SMART DATA FOUNDRY

Synthetic Data Package

Consumer Identity Factsheet

At a glance

- Identity data for 100k people
- ~1,000,000 rows of the synthetic individuals' data reflecting different organisations' view of the same person
- Multiple data issues present in the data to reflect 'real world' data complexity

Accelerate Product Development by:

- Testing and improving your processes for matching individuals across multiple simulated data siloes
- Building and iterating your data governance and processes for sensitive data on high utility simulated PII datapoints - with zero privacy risk
- Implementing, testing, and evolving error checking, validation, and imputation processes – all verifiable against 'ground truth' data

Core data attributes (see metadata for full details)

- 100,000 synthetic individuals' personal data
- Average of 10 different versions of each individual record
- Reliable citizen identifier for each variation, so that a verifiable 'ground truth' is always known and can be validated
- Multiple possible linkage keys provided (e.g. email, phone numbers, name)
- 100% synthetic with no precursor training data, or lineage issues
- Full PII – first, middle, last names for each individual, with DOB, title, and sex
- Addresses in standardised form: line 1, line 2, city/town, postcode, and country
- Land line and up to 2 mobile phone numbers for each person-variation
- Up to two email addresses for each version of the data subject
- Realistic data validation and hygiene issues

Key features

Synthetic current account customers, with personal details

- Synthetic Personal Information for 100,000 Synthetic Data Subjects, with an average of 10 versions of each
- Roman and non-Roman alphabet types present, which may vary between versions
- Multiple data 'mutations' (e.g. input error, transposition, truncation, concatenation, splitting, validation) replicating real data complexity
- Different versions of address, e.g. house name instead of number, or unreliable apartment/flat assignations
- Individuals may also be known by nicknames, middle names, or adopted names
- Common pitfalls like family members sharing names at same address, or leaving/returning home
- Home moves and new addresses

Key benefits

- Our synthetic data is generated from simulations, and contains the same properties as real-world data, removing any privacy issues by generating the data from scratch, meaning no real people, accounts, or transaction are contained in the output
- Available immediately with no need for Privacy Assessments or other Data Governance activity
- Can be tailored to meet exact requirements – please get in touch

Bayes Centre, 47 Potterrow,
Edinburgh EH8 9BT

www.smartdatafoundry.com

SMART DATA FOUNDRY

Synthetic Data

Consumer Identity Factsheet

Example Metadata of each file type

- People

Name	Type	Min	Max	# nulls	Description
citizen_id	string	BG000020C	ZZ999936C	0	An external identifier for the person, typically the national insurance number
dob	string	0000-01-15	9999-10-24	282,133	The person's date of birth, formatted as an ISO-8601 date (YYYY-MM-DD)
first_name	string	ARK	혜진	37	The forename (first name) of the person
middle_name	string		혜진	541,017	The middle name/names of the person (possibly null)
last_name	string		혜진	2,812	The surname (last name, family name) of the person.
title	string	DOCTOR	professor	70,619	The title (Mr, Ms etc.) to use when communicating with the person
sex	string	F	M	324,537	The gender (sex) of the person. Currently only supports M and F.
address_line1	string	3ORMOND DRIVE	§O0W, Cator Close	54,502	First line of address. Free text, mixed case.
address_line2	string	1 Holmwood Gv.	z 29 Old Oak Road	839,015	First line of address. Free text, mixed case.
address_city	string	AIDSTONE	ythe	222,716	Second line of address. Nominally a postal town. Upper case.
address_postcode	string	11 6an	zy8 1ej	77,194	The post code, nominally in standard form: upper case, with a single space separator and no padding.
address_country	string	GB	GB	0	Country.
address_type	string	Delivery	home	216,404	Type of address either home or delivery
land_line	string	(0000) 408 5662	9797 642 7281	182,168	Land line phone number (if any), without country code which is +44 for UK and should have no spaces, punctuation etc.
mobile_number1	string	(0011) 499 8167	97992456840	0	Mobile phone number (if any). Normally starts with 07 for UK and should have no spaces, punctuation etc.
mobile_number2	string	(0024) 252 7377	97906916825	540,586	Second mobile number.
personal_email1	string	!2366@hotmail.net	~-#7063@googlemmail.com	0	First email address (personal)
personal_email2	string	4754@ggmail.net	–7188@ymail.cOm	693,182	Second email address (personal)

Bayes Centre, 47 Potterrow,
Edinburgh EH8 9BT