Spatial Data: Dimensionality Reduction

CS444
Techniques, Lecture 3

In this subfield, we think of a data point as a vector in R^n

(what could possibly go wrong?)

“Linear” dimensionality reduction:

Reduction is achieved by is a single matrix for every point.

Regular Scatterplots

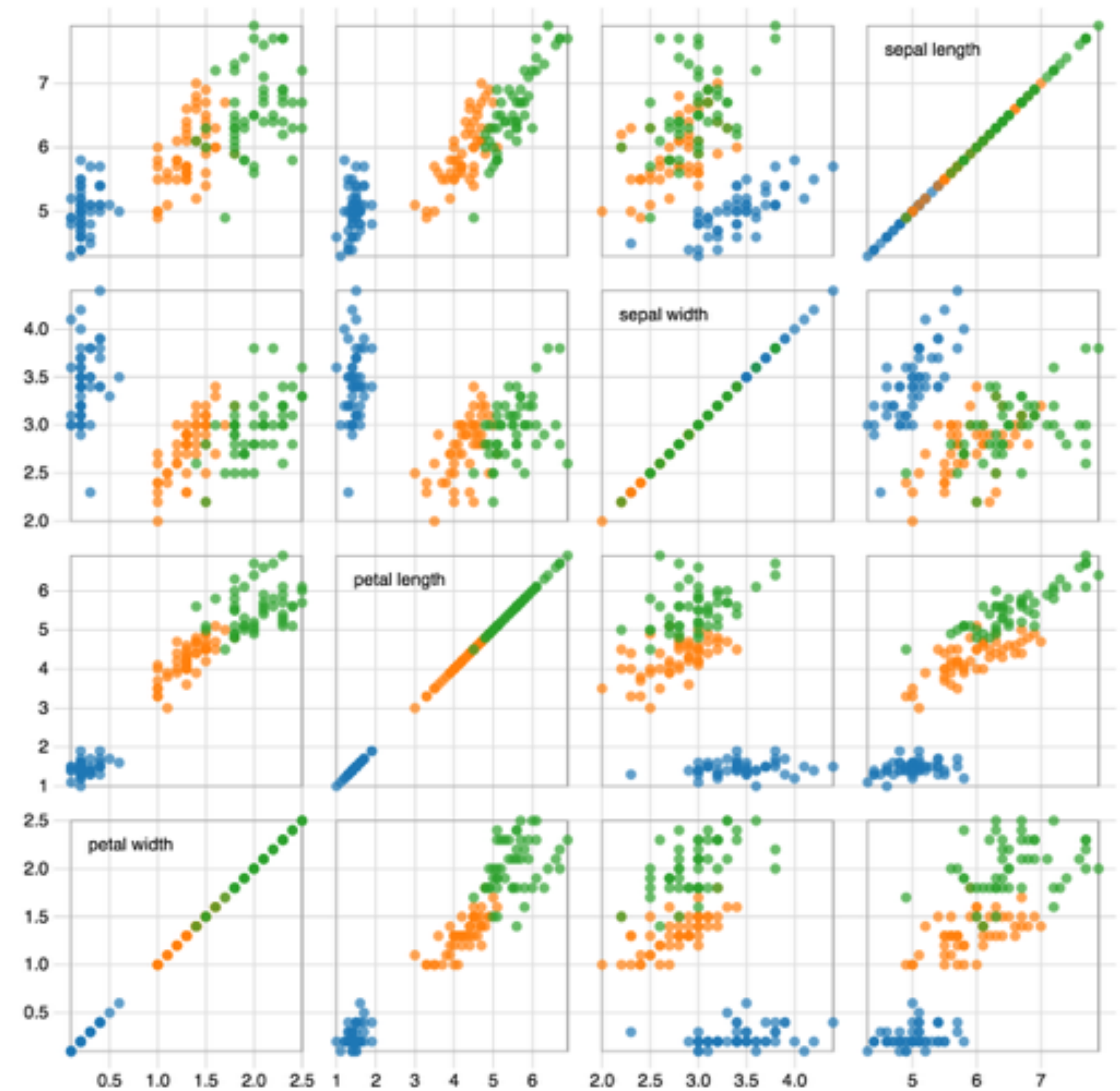
- Every data point is a vector:

$$\begin{bmatrix} v_0 \\ v_1 \\ v_2 \\ v_3 \end{bmatrix}$$

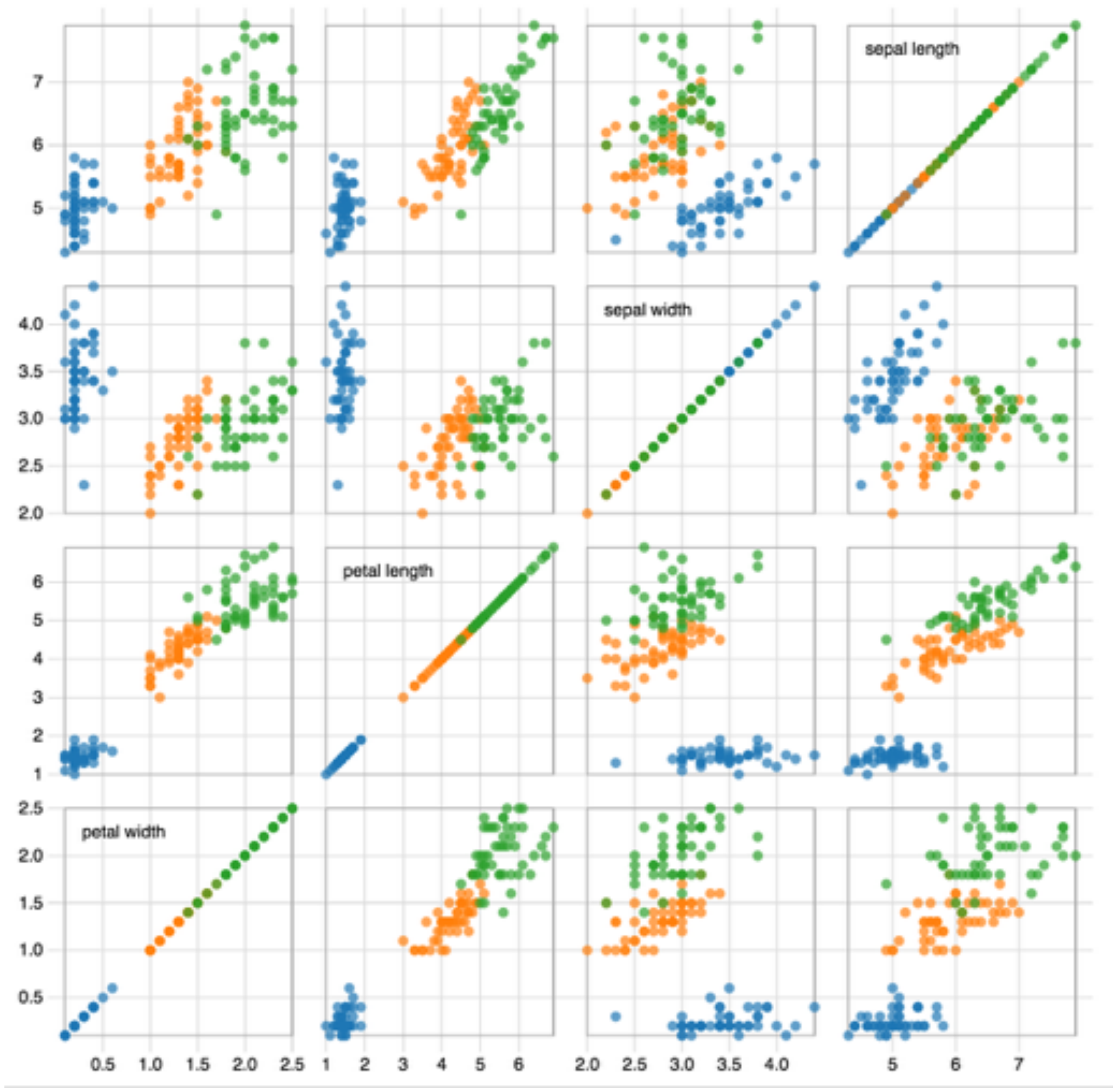
- Every scatterplot is produced by a very simple matrix:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

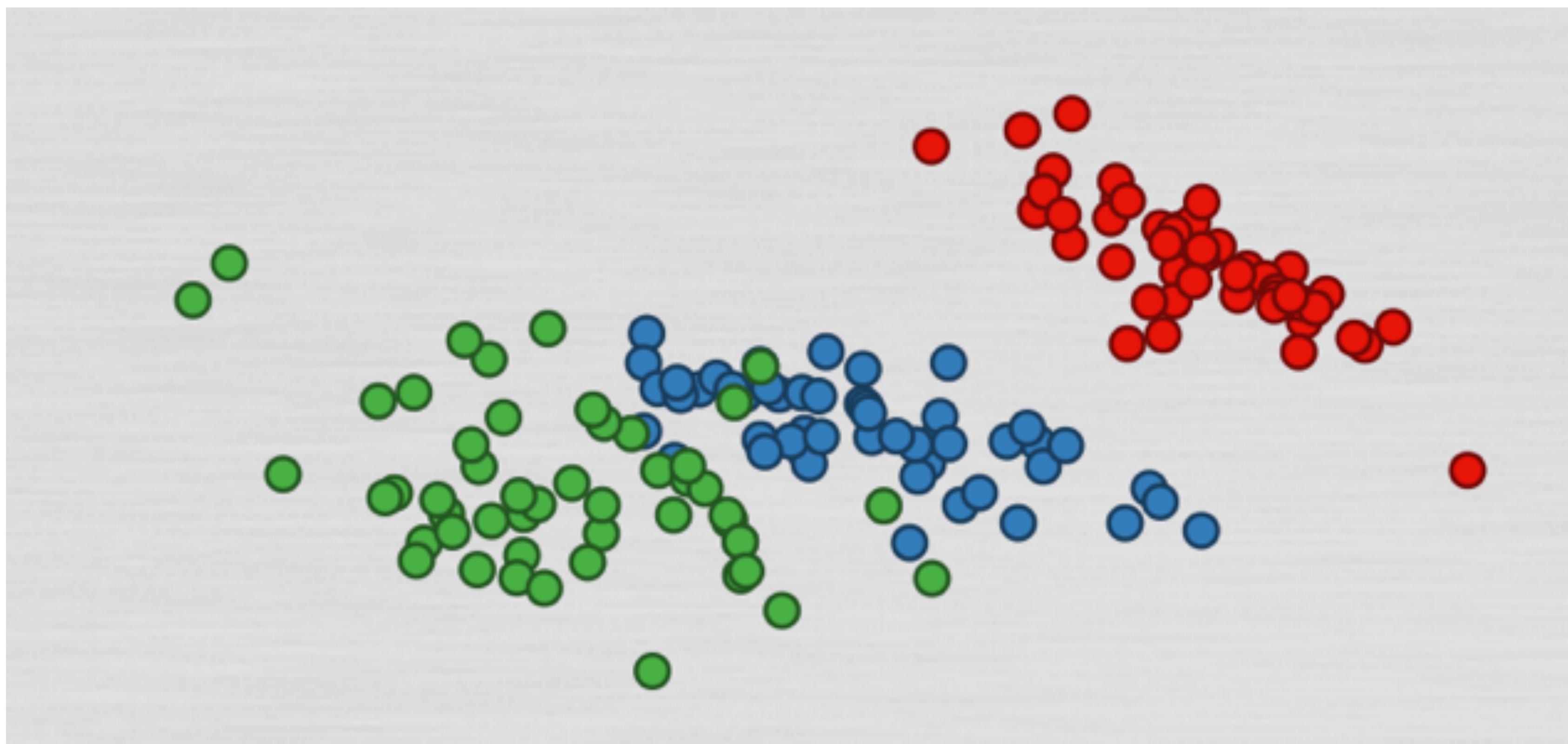
$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$



What about other matrices?



Grand Tour (Asimov, 1985)



http://cscheid.github.io/lux/demos/tour/tour.html

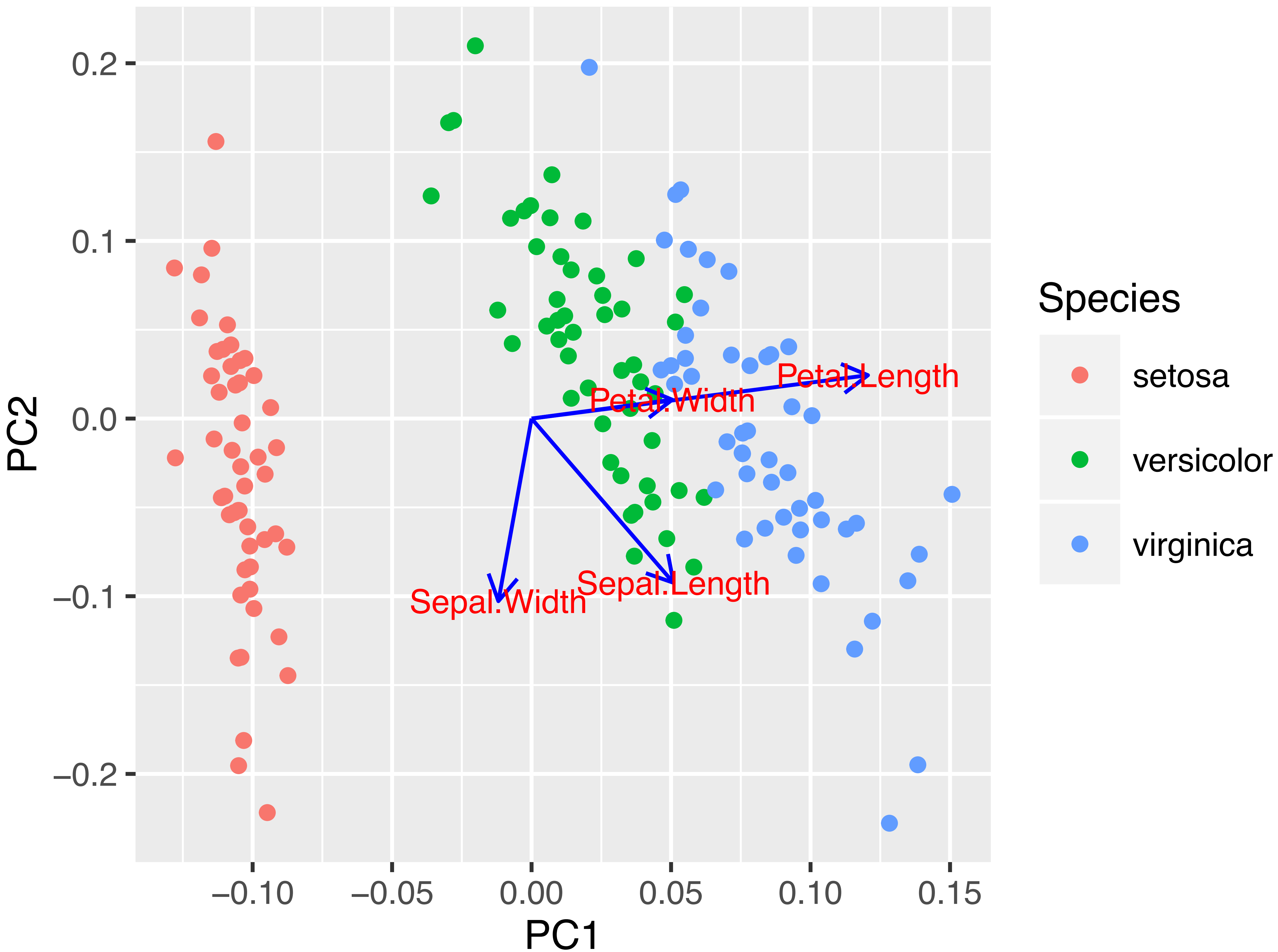
Is there a best matrix?

How do we think about that?

Linear Algebra review

- Vectors
- Inner Products
  - Lengths
  - Angles
- Bases
- Linear Transformations and Eigenvectors

Principal Component Analysis



Principal Component Analysis

- Algorithm:
  - Given data set as matrix X in R^(d x n),
  - Center matrix: $\tilde{X} = X(I - \frac{\vec{1}}{n}\vec{1}^T) = XH$
  - Compute eigendecomposition of $\tilde{X}^T \tilde{X}$
    - $\tilde{X}^T \tilde{X} = U \Sigma U^T$
  - The principal components are the first few rows of

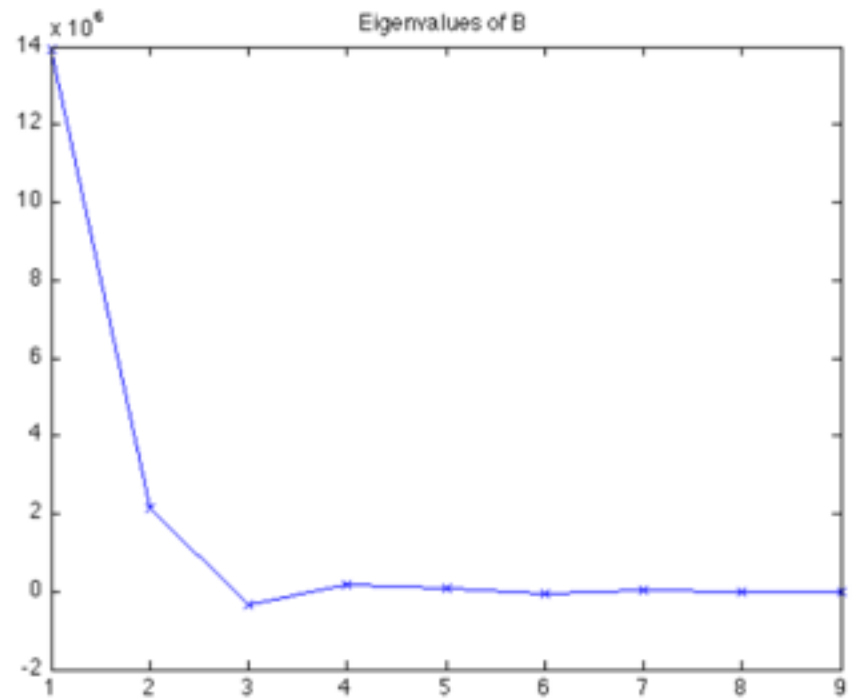
    $U\Sigma^{1/2}$

What if we don’t have coordinates, but distances?

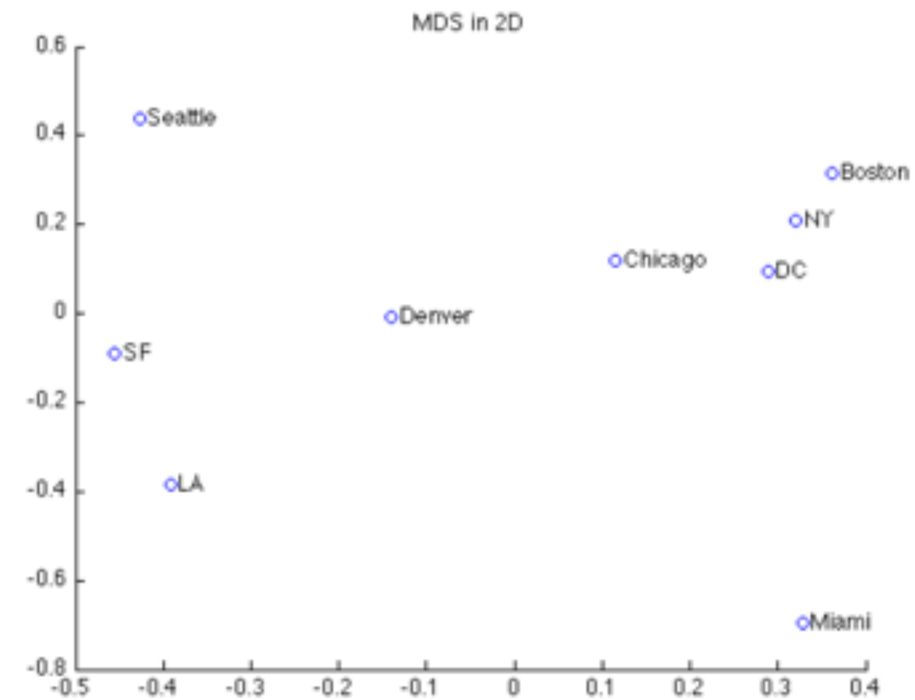
“Classical” Multidimensional Scaling

		1 BOST	2 NY	3 DC	4 MIAM	5 CHIC	6 SEAT	7 SF	8 LA	9 DENV
1	BOSTON	0	206	429	1504	963	2976	3095	2979	1949
2	NY	206	0	233	1308	802	2815	2934	2786	1771
3	DC	429	233	0	1075	671	2684	2799	2631	1616
4	MIAMI	1504	1308	1075	0	1329	3273	3053	2687	2037
5	CHICAGO	963	802	671	1329	0	2013	2142	2054	996
6	SEATTLE	2976	2815	2684	3273	2013	0	808	1131	1307
7	SF	3095	2934	2799	3053	2142	808	0	379	1235
8	LA	2979	2786	2631	2687	2054	1131	379	0	1059
9	DENVER	1949	1771	1616	2037	996	1307	1235	1059	0

(a)



(b)



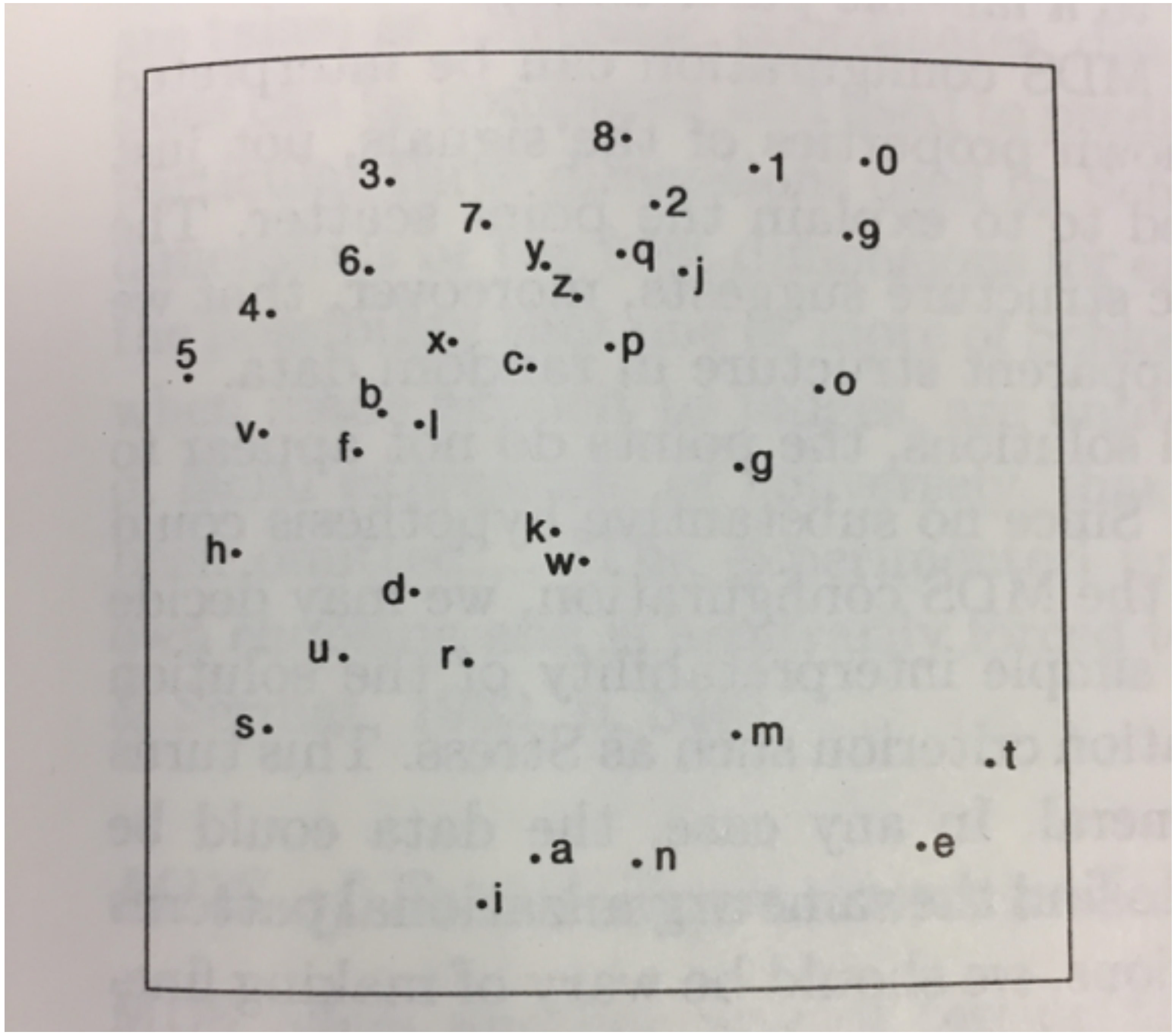
(c)

http://www.math.pku.edu.cn/teachers/yaoy/Fall2011/lecture11.pdf

TABLE 4.2. Confusion percentages between Morse code signals (Rothkopf, 1957); decimal points omitted.

Morse code		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	1	2	3	4	5	6	7	8	9	0	
·—	A	92	4	6	13	3	14	10	13	46	5	22	3	25	34	6	6	9	35	23	6	37	13	17	12	7	3	2	7	5	5	8	6	5	6	2	3	A
—···	B	5	84	37	31	5	28	17	21	5	19	34	40	6	10	12	22	25	16	18	2	18	34	8	84	30	42	12	17	14	40	32	74	43	17	4	4	B
—·—·	C	4	38	87	17	4	29	13	7	11	19	24	35	14	3	9	51	34	24	14	6	6	11	14	32	82	38	13	15	31	14	10	30	28	24	18	12	C
—··	D	8	62	17	88	7	23	40	36	9	13	81	56	8	7	9	27	9	45	29	6	17	20	27	40	15	33	3	9	6	11	9	19	8	10	5	6	D
·	E	6	13	14	6	97	2	4	4	17	1	5	6	4	4	5	1	5	10	7	67	3	3	2	5	6	5	4	3	5	3	5	2	4	2	3	3	E
··—·	F	4	51	33	19	2	90	10	29	5	33	16	50	7	6	10	42	12	35	14	2	21	27	25	19	27	13	8	16	47	25	26	24	21	5	5	5	F
——·	G	9	18	27	38	1	14	90	6	5	22	33	16	14	13	62	52	23	21	5	3	15	14	32	21	23	39	15	14	5	10	4	10	17	23	20	11	G
····	H	3	45	23	25	9	32	8	87	10	10	9	29	5	8	8	14	8	17	37	4	36	59	9	33	14	11	3	9	15	43	70	35	17	4	3	3	H
··	I	64	7	7	13	10	8	6	12	93	3	5	16	13	30	7	3	5	19	35	16	10	5	8	2	5	7	2	5	8	9	6	8	5	2	4	5	I
·———	J	7	9	38	9	2	24	18	5	4	85	22	31	8	3	21	63	47	11	2	7	9	9	9	22	32	28	67	66	33	15	7	11	28	29	26	23	J
—·—	K	5	24	38	73	1	17	25	11	5	27	91	33	10	12	31	14	31	22	2	2	23	17	33	63	16	18	5	9	17	8	8	18	14	13	5	6	K
·—··	L	2	69	43	45	10	24	12	26	9	30	27	86	6	2	9	37	36	28	12	5	16	19	20	31	25	59	12	13	17	15	26	29	36	16	7	3	L
——	M	24	12	5	14	7	17	29	8	8	11	23	8	96	62	11	10	15	20	7	9	13	4	21	9	18	8	5	7	6	6	5	7	11	7	10	4	M
—·	N	31	4	13	30	8	12	10	16	13	3	16	8	59	93	5	9	5	28	12	10	16	4	12	4	16	11	5	2	3	4	4	6	2	2	10	2	N
———	O	7	7	20	6	5	9	76	7	2	39	26	10	4	8	86	37	35	10	3	4	11	14	25	35	27	27	19	17	7	7	6	18	14	11	20	12	O
·——·	P	5	22	33	12	5	36	22	12	3	78	14	46	5	6	21	83	43	23	9	4	12	19	19	19	41	30	34	44	24	11	15	17	24	23	25	13	P
——·—	Q	8	20	38	11	4	15	10	5	2	27	23	26	7	6	22	51	91	11	2	3	6	14	12	37	50	63	34	32	17	12	9	27	40	58	37	24	Q
·—·	R	13	14	16	23	5	34	26	15	7	12	21	33	14	12	12	29	8	87	16	2	23	23	62	14	12	13	7	10	13	4	7	12	7	9	1	2	R
···	S	17	24	5	30	11	26	5	59	16	3	13	10	5	17	6	6	3	18	96	9	56	24	12	10	6	7	8	2	2	15	28	9	5	5	5	2	S
—	T	13	10	1	5	46	3	6	6	14	6	14	7	6	5	6	11	4	4	7	96	8	5	4	2	2	6	5	5	3	3	3	8	7	6	14	6	T
··—	U	14	29	12	32	4	32	11	34	21	7	44	32	11	13	6	20	12	40	51	6	93	57	34	17	9	11	6	6	16	34	10	9	9	7	4	3	U
···—	V	5	17	24	16	9	29	6	39	5	11	26	43	4	1	9	17	10	17	11	6	32	92	17	57	35	10	10	14	28	79	44	36	25	10	1	5	V
·——	W	9	21	30	22	9	36	25	15	4	25	29	18	15	6	26	20	25	61	12	4	19	20	86	22	25	22	10	22	19	16	5	9	11	6	3	7	W
—··—	X	7	64	45	19	3	28	11	6	1	35	50	42	10	8	24	32	61	10	12	3	12	17	21	91	48	26	12	20	24	27	16	57	29	16	17	6	X
—·——	Y	9	23	62	15	4	26	22	9	1	30	12	14	5	6	14	30	52	5	7	4	6	13	21	44	86	23	26	44	40	15	11	26	22	33	23	16	Y
——··	Z	3	46	45	18	2	22	17	10	7	23	21	51	11	2	15	59	72	14	4	3	9	11	12	36	42	87	16	21	27	9	10	25	66	47	15	15	Z
·————	1	2	5	10	3	3	5	13	4	2	29	5	14	9	7	14	30	28	9	4	2	3	12	14	17	19	22	84	63	13	8	10	8	19	32	57	55	1
··———	2	7	14	22	5	4	20	13	3	25	26	9	14	2	3	17	37	28	6	5	3	6	10	11	17	30	13	62	89	54	20	5	14	20	21	16	11	2
···——	3	3	8	21	5	4	32	6	12	2	23	6	13	5	2	5	37	19	9	7	6	4	16	6	22	25	12	18	64	86	31	23	41	16	17	8	10	3
····—	4	6	19	19	12	8	25	14	16	7	21	13	19	3	3	2	17	29	11	9	3	17	55	8	37	24	3	5	26	44	89	42	44	32	10	3	3	4
·····	5	8	45	15	14	2	45	4	67	7	14	4	41	2	0	4	13	7	9	27	2	14	45	7	45	10	10	14	10	30	69	90	42	24	10	6	5	5
—····	6	7	80	30	17	4	23	4	14	2	11	11	27	6	2	7	16	30	11	14	3	12	30	9	58	38	39	15	14	26	24	17	88	69	14	5	14	6
——···	7	6	33	22	14	5	25	6	4	6	24	13	32	7	6	7	36	39	12	6	2	3	13	9	30	30	50	22	29	18	15	12	61	85	70	20	13	7
———··	8	3	23	40	6	3	15	15	6	2	33	10	14	3	6	14	12	45	2	6	4	6	7	5	24	35	50	42	29	16	16	9	30	60	89	61	26	8
————·	9	3	14	23	3	1	6	14	5	2	30	6	7	16	11	10	31	32	5	6	7	6	3	8	11	21	24	57	39	9	12	4	11	42	56	91	78	9
—————	0	9	3	11	2	5	7	14	4	5	30	8	3	2	3	25	21	29	2	3	4	5	3	2	12	15	20	50	26	9	11	5	22	17	52	81	94	

Borg and Groenen, Modern Multidimensional Scaling



Borg and Groenen, Modern Multidimensional Scaling

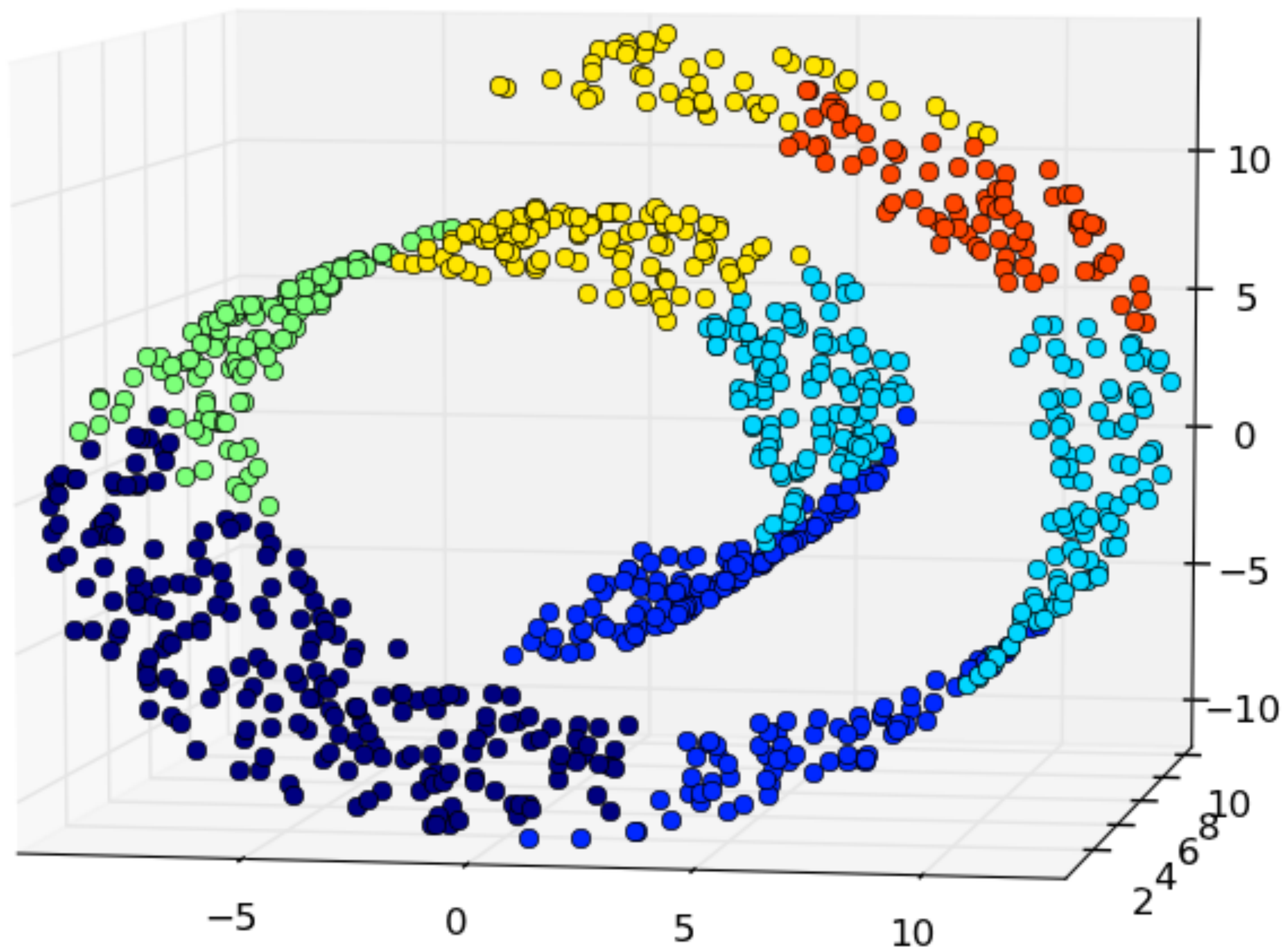
“Classical” Multidimensional Scaling

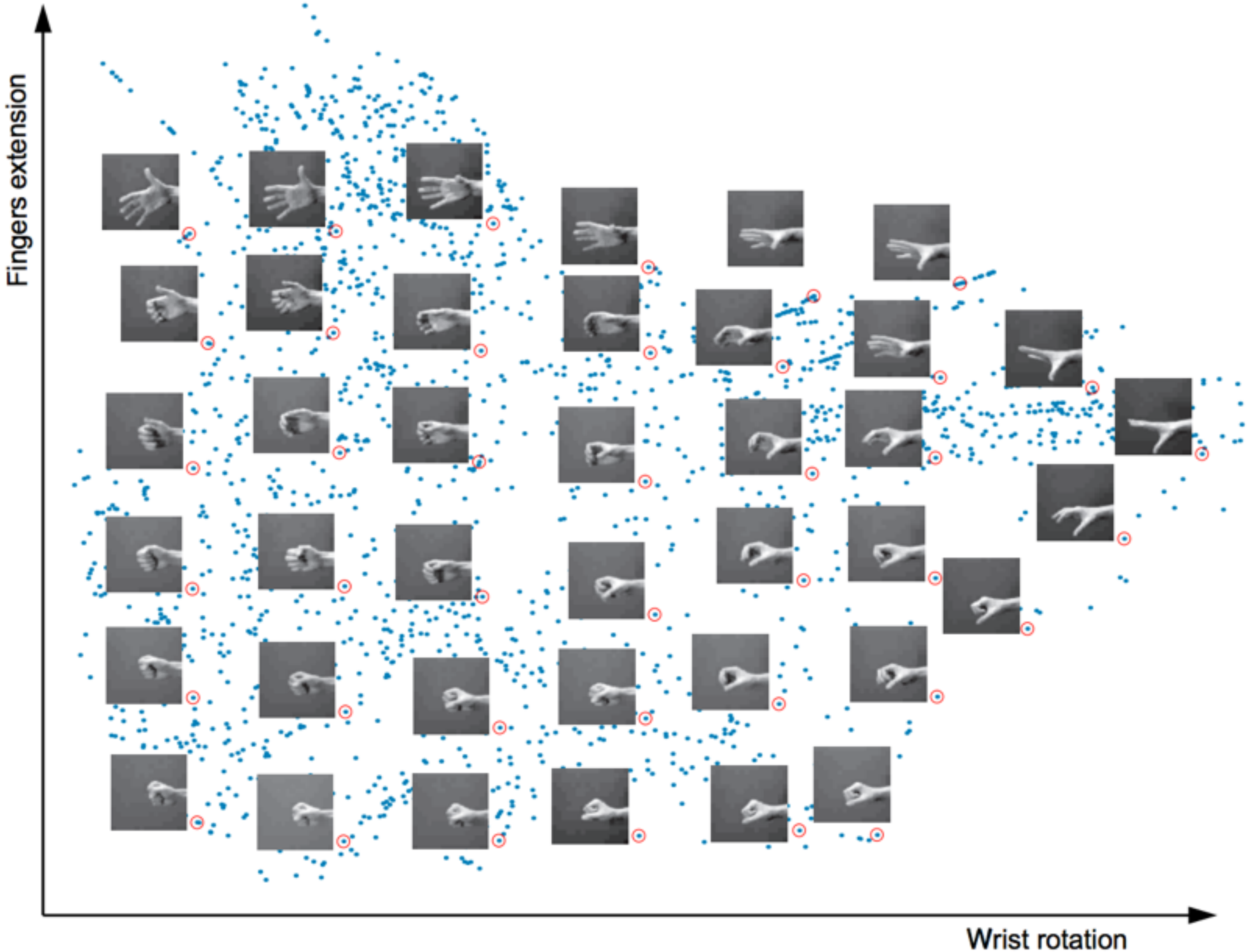
- Algorithm:
- Given $D_{ij} = |X_i - X_j|^2$, create $B = -\frac{1}{2} H D H^T$
- PCA of B is equal to the PCA of X
  - Huh?!

“Nonlinear” dimensionality reduction

(ie: projection is not a matrix operation)

Data might have "high-order" structure





http://isomap.stanford.edu/Supplemental_Fig.pdf

We might want to minimize something else besides “difference between squared distances”

t-SNE: difference between neighbor ordering

Why not distances?

The curse of Dimensionality

- High dimensional space looks **nothing** like low-dimensional space
  - **Most distances become meaningless**