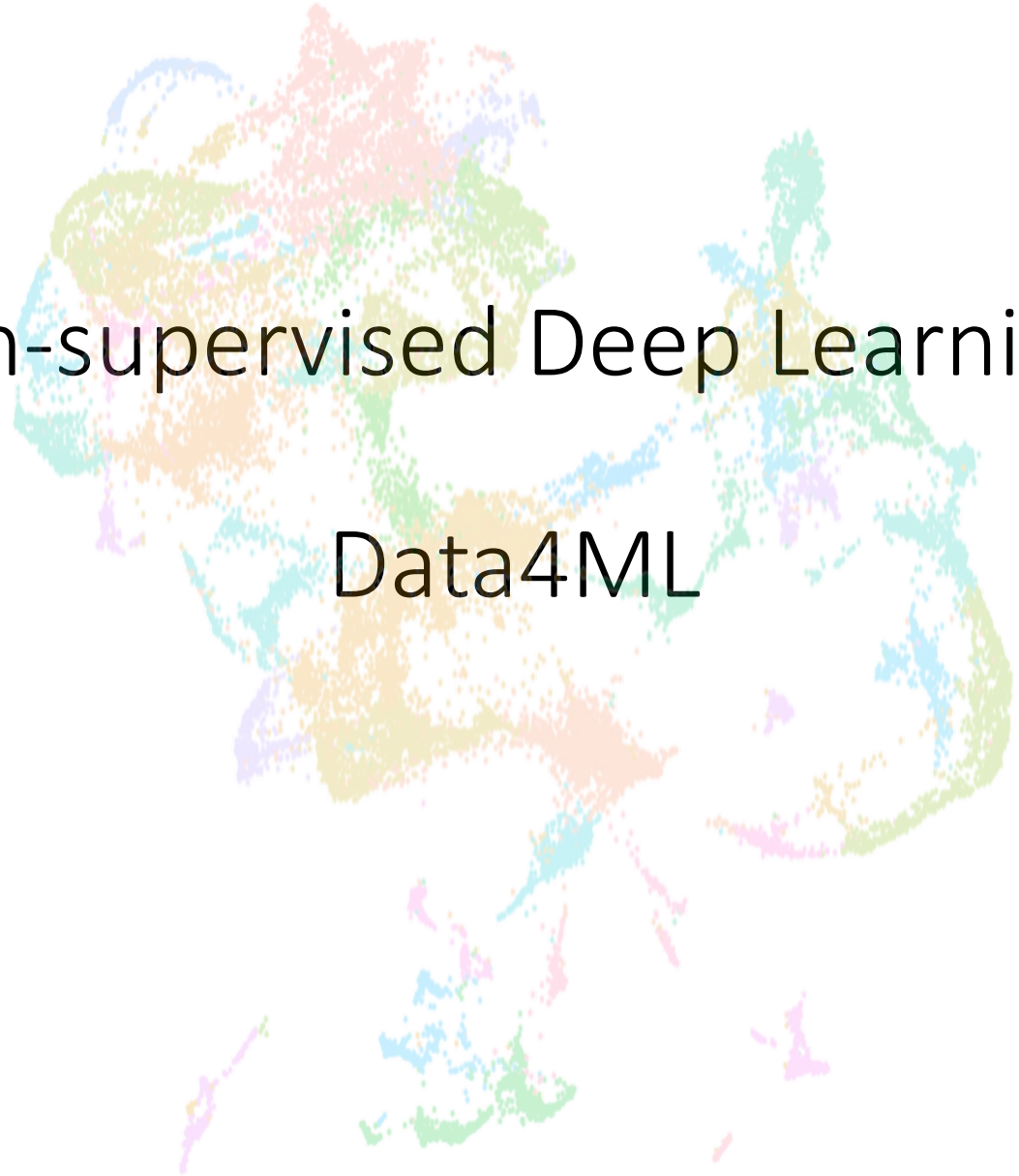
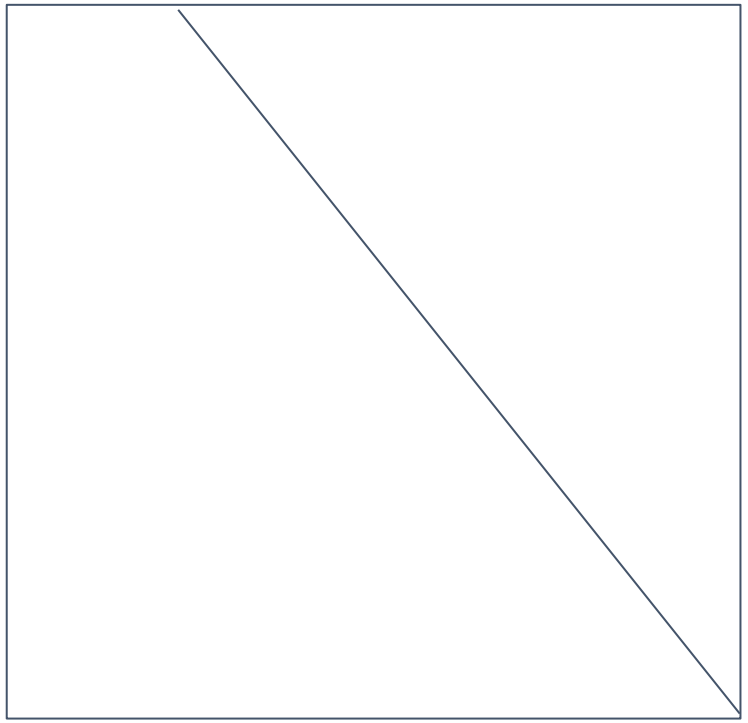


Un-supervised Deep Learning

Data4ML

Reminder Dimensionality Reduction

Picture of a Line
(500x500 pixels=250,000 Numbers)



Equation:
y=m*x+b (2 numbers)

Both have the same information in different formats

It’s often a lot easier to analyze 2 number than 250,000

A Reminder

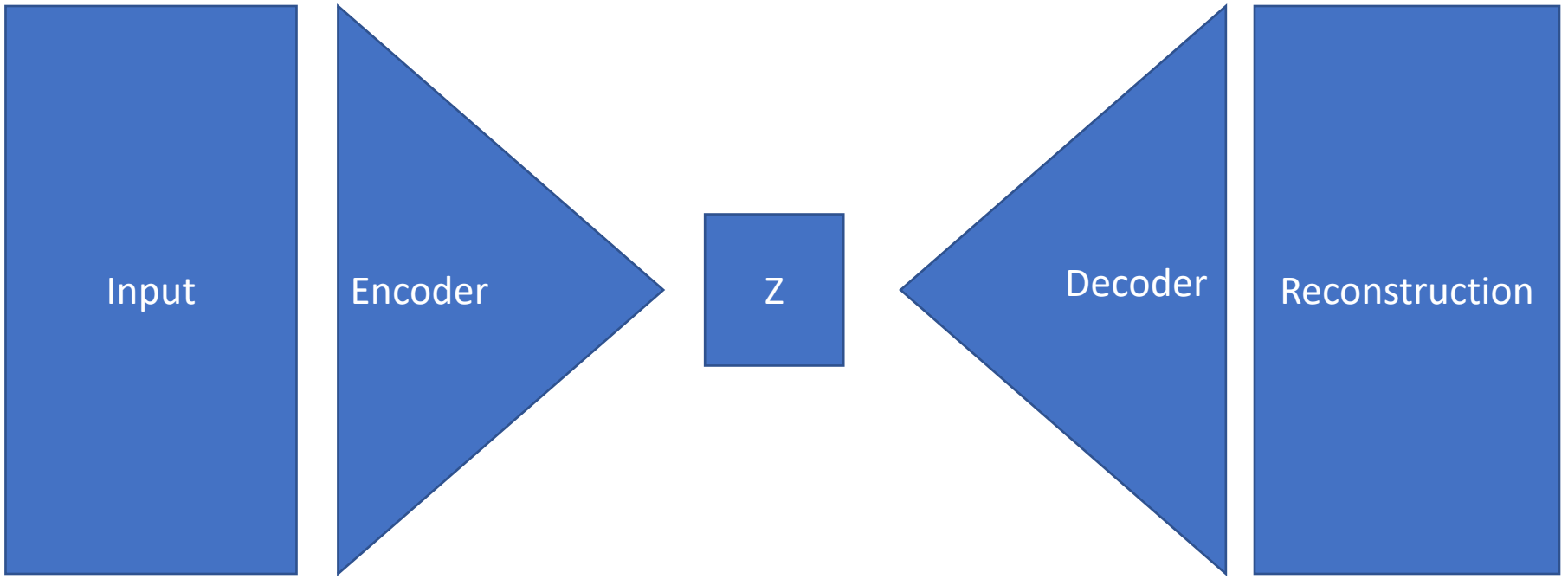
- PCA
  - Linear algorithm that finds the directions with the largest variance
  - You can trust inter and intra cluster distances
  - You may have more important PCA components than you want to plot
  - Lossless until you discard higher PC dimensions
- TSNE/UMAP
  - Relies on nearest-neighbors information
  - You can **not** trust inter and intra cluster distances
  - You can pick the dimension size you want to reduce to use
- What if you want to get the best of both worlds
- What if you want to use un-structured data

Deep Unsupervised Learning Vocab

- Deep learning uses stack of layers to approximate a function. In unsupervised learning that function is a mapping from your input data to a latent space
- **Latent space** – often denoted **z** (like **x** denotes inputs), think of this like your principal components or outputs of UMAP it's then transformed space that normally has a lower dimension than the original.
- **Encoder –** A function that maps x -> z it 'encodes' your input data
- **Decoder –** A function that maps z -> x it 'decodes' your input data
- Algorithms may use some variations one or all of these

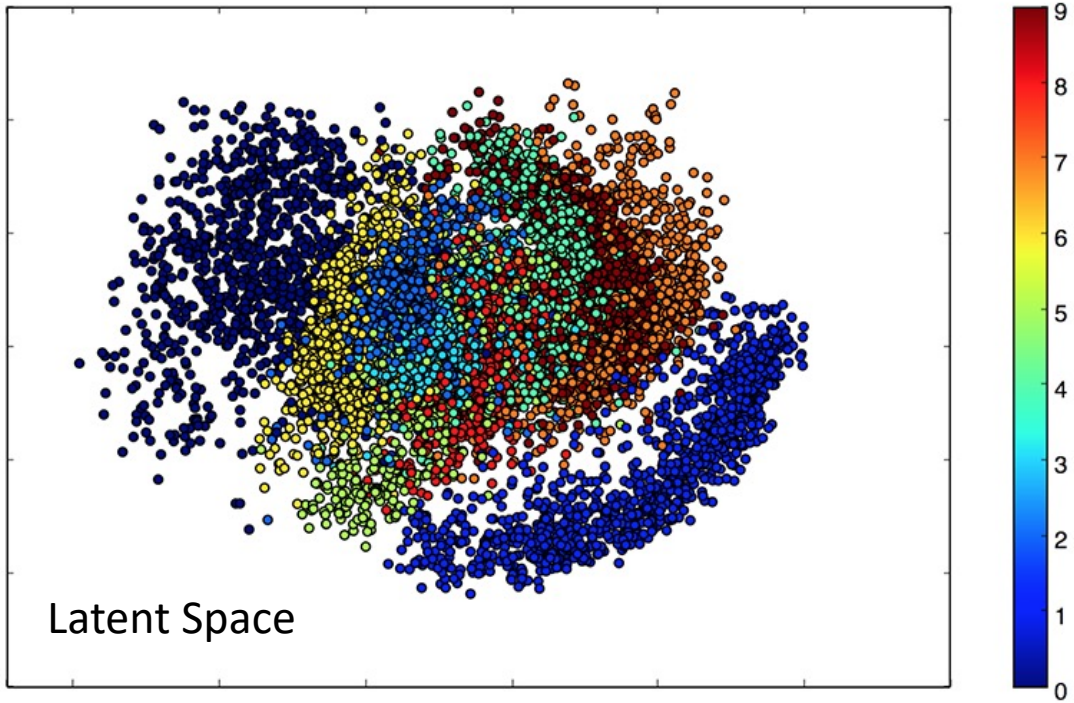
Autoencoder

- A common deep learning model for dimensionality reduction

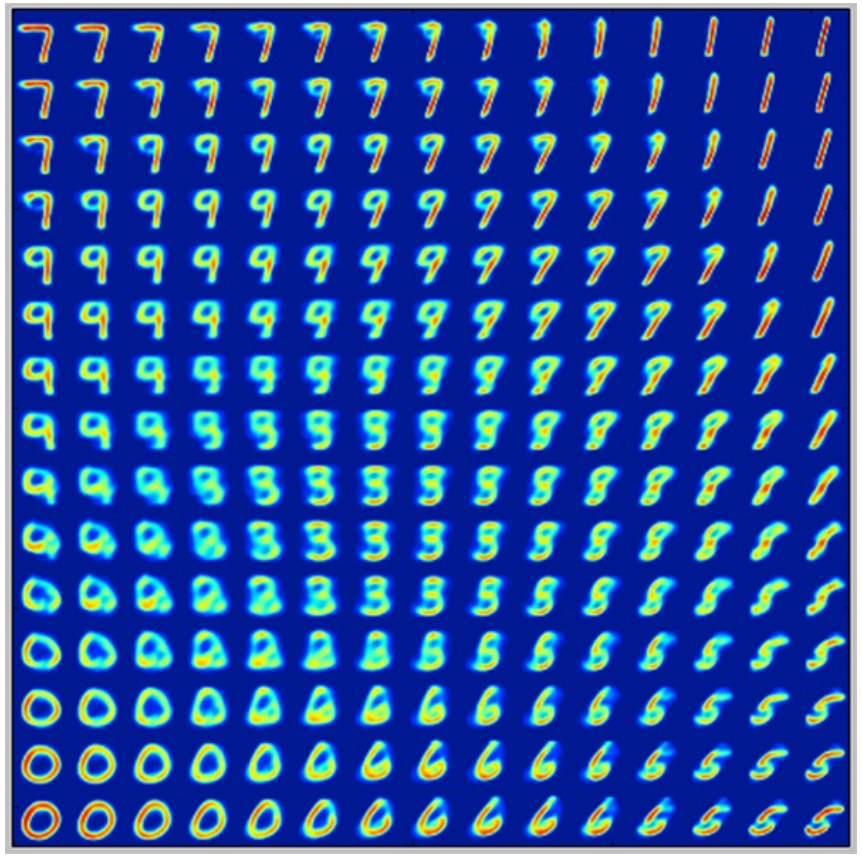


- Trained to find the best way to ‘compress’ all the information into z possible to recreate the original data
- The encoder preforms some dimensionality reduction, and the decoder tries to reverse this process

Encoder



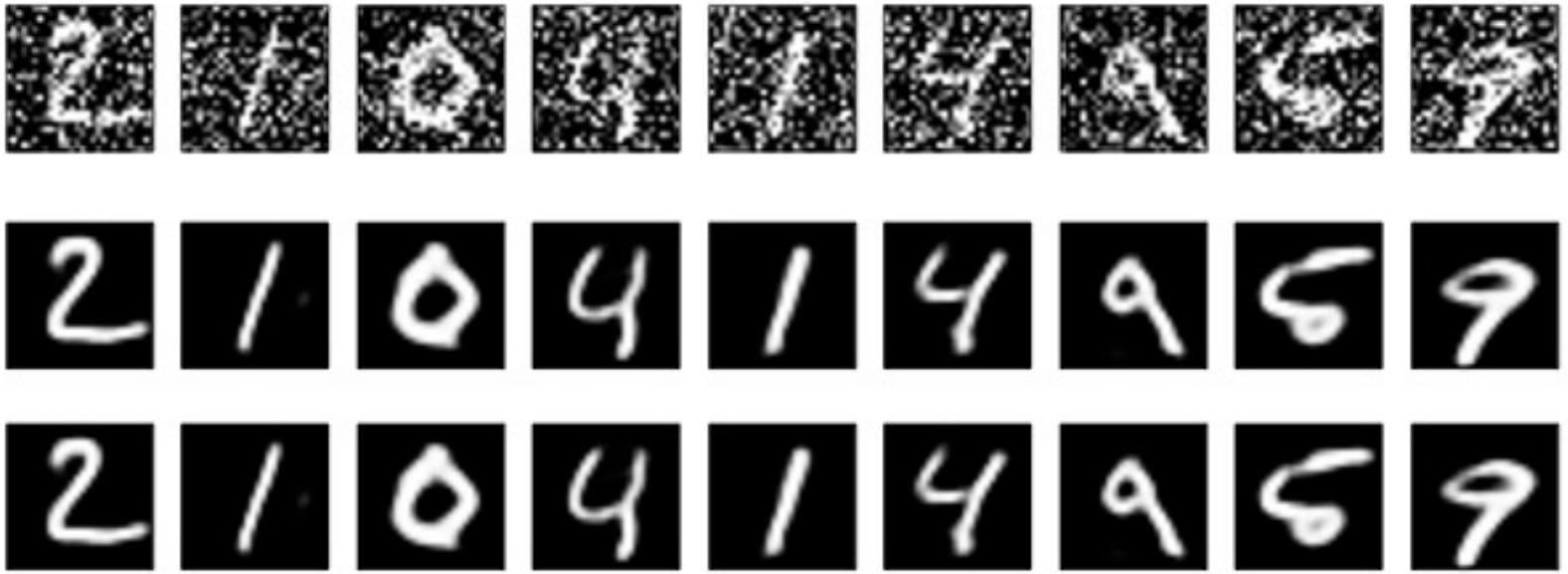
Decoder



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

Denoising

- Another handy feature of Autoencoders is their ability to denoise

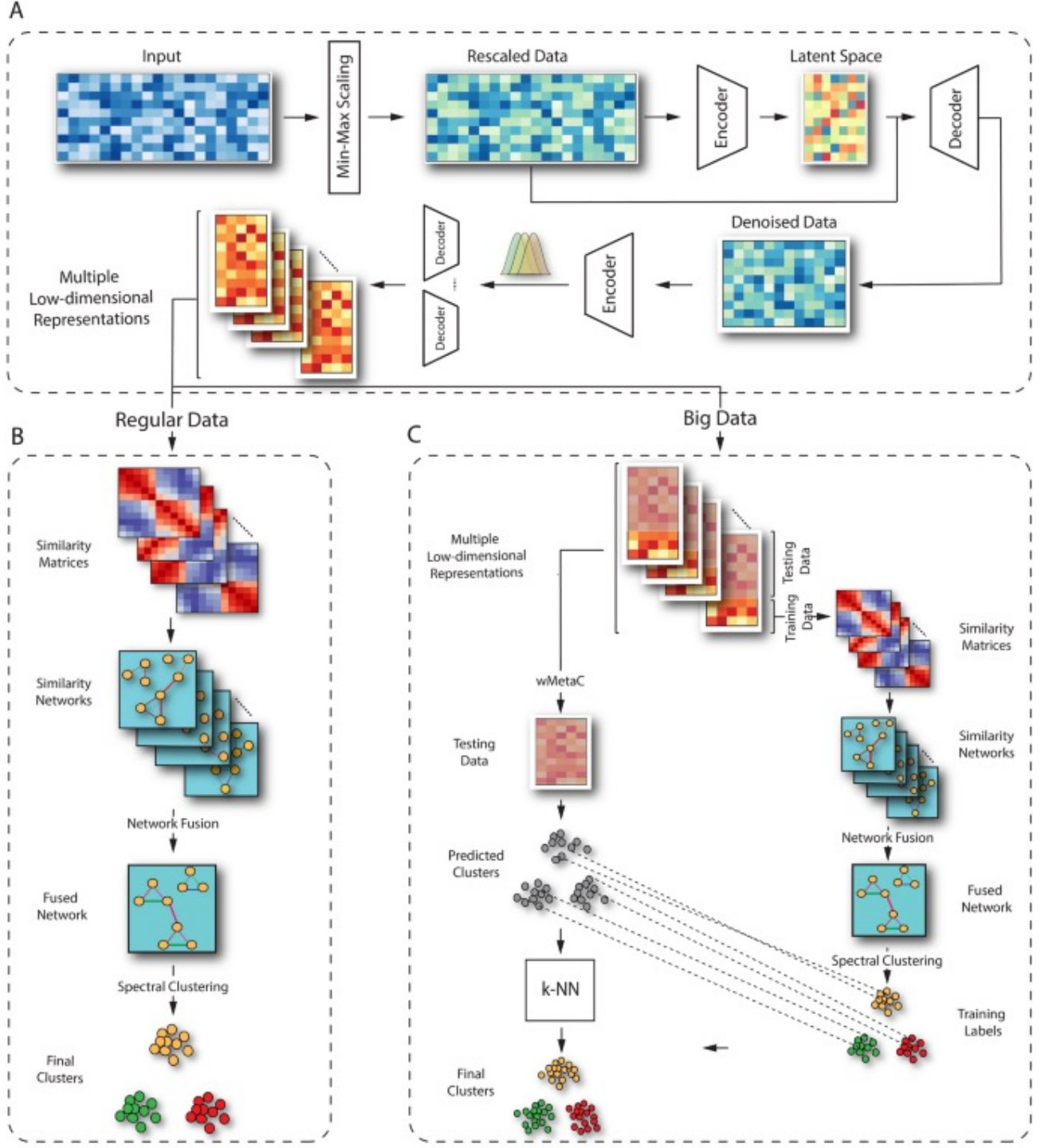


Becoming Available as Tools in Genomics

- https://www.nature.com/articles/s41598-022-14218-6

scCAN: single-cell clustering using autoencoder and network fusion

Lots of exploration in this space, so there are lots of new algorithms to try, this one has a convenient R package

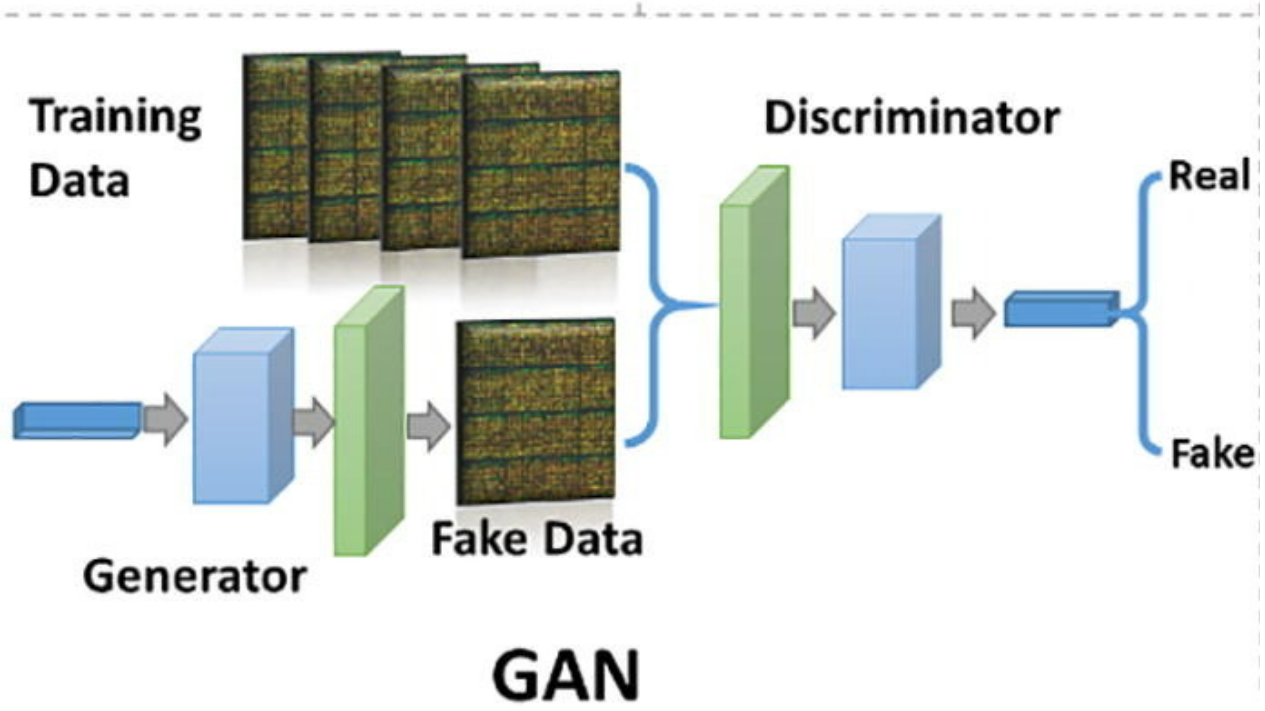


Generative Adversarial Learning (GANS)

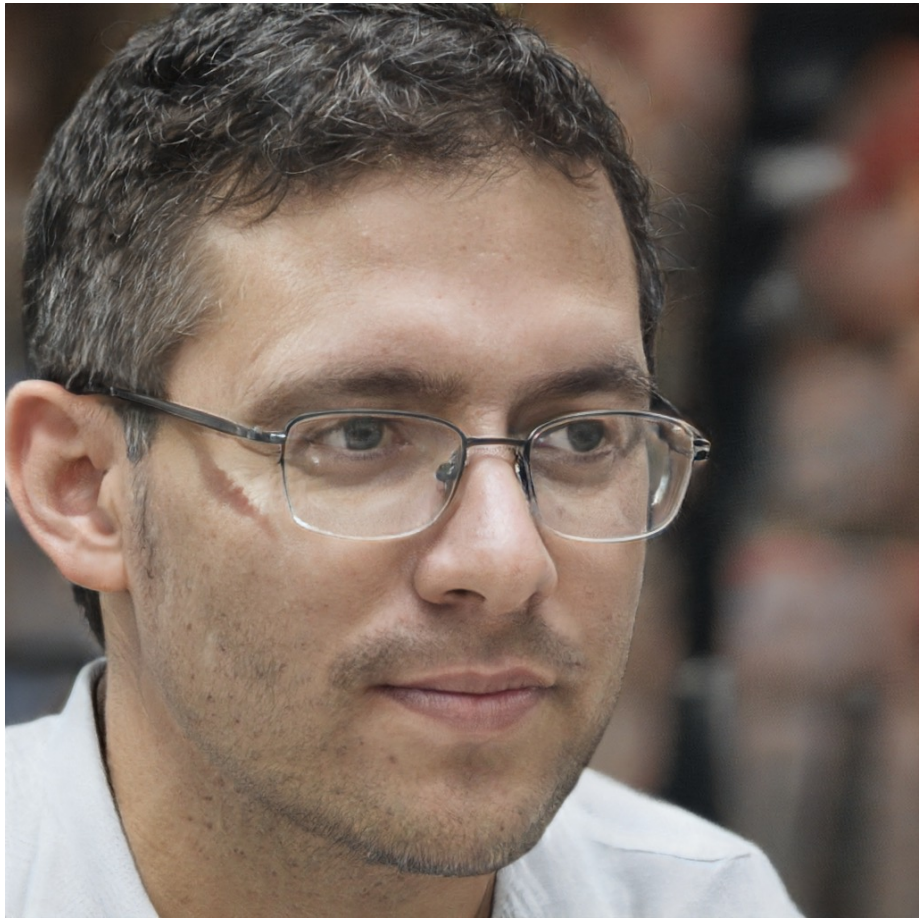
- A different model with only a generator, its goal is to generate new examples that have the same properties as your data.
- Usually Better reconstructions that auto-encoders
- Usually worse latent spaces
- Application lean toward generating new examples
  - Therapeutic Proteins
  - Genomes
  - Etc.

Example https://thisstartupdoesnotexist.com/

https://www.sciencedirect.com/science/article/pii/S2001037020303068



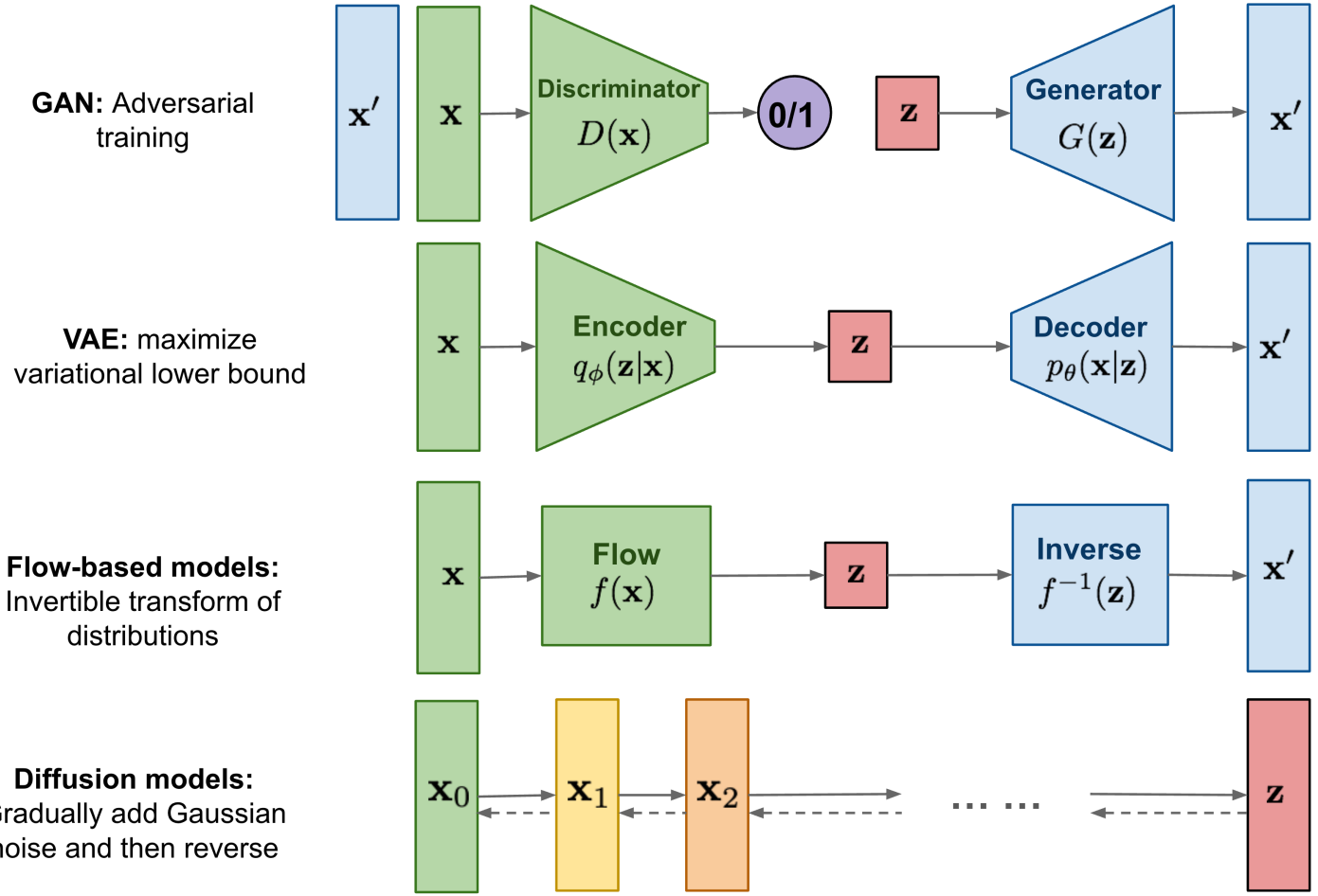
These Networks have produced incredible results for images



https://thispersondoesnotexist.com/

https://www.craiyon.com/

Other Models



https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

Conclusions

- Lots of interesting capabilities coming out of deep learning
- Exactly how this will impact genomics is being figured out as we speak
- Keep an eye out and try new methods as they come out