Project no. 826278

SERUMS

Research & Innovation Action (RIA)
SECURING MEDICAL DATA IN SMART-PATIENT HEALTHCARE SYSTEMS

Report on Refined Data Masking, Data Fabrication and Semantic-Preserving Encryption
D4.2

Due date of deliverable: 31st October 2020

Start date of project: 1st January 2019

Type: Deliverable
WP number: WP4

Responsible Institution: IBM
Editor and editor's address: Michael Vinov (vinov@il.ibm.com)
Partners Contributing: UCL, USTAN, SCCH, ZMC, UCY, FCRB

Approved by:

Version 0.1

Project co-founded by the European Commission within the Horizon H2020 Programme		
Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Release History

Release No.	Date	Author(s)	Release Description/Changes made
V0.1	14/10/20	Michael Vinov (IBM)	Initial draft
V0.2	20/10/20	Thomas Given-Wilson (UCL) Eduard Baranov (UCL)	Added Quality Verification Section
V0.3	25/10/20	Michael Vinov (IBM)	Updated Section 2
V0.4	26/10/20	Michael Vinov (IBM)	Updated Section 4
V1.0	27/10/20	Michael Vinov (IBM)	Review-ready version

SERUMS Consortium

Partner 1	**University of St Andrews**
Contact Person	Name: Juliana Bowles Email: jkfb@st-andrews.ac.uk
Partner 2	**Zuyderland Medisch Centrum**
Contact Person	Name: Cindy Wings Email: c.wings@zuyderland.nl
Partner 3	**Accenture B.V.**
Contact Person	Name: Bram Elshof, Wanting Huang Email: bram.elshof@accenture.com, wanting.huang@accenture.com
Partner 4	**IBM Israel Science & Technology Ltd.**
Contact Person	Name: Michael Vinov Email: vinov@il.ibm.com
Partner 5	**Sopra-Steria**
Contact Person	Name: Andre Vermeulen Email: andreas.vermeulen@soprasteria.com
Partner 6	**Université Catholique de Louvain**
Contact Person	Name: Axel Legay Email: axel.legay@uclouvian.be
Partner 7	**Software Competence Centre Hagenberg**
Contact Person	Name: Michael Rossbory Email: michael.rossbory@scch.at
Partner 8	**University of Cyprus**
Contact Person	Andreas Pitsillides Email: andreas.pitsillides@ucy.ac.cy
Partner 9	**Fundació Clínic per a la Recerca Biomèdica**
Contact Person	Name: Santiago Iriso Email: siriso@clinic.cat
Partner 10	**University of Dundee**
Contact Person	Name: Vladimir Janjic Email: vjanjic001@dundee.ac.uk

Table of Contents

Executive Summary

Securing Medical Data in Smart Patient-Centric Healthcare Systems (SERUMS) is a research project supported by the European Commission (EC) under the Horizon 2020 program. This document is the second deliverable of Work Package 4: "Secure and Privacy-Preserving Data Communication". The leader of this work package is IBM, with involvement from the following partners: UCL, USTAN, SCCH, ZMC, UCY and FCRB. The goal of this work package is to explore and develop techniques and mechanisms to ensure the security and protection of the personal medical data that is shared as part of a coherent smart healthcare system. The objectives of WP4 are to:

- develop advanced data masking and synthetic data fabrication technologies to enable sharing of personal medical data between components of the Smart Health Centre system developed in WP6;
- develop metrics and techniques to verify both the security and the functional properties of the advanced data analytics and the Serums patient-centric Smart Health Centre system;
- explore and develop technology for encrypting information while preserving certain required semantics, in order to enable advanced data analytics while adhering to privacy regulations.

This deliverable entitled "Report on Refined Data Masking, Data Fabrication and Semantic-Preserving Encryption" is the second deliverable of the WP4. It describes enhanced versions of the data masking and data fabrication technologies that are used in the project to enable sharing of personal healthcare data between the project partners and development of the Smart Health Centre. The deliverable report also describes an enhanced version of the technology to verify the quality of fabricated synthetic data on interim versions of the data analytics and authentication tools and an enhanced version of the semantic-preserving data encryption technology to enable and facilitate the application of the Serums advanced data analytics on personal medical data, while fully adhering to necessary privacy regulations.

1 Introduction

1.1 Role of the Deliverable

The aim of this deliverable is to report and describe the design and development of enhanced versions of the data masking, data fabrication, data quality verification and semantic-preserving data encryption technologies. All these technologies are used to explore and develop techniques and mechanisms to ensure the security and protection of the personal medical data that is shared as part of a coherent smart health-care system and to enable and facilitate the application of the Serums advanced data analytics on personal medical data, while fully adhering to necessary privacy regulations.

1.2 Relationship to Other SERUMS Deliverables

Tasks 4.1 and 4.2 of WP4 are closely related to the work done in WP2 – "Smart Patient Record Construction". Masked data and synthetic fabricated data of WP4 is formatted based on the Smart Patient Record format definition developed in WP2. T4.3 of WP4 is closely related to WP2 and WP5. The data technology developed in T4.3 will be applied to verify quality of fabricated medical data and its usage for data analytics and authentication tools of WP2 and WP5. In addition, the output of T4.4 will be used by the WP2 of the project.

1.3 Structure of this Document

This document is structured as follows: *Chapter 2* describes the enhanced version of IBM's Data Fabrication Technology and its usage for fabrication of the project synthetic medical data. *Chapter 3* describes the methodology that is used for verification of the fabricated data quality and its usage for development and testing of the project advanced data analytics and user authentication tools. *Chapter 4* provides a description of the semantic- and privacy-preserving encryption methodology that is used to enable and facilitate the application of the project advanced data analytics on personal medical data, while fully adhering to necessary privacy regulations. *Chapter 5* concludes the deliverable.

2 Data Masking and Synthetic Data Fabrication

2.1 Introduction to DFP

IBM's Data Fabrication Platform (DFP) [1][2] is a web based central platform for generating high-quality data for testing, development, and training. The platform provides a consistent and organizational wide methodology for creating test data. The methodology used is termed "rule guided fabrication".
The primary DFP use case for fabricating synthetic data contains two actors: a user (initiator) and Database/File (participator). This use case includes two sub-use cases: data requirements modelling and data generation. The data requirements use case includes three sub-use cases: resources and structure definitions, constraint rules definitions and fabrication configuration definitions. The data structure for databases (schema, tables, columns, etc.) is automatically imported, however structural hierarchy of data elements (structs, arrays, tables, fields, types) need to be manually defined by the user. The constraint rules are required to construct a model of the data and thus enable creation of meaningful realistic data vales. Input and output resources are standard relational databases (e.g., DB2, Oracle, PostgreSQL, SQLite), standard file formats (e.g., Flat file, XLS, CSV, XML, JSON) and streaming via MQTT protocol.

More detailed description of the DFP tool is available in the D4.1 document "Report on Initial Data Masking, Data Fabrication and Semantic-Preserving Encryption" of the project.

2.2 DFP Enhancements

During the second year of the SERUMS project the Data Fabrication Platform technology has been significantly enhanced to enable improved user experience and fabrication of more complex synthetic data. The major enhancements of the tool include:

1. New GUI to enable improved user experience and typo-free rules modelling,
2. PRB Solver Parallel Edition to considerably improve the tool performance and enable creation of synthetic big data,
3. Support for new operators to support modelling of new fabrication rules.

Below is a mode detailed description of the above improvements.

New GUI

The new updated DFP tool version includes a new implementation of a web-based graphical user interface. The new GUI provides a table-based view of the modelled data fields/columns. A set of fabrication rules is virtually associated with each field/column. Figure 1 shows the main table-based view of the new GUI.

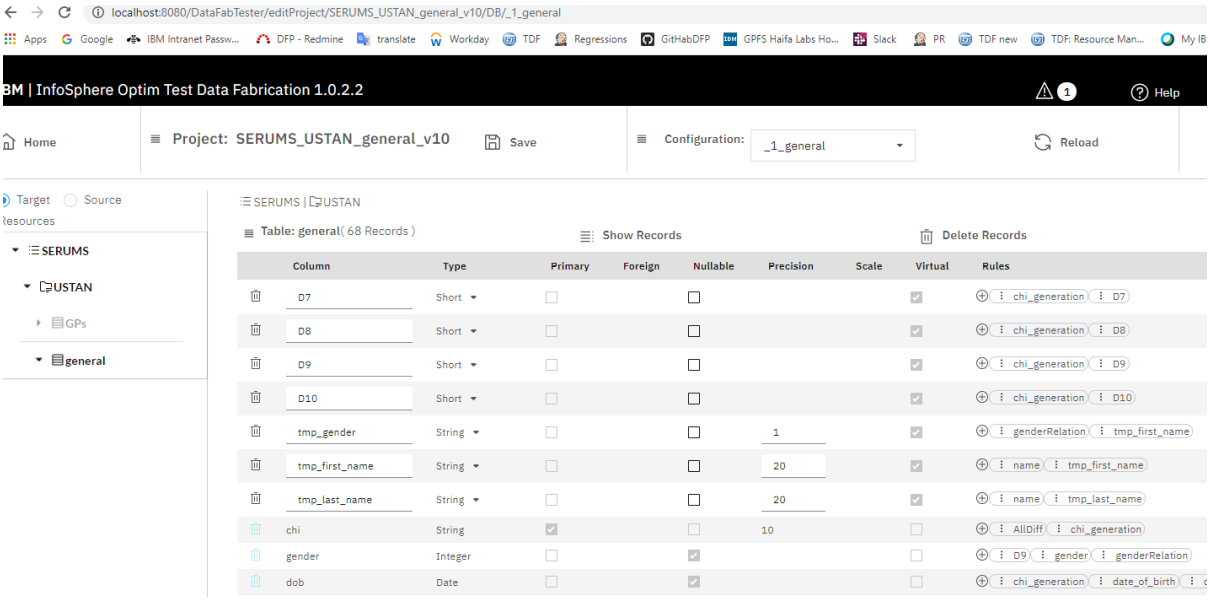

Figure 1 New Data Fabrication Platform GUI

Besides new interface design and look-and-feel, the new GUI includes a new component called Rule Editor. The new component enables graphical form-based definition of the data fabrication rules/constraints instead of textual definition of the constraints using the tool modelling language available in the previous versions of the GUI. The new Rule Editor significantly improves the modelling efficiency and enables faster learning curve for new users. Both graphical and textual representations of a rule that is being edited are represented at the same rule modelling window as shown at the Figure 2 below. The GUI users may choose which rule definition method to choose based on their personal preferences and modelling experience. The graphical and textual rule representations are kept fully synchronized.

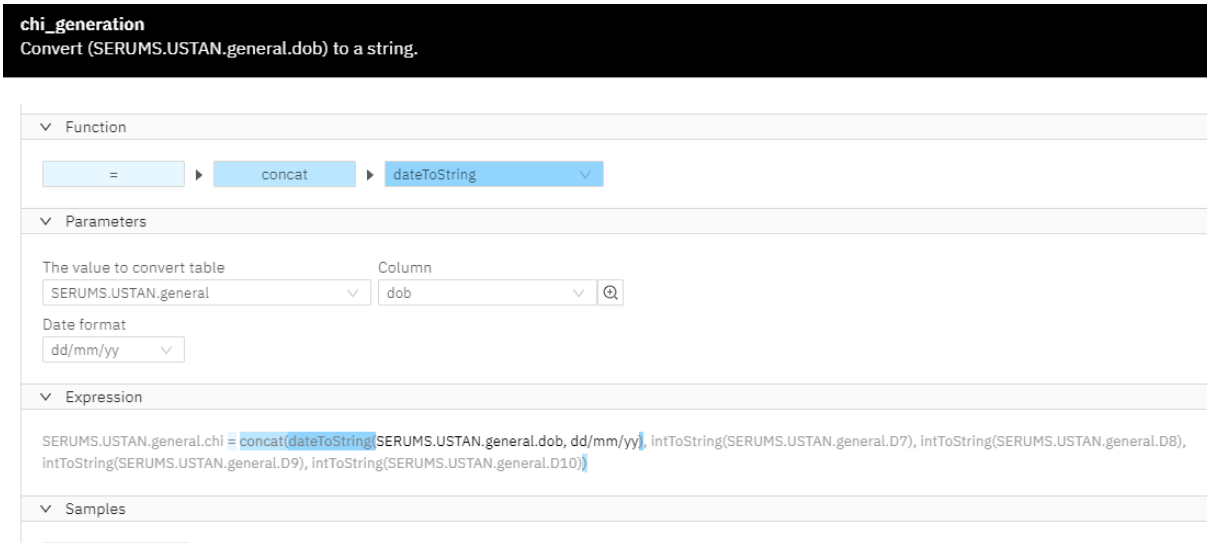

Figure 2 Rule Editor Window

Based on the available meta-data info (e.g. table and column names, column types), the Rule Editor provides a drop-down list of available names for the rule operators and operands (table/column names) as shown on Figure 3 and Figure 4.

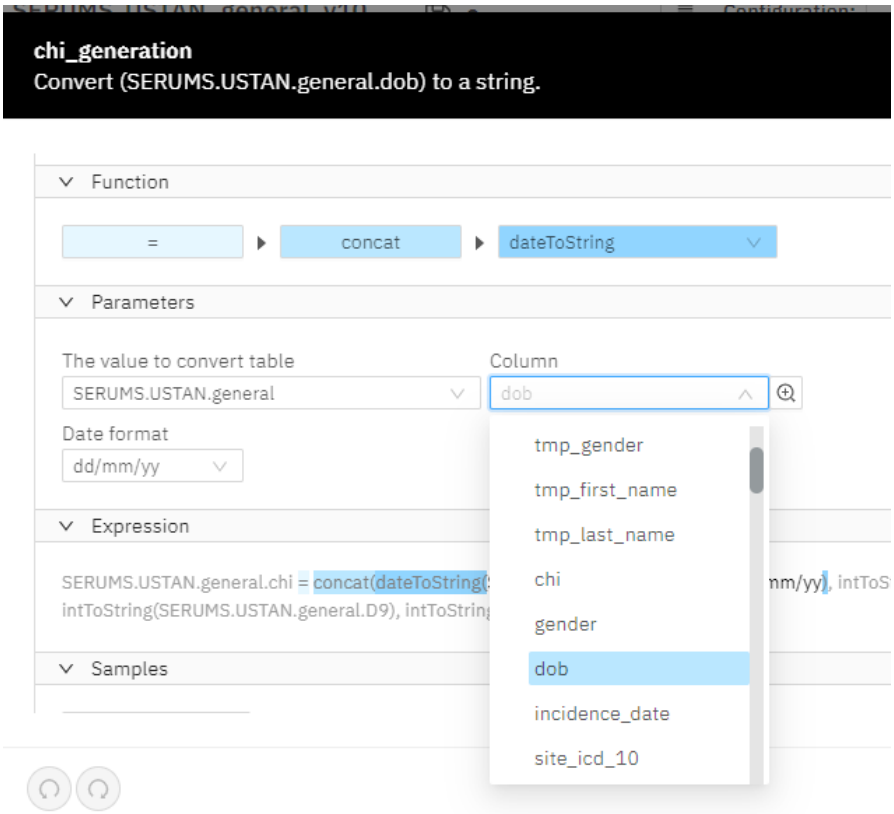

Figure 3 Rules Editing Example (a)

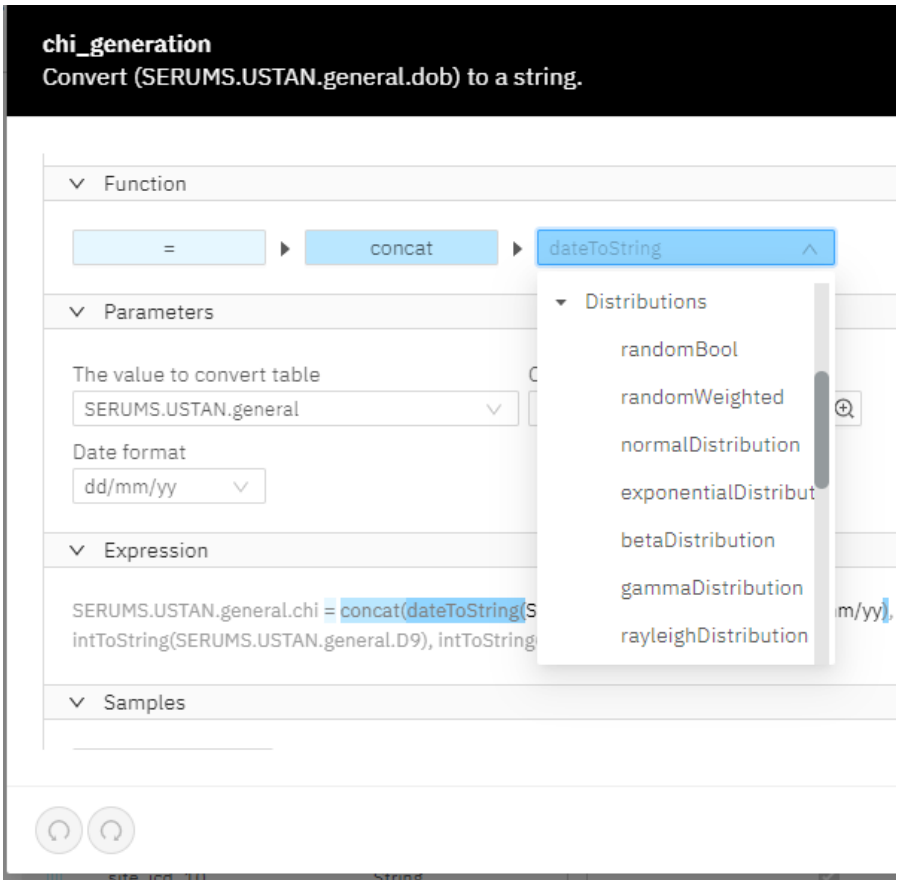

Figure 4 Rules Editing Example (b)

Moreover, when a new rule is added to a table column, all operand fields of the rule get default values based on the column name and its location in the resource hierarchy.

Users can also zoom-in and zoom-out to complex nested rule definitions to emphasis a specific sub-expression of the rule that is currently edited. Figure 5 shows a Rule Editor window for a complex rule. Rule definition and meaning appear on top of a rule definition window right below the rule name.

The new Rule Editor also enables to run a single rule right from the editing window to get its sample results. Each invocation of the "Fabricate sample" option (see Figure 6 below) provides different random results of the rule. This new feature significantly simplifies fabrication rules definition and avoid many definition errors.

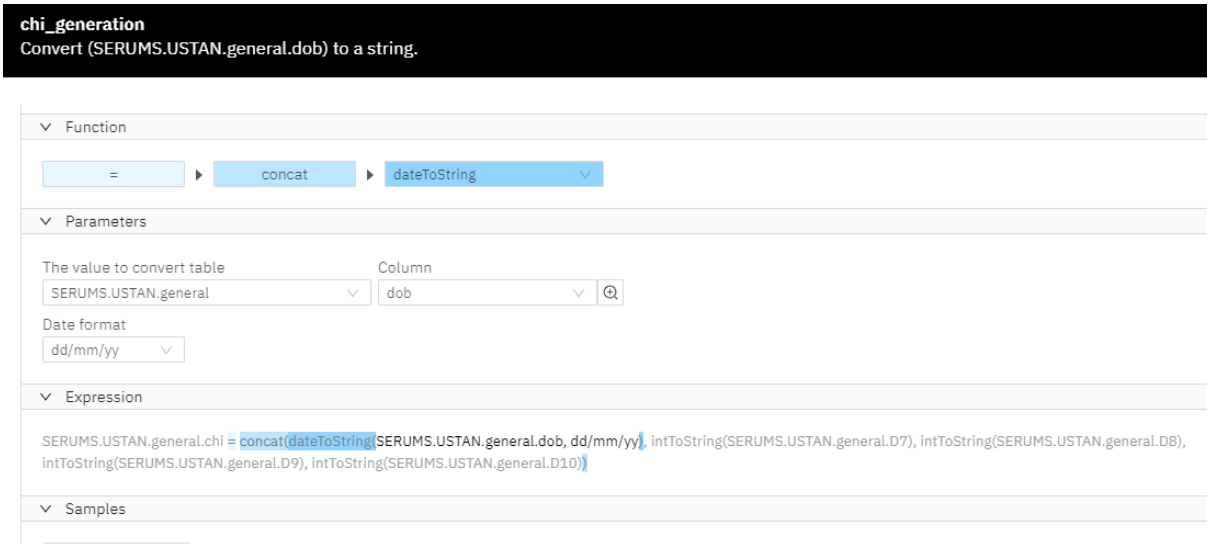

Figure 5 Complex Rule Example

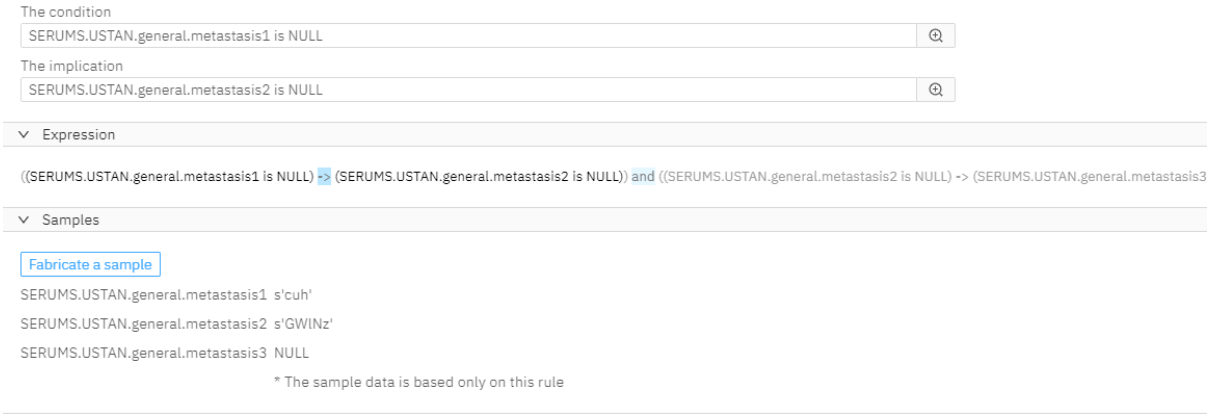

Figure 6 Fabrication of a Sample Rule Result

PRB Solver Parallel Edition

The new Data Fabrication Platform version includes a new enhanced version of the PRB CSP Solver. The new Solver enables to simultaneously solve several CSP problems and thus to concurrently create values for several table rows. This new fabrication mechanism significantly improves the tool performance and enables fabrication of big data.

The new Solver consists of one Manager fabrication process and several Child fabrication processes as shown on Figure 7 below. The responsibilities of the processes are as follows:

- The Manager fabrication process is a singleton. It is responsible for creating all Children fabrication processes and managing them. It is also responsible for managing data dependencies between different CSP problems, communication with

the Fabrication Core, sharing available Data Base connections with the Children processes, and returning CSP solution results to the Core.

- A Child fabrication process is responsible for solving a single CSP problem and fabricating data for a singe table row. When done, it reports a solution to the Manager fabrication process.

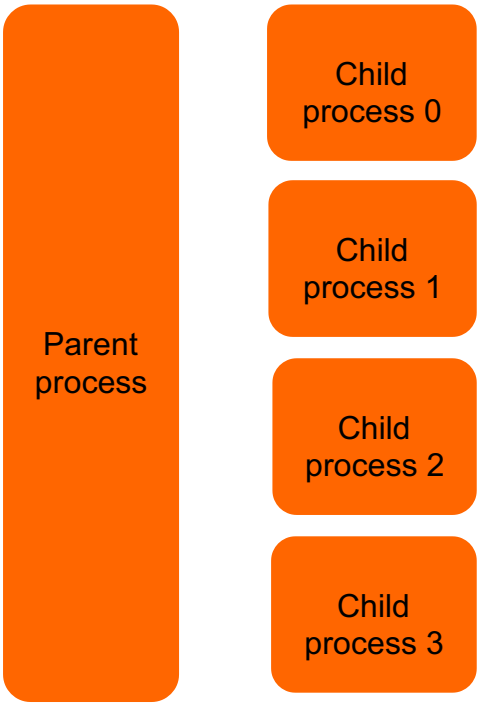

Figure 7 PRB Solver Parallel Edition Overview

PRB Solver – New Operators

The following new operators has been implemented in the new PRB Solver version to better support requirements of fabricating realistic healthcare data.

- `Correlation` operator
  The correlation operator defines a statistical relationship between random variables. It commonly refers to the degree to which a pair of variables are linearly related.
- `Coverage` operator
  Support for optional new `'per'` parameter has been added to the `Coverage` operator. The new parameter defines a subset of the table rows for which the operator is applied.
- Support for several new data distribution operators has been implemented in the Solver – `Exponential, Beta, Gamma, Rayleigh, LogNorm, Weibull, Poisson`.

2.3 Data Masking

Data masking is a well-known method of creating a structurally similar but inauthentic version of an organization's data that can be used for purposes such as software testing, software development and user training. The purpose is to protect the actual personal or sensitive data while having a functional substitute for occasions when the real data is not required. In data masking, the format of data remains the same, only the values are changed. The data may be altered in several ways, including encryption, character shuffling, and character or word substitution.

It was the SERUMS consortium decision that most of the data used for the development and testing of the SERUMS data analytics, user authentication technologies and its patient-centric healthcare system will be synthetic data fabricated by IBM's Data Fabrication Technology described in Section 2.1 above. Moreover, usage of synthetic realistic data solves a known weakness of the data masking approach – its reversibility and a need for the real data access. In case synthetic fabricated data will not be sufficient or "good enough" for the development and testing requirements of the project, we will consider applying the same IBM's DFP tool to produce masked data from the project use-cases real data.

3 Verification of Fabricated Data Quality

Fabricated data is going to be used in other parts of the Serums project for testing and verification. Therefore, the quality of fabricated data is crucial. The data fabrication techniques ensure that the fabricated data is correct. However, there is no guarantees that the fabricated data is "realistic", i.e. it matches data patterns and have hidden intrinsic dependencies of real data.

To address this challenge, we have developed and improved an approach based on ML algorithms as initially reported in deliverable D4.1. The ML algorithms are trained on real and fabricated data sets identifying distinguishing features. High accuracy of the distinguisher indicates the presence of some dependencies intrinsic only to real or to fabricated data. Decision-tree (DT) based ML algorithms are capable to provide feedback on the significant features that allow us to improve the data fabrication. Once the data fabrication rules are updated based on the feedback, the process is be repeated until the ML cannot effectively distinguish between real and fabricated data. The diagram below shows the overall process.

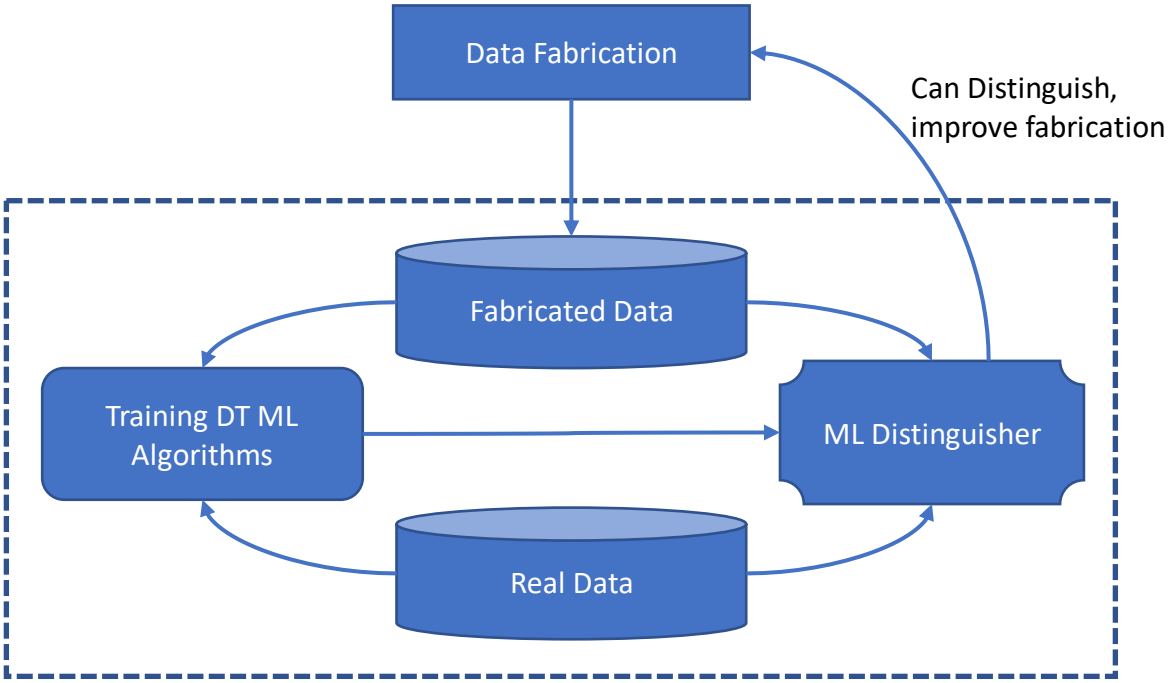

Figure 8 ML-based Data Quality Estimation

We have implemented the approach with the following ML training algorithms: AdaBoost, Gini, Entropy, and Random Forest. These all allow for information about the distinguishing between real and fabricated data to be easily extracted. That is, to understand what features are most significant to distinguish real from fabricated data.

The approach starts by finding a baseline by randomly separating real data into two classes; half "real" and half "fabricated". Since both classes come from the same (real) dataset no

distinguishing features are expected, i.e. accuracy should be close to 50%. This indicates the accuracy that should be achieved when fabricated data is indistinguishable from real data (see below).

The main approach is then an iteration until the baseline is achieved as follows. Fabricated data is generated and the ML is tested on with both real and fabricated data being correctly labelled. These are then treated in the usual manner; split into training and testing sets, and then a distinguisher generated to try and distinguish real from fabricated data. If the distinguisher is effective (i.e. not close to the baseline accuracy), the ML output is examined to determine which data fabrication rules are performing poorly. These rules are then used to generate new fabricated data and the process repeated.

Note that in practice iterations can also include the inclusion of more fields and more data, so results may not always show direct improvement in accuracy.

The approach is being used by USTAN. Running code on their own systems ensures privacy and isolation of patient medical data. USTAN attempted to generate fabricated data for 4 tables: Demographics, Diagnosis, smr01s, smr06s. The baseline results are shown in Table 1 below.

Table 1. Baseline for Data Fabrication

Table	Gini	Entropy	AdaBoost	Random Forest
Demograpics	0.5269	0.516	0.5215	0.56639
Diagnosis	0.5335	0.5240	0.5229	0.5775
smr01s	0.479	0.4820	0.488	0.4486
smr06s	0.4150	0.422	0.458	0.3895

Tables below present the results obtained after 3 iterations of the data fabrication improvement. After each training of the ML distinguisher, the significant features that had the most effect on the distinguisher were examined and were used to update the fabrication rules. Note that at each iteration the number of fabricated fields in each table have been increases that have caused higher achieved accuracies in latter iterations on some of the tables.

Table 2. Results after Iteration 1

Table	Gini	Entropy	AdaBoost	Random Forest
Demograpics	0.8109	0.8098	0.8301	0.8252
Diagnosis	0.9020	0.8997	0.9038	0.8982
smr01s	0.9846	0.9869	0.9887	0.9883
smr06s	0.9313	0.9346	0.9476	0.9356

Table 3. Results after Iteration 2

Table	Gini	Entropy	AdaBoost	Random Forest
Demograpics	0.6226	0.6260	0.5903	0.6254
Diagnosis	0.9323	0.9334	0.9284	0.9229

smr01s	0.7761	0.7820	0.7281	0.7819
smr06s	0.7666	0.7567	0.7084	0.7973

ZMC has also started experimenting with the ML distinguisher on their own data sets. Results and information on their progress will be reported in the next deliverable.

Table 4. Results after Iteration 3

Table	Gini	Entropy	AdaBoost	Random Forest
Demograpics	0.6306	0.6295	0.5958	0.6313
Diagnosis	0.5951	0.5948	0.5826	0.6811
smr01s	0.7164	0.7204	0.7227	0.7432
smr06s	0.8483	0.8333	0.803	0.8618

4 Semantic-preserving Data Encryption

Fully Homomorphic Encryption (FHE), is a cryptographic technique that allows to perform operations on encrypted data that are equivalent to directly manipulating the plaintext. Performing analytics over encrypted data has an intrinsic tradeoff: Accuracy-Security-Performance. Accuracy is measured against the accuracy of comparable plaintext analytics; Security is measured in terms of the ability to deduce information about the private encrypted data; Performance is measured against the time and storage performance of comparable plaintext analytics. For complex tasks, in most cases, at least one of these elements is sacrificed for the others.

All existing FHE schemes have the property that the encrypted data contains noise, and this noise increases when this data is manipulated. When performing long computations, this noise needs to be cleaned every once in a while. This can be done in one of two ways. One way is to interact with the client (the owner of the data who encrypted the data in the first place) as follows: Every time a ciphertext accumulates too much noise it is sent back to the client, where it is decrypted, encrypted again, and returned. Decrypting cleans the noise and encrypting again creates a fresh ciphertext with minimal noise. Another way, completely non-interactive, is to use an operation called Bootstrapping which cleans the noise. This operation is computationally expensive and currently not available in most FHE schemes implementations.

In IBM we set out to perform different analytic task over encrypted data with both interactive and non-interactive methods. We were able to implement inference over encrypted data with encrypted NN models of greater and greater size starting from two layers and moving to quite deep networks. We were also able to classify encrypted data using decision trees, and evaluate SQL queries over encrypted Databases. We were also able to train a given Neural Network (NN) under FHE (Fully Homomorphic Encryption). This NN includes an array of connected components, among which are, a Convolution Layer, a Fully Connected Layer, a Dropout Layer, and more.

All of these example were examined using the Accuracy-Security-Performance tradeoff, with the security and accuracy legs set, optimizing performance to reach a 'usability-threshold'.

5 Conclusions

The aim of this deliverable D4.2 is to report and describe the design and development of refined versions of the project data masking, data fabrication, data quality verification and semantic-preserving data encryption technologies. All these technologies are used to explore and develop techniques and mechanisms to ensure the security and protection of the personal medical data that is shared as part of a coherent smart health-care system and to enable and facilitate the application of the Serums advanced data analytics on personal medical data, while fully adhering to necessary privacy regulations.

First, the document describes IBM's Data Fabrication Technology and the tool enhancements implemented during the second project year to improve user experience and enable modeling and fabrication of more complex medical data for the project use-cases. Further, the document describes an extended version of our approach to estimating the quality of fabricated synthetic data to ensure that all data analytics and user authentication tools developed by Serums consortium will be fully applicable for real medical data in the future. The document also describes our approach to Fully Homomorphic Encryption to be able to apply the advanced data analytics and machine-learning algorithms for analyzing encrypted personal data.

This document is the second deliverable of Work Package 4: "Secure and Privacy-Preserving Data Communication". Deliverable D4.3 will describe more advanced final versions of all the above technologies and their application for the development of the Serums smart patient-centric healthcare system.

References

[1] “Create high-quality test data while minimizing the risks of using sensitive production data.” *IBM InfoSphere Optim Test Data Fabrication*, IBM, 2017, https://www.ibm.com/il-en/marketplace/infosphere-optim-test-data-fabrication.

[2] “Test Data Fabrication.” *Security and Data Fabrication*, IBM Research, 2011, https://www.research.ibm.com/haifa/dept/vst/eqt_tdf.shtml.

[3] “Constraint Satisfaction.” IBM Haifa Research, IBM, 2002, https://www.research.ibm.com/haifa/dept/vst/csp.shtml.

[4] Y. Richter, Y. Naveh, D. L. Gresh, and D. P. Connors (2007), "Optimatch: Applying Constraint Programming to Workforce Management of Highly-skilled Employees", International Journal of Services Operations and Informatics (IJSOI), Vol 3, No. 3/4, pp. 258 - 270.

[5] Y. Naveh, Y. Richter, Y. Altshuler, D. Gresh, and D. Connors (2007), "Workforce Optimization: Identification and Assignment of Professional Workers Using Constraint Programming", IBM J. R&D.

[6] Y. Naveh, M. Rimon, I. Jaeger, Y. Katz, M. Vinov, E. Marcus, and G. Shurek (2006), "Constraint-Based Random Stimuli Generation for Hardware Verification", AI magazine Vol 28 Number 3.

[7] E. Bin, R. Emek, G. Shurek, and A. Ziv (2002). "Using a constraint satisfaction formulation and solution techniques for random test program generation", IBM Systems Journal, 2002.

[8] Merav Aharoni, Odellia Boni, Ari Freund, Lidor Goren, Wesam Ibraheem, Tamir Segev (2015), "Rectangle Placement for VLSI Testing", CPAIOR 2015: 18-30

[9] O. Boni, F. Fournier, N. Mashkif, Y. Naveh, A. Sela, U. Shani, Z. Lando, A. Modai (2012) "Applying Constraint Programming to Incorporate Engineering Methodologies into the Design Process of Complex Systems" Proceedings of the Twenty-Fourth Conference on Innovative Applications of Artificial Intelligence, Toronto, Ontario, Canada. AAAI 2012.

[10] Y. Ben-Haim, A. Ivrii, O. Margalit and A. Matsliah (2012) "Perfect Hashing and CNF Encodings of Cardinality Constraints", SAT 2012, Trento, Italy.

[11] E. Bin, O. Biran, O. Boni, E. Hadad, E. K. Kolodner, Y. Moatti, D. H. Lorenz (2011), "Guaranteeing High Availability Goals for Virtual Machine Placement", ICDCS 2011.

[12] Jeonghee Shin, John A Darringer, Guojie Luo, Merav Aharoni, Alexey Y Lvov, G Nam, Michael B Healy (2011), "Floorplanning challenges in early chip planning", SOCC Conference, 2011 IEEE International, pp. 388—393

[13] Y. Naveh (2010). "The Big Deal, Applying Constraint Satisfaction Technologies Where it Makes the Difference". Proceedings of the Thirteenth International Conference on Theory and Applications of Satisfiability Testing (SAT'10).

[14] S. Asaf, H. Eran, Y. Richter, D. P Connors, D. L. Gresh, J. Ortega, M. J. Mcinnis (2010). "Applying Constraint Programming to Identification and Assignment of Service Professionals". Accepted for presentation in The 16th International Conference on Principles and Practice of Constraint Programming (CP2010). The paper received the Best Application Paper Award.

[15] B. Dubrov, H. Eran, A. Freund, E. F. Mark, S. Ramji, and T. A. Schell, (2009). "Pin Assignment Using Stochastic Local Search Constraint Programming" in Proceedings of the 15th International Conference on Priniciples and Practice of Constraint Programming (CP'09), Edited by Ian P. Gent, pp 35-49.

[16] Y. Richter, Y. Naveh, D. L. Gresh, and D. P. Connors (2007), "Optimatch: Applying Constraint Programming to Workforce Management of Highly-skilled Employees", IEEE/INFORMS International Conference on Service Operations and Logistics, and Informatics (SOLI), Philadelphia, pp. 173-178.

[17] S. Sabato and Y. Naveh (2007), "Preprocessing Expression-based Constraint Satisfaction Problems for Stochastic Local Search", Proceedings of The Fourth International Conference on Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems (CP-AI-OR).

[18] Y. Naveh, M. Rimon, I. Jaeger, Y. Katz, M. Vinov, E. Marcus, and G. Shurek (2006), "Constraint-Based Random Stimuli Generation for Hardware Verification", IAAI 2006.

[19] Y. Richter, A. Freund, and Y. Naveh (2006), "Generalizing AllDifferent: The SomeDifferent constraint", Proceedings of the 12 International Conference on Principles and Practice of Constraint Programming - CP 2006, Lecture Notes in Computer Science, Volume 4204, pages 468-483.

[20] Y. Naveh and R. Emek (2006). "Random stimuli generation for functional hardware verification as a CP application - a demo", IAAI 2006.

[21] Y. Naveh (2005). "Stochastic solver for constraint satisfaction problems with learning of high-level characteristics of the problem topography" CP 2005

[22] F. Geller and M. Veksler (2005), "Assumption-based pruning in conditional CSP", in van Beek, P., ed., CP, "Principles and Practice of Constraint Programming - CP 2005" of Lecture Notes in Computer Science (3709), 241-255 Springer.

[23] R. Dechter, K. Kask, E. Bin, and R. Emek (2002). "Generating random solutions for constraint satisfaction problems", AAAI 2002.

[24] D. Lewin, L. Fournier, M. Levinger, E. Roytman, G. Shurek (1995). "Constraint Satisfaction for Test Program Generation", Internat. Phoenix Conf. on Computers and Communications, March 1995.

[25] Juvekar, C., Vaikuntanathan, V., Chandrakasan, A.: Gazelle: A low latency framework for secure neural network inference. arXiv preprint arXiv:1801.05507 (2018).