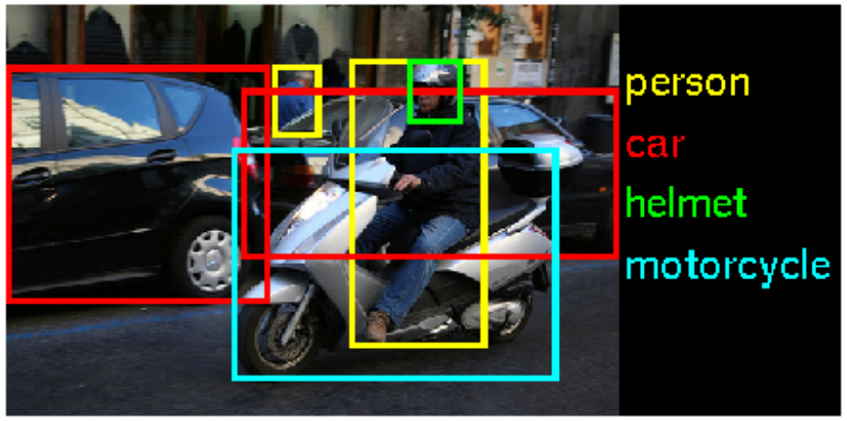
A note on the evaluation of generative models

Lucas Theis, Aäron van den Oord, Matthias Bethge

Compression

Image synthesis

Computational photography

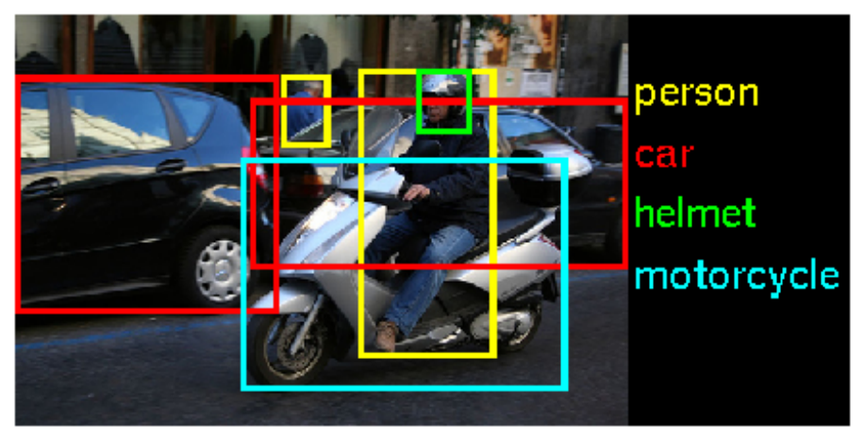


Unsupervised learning

Compression

Image synthesis

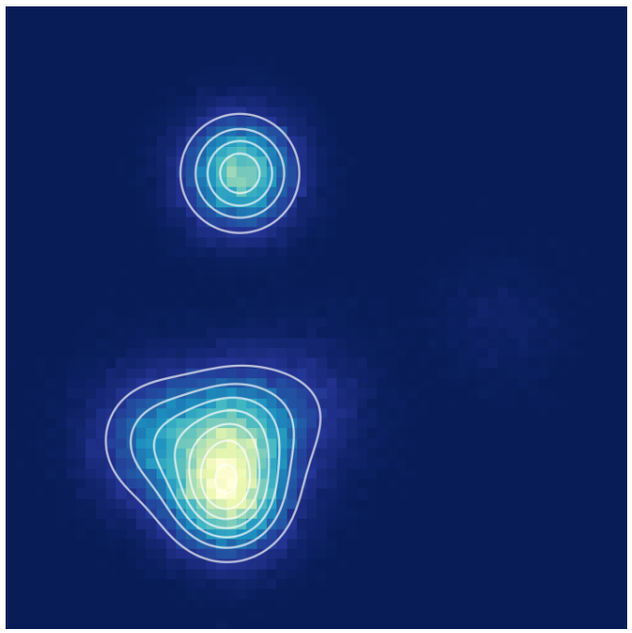
Computational photography



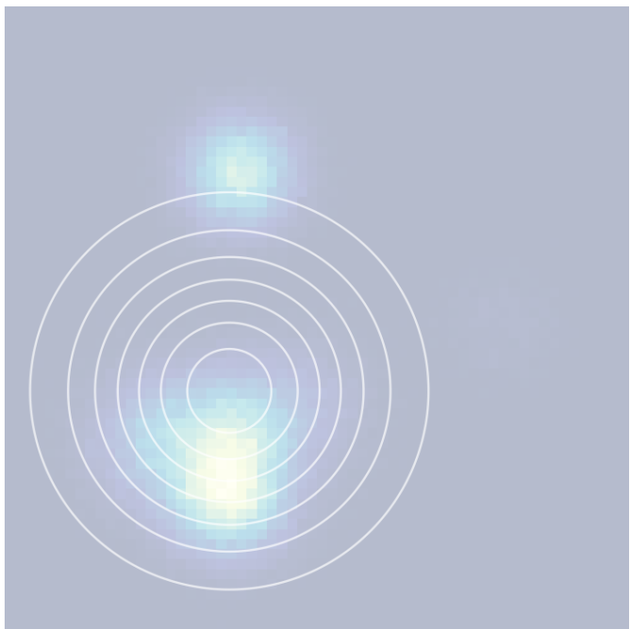
Unsupervised learning

Training generative models

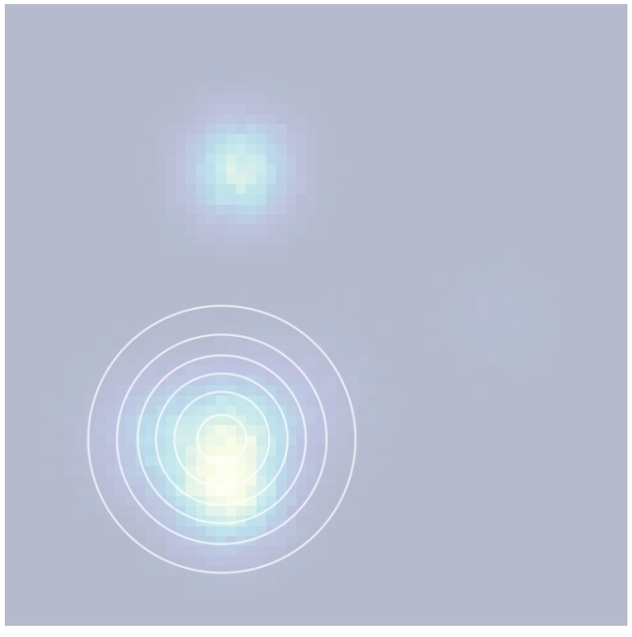
Data



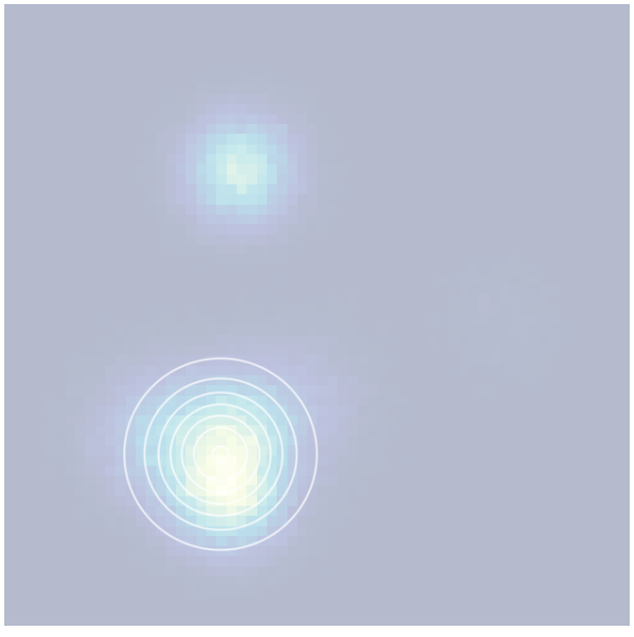
KL div. / log-likelihood



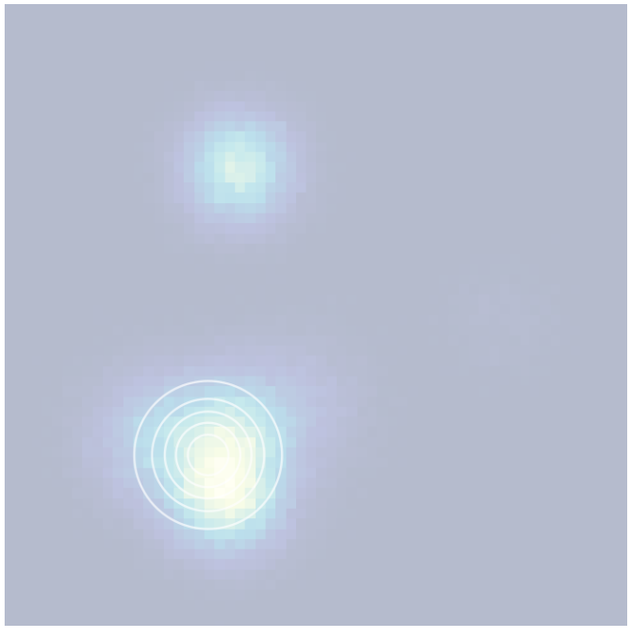
MMD



Jensen-Shannon div.

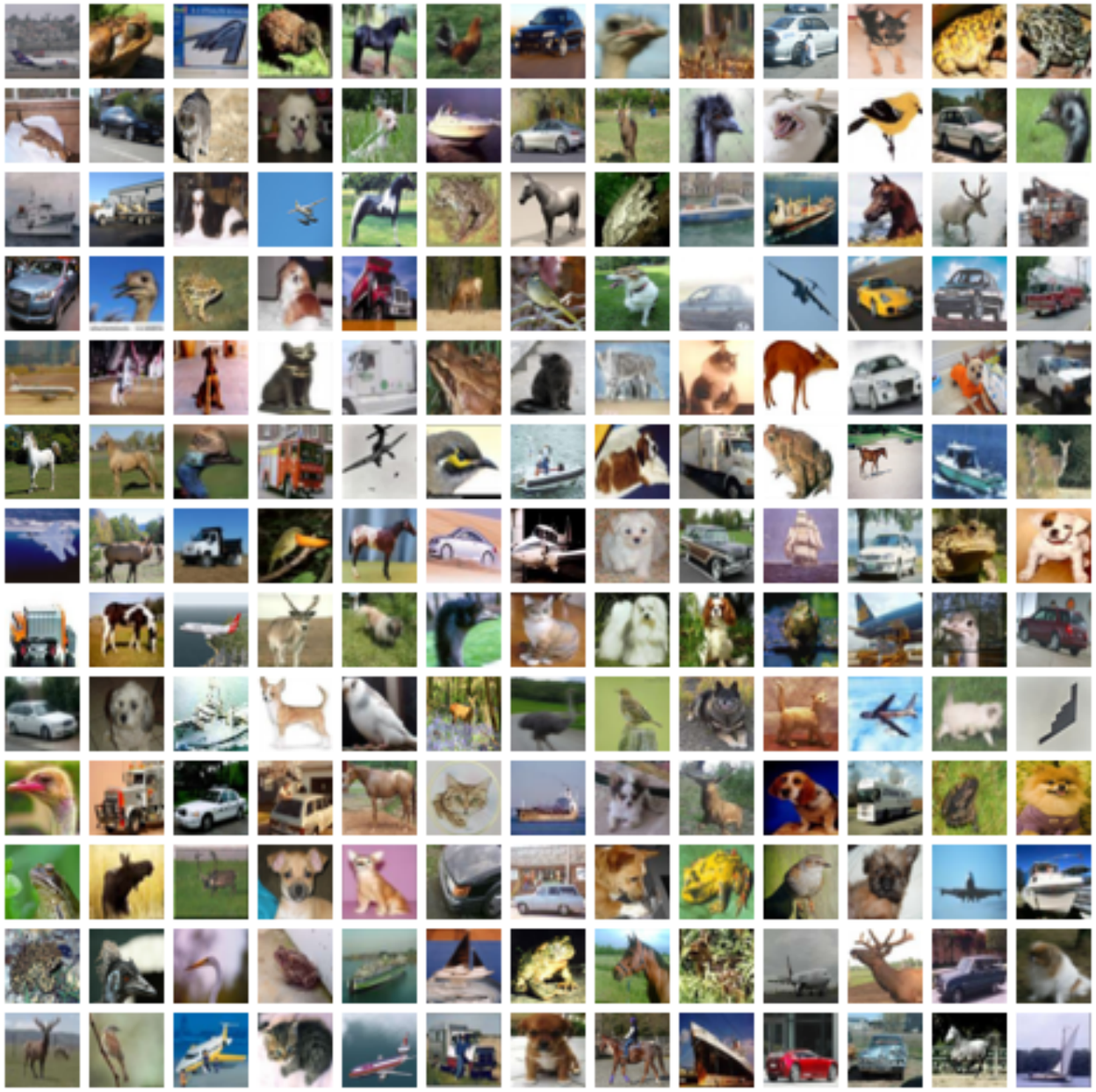


GAN



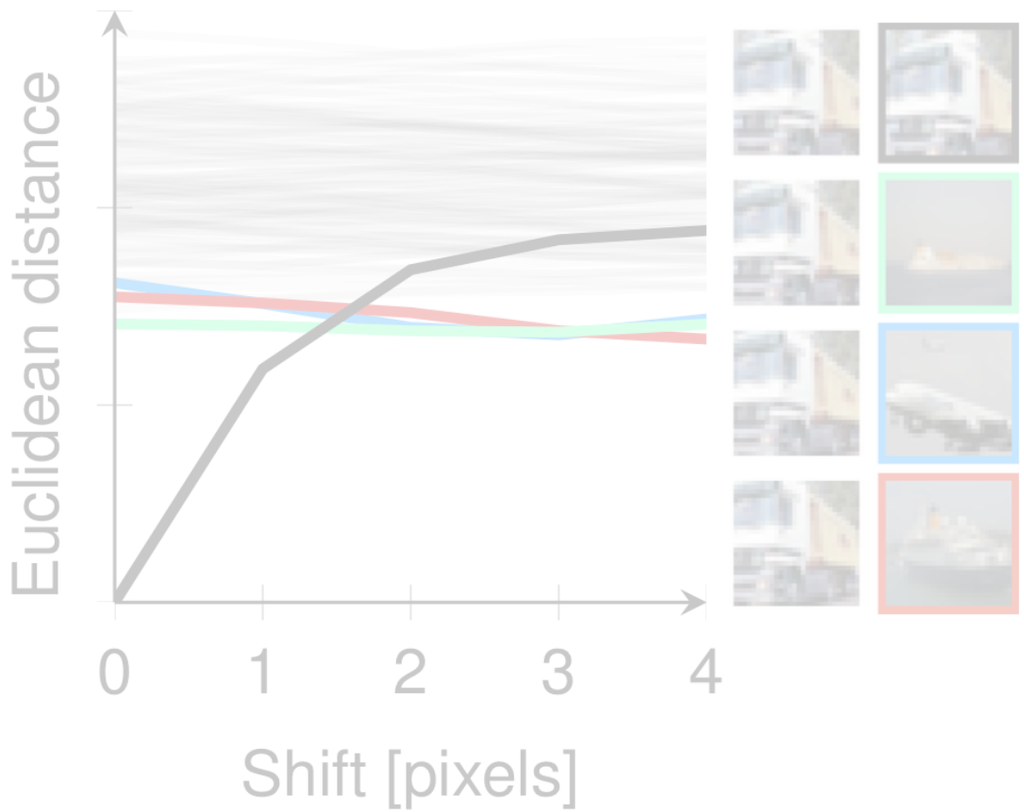
Evaluating generative models

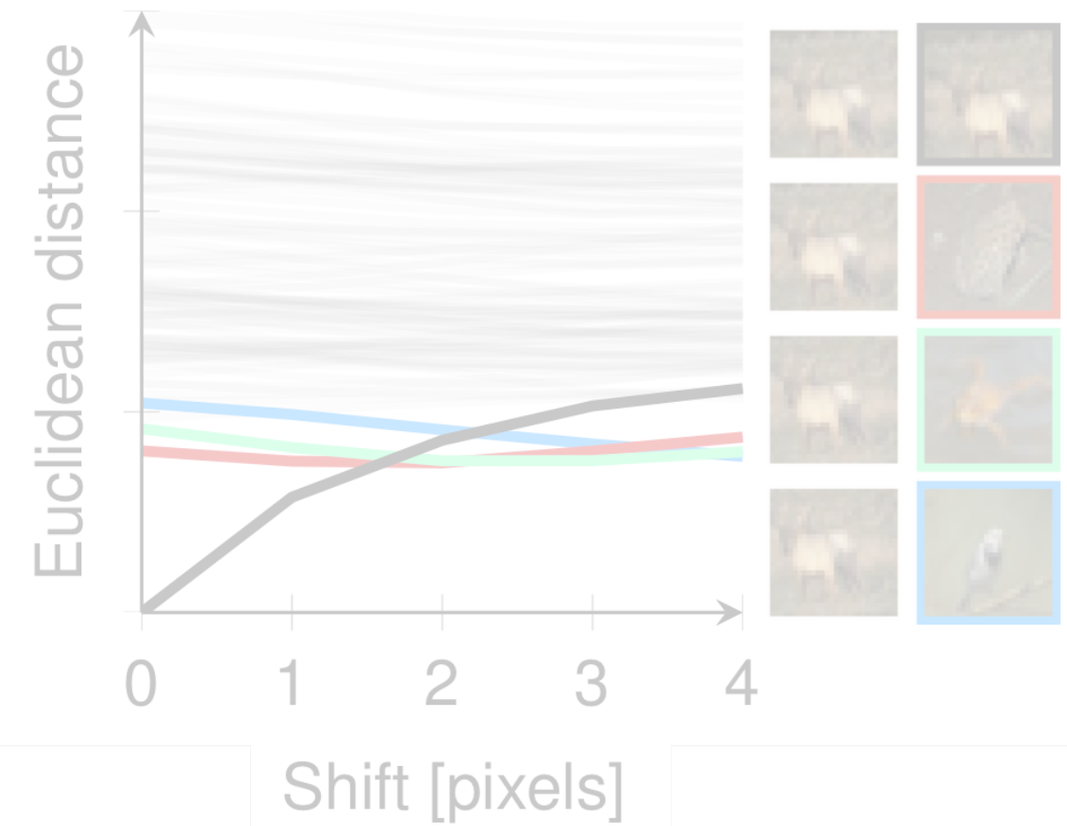
Sample plausibility



$$q(\mathbf{x}) = \frac{1}{N} \sum_n \delta(\mathbf{x} - \mathbf{x}_n)$$

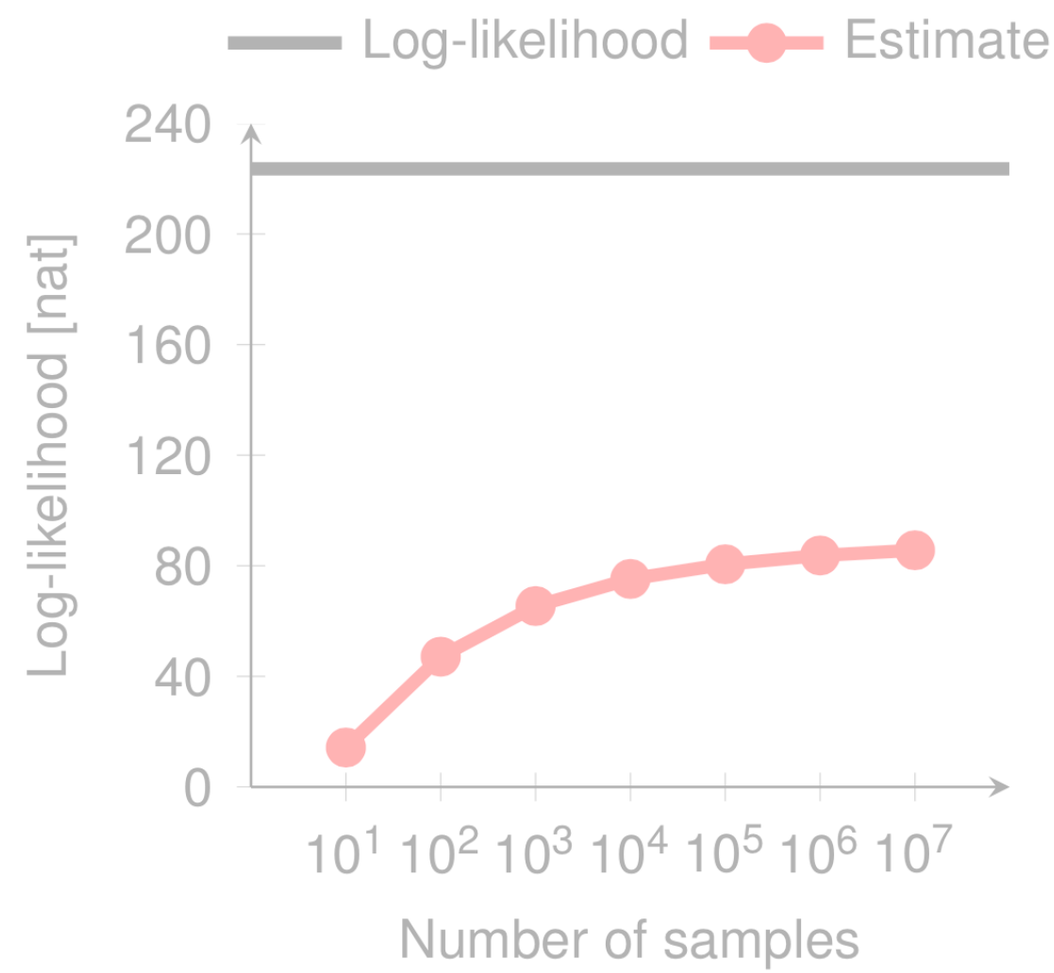
$$\mathrm{argmin}_i d(\mathbf{x}_i, \mathbf{s})$$





Parzen window estimates

$$q(\mathbf{x}) = \frac{1}{N} \sum_n \mathcal{N}(\mathbf{x}; \mathbf{x}_n, \varepsilon^2 \mathbf{I})$$



Parzen window estimates

Model	Parzen est. [nat]
Stacked CAE	121
DBN	138
GMMN	147
Deep GSN	214
Diffusion	220
GAN	225
True distribution	**243**
GMMN + AE	282
k-means	313

(see also Sohl-Dickstein et al., 2015)

Log-likelihood

$$\mathbf{y} = \mathbf{x} + \mathbf{u}, \quad \mathbf{u} \sim U([0,1[^D)$$

$$Q(\mathbf{x}) = \int_{[0,1[^D} q(\mathbf{x} + \mathbf{u})\, d\mathbf{u}$$

$$-\int p(\mathbf{y}) \log q(\mathbf{y})\, d\mathbf{y} \geq \underbrace{-\sum_{\mathbf{x}} P(\mathbf{x}) \log Q(\mathbf{x})}_{\text{Compression performance (bits/nats)}}$$

(see also Uria et al., 2013)

Samples and log-likelihood

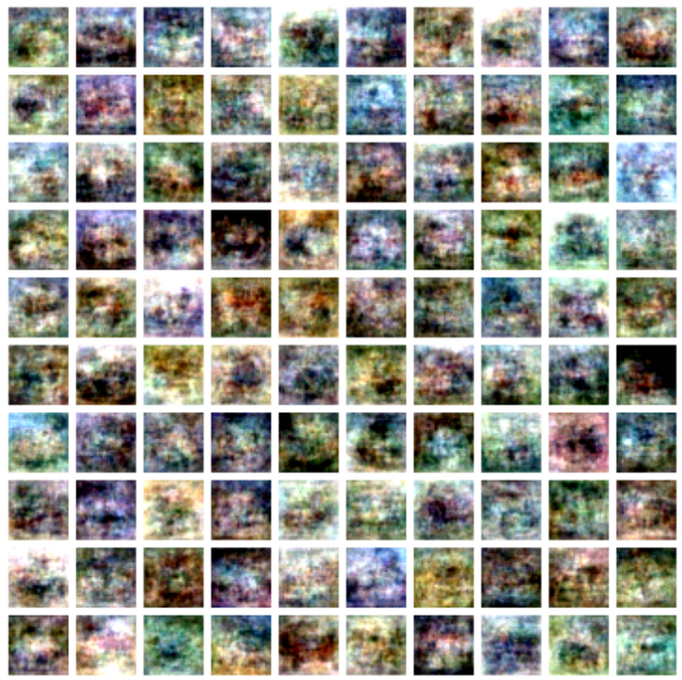
$$\log\left[0.01p(\mathbf{x}) + 0.99q(\mathbf{x})\right] \geq \log\left[0.01p(\mathbf{x})\right]$$
$$= \log p(\mathbf{x}) - \log 100$$
$$\approx \log p(\mathbf{x}) - 4.61$$

(van den Oord and Dambre, 2015)

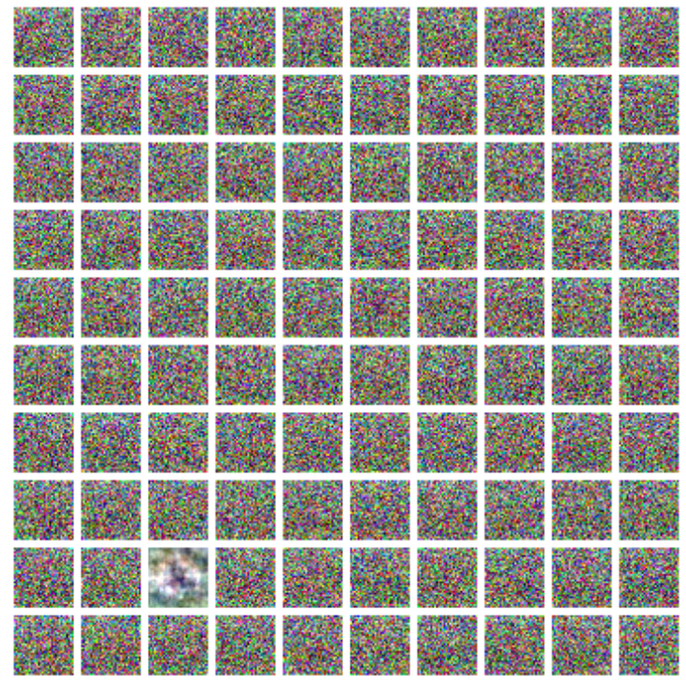
99%

Gaussian | Isotropic Gaussian | Lookup table

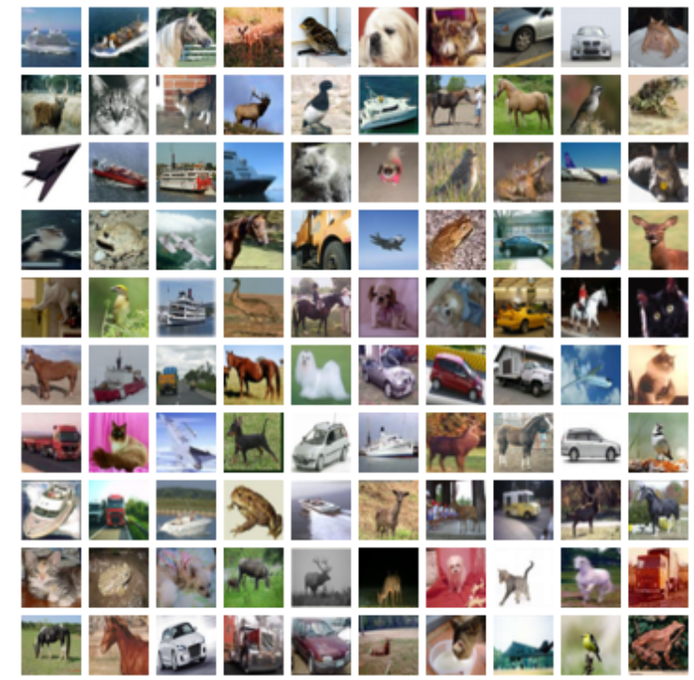
Gaussian



18,613 (bits/image)



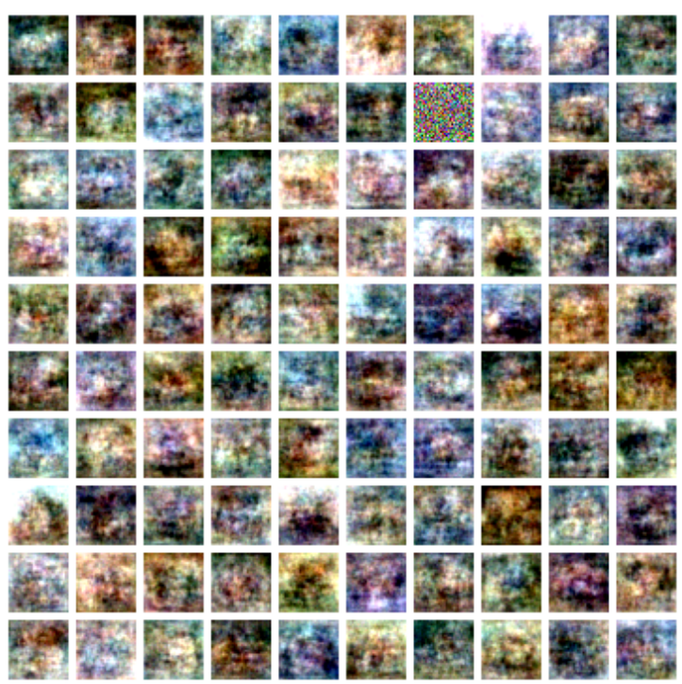
18,620 (bits/image)



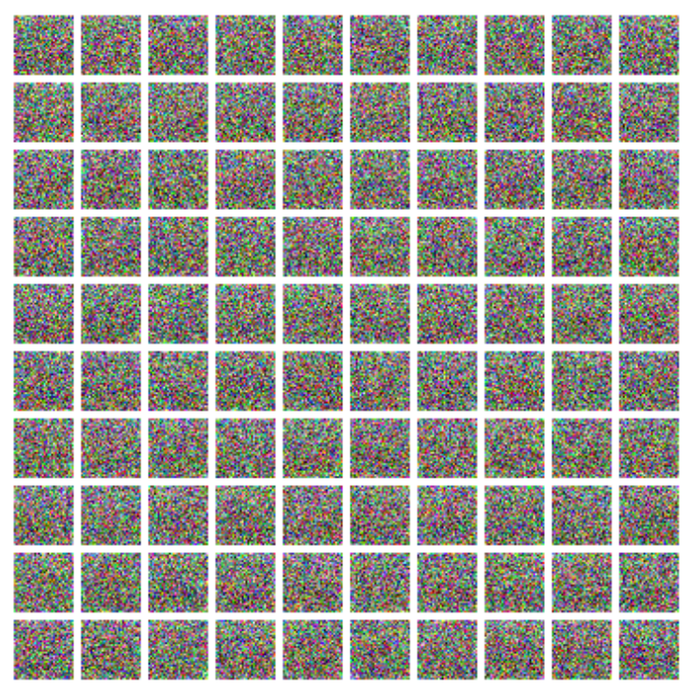
18,620 (bits/image)

1%

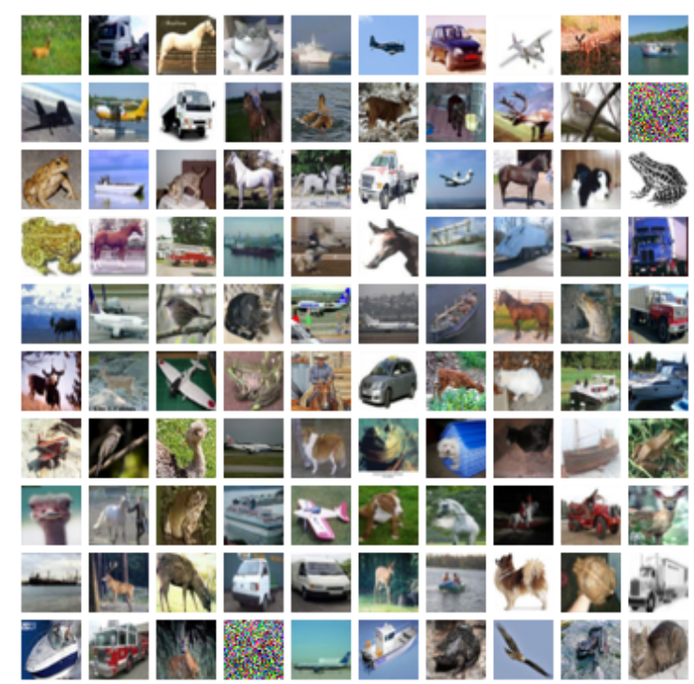
Isotropic Gaussian



18,620 (bits/image)



29,557 (bits/image)



29,564 (bits/image)

Samples and applications

$$P(y, \mathbf{x}) = 0.01 p(\mathbf{x}) p(y \mid \mathbf{x}) + 0.99 q(\mathbf{x}) q(y \mid \mathbf{x})$$

$$\ln p(\mathbf{x}) \gg \ln q(\mathbf{x})$$

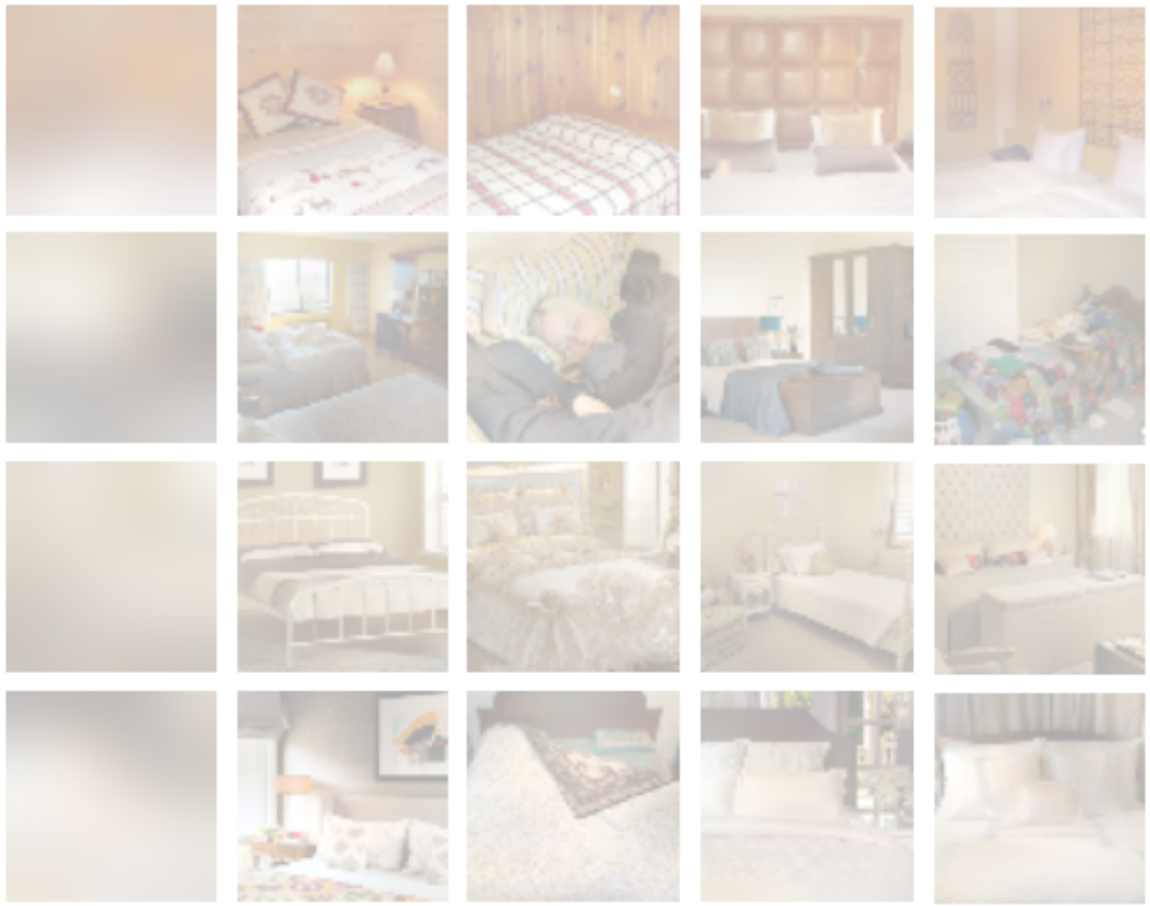
$$P(y \mid \mathbf{x}) = \alpha p(y \mid \mathbf{x}) + (1 - \alpha) q(y \mid \mathbf{x}) \approx p(y \mid \mathbf{x})$$

Conclusions

- **Evaluate generative models on applications**
  - **Compression:** log-likelihood
  - **Content gen./rec.:** samples/psychophysics (e.g. H. Gerhard et al., 2013)
  - **Unsupervised feature learning:** supervised task (e.g., classification)

- Recommendations:
  - Do not use Parzen window estimates
  - Do not rely on nearest neighbors to test for overfitting
  - Use samples where relevant for application and as a diagnostic tool but do not rely on it as a proxy for other tasks

LAPGAN (Denton et al., 2015)



LSUN (bedroom)

Since 1699, when French explorers landed at the great bend of the Mississippi River and celebrated the first Mardi Gras in North America, New Orleans has brewed a fascinating melange of cultures. It was French, then Spanish, then French again, then sold to the United States. Through all these years, and even into the 1900s, others arrived from everywhere: Acadians (Cajuns), Africans, indige-