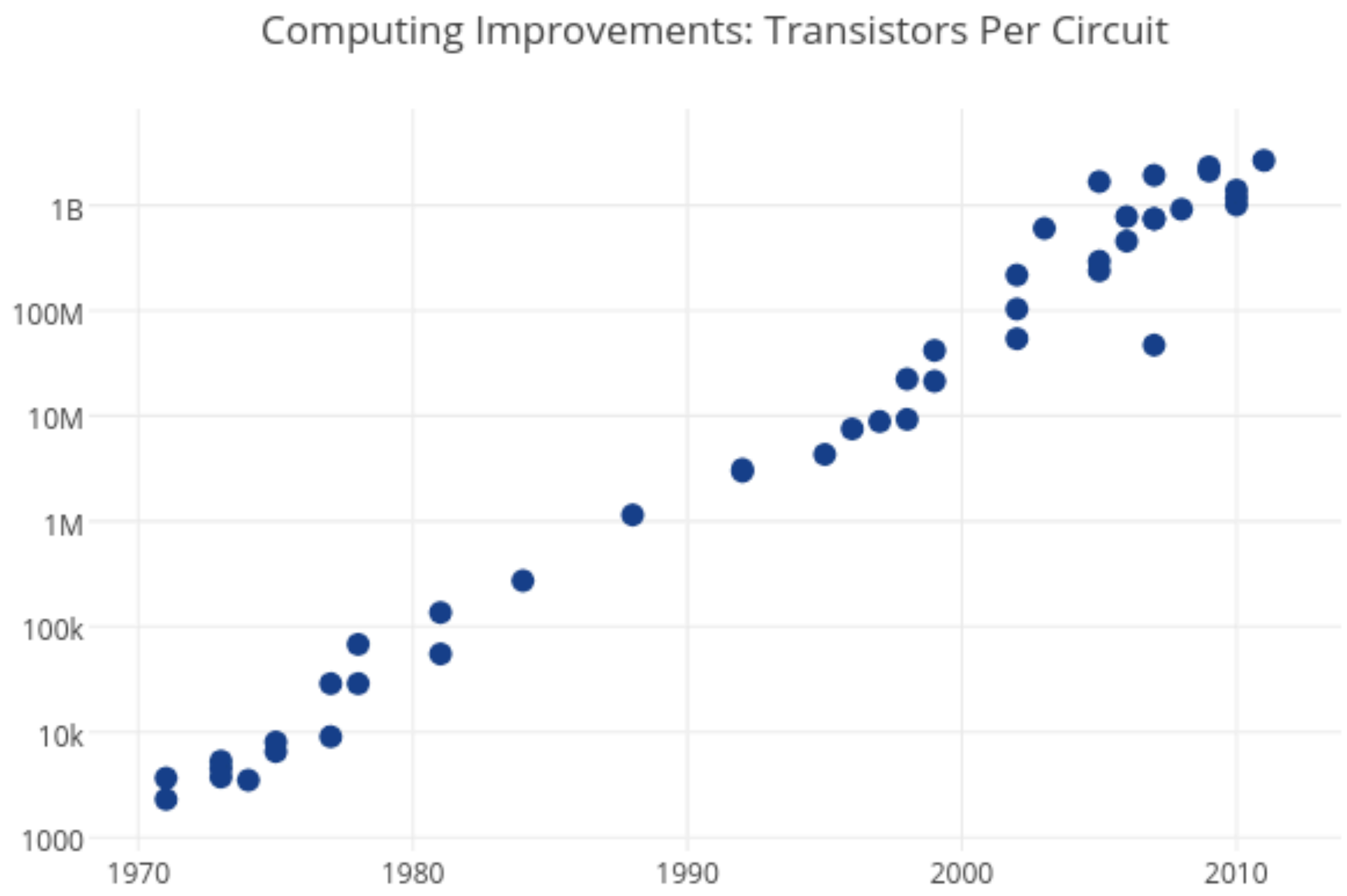
Carnegie Mellon University
Heinz College

95-865 Australia Lecture 2: Finding Possibly Related Entities, Visualizing High-Dimensional Vectors, Dimensionality Reduction

George Chen

Continuous Measurements

- So far, looked at relationships between *discrete* outcomes
- For pair of *continuous* outcomes, use a **scatter plot**

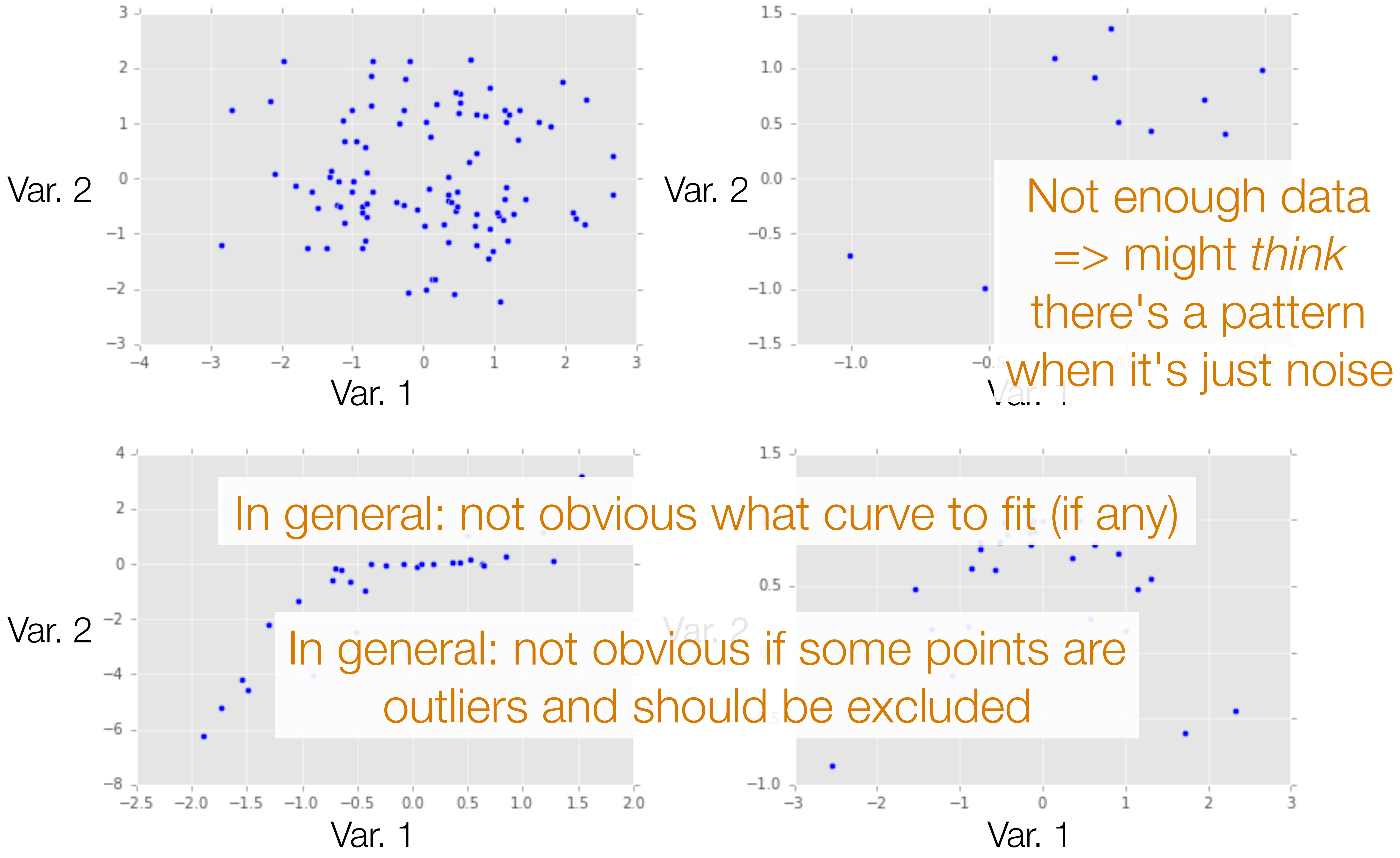


Of course, not all trends look like a line

(so don't just do linear regression!)

Image source: https://plot.ly/~MattSundquist/5405.png

The Importance of Staring at Data

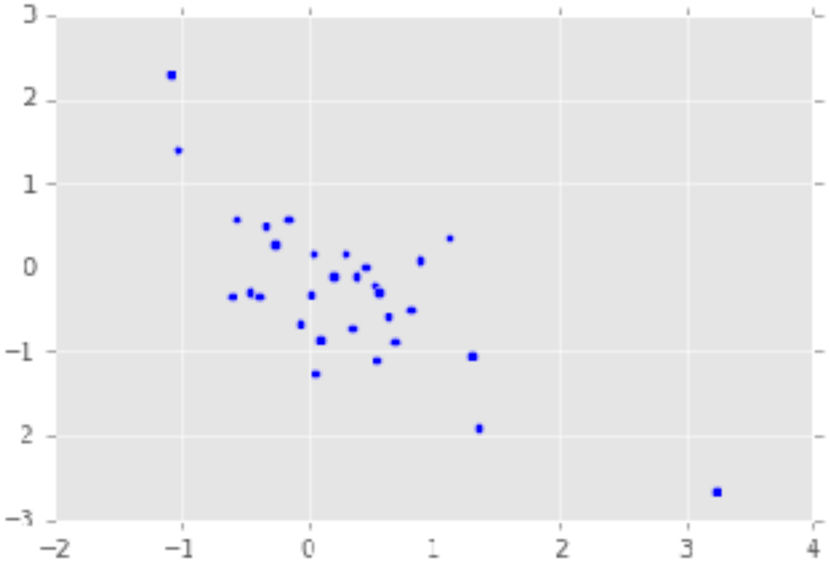


Not enough data => might *think* there's a pattern when it's just noise

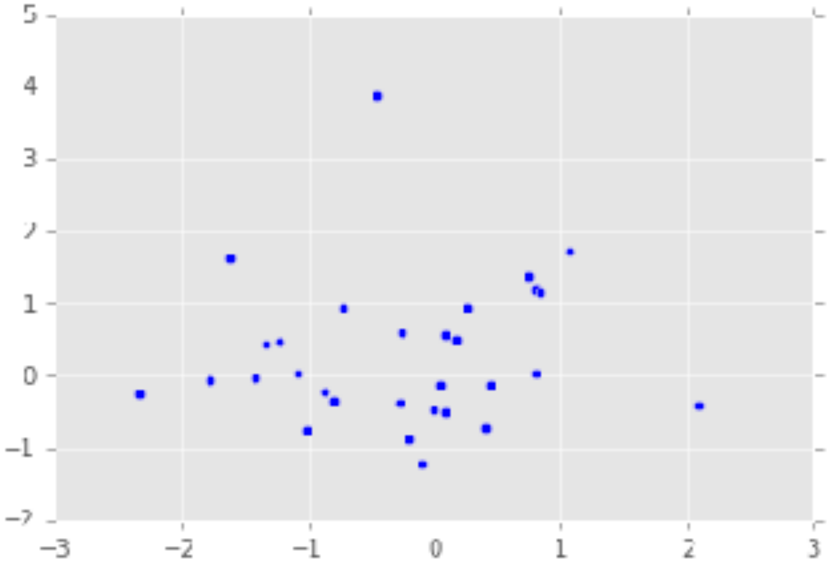
In general: not obvious what curve to fit (if any)

In general: not obvious if some points are outliers and should be excluded

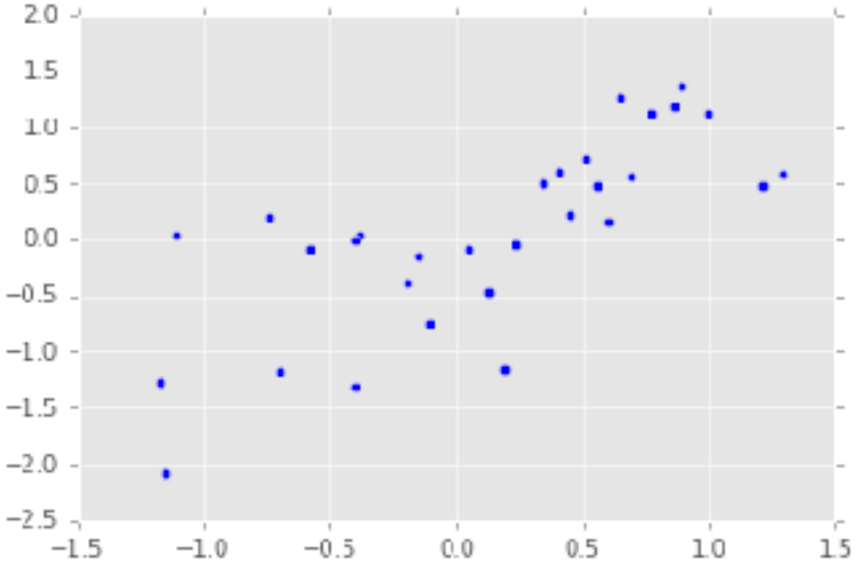
Correlation



Negatively correlated



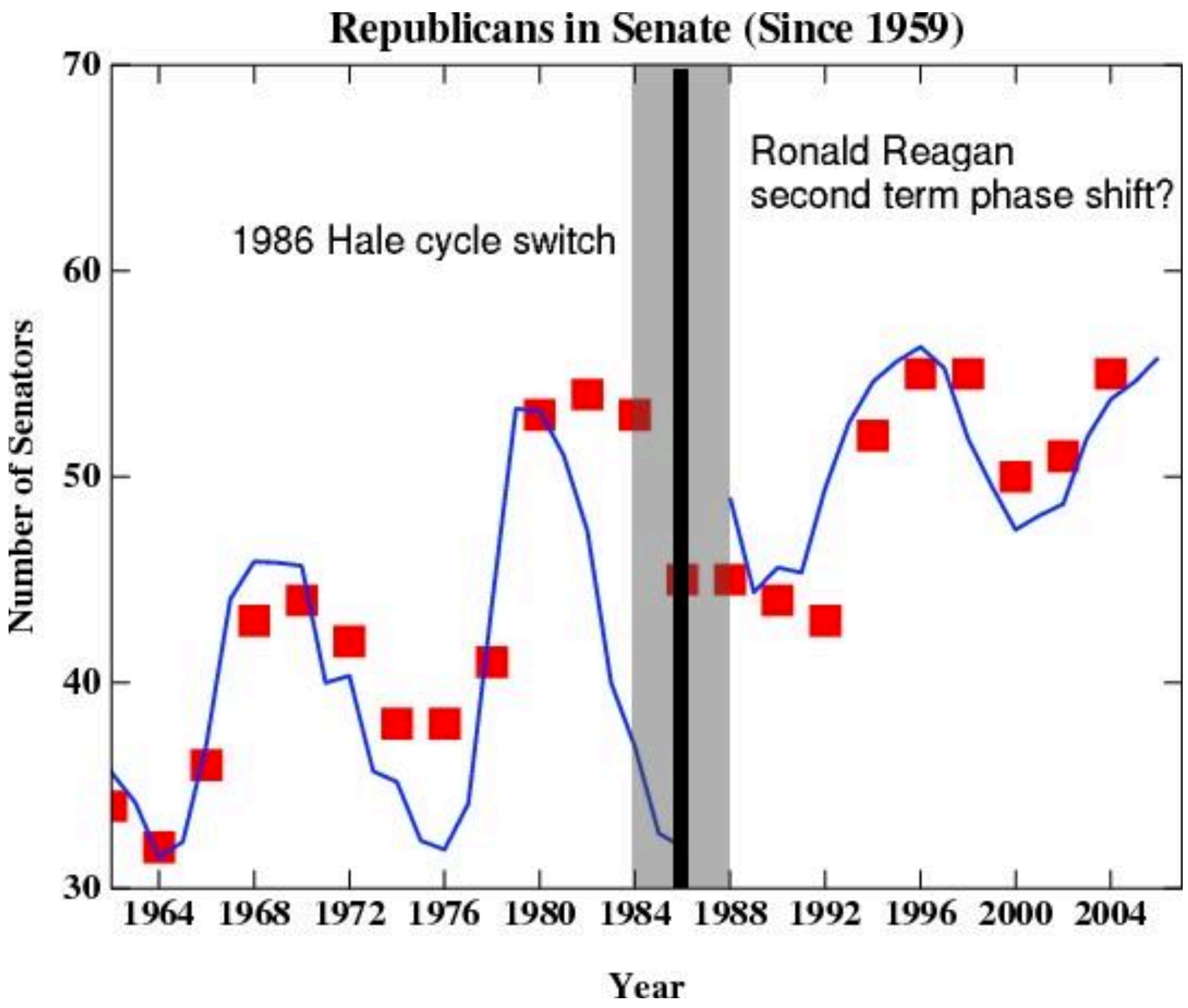
Not really correlated



Positively correlated

Beware: Just because two variables appear correlated doesn't mean that one can predict the other

Correlation ≠ Causation



Blue: Scaled sunspot number (inverted after Reagan's 2nd term)

Red: Number of Republican senators

Moreover, just because we find correlation in data doesn't mean it has predictive value!

Image source: http://www.realclimate.org/index.php/archives/2007/05/fun-with-correlations/

Important: At this point in the course, we are finding *possible* relationships between two entities

We are *not* yet making statements about prediction (we'll see prediction later in the course)

We are *not* making statements about causality (beyond the scope of this course)

Causality

Studies in 1960's: Coffee drinkers have higher rates of lung cancer

Can we claim that coffee is a cause of lung cancer?

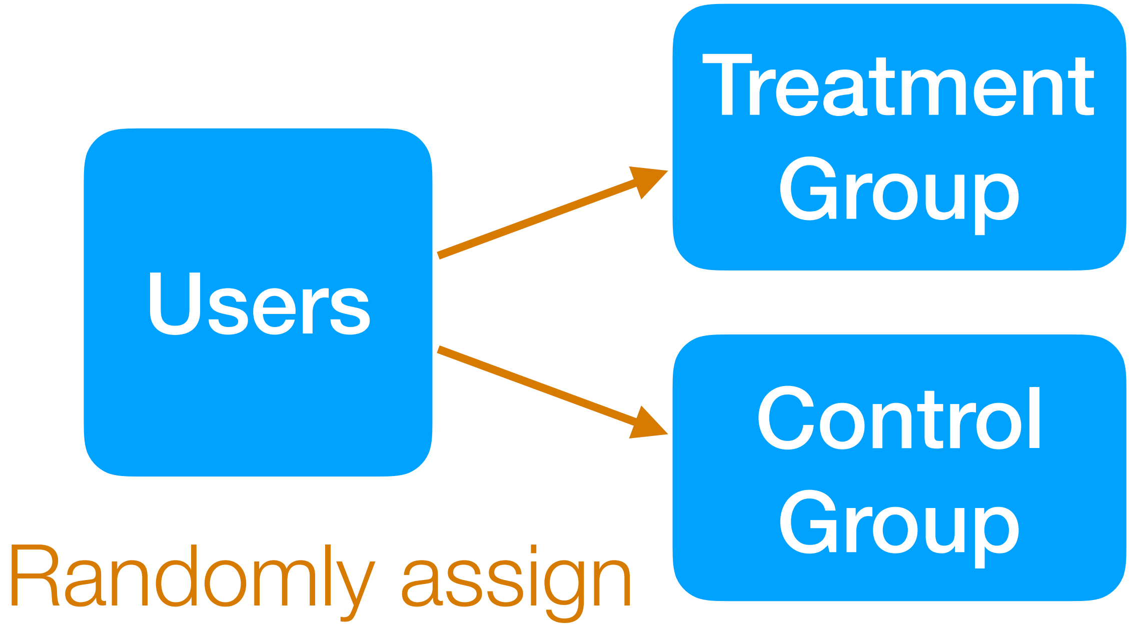
Back then: coffee drinkers also tended to smoke more than non-coffee drinkers (smoking is a **confounding variable**)

To establish causality, groups getting different treatments need to appear similar so that the only difference is the treatment

Image source: George Chen

Establishing Causality

If you control data collection



Compare outcomes of two groups

Randomized controlled trial (RCT)
also called **A/B testing**

Example: figure out webpage layout to maximize revenue (Amazon)

Example: figure out how to present educational material to improve learning (Khan Academy)

If you do not control data collection

In general: *not* obvious establishing what caused what

95-865

Part I: Exploratory data analysis

Identify structure present in "unstructured" data

- Frequency and co-occurrence analysis *Basic probability & statistics*
- Visualizing high-dimensional data/dimensionality reduction
- Clustering
- Topic modeling (a special kind of clustering)

Part II: Predictive data analysis

Make predictions using structure found in Part I

- Classical classification methods
- Neural nets and deep learning for analyzing images and text

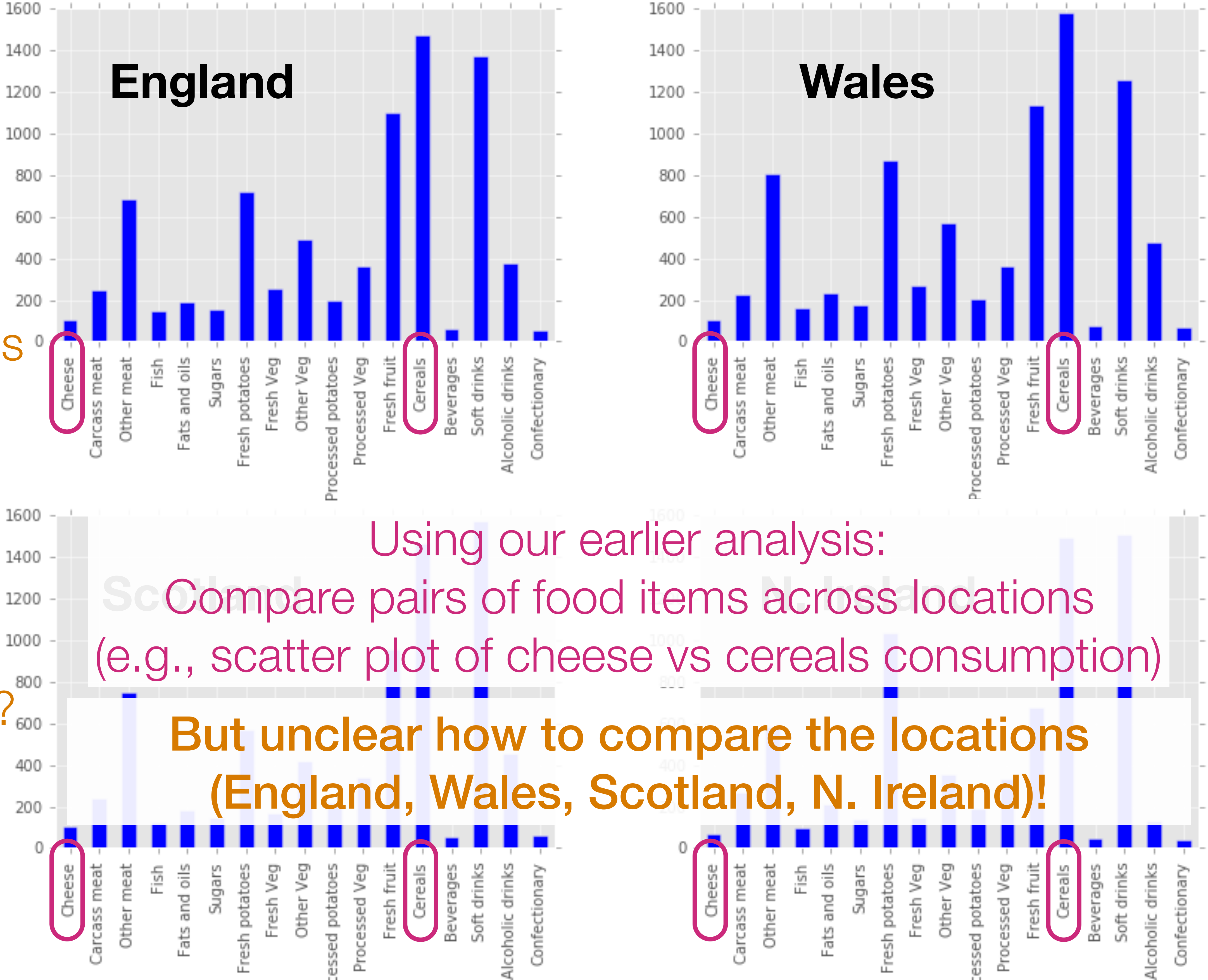
Visualizing High-Dimensional Vectors

The next two examples are drawn from:
http://setosa.io/ev/principal-component-analysis/

Visualizing High-Dimensional Vectors

Imagine we had hundreds of these

How to visualize these for comparison?



Using our earlier analysis:
Compare pairs of food items across locations
(e.g., scatter plot of cheese vs cereals consumption)

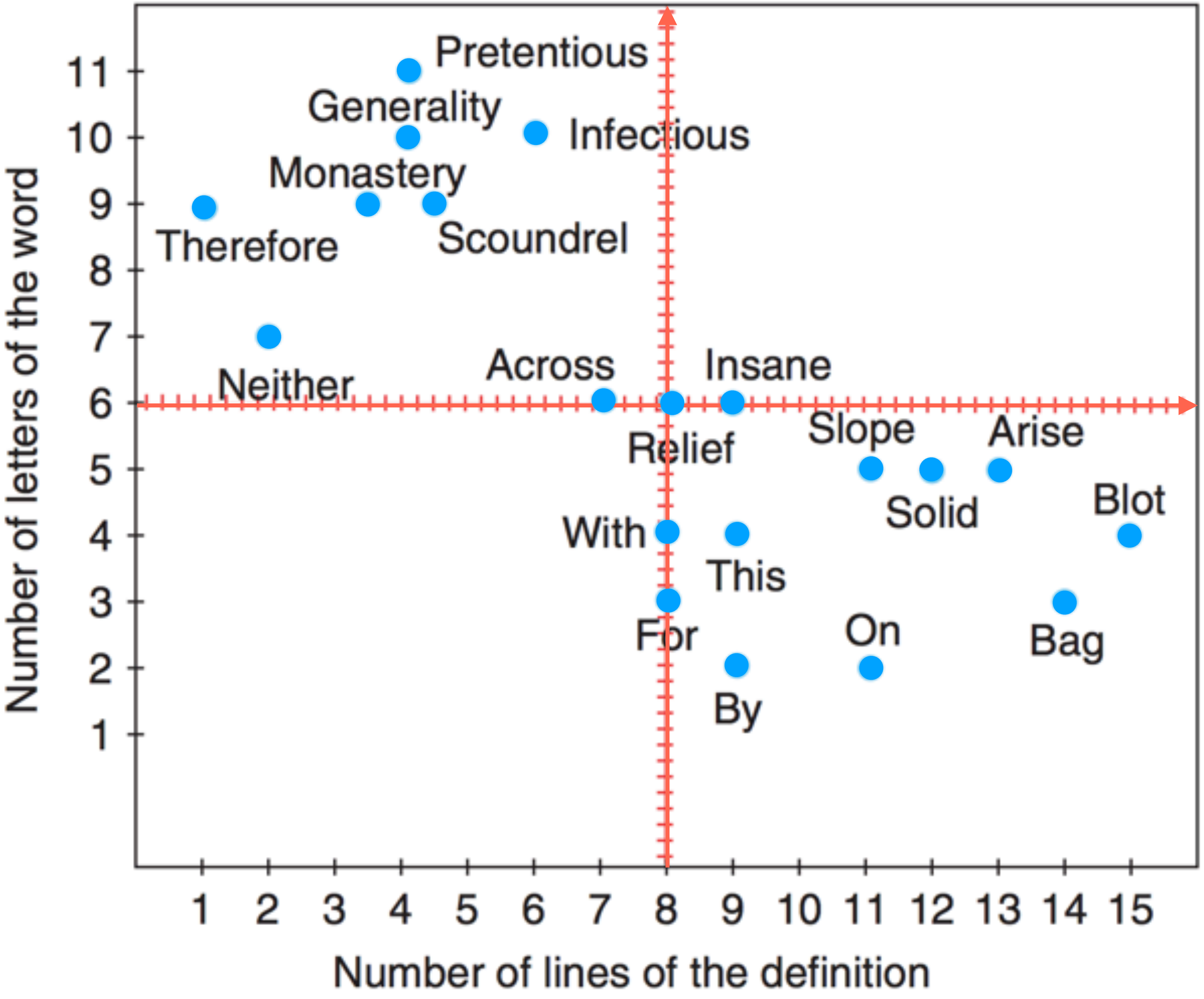
But unclear how to compare the locations (England, Wales, Scotland, N. Ireland)!

The issue is that as humans we can only really visualize up to 3 dimensions easily

Goal: Somehow reduce the dimensionality of the data preferably to 1, 2, or 3

Principal Component Analysis (PCA)

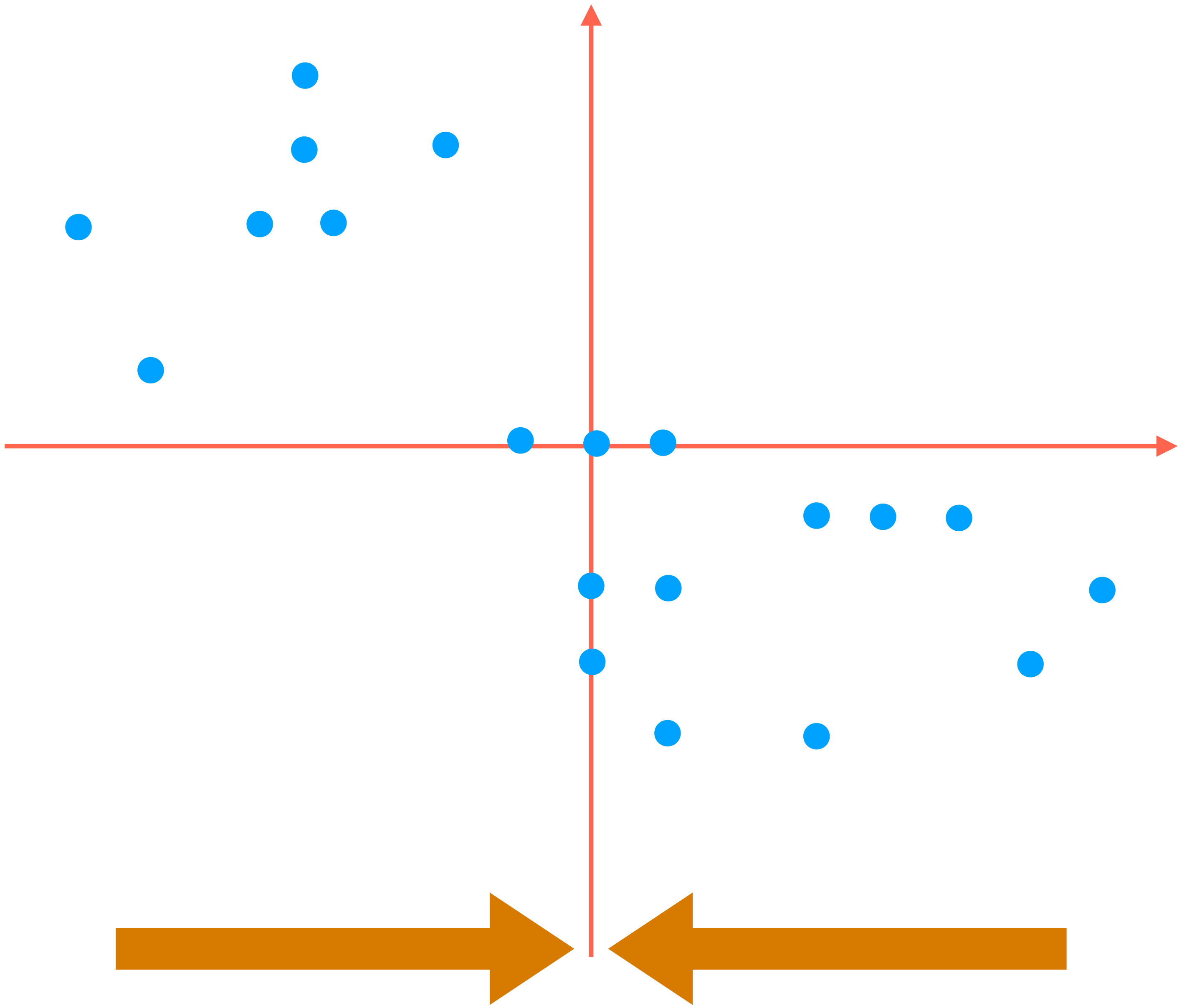
How to project 2D data down to 1D?



Hervé Abdi and Lynne J. Williams. Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics. 2010.

Principal Component Analysis (PCA)

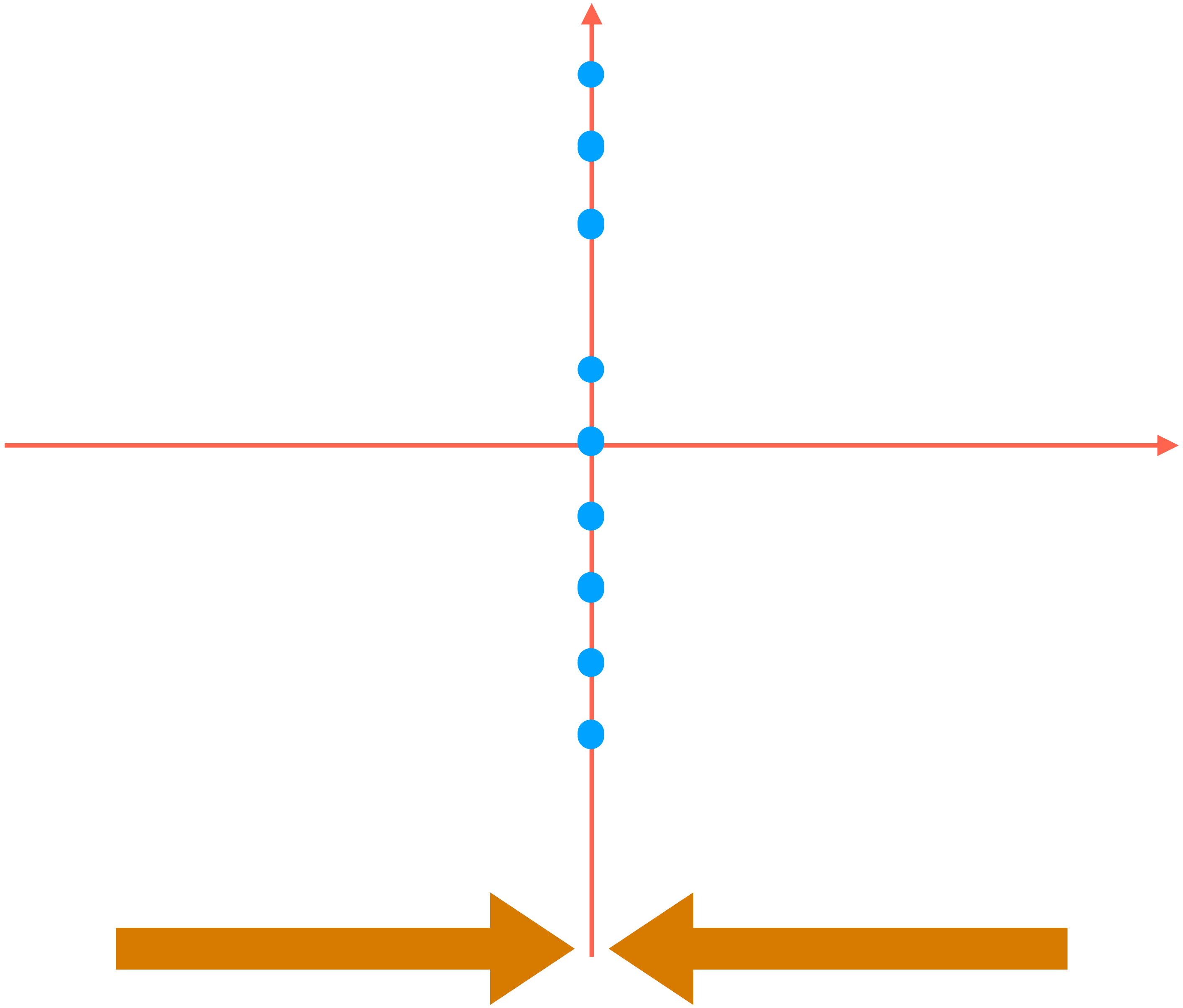
How to project 2D data down to 1D?



Simplest thing to try: flatten to one of the red axes

Principal Component Analysis (PCA)

How to project 2D data down to 1D?

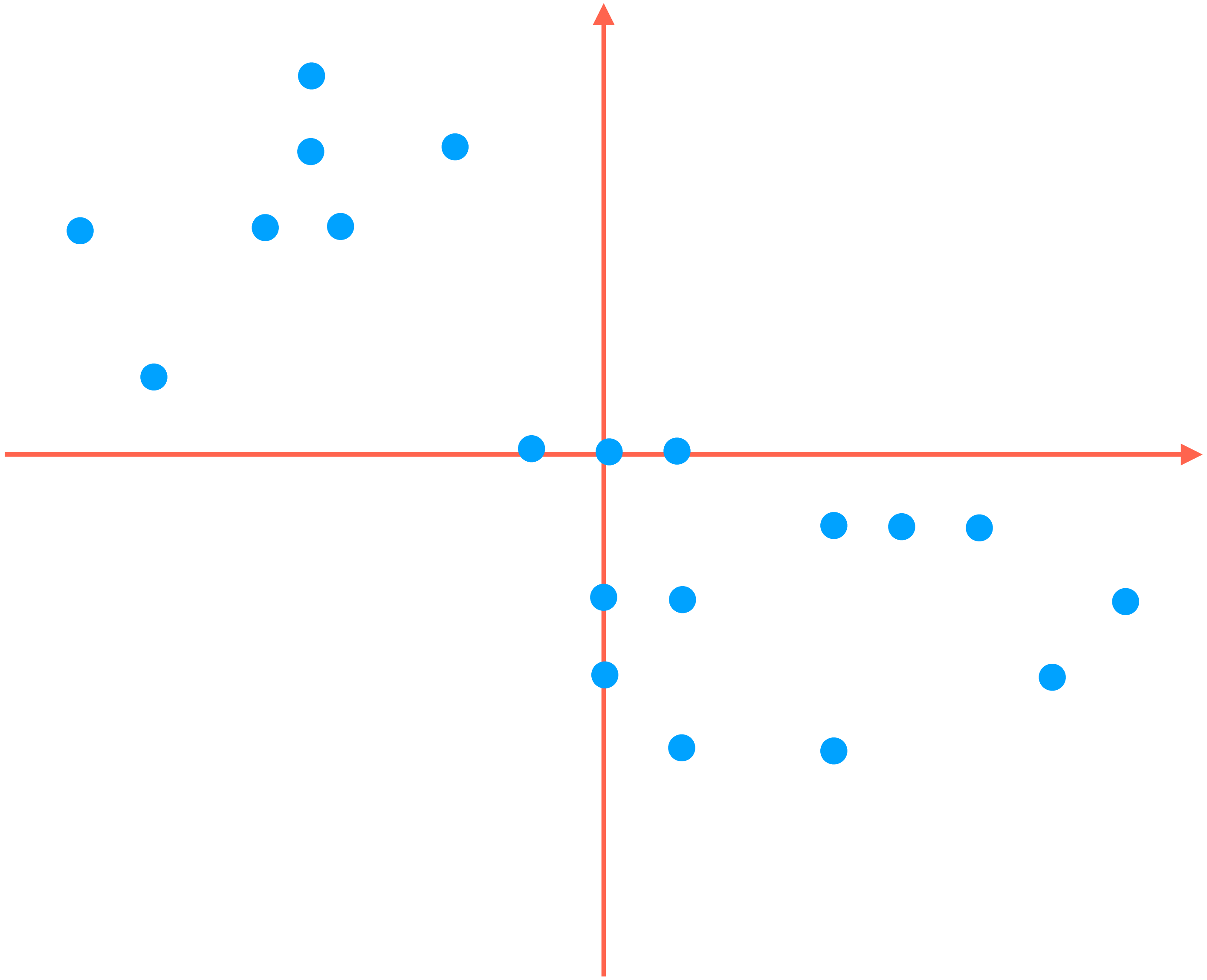


Simplest thing to try: flatten to one of the red axes

(We could of course flatten to the other red axis)

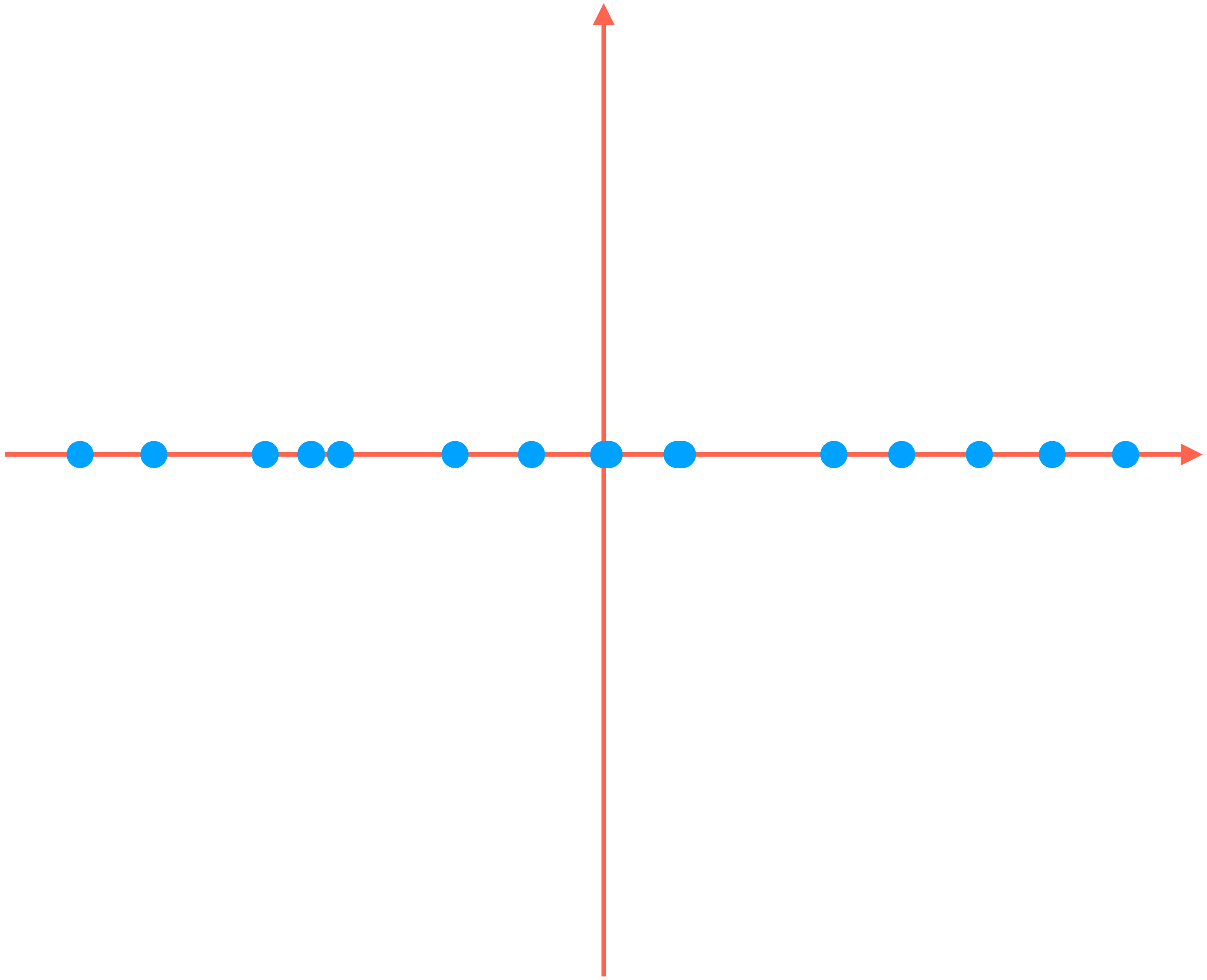
Principal Component Analysis (PCA)

How to project 2D data down to 1D?



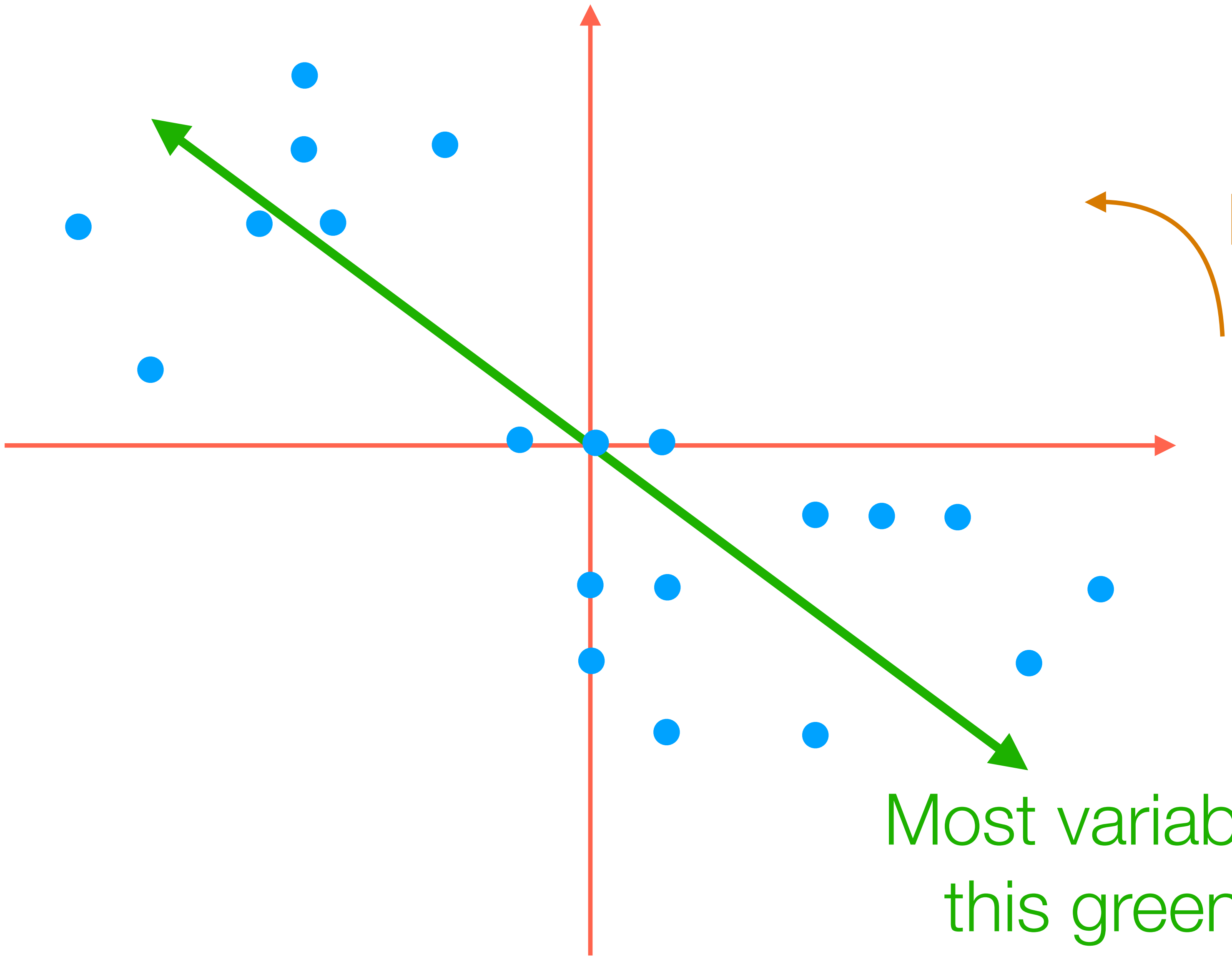
Principal Component Analysis (PCA)

How to project 2D data down to 1D?



Principal Component Analysis (PCA)

How to project 2D data down to 1D?



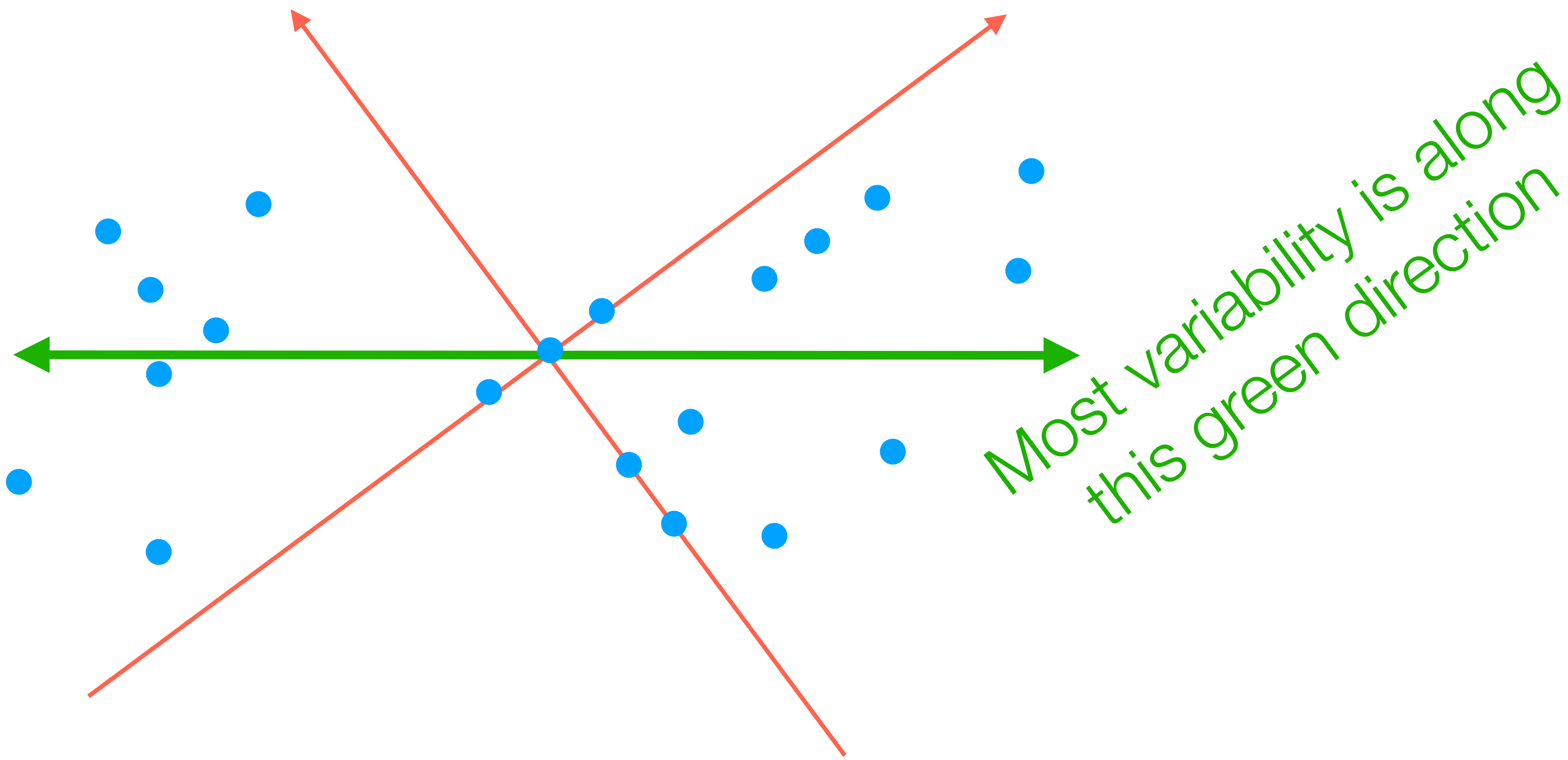
Rotate!

Most variability is along this green direction

But notice that most of the variability in the data is *not* aligned with the red axes!

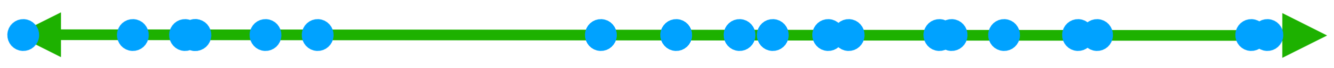
Principal Component Analysis (PCA)

How to project 2D data down to 1D?



Principal Component Analysis (PCA)

How to project 2D data down to 1D?



Most variability is along this green direction

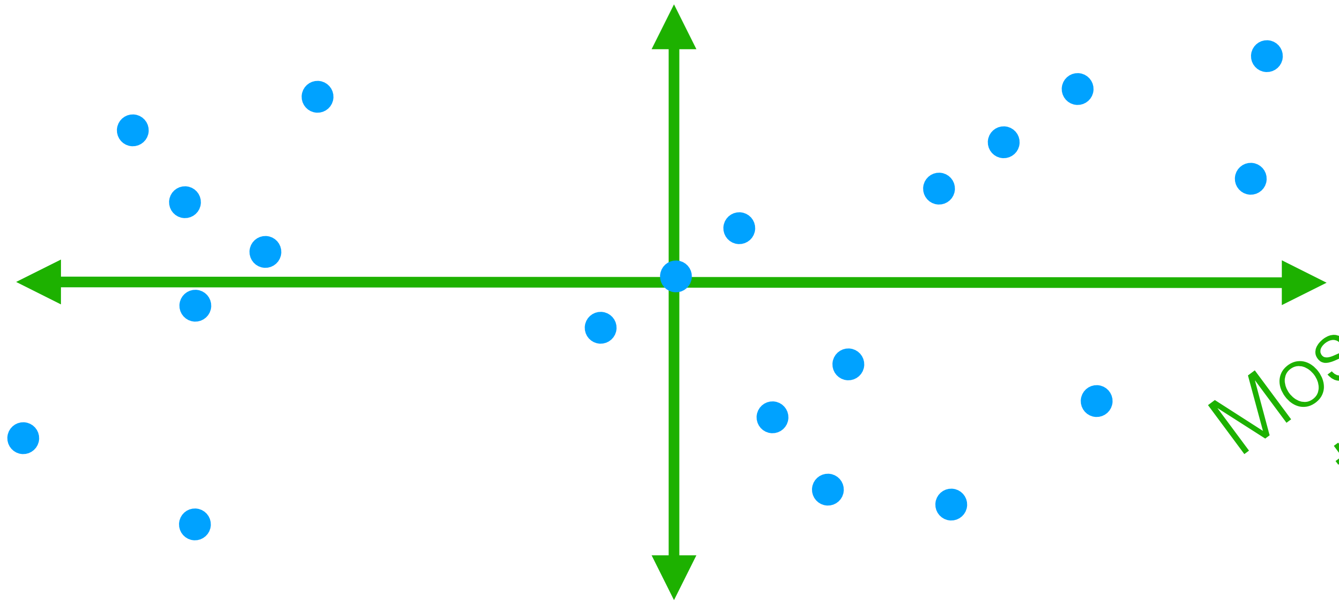
The idea of PCA actually works for 2D ➔ 2D as well (and just involves rotating, and not “flattening” the data)

Principal Component Analysis (PCA)

~~How to project 2D data down to 1D?~~

How to rotate 2D data so 1st axis has most variance

before “flattening”



Most variability is along this green direction

The idea of PCA actually works for 2D ➔ 2D as well
(and just involves rotating, and not “flattening” the data)

2nd green axis chosen to be 90° (“orthogonal”) from first green axis

Principal Component Analysis (PCA)

- Finds top *k* orthogonal directions that explain the most variance in the data
  - 1st component: explains most variance along 1 dimension
  - 2nd component: explains most of remaining variance along next dimension that is orthogonal to 1st dimension
  - …
- “Flatten” data to the top *k* dimensions to get lower dimensional representation (if *k* < original dimension)

Principal Component Analysis (PCA)

3D example from:
http://setosa.io/ev/principal-component-analysis/

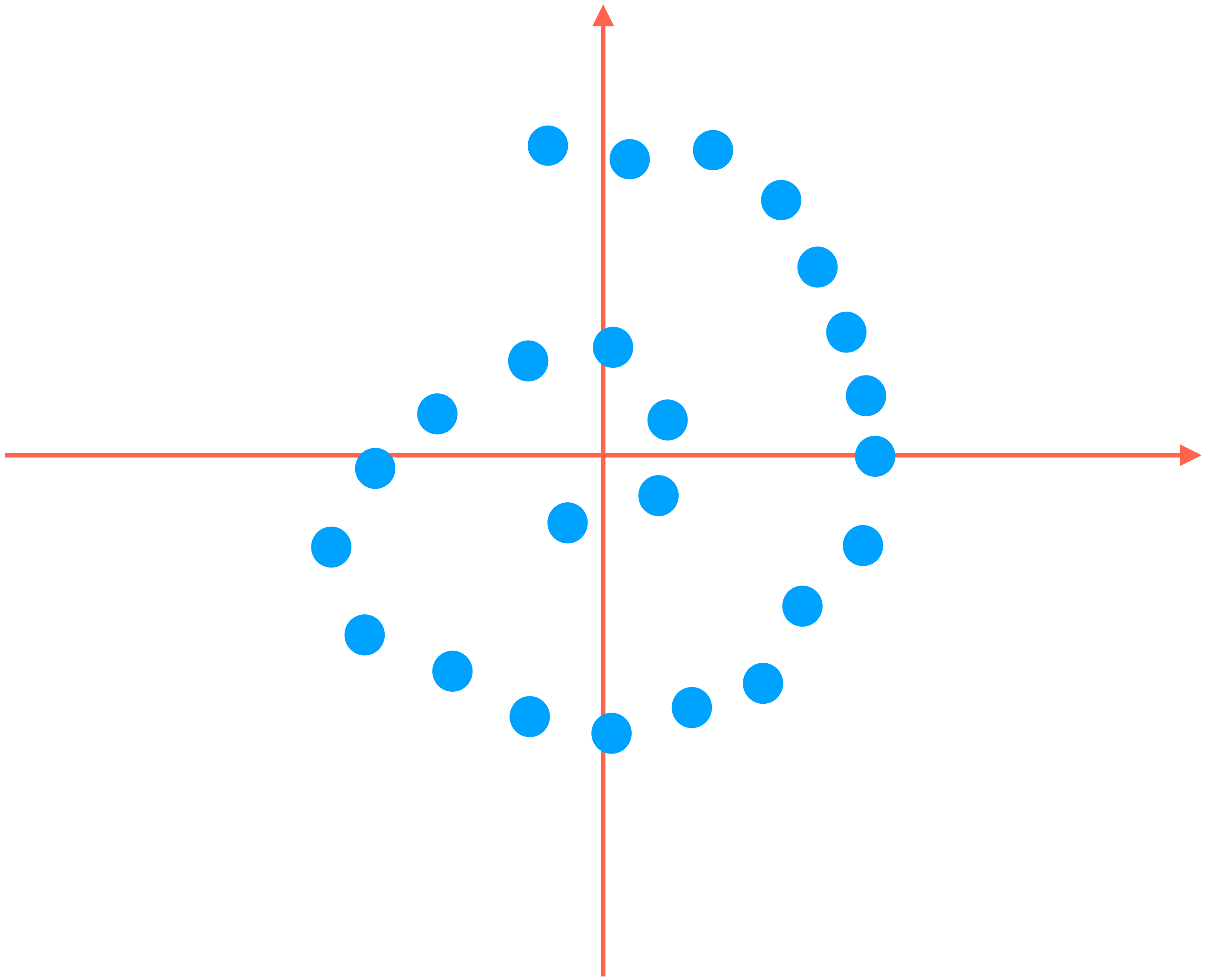
Principal Component Analysis (PCA)

Demo

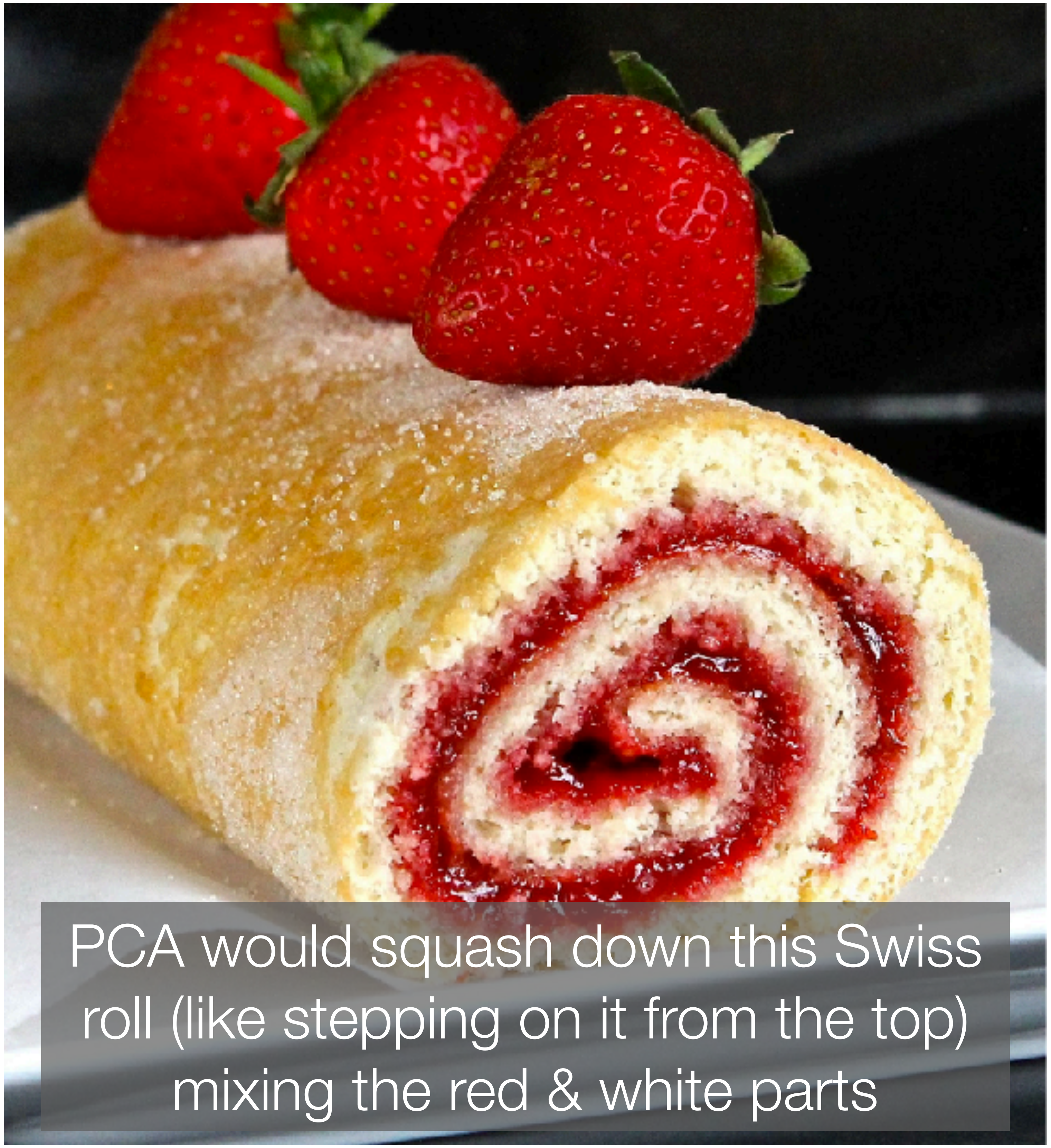
PCA reorients data so axes explain variance in “decreasing order”
➔ can “flatten” (*project*) data onto a few axes that captures most variance

Image source: http://4.bp.blogspot.com/-USQEgoh1jCU/VfncdNOETcI/AAAAAAAAGp8/
Hea8UtE_1c0/s1600/Blog%2B1%2BIMG_1821.jpg

2D Swiss Roll



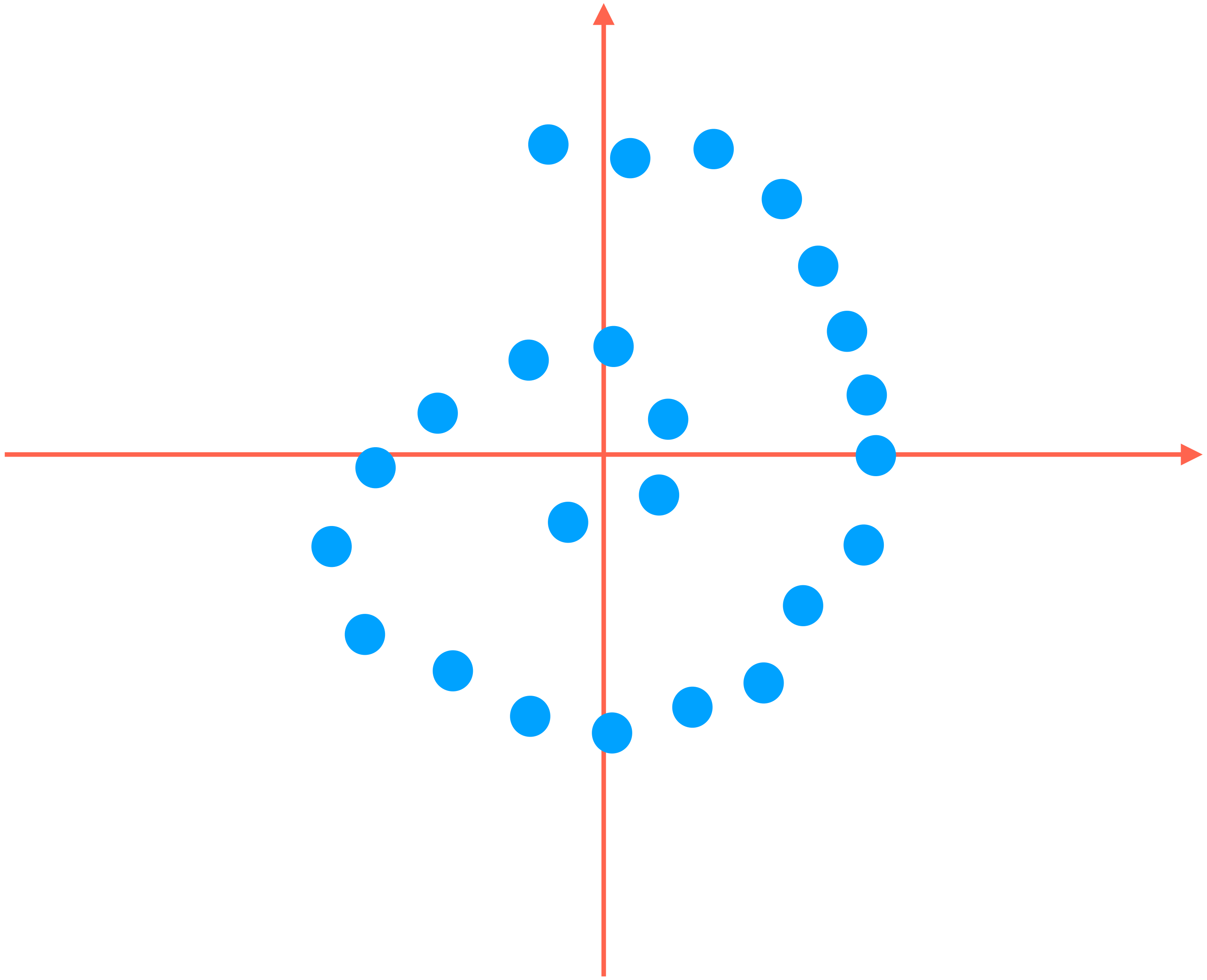
PCA would just flatten this thing and *lose the information that the data actually lives on a 1D line that has been curved!*



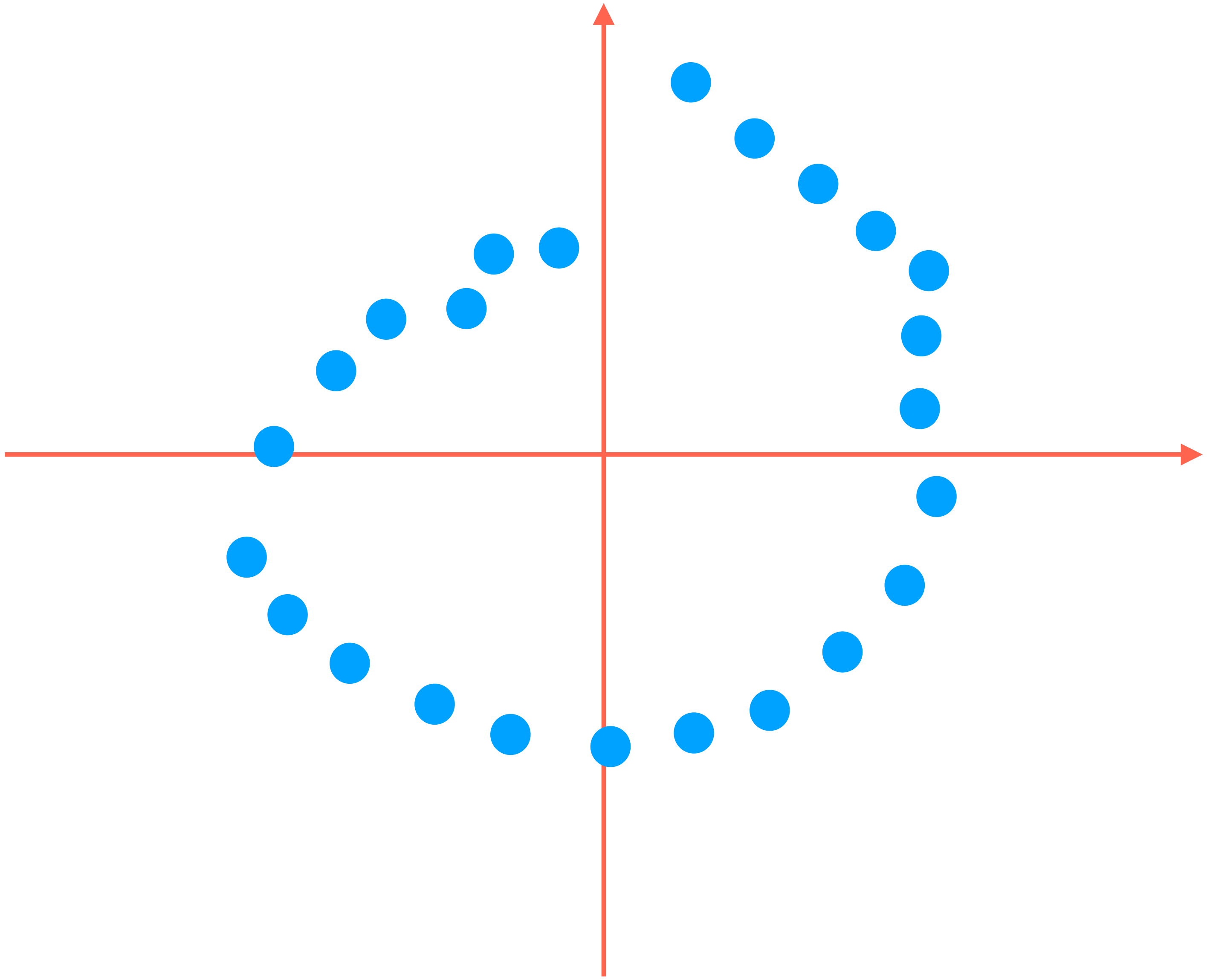
PCA would squash down this Swiss roll (like stepping on it from the top) mixing the red & white parts

Image source: http://4.bp.blogspot.com/-USQEgoh1jCU/VfncdNOETcI/AAAAAAAAGp8/Hea8UtE_1c0/s1600/Blog%2B1%2BIMG_1821.jpg

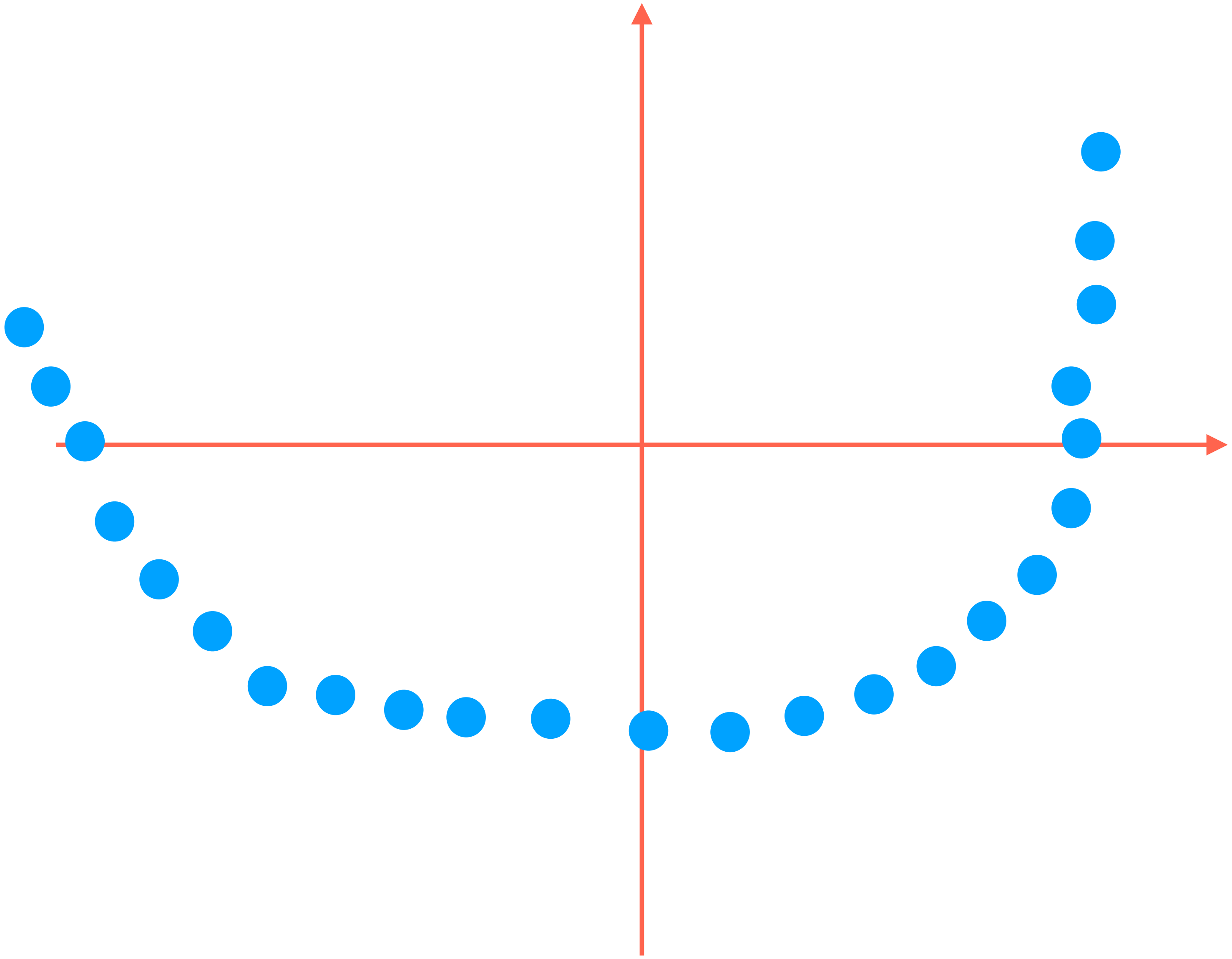
2D Swiss Roll



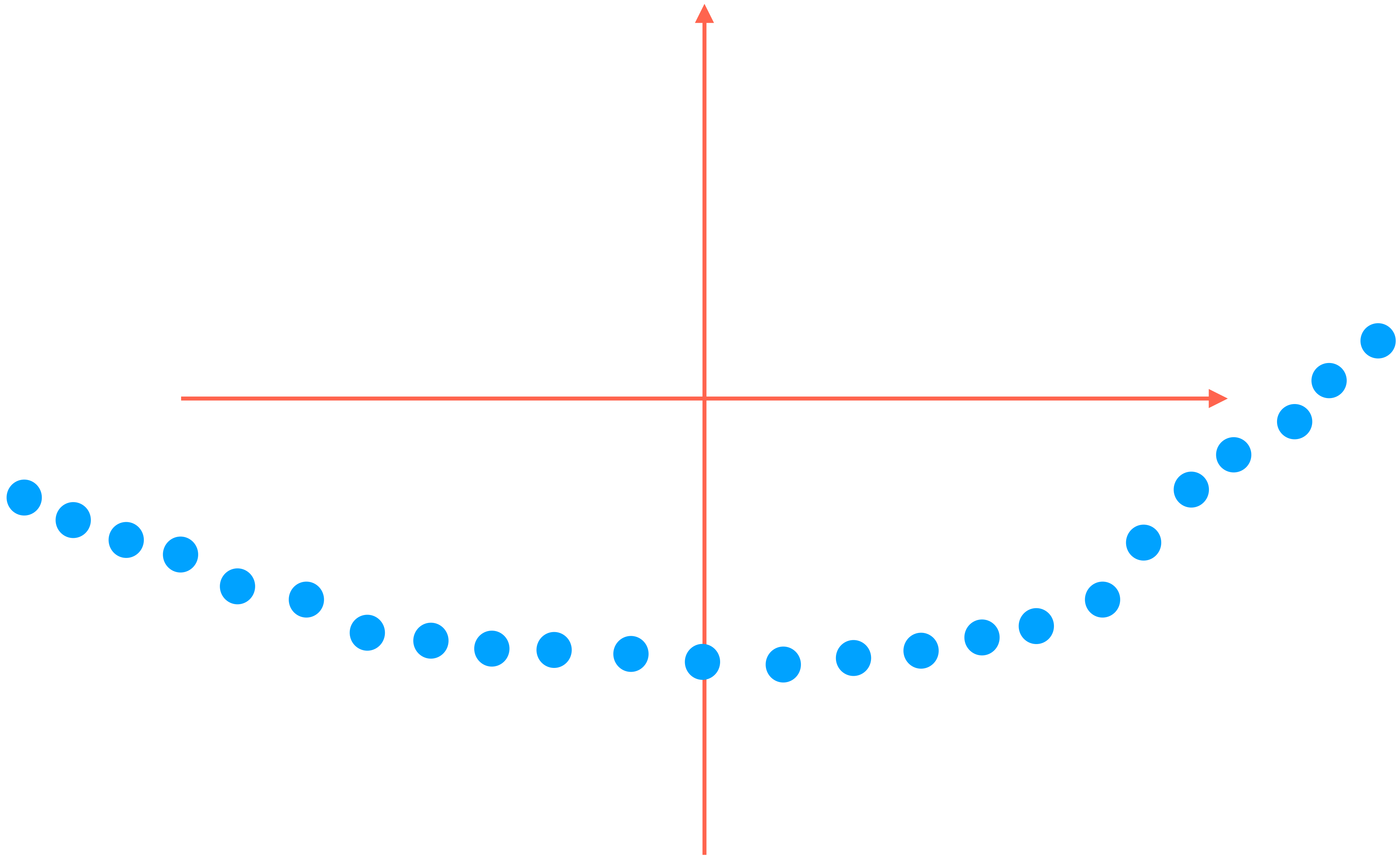
2D Swiss Roll



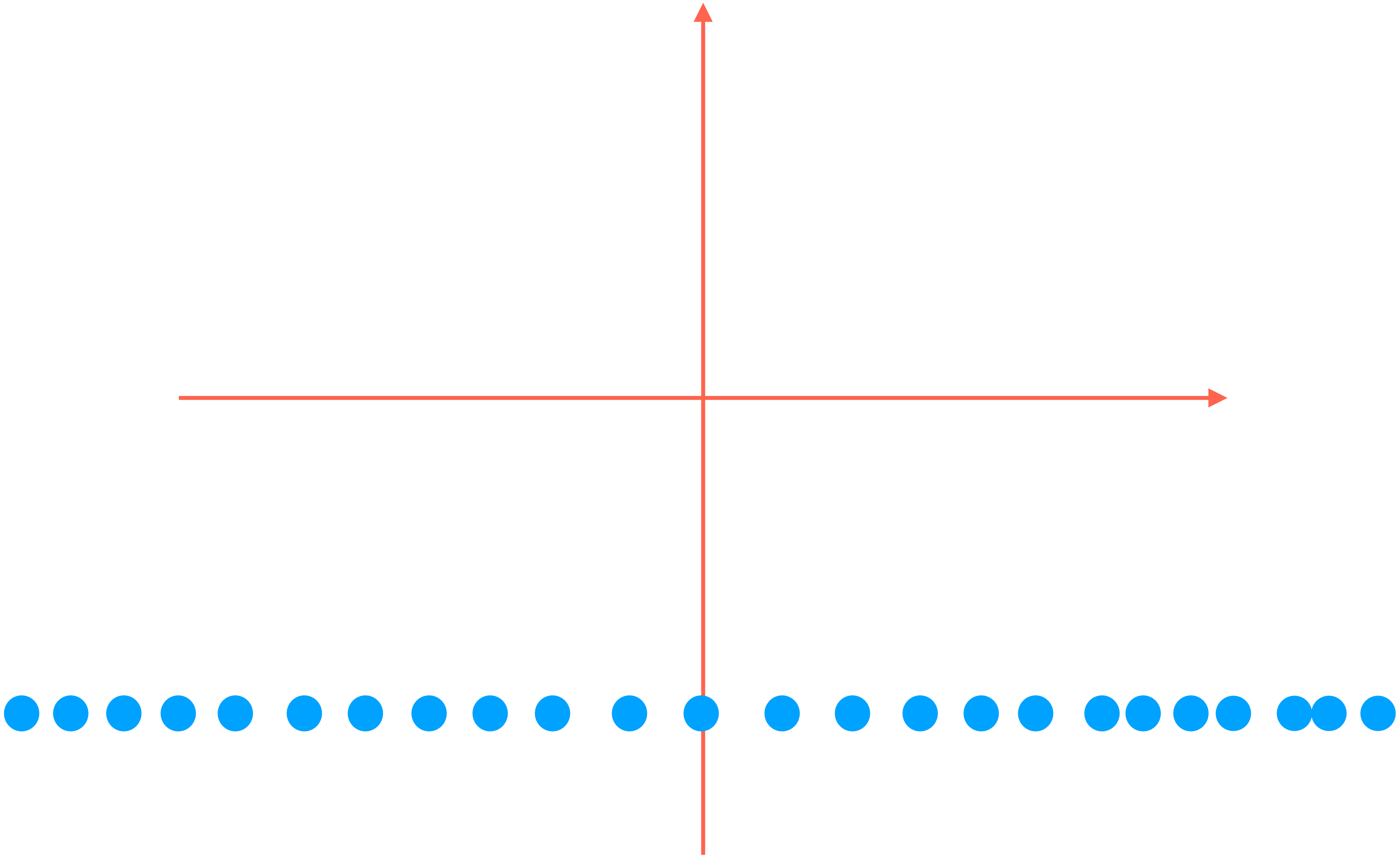
2D Swiss Roll



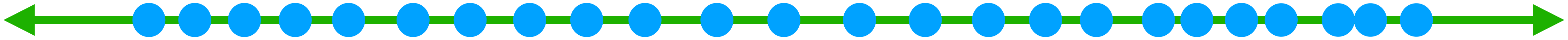
2D Swiss Roll



2D Swiss Roll

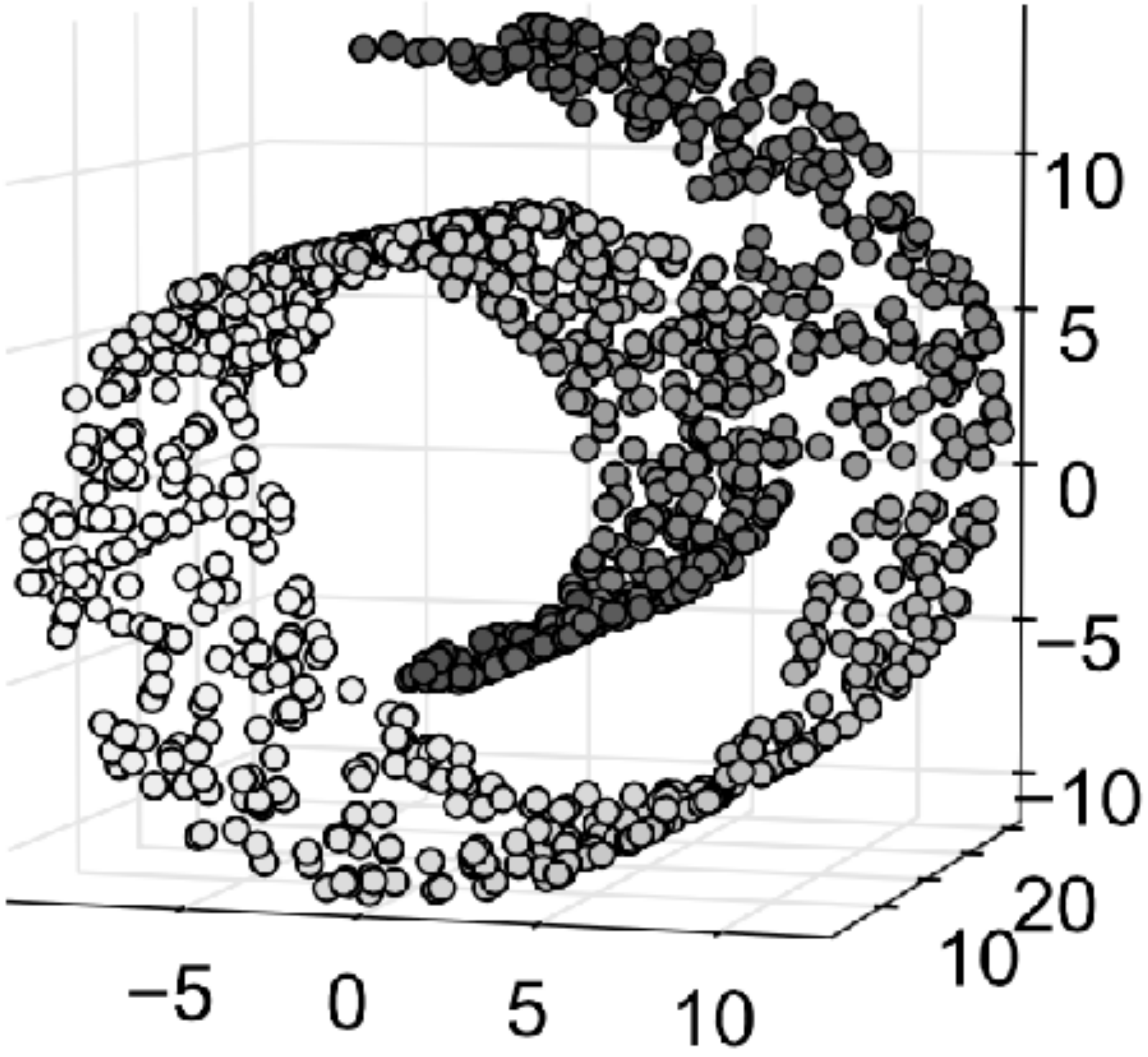


2D Swiss Roll



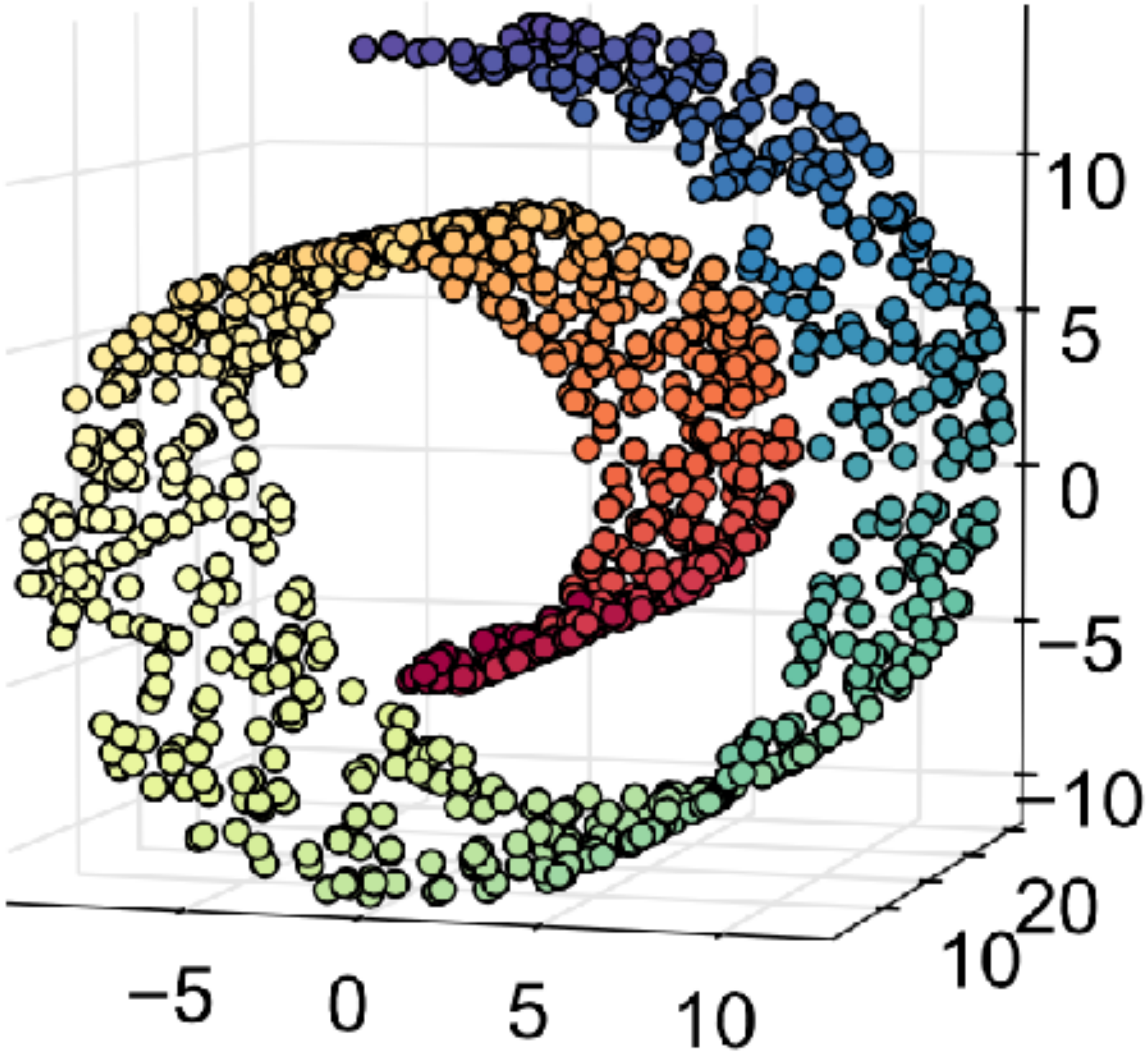
This is the desired result

3D Swiss Roll



Projecting down to any 2D plane puts points that are far apart close together!

3D Swiss Roll

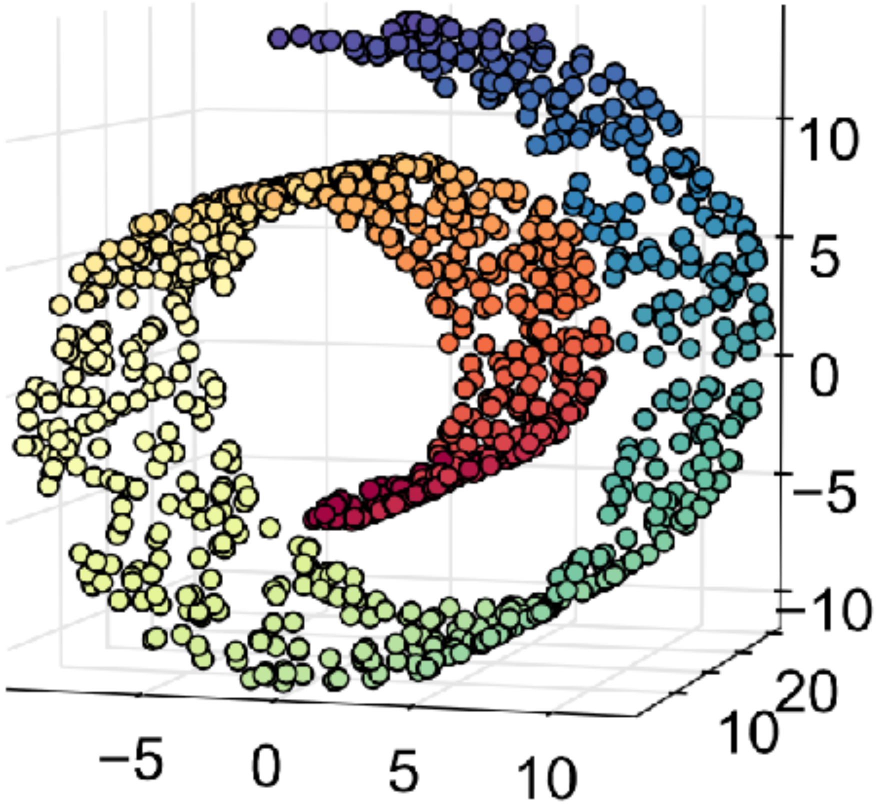


Projecting down to any 2D plane puts points that are far apart close together!

Goal: Low-dimensional representation where similar colored points are near each other (we don't actually get to see the colors)

Manifold Learning

- Nonlinear dimensionality reduction (in contrast to PCA which is linear)
- Find low-dimensional "manifold" that the data live on



Basic idea of a manifold:

1. Zoom in on any point (say, *x*)

2. The points near *x* look like they're in a lower-dimensional Euclidean space (e.g., a 2D plane in Swiss roll)

Do Data Actually Live on Manifolds?

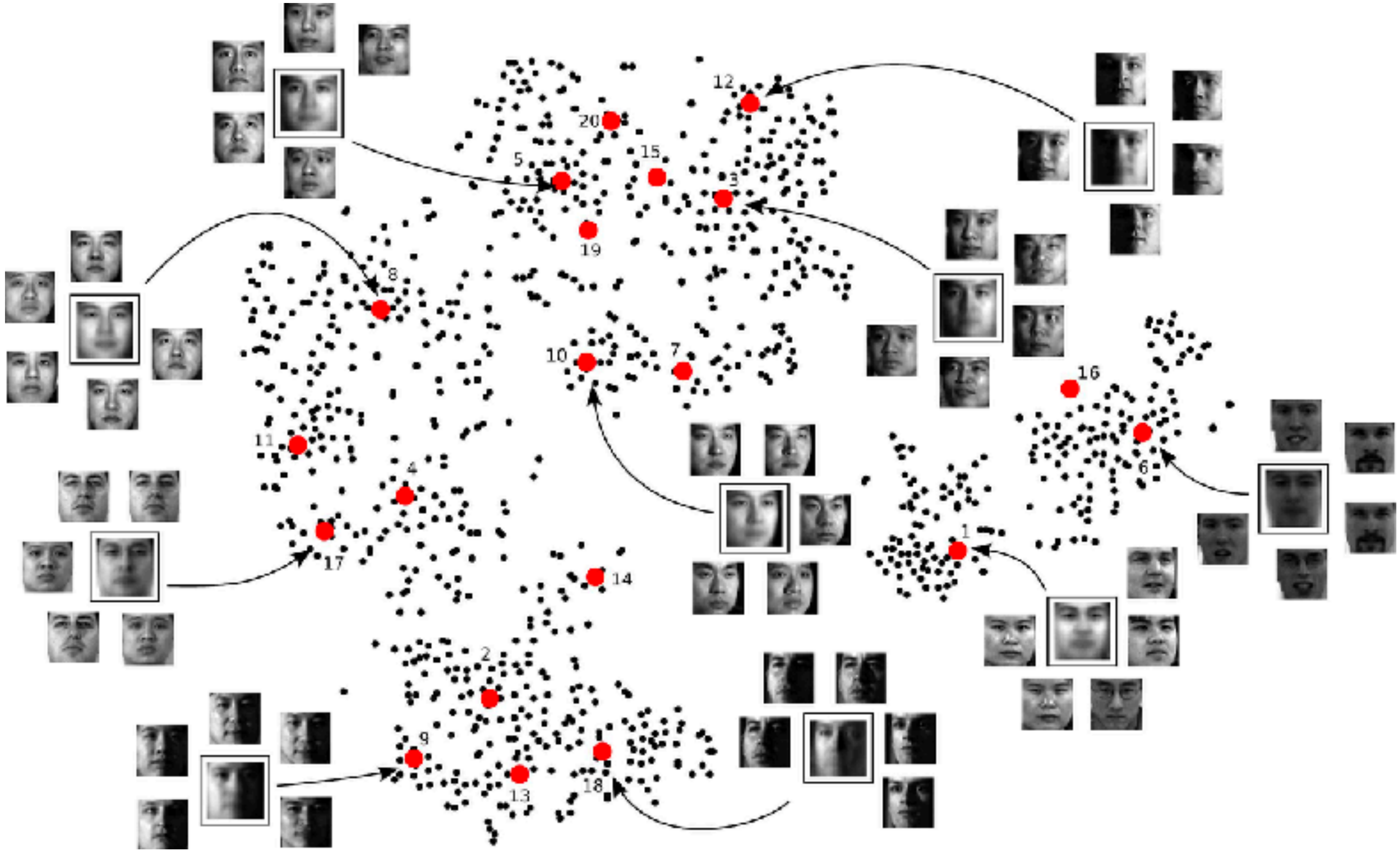
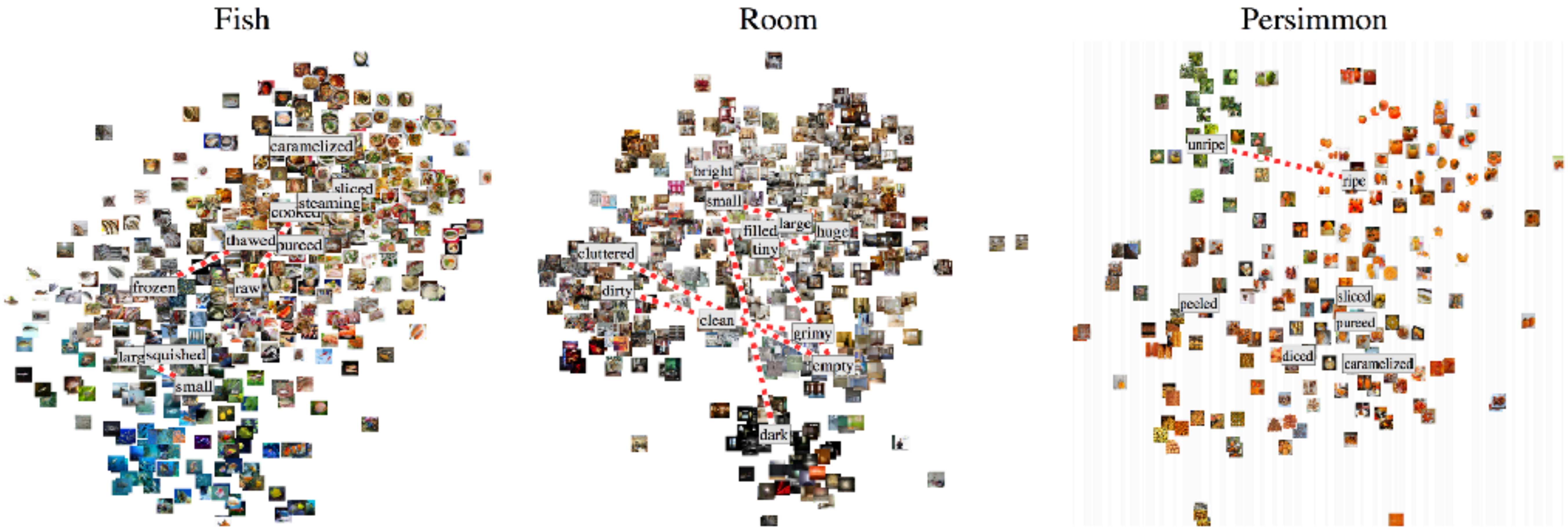


Image source: http://www.columbia.edu/~jwp2128/Images/faces.jpeg

Do Data Actually Live on Manifolds?



Phillip Isola, Joseph Lim, Edward H. Adelson. Discovering States and Transformations in Image Collections. CVPR 2015.

Do Data Actually Live on Manifolds?

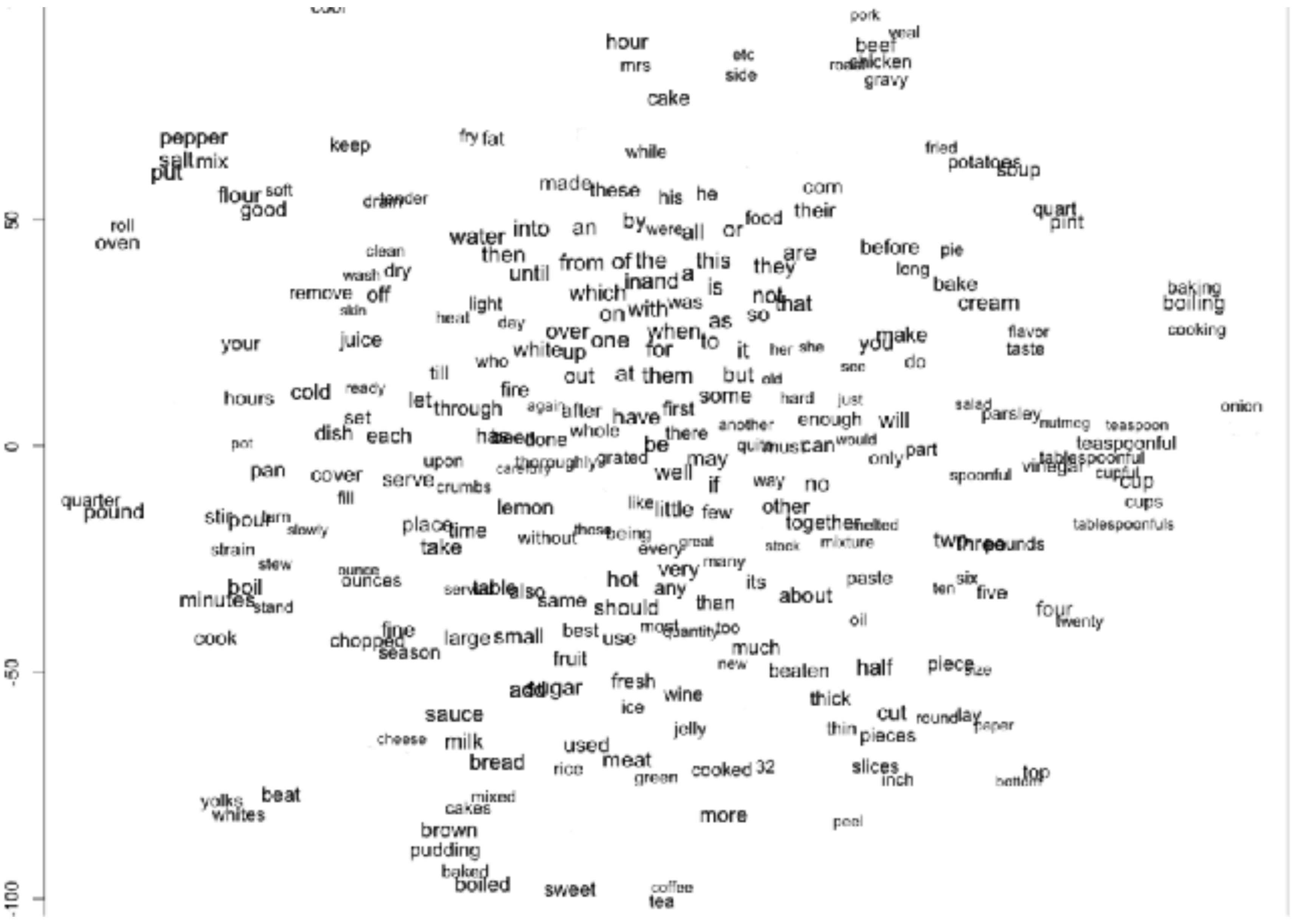
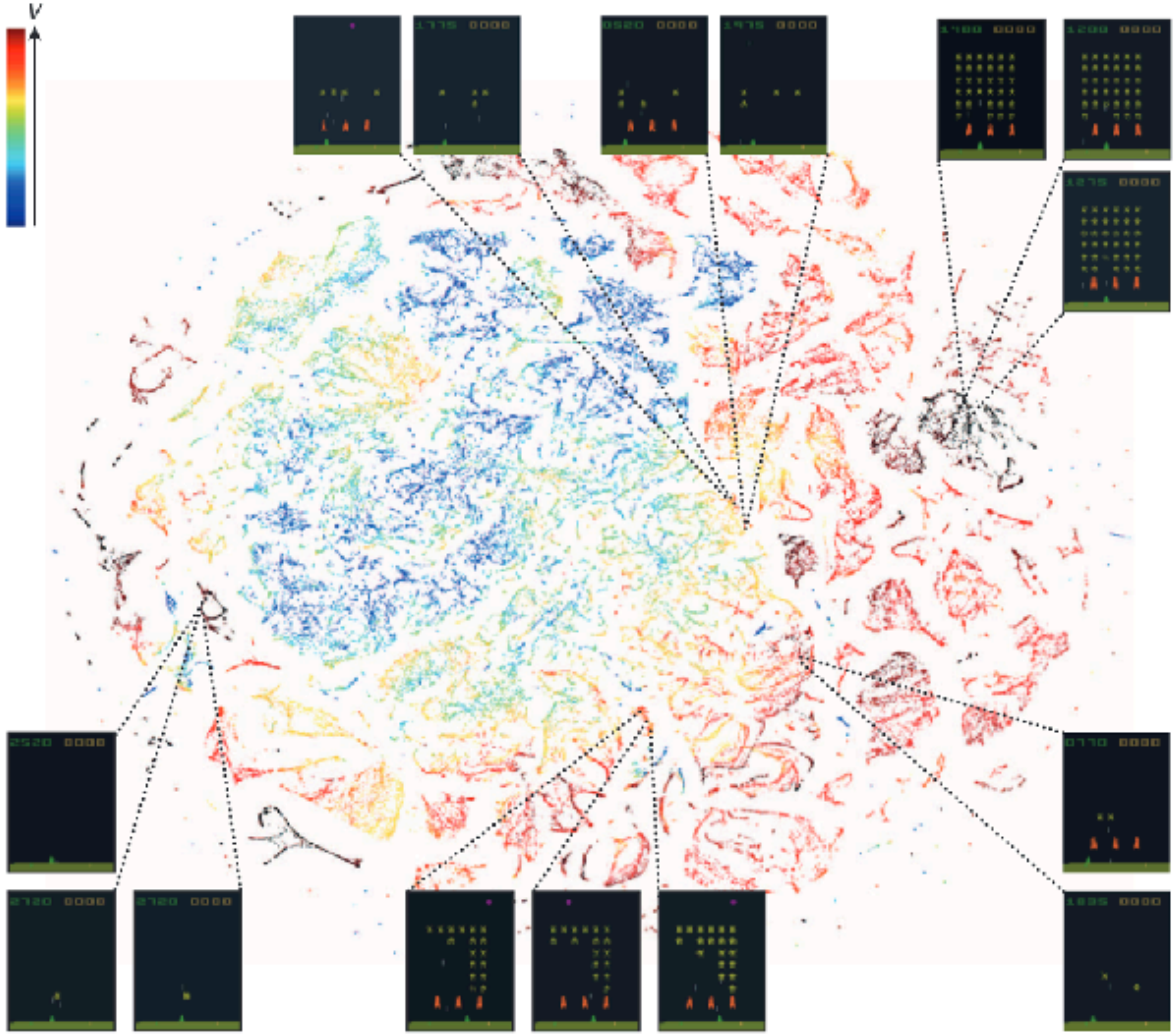


Image source: http://www.adityathakker.com/wp-content/uploads/2017/06/word-embeddings-994x675.png

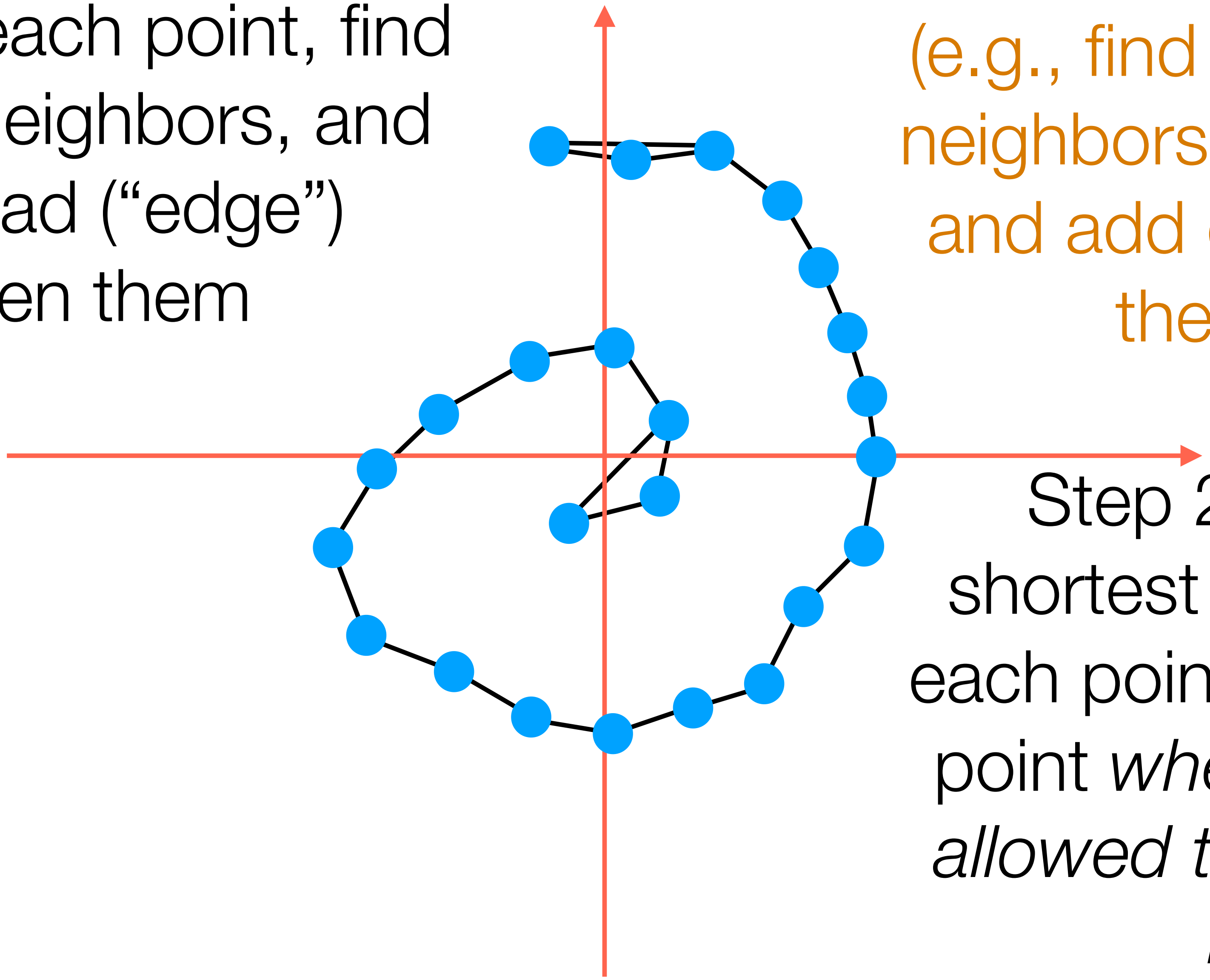
Do Data Actually Live on Manifolds?



Mnih, Volodymyr, et al. Human-level control through deep reinforcement learning. Nature 2015.

Manifold Learning with Isomap

Step 1: For each point, find its nearest neighbors, and build a road (“edge”) between them



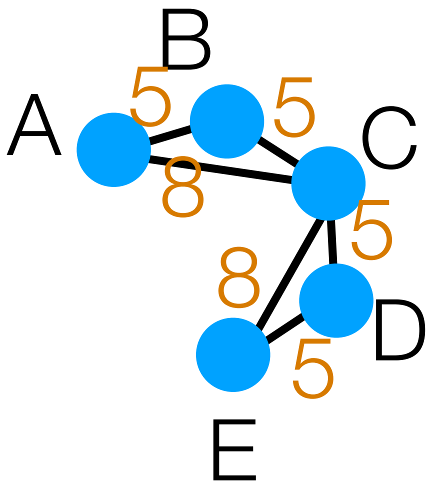
(e.g., find closest 2 neighbors per point and add edges to them)

Step 2: Compute shortest distance from each point to every other point *where you’re only allowed to travel on the roads*

Step 3: It turns out that given all the distances between pairs of points, we can compute what the points should be (the algorithm for this is called *multidimensional scaling*)

Isomap Calculation Example

In orange: road lengths



2 nearest neighbors of A:	B, C
2 nearest neighbors of B:	A, C
2 nearest neighbors of C:	B, D
2 nearest neighbors of D:	C, E
2 nearest neighbors of E:	C, D

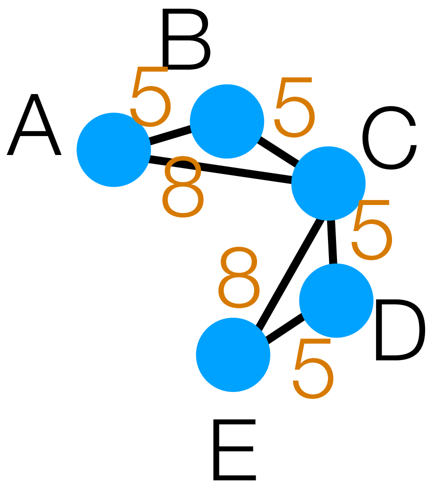
Build "symmetric 2-NN" graph (add edges for each point to its 2 nearest neighbors)

Shortest distances between every point to every other point *where we are only allowed to travel along the roads*

	A	B	C	D	E
A					
B					
C					
D					
E					

Isomap Calculation Example

In orange: road lengths



2 nearest neighbors of A:	B, C
2 nearest neighbors of B:	A, C
2 nearest neighbors of C:	B, D
2 nearest neighbors of D:	C, E
2 nearest neighbors of E:	C, D

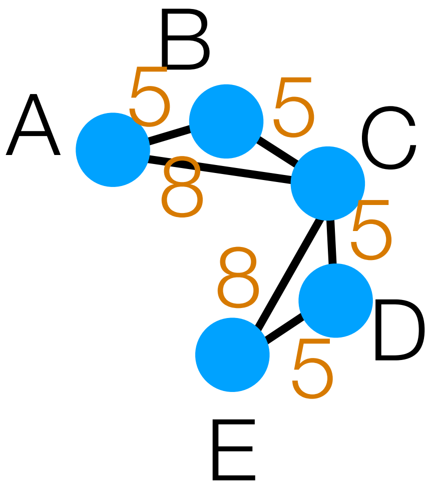
Build "symmetric 2-NN" graph (add edges for each point to its 2 nearest neighbors)

Shortest distances between every point to every other point *where we are only allowed to travel along the roads*

	A	B	C	D	E
A	0				
B		0			
C			0		
D				0	
E					0

Isomap Calculation Example

In orange: road lengths



2 nearest neighbors of A: B, C

2 nearest neighbors of B: A, C

2 nearest neighbors of C: B, D

2 nearest neighbors of D: C, E

2 nearest neighbors of E: C, D

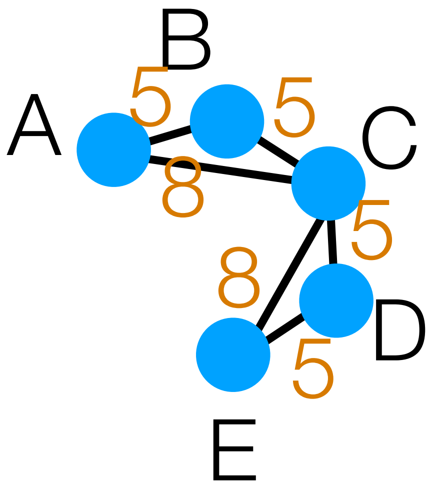
Build "symmetric 2-NN" graph (add edges for each point to its 2 nearest neighbors)

Shortest distances between every point to every other point *where we are only allowed to travel along the roads*

	A	B	C	D	E
A	0	5			
B		0	5		
C			0	5	
D				0	5
E					0

Isomap Calculation Example

In orange: road lengths



2 nearest neighbors of A: B, C

2 nearest neighbors of B: A, C

2 nearest neighbors of C: B, D

2 nearest neighbors of D: C, E

2 nearest neighbors of E: C, D

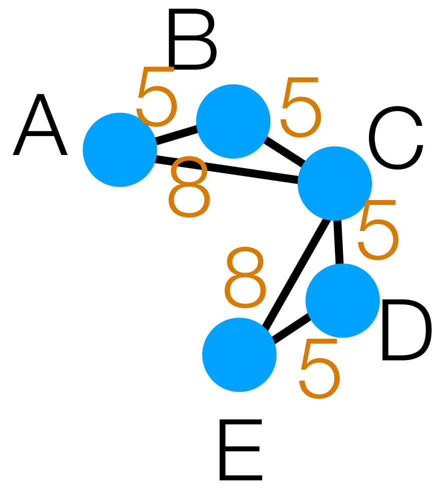
Build "symmetric 2-NN" graph (add edges for each point to its 2 nearest neighbors)

Shortest distances between every point to every other point *where we are only allowed to travel along the roads*

	A	B	C	D	E
A	0	5	8		
B		0	5		
C			0	5	
D				0	5
E					0

Isomap Calculation Example

In orange: road lengths



2 nearest neighbors of A: B, C

2 nearest neighbors of B: A, C

2 nearest neighbors of C: B, D

2 nearest neighbors of D: C, E

2 nearest neighbors of E: C, D

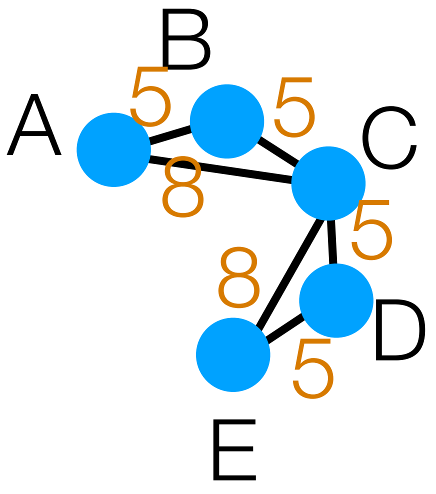
Build "symmetric 2-NN" graph (add edges for each point to its 2 nearest neighbors)

Shortest distances between every point to every other point *where we are only allowed to travel along the roads*

	A	B	C	D	E
A	0	5	8	13	
B		0	5		
C			0	5	
D				0	5
E					0

Isomap Calculation Example

In orange: road lengths



2 nearest neighbors of A: B, C

2 nearest neighbors of B: A, C

2 nearest neighbors of C: B, D

2 nearest neighbors of D: C, E

2 nearest neighbors of E: C, D

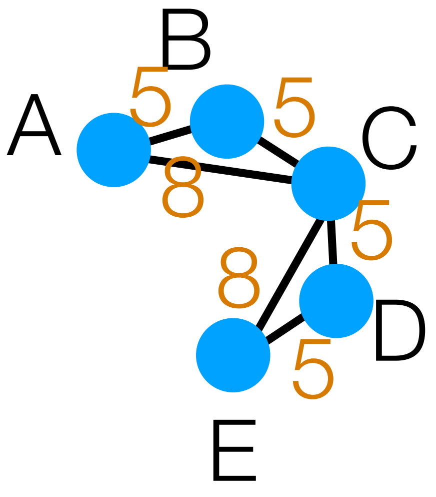
Build "symmetric 2-NN" graph (add edges for each point to its 2 nearest neighbors)

Shortest distances between every point to every other point *where we are only allowed to travel along the roads*

	A	B	C	D	E
A	0	5	8	13	16
B		0	5		
C			0	5	
D				0	5
E					0

Isomap Calculation Example

In orange: road lengths



2 nearest neighbors of A: B, C

2 nearest neighbors of B: A, C

2 nearest neighbors of C: B, D

2 nearest neighbors of D: C, E

2 nearest neighbors of E: C, D

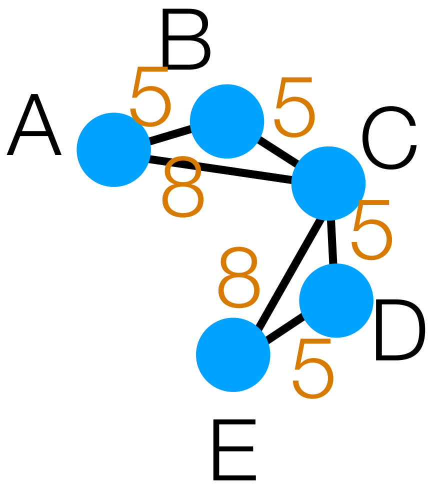
Build "symmetric 2-NN" graph (add edges for each point to its 2 nearest neighbors)

Shortest distances between every point to every other point *where we are only allowed to travel along the roads*

	A	B	C	D	E
A	0	5	8	13	16
B		0	5	10	
C			0	5	
D				0	5
E					0

Isomap Calculation Example

In orange: road lengths



2 nearest neighbors of A: B, C

2 nearest neighbors of B: A, C

2 nearest neighbors of C: B, D

2 nearest neighbors of D: C, E

2 nearest neighbors of E: C, D

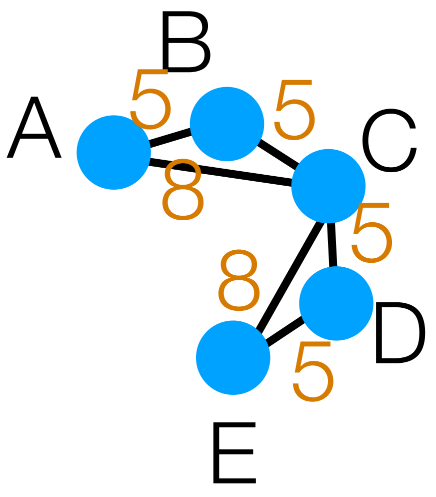
Build "symmetric 2-NN" graph (add edges for each point to its 2 nearest neighbors)

Shortest distances between every point to every other point *where we are only allowed to travel along the roads*

	A	B	C	D	E
A	0	5	8	13	16
B		0	5	10	13
C			0	5	
D				0	5
E					0

Isomap Calculation Example

In orange: road lengths



2 nearest neighbors of A: B, C

2 nearest neighbors of B: A, C

2 nearest neighbors of C: B, D

2 nearest neighbors of D: C, E

2 nearest neighbors of E: C, D

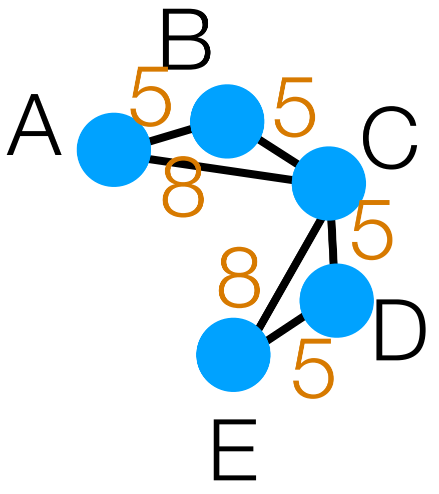
Build "symmetric 2-NN" graph (add edges for each point to its 2 nearest neighbors)

Shortest distances between every point to every other point *where we are only allowed to travel along the roads*

	A	B	C	D	E
A	0	5	8	13	16
B		0	5	10	13
C			0	5	8
D				0	5
E					0

Isomap Calculation Example

In orange: road lengths



2 nearest neighbors of A: B, C

2 nearest neighbors of B: A, C

2 nearest neighbors of C: B, D

2 nearest neighbors of D: C, E

2 nearest neighbors of E: C, D

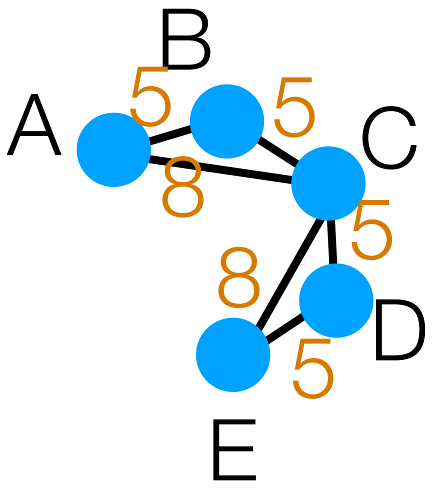
Build "symmetric 2-NN" graph (add edges for each point to its 2 nearest neighbors)

Shortest distances between every point to every other point *where we are only allowed to travel along the roads*

	A	B	C	D	E
A	0	5	8	13	16
B	5	0	5	10	13
C	8	5	0	5	8
D	13	10	5	0	5
E	16	13	8	5	0

Isomap Calculation Example

In orange: road lengths



2 nearest neighbors of A:	B, C
2 nearest neighbors of B:	A, C
2 nearest neighbors of C:	B, D
2 nearest neighbors of D:	C, E
2 nearest neighbors of E:	C, D

Build "symmetric 2-NN" graph (add edges for each point to its 2 nearest neighbors)

Shortest distances between every point to every other point *where we are only allowed to travel along the roads*

	A	B	C	D	E
A	0	5	8	13	16
B	5	0	5	10	13
C	8	5	0	5	8
D	13	10	5	0	5
E	16	13	8	5	0

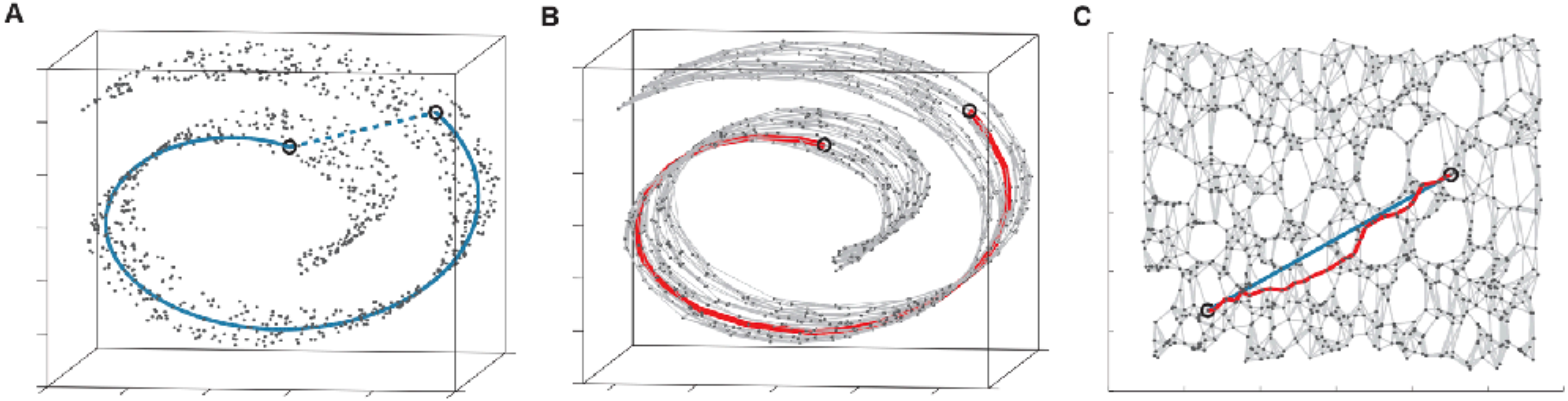
This matrix gets fed into *multidimensional scaling* to get 1D version of A, B, C, D, E

The solution is not unique!

Isomap Calculation Example

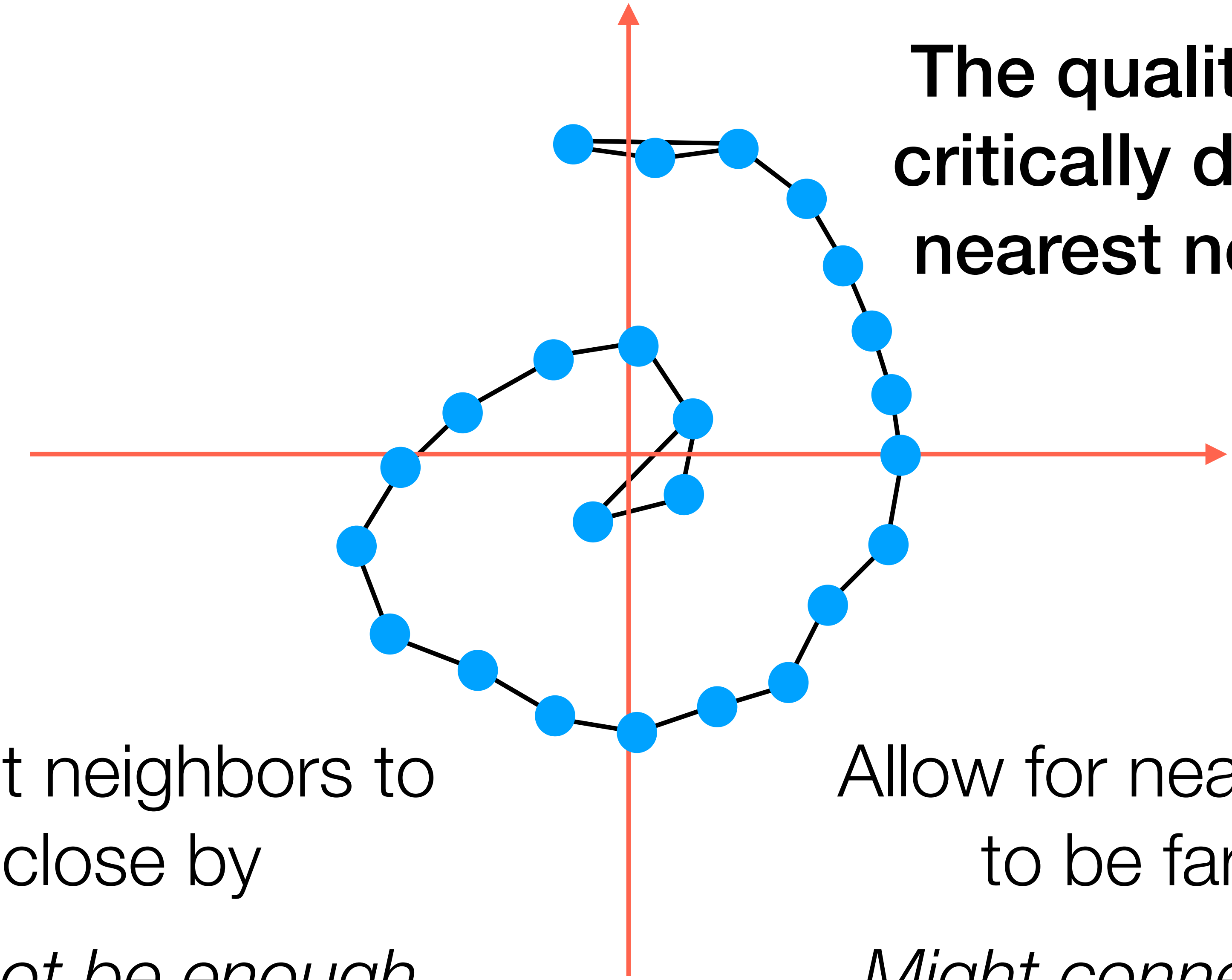
Multidimensional scaling demo

3D Swiss Roll Example



Joshua B. Tenenbaum, Vin de Silva, John C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science 2000.

Some Observations on Isomap



The quality of the result critically depends on the nearest neighbor graph

Ask for nearest neighbors to be really close by

There might not be enough edges

Allow for nearest neighbors to be farther away

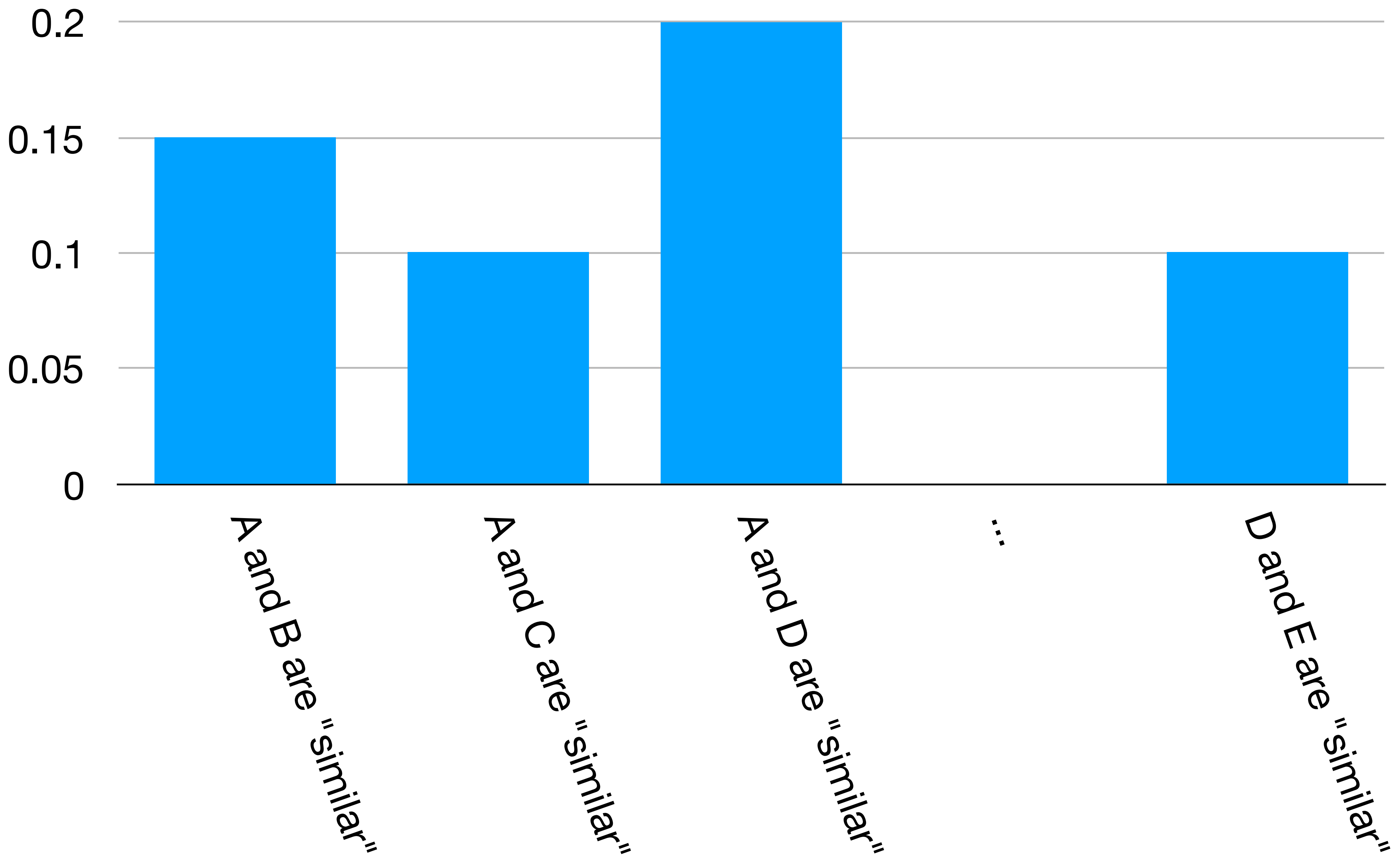
Might connect points that shouldn't be connected

In general: try different parameters for nearest neighbor graph construction when using Isomap + visualize

t-SNE
(t-distributed stochastic neighbor embedding)

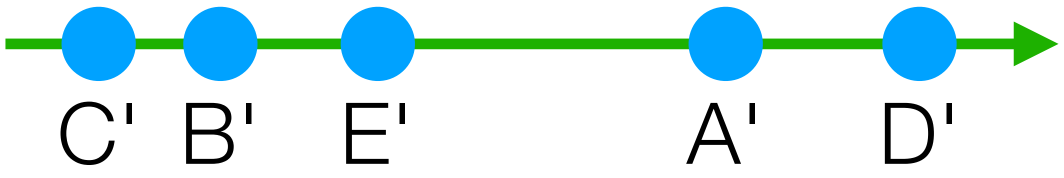
t-SNE High-Level Idea #1

- Don't use deterministic definition of which points are neighbors
- Use probabilistic notation instead

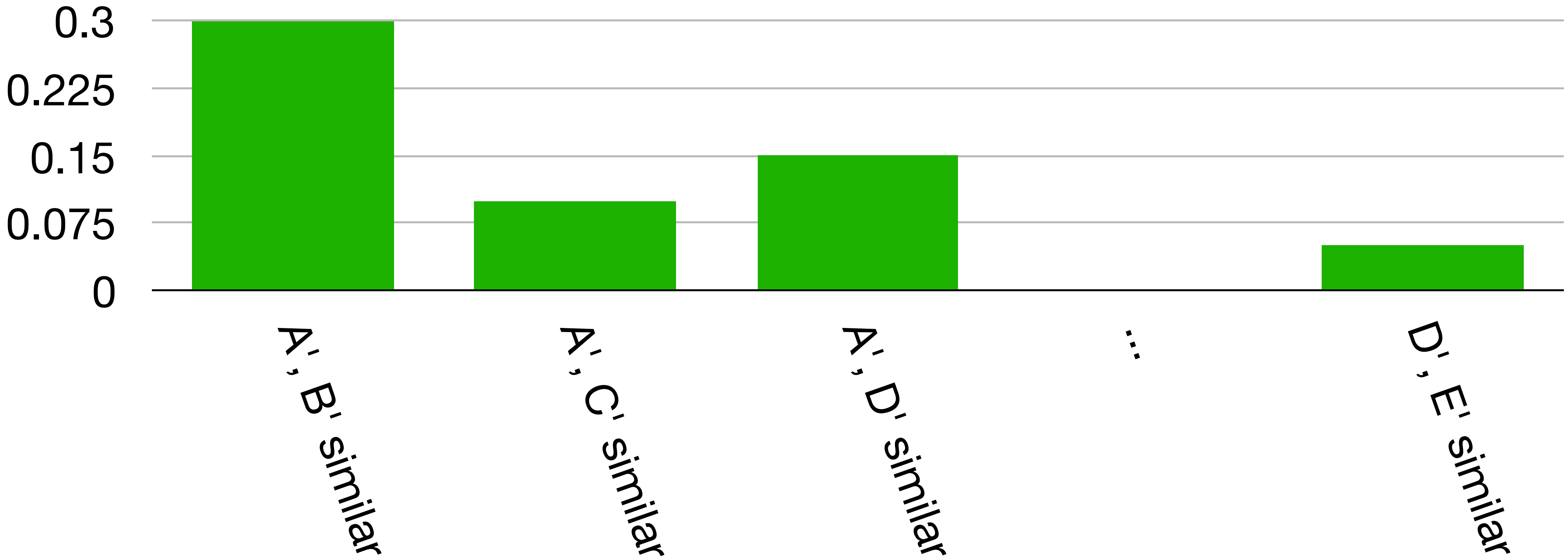


t-SNE High-Level Idea #2

- In low-dim. space (e.g., 1D), suppose we just randomly assigned coordinates as a candidate for a low-dimensional representation for A, B, C, D, E (I'll denote them with primes):

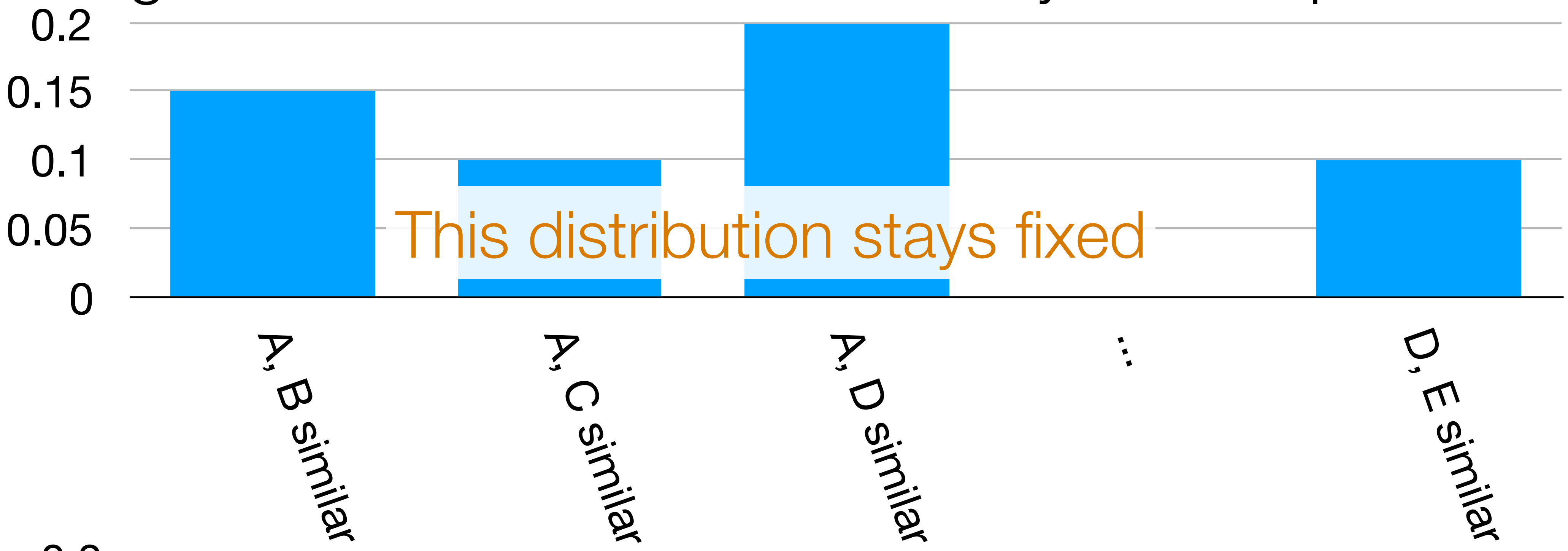


- With any such candidate choice, we can define a probability distribution for these low-dimensional points being similar

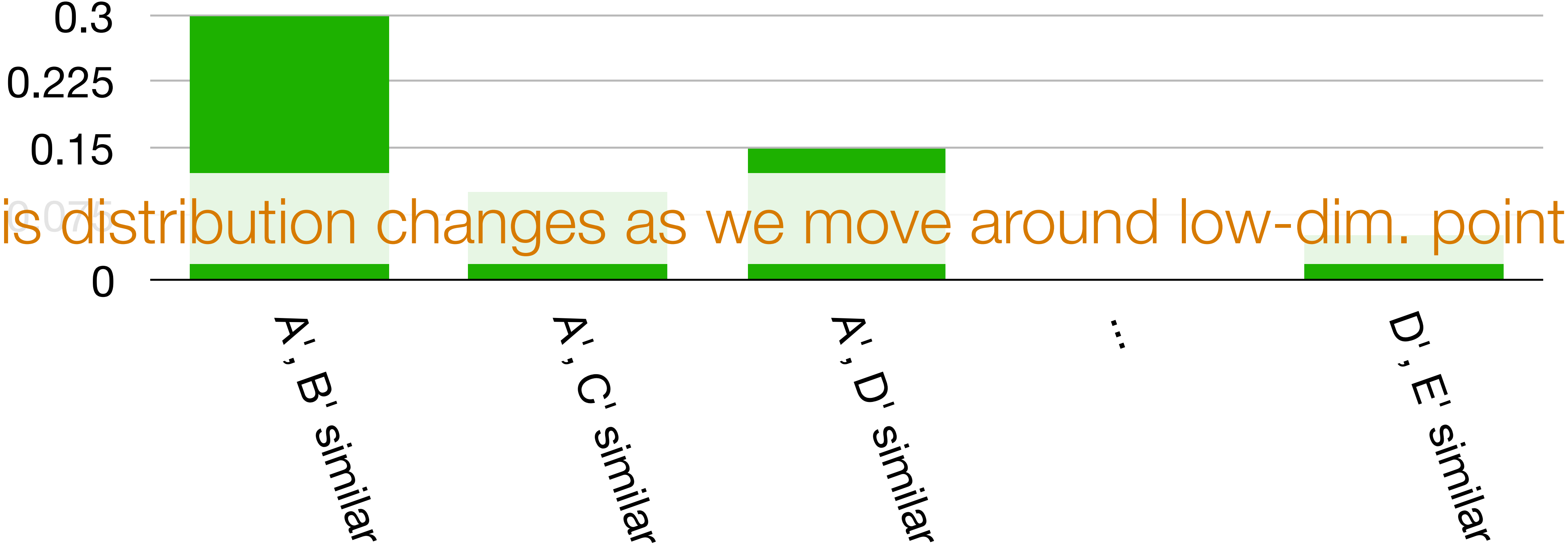


t-SNE High-Level Idea #3

- Keep improving low-dimensional representation to make the following two distributions look as closely alike as possible



This distribution stays fixed



This distribution changes as we move around low-dim. points

Technical Detail for t-SNE

Fleshing out high level idea #1

Suppose there are *n* high-dimensional points $x_1, x_2, \ldots, x_n$

For a specific point *i*, point *i* picks point *j* (≠ *i*) to be a neighbor with probability:

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

σ_i (depends on *i*) controls the probability in which point *j* would be picked by *i* as a neighbor (think about when it gets close to 0 or when it explodes to ∞)

σ_i is controlled by a knob called 'perplexity'
(rough intuition: it is like selecting small vs large neighborhoods for Isomap)

Points *i* and *j* are "similar" with probability: $p_{i,j} = \dfrac{p_{j|i} + p_{i|j}}{2n}$

This defines the earlier blue distribution

Technical Detail for t-SNE

Fleshing out high level idea #2

Denote the *n* low-dimensional points as x_1', x_2', …, x_n'

Low-dim. points *i* and *j* are "similar" with probability: $q_{i,j} = \dfrac{\frac{1}{1+\|x_i' - x_j'\|^2}}{\sum_{k \neq m} \frac{1}{1+\|x_k' - x_m'\|^2}}$

This defines the earlier green distribution

Fleshing out high level idea #3

Approximately minimize (with respect to $q_{i,j}$) the following cost:

$$\sum_{i \neq j} p_{i,j} \log \frac{p_{i,j}}{q_{i,j}}$$

This cost is called the “KL divergence” between distributions ***p*** and ***q***

Manifold Learning with t-SNE

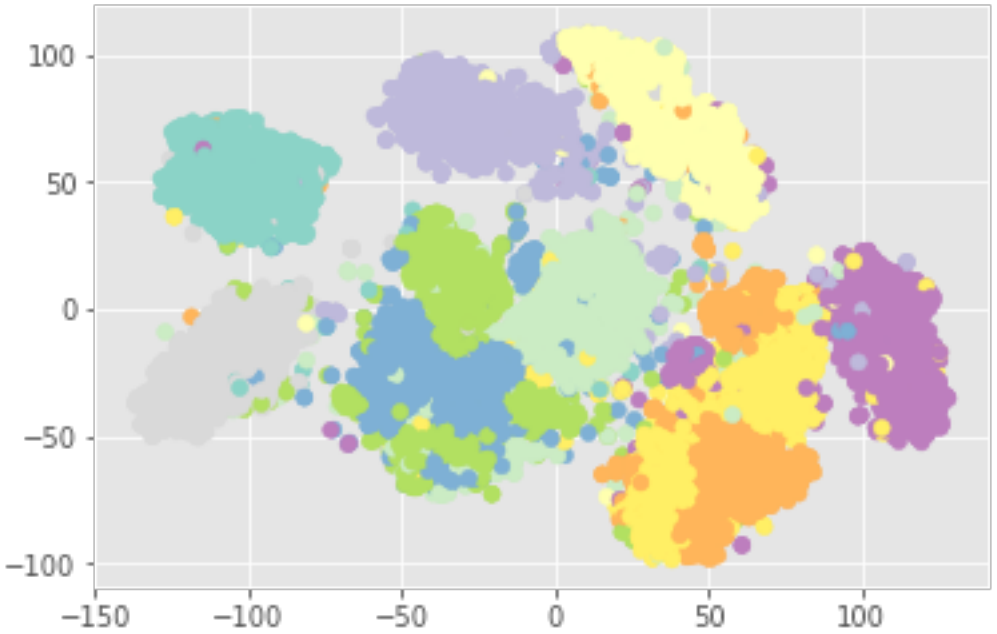
Demo

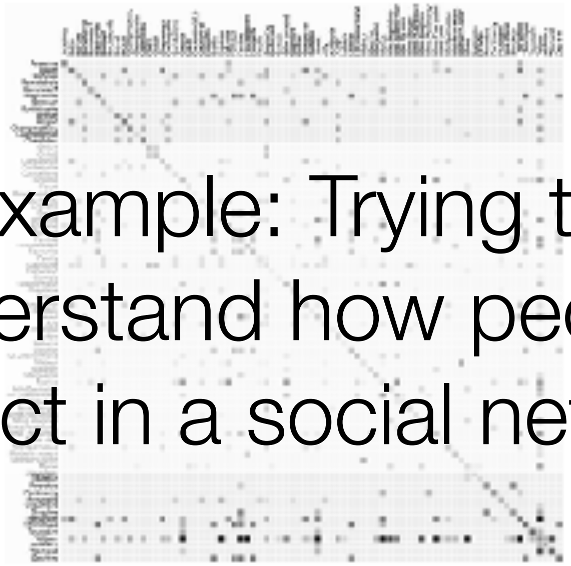
t-SNE Interpretation

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

Visualization

is a way of debugging data analysis!





Example: Trying to understand how people interact in a social network

Important:
Handwritten digit demo was a **toy example** where we know which images correspond to digits 0, 1, … 9

Many real UDA problems:
The data are **messy** and it's not obvious what the "correct" labels/answers look like, and "correct" is ambiguous!

This is largely why I am covering "supervised" methods (require labels) *after* "unsupervised" methods (don't require labels)

Top right image source: https://bost.ocks.org/mike/miserables/

Dimensionality Reduction for Visualization

- There are *many* methods (I've posted a link on the course webpage to a scikit-learn example using ~10 methods)
- PCA is very well-understood; the new axes can be interpreted
- Nonlinear dimensionality reduction: new axes may not really be all that interpretable (you can scale axes, shift all points, etc)
- PCA and t-SNE are good candidates for methods to try first
- If you have good reason to believe that only certain features matter, of course you could restrict your analysis to those!