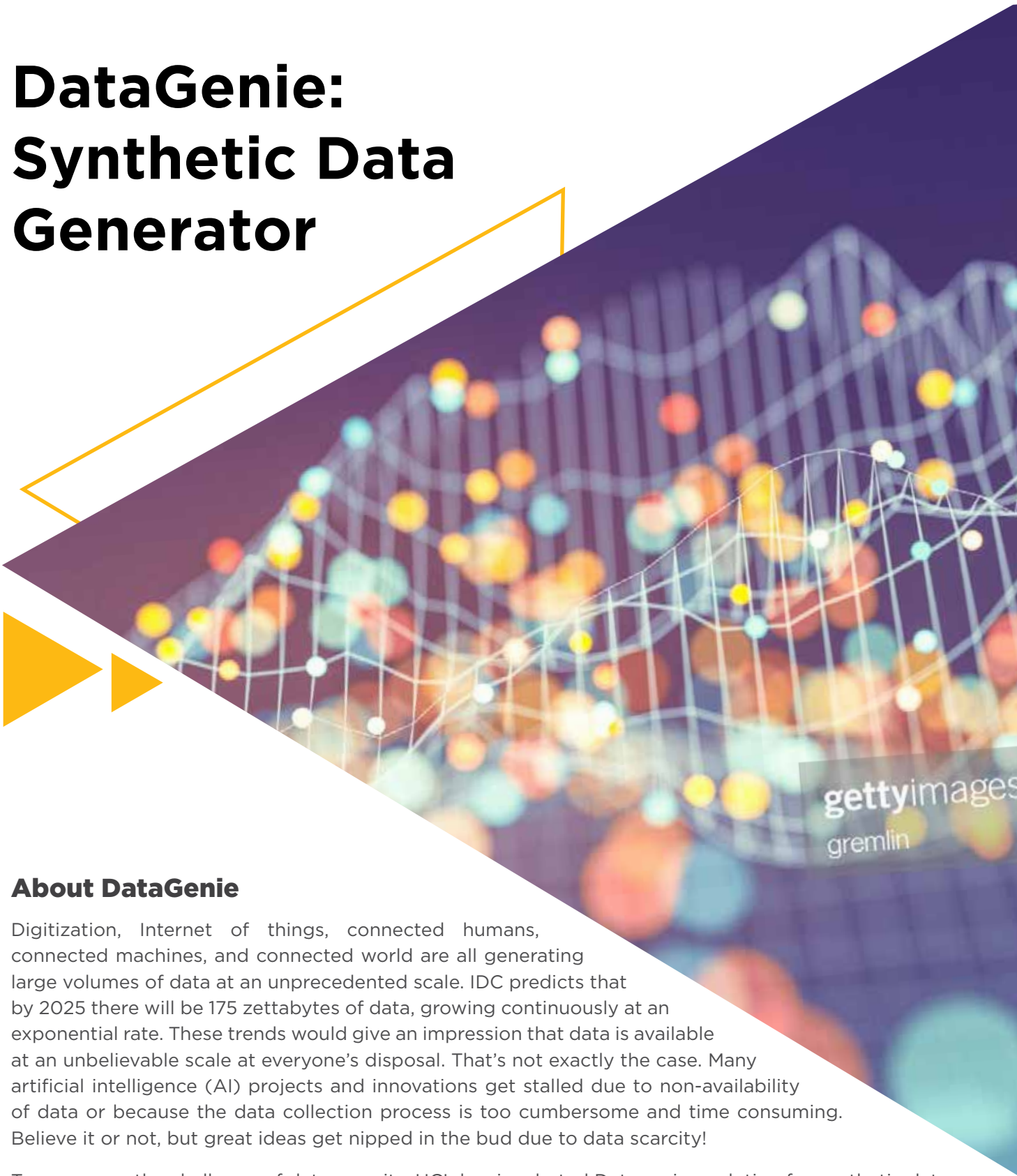
www.hcltech.com

DataGenie: Synthetic Data Generator



About DataGenie

Digitization, Internet of things, connected humans, connected machines, and connected world are all generating large volumes of data at an unprecedented scale. IDC predicts that by 2025 there will be 175 zettabytes of data, growing continuously at an exponential rate. These trends would give an impression that data is available at an unbelievable scale at everyone's disposal. That's not exactly the case. Many artificial intelligence (AI) projects and innovations get stalled due to non-availability of data or because the data collection process is too cumbersome and time consuming. Believe it or not, but great ideas get nipped in the bud due to data scarcity!

To overcome the challenge of data scarcity, HCL has incubated Datagenie - solution for synthetic data generation. This solution focuses on generating structured tabular data and images. A minimal input data prepared in consultation with an SME or collected from the field is required for generating large volumes of tabular data, which are of the categorical, continuous, or time series type. Such data can be used for modeling classification, regression, or forecasting problems. DataGenie can also help to augment images and generate images. Around a hundred images are required as input for generating thousands of output images.

Solution Overview

In DataGenie, classification and regression trees are used for the generation of structured data and augmentation is used to generate the image-based data. The following figure shows the whole process of generating structured data.

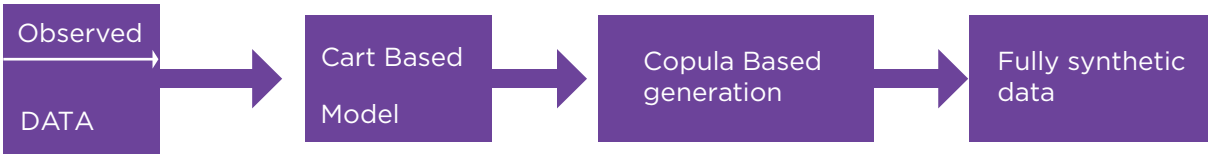


Fig1: Structured data generation process

For unstructured data like images, Datagenie uses augmentation to generate new samples of images. This has been implemented and demonstrated in various kinds of augmentations. The generated images and list of augmentations are discussed in the subsequent sections.

Key Features

DataGenie is able to generate structured as well as unstructured data for analytical purposes.

The usefulness of data can be proved by looking at various data comparison metrics which show the similarity between the observed and the synthetic data. It can generate data for all purposes like

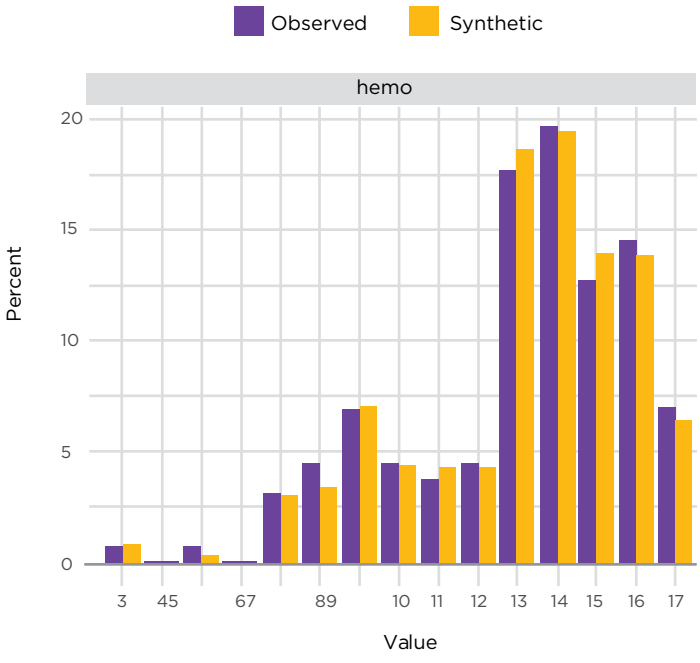
- Classification
- Regression
- Time series-based data

DataGenie can also compare the generated data with the observed data through histograms and difference of standard coefficients. It also has the capability to generate the image data with the help of various augmentations.

Results achieved

DataGenie has been deployed in generating data for the following use cases which helped in training the models with a reasonable amount of data, and resulted in improved model performance. The generated data resulted in kick starting innovation, which otherwise, would not have been possible.

The following histogram shows the results achieved on a kidney dataset which had only 150 rows in the start. Through the use of DataGenie; the dataset was expanded to 5000 rows. The following histograms show the comparison between the observed and synthetic data.



DataGenie can also be used to generate images. The images can be generated through many augmentations. Some of the augmentation are as follows.

- **Rotation:** In this augmentation the images are rotated in given range of angle. Following image was rotated by (-25 to 25 degrees).

Fig3: Rotation Augmentation on Car Image

- **Translation:** Datagenie can translate the image from its actual position and get augmented one as in the following case the image was translated by 20 pixels towards left.

Fig4: Translation Augmentation on Car Image

- **Additive Gaussian Noise:** A typical model of image noise is Gaussian, additive, independent at each pixel, and independent of the signal intensity, caused primarily by Johnson–Nyquist noise (thermal noise), including that which comes from the reset noise of capacitors. Amplifier noise is a major part of the “read noise” of an image sensor, that is, of the constant noise level in dark areas of the image. It can also add this noise to the given image for augmentation as given in the following example.

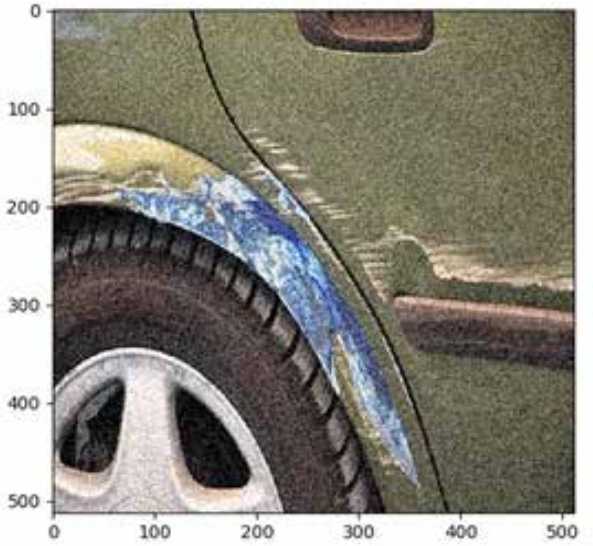


Fig5: Gaussian Noise Augmentation on Car Image

- **Color Space Transformation:** A color space is a specific organization of colors. In combination with physical device profiling, it allows for reproducible representations of color, in both analog and digital representations. A color space may be arbitrary, with colors assigned to a set of physical color swatches and corresponding assigned color names or numbers such as with the Pantone collection, or structured mathematically as with the NCS System, Adobe RGB and sRGB. RGB to HSV transformation is used for augmentation the following example shows this.

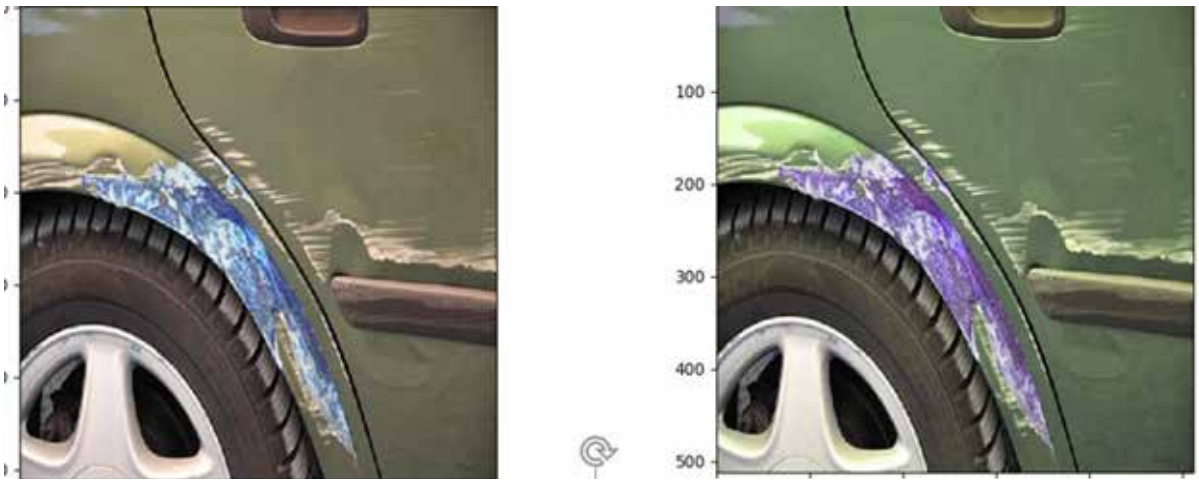


Fig6: Color space transformation showing original and generated Image

- **Grayscale levels augmentation:** Augmented images of various levels of grayscales can be generated as well. This is demonstrated in the following example.

Fig7: Various levels of Grayscale augmentation

- **Crop and Pad:** Data Genie Augmenter that crops/pads images by defined amounts in pixels or percent (relative to input image size). Following example shows crop and pad plus rotation.

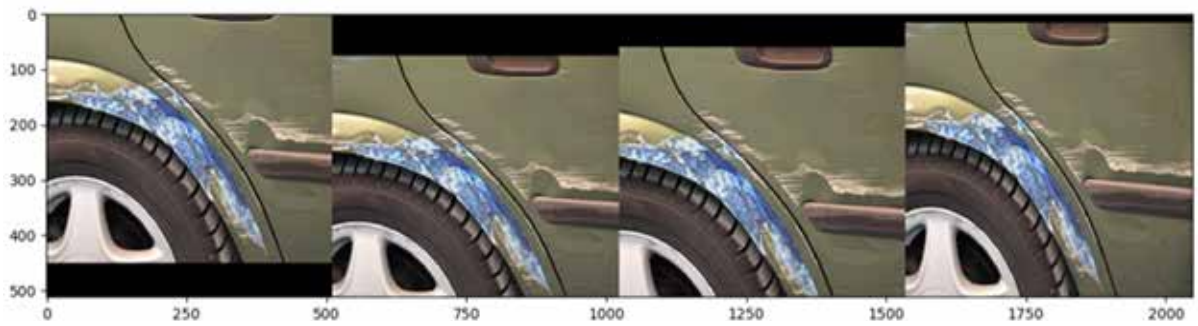
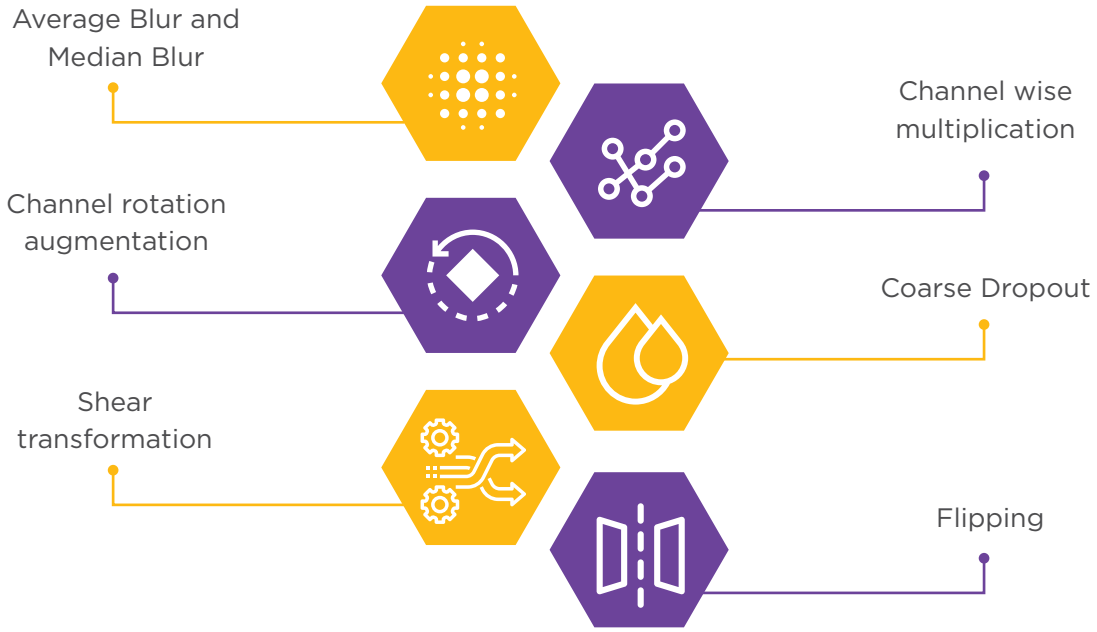


Fig8: Crop and Pad Augmentation

Additionally, following augmentations can also be achieved.



DataGenie has also been successfully used to augment medical images for detection of knee joint detection and localization.

Conclusion

DataGenie holds a lot of promise in highly regulated industries like financial services, medical, health care, clinical trials etc. Image augmentation and generation in combination with transfer learning, holds huge promise when getting real data from the field is a challenge.

Though synthetic data is not a complete substitute for real data, since it would be practically impossible to cover all real-world scenarios, it helps to kick start AI projects, while real data collection progresses on the field. Labelling effort can also be saved since labelling can be a part of the generation process itself. Training deep learning models with synthetic data and real data will help to protect the model against adversarial attacks and improve data security and the robustness of the models. The model is exposed to new types of data which is a little different from real data so that overfitting issues are taken care of.