

Synthetic Data: An introductory workshop

Dr. J. Kasmire

Senior Research Fellow

Cathie Marsh Institute/UK Data Service

Table of Contents for this workshop

Synthetic data – what it is and is not

Fidelity – an important concept for synthetic data

Uses of synthetic data

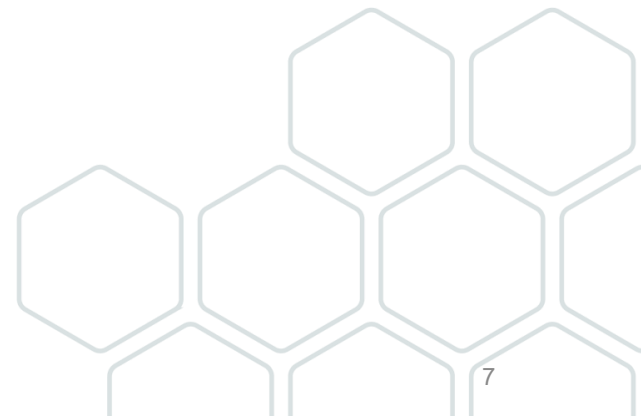
Generating synthetic data

Break

Hands on session/code demo



Synthetic data



What is synthetic data?

Synthetic data is any data that is generated rather than observed.

Examples of synthetic data 1

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed at turpis vitae velit euismod aliquet. Pellentesque et arcu. Nullam venenatis gravida orci. Pellentesque et arcu. Nam pharetra. Vestibulum viverra varius enim.

Nam laoreet dui sed magna. Nunc in turpis ac lacus eleifend sagittis. Pellentesque ac turpis. Aliquam justo lectus, iaculis a, auctor sed, congue in, nisl. Aenean luctus vulputate turpis. Mauris urna sem, suscipit vitae, dignissim id, ultrices sed, nunc.

Phasellus nisi metus, tempus sit amet, ultrices ac, porta nec, felis. Quisque malesuada nulla sed pede volutpat pulvinar. Sed non ipsum. Mauris et dolor. Pellentesque suscipit accumsan massa. In consectetur, lorem eu lobortis egestas, velit odio imperdiet eros, sit amet sagittis nunc mi ac neque. Nulla blandit justo a metus. Ut venenatis.

Beao, Public domain, via Wikimedia Commons

Examples of synthetic data 2

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed at turpis vitae velit euismod aliquet. Pellentesque et arcu. Nullam venenatis gravida orci. Pellentesque et arcu. Nam pharetra. Vestibulum viverra varius enim.

Nam laoreet dui sed magna. Nunc in turpis ac lacus eleifend sagittis. Pellentesque ac turpis. Aliquam justo lectus, iaculis a, auctor sed, congue in, nisl. Aenean luctus vulputate turpis. Mauris urna sem, suscipit vitae, dignissim id, ultrices sed, nunc.

Phasellus nisi metus, tempus sit amet, ultrices ac, porta nec, felis. Quisque malesuada nulla sed pede volutpat pulvinar. Sed non ipsum. Mauris et dolor. Pellentesque suscipit accumsan massa. In consectetur, lorem eu lobortis egestas, velit odio imperdiet eros, sit amet sagittis nunc mi ac neque. Nulla blandit justo a metus. Ut venenatis.

Beao, Public domain, via Wikimedia Commons



Figure uploaded by [Claus Aranha](#)
<https://doi.org/10.48550/arXiv.1904.09135>

Examples of synthetic data 3

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed at turpis vitae velit euismod aliquet. Pellentesque et arcu. Nullam venenatis gravida orci. Pellentesque et arcu. Nam pharetra. Vestibulum viverra varius enim.

Nam laoreet dui sed magna. Nunc in turpis ac lacus eleifend sagittis. Pellentesque ac turpis. Aliquam justo lectus, iaculis a, auctor sed, congue in, nisl. Aenean luctus vulputate turpis. Mauris urna sem, suscipit vitae, dignissim id, ultrices sed, nunc.

Phasellus nisi metus, tempus sit amet, ultrices ac, porta nec, felis.

Quisque malesuada nulla sed pede volutpat pulvinar. Sed non ipsum. Mauris et dolor. Pellentesque suscipit accumsan massa. In consectetur, lorem eu lobortis egestas, velit odio imperdiet eros, sit amet sagittis nunc mi ac neque. Nulla blandit justo a metus. Ut venenatis.

Beao, Public domain, via Wikimedia Commons



Anderson Rodrigues, from Pexels



Figure uploaded by [Claus Aranha](#)
<https://doi.org/10.48550/arXiv.1904.09135>

Examples of synthetic data 4

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed at turpis vitae velit euismod aliquet. Pellentesque et arcu. Nullam venenatis gravida orci. Pellentesque et arcu. Nam pharetra. Vestibulum viverra varius enim.

Nam laoreet dui sed magna. Nunc in turpis ac lacus eleifend sagittis. Pellentesque ac turpis. Aliquam justo lectus, iaculis a, auctor sed, congue in, nisl. Aenean luctus vulputate turpis. Mauris urna sem, suscipit vitae, dignissim id, ultrices sed, nunc.

Phasellus nisi metus, tempus sit amet, ultrices ac, porta nec, felis.

Quisque malesuada nulla sed pede volutpat pulvinar. Sed non ipsum. Mauris et dolor. Pellentesque suscipit accumsan massa. In consectetur, lorem eu lobortis egestas, velit odio imperdiet eros, sit amet sagittis nunc mi ac neque. Nulla blandit justo a metus. Ut venenatis.

Beao, Public domain, via Wikimedia Commons



Anderson Rodrigues, from Pexels

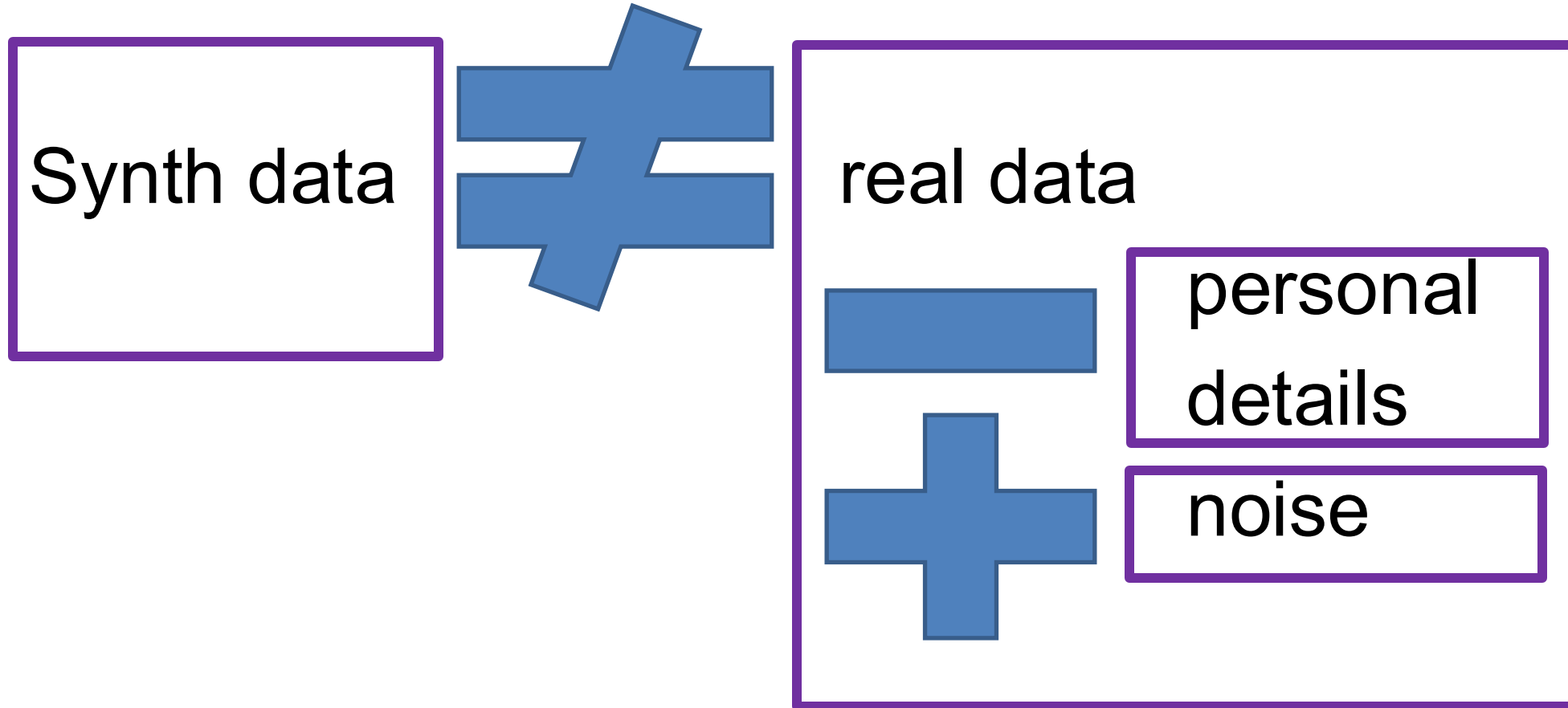


Figure uploaded by [Claus Aranha](#)
<https://doi.org/10.48550/arXiv.1904.09135> u



Armando Are, from Pexels

What is not synthetic data?

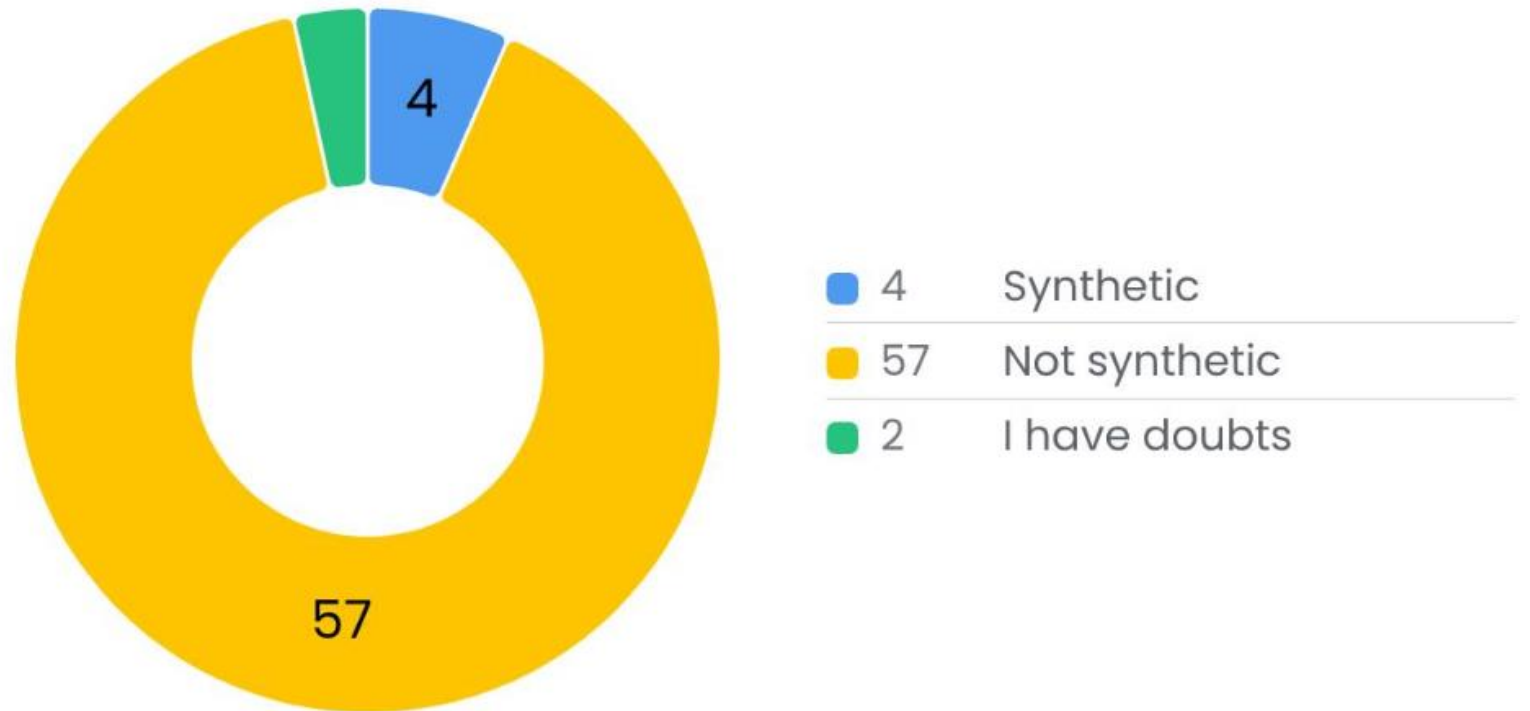


Terminology is important

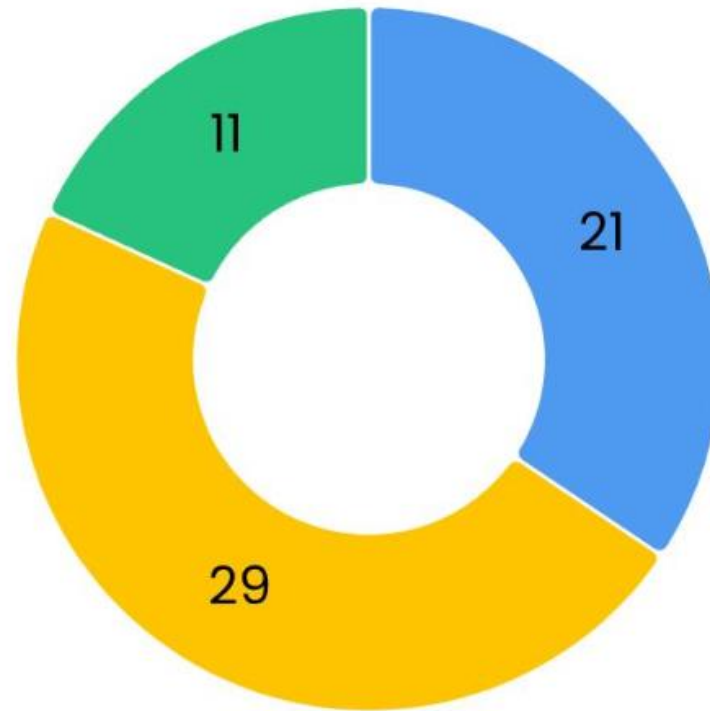
Some people use 'synthetic data' when they mean 'anonymised', 'de-personalised', etc.

They are wrong.

Is data from a cycle lane sensor synthetic?

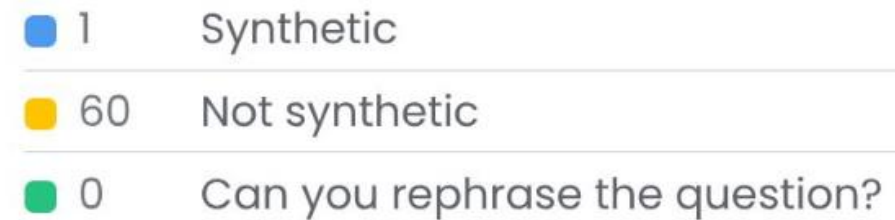
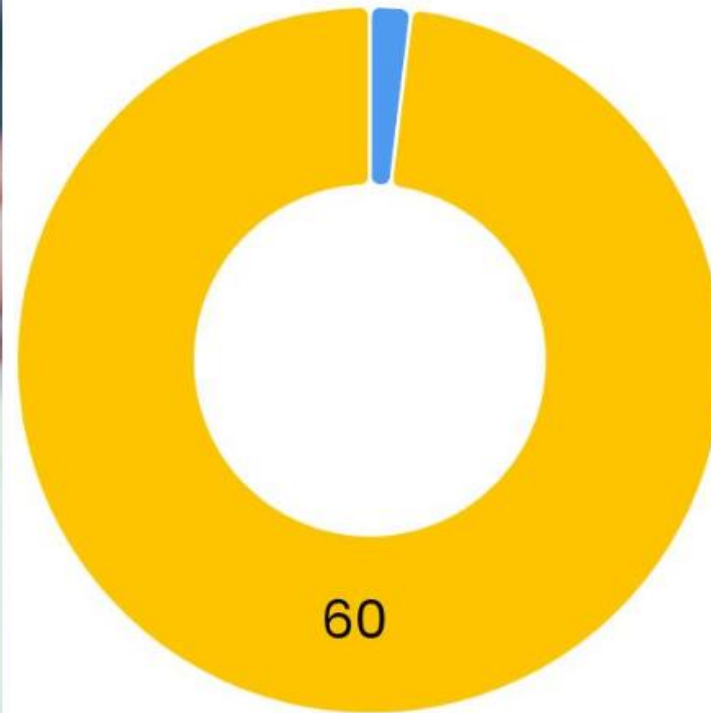


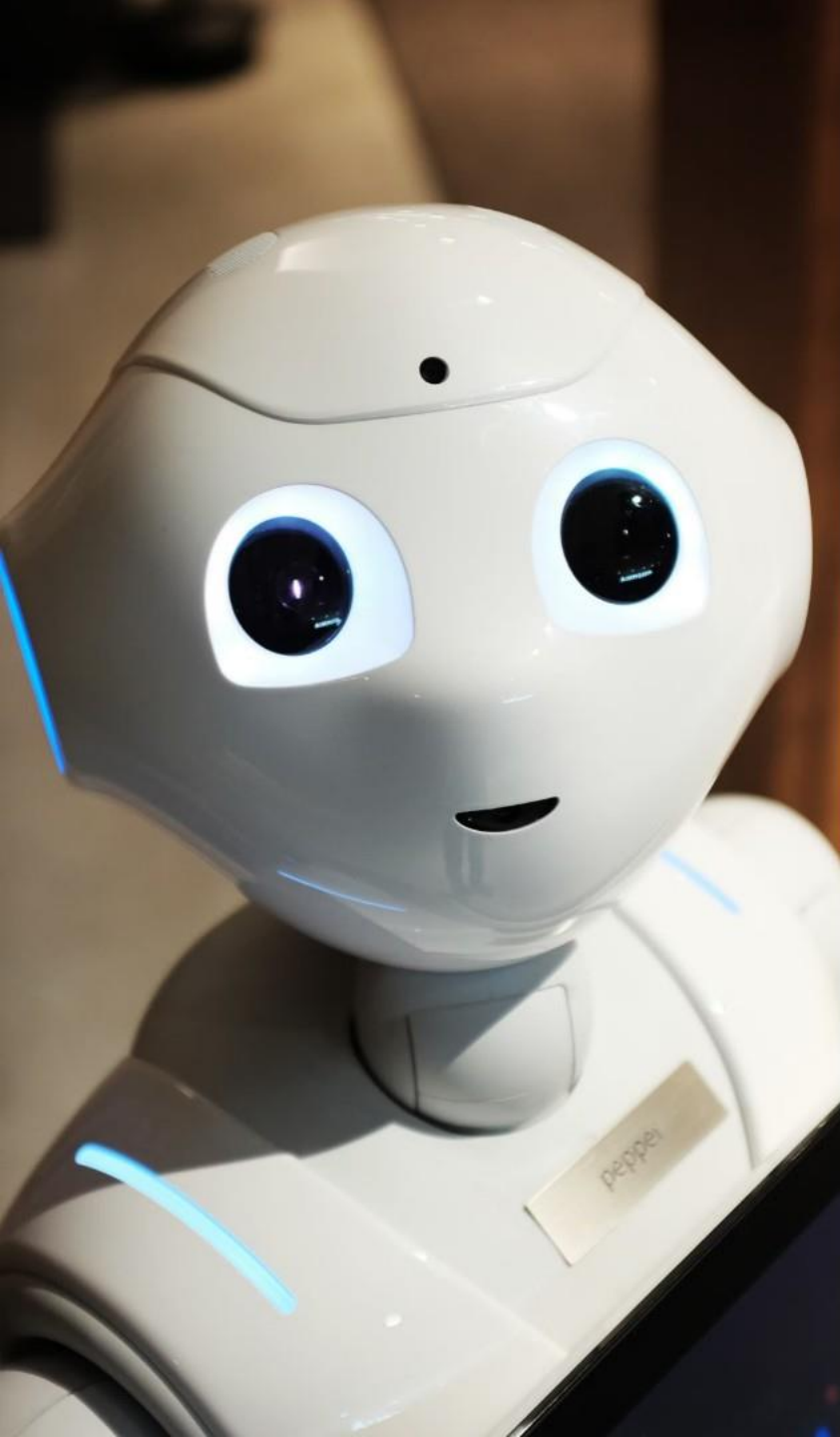
Are the predictions from weather forecasts synthetic data?



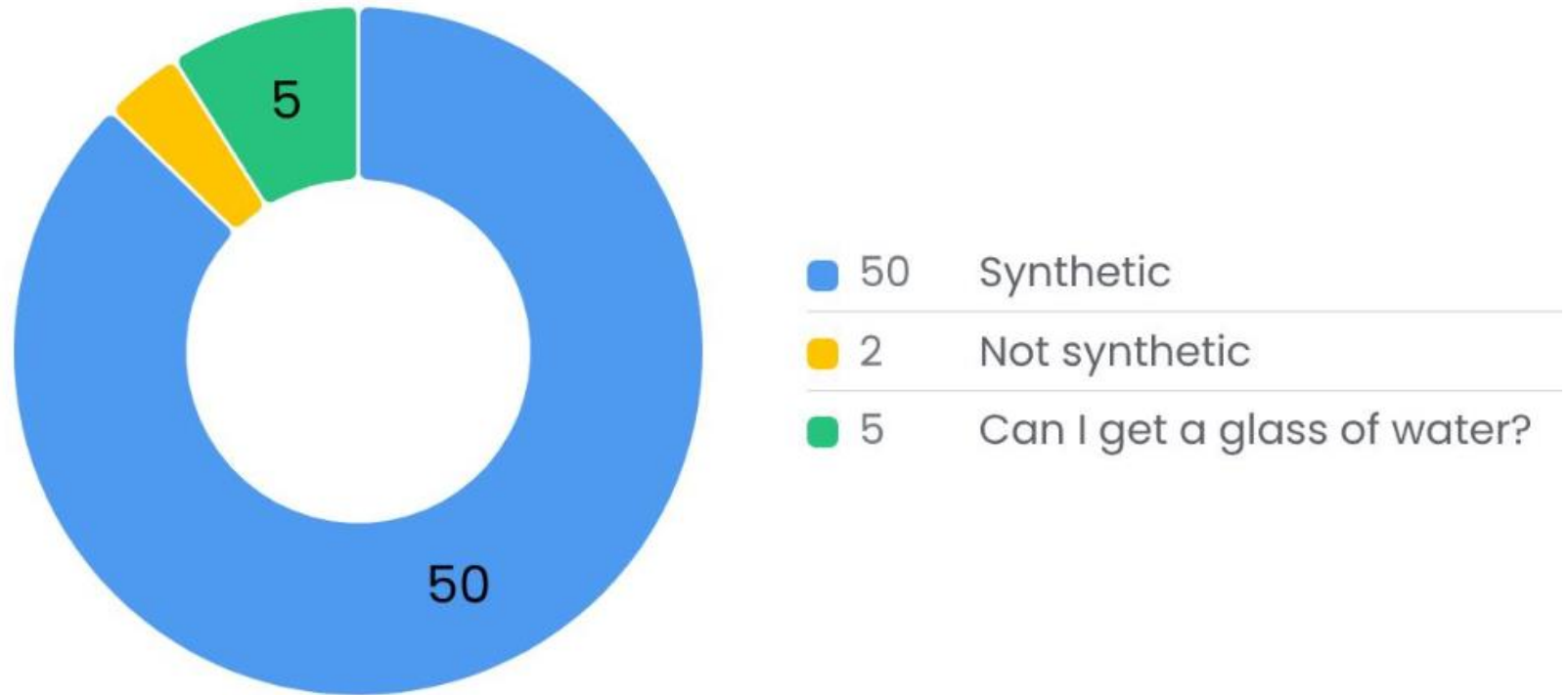
21	Synthetic
29	Not Synthetic
11	I... uh... not sure.

Is census microdata synthetic?



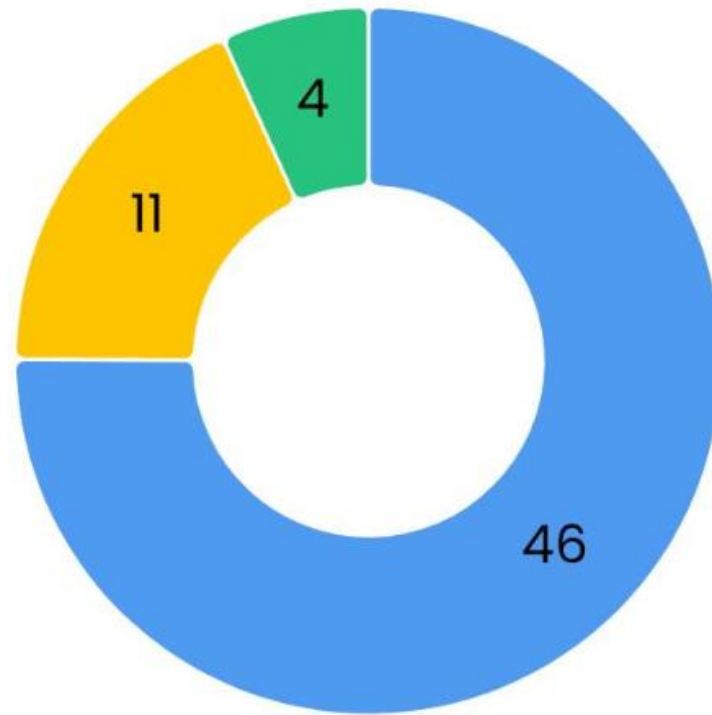


Is the output from Chat GPT synthetic?



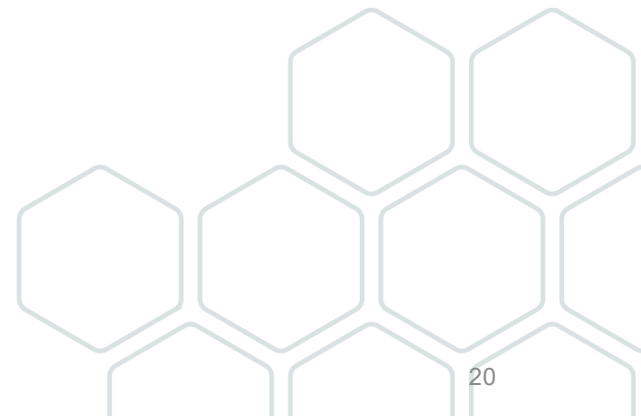


Is stuff you just made up in your head synthetic?



46	Synthetic
11	Not synthetic
4	What IS that behind you?!?

Fidelity



What is fidelity?

Fidelity means faithfulness.

But data faithfulness is not binary.

Synthetic data can be faithful on:

- number and/or type of variables?
- mean, range, standard deviation, distribution, etc.?
- documentation?
- volume of data?
- relationships between variables?
- other features?

Example

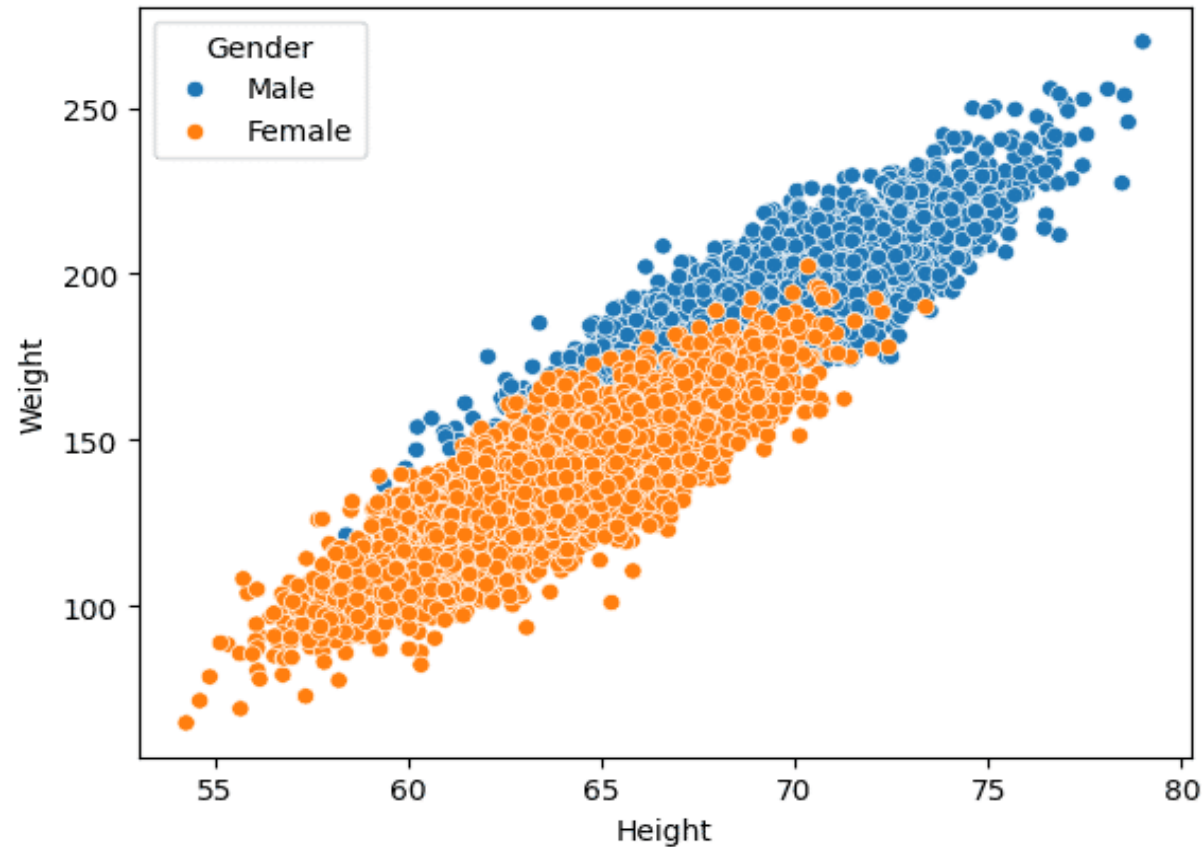
Real

Q	X	Y	Z
1	3	45	100
0	9	40	200
0	10	40	300
1	9	55	200

Synthetic

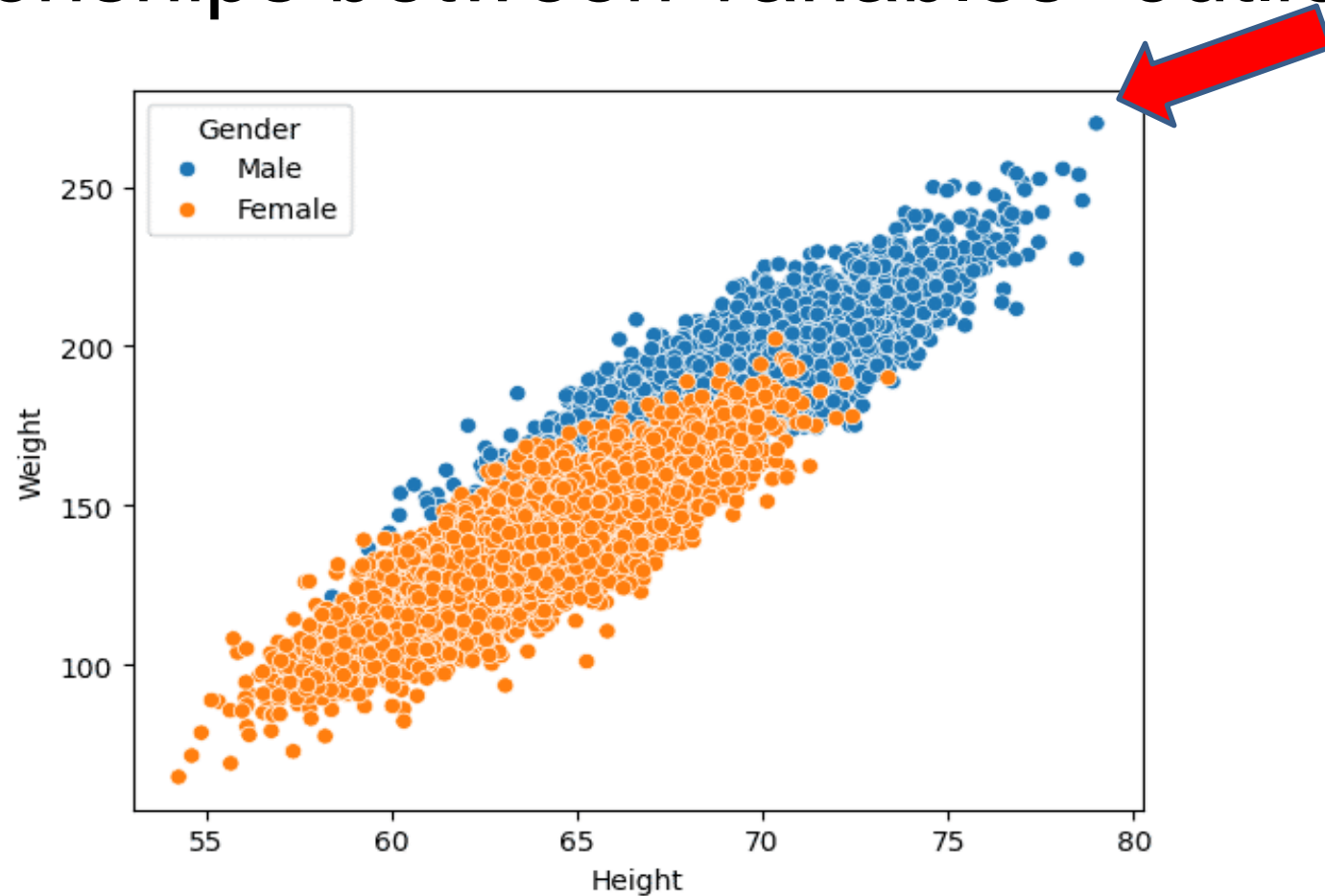
Q	X	Y	Z
3	3	45	100.63
3	6	45	234.98
4	5	45	297.22
4	8	45	174.99

Relationships between variables



Graph created by J. Kasmire using Python and Seaborn.
Data sources: height and weight data from Kaggle.com originally scraped and compiled from Statistics Online Computational Resources.

Relationships between variables -outliers



Graph created by J. Kasmire using Python and Seaborn.
Data sources: height and weight data from Kaggle.com originally scraped and compiled from Statistics Online Computational Resources.

Risk?

Synthetic data has no disclosure risk, because there is no *real* data that can be disclosed.

But (near) matches of high-fidelity data have an indication risk.

High-fidelity synthetic data

Nothing can be 100% faithful or it would just be identical to the real-world data.

Must be custom built to suit the particular dataset, research question, use case and generation method.

Need to clarify exactly *which* features of the original need to be faithful, which can be unfaithful, and how.

Greater fidelity is not always better

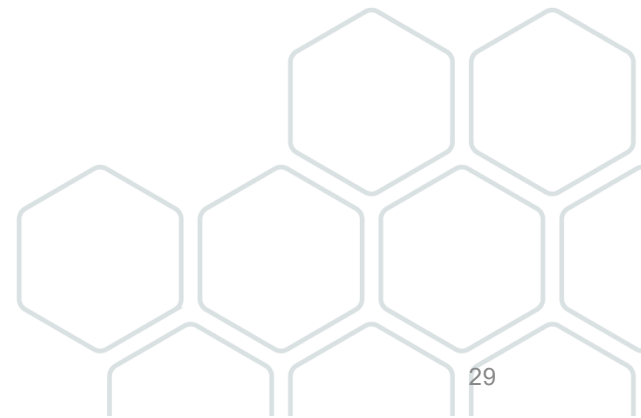
Some synthetic data is more useful if it is NOT faithful to the original in some ways.

Example:

Goal = AI tool to categorise skin lesion pictures by risk

Problem = real data has relatively few photos of skin lesions on POC

Purposes for synthetic data



Preview



Flávia Vicentini, from Pexels



1- Preview

Small

demonstrates structure/style

These should be faithful to:

- Number and type of fields
- Format
- Features that make the data unique
- May indicate features/relationships of interest

Proof of concept



Markus Winkler, from Pexels



2 - Proof of concept

Sufficiently large

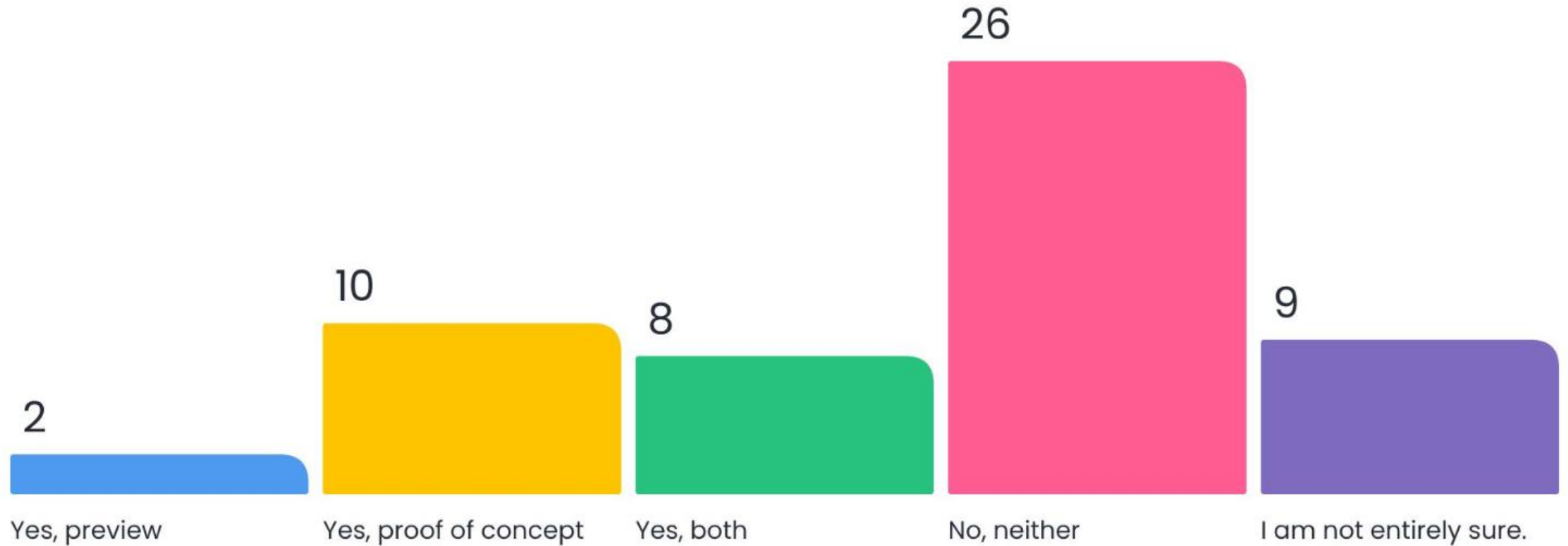
usefully demonstrates that:

- Concepts theoretically could work on the real data
- Visualisations/maps/outputs are useful

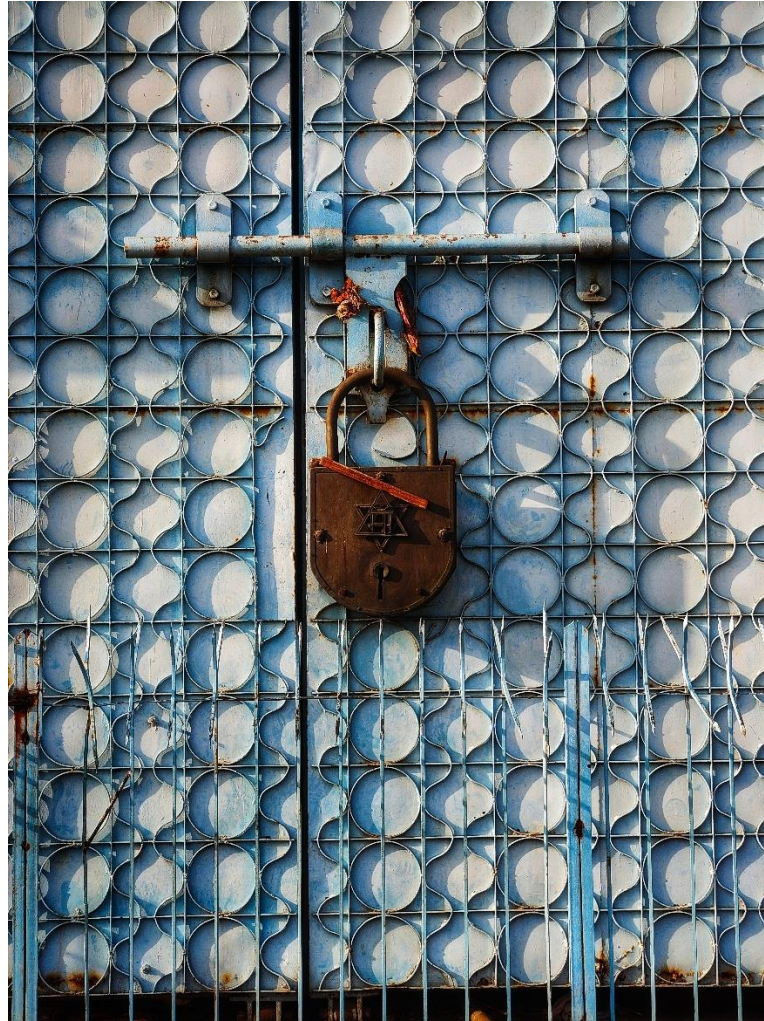
May be:

- Preliminary step
- Called a “toy dataset”

Have you ever used data for preview or proof of concept purposes?



Availability



Mr Cool Hitman, from Pexels



3 - Availability

Sufficiently large

For example:

- Not currently available (in the right format)
- Rare or very negative events
- Unethical situations
- Unfeasible requirements

May be preliminary steps or may be the whole research.

Presentation



Kampus Production, from Pexels

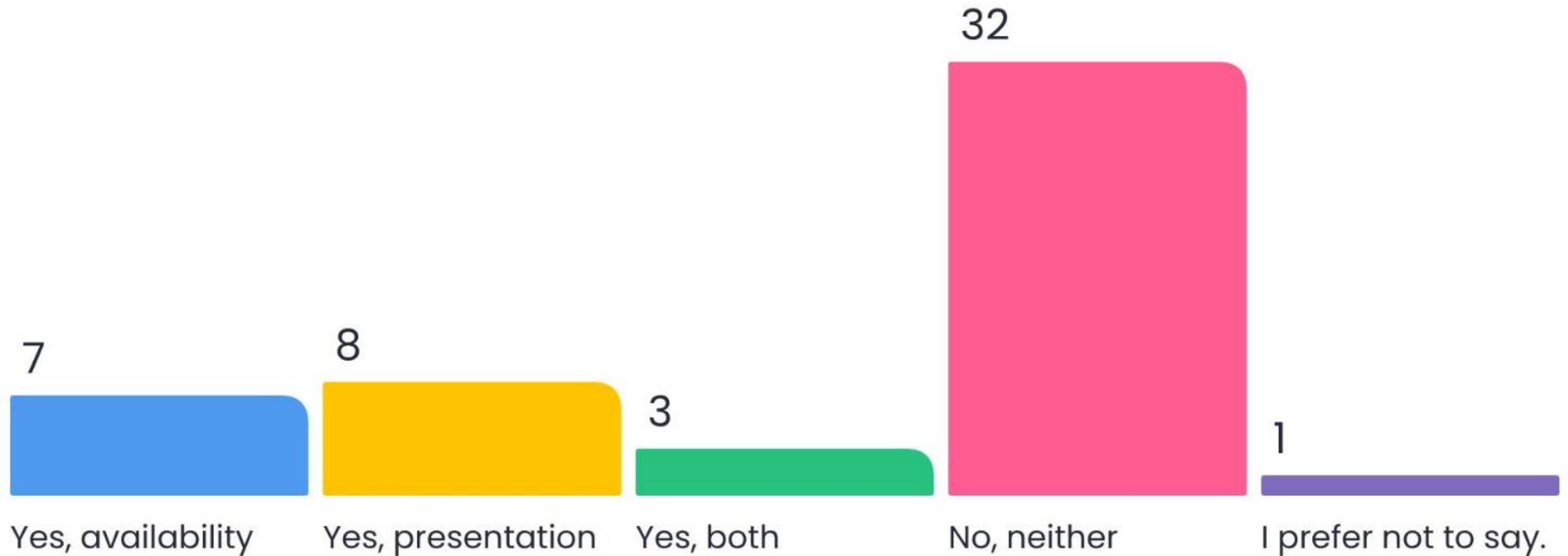


4 - Presentation

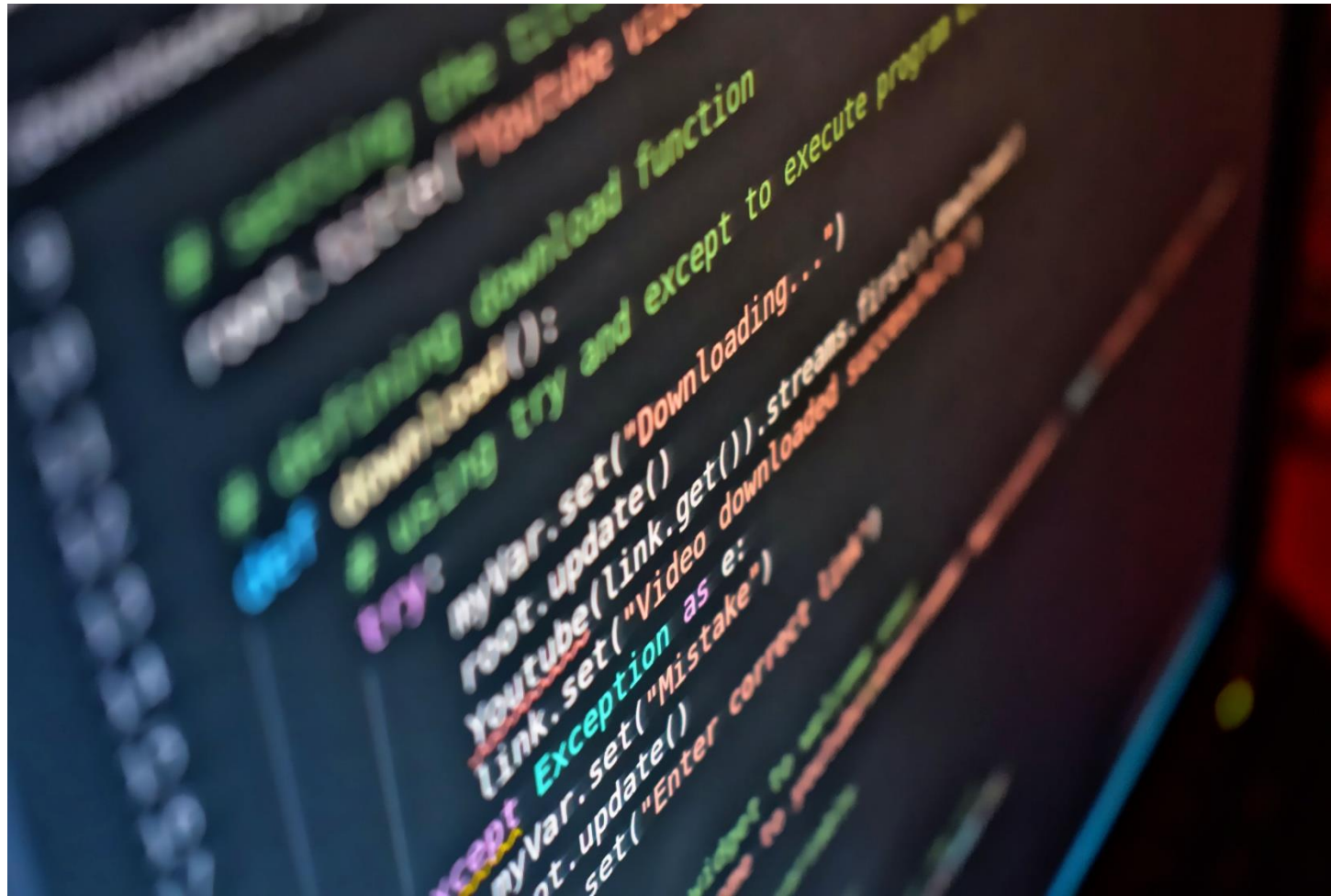
Sufficiently large

- Representative in relevant ways
- Take care if it could be mistaken as real
- Thoroughly test it

Have you ever used data for availability or presentation purposes?



Code Development



Godfrey Atima, from Pexels

5 - Code development

Sufficiently large

Deliberately unfaithful

- Test if code runs under all assumptions
- Test that code, outputs and documentation are clear and useful
- Ensure code could be run by others

Remote work



Mert Coşkun, from Pexels



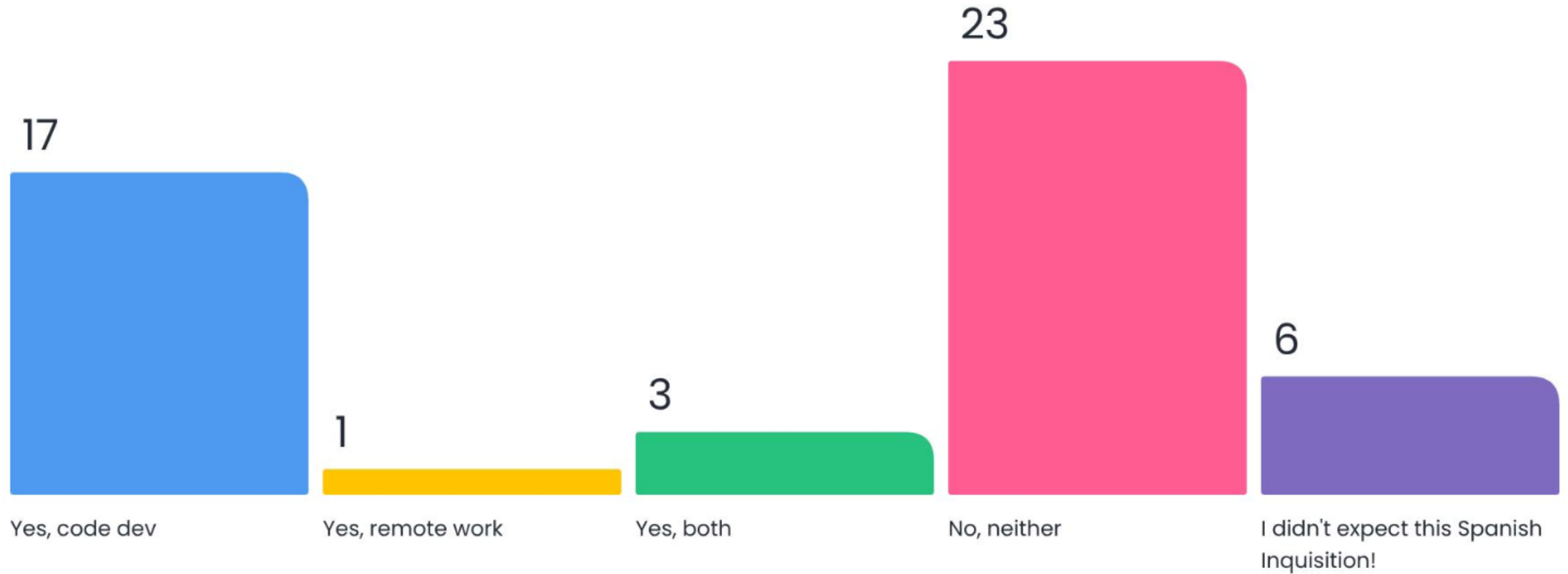
6- Remote work

Probably large and medium or high fidelity

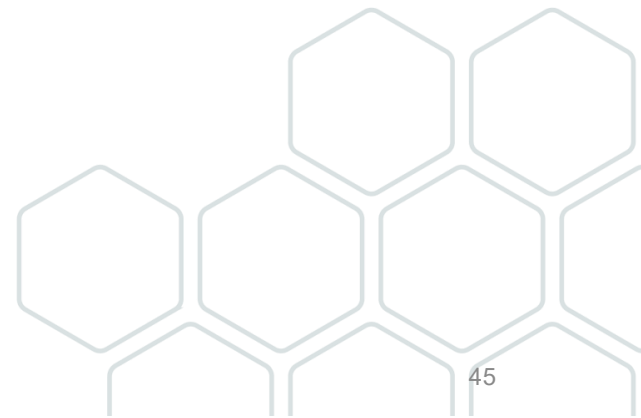
Should be:

- Portable/workable in diverse computing environments,
- As faithful as necessary for useful analysis
- Clearly useful for reproducing results
- Communicated accurately.

Have you ever used data for code dev or remote work purposes?



How to generate synthetic data



Handmade

If few examples needed, synthetic data can be “made up” from imagination

- **Obviously synthetic** or not
- **Representative** or not
- **Other**

Random/Nonsense

Output generated by random generators

- **Number(s)** - most programming packages have multiple options
- **Strings** – import some basic packages and write basic loops
- **Combination/structured** – may need to import multiple packages and write mini-programmes or scripts to generate the right combinations or structures that you need

Machine learning

Output generated by trained machine learning models

- **Supervised** methods (linear regression, decision trees, random forest, neural networks, etc.)
- **Classification** methods (logistic regression, support vector machine, naïve bayes, decision trees, random forest, neural networks, etc.)
- **Unsupervised** methods (clustering, dimensionality reduction, principal component analysis, etc.)

Simulation

Output generated by generative simulation models

- **Artificial environments** like wind tunnels, wave pools, vacuum tanks, etc. in which inputs and outputs are measured in a controlled environment meant to mimic the real-world
- **Computer simulations** in which real and/or simulated actors/forces/situations are applied within simulated environments with inputs and outputs can be measured as needed

Synthetic data conclusions

Generated, NOT anonymised.

Fidelity matters but

- Truly high fidelity is not feasible.
- Higher fidelity is not always better.

Many purposes for synthetic data.

Many ways to generate synthetic data.

Key for reproducibility.



What impediments or knowledge gaps still prevent you from using synthetic data? 1

Coding skills

lack of experience

get to know the data type

How to generate in the case of lack of data

How to generate it

Haven't considered it yet.

No experience with creating custom datasets using code

Good R package to generate data

Time

Tough to wrap my head around how I would use it in practice

How to generate it

Essentially none

I have access to plenty of real data

Skills

How to generate it in the specific context

understanding how to develop

Just now considering

Don't know how to generate appropriate data for my field

Hardware for generating

Not sure if I could trust

What impediments or knowledge gaps still prevent you from using synthetic data? 2

Limitations for nested lists/dictionary.

Knowing exactly how to specify the data and how to use the software

Synthetic data that makes sense for my usecase very complex to create

Lots of variables to consider and risk of over fitting /simulating

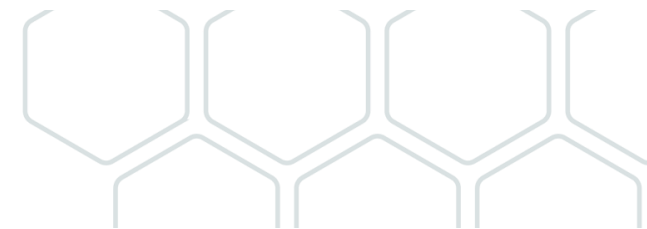
lack of advanced coding skills for simulation or some machine learning representativeness?

How it works with research questions

I don't know enough about it yet.

how to assess risk in generated data

I have not found good models for good fidelity levels, I have made up data not sure if i used the models properly



Further reading/listening

Data in Government Blog tinyurl.com/Synth-DataInGov

The unreasonable effectiveness of synthetic data with Daeil Kim
tinyurl.com/Synth-Podcast

Former UKDS team Medium article tinyurl.com/Synth-Blogpost

Synthetic data estimation for the UK longitudinal studies
<https://tinyurl.com/5hc96ukr>

Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation <https://tinyurl.com/2sdjz8zv>

Thank you.

Dr. Jools Kasmire

Email j.kasmire@manchester.ac.uk