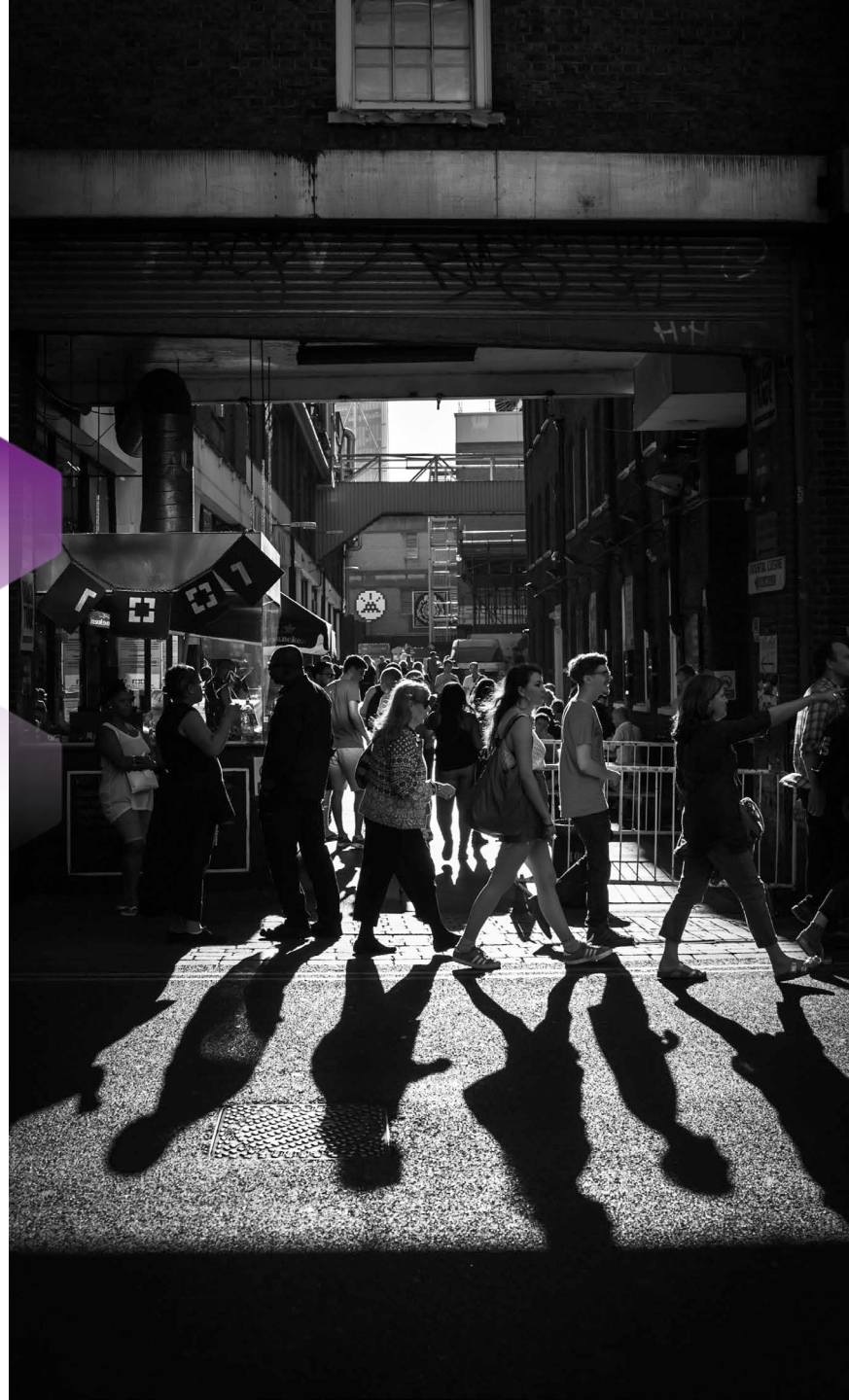
UK Data Service

Introduction to synthetic data for Trusted Research Environments data experts



18 November 2024

Cristina Magder, Data Collections Development Manager

UK Data Service

UKRI Economic and Social Research Council

Today's programme

- Introduction to Balancing the Data Scales Project.
- Introduction to Synthetic Data:
  - What is synthetic data?
  - What types of synthetic data are there?
  - For what purposes can synthetic data be used?
  - How can synthetic data be generated?
- Q&A.
- Break (10 mins).
- Live Coding Demonstration.
- Q&A.

UK Data Service (UKDS)

Hosts UK's largest collection of social, economic and population research data.

Provide users with access, support, guidance and training to facilitate high quality social and economic research and education.

A partnership between UK Data Archive at University of Essex, Cathie Marsh Institute for Social Research at University of Manchester, Jisc, UK, EDINA from University of Edinburgh, and the Department of Information Studies and Centre for Advanced Spatial Analysis at University College London.

Support the development of best practices for data preservation and sharing standards.

ukdataservice.ac.uk

Stats about UKDS

~ 9,700 data collections.

~ 250 new data collections and new editions added each year.

~ 48,000 registered users.

~ 130,000 data accesses worldwide p.a.

UKDS data collections are accessed every 6 minutes

24 x 7 x 365

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.

Why synthetic data?

SIPHER case study: synthetic population

SIPHER Consortium. (2023, June 15). *What is a synthetic population? SIPHER's synthetic data explained* [Video]. YouTube. Visit the SIPHER website.

SIPHER case study

The SIPHER Consortium is addressing health inequalities by linking social determinants (e.g., income, housing, education) with health outcomes.

Using spatial microsimulation, the SIPHER project created a synthetic population, reflecting the adult population in Great Briain based on Understanding Society and UK Census data. It offers a "digital twin" of these populations, supporting innovative research while protecting privacy.

- Enables simulation models to assess the impact of policies on different population subgroups.
- Helps policymakers evaluate public health interventions and outcomes across regions.

Data collection available through the UKDS **SIPHER Synthetic Population for Individuals in Great Britain, 2019-2021 (SN9277)** alongside the replication package.

Read the full case study and explore the interactive dashboard.

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

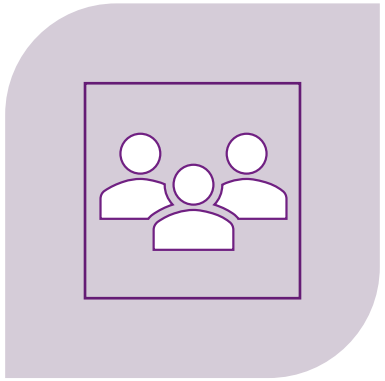
Data, metadata and documentation

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 dimension

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.

Literature review highlights: Legal dimension

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

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.

Literature review highlights: Current gaps

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 with data owners: Background

Main objectives:

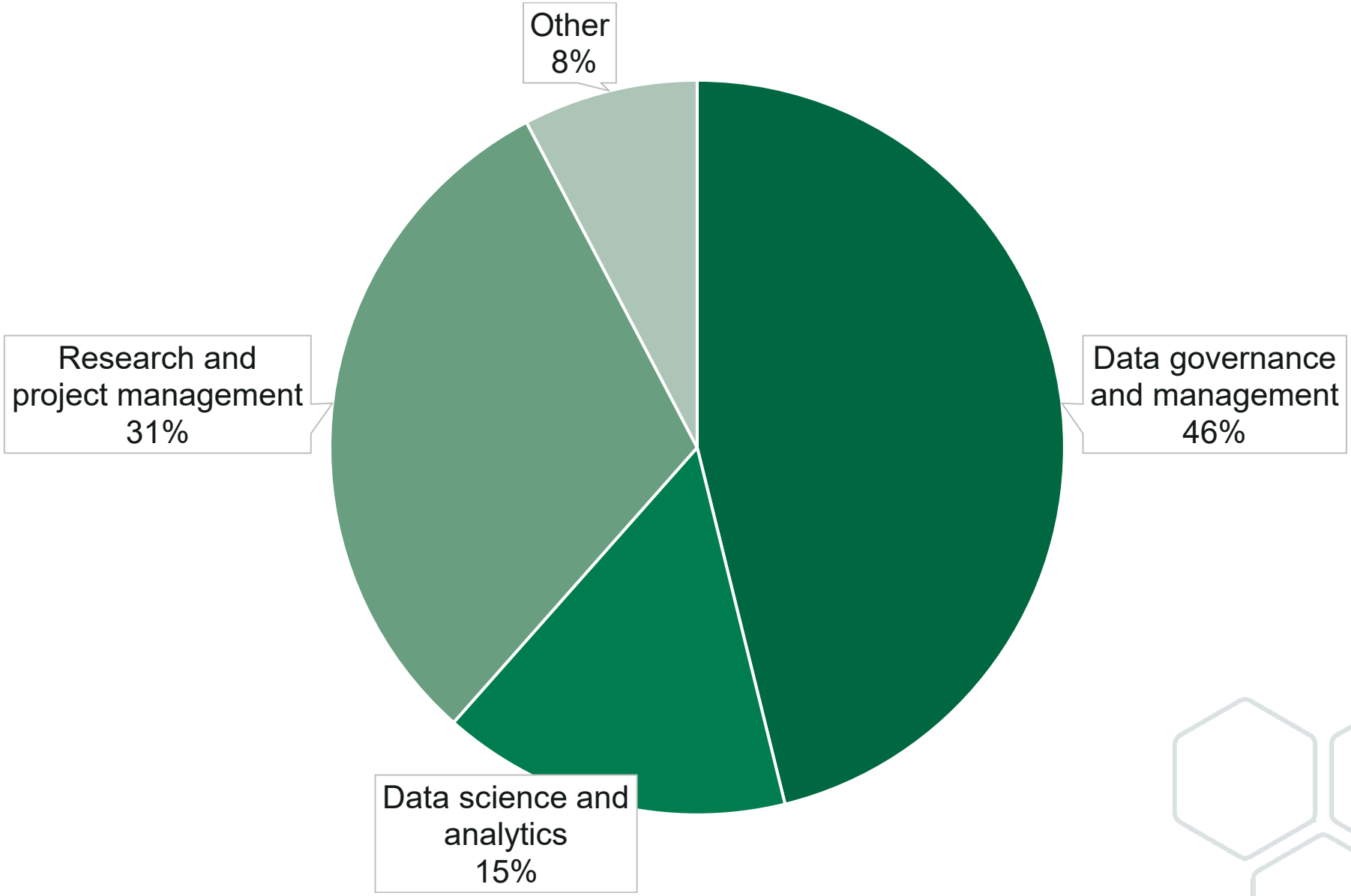
- Understand current synthetic data production and sharing practices.
- Identify challenges (technical, operational, financial) faced by data creators.
- Explore the benefits, future trends, and support needs in the data producer community.

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

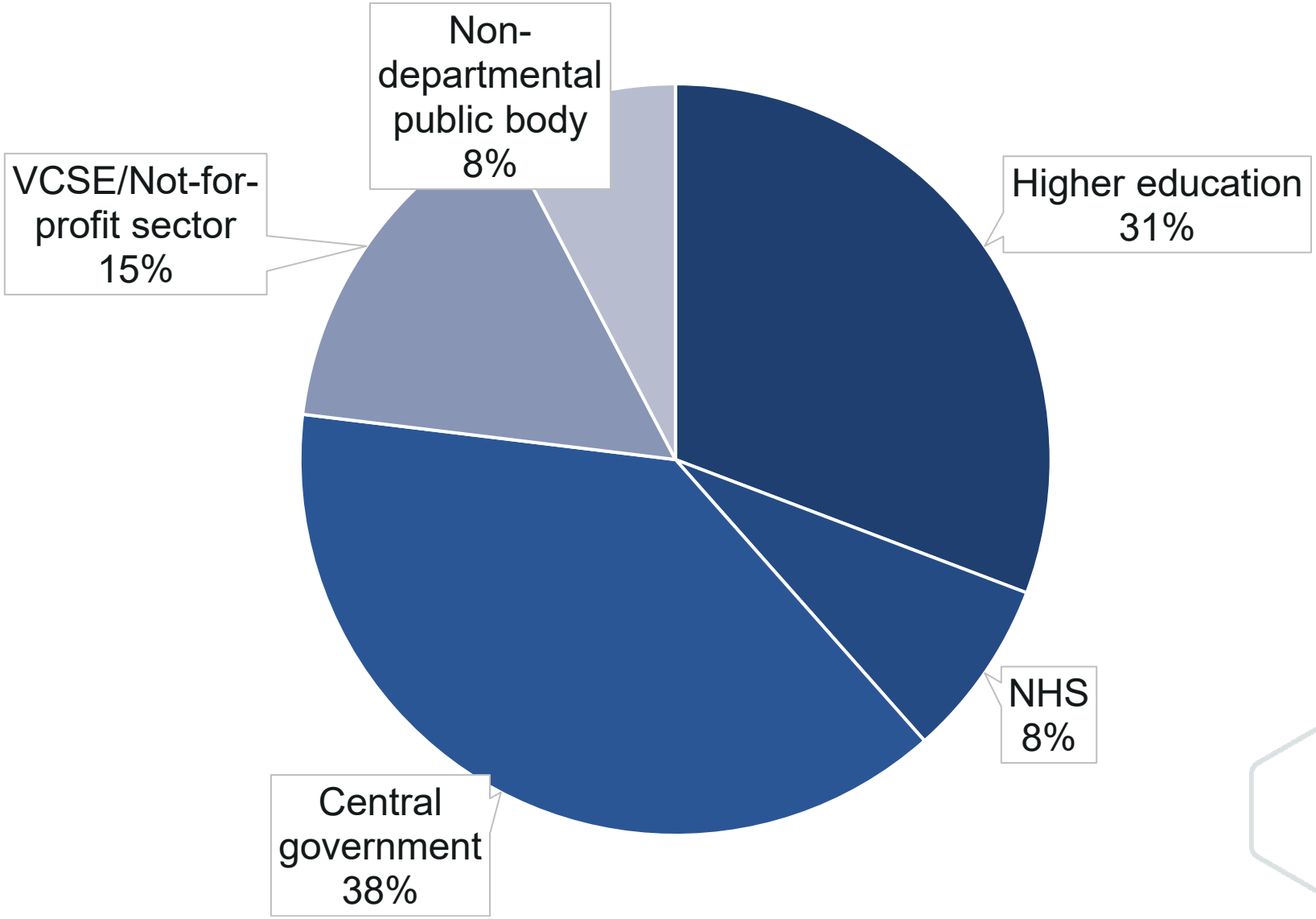
UK Data Service

Preliminary survey findings

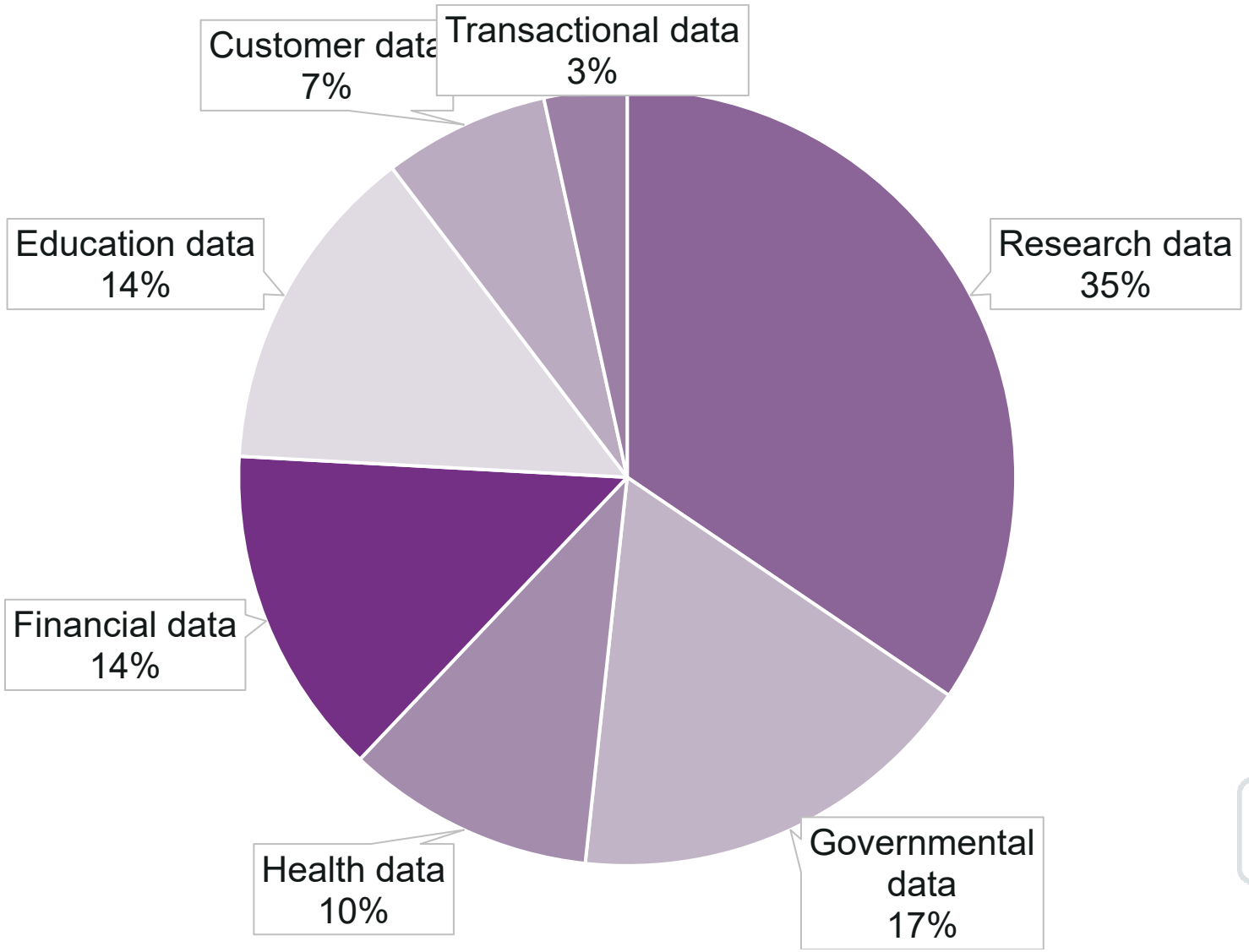
Survey with data owners: Respondents' role



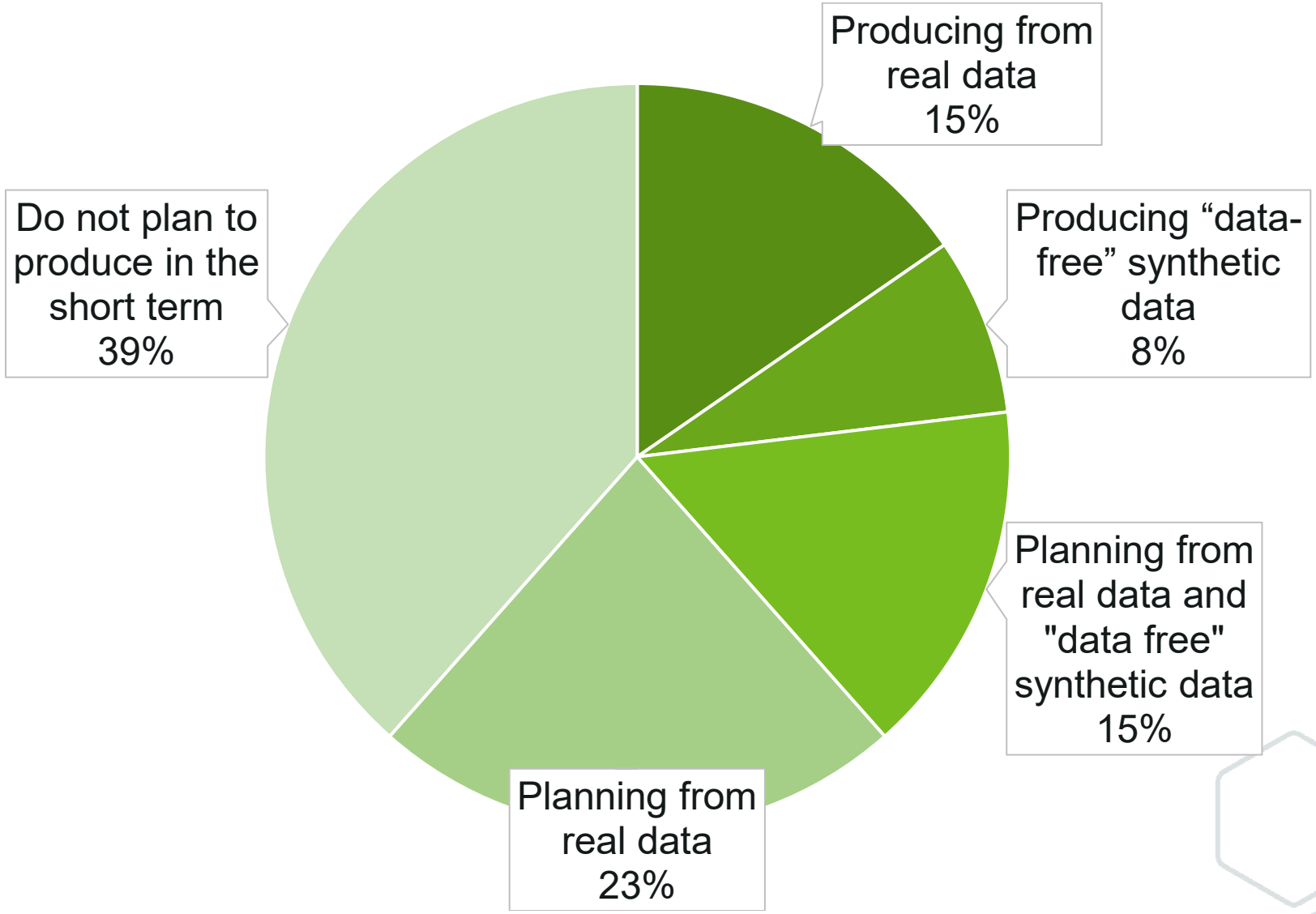
Survey with data owners: Sector



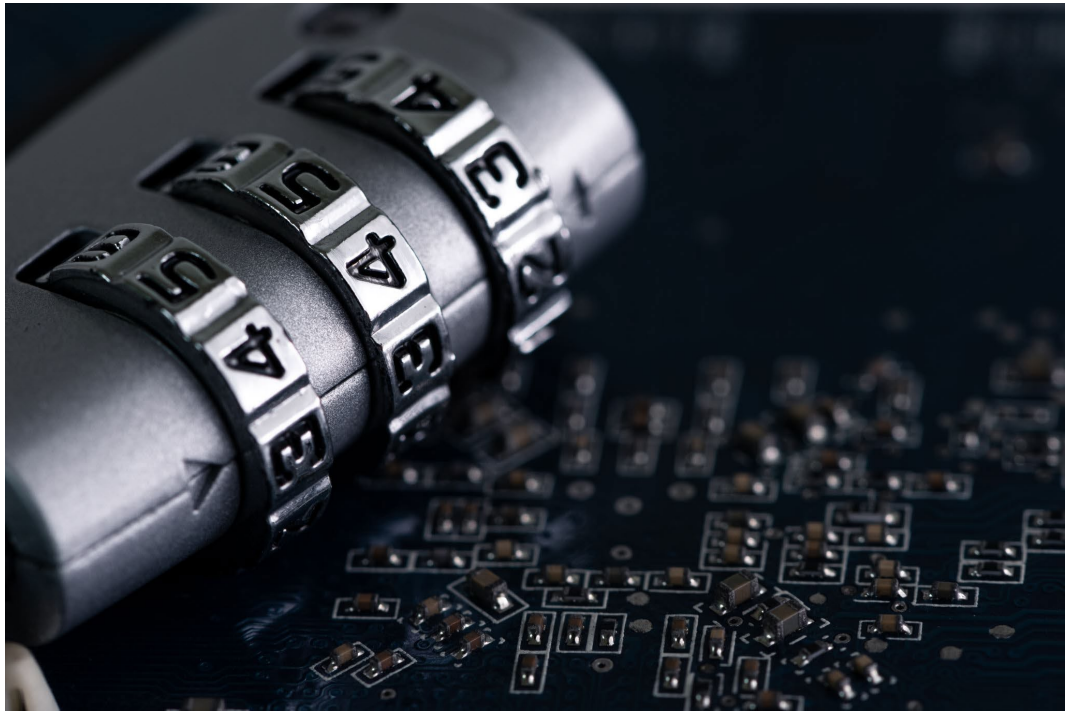
Survey with data owners: Real data production



Survey with data owners: Synthetic data production



Survey with data owners: Benefits



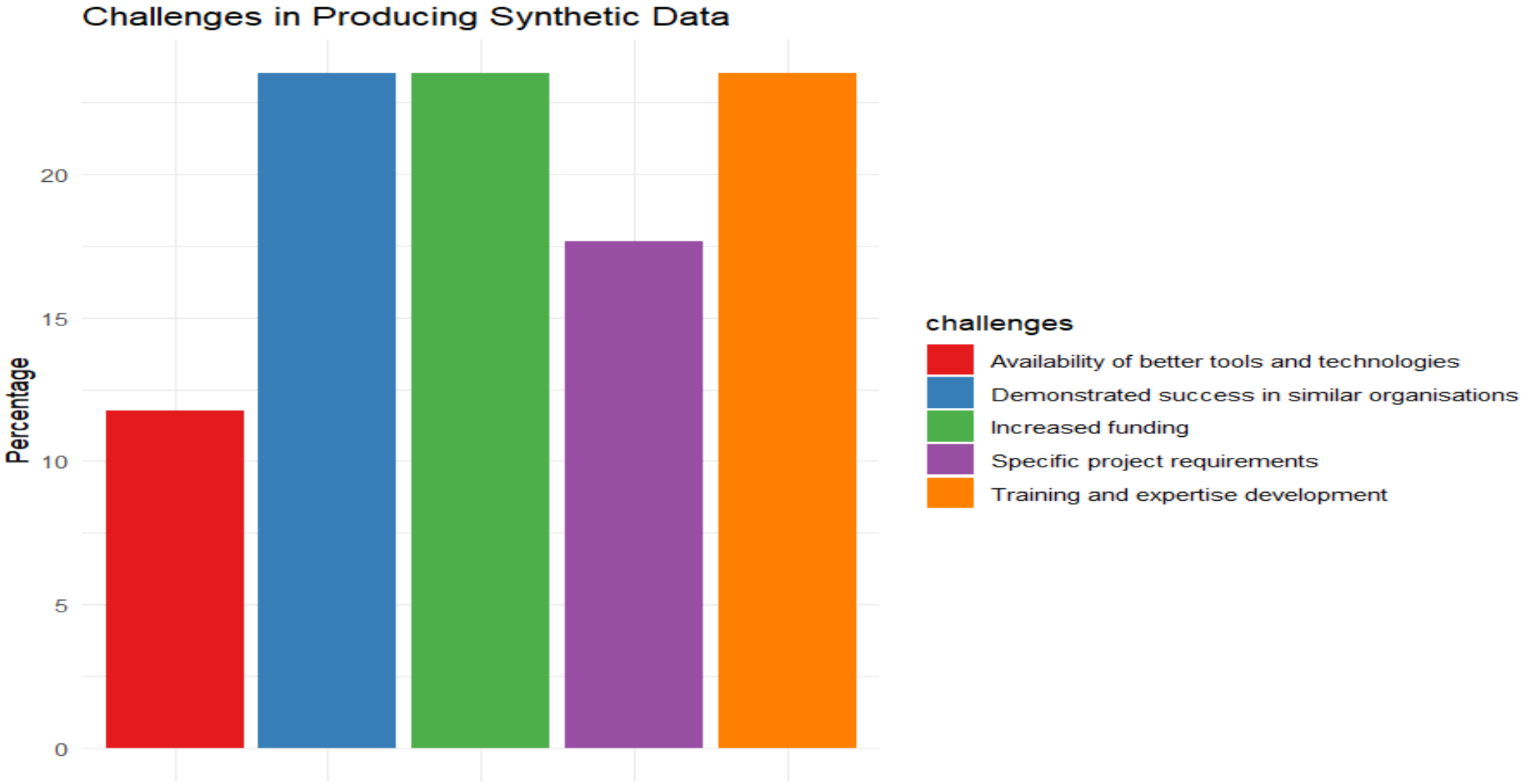
"Enhanced privacy, easier compliance with regulations.“

“[…]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).”

“[…] better data access requests and fewer failed projects/applications.”

"Increased data sharing, accelerated data access."

Survey with data owners: Challenges



Focus group with TRE representatives

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

- Practical implications
- Challenges
- Opportunities

Focus group invitation

We will be conducting an **online** focus group with TRE representatives on the **11th of December** 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