

DH04

Text as Data

in the Context of Anonymising Clinical Study Reports

Lukasz Kniola

Biogen, UK

Woo Song, Tim Perin, Nitin Chaudhary

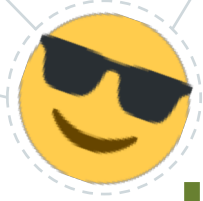
Xogene, USA



Disclaimer

The scope of this presentation is to present the opinions and suggestions of the authors. The interpretations of standards and procedures presented are those of the authors and are not necessarily correct. Any views, conclusions and recommendations are those of the authors, and they do not represent the positions of their employers.

The challenge



**Datasets
de-identified
successfully**

**Documents
a different
story**

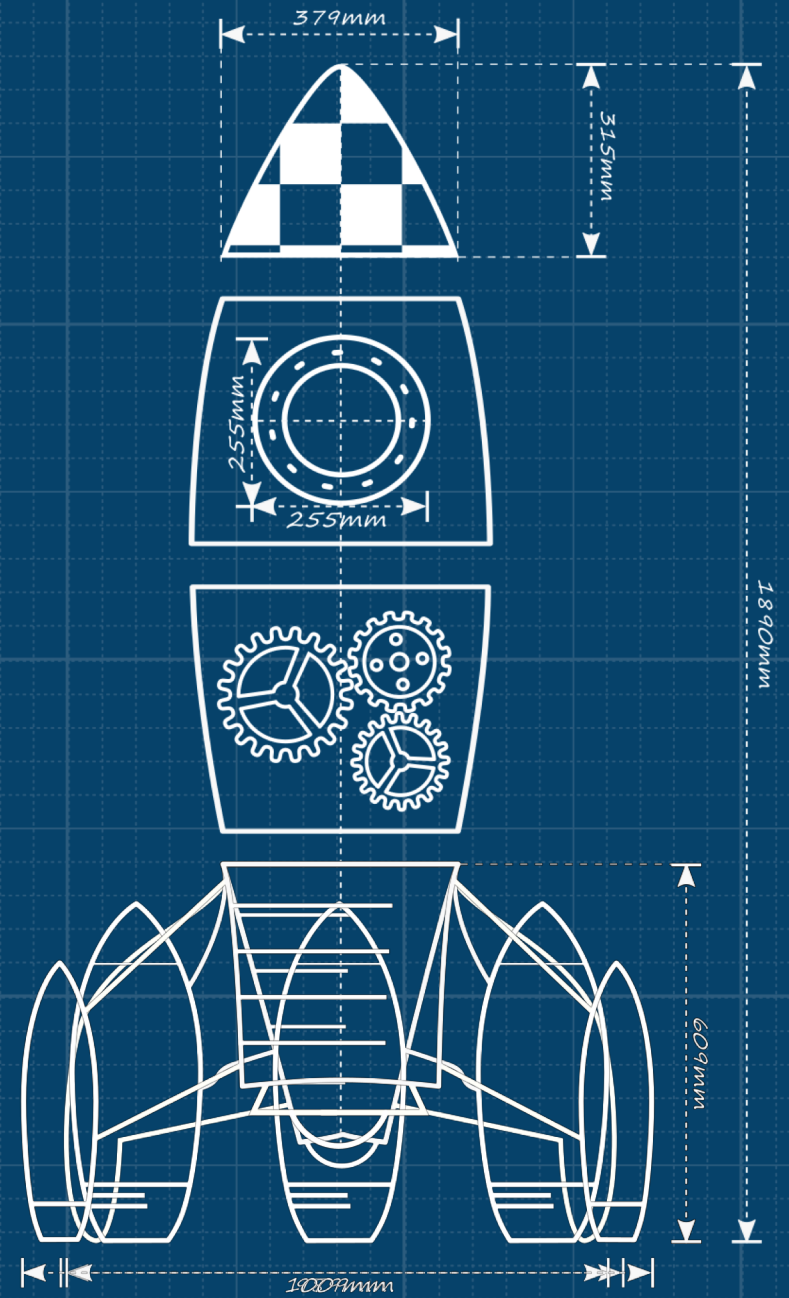
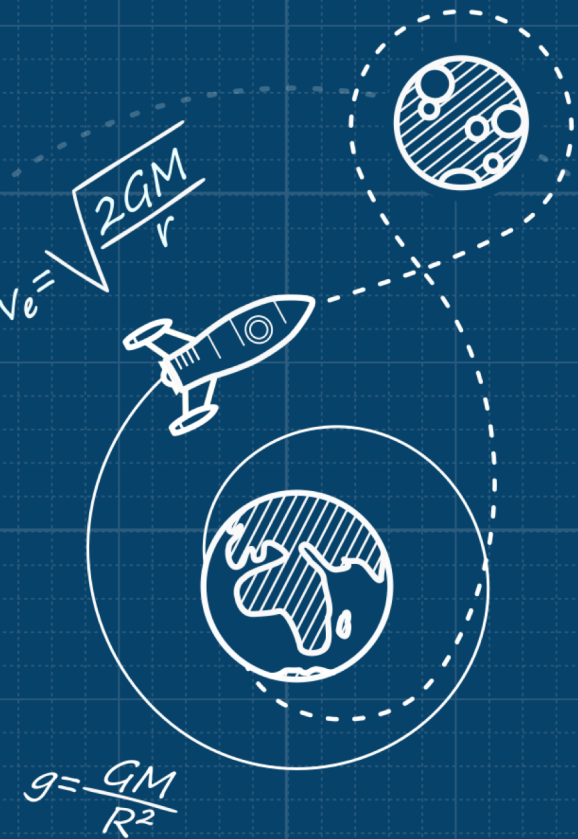


The Answer

A prototype tool that
reads, processes and
anonymises reports



Modular Design



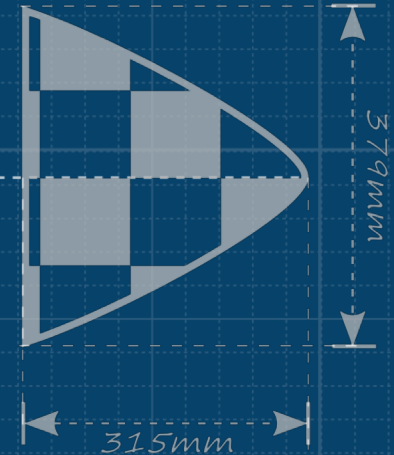
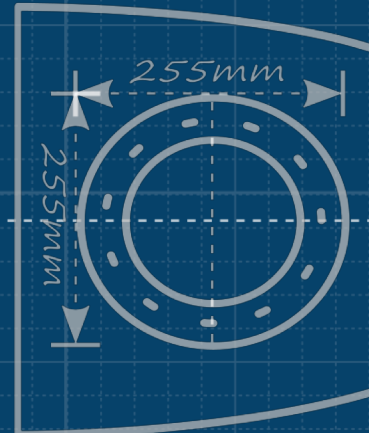
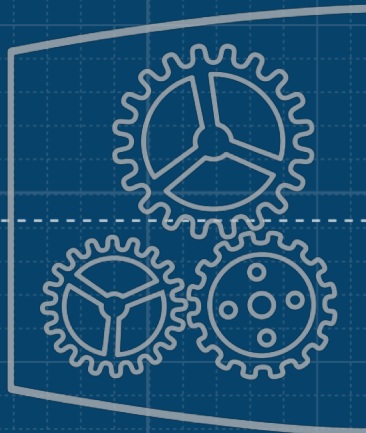
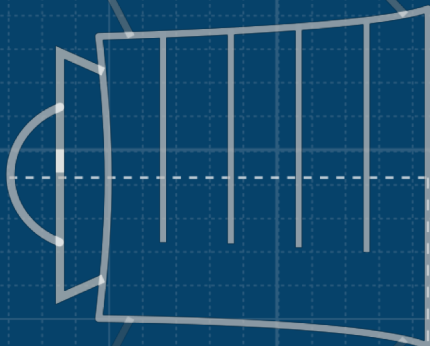
Modular Design

*Text
Extraction
Module*

NLP
Module*

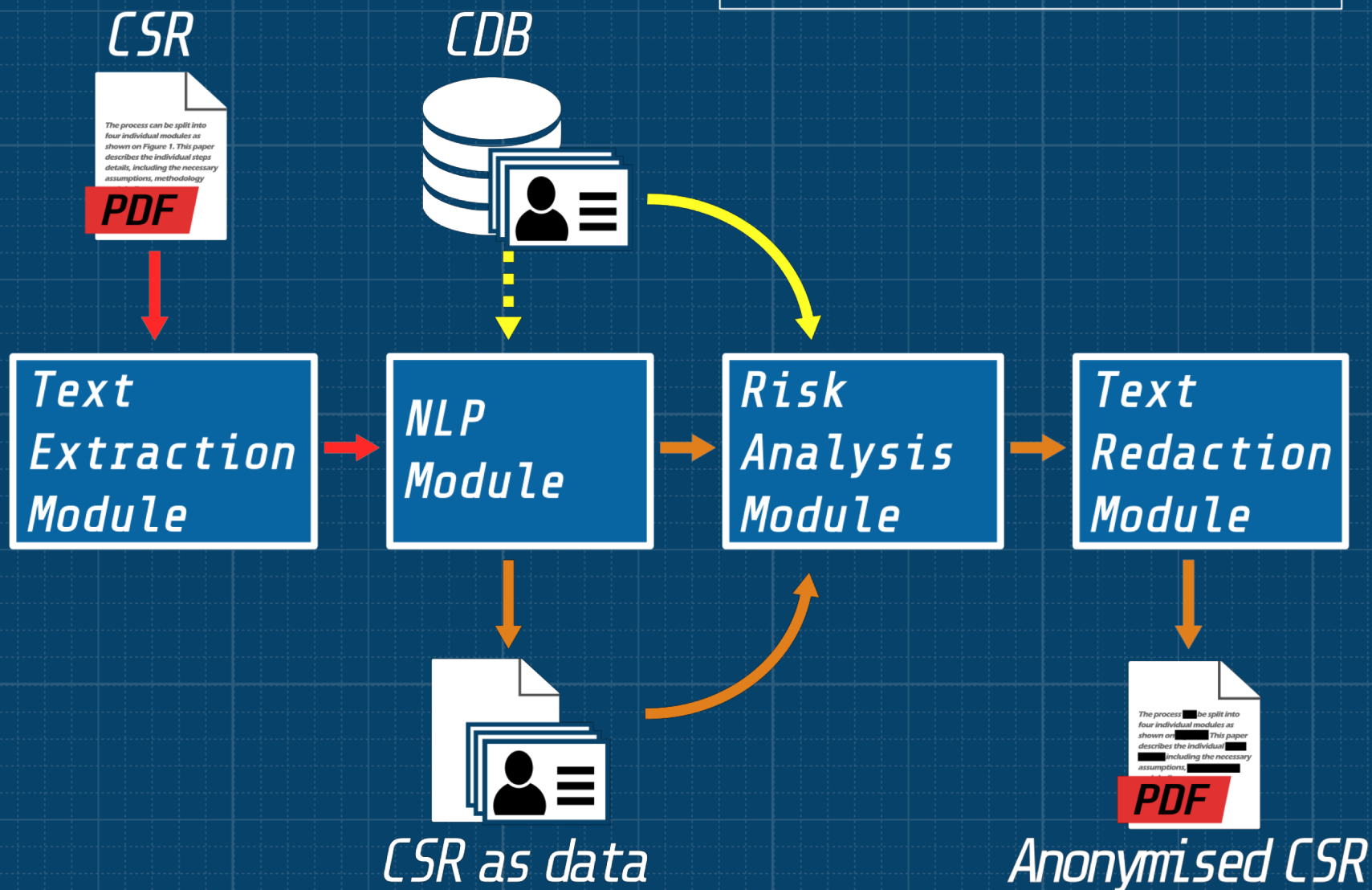
*Risk
Analysis
Module*

*Text
Redaction
Module*



**Natural Language Processing*

Process/Data Flow



Tools

PDF SDK

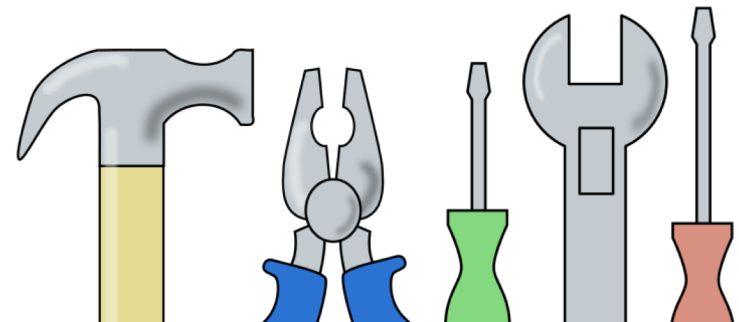
PDFTRON

NLP + ML SDK

spaCy

All coding in

 **python**TM





1.

Extracting text from PDF

What could possibly go wrong?

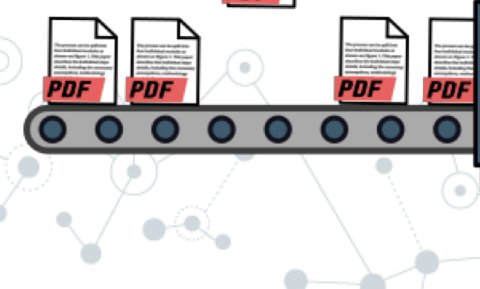
Generate a PDF

Export as PDF

Print to PDF

Generate

- shown. This paper describes the individual steps details, including the methodology, assumptions, methods, and challenges.



A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines. The nodes are represented by circles of varying sizes, some with concentric rings, and the lines are thin and grey. The diagram is partially cut off by the left edge of the slide.

2.

Processing text

Finding candidates for
anonymisation/redaction

A decorative network diagram in the bottom-right corner, similar to the one in the top-left. It shows a complex web of interconnected nodes and lines, with nodes represented by circles of varying sizes and lines as thin grey connections. The diagram is partially cut off by the right edge of the slide.

Identifiers in the text

- ◎ Personal info: **name, phone no., address, SSN**, etc.
- ◎ IDs: **Subject ID, test ID, site ID, lab ID**, etc.
- ◎ Demographic info: **sex, age, race, ethnicity**
- ◎ Geographic info: **country, region, continent**
- ◎ Body measurements: **height, weight, BMI**, etc.
- ◎ **AE, CM, MH**, etc.
- ◎ Sensitive info: **pregnancy, mental health, substance abuse**, etc.

◎ Dates

Simple methods

IF this THEN that

ELSE IF this THEN that

ELSE IF this THEN that

ELSE IF this THEN that

ELSE IF this THEN that

ELSE IF this THEN that

ELSE IF this THEN that

ELSE IF this THEN that

ELSE IF this THEN that

ELSE IF this THEN that

ELSE IF this THEN that

ELSE IF this THEN that

ELSE IF this THEN that

ELSE IF this THEN that

ELSE IF this THEN that

Regular Expressions (RegEx)

"Lukasz"

r"lukasz"i

r"[l|ł]u[c|k]asz?[a-z]*"i

Lukasz

Lucas

Lukas

Łukaszek

Łukasz Kniola

Lucasz

lukasz kniola@biogen.com

Regular Expressions (RegEx)

regex101

REGULAR EXPRESSION

`i r" \d+[-]ye?a?rs?[-]old\b`

8 m

TEST STRING

SWIT

21 years old
22 Years Old
23-year-old
24 year-old
25 yeaear-old
26 yr-old
27 yrs-old
28-yr-old
29-yrs-old
30 years older

MATCH INFORMATION

Match 1

Full match

0-12 `21 years old`

Match 2

Full match

13-25 `22 Years Old`

Match 3

Full match

26-37 `23-year-old`

Match 4

Full match

38-49 `24 year-old`

Match 5

Full match

64-73 `26 yr-old`

Match 6

Full match

74-84 `27 yrs-old`

Regular Expressions (RegEx)

 regex101    

REGULAR EXPRESSION 5 matches, 379 steps

```
r"[^-\]\b(?:P<SUBJID>\d{3}-\d{3})\b[^\-]
```

TEST STRING SWITCH TO UN

* 101-221 and 110-091 met on 2001-06-11
They called 202-110 on 1-555-101-001
and asked to find 310-001 at 122 High St
Information can be found on pages 302-304

SUBSTITUTION

MATCH INFORMATION

Match 1

Full match	1-10	` 101-221 `
Group `SUBJID`	2-9	` 101-221 `

Match 2

Full match	13-22	` 110-091 `
Group `SUBJID`	14-21	` 110-091 `

Match 3

Full match	52-61	` 202-110 `
Group `SUBJID`	53-60	` 202-110 `

Match 4

Full match	96-105	` 310-001 `
Group `SUBJID`	97-104	` 310-001 `

Match 5

Full match	157-166	` 302-304 `
Group `SUBJID`	158-165	` 302-304 `

Value Lists based on CDB

Clinical Database

USUBJID	AETERM
X01-001	Fever
X01-001	Cough
X01-001	Nausea
X01-003	Headache
X01-003	Fever
X02-004	Cough
X02-004	Headache
X02-004	Nausea

Value Lists

USUBJID	AETERM
X01-001	Cough
X01-003	Fever
X02-004	Nausea
	Headache

Report Text

Lorem ipsum dolor sit amet, consectetur **X01-003** sed do **fever** incidunt ut labore et **headache** magna aliqua. Ut enim ad minim **X01-001**, quis nostrud ullamco **cough** nisi aliquip ex ea **nausea** consequat. Duis aute irure **fever** in reprehenderit in voluptate cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident.

Lorem ipsum dolor **X01-003**, constetur adipiscing aliqua. Ut enim ad **headache** veniam, quis nostrud. Duis aute irure **fever** in reprehenderit in voluptate. Lorem ipsum **X02-004** amet, consectetur, sed do **cough** ut labore et **nausea** aliqua.

Lorem ipsum dolor, **X01-003** et **X02-004** adipisce sed do incididunt labore et **headache** magna aliqua. Ut enim ad minim, quis nostrud exercitation, aute irure **fever** in reprehenderit in voluptate cillum

Natural Language Processing (NLP)

Tokenization

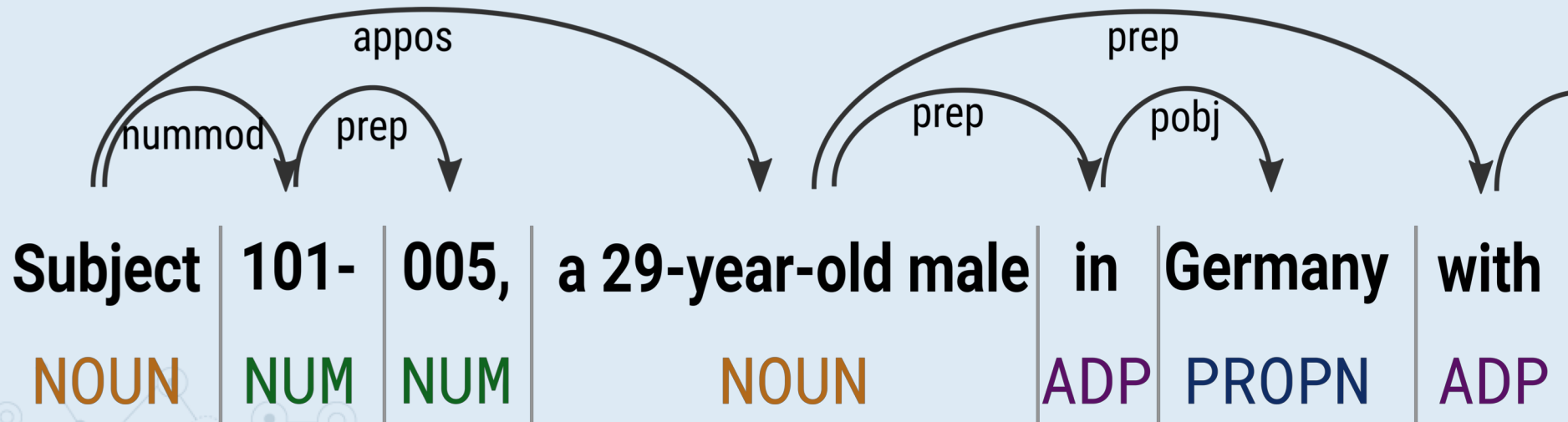
Stemming

Lemmatization

Part of Speech (POS) Tagging

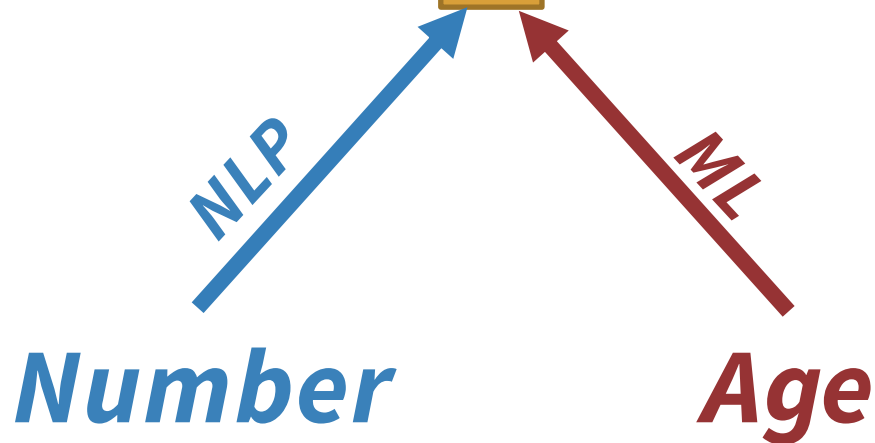
Named Entity Recognition

Chunking



Machine Learning (ML)

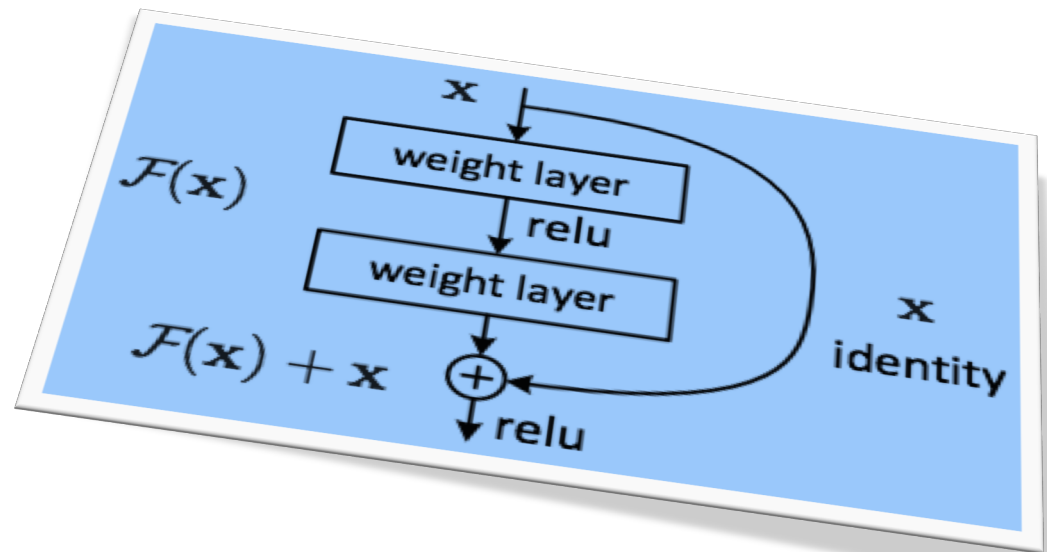
“The subject was **43** at the time of the event.”



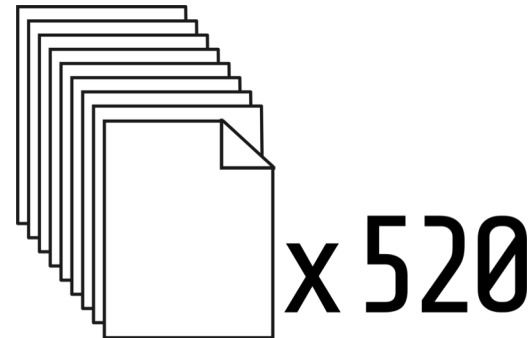
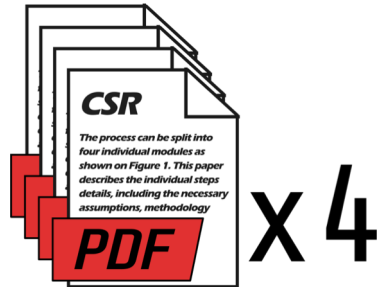
Machine Learning (ML)

spaCy

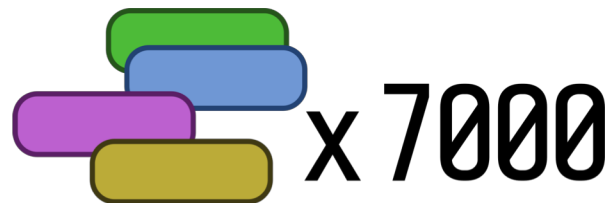
- Residual CNN architecture
Convolutional Neural Network
Layers with Residual Connections
- NER Tagging
Named Entity
Recognition



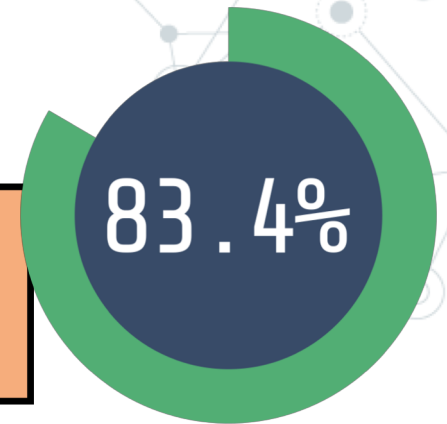
Machine Learning (ML)



manually tagged
patient ID, age, gender, race/ethnicity, date



Machine Learning (ML)



Subject **101-005** SUBJECTID, a **29-year-old** AGE **male** GENDER in **Germany** GPE with Coding Error Disease, was seen for pyrexia on **03 September 2017** DATE, and fatigue on **08 September 2018** DATE.

Text parametrising

Start position	Page number	Text	Length	Entity type
194	93	101-002	7	SUBJECTID
429	93	101-002	7	SUBJECTID
440	93	30-year-old	11	AGE
452	93	male	4	GENDER
520	93	21 August 2017	14	DATE
722	93	Injection Site Erythema	23	AEDECOD
896	93	03 July 2017	12	DATE
...	...	...	...	...



3.

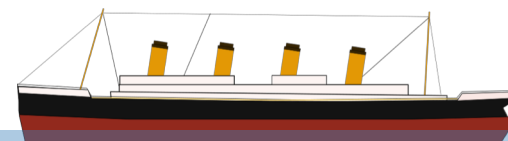
Re-identification risk assessment

Selecting anonymisation/redaction
rules



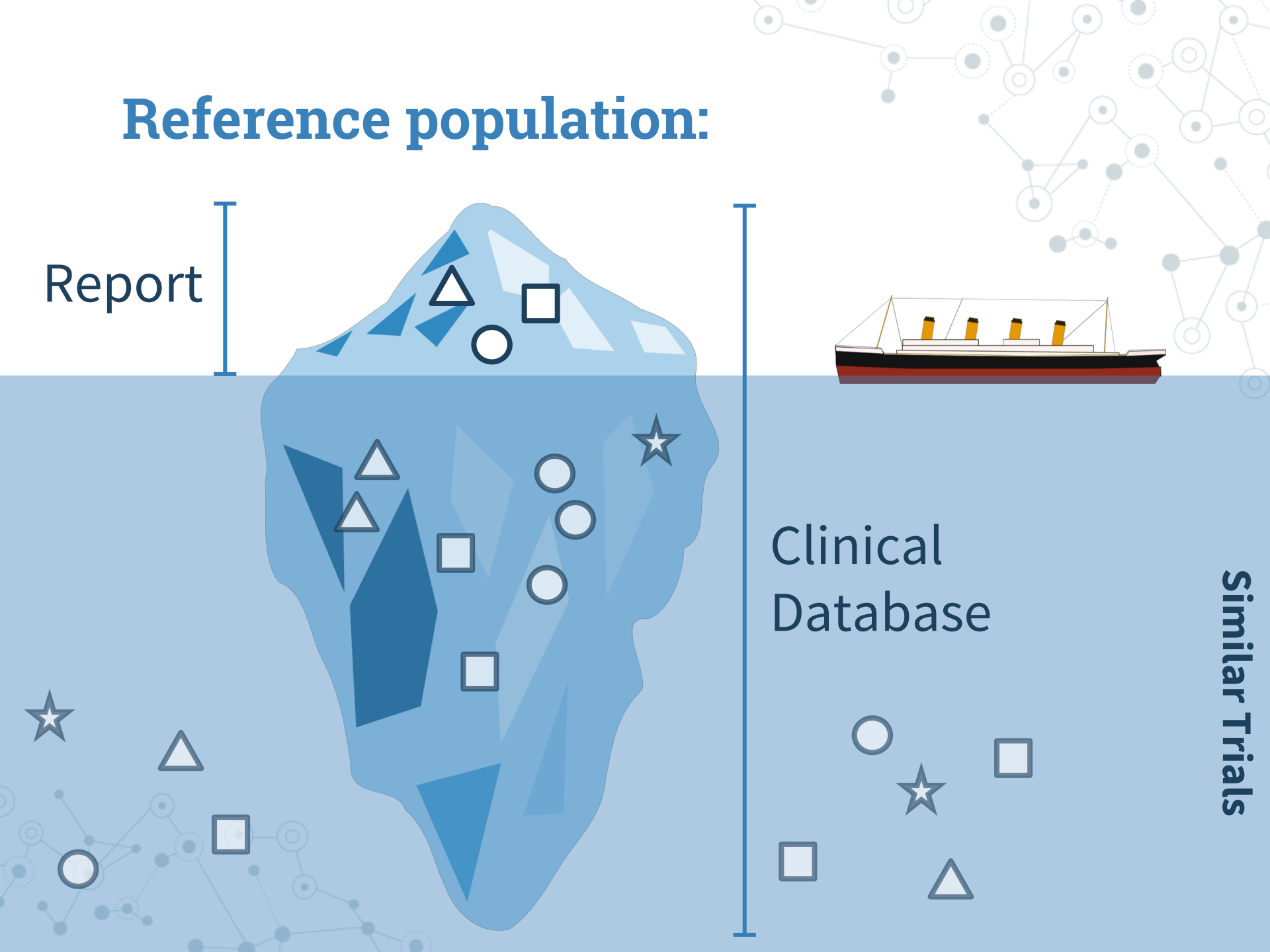
Reference population:

Report

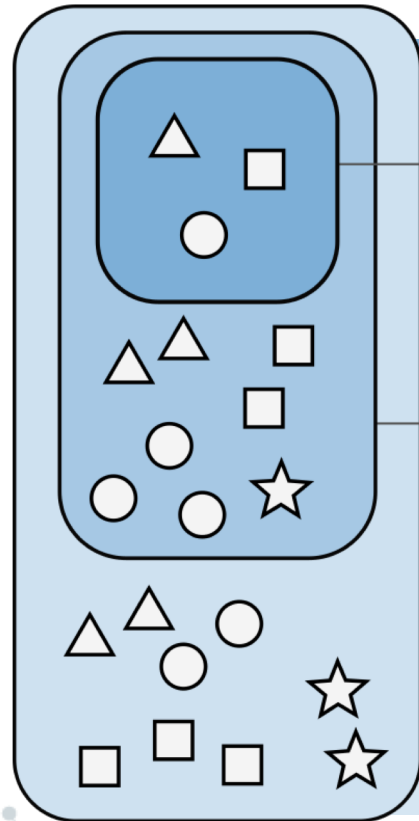


Clinical
Database

Similar Trials



Reference population:



A diagram of a reference population container, represented as a large rounded rectangle with a smaller rounded rectangle inside. The inner rectangle contains three shapes: a triangle, a square, and a circle. The outer rectangle contains a total of 14 shapes: 3 triangles, 3 squares, 3 circles, and 5 stars. Lines connect the labels 'Report', 'Clinical Database', and 'Similar Trials' to the inner, middle, and outer rectangles respectively.

				
<i>Report</i>	1/1	1/1	1/1	-
<i>Clinical Database</i>	1/3	1/3	1/4	1/1
<i>Similar Trials</i>	1/5	1/6	1/6	1/3

Risk Assessment Methodology

CONTEXT

**CONTROLLED
SHARING**

**PUBLIC
RELEASE**

REFERENCE
POPULATION

**TRIAL/SUBMISSION
*DATABASE***

SIMILAR TRIALS

RISK
MEASURE

**AVERAGE
RISK**

**MAXIMUM
RISK**

THRESHOLD

0.09

OTHER

Reducing Risk :: De-ID Rules

Pseudonyms:

- ◎ Subject ID, Site ID, etc.

Groups or bands:

- ◎ Age to Age bands
- ◎ Country to Region
- ◎ Race – retain common values, pool rare values (e.g. White, Non-White)

◎ Preferred Term to SOC,



Reducing Risk :: De-ID Rules

Random value from within the band:

	Original Value		Band		New value
AGE:	41	⇒	40-44	⇒	42
Country:	Germany	⇒	Western Europe	⇒	France
Height:	181cm	⇒	180-199cm	⇒	193cm

Match format. Hide in plain sight.

Building the set rules for document anonymisation

Start pos.	Page #	Text	Entity type	Anonymisation rule	Replacement/overlay
194	93	101-002	SUBJECTID	REPLACE	X0014
429	93	101-002	SUBJECTID	REPLACE	X0014
440	93	30-year-old	AGE	REDACT	-
452	93	male	GENDER	KEEP	
520	93	21 August 2017	DATE	REPLACE	30 August 2017
722	93	Injection Site Eryth...	AEDECOD	KEEP	
896	93	03 July 2017	DATE	REPLACE	12 July 2017
...	...	...	...	...	...

A decorative network diagram in the top-left corner, consisting of various sized circles (nodes) connected by thin lines (edges). Some nodes are solid grey, while others are hollow with a grey outline. The diagram is partially cut off by the left edge of the slide.

4.

Modifying PDF

Editing and redacting the source document programmatically

A decorative network diagram in the bottom-right corner, similar to the one in the top-left, with nodes and connecting lines. It is also partially cut off by the right edge of the slide.

Redacting vs Modifying

Subject Details

Subject ID: **101-005**

Site Number: **101**

Age at Enrolment: **29**

Ethnicity/Race: **White**

Country: **Germany**

Sex: **Male**

Treatment

Administered:

2017-07-11 to

2018-07-07

Narrative

Serious Adverse Event: Fever

MedDRA Term: Pyrexia

Days on Study at Last Dose: 362

Subject **101-005**, a **29**-year-old male in **Germany** with Coding Error, was seen for pyrexia on **03 September** 2017, and fatigue on **08 September** 2018. At the time of the events, he was enrolled in Study PHUSE2018 and was receiving placebo every 4 weeks.

The subject's relevant medical history included **acne**. Relevant concomitant medications at the time of the event were caffeine supplement, Vitamin C, and topical antiseptics. No subject allergies were reported.

Serious Adverse Event: Fever

The subject had received his first placebo dose on **11 July** 2017 approximately 55 days prior to the onset of fever.

Redacting vs Modifying

Subject Details

Subject ID: [REDACTED]

Site Number: **101**

Age at Enrolment: [REDACTED]

Ethnicity/Race: **White**

Country: [REDACTED]

Sex: **Male**

Treatment

Administered:

2017-[REDACTED] to

2018-[REDACTED]

Narrative

Serious Adverse Event: Fever

MedDRA Term: Pyrexia

Days on Study at Last Dose: 362

Subject [REDACTED], a [REDACTED]-year-old male in [REDACTED] with Coding Error, was seen for pyrexia on [REDACTED] 2017, and fatigue on [REDACTED] 2018. At the time of the events, he was enrolled in Study PHUSE2018 and was receiving placebo every 4 weeks.

The subject's relevant medical history included [REDACTED]. Relevant concomitant medications at the time of the event were caffeine supplement, Vitamin C, and topical antiseptics. No subject allergies were reported.

Serious Adverse Event: Fever

The subject had received his first placebo dose on [REDACTED] 2017 approximately 55 days prior to the onset of fever.

Redacting vs Modifying

Subject Details

Subject ID: **X0021**

Site Number: **101**

Age at Enrolment: **25-29**

Ethnicity/Race: **White**

Country: **Western Europe**

Sex: **Male**

Treatment

Administered:

2017-07-21 to

2018-07-17

Narrative

Serious Adverse Event: Fever

MedDRA Term: Pyrexia

Days on Study at Last Dose: 362

Subject **X0021**, a **25-29**-year-old male in **Western Europe** with Coding Error, was seen for pyrexia on **13 September** 2017, and fatigue on **18 September** 2018. At the time of the events, he was enrolled in Study PHUSE2018 and was receiving placebo every 4 weeks.

The subject's relevant medical history included **skin condition**. Relevant concomitant medications at the time of the event were caffeine supplement, Vitamin C, and topical antiseptics. No subject allergies were reported.

Serious Adverse Event: Fever

The subject had received his first placebo dose on **21 July** 2017 approximately 55 days prior to the onset of fever.

A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines. The nodes are represented by circles of varying sizes, some with concentric rings, and the lines are thin and grey. The diagram is partially cut off by the left edge of the slide.

5. **Conclusions**



It's a start...



- Modular approach
- Toolkits available
- ML model quality
- NLP+ML evolving



The background of the slide features a light gray network pattern. It consists of numerous small circles, some solid and some hollow, connected by thin lines, creating a web-like structure across the entire image.

Thanks

Any questions?

lukasz.kniola@biogen.com

woo@xogene.com

Resources



<https://www.python.org>



<https://www.pdftron.com/pdfnet>



<https://spacy.io/>

regular expressions ₁₀₁

<https://regex101.com>