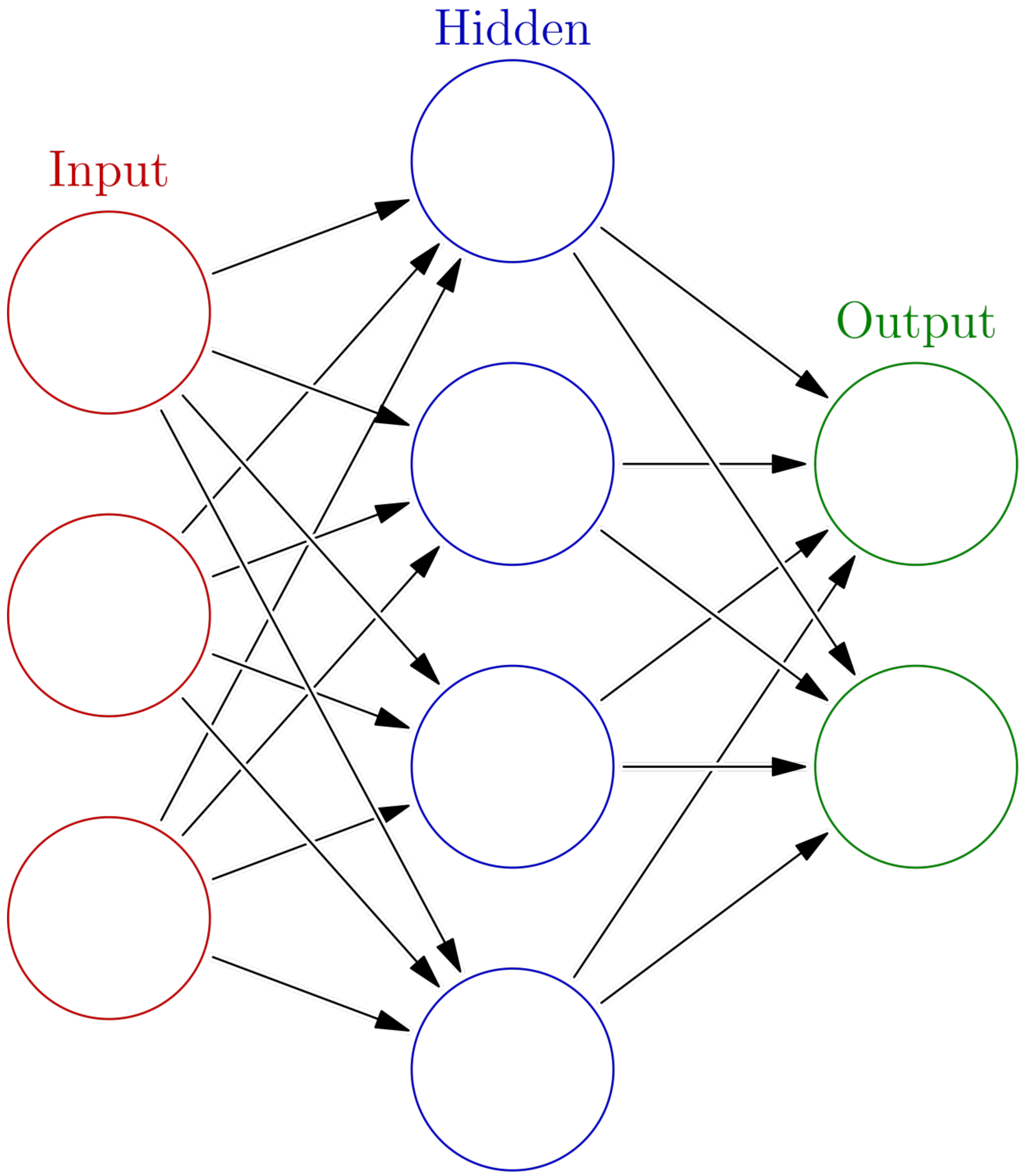
What is the INTRINSIC DIMENSION of Your Data?

University of Washington Seminar

Wenhao Chai

University of Washington Information Processing Lab Seminar	Seattle, WA
What is the Intrinsic Dimension of Your Data?	Jan 2025
Bridging the Parallel Decoding of LLMs with the Diffusion Process	Oct 2024
DPO and RLHF for Large Language Model Post-training	Apr 2024
Vision Representation Learning from Sythetic Data	Jan 2024
From Large Language Models to Large Multi-modal Models	Nov 2023

Intrinsic Dimension



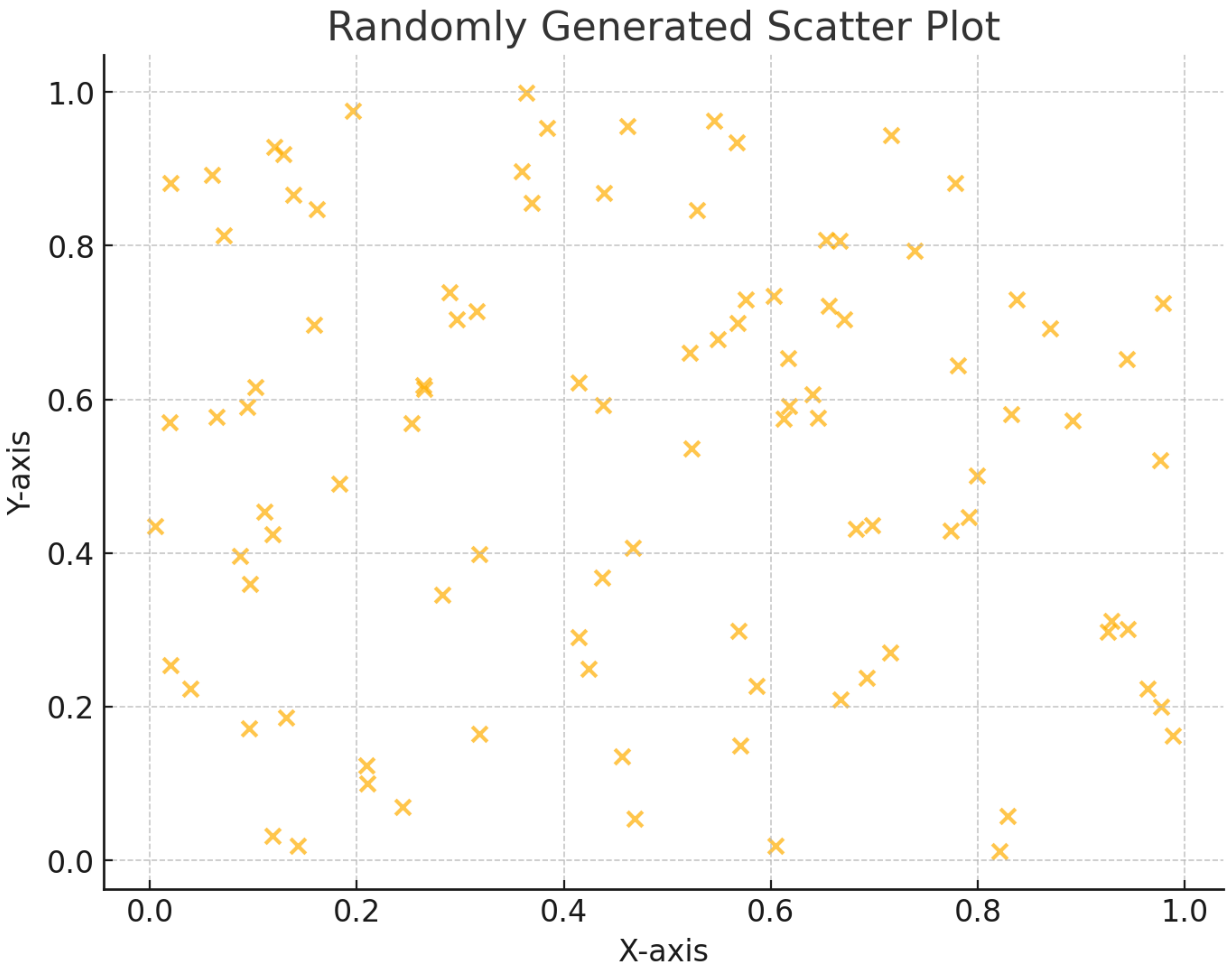
Hidden dim happened in NN design

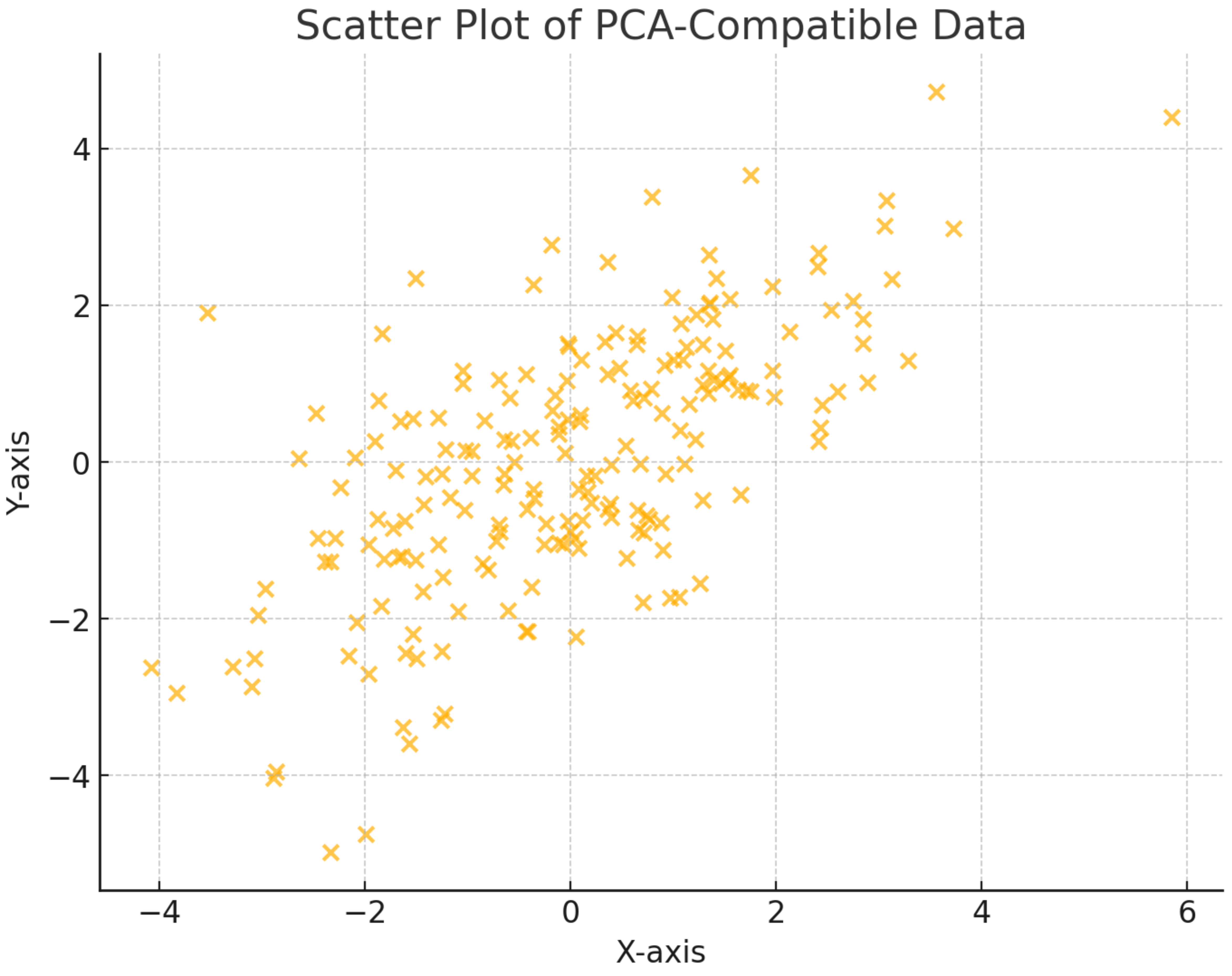
However, NN is overparameterized

The **intrinsic dimension** for a data set can be thought of as the number of variables needed in a minimal representation of the data

How to measure the intrinsic dimension of the original data?

Toy Data Example





More Complex Data Example

ImageNet

14 million images

more than 20,000 categories

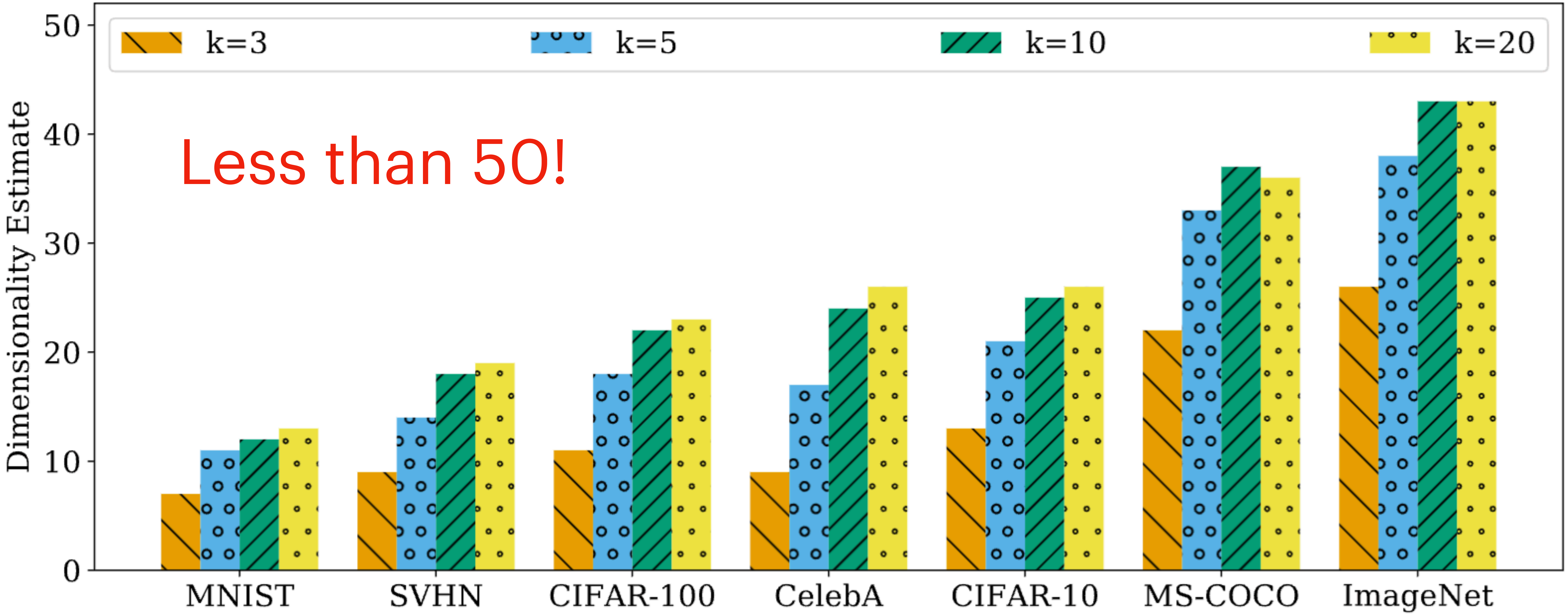
224 x 224 resolution

150,528 pixels per image

What is the intrinsic dimension of that?

Let's guess!

Intrinsic Dimension of Some Dataset



* k is a hyper-param they used

Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., & Goldstein, T. The Intrinsic Dimension of Images and Its Impact on Learning. In *International Conference on Learning Representations*.

NeurIPS 2004

Maximum Likelihood Estimation of Intrinsic Dimension

Elizaveta Levina
Department of Statistics
University of Michigan
Ann Arbor MI 48109-1092
elevina@umich.edu

Peter J. Bickel
Department of Statistics
University of California
Berkeley CA 94720-3860
bickel@stat.berkeley.edu

Published as a conference paper at ICLR 2021

THE INTRINSIC DIMENSION OF IMAGES AND ITS IMPACT ON LEARNING

Phillip Pope[1], Chen Zhu[1], Ahmed Abdelkader[2], Micah Goldblum[1], Tom Goldstein[1]
[1]Department of Computer Science, University of Maryland, College Park
[2]Oden Institute for Computational Engineering and Sciences, University of Texas at Austin
{pepope,chenzhu}@umd.edu, akader@utexas.edu, {goldblum,tomg}@umd.edu

Notation and Assumption

$P \subset \mathbb{R}^N$ data point

$M \subseteq \mathbb{R}^N$ manifold

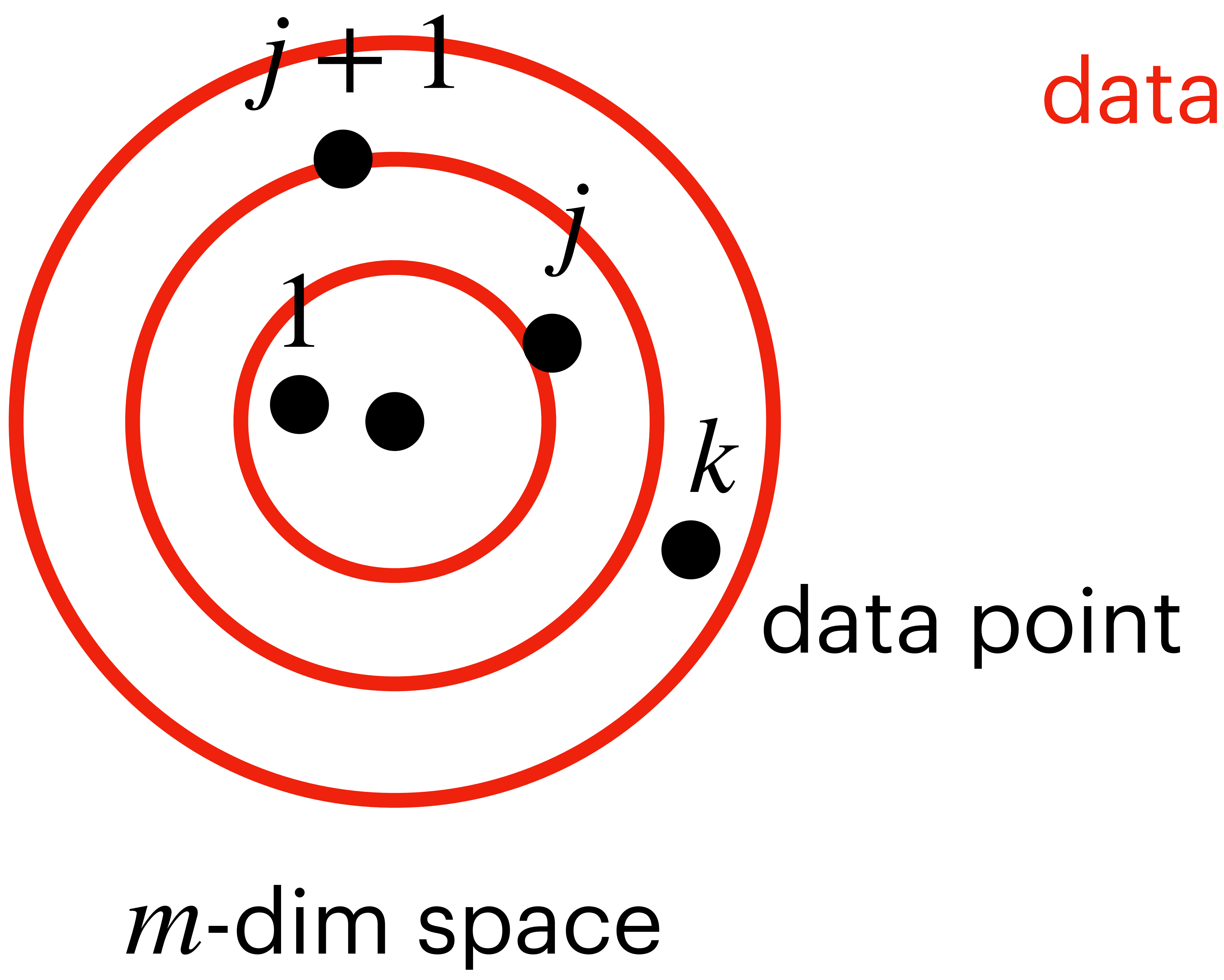
$m = \dim(M) \ll N$ intrinsic dimension

density is constant within small neighborhoods Local uniformity assumption

Before Math …

Find a relationship between some var and intrinsic dimension m

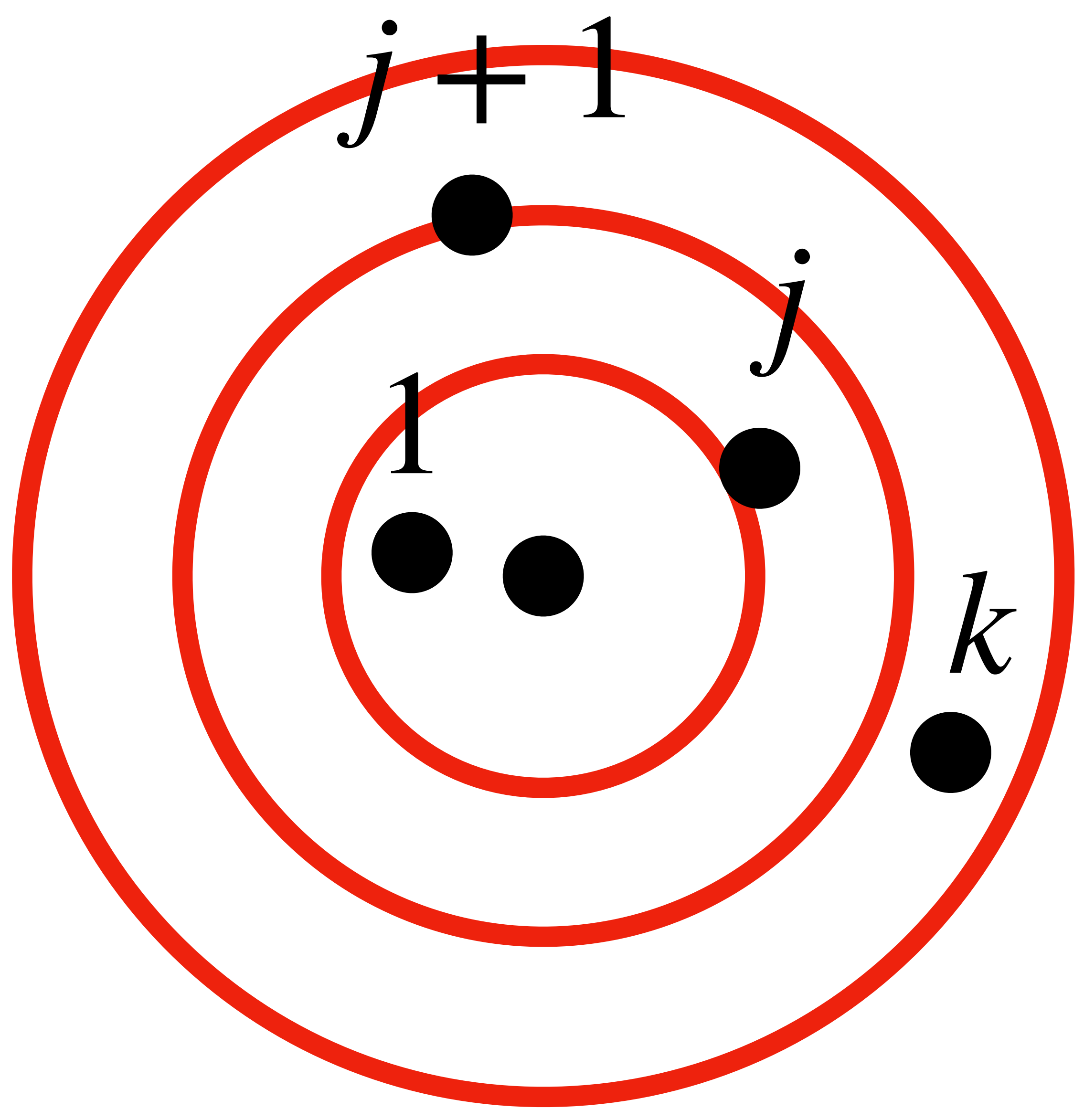
data density / distance



$$\mathbb{E}(\text{number of points}) = \rho V_m(r) \propto r^m$$

e.g. $V_2(r) = \pi r^2, V_3(r) = \dfrac{4}{3}\pi r^3$

Maximum Likelihood Estimation of Poisson Process



k-nearest neighbor

We observe $r_1, r_2, r_3, \ldots, r_k$

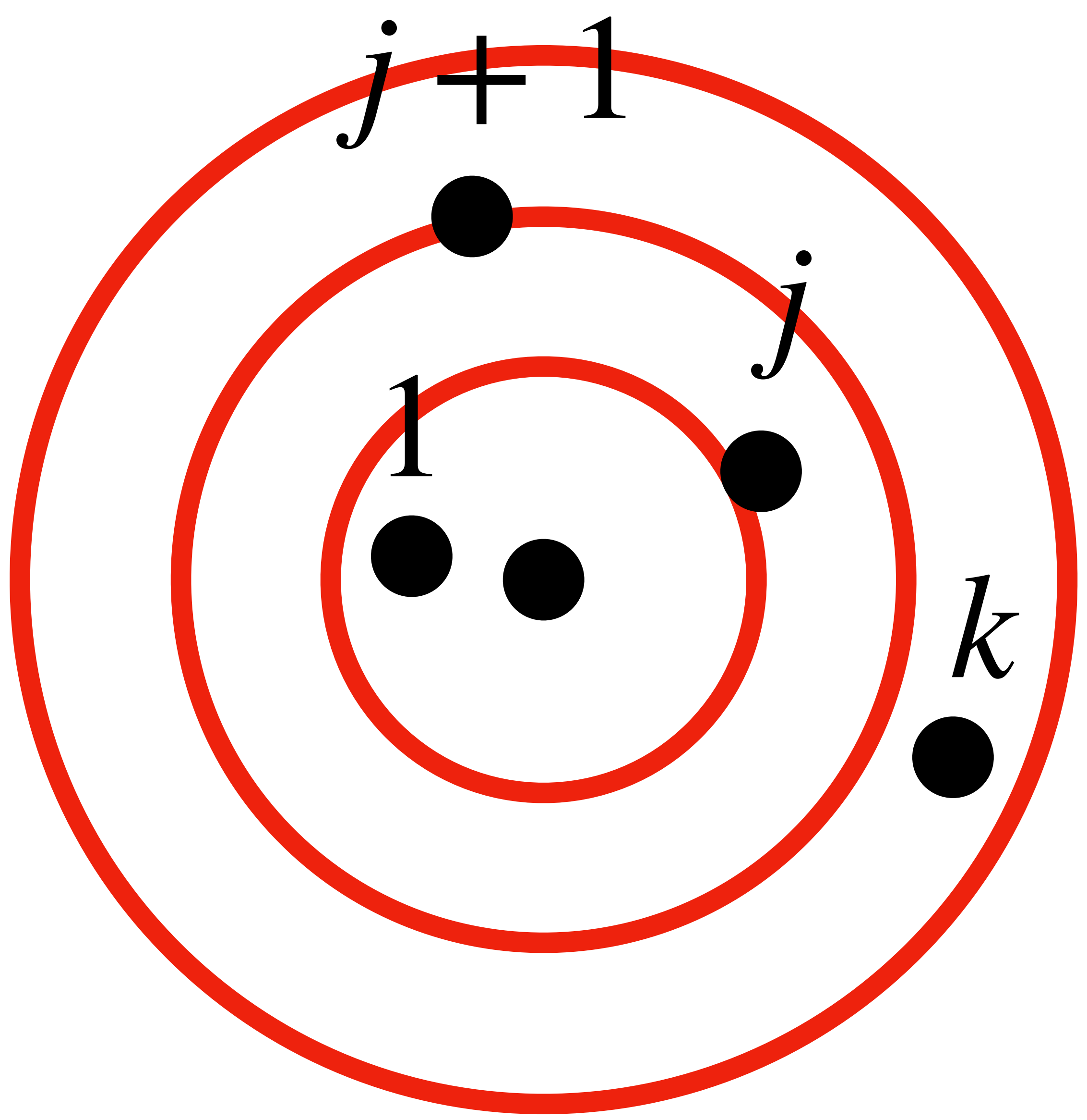
$$\mathbb{E}(\text{number of points}) = N(r) = \rho V_m(r) \propto r^m$$

$$\lambda(r) \propto \frac{d}{dr}\left[r^m\right] = m \cdot r^{m-1}$$

$$P(N(r)) \propto \exp\left(-\int_0^R \lambda(r)\, dr\right) \prod_j \lambda(r_j),$$

$$L(m) = \int_0^R \log \lambda(r) dN(r) - \int_0^R \lambda(r) dr$$

Maximum Likelihood Estimation of Poisson Process



k-nearest neighbor

$$L(m) = \int_0^R \log \lambda(r) dN(r) - \int_0^R \lambda(r) dr$$

$$\frac{\partial L}{\partial m} = 0$$

$$\hat{m} = \left[\frac{1}{k-1} \sum_{j=1}^{k-1} \log \frac{r_k}{r_j} \right]^{-1}$$

Validating Dimension Estimation with Synthetic Data

Prepare:

Pretrained BigGAN with 128-dim latent

set d when the others are 0 to make data

k	$\bar{d}$						
	2	4	8	16	32	64	128
3	1.1	2.6	6.1	10.5	16.0	20.0	20.0
4	1.5	3.6	8.2	14.0	21.0	26.0	26.0
5	1.7	4.1	9.3	15.7	23.5	28.7	28.5
6	1.8	4.4	9.9	16.6	24.9	30.3	29.9
7	1.9	4.6	10.4	17.2	25.8	31.2	30.6
8	1.9	4.7	10.7	17.6	26.4	31.7	31.1
9	2.0	4.9	10.9	18.0	26.8	31.9	31.5
10	2.0	5.0	11.1	18.2	27.1	32.1	31.7
15	2.1	5.3	11.6	18.8	27.8	32.3	31.7
20	2.2	5.5	11.8	19.0	27.9	31.9	31.3
25	2.2	5.7	12.0	19.2	27.9	31.5	30.8

GT

Why We Need to Know Intrinsic Dimension?

Measuring the difficulty in terms of classification

Dataset	MNIST	SVHN	CIFAR-100	CelebA	CIFAR-10	MS-COCO	ImageNet
MLE (k=3)	7	9	11	9	13	22	26
MLE (k=5)	11	14	18	17	21	33	38
MLE (k=10)	12	18	22	24	25	37	43
MLE (k=20)	13	19	23	26	26	36	43
SOTA Accuracy	99.84	99.01	93.51	-	99.37	-	88.55

Measuring the difficulty in terms of diffusion generation

Number of diffusion steps $= O(d)$

Linear Convergence of Diffusion Models Under the Manifold Hypothesis

Peter Potaptchik*, Iskander Azangulov*, and George Deligiannidis

University of Oxford
{surname}@stats.ox.ac.uk

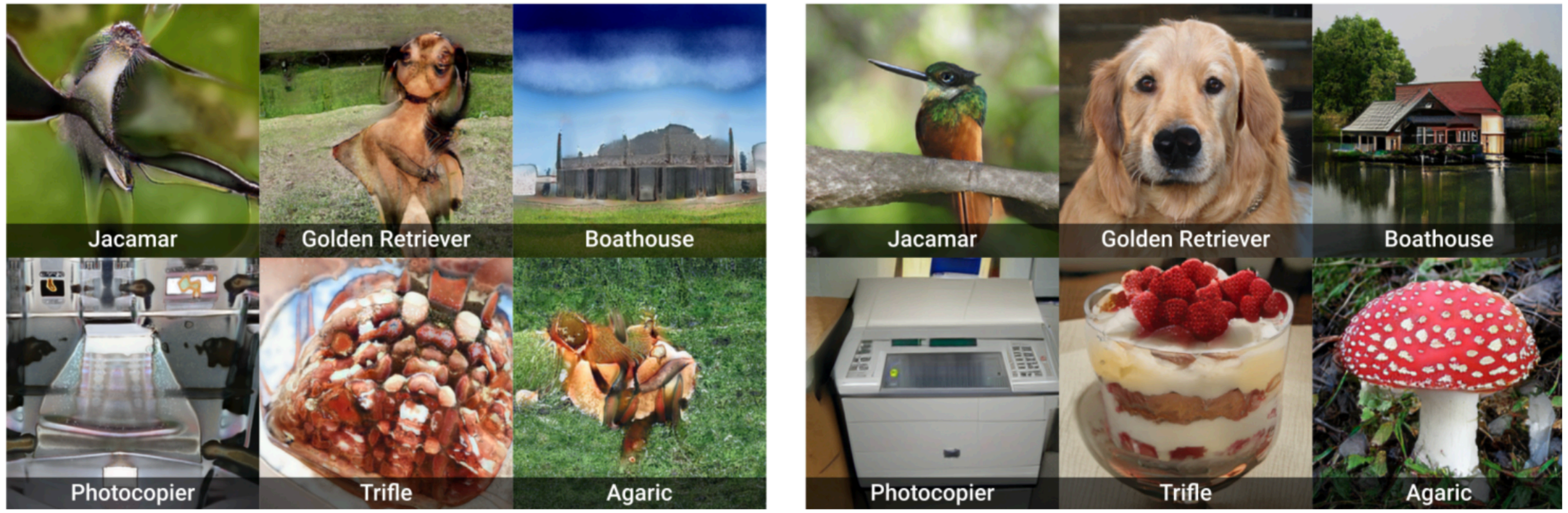
Why We Need to Know Intrinsic Dimension?

Guidance of GAN design

Accordingly, a latent code of size 512 is highly redundant, making the mapping network's task harder at the beginning of training. Consequently, the generator is slow to adapt and cannot benefit from Projected GAN's speed up. We therefore reduce StyleGAN's latent code z to 64

StyleGAN-XL: Scaling StyleGAN to Large Diverse Datasets

AXEL SAUER, KATJA SCHWARZ, and ANDREAS GEIGER
University of Tübingen and Max Planck Institute for Intelligent Systems, Tübingen, Germany

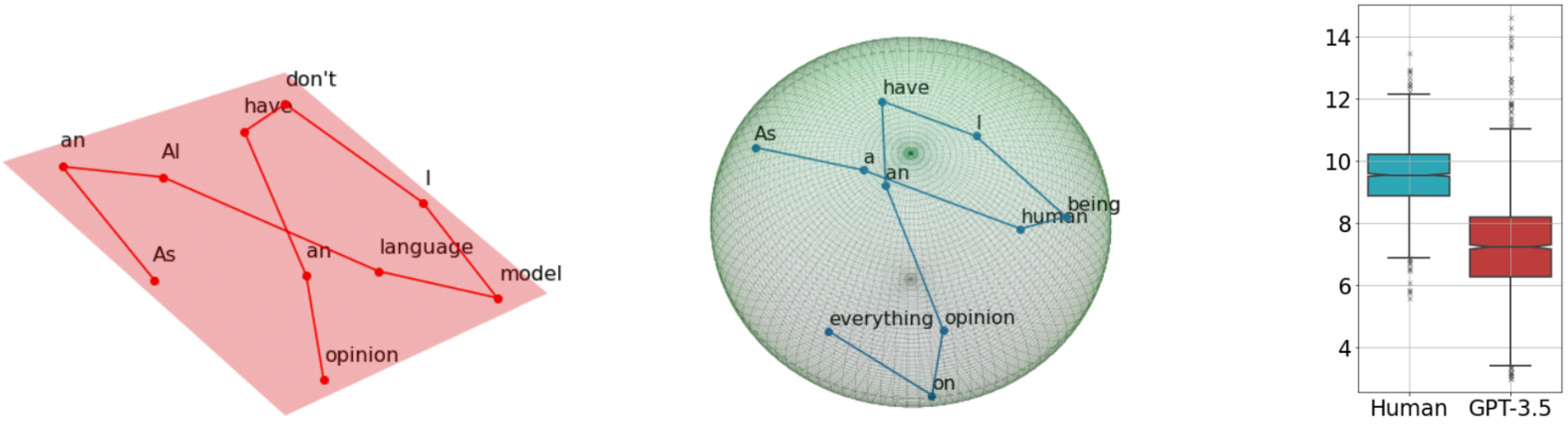


Why We Need to Know Intrinsic Dimension?

Detect AI-generated content

Intrinsic Dimension Estimation for Robust Detection of AI-Generated Texts

Eduard Tulchinskii[1], **Kristian Kuznetsov**[1], **Laida Kushnareva**[2], **Daniil Cherniavskii**[3], **Sergey Nikolenko**[5], **Evgeny Burnaev**[1,3], **Serguei Barannikov**[1,4], **Irina Piontkovskaya**[2]
[1]Skolkovo Institute of Science and Technology, Russia;
[2]AI Foundation and Algorithm Lab, Russia;
icial Intelligence Research Institute (AIRI), Russia;[4]CNRS, Université Paris Cité, France;
[5]St. Petersburg Department of the Steklov Institute of Mathematics, Russia



(a) AI generated (b) human written (c)