The importance and pitfalls of pseudonymization

Lisa Becker

Machine Learning Engineer

Working Group Lead - Speech / Audio

/lisabecker

/becker-lisa

ML6

“87% of the U.S. population is uniquely identified by
📅 date of birth
👤 gender
🏙️ postal code.”

(Latanya Sweeney, 2000)

“> 99% of U.S. population is uniquely identified by 15 random quasi-identifiers in any dataset.”

(Rocher et al., 2019)

GDPR

ML6

 |

General Data Protection Regulation in a nutshell

- Enforceable since 2018
- Regulates EU/EEA law on data protection and privacy
- **Goal**: Enhancing the individual’s control and rights over their **personal data**:
- “**Any information** from which a person (a data subject) **can be identified or potentially identified**” needs to be pseudonymized, for example:

Names, nicknames, ID numbers, location, physical, physiological, genetic, mental, economic data, or cultural or social identity

- Exceptions:
  - Explicit consent, social security & protection, substantial public interest, trade unions or religions, doctors, courts or lawyers
  - If identifiable information is permanently removed
- **GDPR does not prescribe pseudonymization technique**

Difference between pseudonymization and anonymization

Pseudonymization:
data *can be <u>re-identified</u>* with the help of an identifier (=additional information)
→ stays personal data

Anonymization:
<u>permanent replacement</u> of sensitive data with unrelated characters
→ no personal data anymore

On this day, 17th of April 2021
Before me, Nota spqyuayeahre ,

Appeared: Abe Ross, John Oliver
born on 12th of December 1975, …

On this day, 17th of April 2021
Before me, Notary Danny McGraw,

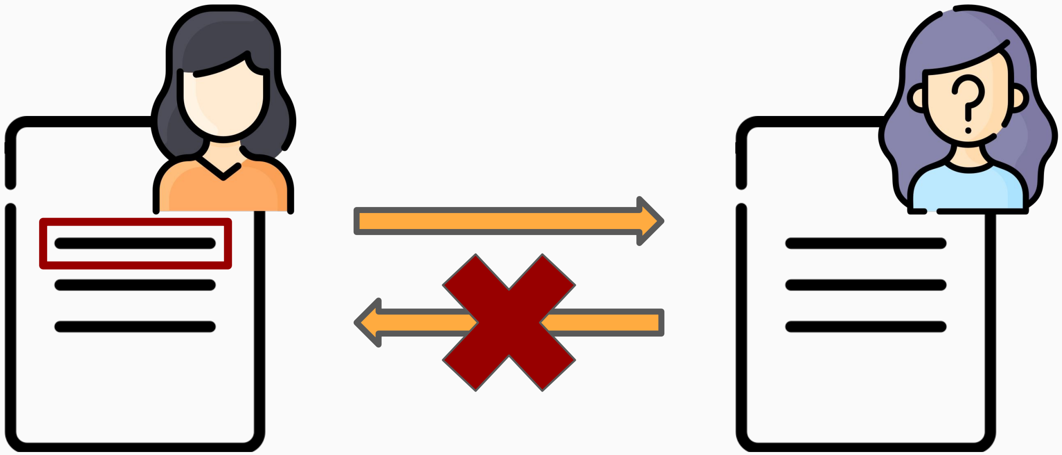
Appeared: John Oliver
born on 12th of December 1975, …

3 pillars of pseudonymization

What to pseudonymize 🔍

How to pseudonymize 🎲

Averting attacks ⚠️



ML6

What to pseudonymize 🔍

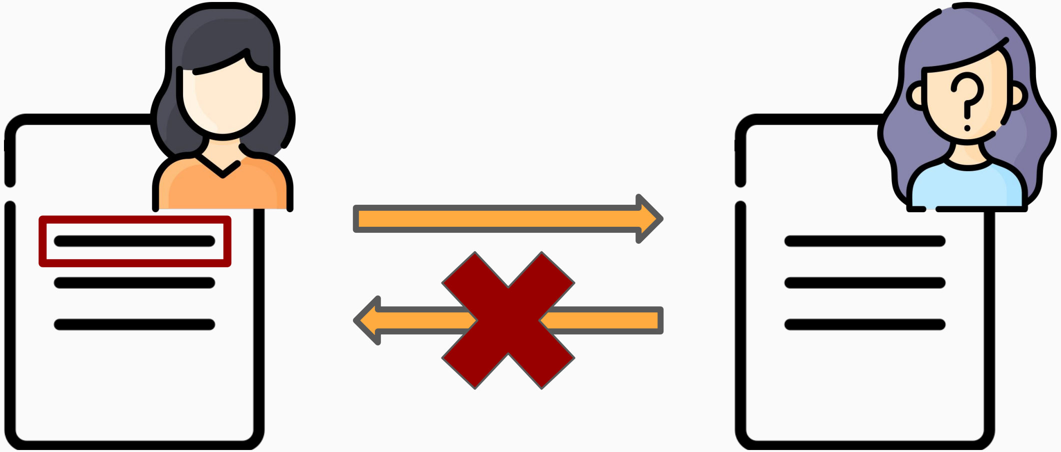
- **RegExes**:
  - E-Mail-Addresses
  - Phone numbers
  - Date / Time
  - Events / Companies / … ?
  - Names?
  - Addresses? → different countries?
- **Models**
- **Combination of both (Hybrid ML)**

3 pillars of pseudonymization

What to pseudonymize 🔍 ✔

How to pseudonymize 🎲

Averting attacks ⚠️



How to pseudonymize 🎲

Single identifier pseudonymization

- Counter
- Random number generators
- Cryptographic hash functions
- Message authentication code (MAC)

How to pseudonymize 🎲

Pseudonymization policy

- Deterministic pseudonymization (same across documents)

On this day, 17th of April 2021
Before me, Notary Danny McGraw,
John Oliver
Appeared: Abe Ross, born on 12th of December 1975,
John Oliver
Abe Ross declares to have sufficient funds.

On this day, 27th of March 1995
Before me, Notary Julien Schuermans,
John Oliver
Appeared: Abe Ross, born on 12th of December 1975,

To buy the property, located at 123 Fake Street, Phoenix, for the agreed upon price of €125.000.

ML6

How to pseudonymize

Pseudonymization policy

- Deterministic pseudonymization (same across documents)
- **Document-randomized pseudonymization (same within document)**

On this day, 17th of April 2021
Before me, Notary Danny McGraw,
John Oliver
Appeared: Abe Ross, born on 12th of December 1975,
John Oliver
Abe Ross declares to have sufficient funds.

On this day, 27th of March 1995
Before me, Notary Julien Schuermans,
Tim Esser
Appeared: Abe Ross, born on 12th of December 1975,

To buy the property, located at 123 Fake Street, Phoenix, for the agreed upon price of €125.000.

How to pseudonymize 🎲

Pseudonymization policy

- Deterministic pseudonymization (same across documents)
- Document-randomized pseudonymization (same within document)
- **Fully-randomized pseudonymization (never same)**

On this day, 17th of April 2021
Before me, Notary Danny McGraw,
John Oliver
Appeared: Abe Ross, born on 12th of
December 1975,
Tim Esser
Abe Ross declares to have sufficient funds.

ML6

How to pseudonymize 🎲

Other problems:

- Gender
  - Coreference
  - E-Mail-Addresses
- Scanned documents:
  - OCR errors: L1sa → might not be identified as name
- Black boxes instead of text
  - Missing information
  - Length of original data known

On this day, [redacted]

Before me, [redacted]

Appeared: [redacted] born on 12th of December 1975,

To buy the property, located at [redacted] for the agreed upon price of [redacted]

→ **privacy versus utility**

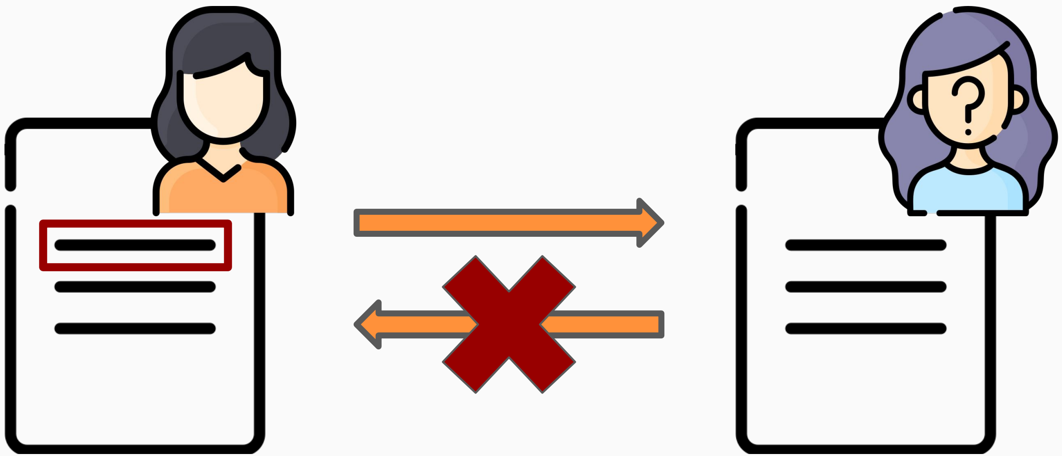
ML6

3 pillars of pseudonymization

What to pseudonymize 🔍 ✔

How to pseudonymize 🎲 ✔

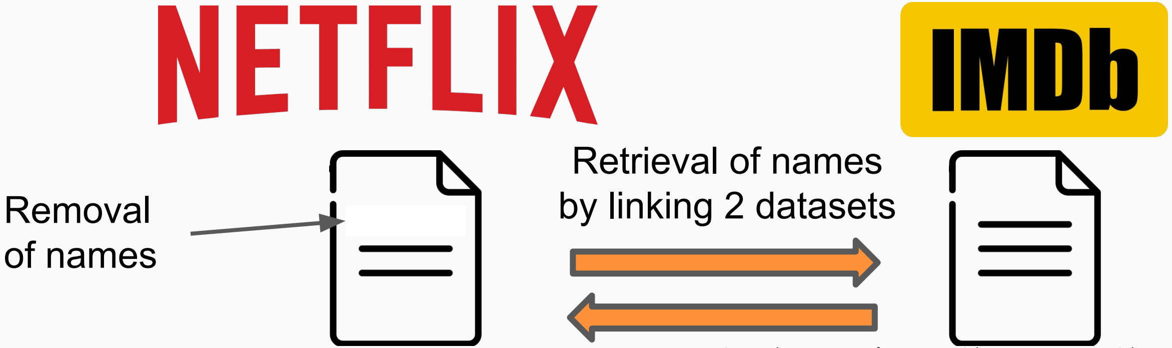
Averting attacks ⚠️



Averting attacks ⚠️

Linkage Attacks

- Re-identification
- Combining data by linking multiple datasets
- Quasi-identifiers: Pieces of information that aren’t themselves unique identifiers but become so through combination
- Example:



Averting attacks ⚠️

k-anonymity:

Quasi-identifiers have to reach **k-anonymity** through transformation

- Even with auxiliary information, each individual is still **indistinguishable from at least k-1 other individuals**

2 common methods:

- **Suppression**: Replacement of values of certain attributes with the same value (like nationality through *)
- **Generalization**: Replacement of values of certain attributes with broader category (like numbers through number ranges: 28 through <30)

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Averting attacks ⚠️

Table is 4 anonymous (zip-code, age, nationality):

- For any combination of these attributes, there are at least 3 rows with those exact attributes.

Other attacks against k-anonymity:

- **Homogeneity Attack**:

Attacker knows that Bob is admitted to hospital (31 y/o in 13053). Bob's record number is: 9, 10, 11 or 12.
All patients have same condition.
Conclusion: Bob has cancer.

- **Background Knowledge Attack**:

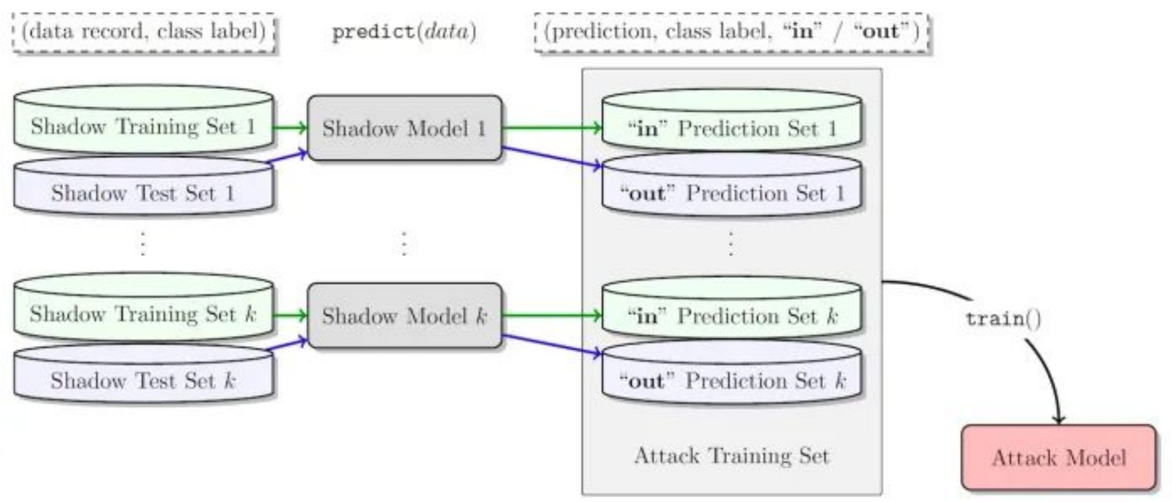
Attacker knows that Umeko (🇯🇵, 21 y/o in 13068) is at same hospital and **heart diseases are rare in Japan**.
Conclusion: Umeko probably has a viral infection.

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Averting attacks ⚠️

Membership Inference Attacks

- Attacker knows the model's algorithm and architecture **or service used to create the model**
- **Goal**: Observing the behavior of target models: **prediction of input data**
- Training of 'shadow models' to predict whether sample was part of model's training data
- Predictions of 'shadow model(s)' used to train membership inference attack model



Pseudonymization is hard

What to pseudonymize 🔍 ✓

How to pseudonymize 🎲 ✓

Averting attacks ⚠️ ✓

There is no 'one-size-fits-all' approach:

- Privacy versus utility
- Depends on the use case

Pseudonymization use case

Regexes

List of geolocations

Random number generator

On this day, 17th of April 2021
Before me, Notary Danny McGraw,

Appeared: Abe Ross,
born on 12th of December 1975,
To buy the property,
located at 123 Fake Street, Phoenix, for
the agreed upon price of €125.000.

On this day, **10th of January 1996**
Before me, Notary Danny McGraw,

Appeared: Abe Ross,
born on 12th of December 1975,
To buy the property, located at
EvenFakerStreet 987, **New York**, for the
agreed upon price of €**285.000**.

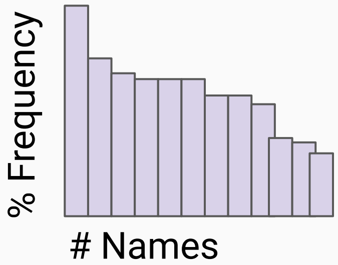
Pseudonymization use case

Long list of names

For **recognizing** as many names as possible

For **replacing** with names **common in Belgium**

Short list of names



Name of

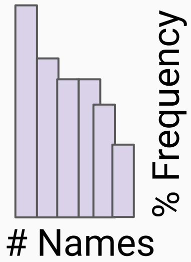
0.42% frequency

Before me,
Notary Danny McGraw.

Before me,
Notary Bart Derudder.

Name of

0.42% frequency



Pseudonymization use case

List of names 🧍

Ik heet Ben.

My name is Ben.

Je suis madame Le.

I'm mrs. Le.

Pseudonymization use case

List of names

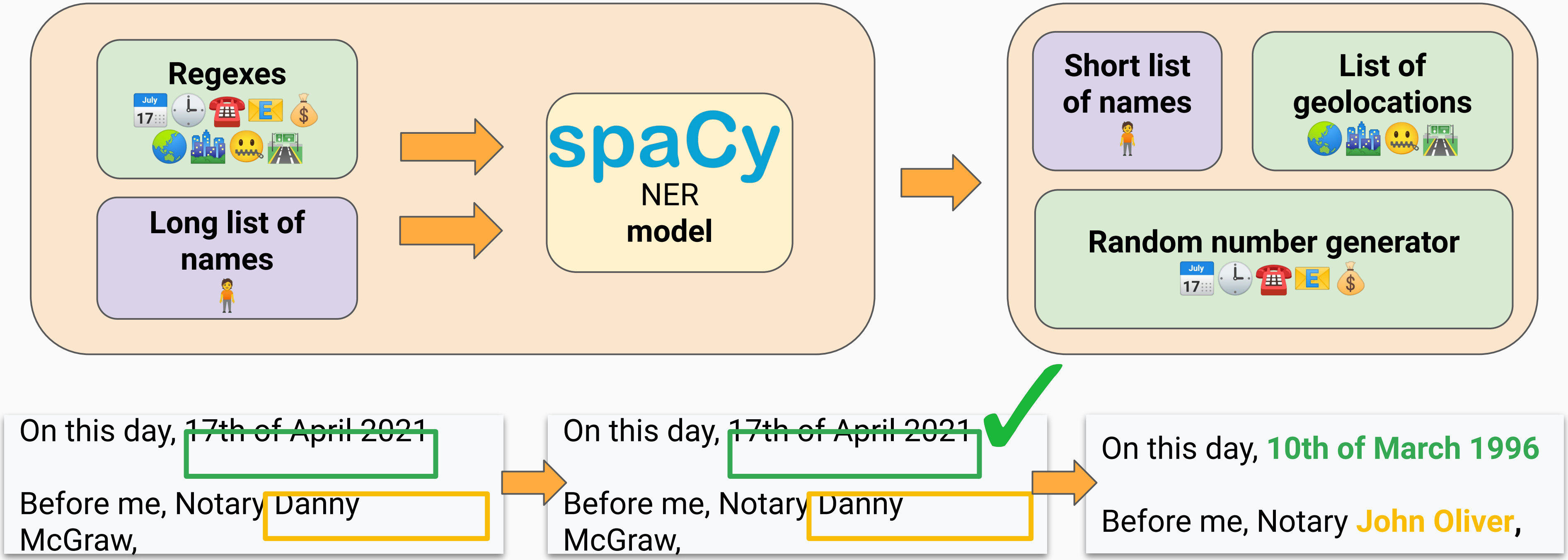
spaCy
NER **model**

→ names shorter than 4 letters not captured by RegEx but NER model.

Ik ben Lisa.
I am Lisa.

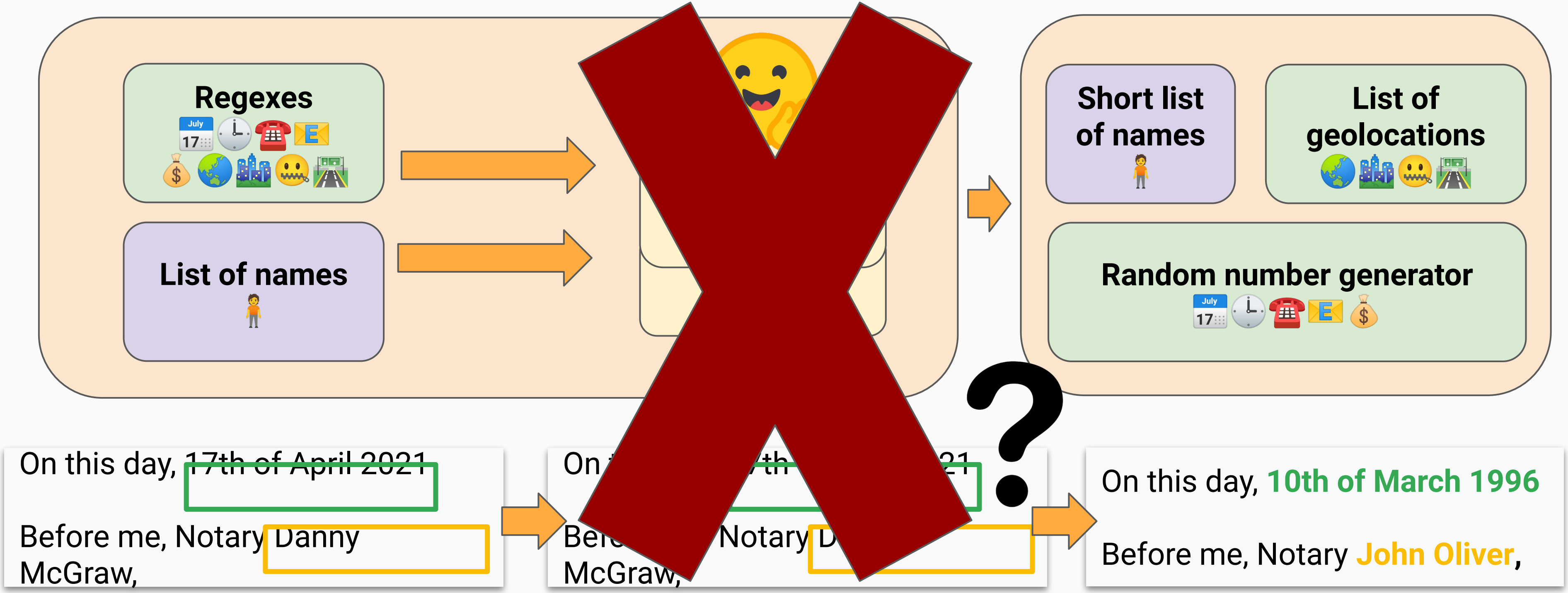
Le chien est mignon.
The dog is cute.

Pseudonymization use case



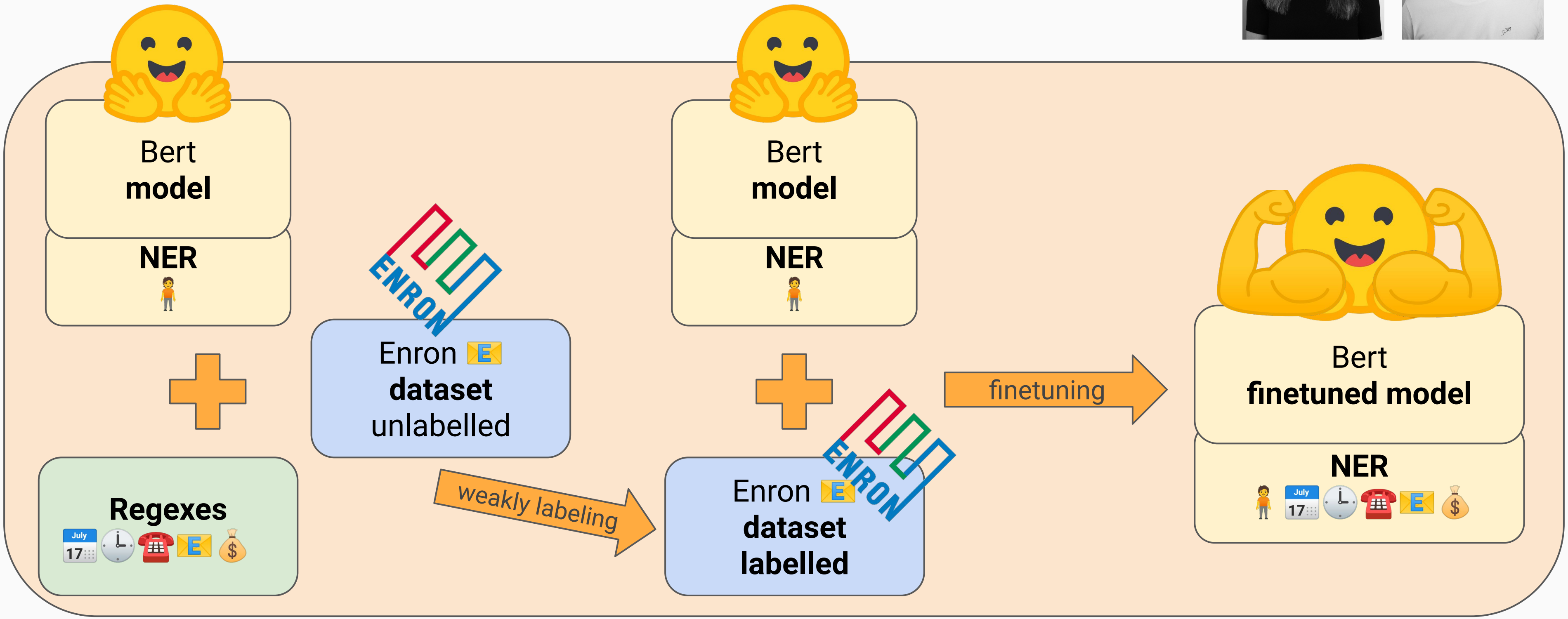
Anonymization use case.

~~Annanymization~~.



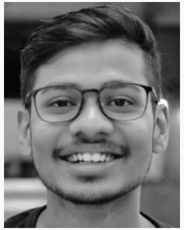
ML6

Pseudonymization demo.



Address NER Model.

(e.g. to improve pseudonymization demo)



The first of its kind **open sourced!**

Bert
model

NER

\+

Ultra-Fine
Entity Typing
dataset

finetuning

Bert
finetuned model

🌏🏙️

improvements
for the future

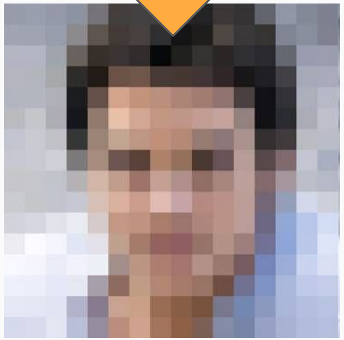
🌏🏙️🤐🛣️

Anonymization of video

Input video

Object detection **Model**

Deepstream Video processing



Recap 💡

Scan me to give Lisa feedback!🙏
Or visit: s.truqu.com/0O7zxn

Importance:

- Pseudonymization/anonymization is important for the individual's privacy and safety
- GDPR regulates data privacy
- ML6 decides on use-case basis whether data is pseudonymized or anonymized
  → if pseudonymization: adhere to GDPR
  (less safe, more useful)
  → if anonymization: freedom!
  (more safe, less useful)

Pitfalls:

- Use-case-specific
- Trade-off: Privacy versus utility
- 3 pillars of pseudonymization:
  - What to pseudonymize 🔍
  - How to pseudonymize 🎲
  - Averting attacks ⚠️
- Never 100% rock-solid

In EU highly regulated on paper, but difficult in practice!