

Synthetic Data: An introductory workshop

Dr. J. Kasmire
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

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae

Examples of synthetic data 2

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae



Examples of synthetic data 3

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident sunt in culpa qui officia deserunt mollit anim id est laborum sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae



8

Min 1

Max 10

GENERATE

▼

Examples of synthetic data 4

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae



8

Min 1

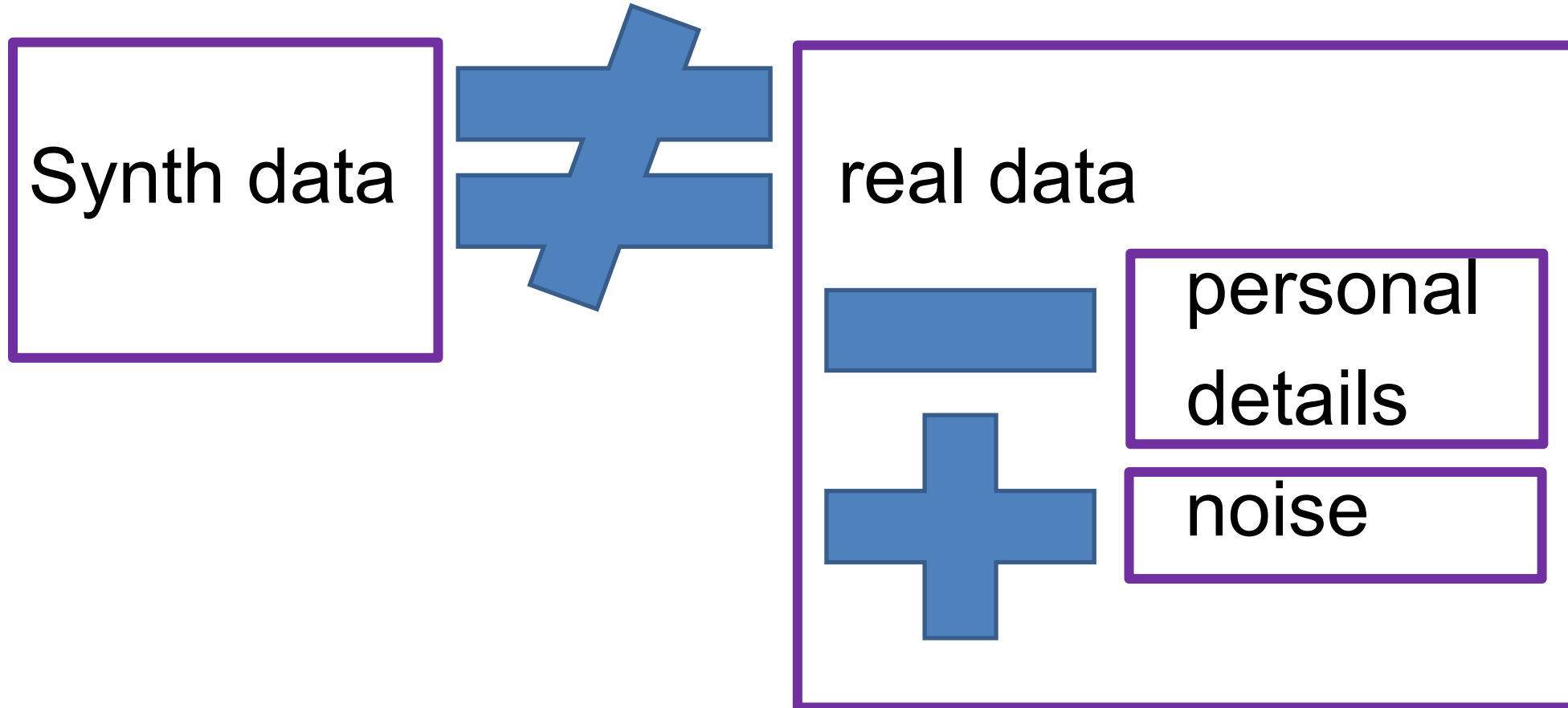
Max 10

GENERATE

▼



What is not synthetic data?



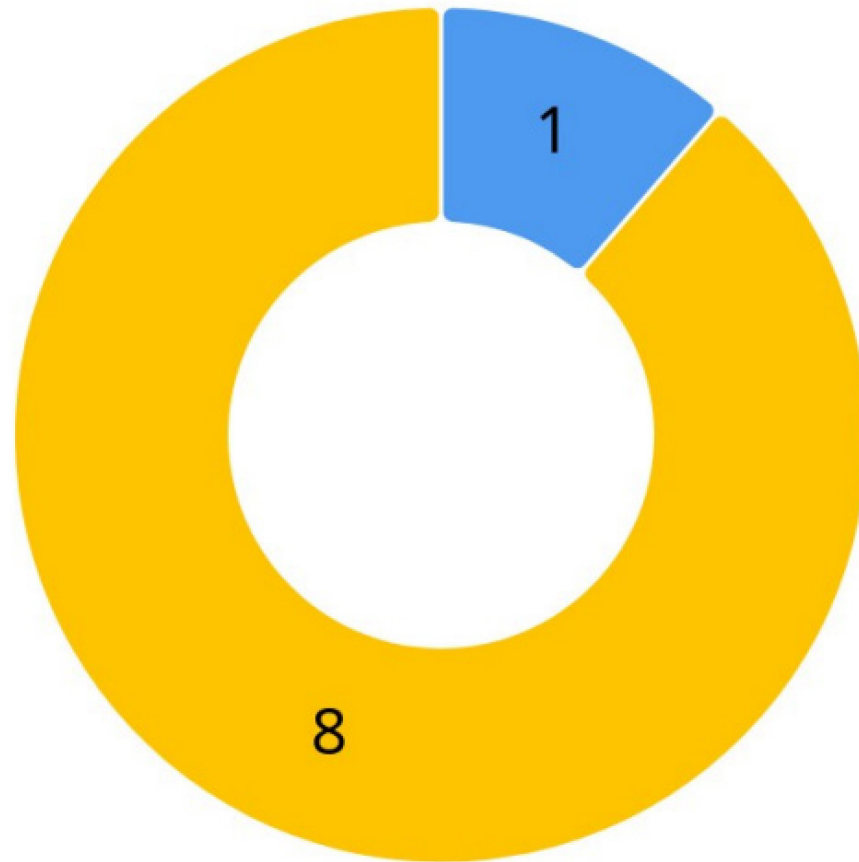
Terminology is important

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

They are wrong.



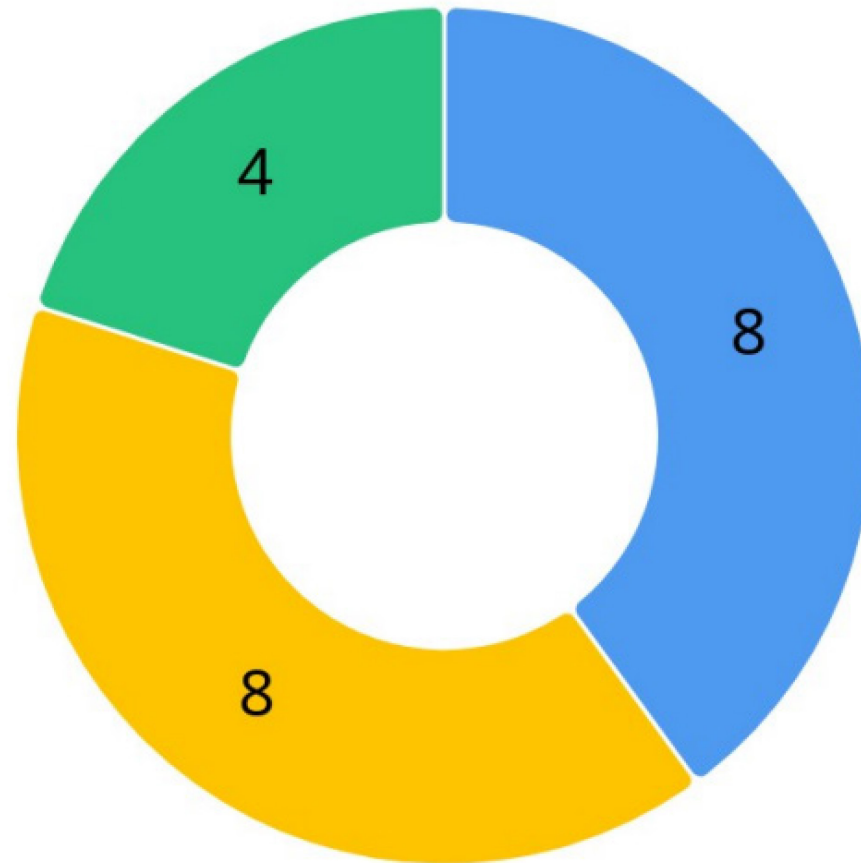
Is data from a cycle lane sensor synthetic?



- 1 Synthetic
- 8 Not synthetic
- 0 I have doubts



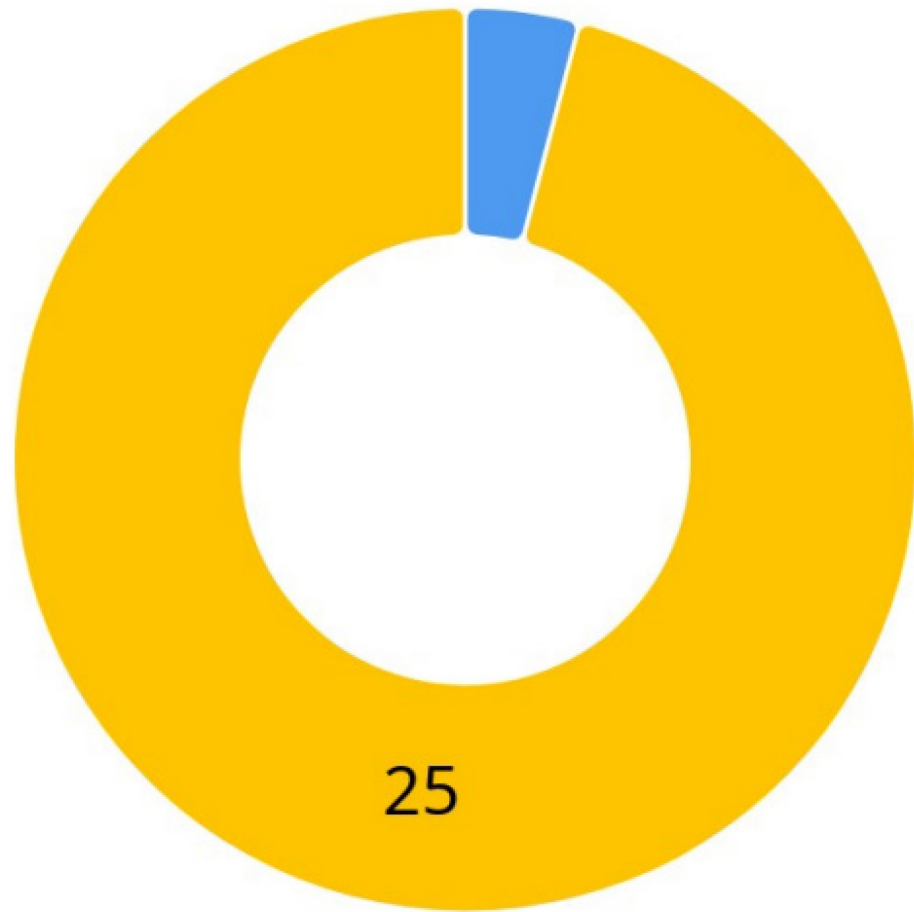
Are the predictions from weather forecasts synthetic data?



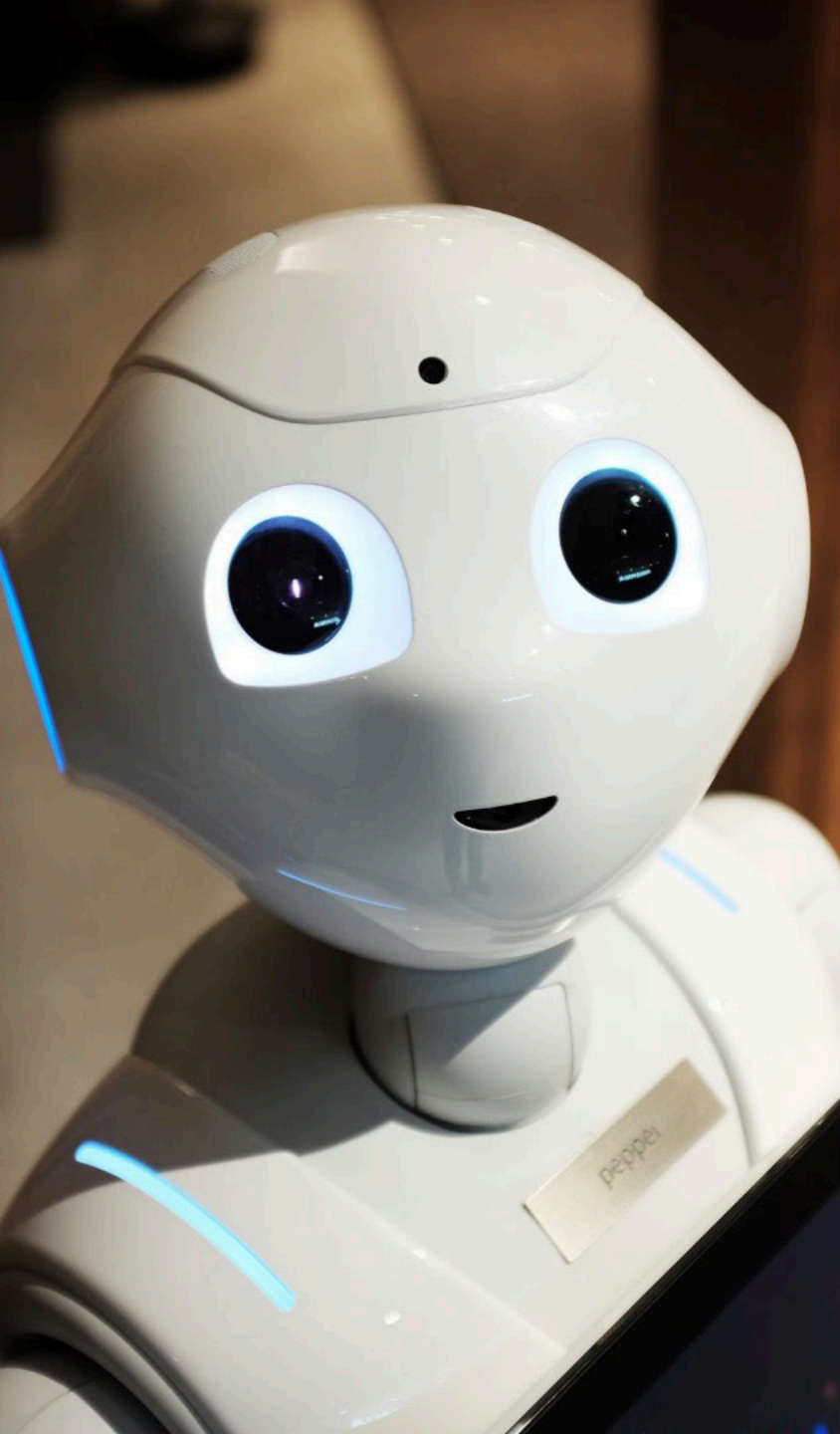
- 8 Synthetic
- 8 Not Synthetic
- 4 I... uh... not sure.



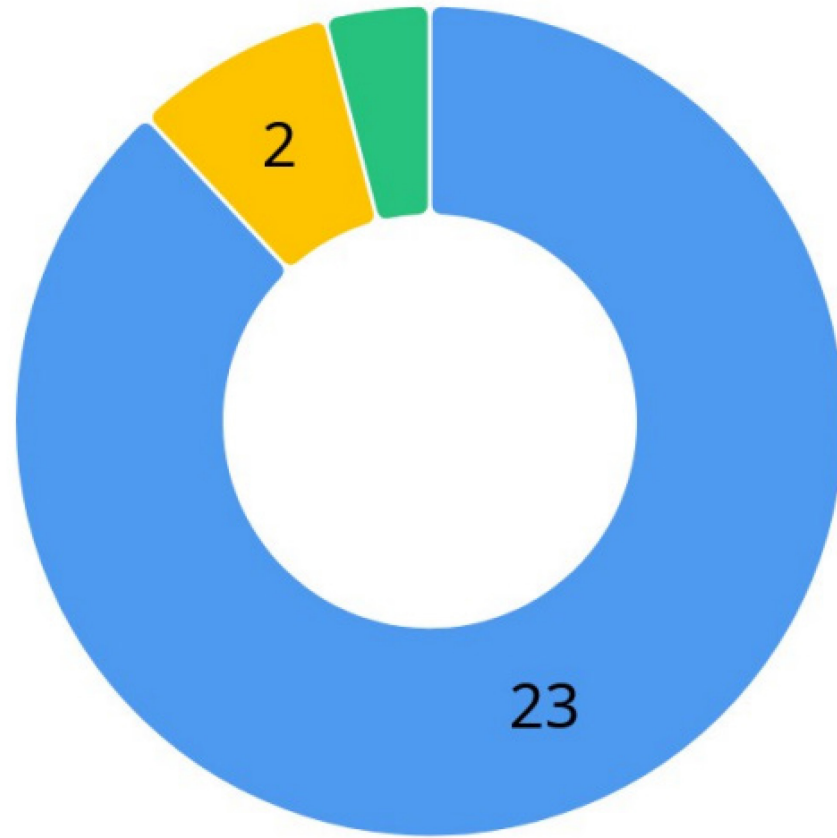
Is census microdata synthetic?



- 1 Synthetic
- 25 Not synthetic
- 0 Can you rephrase the question?



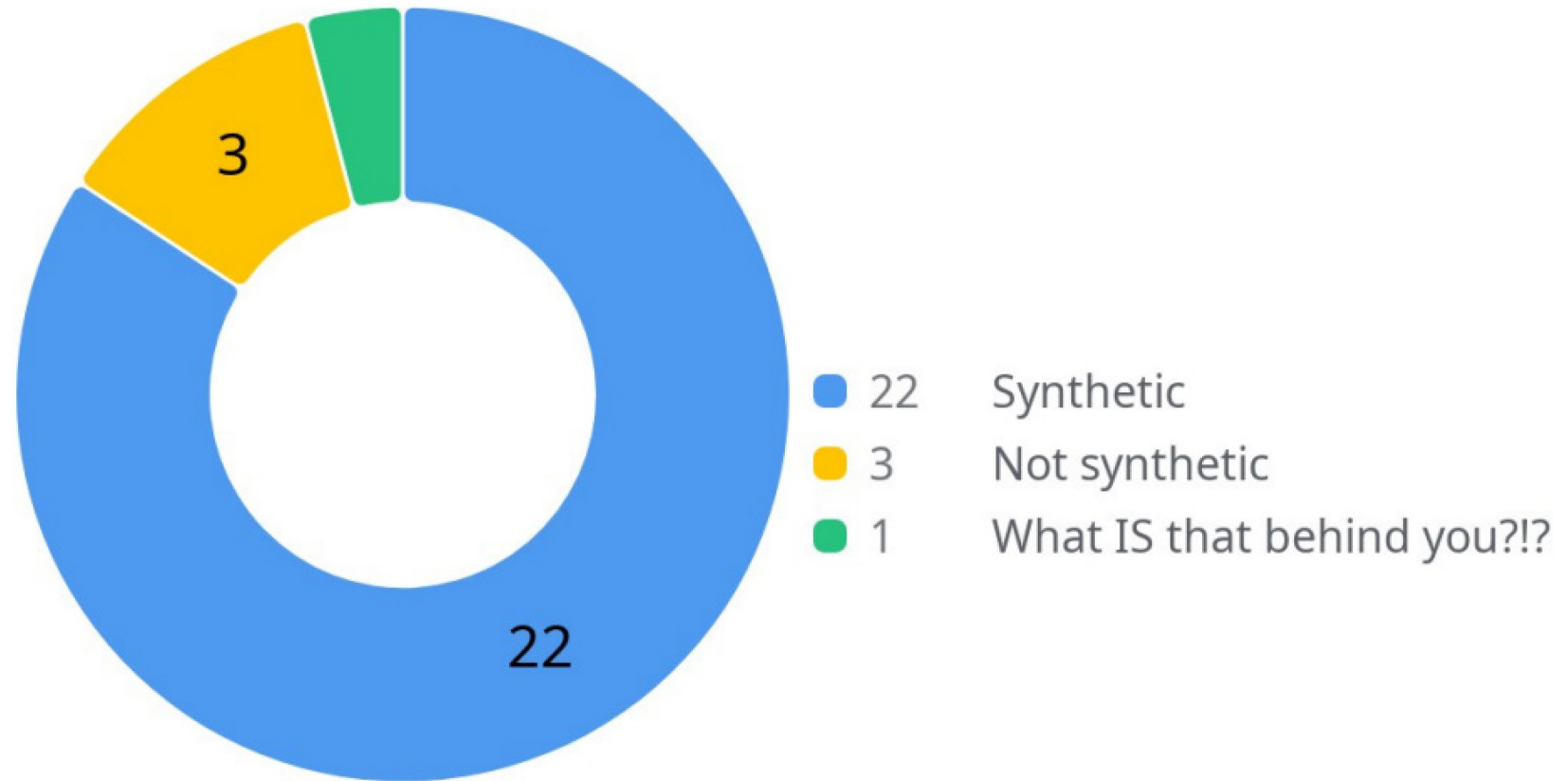
Is the output from Chat GPT synthetic?



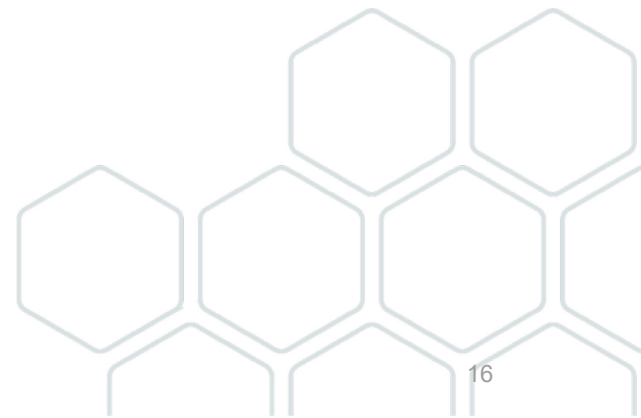
- 23 Synthetic
- 2 Not synthetic
- 1 Can I get a glass of water?



Is stuff you just made up in your head synthetic?



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

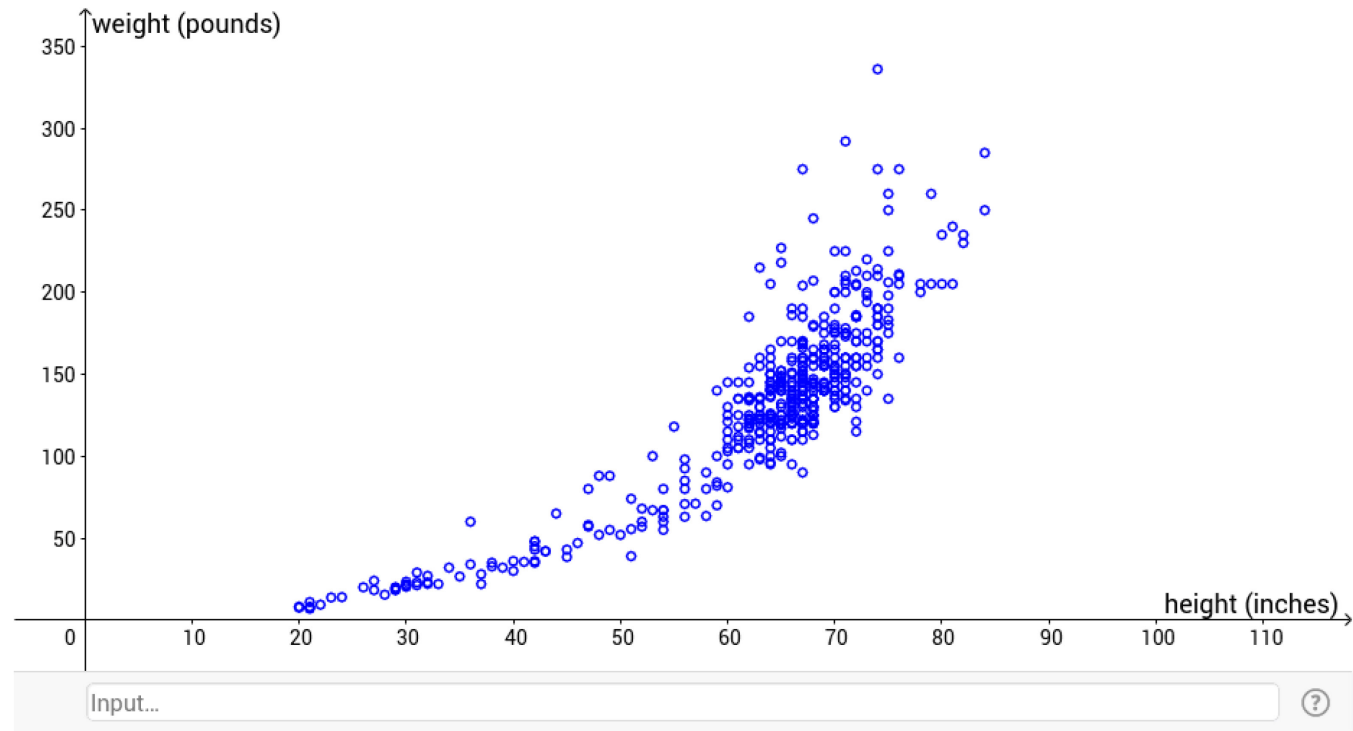
Real

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

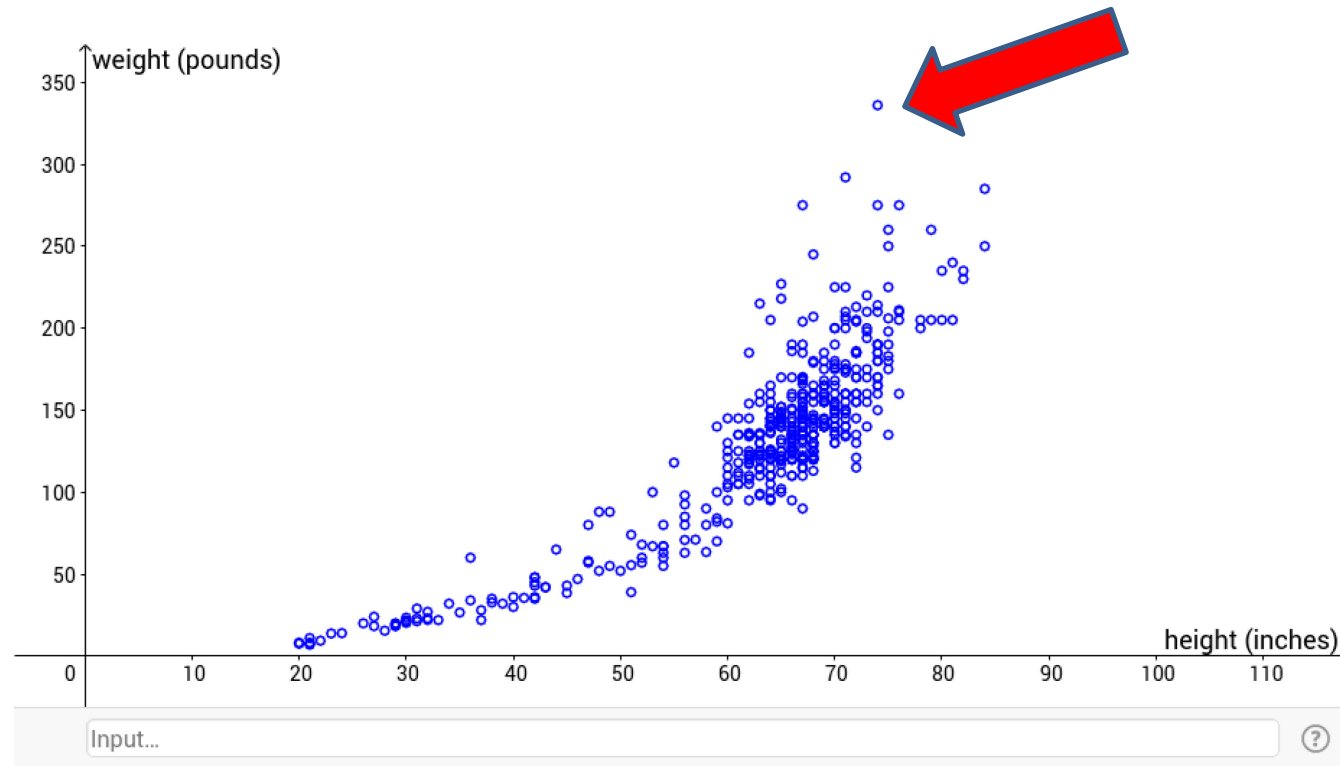
Synthetic

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



Relationships between variables and outliers



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

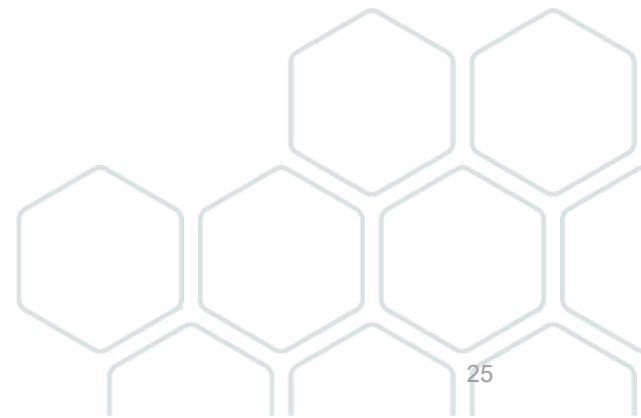
Some synthetic data is more useful if it is NOT faithful to the original in some ways.

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

Data Import Wizard - Step 1 of 5

Data Preview

Data Preview

Header After Skip Skip Rows: 0

Format: csv Preview Row Limit: 100

Encoding: Default

Delimiter: , Line Terminator: environment default

Left Enclosure: none Right Enclosure: none

COLUMN1	COLUMN2	COLUMN3	COLUMN4	COLUMN5
1	Office Equip...	25-MAY-2012	1201	13-JUN-2012
2	Computer S...	18-JUN-2012	1201	27-JUN-2012
3	Travel Expe...	26-JUN-2012	1340	11-JUL-2012

Help < Back Next > Finish Cancel



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



TOY DATASETS

Datasets



2 - Proof of concept

Sufficiently large

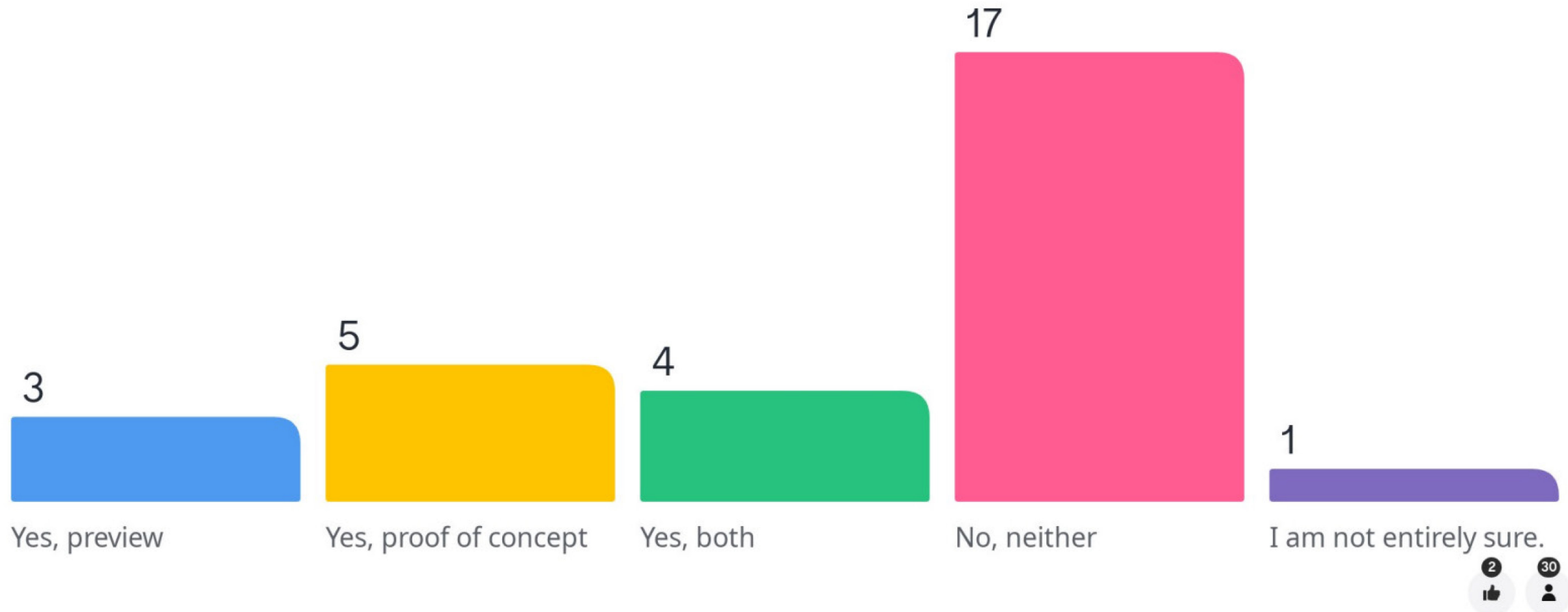
usefully demonstrates that:

- 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



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

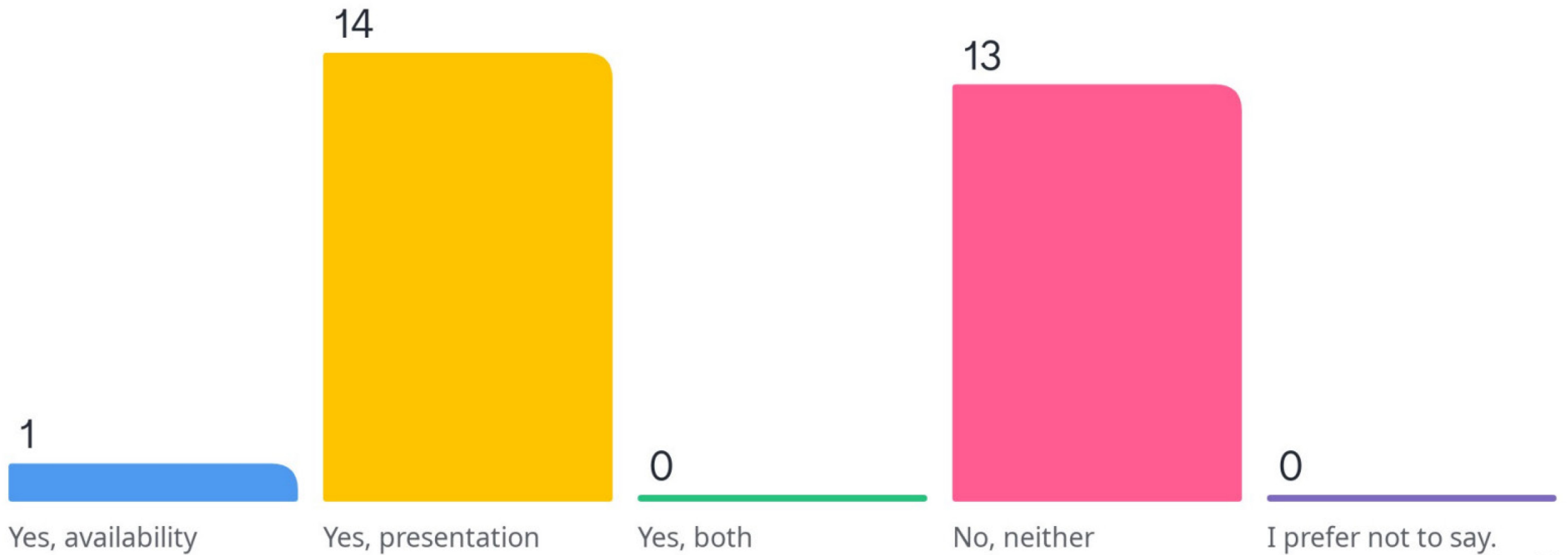


4 - Presentation

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

```
        'replace_interests' => false,  
        'send_welcome'     => false,  
    });  
  
    if (array('error', $result)) {  
        $arResult = array ('response'=>'error', 'message'  
        * (  
        $arResult = array ('response'=>'success');  
    }  
  
    $arResult($arResult);
```



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



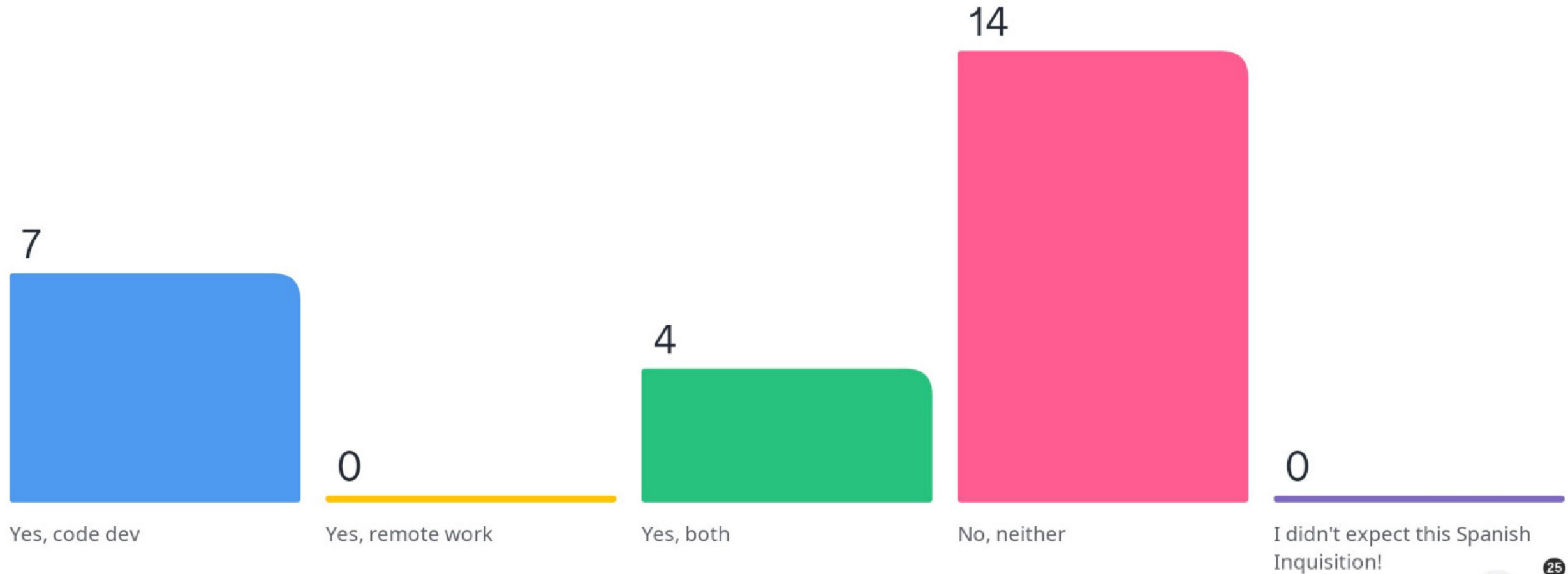
6- Remote work

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.



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. J. Kasmire

Email j.kasmire@manchester.ac.uk

 @JKasmireComplex on Twitter

@UKDataService on Twitter