

# Inter-Annotator Agreement

LIN 313 Language and Computers  
UT Austin Fall 2025  
Instructor: Gabriella Chronis

# Admin

- HW 2 Due today
- HW 3 posted Wednesday (Due 10/15)
- read + annotate for Friday: Excavating AI

# Overview

- Review potential annotation schemes
- Lecture: Building a dataset
  - annotation guidelines
  - What is inter-annotator agreement
- Lab: test annotations & calculating agreement

# Data Quality

High Quality Data is

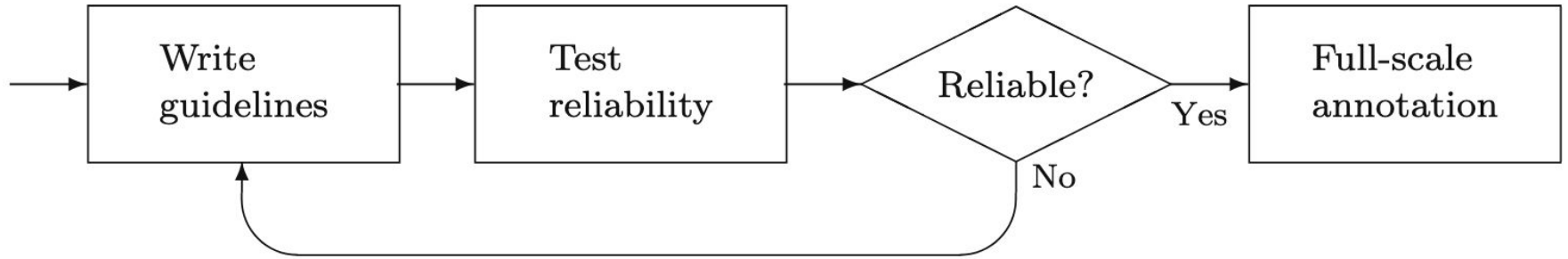
1. accuracy
2. complete
3. consistent
4. relevant to the problem

# Garbage in Garbage out

What are the consequences of low-quality data?

If inaccurate, incomplete, or poor-quality data is provided to a system, it will inevitably produce flawed, unreliable, or incorrect results, regardless of how sophisticated the system itself is

# Dataset Creation



**Fig. 1** Iterative reliability testing

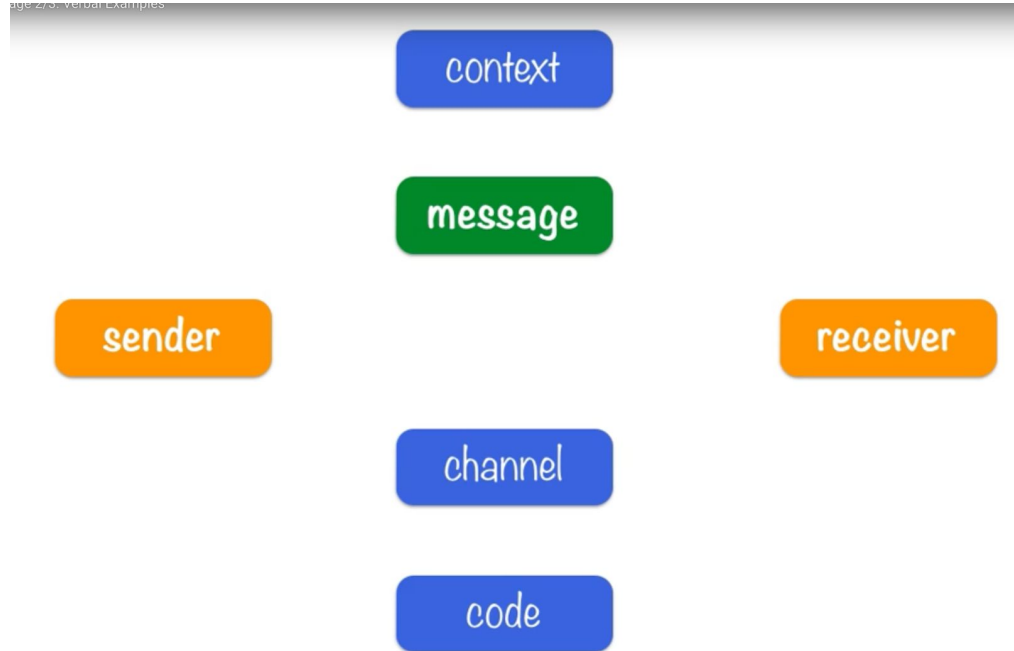
# Classification Guidelines

instructions to annotator

strike a balance between prescriptive and open-ended

- we want to avoid ambiguity
- but we want to elicit intuitive judgments of the speaker
  - "what X means to me" vs "what the researchers mean by X"

# Remember the elements of the speech situation:



- is the sentence creepy vs wet?
- is the speaker creepy vs wet?
- is the topic creepy vs wet?



# Annotation Guidelines: an example

**not prescriptive enough:** "creepy vs wet"

**less ambiguous:** "The thing being talked about is creepy vs wet"

**maybe too prescriptive:** is the person being spoken to experiencing the sensation of creepy vs wet where creepy is "an unpleasant feeling of fear or unease" and "wet" is "covered or saturated with water or another liquid"

# Inter-annotator agreement

There are many ways to measure how close two raters are. We will use Cohen's Kappa.

- two raters
- nominal judgments (versus ordinal, aka on a scale)
- have judgments for both raters for all items

A \ B	Yes	No
Yes	a	b
No	c	d

e.g.

A \ B	Yes	No
Yes	20	5
No	10	15

A \ B	Yes	No
Yes	a	b
No	c	d

The observed proportionate agreement is:

$$p_o = \frac{a + d}{a + b + c + d} = \frac{20 + 15}{50} = 0.7$$

e.g.

A \ B	Yes	No
Yes	20	5
No	10	15

A \ B	Yes	No
Yes	a	b
No	c	d

The observed proportionate agreement is:

$$p_o = \frac{a + d}{a + b + c + d} = \frac{20 + 15}{50} = 0.7$$

So the expected probability that both would say yes at random is:

$$p_{\text{Yes}} = \frac{a + b}{a + b + c + d} \cdot \frac{a + c}{a + b + c + d} = 0.5 \times 0.6 = 0.3$$

Similarly:

$$p_{\text{No}} = \frac{c + d}{a + b + c + d} \cdot \frac{b + d}{a + b + c + d} = 0.5 \times 0.4 = 0.2$$

Overall random agreement probability is the probability that they agreed on either Yes or No, i.e.:

$$p_e = p_{\text{Yes}} + p_{\text{No}} = 0.3 + 0.2 = 0.5$$

So now applying our formula for Cohen's Kappa we get:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = \frac{0.7 - 0.5}{1 - 0.5} = 0.4$$

e.g.

A \ B	Yes	No
Yes	20	5
No	10	15

# Cohen's Kappa

$P_0$  = relative observed agreement

$P_e$  = probability of random agreement

$$\kappa = \frac{p_0 - p_e}{1 - p_e},$$

# Why is it designed this way?

$$\kappa = \frac{p_0 - p_e}{1 - p_e},$$

- what is the value if you and I agree on everything and we both split our answers half and half?
- What is the value if you and I agree on everything and we both choose 70% a and 30% b?
- what is the value if you and i agree on 50% of items and we both split our answers half and half?

# What does it mean?

Cohen's Kappa	Interpretation
0	No agreement
0.10 - 0.20	Slight agreement
0.21 - 0.40	Fair agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Substantial agreement
0.81 - 0.99	Near perfect agreement
1	Perfect agreement

Calculator: