Carnegie Mellon University
Heinz College

95-865 Unstructured Data Analytics

Week 2: Finding possibly related entities, visualizing high-dimensional data (PCA, Isomap)

George Chen

Co-Occurrences

For example: count # news articles that have different named entities co-occur

	Apple	Facebook	Tesla
Elon Musk	10	15	300
Mark Zuckerberg	500	10000	500
Tim Cook	200	30	10

Big values ➔ *possibly* related named entities

How to downweight "Mark Zuckerberg" if there are just way more articles that mention him?

Key idea: what would happen if people and companies were independent?

	Apple	Facebook	Tesla
Elon Musk	10	15	300
Mark Zuckerberg	500	10000	500
Tim Cook	200	30	10

Probability of drawing "Elon Musk, Apple"?

Probability of drawing a card that says "Apple" on it?

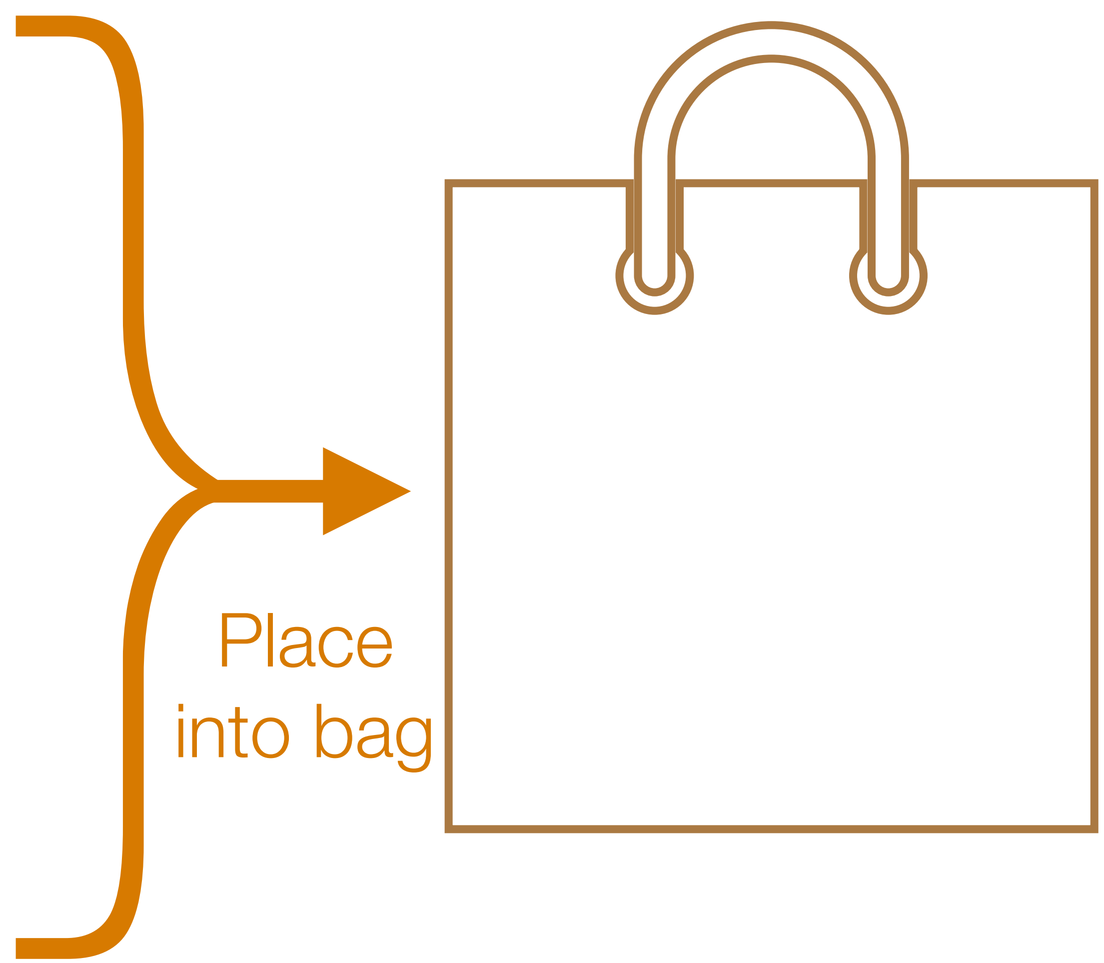
10 of these cards: Elon Musk, Apple

15 of these cards: Elon Musk, Facebook

300 of these cards: Elon Musk, Tesla

⋮

10 of these cards: Tim Cook, Tesla



Co-occurrence table

	Apple	Facebook	Tesla
Elon Musk	10	15	300
Mark Zuckerberg	500	10000	500
Tim Cook	200	30	10

Total: 11565

Joint probability table

	Apple	Facebook	Tesla
Elon Musk	10 /11565	15 /11565	300 /11565
Mark Zuckerberg	500 /11565	10000 /11565	500 /11565
Tim Cook	200 /11565	30 /11565	10 /11565

sum to get P(Elon Musk)

Total: 11565

Joint probability table

	Apple	Facebook	Tesla	
Elon Musk	0.00086	0.00130	0.02594	**0.02810**
Mark Zuckerberg	0.04323	0.86468	0.04323	**0.95115**
Tim Cook	0.01729	0.00259	0.00086	**0.02075**
	0.06139	**0.86857**	**0.07004**	

Recall: if events A and B are independent, $P(A, B) = P(A)P(B)$

Joint probability table **if people and companies were independent**

	Apple	Facebook	Tesla	
Elon Musk	0.00173	0.02441	0.00197	**0.02810**
Mark Zuckerberg	0.05839	0.82614	0.06662	**0.95115**
Tim Cook	0.00127	0.01802	0.00145	**0.02075**
	0.06139	**0.86857**	**0.07004**	

Recall: if events A and B are independent, $P(A, B) = P(A)P(B)$

What we actually observe

	Apple	Facebook	Tesla
Elon Musk	0.00086	0.00130	0.02594
Mark Zuckerberg	0.04323	0.86468	0.04323
Tim Cook	0.01729	0.00259	0.00086

What should be the case if people are companies are independent

	Apple	Facebook	Tesla
Elon Musk	0.00173	0.02441	0.00197
Mark Zuckerberg	0.05839	0.82614	0.06662
Tim Cook	0.00127	0.01802	0.00145

Pointwise Mutual Information (PMI)

Probability of A and B co-occurring

$$\frac{\boxed{P(A, B)}}{\boxed{P(A)\ P(B)}}$$

if equal to 1
➔ A, B are indep.

Probability of A and B co-occurring *if they were independent*

PMI(A, B) is defined as the log of the above ratio

PMI measures (the log of) a ratio that says how far A and B are from being independent

Looking at All Pairs of Outcomes

- PMI measures how P(A, B) differs from P(A)P(B) using a **log ratio**
- **Log ratio** isn't the only way to compare!
- Another way to compare:

$$\frac{[\ P(A, B) - P(A)\ P(B)\ ]^2}{P(A)\ P(B)}$$

$$\text{Phi-square} = \sum_{A,\,B} \frac{[\ P(A, B) - P(A)\ P(B)\ ]^2}{P(A)\ P(B)}$$

$$\text{Chi-square} = N \times \text{Phi-square}$$

N = sum of all co-occurrence counts

Phi-square is between 0 and min(#rows, #cols)-1

0 ➔ pairs are all indep.

Measures how close *all* pairs of outcomes are close to being indep.

PMI/Phi-Square/Chi-Square Calculation

Demo

Co-occurrence Analysis Applications

- If you're an online store/retailer:
  anticipate *when* certain products are likely to be purchased/rented/consumed more
    - Products & dates
- If you have a bunch of physical stores:
  anticipate *where* certain products are likely to be purchased/rented/consumed more
    - Products & locations
- If you're the police department:
  create "heat map" of where different criminal activity occurs
    - Crime reports & locations

Co-occurrence Analysis Applications

- If you're an online store/retailer:
  anticipate *when* certain products are likely to be purchased/
  re
  -
- If y
  an ... sed/
  re
  -
- If y
  cr ... curs
  - Crime reports & locations

Examples of data to take advantage of:

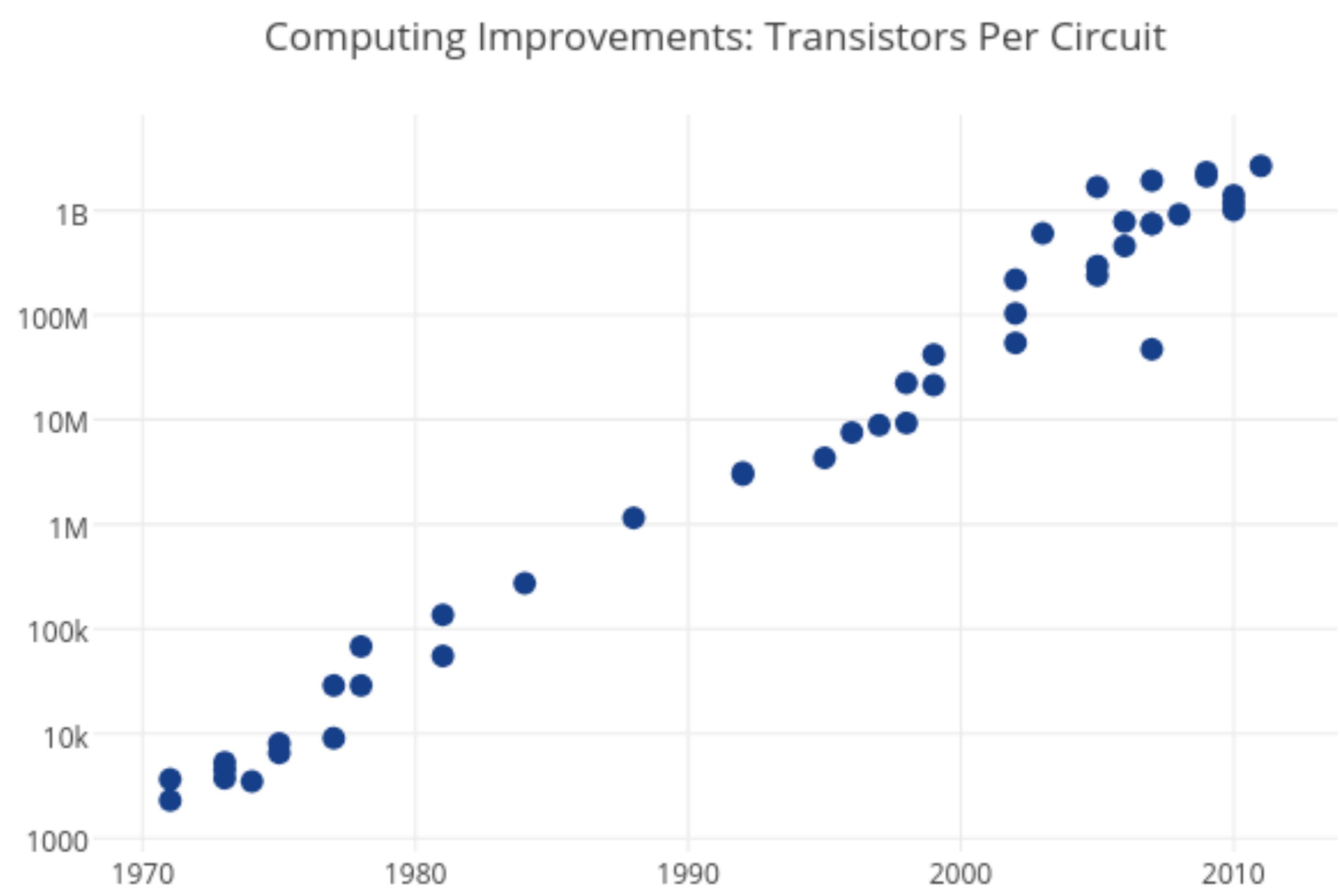
- data collected by your organization
- social networks
- news websites
- blogs

Web scraping frameworks can be helpful:

- Scrapy
- Selenium (great with JavaScript-heavy pages)

Continuous Measurements

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

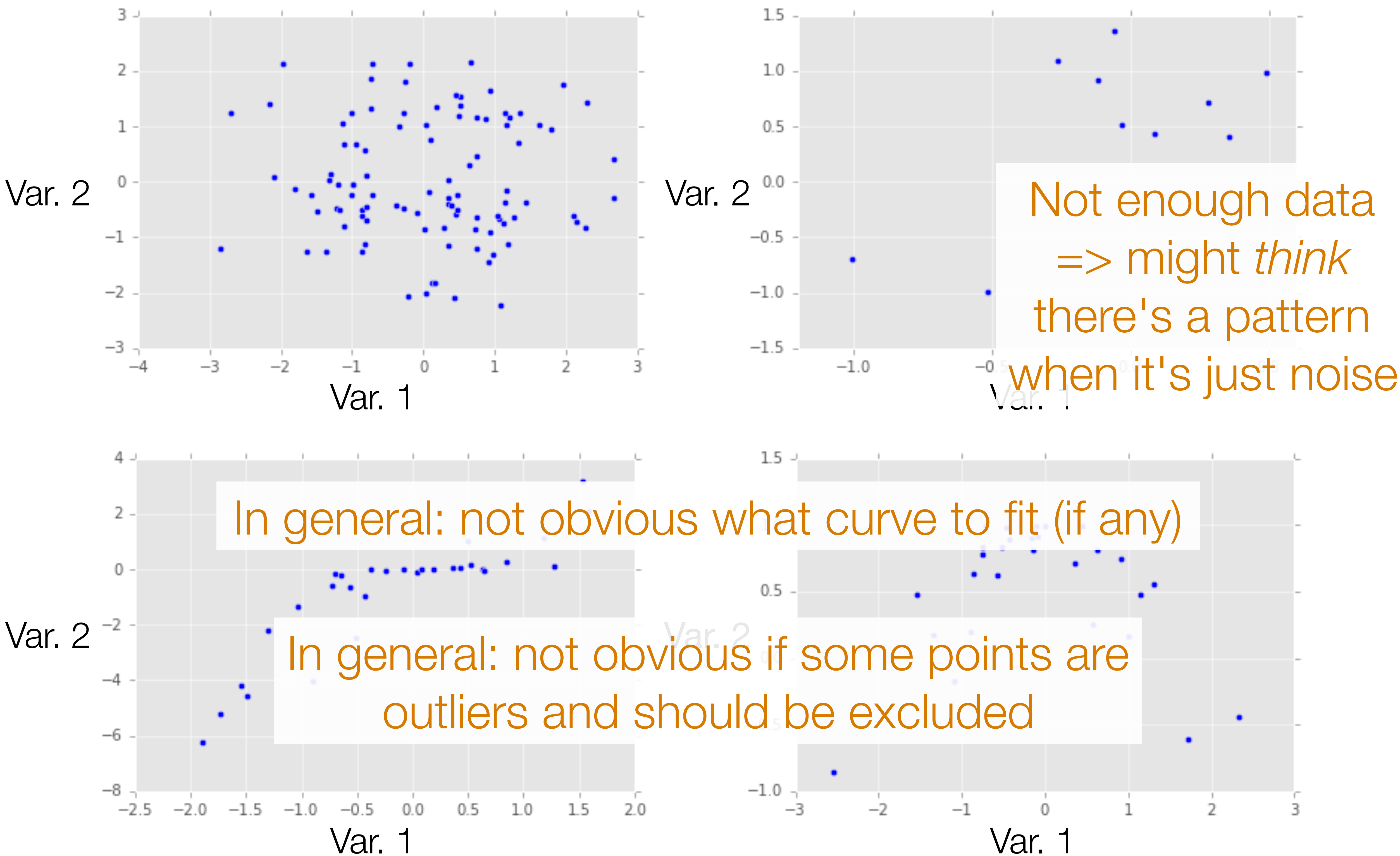


Of course, not all trends look like a line

(so don't just do linear regression!)

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

The Importance of Staring at Data

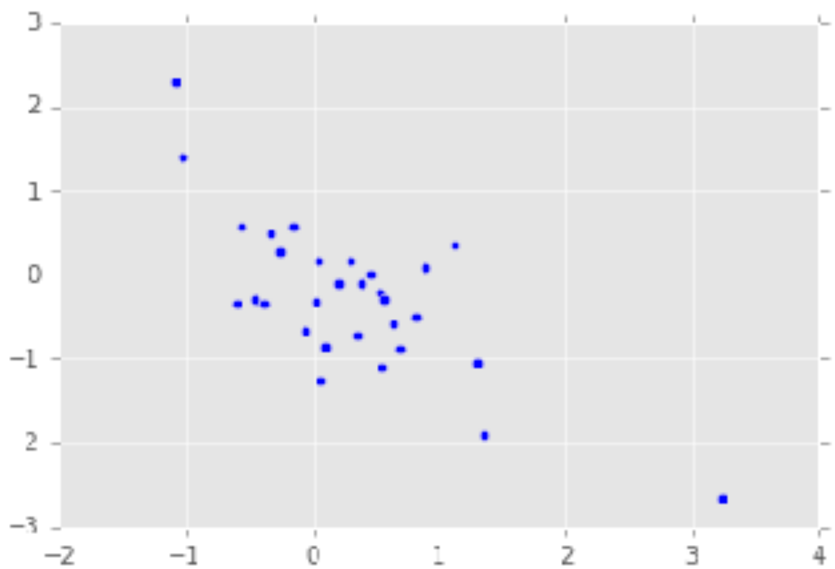


Not enough data => might *think* there's a pattern when it's just noise

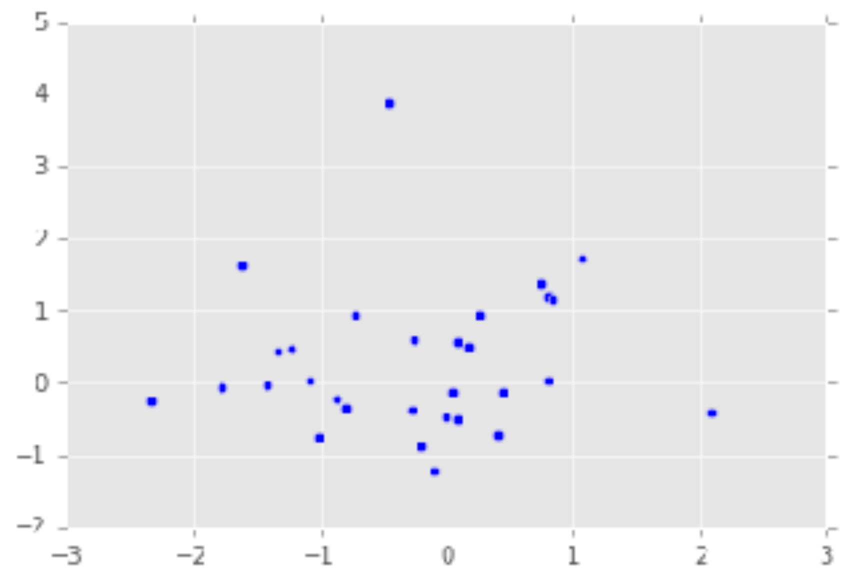
In general: not obvious what curve to fit (if any)

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

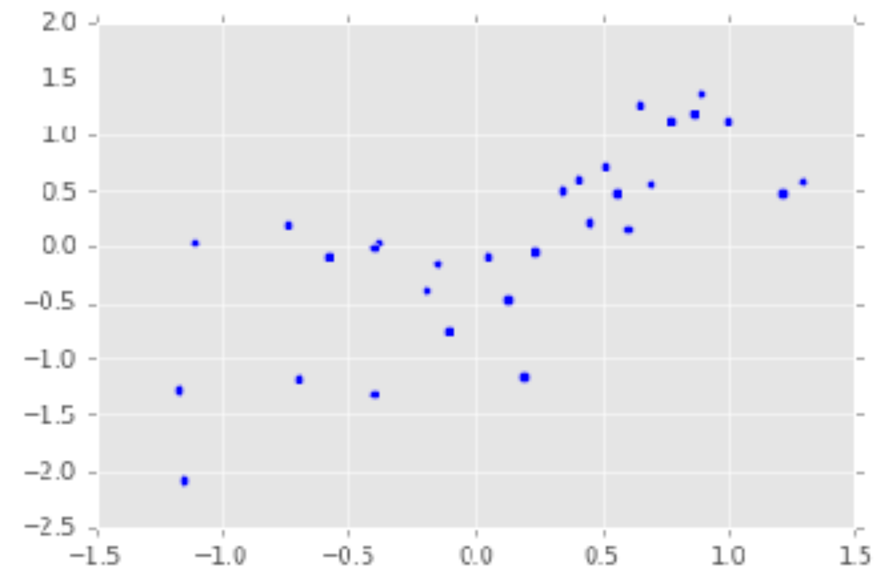
Correlation



Negatively correlated



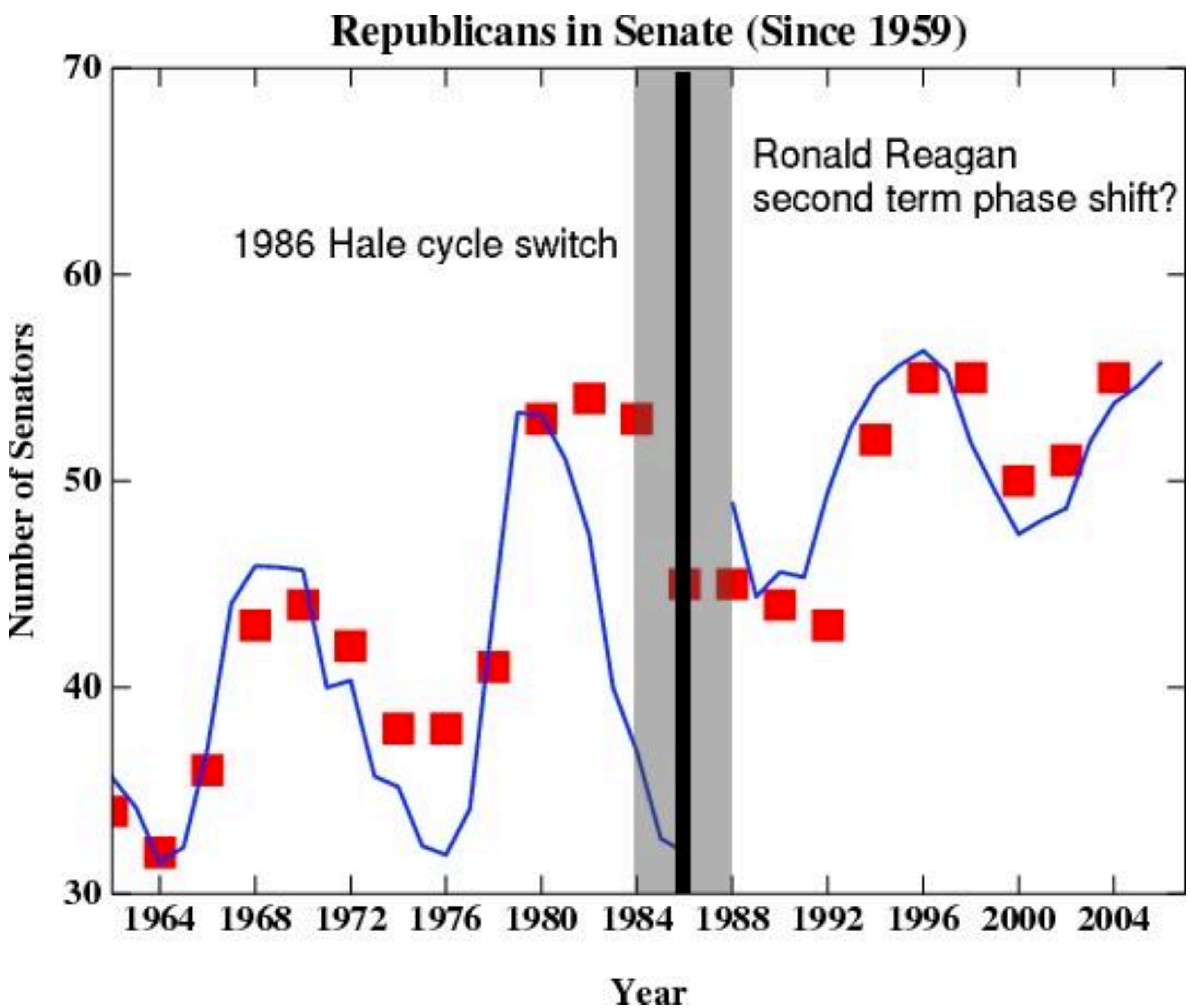
Not really correlated



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?

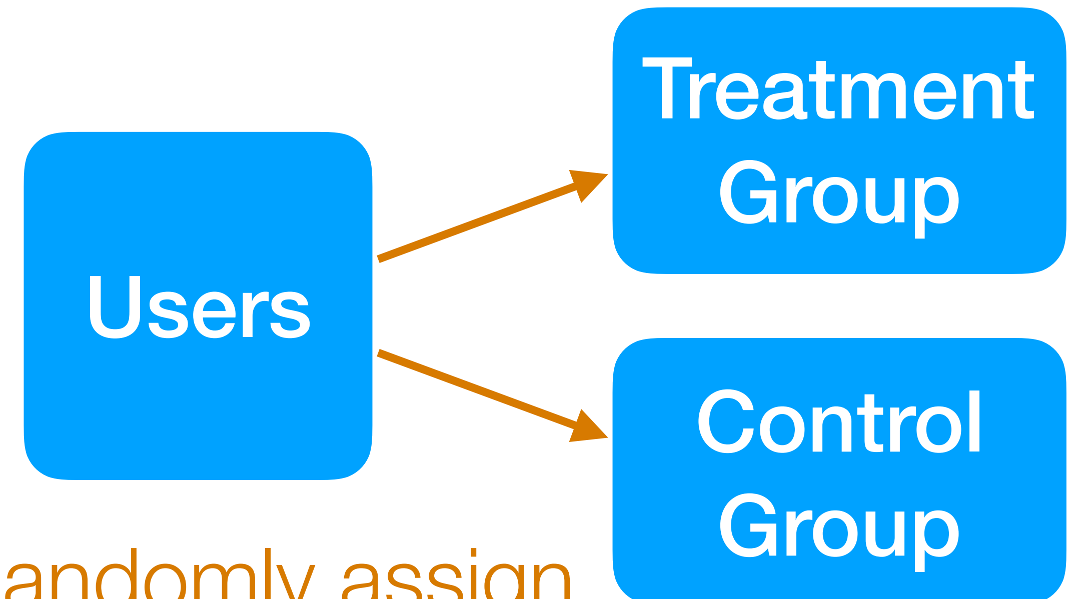
Back then: coffee drinkers also tended to smoke more than non-coffee drinkers (smoking is a **confounding variable**)

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



Randomly assign

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

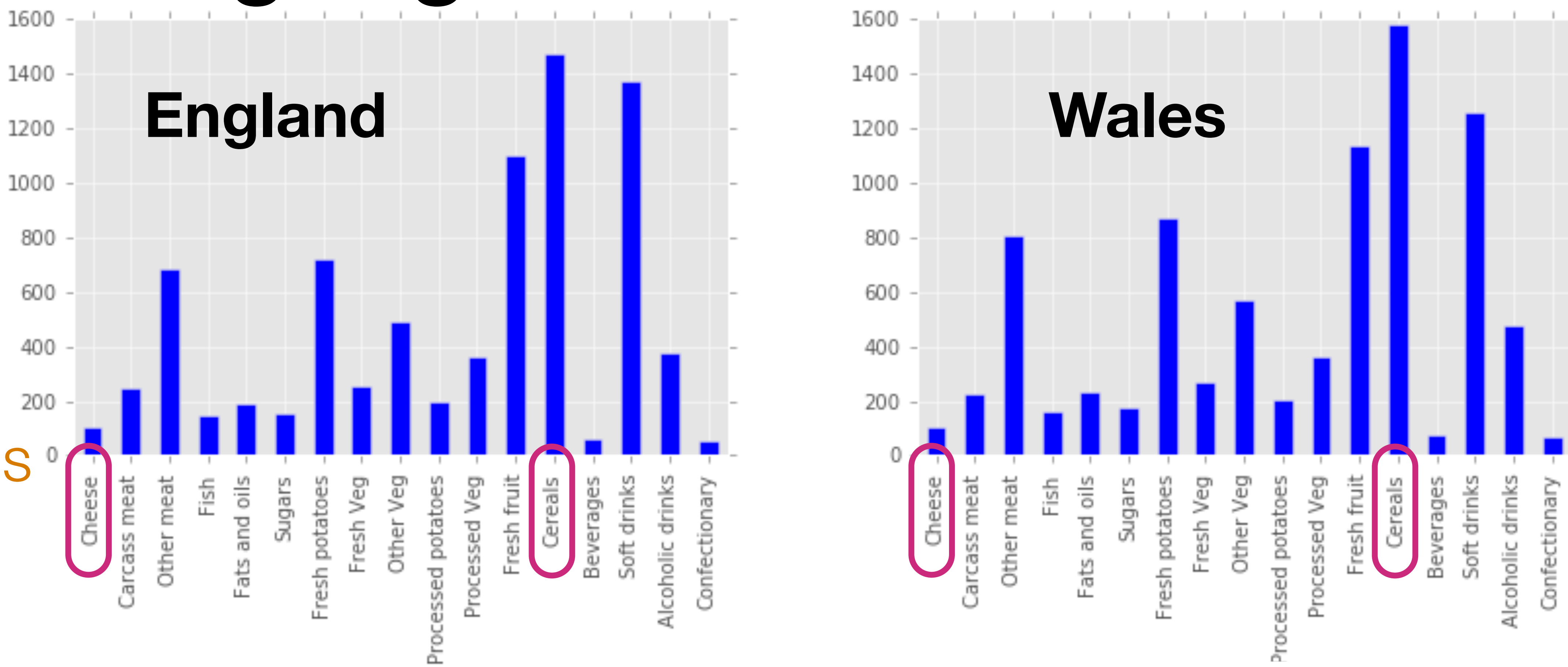
Make predictions using structure found in Part I

- 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)

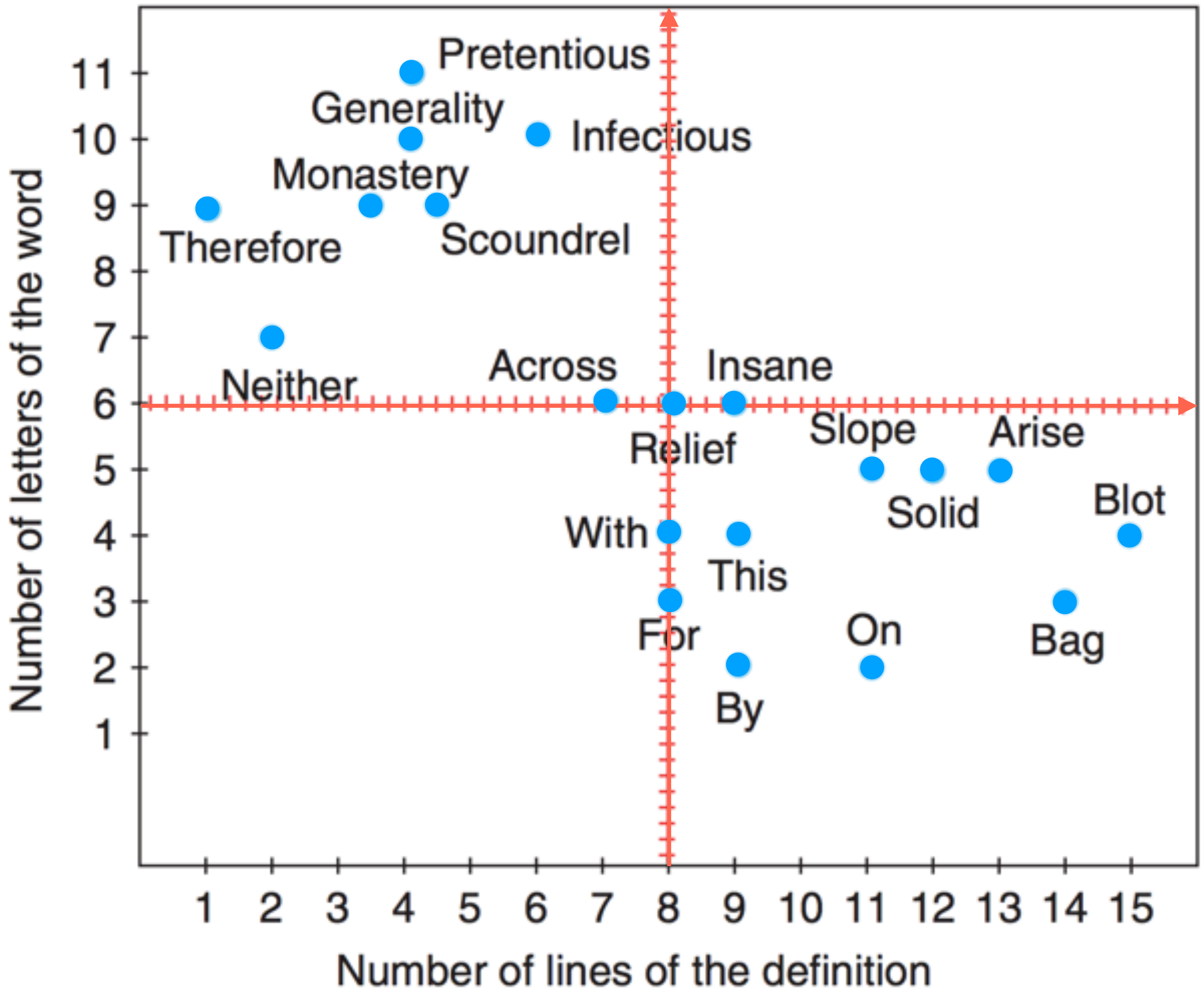
But unclear how to compare the locations (England, Wales, Scotland, N. Ireland)!

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)

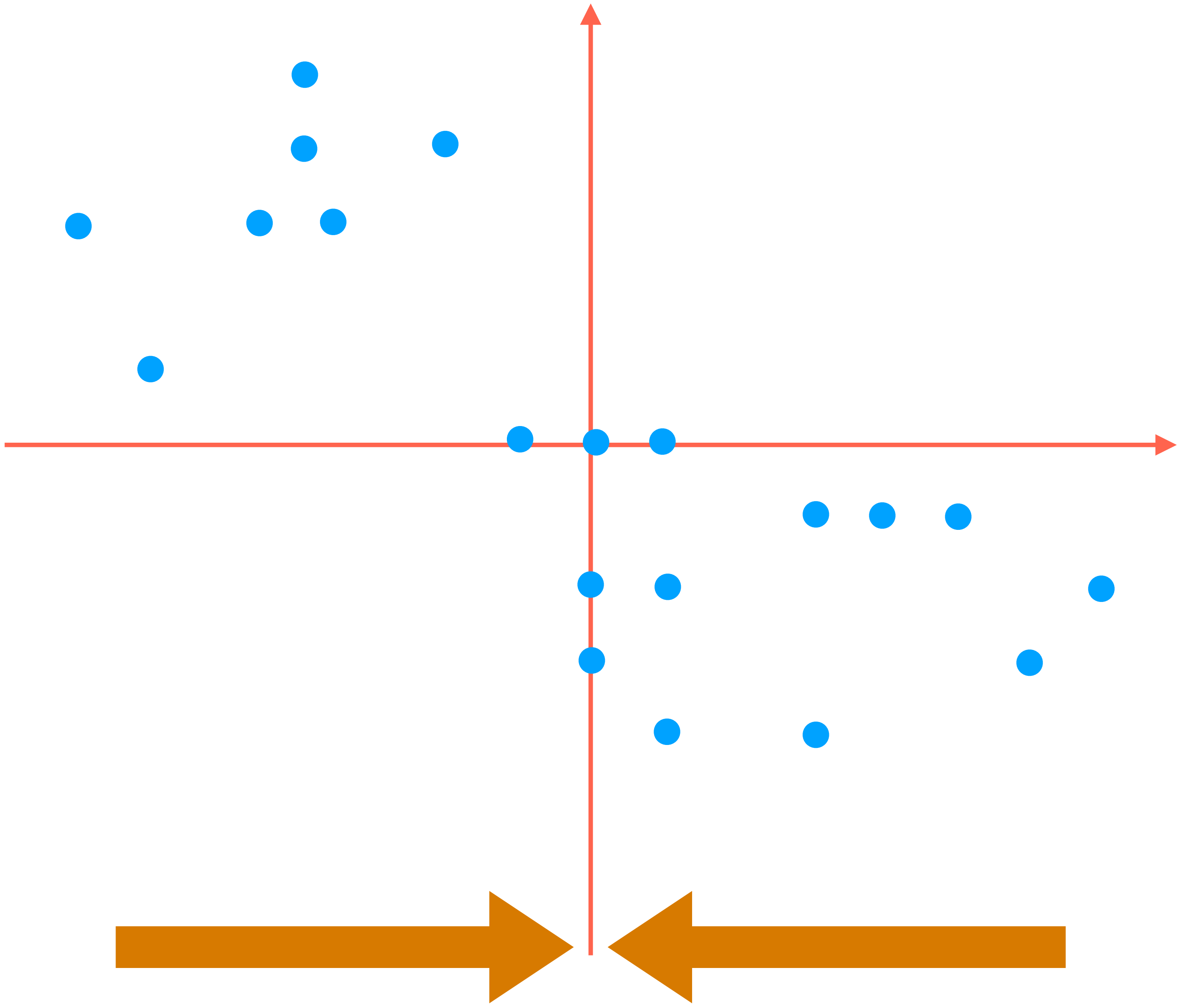
How to project 2D data down to 1D?



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

Principal Component Analysis (PCA)

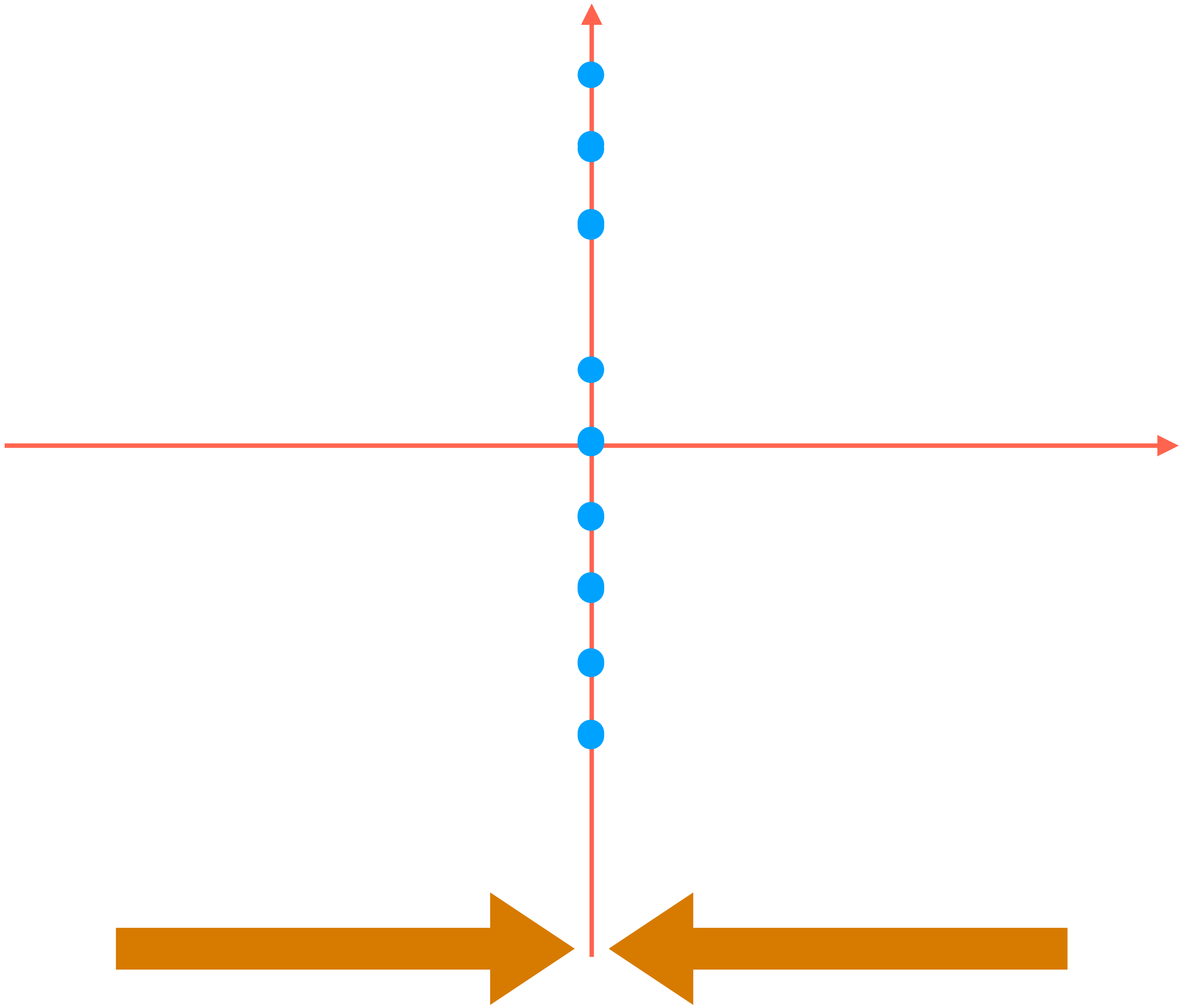
How to project 2D data down to 1D?



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

Principal Component Analysis (PCA)

How to project 2D data down to 1D?

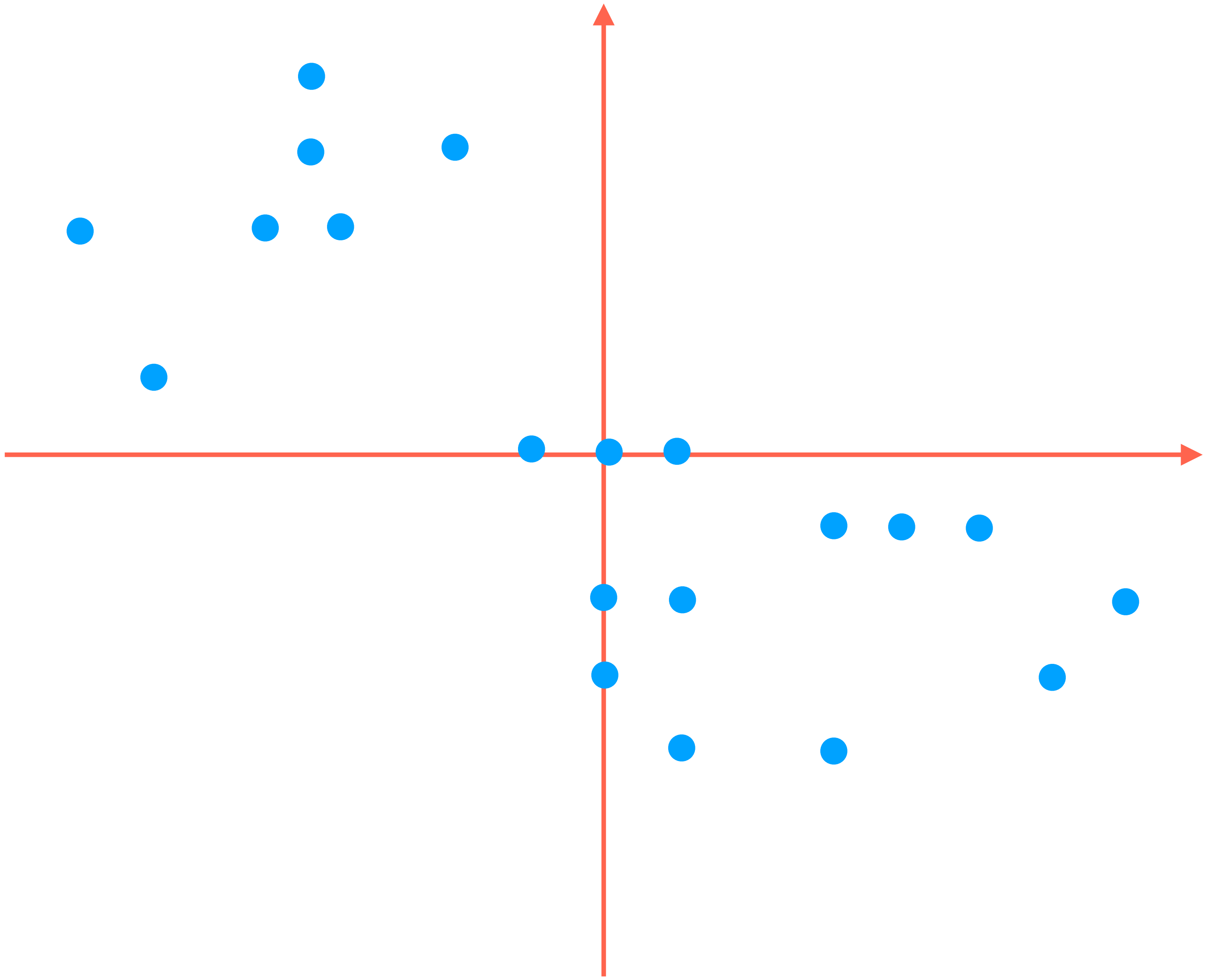


Simplest thing to try: flatten to one of the red axes

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

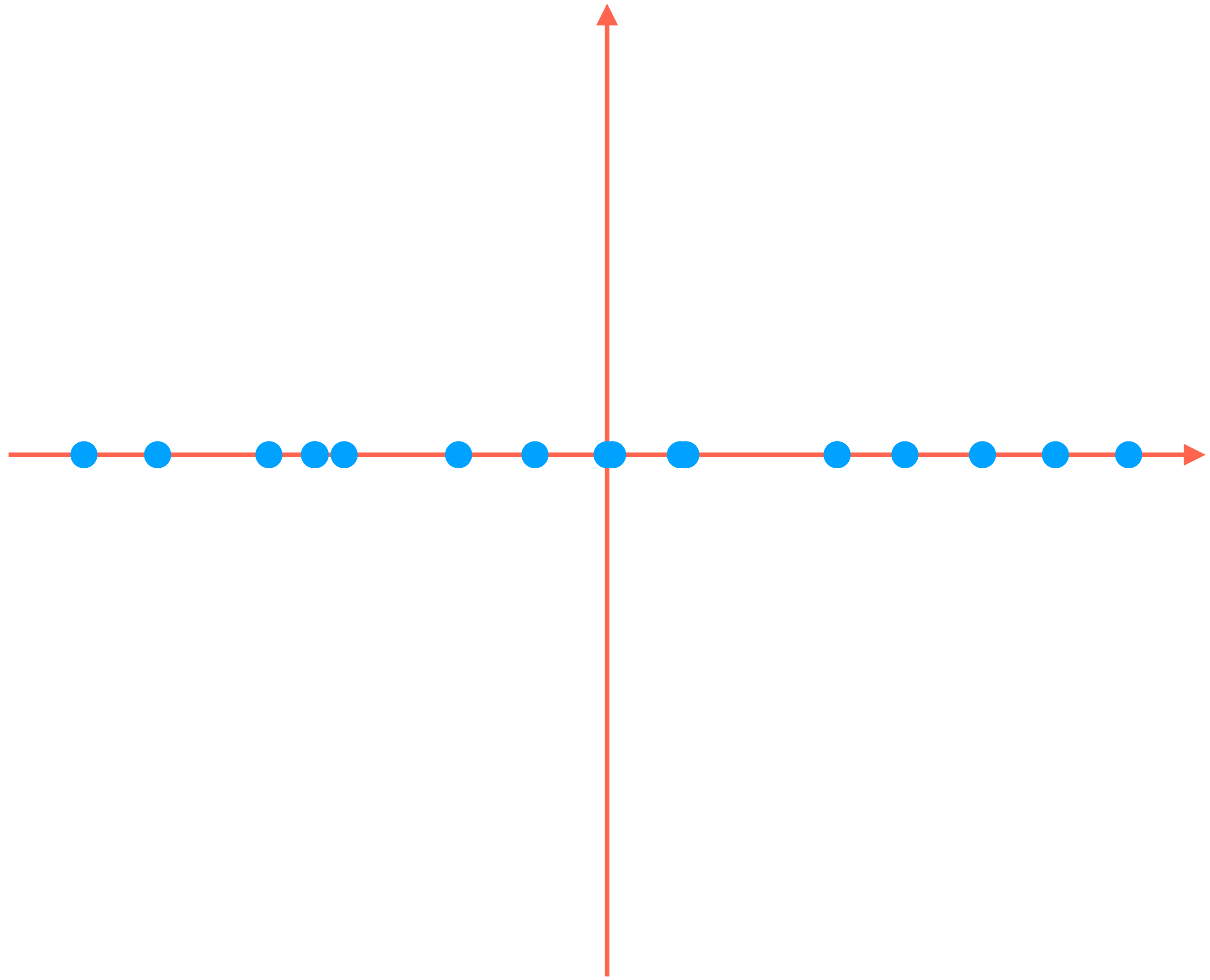
Principal Component Analysis (PCA)

How to project 2D data down to 1D?



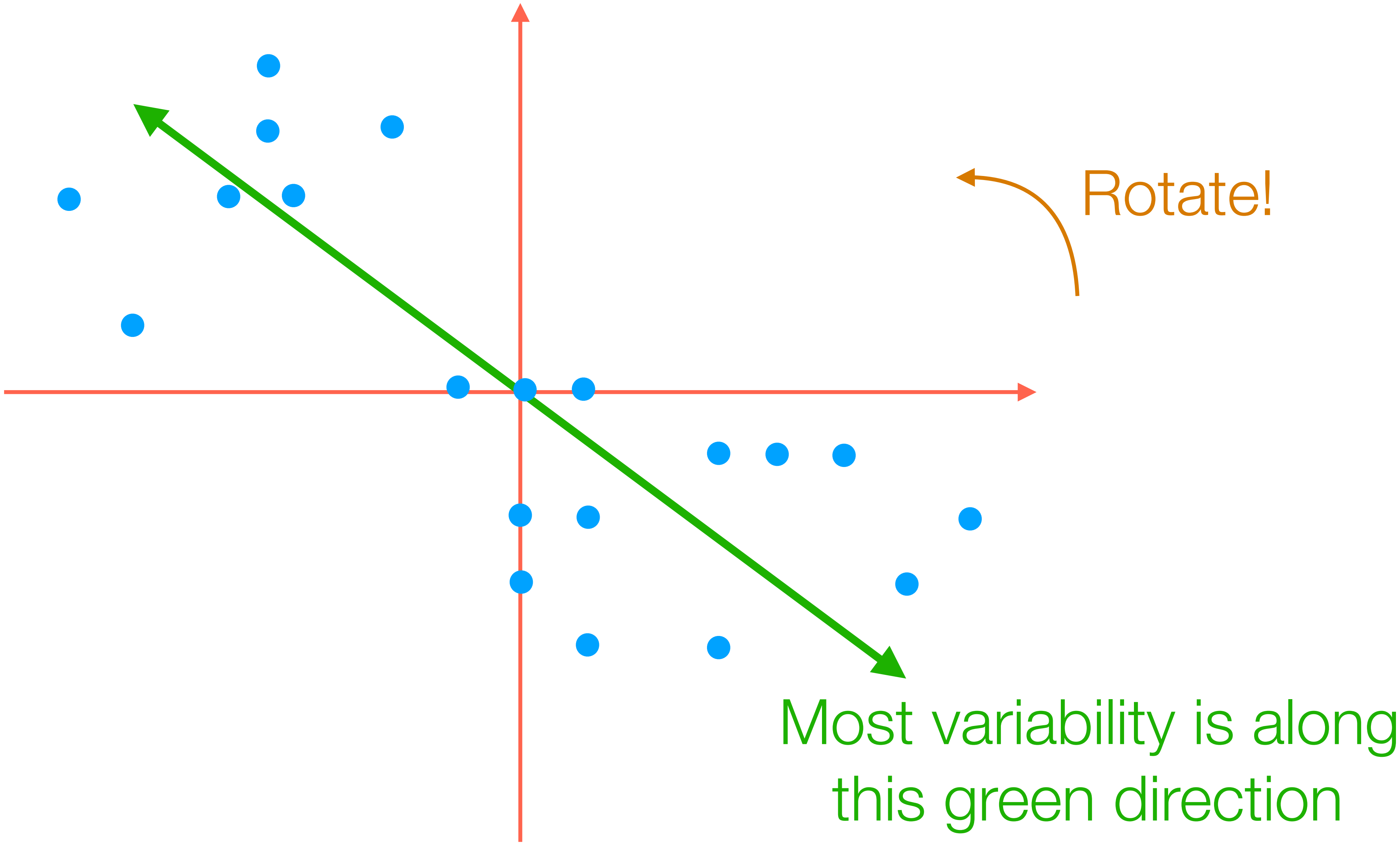
Principal Component Analysis (PCA)

How to project 2D data down to 1D?



Principal Component Analysis (PCA)

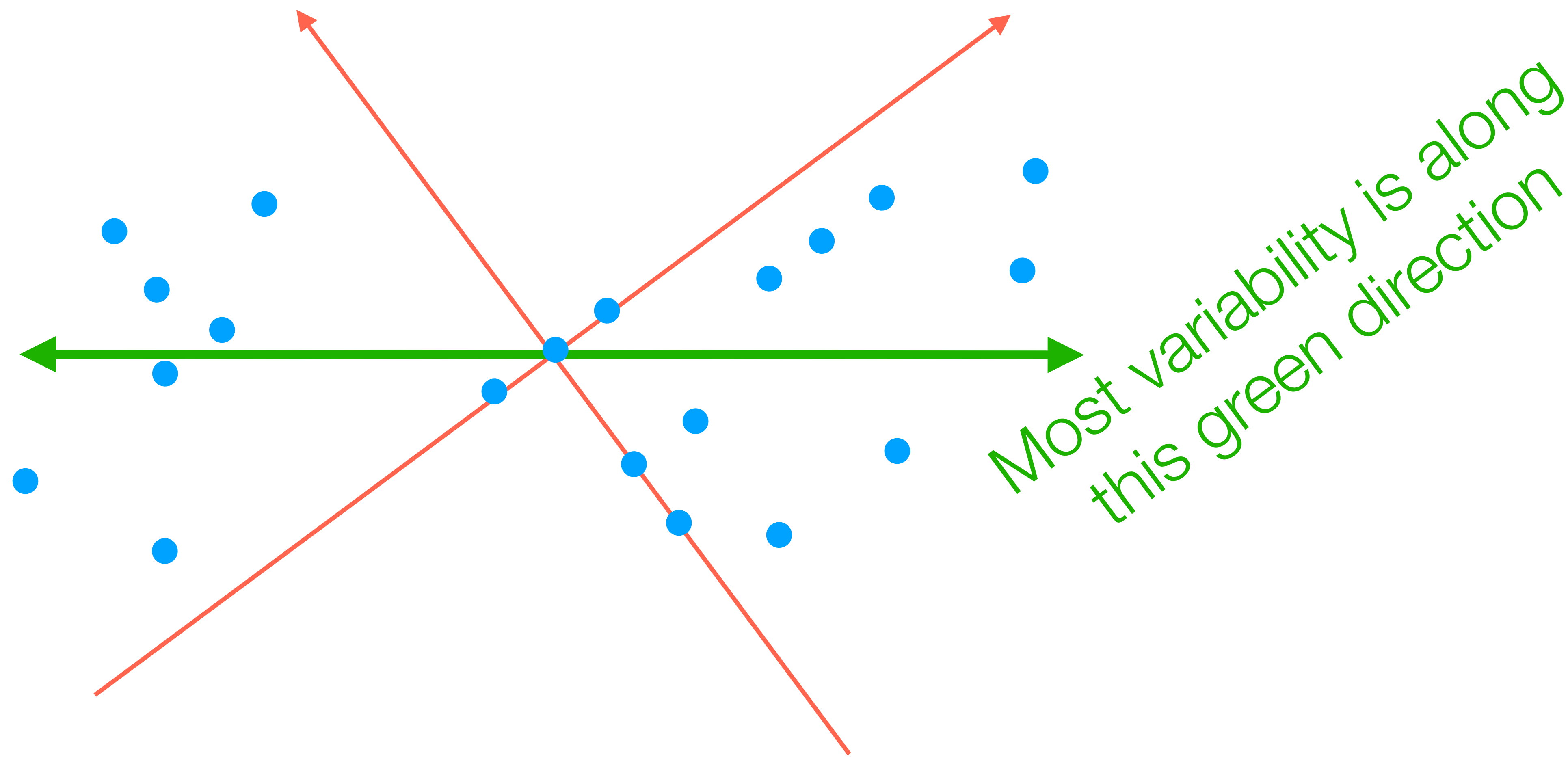
How to project 2D data down to 1D?



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

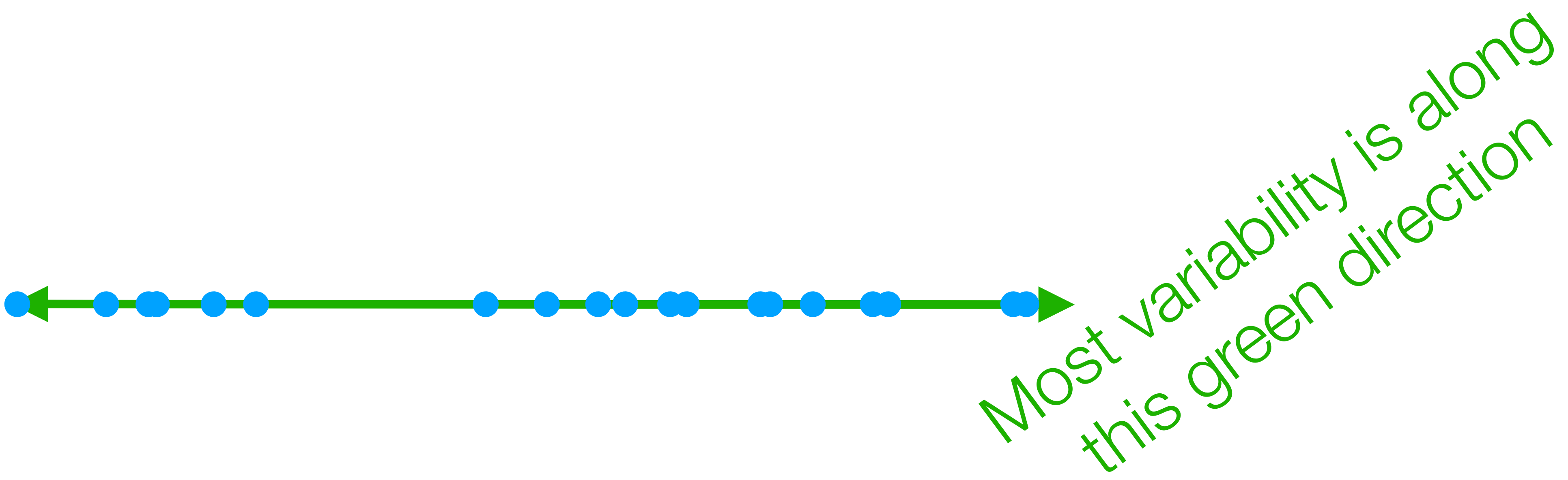
Principal Component Analysis (PCA)

How to project 2D data down to 1D?



Principal Component Analysis (PCA)

How to project 2D data down to 1D?

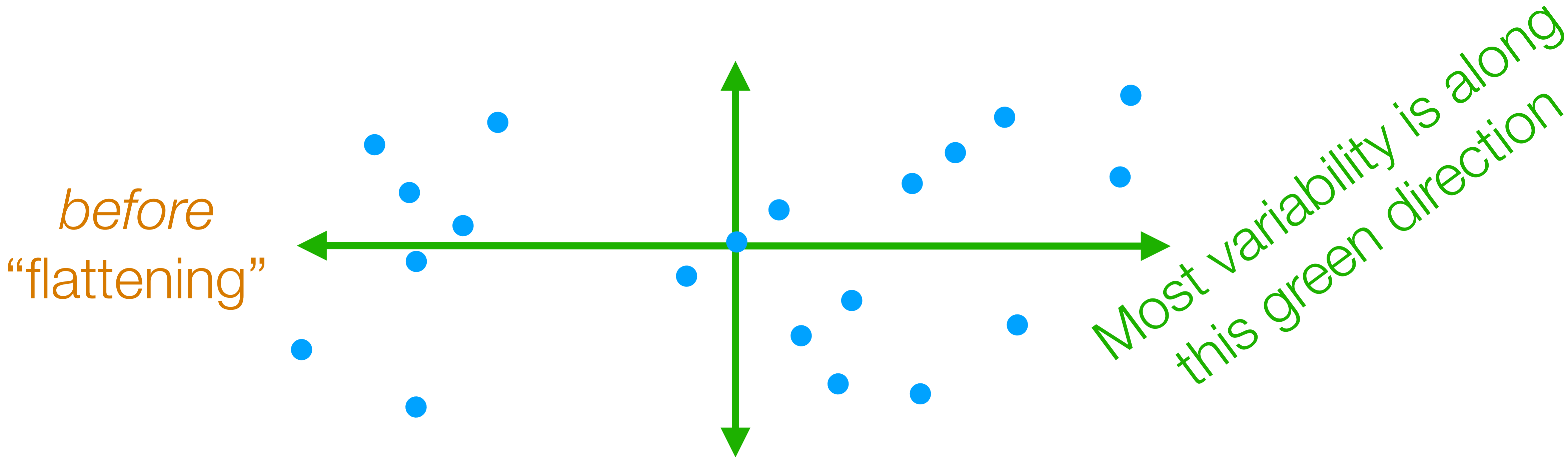


The idea of PCA actually works for 2D ➔ 2D as well
(and just involves rotating, and not “flattening” the data)

Principal Component Analysis (PCA)

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

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



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/

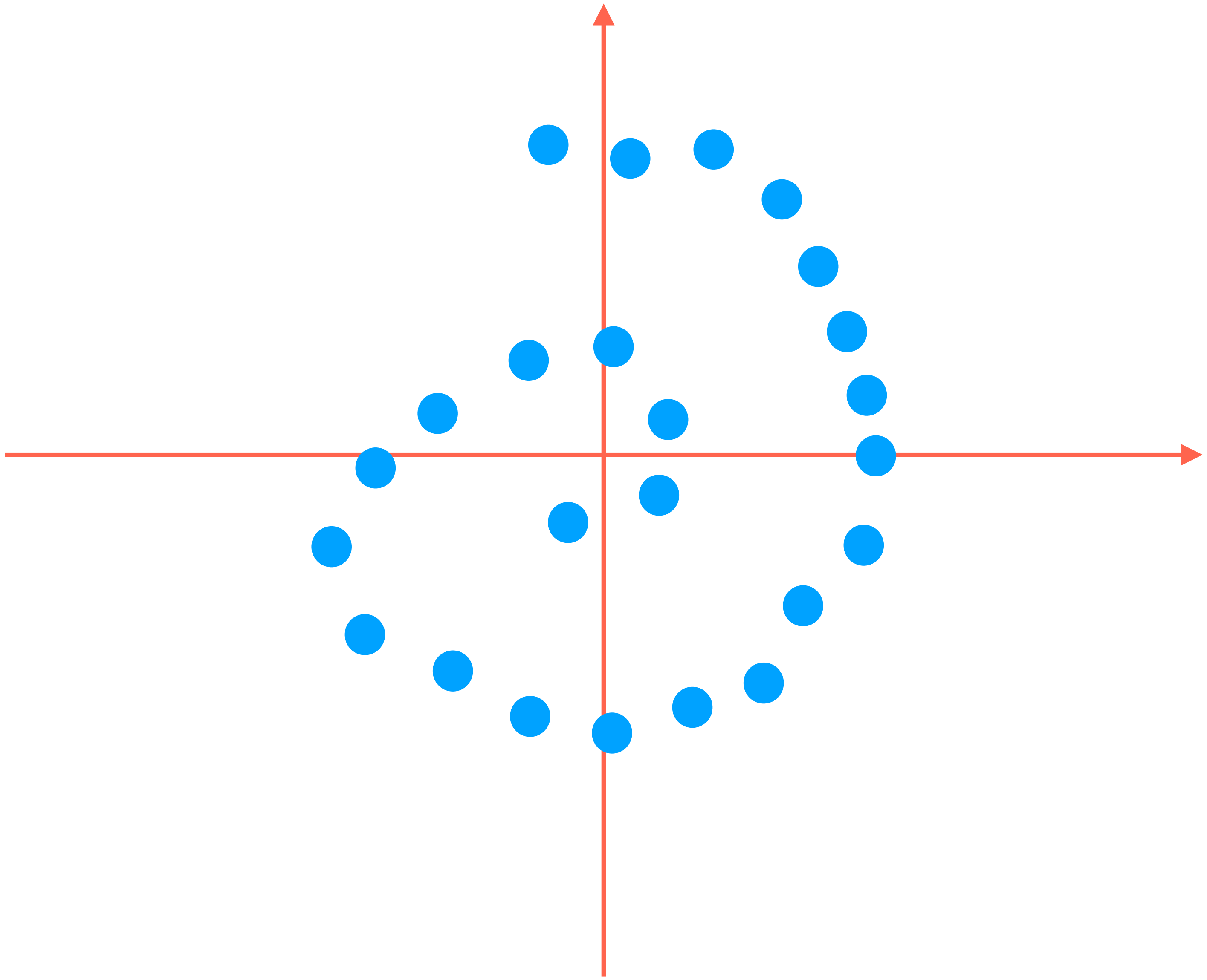
Principal Component Analysis (PCA)

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!

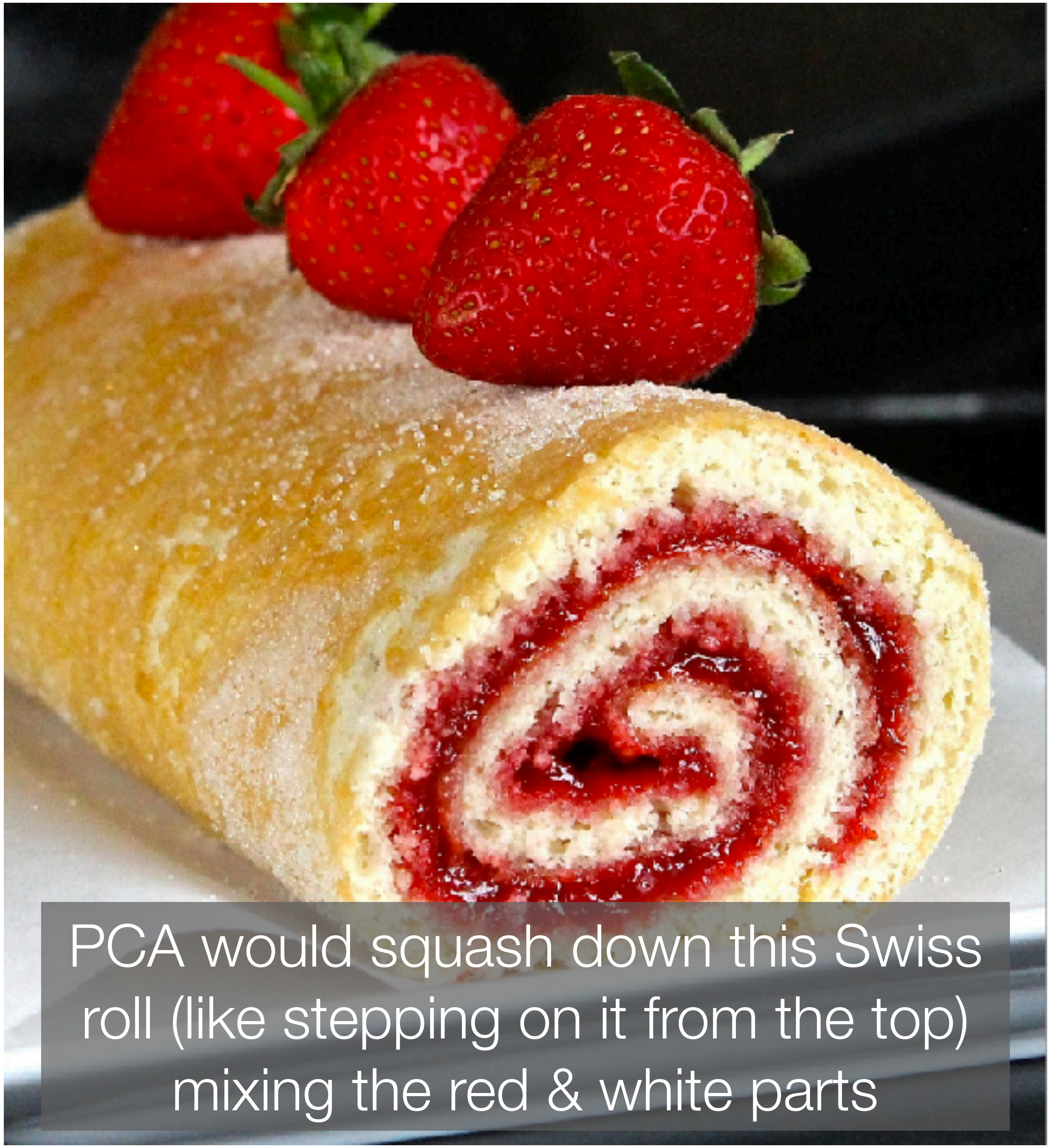
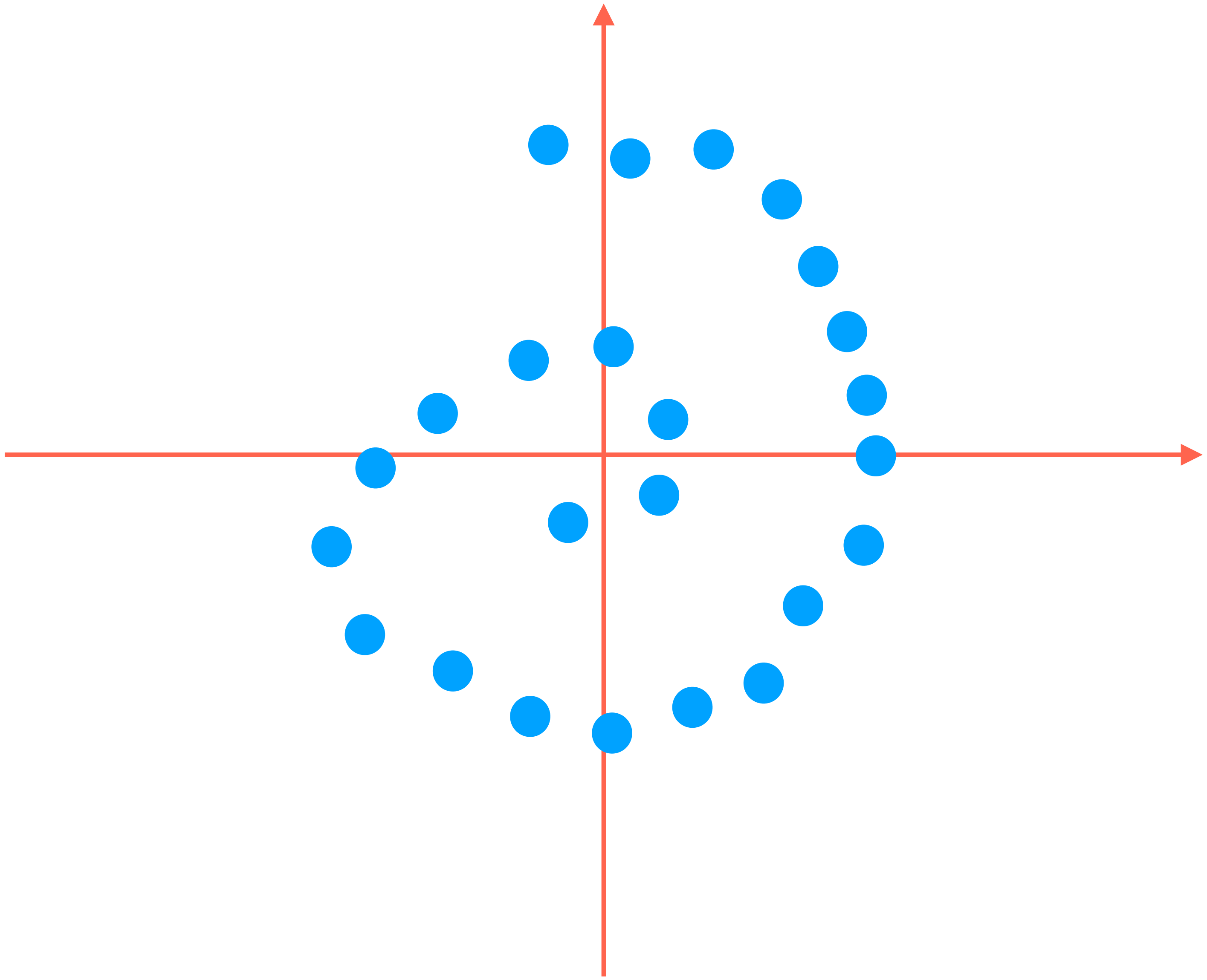
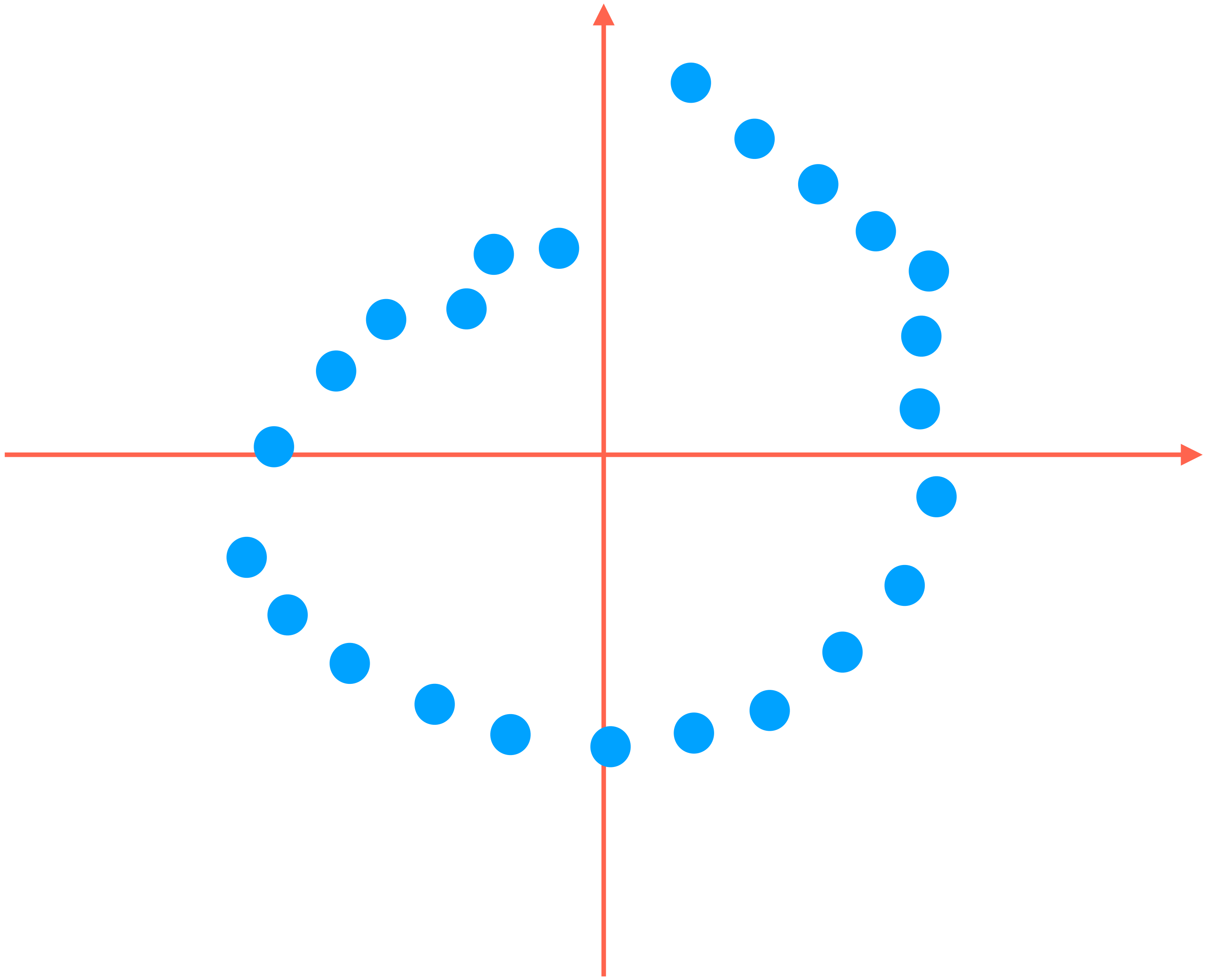


Image source: http://4.bp.blogspot.com/-USQEgoh1jCU/VfncdNOETcI/AAAAAAAAGp8/Hea8UtE_1c0/s1600/Blog%2B1%2BIMG_1821.jpg

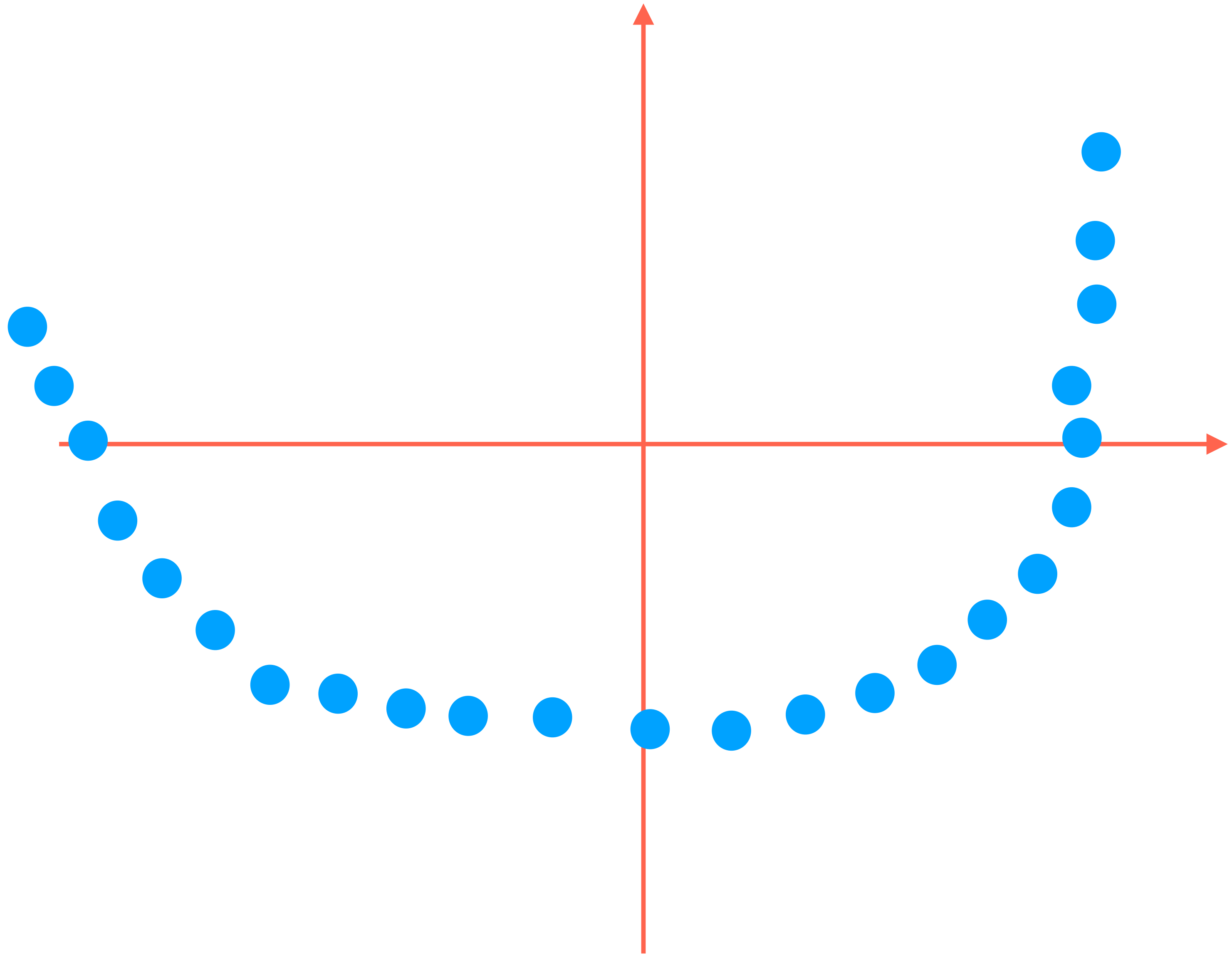
2D Swiss Roll



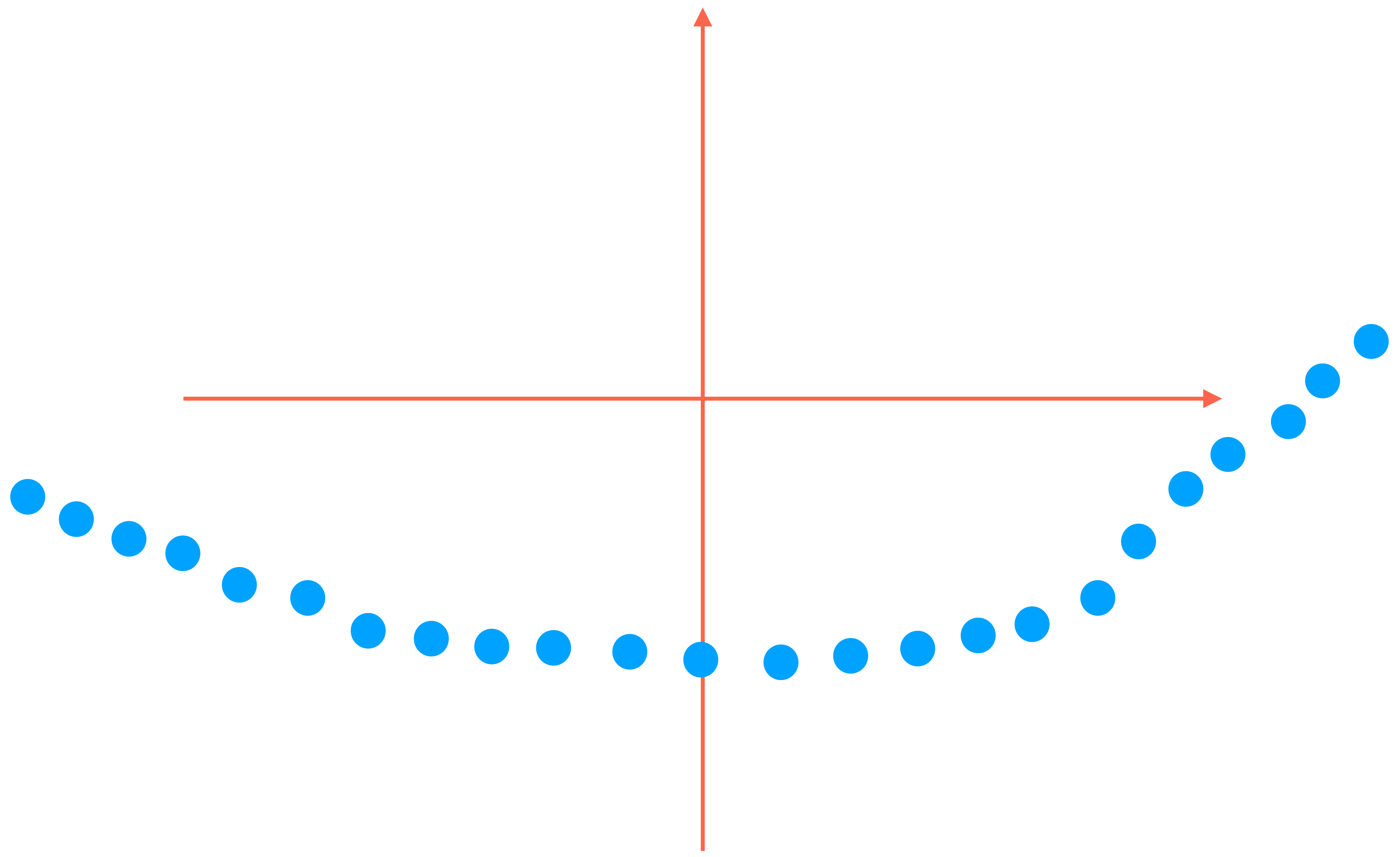
2D Swiss Roll



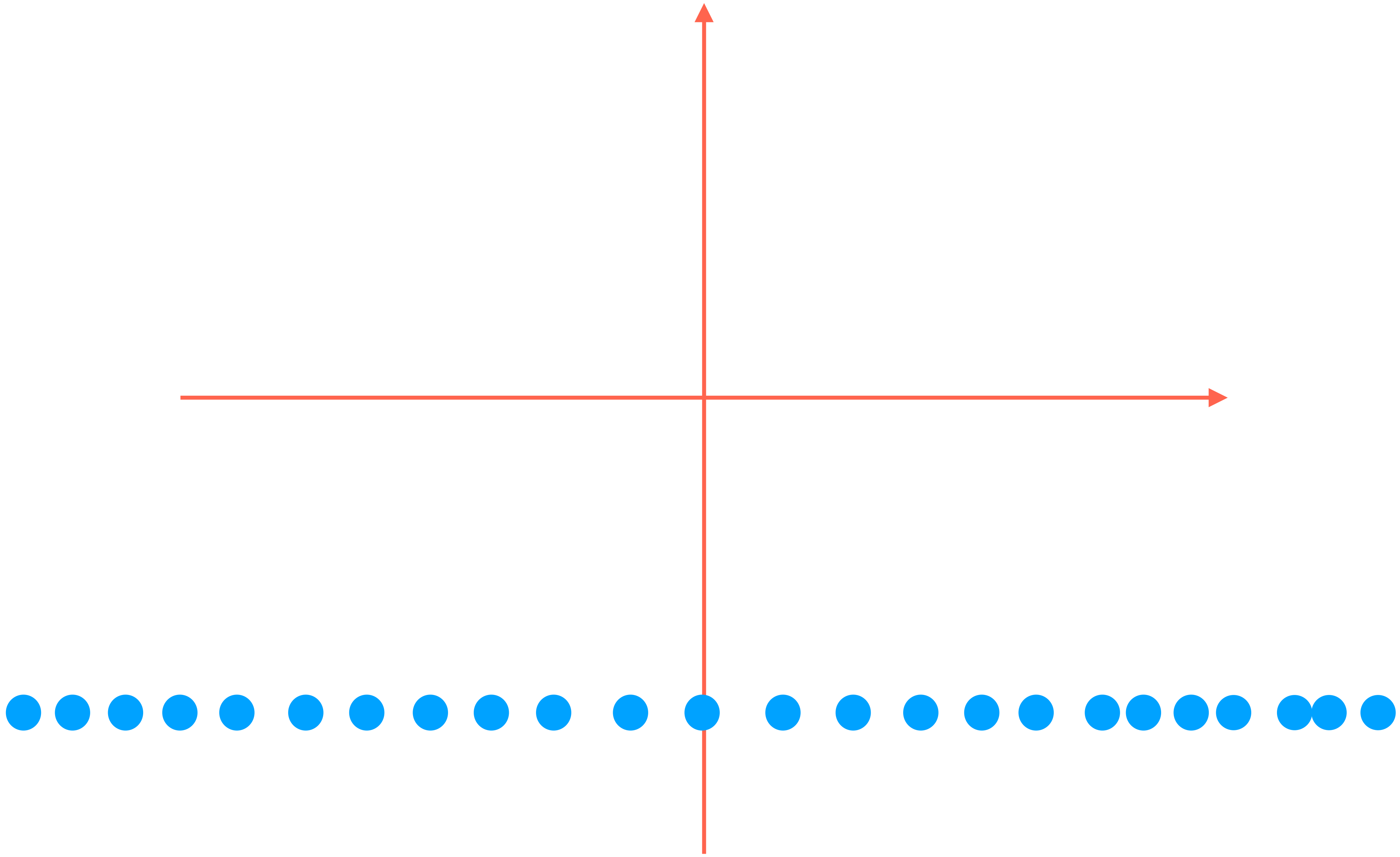
2D Swiss Roll



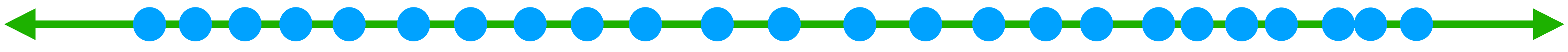
2D Swiss Roll



2D Swiss Roll



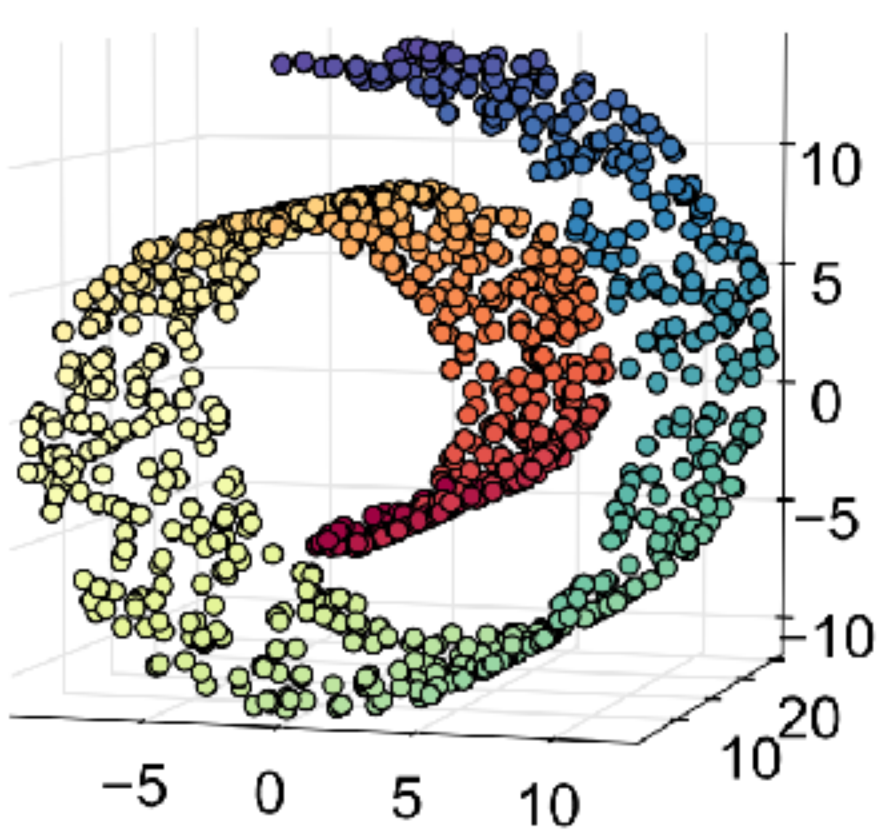
2D Swiss Roll



This is the desired result

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?

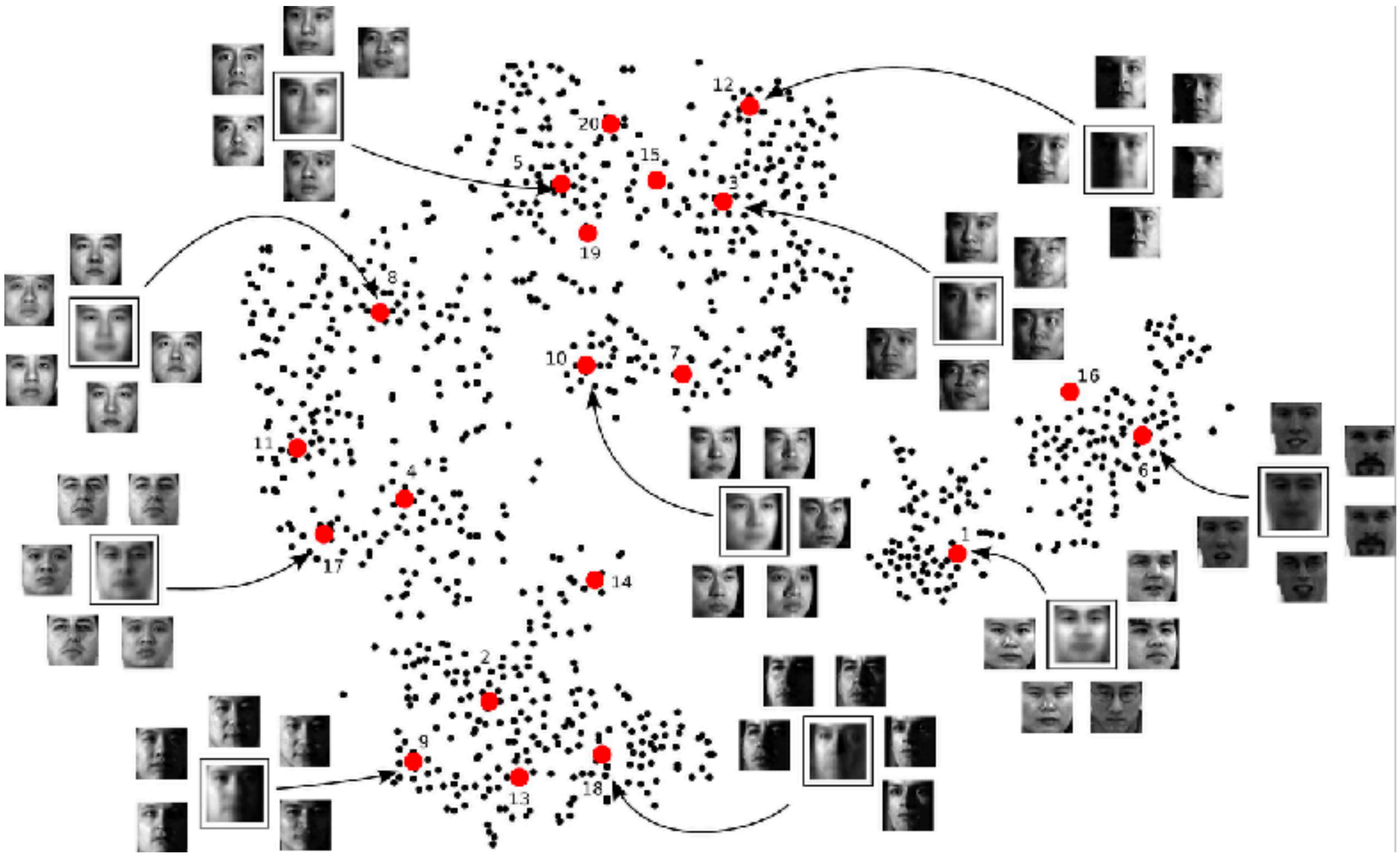


Image source: http://www.columbia.edu/~jwp2128/Images/faces.jpeg

Do Data Actually Live on Manifolds?

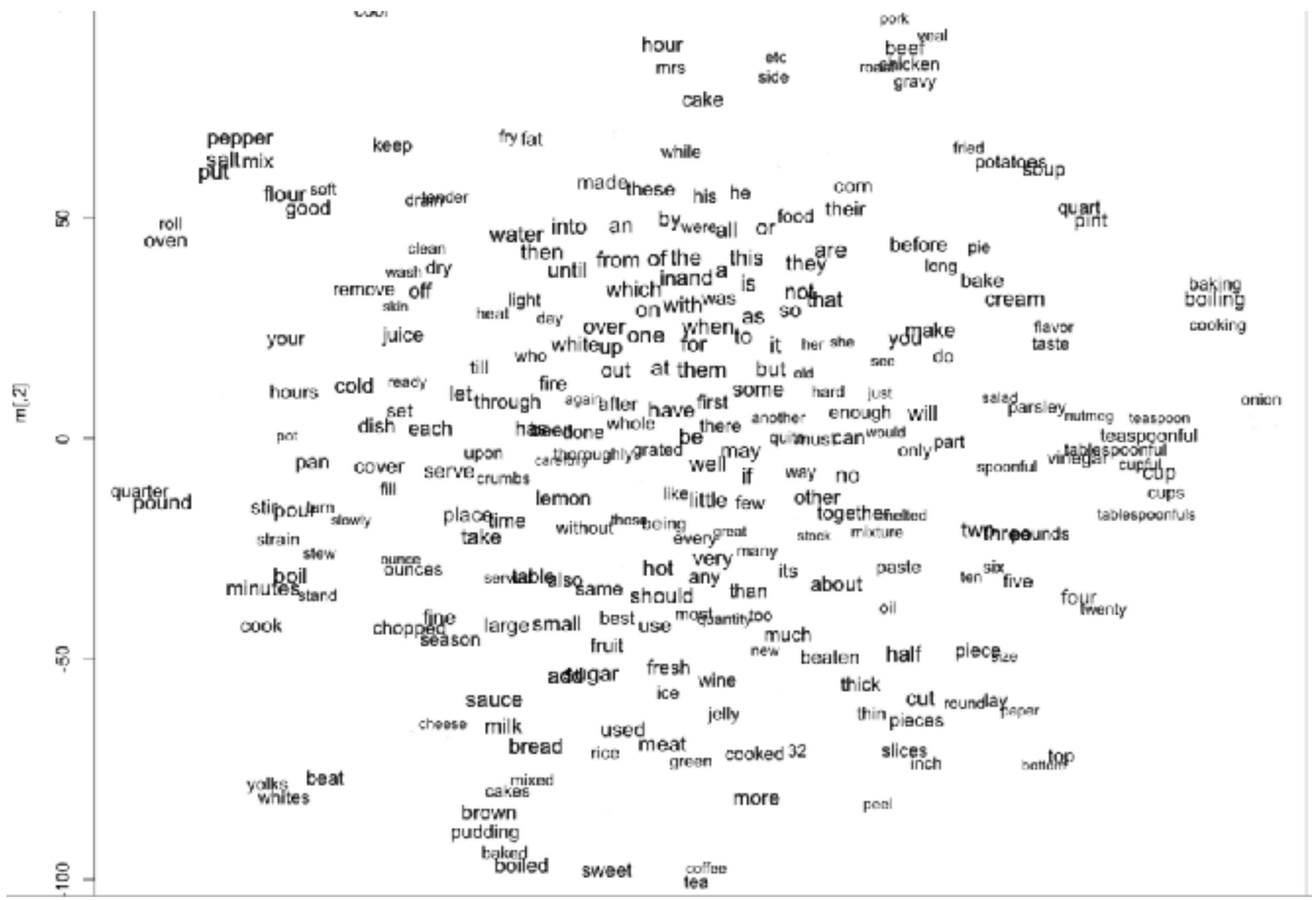
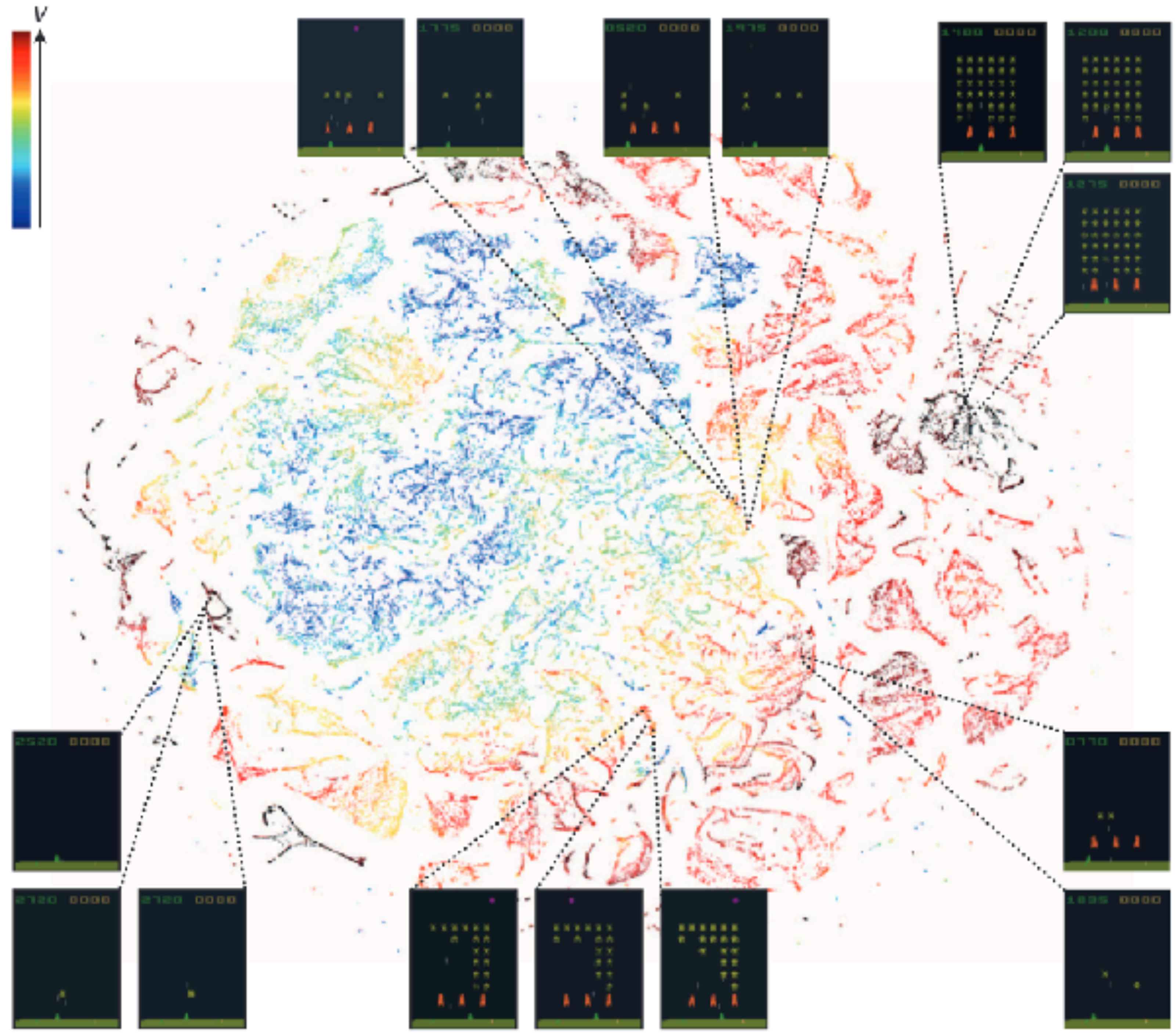


Image source: http://www.adityathakker.com/wp-content/uploads/2017/06/word-embeddings-994x675.png

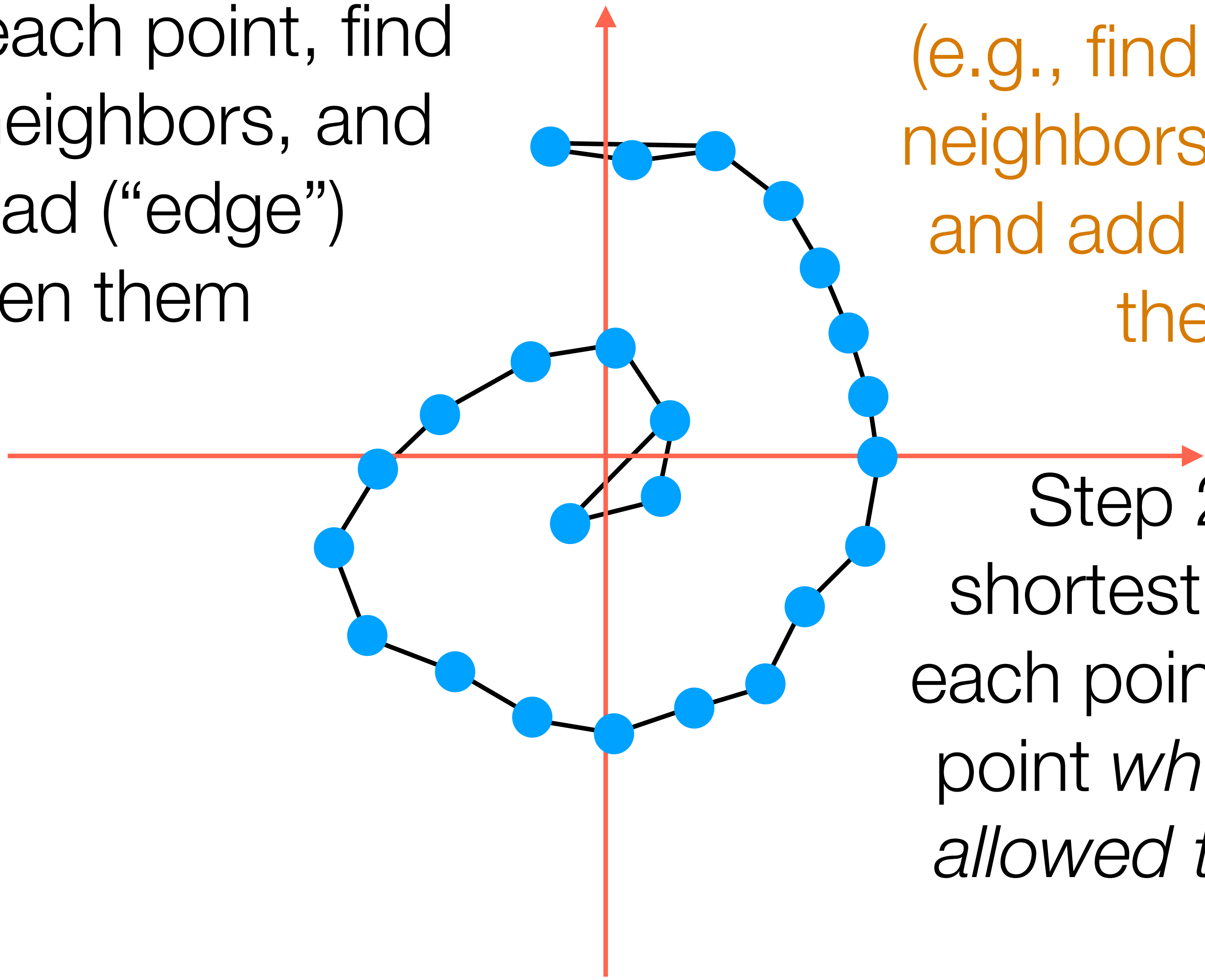
Do Data Actually Live on Manifolds?



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



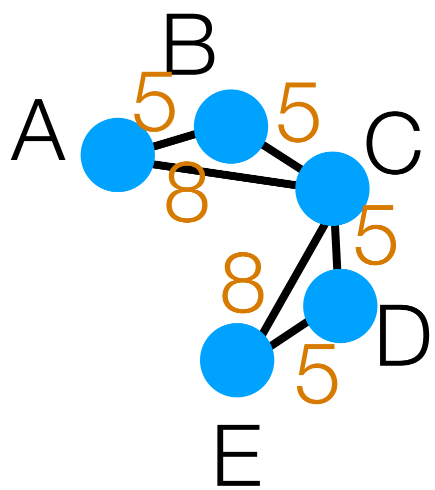
(e.g., find closest 2 neighbors per point and add edges to them)

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

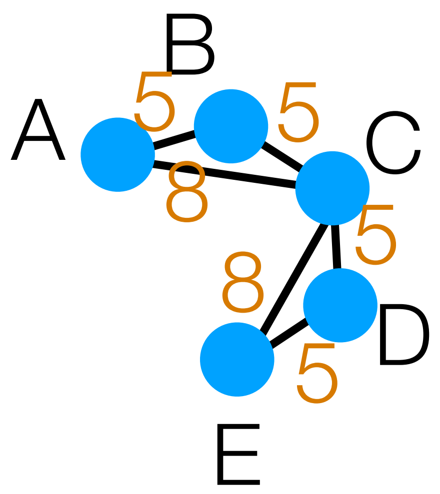
Build "symmetric 2-NN" graph (add edges for each point to its 2 nearest neighbors)

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

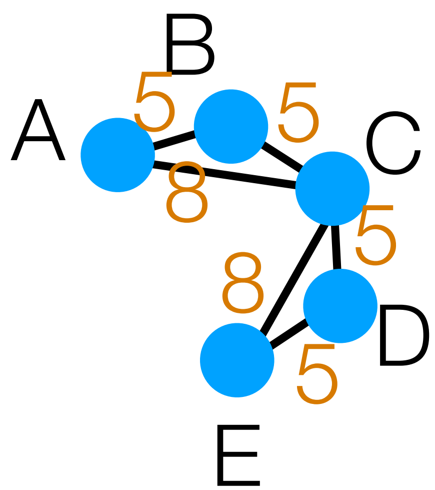
Build "symmetric 2-NN" graph (add edges for each point to its 2 nearest neighbors)

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

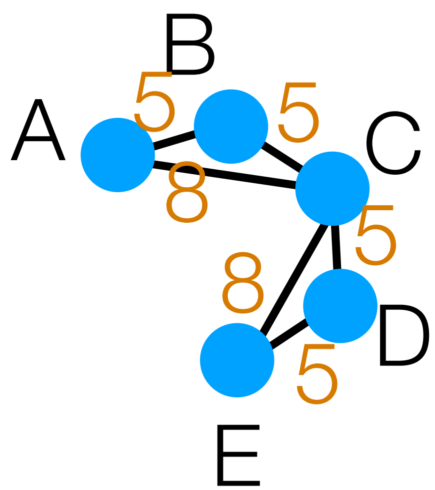
Build "symmetric 2-NN" graph (add edges for each point to its 2 nearest neighbors)

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

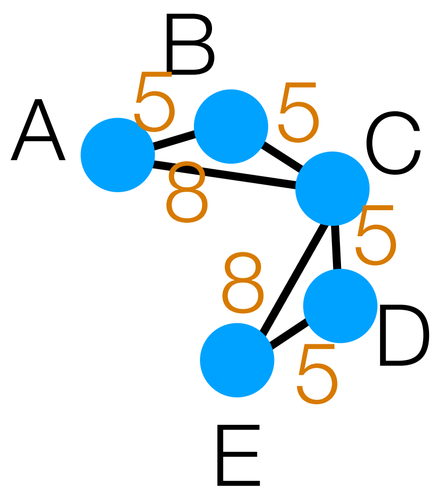
Build "symmetric 2-NN" graph (add edges for each point to its 2 nearest neighbors)

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

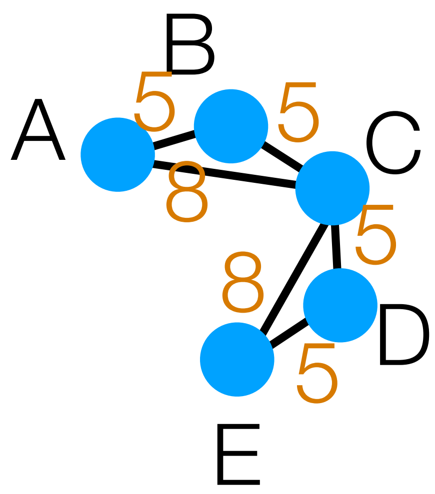
Build "symmetric 2-NN" graph (add edges for each point to its 2 nearest neighbors)

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

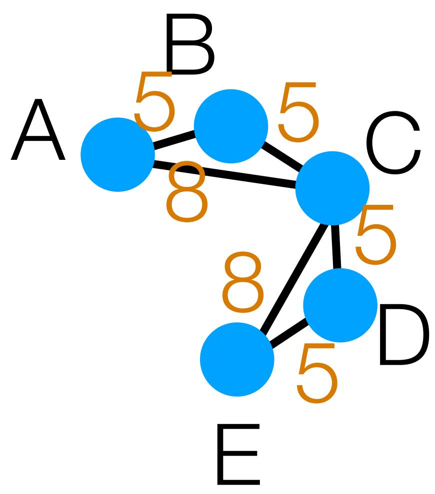
Build "symmetric 2-NN" graph (add edges for each point to its 2 nearest neighbors)

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

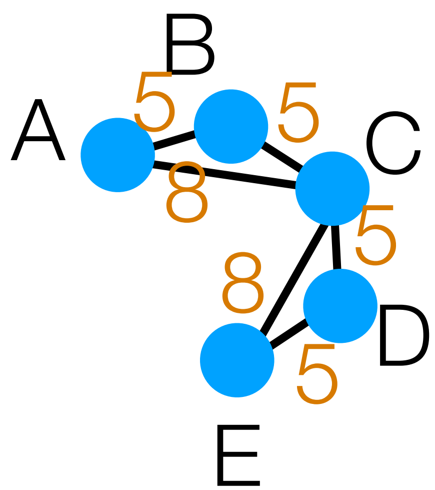
Build "symmetric 2-NN" graph (add edges for each point to its 2 nearest neighbors)

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

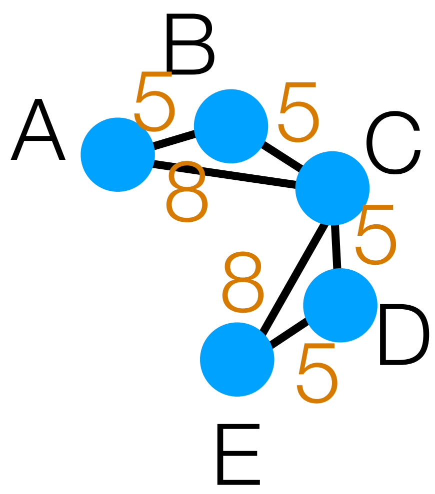
Build "symmetric 2-NN" graph (add edges for each point to its 2 nearest neighbors)

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

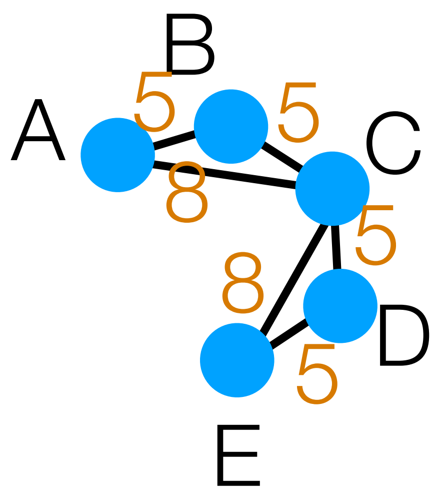
Build "symmetric 2-NN" graph (add edges for each point to its 2 nearest neighbors)

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

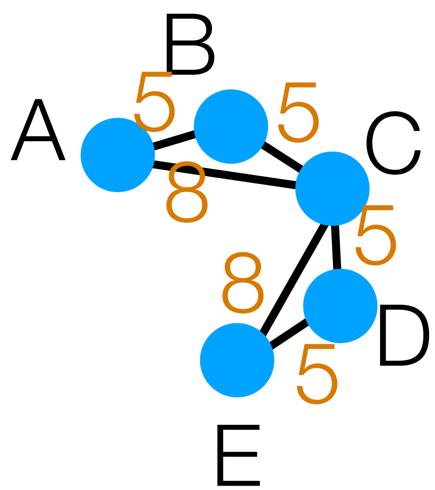
Build "symmetric 2-NN" graph (add edges for each point to its 2 nearest neighbors)

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

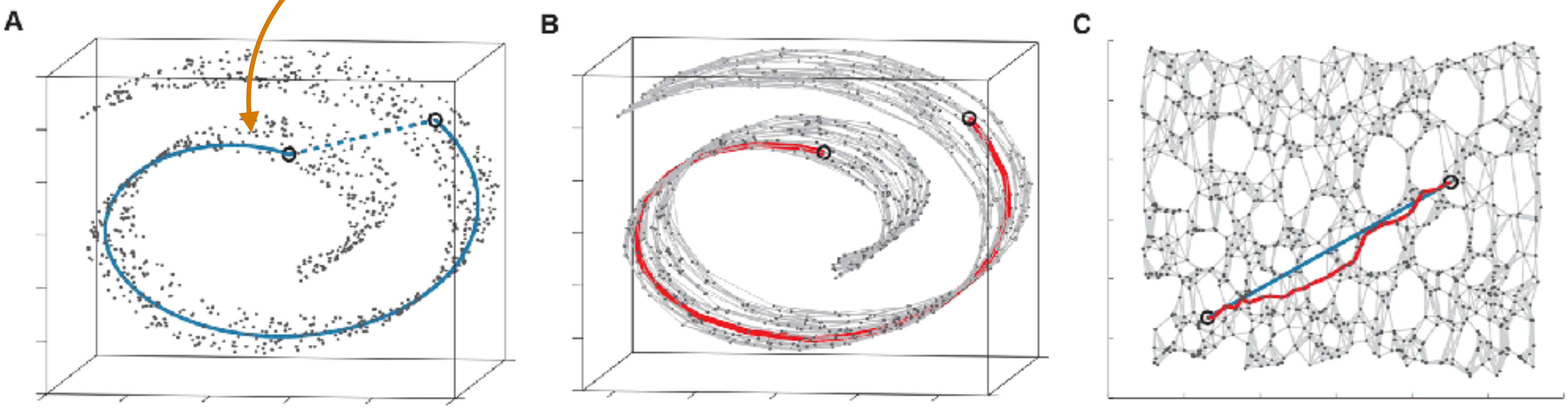
The solution is not unique!

Isomap Calculation Example

Demo

3D Swiss Roll Example

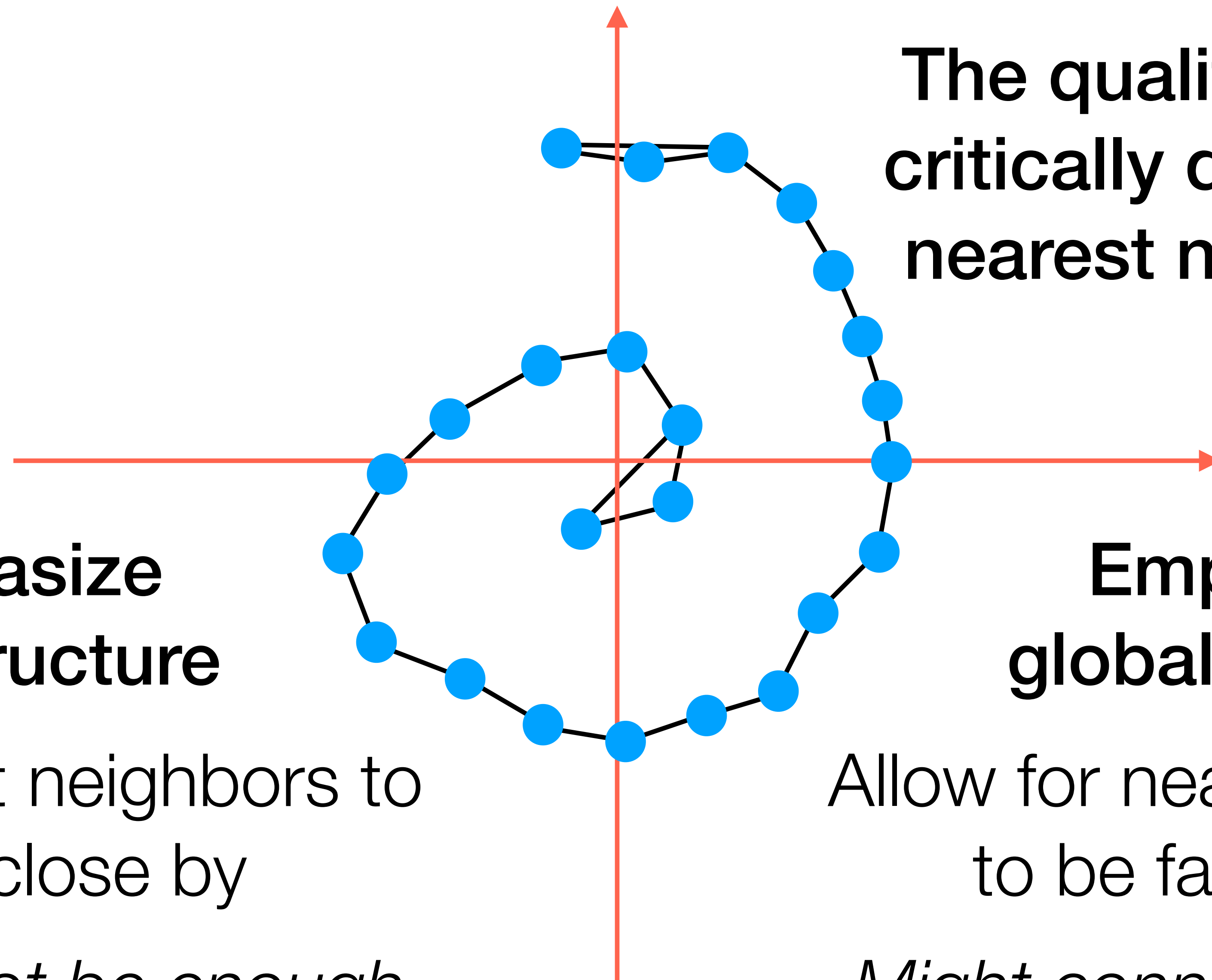
Key idea: true distance on manifold is the blue line



We’re approximating the blue line with the red line
(poor choice of # nearest neighbors can make approximation bad)

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

Emphasize local structure

Ask for nearest neighbors to be really close by

There might not be enough edges

Emphasize global structure

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