ATHENIA
AI LABS

MICROSOFT PRESIDIO

Data Protection and De-identification SDK

01 - NLP

Field of AI. Interaction between computers and human language. It enables machine to understand, interpret and respond to text or spoken words same as the humans do.

02 - PII

Personally identifiable Information.
Data that can be used to identify an individual, directly or indirectly (Names, email, IP addresses, phone numbers etc...)

De-identification

Process of removing or altering information

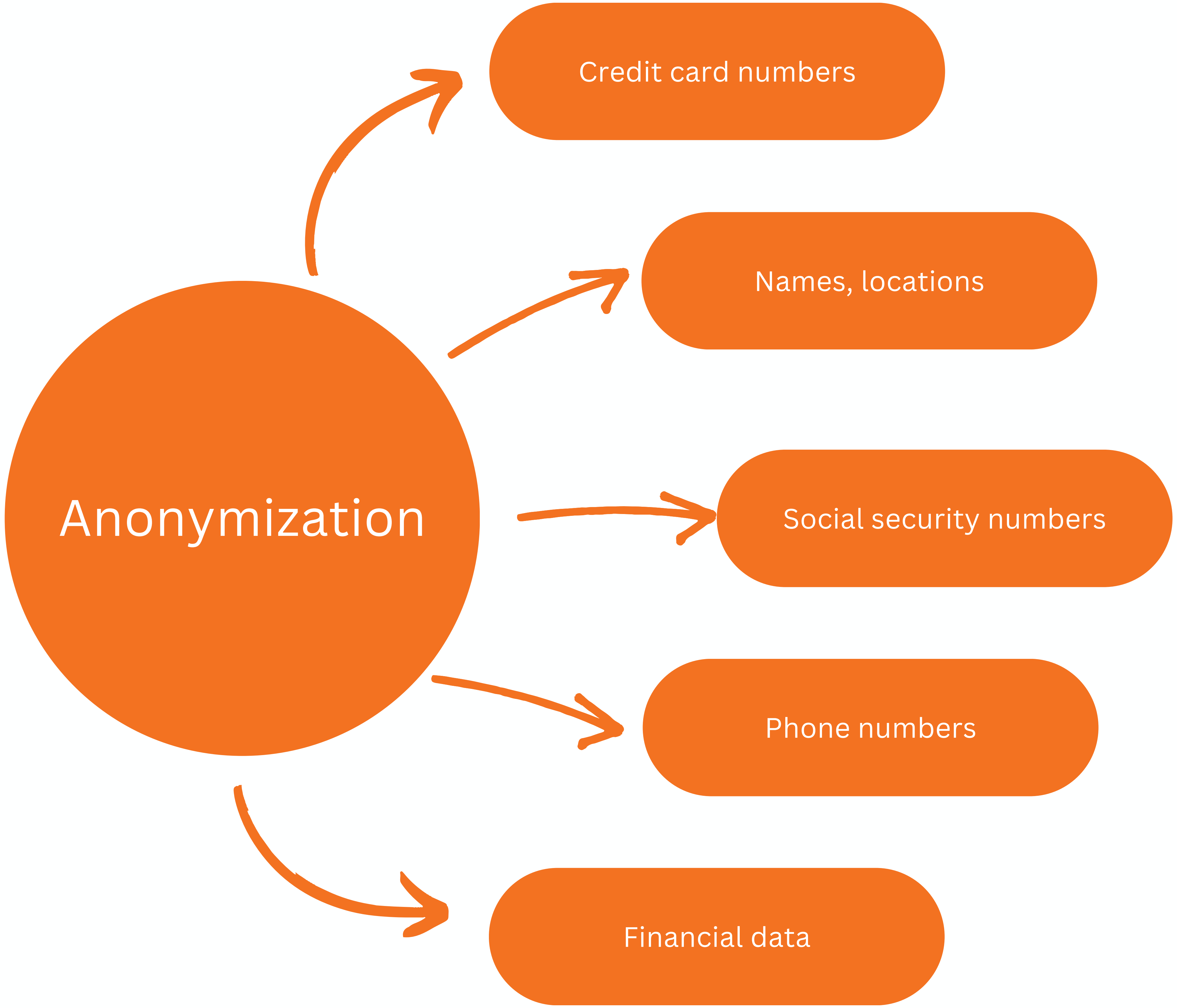
Hi my name is **Ajin** and my number is **123456**

Hi my name is **<Person>** and my number is **<Phone no.>**

Hi my name is **David** and my number is **98388141**

Presidio

Is a open-source Python package which helps to ensure sensitive data is properly managed and governed.



Use Cases

- ✓ Organizations need to comply with regulations like GDPR, HIPAA, and CCPA, which require the protection of PII.
- ✓ Sharing data with third parties, such as analytics firms or business partners, often requires the removal of sensitive information to protect individuals' privacy.
- ✓ Healthcare providers need to anonymize patient records for research and data analysis without compromising patient privacy.
- ✓ Law firms and legal departments manage documents containing sensitive client information that must be protected during the legal process.

Installation

- **Analyze: Analyses the text/image**
- **Anonymizes: Masks/Alter the analyzed text**

PII anonymization on text

spaCy:

```
pip install presidio_analyzer
pip install presidio_anonymizer
python -m spacy download en_core_web_lg
```

Transformers:

```
pip install "presidio_analyzer[transformers]"
pip install presidio_anonymizer
python -m spacy download en_core_web_sm
```

stanza:

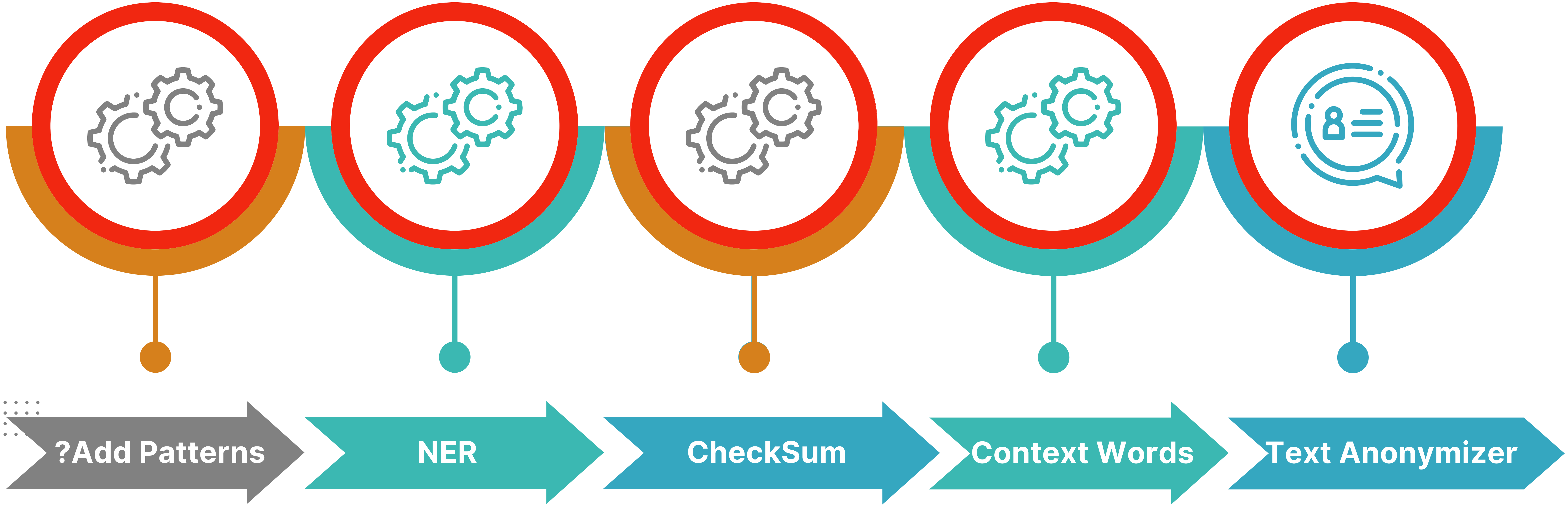
```
pip install "presidio_analyzer[stanza]"
pip install presidio_anonymizer
```

PII redaction in images

```
pip install presidio_image_redactor
# Presidio image redactor uses the presidio-analyzer
# which requires a spaCy language model:
python -m spacy download en_core_web_lg
```

We can also use Presidio with **langChain**

Presidio Workflow



Regex pattern recognition
eg: \b[6-9]\d{9}\b

Named Enitity Recognition uses nlp models such as spaCy, flair, etc...

Validate patterns

Increase the detection confidence

Multiple anonymization techniques

Types of Pattern Recognition

Deny List

Will pass a short list of tokens which should be marked as PII if detected

Using Regex

Detects any regular expression pattern

Rule based logic recognizer

Let's say we also would like to detect numbers within words, e.g. "Number One". We can leverage the underlying spaCy token attributes, or write our own logic to detect such entities.

Supporting new languages

Can be acheived through NLPEngine

Faker

Faker is a Python package that generates fake data for you.

```
fake.name()
fake.address()
fake.email()
fake.text()
```

```
fake.country()
fake.latitude()
fake.longitude()
fake.url()
```

Change Language

```
fake = Faker('hi_IN')
fake.sentence(ext_word_list = word_list)
```

```
document = """Date: October 19, 2021
 Witness: John Doe
 Subject: Testimony Regarding the Loss of Wallet
 Testimony Content:
 Hello Officer,
 My name is John Doe and on October 19, 2021, my wallet was stolen in the vicinity of Kilmarnock during a bike trip. This wallet contains some very important things to me.
 Firstly, the wallet contains my credit card with number 4111 1111 1111 1111, which is registered under my name and linked to my bank account, PL61109010140000071219812874.
 Additionally, the wallet had a driver's license - DL No: 999000680 issued to my name. It also houses my Social Security Number, 602-76-4532.
 What's more, I had my polish identity card there, with the number ABC123456.
 I would like this data to be secured and protected in all possible ways. I believe It was stolen at 9:30 AM.
 In case any information arises regarding my wallet, please reach out to me on my phone number, 999-888-7777, or through my personal email, johndoe@example.com.
 Please consider this information to be highly confidential and respect my privacy.
 The bank has been informed about the stolen credit card and necessary actions have been taken from their end. They will be reachable at their official email, support@bankname.com.
 My representative there is Victoria Cherry (her business phone: 987-654-3210).
 Thank you for your assistance,
 John Doe"""


def analyze(text):
  analyzer_results = analyzer.analyze(document, language="en")
  for res in analyzer_results:
    print((document[res.start:res.end], res.start, res.end))

analyze(document)
```

```
[('4111 1111 1111 1111', 360, 379),
('PL61109010140000071219812874', 446, 474),
('johndoe@example.com', 950, 969),
('support@bankname.com', 1216, 1236),
('987-654-3210', 1303, 1315),
('October 19, 2021', 6, 22),
('John Doe', 33, 41),
('John Doe', 142, 150),
('October 19, 2021', 158, 174),
('Kilmarnock', 216, 226),
('602-76-4532', 606, 617),
('9:30 AM', 800, 807),
('Victoria Cherry', 1266, 1281),
('John Doe', 1353, 1361),
('999-888-7777', 906, 918),
('987-654-3210', 1303, 1315),
('example.com', 958, 969),
('bankname.com', 1224, 1236),
('999000680', 535, 544),
('999000680', 535, 544),
('999000680', 535, 544)]
```

Lets Anonymize

```
def anonymize(text, analyzer_results):
  anonymized_text = anonymizer.anonymize(text=text,
analyzer_results=analyzer_results)
  print(anonymized_text)


anonymize(document, analyzer_results)
```

text: Date: <DATE_TIME>

Witness: <PERSON>

Subject: Testimony Regarding the Loss of Wallet

Testimony Content:

Hello Officer,

My name is <PERSON> and on <DATE_TIME>, my wallet was stolen in the vicinity of <LOCATION> during a bike trip. This wallet contains some very important things to me.

Firstly, the wallet contains my credit card with number <CREDIT_CARD>, which is <IN_PAN> under my name and linked to my bank account, <IBAN_CODE>.

Additionally, the wallet had a driver's license - DL No: <US_DRIVER_LICENSE> issued to my name. It also houses my Social Security Number, <US_SSN>.

What's more, I had my polish identity card there, with the number *ABC123456*.

I would like this data to be secured and protected in all possible ways. I believe It was stolen at <DATE_TIME>.

In case any information arises regarding my wallet, please reach out to me on my phone number, <PHONE_NUMBER>, or through my personal email, <EMAIL_ADDRESS>.

Please consider this information to be highly confidential and respect my privacy.

The bank has been informed about the stolen credit card and necessary actions have been taken from their end. They will be reachable at their official email, <EMAIL_ADDRESS>.

My representative there is <PERSON> (her business phone: <UK_NHS>).

Thank you for your assistance,

<PERSON>

ADD a REGEX PATTERN

```
def anonymize_polish_id(text):
    # Define the pattern for Polish ID
    polish_id_pattern = Pattern(
        name="polish_id_pattern",
        regex="[A-Z]{3}\d{6}",
        score=1.0,
    )

    # Create a recognizer using the defined pattern
    polish_id_recognizer = PatternRecognizer(
        supported_entity="POLISH_ID", patterns=[polish_id_pattern]
    )

    # Initialize the AnalyzerEngine and AnonymizerEngine
    analyzer = AnalyzerEngine()
    anonymizer = AnonymizerEngine()

    # Add the custom recognizer to the analyzer's registry
    analyzer.registry.add_recognizer(polish_id_recognizer)

    # Analyze the text to detect entities
    analyzer_results = analyzer.analyze(document=text, language="en")

    # Anonymize the detected entities in the text
    anonymized_results = anonymizer.anonymize(
        text=text,
        analyzer_results=analyzer_results,
    )

    print(anonymized_results.text)


anonymize_polish_id(document)
```

```
What's more, I had my polish identity card there, with the number <POLISH_ID>.
```

Issues! Issues! Issues!

Cannot directly use pdf format

Alternative: Use any other python package for text extraction. E.g: PyPDF2, Pdfminer.six, etc...

Choosing Open-Source NER/NLPEngine

Out of the box models are a **starting point**, not a **final destination.** We might need to **fine tune** these models for our particular needs.

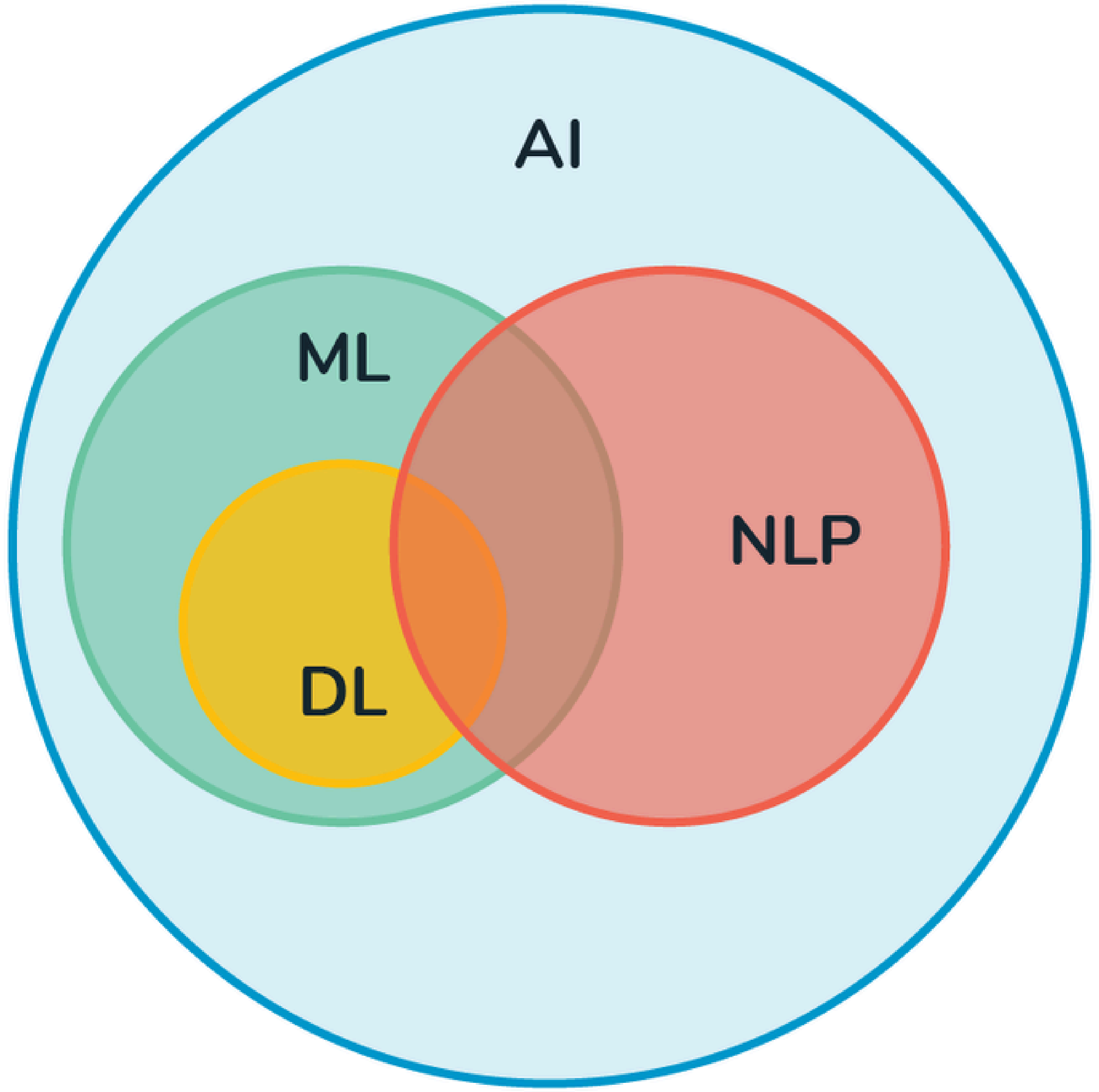
01 Small spaCy model

02 Large spacy Model

03 Flair

04 Stanford’s Core NLP

05 Yet to research on!



ATHENIA
AI LABS

Conclusion

Microsoft Presidio stands out as a robust and versatile tool for handling sensitive information within text. Through this demonstration, I have highlighted its capability to accurately identify and anonymize various types of Personally Identifiable Information (PII) and their drawbacks, what improvment we can do etc... As data privacy regulations continue to tighten globally, tools like Microsoft Presidio are invaluable in ensuring compliance and protecting individuals' sensitive information. This demonstration underscores Presidio's practical utility and effectiveness, making it a critical component in any data protection strategy.

ATHENIA
AI LABS

GOT QUESTIONS?

Thank You