Data Anonymization for Local Crisis Response

a project by Aussie Frost for DSCI 410L

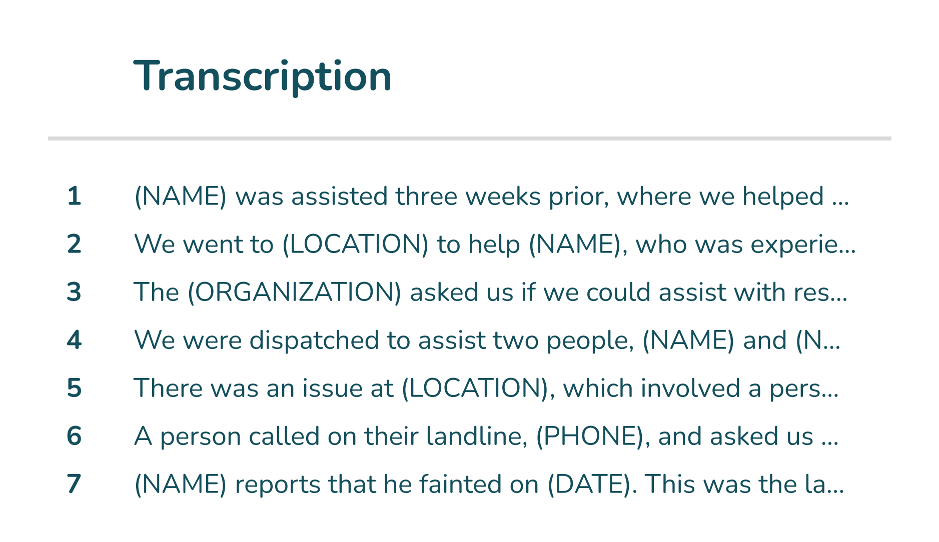
Background

- organizations are collecting data all the time
- this data can be analyzed to provide important insights
- CAHOOTS wants to share data with researchers *without compromising privacy*

Research question

Can a script reliably **remove identifying information** such as names, phone numbers, and addresses **from case narratives**?

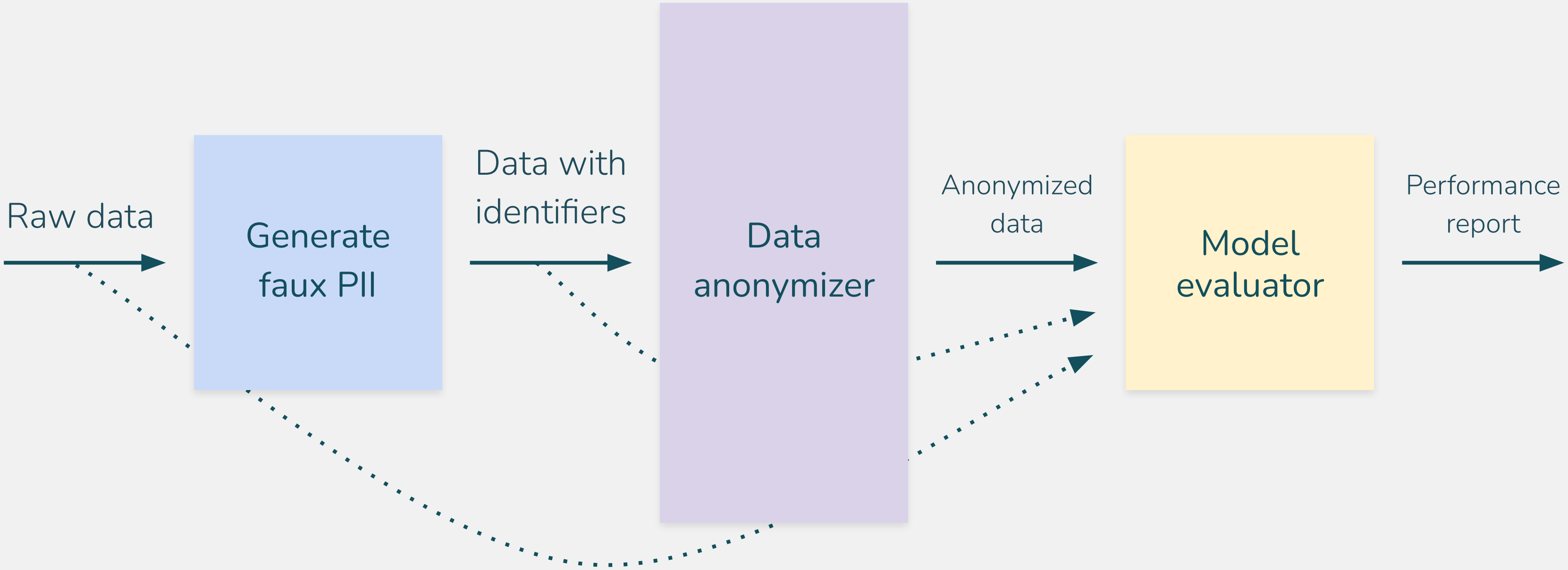
Data – Case narratives



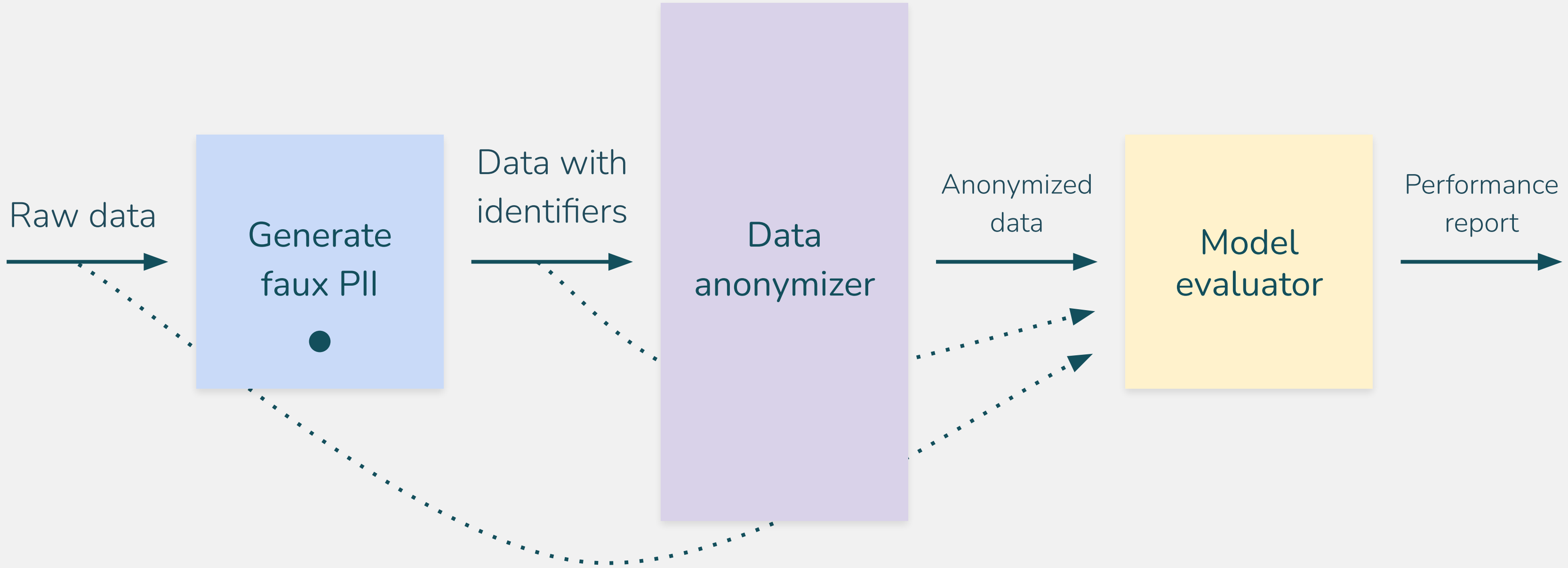
	Transcription
1	(NAME) was assisted three weeks prior, where we helped …
2	We went to (LOCATION) to help (NAME), who was experie…
3	The (ORGANIZATION) asked us if we could assist with res…
4	We were dispatched to assist two people, (NAME) and (N…
5	There was an issue at (LOCATION), which involved a pers…
6	A person called on their landline, (PHONE), and asked us …
7	(NAME) reports that he fainted on (DATE). This was the la…

Hand-anonymized Case Narratives courtesy of CAHOOTS

Methods – Anonymization script process

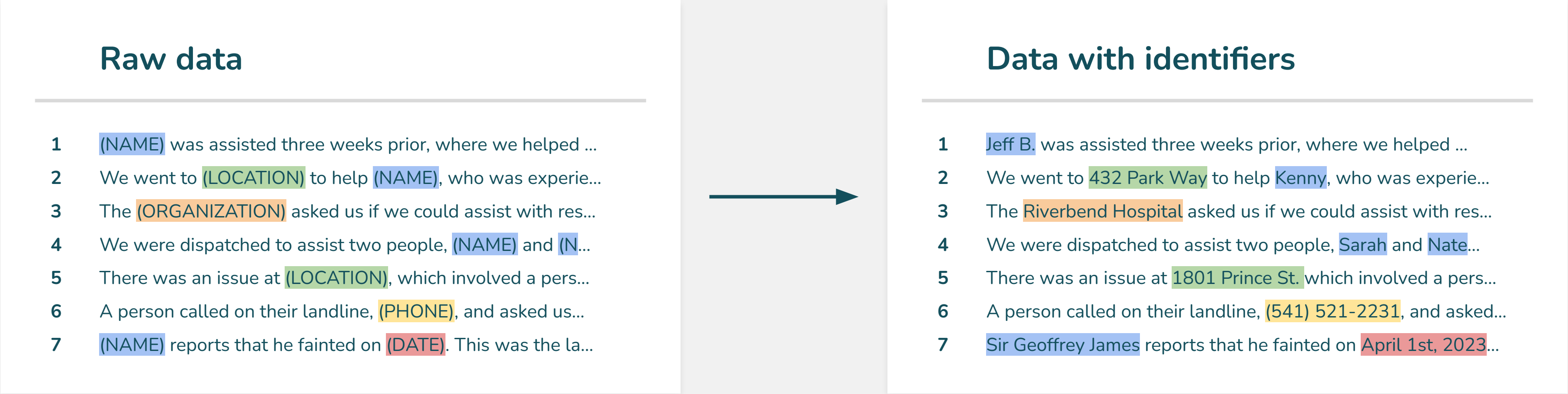


Methods – Anonymization script demo



Methods – Anonymization script demo

Generate faux PII | Replaces labels with randomly generated personal info



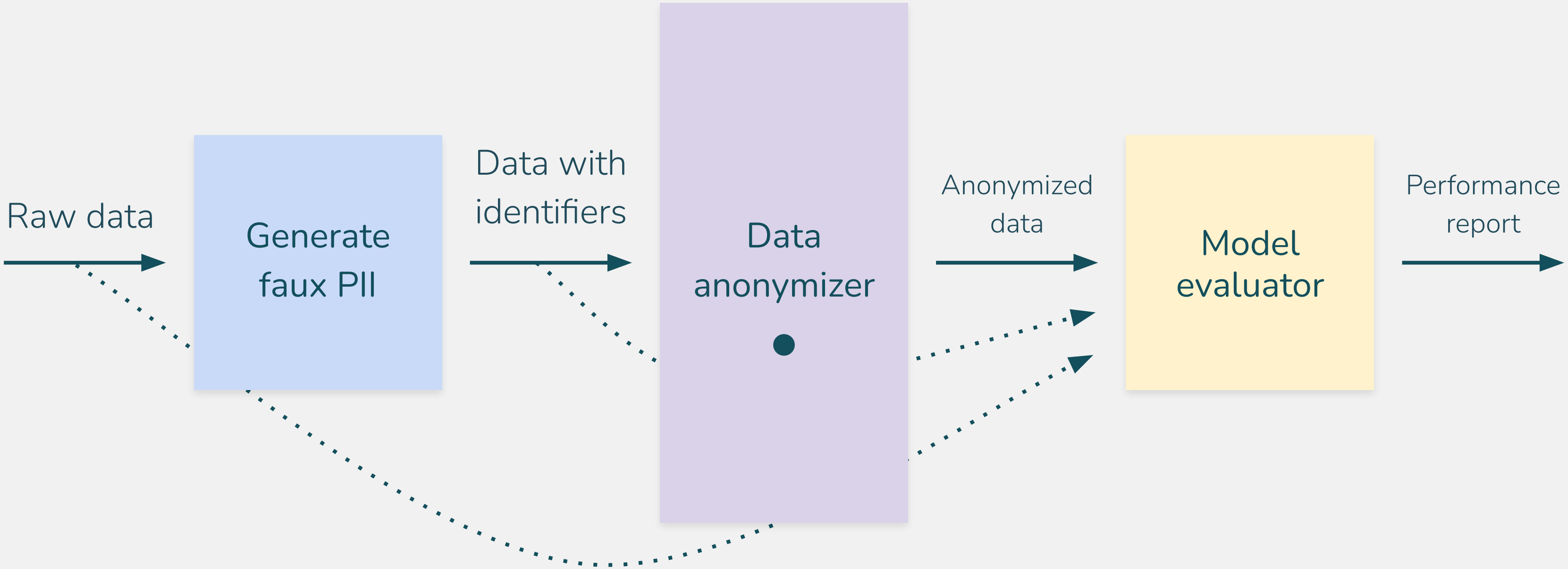
Raw data

1 (NAME) was assisted three weeks prior, where we helped …
2 We went to (LOCATION) to help (NAME), who was experie…
3 The (ORGANIZATION) asked us if we could assist with res…
4 We were dispatched to assist two people, (NAME) and (N…
5 There was an issue at (LOCATION), which involved a pers…
6 A person called on their landline, (PHONE), and asked us…
7 (NAME) reports that he fainted on (DATE). This was the la…

Data with identifiers

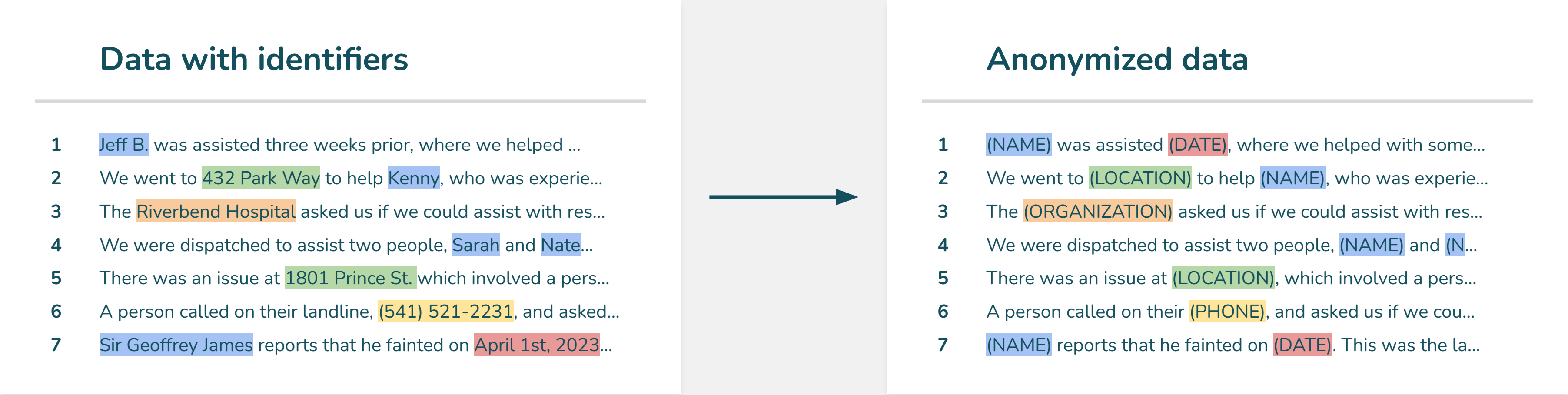
1 Jeff B. was assisted three weeks prior, where we helped …
2 We went to 432 Park Way to help Kenny, who was experie…
3 The Riverbend Hospital asked us if we could assist with res…
4 We were dispatched to assist two people, Sarah and Nate…
5 There was an issue at 1801 Prince St. which involved a pers…
6 A person called on their landline, (541) 521-2231, and asked…
7 Sir Geoffrey James reports that he fainted on April 1st, 2023…

Methods – Anonymization script demo



Methods – Anonymization script demo

Data anonymizer — Uses named entity recognition, RegEx, and predefined terms to identify PII



Data with identifiers

1 Jeff B. was assisted three weeks prior, where we helped ...

2 We went to 432 Park Way to help Kenny, who was experie...

3 The Riverbend Hospital asked us if we could assist with res...

4 We were dispatched to assist two people, Sarah and Nate...

5 There was an issue at 1801 Prince St. which involved a pers...

6 A person called on their landline, (541) 521-2231, and asked...

7 Sir Geoffrey James reports that he fainted on April 1st, 2023...

Anonymized data

1 (NAME) was assisted (DATE), where we helped with some...

2 We went to (LOCATION) to help (NAME), who was experie...

3 The (ORGANIZATION) asked us if we could assist with res...

4 We were dispatched to assist two people, (NAME) and (N...

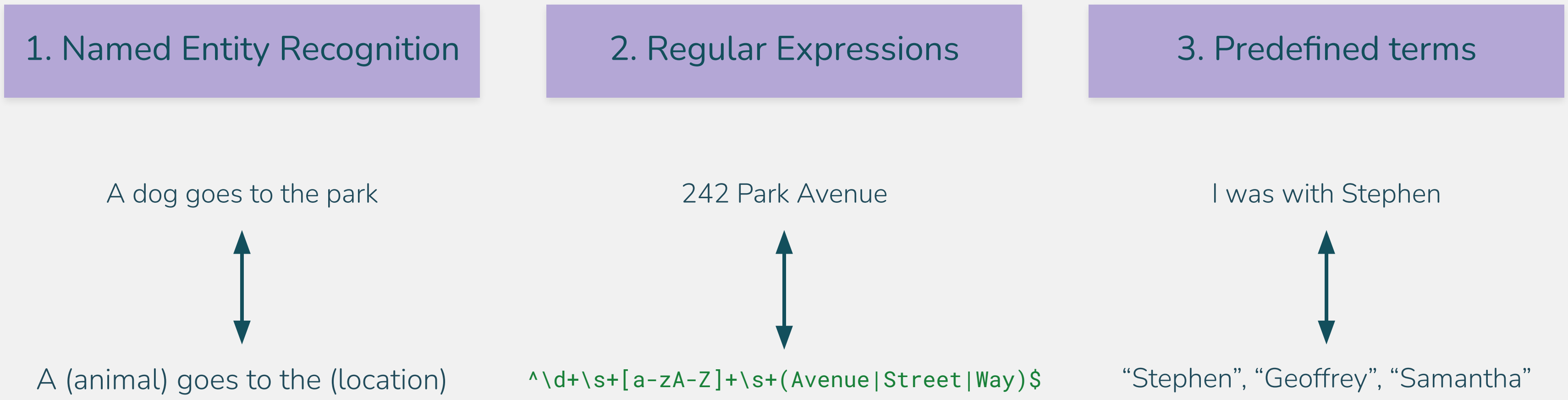
5 There was an issue at (LOCATION), which involved a pers...

6 A person called on their (PHONE), and asked us if we cou...

7 (NAME) reports that he fainted on (DATE). This was the la...

Methods – Anonymization script demo

Data anonymizer	Uses named entity recognition, RegEx, and predefined terms to identify PII



1. Named Entity Recognition

A dog goes to the park

A (animal) goes to the (location)

2. Regular Expressions

242 Park Avenue

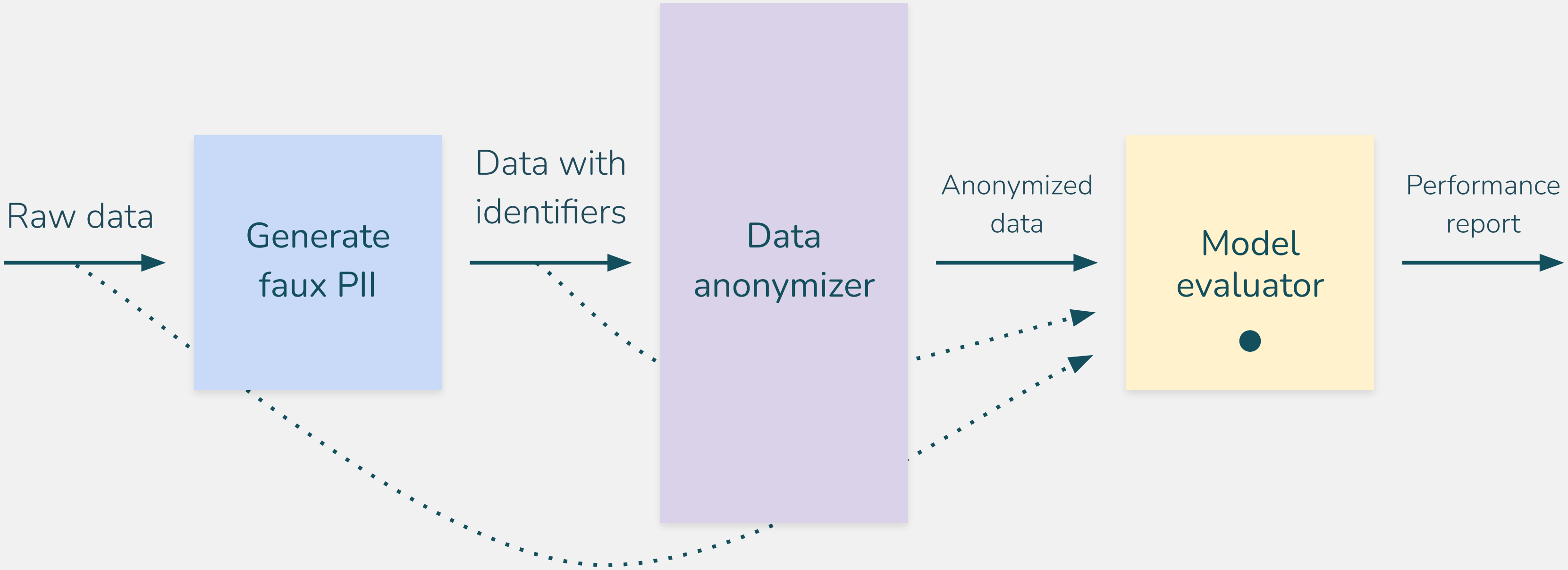
`^\d+\s+[a-zA-Z]+\s+(Avenue|Street|Way)$`

3. Predefined terms

I was with Stephen

“Stephen”, “Geoffrey”, “Samantha”

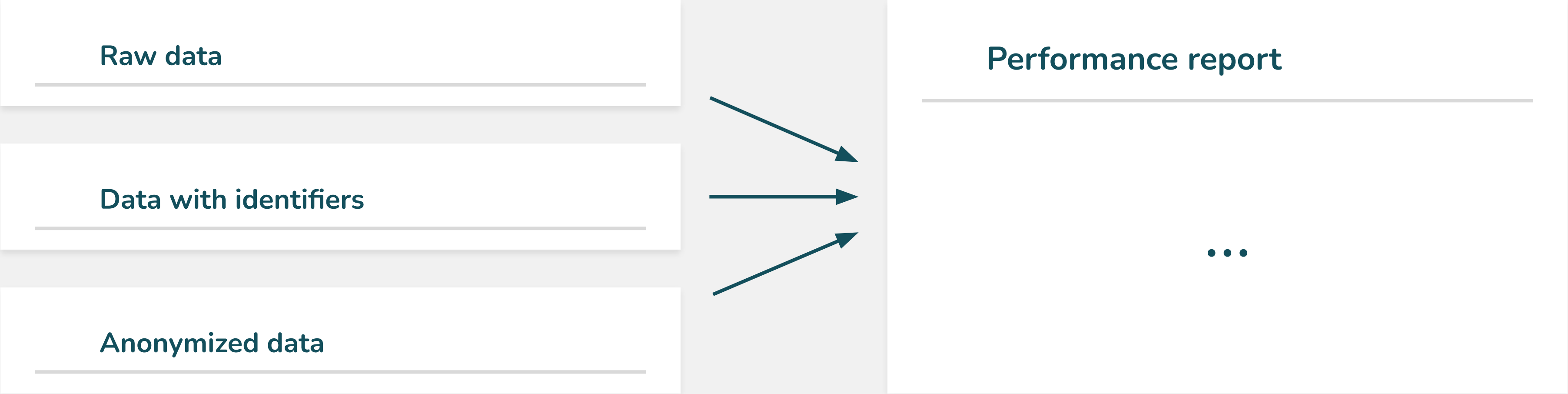
Methods – Anonymization script demo



Methods – Anonymization script demo

Model evaluator

Uses resulting data to evaluate model performance



Raw data

Data with identifiers

Anonymized data

Performance report

…

Results

Performance metrics (anonymized vs raw data)

Label	Recall	Precision
Name	0.92	0.39
Location	0.95	0.34
Date	1.0	0.04

Interpretation: Script correctly identifies most PII (high recall), but also classifies some non-PII as PII (low precision).

Discussion – Next steps

Findings:

- case narrative data anonymization is feasible with my script
  - 97.5% similarity between labeled datasets
  - PII removal is tricky with edge-cases

Future directions:

- lexical analysis – we can analyze the anonymized data for context
  - useful for understanding organization operations and impact

Acknowledgements

Thank you to Rori Rohlfs for providing the structure of this course

Thank you to peers for feedback and advice

Thank you to DS4SJ research group for providing collaborative insights

Thank you to CAHOOTS for providing the platform for this research project

thank you for attending!

Aussie Frost | ausdfrost@gmail.com