Deconstructing Data Science

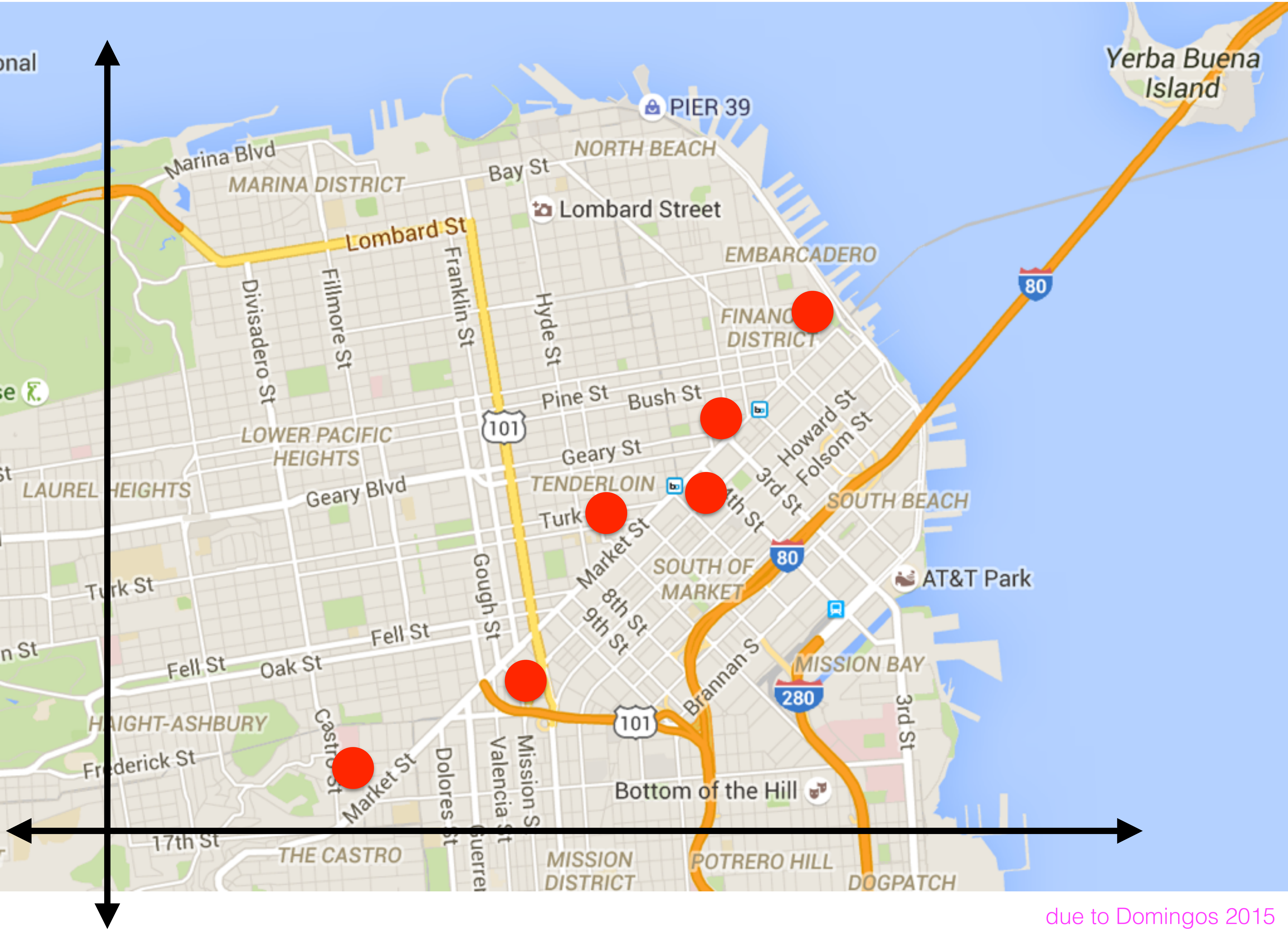
David Bamman, UC Berkeley

Info 290
Lecture 14: Principal Component Analysis

Mar 9, 2016

Unsupervised Learning

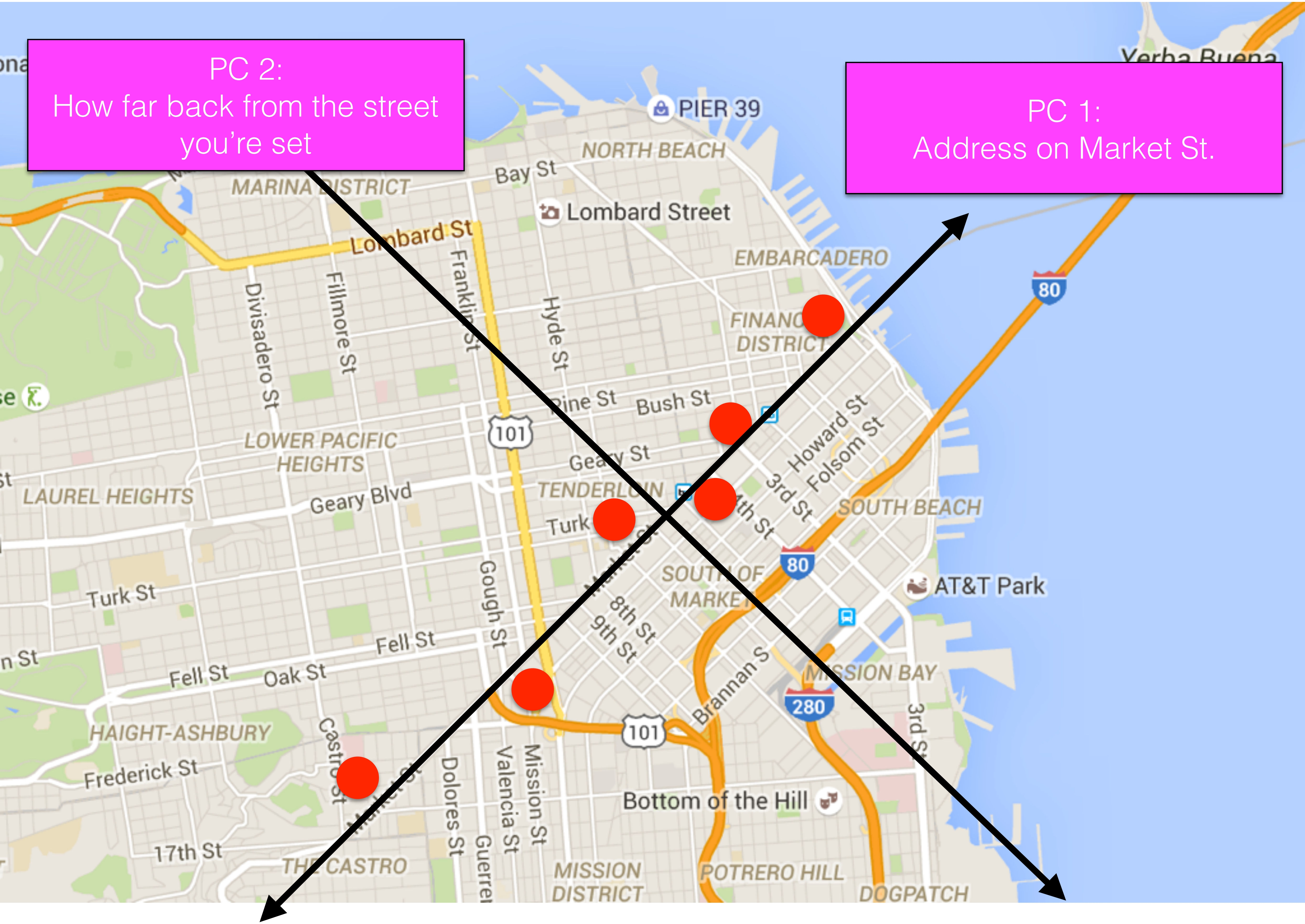
- Unsupervised learning finds *structure* in data.
    - clustering data into groups
    - discovering “factors”



due to Domingos 2015

Principal Component Analysis

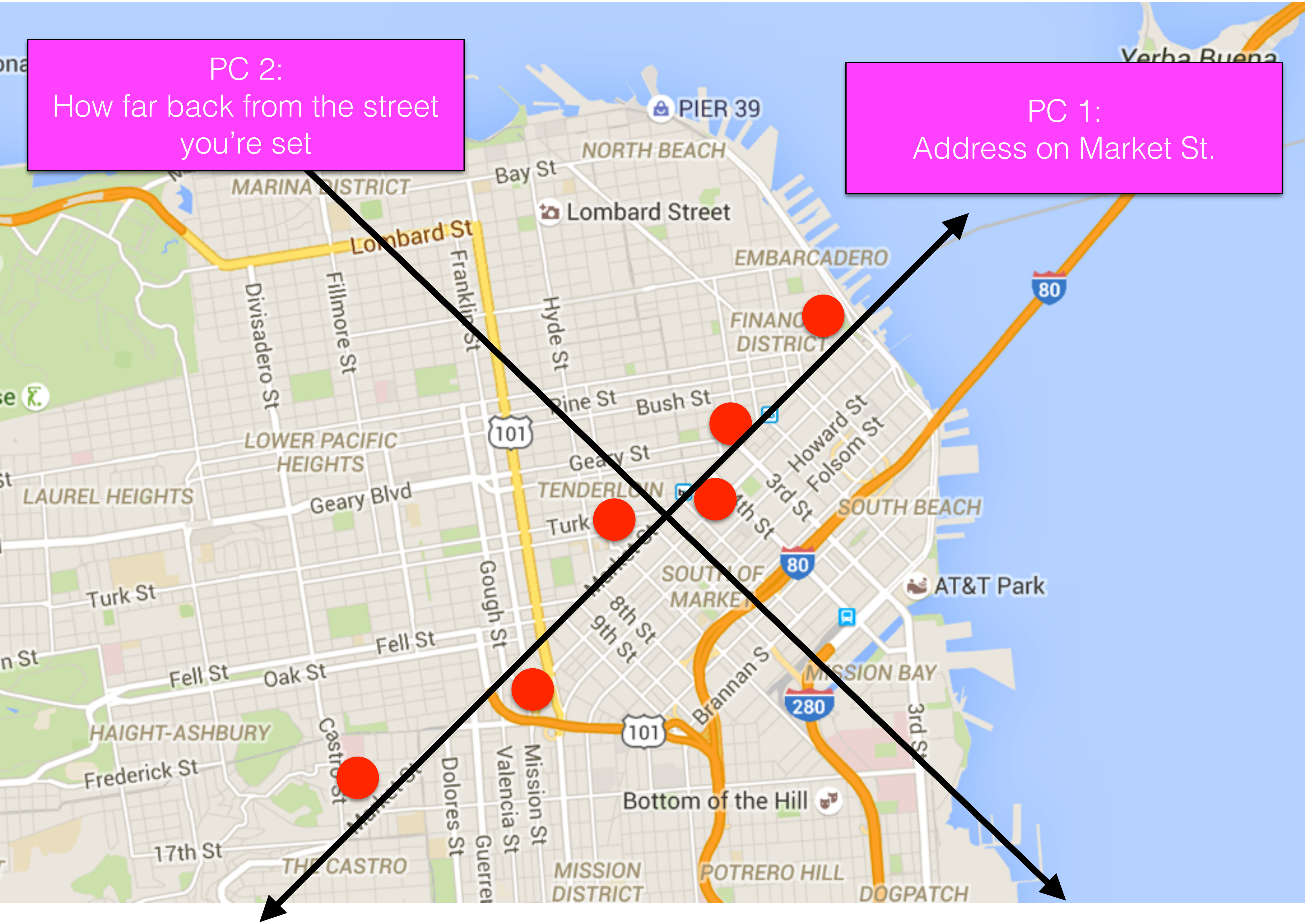
- Method for transforming a set of original (possible correlated) observations into new (uncorrelated) values.


PC 2:
How far back from the street you're set
PC 1:
Address on Market St.
PIER 39
NORTH BEACH
Bay St
MARINA DISTRICT
Lombard Street
Lombard St
EMBARCADERO
Divisadero St
Fillmore St
Franklin St
Hyde St
FINANCIAL DISTRICT
Pine St
Bush St
Howard St
Folsom St
LOWER PACIFIC HEIGHTS
Geary St
3rd St
LAUREL HEIGHTS
Geary Blvd
TENDERLOIN
SOUTH BEACH
Turk St
Gough St
SOUTH OF MARKET
AT&T Park
8th St
9th St
Fell St
Oak St
MISSION BAY
Brannan St
HAIGHT-ASHBURY
Frederick St
Castro St
Dolores St
Valencia St
Mission St
Bottom of the Hill
17th St
THE CASTRO
MISSION DISTRICT
POTRERO HILL
DOGPATCH
Yerba Buena
80
101
280

- Original values: latitude and longitude (very strong correlation for these data points)
- Transformed values: street address and distance from street (no correlation)

Dimensionality reduction

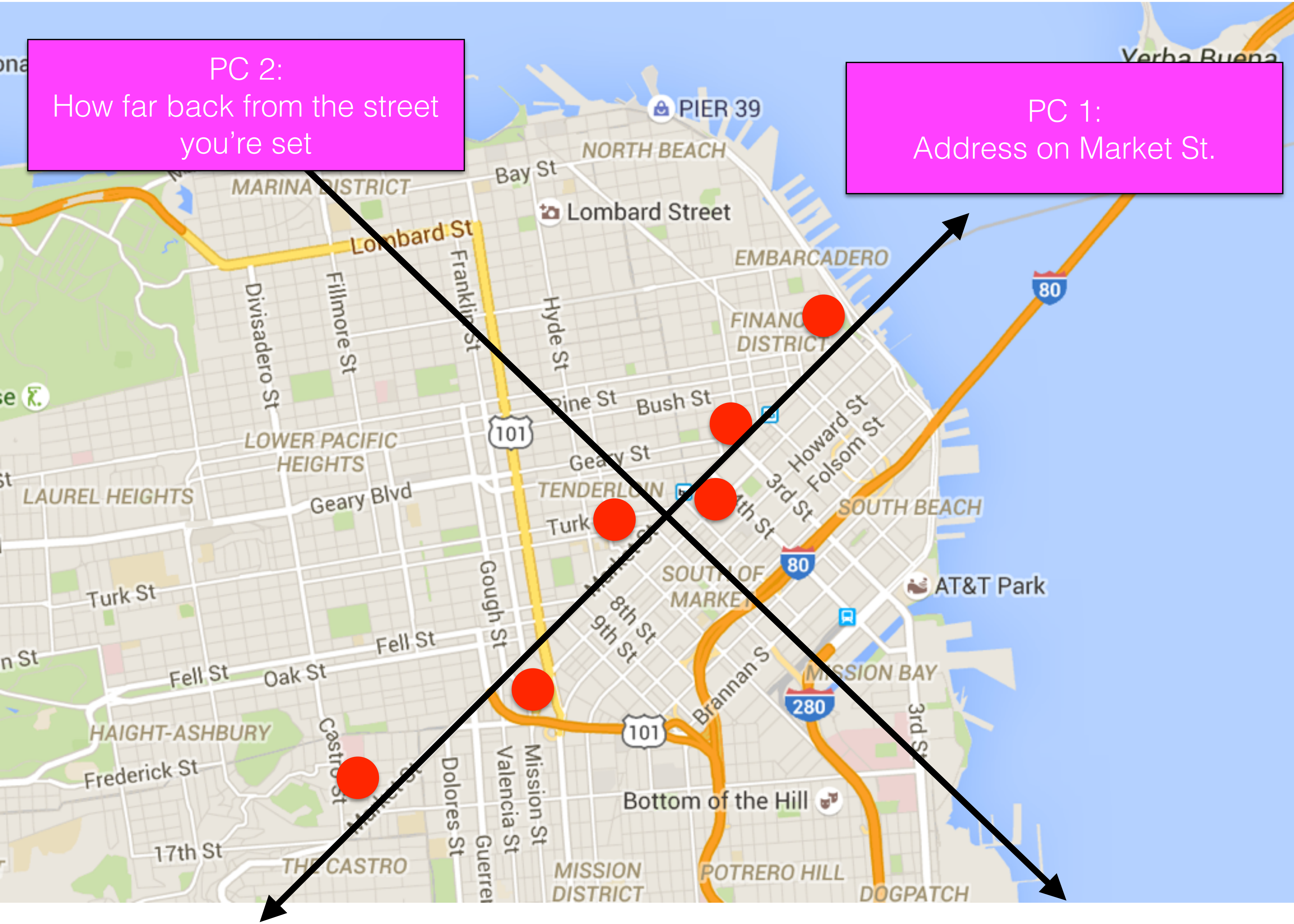
- Perhaps we only want to capture represent each data point with a single feature (or a smaller number of features n than the original representation)
- We can represent each point by the first n principal components


PC 2:
How far back from the street you're set
PC 1:
Address on Market St.
PIER 39
NORTH BEACH
Bay St
Lombard Street
Lombard St
EMBARCADERO
Divisadero St
Fillmore St
Hyde St
Pine St
Bush St
Howard St
Folsom St
3rd St
LOWER PACIFIC HEIGHTS
LAUREL HEIGHTS
Geary Blvd
TENDERLOIN
SOUTH BEACH
Turk St
Gough St
8th St
9th St
AT&T Park
Fell St
Oak St
MISSION BAY
HAIGHT-ASHBURY
Frederick St
Dolores St
Valencia St
Mission St
Bottom of the Hill
17th St
THE CASTRO
MISSION DISTRICT
POTRERO HILL
DOGPATCH
101
80
280

Why?

Main idea

- Each principal component (1 … F) is the axis that exhibits them most variance in the data and is uncorrelated (orthogonal) with earlier PCs
- The first PC explains the most variance; the second PC explains the most remaining variance, etc.


PC 2:
How far back from the street you're set
PC 1:
Address on Market St.
PIER 39
NORTH BEACH
Bay St
Lombard Street
MARINA DISTRICT
Lombard St
EMBARCADERO
Divisadero St
Fillmore St
Franklin St
Hyde St
FINANCIAL DISTRICT
Pine St
Bush St
101
LOWER PACIFIC HEIGHTS
Geary St
Howard St
Folsom St
3rd St
LAUREL HEIGHTS
Geary Blvd
TENDERLOIN
SOUTH BEACH
Turk
80
Turk St
Gough St
SOUTH OF MARKET
AT&T Park
8th St
9th St
Fell St
Fell St
Oak St
Brannan St
MISSION BAY
280
HAIGHT-ASHBURY
Castro St
Frederick St
Dolores St
Valencia St
Mission St
101
3rd St
Bottom of the Hill
17th St
THE CASTRO
MISSION DISTRICT
POTRERO HILL
DOGPATCH

Variance

$$var(X) = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(X_i - \bar{X})}{n-1}$$

Covariance

$$cov(X, Y) = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

$$var(X) = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(X_i - \bar{X})}{n-1}$$

$$cov(X, Y) = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

$$var(X) = cov(X, X)$$

$$cov(X, Y) = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

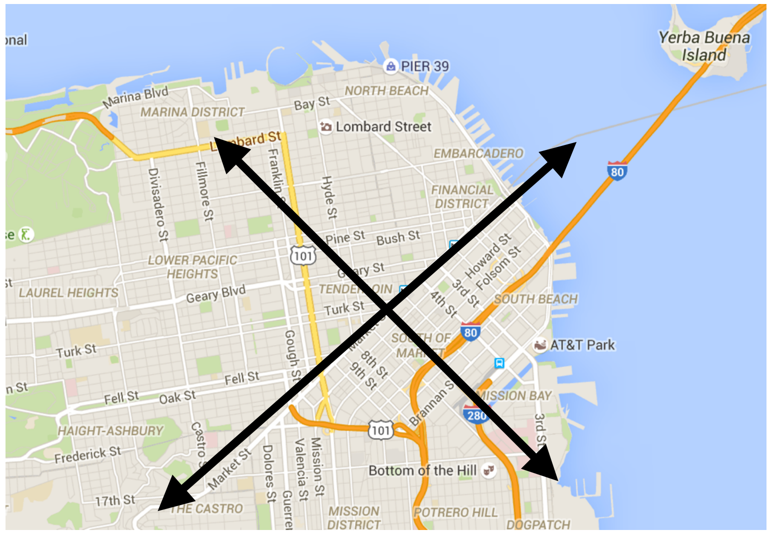
X-mean(X)	Y1-mean(Y1)	Y2-mean(Y2)	Y3-mean(Y3)	Y4-mean(Y4)
4	4.33	-4	0	2.67
0	0.33	0	0	-5.33
-4	-4.67	4	0	2.67
Cov(X,Y)	18	-16	0	0

bit.ly/1LcnZH6

[http://winedarksea.org/wp-content/uploads/2015/07/mydata.csv]

OLS vs. PCA

Vectors



- Each of these axes is a vector
- It doesn't matter how long it is — it still points in the same direction

Vectors

- Each of these axes is a vector

- It doesn't matter how long it is — it still points in the same direction

Eigenvectors

3	1
1	3

matrix A

- Eigenvectors are those axes
- A square matrix of n dimension has n eigenvectors
- Each eigenvector is uncorrelated with the others (orthonormal)

PCA

3	1
1	3

- By finding the eigenvectors of the covariance matrix, we are finding the dimensions along which the greatest variance of the data is explained

PCA

1. Center the data (subtract the mean from each variable)

2. Calculate the covariance matrix from that centered data

3. Find the eigenvectors for that covariance matrix

PCA

	PC1	PC2	PC3	PC4	PC5
1					
2					
3					
4					
5					

what you get out of PCA are a set of principle components with which you can transform your original data

PCA

data point

1
2
3
4
5

X

	PC1	PC2	PC3	PC4	PC5
1					
2					
3					
4					
5					

original feature dimension

= original data point in transformed space

PCA

data point

1
2
3
4
5

X

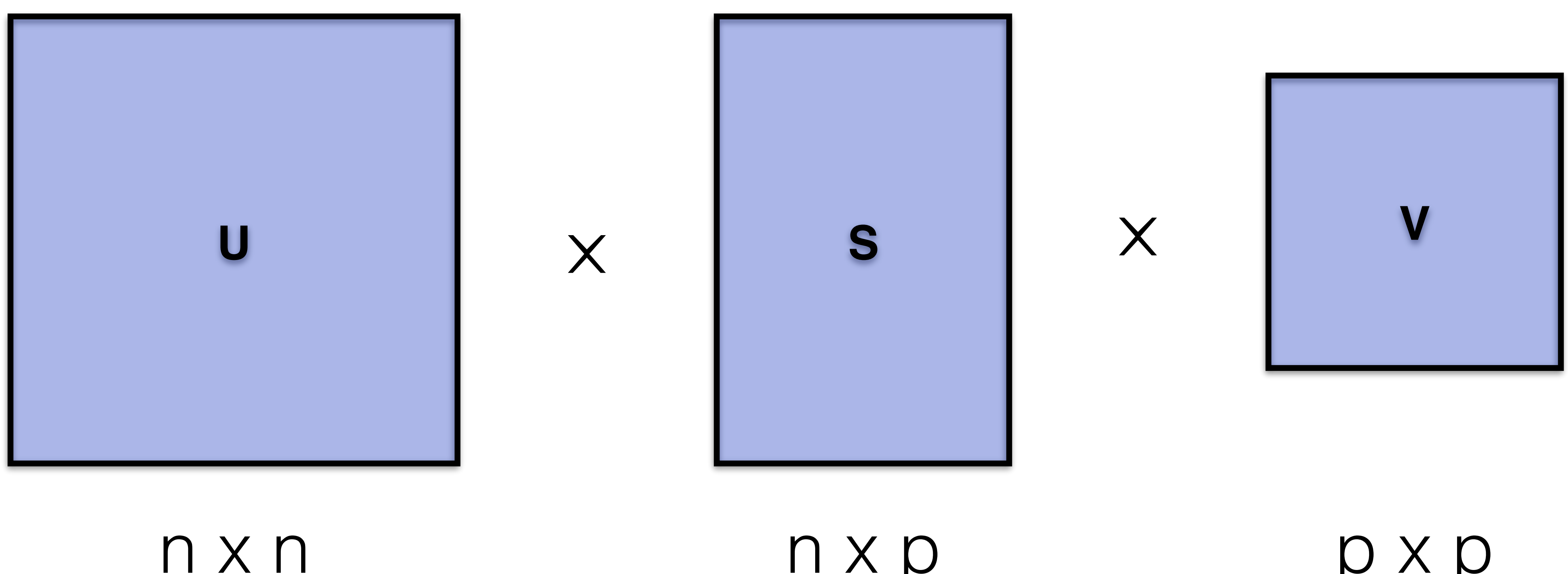
	PC1 PC2	PC3 PC4 PC5
1		
2		
3		
4		
5		

original feature dimension

= data point with fewer dimensions

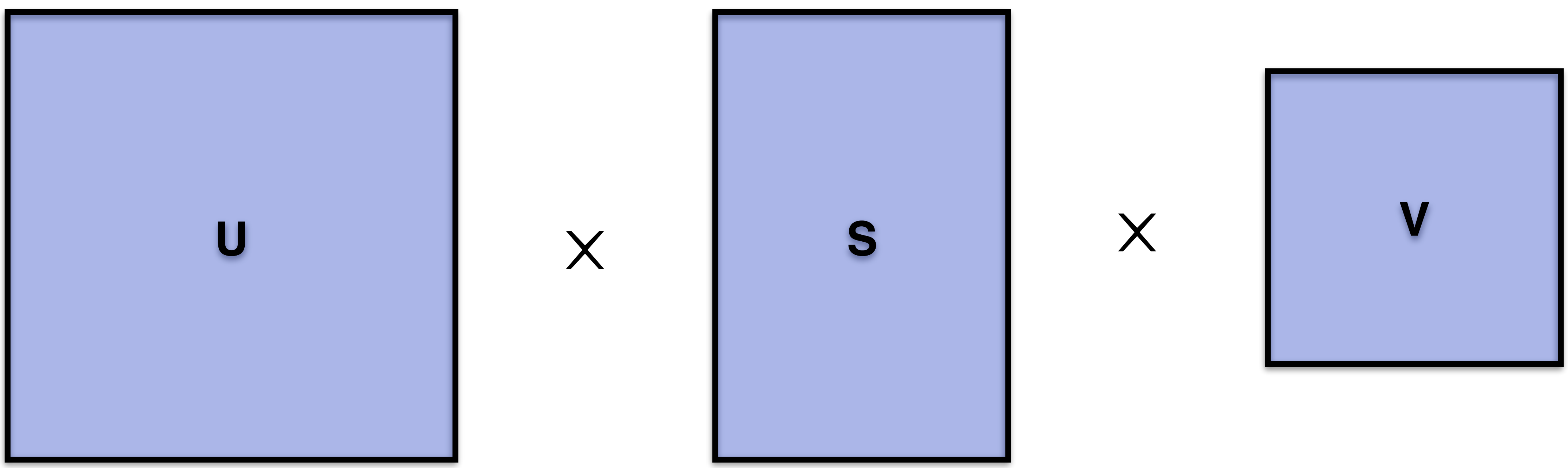
Singular value decomposition

- Any nxp matrix X can be decomposed as the product of three matrices:



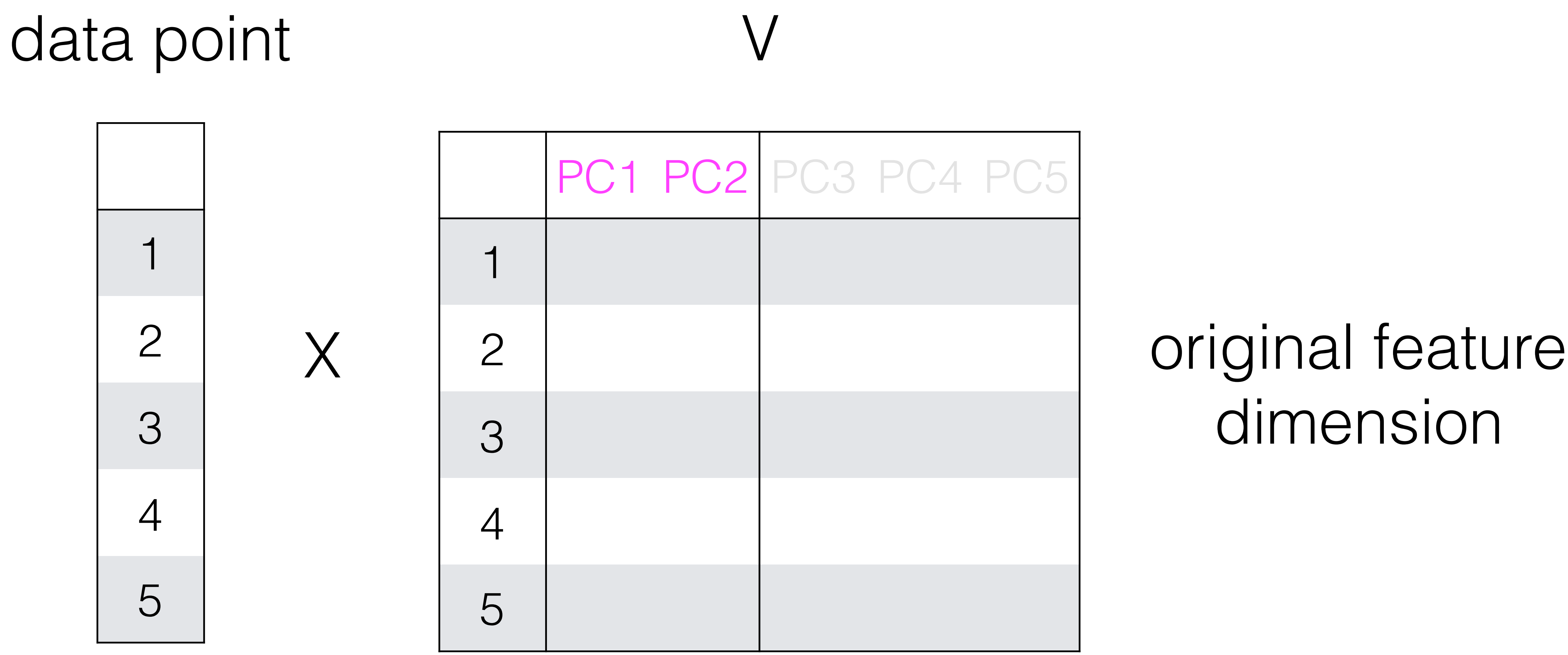
SVD

- Any nxp matrix X can be decomposed as the product of three matrices:



V = the eigenvectors of the original covariance matrix

SVD for PCA



= data point with fewer dimensions

PCA vs SVD

- Calculating the eigenvectors of the covariance matrix requires forming $X^TX = \mathbb{R}^{p \times p}$. (not feasible for lots of features)

- Lots of fast ways of solving SVD

	feat 1	feat 2	feat 3	feat 4	feat 5
item 1					
item 2					
item 3					
item 4					
item 5					

item-feature matrix

Latent semantic analysis

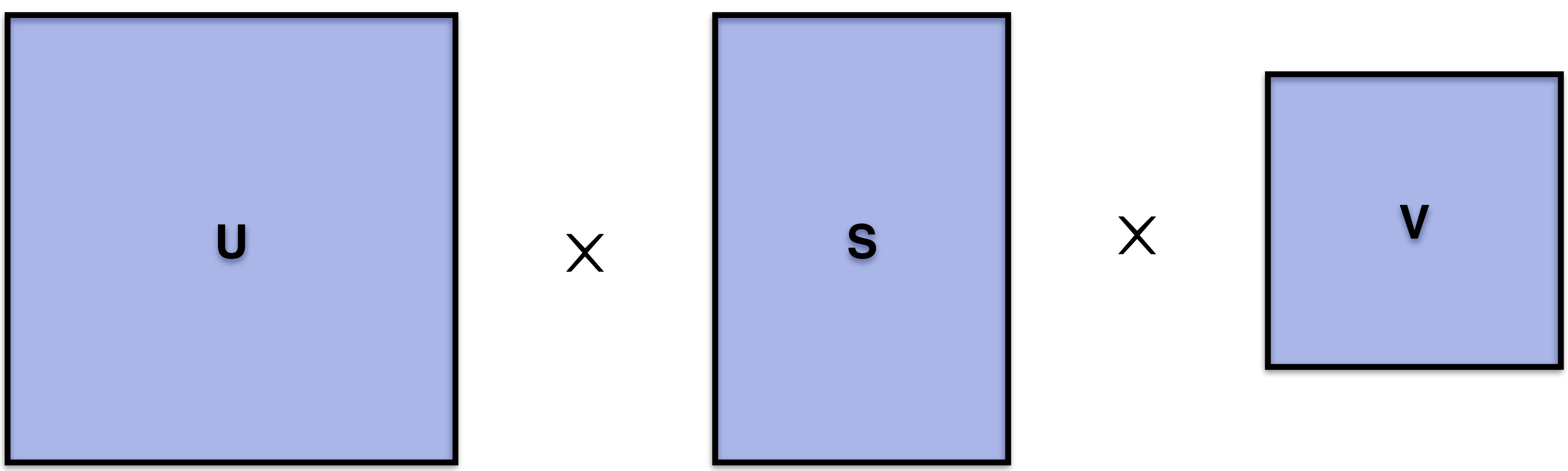
	doc 1	doc 2	doc 3	doc 4	doc 5
the					
dog					
ice					
eats					
cream					

term-document matrix
(typically weighted by tf-idf)

"documents"

"words"

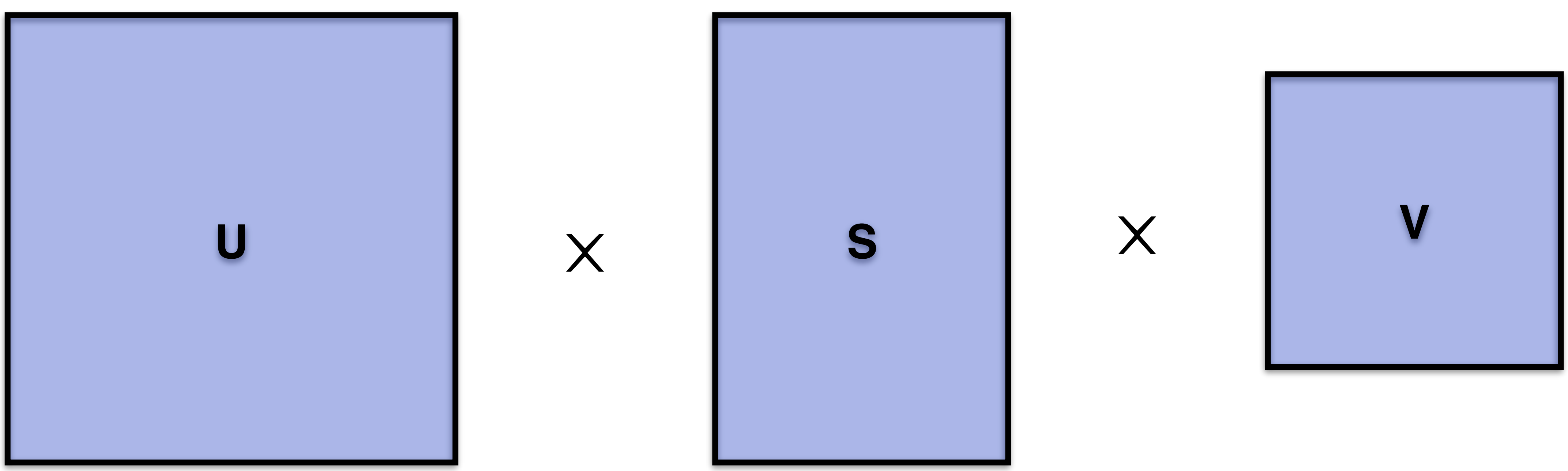
	doc 1	doc 2	doc 3	doc 4	doc 5
the					
dog					
ice					
eats					
cream					



Latent semantic analysis

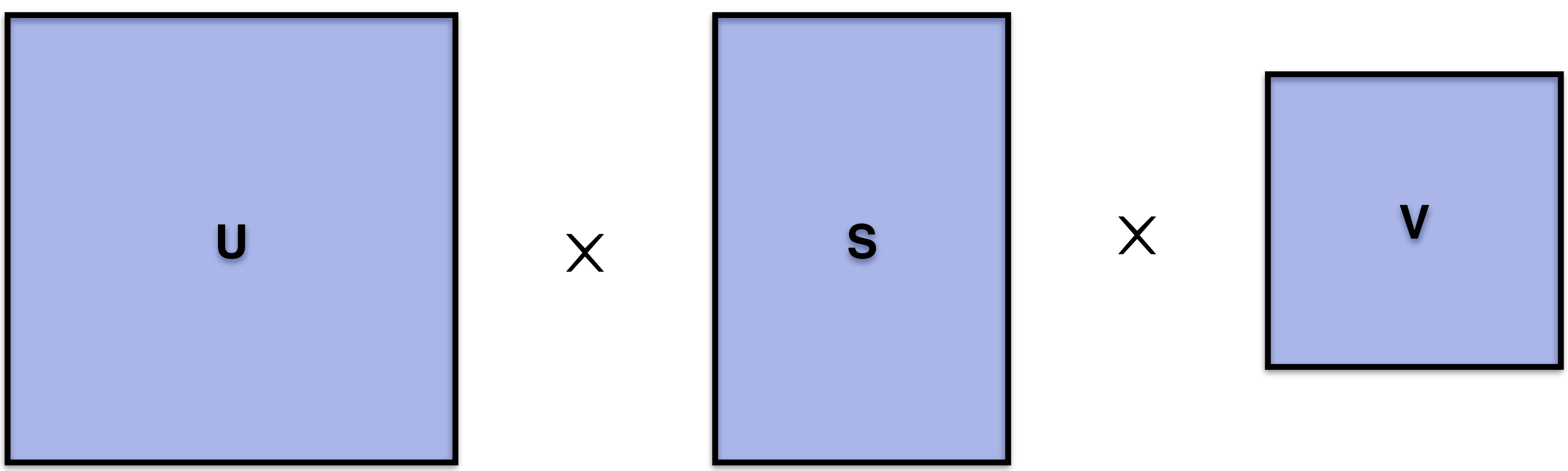
V = the eigenvectors of the document x word covariance matrix

V = the eigenvectors of the word x document covariance matrix



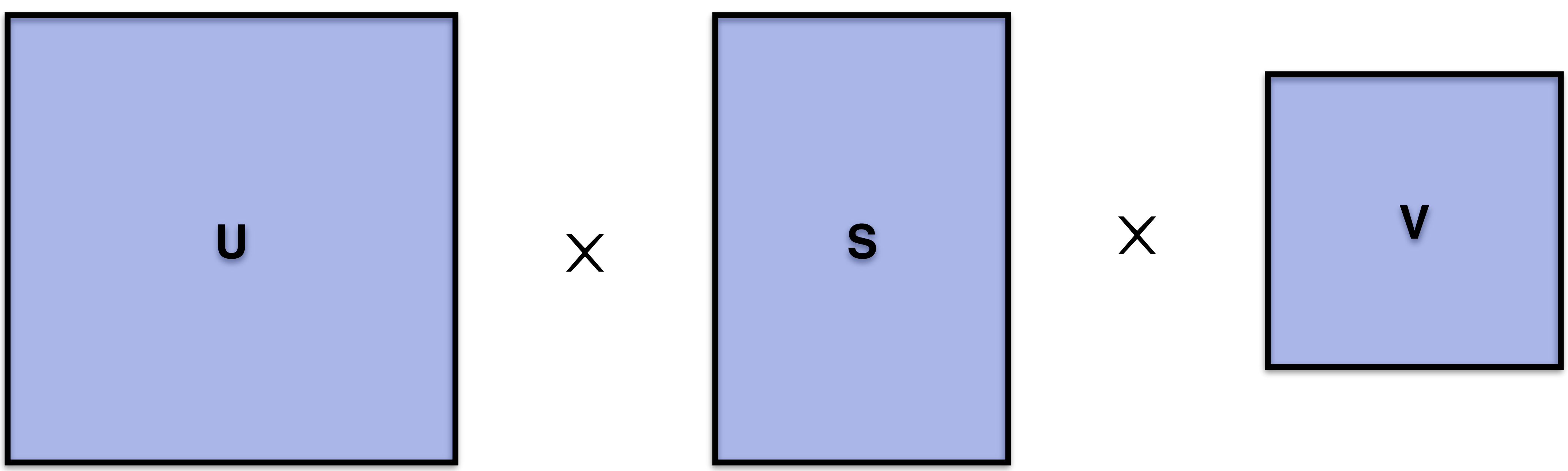
Latent semantic analysis

If we wanted to transform the original matrix into a new, low-dimensional space, we could simply multiply it by V (as before)



Latent semantic analysis

But the individual matrices themselves also give us a low-dimensional representation of the features (and documents)



Word2Vec

“context”

“word”

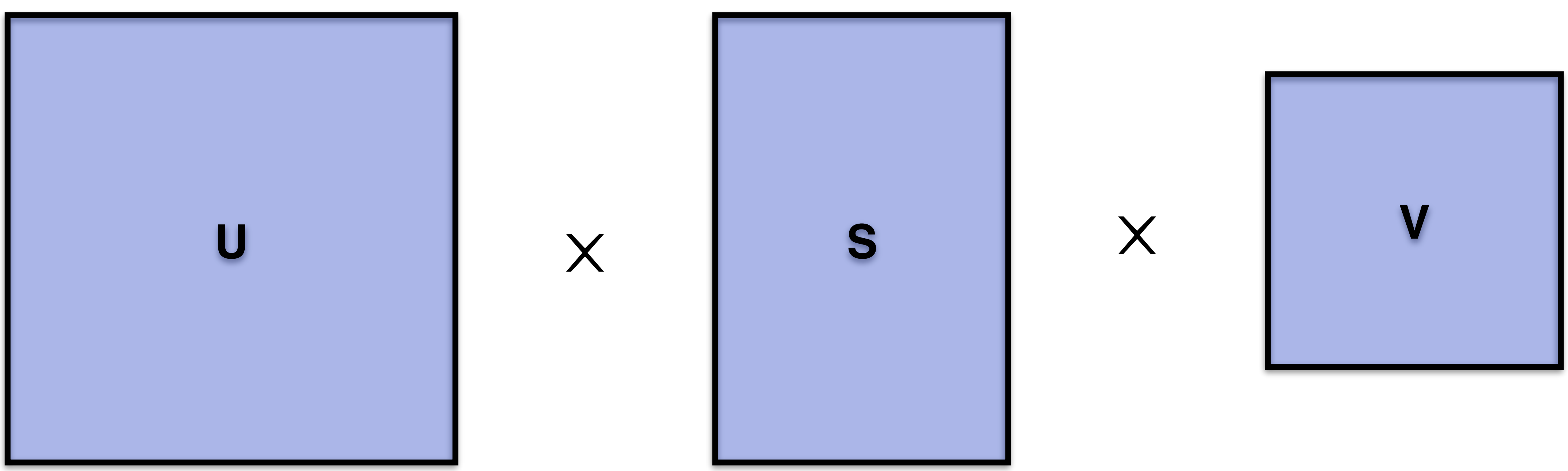
	dog	cat	hot	ice	summer
the					
dog					
ice					
eats					
cream					

word-context matrix
(weighted by pointwise mutual information)

“context”

“word”

	dog	cat	hot	ice	summer
the					
dog					
ice					
eats					
cream					



Levy and Goldberg (2014)

http://mybinder.org/repo/dbamman/dds