PAT 464/564 (Winter 2026)

Generative AI for Music & Audio Creation

Lecture 12: Variational Autoencoders

Instructor: Hao-Wen Dong

Representative Types of Deep Generative Models

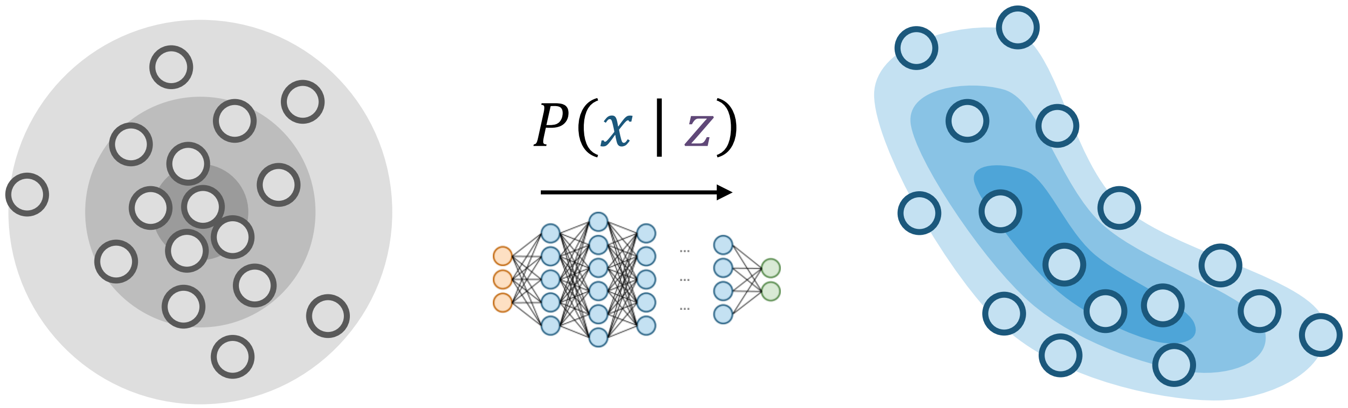
- **Deep autoregressive models**
  - Recurrent neural network (RNN)
  - Long short-term memory (LSTM)
  - Transformer model
- **Deep latent variable models**
  - Variational autoencoder (VAE) **Today's topic!**
  - Generative adversarial network (GAN)
  - Diffusion model
  - Flow-based model
- *And many others…*

Deep Latent Variable Models

Deep Latent Variable Models

- **Intuition**: Learn to map a known distribution to the data distribution

Known distribution **Data distribution**

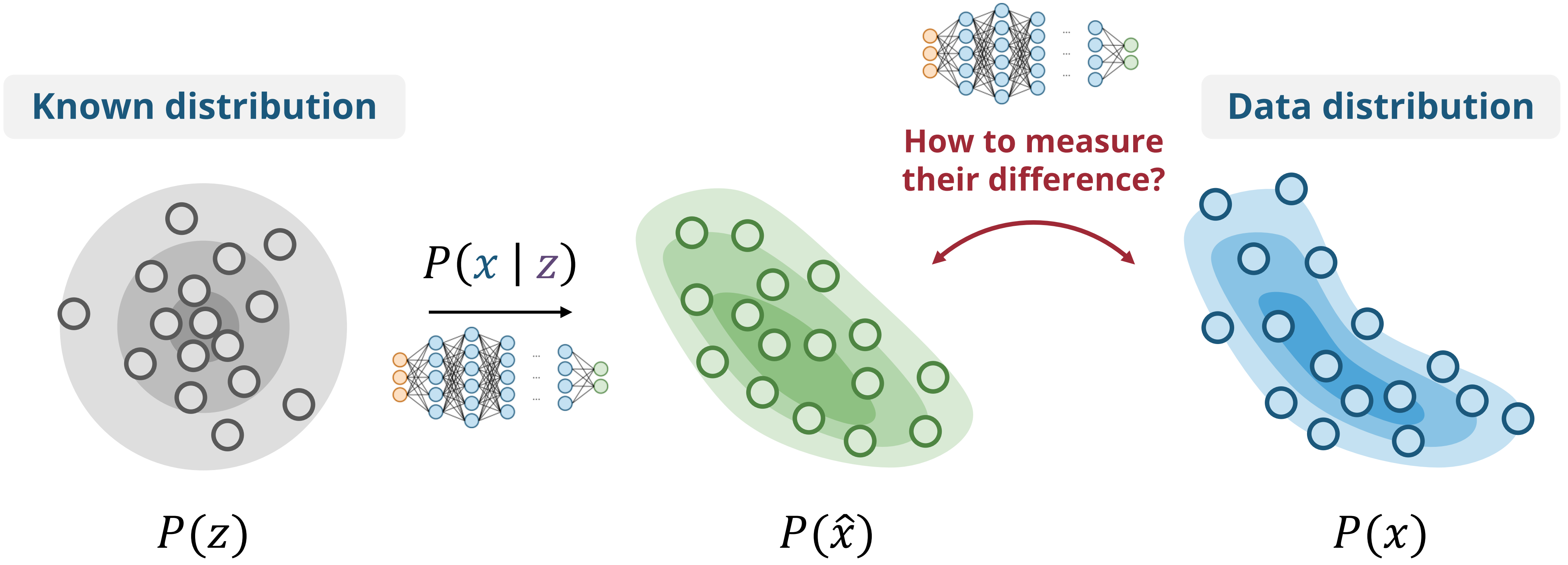


$P(z)$ $P(x)$

$$P(x) = P(z)\ P(x \mid z)$$

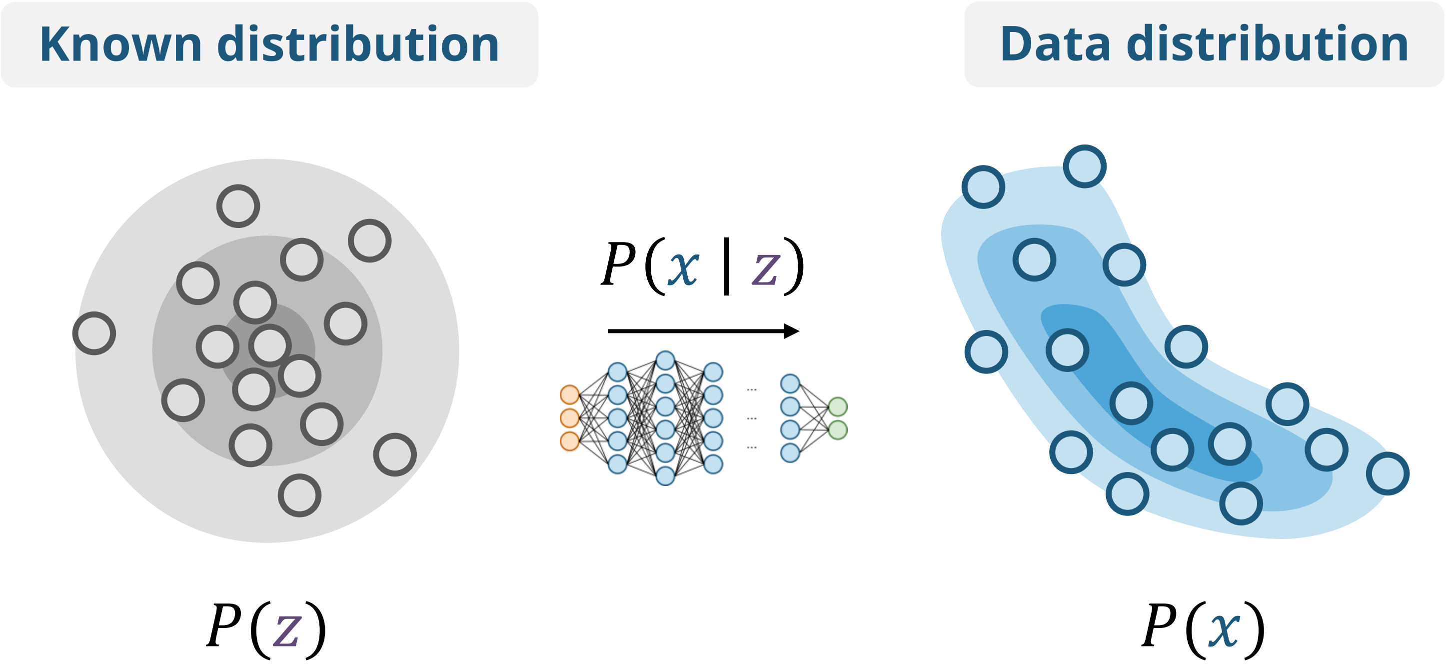
Deep Latent Variable Models

- **Intuition**: Learn to map a known distribution to the data distribution



Deep Latent Variable Models

- **Intuition**: Learn to map a known distribution to the data distribution



$$P(x) = P(z)\ P(x \mid z)$$

Deep Latent Variable Models

- **Intuition**: Learn to map a known distribution to the data distribution

What we want the model to learn!

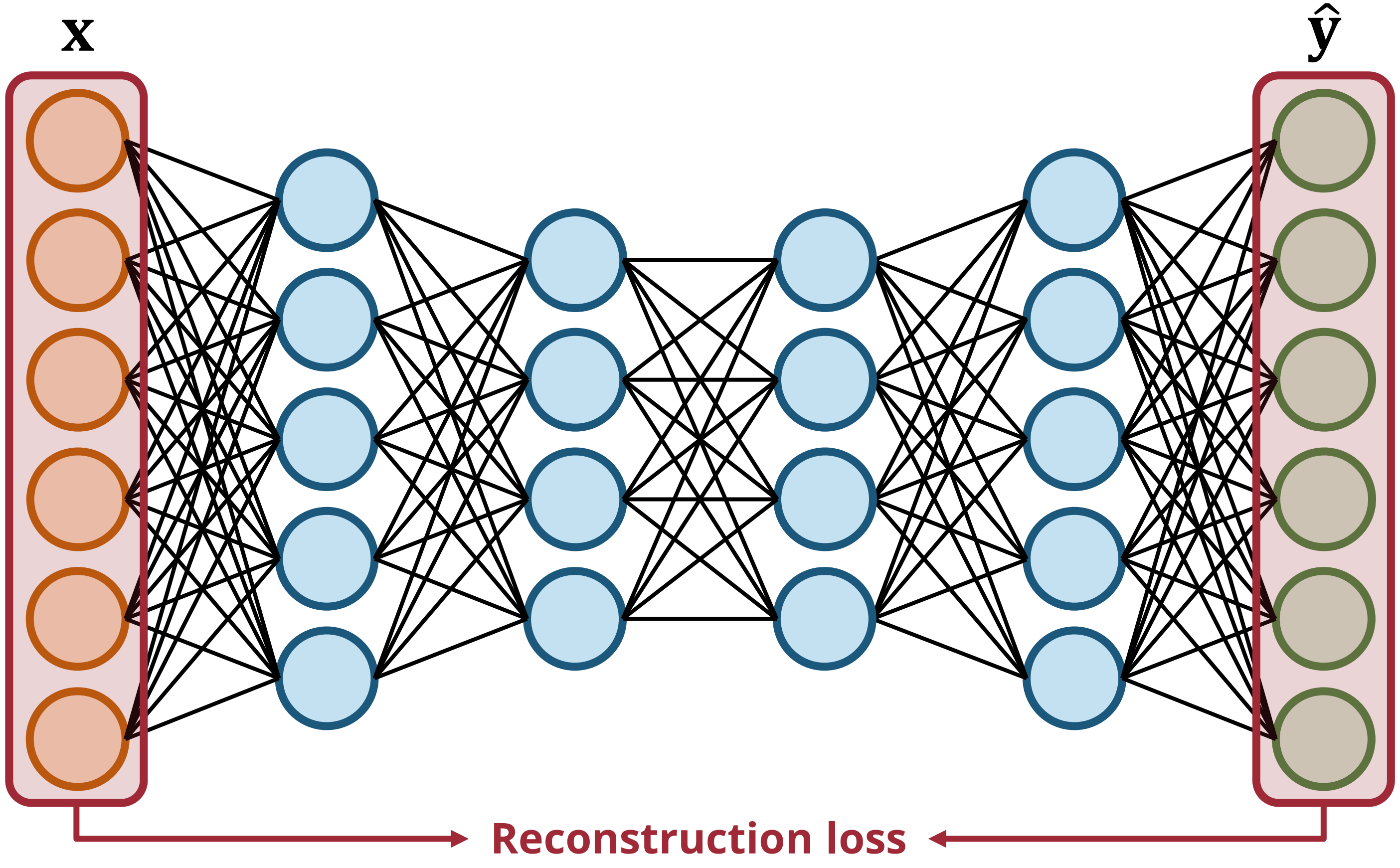
$$P(x) = P(z)\boxed{P(x \mid z)}$$

Data distribution **Latent distribution**

Autoencoders

Autoencoders

- A neural network where the **input and output are the same**



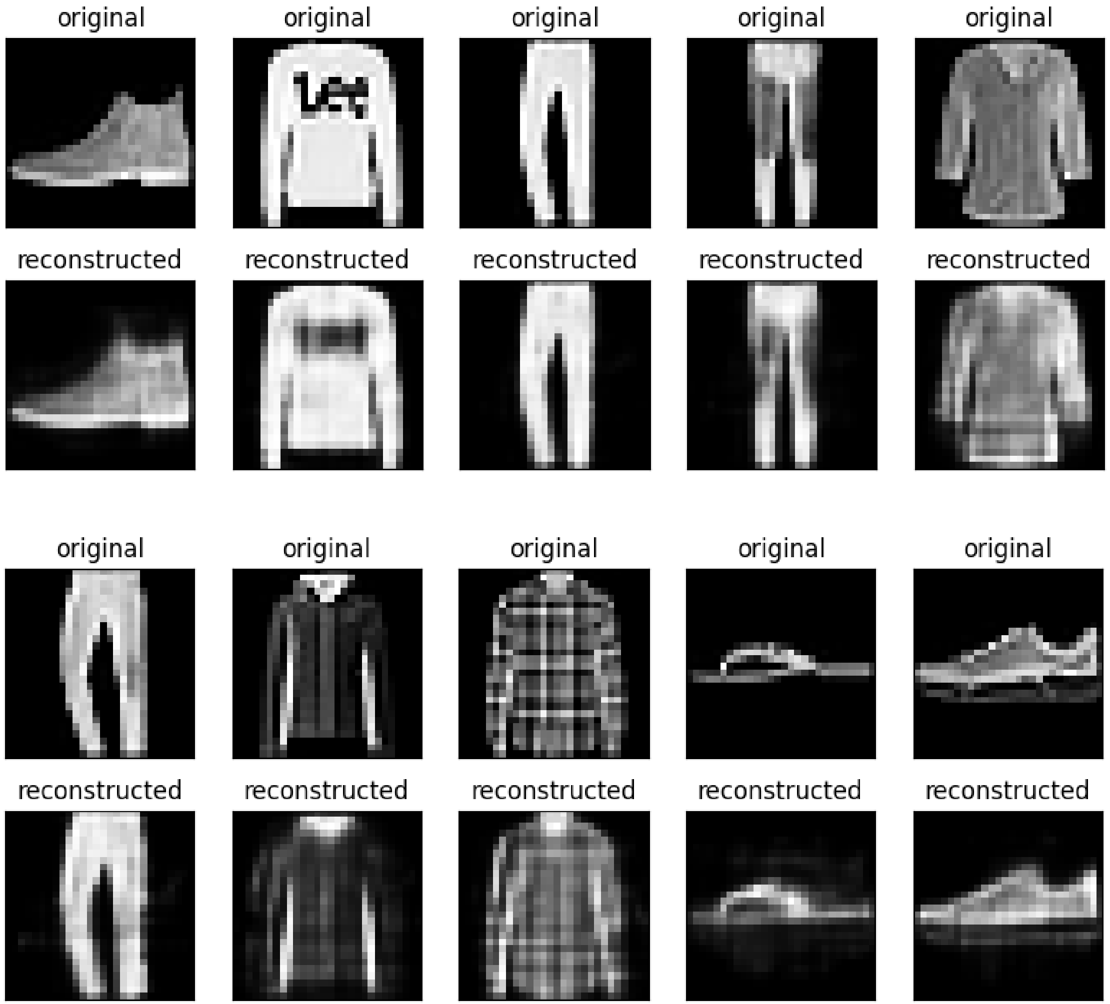
Autoencoders: Reconstruction Examples

Original

Reconstructed

Original

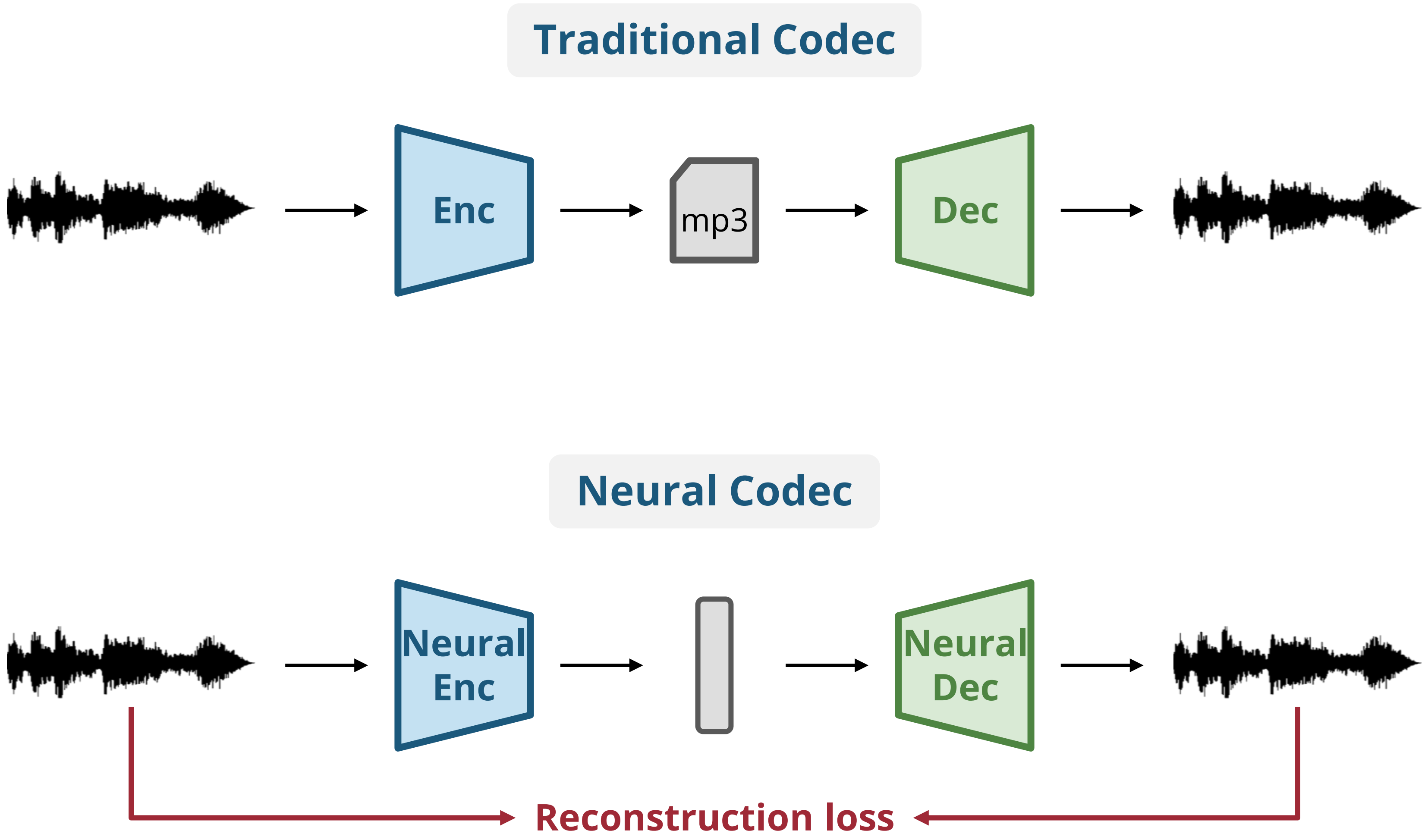
Reconstructed



(Source: tensorflow.org)

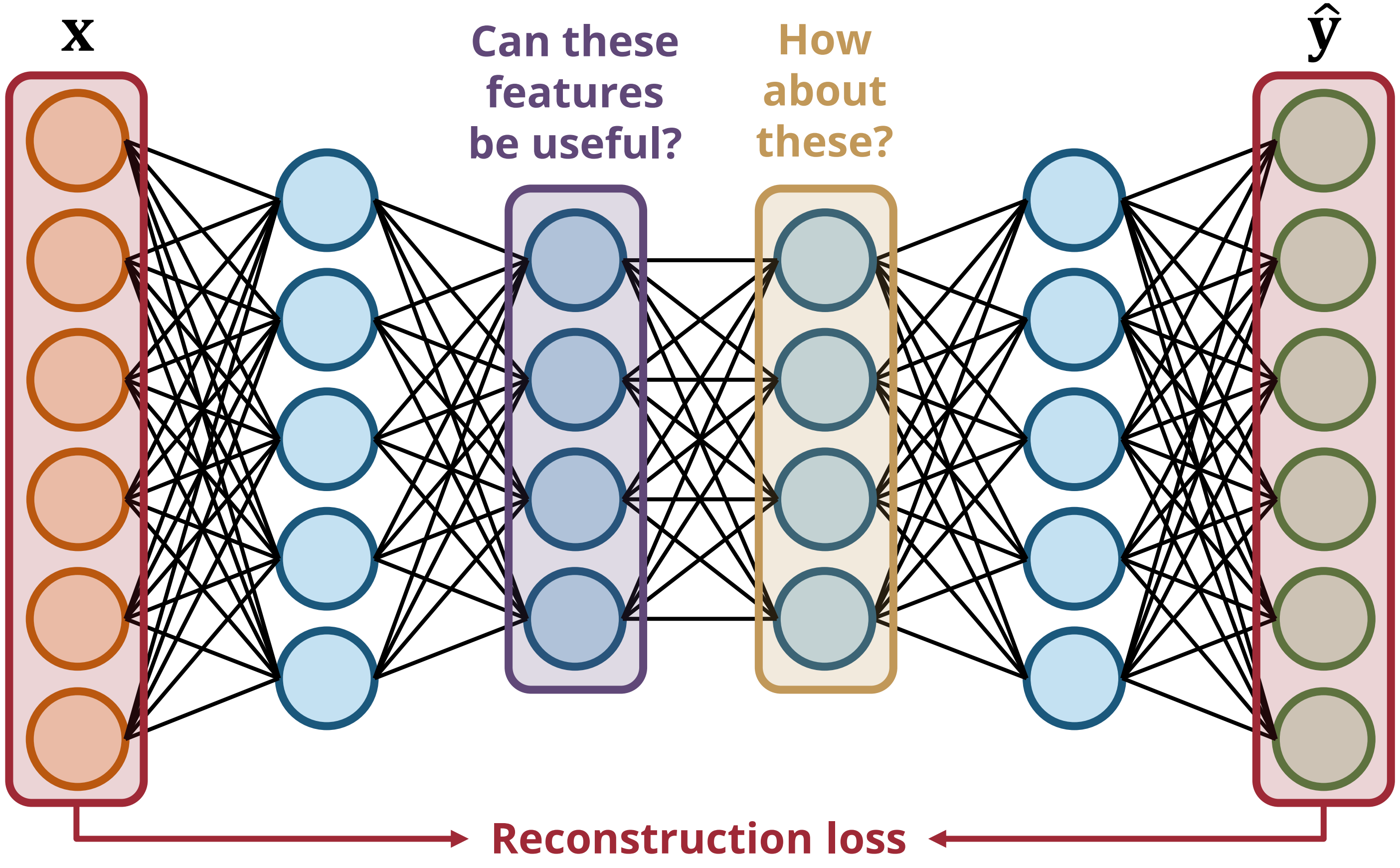
tensorflow.org/tutorials/generative/autoencoder

Codec is an Autoencoder



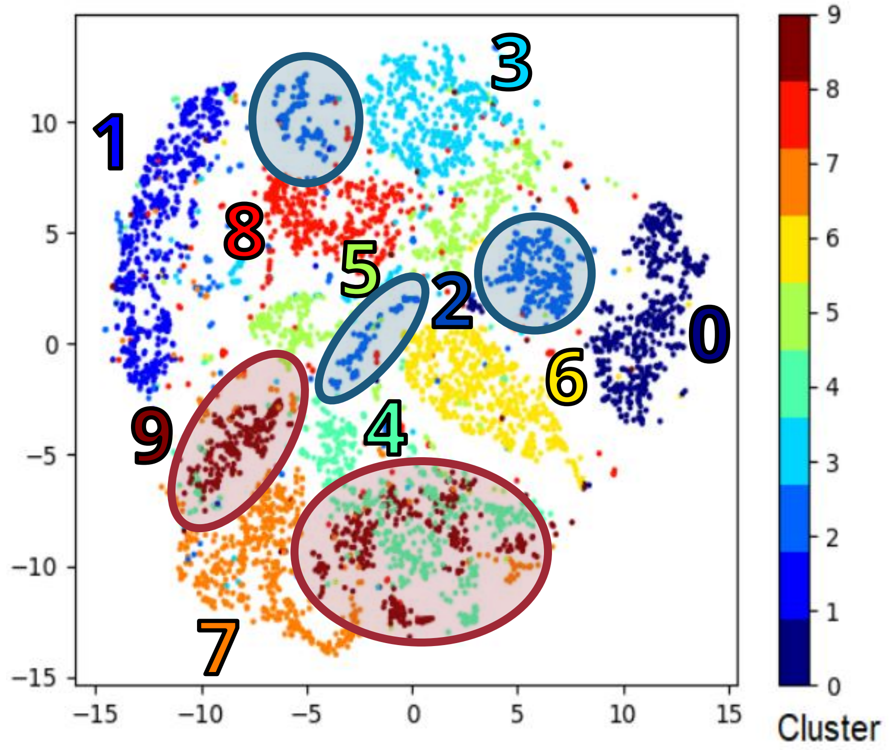
Autoencoders

- A neural network where the **input and output are the same**

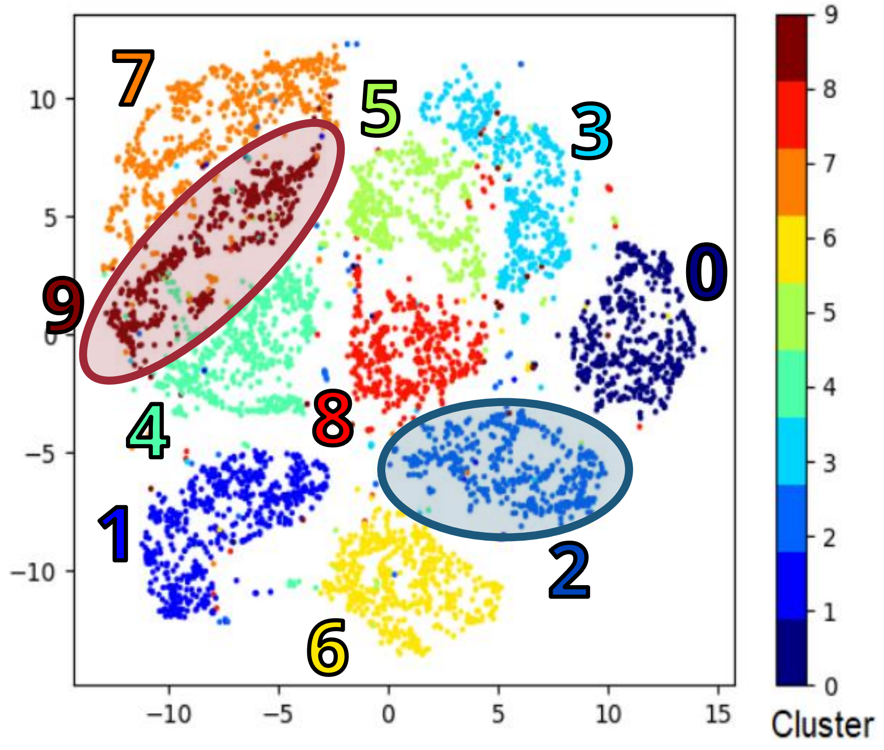


Unsupervised Clustering with an Autoencoder

Raw pixels



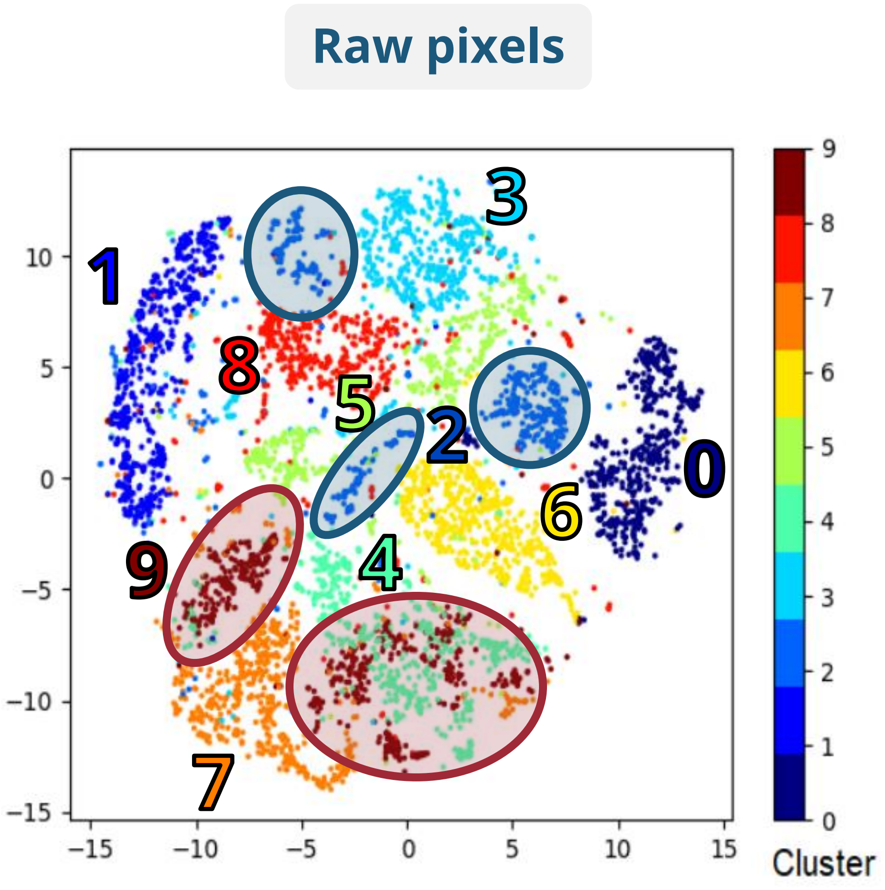
AE-encoded



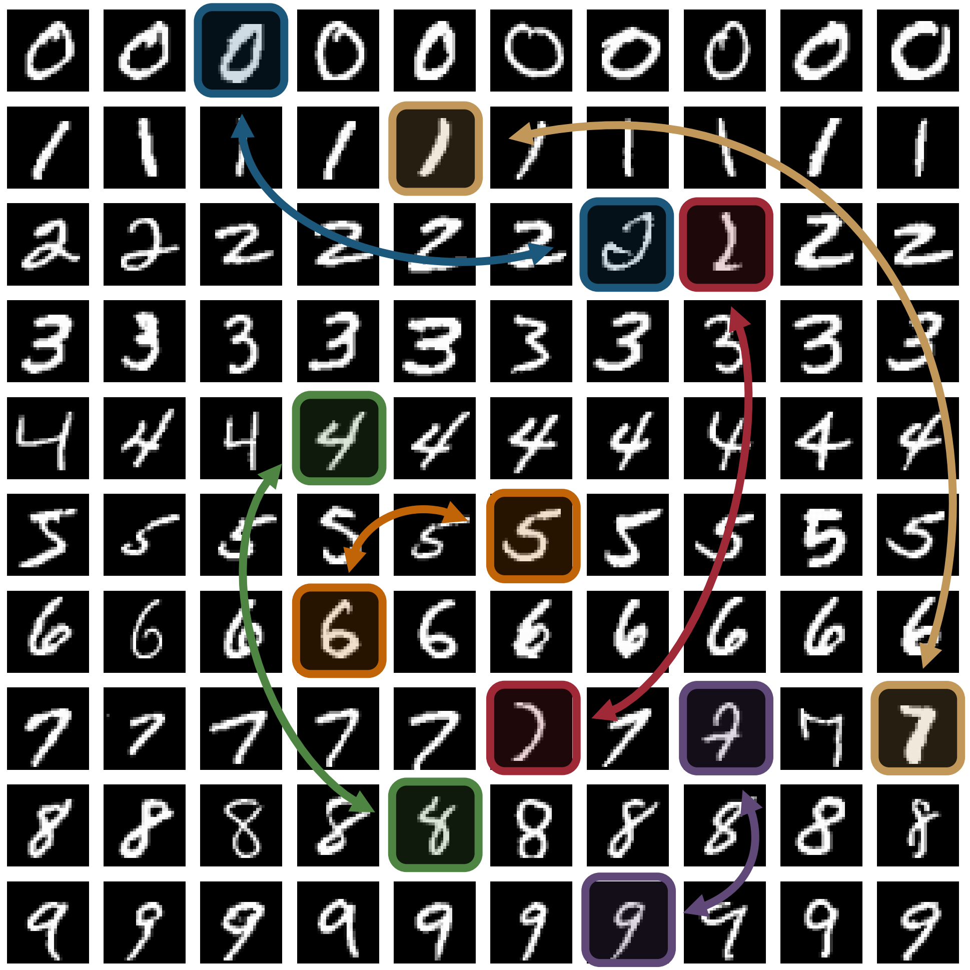
(Source: Aljalbout et al., 2020)

Elie Aljalbout, Vladimir Golkov, Yawar Siddiqui, Maximilian Strobel, and Daniel Cremers, "Clustering with Deep Learning: Taxonomy and New Methods," *arXiv preprint arXiv:1801.07648*, 2018.

Unsupervised Clustering with an Autoencoder



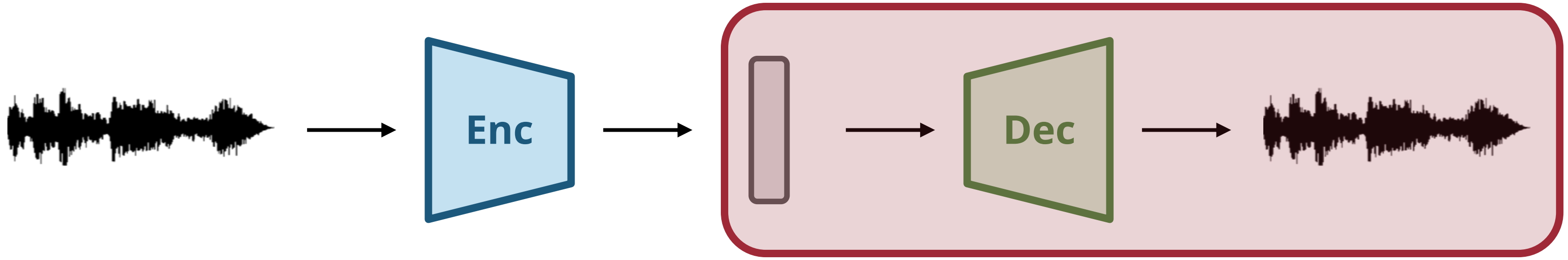
(Source: Aljalbout et al., 2020)



Elie Aljalbout, Vladimir Golkov, Yawar Siddiqui, Maximilian Strobel, and Daniel Cremers, "Clustering with Deep Learning: Taxonomy and New Methods," *arXiv preprint arXiv:1801.07648*, 2018.

Variational Autoencoder (VAE)

Autoencoder

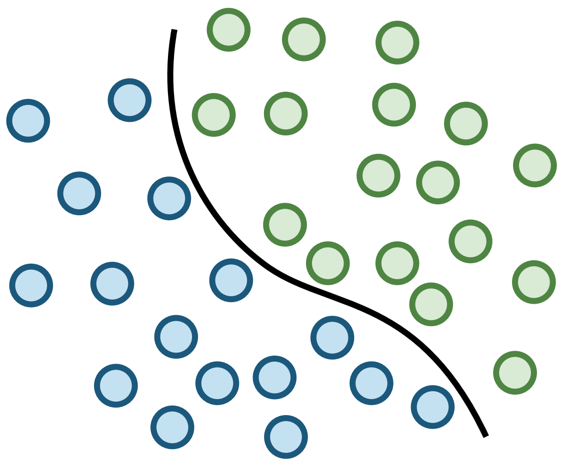


Isn’t this like a generative model?

What exactly is a generative model?

Discriminative vs Generative Models

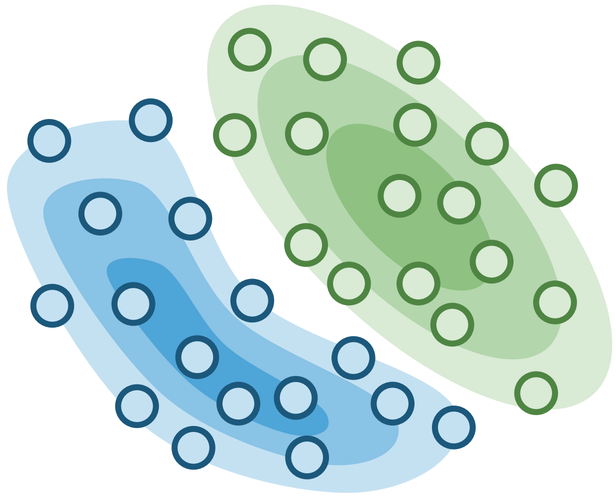
Discriminative



Discriminative models learn the decision boundary

$P(y|x)$

Generative



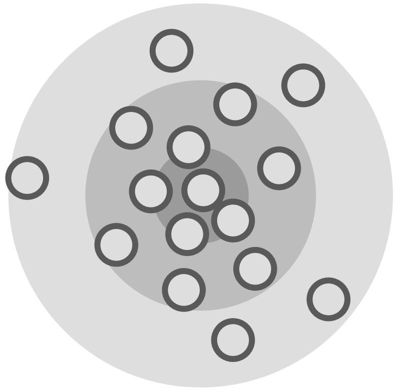
Generative models learn the underlying distribution

$P(x)$ or $P(x|y)$

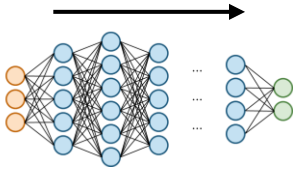
Generating Data from a Random Distribution

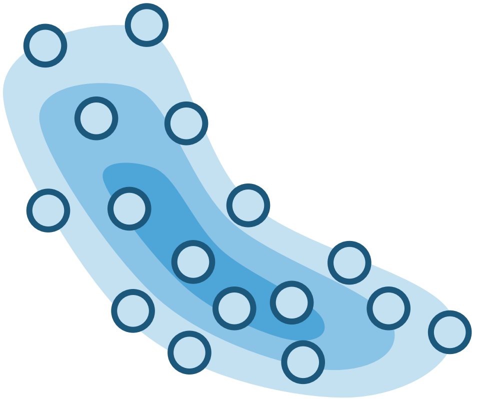
Random distribution

Data distribution



$P(z)$

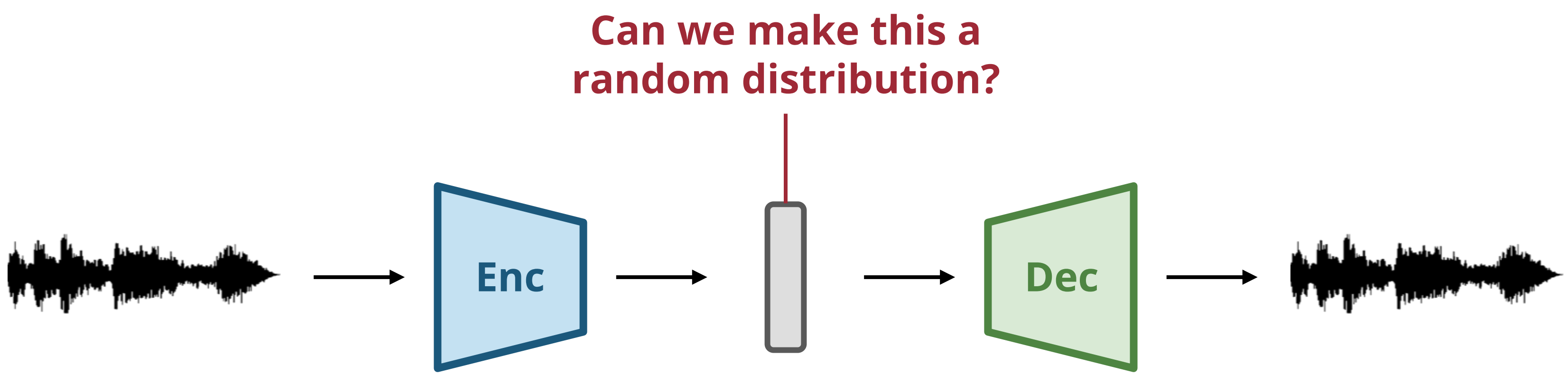




$P(x)$

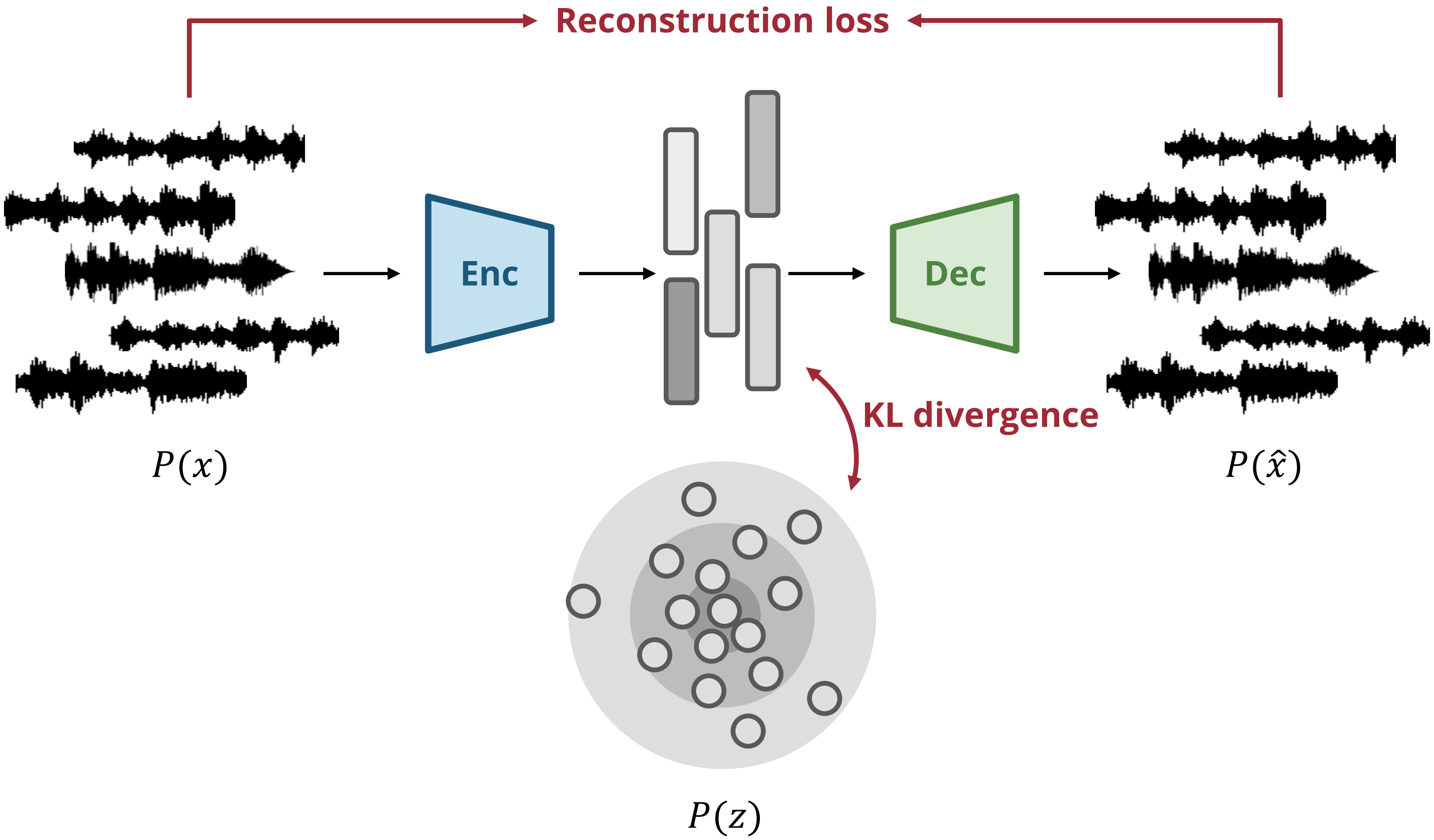
If we can learn this mapping, we can then generate new samples from the data distribution

Variational Autoencoder (VAE)



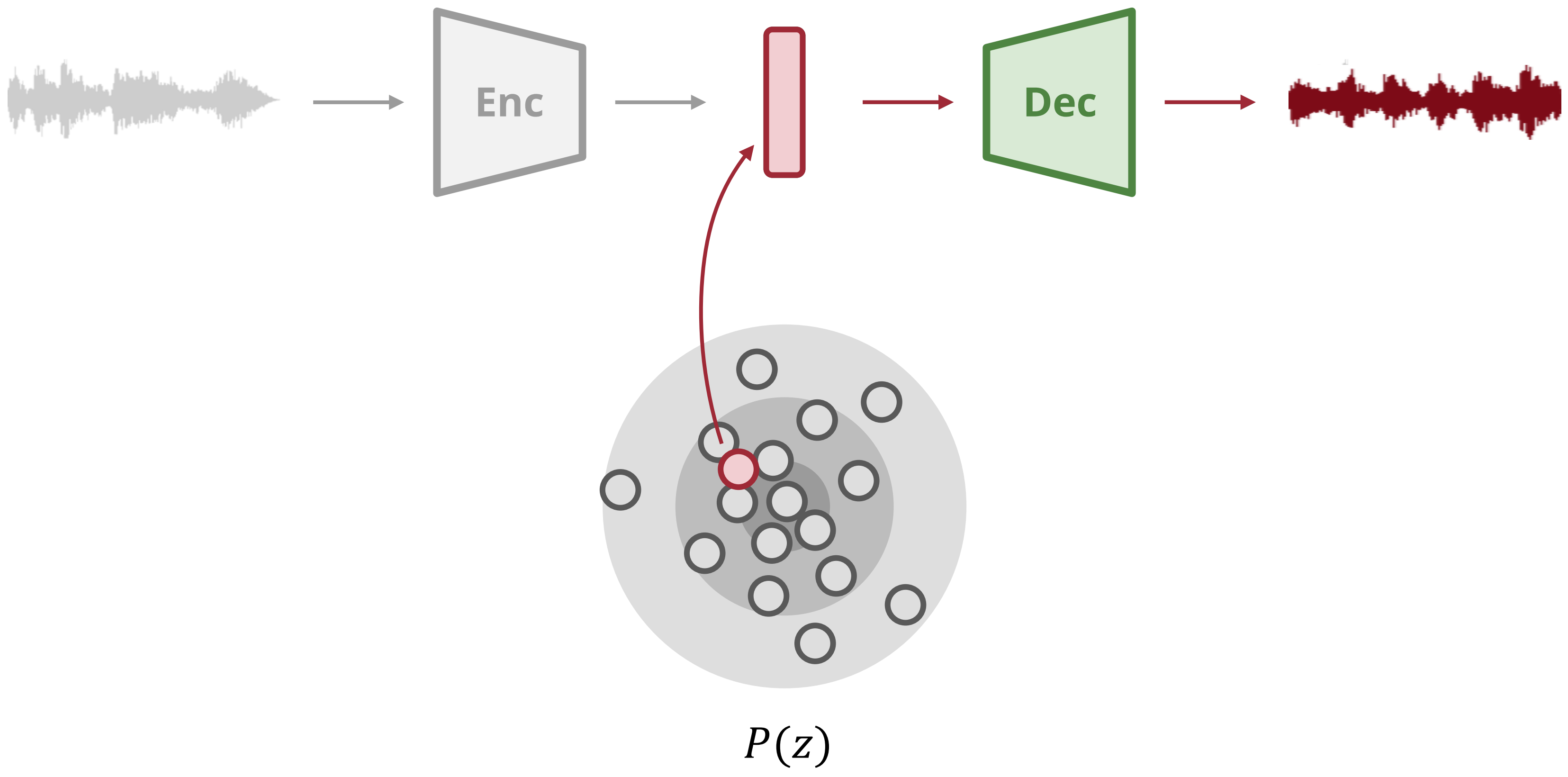
Diederik P Kingma and Max Welling, "Auto-Encoding Variational Bayes," *ICLR*, 2014.

Variational Autoencoder (VAE): Training



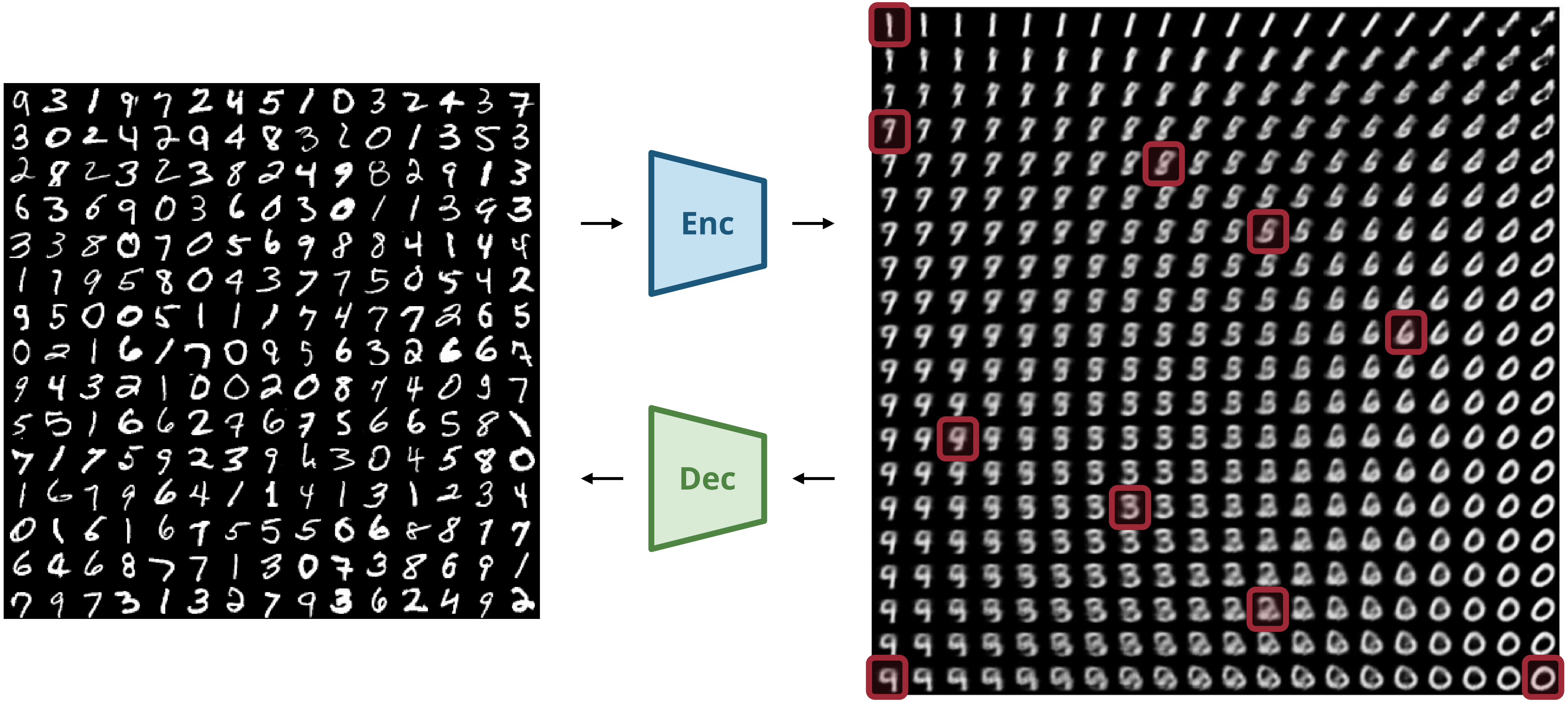
Diederik P Kingma and Max Welling, "Auto-Encoding Variational Bayes," *ICLR*, 2014.

Variational Autoencoder (VAE): Generation



Diederik P Kingma and Max Welling, “Auto-Encoding Variational Bayes,” *ICLR*, 2014.

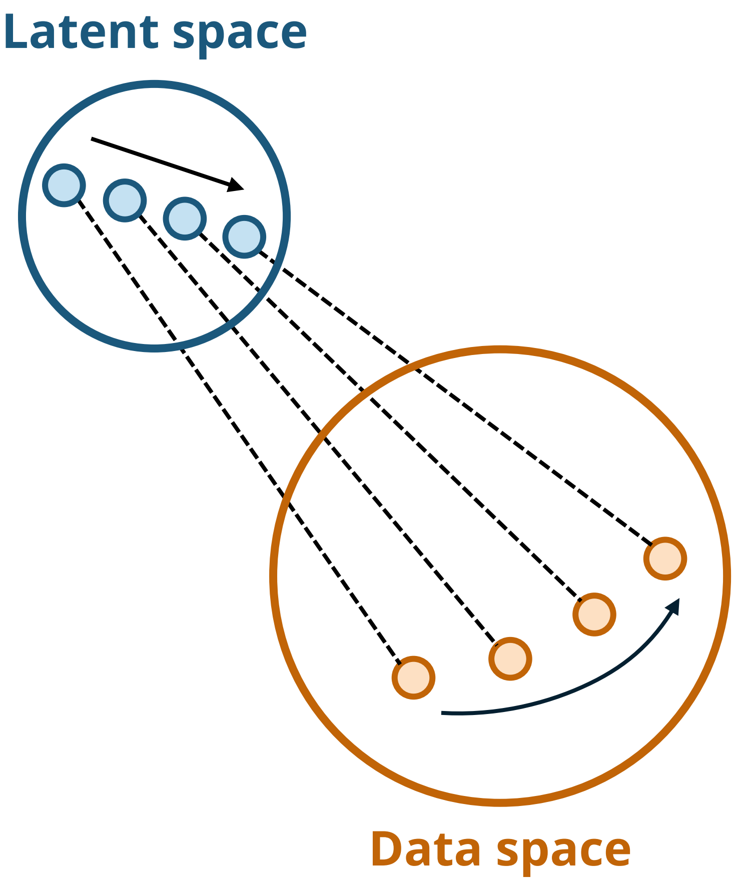
What does a VAE learn?

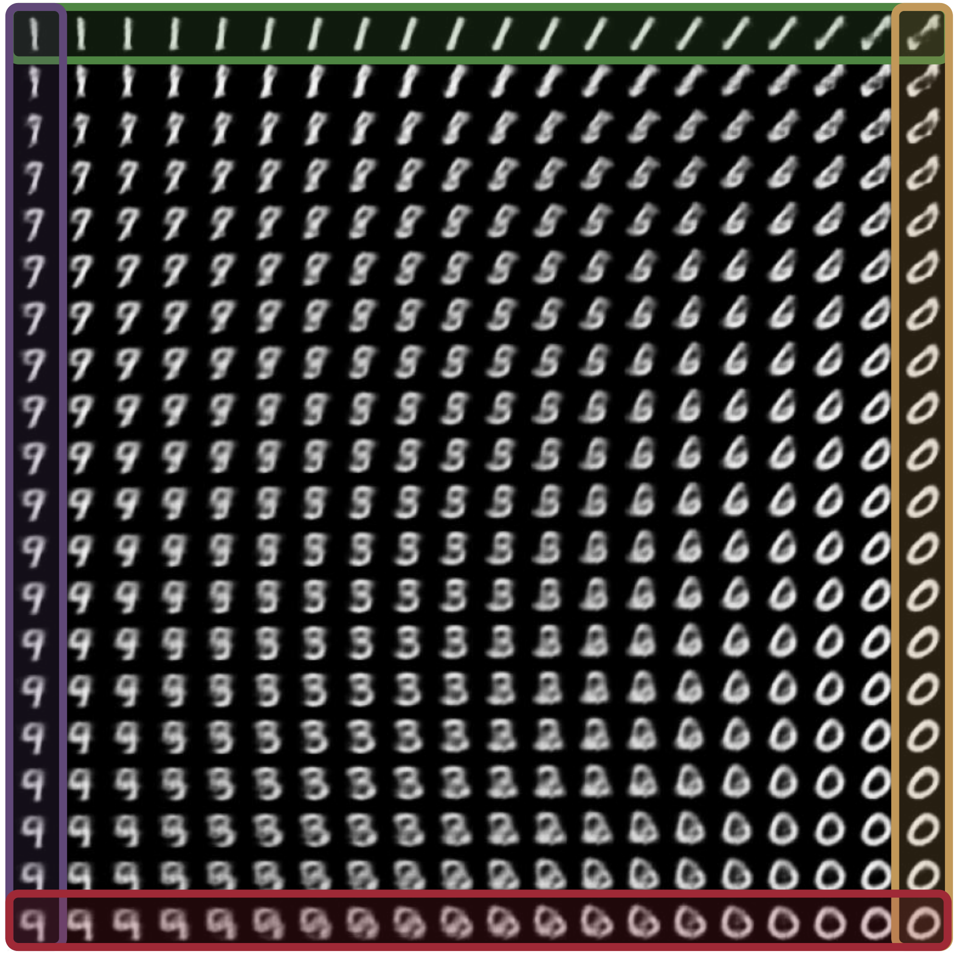


(Source: tensorflow.org)

tensorflow.org/tutorials/generative/cvae

Latent Space Interpolation of a VAE

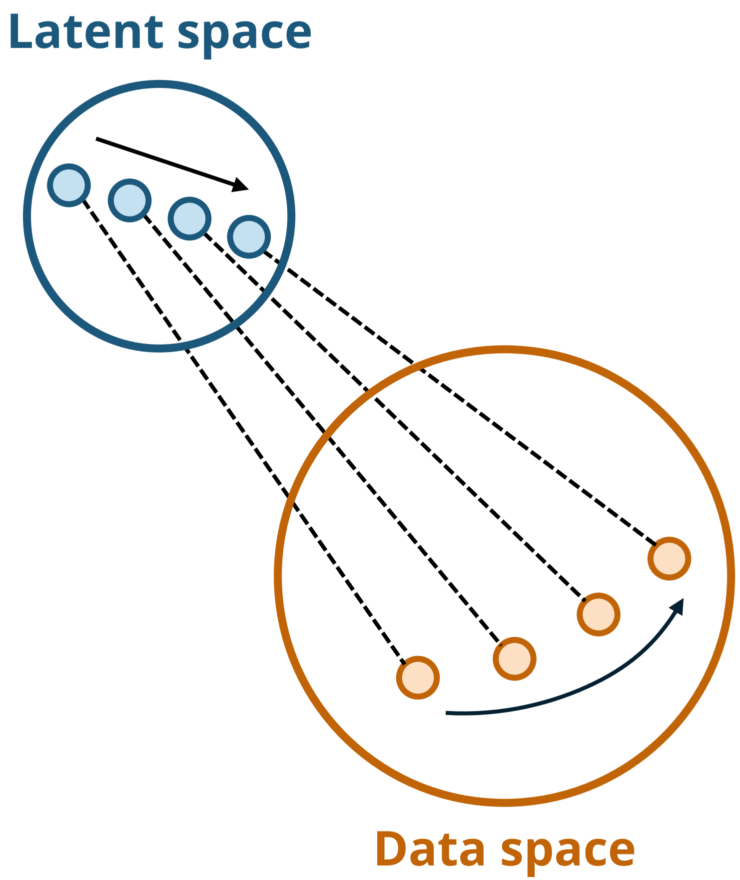


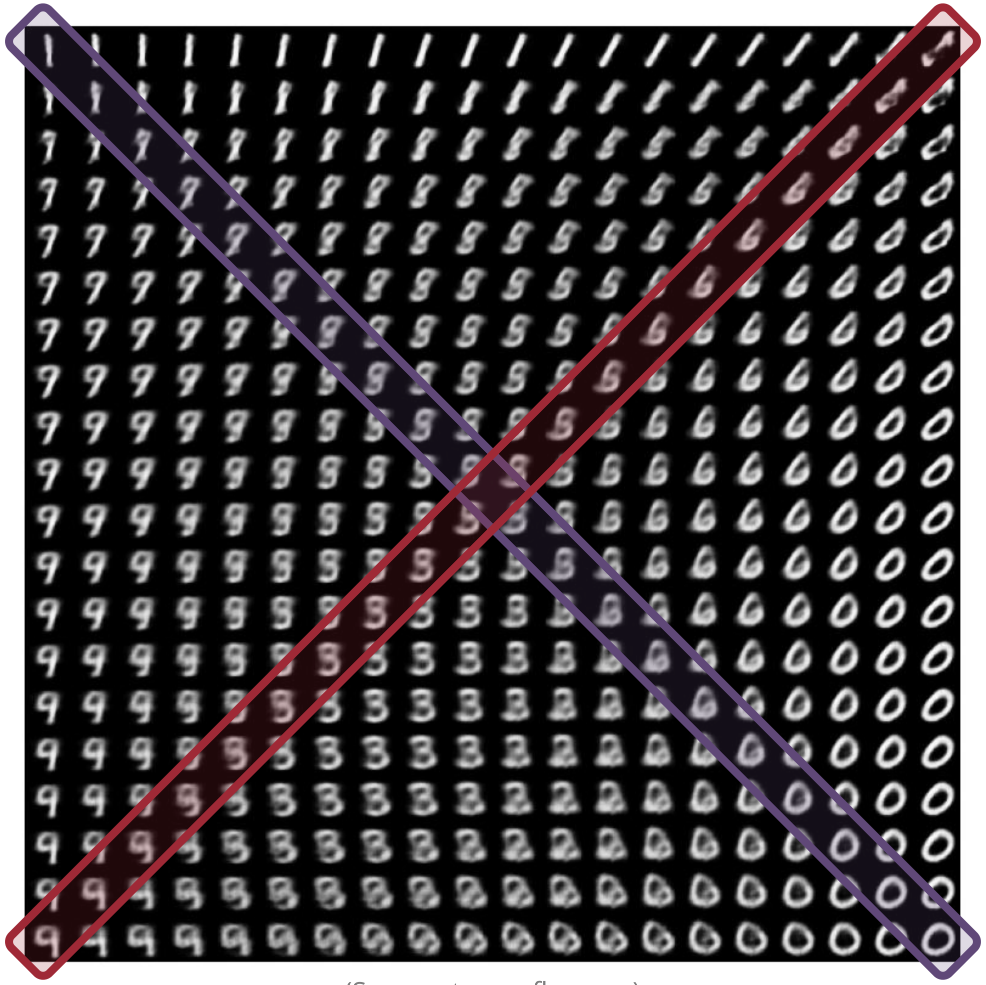


(Source: tensorflow.org)

tensorflow.org/tutorials/generative/cvae

Latent Space Interpolation of a VAE

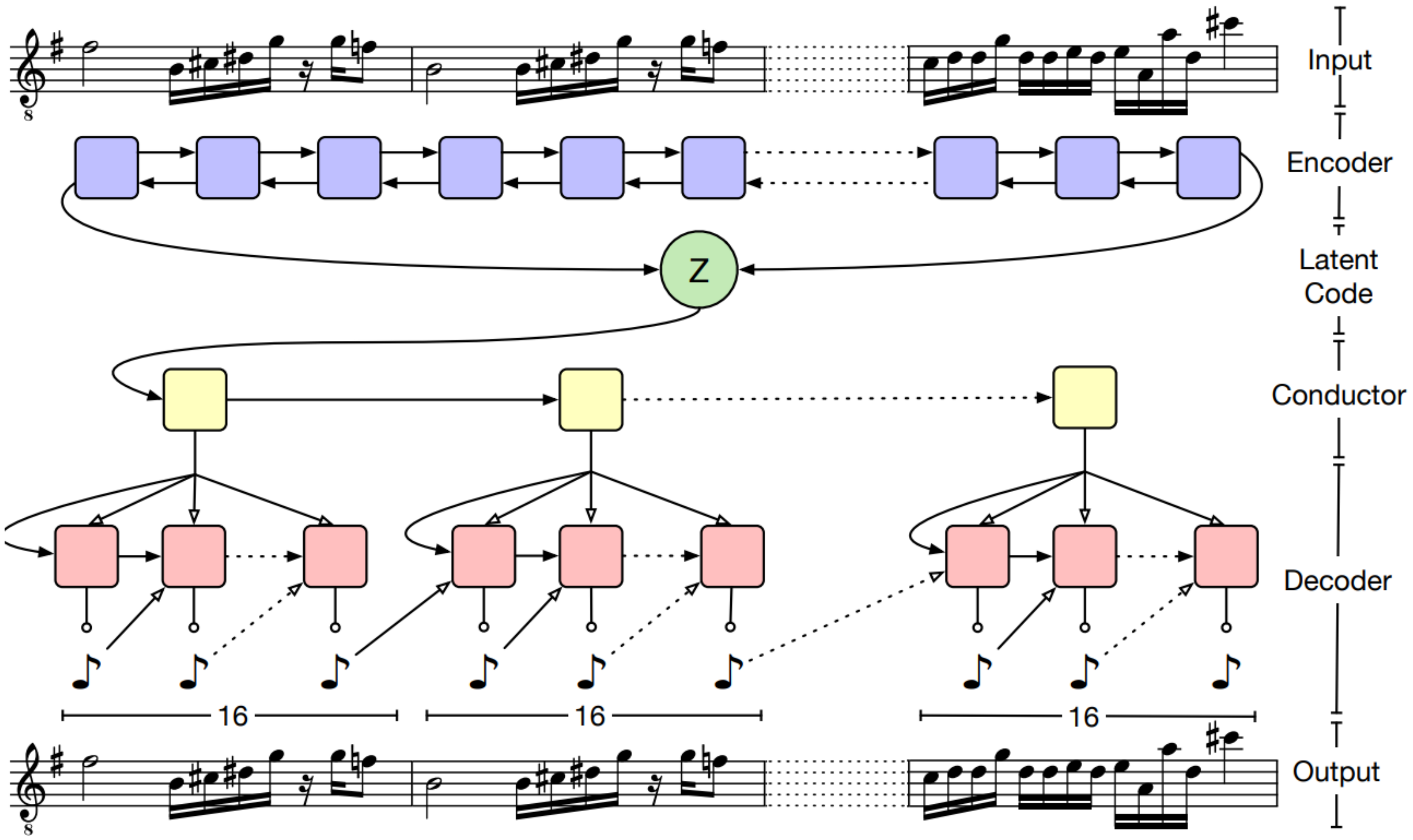




(Source: tensorflow.org)

tensorflow.org/tutorials/generative/cvae

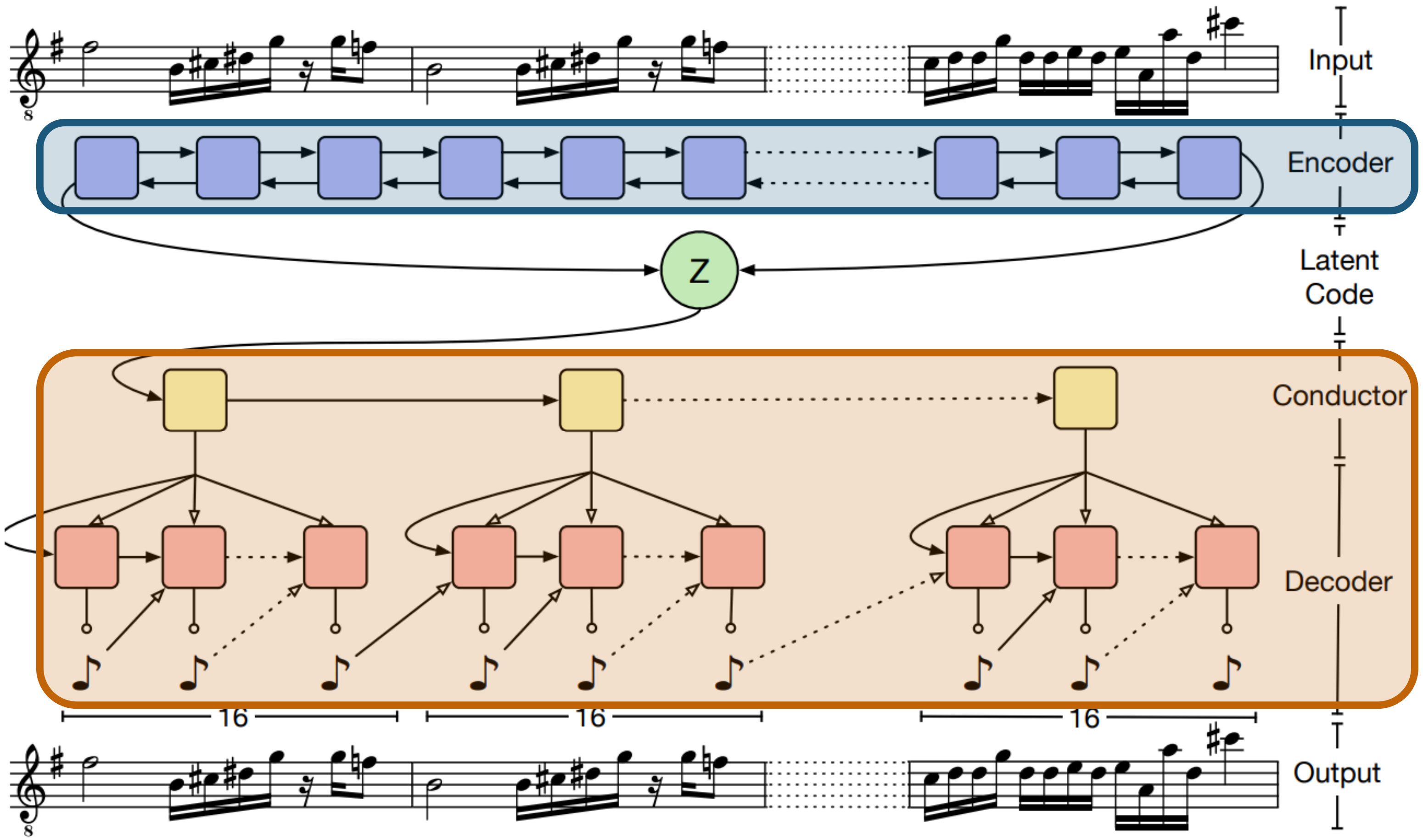
MusicVAE: A VAE for Symbolic Music (Roberts et al., 2018)



(Source: Roberts et al., 2018)

Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck, "A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music," *ICML*, 2018.

MusicVAE: A VAE for Symbolic Music (Roberts et al., 2018)



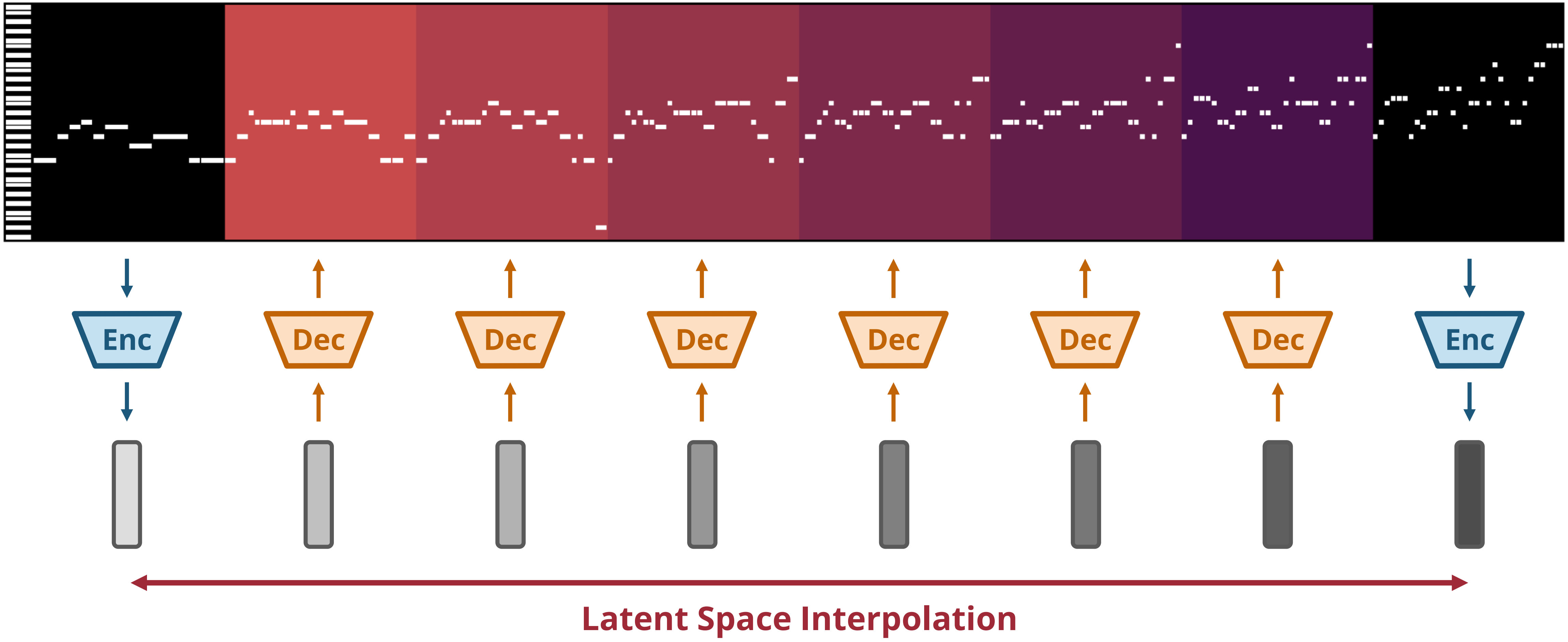
Bidirectional encoder

Hierarchical decoder

(Source: Roberts et al., 2018)

Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck, "A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music," *ICML*, 2018.

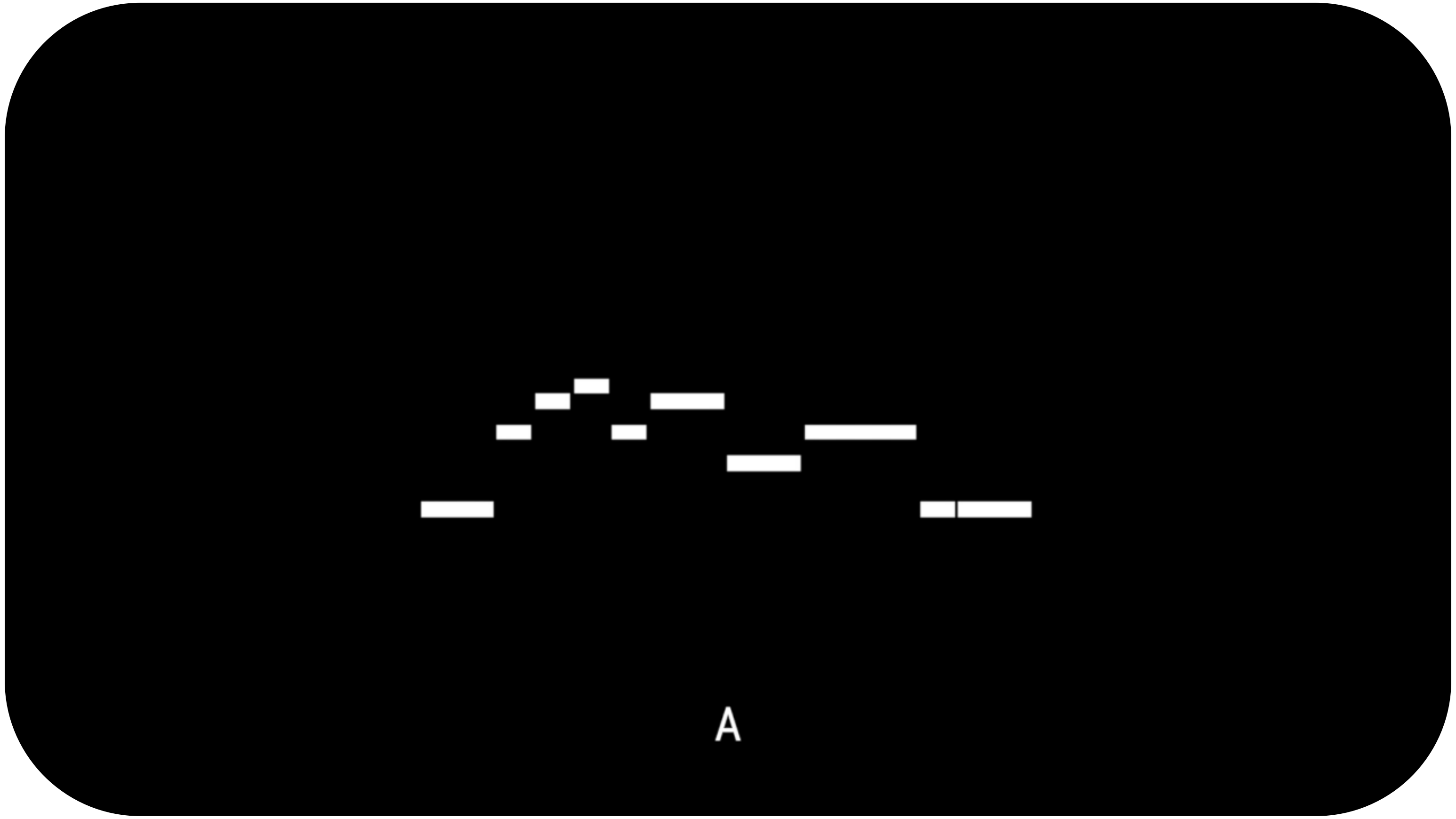
Latent Space Interpolation for MusicVAE (Roberts et al., 2018)



(Source: Roberts et al., 2018)

Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck, "A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music," *ICML*, 2018.

Latent Space Interpolation for MusicVAE (Roberts et al., 2018)



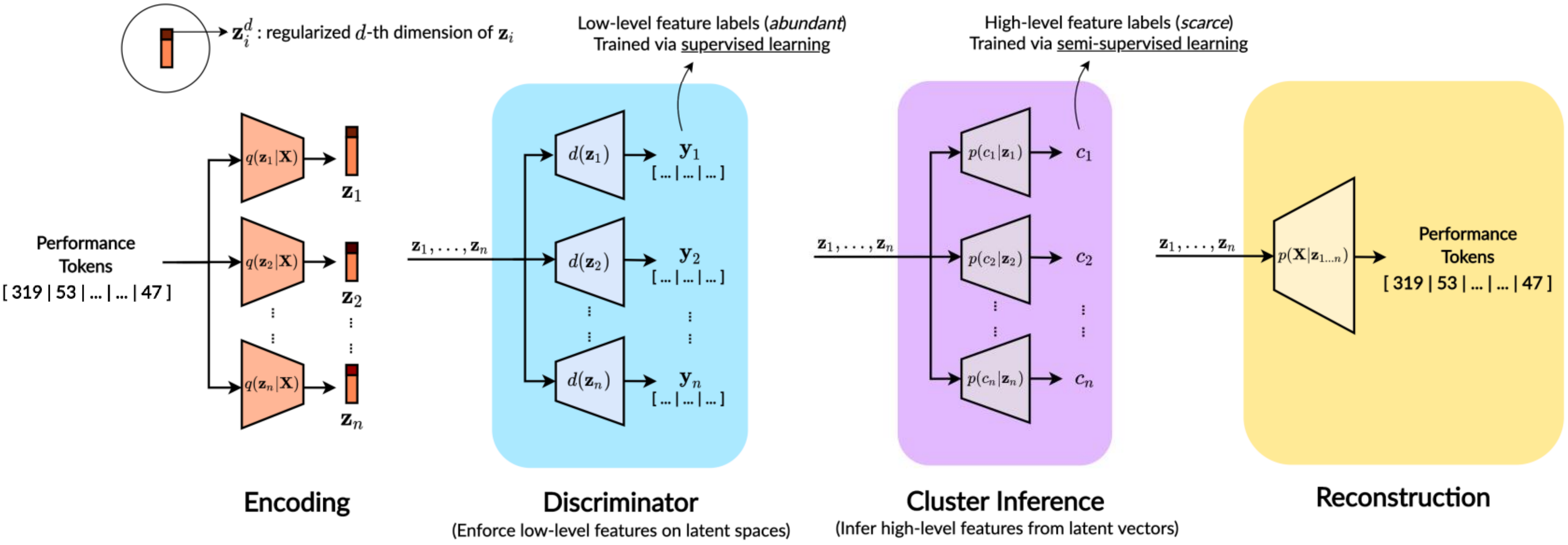
(Source: Roberts et al., 2018)

goo.gl/magenta/musicvae-examples

Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck, "A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music," *ICML*, 2018.

Disentangling the Latent Variables

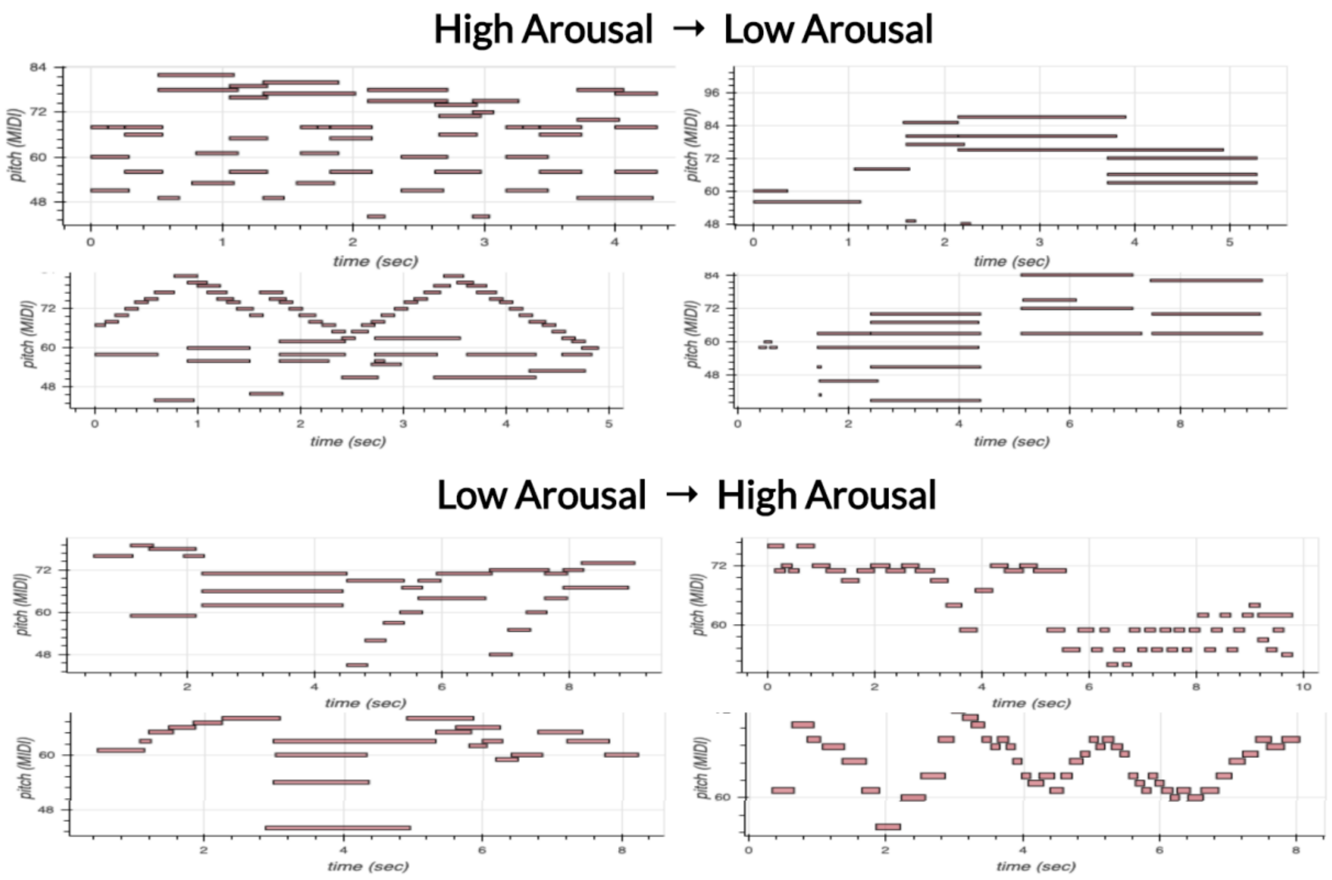
Music FaderNet (Tan & Herremans, 2020)



(Source: Tan & Herremeans, 2020)

Hao Hao Tan and Dorien Herremans, "Music FaderNets: Controllable Music Generation Based On High-Level Features via Low-Level Feature Modelling," ISMIR, 2020.

Music FaderNet (Tan & Herremans, 2020)



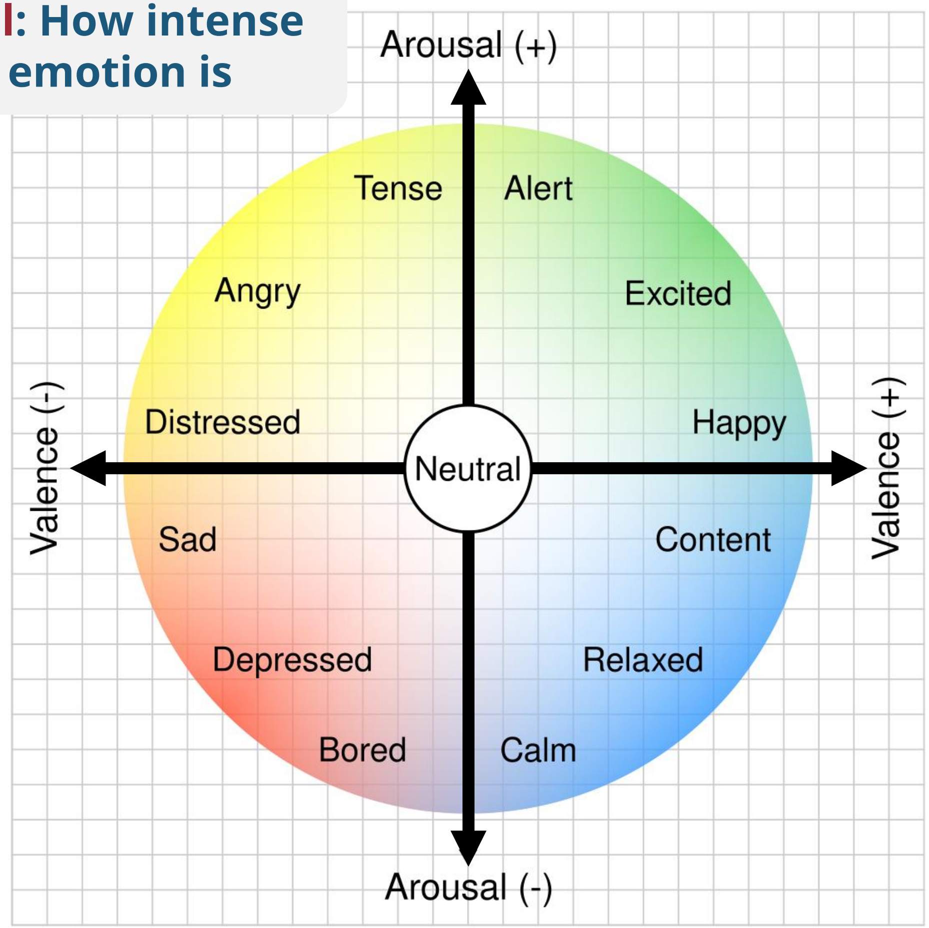
(Source: Tan & Herremeans, 2020)

music-fadernets.github.io

Hao Hao Tan and Dorien Herremans, "Music FaderNets: Controllable Music Generation Based On High-Level Features via Low-Level Feature Modelling," ISMIR, 2020.

Valence-Arousal Model for Emotion

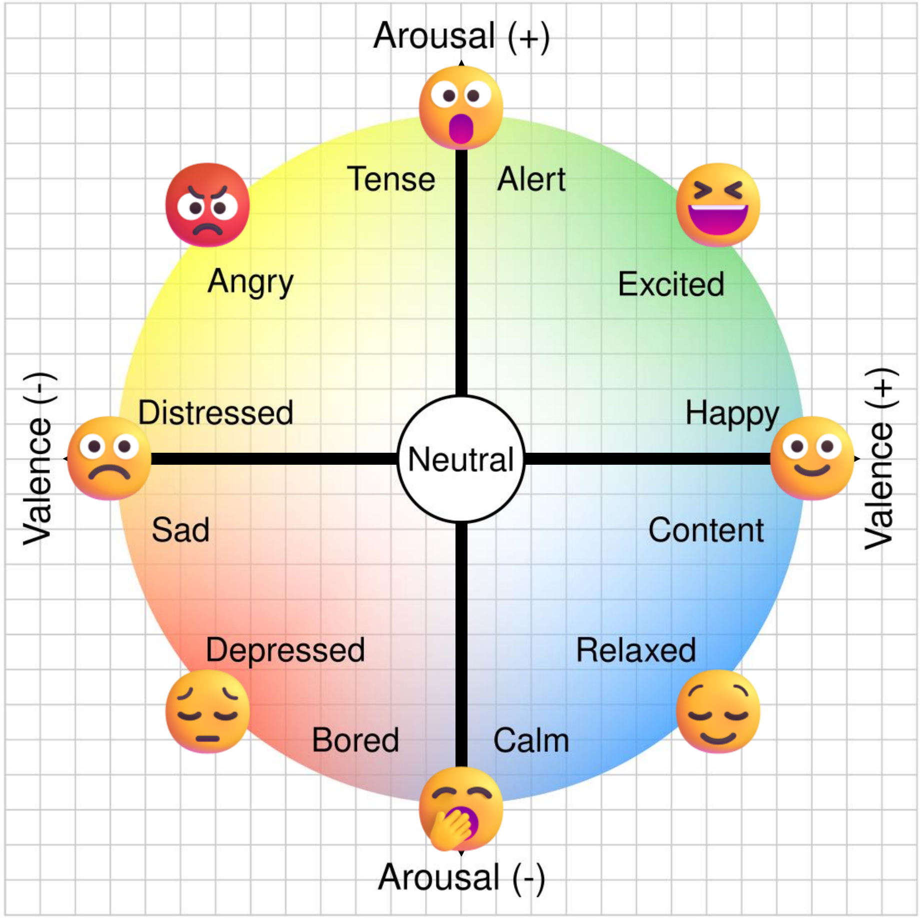
Arousal: How intense the emotion is



Valence: How pleasant the emotion is

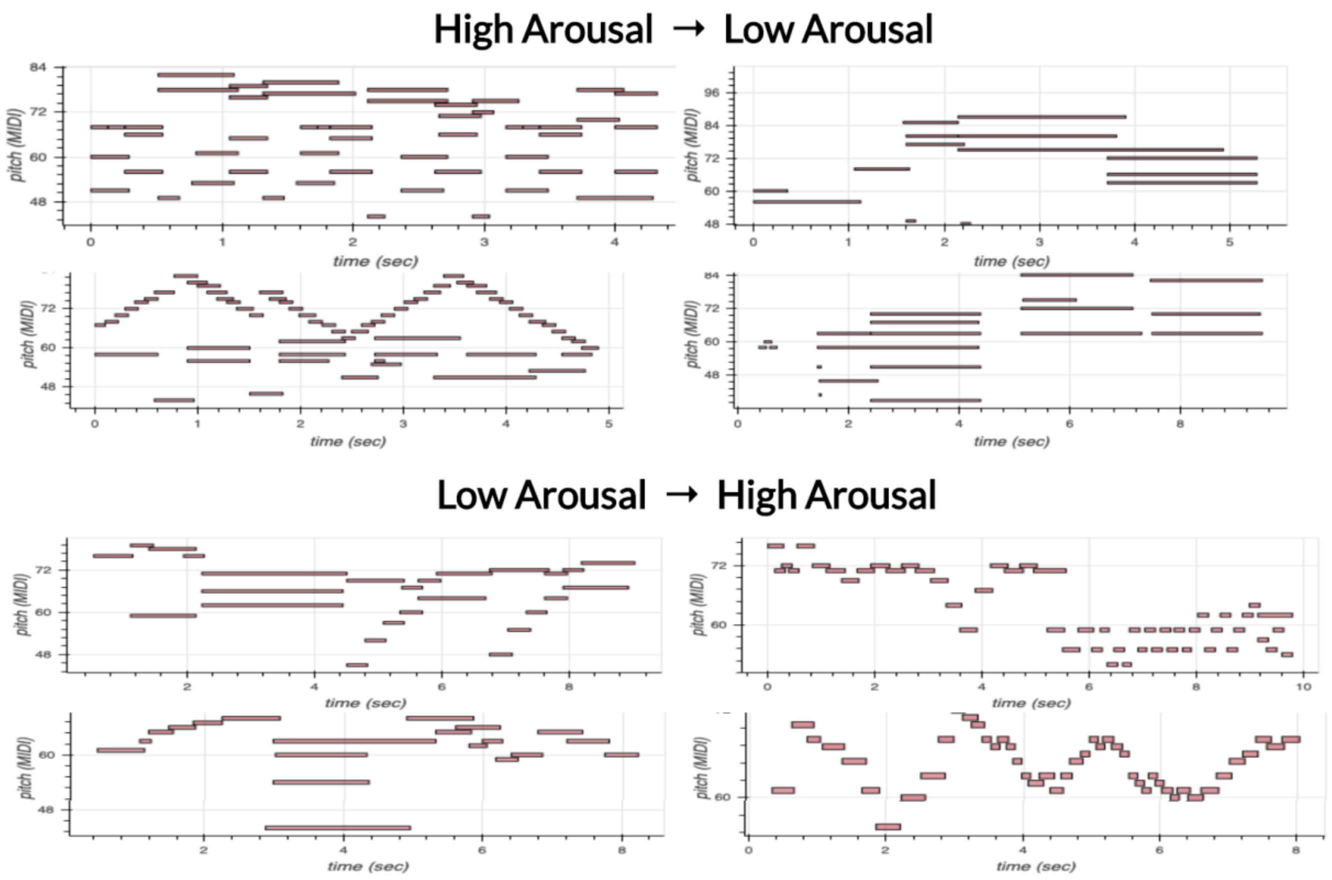
(Source: mrAnmol)

Valence-Arousal Model for Emotion



(Source: mrAnmol)

Music FaderNet (Tan & Herremans, 2020)

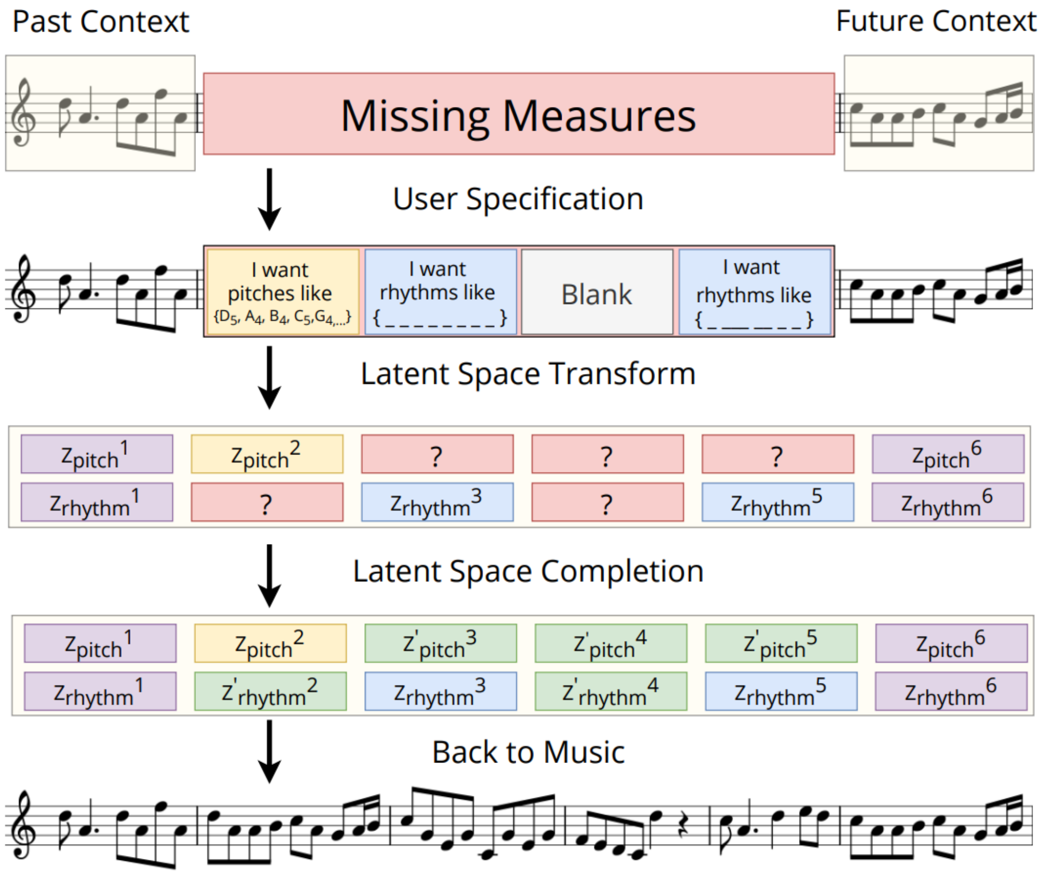


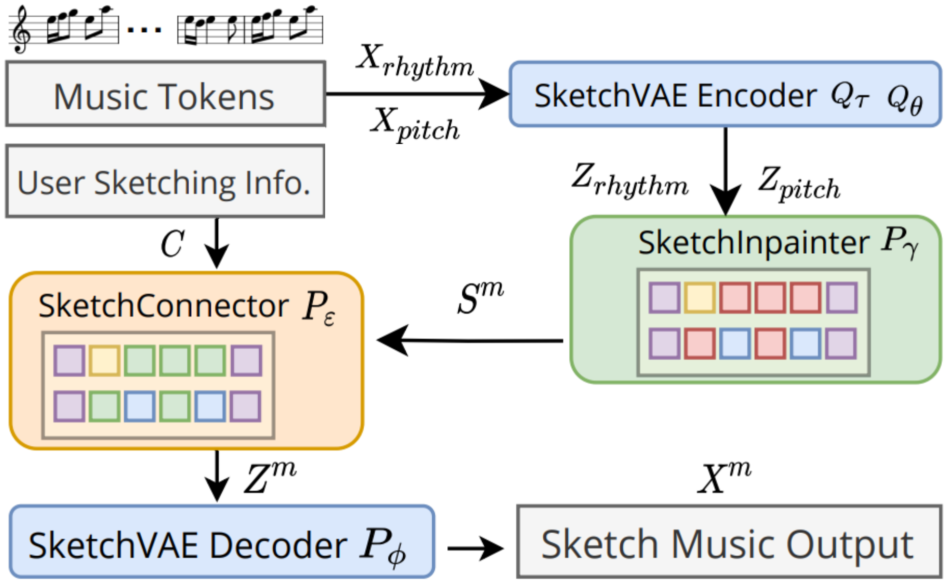
(Source: Tan & Herremeans, 2020)

music-fadernets.github.io

Hao Hao Tan and Dorien Herremans, "Music FaderNets: Controllable Music Generation Based On High-Level Features via Low-Level Feature Modelling," ISMIR, 2020.

Music SketchNet (Chen et al., 2020)

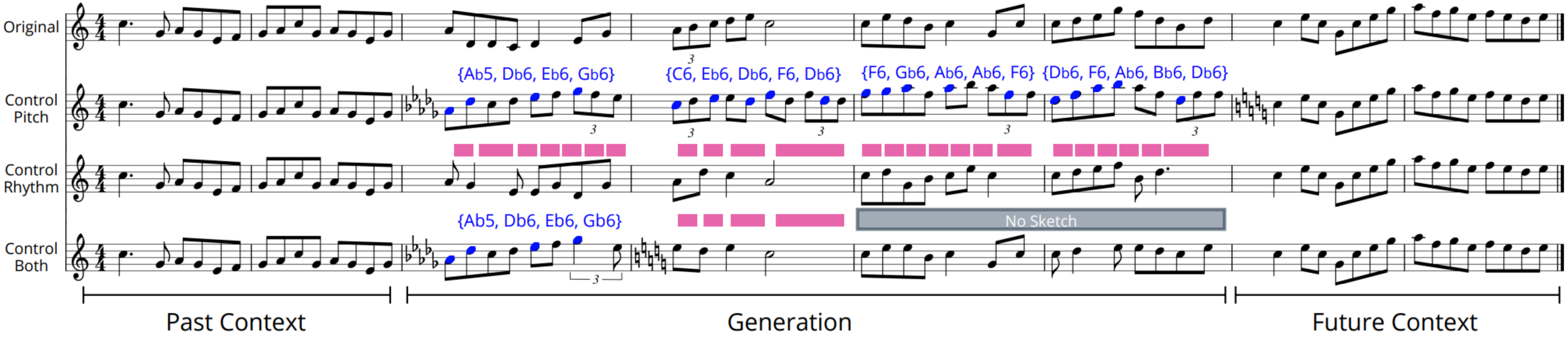




(Source: Chen et al., 2020)

Ke Chen, Cheng-i Wang, Taylor Berg-Kirkpatrick, and Shlomo Dubnov, "Music FaderNets: Controllable Music Generation Based On High-Level Features via Low-Level Feature Modelling," *ISMIR*, 2020.

Music SketchNet (Chen et al., 2020)



(Source: Chen et al., 2020)

Ke Chen, Cheng-i Wang, Taylor Berg-Kirkpatrick, and Shlomo Dubnov, "Music FaderNets: Controllable Music Generation Based On High-Level Features via Low-Level Feature Modelling," *ISMIR*, 2020.

Piano Genie

Piano Genie (Donahue et al., 2019)

youtu.be/YRb0XAnUpIk & magenta.tensorflow.org/pianogenie

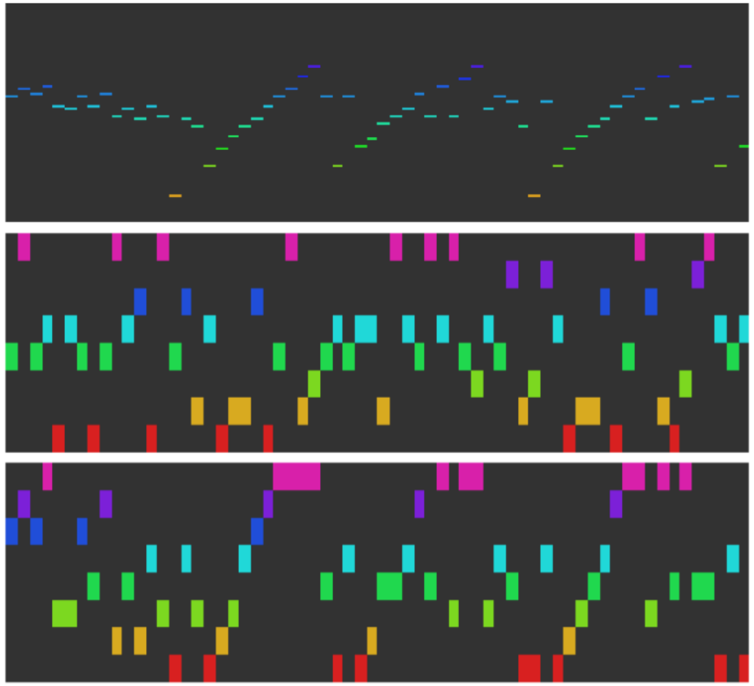
Chris Donahue, Ian Simon, and Sander Dieleman, "Piano Genie," *IUI*, 2019.

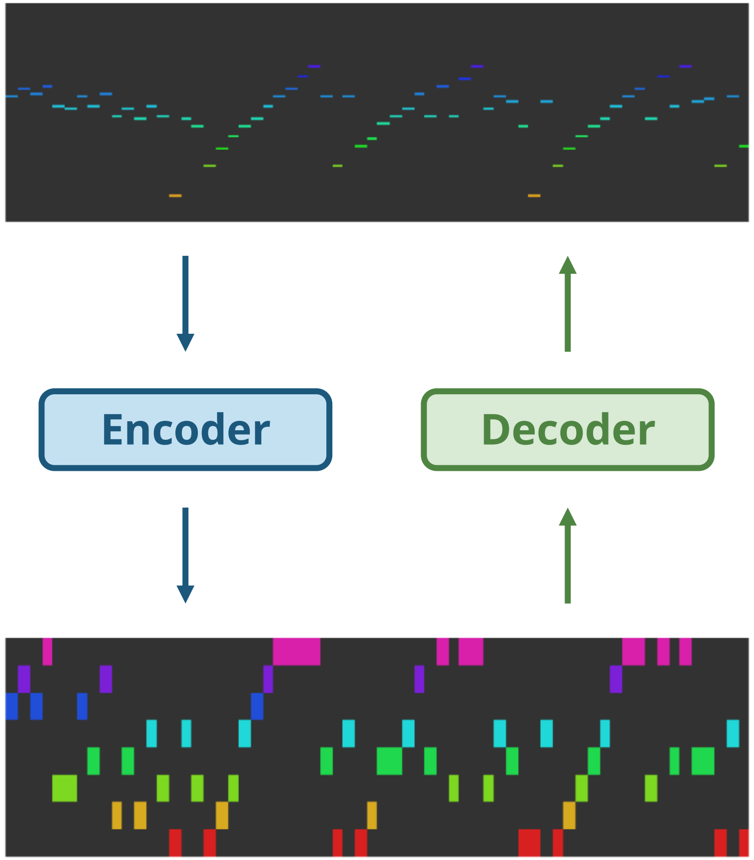
Piano Genie (Donahue et al., 2019)

Input melody

Baseline

Proposed





(Source: Donahue et al., 2019)

Chris Donahue, Ian Simon, and Sander Dieleman, "Piano Genie," *IUI*, 2019.

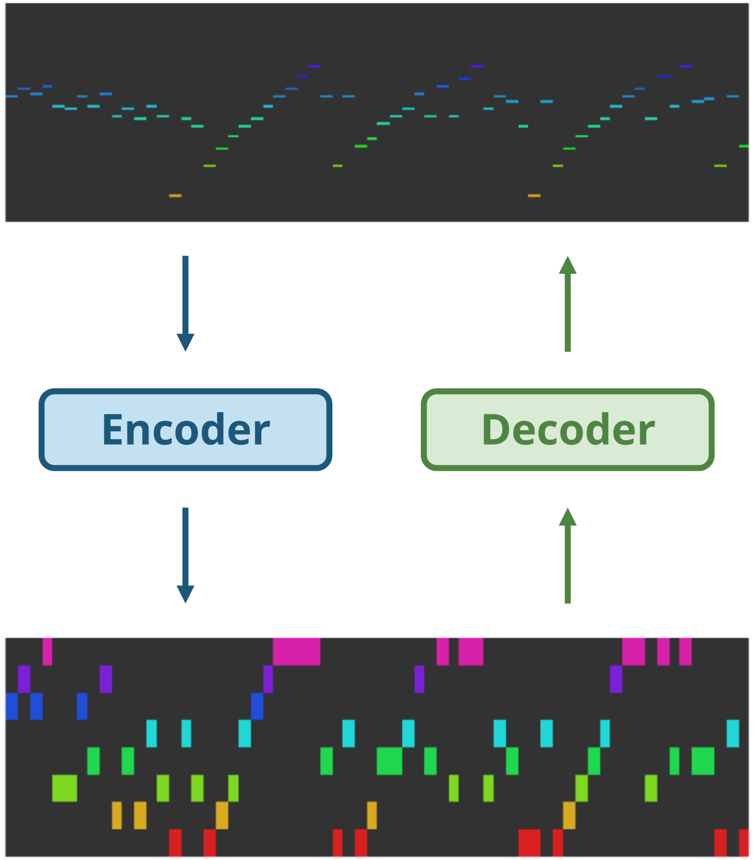
Piano Genie (Donahue et al., 2019)

$$L = L_{\text{recons}} + L_{\text{margin}} + L_{\text{contour}}$$

$$L_{\text{recons}} = -\Sigma \log P_{\text{dec}}(\boldsymbol{x}|\text{enc}(\boldsymbol{x}))$$

$$L_{\text{margin}} = \Sigma \max(|\text{enc}_s(\boldsymbol{x})| - 1, 0)^2$$

$$L_{\text{contour}} = \Sigma \max(1 - \Delta\boldsymbol{x}\Delta\text{enc}_s(\boldsymbol{x}), 0)^2$$



(Source: Donahue et al., 2019)

Chris Donahue, Ian Simon, and Sander Dieleman, "Piano Genie," *IUI*, 2019.

Fruit Genie (2019)

youtu.be/HoVs4kC68no

magenta.tensorflow.org/fruitgenie

Fruit Genie Live (2019)

youtu.be/L4wvXrPmIkU

magenta.tensorflow.org/fruitgenie

Variational Autoencoders for Audio

Four Paradigms of Music Generation

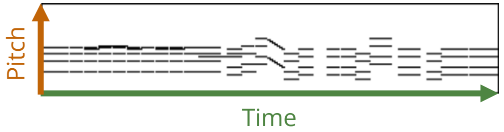
Symbolic music generation

Text-based

```
Program_change_0,
Note_on_60, Time_shift_2, Note_off_60,
Note_on_60, Time_shift_2, Note_off_60,
Note_on_76, Time_shift_2, Note_off_67,
Note_on_67, Time_shift_2, Note_off_67,
...
```

MIDI

Image-based



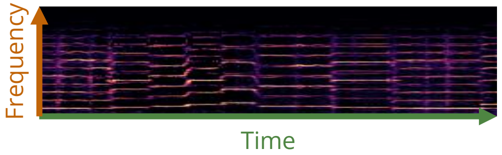
Piano roll

Audio-domain music generation

Time series-based

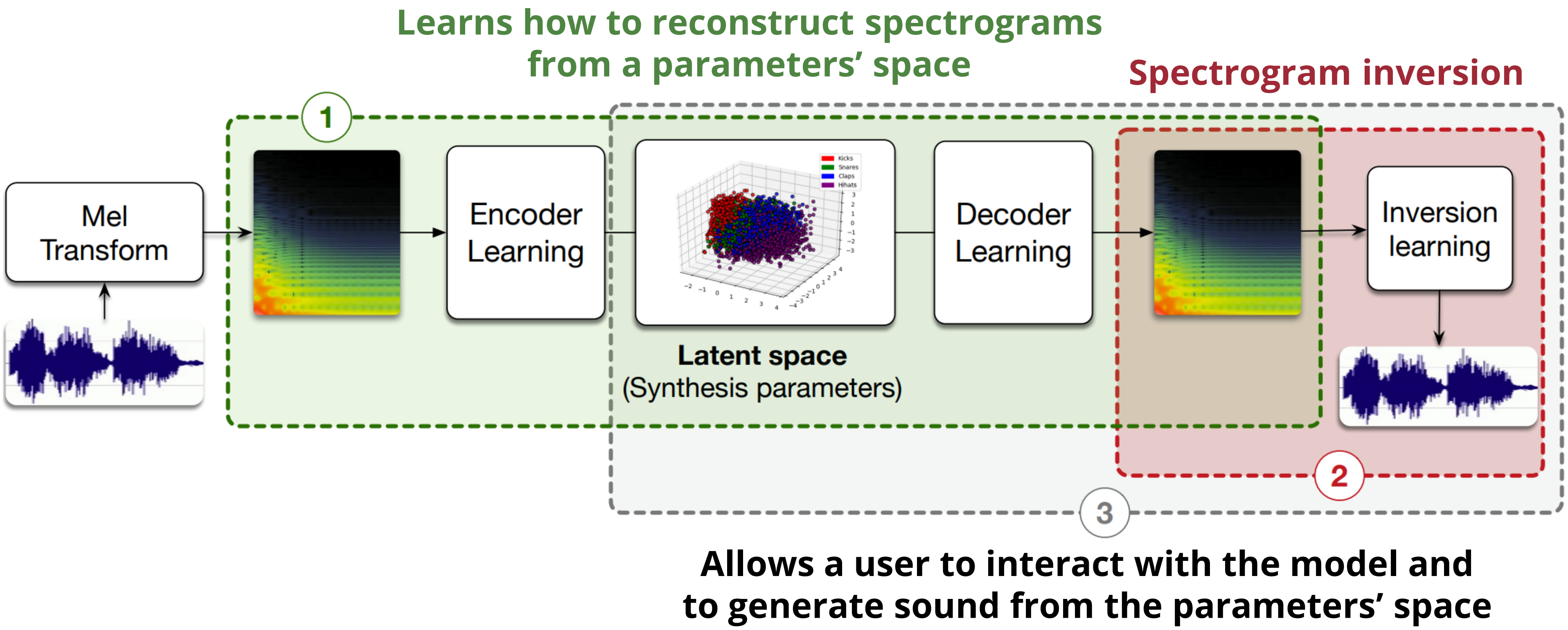
Waveform

Image-based



Spectrogram

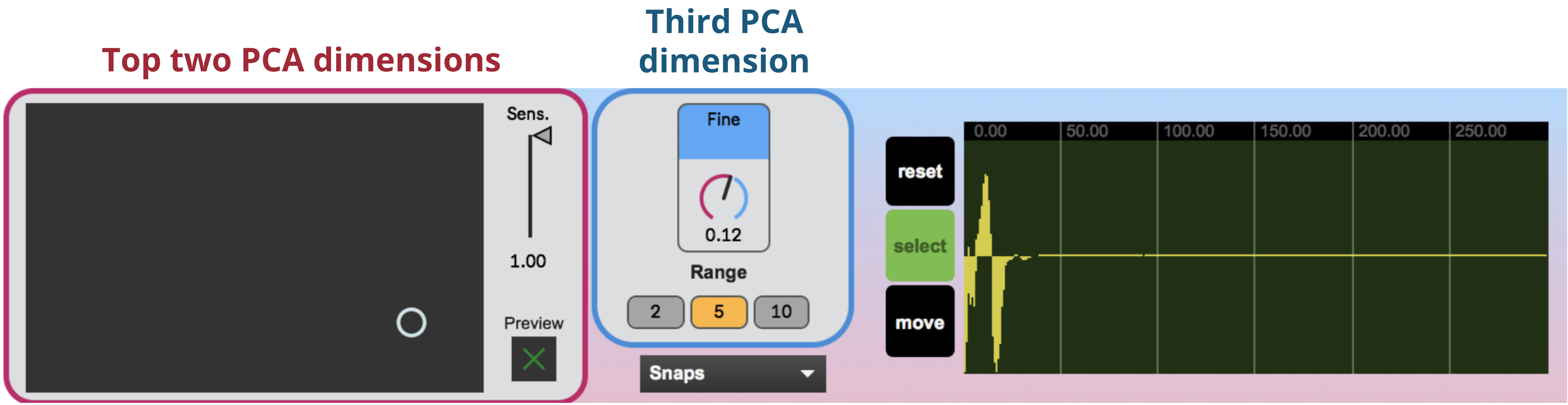
Neural Drum Machine (Aouameur et al., 2019)



(Source: Aouameur et al., 2019)

Cyran Aouameur, Philippe Esling, and Gaëtan Hadjeres, "Neural Drum Machine : An Interactive System for Real-time Synthesis of Drum Sounds," *ICCC*, 2019.

Neural Drum Machine (Aouameur et al., 2019)

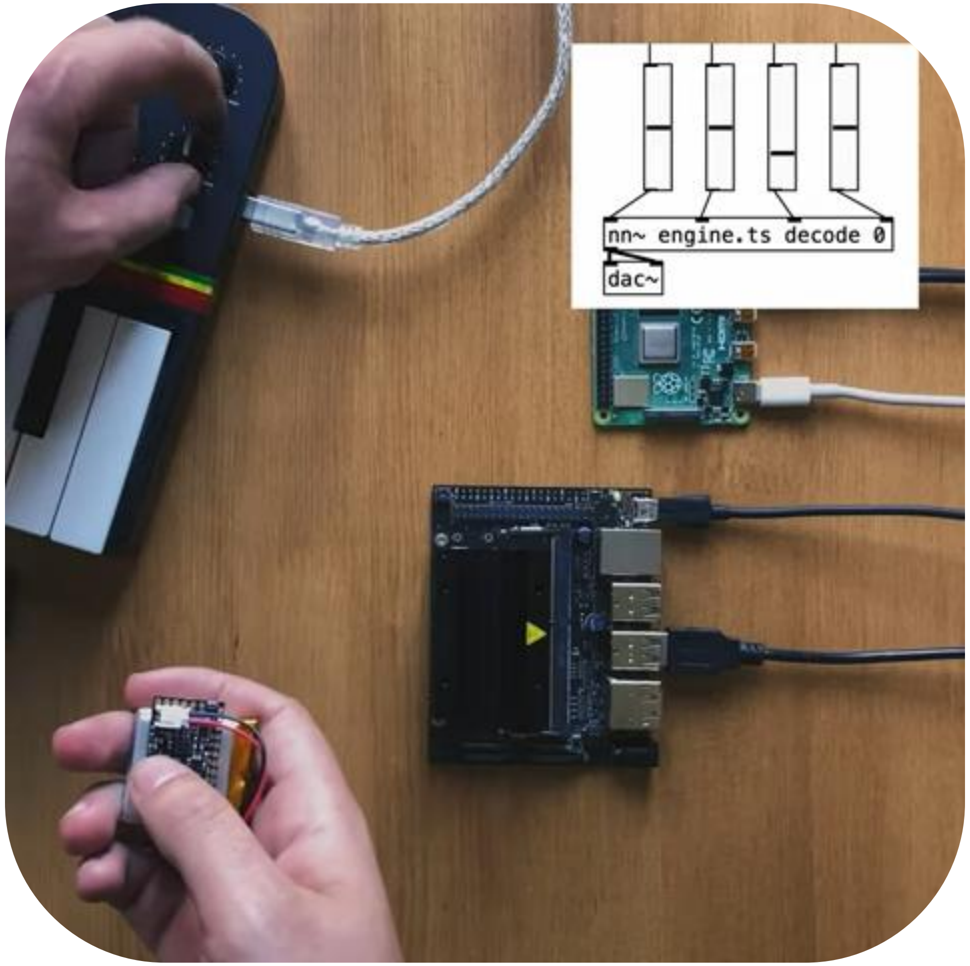


(Source: Aouameur et al., 2019)

drive.google.com/file/d/1DDo0_KnwkWirCM4t0PT8cp6uotsfuufj/view

Cyran Aouameur, Philippe Esling, and Gaëtan Hadjeres, "Neural Drum Machine : An Interactive System for Real-time Synthesis of Drum Sounds," *ICCC*, 2019.

RAVE: Real-time Audio Synthesis (Caillon & Esling, 2022)

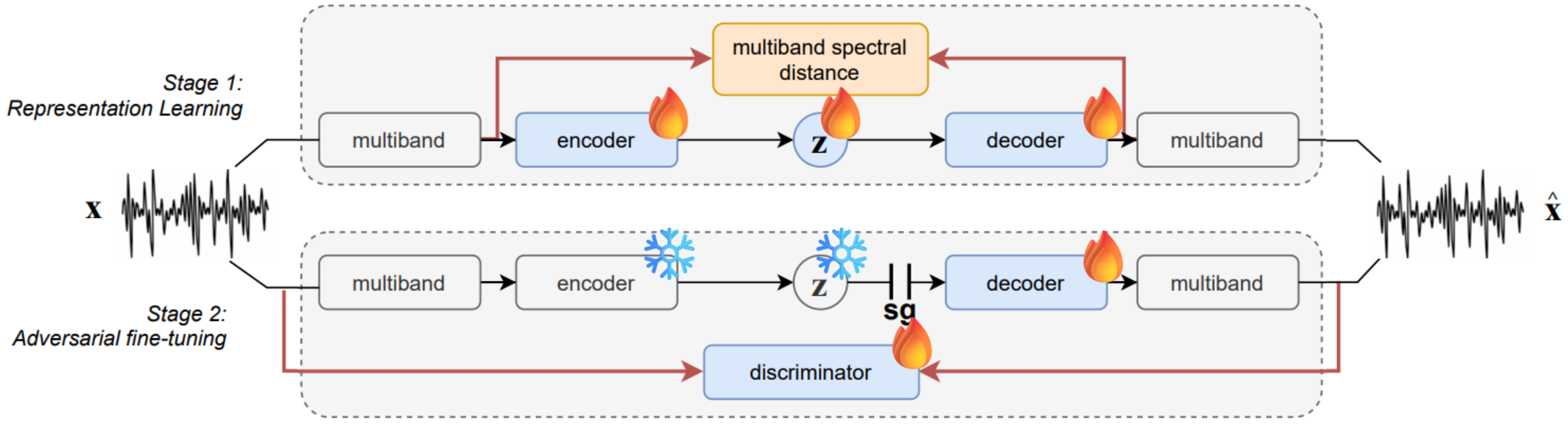


youtu.be/jAIRf4nGgYI

Antoine Caillon and Philippe Esling, "RAVE: A variational autoencoder for fast and high-quality neural audio synthesis," *arXiv preprint arXiv:2111.05011*, 2021.

RAVE: Real-time Audio Synthesis (Caillon & Esling, 2022)

16-band decomposition, 48kHz



(Source: Caillon & Esling, 2021)

github.com/acids-ircam/RAVE

Antoine Caillon and Philippe Esling, "RAVE: A variational autoencoder for fast and high-quality neural audio synthesis," *arXiv preprint arXiv:2111.05011*, 2021.

RAVE: Real-time Audio Synthesis (Caillon & Esling, 2022)

Model	CPU synthesis	GPU synthesis
NSynth	18 Hz	57 Hz
SING	304 kHz	9.8 MHz
RAVE (Ours) w/o multiband	38 kHz	3.7 MHz
RAVE (Ours)	**985 kHz**	**11.7 MHz**

Realtime capable on CPUs & GPUs

(Source: Caillon & Esling, 2021)

anonymous84654.github.io/RAVE_anonymous

Antoine Caillon and Philippe Esling, "RAVE: A variational autoencoder for fast and high-quality neural audio synthesis," *arXiv preprint arXiv:2111.05011*, 2021.

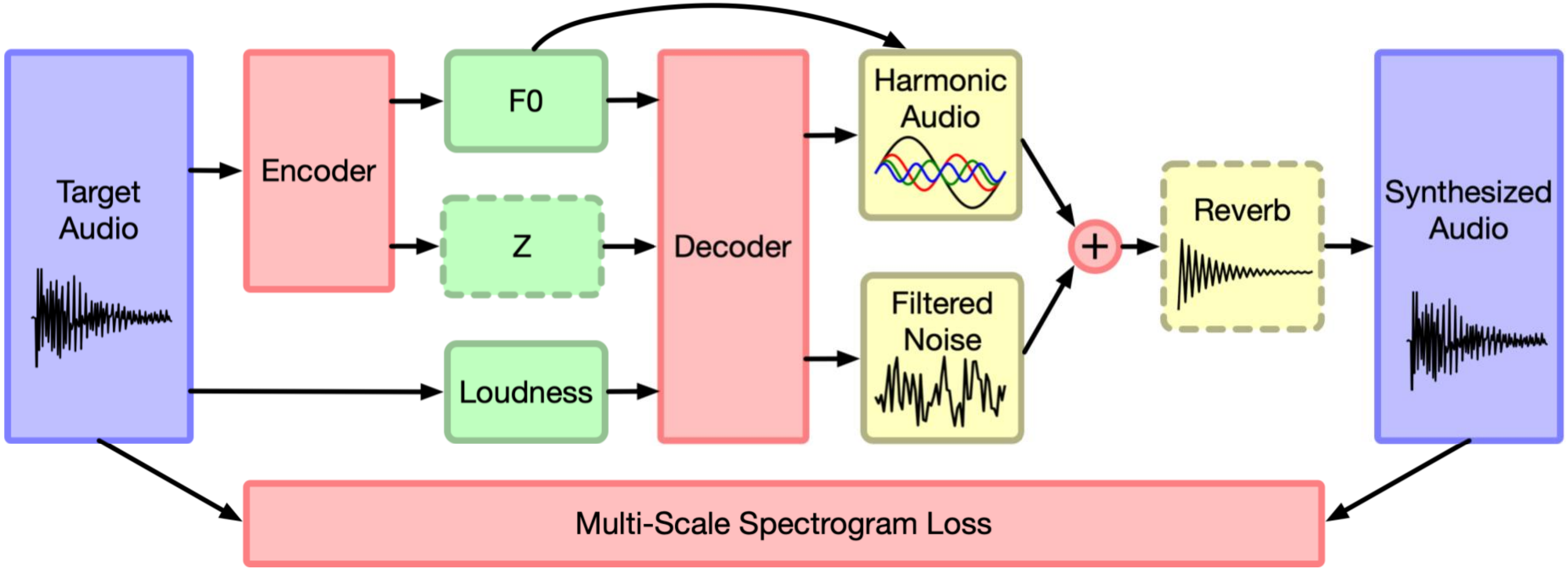
RAVE: Real-time Audio Synthesis (Caillon & Esling, 2022)

youtu.be/dMZs04TzxUI

Antoine Caillon and Philippe Esling, "RAVE: A variational autoencoder for fast and high-quality neural audio synthesis," *arXiv preprint arXiv:2111.05011*, 2021.

Differentiable DSP

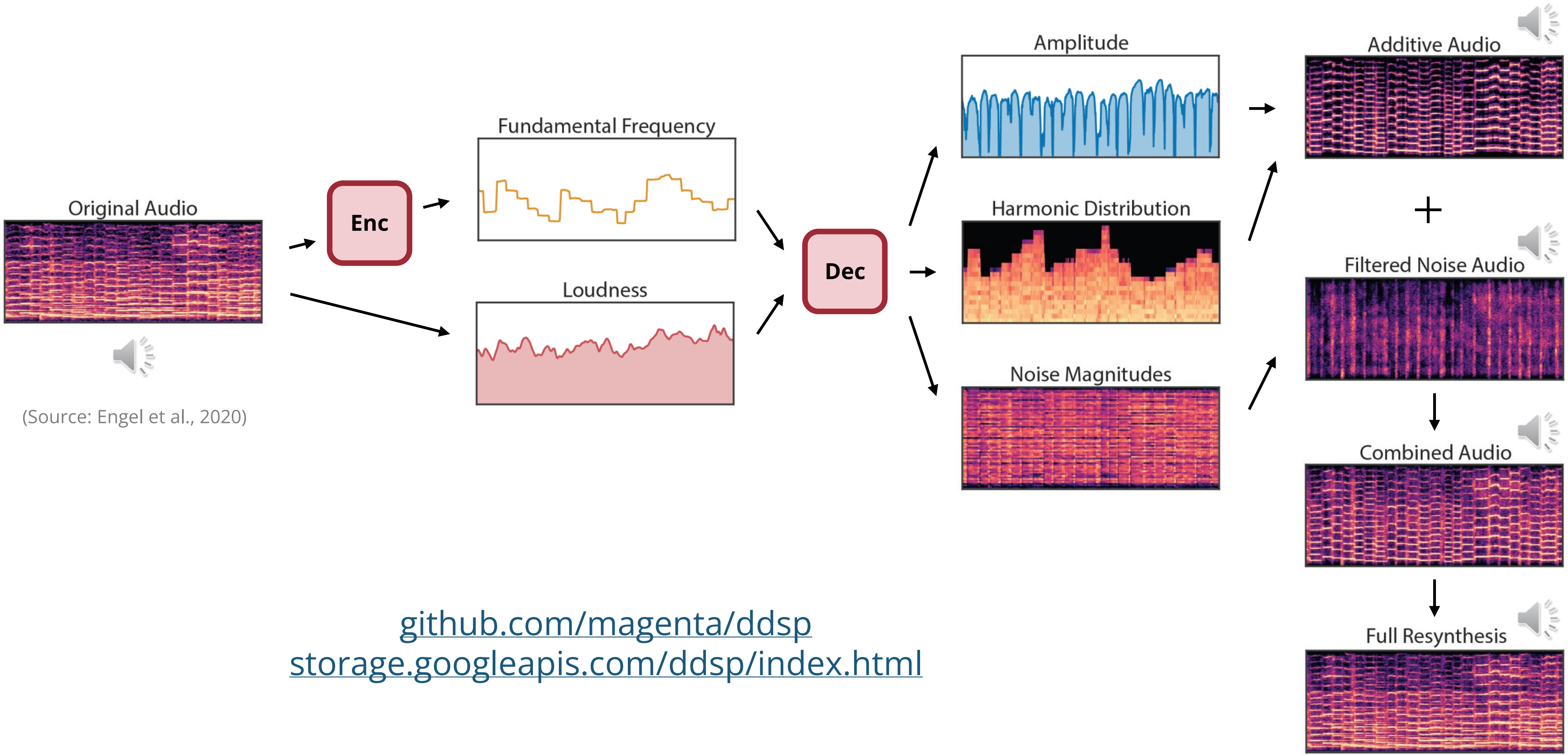
Differentiable DSP (DDSP) (Engel et al., 2020)



(Source: Engel et al., 2020)

Jesse Engel, Lamtharn (Hanoi) Hantrakul, Chenjie Gu, Adam Roberts, "DDSP: Differentiable Digital Signal Processing," *ICLR*, 2020.

Differentiable DSP (DDSP) (Engel et al., 2020)



(Source: Engel et al., 2020)

github.com/magenta/ddsp
storage.googleapis.com/ddsp/index.html

Jesse Engel, Lamtharn (Hanoi) Hantrakul, Chenjie Gu, Adam Roberts, "DDSP: Differentiable Digital Signal Processing," *ICLR*, 2020.

Entering Demons & Gods by Yaboi Hanoi (2022)

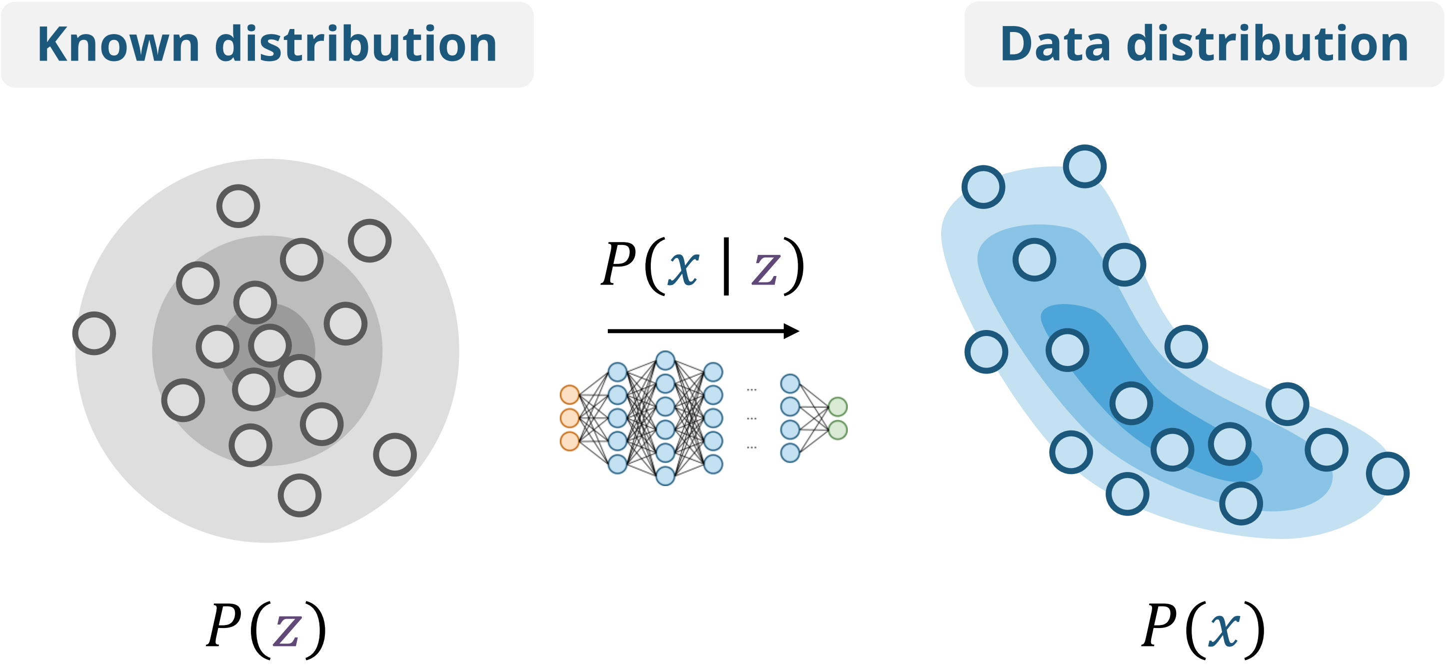
youtu.be/PbrRoR3nEVw

soundcloud.com/yaboihanoi/enter-demons-and-gods

Recap

Deep Latent Variable Models

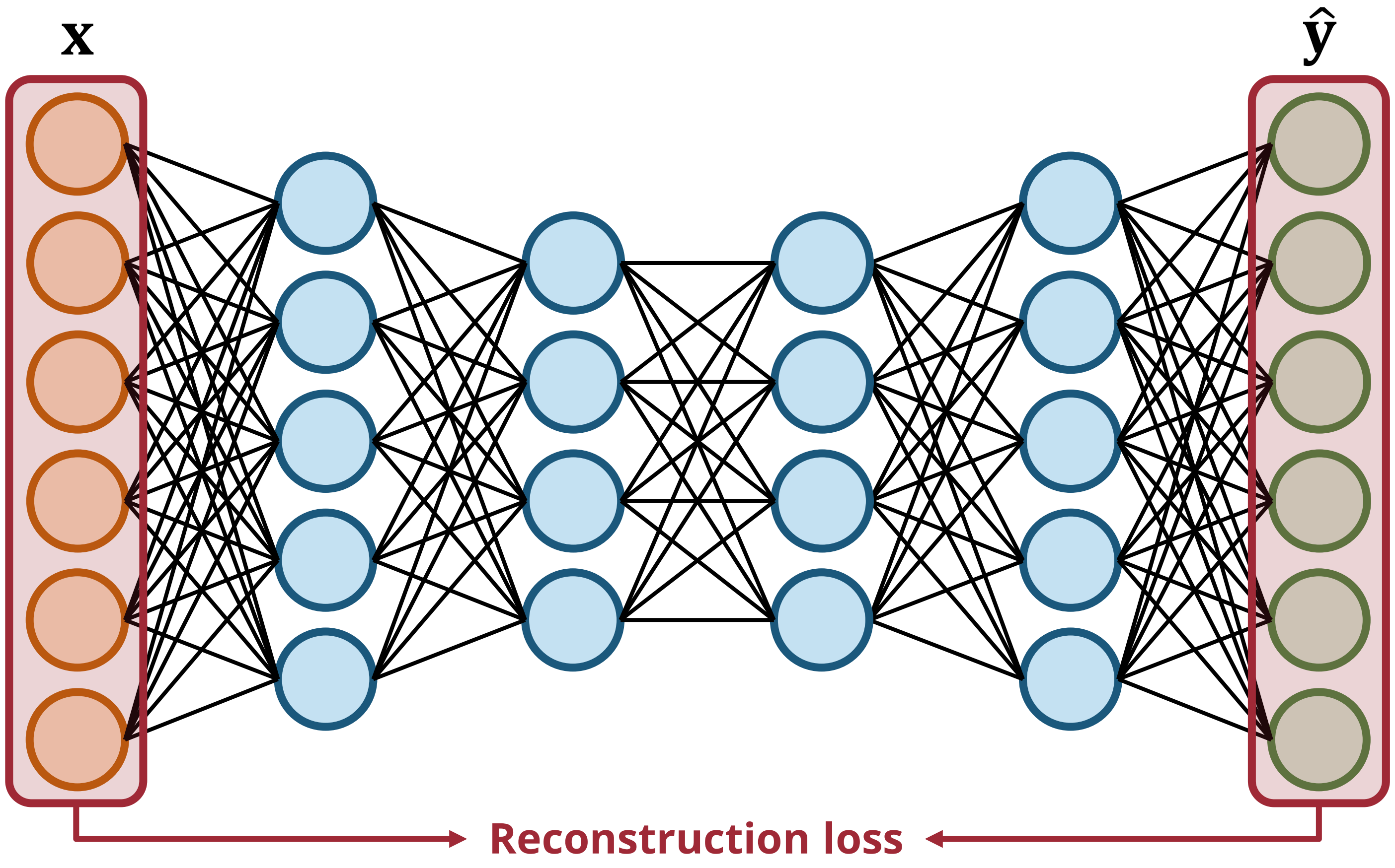
- **Intuition**: Learn to map a known distribution to the data distribution



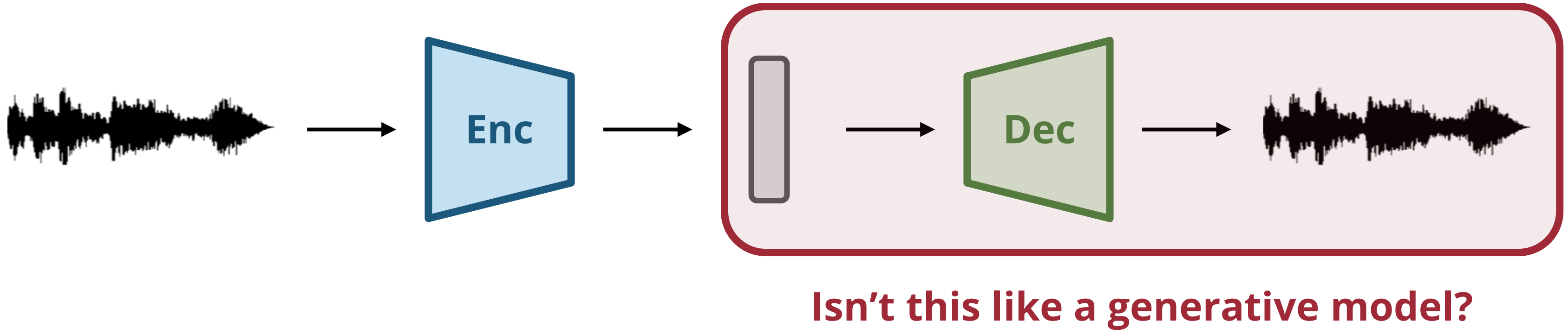
$$P(x) = P(z)\, P(x \mid z)$$

Autoencoders

- A neural network where the **input and output are the same**

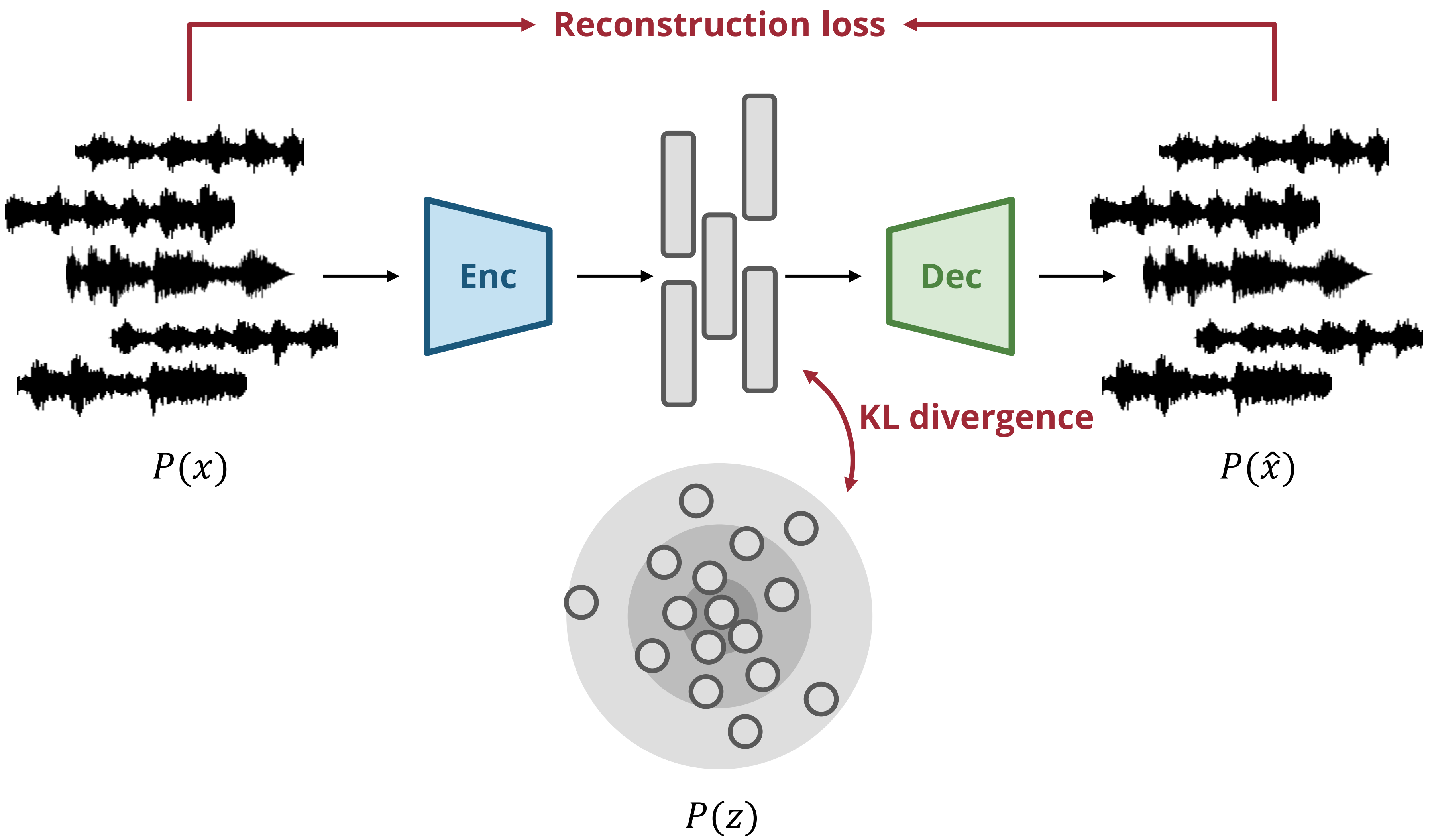


Autoencoder



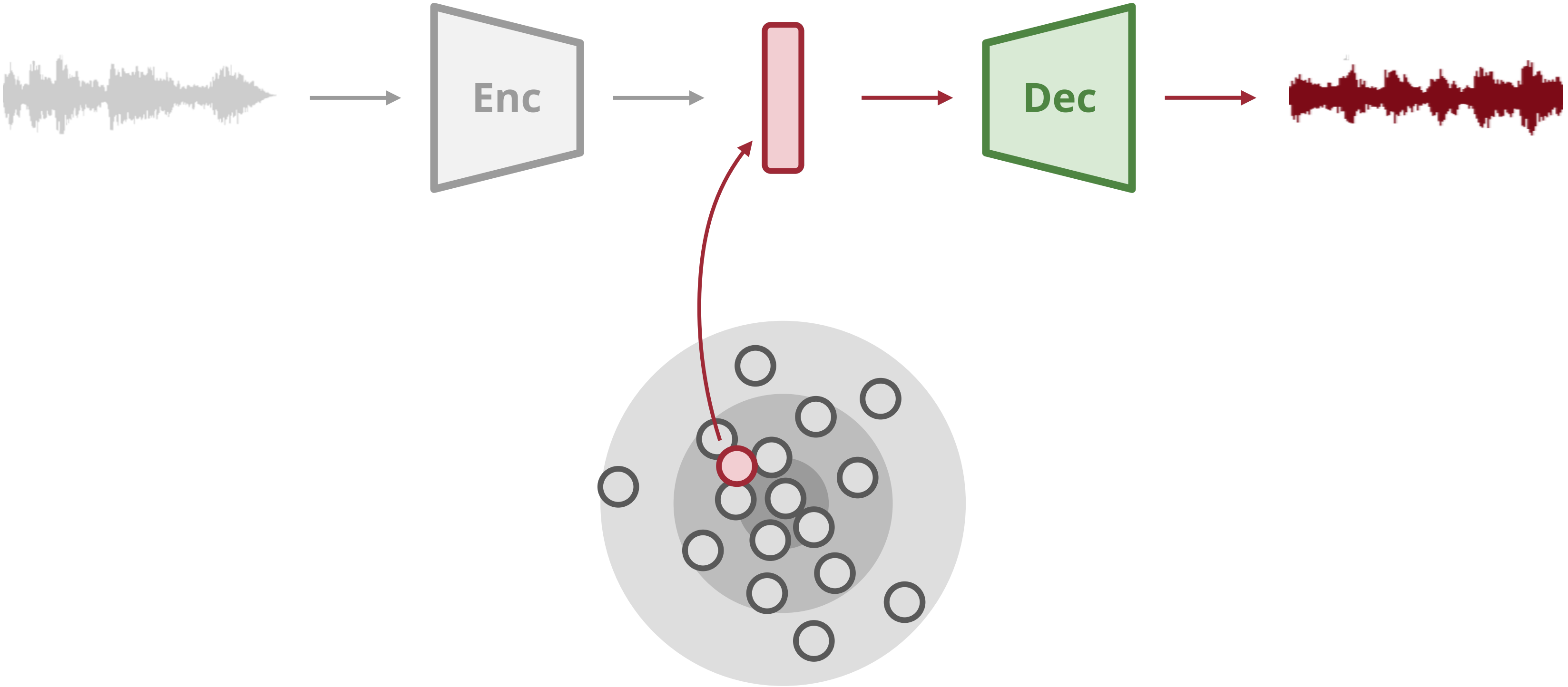
What exactly is a generative model?

Variational Autoencoder (VAE): Training



Diederik P Kingma and Max Welling, "Auto-Encoding Variational Bayes," *ICLR*, 2014.

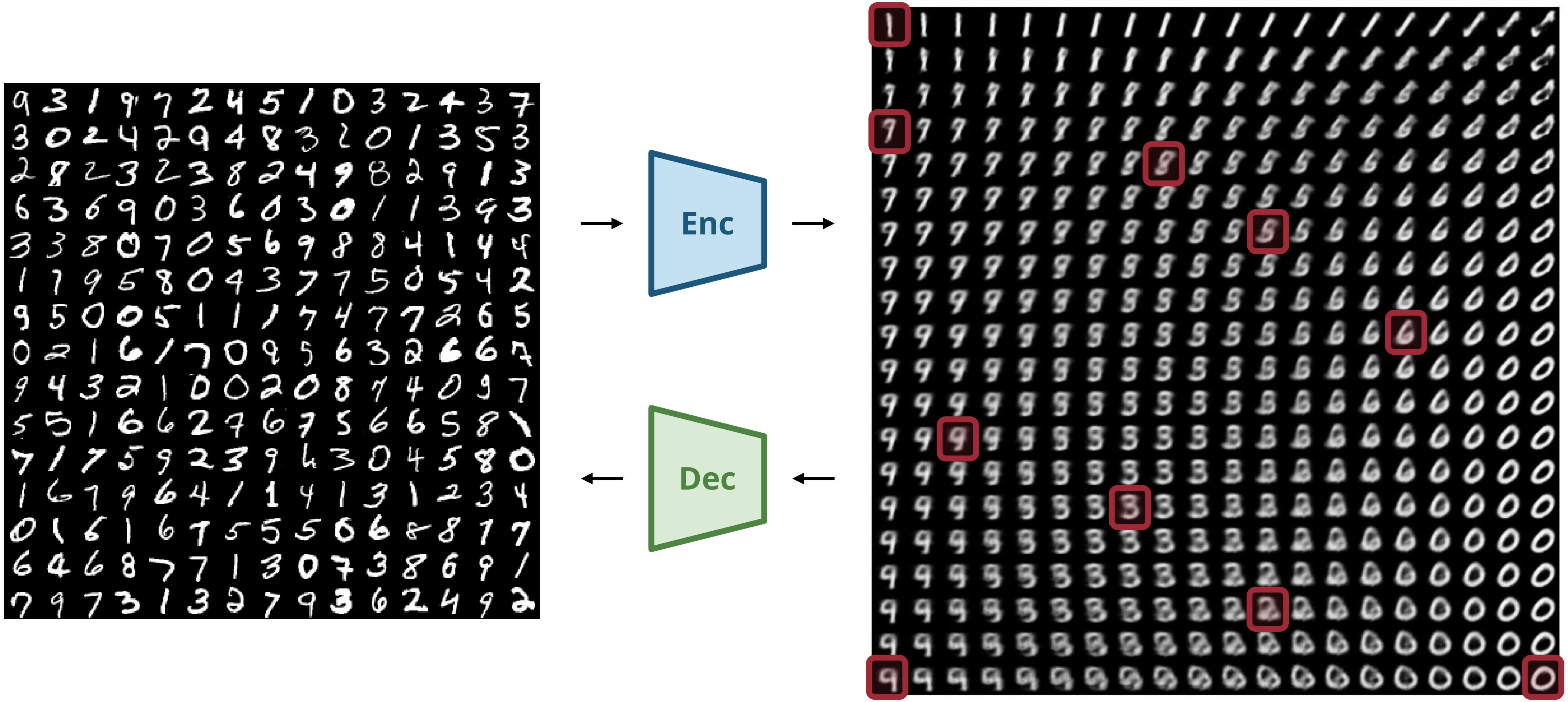
Variational Autoencoder (VAE): Generation



$P(z)$

Diederik P Kingma and Max Welling, "Auto-Encoding Variational Bayes," *ICLR*, 2014.

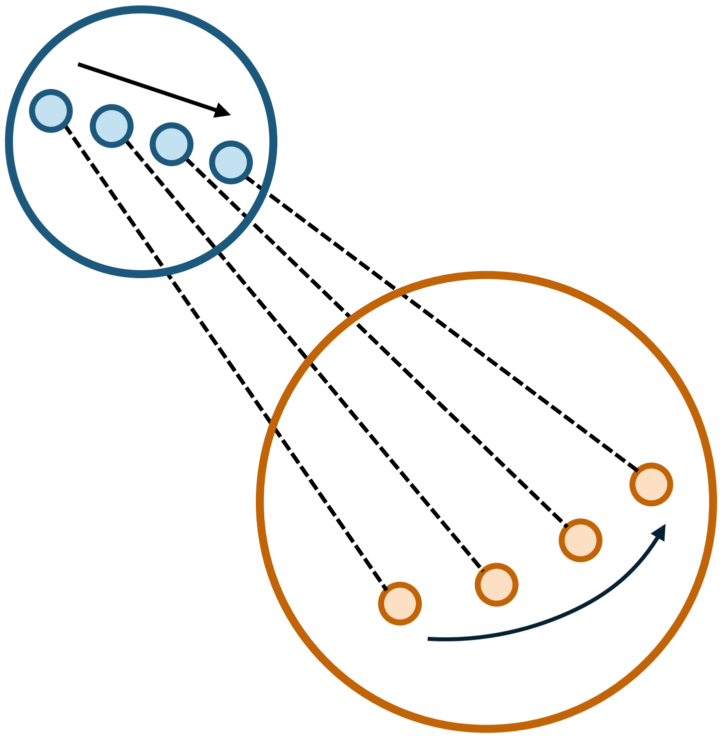
What does a VAE learn?

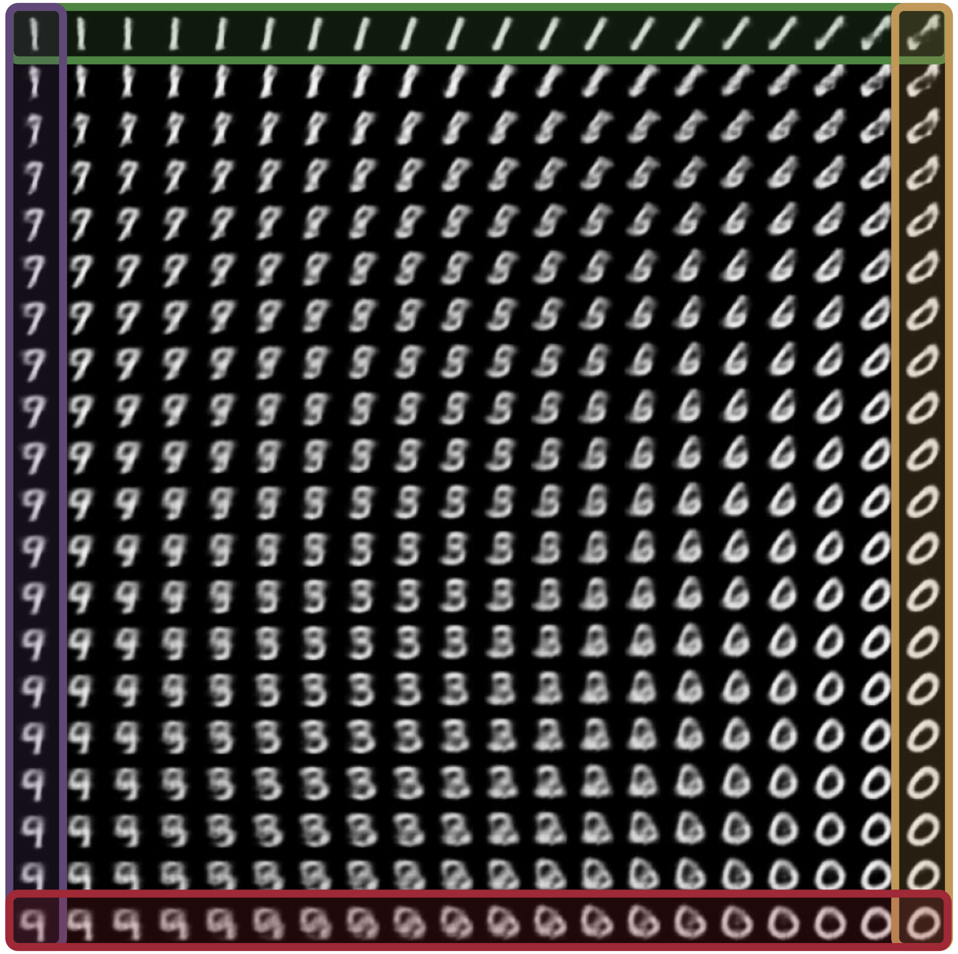


(Source: tensorflow.org)

tensorflow.org/tutorials/generative/cvae

Latent Space Interpolation of a VAE

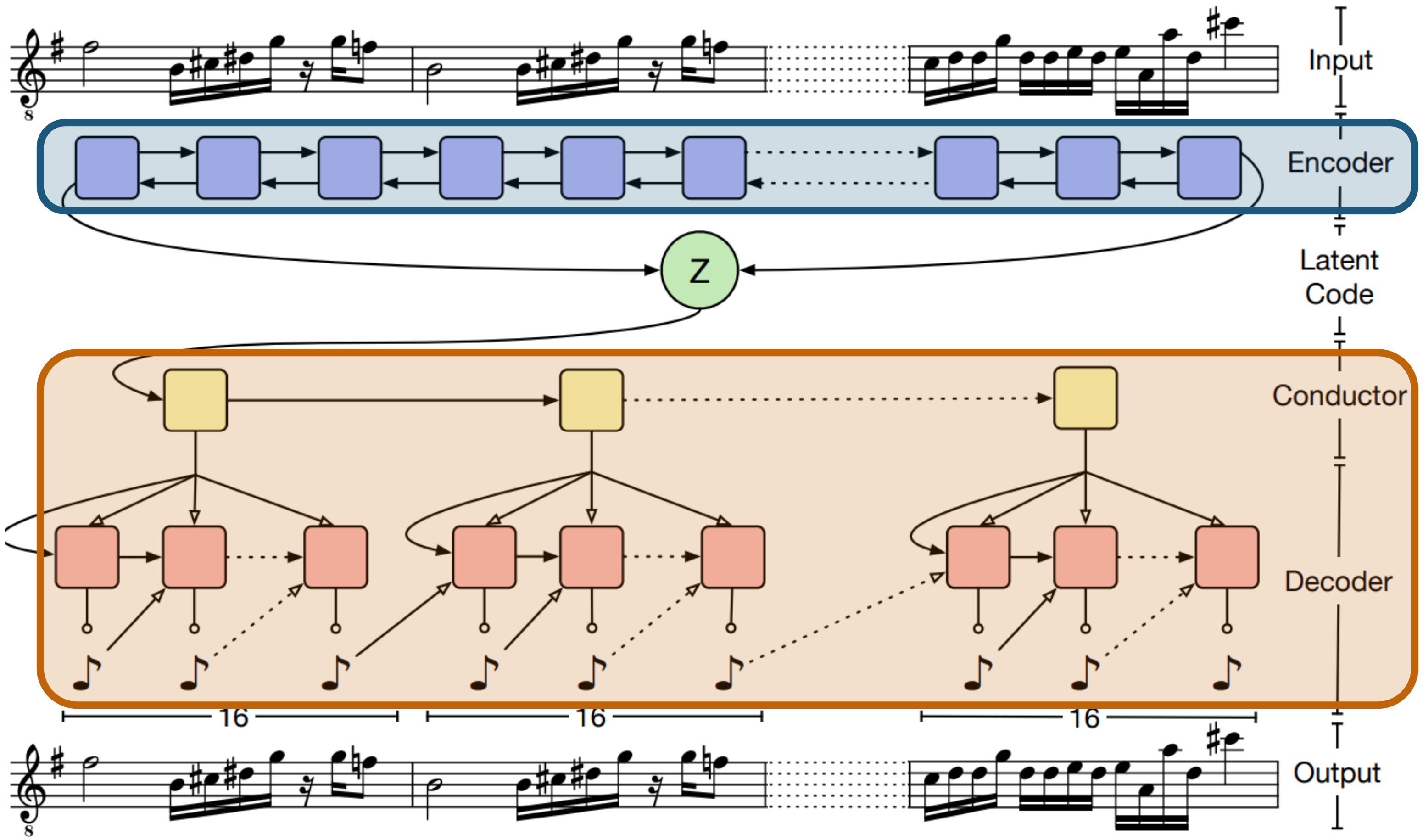




(Source: tensorflow.org)

tensorflow.org/tutorials/generative/cvae

MusicVAE: A VAE for Symbolic Music (Roberts et al., 2018)



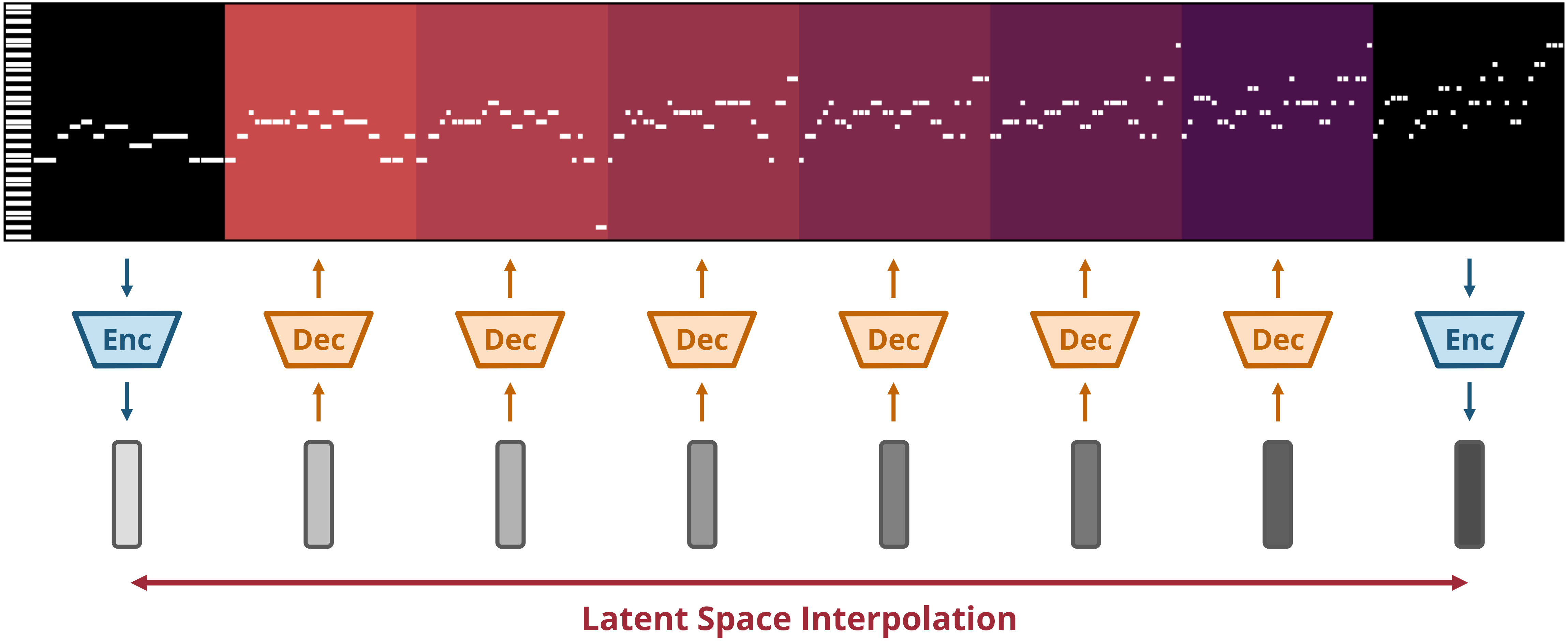
Bidirectional encoder

Hierarchical decoder

(Source: Roberts et al., 2018)

Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck, "A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music," *ICML*, 2018.

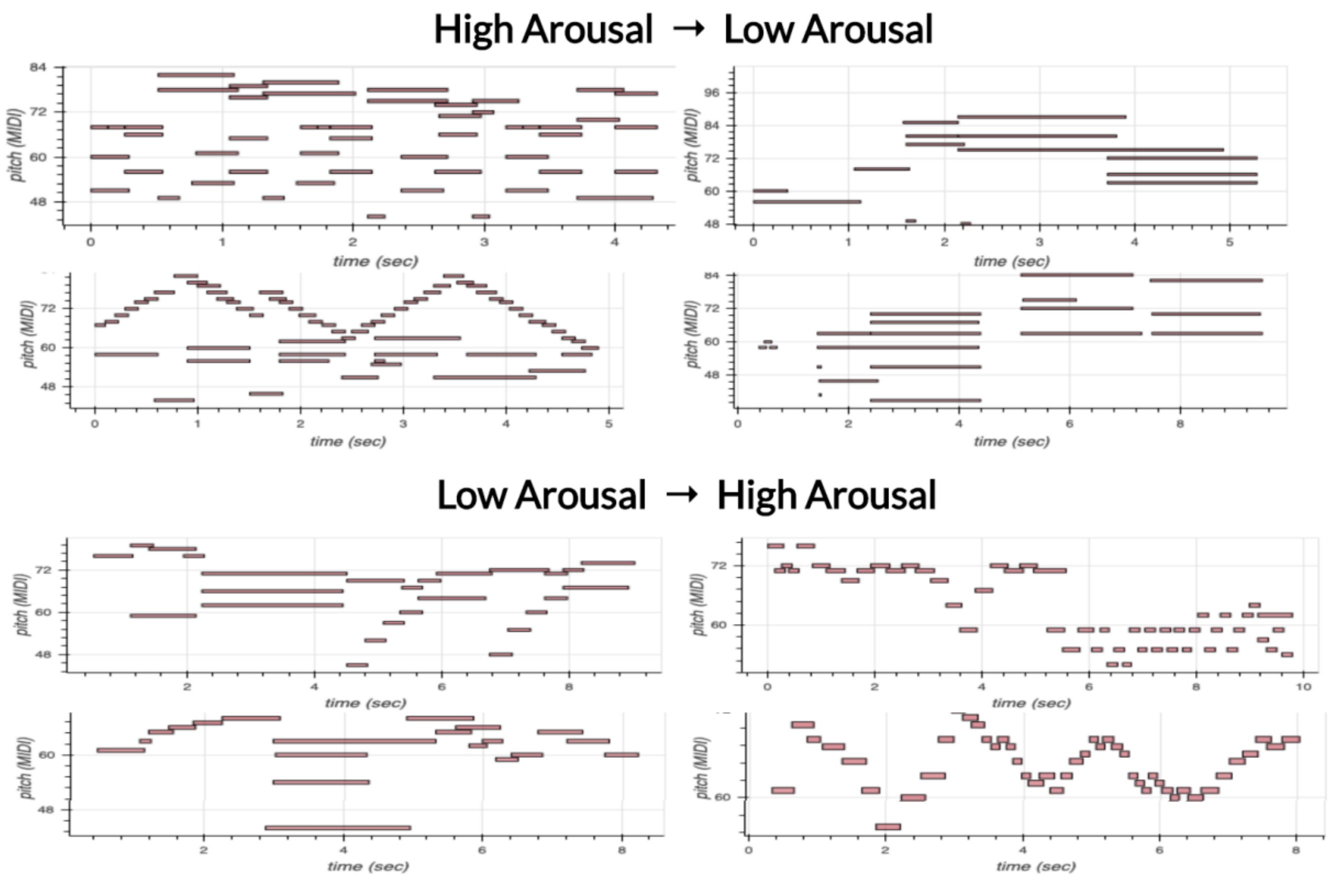
Latent Space Interpolation for MusicVAE (Roberts et al., 2018)



(Source: Roberts et al., 2018)

Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck, "A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music," *ICML*, 2018.

Music FaderNet (Tan & Herremans, 2020)

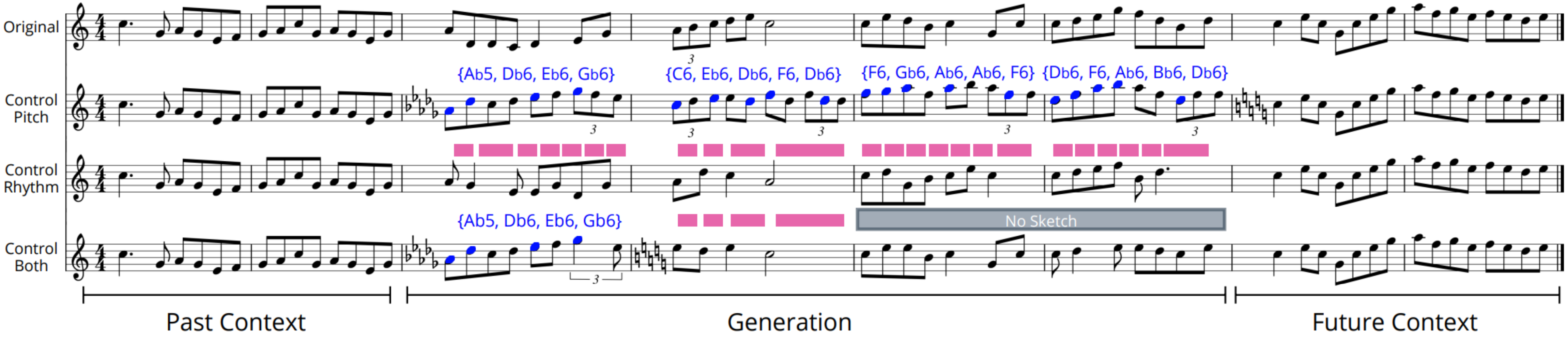


(Source: Tan & Herremeans, 2020)

music-fadernets.github.io

Hao Hao Tan and Dorien Herremans, "Music FaderNets: Controllable Music Generation Based On High-Level Features via Low-Level Feature Modelling," ISMIR, 2020.

Music SketchNet (Chen et al., 2020)



(Source: Chen et al., 2020)

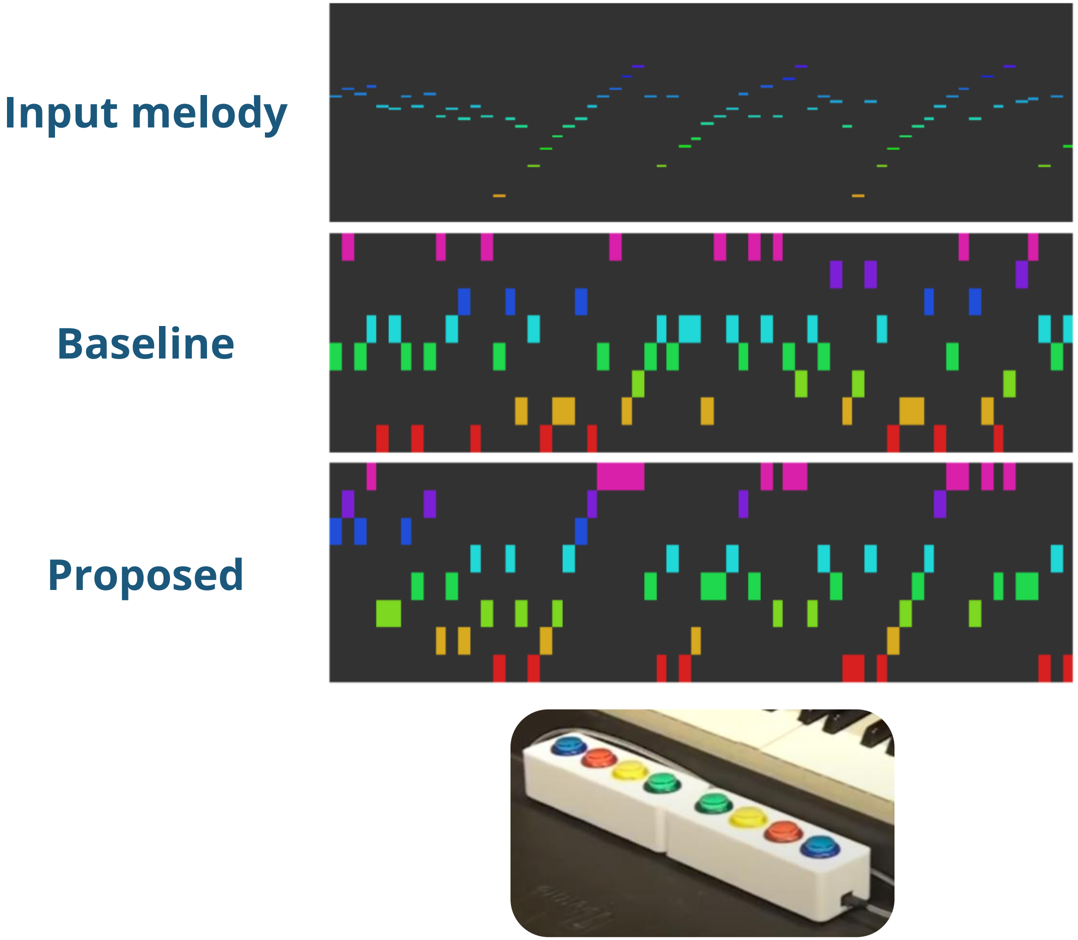
Ke Chen, Cheng-i Wang, Taylor Berg-Kirkpatrick, and Shlomo Dubnov, "Music FaderNets: Controllable Music Generation Based On High-Level Features via Low-Level Feature Modelling," *ISMIR*, 2020.

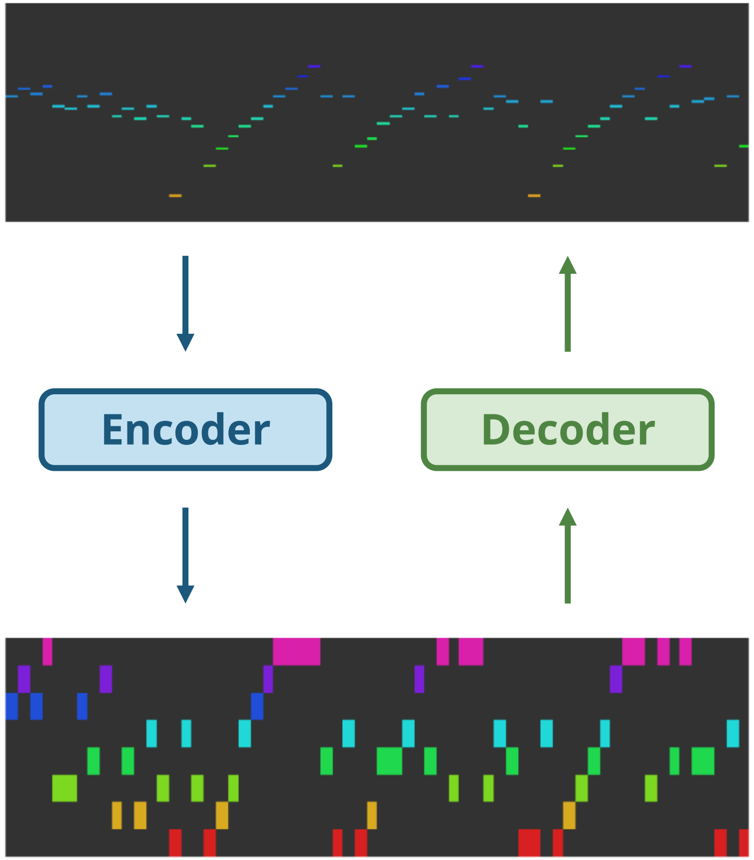
Piano Genie (Donahue et al., 2019)

youtu.be/YRb0XAnUpIk & magenta.tensorflow.org/pianogenie

Chris Donahue, Ian Simon, and Sander Dieleman, "Piano Genie," *IUI*, 2019.

Piano Genie (Donahue et al., 2019)





(Source: Donahue et al., 2019)

Chris Donahue, Ian Simon, and Sander Dieleman, "Piano Genie," *IUI*, 2019.

Neural Drum Machine (Aouameur et al., 2019)

Top two PCA dimensions

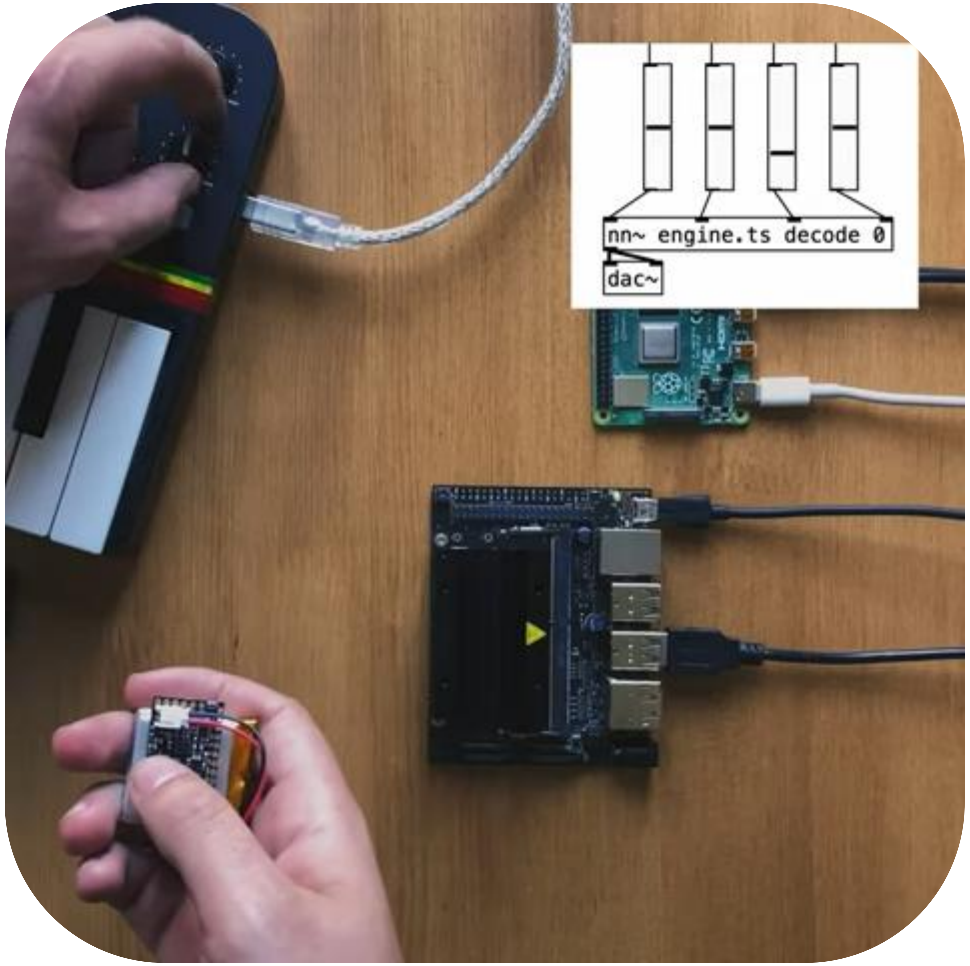
Third PCA dimension

(Source: Aouameur et al., 2019)

drive.google.com/file/d/1DDo0_KnwkWirCM4t0PT8cp6uotsfuufj/view

Cyran Aouameur, Philippe Esling, and Gaëtan Hadjeres, "Neural Drum Machine : An Interactive System for Real-time Synthesis of Drum Sounds," *ICCC*, 2019.

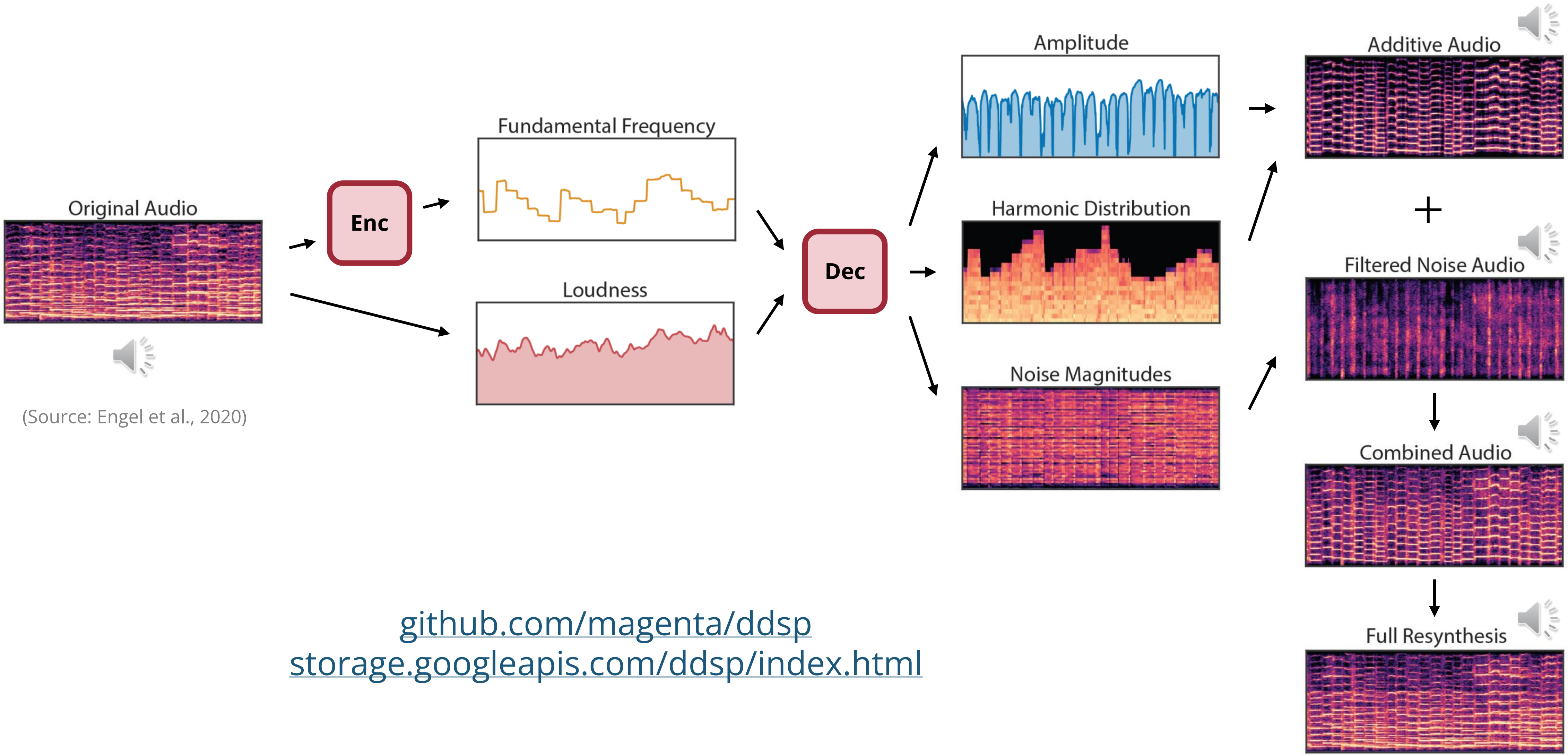
RAVE: Real-time Audio Synthesis (Caillon & Esling, 2022)



youtu.be/jAIRf4nGgYI

Antoine Caillon and Philippe Esling, "RAVE: A variational autoencoder for fast and high-quality neural audio synthesis," *arXiv preprint arXiv:2111.05011*, 2021.

Differentiable DSP (DDSP) (Engel et al., 2020)



(Source: Engel et al., 2020)

github.com/magenta/ddsp
storage.googleapis.com/ddsp/index.html

Jesse Engel, Lamtharn (Hanoi) Hantrakul, Chenjie Gu, Adam Roberts, "DDSP: Differentiable Digital Signal Processing," *ICLR*, 2020.

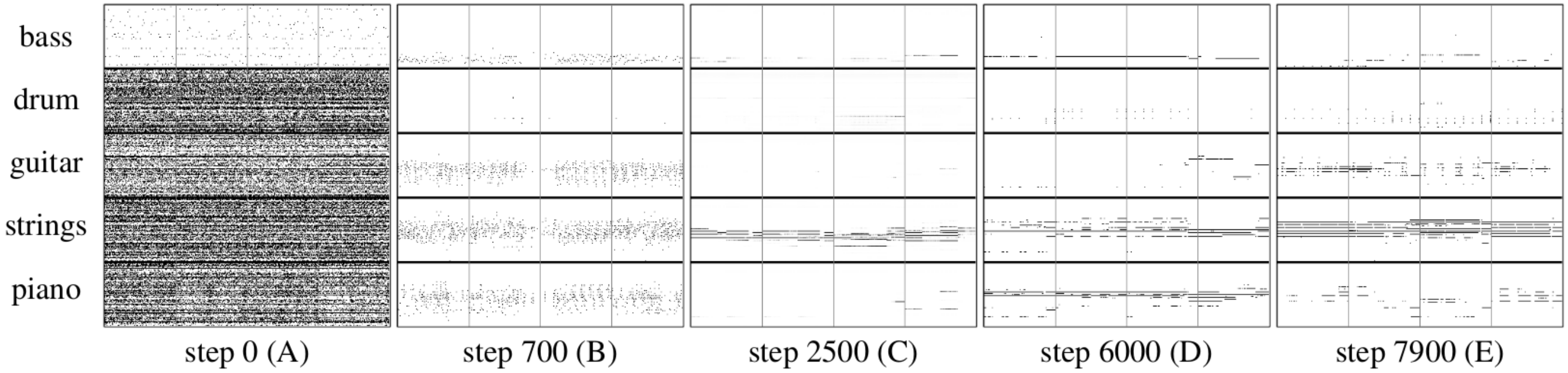
Entering Demons & Gods by Yaboi Hanoi (2022)

youtu.be/PbrRoR3nEVw

soundcloud.com/yaboihanoi/enter-demons-and-gods

Next Lecture

Generative Adversarial Nets



(Source: Dong et al., 2018)