UK Data Service

Synthetic Data – Redaction & Masking

Joseph Allen, Research Associate

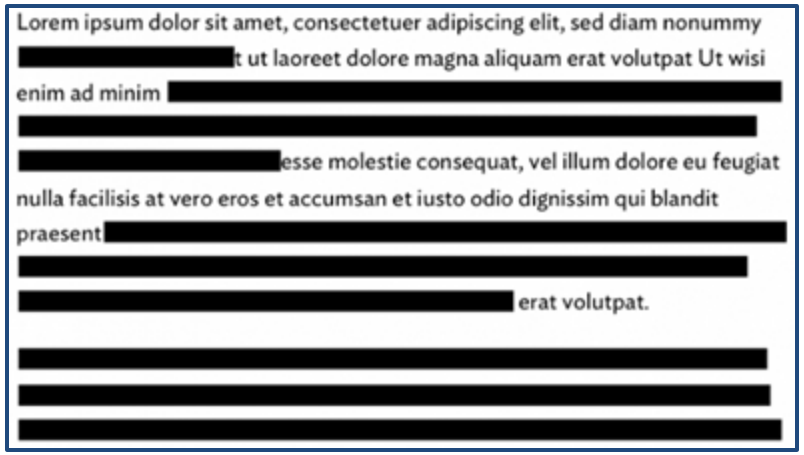
25 May 2021

Last time on Synthetic Data

1. What is Synthetic Data?
2. Why should we make synthetic data?
3. Examples, benefits and purposes of synthetic data.
4. Features of synthetic data.

This time on Synthetic Data

- Redaction - removing data entirely.
- Masking - replacing data with new, fake data.



Masking

- Replacing data, or parts of data with generated information.
  - Make up names yourself.
  - Use a data generation tool like Faker or Mockaroo.

ID	first_name	last_name	email	gender	address
2999	Grazia	Midden	gmidden@163.com	F	10 Pawling Center, M3 6GA

Masking Example

- Replace Grazia and Midden with a generated name.
- We have masked **first_name** & **last_name.**
- Notice email still holds real data on Grazia Midden.

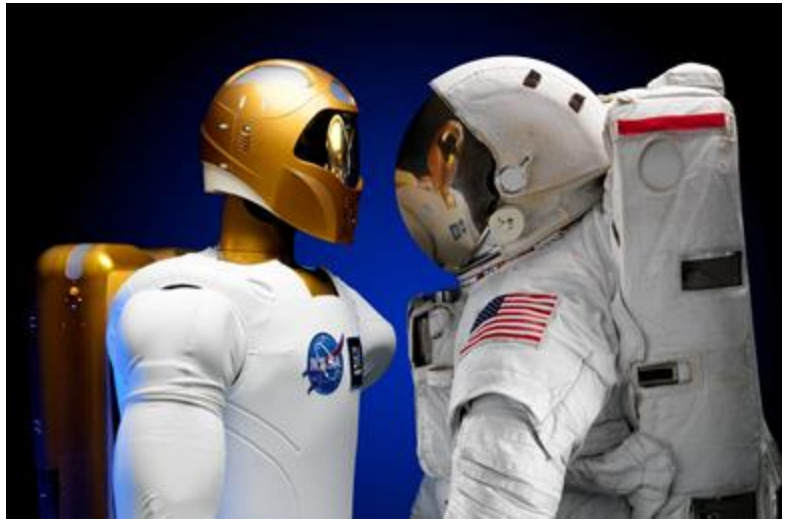
ID	first_name	last_name	email	gender	address
2999	Joe	Allen	gmidden@163.com	F	10 Pawling Center, M3 6GA

Simulation vs Masking

Simulation	Masking
Modern algorithms	Curated source of synthetic data
Irreversible	May be reversible
Requires technical investment	Relatively simple
Represents distributions of the real data	Quick sampling for GDPR compliance

Synthesis

- Simulation - applies modern algorithms to generate new data.
- Masking - replaces data with synthetic data from a curated source.



Reversibility

- Simulation - Uses noise to generate new, fake data. Not reversible.
- Masking - Uses real data. Potentially reversible.

Difficulty

- Simulation - Requires a large investment to get right.
- Masking - Requires a small technical effort.

Statistics

- Simulation - Maintains statistical distributions.
- Masking - Random, but realistic values.

Best practices

- Explore your data fully.
- Understand the context of data before synthesis.
- Use irreversible methods when possible.
- Persist the structure of data.

GDPR Compliance

Article 6 (4-e): “the existence of appropriate safeguards, which may include encryption or pseudonymization.”

Warning

- You likely do not have the right to distribute data.
- Ask your data provider.

2020 Census Manchester

Case Study – the data

ID	first_name	last_name	email	gender	address
2999	Grazia	Midden	gmidden@163.com	F	Apt 38 Sillivan Way Manchester M36GD
3000	Andy	DuFrens	adufrens@123.com	M	66 Myrtle Road London N13 5QX

Case Study - ID

ID	first_name	last_name	email	gender	address
2999	Grazia	Midden	gmidden@163.com	F	Apt 38 Sillivan Way Manchester M36GD
3000	Andy	DuFrens	adufrens@123.com	M	66 Myrtle Road London N13 5QX

Reindexed ID

ID	first_name	last_name	email	gender	address
9000	Grazia	Midden	gmidden@163.com	F	Apt 38 Sillivan Way Manchester M36GD
9001	Andy	DuFrens	adufrens@123.com	M	66 Myrtle Road London N13 5QX

Case Study - Names

ID	first_name	last_name	email	gender	address
9000	Grazia	Midden	gmidden@163.com	F	Apt 38 Sillivan Way Manchester M36GD
9001	Andy	DuFrens	adufrens@123.com	M	66 Myrtle Road London N13 5QX

New Names

ID	first_name	last_name	email	gender	address
9000	Georgina	Shicky	gmidden@163.com	F	Apt 38 Sillivan Way Manchester M36GD
9001	Joseph	Allen	adufrens@123.com	M	66 Myrtle Road London N13 5QX

Case Study - Email

ID	first_name	last_name	email	gender	address
9000	Georgina	Shicky	gmidden@163.com	F	Apt 38 Sillivan Way Manchester M36GD
9001	Joseph	Allen	adufrens@123.com	M	66 Myrtle Road London N13 5QX

Connected Email

ID	first_name	last_name	email	gender	address
9000	Georgina	Shicky	gshicky@123.com	F	Apt 38 Sillivan Way Manchester M36GD
9001	Joseph	Allen	jallen@678.com	M	66 Myrtle Road London N13 5QX

Case Study - Gender

ID	first_name	last_name	email	gender	address
9000	Georgina	Shicky	gshicky@123.com	F	Apt 38 Sillivan Way Manchester M36GD
9001	Joseph	Allen	jallen@678.com	M	66 Myrtle Road London N13 5QX

Case Study - Address

ID	first_name	last_name	email	gender	address
9000	Georgina	Shicky	gshicky@123.com	F	Apt 38 Sillivan Way Manchester M36GD
9001	Joseph	Allen	jallen@678.com	M	66 Myrtle Road London N13 5QX

Synthetic Address

ID	first_name	last_name	email	gender	address
9000	Georgina	Shicky	gshicky@123.com	F	Apt 66 Filey Road Manchester M3 6A
9001	Joseph	Allen	jallen@678.com	M	66 Chancery lane London N13 5QX

Case Study - Redaction

ID	first_name	last_name	email	gender	address
9000	Georgina	Shicky	gshicky@123.com	F	Apt 66 Filey Road Manchester M3 6A
9001	Joseph	Allen	jallen@678.com	M	66 Chancery lane London N13 5QX

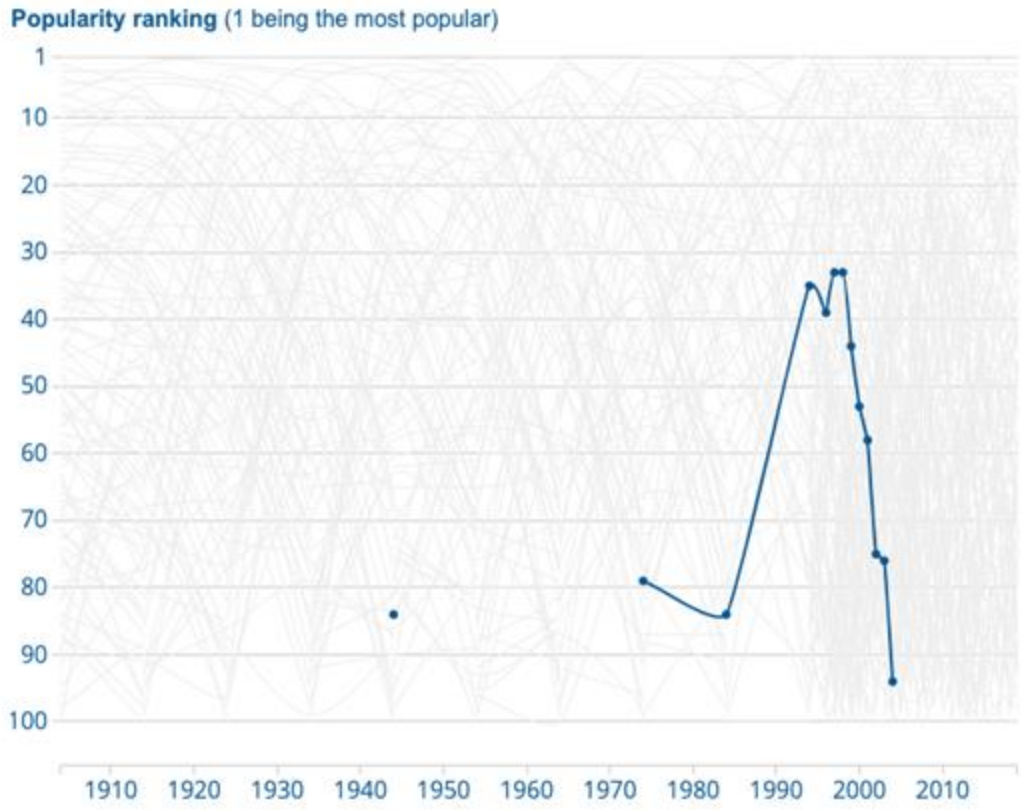
Case Study - Synthesis Process

- Generate a new index
- Generate a first_name and last_name
- Generate a realistic email & gender
- Generate a new address

ID	first_name	last_name	email	gender	address
9000	Georgina	Shicky	gshicky@123.com	F	Apt 66 Filey Road Manchester M3 6A
9001	Joseph	Allen	jallen@678.com	M	66 Chancery lane London N13 5QX

What’s in a name?

Ranking of the name “Georgina” over time



Name - Assumptions

ID	ethnicity	Age	hasEmail	gender	address
9000	Italian	20-30	True	F	Apt 66 Filey Road Manchester M3 6A

Case Study - Final data

ID	gender	address
9000	F	M3

Case Study - Documentation

- We’ve been forced to make a lot of assumptions:
  - F is female.
  - M is male.
  - IDs must have 4 digits .
  - No rows are missing.
  - ID, gender and a coarsened address are sufficient.

Case Study - Redaction Process

- Redacted rows not in Manchester.
- Redacted free text columns - first name, last name and email.
- Redacted all address data beyond the city and top level postcode.

ID	gender	address
9000	F	M3

Disclosure Control - Masking

Masking Techniques 1

- Substitution - Replacing data with data from a source.
- Shuffling - Changing the order of data.
- Variance - Adding small amounts of noise to data.
- Encryption - Traditional encryption using a key.

Masking Techniques 2

- Scrambling - Changing the order of characters in data.
- Nulling out - replacing data with null values.
- Masking out - replacing parts of the data with a clear fake value.

Sharing Data

Method	Suitable for unauthorized users?
Substitution	Maybe
Shuffling	No
Variance	No
Encryption	Maybe
Scrambling	No
Nulling out	Maybe
Masking out	Maybe

Substitution

Method	Anonymous	Indistinguishable	Technical	Practical
Substitution	Maybe	Maybe	Yes	Yes

Name	Age
Grazia Midden	24
Andy Otto	45

Name	Age
Andy Dufrens	24
Joseph Allen	45

Shuffling

Method	Anonymous	Indistinguishable	Technical	Practical
Shuffling	No	No	Yes	No

Name	Age
Grazia Midden	24
Andy Otto	45

Name	Age
Andy Otto	24
Grazia Midden	45

Variance

Method	Anonymous	Indistinguishable	Technical	Practical
Variance	No	Yes	Yes	Yes

Name	Age
Grazia Midden	24
Andy Otto	45

Name	Age
Grazia Midden	26
Andy Otto	44

Encryption

Method	Anonymous	Indistinguishable	Technical	Practical
Encryption	Maybe	No	Yes	Yes

Name	Age
Grazia Midden	24
Andy Otto	45

Name	Age
Rw10”£	24
S)!”SS	45

Scrambling

- Randomly rearranging the characters in a string.

Method	Anonymous	Indistinguishable	Technical	Practical
Scrambling	No	No	Yes	No

Name	Age
Grazia Midden	24
Andy Otto	45

Name	Age
zaairG dMinde	24
dyAn tOto	45

Nulling out

- Replace data with null values.
- Similar to redaction, but shows we intentionally removed the data.

Name	Age
Grazia Midden	24
Andy Otto	45

Name	Age
NULL	24
NULL	45

Masking out

- Replace some of the data with fake data.
- Commonly used to show your credit card number as
  - XXXX XXXX XXXX 1234

Name	Age
Grazia Midden	24
Andy Otto	45

Name	Age
NAME	24
NAME	45

Masking Comparison

Method	Anonymous	Indistinguishable	Technical	Practical
Substitution	Maybe	Maybe	Yes	Yes
Shuffling	No	No	Yes	No
Variance	No	Yes	Yes	Yes
Encryption	Maybe	No	Yes	Yes
Scrambling	No	No	Yes	No
Nulling out	Maybe	No	Maybe	Yes
Masking out	Maybe	No	Maybe	Yes

Disclosure Control - Redaction

Redaction

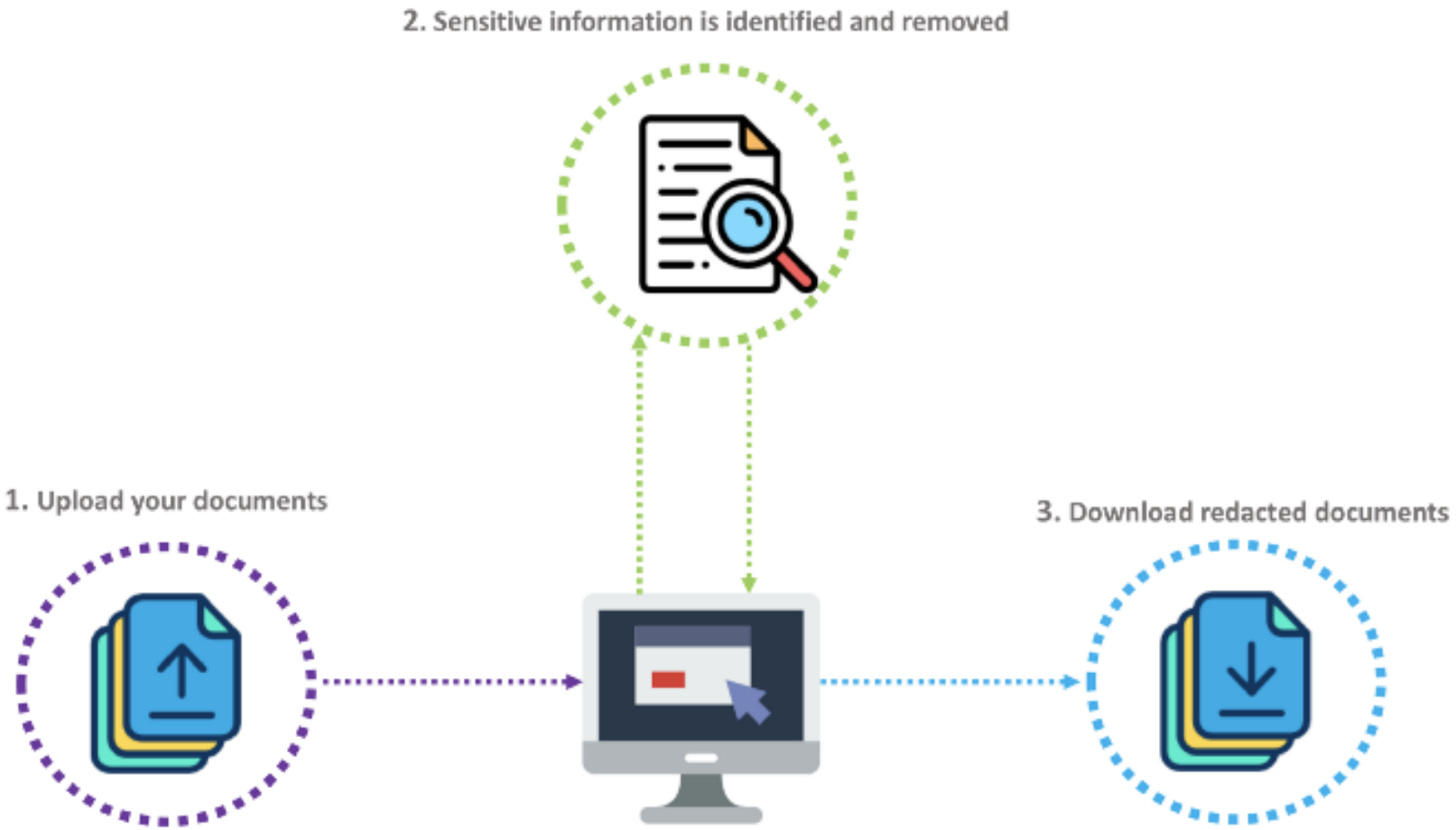
Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy [redacted]t ut laoreet dolore magna aliquam erat volutpat Ut wisi enim ad minim [redacted] [redacted] [redacted]esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent [redacted] [redacted] [redacted] erat volutpat.

[redacted] [redacted] [redacted]

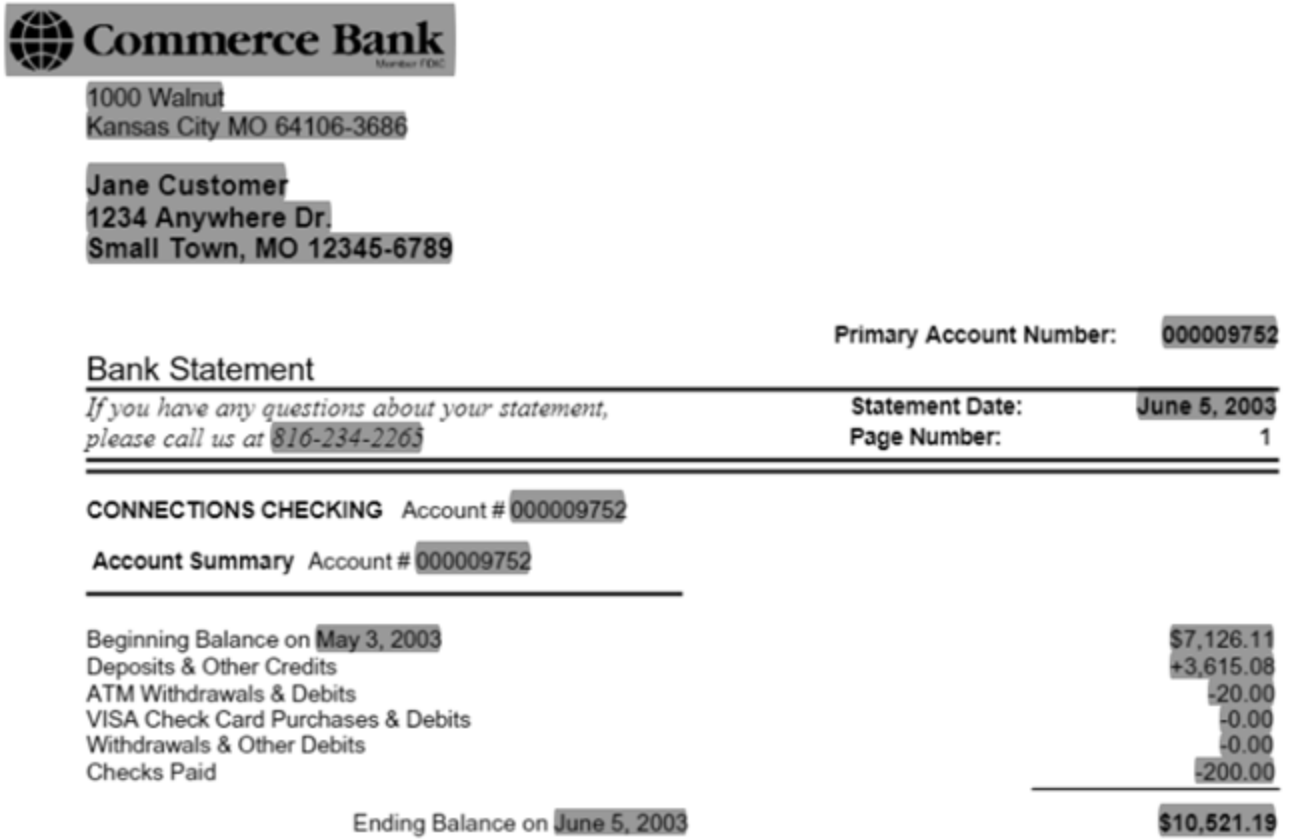
What is Redaction?

- Redaction - Removing sensitive data from our dataset
  - Sensitive - military bases and hospitals in a dataset.
  - Outliers - sometimes being an outlier is enough to identify individuals.

Automated Redaction



Redacted Bank Statement



Commerce Bank
Member FDIC

1000 Walnut
Kansas City MO 64106-3686

Jane Customer
1234 Anywhere Dr.
Small Town, MO 12345-6789

Primary Account Number: 000009752

Bank Statement

If you have any questions about your statement, please call us at 816-234-2265

Statement Date: June 5, 2003
Page Number: 1

CONNECTIONS CHECKING Account # 000009752

Account Summary Account # 000009752

Beginning Balance on May 3, 2003	$7,126.11
Deposits & Other Credits	+3,615.08
ATM Withdrawals & Debits	-20.00
VISA Check Card Purchases & Debits	-0.00
Withdrawals & Other Debits	-0.00
Checks Paid	-200.00
Ending Balance on June 5, 2003	$10,521.19

Sisters Dataset

Student ID	Average grade	Number of sisters
1	B	1
2	A	2
3	A	11
4	B	1

Student 3

- Knowing which student has 11 sisters might reveal the identity and grade of this record.

Student ID	Average grade	Number of sisters
1	B	1
2	A	2
3	A	11
4	B	1

Student 2

- What about Student 2?
  - Student 2 could be uniquely identified as the only student with 2 sisters.

Student ID	Average grade	Number of sisters
1	B	1
2	A	2
3	A	11
4	B	1

Protected students

- Students 1 and 4 cannot be uniquely identified.

Student ID	Average grade	Number of sisters
1	B	1
2	A	2
3	A	11
4	B	1

Outliers

- Outliers can be the most important
- Only with context, can we redact appropriately

Redaction - Row

- We remove our obvious outlier, with 11 sisters.
- Implied correlation between grades and number of sisters.

Student ID	Average grade	Number of sisters
1	B	1
2	A	2
4	B	1

Redaction - Columns

- Free Text is rarely useful in Machine Learning.
- They can become useful with pre-processing.

ID	Age	Name
1	12	Pattie
2	15	Daren
3	13	Shara
4	8	Andy

Redaction - Part

- In some cases redacting only part of the data may be useful.
  - Joe Allen -> Joe
  - 42 Wallaby Way -> Wallaby Way
  - 1234 5678 1234 1234 -> XXXX XXXX XXXX 1234
- Coarsening, Nulling Out and Redaction all look the same.

Conclusion

- Redaction is a quick-fix to remove protected fields.
- Masking provides context, whilst requiring some technical skills.
- Use the substitution method with a data generation library for protected fields.
- Use variance sparingly on numeric fields.

Next time on Synthetic Data

- Coarsening - Reducing precision of data
- Mimicking - generate a dataset that closely matches the real dataset but does not contain exactly the same entries
- Simulation - Generating new data from observed patterns.

Further Reading

- Data in Government Blog tinyurl.com/Synth-DataInGov
- What is Data Masking? research.aimultiple.com/data-masking
- Revolutionising Redaction towardsdatascience.com/revolutionising-redaction-my-final-year-project-fe664e28ef84

Questions

Twitter @JosephAllen1234

Joseph Allen

Email Joseph.allen@manchester.ac.uk

UK Data Service