

Quiz instructions

COMP 4446 / 5046
Lecture 10, 2025



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) and square(s): ●

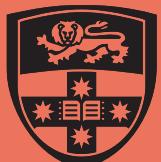
If you make a mistake, put an X over the filled in circle / square: ✗

COMP 4446 / 5046

Lecture 10: Data – Annotation and Crowdsourcing

Jonathan K. Kummerfeld

Semester 1, 2025



THE UNIVERSITY OF
SYDNEY

[menti.com 1552 7067](https://menti.com/15527067)

[Roundly-condemned
headlines initiative
shuttered indefinitely.]

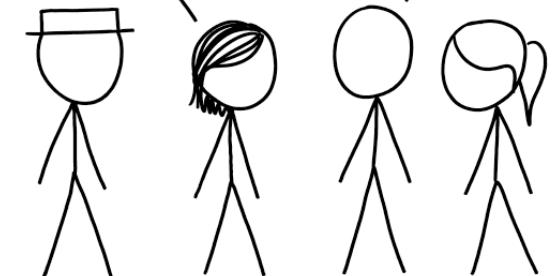
Headline Words

MAYBE ROB SHOULDN'T HOST THE PARTY. HE HAS CATS AND SOME OF US ARE ALLERGIC.

WOW, MAJOR SNUB FOR
WIDELY-TOUTED TOP SPOT
AS LAVISH GALA BID NIXED.

WHY ARE YOU TALKING
SO WEIRD? PLEASE STOP.

ILL-ADVISED SCHEME MULLED
AS TENSION MOUNTS AMID
GROWING BACKLASH.



MY PROJECT TO SPEAK ONLY IN WEIRD
HEADLINE WORDS DIDN'T LAST LONG.

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



Assignment 4

Data Sources

Annotation

Crowdsourcing

Workshop Preview

Due in 8 days

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

Apologies for delay on reference result release



[menti.com 1552 7067](https://menti.com/15527067)



Data Sources

Annotation

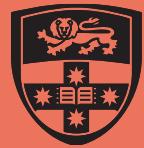
Crowdsourcing

Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

Data Sources



Data Sources

Annotation

Crowdsourcing

Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

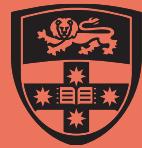
Where does the data to train these models come from?

Explore with an example:



FineWeb

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



Data Sources

Annotation

Crowdsourcing

Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

Step 1: Download data

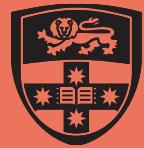


250 billion pages
17 years

mainly English data

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>

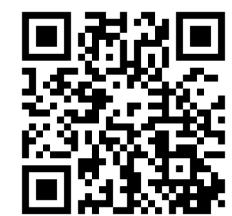


Data Sources

Annotation

Crowdsourcing

Workshop Preview



menti.com 1552 7067

Step 2: Filter the data

exclude some website

1. Url Filtering, removing URLs as subwords detect
2. Trafilatura text extraction from Crawl'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

*use model to
detect which web
is good*

ous and NSFW websites, using both block-list as well

as Crawl's warc files

removing any document with en language score lower than 0.65

Gopher Repetition / Quality

C4 Quality filters except terminal_punct rule

FineWeb custom filters, consisting of heuristics for removing list-like documents, documents with repeated lines and documents with likely wrong line formatting.

MinHash deduplication with each crawl deduplicated individually (5-grams, 14x8 hash functions)

PII Formatting to anonymize email and public IP addresses

✓ avoid duplicate data



Data Sources

Annotation

Crowdsourcing

Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

Innovation in FineWeb - choosing filters by training

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



Data Sources

Annotation

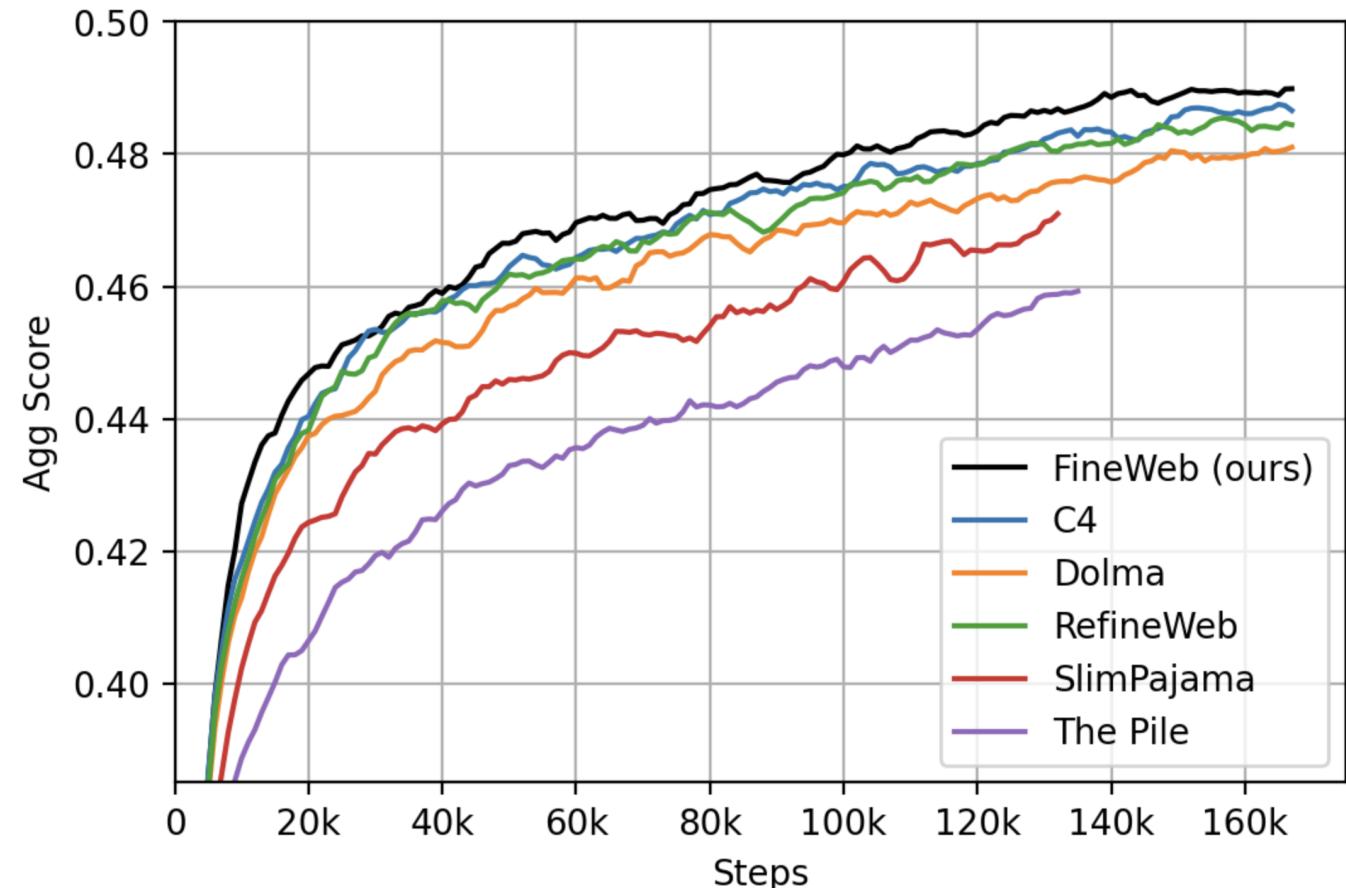
Crowdsourcing

Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

Innovation in FineWeb - choosing filters by training



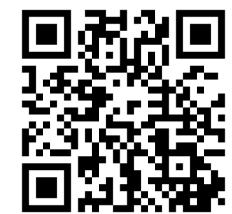


Data Sources

Annotation

Crowdsourcing

Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

LLM developers seek more + better data

BERT - Books scraped from the internet (16 Gb)

RoBERTa - Add more data! (160 Gb)

↑ prop of data

'High quality'

- Wikipedia
- News
- Scientific papers

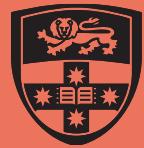
Vs.

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*



Data Sources

Annotation
Crowdsourcing
Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

What about other tasks?

Named Entity Recognition

Coreference Resolution

Syntactic Parsing

Sentiment Analysis

Translation

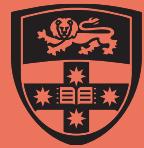
....

from

Some translated data is available:

Canadian Hansards
European Parliament
United Nations

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



internet related chat

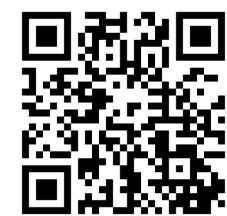
General approach, through an example:
IRC Dialogue Disentanglement Dataset

Data Sources

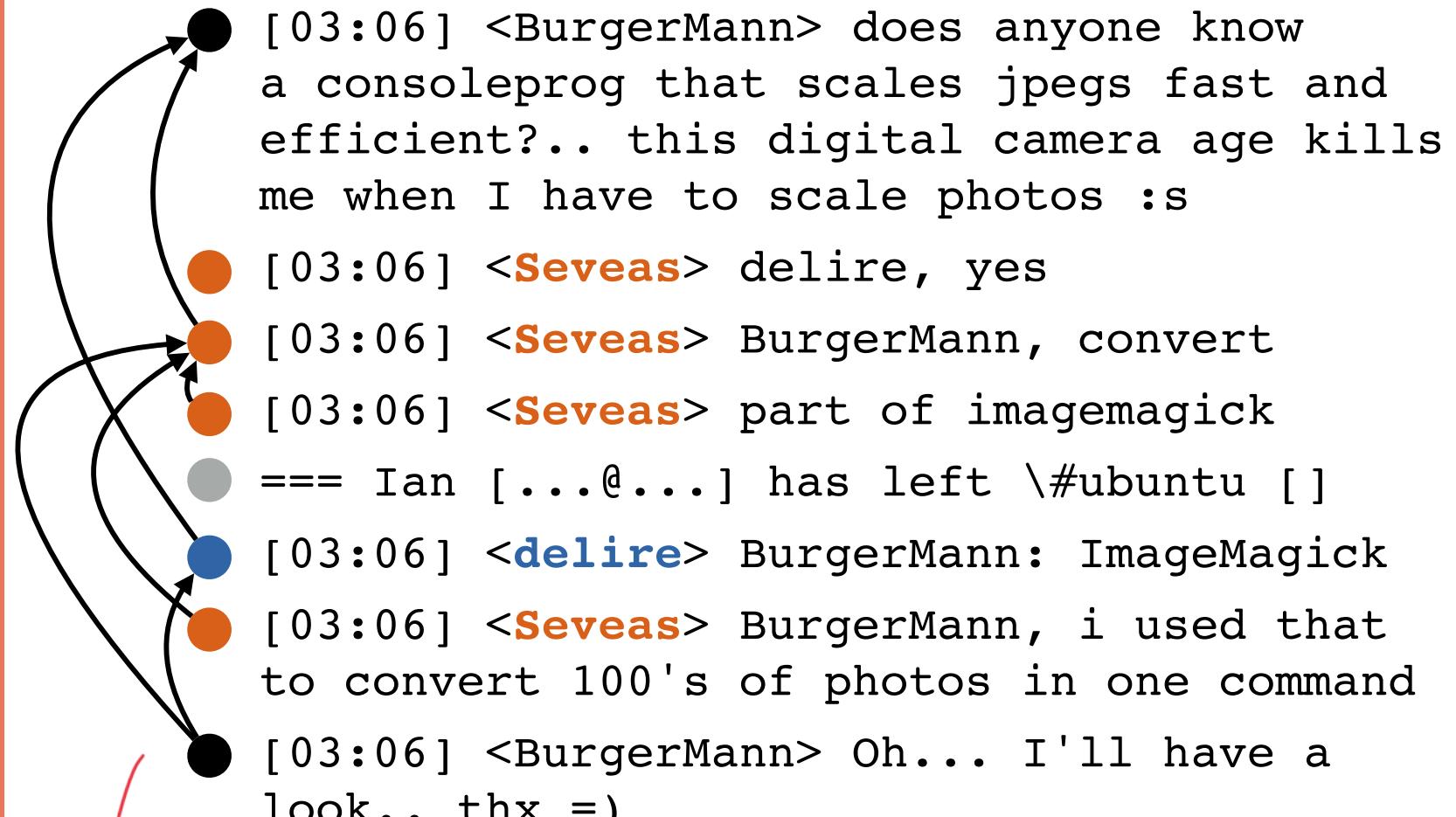
Annotation

Crowdsourcing

Workshop Preview

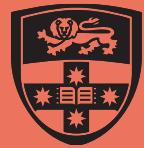


[menti.com 1552 7067](https://menti.com/15527067)



Chat at same time

show which msg related to which



Data Sources

Annotation

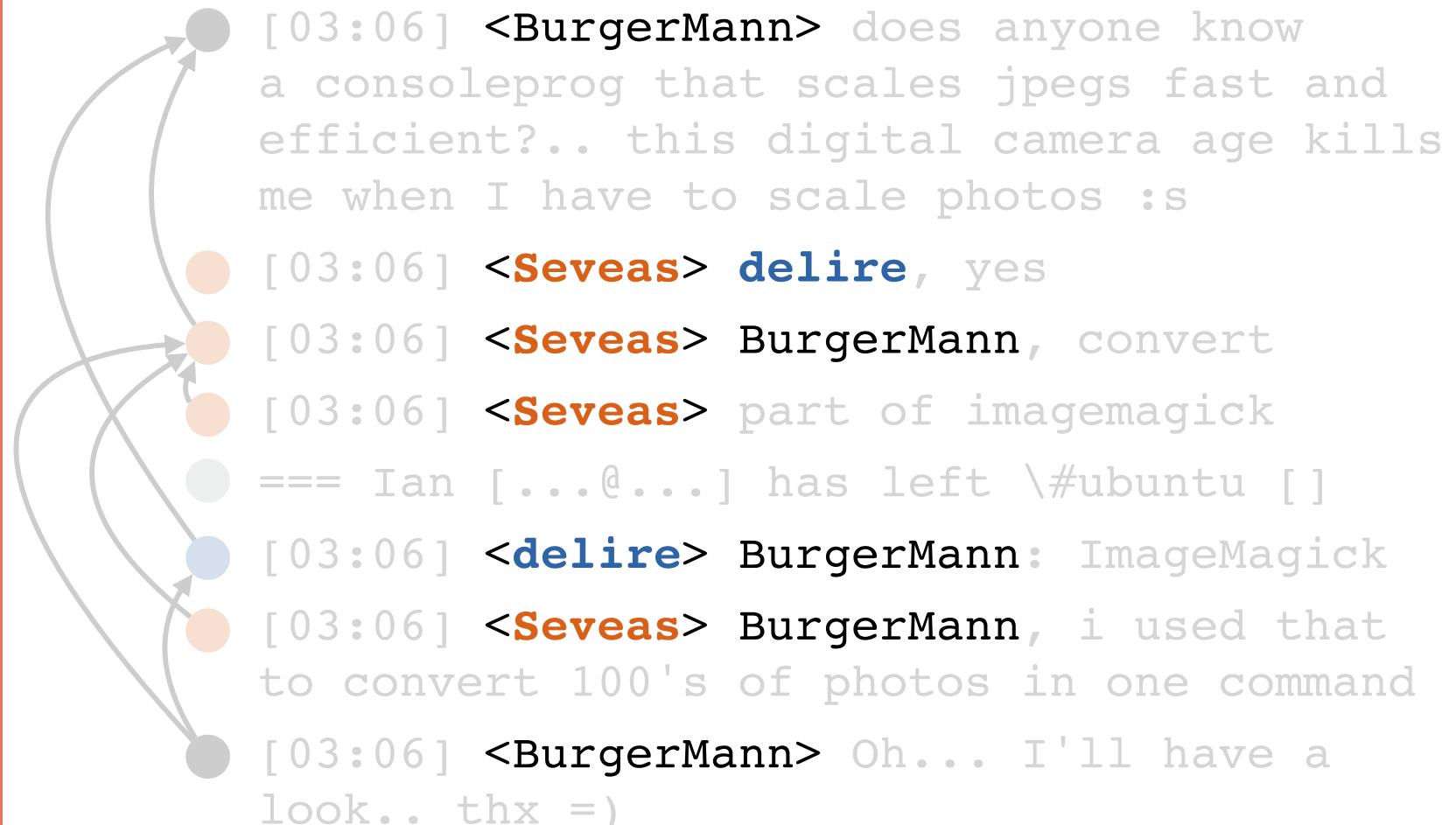
Crowdsourcing

Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

General approach, through an example: IRC Dialogue Disentanglement Dataset





Data Sources

Annotation

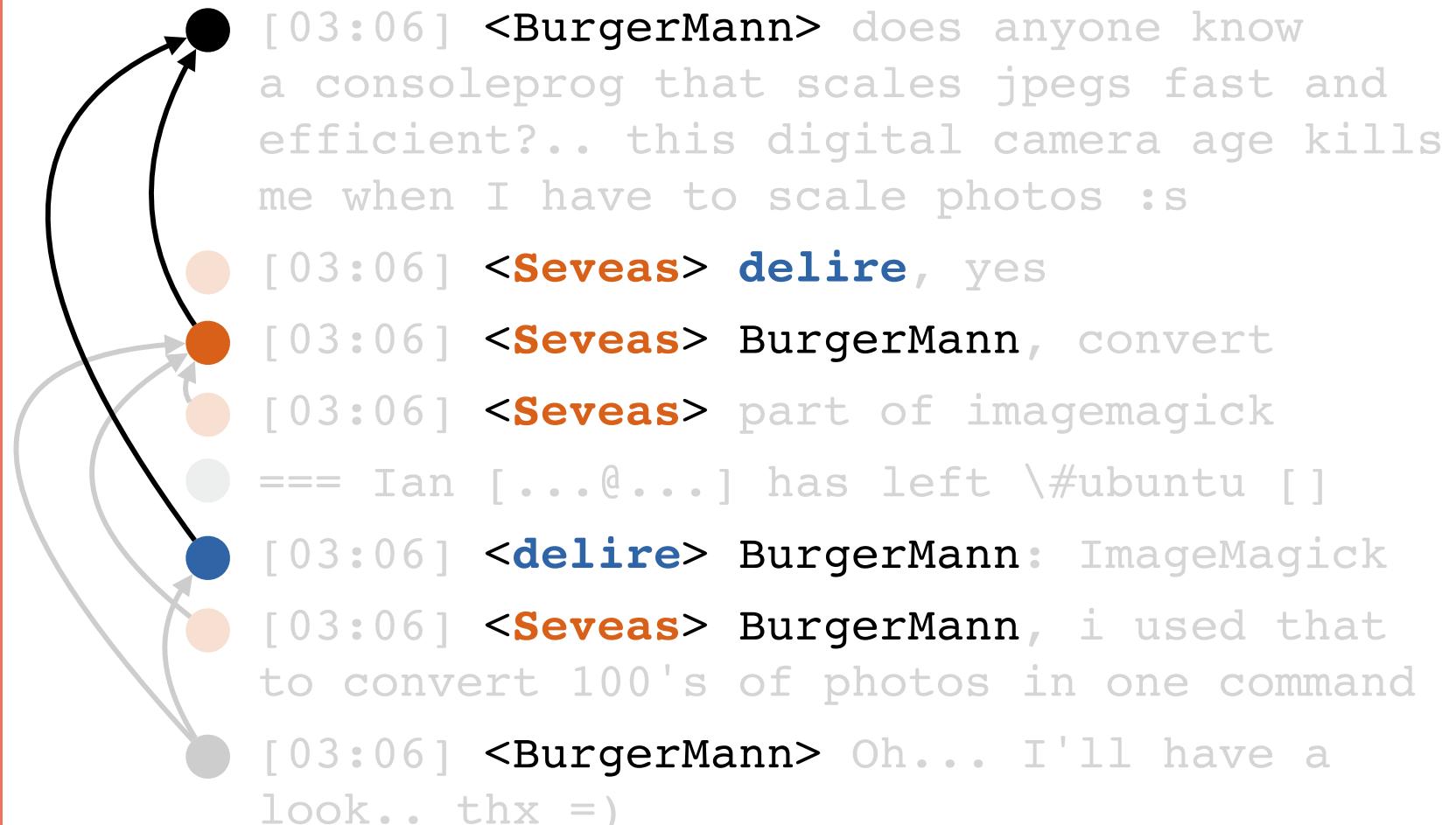
Crowdsourcing

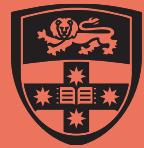
Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

General approach, through an example: IRC Dialogue Disentanglement Dataset





Data Sources

Annotation

Crowdsourcing

Workshop Preview

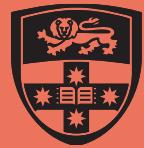


[menti.com 1552 7067](https://menti.com/15527067)

General approach, through an example: IRC Dialogue Disentanglement Dataset

Raw data:



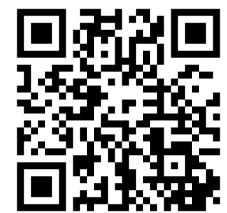


Data Sources

Annotation

Crowdsourcing

Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

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

```
1. uma2: ~/research/slate (Python)
[19:58] <wafflejock> corba, well depends on where on the local disk you're trying to save, if it's your home folder should be fine
[19:58] <corba> MWM, i tried with NTFS and ext4
[19:58] <wafflejock> corba, someone here suggests thunar-shares-plugin but from 2010 https://ubuntuforums.org/showthread.php?t=1582665
[19:59] <MWM> oh wait... heard you wrong. You are booted to xubuntu and are trying to get to windows right? how did you set up samba?
[19:59] <corba> wafflejock, nope, its a separate disk
[19:59] <corba> wafflejock, ill look into that thanks
[19:59] <MWM> did you write your share into /etc/samba/smb.conf?
[19:59] <corba> MWM, i didn't write the share, shouldnt i only have to do that if i was trying to share from xubuntu?
[20:01] <corba> i also tried with gigolo
[20:01] <MWM> oh jeeze: I keep crossing what you are saying iun my head. I got it mixed up again and was right the first time. I dont know how to get Windows from xubuntu... only xubuntu from windows. sorry
[20:02] <corba> no prob :)
== nicomach2s is now known as nicomachus
[20:04] <corba> ive done it a thousand times in cinnamon but this is an old computer and i have to use xfce or lxde, i dont know if i can run mate...
[20:04] <wafflejock> ah yeah the shares plugin appears to be for adding shared folders, might be useful but not for this issue
[20:05] <corba> not only that, im having problems downloading it form the
```



Data Sources

Annotation

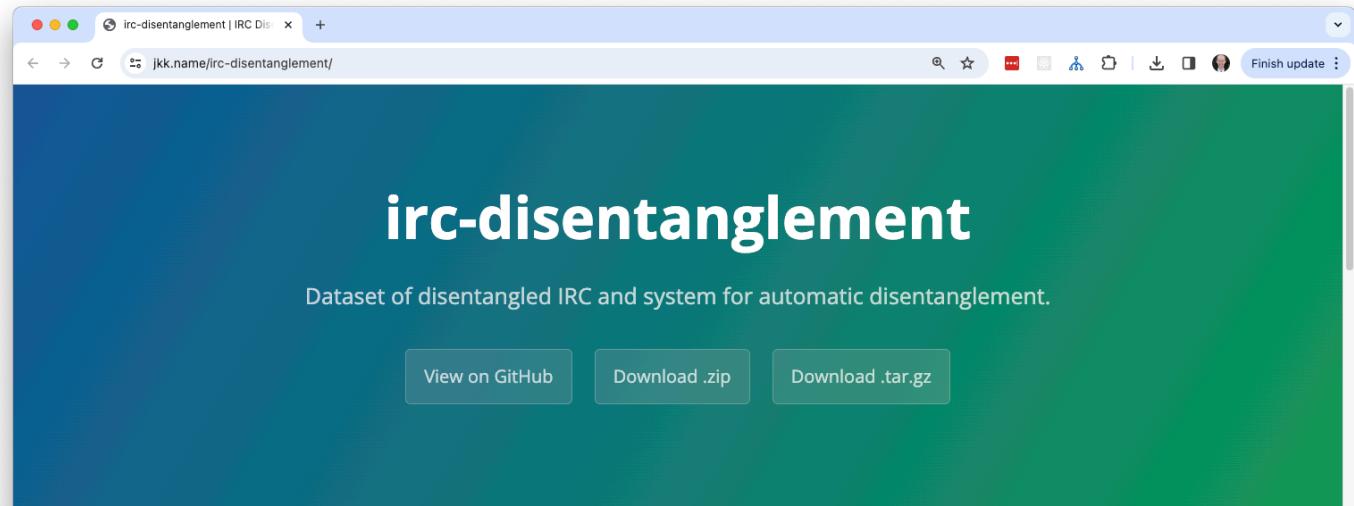
Crowdsourcing

Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

General approach, through an example: IRC Dialogue Disentanglement Dataset



The screenshot shows a web browser window with the title "irc-disentanglement | IRC Disentanglement Dataset". The URL in the address bar is "jkk.name/irc-disentanglement/". The main content features a large green gradient header with the text "irc-disentanglement" in white. Below it, a sub-header reads "Dataset of disentangled IRC and system for automatic disentanglement." Three buttons are visible: "View on GitHub", "Download .zip", and "Download .tar.gz".

irc-disentanglement

Dataset of disentangled IRC and system for automatic disentanglement.

[View on GitHub](#) [Download .zip](#) [Download .tar.gz](#)

This repository contains data and code for disentangling conversations on IRC, as described in the following two papers:

- [A Large-Scale Corpus for Conversation Disentanglement](#), Jonathan K. Kummerfeld, Sai R. Gouravajhala, Joseph Peper, Vignesh Athreya, Chulaka Gunasekara, Jatin Ganhotra, Siva Sankalp Patel, Lazaros Polymenakos, and Walter S. Lasecki, ACL 2019
- [Chat Disentanglement: Data for New Domains and Methods for More Accurate Annotation](#), Sai R. Gouravajhala, Andrew M. Vernier, Yiming Shi, Zihan Li, Mark Ackerman, Jonathan K. Kummerfeld

Conversation disentanglement is the task of identifying separate conversations in a single stream of messages. For example, the image below shows two entangled conversations and an annotated graph structure (indicated by lines and colours). The example includes a message that receives



Data Sources

Annotation

Crowdsourcing

Workshop Preview



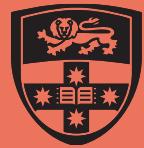
menti.com 1552 7067

General approach, through an example: IRC Dialogue Disentanglement Dataset

The screenshot shows a web browser window with the title "Evaluation" from the "irc-disentanglement | IRC Disentanglement Dataset" website. The page content includes:

- Evaluation**: We provide three scripts for evaluation.
- Graph Evaluation**: This calculates precision, recall, and F-score over edges.
Command: `python3 tools/evaluation/graph-eval.py --gold gold-files --auto system-files`
- The expected format for output files is:
Format: `anything.../filename:line-number line-number -`
- For example:
Example output:

```
blah-blah/2004-11-15.annotation.txt:1000 1000 -
blah-blah/2004-11-15.annotation.txt:1001 1001 -
blah-blah/2004-11-15.annotation.txt:1002 1002 -
blah-blah/2004-11-15.annotation.txt:1003 1003 -
yaddah_yaddah/2004-11-15.annotation.txt:1004 1003 -
yaddah_yaddah/2004-11-15.annotation.txt:1005 1003 -
```
- Conversation Evaluation**: This calculates a range of cluster metrics. It requires the [Google OR Tools](#).
Command: `python3 tools/evaluation/conversation-eval.py gold-file system-file`
- The expected format for output files has each cluster on one line:
Format: `anything.../filename:line-number line-number line-number line-number ...`
- For example:

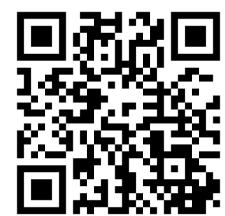


Data Sources

Annotation

Crowdsourcing

Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

Datasets can have massive impact

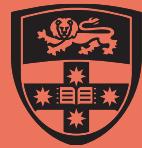
Penn Treebank - syntactic annotation, 11,000 citations

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



Data Sources

Annotation

Crowdsourcing

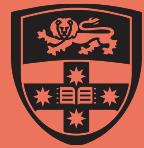
Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

My most cited paper is a dataset

	TITLE	CITED BY	YEAR
<input type="checkbox"/>	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
<input type="checkbox"/>	Improving Text-to-SQL Evaluation Methodology C Finegan-Dollak, JK Kummerfeld, L Zhang, K Ramanathan, ... ACL	328	2018
<input type="checkbox"/>	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
<input type="checkbox"/>	Factors Influencing the Surprising Instability of Word Embeddings L Wendlandt, JK Kummerfeld, R Mihalcea NAACL	166	2018
<input type="checkbox"/>	A Large-Scale Corpus for Conversation Disentanglement JK Kummerfeld, SR Gouravajhala, J Peper, V Athreya, C Gunasekara, ... ACL	135*	2019
<input type="checkbox"/>	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
<input type="checkbox"/>	Tools for Automated Analysis of Cybercriminal Markets RS Portnoff, S Afroz, G Durrett, JK Kummerfeld, T Berg-Kirkpatrick, ... WWW	114	2017
<input type="checkbox"/>	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

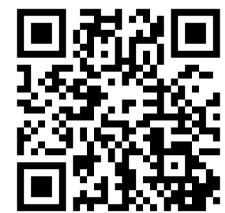


Data Sources

Annotation

Crowdsourcing

Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

Evaluate on multiple tasks at once - avoid overfitting to one

*get multiple data sets 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

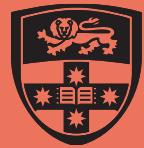
MultiNLI

Question NLI

Recognizing Textual Entailment

Winograd NLI

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

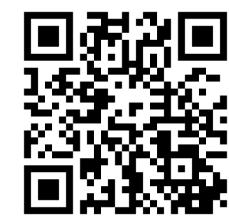


Data Sources

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”

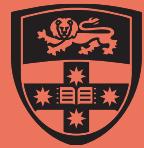


"If the eighth day of the month comes on a Tuesday, what is the date of the following Friday?"
Hans would answer by tapping his hoof eleven times

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

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

Wikipedia



Data Sources

Annotation

Crowdsourcing

Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

Other risks - Shortcuts / Validity

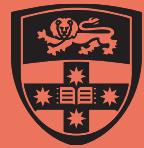
What does Macduff do to Macbeth?

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

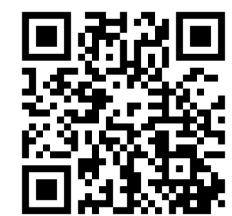


Data Sources

Annotation

Crowdsourcing

Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

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.”

Where ...

Who ...

When ...

short cut from input

Only one option in many cases!

From the UT Austin NLP class



Data Sources

Annotation

Crowdsourcing

Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

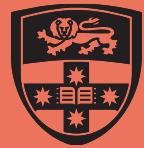
Other risks - Shortcuts / Validity

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



Data Sources

Annotation

Crowdsourcing

Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

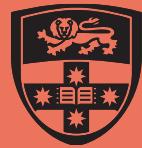
Other risks - Shortcuts / Validity

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 test cases

From the UT Austin NLP class

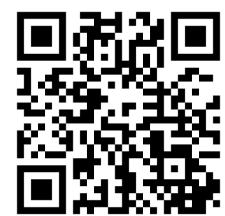


Data Sources

Annotation

Crowdsourcing

Workshop Preview

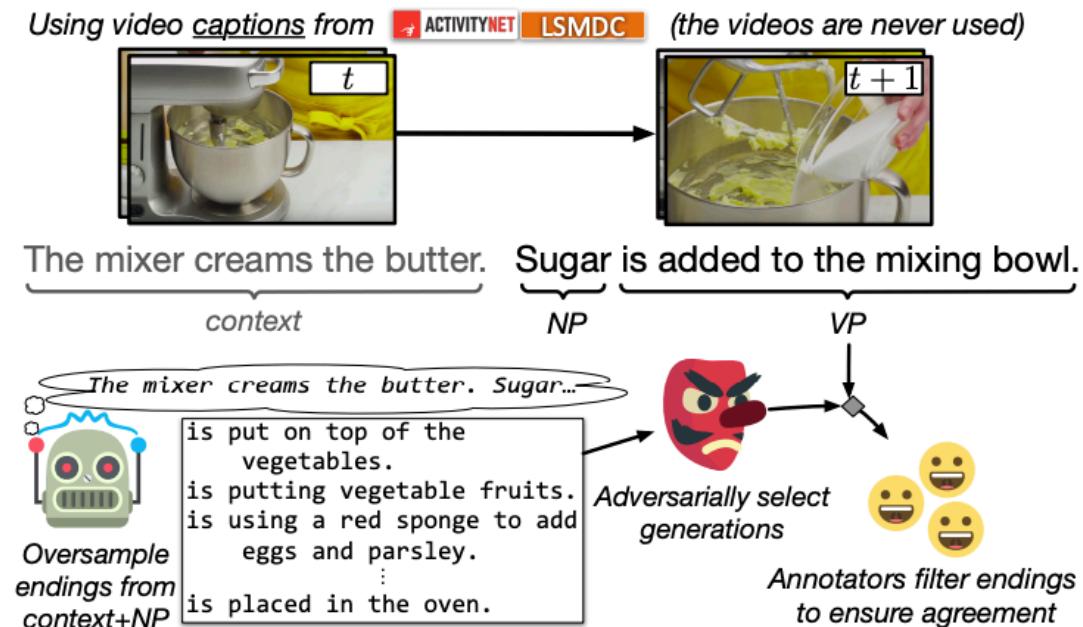


[menti.com 1552 7067](https://menti.com/15527067)

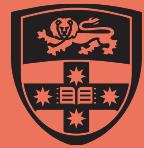
Other risks - Does our task have statistical power?

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

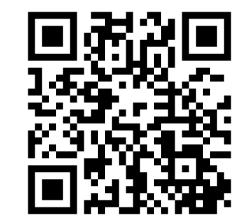


Data Sources

Annotation

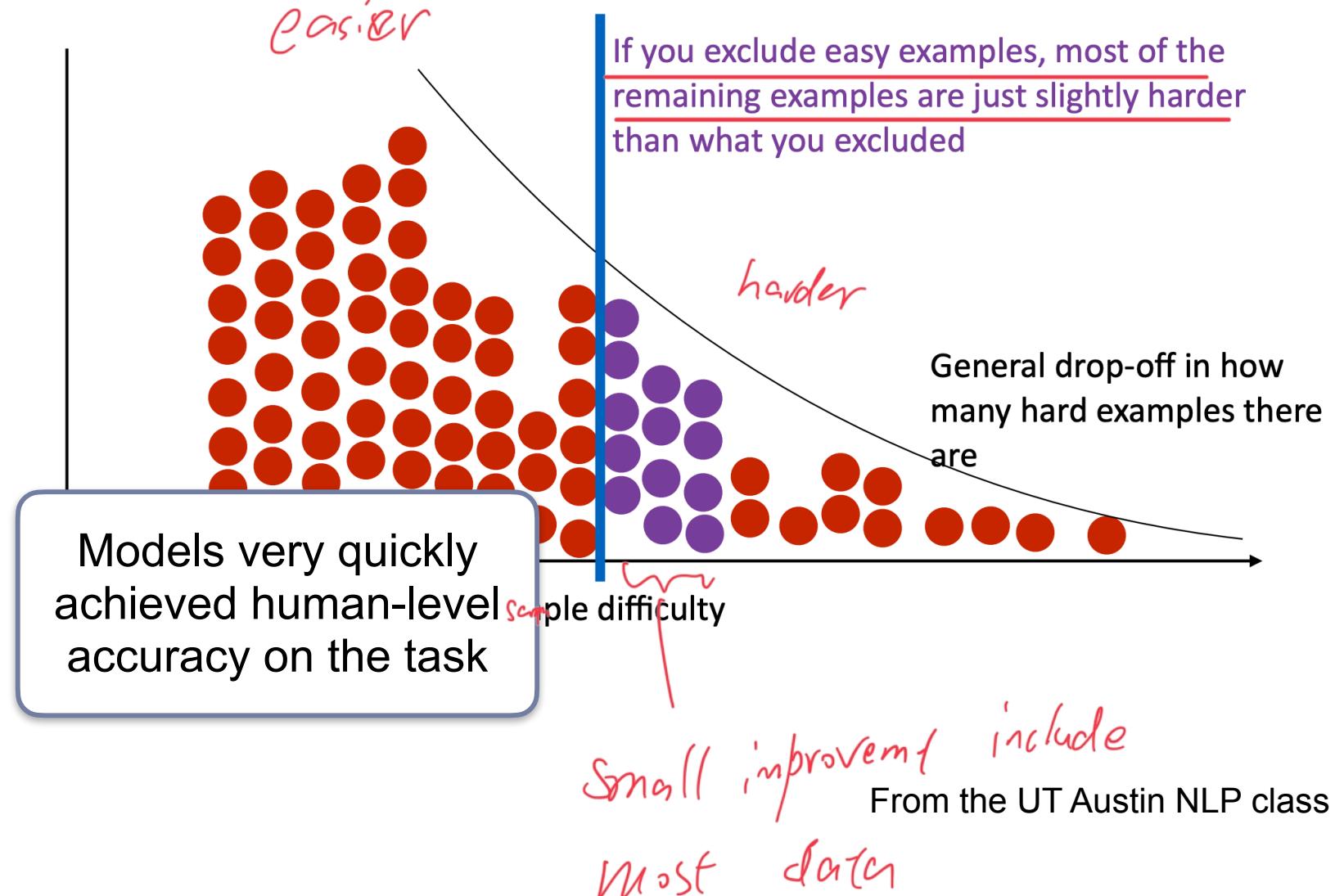
Crowdsourcing

Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

Other risks - Does our task have statistical power?





Data Sources

Annotation

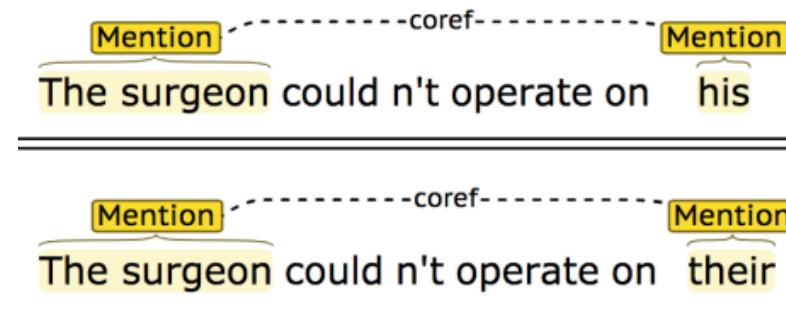
Crowdsourcing

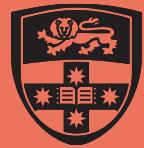
Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

Other risks - Social bias





Data Sources

Annotation

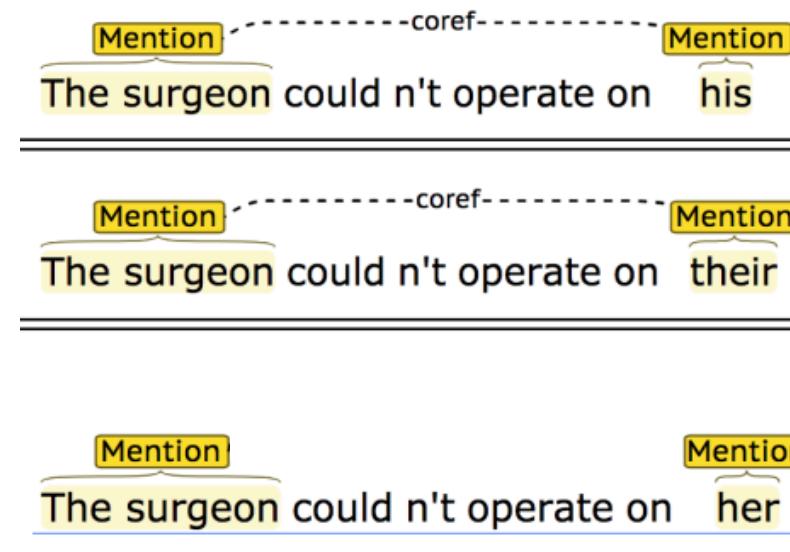
Crowdsourcing

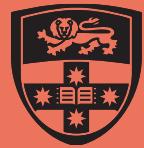
Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

Other risks - Social bias





Data Sources

Annotation

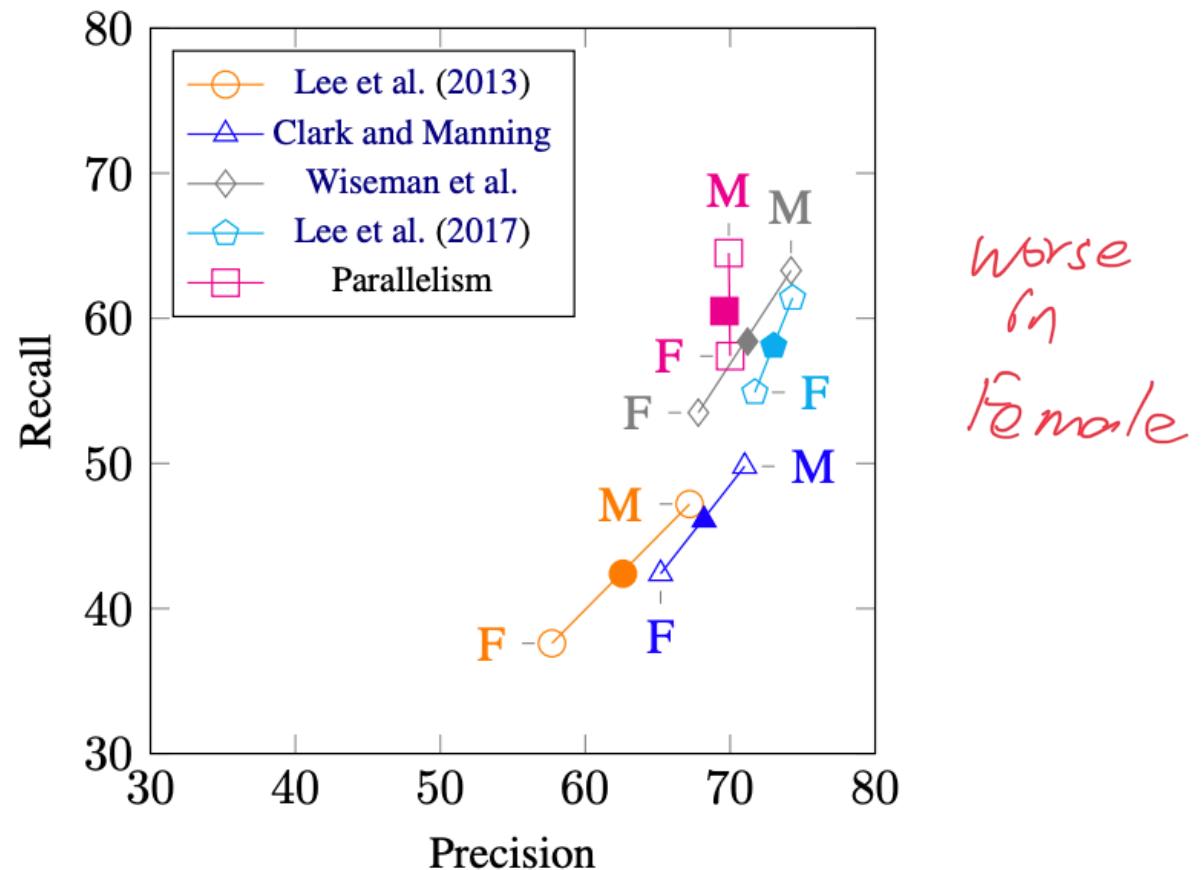
Crowdsourcing

Workshop Preview

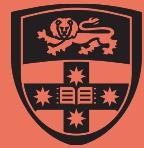


[menti.com 1552 7067](https://menti.com/15527067)

Other risks - Social bias



Webster et al., (2018)



Data Sources

Annotation

Crowdsourcing

Workshop Preview



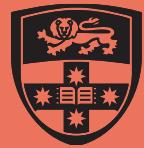
[menti.com 1552 7067](https://menti.com/15527067)

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)



Data Sources

Annotation

Crowdsourcing

Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

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



Data Sources

Annotation

Crowdsourcing

Workshop Preview

*Understand what
does each property means*

What makes a good dataset?

- 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

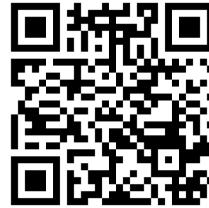
- Testable

- Publicly licensed

- Avoids social bias



[menti.com 1552 7067](https://menti.com/15527067)



Constructive

SPAMMERS ARE BREAKING TRADITIONAL CAPTCHAS WITH AI, SO I'VE BUILT A NEW SYSTEM. IT ASKS USERS TO RATE A SLATE OF COMMENTS AS "CONSTRUCTIVE" OR "NOT CONSTRUCTIVE."



THEN IT HAS THEM REPLY WITH COMMENTS OF THEIR OWN, WHICH ARE LATER RATED BY OTHER USERS.



BUT WHAT WILL YOU DO WHEN SPAMMERS TRAIN THEIR BOTS TO MAKE AUTOMATED CONSTRUCTIVE AND HELPFUL COMMENTS?



MISSION.
FUCKING.
ACCOMPLISHED.



[“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

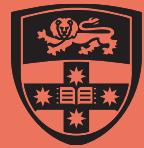
Crowdsourcing

Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

Annotation



Data Sources

Annotation

Crowdsourcing

Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

What to annotate?

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

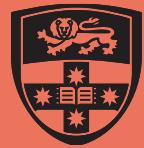
Created data - may have unintended patterns or biases

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)



Data Sources

Annotation

Crowdsourcing

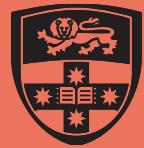
Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

Data preprocessing

All the steps we saw for FineWeb



Data Sources

Annotation

Crowdsourcing

Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

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



Data Sources

Annotation

Crowdsourcing

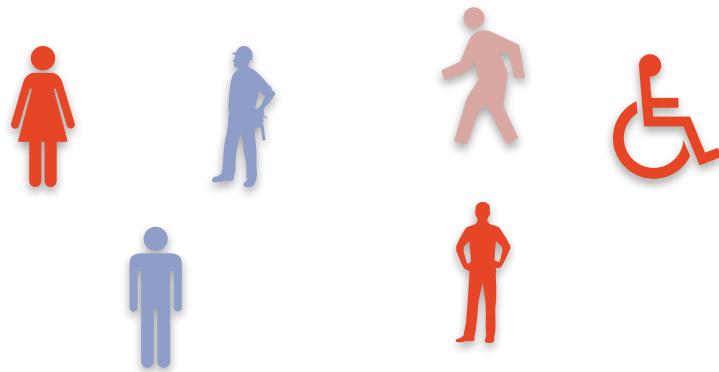
Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

Annotation guidelines

Risk? Each annotator handles cases differently



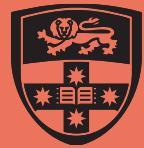
Instructions

- Simple
- Clear
- Concise
- Consistent
- Generalisable

Examples

- Common cases
- Rare cases

Refine over time



Data Sources

Annotation

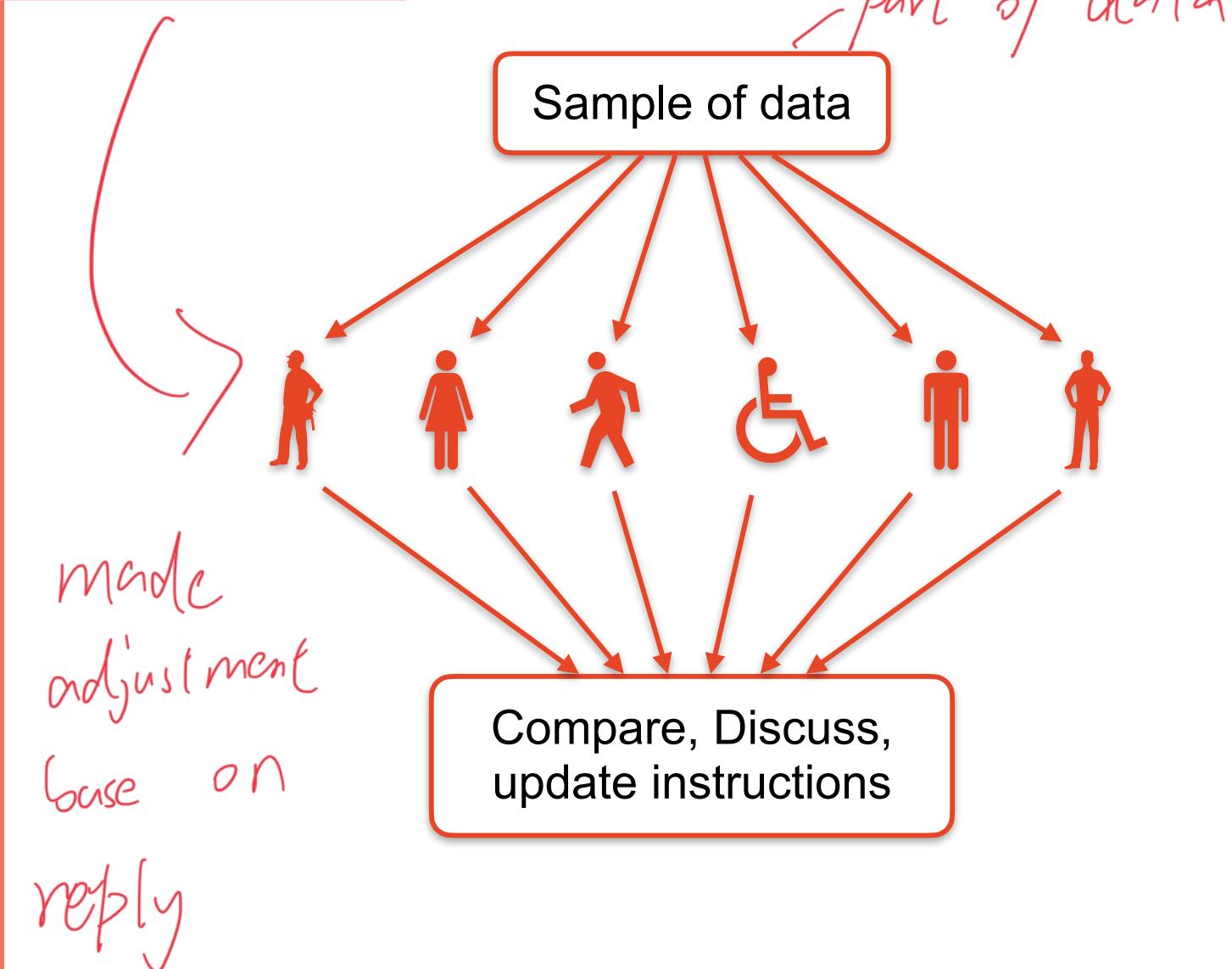
Crowdsourcing

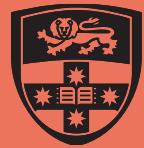
Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

Pilot annotation





Data Sources

Annotation

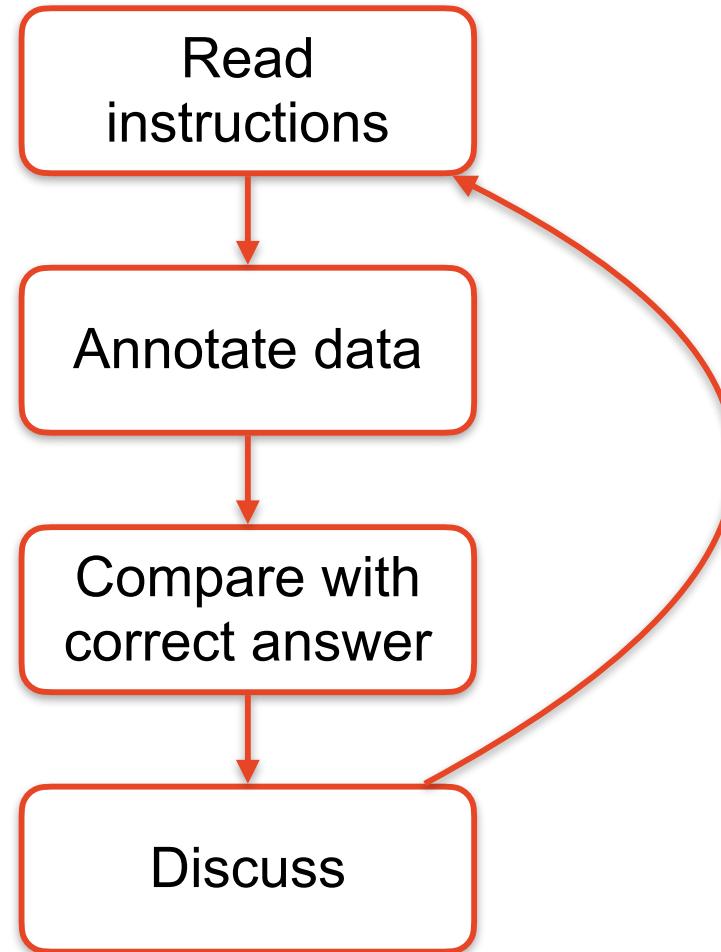
Crowdsourcing

Workshop Preview

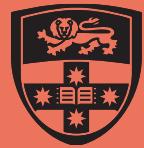


[menti.com 1552 7067](https://menti.com/15527067)

Train annotators



Possibly update the instructions in this process too



Data Sources

Annotation

Crowdsourcing

Workshop Preview



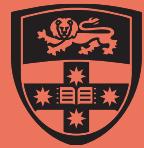
[menti.com 1552 7067](https://menti.com/15527067)

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

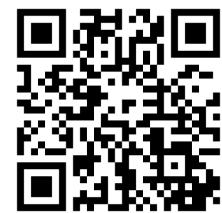


Data Sources

Annotation

Crowdsourcing

Workshop Preview



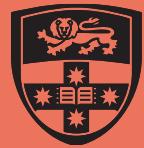
[menti.com 1552 7067](https://menti.com/15527067)

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



Data Sources

Annotation

Crowdsourcing

Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

Adjudication

Filter to cases where there is a disagreement

Look at them together and determine a label

```
1. uma2: ~/research/slate (Python)
[19:58] <wafflejock> corba, well depends on where on the local disk you're trying to save, if it's your home folder should be fine
[19:58] <corba> MWM, i tried with NTFS
[19:58] <wafflejock> corba, someone I from 2010 https://ubuntuforums.org/sh
[19:59] <MWM> oh wait... heard you w are trying to get to windows right? I
[19:59] <corba> wafflejock, nope, its a separate disk
[19:59] <corba> wafflejock, ill look into that thanks
[19:59] <MWM> did you write your share into /etc/samba/smb.conf?
[19:59] <corba> MWM, i didn't write the share, shouldnt i only have to do that if i was trying to share from xubuntu?
[20:01] <corba> i also tried with gigolo
[20:01] <MWM> oh jeeze: I keep crossing what y got it mixed up again and was right the first Windows from xubuntu... only xubuntu from wind
[20:02] <corba> no prob :)
==== nicomach2s is now known as nicomachus
[20:04] <corba> ive done it a thousand times in cinnamon but this is an old computer and i have to use xfce or lxde. i dont know if i can run mate...
[20:04] <wafflejock> shared folders, might
[20:05] <corba> not o
```

Blue & Cyan: Options

Green: Current line

Red: Other lines with disagreements

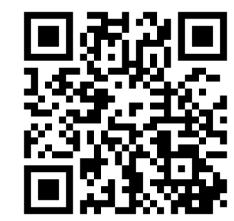


Data Sources

Annotation

Crowdsourcing

Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

Measuring agreement

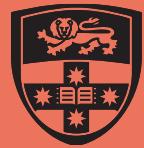
We have some labels!

But were we consistent?

maybe consistent make
some mistake

Note:

consistent != good
consistent != useful



Data Sources

Annotation

Crowdsourcing

Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

Measuring agreement - Cohen's Kappa

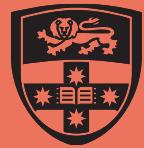
		Annotator 1	
		Label A Label B	
		Label A	7
		Label B	81
Annotator 2			
7 + 8			
$\frac{7+8}{100}$			

$$p_e = \frac{P(A1 = A)P(A2 = A) + P(A1 = B)P(A2 = B)}{100}$$
$$= \frac{0.15 * 0.11 + 0.85 * 0.89}{100}$$
$$= 0.773$$

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$
$$= \frac{0.88 - p_e}{1 - p_e}$$
$$= \frac{0.88 - 0.773}{1 - 0.773}$$
$$= 0.471$$

$$\frac{8+81}{100}$$
$$\frac{4+81}{100}$$

Berkeley NLP course



Data Sources

Annotation

Crowdsourcing

Workshop Preview



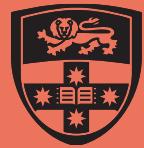
[menti.com 1552 7067](https://menti.com/15527067)

Measuring agreement - Cohen's Kappa

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

$$\begin{aligned}\kappa &= \frac{p_o - p_e}{1 - p_e} \\ &= \frac{0.98 - 0.961}{1 - 0.961} \\ &= 0.487\end{aligned}$$

$$\begin{aligned}p_e &= P(A1 = A)P(A2 = A) + P(A1 = B)P(A2 = B) \\ &= 0.02 * 0.02 + 0.98 * 0.98 \\ &= 0.961\end{aligned}$$



Data Sources

Annotation

Crowdsourcing

Workshop Preview



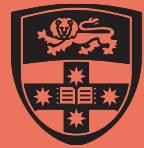
[menti.com 1552 7067](https://menti.com/15527067)

Measuring agreement - Cohen's Kappa

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

$$\begin{aligned}\kappa &= \frac{p_o - p_e}{1 - p_e} \\ &= \frac{0.98 - 0.5}{1 - 0.5} \\ &= 0.96\end{aligned}$$

$$\begin{aligned}p_e &= P(A1 = A)P(A2 = A) + P(A1 = B)P(A2 = B) \\ &= 0.5 * 0.5 + 0.5 * 0.5 \\ &= 0.5\end{aligned}$$



Data Sources

Annotation

Crowdsourcing

Workshop Preview



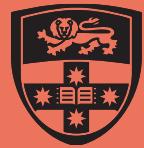
[menti.com 1552 7067](https://menti.com/15527067)

Measuring agreement - Cohen's Kappa

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

$$\begin{aligned}\kappa &= \frac{p_o - p_e}{1 - p_e} \\ &= \frac{0.5 - 0.5}{1 - 0.5} \\ &= 0\end{aligned}$$

$$\begin{aligned}p_e &= P(A1 = A)P(A2 = A) + P(A1 = B)P(A2 = B) \\ &= 0.5 * 0.5 + 0.5 * 0.5 \\ &= 0.5\end{aligned}$$



Data Sources

Annotation

Crowdsourcing

Workshop Preview

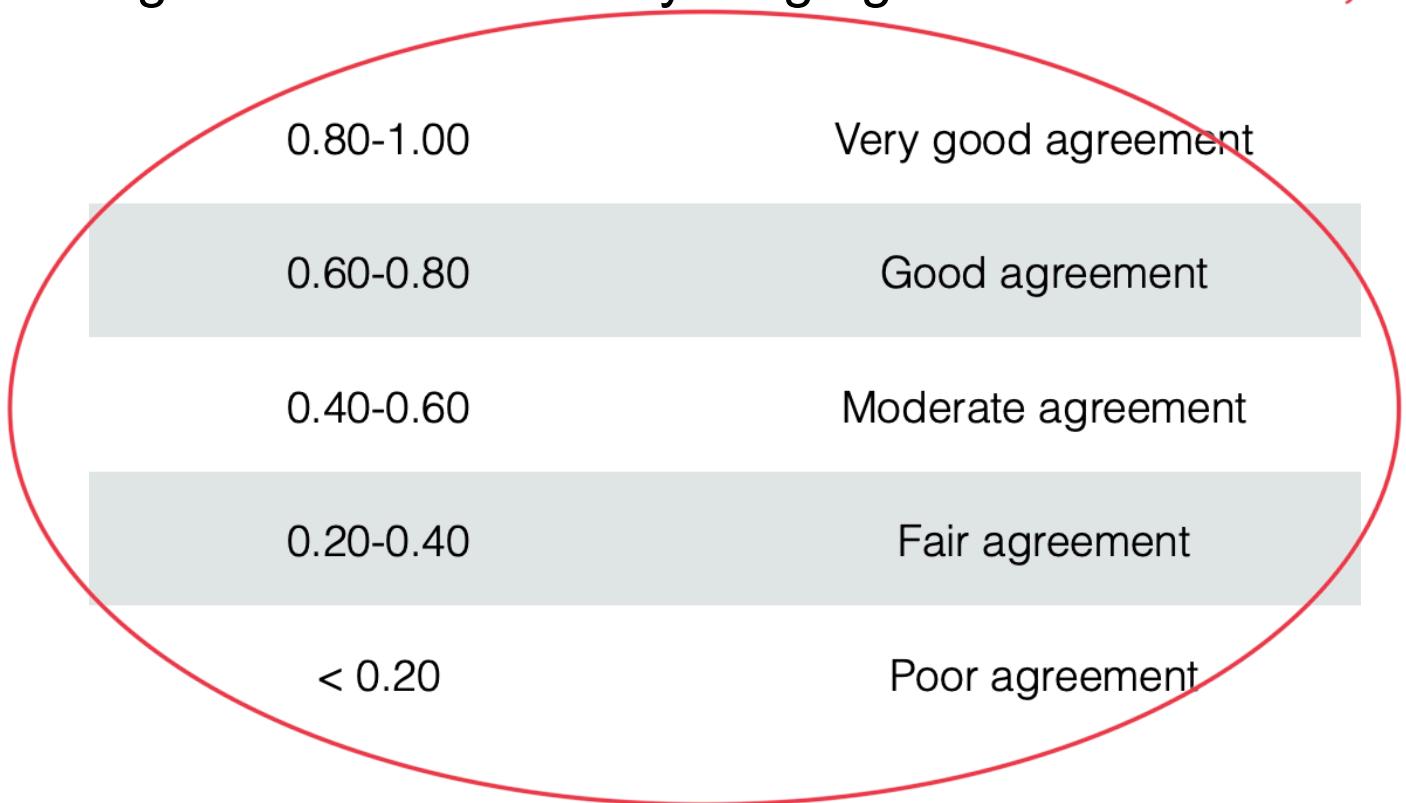


[menti.com 1552 7067](https://menti.com/15527067)

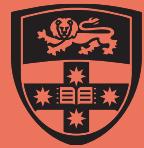
Measuring agreement - Cohen's Kappa

What is a good value? One very rough guide:

meaning of κ



Berkeley NLP course



Data Sources

Annotation

Crowdsourcing

Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

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

Overall agreement

Average over items

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

Chance agreement

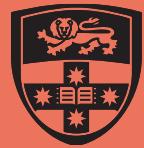
Average over classes

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

n_{ij}

Overlap in labels

Squared proportion assigned to this class



Data Sources

Annotation

Crowdsourcing

Workshop Preview

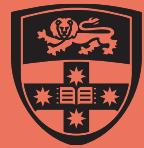


[menti.com 1552 7067](https://menti.com/15527067)

Measuring agreement - Krippendorf's Alpha

more general

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



Data Sources

Annotation

Crowdsourcing

Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

What do you do if agreement is low?

Investigate & Iterate!

those question

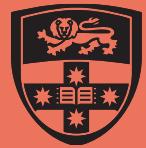


Unclear
guidelines?

Careless
annotators?

Ambiguous
cases?

Subjective
cases?



Data Sources

Annotation

Crowdsourcing

Workshop Preview



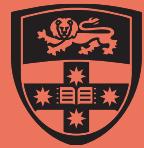
[menti.com 1552 7067](https://menti.com/15527067)

Releasing data - Hosting

GitHub

HuggingFace Datasets

Your own website



Data Sources

Annotation

Crowdsourcing

Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

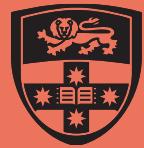
Releasing data - Licenses

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?



Data Sources

Annotation

Crowdsourcing

Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

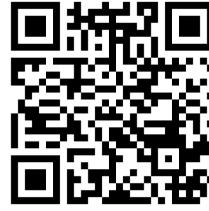
Annotation is human subjects research

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

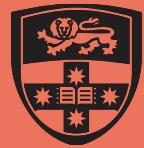
Crowdsourcing

Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

Crowdsourcing



COMP 4446 / 5046
Lecture 10, 2025

Data Sources

Annotation

Crowdsourcing

Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

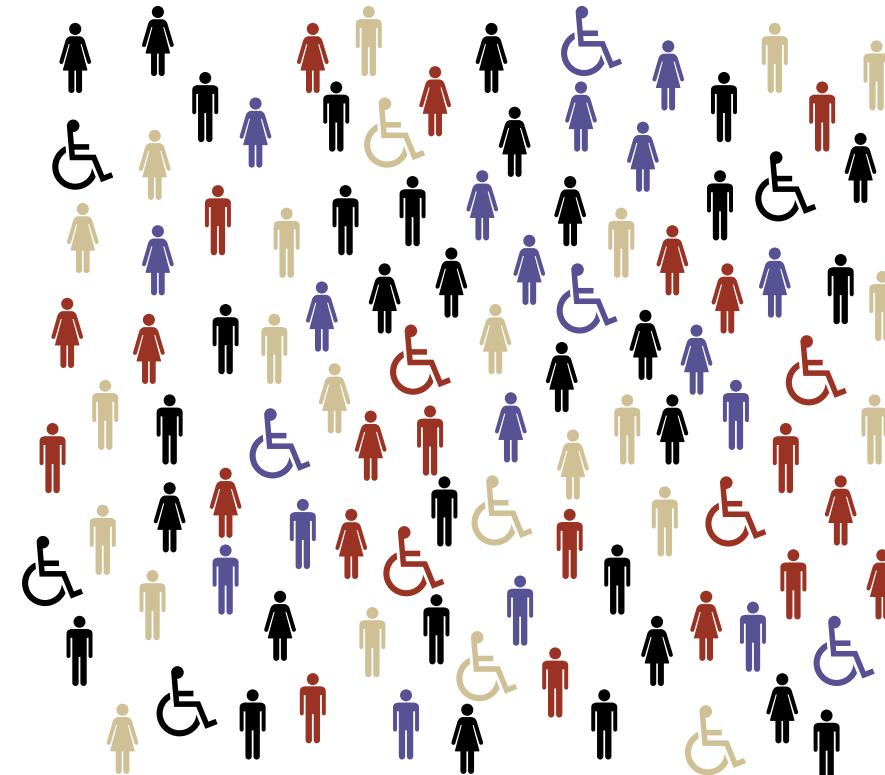


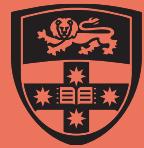


Data Sources
Annotation
Crowdsourcing
Workshop Preview

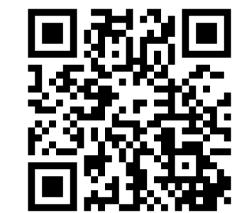


[menti.com 1552 7067](https://menti.com/15527067)

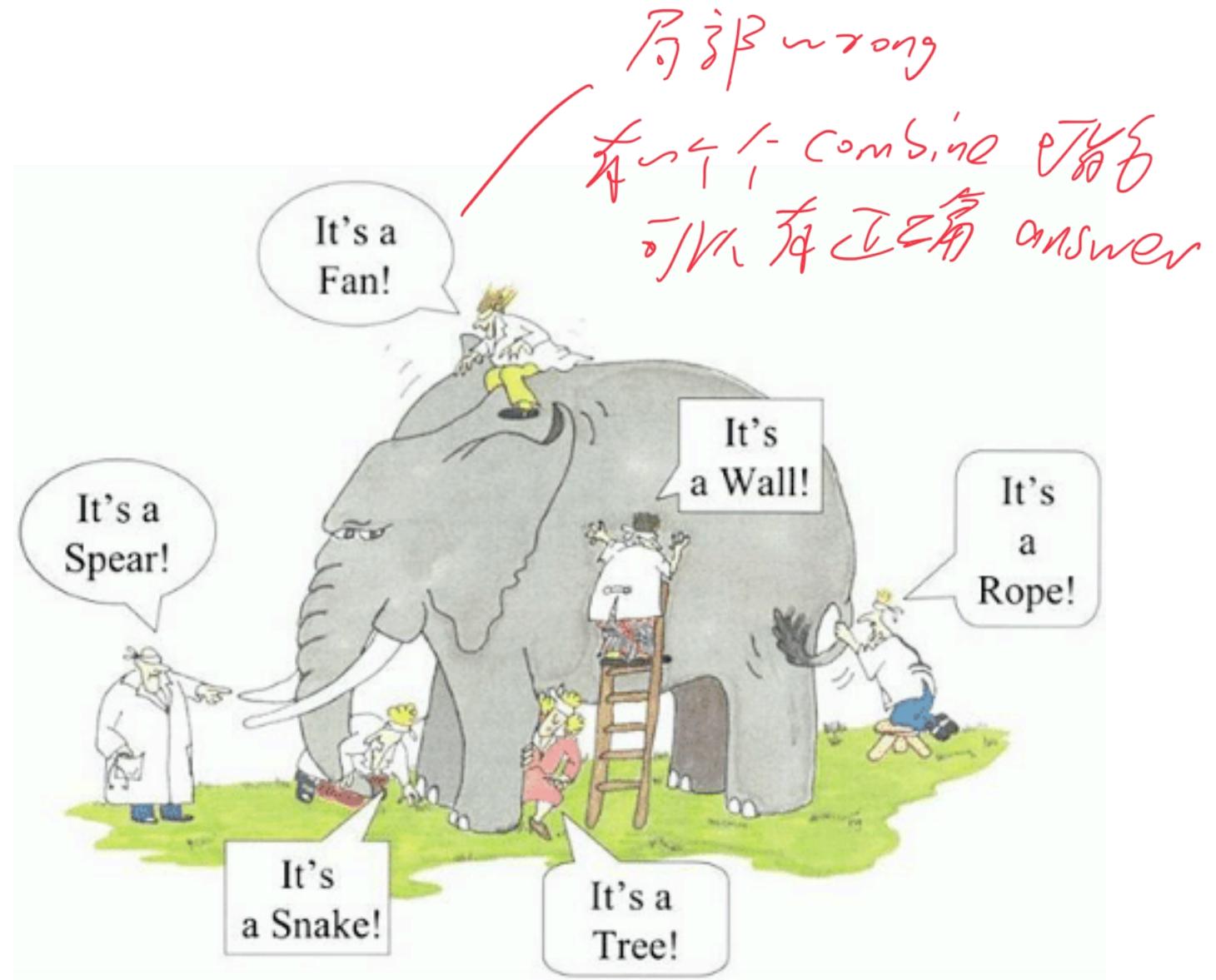


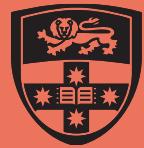


Data Sources
Annotation
Crowdsourcing
Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)





Data Sources
Annotation
Crowdsourcing
Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

800 Guesses at the 1906 country fair in Plymouth

Median: 1207 lb

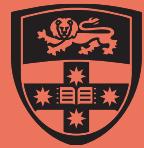
Mean: 1197 lb

True: 1198 lb

Combined
Answer
works



PEC Large Livestock Scales, Amazon



Data Sources
Annotation
Crowdsourcing
Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

Paid, small tasks

Choose the best category for this image [View Instructions↓](#)

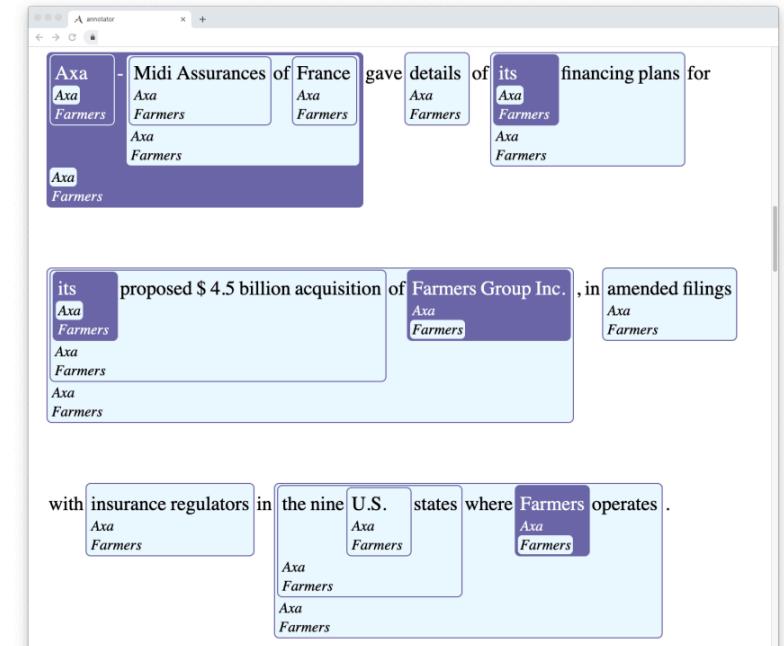


Select the room location in home for this picture. Seating areas outside are outside not living. Offices or dens are living not bedrooms. Bedrooms should contain a bed in the picture.

- kitchen
- living
- bath
- bed
- outside

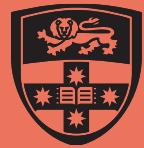
You must ACCEPT the HIT before you can submit the results.

Microwork



amazon
mechanical turk





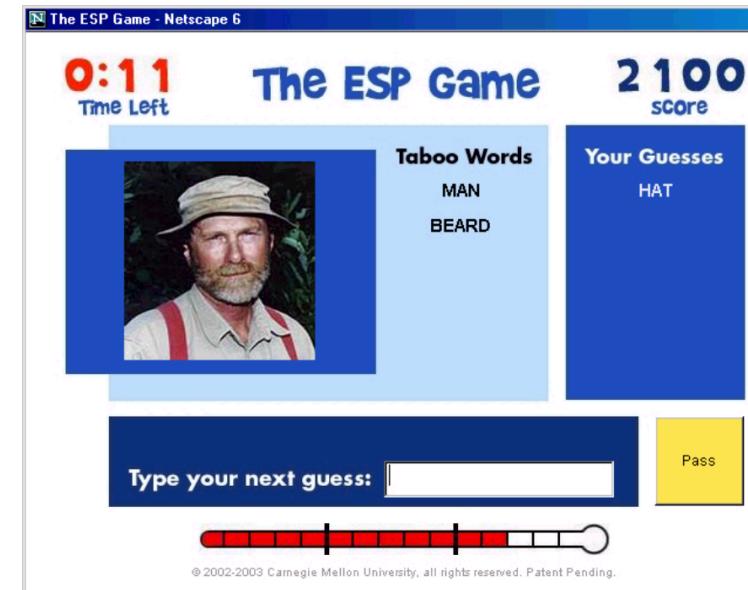
Data Sources
Annotation
Crowdsourcing
Workshop Preview

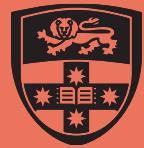


[menti.com 1552 7067](https://menti.com/15527067)

Games with a Purpose

Unpaid, make task fun





Data Sources

Annotation

Crowdsourcing

Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

Games with a Purpose

Competition with prize money

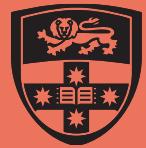
PHRASE•••DETECTIVES

Sherlink Holmes went to **the shop**.
He couldn't believe **it** was closed.

If you think that the word in orange, **It**, refers to the same object as the phrase in blue, **the shop**, you should say that you agree with this decision.

[Further instructions](#)





Data Sources

Annotation

Crowdsourcing

Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

Citizen Science / Collaboration

Unpaid, maybe fun? Satisfying



WIKIPEDIA
The Free Encyclopedia



Data Sources

Annotation

Crowdsourcing

Workshop Preview



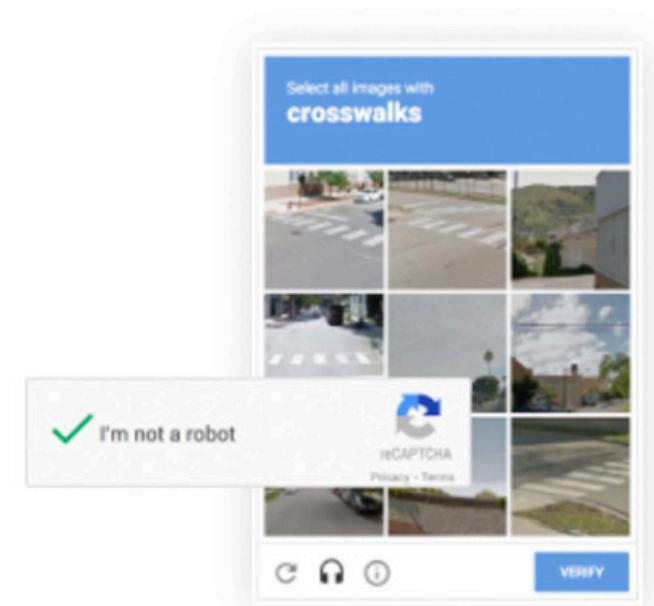
[menti.com 1552 7067](https://menti.com/15527067)

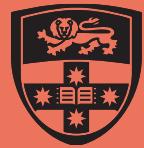
Implicit Work

Unpaid, not fun,
provides a benefit



Only know ^{one} real label
One is for check
One is for collect data





Data Sources
Annotation
Crowdsourcing
Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

Paid, small tasks

Choose the best category for this image [View Instructions↓](#)

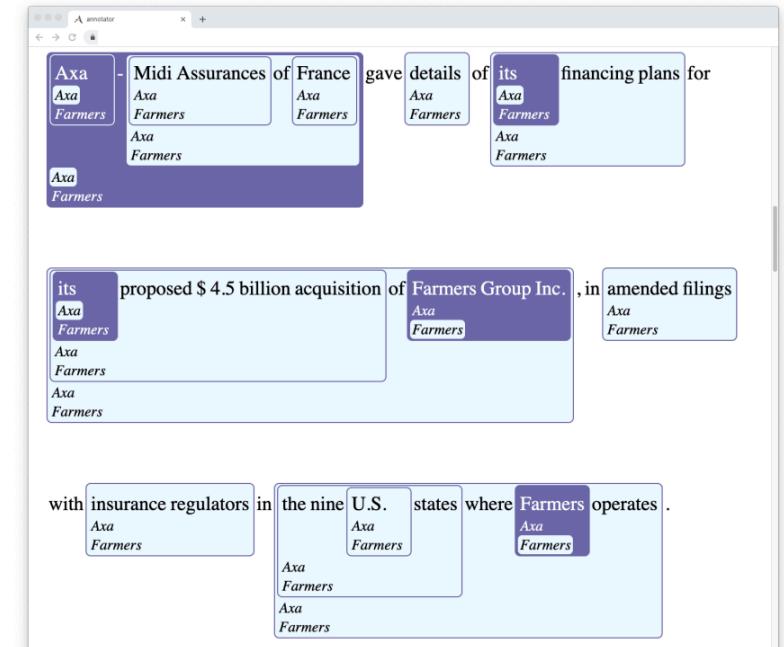


Select the room location in home for this picture. Seating areas outside are outside not living. Offices or dens are living not bedrooms. Bedrooms should contain a bed in the picture.

- kitchen
- living
- bath
- bed
- outside

You must ACCEPT the HIT before you can submit the results.

Microwork



amazon
mechanical turk





COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

Workshop Preview



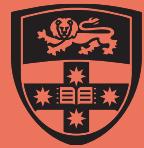
COMP 4446 / 5046
Lecture 10, 2025

Data Sources
Annotation
Crowdsourcing
Workshop Preview



[menti.com 1552 7067](https://menti.com/15527067)

Work on Assignment 4
+
Ask questions



Data Sources
Annotation
Crowdsourcing
Workshop Preview



menti.com 1552 7067

Muddy Card

Open shortly, closes at 7:05pm

[https://saipll.shinyapps.io/
student-interface/](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