APS1070

Foundations of Data Analytics and Machine Learning

Summer 2020

Wed June 24 / Week 7:

- *Principle Component Analysis (PCA)*

Jason Riordon, PhD

News

- Midterm quiz in progress
- Project 3 Tutorial tomorrow, Q&A session next week
- No class next week – Canada Day!

Slide Attribution

These slides contain materials from various sources. Special thanks to Scott Sanner, Marc Deisenroth and Josh Starmer.

Principle Component Analysis (PCA)

High-Dimensional Data

- Real world data is often high-dimensional!
- Challenge: difficult to **analyze**, **visualize** and **interpret**

Properties of High-Dimensional Data

- Many dimensions are unnecessary
- Data often lives on a low-dimensional manifold
- Dimensionality reduction finds the relevant dimensions

These slides are adapted from

StatQuest

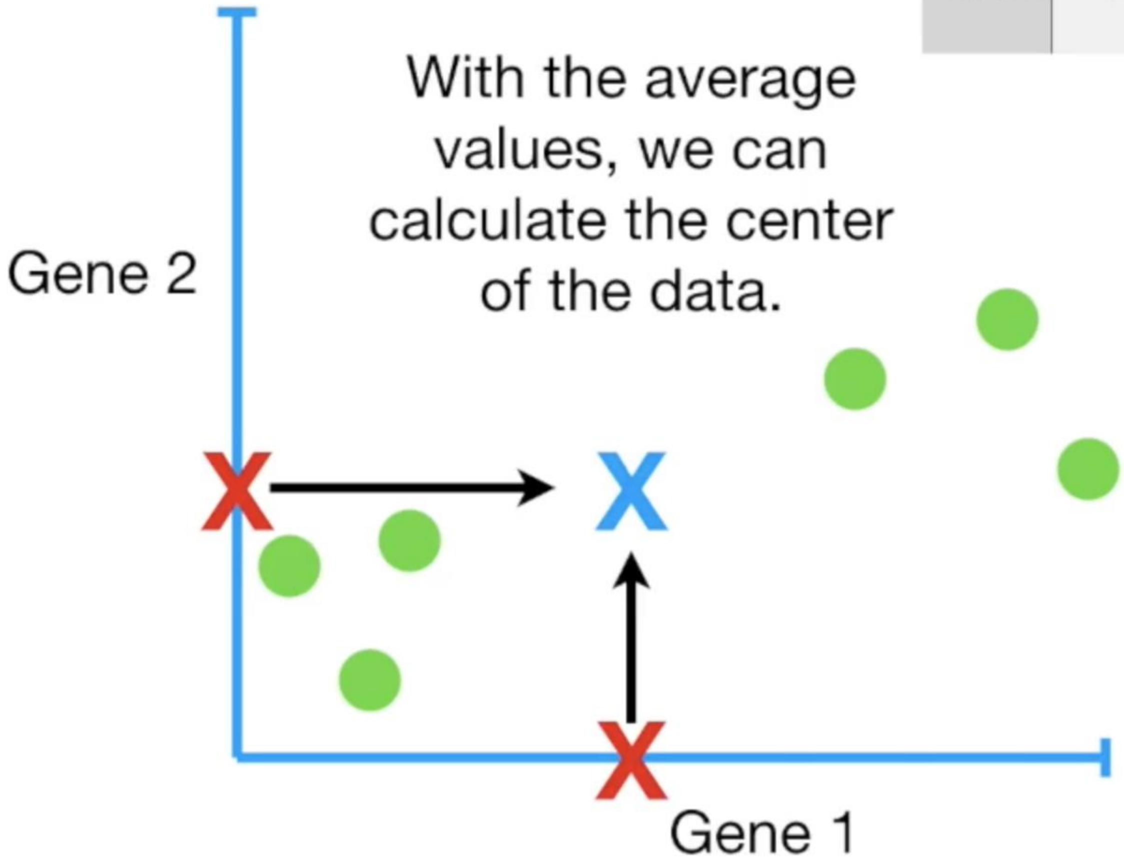
https://www.youtube.com/watch?v=FgakZw6K1QQ

Data

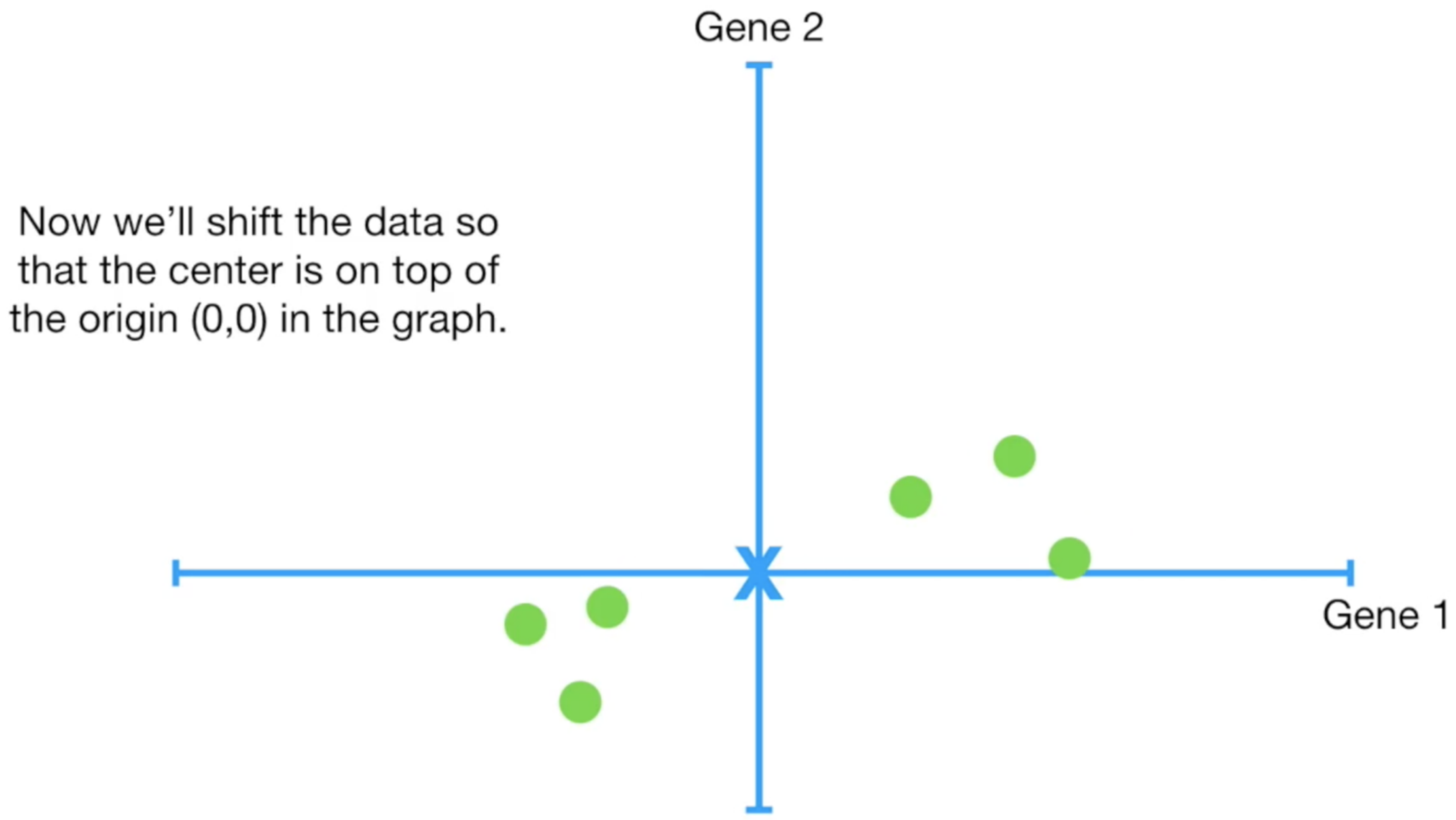
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1

Average

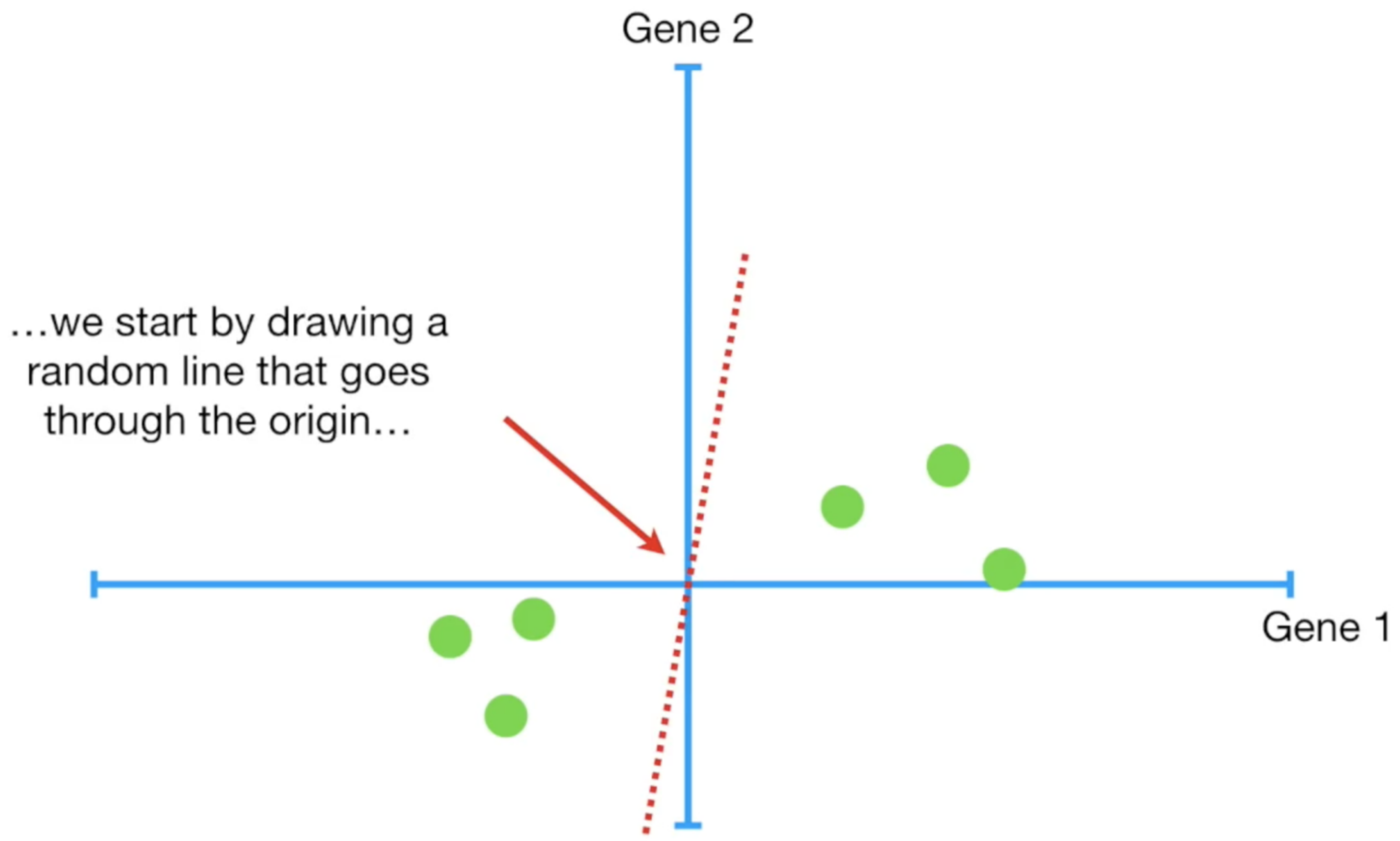
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1



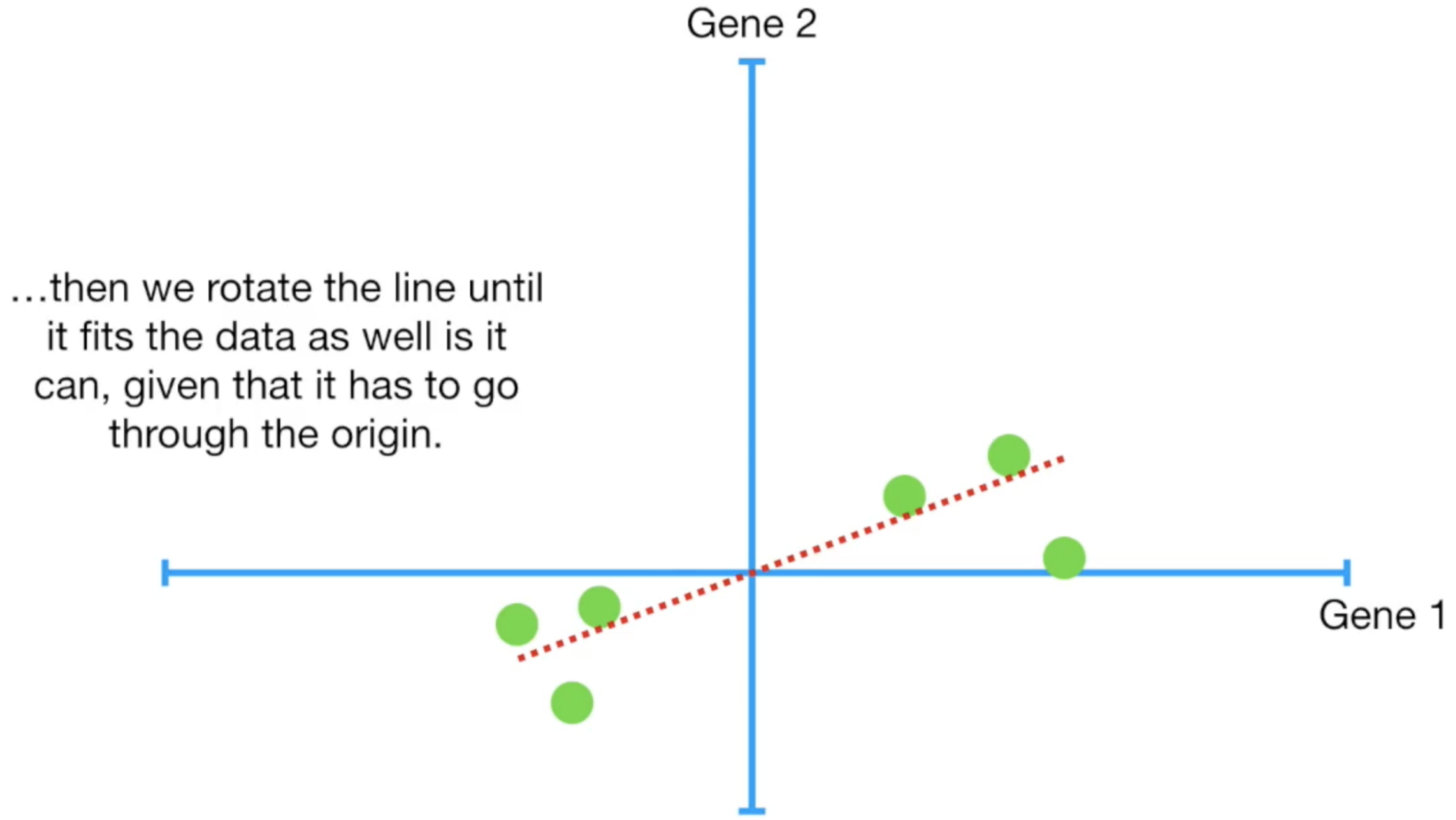
Shift the data



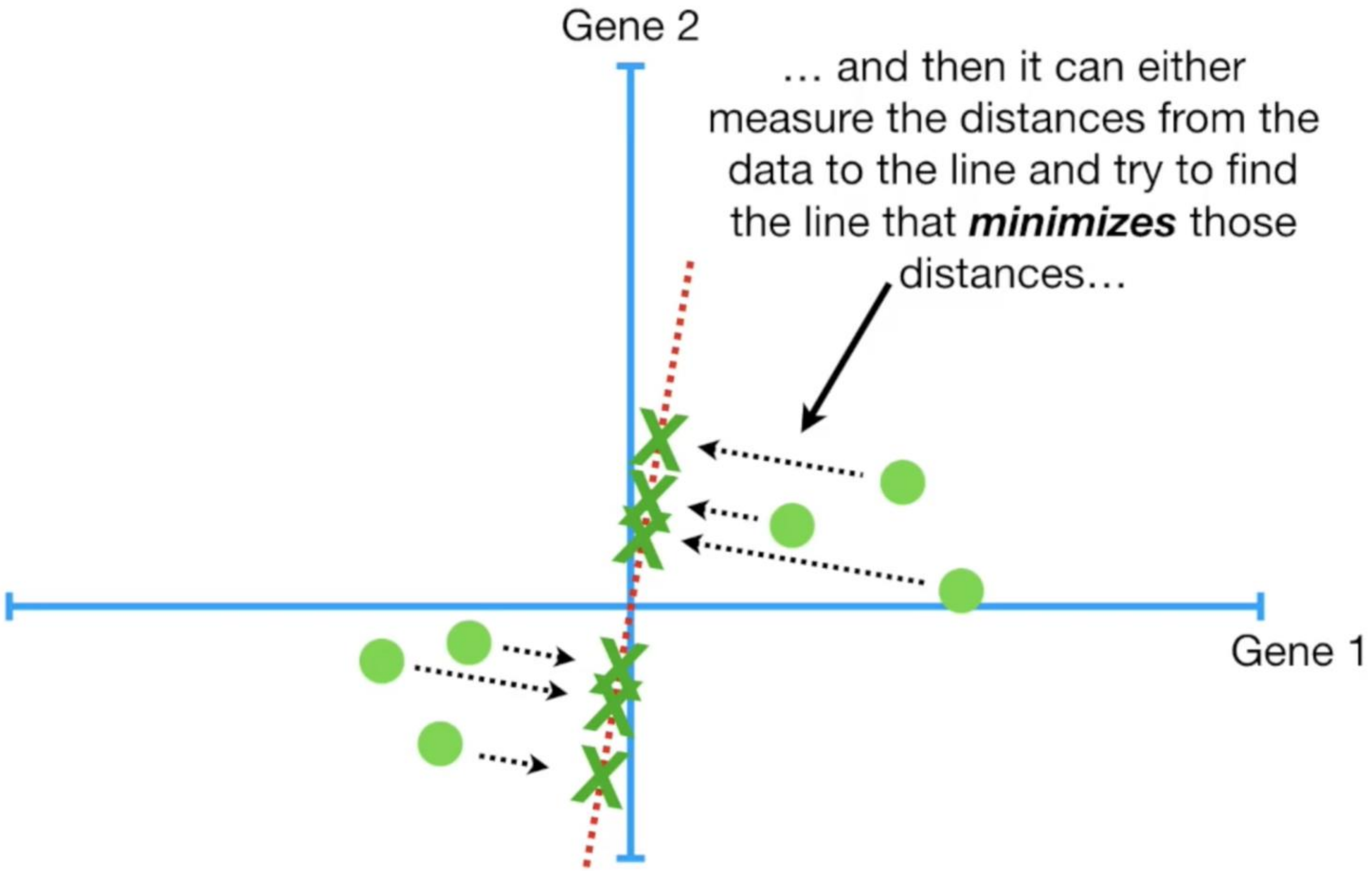
Draw a line crossing the origin (any orientation)



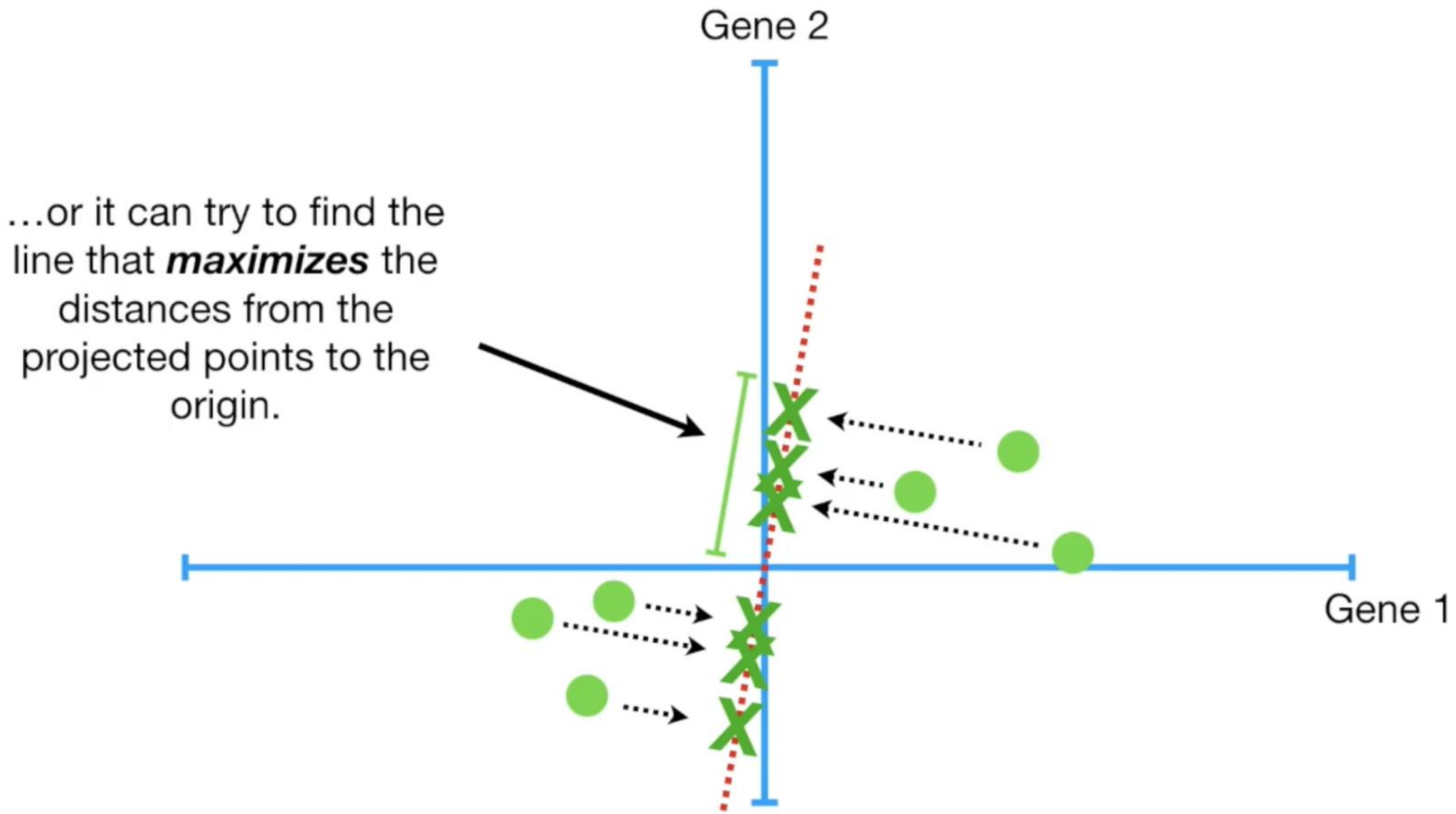
Target: Fit the data as best as a line can



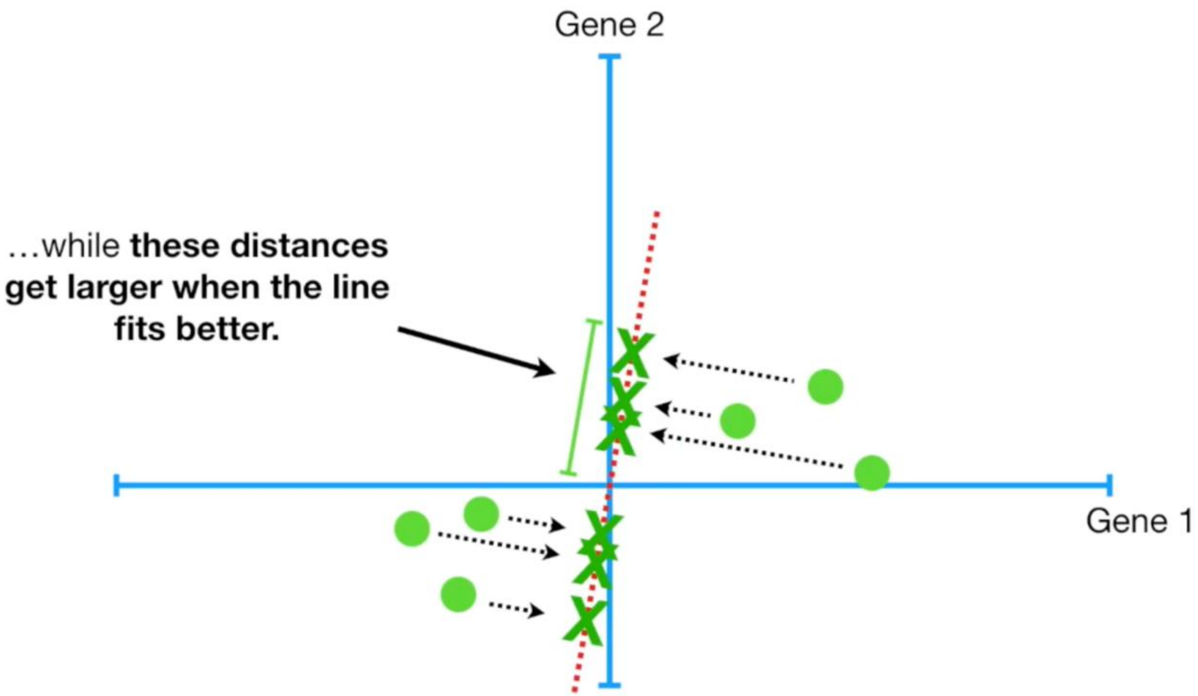
How do we know we have the best fit?

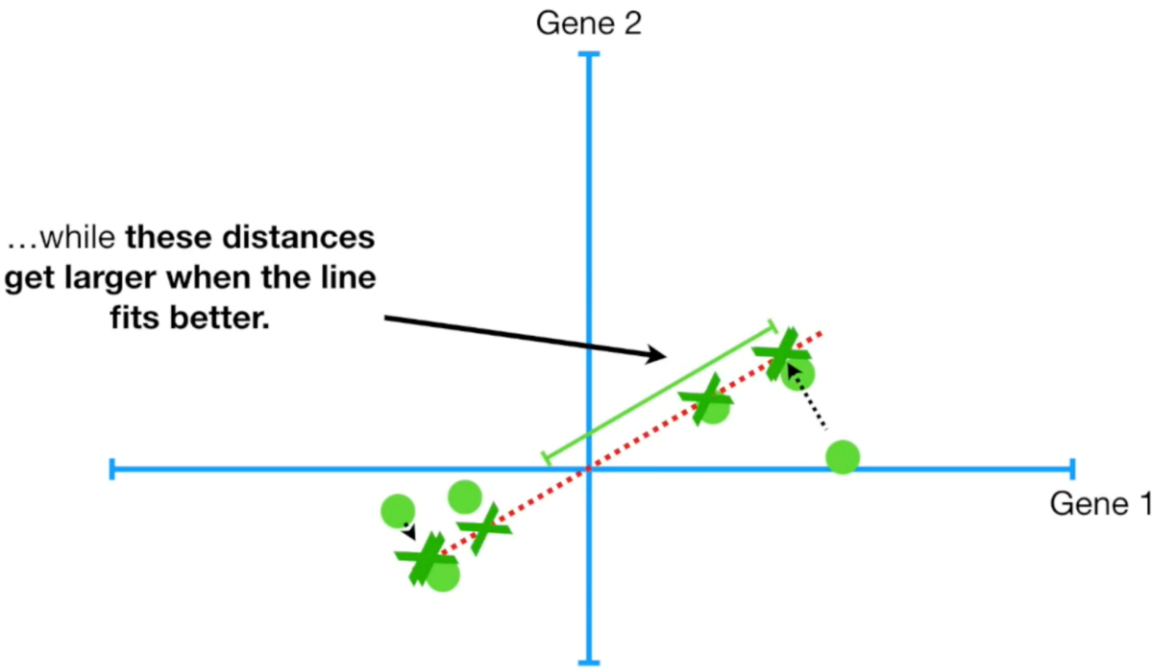


How we know the best fit?

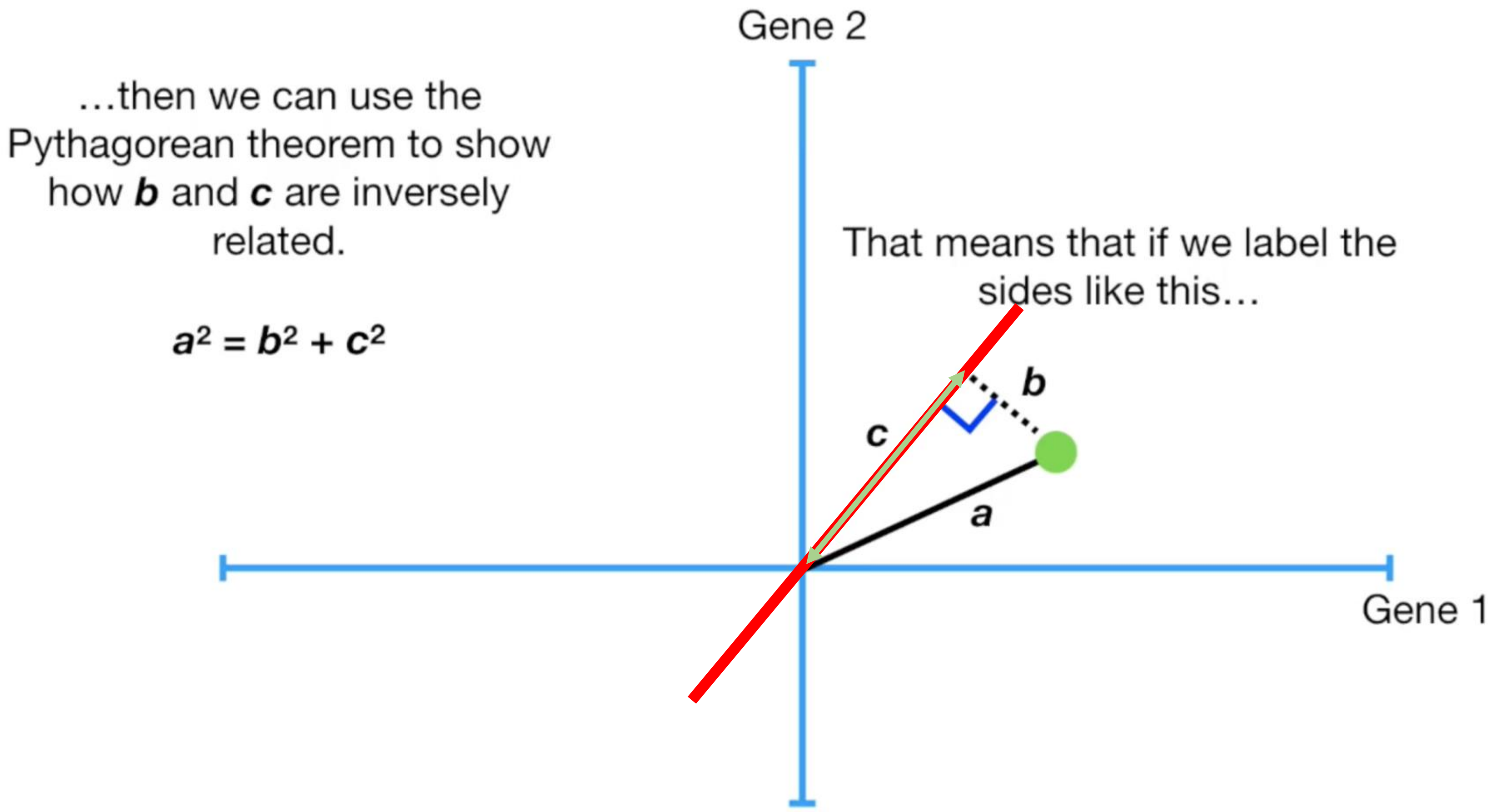


Intuition


Gene 2
...while these distances get larger when the line fits better.
Gene 1


Gene 2
...while these distances get larger when the line fits better.
Gene 1

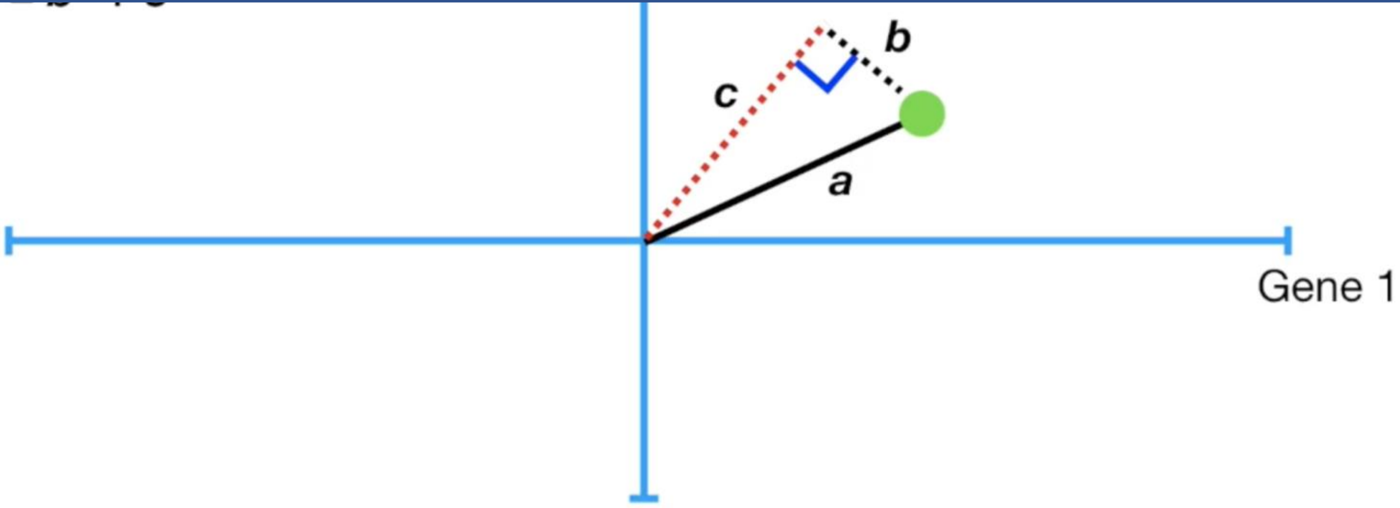
Why?



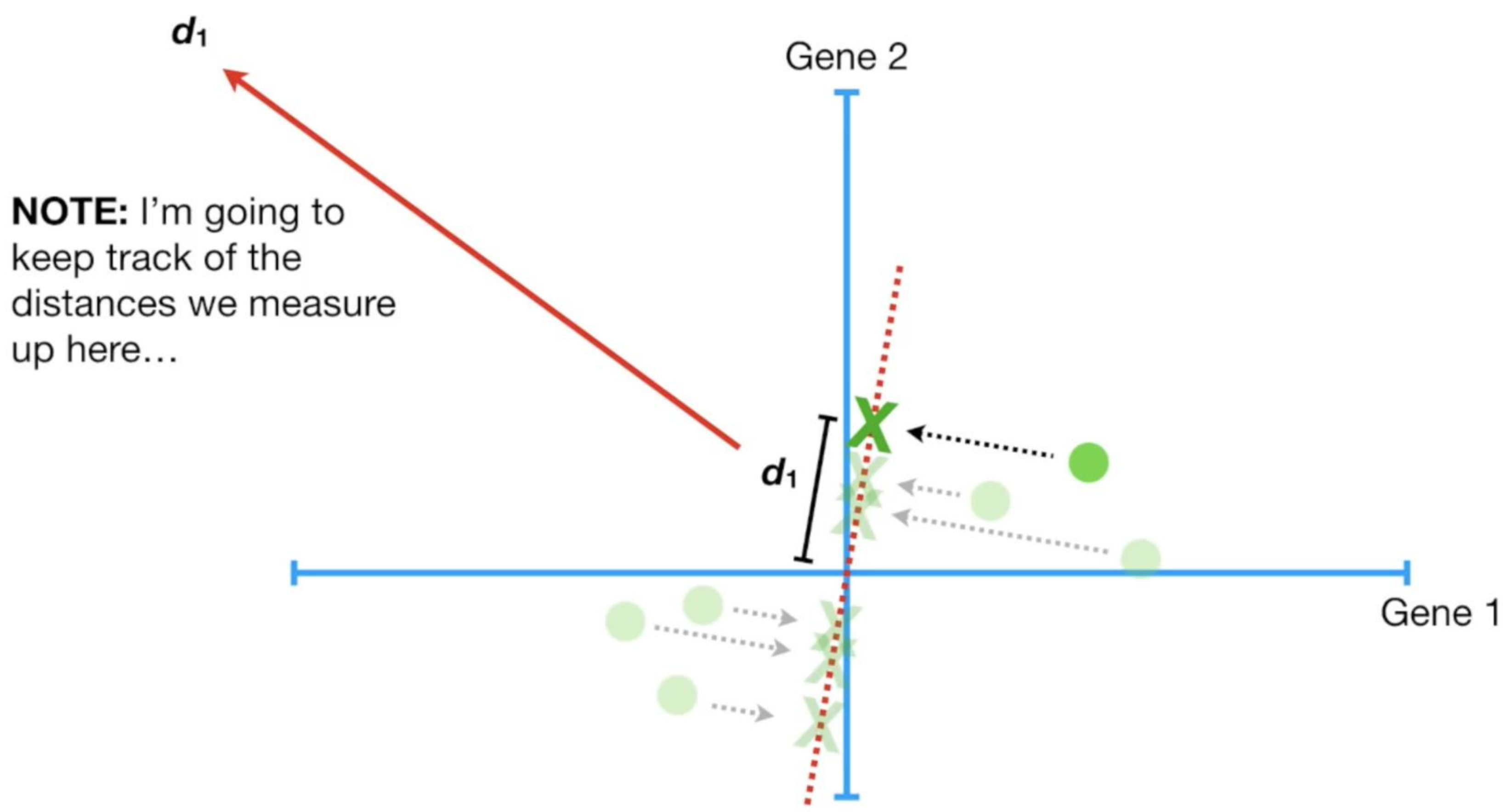
Gene 2
…then we can use the Pythagorean theorem to show how ***b*** and ***c*** are inversely related.
$a^2 = b^2 + c^2$
That means that if we label the sides like this…
b
c
a
Gene 1

Why?

Objective: Minimize b or Maximize c

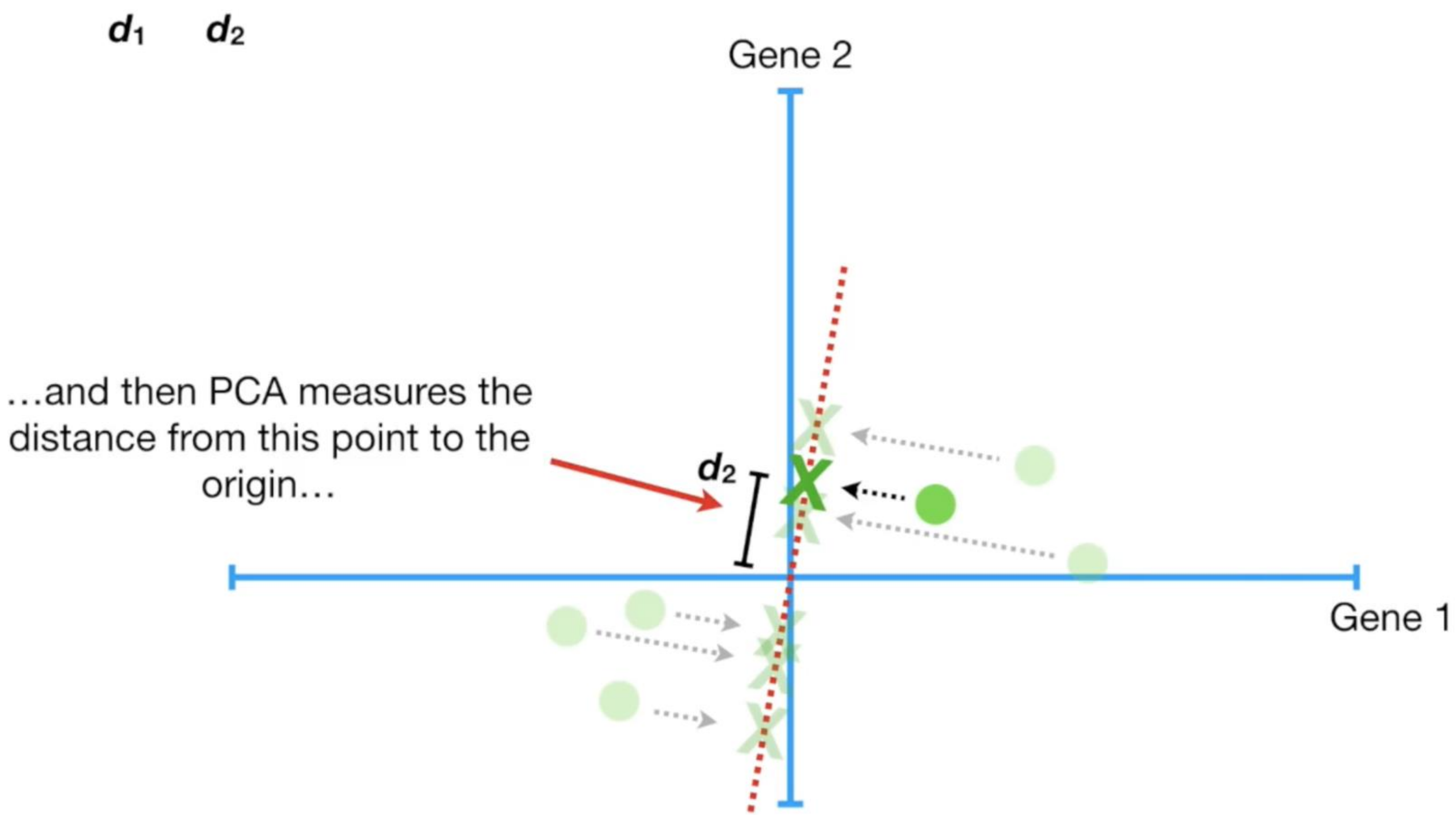


Distances



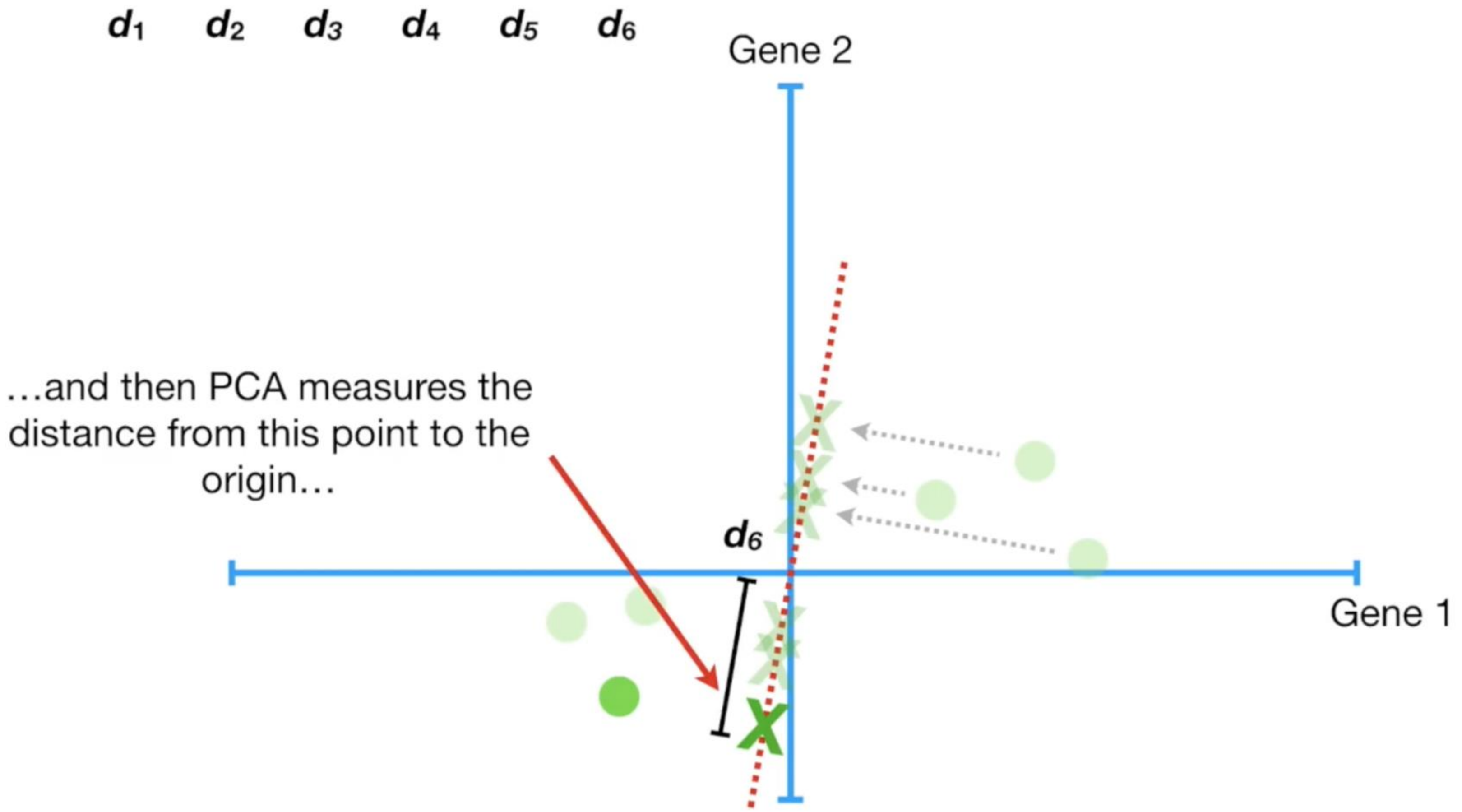
d1
Gene 2
NOTE: I'm going to keep track of the distances we measure up here…
d1
Gene 1

Distances



d1 d2
Gene 2
...and then PCA measures the distance from this point to the origin...
d2
Gene 1

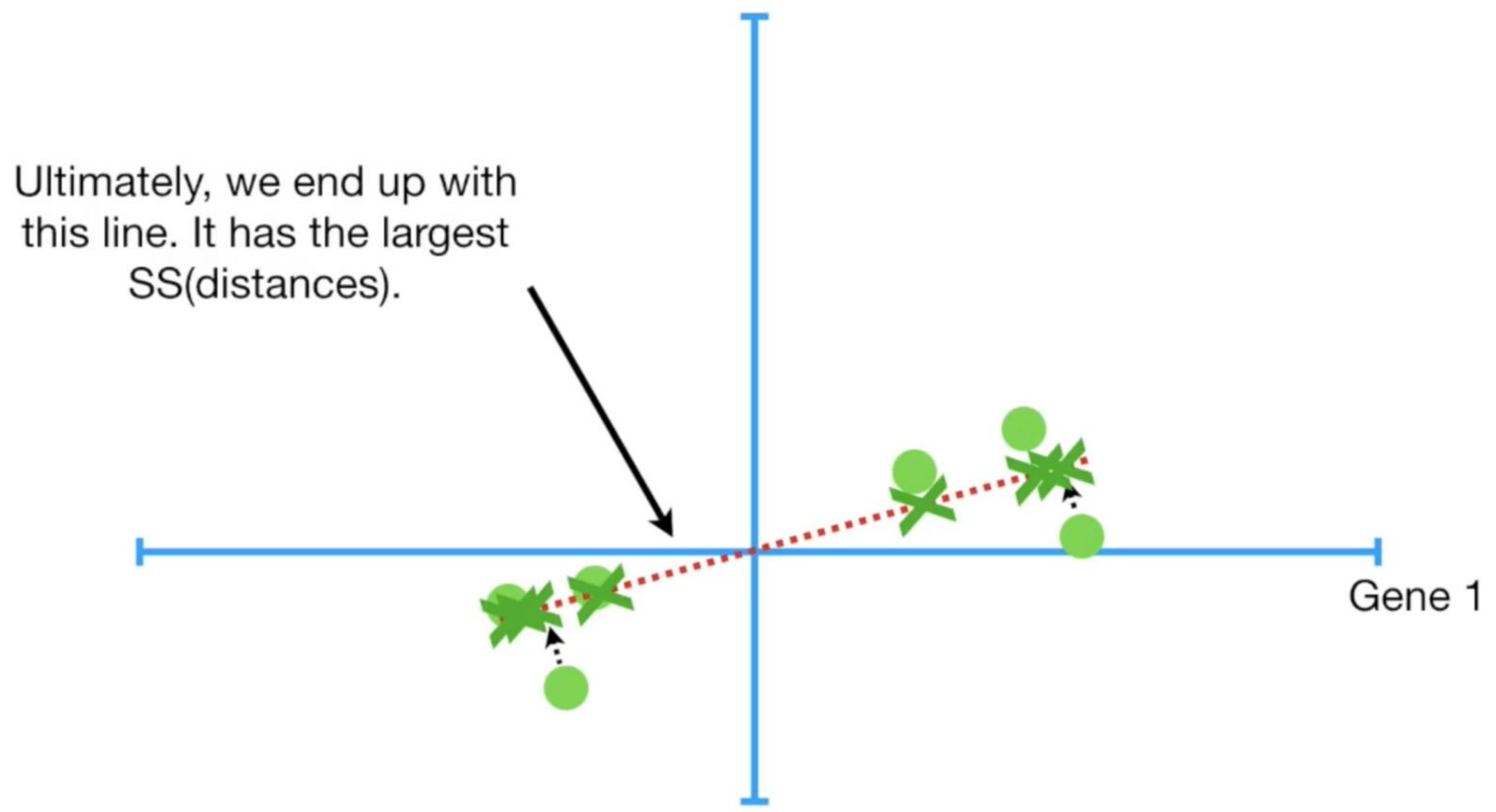
Distances



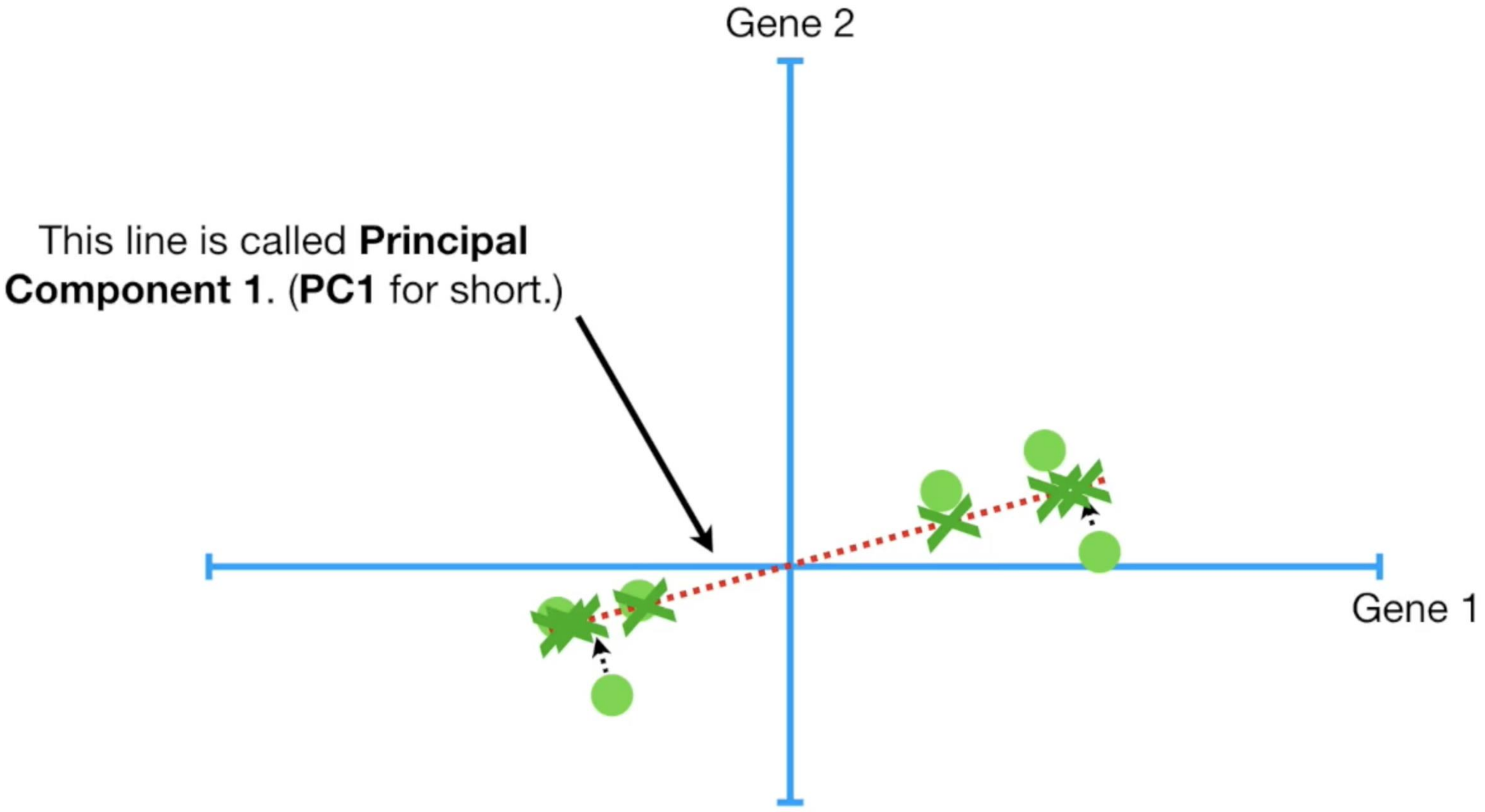
d1 d2 d3 d4 d5 d6
Gene 2
...and then PCA measures the distance from this point to the origin...
d6
Gene 1

Maximize the Sum of squared distances

$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2$ = sum of squared distances = SS(distances)

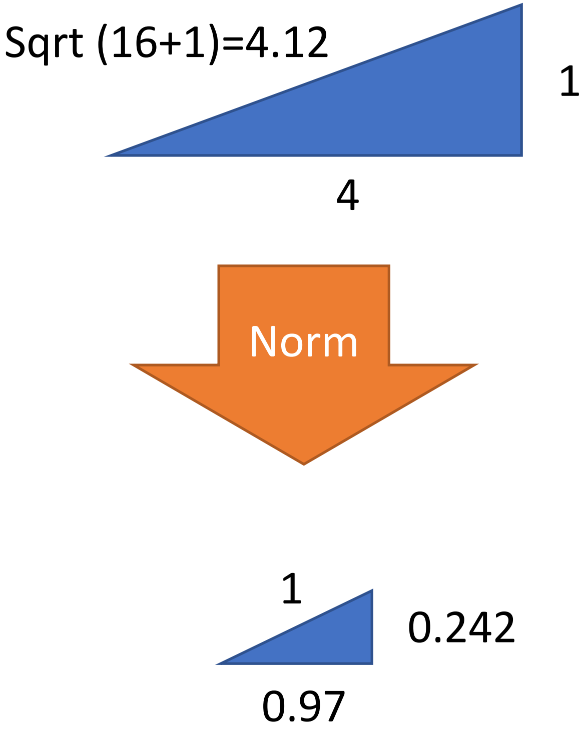


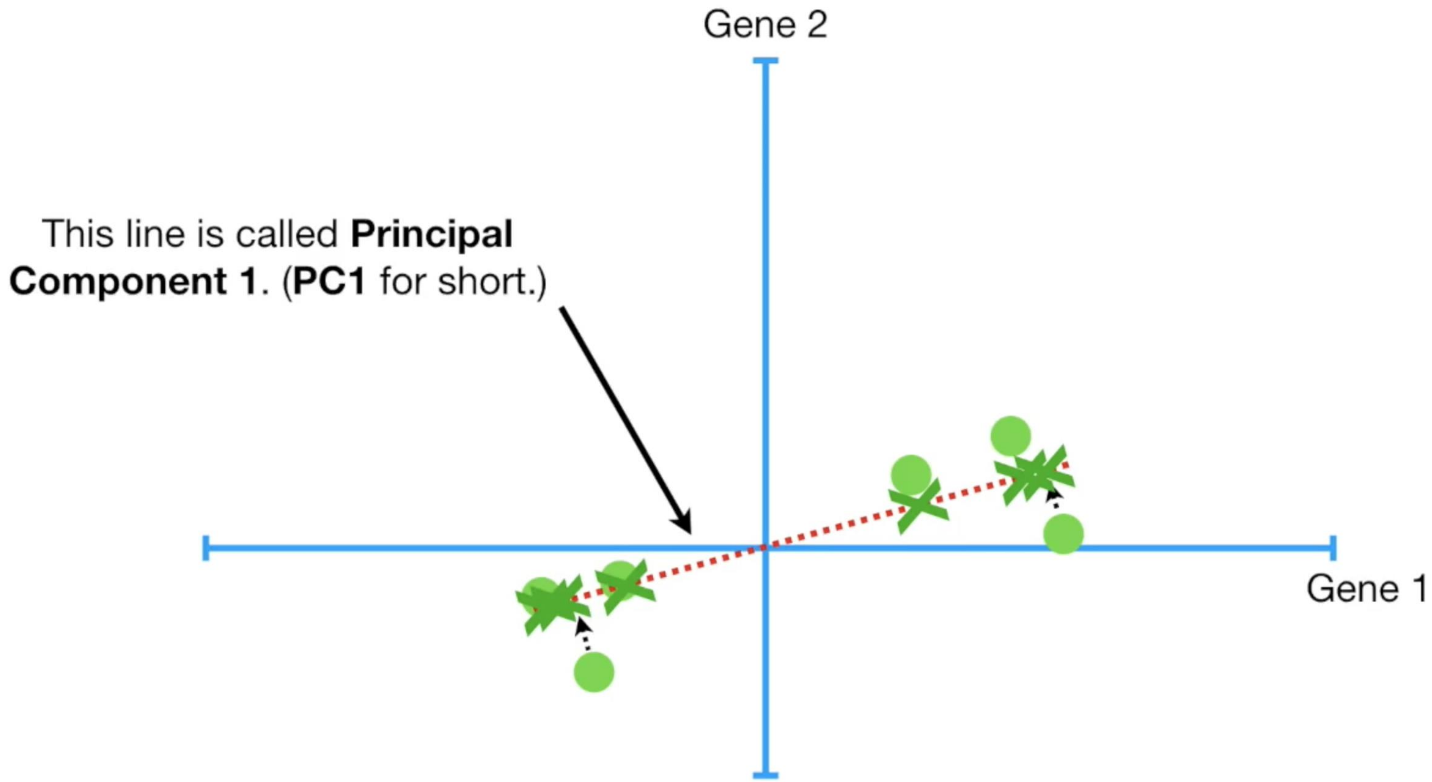
First Principal Component



First Principal Component

- The slope of the PC1 = 0.25

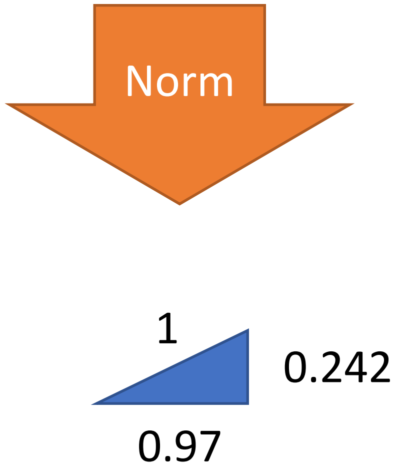


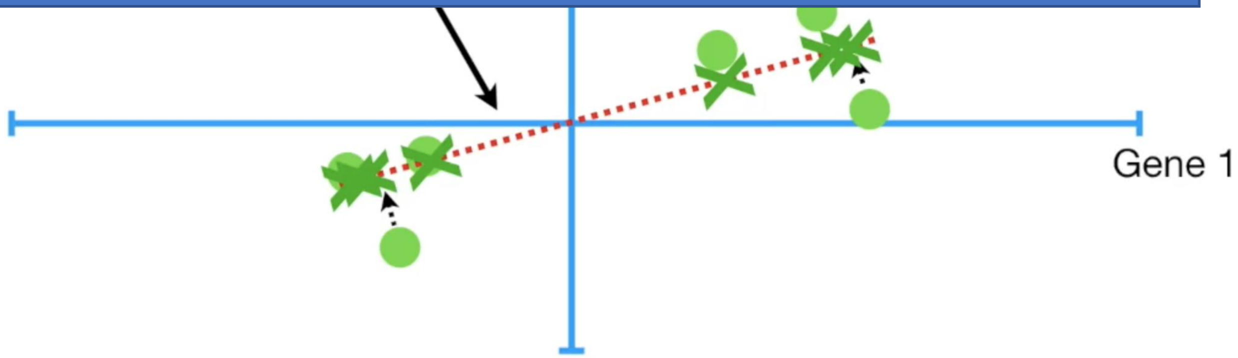


First Principal Component

Eigenvector for PC1 = [0.97, 0.242]

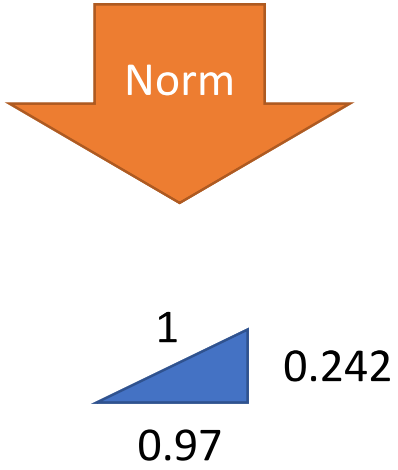
Other name: Singular vector

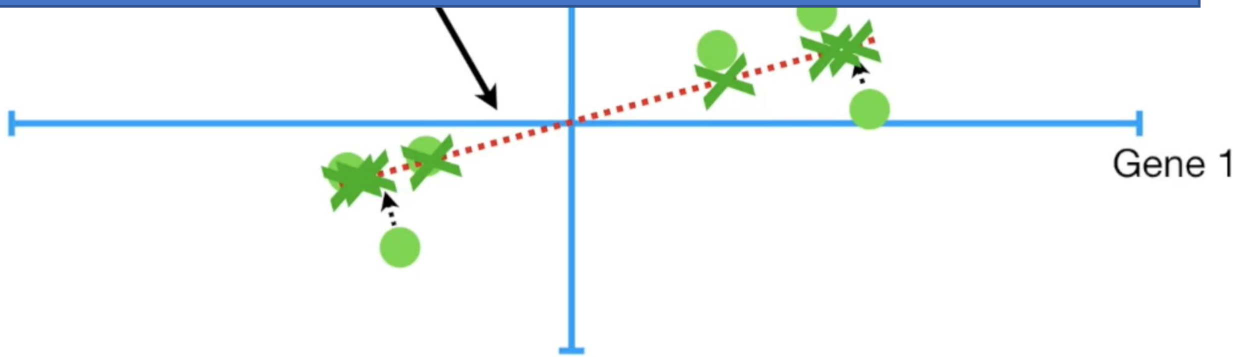




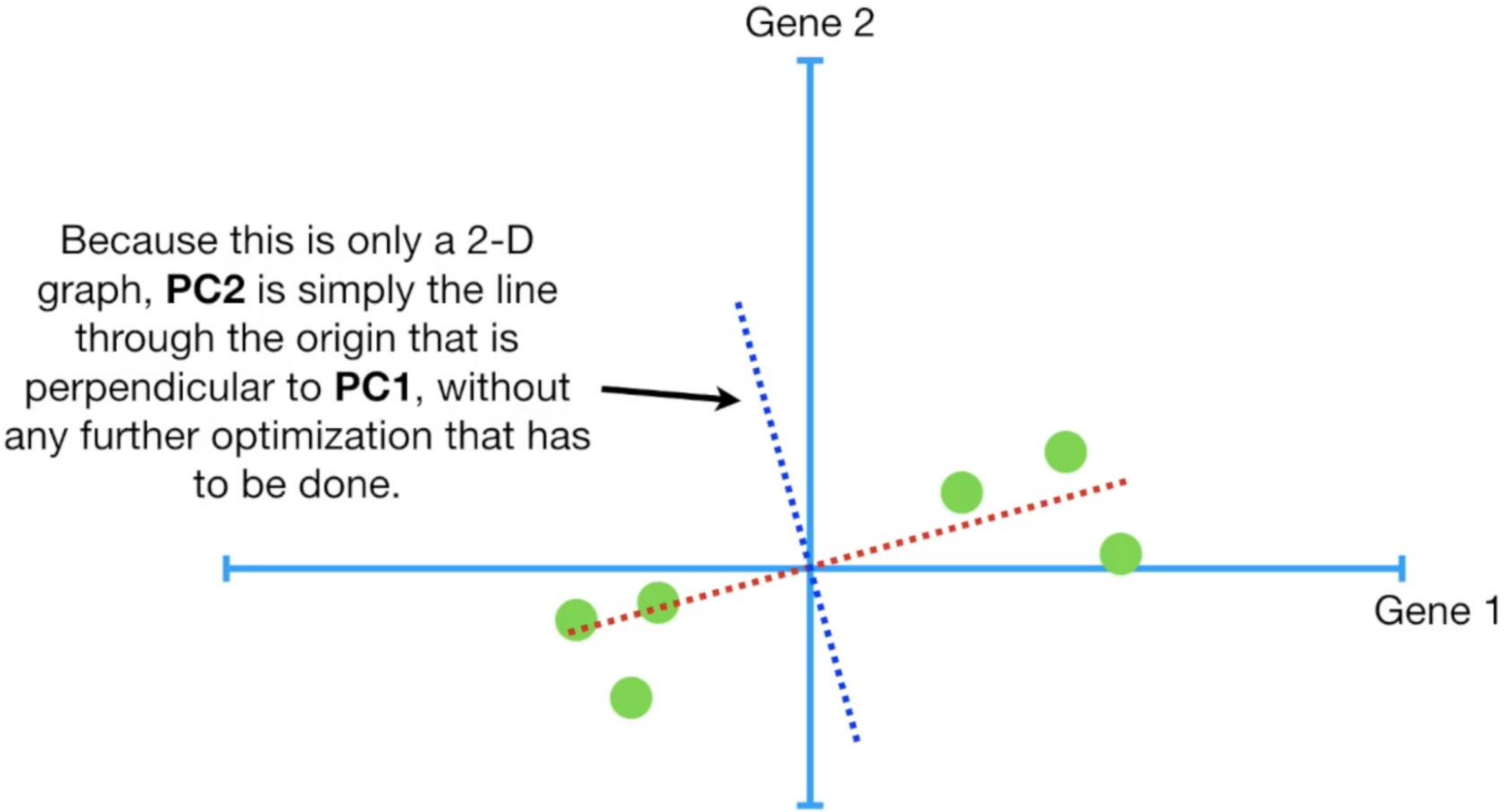
First Principal Component

Sum of squared distances for PC1 = **Eigenvalue** for PC1
Sqrt(Eigenvalue for PC1) = Singular Value for PC1

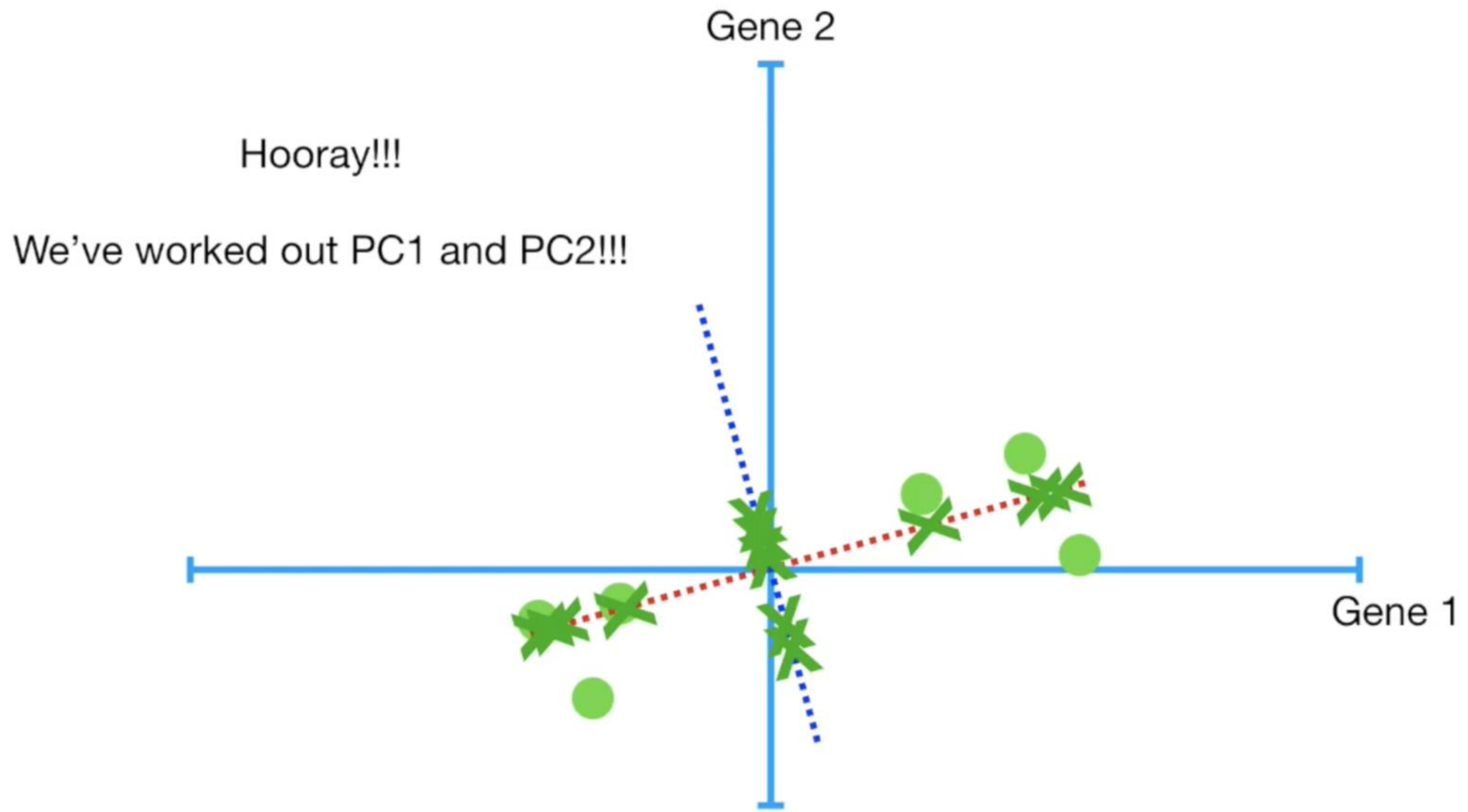




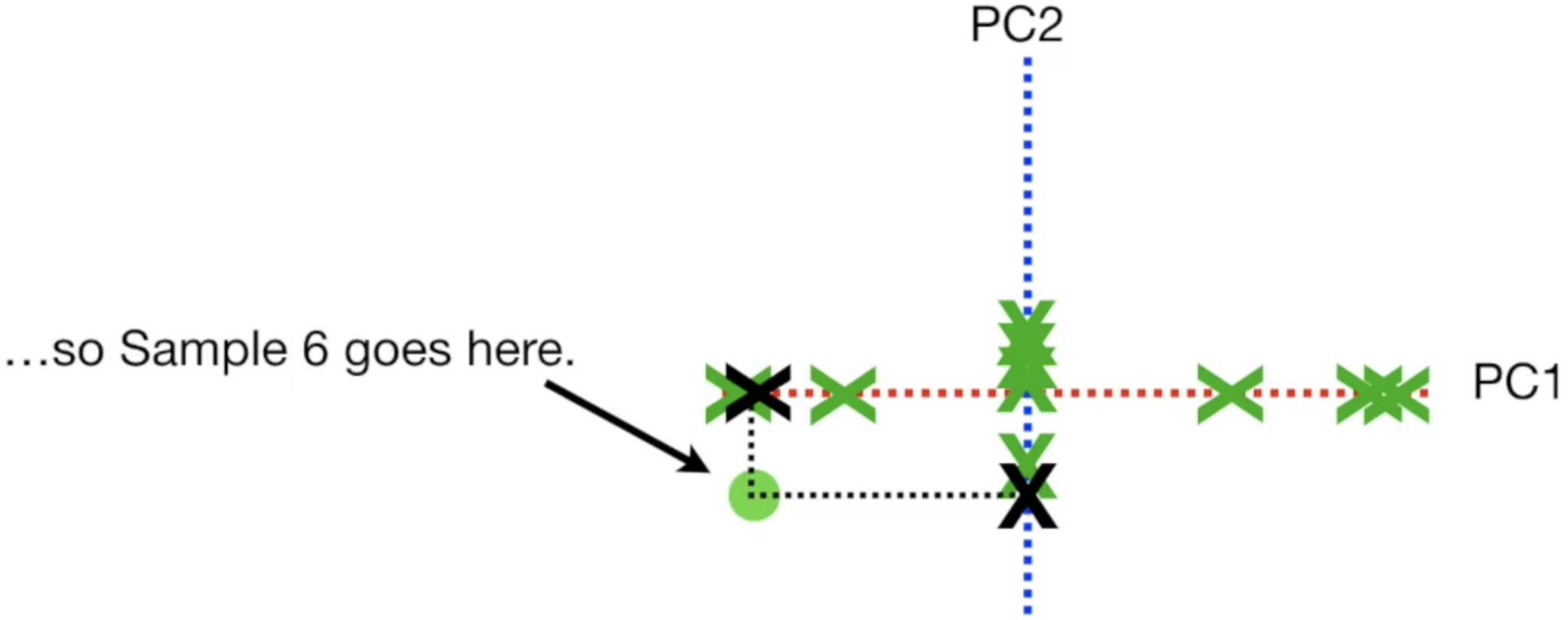
PC2


Gene 2
Because this is only a 2-D graph, **PC2** is simply the line through the origin that is perpendicular to **PC1**, without any further optimization that has to be done.
Gene 1

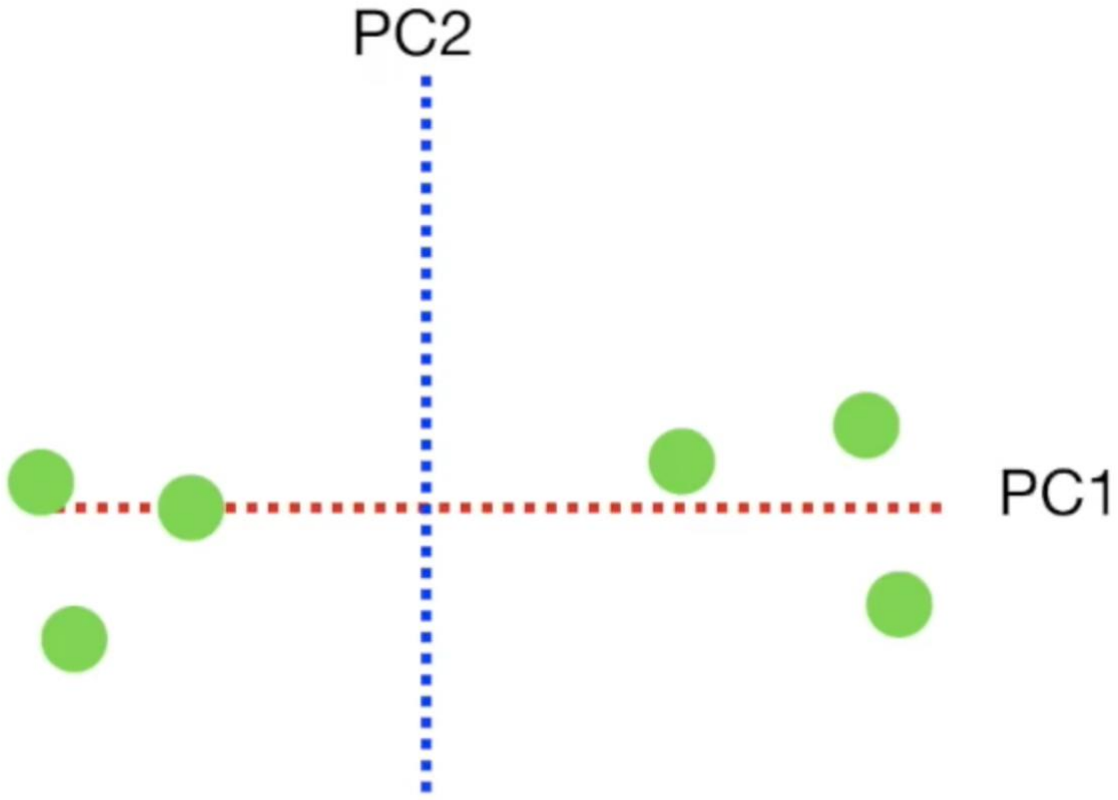
Projection - PC1 and PC2



Reconstruction

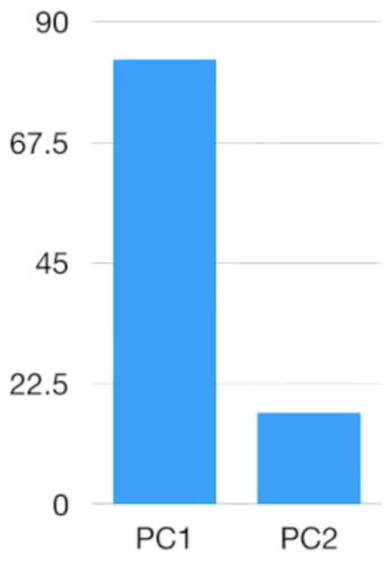


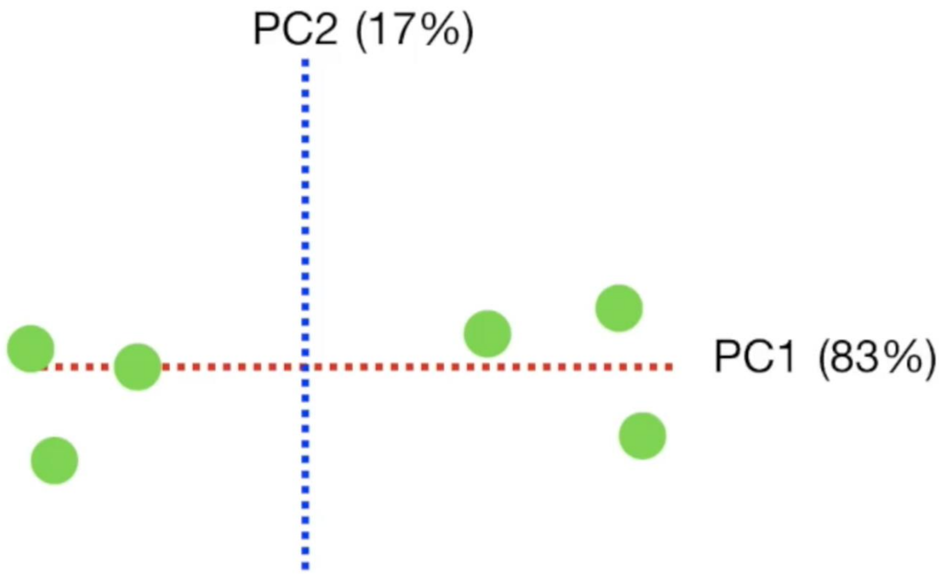
Reconstruction



Scree Plot

TERMINOLOGY ALERT!!!! A **Scree Plot** is a graphical representation of the percentages of variation that each PC accounts for.



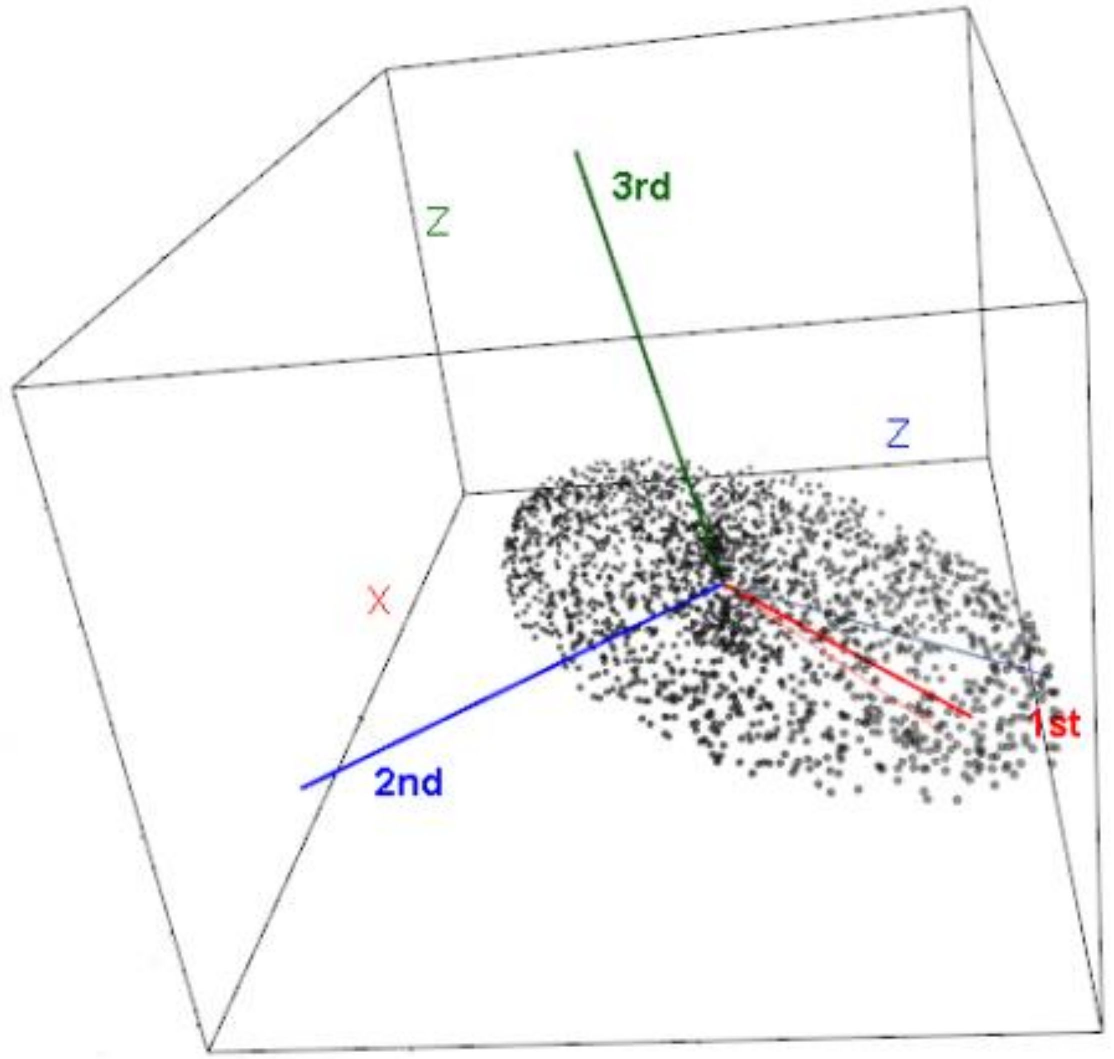


What happens beyond 2D?

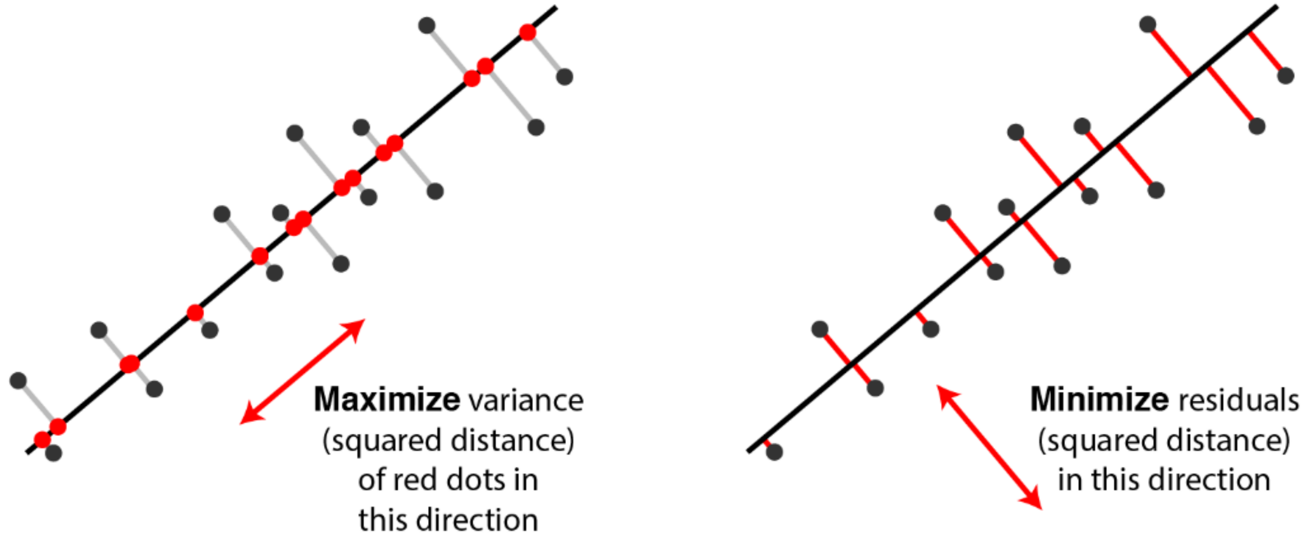
- When finding PC2:
  - Should cross the origin
  - Be perpendicular to PC1
  - **Minimizes the sum of squared distances**

How to obtain PCA?

- We rigidly rotate the coordinate axes to new 'natural' positions (principal axes) such that:
  - Principal axis 1 corresponds to the highest variance, axis 2 has the next highest variance, . . . , and axis *p* has the lowest variance.
  - All principal axes are uncorrelated.



In summary…



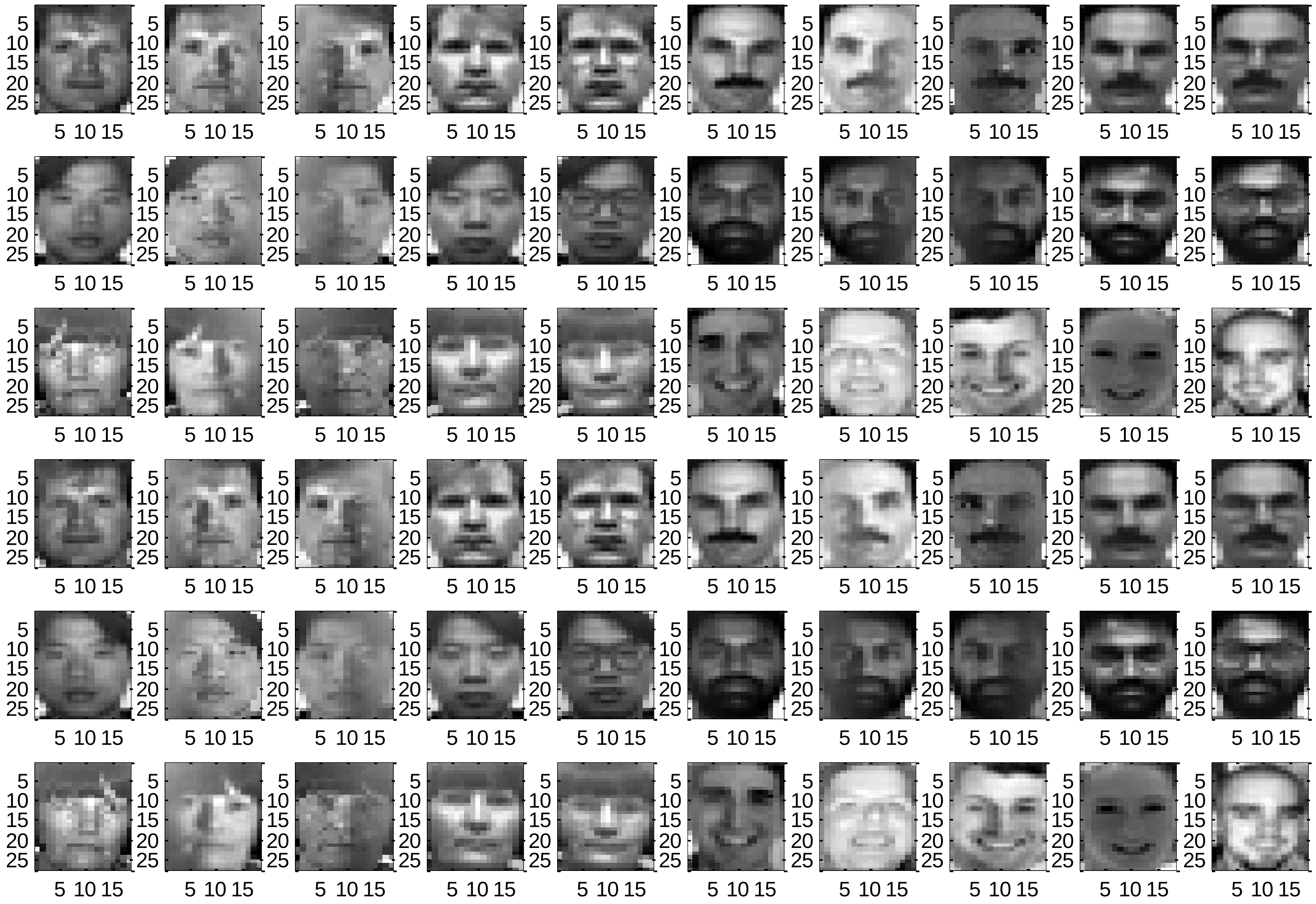
- Project D-dimensional data onto an M-dimensional subspace that retains as much information as possible
- Informally: information = diversity = variance

Graphic – Alex Williams

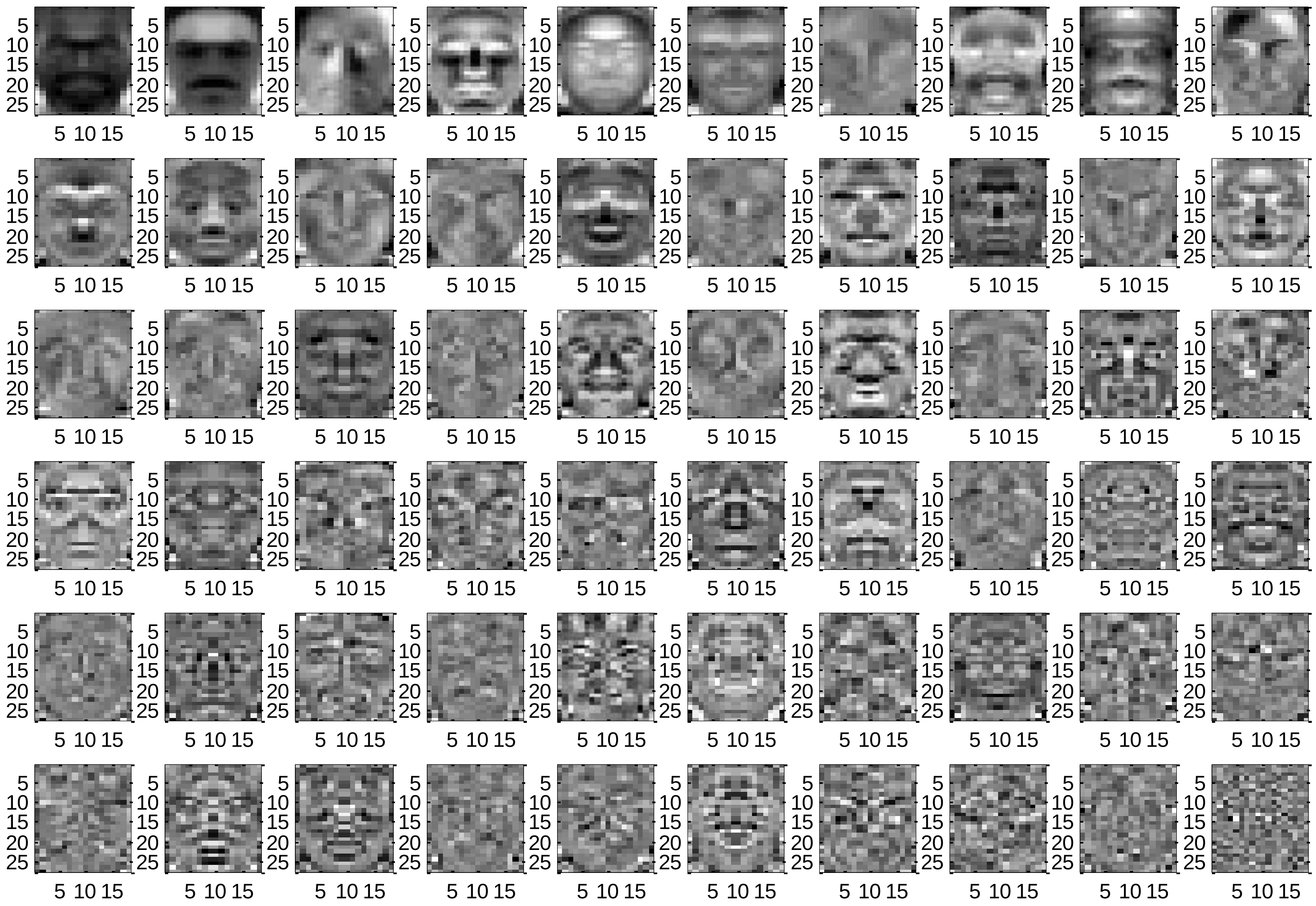
What About Images?

- Images? Yes, we can use PCA! (Turk, Pentland, 1991)
- Each 150 pixel × 150 pixel image can be represented by a 22,500-long vector

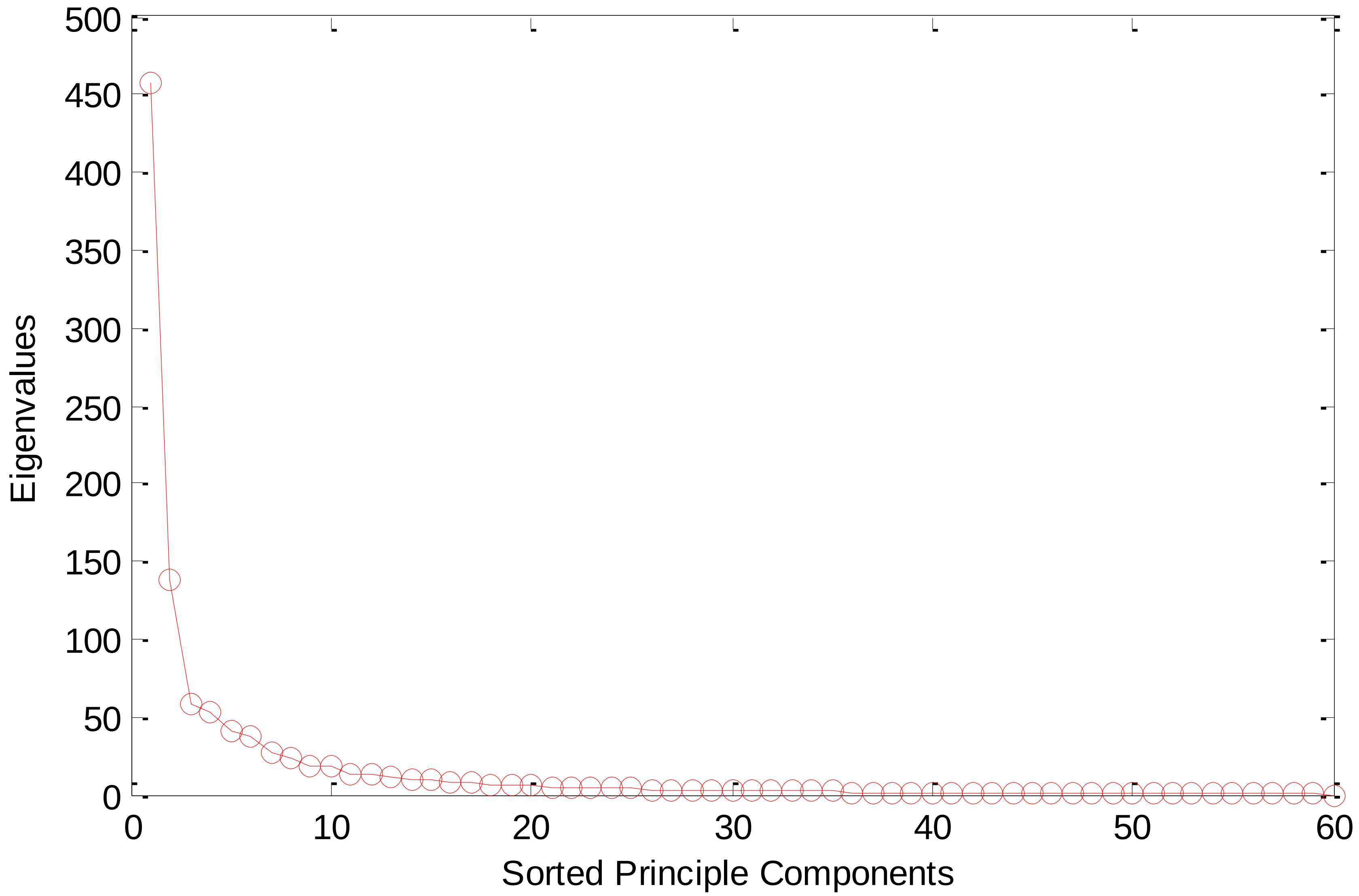
Normalized Face Data



The Eigenfaces

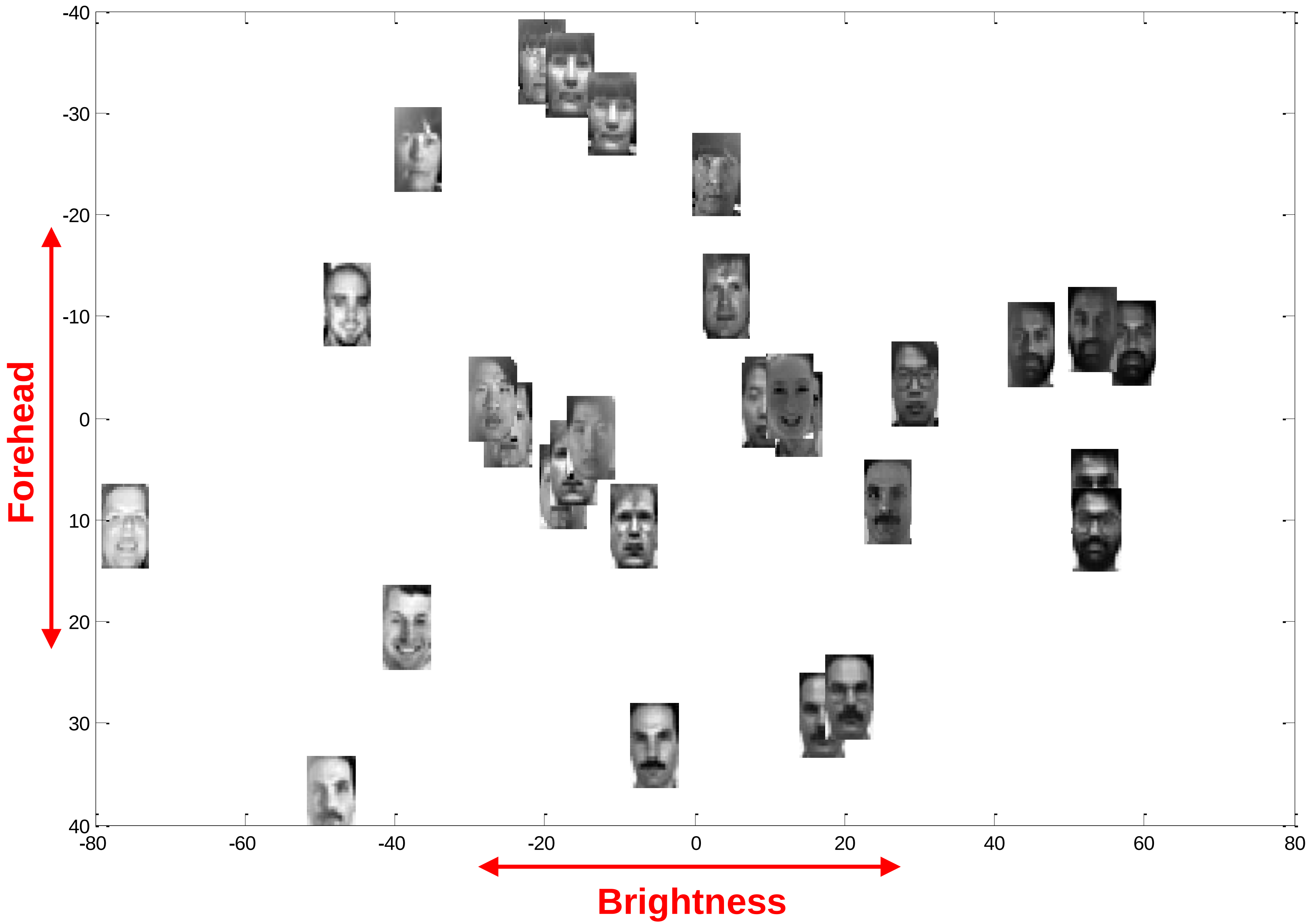


Eigenvalues for each Eigenface

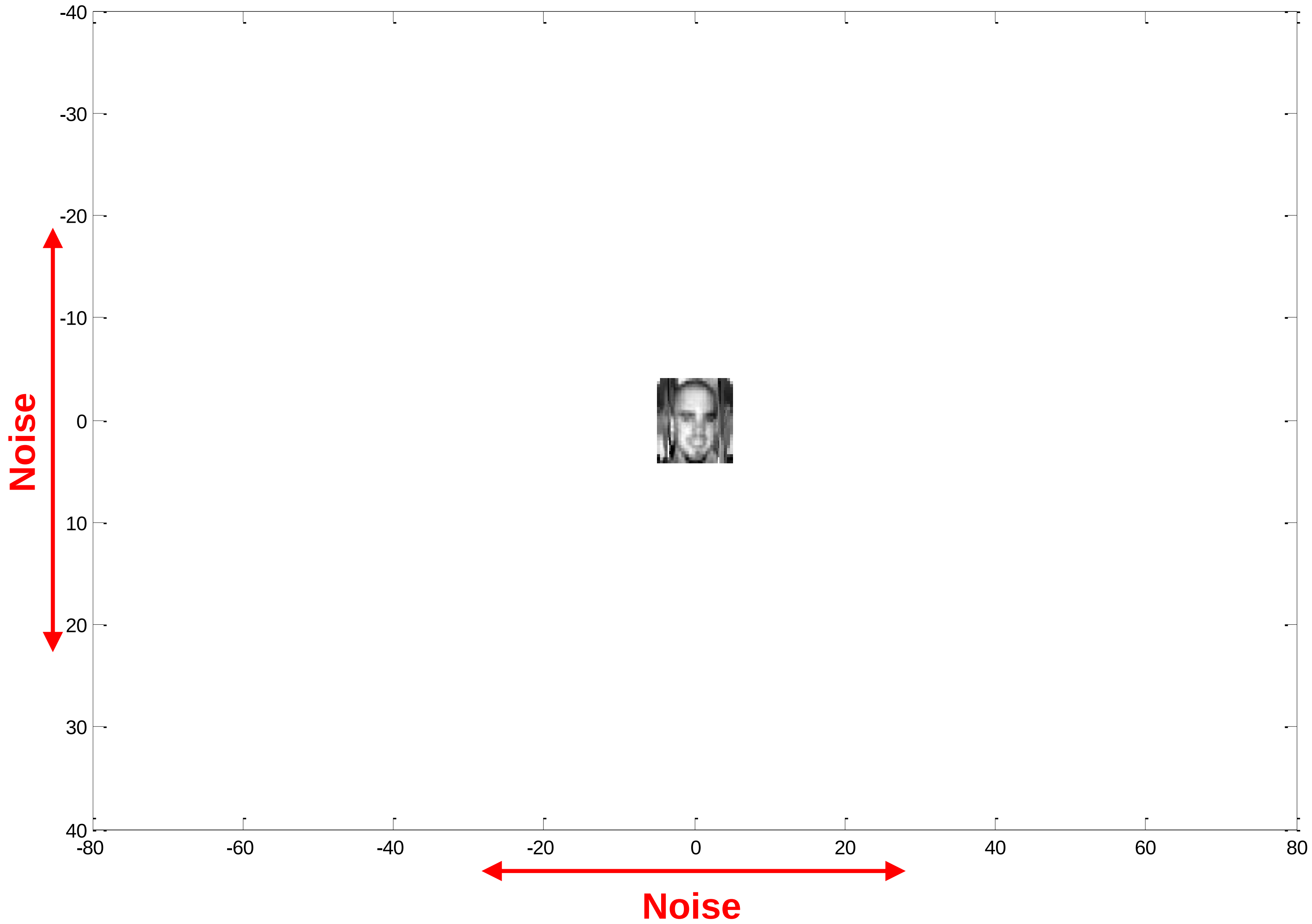


Reconstructing a Face from PCs

Faces in PC-space (Dims 1,2)



Faces in PC-space (Dims 59,60)

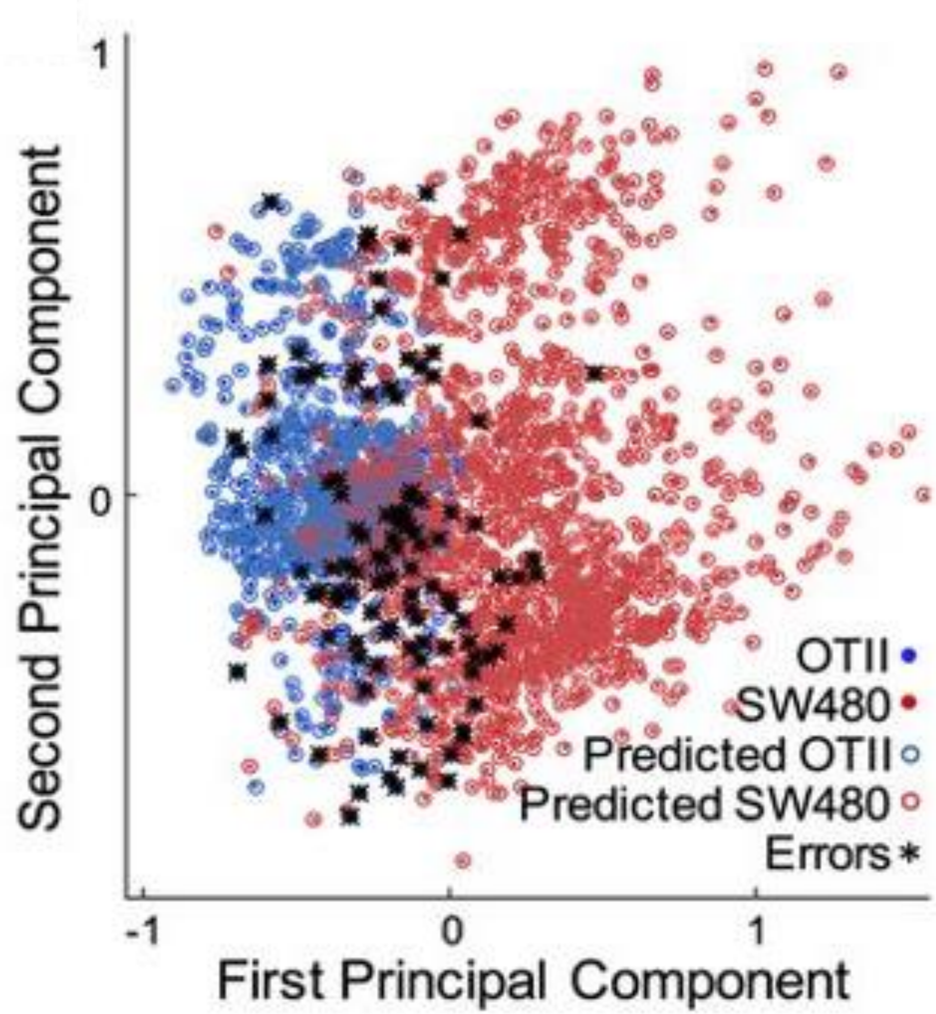


PCA Overview

1. Standardize the data.
2. Obtain the Eigenvectors and Eigenvalues
3. Choose the k eigenvectors that correspond to the k largest eigenvalues
4. Construct the projection matrix $\mathbf{W}$ from the selected k eigenvectors.
5. Transform the original dataset $\mathbf{X}$ via $\mathbf{W}$ to obtain a k-dimensional feature subspace $\mathbf{Y}$.

Why use PCA?

- Speed-up machine learning!
- Data Visualization!



From Chen et al., *Deep Learning in Label-free Cell Classification*, 2016

Additional Reading:
Mathematics for Machine Learning – Chapter 10