

## LECTURE 1

# Natural language Annotation for Machine Learning

---

Course Introduction

Importance of Annotation

Annotated Corpora

# Course Structure

---

## Sessions

1. A 'lecture', where we share **theoretical** concepts.
2. A lab session, where we experiment with **practical** methods that implement theoretical concepts.

## Assessment

1. Bi-weekly practical exercises
2. End exam on theoretical concepts through practice.

# What is Annotation?

annotation

/anə'teɪʃ(ə)n/

*noun*

noun: **annotation**; plural noun: **annotations**

1. a note by way of explanation or comment added to a text or diagram.  
"marginal annotations"
2. The action of annotating a text or diagram.  
"annotation of prescribed texts"

# Annotation vs Machine Learning

Often, unfairly, NL annotation is overshadowed by ML

## Natural language annotation

- Time consuming
- Writing and revising annotation guidelines for non-experts
- More frustrating to control annotation quality
- Leads to relatively few publications
- Machine learning *desperately* needs it

## Machine learning

- Relatively short-term
- Writing code for machines
- More fun to push the score and compete with other systems
- Leads to more publications (~diff algorithm x different task)
- *No high results* without good data

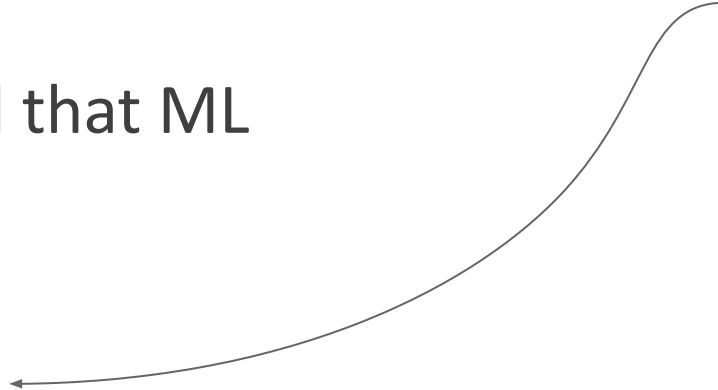
# Why annotation

ML algorithms usually work better when they are provided with:

- Enough Data
- **Quality Data**
- Pointers to what is relevant information

Data needs to be prepared that ML algorithms learn easily:

- Collect data
- Add metadata:



**Any metadata tag used to mark up elements of the dataset is called an annotation**

# Natural language processing

<b>Machine translation</b>	Automatically translating from one language into another
<b>Speech recognition</b>	Convert speech to text
<b>Question answering</b>	Give answers based on Knowledge source
<b>Summarization</b>	Produce a <b>logical</b> summary from collections.
<b>Document classification</b>	Identify an application-dependent category of a document
<b>Fact checking</b>	Check correctness of a given statement.

# Natural language processing

<b>Segmentation</b>	Detect token or sentence boundaries
<b>Part-of-speech tagging</b>	Guess a grammatical category of a token
<b>Syntactic parsing</b>	recognize a sentence and assign an underlying syntactic structure to it
<b>Word sense disambiguation</b>	detect a sense of a word
<b>Semantic parsing</b>	Assign a structure to a sentence that reflects its (approximate) meaning
<b>Natural language inference</b>	identify semantic relation between meanings of natural language sentences

# Layers of linguistic description

**Syntax** - how words and phrases are combined into phrases and sentences, respectively;

**Semantics** - meanings of phrases/sentences and relations over these meanings;

**Morphology** - smallest units of language that has meaning or function (aka morphemes);

**Phonology** - sound patterns of a particular language;

**Phonetics** - sounds of human speech, and how they are made and perceived;

Aminata was feeling "pretty frustrated" as she made her way home from her law studies class. The 19-year-old had been excited to vote for the first time in the French presidential election and had been glued to it on social media, but her candidate, the hard-left Jean-Luc Mélenchon ...



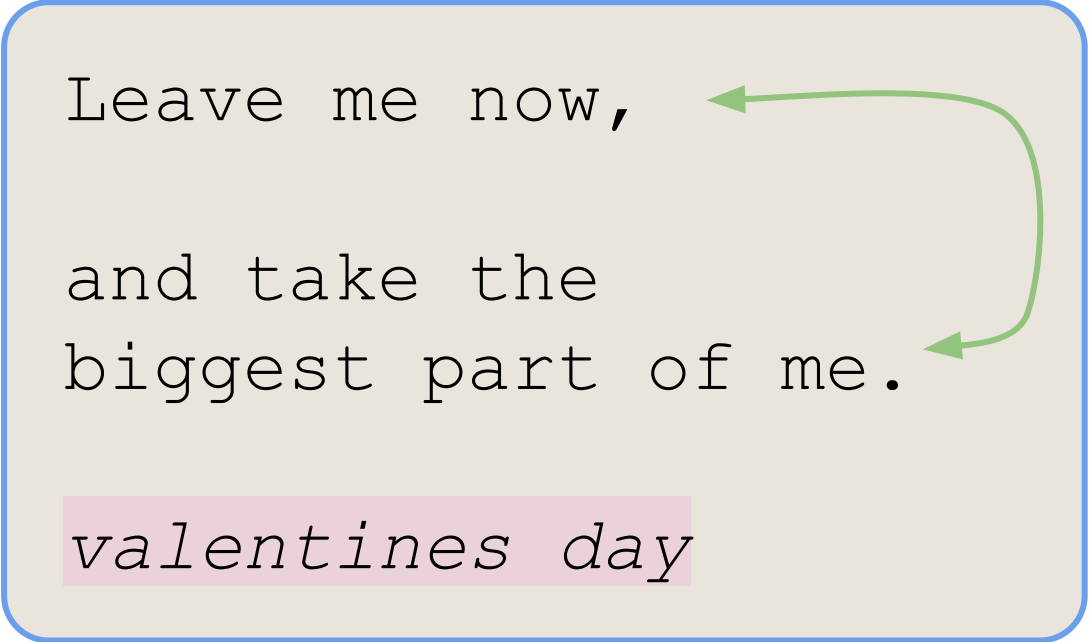
# Layers of linguistic description (II)

**Lexicon** - words and phrases used in a language (vocabulary);

**Discourse analysis** - exchanges of information and the flow of information across sentence boundaries;

**Pragmatics** - how the context of text affects the meaning of an expression and how to recover a hidden/presupposed meaning;

**Text structure analysis** - how narratives and other textual styles are constructed to make larger textual compositions;



Leave me now,  
and take the  
biggest part of me.

*valentines day*

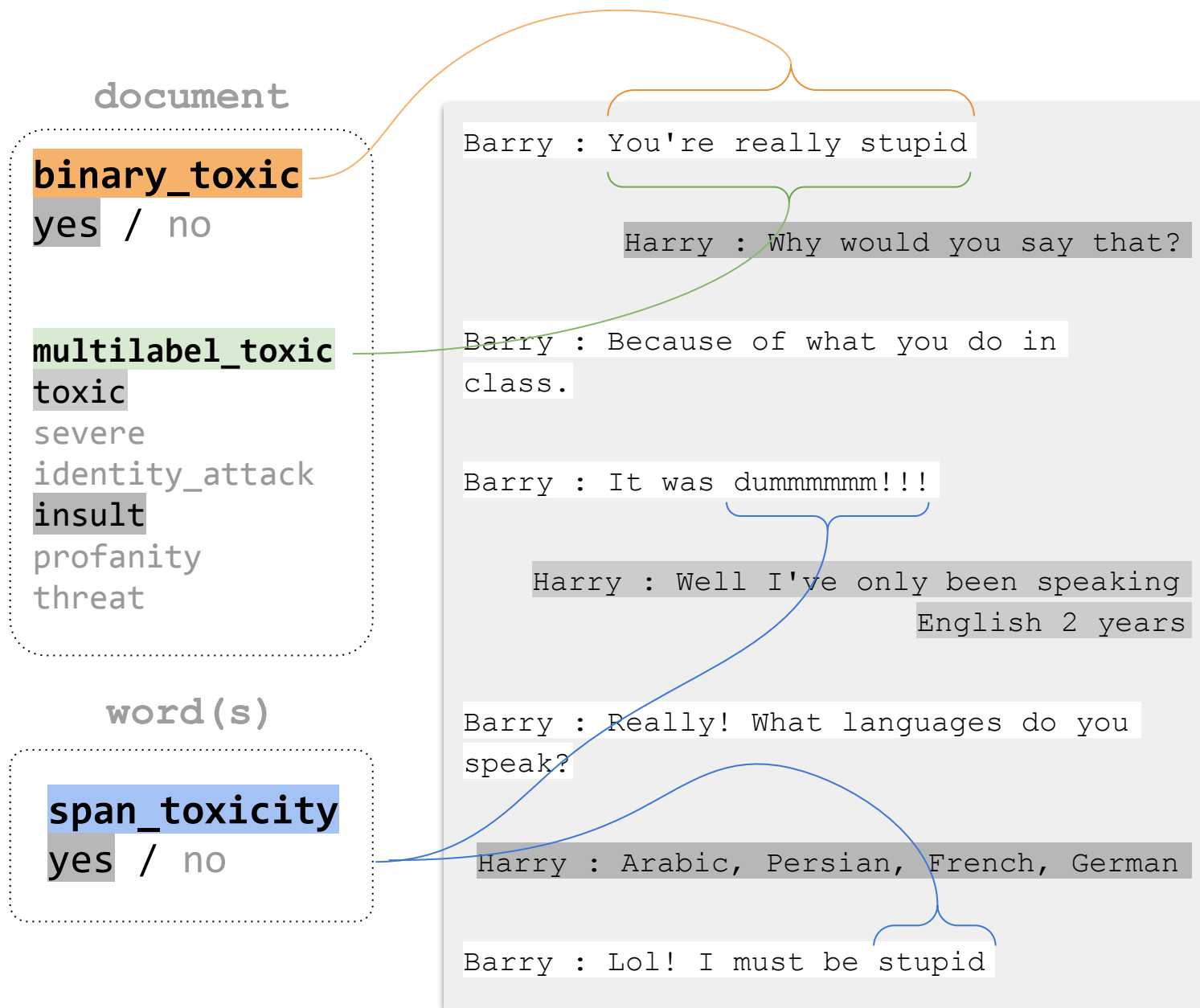
# Types of Annotation

---

# Toxic Language Detection

“threatening, insulting, or off-topic language that is likely to make a person leave a discussion”

(Wulczyn et al., 2017)



# Part of Speech Annotation

**Part of speech** – a token-based; It serves as an initial word sense disambiguation.

N... typically indicates a noun

V... typically indicates a verb

J... typically indicates an adjective

NP... often means a proper noun

NN... often means an ordinary (common) noun

VB... often means part of the verb BE

VH... often means part of the verb HAVE

VV... often means part of a lexical verb (e.g. *play, run*)

	text	lemma_	pos_	tag_	dep_	shape_	is_alpha	is_stop
0	This	this	DET	DT	nsubj	Xxxx	True	True
1	is	be	AUX	VBZ	ROOT	xx	True	True
2	a	a	DET	DT	det	x	True	True
3	sentence	sentence	NOUN	NN	attr	xxxx	True	False
4	about	about	ADP	IN	prep	xxxx	True	True
5	the	the	DET	DT	det	xxx	True	True
6	greatness	greatness	NOUN	NN	pobj	xxxx	True	False
7	of	of	ADP	IN	prep	xx	True	True
8	cats	cat	NOUN	NNS	pobj	xxxx	True	False
9	.	.	PUNCT	.	punct	.	False	False

# Kinds of annotation (II)

## Semantic typing

Token/phrase-based; It denotes a type from a reserved vocabulary or ontology.

A common form of which is Named-entity-recognition.

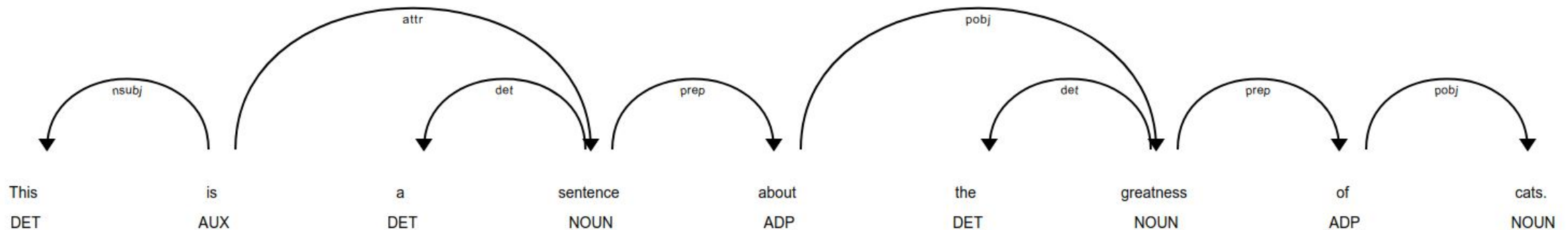
Stratford **GPE** -upon-avon is the birthplace of Shakespeare **PERSON** .

# Kinds of annotation (III)

## Dependency Parsing

Recognise a tree of relationships between words, based on their grammatical roles.

	text	lemma_	pos_	tag_	dep_	shape_	is_alpha	is_stop
0	This	this	DET	DT	nsubj	Xxxx	True	True
1	is	be	AUX	VBZ	ROOT	xx	True	True
2	a	a	DET	DT	det	x	True	True
3	sentence	sentence	NOUN	NN	attr	xxxx	True	False
4	about	about	ADP	IN	prep	xxxx	True	True
5	the	the	DET	DT	det	xxx	True	True
6	greatness	greatness	NOUN	NN	pobj	xxxx	True	False
7	of	of	ADP	IN	prep	xx	True	True
8	cats	cat	NOUN	NNS	pobj	xxxx	True	False
9	.	.	PUNCT	.	punct	.	False	False



# Machine learning with language data

---

What's the difference between token-level, span-level, and document-level annotation?

# Machine learning with language data

Three major types of ML algorithms wrt annotated corpus:

## **Supervised learning**

- Learns a hypothesis that maps inputs to a pre-defined structure or a sequence/set of labels
- Needs annotated data for training

## **Unsupervised learning**

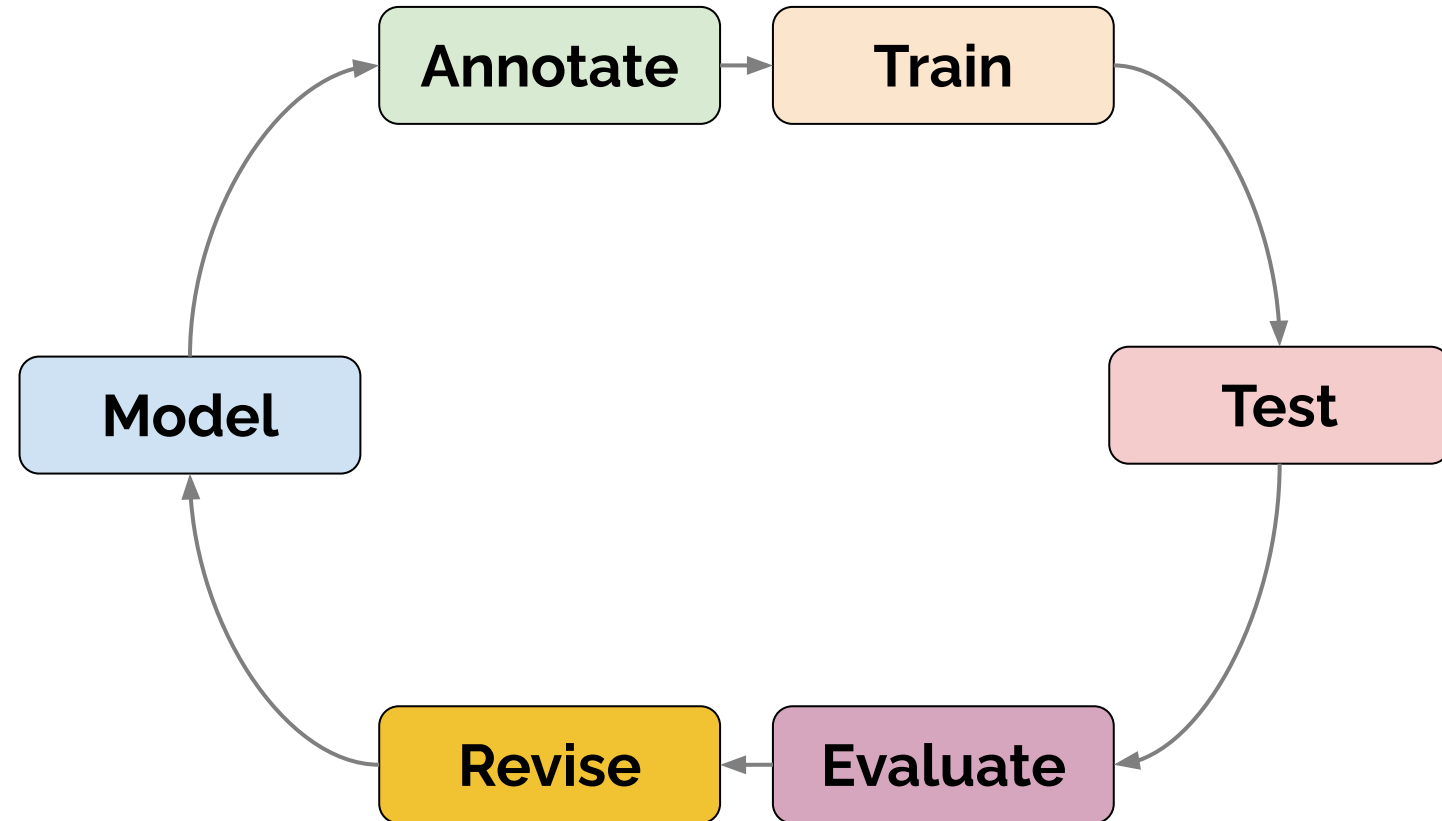
- Learn structure from unlabeled data

## **Semi-supervised learning**

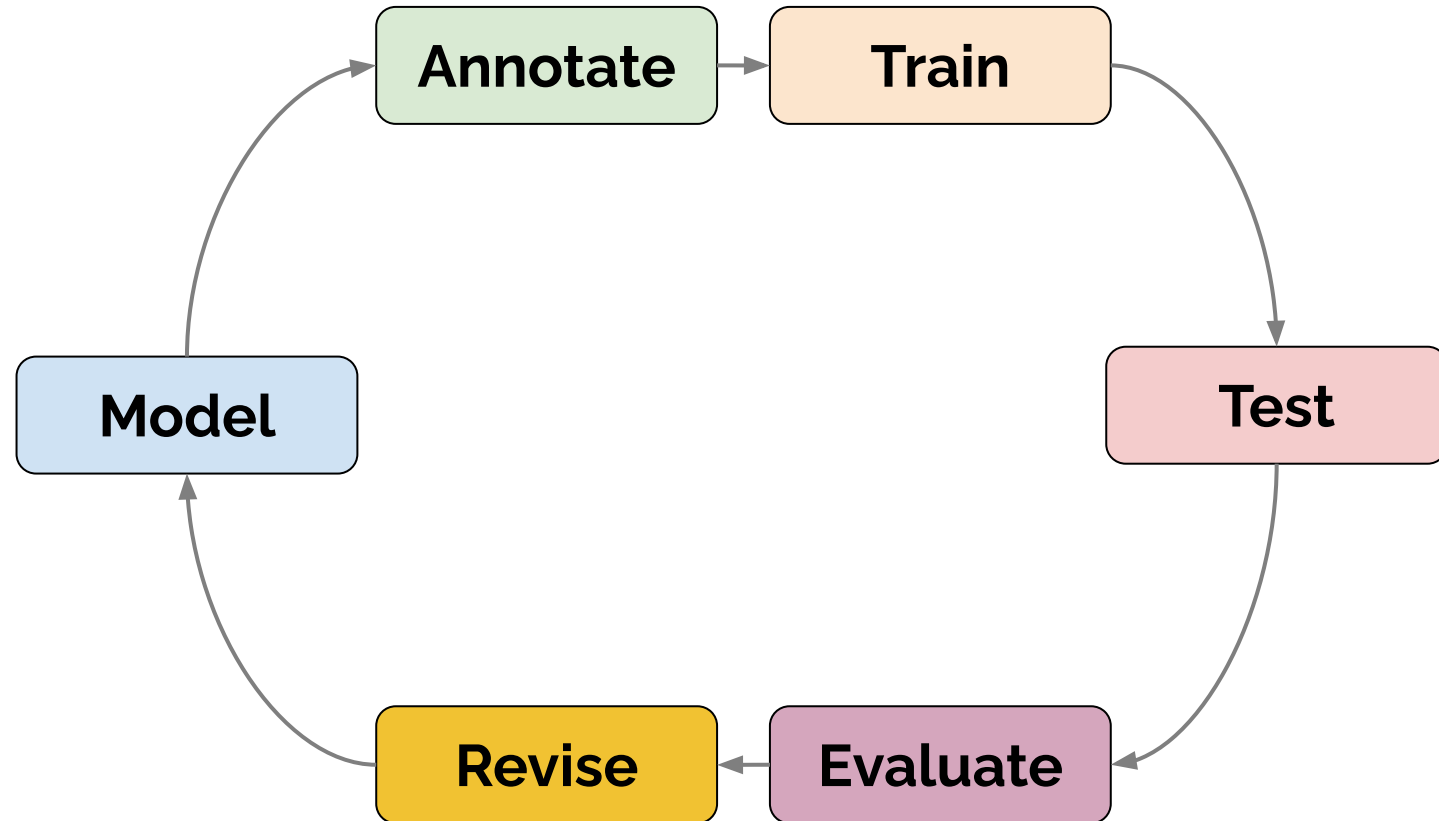
- Combination of the two former



# MATTER cycle



# MATTER cycle



**Model** the phenomenon

**Annotate** with the specification

**Train** algorithms

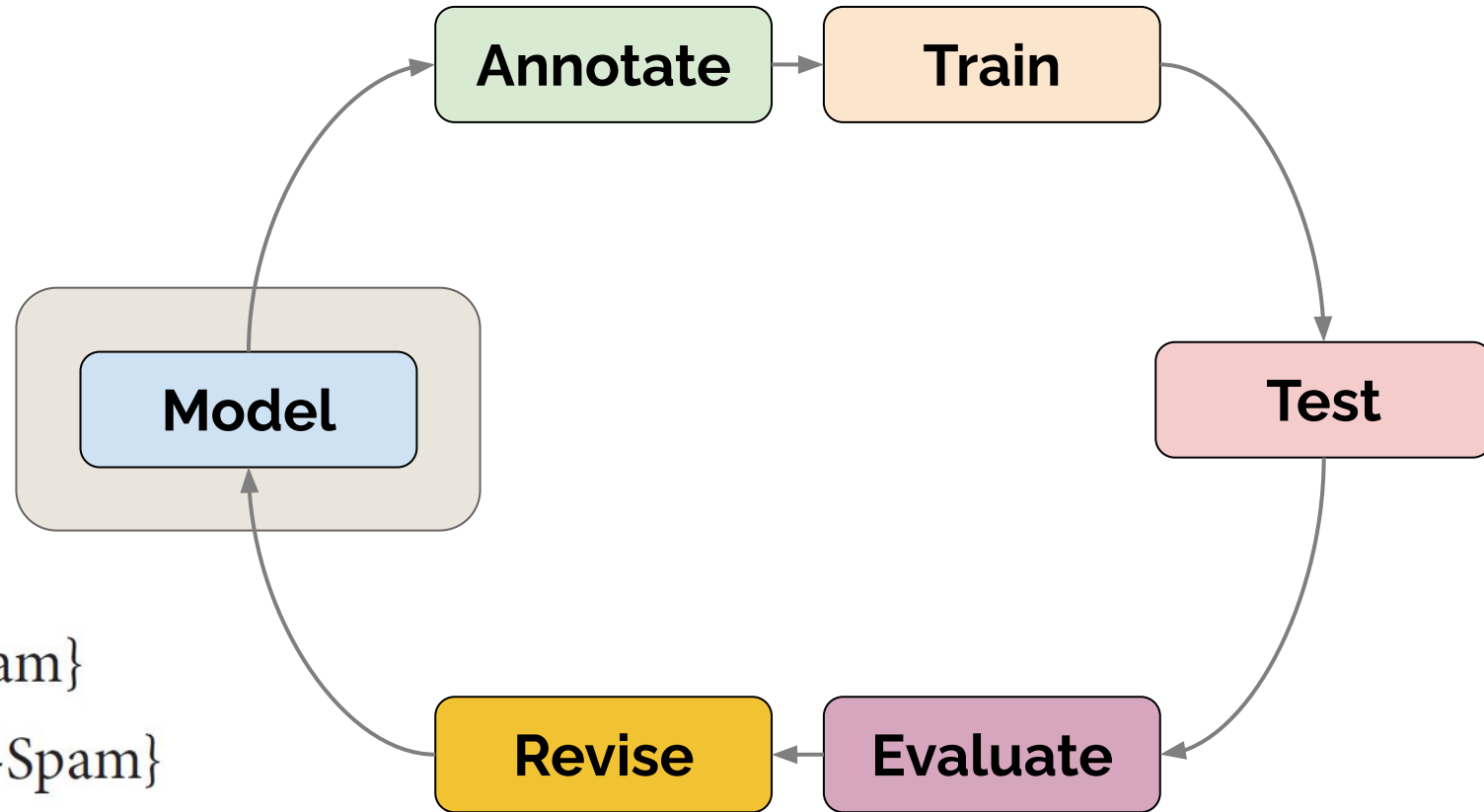
**Test** them on unseen data

**Evaluate** the results

**Revise** the model and algorithms

# MATTER: Model the phenomenon

- *Terms*
- *Relations*
- *Interpretation*



$T = \{\text{Document\_type}, \text{Spam}, \text{Not-Spam}\}$

$R = \{\text{Document\_type} ::= \text{Spam} \mid \text{Not-Spam}\}$

$I = \{\text{Spam} = \text{"something we don't want!"},$

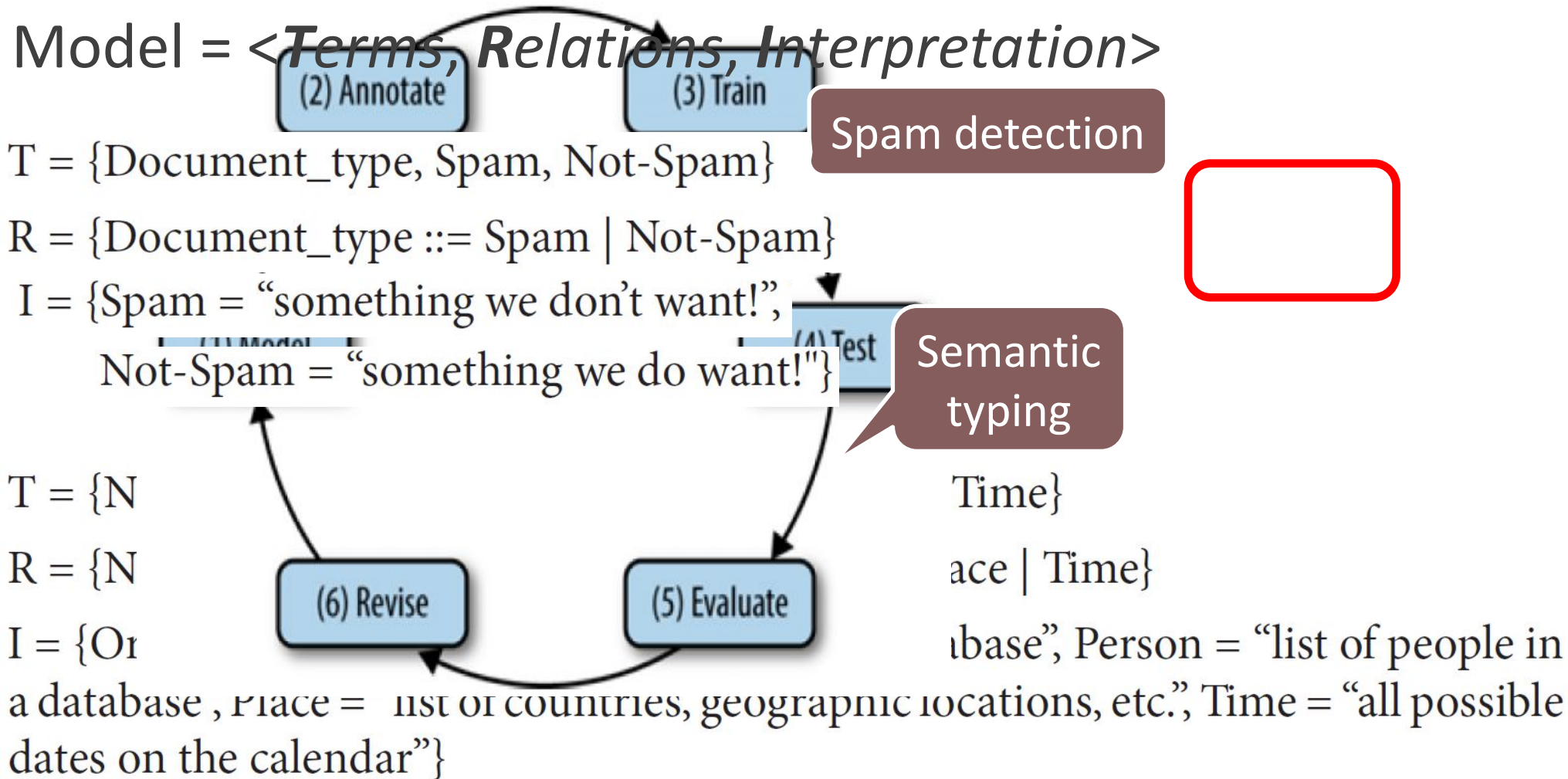
$\text{Not-Spam} = \text{"something we do want!"}\}$

# We are talking about modelling a phenomenon.

---

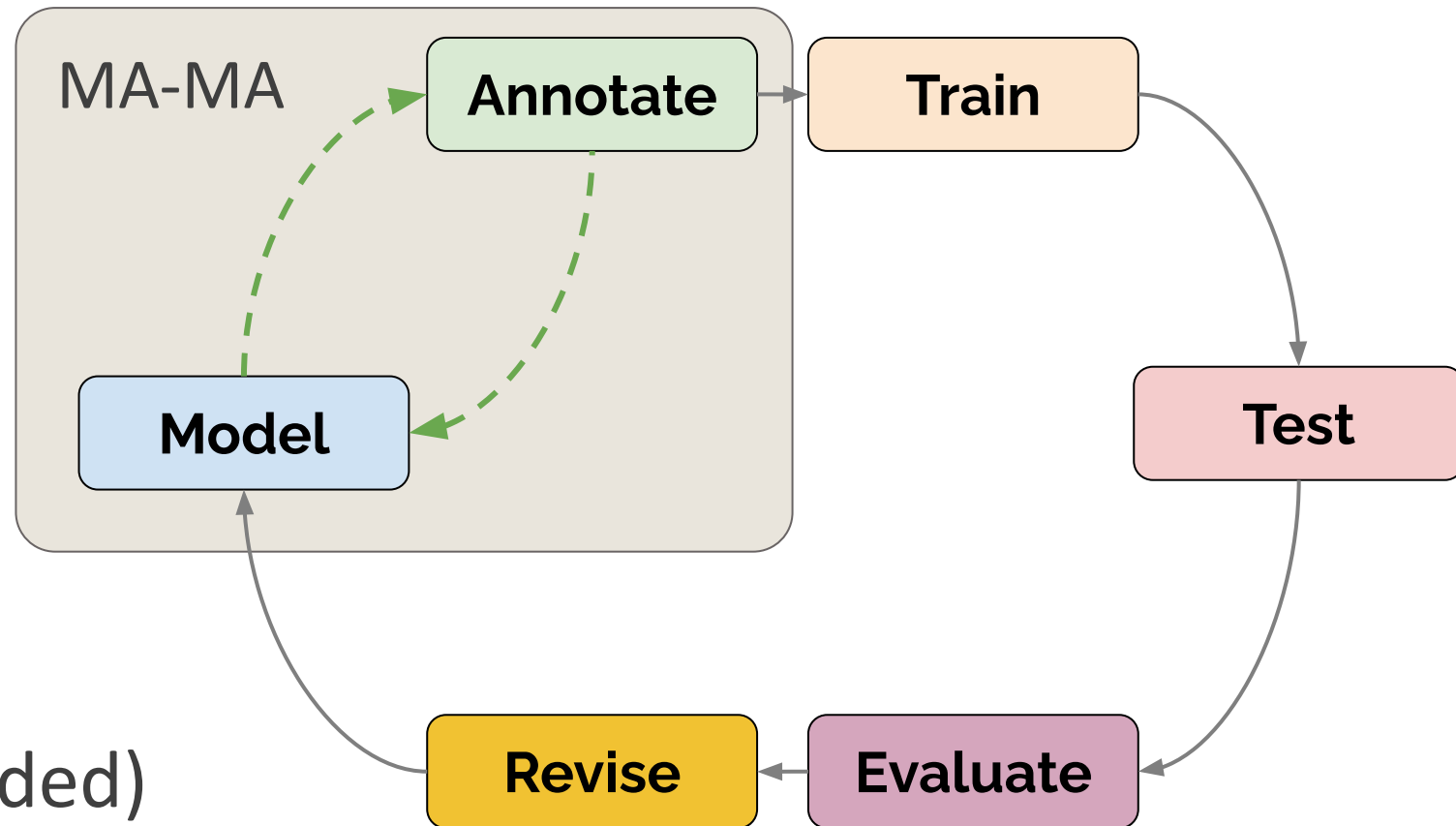
What model are we not talking about?

# MATTER: Model the phenomenon



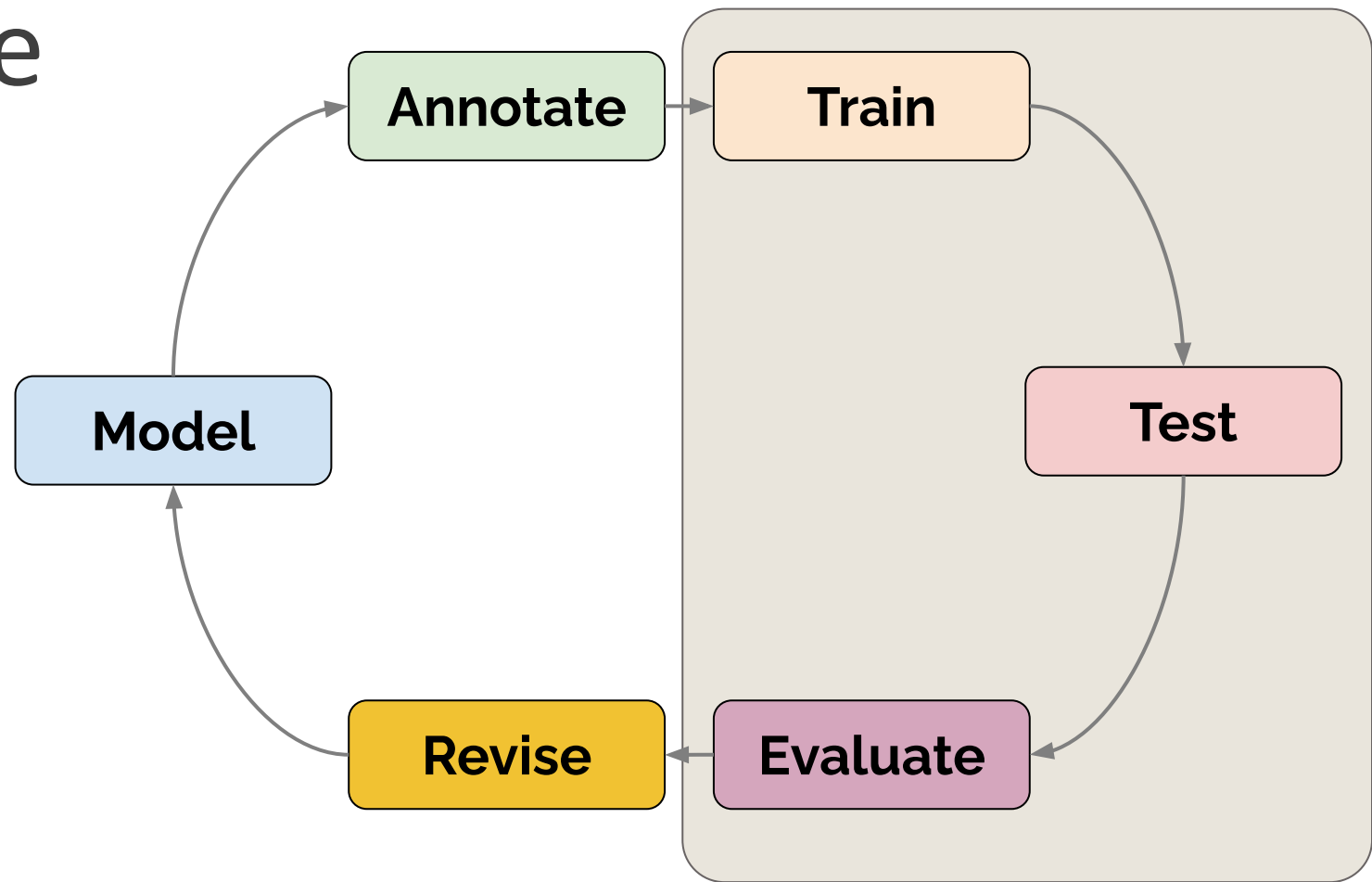
# MATTER: Annotate with the specification

- Design specification
  - Consuming tags
  - Non-consuming tags
- Write guidelines
  - Span of the tag
- Annotate
- Evaluate & revise (if needed)
  - Understanding or just agreement?
- Adjudication & gold standard

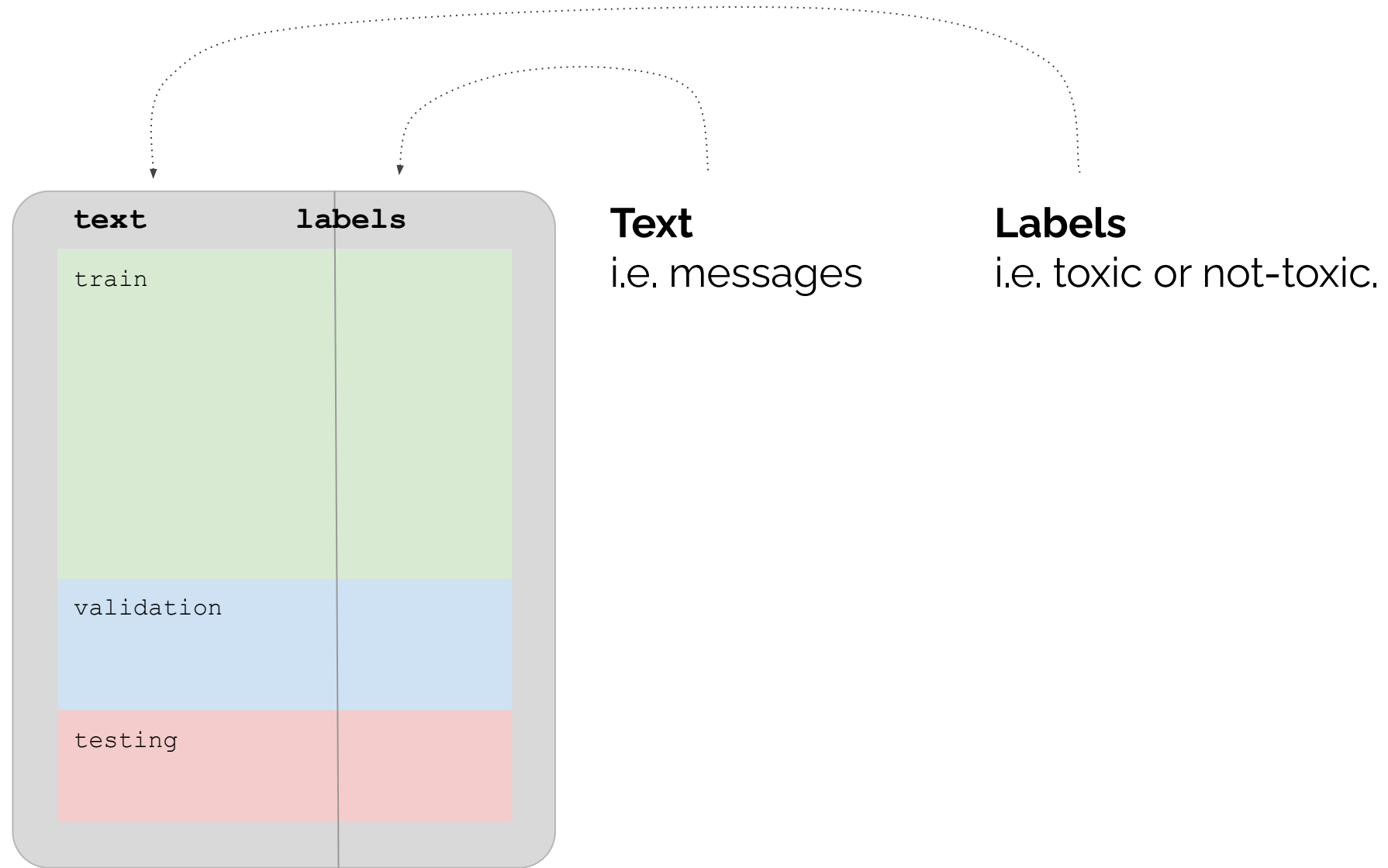


# MATTER:

Train, test, evaluate

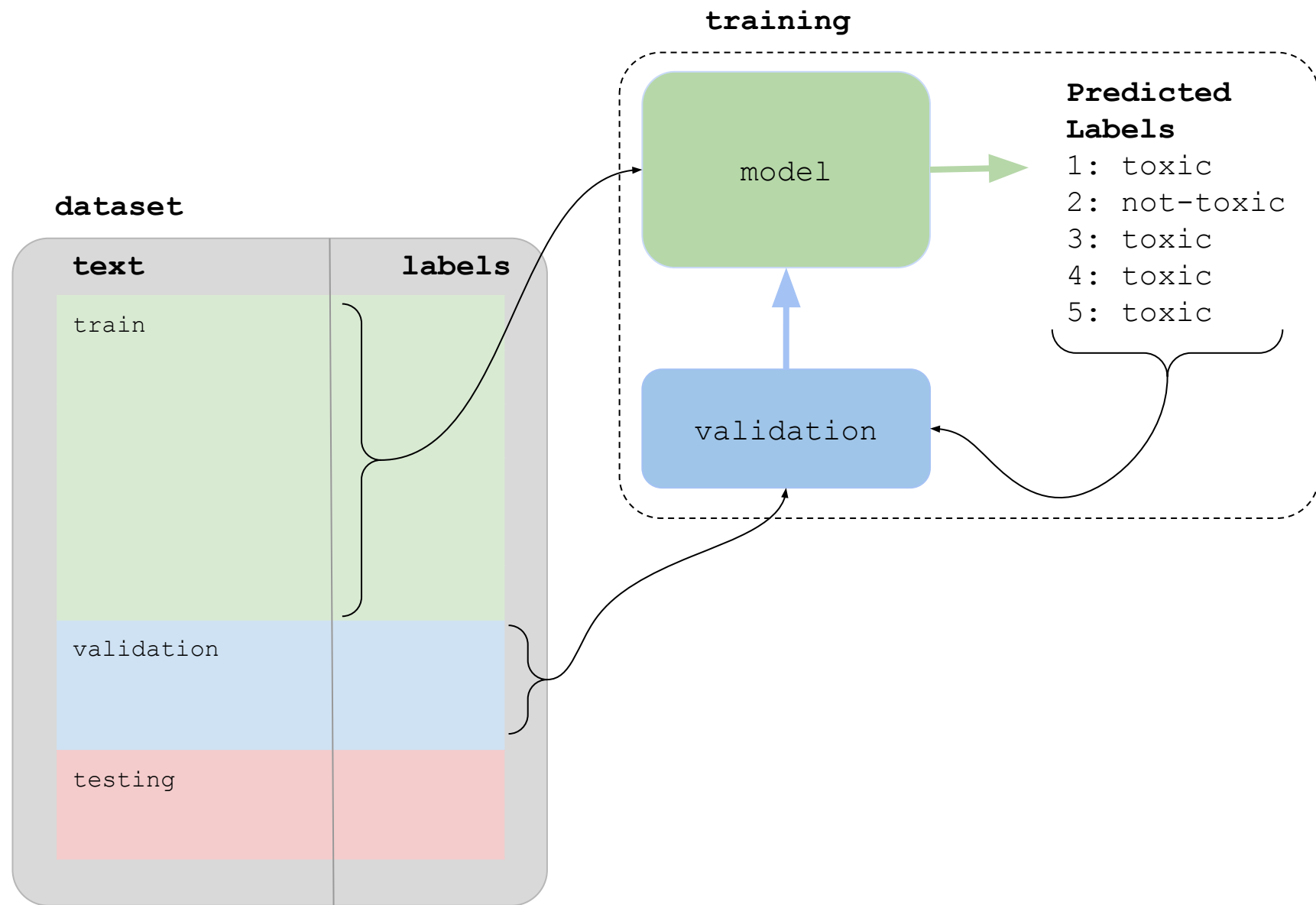


# A dataset



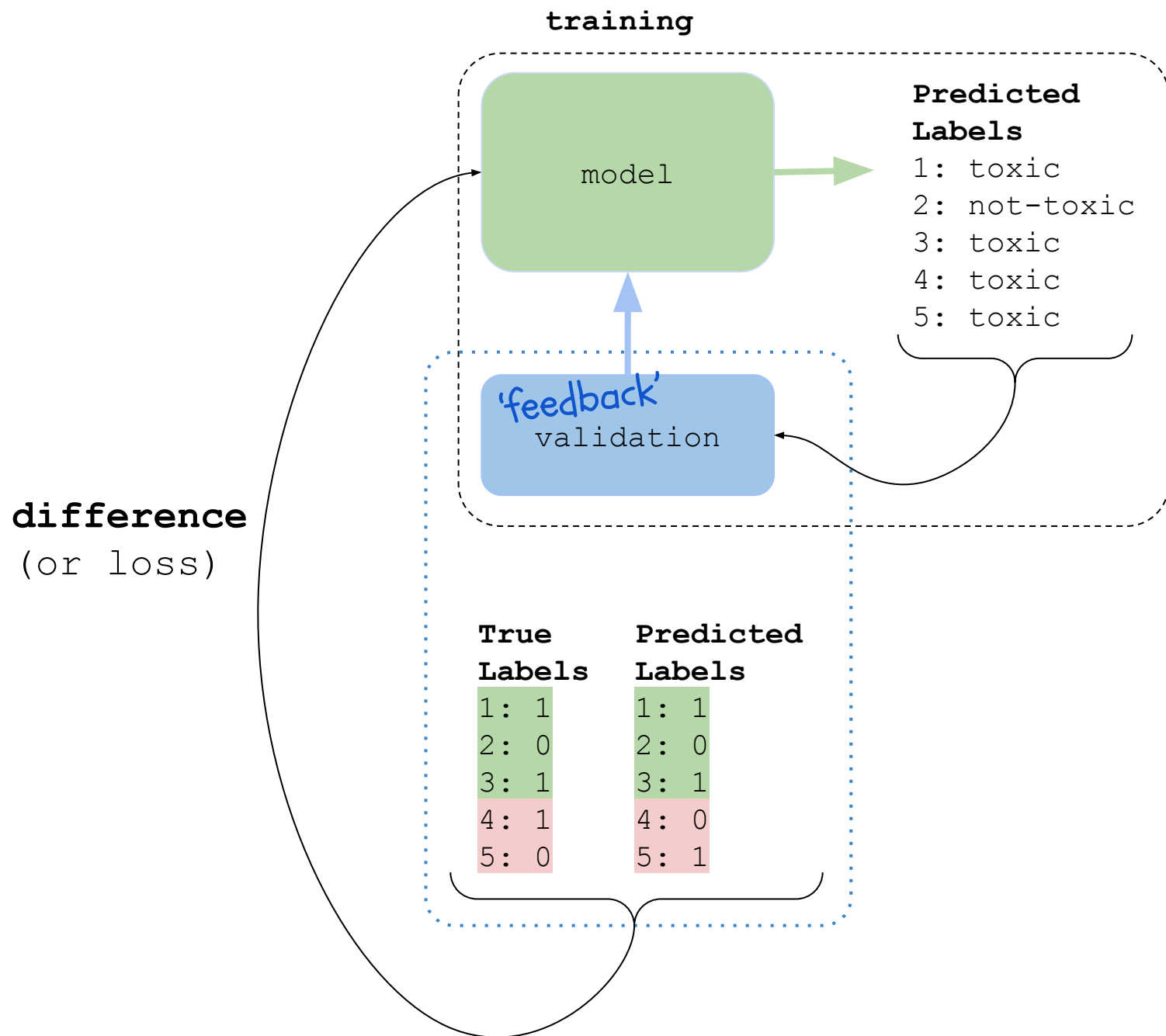


# Training a Machine Learning Model

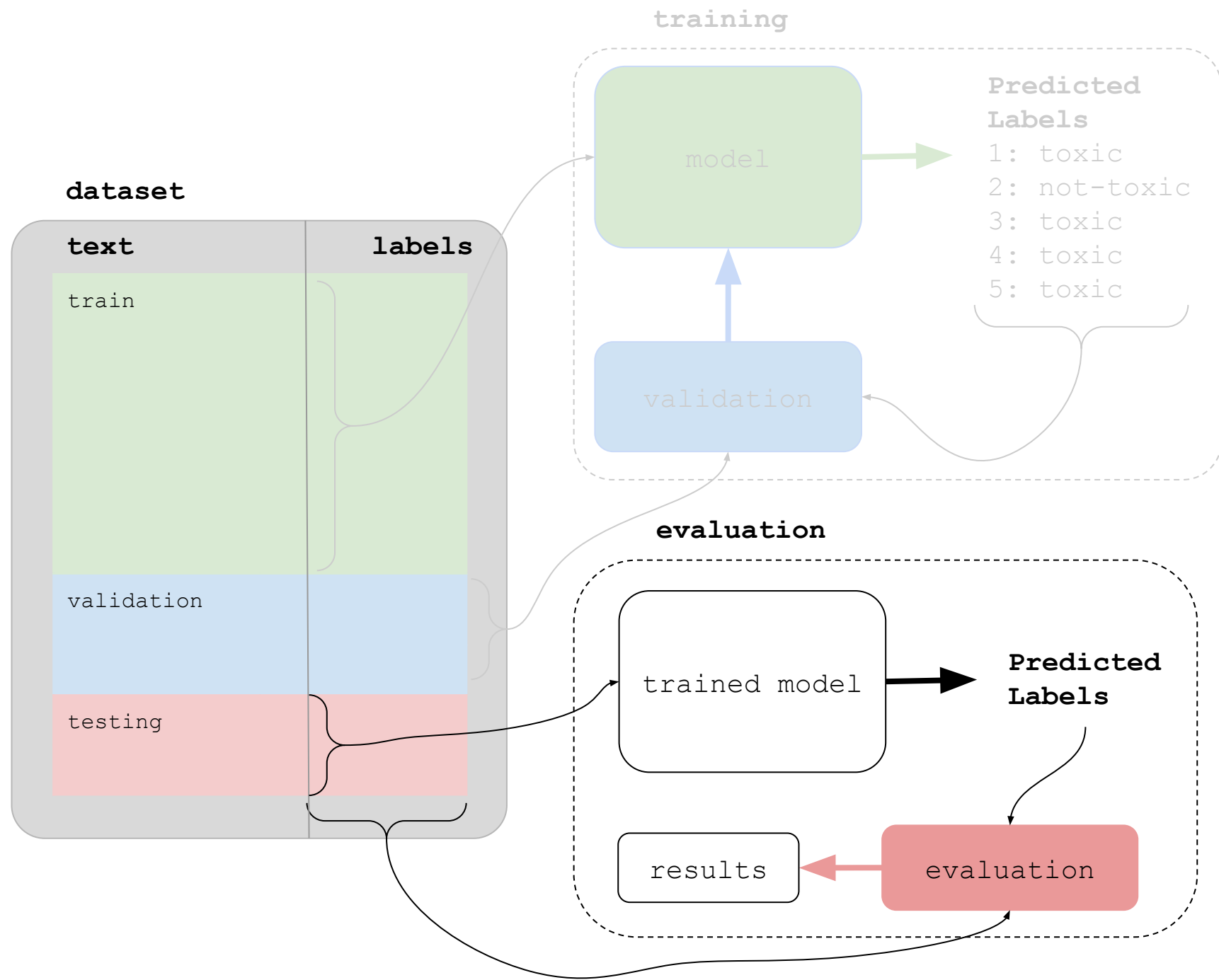


# Validation

The model improves its predictions based on the difference between the true labels and predicted labels.



# Evaluation

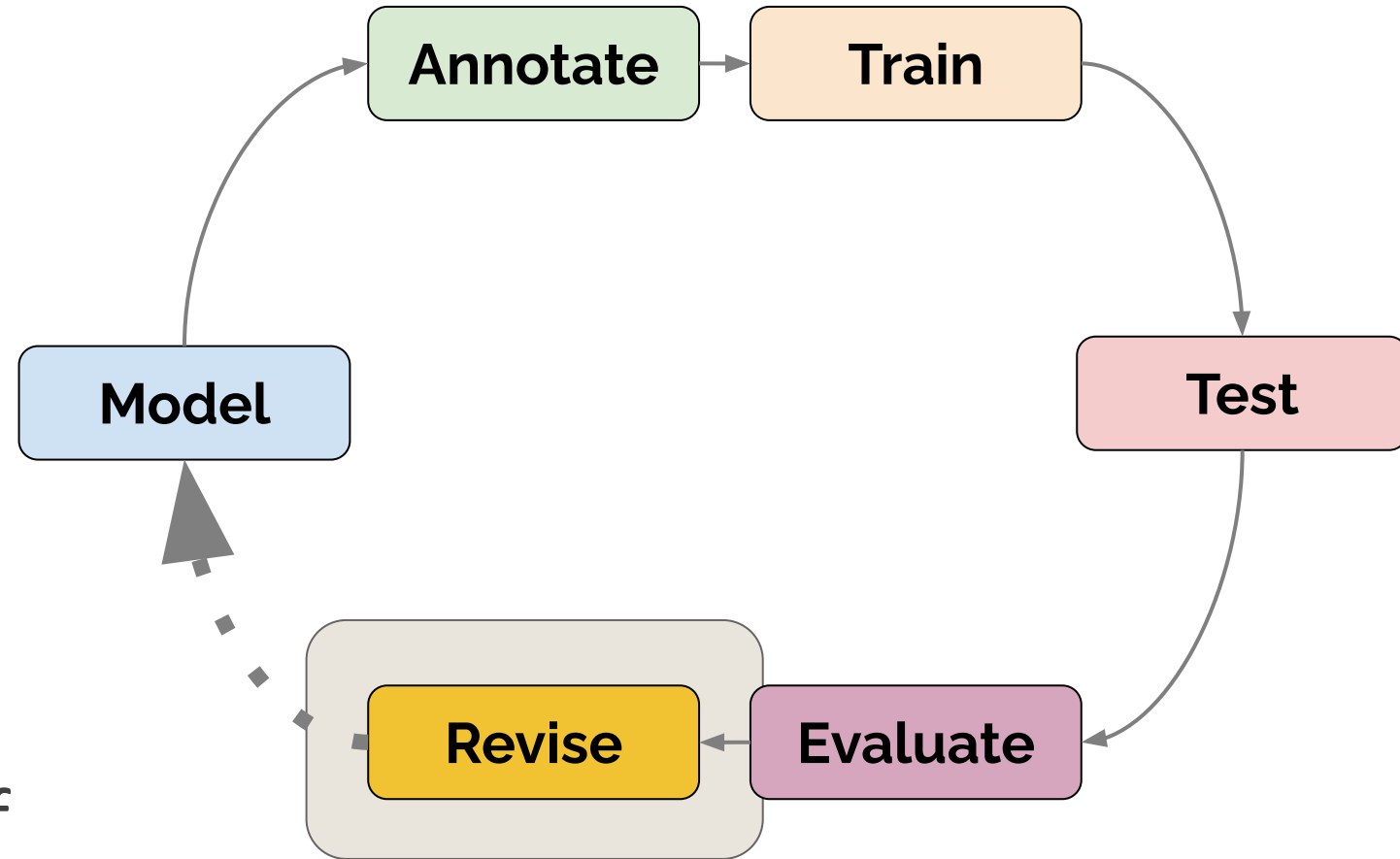


# MATTER: Revise

Possible revisions:

- Introduce a new tag or type, e.g., add Event type
- Split an existing tag or type, e.g., distinguish geopolitical from nongeopolitical
- Collect more data

The MATTER cycle restarts if revision is made



# Take a break?

---

# The Corpus

---

## What Is a Corpus?

A corpus is a collection of machine-readable texts that have been produced in a natural communicative setting. They have been sampled to be *representative and balanced* with respect to particular factors; for example, by genre—newspaper articles, literary fiction, spoken speech, blogs and diaries, and legal documents. A corpus is said to be “*representative of a language variety*” if the content of the corpus can be generalized to that variety (Leech 1991).

This is not as circular as it may sound. Basically, if the content of the corpus, defined by specifications of linguistic phenomena examined or studied, reflects that of the larger population from which it is taken, then we can say that it “*represents that language variety*.”

# Corpus linguistics

Corpus linguistics studies language as expressed in a (large) collection of “real world” texts.

It has caused heated debates (in the past):

A corpus can't describe a natural language entirely

Corpus linguists study real language, other linguists just sit at their coffee table and think of wild and impossible sentences

Natural language is infinite.





# Timeline of corpus linguistics

## 1960s

- Brown corpus (Kucera & Francis)
  - First broadly available balanced corpus
  - 500 text samples 2000 words each (1961 English)

## 1970s

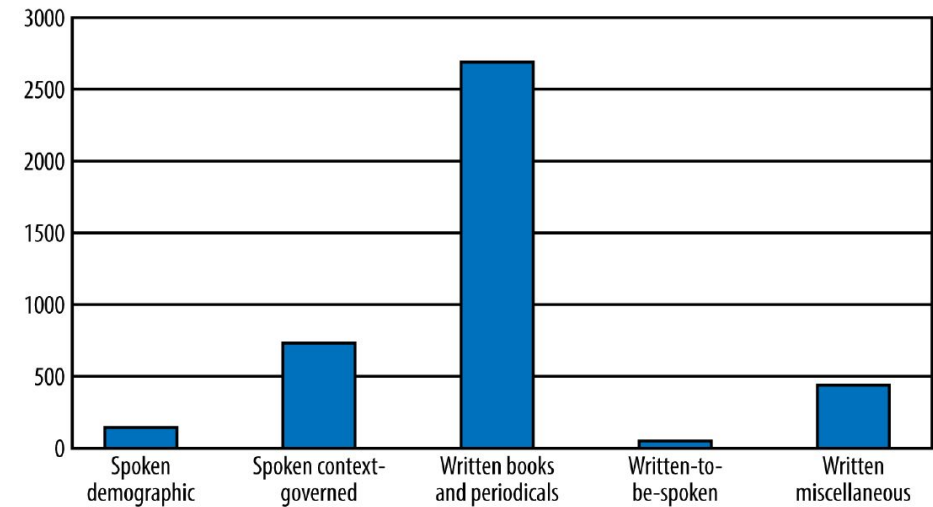
- Statistical models for speech recognition from speech corpora
- Vector space model for document indexing
- London-Lund Corpus (LLC)

# Timeline of corpus linguistics (II)

1990s

- Penn TreeBank (Marcus et al. 1993)
  - Tagged and parsed sentences
  - 4.5 M words (naturally occurring English)
- British National Corpus (BNC)
  - 100 M words (spoken & written British English)

```
(S (NP (NNP John))  
  (VP (VPZ loves)  
      (NP (NNP Mary))))  
(..))
```



# Timeline of corpus linguistics (III)

2000s

- Gutenberg Corpus
  - 25,000 free electronic books
  - Spoken, fiction, popular magazines, newspapers, and academic texts.
- Google N-gram Corpus
  - 1 trillion word tokens from public web pages
  - Up to five n-grams for each word token + frequencies
- The Text Encoding Initiative (TEI)
  - develop and maintain a standard for the representation of texts in digital form

# Timeline of corpus linguistics (IV)

## 2010s

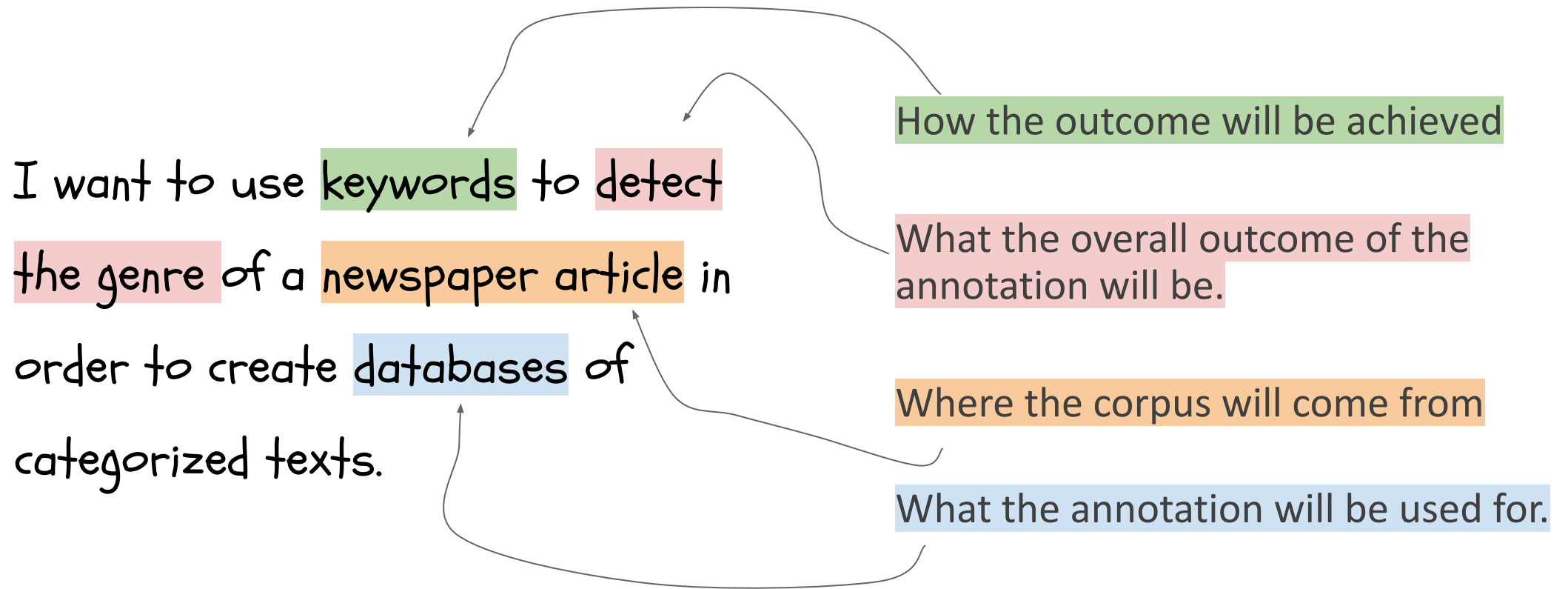
- Twitter, Facebook, Blogs
- Tatoeba (translations from foreign language learners)
- Stanford Natural Language Inference (SNLI) dataset
  - 570K human written English sentence pairs
- Universal dependencies
  - >100 treebanks (~70 languages)
- Groningen Meaning Bank (EN) & Parallel Meaning Bank (EN, NL, DE, IT)
  - Texts annotated with rich formal meaning representations
- International standards organizations, such as ISO, begin to recognize and co-develop text encoding formats for corpus annotation efforts



# Annotation in Practice

---

# Step 1. Define annotation goal



# Step 2. Refine the Annotation Process

Informativity vs correctness

Make annotation easy for annotators,  
recover the rest of relations  
automatically

Annotation that is not too difficult for  
annotators to complete accurately

Annotation that is most useful for  
your task

Twitter adopts 'poison  
pill' plan to shield  
itself from Elon Musk  
takeover.

Tesco starts selling  
lateral flow kits as free  
testing in England ends

# Step 2. Refine the Annotation Process

## Scope of annotation

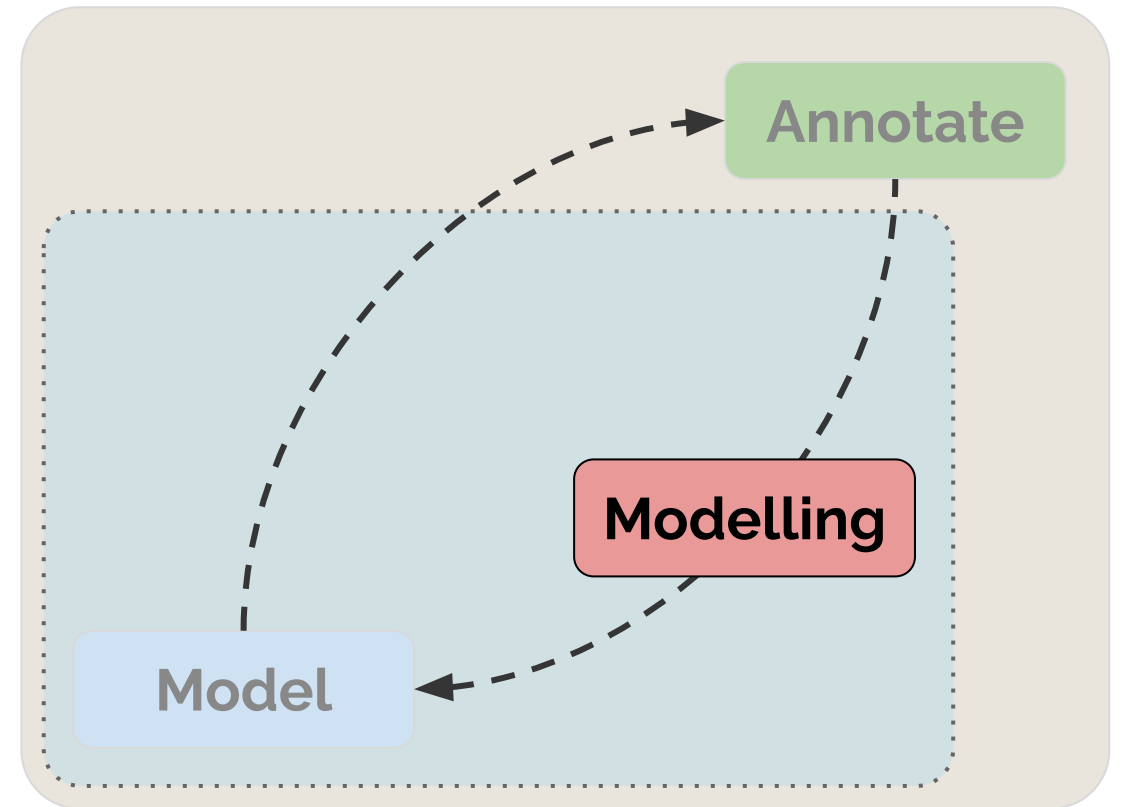
Example questions:

- How specific are categories?
- What are/ are not keywords?
- Is article context relevant?

Practicalities

- How long should an annotation take?
- What tools will the annotator use?
- What if the annotator is not sure?

MA-MA





# Refining the goal (Data Preparation)

## Sources of the Corpus

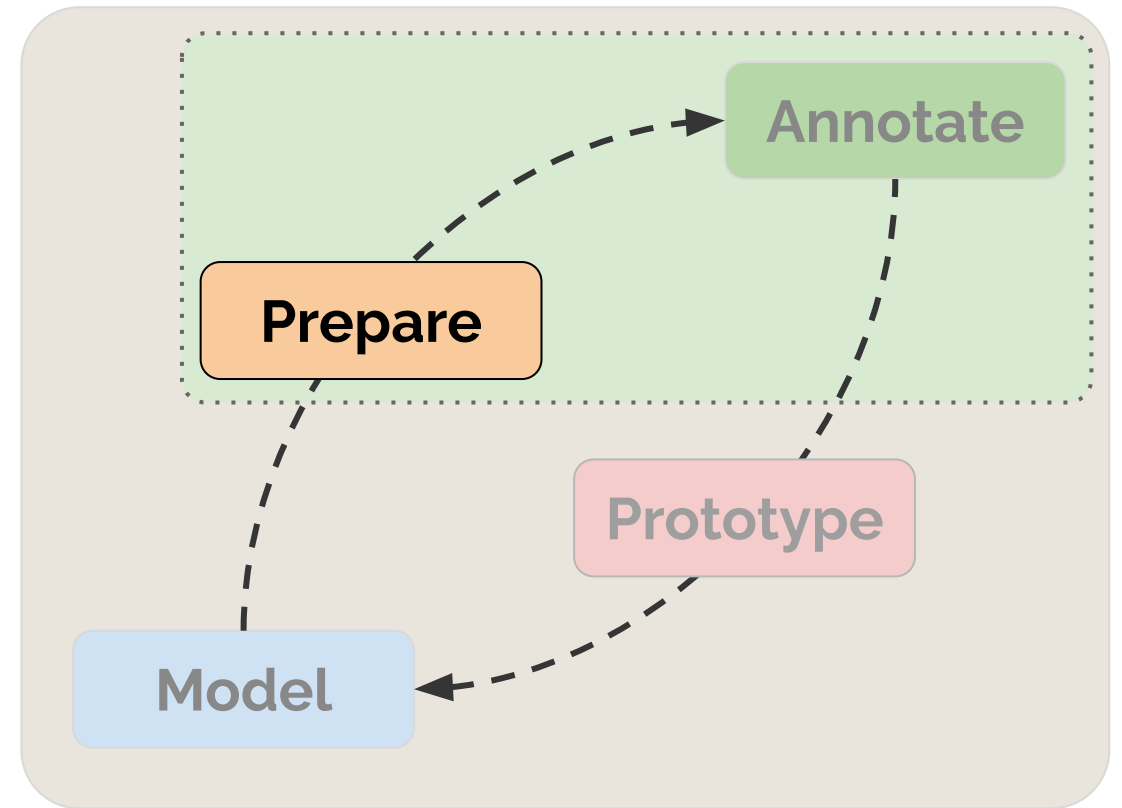
Example questions (newspaper genre classification):

- Include similar sources or not (e.g., NYT vs WSJ)?
- Are blogs news sources?
- Exclusively online newspapers?
- Only written articles or also transcripts of broadcasts?

Example questions (temporal annotation):

- Which publication styles?

## MA-MA



# Step 4. Collect Data

How representative is the data?  
(~coverage)

How balanced is the data?  
(~distribution)

Practically:

- Internet
- Elicit from people: experts or crowd
- Read speech vs spontaneous speech

# Size of the corpus

Depends on several factors:

- Complexity of annotation task
- Time
- Money
- Limited resources
- Type of NLP task

**Have a sample corpus that has examples of all the phenomena that you are expecting to be relevant to your task**

Corpus	Estimated size
ClueWeb09	1,040,809,705 web pages
British National Corpus	100 million words
American National Corpus	22 million words (as of the time of this writing)
TempEval 2 (part of SemEval 2010)	10,000 to 60,000 tokens per language dataset
Penn Discourse TreeBank	1 million words
i2b2 2008 Challenge—smoking status	502 hospital discharge summaries
TimeBank 1.2	183 documents; 61,000 tokens
Disambiguating Sentiment Ambiguous Adjectives (Chinese language data, part of SemEval 2010)	4,000 sentences



# Step 3. Search for related work

## Language Resources

- The Linguistic Data Consortium (LDC)
- The European Language Resources Association (ELRA)

## Conferences and organizations:

- Association for Computational Linguistics (ACL)
- Language Resources and Evaluation Conference (LREC)
- Conference on Computational Linguistics (COLING)
- American Medical Informatics Association (AMIA)
- Various NLP challenges (aka shared tasks)
  - Semeval, CoNLL Shared Task, i2b2 NLP Shared Task

# Step 5. Evaluate Annotation

---

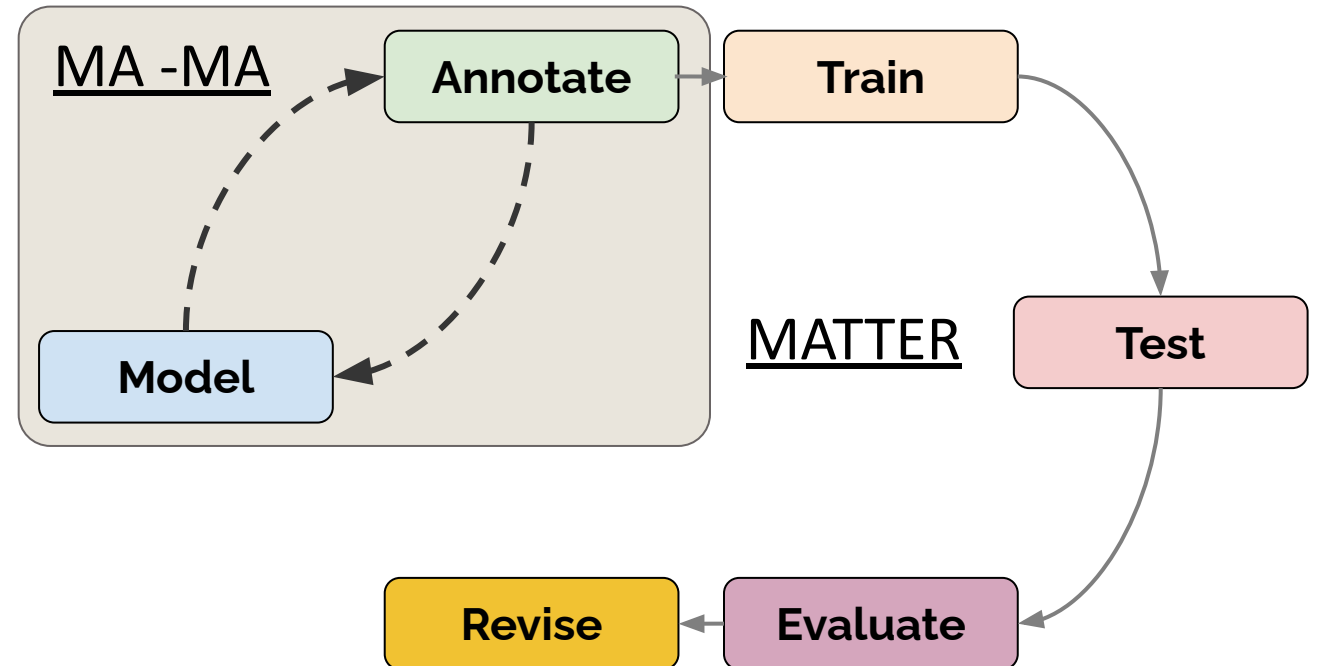
Can I trust these labels?

Do annotators agree?

...

# Wrapping up

- Annotation helps ML
- Corpus linguistics
- Kinds of annotations
- MA..MATTER cycle
- Annotation goal



# Bonus: Mini Demo on NLTK Corpus

# Clarification: Inter-annotator agreement

Multilabel Task		Krippendorff's $\alpha$	Cohen's $\kappa$
Hate Speech	Ross 2016	.18	
Cyberbullying	Van Hee et al.2015		.19 - .69
Racism	Nobata et al. 2016		.456
Toxic Language	Wulczyn 2017	.4297	
This thesis		<b>.4243 (.18 - 73)</b>	<b>.6293</b>

Rater agreement scores for annotated data are given as both **Cohen's  $\kappa$**  and **Krippendorff's  $\alpha$** .

Krippendorff's alpha is not ideal on this nominal multilabel data. However, it does facilitate comparison to the relevant.

		IDENTITY	INSULT	PROFANE	SEVERE	THREAT	TOXIC
Cohen's Kappa	IDENTITY	0.51					
	INSULT	-0.09	0.57				
	OBSCENE	-0.03	-0.16	0.59			
	SEVERE	-0.06	-0.17	-0.13	0.62		
	THREAT	0.00	-0.04	-0.06	-0.06	0.52	
	TOXIC	-0.07	-0.26	-0.19	-0.14	-0.11	0.60