Open Mathematical Problems in Dimension Reduction Methods and Energy Landscapes

Antonio Auffinger

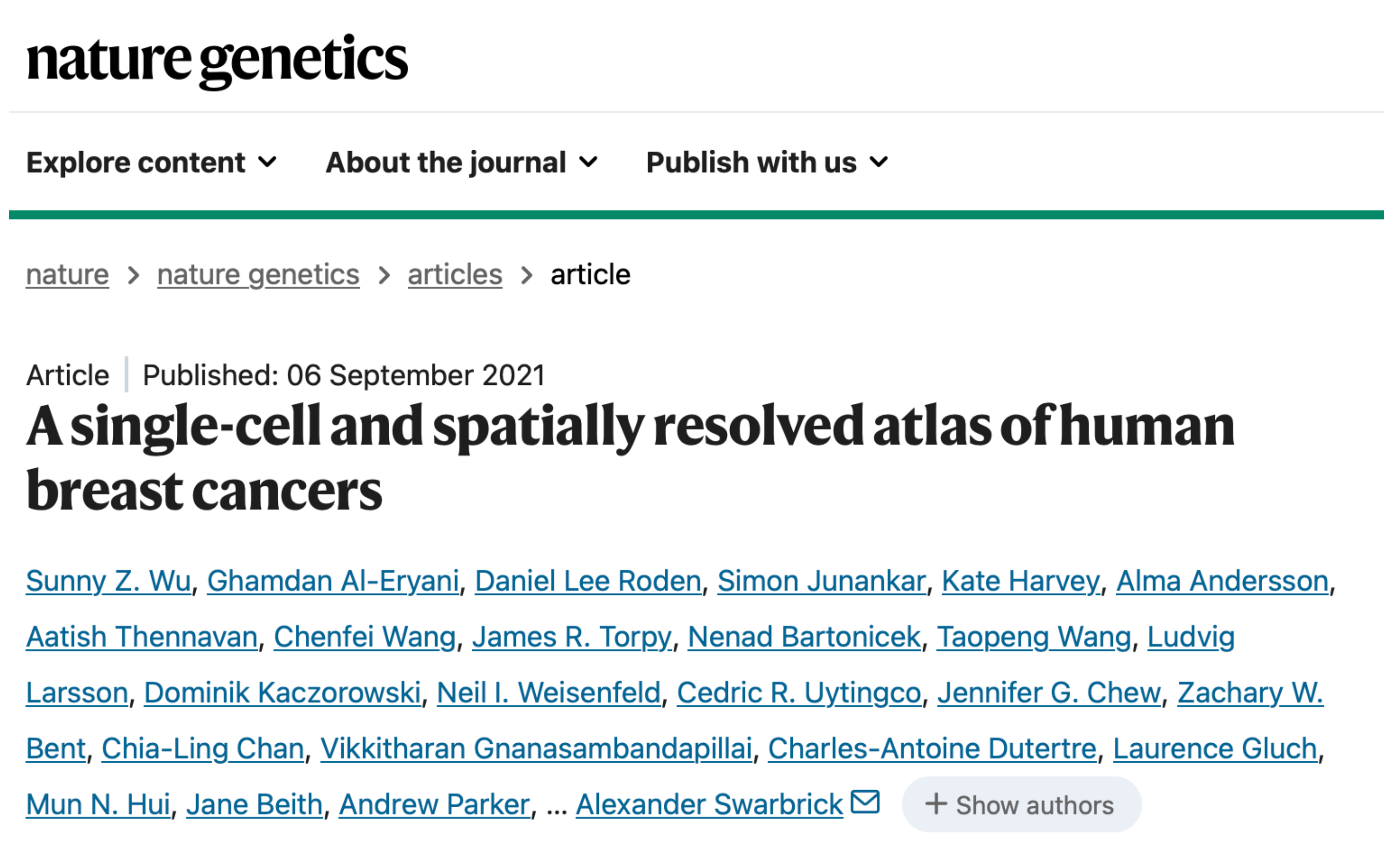
Mathematics of Cancer: Open Mathematical Problems

IPAM, FEBRUARY 23 - 27, 2026

National Institute for Theory and Mathematics in Biology

Funded by the U.S. National Science Foundation and the Simons Foundation

Northwestern University



nature genetics

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature > nature genetics > articles > article

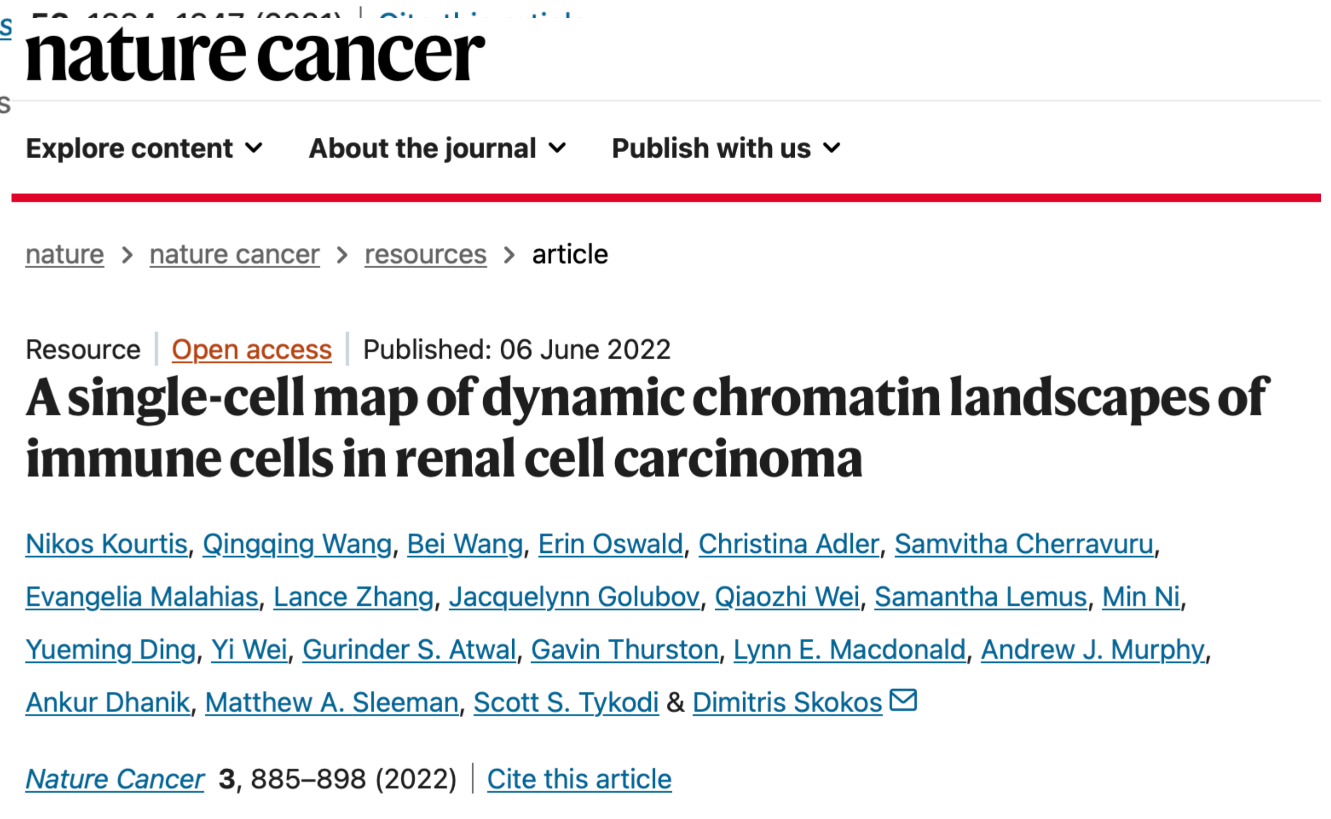
Article | Published: 06 September 2021

A single-cell and spatially resolved atlas of human breast cancers

Sunny Z. Wu, Ghamdan Al-Eryani, Daniel Lee Roden, Simon Junankar, Kate Harvey, Alma Andersson, Aatish Thennavan, Chenfei Wang, James R. Torpy, Nenad Bartonicek, Taopeng Wang, Ludvig Larsson, Dominik Kaczorowski, Neil I. Weisenfeld, Cedric R. Uytingco, Jennifer G. Chew, Zachary W. Bent, Chia-Ling Chan, Vikkitharan Gnanasambandapillai, Charles-Antoine Dutertre, Laurence Gluch, Mun N. Hui, Jane Beith, Andrew Parker, ... Alexander Swarbrick ✉ + Show authors

Nature Genetics

134k Accesses



nature cancer

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature > nature cancer > resources > article

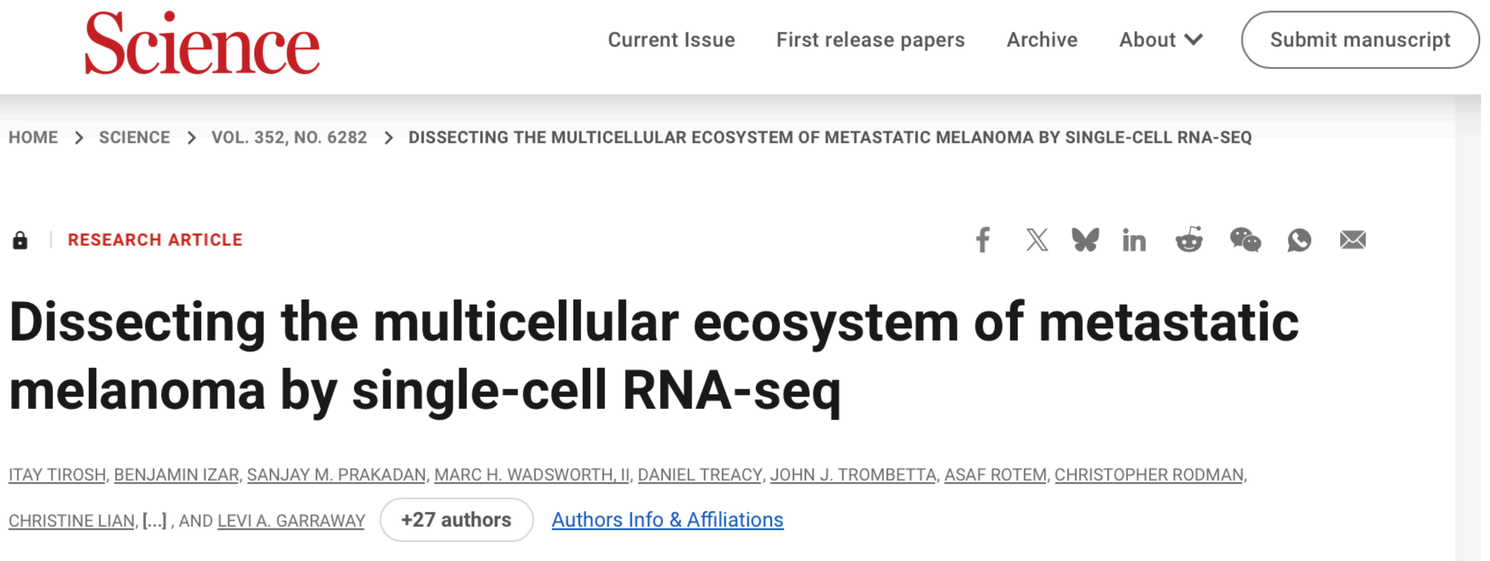
Resource | Open access | Published: 06 June 2022

A single-cell map of dynamic chromatin landscapes of immune cells in renal cell carcinoma

Nikos Kourtis, Qingqing Wang, Bei Wang, Erin Oswald, Christina Adler, Samvitha Cherravuru, Evangelia Malahias, Lance Zhang, Jacquelynn Golubov, Qiaozhi Wei, Samantha Lemus, Min Ni, Yueming Ding, Yi Wei, Gurinder S. Atwal, Gavin Thurston, Lynn E. Macdonald, Andrew J. Murphy, Ankur Dhanik, Matthew A. Sleeman, Scott S. Tykodi & Dimitris Skokos ✉

Nature Cancer 3, 885–898 (2022) | Cite this article

25k Acc



Science

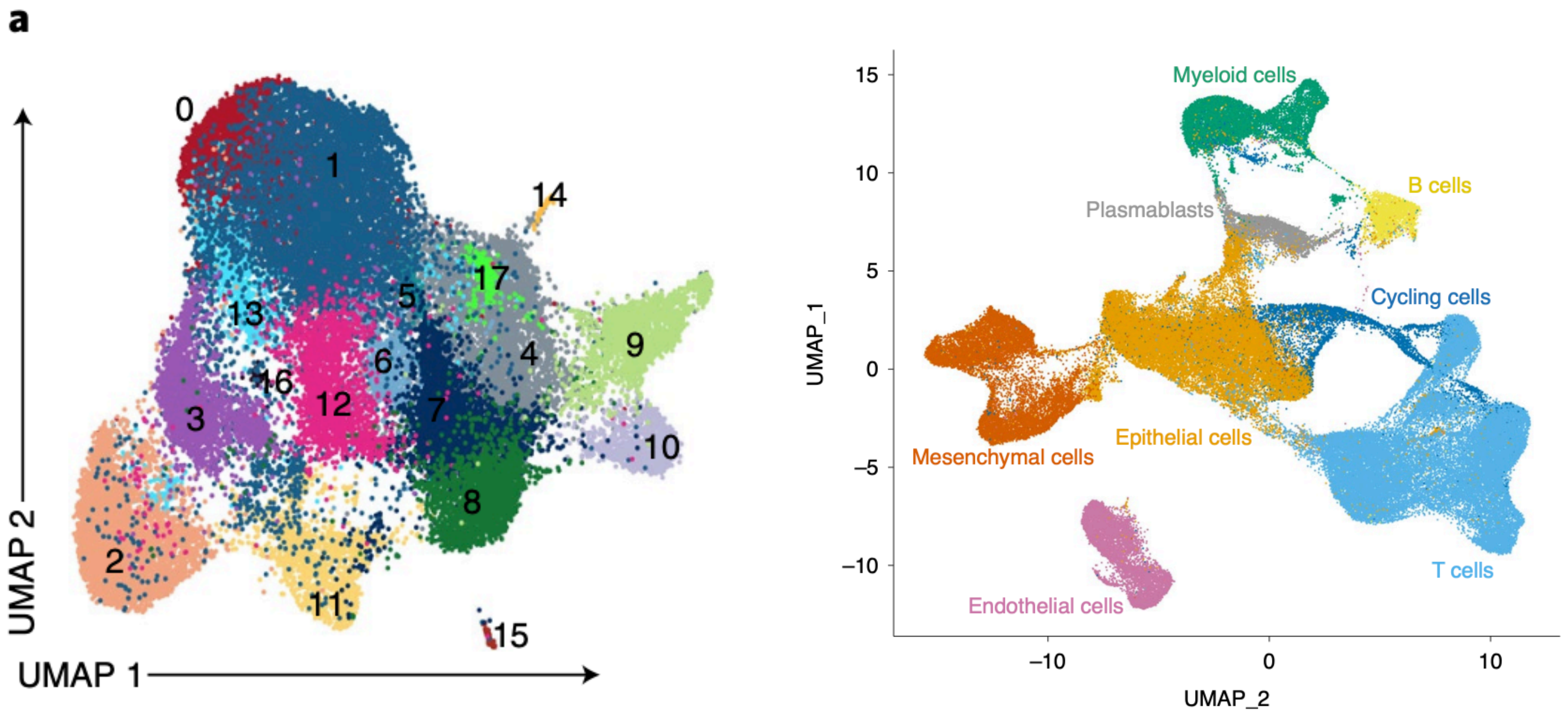
Current Issue First release papers Archive About ⌄ Submit manuscript

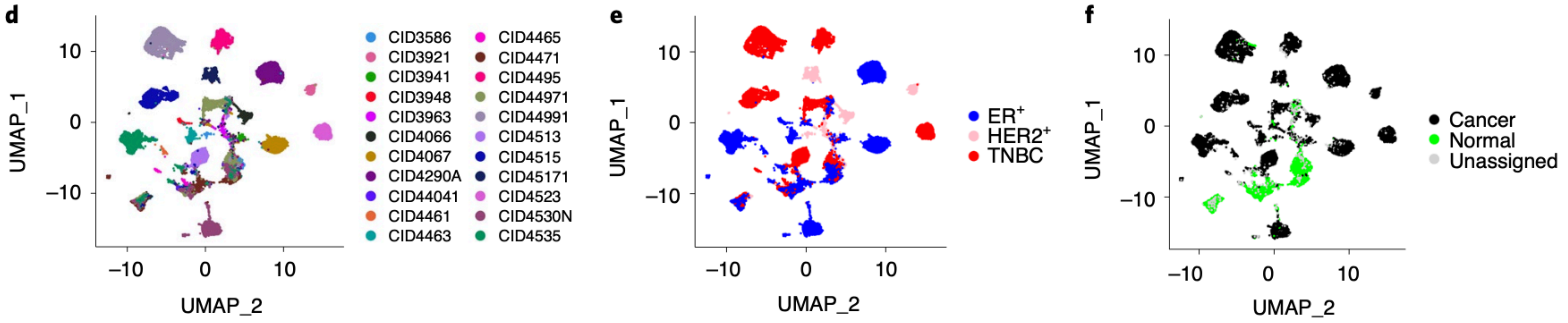
HOME > SCIENCE > VOL. 352, NO. 6282 > DISSECTING THE MULTICELLULAR ECOSYSTEM OF METASTATIC MELANOMA BY SINGLE-CELL RNA-SEQ

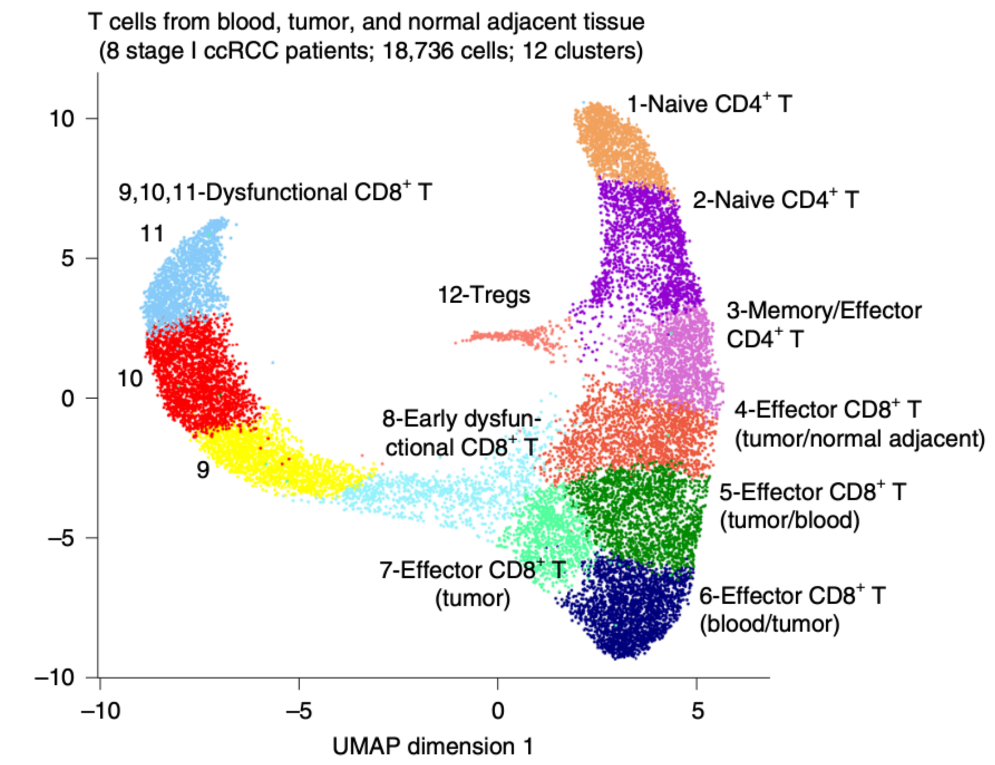
RESEARCH ARTICLE

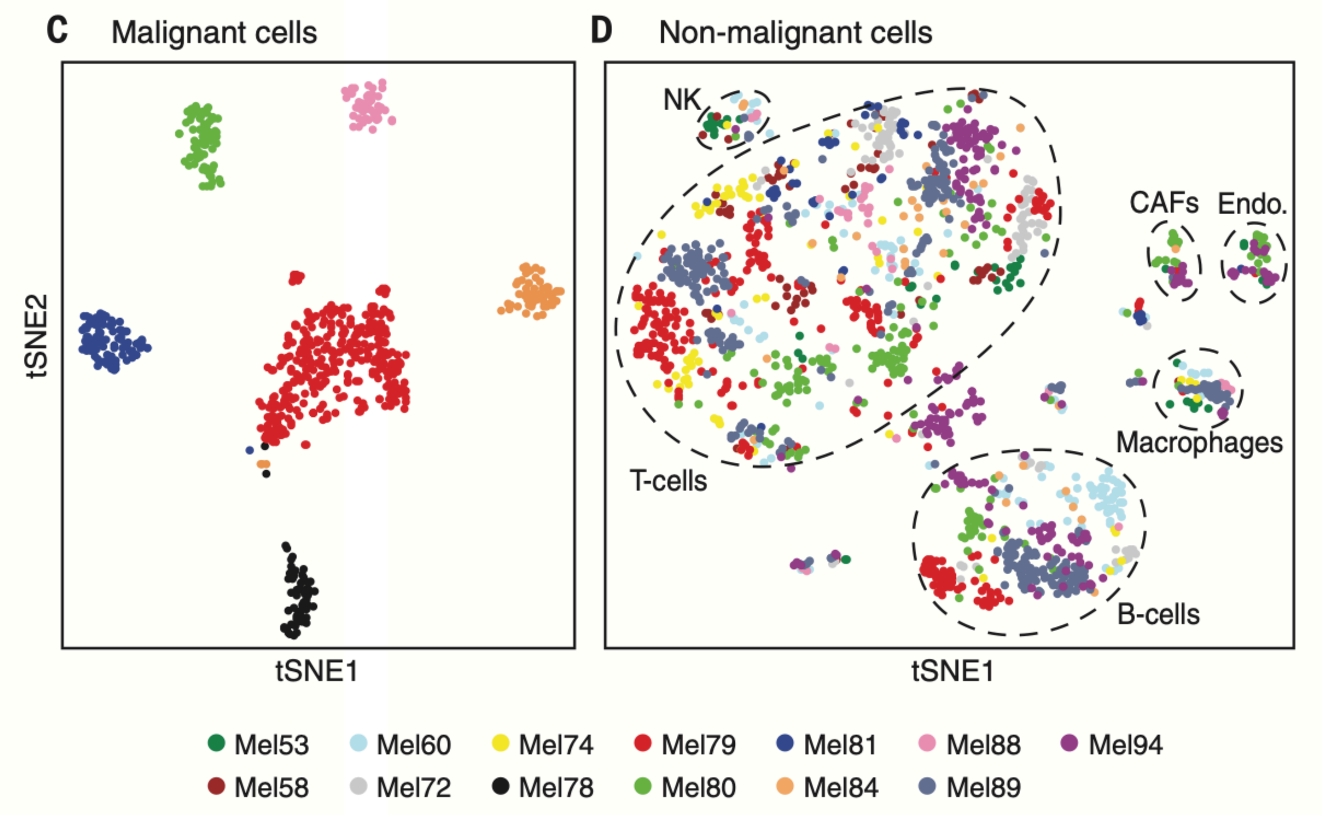
Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq

ITAY TIROSH, BENJAMIN IZAR, SANJAY M. PRAKADAN, MARC H. WADSWORTH, II, DANIEL TREACY, JOHN J. TROMBETTA, ASAF ROTEM, CHRISTOPHER RODMAN, CHRISTINE LIAN, [...], AND LEVI A. GARRAWAY +27 authors Authors Info & Affiliations









Dimension Reduction Problem

Main problem: Given a set of n high-dimensional points

$$X_1, \ldots, X_n \in \mathbb{R}^d$$

Find a low dimensional collection of points

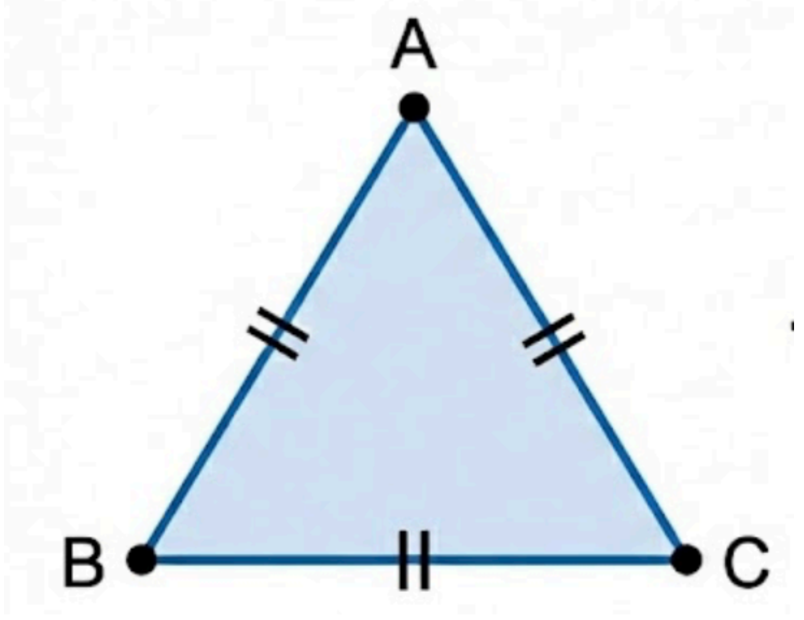
$$Y_1, \ldots, Y_n \in \mathbb{R}^s, \quad s < d$$

that captures most of the geometry of $\{X_i : i = 1, \ldots, n\}$.

Dimension Reduction Problem

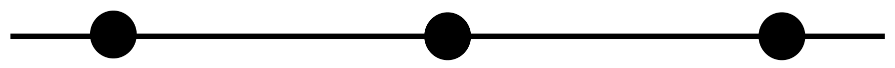
Main problem: Given a set of n high-dimensional points

$$X_1, \ldots, X_n \in \mathbb{R}^d$$



Find a low dimensional collection of points

$$Y_1, \ldots, Y_n \in \mathbb{R}^s, \quad s < d$$



that captures most of the geometry of $\{X_i : i = 1, \ldots, n\}$.

- *Ill-posed.*
- *Input* $d \gg 1$, *Output* $s = 2$.

Johnson-Lindenstrauss's Lemma

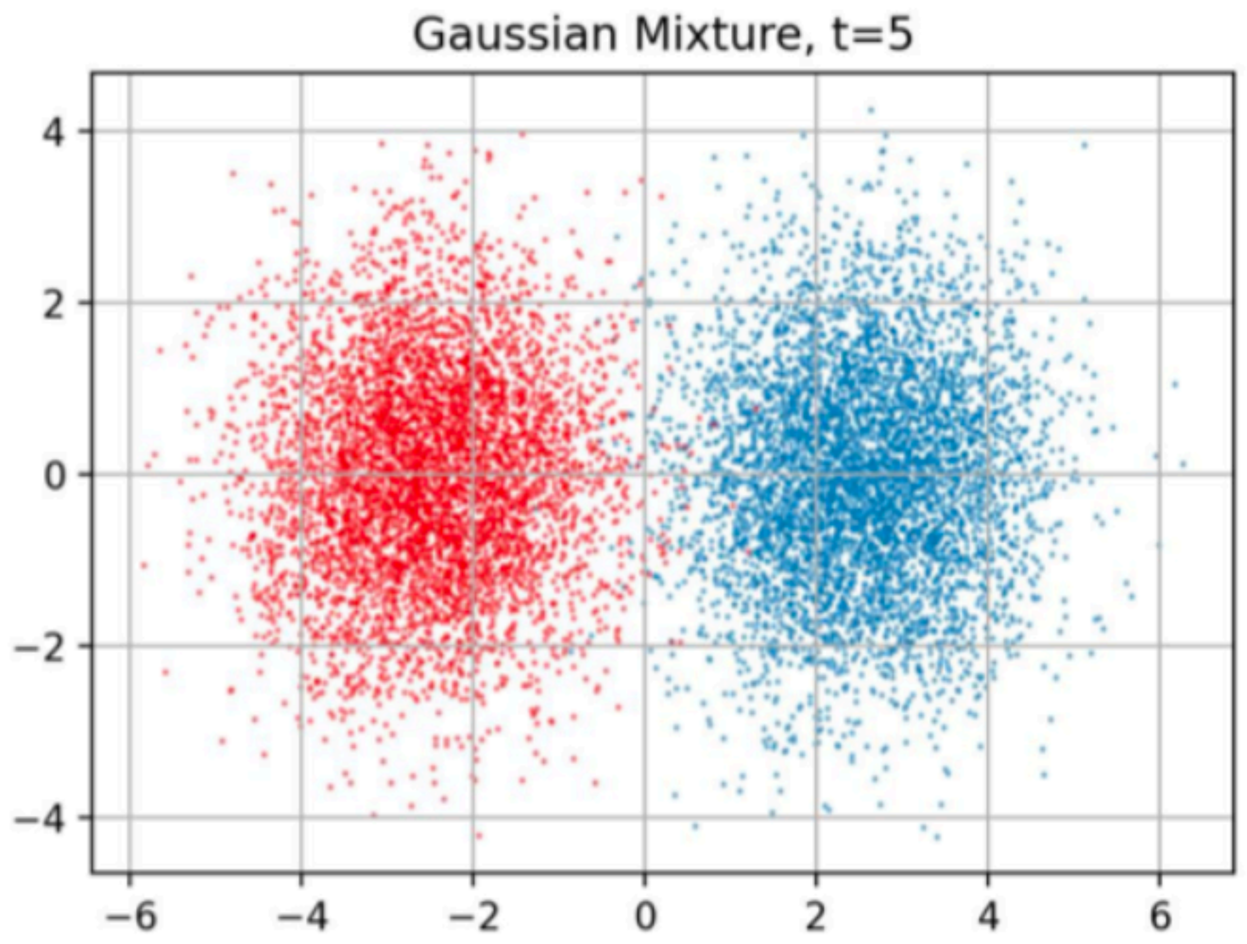
Theorem (Johnson–Lindenstrauss, 1984)

Let $X \subset \mathbb{R}^d$ be a set of n points. There exists some universal constant $C > 0$ such that, given $\varepsilon > 0$, for $s \geq \frac{C \log n}{\varepsilon^2}$ there exists a linear map $f : \mathbb{R}^d \to \mathbb{R}^s$ such that, for all $x, x' \in X$

$$(1-\varepsilon)\|x - x'\|^2 \leq \|f(x) - f(x')\|^2 \leq (1+\varepsilon)\|x - x'\|^2.$$

- $n = 10,000$, $\varepsilon = 1/10$ leads to $s \approx 250,000$.
- Probabilistic method. Solution: a $d \times s$ random matrix works with high probability.

Two illustrative examples: MNIST and Gaussian noise



MNIST

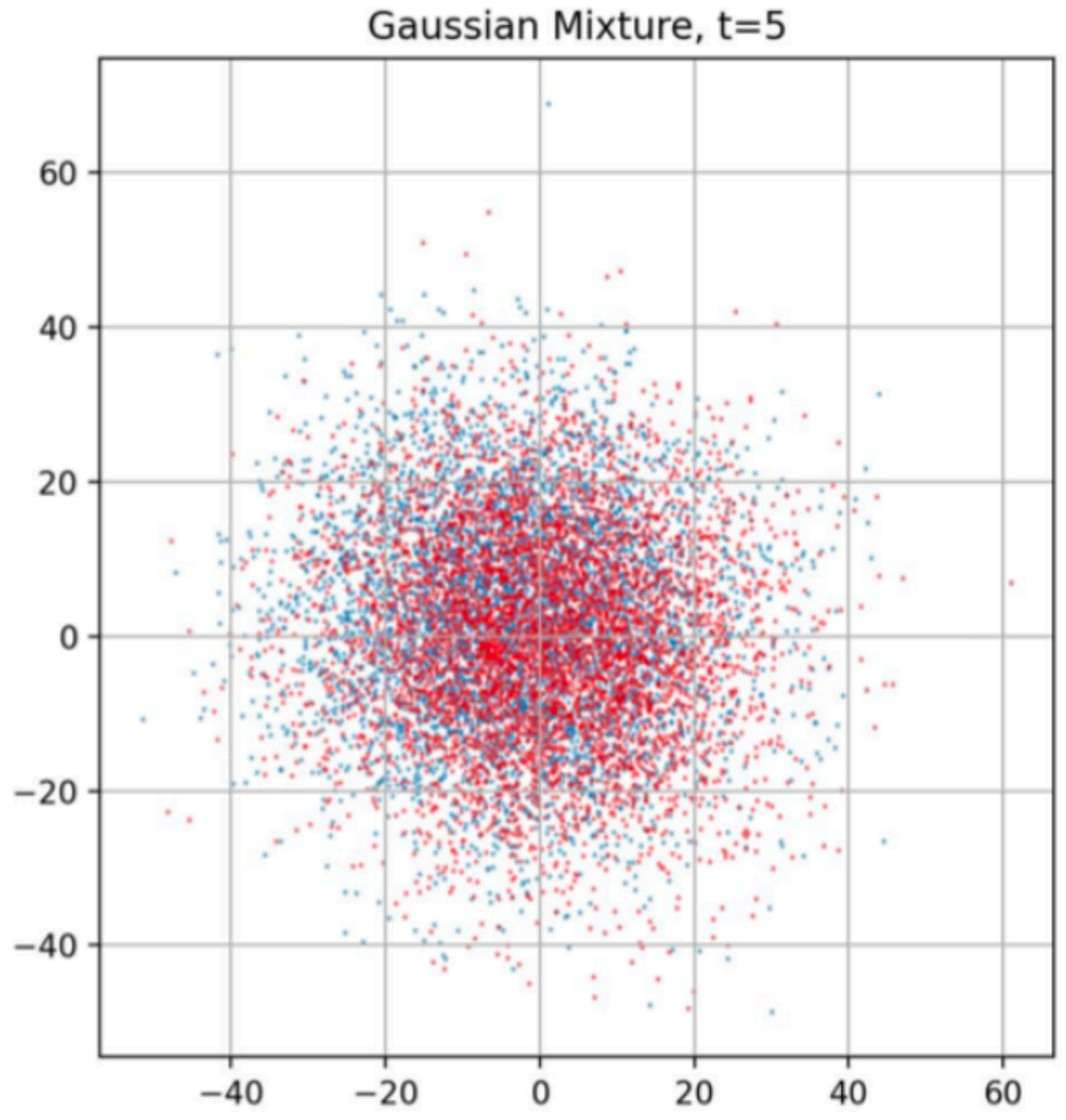
Figure: 28 × 28 pixel boxes.

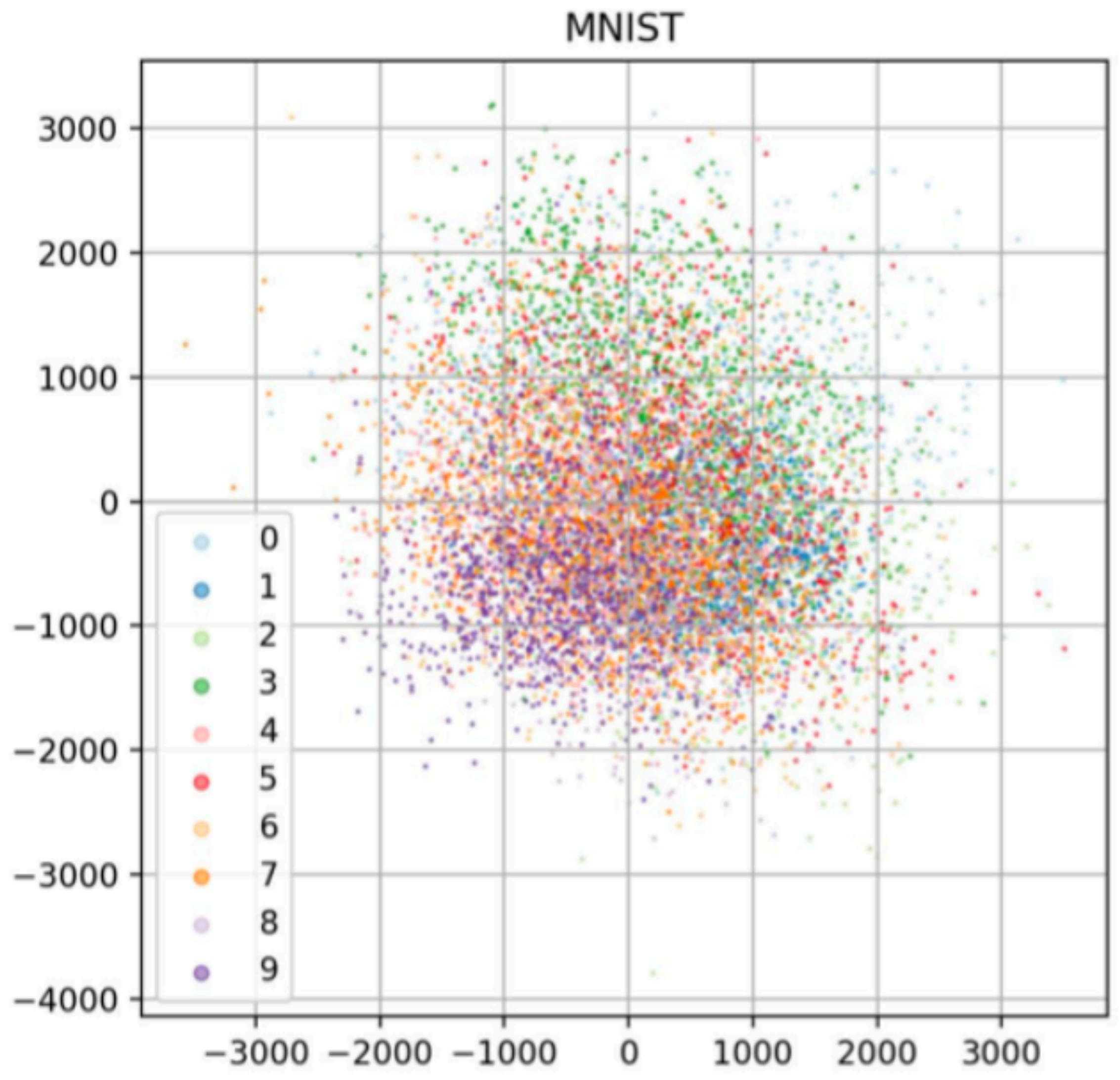
- MNIST: Each number corresponds to a vector in $\mathbb{R}^d$ with $d = 28 \times 28 = 784$.
- Gaussian Noise: $5,000$ iid standard Gaussians in dimension 784 and $5,000$ iid Gaussian centered at te_1, $t > 0$.
- Visualize different clusters.

Johnson-Lindenstrauss's Lemma

- What happens if we use Johnson-Lindenstrauss's result with output dimension 2?

Using a 2 × 758 random matrix A with *iid* Bernoulli ± 1 entries. $Y = AX$.

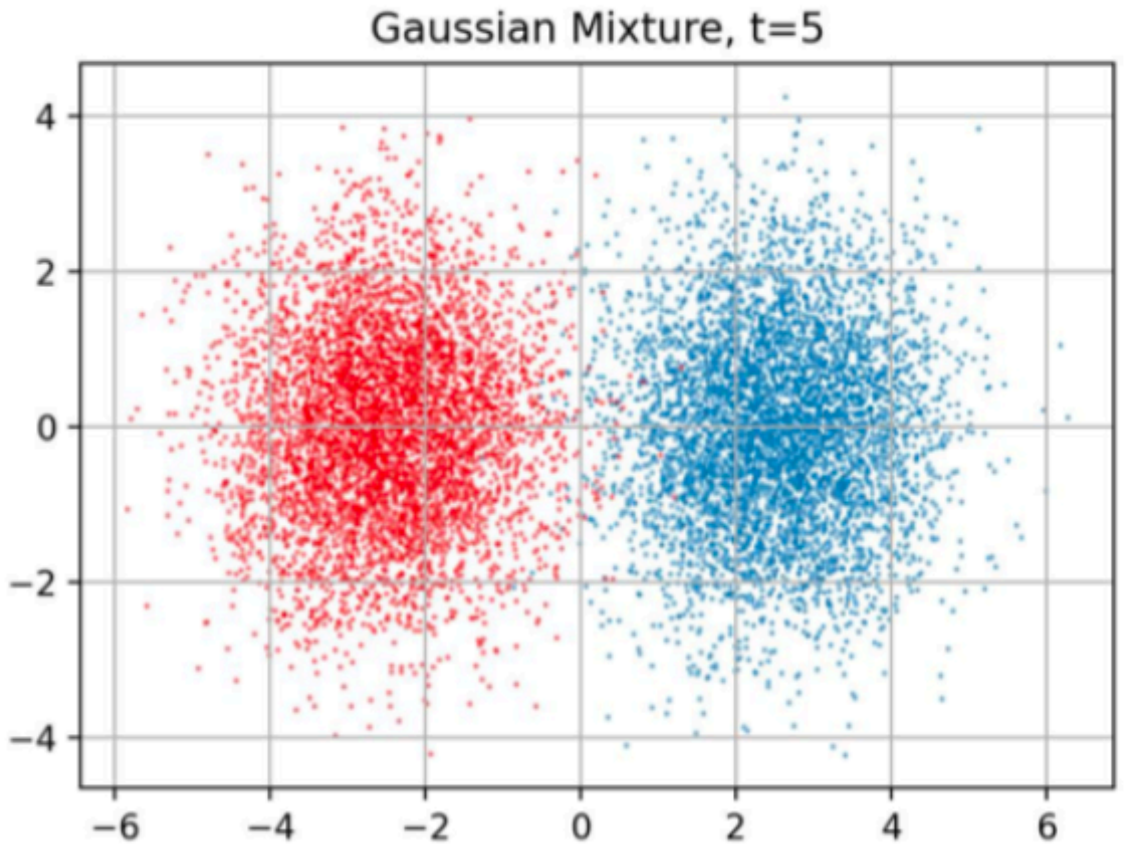


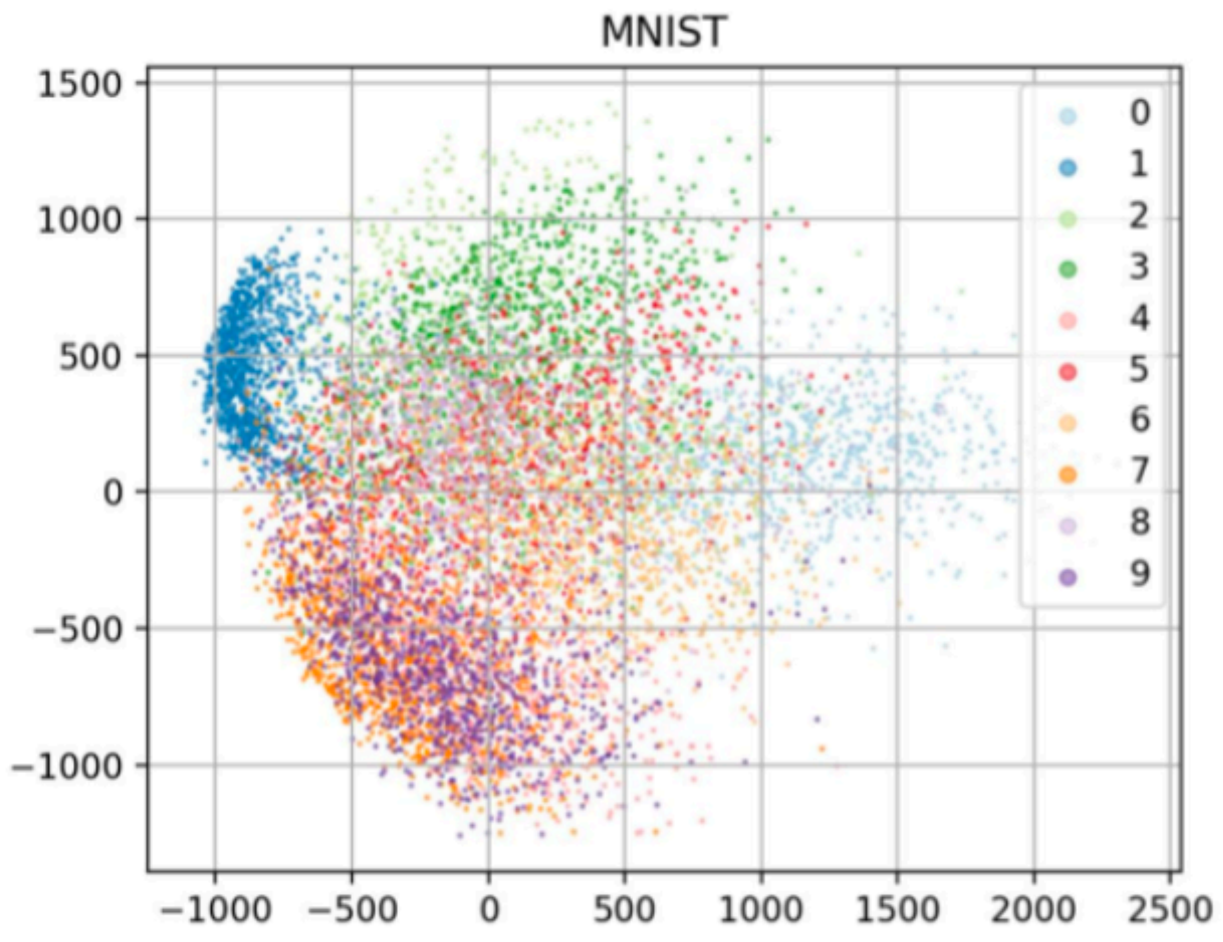


Poor performance as expected.

Principal Component Analysis (Pearson 1901)

- Write your input as a $n \times d$ matrix X.
- Construct the matrix $C = XX^t$.
- Project the input data X onto the subspace spanned by the top 2 eigenvectors of XX^T.



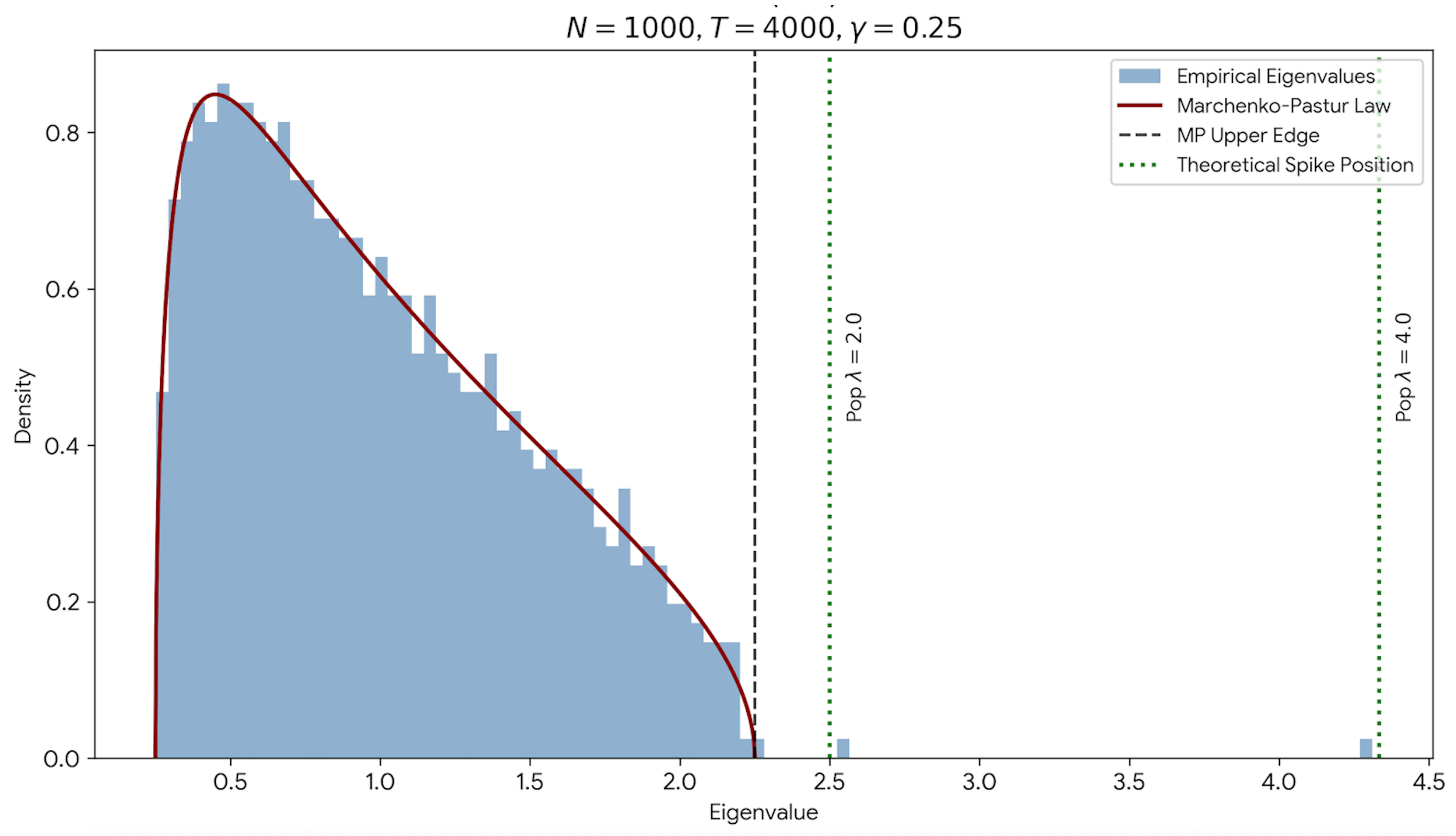


- Top eigenvectors are called the principal components.

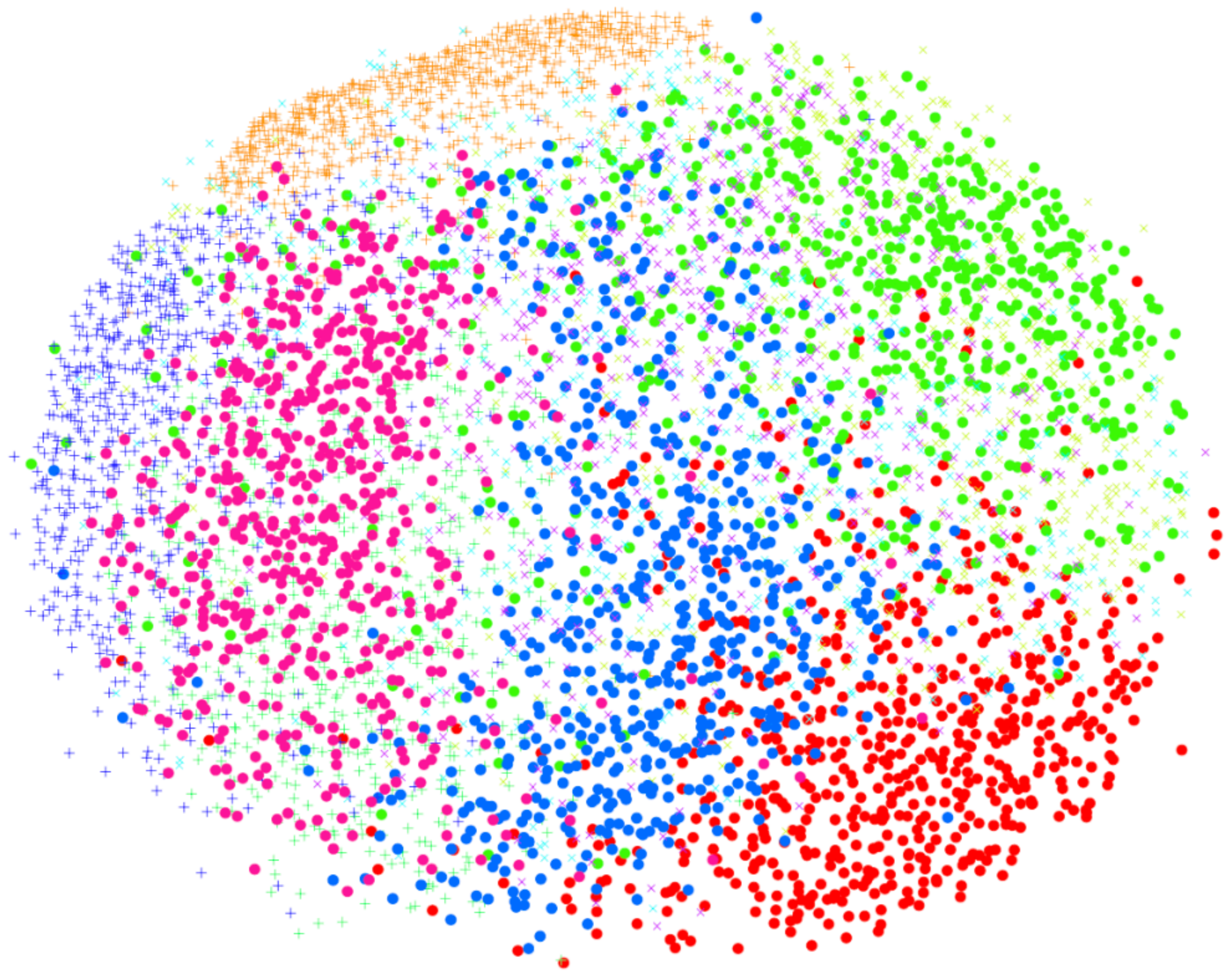
- Why and when does it work?
- Let X be a matrix of independent random variables (say Gaussian).
- (Baik-Ben Arous-Péché, 2002) The top eigenvector v_1 of the matrix $Z = XX^T + te_1e_1^T$ goes through a phase transition.

$$\langle v_1, e_1 \rangle \to \begin{cases} 0, & t < 1 \\ c(\gamma) > 0, & t > 1. \end{cases}$$

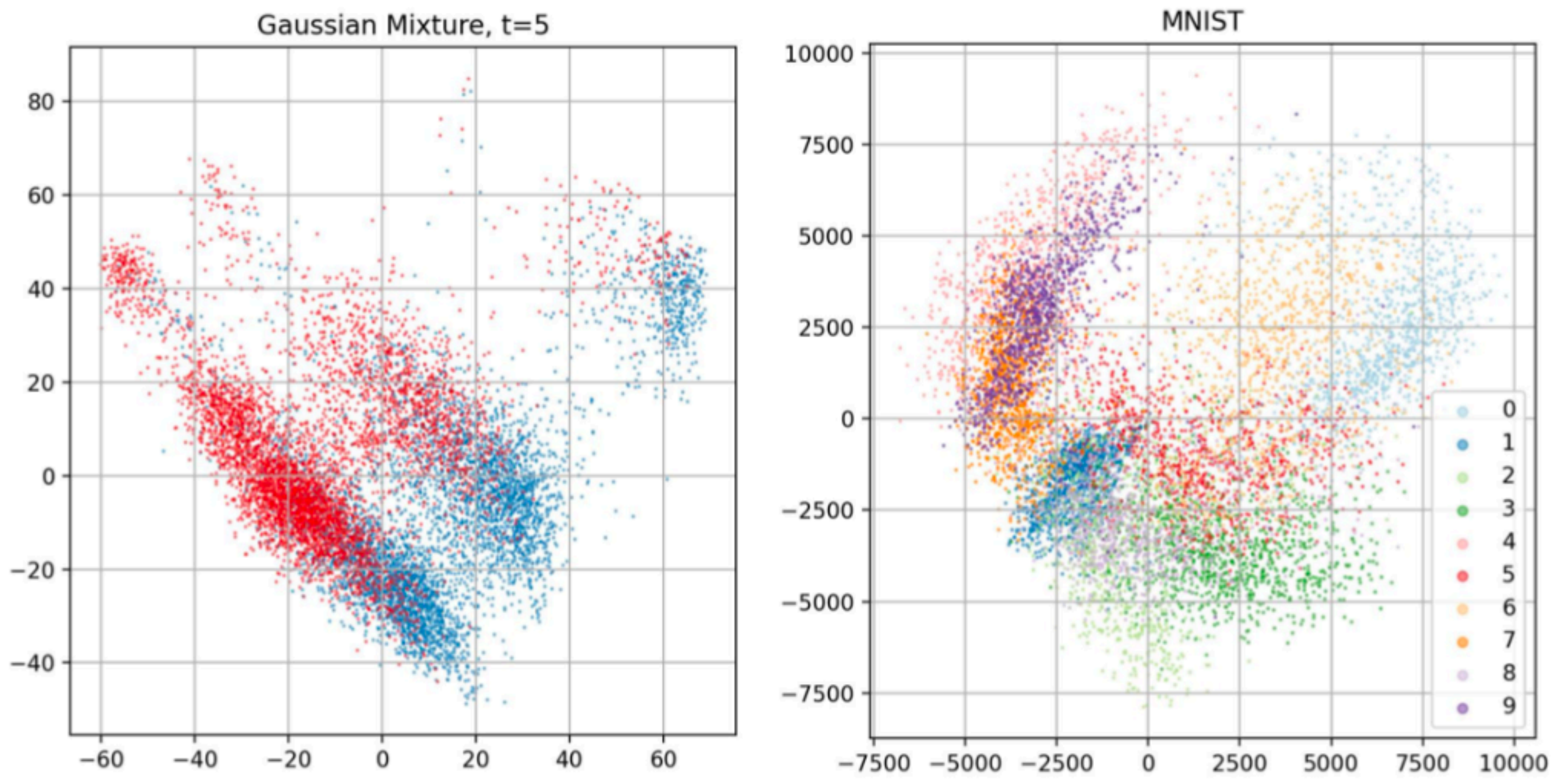
as $n \to \infty$ where $\gamma = d/n \in (0, 1)$.



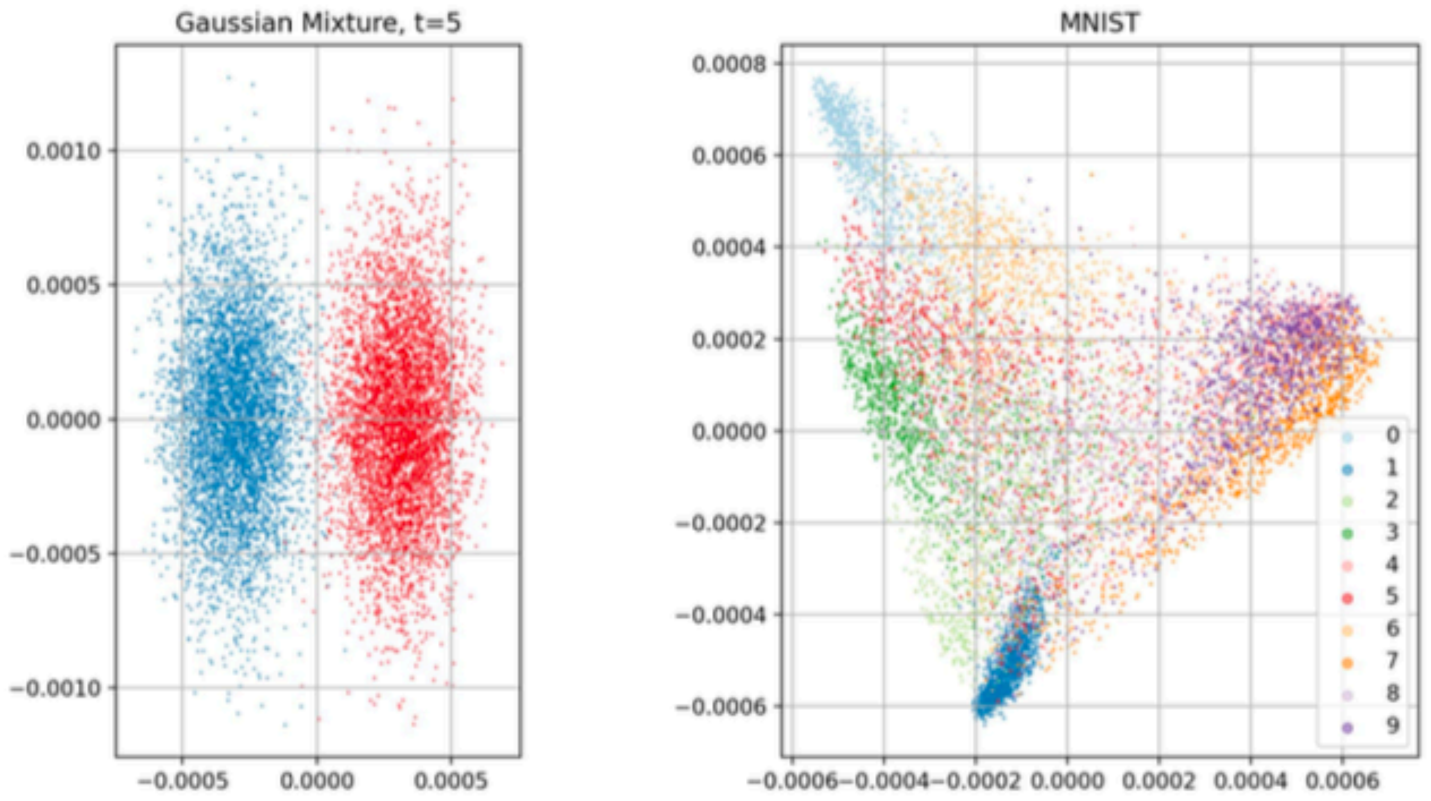
Sammon (1969)



Isomaps (Tenenbaum, de Silva, Langford 2000)

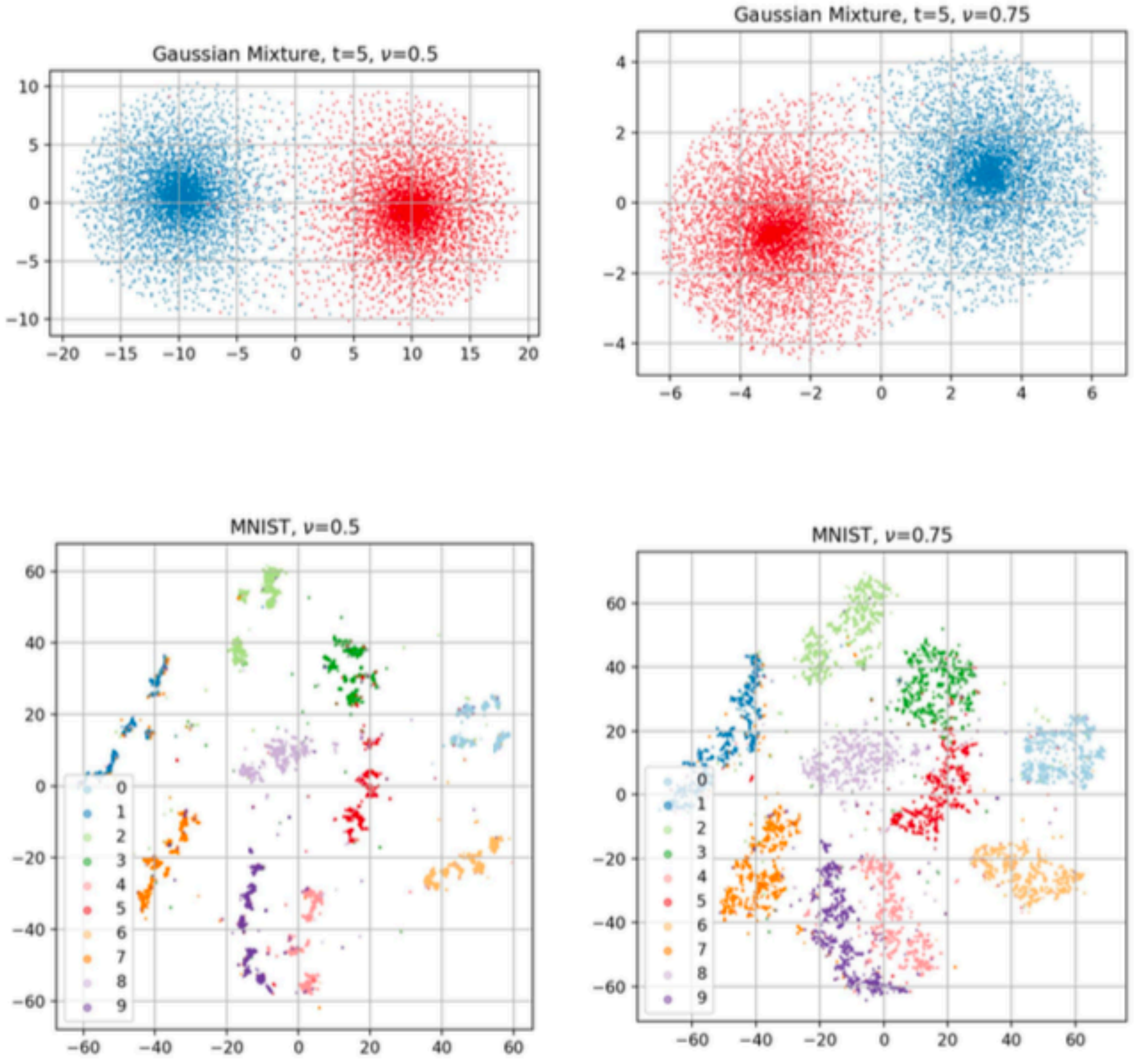


Laplacian Eigenmaps (2003)

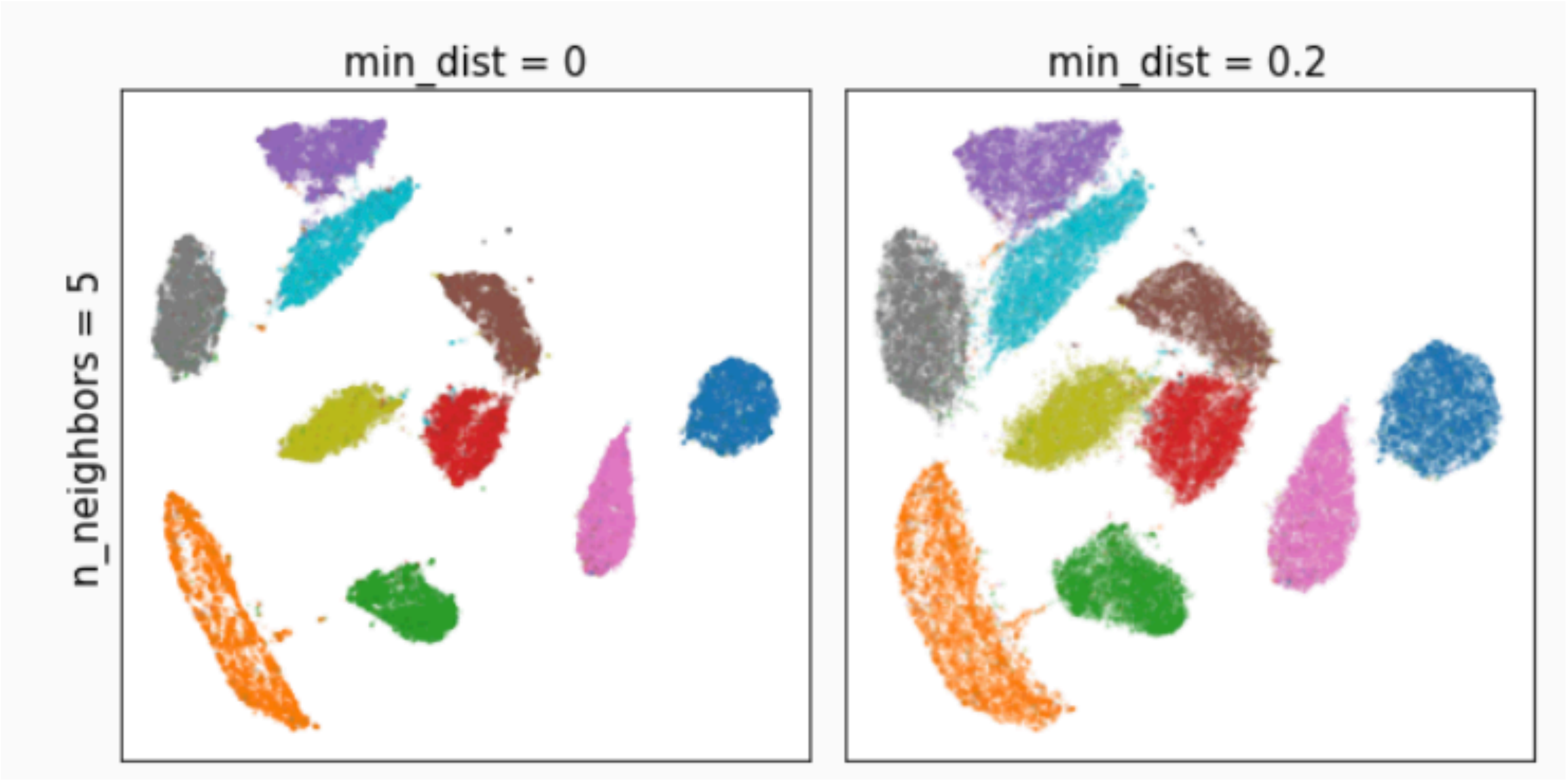


t-Stochastic Neighbor Embedding

(van Der Maaten & Hinton, 2008)



UMAP (McInnes & Healy, 2018)



What is t-SNE?

- **Main idea.** Turn a configuration of points into probability distributions and force these distributions to be similar.

- The collection $X = \{X_1, \ldots, X_n\}$ is mapped to a sequence of probability measures

$$\mu_X = (\mu_1, \ldots, \mu_n).$$

- The measure μ_i is supported on $\{x_j\}, j \neq i$ and represents the likelihood that point i will choose point j (or the similarity between i and j).

- Do the same for $(Y_1, \ldots, Y_n)$ and create $\mu_Y = (\nu_1, \ldots, \nu_n)$.

What is t-SNE?

- Given μ_X optimize the choice of $(Y_1, \ldots, Y_n)$ by minimizing the Kullbeck-Leibler divergence (or relative entropy) between μ_X and μ_Y.

$$d(\mu_X, \mu_Y) = \sum_{i=1}^{n} D_{KL}(\mu_i, \nu_i), \quad D_{KL}(\mu, \nu) = \int \log\left(\frac{d\mu}{d\nu}\right) d\nu.$$

KL Divergence (Kullback & Leibler, 1951)

- $D_{KL}(\mu_i, \nu_i) \geq 0$.
- $D_{KL}(\mu_i, \nu_j) = \sum_j p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right)$.
- $D_{KL}(\mu_i, \nu_i) = 0$ if and only if $\mu_i = \nu_i$.

How to choose the measures?

- Gaussian kernel for the input

$$\mu_i(x_j) = p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{j\neq i} \exp(-\|x_i - x_j\|^2/2\sigma_i^2)}$$

  with $p_{i|i} = 0$ and set $p_{ij} = (p_{j|i} + p_{i|j})/2$.

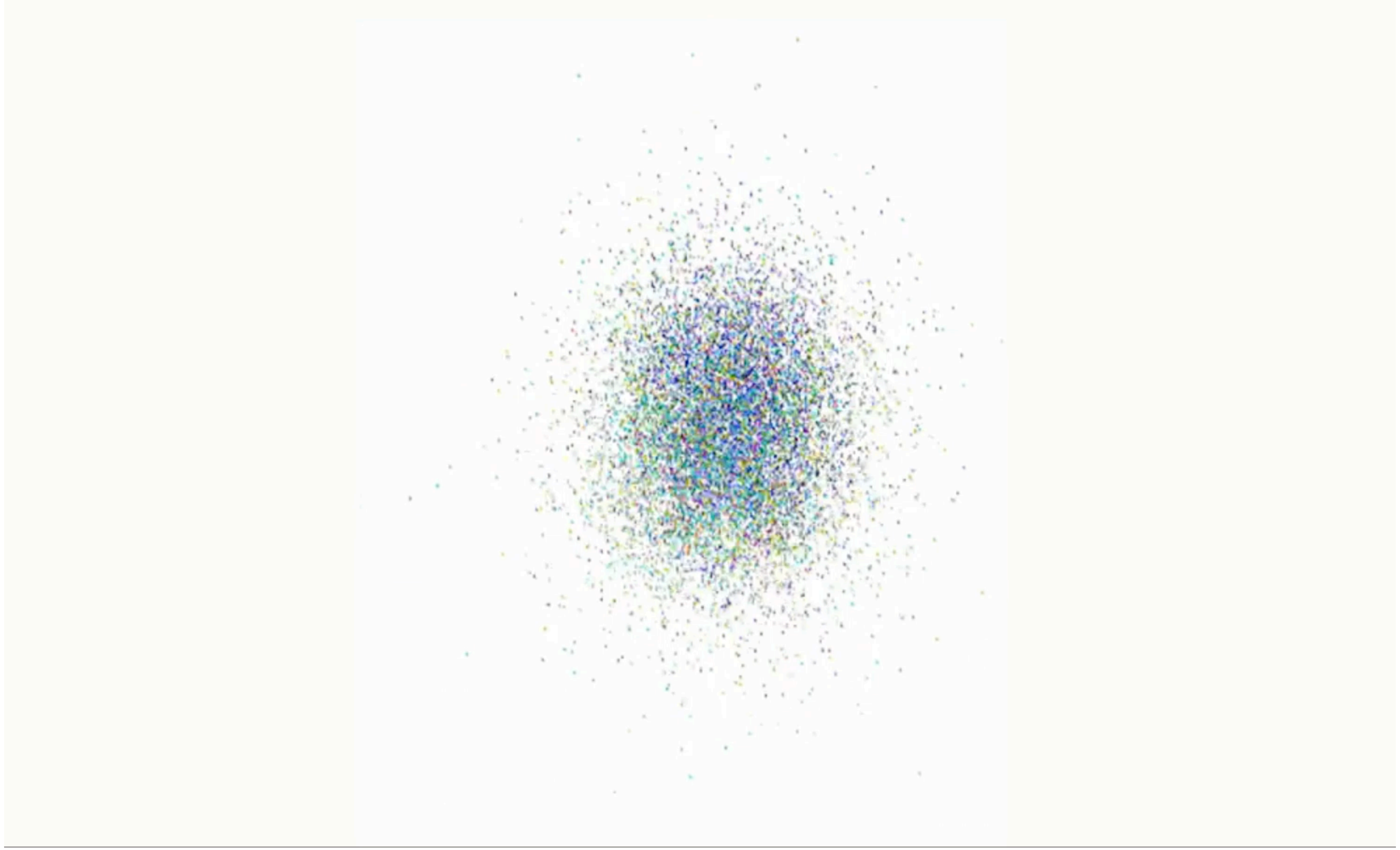
- A t-distribution for the output variables

$$\nu_i(y_j) = q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{j\neq i}(1 + \|y_i - y_j\|^2)^{-1}}$$

so we want to minimize

$$L_n(X,Y) = L_n(Y|X) = \sum_{i\neq j} p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right)$$

over the choice of q_{ij} coming from $Y \in (R^2)^n$ (non-linear constraint).



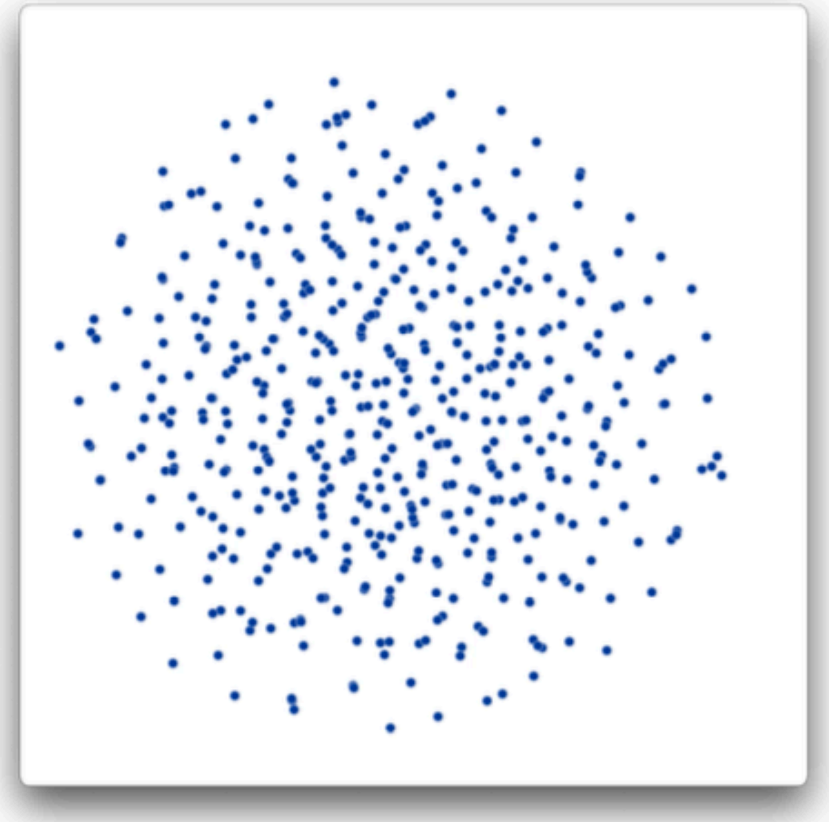
Remarks/Previous work

- SNE - Hinton & Roweis 2002
- van de Maaten & Hinton, 2008 (t-SNE)
- 56k+ citations. Wide range of use.

In mathematics:

- Linderman & Steinerberger, 2017
- Arora, Hu & Kothari, 2018
- Cai & Ma, 2022
- A. & Fletcher, 2023, A. & Gu 2025
- Jeong & Wu, Weinkove, 2024
- Bergam, Snoeck, & Verma, 2025

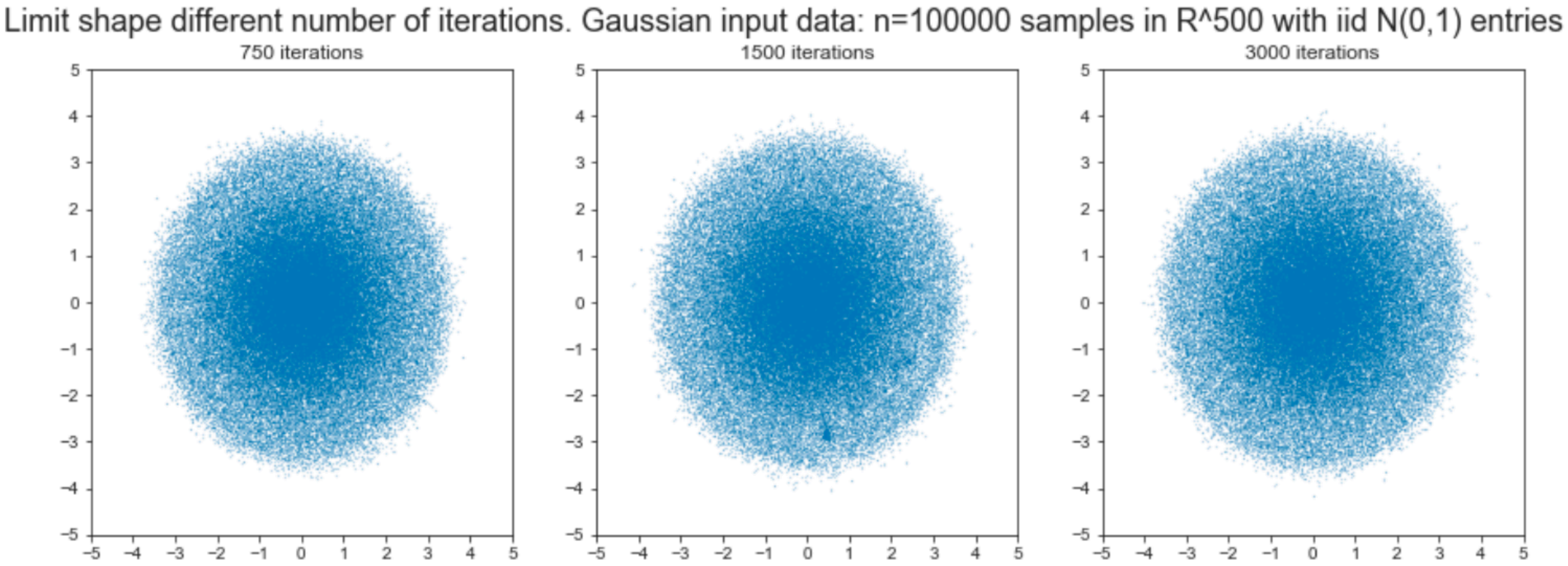
Question 1: Pure noise $\xrightarrow{\text{t-SNE}}$???



Question 2: Signal + noise – how strong should be a signal so we can detect? Recovery?

Pure noise simulations

- Set (X_i) to be independent identically distributed standard Gaussians.



Limit shape different number of iterations. Gaussian input data: n=100000 samples in R^500 with iid N(0,1) entries

750 iterations

1500 iterations

3000 iterations

Back to t-SNE

Input: $X \in (\mathbb{R}^d)^N$, Output: $Y \in (\mathbb{R}^2)^N$

$$\text{Goal:} \quad \min_Y \sum_{i \neq j} p_{ij}(X) \log \left(\frac{p_{ij}(X)}{q_{ij}(Y)} \right)$$

Perplexity

-
$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{j \neq i} \exp(-\|x_i - x_j\|^2 / 2{\sigma_i}^2)}$$

- The choice of σ_i balances local/large scale density of the original dataset.
- Choose σ_i so that

$$-\sum_{j \neq i} p_{j|i} \log p_{j|i} = \mathrm{Ent}(\mu_i) = \log \mathrm{Perp}.$$

- Perp is a parameter of the model called Perplexity.

Theorem (A.-Fletcher ’23, A.-Gu ’25)

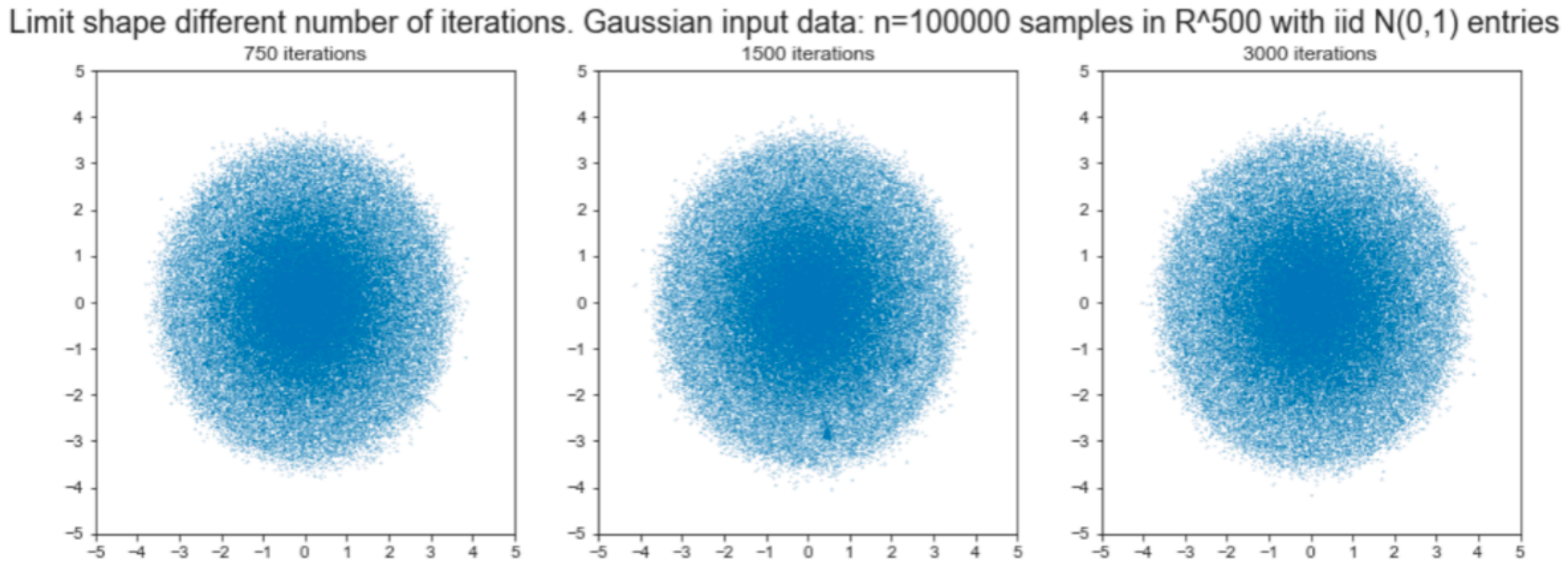
Let $X_i \in \mathbb{R}^d$ be independent, identically distributed r.v. with distribution μ_X with finite exponential moments. For $\rho \in (0,1)$, let $Y_i \in \mathbb{R}^s$ be the output of t-SNE with Perp $= n\rho$. Then

$$\lim_{n\to\infty} \inf_Y L_n(X,Y) = \inf_{\mu\in\mathcal{P}_X} I_\rho(\mu).$$

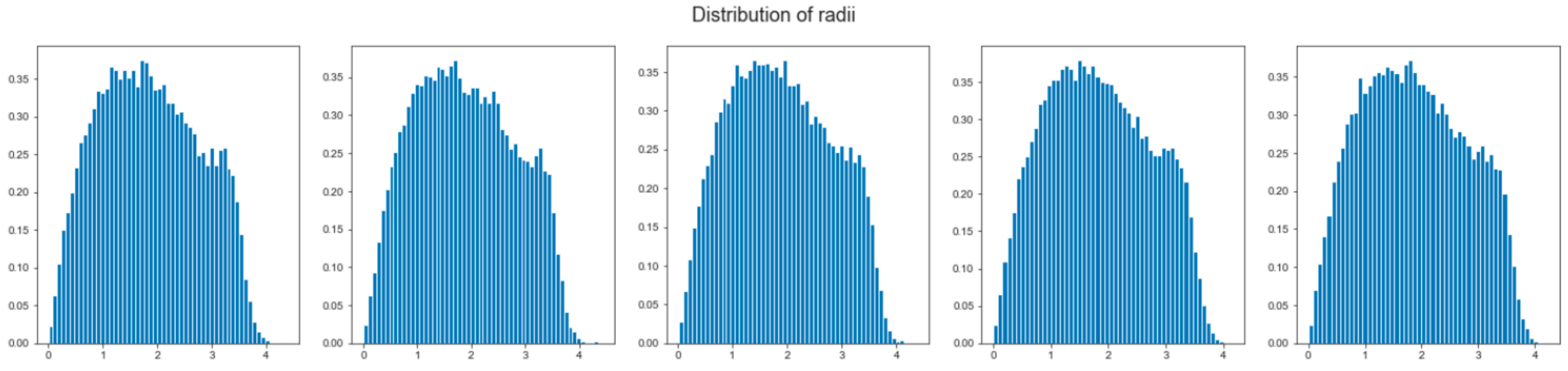
Moreover, there exists a sub-sequence n_k such that m_{n_k} converges weakly to μ^* and

$$I_\rho(\mu^*) = \inf_{\mu\in\mathcal{P}_X} I_\rho(\mu).$$

If μ_X has compact support, then μ^* has compact support.



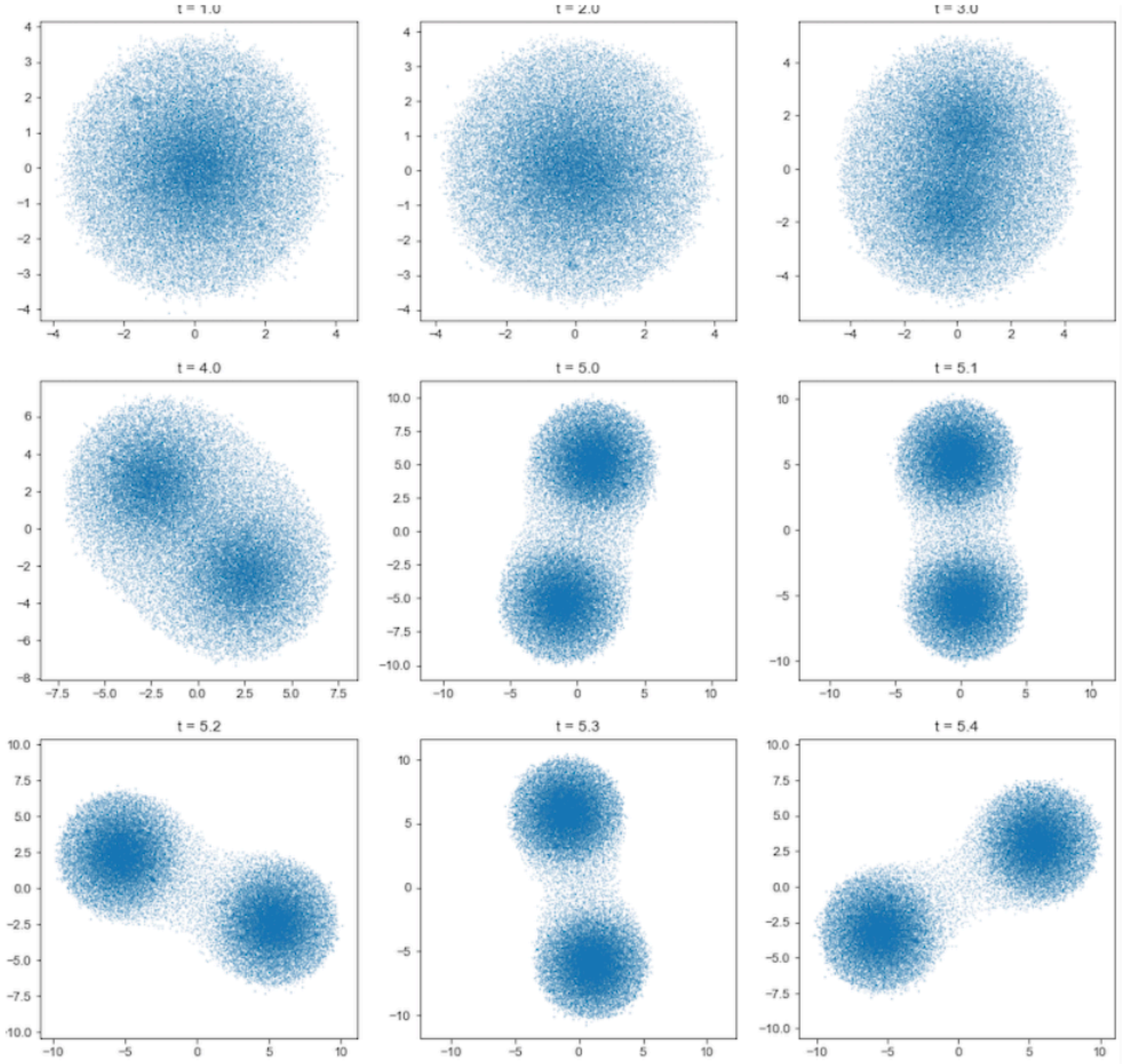
Limit shape different number of iterations. Gaussian input data: n=100000 samples in R^500 with iid N(0,1) entries

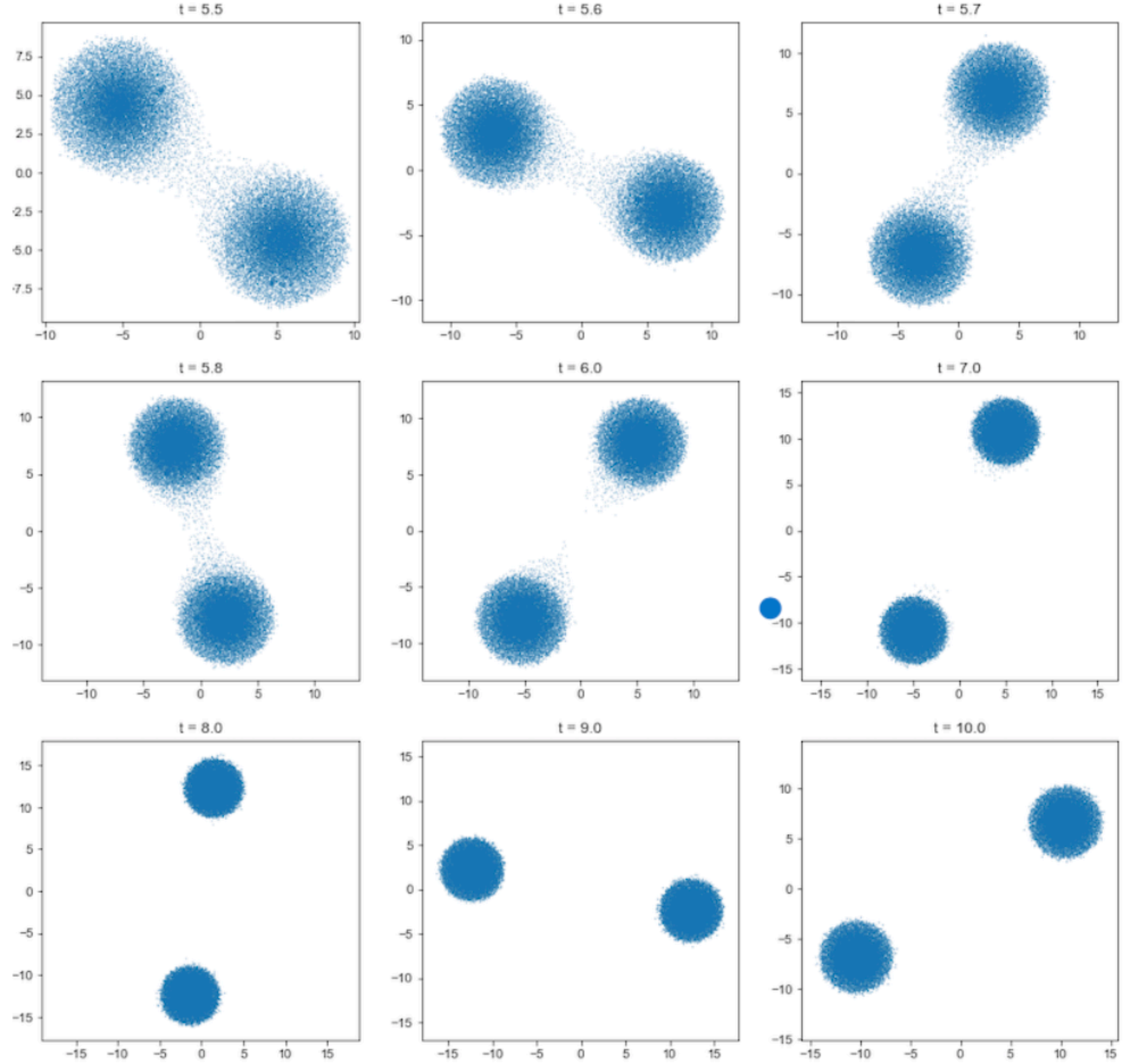


Distribution of radii

Adding a signal

Take $\mu_{X_i} \sim \frac{1}{2}\delta_Z + \frac{1}{2}\delta_{W+te_1}$, Z, W i.i.d. Gaussian.





There exists $t_c = t_c(d, \rho)$ such that $\mu^*(t)$ is connected if and only if $t \leq t_c$.

A continuous problem - limiting perplexity

Let $\mu \in \mathcal{M}(\mathbb{R}^d)$. For $\sigma \in \mathbb{R}$, $\rho \in (0,1)$ as

$$
\begin{aligned}
&F_{\rho,\mu}(x,\sigma) \\
&= \int \frac{p_\sigma(x,x')}{\int p_\sigma(x,x')\,d\mu(x')} \log\left(\frac{p_\sigma(x,x')}{\int p_\sigma(x,x')\,d\mu(x')}\right) d\mu(x') + \log \rho.
\end{aligned}
$$

with $p_\sigma(x,x') = \exp(-\|x-x'\|^2/2\sigma)$.

Let $\sigma^*(x)$ be defined as the unique solution of

$$F_{\rho,\mu}(x, \sigma^*_{\rho,\mu}(x)) = 0.$$

Now, given a function $\psi : \mathbb{R}^d \to \mathbb{R}$, define the function $p_\psi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ as

$$P_\mu(x, x') = \frac{1}{2}\left(\frac{p_{\sigma^*(x)}(x, x')}{\int p_{\sigma^*(x)}(x, x') d\mu(x')} + \frac{p_{\sigma^*(x')}(x, x')}{\int p_{\sigma^*(x')}(x, x')\, d\mu(x)}\right)$$

and also let $q : \mathbb{R}^s \times \mathbb{R}^s \to \mathbb{R}$ be given by

$$q(y, y') = \frac{(1 + \|y - y'\|^2)^{-1}}{\iint (1 + \|y - y'\|^2)^{-1}\, d\mu(y) d\mu(y')}.$$

Set

$$I_\rho(\mu) = \iint P_\mu(x, x') \log\left(\frac{P_\mu(x, x')}{q(y, y')}\right)\, d\mu(x, y) d\mu(x', y')$$

Some Open Challenges

1) Highly non-convex: when/where to stop?

2) Quantitative, error estimates

3) Different Signals; how to choose Perplexity

4) Noise dependence? Universality class?

5) UMAP: no mathematical results