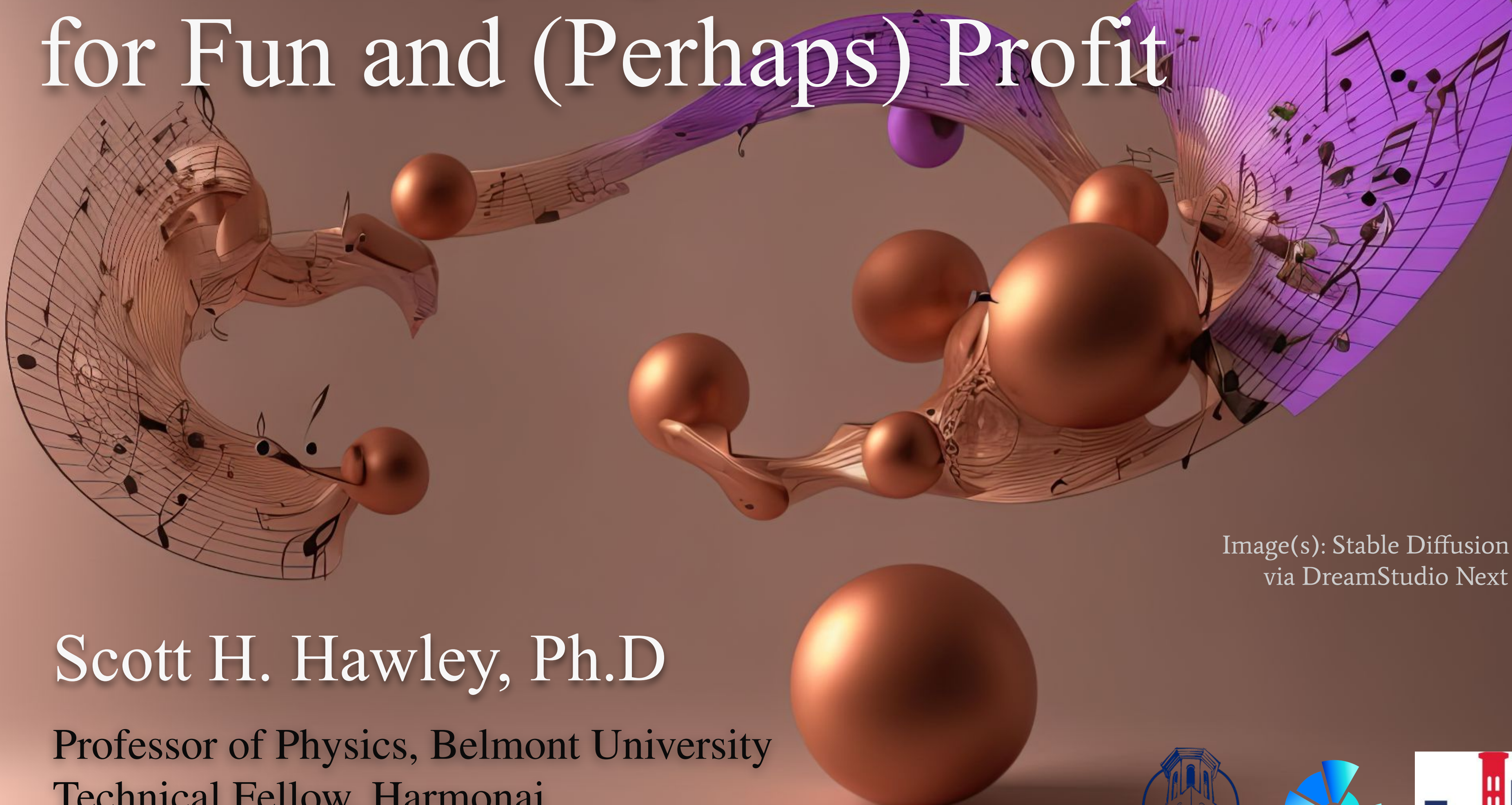
Mathemangling Audio Embeddings for Fun and (Perhaps) Profit



Image(s): Stable Diffusion XL via DreamStudio Next

Scott H. Hawley, Ph.D

Professor of Physics, Belmont University
Technical Fellow, Harmonai
Senior Data Fellow, Belmont Data Collaborative

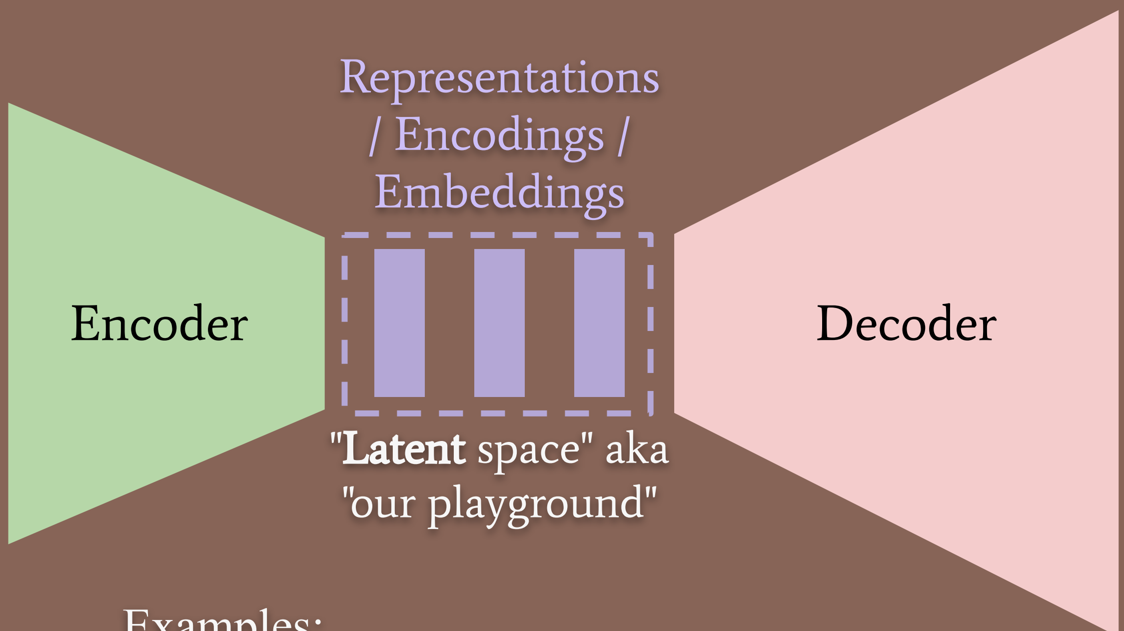
BELMONT UNIVERSITY

BELMONT DATA COLLABORATIVE

March 20, 2023, Data Science Nashville

Idea: Gettin' Fiddly With It…

Fun with Encodings / Representations of Music(al Audio)



Examples:
Sheet Music, Piano Rolls, MIDI,
Time Domain Samples, "Spectral Methods",
…Neural Networks

Fiddling With "Found Models"

Trend: People training very large models, lots of data & compute

esp. Getting an audio *decoder* to sound "good" is *difficult* and very *time-consuming*

Idea: What if we could "parasitically" manipulate the encodings* of large pretrained audio autoencoders to get them to do what we want?

Could make do with less data, less time.
(Maybe even "zero shot" style transfer?)

The Thread of This Talk…

Music representation learning for the purpose of audio (re-)*synthesis*.

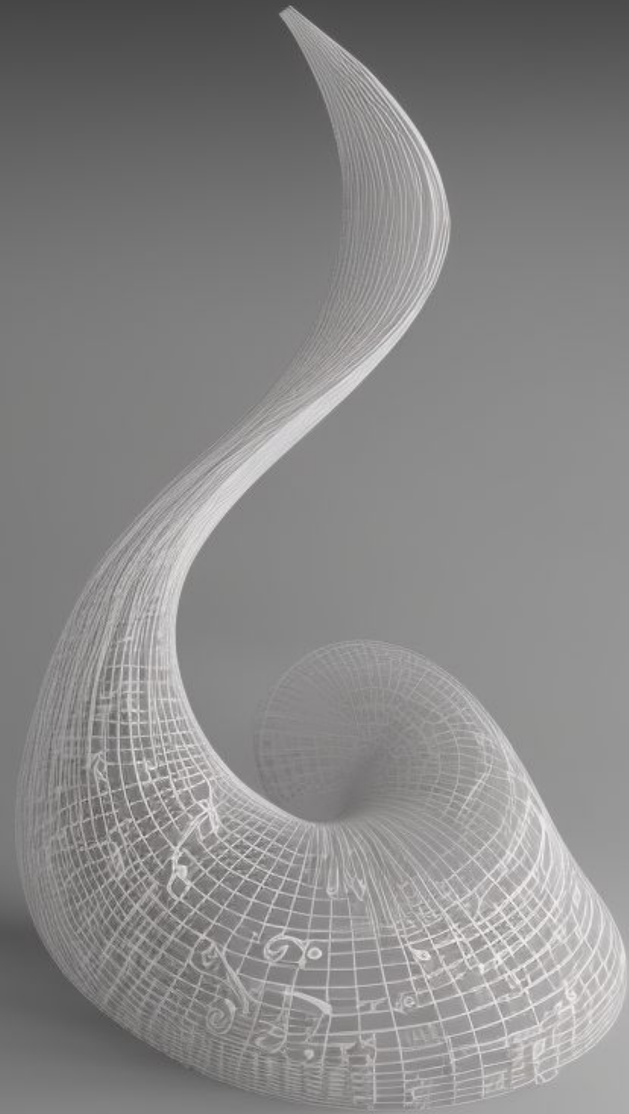
i.e., information theory, encodings, machine learning, music performance ...and some **geometry**!

Similarities to Music Information Retrieval (MIR) - Youngmoo Kim's group

..But differences:

- We have a decoder, not just encoder
- Our output has to "sound good"!

"We" = Harmonai.org: audio-lovin' open source group

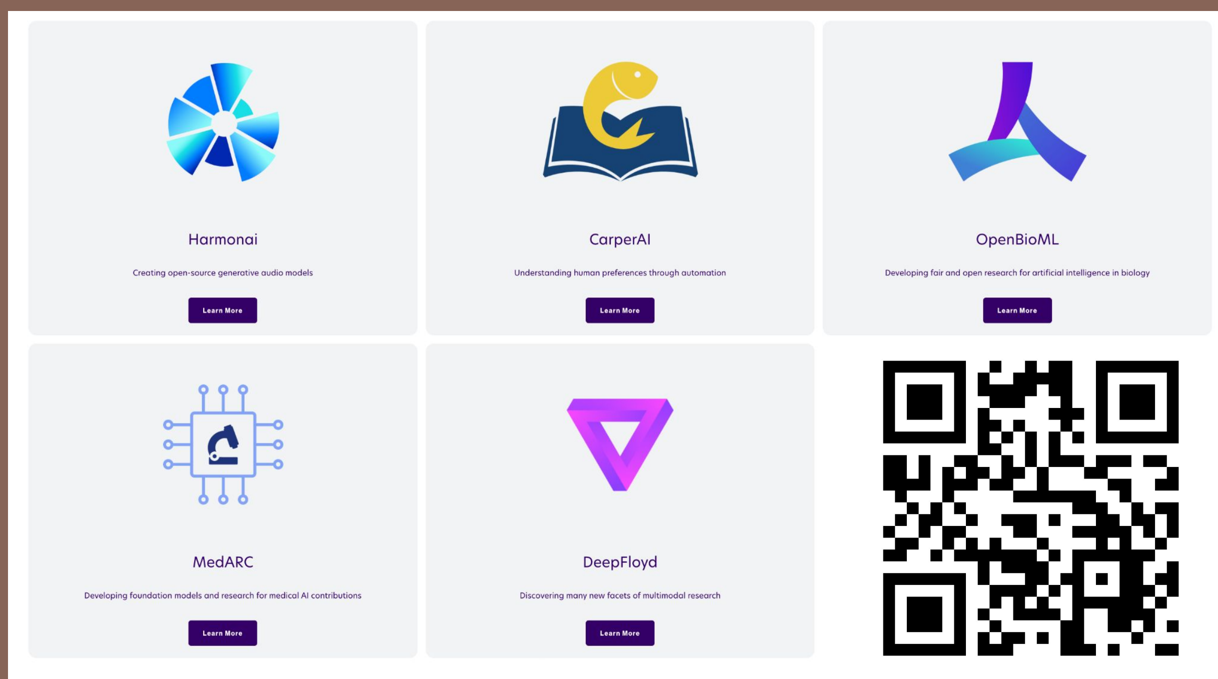


Harmonai(.org)

Notebooks & more https://github.com/harmonai-org

Open-source collective of ML-audio enthusiasts, researchers, hobbyists, musicians, engineers – lots of pro EDM DJs!

Supported as part of the Stability AI "family" of research orgs:



Credits

Work presented here is in concert with

Zach Evans, CJ Carr, Flavio Schneider, Harmonai server users (audio diffusion)

Christian Steinmetz (effects & representations)

Max Ortner, Nils Demerle', Antoine Caillon (geometry, RAVE)

Thread of the Speaker's Career

- PhD & Postdoc: Computational Physics: Solving Einstein's General Relativity (GR) on supercomputers: ~1996 to ~2008
- Playing music
- Teaching acoustics & electronics to Audio Engineering Technology (AET) students as prof. at Belmont University since 2006
- Exposed to machine learning (ML) for audio in 2013 and *got sucked in!*
- Got involved with Harmonai doing ML for musical audio on supercomputers, May 2022... & now even using some GR!

Ways to [Represent] 'Music' Info

> Sheet Music

Piano Rolls

MIDI

Time Domain Samples

"Spectral Methods"

...

Neural Networks

Ways to Represent 'Music' Info

Sheet Music

> Piano Rolls

MIDI

Time Domain Samples

Spectrograms

Ways to Represent 'Music' Info

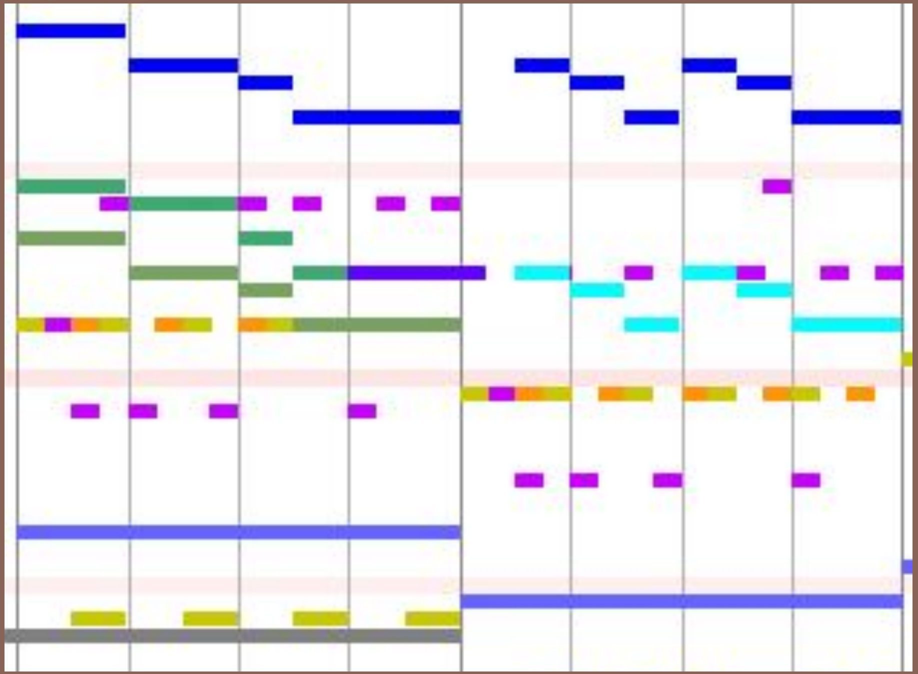
Sheet Music

Piano Rolls

> MIDI

Time Domain Samples

Spectrograms



Can control "semantic" / meaningful-to-humans info, such as instrument type, attack velocity,...

Requires crafting of virtual instruments

Ways to Represent 'Music' Info

Sheet Music

Piano Rolls

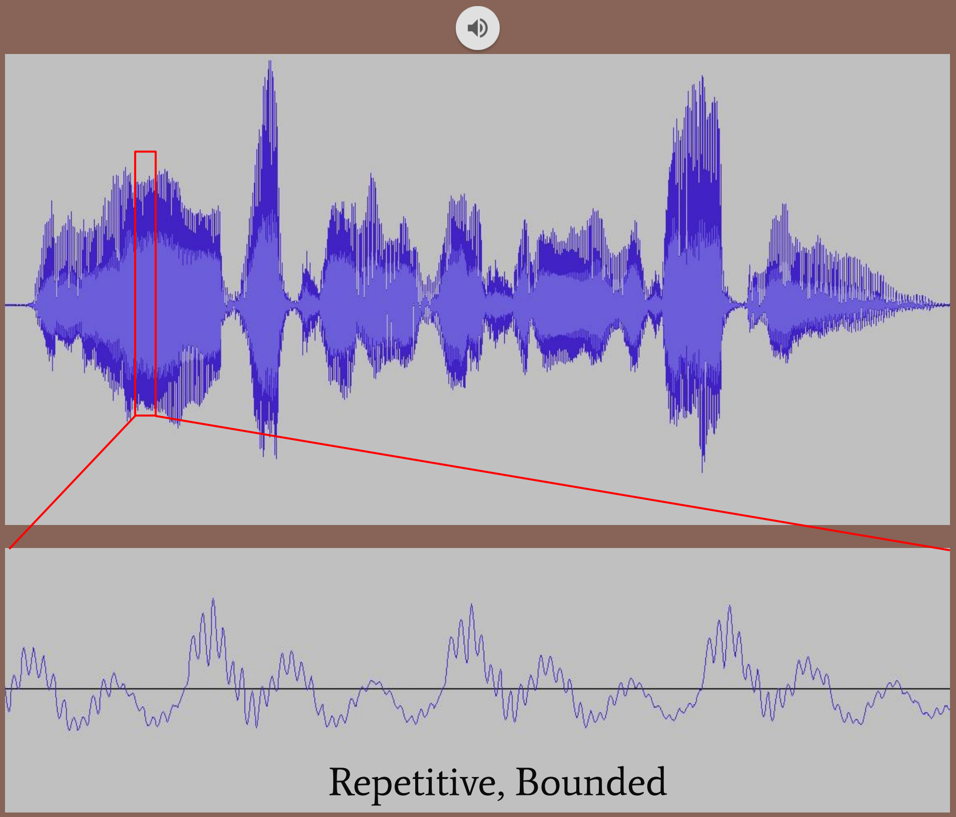
MIDI

> Time Domain Samples

"Spectral Methods"

...

Neural Networks



Instead, use sum of sine & cosine functions at different frequencies, i.e. "**Fourier basis**" → More generally, "Spectral Methods"

Ways to *Approximately* Represent 'Music' Info

Sheet Music

Piano Rolls

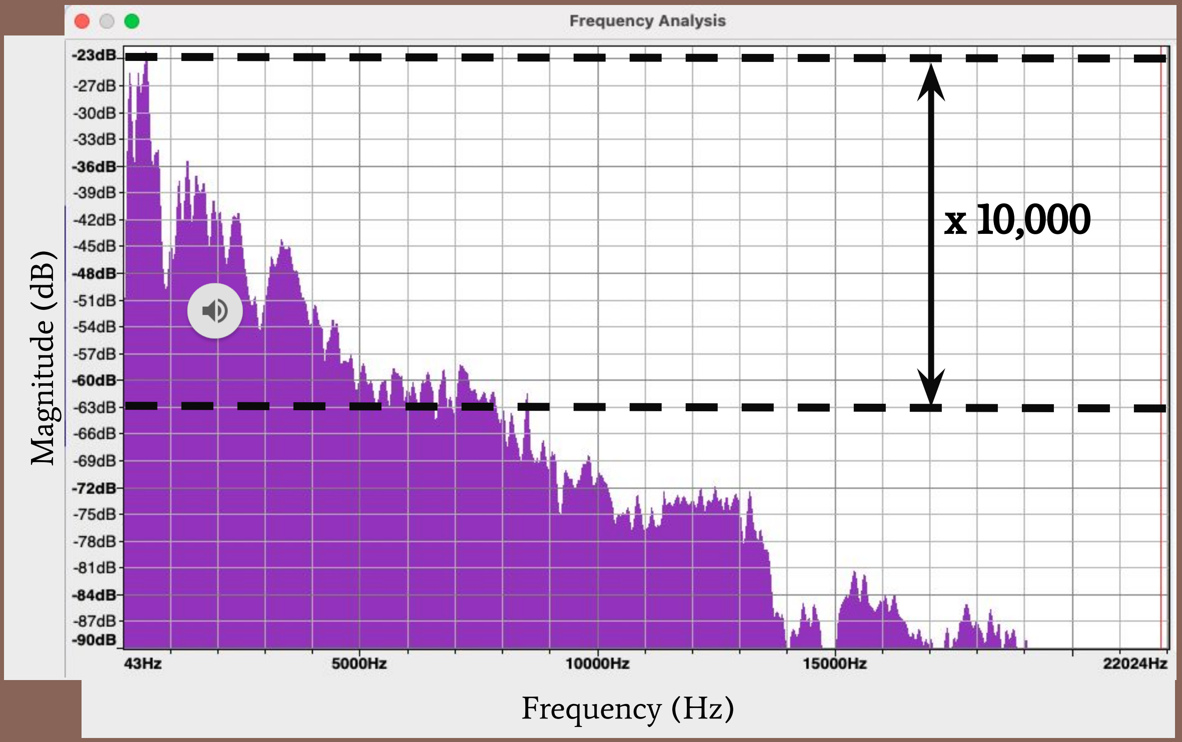
MIDI

Time Domain Samples

> "Spectral Methods"

...

Neural Networks



Also, we may "throw out" phase info,
e.g., human perception may ignore it

Ways to *Approximately* Represent 'Music' Info

Sheet Music

Piano Rolls

MIDI

Time Domain Samples

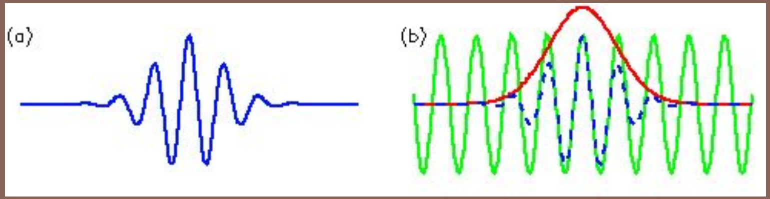
> "Spectral Methods"

...

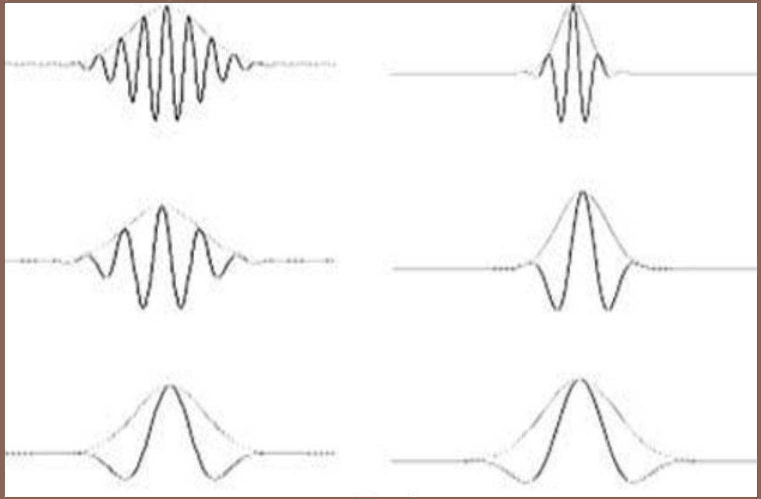
Neural Networks

Q: What if signal is more of a "blip" (transient)?
e.g., EKG data, seismology

A: For *sparser* representation(s), multiply Fourier sinusoids by compact "window" to make "wavelets":



Torrence & Compo, Wavelet Analysis



"A Review for Face Recognition using Gabor Wavelet Transform," Akshat Agrawal, 2017

Ways to *Approximately* Represent 'Music' Info

Sheet Music

Piano Rolls

MIDI

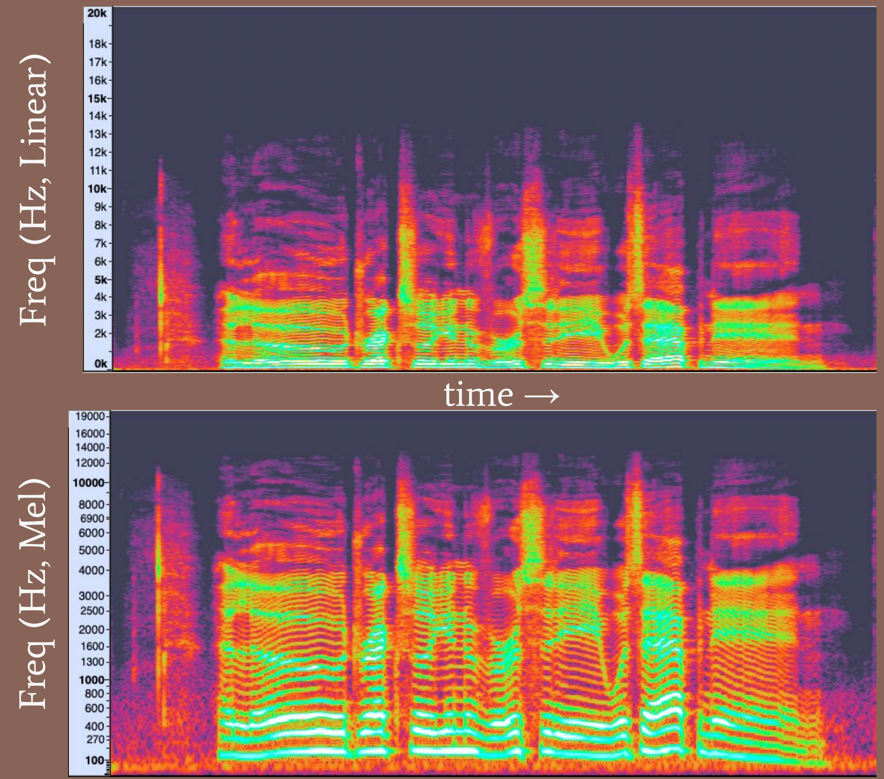
Time Domain Samples

> "Spectral Methods"

...

Neural Networks

Amplitudes of sinusoids over short times = "Short Time Fourier Transform (STFT)" aka "Spectrograms"



Humans discern *ratios* of freq, thus representation via *logarithmic* Mel scale tends to match human perception better than linear freq. scale

Typical Musical Audio Production Workflows…

…use some combination of

- **Time Domain** - e.g. reverb, dynamic range compression
- **Spectral Methods** - e.g., EQ filters
- MIDI (maybe) - e.g., virtual instruments, drum replacement,

i.e., "simple" tools, which…

…are *interpretable* algorithmically, but make it difficult to effect **complex** changes that are "semantically meaningful" to humans.

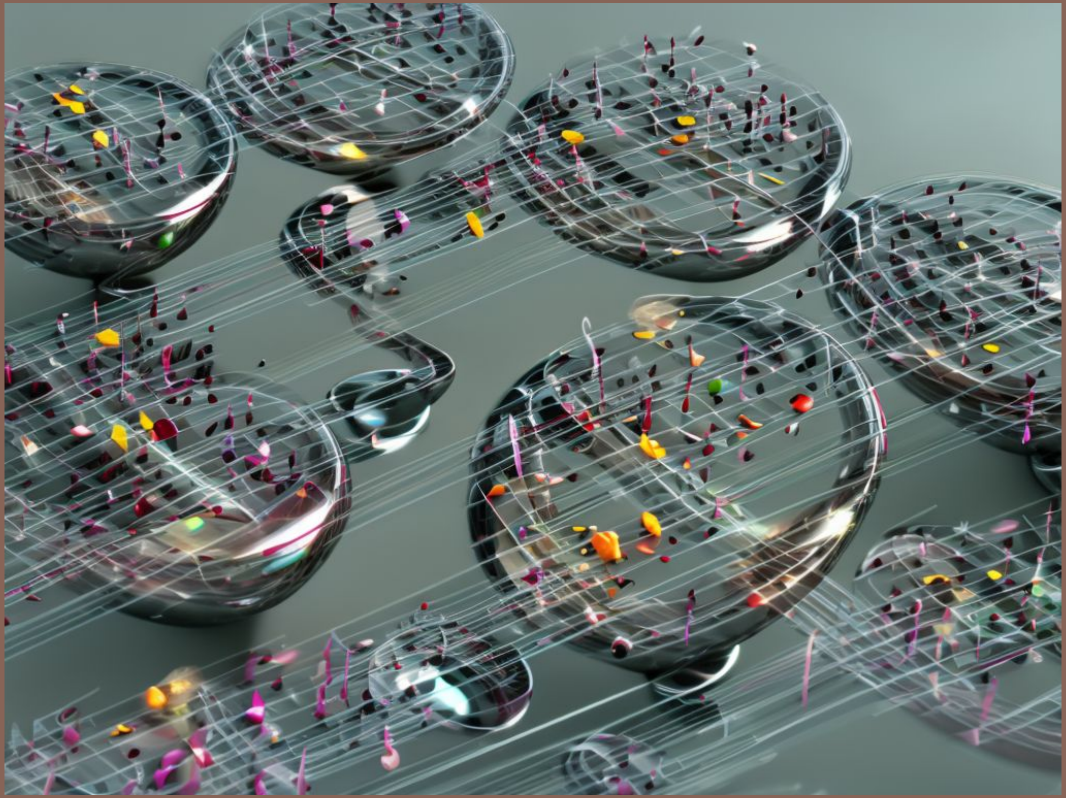
- source separation
- mixing/arranging based on qualitative criteria, e.g. "mood"
- "compositionality" - combining/replacing "attributes" of sound

So… Which Representations Most Closely Match Those of Humans *and* Allow Us to Do Cool Music Production (with Machines)?

e.g., Can We Build a "MIDI-like" Representation of General Musical Audio?

Audio sources as "objects" that can be manipulated independently?

Attributes such as style and content that can be interchanged – or combined – at will?



ML Audio Wizardry, 10 years ago

In 2013, Accusonus' "**DrumAtom**" plugin performed drum leakage removal, i.e. audio source separation. (*This* is what got me into ML!)

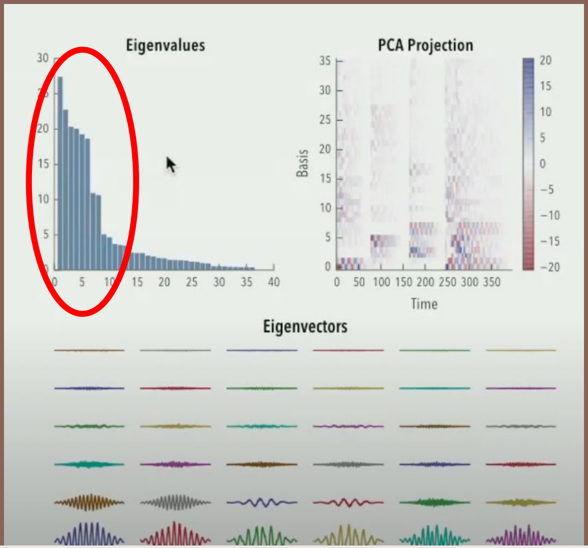
DrumAtom "learned" a (sparse) representation of **what makes a kick, snare, tom, etc.** from data, via "Non-negative Matrix Factorization" (NMF)

"brings new thinking to the old challenge of cutting out drum spill, and is often capable of producing results where existing processors fall short." – Sam Inglis, Sound on Sound, 2016

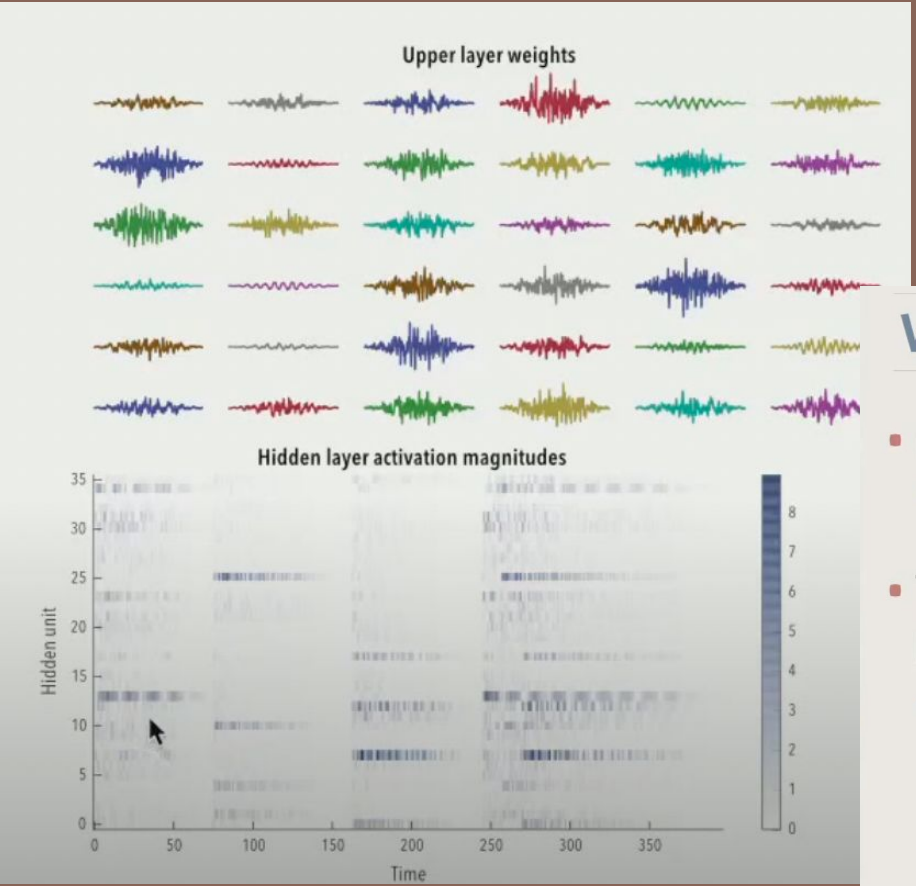
*Accusonus acquired by Meta in 2022 for "between €70 & €100 million."

And NMF and Neural Networks are "All the Same"

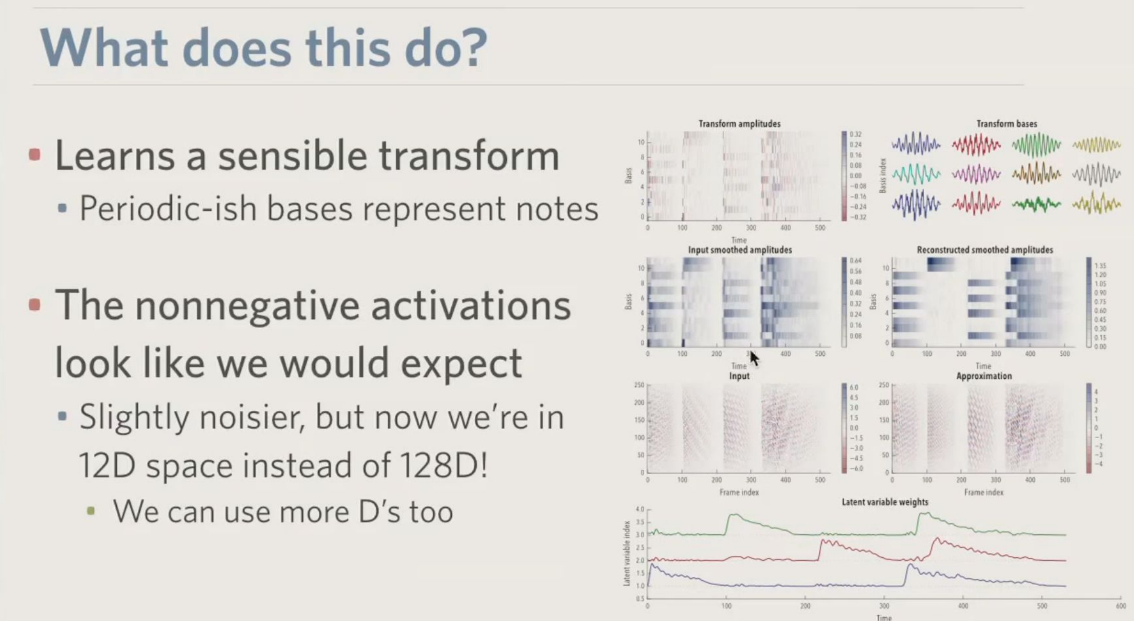
from **Paris Smaragdis**, "[NMF? Neural Nets? It's all the same](#)," SANE **2015**



"Activations" (i.e. amplitudes) are not simple to interpret mathematically – just various kinds of wiggles – but represent semantically meaningful musical info (to humans)

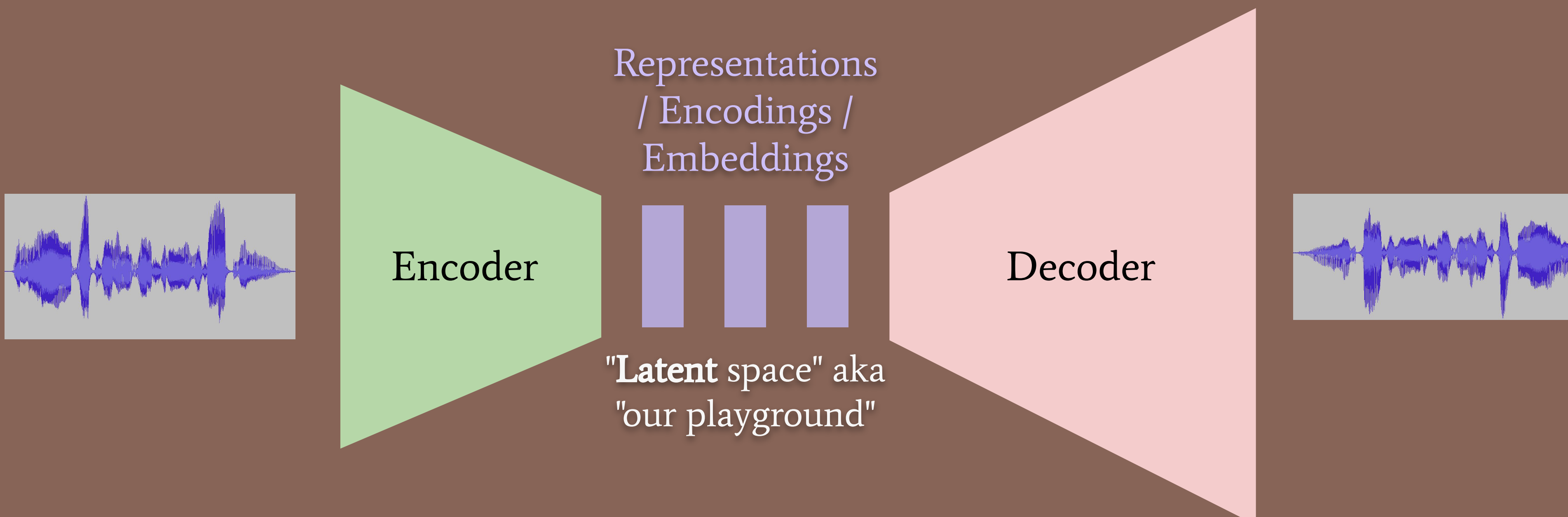


(Paragdis was hired by Peter Jackson to do source sep for Beatles pic "Get Back")



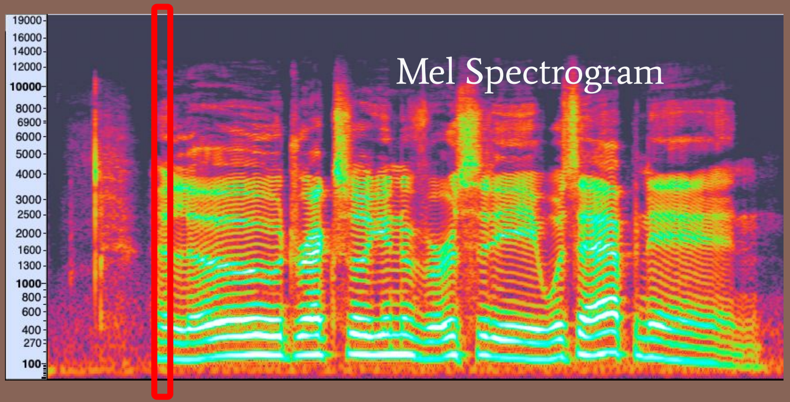
The General Picture: "AutoEncoders"

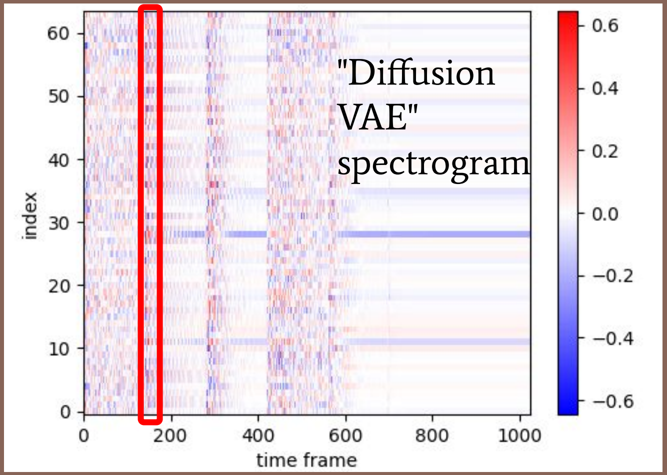
Technically only "auto" if output=input, "encoder-decoder" or "coder-decoder" ("codec"!), but people often say "autoencoder" anyway



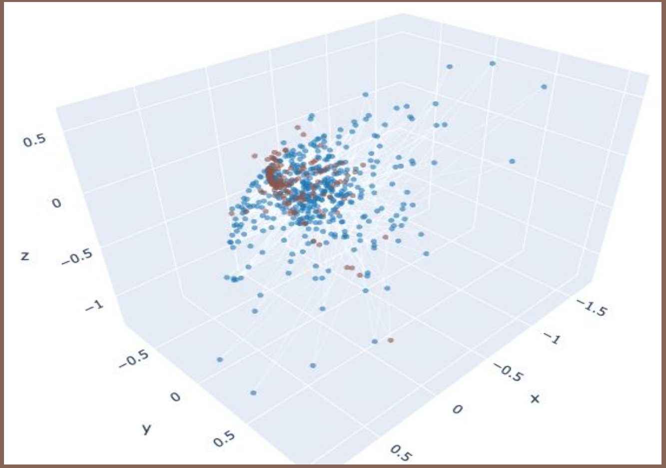
Could be spectrogram-based, or neural-network based, or...etc.

"Spectrograms" ⇔ "Vectors" ⇔ "Embeddings"





1. Take each "column" of "pixels" → array of numbers
2. Treat array as coordinates of a point in many-dimensional space, e.g. D=128
3. "vector" goes from origin to point
4. Regard this "vector space" as having *geometric* properties -> "embeddings".
   a. Time evolution = "trajectory" through "latent space"



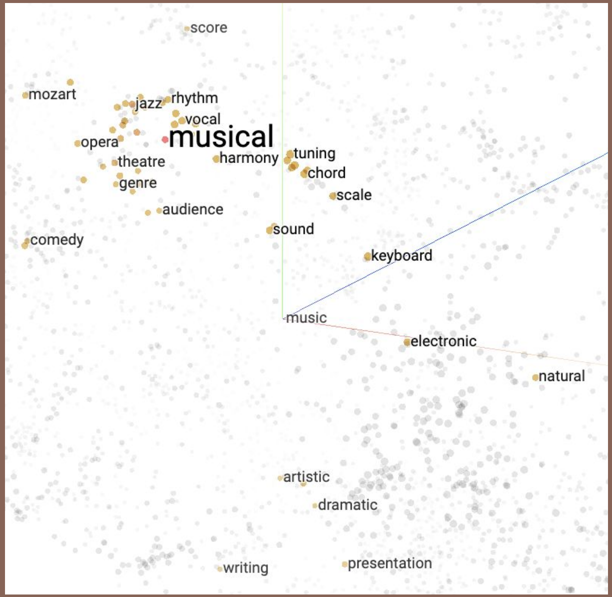
All of these are "representations"

Inspiring: Word Vectors

Language models operate on "tokens" representing words as vectors in big space

Train a big language model to either predict the next word ("autoregressive") or fill in missing ("masked") words

—> Find that the space has "semantic" structure!



https://projector.tensorflow.org/

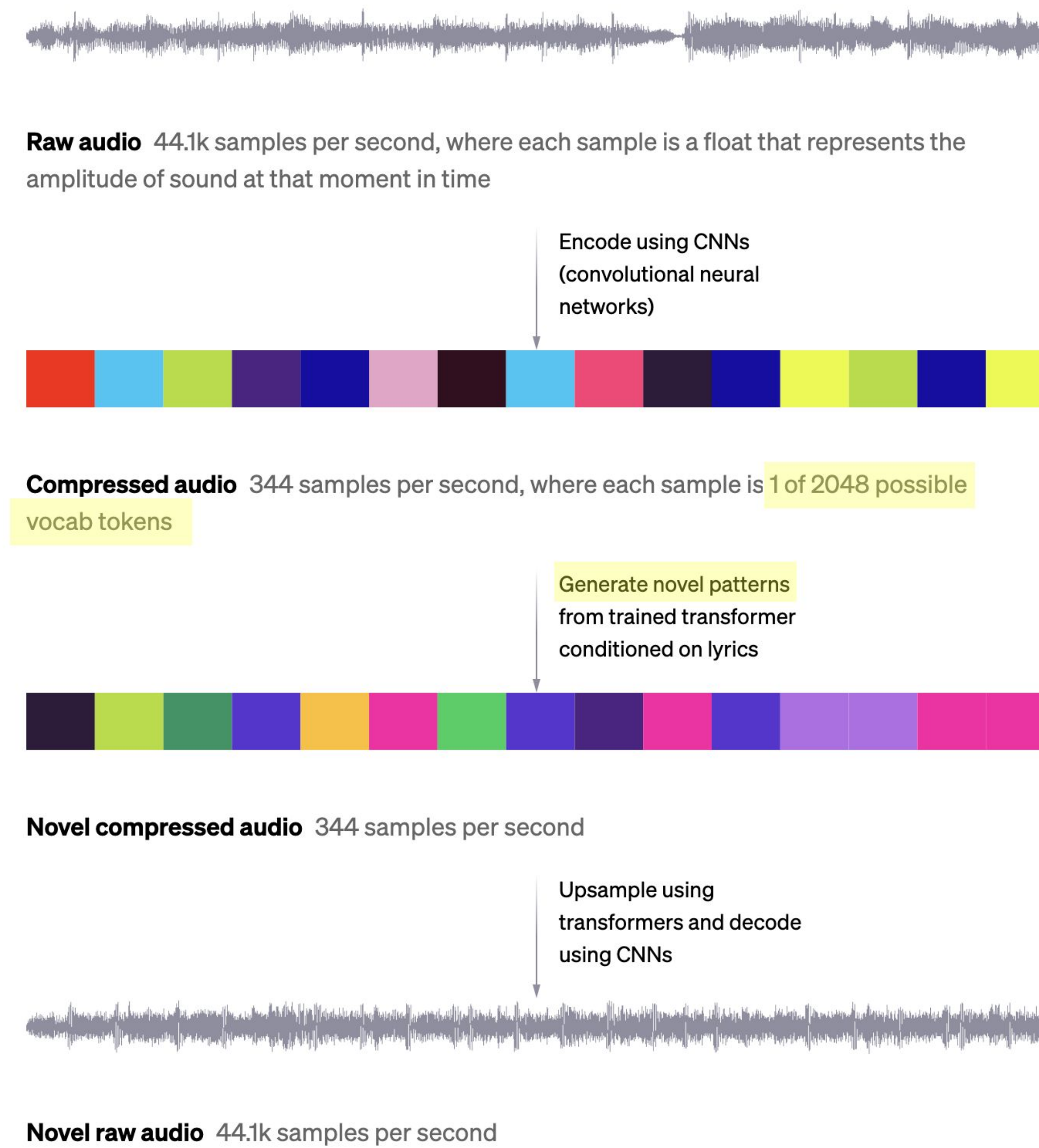
Famous results:

"king" – "man" + "woman" ≅ "queen"

(Country – Capital) is (nearly) invariant

OpenAI's Jukebox

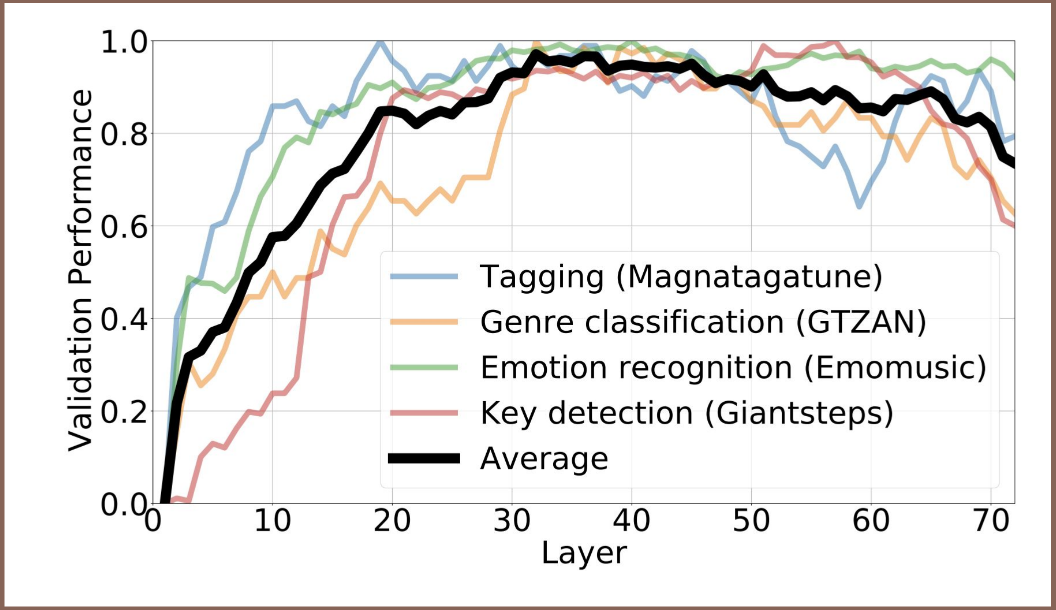
- "Language model" for audio: Treat representations as "tokens," i.e. "word vectors".
- "Autoregressive": predicts next "token"
- Also conditioned on *lyrics!*



Inspiration: Jukebox Representations!

"Codified audio language modeling learns useful representations for music information retrieval", Castellon, Donahue & Liang, 2021

Found that Jukebox's representations – trained for synthesis – containing "semantically useful" info about the music: genre, mood,...



What happens if we manipulate the embeddings?

Music Info Retrieval

- Besides MIR, can we operate on the embeddings to get new sounds?
- Can we do "music processing" in latent space?
- Is there musically something akin to "king – man + woman = queen"?
- i.e. is there an "Audio Algebra"?

- Besides Jukebox, which models should we consider?
  - Diffusion VAE (DVAE), Zach Evans et al
  - RAVE, Antoine Caillon et al
  - ...(your model here)

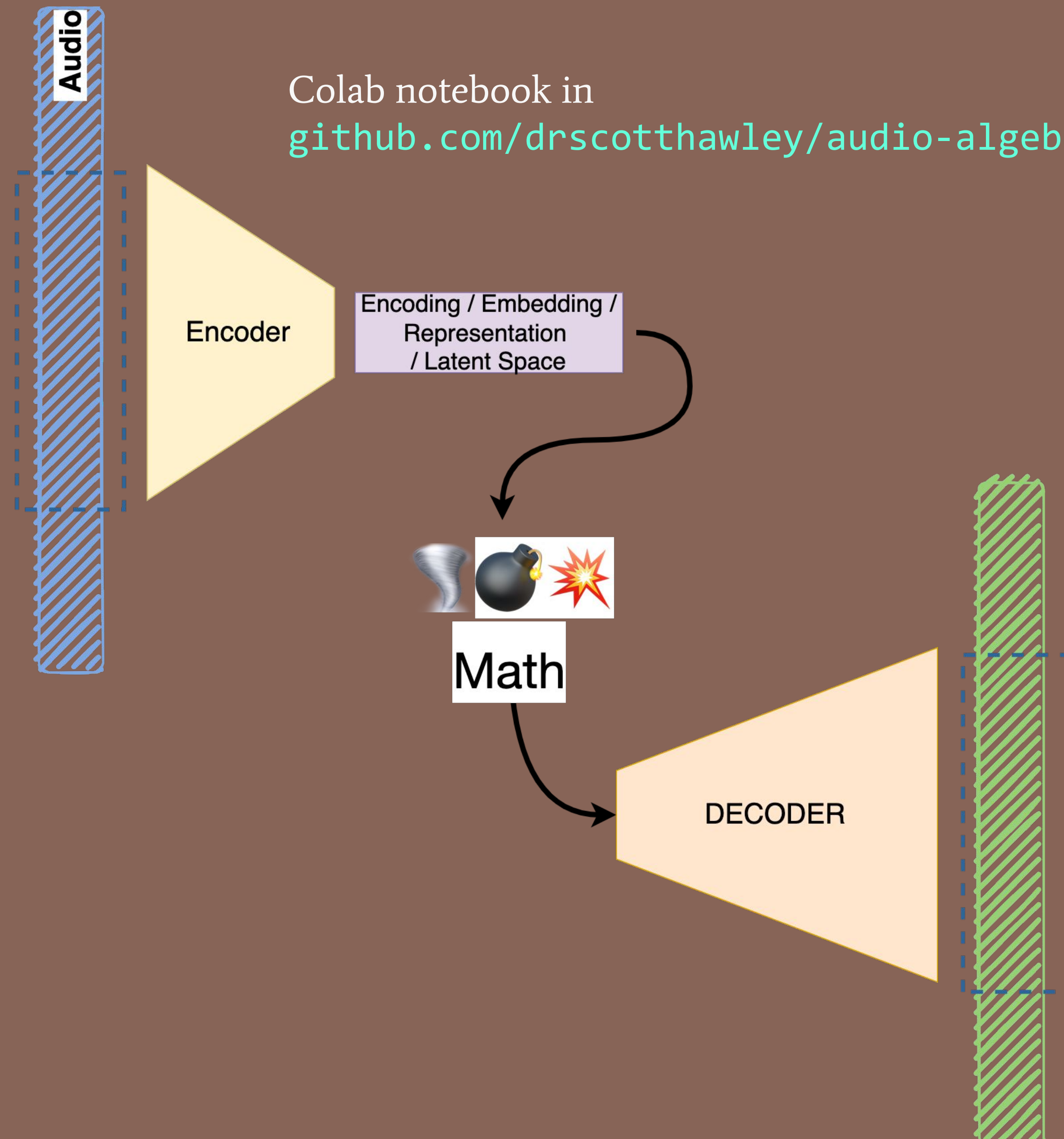
Destructo!

Colab notebook in
`github.com/drscotthawley/audio-algebra`

Idea: Encode audio, manipulate the representations, decode, see what it sounds like

Just various math operations that "mathemangle" the embeddings...

...Sound pretty awful.

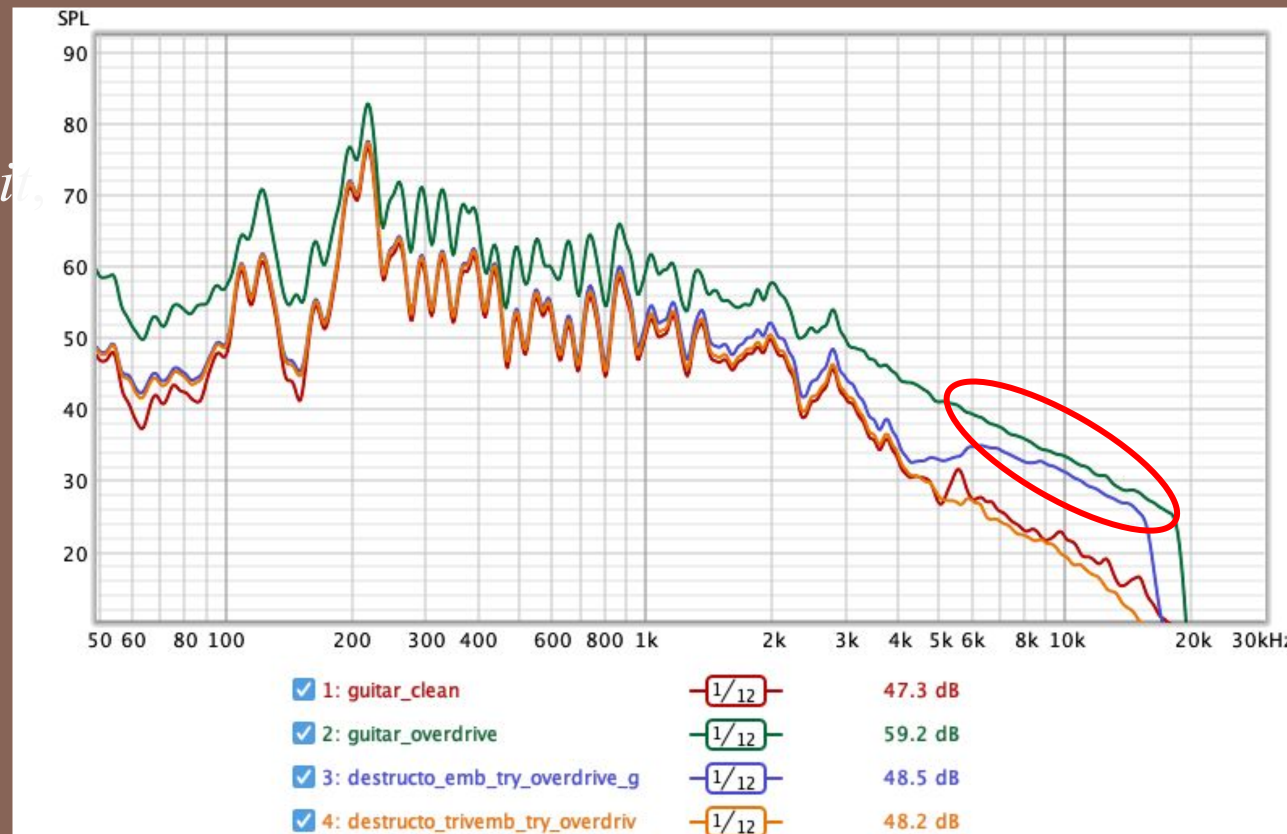


Processing: "Destructo" Colab

What about effects? $\langle \mathbf{piano}_{\mathbf{wet}} — \mathbf{piano}_{\mathbf{dry}} \rangle + \mathbf{guitar}_{\mathbf{dry}} = ?$

…for overdrive & reverb.

Overdrive: maybe *a bit,* for 6-16 kHz:

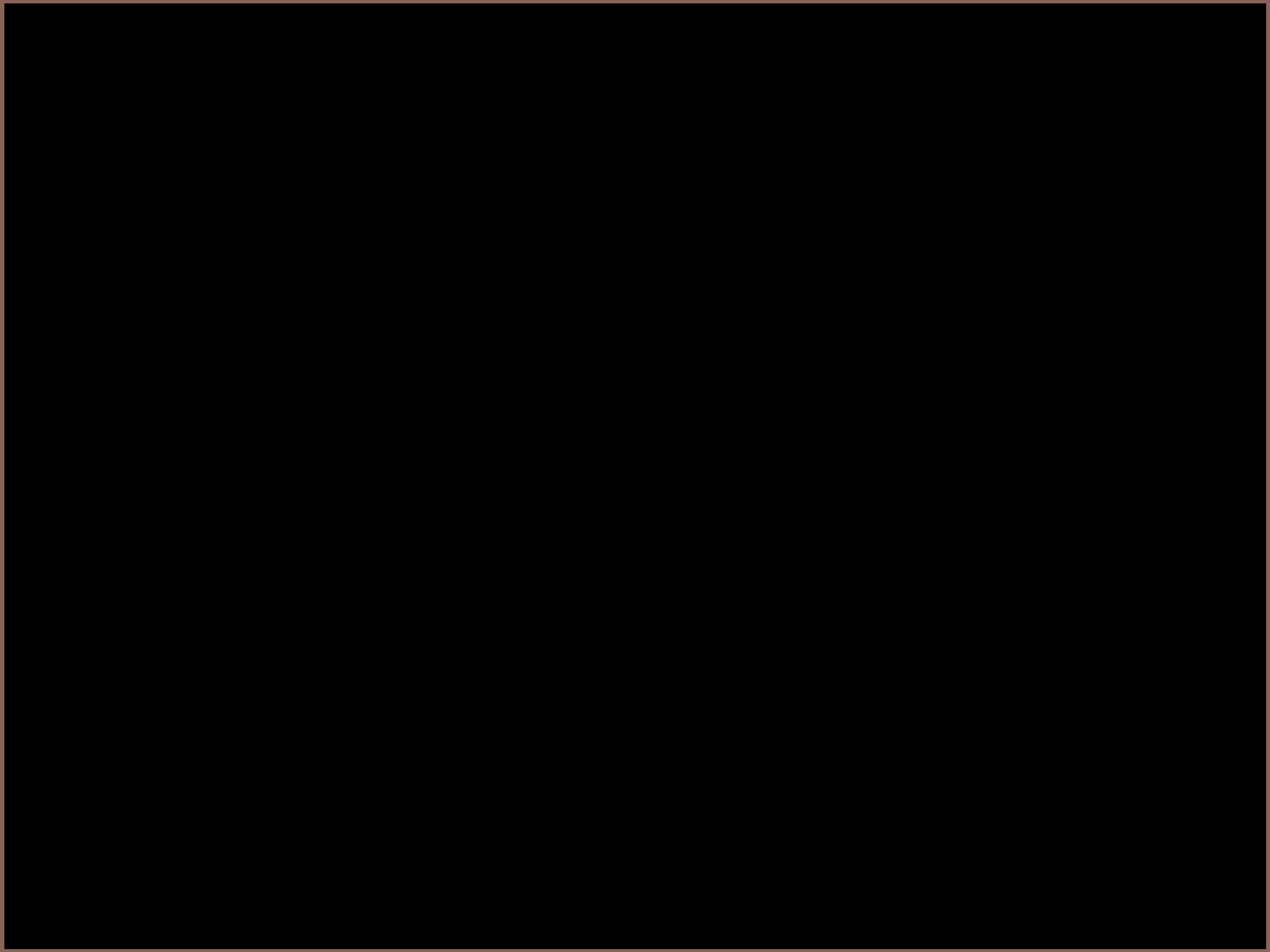


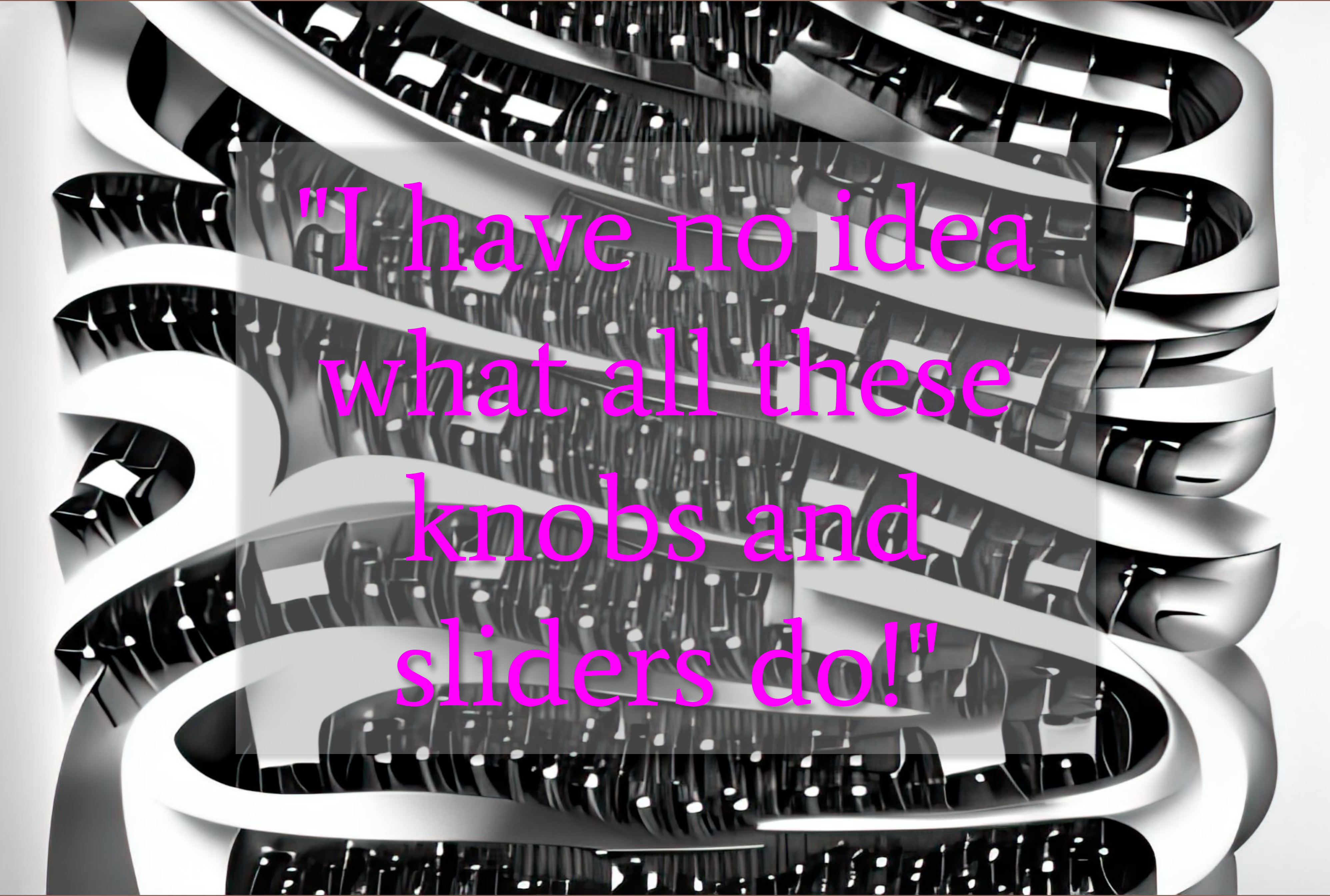
Reverb: Not yet

How to get "better" embeddings? or better *control* of embeddings?

Directions not Uniform (new result!)

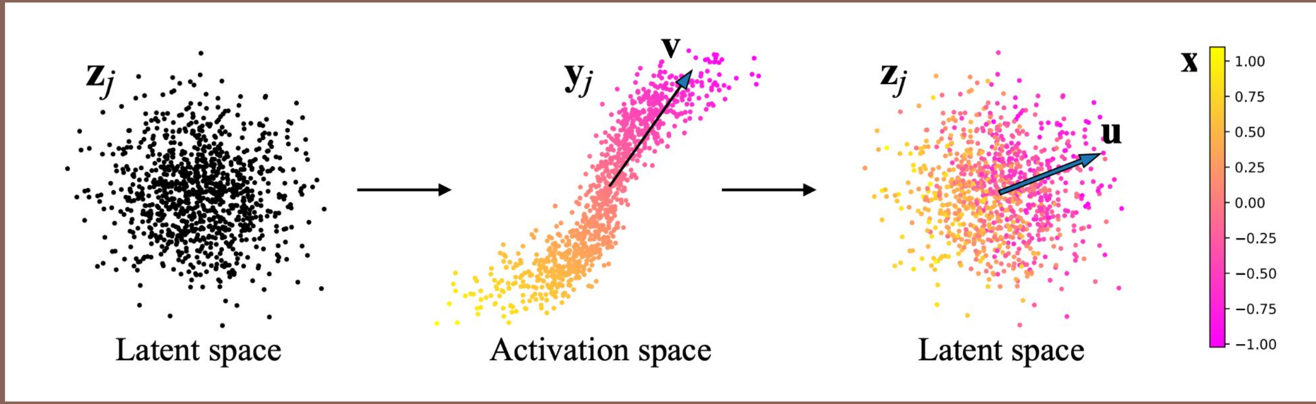
"More Distortion" for guitar (left) & piano (right):




"I have no idea what all these knobs and sliders do!"

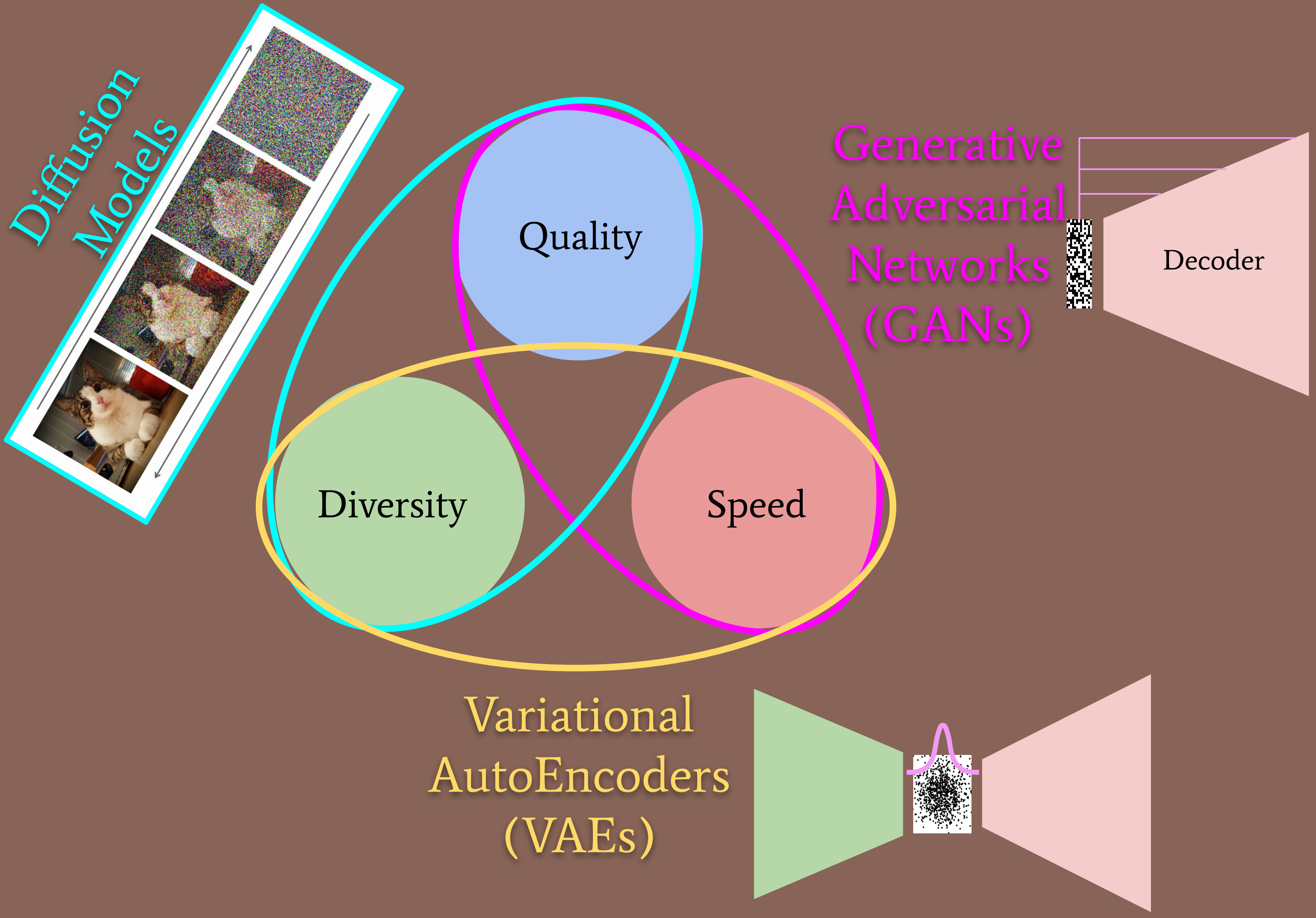
Idea: *Discover* Meaningful Directions

"GANSpace: Discovering Interpretable GAN Controls," Härkönen et al, 2020
Take **general model** trained for **synthesis**, measure which directions cause the most variance, inspect results -> Produce (smaller number of) knobs / sliders

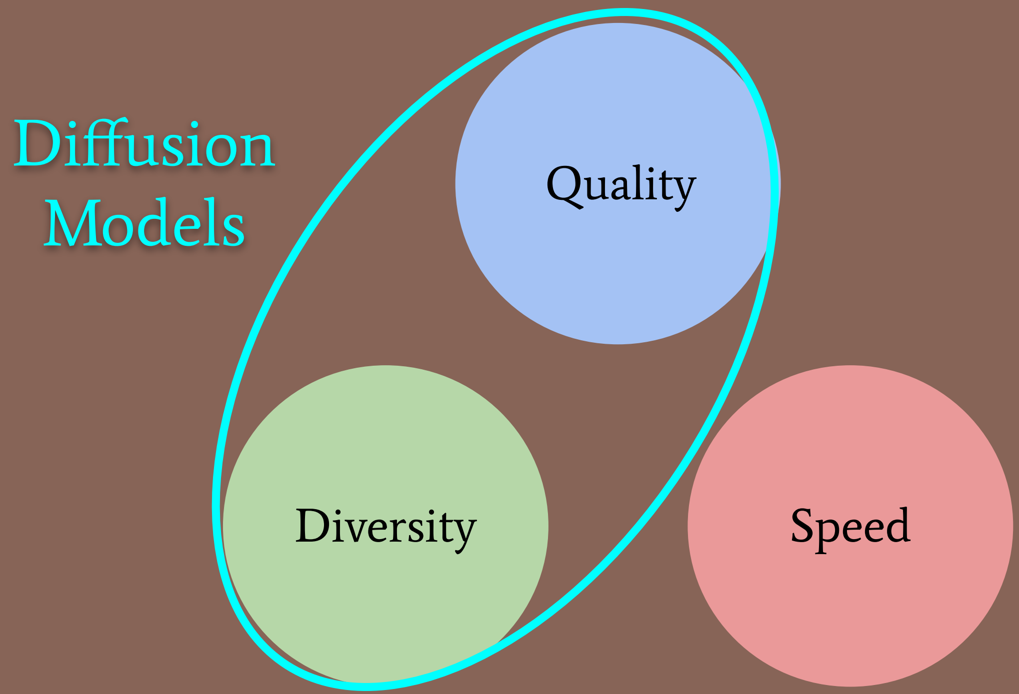


WANTED: "Fluffiness" knob for AUDIO!

Generative Models: Pick Two?



Dance Diffusion (& Stable Diffusion, Riffusion, etc)



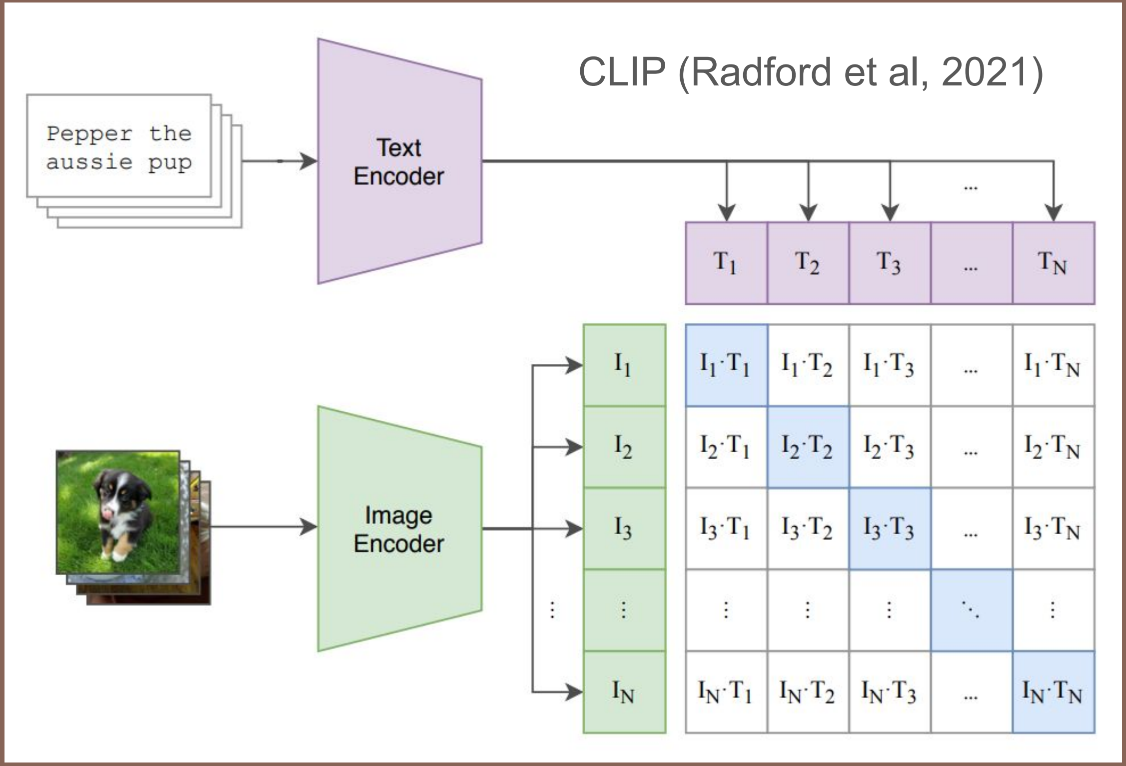
Each act of generation requires "sampling," i.e. diff. eq. solve

Idea: Joint Text-Image Embeddings

Two separate encoder models: text encoder & image encoder

Made such that text-image pairs that "go together" map to the same location in embedding space.

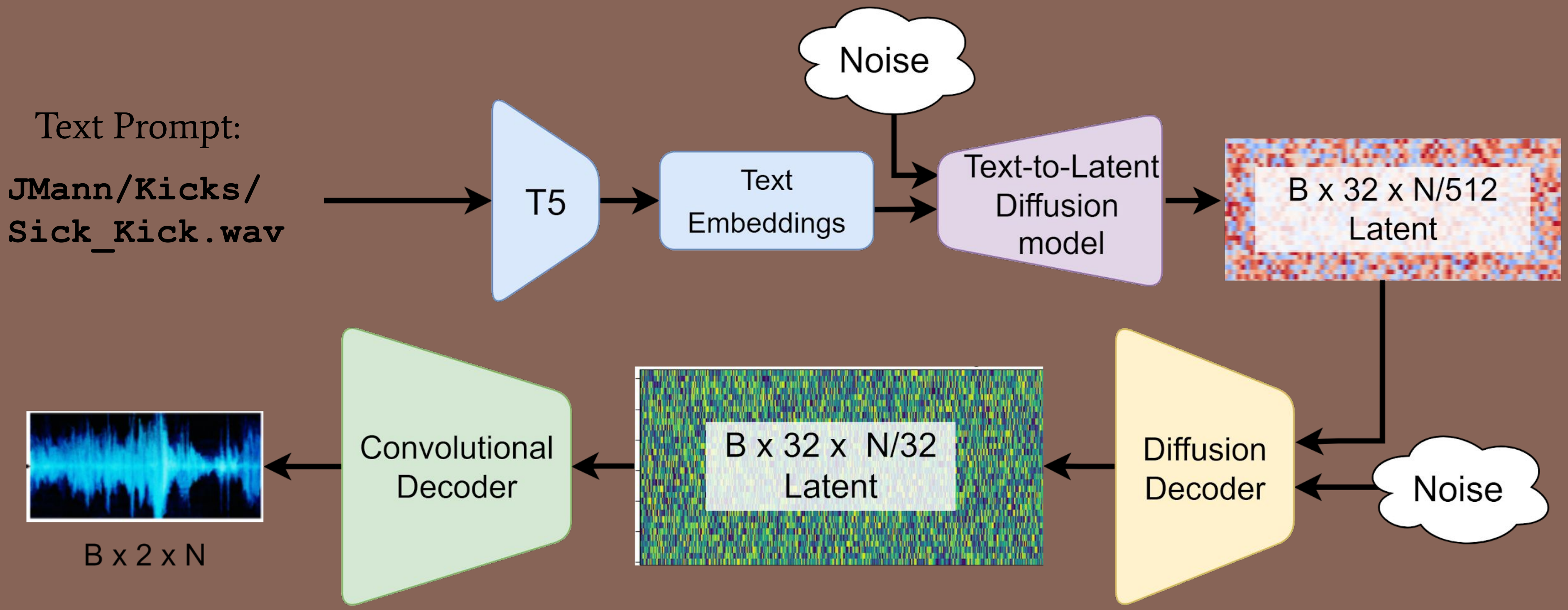
- Can be used to **guide*** denoising process for text-to-image generation



"Guidance is an add on". CLIP guidance examples:
BigSleep (GAN), VQGAN+CLIP, Stable Diffusion,...

Dance Diffusion / Sample Diffusion

Two "levels" of latents



"Sample Diffusion"

Audio Examples (2s)

Prompt = (Make up a fake file path)

Note: Dataset is dominated by electronic music (dubstep, DnB,...)

- "FakeDJ/Kicks/Sick_Kick.wav"
- "FakeDJ/Snares/Distorted Snare.wav"
- "FakeDJ/Toms/Tom Roll.wav"
- "FakeDJ/Strings/Quartet Pad.wav"
- "Breaks/Amen_Break.wav"
- "FakeDJ/AwesomeCuts/Hybrid Trap"
- "Samples/Special_Effects/Explosion 01.wav"
- "Sounds of Fumbling/Damn I Dropped a Bunch of Stuff on the Floor.flac"
- "Loops/Bass/Bass Loop 1 150 BPM.wav"
- "Loops/Bass/Bass Loop 1 100 BPM.wav"

RAVE (Caillon et al, IRCAM)

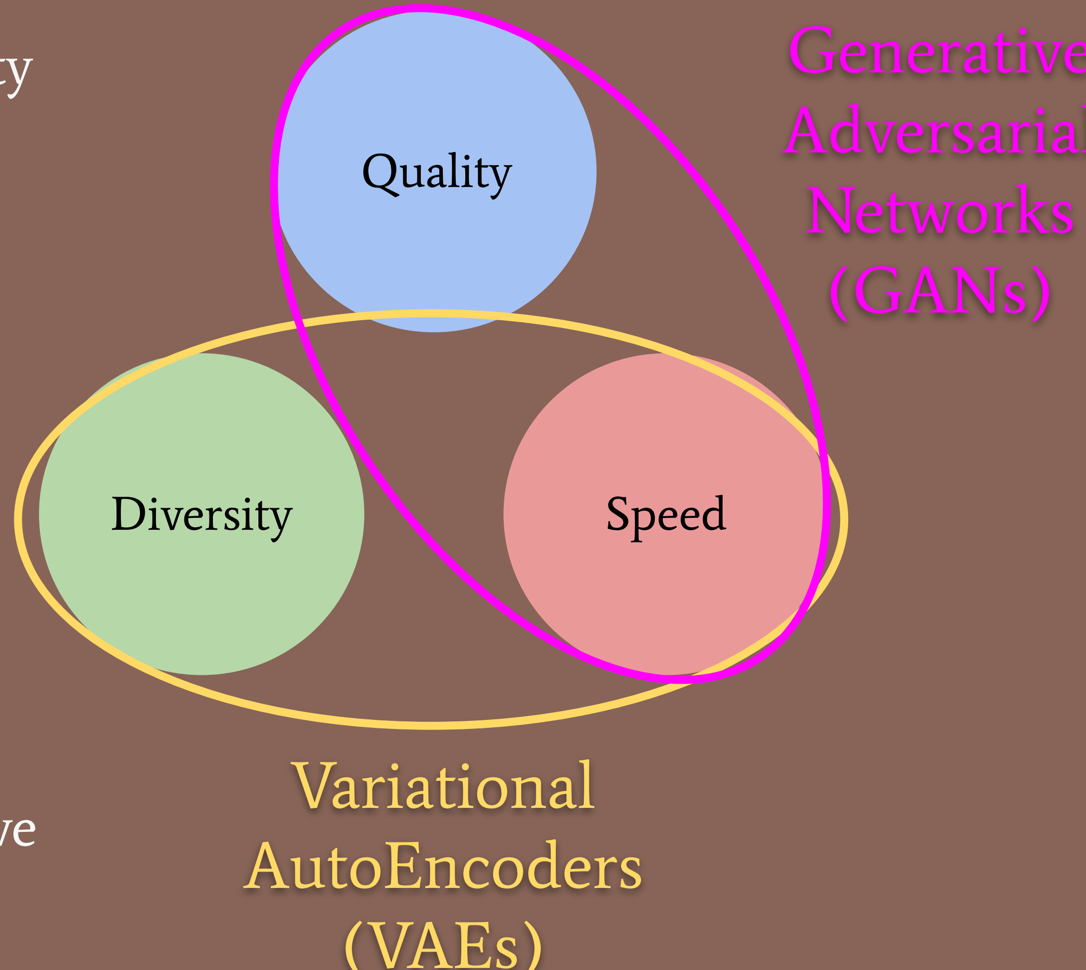
RAVE is a VAE, *but* final training stage is performed using a GAN-like loss for better quality

RAVE auto- prioritizes "interesting" directions

Runs *fast*, is lightweight: Can run in real time on Raspberry Pi!

Often used for style transfer- change instrument/voice type

Can also train an autoregressive "prior" to predict trajectories through latent space!



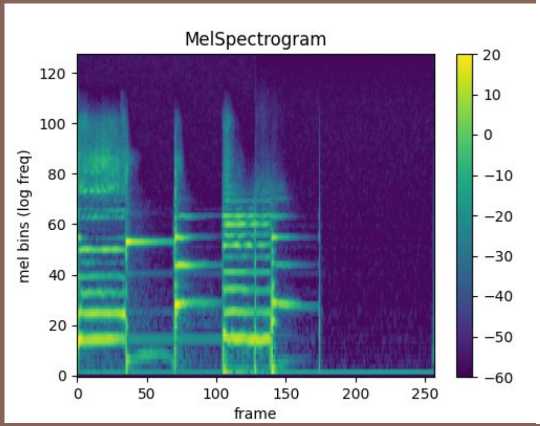
RAVE

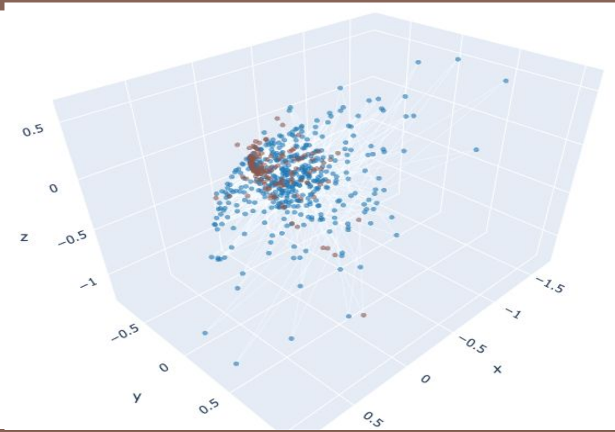

LAB
SCIENCES ET TECHNOLOGIES
DE LA MUSIQUE ET DU SON
Antoine Caillon
Apprentissage temporel hiérarchique
pour la synthèse audio neuronale de la musique
Équipe Représentations musicales

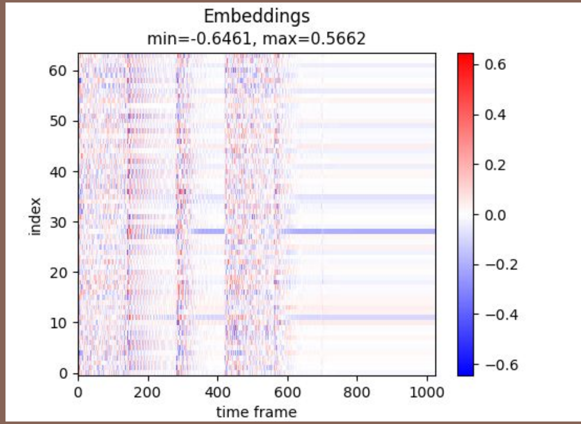
Visualizing embeddings: `aeiou.viz`

`pip install aeiou`

Before:

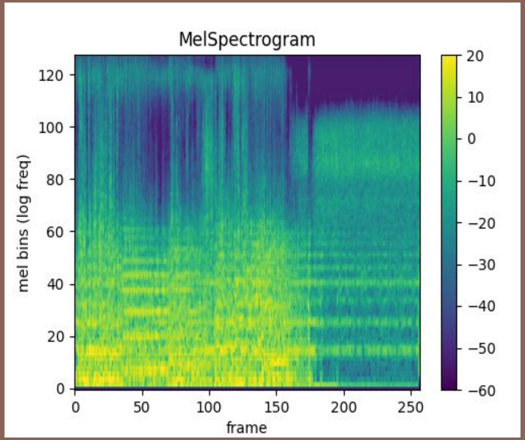


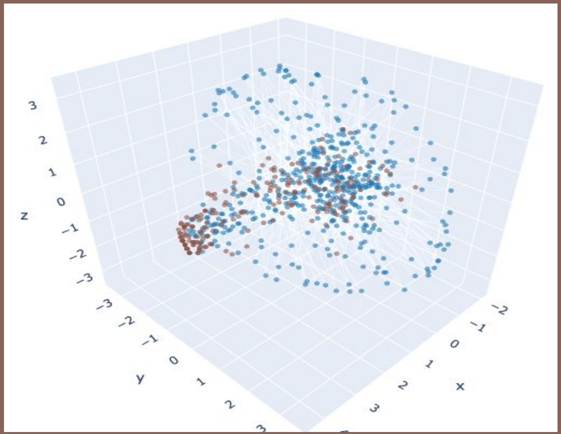


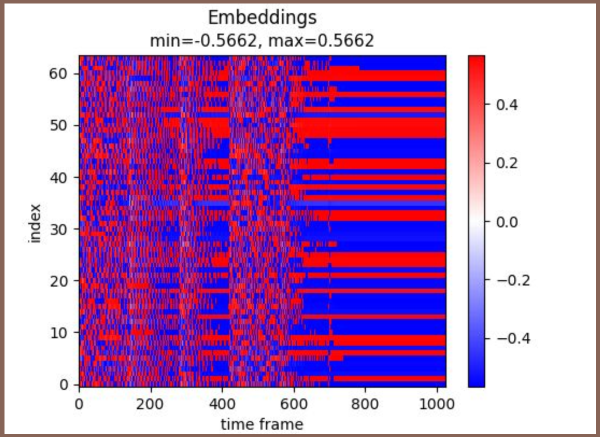


—> Destructo! <—

After:







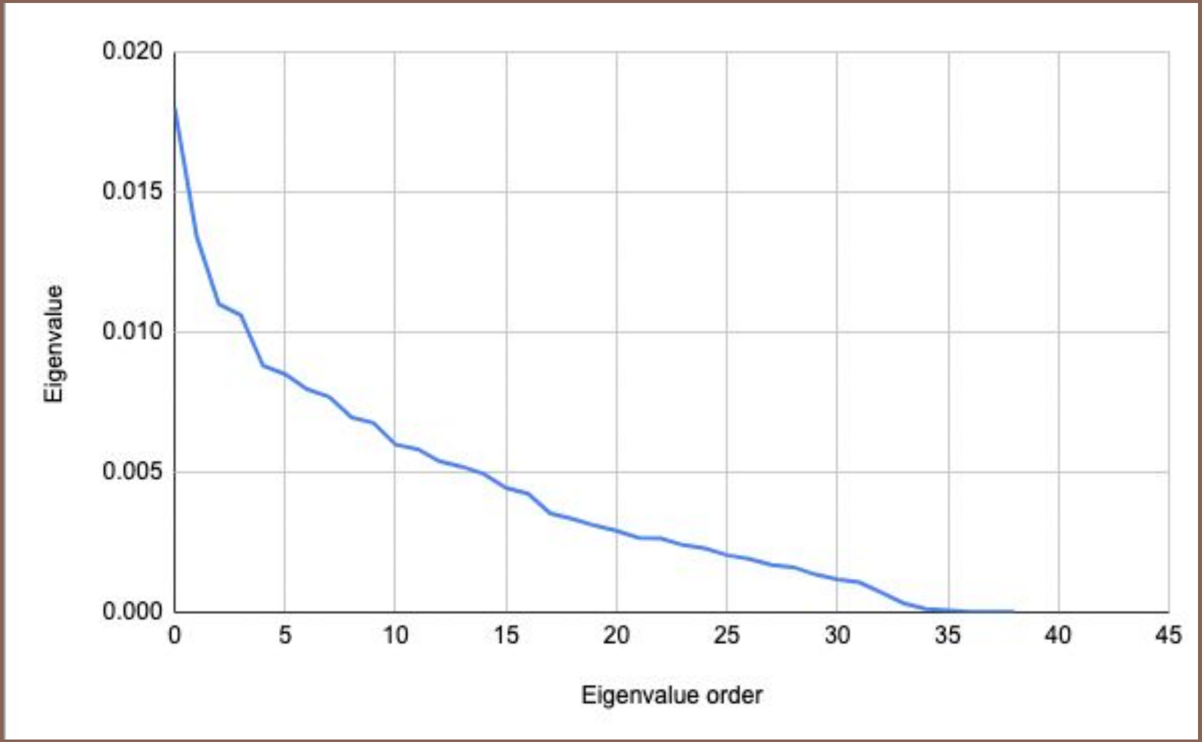
Better Embeddings: Masked AutoEncoder?

Masking (removing sections of inputs & having model learn to "fill in" missing data) can help language models (e.g. BERT) learn "semantically meaningful" embeddings.

Q: What about audio?   A: Depends on the autoencoder!

Principal 'directions'/'knobs'/'sliders' for DVAE

64 dimensions (+ time), but really only the first 32 do anything.



Create 'Bespoke' Embedding Spaces: `audio-algebra`

"Splice" into the middle of pre-trained autoencoders that are optimized for synthesis – whose representations may not be 'semantic' –

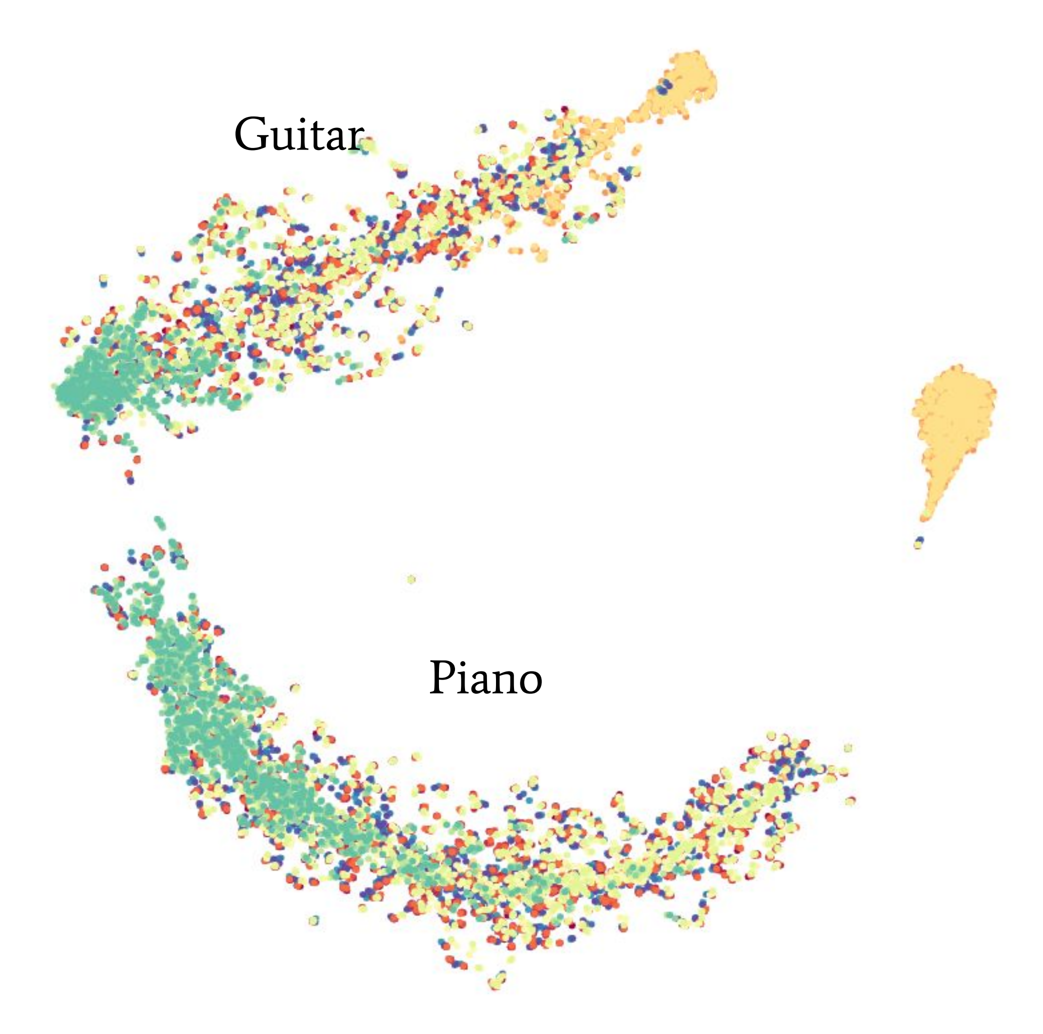
Invertibly map to a new space, with 'custom' similarity measure(s):

- instruments
- **audio effects**
- mixing

Potential "Map" of Audio Attributes

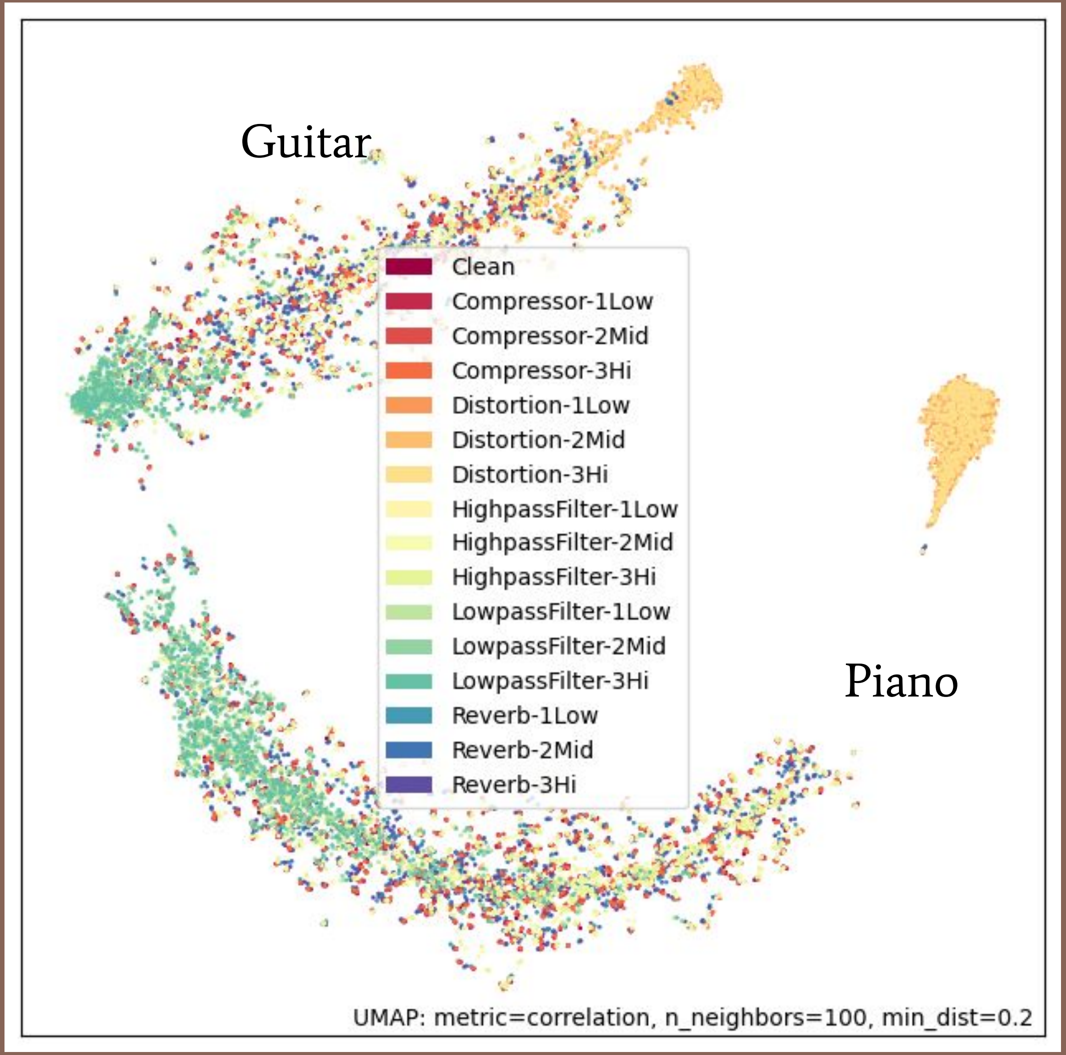
EDA: Real "Map" of Audio Effect Attributes

(UMAP)



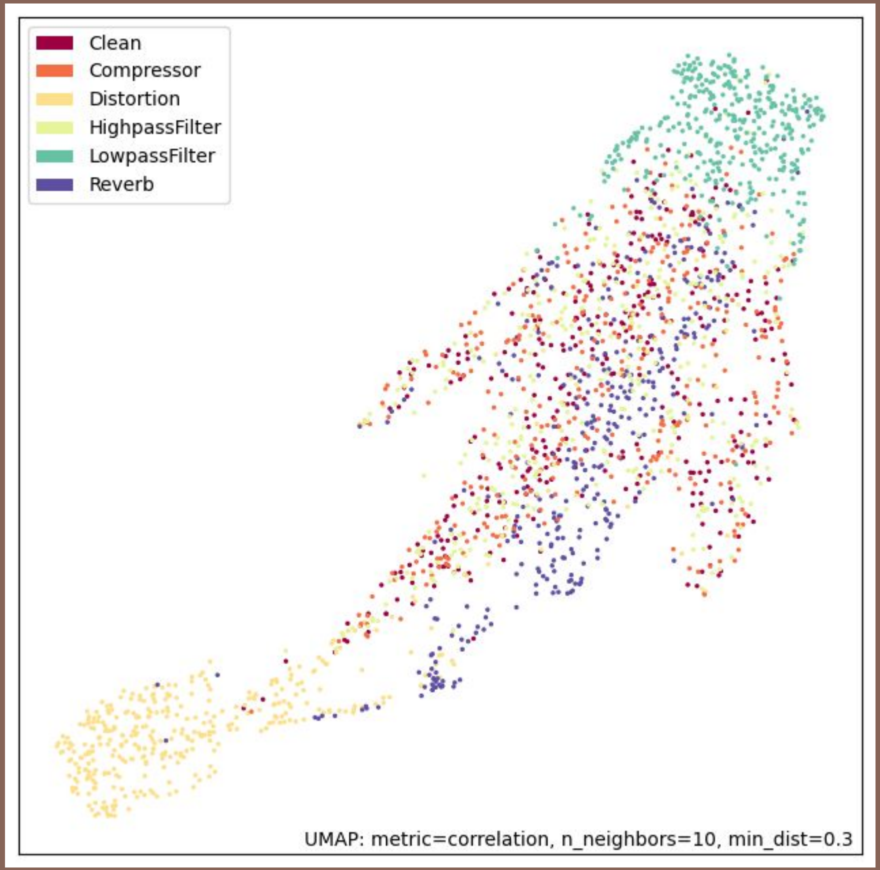
EDA: Real "Map" of Audio Effect Attributes

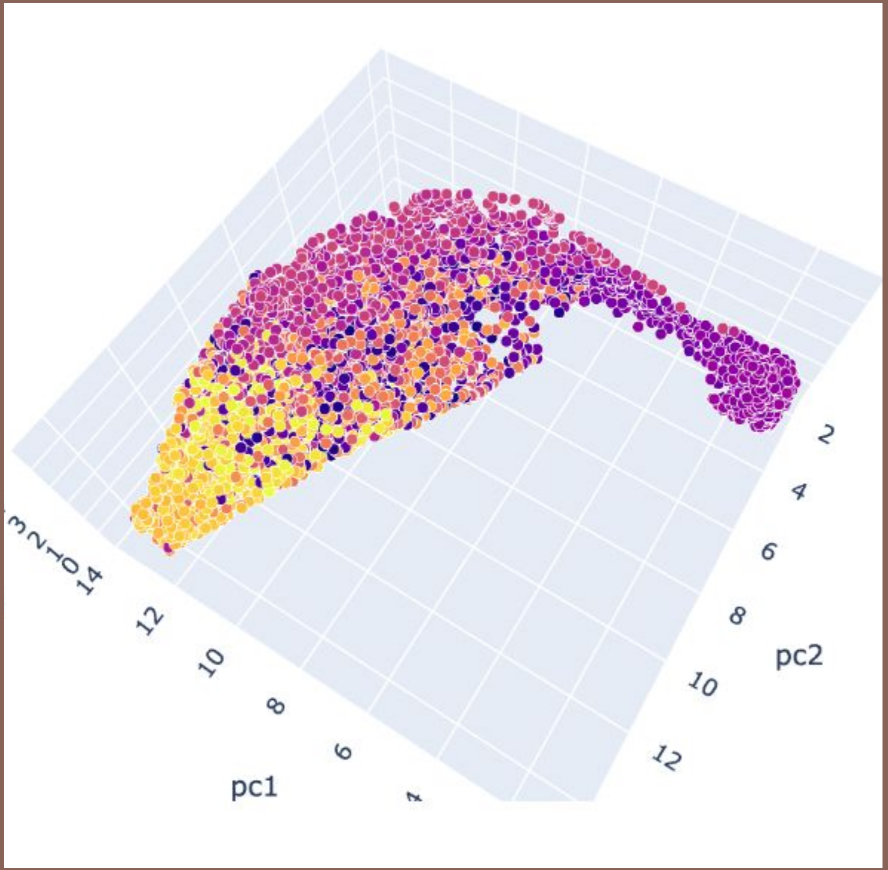
(UMAP)



"Seeing Animal Shapes in UMAP Embeddings"

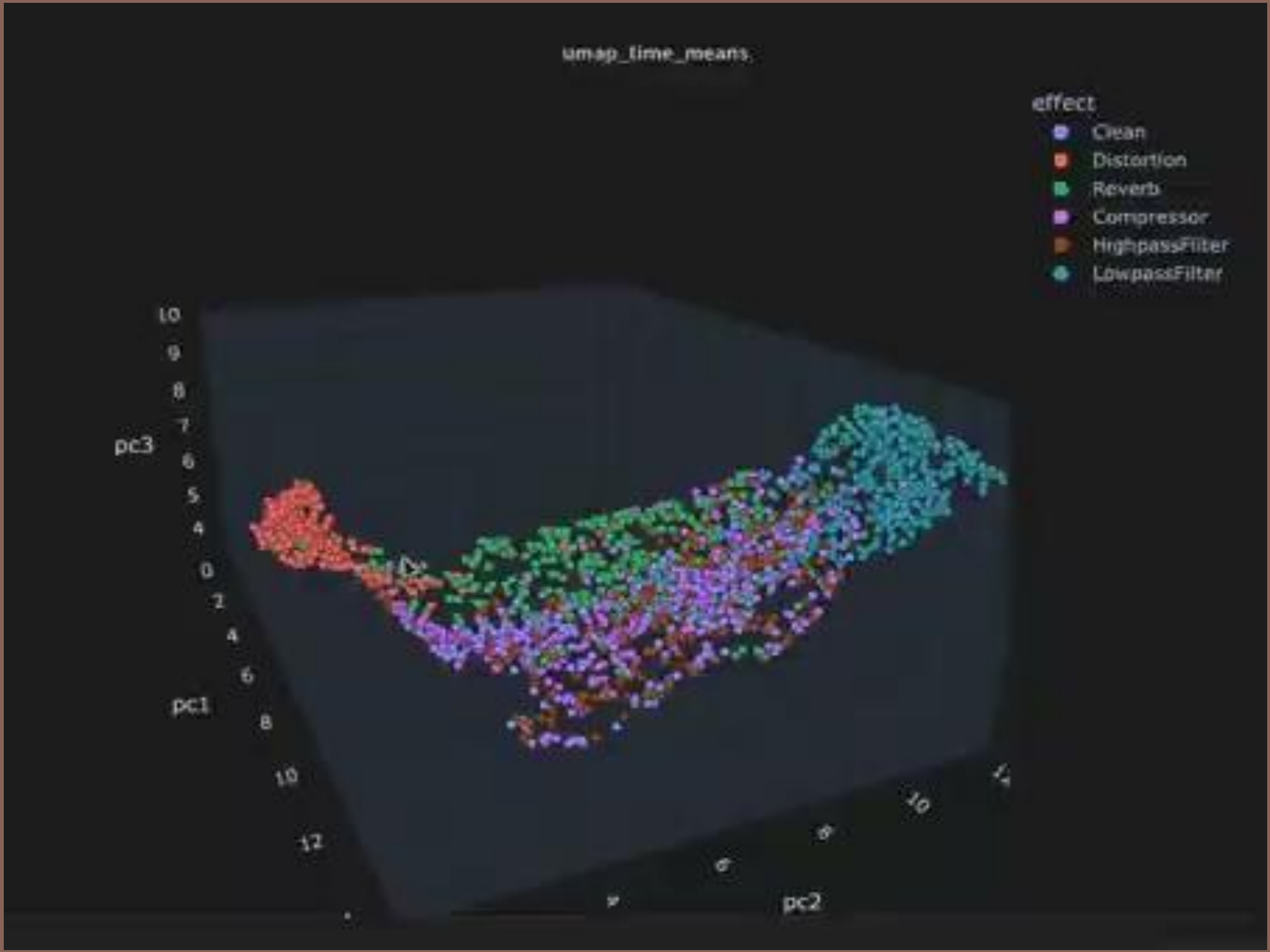
Fish!





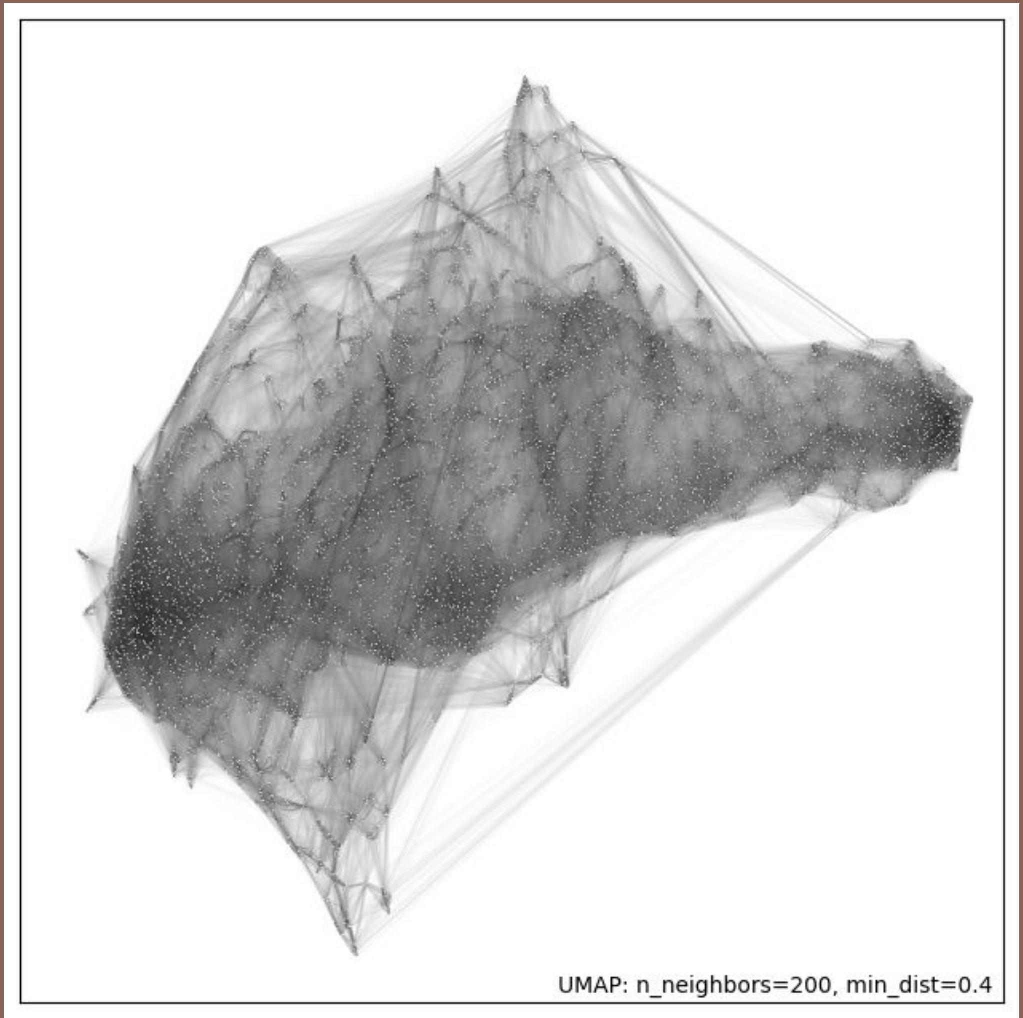
"Seeing Animal Shapes in UMAP Embeddings"

Seal!



"Seeing Animal Shapes in UMAP Embeddings"

Horsie!

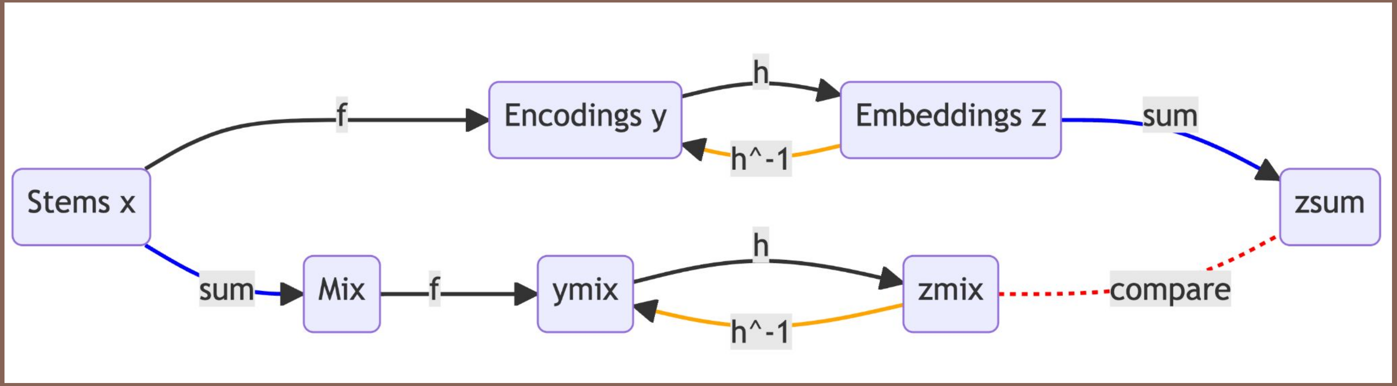


Idea: 'Bespoke' Embedding Spaces

Map the autoencoder's latent space to some new metric space, for applications:

- Mixing:

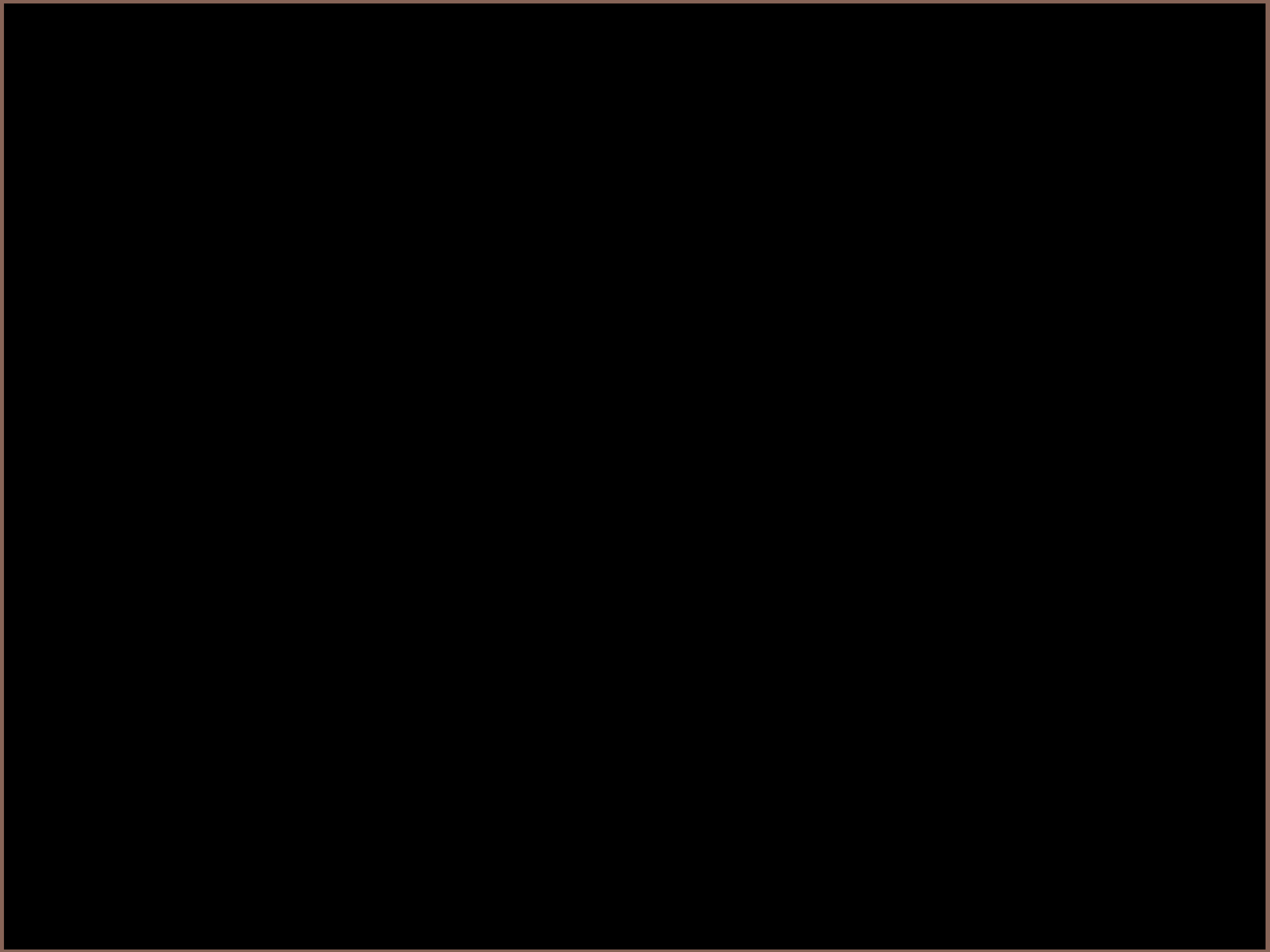
"mix of embeddings" = "embedding of mix"



Instead of Contrastive Loss on z's, use VICReg (Bardes, Ponce & Lecun, 2021)

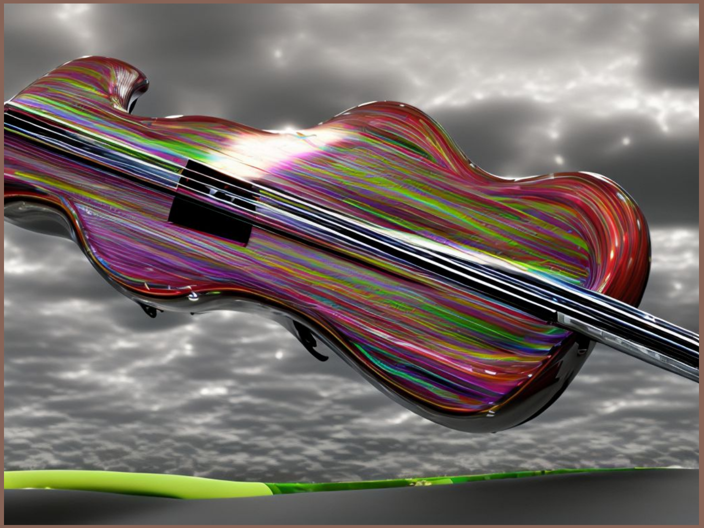
Idea: 'Bespoke' Embedding Spaces

By viewing the latents space as a real space



Closing Thoughts

It's a very exciting time! Moving toward expressive, intuitive, general musical audio production interfaces!

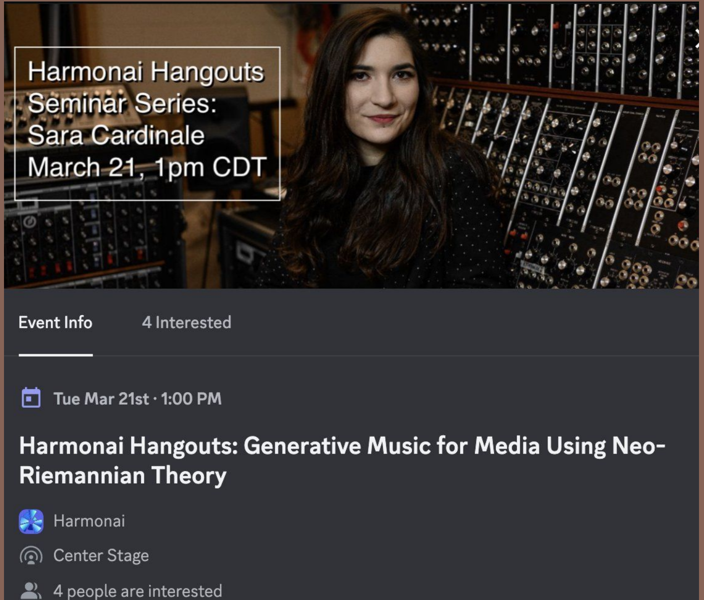


Check out Harmonai.org, e.g. our Discord servers

- Harmonai

- Harmonai R&D

- Talks most Tuesdays at 1pm.

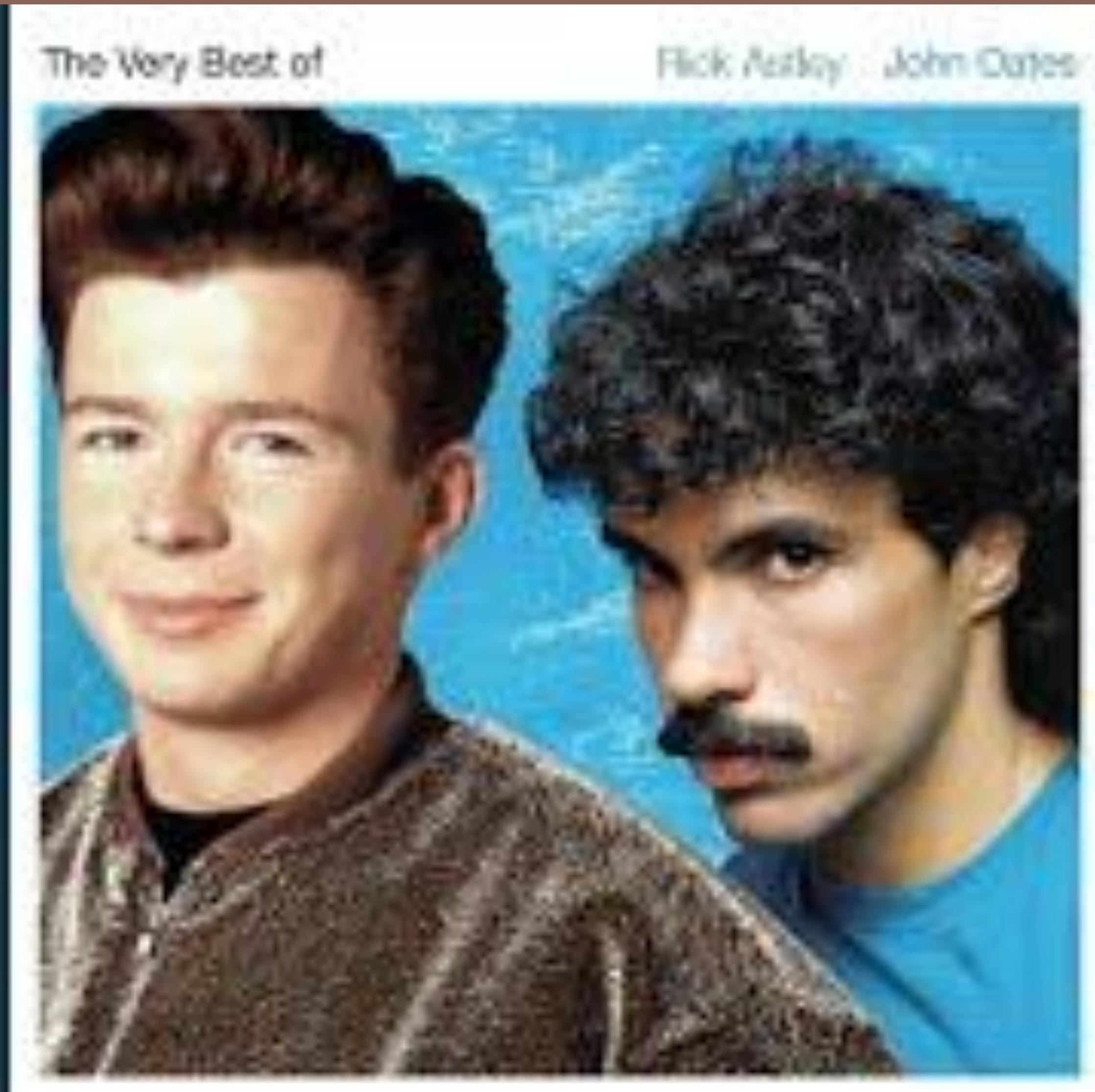


Extra unused slides follow…

(Audio effects by other models)

"Reverberation" by Stable Diffusion XL

And DJ "DataLoader" is a fake!

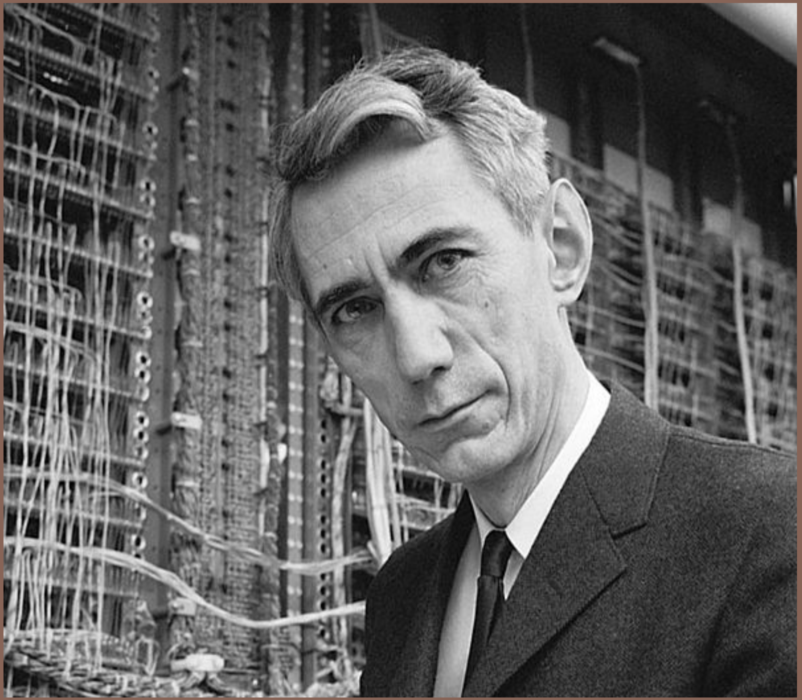


How Much Information to Encode, to Whom?

Claude Shannon: "Information" as a notion of "surprise", i.e., that which could not be guessed

High information:

- General Inputs: Encoding any possible range of inputs for systems that assume nothing about each other
- Details of "noise" (instead of throwing them out)



If your encoder & decoder have a lot in common, you can send minimal information

e.g.: Google's Lyra streaming speech codec (2021): model of human voice

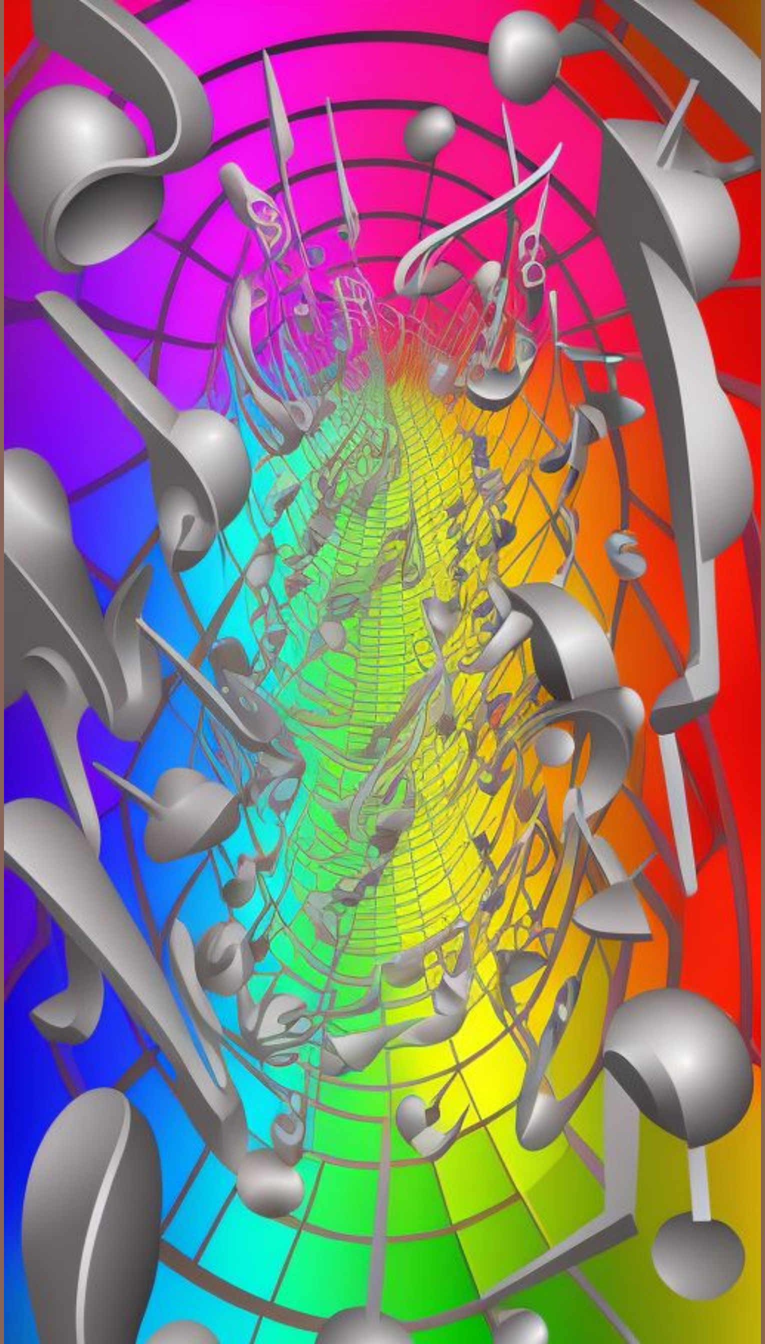
Aside: Consider Physical Tone Generation

"Shaping Noise": Broadband excitation, filtered by resonances in the instrument

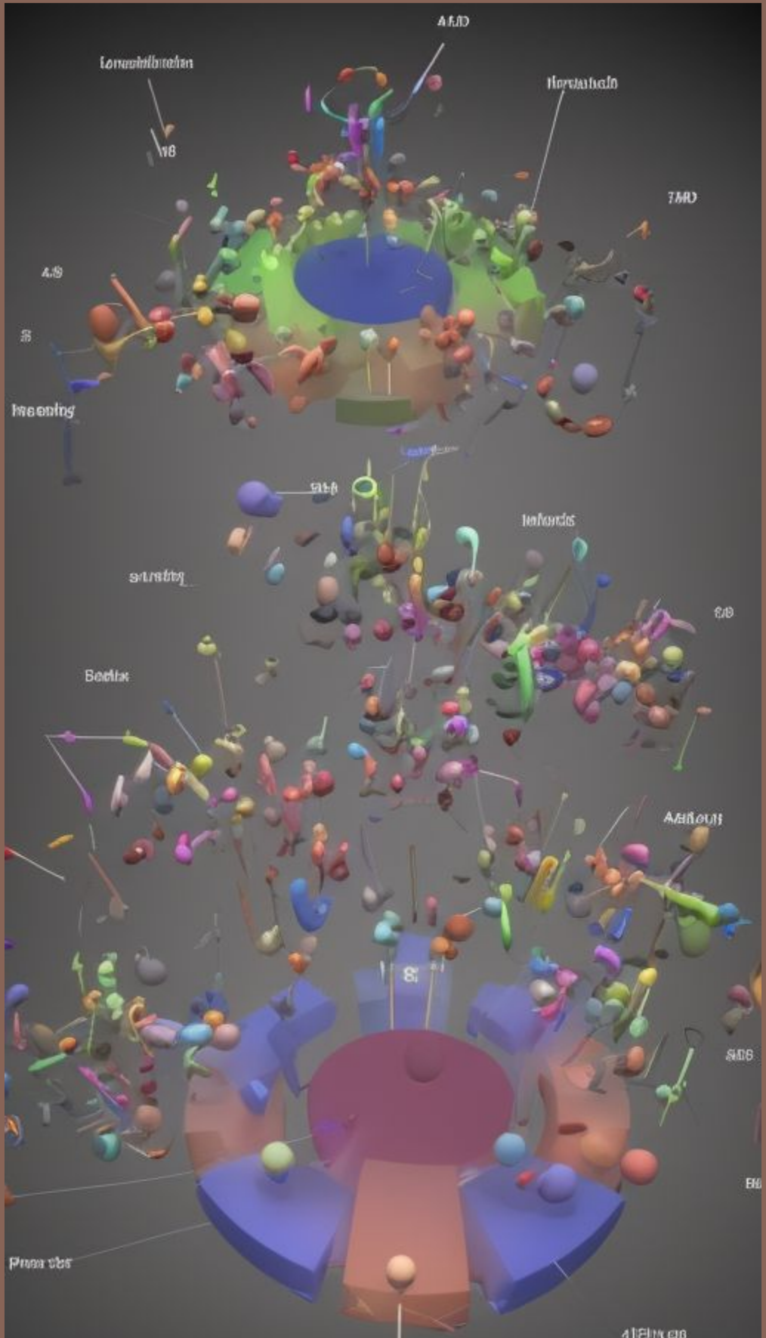
Woodwinds & brass: breath / reed
Guitars, piano: "percussive" impulse

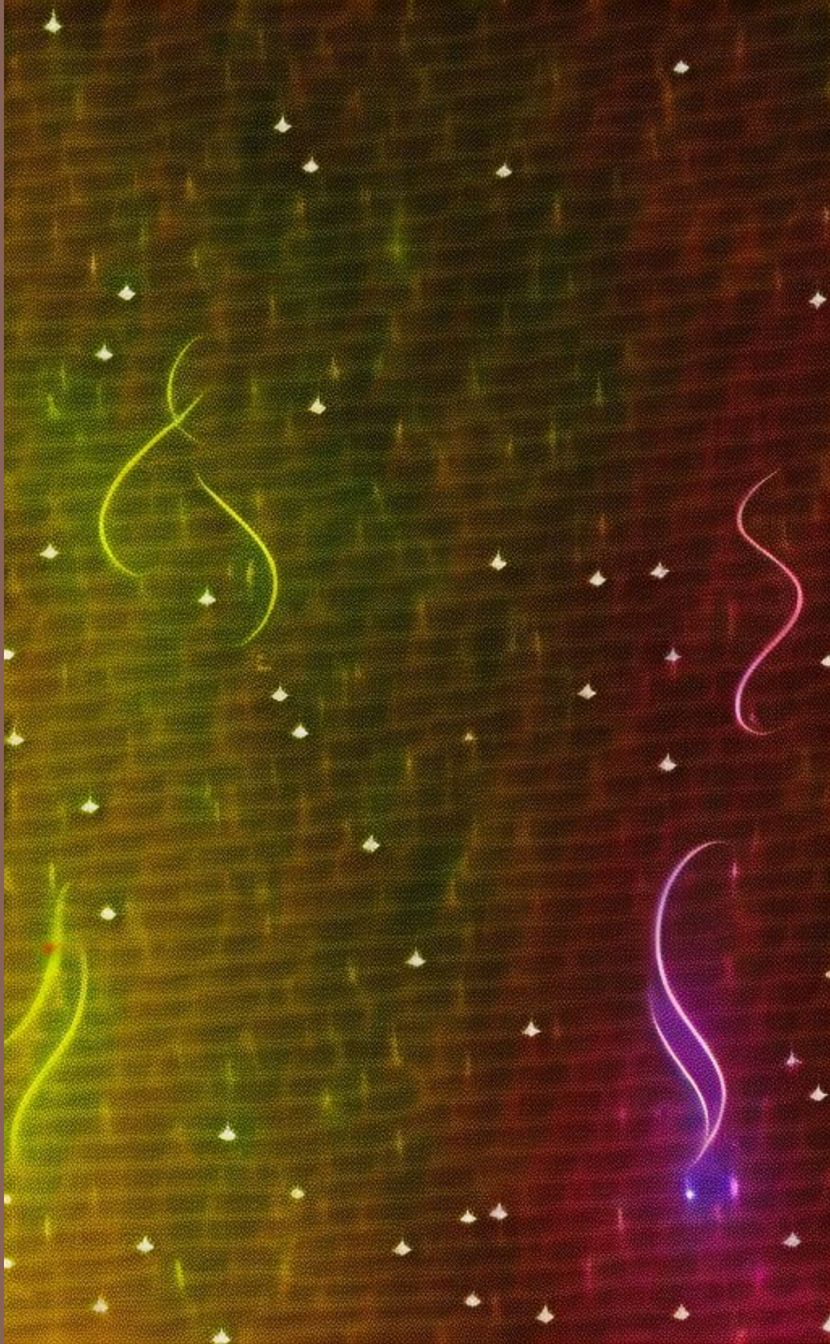
Phase is often random or immaterial:
Similar but different noise / impulse may sound indistinguishable

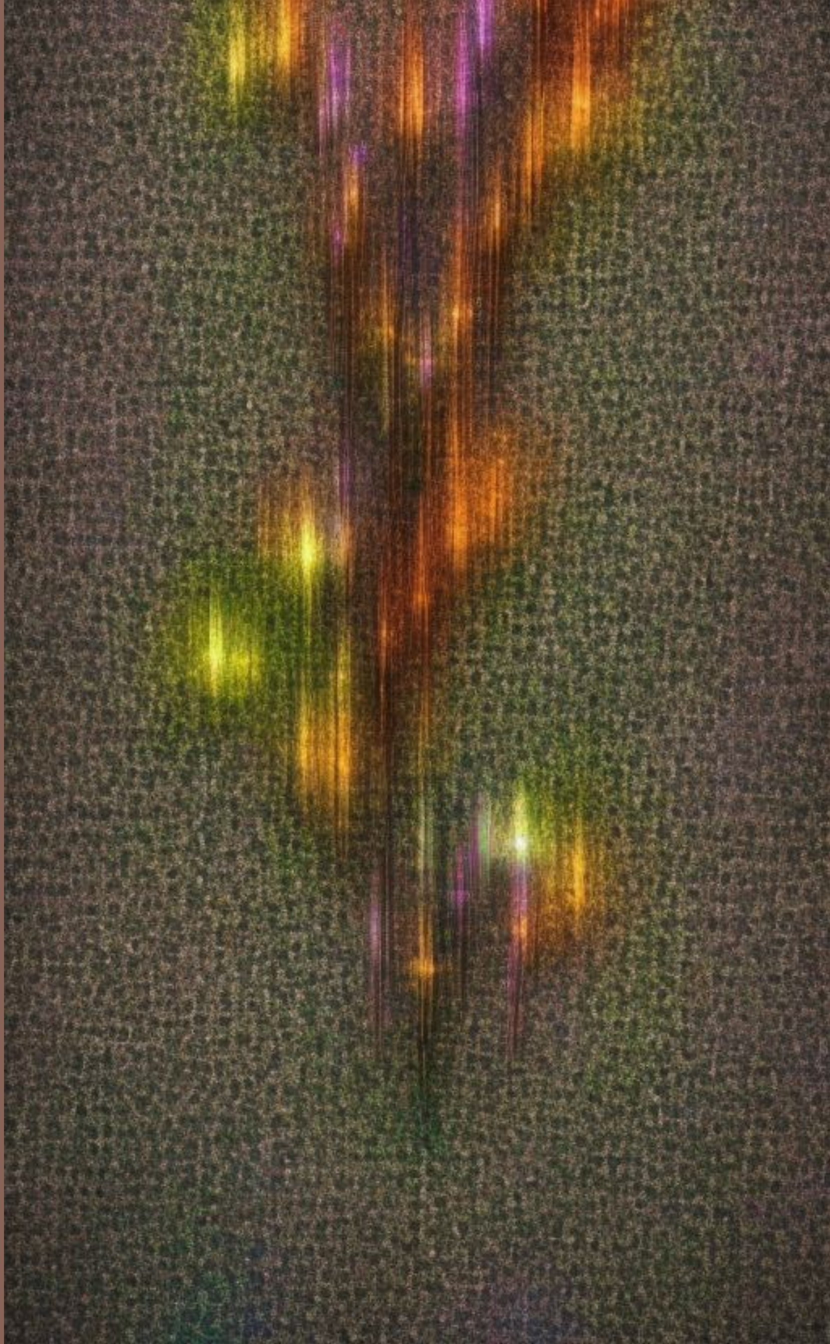
More
SD
Artwork

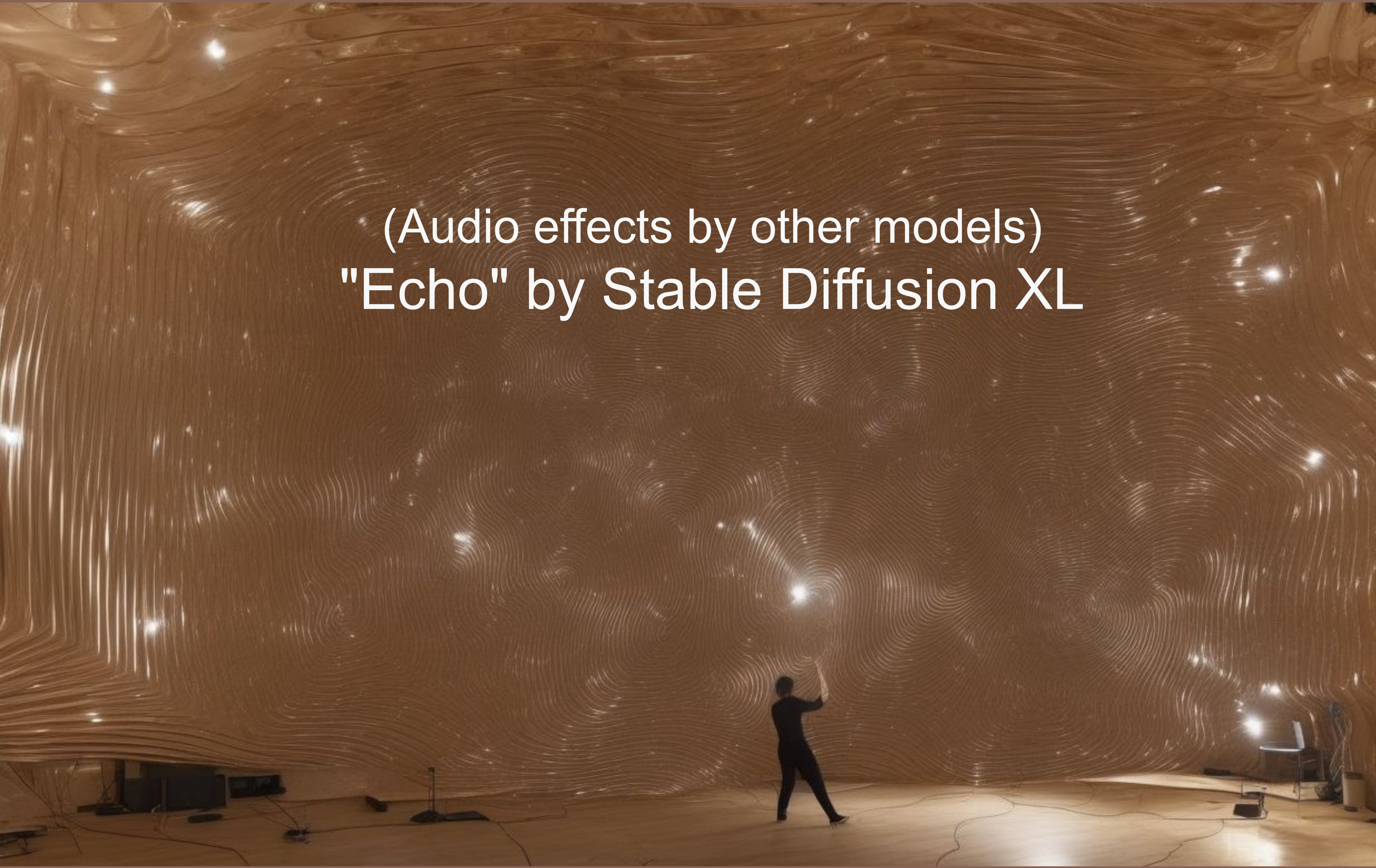



"Ricci Scalar" by Stable Diffusion 2.0








(Audio effects by other models)
"Echo" by Stable Diffusion XL