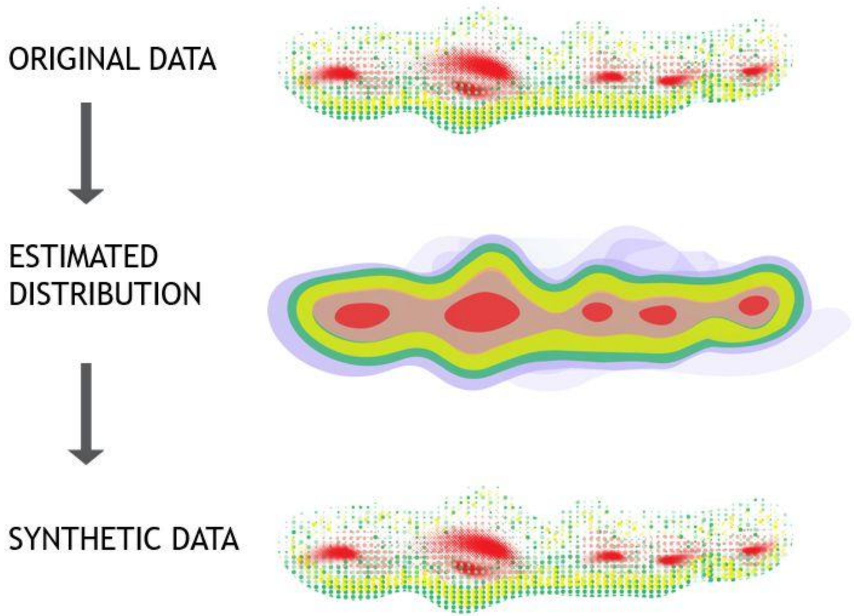
clearbox AI

Torino, 20/05/2024

Preserving the Privacy of Personal Data: A Practical Approach with Nerpii

humans && machines

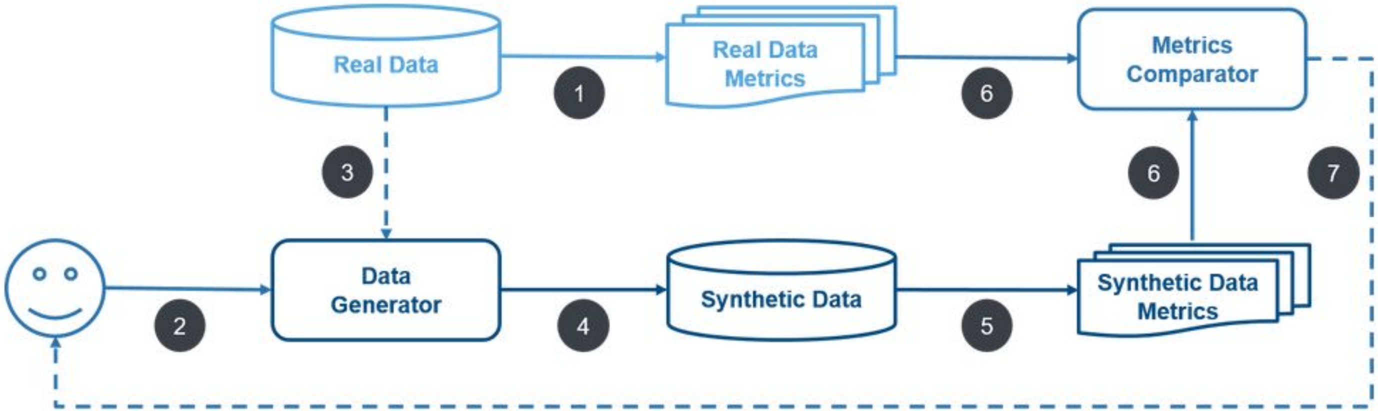
What are Synthetic Data?



- Synthetic data is artificially generated using artificial intelligence algorithms on real data samples.

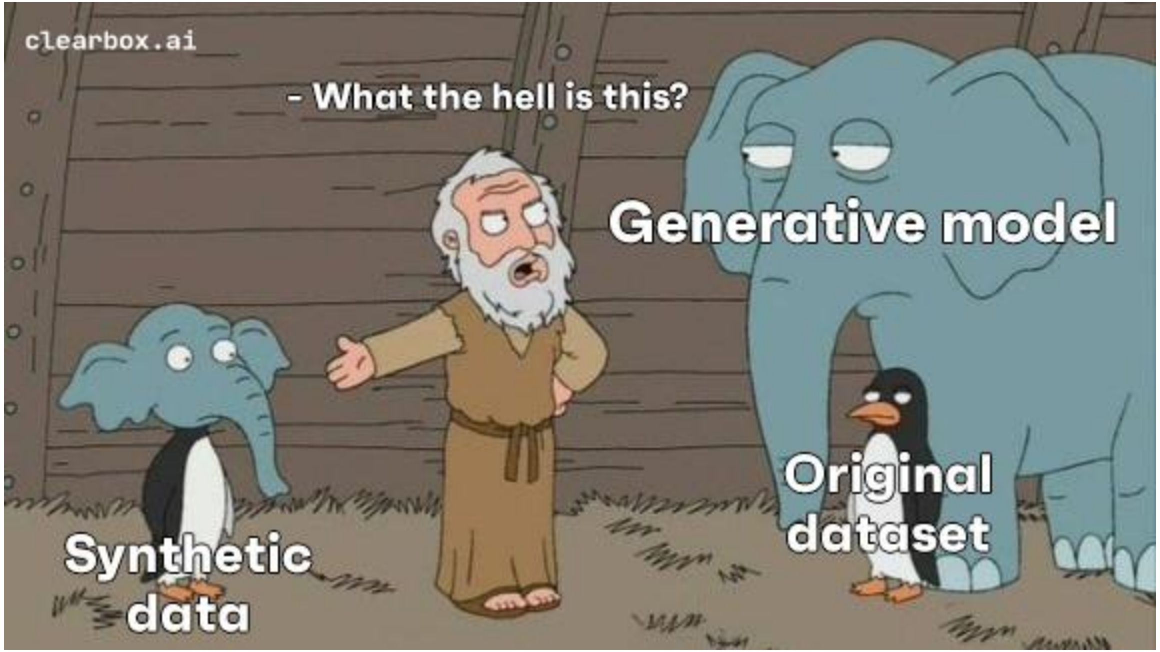
- They possess the same statistical properties and predictive capabilities as the real data from which they were generated.

How are synthetic data generated?



Source: J. P. Morgan https://www.jpmorgan.com/technology/artificial-intelligence/initiatives/synthetic-data

Synthetic Data Generation

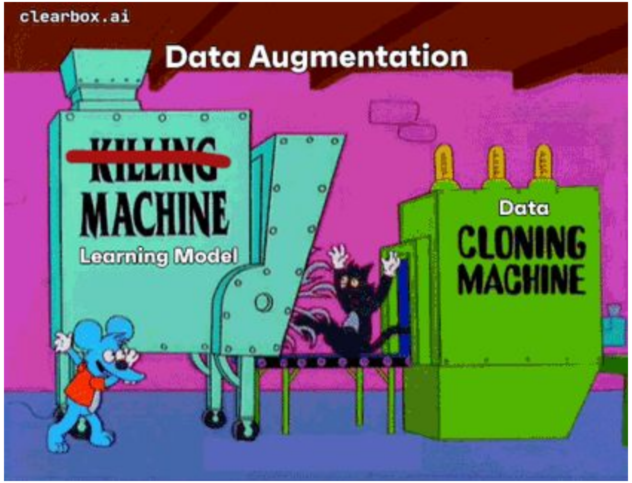


Why Synthetic Data?

Data Augmentation

↓

Utility/Fidelity



Why Synthetic Data?

Data Sharing

↓

Privacy

Data Anonymization

All the traditional techniques require a more or less complex transformation of the original sensitive data into anonymised data. This means that there is a **1-to-1 mapping** between the original data and the synthetic data, so it might be possible to find an **inverse transformation** that allows one to get back the original data from the synthetic data, although this might be very difficult.

On the other hand, with a good synthetic data generation, new data records are generated from a distribution learnt by a (ML) model, **there is by construction no unique mapping between the synthetic records and the original ones**. This is true in general, but we should not underestimate the risks.

Identity Disclosure

Let's say **s** is a record in our private synthetic dataset **S**. If an attacker is able, with or without any **background knowledge**, to assign a real identity to that record, we call this an **identity disclosure**.

Assessing the risk and relevance of an identity disclosure, we must take into account the background knowledge and resulting **information gain** of the attacker. Did the attacker learn something new about the identified subject?

We should be particularly concerned only with the risks derived from identity disclosures where the information gain is greater than zero. A good synthetic data generation protects from these **meaningful identity disclosures**.

Utility vs Privacy

Ideally Synthetic Data should be:

- globally (statistically) very similar AND
- individual record wise very different

with respect to the Original Data.

Direct and Indirect Identifiers

Identifiers are personal attributes that can be used to help identify an individual.

Direct identifiers: Identifiers that are unique to a single individual, such as Name, Surname, Social Security numbers (Codice Fiscale), passport numbers, IP addresses and IBANs. These are called **Personal Identifiable Information (PII)**.

Indirect identifiers The remaining kinds of identifiers, personal attributes that are not unique to a specific individual on their own such as height, gender, native country, city of residence, and more. Indirect identifiers can often be used in combination to single out an individual's records.

Privacy Preserving Techniques on Textual Data

- **Replace Personally Identifiable Information (PII)**:
  - Replace names, addresses, and other identifiable information with generic placeholders (e.g., “John Doe“, “123 Main St“).
- **Tokenization:**
  - Break down text into tokens and replace sensitive words or phrases with random identifiers or tokens.
  - Preserves the structure and context of the text while masking sensitive information.
- **Hashing or Encryption:**
  - Apply cryptographic techniques like hashing or encryption to obscure sensitive data.
  - Hash functions convert data into fixed-size strings of characters, making it difficult to reverse-engineer original values.
  - Encryption algorithms encode data in a way that only authorized parties with access to a decryption key can decipher.

NERPII

NERPII is a Python library to perform Named Entity Recognition and generate Personal Identifiable Information.

When working with banks or healthcare companies, personal information is abundant, necessitating the anonymization of this data.

The idea for NERPII stems from the need to preserve the privacy of personal information contained in **structured data**, such as Excel or CSV files.

Named Entity Recognition

Named Entity Recognition (NER) is a natural language processing technique that focuses on identifying and classifying specific entities or named entities in data, such as LOCATION, ORGANIZATION or PERSON to words in texts.

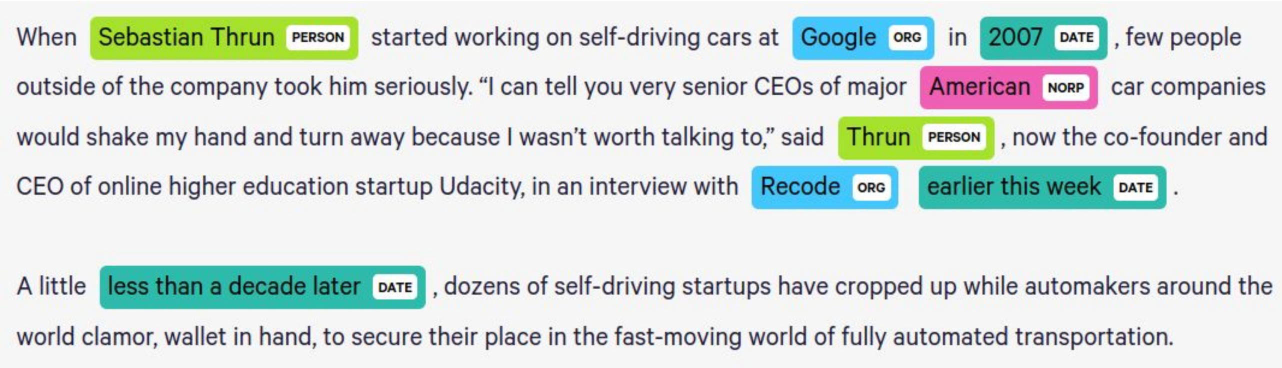


Image source: https://demos.explosion.ai/displacy-ent

NERPII Library - Architecture

NERPII contains two classes:

NamedEntityRecognizer

to perform Named Entity Recognition on structured data, typically in the form of a CSV file

FakerGenerator

to generate synthetic Personal Identifiable Information

NamedEntityRecognizer

This class assigns a named entity to the columns of a dataset to obtain information about their contents.

To do this, it uses:

- **Presidio**, an SDK from Microsoft
- NLP model (***dslim/bert-base-NER*** for English and ***osiria/bert-italian-uncased-ner*** for Italian), available on HuggingFace.
- some manual processes based on regex.

These two models, trained to perform NER, check each row in a column and try to assign an entity to each row. The final entity assigned to the column will be the most frequent entity.

NERPII Library - How to use it - 1

```
pip install nerpii
```

```
from nerpii.named_entity_recognizer
import NamedEntityRecognizer
```

```
recognizer = NamedEntityRecognizer('./csv_path.csv', lang='en')
```

```
recognizer.assign_entities_with_presidio()
recognizer.assign_entities_manually()
recognizer.assign_organization_entity_with_model()
```

```
recognizer.dict_global_entities
```

Results of the NamedEntityRecognizer


```
{'name': {'entity': 'PERSON', 'confidence_score': 0.9127725856697819},
'address': {'entity': 'ADDRESS', 'confidence_score': 0.8926174496644296},
'city': {'entity': 'LOCATION', 'confidence_score': 0.8731343283582089},
'zip': {'entity': 'ZIPCODE', 'confidence_score': 1.0},
'phone': {'entity': 'PHONE_NUMBER', 'confidence_score': 0.888},
'email': {'entity': 'EMAIL_ADDRESS', 'confidence_score': 1.0}}
```

FakerGenerator

After knowing which entity types our columns in the dataset contain, NERPII regenerates data for those columns containing PII, such as names, addresses, phone numbers etc.

For this generation it uses **Faker**, a Python package that generates fake data to anonymize PII.

NERPII Library - How to use it - 2

```
from nerpii.faker_generator import
FakerGenerator
```

```
generator = FakerGenerator(dataset, recognizer.dict_global_entities,  lang='en')
```

```
generator.get_faker_generation()
```

Results of the FakerGenerator

	first_name	last_name	company_name	address	city	county	state	zip	phone1	phone2	email
0	James	Butt	Benton, John B Jr	6649 N Blue Gum St	New Orleans	Orleans	LA	70116	504-621-8927	504-845-1427	jbutt@gmail.com
1	Josephine	Darakjy	Chanay, Jeffrey A Esq	4 B Blue Ridge Blvd	Brighton	Livingston	MI	48116	810-292-9388	810-374-9840	josephine_darakjy@darakjy.org
2	Art	Venere	Chemel, James L Cpa	8 W Cerritos Ave #54	Bridgeport	Gloucester	NJ	8014	856-636-8749	856-264-4130	art@venere.org
3	Lenna	Paprocki	Feltz Printing Service	639 Main St	Anchorage	Anchorage	AK	99501	907-385-4412	907-921-2010	lpaprocki@hotmail.com
4	Donette	Foller	Printing Dimensions	34 Center St	Hamilton	Butler	OH	45011	513-570-1893	513-549-4561	donette.foller@cox.net
5	Simona	Morasca	Chapman, Ross E Esq	3 Mcauley Dr	Ashland	Ashland	OH	44805	419-503-2484	419-800-6759	simona@morasca.com
6	Mitsue	Tollner	Morlong Associates	7 Eads St	Chicago	Cook	IL	60632	773-573-6914	773-924-8565	mitsue_tollner@yahoo.com
7	Leota	Dilliard	Commercial Press	7 W Jackson Blvd	San Jose	Santa Clara	CA	95111	408-752-3500	408-813-1105	leota@hotmail.com
8	Sage	Wieser	Truhlar And Truhlar Attys	5 Boston Ave #88	Sioux Falls	Minnehaha	SD	57105	605-414-2147	605-794-4895	sage_wieser@cox.net
9	Kris	Marrier	King, Christopher A Esq	228 Runamuck Pl #2808	Baltimore	Baltimore City	MD	21224	410-655-8723	410-804-4694	kris@gmail.com

Table 1: Real Data

	first_name	last_name	company_name	address	city	county	state	zip	phone1	phone2	email
0	Derek	Rogers	Benton, John B Jr	8756 Melissa Prairie	Markberg	Orleans	SC	16556	491-985-4242x3796	001-471-611-7788x69862	derek.rogers@yahoo.com
1	Douglas	Benitez	Chanay, Jeffrey A Esq	4086 Stuart Road Suite 611	Hernandezshire	Livingston	MO	64129	(297)198-8232x773	0352799906	douglas.benitez@hotmail.com
2	Franklin	Rodriguez	Chemel, James L Cpa	50083 Kiara Village	Aaronberg	Gloucester	TX	60793	+1-029-117-3288x07324	816-789-4684x07100	franklin.rodriguez@yahoo.com
3	Daniel	Bush	Feltz Printing Service	453 Parker Points	Brandymouth	Anchorage	TX	13204	+1-955-200-1727x55320	(217)516-2281	daniel.bush@hotmail.com
4	Lee	Torres	Printing Dimensions	912 Brown Curve Apt. 782	New Garretthaven	Butler	TX	93264	568.085.0635x308	564.103.1190x1408	lee.torres@gmail.com
...	...	...	...	...	...	...	...	...	...	...	...
495	Raymond	Macias	Inner Label	26664 Cain Meadow	Williamsfurt	Ada	NH	61976	384-631-3190	+1-238-503-8197x50635	raymond.macias@gmail.com
496	Tanner	Flores	Hermar Inc	30903 Webb Parks	West Calvin	Elkhart	WV	35474	495-620-7576x979	+1-882-751-7106x461	tanner.flores@hotmail.com
497	Garrett	Perry	Simonton Howe & Schneider Pc	9852 Walters Shoal	Franklinbury	Box Butte	ME	81011	+1-041-980-2373x6347	(174)238-8279x2745	garrett.perry@gmail.com
498	Richard	Smith	Warehouse Office & Paper Prod	922 Beth Drives Suite 294	Melaniefort	King	AR	14946	288.351.0492x77857	422-416-2622	richard.smith@gmail.com
499	Ronald	Sheppard	Affiliated With Travelodge	74567 Lori Field Apt. 638	Weaverhaven	Orange	CA	05555	1786814253	(735)510-6062	ronald.sheppard@yahoo.com

Table 2: Synthetic Data

Future Directions and Conclusion for Nerpii

A promising direction for the future development of this library is

- to adapt it to other languages in addition to English and Italian
- to expand the number of Named Entities recognized by the NamedEntityRecognizer and those regenerated by the FakerGenerator.
- to examine the aspect of coherence among different columns.

NERPII combines the remarkable potential of Named Entity Recognition with the power of synthetic data in order to preserve the privacy of the Personal Identifiable Information contained in structured data.

Useful links

NERPII

Slides

clearbox AI
humans && machines

Thanks!

Feel free to contact us:

www.clearbox.ai

info@clearbox.ai
simona@clearbox.ai

@ClearboxAI