Data Collection

Natalie Parde, Ph.D.

Department of Computer Science

University of Illinois at Chicago

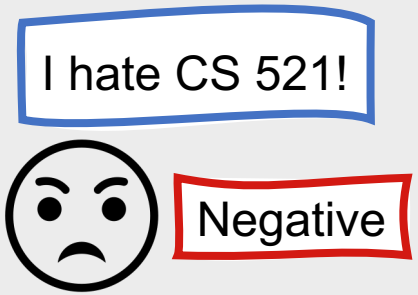
CS 521: Statistical Natural Language Processing

Spring 2022

Many slides adapted from: Pustejovsky, J., & Stubbs, A. (2012). *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. O'Reilly Media, Inc.

What is data collection?

- The process of curating data and assigning labels to it
- Essential for statistical natural language processing!



Text collections for NLP tasks are typically referred to as corpora.

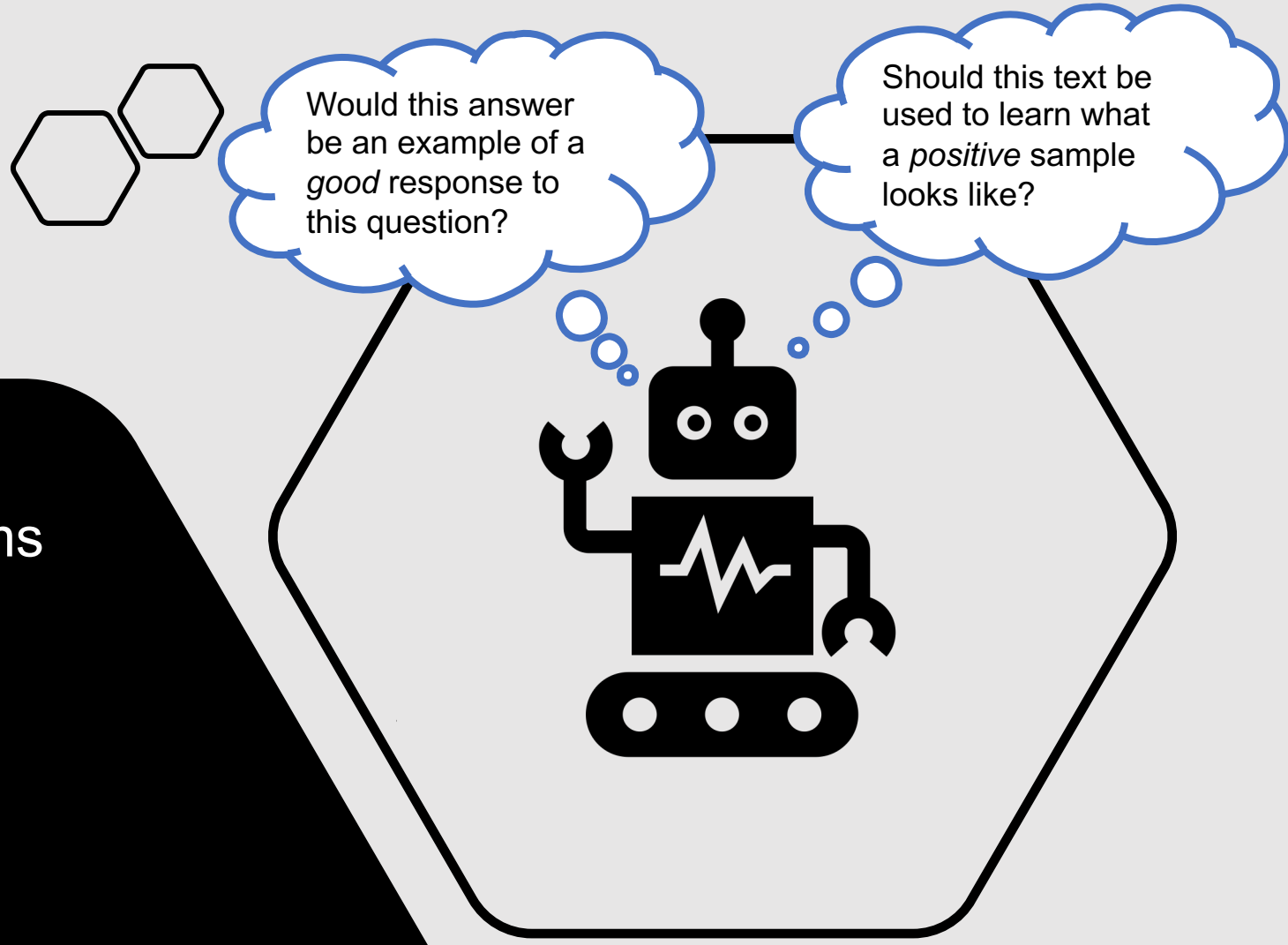
- Singular: **Corpus**
- Corpora come in many forms
  - Word-level annotations
  - Sentence-level annotations
  - Document-level annotations

Corpora can also focus on many different topics.

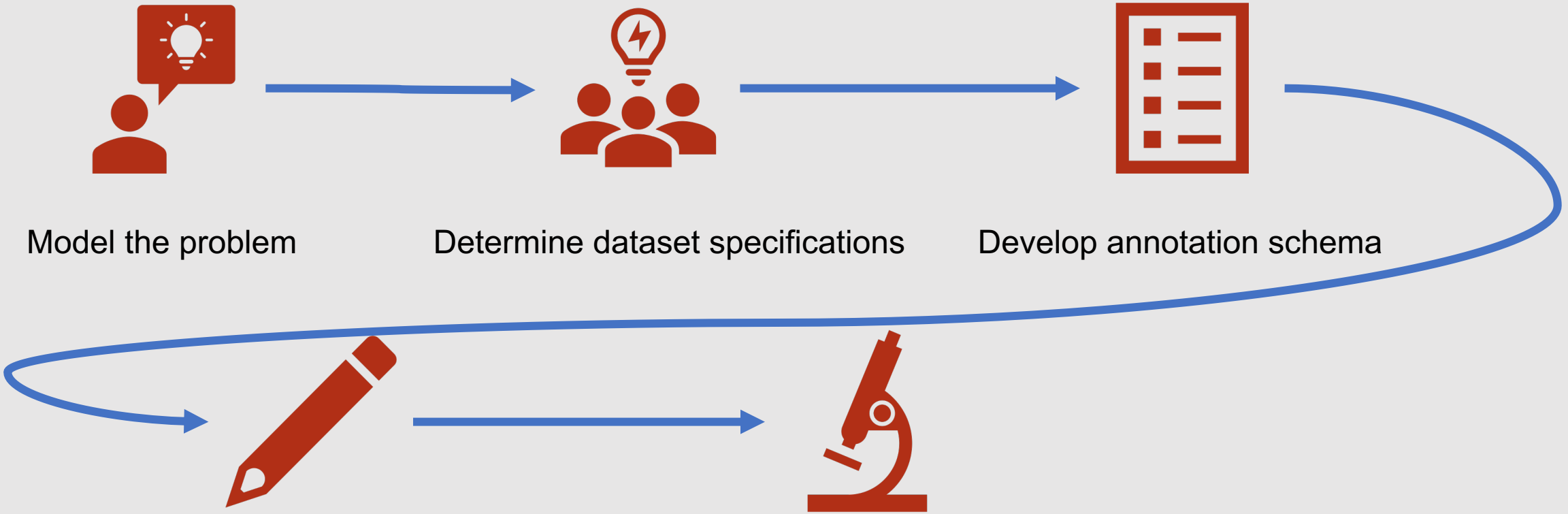
- **Sentiment Analysis**
  - IMDB Movie Reviews: http://ai.stanford.edu/~amaas/data/sentiment/
  - Amazon Reviews: https://snap.stanford.edu/data/web-Amazon.html
- **Question Answering**
  - WikiQA Corpus: https://www.microsoft.com/en-us/download/details.aspx?id=52419
  - Jeopardy! Corpus: https://www.reddit.com/r/datasets/comments/1uyd0t/200000_jeopardy_questions_in_a_json_file/
- **Syntax**
  - Penn Treebank: https://catalog.ldc.upenn.edu/LDC99T42
- **Dialogue Act Prediction**
  - Switchboard Dialog Act Corpus: http://compprag.christopherpotts.net/swda.html
- …and many more!

Why is data collection necessary?

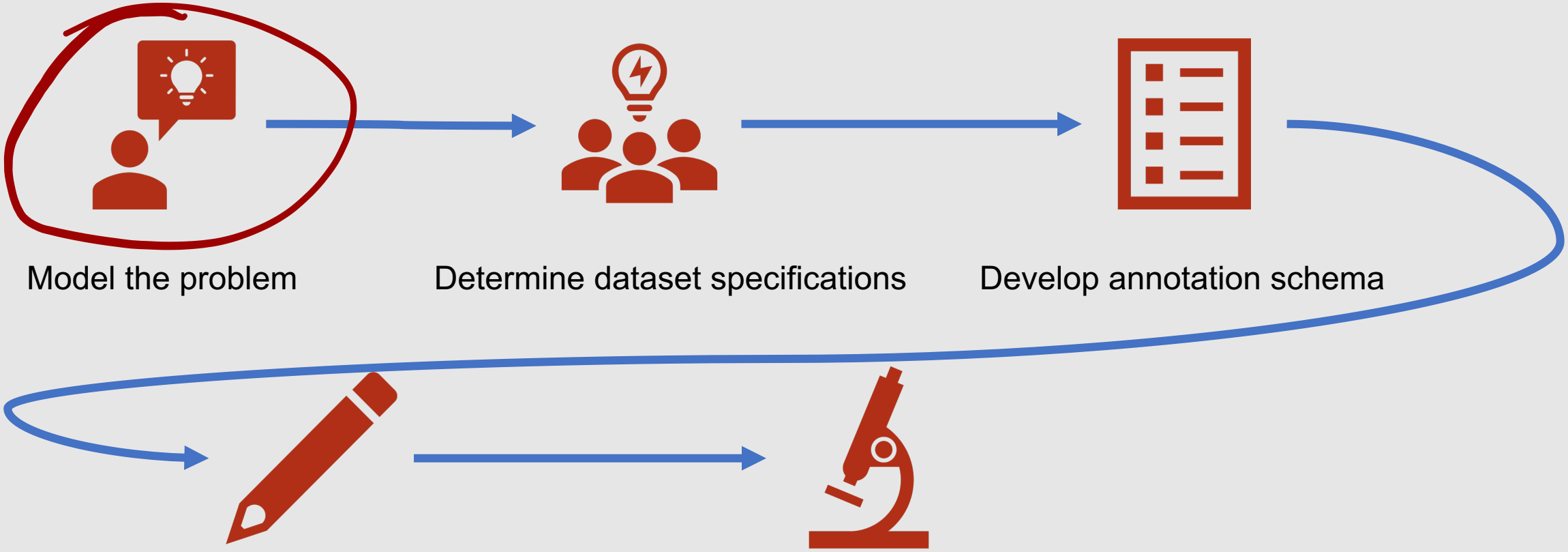
- Supervised learning algorithms require human assistance!
- Models need to know what to learn in order to succeed



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

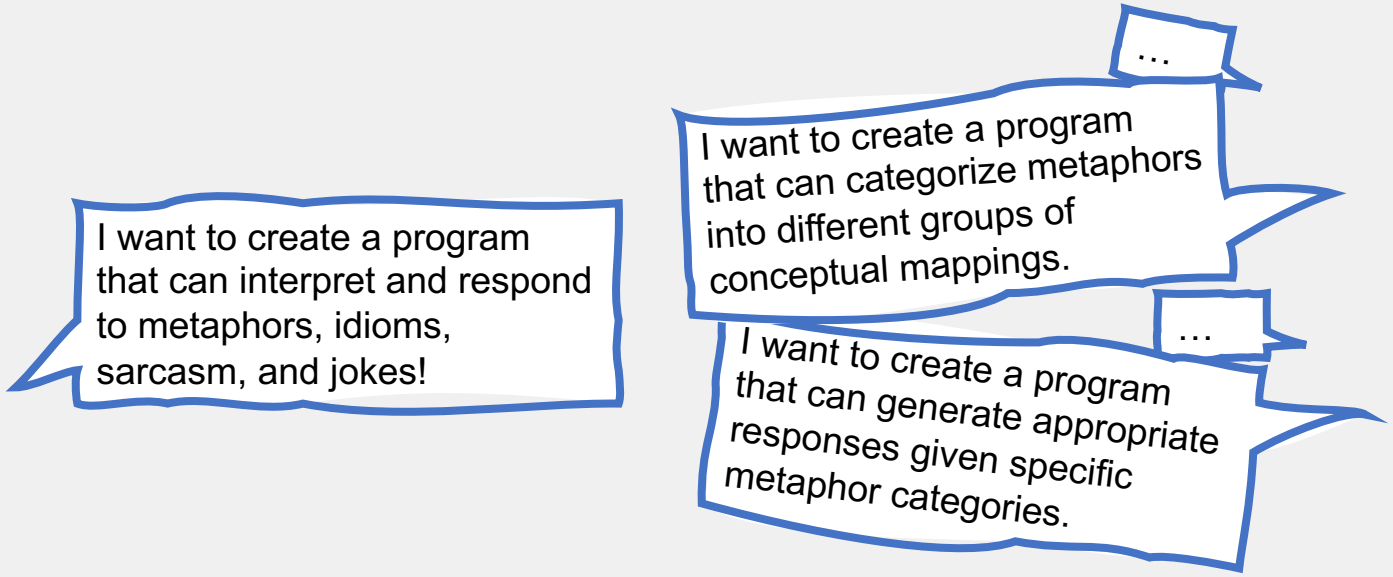
Goal Definition

Write a statement of purpose

Expand upon how you will achieve it

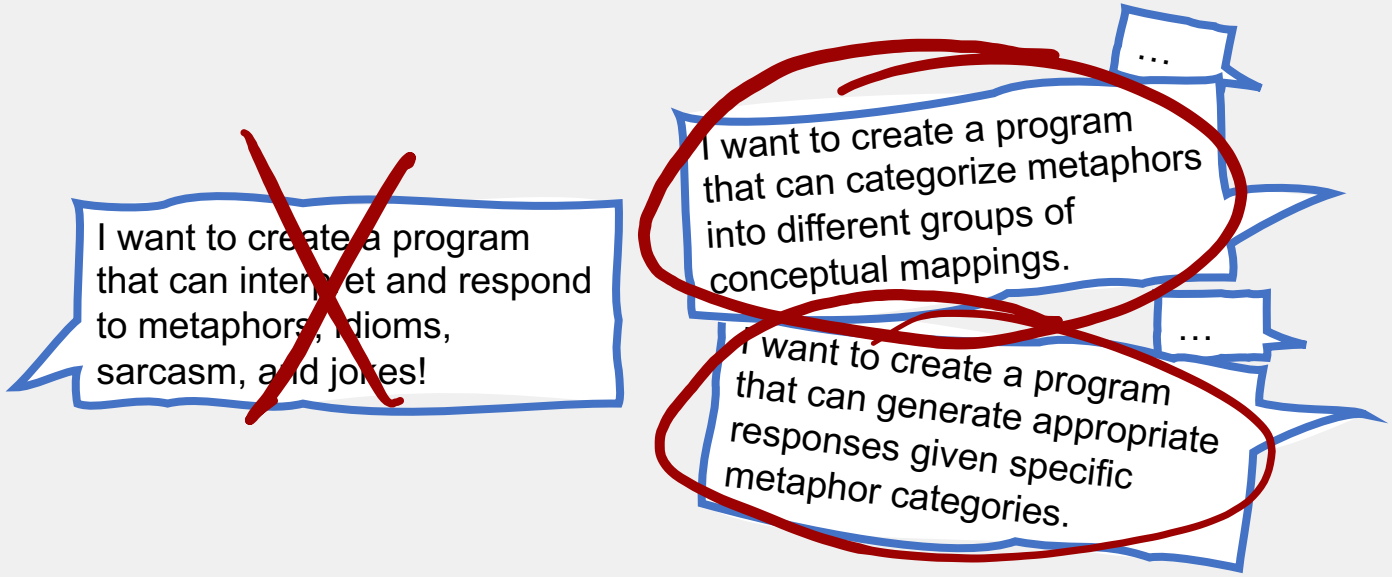
Statement of Purpose

- 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



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, 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?

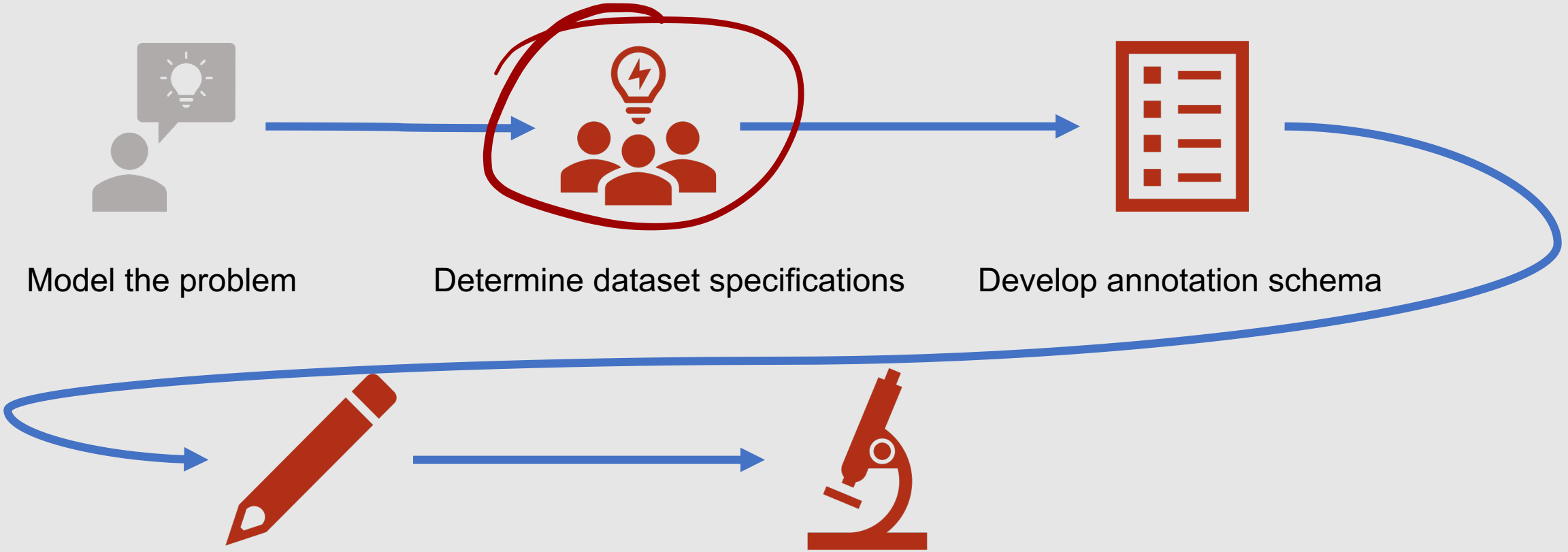
Scope of the Annotation Task

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?

Typical Data Collection Pipeline



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
  - https://semeval.github.io/SemEval2022/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

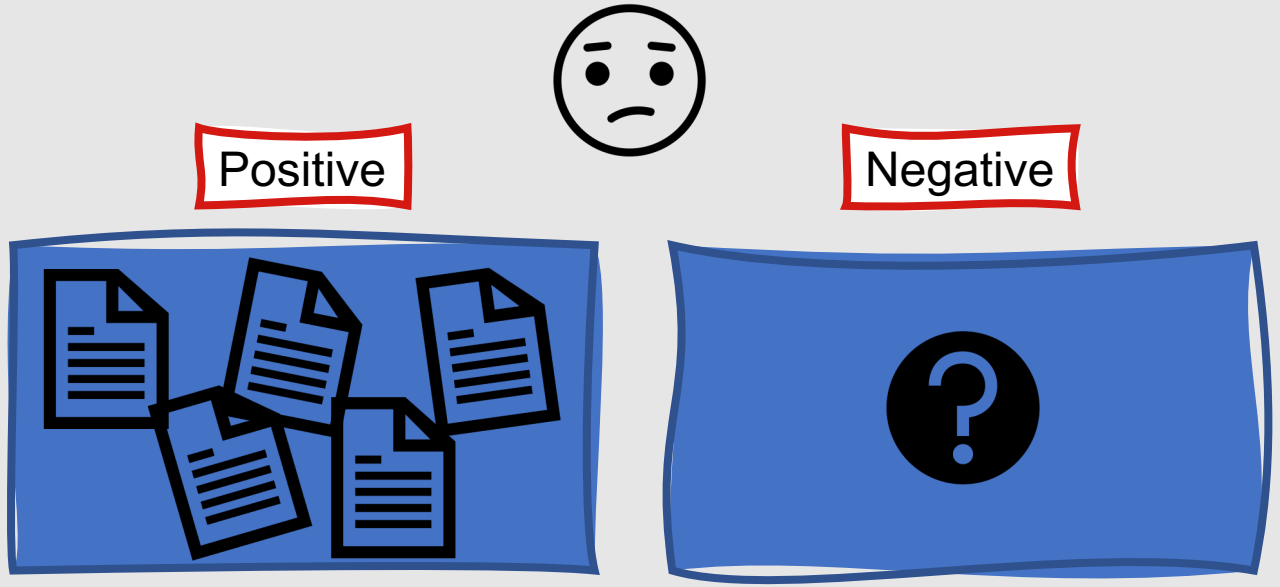
Collecting Data from People

- 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

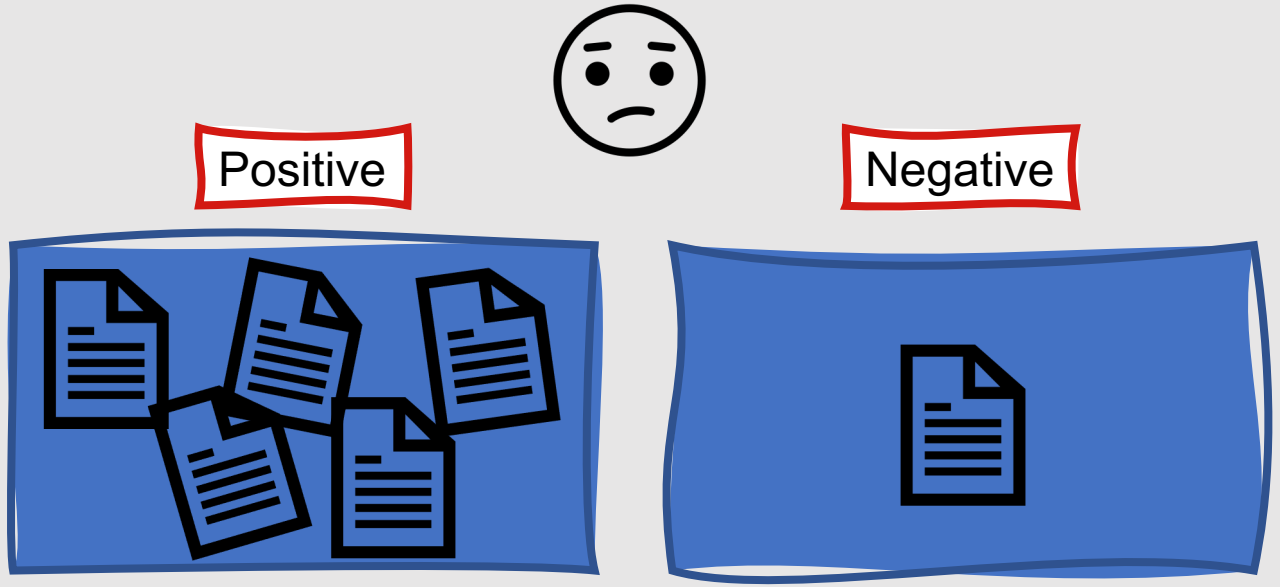
Balanced: The corpus contains realistic (in many cases, equal is ideal) distributions of all text categories



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



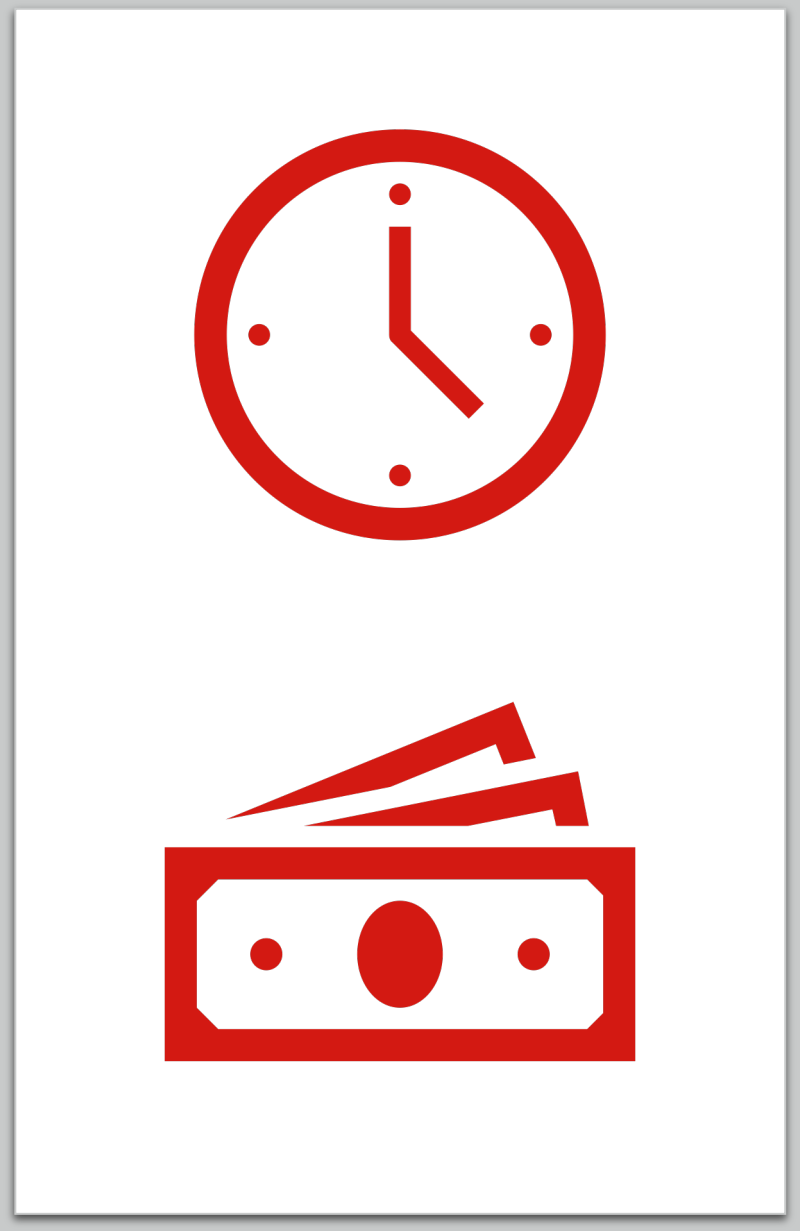
Achieving a Representative and Balanced Corpus

Representative: The corpus contains samples of all text categories

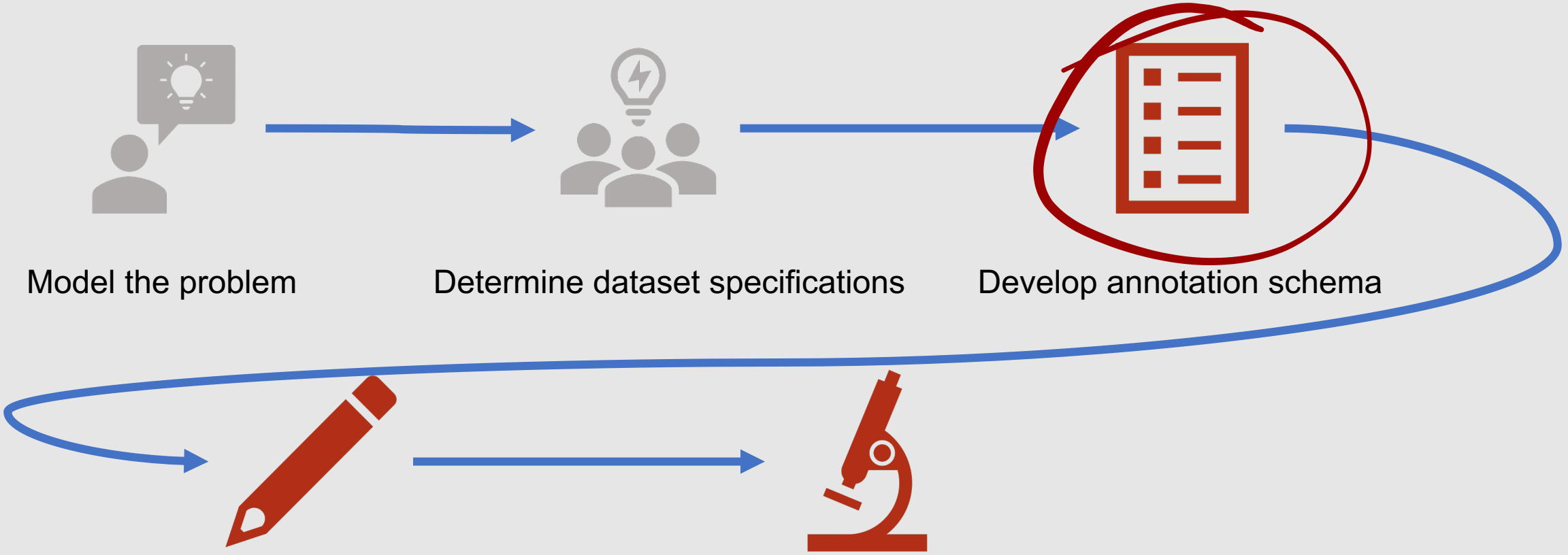
Balanced: The corpus contains realistic (in many cases, equal is ideal) distributions of all text categories

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



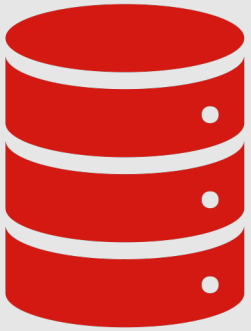
Typical Data Collection Pipeline



You know what you want your annotators to do …but how will they do it?

- What will the annotations **look like**?
- How are **different types of tasks** represented differently?
- How can you **maximize your corpus's usefulness** to the broader research community?
- What **constraints** are placed on your annotation environment and data format?
  - Consider both annotators and downstream machine learning task

Keep your data accessible!



Intricate database systems are difficult to set up (both for you, and for other people)

Complex XML schemas can be difficult for other researchers to understand

Defining Annotation Categories

Determine	Determine a reasonable number of annotation categories • Remember …lots of categories → more potential for errors and disagreement
Define	Concisely define each category
Include	If possible, include an example

Case Example: Emotion Classification

Natalie wants to collect emotion labels for a classification task. She has a dataset of 1000 tweets, and wants to label each tweet with the most dominant emotion that it expresses.

I'm so happy I get to take CS 521!!!! #bestclassever

oh no I registered late and got stuck in CS 521 😭

Wow I had no idea that UIC had a statistical NLP class 😮

Before she begins, she needs to decide on a reasonable set of annotation labels….

Case Example: Emotion Classification

Natalie wants to collect emotion labels for a classification task. She has a dataset of 1000 tweets, and wants to label each tweet with the most dominant emotion that it expresses.

I'm so happy I get to take CS 521!!!! #bestclassever

oh no I registered late and got stuck in CS 521 😭

Wow I had no idea that UIC had a statistical NLP class 😮

Before she begins, she needs to decide on a reasonable set of annotation labels. After doing some research, she decides to use Ekman's six basic emotion categories because they are easy to explain, not too numerous, and used in other datasets as well.

Happiness | Sadness | Surprise | Fear | Anger | Disgust

Case Example: Emotion Classification

Natalie wants to collect emotion labels for a classification task. She has a dataset of 1000 tweets, and wants to label each tweet with the most dominant emotion that it expresses.

I'm so happy I get to take CS 521!!!! #bestclassever

oh no I registered late and got stuck in CS 521 😭

Wow I had no idea that UIC had a statistical NLP class 😮

Happiness | Sadness | Surprise | Fear | Anger | Disgust

Now that she's decided on these labels, she needs to concisely define each one….

Case Example: Emotion Classification

Natalie wants to collect emotion labels for a classification task. She has a dataset of 1000 tweets, and wants to label each tweet with the most dominant emotion that it expresses.

I'm so happy I get to take CS 521!!!! #bestclassever

oh no I registered late and got stuck in CS 521 😭

Wow I had no idea that UIC had a statistical NLP class 😮

Now that she's decided on these labels, she needs to concisely define each one, so she creates the following table.

Happiness	Sadness	Surprise	Fear	Anger	Disgust
Indicates happiness or joy.	Indicates sadness or disappointment.	Indicates surprise or shock.	Indicates fear, stress, or panic.	Indicates anger or resentment.	Indicates disgust or unpleasantness.

Case Example: Emotion Classification

Natalie wants to collect emotion labels for a classification task. She has a dataset of 1000 tweets, and wants to label each tweet with the most dominant emotion that it expresses.

I'm so happy I get to take CS 521!!!! #bestclassever

oh no I registered late and got stuck in CS 521 😭

Wow I had no idea that UIC had a statistical NLP class 😮

To make sure her guidelines are clear, she also adds an example for each category.

Happiness	Sadness	Surprise	Fear	Anger	Disgust
Indicates happiness or joy.	Indicates sadness or disappointment.	Indicates surprise or shock.	Indicates fear, stress, or panic.	Indicates anger or resentment.	Indicates disgust or unpleasantness.
Yay I finished my homework three days before the deadline! #sohappy	I forgot to do a final project and failed my class 😕	Wait there was an exam?!?!	Someone in this room just lost their tarantula so current status is terrified #omg	I am so mad that I have to retake this class for a fifth time	Ew gross the person next to me is eating raw eggs 🥚

Include Clear Instructions

Be as concise as possible

Remember that your annotators will most likely be much less familiar with the task than you …you may need to state things that seem obvious!

Case Example: Emotion Classification

Natalie wants to collect emotion labels for a classification task. She has a dataset of 1000 tweets, and wants to label each tweet with the most dominant emotion that it expresses.

I'm so happy I get to take CS 521!!!! #bestclassever

oh no I registered late and got stuck in CS 521 😭

Wow I had no idea that UIC had a statistical NLP class 😮

Select the most dominant emotion expressed by the tweet below. Base your decision only on the information provided in the tweet text.
<tweet>

Happiness	Sadness	Surprise	Fear	Anger	Disgust
Indicates happiness or joy.	Indicates sadness or disappointment.	Indicates surprise or shock.	Indicates fear, stress, or panic.	Indicates anger or resentment.	Indicates disgust or unpleasantness.
Yay I finished my homework three days before the deadline! #sohappy	I forgot to do a final project and failed my class 😕	Wait there was an exam?!?!	Someone in this room just lost their tarantula so current status is terrified #omg	I am so mad that I have to retake this class for a fifth time	Ew gross the person next to me is eating raw eggs 🥚

Remember, developing a good annotation schema is often an iterative process!

- **Start by giving your annotators a small sample of your data** and checking:
  - Inter-annotator agreement
  - Agreement with what you'd like to see
- **Ask annotators for preliminary feedback**
  - Was there a category that seemed to be obviously missing?
  - Were there categories that seemed to overlap?
  - Was the annotation format reasonable?

Case Example: Emotion Classification

Using the scheme she'd developed, Natalie collected annotations for 50 tweets. Her inter-annotator agreement was lower than expected, and a lot of tweets seemed to be incorrectly labeled with "happiness." She asked her annotators for some feedback.

Select the most dominant emotion expressed by the tweet below. Base your decision only on the information provided in the tweet text.
<tweet>

Happiness	Sadness	Surprise	Fear	Anger	Disgust
Indicates happiness or joy.	Indicates sadness or disappointment.	Indicates surprise or shock.	Indicates fear, stress, or panic.	Indicates anger or resentment.	Indicates disgust or unpleasantness.
Yay I finished my homework three days before the deadline! #sohappy	I forgot to do a final project and failed my class 🙁	Wait there was an exam?!?!	Someone in this room just lost their tarantula so current status is terrified #omg	I am so mad that I have to retake this class for a fifth time	Ew gross the person next to me is eating raw eggs 🥚

Case Example: Emotion Classification

Using the scheme she'd developed, Natalie collected annotations for 50 tweets. Her inter-annotator agreement was lower than expected, and a lot of tweets seemed to be incorrectly labeled with "happiness." She asked her annotators for some feedback.

Sometimes there didn't really seem to be any emotion expressed by the tweet, so I just chose "happiness."

I wasn't sure what the difference was between "surprise" and "fear."

Select the most dominant emotion expressed by the tweet below. Base your decision only on the information provided in the tweet text.
<tweet>

Happiness	Sadness	Surprise	Fear	Anger	Disgust
Indicates happiness or joy.	Indicates sadness or disappointment.	Indicates surprise or shock.	Indicates fear, stress, or panic.	Indicates anger or resentment.	Indicates disgust or unpleasantness.
Yay I finished my homework three days before the deadline! #sohappy	I forgot to do a final project and failed my class 🙁	Wait there was an exam?!?!	Someone in this room just lost their tarantula so current status is terrified #omg	I am so mad that I have to retake this class for a fifth time	Ew gross the person next to me is eating raw eggs 🥚

Case Example: Emotion Classification

Taking these descriptions into account, Natalie updated her description of “fear.” She also added a new annotation category, “neutral.”

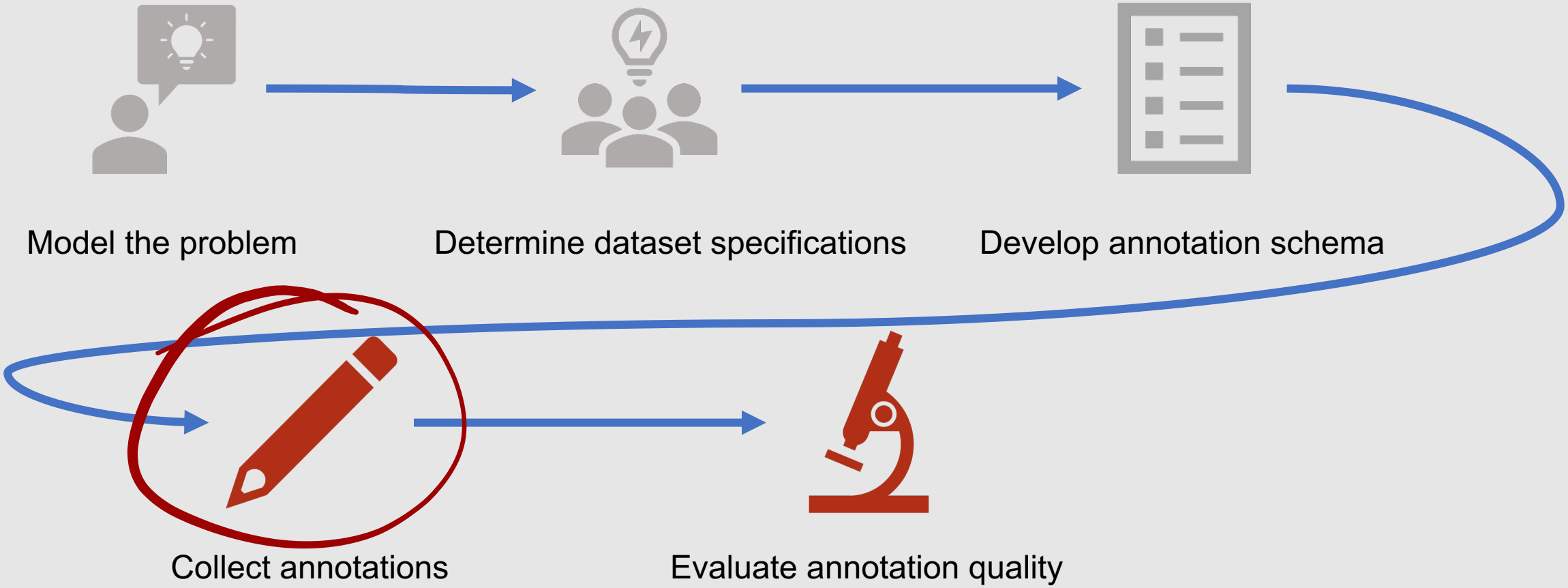
Select the most dominant emotion expressed by the tweet below. Base your decision only on the information provided in the tweet text.
<tweet>

Happiness	Sadness	Surprise	Fear	Anger	Disgust	Neutral
Indicates happiness or joy.	Indicates sadness or disappointment.	Indicates surprise or shock.	Indicates fear, terror, or dread.	Indicates anger or resentment.	Indicates disgust or unpleasantness.	Used no obvious emotion.
Yay I finished my homework three days before the deadline! #sohappy	I forgot to do a final project and failed my class 😕	Wait there was an exam?!?!	Someone in this room just lost their tarantula so current status is terrified #omg	I am so mad that I have to retake this class for a fifth time	Ew gross the person next to me is eating raw eggs 🥚	Slides from yesterday’s class are available now.

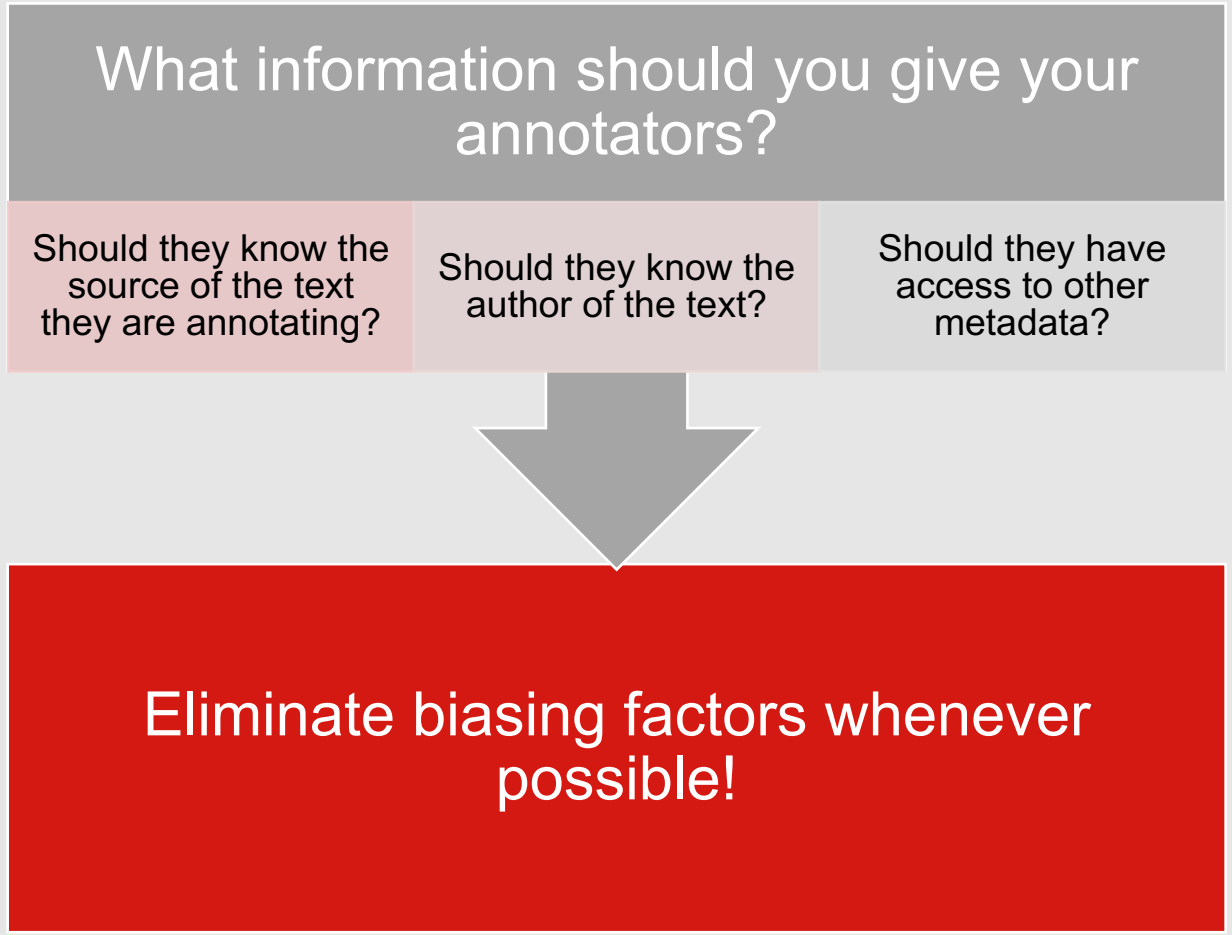
For more tips on creating good annotation schema, check out guidelines for other popular corpora in your area.

- If there are well-defined standards in your annotation area, use those to the extent possible
- If you deviate from a well-defined standard, make sure to justify your reasons for doing so (others interested in using the corpus will probably ask)

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?

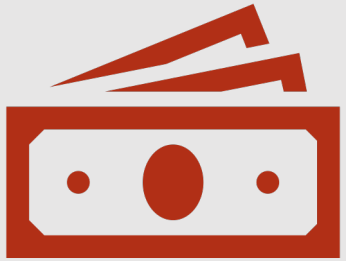
- Background expertise
- Language competency
- Demographic characteristics

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:

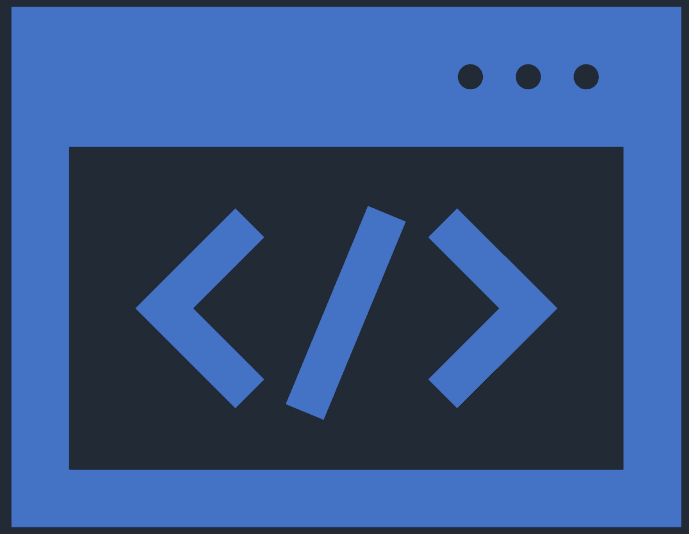
- Amazon Mechanical Turk: https://www.mturk.com/
- Appen: https://appen.com/

Annotation Environments

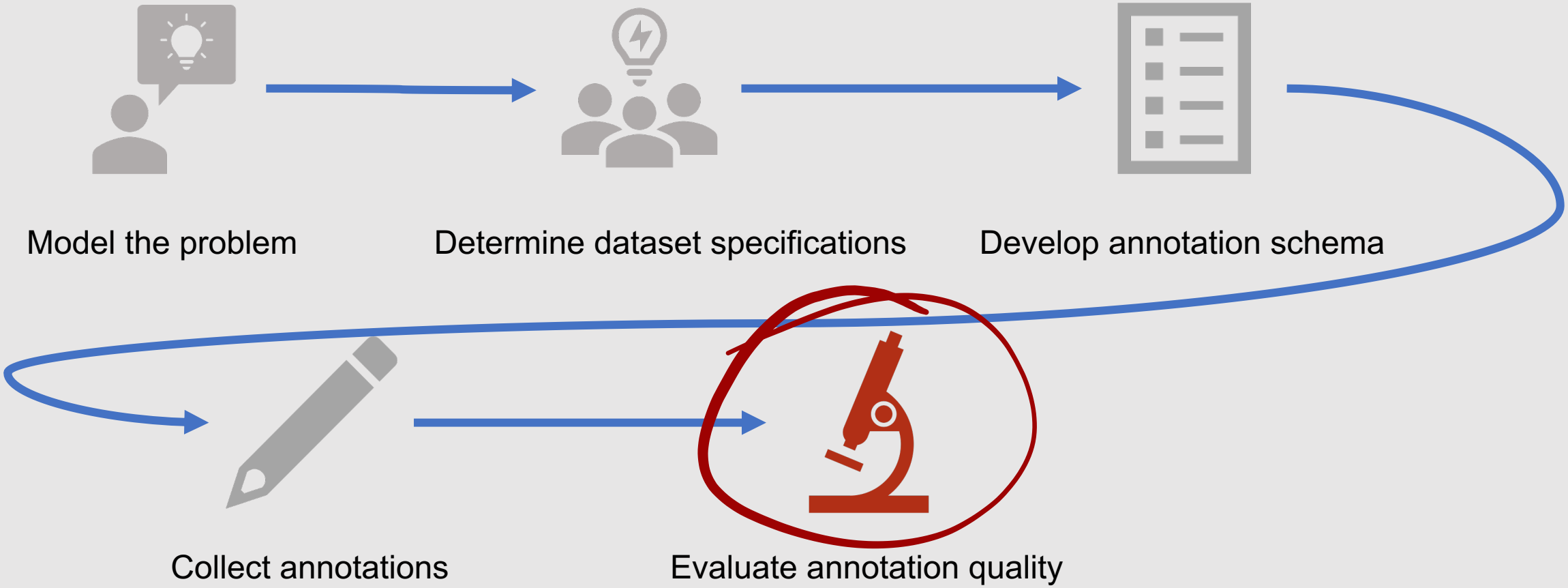
- Many different tools exist!
  - Multipurpose Annotation Environment (MAE)
    - https://github.com/nathan2718/mae-annotation-1
  - General Architecture for Text Engineering (GATE)
    - https://gate.ac.uk/sale/tao/split.html
  - WebAnno
    - https://webanno.github.io/webanno/
  - INCEpTION
    - https://inception-project.github.io/

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



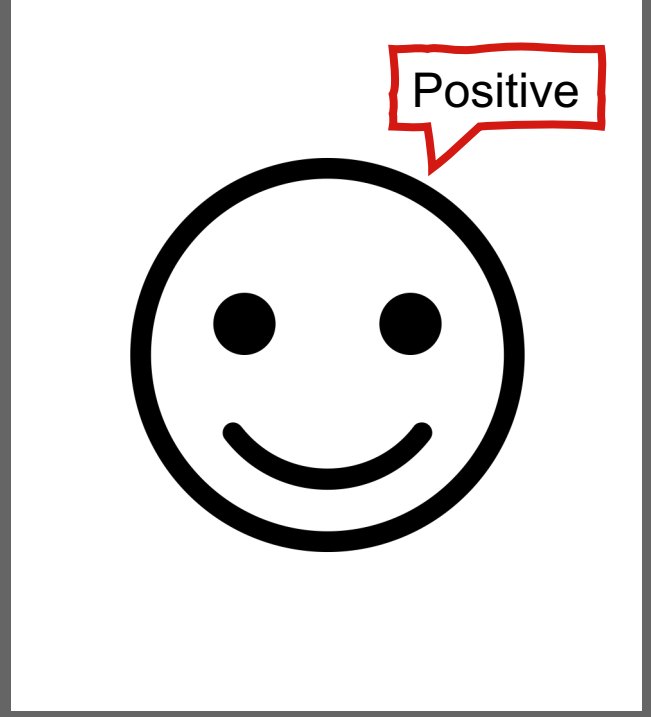
Typical Data Collection Pipeline

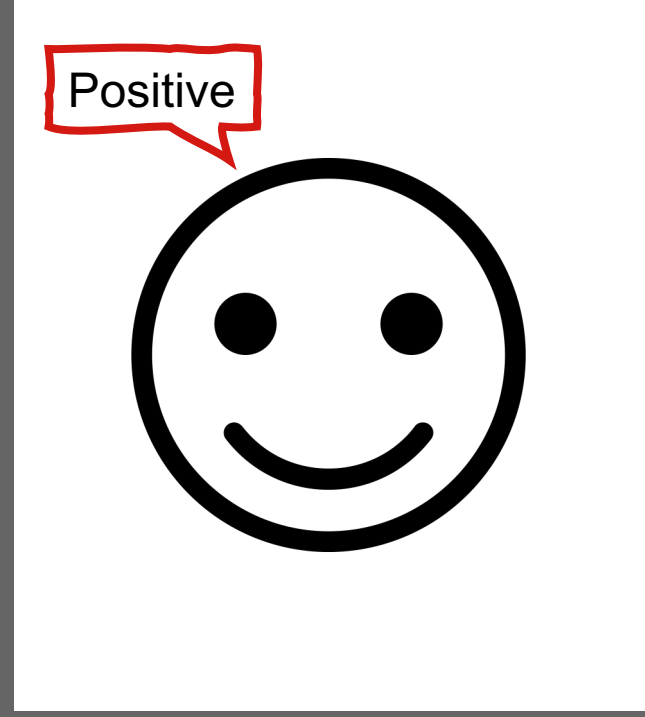


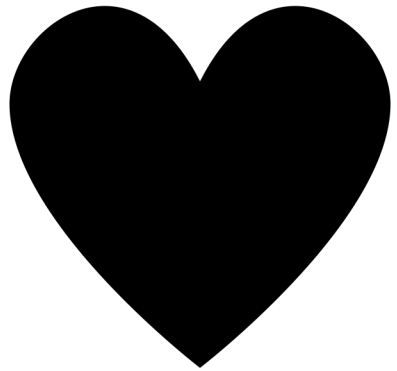
Inter-Annotator Agreement (IAA)

- Collect labels from multiple annotators for the same data instances
- Determine how well the annotators agreed with one another

- Why is this important?
  - Good IAA scores ensure that:
    - Your annotation scheme effectively models your problem
    - Your work is reproducible







How is IAA computed?

Percent agreement?

- Doesn't consider random chance agreement 😕

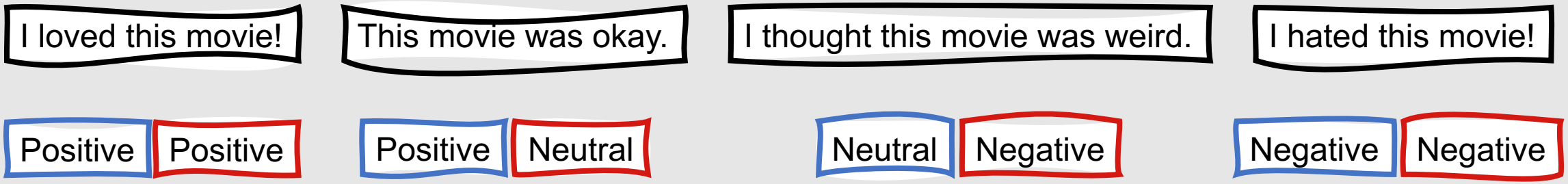
Most common metrics:

- Cohen's Kappa
- Krippendorff's Alpha

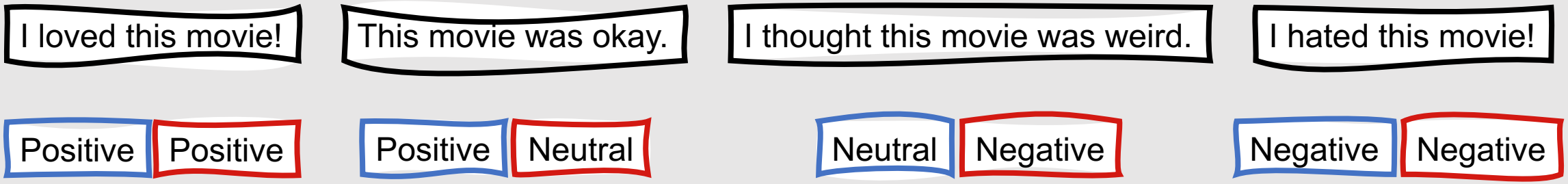
Cohen's Kappa

- Measures the agreement between two annotators, while considering the possibility of chance agreement
  - $\kappa = \frac{p_r - p_e}{1 - p_e}$
    - where p_r is the relative observed agreement between annotators, and p_e is the expected agreement between annotators, if each selected a label randomly

Example: Cohen's Kappa

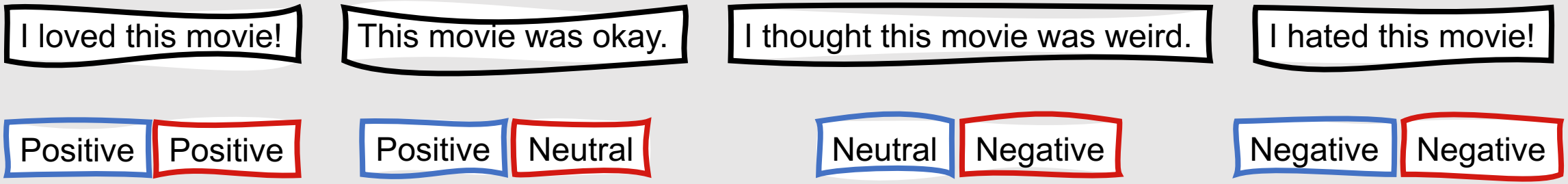


Example: Cohen's Kappa



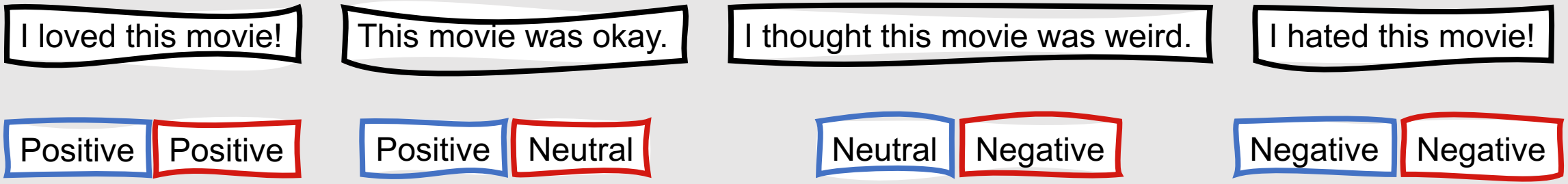
		Annotator B		
		Positive	Neutral	Negative
Annotator A	Positive			
	Neutral			
	Negative			

Example: Cohen's Kappa



		Annotator B		
		Positive	Neutral	Negative
Annotator A	Positive	1	1	0
	Neutral	0	0	1
	Negative	0	0	1

Example: Cohen's Kappa

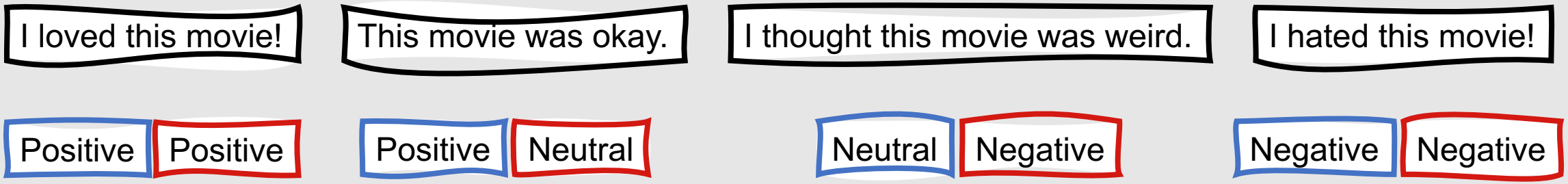


p_r = actual observed agreement

$$p_r = \frac{1+1}{1+1+1+1} = 0.5$$

		Annotator B		
		Positive	Neutral	Negative
Annotator A	Positive	1	1	0
	Neutral	0	0	1
	Negative	0	0	1

Example: Cohen's Kappa



p_r = actual observed agreement

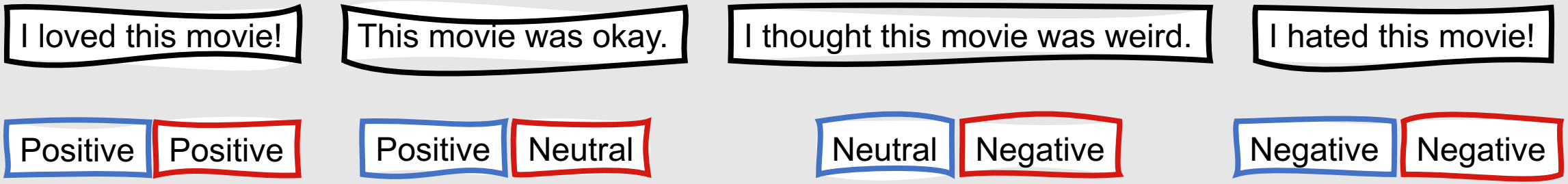
$$p_r = \frac{1+1}{1+1+1+1} = 0.5$$

p_e = expected chance agreement

Annotator A used "positive" 2 times (0.5 of all annotations)
Annotator B used "positive" 1 time (0.25 of all annotations)

		Annotator B		
		Positive	Neutral	Negative
Annotator A	Positive	1	1	0
	Neutral	0	0	1
	Negative	0	0	1

Example: Cohen's Kappa



p_r = actual observed agreement

$$p_r = \frac{1+1}{1+1+1+1} = 0.5$$

p_e = expected chance agreement

Annotator A used "positive" 2 times (0.5 of all annotations)
Annotator B used "positive" 1 time (0.25 of all annotations)
expected chance agreement: 0.5 * 0.25 = 0.125

		Annotator B		
		Positive	Neutral	Negative
Annotator A	Positive	1	1	0
	Neutral	0	0	1
	Negative	0	0	1

Example: Cohen's Kappa

I loved this movie!	This movie was okay.	I thought this movie was weird.	I hated this movie!
Positive Positive	Positive Neutral	Neutral Negative	Negative Negative

p_r = actual observed agreement

$$p_r = \frac{1+1}{1+1+1+1} = 0.5$$

p_e = expected chance agreement

p_e("positive") = 0.125

Annotator A used "neutral" 1 time (0.25 of all annotations)

Annotator B used "neutral" 1 time (0.25 of all annotations)

expected chance agreement: 0.25 * 0.25 = 0.0625

		Annotator B		
		Positive	Neutral	Negative
Annotator A	Positive	1	1	0
	Neutral	0	0	1
	Negative	0	0	1

Example: Cohen's Kappa

I loved this movie!	This movie was okay.	I thought this movie was weird.	I hated this movie!
Positive Positive	Positive Neutral	Neutral Negative	Negative Negative

p_r = actual observed agreement

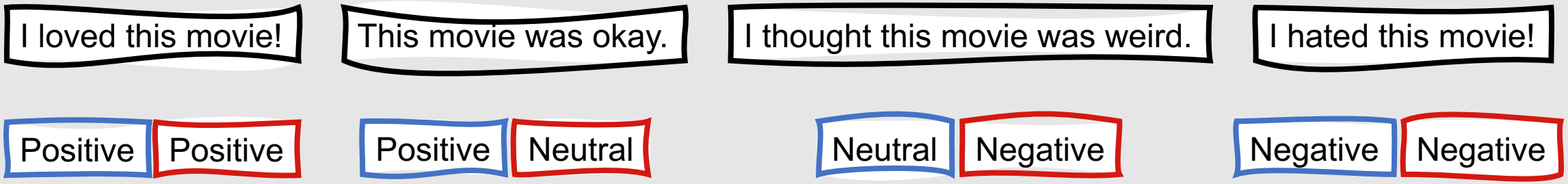
$$p_r = \frac{1+1}{1+1+1+1} = 0.5$$

p_e = expected chance agreement
p_e("positive") = 0.125, p_e("neutral") = 0.0625

Annotator A used "negative" 1 time (0.25 of all annotations)
Annotator B used "negative" 2 times (0.5 of all annotations)
expected chance agreement: 0.25 * 0.5 = 0.125

		Annotator B		
		Positive	Neutral	Negative
Annotator A	Positive	1	1	0
	Neutral	0	0	1
	Negative	0	0	1

Example: Cohen's Kappa



p_r = actual observed agreement

$$p_r = \frac{1+1}{1+1+1+1} = 0.5$$

p_e = expected chance agreement
p_e("positive") = 0.125, p_e("neutral") = 0.0625,
p_e("negative") = 0.125

$$p_e = 0.125 + 0.0625 + 0.125 = 0.3125$$

		Annotator B		
		Positive	Neutral	Negative
Annotator A	Positive	1	1	0
	Neutral	0	0	1
	Negative	0	0	1

Example: Cohen's Kappa

I loved this movie!	This movie was okay.	I thought this movie was weird.	I hated this movie!
Positive Positive	Positive Neutral	Neutral Negative	Negative Negative

p_r = actual observed agreement

$$p_r = \frac{1+1}{1+1+1+1} = 0.5$$

p_e = expected chance agreement

$$p_e = 0.125 + 0.0625 + 0.125 = 0.3125$$

$$\kappa = \frac{p_r - p_e}{1 - p_e} = \frac{0.5 - 0.3125}{1 - 0.3125} = \frac{0.1875}{0.6875} = 0.27$$

		Annotator B		
		Positive	Neutral	Negative
Annotator A	Positive	1	1	0
	Neutral	0	0	1
	Negative	0	0	1

What if each instance was annotated by more than two annotators?

- Fleiss's Kappa
  - $\kappa = \frac{p - \overline{p_e}}{1 - \overline{p_e}}$
    - where $\bar{p}$ is the average of the percentage of annotators who agree, and $\overline{p_e}$ is the average of the percentages of annotators expected to agree by chance
- Krippendorff's Alpha
  - $\alpha = \frac{p_\alpha - p_e}{1 - p_e}$
    - where p_α is a weighted percent agreement, and p_e is a weighted percent chance agreement
  - Computationally expensive behind the scenes!

Interpreting Kappa Values

- What is a “good” kappa value?
  - Depends on the task complexity and objectivity
- In general, most researchers adhere to the following (Landis and Koch, 1977):
  - $\kappa \leq 0$: Poor agreement
  - $0.00 < \kappa < 0.20$: Slight agreement
  - $0.20 \leq \kappa < 0.40$: Fair agreement
  - $0.40 \leq \kappa < 0.60$: Moderate agreement
  - $0.60 \leq \kappa < 0.80$: Substantial agreement
  - $0.80 \leq \kappa$: Perfect (or at least, extremely good) agreement

Creating a Gold Standard

Once you're satisfied with your IAA scores, how do you select final labels for data that has been annotated by multiple people?

If in agreement, use that label

If in disagreement, adjudicate!

Select an adjudicator who is already very familiar with the task (usually someone who was involved in creating the annotation guidelines)

Adjudication Guidelines

- Allocate plenty of time for adjudication
- Don’t feel pressured to go with the majority, in cases with more than two annotators
  - Annotators may have agreed due to random chance
- If using multiple adjudicators, compute IAA between them to make sure they’re on the right track

After your data has been adjudicated, your corpus is complete!

Make sure to document the process well

If publishing the corpus, make sure the data and annotations are in a clean, organized format that is easy to use by other researchers

Summary: Data Collection

- Data collection is the process of **curating data and assigning labels** to it
- Data can be collected from existing text, recordings, or other online samples, or it can be collected directly from people completing a specific task
- Annotations can be collected from in-person annotators, or **crowdsourced** online
- **Inter-annotator agreement (IAA)** indicates how well individual annotators agreed with one another when labeling a dataset
- A common metric for computing IAA is **Cohen's Kappa**
- Once all data has been collected, an **adjudicator** determines the final labels for instances about which individual annotators disagreed
- The final set of adjudicated labels is referred to as the **gold standard**