Synthetic data for public good

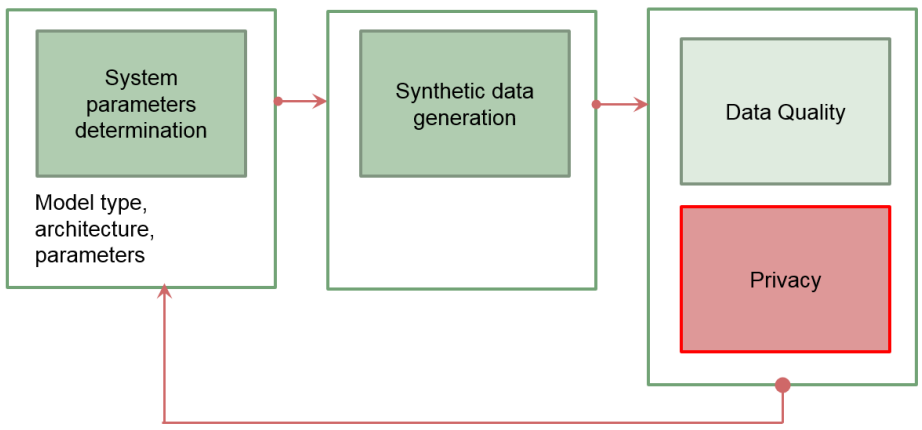
C. Joshi*, I. Kaloskampis*, D. Pugh*, A. Noyvirt, L. Benedikt, L. Nolan

(*) these authors contributed equally to this work

Introduction

We develop a system which generates synthetic data to replace real data for processing and analysis. This is useful when the real data is sensitive (e.g. microdata, medical records, defence data).

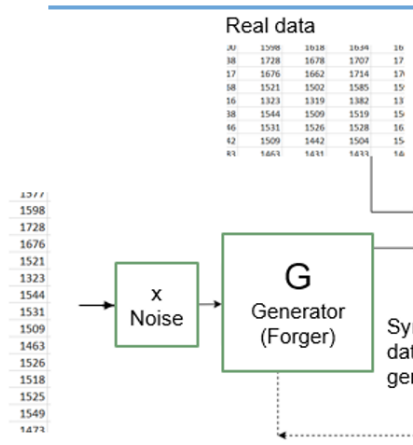
- Generative Adversarial Networks (GAN)
- (Variational) Auto-encoder
- Auto-regressive models
- Conventional generators (SMOTE etc.)



Our system utilises robust statistical, machine learning and state of the art deep learning techniques.

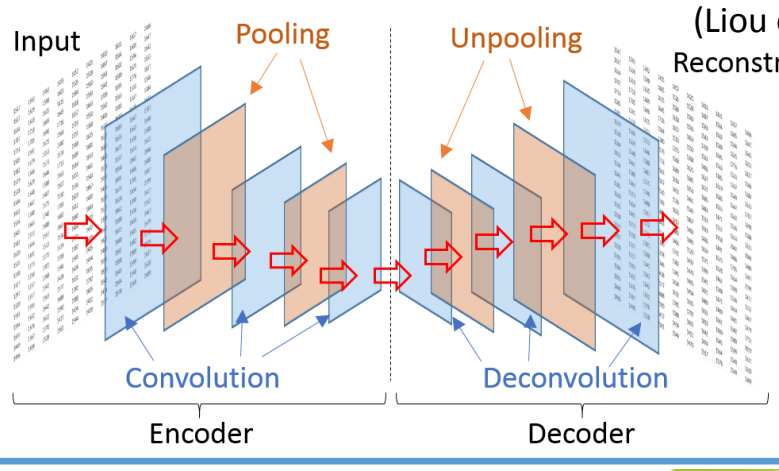
Safer, easier and faster sharing of sensitive data between research communities and ONS.

Models

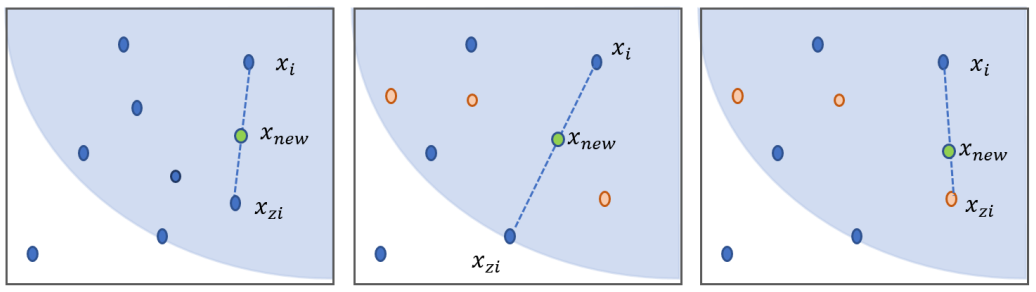


Generative Adversarial Networks (Goodfellow et al. 2014)

Autoencoders (Liou et al. 2008)



SMOTE (Chawla et al. 2002)

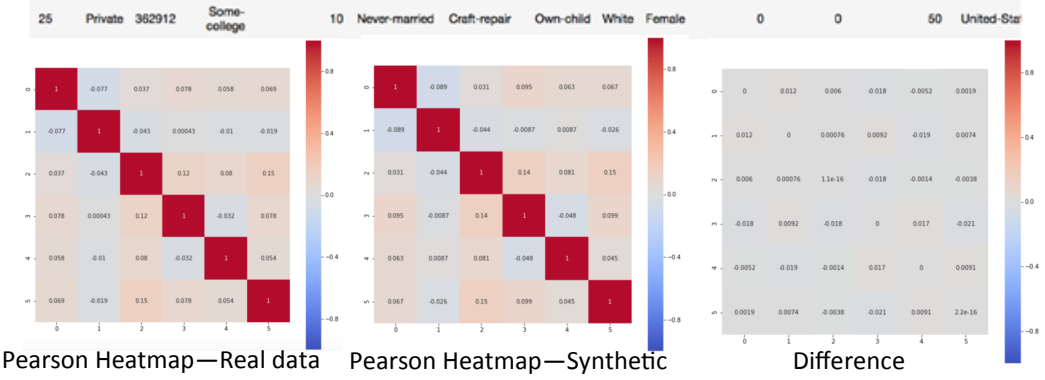


SMOTE *Borderline 2* *Borderline 1*

Results

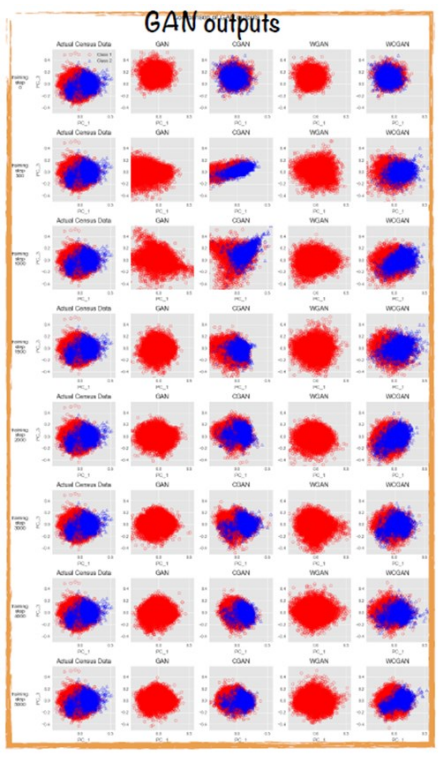
Adult Census Dataset (UCI Repository)

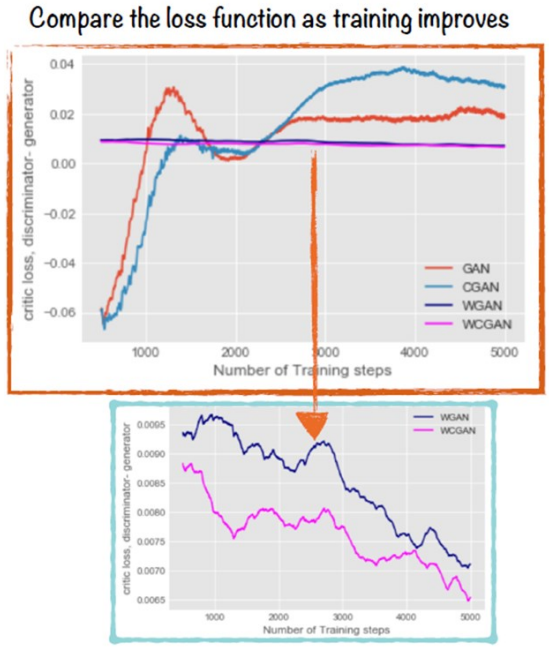
age	workclass	fnlwgt	education	education.num	marital.status	occupation	relationship	race	sex	capital.gain	capital.loss	hours.per.week	native.coun
59	Private	108496	Some-college	10	Divorced	Craft-repair	Not-in-family	White	Male	2907	0	40	United-Sta
40	Private	145160	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	40	United-Sta
46	Private	165953	Bachelors	13	Married-civ-spouse	Adm-clerical	Husband	White	Male	0	0	45	United-Sta
39	Private	216552	Masters	14	Married-civ-spouse	Adm-clerical	Husband	White	Male	0	0	40	United-Sta
25	Private	362912	Some-college	10	Never-married	Craft-repair	Own-child	White	Female	0	0	50	United-Sta



Pearson Heatmap—Real data | Pearson Heatmap—Synthetic | Difference

Model	Dataset	Application	Testing for...	Result
GAN	OCS Adult Census (numerical variables)	Synthesis	Data Quality – Pearson Correlation	0.085802
WGAN				0.064000
SMOTE				0.011192
SMOTE B1			absolute mean difference between real & synthetic correlation matrices	0.011192
SMOTE B2				0.009105
ADASYN				0.012401
Autoencoder				0.008293
Variational Autoencoder				0.211314





WGAN, WCGAN further training seems to improve the results in contrast to GAN and CGAN