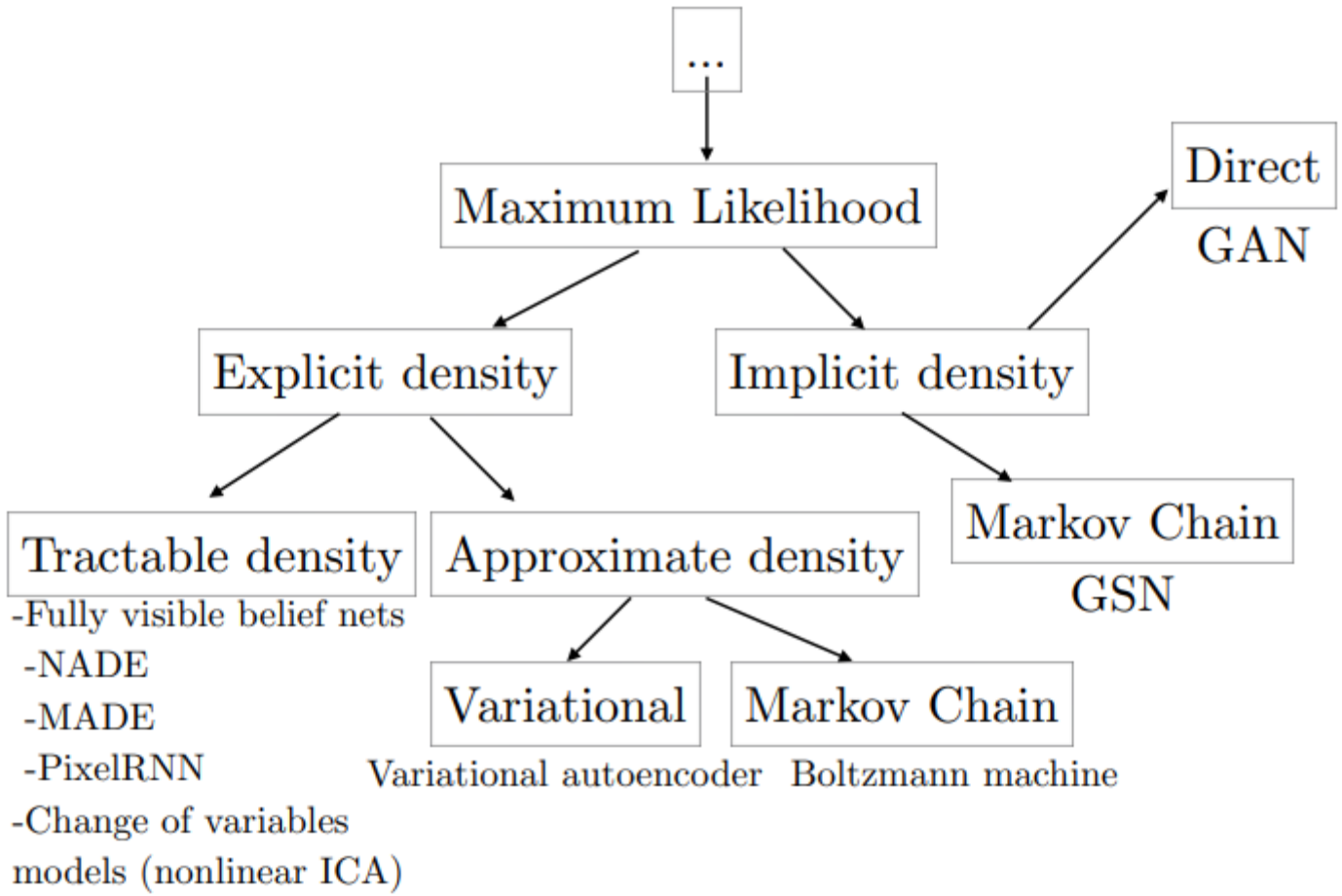
Generative AI Models
ECE 598 LV – Lecture 9

Lav R. Varshney

17 February 2022



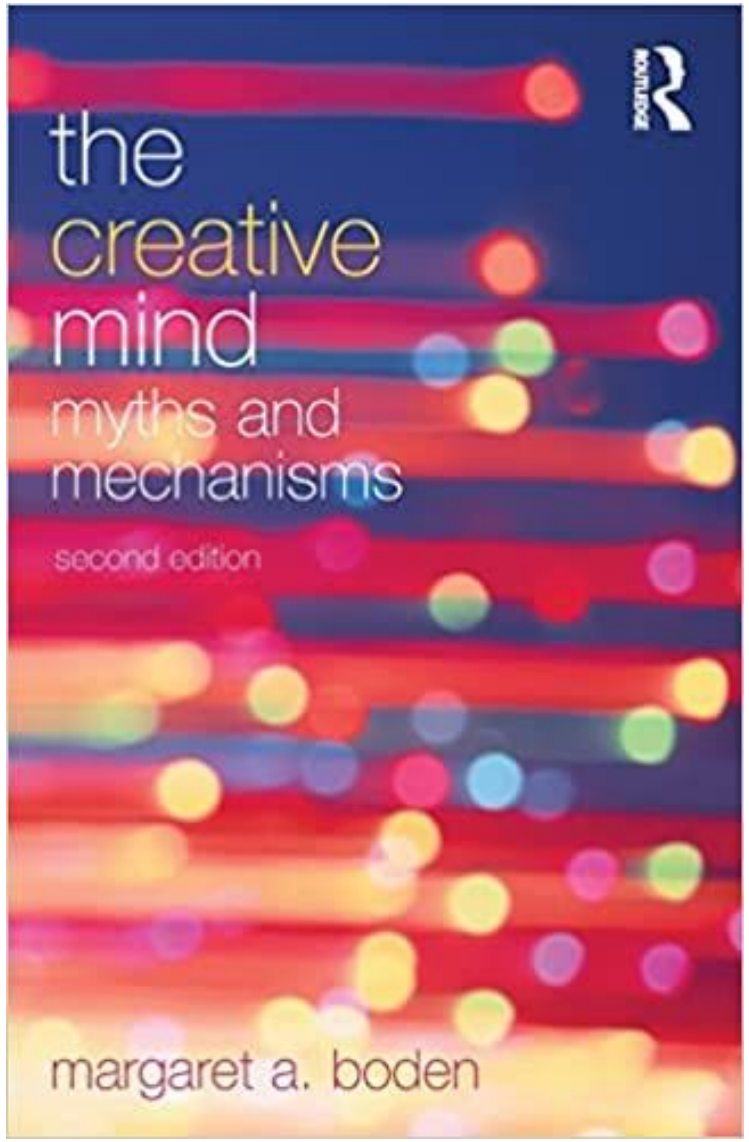
https://arxiv.org/pdf/1701.00160.pdf

https://thispersondoesnotexist.com/

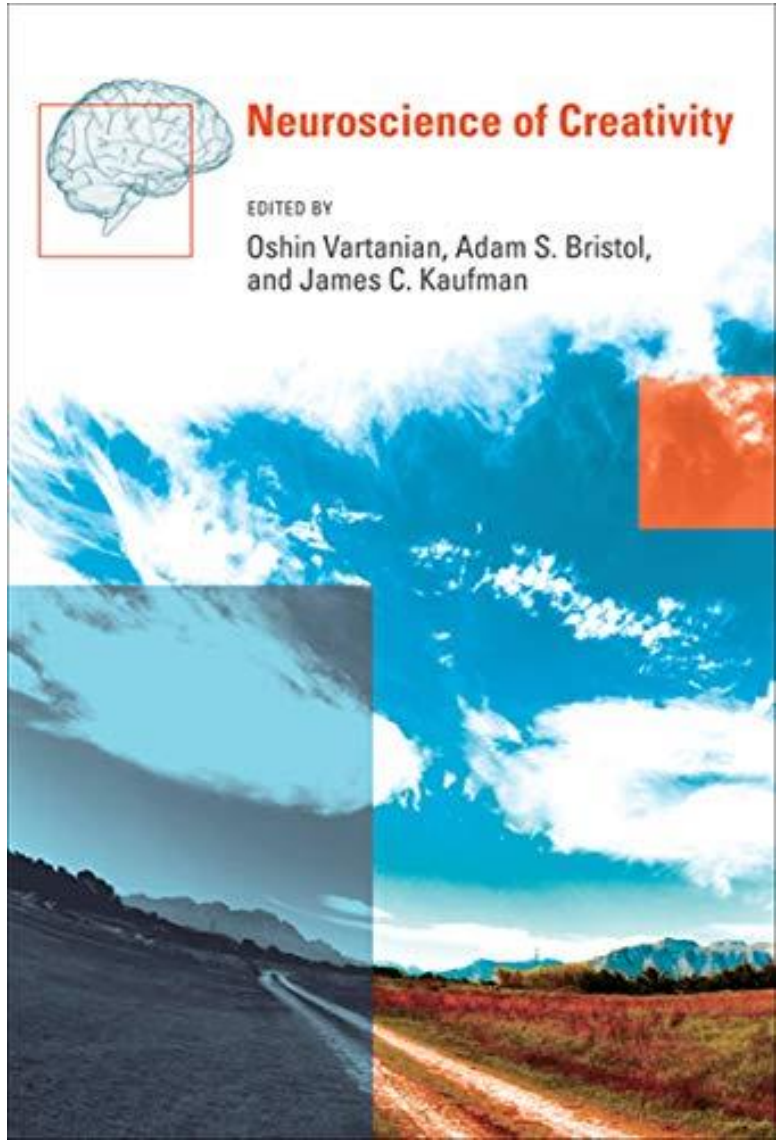
Mathematical limit theorems for computational creativity

L. R. Varshney

Creativity is the generation of an idea or artifact judged to be novel and high-quality by a knowledgeable social group, and is often said to be the pinnacle of intelligence. Several computational creativity systems of various designs are now being demonstrated and deployed. These myriad design possibilities raise the natural question: Are there fundamental limits to creativity? Here, we define a mathematical abstraction to capture key aspects of combinatorial creativity and study fundamental tradeoffs between novelty and quality. The functional form of this fundamental limit resembles the capacity-cost relationship in information theory, especially when measuring novelty using Bayesian surprise—the relative entropy between the empirical distribution of an inspiration set and that set updated with the new idea or artifact. As such, we show how information geometry techniques provide insight into the limits of creativity and find that the maturity of the creative domain directly parameterizes the fundamental limit. This result is extended to the case when there is a diverse audience for creativity and when the quality function is not known but must be estimated from samples.


the
creative
mind
myths and
mechanisms
second edition
margaret a. boden


R. KEITH SAWYER
the
SCIENCE
of
HUMAN INNOVATION
Explaining Creativity
SECOND EDITION
OXFORD


Neuroscience of Creativity
EDITED BY
Oshin Vartanian, Adam S. Bristol,
and James C. Kaufman

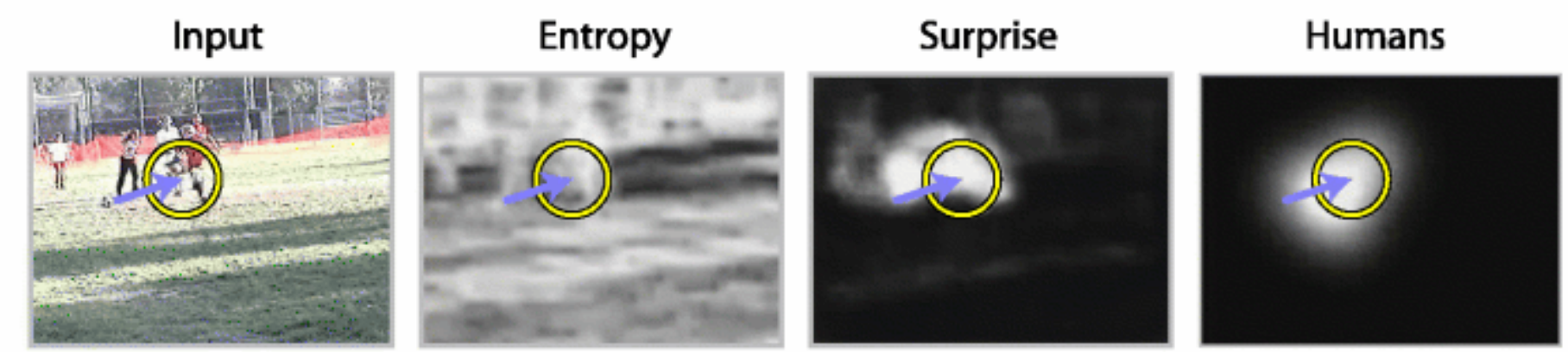
Towards a formalism

Creativity is the generation of an **artifact** that is judged to be **novel** and also to be appropriate, **useful**, or valuable by a suitably **knowledge**able social group.

Bayesian surprise as novelty

$$S(R,\mathcal{B}) = D\left(P_{B|R}||P_B\right) = \int_{\mathcal{B}} P_{B|R} \log \frac{P_{B|R}}{P_B} dB$$

- newly created recipe
- personalized repository of prior food experience
- posterior beliefs
- prior beliefs



[Itti and Baldi, 2006]

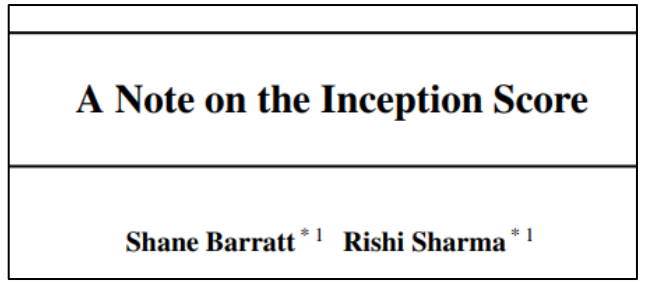
Basic Tradeoff in Creativity: Average Case

Novelty-Quality tradeoff in Creativity

$$S(Q) = \max_{P_A(\alpha):E[q(A)]\geq Q} E[s(A,\Theta)]$$

Lemma [Varshney, 2013]

$$E[s(A,\Theta)] = I(A,\Theta).$$


A Note on the Inception Score

Shane Barratt [*1] Rishi Sharma [*1]

Corollary

$$S(Q) = \max_{P_A(\alpha):E[q(A)]\geq Q} I(A,\Theta)$$

(Shannon's capacity-cost function)

2.1. Desiderata

Before delving into the explanation of evaluation measures, first I list a number of desired properties that an efficient GAN evaluation measure should fulfill. These properties can serve as meta measures to evaluate and compare the GAN evaluation measures. Here, I emphasize on the qualitative aspects of these measures. As will be discussed in Section 3, some recent works have attempted to compare the meta measures quantitatively (*e.g.* computational complexity of a measure). An efficient GAN evaluation measure should:

1. favor models that generate high fidelity samples (*i.e.* ability to distinguish generated samples from real ones; discriminability),
2. favor models that generate diverse samples (and thus is sensitive to over-fitting, mode collapse and mode drop, and can undermine trivial models such as the memory GAN),
3. favor models with disentangled latent spaces as well as space continuity (*a.k.a* controllable sampling),
4. have well-defined bounds (lower, upper, and chance),
5. be sensitive to image distortions and transformations. GANs are often applied to image datasets where certain transformations to the input do not change semantic meanings. Thus, an ideal measure should be invariant to such transformations. For instance, score of a generator trained on CelebA face dataset should not change much if its generated faces are shifted by a few pixels or rotated by a small angle.
6. agree with human perceptual judgments and human rankings of models, and
7. have low sample and computational complexity.

In what follows, GAN measures will be discussed and assessed with respect to the above desiderata, and a summary will be presented eventually in Section 3. See Table 2.

https://arxiv.org/pdf/1802.03446.pdf

	Measure	Description
Quantitative	1. Average Log-likelihood [18, 22]	• Log likelihood of explaining realworld held out/test data using a density estimated from the generated data (*e.g.* using KDE or Parzen window estimation). $L = \frac{1}{N}\sum_i \log P_{model}(\mathbf{x}_i)$
	2. Coverage Metric [33]	• The probability mass of the true data "covered" by the model distribution $C := P_{data}(dP_{model} > t)$ with t such that $P_{model}(dP_{model} > t) = 0.95$
	3. Inception Score (IS) [3]	• KLD between conditional and marginal label distributions over generated data. $\exp\left(\mathbb{E}_{\mathbf{x}}\left[\mathbb{KL}\left(p\left(\mathrm{y} \mid \mathbf{x}\right) \parallel p\left(\mathrm{y}\right)\right)\right]\right)$
	4. Modified Inception Score (m-IS) [34]	• Encourages diversity within images sampled from a particular category. $\exp(\mathbb{E}_{\mathbf{x}_i}[\mathbb{E}_{\mathbf{x}_j}[(\mathbb{KL}(P(y\|\mathbf{x}_i)\|P(y\|\mathbf{x}_j))]])$
	5. Mode Score (MS) [35]	• Similar to IS but also takes into account the prior distribution of the labels over real data. $\exp\left(\mathbb{E}_{\mathbf{x}}\left[\mathbb{KL}\left(p\left(y \mid \mathbf{x}\right) \parallel p\left(y^{train}\right)\right)\right] - \mathbb{KL}\left(p\left(y\right) \parallel p\left(y^{train}\right)\right)\right)$
	6. AM Score [36]	• Takes into account the KLD between distributions of training labels vs. predicted labels, as well as the entropy of predictions. $\mathbb{KL}(p(y^{\text{train}}) \parallel p(y)) + \mathbb{E}_{\mathbf{x}}\left[H(y\|\mathbf{x})\right]$
	7. Fréchet Inception Distance (FID) [37]	• Wasserstein-2 distance between multi-variate Gaussians fitted to data embedded into a feature space $FID(r,g) = \|\mu_r - \mu_g\|_2^2 + Tr(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{\frac{1}{2}})$
	8. Maximum Mean Discrepancy (MMD) [38]	• Measures the dissimilarity between two probability distributions P_r and P_g using samples drawn independently from each distribution. $M_k(P_r, P_g) = \mathbb{E}_{\mathbf{x},\mathbf{x}'\sim P_r}[k(\mathbf{x},\mathbf{x}')] - 2\mathbb{E}_{\mathbf{x}\sim P_r,\boldsymbol{y}\sim P_g}[k(\mathbf{x},\boldsymbol{y})] + \mathbb{E}_{\boldsymbol{y},\boldsymbol{y}'\sim P_g}[k(\boldsymbol{y},\boldsymbol{y}')]$
	9. The Wasserstein Critic [39]	• The critic (*e.g.* an NN) is trained to produce high values at real samples and low values at generated samples $\hat{W}(\mathbf{x}_{test}, \mathbf{x}_g) = \frac{1}{N}\sum_{i=1}^{N} \hat{f}(\mathbf{x}_{test}[i]) - \frac{1}{N}\sum_{i=1}^{N} \hat{f}(\mathbf{x}_g[i])$
	10. Birthday Paradox Test [27]	• Measures the support size of a discrete (continuous) distribution by counting the duplicates (near duplicates)
	11. Classifier Two Sample Test (C2ST) [40]	• Answers whether two samples are drawn from the same distribution (*e.g.* by training a binary classifier)
	12. Classification Performance [1, 15]	• An indirect technique for evaluating the quality of unsupervised representations (*e.g.* feature extraction; FCN score). See also the GAN Quality Index (GQI) [41].
	13. Boundary Distortion [42]	• Measures diversity of generated samples and covariate shift using classification methods.
	14. Number of Statistically-Different Bins (NDB) [43]	• Given two sets of samples from the same distribution, the number of samples that fall into a given bin should be the same up to sampling noise
	15. Image Retrieval Performance [44]	• Measures the distributions of distances to the nearest neighbors of some query images (*i.e.* diversity)
	16. Generative Adversarial Metric (GAM) [31]	• Compares two GANs by having them engaged in a battle against each other by swapping discriminators or generators. $p(\mathbf{x}\|y=1;M_1^{\cdot})/p(\mathbf{x}\|y=1;M_2^{\cdot}) = \left(p(y=1\|\mathbf{x};D_1)p(\mathbf{x};G_2)\right)/\left(p(y=1\|\mathbf{x};D_2)p(\mathbf{x};G_1)\right)$
	17. Tournament Win Rate and Skill Rating [45]	• Implements a tournament in which a player is either a discriminator that attempts to distinguish between real and fake data or a generator that attempts to fool the discriminators into accepting fake data as real.
	18. Normalized Relative Discriminative Score (NRDS) [32]	• Compares n GANs based on the idea that if the generated samples are closer to real ones, more epochs would be needed to distinguish them from real samples.
	19. Adversarial Accuracy and Divergence [46]	• Adversarial Accuracy. Computes the classification accuracies achieved by the two classifiers, one trained on real data and another on generated data, on a labeled validation set to approximate $P_g(y\|\mathbf{x})$ and $P_r(y\|\mathbf{x})$. Adversarial Divergence: Computes $\mathbb{KL}(P_g(y\|\mathbf{x}), P_r(y\|\mathbf{x}))$
	20. Geometry Score [47]	• Compares geometrical properties of the underlying data manifold between real and generated data.
	21. Reconstruction Error [48]	• Measures the reconstruction error (*e.g.* L_2 norm) between a test image and its closest generated image by optimizing for z (*i.e.* $min_{\mathbf{z}}\|G(\mathbf{z}) - \mathbf{x}^{(test)}\|^2$)
	22. Image Quality Measures [49, 50, 51]	• Evaluates the quality of generated images using measures such as SSIM, PSNR, and sharpness difference
	23. Low-level Image Statistics [52, 53]	• Evaluates how similar low-level statistics of generated images are to those of natural scenes in terms of mean power spectrum, distribution of random filter responses, contrast distribution, etc.
	24. Precision, Recall and F_1 score [23]	• These measures are used to quantify the degree of overfitting in GANs, often over toy datasets

Qualitative	1. Nearest Neighbors	• To detect overfitting, generated samples are shown next to their nearest neighbors in the training set
	2. Rapid Scene Categorization [18]	• In these experiments, participants are asked to distinguish generated samples from real images in a short presentation time (*e.g.* 100 ms); *i.e.* real v.s fake
	3. Preference Judgment [54, 55, 56, 57]	• Participants are asked to rank models in terms of the fidelity of their generated images (*e.g.* pairs, triples)
	4. Mode Drop and Collapse [58, 59]	• Over datasets with known modes (*e.g.* a GMM or a labeled dataset), modes are computed as by measuring the distances of generated data to mode centers
	5. Network Internals [1, 60, 61, 62, 63, 64]	• Regards exploring and illustrating the internal representation and dynamics of models (*e.g.* space continuity) as well as visualizing learned features

3. **Inception Score (IS).** Proposed by Salimans *et al.* [3], it is perhaps the most widely adopted score for GAN evaluation (*e.g.* in [67]). It uses a pre-trained neural network (the Inception Net [68] trained on the ImageNet [69]) to capture the desirable properties of generated samples: *highly classifiable* and *diverse* with respect to class labels. It measures the average KL divergence between the conditional label distribution $p(y|\mathbf{x})$ of samples (expected to have low entropy for easily classifiable samples; better sample quality) and the marginal distribution $p(y)$ obtained from all the samples (expected to have high entropy if all classes are equally represented in the set of samples; high diversity). It favors low entropy of $p(y|\mathbf{x})$ but a large entropy of $p(y)$.

$$\exp\left(\mathbb{E}_{\mathbf{x}}\left[\mathbb{KL}\left(p\left(\mathrm{y}\mid\mathbf{x}\right)\parallel p\left(\mathrm{y}\right)\right)\right]\right)=\exp\left(H(y)-\mathbb{E}_{\mathbf{x}}\left[H(y|\mathbf{x})\right]\right), \quad (1)$$

where $p\left(\mathrm{y}\mid\mathbf{x}\right)$ is the conditional label distribution for image $\mathbf{x}$ estimated using a pretrained Inception model [68], and $p\left(\mathrm{y}\right)$ is the marginal distribution: $p\left(\mathrm{y}\right)\approx{}^{1}/_{\mathrm{N}}\sum_{n=1}^{\mathrm{N}}p\left(\mathrm{y}\mid\mathbf{x}_{n}=G\left(\mathbf{z}_{n}\right)\right)$. $H(\mathbf{x})$ represents entropy of variable $\mathbf{x}$.

The Inception score shows a reasonable correlation with the quality and diversity of generated images [3]. IS over real images can serve as the upper bound. Despite these appealing properties, IS has several limitations:

(a) First, similar to log-likelihood, it favors a "memory GAN" that stores all training samples, thus is unable to detect overfitting (*i.e.* can be fooled by generating centers of data modes [46]). This is aggravated by the fact that it does not make use of a holdout validation set.
(b) Second, it fails to detect whether a model has been trapped into one bad mode (*i.e.* is agnostic to mode collapse). Zhou *et al.* [36], however, shows results on the contrary.
(c) Third, since IS uses Inception model that has been trained on ImageNet with many object classes, it may favor models that generate good objects rather realistic images.
(d) Fourth, IS only considers P_g and ignores P_r. Manipulations such as mixing in natural images from an entirely different distribution could deceive this score. As a result, it may favor models that simply learn sharp and diversified images, instead of P_r [26][2].
(e) Fifth, it is an asymmetric measure.
(f) Finally, it is affected by image resolution. See Fig. 2.

Measure		Desiderata: Discriminability	Detecting Overfitting	Disentangled Latent Spaces	Well-defined Bounds	Perceptual Judgments	Sensitivity to Distortions	Comp. & Sample Efficiency
1. Average Log- likelihood	[18, 22]	low	low	-	[-∞, ∞]	low	low	low
2. Coverage Metric	[33]	low	low	-	[0, 1]	low	low	-
3. Inception Score (IS)	[3]	high	moderate	-	[1, ∞]	high	moderate	high
4. Modified Inception Score (m-IS)	[34]	high	moderate	-	[1, ∞]	high	moderate	high
5. Mode Score (MS)	[35]	high	moderate	-	[0, ∞]	high	moderate	high
6. AM Score	[36]	high	moderate	-	[0, ∞]	high	moderate	high
7. Fréchet Inception Distance (FID)	[37]	high	moderate	-	[0, ∞]	high	high	high
8. Maximum Mean Discrepancy (MMD)	[38]	high	low	-	[0, ∞]	-	-	-
9. The Wasserstein Critic	[39]	high	moderate	-	[0, ∞]	-	-	low
10. Birthday Paradox Test	[27]	low	high	-	[1, ∞]	low	low	-
11. Classifier Two Sample Test (C2ST)	[40]	high	low	-	[0, 1]	-	-	-
12. Classification Performance	[1, 15]	high	low	-	[0, 1]	low	-	-
13. Boundary Distortion	[42]	low	low	-	[0, 1]	-	-	-
14. NDB	[43]	low	high	-	[0, ∞]	-	low	-
15. Image Retrieval Performance	[44]	moderate	low	-	*	low	-	-
16. Generative Adversarial Metric (GAM)	[31]	high	low	-	*	-	-	moderate
17. Tournament Win Rate and Skill Rating	[45]	high	high	-	*	-	-	low
18. NRDS	[32]	high	low	-	[0, 1]	-	-	poor
19. Adversarial Accuracy & Divergence	[46]	high	low	-	[0, 1], [0, ∞]	-	-	-
20. Geometry Score	[47]	low	low	-	[0, ∞]	-	low	low
21. Reconstruction Error	[48]	low	low	-	[0, ∞]	-	moderate	moderate
22. Image Quality Measures	[49, 50, 51]	low	moderate	-	*	high	high	high
23. Low-level Image Statistics	[52, 53]	low	low	-	*	low	low	-
24. Precision, Recall and F_1 score	[23]	low	high	✓	[0, 1]	-	-	-

	Quantitative /Analysis /Optimization	Qualitative	Overfitting /Memorization	Latent Space Disentan-glement	Deepfake Detection
FID & IS Variants					
Spatial FID (sFID)	**Analysis**				
Class-aware FID (CAFD)	✓		✓		
Conditional FID	✓		✓		
Fast FID	**Optimization**				
Memorization-informed FID (MiFID)	✓		✓		
Unbiased FID and IS	✓				
Clean FID	**Analysis**				
Fre'chet Video Distance (FVD)	✓				
Methods based on Self-supervised Learned Representations	**Analysis**				
Methods based on Analysing Data Manifold					
Local Intrinsic Dimensionality (LID)	✓				
Intrinsic Multi-scale Distance (IMD)	✓				
Perceptual Path Length (PPL)	✓	✓		✓	
Linear Separability in Latent Space	✓			✓	
Classification Accuracy Score (CAS)	✓				
Non-Parametric Tests to Detect Data-Copying	✓		✓		
Measures that Probe Generalization	**Analysis**		✓		
New Ideas based on Precision and Recall (P&R)					
Density and Coverage	✓		✓		
Alpha Precision and Recall	✓		✓		
Duality GAP Metric	✓				
Spectral Methods	✓	✓			✓
Caption Score (CapS)	✓				
Human Eye Perceptual Evaluation (HYPE)		✓			
Neuroscore		✓			
GAN Steerability & Dissection		✓		✓	
A Universal Fake vs. Real Detector		✓			✓

https://arxiv.org/pdf/2103.09396.pdf