TEXT ANONYMIZATION WITH EXPLICIT MEASURES OF DISCLOSURE RISK

PRIVACY AND DATA PROTECTION

- Privacy is a fundamental right for every individual and an essential component for a democratic society (Art. 8 of the European Convention on Human Rights, Art. 12 of the Universal Declaration of Human Rights)
- Privacy ≠ Data Protection
- Regulations like the GDPR were established to ensure that every natural person *(data subject)* has the right to control their personal data and anyone who processes such data *(data processor, data controllers*) has to abide by the rules

PERSONALLY IDENTIFIABLE INFORMATION

- Personally Identifiable Information (PII) is information tied to an identifiable individual
  - Direct identifiers (e.g. name, passport number, social security number)
  - Quasi identifiers (e.g. date of birth, nationality)
- By combining just zip codes, gender, and birth dates, one could identify 63-87% percent of the U.S. population
- To use datasets containing personal information, they need to be *anonymized*
  - Direct and quasi identifiers need to be removed or modified (e.g. deletion, generalisation etc.)

RELATED WORK

- ***Natural Language Processing (NLP)*** approaches (mainly medical data, NER based approaches)

  + They take the complexity of the nature of language into account

  - They anonymize all identifiers equally, without the possibility of parameterising according to their actual disclosure risk

  - They require labelled data, which is a costly and time-consuming task. Also, difficult to release and share with the research community due to regulations

  - Not all personal information can be grouped into a NE type

RELATED WORK

- ***Privacy Preserving Data Publishing (PPDP)*** (k-anonymity, t-plausibility, C-sanitized):

  + They allow for parameterization of the trade-off between data protection and data utility

  - They treat language as a flat entity

RELATED WORK

- ***Differential privacy***:

  + They provide strong privacy guarantees by ensuring that the computed statistics of the data cannot be exploited for identifying an individual

  - They treat the text as a distribution of words, disregarding semantics

  - It has been argued that such approaches result in poor quality data

OUR APPROACH

- Combine NLP and PPDP approaches
- Move past a NER based approach to text anonymization, by focusing on *text anonymization with explicit disclosure risk*
- Automatically annotate text for personal information

Wikipedia

Wikipedia biography dataset

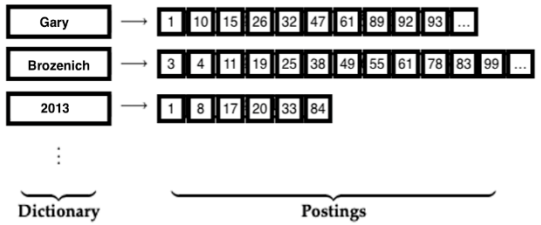
"*Gary Brozenich is a visual effects supervisor. Brozenich and ...*"

Wikidata

Gazetteer

[('Gary', 'Brozenich'), ('The', 'Lone', 'Ranger'), ...]

Inverted Index



Gazetteer annotated entities

Gary Brozenich is a visual effects supervisor .

Brozenich and his fellow visual effects artists were nominated for an Academy Award for Best Visual Effects for the 2013 film The Lone Ranger .

Greedy k-anonymity function

Random k-anonymity function

Sequence labelling model

Sequence labelling model

WIKIPEDIA

- We started working with a Wikipedia biography dataset
  - Filtered the entries to consist of only individuals (not bands, groups, finctional characters etc.)
  - Kept only entries with valid ids
  - Kept only name and summary (first paragraph) of the article
- Already split into training (80%), development (10%), and test (10%)

DATASET STATISTICS

	Before	After
Train	582.659	411.623
Valid	72.831	51.498
Test	72.818	51.496
Total	*728.308*	*514.617*

Table 1: # of entries before and after filtering

	# Tokens	# Sentences	Shortest	Longest	Average
Train	1.837.049	33.543.087	1	122	4
Valid	230.902	4.209.091	1	65	4
Test	227.208	4.143.743	1	85	4

Table 2: Datasets' statistics

BACKGROUND KNOWLEDGE

- Wikidata as a knowledge base (KB) that one could use to identify an individual
- Downloaded and filtered a wikidata dump file to include only individuals
- Each of the entries contained the following information:
  - id (e.g. '*Q6680829*'), a unique identifier for each item
  - labels (e.g. '*Cambridge*'), a small unit of information
  - descriptions (e.g. *'city in Massachusetts, United States'*), information to disambiguate labels
  - aliases (e.g. *'Barry Obama'),* alternative names, like nicknames, acronyms, or translations
  - claims, which consist of a property (e.g. *'cause of death')* and a value (e.g. *'epiglottitis')*
  - sitelinks, name of the equivalent Wikipedia article page

LINGUISTIC VARIANCE

- Dates
  - Dates in Wikidata only follow the *yyyy-mm-dd* format, while in Wikipedia the dates come in many formats (e.g. 24 April 2009, April 24, 2009, April 24 2009, April 24, 2009, April 2009, 2009 etc.). For each date in Wikidata then, the Wikipedia formatted equivalents of it were added next to the original date
- Names
  - Sentences in Wikipedia articles often refer to the individual with parts of their names (e.g. last name, first & last name only, etc.). For each full name in Wikidata, first name, last name, and any middle names were also added to the inverted index as separate strings

LINGUISTIC VARIANCE

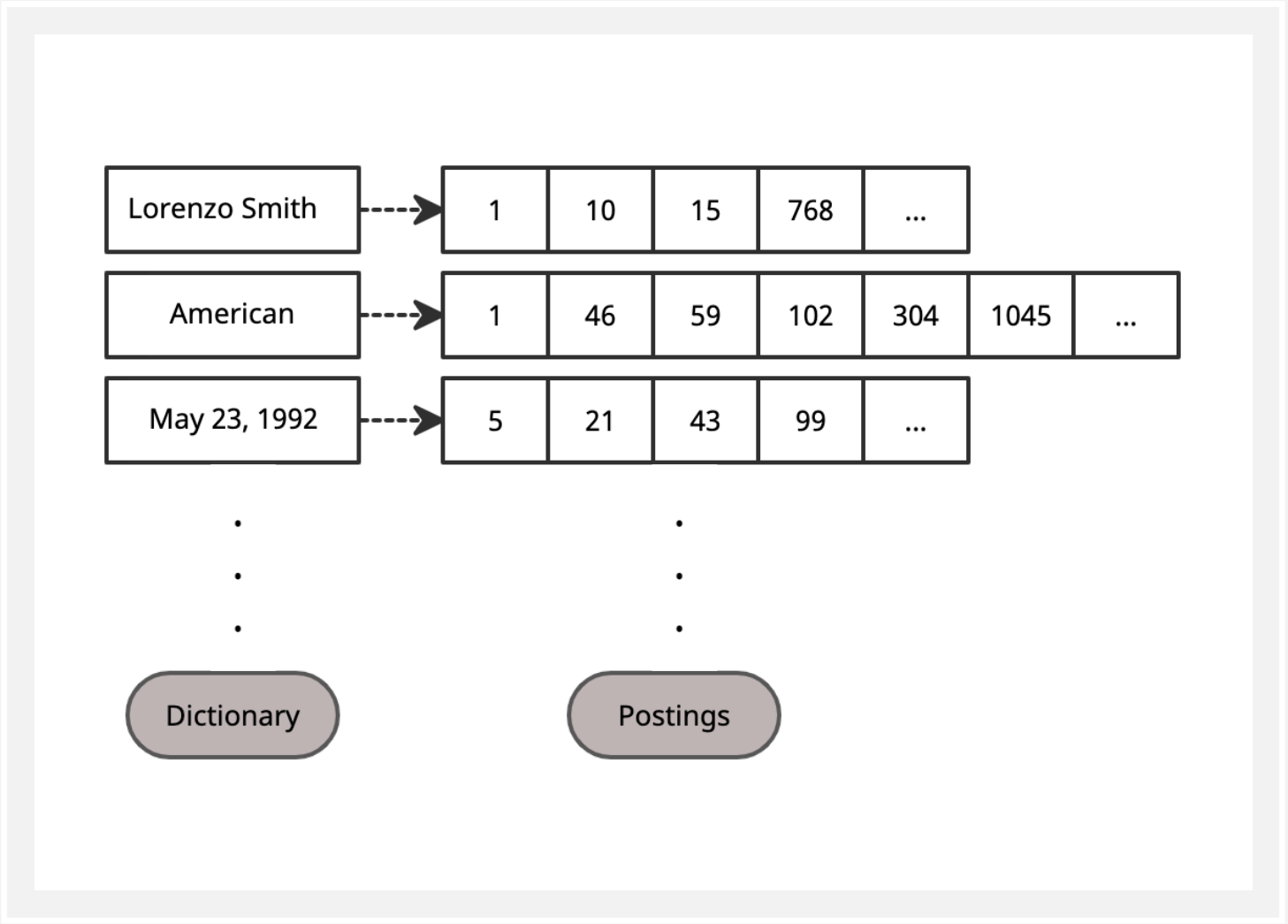
- Nationality-Country
  - It is more common to have a person's nationality in Wikipedia (e.g.' Austrian poet') while the country name is more common in Wikidata (e.g.' place of birth': ' Austria'). For each nationality then, the equivalent country name was added to the inverted index
- Official country names
  - To account for cases such as 'Hellenic Republic' which is the official country name, and 'Greece' which is more commonly used

LINGUISTIC VARIANCE

- Unique tokens before variance:
  - *119.023.317*
- Unique tokens after variance:
  - *172.567.265*

INVERTED INDEX

- An *inverted index*, which is an index used to map back terms to the documents they occurred, consisting of a dictionary and its postings
- The Wikidata file was further reformatted into key (person name) - value (list of strings) pairs so as to be used for its creation

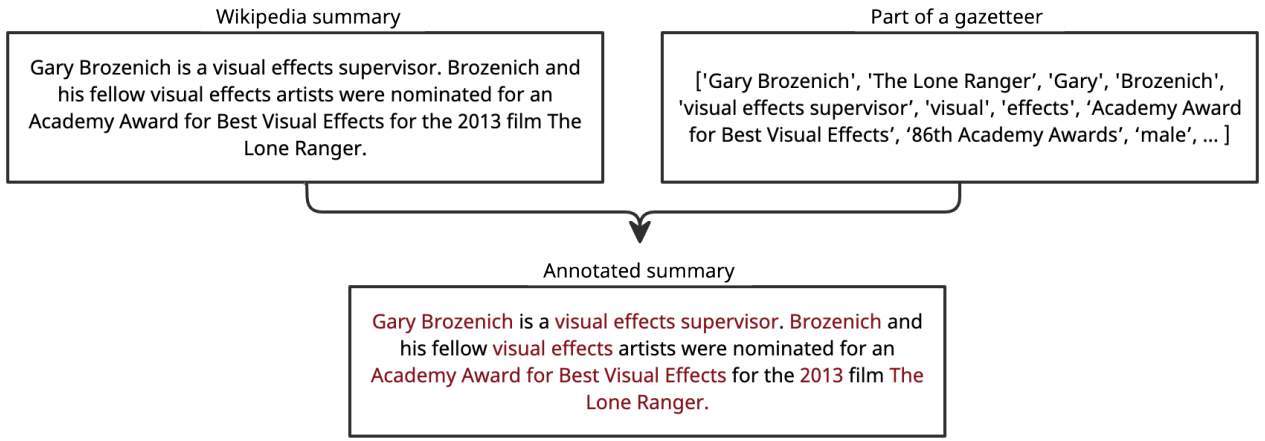


INVERTED INDEX

- We decided on a Boolean retrieval model in which one can query an inverted index using terms which are combined with Boolean operators (Boolean AND in our case)
- For example for the query '*Lorenzo Smith AND American*' the function will:
  - Locate *Lorenzo Smith* and retrieve its postings
  - Locate *American* and retrieve its postings
  - Intersect the two posting lists and return the result

GAZETTEER

- To be able to query the inverted index we needed to find a mode of extracting terms from the Wikipedia summaries
- We implemented a *gazetteer,* which can search in a document and find occurrences of items present in a list of words, phrases etc
- These items were then used to annotate the Wikipedia summaries



GAZETTEER

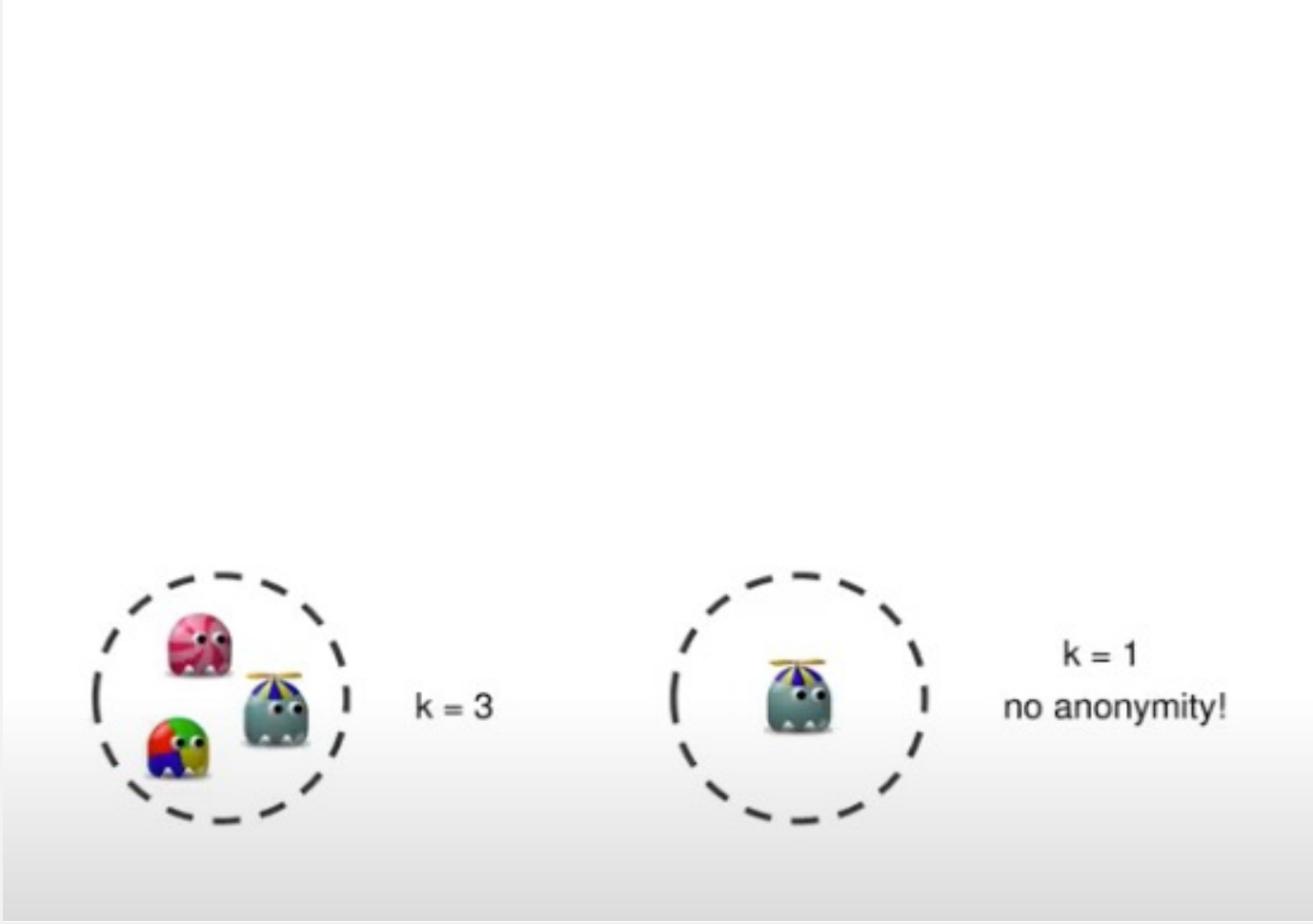
- We used these terms to form combinations and query the inverted index to see how many individuals could be identified
- For the individuals that were not identified we observed a high mismatch between information in Wikipedia and in Wikidata

	Total	# of identified
Train	411.623	401.517
Valid	51.498	50.611
Test	51.496	50.550
Total	*514.617*	*502.678*

Table 3: # of identified individuals

K-ANONYMITY FUNCTION

- Ensure that there are at least k individuals in the dataset that share the same attributes
- With k-anonymity, the individuals are basically 'hidden' be being part of a larger group.
- We decided on *k=5*



GREEDY FUNCTION

FOR summary

REPEAT

FOR combination in inverted index items combinations

DO intersection of posting lists

IF intersection > 5 **THEN CONTINUE**

ELSE

DO remove item with smaller posting list

Summary

Lorenzo Smith (born May 23, 1972) is an American singer-songwriter who has released three albums

Greedy Anonymized Summary

********** *(born* ********** *)* *is an* ********** *singer-songwriter who has released three albums.*

RANDOM FUNCTION

FOR summary

REPEAT

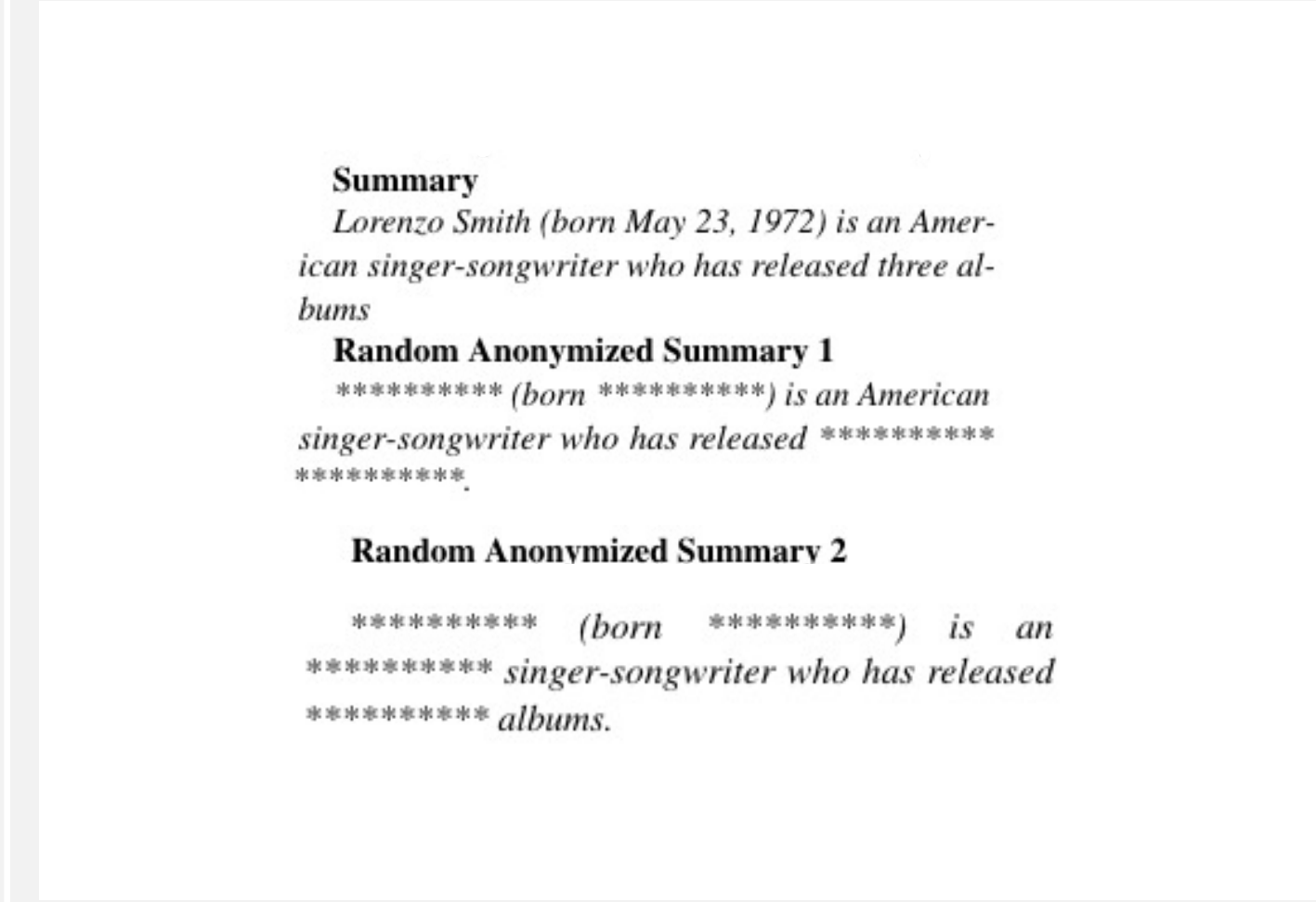
FOR combination in inverted index items combinations

DO intersection of posting lists

IF intersection > 5 **THEN CONTINUE**

ELSE

DO remove random item



Summary

Lorenzo Smith (born May 23, 1972) is an American singer-songwriter who has released three albums

Random Anonymized Summary 1

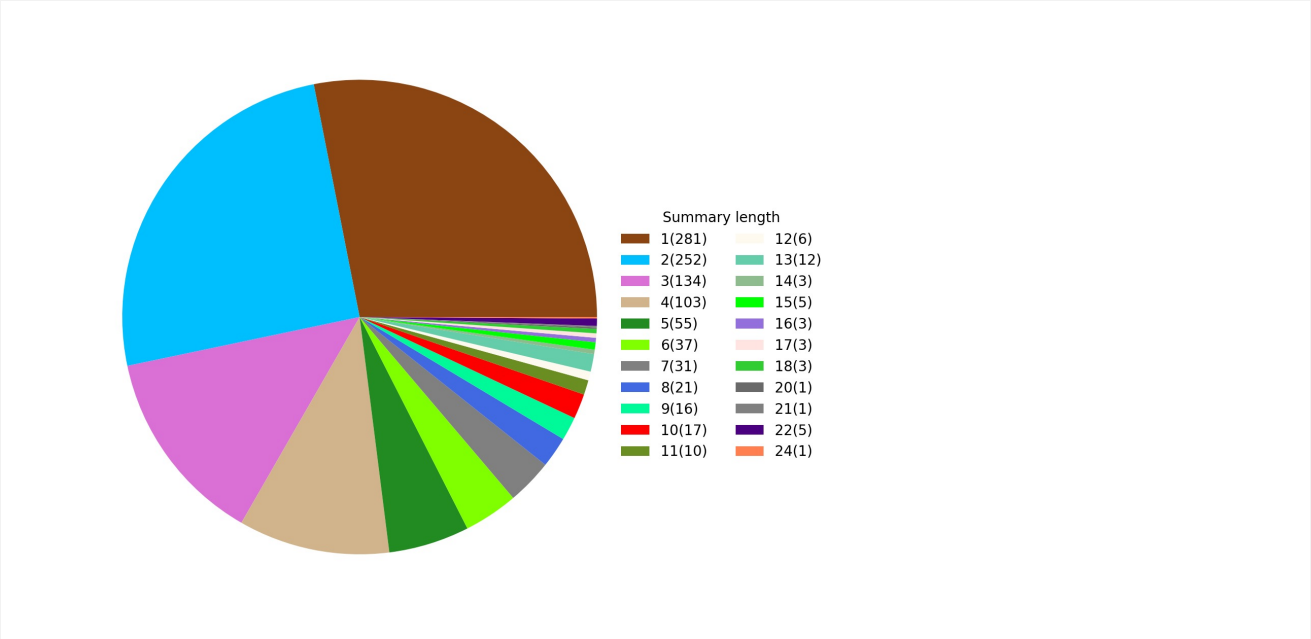
********** *(born* ********** *) is an American singer-songwriter who has released* ********** ********** .

Random Anonymized Summary 2

********** *(born* ********** *) is an* ********** *singer-songwriter who has released* ********** *albums.*

MANUAL ANNOTATION

- 1000 summaries
- Following the same distribution as the test dataset
- Some post-processing to make the text cleaner for the annotators
- SpaCy and index terms combined
- 100 annotated by 2 annotators, the rest split between them



MANUAL ANNOTATION

Task: Annotate this biographical text to conceal the identity of the main person: darrell griffith

Darrell Steven Griffith (born June 16, 1958), also known by his nickname Dr. Dunkenstein, is an American former basketball player who spent his entire professional career with the Utah Jazz of the National Basketball Association. He played collegiately at the University of Louisville. He is widely regarded as one of the greatest college basketball players of all time.

[Annotator: , Individual to protect: darrell griffith]

NB:This text output is just for reviewing purposes, do not annotate this file!
Task: Annotate this biographical text to conceal the identity of the main person: darrell griffith

********** (born **********), also known by his nickname **********, is an ******* former basketball player who spent his entire professional career with the ********** . He played collegiately at the ********** . He is widely regarded as one of the greatest college basketball players of all time.

[Annotator: , Individual to protect: darrell griffith]

MODEL TRAINING

- We used the automatic annotations from the greedy and the random functions to train two BERT models with a linear inference layer on top (GreedyBERT, RandomBERT)
- Following the metrics proposed in the SemEval-13 task 9, we calculated precision, recall, and F1-score on the entity level, with two different levels of strictness:
  - **Exact:**
    - The boundaries of the prediction match the boundaries of the true string in an exact manner.
  - **Partial:**
    - There is a partial match between the prediction and the true string regarding boundaries.

RESULTS

	Precision		Recall		F1-score	
	Exact	Partial	Exact	Partial	Exact	Partial
GreedyBERT - Dev set						
MASK	0.765	0.784	0.797	0.817	0.781	0.800
RandomBERT - Dev set						
MASK	0.769	0.788	0.786	0.805	0.777	0.796
GreedyBERT - Test set						
MASK	0.768	0.784	0.804	0.821	0.786	0.802
RandomBERT - Test set						
MASK	0.771	0.786	0.792	0.810	0.781	0.797

Table 5: Entity level evaluation of BERT models

RESULTS

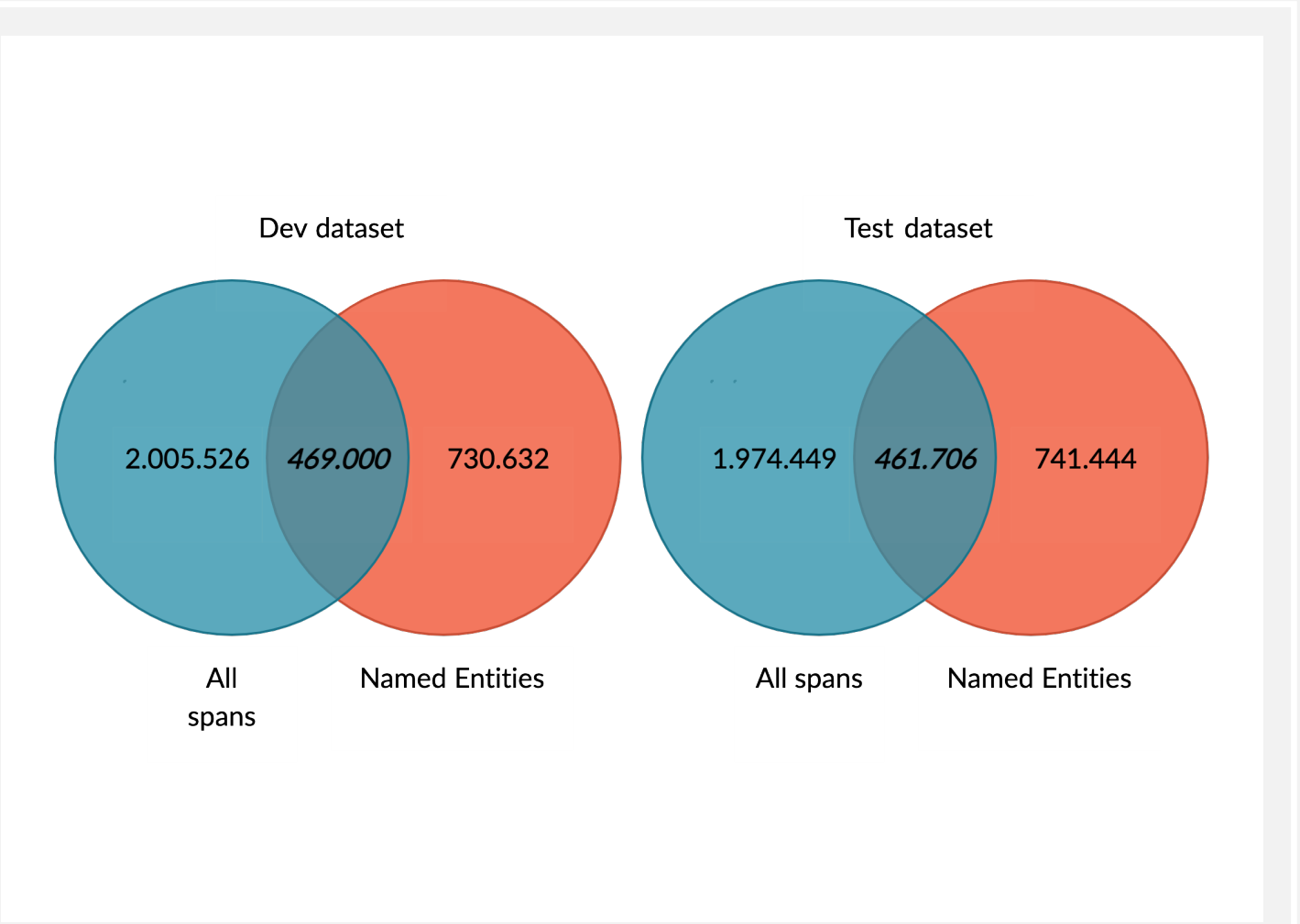
- The performance of a model trained on data derived from a process like the one we described is directly tied to the quality and quantity of information in the background knowledge
- The larger the KB, the less chance there is to miss a piece of information that could lead to identification of an individual
- Despite the overlap of information in Wikipedia and Wikidata, as well as the linguistic variance we decided to add to the KB, there was still some mismatch between them
  - This resulted in some of the sensitive entities missing or being partially annotated by the functions.
- *No 'right' answer*

RESULTS

- *Tom Jasper's* Wikipedia article starts with his full name '*Thomas D. Jasper …*'
- This exact information is not present in his Wikidata page, but the KB makes up for it by having general knowledge '*Thomas*' and 'Jasper' but not 'D.'
- Thus, instead of the entity being annotated as *Thomas D. Jasper*, it was annotated *Thomas D. Jasper*

SPACY'S NER

- We then decided to run SpaCy's NER on the development and test dataset to see whether named-entities only contained enough information to identify an individual, framing, thus the system as a sensitive information identifier



SPACY'S NER

- The low overlap, and the low scores suggest that masking NEs is not enough to protect and individual from being identified
- There are many direct and quasi identifiers that are not NEs, and could be used by an attacker

	Precision	Recall	F1-score
Dev set	0.634	0.233	0.340
Test set	0.631	0.232	0.339

Table 6: Standard evaluation for SpaCy's NER

TAB DATASET

- Since our models were trained on data from a generic domain, we decided to test our models on data from the legal domain
- ECHR court cases, manually annotated, high inter-annotator aggreement

	Precision		Recall		F1-score	
	Exact	Partial	Exact	Partial	Exact	Partial
			GreedyBERT			
Ent_MASK	0.094	0.108	0.488	0.563	0.157	0.181
Token_MASK	0.333	-	0.653	-	0.441	-
			RandomBERT			
Ent_MASK	0.093	0.109	0.486	0.573	0.156	0.184
Token_MASK	0.328	-	0.623	-	0.430	-

Table 7: Entity and token level evaluation on the ECHR cases

TAB DATASET

- The vocabulary, the length, and the types of direct identifiers one can find in a Wikipedia summary and a court case are vastly different, which is reflected in the performance of the models
  - For example, a type of direct identifier one can find in the TAB dataset are court case numbers, information which is never present in Wikipedia and Wikidata.
- The model's recall is not as low as one would expect, especially if we consider the token-level score instead of the entity level one, which is much higher.

PREDICTION EXAMPLES

- Looking at the actual output of the model we see that dates and names were entities the model managed to mask correctly, in most cases

Part of court case text:

... Mr Galip Yalman, Mr Bahattin Sarısoy, Mr Osman Çağlayan and Mr Yusuf Camca ... on 29 November 1996.

Annotation of text

[...] **********, **********, ********** and ********** ... on **********.

Model Prediction

[...] **********, **********, ********** and ********** ... on **********.

PREDICTION EXAMPLES

- We also observe a trend of over-masking

Part of court case text:

The applicants were represented by Mr S. Esmer, a lawyer practicing in Ankara.

Annotation of text

The applicants were represented by **********, a lawyer practicing in Ankara.

Model Prediction

The applicants were represented by ********** S. **********, a ********** practicing in **********.

PREDICTION EXAMPLES

- Despite seeing that the performance was not optimal, as was expected, by going through some of the examples we can see the usefulness of the approach in capturing generic sensitive information, even in texts of different domains.
- If one was to use this model in such cases, terms that went undetected and are domain specific are easily caught as a post-processing step (e.g. court case numbers).

CONCLUSION

- We propose a method of automatically annotating text data for sensitive information with explicit measures of disclosure risk, decided by our background knowledge
- NEs are not always relevant to disclosure risk
- Generic anonymization models can detect some personal information even in out-of-domain data