GLOBAL OPEN FINANCE

CENTRE OF EXCELLENCE

Reaping Real Social and Societal Benefits from Synthetic Financial Data

Nick Radcliffe
Chief Data Scientist

https://www.globalopenfinance.com/Synthetic-Data-Fintech2021/

Fintech 2021 • 16 September 2021 • Dynamic Earth, Edinburgh

GOFCoE

GLOBAL OPEN FINANCE CENTRE OF EXCELLENCE

- Aims safely to unlock the power of financial data for social and societal benefit without compromising the privacy or security of individuals, firms etc.
- Works with fintechs, banks, other financial institutions and government to produce economic insights and to make data safely available to researchers, within the University's Data Safe Haven
- Strong information governance & statistical disclosure controls
- Data only shared with clear legal basis; safe people; safe processes, safe places.
- Partners include banks, fintechs, credit bureaux, regulators and various levels of government
- Focus areas include impact of COVID-19 on UK Citizens and Businesses; Tackling Algorithmic Bias; Fintech enablement through Innovation Environment; Tackling social exclusion; and sustainability.

What?

Real Data

- Actual data describing real people/events/entities/businesses/transactions etc.

PROS:

- It's real *(though can still be inaccurate/biased/partial/misleading)*

CONS:

- Privacy (especially for personal data),
- Access, Restrictions,
- Ownership/Licenses/Rights/Restrictions
- May reflect real-world biases, inequalities etc.
- Limited to what has actually happened/known so some limitations on scenario planning/*What if?* etc.
  (e.g. climate change, policy change etc.)
- Sometimes limited volume

Synthetic Data

GLOBAL OPEN FINANCE
CENTRE OF EXCELLENCE

- “Fake” / “made up” / “synthesized” data, (usually) sharing some of the characteristics of some real data
- Sometimes talk about “synthetic doubles” of (real) datasets, which have many of the same characteristics as the real data, but which pertain entirely to fake/ imaginary/non-real people/events/entities/businesses/transactions.
- Usually, the ideal is to have “all” the same “statistical” properties, so that — for example — you can build almost identical models on the “fake” data as on the real data.
- Sometimes, you actually want to change some properties, e.g. to model scenarios, remove historical biases etc.

The adequacy/utility of a synthetic dataset depends on the intended use.

Whatever the claims, no synthetic double can ever be equivalent to the real data for all purposes.

But for in particular circumstances, synthetic data can be better than real data.

Why?

Why is GOFCoE Interested in Synthetic Data?

GLOBAL OPEN FINANCE CENTRE OF EXCELLENCE

- Innovation Environment — *We want to make realistic data available to fintechs for testing, development and validation without sharing real data*
- Our own Data Science Work — *"Safer data in a less locked-down environment"*
  - development of code/algorithms;
  - sharing of code for testing (with data);
  - potentially even results generation in some circumstances.
- Possibly
  - work on algorithmic bias — *"What if we hadn't discriminated in the past?"*
  - model robustness — *"What if we left the EU?"*
  - climate change — *"What if Earth became uninhabitable for humans?"*
  - financial crime etc. — *"What if someone tried this?"*

How?

Broad Approaches

GLOBAL OPEN FINANCE
CENTRE OF EXCELLENCE

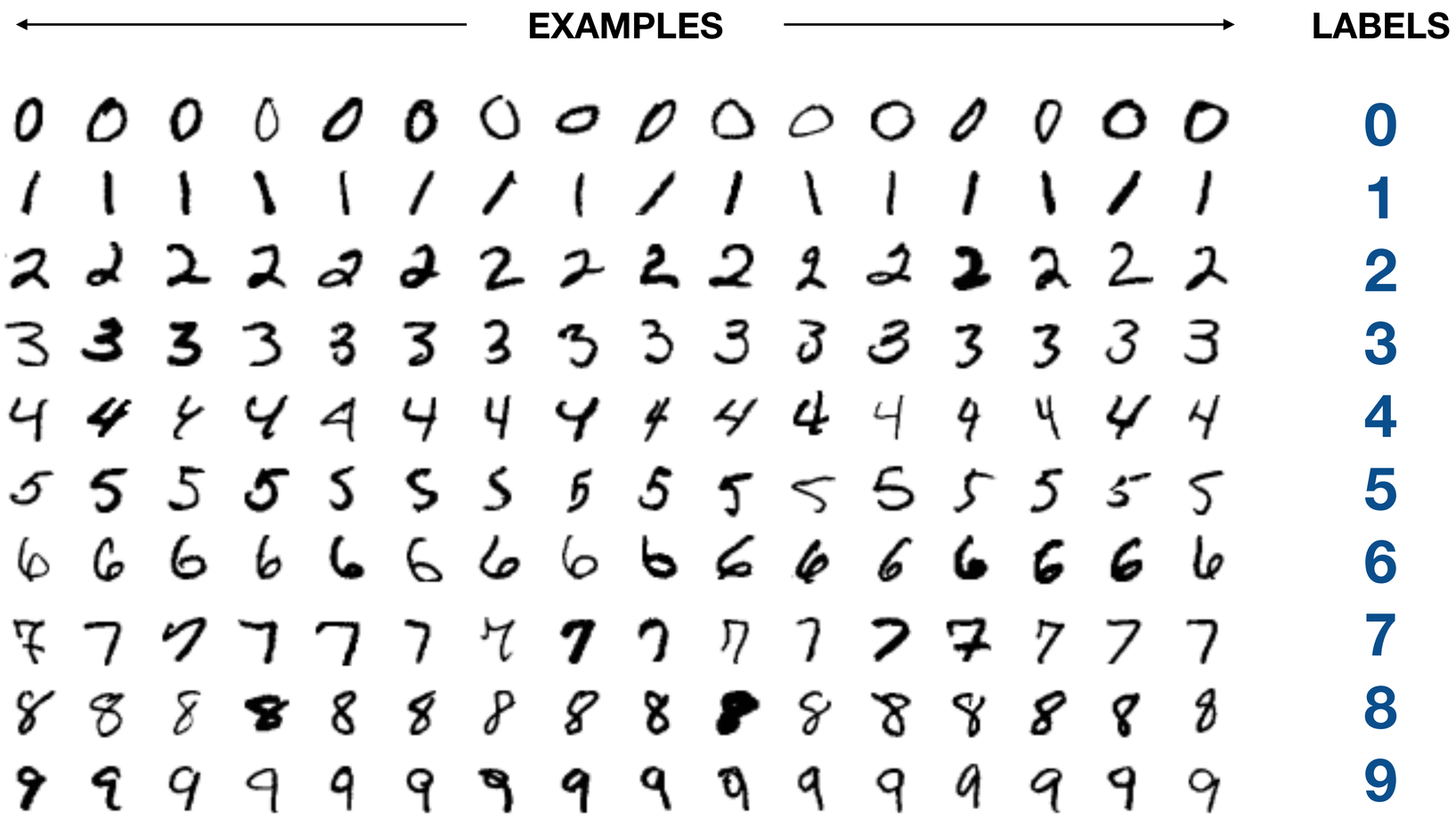
- Pure Random: *Really just copy field names and types, possibly ranges, and generate random stuff. OK for some code writing, testing, performance testing etc.*
- Constrained Data Generation: *Characterize the data with constraints etc. and match the constraints. More realistic, but misses correlations.*
- Simulation/“Agent-based Modelling”. *Create fake people/entities with characteristics and simulate their behaviour to generate transactions, events interactions etc. This is what we did before we had real data. Gets you further, but is hard work and only achieves limited realism.*
- Model-Based Generation (“synthetic doubles"). *Use machine learning (particularly neural nets/deep learning) to learn data and then “reverse the mapping” to generate synthetic data.*

Machine Learning

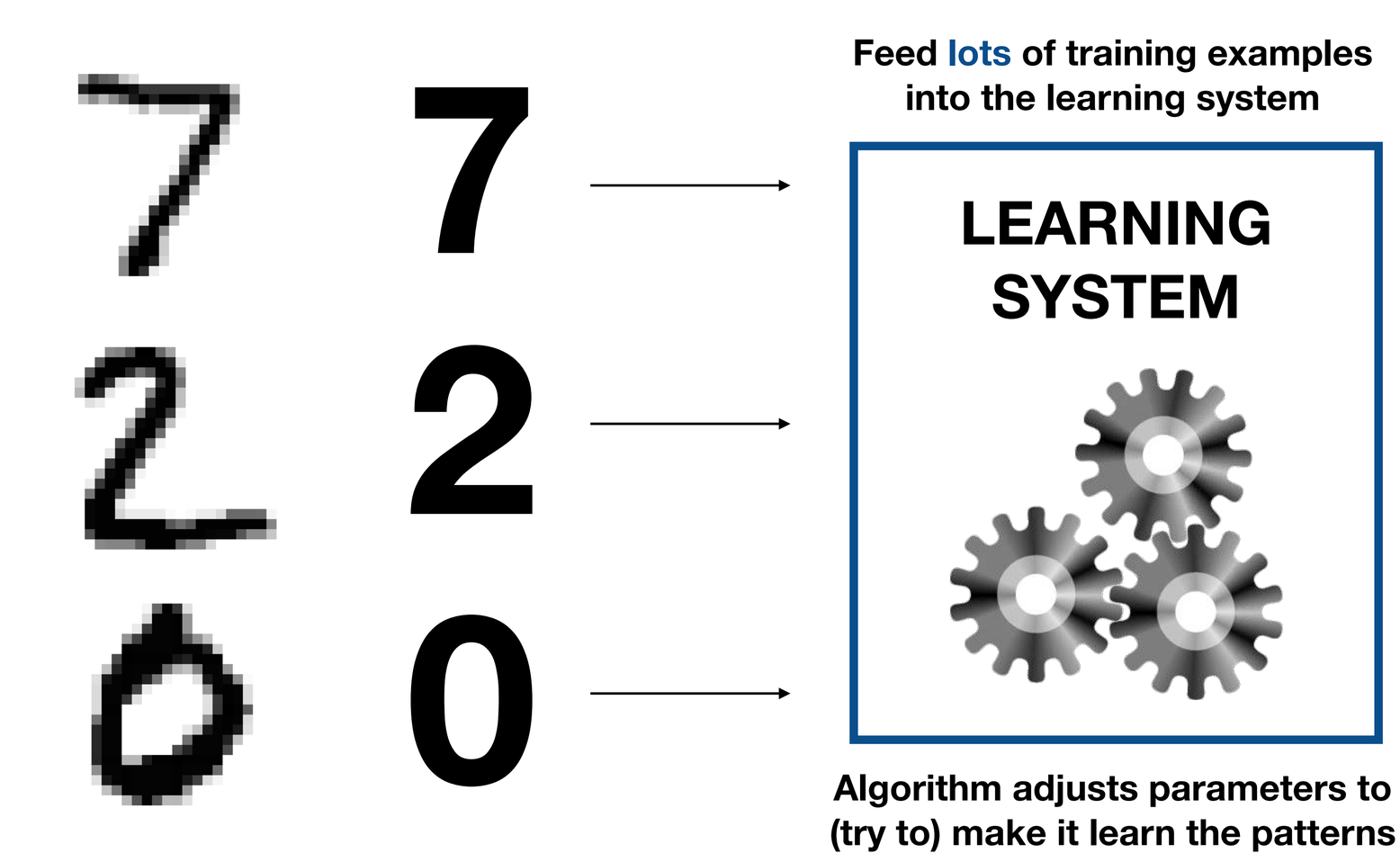
Supervised Learning

GLOBAL OPEN FINANCE
CENTRE OF EXCELLENCE

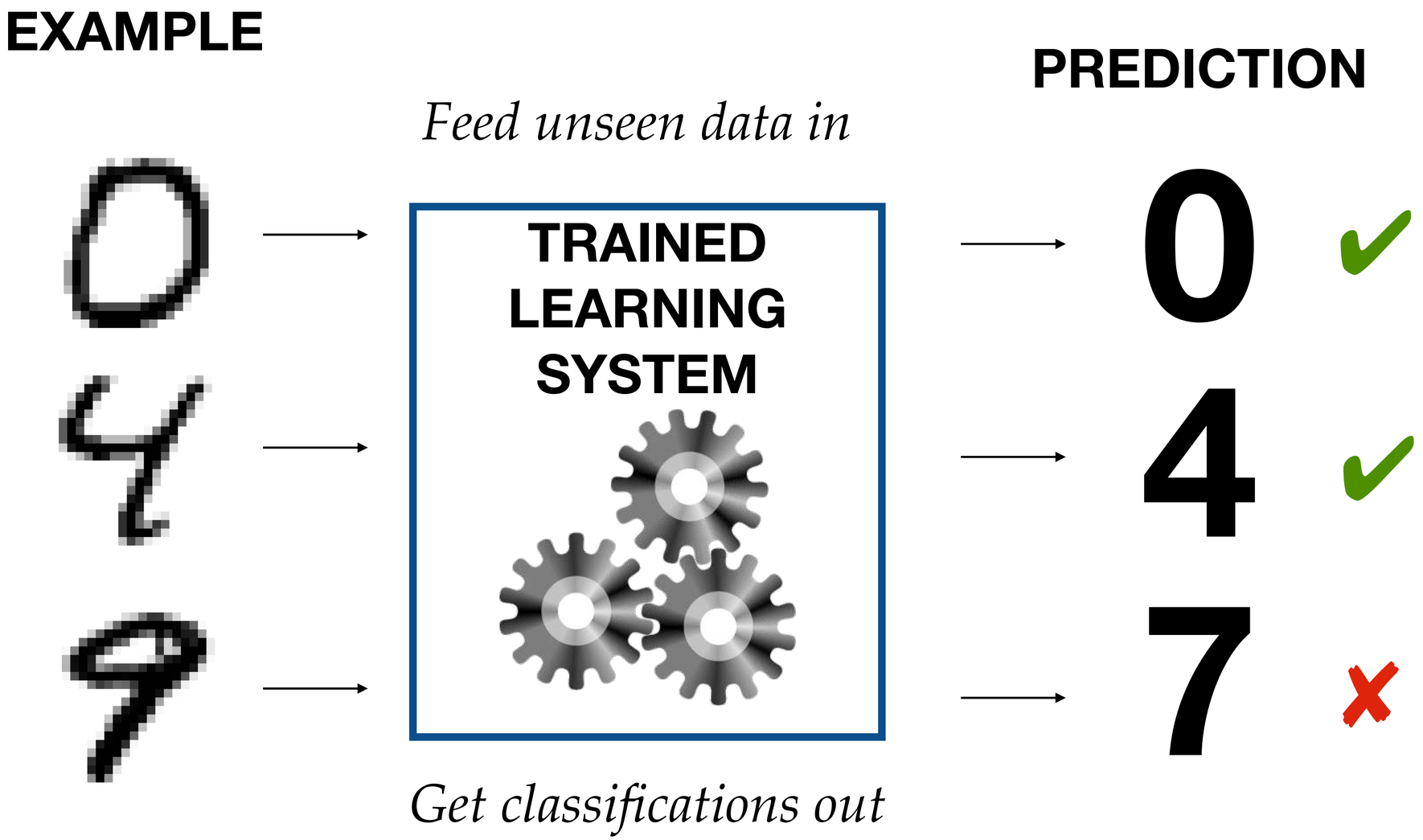
CORPUS OF "LABELLED" EXAMPLES



Training



Classification



Keep algorithmically twiddling the knobs until the network has learnt to do the classification sufficiently accurately!

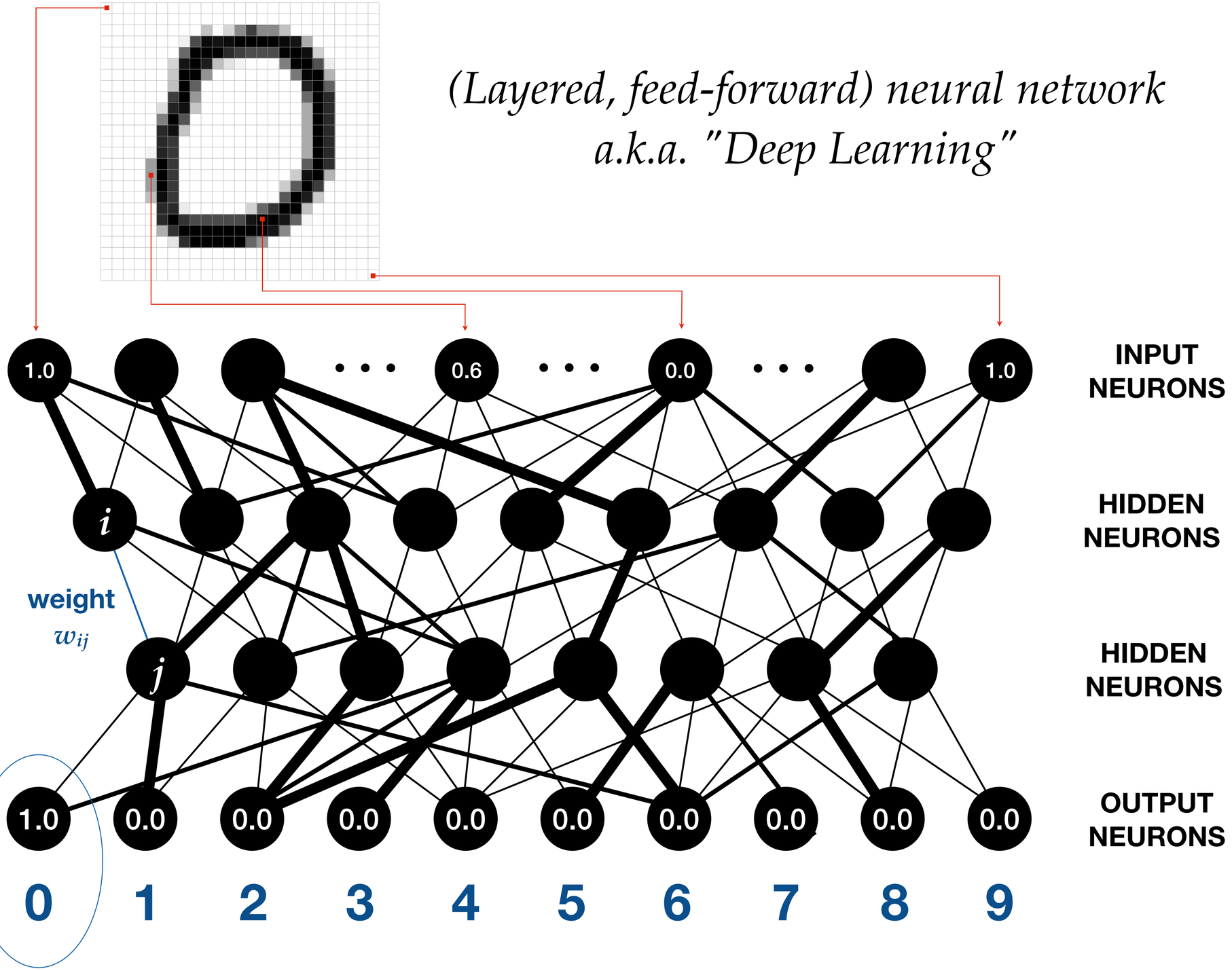
. . . BUT MAYBE YOU'D LIKE A LITTLE MORE DETAIL ON THAT "LEARNING" BIT ...

How to draw an owl

1.

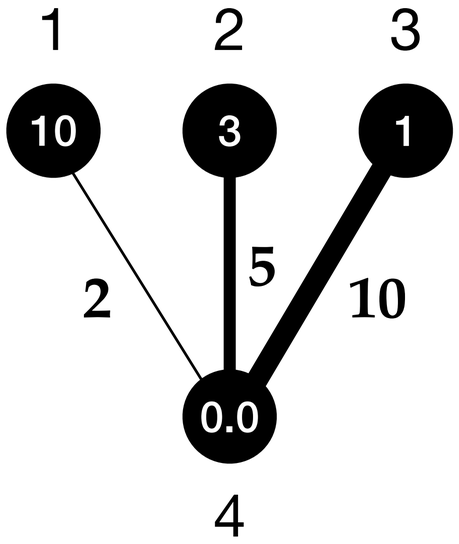
1. Draw some circles

(Layered, feed-forward) neural network
a.k.a. "Deep Learning"



FORWARD PASS

The output of each node is a function of a weighted sum of its inputs



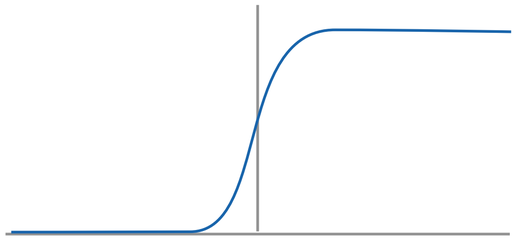
bias (optional)

$$v_j = f\left(\sum v_i w_{ij} + \Theta_j\right)$$

$$v_4 = f\left(v_1 w_{14} + v_2 w_{24} + v_3 w_{34}\right)$$

$$v_4 = f\left(10 \times 2 + 3 \times 5 + 1 \times 10\right)$$

$$f(x) = \frac{1}{1 + \exp(-kx)}$$



Back Propagation

BACKWARD PASS ↑

An error is calculated for each output node, and an adjustment for each weight is calculated to reduce that error.

Errors "propagate backwards" through the network (hence "back propagation")

$$\Delta w_{ij} = \alpha(t_j - v_j)f'(h_j)v_i$$

- Δw_{ij}: CHANGE IN WEIGHT ij
- α: LEARNING RATE
- $(t_j - v_j)$: ERROR AT NODE j
- f': DERIVATIVE OF f
- (h_j): SUM OF INPUTS FOR NODE j
- v_i: ACTUAL OUTPUT NODE j

Back Propagation

BACKWARD PASS ↑

An error is calculated for each output node, and an adjustment for each weight is calculated to reduce that error.

Errors "propagate backwards" through the network (hence "back propagation")

CALCULUS!

WE ❤

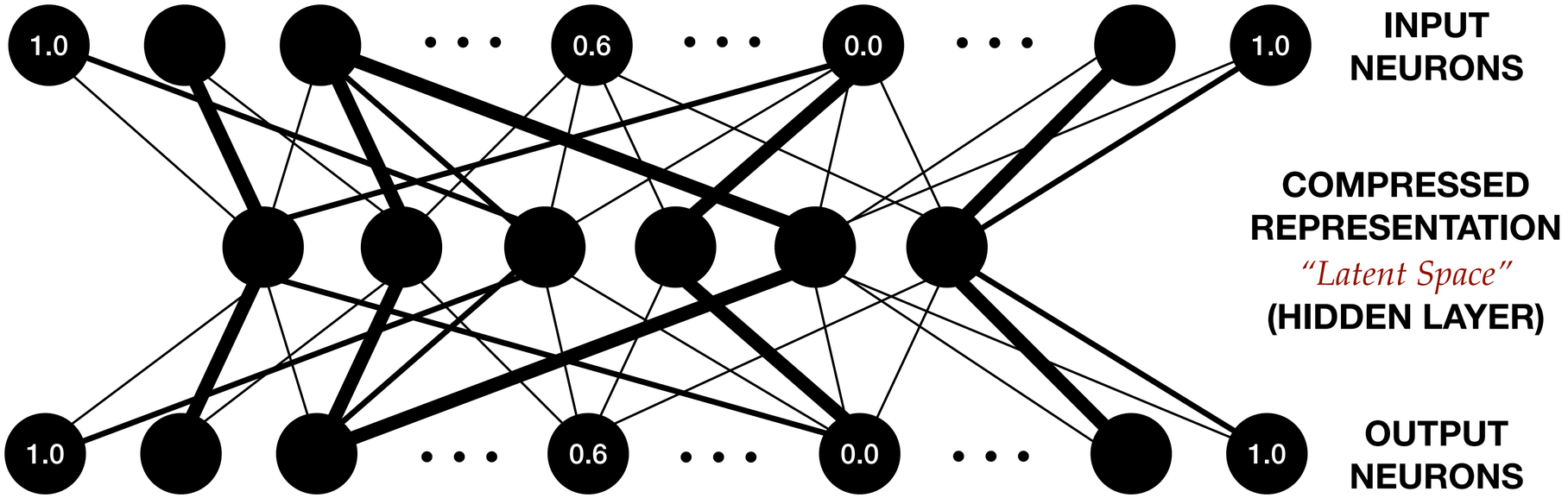
$$\Delta w_{ij} = \alpha(t_j - v_j)f'(h_j)v_i$$

- Δw_{ij}: CHANGE IN WEIGHT ij
- α: LEARNING RATE
- $(t_j - v_j)$: ERROR AT NODE j
- f': DERIVATIVE OF f
- (h_j): SUM OF INPUTS FOR NODE j
- v_i: ACTUAL OUTPUT NODE j

Generative Adversarial Networks

Autoencoder

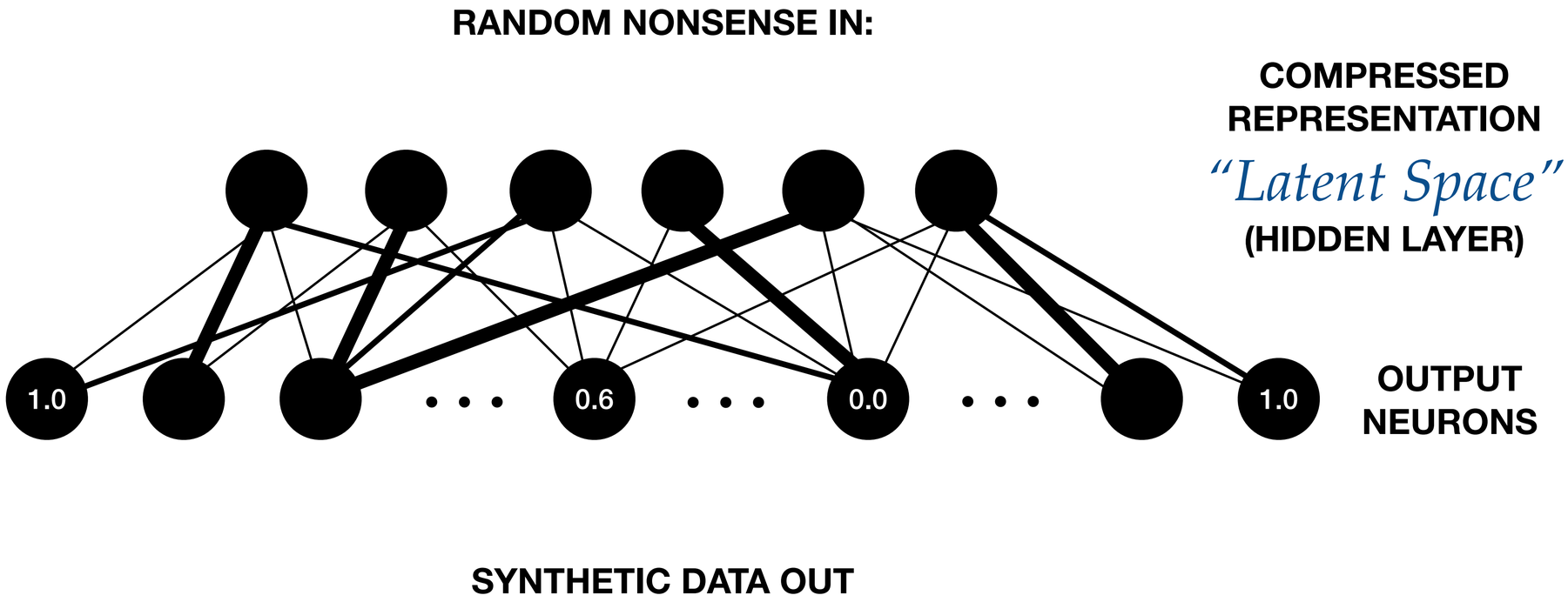
Train a network to reproduce input patterns on the output nodes



What's the point in that?
Well, the compressed representation captures the correlations in the input data.
If you throw away the input layer and just directly set patterns in the hidden layer,
it will tend to generate outputs that are similar to real inputs.

Map from Latent Space

(Very roughly)

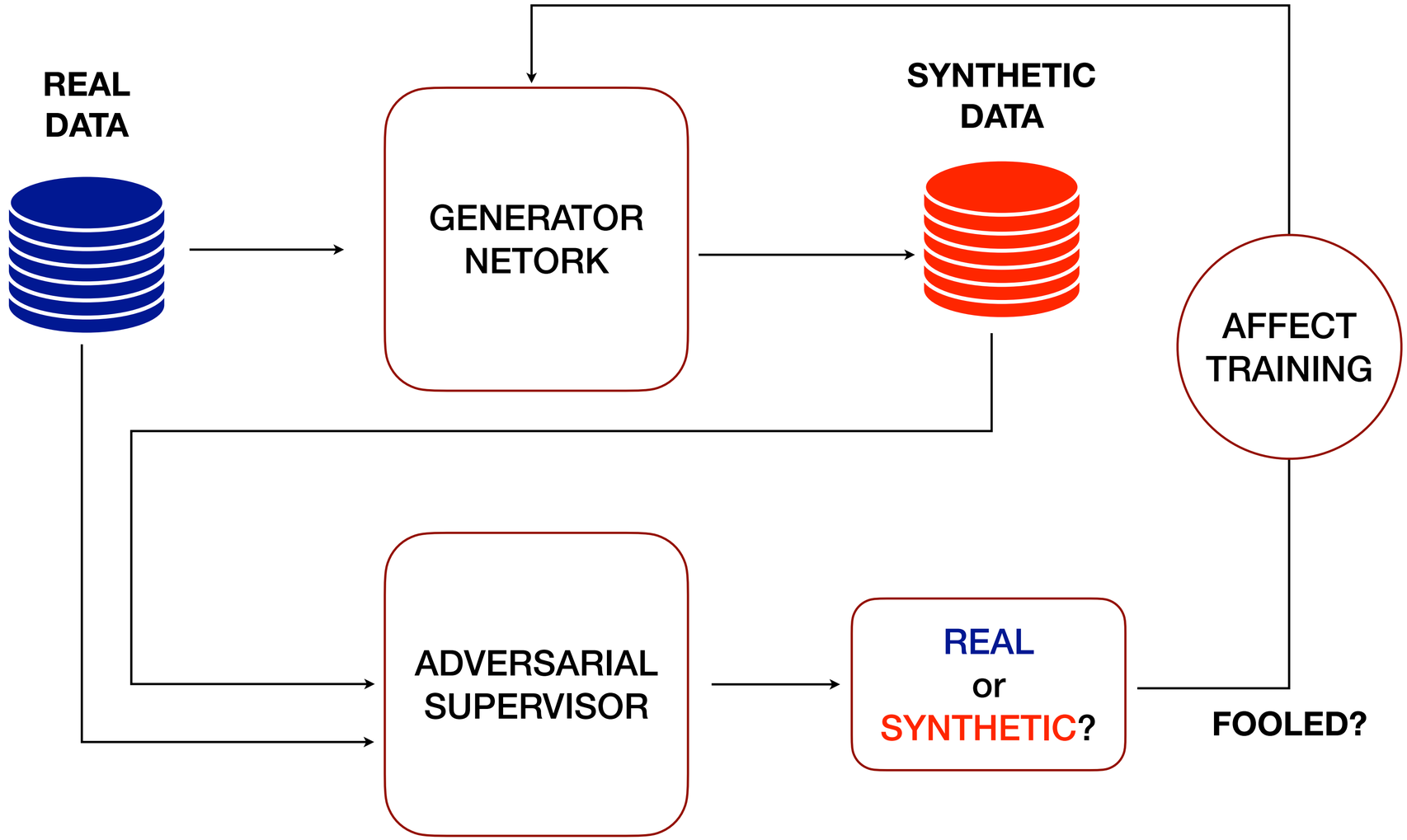


Generative Adversarial Networks (GANs)

GLOBAL OPEN FINANCE CENTRE OF EXCELLENCE

- Use two models — often both neural networks — one as a generator and one as a “supervisor”.
- The supervisor model is fed real data and outputs from the synthetic data generator and tries to distinguish real and synthetic data. Essentially, it is “rewarded” when it correctly identifies fake data as fake or real data as real, and is penalised when it gets it wrong.
- The generative network learns the mapping from latent space to plausible outputs, updating its weights to reward it when it fools the supervisor network, and penalise it when the supervisor network gets it right.
- The two “co-evolve”, driving each other to ever better performance.

Generative Adversarial Networks (GANs)



Issues (It's not a Free Lunch)

GLOBAL OPEN FINANCE
CENTRE OF EXCELLENCE

- Even if the data generator gets to the point where the supervisor network can't distinguish its synthetic data from real data, that doesn't mean *nothing* can.
- You have to worry about leakage of real data into the synthetic data generator. (How do you know it won't suddenly synthesize Nicola Sturgeon?)
- This is all pretty bleeding edge: it's not clear whether it's ready for the prime time. We'll need a mix of techniques and they won't work in all circumstances.
- Even if the approach works well for fairly simple single tables, and picks up most of the key correlations, strong relations between fields (constraints) or rows are a challenge, and linked datasets are even harder for current approaches (GANS, Bayesian Networks etc.)

Why GOFCoE Is Well Placed in this

- GOFCoE can (perhaps) create synthetic doubles of (some of) its data resources and make them available to people for model building etc.
- People can then build their models, reports, systems etc. with them, but won't be certain how similar the synthetic data is to the real data
- GOFCoE can potentially then import the models/reporting systems/whatever into the Safe Haven and test them against the real data, and check that the results on the synthetic data are consistent with those on the real data.

GOFCoE's Main Interest

- GOFCoE is currently learning about synthetic data, evaluating both open source/academic implementations (SDV, Synthpop) and commercial systems from (inter alia) Hazy & Diveplane.
- GOFCoE's primary interest is in evaluating the safety and utility of synthetic data generators and getting to a position where we understand what data we can safely synthesize and share while giving strong and well-grounded assurances about privacy preservation.
- There is undoubtedly hype around synthetic data at the moment, and we do not believe for a moment that any current system will take in data and safely produce synthetic doubles that are safe and broadly useful; but over time, we expect the power, safety and utility of synthetic data to increase.
- We intend to be there, using it to help unlock the power for financial data to do good safely.

GLOBAL OPEN FINANCE

CENTRE OF EXCELLENCE

https://www.globalopenfinance.com/Synthetic-Data-Fintech2021/

(zero, not "oh")

@njr0

@GOFCoE

n.radcliffe@ed.ac.uk