Computational Systems Biology
Deep Learning in the Life Sciences

6.802 6.874 20.390 20.490 HST.506

David Gifford

Lecture 6

February 25, 2019

The Zen of PCA, t-SNE, and Autoencoders

http://mit6874.github.io

What's on tap today!

- Embedding data in a lower dimensional space
- Linear reduction of dimensionality
  - Principle Component Analysis
- Non-linear embedding
  - t-distributed Stochastic Network Embedding (t-SNE)
  - Autoencoders

Dimensionality reduction has multiple applications

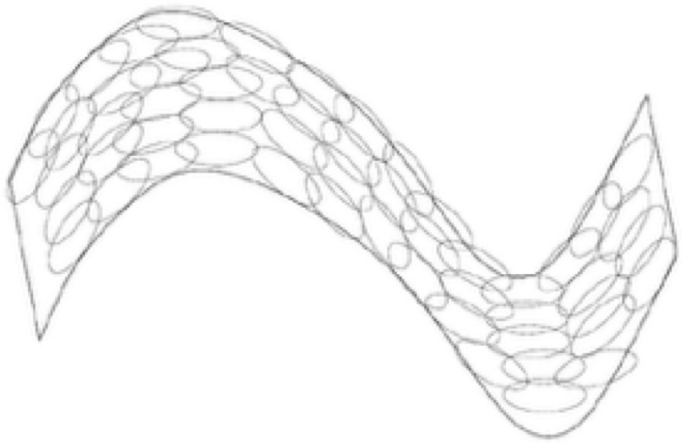
- Uses:
  - Data Visualization
  - Data Reduction
  - Data Classification
  - Trend Analysis
  - Factor Analysis
  - Noise Reduction

- Examples:
  - How many unique “sub-sets” are in the sample?
  - How are they similar / different?
  - What are the underlying factors that influence the samples?
  - Which time / temporal trends are (anti)correlated?
  - Which measurements are needed to differentiate?
  - How to best present what is “interesting”?
  - Which “sub-set” does this new sample rightfully belong?

A manifold is a topological space that locally resembles Euclidean space near each point

A manifold embedding is a structure preserving mapping of a high dimensional space into a manifold

Manifold learning learns a lower dimensional space that enables a manifold embedding

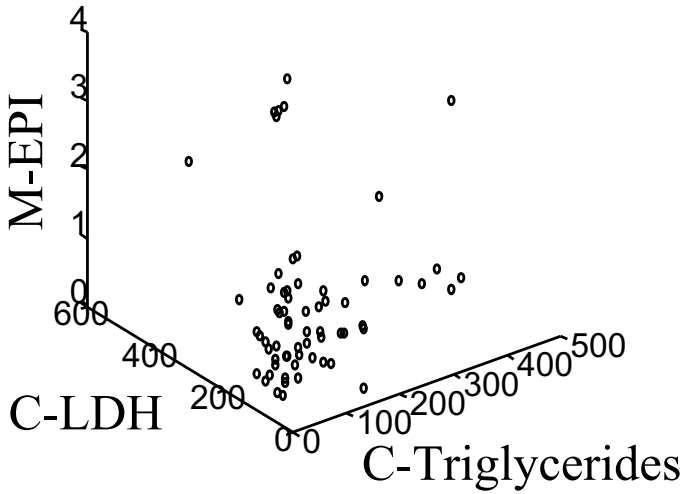


1. Principal Component Analysis

Example data

- Example: 53 Blood and urine measurements (wet chemistry) from 65 people (33 alcoholics, 32 non-alcoholics)

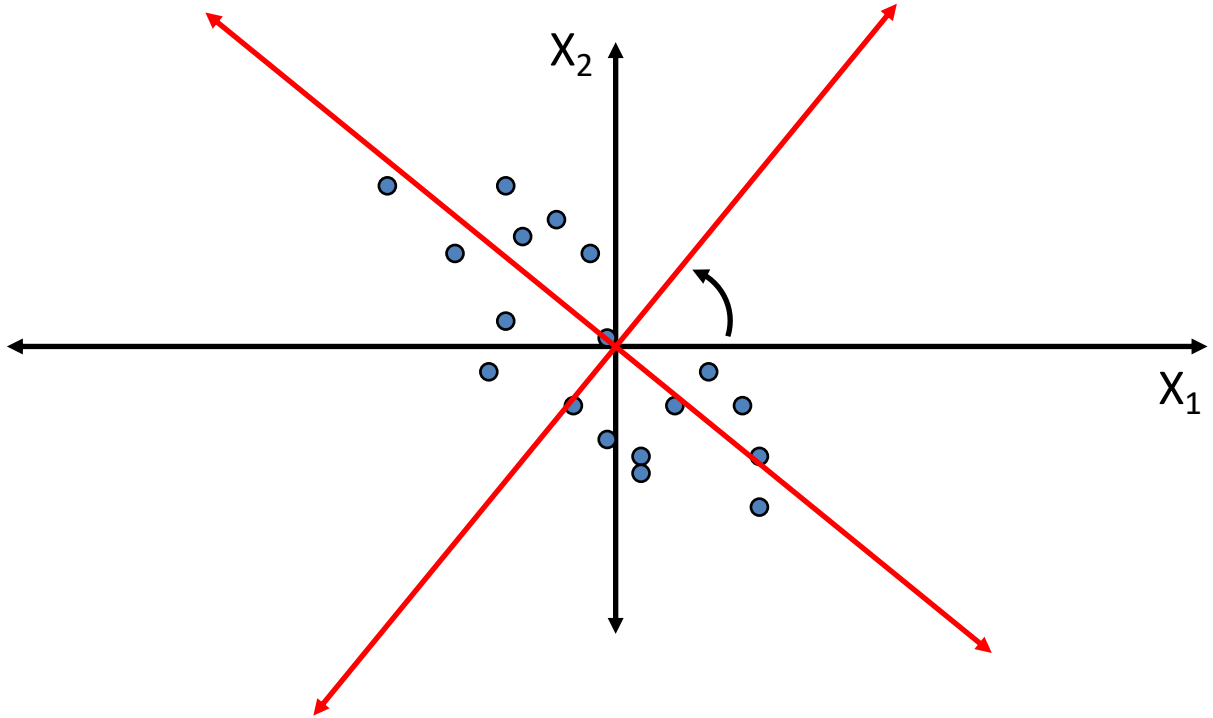
- Trivariate plot



	H-WBC	H-RBC	H-Hgb	H-Hct	H-MCV	H-MCH	H-MCHC
A1	8.0000	4.8200	14.1000	41.0000	85.0000	29.0000	34.0000
A2	7.3000	5.0200	14.7000	43.0000	86.0000	29.0000	34.0000
A3	4.3000	4.4800	14.1000	41.0000	91.0000	32.0000	35.0000
A4	7.5000	4.4700	14.9000	45.0000	101.0000	33.0000	33.0000
A5	7.3000	5.5200	15.4000	46.0000	84.0000	28.0000	33.0000
A6	6.9000	4.8600	16.0000	47.0000	97.0000	33.0000	34.0000
A7	7.8000	4.6800	14.7000	43.0000	92.0000	31.0000	34.0000
A8	8.6000	4.8200	15.8000	42.0000	88.0000	33.0000	37.0000
A9	5.1000	4.7100	14.0000	43.0000	92.0000	30.0000	32.0000

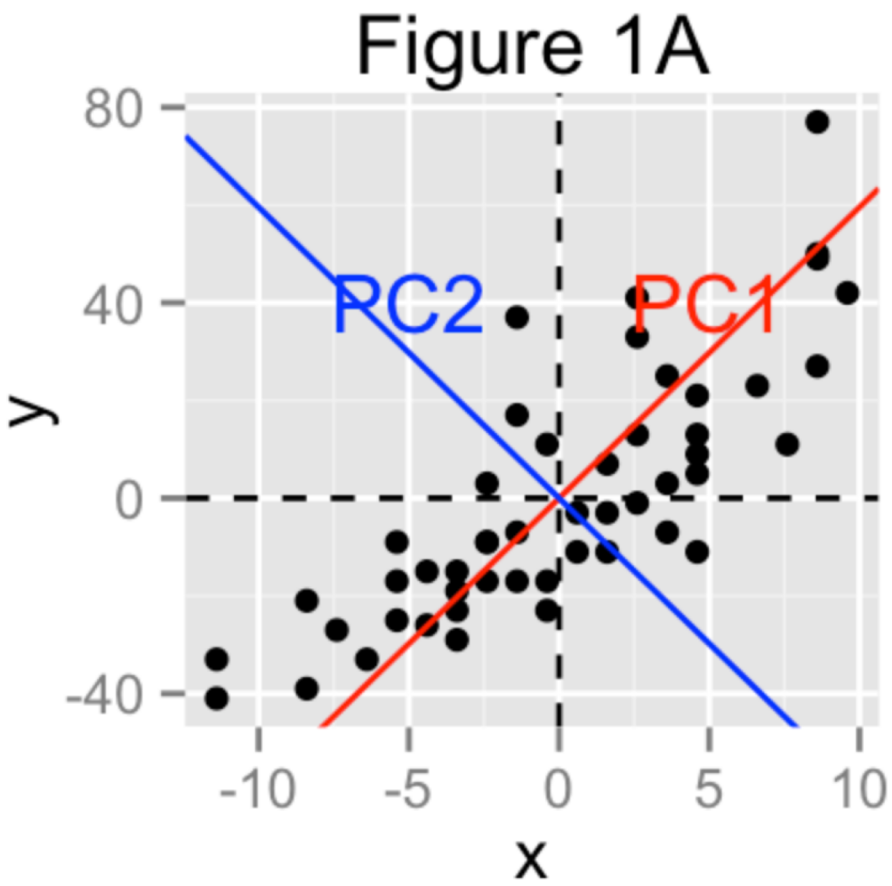
Principal Component = axis of greatest variability

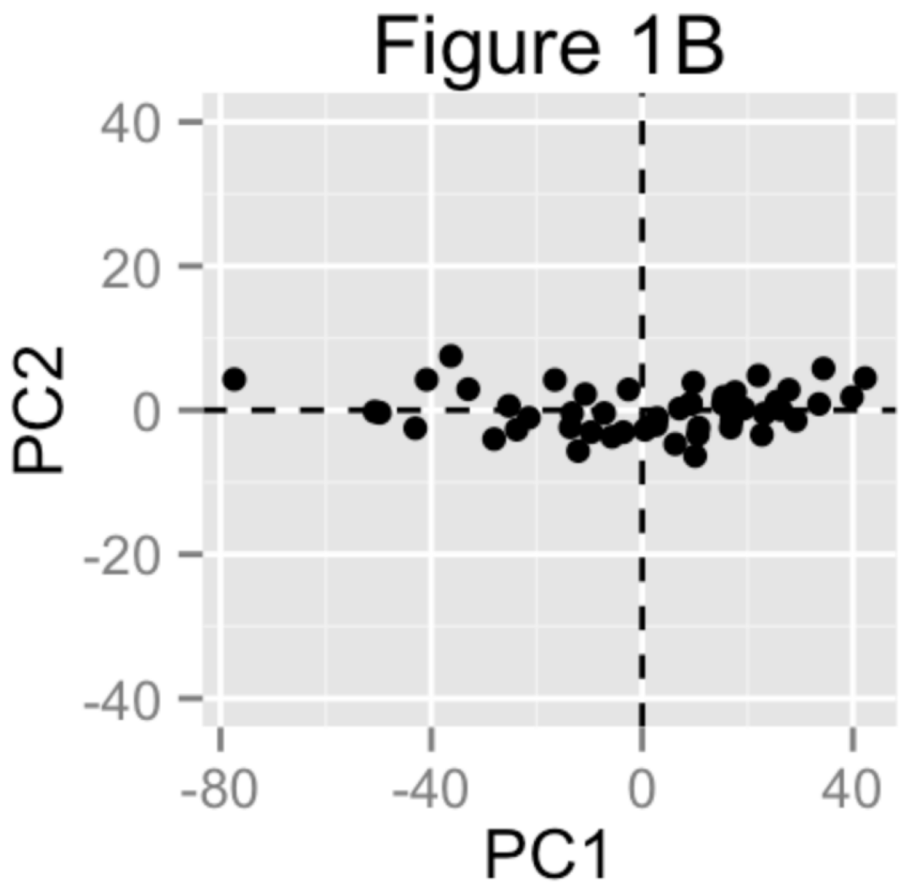
Suppose we have a population measured on p random variables $X_1,\ldots,X_p$. Note that these random variables represent the p-axes of the Cartesian coordinate system in which the population resides. Our goal is to develop a new set of p axes (linear combinations of the original p axes) in the directions of greatest variability:



This is accomplished by rotating the axes.

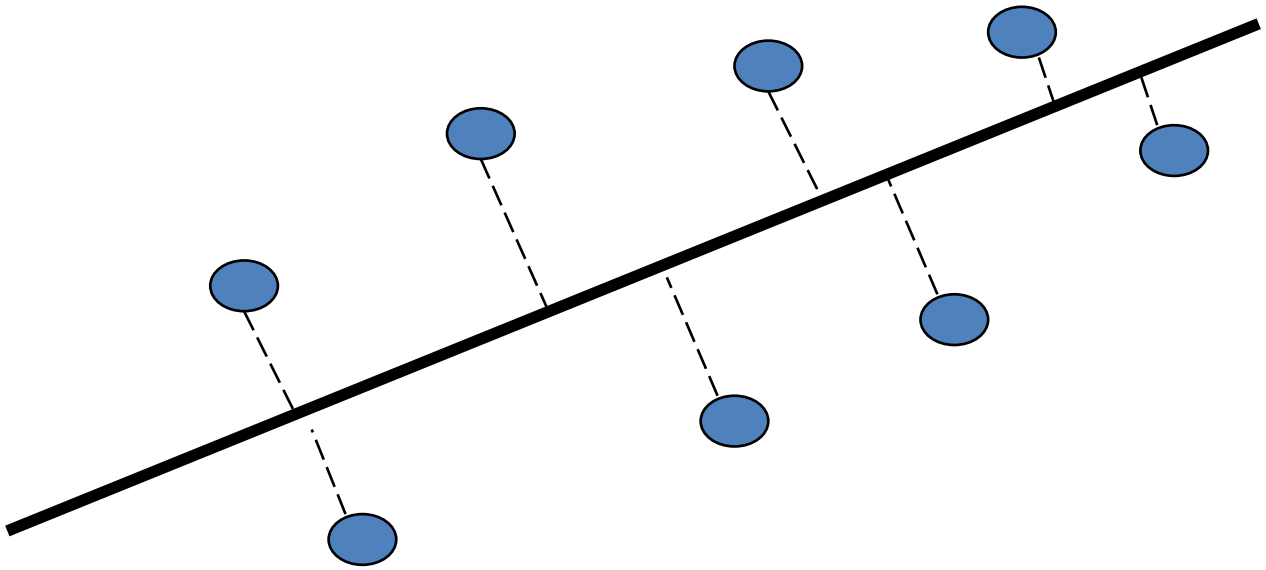
Data projected onto PC1





Selecting Principal Components

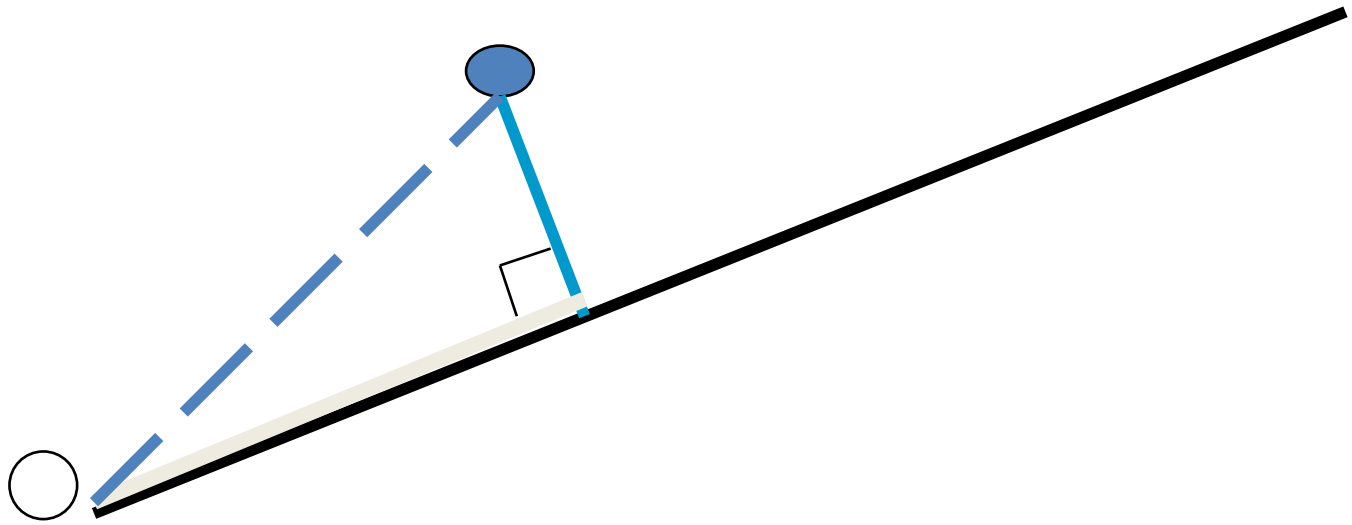
- Given m points in a n dimensional space, for large n, how does one project on to a 1 dimensional space?
- Formally, minimize sum of squares of distances to the line.



- Why sum of squares? Because it allows fast minimization, assuming the line passes through 0

We maximize the sum of square of projections to minimize the sum of squares distances

- Minimizing sum of squares of distances to the line is the same as maximizing the sum of squares of the projections on that line, thanks to Pythagoras.



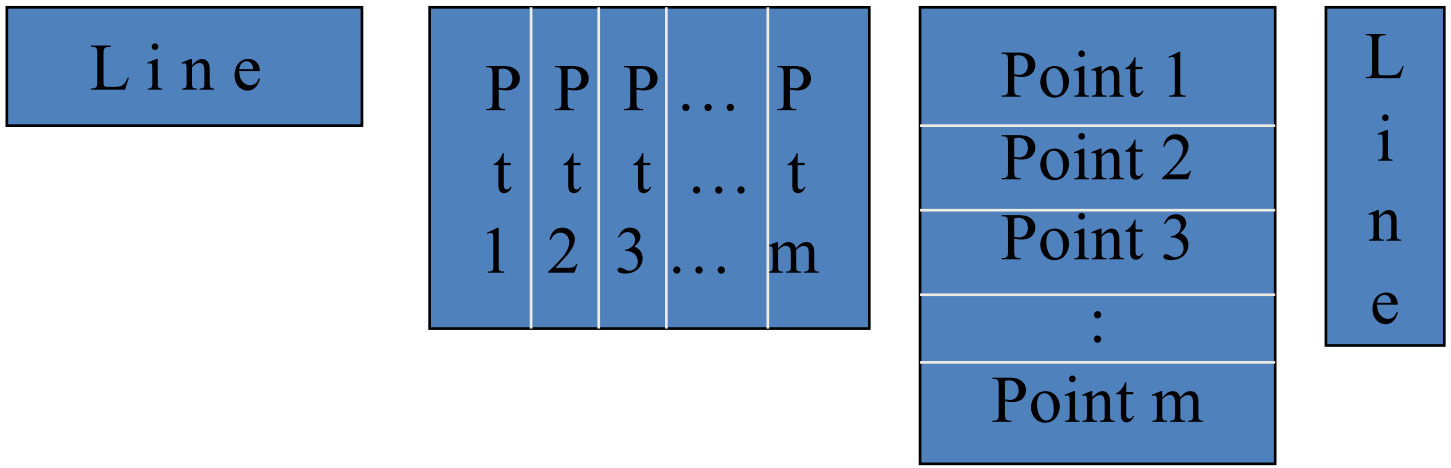
Sum of squares of projections for all m points onto vector x in matrix form

$$\max(v^T A^T A v), \text{ subject to } v^T v = 1$$

$$\Sigma = A^T A$$

$$v^T \Sigma v = \lambda^2$$

$$\Sigma v = \lambda^2 v$$



v^T A^T A v

[1, n] [n, m] [m, n] [n, 1]

Principle Component Analysis (PCA)

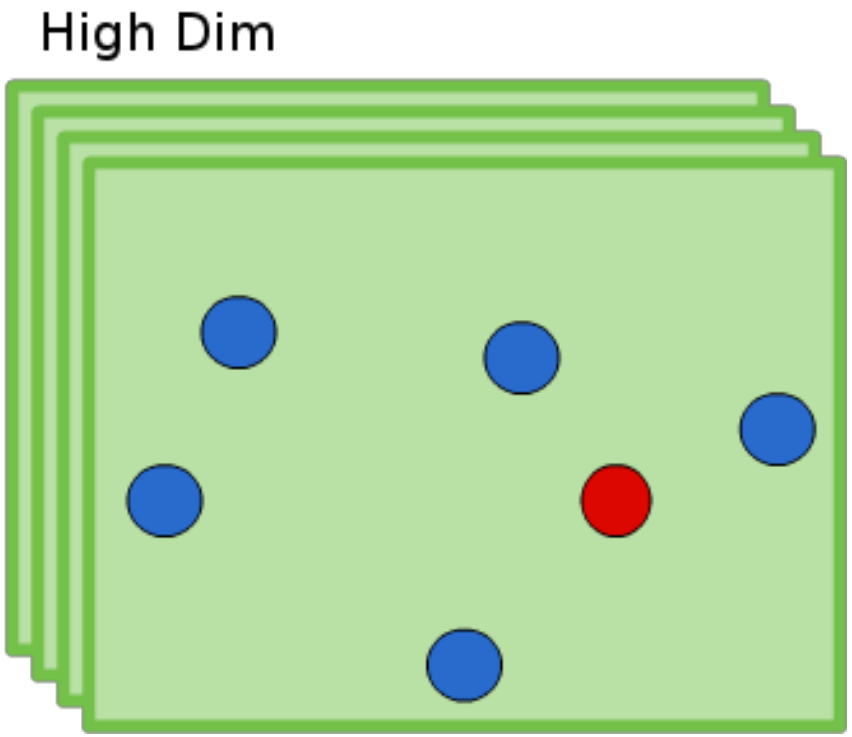
- How do we find the eigenvectors v_i?
- We use singular value decomposition to decompose Σ into an orthogonal rotation matrix U and a diagonal scaling matrix S:

$$\begin{aligned} \Sigma &= USU^T && (22) \\ \Sigma U &= (USU^T)U && (23) \\ &= US && (24) \end{aligned}$$

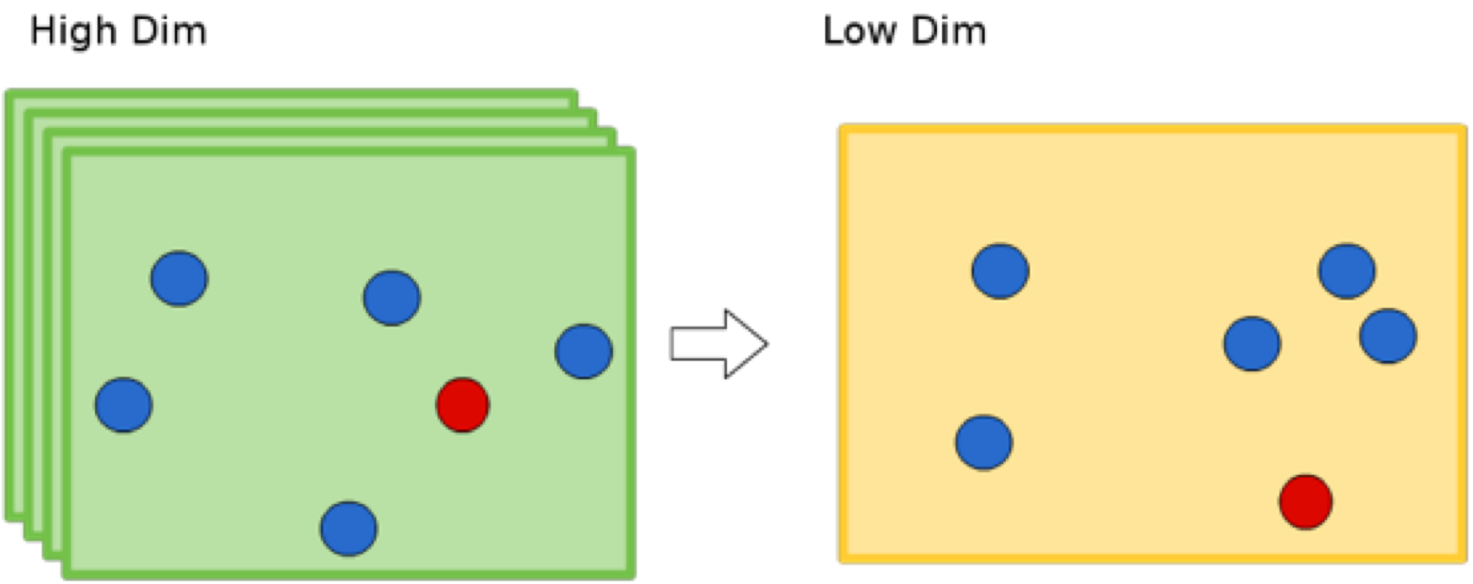
- The columns of U are the v_i, and S is the diagonal matrix of eigenvalues λ_i^2

2. tSNE non-linear embedding

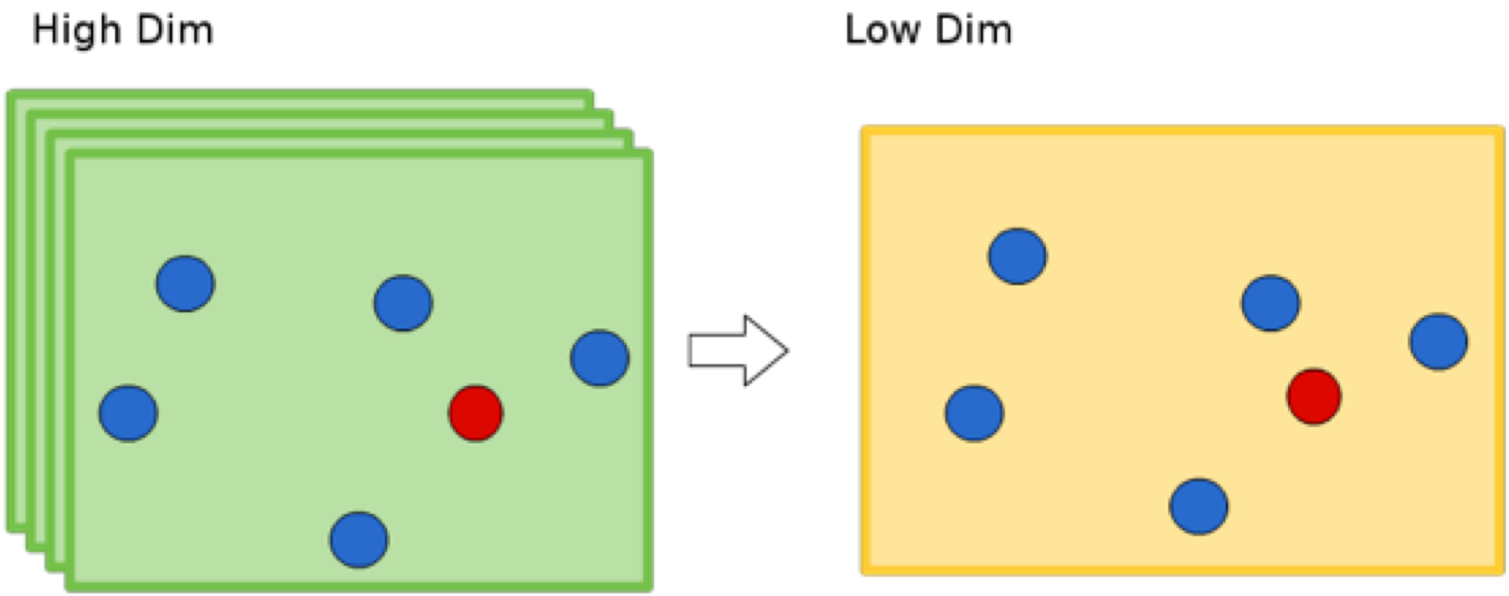
Distance Preservation Neighbor Preservation



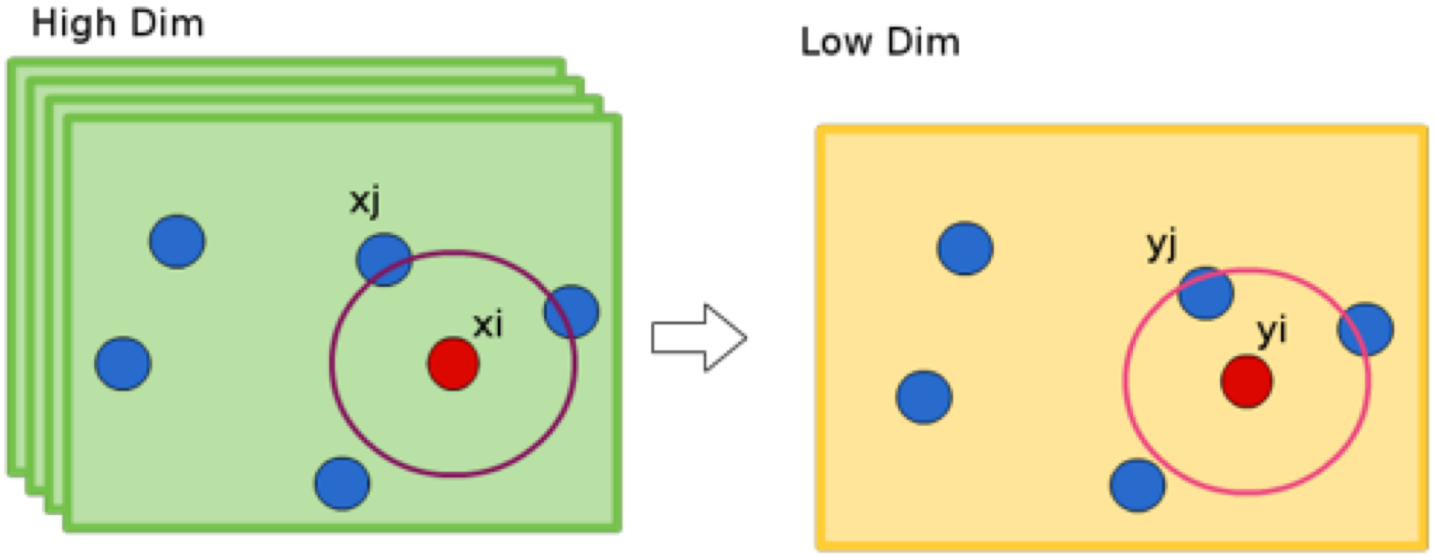
Neighborhood not preserved



Neighborhood preserved



Measure pairwise distances in high dimensional space



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

Set the bandwidth σ_i such that the conditional has a fixed perplexity (effective number of neighbors) $Perp(P_i) = 2^{H(P_i)}$, typical value is about 5 to 50

We want to choose an embedding that minimizes divergence between low and high dimension similarities

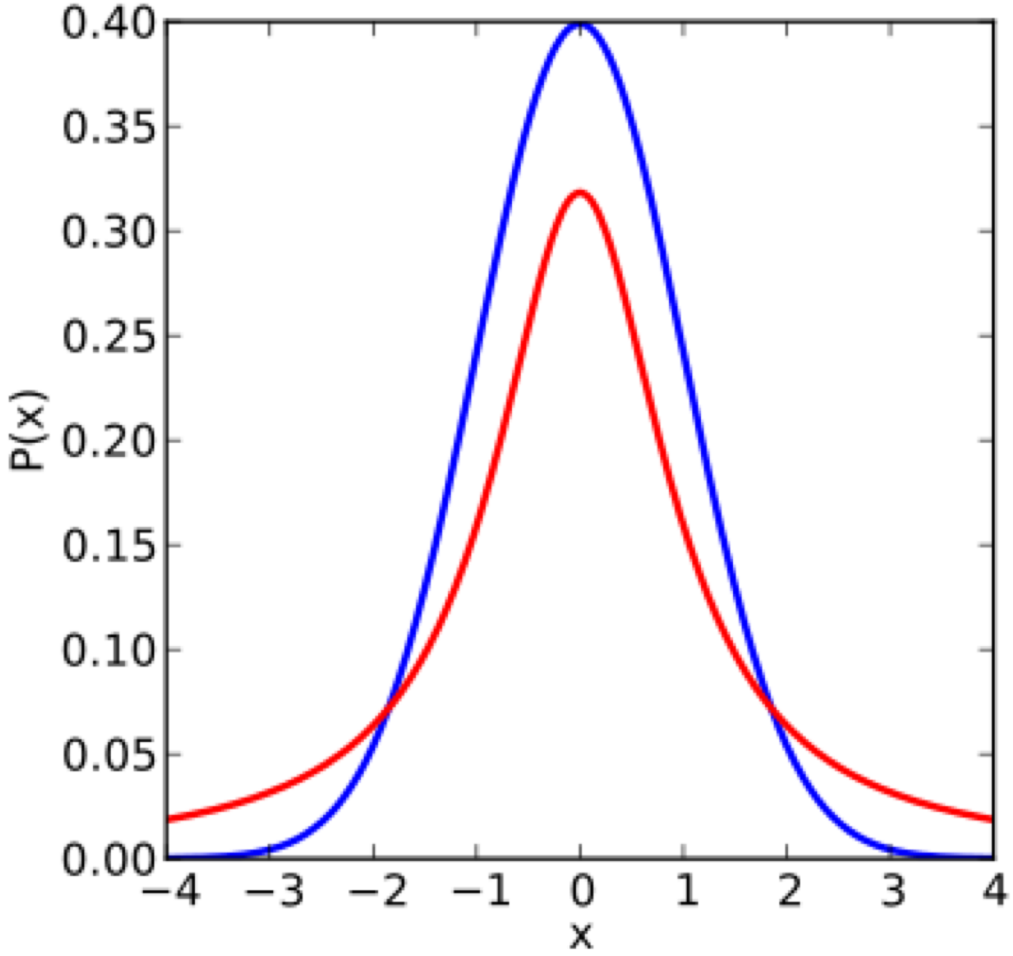
- Similarity of datapoints in High Dimension

$$p_{ij} = \frac{exp(-||x_i - x_j||^2/2\sigma^2)}{\sum_{k \neq l} exp(-||x_l - x_k||^2/2\sigma^2)}$$

- Similarity of datapoints in Low Dimension

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k \neq l}(1 + ||y_k - y_l||^2)^{-1}}$$

Low dimensional embedding using a Student t-distribution to avoid overcrowding



Red – Student t-distribution (1 degree of freedom)
Blue - Gaussian

We can use gradient methods to find an embedding

- Cost function

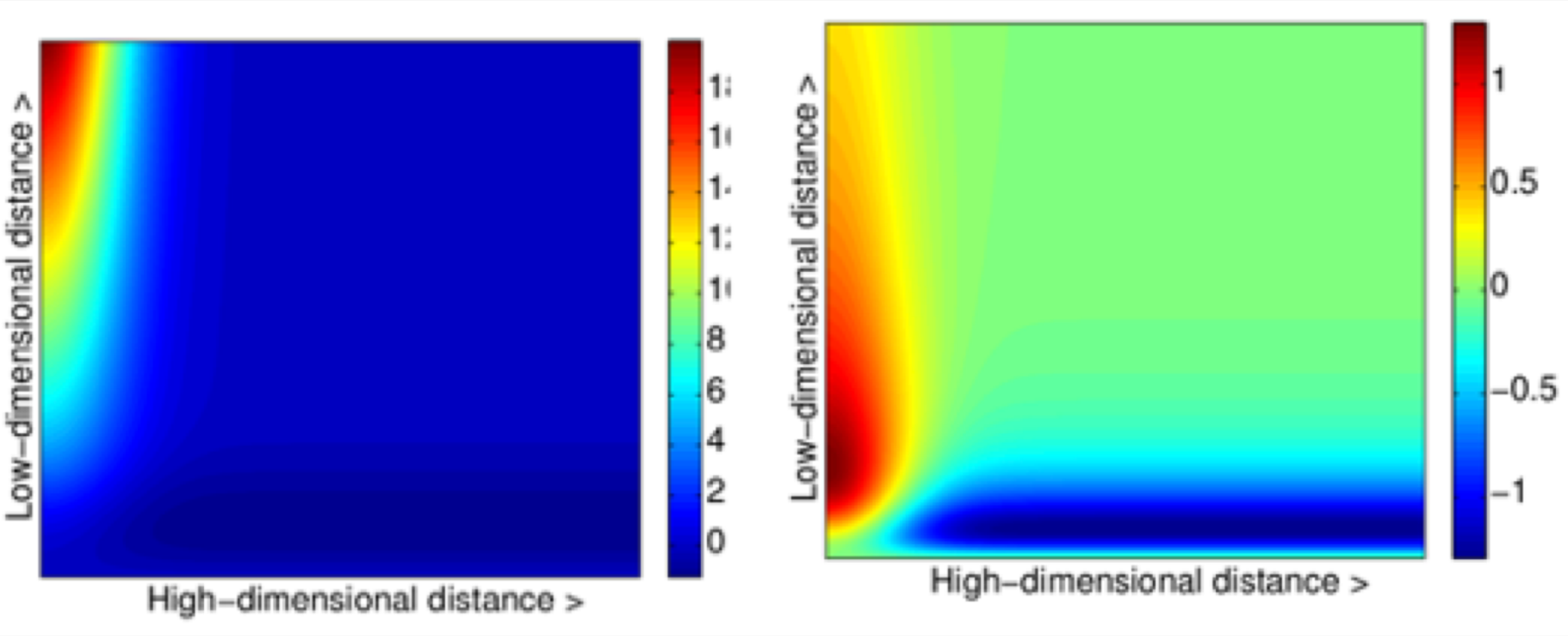
$$C = KL(P||Q) = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}}$$

  - Large p_{ij} modeled by small q_{ij}: Large penalty
  - Small p_{ij} modeled by large q_{ij}: Small penalty
  - t-SNE mainly preserves local similarity structure of the data

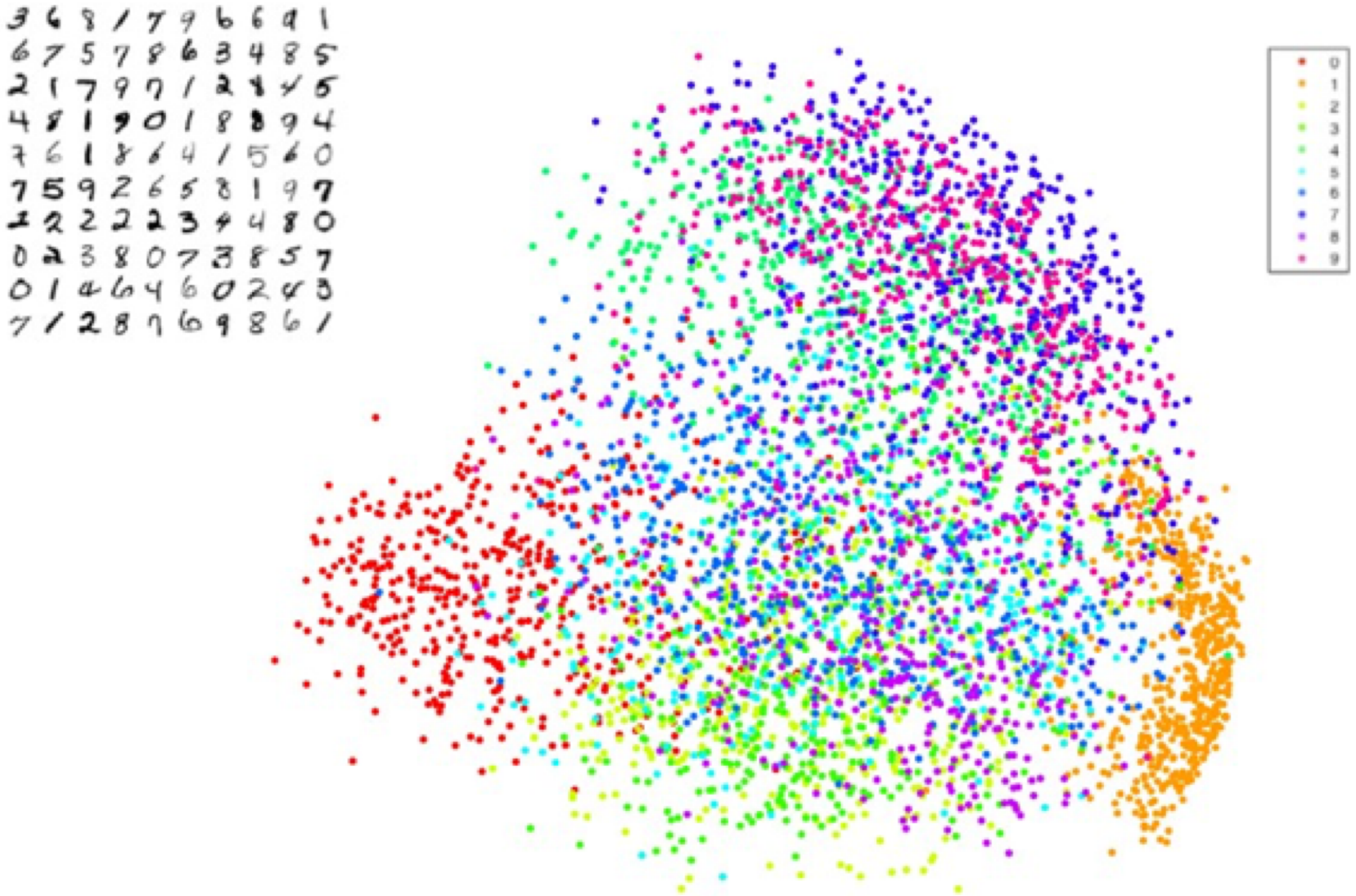
- Gradient

$$\frac{\partial C}{\partial y_i} = 4 \sum_{j \neq i} (p_{ij} - q_{ij})(1 + ||y_i - y_j||^2)^{-1}(y_i - y_j)$$

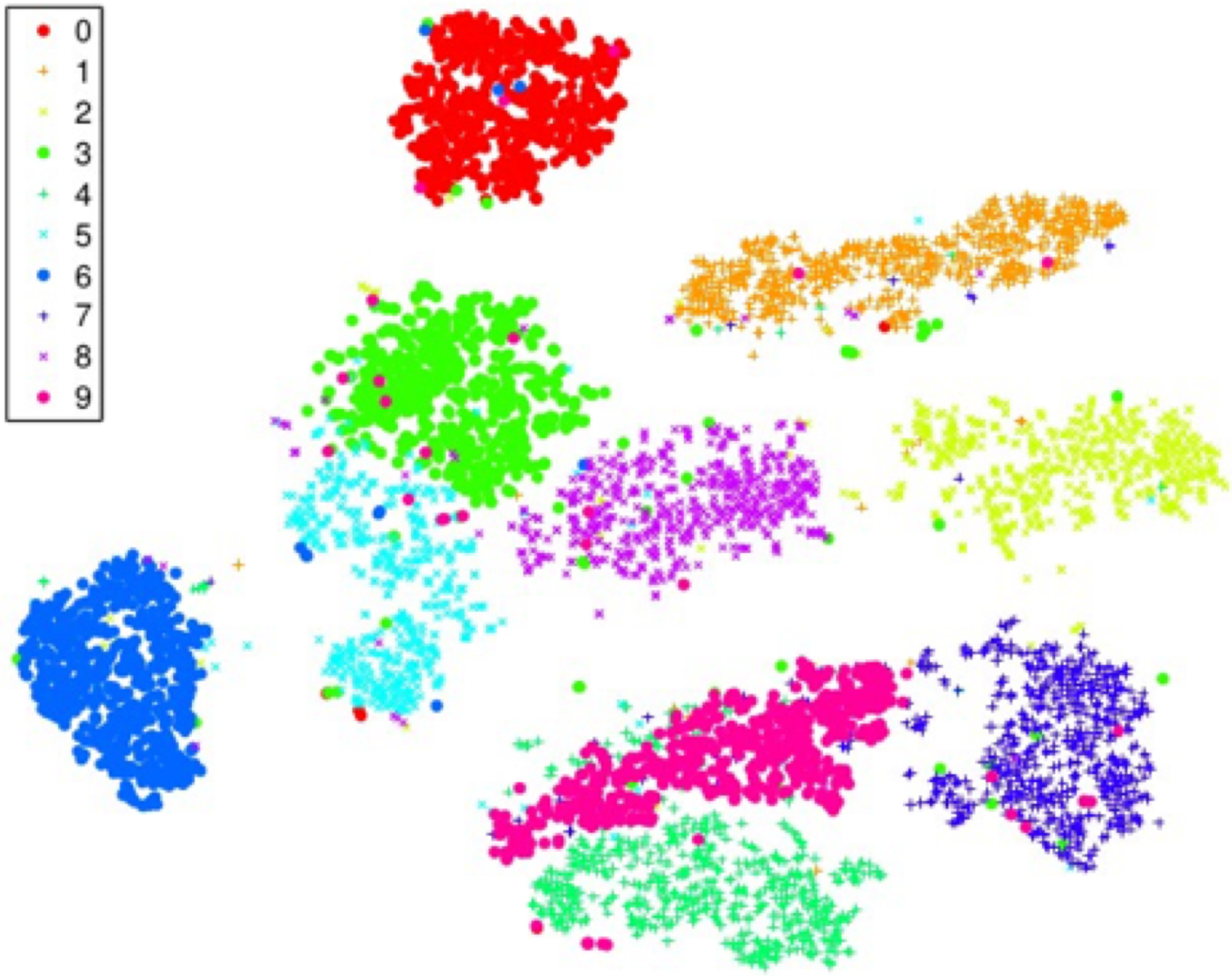
Interpretation of SNE (left) and t-SNE (right) gradients



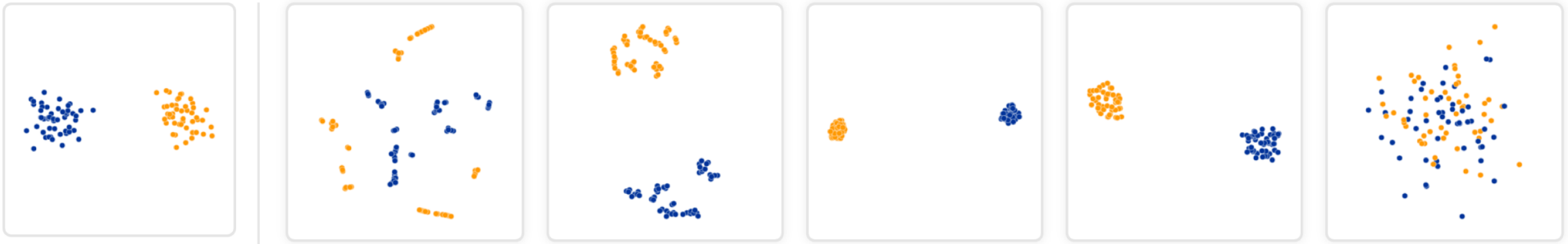
PCA of MNIST digits



t-SNE of MNIST digits



Perplexity matters



Original

Perplexity: 2
Step: 5,000

Perplexity: 5
Step: 5,000

Perplexity: 30
Step: 5,000

Perplexity: 50
Step: 5,000

Perplexity: 100
Step: 5,000

https://distill.pub/2016/misread-tsne/

Number of steps matter

Original

Perplexity: 30
Step: 10

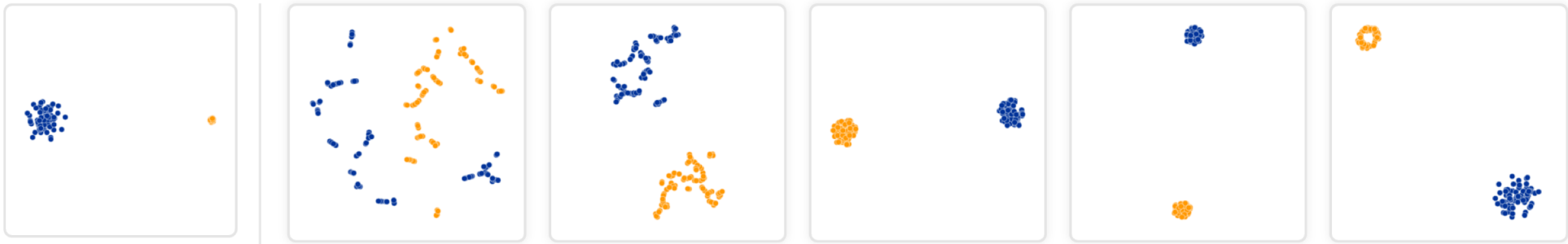
Perplexity: 30
Step: 20

Perplexity: 30
Step: 60

Perplexity: 30
Step: 120

Perplexity: 30
Step: 1,000

Cluster sizes are not meaningful



Original

Perplexity: 2
Step: 5,000

Perplexity: 5
Step: 5,000

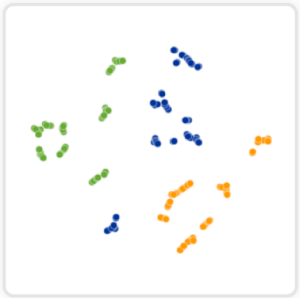
Perplexity: 30
Step: 5,000

Perplexity: 50
Step: 5,000

Perplexity: 100
Step: 5,000

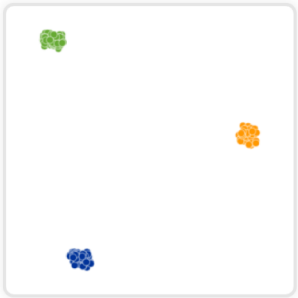
Distance is not always preserved

Original



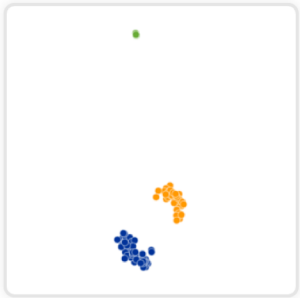
Perplexity: 2
Step: 5,000

Perplexity: 5
Step: 5,000



Perplexity: 30
Step: 5,000

Perplexity: 50
Step: 5,000



Perplexity: 100
Step: 5,000

False clusters may appear

Original

Perplexity: 2
Step: 5,000

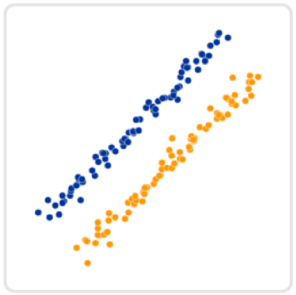
Perplexity: 5
Step: 5,000

Perplexity: 30
Step: 5,000

Perplexity: 50
Step: 5,000

Perplexity: 100
Step: 5,000

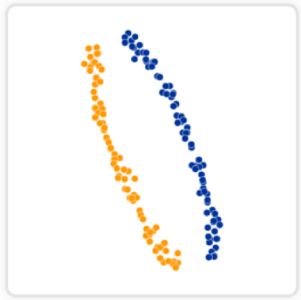
Relationships are not always preserved



Original

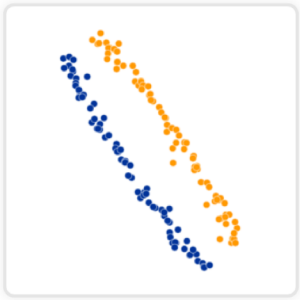
Perplexity: 2
Step: 5,000

Perplexity: 5
Step: 5,000



Perplexity: 30
Step: 5,000

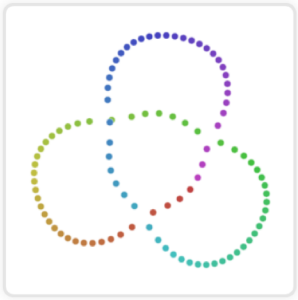
Perplexity: 50
Step: 5,000



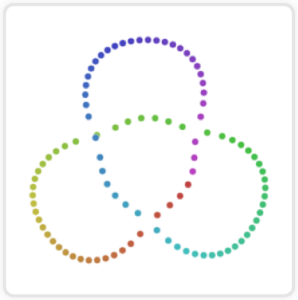
Perplexity: 100
Step: 5,000

Different runs may produce similar results…

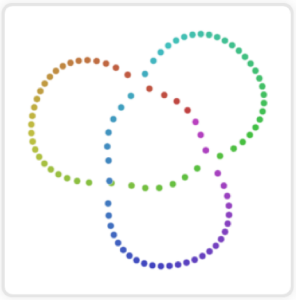
Original



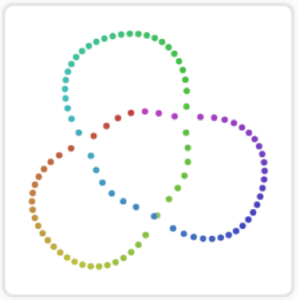
Perplexity: 50
Step: 5,000



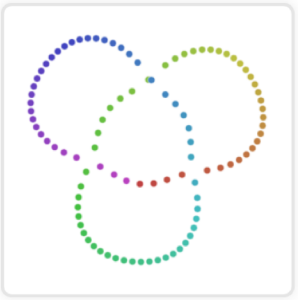
Perplexity: 50
Step: 5,000



Perplexity: 50
Step: 5,000



Perplexity: 50
Step: 5,000



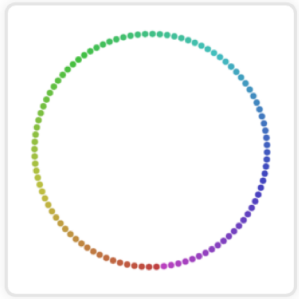
Perplexity: 50
Step: 5,000

Or not…

Original

Perplexity: 2
Step: 5,000

Perplexity: 2
Step: 5,000

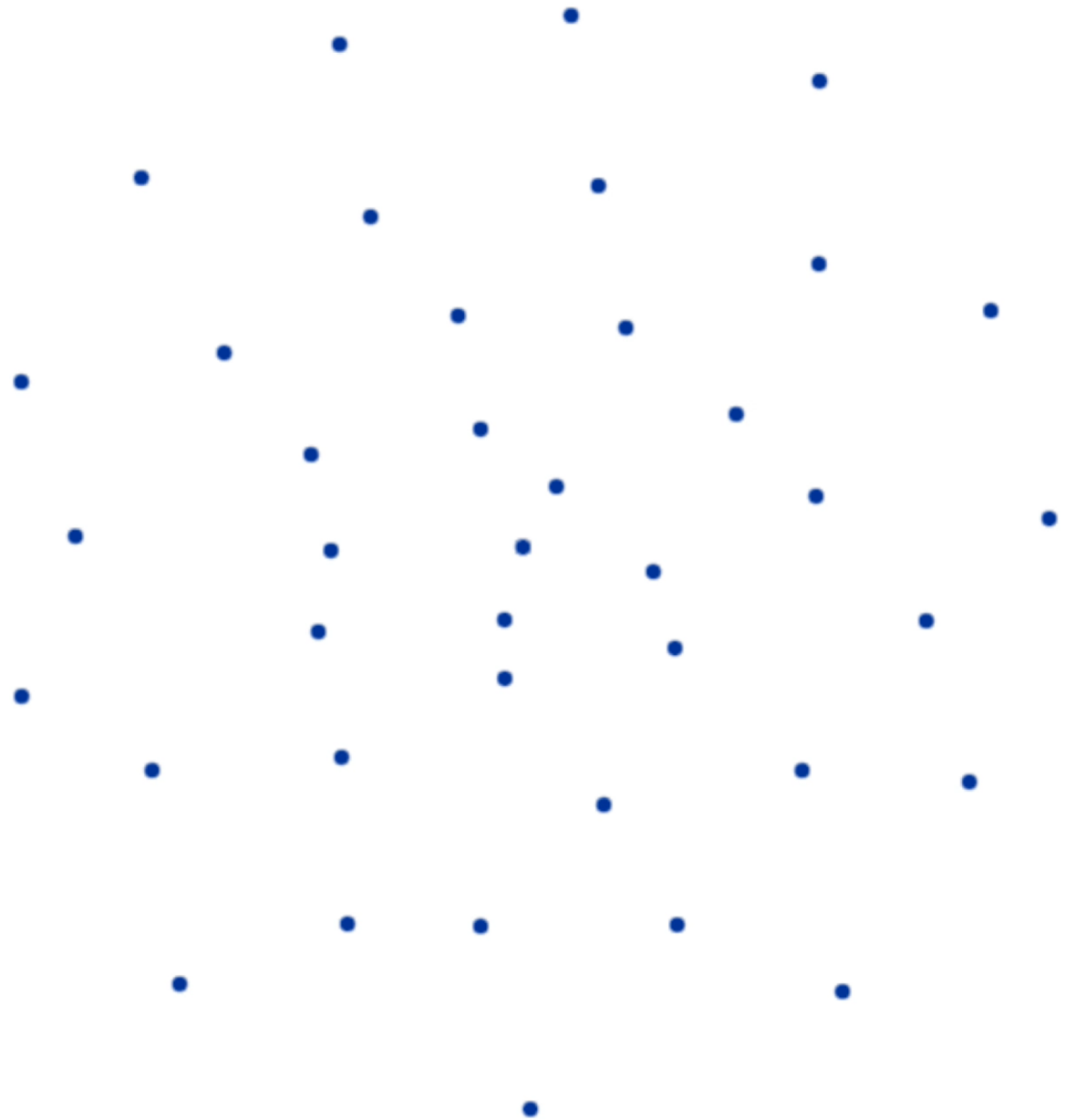


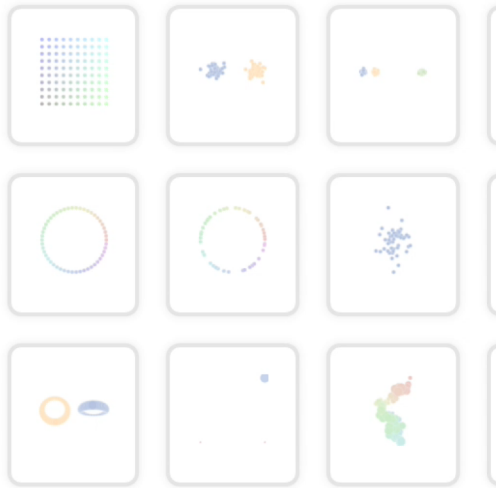
Perplexity: 2
Step: 5,000

Perplexity: 2
Step: 5,000

Perplexity: 2
Step: 5,000

t-SNE of equidistant points



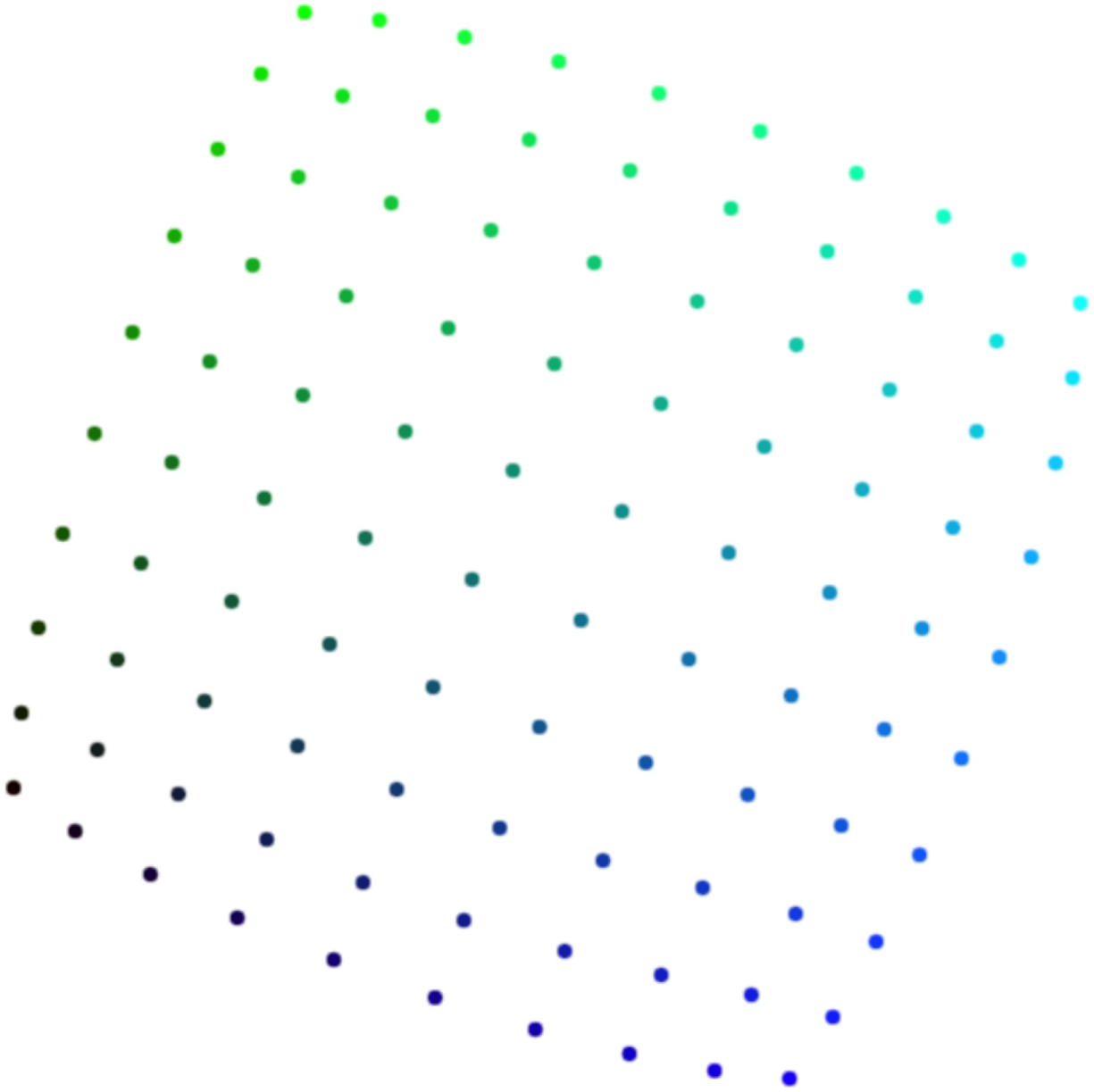


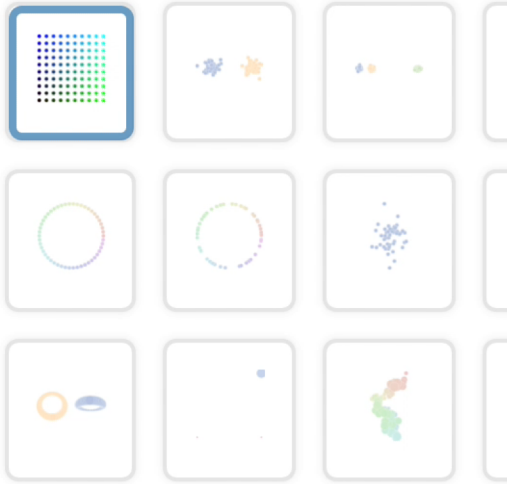
Number Of Points 50

Perplexity 21

Epsilon 5

t-SNE of square grid



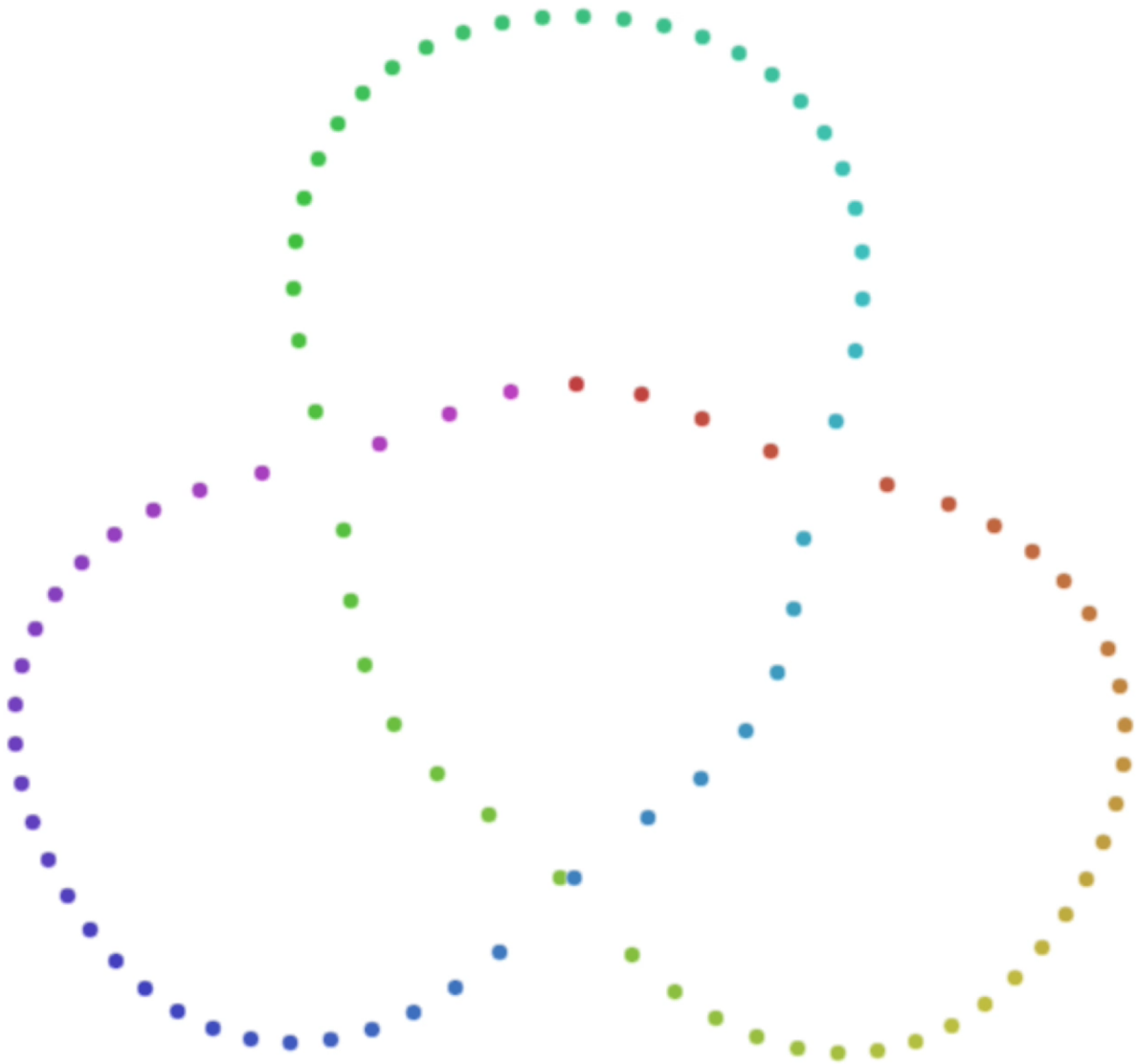


Points Per Side 10

Perplexity 50

Epsilon 5

t-SNE of 3D Knot



Step
2,140

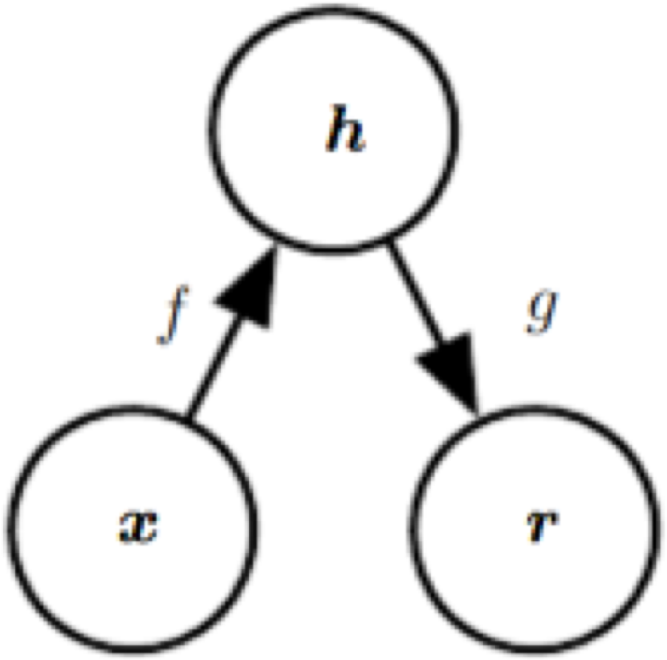
Number Of Points 100

Perplexity 50

Epsilon 5

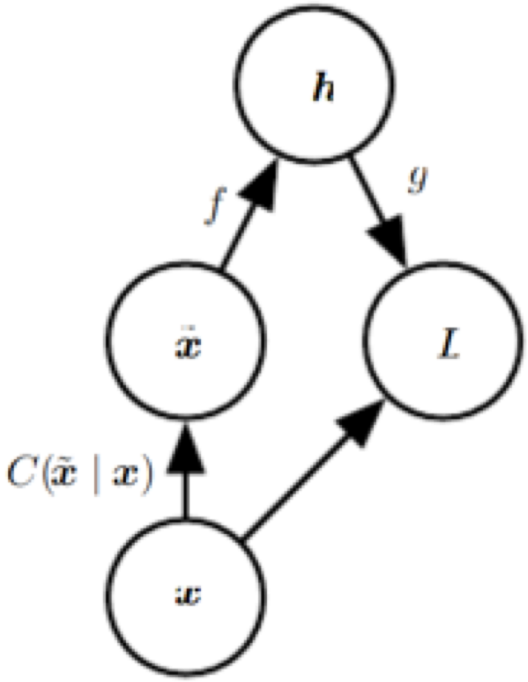
3. Autoencoders embed data into a latent space

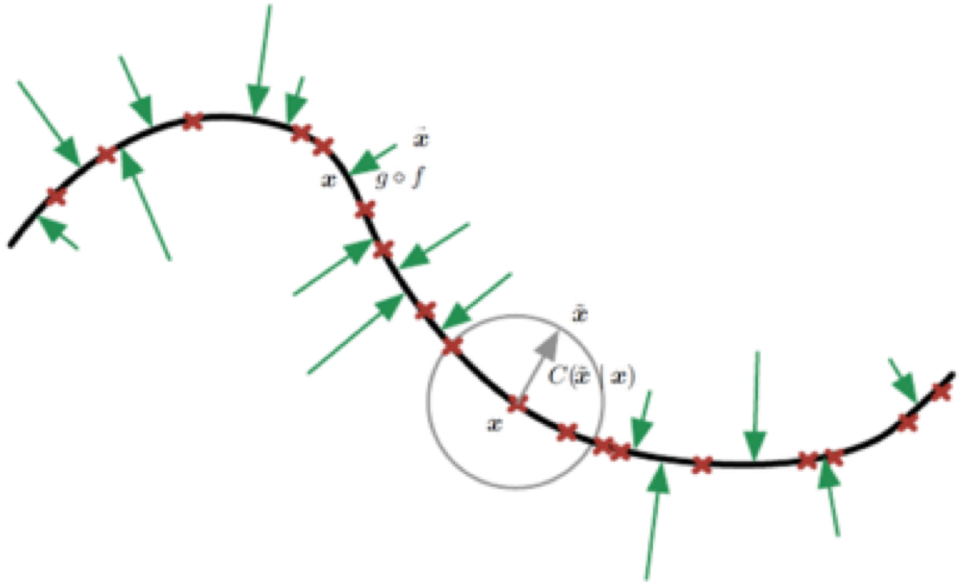
Autoencoders learn a latent representation of data



$$L(\boldsymbol{x}, g(f(\boldsymbol{x})))$$

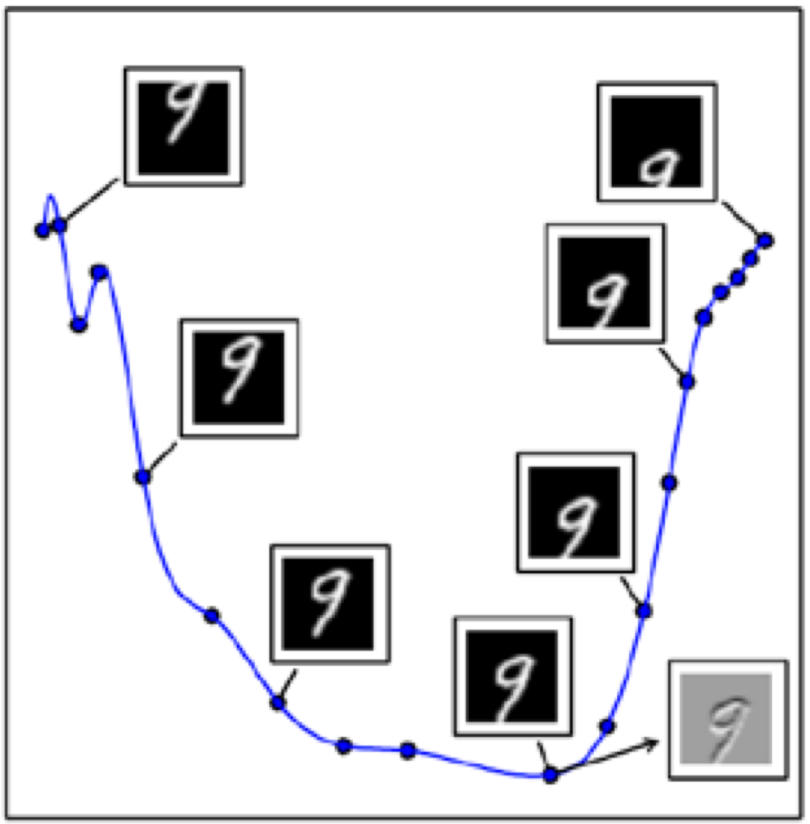
Denoising autoencoders recover signal corrupted by noise

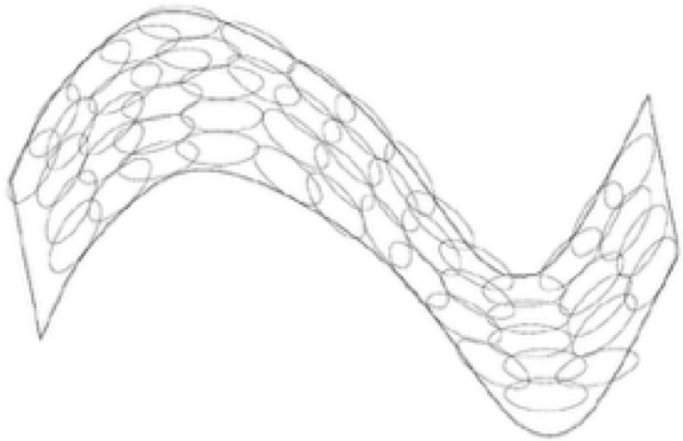




$$L(\boldsymbol{x}, g(f(\tilde{\boldsymbol{x}}))),$$

We can learn manifolds with autoencoders



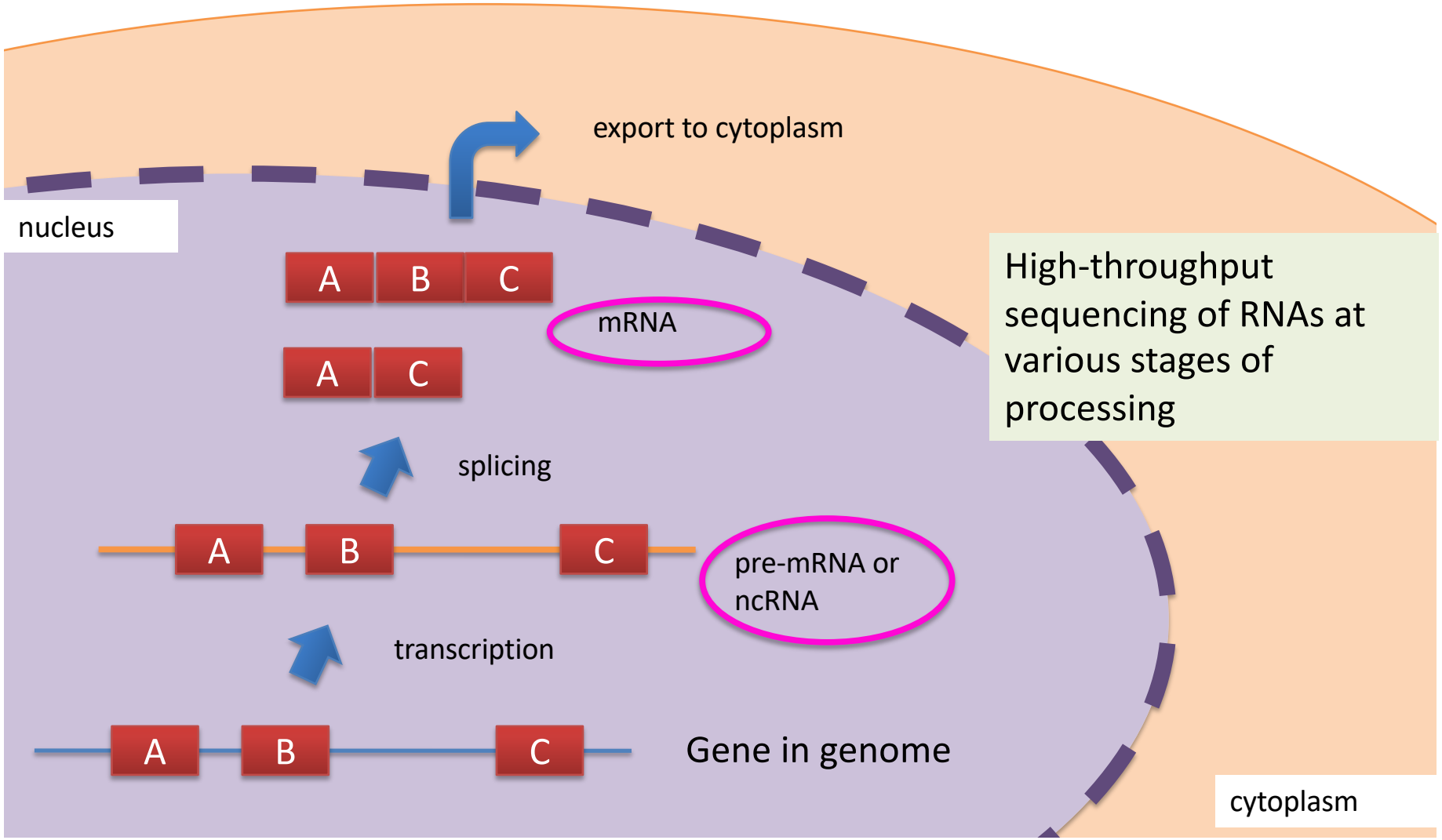


Interesting on-line demos

http://dpkingma.com/sgvb_mnist_demo/demo_old.html
http://elf-project.sourceforge.net/autoencoder.html
http://vdumoulin.github.io/morphing_faces/online_demo.html

4. RNA-seq data has 3,000 – 20,000 gene expression levels per sample

RNA-Seq characterizes RNA molecules

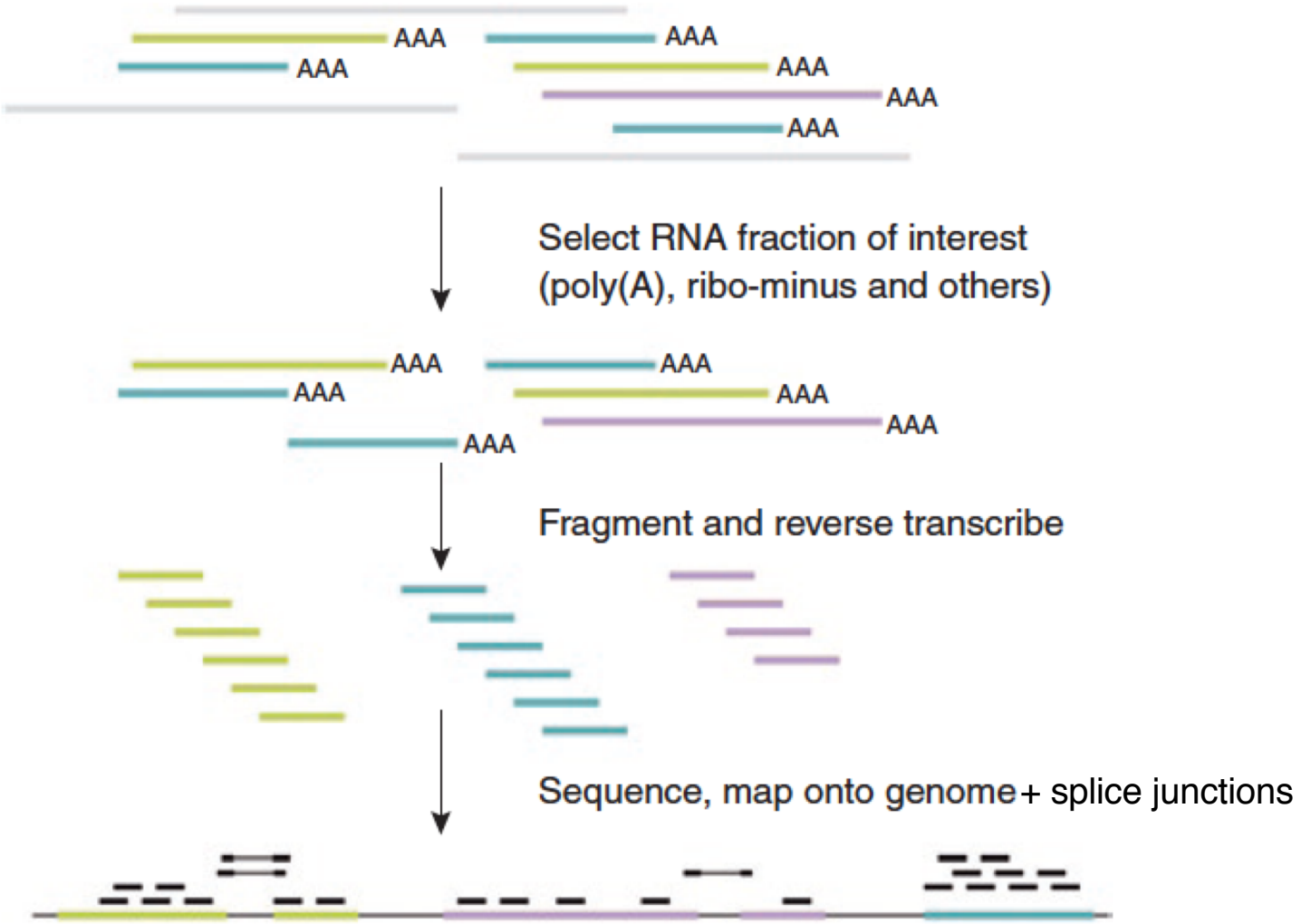


High-throughput sequencing of RNAs at various stages of processing

Slide courtesy Cole Trapnell

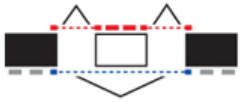
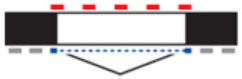
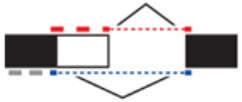
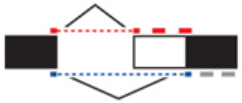
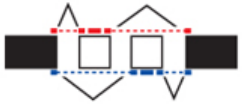
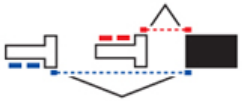
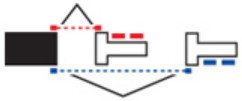
RNA-Seq: millions of short reads from fragmented mRNAs

Extract RNA from cells/tissue



Pepke et. al. *Nature Methods* 2009

Pervasive tissue-specific regulation of alternative mRNA isoforms.

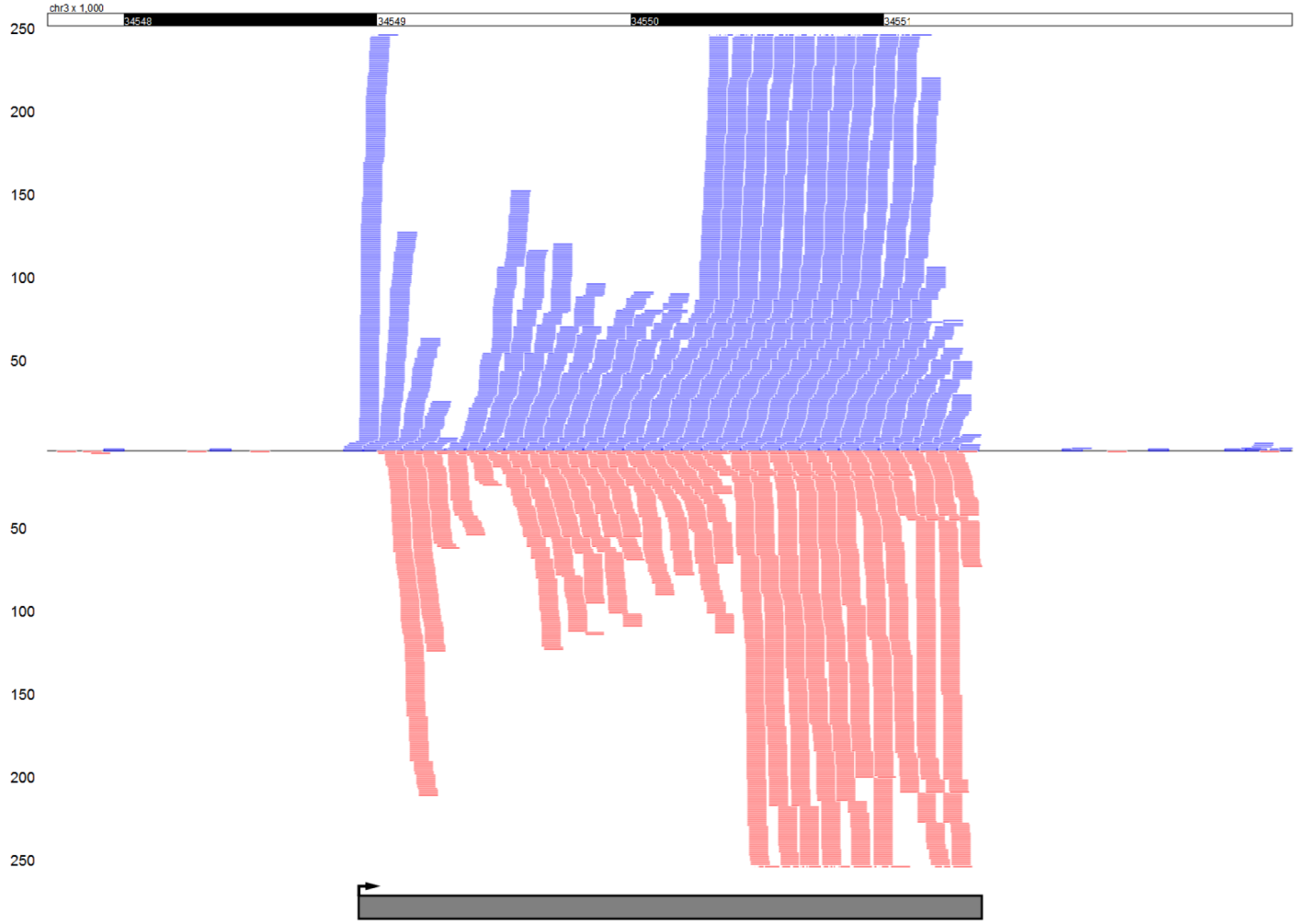
Alternative transcript events		Total events ($\times 10^3$)	Number detected ($\times 10^3$)	Both isoforms detected	Number tissue-regulated	% Tissue-regulated (observed)	% Tissue-regulated (estimated)
Skipped exon		37	35	10,436	6,822	65	72
Retained intron		1	1	167	96	57	71
Alternative 5′ splice site (A5SS)		15	15	2,168	1,386	64	72
Alternative 3′ splice site (A3SS)		17	16	4,181	2,655	64	74
Mutually exclusive exon (MXE)		4	4	167	95	57	66
Alternative first exon (AFE)		14	13	10,281	5,311	52	63
Alternative last exon (ALE)		9	8	5,246	2,491	47	52
Tandem 3′ UTRs		7	7	5,136	3,801	74	80
Total		105	100	37,782	22,657	60	68

Constitutive exon or region — Body read — Junction read — pA Polyadenylation site

Alternative exon or extension — Inclusive/extended isoform — Exclusive isoform — Both isoforms

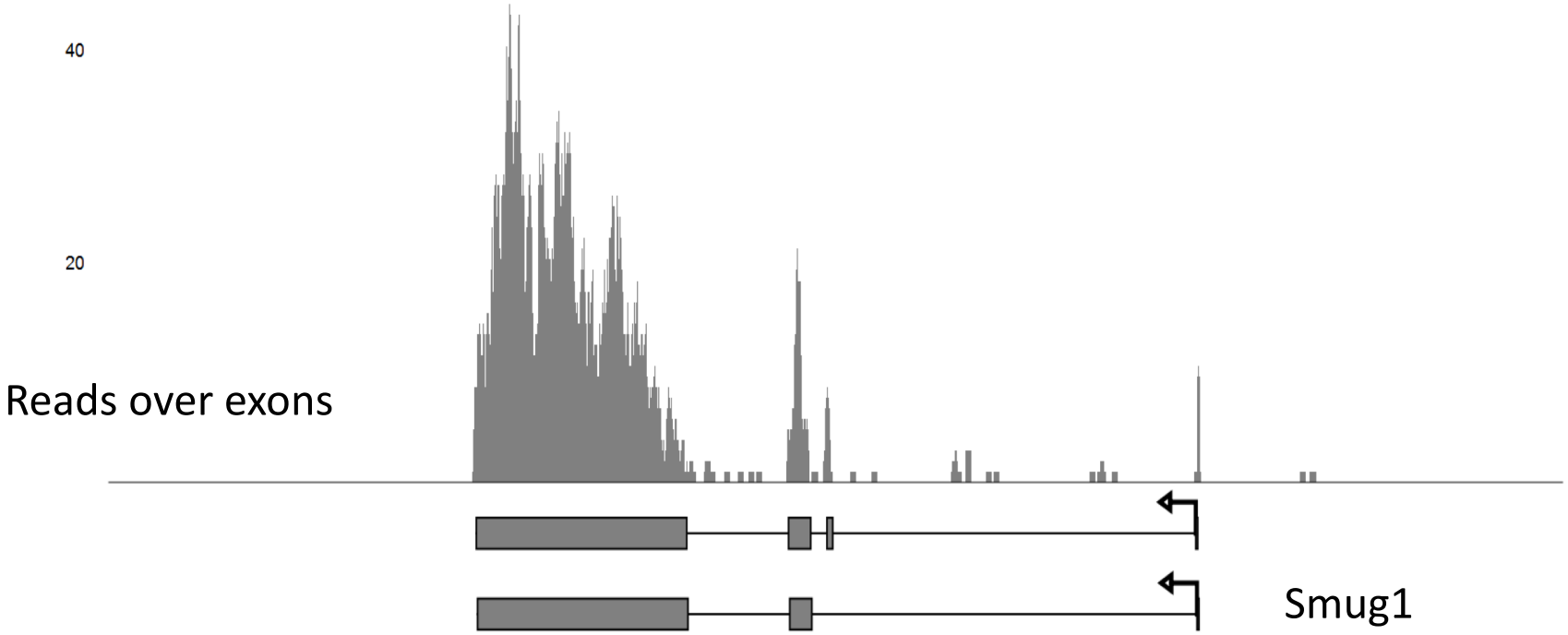
ET Wang *et al. Nature* **000**, 1-7 (2008)
doi:10.1038/nature07509

As crude measure of expression is RPKM – reads per thousand

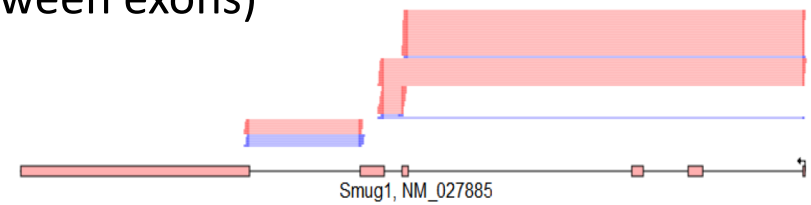


Sox2

RNA-seq reads map to exons and across exons



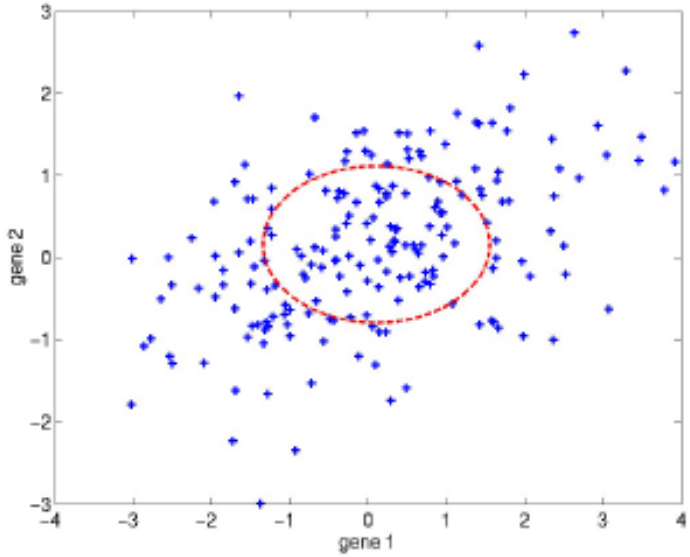
Junction reads (split between exons)



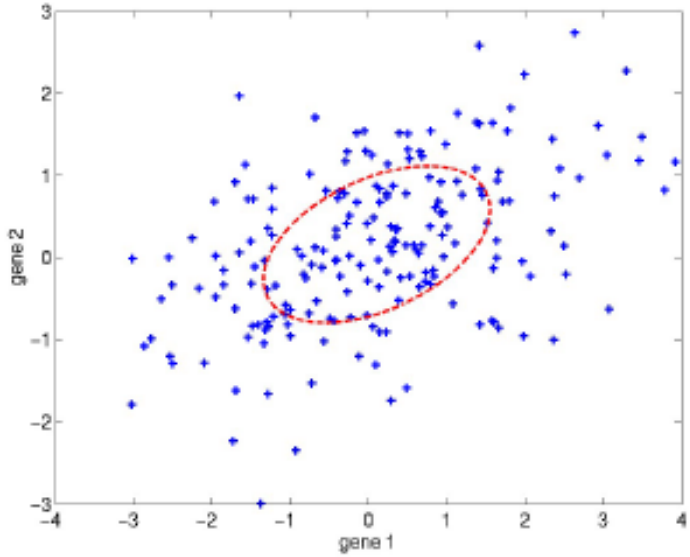
5. The significance of differential expression

Statistical tests: example

- The alternative hypothesis H_1 is more expressive in terms of explaining the observed data



null hypothesis



alternative hypothesis

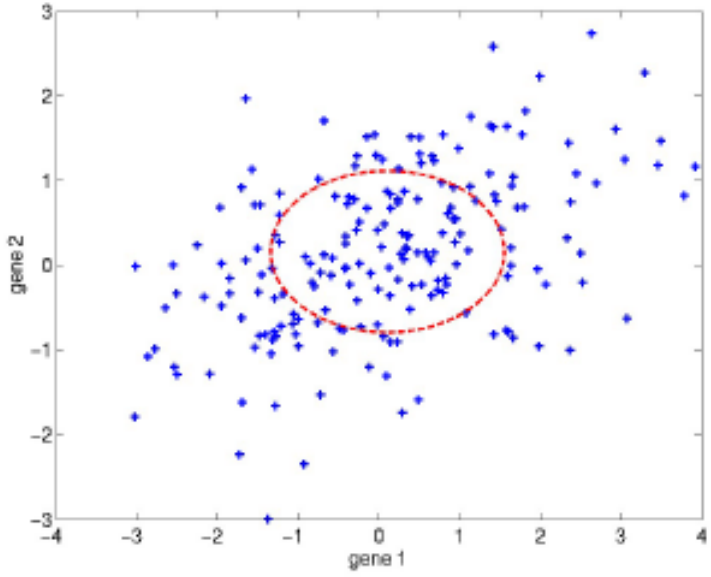
- We need to find a way of testing whether this difference is significant

Degrees of freedom

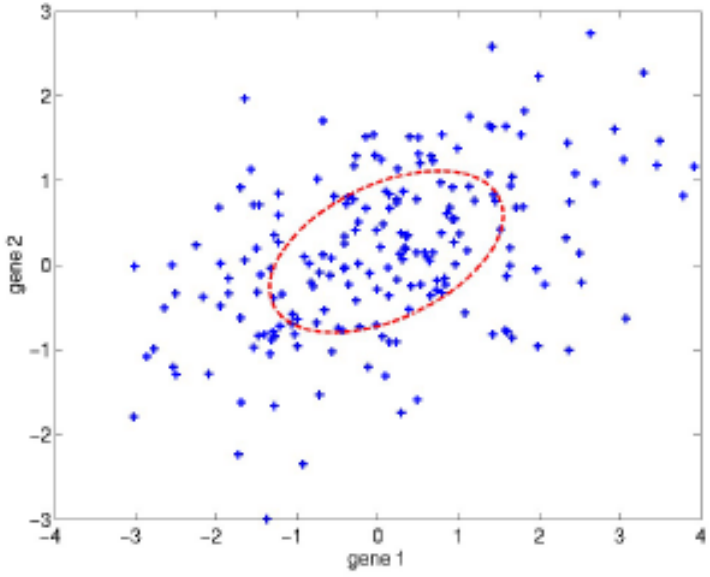
- How many degrees of freedom do we have in the two models?

$$H_0: \quad \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \right)$$

$$H_1: \quad \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$



H_0



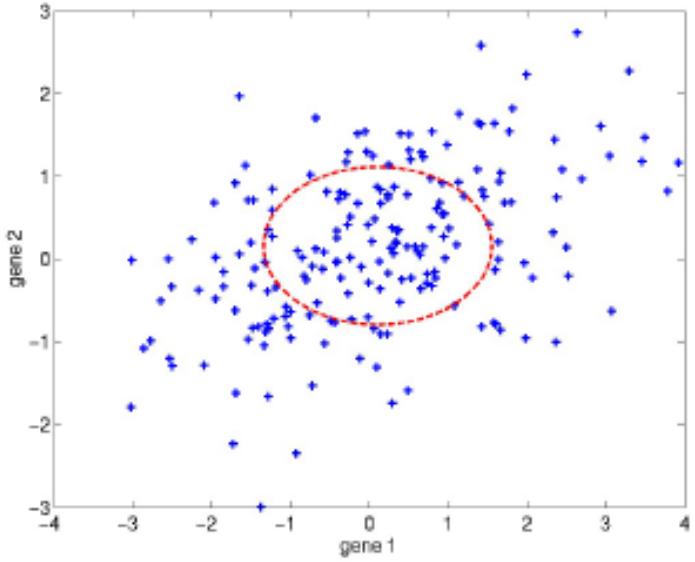
H_1

Degrees of freedom

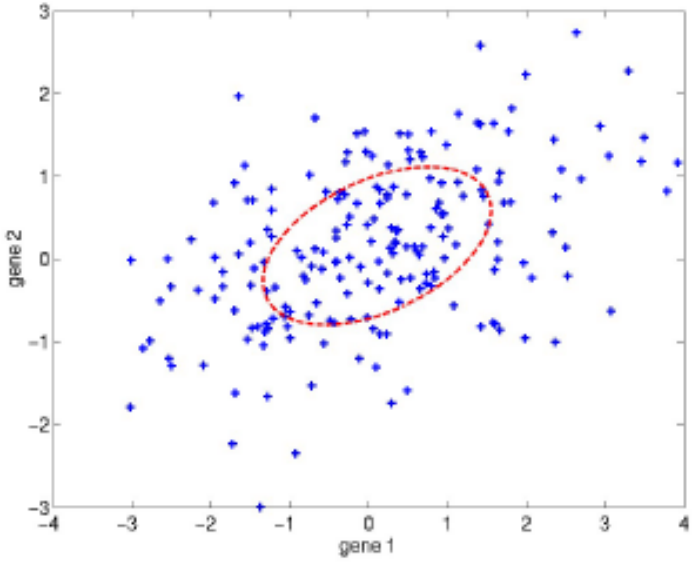
- How many degrees of freedom do we have in the two models?

$$H_0: \quad \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \right)$$

$$H_1: \quad \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$



H_0



H_1

- The observed data overwhelmingly supports H_1

Test statistic

- Likelihood ratio statistic

$$T(X^{(1)}, \ldots, X^{(n)}) = 2 \log \frac{P(X^{(1)}, \ldots, X^{(n)} | \hat{H}_1)}{P(X^{(1)}, \ldots, X^{(n)} | \hat{H}_0)} \qquad (1)$$

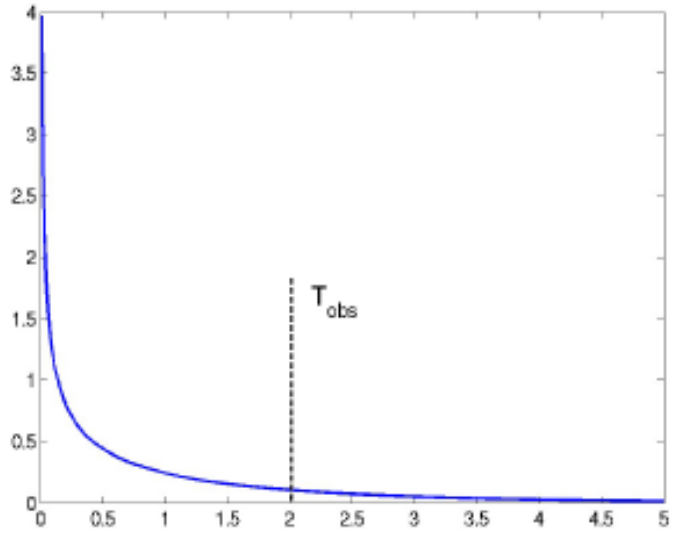
  Larger values of T imply that the model corresponding to the null hypothesis H_0 is much less able to account for the observed data

- To evaluate the P-value, we also need to know the sampling distribution for the test statistic

  In other words, we need to know how the test statistic $T(X^{(1)}, \ldots, X^{(n)})$ varies if the null hypothesis H_0 is correct

Test statistic cont'd

- For the likelihood ratio statistic, the sampling distribution is χ^2 with degrees of freedom equal to the difference in the number of free parameters in the two hypotheses



- Once we know the sampling distribution, we can compute the P-value

$$p = Prob(\, T(X^{(1)}, \ldots, X^{(n)}) \geq T_{obs} \,|\, H_0 \,) \tag{2}$$

FIN - Thank You