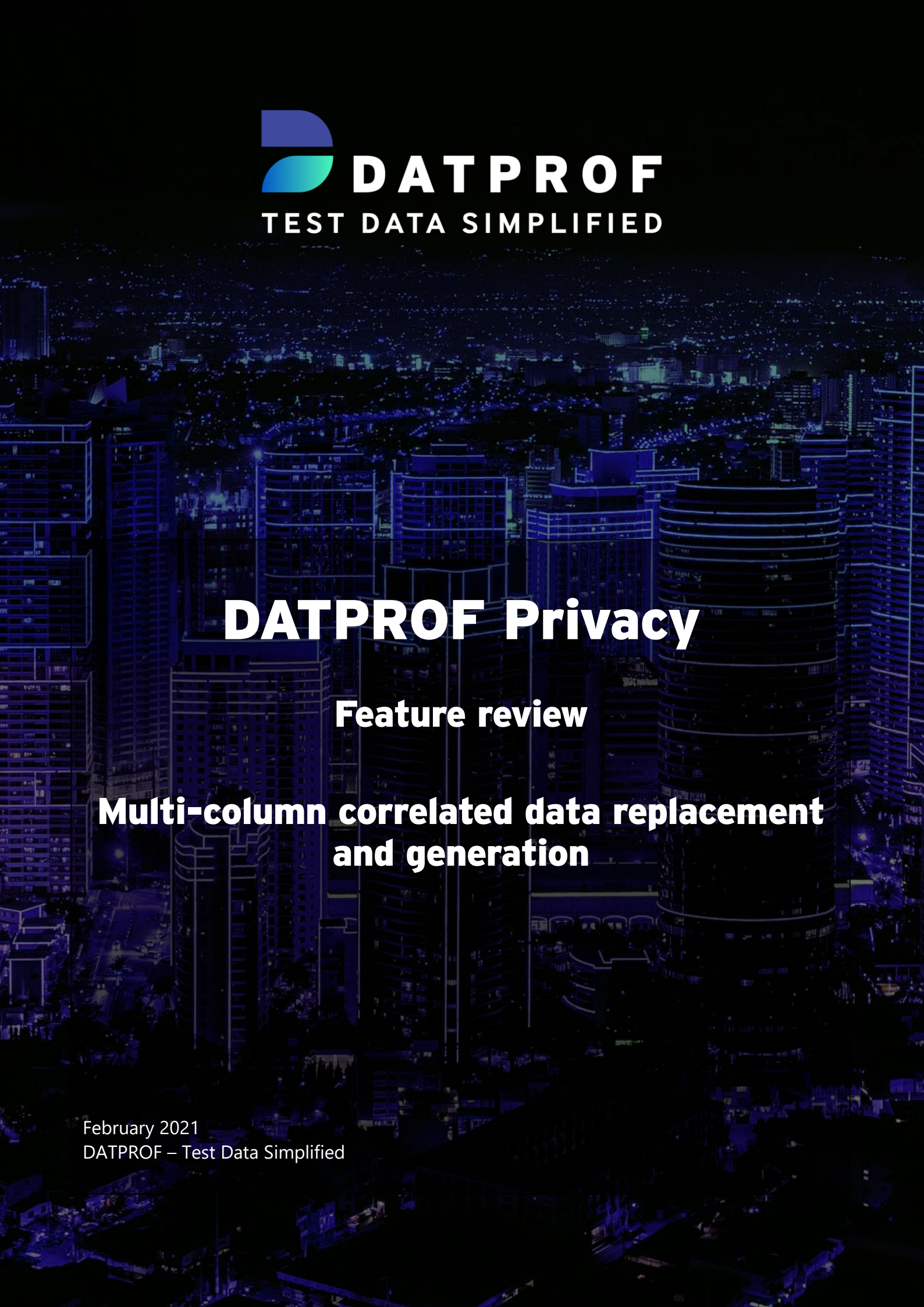


DATPROF Privacy

Feature review

Multi-column correlated data replacement and generation

February 2021
DATPROF – Test Data Simplified

Introduction

For more than seven years DATPROF Privacy has helped customers around the world address the compliance challenges faced when holding Personally Identifiable Information (PII) in databases used for Development, Testing, Training, Quality Assurance and User Acceptance Testing. There are often more than five copies of any one database in these lower environments leading to extensive exposure for data escape (or, bluntly, theft) and the subsequent catastrophic impact on the business when such breaches are made public, which data protection legislation obliges the organisation to do. This leads to financial penalties and a loss of public confidence.

As part of the Test Data Provisioning framework, DATPROF Privacy remediates existing data content using data masking or data generation techniques. But what happens when you have a set of columns in a table that have related content, for instance town, county, state and zip code?

It's for this reason that we've developed our new Privacy generator feature "Value from multi-column seed file". This technique allows you to match one or more columns in a table with corresponding columns in a pre-prepared .CSV file and will result in realistically related replacement or generated values in the row.

This is of particular importance to applications which may use such correlated information as components of an external lookup or as factors in determining an action such as a risk based analysis based upon an address or geo-location.

Additionally, it's always good practice to deliver an environment where the data "looks right"; it's "joined up properly"; it doesn't look "broken".

The "Value from multi-column seed file" delivers this in an intuitive, easy to use way.

Getting started

Multi-column datasets can be either user-defined or provided by your DATPROF Implementation Consultant from our extensive library of replacement candidates evolved over years of masking projects. All replacement data is public domain and therefore not liable to copyright infringement.

The example used in this document is a .CSV which contains US Town, County, State, Abbreviated State, ZipCode, Country, Latitude and Longitude. For instance

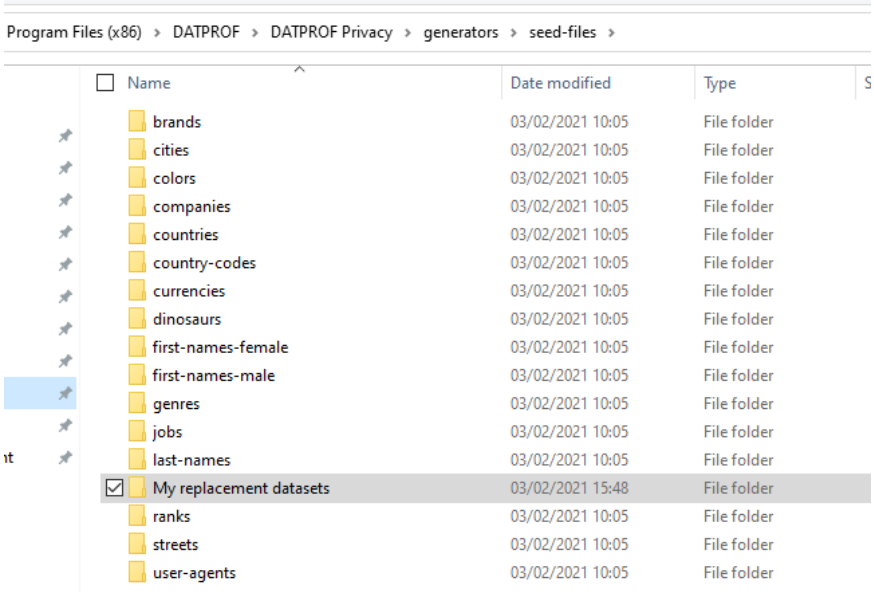
Memphis	Shelby	Tennessee	TN	38114	US	35.0981	-89.9825
Memphis	Shelby	Tennessee	TN	38115	US	35.0579	-89.864
Memphis	Shelby	Tennessee	TN	38116	US	35.0303	-90.0123
Memphis	Shelby	Tennessee	TN	38117	US	35.1124	-89.9034

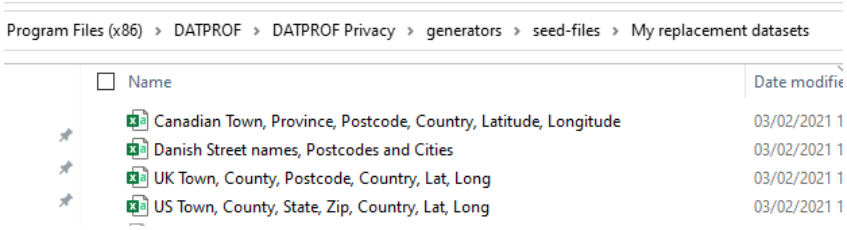
This file is available upon request.

*** For Elvis fans, the circled ZipCode encompasses Graceland ***

Planning and preparation is critical in provisioning your test data, whether it be masking copies of production to act as Test Data Masters or generating column/row data in one or more tables.

So the first step is to source and validate your replacement .CSV and place it in a convenient directory. We would recommend that you create a directory in the DATPROF Privacy installation directory to ensure they are available to all DATPROF Privacy developers:

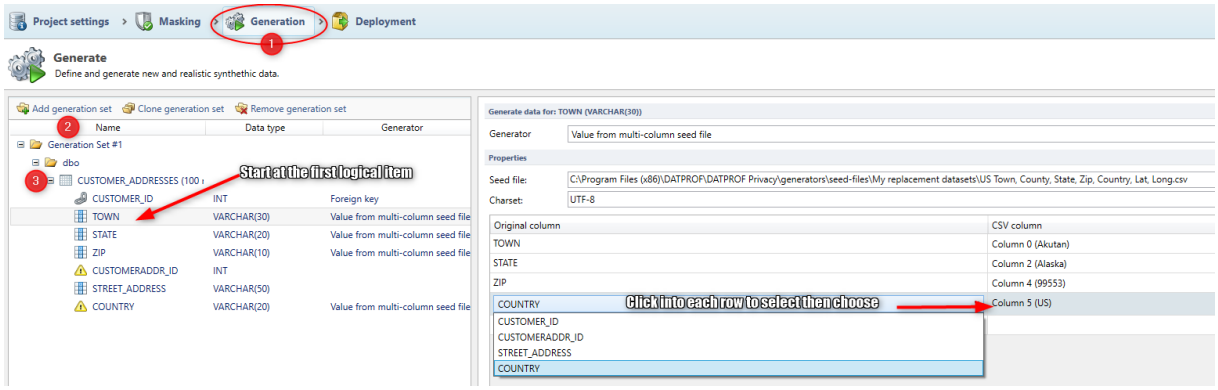




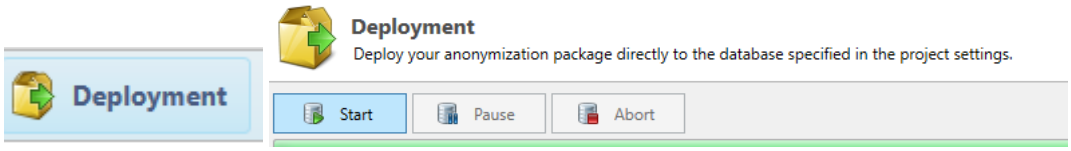
Adding new rows in a table using the "Value from multi-column seed file" generator

For further information on Synthetic Data Generation please request your copy of "DATPROF Privacy - Synthetic Data Generator 1.0" from your DATPROF Representative.

For the purpose of this paper, a new "CUSTOMER_ADDRESSES" table was created in our demo database which contains typical address fields. It also has a Foreign Key constraint to the CUSTOMERS table related by CUSTOMER_ID. The Primary Key will be a sequence starting with "1" and we want to fill the address fields with related items. The source seed file is the US address dataset (shown in the directory listing above) and you'll see the fields, plus content from the first row of the file, in the rule configuration:



Let's run it:



Our result set is 100,000 new address rows, appropriately populated:

Results | Messages

	CUSTOMERADDR_ID	CUSTOMER_ID	STREET_ADDRESS	TOWN	STATE	ZIP	COUNTRY
1	1	93601	16Th Ave E	Meridian	Mississippi	39304	US
2	2	93602	Woodlawn Ave Ne	Saint Paul Island	Alaska	99660	US
3	3	93603	Sw Hinds St	Hampton	Virginia	23664	US
4	4	93604	S Albro Pl	Oklahoma City	Oklahoma	73142	US
5	5	93605	Northwood Rd Nw	Mauk	Georgia	31058	US
6	6	93606	54Th Ave Sw	Deerfield	Michigan	49238	US
7	7	93607	43Rd Ave S	Madrid	Iowa	50156	US
8	8	93608	Shilshole Ave Nw	Big Sandy	Texas	75755	US
9	9	93609	Sw Raymond St	Las Vegas	Nevada	89124	US
10	10	93610	Mercer St I5 Express Rp	Salem	New Hampshire	3079	US

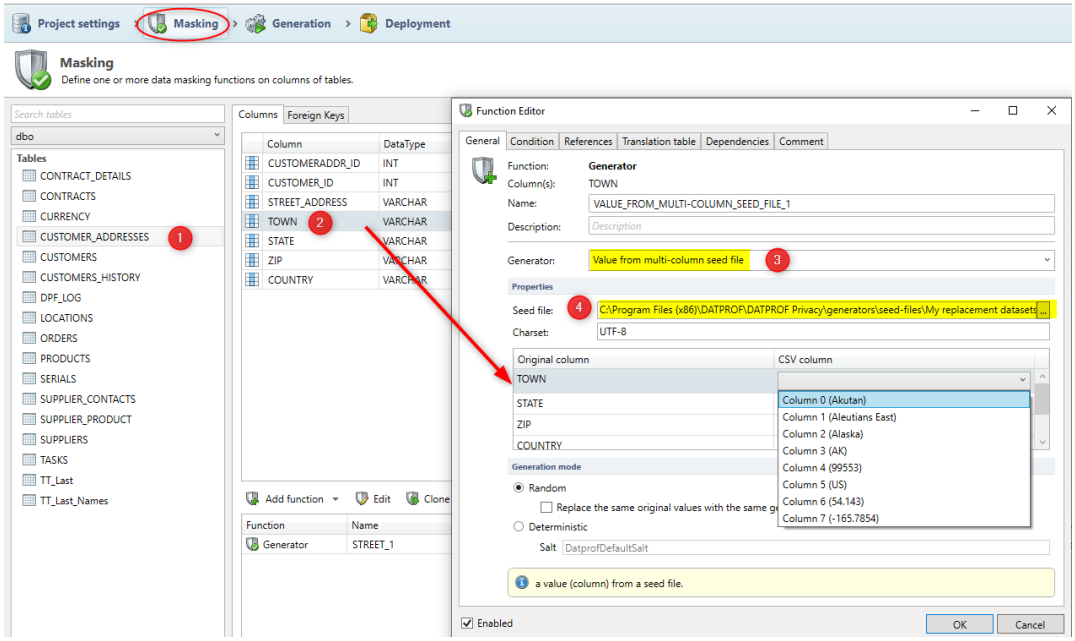
*** The CUSTOMER_ID foreign key content was drawn from the parent table by specifying the "Foreign Key" generator. This guarantees referential and data integrity. ***

Masking using the "Value from multi-column seed file" generator

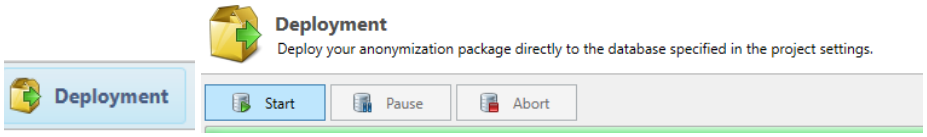
Continuing our story, the data generated above has been used for Unit Testing and the update was rolled out to LIVE. The table is now being populated in LIVE so it's real PII data and therefore must be masked when it is being provisioned back into the Lower Environments.

Like all masking functions in DATPROF Privacy the start point is a column in the target table. With correlated replacement it's generally best to start with the first column in the logical grouping. In this scenario it's STREET which will be a straight-forward replacement. Next are the related items TOWN, STATE, ZIP and COUNTRY. So, we start with TOWN and work our way down the list, in this case double-clicking the original column to present a picklist of the column names not already configured in this specific masking function.

You can equate these steps to your own data structures and replacement content:



Once configured, execute the project:



Our result set is the 100,000 rows of address data, suitably masked with appropriate, related values:

Results | Messages

	CUSTOMERADDR_ID	CUSTOMER_ID	STREET_ADDRESS	TOWN	STATE	ZIP	COUNTRY
1	1	93601	Marcus Ave S	Edmond	Oklahoma	73083	73083
2	2	93602	Nw 101St St	North Stonington	Connecticut	6359	6359
3	3	93603	Northgate Wy I5 Rp	Oceanside	California	92049	92049
4	4	93604	W Olympic Pl	Noyes	Minnesota	56740	56740
5	5	93605	25Th Pl W	Shelby	North Carolina	28150	28150
6	6	93606	N 79Th St	Livingston	Illinois	62058	62058
7	7	93607	W Ohman Pl	Brooklyn	New York	11217	11217
8	8	93608	47Th Ave W	La Harpe	Illinois	61450	61450
9	9	93609	Renton Ave S	San Jose	California	95172	95172
10	10	93610	S Morgan Pl	Memphis	Missouri	63555	63555

Summary

The multi-column seed file generator is a powerful function within DATPROF Privacy which delivers both realistic test data content and volume.

The ability to define column content using the "out of the box" replacement datasets, as well as "user-defined" datasets and seed files will help you achieve your test data provisioning goals by delivering credible, useable, testing content.

About DATPROF

Behind all products and services there is a team that builds and supports it. From phones and cars, to bank accounts and health insurance, there are developers and testers working across the world that rely on good quality test data.

It is for these people that we build our products to enable them to focus on what's really important to them. By giving them the right data, at the right place and at the right time they can innovate faster and deliver their products with the highest quality. In the end *everyone* benefits.

We **simplify** getting the **right test data** in the **right place** at the **right time**!