Balancing the data scales: A cost-benefit analysis of low-fidelity synthetic data for data owners and providers

17 September 2024

Cristina Magder, Data Collections Development Manager

UK Data Service

Synthetic data…

- … “are artificially generated data that are made to resemble real-world, often sensitive, data.” ONS
- … “are microdata records created to improve data utility while preventing disclosure of confidential respondent information.” US Census Bureau
- … “is computer-generated information designed to improve AI models, protect sensitive data, and mitigate bias.” IBM Research
- ... “is data that has been generated using a purpose built mathematical model or algorithm, with the aim of solving a (set of) data science task(s).” Royal Society

Project context

Evaluate the cost-benefit dynamics of synthetic data for data owners and Trusted Research Environments (TREs).

Mixed method approach.

Principal Investigator: Cristina Magder

Co-Investigators: Maureen Haaker, Jools Kasmire, Hina Zahid

Researcher: Melissa Ogwayo

8 April 2024 – 31 March 2025

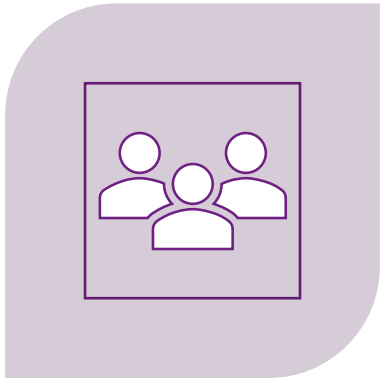
Data, metadata and documentation

Synthetic data created from real data

“Data free” synthetic data

Project objectives

EXPLORE EFFICIENCY GAINS



ASSESS DATA SHARING STRATEGIES

EVALUATE THE COST SPECTRUM

Work packages

Literature review

Survey with data owners

Case studies with providers of synthetic data

Focus group with TRE representatives

Literature review highlights: ethical and legal dimensions

No established legal or ethical frameworks to regulate its use.

Several ethical considerations emerged from the findings of the review:

- informed consent
- data quality and bias
- transparency
- accountability
- confidentiality, privacy and disclosure

No clear legislation surrounding the use of synthetic data

Some key considerations include:

- Generation and processing of synthetic datasets should be treated separately.
- Generative models using personal data for synthetic data are subject to UK GDPR/GDPR.
- Synthetic datasets, containing only artificial attributes, are not subject to UK GDPR/GDPR.

Literature review highlights: usage and current gaps

Synthetic data is used across industries for various applications, including:

- Machine learning and AI training.
- Enhancing datasets for better model performance.
- Creating balanced datasets to reduce biases.
- Protecting privacy by replacing real sensitive data.
- Testing and development.
- Healthcare, financial services, and education.

Gaps in the current knowledge around synthetic data generation

- No standardised methods for generating synthetic data, which results in a lack of common evaluation metrics to assess its quality, utility, and re-identification risks.
- Lack of established legal and ethical frameworks to govern the utilisation of synthetic data.
- Lack of established benchmarks or standardised methods for validating synthetic data to ensure the quality and reliability.

Survey background

Assess the readiness and perceptions of data creators toward synthetic data production and dissemination.

Main objectives:

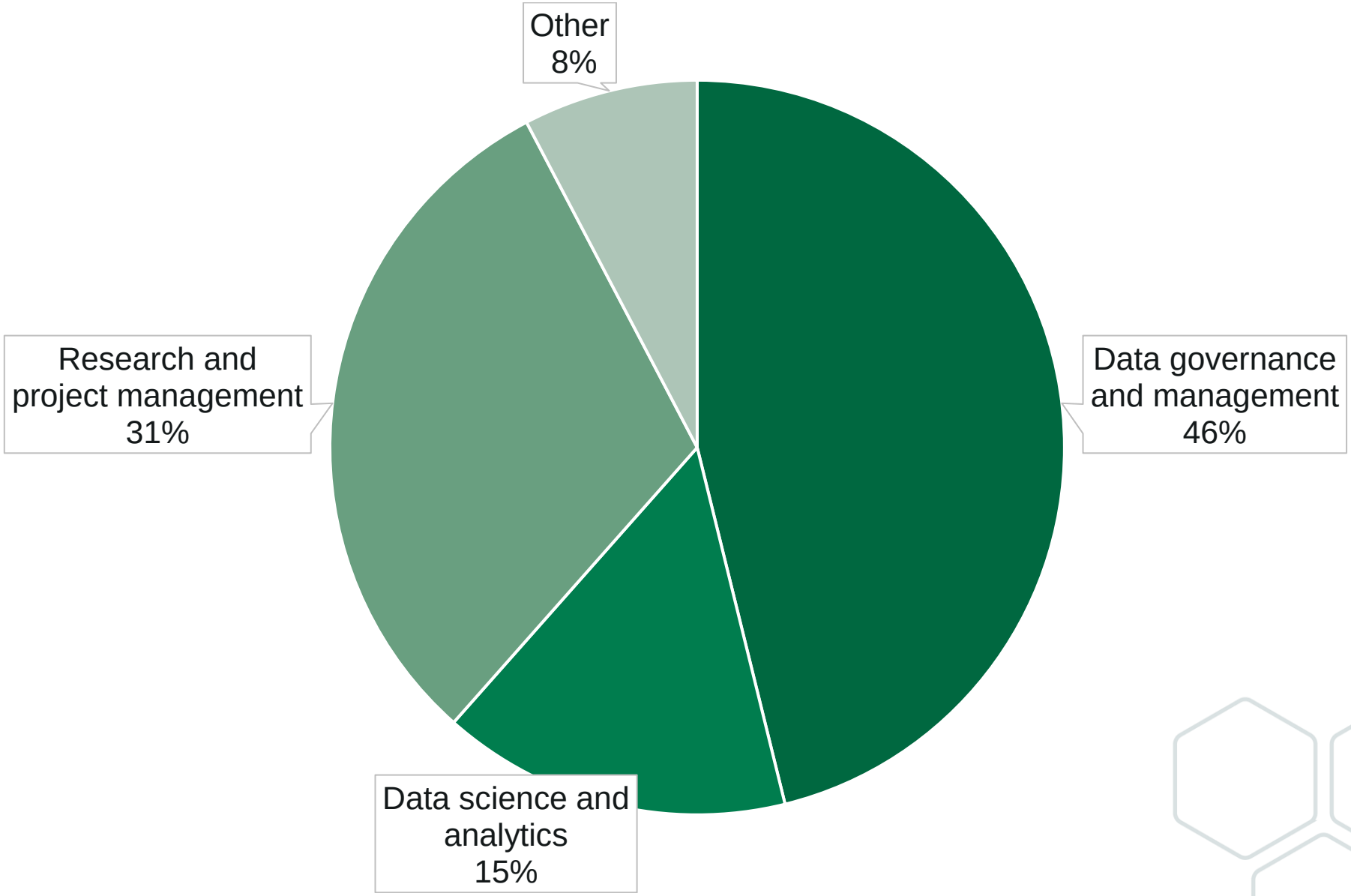
- Understand current synthetic data production and sharing practices.
- Identify challenges (technical, operational, financial) faced by data creators.
- Explore the benefits, future trends, and support needs in the data producer community.

Open for data owners, producers, and management teams across sectors like government, higher education, healthcare, and private enterprises.

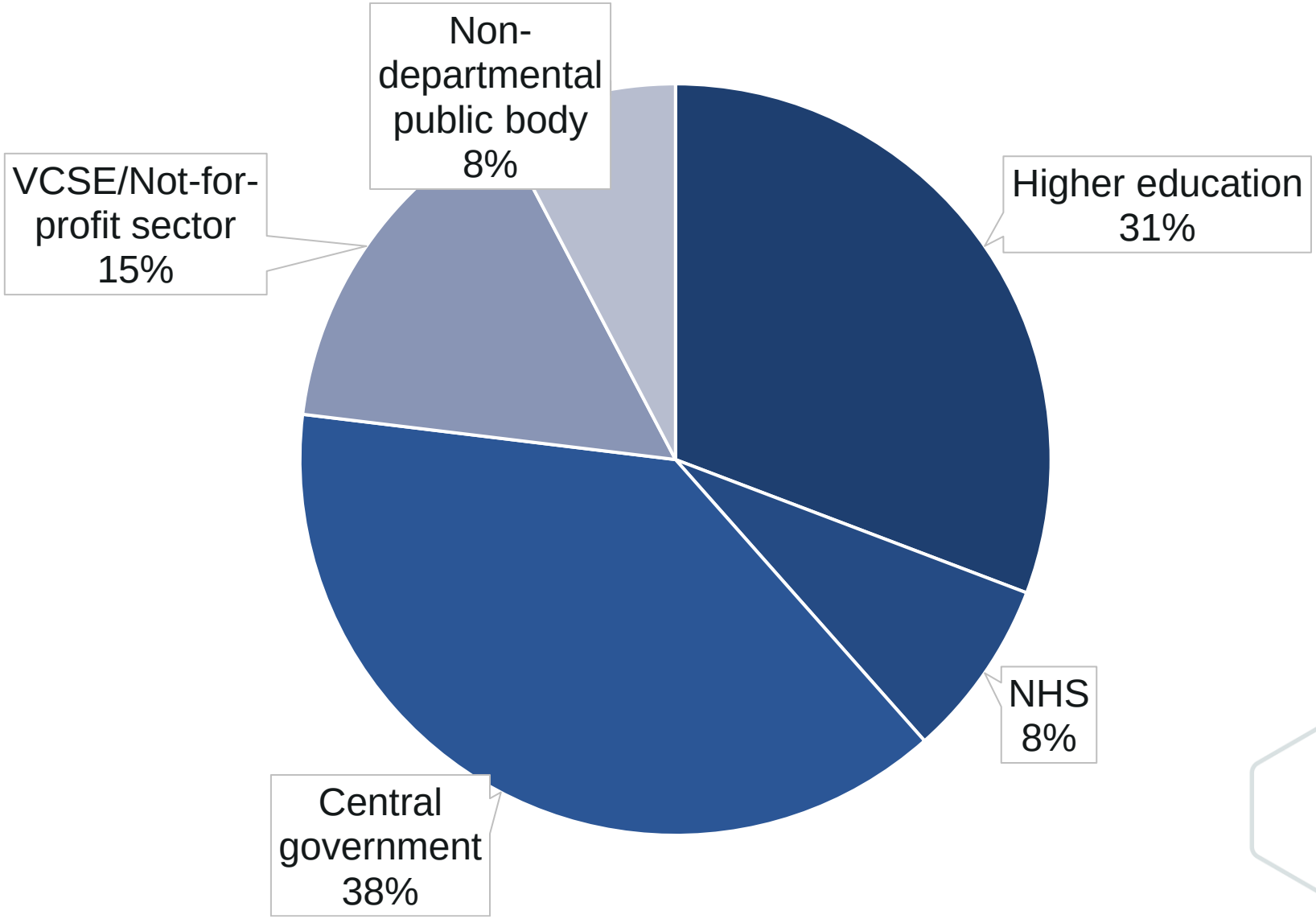
UK Data Service

Preliminary survey findings

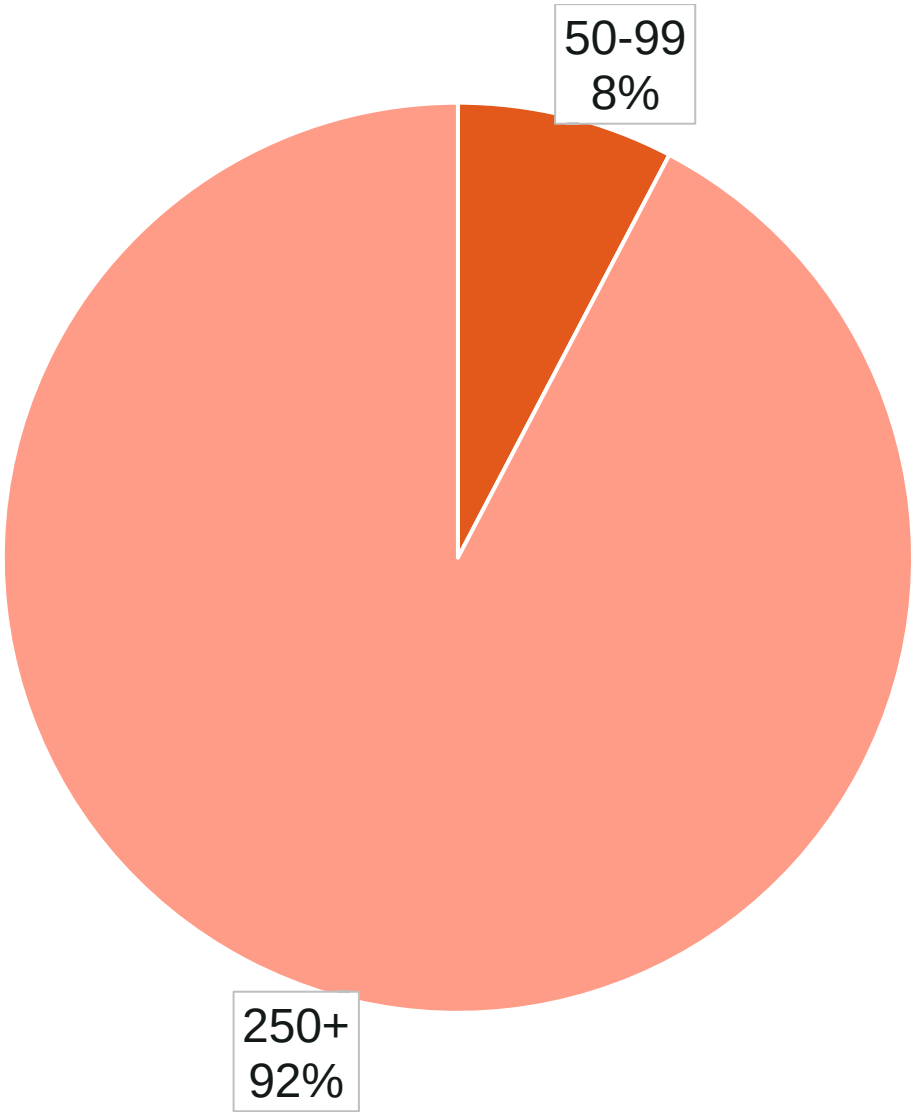
Survey results to date: respondents' role



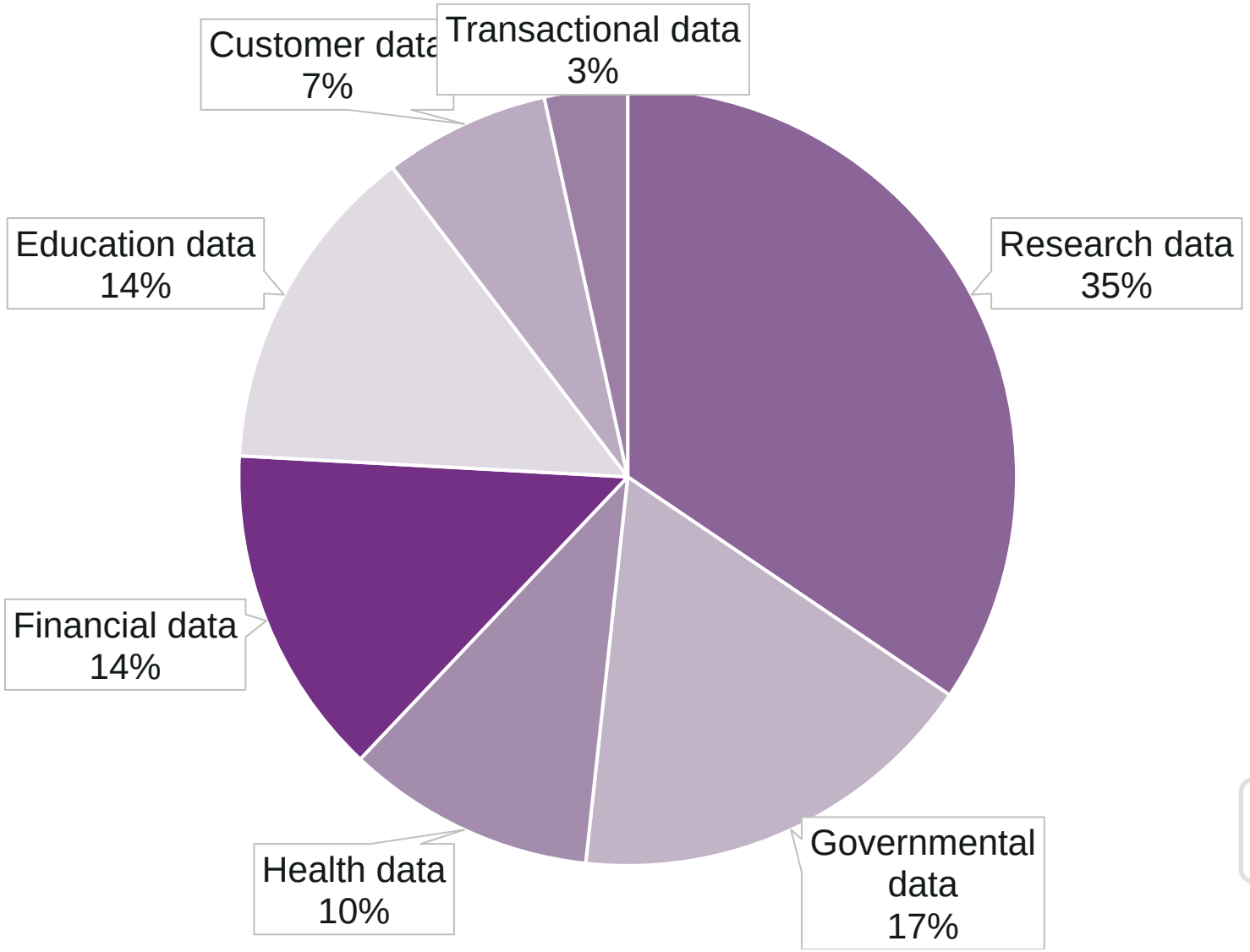
Survey results to date: sector



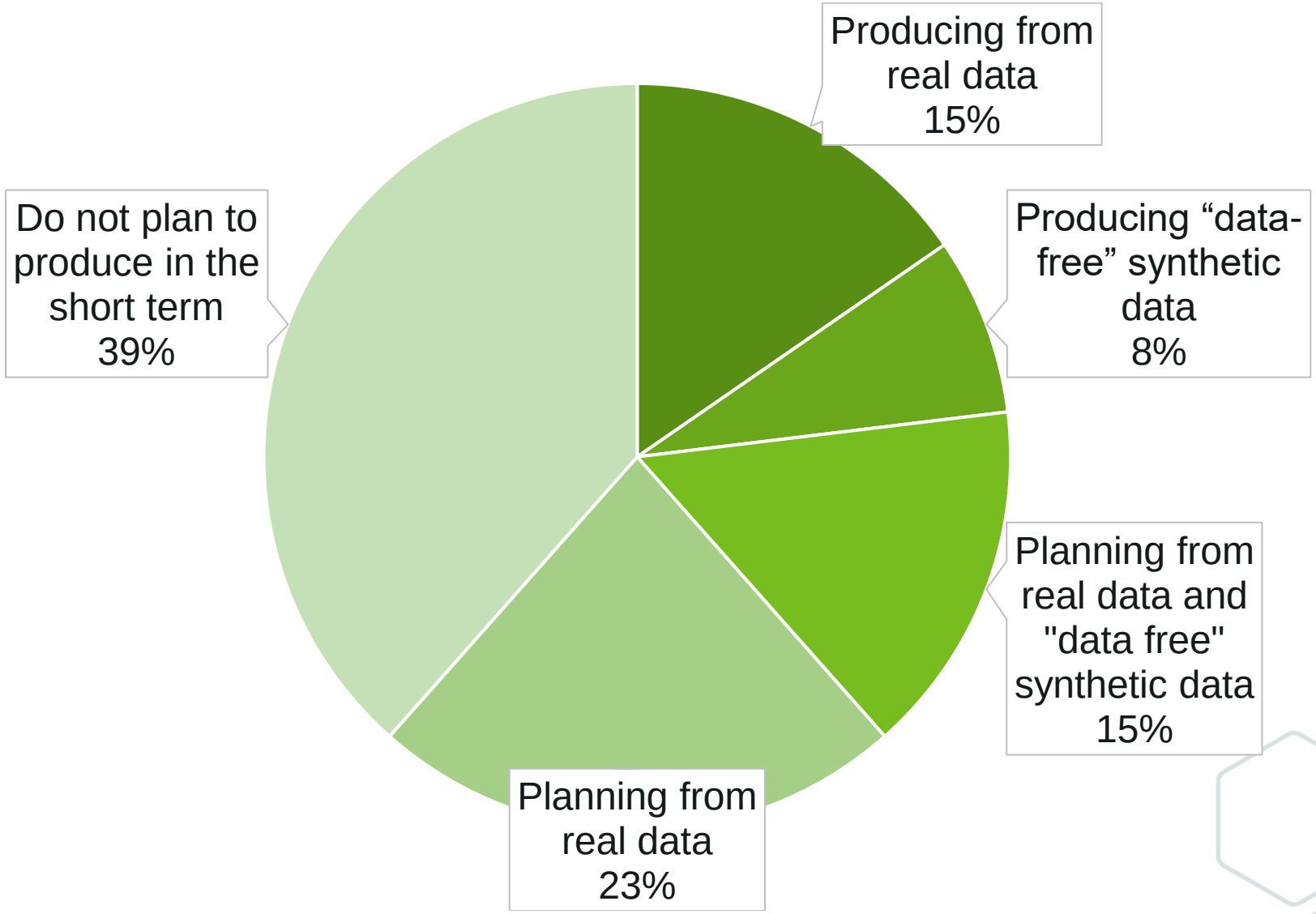
Survey results to date: organisation size



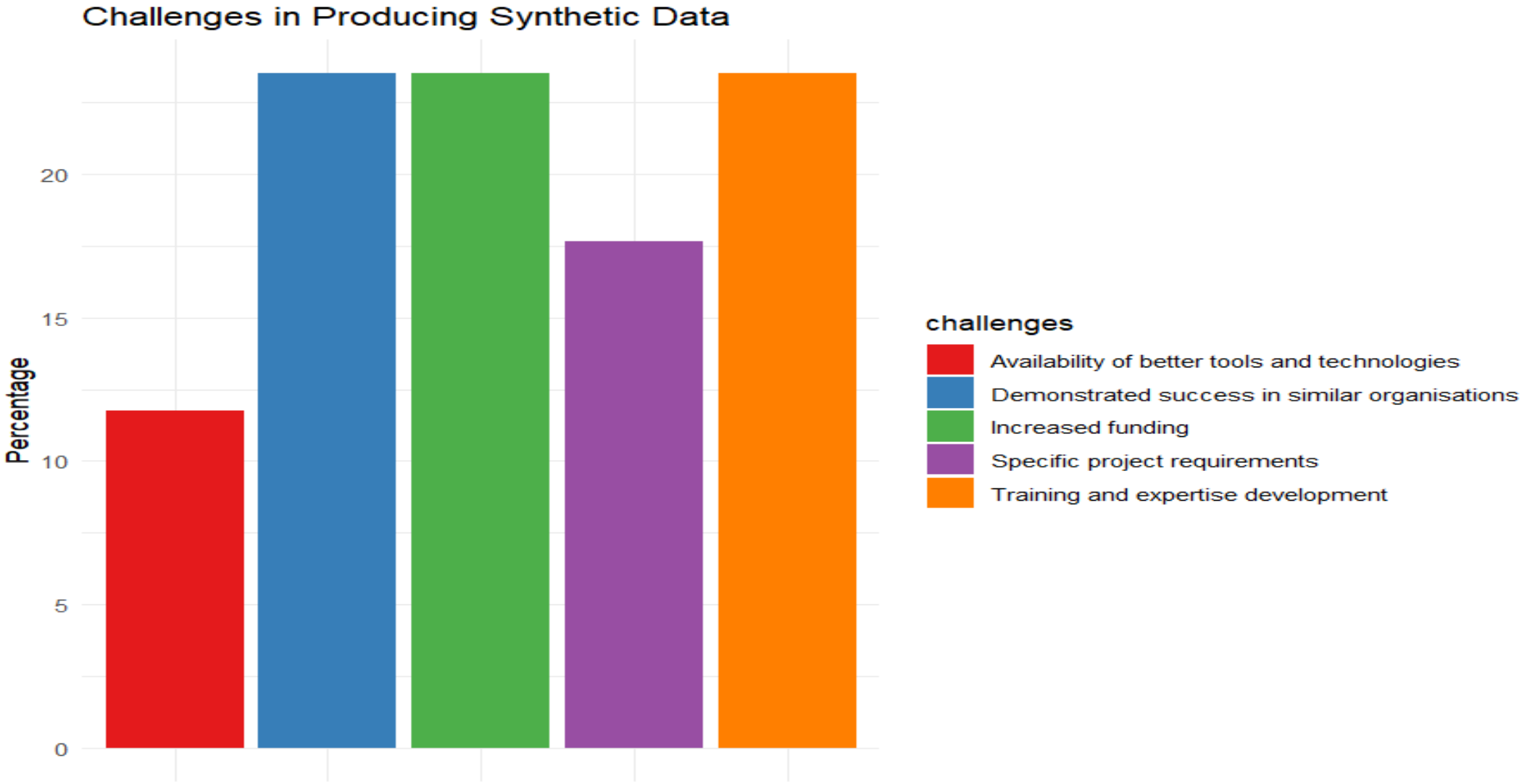
Survey results to date: real data production



Survey results to date: synthetic data production



Survey results to date : Challenges in Producing Synthetic Data



Survey results to date: More detail on the Challenges in Producing Synthetic Data

"Ensuring and explaining the demonstration of GDPR compliance is a technical challenge in this context. Additionally, dissemination and technical expertise, while listed, are not found within the team but rather within stakeholders, authorities, and collaborators"

This was stated by a person in data governance and management within the central government sector, which produces research data, governmental data, financial data, educational data, and customer data, when asked about the technical challenges the organization has faced.

"Initial investment costs, ongoing operational costs and cost of computational resources."

This was mentioned by a person in data governance and management in the NHS, working with health data, as further detail on financial costs.

"Lack of clear, authoritative evidence on the privacy risks of high-fidelity data, and of the utility of any analysis conducted on it. Unless convinced that valid analysis can be conducted on synthetic data with no privacy risk, we prefer to leave it to TREs holding our data to explore the possibilities of synthetic data."

This was shared by an analytical external engagement lead at a central government organization, which deals with governmental and customer data. They explained the reasons and challenges that keep them in the planning or assessing phase of synthetic data production and not yet generating it.

Survey results to date: Data owners on benefits of producing synthetic data

"Enhanced privacy, easier compliance with regulations."

A data governance and management specialist in the NHS, working with health data

"Enhanced privacy, better provision for analysts working with large, complex datasets (e.g., they don't have to invest in and wait for access to understand datasets readily and what can be done with them—it's incomparably better than just reading a user guide or metadata). For the same reasons, better data access requests and fewer failed projects/applications."

A data governance and management specialist in the central government sector that produces research data, governmental data, financial data, educational data, and customer data,

"Increased data sharing, accelerated data access."

A data science and analytics professional in the central government sector dealing with governmental data and transactional data.

Circulate to your data owners

For circulation to **data owners and producers.**

Approximately 25 minutes to complete.

The survey is anonymous, and no personally identifiable information is collected.

Please complete by no later than **29 September 2024**.

Introduction to synthetic data event

18 Nov 2024 10:00 am - 12:30 pm GMT via Zoom

The workshop will introduce the fundamentals of synthetic data, with a particular focus on its relevance to TREs and data owners.

Participants will have access to all workshop materials, including slide decks and Jupyter notebooks, via a GitHub repository.

No formal prerequisites are required to attend. However, those who wish to actively participate in the coding demonstration should have:
• Access to a computer with Python installed, or an online Python environment.
• Basic Python knowledge (e.g., loading packages, handling data, etc.).

Read more and book your place

Focus group with TRE representatives

Explore the operational dimensions of synthetic data usage within secure environments to understand:

- Practical implications
- Challenges
- Opportunities

Focus group with TREs

We will be conducting an online focus group with TRE representatives on the **20th of November** at 10:00am to 12:30pm GMT (UTC +0).

Feel free to scan the QR code to register for the focus group !

UK Data Service

Thank you!

dcmagd@essex.ac.uk
datasharing@ukdataservice.ac.uk
beta.ukdataservice.ac.uk/help