Quiz instructions

Get a blue or black pen out now, before we hand out the quiz.

Do not talk or use electronic devices once we start handing out the quiz.

Leave the quiz facing down until we say it is time to start.

There are two versions of the quiz. The people either side of you must have a different coloured quiz to you.

When we say it is time to stop, hand your quiz to the end of your row.

Make sure you write your name and SID.

This is a closed-book, closed-note quiz. No electronic devices may be used in any way.

Note your responses by completely filling in the relevant circle(s) an square(s): ●

If you make a mistake, put an X over the filled in circle / square: ⊗

COMP 4446 / 5046
Lecture 10: Data – Annotation and Crowdsourcing

Jonathan K. Kummerfeld

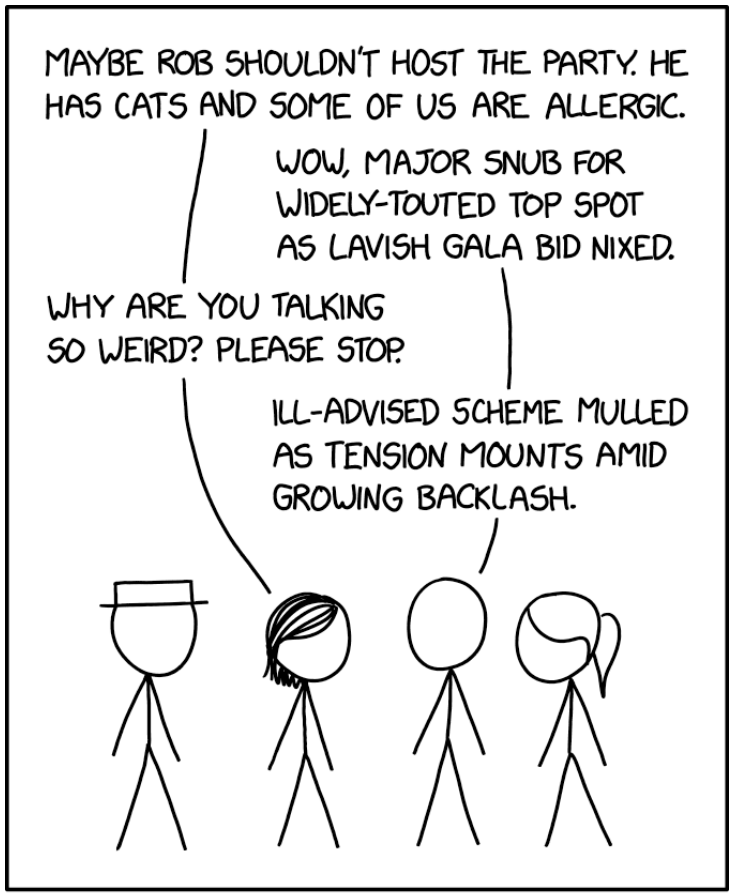
Semester 1, 2025

THE UNIVERSITY OF SYDNEY

menti.com 1552 7067

[Roundly-condemned headlinese initiative shuttered indefinitely.]

Headline Words



Source: https://xkcd.com/2584/

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Assignment 4

Due in 8 days

Teams with the top 3 results will be invited to describe their approach

Apologies for delay on reference result release

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Data Sources

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Where does the data to train these models come from?

Explore with an example:

https://huggingface.co/datasets/HuggingFaceFW/fineweb

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Step 1: Download data

mainly English data

250 billion pages
17 years

language	%
eng	46.4328
deu	5.8355
rus	5.5709
jpn	4.7475
fra	4.6351
spa	4.6199
zho	3.7971
ita	2.7243
<unknown>	2.7090
nld	2.1566
pol	1.7413
por	1.1169
ces	1.0611
vie	1.0570

https://huggingface.co/datasets/HuggingFaceFW/fineweb

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Step 2: Filter the data

exclude some website

Other work does this with a model-based detector, but that has risks

1. Url Filtering, remov ous and NSFW websites, using both block-list as well as subwords detect
2. Trafilatura text extra nCrawl's warc files
3. FastText LanguageFilter, removing any document with en language score lower than **0.65**
4. Quality filtering
    1. Gopher Repetition / Quality
    2. C4 Quality filters except `terminal_punct` rule
    3. FineWeb custom filters, consisting of heuristics for removing list-like documents, documents with repeated lines and documents with likely wrong line formatting.
5. MinHash deduplication with each crawl deduplicated individually (5-grams, 14x8 hash functions)
6. PII Formatting to anonymize email and public IP addresses

avoid duplicate data

use model to detect which web is good

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Innovation in FineWeb - choosing filters by training

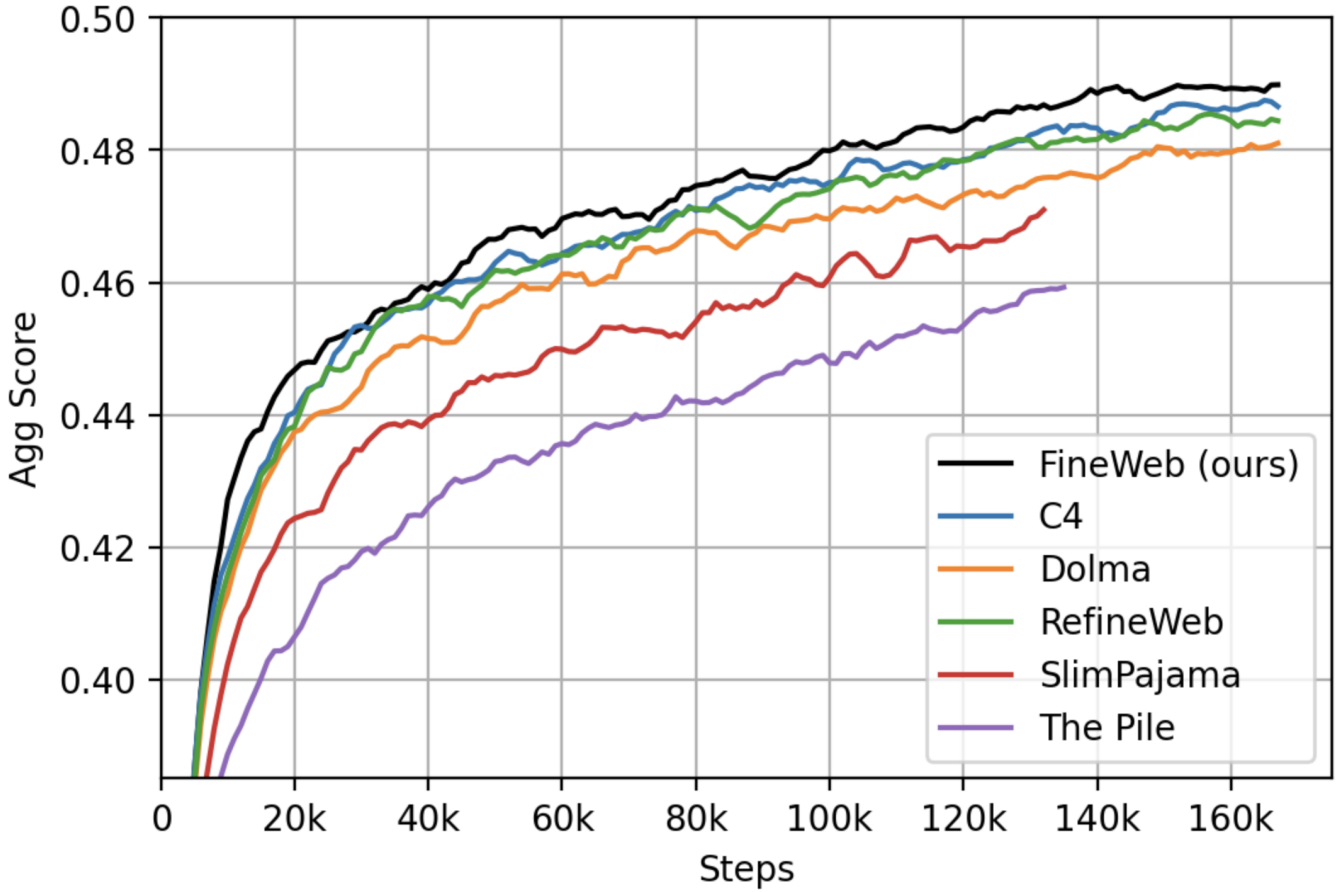
Vary the filters → use different filters to find which one is best
Train models (1.8b parameters each!)
Measure performance

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Innovation in FineWeb - choosing filters by training



COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

LLM developers seek more + better data

BERT - Books scraped from the internet (16 Gb)

RoBERTa - Add more data! (160 Gb)

2 prop of data

'High quality'
- Wikipedia
- News
- Scientific papers

Vs.

Scale:
- The web

"Textbooks Are All You Need", 2023

↳ small data with very good data could also work but still not better than many data

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

What about other tasks?

Named Entity Recognition

Coreference Resolution

Syntactic Parsing

Sentiment Analysis

Translation

….

Annotated data isn't available for free to download!
…mostly

Some translated data is available:

Canadian Hansards
European Parliament
United Nations

from

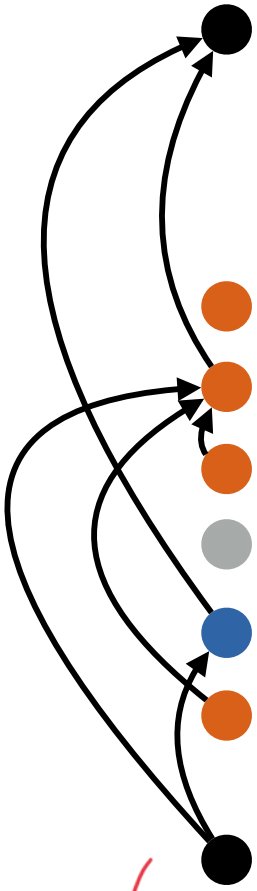
COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

internet related chat

General approach, through an example:
IRC Dialogue Disentanglement Dataset



```
[03:06] <BurgerMann> does anyone know
a consoleprog that scales jpegs fast and
efficient?.. this digital camera age kills
me when I have to scale photos :s
[03:06] <Seveas> delire, yes
[03:06] <Seveas> BurgerMann, convert
[03:06] <Seveas> part of imagemagick
=== Ian [...@...] has left \#ubuntu []
[03:06] <delire> BurgerMann: ImageMagick
[03:06] <Seveas> BurgerMann, i used that
to convert 100's of photos in one command
[03:06] <BurgerMann> Oh... I'll have a
look.. thx =)
```

Chat at same time

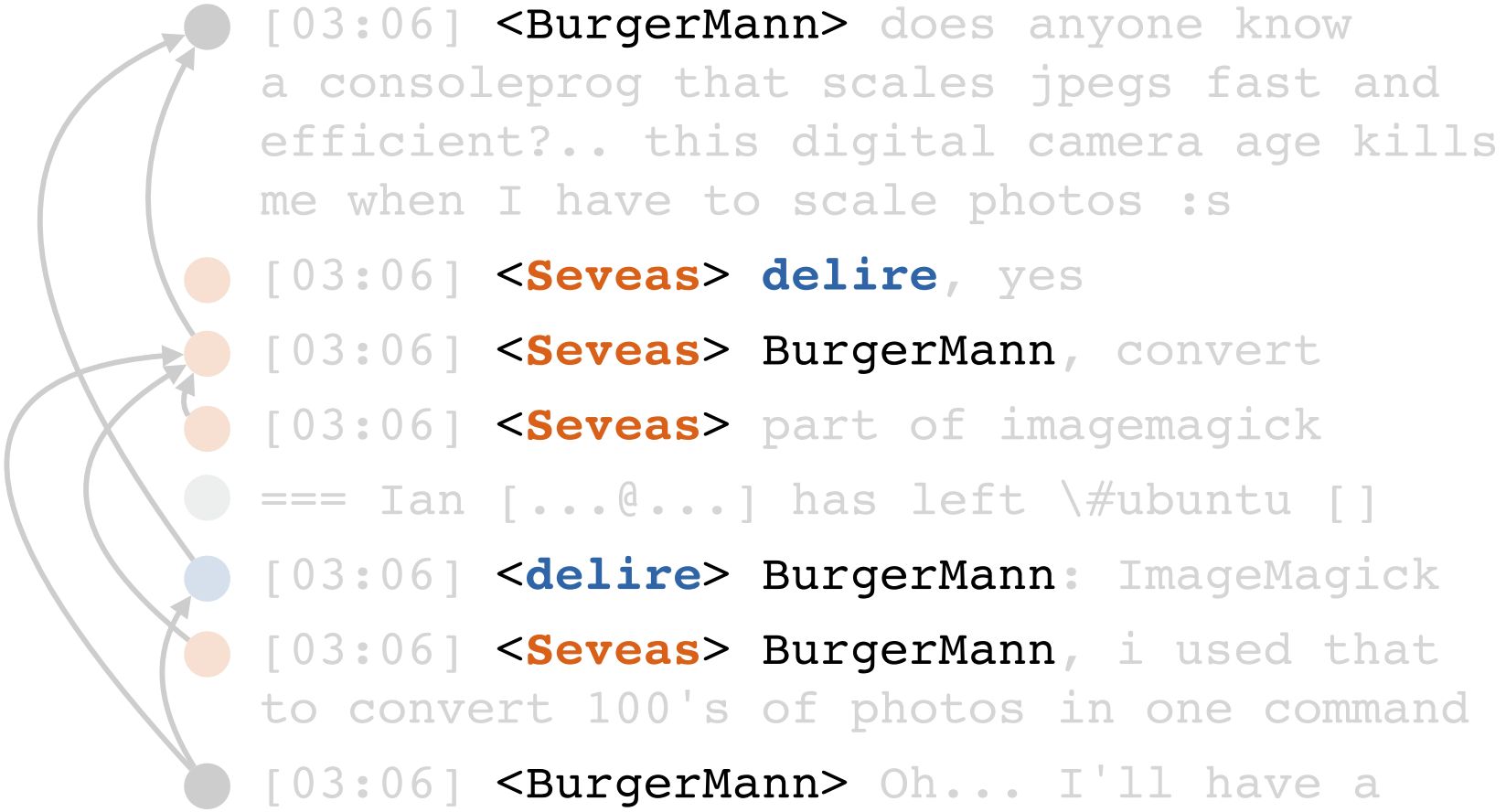
show which msg related to which

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

General approach, through an example: IRC Dialogue Disentanglement Dataset

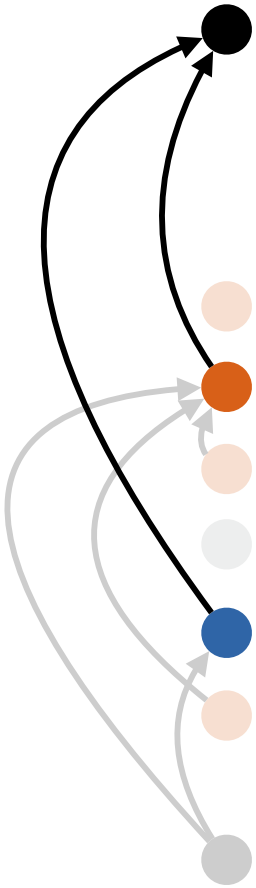


COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

General approach, through an example: IRC Dialogue Disentanglement Dataset



```
[03:06] <BurgerMann> does anyone know a consoleprog that scales jpegs fast and efficient?.. this digital camera age kills me when I have to scale photos :s
[03:06] <Seveas> delire, yes
[03:06] <Seveas> BurgerMann, convert
[03:06] <Seveas> part of imagemagick
=== Ian [...@...] has left \#ubuntu []
[03:06] <delire> BurgerMann: ImageMagick
[03:06] <Seveas> BurgerMann, i used that to convert 100's of photos in one command
[03:06] <BurgerMann> Oh... I'll have a look.. thx =)
```

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

General approach, through an example: IRC Dialogue Disentanglement Dataset

Raw data:

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

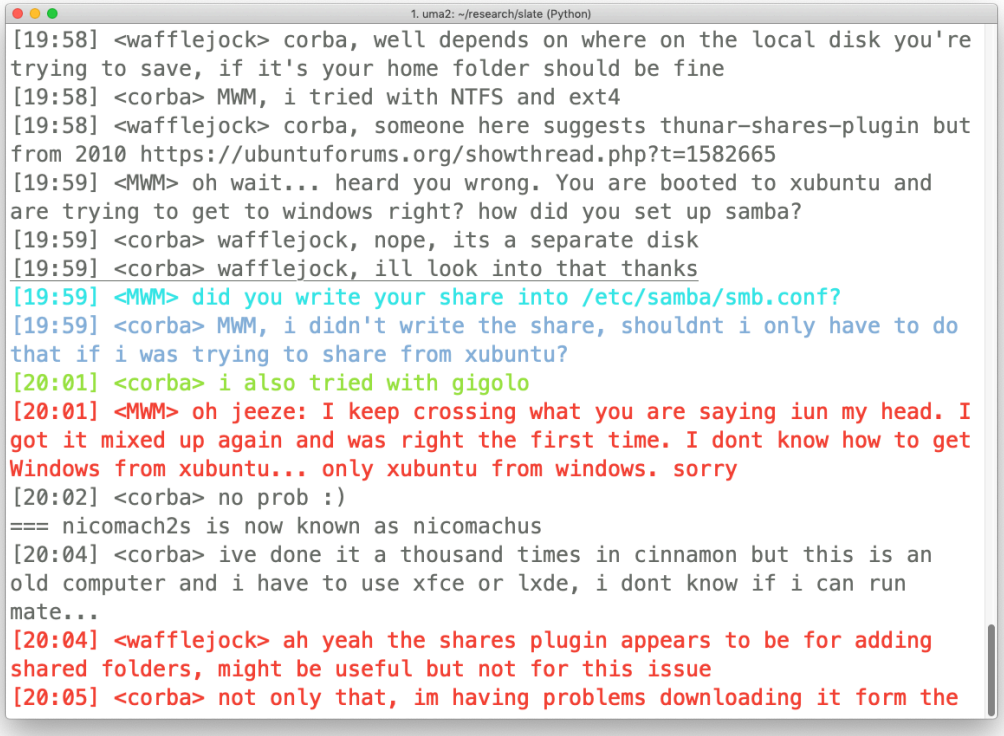
menti.com 1552 7067

General approach, through an example: IRC Dialogue Disentanglement Dataset

Annotation

~250 hours of work

(We'll discuss this more after the break)



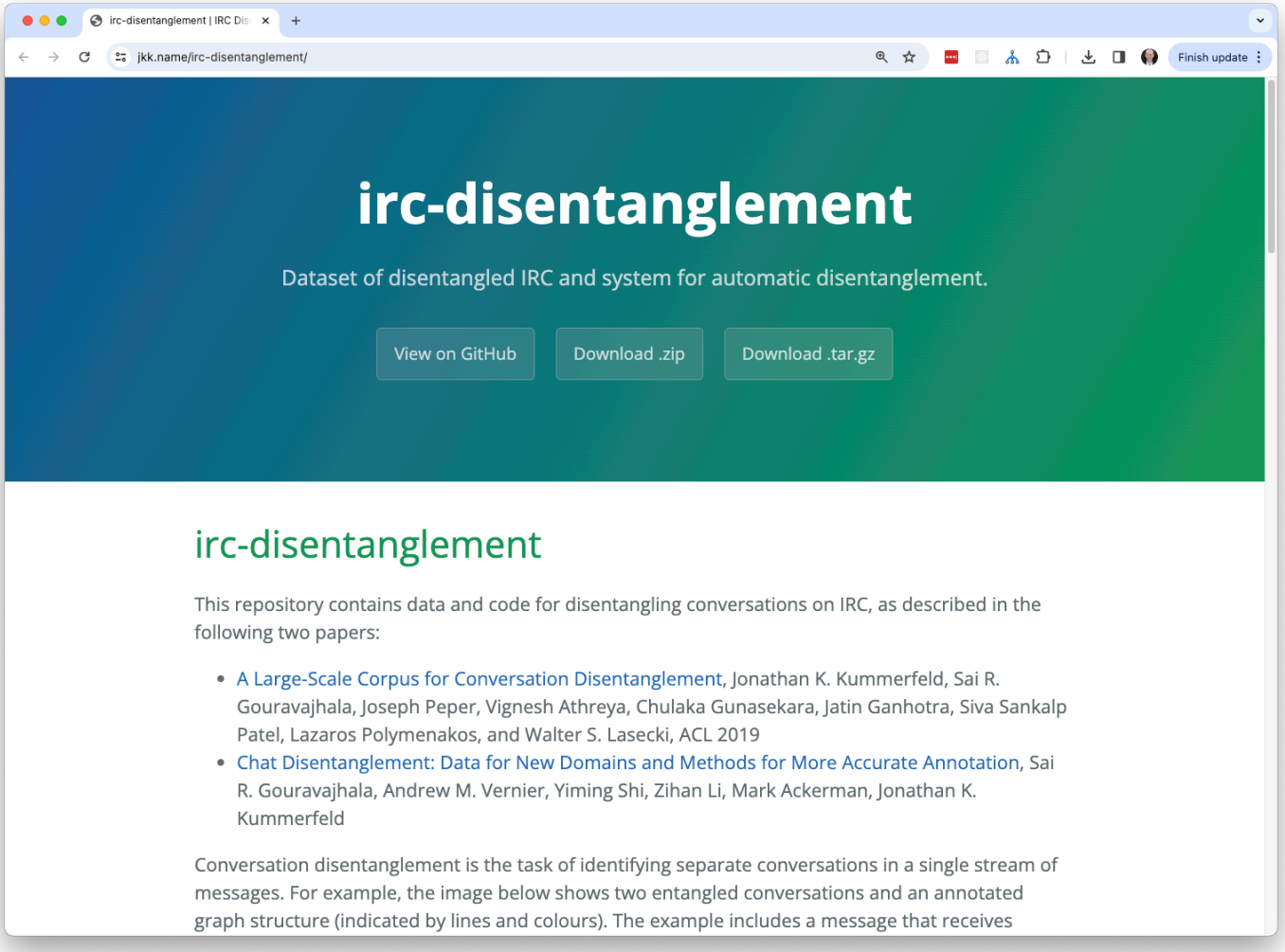
analyze which one respond to which one

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

General approach, through an example: IRC Dialogue Disentanglement Dataset

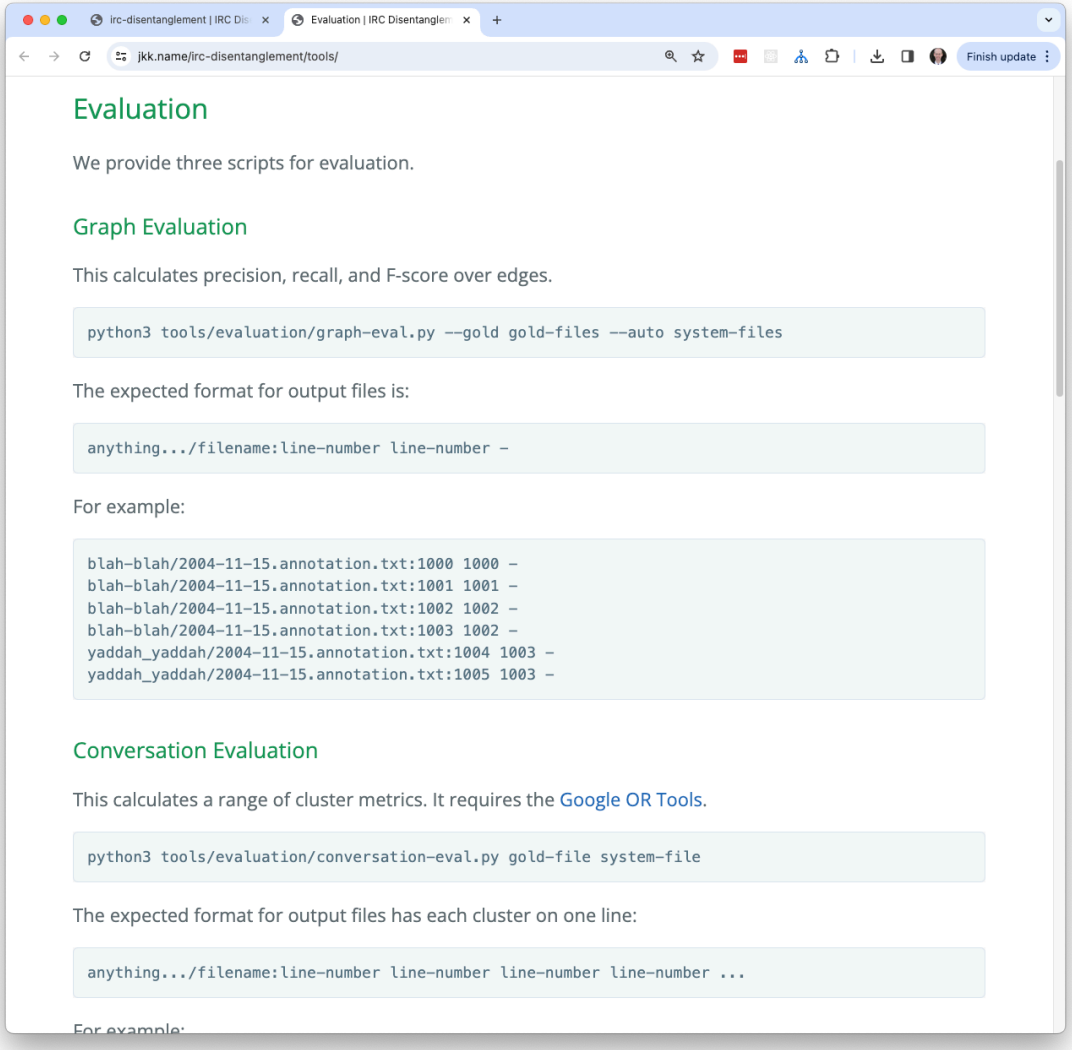


COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

General approach, through an example: IRC Dialogue Disentanglement Dataset



COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Datasets can have massive impact

Penn Treebank - syntactic annotation, 11,000 citations

OntoNotes - coreference, 2,000 citations (across two papers)

SQuAD - question answering, 8,000 citations

Stanford Sentiment Treebank - 9,000 citations

For context, all of my research over the last 17 years has 3,100 citations

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

My most cited paper is a dataset

TITLE	CITED BY	YEAR
An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction S Larson, A Mahendran, JJ Peper, C Clarke, A Lee, P Hill, JK Kummerfeld, ... EMNLP (Short Paper)	612	2019
Improving Text-to-SQL Evaluation Methodology C Finegan-Dollak, JK Kummerfeld, L Zhang, K Ramanathan, ... ACL	328	2018
Spatiotemporal Hierarchy of Relaxation Events, Dynamical Heterogeneities, and Structural Reorganization in a Supercooled Liquid R Candelier, A Widmer-Cooper, JK Kummerfeld, O Dauchot, G Biroli, ... Physical review letters 105 (13), 135702	202	2010
Factors Influencing the Surprising Instability of Word Embeddings L Wendlandt, JK Kummerfeld, R Mihalcea NAACL	166	2018
A Large-Scale Corpus for Conversation Disentanglement JK Kummerfeld, SR Gouravajhala, J Peper, V Athreya, C Gunasekara, ... ACL	135 *	2019
Parser Showdown at the Wall Street Corral: An Empirical Investigation of Error Types in Parser Output JK Kummerfeld, D Hall, JR Curran, D Klein EMNLP	120	2012
Tools for Automated Analysis of Cybercriminal Markets RS Portnoff, S Afroz, G Durrett, JK Kummerfeld, T Berg-Kirkpatrick, ... WWW	114	2017
Overview of the seventh dialog system technology challenge: Dstc7 LF D'Haro, K Yoshino, C Hori, TK Marks, L Polymenakos, JK Kummerfeld, ... Computer Speech & Language 62, 101068	102 *	2020

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Evaluate on multiple tasks at once - avoid overfitting to one

get multiple dataset together
give a better DB?

GLUE - General Language Understanding Evaluation

The Corpus of Linguistic Acceptability
The Stanford Sentiment Treebank
Microsoft Research Paraphrase Corpus
Semantic Textual Similarity Benchmark
Quora Question Pairs
MultiNLI
Question NLI
Recognizing Textual Entailment
Winograd NLI

One gripe - people cite a benchmark suite, but not the datasets that went into it

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Other risks - Shortcuts / Validity

Is there a shortcut - a way to get the right answer without doing the task?

"Clever Hans"

"If the eighth day of the month comes on a Tuesday, what is the date of the following Friday?"
Hans would answer by tapping his hoof eleven times

But only if it could see the questioner and the questioner knew the answer

Wikipedia

↓use human reaction to answer question, similar to a short cut

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Other risks - Shortcuts / Validity

What does Macduff do to Macbeth?

What becomes of Macbeth?

What violent act does Macduff perform upon Macbeth?

The same answer in every case, but some are easier than others

From the UT Austin NLP class

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Other risks - Shortcuts / Validity

"On October 19, 1512, Luther was awarded his doctorate of theology and, on October 21, 1512, was received into the senate of the theological faculty of the University of Wittenberg. He spent the rest of his career in this position at the University of Wittenberg."

shortcut from input

Where …
Who …
When …

Only one option in many cases!

From the UT Austin NLP class

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Other risks - Shortcuts / Validity

One hint is whether a model can do well without the question and/or context?

For example, for the Stanford Natural Language Inference task, looking just at the hypothesis:

	Hyp-only model	Majority class
SNLI	69.17	33.82
MNLI-1	55.52	35.45
MNLI-2	55.18	35.22

From the UT Austin NLP class

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Other risks - Shortcuts / Validity

What to do?

- Get text from real world use
- Make hard cases
  - Wrong answers that have many matching words with the input or true answer
  - Make minimal edits to the true answer that make it wrong

we can test by providing harder testcases

From the UT Austin NLP class

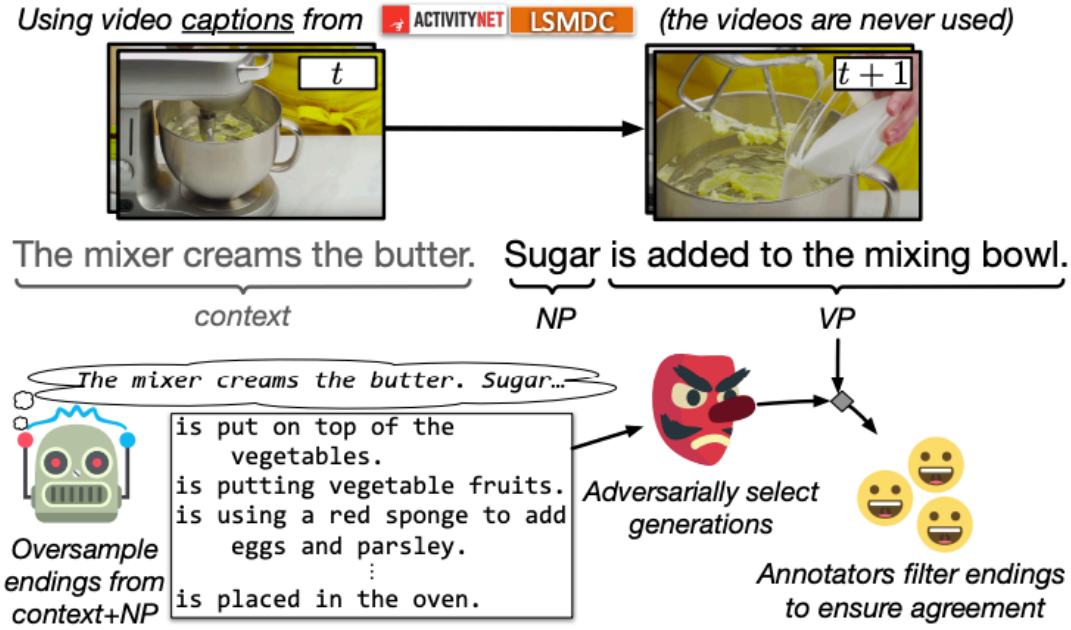
COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Other risks - Does our task have statistical power?

A girl is going across a set of monkey bars. She
a) jumps up across the monkey bars.
b) struggles onto the monkey bars to grab her head.
c) gets to the end and stands on a wooden plank.
d) jumps up and does a back flip.



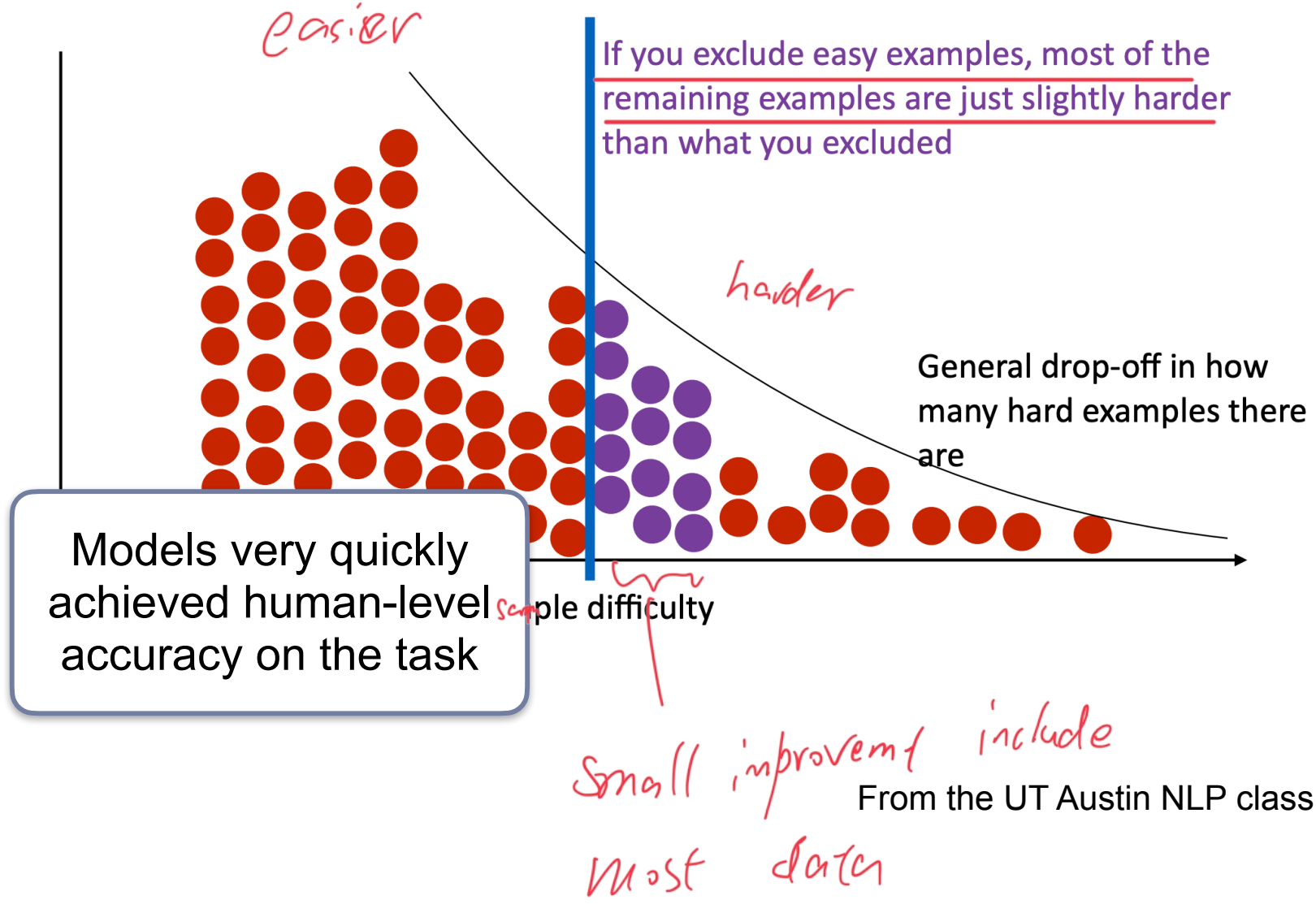
From the UT Austin NLP class

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Other risks - Does our task have statistical power?



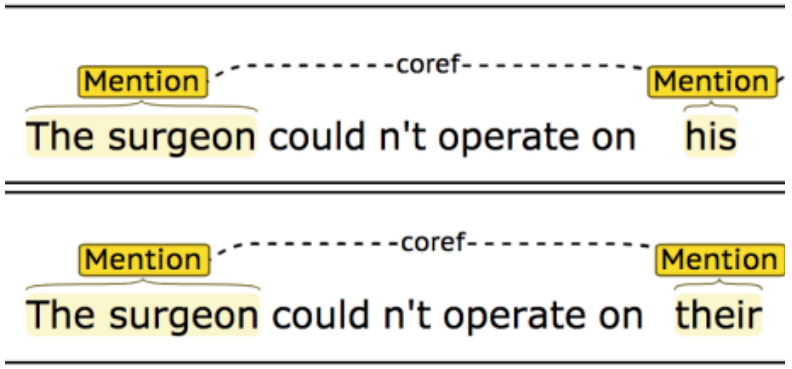
From the UT Austin NLP class

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Other risks - Social bias

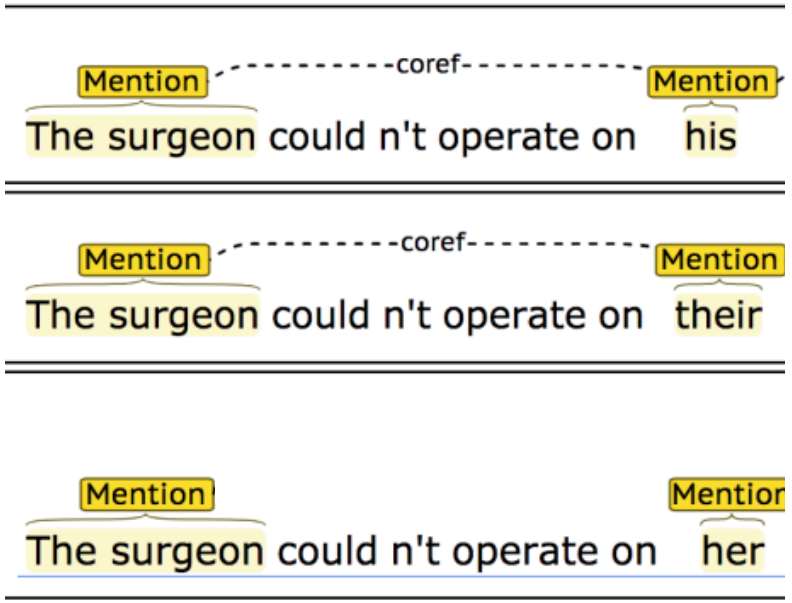


COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Other risks - Social bias

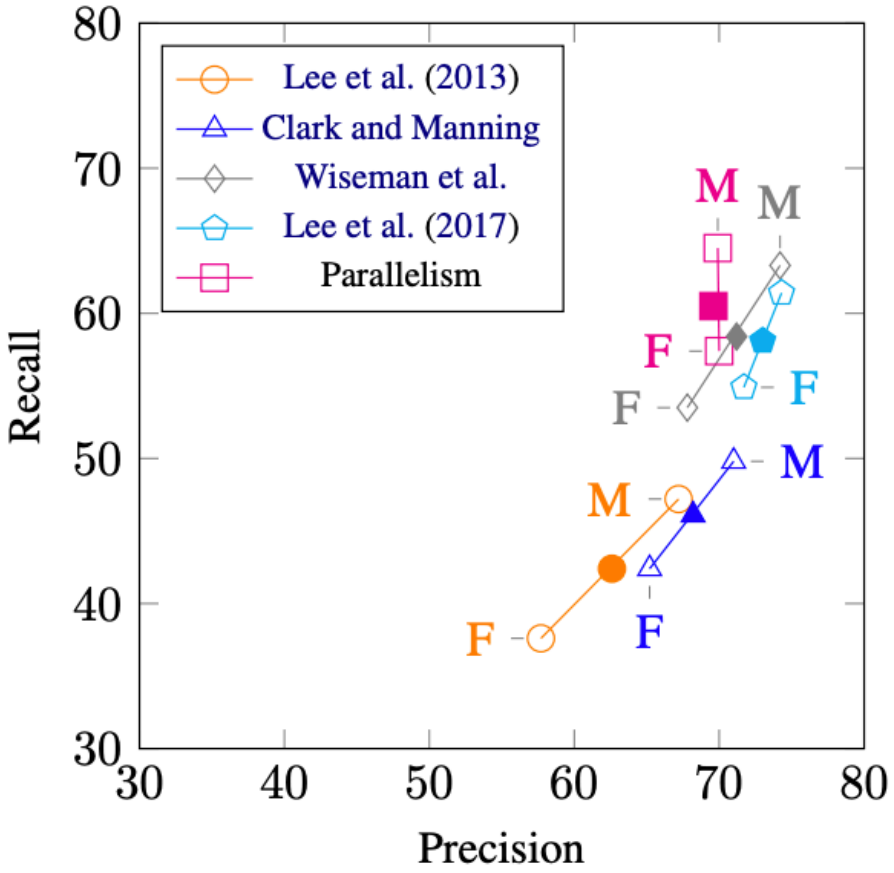


COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Other risks - Social bias



worse for female

Webster et al., (2018)

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Other risks - Social bias

OntoNotes
- 80% of gendered pronouns are male
- Male mentions are twice as likely to contain a job title

Zhao et al. (2018)

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Other risks - Is this a task we want to create?

One research group introduced this task:

Given a photo of a face,
predict if the person is a criminal !!!

1. Major ethical concerns here
2. This task doesn't make sense - we have no reason to believe it is possible to do better than chance

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

What makes a good dataset?

understand what does each property means

- Validity
  Models that perform well on the task will perform well in the real world

- Reliability
  Data is accurately labeled

- Statistical power
  Enough examples, including enough hard ones, to provide a meaningful comparison

- Testable

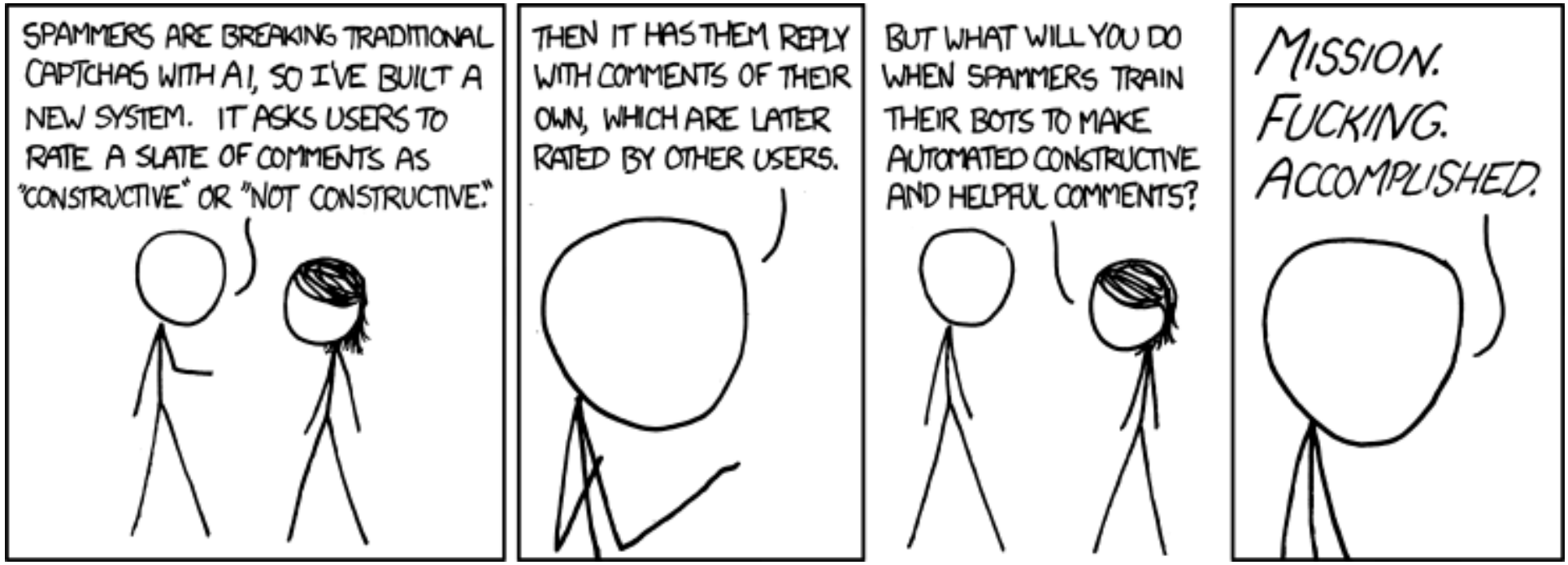
- Publicly licensed

- Avoids social bias

3 minute Break - stretch and visit Menti

menti.com
1552 7067

Constructive



["And what about all the people who won't be able to join the community because they're terrible at making helpful and constructive co-- ... oh."]

Source: https://xkcd.com/810/

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Annotation

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

What to annotate?

Real world data - may only rarely contain the phenomena we care about or may have biases

Created data - may have unintended patterns or biases

Spider (Yu et al., 2018)

NL: *List the emails of the professionals who live in the state of Hawaii or the state of Wisconsin.*

SQL:
```
select email_address from
professionals where state =
'Hawaii' or state = 'Wisconsin';
```

Suhr et al. (2020)

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Data preprocessing

All the steps we saw for FineWeb

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Tools for annotation

Benefits:
- Consistency
- Support

Annotate from scratch

Chocolate has been eaten for ….

Edit auto-generated labels

NN VBZ VBN VBN IN
Chocolate has been eaten for ….

Saves time, but could increase error if annotators start accepting everything

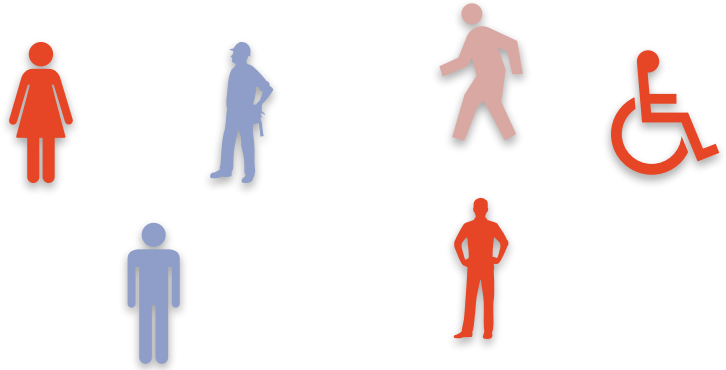
COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Annotation guidelines

Risk? Each annotator handles cases differently



Instructions
- Simple
- Clear
- Concise
- Consistent
- Generalisable

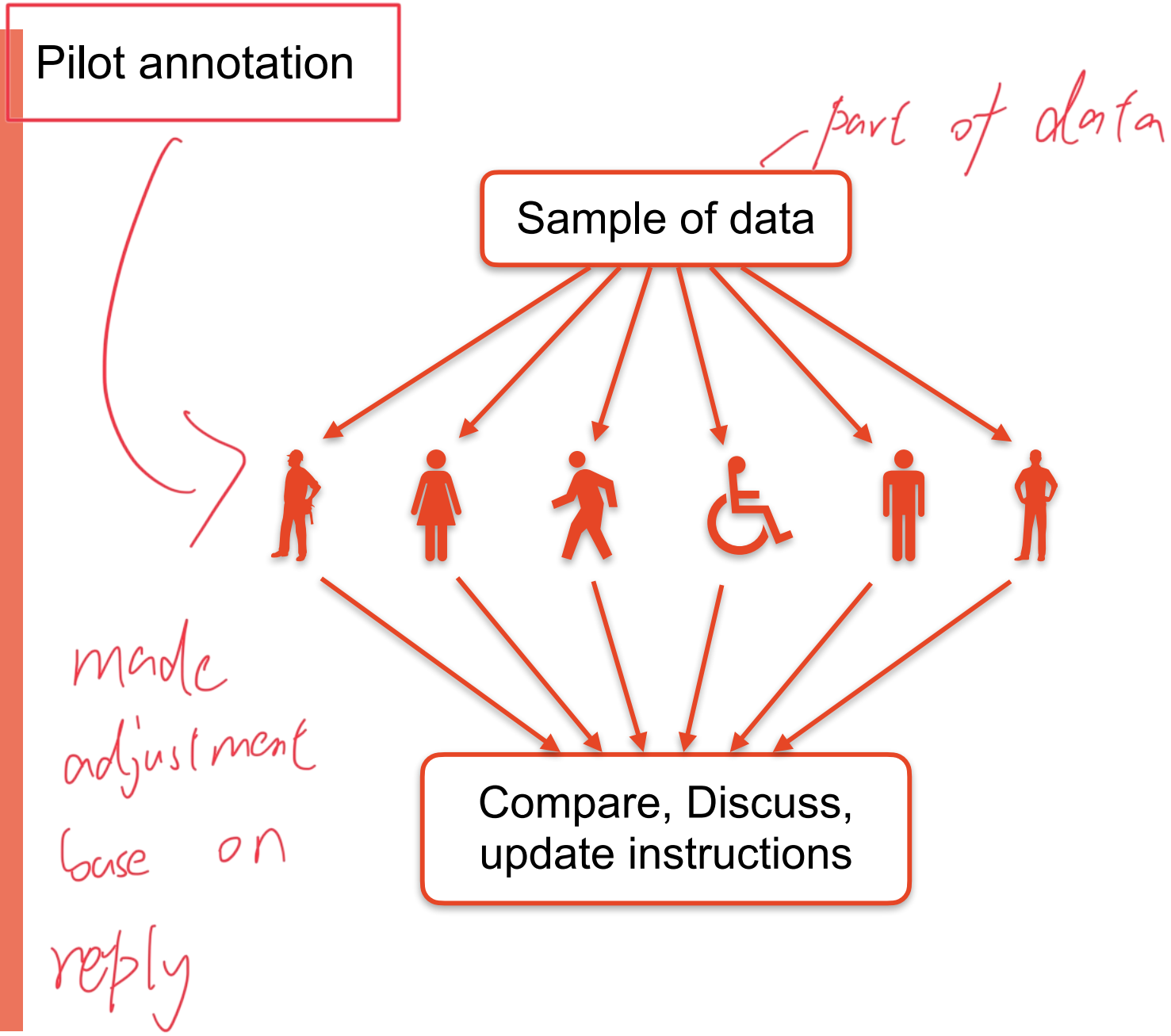
Examples
- Common cases
- Rare cases

Refine over time

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

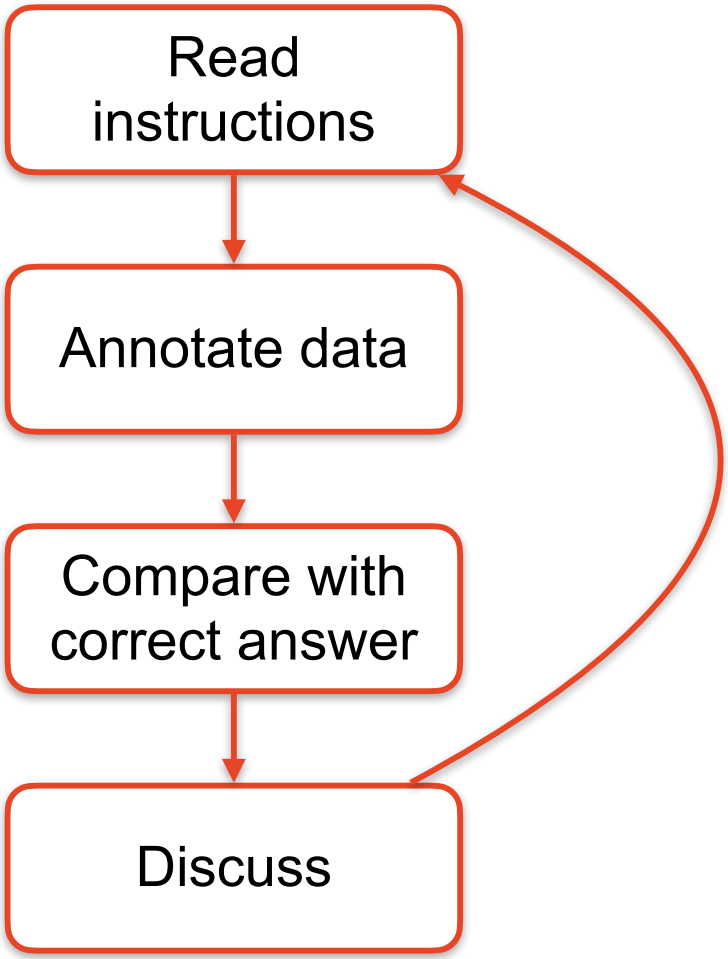


COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Train annotators



Possibly update the instructions in this process too

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Full annotation

How many people annotate each example? Tradeoff between the amount of data and annotation quality.

Training data: 1 label per example

Test data: multiple labels + adjudication

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Full annotation

Attention checks - insert out of place content to check the person is focused

> "… Ignore the rest of this paragraph, don't annotate anything …"

Consistency checks - have some fraction of examples be done by other people and raise an error if there is too much inconsistency

COMP 4446 / 5046
Lecture 10, 2025

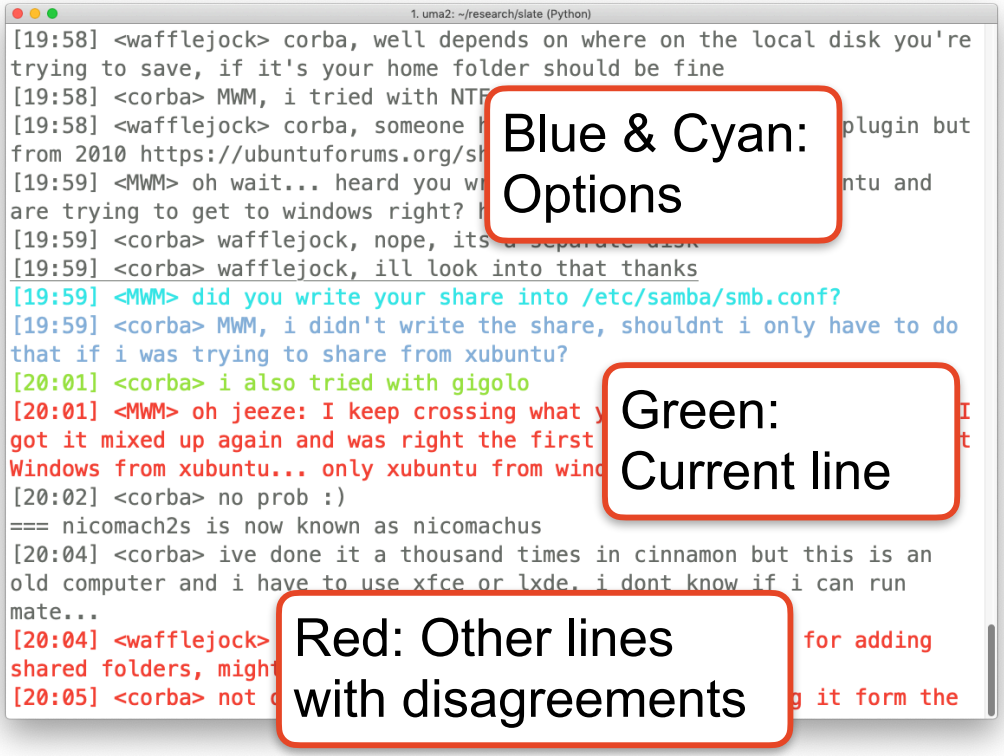
Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Adjudication

Filter to cases where there is a disagreement

Look at them together and determine a label



COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Measuring agreement

We have some labels!

But were we consistent?

maybe consistent make same mistake

Note:
consistent != good
consistent != useful

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Measuring agreement - Cohen's Kappa

Annotator 1

Annotator 2	Label A	Label B
Label A	7	4
Label B	8	81

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

$$= \frac{0.88 - p_e}{1 - p_e}$$

$$= \frac{0.88 - 0.773}{1 - 0.773}$$

$$= 0.471$$

$$p_e = P(A1 = A)P(A2 = A) + P(A1 = B)P(A2 = B)$$

$$= 0.15 * 0.11 + 0.85 * 0.89$$

$$= 0.773$$

Berkeley NLP course

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Measuring agreement - Cohen's Kappa

		Annotator 1	
		Label A	**Label B**
Annotator 2	**Label A**	1	1
	Label B	1	97

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

$$= \frac{0.98 - 0.961}{1 - 0.961}$$

$$= 0.487$$

$$p_e = P(A1 = A)P(A2 = A) + P(A1 = B)P(A2 = B)$$
$$= 0.02 * 0.02 + 0.98 * 0.98$$
$$= 0.961$$

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Measuring agreement - Cohen's Kappa

		Annotator 1	
		Label A	**Label B**
Annotator 2	**Label A**	49	1
	Label B	1	49

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

$$= \frac{0.98 - 0.5}{1 - 0.5}$$

$$= 0.96$$

$$p_e = P(A1 = A)P(A2 = A) + P(A1 = B)P(A2 = B)$$

$$= 0.5 * 0.5 + 0.5 * 0.5$$

$$= 0.5$$

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Measuring agreement - Cohen's Kappa

		Annotator 1	
		Label A	**Label B**
Annotator 2	**Label A**	25	25
	Label B	25	25

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

$$= \frac{0.5 - 0.5}{1 - 0.5}$$

$$= 0$$

$$p_e = P(A1 = A)P(A2 = A) + P(A1 = B)P(A2 = B)$$

$$= 0.5 * 0.5 + 0.5 * 0.5$$

$$= 0.5$$

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Measuring agreement - Cohen's Kappa

meaning of k

What is a good value? One very rough guide:

0.80-1.00	Very good agreement
0.60-0.80	Good agreement
0.40-0.60	Moderate agreement
0.20-0.40	Fair agreement
< 0.20	Poor agreement

Berkeley NLP course

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Measuring agreement - Fleiss' Kappa

Similar idea, but for more than 2 annotators

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

Number of annotators who labeled item i with label j n_{ij}

Overall agreement

$$P_o = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{n(n-1)} \sum_{j=1}^{K} n_{ij}(n_{ij} - 1) \right)$$

Average over items

Overlap in labels

Chance agreement

$$P_e = \sum_{j=1}^{K} \left(\frac{1}{Nn} \sum_{i=1}^{N} n_{ij} \right)^2$$

Average over classes

Squared proportion assigned to this class

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Measuring agreement - Krippendorf's Alpha

more general

More general approach that supports other data types (e.g., ranks, ratios, intervals)

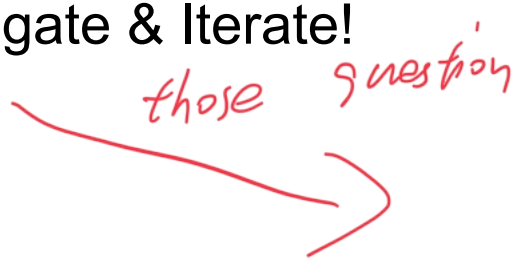
COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

What do you do if agreement is low?

Investigate & Iterate!



Careless annotators?

Unclear guidelines?

Ambiguous cases?

Subjective cases?

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Releasing data - Hosting

GitHub

HuggingFace Datasets

Your own website

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Releasing data - Licenses

Do you have consent to release the data?

Have you anonymised the data where appropriate?

Do you have the right to share the data?

Under what license?

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

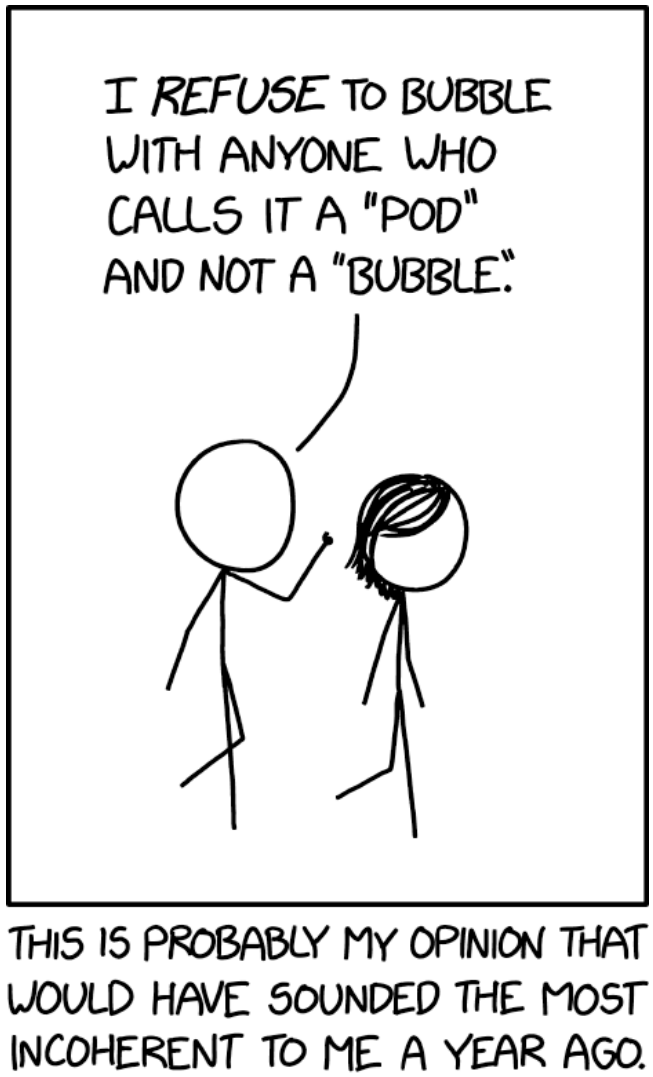
Annotation is human subjects research

Universities - Seek ethics approval

Companies - See guidelines

3 minute Break - stretch and visit Menti

menti.com
1552 7067



Pods vs Bubbles

["Canada's travel restrictions on the US are 99% about keeping out COVID and 1% about keeping out people who say 'pod.'"]

Source: https://xkcd.com/2339/

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Crowdsourcing

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

COMP 4446 / 5046
Lecture 10, 2025

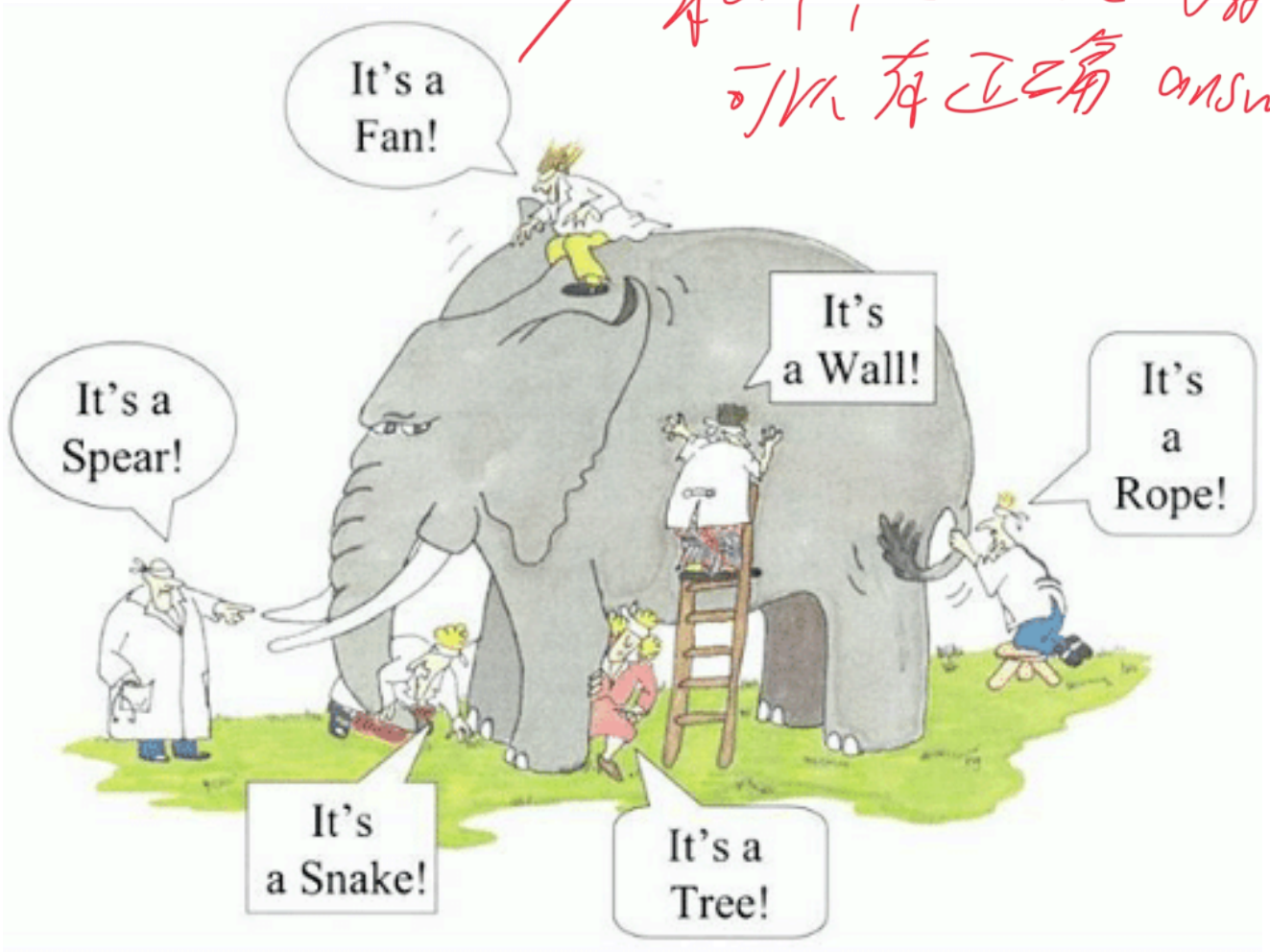
Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067



COMP 4446 / 5046
Lecture 10, 2025

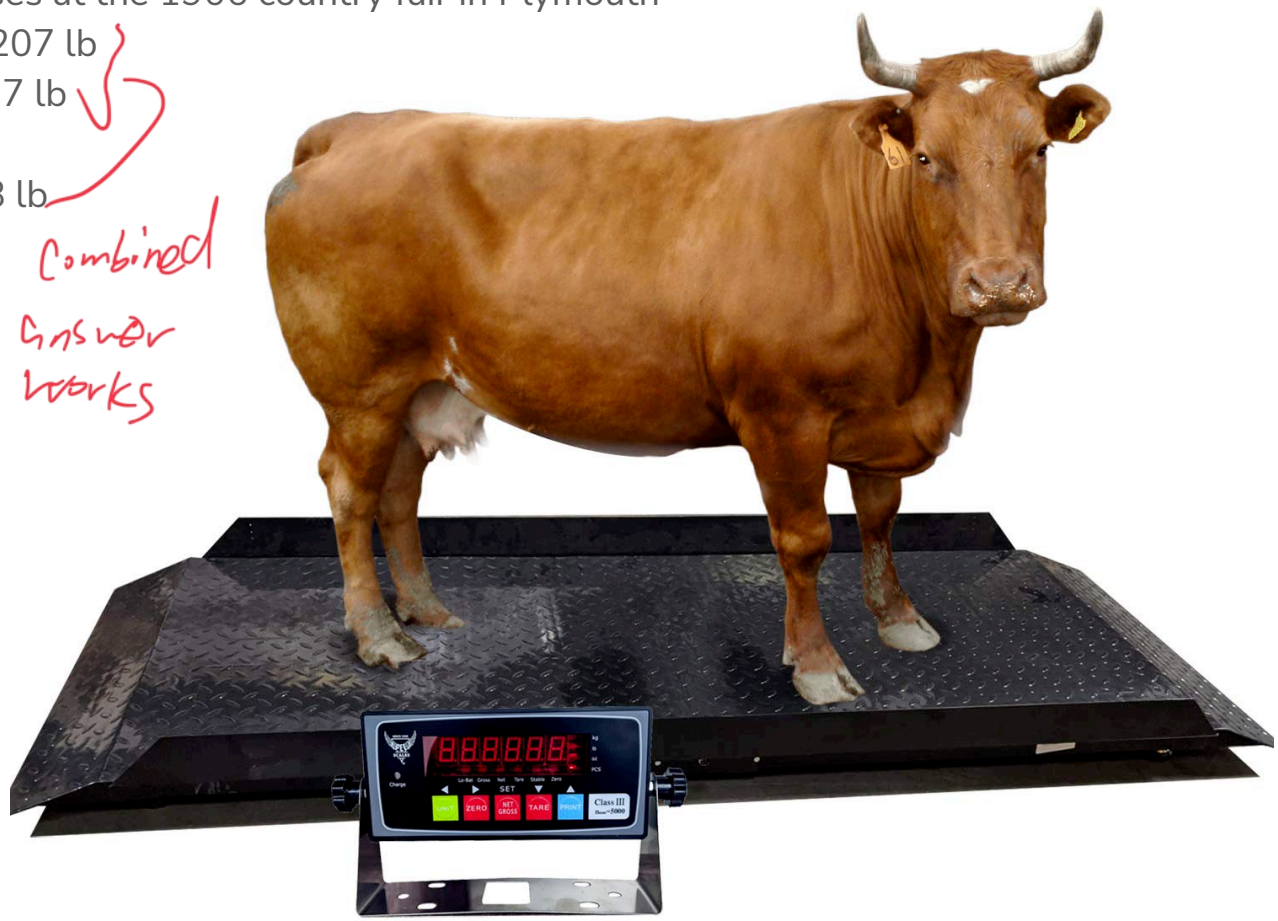
Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

800 Guesses at the 1906 country fair in Plymouth
Median: 1207 lb
Mean: 1197 lb

True: 1198 lb

Combined answer works



PEC Large Livestock Scales, Amazon

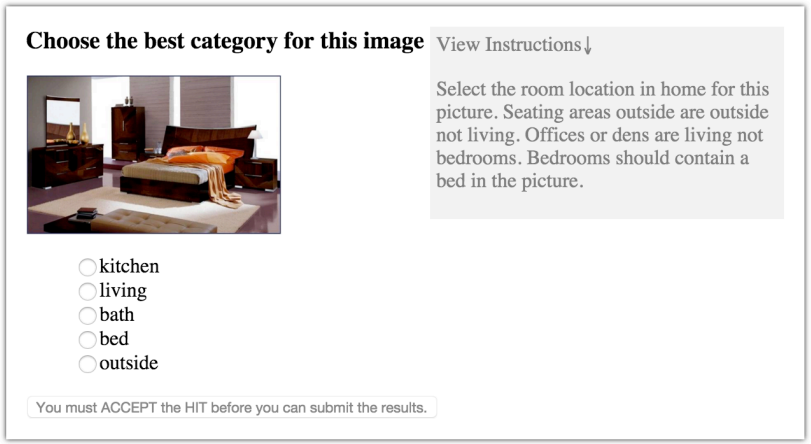
COMP 4446 / 5046
Lecture 10, 2025

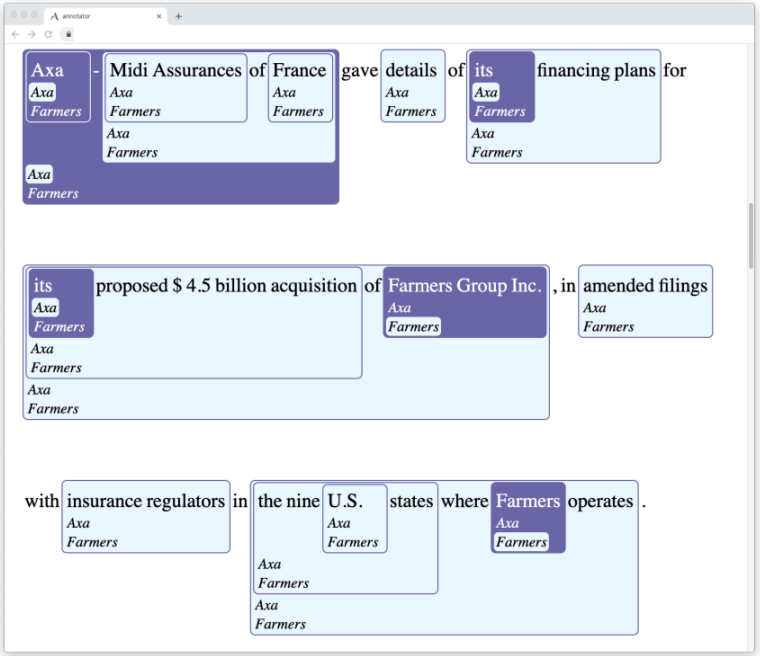
Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Microwork

Paid, small tasks





amazon mechanical turk

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Games with a Purpose

Unpaid, make task fun

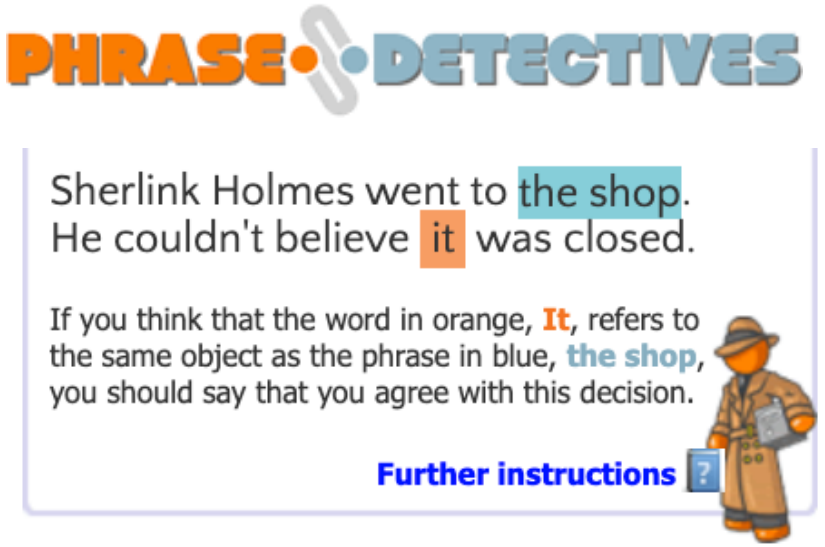
COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Games with a Purpose

Competition with prize money



COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Citizen Science / Collaboration

Unpaid, maybe fun? Satisfying

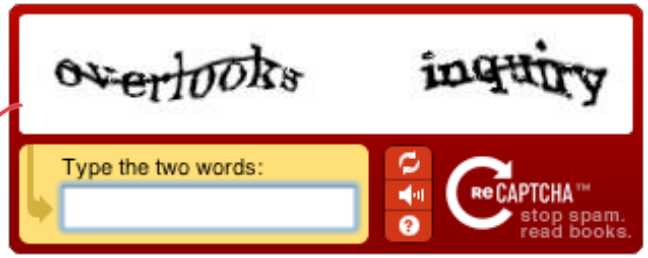
COMP 4446 / 5046
Lecture 10, 2025

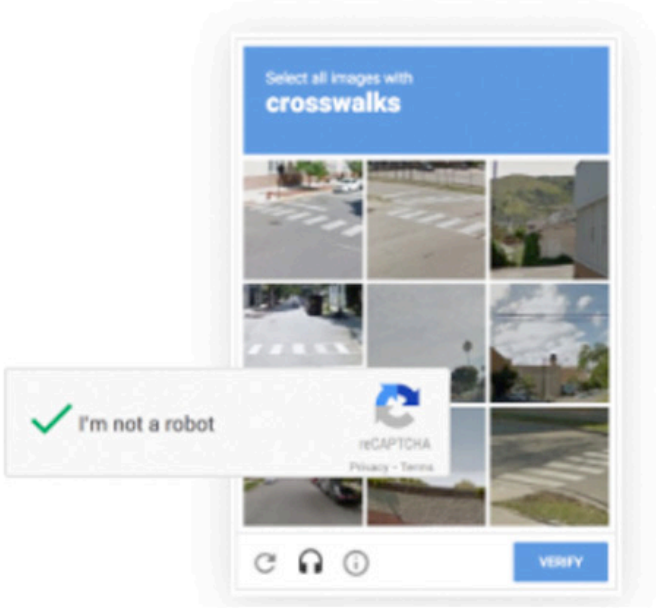
Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Implicit Work

Unpaid, not fun, provides a benefit





Only know one real label

one is for check

one is for collect data

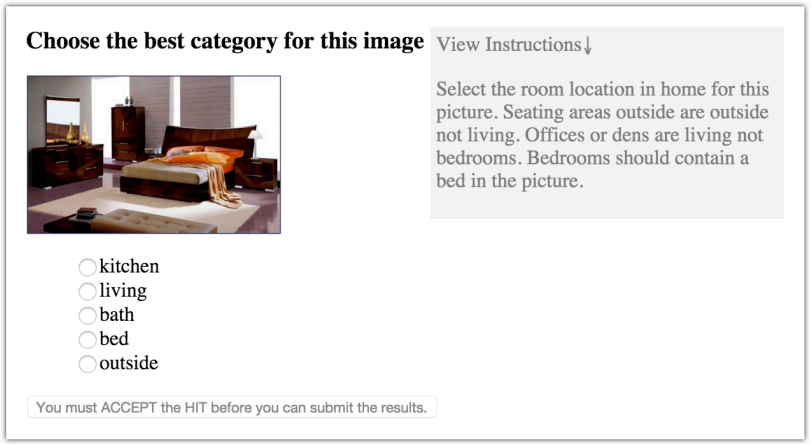
COMP 4446 / 5046
Lecture 10, 2025

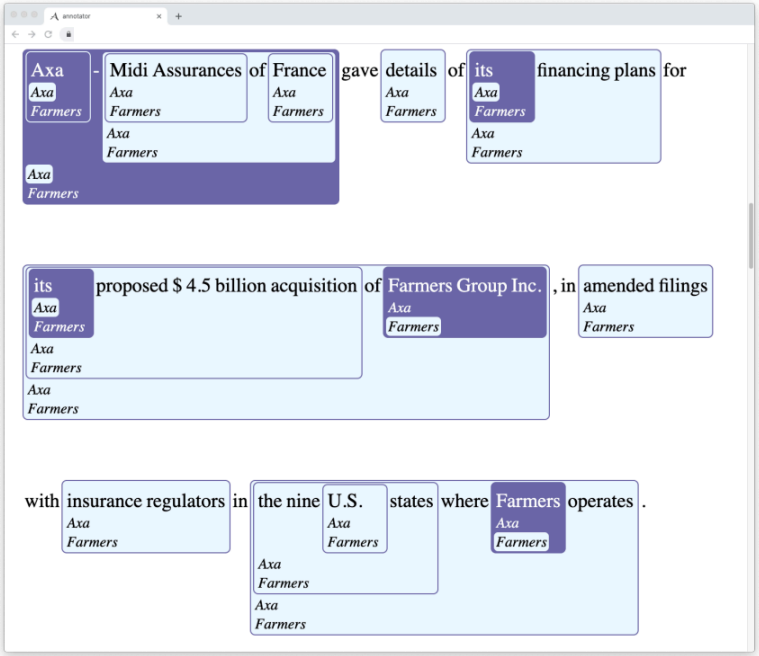
Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Microwork

Paid, small tasks





amazon mechanical turk

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Workshop Preview

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Work on Assignment 4
+
Ask questions

COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview

menti.com 1552 7067

Muddy Card

Open shortly, closes at 7:05pm

https://saipll.shinyapps.io/student-interface/

If you do not wish to participate in the study, use the Ed form instead

Go to Ed → Lessons → Muddy Cards Lecture 10