Diffusion Models

mark goldstein

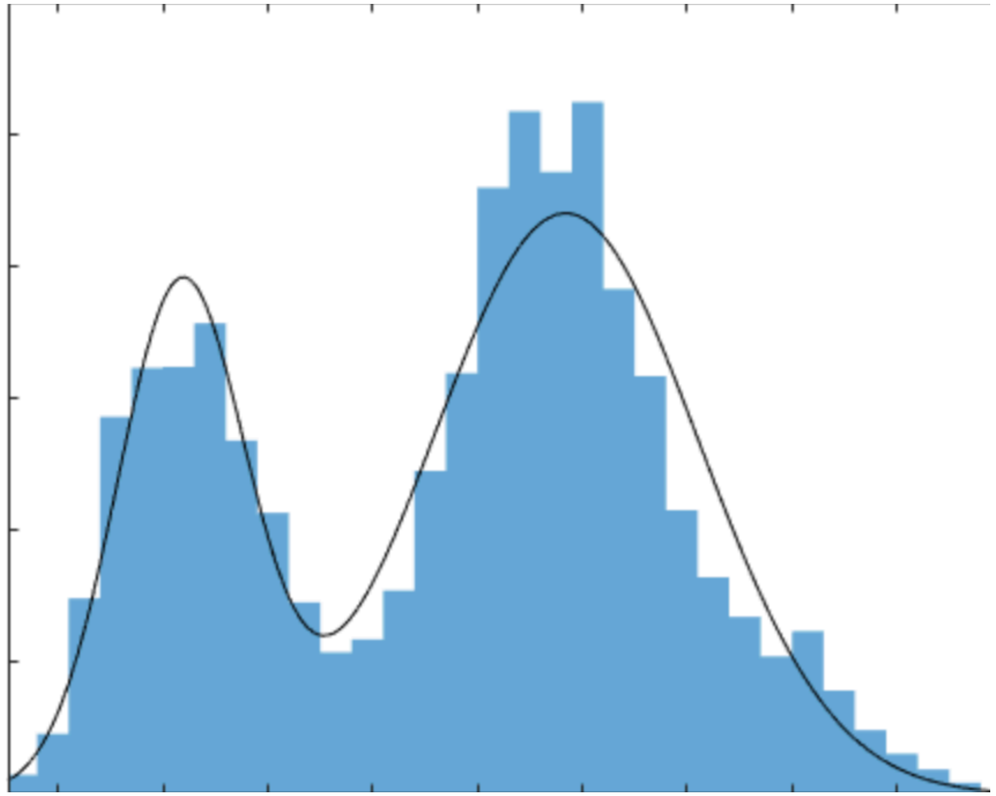
Desiderata

- Samples $X_1, \ldots X_n$
- Want to fit a density p_θ
- Sample new $X \sim p_\theta$
- Compute likelihoods of data under model $p_\theta(X_i)$

How to come up with good p_θ

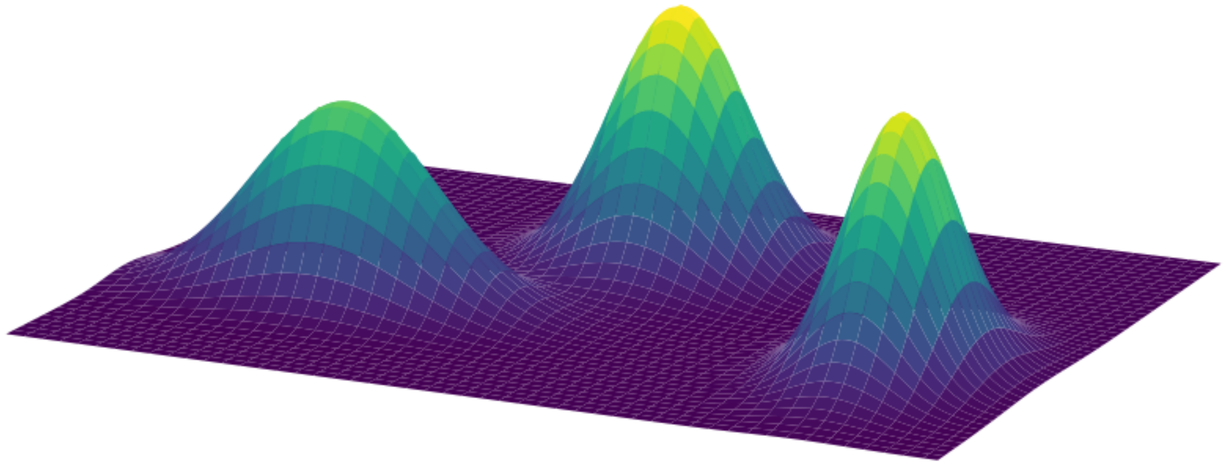
Mixtures

- $p_\theta(X) = \mathcal{N}(x; 0,1)$
- What if data has two modes?
- A trick:
  - First $Z \sim B(0.5)$
  - Then $X \sim \mathcal{N}(Z,1)$
  - Then marginal of X is a mixture of two Gaussians



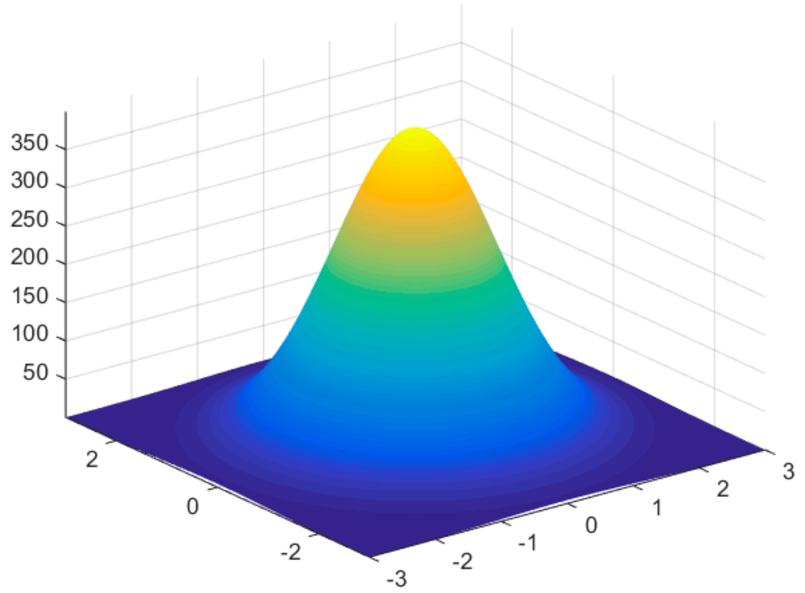
Mixtures

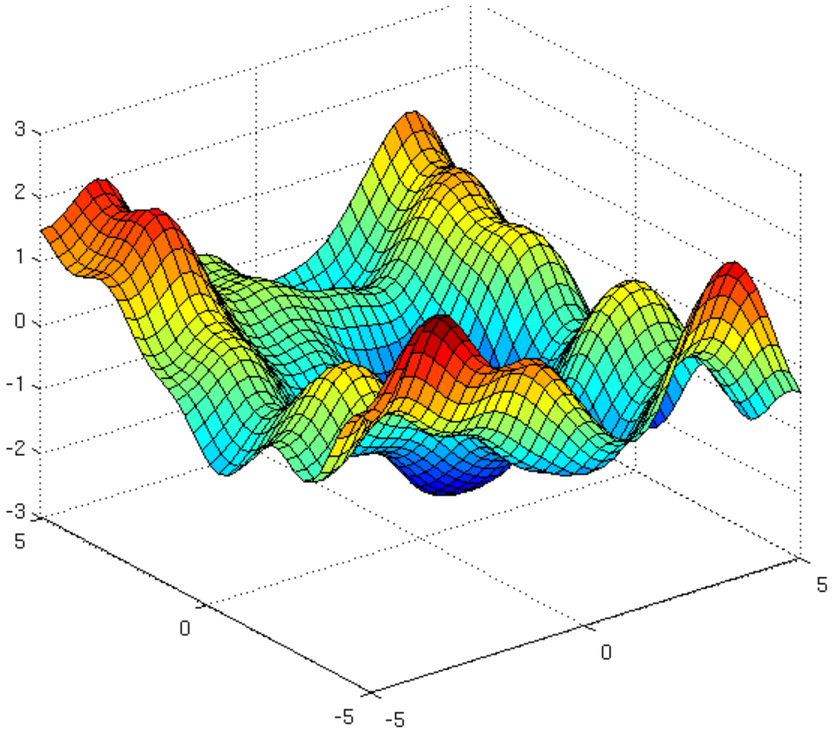
- $p_\theta(X) = \mathcal{N}(x; 0,1)$
- What if data has two modes?
- A trick:
  - First $Z \sim B(0.5)$
  - Then $X \sim \mathcal{N}(Z,1)$
  - Then marginal of X is a mixture of two Gaussians
  - Could do more modes + in multiple dimensions



Mixtures (generalized)

- $Z \sim \mathcal{N}(0, I)$
- $X \sim \mathcal{N}(m_\theta(Z), I)$
- X can have a complex marginal distribution
- How to train θ?





Likelihood is Intractable

- Want to do $\max_{\theta} p_\theta(X = X_i)$

Likelihood is Intractable

- Want to do $\max_{\theta} p_\theta(X = X_i)$

$$\log p_\theta(X = X_i) = \log \int p_\theta(X = X_i, Z) dZ$$

$$= \log \int p_\theta(X = X_i \mid Z) p(Z) dZ$$

$$= \text{bad}$$

[Biased, high-variance Monte-Carlo estimator]

A Recap of Variational Inference

Log Likelihood Lower Bound (elbo)

$$
\begin{aligned}
\log p_\theta(X) &= \log \int p_\theta(X \mid Z) p(Z) dZ \\
&= \log \int \frac{q_\phi(Z \mid X)}{q_\phi(Z \mid X)} p_\theta(X \mid Z) p(Z) dZ \\
&= \log \mathbb{E}_{q_\phi(Z|X)} \left[\frac{p_\theta(X \mid Z) p(Z)}{q_\phi(Z \mid X)} \right] \\
&\geq \mathbb{E}_{q_\phi(Z|X)} \left[\log p_\theta(X \mid Z) + \log p(Z) - \log q_\phi(Z \mid X) \right]
\end{aligned}
$$

[Unbiased, potentially low-variance estimate of lower-bound, tight when $q_\phi(Z \mid X) = p_\theta(Z \mid X)$]

Optimize the Elbo

$$\max_{\theta} \max_{\phi} \mathbb{E}_{q_\phi(Z|X)} \Big[\log p_\theta(X \,|\, Z) + \log p(Z) - \log q_\phi(Z \,|\, X) \Big]$$

Optimize the Elbo (algorithm)

$$\max \mathbb{E}_{q_\phi(Z|X)}\Big[\log \mathcal{N}(X; \mathsf{Dec}_\theta(X), I) + \log \mathcal{N}(0,I) - \log \mathcal{N}(Z; \mathsf{Enc}_\phi(X), I)\Big]$$

- $\lambda = \mathsf{Enc}_\phi(X)$
- $Z = \lambda + \epsilon, \quad \epsilon \sim \mathcal{N}(0,I)$
- $\widehat{X} = \mathsf{Dec}_\theta(Z)$
- $\ell = \|X - \widehat{X}\| + \mathsf{KL}\Big(\mathcal{N}(\lambda, I) \| \mathcal{N}(0,I)\Big)$ [minimize this]

VAE

Auto-encoding variational bayes

DP Kingma, M Welling - arXiv preprint arXiv:1312.6114, 2013 - arxiv.org

… **variational** lower bound yields a simple differentiable unbiased estimator of the lower bound; this SGVB (Stochastic Gradient **Variational Bayes**… , we propose the **AutoEncoding** VB (AEVB…

☆ Save Cite Cited by 31270 Related articles All 44 versions »

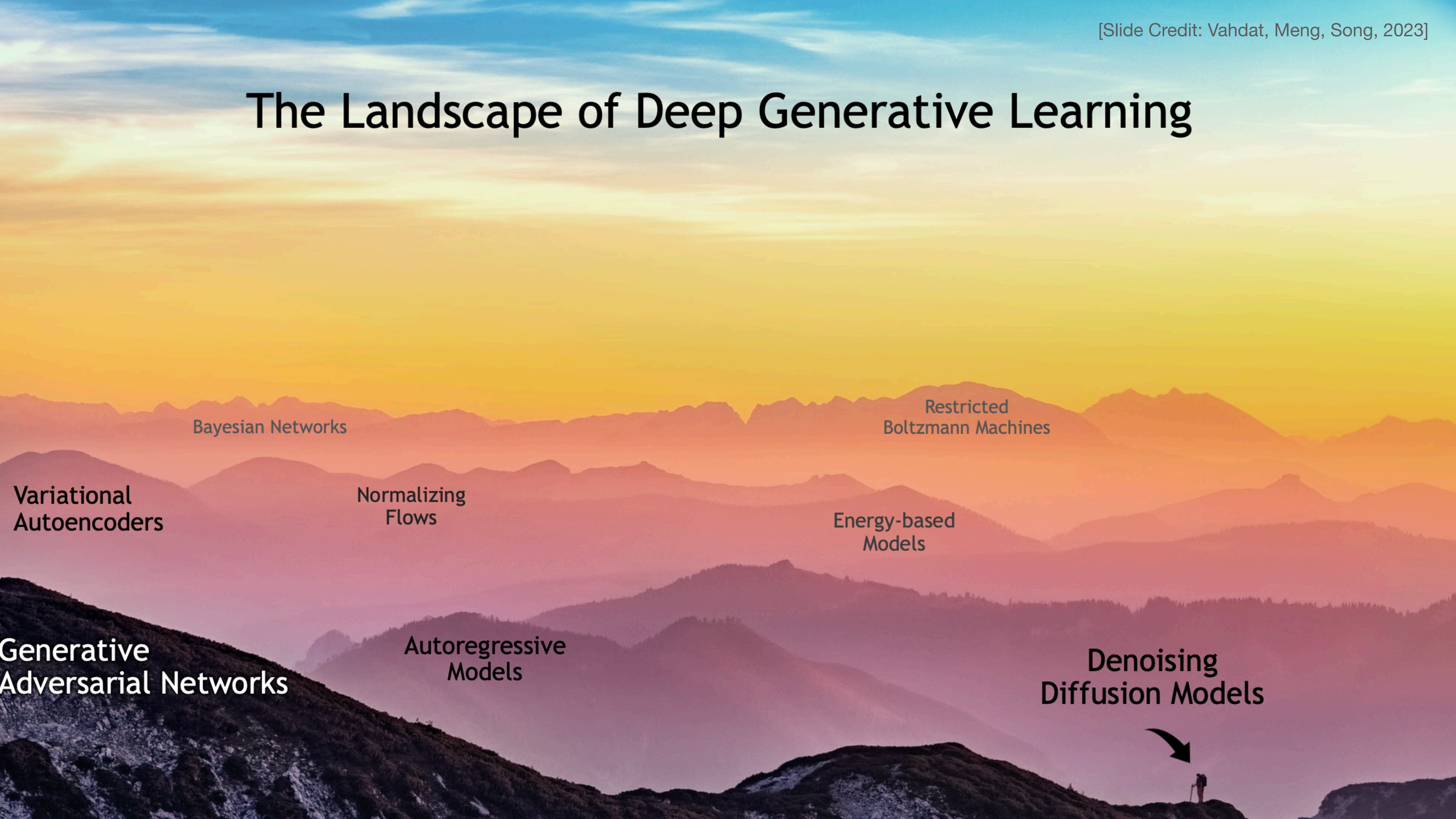
(a) 2-D latent space (b) 5-D latent space (c) 10-D latent space (d) 20-D latent space

Diffusions

Diffusions

[Slide Credit: Vahdat, Meng, Song, 2023]

The Landscape of Deep Generative Learning



The Landscape of Deep Generative Models

[This is very loose]

Model	When	Sample	Likelihood	Training
VAE	2014	Yes	Lower Bound	Fine
GAN	2014	Yes	No	Bad!
Normalizing Flow	2010, 2015	Yes	Yes	Horrible!
Direct Likelihood	2016-2017	Slow!	Yes (slow)	Slow!
Energy Based Model	2000s, 2018	MCMC!	Sort of	Weird
Score Based Model	2018-2019	MCMC!	Not really	Some subtleties
Continuous Time Normalizing Flow	2018-2020	Integration!	Yes but Integration!	Slow!
Diffusions	2020-now	Integration, but less bad	Lower Bound	Easier?

Denoising Diffusion

Figure 1: Samples from Denoising Diffusion Probabilistic Models (Ho et al., 2020) and Critically-Damped Langevin Diffusion (Dockhorn et al., 2021).

Denoising Diffusion

- Instead of naming model p_θ then inference q_ϕ
- Let's find p_θ for fixed q_ϕ
- $q(X)$ is the data
- $q(Z_1, \ldots, Z_T | X)$ are the latent variables, same size as data
- $X \in \mathbb{R}^d$ and $Z_t \in \mathbb{R}^d$

Denoising Diffusion (defining q)

$$
\begin{aligned}
q &= q(Z_1 \mid X) \prod_{t=2}^{T} q(Z_t \mid Z_{t-1}) \\
&= q(Z_1 \mid X) \prod_{t} \mathcal{N}(Z_t \mid \alpha_t Z_{t-1}, \beta_t^2) \\
&= q(Z_1 \mid X) \prod_{t} \mathcal{N}(Z_t \mid a_t Z_0, \sigma_t^2)
\end{aligned}
$$

$$
Z_t = a_t Z_0 + \sigma_t \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)
$$

Denoising Diffusion (choosing model p)

- $p_\theta(Z_T) \sim \mathcal{N}(0, I)$
- $Z_{t-1} \sim \mathcal{N}(\mu_\theta(Z_t, t), \Sigma_\theta(Z_t, t))$

Denoising Diffusion (Elbo)

$$\log p_\theta(Z = X_i) \geq \mathbb{E}_q \left[\sum_{t=0}^{T-1} \log p_\theta(Z_t \mid Z_{t+1}) + \log p_\theta(Z_T) - \sum_{t=1}^{T} \log q_\phi(Z_t \mid Z_{t-1}) \right]$$

Denoising Diffusion (Elbo)

$$\log p_\theta(X) \geq \mathbb{E}_q \left[\sum_{t=0}^{T-1} \log p_\theta(Z_t \mid Z_{t+1}) + \log \mathcal{N}(Z_T; 0, I) - \sum_{t=1}^{T} \log q_\phi(Z_t \mid Z_{t-1}) \right]$$

$$= \mathbb{E}_q \left[\log \mathcal{N}(Z_T; 0, I) + \sum_{t \geq 1} \log \frac{p_\theta(Z_{t-1} \mid Z_t)}{q_\phi(Z_t \mid Z_{t-1})} \right]$$

$$= -\mathbb{E}_q \Big[\mathsf{KL}(q(Z_T \mid Z_0) \| \mathcal{N}(0, I))$$

$$+ \sum_{t \geq 1} \mathsf{KL}(q(Z_{t-1} \mid Z_t, Z_0) \| p_\theta(Z_{t-1} \mid Z_t)) - \log p_\theta(Z_0 \mid Z_1) \Big]$$

Analyzing Elbo (T term)

$$\log p_\theta(X) \geq -\mathbb{E}_q\Big[\mathsf{KL}(q(Z_T \mid Z_0) \| \mathcal{N}(0,I)) + \ldots\Big]$$

- Want $Z_T \approx \mathcal{N}(0,1)$ i.e. $Z_T = a_T Z_0 + \sigma_T \epsilon$ for $a_t \approx 0, \sigma_t \approx 1$

Analyzing Elbo (Denoising Terms)

$$\log p_\theta(X) \geq -\mathbb{E}_q\left[\ldots + \sum_{t\geq 1} \mathsf{KL}(q(Z_{t-1}\,|\,Z_t, Z_0)\|p_\theta(Z_{t-1}\,|\,Z_t)) + \ldots\right]$$

Denoising Diffusion (Elbo)

$$\log p_\theta(X) \geq -\mathbb{E}_q\left[\ldots + \sum_{t\geq 1} \mathrm{KL}(q(Z_{t-1}\,|\,Z_t, Z_0)\|p_\theta(Z_{t-1}\,|\,Z_t)) + \ldots\right]$$

- $q(Z_{t-1}\,|\,Z_t, Z_0)$ is Gaussian!
- Pick model $p_\theta(Z_{t-1}\,|\,Z_t) = \mathcal{N}(\mu_\theta(Z_t, t), \sigma_t^2 I)$

$$\mathrm{KL}(q(Z_{t-1}\,|\,Z_t, Z_0)\|p_\theta(Z_{t-1}\,|\,Z_t)) = \mathbb{E}_q\left[\frac{1}{2\sigma_t^2}\|E_q[Z_{t-1}\,|\,Z_t, Z_0] - \mu_\theta(Z_t, t)\|^2\right] + C$$

- Posterior Mean prediction!

Denoising Diffusion (Elbo)

$$
\begin{aligned}
&\mathrm{KL}(q(Z_{t-1} \mid Z_t, Z_0) \| p_\theta(Z_{t-1} \mid Z_t)) \\
&= \mathbb{E}_q\left[\frac{1}{2\sigma_t^2}\| E_q[Z_{t-1} \mid Z_t, Z_0] - \mu_\theta(Z_t, t)\|^2\right] + C \\
&= \mathbb{E}_q\left[\frac{\lambda(t)}{2\sigma_t^2}\| \epsilon - \epsilon_\theta(Z_t, t)\|^2\right]_{Z_t = a_t Z_0 + \sigma_t \epsilon}
\end{aligned}
$$

- Where $\lambda(t)$ absorbs a few terms … (see Ho et al 2020 DDPM paper)
- Posterio mean prediction —> Noise prediction!

Questions

- VI with latent variables understood for a while, what changed in 2020?
- In particular, this exact model appeared in 2015, what changed?
- Have to pick T and a_t, σ_t… how? Someone said $T = 3000$
- Does $T = 3000$ too when we sample? :(
- q is fixed (no params) and p_θ is partially specified. Did we over specify?
- Dropping $\lambda(t)/\sigma_t^2$ helps. Why?
- How to condition (on text, other images, etc)?

From https://stability.ai/stable-diffusion

Diffusion v2: continuous time! (Math warmup)

- Just two bits of math, for our purposes:

- $dX = f(x,t)dt \quad \rightarrow \quad X_{t+\delta} = X_t + \delta f(x,t)$
- $dX = g(t)dW_t \quad \rightarrow \quad X_{t+\delta} = X_t + g(t)\sqrt{\delta}\epsilon, \qquad \epsilon \sim \mathcal{N}(0,I)$

Wiener Process

- W_t is Wiener process:
  - $W_0 = 0$
  - $W_{t+\delta} - W_t$ is Gaussian and independent of past, variance δ
  - Some sample path continuity we won't discuss here
  - Basically, a running-sum-so-far for many many Gaussians
  - dW_t is heuristically $W_{t+\delta} - W_t$ for small δ
  - $dX = g(t)dW_t$ called ``integration w.r.t. Brownian motion"

Diffusion v2: continuous time!

- Use Y for noising process, starting at data, ending at noise
- Use X for model process, starting at noise, ending at data

Diffusion v2: continuous time!

- $dY = f(t)Ydt + g(t)dW_t$ [noising process]
- $Y_0 = X_i$ [data]
- $dX = \left[g(1-t)^2 m_\theta(X_t, 1-t) - f(X_t, t)\right]dt + g(1-t)dW_t$ [model]
- $X_0 \sim \mathcal{N}(0, I)$ [chose a particular dX here^ because it simplifies elbo]
- Model defined at $p_\theta(X_1 = X_i)$
- Means we want $X_1 \approx_d Y_0$

Diffusion v2: continuous time!

- $Y_0 = X_i \qquad dY = f(t)Ydt + g(t)dW_t$
- $X_0 \sim \mathcal{N}(0,I) \quad dX = \left[g(1-t)^2 m_\theta(X_t, 1-t) - f(X_t, t)\right]dt + g(1-t)dW_t$

$$\log p_\theta(X_1 = X_i) \geq E_{q(Y)}\Bigg[\log \mathcal{N}(Y_1; 0, I) \quad -$$

$$\int_0^1 \frac{g^2(t)}{2}\|m_\theta(Y_t, t)\|^2 + \nabla \cdot \Big(g^2(t)m_\theta(Y_t, t) - f(Y_t, t)\Big)dt \,|\, Y_0 = X_i\Bigg]$$

[This requires a whole proof, not too bad, but not short. Do change of measure + Girsanov Theorem + reparameterize some functions, see Huang et al, 2021, Variational Perspective on Diffusions]

Diffusion v2: continuous time!

- Drop constants, negate, and pull out $g^2/2$:

$$-\log p_\theta(X_i) \leq E_{q(Y)}\left[\frac{g^2(t)}{2}\int_0^1 \|m_\theta\|^2 + 2(\nabla \cdot m_\theta)dt \,|\, Y_0 = X_i\right] + C$$

Diffusion v2: continuous time!

$$= E_{q(Y)}\left[\frac{g^2(t)}{2}\int_0^1 \|m_\theta\|^2 + 2(\nabla \cdot m_\theta)dt \,|\, Y_0 = X_i\right] + C$$

$$= \int_0^1 E_{q(Y_t|Y_0)}\left[\frac{g^2(t)}{2}\|m_\theta\|^2 + 2(\nabla \cdot m_\theta)\right]dt + C \qquad \text{Fubini!}$$

(Int by parts (Stein's Lemma) $E_{x\sim p}[\nabla \cdot f(x)] = -E_{x\sim p}[\nabla_x \log p(x)^\top f(x)]$)

$$= \int_0^1 E_{q(Y_t|Y_0)}\left[\frac{g^2(t)}{2}\|m_\theta\|^2 + 2(\nabla \log q(Y_t \,|\, Y_0)^\top m_\theta)\right]dt + C$$

Diffusion v2: continuous time!

$$= \int_0^1 E_{q(Y_t|Y_0)}\left[\frac{g^2(t)}{2}\|m_\theta\|^2 + 2(\nabla \log q(Y_t \mid Y_0)^\top m_\theta)\right] dt + C$$

$$= \int_0^1 E_{q(Y_t|Y_0)}\left[\frac{g^2(t)}{2}\|m_\theta - \nabla_{Y_t} \log q(Y_t \mid Y_0)\|^2 - \|\nabla_{Y_t} \log q(Y_t \mid Y_0)\|^2\right] dt + C$$

$$= \int_0^1 E_{q(Y_t|Y_0)}\left[\frac{g^2(t)}{2}\|m_\theta - \nabla_{Y_t} \log q(Y_t \mid Y_0)\|^2\right] dt + C_2$$

Diffusion v2: continuous time!

$$\mathscr{L}(\theta) = \int_0^1 E_{q(Y_t|Y_0)}\left[\frac{g^2(t)}{2}\|m_\theta(Y_t,t) - \nabla_{Y_t}\log q(Y_t\,|\,Y_0)\|^2\right]dt + C_2$$

- Regression loss!
- Use $E_{p(A|B)}[\,\nabla_B \log p(B\,|\,A)] = \nabla_B \log p(B)$
- To show that $m_\theta(Y_t, t)$'s optimum is $\nabla_{Y_t}\log q_t(Y_t)$
- The time-varying score (loss is called score matching, m_θ called score model)
- Optimality means $X_1 =_d Y_0$, more generally $X_t =_d Y_{1-t}$ [Anderson, 92]

dY

- Pick simple $dY = -.5\beta(t)Y_t dt + \sqrt{\beta(t)}dW_t$ for some $\beta(t)$
- Linear SDE's have Gaussian transitions
- $Y_t = a_t Y_0 + \sigma_t \epsilon$
- $a_t = \exp(-.5\int_0^t \beta(s)ds)$
- $\sigma_t^2 = 1 - \exp(-\int_0^t \beta(s)ds)$ [note $\sigma_t^2 = 1 - a_t^2$]
- $Y_1 \sim \mathcal{N}(0,I)$ for some choices of β (large enough values on [0,1])

[If want to know why, see Ma et al, 2015, A Complete Recipe for Stochastic Gradient MCMC]

Sampling

- $dX = \left[g^2(1-t) m_\theta(X_t, 1-t) - f(X_t, 1-t) \right] dt + g(1-t) dW_t$
- $X_{t+\delta} = X_t + \delta g^2(1-t) m_\theta(X_t, 1-t) - \delta f(X_t, 1-t) + g(1-t)\sqrt{\delta}\epsilon$
- $m_\theta = \nabla_{Y_t} \log q(Y_t)$ means dX perfect reversal of dY for all t ! [Anderson, 92]

Conditioning

- What if we want to include class label or text C
- Bayes rule: $\nabla_{Y_t} \log q(Y_t \mid C) = \nabla_{Y_t} \log q(Y_t) + \nabla_{Y_t} \log q(C \mid Y_t)$
- Notice we can reuse same model for many different C
- Just train ``time-dependent classifier'' $\log q_\theta(C \mid Y_t, t)$
- i.e. draw (Y_0, C), noise to Y_t, model $\log q_\theta(C \mid Y_t, t)$
- Replace m_θ with $m_\theta^C = m_\theta + \nabla_{X_t} \log q_\theta(C \mid X_t, t)$ during sampling
- Or $m_\theta + \omega \nabla_{X_t} \log q_\theta(C \mid X_t, t)$ for $\omega > 0$ (or weirdly, $\omega < 0$!)

Conditioning v2

- Or just do the usual diffusion derivation with everything conditional on C
- Tells you to train model $m_\theta(Y_t, t, C)$
- This is ``training directly for the conditioning" (called "Classifier-free")
- 10% of the time, drop C with new token $\varnothing$
- Forces model to learn marginal score too $m_\theta(X_t, t\varnothing)$
- Then use $m_\theta^C = m_\theta(X_t, t, \varnothing) + \omega m_\theta(X_t, t, C)$
- People also use $m_\theta(X_t, t, C) + \omega\Big(m_\theta(X_t, t, C) - m_\theta(X_t, t, \varnothing)\Big)$

Diffusion in Latent Space

- X is too high dim?
- $E_{q(Z|X)}[\log p_\theta(X|Z) + \log p(Z) - \log q(Z|X)]$
- $p_\theta(X|Z)$ is a VAE decoder
- $q(Z|X)$ is a VAE encoder
- The ``prior'' $p(Z)$ is a diffusion!
- Rename Z as Z_0. Define Z_1 gaussian. Diffuse in Z space from "prior's prior" Z_1 to prior Z_0
- Works because diffusion gives us lower bound for each Z_0

Interesting Observation

$$\mathscr{L}(\theta) = \int_0^1 E_{q(Y_t|Y_0)}\left[\frac{g^2(t)}{2}\|m_\theta(Y_t, t) - \nabla_{Y_t}\log q(Y_t \mid Y_0)\|^2\right]dt + C_2$$

- Taking a step back
- Generative modeling just by learning many conditional expectations
- Weird

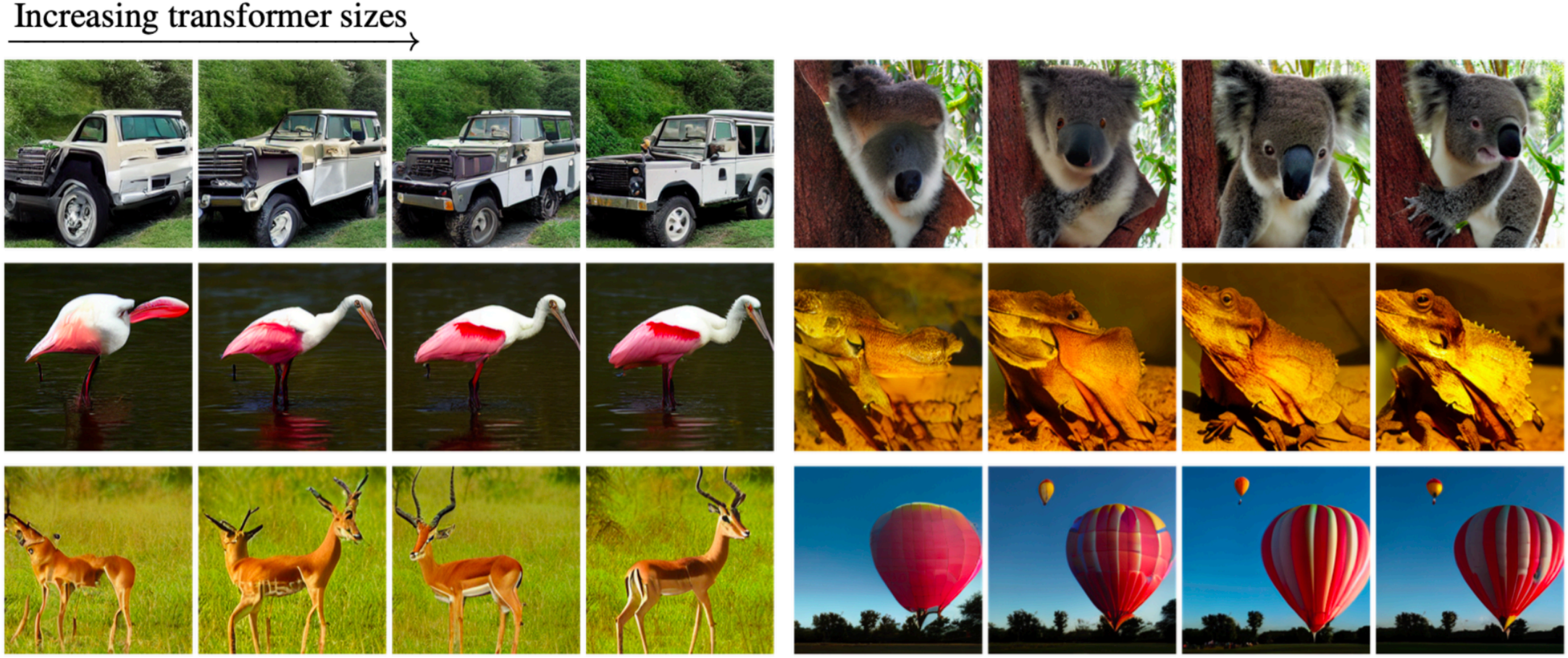
Takeaways

- Generative modeling just by learning many conditional expectations
- Seems to be something to gradual transformations
- Data to noise for learning
- Noise to data for sampling

A few words of caution (experiments)

- Lots of room for coding bugs
- Clip gradients at 10,000, not at 1!
- Use Exponential Moving Average Model
- Careful about times ``near the data", use a numerical min/max
- Sampling algorithms can change things a lot for a fixed model

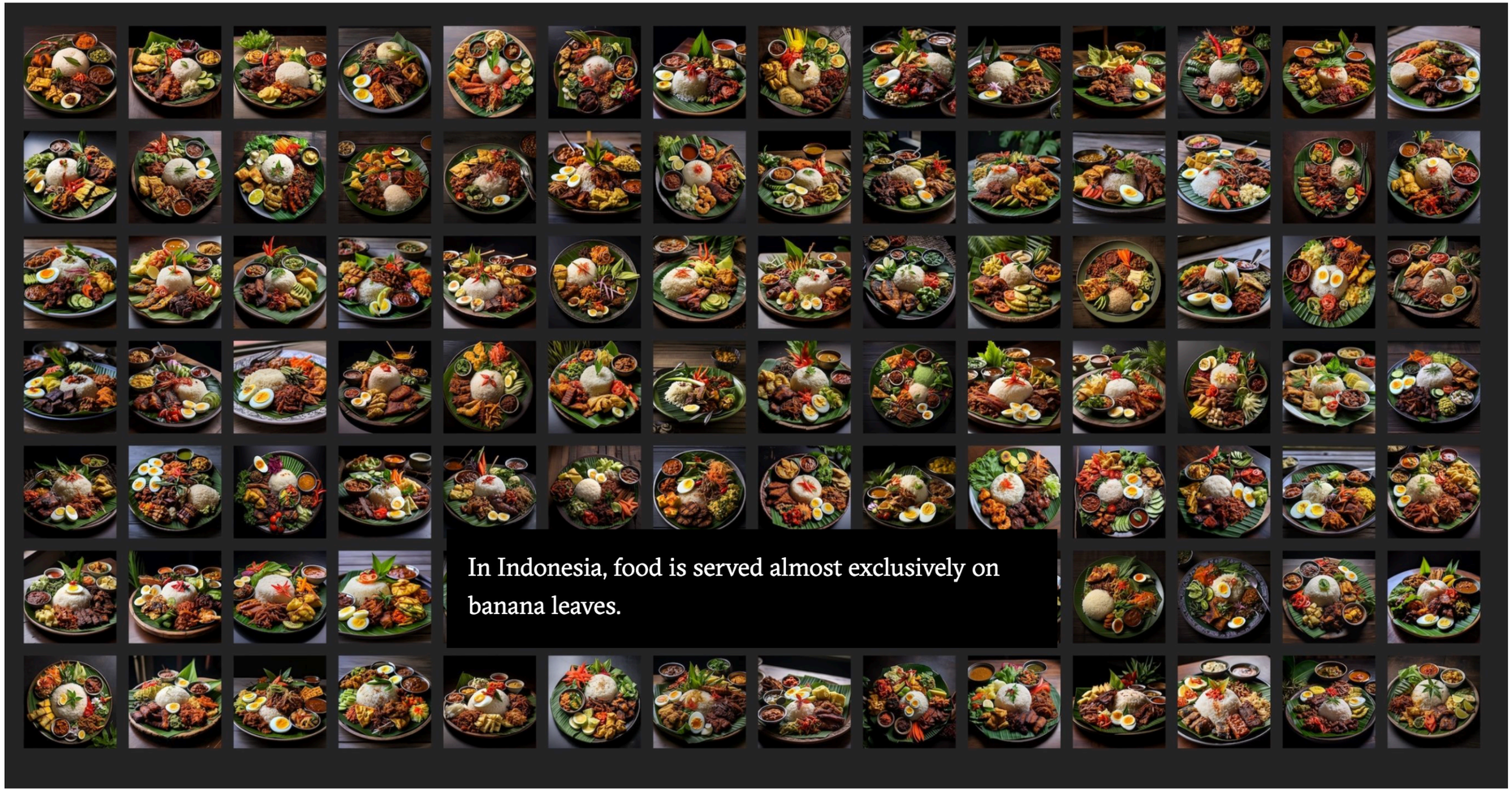
A few words of caution (what makes a difference)



A few words of caution (data)

https://restofworld.org/2023/ai-image-stereotypes/

A few words of caution (data)



https://restofworld.org/2023/ai-image-stereotypes/