Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Data privacy & unstructured data = ?

Pierre Lison
plison@nr.no

27.04.2022

Finse Cyber-security Winter School

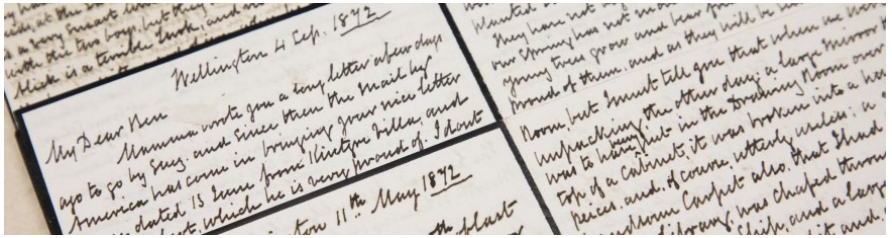
Outline

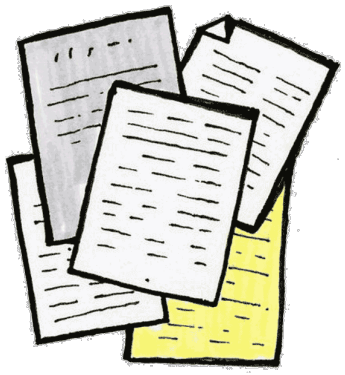
- What is unstructured data?
- Can unstructured data be anonymized?
- De-identification methods for text
  - Machine learning models
  - Concrete example
- De-identification methods for images and speech

Outline

- **What is unstructured data?**
- Can unstructured data be anonymized?
- De-identification methods for text
  - Machine learning models
  - Concrete example
- De-identification methods for images and speech

Unstructured data is everywhere





Industry estimates: up to 80% of the world's data in unstructured format (Gartner)

Unstructured data

= Umbrella term for various types of data that do not follow a predefined *data model*

- **Examples**: text documents, pictures, web pages, audio/video recordings, emails, etc.

←→ Often contrasted with **tabular data,** defined with *a fixed set of attributes*, where each attribute is associated with a *predefined range of possible values*

Note: Many datasets actually "semi-structured" (combining e.g. numeric attributes with free-form fields)

Unstructured data

	Person name	Date of birth	Gender	Nationality	Vaccination Status
1	Peter Higgs	30.07.1975	Male	British	2 shots
2	Andreas Sauner	02.10.1981	Male	German	No shot
3	Laurence Barrière	03.10.1957	Female	French	1st shot

VS

Peter Higgs, born on July 30, 1975, is a UK national and has already received 2 shots of the vaccine, while his German colleague Andreas Sauner, who will celebrate his 40th birthday on October 2, did not yet receive any shot. Meanwhile, their common acquaintance Laurence Barrière recently got her first vaccine shot. Mrs. Barrière is French and will turn 64 years old on October 3.

Outline

- What is unstructured data?
- **Can unstructured data be anonymized?**
- De-identification methods for text
  - Machine learning models
  - Concrete example
- De-identification methods for images and speech

International Data Privacy Law, 2022, Vol. 00, No. 0

The GDPR and unstructured data: is anonymization possible?

Emily M. Weitzenboeck*, Pierre Lison**,
Malgorzata Cyndecka***, and Malcolm Langford***

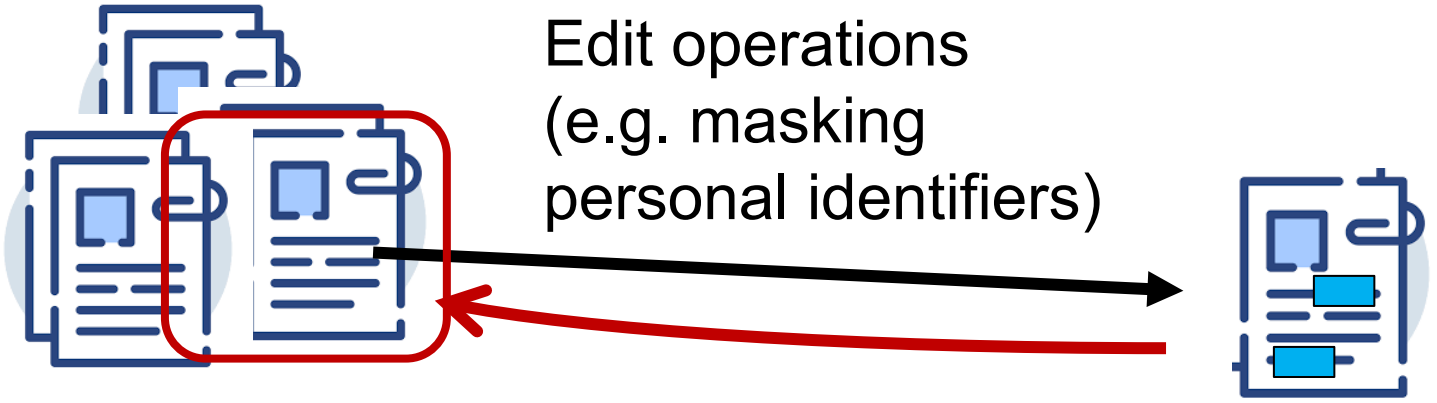
Our answer: it depends on how one interprets GPDR!

- *Strict interpretation*: no, unless the original data is deleted
- *Risk-based interpretation*: difficult, but possible

Anonymisation of unstructured data

Main problem: requirement of unlikability with original dataset

➔ If one has access to the original data, it is quite easy to do a *phrase search* to find back the original document



A simple experiment:

1. The applicant [Mr Colin Joseph O'Brien] was born in 1955 and lives in Bridgend.

2. His wife died on 29 April 1999 leaving two children, born in 1989 and 1991.

3. In 1999 the applicant enquired about widows' benefits and he was informed that he was not entitled to such benefits.

4. In early 2000 the applicant applied for widows' benefits again and on 13 March 2000 the Benefits Agency rejected his claim.

5. He lodged an appeal against this decision on 16 March 2000 and this appeal was struck out on 23 May 2000 on the basis that it was misconceived.

6. On 16 May 2000 the applicant made an oral claim for Widow's Bereavement Allowance to the Inland Revenue. On 23 May 2000 he was informed that his claim could not be accepted because there was no basis in domestic law allowing widowers to claim this benefit. The applicant was advised that an appeal against this decision would be bound to fail.

7. The applicant received child benefit in the sum of GBP 100 per month.

[excerpt from a court case from the ECHR, nr. 61391/00]

After masking (quasi-)identifiers

1. The applicant [***] was born in *** and lives in ***

2. His wife died on *** leaving *** children, born in ***

3. In *** the applicant enquired about widows' benefits and he was informed that he was not entitled to such benefits.

4. In *** the applicant applied for widows' benefits again and on *** the *** rejected his claim.

5. He lodged an appeal against this decision on *** and this appeal was struck out on *** on the basis that it was misconceived.

6. On *** the applicant made an oral claim for Widow's Bereavement Allowance to the Inland Revenue. On *** he was informed that his claim could not be accepted because there was no basis in domestic law allowing widowers to claim this benefit. The applicant was advised that an appeal against this decision would be bound to fail.

7. The applicant received child benefit in the sum of *** per month.

[Masking decisions made by law students from UiO]

Is it anonymous?

1. The applicant [***] was born in *** and lives in ***

2. His wife died on *** leaving *** children, born in ***

3. In *** the applicant enquired about widows' benefits and he was informed that he was not entitled to such benefits.

4. In *** the applicant applied for widows' benefits again and on *** the *** rejected his claim.

5. He lodged an appeal against this decision on *** and this appeal was struck out on *** on the basis that it was misconceived.

6. On *** the applicant made an oral claim for Widow's Bereavement Allowance to the Inland Revenue. On *** he was informed that his claim could not be accepted because there was no basis in domestic law allowing widowers to claim this benefit. The applicant was advised that an appeal against this decision would be bound to fail.

7. The applicant received child benefit in the sum of *** per month.

Combination of "*rejected his claim*" and "*could not be accepted*" also occurs one in the full dataset

"*was advised that an appeal against*" appears only once in a collection of 13,759 court cases!

Full anonymisation

1. The applicant [***] was born in *** and lives *** ***

2. *** *** *** *** *** *** two *** *** *** *** *** *** ***

3. In *** *** *** *** *** *** *** *** *** *** *** *** *** was *** *** *** *** *** ***

4. In *** the applicant *** *** *** *** *** *** *** *** the *** *** his *** ***

5. *** *** an *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** the *** that it was *** ***

6. *** *** *** *** *** *** *** *** for *** *** *** *** the *** *** *** *** *** *** *** *** *** *** *** could *** *** *** *** *** *** *** *** in *** law *** *** to *** this *** *** *** *** *** *** *** *** *** *** this *** *** *** *** to *** ***

7. The *** *** *** *** in the *** *** *** *** *** ***

Theoretically possible to remove all unique phrases… but the result is worthless

Text anonymisation

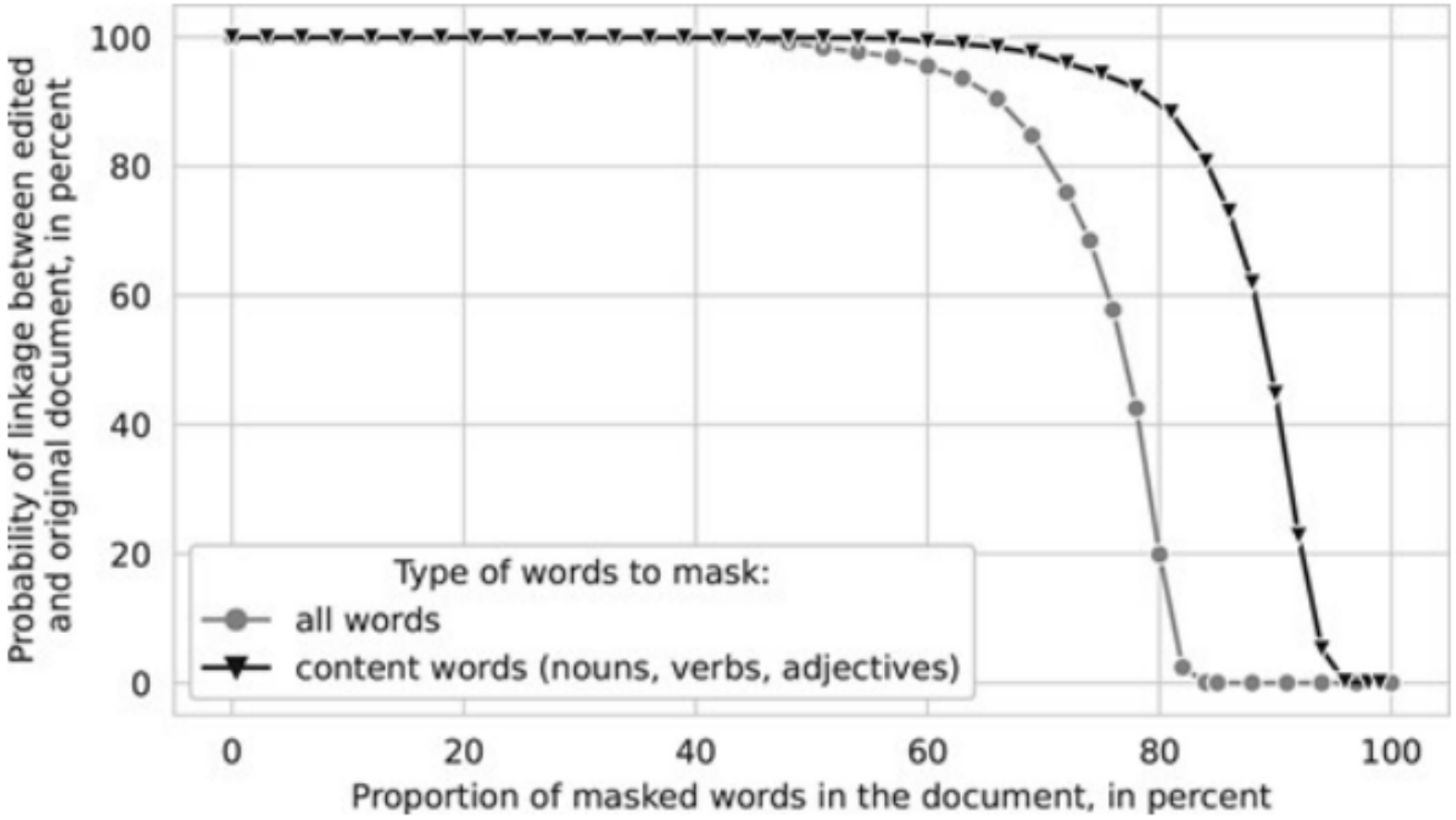


Image anonymisation

- Experiment with 7570 chest X-ray images from the US National Library of Medicine
- Not possible to recognize a person directly from the X-ray … but the hospital where the X-ray was taken has access to the patient name
- Can we “distort” the image to ensure non-linkability?

Image anonymisation

Can we “distort” the image to ensure non-linkability?

➔ In theory, yes, for instance by introducing artificial noise

… but the image is essentially destroyed in the process:

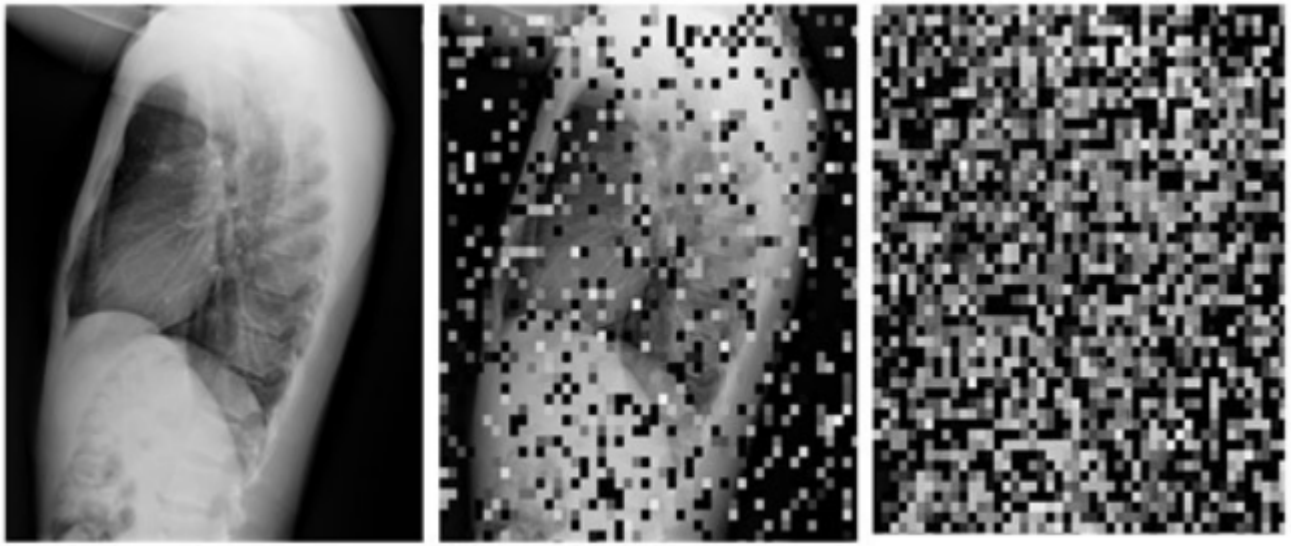
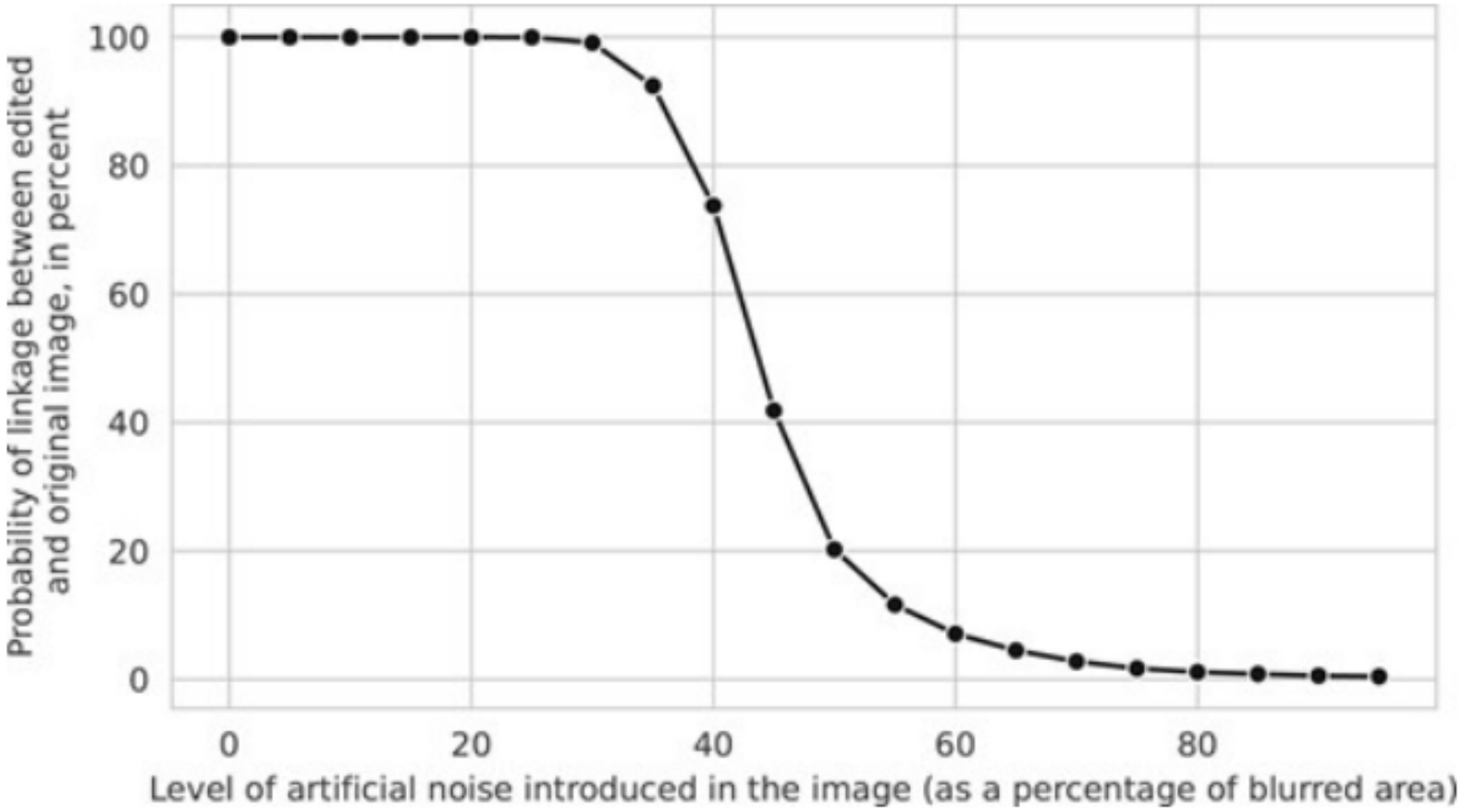


Image anonymisation



Outline

- What is unstructured data?
- Can unstructured data be anonymized?
- **De-identification methods for text**
  - **Machine learning models**
  - **Concrete example**
- De-identification methods for images and speech

De-identification methods

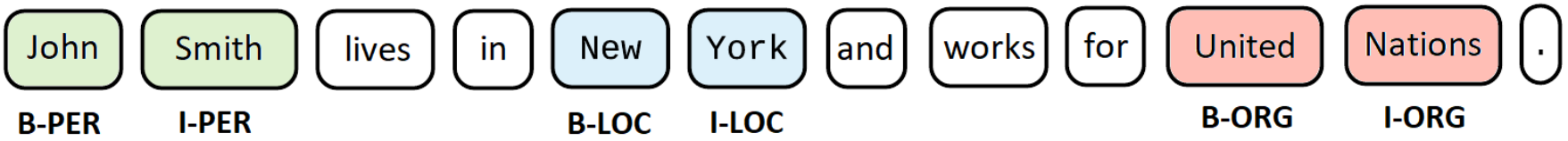
- Simplest approach: **handcrafted patterns**
  - Regex to detect numbers, dates etc
  - Gazetteers to detect occurrences of specific words/phrases (compiled in a list)

Challenges:

- Difficult to cover all types of entities
- Cannot handle ambiguities (e.g. «Stein»)

De-identification methods

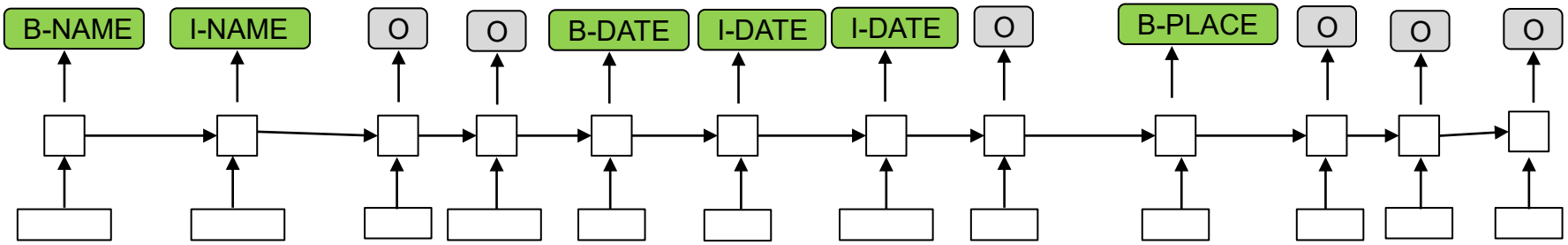
- Nowadays, data-driven approaches (often based on deep neural networks) are dominant
- De-identification as a **sequence labelling problem**:



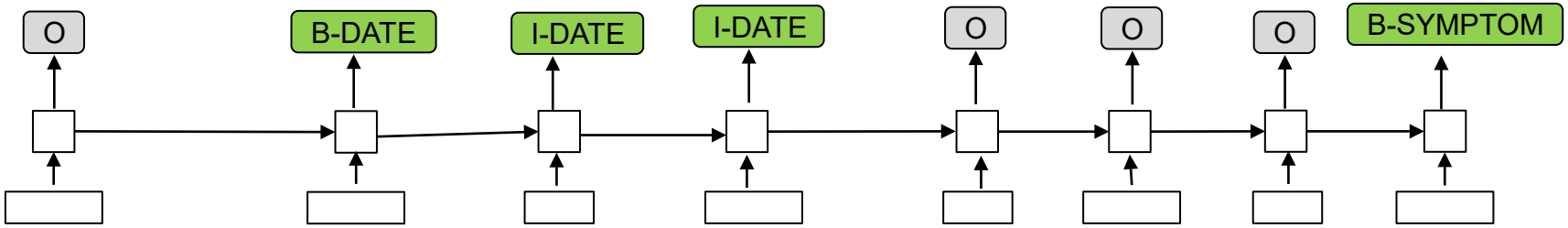
- **BIO scheme**: **B**(eginning), **I**(nside) or (**O**)ut of an entity
- Models must be trained from (labelled) data

Sequence labelling

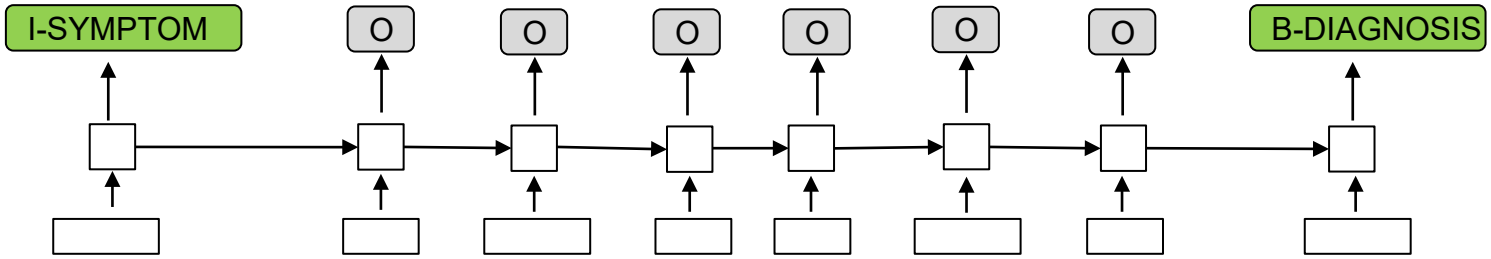
- **B**eginning
- **I**nside
- **O**ut



Ola Normann , født 2. april 1978 i Drammen , kom til



akuttmottaket mandag 5. september på grunn av sterke



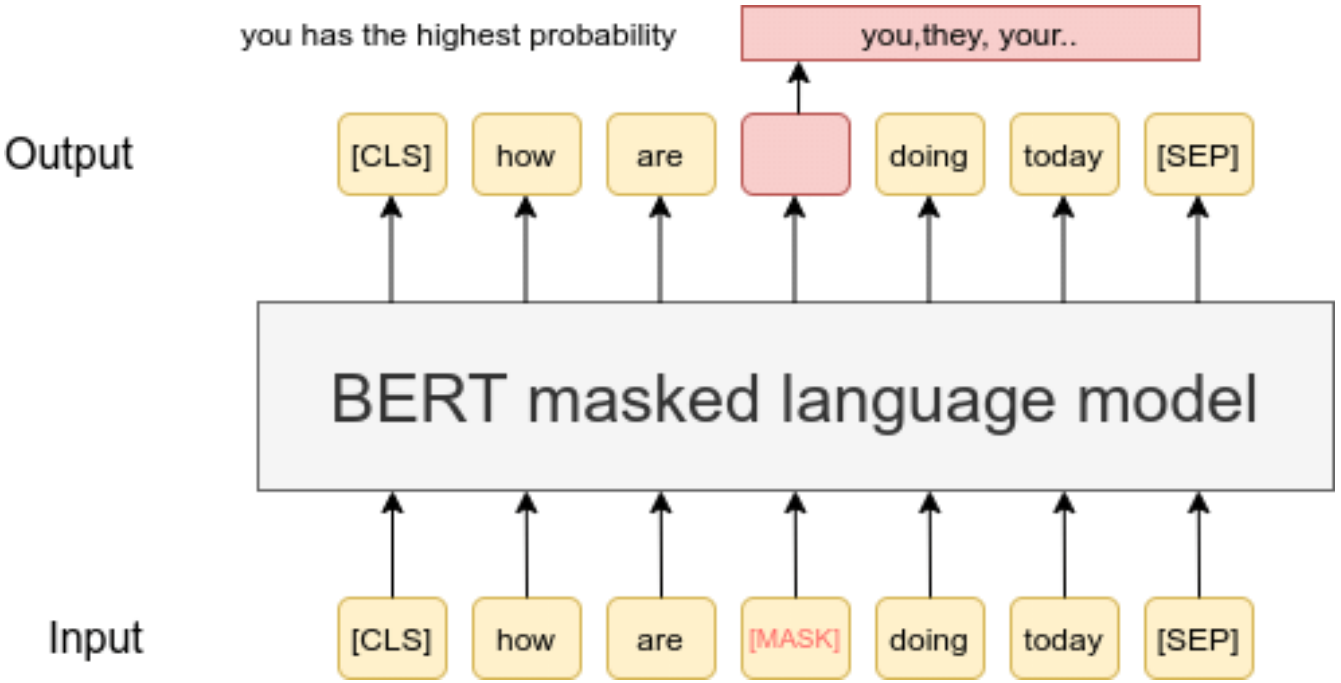
brystsmerter som viste seg å være en blodpropp….

Sequence labelling

[here, do some live coding to show how to use Spacy]

Neural language models

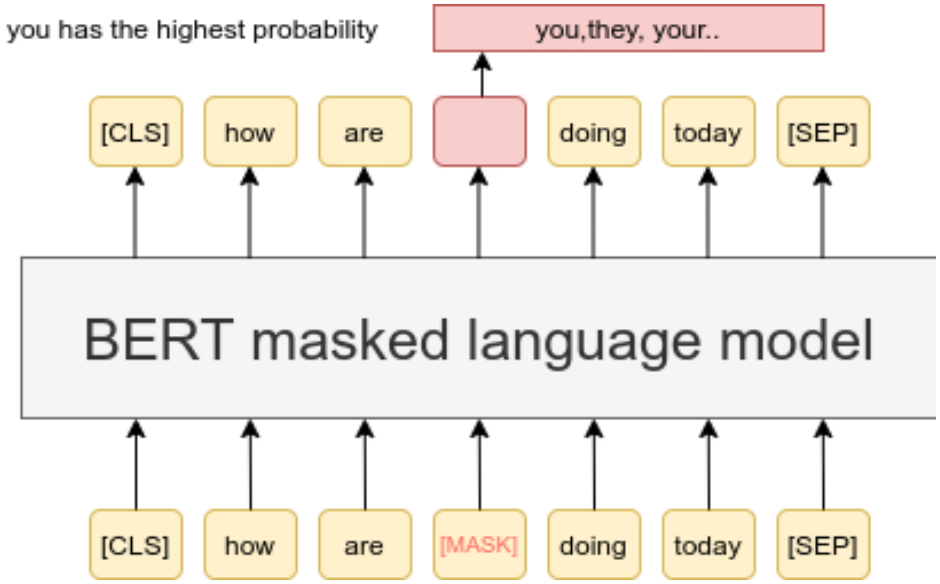
The best sequence labelling models are typically not learned “from scratch”, but **fine-tuned** from existing, large language models like BERT, GPT3 or T5



Neural language models

= Large neural models, with dozens of layers and billions of parameters

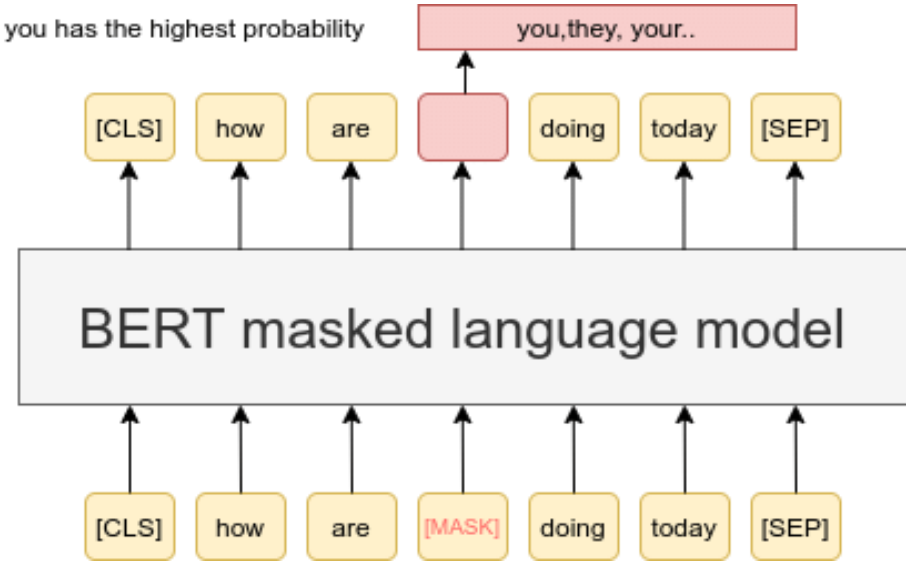
Each token represented by a numerical vector (=embedding)



- Those word vectors are computed through multiple transformer layers where each token can be «influenced» by its neighbours using an attention mechanism
- **Task**: predict missing words!

Fine-tuning

- When doing fine-tuning of an existing model, we remove the top layer (specific to the word prediction task), but keep the other layers
- We then add a new layer specific to our problem (like detecting named entities), and train this model on domain-specific data



Fine-tuning

[here, do some live coding with simpletransformers]

Pros & cons of sequence labelling approaches (for data privacy)

+ Good performance for detecting entities likes names, places, organisations, etc.

+ Can take context into account (to e.g. resolve ambiguities)

But:

- Does not remove enough (typically limited to predefined categories of entities, like named entities)

- May remove too much (remove all entities without considering the actual disclosure risk)

NLP for other privacy-enhancing tasks

- NLP models can also be used to *obfuscate* specific demographic attributes of the author (like gender or ethnicity) from texts

- Or produce *synthetic* texts (based on an original corpus) with privacy guarantees
- Or train neural language models that do not leak personal data from the training set

[For details, see e.g. Lison et al, "*Anonymisation Models for Text Data: State of the Art, Challenges and Future Directions*", ACL 2021]

PPDP methods

Outside NLP, the field of privacy-preserving data publishing (PPDP) has also developed several text de-identification methods

Inputs:

- Document *d* (represented as collection of terms)
- Individuals/entities *C* to protect in *d*
- Background knowledge *K*

Output:

Edited document *d'* such that the remaining terms no longer identify anyone from *C*

Chakaravarthy et al. (2008), Cumby & Ghani (2011), Anandan et al. (2012), Sánchez and Batet (2016, 2017)

C-sanitize

Sánchez and Batet (2016, 2017)

Inputs:

- Document *d* (represented as collection of terms)
- Individuals/entities *C* to protect in *d*
- Background knowledge *K*

Output:

Edited document *d'* such that the remaining terms no longer identify anyone from *C*

- Information-theoretic approach based on pointwise mutual information (PMI)
- PMI estimated from web occurrence counts (background knowledge = "the web")

PPDP methods

+ Explicit account of *disclosure risk* based on a privacy model (often k-anonymity)
+ Not limited to predefined entity types

But:

- Document reduced as a “bag of term”
- Restricted types of semantic inferences
- Scalability issues

CLEANUP project

CLEANUP:

Machine Learning for the Anonymisation of Unstructured Personal Data

Goal: reconcile NLP and PPDP approaches to text anonymization!

General procedure:

1. Use a neural model to detect personal (quasi-)identifiers
2. Compute various estimators of disclosure risk
3. Search for set of **edit operations** (mask or generalise) that can ensure a disclosure risk below a given threshold, yet minimize the semantic loss

The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for Text Anonymization

Ildikó Pilán*
Norwegian Computing Center,
Oslo, Norway

Pierre Lison*
Norwegian Computing Center,
Oslo, Norway

Lilja Øvrelid
Language Technology Group,
University of Oslo, Norway

Anthi Papadopoulou
Language Technology Group,
University of Oslo, Norway

David Sánchez
Universitat Rovira i Virgili, CYBERCAT,
UNESCO Chair in Data Privacy, Spain

Montserrat Batet
Universitat Rovira i Virgili, CYBERCAT,
UNESCO Chair in Data Privacy, Spain

+ 12 law students involved in the annotation (about 1000 hours in total!)

1278 court cases from the ECHR annotated for personal information:

- Semantic type
- Masking decision
- Confidential attributes
- Co-reference relations

The TAB corpus

We also propose *new evaluation metrics* dedicated to text anonymization

Why? Three reasons:

1. Not all personal identifiers are equally important to mask!
2. A (direct or indirect) identifier is only «protected» if all its occurrences are masked
3. Multiple *possible solutions* to a given anonymisation problem

Outline

- What is unstructured data?
- Can unstructured data be anonymized?
- De-identification methods for text
  - Machine learning models
  - Concrete example
- **De-identification methods for images and speech**

De-identification of images

Large range of personal data “expressed” in images

De-identification of images

- The detection of human faces using deep neural nets can now be done with high accuracy
  - Once the bounding box of a face is extracted, one can easily apply a blurring filter to it
- Same for specific items such as vehicle registration plates
- But more indirect «cues» (clothing etc.) harder to handle

De-identification of images

[here, do some live coding with face_recognition]

De-identification of recordings

- Speech is by definition personal data: one can recognize a person by their voice
- It is possible to distort the voice of an individual to make it harder to identify
- But it does **not** make the recording anonymous:
  - Individual speaking patterns, word choices etc
  - Content in what is being said

Questions, comments?