Dimensionality Reduction

Darren Reger Lecture for Galvanize DSI

What's My Dimensionality?

8 features → dimensionality of 8

	mpg	cylinders	displacement	horsepower	weight	acceleration	model	origin	car_name
0	18	8	307	130.0	3504	12.0	70	1	chevrolet chevelle malibu
1	15	8	350	165.0	3693	11.5	70	1	buick skylark 320
2	18	8	318	150.0	3436	11.0	70	1	plymouth satellite
3	16	8	304	150.0	3433	12.0	70	1	amc rebel sst
4	17	8	302	140.0	3449	10.5	70	1	ford torino

handwritten digits made of images of 28 × 28 pixels (horizontally × vertically)

28 × 28 = 784 pixels are used to represent a handwritten digit → 784 features → dimensionality of 784

"dimensionality" = "number of dimensions" = "number of features/predictors"

Dim Reducers

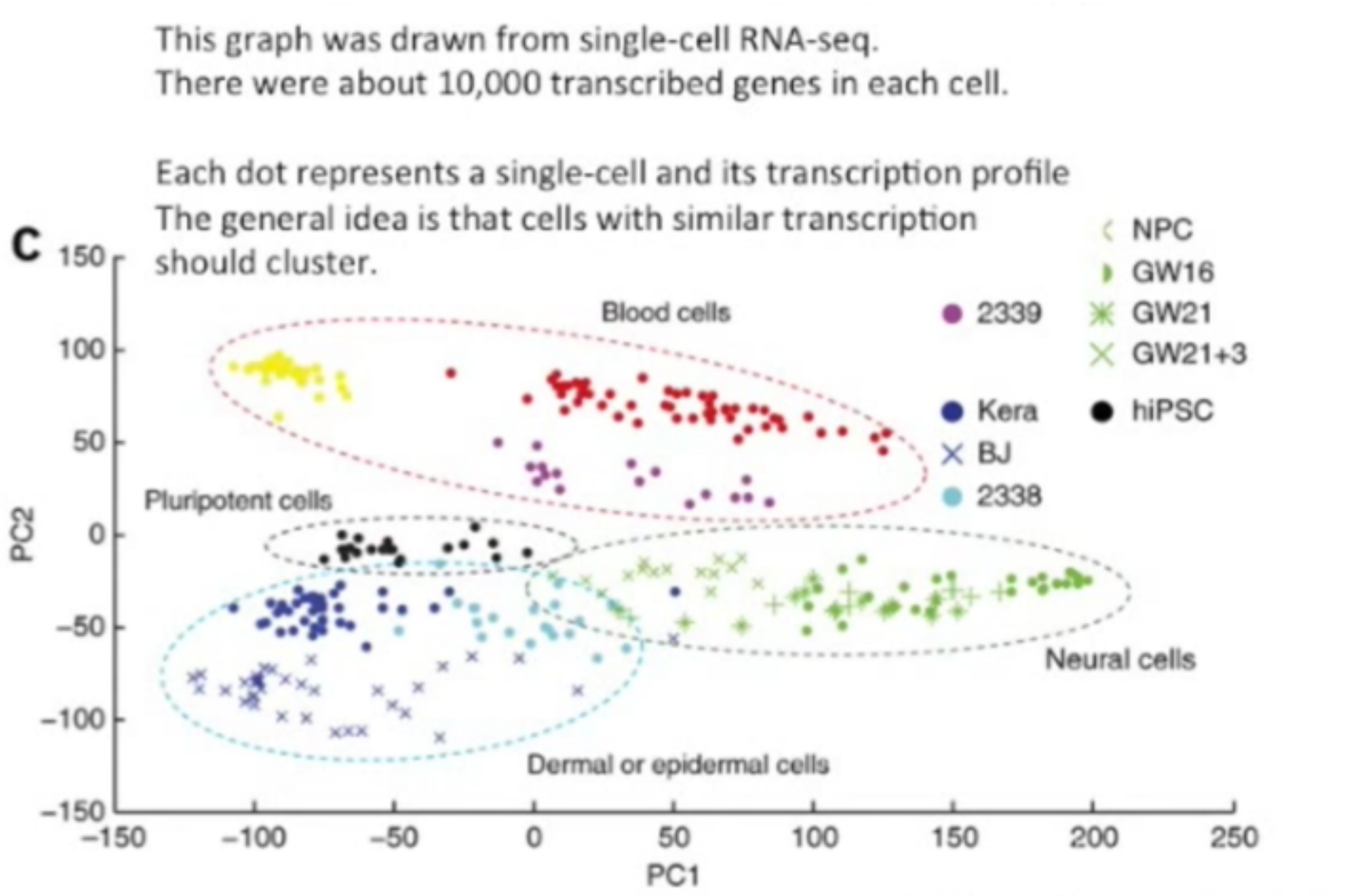
- Lasso
- Stepwise Selection
- Relaxed Lasso
- **PCA**

Why Reduce Dimensionality?

Do you have any of the following?

1. A need to visualize data in 2D?
2. Redundant and/or correlated features?
3. More features than you know what to do with?
4. So many features that storing all your data is taking too much space?

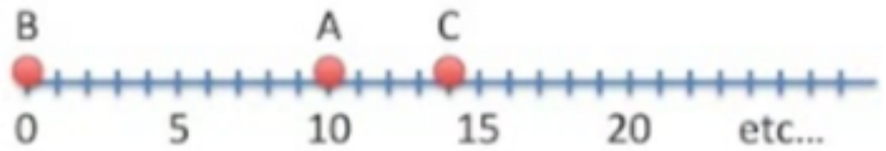
PCA in Action



Pollen et al. Nature Biotechnology 2014

Dimensions

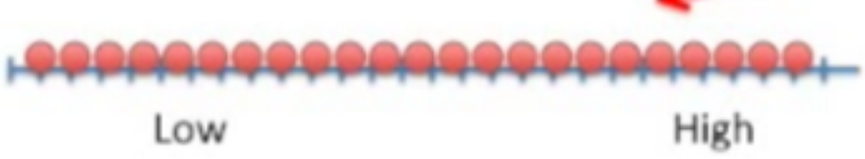
1-Dimension (1-D) = a number line



A pretend RNA-seq data set for a single cell:

Gene:	Reads:
A	10
B	0
C	14
...	...

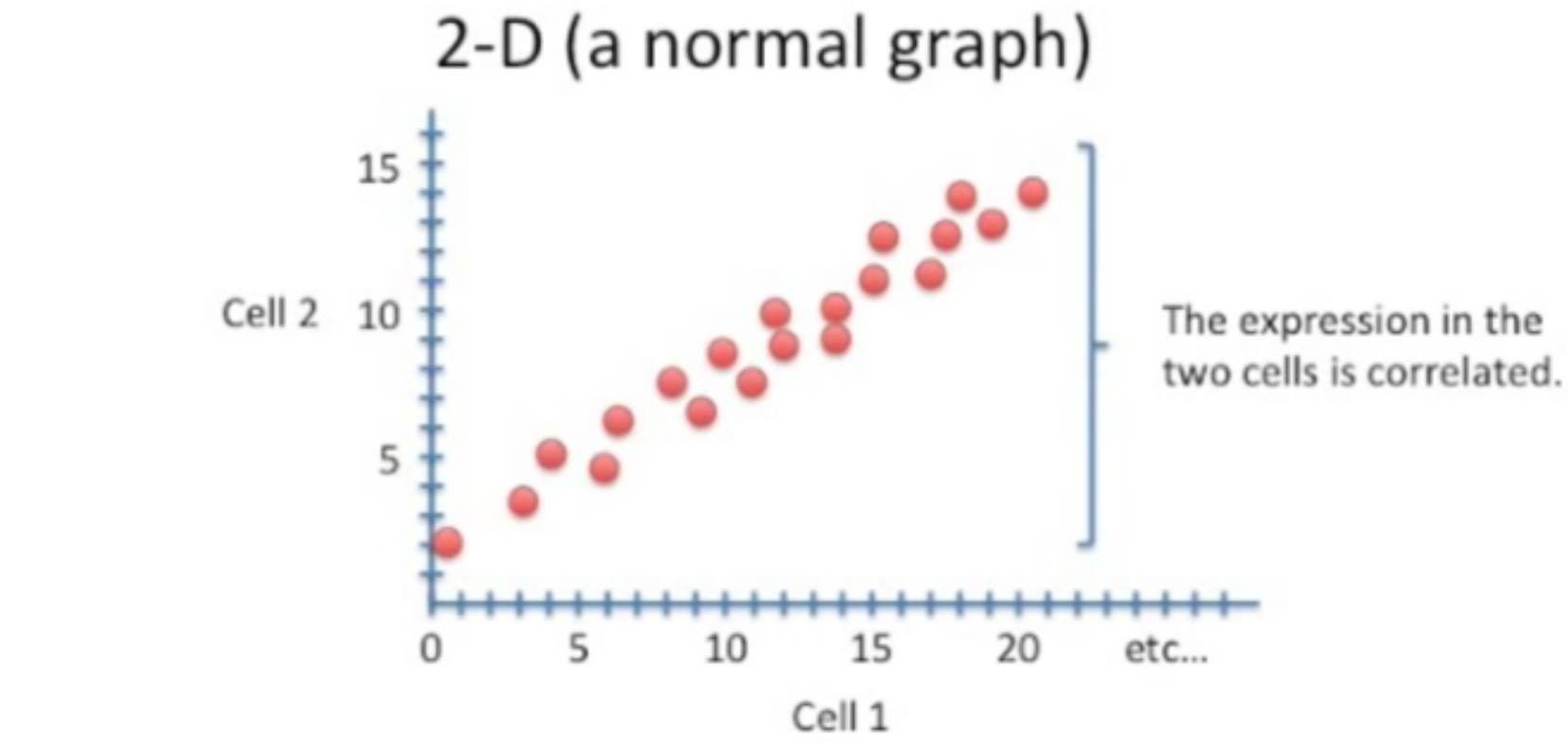
If we plotted all genes, we might see something like this or this.



A uniform distribution of transcripts

A non-uniform distribution of transcripts
(some genes are low, some are high)

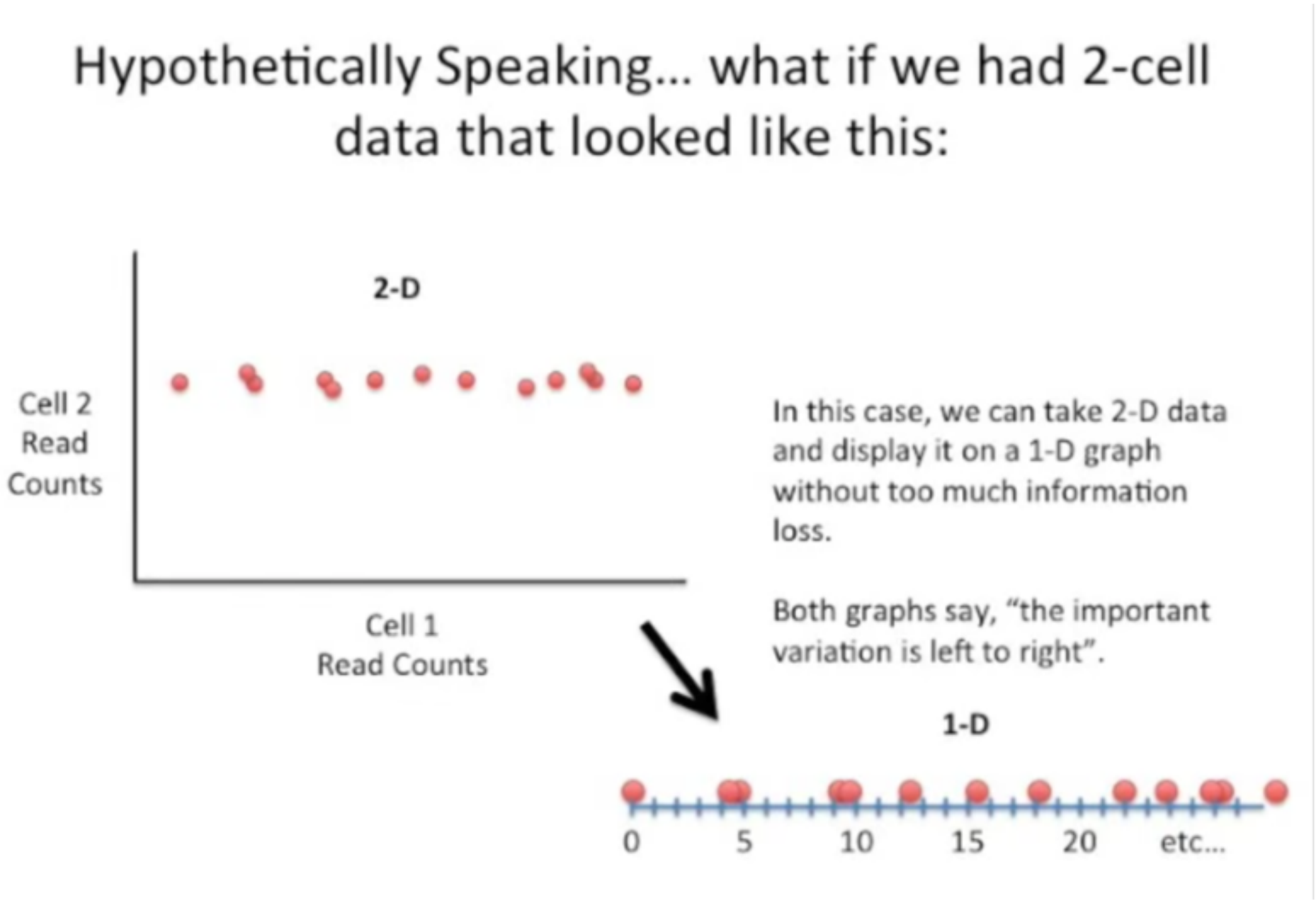
2-Dimensions Still Easy



A pretend RNA-seq data set for two single cells:

Gene:	Cell1 Reads:	Cell2 Reads:
A	10	8
B	0	2
C	14	10
...	...	...

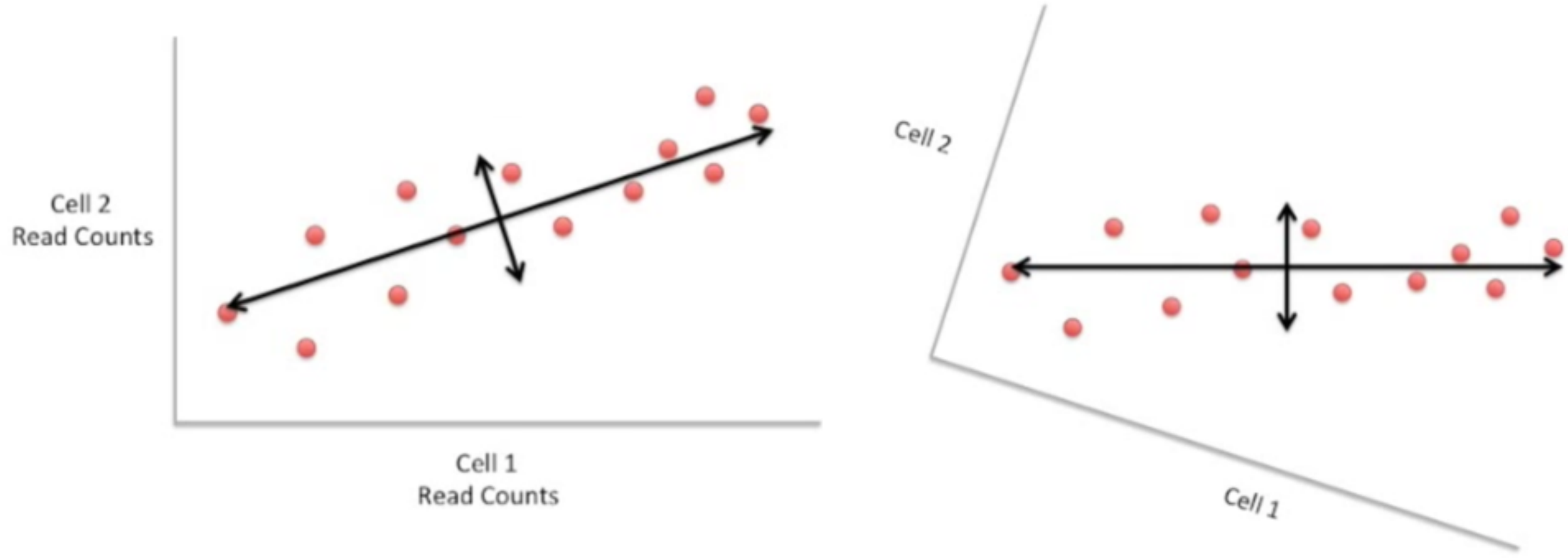
Do We Need Them All?



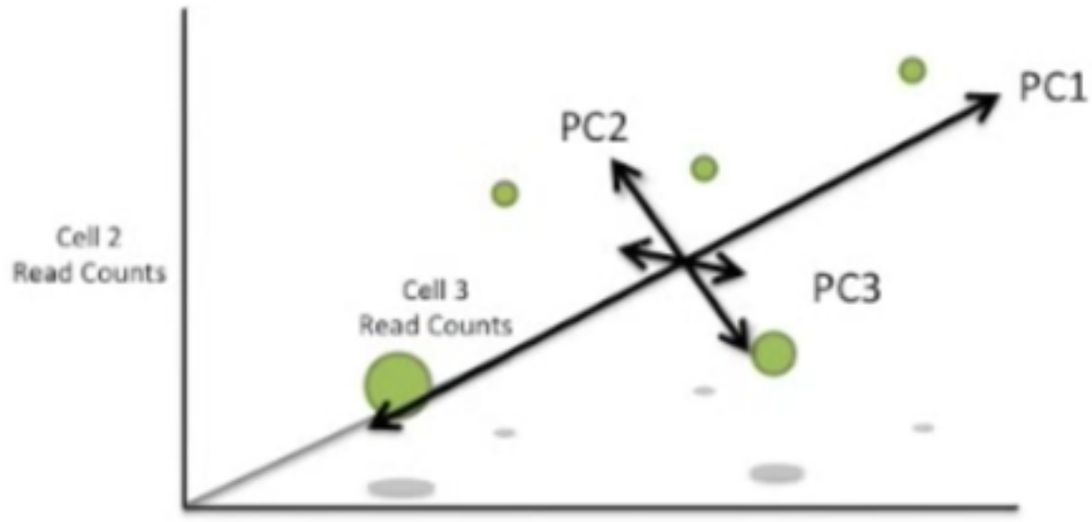
Teapot and TV Discussion

https://clara.io/view/8d9a8181-f1ce-4340-b24f-e36bbaf318f7/webgl

What Does PCA Do?



Extension to 3 Dims



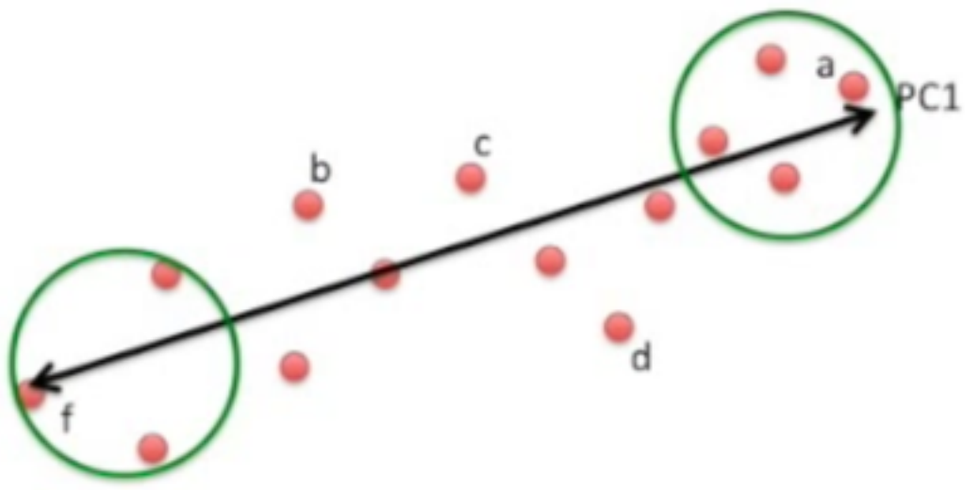
Just like before, PC1 would span the direction of the most variation.
PC2 would span the direction of the 2nd most variation.
However, since we have another direction we can have variation, we need another PC.

PC3 spans the direction of the 3rd most variation.

What if we had 1000 dims?

How to get loadings

The length and direction of PC1 is mostly determined by the circled genes.



Some genes have more influence on PC1 than others.

Gene	Influence on PC1	In numbers
a	high	10
b	low	0.5
c	low	0.2
d	low	-0.2
e	high	13
f	high	-14
...	...	

Genes with little influence on PC1 get values close to zero, and genes with more influence get numbers further from zero.

How to get the features in PCA space

Using the two Principle Components to plot cells

Combining the read counts for all genes in a cell to get a single value.

The original read counts

Gene	Cell1	Cell2
a	10	8
b	0	2
c	14	10
d	33	45
e	50	42
f	80	72
g	95	90
h	44	50
i	60	50
etc	etc	etc

PC1

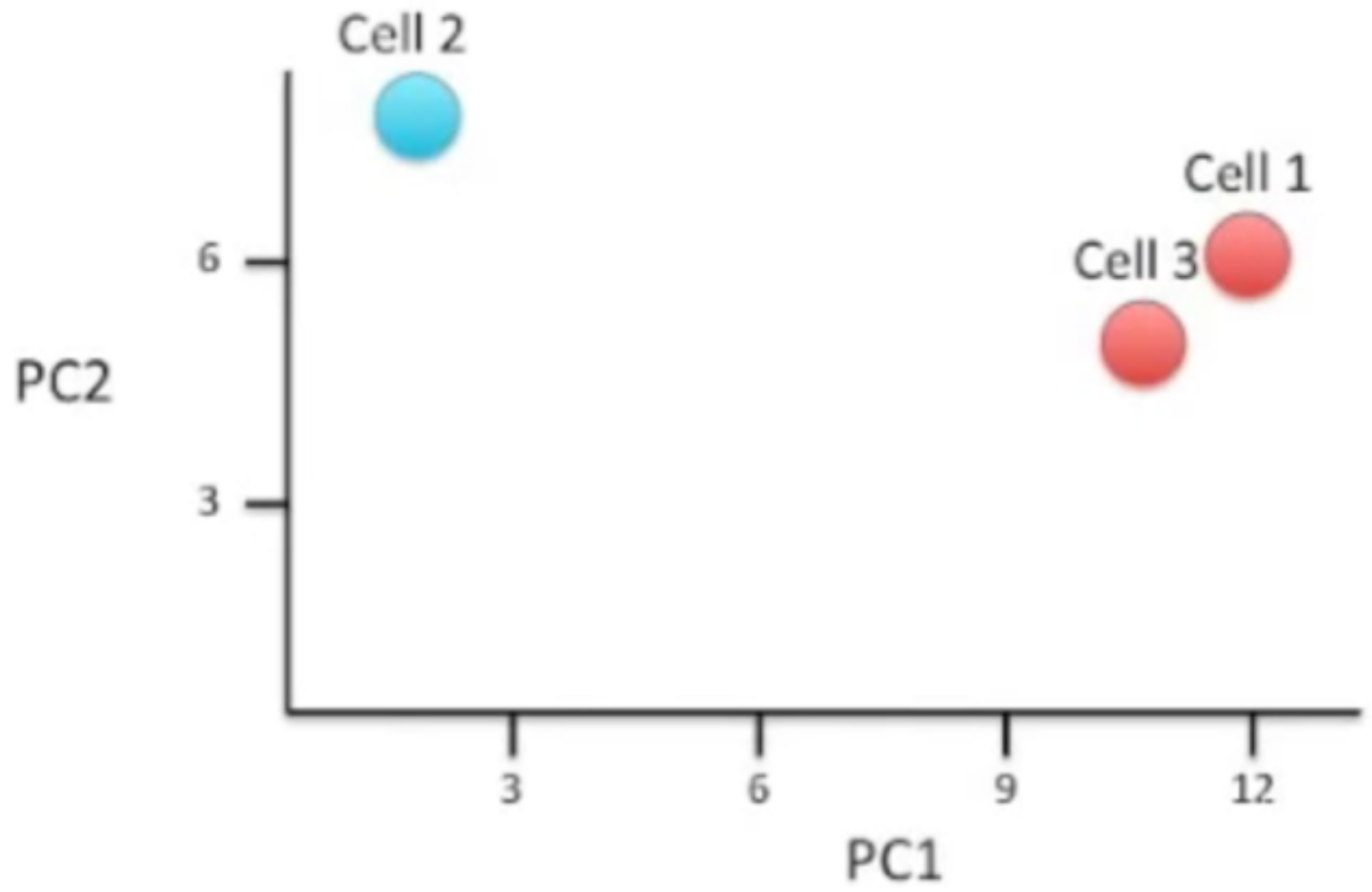
Gene	Influence on PC1	In numbers
a	high	10
b	low	0.5
c	low	0.2
d	low	-0.2
e	high	13
f	high	-14
...	...	

PC2

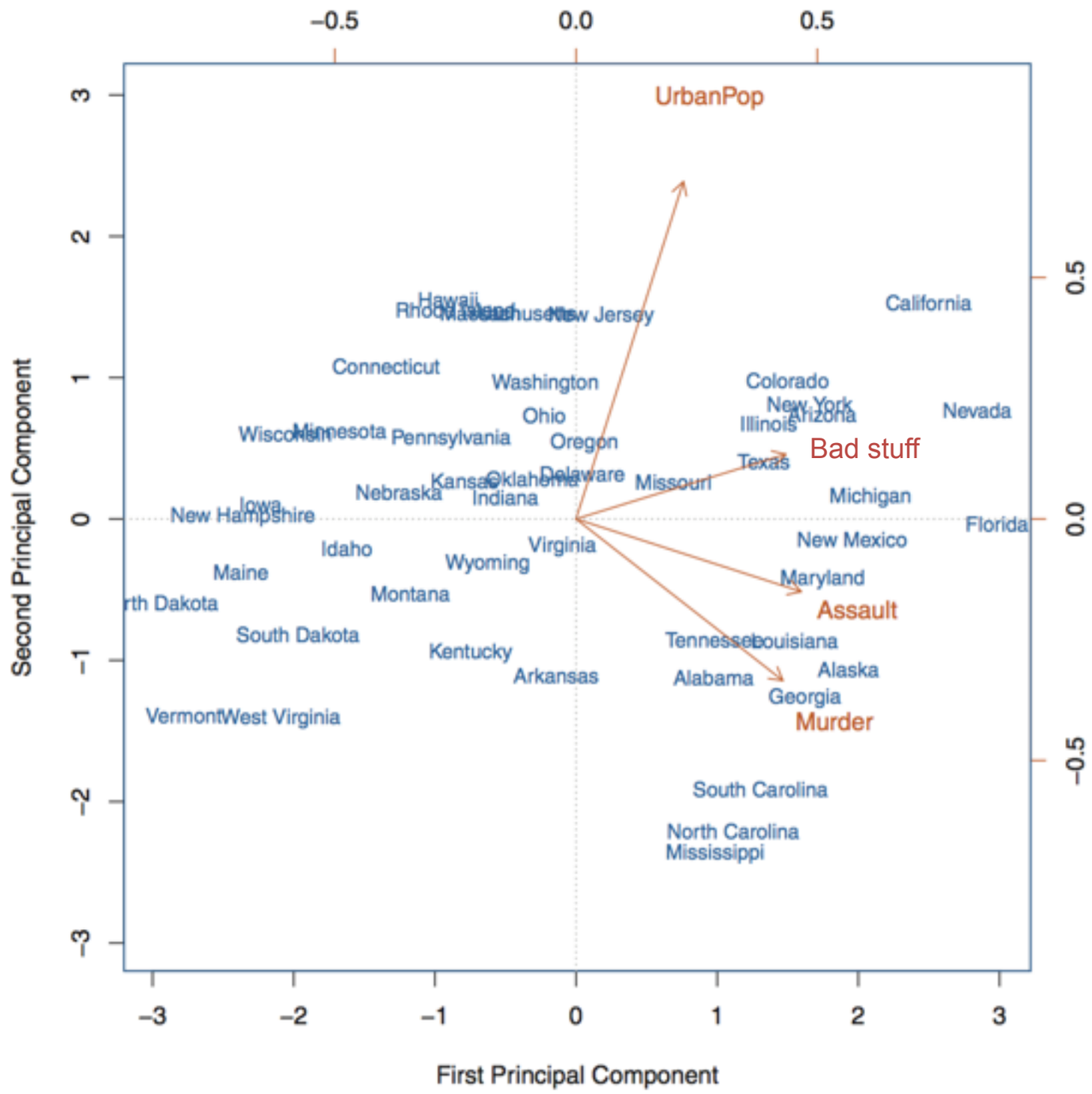
Gene	Influence on PC2	In numbers
a	medium	3
b	high	10
c	high	8
d	high	-12
e	low	0.2
f	low	-0.1
...	...	

Cell1 PC1 score = (read count * influence) + ... for all genes

Visualizing the Features in PCA Space

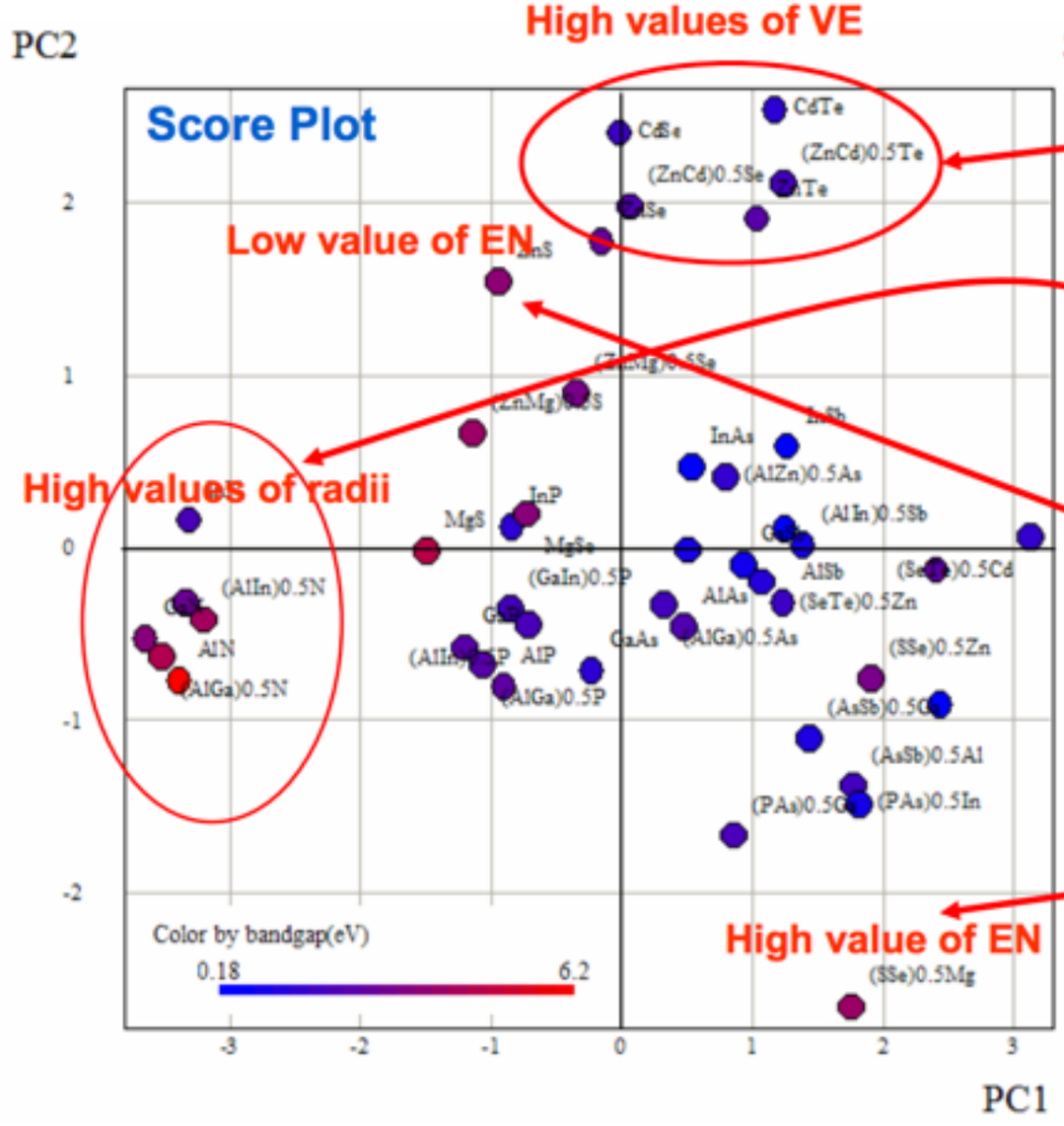


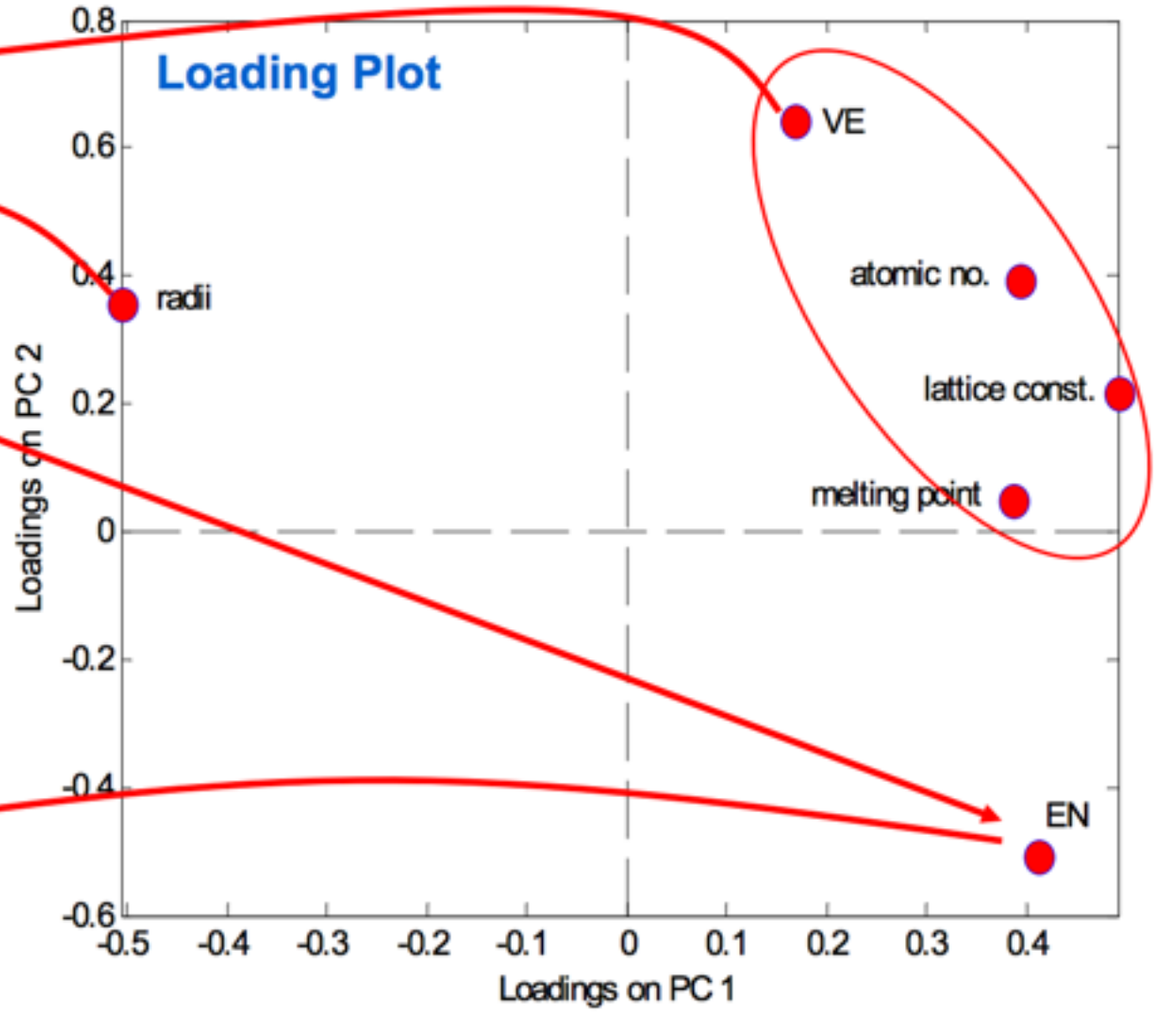
Loading Plots



Example: Interpretation of scores and loadings with semiconductor data (Continued)

By comparing the score and loading plot,
We can identify the relationships between
Samples and variables



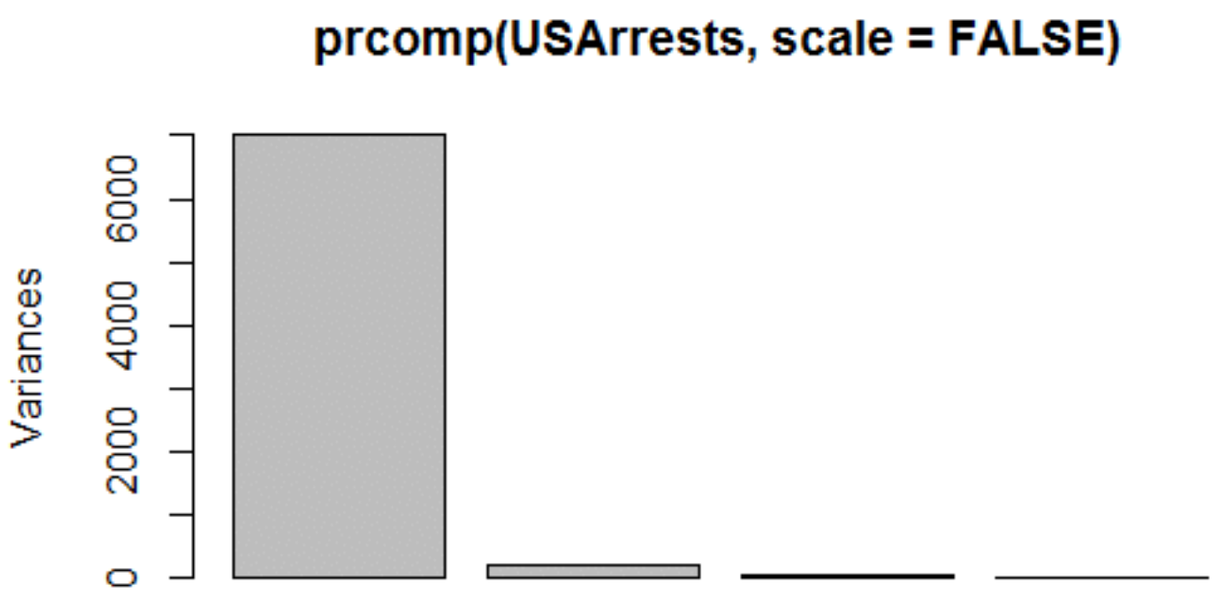


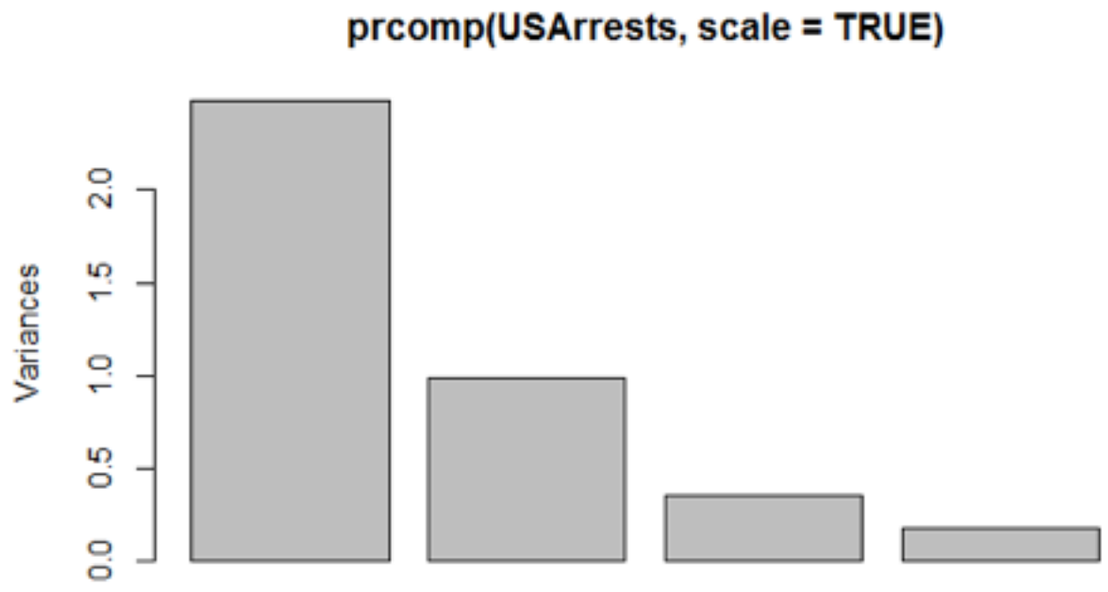
On the score plot,
"Sit together": similar behavior between samples
ex.) Nitrides

On the loading plot,
"Sit together": similar behavior between variables
ex.) VE, atomic no. lattice const., and melting point

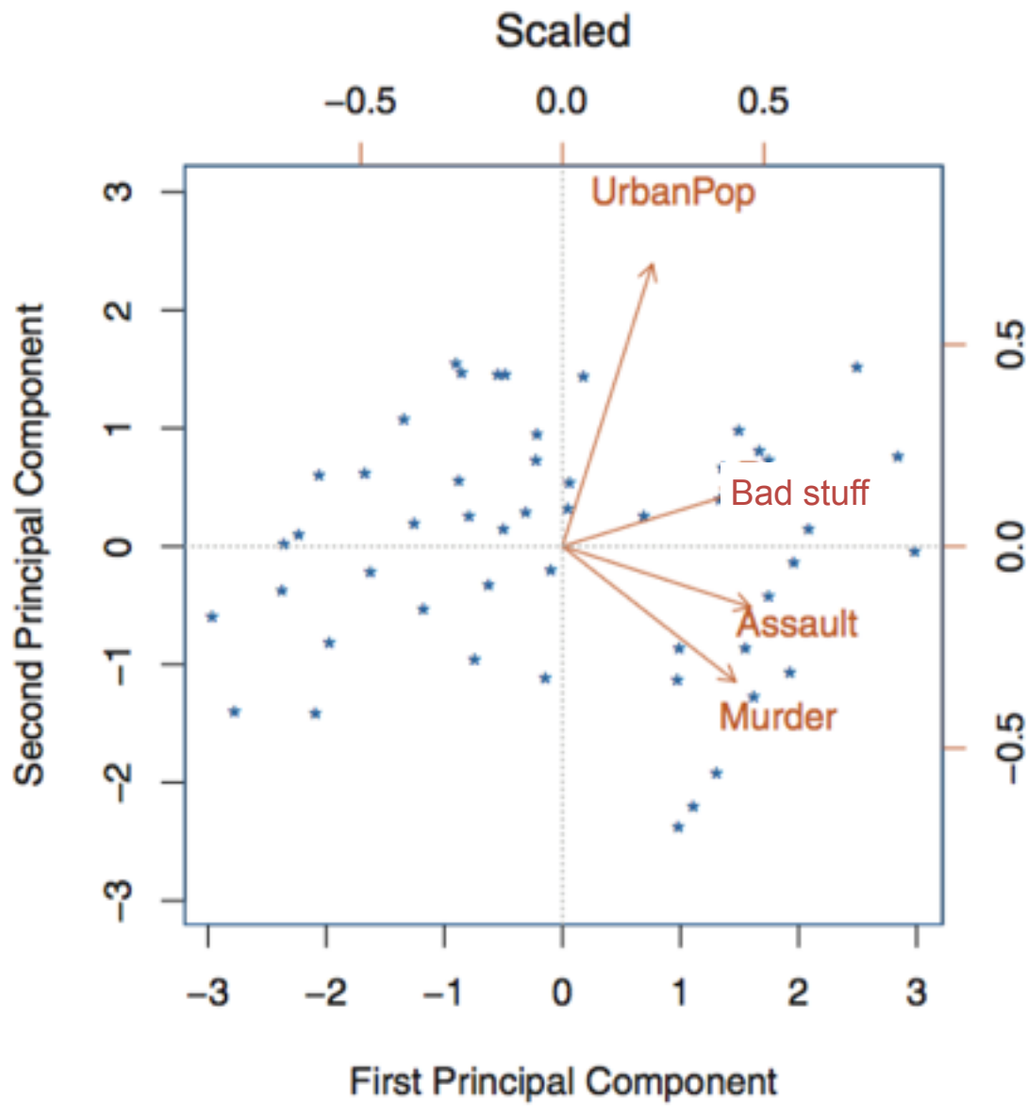
Why Standardize?

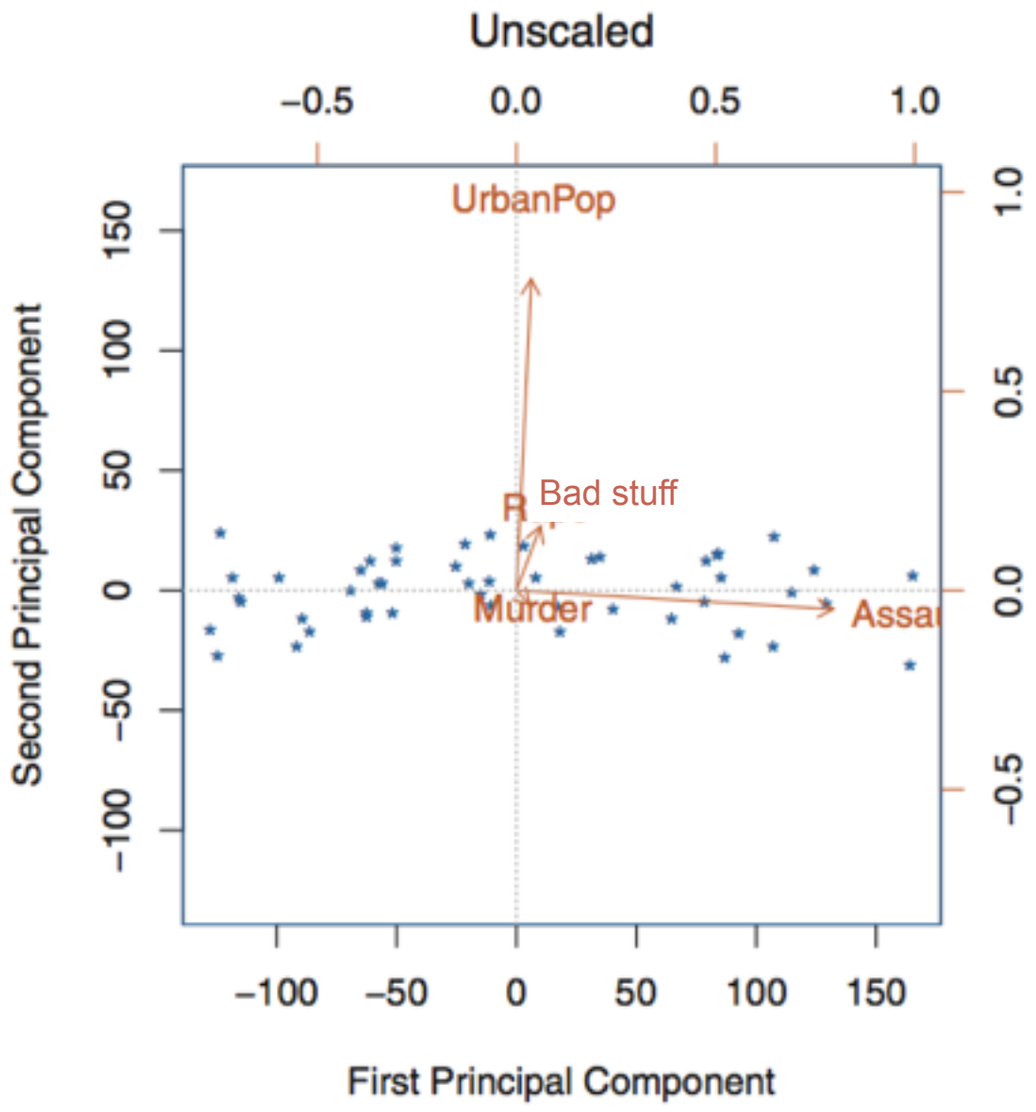
	Murder	Assault	UrbanPop	Bad Stuff
Murder	18.97	291.06	4.38	22.99
Assault	291.06	6945.16	312.27	519.26
UrbanPop	4.38	312.27	209.51	55.76
Bad Stuff	22.99	519.26	55.76	87.72





Why Standardize?





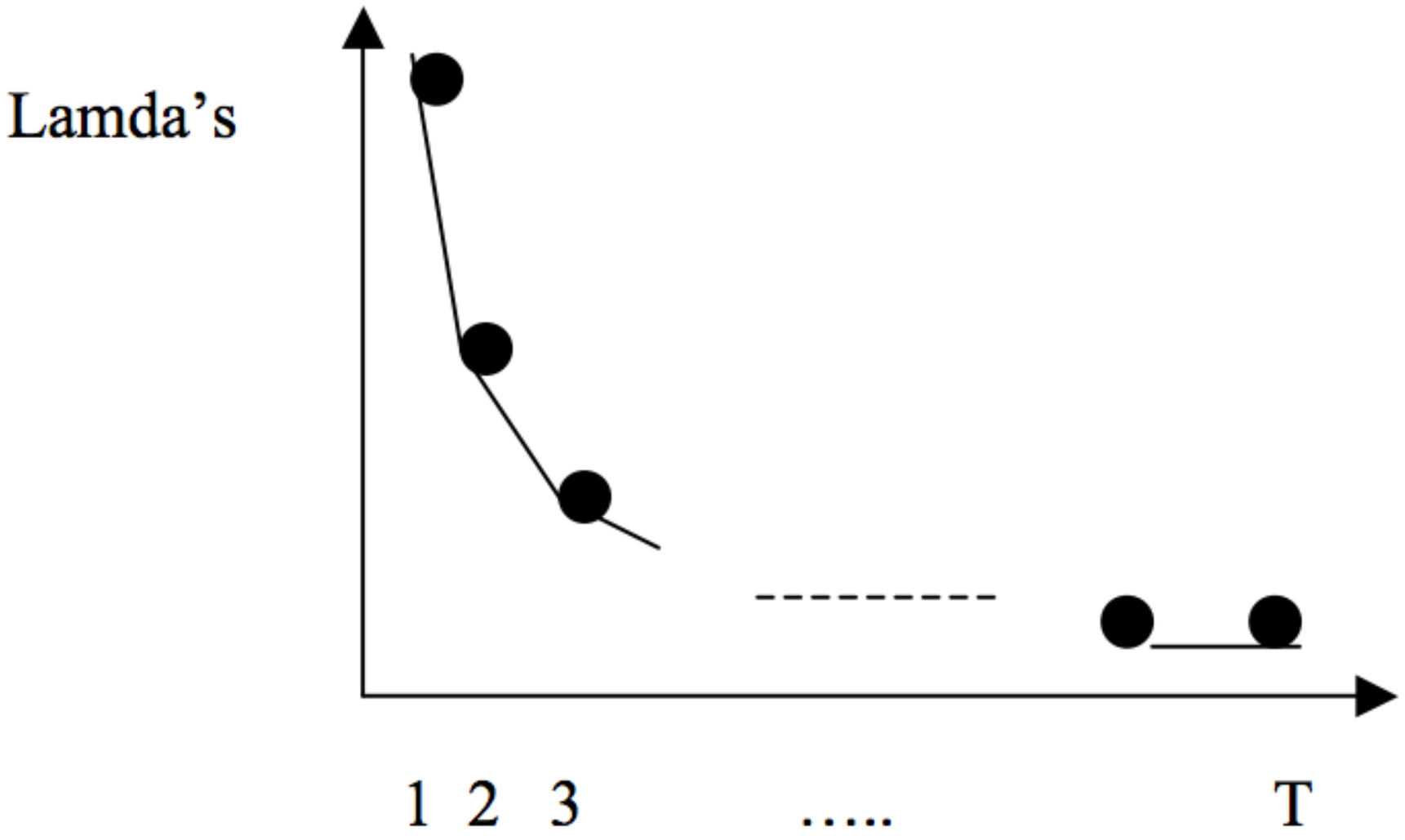
PCA Summary

Math to come in the afternoon

- create the centered design matrix X (n rows/observations × p columns/features)
  - (meaning that each column vector is centered around its mean)
- calculate the covariance matrix X^TX (a p × p square matrix)
- the principal components are the eigenvectors of the covariance matrix; the principal components' variance (σ^2) is
  - ordering the principal components/eigenvectors by decreasing variance/eigenvalue, you get an orthogonal basis capturing the directions of the most-to-least variance of your data

Scree Plots

Scree plot: eigenvalues in non-increasing order



Scree Plots

1. Consider the proportion of explained variability, and retain as many directions as needed to explain a selected proportion
2.

$$\frac{\lambda_1}{\sum_{j=1}^{T}\lambda_j} \quad \frac{\lambda_2}{\sum_{j=1}^{T}\lambda_j} \quad \cdots \quad \frac{\lambda_T}{\sum_{j=1}^{T}\lambda_j}$$

One by one

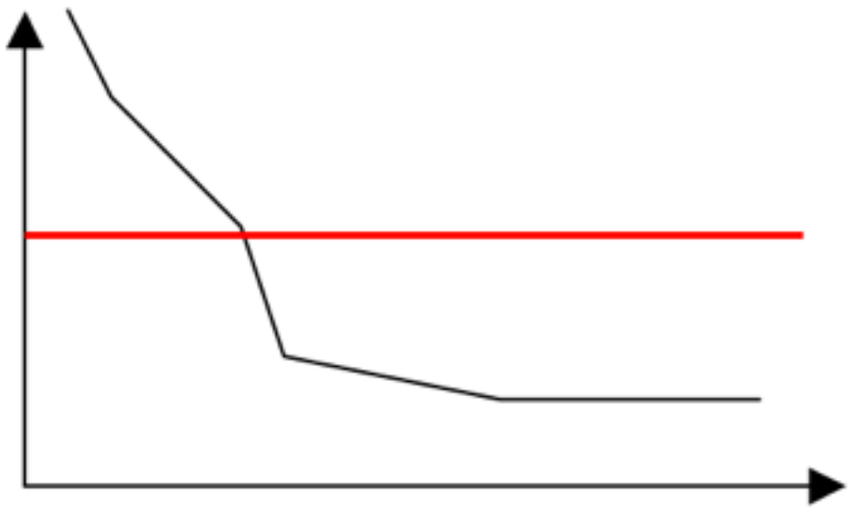
$$\frac{\lambda_1}{\sum_{j=1}^{T}\lambda_j} \quad \frac{\lambda_1+\lambda_2}{\sum_{j=1}^{T}\lambda_j} \quad \cdots \quad \frac{\sum_{j=1}^{T}\lambda_j}{\sum_{j=1}^{T}\lambda_j}=1$$

Cumulative. Stop when you reach, say, .80 i.e. 80%

Scree Plots

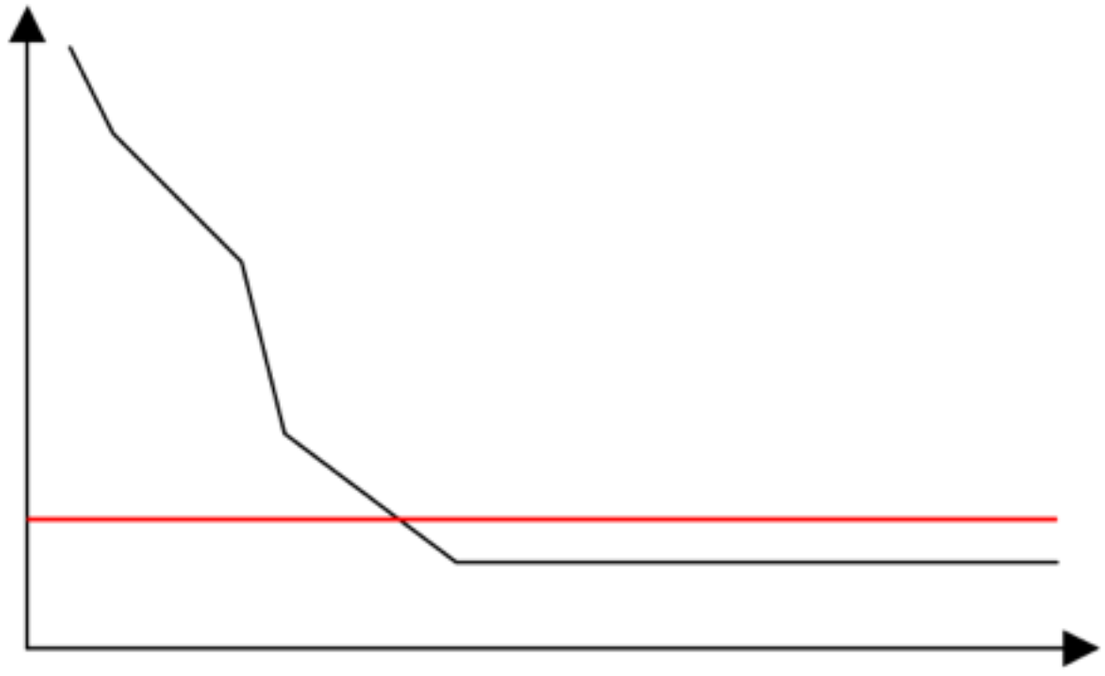
3. Consider the average explained variability per component, and retain directions with an explanatory capability above average – on the scree plot:

$$\overline{\lambda} = \frac{1}{T}\sum_{j=1}^{T} \lambda_j$$



Scree Plots

4. Look for bends in the scree plot. If there is a clear bend, keep directions associated with eigenvalues before the bend – those afterwards have comparable, small(er) size (smaller the more they are)



Should I Reduce Dimensions?

- The correlation matrix shows… correlations
- Dimensionality reduction is good when there is correlation
- Rule of thumb: If many pairwise correlations have a magnitude greater than 0.3, PCA will probably work

Correlation Matrix

	Red Meat	White Meat	Eggs	Milk	Fish	Cereals	Starch	Nuts	Fruits & Vegetables
Red Meat	1	0.153	0.58561	0.50293	0.06096	-0.49988	0.13543	-0.34945	-0.07422
White Meat	0.153	1	0.62041	0.28148	-0.23401	-0.4138	0.31377	-0.63496	-0.06132
Eggs	0.58561	0.62041	1	0.57553	0.06557	-0.71244	0.45223	-0.55978	-0.04552
Milk	0.50293	0.28148	0.57553	1	0.13788	-0.59274	0.22241	-0.62109	-0.40836
Fish	0.06096	-0.23401	0.06557	0.13788	1	-0.52423	0.40385	-0.14715	0.26614
Cereals	-0.49988	-0.4138	-0.71244	-0.59274	-0.52423	1	-0.53326	0.651	0.04655
Starch	0.13543	0.31377	0.45223	0.22241	0.40385	-0.53326	1	-0.47431	0.08441
Nuts	-0.34945	-0.63496	-0.55978	-0.62109	-0.14715	0.651	-0.47431	1	0.37497
Fruits & Vegetables	-0.07422	-0.06132	-0.04552	-0.40836	0.26614	0.04655	0.08441	0.37497	1

Interpreting PCA

AP Points from Week 6				Offense				Defense			
TEAM	AP Points	Wins	Losses	Yards/Game	Pass Y/G	Rush Y/G	Points/G	Yards/Game	Pass Y/G	Rush Y/G	Points/G
Alabama	1,514	5	-	484	253	232	44	256	188	68	13
Ohio State	1,451	4	-	576	244	332	57	238	140	98	9
Clemson	1,403	5	-	463	296	168	35	288	160	129	16
Michigan	1,334	5	-	444	234	210	44	248	135	112	12
Washington	1,234	5	-	441	242	199	45	299	177	122	13
Houston	1,233	5	-	506	305	201	44	250	208	42	11
Louisville	1,160	4	1	659	350	309	58	324	186	138	26
Texas A&M	1,113	5	-	521	262	259	39	388	253	135	15
Tennessee	1,045	5	-	382	207	175	33	361	212	149	23
Miami	909	4	-	474	242	233	47	253	138	116	11
Wisconsin	882	4	1	360	199	162	26	291	201	90	12
Nebraska	821	5	-	473	238	234	37	347	195	152	18
Baylor	805	5	-	568	278	290	43	341	166	175	19
Ole Miss	712	3	2	490	333	157	42	449	234	215	30
Stanford	711	3	1	310	149	161	20	359	234	125	20
Arkansas	528	4	1	443	246	197	36	374	211	163	23
North Carolina	497	4	1	484	348	136	40	459	222	237	31
Florida	391	4	1	407	246	161	28	230	140	91	12
Boise State	385	4	-	475	298	177	34	358	286	72	18
Oklahoma	324	2	2	493	294	199	40	429	299	131	35
Colorado	276	4	1	531	313	219	43	290	149	141	21
West Virginia	240	4	-	505	318	187	29	419	227	191	20
Florida State	230	3	2	509	268	240	41	438	247	191	35
Utah	86	4	1	431	257	174	26	324	207	117	18
Virginia Tech	85	3	1	449	254	196	40	264	151	114	19

Interpreting PCA

principal Components	PC1	PC2
% of Varience Explained by Component	41%	29%
% of Varience Running Total	41%	70%
Wins	0.35	0.05
Losses	(0.39)	(0.02)
O_YDS/G	0.00	0.57
O_P YDS/G	(0.18)	0.41
O_R YDS/G	0.18	0.44
O_PTS/G	0.12	0.51
D_YDS/G	(0.45)	0.04
D_P YDS/G	(0.35)	(0.08)
D_R YDS/G	(0.34)	0.14
D_PTS/G	(0.46)	0.13

Singular Value Decomposition & Random Asides About PCA

PCA Overall

When to use:

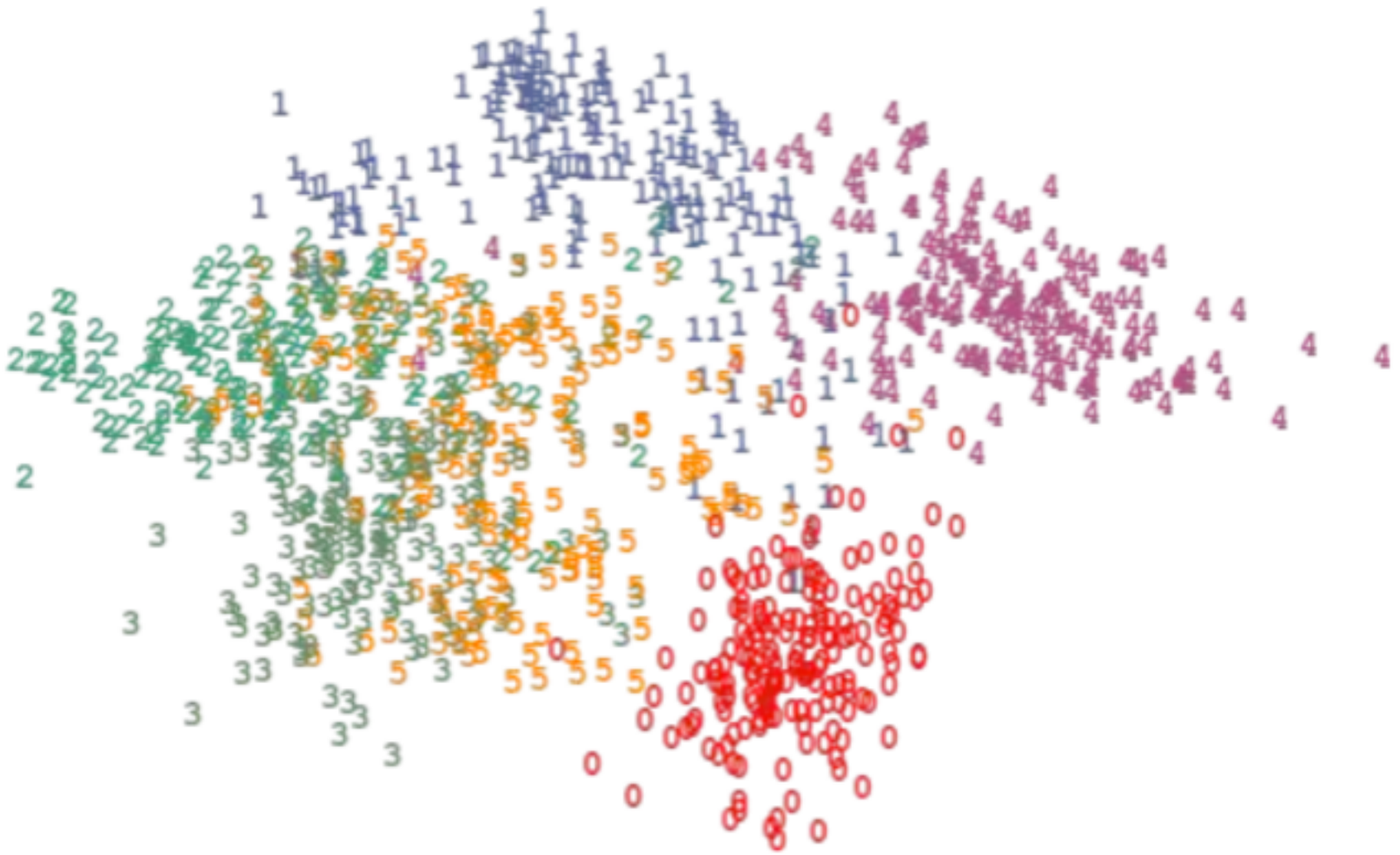
- kNN on high dimensional data
- Clustering on high dimensional data
- Visualization of un-visualizable data
- Working with images (but too lazy to use neural networks)

When to not use:

- Need interpretability of results
- Reducing dimensions isn't helpful (OLS with few predictors to start)

kNN and PCA

k-NN after PCA on MNIST to classify digits

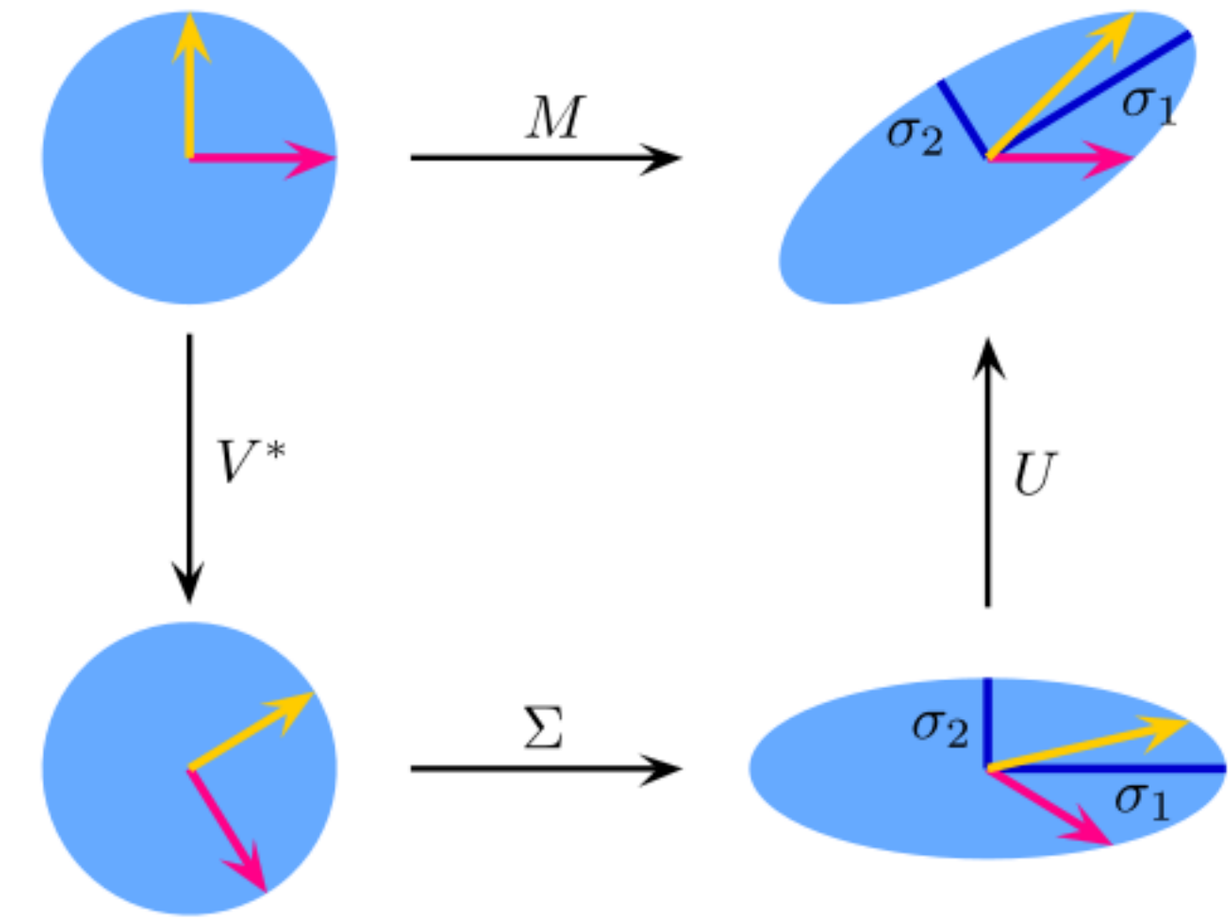


we can use the top principal components resulting from PCA (up to 4, remember?) as features to train a k-NN classifier to classify the handwritten digits

Interview Question:

Compare Lasso Regression and Logistic Regression Using PCA.

You Down With SVD?



$$M = U \cdot \Sigma \cdot V^*$$

Application of SVD