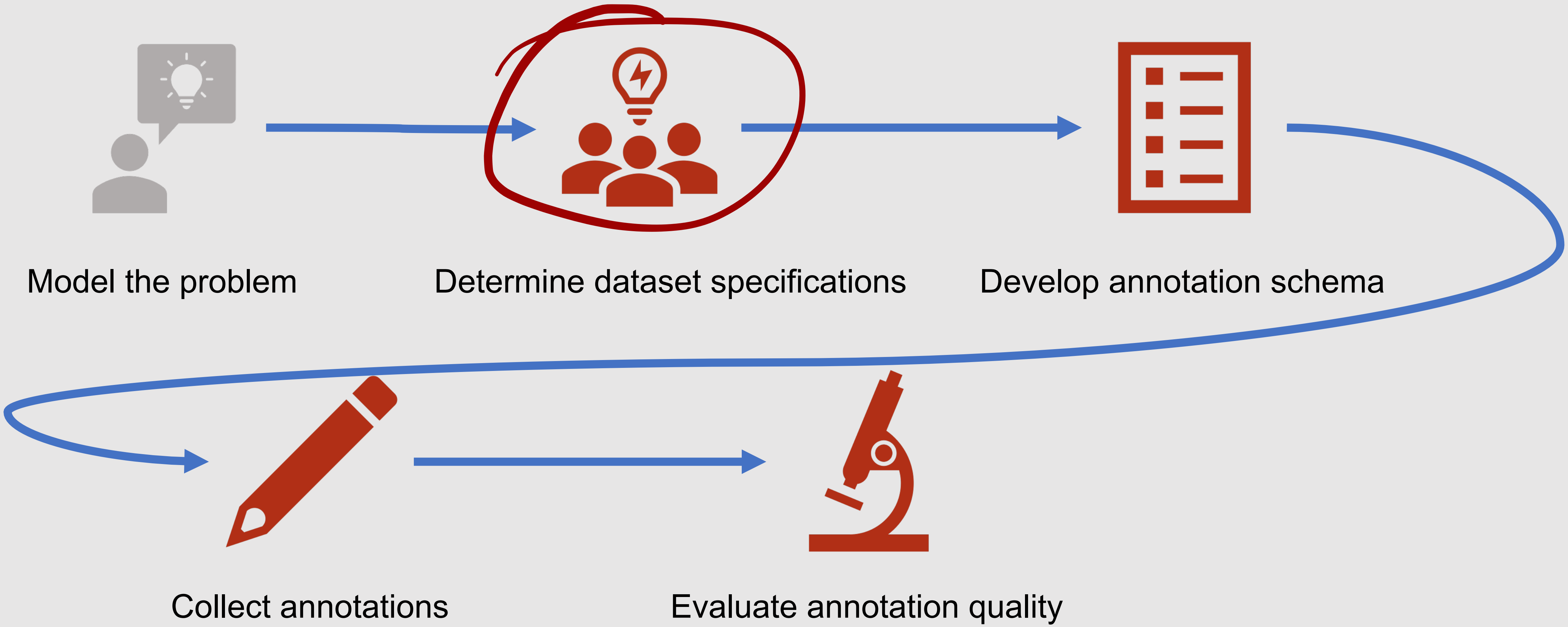
Determining Dataset Specifications

Natalie Parde

UIC CS 521

Typical Data Collection Pipeline



Model the problem

Determine dataset specifications

Develop annotation schema

Collect annotations

Evaluate annotation quality

Dataset Specifications

- Conduct background research!
  - See what related corpora (if any) already exist for your task
    - Same task, different domain?
    - Same domain, different task?
- Review existing annotation schemes in areas close to your planned dataset

Where to look for related corpora?

Linguistic Data Consortium	https://www.ldc.upenn.edu/
European Language Resources Association	http://www.elra.info/en/
Linguistic Resources and Evaluation Map	http://lremap.elra.info/
Google Dataset Search	https://toolbox.google.com/datasetsearch
AWS Open Data Registry	https://registry.opendata.aws/

Other places to search….

NLP Conferences

- LREC
  - http://www.lrec-conf.org/
- ACL Anthology
  - https://www.aclweb.org/anthology/

NLP Challenges

- SemEval
  - http://alt.qcri.org/semeval2020/index.php?id=tasks
- CoNLL Shared Task
  - https://www.conll.org/previous-tasks

Determine Dataset Sources

- Ensure that sources are representative of the domain you're trying to model
- Make sure to fully document:
  - From where the data was obtained
  - Why it was selected
- Try to keep the corpus **balanced** across your desired annotation categories

Planning to make the dataset public?

- Make sure you have permission!
- Decide what type of license you will use

Common Data Sources in NLP

Books
Project Gutenberg

News Articles
News websites

Blogs

Social Media
Twitter

People

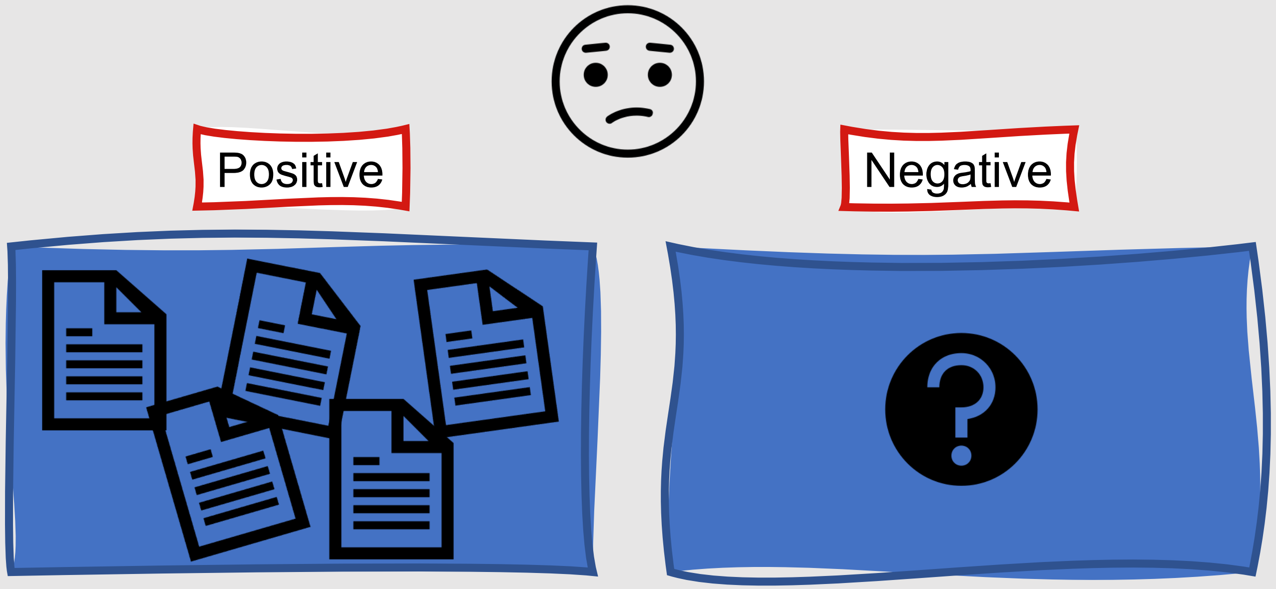
Collecting Data from People

- Common when you need to elicit speech or text samples from humans performing some specific task
- Usually requires **IRB approval**
- Two main types of data can be collected from people:
  - **Read Speech:** Have each person read the same set of sentences or words out loud
  - **Spontaneous Speech/Text:** Give people prompts or ask them open-ended questions, and collect their responses

Achieving a Representative and Balanced Corpus

Representative: The corpus contains samples of all text categories

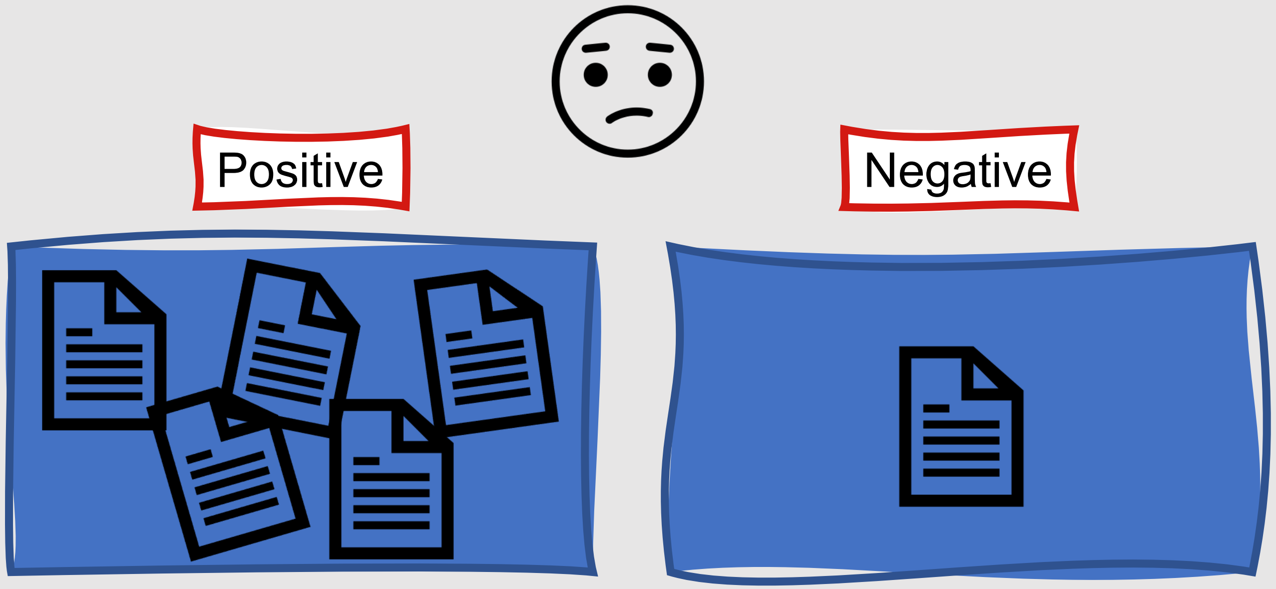
Balanced: The corpus contains realistic (in many cases, equal is ideal) distributions of all text categories



Achieving a Representative and Balanced Corpus

Representative: The corpus contains samples of all text categories

Balanced: The corpus contains realistic (in many cases, equal is ideal) distributions of all text categories



Achieving a Representative and Balanced Corpus

Representative: The corpus contains samples of all text categories

Balanced: The corpus contains realistic (in many cases, equal is ideal) distributions of all text categories

Positive
Negative

Corpus Size

- How much data are you going to collect and annotate?
  - Generally, more data → better
  - However, more data also → more time and money
- **Start small** and see how your annotation task and guidelines work before scaling up
- **Refer to similar existing corpora** to get a general idea of desired corpus size