Private synthetic data with GANs, autoencoders & SMOTE

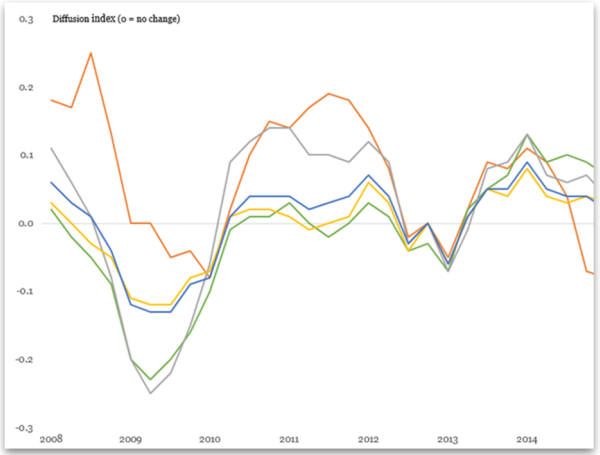
I. Kaloskampis, D. Pugh, C. Joshi, A. Eidukas, G. Clews, J. Rees, J. Latham, L. Nolan.

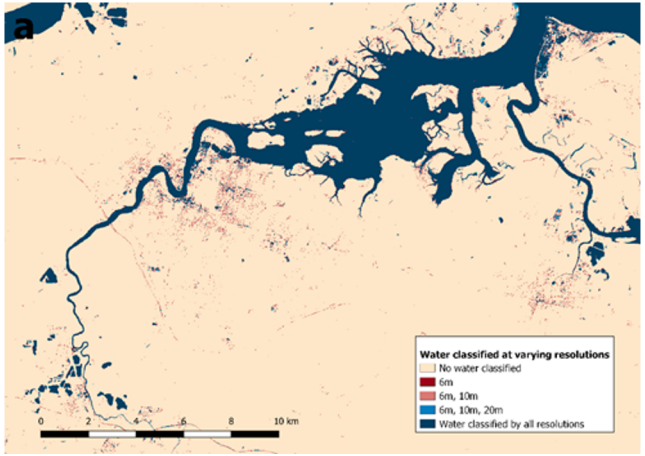
Applied Machine Learning Days
January 2020

Data Science Campus

- Utilise state-of-the art machine & deep learning
  - Solve complex data science problems within ONS, Government
  - Experienced data scientists (academia, industry)
- Training & mentoring

Data Science Campus Projects





Faster economic indicators

Economics, time series

Water project

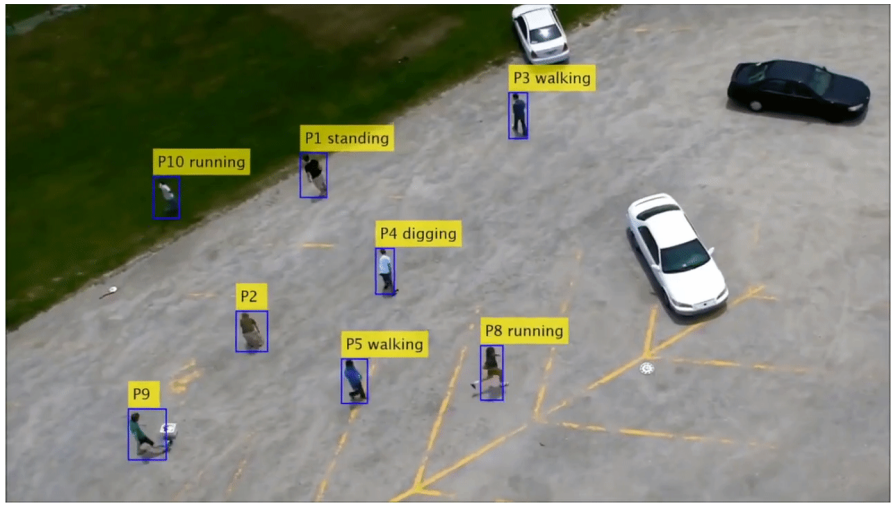
GIS, computer vision, deep learning

Pygrams

NLP, time series

Ioannis

Senior Data Scientist, Data Science Campus



- Synthetic data, Privacy, Economics
- Data confidentiality Champion
- PhD in activity recognition with machine learning
- Post-doc in anomaly detection, in collaboration with MOD
- Individual identification of white sharks using CV/PR (2008):

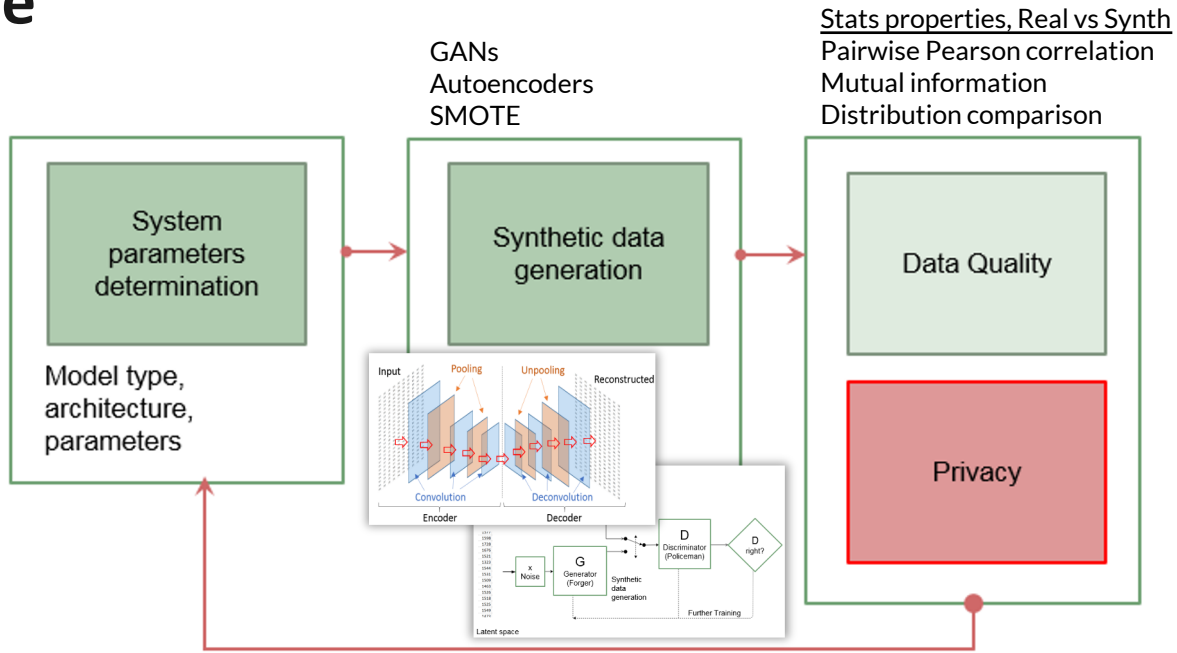
Synthetic data project overview

Image credit: Christies.com

- Problems with using real data
  - Data sensitivity, availability, missing data
- Solution: Use synthetic data
- Impact
  - Easier, faster and safer to share sensitive data
  - Data sharing/privacy for medical, defence apps etc.
  - Online tool for synthetic data generation in the future

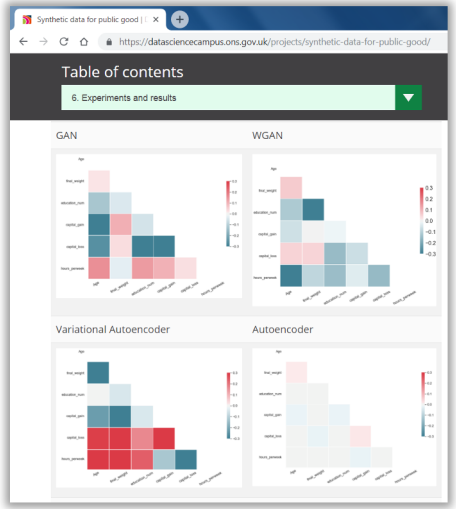
Processing Pipeline

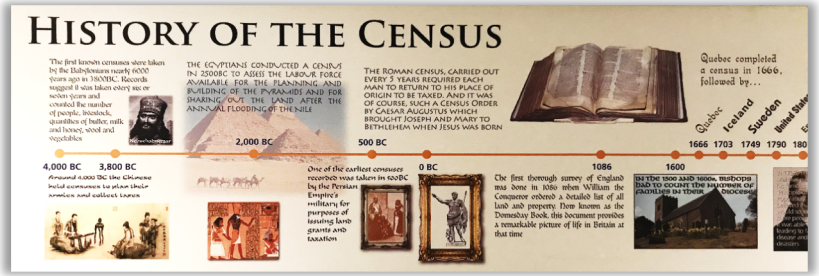
- State-of-the-art machine learning algorithms
- Latest ideas in Privacy
- Working closely with Academia



Outputs so far...

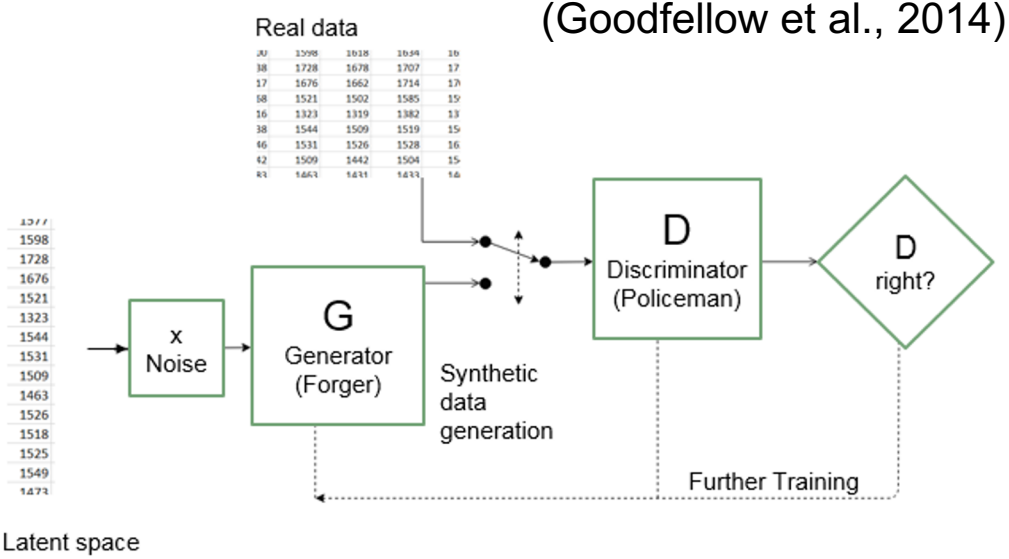
- 4 reports, Conference presentations
- Synthetic data for Census
- BEIS projects
  - NEED & National Household Model
- Mentoring projects
  - Annual Business Survey (MDR)
  - Differential Privacy (Cardiff Uni, MDR)





Generative Adversarial Networks

(Goodfellow et al., 2014)



	Advantages	Disadvantages
GAN	Generated data are not sampled from the real data, which can give privacy properties.	Difficult to tune, can take a long time to train.
Autoencoder	Good results for standard hyperparameter values.	Difficult to tune, especially variational autoencoders.
SMOTE	Standard package, good results without the need for parameter optimisation.	Generated data are sampled from the real data.

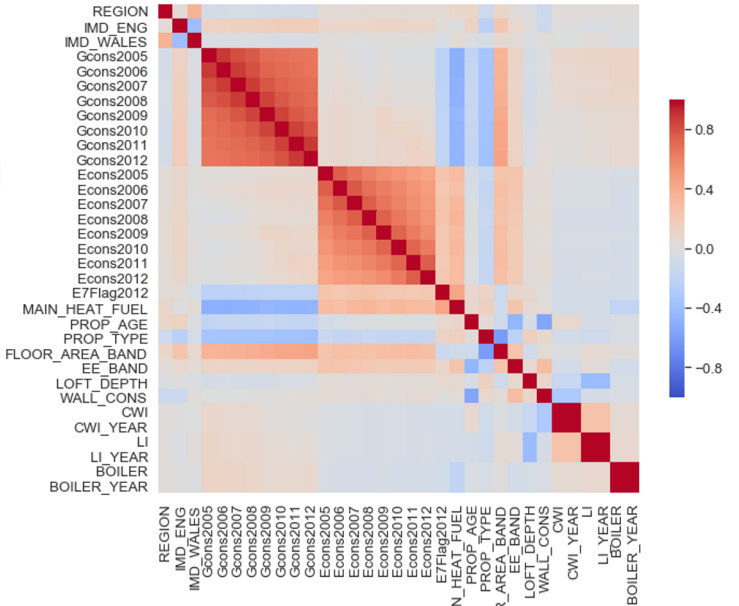
BEIS NEED dataset

Department for Business, Energy & Industrial Strategy

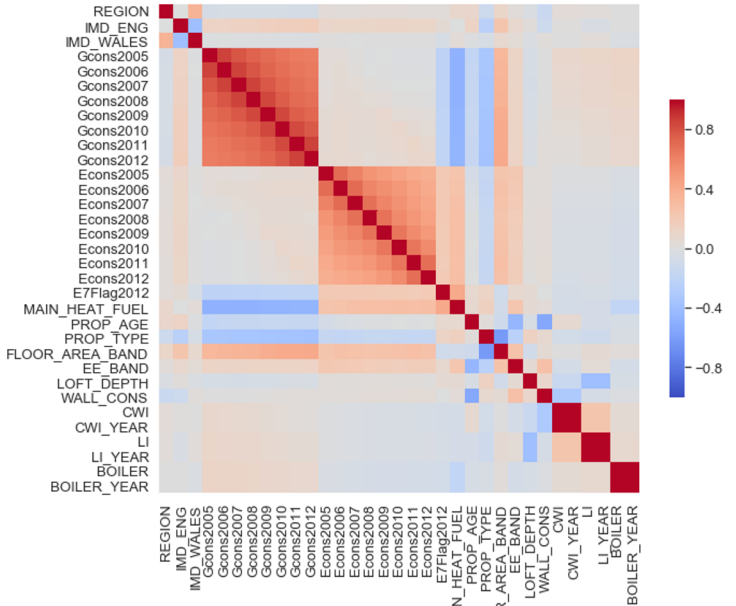
- National Energy Efficiency Data-Framework
- 33 variables, numerical & categorical
- Temporal dependencies (time series)
- Method utilised: autoencoders

BEIS NEED dataset – Pearson heatmaps

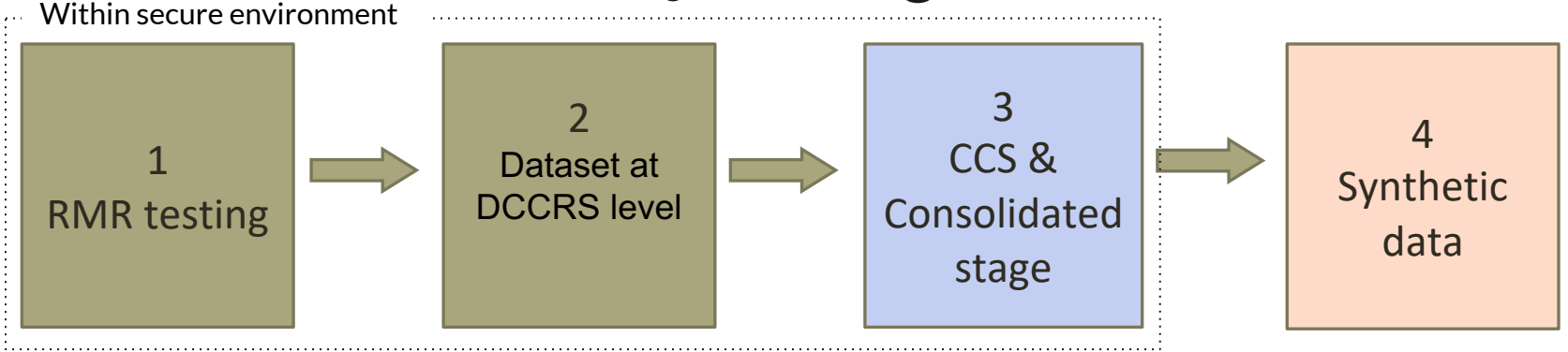
Real



Synthetic



Census - Project stages



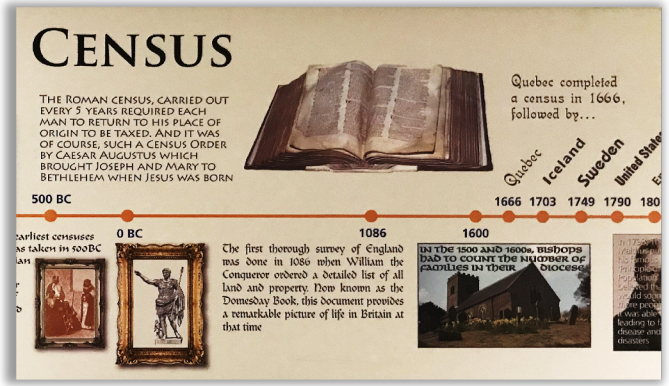
- Use 2011 Census
- Convert structure to match 2021 Census
- Persons & Households datasets
- Resolve Multiple Responses (RMR)

- Data Capture & Coding Requirements Specification (DCCRS)
- Sets out the rules for coding questions

- Census Coverage Survey (CCS): 1% sample datafile to accompany the Census
- Consolidated stage: post merge of online & paper data

- Maintain statistical properties of original data
- Using GANs, SMOTE etc.

Challenges



- Interfacing between teams
  - Specifications, evolving project, rapid prototyping, human resources
- Technical
  - ONS Secure Environment
  - Dataset linking, closely match distributions
- Timeline!
  - Stage 1: 2 weeks, Stage 2: 4 weeks

Next steps

- Privacy evaluation (with Cardiff University)
- Disclosure control (with ONS MDR)
  - Privacy/utility trade-off
- Software package
- From research to policy

https://datasciencecampus.ons.gov.uk/projects/synthetic-data-for-public-good/
https://datasciencecampus.github.io/balancing-data-with-smote/