Generating Tabular Data Using Generative Adversarial Networks

By: Joe and Yunus

English

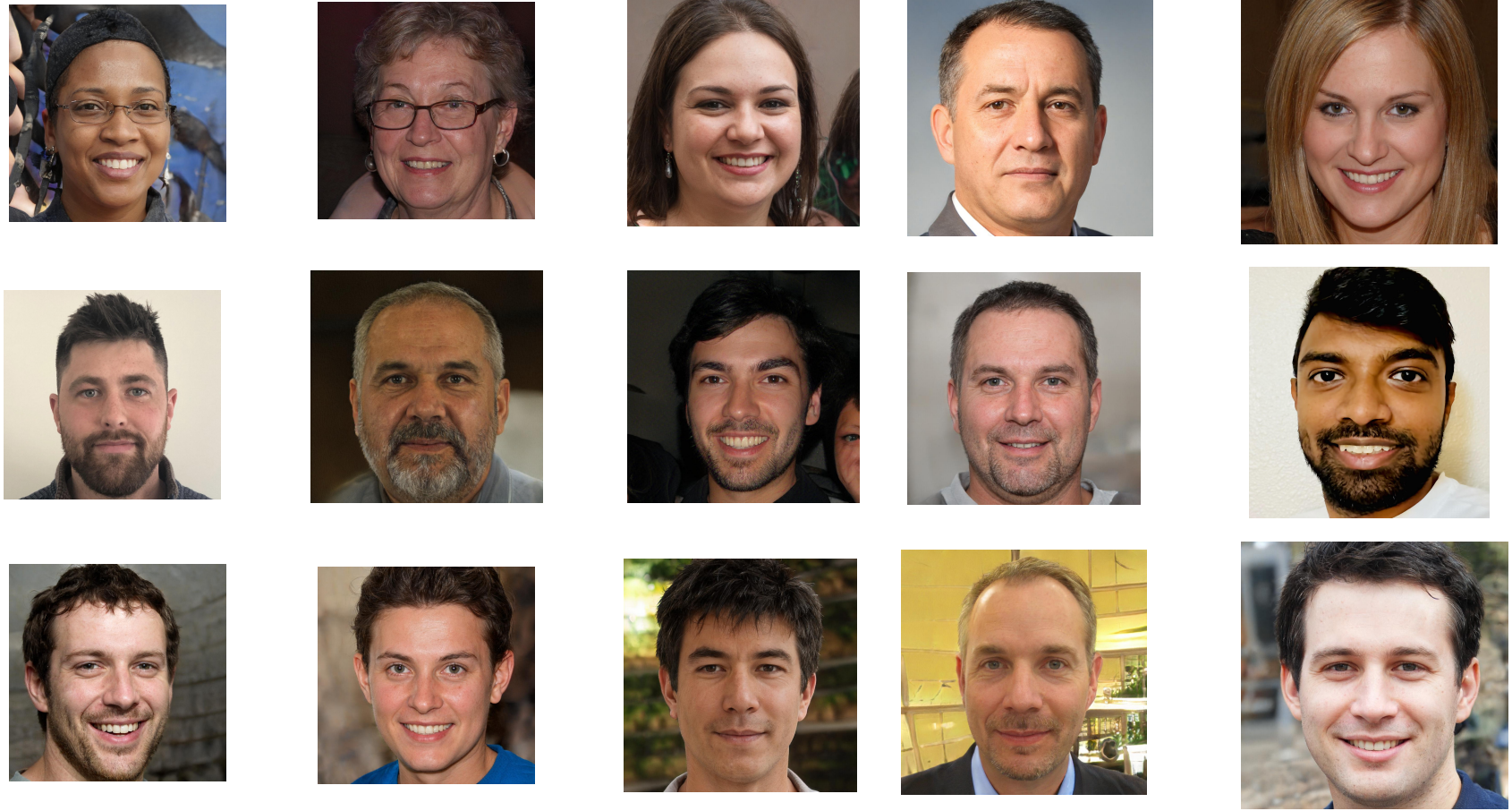
Random Face Generator (This Person Does Not Exist)

Generate random human face in 1 click and download it! AI generated fake person photos: man, woman or child.

this-person-does-not-exist.com/enl

Refresh Image | Download

Motivation: Let's look at faces generated by this-person-does-not-exist.com



GANs make astonishing things, but…

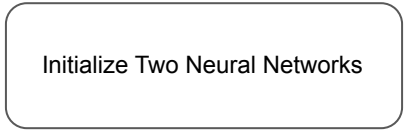
Can they generate tabular data?

Fun Facts

- Generator and Discriminator are pitted against each other (Adversaries)
- Discriminator is trained on real data from the domain set to improve its distinguishing prowess
- Generator produces fake samples, aiming to fool the discriminator through iterations
- Model is successful when the discriminator cannot distinguish between real and fake samples

Fun Opinions

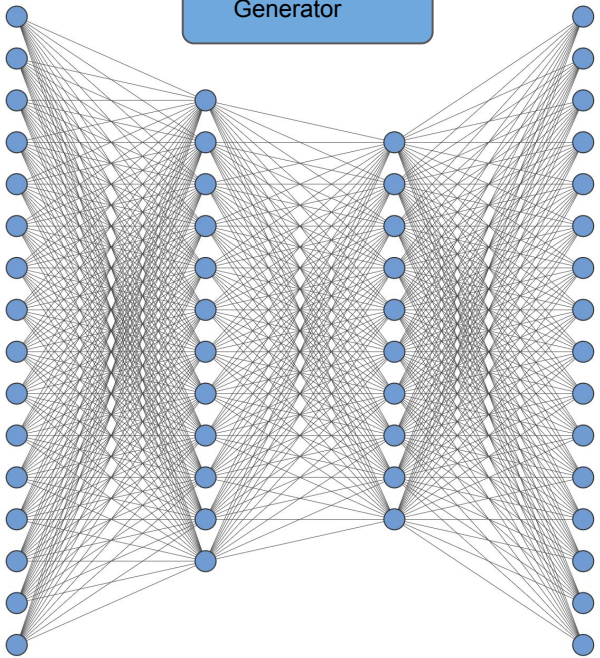
- GANs are a very successful at generating images
- We think there are a substantial number of unrealized applications for this
- Leveraging their aptitude to learn and generate images GANs as they almost certainly have the potential to generate tabular data
- Generating additional data could be useful in any area requiring additional data such as clinical trials, logistics, finance, insurance, and much more

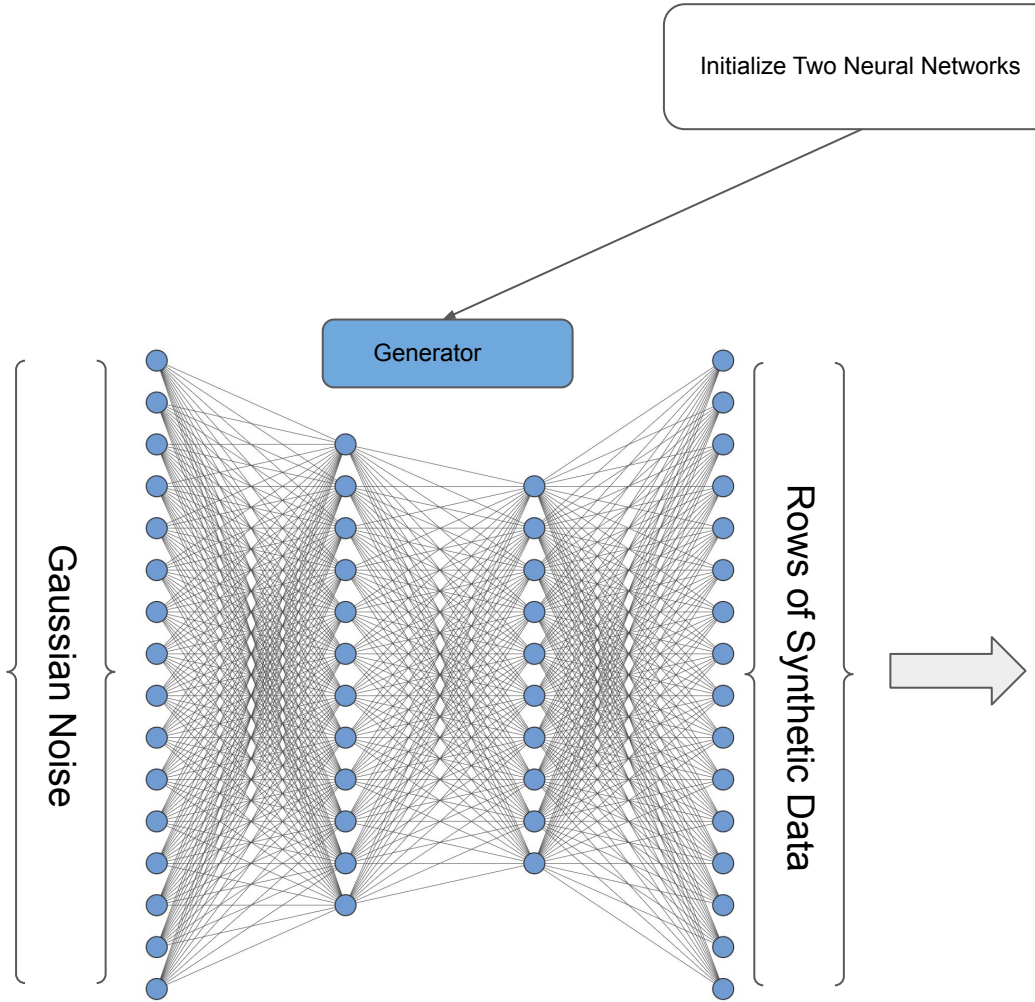

Initialize Two Neural Networks

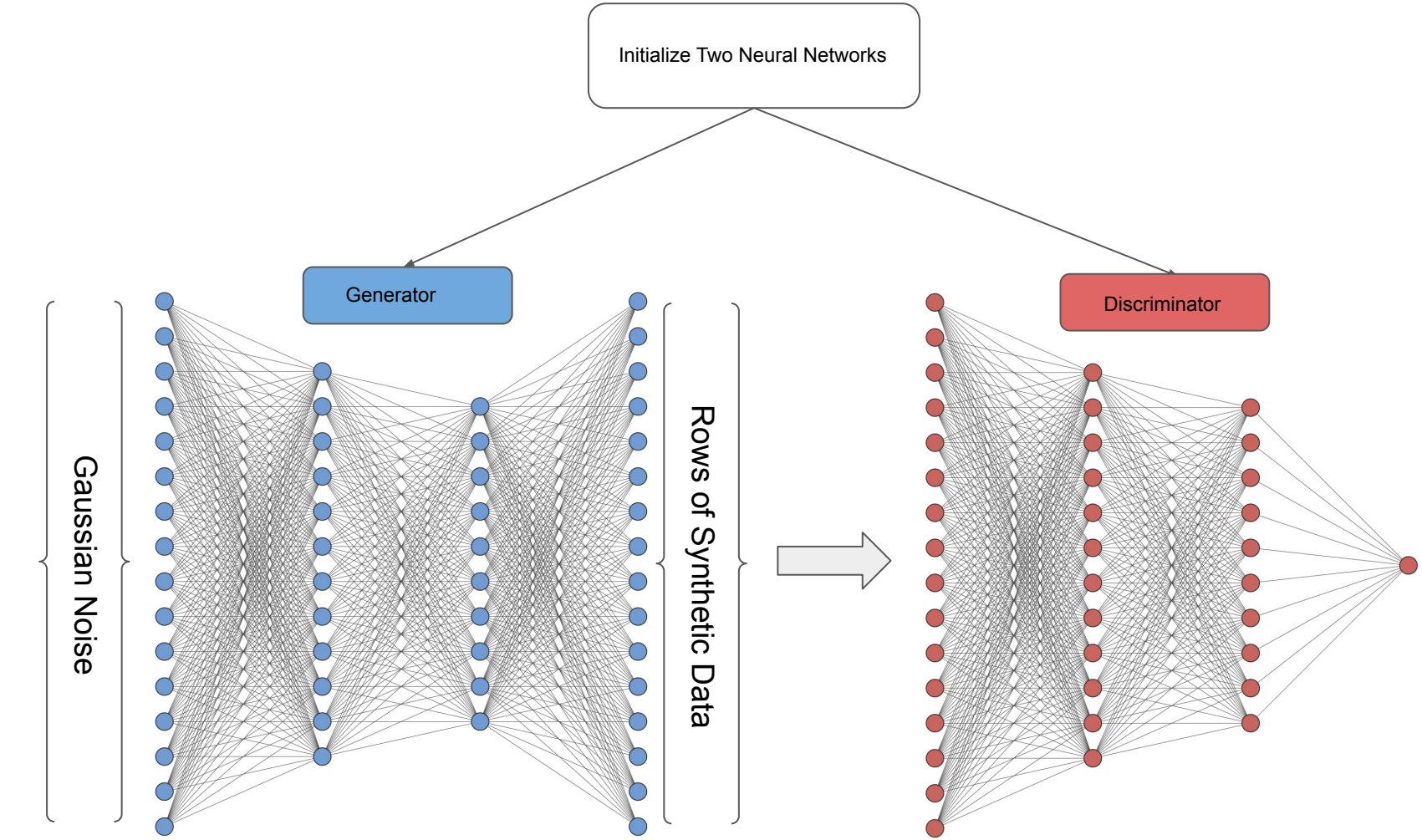
Initialize Two Neural Networks

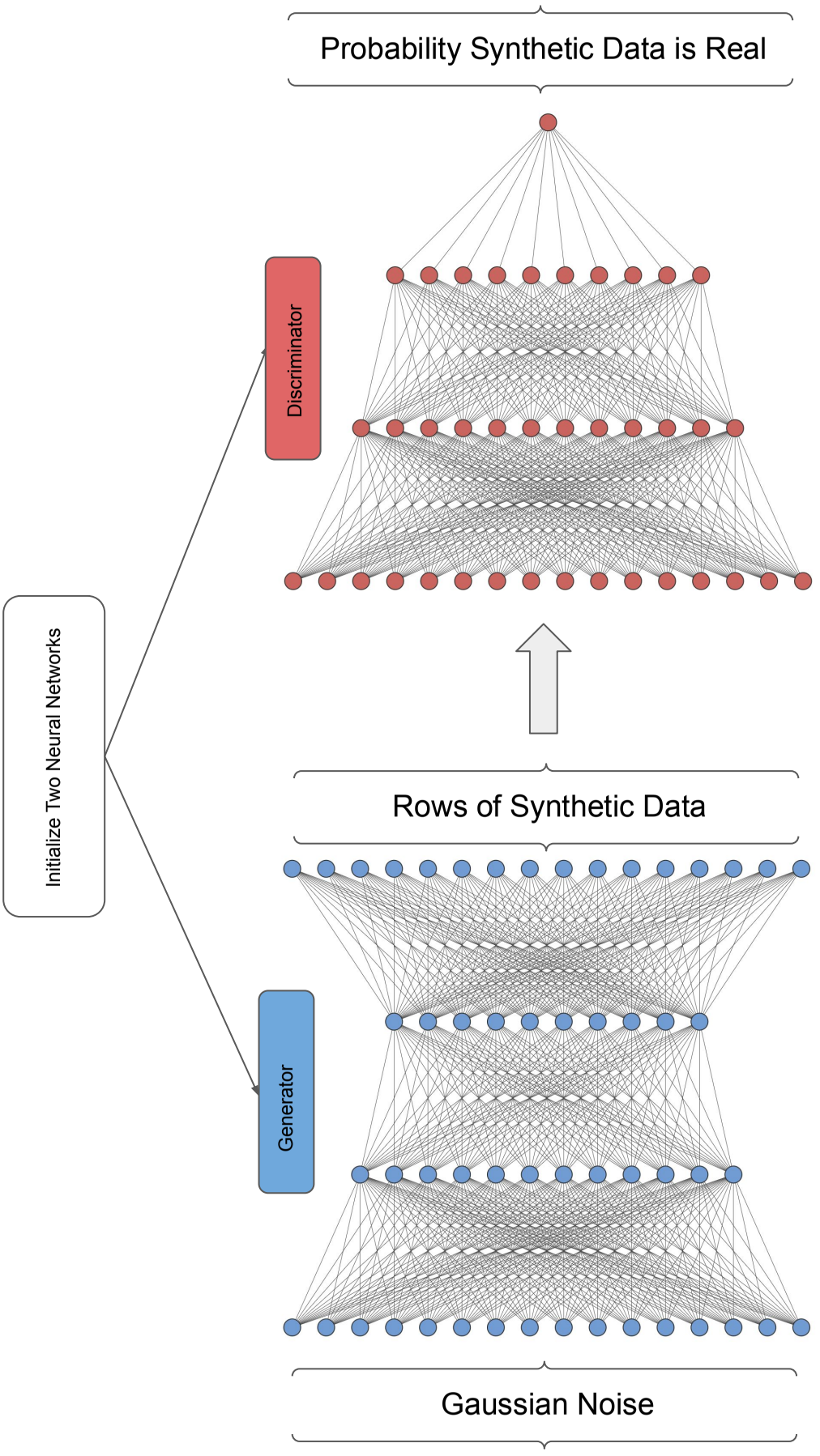
Generator

Gaussian Noise




Initialize Two Neural Networks
Generator
Gaussian Noise
Rows of Synthetic Data


Initialize Two Neural Networks
Generator
Discriminator
Gaussian Noise
Rows of Synthetic Data


Probability Synthetic Data is Real
Discriminator
Initialize Two Neural Networks
Rows of Synthetic Data
Generator
Gaussian Noise

Training a GAN to Generate Tabular Data

Minimal Example				
rdelay	age	rasp3	td	expdd
-0.1526527	-1.2121831	-0.4472136	-0.4923660	1.4494064
-0.6383658	0.2020305	0.6708204	-0.0820610	-0.0607479
1.6514244	-0.8081220	-1.5652476	-0.3282440	-0.1689550
-0.9159161	1.2121831	0.6708204	-0.8206099	-1.3592337
0.0555101	0.6060915	0.6708204	1.7232809	0.1395303

Minimal Example

Label Real Data

rdelay	age	rasp3	td	expdd	is_real
-0.1526527	-1.2121831	-0.4472136	-0.4923660	1.4494064	1
-0.6383658	0.2020305	0.6708204	-0.0820610	-0.0607479	1
1.6514244	-0.8081220	-1.5652476	-0.3282440	-0.1689550	1
-0.9159161	1.2121831	0.6708204	-0.8206099	-1.3592337	1
0.0555101	0.6060915	0.6708204	1.7232809	0.1395303	1

Minimal Example

Generate fake data, append to real, and train GAN

rdelay	age	rasp3	td	expdd	is_real
-0.1526527	-1.2121831	-0.4472136	-0.4923660	1.4494064	1
-0.6383658	0.2020305	0.6708204	-0.0820610	-0.0607479	1
1.6514244	-0.8081220	-1.5652476	-0.3282440	-0.1689550	1
-0.9159161	1.2121831	0.6708204	-0.8206099	-1.3592337	1
0.0555101	0.6060915	0.6708204	1.7232809	0.1395303	1
-1.2081298	-0.1415998	-0.9334291	-0.3736803	0.7781137	0
-0.6549551	0.4887839	2.1236970	1.0688486	-0.9533416	0
0.6133733	0.5632056	0.9090887	-0.2830115	0.2700140	0
-1.2241351	-0.9369381	-0.7058276	-0.5578974	-1.8549658	0
-0.3605781	1.3908176	-1.1527232	-1.4189711	0.1101485	0

Minimal Example

Estimated probability data is real

	rdelay	age	rasp3	td	expdd	probability_is_real
Real Data	-0.1526527	-1.2121831	-0.4472136	-0.4923660	1.4494064	0.9679279
	-0.6383658	0.2020305	0.6708204	-0.0820610	-0.0607479	0.9929178
	1.6514244	-0.8081220	-1.5652476	-0.3282440	-0.1689550	0.9890847
	-0.9159161	1.2121831	0.6708204	-0.8206099	-1.3592337	0.9710315
	0.0555101	0.6060915	0.6708204	1.7232809	0.1395303	0.9555076
Synthetic Data	-1.5047081	0.4131216	0.7210116	0.6008804	-1.5889401	0.0487107
	1.5336657	-0.5486535	1.5022655	-0.3632001	0.1820924	0.0209049
	1.1543272	1.2725888	2.2691895	1.0132900	1.9007278	0.0337531
	2.0620126	-0.5945645	-0.5089318	0.4417439	1.1727553	0.0344016
	-0.6610273	-0.5727584	-0.6632418	0.1115845	-0.7554694	0.0385241

Minimal Example

Estimated probability data is real

	rdelay	age	rasp3	td	expdd	probability_is_real
Real Data	-0.1526527	-1.2121831	-0.4472136	-0.4923660	1.4494064	0.9664078
	-0.6383658	0.2020305	0.6708204	-0.0820610	-0.0607479	0.9579901
	1.6514244	-0.8081220	-1.5652476	-0.3282440	-0.1689550	0.9872691
	-0.9159161	1.2121831	0.6708204	-0.8206099	-1.3592337	0.9968588
	0.0555101	0.6060915	0.6708204	1.7232809	0.1395303	0.9848552
Synthetic Data	1.2298782	0.1595506	-0.3525315	-0.4391252	-0.0648598	0.1573992
	0.2666857	0.8717755	0.4242235	0.0354590	2.8259816	0.1935930
	0.4419240	0.6767167	-0.0340592	0.6599349	-0.0367753	0.1456098
	0.1619623	-0.0378117	-0.9497960	-0.1654065	1.1924868	0.2089608
	-1.1857709	-0.6975662	1.3239881	2.2340208	-0.6295113	0.2273929

Minimal Example

Estimated probability data is real

	rdelay	age	rasp3	td	expdd	probability_is_real
Real Data	-0.1526527	-1.2121831	-0.4472136	-0.4923660	1.4494064	0.9632579
	-0.6383658	0.2020305	0.6708204	-0.0820610	-0.0607479	0.9795990
	1.6514244	-0.8081220	-1.5652476	-0.3282440	-0.1689550	0.9673142
	-0.9159161	1.2121831	0.6708204	-0.8206099	-1.3592337	0.9646654
	0.0555101	0.6060915	0.6708204	1.7232809	0.1395303	0.9695260
Synthetic Data	-1.7909718	0.1406559	-0.9071052	-1.2686191	-1.1364248	0.4309019
	-2.1082996	-1.8685017	2.1539457	-0.6920305	0.1153666	0.3207892
	1.2230039	1.8194209	-0.9003389	0.2042871	0.5771646	0.4275672
	-0.4258058	0.6174860	0.0180184	0.6808070	-0.6989342	0.3991783
	0.2051817	1.1343274	-0.1842599	1.4712985	0.7347378	0.3174398

Minimal Example

Estimated probability data is real

rdelay	age	rasp3	td	expdd	probability_is_real
-0.1526527	-1.2121831	-0.4472136	-0.4923660	1.4494064	0.9925000
-0.6383658	0.2020305	0.6708204	-0.0820610	-0.0607479	0.9977413
1.6514244	-0.8081220	-1.5652476	-0.3282440	-0.1689550	0.9833075
-0.9159161	1.2121831	0.6708204	-0.8206099	-1.3592337	0.9935970
0.0555101	0.6060915	0.6708204	1.7232809	0.1395303	0.9867053
0.1351111	1.0289204	-1.7491802	1.1701784	0.8917155	0.7776264
0.2644418	0.3564599	-0.6883506	-2.2317054	0.6318882	0.6469500
0.6625127	0.6405424	0.2892740	0.5008004	0.2006482	0.8975241
-0.5822321	0.2885843	1.1049722	-1.3624818	-0.3435825	0.5485536
-0.9422829	0.1078699	0.2500177	-1.6098746	1.6449217	0.8015168

Real Data

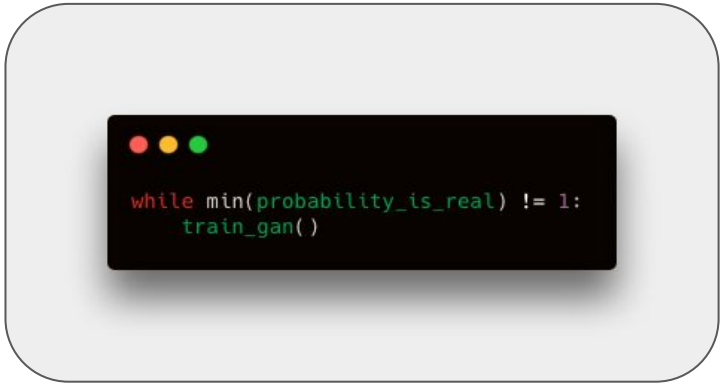
Synthetic Data

Minimal Example

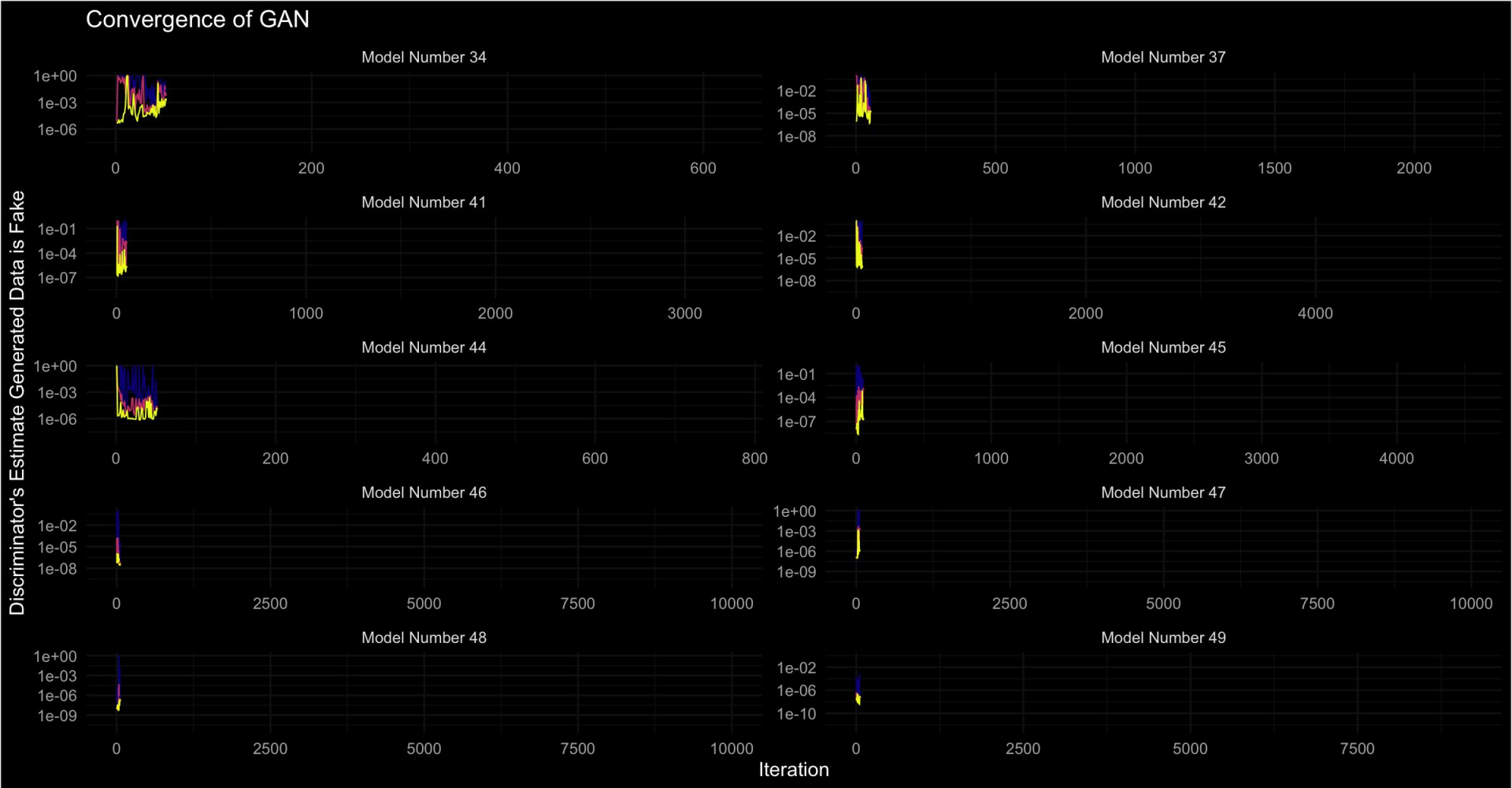
Train GAN until generated data is indistinguishable from real data

Real Data

Synthetic Data



probability_is_real
0.9998925
0.9926733
0.9979285
0.9902920
0.9924870
0.9971239
0.9975592
0.9950743
0.9940353
0.9910127

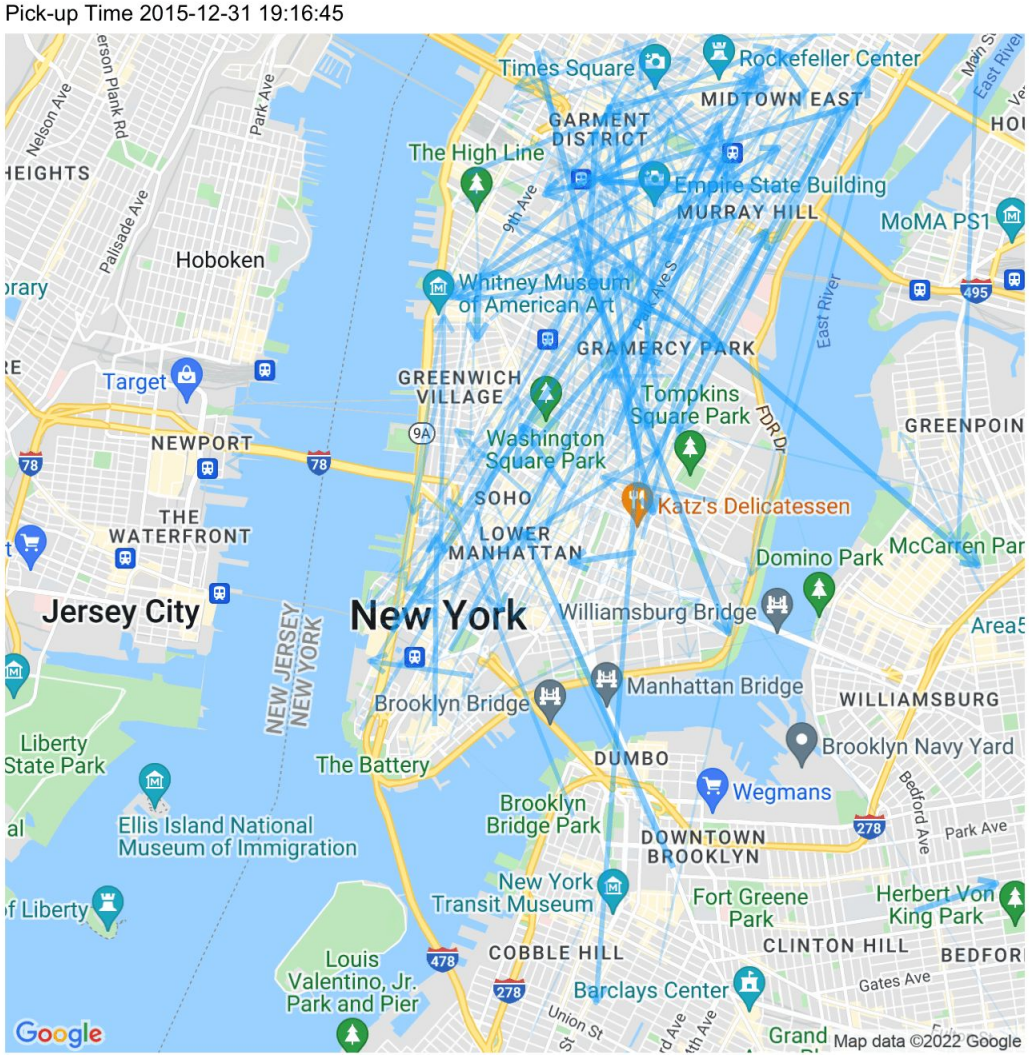

Convergence of GAN
Model Number 34
Model Number 37
Model Number 41
Model Number 42
Model Number 44
Model Number 45
Model Number 46
Model Number 47
Model Number 48
Model Number 49
Discriminator's Estimate Generated Data is Fake
Iteration
Summary Statistic
Max
Mean
Min

New York City Taxi data.

Shown is a small subset of taxi rides in the year 2016

- Pick up time and location
- Drop-off time and location

Permits filtering on simple heuristics and visual inspection.



Challenges

- Plausible data are mixed in with a significant amount of non plausible data. For 14th month, and -5th day.

Substantial filtering on heuristics.

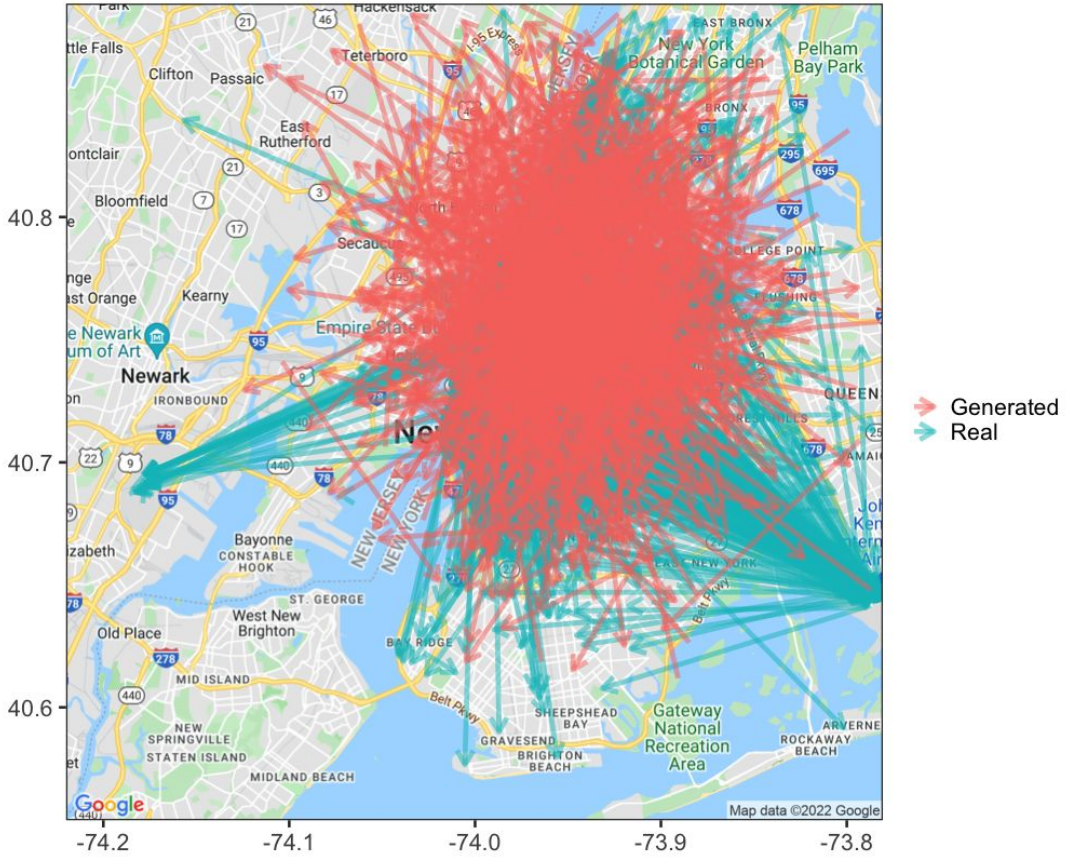
- Trained GANs generate only several distinct data points.

Retrain GAN for each generated data point

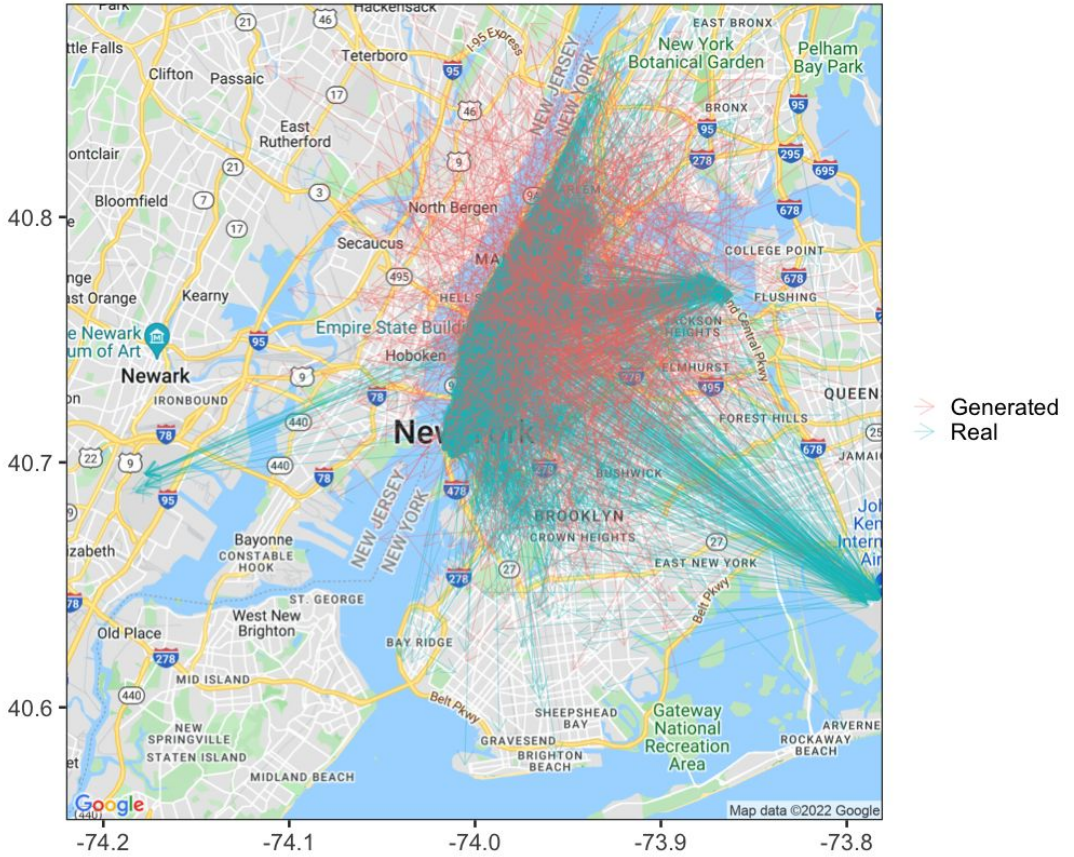
year_pickup_datetime	quarter_pickup_datetime	month_pickup_datetime	dayofweek_pickup_datetime
2110.929	96.17413	490.38092	574.53296
2123.7434	119.4954	589.7978	755.8383
2124.5603	120.58361	595.75085	724.0522
2125.4766	105.55681	536.96686	657.0043

year_pickup_datetime	month_pickup_datetime	dayofweek_pickup_datetime	dayofmonth_pickup_datetime
2017	4	5	16
2017	4	5	16
2017	4	5	16
2017	4	5	16

- Generated data follow a different structure. In the real data, a significant number of taxi rides are from the airport to midtown. In the generated data most of the rides are from midtown to other areas in midtown.
- Curation is required, limiting applications to visual inspection, however a large portion of these are plausible.



- Generated data follow a different structure. In the real data, a significant number of taxi rides are from the airport to midtown. In the generated data most of the rides are from midtown to other areas in midtown.
- Curation is required, limiting applications to visual inspection, however a large portion of these are plausible.



Challenges remaining

- Uses significant resources, likely competitive with bootstrapping
- ideally, this would be something trained once
- Curation, although common in GANs, is not ideal. This limits the use cases
- Requires different architecture for different data sets.

Conclusion, further study

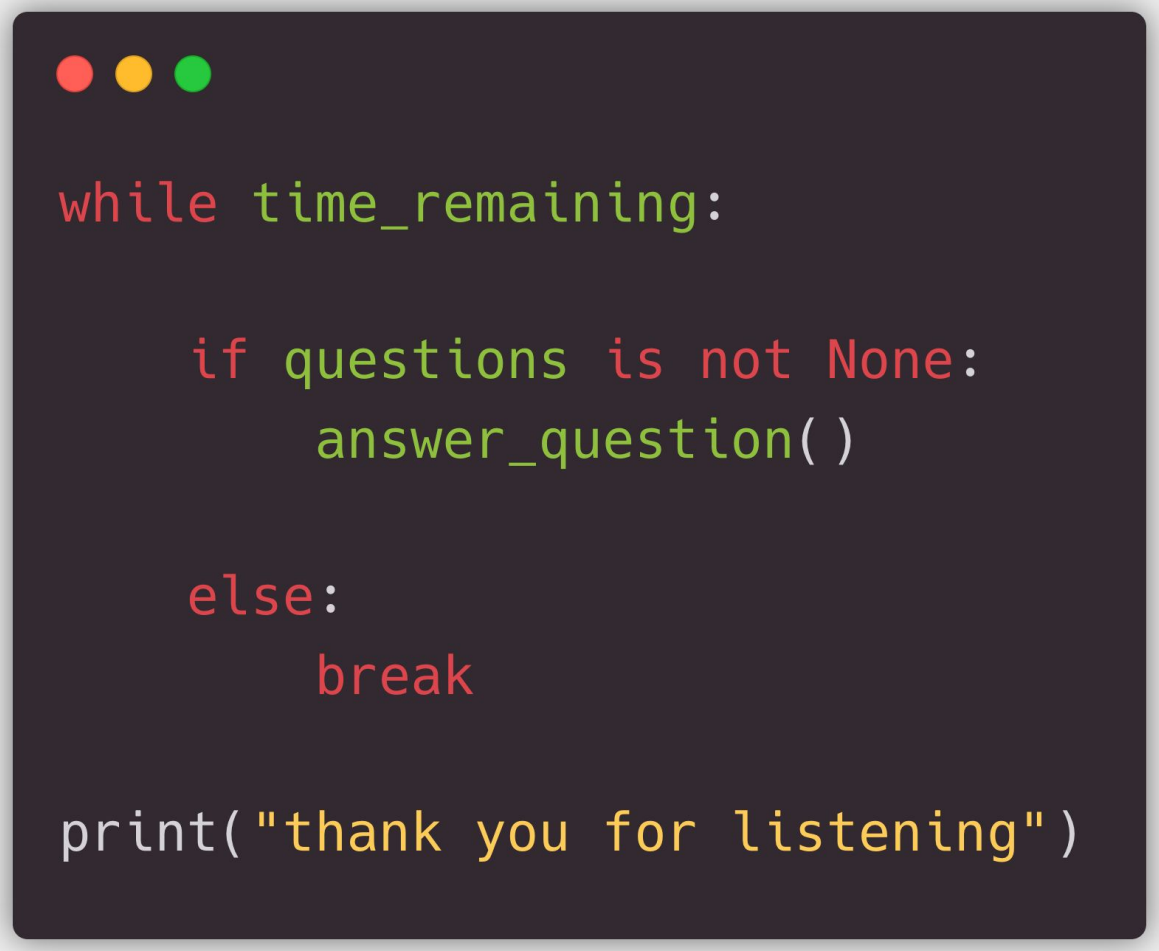
- We found a fabulously expensive way to generate synthetic data.
- On Google Colab’s Tesla P100 this takes 18-24 hours to run. Generating 30MB/hour, but only a small amount is useable.
- This is a way to generate synthetic samples without using statistical methods
- Further study would included convolutional layers

References

[1] Hung Ba.
Improving detection of credit card fraudulent transactions using generative adversarial networks.
CoRR, abs/1907.03355, 2019.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio.
Generative adversarial nets.
In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27.
Curran Associates, Inc., 2014.

[3] New York City Taxi and Limousine Commission.


while time_remaining:
if questions is not None:
answer_question()
else:
break
print("thank you for listening")