

Variational Autoencoder (VAE)

Deep-people #10

音楽生成 is difficult (言い訳タイム)

2

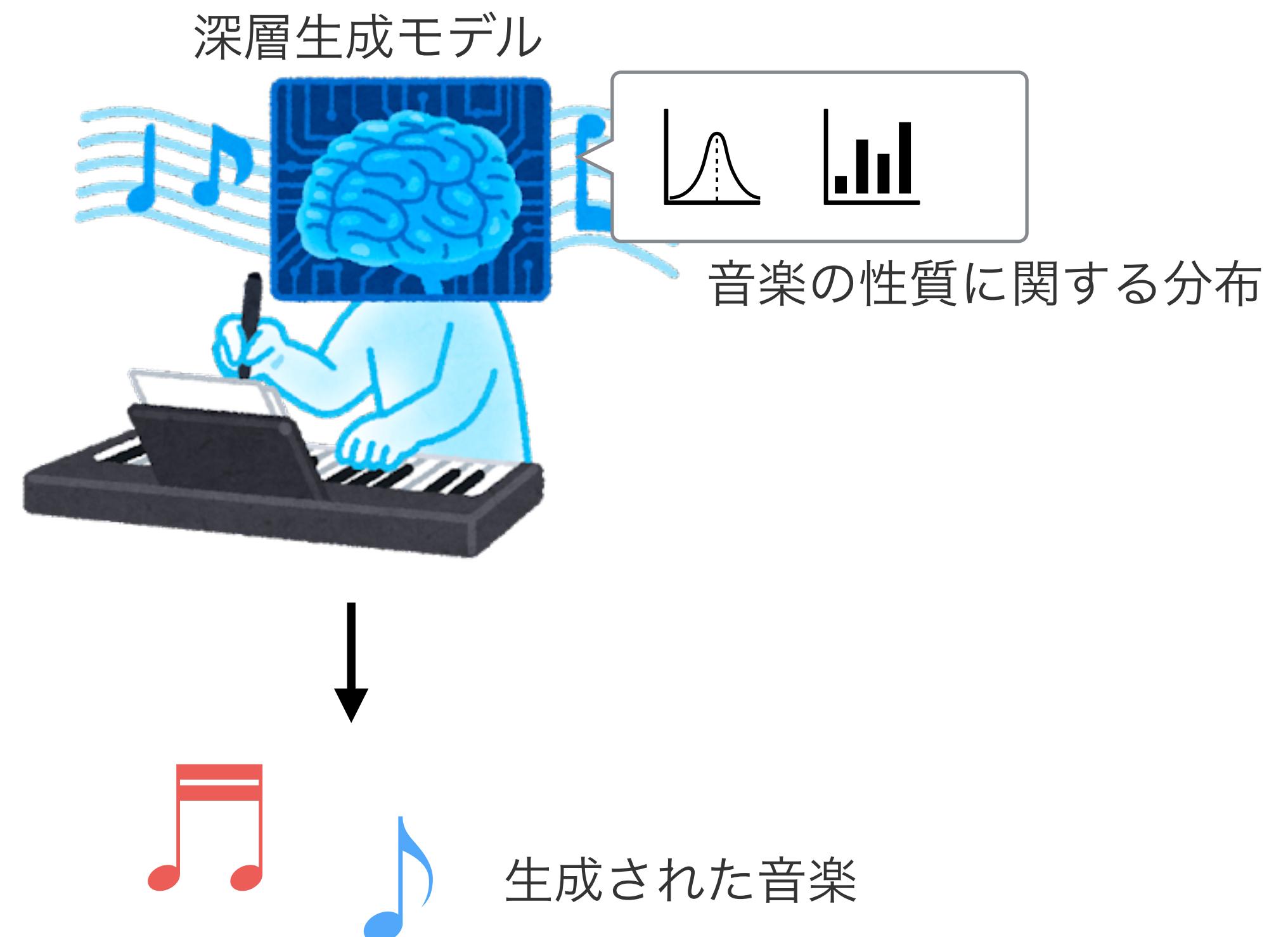


- 今回の実習は、**画像のVAE**です
- 音楽生成VAEはあまりにも初心者殺しのハマリポイントが多かったので、今回一気に説明すると爆発します（し、実装も標準化されてなくてどうしようか決められませんでした）
- やるとしたら別途（おそらく秋ABはじめ or 夏季休業中のどこか）に集中講義としてやる、くらいでしょうか
- 度重なるやるやる詐欺で申し訳ございません

深層生成モデル

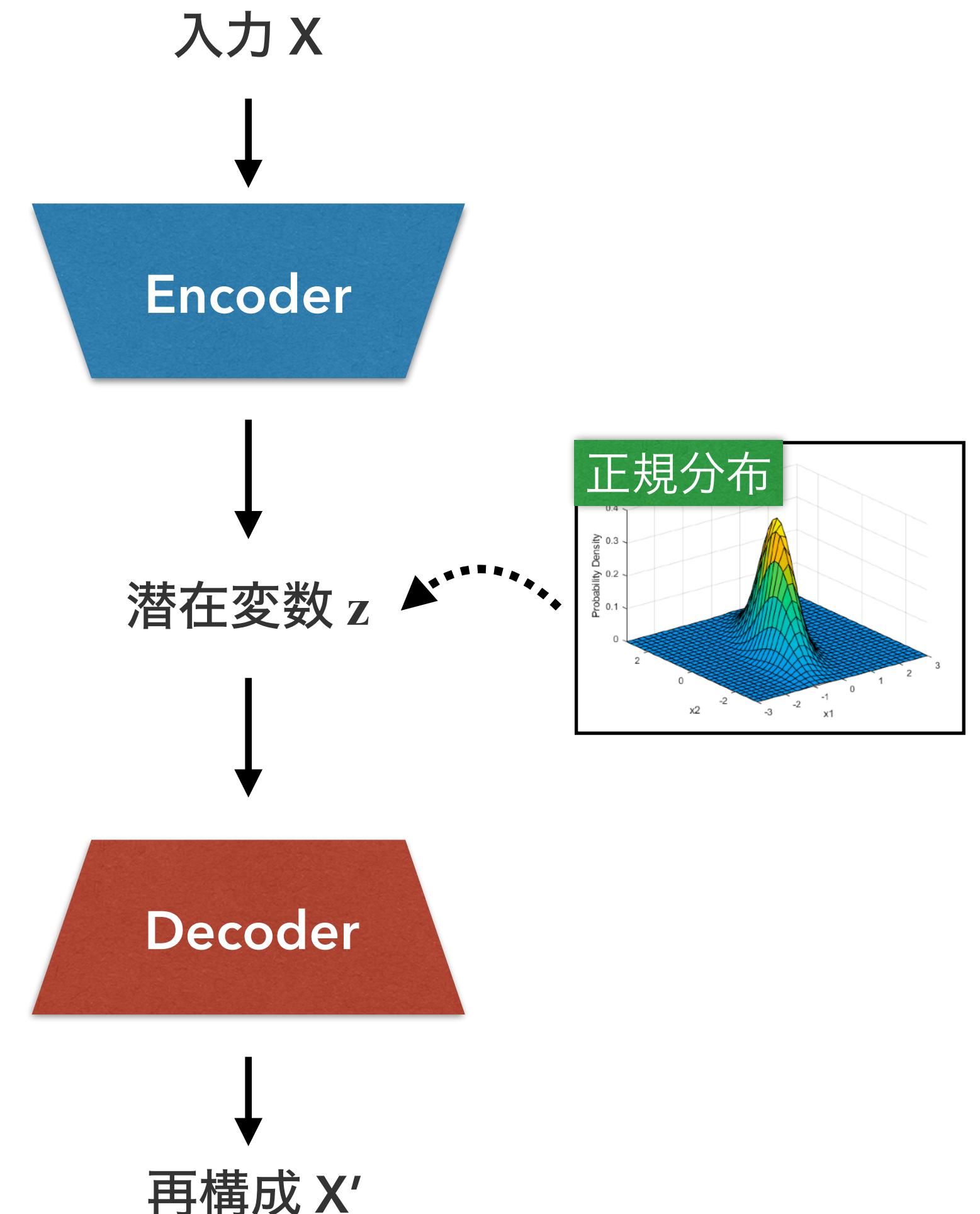
深層学習によってデータの生成を行うモデル

- データに関する性質をモデリングし、それに基づき新しいデータを作りだす
- VAE, GAN, 自己回帰モデル, Flow, Diffusionモデル等



変分自己符号化器 ; Variational auto encoder (VAE)

- ベイズ推論（変分ベイズ）の考えを取り入れたオートエンコーダ
- 潜在変数 z の分布に正規分布を仮定し、生成物同士の補間を可能にした
- 現在、最先端の研究でも広く用いられる



VAEが担うべき以下の2つを最適化する

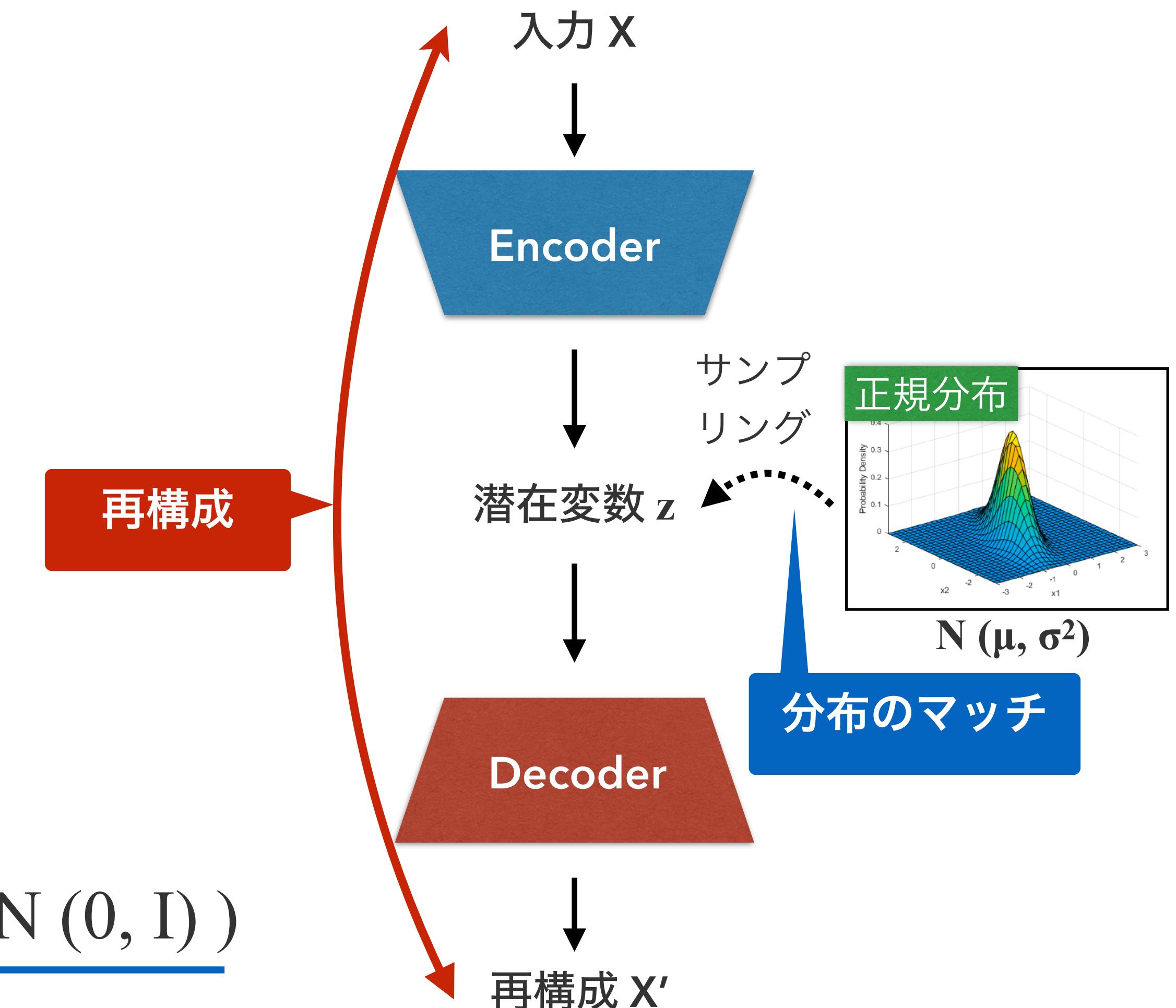
1. VAEの入力を再構成すること
2. VAEによって入力から得た潜在変数が,

事前分布にしたがうこと

(この詳細は補足スライドを参照)

損失関数

$$L(W; x) = \underline{\|x - x'\|^2} + \underline{KLD (N(\mu(x), \sigma(x)) \| N(0, I))}$$



2つの分布間の距離を表すもの

- 2つの分布 $p(x)$ と $q(x)$ に対して,

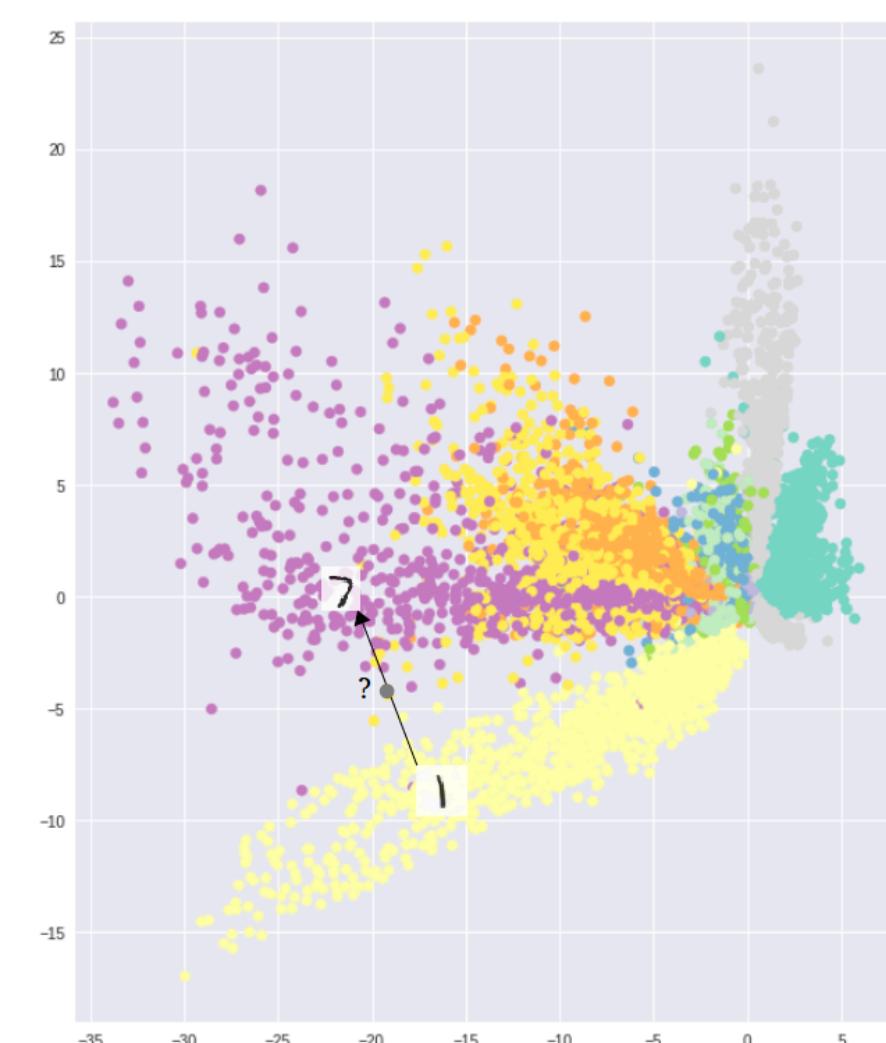
$$KLD(p(x) \parallel q(x)) \equiv \int p(x) \ln \frac{p(x)}{q(x)} dx = \mathbb{E}_{p(x)}(\ln p(x) - \ln q(x))$$

- 2つの分布の近さを表すと言われている（※距離の公理を満たしていないので厳密には距離ではない）
- 非対称であることに注意 : $KLD(p(x) \parallel q(x)) \neq KLD(q(x) \parallel p(x))$

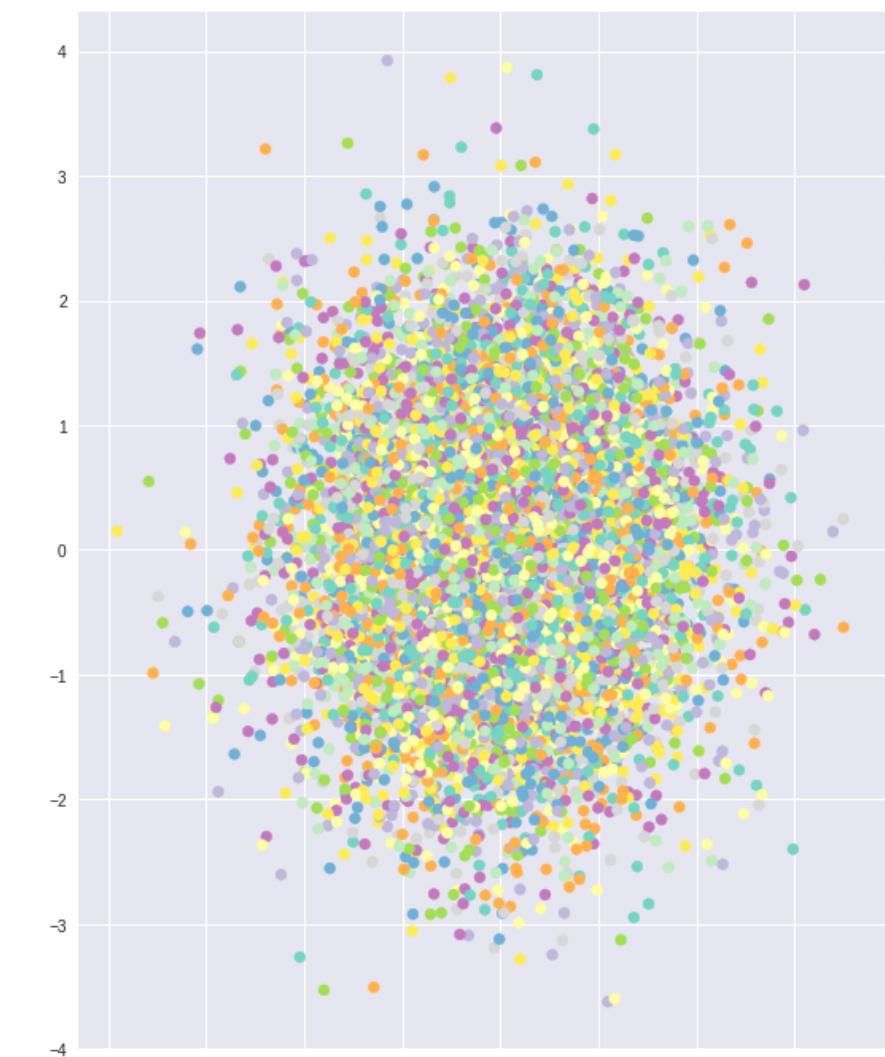
VAEの潜在変数空間

7

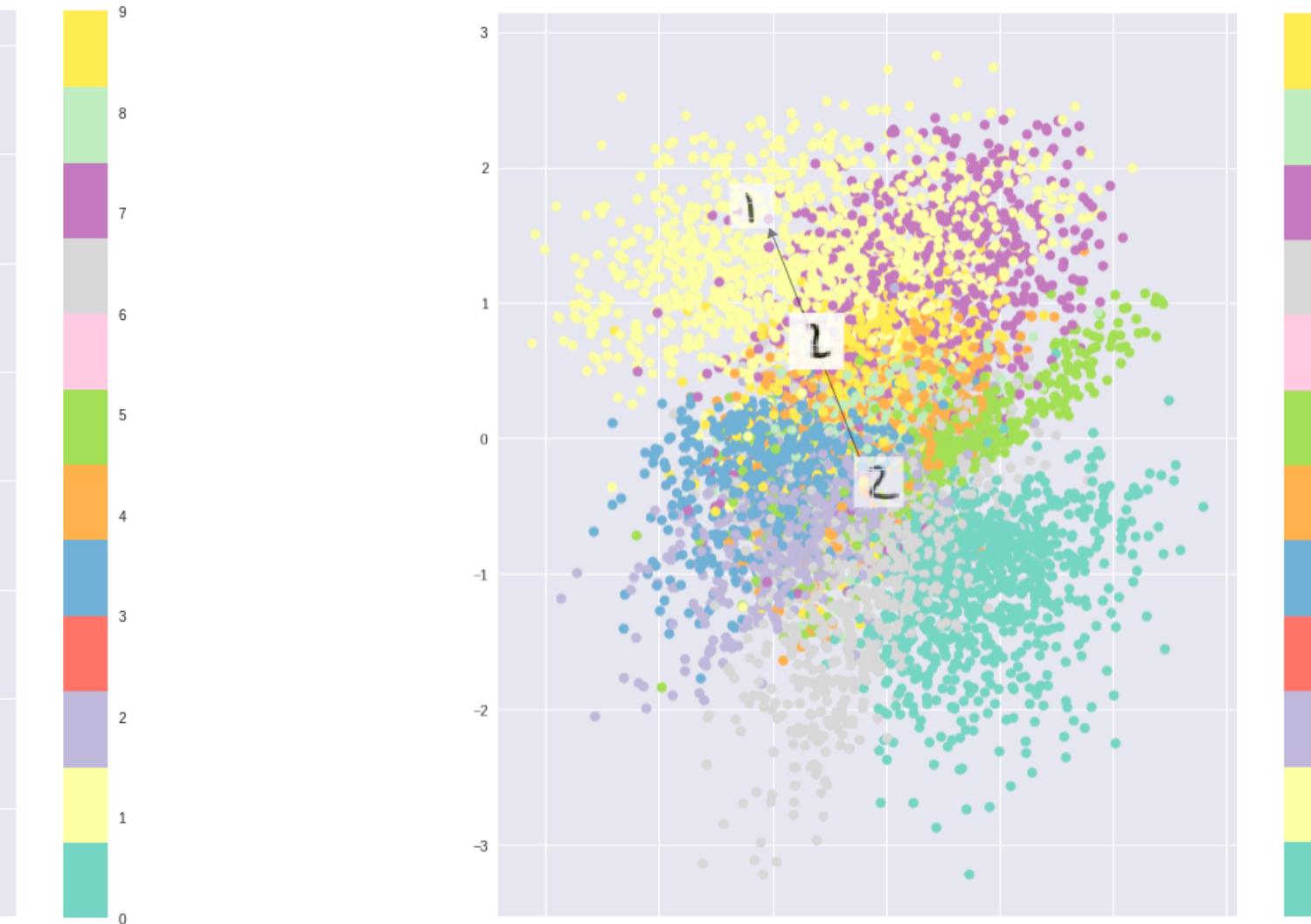
- 両方の制約によって、連續的かつラベルごとに別れたクラスタを形成する分布ができる



再構成のみ



KLDのみ



両方

Re-parameterization trick

Encoderによって得た分布の平均に、分布の分散に従う

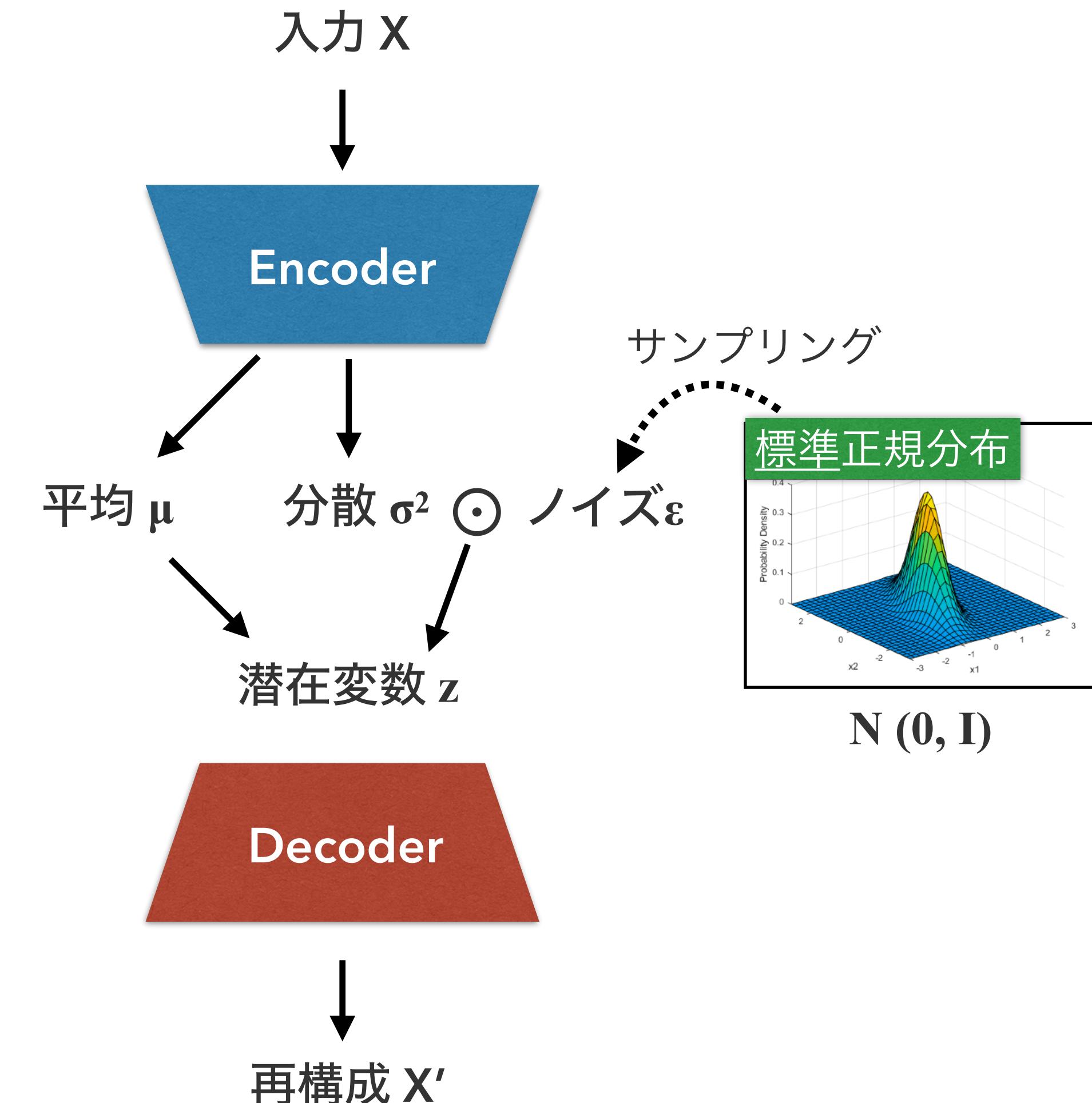
ランダムノイズを重畳する

-> サンプリングが逆伝播計算不可能なために

Encoderに勾配が伝わらない問題を、これにより解決

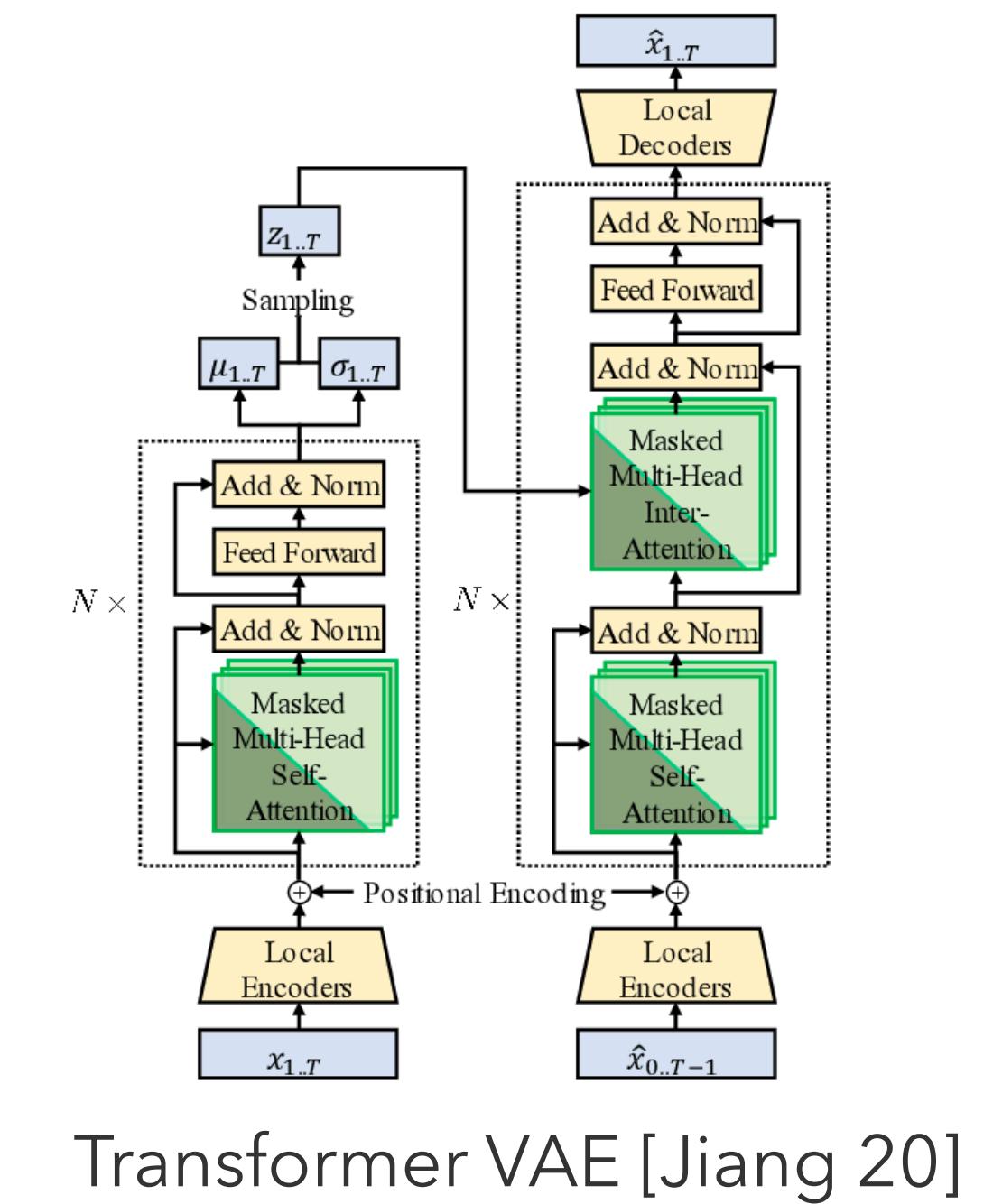
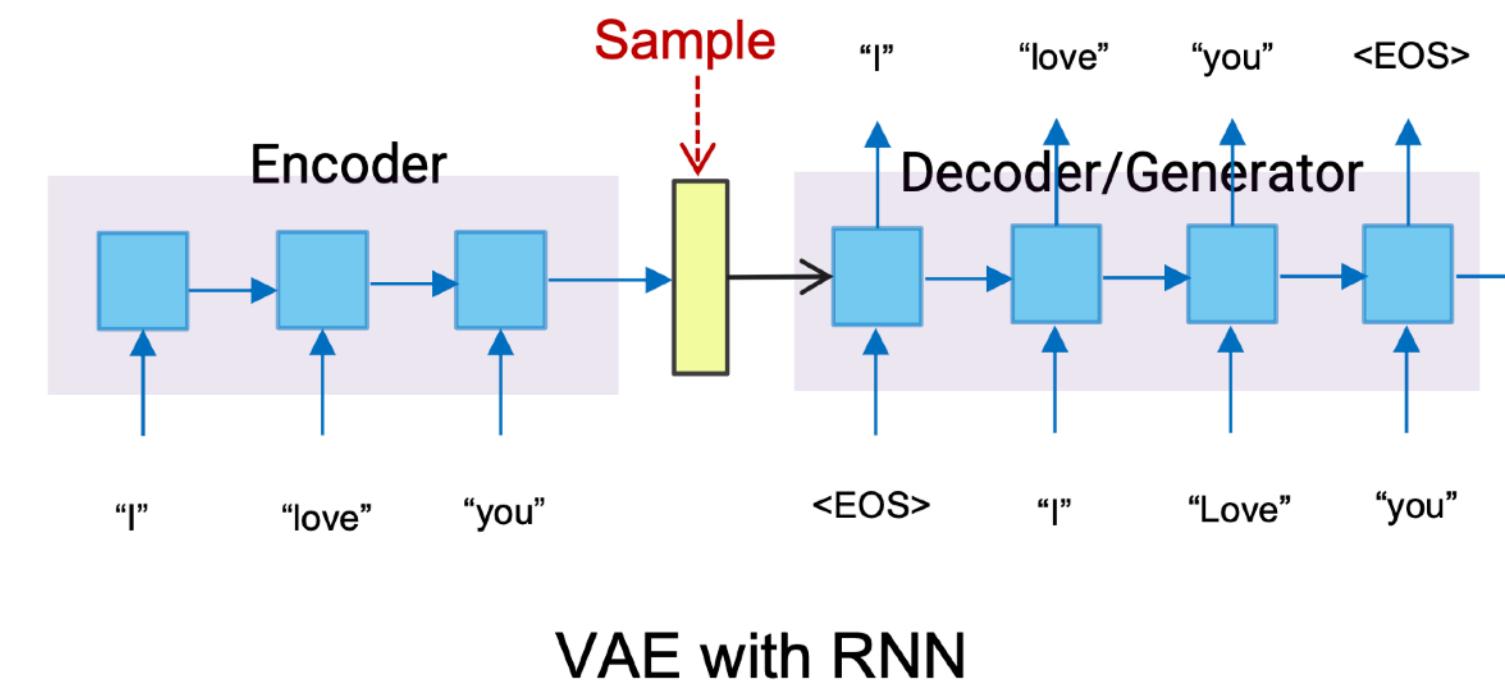
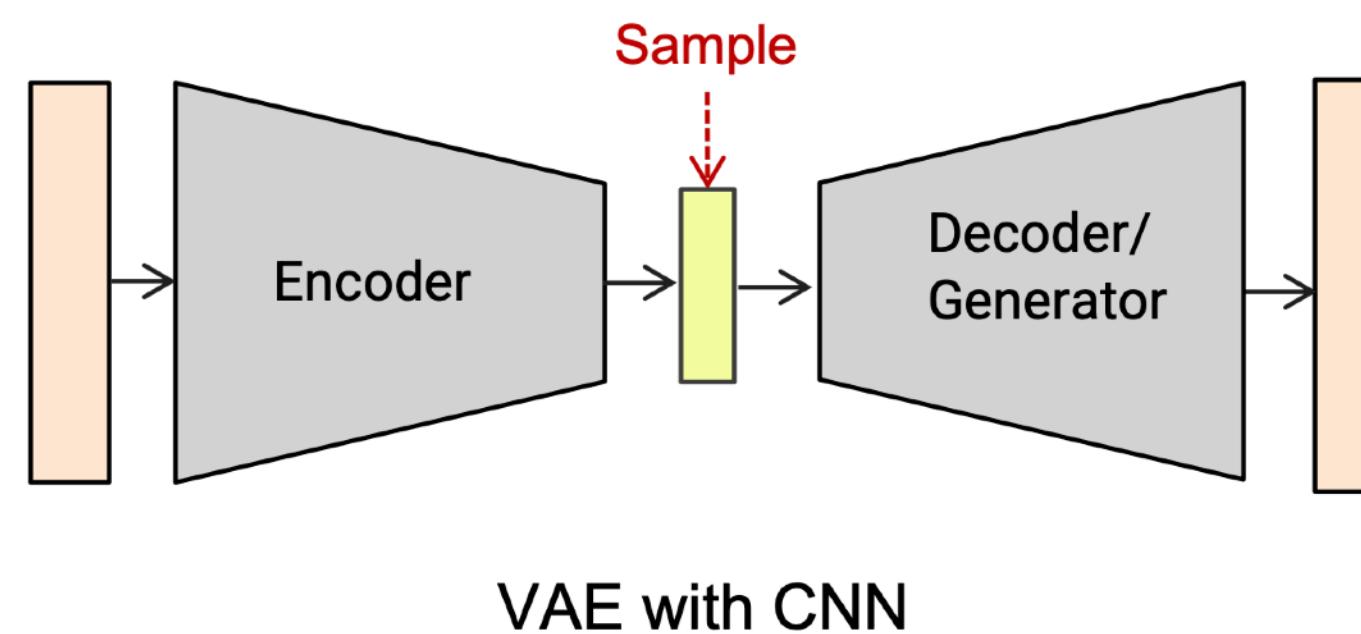
$$z = \mu(x) + \epsilon \odot \Sigma(x)$$

$$\epsilon \sim \mathcal{N}(\mathbf{0}, I)$$



EncoderもDecoderも、 パーツが何かとは独立

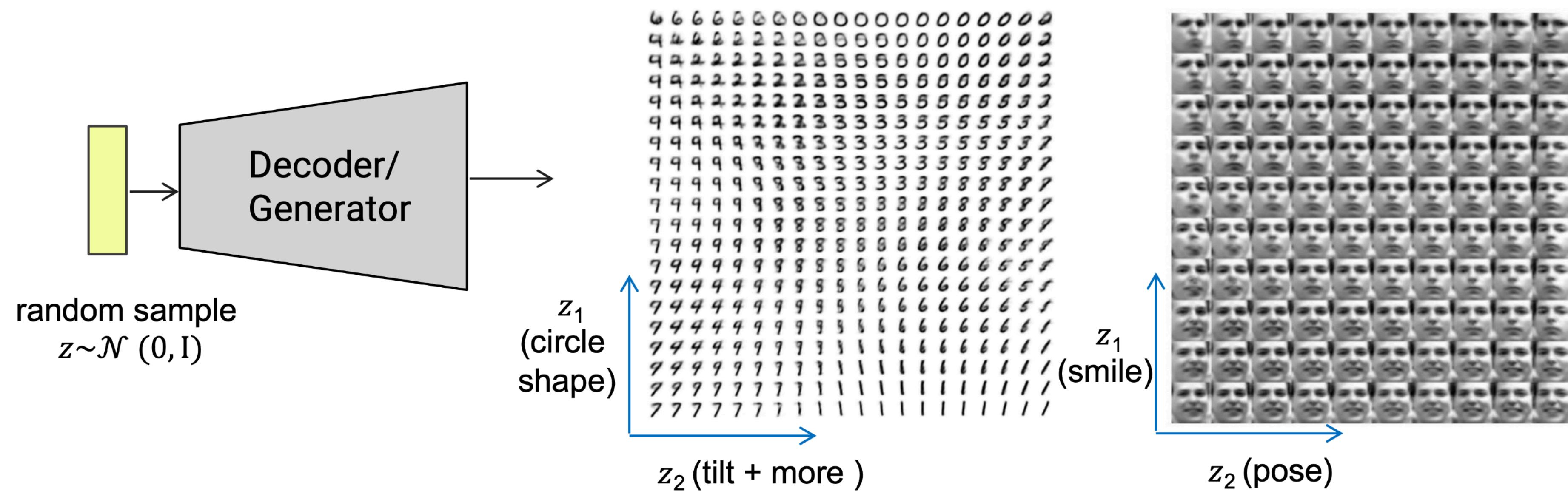
- 全結合層, CNN, RNN, Transformer etc...
- 対象によって適切なパートを選ぼう



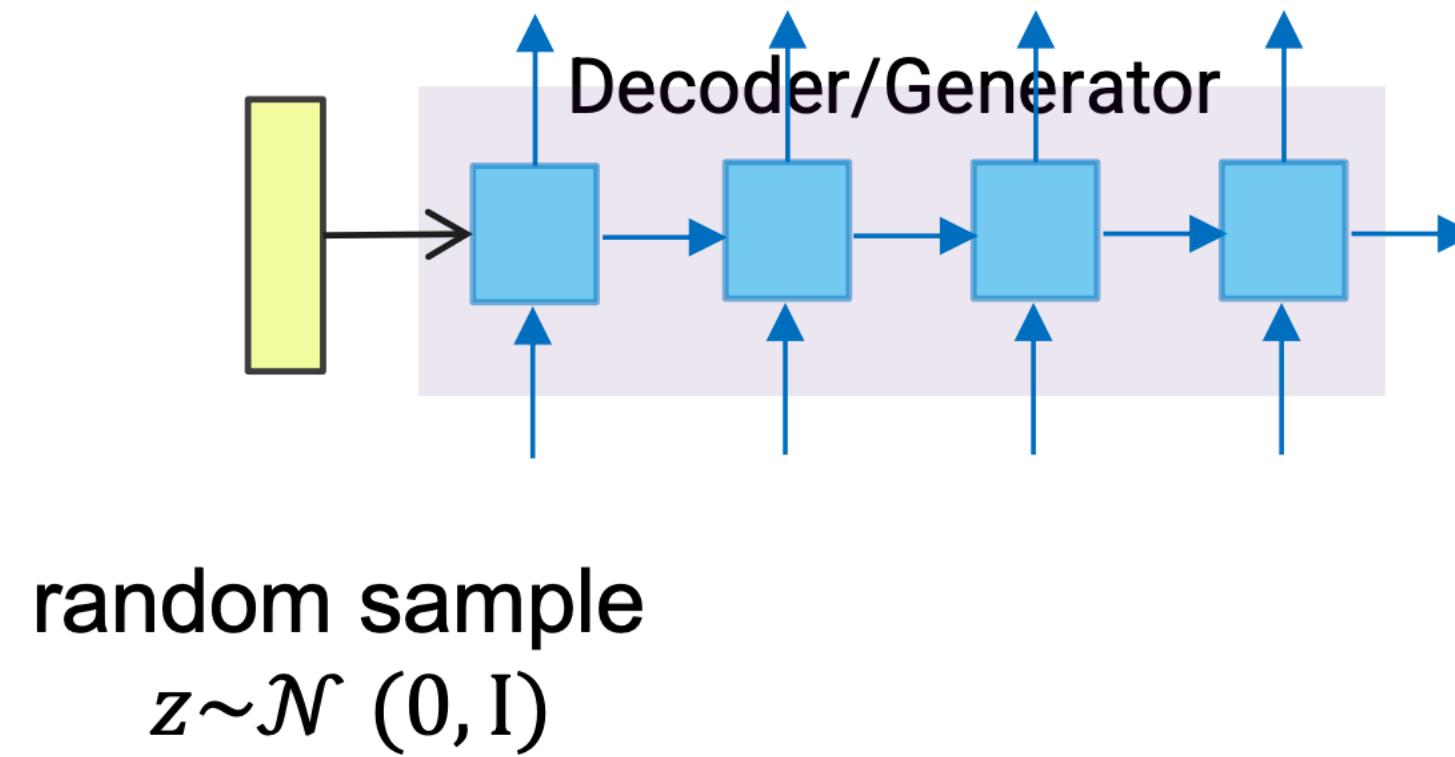
VAEの性質：多様体学習

10

- 潜在空間内でデータが集中している部分は局所的に連続している
 - 多様体学習：低次元空間にうまく展開するよう学習を行う
 - ある次元の値を動かすと、意味的に変化することが期待できる



- 自然言語でも同様の学習が可能



no .
he said .
“ no , ” he said .
“ no , ” i said .
“ i know , ” she said .
“ thank you , ” she said .
“ come with me , ” she said .
“ talk to me , ” she said .
“ do n’t worry about it , ” she said .

this was the only way .
it was the only way .
it was her turn to blink .
it was hard to tell .
it was time to move on .
he had to do it again .
they all looked at each other .
they all turned to look back .
they both turned to face him .
they both turned and walked away .

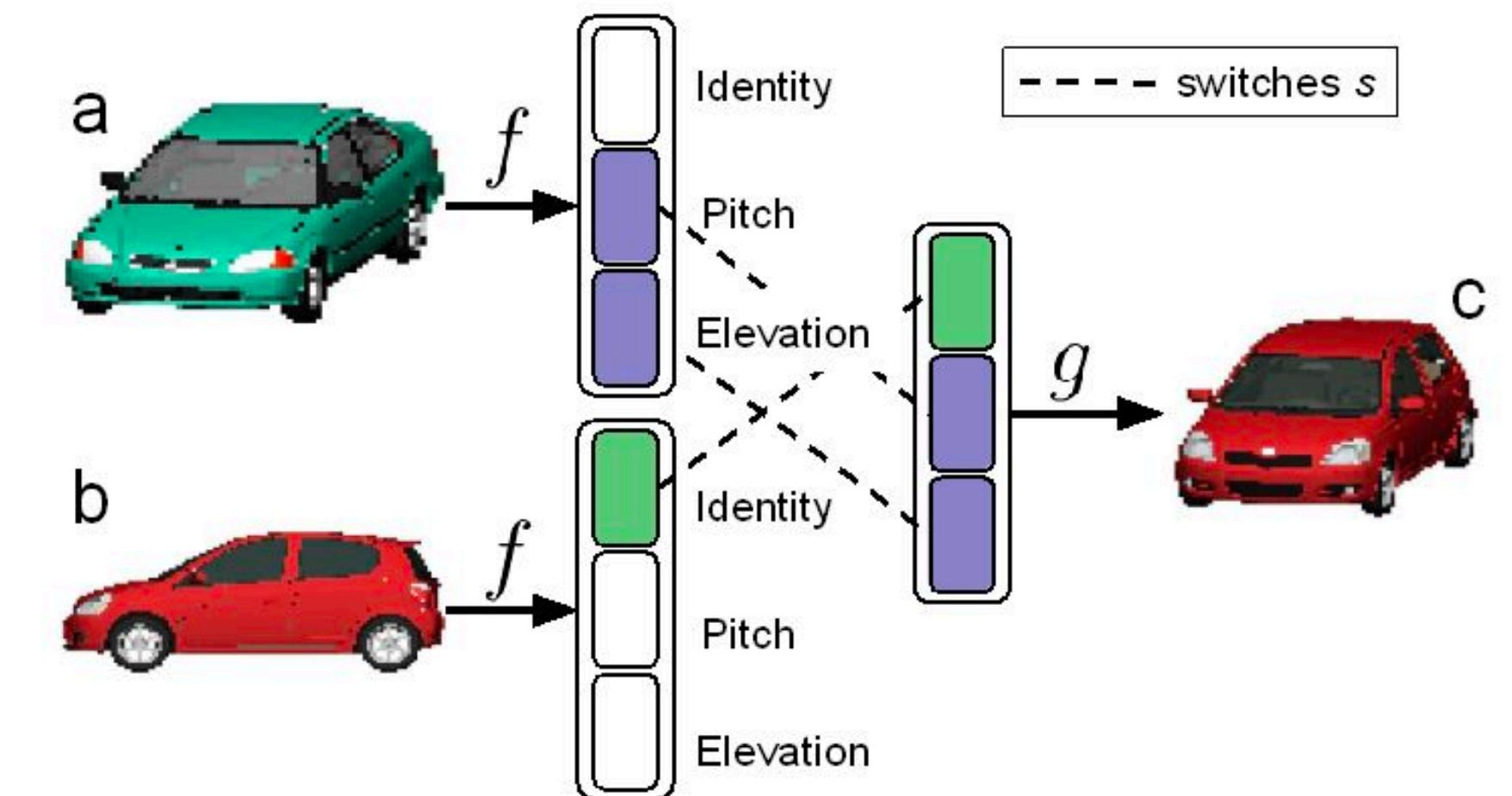
Interpolated sentences between pairs of random points in the latent space z

Disentangle: 特徴量のときほぐし

データは独立する複数の要因から生成されると仮定
 →データを構成する要素を分解する

- Ex.
- 顔画像 -> (肌色, 顔の向き, 表情, 年齢 etc.)
- 車画像 -> (形状, 色, 向き etc.)
- 楽器音 -> (音色, ピッチ, 奏法 etc.)
- 声 -> (話者, 発話内容, 感情 etc.)

Learning a disentangled representation

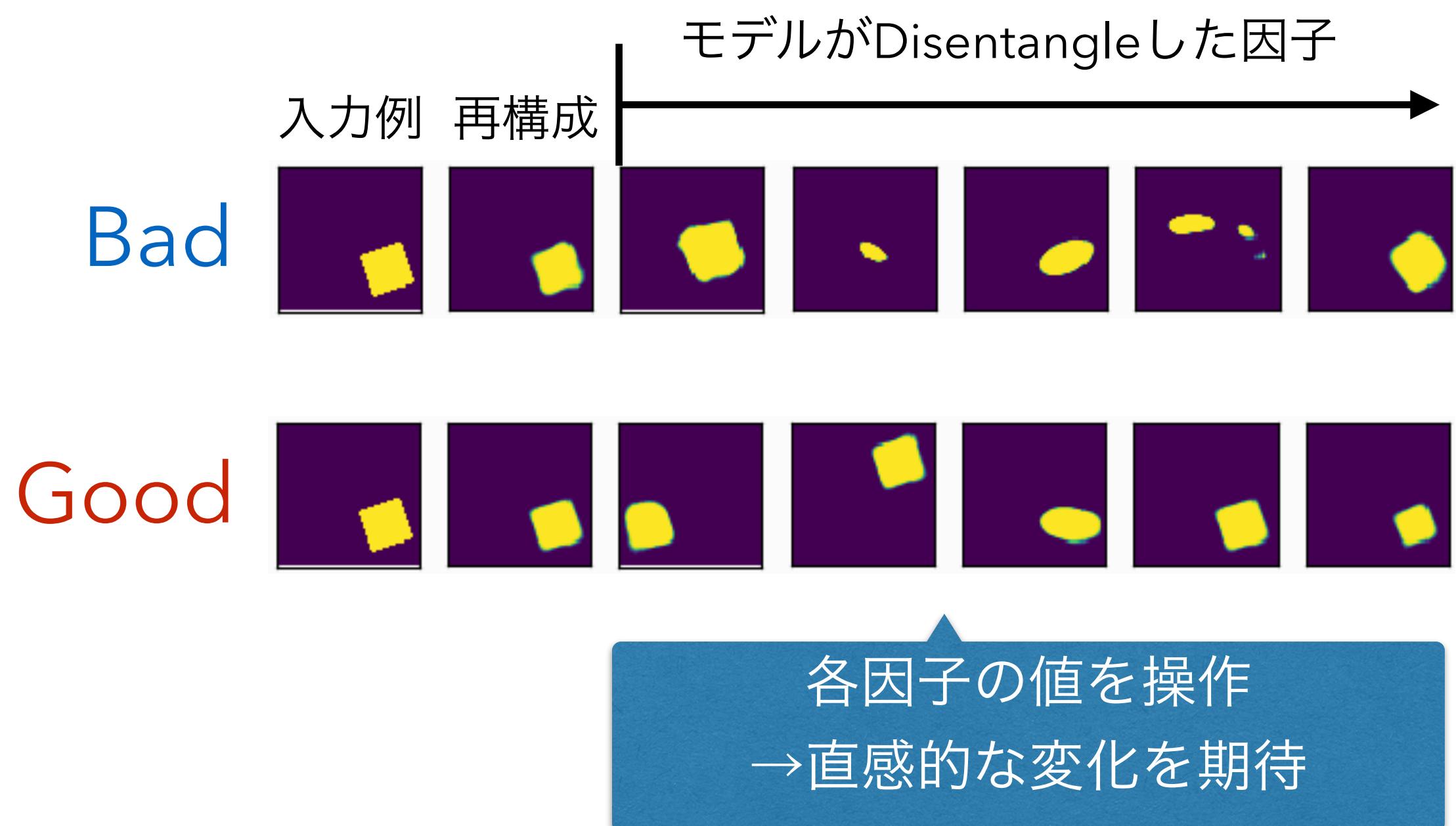


Reed, Scott E. et al. "Deep Visual Analogy-Making." NIPS (2015).

Ex. 図形画像のdisentangle

13

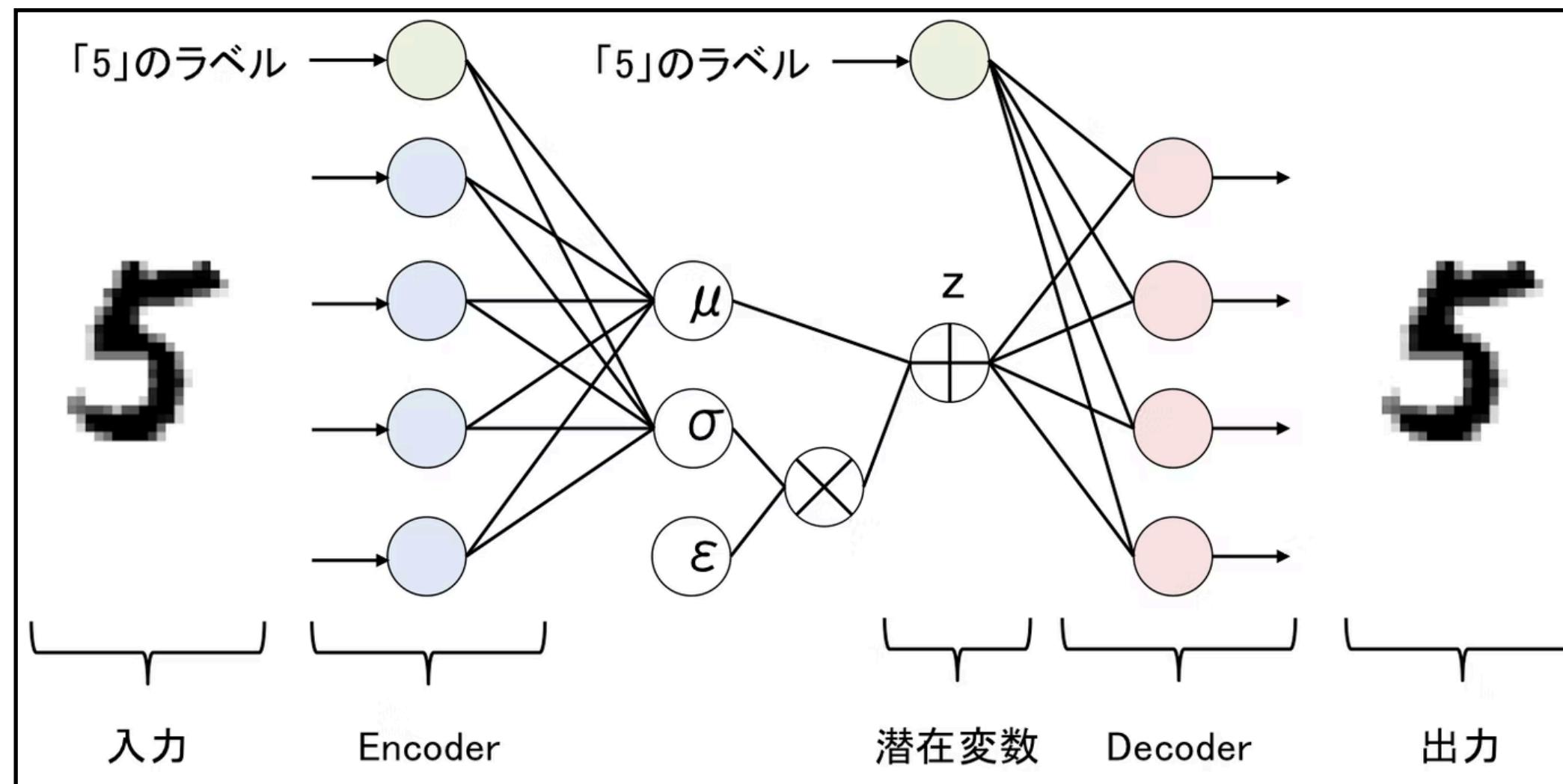
- ・ 図形がサイズ・位置を変えて配置された
画像のデータセットのDisentangle
- ・ モデルがデータの各要素
(縦位置・横位置・大きさ・角度) の
4つを独立に捉える必要がある



ラベルでの制御を行うVAE

14

- Conditional VAE (M2 model) によって、条件つきの生成が可能
- 基本的にベクトルの形にできるものであれば、なんでも指定可能



<https://qiita.com/kenchin110100/items/7ceb5b8e8b21c551d69a>

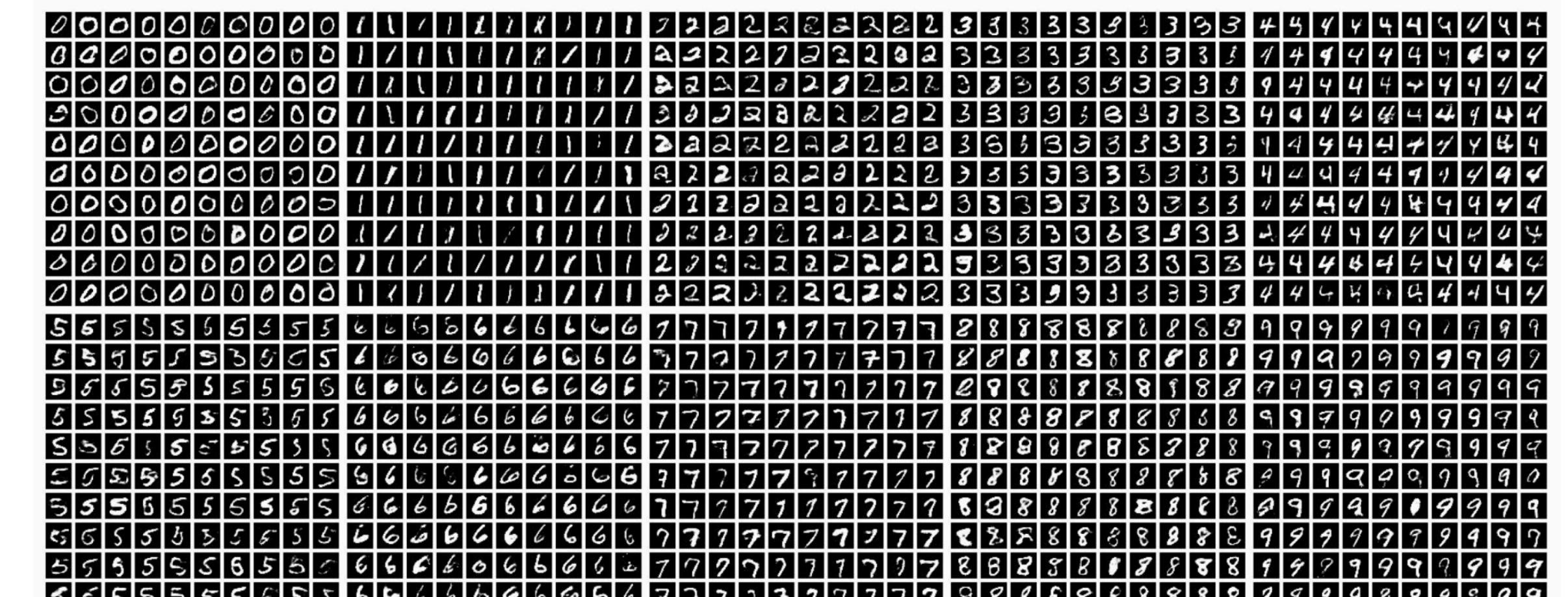


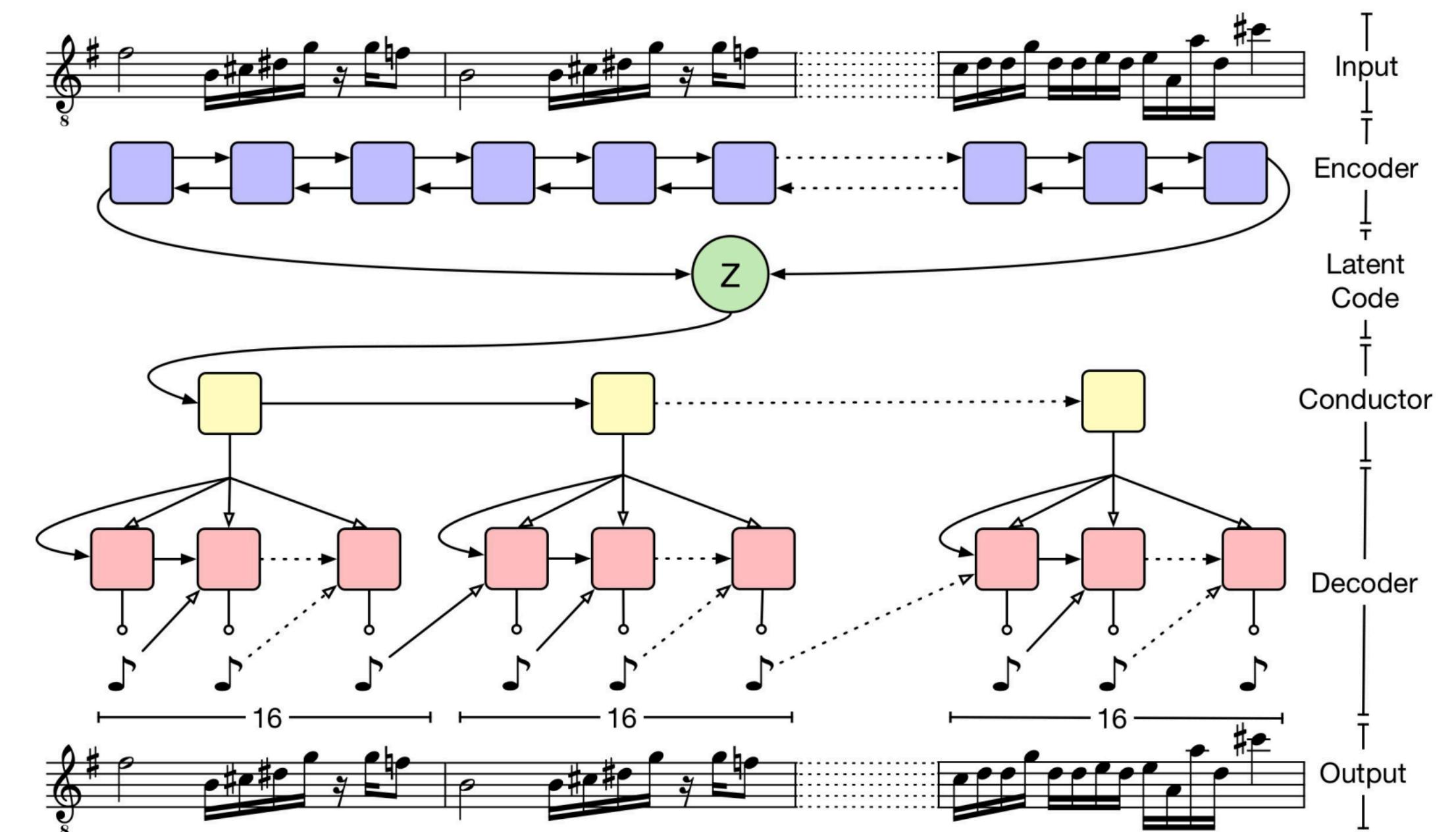
Figure 6: Conditional samples obtained by fixing the class label and varying z (for variant 3 with 3000 supervised examples)

MNISTデータセットの各ラベルで条件つけた生成。
筆跡のスタイル（数字と独立）が z によって表現
<https://pyro.ai/examples/ss-vae.html>

音楽での事例

メロディの生成を行うモデル

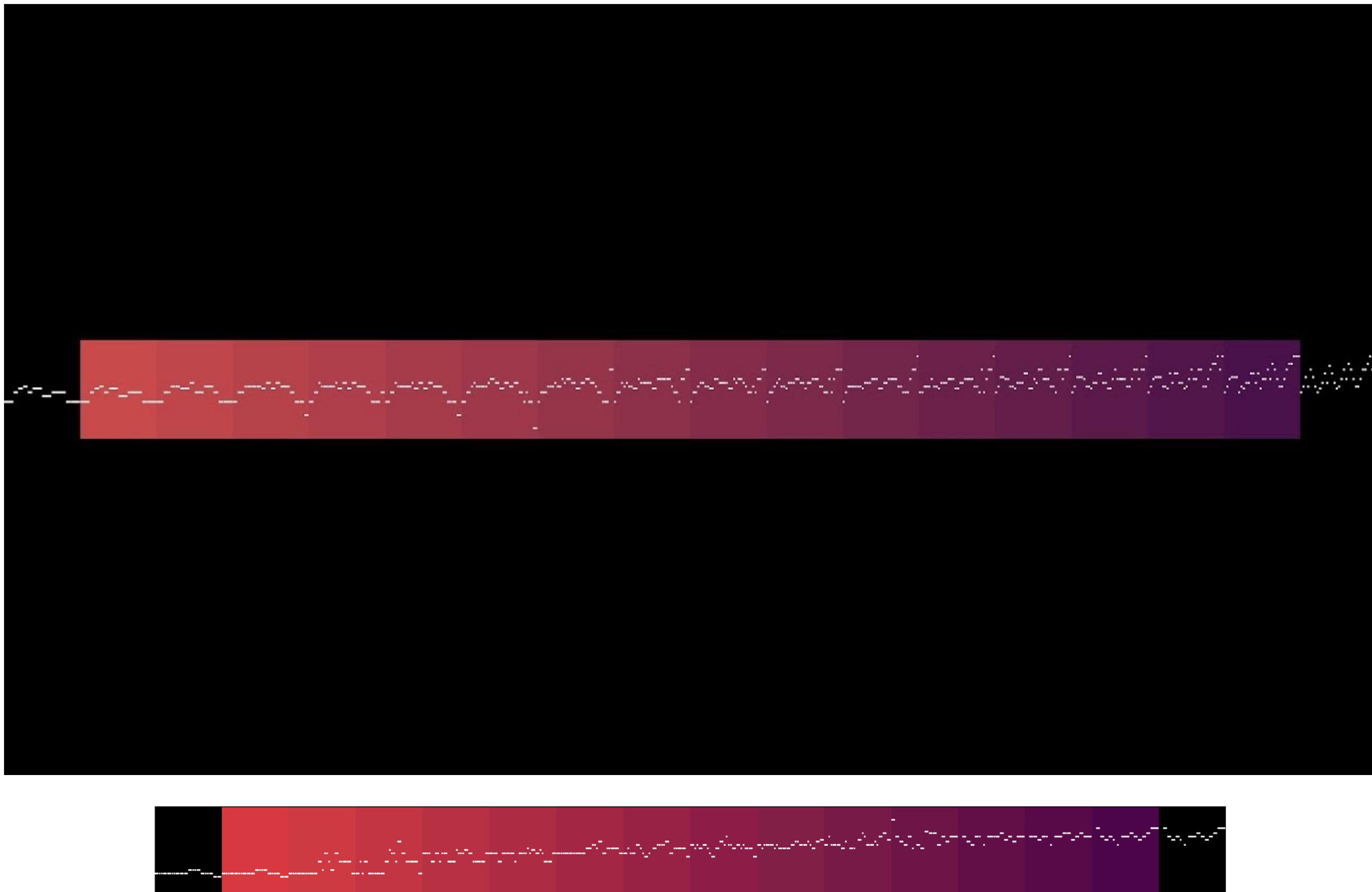
- エンコーダー : Bidirectional RNN
- デコーダー : 階層的に設計
 - **Conductor RNN**: z から, 小節単位の潜在変数を生成
 - **RNNLM**: Conductor RNNからの情報をもとに, 自己回帰的に音符を生成



MusicVAEが生成する音楽

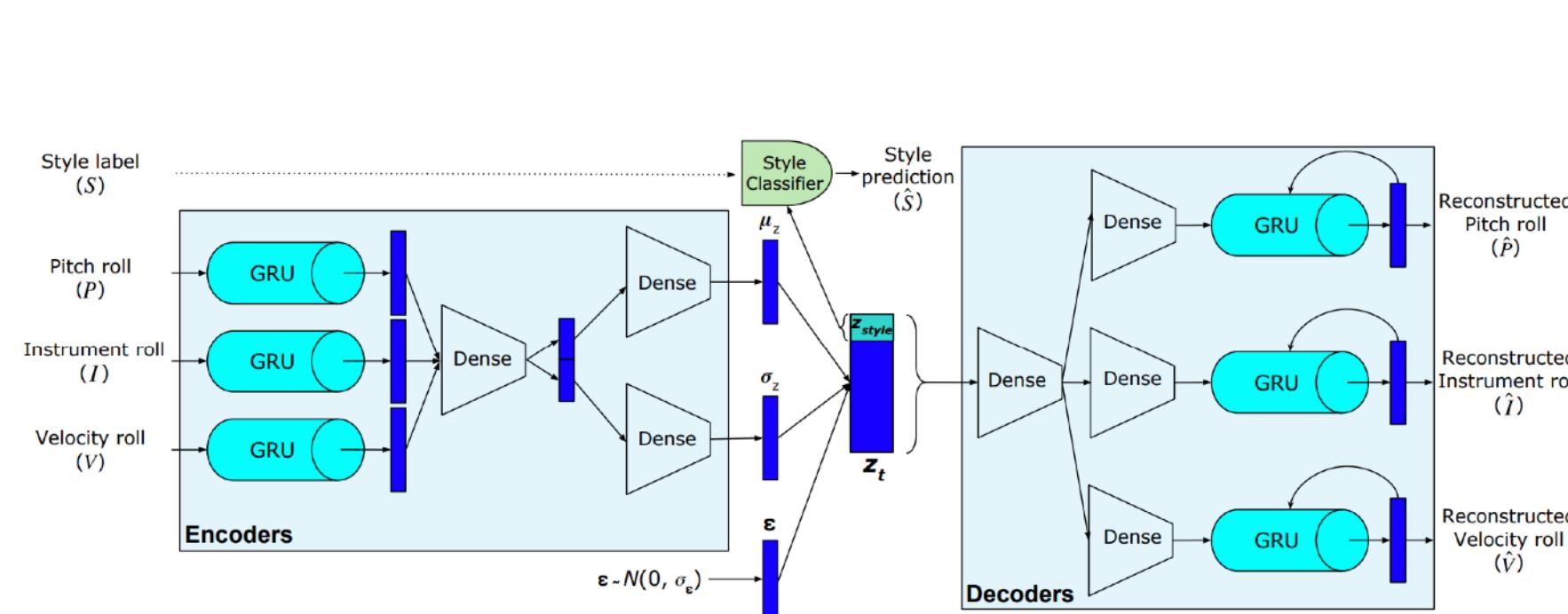
17

- 潜在変数の操作でメロディの補間が可能

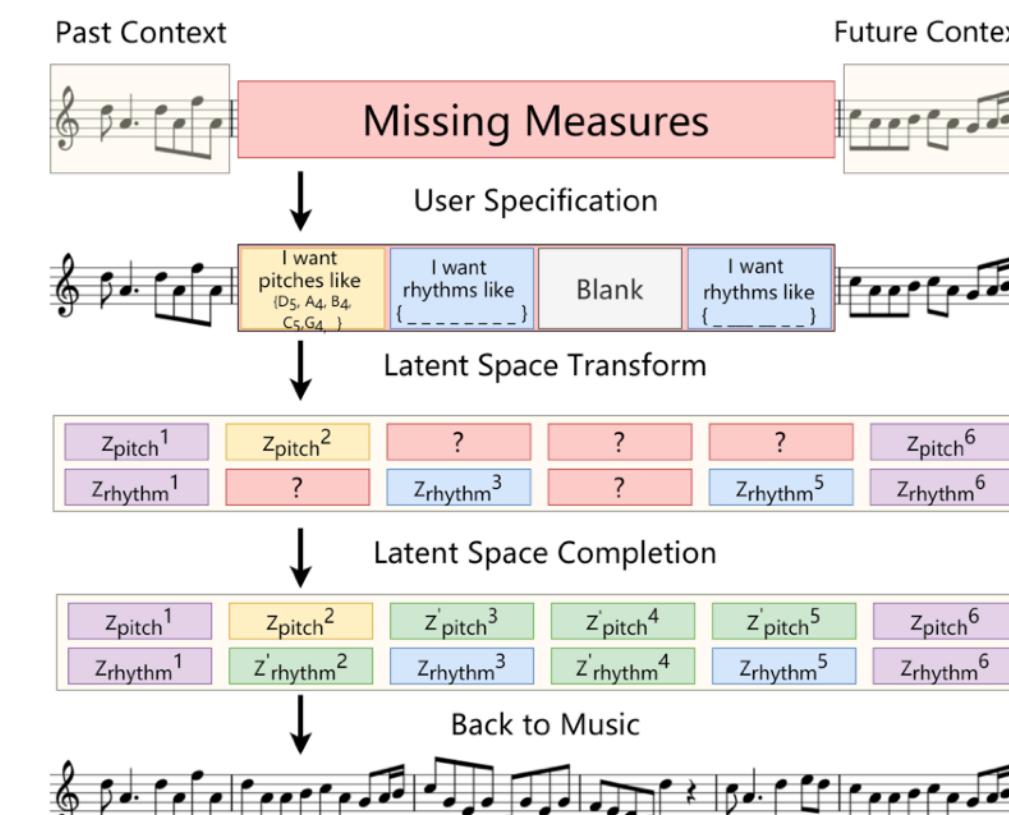


その他の事例

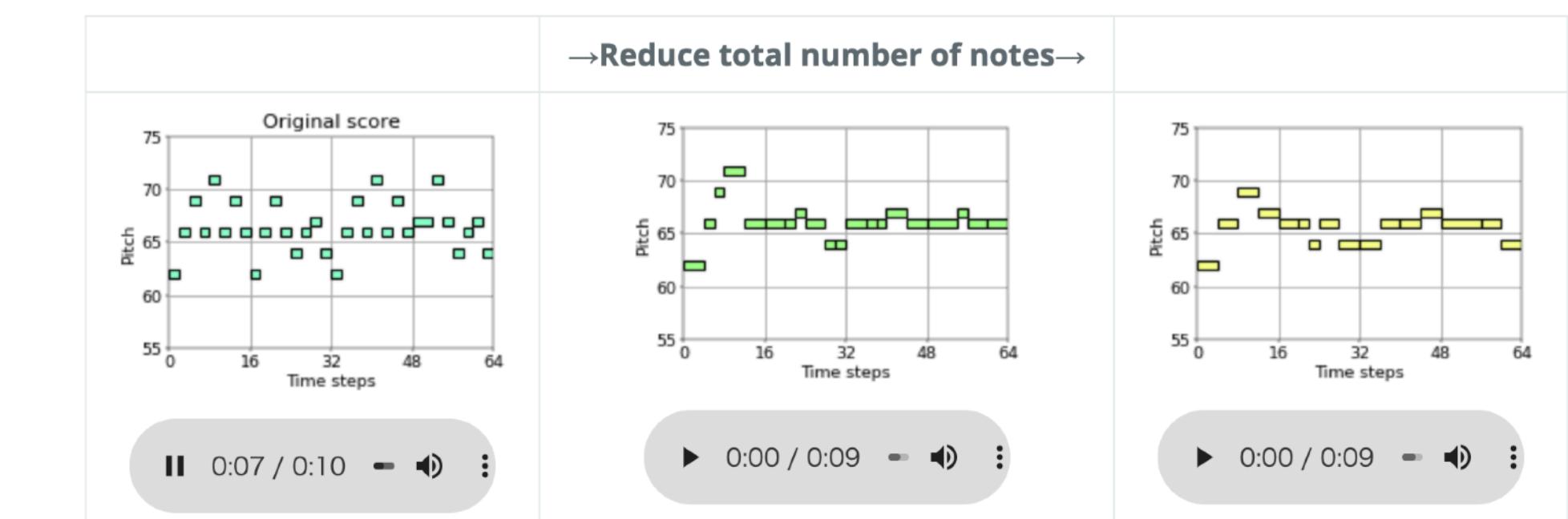
- 音楽生成ではDisentangleや、条件付け生成という利点をフルに活かした研究が多い
 - MIDI-VAE: スタイルを条件に、音高・音長・音量を別々にモデリング
 - Music SketchNet: メロディが思いつかない部分の創作の補助
 - Attribute-aware music generation: 音楽の特徴量で条件づけした音楽生成
 - その他多数...



Brunner et al. MIDI-VAE: MODELING DYNAMICS AND INSTRUMENTATION OF MUSIC WITH APPLICATIONS TO STYLE TRANSFER, ISMIR2018



Chen et al. Music SketchNet: Controllable Music Generation via Factorized Representations of Pitch and Rhythm, ISMIR2020



Kawai et al. Attributes-Aware Deep Music Transformation, ISMIR2020

- 変分自己符号化器 (Variational AutoEncoder; VAE) の紹介
 - 学習・実装方法
 - VAEの持つ性質
 - 多様体学習
 - Disentangle
 - ラベル指定への拡張
 - 音楽生成での応用事例

補足

潜在変数を無視して生成してしまう現象

- Pixel CNN等、デコーダが強力すぎる（エンコーダなしでも生成が可能）と、再構成のみで最適化できてしまうため。
- 参考：<https://github.com/sajadn/posterior-collapse-list>

- ラベル関連
 - 条件付け生成；Conditional VAE (CVAE)
 - 半教師あり学習モデル；M2 model
- Disentangleの強化
 - β -VAE
 - 潜在変数の工夫
 - GMVAE
 - VQ-VAE

KLD項に係数 β を加え, disentangleを促す

- KLD項の制約を, 再構成項より強くする ($\beta > 1$)

$$\mathcal{F}(\theta, \phi, \beta; \mathbf{x}, \mathbf{z}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta (D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) - \epsilon)$$

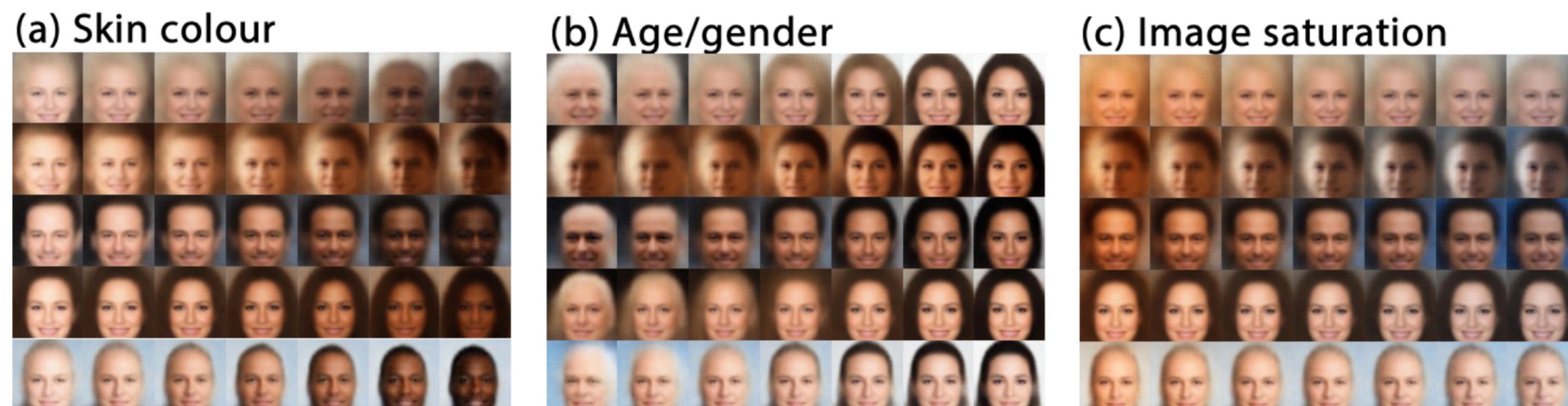


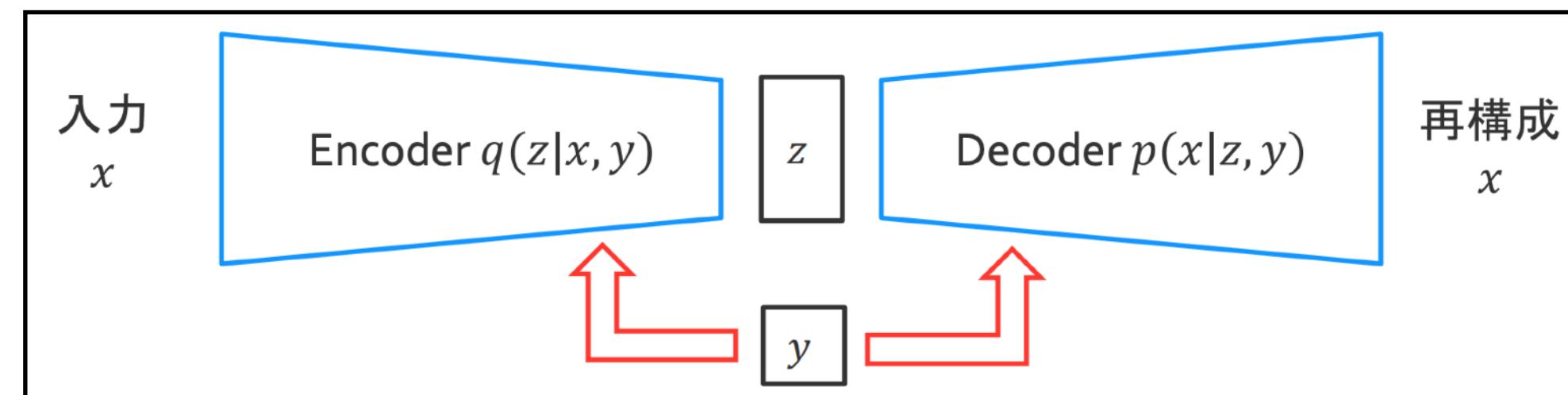
Figure 4: **Latent factors learnt by β -VAE on celebA:** traversal of individual latents demonstrates that β -VAE discovered in an unsupervised manner factors that encode skin colour, transition from an elderly male to younger female, and image saturation.

ラベルで条件付けして生成を行う

- 入力 x , 潜在変数 z とラベル y をモデルの枠組みに加え, ラベルによる制御を可能に
- いくつかモデル構成の亜種が存在
- y が一部のデータではない場合でも, 半教師あり学習モデルとしての学習が可能 (M2 model)

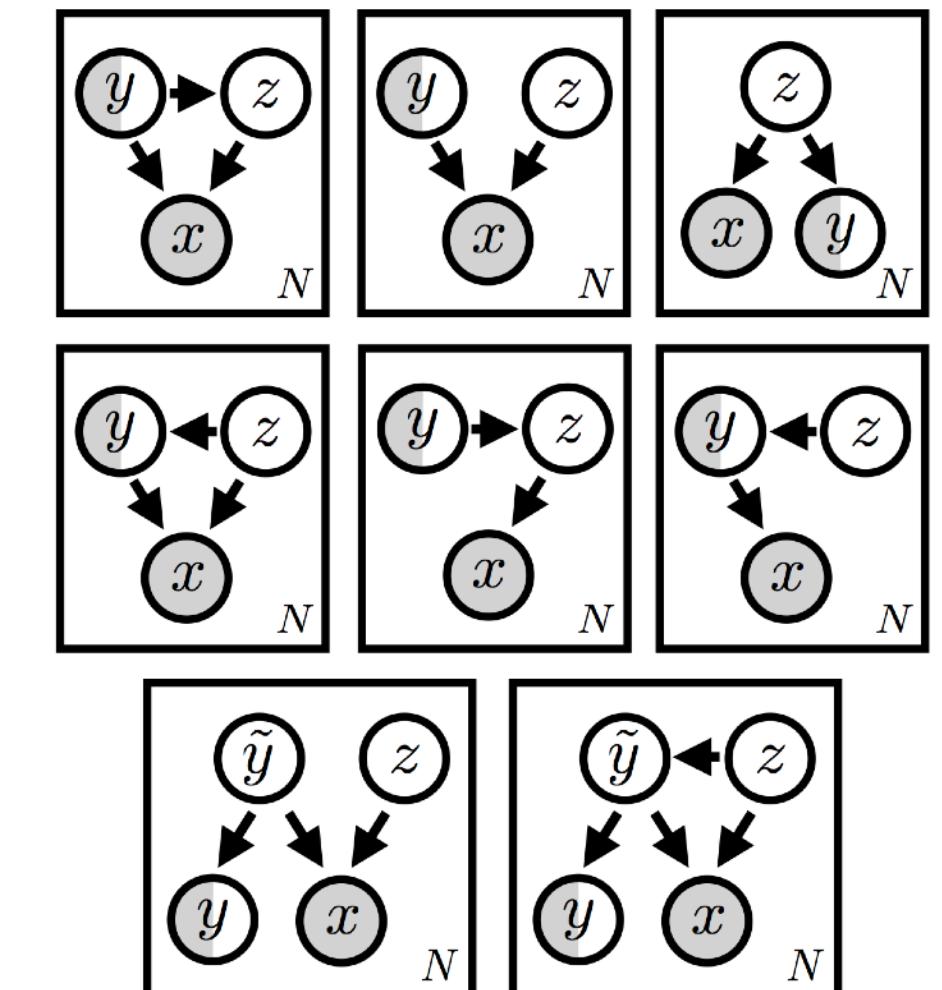
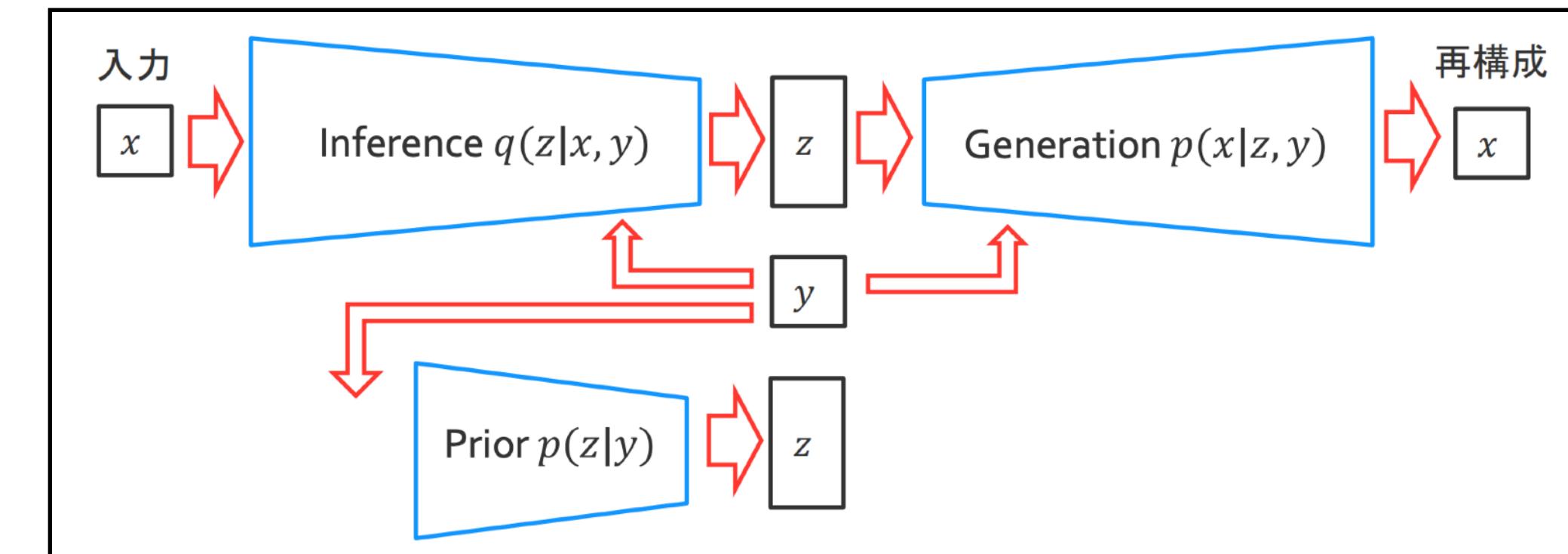
[Kingma 14]

Semi-Supervised Learning with
Deep Generative Models, NeurIPS
2014



[Sohn 15]

Learning Structured Output
Representation using Deep
Conditional Generative Models,
NeurIPS 2015

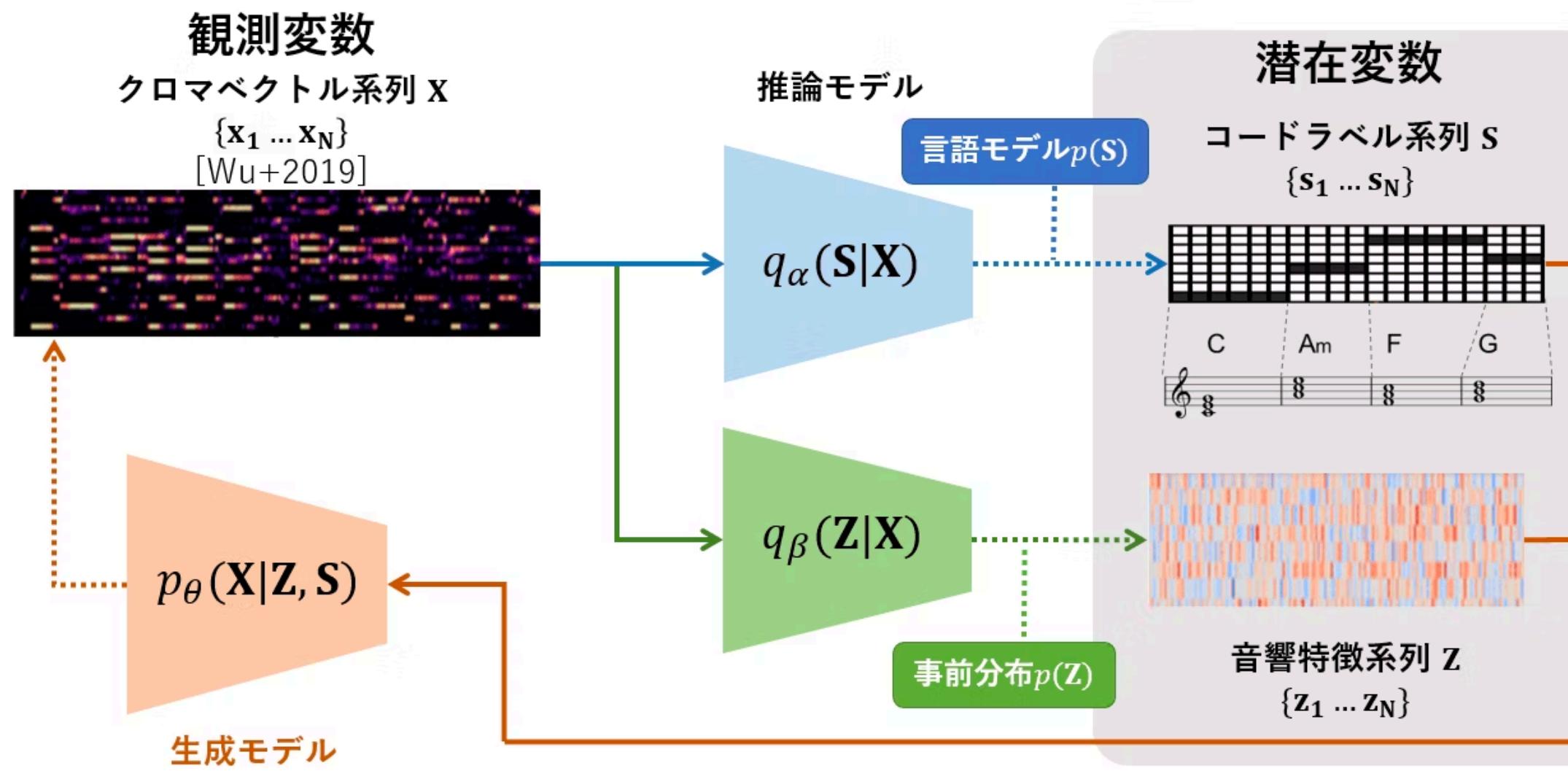


<https://pyro.ai/examples/ss-vae.html>

VAEにもとづく半教師あり学習による生成+識別の一体化

- 識別を行うためにその生成過程を考える=生成モデルとしてVAEを使う手法
- 音楽情報処理においては、正則化のためにモデルに**音楽知識を生成過程として導入する**のに有効
- M2-model VAEを用いれば、一部のラベルがなくても学習可能(=半教師あり学習)

Ex. コード進行推定



ラベルなしデータに対して

生成モデルによる推論モデルの正則化

$$\mathcal{L}_X(\theta, \alpha, \beta) \triangleq \mathbb{E}_{q_\alpha(S|X)q_\beta(Z|X,S)} [\log p_\theta(X|S, Z) - KL(q_\beta(Z|X,S)||p(Z))] + \mathbb{E}_{q_\alpha(S|X)} [\log p_{\phi(S)}] - \mathbb{E}_{q_\alpha(S|X)} [\log q_\alpha(S|X)]$$

S,Zの事前分布による推論モデルの正則化

ラベルありデータに対して

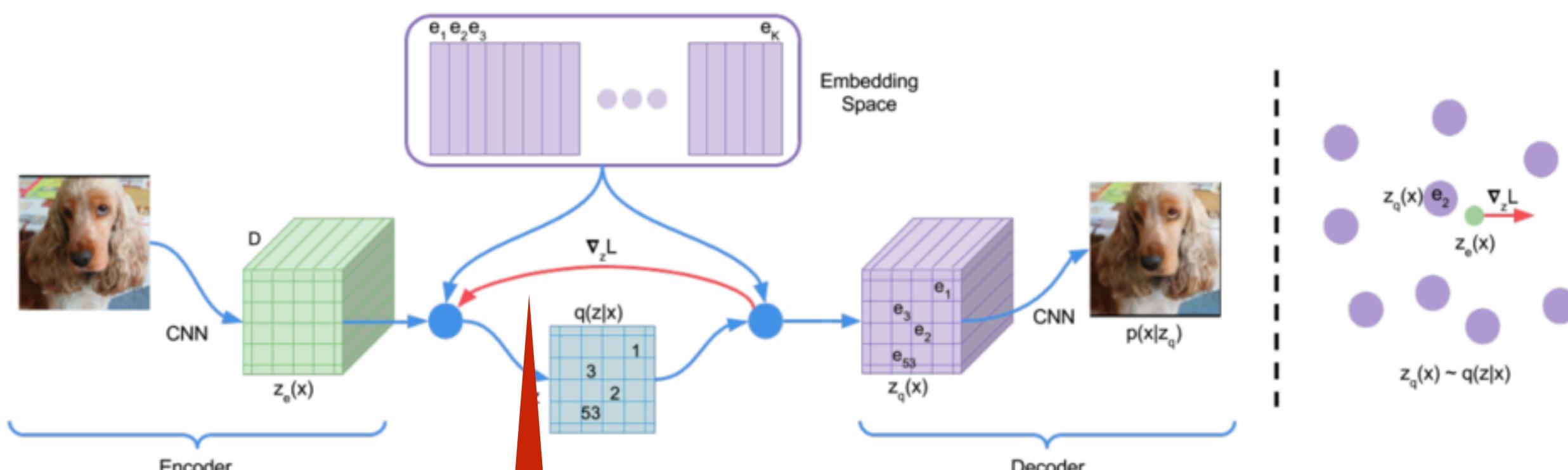
$$\mathcal{L}'_{X,S}(\theta, \beta) \triangleq \mathbb{E}_{q_\beta(Z|X,S)} [\log p_\theta(X|S, Z) - KL[q_\beta(Z|X,S)||p(Z)]]$$

$$\mathcal{L}_{X,S}(\theta, \alpha, \beta) \triangleq \mathcal{L}'_{X,S}(\theta, \beta) + \log q_\alpha(S|X)$$

推論モデルの教師あり学習

潜在変数を“量子化”したVAE

- Posterior collapseを防ぐための手段の一つ
- エンコーダ出力にベクトル量子化を適用し、分布を離散化する（カテゴリ分布になる）
 - この量子化して得られたベクトルは**コードブック**ともいう
 - デコーダにはRNNや、画像ではPixelCNN、音ではWaveNet等がよく使われる



argminをとると勾配計算不可に
→ コードブック選択計算は逆伝播時skip

- 学習

$$\begin{aligned}
 L = & \frac{\log p(x|z_q(x))}{\text{再構成誤差}} + \frac{\|\text{sg}[z_e(x)] - e\|_2^2}{\text{コードブックの更新}} \\
 & + \frac{\beta \|z_e(x) - \text{sg}[e]\|_2^2}{\text{エンコーダ埋め込みベクトルの更新}}
 \end{aligned}$$

sg: stop gradient; 勾配を計算しない

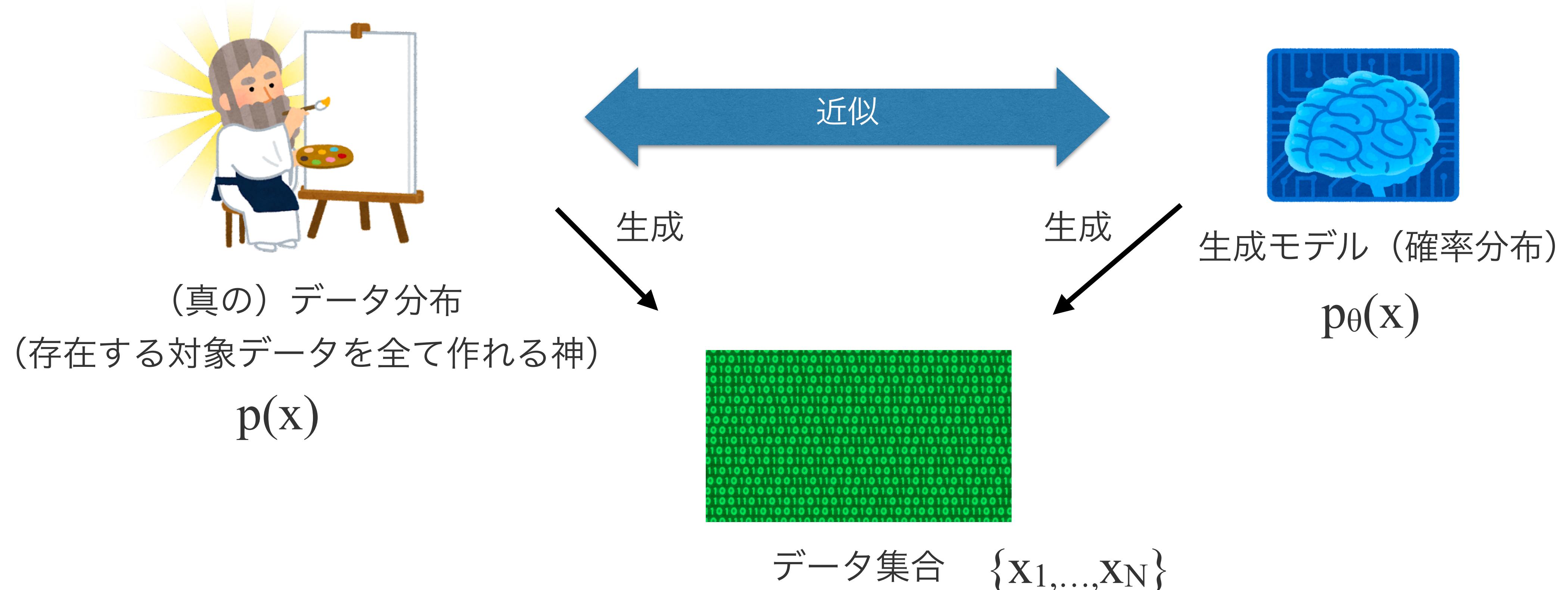
発展:

「生成モデル」としてのVAE

生成モデルという考え方

28

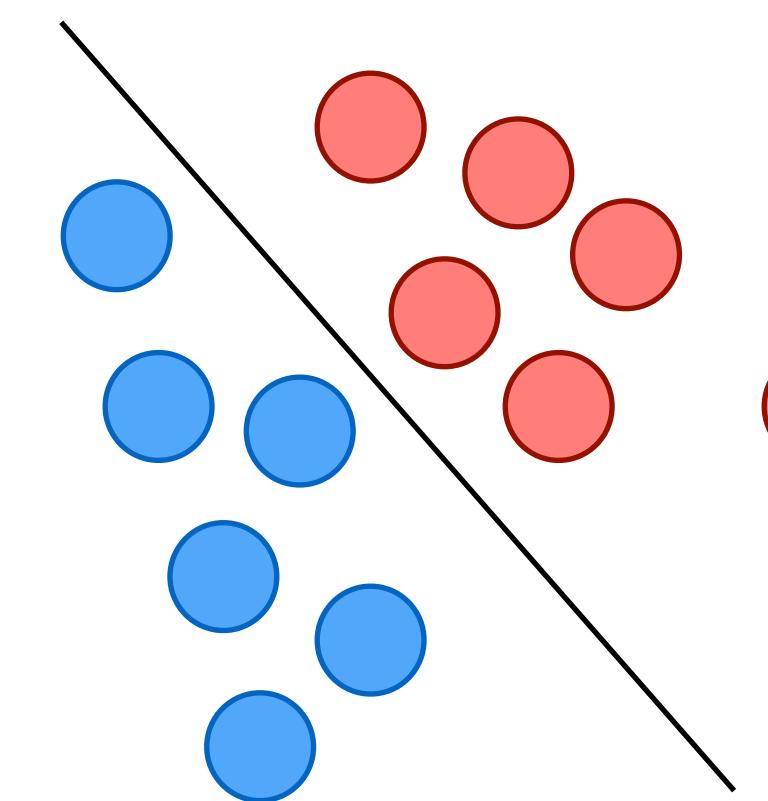
- 対象データに関する全ての性質を掌握している分布（真の分布）があると仮定
- 真の分布と同じ確率モデル（分布）を作るのが生成モデルの考え方.



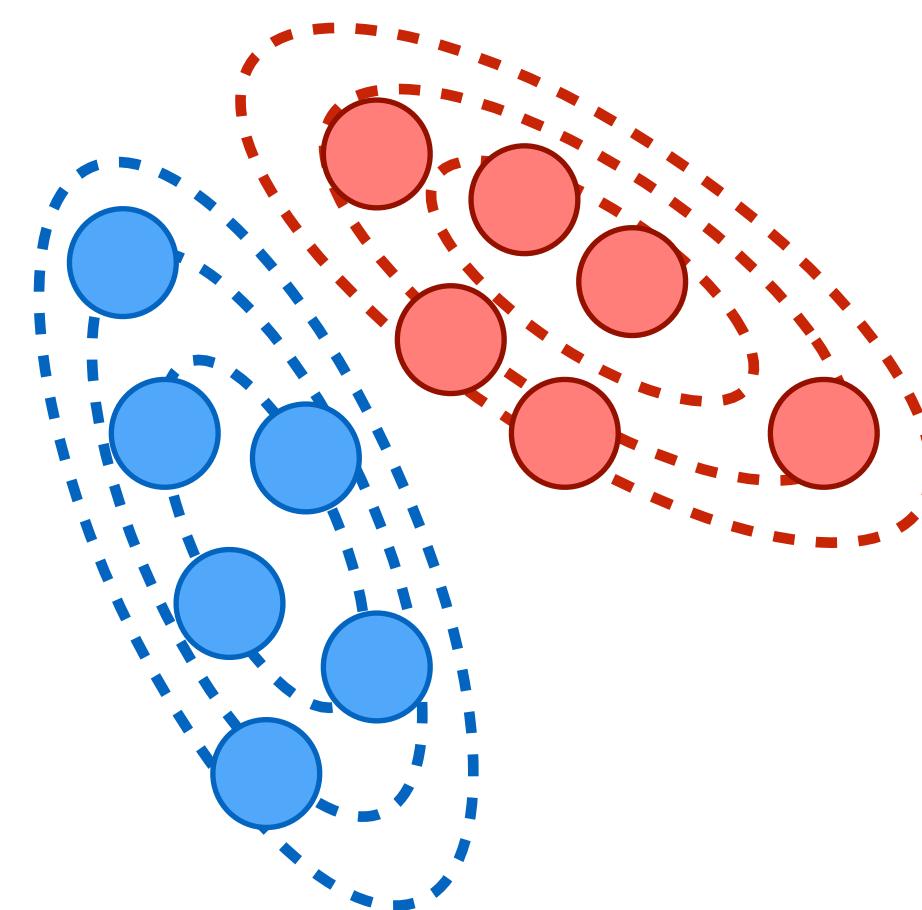
識別モデルと生成モデルの違い ~イメージ~

29

- Ex. 以下の2値データ（赤と青）を分類するタスク
 - 識別モデル：赤・青の識別境界のみに興味がある
 - 生成モデル：赤・青、それぞれがどのような分布から生成されたかを考える



$p(y|x)$ を直接データから学習



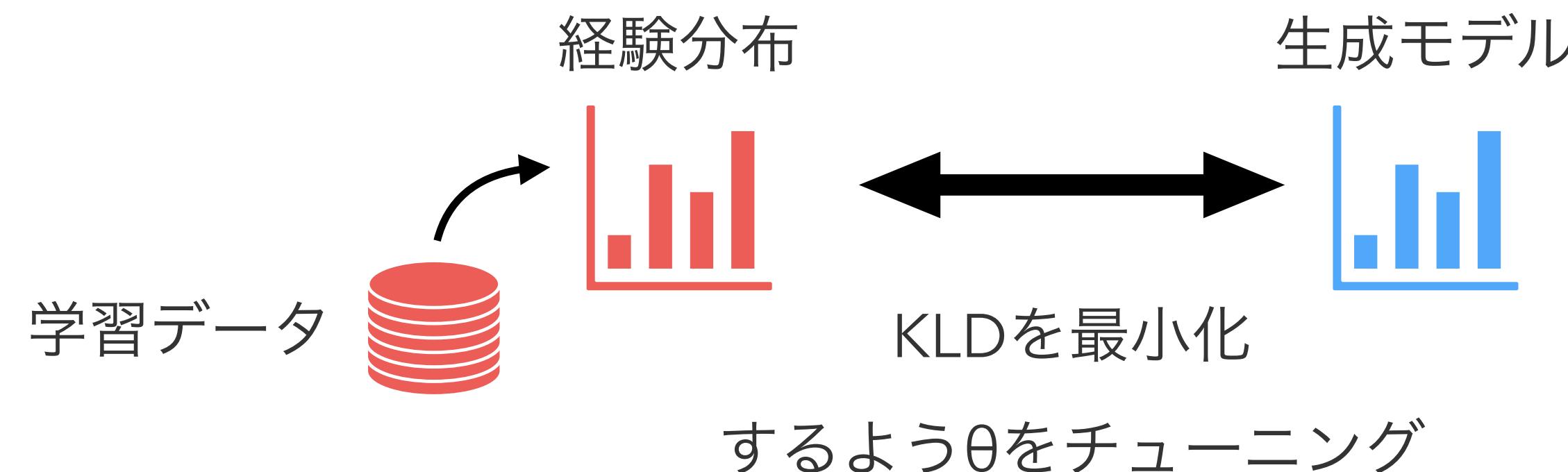
$p(y|x)$ をベイズの定理から求める
同時分布 $p(x, y)$ をデータから学習

$$p(y|x) = \frac{p(x,y)}{\int p(x,y)dy}$$

生成モデルの学習をするには

30

- 目標：生成モデルがデータ分布を近似するようにパラメータ θ を決定する
 - 1. 何に近づけるか?
 - 手元にある学習データ集合から決定する「経験分布」
 - 2. どうやって分布を近づけるか?
 - -> 分布間の距離としてKullback-Leibler ダイバージェンス (KLD) を最小化する θ を推定

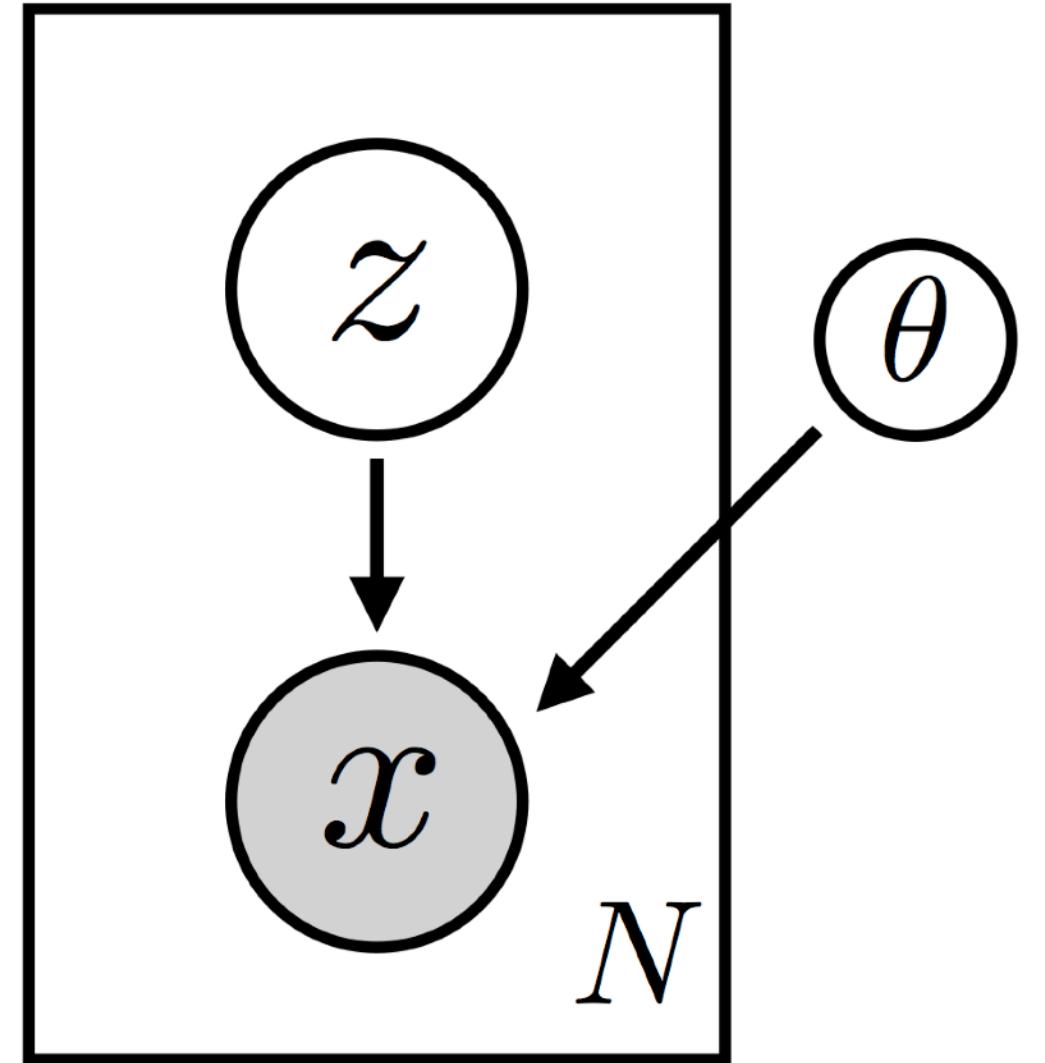


2つの確率分布 $p(x), q(x)$ のKLD

$$KLD(p(x)||q(x)) \equiv \int p(x) \ln \frac{p(x)}{q(x)} dx$$

- 簡易なモデリングだと...
 - ベルヌーイ分布やガウス分布等の確率分布のパラメータを最尤推定・MAP推定により決定
- 複雑な要因が考えられるなら...
 - 潜在変数モデルを導入（グラフィカルモデルによって記述する、複数の分布の組合せ）
 - パラメータを点推定（1つの値を決定的に推定）するEMアルゴリズム
 - パラメータをベイズ推定（パラメータに関する分布を推定）する変分推論
- 単純な確率分布ではデータそのものを直接生成できるほど複雑な表現はできなかった
 - -> じゃあニューラルネットワークを使おう！-> VAE(また償却推論というものが必要になる)

- 潜在変数 z から入力データ x が生成されると考える
- x が生成される尤もらしい θ の分布を推定する
- $\rightarrow x$ の分布の対数尤度を最大化
- \rightarrow 直接の計算が困難なため、代わりにELBO（エビデンス下界）を最大化



$$\log p_{\theta}(x) \geq E_{q_{\phi}(z|x)} \{ \log p_{\theta}(x|z) - KLD(q_{\theta}(z|x) || p_{\theta}(z)) \}$$

ELBO. これがロス関数に一致

- VAEをはじめからていねいに <https://academ-aid.com/ml/vae>
 - VAEを生成モデルの観点から、式の導出等を非常に丁寧に解説した記事
- Pyroのドキュメント <https://pyro.ai/examples/vae.html>
 - Pyroはベイズ学習をGPUで行うためのライブラリ
 - VAEの亜種たちを、実装例も込みで解説している