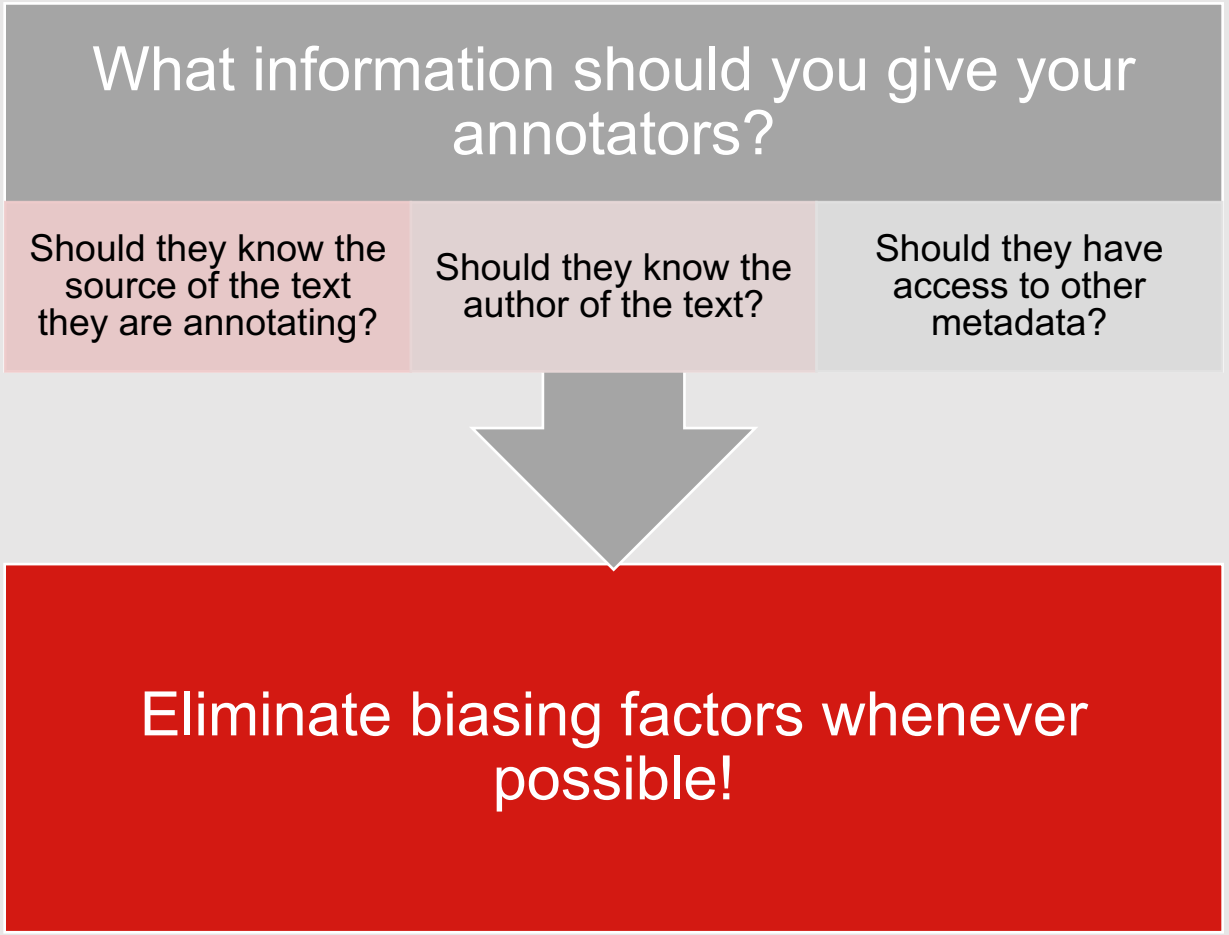
Collecting Annotations

Natalie Parde
UIC CS 521

Typical Data Collection Pipeline



Preparing Data for Annotation



This review received 3 stars

This tweet was written by a user from North America

This article was flagged as being biased

Example Biasing Factors

Preprocessed Data

- Should you give annotators data that already has some information marked up?
  - Presenting annotators with too much information can lead to confusion
  - However, some information can be useful
- For some tasks, you can automatically assign labels and then ask annotators to correct them

Organizing Annotations

- Decide ahead of time how your annotations will be formatted and stored
  - Markup labels
  - CSV file
- Make sure you have a consistent internal system for linking annotations to source and metadata

Selecting Annotators

Does the annotation task require any specialized knowledge?

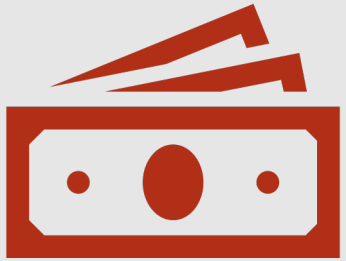
- Background expertise
- Language competency
- Demographic characteristics

What resources do you have available?

- Time
- Money
- Dataset size

Specialized Knowledge

- For tasks requiring close reading or familiarity with colloquialisms, **native proficiency in the target language** may be necessary
- For tasks using some domain-specific data, **advanced training** may be necessary to comprehend the text
- For tasks concerned with specific subsets of language, **residence in specific regions** may be necessary to understand the task



Resource Availability

- Most people can only focus on an annotation task for a few hours at a time
- Annotators will get better at the task with practice
- Financial resources may limit your ability to hire experts

In-Person Annotators

Pros:	Generally available for longer periods of time Can provide feedback one-on-one Easier to provide with specific training
Cons:	Take longer to complete annotations May be more subject to bias (e.g., from close knowledge of the project or the other annotators)

Crowdsourced Annotators

Pros:	Generally faster Less likely to be biased by close knowledge of the project/other annotators
Cons:	Generally less invested in the annotation task Cannot easily be trained with task-specific knowledge Minimal room for feedback May only complete a small number of annotations

Where to find annotators?

In-Person:

- Friends
- Lab mates
- Undergraduates studying linguistics or psychology
- Individuals with task-specific expertise (e.g., medical doctors if annotations are needed for clinical notes)

Crowdsourced:

- Amazon Mechanical Turk: https://www.mturk.com/
- Appen: https://appen.com/

Annotation Environments

- Many different tools exist!
  - Multipurpose Annotation Environment (MAE)
    - https://github.com/nathan2718/mae-annotation-1
  - General Architecture for Text Engineering (GATE)
    - https://gate.ac.uk/sale/tao/split.html
  - WebAnno
    - https://webanno.github.io/webanno/
  - INCEpTION
    - https://inception-project.github.io/

Key Considerations

- Make sure that the annotation environment:
  - Works on all the computers you and your annotators will be using
  - Supports the type of annotations you need
  - Includes any extra support features you need
- **Don’t neglect UI elements!**
  - Try to ensure that your annotation guidelines are easily accessible
  - Make sure that the environment is easy to install and easy to use