ACTUARIAL COLLOQUIUM PARIS 2020

Generative Adversarial Networks for Actuarial Use

Kwanda Ngwenduna, MSc Data Science Candidate, Wits University, South Africa

Rendani Mbuvha, Lecturer & Machine Learning Researcher, Wits University, South Africa

France, Paris, May 13th, 2020

About the speaker

ACTUARIAL COLLOQUIUM PARIS 2020

Kwanda Ngwenduna

- MSc Data Science Candidate, Wits University, South Africa
- Member of Actuarial Society of South Africa Council
- Past Treasurer for the Association of South African Black Actuarial Professionals
- Double Honours: Actuarial Science & Mathematical Statistics with Distinction
- 15 actuarial board exams with SA/F200 remaining

Work Experience

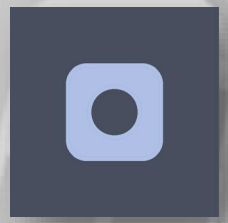
- Munich Re of Africa: predictive analytics, experience analysis, pricing & reinsurance modelling
- Vodacom: advanced analytics, time series projections & machine learning
- Currently: ABSA Life – bancassurance pricing & experience analysis

About the speaker

ACTUARIAL COLLOQUIUM PARIS 2020

Rendani Mbuvha

- Lecturer and Machine Learning Researcher, Wits University, South Africa
- FASSA, CERA
- MSc Machine Learning, KTH Royal Institute of Technology, Stockholm (Sweden)
- BSc Actuarial Science and Statistics and BSc Honours in Actuarial Science, University of Cape Town



Work Experience

- Discovery Health, Actuarial Analyst
- Absa Life, Product Development Actuary

AIMS & OBJECTIVES

- **Explicit models**
  - Approximate
  - Tractable
- **Implicit models**
  - Direct
  - Markov

Overview of generative models

- Theory
- GAN variants
- Challenges
- Traditional domain areas

Generative Adversarial Networks

Example GAN Application on Fraud data

- Predictive modelling in Python
  - Synthetic methods vs. GANs
- Open challenges
- Future improvements
- GANs for actuarial use

Applications of GANs

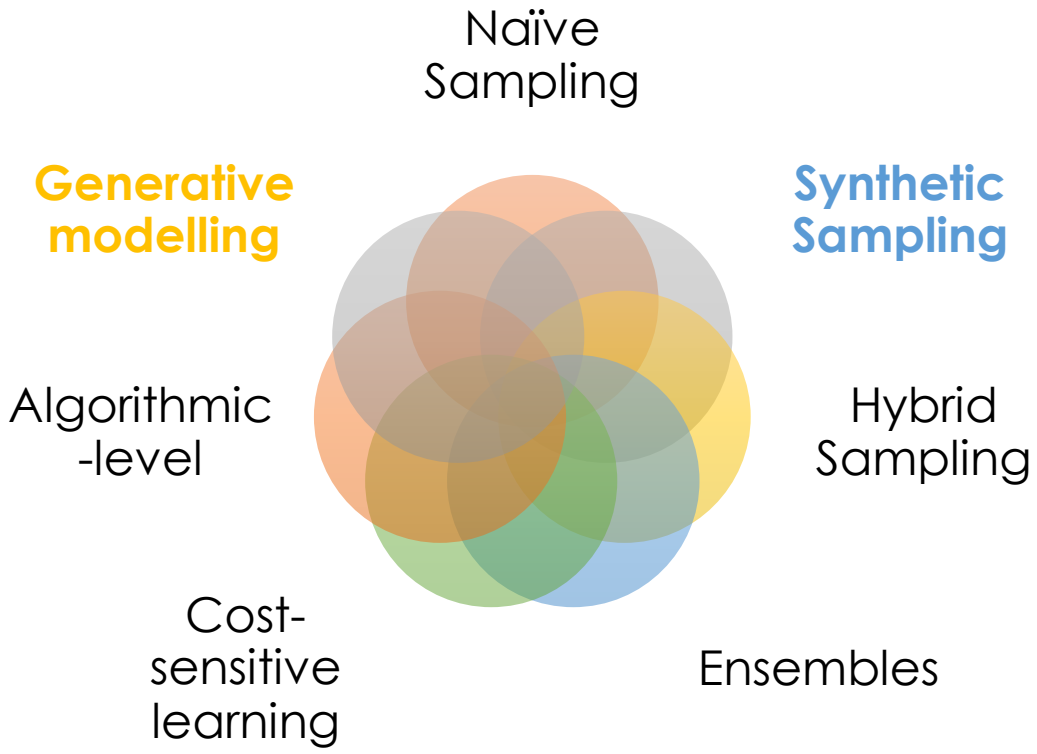
- Computer Vision
- Semi-supervised learning
- Data manipulation
- Domain adaptation
- Sequences, imputation, etc.

CHALLENGES

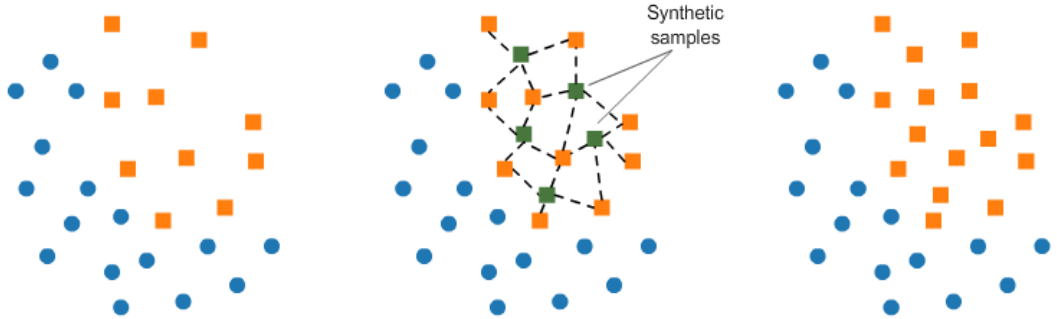
Class imbalance, Limited data sets, Missing data, Domain shift, New territories, Joint distributions.

Data collection costs, absolute rarity, nature of the data space etc.

SOLUTIONS TO TACKLE IMBALANCE, LIMITED DATA SETS & MISSING DATA



- Synthetic sampling is the default method



- But generative models are becoming more popular
- This research is concerned with a popular implicit generative model i.e. Generative Adversarial Network (GAN)

Picture: DataCamp

GENERATIVE MODELS

Given samples from a data density, x ~ q(x), we would like to learn a generative model with density p that matches the data density q. Then we can sample from p.

Why Generative Models

- We know of discriminative models
  - Given an **X**, predict a label **Y**
  - Estimates **P(Y | X)**
- Discriminative models have several disadvantages
  - Can't model **P(X)**, i.e. the probability of seeing a certain point
  - Thus, can't sample from **P(X)**, i.e. **can't generate new data**
- Generative models (in general) cope with all of above
  - Can model **P(X)**
  - Can generate new data

GENERATIVE MODELS

- A generative model is any type of model that captures the training data distribution
- From this, we can sample new examples
- There are 2 types of density estimation approaches in use
- We are interested in GANs as these are better
  - *Better sample quality*
  - *No distributional assumptions*
  - *Sampling in parallel*
  - *No Markov chains*
  - *No variational bounds*
  - *Robust to over-fitting*
  - *No maximum likelihood estimation*
  - *Wider applications*

Explicit density	Approximate	Variational Inference	Variational Autoencoder
		Markov chain	Boltzmann Machine
	Tractable	Fully Visible Belief Nets	NADE MADE PixelCNN/RNN
		Change of variable models	Non-linear independent component analysis
Implicit density	Direct	**GAN**	**Vanilla GAN** **GAN variants**
		Generative Moment Matching Network	
	Markov	Generative Stochastic Networks	

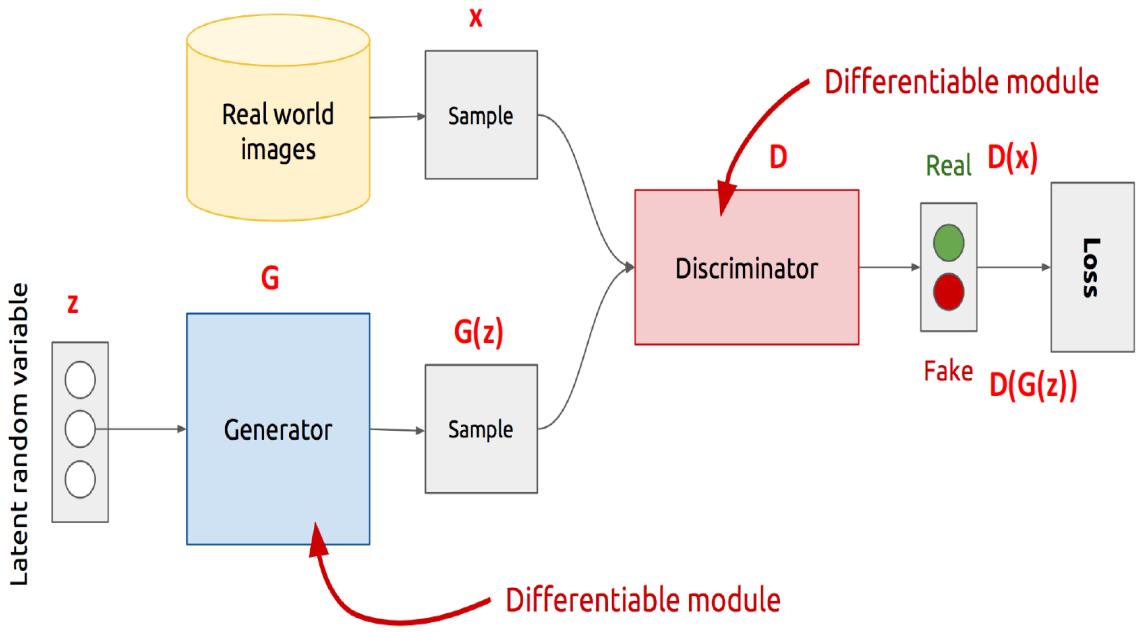
GENERATIVE ADVERSARIAL NETWORKS

"…the most interesting idea in the last 10 years in machine learning..."

- Facebook's AI research director Yann LeCun

Generative Adversarial Network

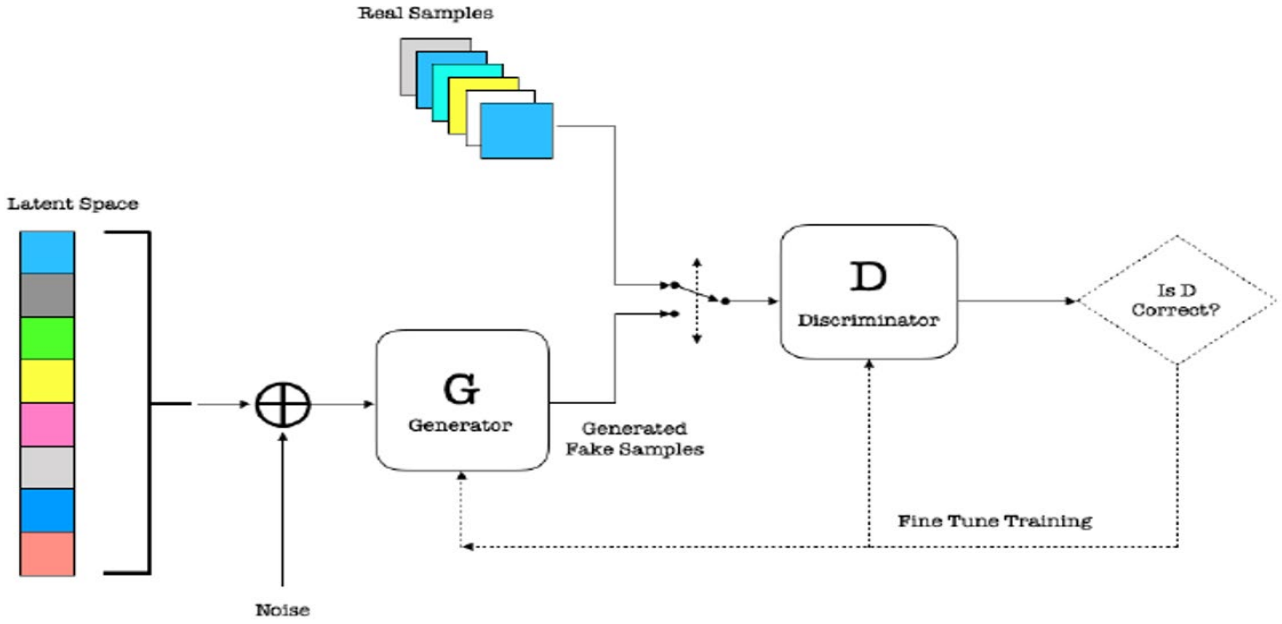
- Generative
  - Learn a generative model
- Adversarial
  - Trained in an adversarial setting
- Network
  - Use Deep Neural Networks



https://www.slideshare.net/xavigiro/deep-learning-for-computer-vision-generative-models-and-adversarial-training-upc-2016

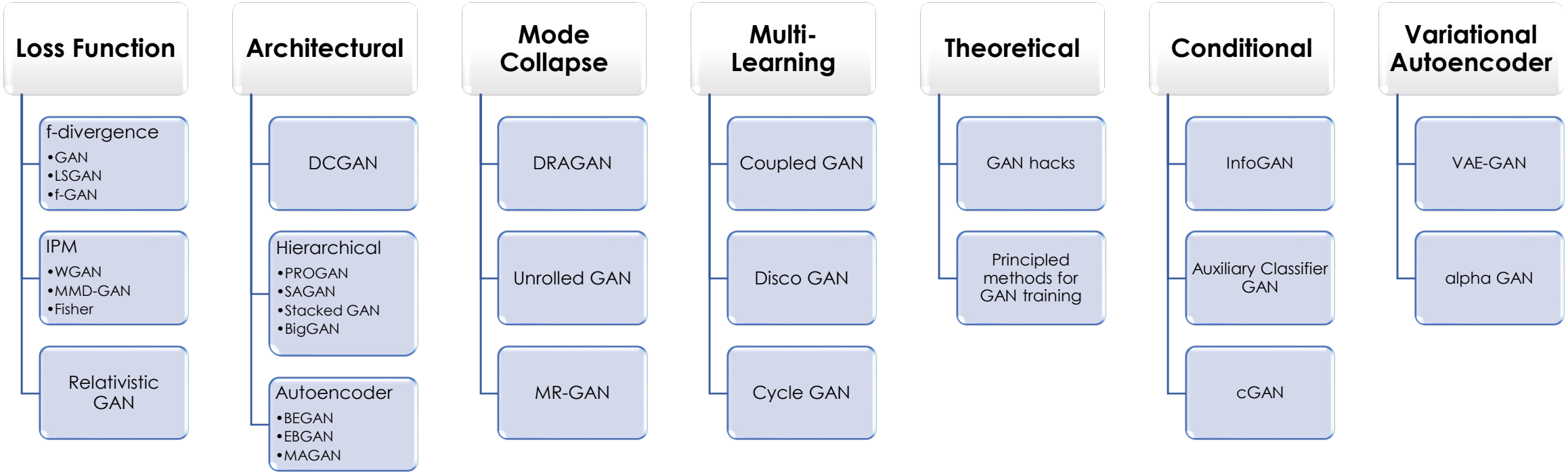
GENERATIVE ADVERSARIAL NETWORKS

- Generator (G) tries to fool the Discriminator (D)
- G generates new instances from a latent space Z
- D is a classifier i.e. real vs. fake
  - D tries not to be fooled
  - D evaluates instances for authenticity
- Models are trained simultaneously
  - As G gets better, D has a more challenging task
  - As D gets better, G has a more challenging task
- Issues?
  - Mode collapse, convergence, vanishing gradient, unstable, etc.
- Wide interest and extensive applications since 2014



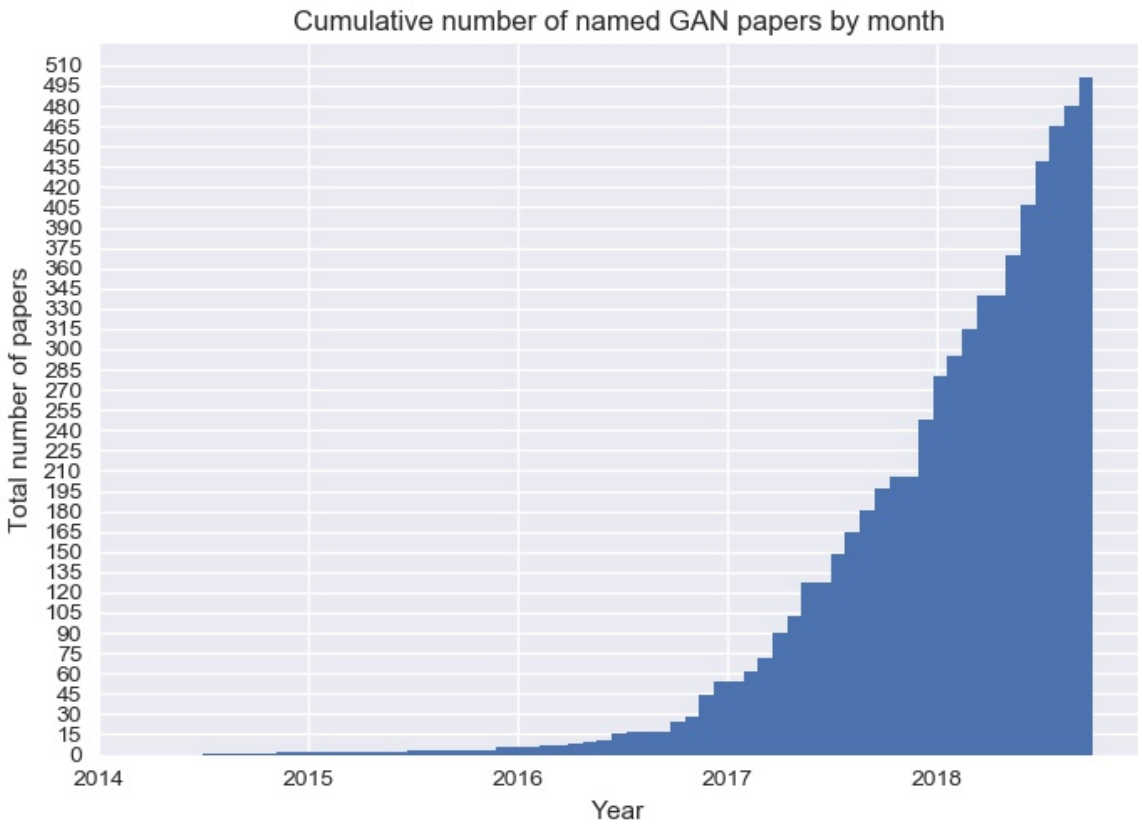
$$\min_{\theta_g} \max_{\theta_d} \mathbb{E}_{X \sim p_{data}(x)} \left[\log D_{\theta_d}(x) \right] + \mathbb{E}_{Z \sim p_z(z)} \left[\log(1 - D_{\theta_d}(G_{\theta_g}(z))) \right]$$

GAN TAXONOMY

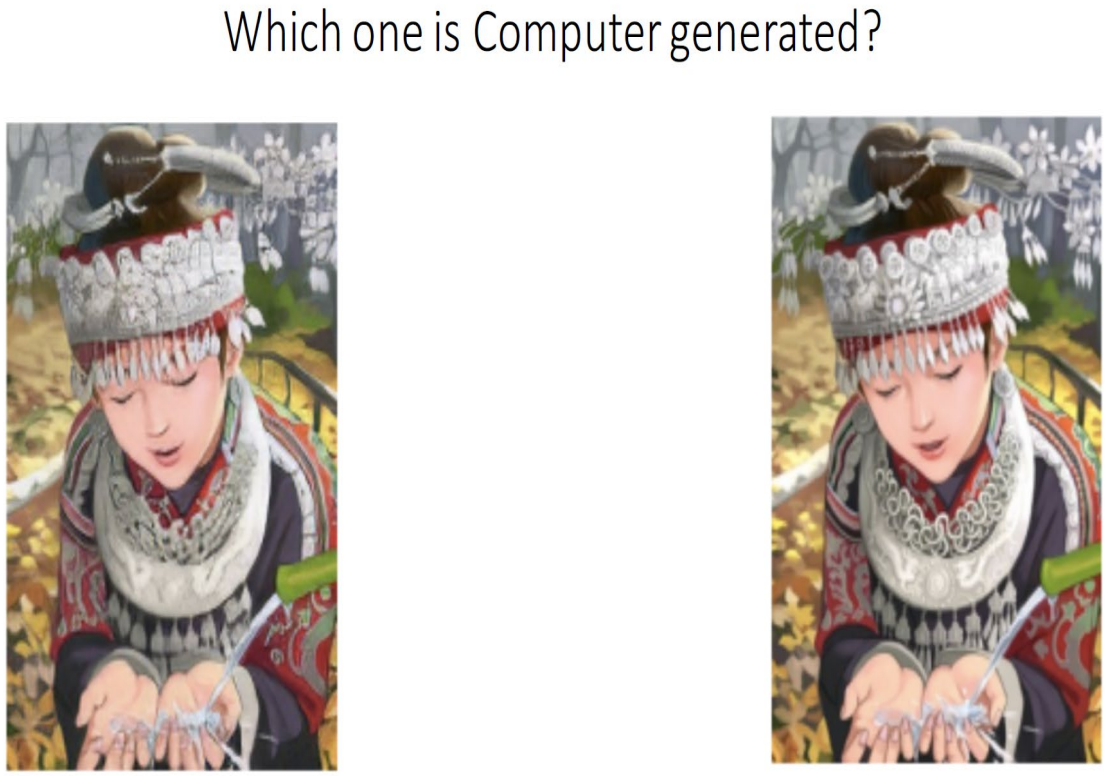


GANs are notoriously difficult to train, unstable convergence and mode collapse problem. Lots of variants improve these challenges. Wasserstein GAN (WGAN) has become the most well studied/used.

GAN POPULARITY & MAGIC



https://github.com/hindupuravinash/the-gan-zoo



Ledig, Christian, et al. "Photo-realistic single image super-resolution using a generative adversarial network." *arXiv preprint arXiv:1609.04802* (2016).

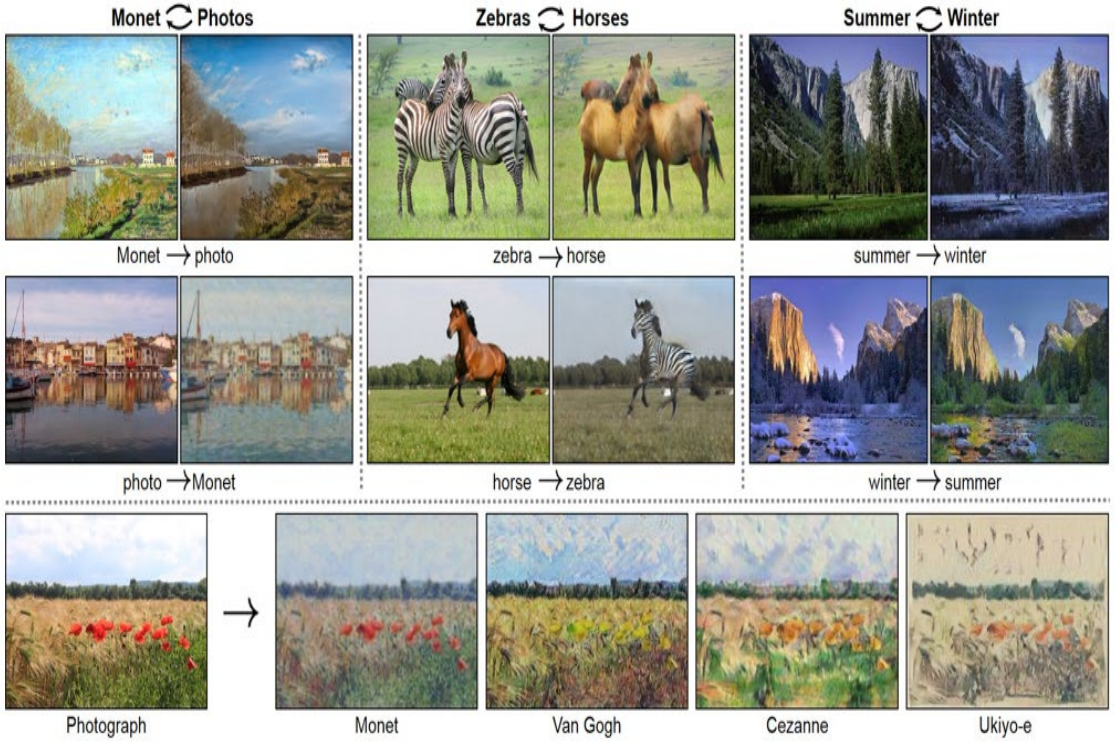
APPLICATION OF GANs

Since 2014, GANs have become popular with researchers.

We consider their varied applications in a number of domains.

COMPUTER VISION

- Image generation (CGAN, DCGAN)
- Image translation (Cycle GAN)
- Super-resolution (SRGAN, ProGAN)
- Face aging (StarGAN, AgeC-GAN)
- Texture synthesis (Style GAN)
- Video (MoCoGAN, Pose-GAN)
- Text-to-images (StackGAN, PSGAN)
- Image Synthesis (BigGAN, ProGAN, SAGAN)
- Object detection (Perceptual GAN)



Picture: Cycle GAN paper

Applications in short-term insurance e.g. telematics, weather patterns, crime spots, risk assessment, claims etc.

DATA AUGMENTATION

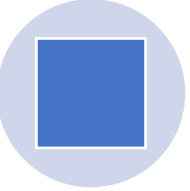
Boost training data sets and thereby the accuracy of predictive models

Train a GAN and use the generator to create new examples

Add samples to augment the training data set

Common on augmenting image data sets but fewer applications on tabular data sets such as mortality data

Data Augmentation GANs (DAGAN, BAGAN, TableGAN)



We show an example of this implemented in Python using a number of data sets

Applications in experience analysis, predictive modelling and pricing. Build a machine learning model, augment training data sets and get better predictive power i.e. rich data ~ powerful models.

TIME SERIES

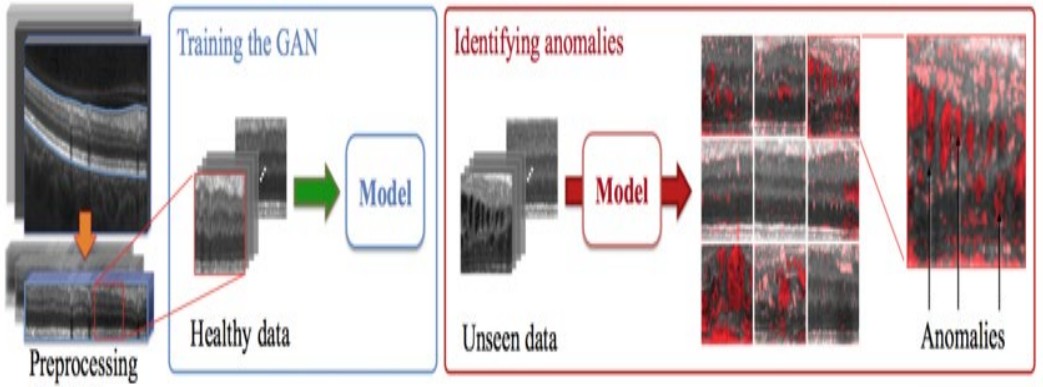
Traditional time series models such as ARIMA & GARCH make strong data, dependence structure & model assumptions

GAN offers the use of recurrent neural networks to generate time series data using Recurrent Conditional GANs

Stochastic modelling, economic modelling, stock prediction, mortality forecasting, Value-at-Risk, financial projections etc.

ANOMALY DETECTION

- Identify rare items or events which are suspicious from the majority of the data
- The generator models the distribution and create fake data
- The discriminator judges if it's close to the training data
- What happens to the judging part if we give it some sample not seen from the training distribution?
- AnoGAN/GANomaly for anomaly detection of medical images. Extension to other data sets.

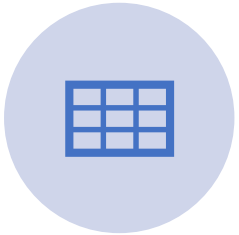


Picture: AnoGAN

Unsupervised learning applications, potentially without having to do cluster analysis which makes rather strong assumptions than GANs.

SEMI-SUPERVISED LEARNING

SEMI-SUPERVISED GAN TRAINS THE DISCRIMINATOR INTO A CLASSIFIER WHICH CAN ACHIEVE SUPERIOR CLASSIFICATION ACCURACY FROM AS FEW LABELED EXAMPLES AS POSSIBLE



REDUCTION IN DEPENDENCY OF CLASSIFICATION TASKS ON ENORMOUS LABELED DATA SETS

DISCRIMINATIVE MODELLING

ALTERNATIVE CLASSIFIER

SHOWN TO GENERATE BETTER THAN A COMPARABLE, FULLY-SUPERVISED CLASSIFIER

THIS HAS BEEN LAUDED AS THE MOST USEFUL GAN APPLICATION WITH GOOD PERFORMANCE WITH A SMALL NUMBER OF LABELS ON DATA SETS. VERY EFFECTIVE ON LIMITED, IMBALANCED DATA SETS.

MISSING DATA IMPUTATION

Missing data causes issues in standard analytic methods

This is almost always bound to happen in any model or analysis

There are many reasons data may be missing and it is appropriate that this is dealt with correctly

Single imputation vs. multiple imputation

Missing data mechanisms e.g. missing at random, missing completely at random and missing not at random

Percentage of missingness

GAIN/MisGAN appear better than imputation

Especially with complicated missing data mechanisms and higher missing data percentages

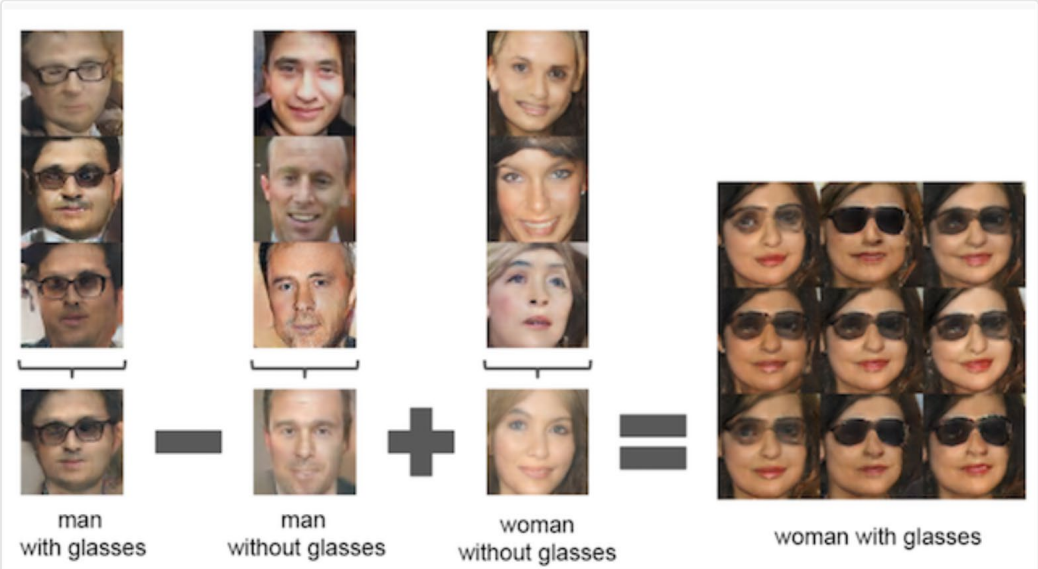
No dependence on missing data mechanism and model distributions

PRIVACY PRESERVATION

- Data of a lot of companies can be a secret, confidential or sensitive (such as medical diagnosis data)
- Sometimes need to share with third parties
- Can share salient features, sampled from a generative model
- This way we won't share exact data, just something that looks like it
- Privacy-preservation GANs accomplish this task (still relatively new)
- Useful for synthetic model points to be fed through any model, without needing policy details

OTHER DOMAIN AREAS

- Domain adaptation e.g. training & testing data from different distributions
- Learning joint distribution e.g. morbidity vs. mortality ~ single model
- Attention prediction e.g. focus on underlying key business
- Reinforcement learning e.g. stock optimization
- Fashion e.g. create new styles and matching
- Video-to-video synthesis (Vid2VID)
- Arts (GANGogh, GauGAN)
- Music (MuseGAN)
- Games e.g. Pokémon games
- e-Commerce
- Drug discovery & development (ChemGAN)
- Content creation etc.



Example of Vector Arithmetic for GAN-Generated Faces.Taken from Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, 2015.

- Check out more applications:
  - https://github.com/nashory/gans-awesome-applications
  - https://machinelearningmastery.com/impressive-applications-of-generative-adversarial-networks/

EXPERIMENTS

- We demonstrate a GAN application for data augmentation, tackling imbalance and limited data sets.

- Subsequently, we show that samples generated by the GAN boost the training data and lead to better predictive models.

DATA

- 5 data sets from the UCL repository
- Standardized
- Training-testing split (75%/25%)
- Python for analysis
- Train a Logistic Regression model on the baseline data and over-sampled data sets to predict each class
  - Before GAN, SMOTE, After GAN comparisons
- Test on the same unseen test data to compute the Precision, Recall, F1-Score, AUPRC and AUC

Data set	# Majority	# Minority	# Features
Credit card fraud	284807	492	31
Diabetes	500	268	8
Glass identification	144	70	9
Credit scoring	700	300	20
Breast Cancer	357	212	9

$$\log\left(\frac{h_\theta(x^{(d)})}{1-h_\theta(x^{(d)})}\right) = \theta_0 + \sum_{i=1}^{d}\theta_i X_i,\ 0 < h_\theta(x^{(d)}) < 1$$

METHODOLOGY

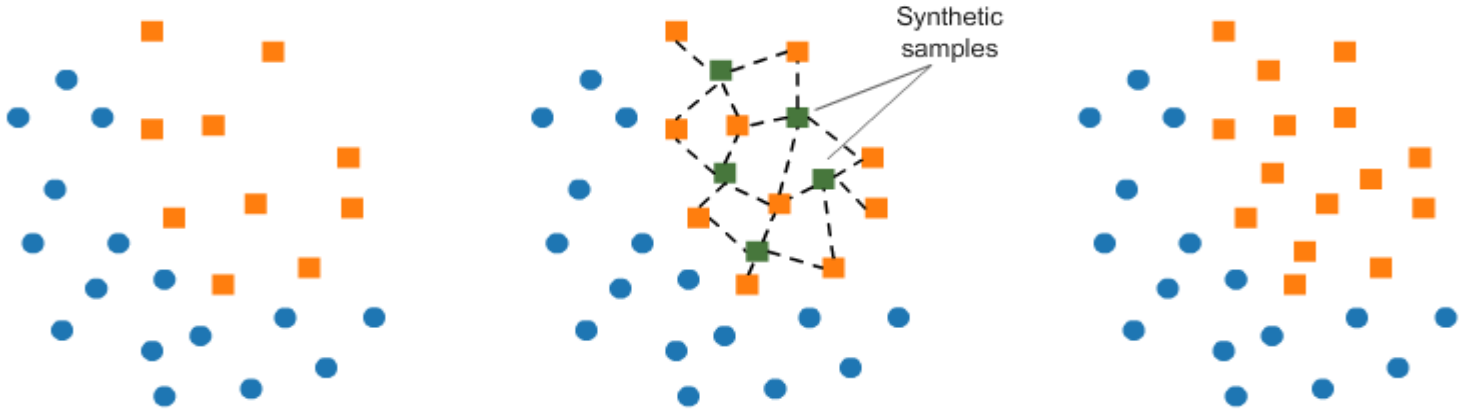
- Synthetic Over-sampling Technique (SMOTE) using R
- GAN training in Python using TensorFlow
- Predict minority classes of various data sets
- Over-sample minority cases using SMOTE and GAN
- Train a classifier and compare performance
- Statistical significance testing using Friedman/Post-hoc Nemenyi test

SYNTHETIC OVER-SAMPLING

- Synthetic Over-saMpling TEchnique (SMOTE), Chawla et al. (2002)
  - State-of-the-art solution and widely popular
  - Interpolation approach using k-nearest neighbour (k-NN): x random instance, y its random k-NN and $\alpha \sim U(0,1)$; then a new instance is generated as:

$$s = x + \alpha(y - x)$$

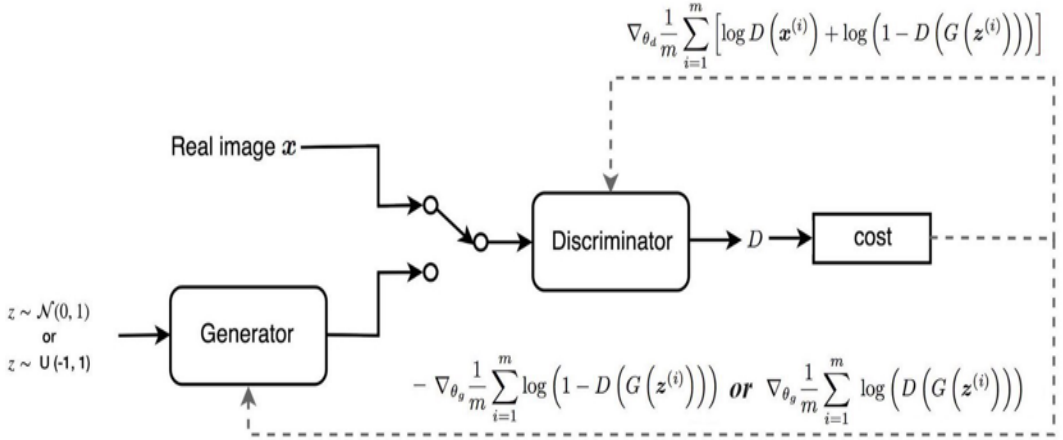
- Issues – over-lapping, noise, feature space, not based on true distribution, alters distribution, over-fitting, subjective

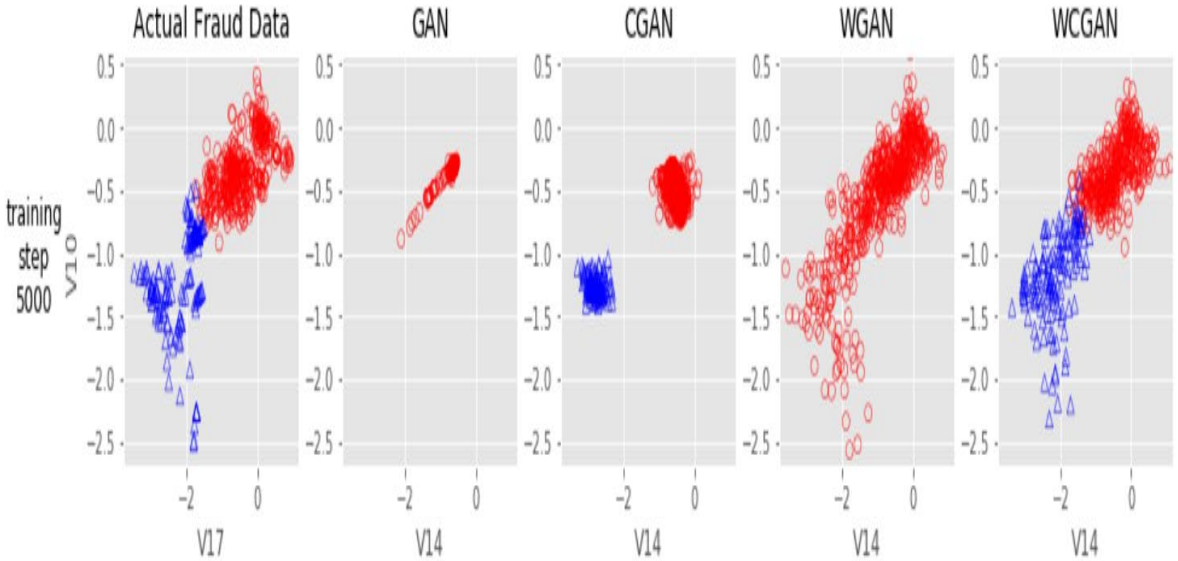


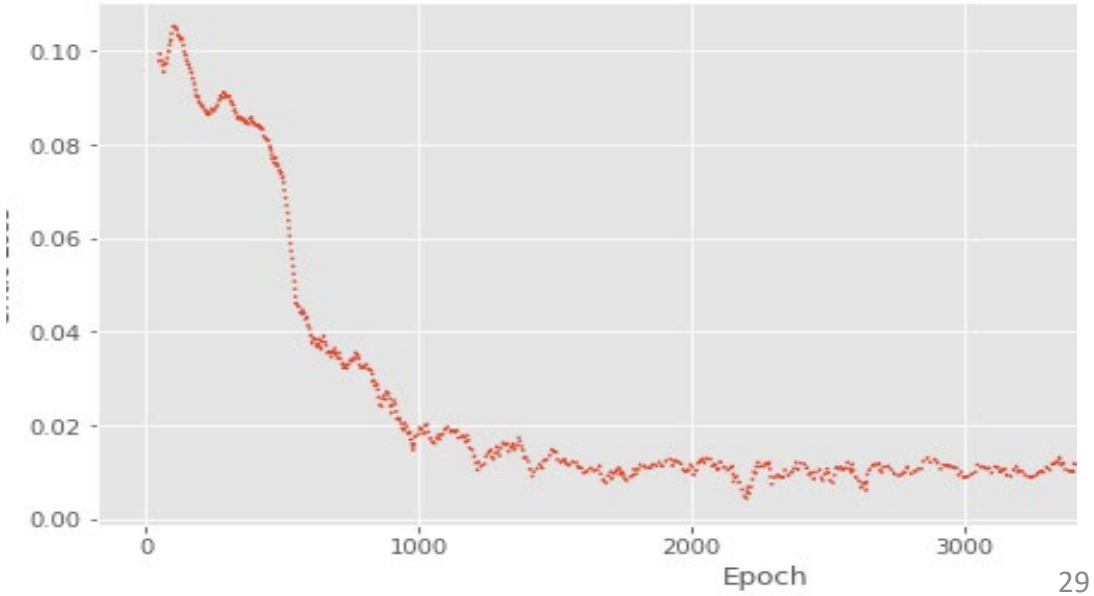
SMOTE is a pioneering method in imbalanced learning but it has issues. Other density estimation approaches have been proposed, including generative modelling. But most of these also make strong assumptions about the data distribution.

GAN TRAINING

- Tried a number of GAN architectures on generating realistic minority data sets
  - Original GAN suffers from mode collapse & non-convergence
  - WGAN is widely popular and remedies most of the issues
- Generated samples appear closer to real for WCGAN







EVALUATION METRICS

- Accuracy = (TN + TP)/(TN+FN+FP+TP)
- Precision = TP/(TP+FP) i.e. completeness
- Recall = TP/(TP+FN) i.e. exactness
- F1-Scoree: harmonic mean between precision and recall
- Receiver Operating Characteristic (ROC)
  - Trade-off between TP and FP rate
- **Area under ROC (AUC)**
- **Area under the Precision-Recall curve (AUPRC)**

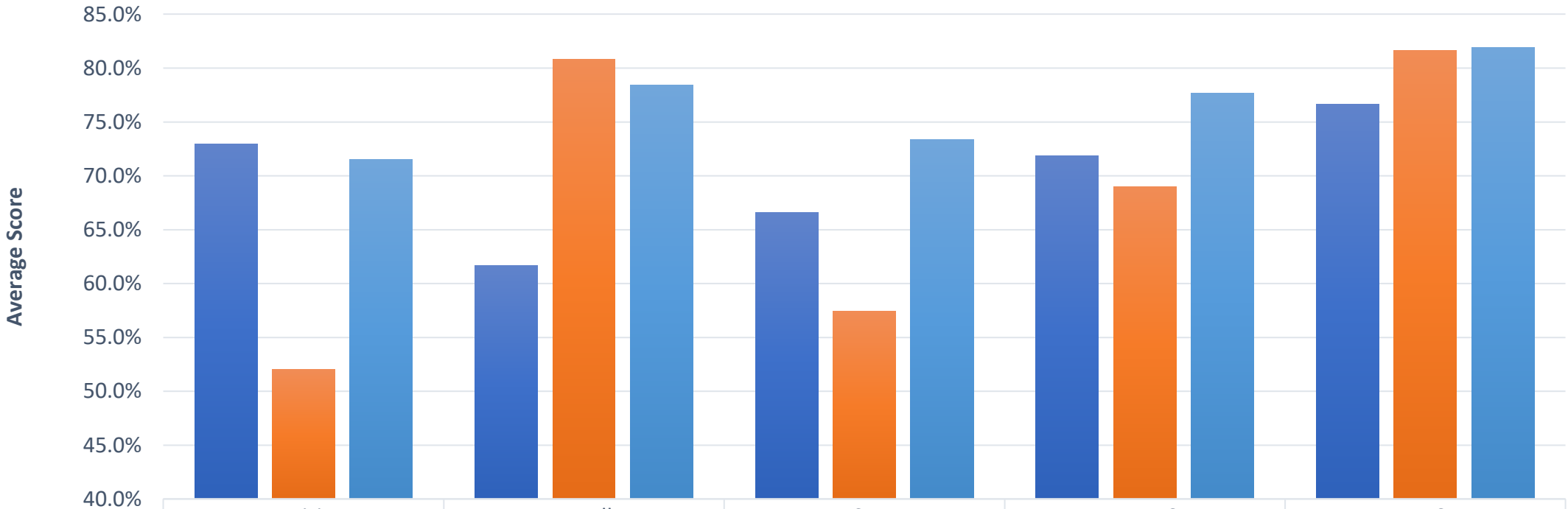
		Predicted Class	
		No	Yes
Observed Class	No	TN	FP
	Yes	FN	TP

TN	**True Negative**
FP	**False Positive**
FN	**False Negative**
TP	**True Positive**

Closer the metrics are to 100%, the better the model

APPLICATION

RESULTS



	Precision	Recall	F1-Score	AUPRC	AUC
Before	73.0%	61.7%	66.6%	71.9%	76.7%
After SMOTE	52.0%	80.8%	57.4%	69.0%	81.6%
After GAN	71.5%	78.4%	73.4%	77.7%	81.9%

STATISTICAL SIGNIFICANCE TESTING

- Focus on the AUPRC and then test if WGAN is statistically better than SMOTE
- 3 of the 5 data sets suggest a difference as p-values<5%
- Significant superiority of WGAN over SMOTE
- SMOTE alters the data distribution, creates over-lapping cases & not based on the true underlying distribution
- WGAN makes no assumptions about the data distribution & leads to better predictive power than SMOTE

Data set	P-value	Significance
Credit Card Fraud	2.9560e-23	Yes
Pima Indians Diabetes	0.188386	No
German Credit Scoring	1.0683e-11	Yes
Glass Identification	0.465622	No
Breast Cancer Wisconsin	4.0085e-12	Yes

Table 9: Results for Friedman's test

Test	Credit Card Fraud	German Credit	Breast Cancer
WCGAN-GP vs. SMOTE	0.001000	0.003000	0.001000

Table 10: Results for the post-hoc test

CONCLUSIONS

SUMMARY

- GANs are powerful generative models applicable in many domains, better than other generative models
- Implementations in Keras, Pytorch & Tensorflow
- GAN augmented data sets achieved better predictive power than original models or SMOTE
- Scope for actuarial use
  - Use augmentation to boost predictive models (easy immediate use)
  - Time series generation instead of stochastic simulations (interesting scope)
  - Anomaly detection
  - Discriminative modelling/semi-supervised learning (probably the most important)
  - Domain adaptation
  - Joint distribution
  - Learn new cases or something rare
  - Produce new data in data-limited situations

LIMITATIONS & FUTURE WORK

- We considered binary cases and less complex data sets
- Open challenges & new directions
  - Largely fixed with lots of GAN variants
  - GANs for discrete data
  - GAN measure of uncertainty and evaluation metrics
  - Fake data created by GANs – problems?
  - Verified artificial intelligence
- Future of GANs will be characterised by:
  - Advancements in deep learning;
  - Impressive results and their use in commercial applications;
  - Maturation of the training process; and
  - Open acceptance

Picture: www.forbes.co.za

"…in the process of training generative models, we will endow the computer with the understanding of the world and what is made up of."

OpenAI

REFERENCES

- **GAN**: I. Goodfellow et al. "Generative Adversarial Nets". In: Advances in Neural Information Processing Systems. 2014, pp. 2672–2680.
- **GAN Hacks**: *https://github.com/soumith/ganhacks*
- ***GAN Zoo***: https://github.com/hindupuravinash/the-gan-zoo
- ***GAN Applications***: https://github.com/nashory/gans-awesome-applications tps://github.com/nashory/gans-awesome-applications
- **GAN TUTORIAL:** Ian *Goodfellow's NIPS 2016 Tutorial, https://sites.google.com/view/cvpr2018tutorialongans/*
- **Improved GAN training:** T. Salimans et al. "Improved techniques for training GANs". In: Advances in Neural Information Processing Systems. 2016, pp. 2234–2242.
- **SMOTE:** N.V. Chawla et al. "SMOTE: synthetic minority over-sampling technique". In: Journal of Artificial Intelligence Research 16 (2002), pp. 321–357.
- **SMOTE VARIANTS:** A. Fernández et al. "Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary". In: Journal of Artificial Intelligence Research 61 (2018), pp. 863–905.
- **PDFOS:** M. Gao et al. "PDFOS: PDF estimation based over-sampling for imbalanced two-class problems". In: Neurocomputing 138 (2014), pp. 248–259.
- **WGAN:** M. Arjovsky, S. Chintala, and L. Bottou. "Wasserstein GAN". In: arXiv preprint arXiv:1701.07875 (2017).
- **WGAN-GP:** I. Gulrajani et al. "Improved training of Wasserstein gans". In: Advances in Neural Information Processing Systems. 2017, pp. 5767–5777.

Thank you for your attention

ACTUARIAL COLLOQUIUM PARIS 2020

Contact details:

+27833837000

School of Statistics and Actuarial Science,
University of the Witwatersrand,
Braamfontein, Johannesburg, 2001,
South Africa

Sngwenduna@gmail.com
Rendani.mbuvha@wits.ac.za

https://www.actuarialcolloquium2020.com/

ACTUARIAL COLLOQUIUM PARIS 2020

Disclaimer:

The views or opinions expressed in this presentation are those of the authors and do not necessarily reflect official policies or positions of the Institut des actuaires (IA), the International Actuarial Association (IAA) and its Sections.

While every effort has been made to ensure the accuracy and completeness of the material, the IA, IAA and authors give no warranty in that regard and reject any responsibility or liability for any loss or damage incurred through the use of, or reliance upon, the information contained therein. Reproduction and translations are permitted with mention of the source.

Permission is granted to make brief excerpts of the presentation for a published review. Permission is also granted to make limited numbers of copies of items in this presentation for personal, internal, classroom or other instructional use, on condition that the foregoing copyright notice is used so as to give reasonable notice of the author, the IA and the IAA's copyrights. This consent for free limited copying without prior consent of the author, IA or the IAA does not extend to making copies for general distribution, for advertising or promotional purposes, for inclusion in new collective works or for resale.