The magic of dimensionality reduction

Alex Peattie, CTO | Peg

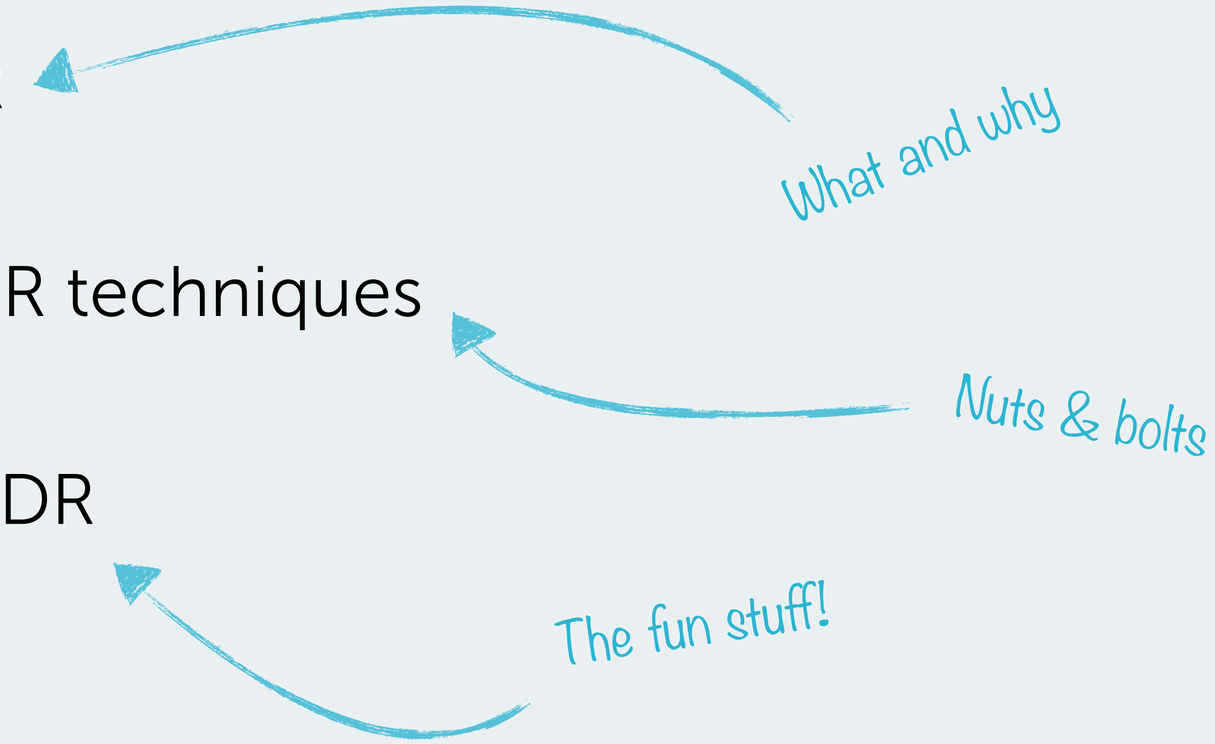
Obligatory poll

Keywords: dimensionality reduction, PCA, LDA, feature selection, feature extraction, embeddings

- I know little / nothing about DR
- I’ve know something about / have played around with DR
- I know about DR already / I regularly work with DR

Schedule

- Overview of DR
- Specifics of 4 DR techniques
- Applications of DR
- Questions



Approach

- Emphasis on intuition and clarity over mathematical rigour
- Emphasis on visualization over other applications
- Code at: https://github.com/alexpeattie/magic

DIMENSIONALITY REDUCTION:

WHAT AND WHY?

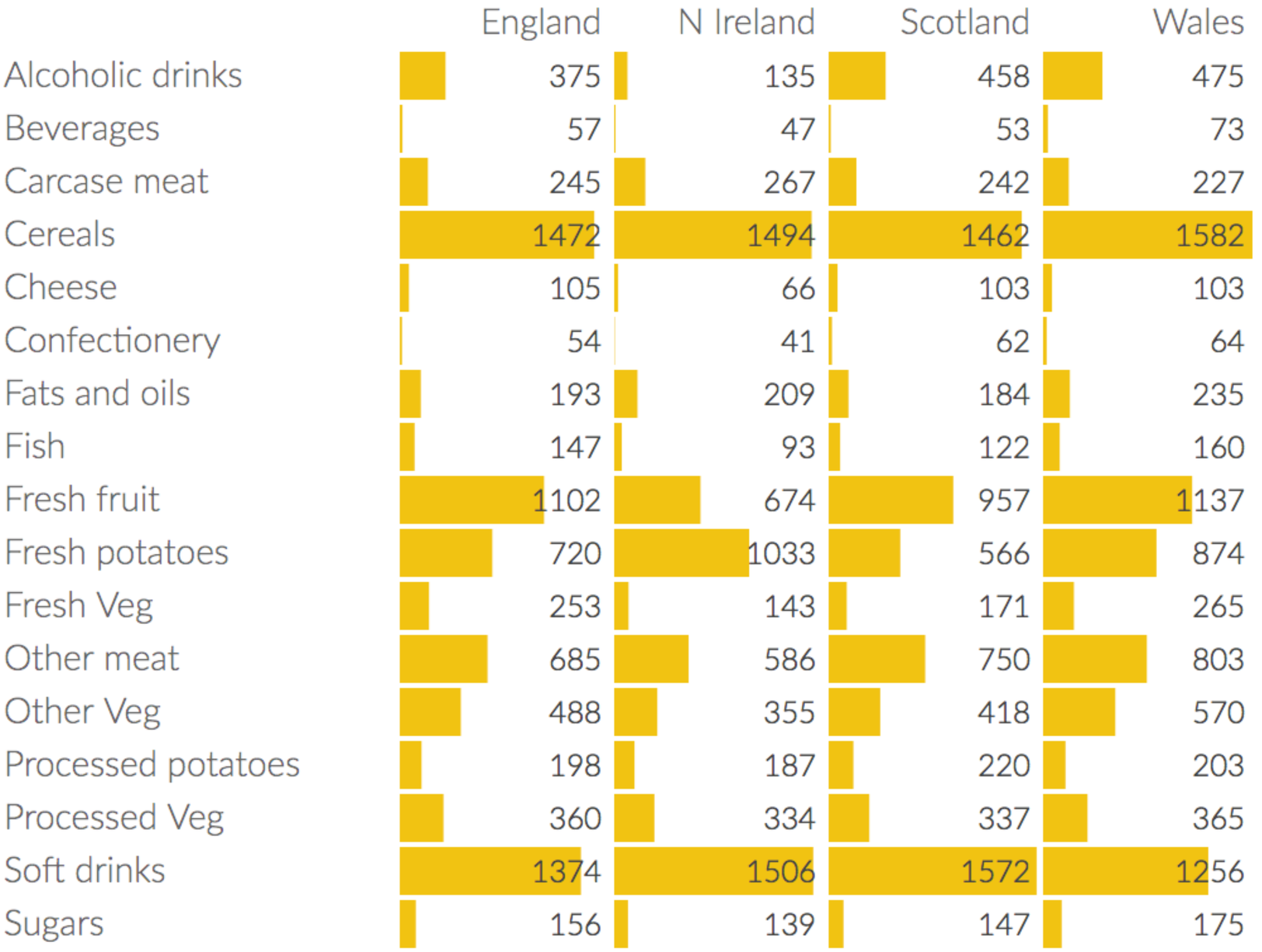
A true story

8.524	150	36.107	2.206	86.364	371.446	551	9.932.931,20 ISK	8.000.000,00 ISK	-1.932.931,20 ISK	-17.396.380,80 ISK	-19,46%	160,00	9,00	1	9,00	89.396.3
100	1	2.508	2	13.542	16.050	2	474.533,60 ISK	730.000,00 ISK	255.466,40 ISK	5.747.994,00 ISK	53,84%	64,00	22,50	1	22,50	10.677.0
	24	5.812	243	13.263	47.029	41	1.129.865,60 ISK	1.340.000,00 ISK	210.134,40 ISK	3.087.689,14 ISK	18,60%	98,00	14,69	1	14,69	16.602.1
1.028	18	4.895		18.387	50.513	89	1.186.402,40 ISK	1.300.000,00 ISK	113.597,60 ISK	1.669.189,22 ISK	9,57%	98,00	14,69	1	14,69	17.432.8
500		1.901	10	4.005	18.008	1	386.829,10 ISK	290.000,00 ISK	-96.829,10 ISK	-2.178.654,75 ISK	-25,03%	64,00	22,50	1	22,50	8.703.6
968	8	4.618	300	10.728	35.496	96	1.179.842,40 ISK	1.200.000,00 ISK	20.157,60 ISK	296.193,31 ISK	1,71%	98,00	14,69	1	14,69	17.336.4
47.196	1.114	159.341		659.556	2.643.862	2.863	49.416.416,80 ISK	50.000.000,00 ISK	583.583,20 ISK	4.201.799,04 ISK	1,18%	200,00	7,20	1	7,20	355.798.2
250	2	2.304	45	9.517	15.021	20	476.823,60 ISK	480.000,00 ISK	3.176,40 ISK	71.469,00 ISK	0,67%	64,00	22,50	1	22,50	10.728.5
1.004		2.612	50	5.000	16.070	10	563.465,40 ISK	560.000,00 ISK	-3.465,40 ISK	-75.608,73 ISK	-0,62%	66,00	21,82	1	21,82	12.293.7
322	1	2.010	121	5.627	20.598	15	505.535,90 ISK	540.000,00 ISK	34.464,10 ISK	763.512,37 ISK	6,82%	65,00	22,15	1	22,15	11.199.5
9.234	191	35.139	2.308	150.537	552.145	472	12.175.706,80 ISK	10.600.000,00 ISK	-1.575.706,80 ISK	-17.589.285,21 ISK	-12,94%	129,00	11,16	1	11,16	135.914.8
9.336	110	36.172	2.509	170.515	462.050	562	12.040.208,90 ISK	9.000.000,00 ISK	-3.040.208,90 ISK	-27.361.880,10 ISK	-25,25%	160,00	9,00	1	9,00	108.361.8
1.598	12	3.597		10.387	43.332	81	1.028.084,50 ISK	1.200.000,00 ISK	171.915,50 ISK	2.062.986,00 ISK	16,72%	120,00	12,00	1	12,00	12.337.0
70.795	1.671	239.012		989.335	3.965.793	4.294	74.124.417,80 ISK	77.800.000,00 ISK	3.675.582,20 ISK	26.464.191,84 ISK	4,96%	200,00	7,20	1	7,20	533.695.8
251	4	2.507	75	4.612	13.041	10	425.920,70 ISK		0,00 ISK	0,00 ISK	0,00%	66,00	21,82	1	21,82	9.292.8
9.118	160	37.087	2.604	120.257	561.026	591	12.164.809,70 ISK	9.500.000,00 ISK	-2.664.809,70 ISK	-28.852.074,95 ISK	-21,91%	133,00	10,83	1	10,83	131.709.2
		4	2		424		4.536,00 ISK	4.500,00 ISK	-36,00 ISK	-10.368,00 ISK	-0,79%	5,00	288,00	1	288,00	1.306.3
10		202		2	3.075	2	35.293,90 ISK	34.500,00 ISK	-793,90 ISK	-228.643,20 ISK	-2,25%	5,00	288,00	1	288,00	10.164.6
12	1	150		358	1.758	2	31.556,40 ISK	3.400,00 ISK	2.443,60 ISK	3.518.784,00 ISK	7,74%	1,00	1.440,00	10	14.400,00	45.441.2
10		25	4	11	10		6.792,50 ISK	7.200,00 ISK	407,50 ISK	586.800,00 ISK	6,00%	1,00	1.440,00	1	1.440,00	9.781.2
3	40	2	18	46		6	109.899,40 ISK	118.800,00 ISK	8.900,60 ISK	4.272.288,00 ISK	8,10%	3,00	480,00	1	480,00	52.751.7
2		4	2	5	2	1	3.174,90 ISK	5.000,00 ISK	1.825,10 ISK	876.048,00 ISK	57,49%	3,00	480,00	1	480,00	1.523.9
4	11	99	4	8		2	36.284,80 ISK	38.200,00 ISK	1.915,20 ISK	919.296,00 ISK	5,28%	3,00	480,00	1	480,00	17.416.7
7		3			4		1.468,80 ISK	1.900,00 ISK	431,20 ISK	206.976,00 ISK	29,36%	3,00	480,00	1	480,00	705.0
3	22	78	1			8	61.883,60 ISK	75.000,00 ISK	13.116,40 ISK	6.295.872,00 ISK	21,20%	3,00	480,00	1	480,00	29.704.1
3		2	3	1	4		3.132,10 ISK	3.100,00 ISK	-32,10 ISK	-15.408,00 ISK	-1,02%	3,00	480,00	1	480,00	1.503.4
22	45	4	3	32		13	118.243,80 ISK	138.000,00 ISK	19.756,20 ISK	9.482.976,00 ISK	16,71%	3,00	480,00	1	480,00	56.757.0
3			3	4		1	3.893,00 ISK	5.400,00 ISK	1.507,00 ISK	723.360,00 ISK	38,71%	3,00	480,00	1	480,00	1.868.6
42	33	5	15	5	12	8	100.344,10 ISK	112.900,00 ISK	12.555,90 ISK	6.026.832,00 ISK	12,51%	3,00	480,00	1	480,00	48.165.1
3			1	7		1	2.316,50 ISK	3.000,00 ISK	683,50 ISK	328.080,00 ISK	29,51%	3,00	480,00	1	480,00	1.111.9
3	14	49	3	51		3	40.383,30 ISK	47.900,00 ISK	7.516,70 ISK	3.608.016,00 ISK	18,61%	3,00	480,00	1	480,00	19.383.9
9		2		1	3		1.775,80 ISK	2.370,00 ISK	594,20 ISK	285.216,00 ISK	33,46%	3,00	480,00	1	480,00	852,3

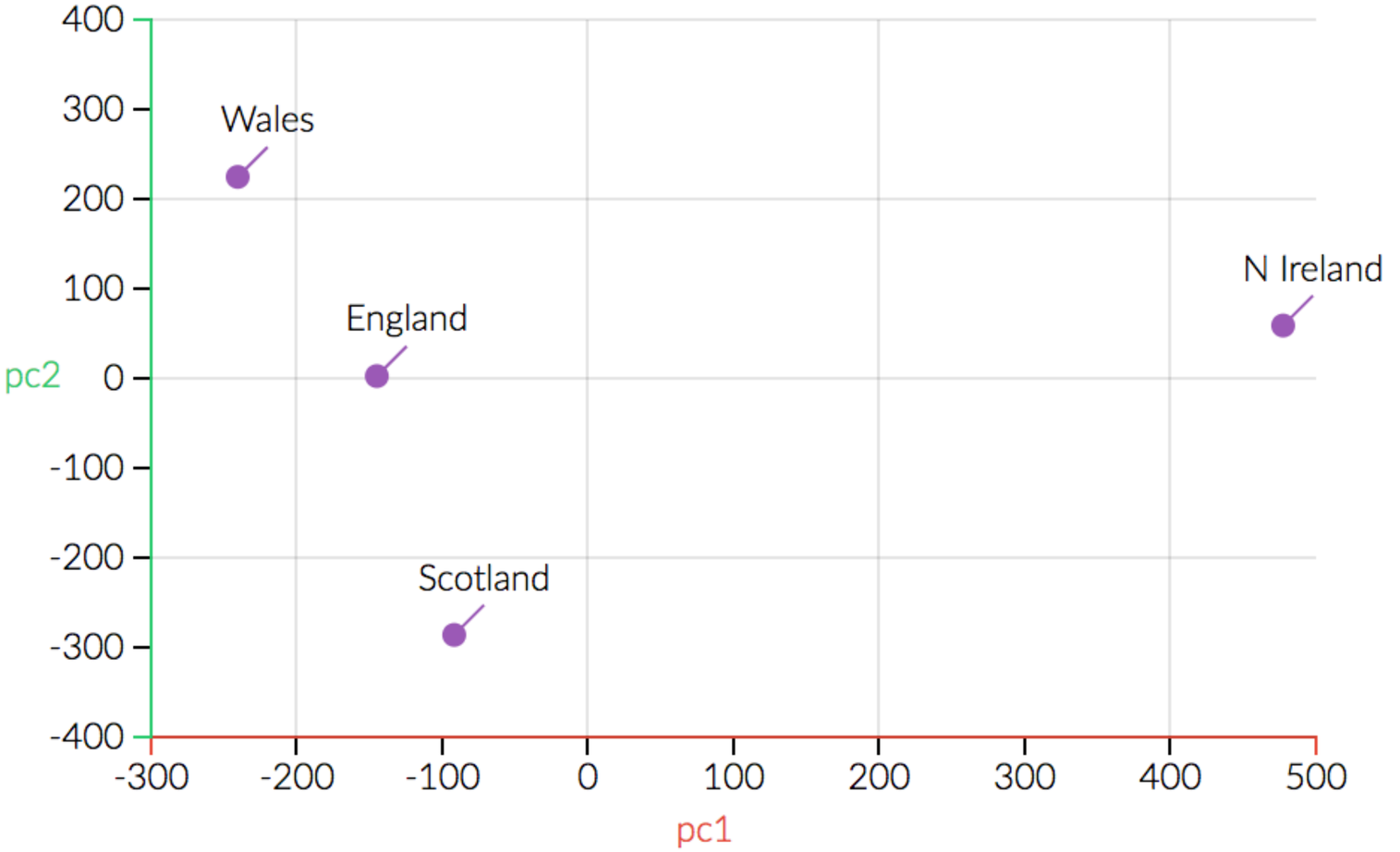

HSBC
NIKE
SAMSUNG
IBM
Disney
Coca-Cola
Virgin
SUBWAY
asos
LACOSTE

Understandable → Meaningful

Average consumption of food types in grams per person per week

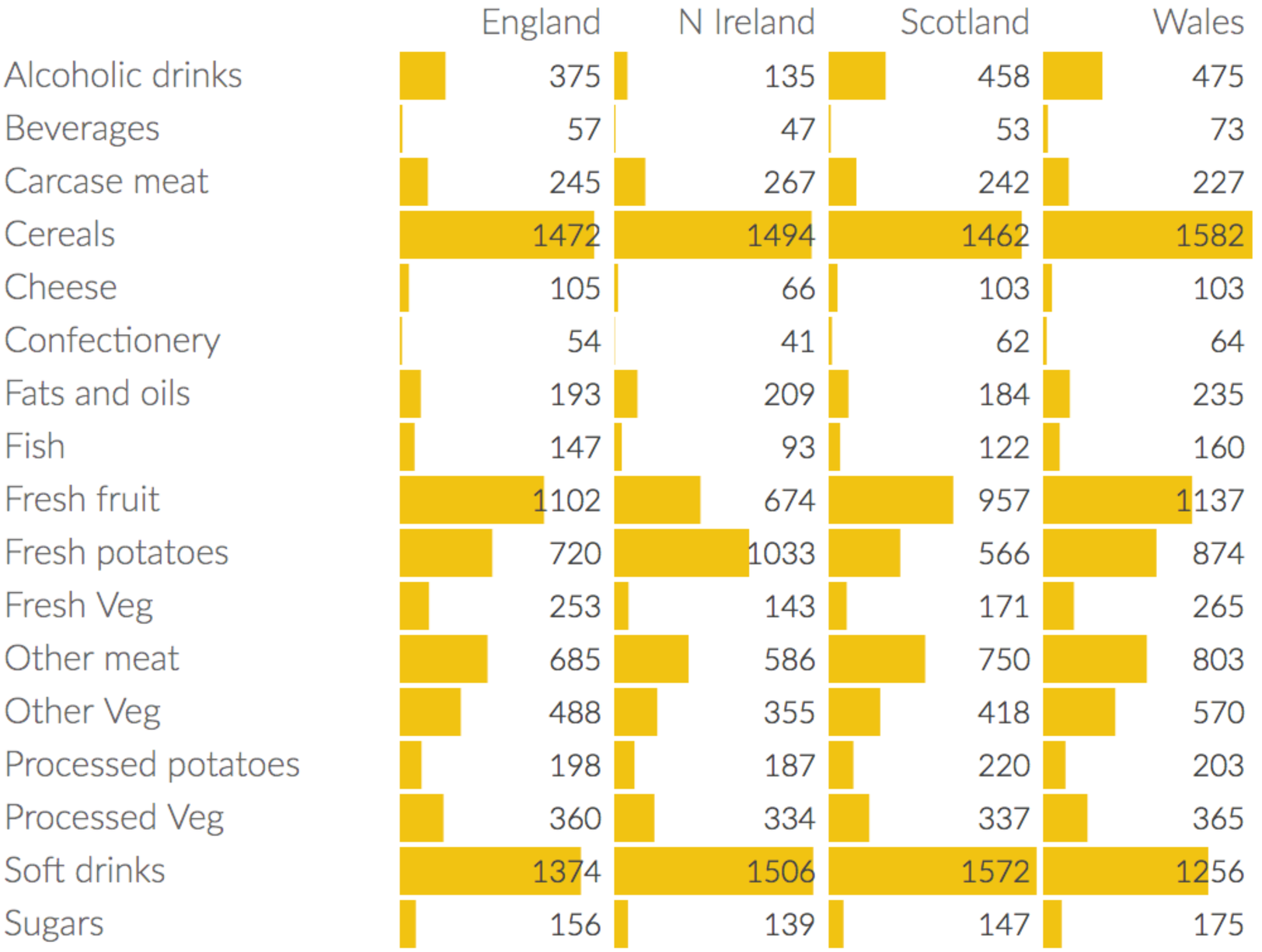


Credit: Victor Powell

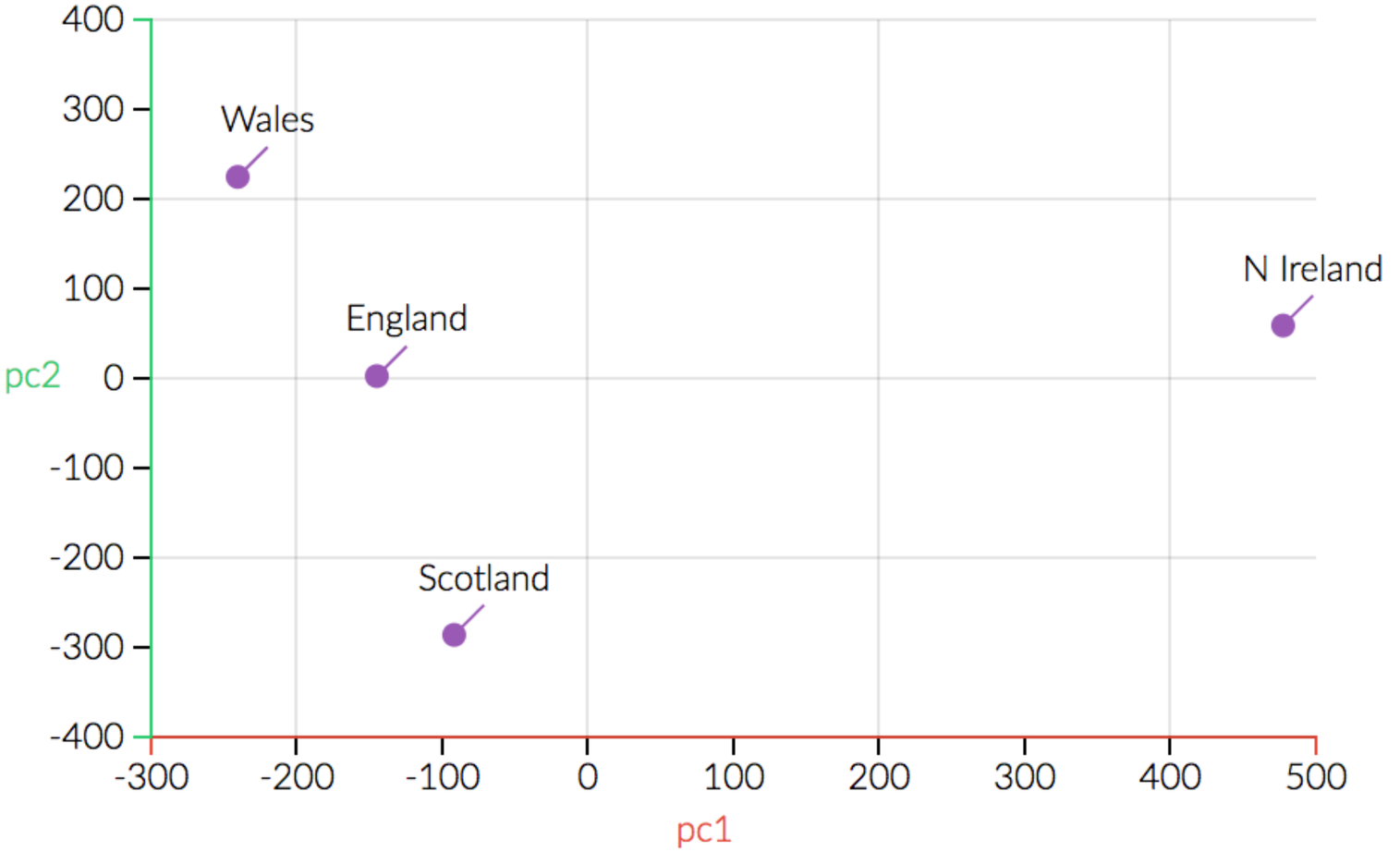


Credit: Victor Powell

Average consumption of food types in grams per person per week



Credit: Victor Powell



Credit: Victor Powell

Understandable → Meaningful

8.524	150	36.107	2.206	86.364	371.446	551	9.932.931,20 ISK	8.000.000,00 ISK	-1.932.931,20 ISK	-17.396.380,80 ISK	-19,46%	160,00	9,00	1	9,00	89.396.3
100	1	2.508	2	13.542	16.050	2	474.533,60 ISK	730.000,00 ISK	255.466,40 ISK	5.747.994,00 ISK	53,84%	64,00	22,50	1	22,50	10.677.0
	24	5.812	243	13.263	47.029	41	1.129.865,60 ISK	1.340.000,00 ISK	210.134,40 ISK	3.087.689,14 ISK	18,60%	98,00	14,69	1	14,69	16.602.1
1.028	18	4.895		18.387	50.513	89	1.186.402,40 ISK	1.300.000,00 ISK	113.597,60 ISK	1.669.189,22 ISK	9,57%	98,00	14,69	1	14,69	17.432.8
500		1.901	10	4.005	18.008	1	386.829,10 ISK	290.000,00 ISK	-96.829,10 ISK	-2.178.654,75 ISK	-25,03%	64,00	22,50	1	22,50	8.703.6
968	8	4.618	300	10.728	35.496	96	1.179.842,40 ISK	1.200.000,00 ISK	20.157,60 ISK	296.193,31 ISK	1,71%	98,00	14,69	1	14,69	17.336.4
47.196	1.114	159.341		659.556	2.643.862	2.863	49.416.416,80 ISK	50.000.000,00 ISK	583.583,20 ISK	4.201.799,04 ISK	1,18%	200,00	7,20	1	7,20	355.798.2
250	2	2.304	45	9.517	15.021	20	476.823,60 ISK	480.000,00 ISK	3.176,40 ISK	71.469,00 ISK	0,67%	64,00	22,50	1	22,50	10.728.5
1.004		2.612	50	5.000	16.070	10	563.465,40 ISK	560.000,00 ISK	-3.465,40 ISK	-75.608,73 ISK	-0,62%	66,00	21,82	1	21,82	12.293.7
322	1	2.010	121	5.627	20.598	15	505.535,90 ISK	540.000,00 ISK	34.464,10 ISK	763.512,37 ISK	6,82%	65,00	22,15	1	22,15	11.199.5
9.234	191	35.139	2.308	150.537	552.145	472	12.175.706,80 ISK	10.600.000,00 ISK	-1.575.706,80 ISK	-17.589.285,21 ISK	-12,94%	129,00	11,16	1	11,16	135.914.8
9.336	110	36.172	2.509	170.515	462.050	562	12.040.208,90 ISK	9.000.000,00 ISK	-3.040.208,90 ISK	-27.361.880,10 ISK	-25,25%	160,00	9,00	1	9,00	108.361.8
1.598	12	3.597		10.387	43.332	81	1.028.084,50 ISK	1.200.000,00 ISK	171.915,50 ISK	2.062.986,00 ISK	16,72%	120,00	12,00	1	12,00	12.337.0
70.795	1.671	239.012		989.335	3.965.793	4.294	74.124.417,80 ISK	77.800.000,00 ISK	3.675.582,20 ISK	26.464.191,84 ISK	4,96%	200,00	7,20	1	7,20	533.695.8
251	4	2.507	75	4.612	13.041	10	425.920,70 ISK		0,00 ISK	0,00 ISK	0,00%	66,00	21,82	1	21,82	9.292.8
9.118	160	37.087	2.604	120.257	561.026	591	12.164.809,70 ISK	9.500.000,00 ISK	-2.664.809,70 ISK	-28.852.074,95 ISK	-21,91%	133,00	10,83	1	10,83	131.709.2
		4	2		424		4.536,00 ISK	4.500,00 ISK	-36,00 ISK	-10.368,00 ISK	-0,79%	5,00	288,00	1	288,00	1.306.3
10		202		2	3.075	2	35.293,90 ISK	34.500,00 ISK	-793,90 ISK	-228.643,20 ISK	-2,25%	5,00	288,00	1	288,00	10.164.6
12	1	150		358	1.758	2	31.556,40 ISK	3.400,00 ISK	2.443,60 ISK	3.518.784,00 ISK	7,74%	1,00	1.440,00	10	14.400,00	45.441.2
10		25	4	11	10		6.792,50 ISK	7.200,00 ISK	407,50 ISK	586.800,00 ISK	6,00%	1,00	1.440,00	1	1.440,00	9.781.2
3	40	2	18	46		6	109.899,40 ISK	118.800,00 ISK	8.900,60 ISK	4.272.288,00 ISK	8,10%	3,00	480,00	1	480,00	52.751.7
2		4	2	5	2	1	3.174,90 ISK	5.000,00 ISK	1.825,10 ISK	876.048,00 ISK	57,49%	3,00	480,00	1	480,00	1.523.9
4	11	99	4	8		2	36.284,80 ISK	38.200,00 ISK	1.915,20 ISK	919.296,00 ISK	5,28%	3,00	480,00	1	480,00	17.416.7
7		3			4		1.468,80 ISK	1.900,00 ISK	431,20 ISK	206.976,00 ISK	29,36%	3,00	480,00	1	480,00	705.0
3	22	78	1			8	61.883,60 ISK	75.000,00 ISK	13.116,40 ISK	6.295.872,00 ISK	21,20%	3,00	480,00	1	480,00	29.704.1
3		2	3	1	4		3.132,10 ISK	3.100,00 ISK	-32,10 ISK	-15.408,00 ISK	-1,02%	3,00	480,00	1	480,00	1.503.4
22	45	4	3	32		13	118.243,80 ISK	138.000,00 ISK	19.756,20 ISK	9.482.976,00 ISK	16,71%	3,00	480,00	1	480,00	56.757.0
3			3	4		1	3.893,00 ISK	5.400,00 ISK	1.507,00 ISK	723.360,00 ISK	38,71%	3,00	480,00	1	480,00	1.868.6
42	33	5	15	5	12	8	100.344,10 ISK	112.900,00 ISK	12.555,90 ISK	6.026.832,00 ISK	12,51%	3,00	480,00	1	480,00	48.165.1
3			1	7		1	2.316,50 ISK	3.000,00 ISK	683,50 ISK	328.080,00 ISK	29,51%	3,00	480,00	1	480,00	1.111.9
3	14	49	3	51		3	40.383,30 ISK	47.900,00 ISK	7.516,70 ISK	3.608.016,00 ISK	18,61%	3,00	480,00	1	480,00	19.383.9
9		2		1	3		1.775,80 ISK	2.370,00 ISK	594,20 ISK	285.216,00 ISK	33,46%	3,00	480,00	1	480,00	852,3


HSBC
NIKE
SAMSUNG
IBM
Disney
Coca-Cola
Virgin
SUBWAY
asos
LACOSTE

What’s going on here?

Feature selection

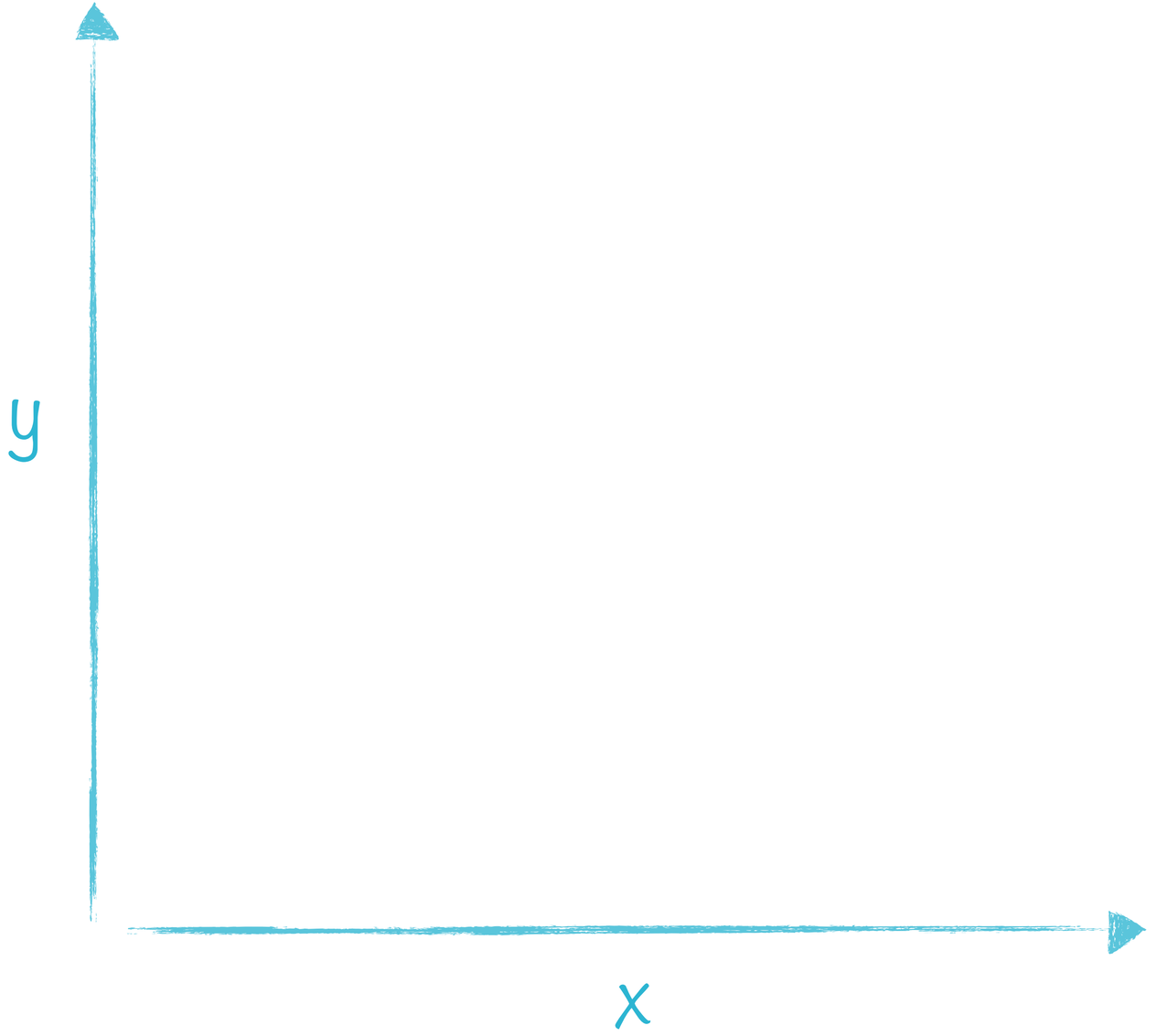
Feature extraction


evo
203
www.evo.co.uk
500+BHP SUPER-ESTATE SHOOTOUT
Beat Lamborghinis, with a wardrobe in the boot
evo
THE THRILL OF DRIVING
CAR OF THE YEAR 2014
Scottish Borders
10 brilliant cars
One winner
Ferrari Fever
WORLD EXCLUSIVE TEST
ENZO v LAFERRARI
CAR OF THE YEAR 2014 ◘ LAFERRARI V ENZO ◘ BENTLEY CONTINENTAL GT3-R ◘ SUPER-E

Name	Horsepower	0-60 (secs)	Max speed	Price	MPG	Road tax
Audi S4	340	5.1	180.1	44,015	37	240
MINI Cooper S	75	10.5	100.6	18,780	51	115
Nissan GT-R	565	2.765	209.8	81,350	24	535
Porsche 911 GT3	381	4.2	193.9	77,891	28	520
Renault Clio Williams	142	7.717	133.1	11,585	55	0
Volkswagen Golf R	296	4.94	163.6	31,255	38	190
BMW i8	357	3.6	198.7	105,580	39	440

Name	Horsepower	0-60 (secs)	Max speed	Price	MPG	Road tax
Audi S4	340	5.1	180.1	44,015	37	240
MINI Cooper S	75	10.5	100.6	18,780	51	115
Nissan GT-R	565	2.765	209.8	81,350	24	535
Porsche 911 GT3	381	4.2	193.9	77,891	28	520
Renault Clio Williams	142	7.717	133.1	11,585	55	0
Volkswagen Golf R	296	4.94	163.6	31,255	38	190
BMW i8	357	3.6	198.7	105,580	39	440

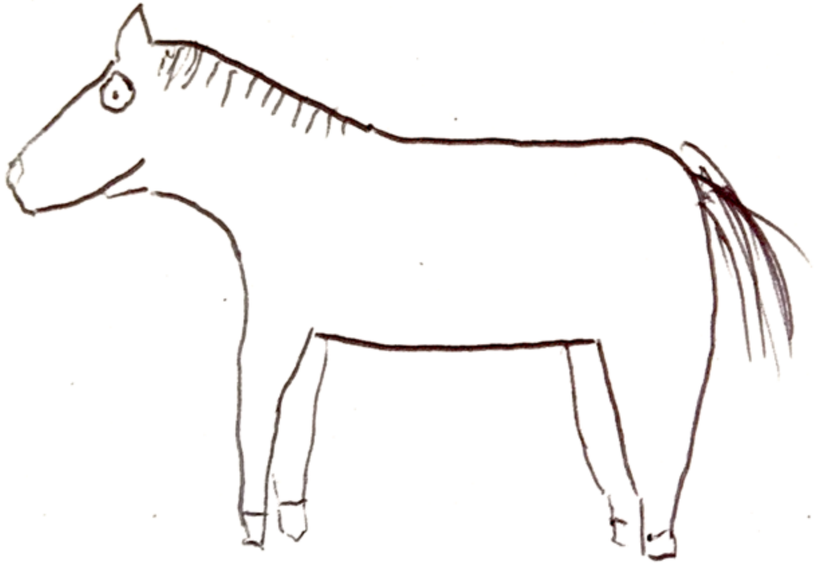

Golf


y
x

Name	Horsepower	0-60 (secs)	Max speed	Price	MPG	Road tax
Audi S4	340	5.1	180.1	44,015	37	240
MINI Cooper S	75	10.5	100.6	18,780	51	115
Nissan GT-R	565	2.765	209.8	81,350	24	535
Porsche 911 GT3	381	4.2	193.9	77,891	28	520
Renault Clio Williams	142	7.717	133.1	11,585	55	0
Volkswagen Golf R	296	4.94	163.6	31,255	38	190
BMW i8	357	3.6	198.7	105,580	39	440

Name	Horsepower	Price
Audi S4	340	44,015
MINI Cooper S	75	18,780
Nissan GT-R	565	81,350
Porsche 911 GT3	381	77,891
Renault Clio Williams	142	11,585
Volkswagen Golf R	296	31,255
BMW i8	357	105,580

Feature selection

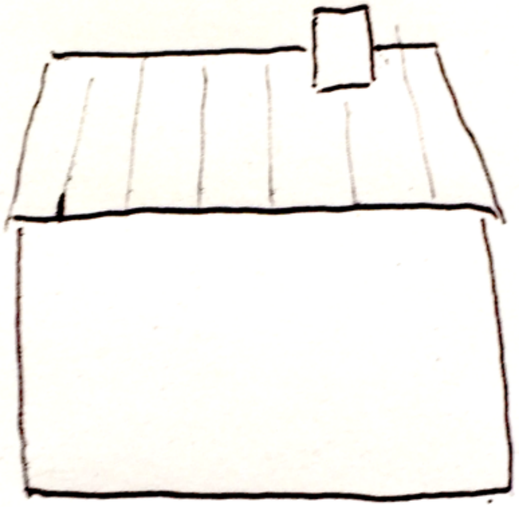


A

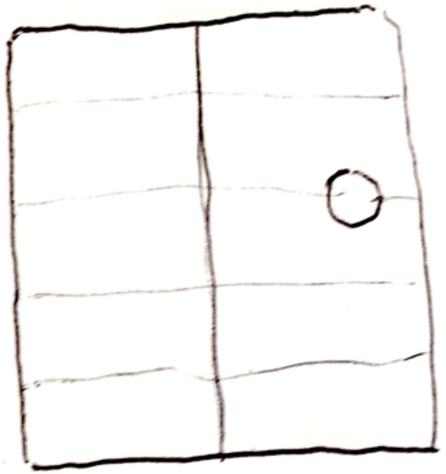
B

C

A



B



C

A

B

C

Name	Horsepower	0-60 (secs)	Max speed	Price	MPG	Road tax
Audi S4	340	5.1	180.1	44,015	37	240
MINI Cooper S	75	10.5	100.6	18,780	51	115
Nissan GT-R	565	2.765	209.8	81,350	24	535
Porsche 911 GT3	381	4.2	193.9	77,891	28	520
Renault Clio Williams	142	7.717	133.1	11,585	55	0
Volkswagen Golf R	296	4.94	163.6	31,255	38	190
BMW i8	357	3.6	198.7	105,580	39	440

Name	Horsepower	Road tax
Audi S4	340	240
MINI Cooper S	75	115
Nissan GT-R	565	535
Porsche 911 GT3	381	520
Renault Clio Williams	142	0
Volkswagen Golf R	296	190
BMW i8	357	440

Name	Max speed (MPH)	Max speed (KMPH)
Audi S4	180.1	289.8
MINI Cooper S	100.6	161.9
Nissan GT-R	209.8	337.6
Porsche 911 GT3	193.9	312.1
Renault Clio Williams	133.1	214.2
Volkswagen Golf R	163.6	263.3
BMW i8	198.7	319.8

Feature extraction

Name	Horsepower	0-60 (secs)	Max speed	Price	MPG	Road tax
Audi S4	340	5.1	180.1	44,015	37	240
MINI Cooper S	75	10.5	100.6	18,780	51	115
Nissan GT-R	565	2.765	209.8	81,350	24	535
Porsche 911 GT3	381	4.2	193.9	77,891	28	520
Renault Clio Williams	142	7.717	133.1	11,585	55	0
Volkswagen Golf R	296	4.94	163.6	31,255	38	190
BMW i8	357	3.6	198.7	105,580	39	440

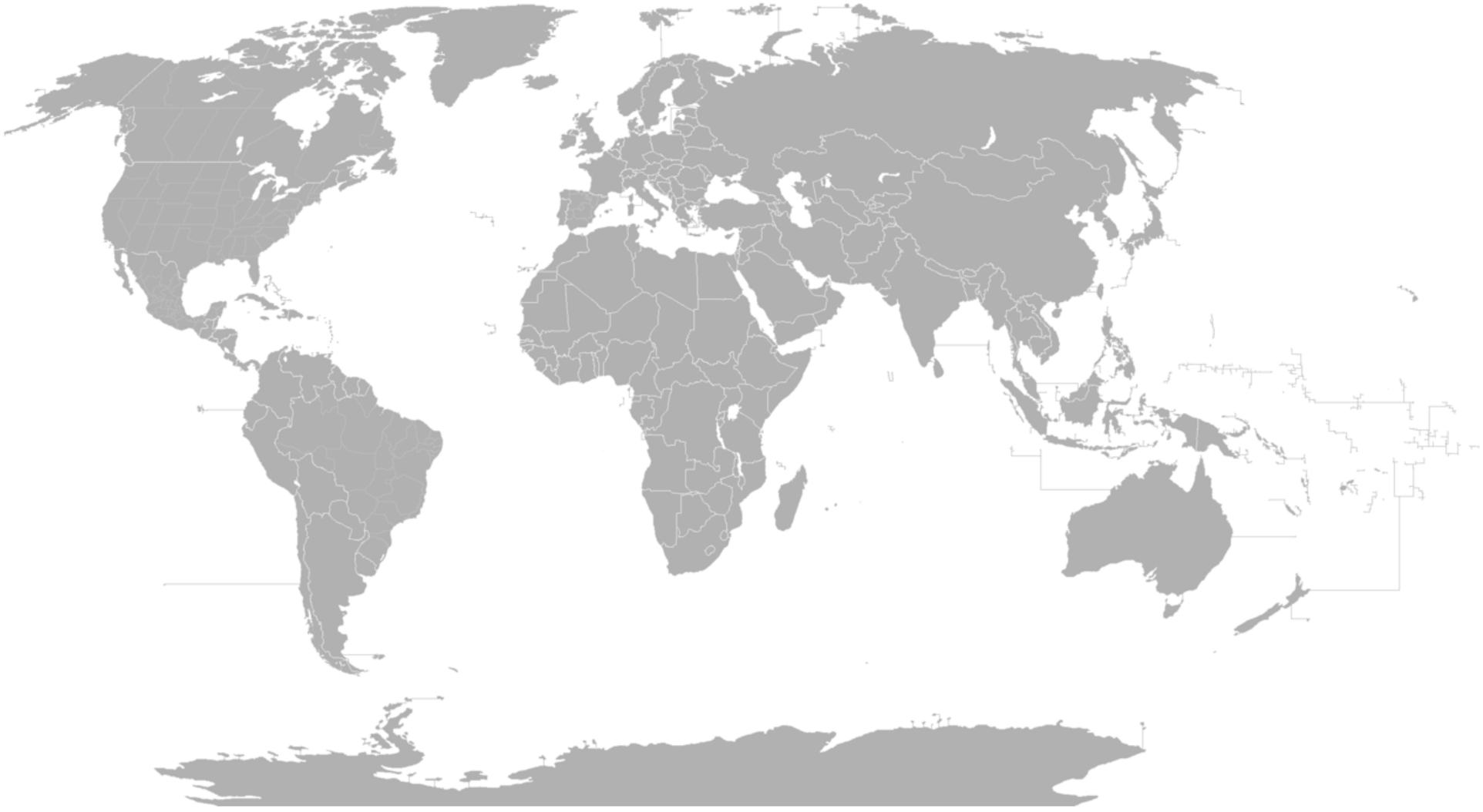
Name	Performance	Cost
Audi S4	6	4
MINI Cooper S	1	2
Nissan GT-R	10	10
Porsche 911 GT3	7	9
Renault Clio Williams	3	1
Volkswagen Golf R	5	3
BMW i8	6	10

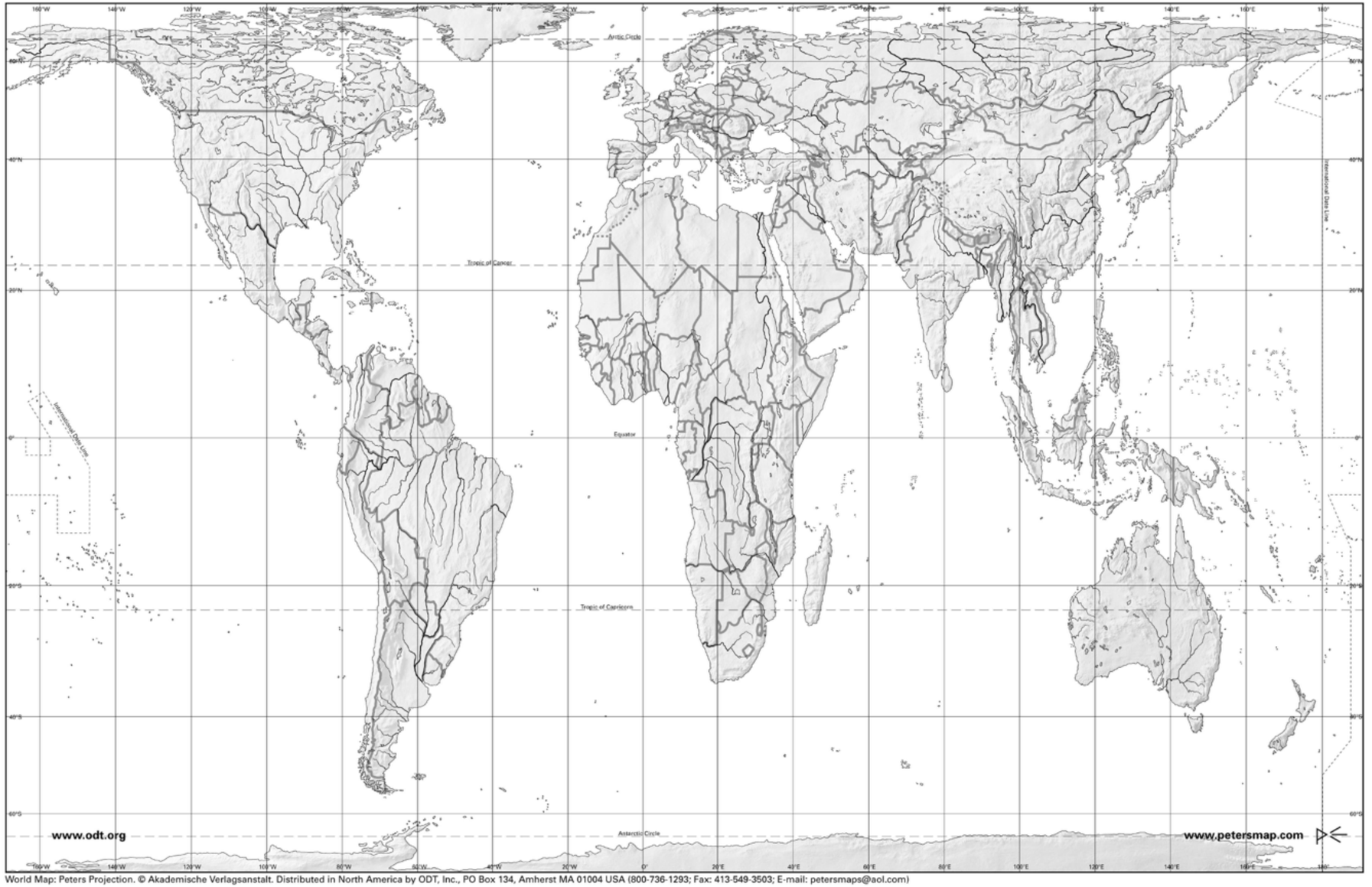
3 ways to think

- Creating new dimensions
- Combining existing dimensions
- Compensating for lost dimensions

Decision time

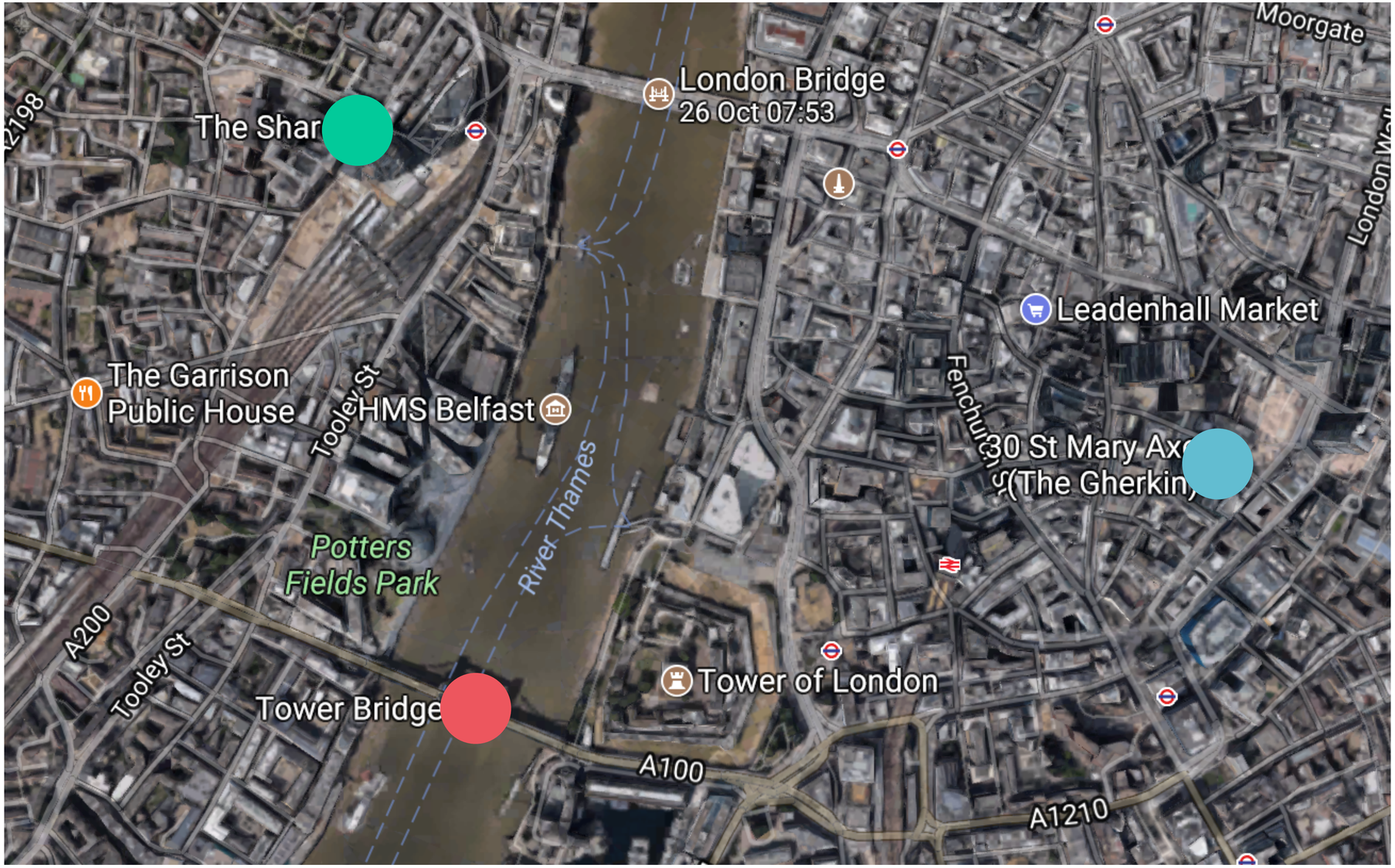
What do I want to preserve?

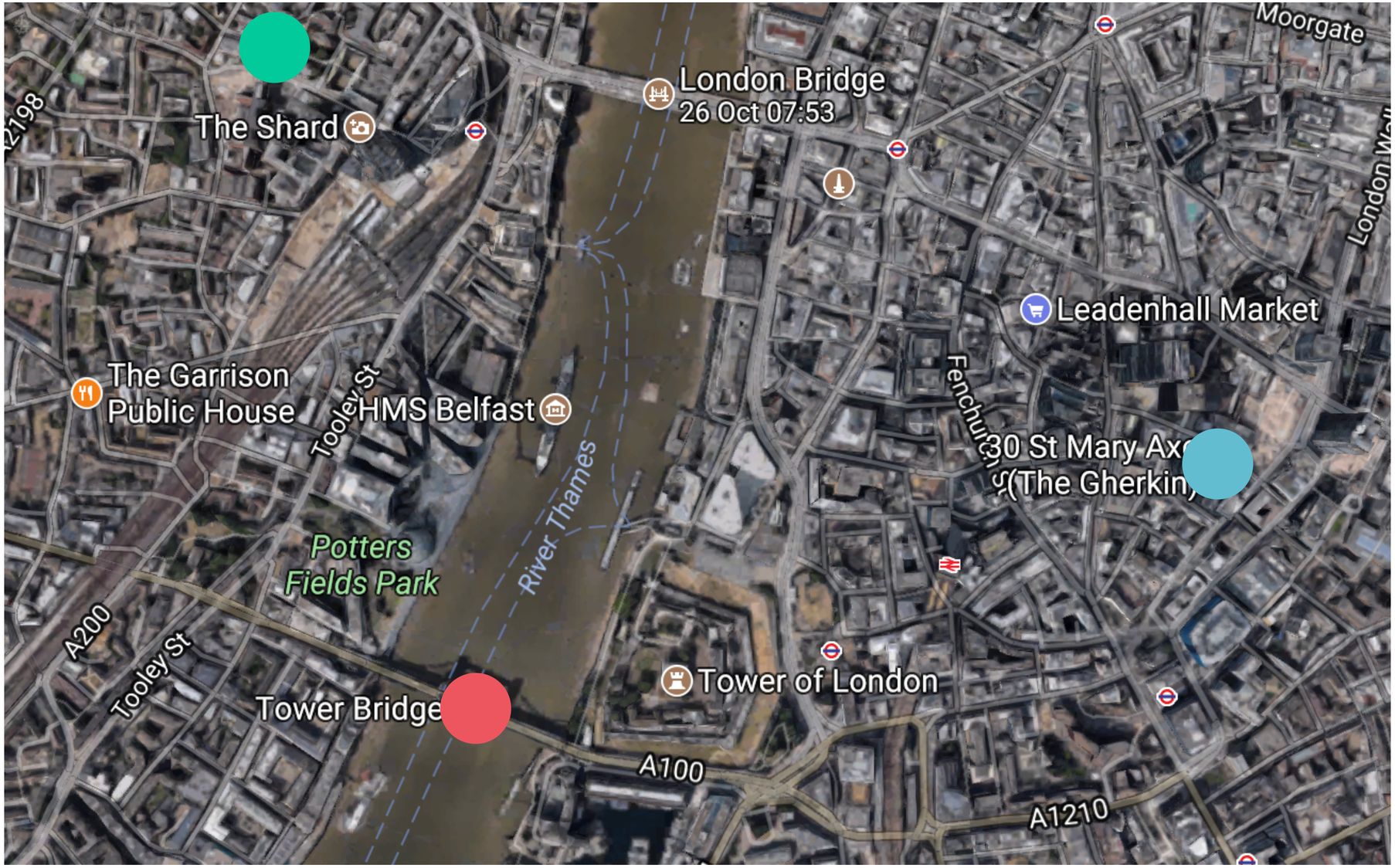


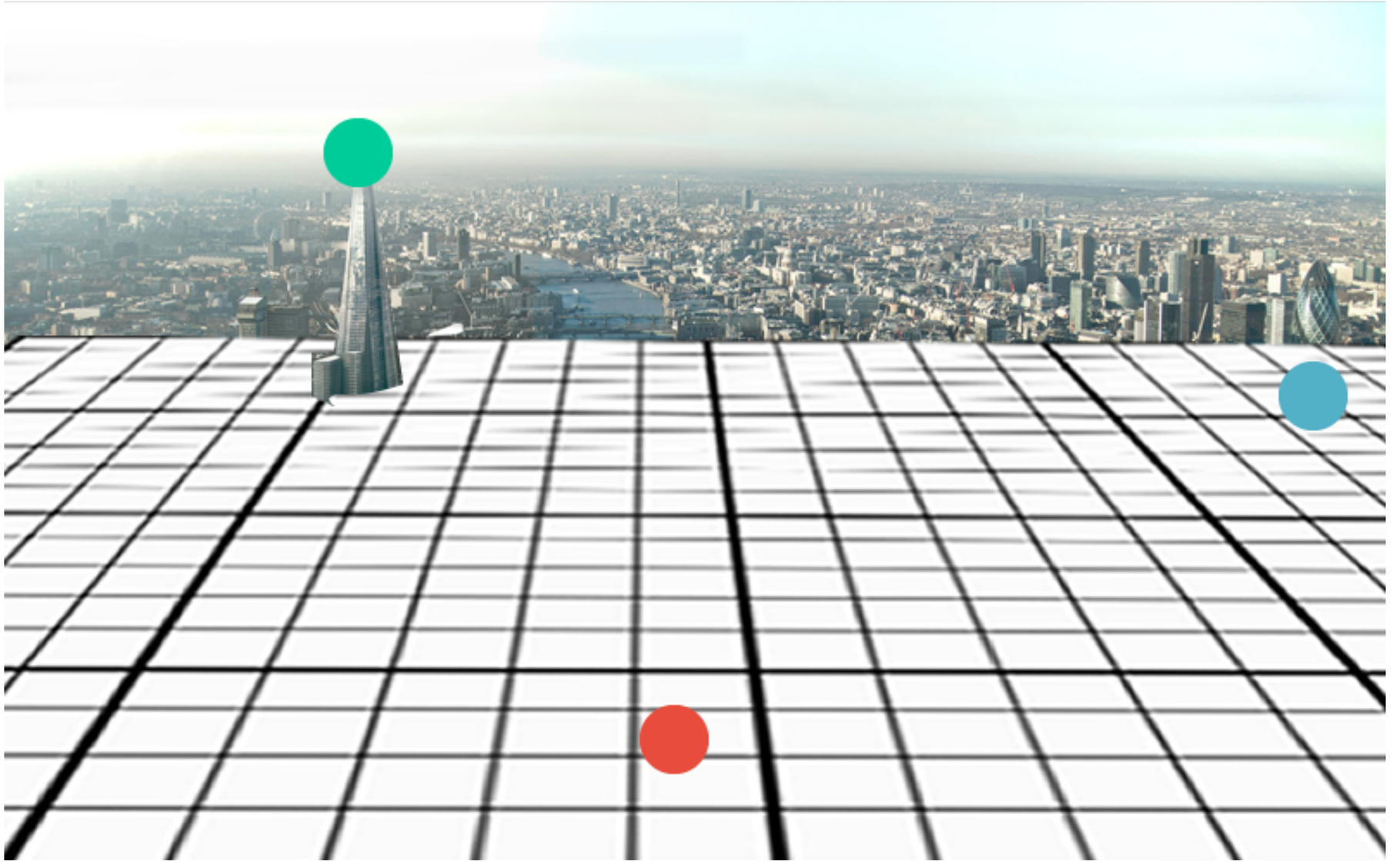


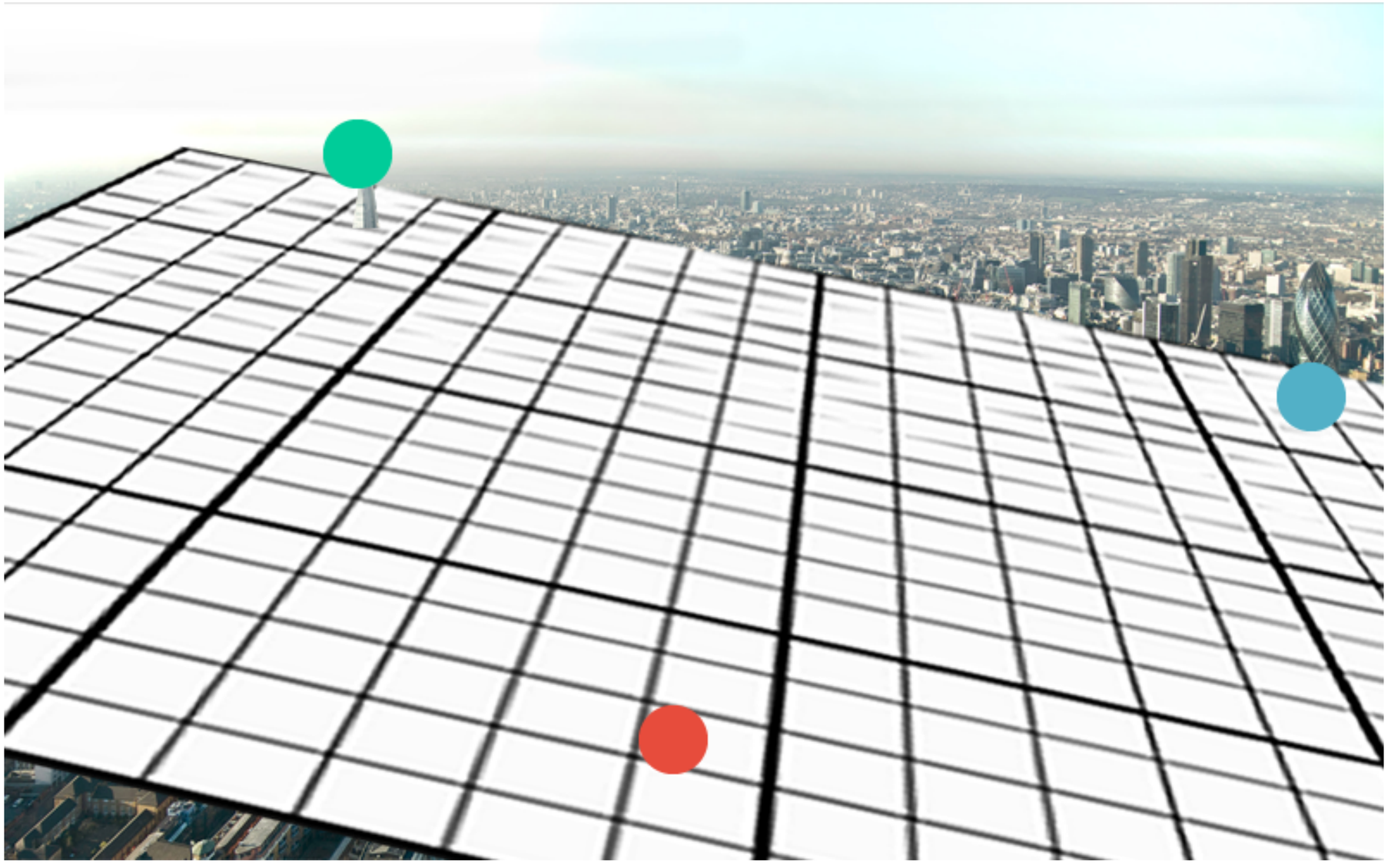
World Map: Peters Projection. © Akademische Verlagsanstalt. Distributed in North America by ODT, Inc., PO Box 134, Amherst MA 01004 USA (800-736-1293; Fax: 413-549-3503; E-mail: petersmaps@aol.com)

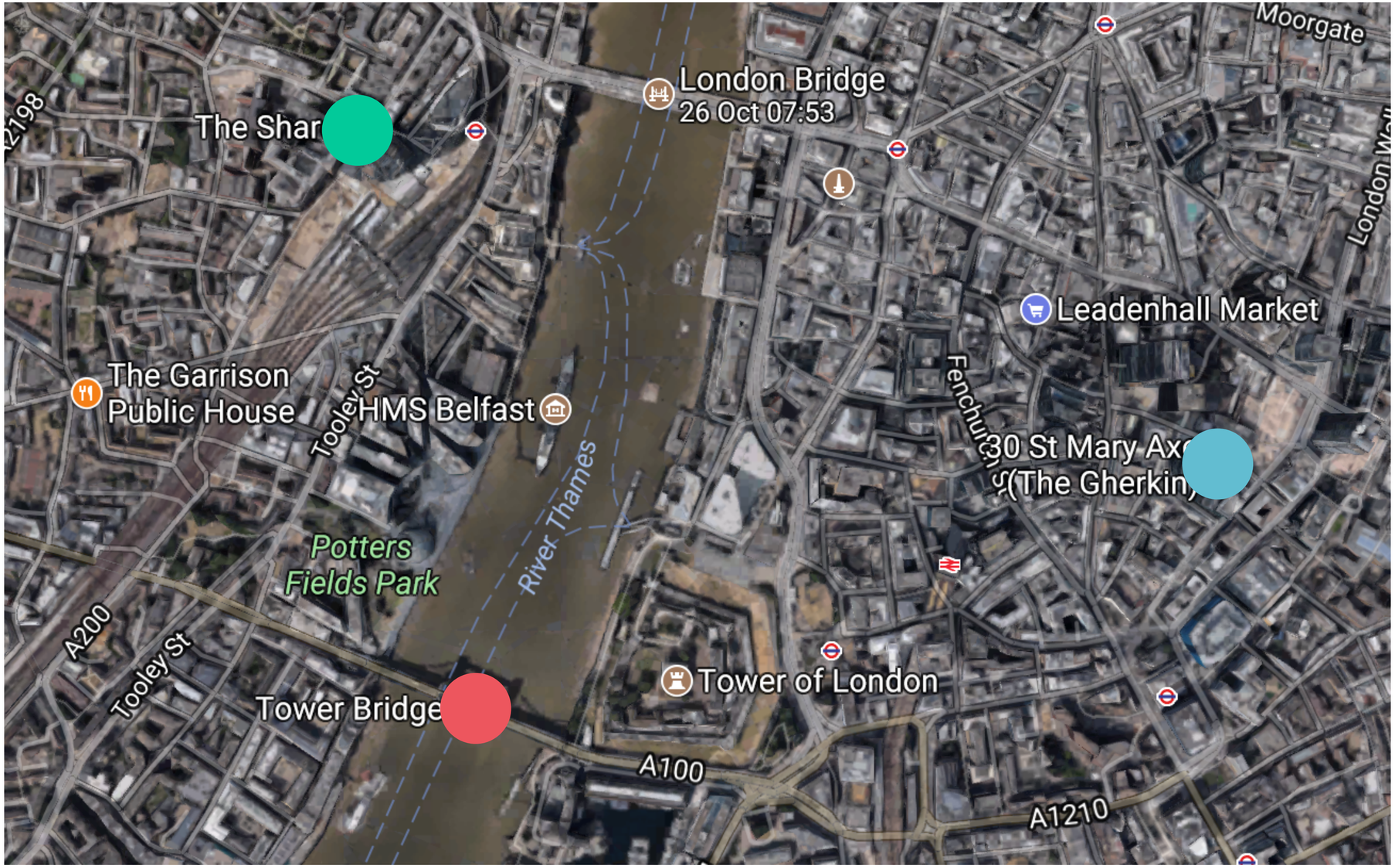
For now: preserve distance

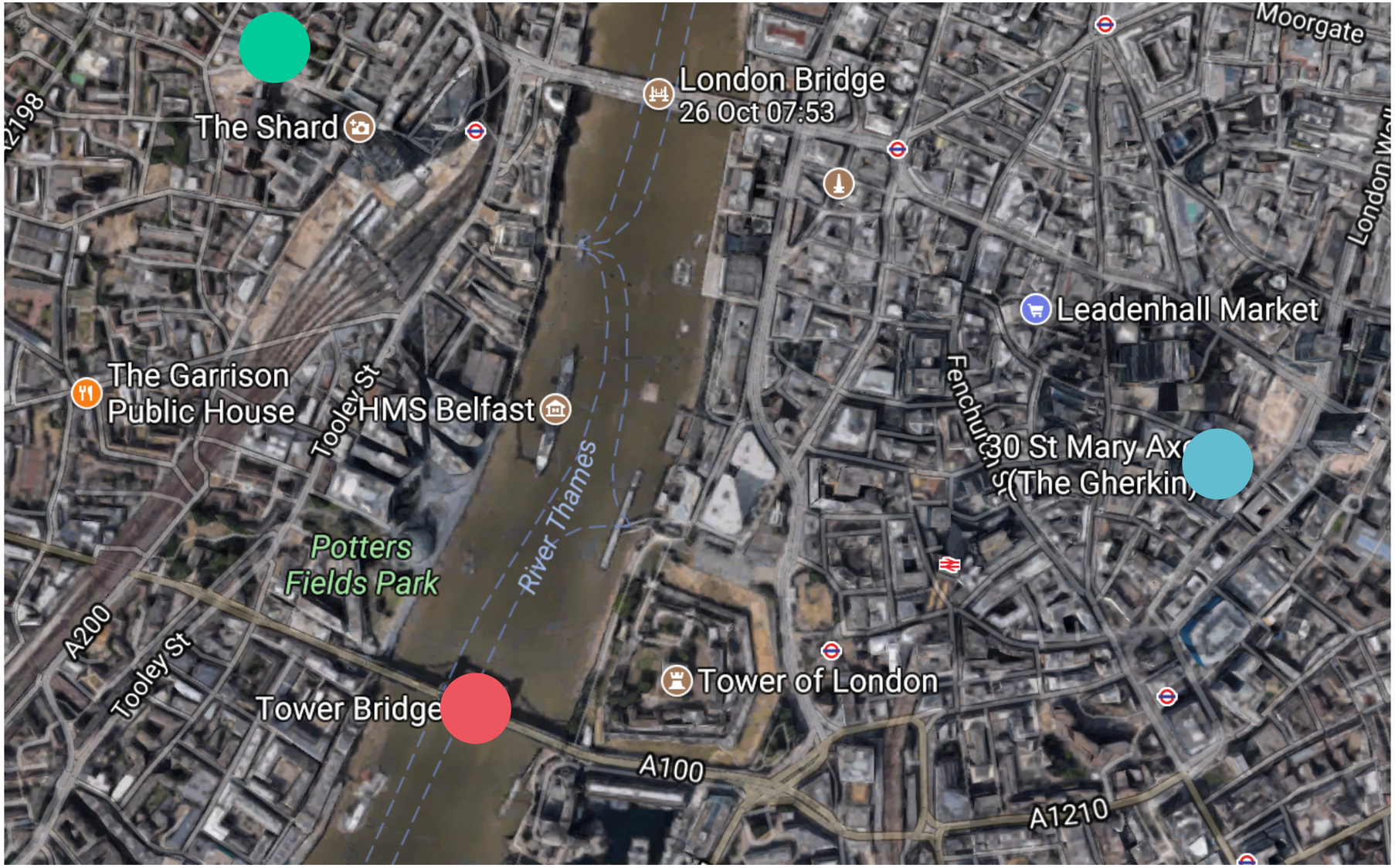

Moorgate
London Bridge
26 Oct 07:53
The Shar
Leadenhall Market
The Garrison
Public House
Tooley St
HMS Belfast
River Thames
Potters
Fields Park
A200
Tooley St
Tower Bridge
Tower of London
A100
A1210


Moorgate
London Bridge
26 Oct 07:53
The Shard
Leadenhall Market
The Garrison
Public House
Tooley St
HMS Belfast
River Thames
Potters
Fields Park
A200
Tooley St
Tower Bridge
Tower of London
A100
A1210






Moorgate
London Bridge
26 Oct 07:53
The Shar
Leadenhall Market
The Garrison
Public House
Tooley St
HMS Belfast
River Thames
Potters
Fields Park
A200
Tooley St
Tower Bridge
Tower of London
A100
A1210


Moorgate
London Bridge
26 Oct 07:53
The Shard
Leadenhall Market
The Garrison
Public House
Tooley St
HMS Belfast
River Thames
Potters
Fields Park
A200
Tooley St
Tower Bridge
Tower of London
A100
A1210

Name	Horsepower	0-60 (secs)	Max speed	Price	MPG	Road tax
Audi S4	340	5.1	180.1	44,015	37	240
MINI Cooper S	75	10.5	100.6	18,780	51	115
Nissan GT-R	565	2.765	209.8	81,350	24	535
Porsche 911 GT3	381	4.2	193.9	77,891	28	520
Renault Clio Williams	142	7.717	133.1	11,585	55	0
Volkswagen Golf R	296	4.94	163.6	31,255	38	190
BMW i8	357	3.6	198.7	105,580	39	440

Name	Horsepower	0-60 (secs)	Max speed	Cost	MPG	Road tax
Audi S4	340	5.1	180.1	4	37	240
MINI Cooper S	75	10.5	100.6	3	51	115
Nissan GT-R	565	2.765	209.8	8	24	535
Porsche 911 GT3	381	4.2	193.9	7	28	520
Renault Clio Williams	142	7.717	133.1	2	55	0
Volkswagen Golf R	296	4.94	163.6	3	38	190
BMW i8	357	3.6	198.7	10	39	440

Name	Horsepower	0-60 (secs)	Max speed	Cost	~~MPG~~	Road tax
Audi S4	340	5.1	180.1	4	~~37~~	240
MINI Cooper S	75	10.5	100.6	2	~~51~~	115
Nissan GT-R	565	2.765	209.8	9	~~24~~	535
Porsche 911 GT3	381	4.2	193.9	8	~~28~~	520
Renault Clio Williams	142	7.717	133.1	1	~~55~~	0
Volkswagen Golf R	296	4.94	163.6	3	~~38~~	190
BMW i8	357	3.6	198.7	10	~~39~~	440

Name	Horsepower	0-60 (secs)	Max speed	Cost	~~MPG~~	~~Road tax~~
Audi S4	340	5.1	180.1	4	~~37~~	~~240~~
MINI Cooper S	75	10.5	100.6	2	~~51~~	~~115~~
Nissan GT-R	565	2.765	209.8	10	~~24~~	~~535~~
Porsche 911 GT3	381	4.2	193.9	9	~~28~~	~~520~~
Renault Clio Williams	142	7.717	133.1	1	~~55~~	~~0~~
Volkswagen Golf R	296	4.94	163.6	3	~~38~~	~~190~~
BMW i8	357	3.6	198.7	10	~~39~~	~~440~~

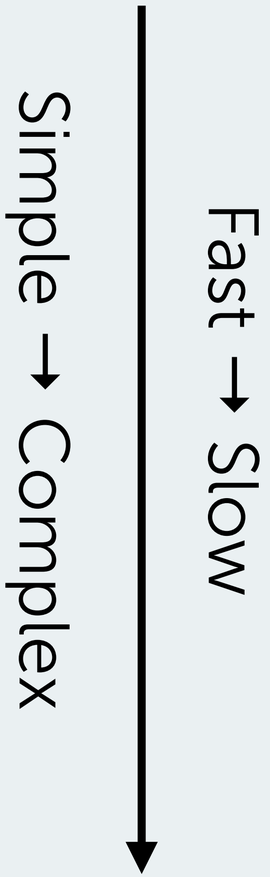
Name	Performance	Cost
Audi S4	6	4
MINI Cooper S	1	2
Nissan GT-R	10	10
Porsche 911 GT3	7	9
Renault Clio Williams	3	1
Volkswagen Golf R	5	3
BMW i8	6	10

Name	~~Performance~~ ???	~~Cost~~ ???
Audi S4	6	4
MINI Cooper S	1	2
Nissan GT-R	10	10
Porsche 911 GT3	7	9
Renault Clio Williams	3	1
Volkswagen Golf R	5	3
BMW i8	6	10

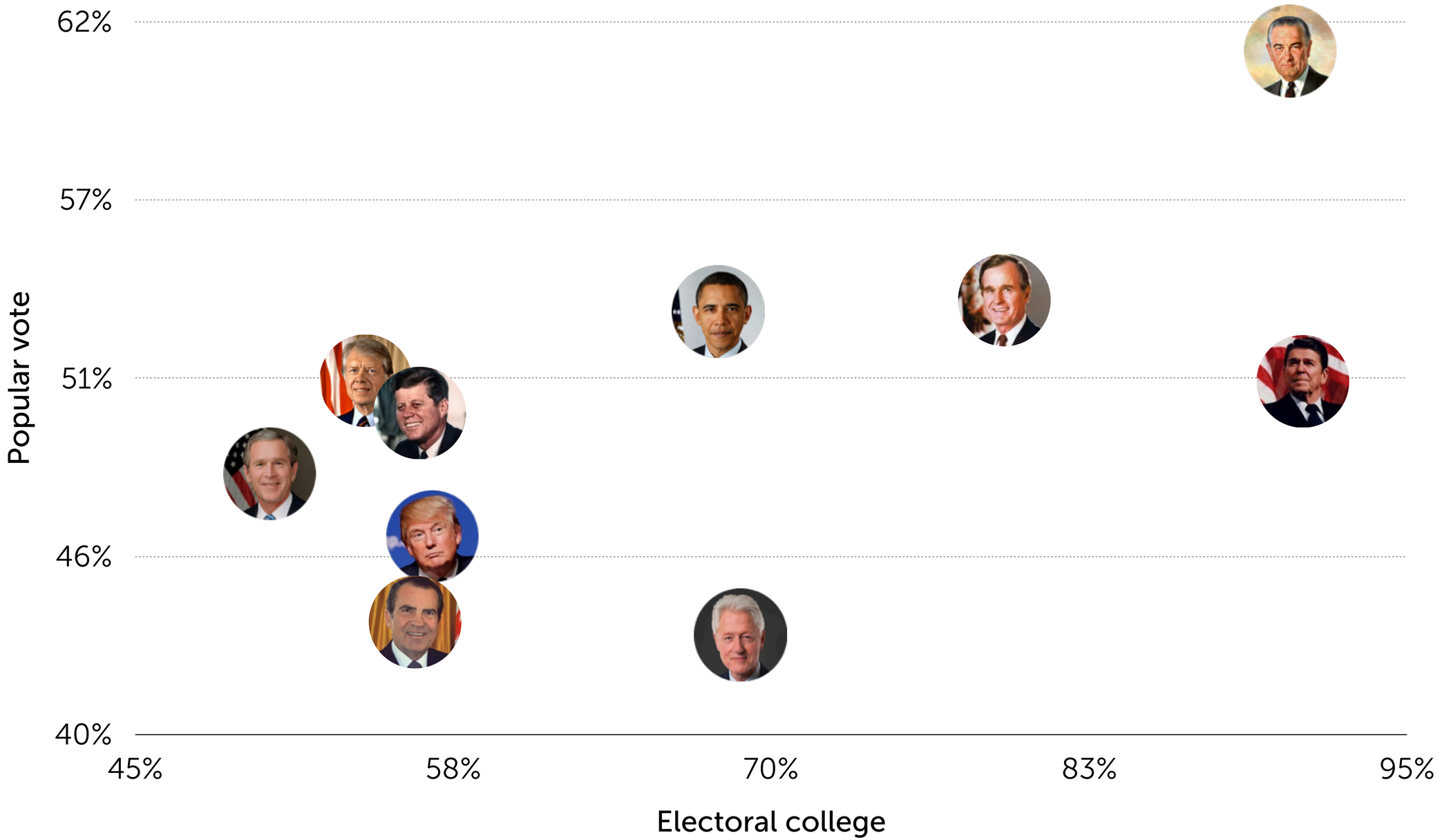
FOUR DIMENSIONALITY REDUCTION TECHNIQUES

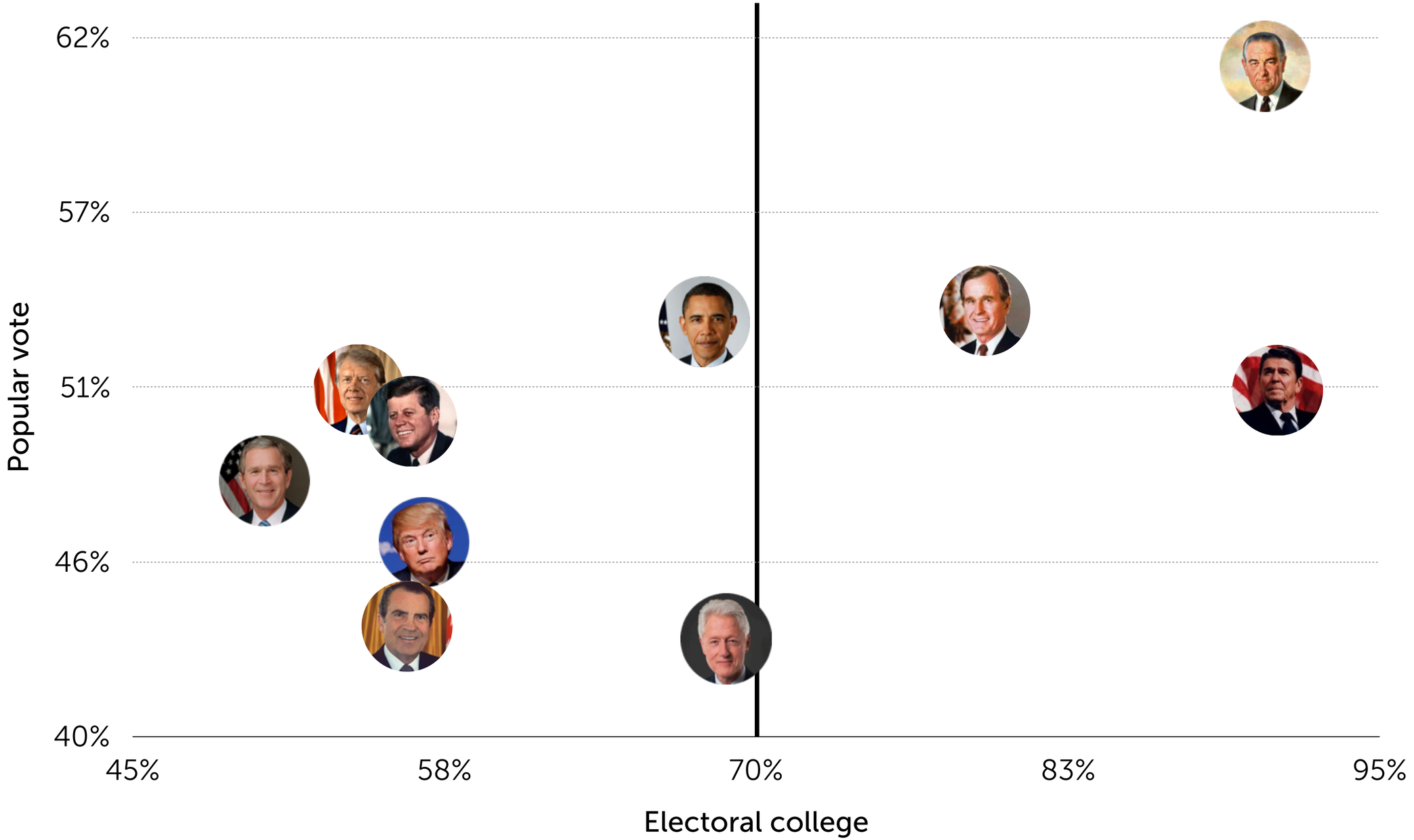
The techniques

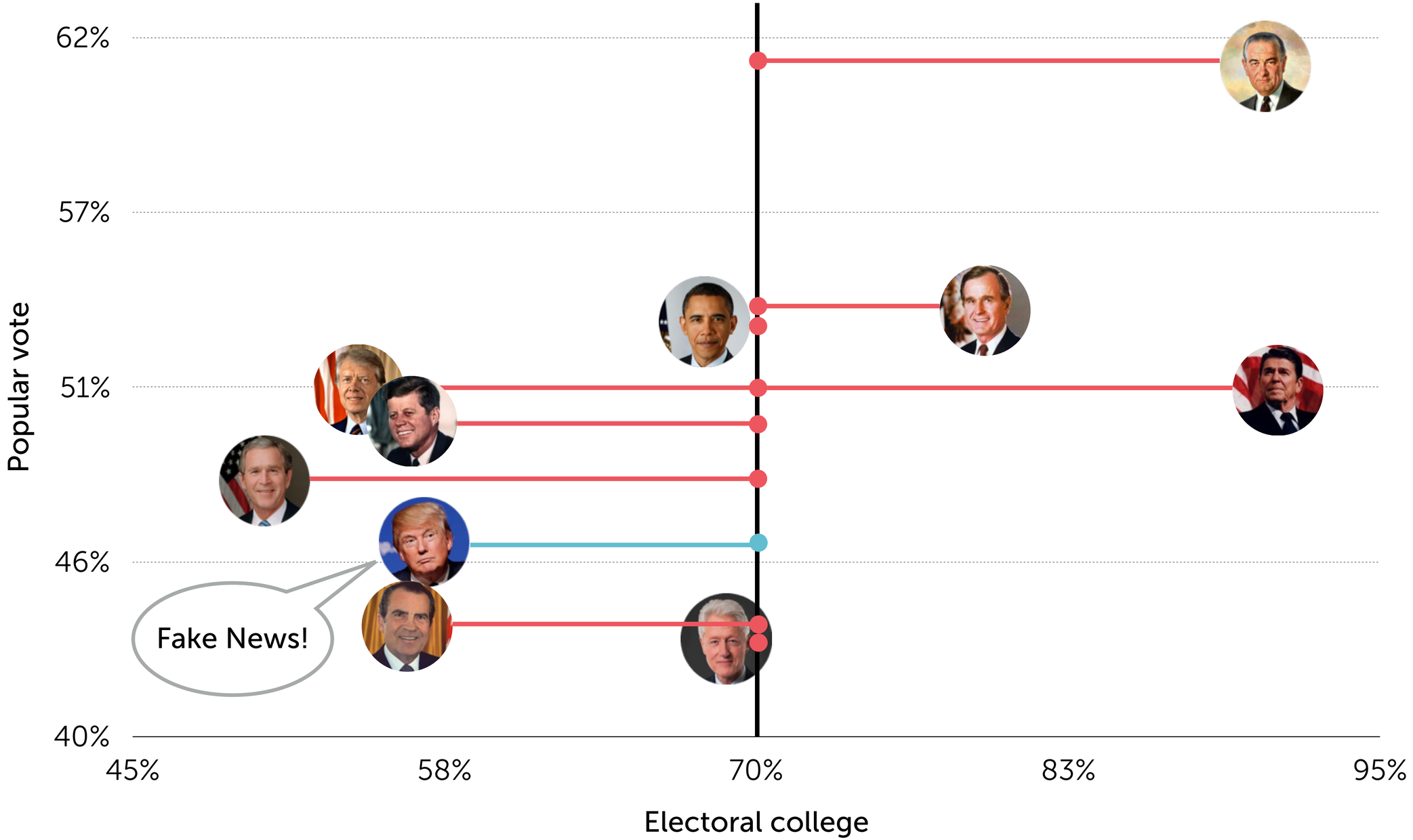
- Random Projection
- Principal Component Analysis
- Isomap
- t-SNE

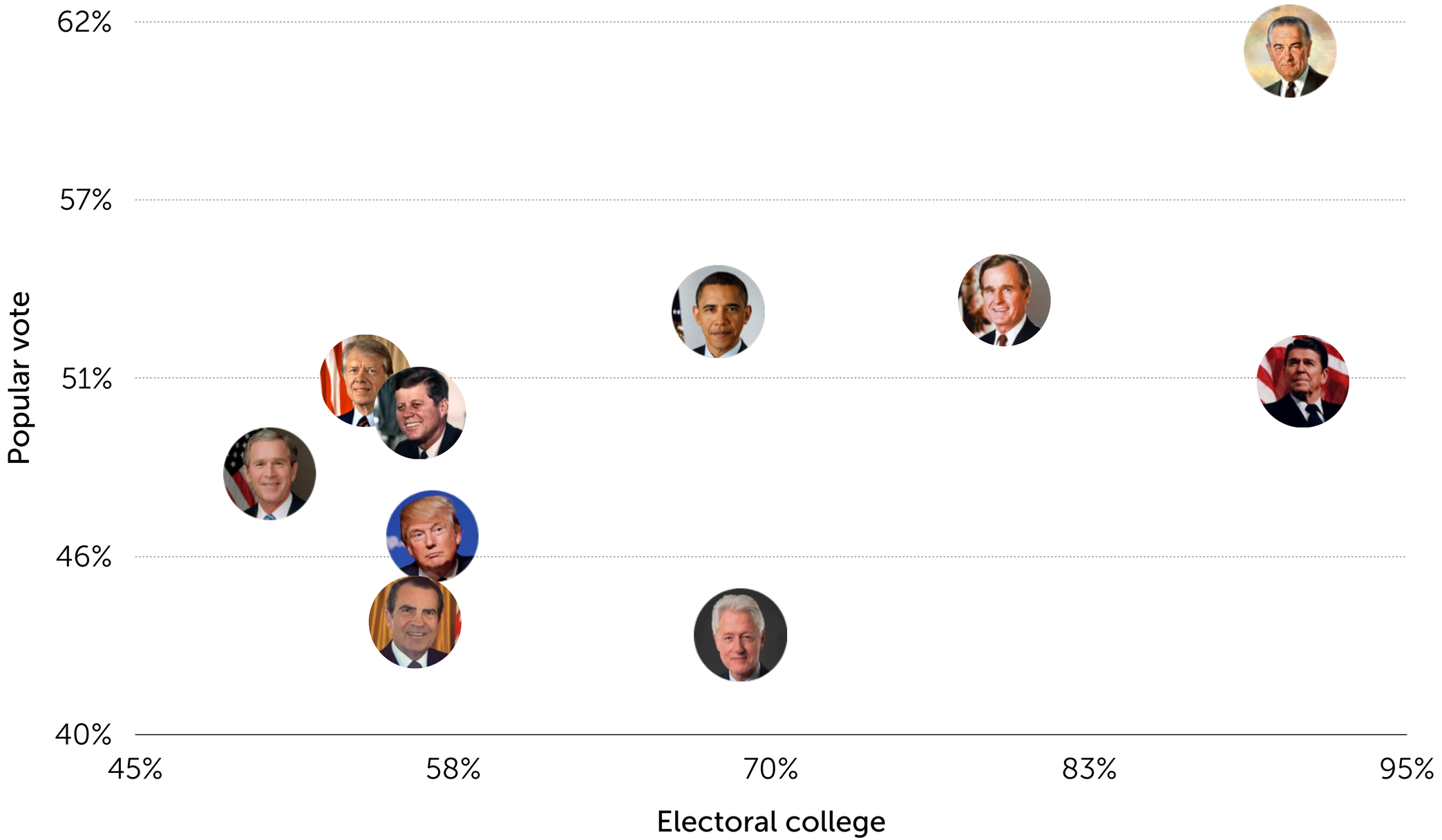


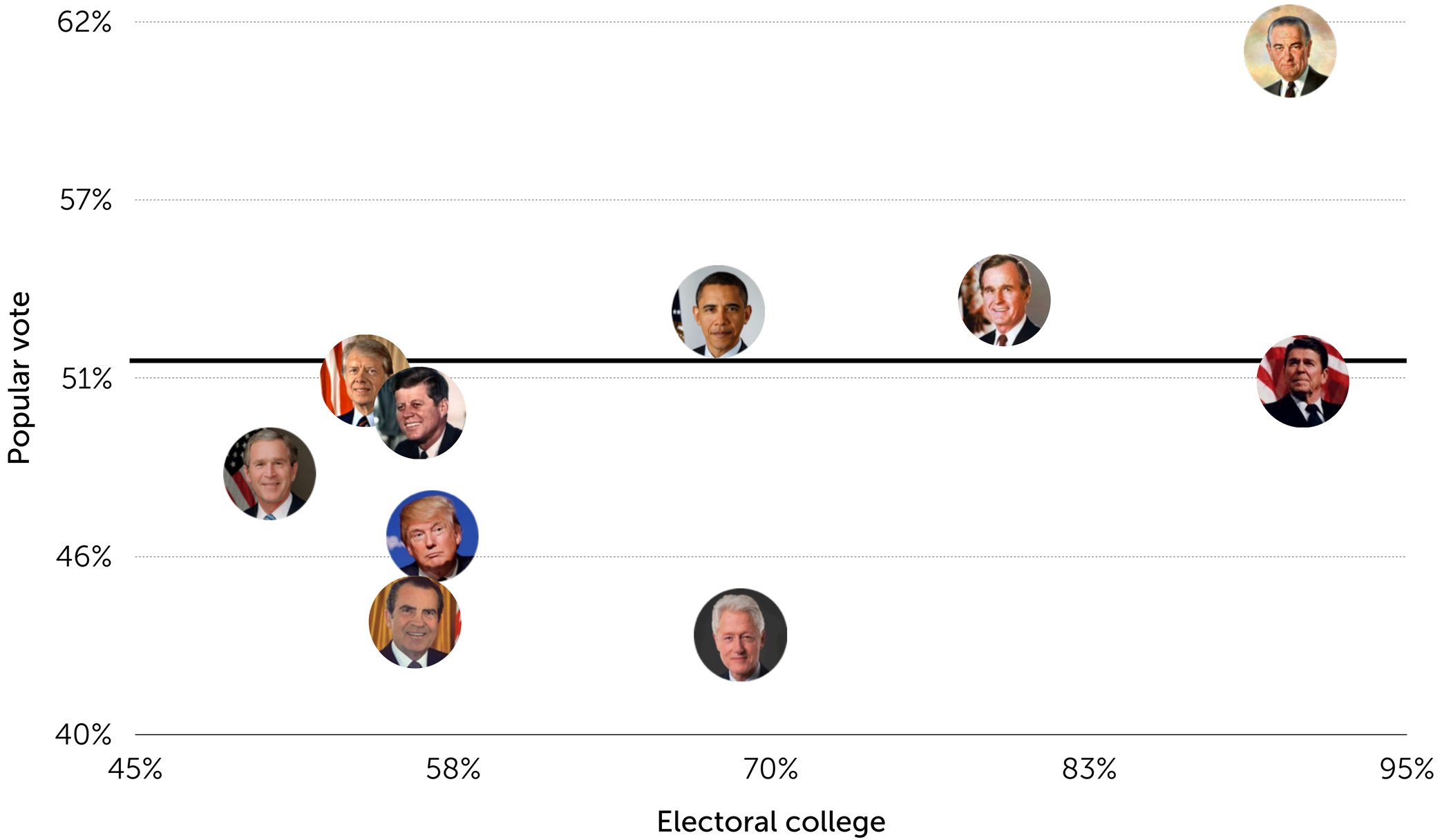
Principal Component Analysis (PCA)

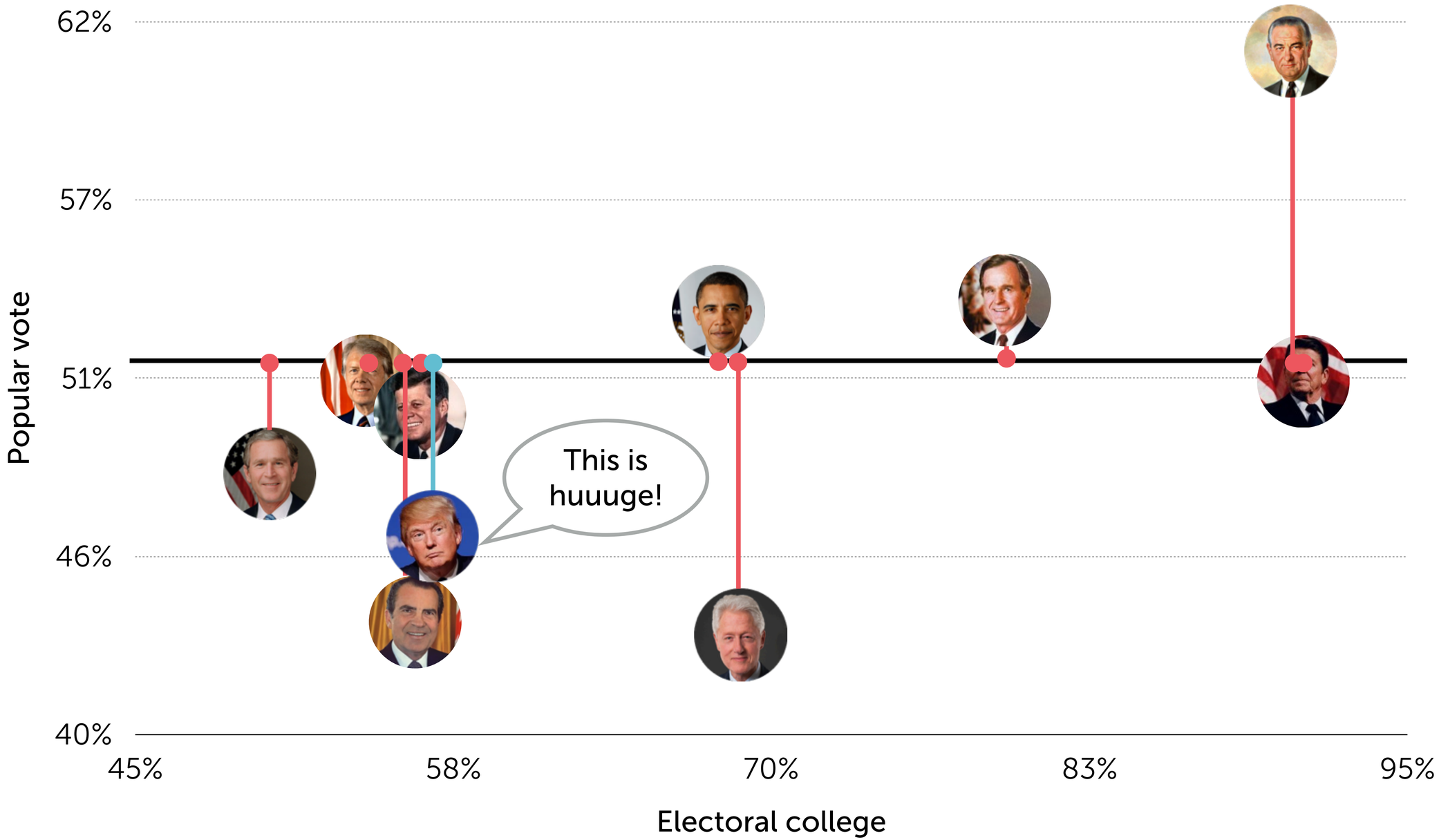

62%
57%
51%
46%
40%
Popular vote
45%
58%
70%
83%
95%
Electoral college

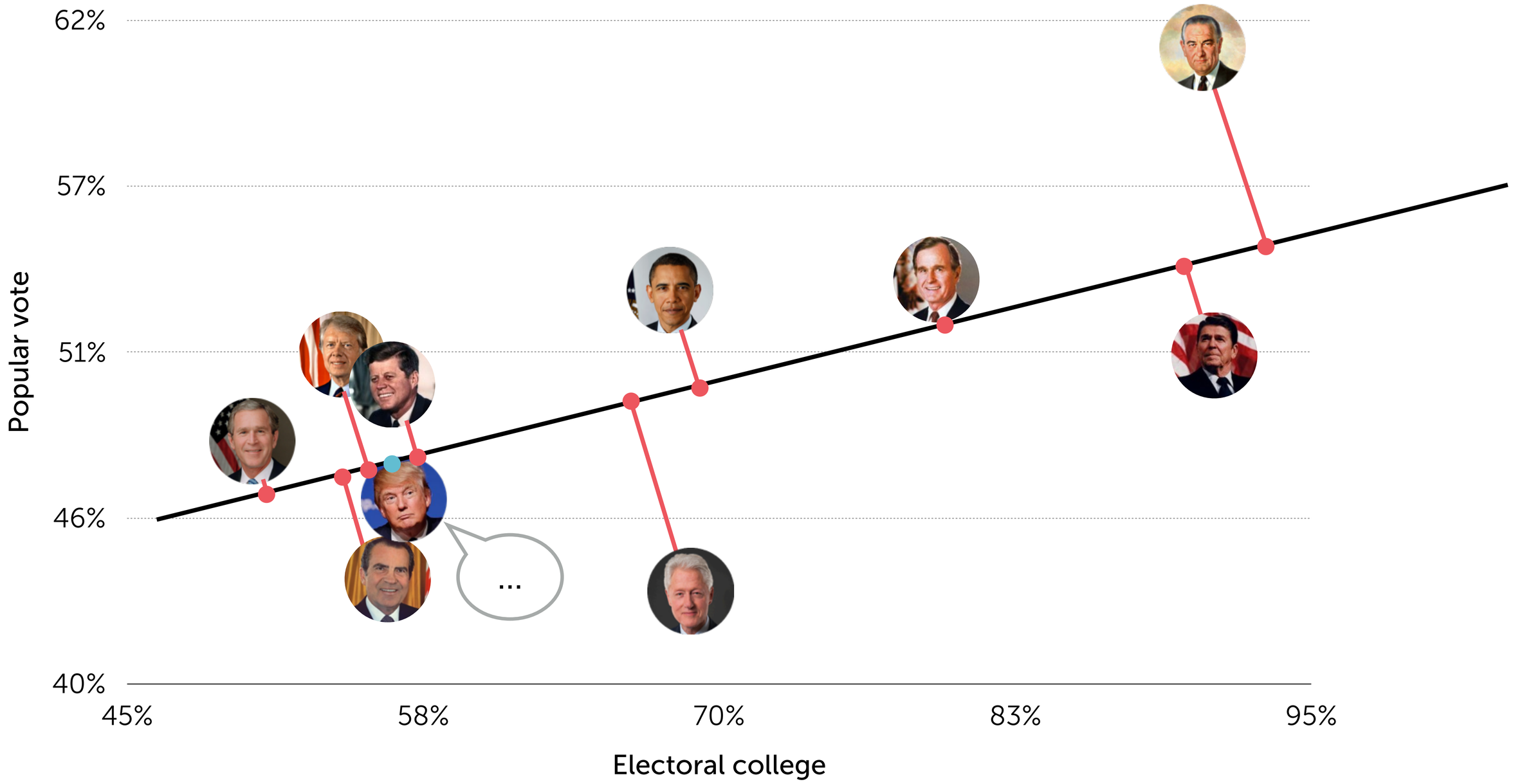

62%
57%
51%
46%
40%
Popular vote
45%
58%
70%
83%
95%
Electoral college


62%
57%
51%
46%
40%
Popular vote
45%
58%
70%
83%
95%
Electoral college
Fake News!

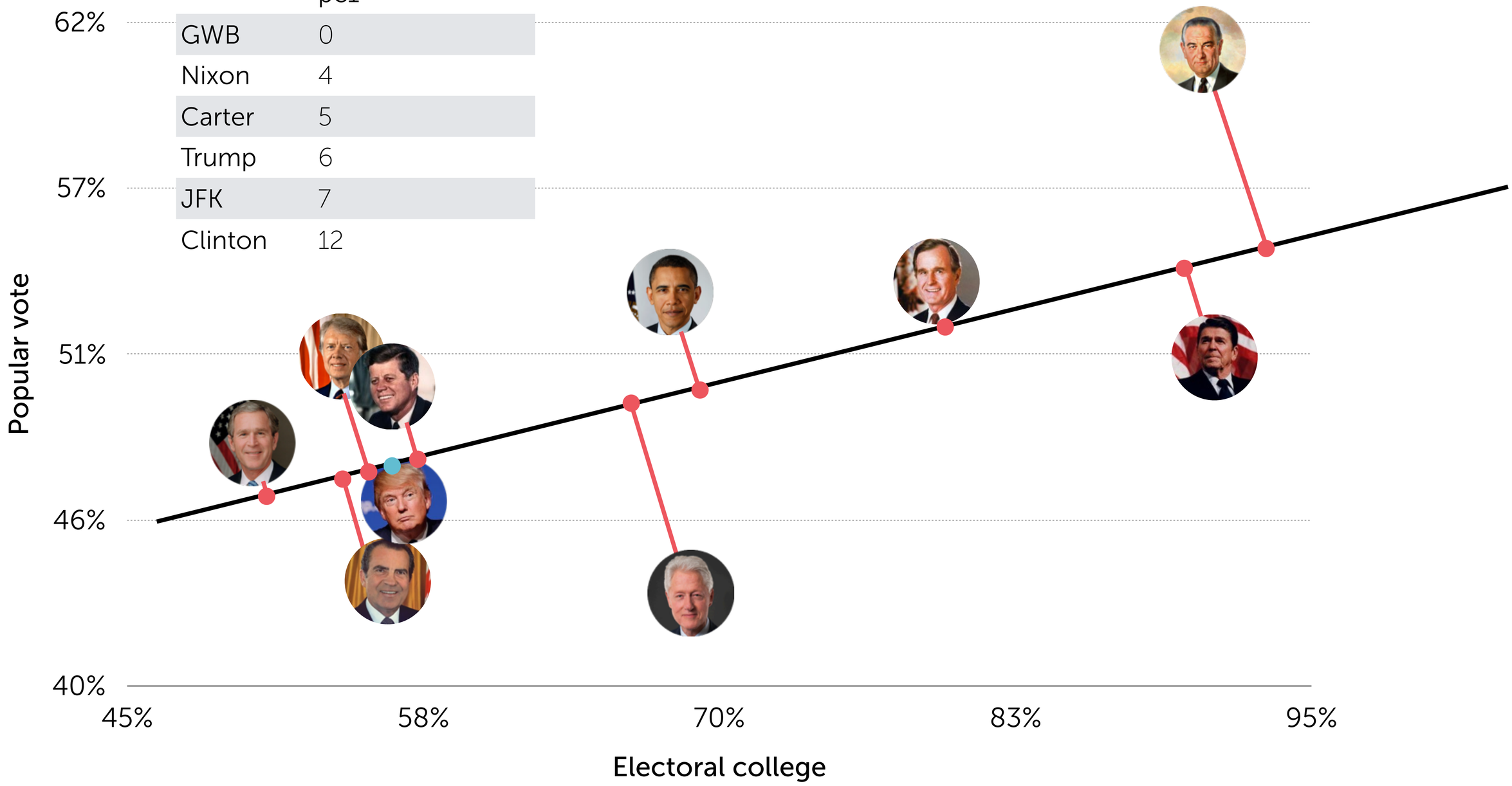

62%
57%
51%
46%
40%
Popular vote
45%
58%
70%
83%
95%
Electoral college


62%
57%
51%
46%
40%
Popular vote
45%
58%
70%
83%
95%
Electoral college

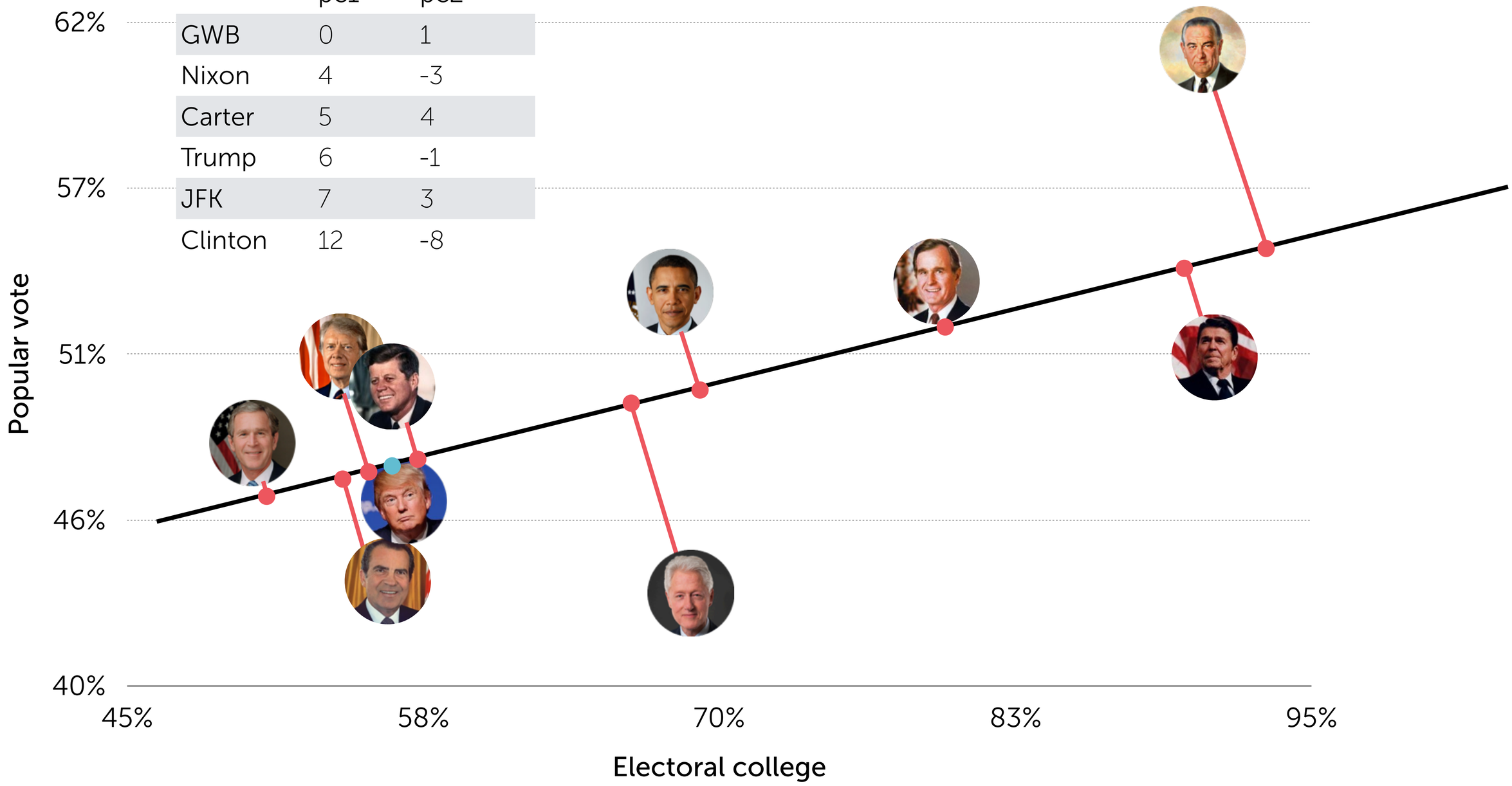

62%
57%
51%
46%
40%
Popular vote
This is huuuge!
45%
58%
70%
83%
95%
Electoral college


62%
57%
51%
46%
40%
Popular vote
45%
58%
70%
83%
95%
Electoral college
...

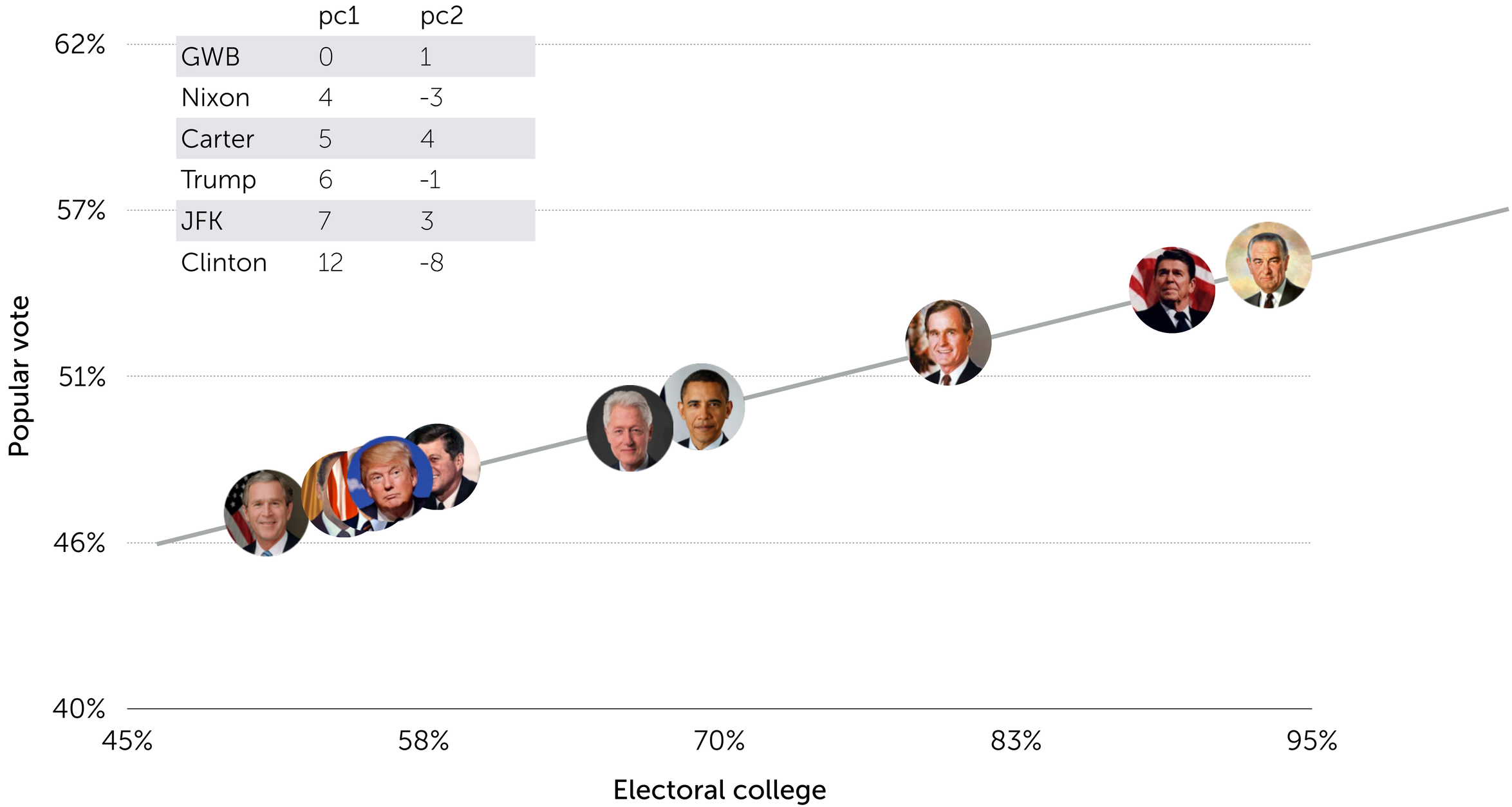
	pc1
GWB	0
Nixon	4
Carter	5
Trump	6
JFK	7
Clinton	12

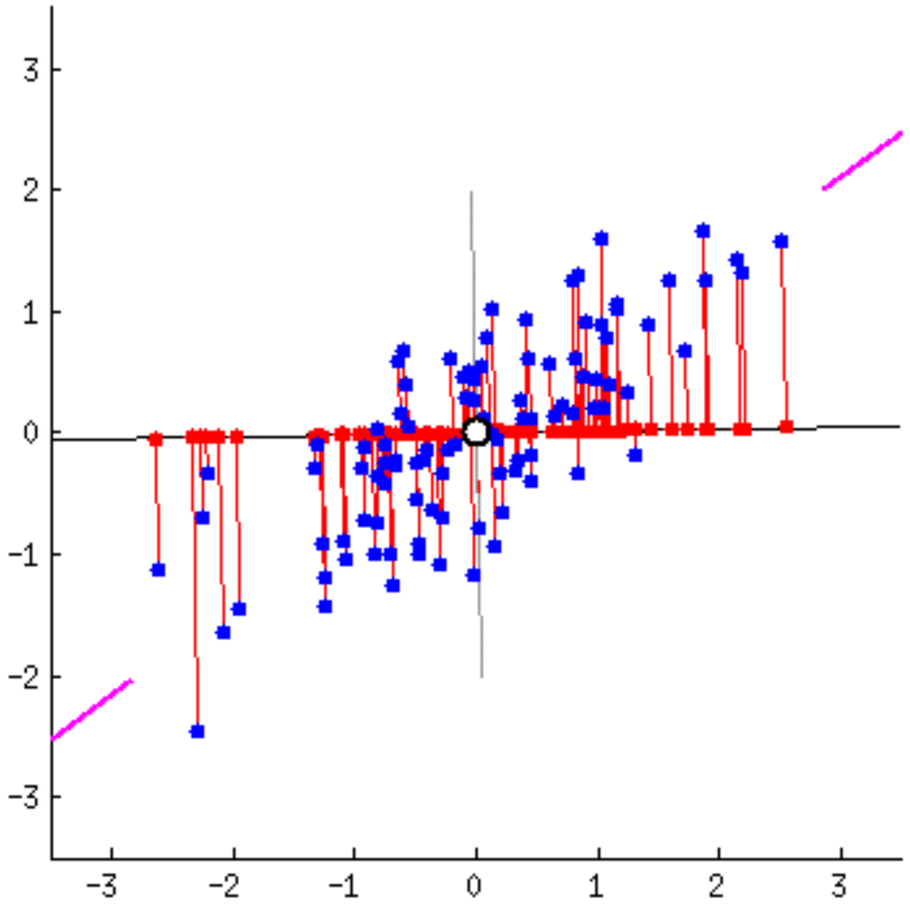


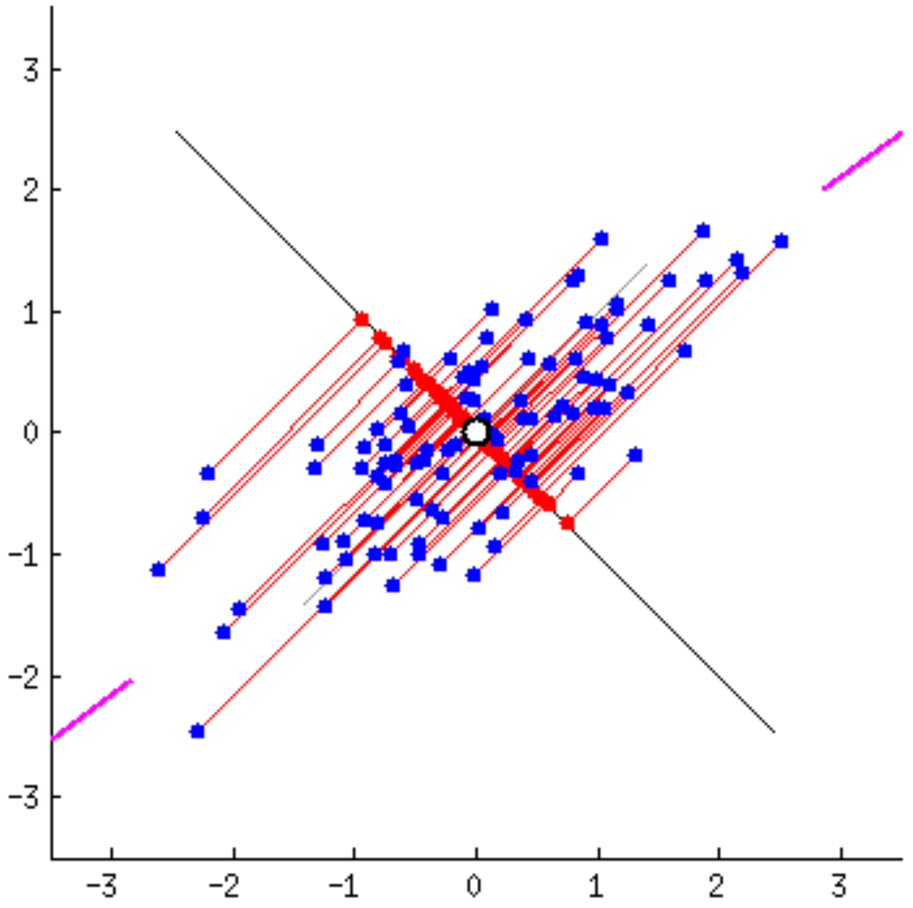
	pc1	pc2
GWB	0	1
Nixon	4	-3
Carter	5	4
Trump	6	-1
JFK	7	3
Clinton	12	-8



	pc1	pc2
GWB	0	1
Nixon	4	-3
Carter	5	4
Trump	6	-1
JFK	7	3
Clinton	12	-8




3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3


3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3

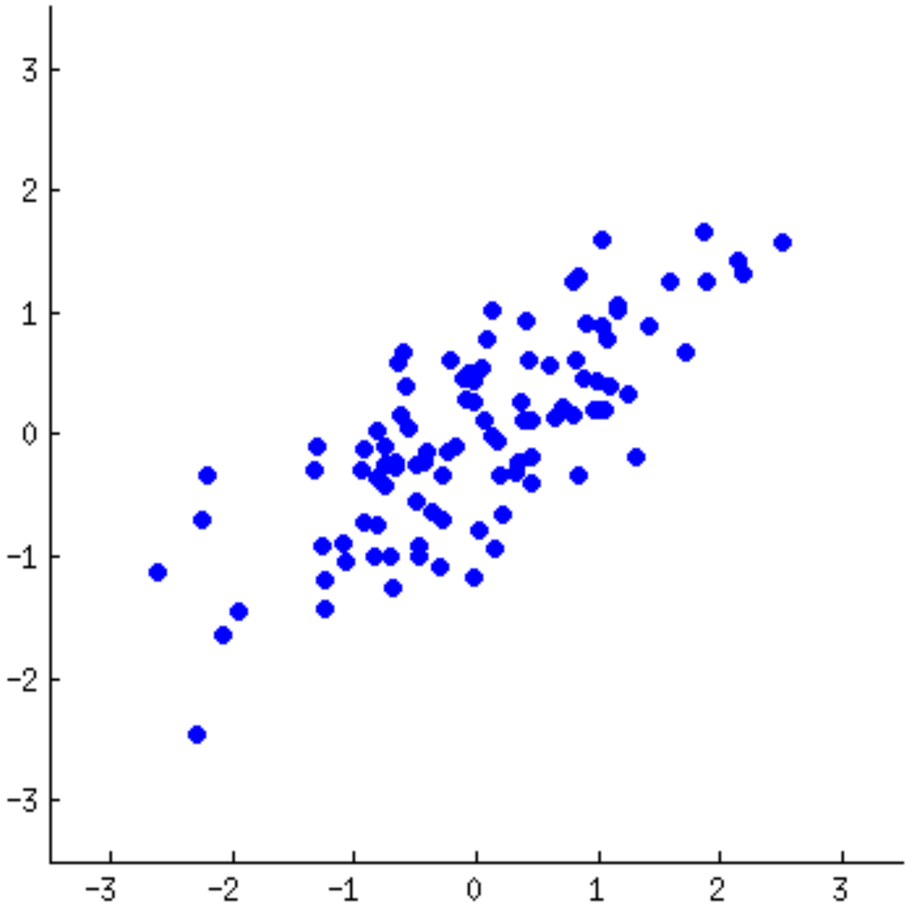
Principal Component Analysis

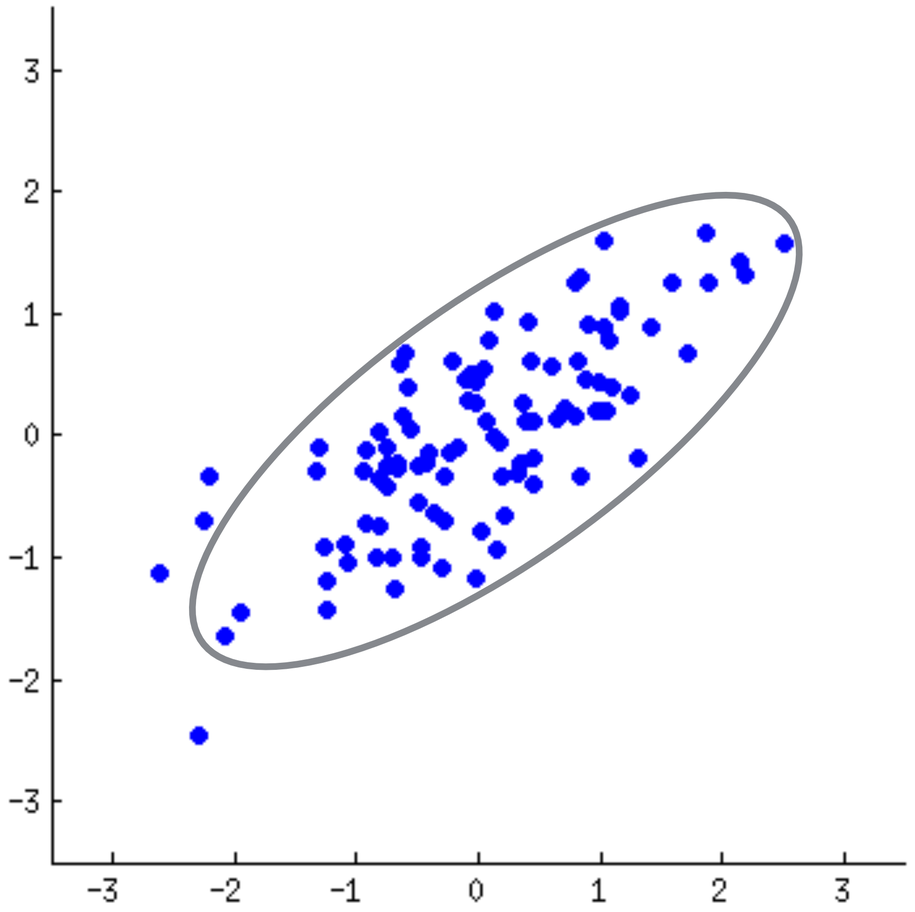
Accomplished with

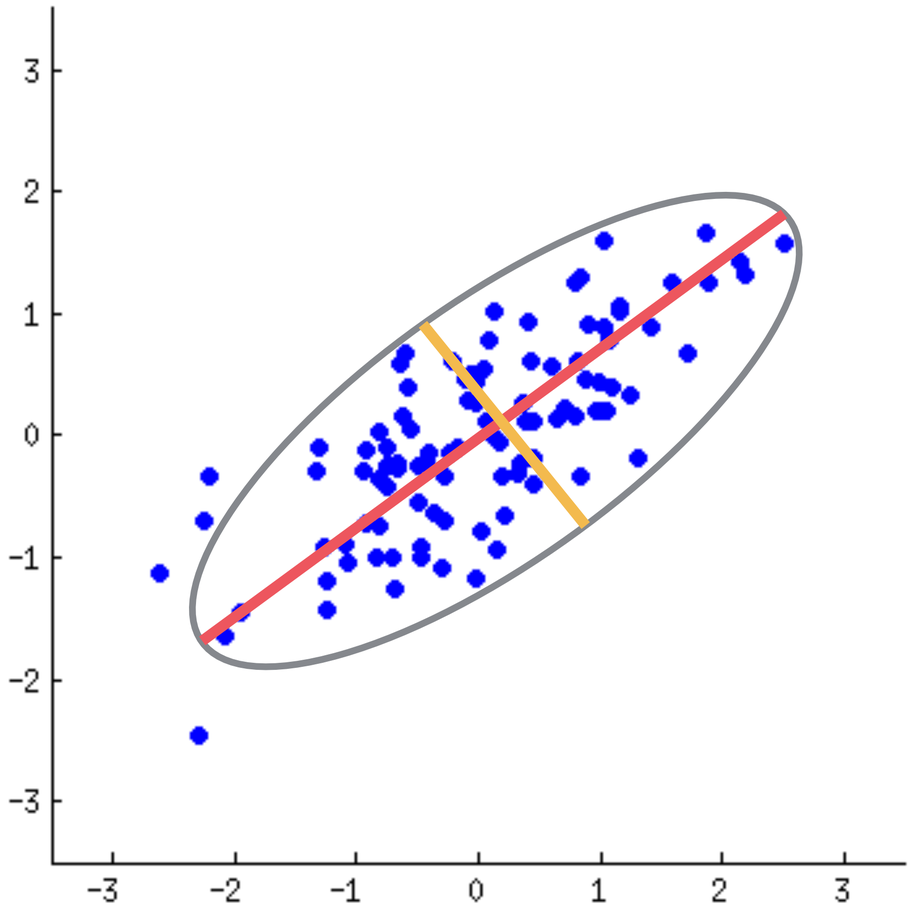
Singular-value decomposition

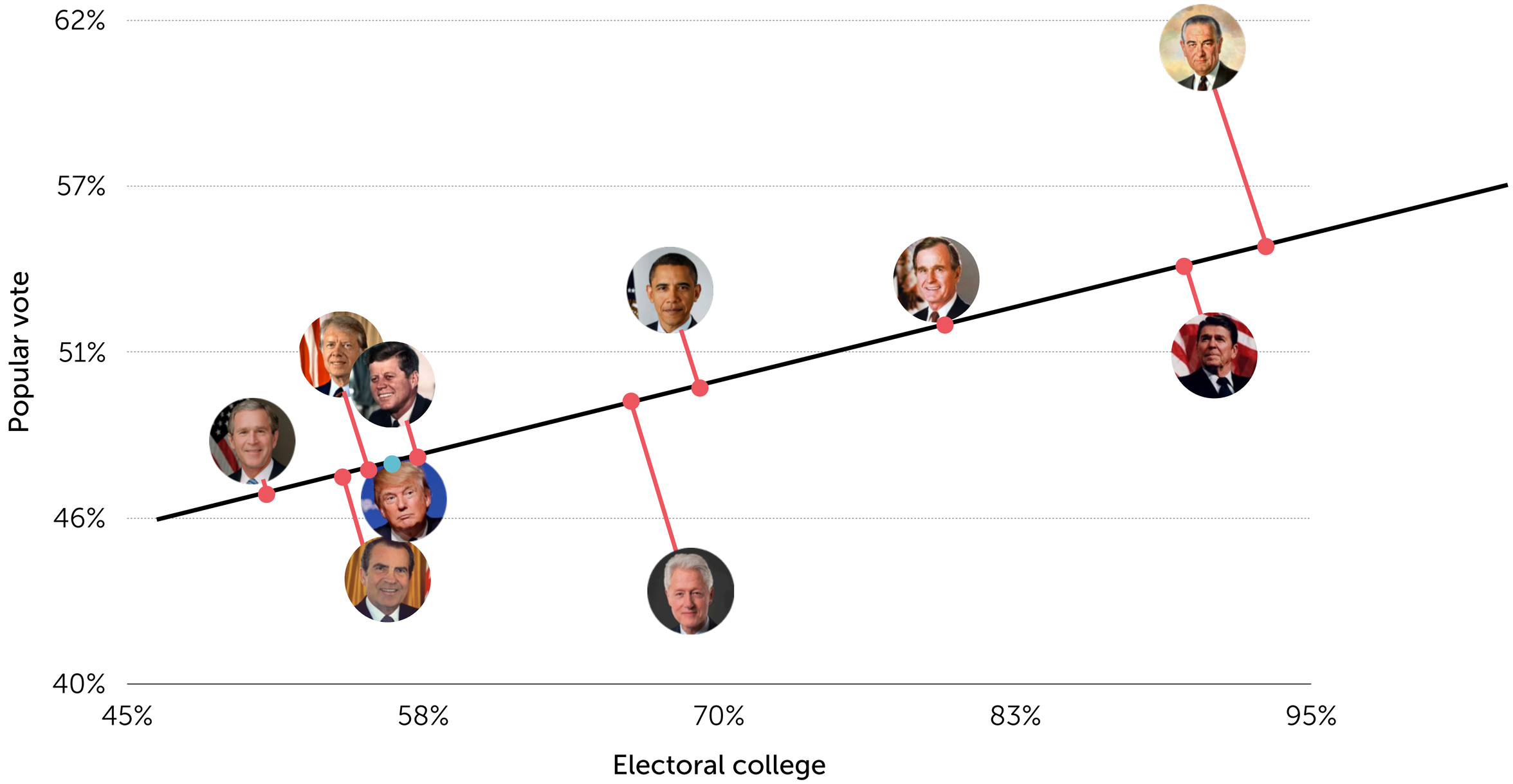
Generalization of

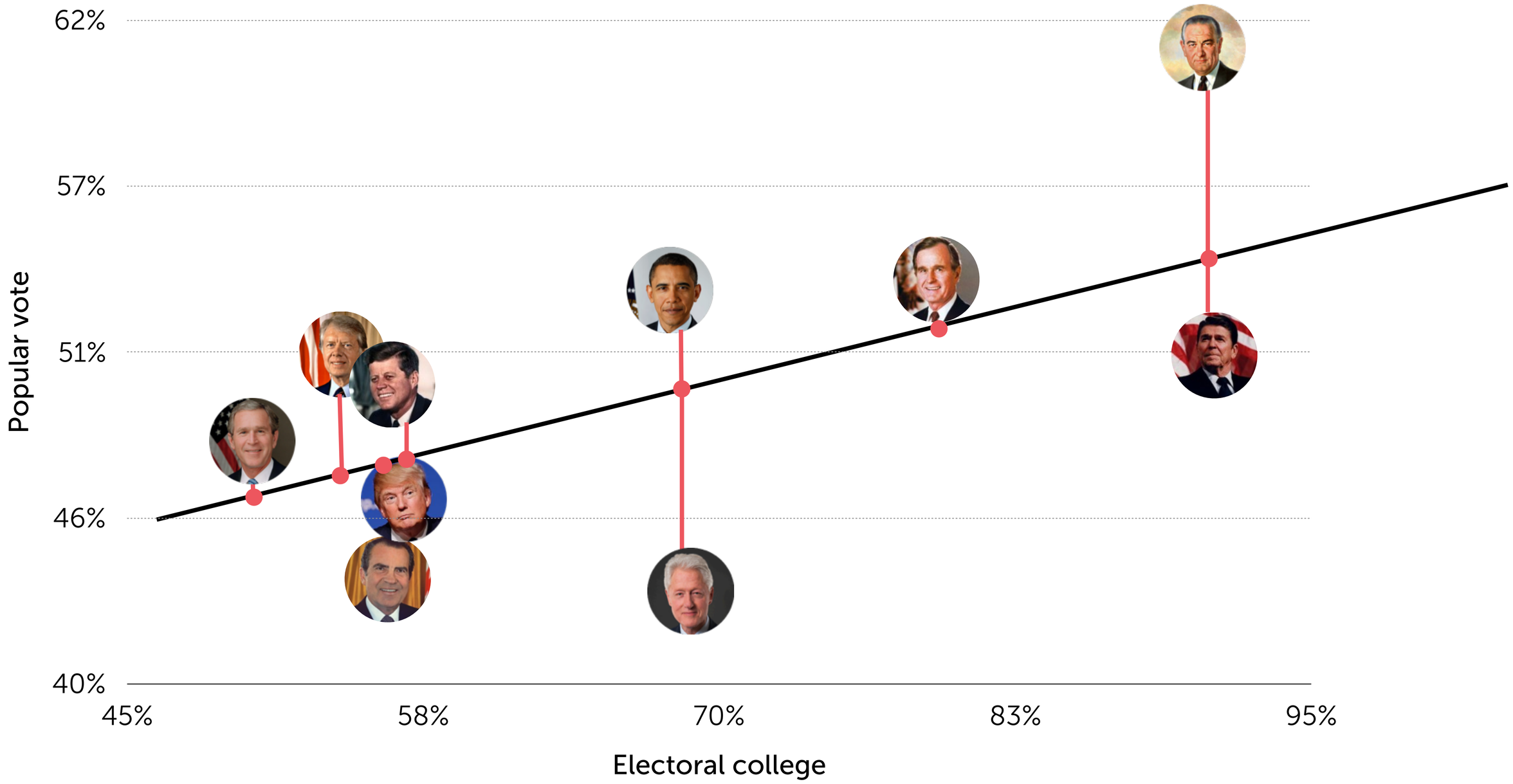
Eigendecomposition

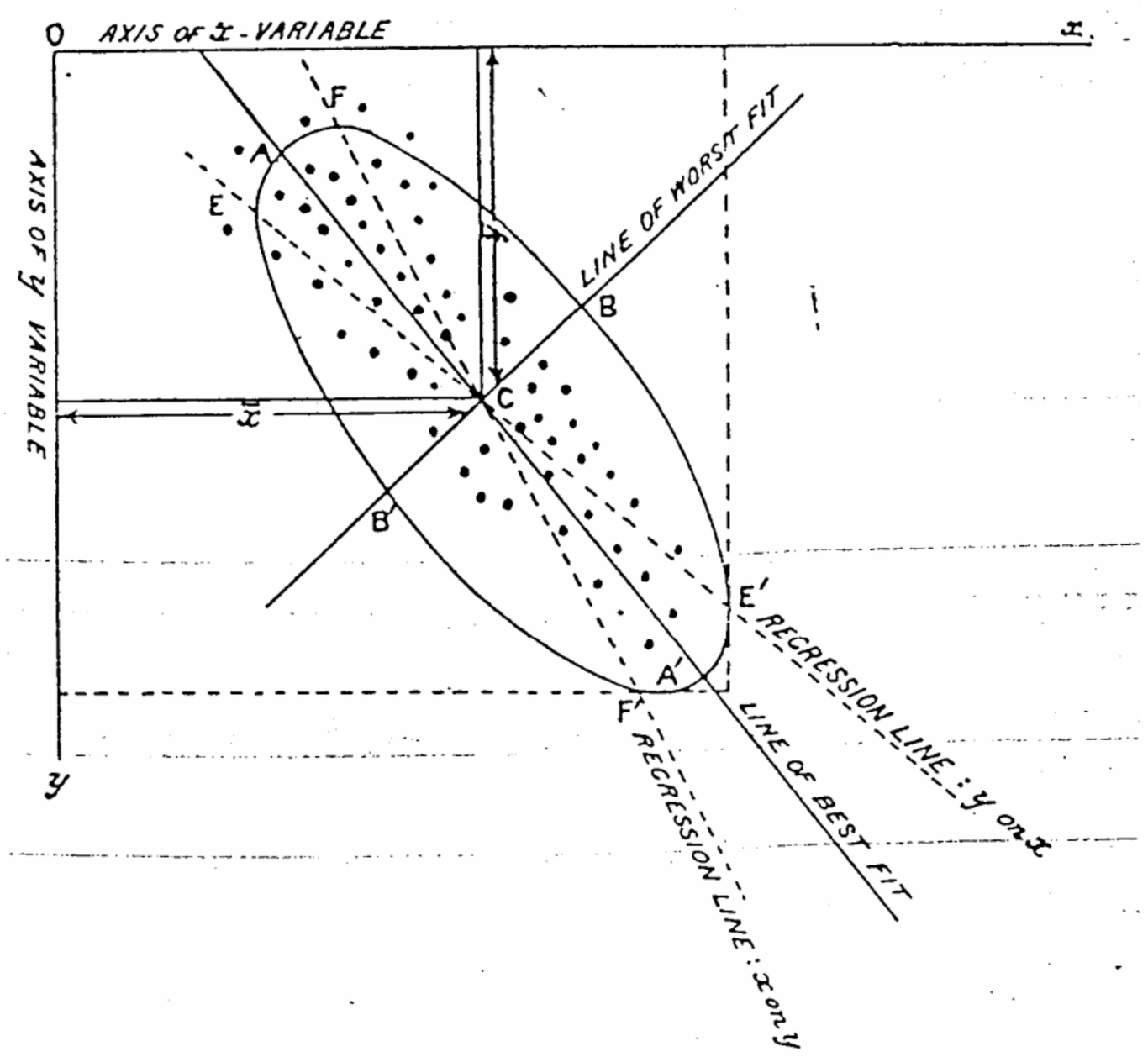

3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3


3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3


3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3

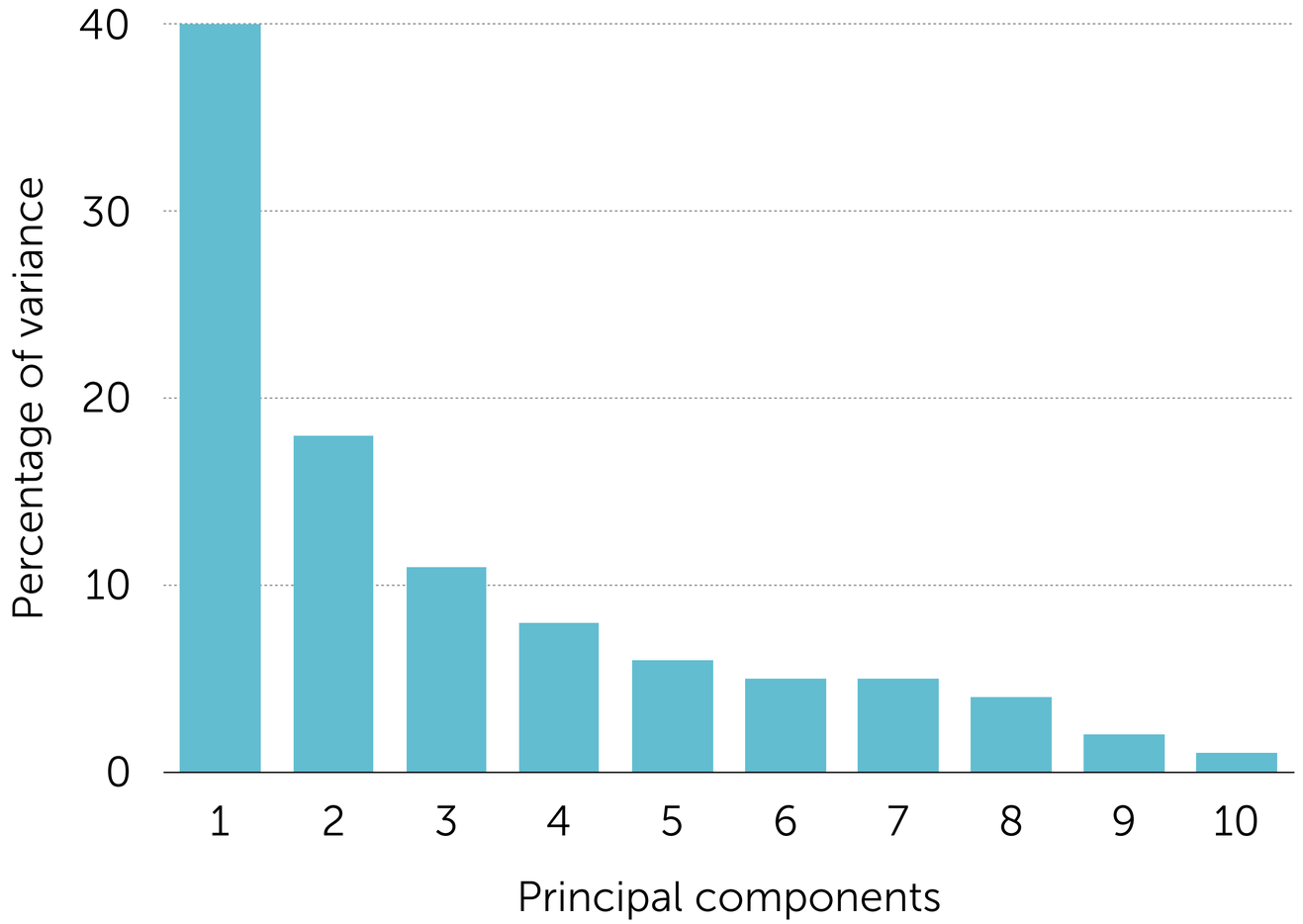

62%
57%
51%
46%
40%
Popular vote
45%
58%
70%
83%
95%
Electoral college


62%
57%
51%
46%
40%
Popular vote
45%
58%
70%
83%
95%
Electoral college


O
AXIS OF x - VARIABLE
x
AXIS OF y VARIABLE
y
F
A
E
LINE OF WORST FIT
B
C
x̄
B'
E'
REGRESSION LINE : y on x
A'
LINE OF BEST FIT
F'
REGRESSION LINE : x on y

	pc1	pc2
GWB	0	1
Nixon	4	-3
Carter	5	4
Trump	6	-1
JFK	7	3
Clinton	12	-8
	...	

Scree plot

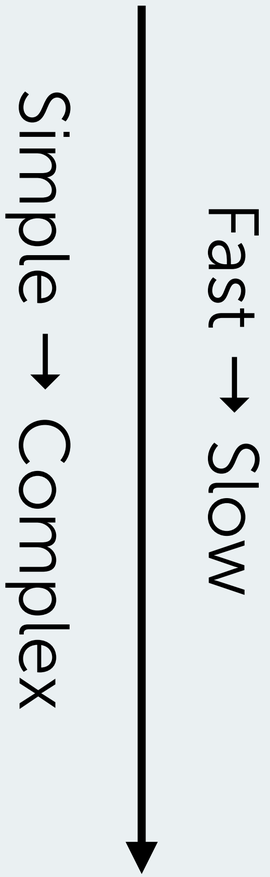

40
30
20
10
0
Percentage of variance
1
2
3
4
5
6
7
8
9
10
Principal components

PCA - Summary

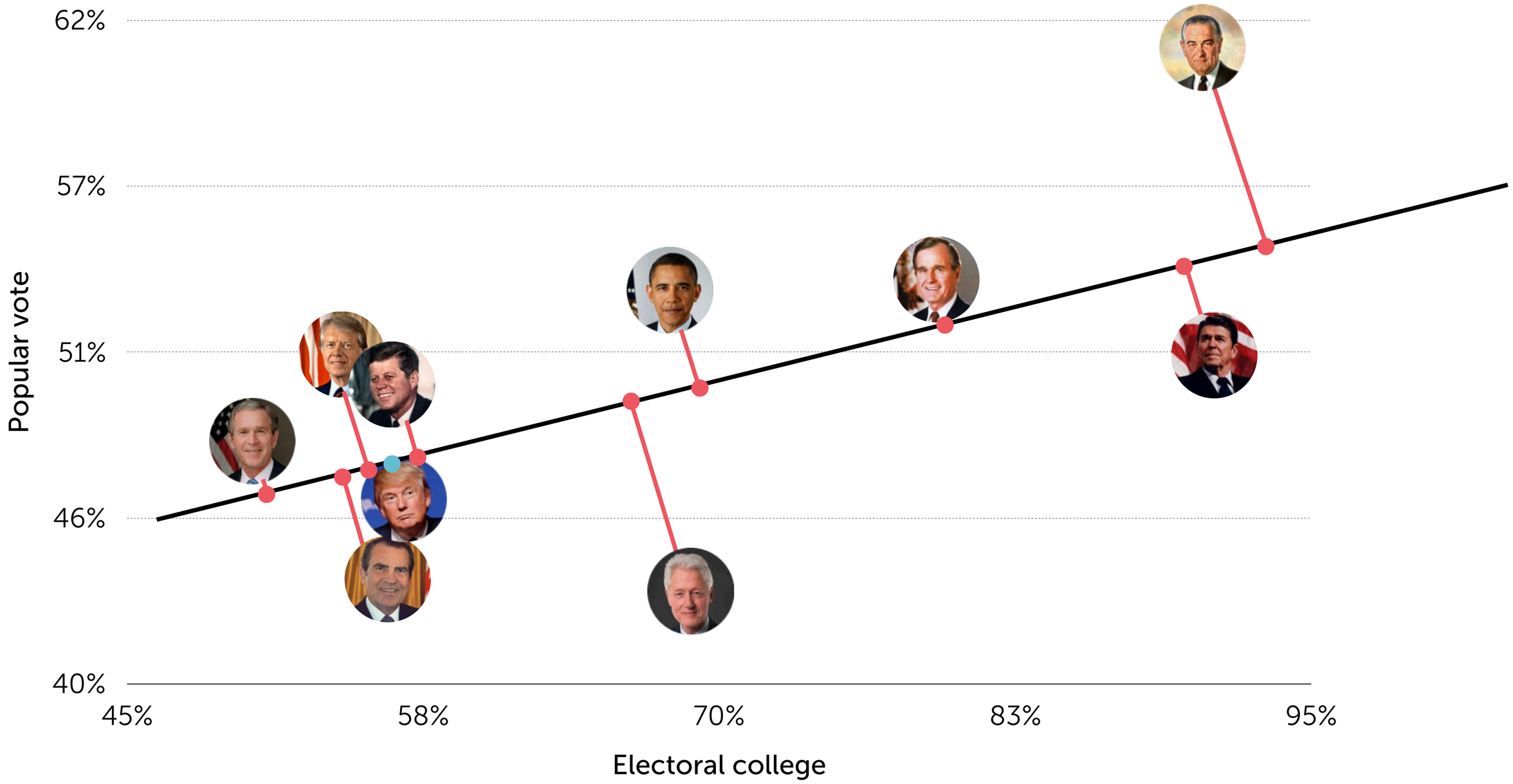
- Preserves: Pairwise distances
- Pros: fast, choose how many dimensions to keep, easy to understand the algorithm
- Cons: relies on data being linear, can still be expensive with large datasets, PCA w/ SVD not very robust to outliers

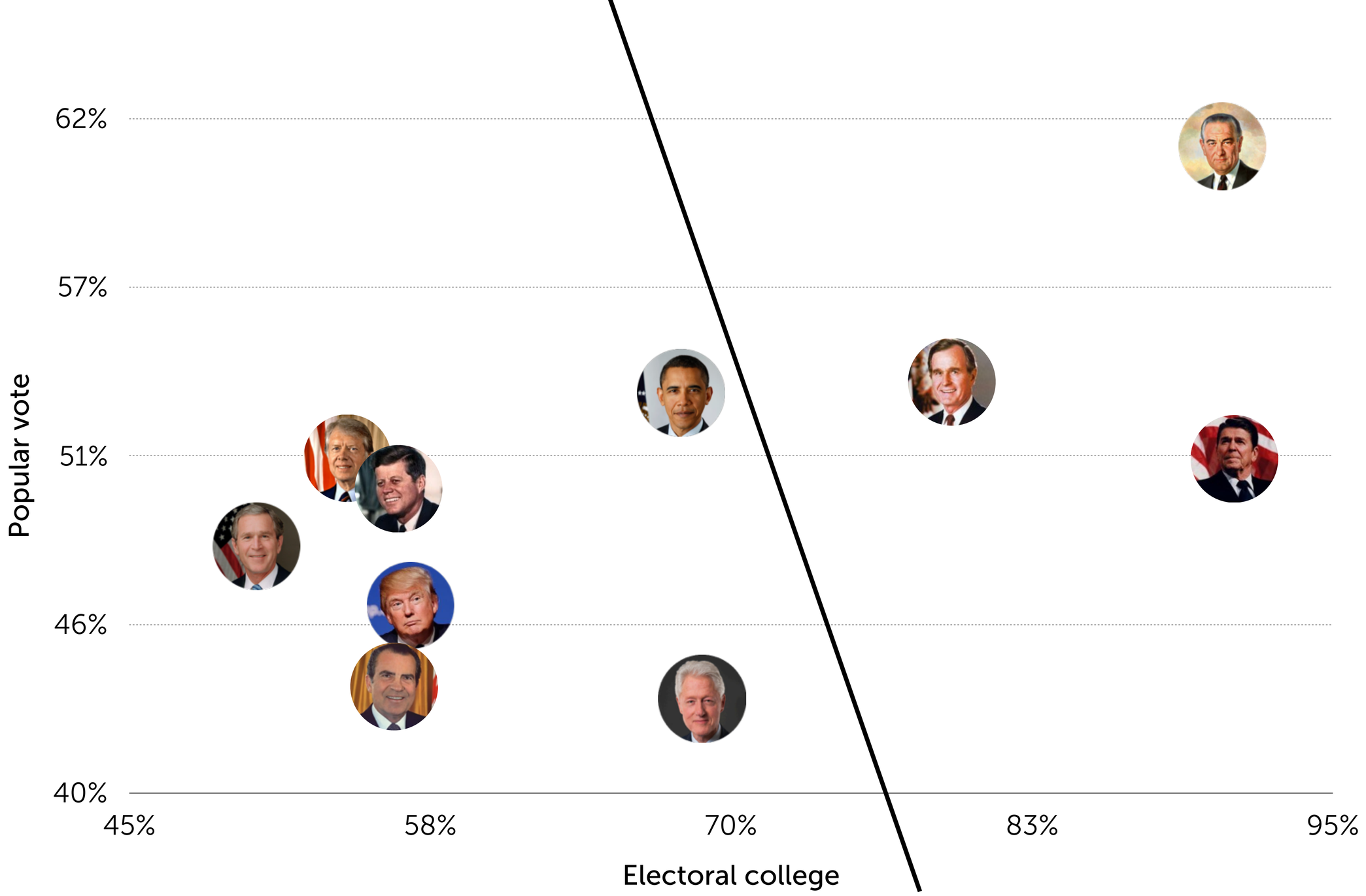
The techniques

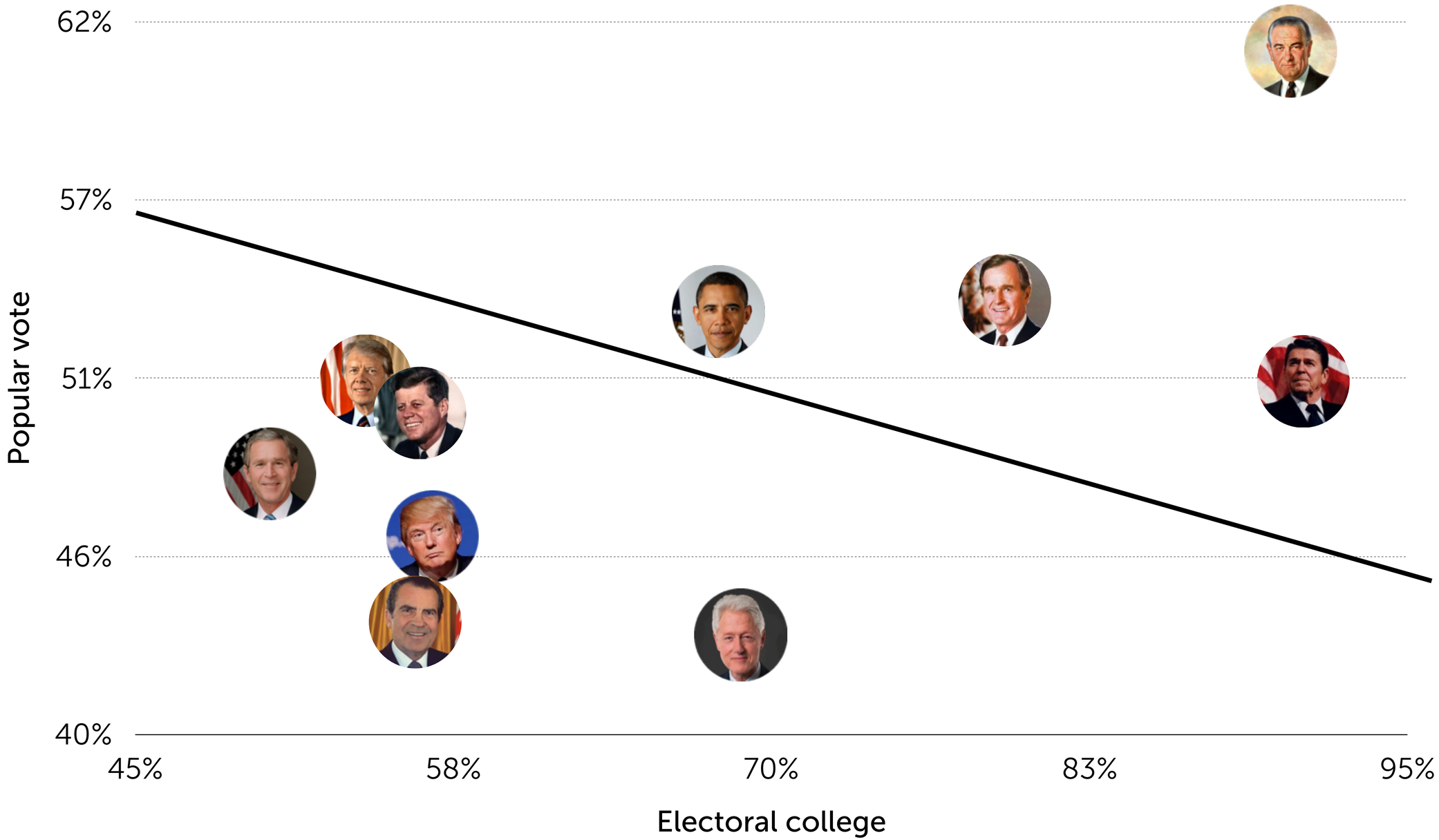
- Random Projection
- ✅ Principal Component Analysis
- Isomap
- t-SNE

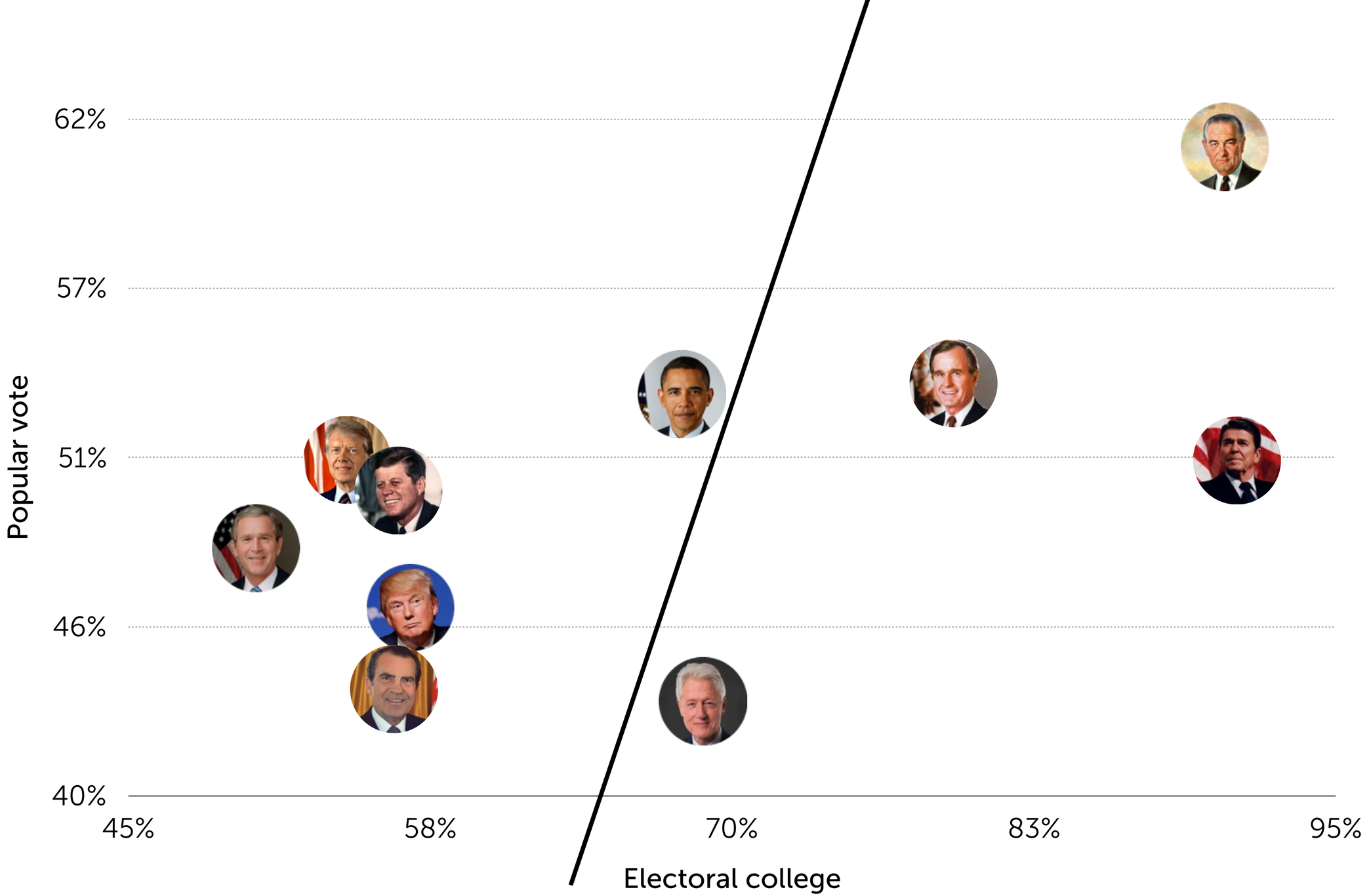


Random Projection


62%
57%
51%
46%
40%
Popular vote
45%
58%
70%
83%
95%
Electoral college


62%
57%
51%
46%
40%
Popular vote
45%
58%
70%
83%
95%
Electoral college


62%
57%
51%
46%
40%
Popular vote
45%
58%
70%
83%
95%
Electoral college


62%
57%
51%
46%
40%
Popular vote
45%
58%
70%
83%
95%
Electoral college

Johnson–Lindenstrauss lemma

Given $0 < \varepsilon < 1$, a set X of m points in $\mathbb{R}^N$, and a number $n > 8\ln(m)/\varepsilon^2$, there is a linear map $f : \mathbb{R}^N \to \mathbb{R}^n$ such that

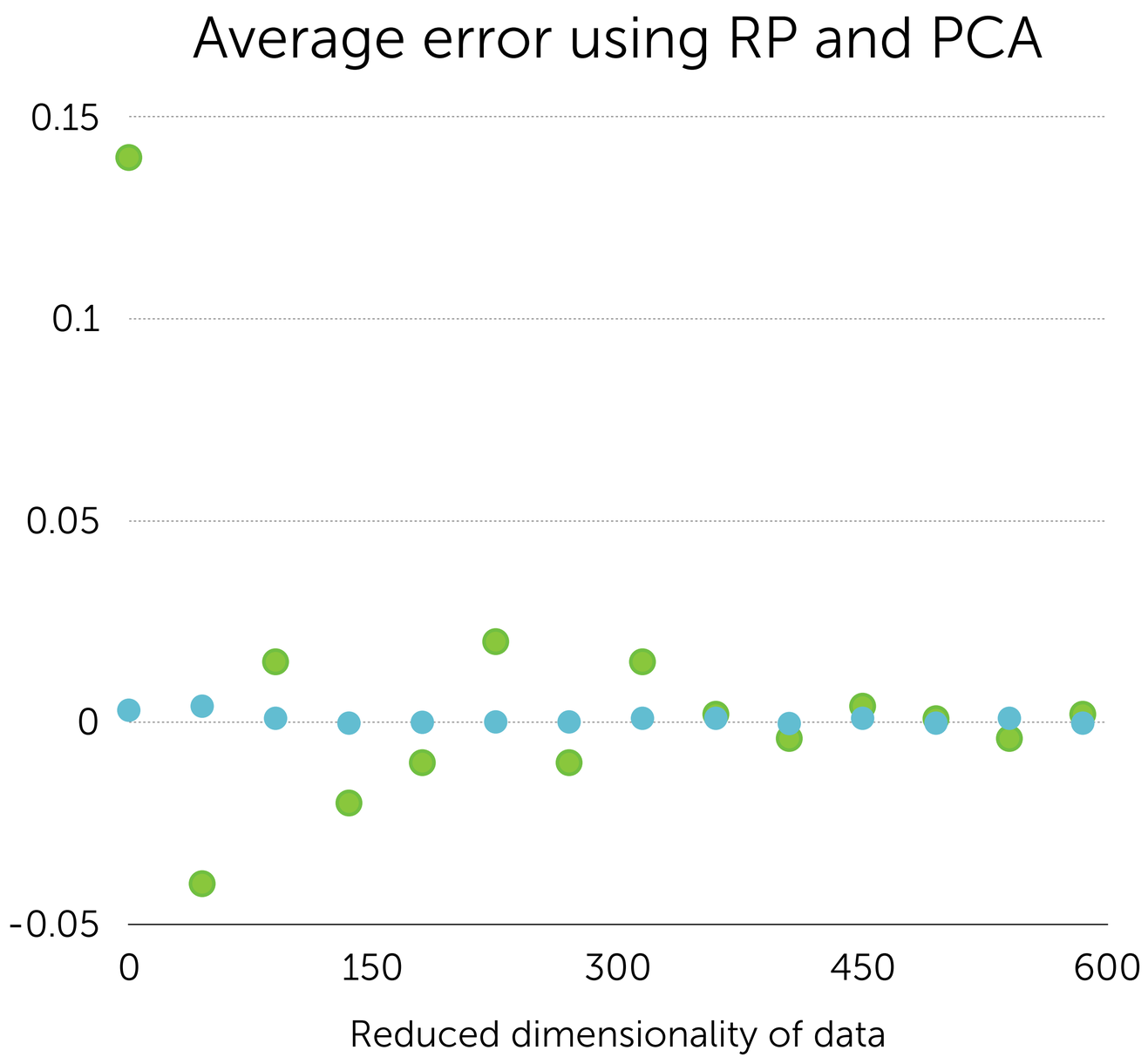
$$(1-\varepsilon)\|u-v\|^2 \leq \|f(u)-f(v)\|^2 \leq (1+\varepsilon)\|u-v\|^2$$

for all $u, v \in X$.

Johnson–Lindenstrauss lemma

The lemma states that a small set of points in a **high-dimensional space** can be embedded into a space of **much lower** (**but still high**) dimension in such a way that distances between the points are **nearly preserved**.

Projection using a random line/plane/hyperplane will perform OK as long as the source and target dimension is high. We typically use normally distributed random numbers.

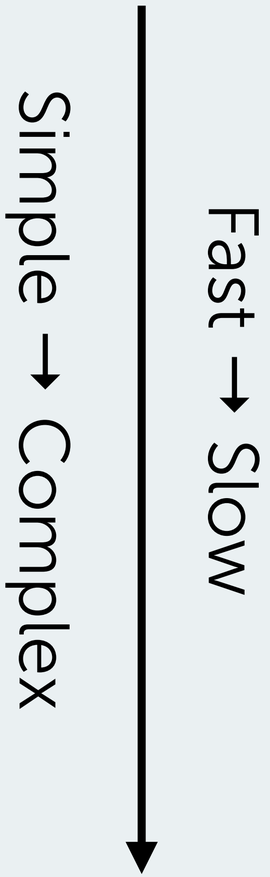

Average error using RP and PCA
0.15
0.1
0.05
0
-0.05
0
150
300
450
600
Reduced dimensionality of data

Random Projection - Summary

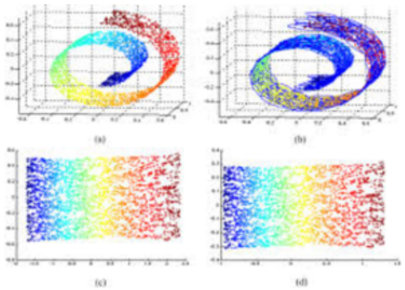
- Preserves: Pairwise distances (sort of)
- Pros: very very fast, doesn’t need to look at the data, can be used nicely in combination with PCA
- Cons: can only take you so far - results start to get weird once we reach low dimensions, like PCA it assumes linearity

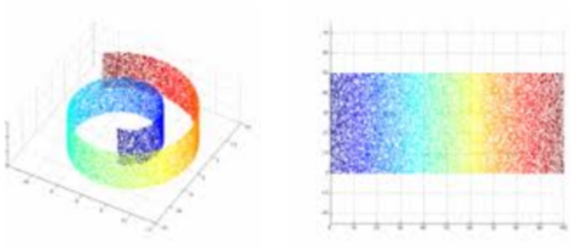
The techniques

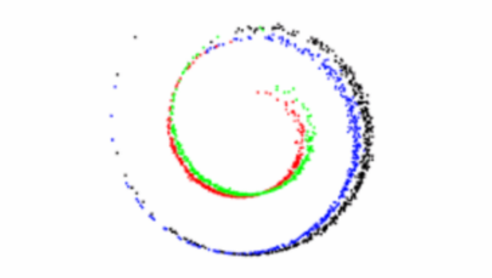
- ✅ Random Projection
- ✅ Principal Component Analysis
- Isomap
- t-SNE

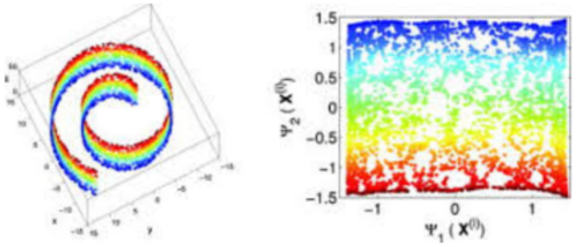


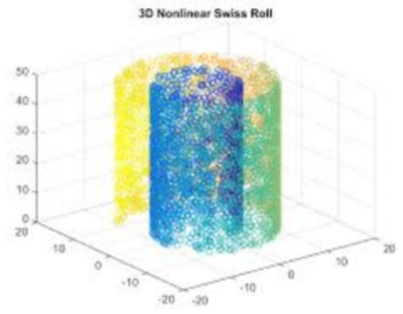
“Swiss roll” problem

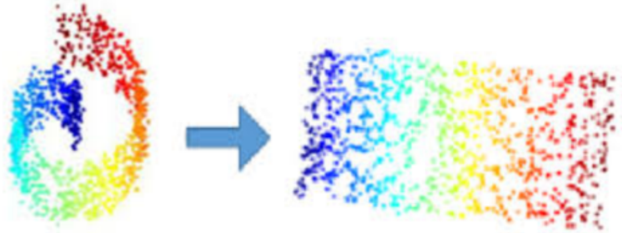


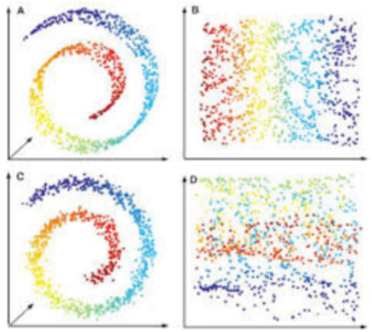


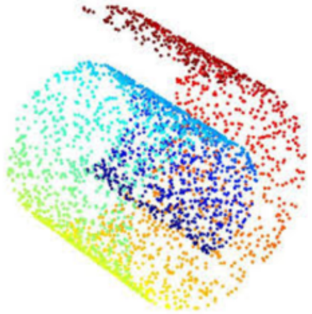


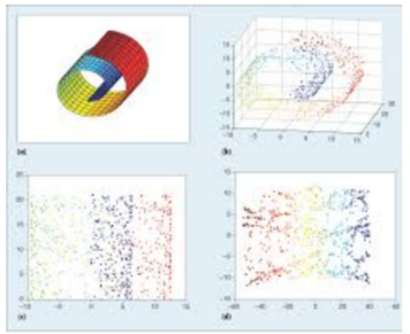


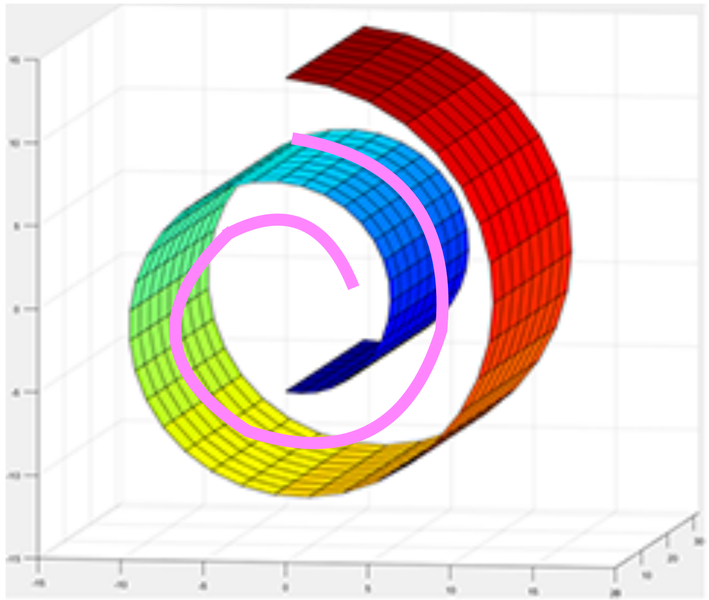


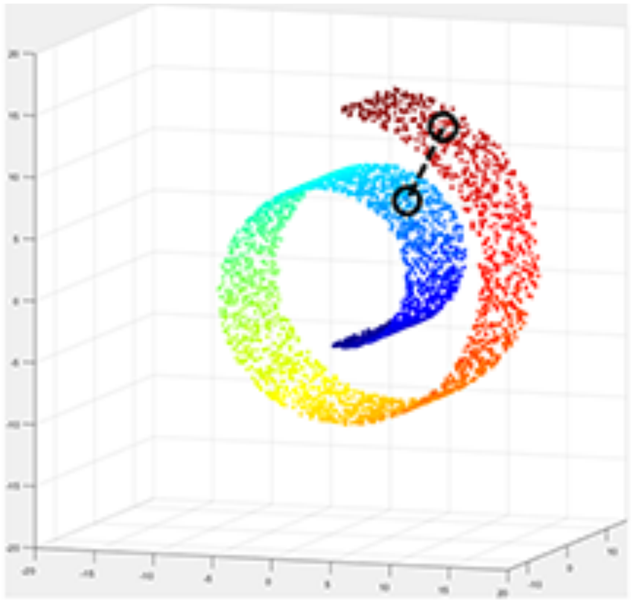


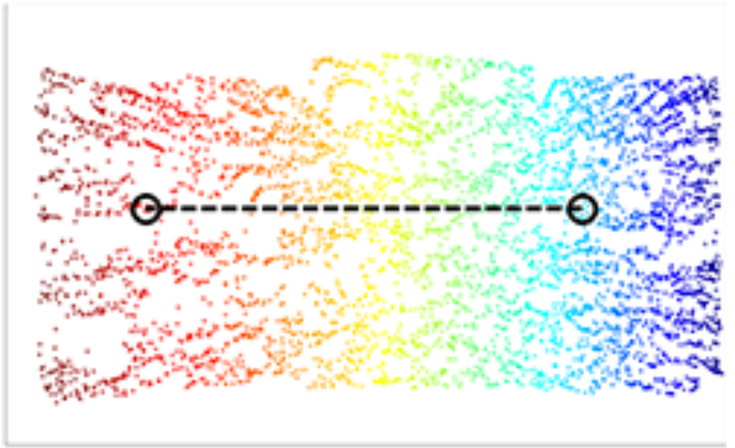


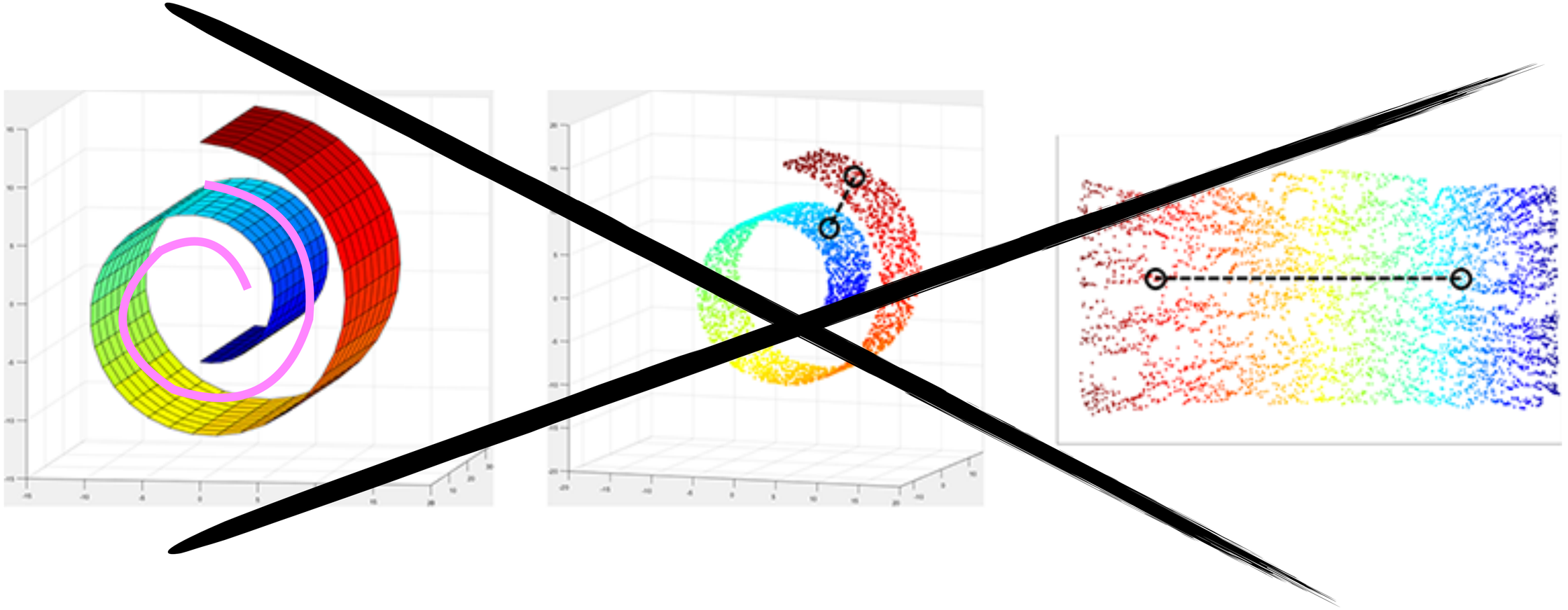


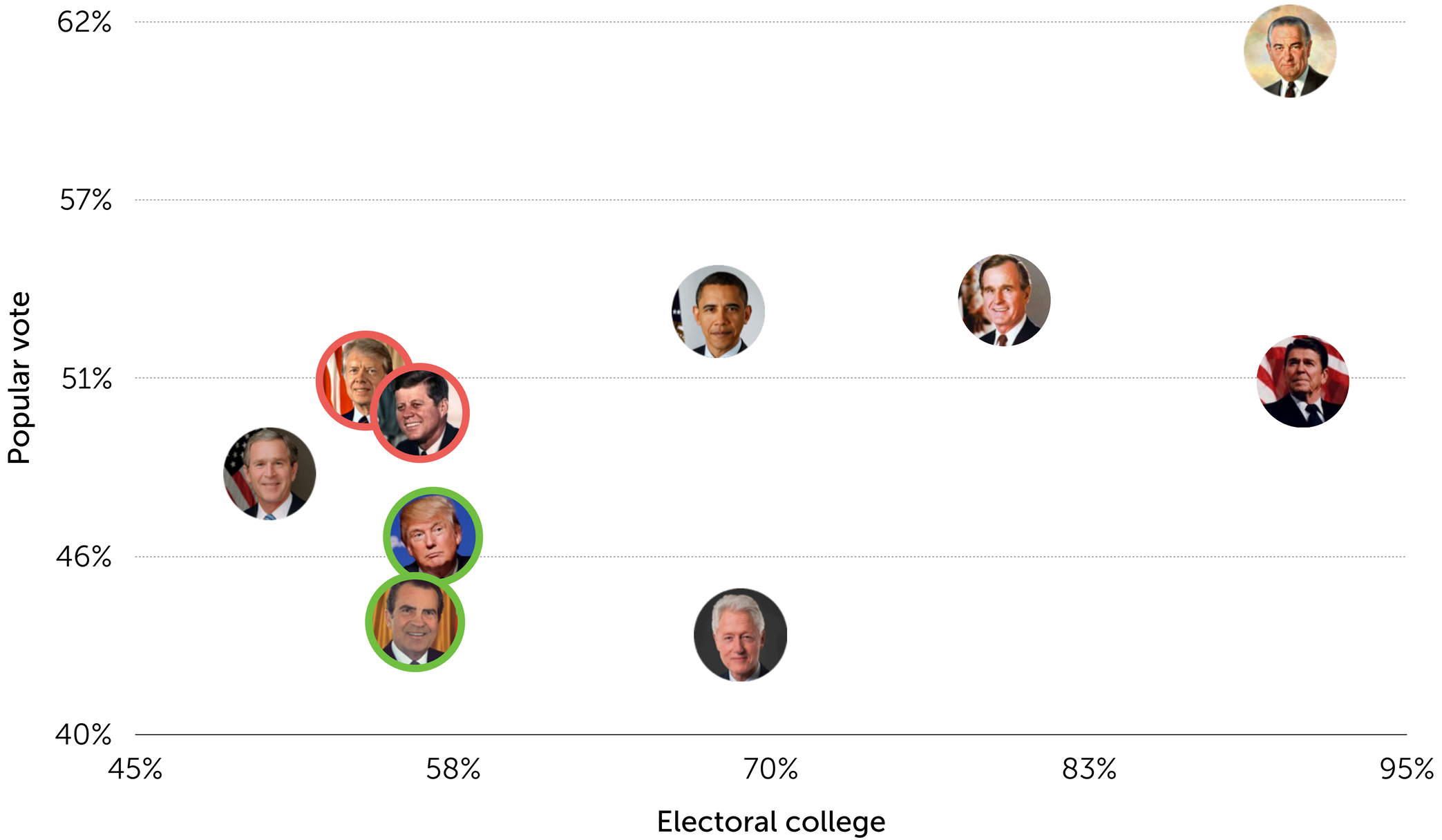


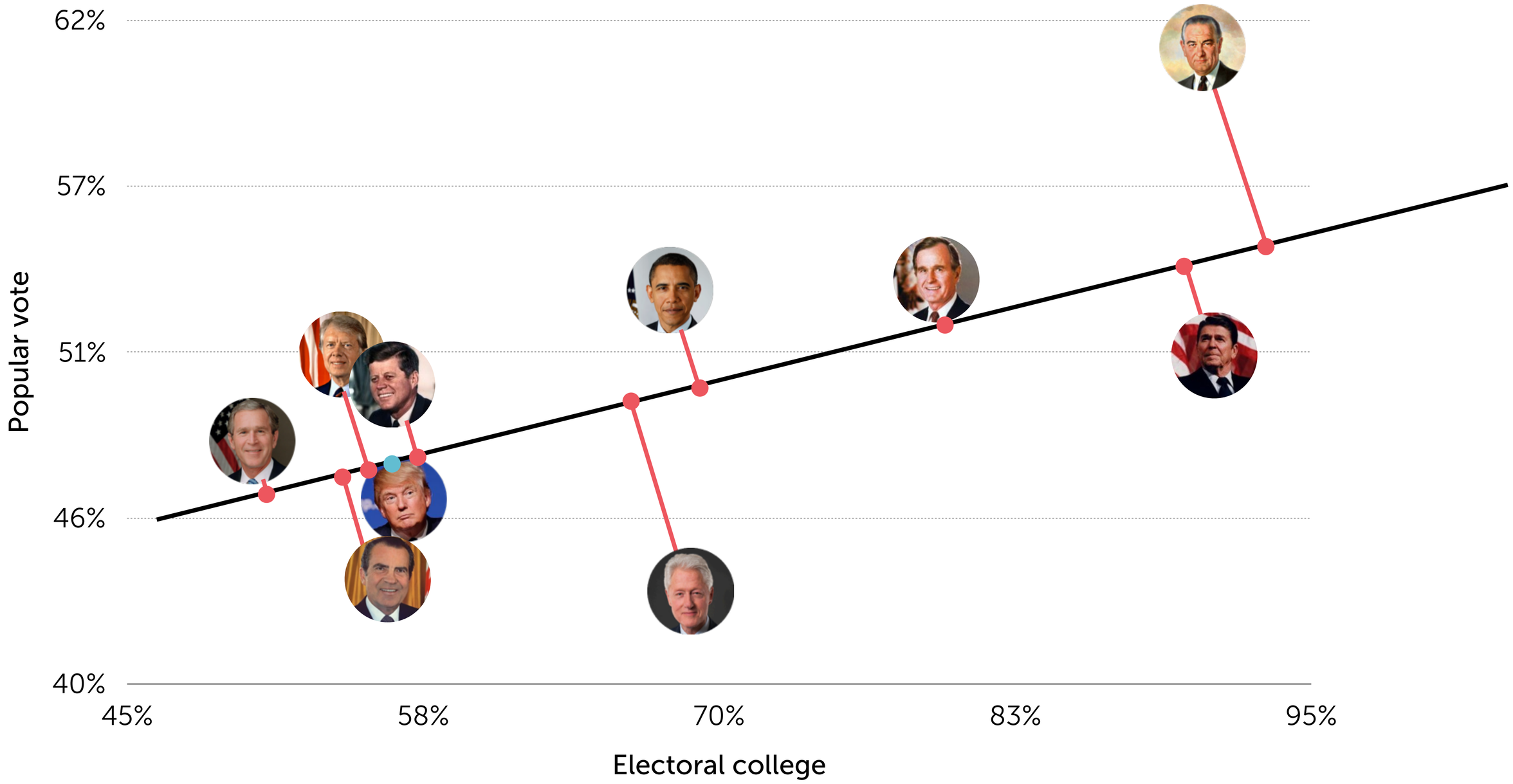


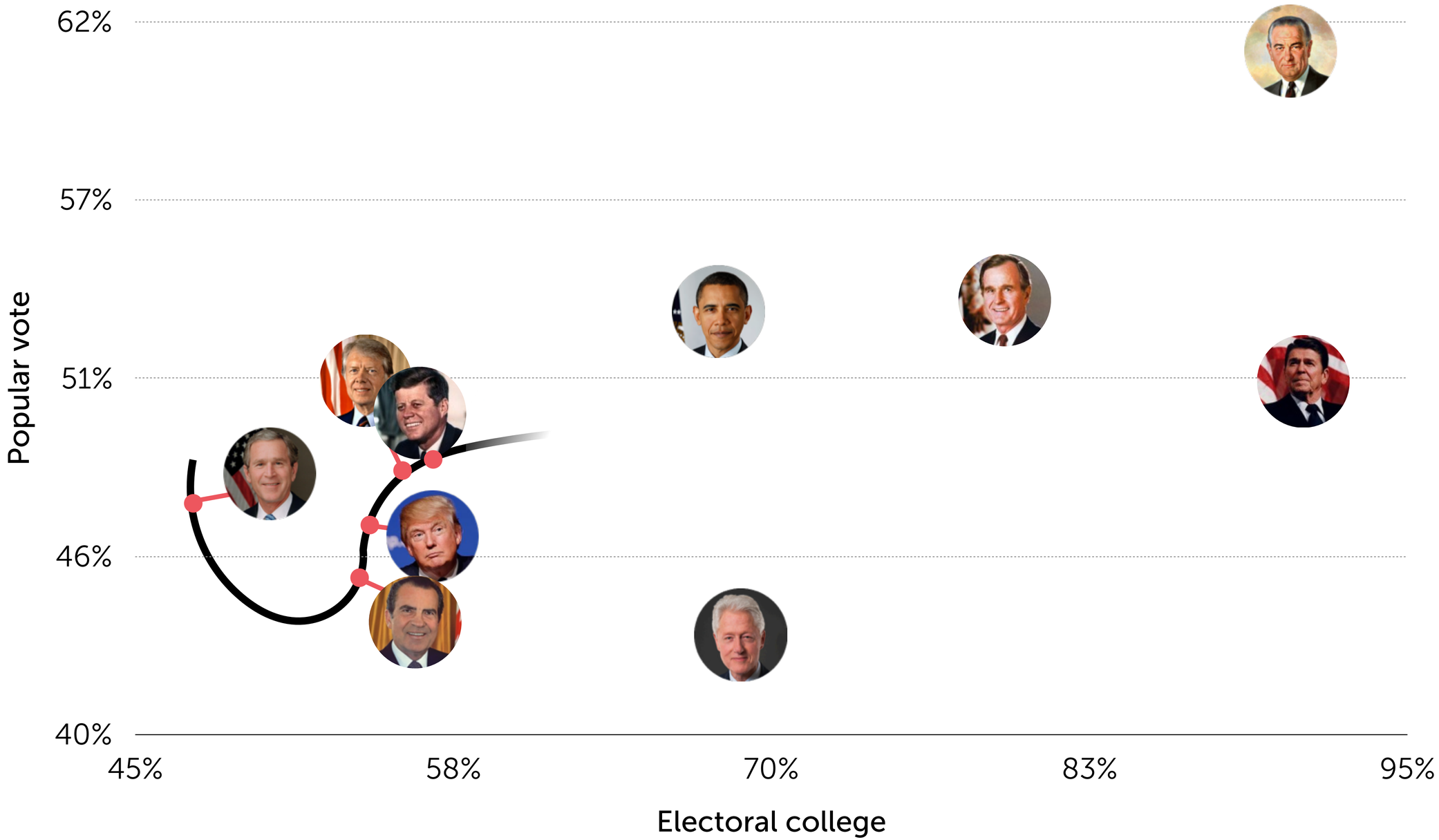






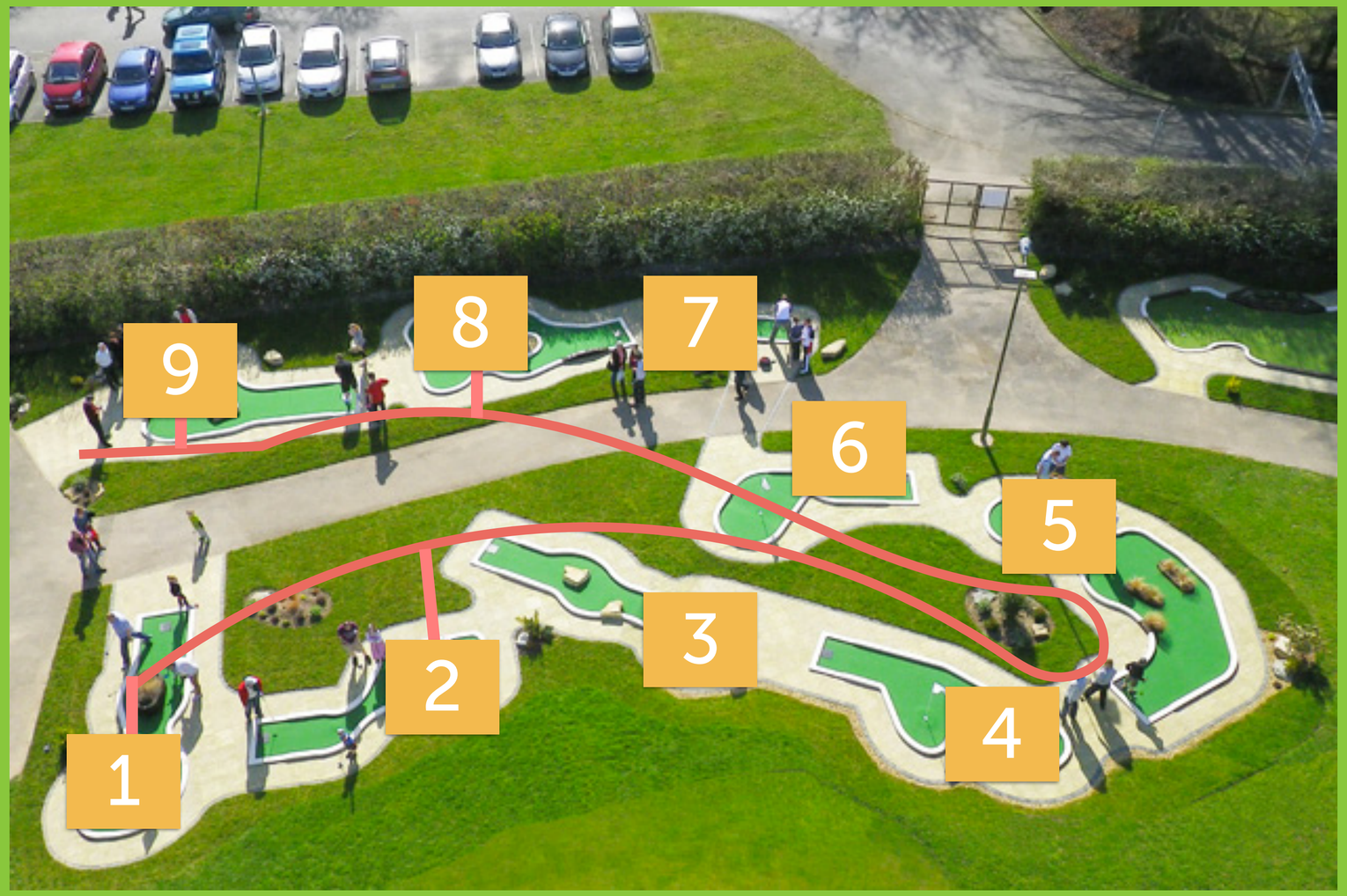

62%
57%
51%
46%
40%
Popular vote
45%
58%
70%
83%
95%
Electoral college


62%
57%
51%
46%
40%
Popular vote
45%
58%
70%
83%
95%
Electoral college

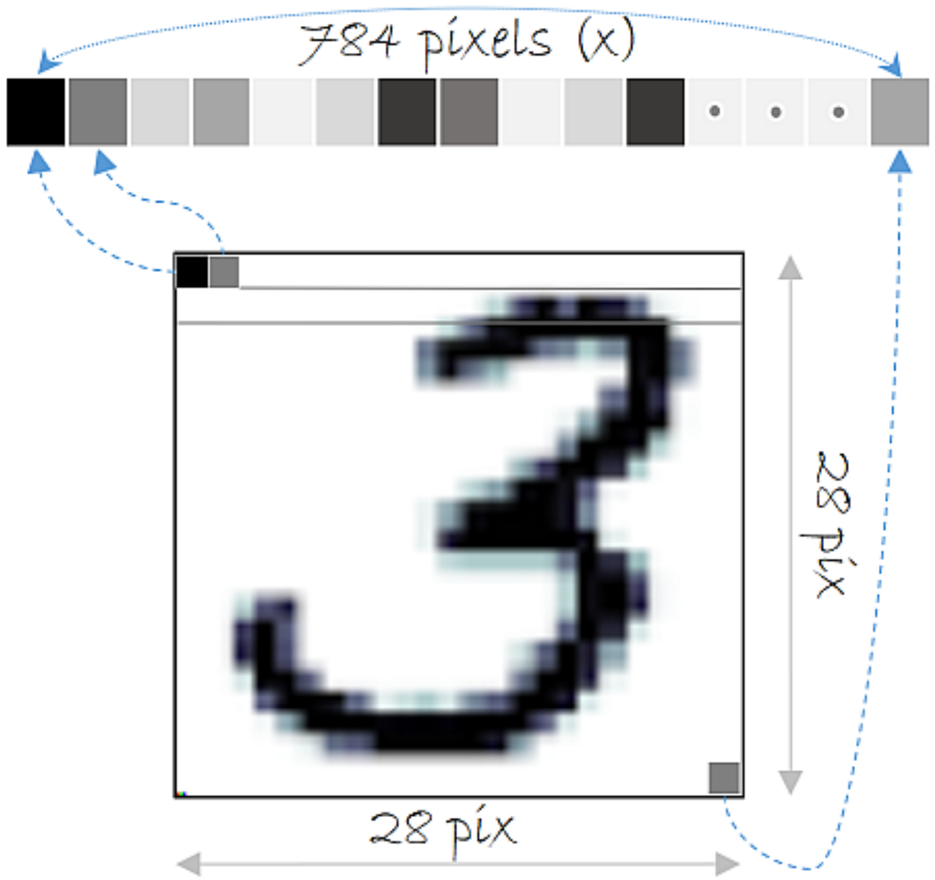

62%
57%
51%
46%
40%
Popular vote
45%
58%
70%
83%
95%
Electoral college


9
8
7
6
5
3
2
4
1

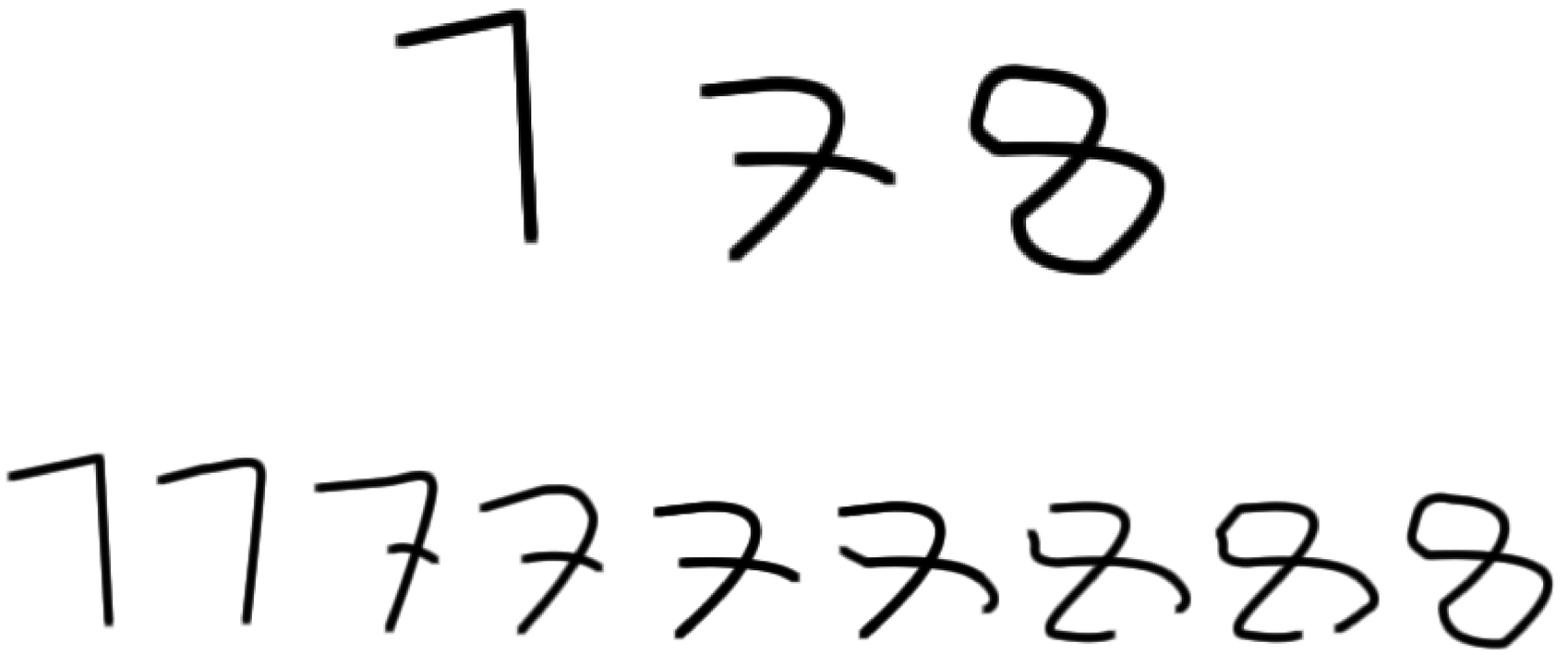

9
8
7
6
5
3
2
4
1

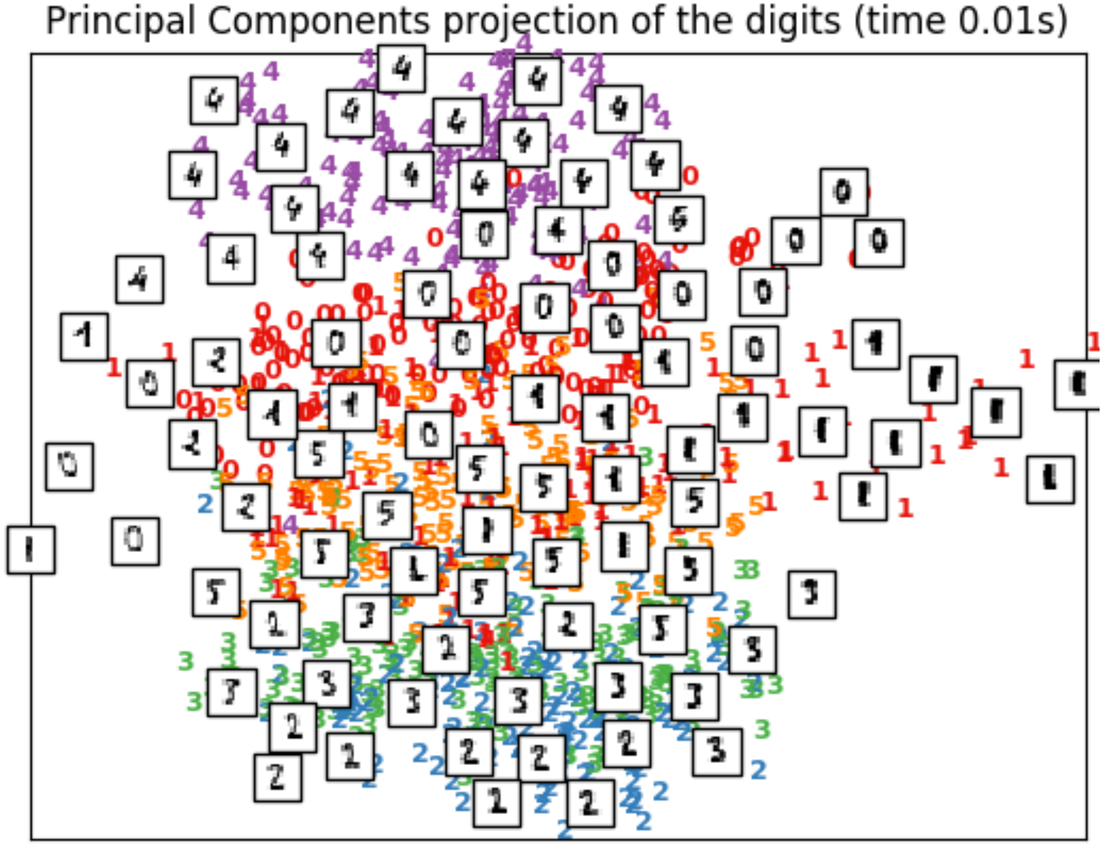

9
8
7
6
5
3
2
4
1

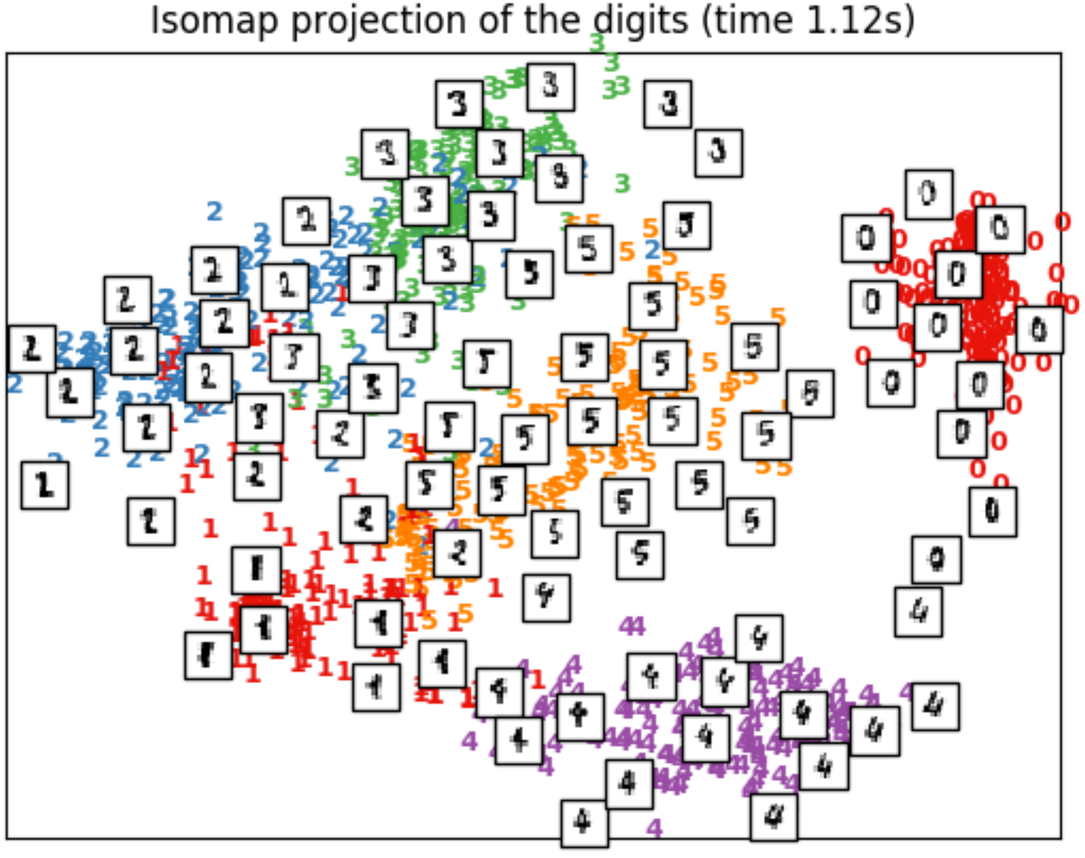
MNIST

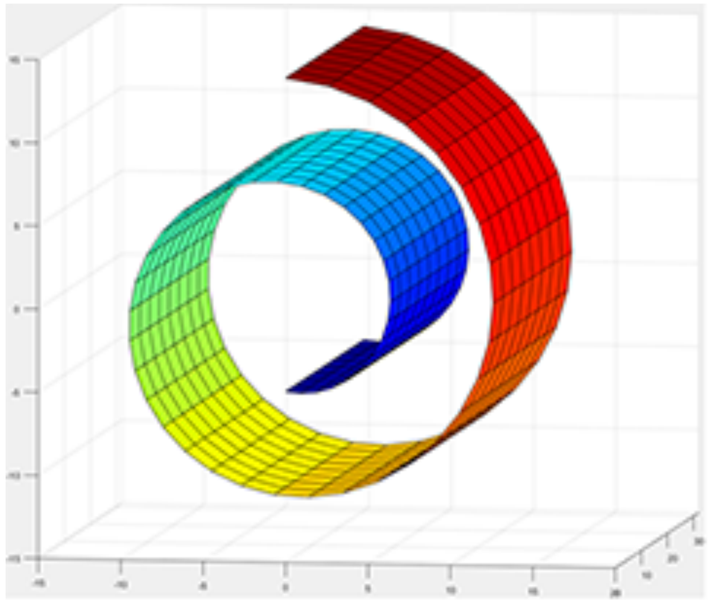


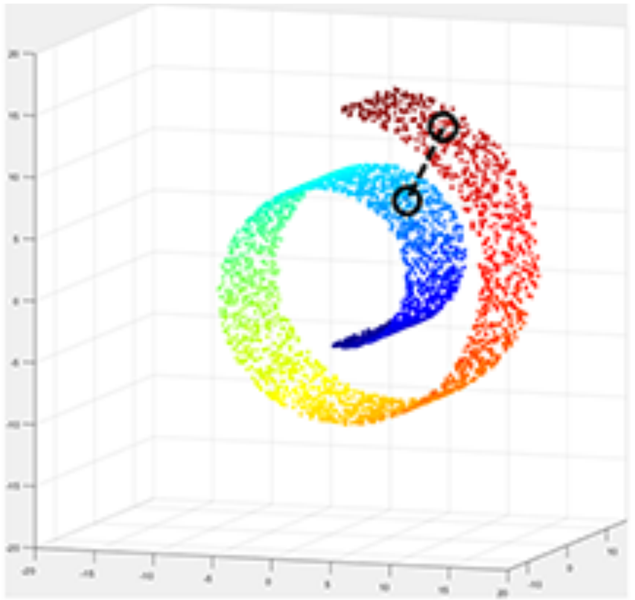
738

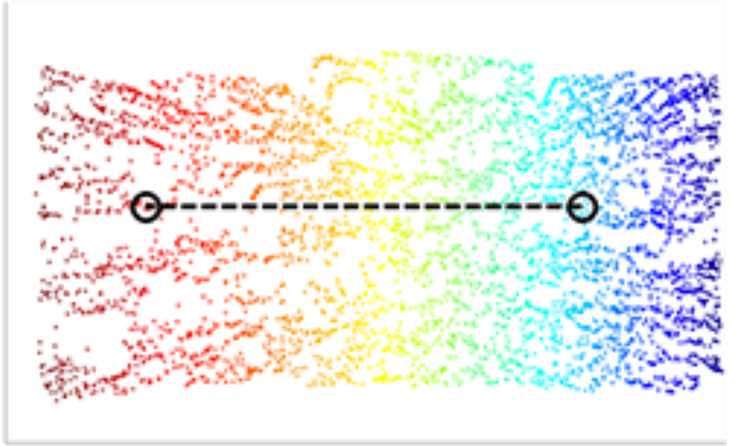



Principal Components projection of the digits (time 0.01s)


Isomap projection of the digits (time 1.12s)





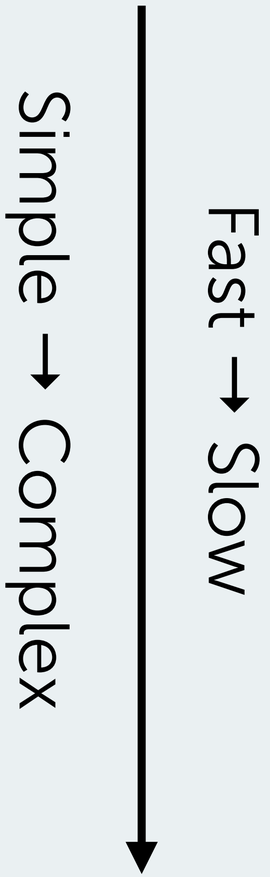


Isomap - Summary

- Preserves: Geodesic distance
- Pros: works well with non-linear data (e.g images), good for visualization
- Cons: much slower than PCA, especially for large data, may not work well for linear data, not as effective for pre-processing (vs. visualization), generally outperformed by t-SNE

The techniques

- ✅ Random Projection
- ✅ Principal Component Analysis
- ✅ Isomap
- t-SNE



t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-SNE - A quick explanation

t-SNE constructs a probability distribution over pairs of high-dimensional objects in such a way that similar objects have a high probability of being picked. It does the same in the low-dimensional space; then it tries to align the two distributions.

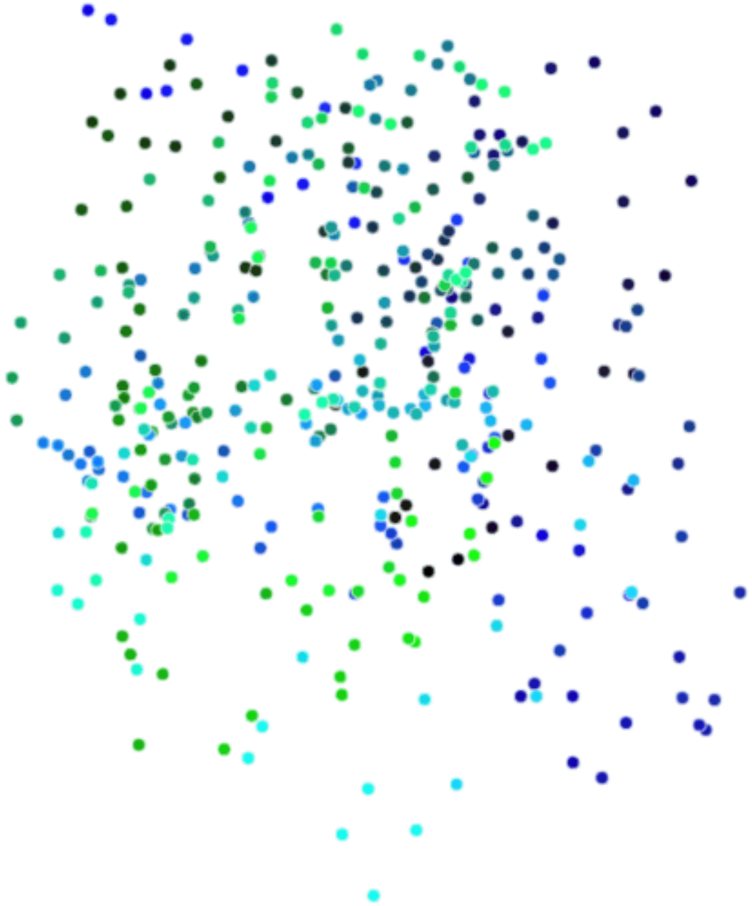
High-dimensional space = Gaussian distribution. 2D space = Student's t-distribution. The fatter tails of a t-distribution compared to a Gaussian help to spread the points more evenly in the 2-dimensional space.

t-SNE - A quick explanation

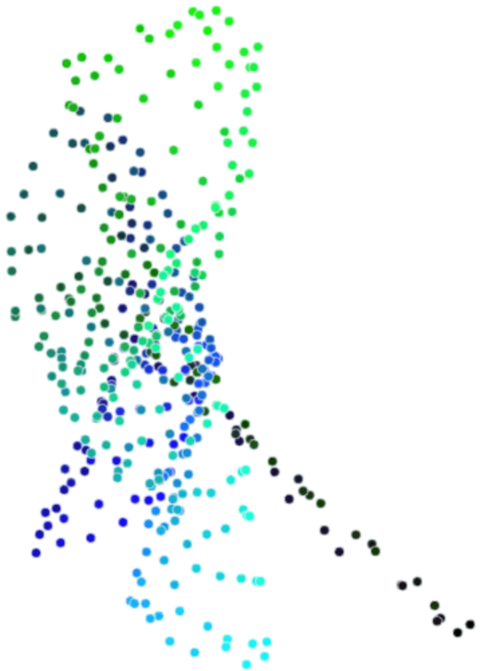
The main parameter controlling the fitting is called **perplexity**. A low perplexity means we care about local scale and focus on the closest other points. High perplexity takes more of a “big picture” approach.

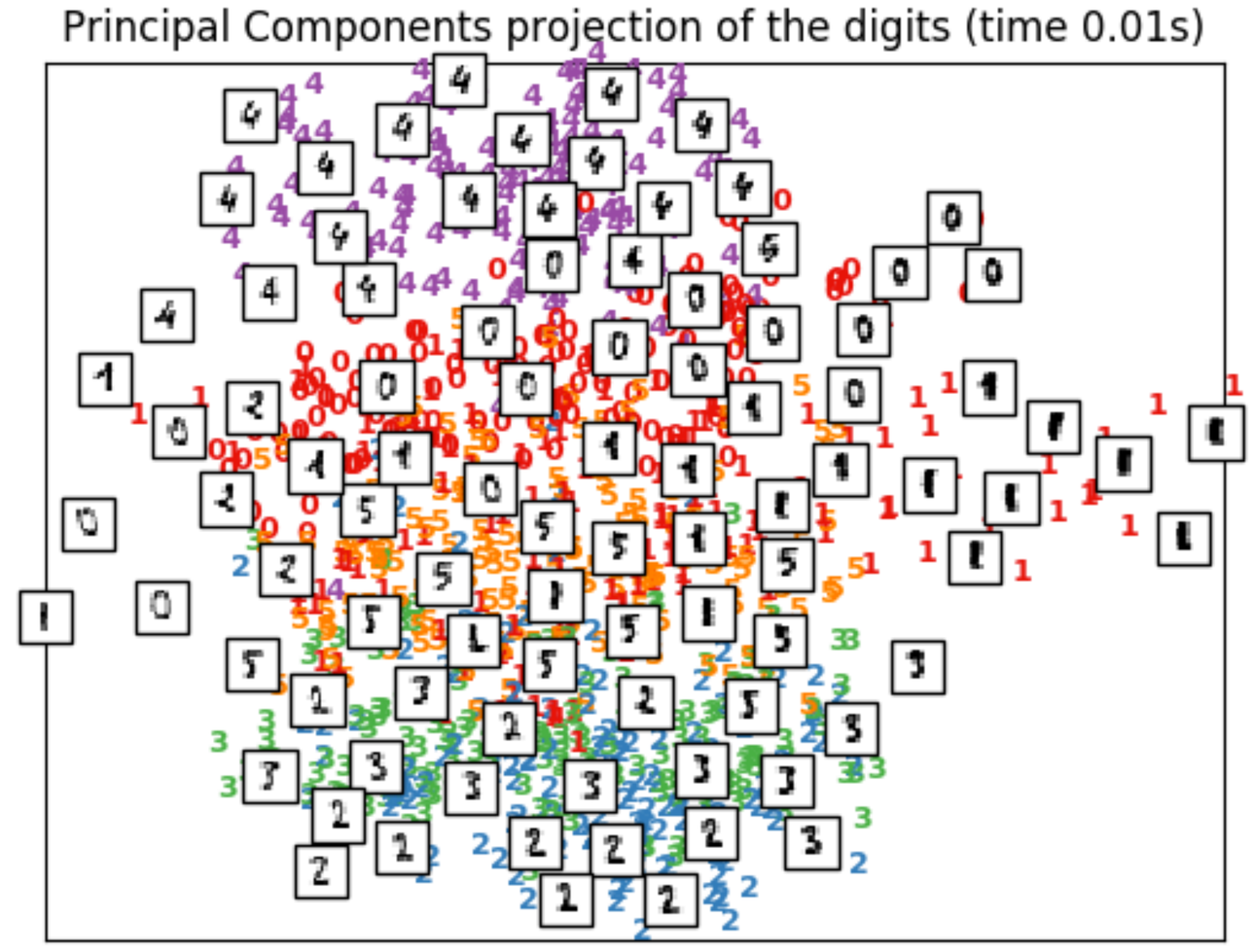
Usually (unless perplexity is very high), t-SNE is focused on **preserving local structure**.

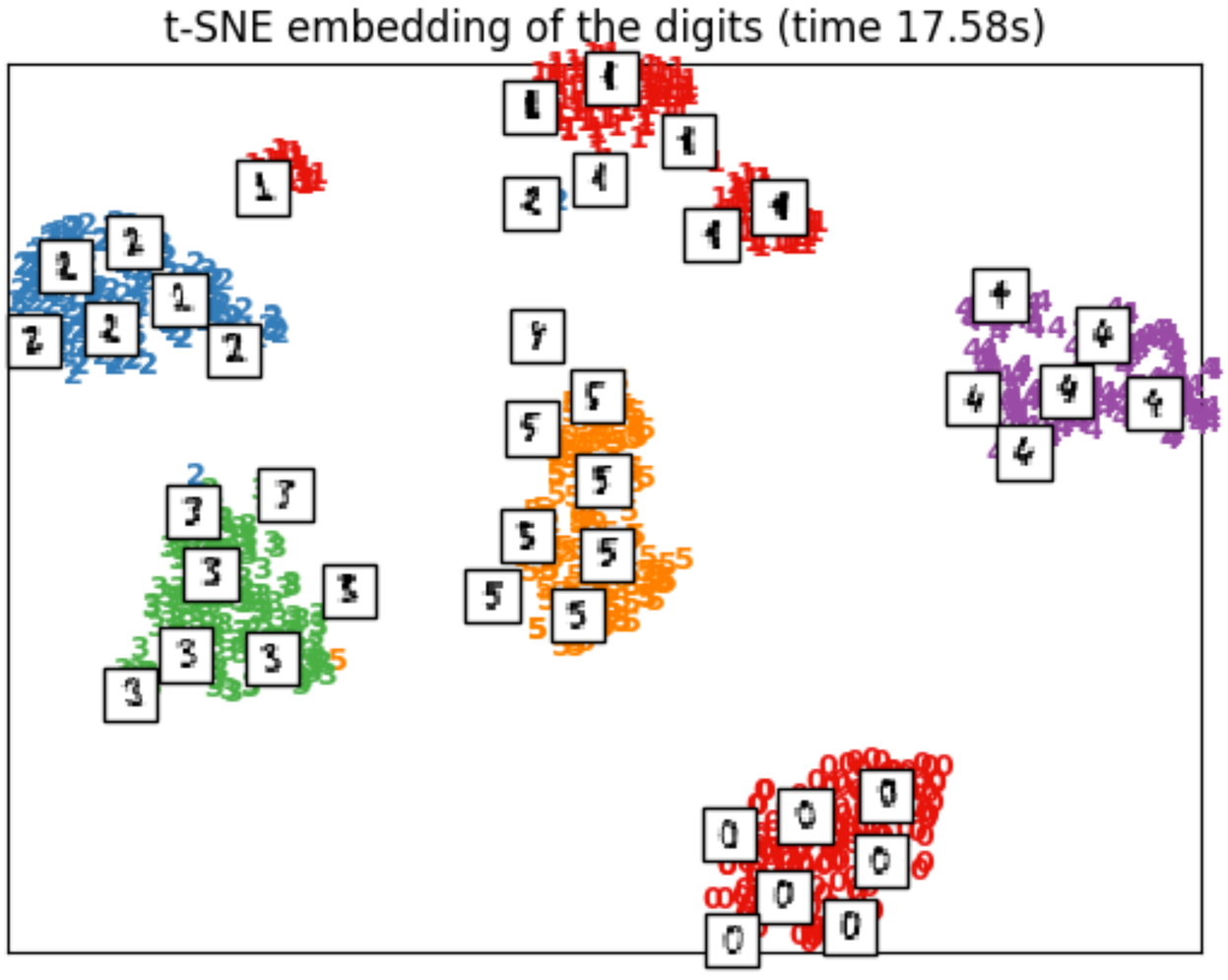
Randomness & t-SNE

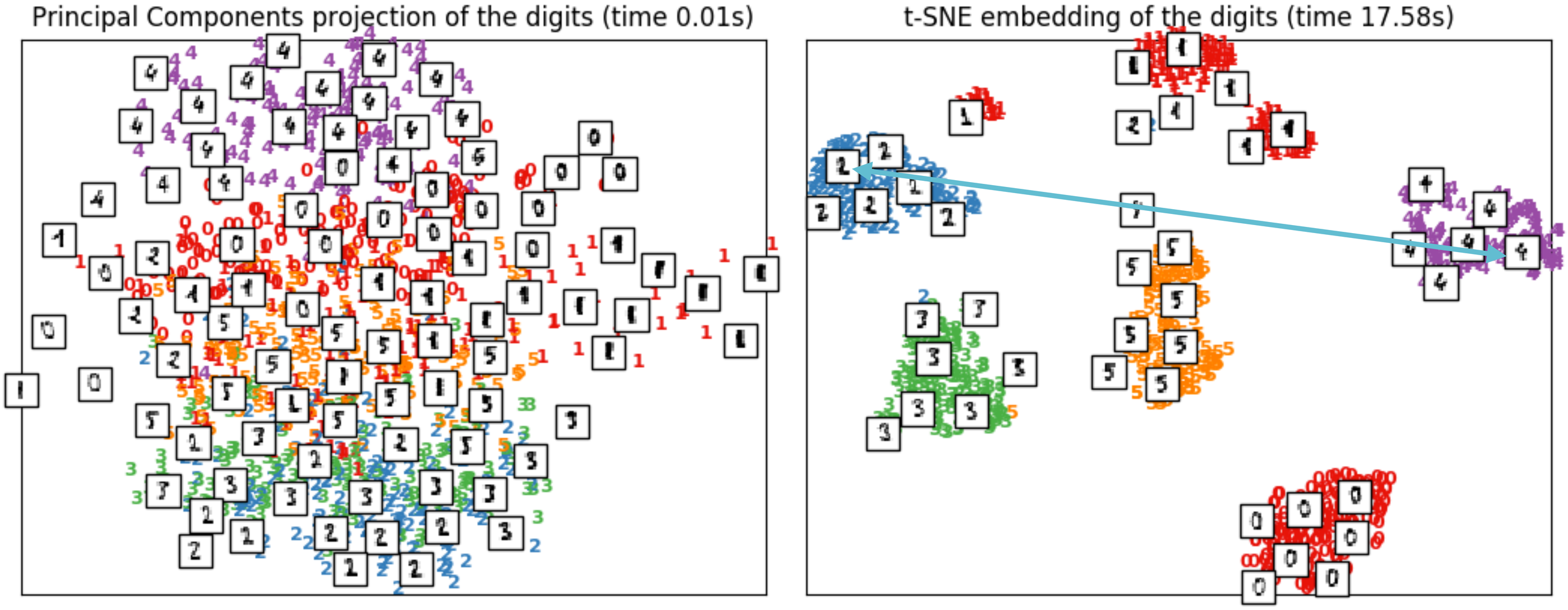


Randomness & t-SNE




Principal Components projection of the digits (time 0.01s)


t-SNE embedding of the digits (time 17.58s)


Principal Components projection of the digits (time 0.01s)
t-SNE embedding of the digits (time 17.58s)

t-SNE & PCA

We'll often use PCA & t-SNE together, because:

a) it'll be faster

b) it will preserve global structure better

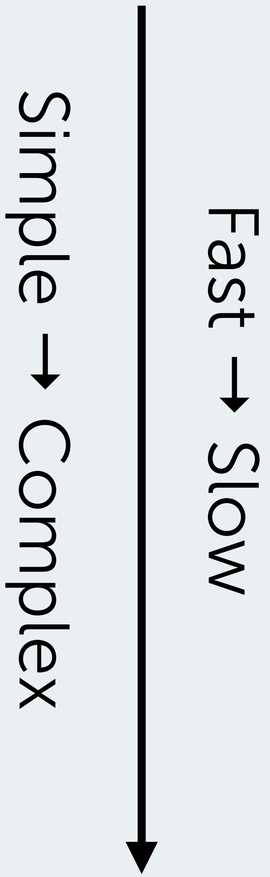
"It is highly recommended to use another dimensionality reduction method (e.g. PCA) to reduce the number of dimensions to a reasonable amount (e.g. 50)" - Scipy docs

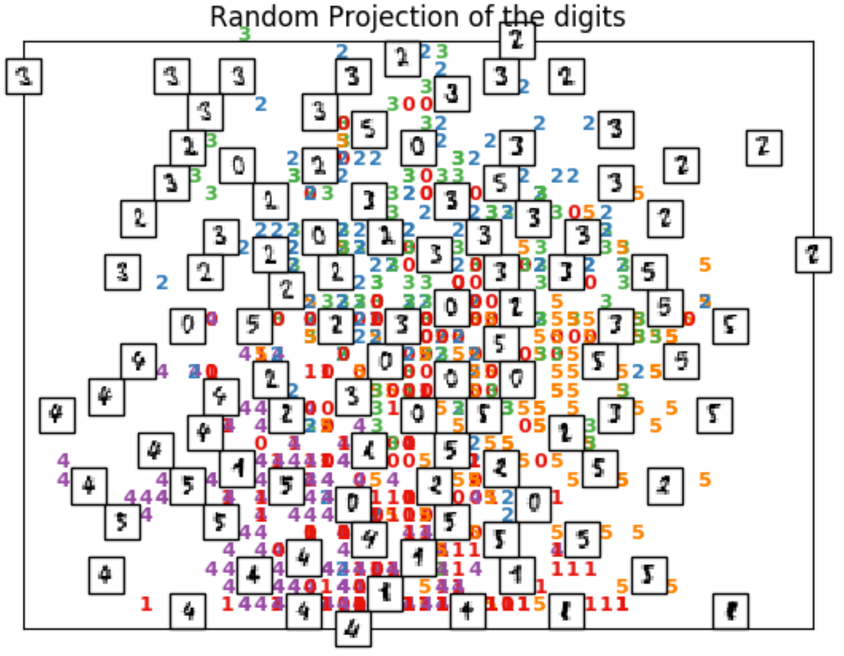
t-SNE - Summary

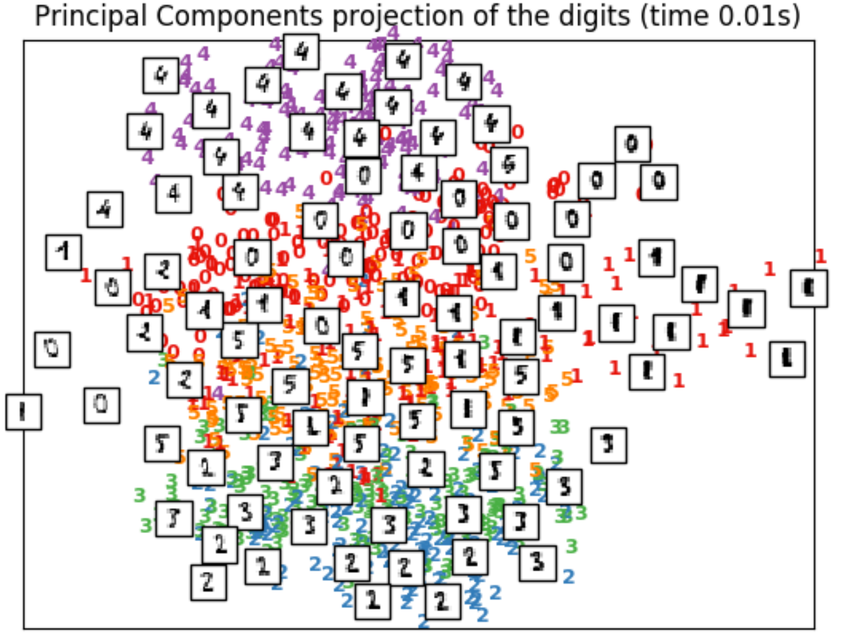
- Preserves: Local structure
- Pros: generally works very well for visualization, can be adjusted with perplexity, works on many kinds of data, pairs well with PCA
- Cons: computationally expensive, element of randomness mean results aren’t guaranteed, doesn’t preserve global structure

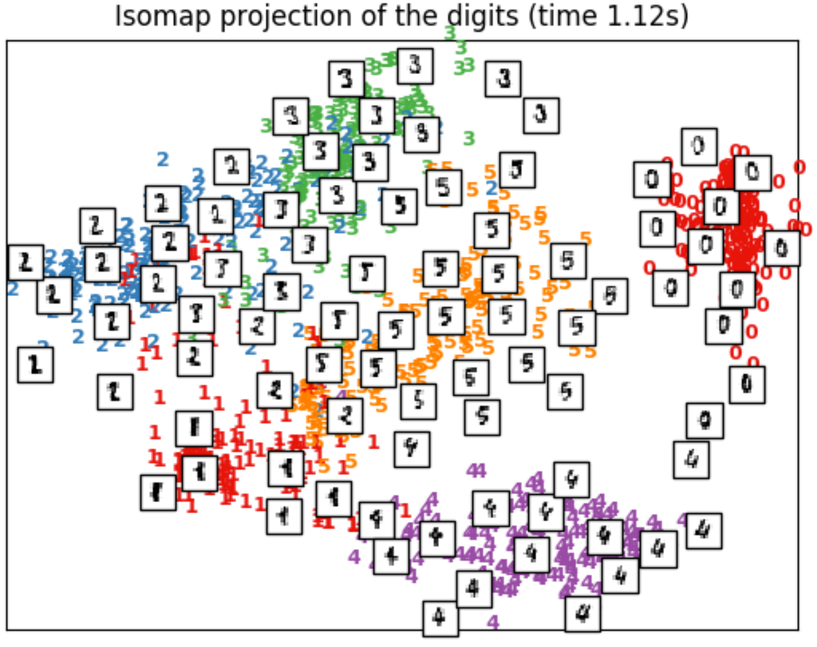
The techniques

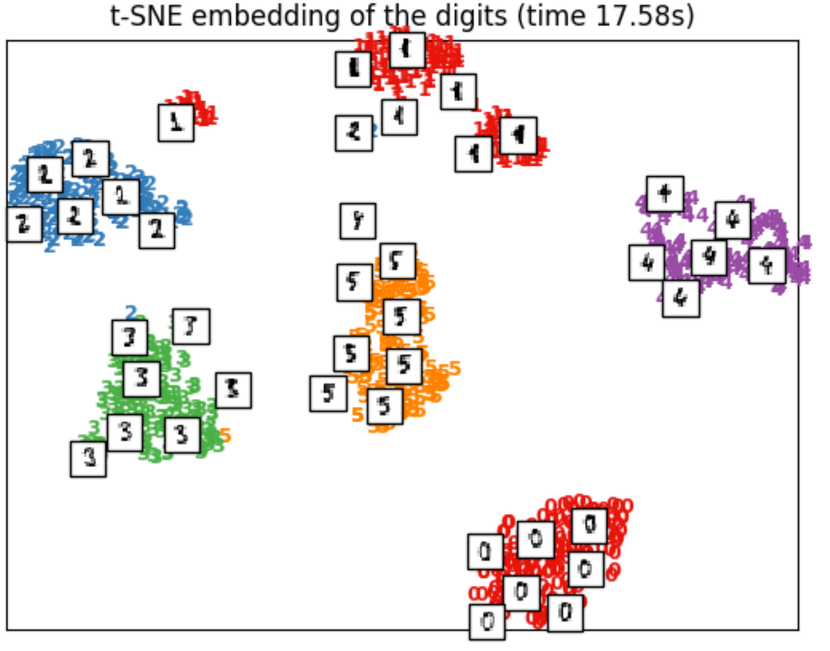
- ✅ Random Projection
- ✅ Principal Component Analysis
- ✅ Isomap
- ✅ t-SNE




Random Projection of the digits

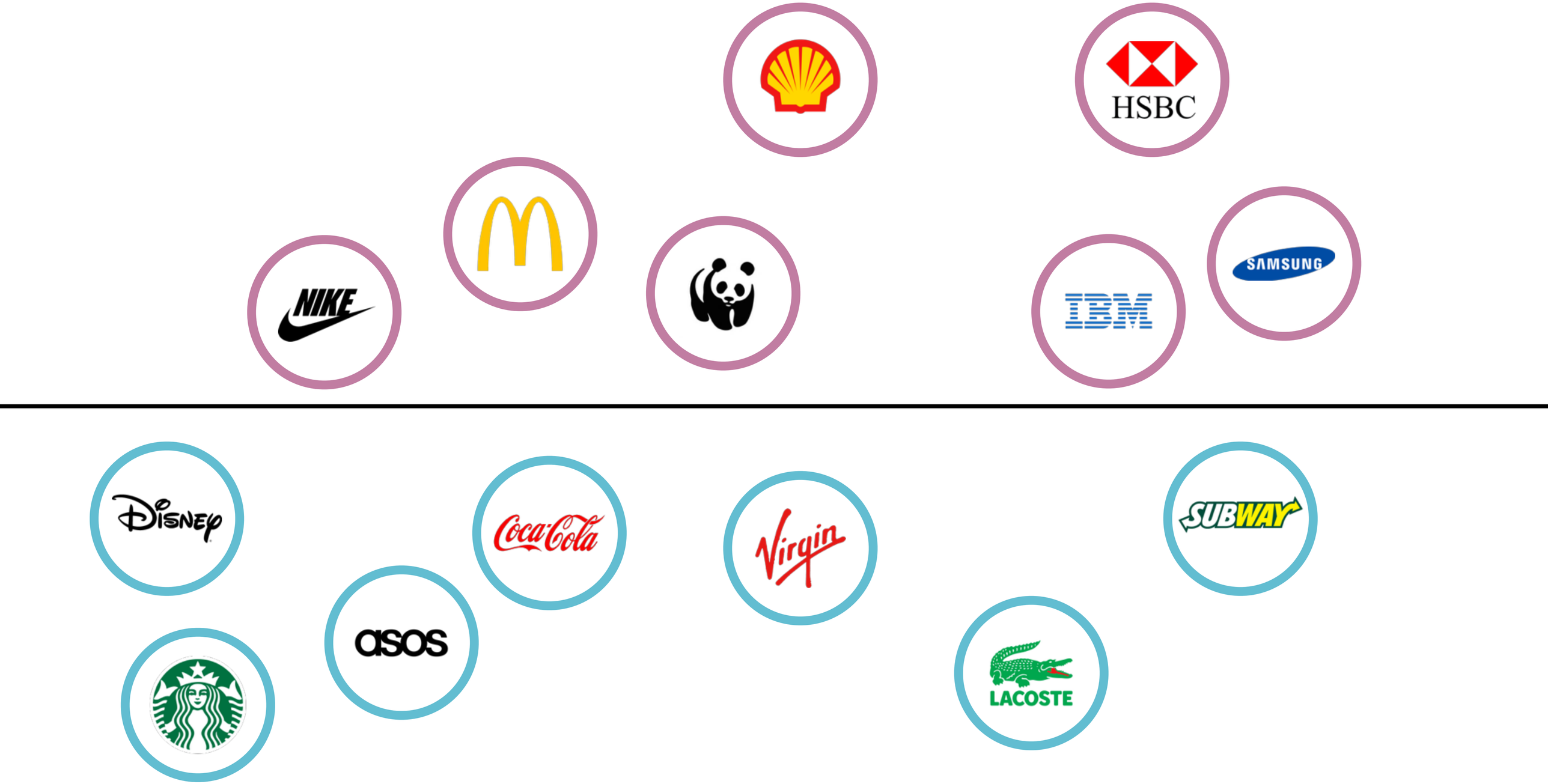

Principal Components projection of the digits (time 0.01s)

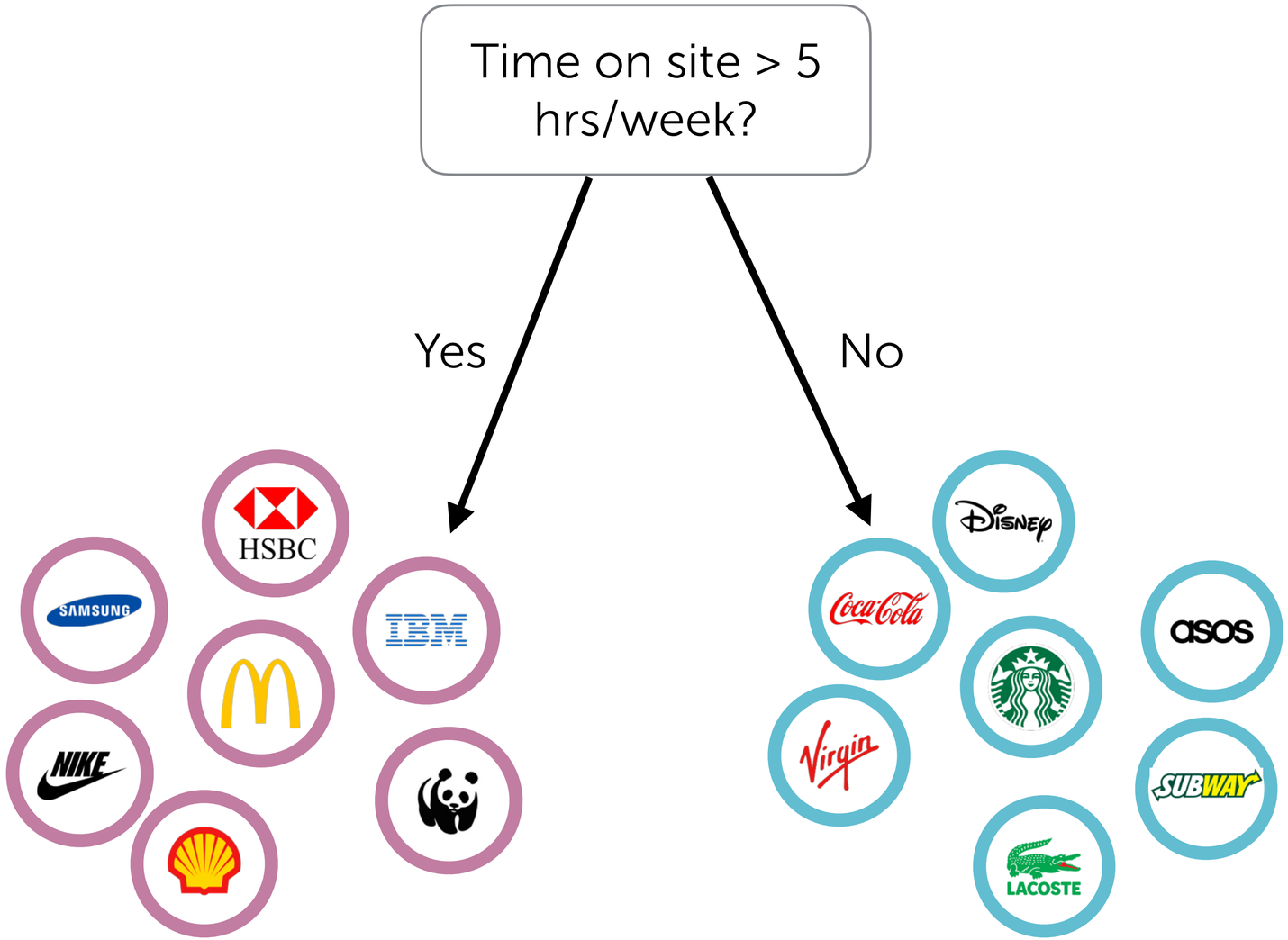

Isomap projection of the digits (time 1.12s)

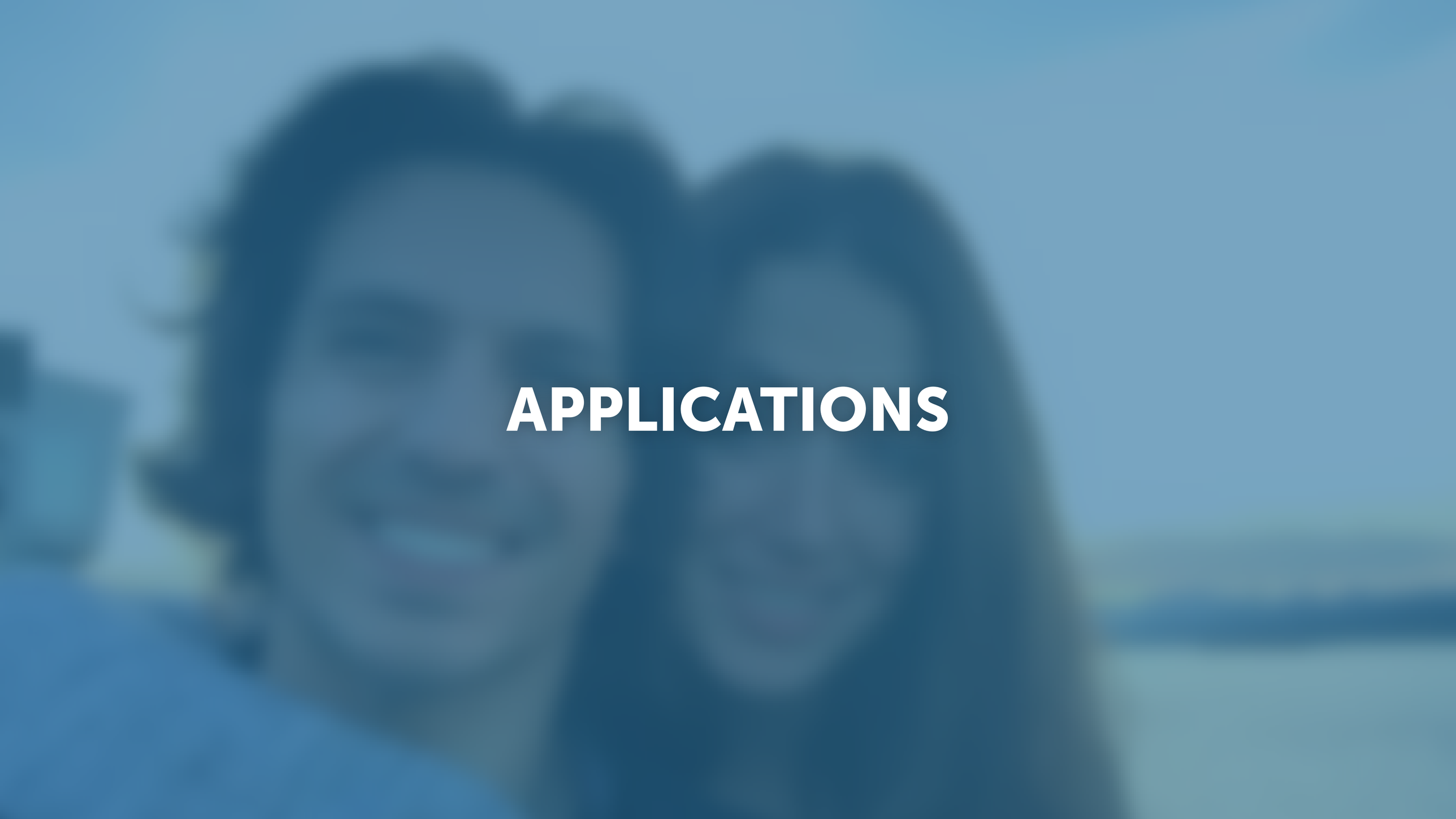

t-SNE embedding of the digits (time 17.58s)

Side note: interpreting DR

Name	~~Performance~~ ???	~~Cost~~ ???
Audi S4	6	4
MINI Cooper S	1	2
Nissan GT-R	10	10
Porsche 911 GT3	7	9
Renault Clio Williams	3	1
Volkswagen Golf R	5	3
BMW i8	6	10


HSBC
NIKE
SAMSUNG
IBM
Disney
Coca-Cola
Virgin
SUBWAY
asos
LACOSTE

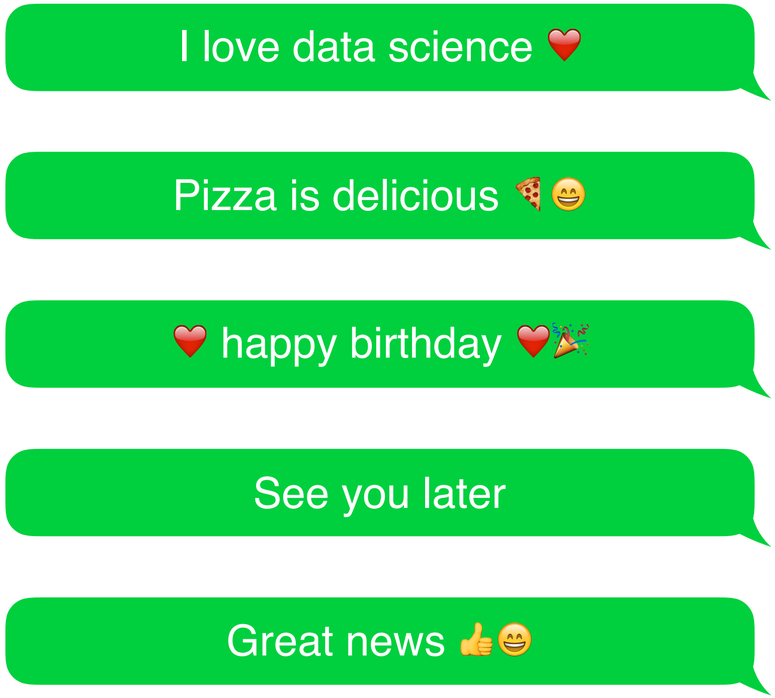

Time on site > 5 hrs/week?
Yes
No
HSBC
SAMSUNG
IBM
NIKE
Disney
Coca-Cola
asos
Virgin
SUBWAY
LACOSTE



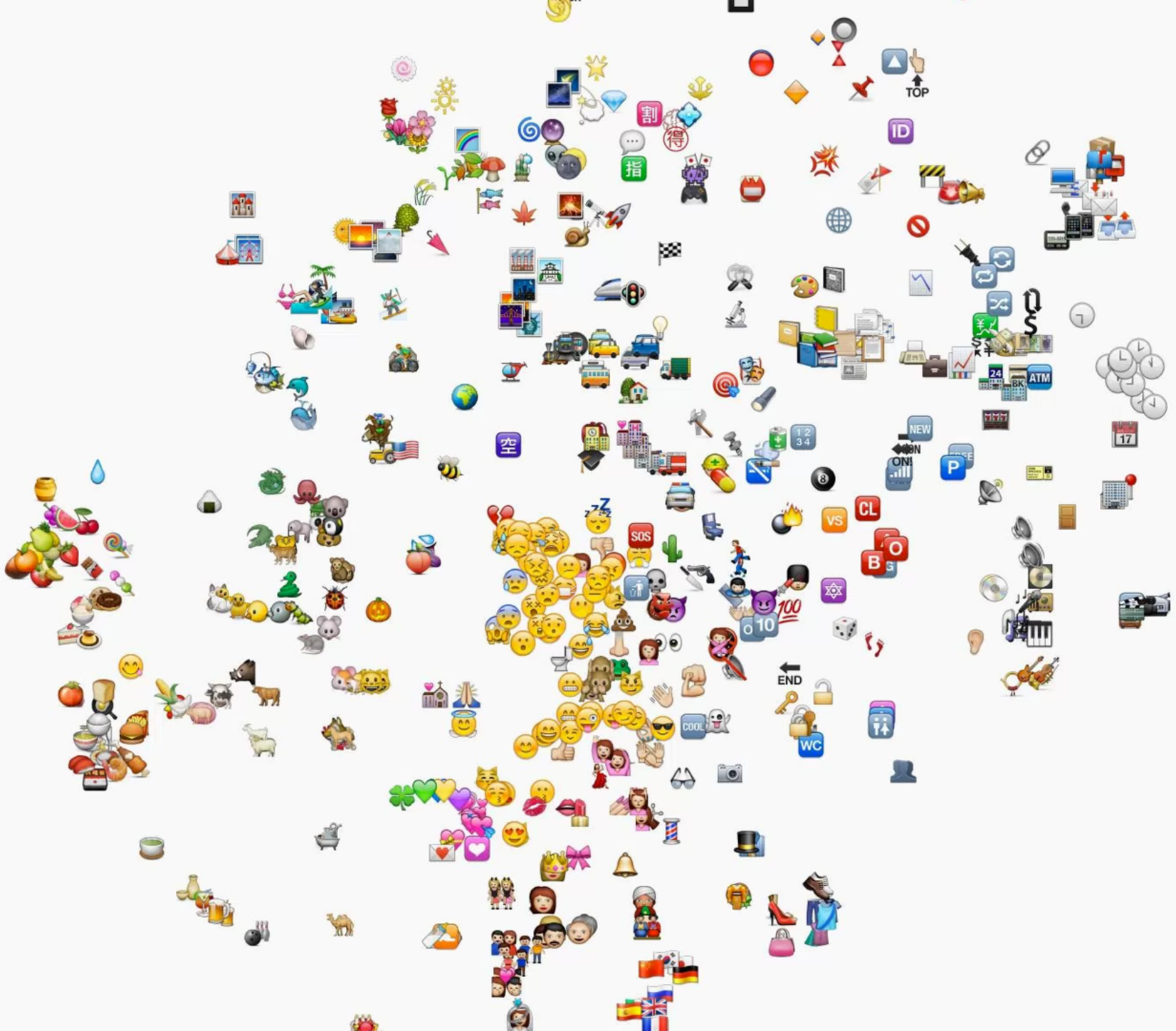
APPLICATIONS

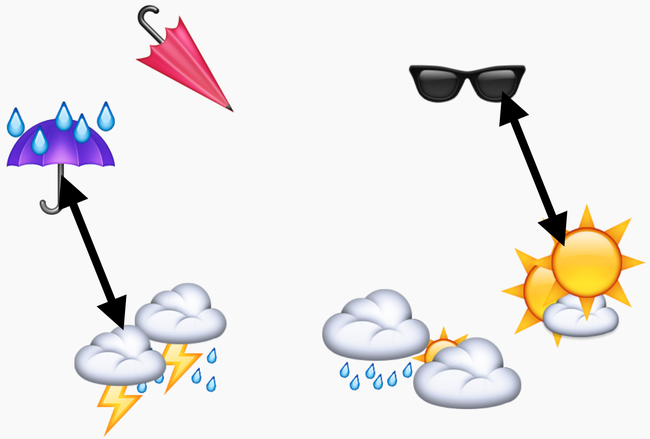
Recommender systems

Recommender systems



❤️	🍕	😄	🎉	…	👍
1	0	0	0	…	0
0	1	1	0	…	0
2	0	0	1	…	0
0	0	0	0	…	0
0	0	1	0	…	1


TOP
ID
割
得
指
ATM
NEW
ON!
17
P
VS
CL
SOS
100
10
END
COOL
WC



Word vectors

Facebook’s Fasttext: ~5 million → 300 dimensions

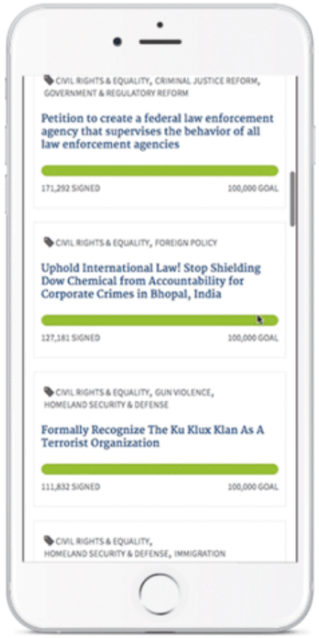
king - man + woman = queen

Word vectors

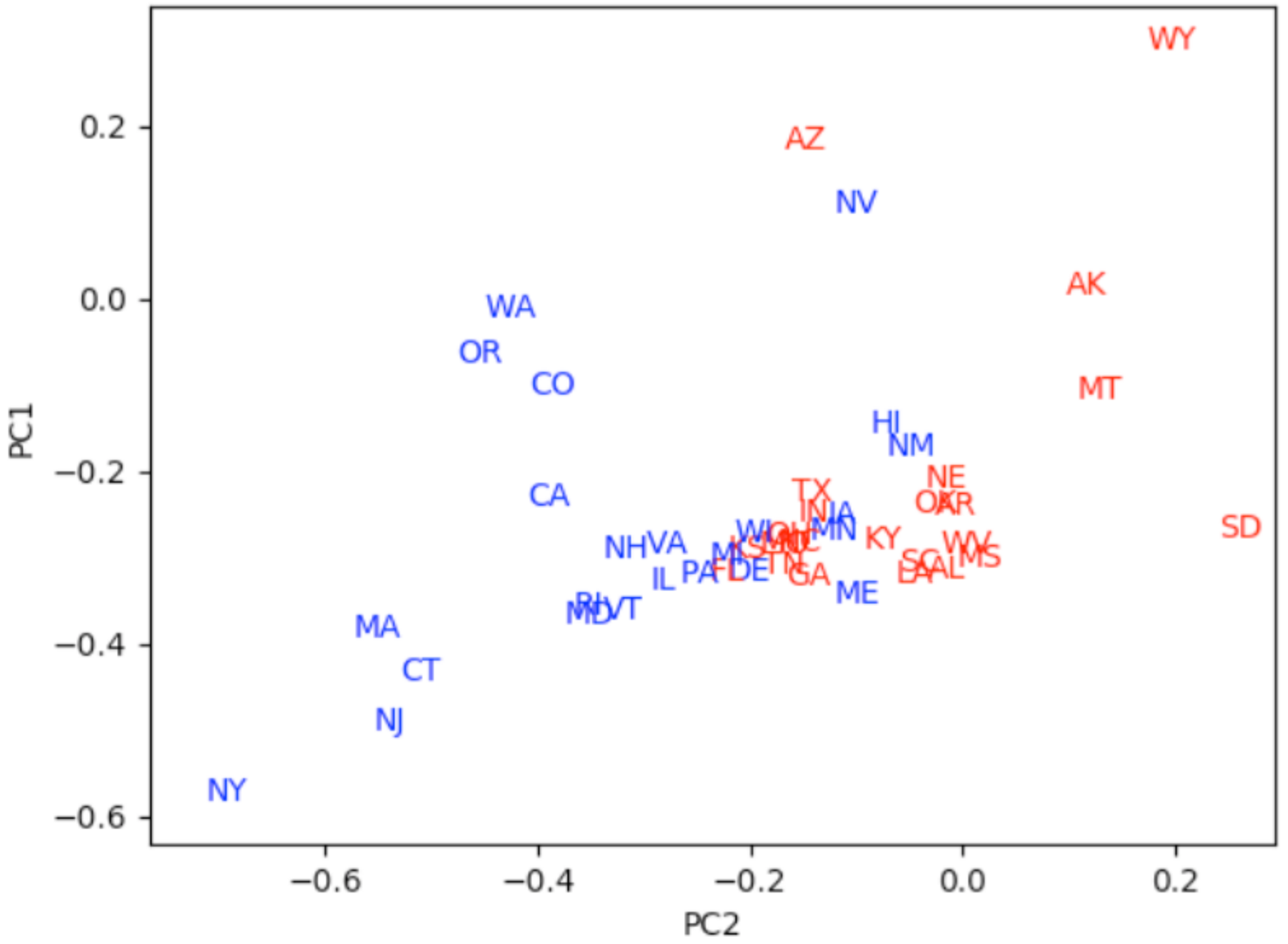
Czech + currency	Vietnam + capital	German + airlines	Russian + river	French + actress
koruna	Hanoi	airline Lufthansa	Moscow	Juliette Binoche
Check crown	Ho Chi Minh City	carrier Lufthansa	Volga River	Vanessa Paradis
Polish zolty	Viet Nam	flag carrier Lufthansa	upriver	Charlotte Gainsbourg
CTK	Vietnamese	Lufthansa	Russia	Cecile De

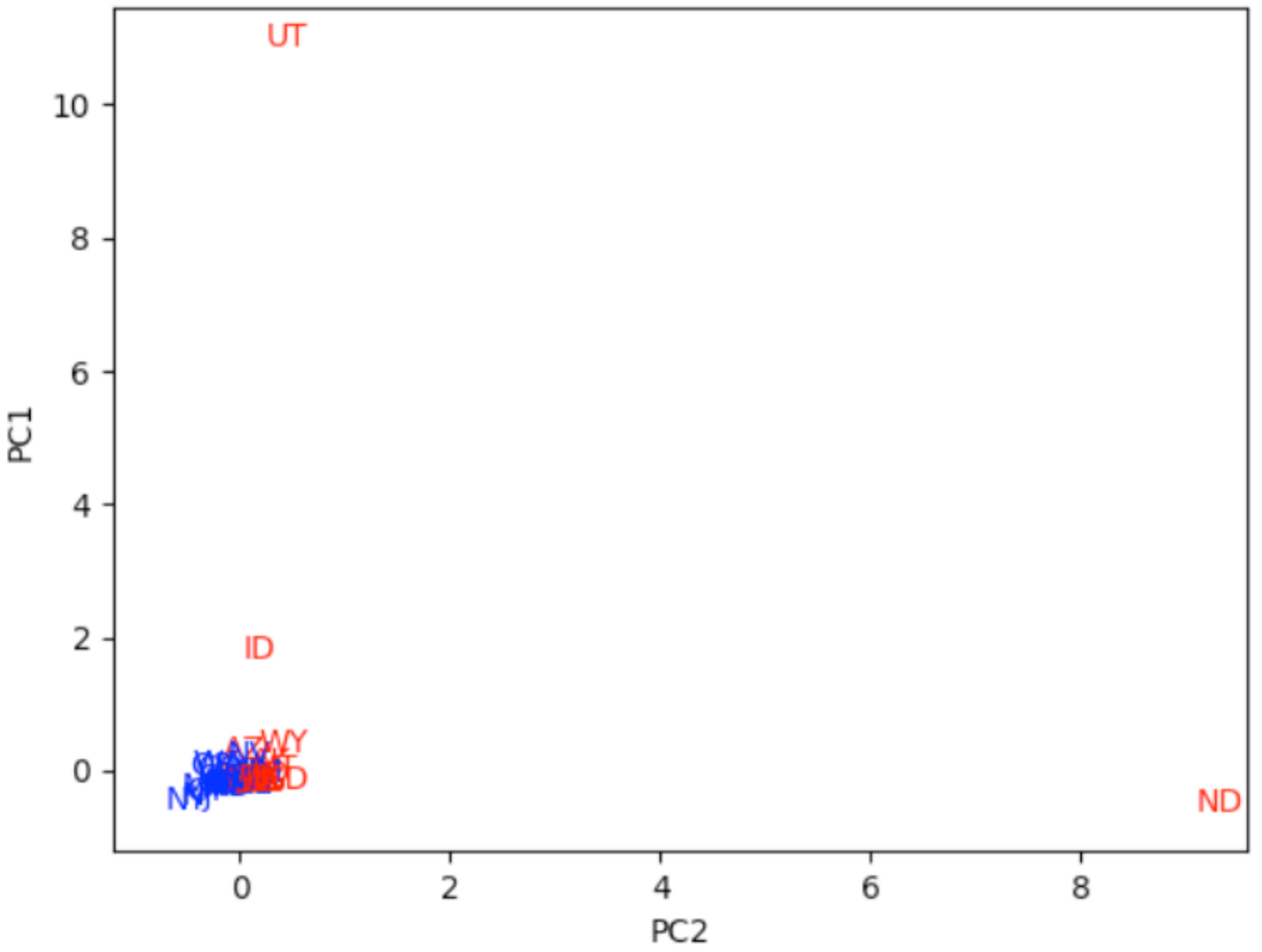
WE the PEOPLE

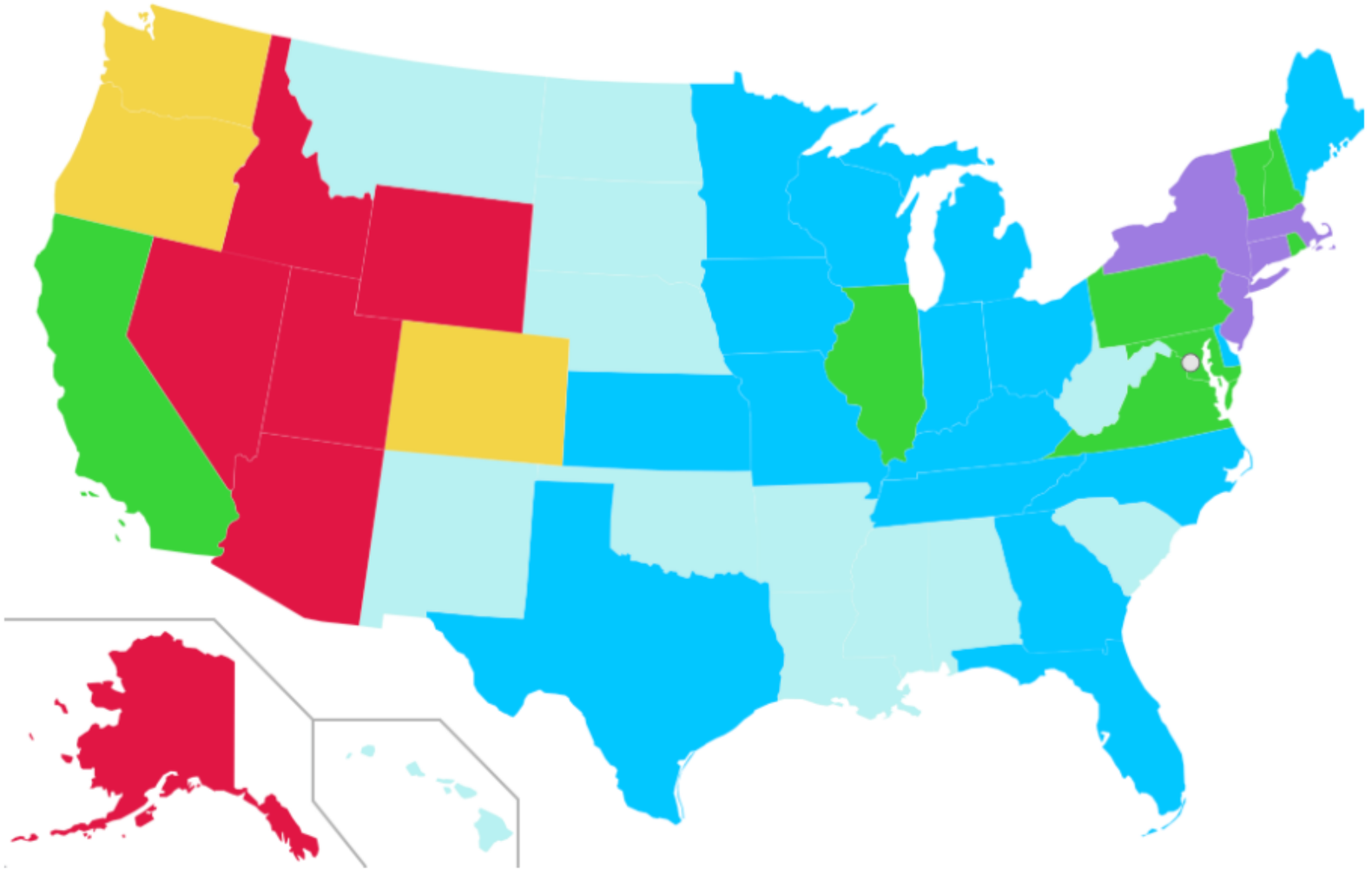
YOUR **VOICE** IN THE WHITE HOUSE



state	4e7b352b4b	4e7b35898d	4e7b362370	4e7b37f611f	4e7b3978b2	4e7b3ea34b	4e7b3ea711	4e7b3f1
AK	0.03519982	0.11827138	0.0183039	0.0183039	0.00281599	0.00140799	0.05068773	0.25343
AL	0.01778341	0.05439631	0.00815945	0.01087926	0.00167373	0.00146452	0.02636129	0.13452
AR	0.01611842	0.06035835	0.01680431	0.00480123	0.00205767	0.00102884	0.02606383	0.17353
AZ	0.04098863	0.10215868	0.02018142	0.01314139	0.00219023	0.00219023	0.0564767	0.21354
CA	0.04093525	0.08901068	0.02582276	0.00593226	0.00233532	0.00252322	0.05865149	0.20738
CO	0.04712483	0.12566621	0.04155734	0.01153266	0.00338026	0.00198839	0.07476344	0.29149
CT	0.04001011	0.08589582	0.02490139	0.00447666	0.00083937	0.00083937	0.05064216	0.22998
DE	0.05345604	0.07350206	0.0345237	0.00890934	0.003341	0	0.04677404	0.17818
FL	0.04027905	0.07588512	0.01676177	0.01081896	0.00167618	0.00096507	0.04099016	0.21709
GA	0.02217742	0.06330645	0.01300403	0.00604839	0.00110887	0.00080645	0.03961694	0.18528
HI	0.02719986	0.07645367	0.01617289	0.02278907	0.00661618	0.00294053	0.04337275	0.14702
IA	0.02232176	0.07188919	0.02232176	0.00623696	0.00164131	0.00131304	0.04825439	0.21993
ID	0.02870663	0.11227483	0.02041361	0.00701718	0.00127585	0.00318963	0.05868911	0.25899
IL	0.03491644	0.08160159	0.02143308	0.00296166	0.00335135	0.00194846	0.0524526	0.21479
IN	0.02853264	0.08606062	0.01850766	0.00431845	0.0015423	0.00092538	0.05074183	0.22903
KS	0.02138012	0.08201554	0.01857617	0.0112158	0.00245346	0.00070099	0.04451271	0.24569
KY	0.02488842	0.07581751	0.01267466	0.00921793	0.00138269	0.00184359	0.0414807	0.17882
LA	0.01808808	0.04632313	0.00904404	0.006397	0.00044117	0.00088235	0.02735271	0.15661
MA	0.04703993	0.13149798	0.04276357	0.00595635	0.00473454	0.00778908	0.0881235	0.31018
MD	0.0386244	0.10530779	0.03689237	0.01160464	0.00207844	0.00242485	0.06581737	0.25651
ME	0.04140441	0.11593234	0.03989879	0.0105393	0.00150561	0.00376404	0.07001109	0.25219

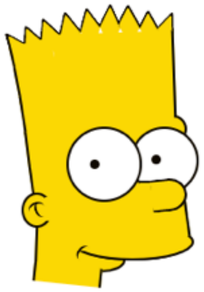

WY
AZ
NV
AK
WA
OR
CO
MT
HI
NM
NE
TX
CA
SD
NH
VA
KY
WV
MS
IL
PA
GA
ME
MA
CT
NJ
NY
0.2
0.0
−0.2
−0.4
−0.6
PC1
−0.6
−0.4
−0.2
0.0
0.2
PC2

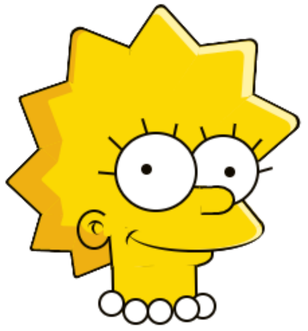

UT
ID
ND
10
8
6
4
2
0
PC1
0
2
4
6
8
PC2

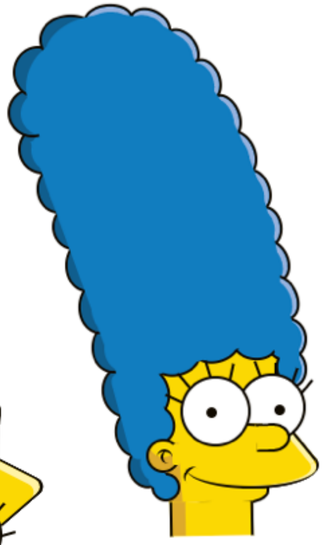


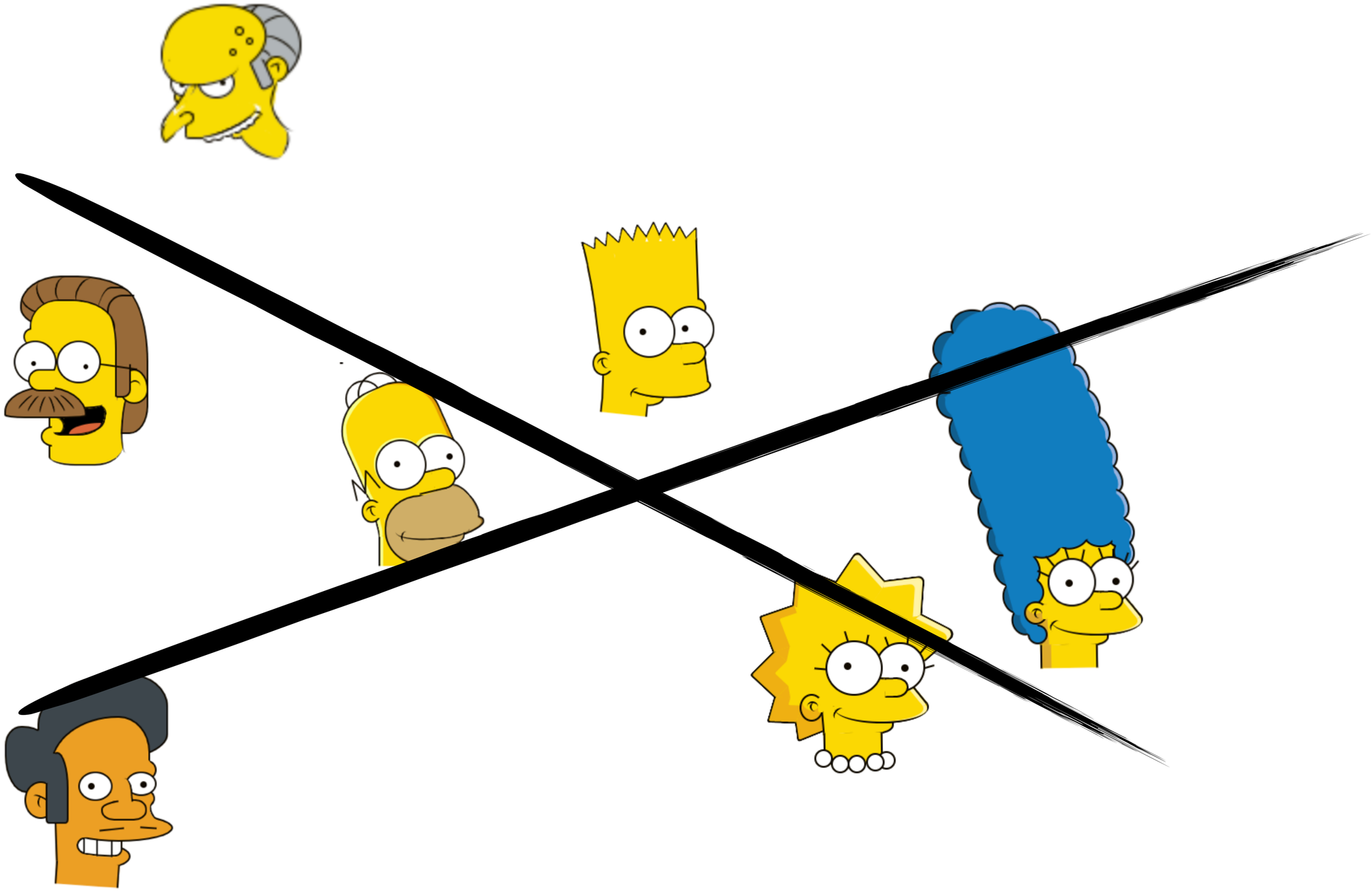
https://www.kaggle.com/wcukierski/the-simpsons-by-the-data

- What episodes do they appear in?
- What scenes do they appear in?
- What words do they use?
- Which other characters do they mention?





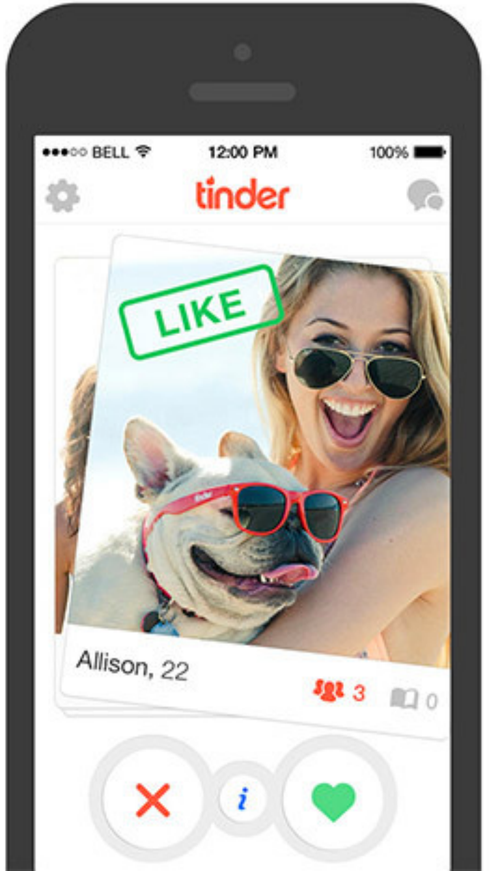


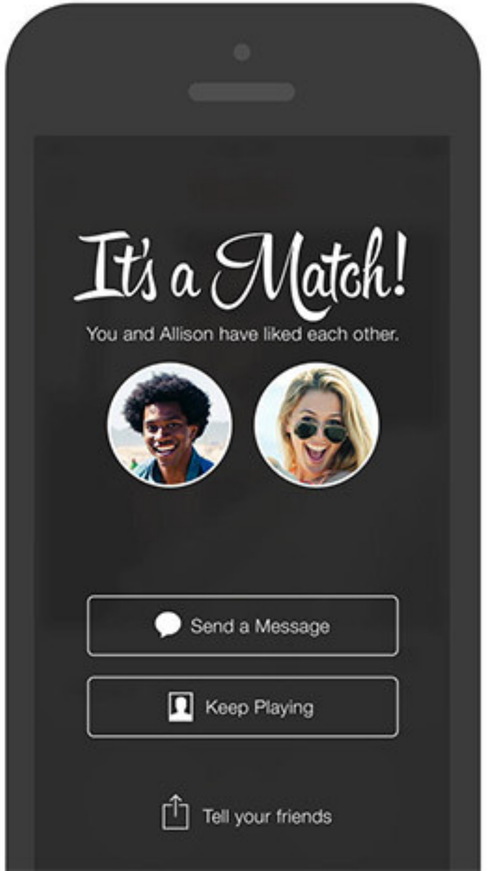


Google Arts & Culture **TSNE MAP**

- Vintage posters
- Lost in space
- Mostly the Brown sisters
- Ballet dancers
- Black & white vs contrasted color
- Rough colors
- Blues
- **Impressionists Hill**
- Landscapes forest
- Horses
- Still life with Arcimboldo
- Still life overview
- Painted in gold
- Shore of portraits
- Japanese scenes of the XIX century
- A Mountain of Sketches

SHOWING ARTWORKS FROM 623 PARTNER COLLECTIONS


BELL
12:00 PM
100%
tinder
LIKE
Allison, 22
3
0


It's a Match!
You and Allison have liked each other.
Send a Message
Keep Playing
Tell your friends

The burning question is, did I solve my own problem? The answer is a resounding yes. In the interest of privacy of my relationship, I won't be publishing too many details on the blog but I will leave you with a picture of myself with the most beautiful woman in the world...

- Justin Long

Questions?