SUGAR
Geometry Based Data Generation

O. Lindenbaum, J.S. Stanley, G. Wolf, S. Krishnaswamy

Yale University

2018

Acknowledgements

This work was done in collaboration with:

Jay Stanley

Guy Wolf

Smita Krishnaswamy

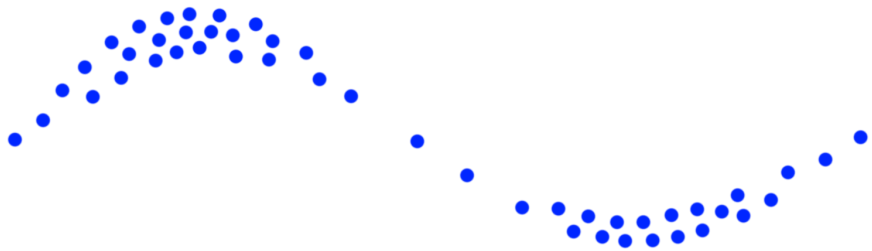
Yale

Research partially funded by grant from the CZI

Introduction & motivation

Traditional models: density based data generation

Generative models typically infer distribution from collected data, and sample it to generate more data.

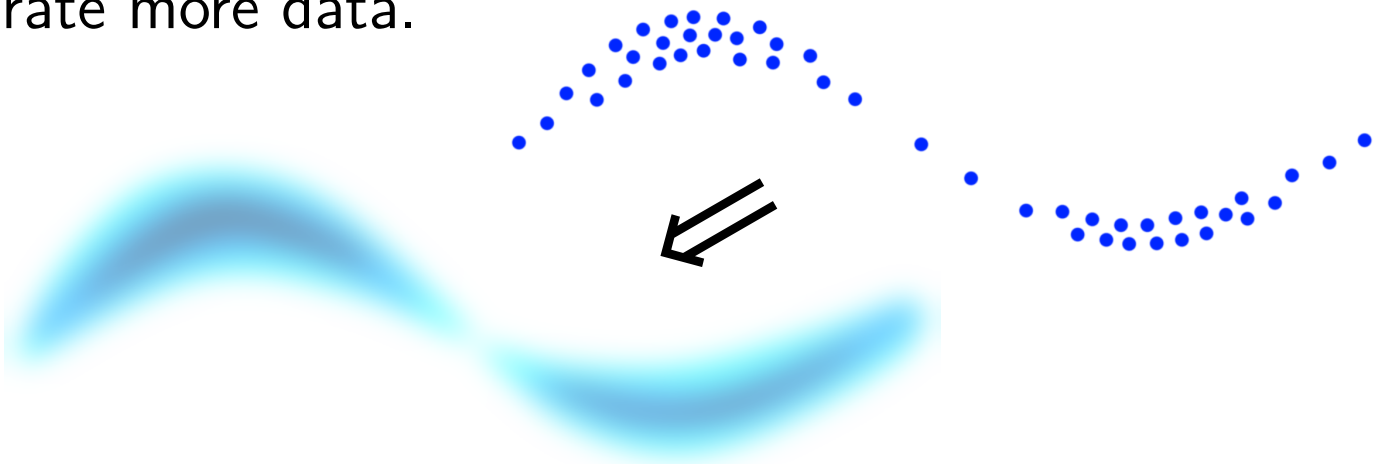


- Biased by sampling density
- May miss rare populations
- Does not preserve the geometry

Introduction & motivation

Traditional models: density based data generation

Generative models typically infer distribution from collected data, and sample it to generate more data.

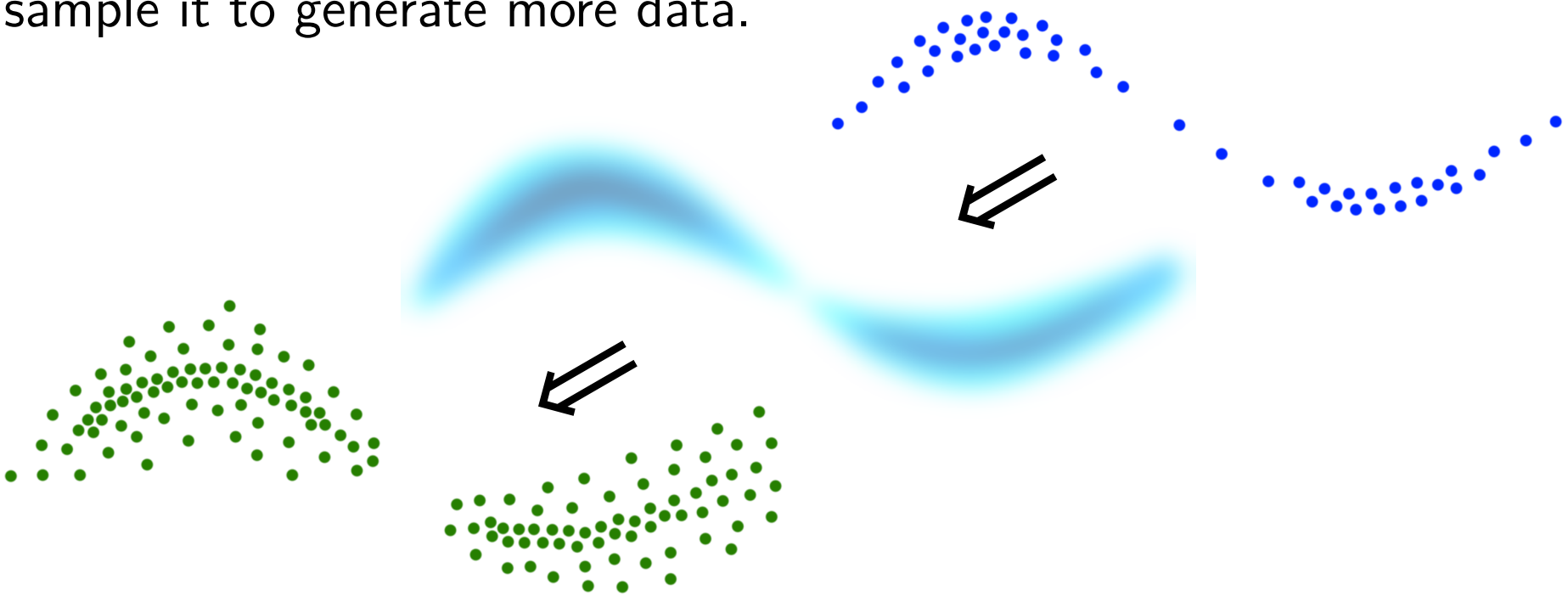


- Biased by sampling density
- May miss rare populations
- Does not preserve the geometry

Introduction & motivation

Traditional models: density based data generation

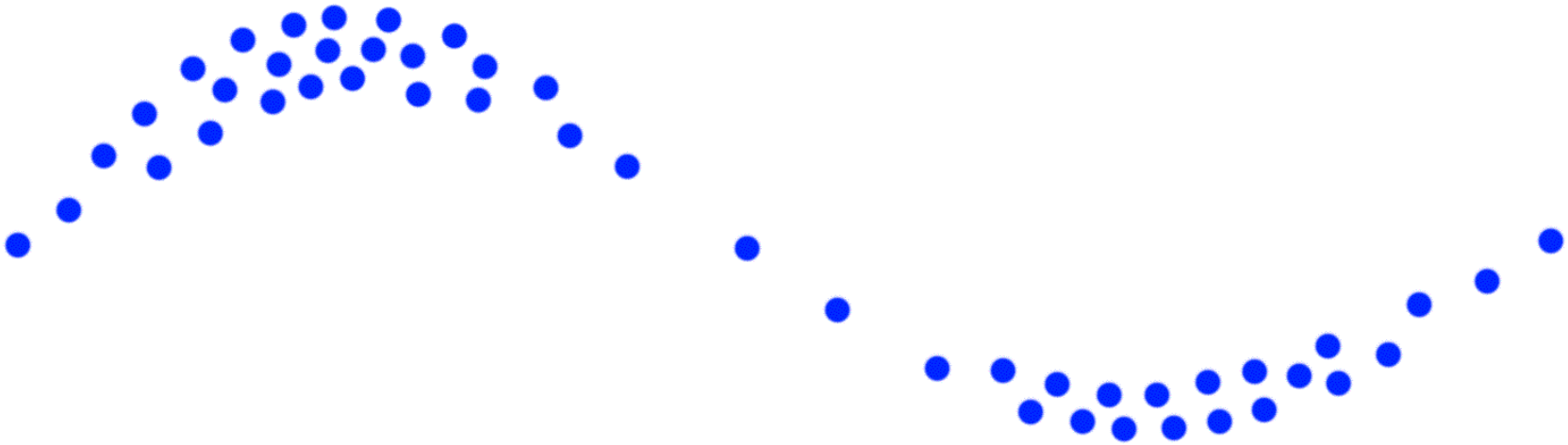
Generative models typically infer distribution from collected data, and sample it to generate more data.



- Biased by sampling density
- May miss rare populations
- Does not preserve the geometry

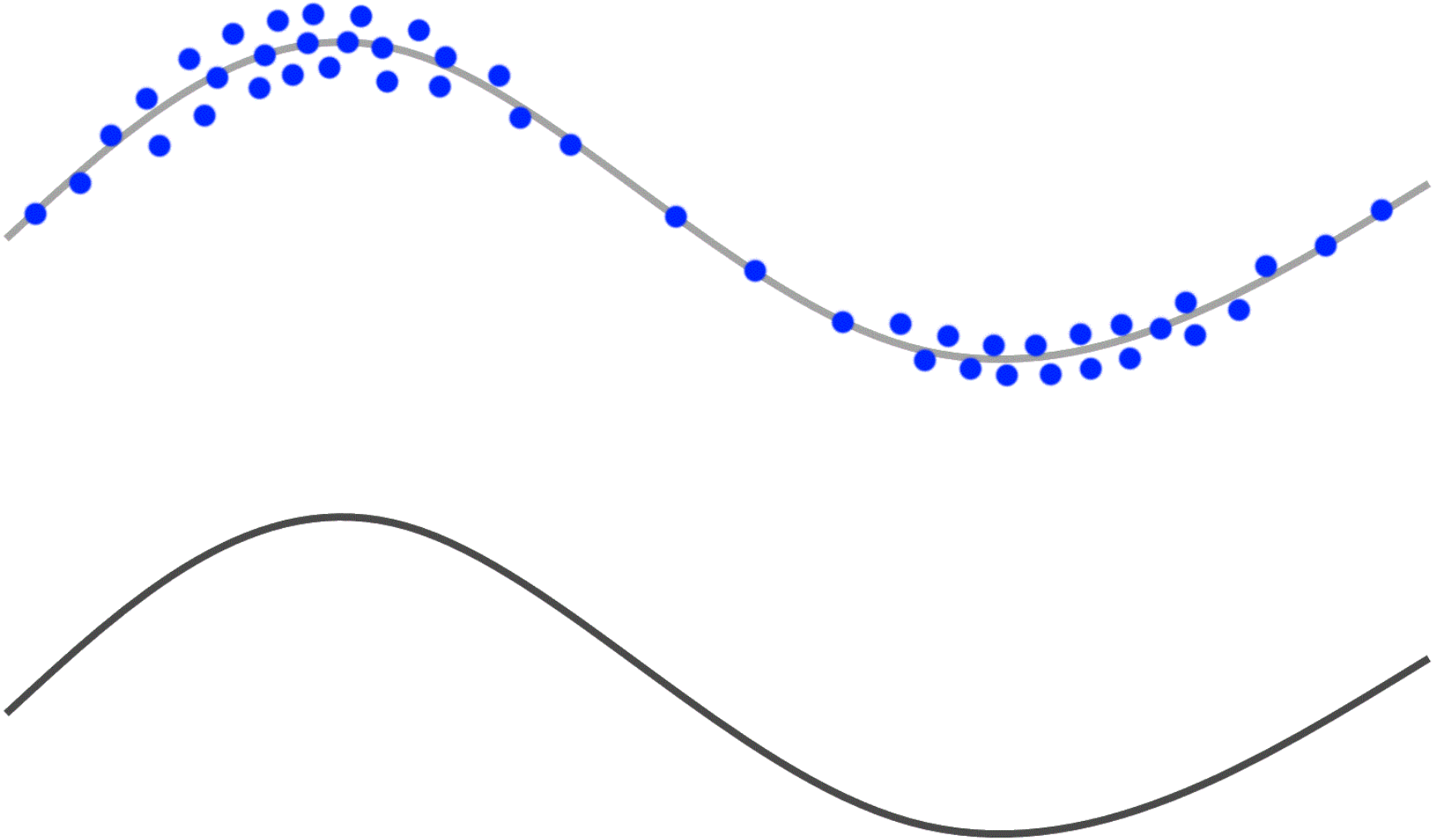
Introduction & motivation

New approach: geometry based data generation



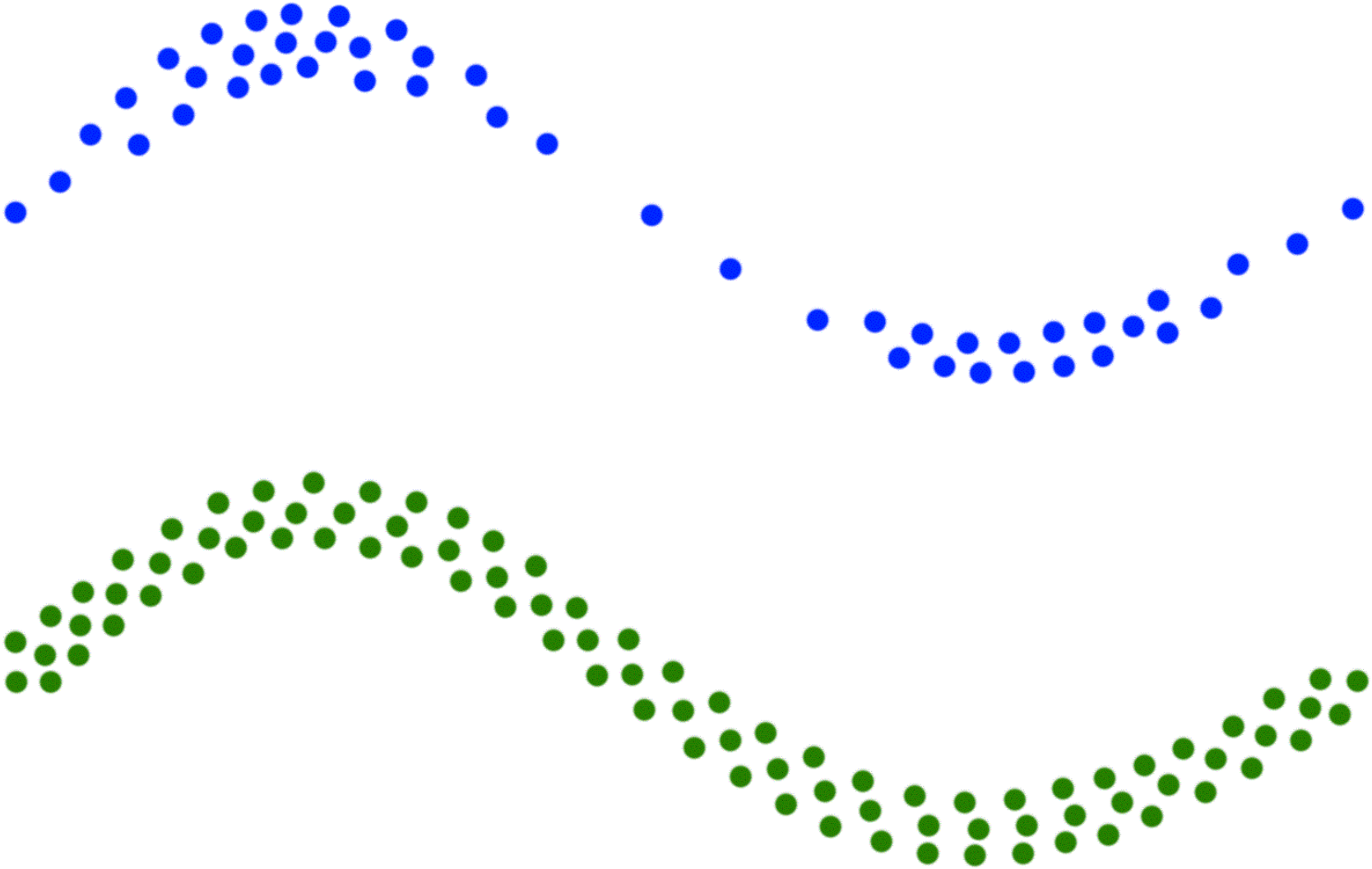
Introduction & motivation

New approach: geometry based data generation



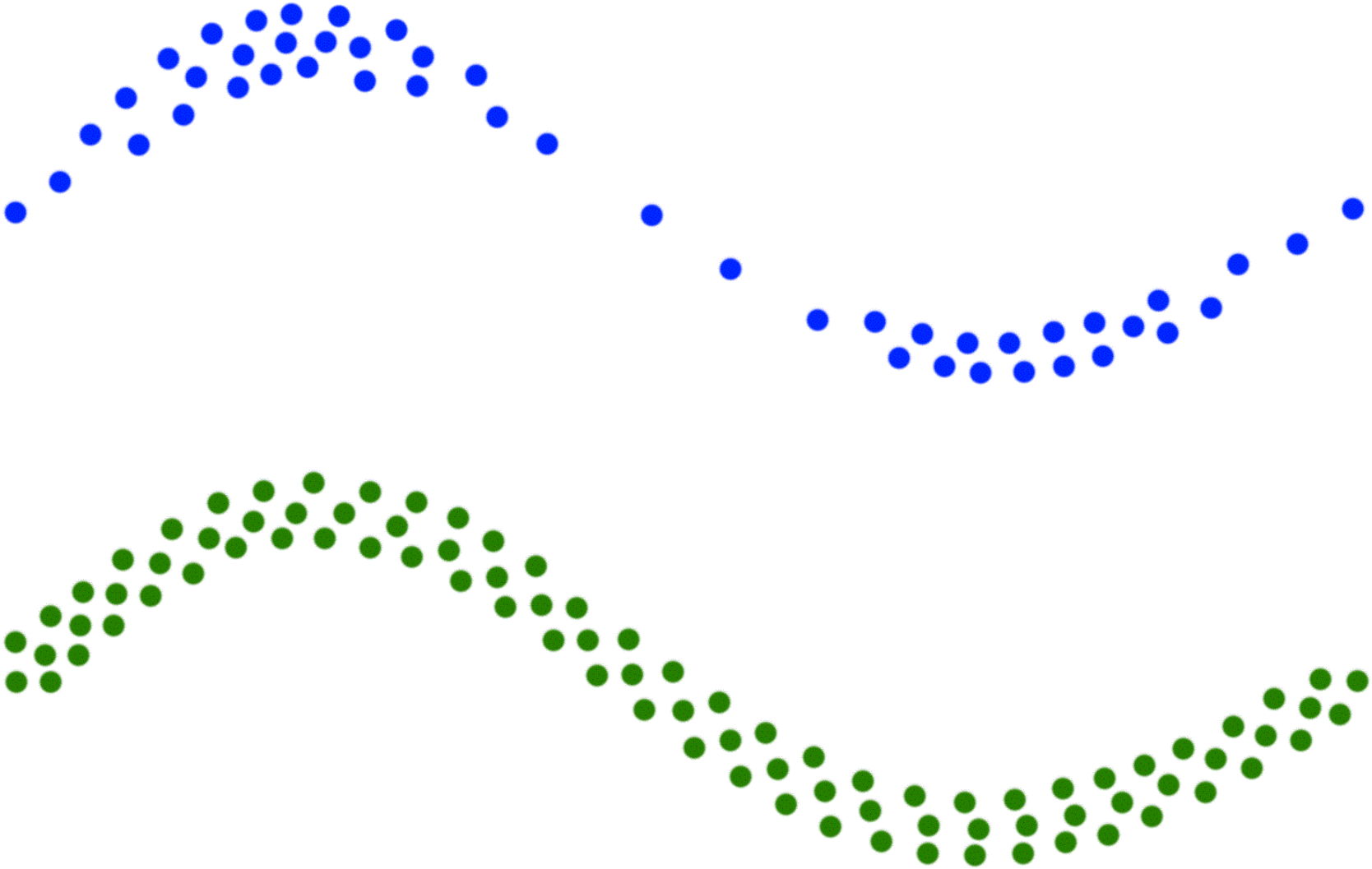
Introduction & motivation

New approach: geometry based data generation



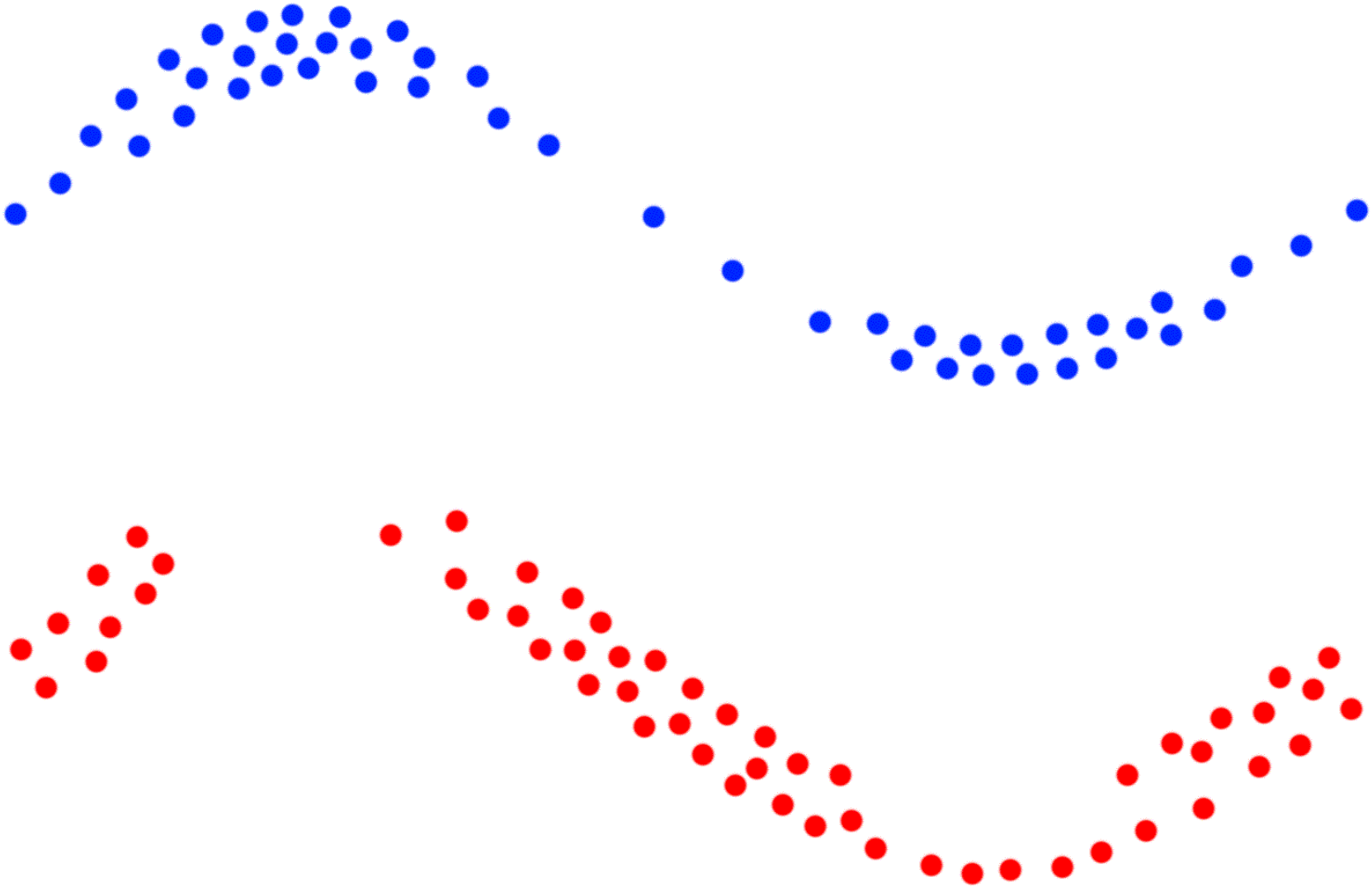
Introduction & motivation

New approach: geometry based data generation



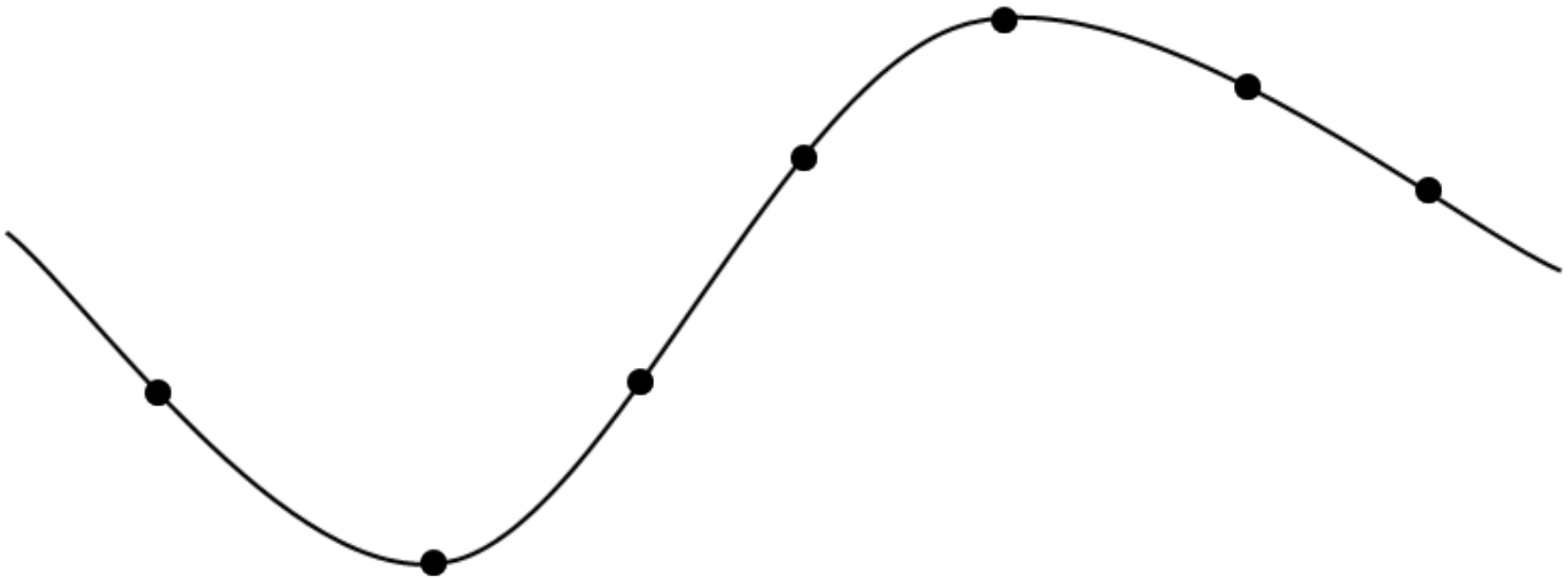
Introduction & motivation

New approach: geometry based data generation



Diffusion geometry

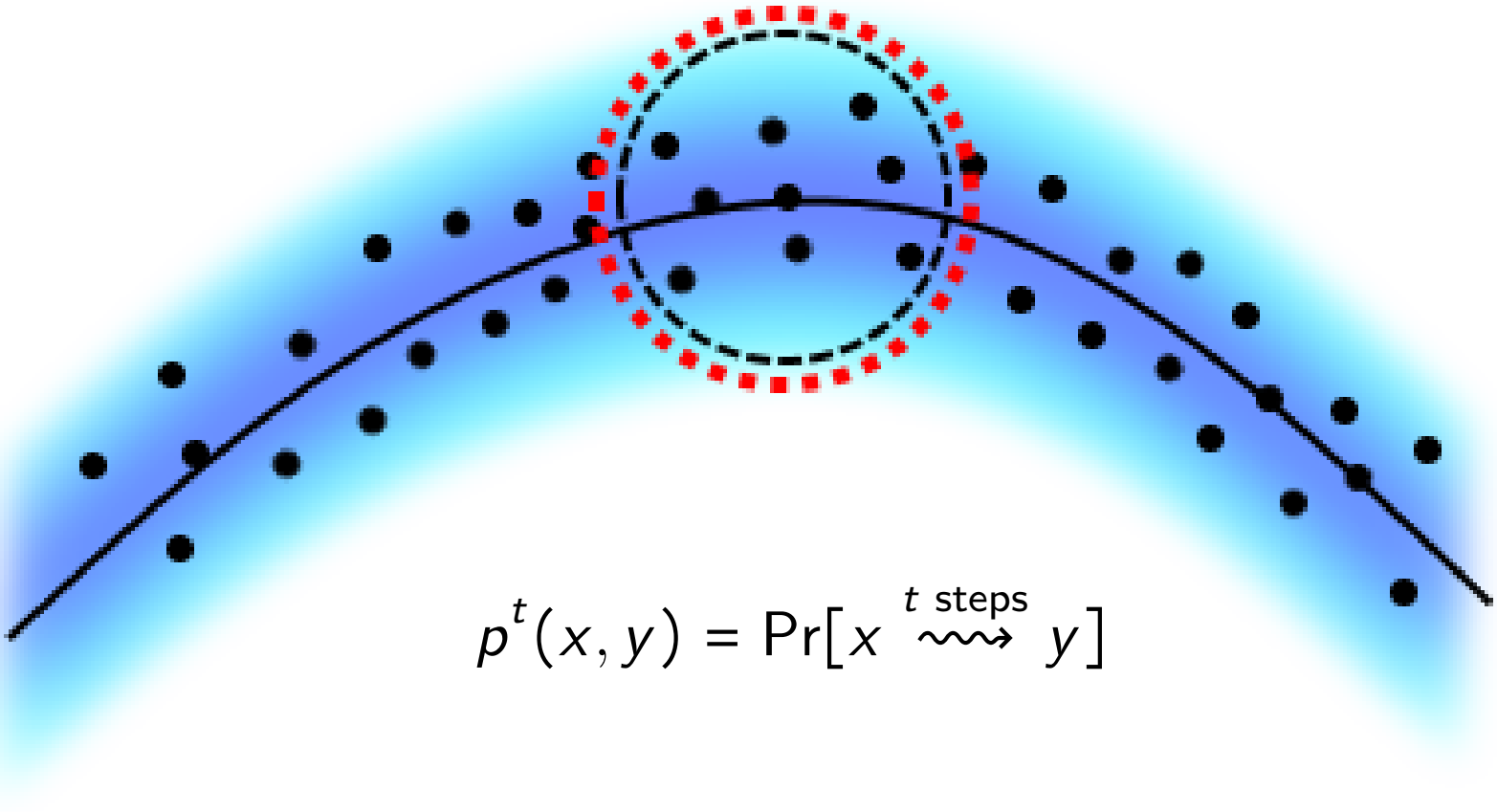
Manifold learning with random walks



- Local affinities $g(x, y) \Rightarrow$ transition probs. $\Pr[x \rightsquigarrow y] = \frac{g(x,y)}{\|g(x,\cdot)\|_1}$
- Markov chain/process $\Rightarrow$ random walks on data manifold

Diffusion geometry

Random walks reveal intrinsic neighborhoods

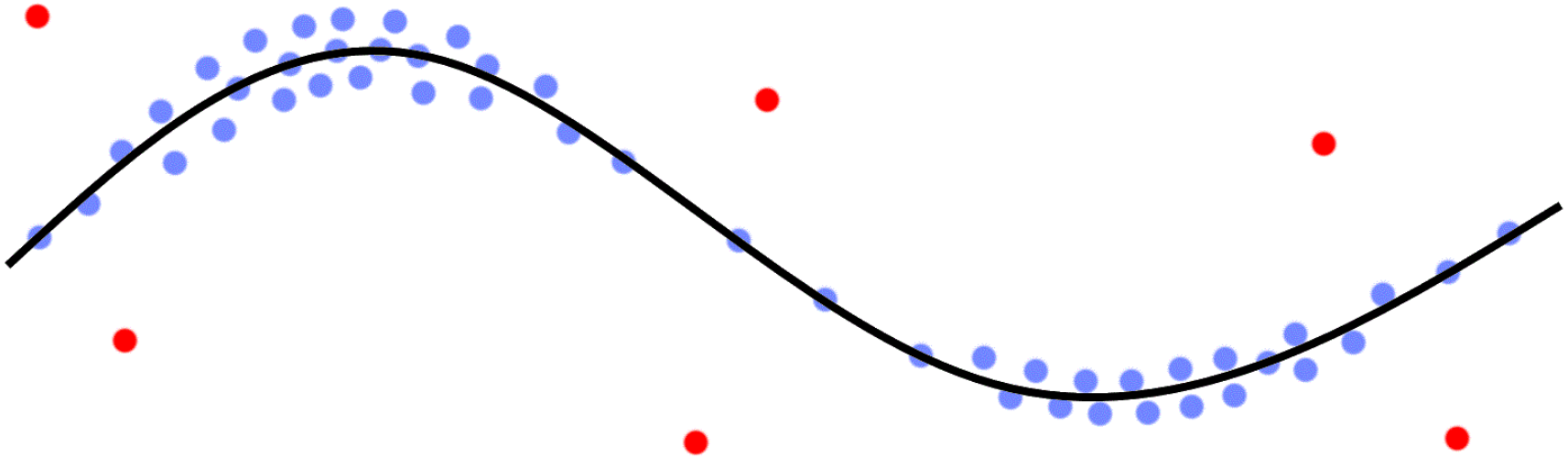


$$p^t(x, y) = \Pr[x \overset{t \text{ steps}}{\rightsquigarrow} y]$$

Data generation with diffusion

Walk toward the data manifold from randomly generated points

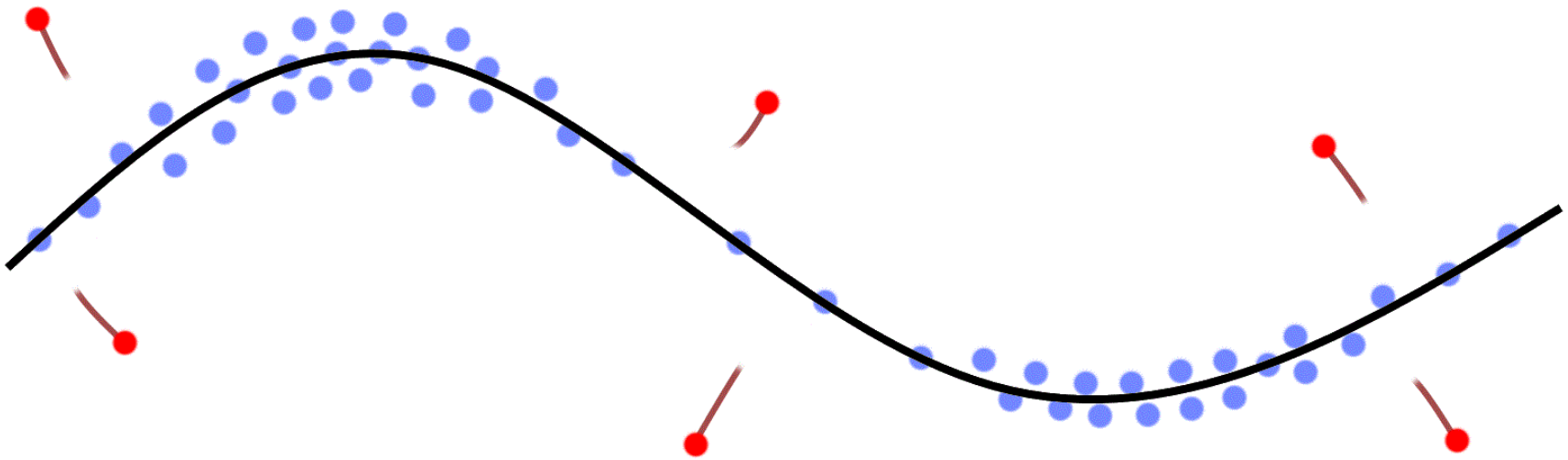
Generate random points:



Data generation with diffusion

Walk toward the data manifold from randomly generated points

Generate random points:

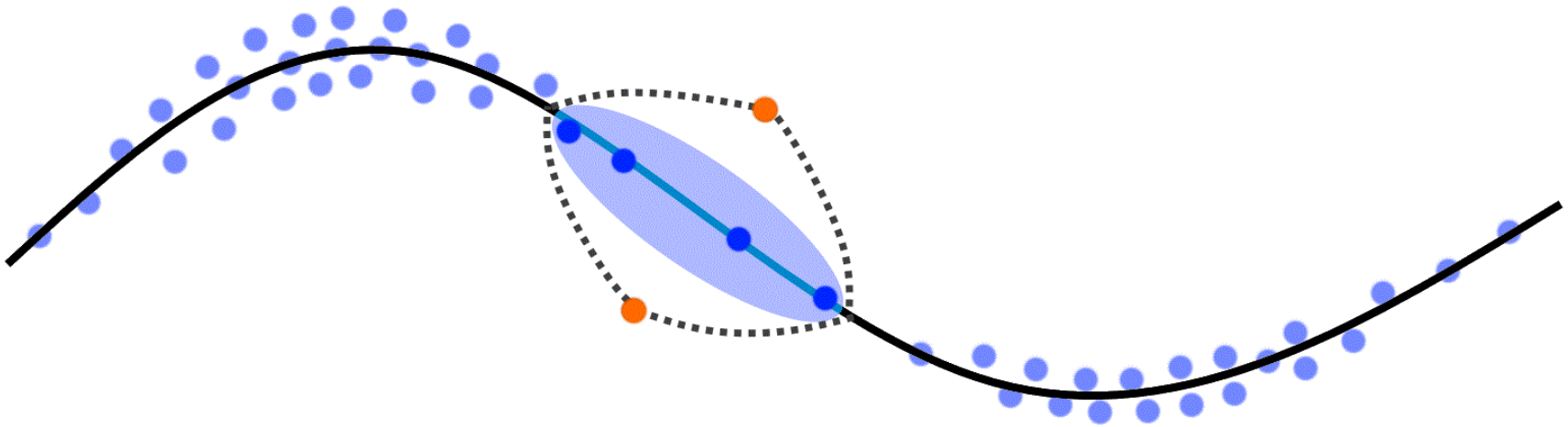


Walk towards the data manifold with diffusion: $x \mapsto \sum_{y \in \text{data}} y \cdot p^t(x, y)$

Data generation with diffusion

Correct density with MGC kernel (Bermanis et al., ACHA 2016)

Separate density/geometry with new kernel: $k(x,y) = \sum_{r \in \text{data}} \frac{g(x,r),g(y,r)}{\text{density}(r)}$

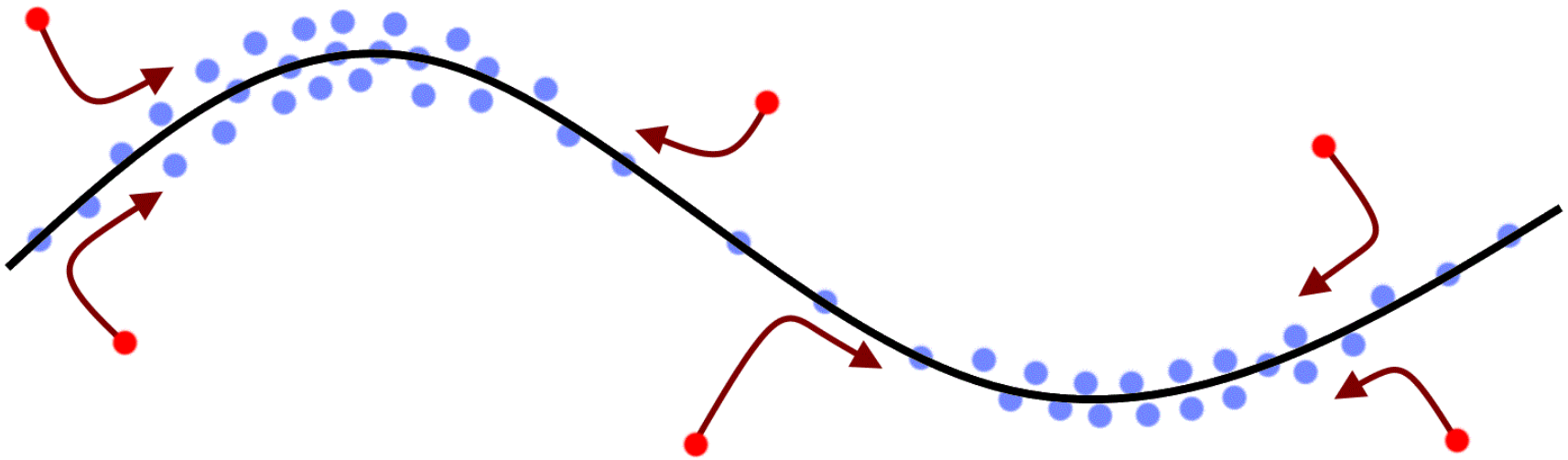


Use new diffusion process $p(x,y) = \frac{k(x,y)}{\|k(x,\cdot)\|_1}$ to walk to the manifold

Data generation with diffusion

Correct density with MGC kernel (Bermanis et al., ACHA 2016)

Separate density/geometry with new kernel: $k(x,y) = \sum_{r \in \text{data}} \frac{g(x,r),g(y,r)}{\text{density}(r)}$



Use new diffusion process $p(x,y) = \frac{k(x,y)}{\|k(x,\cdot)\|_1}$ to walk to the manifold

Data generation with diffusion

Fill sparse areas to create uniform distribution

Question: How should we initialize new points to end up with uniform sampling from the data manifold?

Answer: For each $x \in$ data, initialize $\hat{\ell}(x)$ points sampled from $\mathcal{N}(x, \Sigma_x)$; set $\hat{\ell}$ as the mid-point between the upper & lower bounds in the following proposition.

Proposition

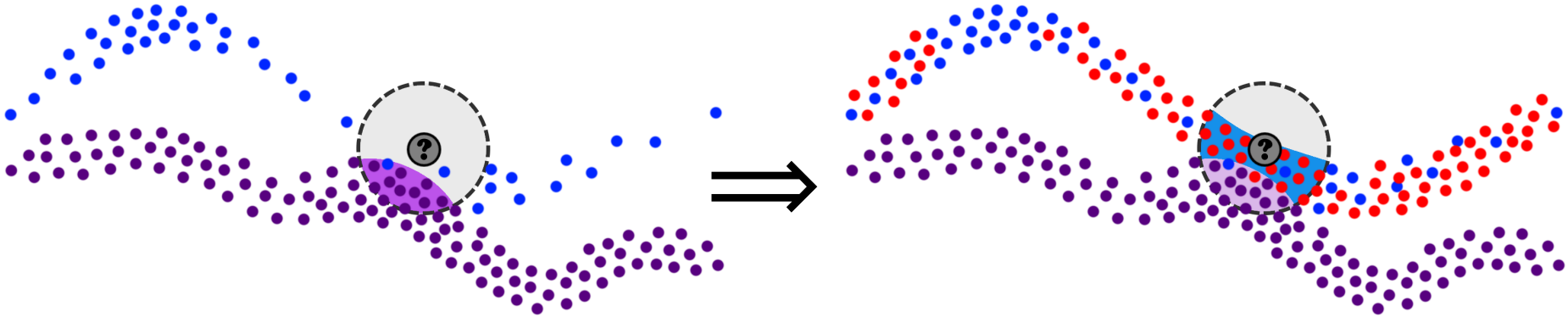
The generation level $\hat{\ell}(x)$ required to equalize density is bounded by

$$\det\left(I + \frac{\Sigma_x}{2\sigma^2}\right)^{\frac{1}{2}} \frac{\max(\hat{d}(\cdot)) - \hat{d}(x)}{\hat{d}(x)+1} - 1 \leq \hat{\ell}(x) \leq \det\left(I + \frac{\Sigma_x}{2\sigma^2}\right)^{\frac{1}{2}} \left[\max(\hat{d}(\cdot)) - \hat{d}(x)\right],$$

where σ is a scale used when defining Gaussian neighborhoods $g(x, y)$ for the diffusion geometry, and $\hat{d}(x) = \|g(x, \cdot)\|_1$ estimates local density.

Applications & results

Alleviating class imbalance in classification



	k-NN			SVM			RUSBoost
	Orig	**SMOTE**	**SUGAR**	**Orig**	**SMOTE**	**SUGAR**	
ACP	0.67	0.76	**0.78**	0.77	0.77	**0.78**	0.75
ACR	0.64	0.73	**0.77**	0.78	0.78	**0.84**	0.81
MCC	0.66	0.74	**0.78**	0.78	0.78	**0.84**	0.80

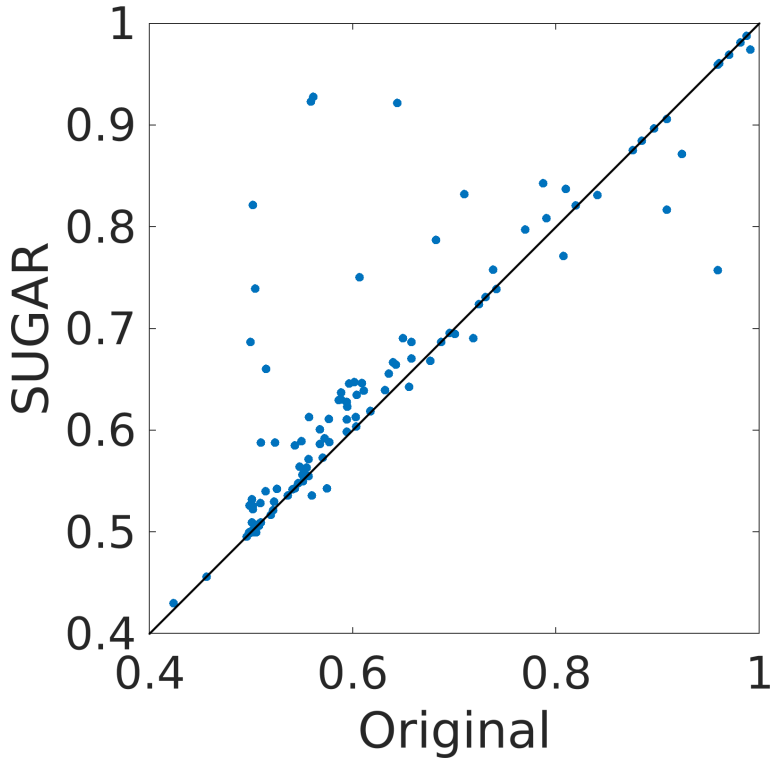
Average class precision/recall (ACP/ACR), and Matthews correlation coefficient (MCC) over **61 imbalanced datasets** (10-fold cross validation).

Applications & results

Density correction improves clustering

Spectral Clustering

Rand index of k-Means

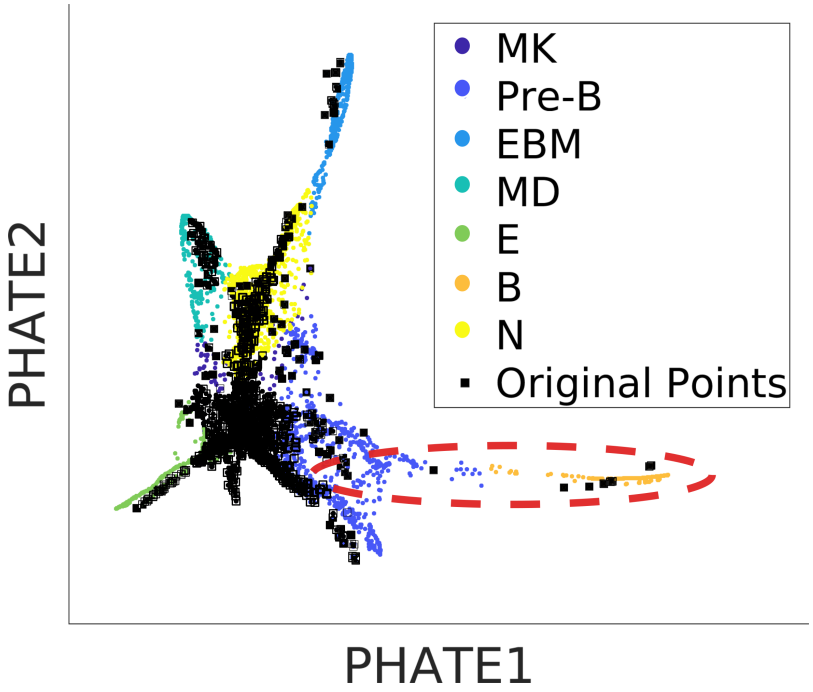


Based on **115 datasets**

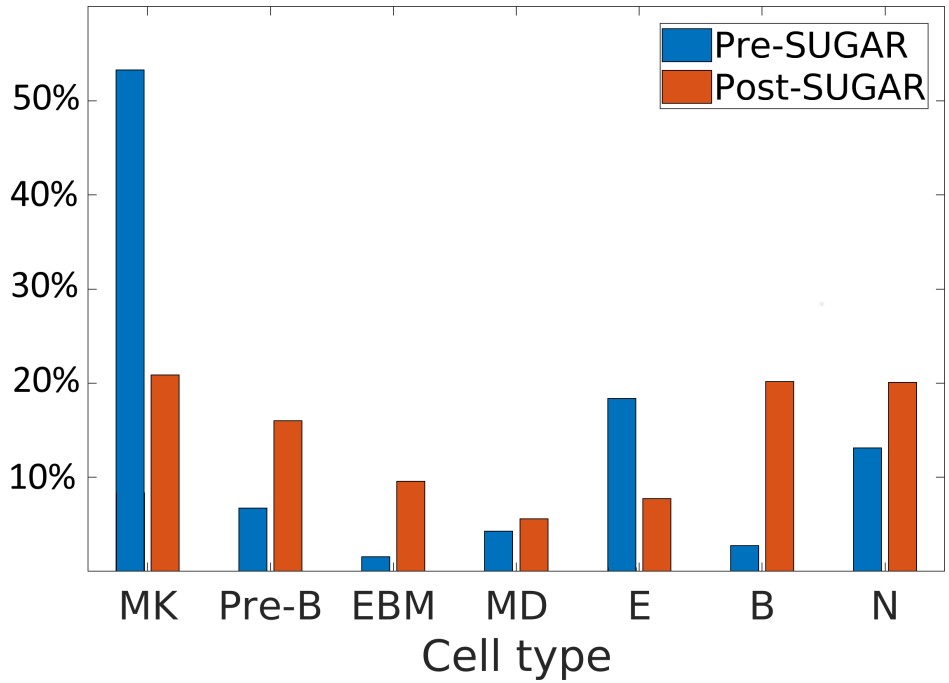
Applications & results

Illuminate hypothetical cell types in single-cell data from Velten et al. 2017

Recovering originally-undersampled lineage in early hematopoeisis:



B-cell maturation trajectory enhanced by SUGAR

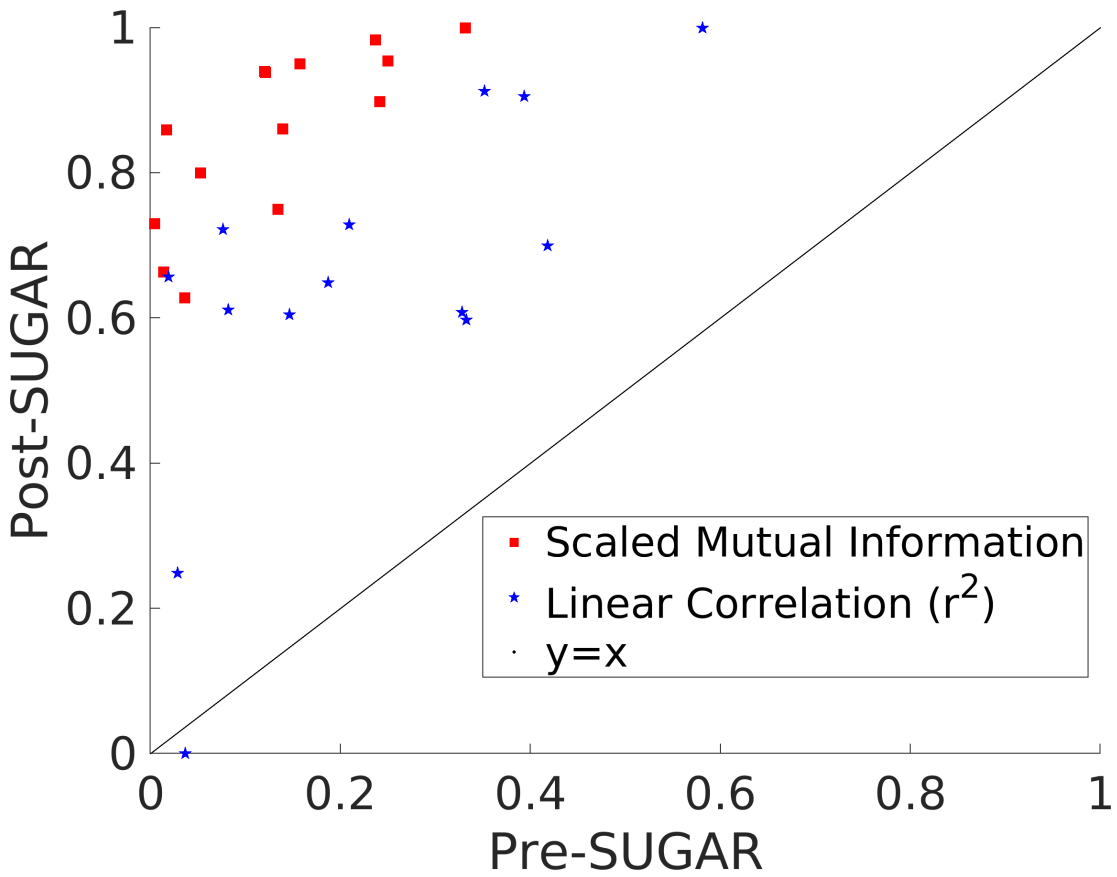


SUGAR equalizes the total cell distribution

Applications & results

Recover gene-gene relationships in single-cell data from Velten et al. 2017

SUGAR improves module correlation and MI identified by Velten et al.

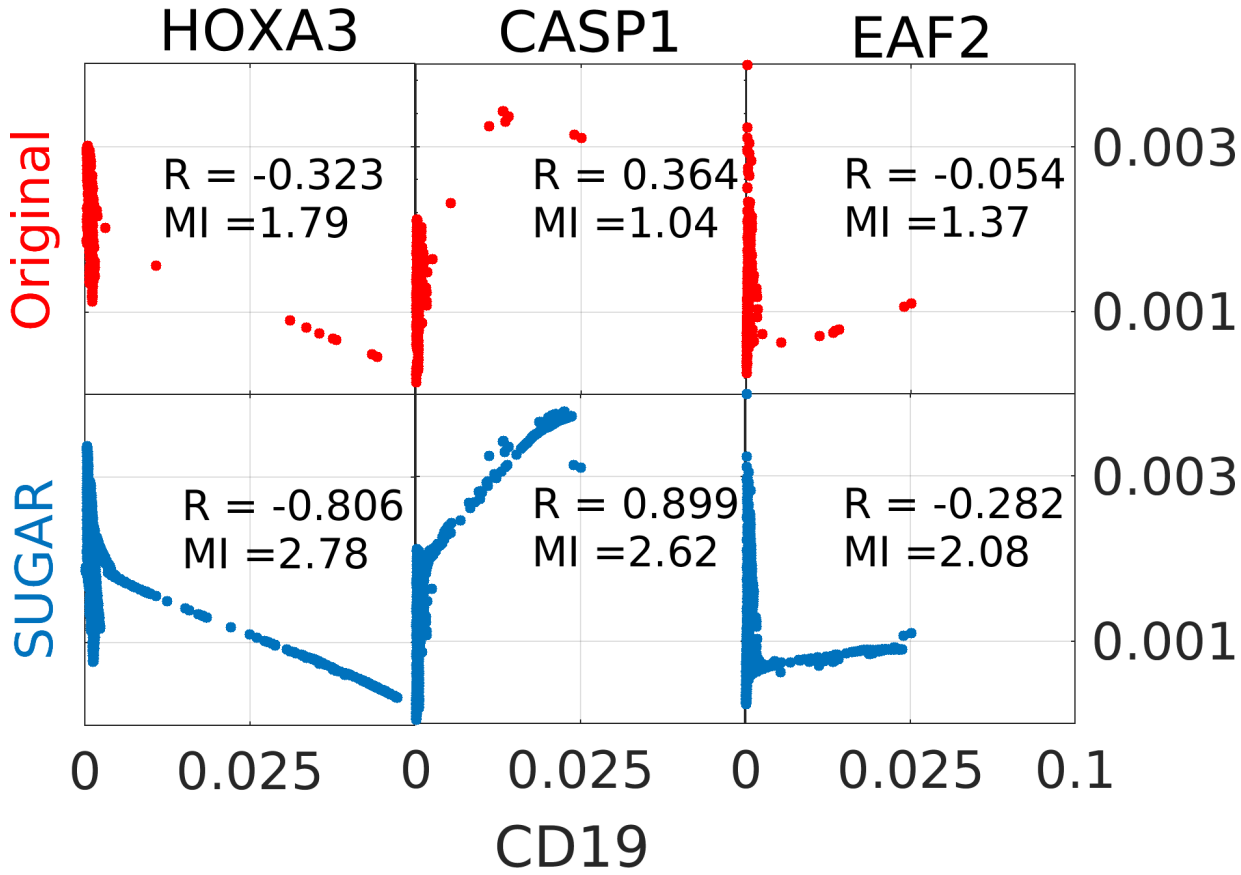


Velten et al., Nature Cell Biology, 19 (2017)

Applications & results

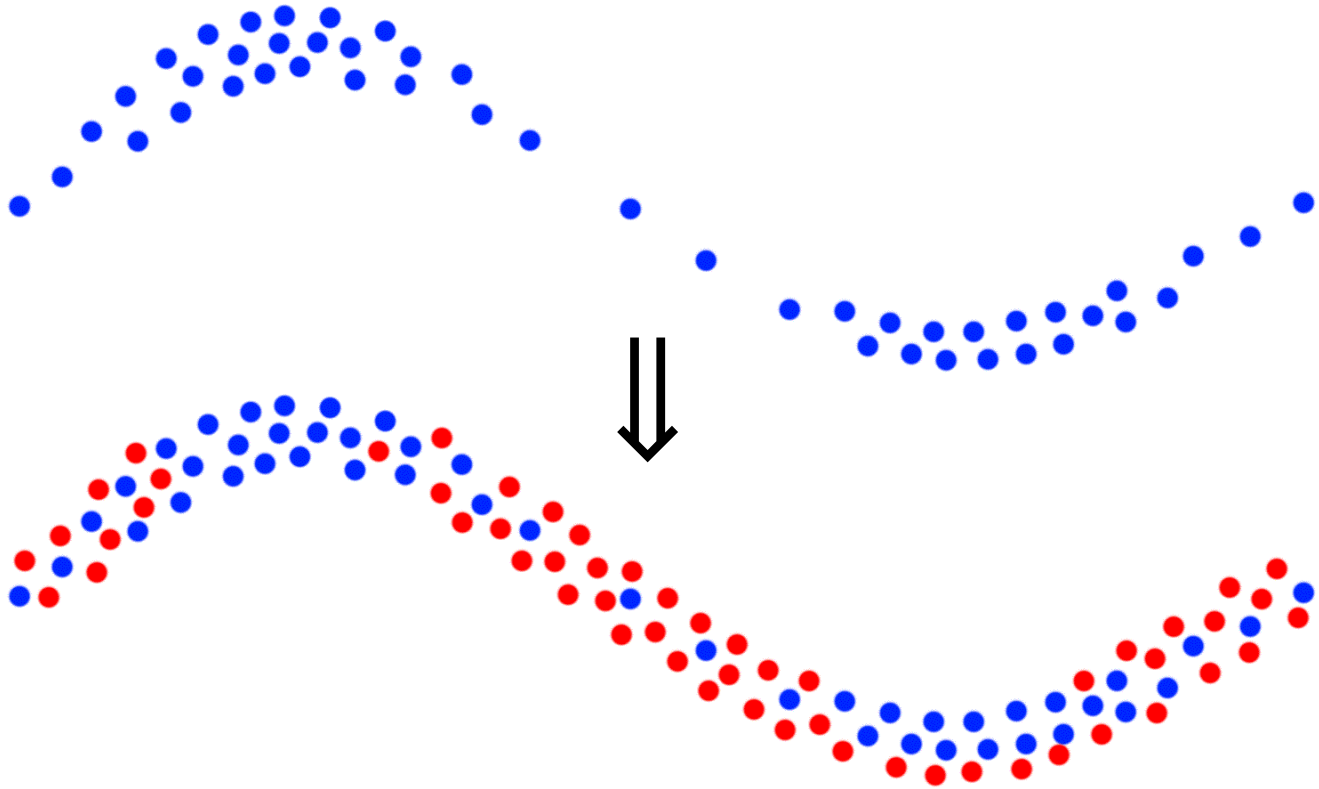
Recover gene-gene relationships in single-cell data from Velten et al. 2017

Generated cells also follow canonical marker correlations



Li et al., Nature communications 7 (2016)

Conclusion



- Generate data over intrinsic geometry rather than distribution
- Alleviate sampling bias in supervised & unsupervised learning
- Enable exploration of sparse (or "hypothetical") data regions