Generating Synthetic Breach Data for the National Cyber Crime Unit using Python

Butterfly Data built a Python tool that generates realistic synthetic data resembling the data exposed in a breach.

Challenge

The National Cyber Crime Unit (NCCU) responds to critical cyber incidents and pursues cybercriminals, with an internal team developing its cyber capabilities.

The NCCU lacked suitable data to test and improve its tools. It needed realistic synthetic data resembling what is exposed in a breach, covering companies in manufacturing, finance and government. The data would be used to test in-house tools, including AI models, and as a honeypot to attract cybercriminals and better understand their activity.

Solution

Butterfly Data built a Python tool using the Faker and Jinja packages to generate synthetic but realistic documents automatically. The tool:

- Used an object-oriented approach, with classes for individuals, employees, projects and companies that share attributes and build on each other
- Reproduced the networks of relationships found in real systems
- Used Jinja to turn Faker-generated data into documents such as invoices and forms
- Generated identifiers such as National Insurance numbers so they could not be mistaken for real ones, with the same care applied to free text such as emails between colleagues

Impact

The NCCU received realistic invoices, asset registers and expense records, customised to its pre-defined parameters, to support tool testing and honeypot work.

Butterfly Data also provided supporting documents recording how each data item was created and the ethical considerations involved in creating and using the synthetic data.

BUTTERFLY DATA

Ready to transform your data?

Book your free discovery call and find out how our bespoke data services and solutions could help you uncover untapped potential and maximise ROI.

hello@butterflydata.co.uk
+44 29 212 00140
butterflydata.co.uk

Get in touch →