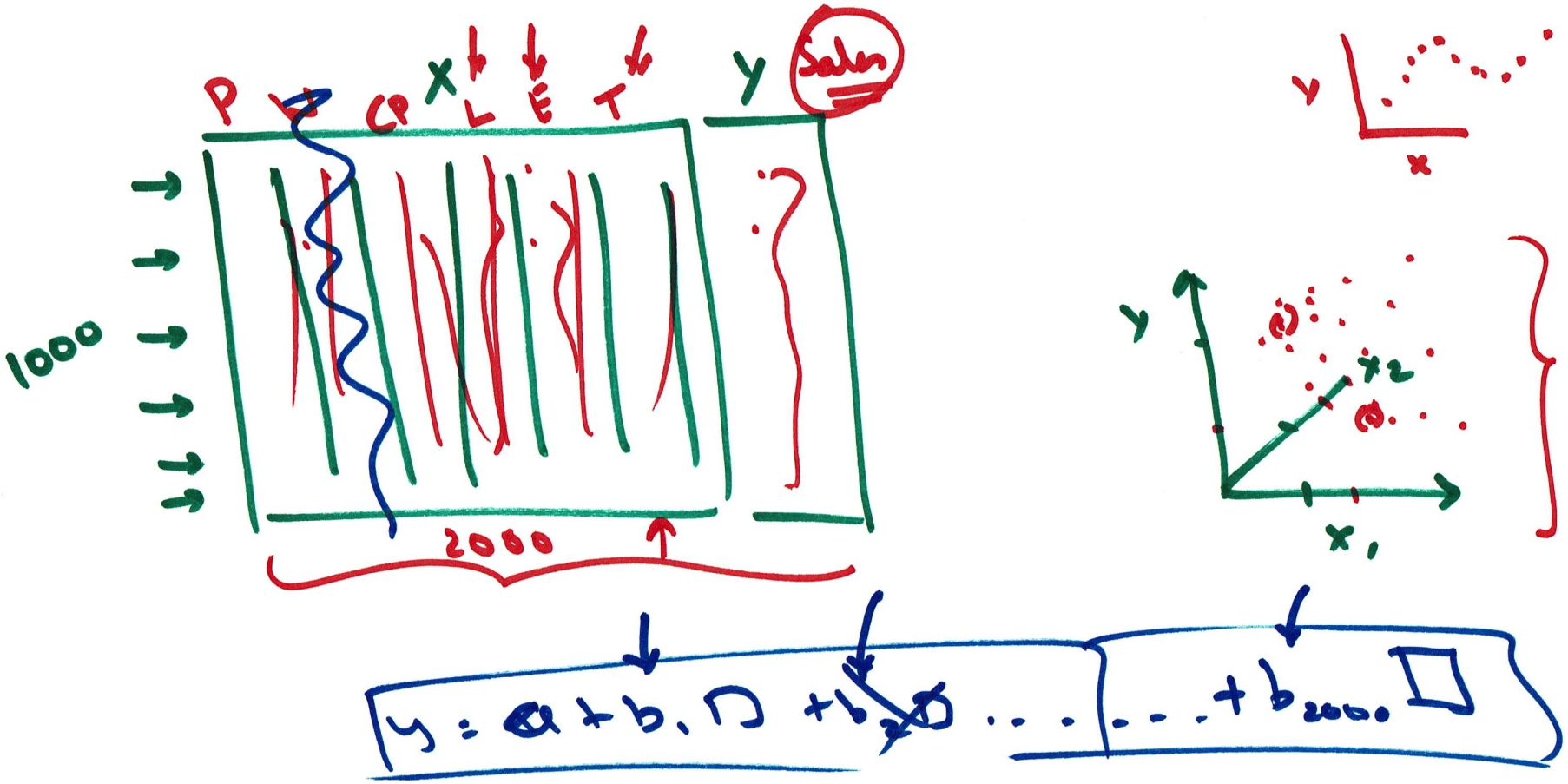
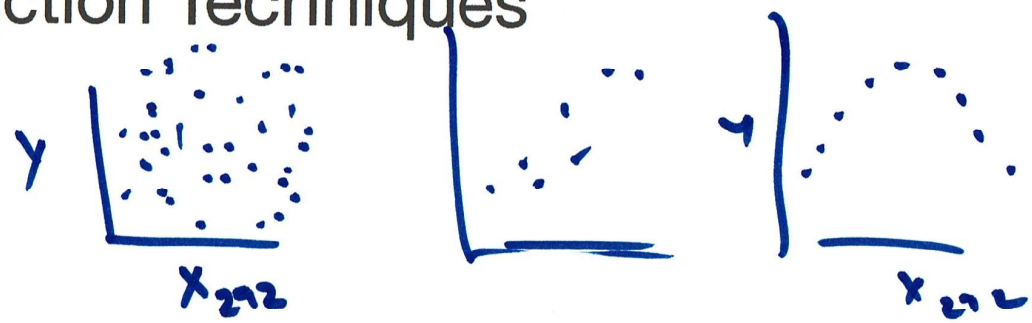
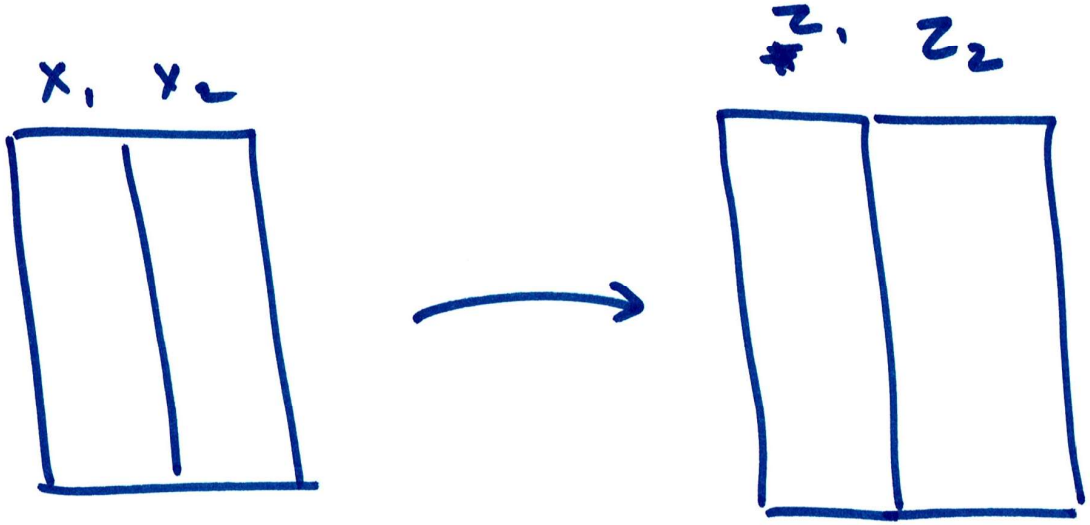
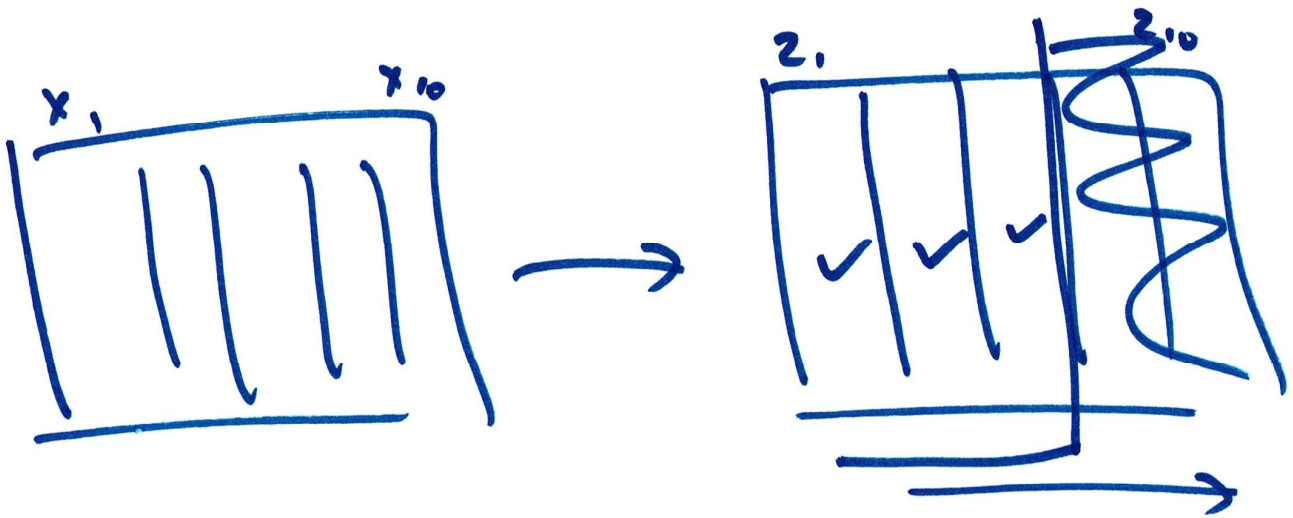

P
CP
X
L
E
T
Y
Salen
1000
2000
y
x
y
x1
x2
y = a + b1 □ + b2 ○ . . . . . + b2000 □

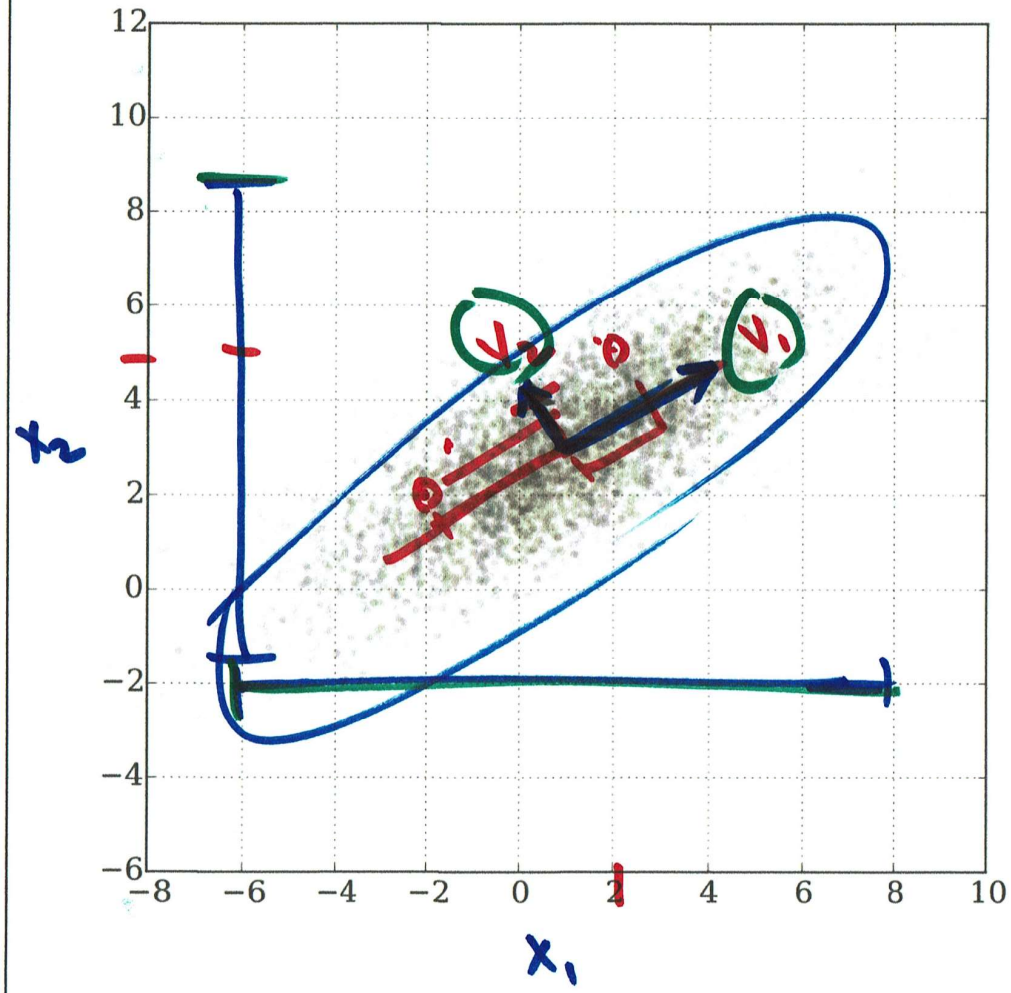
Dim. Reduction Techniques

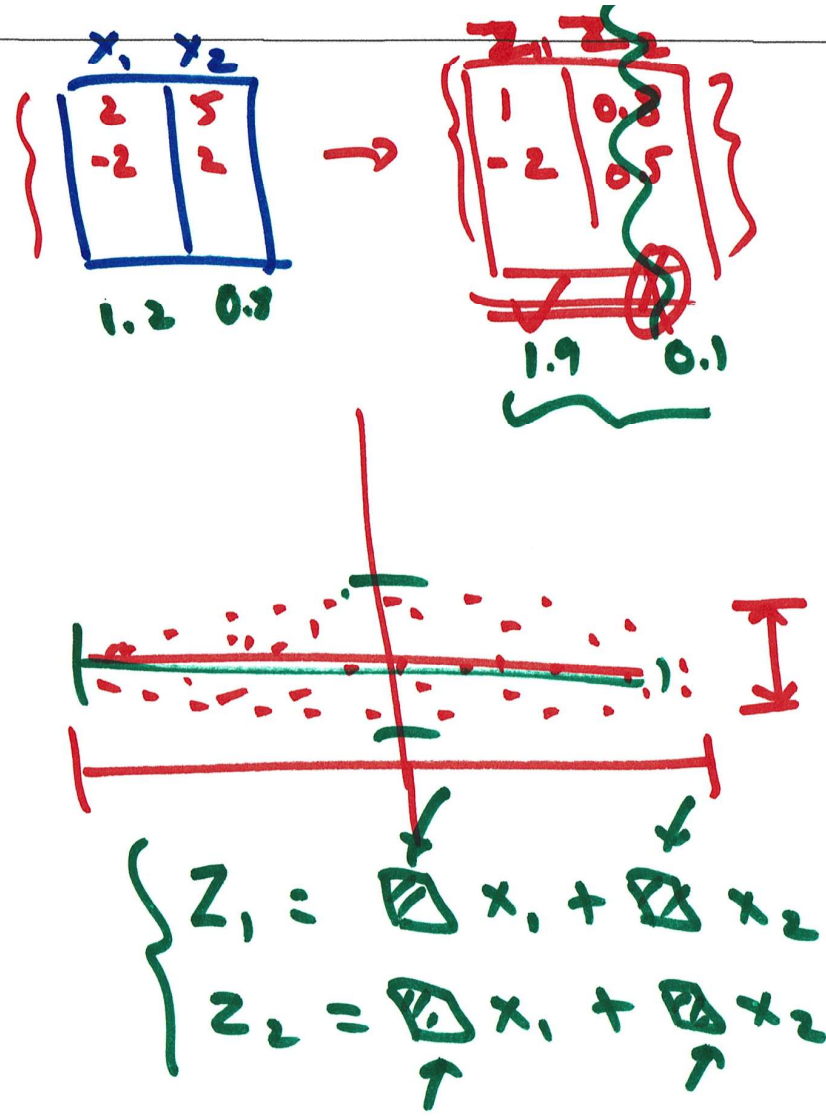


- Feature elimination
  - Simply identify and remove variables (columns) that are not important
  - The disadvantage is that we would gain no insight from those dropped variables and loose any information they contain
- Feature extraction
  - Create a few new variables from the old variables
  - **PCA** Principal Component Analysis: is the most popular feature extraction technique (linear)
  - t-SNE (non-linear)


x_1 x_2
z_1 z_2

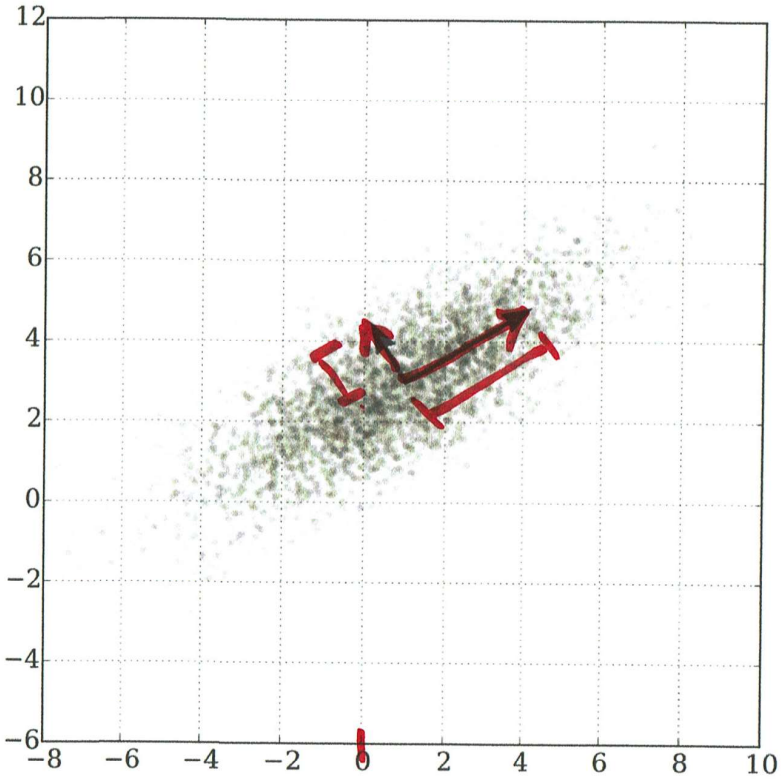

x_1
x_{10}
z_1
z_{10}

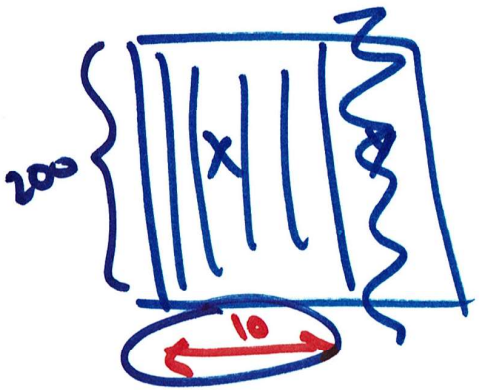

12
10
8
6
4
2
0
−2
−4
−6
−8
−6
−4
−2
0
2
4
6
8
10
x_2
x_1
v_2
v_1


x_1 x_2
2 5
-2 2
1.2 0.8
1 0.5
-2 0.5
1.9 0.1
$z_1 = \square x_1 + \square x_2$
$z_2 = \square x_1 + \square x_2$

PCA

- creates new variables using linear combinations of old variables
- is designed to create variables that are independent of one another
- also manages to tell us how important each of these new variables are
- this "importance", helps us to choose how many variables we will use





$$X_{new} = \frac{X - \mu}{\sigma}$$

$$Cov = Cov(X_{new})$$

Eigen decomposition

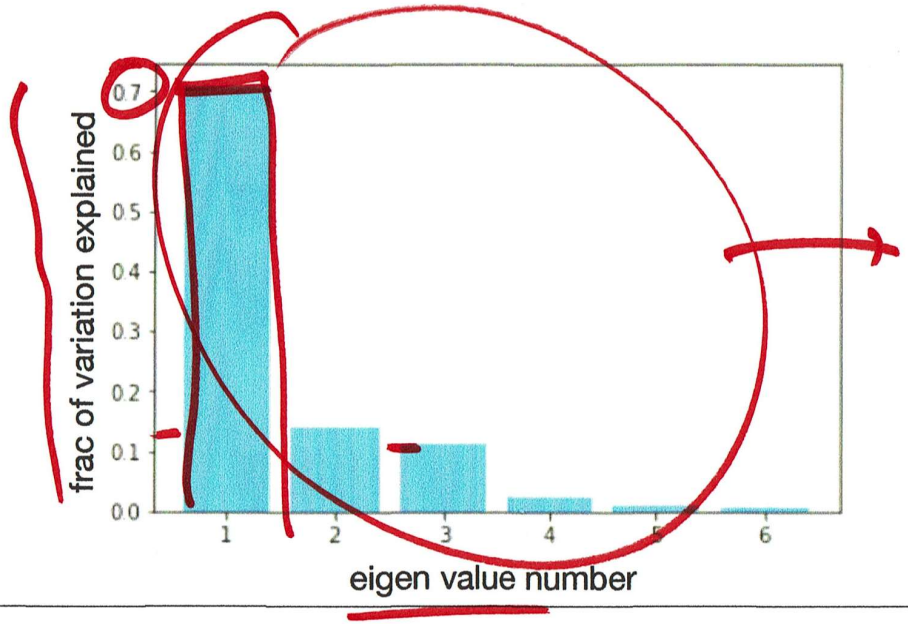
$e_1, e_2, \ldots e_{10}$

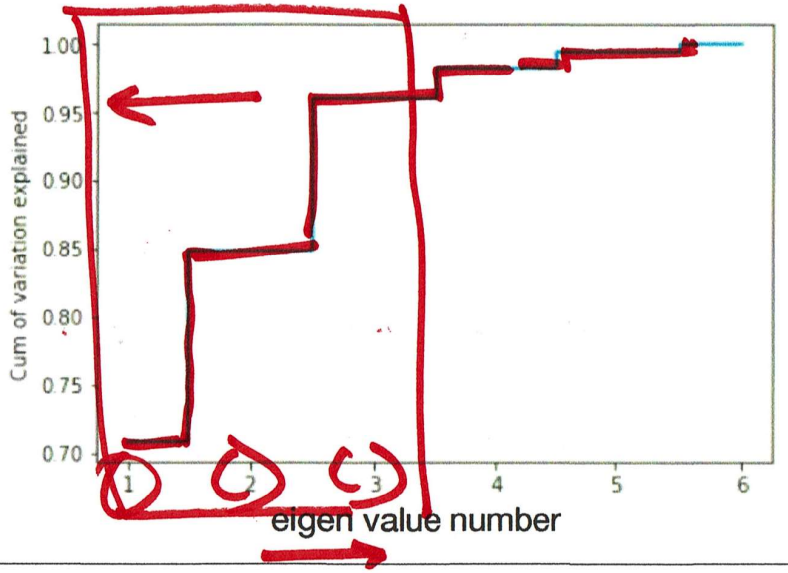
$v_1, v_2 \ldots v_{10}$

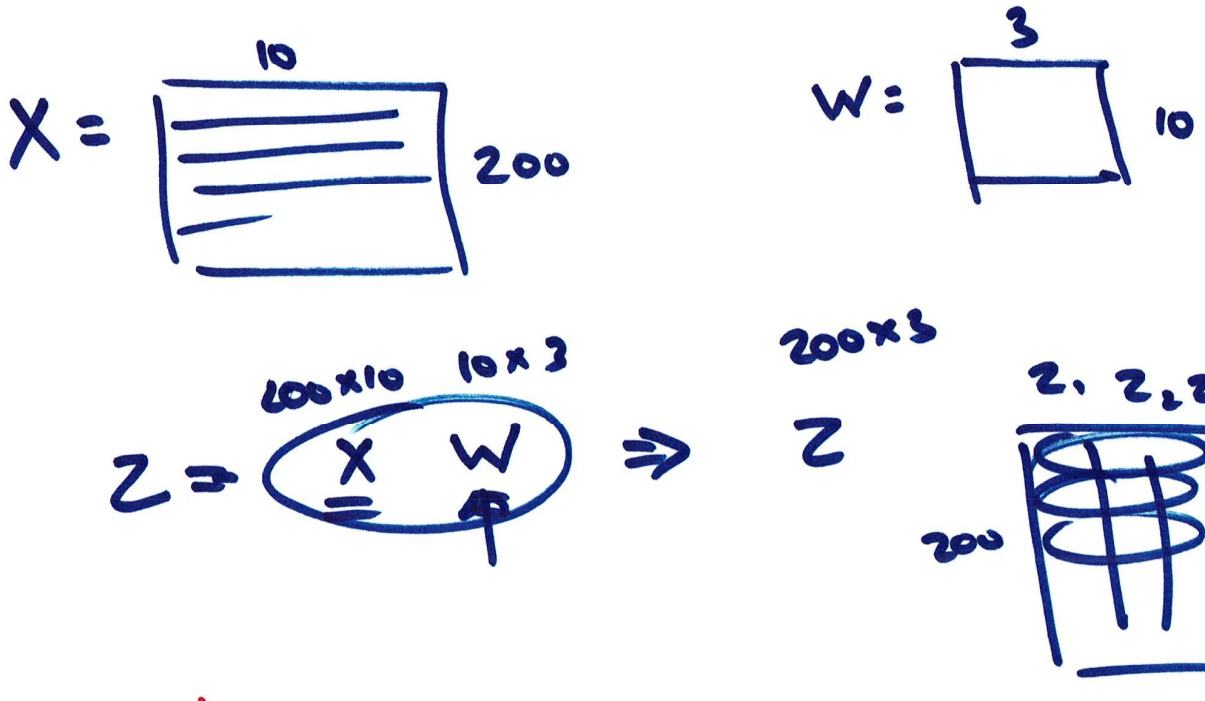
$$W = \begin{pmatrix} v_1 & v_2 & v_3 \end{pmatrix} \quad 10 \times 3$$

$$Z = \quad 10 \times 3$$

- Scale the data and compute the covariance matrix
- Break the covariance matrix into magnitude and direction. Eigen Vectors and the Eigen Values of the covariance matrix can be thought of as the natural axis/directions and magnitudes along those axis, of the data
    - The eigen values also can be used to calculate the percentage of variation explained by each component
- Sort in the eigen values in desending order and calculate the cumulative percentage of variation explained
- Pick the number of principal components you will use
- Transform to new variables






X =
10
200
W =
3
10
Z =
200x10
10x3
X
W
⇒
Z
200x3
z1 z2 z3
200
S = 100 + 20 z1
+ 15 z2
- 10 z3
z1 = □ x1 + □ x2 + □ x3
... x10
z2 = ...
z3 = ...