Pairing human control with generative models for creative content synthesis

Chris Donahue
UC San Diego

How can machine learning be creative?

Supervised learning (not particularly creative)

Data

Example	Label
	Sheep
	Cat
…	…
	Frog

→

Machine learning algorithm

→

Predictive model

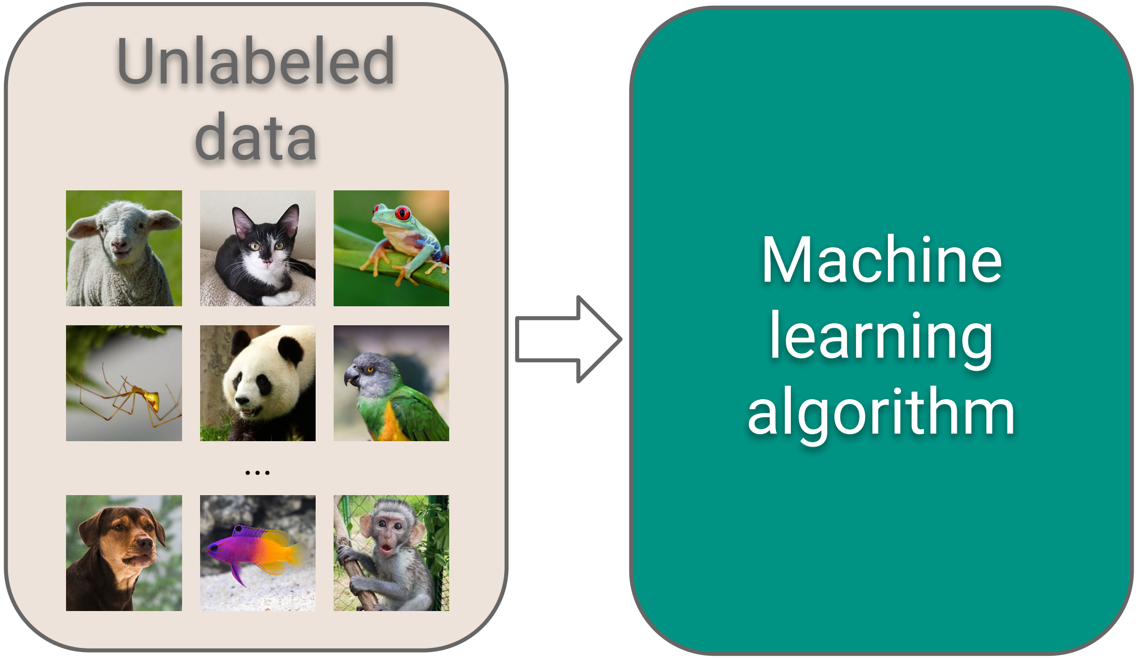
New example

Predicted label

Frog

How can machine learning be creative?

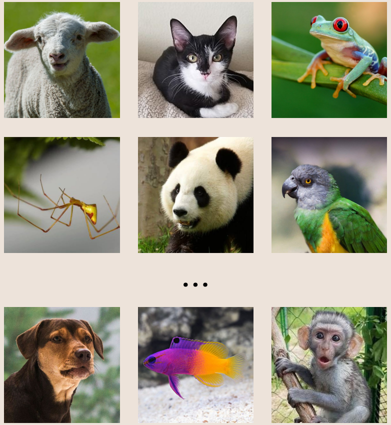
Unsupervised learning (more creative)



How can machine learning be creative?

Unsupervised learning (more creative)

Unlabeled data

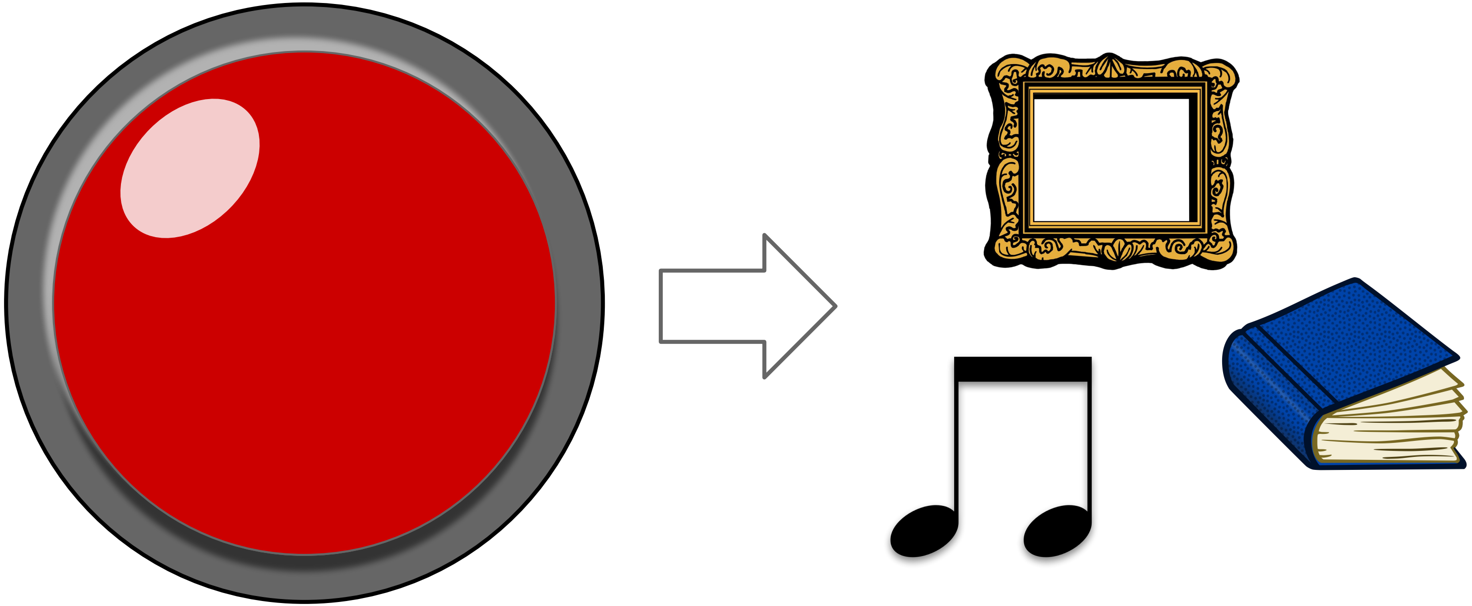


Machine learning algorithm

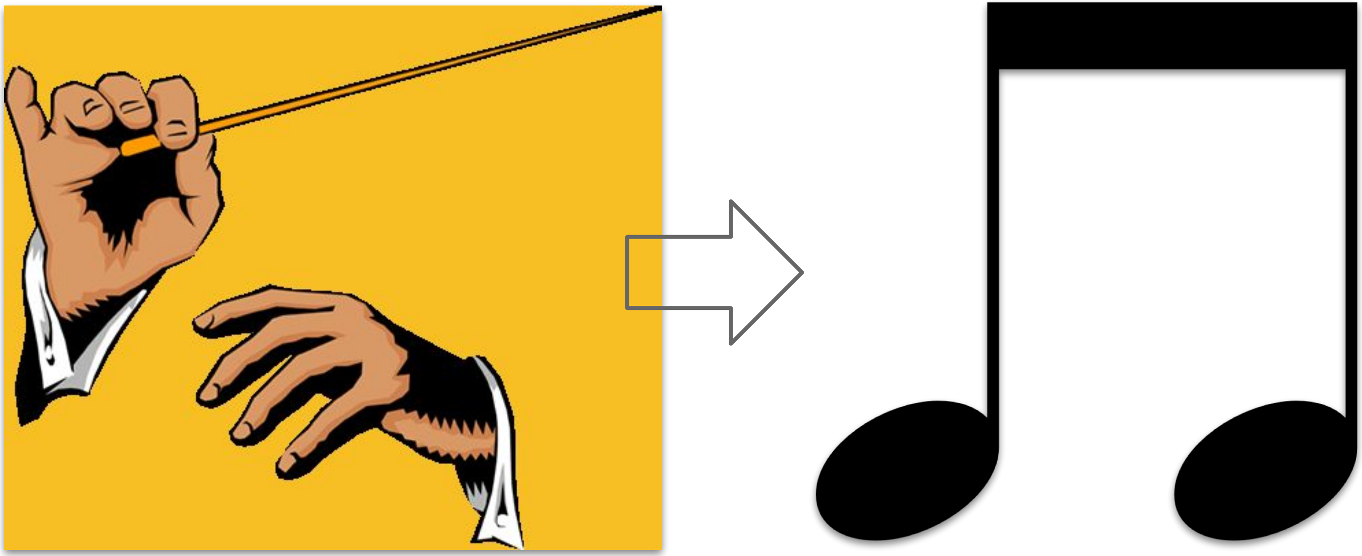
Generative model

BigGAN (Brock et al. 2018); video by Mario Klingemann

Traditional generative modeling framework

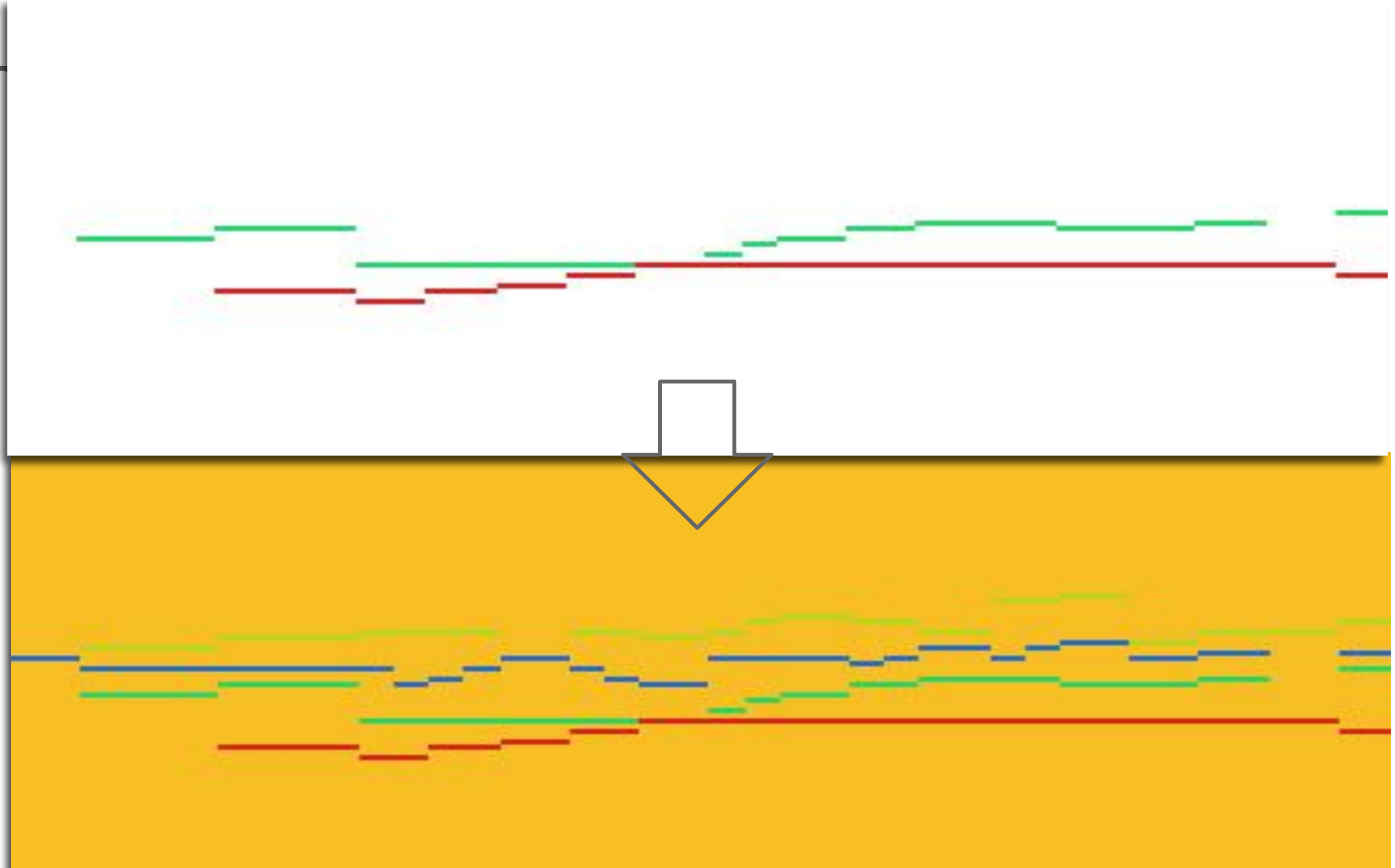


Interactive generative modeling framework



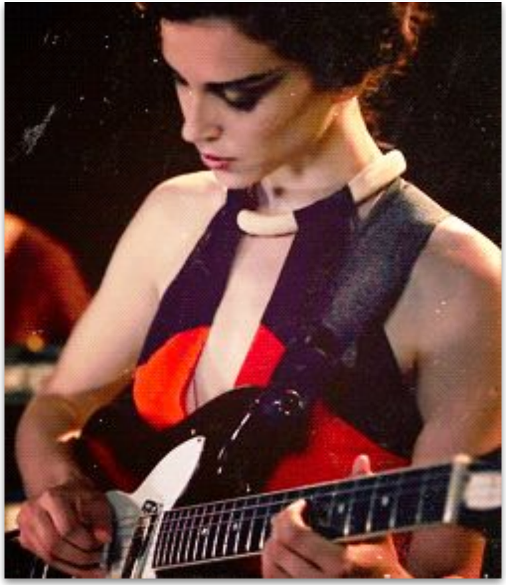
Assisting musicians with interactive ML

Harm



Assisting non-musicians with interactive ML

Help non-musicians *create* music



What do we need for machine learning?

During training

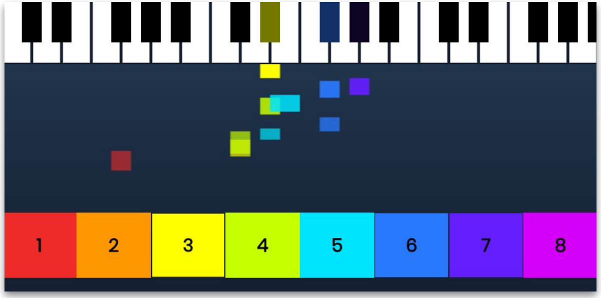
Data

Specialized infrastructure

After training

Commodity hardware

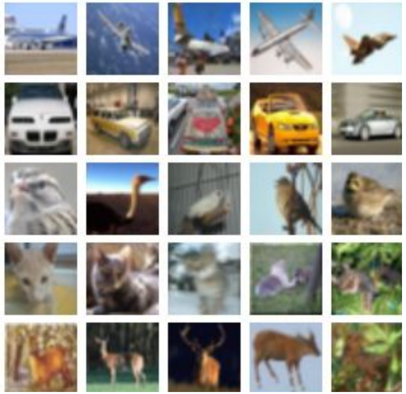
Interface



Popular classes of generative models

- Generative adversarial networks (Goodfellow et al. 2014)
- Variational autoencoders (Kingma et al. 2013)
- Language models (e.g. Markov chains, RNNs)

Each is a different way of modeling data distribution

 $\sim p_{\text{data}}$

Generative adversarial networks

Generative adversarial networks

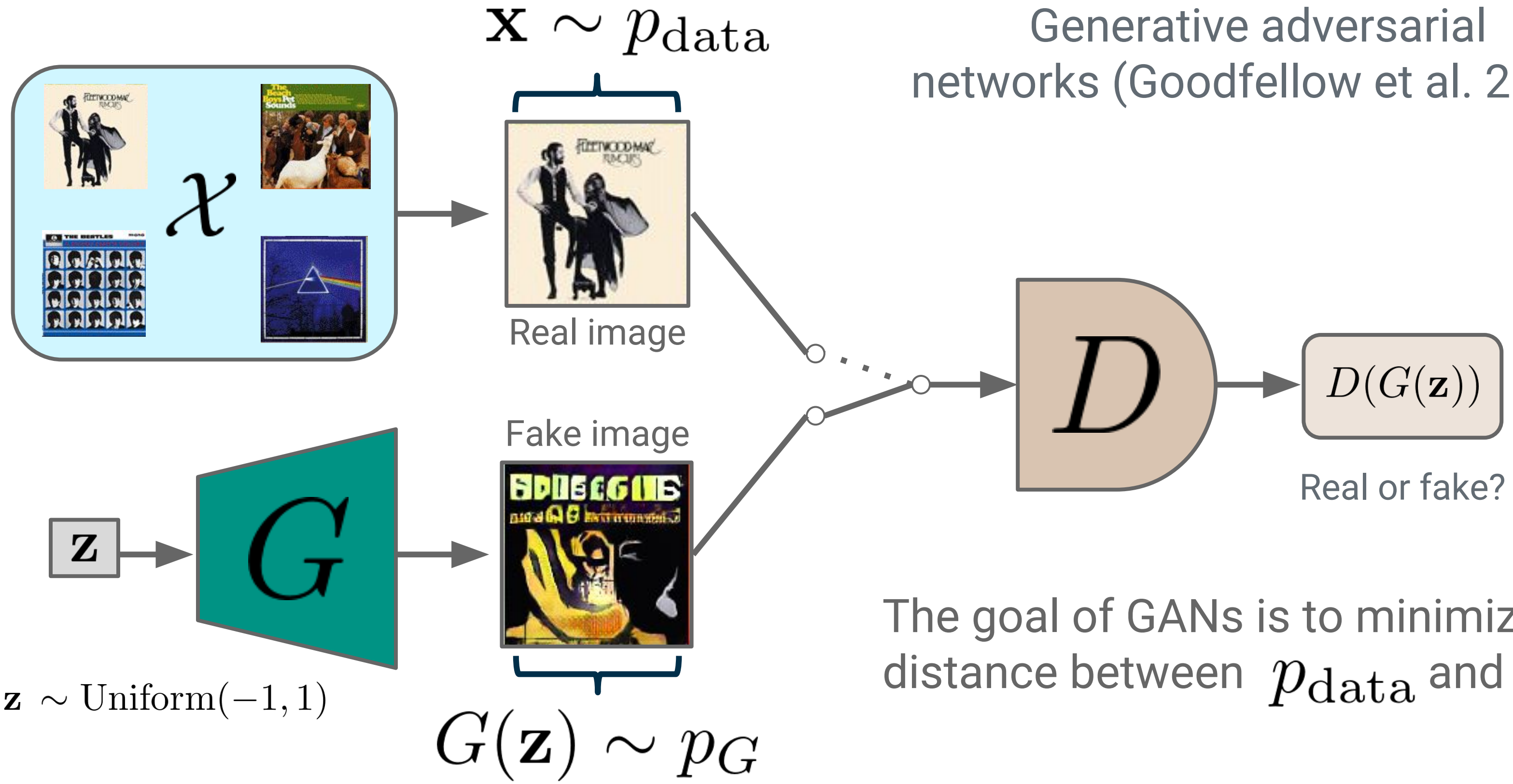
Goodfellow et al. 2014

Radford et al. 2016

Karras et al. 2018

How do GANs work?

Generative adversarial networks (Goodfellow et al. 2014)

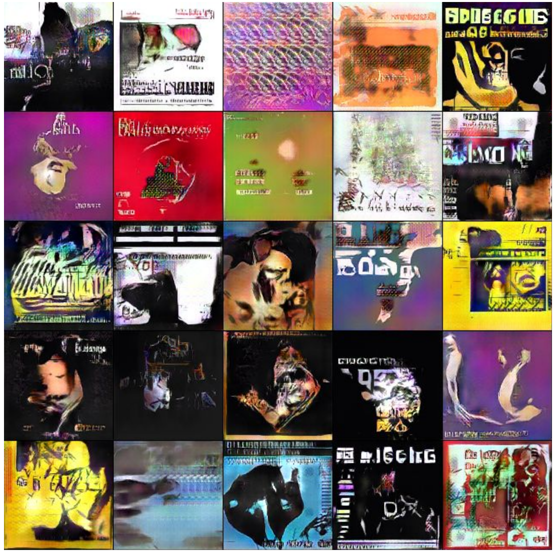


The goal of GANs is to minimize distance between p_{data} and p_G

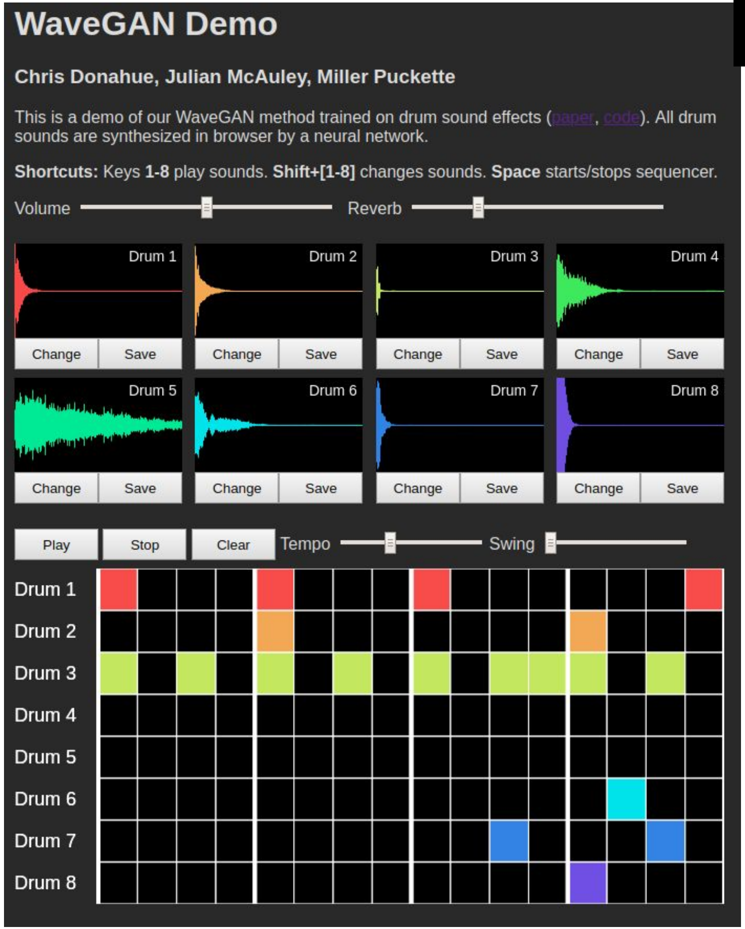
Image generation with GANs

Real images

Generated images



Audio generation with GANs



chrisdonahue.com/wavegan

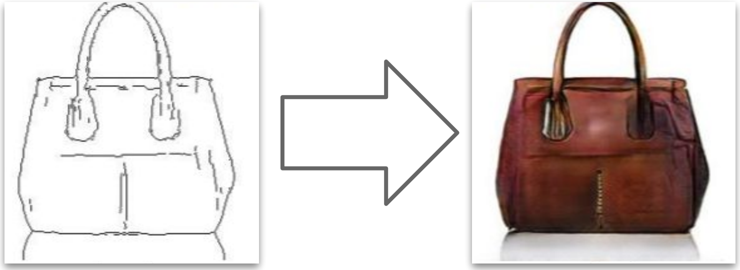
Can we intuitively control GAN generation?

Images synthesized by BigGAN (Brock et al. 2018)

Interpolation
(no modification)

Class conditioning
(limited modification)

“Smart filters”
(heavy modification)



“Smart filters”: pix2pix (Isola et al. 2016)

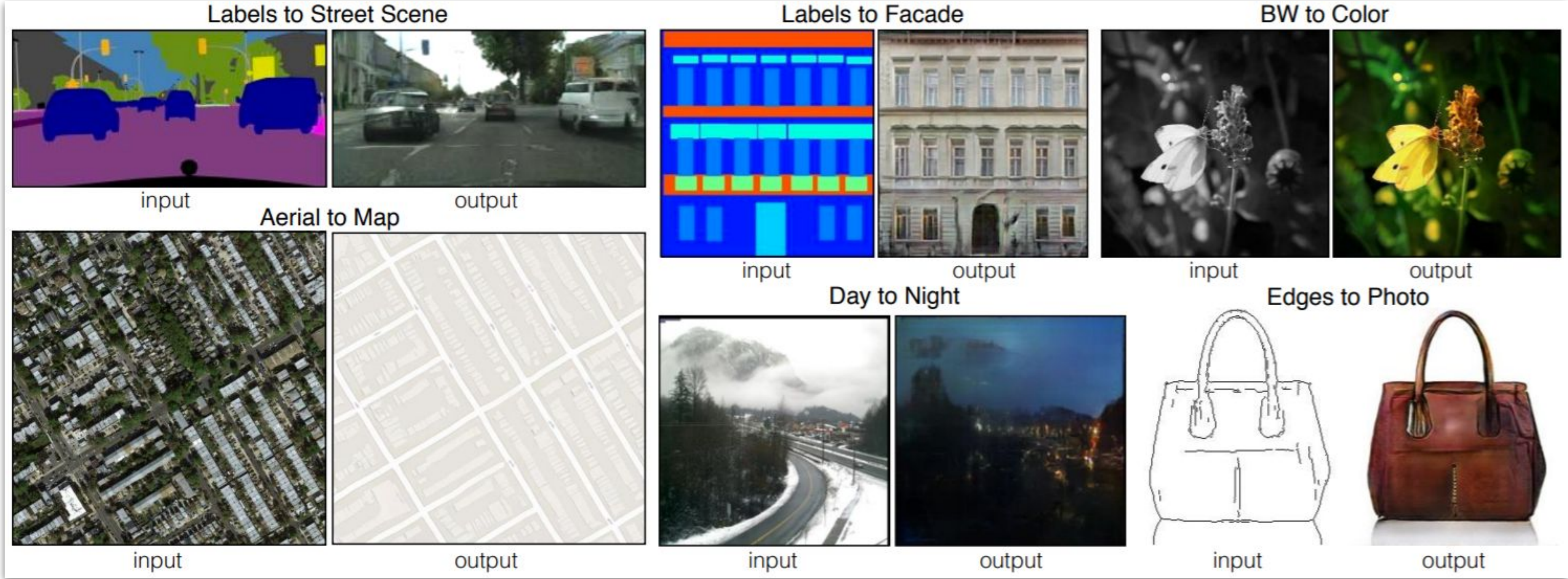
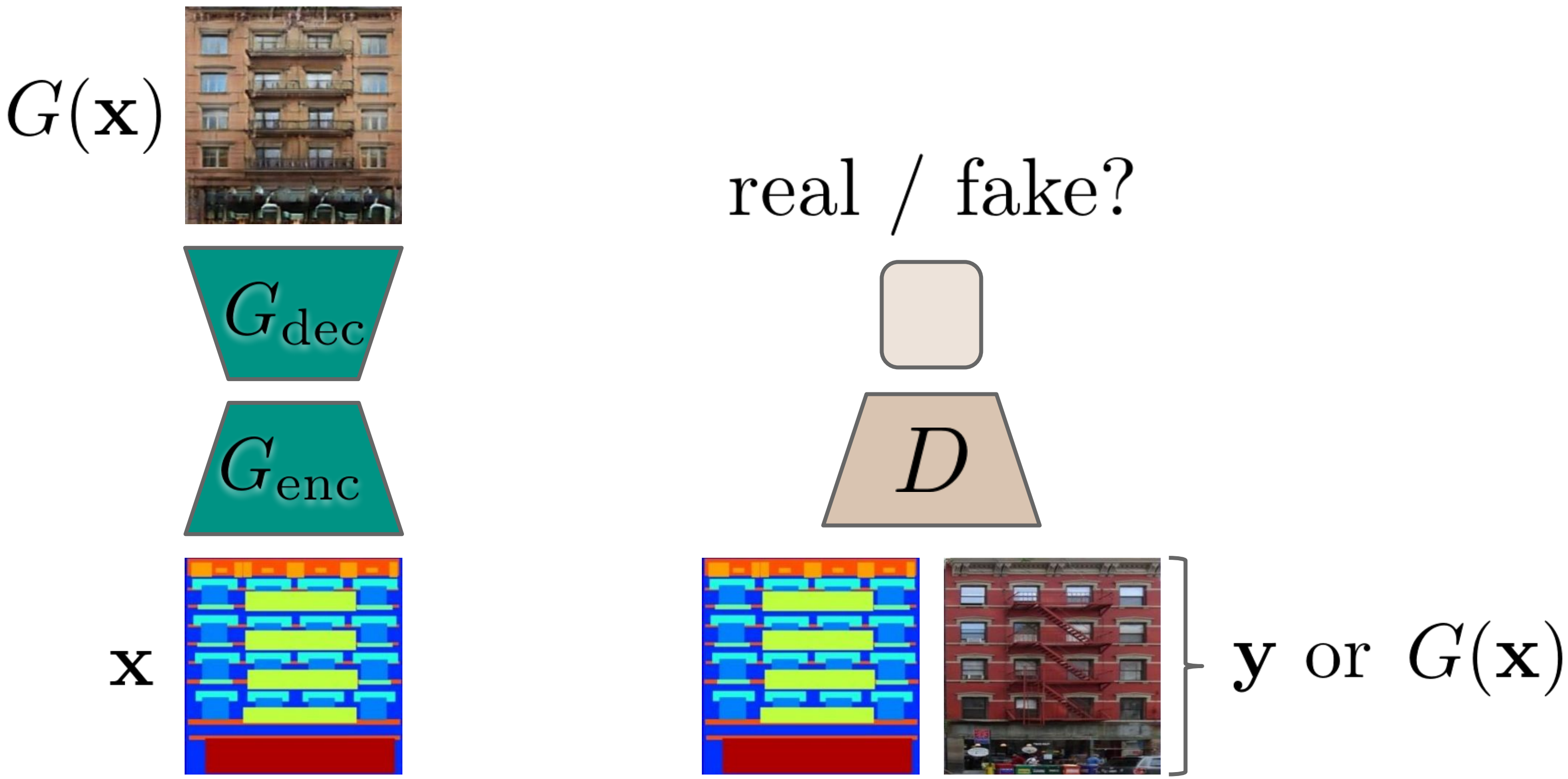
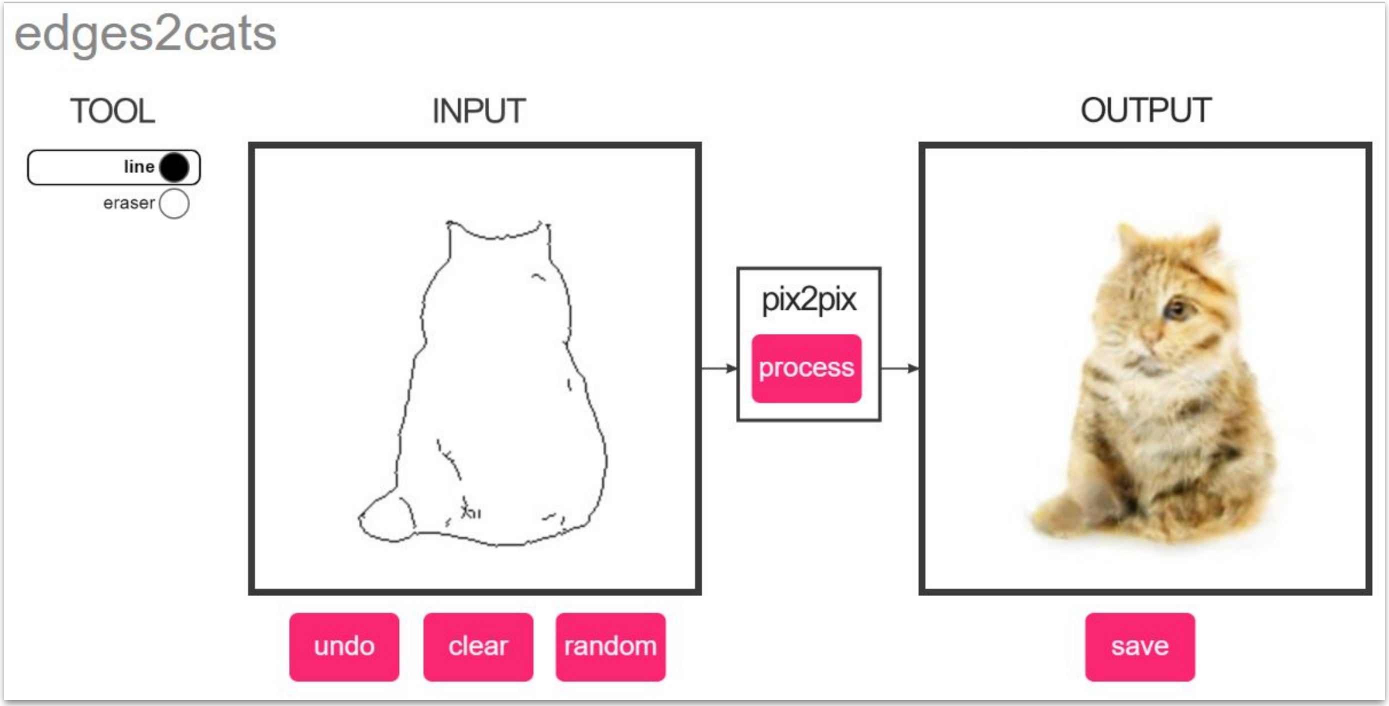


Figure from *Image-to-image translation with conditional adversarial networks (Isola et al. 2016)*

pix2pix (Isola et al. 2016)

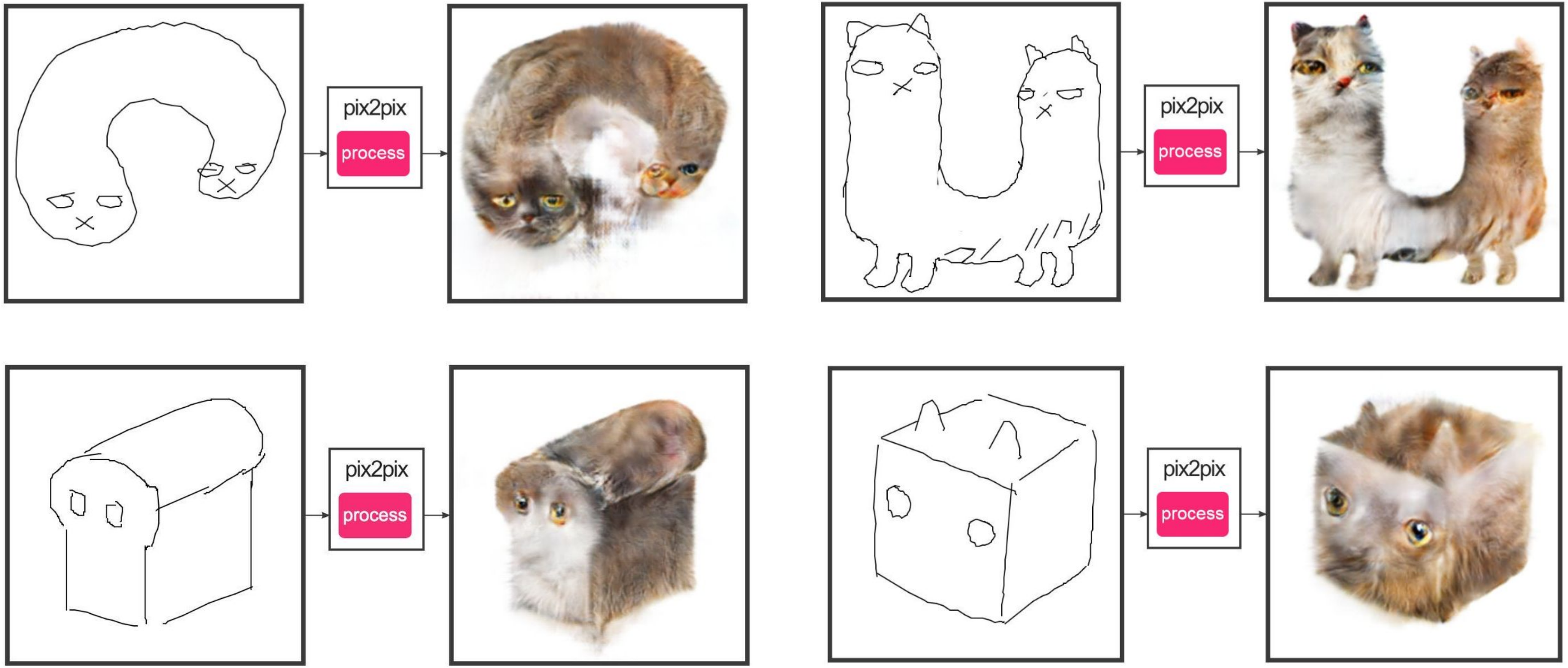


edges2cats



https://affinelayer.com/pixsrv/

Abusing edges2cats



pose2dance

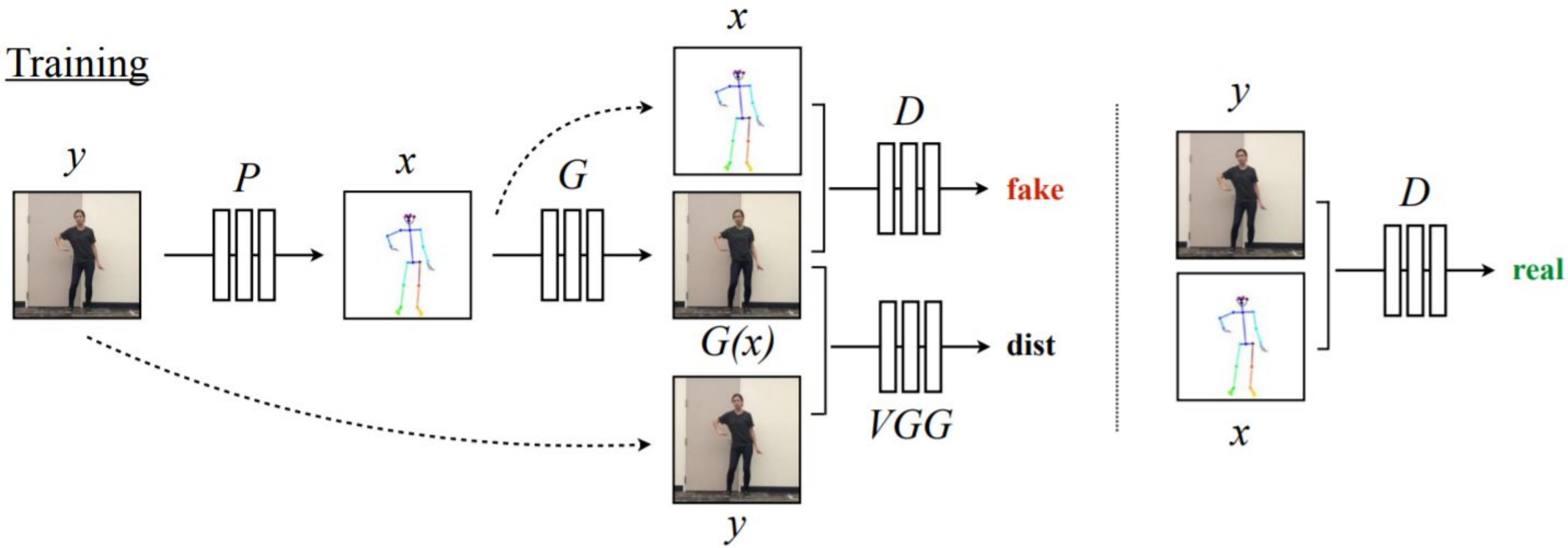


Figure from *Everybody Dance Now* (Chan et al. 2018)

pose2dance


Source Video
Detected Pose
Source to Target 1 Result
Source to Target 2 Result

CycleGAN

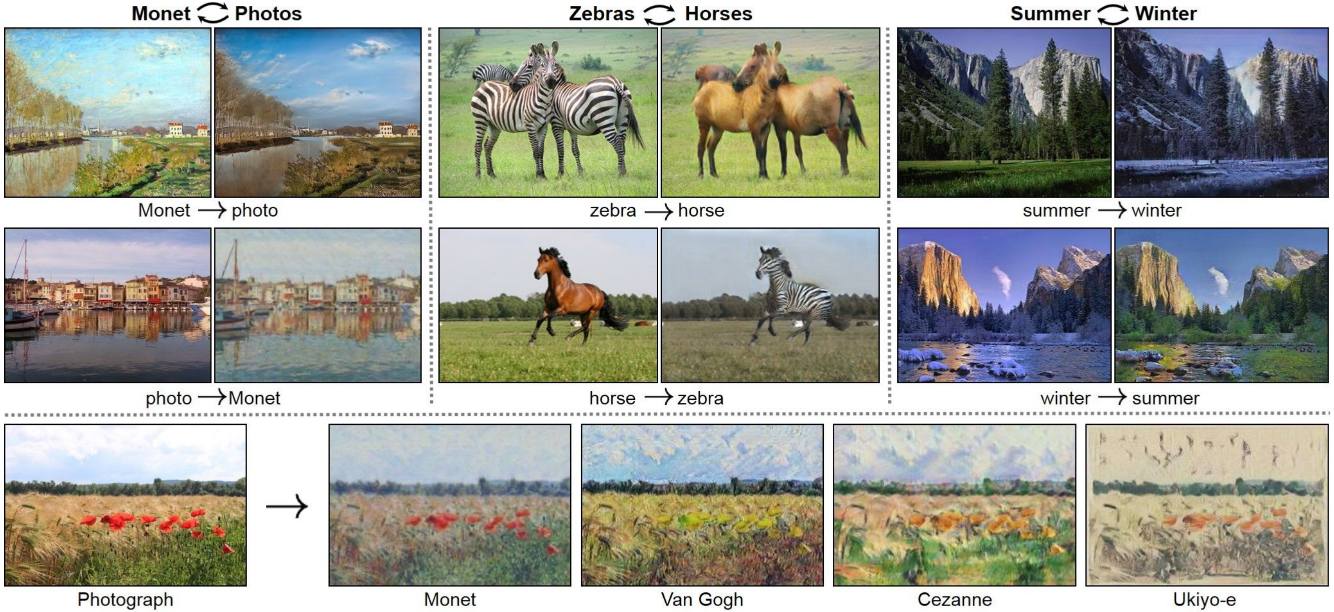


Figure from *Unpaired image-to-image translation using cycle-consistent adversarial networks (Zhu et al. 2017)*

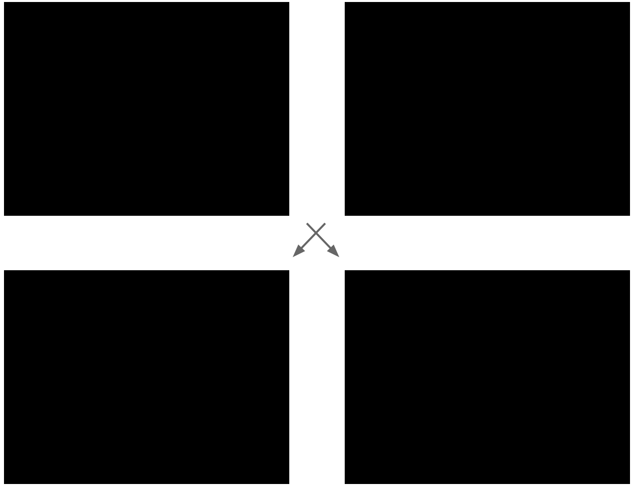
CycleGAN for musical timbre transfer

My attempt
(unpublished)

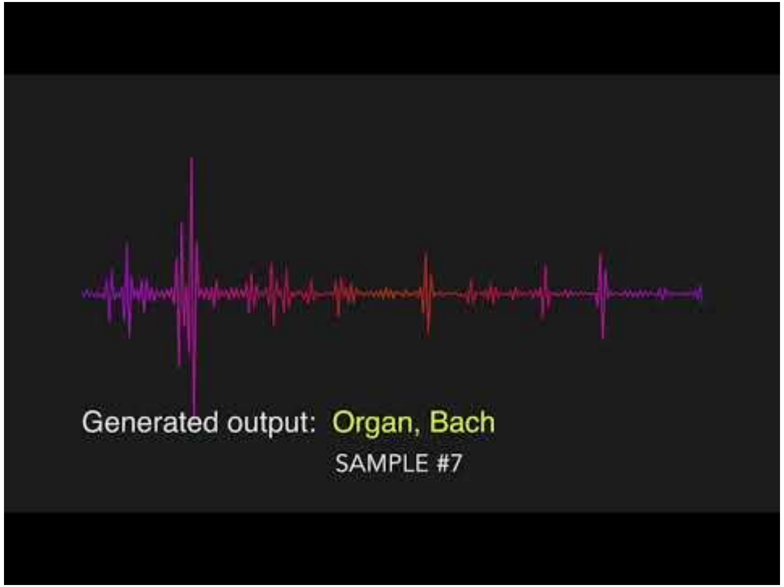
Acoustic Synthetic

Real

Fake



Universal music transformer
(Mor et al. 2018)



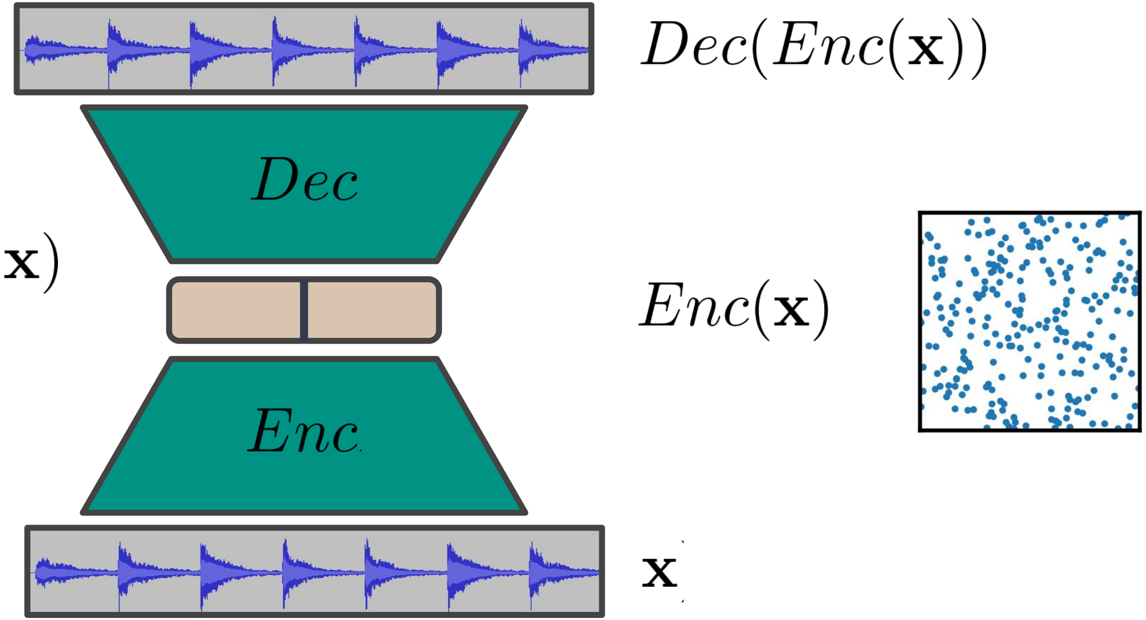
Variational autoencoders

What is an autoencoder?

An autoencoder is (usually) a pair of neural networks which learn to compress data

Minimize:

$$\mathrm{error}(Dec(Enc(\mathbf{x})), \mathbf{x})$$

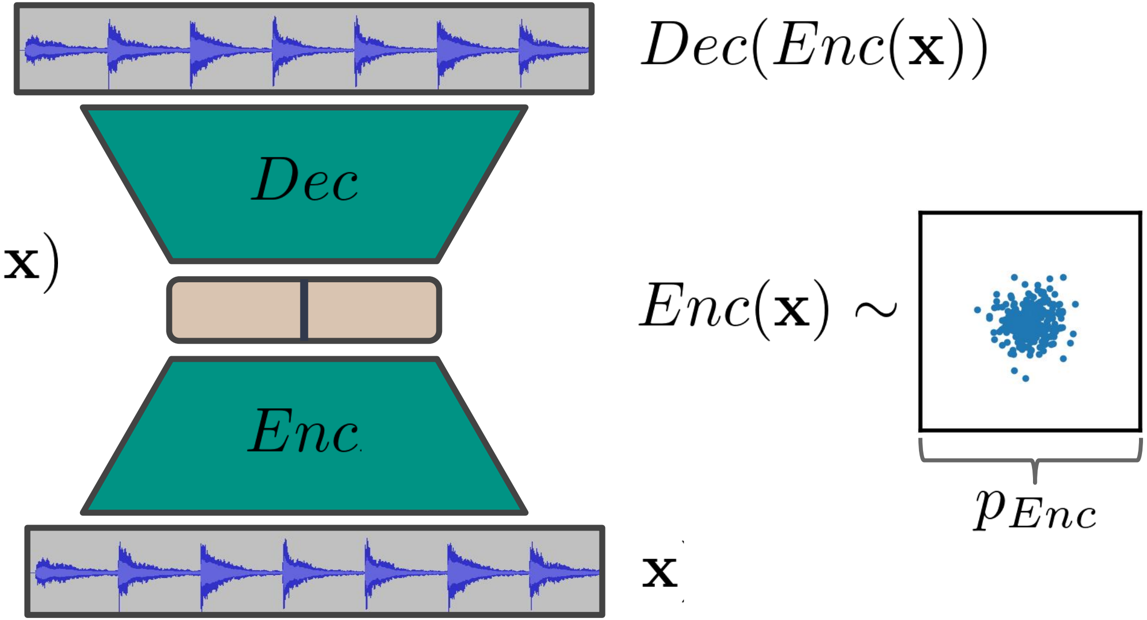


What is a *variational* autoencoder?

A variational autoencoder imposes particular structure on the encoder *latent space* so that we can sample

Minimize:

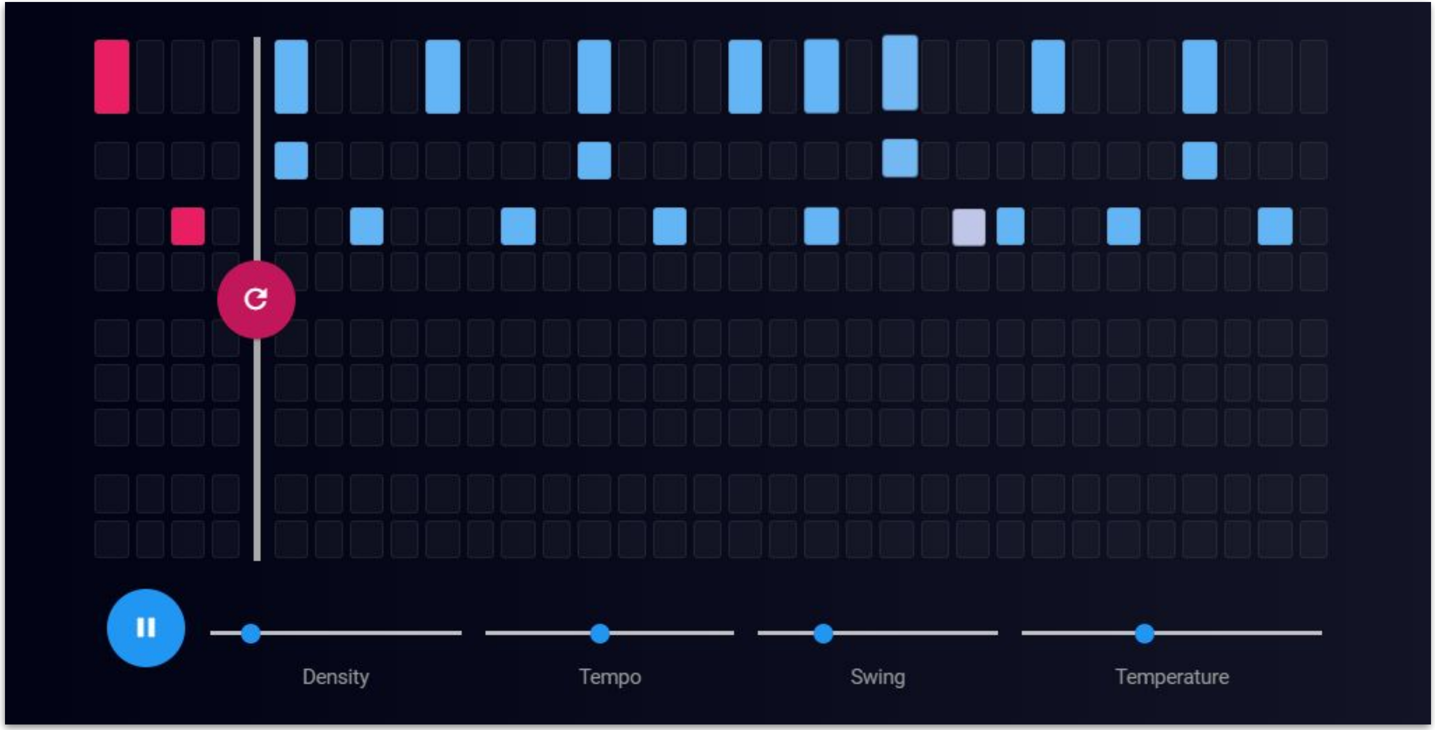
$$\mathrm{error}(Dec(Enc(\mathbf{x})), \mathbf{x})$$
$$+\ \mathrm{distance}(p_{Enc}, p_{\mathbf{z}})$$



SketchRNN (Ha and Eck 2017)

http://magic-sketchpad.glitch.me

Drum VAE (Roberts et al. 2018)



https://codepen.io/teropa/full/RMGxOQ/

Language models

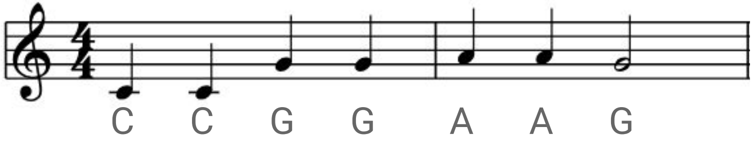
What is a *statistical language model*?

A statistical language model seeks to answer the question "how likely is this sequence" by comparing its statistics to those aggregated from a corpus of training data

How likely is this melody?

What does a language model look like?

Musical sequence

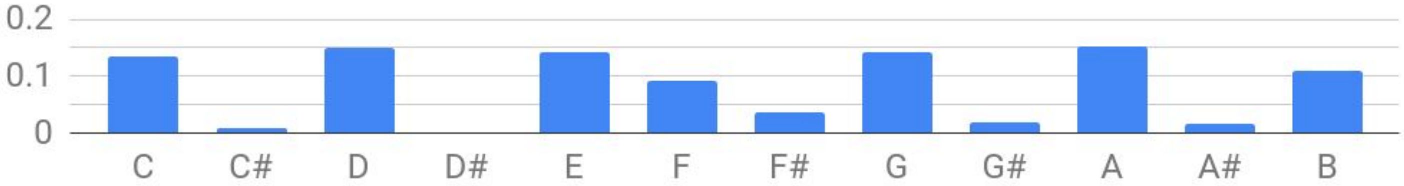


How likely?

Factorization (representation)

$$P(\text{music}) = P(n_1) \cdot P(n_2) \cdot \ldots \cdot P(n_T)$$

Language model

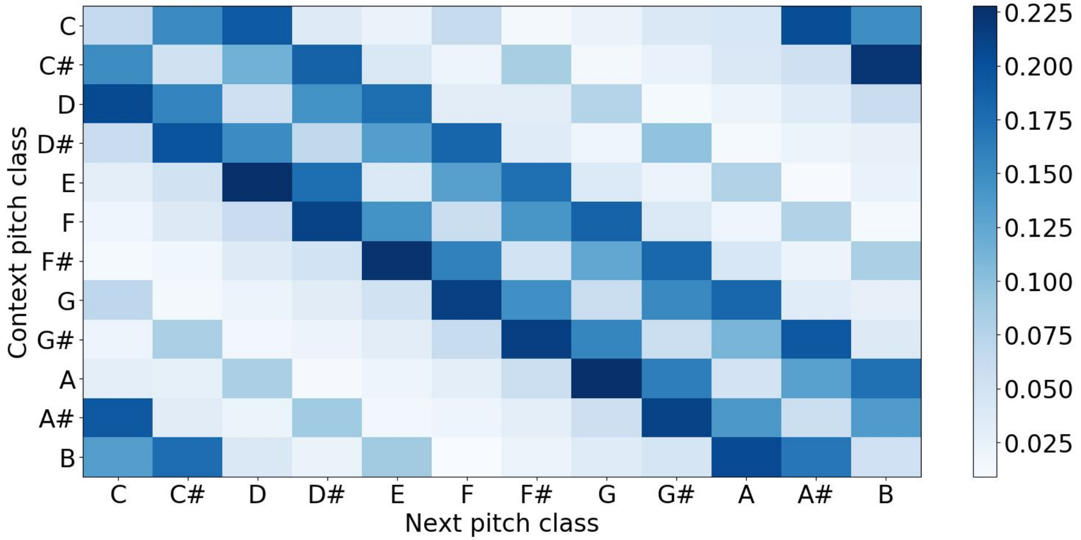


$$P(\text{seq}) = 0.13 \cdot 0.13 \cdot \ldots \cdot 0.14 = 1.18 \cdot 10^{-6}$$

Improving our simple language model

Unigram model $P(n_1) \cdot P(n_2) \cdot \ldots \cdot P(n_T)$

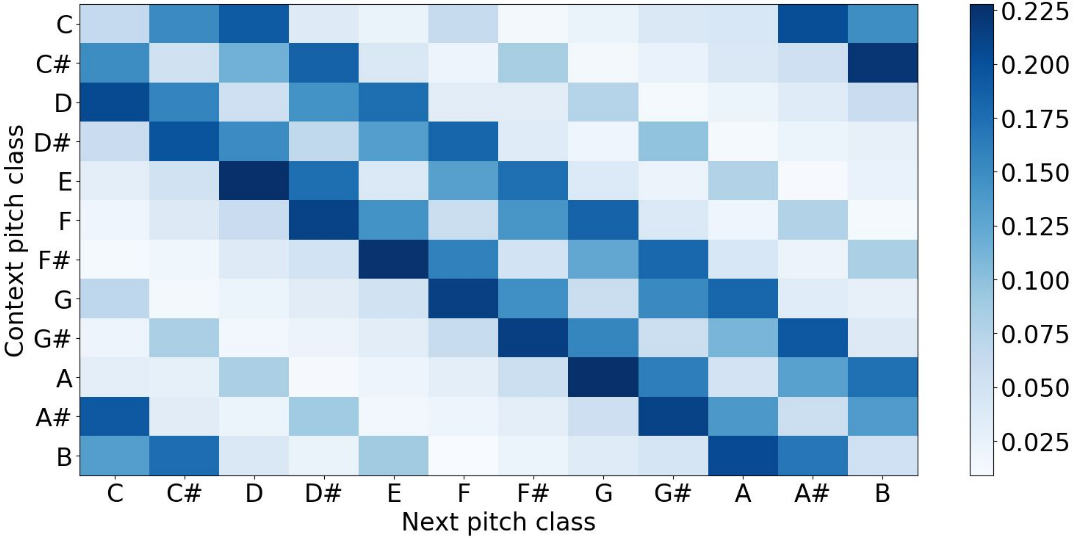
Bigram model (Markov chain) $P(n_1) \cdot P(n_2 \mid n_1) \cdot \ldots \cdot P(n_T \mid n_{T-1})$



Improving our simple language model

Unigram model $P(n_1) \cdot P(n_2) \cdot \ldots \cdot P(n_T)$

Bigram model (Markov chain) $P(n_1) \cdot P(n_2 \mid n_1) \cdot \ldots \cdot P(n_T \mid n_{T-1})$



Generalizing our language model

Simple model

$$P(n_1) \cdot P(n_2) \cdot \ldots \cdot P(n_T)$$

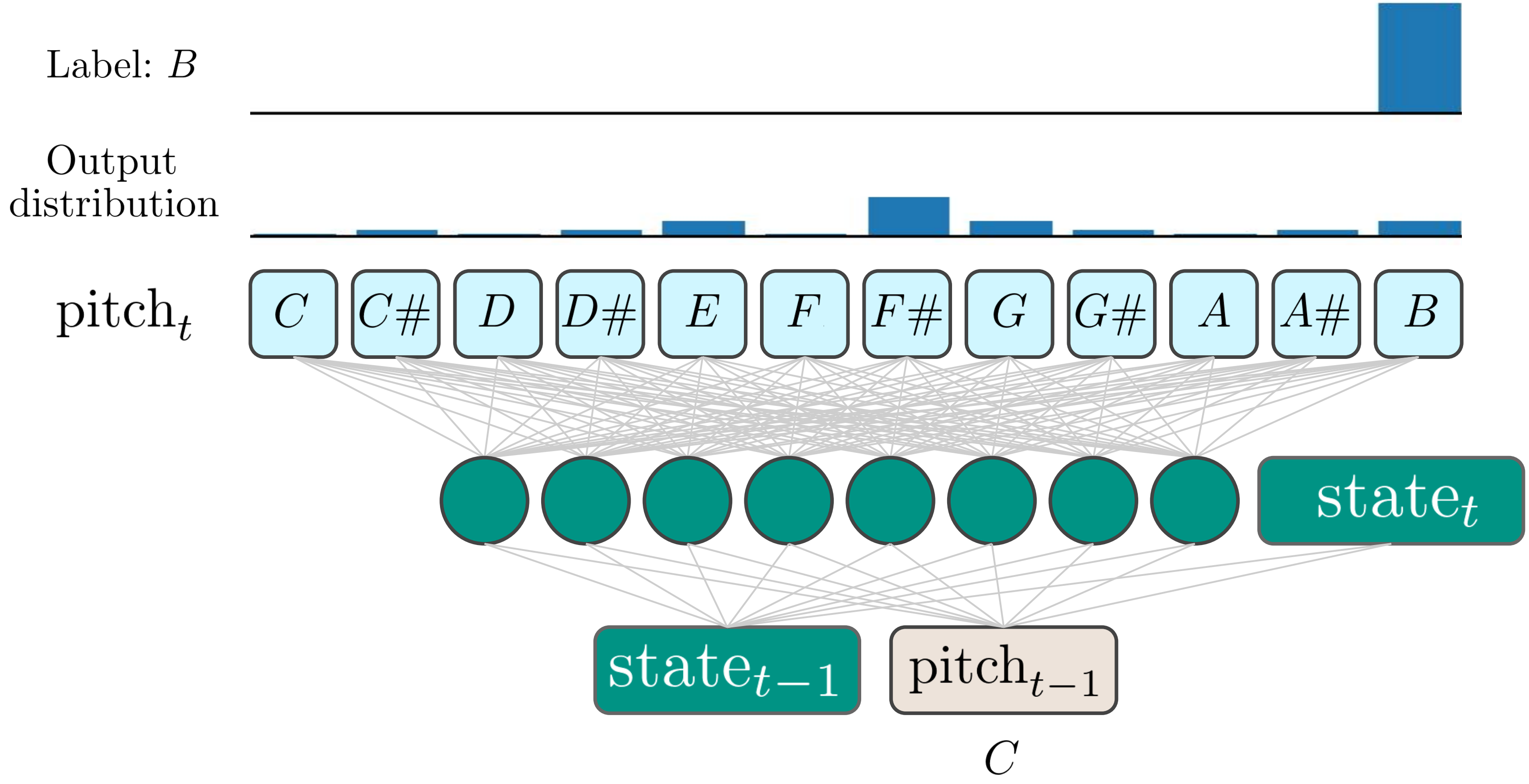
Improved model

$$P(n_1) \cdot P(n_2 \mid n_1) \cdot \ldots \cdot P(n_T \mid n_{T-1})$$

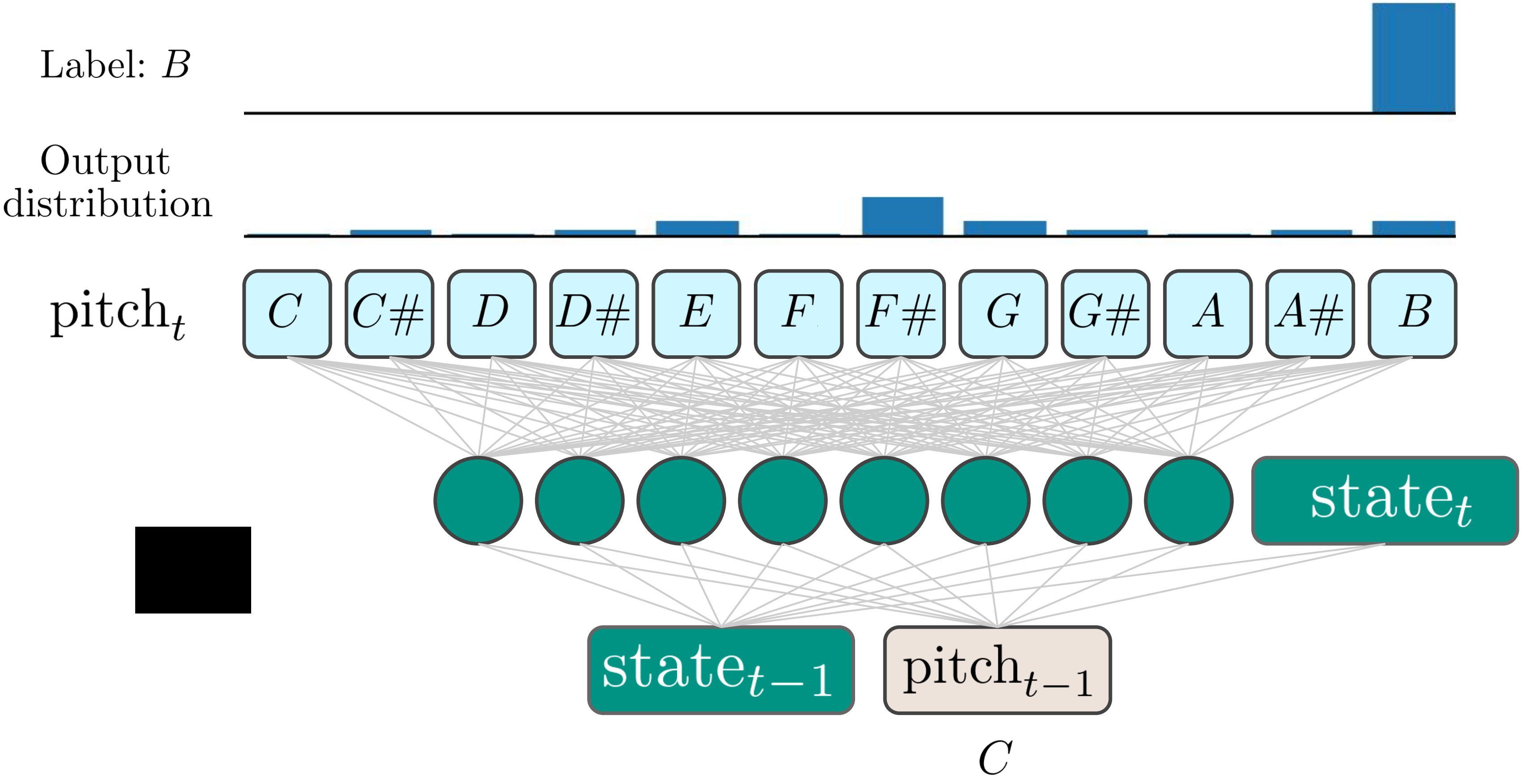
Generalized language model

$$P(n_1) \cdot P(n_2 \mid n_1) \cdot \ldots \cdot P(n_T \mid n_{T-1}, \ldots, n_1)$$

Neural network approach



Neural network approach



Language modeling of piano music

Music Transformer (Huang et al. 2018)

https://magenta.tensorflow.org/music-transformer

Language modeling with gestural control

Piano Genie (Donahue et al. 2018)

https://magenta.tensorflow.org/pianogenie

Demo: http://piano-genie.glitch.me

http://pear-olive.glitch.me

Language modeling of recipes

- 1 ½ teaspoon chicken brown water
- 1 teaspoon dry chopped leaves
- 1/3 cup shallows
- 10 oz brink custard
- ¼ cup bread liquid
- ½ cup baconfroots

Janelle Shane 2019
aiweirdness.com

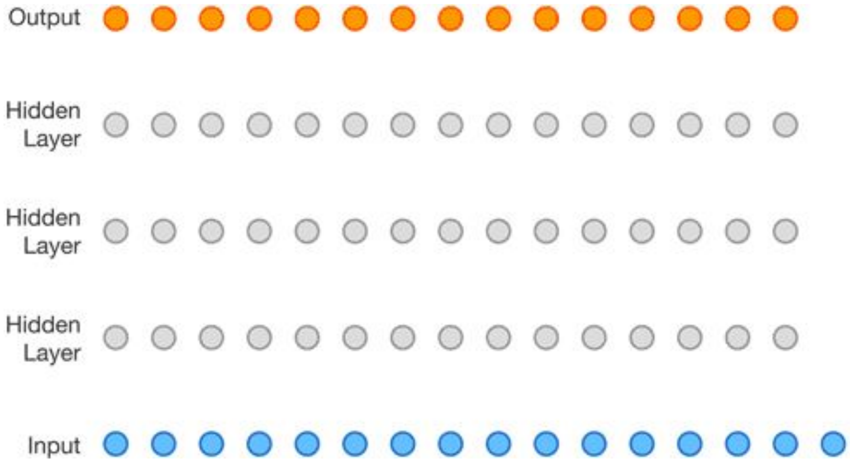
Instructions:

Dice the pulp of the eggplant and put it in a bowl with the vast stark rocks. Whip ½ pint of heavy cream. Add 4 Tbsp. brandy or rum to possibly open things that will never be wholly reported.

Language modeling of *audio*

WaveNet (van den Oord et al 2016)

https://deepmind.com/blog/wavenet-generative-model-raw-audio/

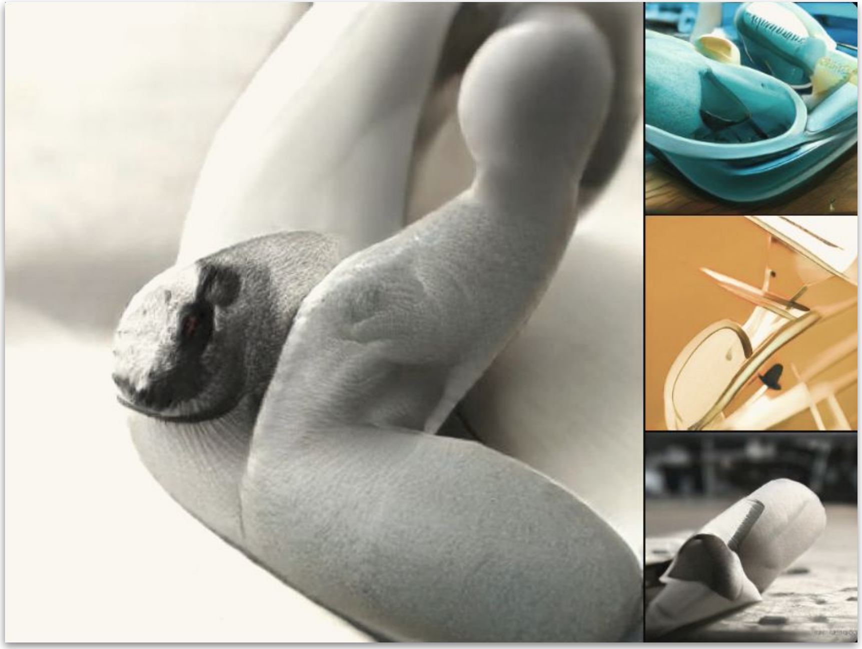


Artists using generative models

Artists using generative models

Helena Serin
Primrose (2019)

Artists using generative models



Mario Klingemann
Untitled (2018)

Artists using generative models

Holly Herndon & Jlin
Godmother (2018)

Thanks!

cdonahue@ucsd.edu

chrisdonahue.com