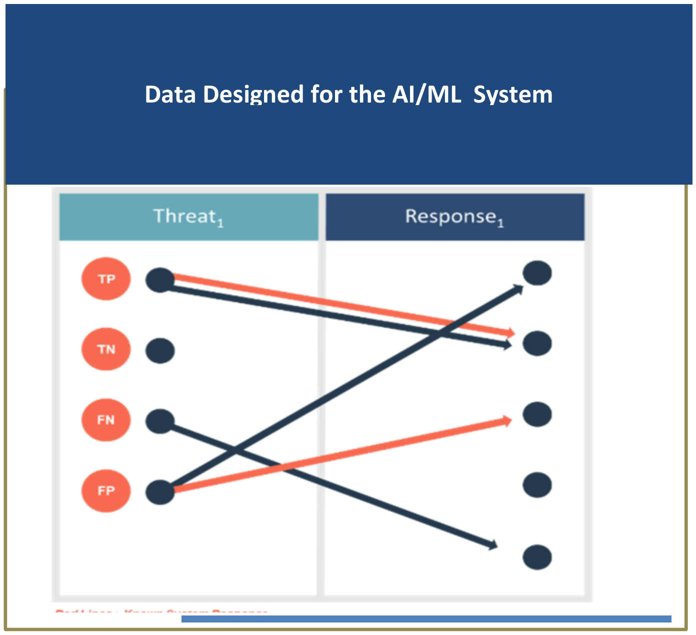
ExactData
Next Generation Systems Testing

How Synthetic Data is Changing the Face of AI/ML

Executive Summary: Use of high fidelity fully synthetic data for testing and training Machine Learning/Artificial Intelligence, ML/AI, solutions will decrease the time and cost to deploy while increasing accuracy. ExactData specializes in creating synthetic data that is designed for the ML/AI system and generated at scale utilizing rules-based algorithms. Though our approach, the synthetic data will be engineered to mirror the production databases and objects, while assuring all the relevant data scenarios and use cases are covered at statistically valid volumes. Furthermore, all datasets will have a known system response ground truth.


Data Designed for the AI/ML System
Threat$_1$
Response$_1$
TP
TN
FN
FP
Red Lines : Known System Response

The Problem: Currently, data sets or test objects are being used that are derived from production data. Such datasets rarely have known ground truth. The data scenarios or use-cases may be present, but where in the data or how often they occur is generally unknown, and the problems of knowing or determining ground truth for these datasets increase directly in proportion to the size of the datasets.

Detecting activity in large datasets poses a problem because generally any given characteristic or signal within a dataset represent a minority class or feature.

As data science has shown, in the absence of known ground, truth these imbalanced datasets can be very difficult to analyze, because ML/AI algorithms tend to show a bias for the majority class or features, leading to misleading conclusions.

The Solution: The use of ExactData's synthetic data generation technologies will enable you to properly train and test these systems by enabling the creation of sufficiently large datasets, with known given relevant data scenarios and use cases covered, and a known ground truth, enabling determining system response classifications as True Positives, TP, True Negatives, TN, False Negative, FN, and False Positive, FP.

Large volumes of relevant data with positive and negative use cases with a known system response enables quantitative outputs from the ML/AI system to be placed in a "confusion matrix", to more easily see how the ML/AI System Under Test, SUT, behaved.

A confusion matrix tells us the rate of FP, FN, TP and TN for a test or predictor. But we can make a confusion matrix only if we know both the predicted values and the true values for a data set. And while this is generally not possible where large production data sets and objects are used, because ground truth is generally not known, this can always be known, and the confusion matrix completed, with the use of synthetic data generation technologies.

A confusion matrix, in predictive analytics, is a two-by-two table.

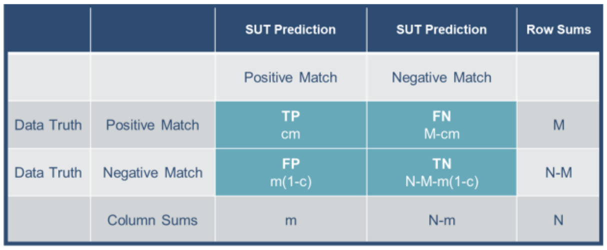
Truth for positive or negative matches is contained in the rows. ML/AI Response positive or negative matches are contained in the columns. The sum of the first row is defined as M, the total number of true positive matches (known from the Truth). ***In typical testing/training of ML/AI systems with real data, this number is never known.***

The total number of positive matches predicted by the SUT is defined as m (observed from SUT output). The total number of elements in the matrix, N, is defined based on the number of elements in the test/training data sets and the nature of the test. Finally, using the Truth, the precision c may be determined as the fraction of the predicted matches that are true positive matches, or c = TP/(TP + FP). When the matrix is completed, we have N = TP + FP + TN + FN.

ML/AI Systems Must Detect Minority Class Activity in Large Datasets

Example Use Case: A ML combined with National Language Processing, NLP, solution to automatically identify technical terms and names associated with complex descriptions of weapon systems buried in various types of data sources. Synthetic data sources would be generated that would simulate localized unstructured electronic documents, and various types of multimedia files including image files in various formats. Weapon systems descriptions would be inserted in various places and varied from the obvious to the very obscure. Outputs would include the test and training files that would include TP, FP, TN, FN and

Testing/training of ML/AI Systems with Real Data, T (Truth) is Never Known

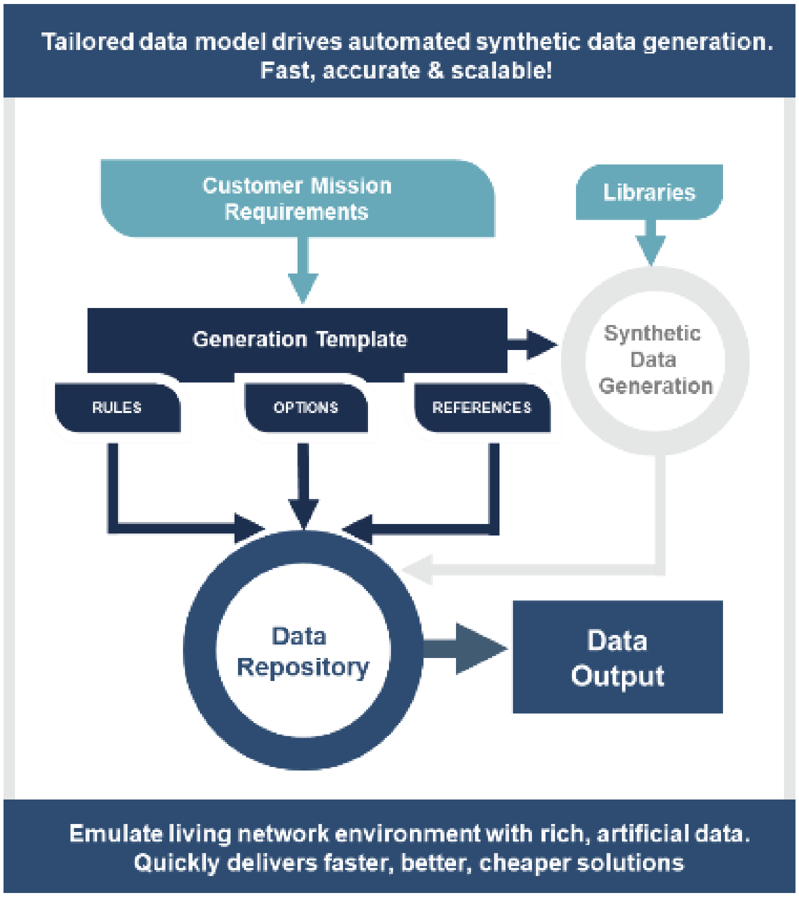


		SUT Prediction	SUT Prediction	Row Sums
		Positive Match	Negative Match	
Data Truth	Positive Match	TP cm	FN M-cm	M
Data Truth	Negative Match	FP m(1-c)	TN N-M-m(1-c)	N-M
	Column Sums	m	N-m	N

the system response or Truth file to enable a systematic and quantitative approach to development and deploying these systems.

The Risk: Risks are minimal in that high fidelity synthetic data generation technologies are proven and have been used in similar applications such as DARPA ADAMS.

The Value: Use of production data sources with manual modifications, is inefficient, costly and cannot be used to measure error rates nor effectively train systems. The difference the use of synthetic data generation technologies will make is faster deployments of solutions with a known level of performance that can be measured. ***You cannot improve something that cannot be measured.*** The payoffs is a strong ROI based on tradition methods realized through reduced development labor and cycle times along with significant risk reduction and cost by reducing the costs of errors.



Tailored data model drives automated synthetic data generation.
Fast, accurate & scalable!

Emulate living network environment with rich, artificial data.
Quickly delivers faster, better, cheaper solutions

How Synthetic Data is Changing the Face of AI