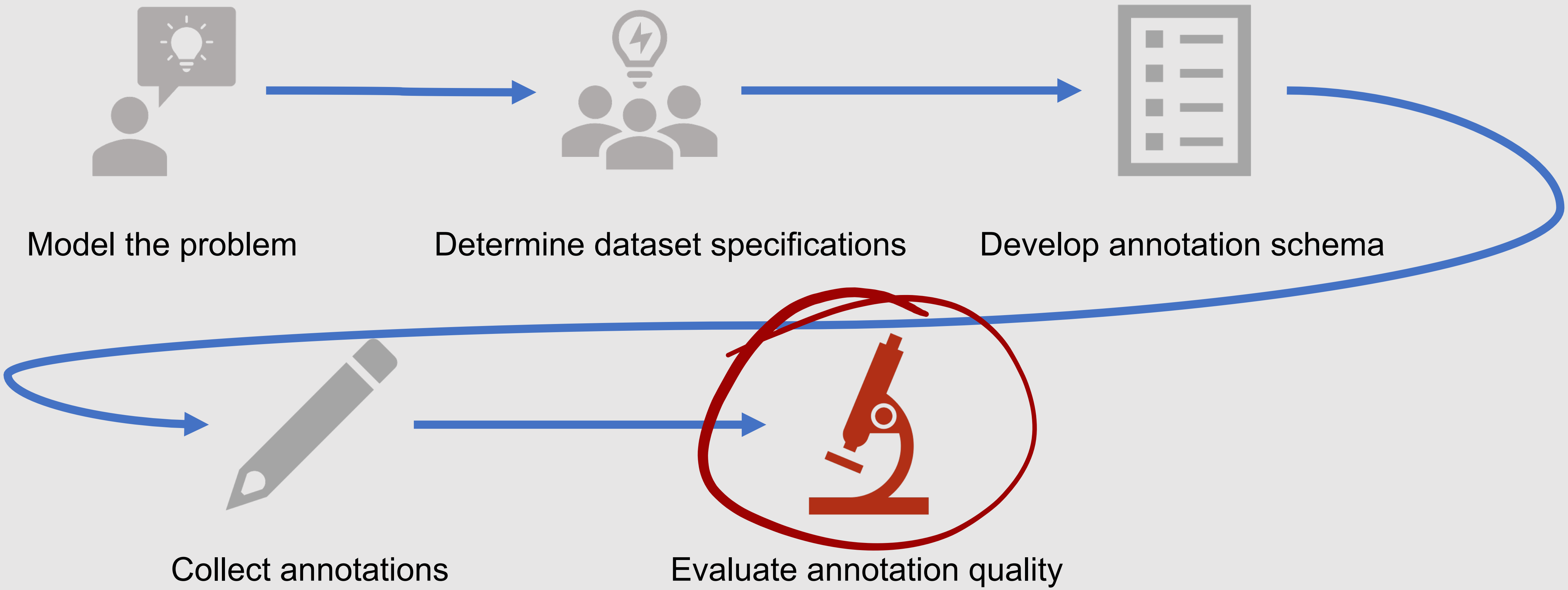
Evaluating Annotation Quality

Natalie Parde
UIC CS 521

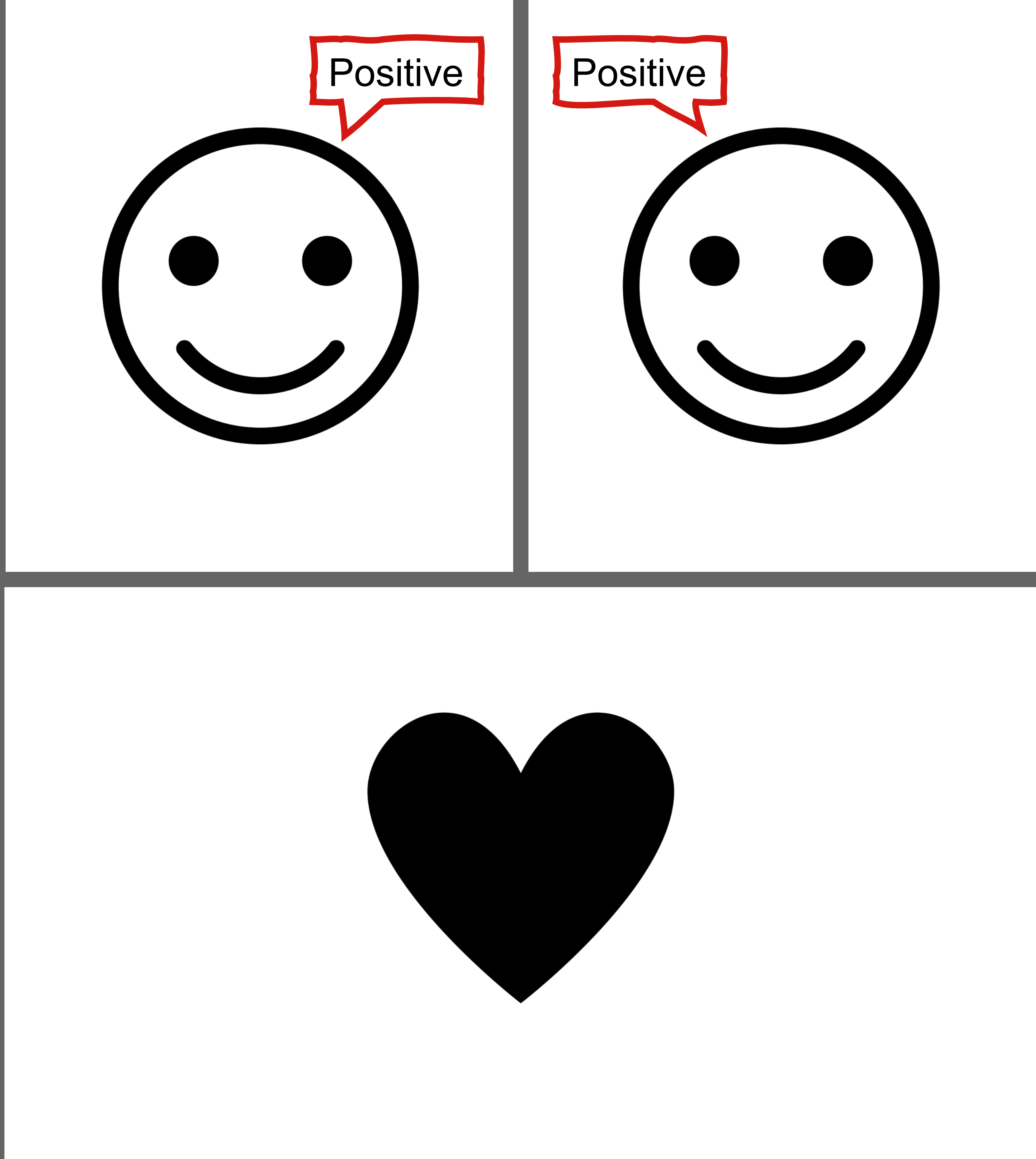
Typical Data Collection Pipeline



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

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

Example: Cohen's Kappa

I loved this movie! | Positive | Positive

This movie was okay. | Positive | Neutral

I thought this movie was weird. | Neutral | Negative

I hated this movie! | Negative | Negative

		Annotator B		
		Positive	Neutral	Negative
Annotator A	Positive			
	Neutral			
	Negative			

Example: Cohen's Kappa

I loved this movie!

Positive Positive

This movie was okay.

Positive Neutral

I thought this movie was weird.

Neutral Negative

I hated this movie!

Negative Negative

		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

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

		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

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

Sentence	Annotator A	Annotator B
I loved this movie!	Positive	Positive
This movie was okay.	Positive	Neutral
I thought this movie was weird.	Neutral	Negative
I hated this movie!	Negative	Negative

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

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,
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! — Positive | Positive

This movie was okay. — Positive | Neutral

I thought this movie was weird. — Neutral | Negative

I hated this movie! — 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
  - $\kappa = \frac{\bar{p} - \overline{p_e}}{1 - \overline{p_e}}$
    - 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
  - $\alpha = 1 - \frac{D_o}{D_e}$
    - where D_o is the observed disagreement, and D_e is the expected chance disagreement
  - 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 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