[TRUST]
[INDUSTRIES]
[PEOPLE]
[COMPETENCE]
[RELIABILITY]
[TECHNOLOGY]
[INNOVATION]
[INDEPENDENT]
[CAN DO]

SYNTHETIC DATA FOR MACHINE LEARNING

Hendrik Niemeyer
PyCon.DE 2017 · Karlsruhe · 25-Oct-2017

ROSEN
empowered by technology

ABOUT ME

Hendrik Niemeyer

Theoretical Physics

Data Scientist

@hniemeye

https://github.com/hniemeyer

hniemeyer@rosen-group.com

THE ROSEN GROUP
NUMBERS AND FIGURES

Company

- Founded in 1981 by Hermann Rosen
- Locations world-wide: over 25
- Employees world-wide: over 3000

Market Position

- Market leader since 2008
- Technology leader since 2005
- Revenue: over 430 Mio. Dollar (2016)
- We work in over 120 countries

Business Portfolio

- Asset Care – Diagnostic and Integrity Solutions
- Enhanced Materials – Intelligent Plastic Solutions
- New Business – Flow Metering Solutions

THE ROSEN GROUP
BUSINESS PORTFOLIO (EXCERPT)

ASSET CARE

Diagnostic Solutions

- Field Products & Services
- Proficient Pipeline Diagnostics
- Advanced Pipeline Diagnostics
- Challenging Pipeline Diagnostics
- NDT Diagnostics
- Industrial Diagnostics

Integrity Solutions

- Integrity Management Systems
- Integrity Management Services

ENHANCED MATERIALS

Intelligent Plastic Solutions

- CoHigh-performance Elastomers
- Coatings
- Smart Plastic Systems

NEW BUSINESS

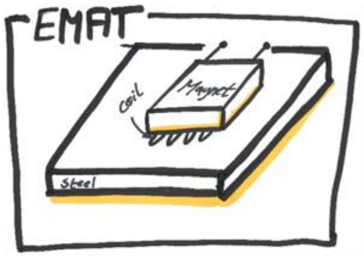
Flow Metering Solutions

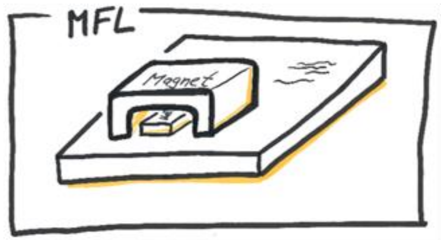
- Steam Flow
- Multiphase Flow
- Advanced Gas Flow
- Standard Industrial Flow
- Commodity Flow
- Novel High End Flow Applications

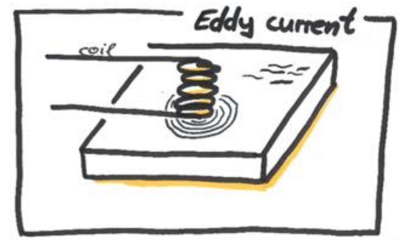
R³ Diagnostic Services

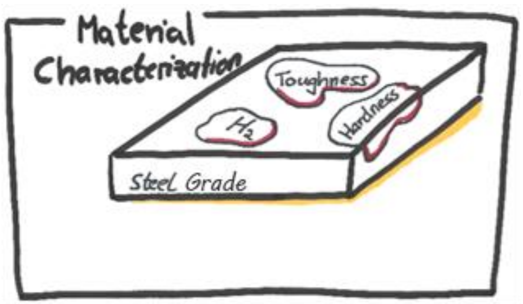
- 365 / 24 / 7 service
- Onshore systems
- Liquid lines
- Pipeline diameters from 6” to 24”
- Pipelines up to 60 miles
- Product temperatures up
  to 150ºF
- Standard wall thicknesses
- 1.5D minimum bend radius
- Product speeds up to 7 mph

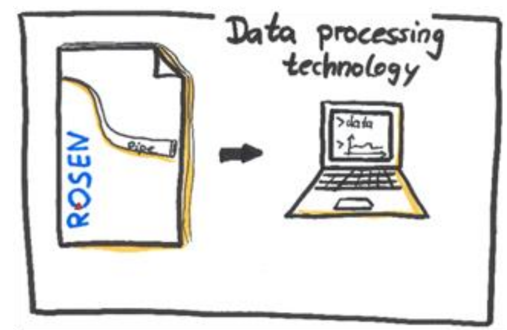
THE ROSEN GROUP
MAIN TECHNOLOGIES

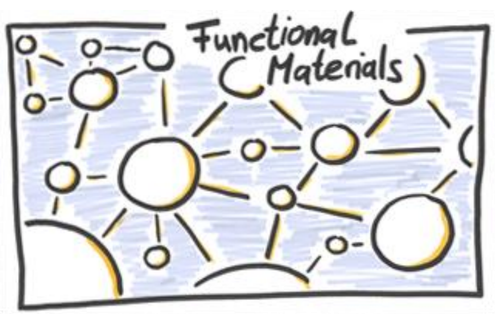


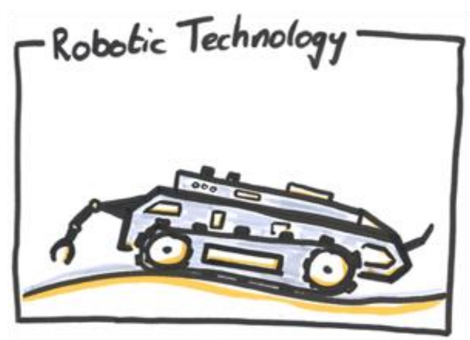


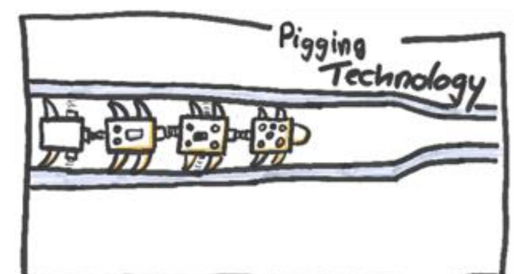










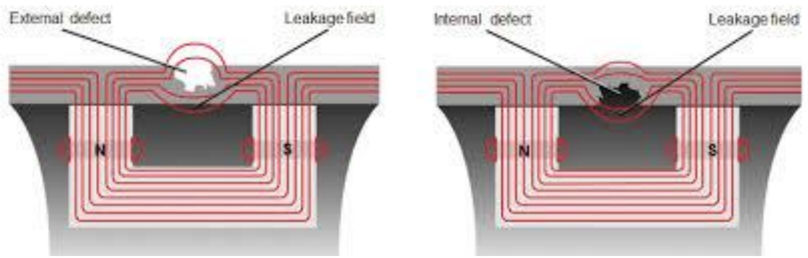


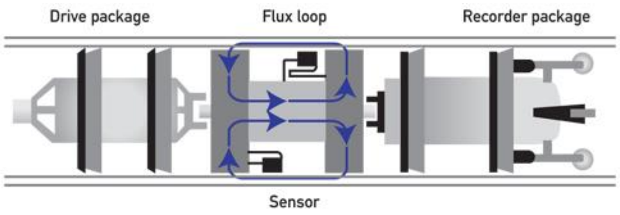
MAGNETIC FLUX LEAKAGE

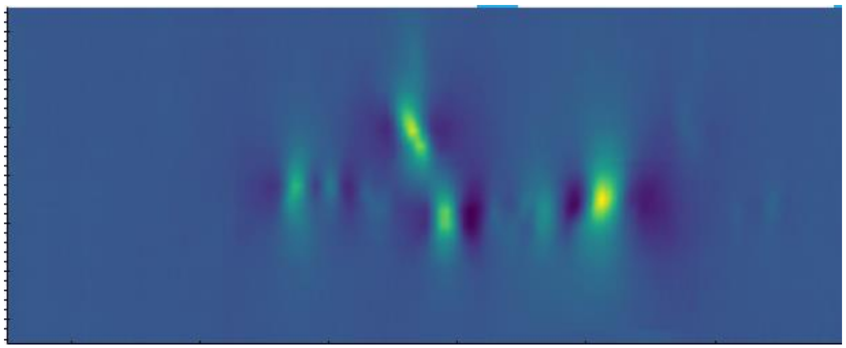
ROSEN
empowered by technology

- Measure volume loss in pipeline wall
- Indirect measurement principle
- Image-like data (2d array of amplitudes)

- Tasks: Detect, classify and estimate defect geometry from measured data.







GROUND TRUTH IN COMPUTER VISION

ROSEN
empowered by technology

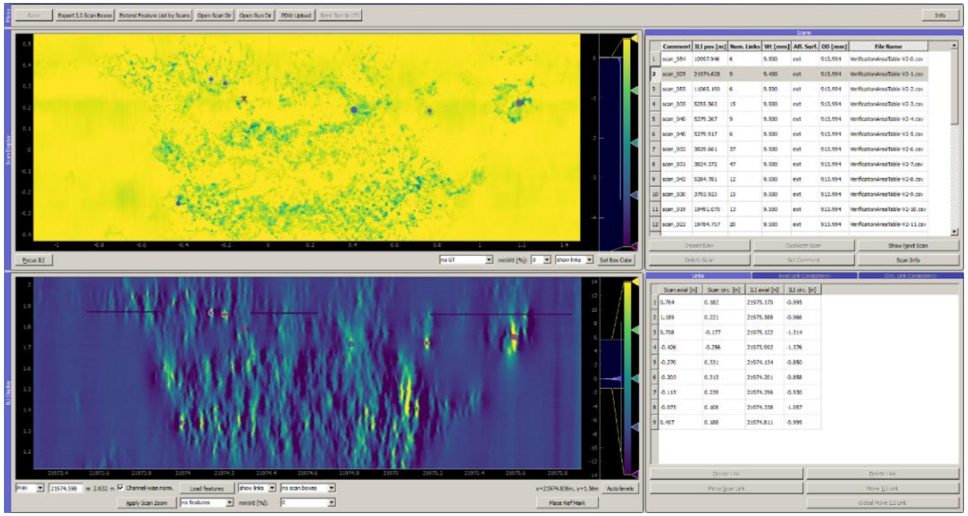
- Acquire a large set of images
- Use Amazon mechanical turk / bachelor students for labeling

The platform provides useful tools to accurately describe your task, specify consensus rules, and the amount you will spend for each item. Roughly, considering a $0.30 reward for each task and only one submission for each item, you could label a 1,000 record-dataset for as little as $300 (plus fees) in a few hours. This might just be cheap, fast, and accurate enough.

OUR GROUND TRUTH – FIELD VERIFICATIONS

ROSEN
empowered by technology

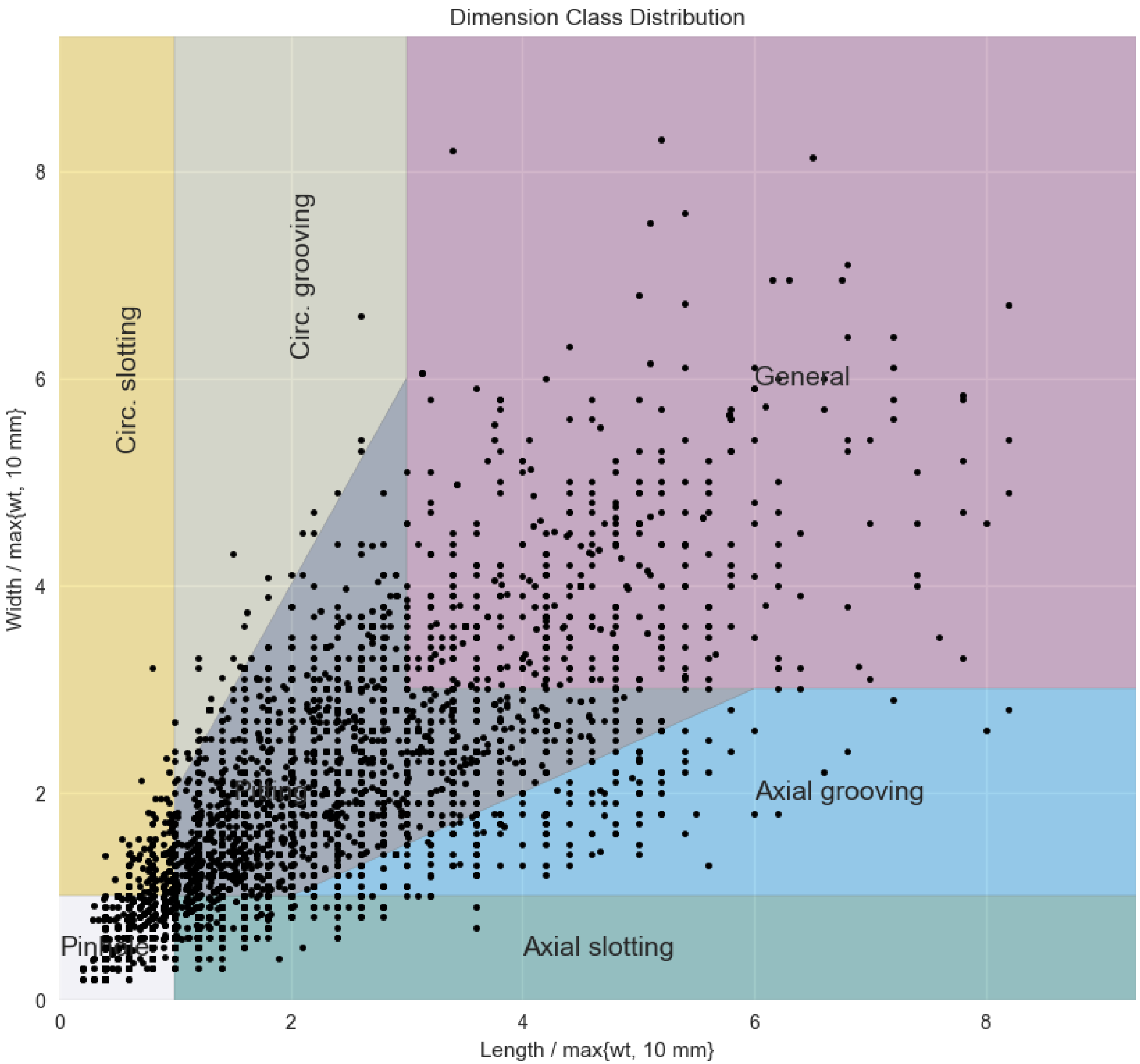
- Pipelines need to be dug up to get access to ground truth
- Defect geometry can be measured using 3d laser scanners
- Alignment to NDT data and labeling by hand



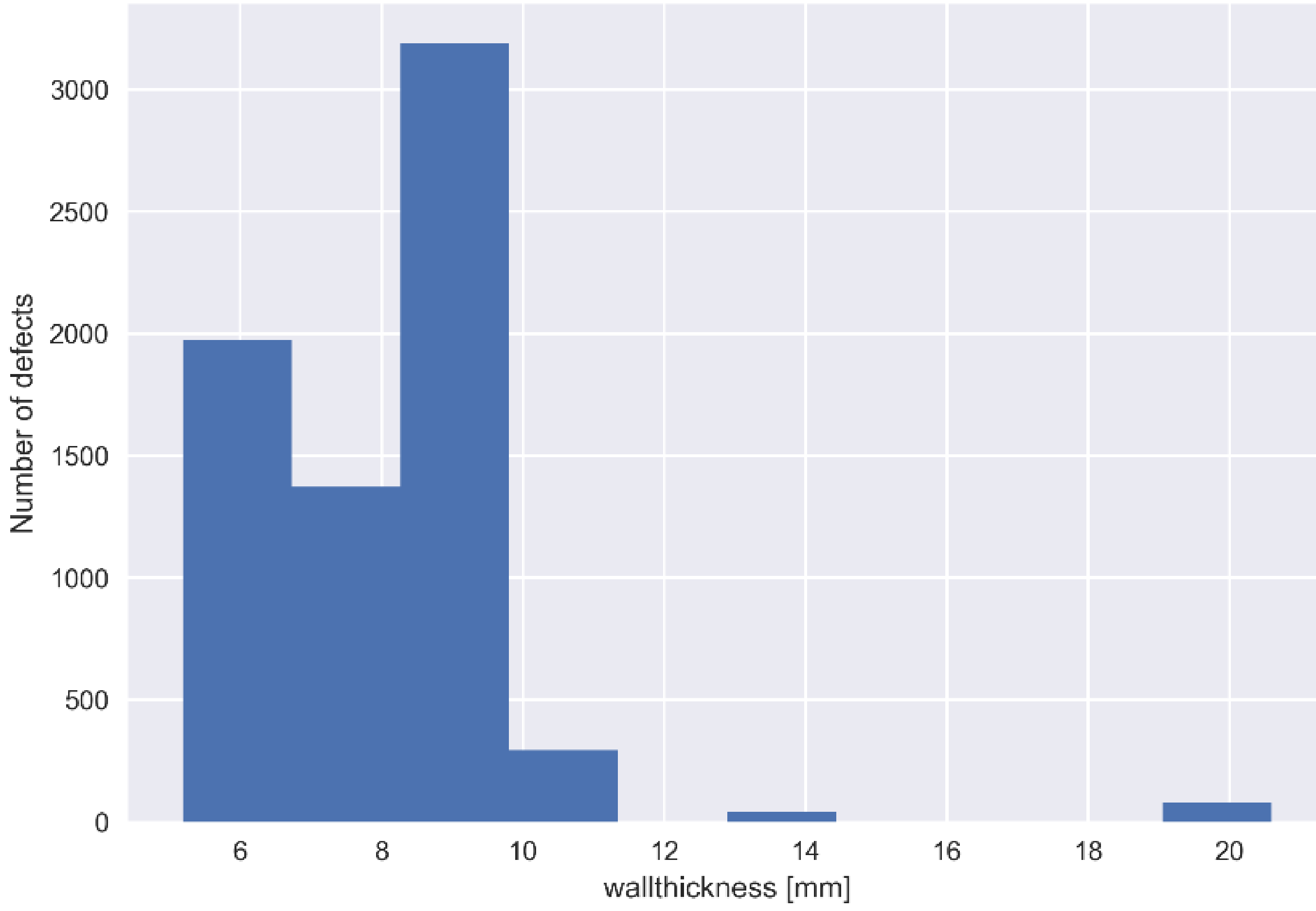
WHY DO WE NEED SYNTHETIC DATA?

Ground truth for image data	Ground truth for inline inspection data
Anyone can label	Labeling by human experts
Small costs	Very high costs
All data can be labeled	Dig up availability
	Inherently imbalanced

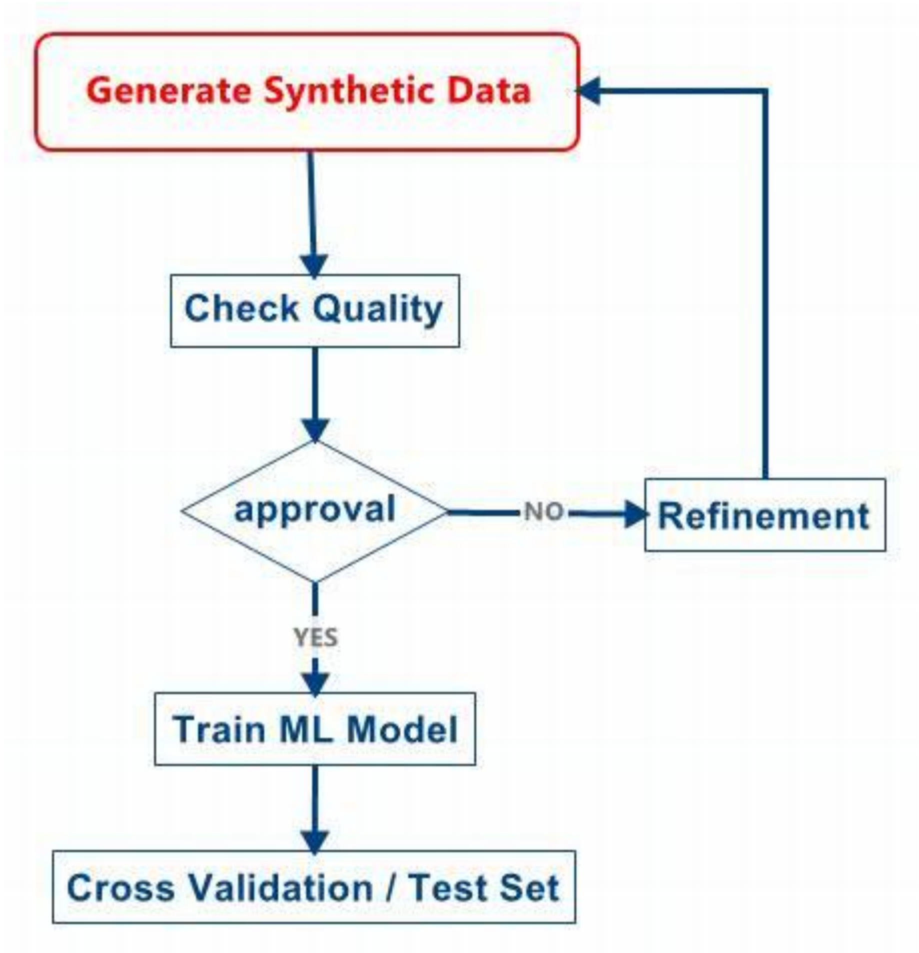
CLASS IMBALANCE



CLASS IMBALANCE



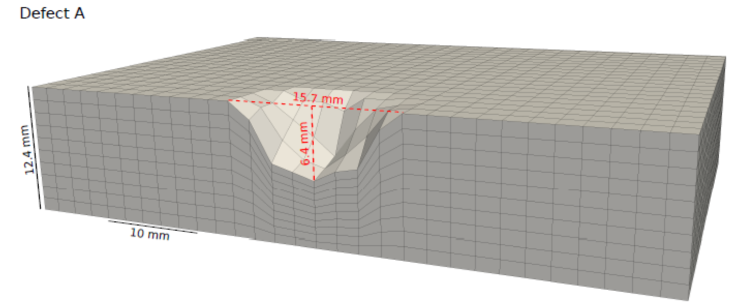
WORKFLOW

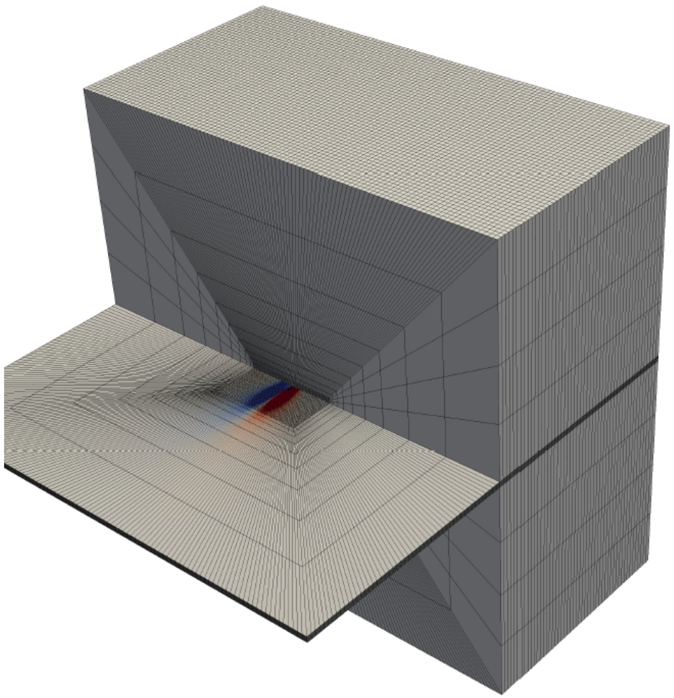

Generate Synthetic Data
Check Quality
approval
NO
Refinement
YES
Train ML Model
Cross Validation / Test Set

SIMULATIONS

- Magnetic flux leakage: Electromagnetic FEM simulations

- No real data or training needed to create model
- Parameters of simulation can be set
- Issue of accuracy versus computing time
- Human expert needed for design of simulation

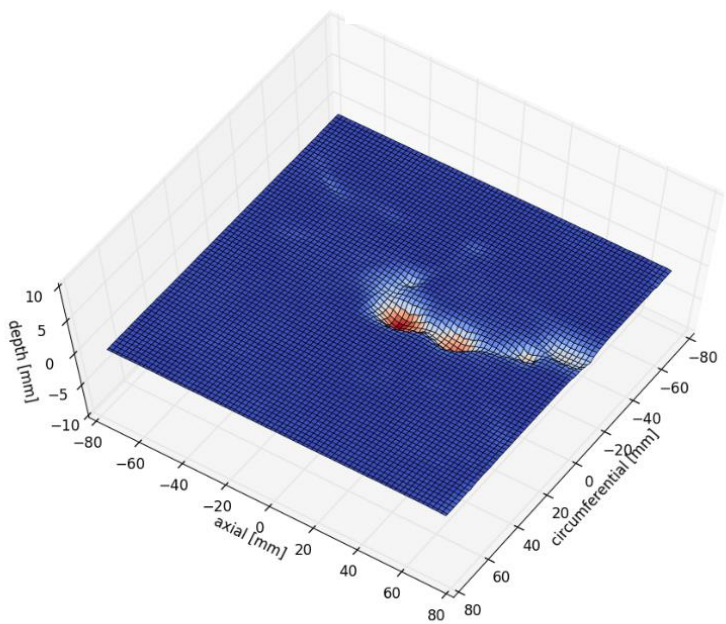


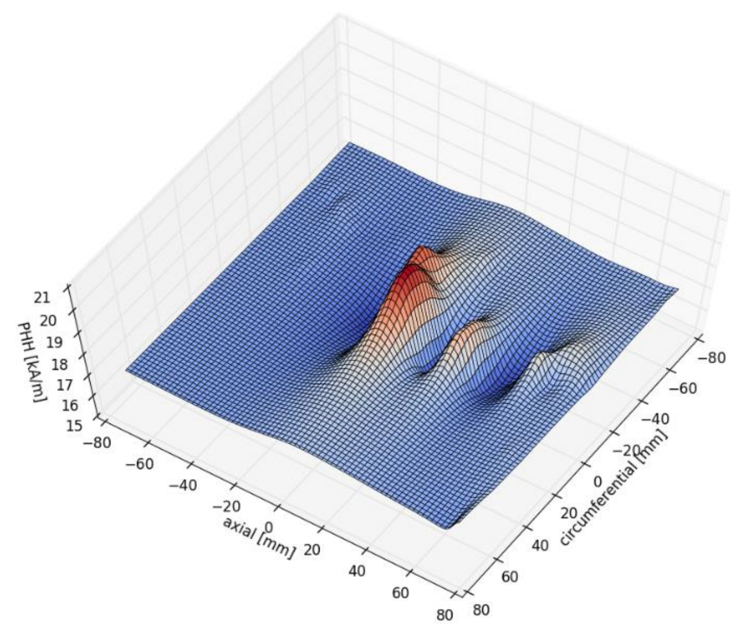


FEM@ROSEN

	Artificial	Scan
Number of FEM simulations	275707	594139
Background magnetization range [kA/m]	10.0 – 30.0	10.0 – 35.0
Wall thickness range [mm]	4.0 – 30.0	4.0 – 30.0
Length range [mm]	4.0 – 49.0	4.0 – 100.0
Width range [mm]	4.0 – 49.0	4.0 – 103.0
Depth range [%]	10.0 – 95.0	9.7 – 97.8

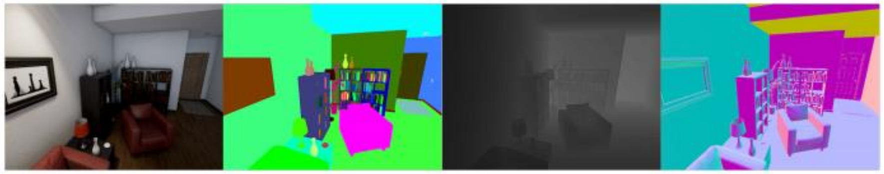
Table 2: Overview of FEM simulation statistics.

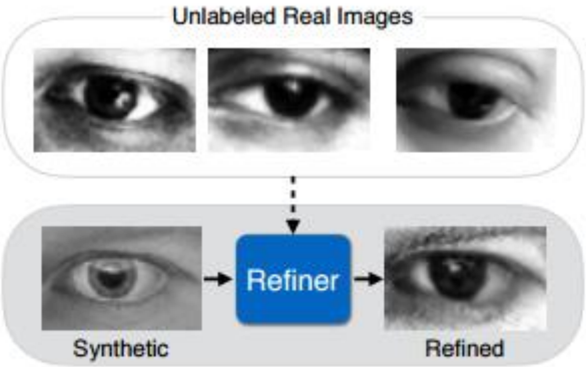




SIMULATIONS

- Using GTA 5 to get semantically labeled traffic data (Richter et. al, ECCV 2016)
- Gaze prediction Shrivastava, (Shrivastava et. al.,arxiv:1612.07828v1)
- UnrealCV (Qui et. al., arxiv:1609.01326v1)





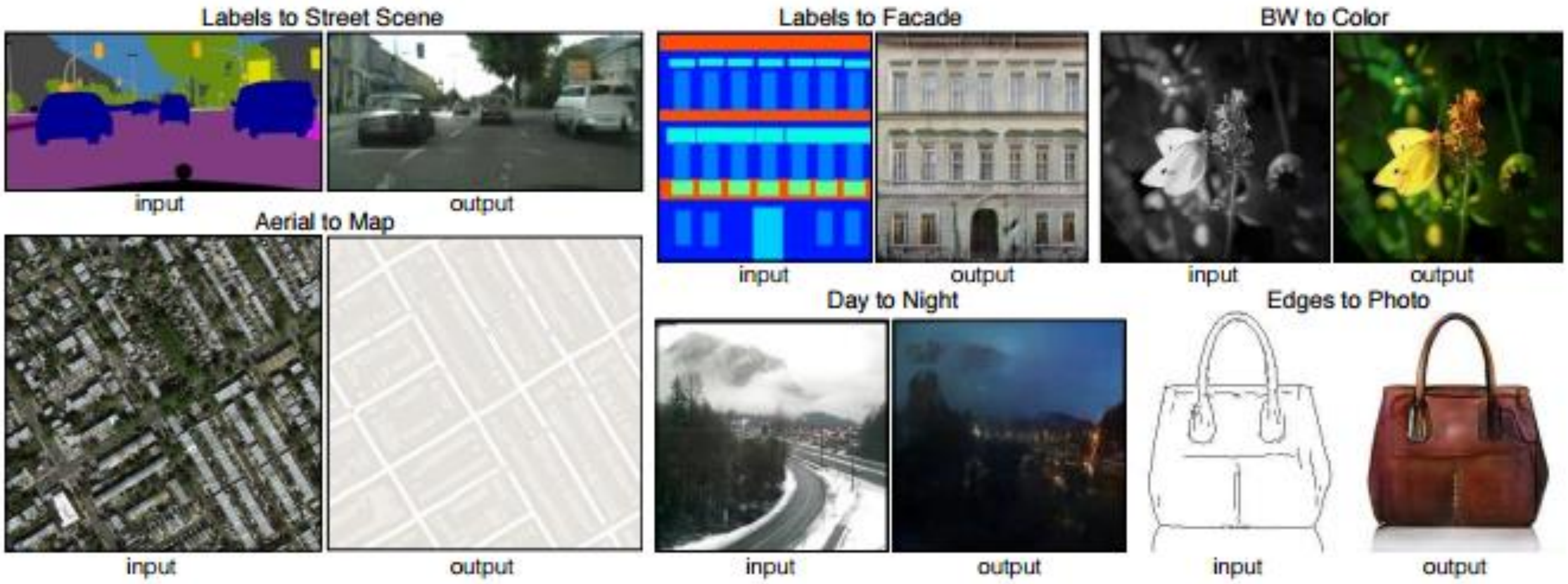
GENERATIVE ADVERSARIAL NETWORKS (GOODFELLOW ET AL, 2014)

- Generator G produces training sample from vector of random noise
- Discriminator D can classify whether a training sample is real or produced by G

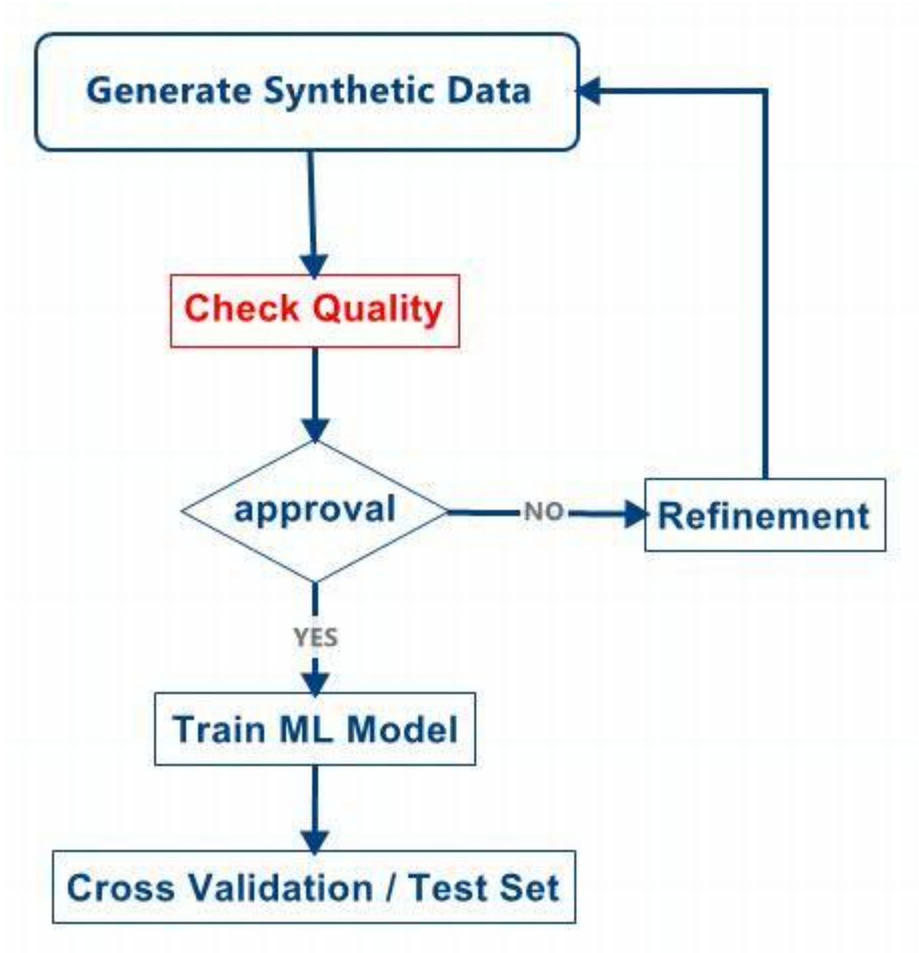
$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))].$$

CONDITIONAL GAN / PIX2PIX (ISOLA ET AL, 2016)

- Replace random noise with meaningful input for the generator G
- Input could be parameters of simulation or even a simulation result for refinement

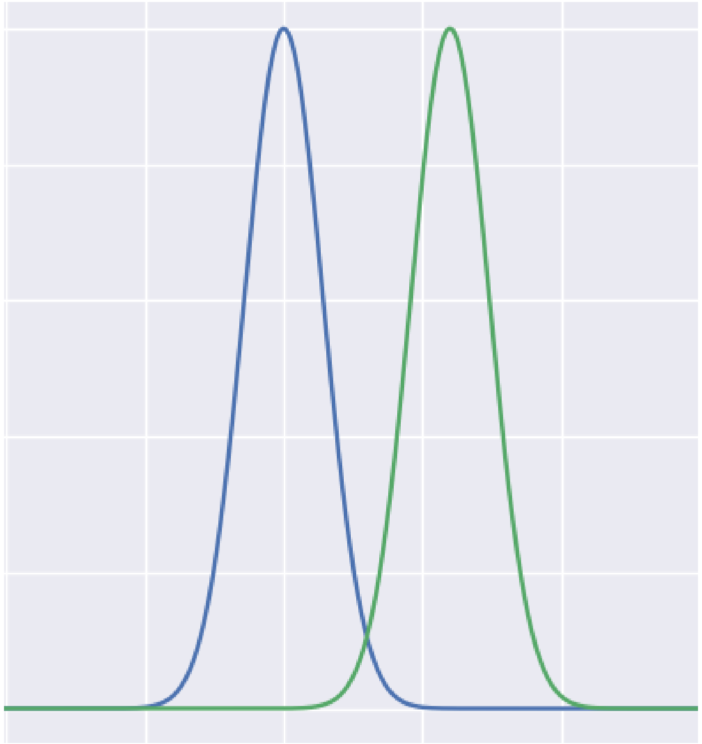


WORKFLOW


Generate Synthetic Data
Check Quality
approval
NO
Refinement
YES
Train ML Model
Cross Validation / Test Set

QUALITY CHECK

- “Synthetic Gap”: Gap between real and synthetic distributions
- ML model might learn artifacts and details specific for the synthetic data and might fail on new real data



T-DISTRIBUTED STOCHASTIC NEIGHBOR EMBEDDING

$$p_{j|i} = \frac{\exp\left(-d(\boldsymbol{x}_i, \boldsymbol{x}_j)/(2\sigma_i^2)\right)}{\sum_{i \neq k} \exp\left(-d(\boldsymbol{x}_i, \boldsymbol{x}_k)/(2\sigma_i^2)\right)}, \quad p_{i|i} = 0,$$

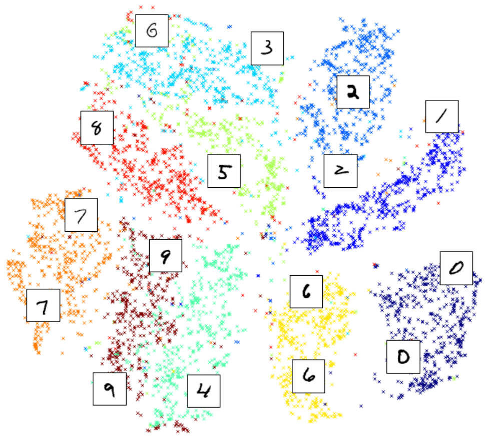
$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}.$$

$$q_{ij} = \frac{(1 + ||\boldsymbol{y}_i - \boldsymbol{y}_j)||^2)^{-1}}{\sum_{k \neq l}(1 + ||\boldsymbol{y}_k - \boldsymbol{y}_l)||^2)^{-1}},$$

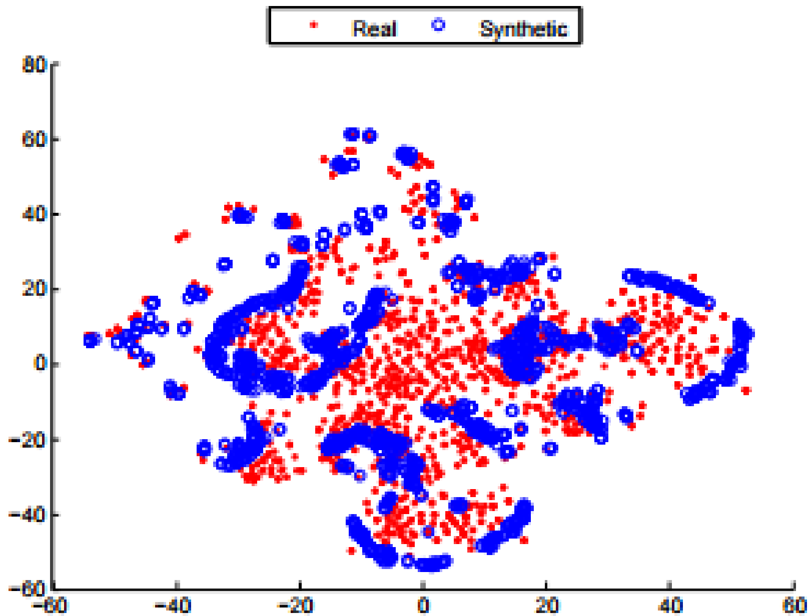
$$KL(P|Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

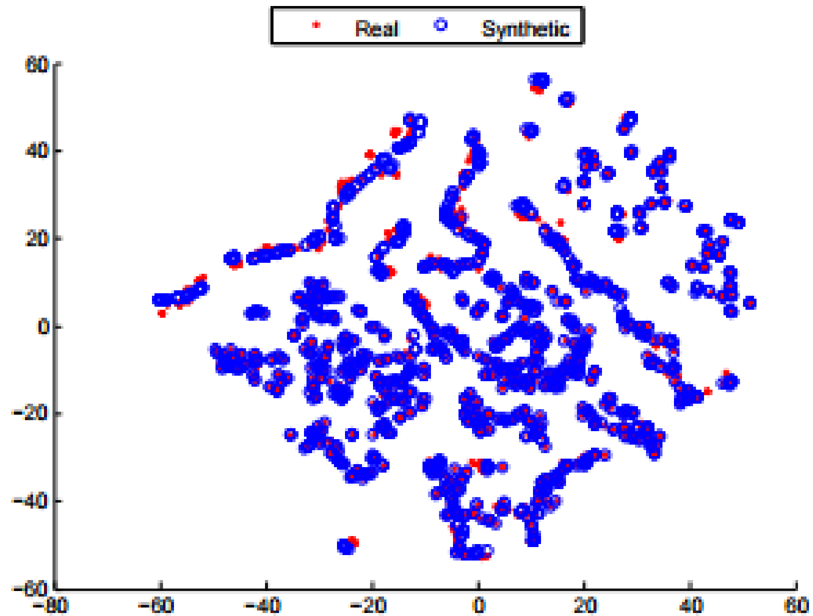
- Learn low-dimensional embedding of feature vector by minimizing KL between similarities in both spaces
- Different distribution in low-dimensional space to compute similarities
- “Tends to cluster points by their classes”
- Van der Maaten, JMLR 2008

MNIST dataset – Two-dimensional embedding of 70,000 handwritten digits with t-SNE



TSNE



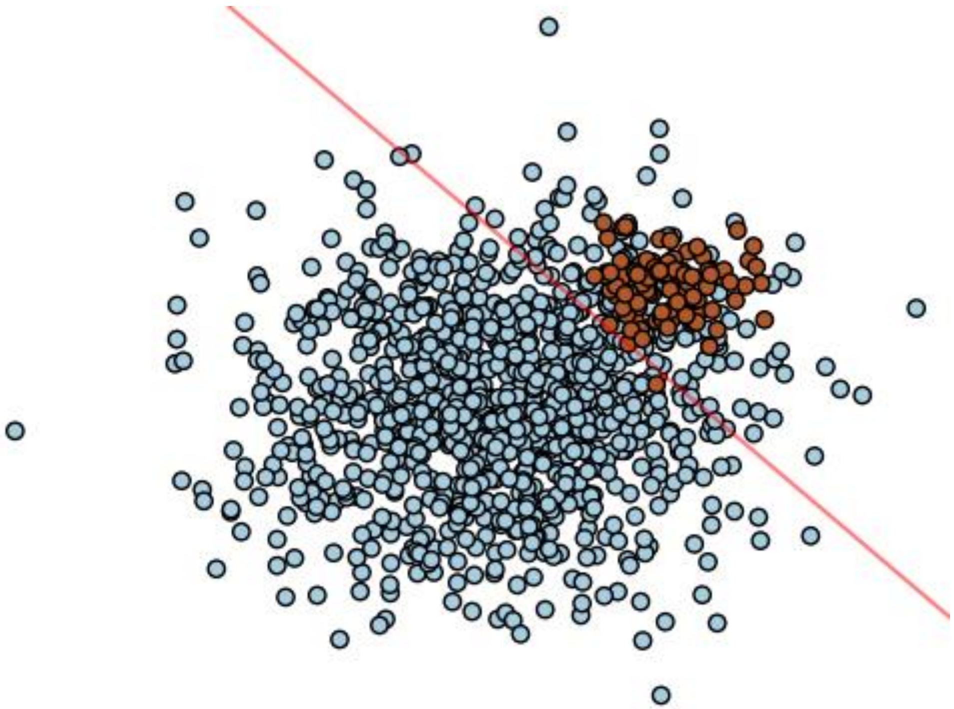


Example from Zhang et al, 2015, arxiv:1503.03163v1 shows how a synthetic gap can be visually identified

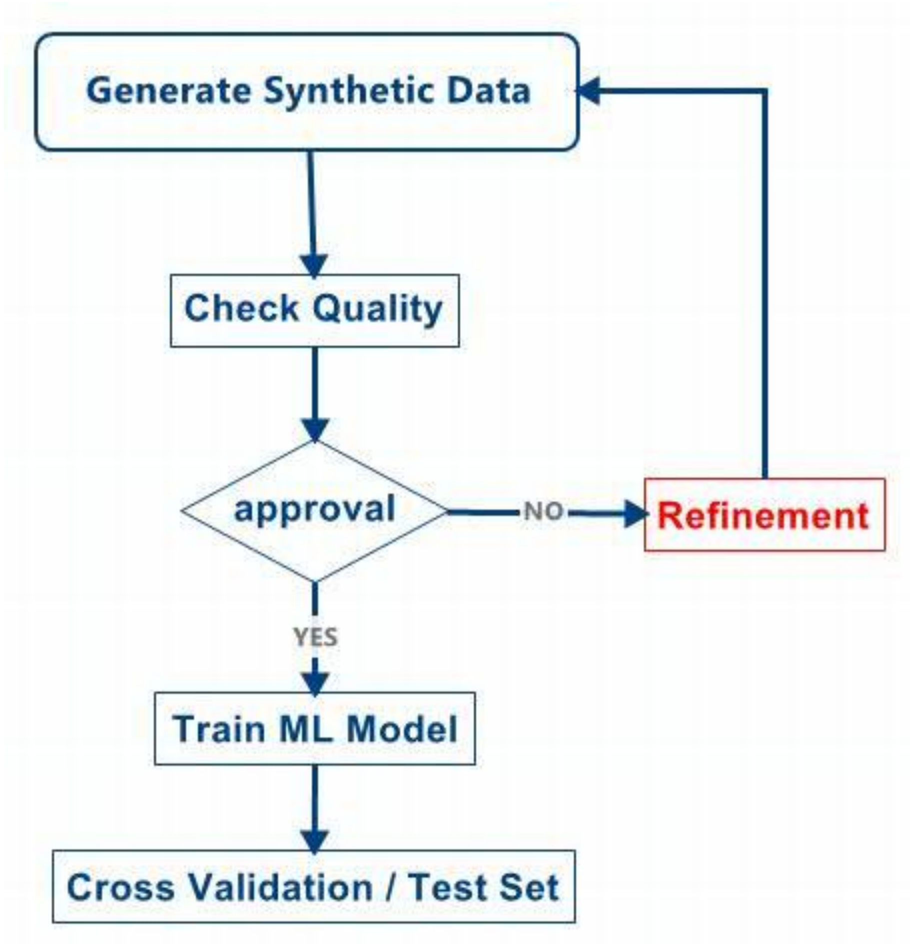
- If synthetic and real data are different enough they might not cluster in a TSNE embedding

CLASSIFIER APPROACH

- Inspired from GANs: Train classifier to discriminate between real and synthetic data
- Failure to do so is a good indicator that there is no synthetic gap
- Classification error is a quantitative measure of the quality of the synthetic data



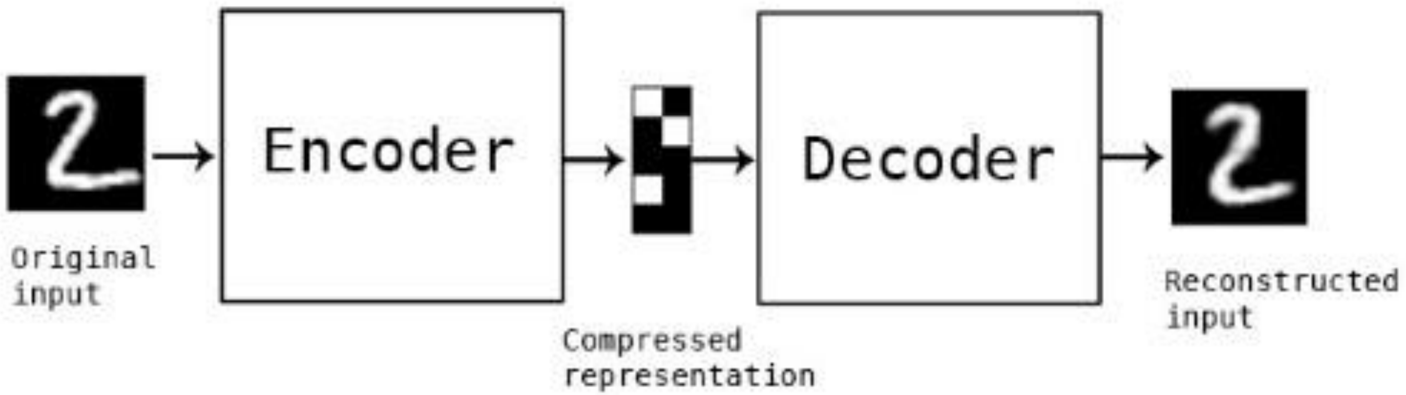
WORKFLOW



Generate Synthetic Data
Check Quality
approval
NO
Refinement
YES
Train ML Model
Cross Validation / Test Set

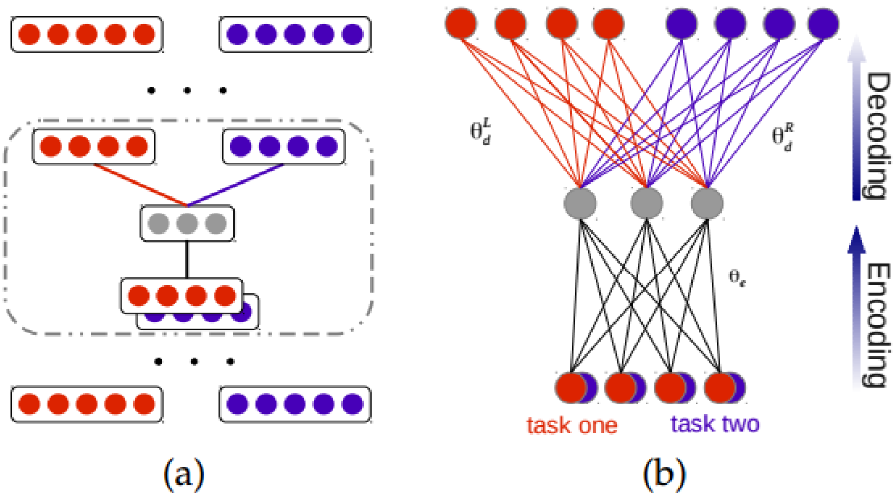
REFINEMENT

- cGAN: Simulation as conditional input and real data as output
- Autoencoder: Learn mapping from synthetic to real data (e.g.: Chen et al, Marginalized Denoising Autoencoders for Domain Adaption, arxiv:1206.4683)



https://blog.keras.io/building-autoencoders-in-keras.html

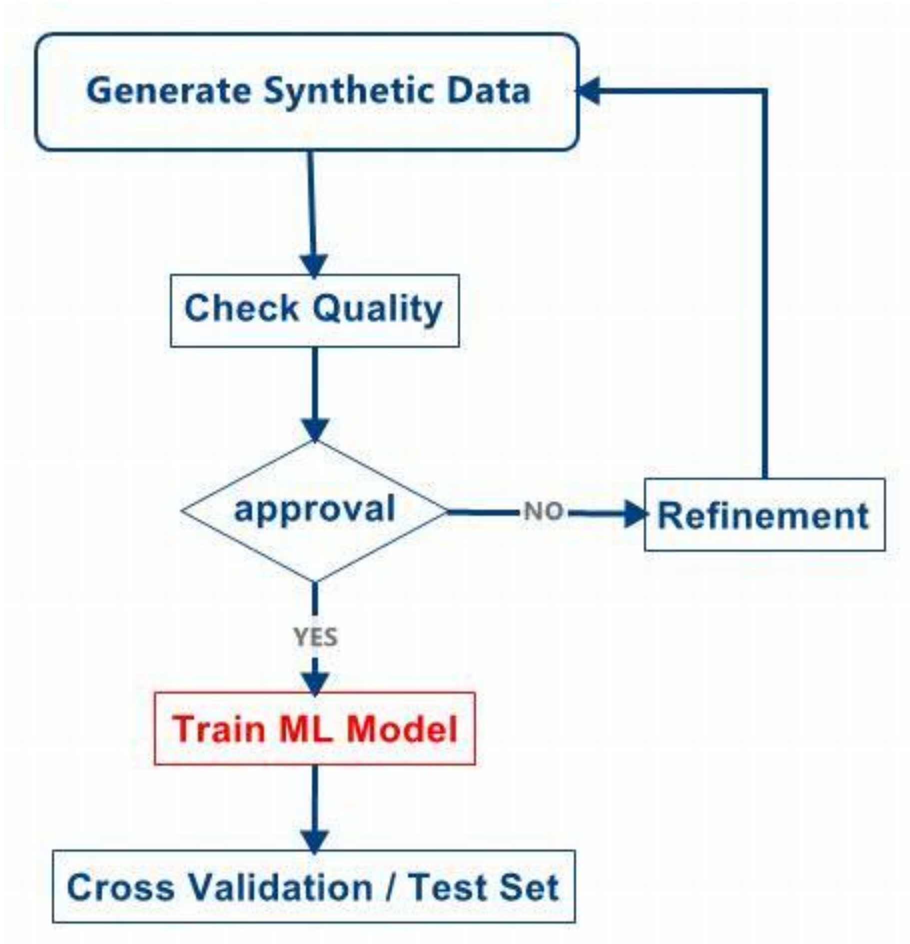
MULTICHANNEL AUTOENCODER



- Learn to tasks at one with shared hidden layer
  - Task One: Input: Synthetic data, Output: Real data
  - Task Two: Input: Real data, Output: Real data

- Zhang et al, 2015, arxiv:1503.03163v1

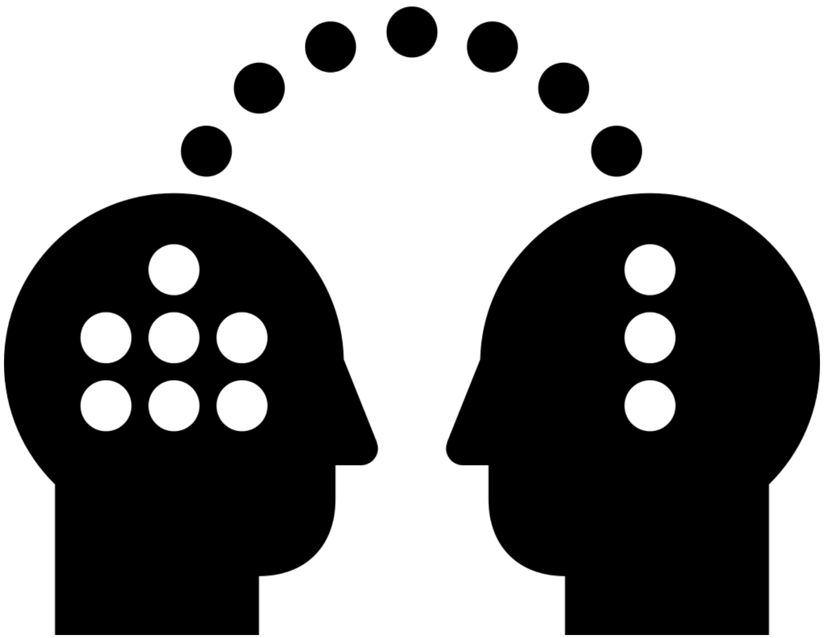
WORKFLOW


Generate Synthetic Data
Check Quality
approval
NO
Refinement
YES
Train ML Model
Cross Validation / Test Set

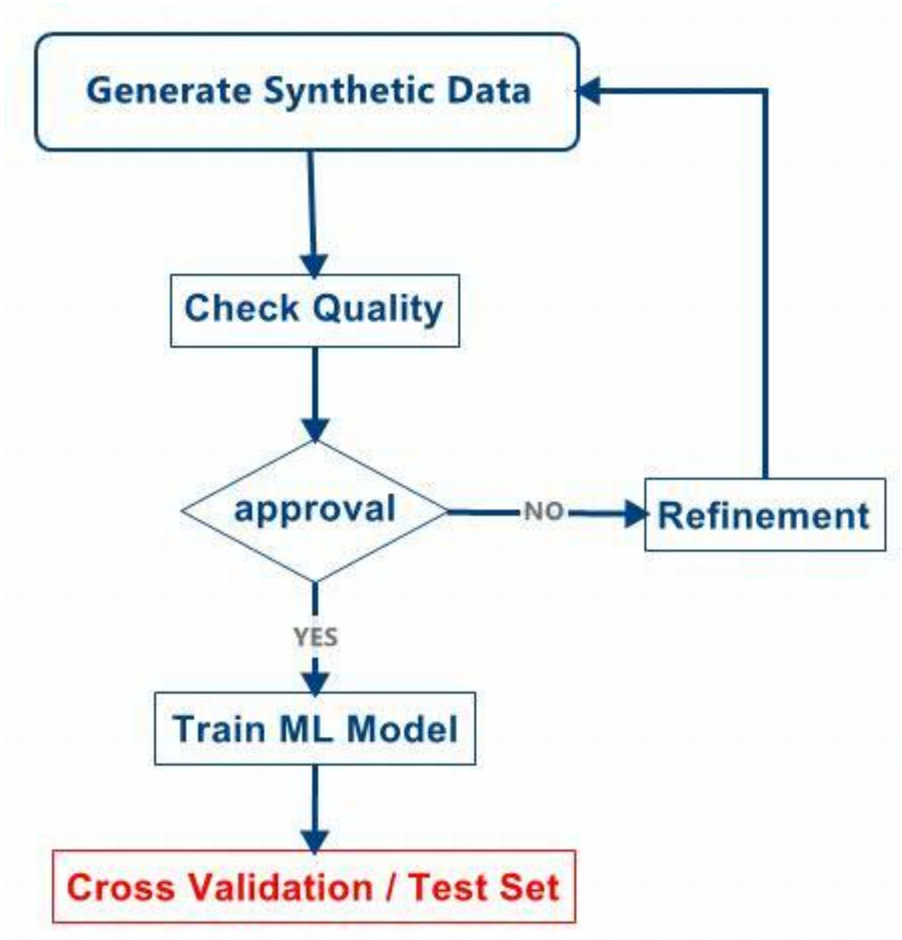
TRANSFER LEARNING / DOMAIN ADAPTION

- Fine-tuning: Train ml model with synthetic data in first iterations and with real data in further iterations
- Modifying the loss function when training on synthetic data :

L= $L_{problem} + L_{similarity}$ (e.g. Haeusser et al, ICCV 2017)

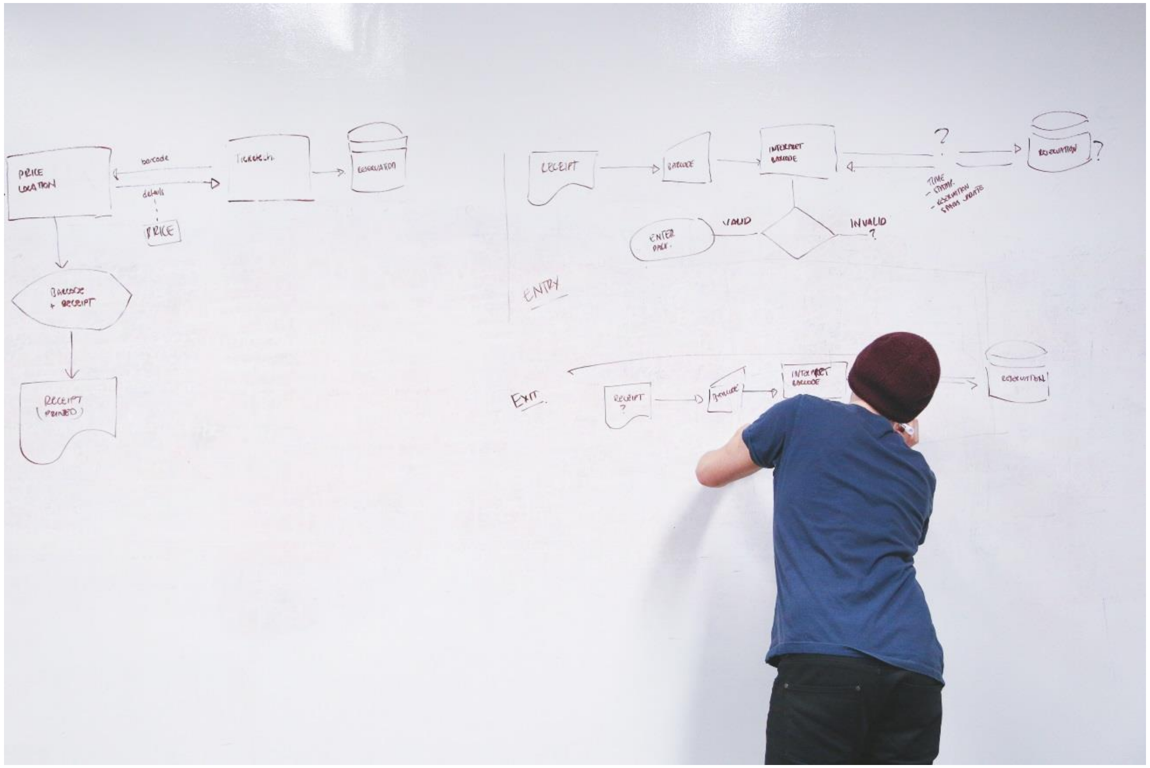


WORKFLOW


Generate Synthetic Data
Check Quality
approval
NO
Refinement
YES
Train ML Model
Cross Validation / Test Set

CROSS VALIDATION AND TESTING

- If there are corresponding pairs of synthetic and real data do not use one for training and the other one for testing
- Synthetic data in a test set is dangerous and should not be done



CONCLUSION

- Synthetic data can be helpful to increase performance of ml model in small data or imbalanced problems
- Gap between the statistical distributions of real and synthetic data needs to be checked

Our success story

- Going from a couple of thousand to more than a million data points using FEM simulations
- Huge increase in performance for automated data evaluation

[TRUST]

[PEOPLE] [INDUSTRIES]

[COMPETENCE]

[RELIABILITY] [TECHNOLOGY]

[INNOVATION]

[INDEPENDENT]

[CAN DO]

THANK YOU FOR JOINING THIS PRESENTATION.

ROSEN

empowered by technology

www.rosen-group.com