ARDUOUS - Developing a data annotation protocol

Ubicomp-ISWC2023

Annotation
ARDUOUS

Who are we?

Prof. Dr. Kristina Yordanova
Head of the Institute for Data Science, University of Greifswald, Germany.
Research areas: human behaviour recognition and modelling,
activity recognition,
knowledge elicitation.
One of the initiators of the ARDUOUS workshop series

Dr. Emma Tonkin
University of Bristol
Expert in Digital Humanities and Data Science for Digital Health.
Part of the SPHERE project (Sensor Platform for HealthcarE in a Residential Environment) - working on data management and analysis, data reproducibility and data ethics

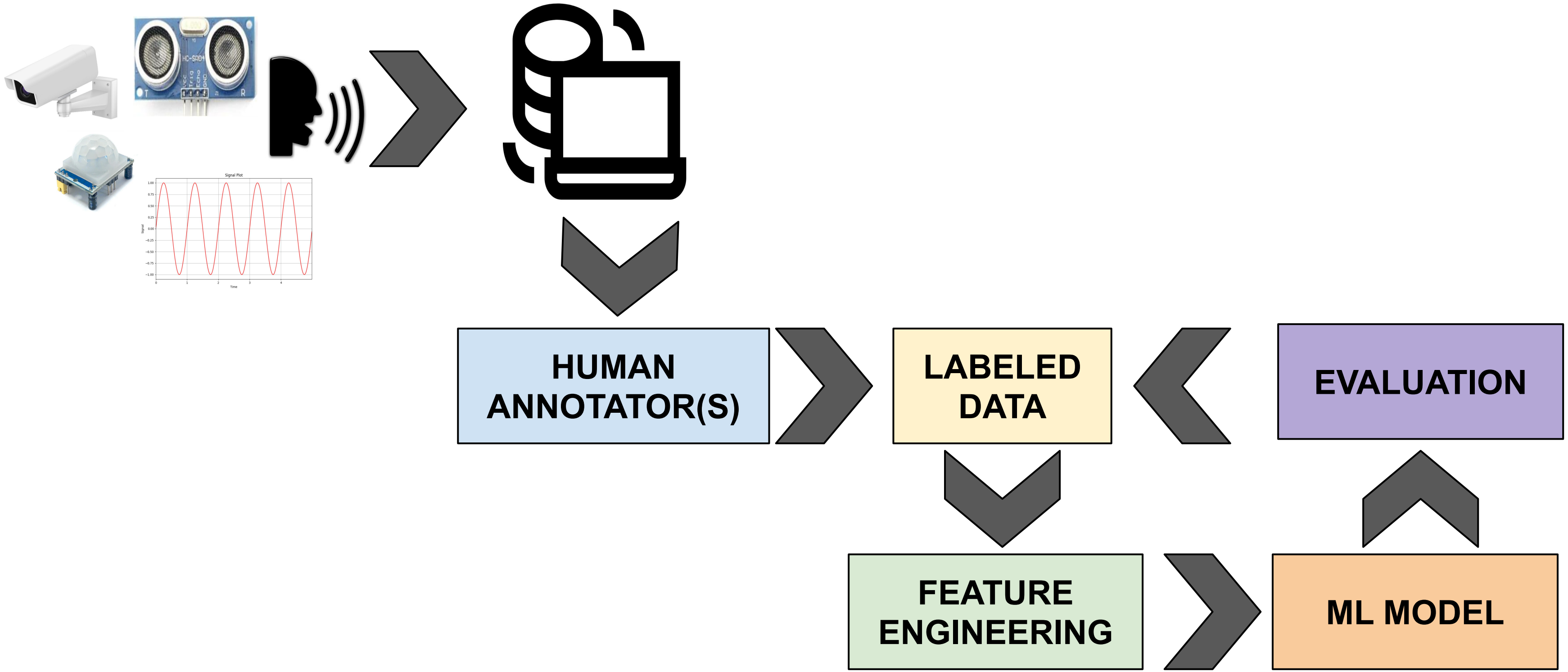
Dr. Gregory Tourte
University of Bristol
Expert in research data management.
Research in Deep time climate modellingmwithin the Bristol Research Initiative for the Dynamic Global Environment
(BRIDGE).

Msc. Teodor Stoev
University of Greifswald, Institute for Data Science
PhD candidate. Research areas: knowledge extraction from heterogeneous data sources, activity recognition and error detection, data annotation.

What is ARDUOUS?

- stands for Annotation of useR Data for UbiquitOUs Systems
- a series of workshops (beginning in 2017)
- topics connected with data annotation in the domain of pervasive & ubiquitous systems (but not limited to these domains)
- included paper sessions, demonstration sessions, short tutorials, keynotes, collaborative tasks

Motivation for ARDUOUS



HUMAN ANNOTATOR(S)

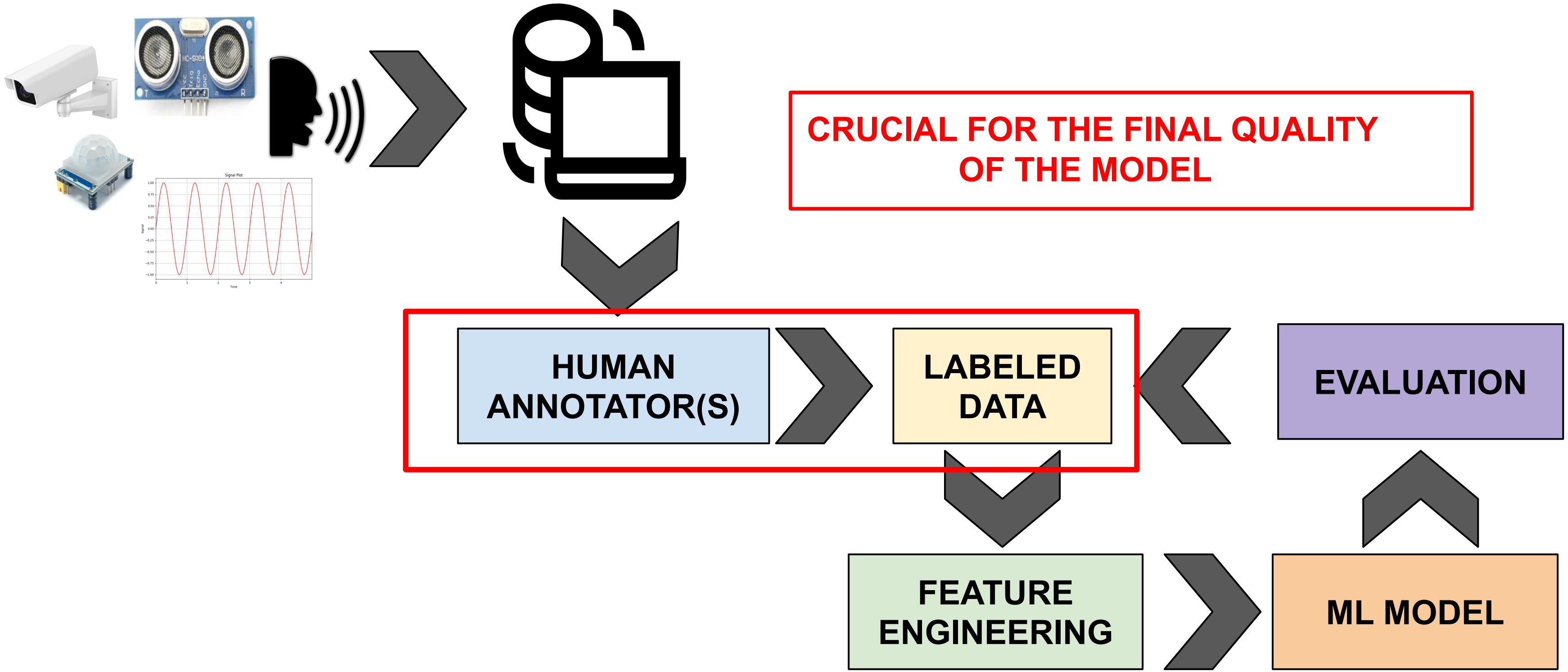
LABELED DATA

EVALUATION

FEATURE ENGINEERING

ML MODEL

Motivation for ARDUOUS



CRUCIAL FOR THE FINAL QUALITY OF THE MODEL

HUMAN ANNOTATOR(S)

LABELED DATA

EVALUATION

FEATURE ENGINEERING

ML MODEL

Motivation for ARDUOUS

- The development of datasets and information about the annotation process are usually not described in detail
- most authors focus on the usage of the data
- detailed information about annotators is usually not provided
- data annotation is crucial for the final outcome of the application
- annotated data is used for proper evaluation

What is data annotation ?

Definition of data annotation

The word "annotation" means (source Cambridge Online Dictionary):

1) “a short explanation or note added to a text or image, or the act of adding short explanations or notes"
2) "a description or piece of information added to data, for example a label saying whether a word is a noun, a verb, etc., or the act of adding this"

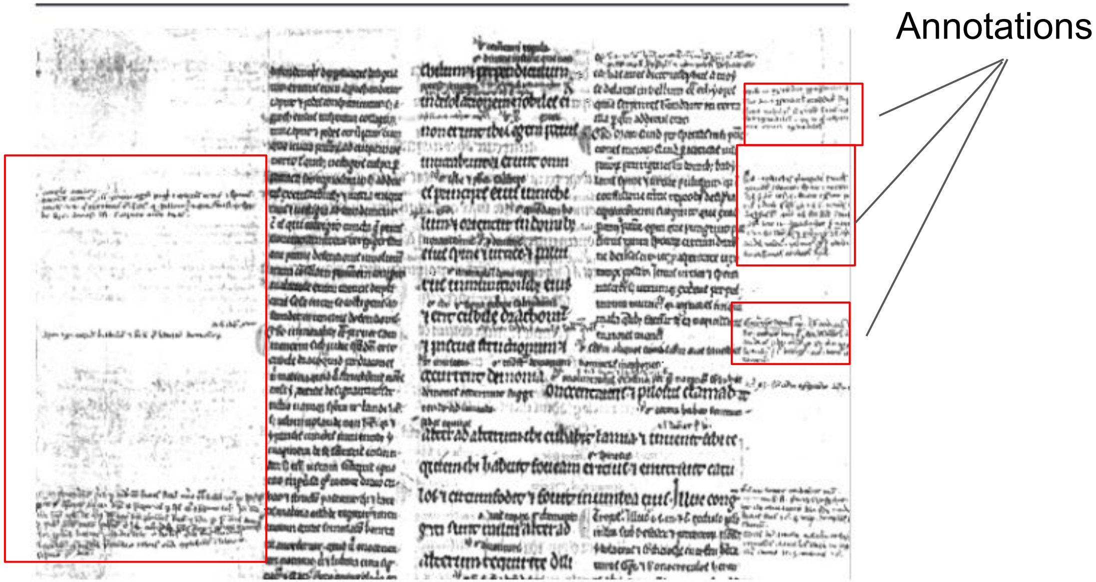
…. but also…

"*the sense-making practice of a given dataset, where annotators assign meaning to data using (pre-defined) labels.*" [1]

‘the goal of compiling data as the basis for ML approaches and automation.’ [2]

[1] Wang, D., Prabhat, S. and Sambasivan, N. ‘Whose AI Dream? In search of the aspiration in data annotation.’ In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. 2022, pp. 1–16.
[2] Pagel, J., Reiter, N., Rösiger, I. and Schulz, S. ‘Annotation als flexibel einsetzbare Methode’. In: Reflektierte Algorithmische Textanalyse. Interdisziplinäre (s) Arbeiten in der CRETA-Werkstatt (2020), pp. 125–141.

Definition of data annotation



Annotations

Bible with commentary (ca. 850-1499)[1]

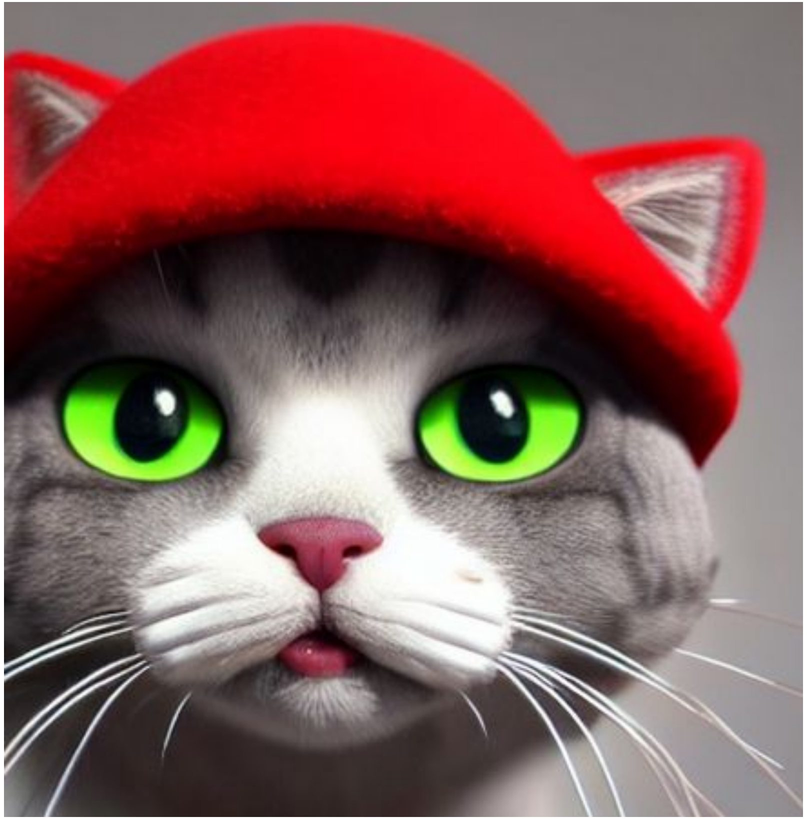
The practice of annotation has immense historical significance. During the medieval period, readers frequently used spaces in manuscripts to annotate and discuss the text, which often transcribed with the primary text. Annotations enabled readers to discuss, critique, and learn from the annotations left behind by earlier readers.

Such annotations are still being used in the domain of Digital Humanities

[1]Philip Bliss collection of medieval manuscript fragments. Courtesy of Department of Special Collections, Stanford University Libraries

Data annotation vs data labelling

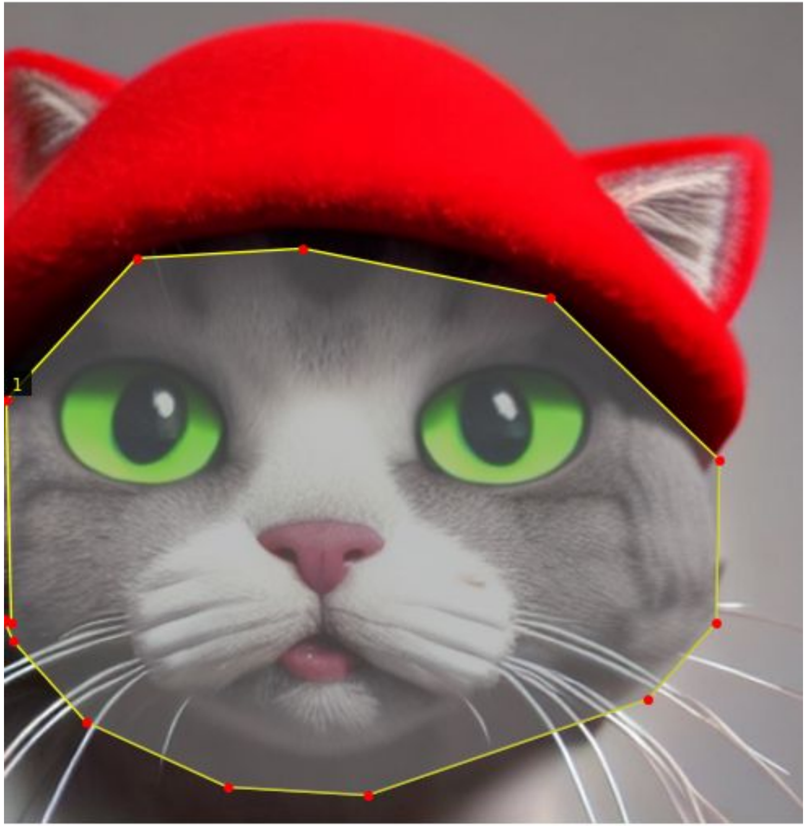
- Often annotation and labels are used interchangeably
- there is no broadly accepted definition, but:
    - both share the common goal of adding metadata to raw data
    - annotations are often considered to include more written comments or marginal notes
    - labels are usually connected with visible features of data which are considered to be concrete



Generated with deepai.org

Label or annotation?

“Cat wearing a red hat”



Label or annotation?

“cat_face”

VIA annotation tool:https://www.robots.ox.ac.uk/~vgg/software/via/via_demo.html

Common ways of annotation?

- In terms of “how“ we annotate data here are several approaches:

- paper-based approaches: common in human sciences (leaving comments alongside texts, interpretations, etc.)
- computer-based approaches: using a mobile device or a computer to annotate datasets (a lot of tools exist)
- voice-based annotation (for instance, while gathering data of an experiment we can leave comments in the form of recordings)

Many software solutions exist for different types of data (we will see some examples later in the tutorial)

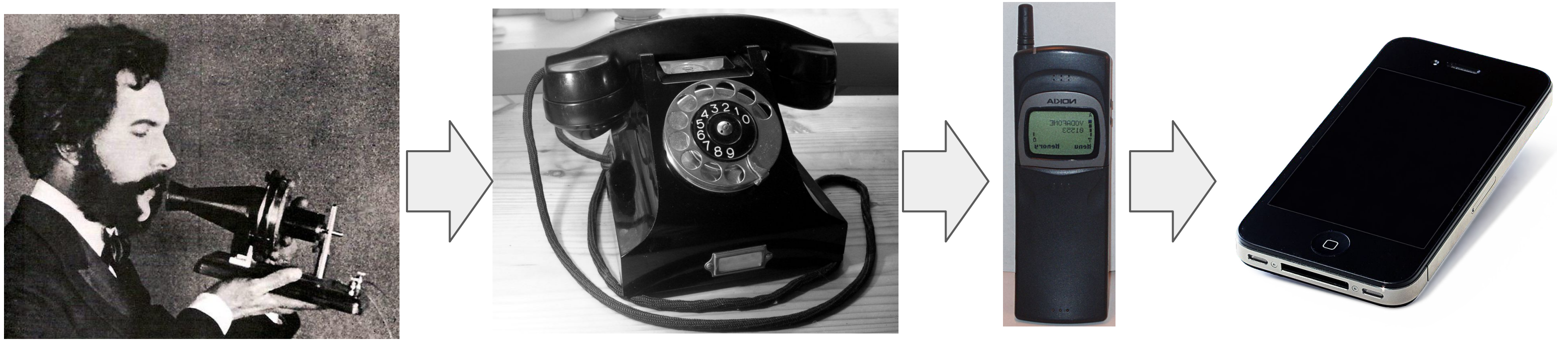
Ground Truth in annotation

The concept of ground truth

- The concept of “ground truth” is often used in the context of annotation
- exact origins of the term are unknown (most probably from the German word *Grundwahrheit* (fundamental truth, first attested in the mid-17th))
- may be understood as a concept relative to the knowledge of the truth concerning a specific question
- the term is aspirational: “the best data available”
- in some domains it is addressed as “gold standard” , or <u>*the best available approximation of the true state which can be used for benchmarking*</u>

Semantic shift, drift and change in ground truth labels

- semantic drift : gradual change in the meaning which may be due to the change in the environment of use of a term



semantic drift of the word “telephone”

Semantic shift, drift and change in labels

- semantic drift : gradual change in the meaning which may be due to the change in the environment of use of a term

semantic drift of the word “web”

Semantic shift, drift and change in labels

- there are also more situations in which labels can change:

  examples:

  - change of the final application and the granularity (e.g. activity recognition)

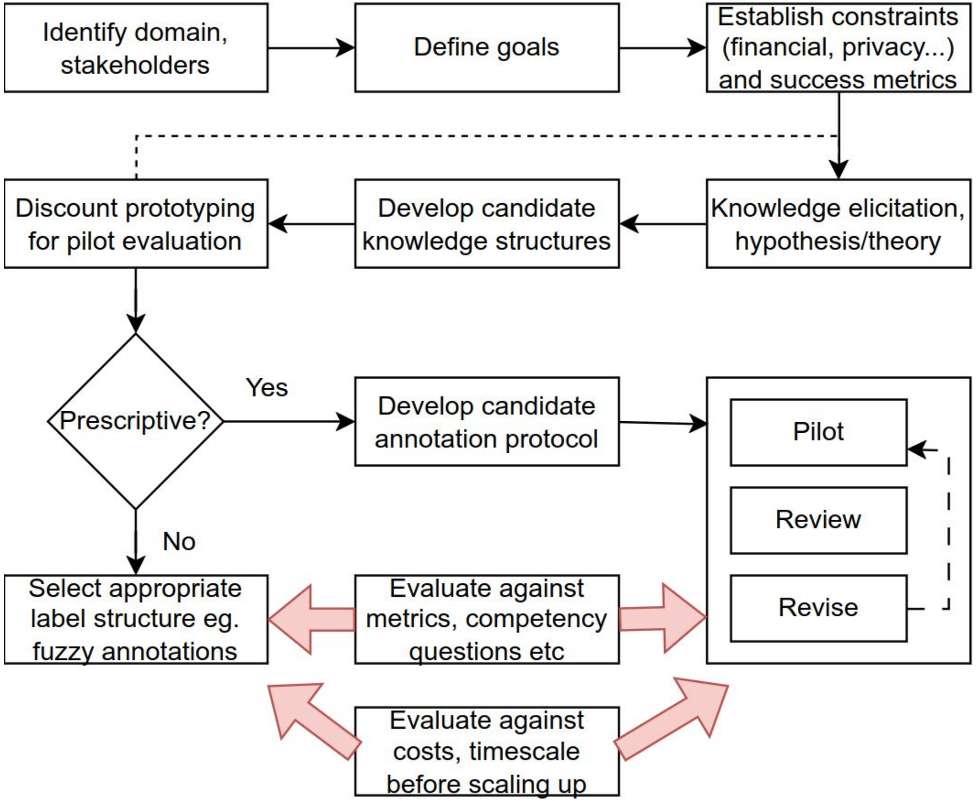
  label: *“take_cup”* to *“take_cup_right”* (“right” indicating hand)

  - labels were based on a concept which changed

  - annotators’ experience with other annotation tasks

Methodology for data annotation protocol

Overall process workflow



Identify domain, stakeholders
Define goals
Establish constraints (financial, privacy...) and success metrics
Knowledge elicitation, hypothesis/theory
Develop candidate knowledge structures
Discount prototyping for pilot evaluation
Prescriptive?
Yes
No
Develop candidate annotation protocol
Pilot
Review
Revise
Select appropriate label structure eg. fuzzy annotations
Evaluate against metrics, competency questions etc
Evaluate against costs, timescale before scaling up

Ground truthing methods and approaches

- How do we get ground truth?

  common ways: domain experts, crowdsourcing, observation-based approaches, automated methods (e.g. using state-of-the art classifiers)

- Who is involved in the process of collection of ground truth?

  domain experts, ML engineers, users ….

- How to elicit (gather) expert knowledge in order to define ground-truth?

Knowledge elicitation

- What is Knowledge elicitation

  "Set of techniques and methods that attempt to elicit the knowledge of a domain expert" (Shadbolt et al. 2015)

  Main idea:

  The idea is that the knowledge elicitor and domain expert work together on the creation of a model which encodes expert's knowledge. This model may represent reality to a varying degree.

  Knowledge elicitation is a subprocess of *knowledge acquisition*

  Foundational assumption: generally, knowledge is held by individuals (for example, in institutions it may be held by employees, volunteers and users)

Shadbolt, Nigel R., et al. "Knowledge elicitation." *Evaluation of human work* (2015): 163-200.

Knowledge Elicitation Techniques - Overview

Two main categories:

- **natural**:

  interviews (structured, semi-structured, unstructured)

  structured: based on a set of pre-arranged prompts provided by the elicitor
  semi-structured: is not restricted to a set of questions, questions might divert during the conversation, more informal
  unstructured: no predefined questions, interviewers might ask questions which come on their mind on the spot

  protocol analysis: a form of “think aloud study” - expert solves a problem in front of the elicitor while explaining the steps and their thought process aloud, elicitor takes notes or records the steps.

Knowledge Elicitation Techniques - Overview

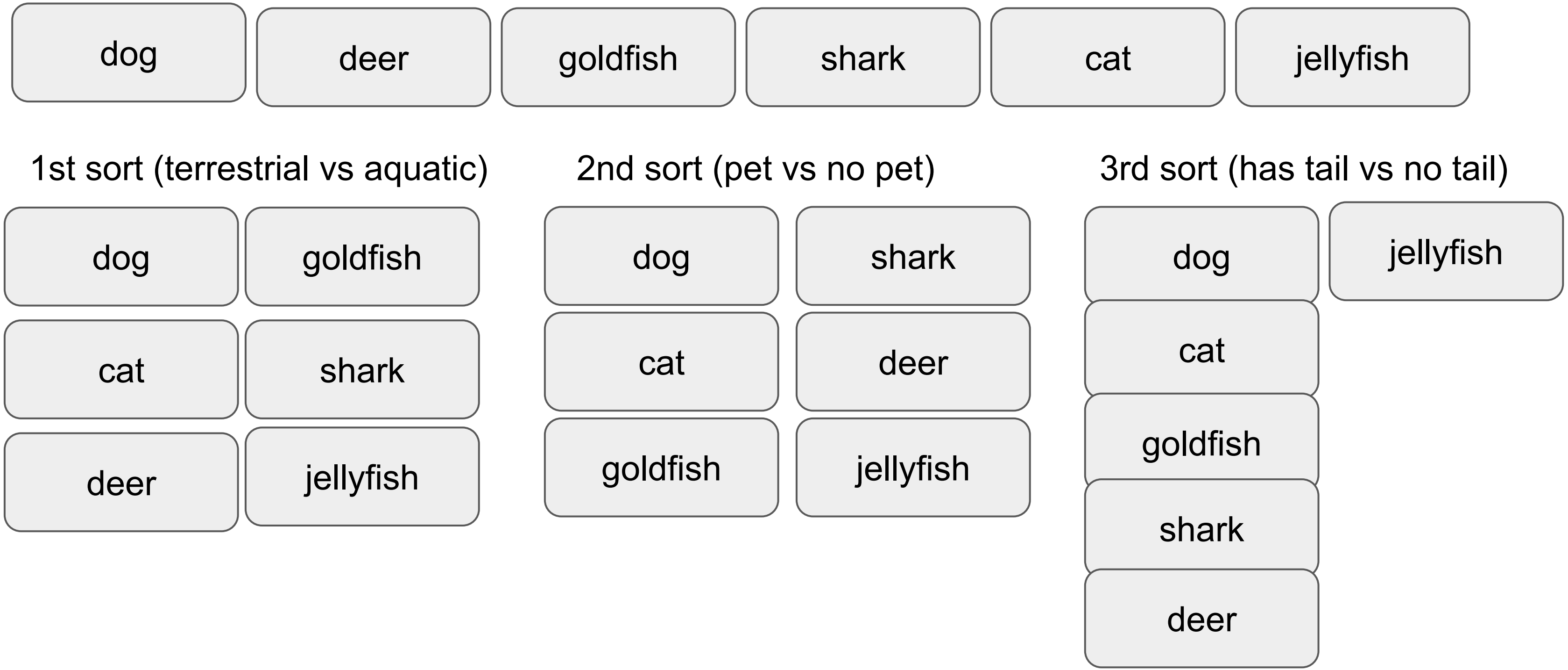
- **contrived (non-natural)** - experts are asked to take part in tasks that they would not ordinarily undertake.
  Two of them are:

  card sorting - often used to elicit categories and relationships. The expert is given cards with concepts and then the expert has to repeatedly sort them in different categories and label the resulted categories. (example on the next slide)

  grid laddering - the interview begins with a seeding term (foundational concept). The participant is then asked to elicit examples of the class
  (“Provide an example of <seeding term> “),
  or to navigate the hierarchy upwards (“What are concept A and B examples of ?”), or across (“what alternative examples are there of this class?”). The expert might also be asked to identify differences (“What is the difference between A and B” ?)
  Result is structured hierarchy.

Knowledge Elicitation Techniques - contrived methods

Example: card sorting (concept sorting) of animals



dog | deer | goldfish | shark | cat | jellyfish

1st sort (terrestrial vs aquatic)

dog	goldfish
cat	shark
deer	jellyfish

2nd sort (pet vs no pet)

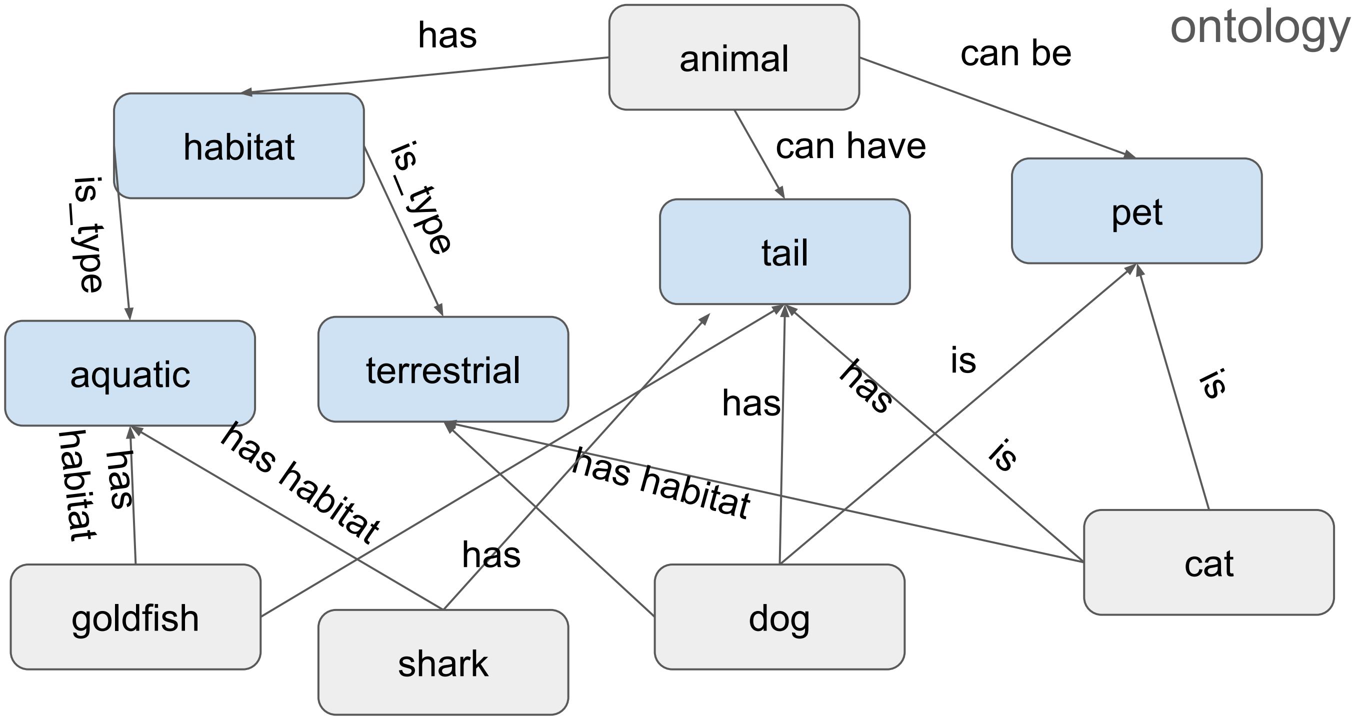
dog	shark
cat	deer
goldfish	jellyfish

3rd sort (has tail vs no tail)

dog	jellyfish
cat	
goldfish	
shark	
deer	

Knowledge Elicitation Techniques - contrived methods

Example: using the resulted piles we can build for instance a knowledge graph* or ontology



*some animals were skipped on purpose (no space)

Identifying an appropriate strategy

Dependent on stakeholders

So, begin with a stakeholder analysis: identify *creators*, *contributors* and *candidate users* of the information.

Subject matter experts are often difficult to contact due to time, finance, resource or organisational constraints (for example: clinical staff, judiciary) and may or may not be inclined to share knowledge*

Consider: Are there any moral, ethical, legal or privacy concerns surrounding the task?

In general: A common starting point is the interview technique

Prototyping methods (e.g. card sort) can be helpful in evaluating knowledge gained at an earlier stage**

* Gavrilova, T. and Andreeva, T. (2012), "Knowledge elicitation techniques in a knowledge management context", *Journal of Knowledge Management*, Vol. 16 No. 4, pp. 523-537. https://doi.org/10.1108/13673271211246112
** Jones, S.R. and Miles, J.C. (1998), The Use of a Prototype System for Evaluating Knowledge Elicitation Techniques. Expert Systems, 15: 83-97. https://doi.org/10.1111/1468-0394.00067

Define the annotation domain

Annotations are typically designed to suit a particular use case, but describe one or more **domains** (‘a realm of the world’)

For example, annotations describing sporting activity might include information about the sport, and require information about **location**, **time** and **causality**

Therefore, as with ontology development more broadly, it is useful to consider whether knowledge structures exist that might be reused

Annotation scope and granularity

Annotation tasks can be lengthy: setting and testing that an appropriate **scope** has been selected and that the information is described to an appropriate **granularity** is helpful in minimising annotation timescale

Example scenario: *a group of annotators may be tempted to annotate every piece of information that they identify in the source material, describing every detail, however slight, in the belief that it may be useful. However, this is time-consuming and the annotators eventually realise that the resulting labels contain more detail than the use case requires.*

Annotation scheme

- Annotation scheme provides information about how the annotations will be represented
- the scheme can be expressed as
  labels (vocabulary) can be controlled (pre-defined set) or uncontrolled (annotators can add labels)
  a **taxonomy** (hierarchical label set), **ontology** (if relationships are included) and **thesaurus** (if some semantic similarities are encoded)
- Example 1: semantic representation in the domain of activity recognition

  labels encoded as string type with information about actions and objects being manipulated: *“take-cup”, “put-cup”, “wash-pot”*

- Example 2: encoding named entities (IOB-format): “Artur is brave”

Artur	I-PER
is	O
brave	O
.	O

I-PER indicates the beginning of an entity “Person”
O indicates “no entity”

Exercise: building a vocabulary

See recipe_task.txt

Labels & Agreement: Inter- & Intra-annotator agreement

- Question: What about annotators’ agreement?
- **Inter-annotator agreement:** how similar are the annotations of different annotators?
- **Intra-annotator agreement:** how similar are the repeated annotations of a single annotator?
- for both “types” of agreements we can use the same metrics for evaluation
- in general, chance-corrected agreement coefficient are preferred

Should annotators agree?

Annotators may disagree for a wide variety of reasons. A few examples below:

- the guidelines are subject to interpretation/unclear, and the annotators have varied understanding of them - **resolvable through iterative discussion and review**
- the problem itself is subject to interpretation, and the annotators disagree with one another for valid reasons - **indicates that the task is difficult, for example, responses vary for cultural or linguistic reasons**
- the source material contains inadequate information to be sure: for example, an image/video is blurry, unclear or the subject is occluded
- annotators are learning as they go - **annotation performance is likely to improve over time**

Studying **annotator subjectivity** may be part of the purpose of the study (descriptive vs prescriptive annotator paradigm*).

* Röttger, P., Vidgen, B., Hovy, D. and Pierrehumbert, J. B. ‘Two contrasting data annotation paradigms for subjective NLP tasks’. In: arXiv preprint arXiv:2112.07475 (2021)

How should we respond to disagreement between annotators?

Options:

- return to, review and rework inconsistent annotations:
  - achieve consensus, improve and harmonise annotator mental models of task
- avoid inconsistent annotators, where appropriate (e.g. MTurk…)
- note that inconsistent labels **may still be reliable,** in the sense that the disagreement identified is an accurate reflection of disagreement between individual annotators’ observations! In some circumstances, labels with low agreement between annotators may be usable as-is.

How do we measure agreement between annotators?

Several common metrics exist:

Percent agreement:

$$\text{Percent agreement} = \frac{\text{number of concordant responses}}{\text{total number of responses}} \times 100$$

Cohen's Kappa: measure of agreement between two annotators which takes into consideration expected agreement resulting by chance. To calculate kappa, we need to calculate the proportion observed agreement (Po) and the proportion expected agreement by chance (Pe)

The formula to calculate Cohen's Kappa $k = \frac{Po - Pe}{1 - Pe}$

The result varies between -1 and 1 (1 being perfect agreement, values below 0 no agreement)

Cohen’s Kappa example

Assume we have the following annotations:

Annot 1 = [C, C, C, D, D, D, D, D]
Annot 2 = [C, C, D, C, C, D, D, D]

Cohen’s Kappa example

Assume we have the following annotations:

Annot 1 = [C, C, C, D, D, D, D, D]
Annot 2 = [C, C, D, C, C, D, D, D]

		Annot 1	
		C	D
Annot 2	C	2	2
	D	1	3

Cohen’s Kappa example

Assume we have the following annotations:

Annot 1 = [C, C, C, D, D, D, D, D]
Annot 2 = [C, C, D, C, C, D, D, D]

		Annot 1	
		C	D
Annot 2	C	2	2
	D	1	3

Observed agreement (diagonal of agreement table) = 5 cases
Proportion observed agreement = 5 / 8 = **0.625**

1) Expected agreement for class C = P(Annot 1 = C) x P(Anot 2 = C)
P(Annot 1 = C) = 3/8 (3 cases out of 8 where Annot 1 put label C)
P(Annot 2 = C) = 4/8 (4 cases out of 8 where Annot 2 put label C)
P (Annot 1 = C) x P(Annot 2 = C) = 3/8 x 4/8 = 12/64 = 0.1875

Cohen’s Kappa example

Assume we have the following annotations:

Annot 1 = [C, C, C, D, D, D, D, D]
Annot 2 = [C, C, D, C, C, D, D, D]

		Annot 1	
		C	D
Annot 2	C	2	2
	D	1	3

Observed agreement (diagonal of agreement table) = 5 cases
Proportion observed agreement = 5 / 8 = **0.625**

1) Expected agreement for class C = P(Annot 1 = C) x P(Anot 2 = C)
P(Annot 1 = C) = 3/8 (3 cases out of 8 where Annot 1 put label C)
P(Annot 2 = C) = 4/8 (4 cases out of 8 where Annot 2 put label C)
P (Annot 1 = C) x P(Annot 2 = C) = 3/8 x 4/8 = 12/64 = 0.1875

2) Expected agreement for class D = P(Annot 1 = D) x P(Anot 2 = D)
P(Annot 1 = D) = 5/8 (5 cases out of 8 where Annot 1 put label D)
P(Annot 2 = D) = 4/8 (4 cases out of 8 where Annot 2 put label D)
P (Annot 1 = D) x P (Annot 2 = D) = 5/8 x 4/8 = 20/64 = 0.3125

Cohen’s Kappa example

Assume we have the following annotations:

Annot 1 = [C, C, C, D, D, D, D, D]
Annot 2 = [C, C, D, C, C, D, D, D]

		Annot 1	
		C	D
Annot 2	C	2	2
	D	1	3

From 1) and 2) :
P (Annot 1 = C) x P(Annot 2 = C) = 3/8 x 4/8 = 12/64 = 0.1875
P (Annot 1 = D) x P (Annot 2 = D) = 5/8 x 4/8 = 20/64 = 0.3125

3) **Expected chance agreement: P (Annot 1 = C) x P(Annot 2 = C) + P (Annot 1 = D) x P (Annot 2 = D) = 0.1875 + 0.3125 = 0.5**

Cohen’s Kappa k = (0.625 - 0.5) / (1 - 0.5) = 0.125 / 0.5 = 0.25

Interpreting Kappa coefficients

Different interpretations exist across the literature

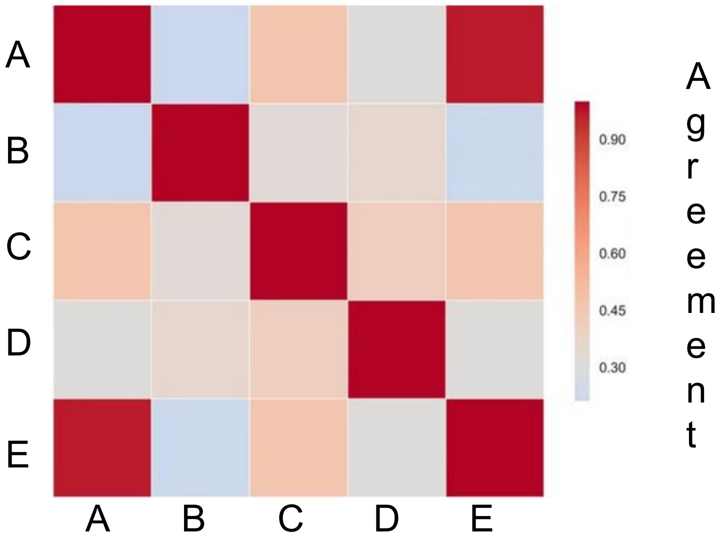
Landis & Koch (1997)	
0.81 - 1.00	almost perfect
0.61 - 0.80	substantial
0.41 - 0.60	moderate
0.21 - 0.40	fair
0.00 - 0.20	slight
<0.00	poor

Altman (1991)	
0.81 - 1.00	very good
0.61 - 0.80	good
0.41 - 0.60	moderate
0.21 - 0.40	fair
0.00 - 0.20	poor

Landis JRKoch, G. G. "The measurement of observer agreement for categorical data." *Biometrics* 33.1 (1977): 159174
Altman DG. Practical Statistics for Medical Research. London: Chapman and Hall; 1991:404–408.

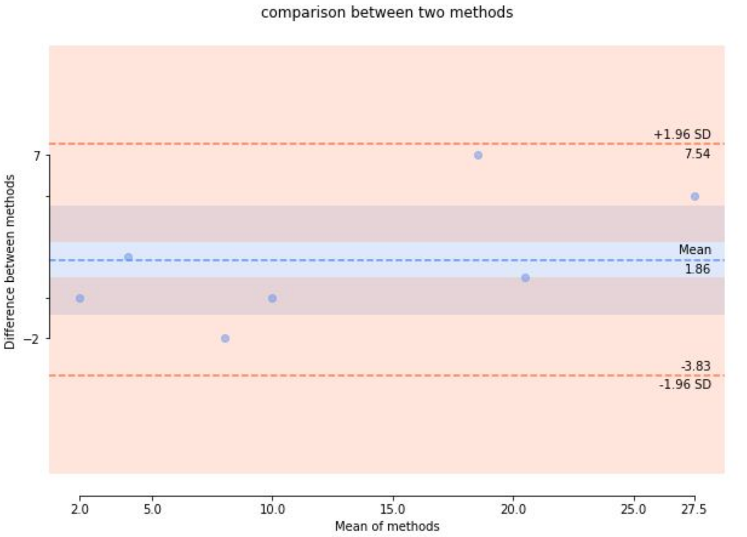
Visualizing agreement

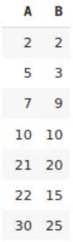
1) Using simple heatmaps (given any agreement metric)



Visualizing agreement

2) If we have annotations which are encoded as numerical values,
we could apply a **Bland–Altman plot** visualizing the difference between two annotations.
- The plot shows the difference between the measurements for instance (A and B) on the y-axis (or A - B) and the average of the measurements on the x-axis.



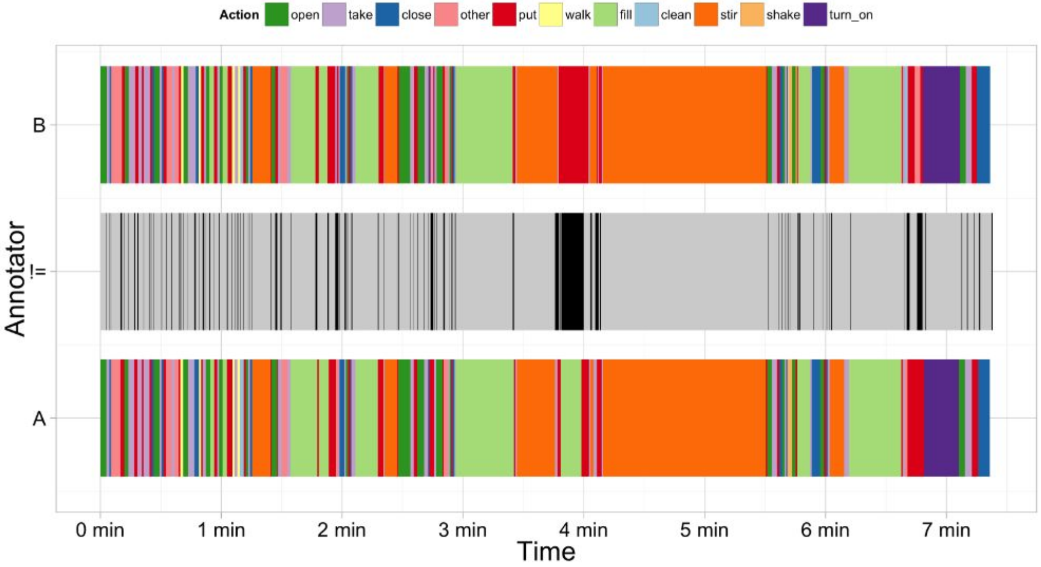


A	B
2	2
5	3
7	9
10	10
21	20
22	15
30	25

Visualizing agreement

2) Visualizing labels in time

if we have a sequential annotation (such as a video log), we could apply visualization of the labels and the differences between annotators at each time step



Validating causal correctness of annotation

- In some cases (such as annotation of activities) the domain contains “naturally” causal relationship. For instance, the action “take cup” causes the action “put cup”.
- How can we validate annotations using such causality?

manually: by checking directly the sequence of annotations one by one

automatically: by checking the correctness of the sequence given a domain model (expressed for example in first order logic)

I will demonstrate how to validate causal annotations using **AI planning**

Factors and biases in annotation tasks

Factors in annotation: Motivation

Main motivations for annotating

Fun

Especially when gamification strategies are applied

Profit

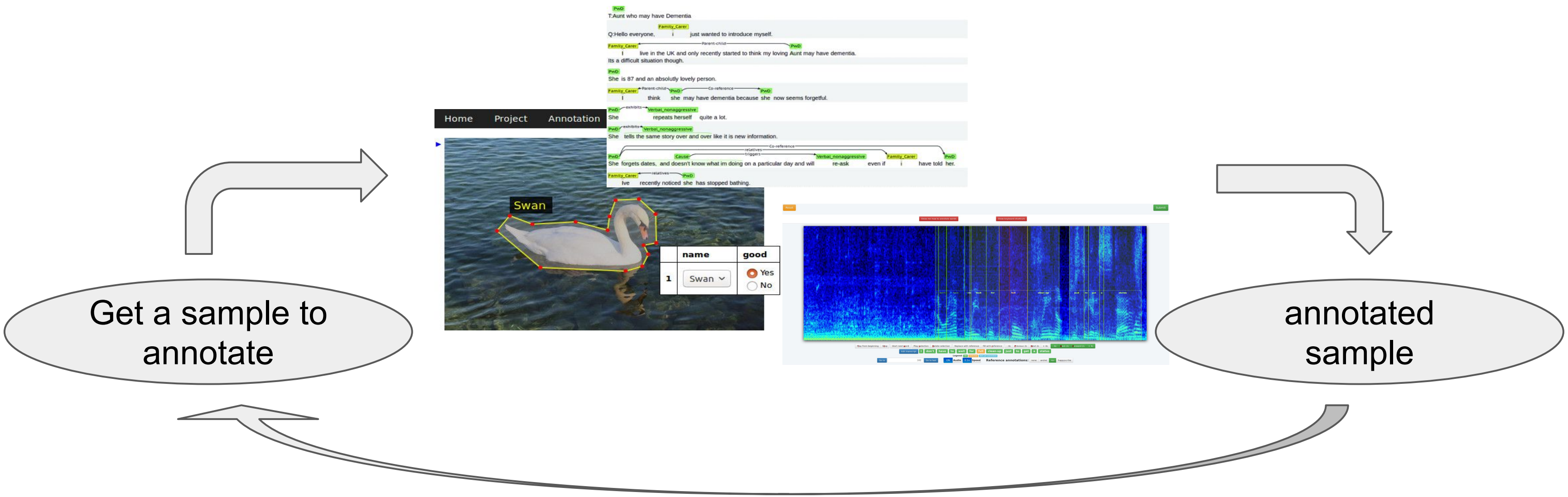
People get paid to annotate (e.g. Amazon Mechanical Turk)

Altruism

doing something good in the name of science/community/system

Annotation outcome can be affected by the motivation of the annotators

Factors in annotation: Repetitiveness



T:Aunt who may have Dementia

Q:Hello everyone, i just wanted to introduce myself.

I live in the UK and only recently started to think my loving Aunt may have dementia.

Its a difficult situation though.

She is 87 and an absolutly lovely person.

I think she may have dementia because she now seems forgetful.

She repeats herself quite a lot.

She tells the same story over and over like it is new information.

She forgets dates, and doesn't know what im doing on a particular day and will re-ask even if i have told her.

Ive recently noticed she has stopped bathing.

	name	good
1	Swan	● Yes ○ No

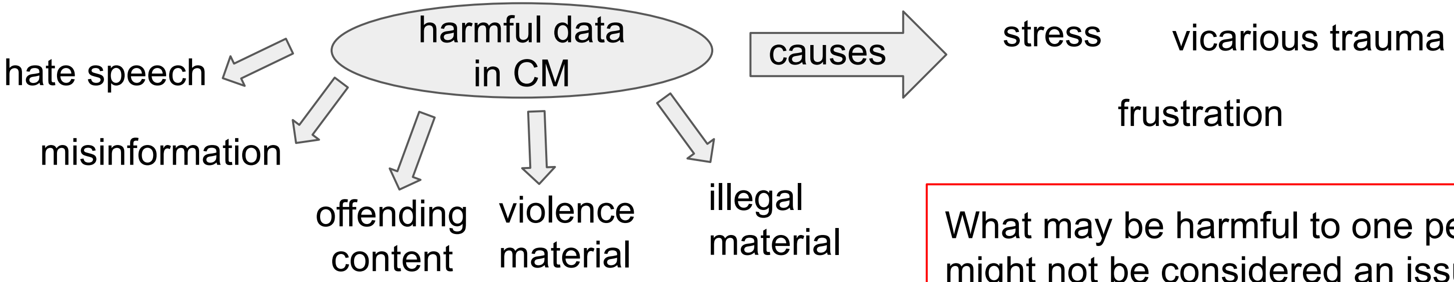
Even when using tools, usually we have the same workflow during the annotation process (although the samples differ)

The repetitive nature of tasks may cause tiredness, drowsiness, and reductions in adaptability and responsiveness

Factors in annotation: coping with harmful data

- **Harmful data**: material which may have harmful/negative effects on some readers/viewers in terms of emotional, psychological and physical well-being and safety (can still be legal !)

- Effects have been well studied in the area of Content Moderation (CM), or the process of detecting contributions that are irrelevant, obscene, illegal, harmful, or insulting with regards to useful or informative contributions.



What may be harmful to one person might not be considered an issue by someone else.

THE WALL STREET JOURNAL.

The Worst Job in Technology: Staring at Human Depravity to Keep It Off Facebook

Social-media giants hire legions of contractors to hunt for pornography, racism and violence in a torrent of posts and videos

By Lauren Weber & Deepa Seetharaman,Dec. 2017, THE WSJ

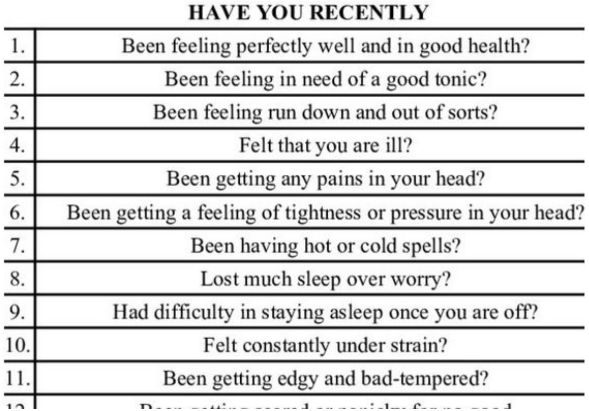
… but, on the other hand, people who do this job may be proud of making the webspace safer for others

Factors in annotation: coping with harmful data

- Assessment mechanisms to evaluate stress on workers (annotators) exist:

  Examples:

  The General Health Questionnaire (GHQ), a self-screening tool relating to psychiatric disorder;



HAVE YOU RECENTLY

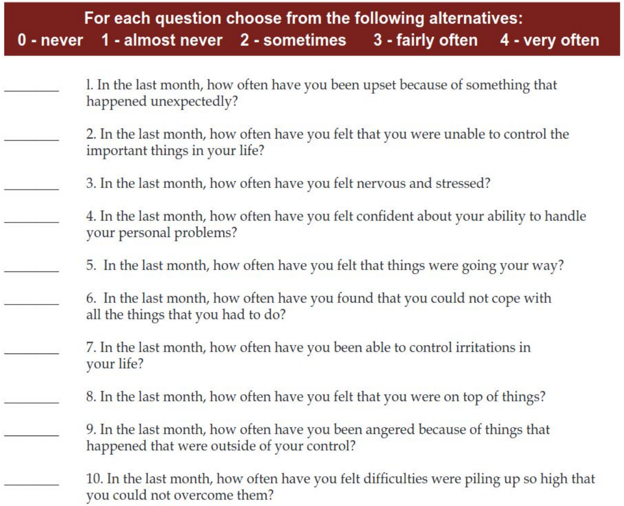
1.	Been feeling perfectly well and in good health?
2.	Been feeling in need of a good tonic?
3.	Been feeling run down and out of sorts?
4.	Felt that you are ill?
5.	Been getting any pains in your head?
6.	Been getting a feeling of tightness or pressure in your head?
7.	Been having hot or cold spells?
8.	Lost much sleep over worry?
9.	Had difficulty in staying asleep once you are off?
10.	Felt constantly under strain?
11.	Been getting edgy and bad-tempered?
12.	Been getting scared or panicky for no good

The first 11 questions of a scaled version of the GHQ (Goldberg and Hillier 1979).

The Perceived Stress Scale, a measurement of stress;
The Oldenburg Burnout Inventory (OLBI);
The Connor-Davidson Resilience Scale (CD-RISC), which measures resilience, against stress.

Factors in annotation: coping with harmful data

The Perceived Stress Scale (PSS), a measurement of stress;



For each question choose from the following alternatives:
0 - never 1 - almost never 2 - sometimes 3 - fairly often 4 - very often

_______ 1. In the last month, how often have you been upset because of something that happened unexpectedly?

_______ 2. In the last month, how often have you felt that you were unable to control the important things in your life?

_______ 3. In the last month, how often have you felt nervous and stressed?

_______ 4. In the last month, how often have you felt confident about your ability to handle your personal problems?

_______ 5. In the last month, how often have you felt that things were going your way?

_______ 6. In the last month, how often have you found that you could not cope with all the things that you had to do?

_______ 7. In the last month, how often have you been able to control irritations in your life?

_______ 8. In the last month, how often have you felt that you were on top of things?

_______ 9. In the last month, how often have you been angered because of things that happened that were outside of your control?

_______ 10. In the last month, how often have you felt difficulties were piling up so high that you could not overcome them?

The 10 questions of the PSS

Further tools:
The Oldenburg Burnout Inventory (OLBI) (link);
The Connor-Davidson Resilience Scale (CD-RISC) - measures resilience, against stress.(link)

Biases in annotation tasks: Biases in annotation instructions

Case study in the domain of Natural Language Understanding (NLU)
Parmar, Mihir, et al. "Don't Blame the Annotator: Bias Already Starts in the Annotation Instructions."

- studied 14 benchmark datasets (mainly question-answering ones)
- authors investigated reoccurring patterns in annotation instructions
- hypothesized that these patterns “might limit the imaginations of annotators while creating samples”
- demonstrated that the repeating patterns are a source of bias in the created samples
- this instruction bias led to inflated assessment of model performance

Biases in annotation tasks: Biases caused by social and demographic factors

Case studies:
Di Domenico et al.:
- domain: emotion recognition
- older people are faster at recognizing positive emotions than negative ones
- older people perceive both happy and angry expressions more intensive

Al Kuwatly et al. :
- aimed to investigate the influence of several demographic characteristics
- methodology: training classifiers on datasets annotated by different groups and testing significance of their performance difference
- results:
  gender: no significant difference, but inter-rater score of females significantly lower;
  first language: classifiers derived from annotations by naive speakers achieved higher scores
  age and education level: evidence that in the particular tasks personal attacks in comments are perceived differently

Di Domenico, Alberto, et al. "Aging and emotional expressions: is there a positivity bias during dynamic emotion recognition?." *Frontiers in psychology* 6 (2015): 1130.
Al Kuwatly, Hala, Maximilian Wich, and Georg Groh. "Identifying and measuring annotator bias based on annotators' demographic characteristics." *Proceedings of the fourth workshop on online abuse and harms*. 2020.

Biases in annotation tasks: Biases caused by pre-annotation

- In some cases a dataset can be pre-annotated, for instance by using weak classifiers on the task at hand. The pre-annotated samples are then reviewed and modified by annotators and additional labels are added if needed

Case study by Fort et al. (2010)

- domain: POS-tagging
- pre-annotation had significantly decreased annotation time
- observed that the accuracy of some annotators dropped
- may be caused by annotators relying too much on the automated tools and being less attentive during the correction phase

Fort, Karën, and Benoît Sagot. "Influence of pre-annotation on POS-tagged corpus development." *The fourth ACL linguistic annotation workshop*. 2010.

Developing annotation guidelines
Best-practices summary

What are annotation guidelines exactly?

- annotation guidelines are designed after the domain knowledge has been elicited and annotation scheme has been chosen
- annotation guidelines deliver information about what and how to annotate
- The development of guidelines is often heavily iterative

Requirements for annotation guidelines

- Annotation guidelines should cover the elicited expert knowledge and provide a clear description of the domain and the application
- Annotation guidelines should be designed to serve the purpose of the final application
- Guidelines need to be as precise as possible, but also as generic as possible – more precisely, the concepts which are subject of the annotation should be understandable enough so that no ambiguity occurs
- Guidelines should include negative and positive examples - e.g. for each category/label inclusion and exclusion criterias should be discussed and example annotations should be provided
- Describe problematic labels / cases
- Incorporate rules for situations where no label is applicable
- There should be a clear definition of the tools which will be used

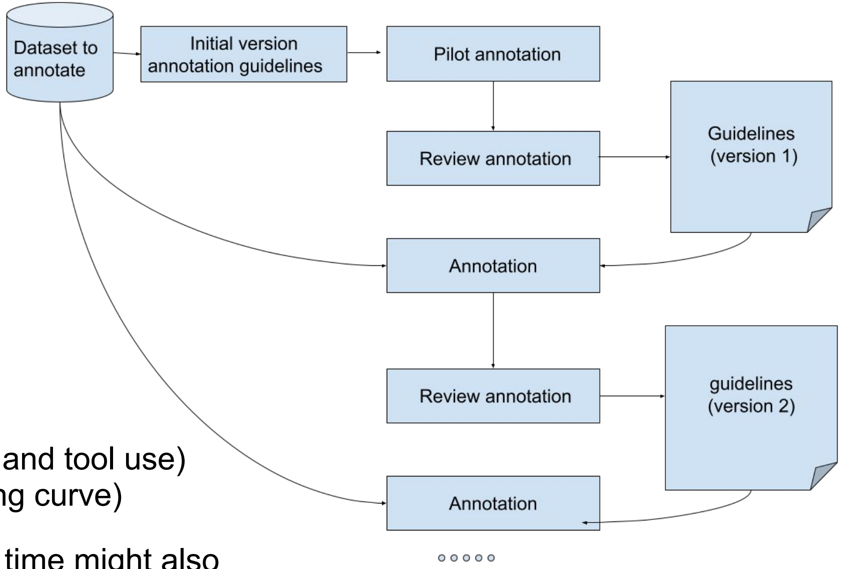
Best practices for creating annotation guidelines

- Annotation guidelines should be developed iteratively with review and refinement subprocesses

1) create a proto version of the annotation manual and choose a tool for the task (software)
2) pilot annotation performed by multiple experts or trained annotators
    - independently
    - no discussions during annotation
    - after done annotating, inspect differences
    - justify decisions
    - document choices & challenging cases
    - identify easy and hard cases
    - refinement of the guidelines
    - meet final decisions of what to annotate

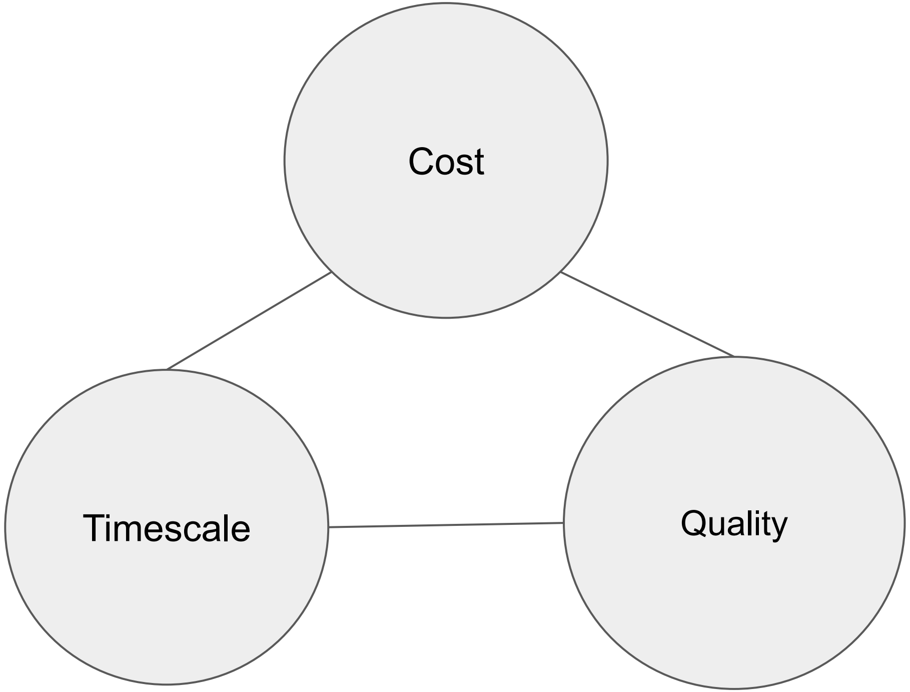
3) create guidelines version 1 (non-pilot)
4) annotate and repeat refinement until an acceptable agreement is reached
5) engage other annotators who will follow the guidelines
    - make sure that annotators have been trained (domain and tool use)
    - new annotators should also do pilot annotation (learning curve)

In some cases measuring intra-annotator agreement over time might also be beneficial (ensures stability over time)



Picture adapted from:
Reiter, Nils. "Anleitung zur Erstellung von Annotationsrichtlinien”. (2020)

Contemporary practices in data annotation: key problems



Cost

Timescale

Quality

Approaches explored here:

- Crowdsourcing and citizen science
- Software assisted/partially automated annotation
- Use of LLMs in data annotation

Crowdsourcing and citizen science

- Addresses problem of **recruiting workers** in manual annotation
- Effort is sourced from broader population, typically via an online platform
- Contributors may be volunteers, or receive some compensation (e.g. micropayment, academic credit, etc)
- Widely used in annotation tasks of various kinds:
  - analysis of texts, video, image data, audio annotation and transcription, functional evaluation tasks, etc.
- Benefit - potentially wide reach for recruitment, low cost
- Concerns - well-understood ethical concerns (for example: low-paid work without income protection, impact on professions affected by use of crowdsourcing) - follow appropriate guidelines when implementing*

* Standing, S. and Standing, C. ‘The ethical use of crowdsourcing’. In: Business Ethics: A European Review 27.1 (2018), pp. 72–80

Machine assisted/partially assisted annotation

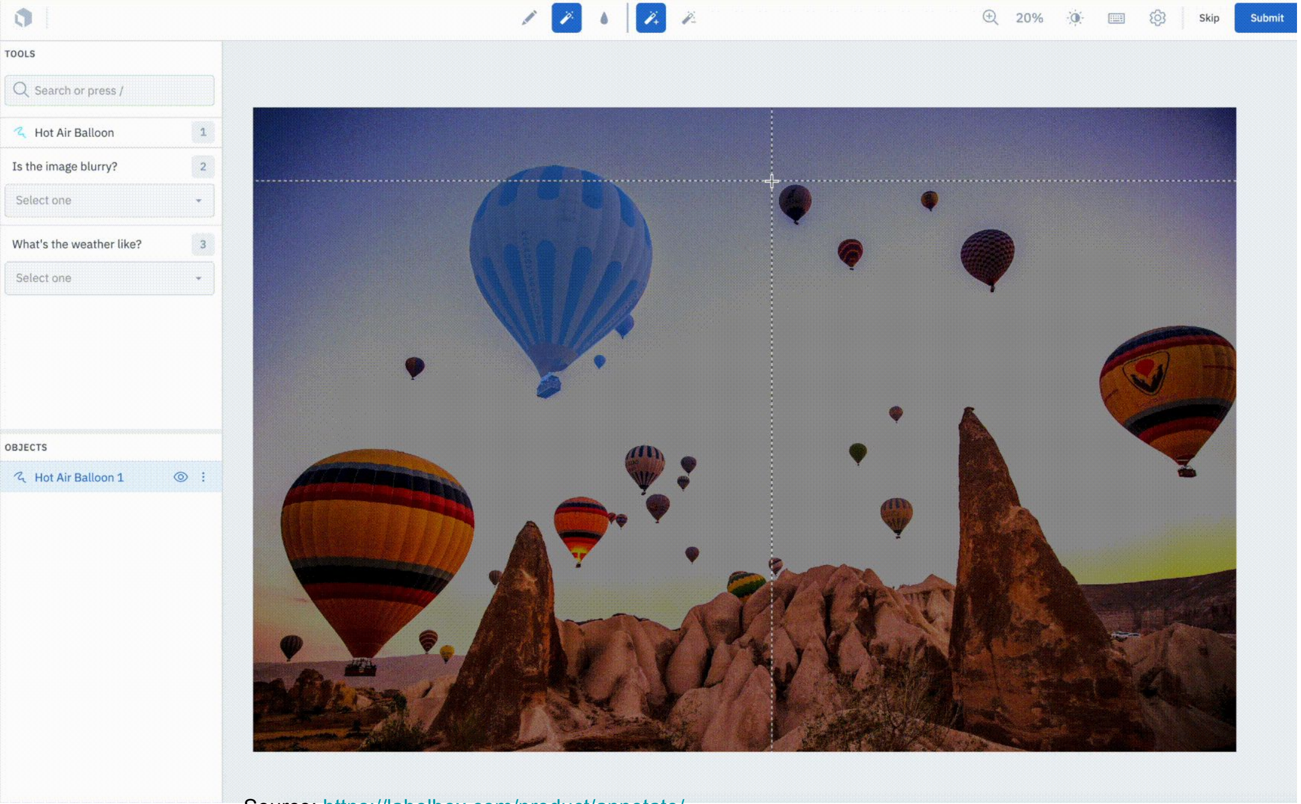
Manual annotation relies on human perception and cognition

Machine assisted annotation has the potential to increase annotator **speed** and **reliability**, hence decreasing cost.

In some domains (e.g. natural-language processing) these technologies are mature and widely used.

Case study: Penn Treebank - semi-automated annotation corrected manually led to performance boost of ~200% in annotation time. Manual annotation had doubled inter-annotator disagreement and ~50% higher error rate.

Machine assisted/partially assisted annotation



Source: https://labelbox.com/product/annotate/

Use of Large Language Models in Data annotation

- Large Language Models (LLMs) (such as Chat GPT) are deep network architectures which can perform a lot of task in the domain of NLU
- “large” is connected to the number of parameters (several billions)
- achieved their prominence (hype) due to their ability to solve tasks given prompts
- Limitations: hallucinations (generate false claims, e.g. citations), expensive to train, difficult to update, unpredictable output
- Can we use these powerful models in data annotation tasks?

Use of Large Language Models in Data annotation

- Several Case studies investigated the usage of LLMs
- Kuzman et al.: explored if ChatGPT can perform genre classification
  - zero-shot approach (without pre-training)
  - ChatGPT performed very good when prompted in English
  - However, poor performance when prompted in Slovenian
  - Conclusion: limitations regarding underrepresented languages
- Gilardi et al. : compared ChatGPT with Amazon MTurk high quality workers
  - Following task were of interest: relevance and topic detection, stance detection (US law related to content moderation), general frame detection
  - ChatGPT outperformed human annotators and was 20x cheaper

Summary: evaluation in LLMs #1

- inter-annotator consistency - LLMs at low temperature respond consistently, but may not be correct, so this is uninformative: compare against existing ground truths generated by human annotators
- LLMs may not do well with more complex tasks in particular*
- Prompt-based refinement may not give stable results**
- ‘Sanitised’ LLMs may deal poorly with language used by marginalised groups***

* Gao, J., Zhao, H., Yu, C. and Xu, R. ‘Exploring the feasibility of chatgpt for event extraction’. In: arXiv:2303.03836 (2023)
** Reiss, M. V. ‘Testing the reliability of chatgpt for text annotation and classification: A cautionary remark’.
In: arXiv preprint arXiv:2304.11085 (2023).
*** Xu et al ‘Detoxifying Language Models Risks Marginalizing Minority Voices’. doi: 10.18653/v1/2021.naacl-main.190.

Summary: evaluation in LLMs #2

Of equal significance are the procedural concerns around LLMs, particularly commercial/online services:

- Lack of clarity surrounding IP issues
- Need to carefully evaluate data reuse policies, if these apply
- Unclear service versioning, low reproducibility, transparency and interpretability (this latter particularly significant where concerns may apply, e.g. automated decision-making services etc)
- Privacy, legal concerns around potential for data capture, reuse in some services
- May not be appropriate for data with high sensitivity
- Consider using local installations where possible/appropriate

Exercise: Machine-assisted annotation

Using the Prodigy data annotation software

https://demo.prodi.gy/?=null&view_id=ner_manual