

..... INTELLIGENT

DATA FAKING

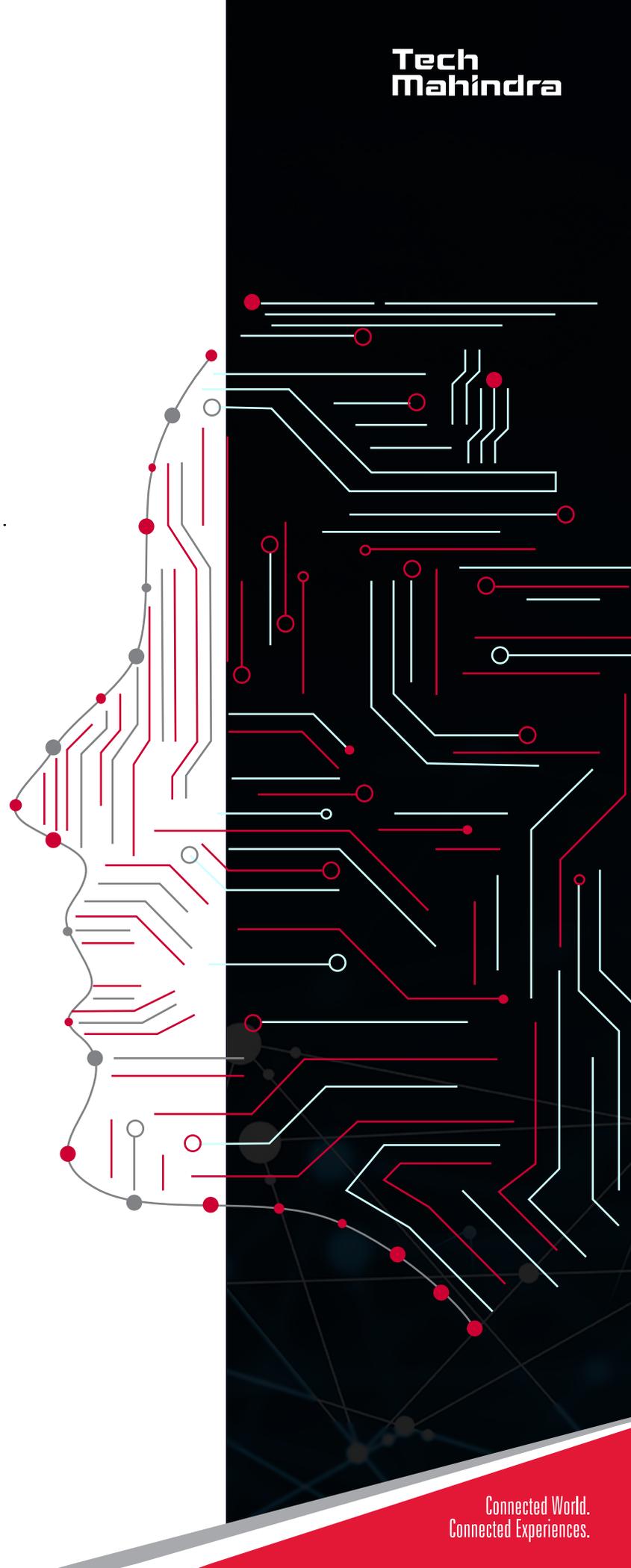
..... USING

AI | ML

Authored by
Kishore Kandula

ABSTRACT: This point of view shares an approach to generate faked data for testing cycles and other non-production purposes by employing Artificial Intelligence (Machine Learning-Natural Language Processing) in order to help organizations avoiding the misuse of data.

.....



Executive Summary

Many organizations inadvertently breach information when they routinely copy sensitive/ regulated or industry specific production data into non-production/Test environments. As a result, data in non-production/Test environment has increasingly become the target of cyber criminals and can be lost or stolen. Data breaches in non-production environments can cause millions of dollars to remediate; entailing irreparable harm to reputation and the brand.

Objective

With this paper, we have shared an approach to generate faked data for testing cycles and other non-production purposes (by employing Artificial Intelligence and ML techniques) in order to help organizations preventing the misuse of the data. To enable this prevention, A lot of packages for NLP (ML) are freely available online namely SPACY, NLTK, CoreNLP. SPACY supports Named Entity Recognition very efficiently to identify specific attributes.



Data Faking

Data Faking is the process of hiding original data with changed content (replaced with special characters or similar data) to be used in Test Environment to perform testing activities. The foremost reason for applying "Faking" to a data field is to protect the data that is classified as Personal Identifiable Data/ Sensitive Data/ Commercially Sensitive Data. Also ensuring that data continues being usable for undergoing further valid test cycles.

■ Why Fake Data?

Organizations share data with other users for a variety of business needs

- Copy production data into test/development environments allowing system administrators to test upgrades, patches and fixes
- Businesses, competitive in nature, require new and improved functionality in the existing production applications. As a result, application developers require an environment mimicking close to production (to build) and test the new functionality; ensuring that the existing functionality does not break
- Retail organizations share customers' Point- Of- Sale data with market researchers to analyze customer buying patterns
- Pharmaceutical or healthcare organizations share patients' data with medical researchers to assess the efficiency of clinical trials and medical treatments

As a result of the cited above reasons, organizations copy millions of sensitive (customer and consumer) data to non-production environments, however, a handful of organizations actually plan and work towards protecting the data when sharing with outsourcers & third parties.

About SpaCy

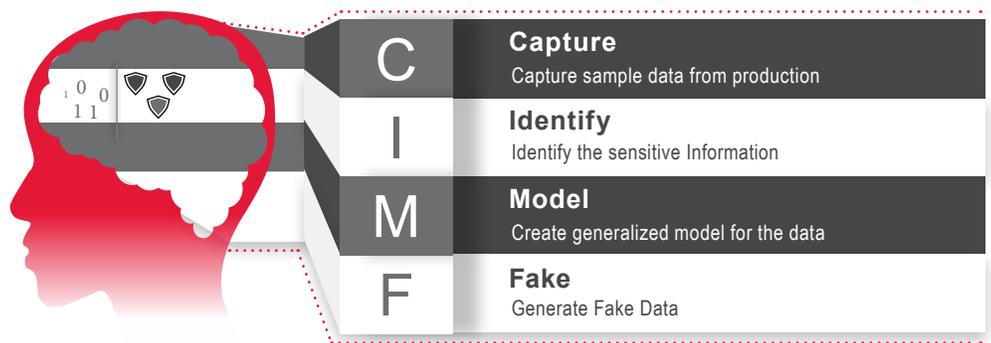
SPACY (<https://spacy.io/>) is an open-source package for advanced Natural Language Processing, written in Python and Cython. The package is published under the MIT license and offers statistical neural network models for English, German, Spanish, Portuguese, French, Italian, Dutch and multi-language NER (Named Entity Recognition), as well as tokenization for various other languages

Advantages of SpaCy

- Rapid as compared to other packages e.g. NLTK (Natural Language Toolkit)
- Features convolutional neural network models for part-of-speech tagging, dependency parsing and named entity recognition
- Easy to use and need not be an NLP Expert to start off with SPACY
- Supports in identifying domain specific data for faking

Proposed Approach for Data Faking

We define a more generic way of data faking which can be used not only for fake data but can also generate consistent data. This recommended approach does not need access to any Database like a SQL server or an Oracle Database, however only a sample dataset in file format of Excel or CSV or Json etc. will suffice.



The approach is as shown in the figure below

CAPTURE

Capture a sample dataset which contains all the data from the production database and export it into formats such as Excel, Json, and CSV etc.

IDENTIFY

In this step we use Natural Language Processing (NLP) based entity recognition to identify the general attributes such as name, organization etc.

MODEL

Create model using SpaCy with Python and attributes which are specific to the project/ domain use TF-IDF (Term Frequency Inverse Document Frequency) and add them to the model generated in the identification phase

FAKE

Once it has identified the attributes, we can use any of the below approaches to fake the sensitive content

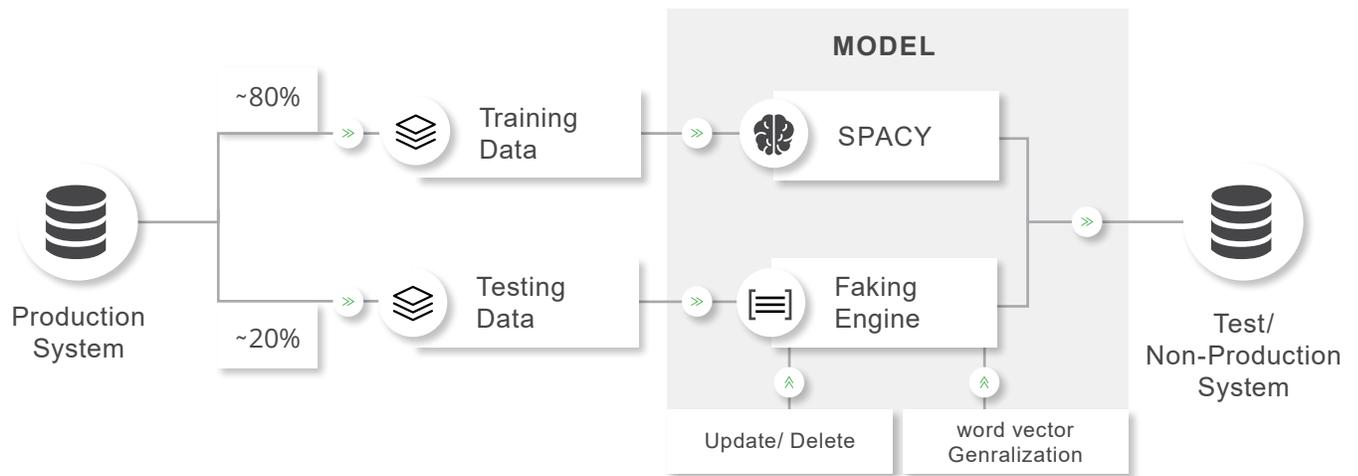
1. Update/Delete

We can define rules for specific entities such as all names must be replaced with **** or any standard name such as John Doe etc.

2. Word Vector Based Generalization

In this data generalization approach, the nearest neighbor of the word in the vector space is utilized to generalize the attribute

■ Technical Approach to create models to Fake Data



SPACY identifies sensitive attributes e.g. SSN, Credit Card details etc. Faking Engines support in applying data faking activity. This approach would help testers to define test data which is relevant to production data.

■ Advantages of Defined Approach

- Models are reusable as SPACY supports multiple languages with minimal rework
- Identify sensitive data automatically for data faking
- Models can be reused multiple domains e.g. Retail, Banking
- Easy to integrate with TDM and Test automation solutions
- Saves up to 70% of efforts in Data Faking
- Minimizes dependency on DBA/ Business Analysts to provide data for testing

As per the present era, AI/ML are getting used in almost all the areas of software development and testing including test data faking and masking. Test Data has proved to be one of the vital feature for testing and test automation therefore usage of AI/ML packages will involve minimal efforts; steering benefits to the business. There are a lot of packages available for ML from open-source community which are easy to combine and to generate

DATA CREATION
DATA FAKING
DATA SLICING ETC.

Kishore Kandula is a technology leader with 17 years of experience in Testing, QA and Automation in Software Service Industry. He has worked with various customers including Banking, Oil and Gas, Manufacturing verticals also managed large teams with proven experience in Test automation, RPA, DevOps and Agile initiatives, Enterprise delivery pipeline for CI/CD/CT. He is frequent with participating in customer workshops, providing the right tools, right framework and required approach to generate early ROI. Kishore has established expertise in setting up end to end automation from design to execution using different tools which include Licensed and Open source also as he is certified in test automation, RPA, and Machine Learning areas.



Kishore Kandula
Practice Head, Digital Assurance Services
Tech Mahindra
 [LinkedIn](#)

Tech Mahindra

Tech Mahindra, herein referred to as TechM provide a wide array of presentations and reports, with the contributions of various professionals. These presentations and reports are for informational purposes and private circulation only and do not constitute an offer to buy or sell any securities mentioned therein. They do not purport to be a complete description of the markets conditions or developments referred to in the material. While utmost care has been taken in preparing the above, we claim no responsibility for their accuracy. We shall not be liable for any direct or indirect losses arising from the use thereof and the viewers are requested to use the information contained herein at their own risk. These presentations and reports should not be reproduced, re-circulated, published in any media, website or otherwise, in any form or manner, in part or as a whole, without the express consent in writing of TechM or its subsidiaries. Any unauthorized use, disclosure or public dissemination of information contained herein is prohibited. Unless specifically noted, TechM is not responsible for the content of these presentations and/or the opinions of the presenters. Individual situations and local practices and standards may vary, so viewers and others utilizing information contained within a presentation are free to adopt differing standards and approaches as they see fit. You may not repackage or sell the presentation. Products and names mentioned in materials or presentations are the property of their respective owners and the mention of them does not constitute an endorsement by TechM. Information contained in a presentation hosted or promoted by TechM is provided "as is" without warranty of any kind, either expressed or implied, including any warranty of merchantability or fitness for a particular purpose. TechM assumes no liability or responsibility for the contents of a presentation or the opinions expressed by the presenters. All expressions of opinion are subject to change without notice.



www.techmahindra.com



connect@techmahindra.com



www.youtube.com/user/techmahindra09



www.facebook.com/TechMahindra



www.twitter.com/Tech_Mahindra



www.linkedin.com/company/tech-mahindra