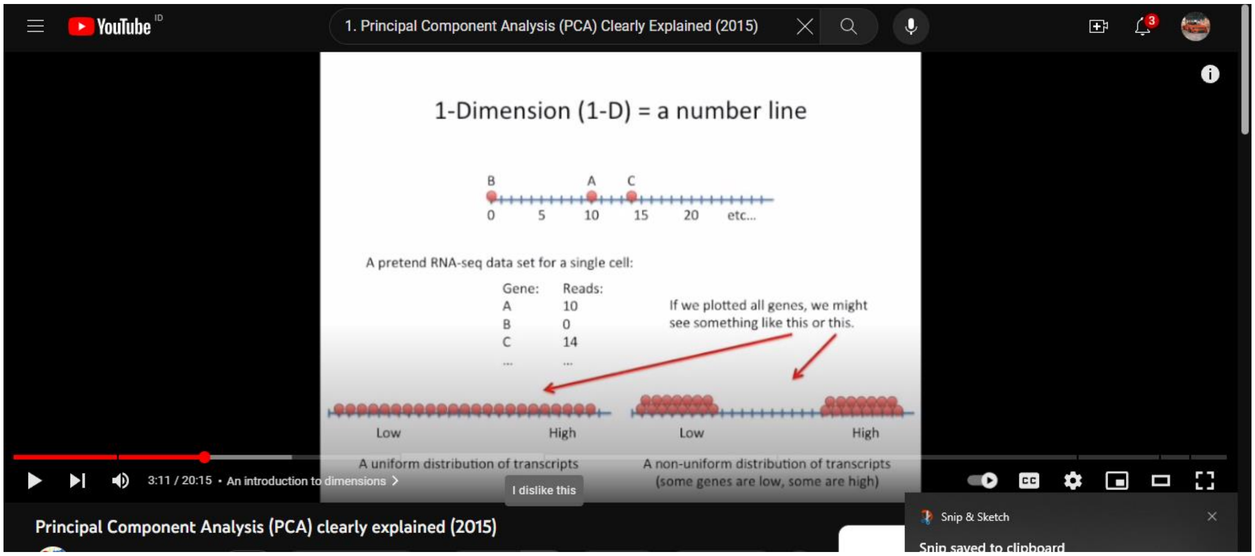
Nama : Arfara Yema Samgusdian

NIM : 1103202004

Kelas : TK-44-G4

1. Principal Component Analysis (PCA) Clearly Explained (2015)


YouTube ID
1. Principal Component Analysis (PCA) Clearly Explained (2015)
B
A
C
0
5
10
15
20
etc...
Low
High
Low
High
I dislike this
3:11 / 20:15 • An introduction to dimensions
Snip & Sketch
Snip saved to clipboard

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

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

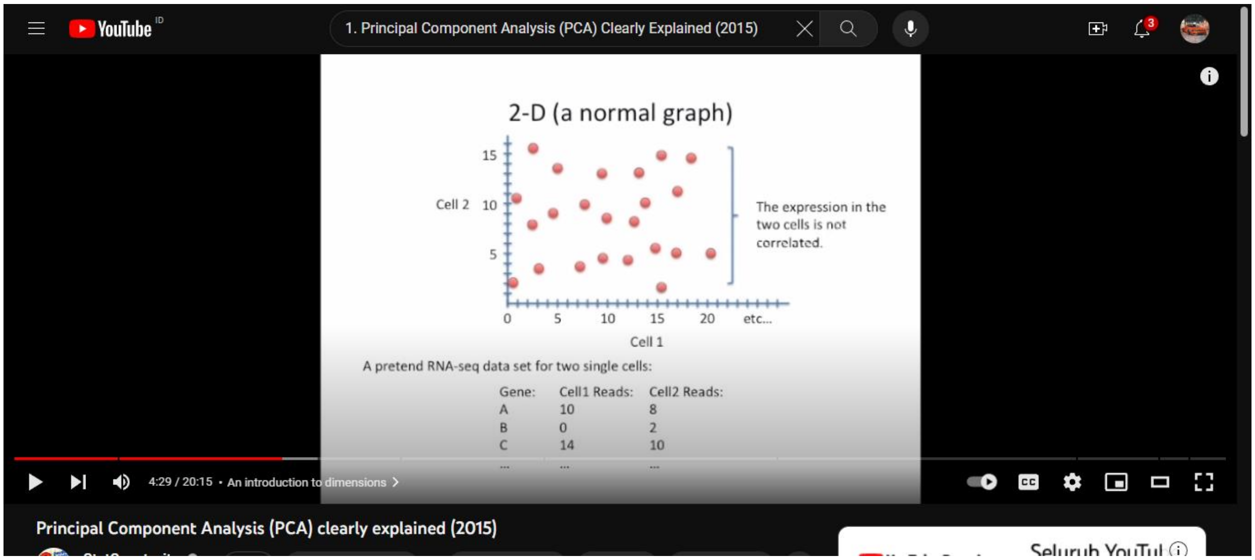
Gene:	Reads:
A	10
B	0
C	14
...	...

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)

Principal Component Analysis (PCA) clearly explained (2015)


YouTube ID
1. Principal Component Analysis (PCA) Clearly Explained (2015)
15
Cell 2
10
5
0
5
10
15
20
etc...
Cell 1
4:29 / 20:15 • An introduction to dimensions
Seluruh YouTul

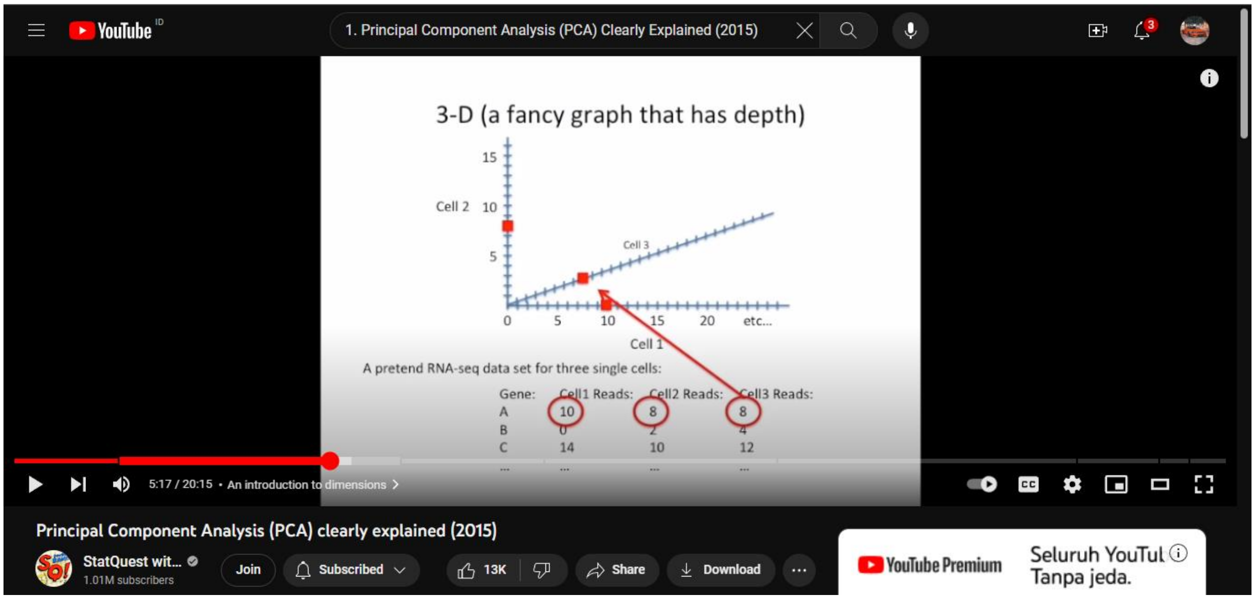
2-D (a normal graph)

The expression in the two cells is not correlated.

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

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

Principal Component Analysis (PCA) clearly explained (2015)

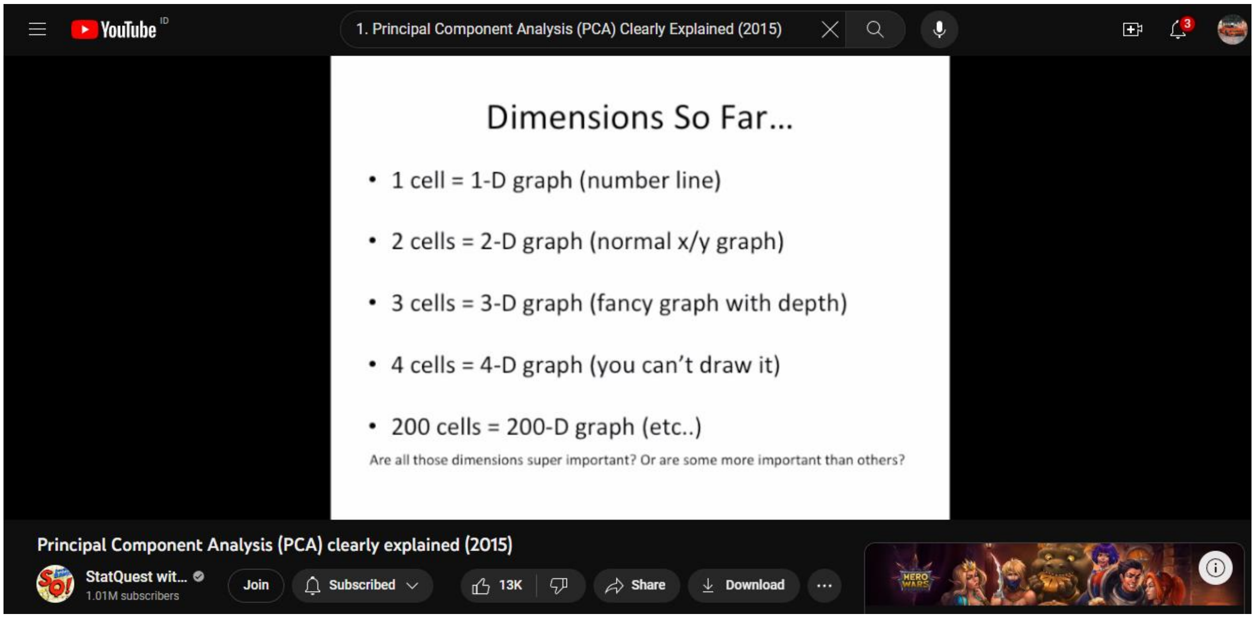


3-D (a fancy graph that has depth)

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

Gene:	Cell1 Reads:	Cell2 Reads:	Cell3 Reads:
A	10	8	8
B	0	2	4
C	14	10	12
…	…	…	…

Principal Component Analysis (PCA) clearly explained (2015)

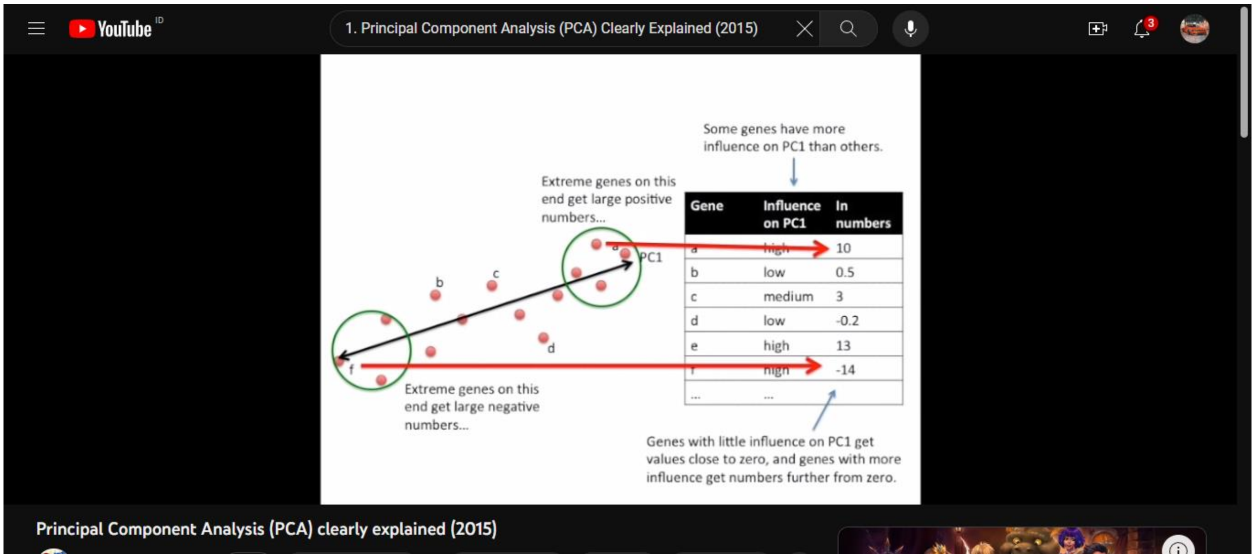


Dimensions So Far…

- 1 cell = 1-D graph (number line)
- 2 cells = 2-D graph (normal x/y graph)
- 3 cells = 3-D graph (fancy graph with depth)
- 4 cells = 4-D graph (you can't draw it)
- 200 cells = 200-D graph (etc..)

Are all those dimensions super important? Or are some more important than others?

Principal Component Analysis (PCA) clearly explained (2015)


YouTube
1. Principal Component Analysis (PCA) Clearly Explained (2015)
b
c
a
PC1
d
f

Some genes have more influence on PC1 than others.

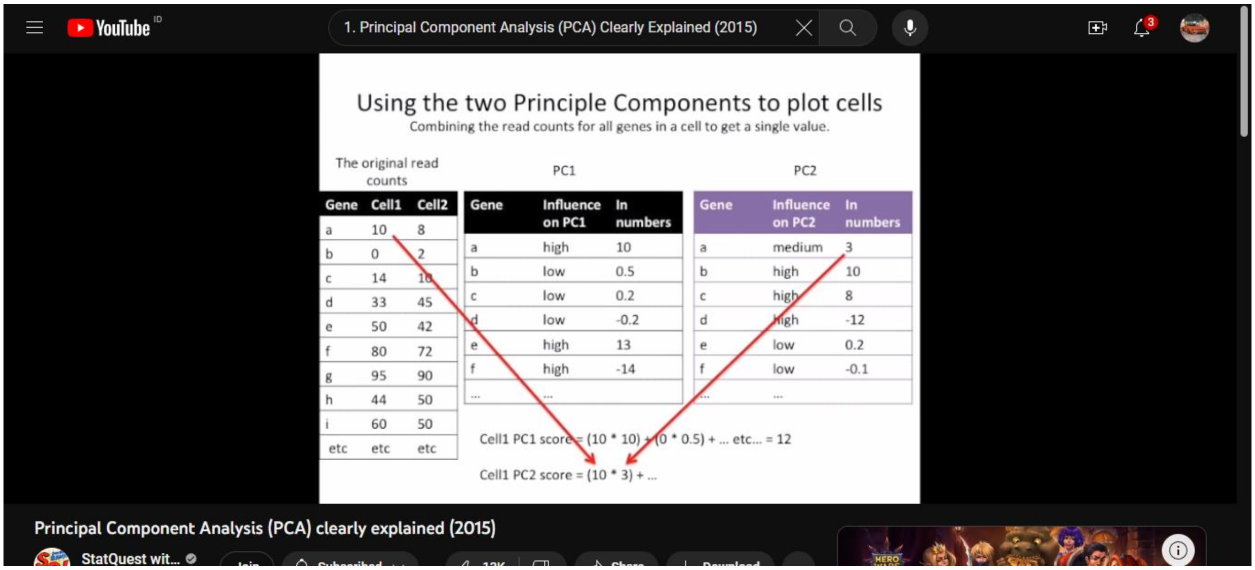
Extreme genes on this end get large positive numbers...

Gene	Influence on PC1	In numbers
a	high	10
b	low	0.5
c	medium	3
d	low	-0.2
e	high	13
f	high	-14
...	...	

Extreme genes on this end get large negative numbers...

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

Principal Component Analysis (PCA) clearly explained (2015)


YouTube
1. Principal Component Analysis (PCA) Clearly Explained (2015)
StatQuest wit...

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

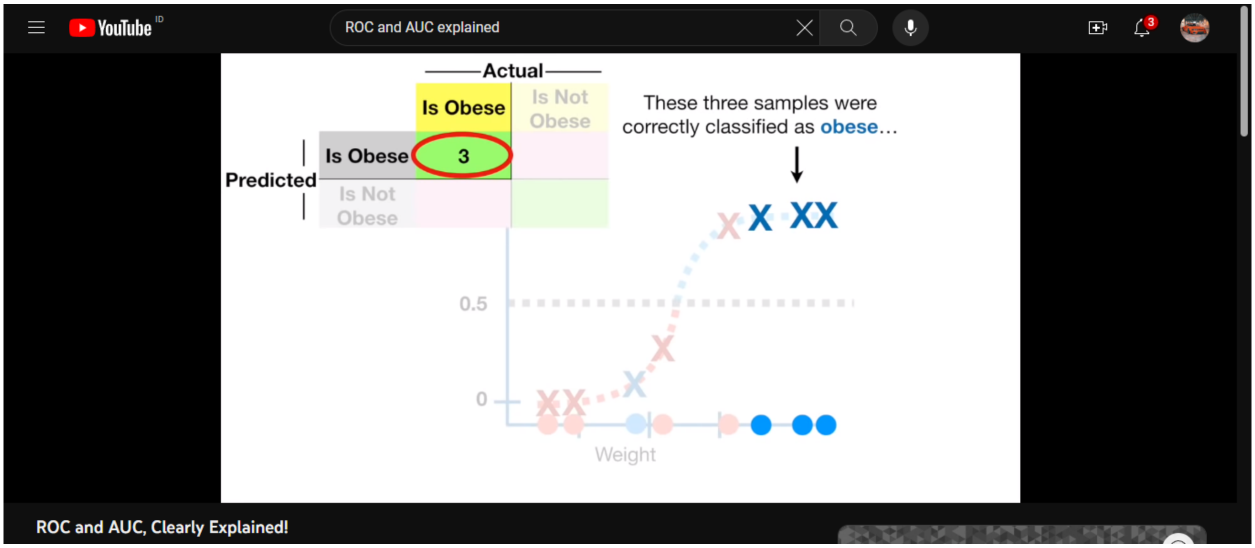
Gene	Influence on PC2	In numbers
a	medium	3
b	high	10
c	high	8
d	high	-12
e	low	0.2
f	low	-0.1
...	...	

Cell1 PC1 score = (10 * 10) + (0 * 0.5) + ... etc... = 12

Cell1 PC2 score = (10 * 3) + ...

Principal Component Analysis (PCA) clearly explained (2015)

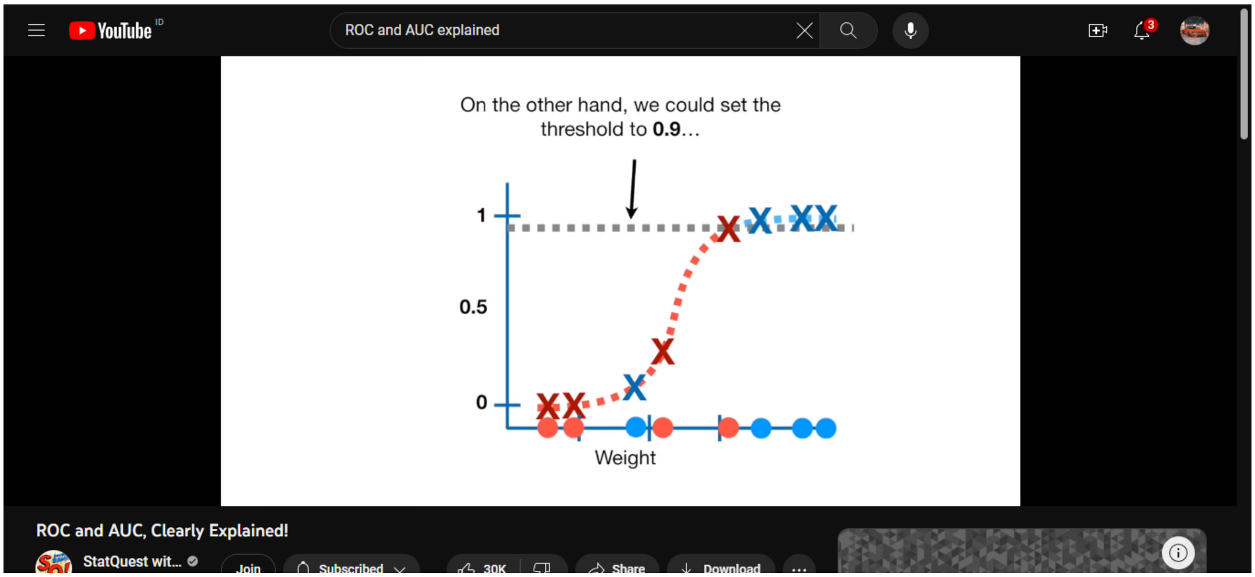
2. ROC and AUC explained



		Actual	
		Is Obese	Is Not Obese
Predicted	Is Obese	3	
	Is Not Obese		

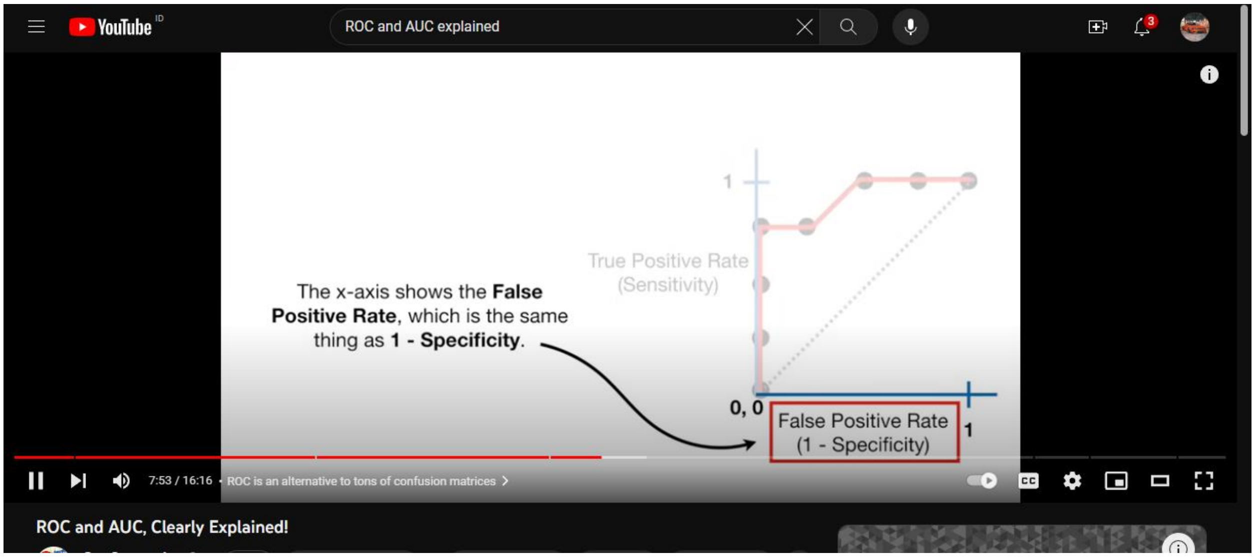
These three samples were correctly classified as **obese**…

ROC and AUC, Clearly Explained!



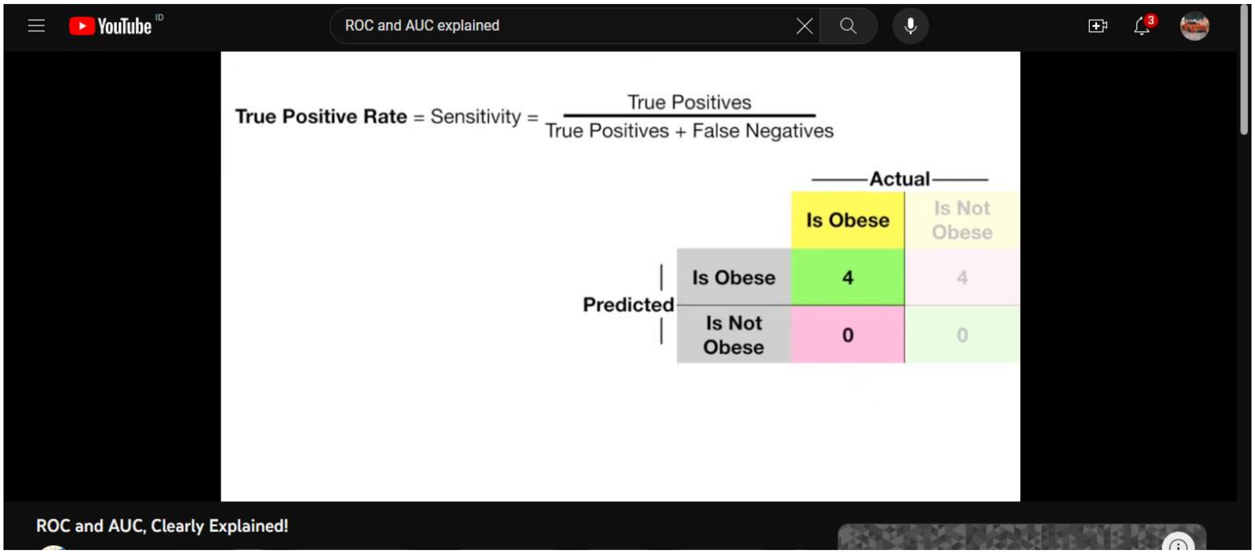
On the other hand, we could set the threshold to **0.9**…

ROC and AUC, Clearly Explained!



The x-axis shows the **False Positive Rate**, which is the same thing as **1 - Specificity**.

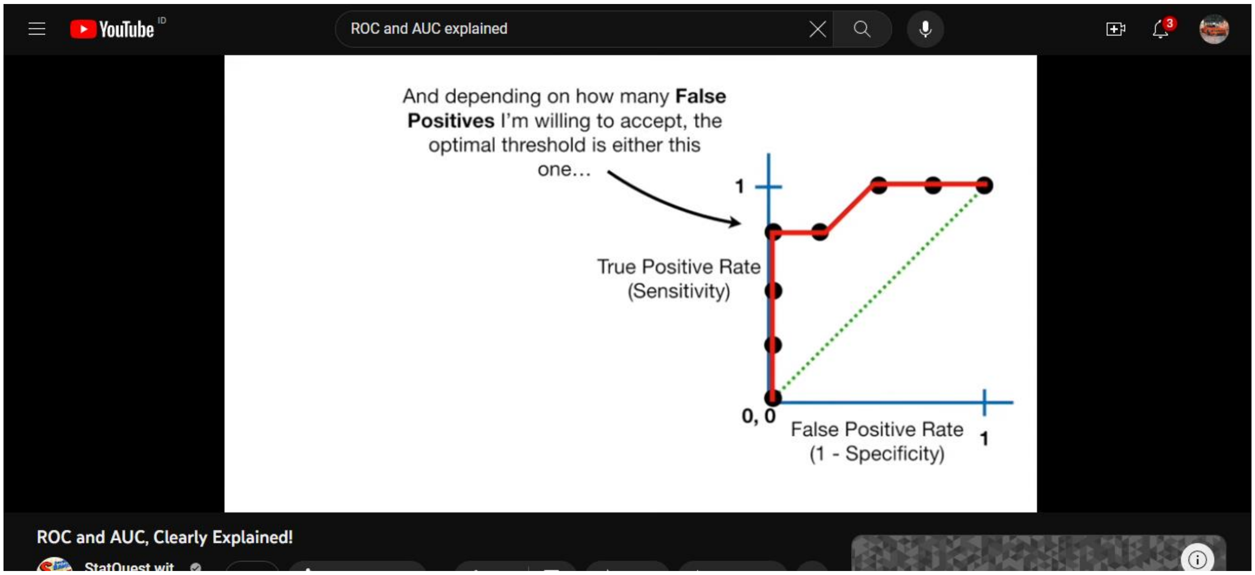
ROC and AUC, Clearly Explained!



$$\textbf{True Positive Rate} = \text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$

		Actual	
		Is Obese	Is Not Obese
Predicted	**Is Obese**	4	4
	Is Not Obese	0	0

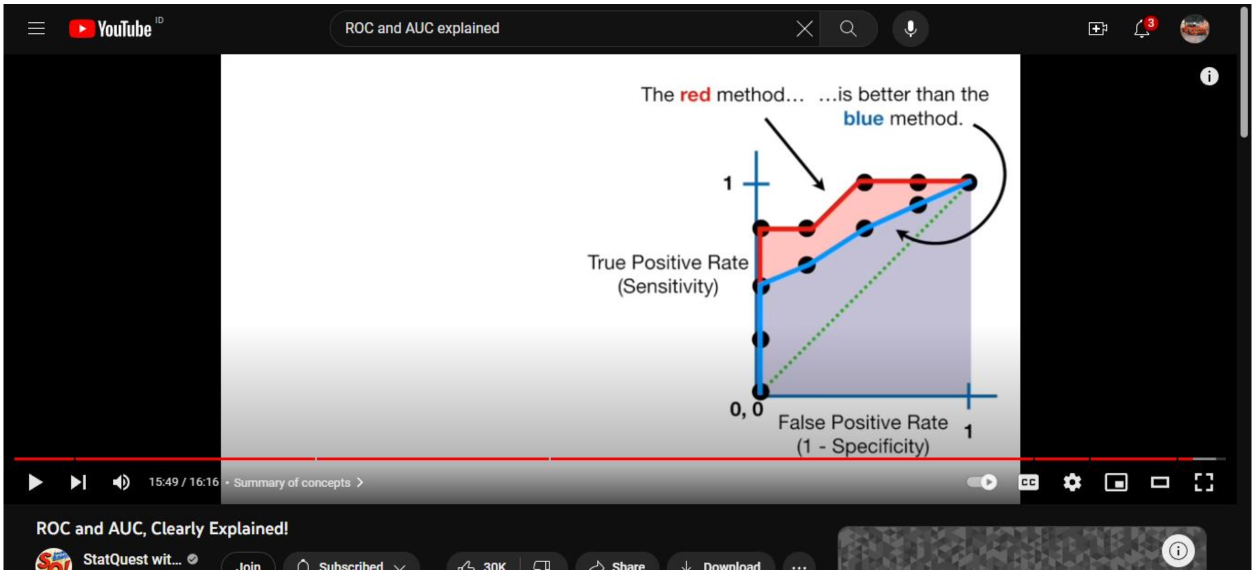
ROC and AUC, Clearly Explained!



And depending on how many **False Positives** I'm willing to accept, the optimal threshold is either this one…

ROC and AUC, Clearly Explained!

StatQuest wit...

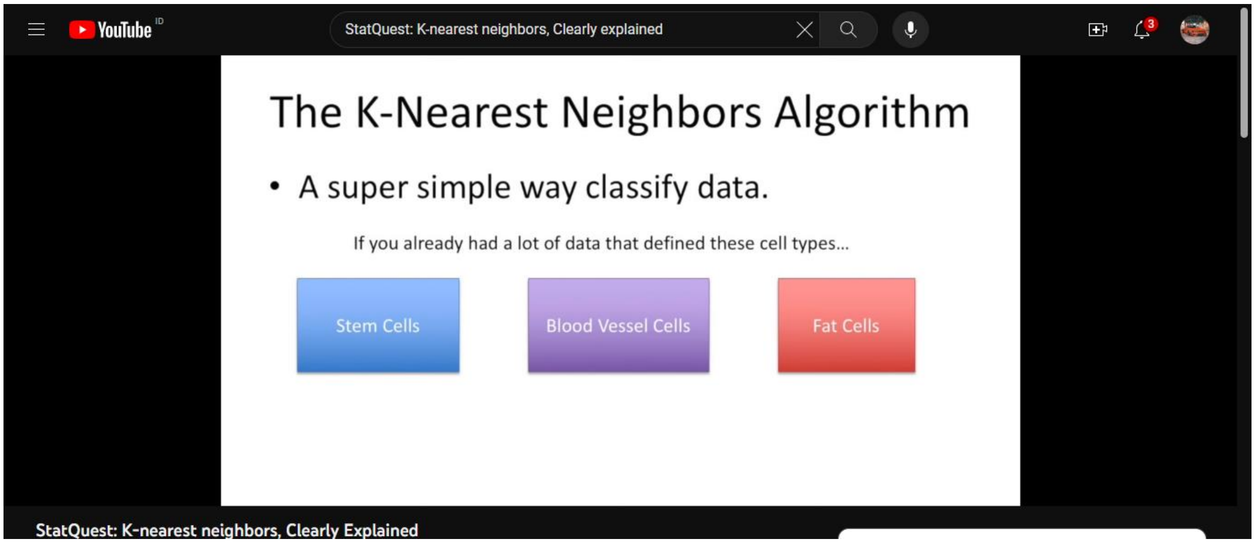


The **red** method… …is better than the **blue** method.

ROC and AUC, Clearly Explained!

StatQuest wit...

3. StatQuest: K-nearest neighbors, Clearly explained



The K-Nearest Neighbors Algorithm

- A super simple way classify data.

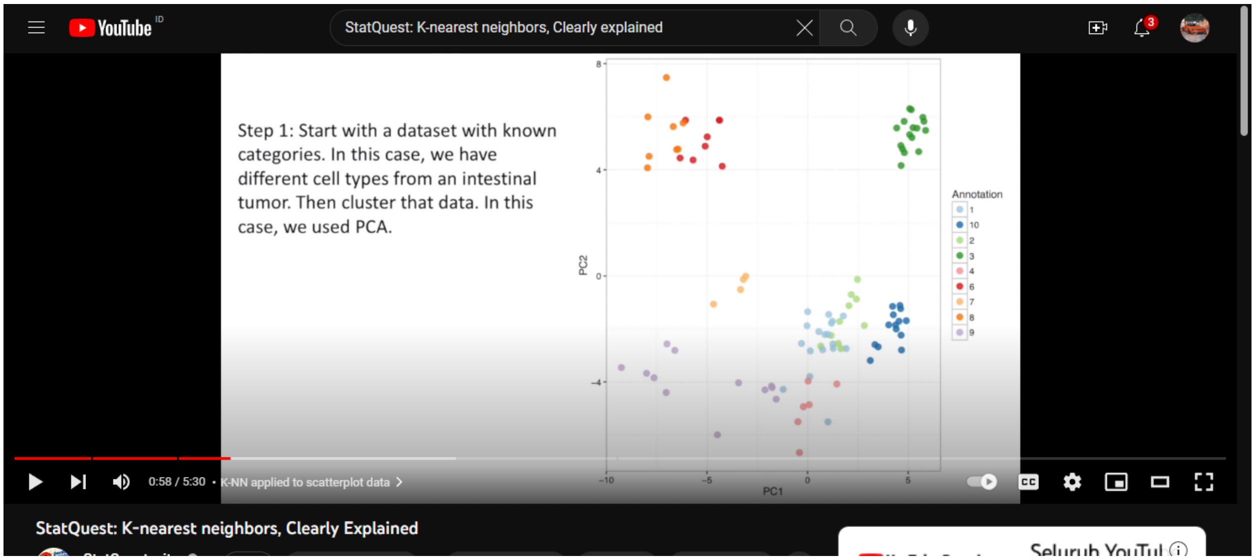
If you already had a lot of data that defined these cell types…

Stem Cells

Blood Vessel Cells

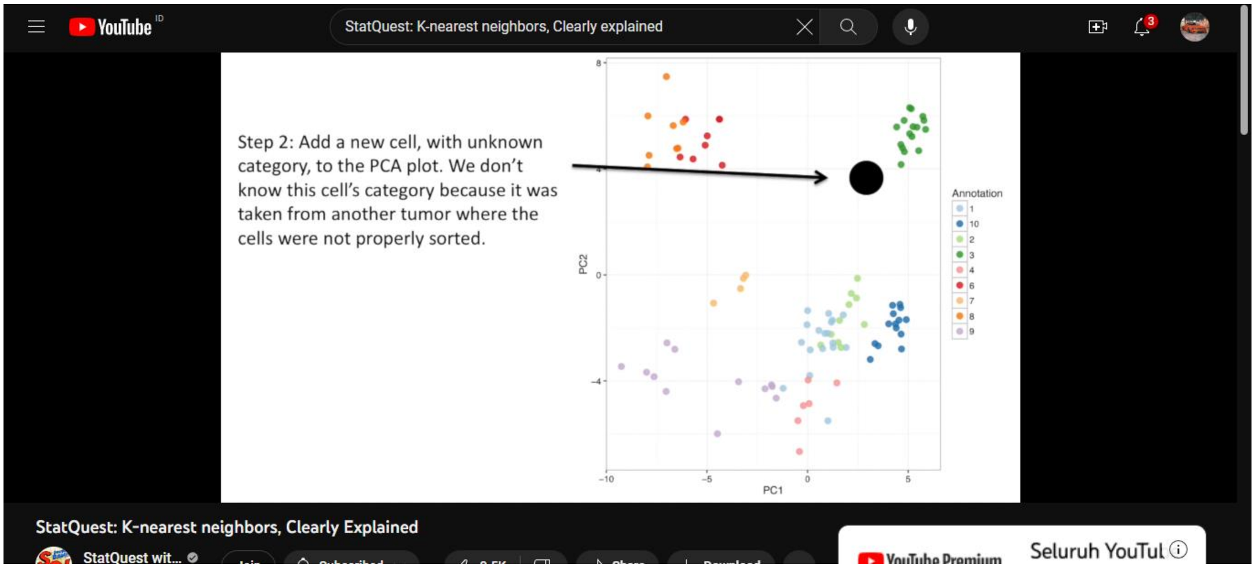
Fat Cells

StatQuest: K-nearest neighbors, Clearly Explained


YouTube ID
StatQuest: K-nearest neighbors, Clearly explained
8
4
0
−4
PC2
−10
−5
0
5
PC1
Annotation
1
10
2
3
4
6
7
8
9
0:58 / 5:30 • K-NN applied to scatterplot data
Seluruh YouTul

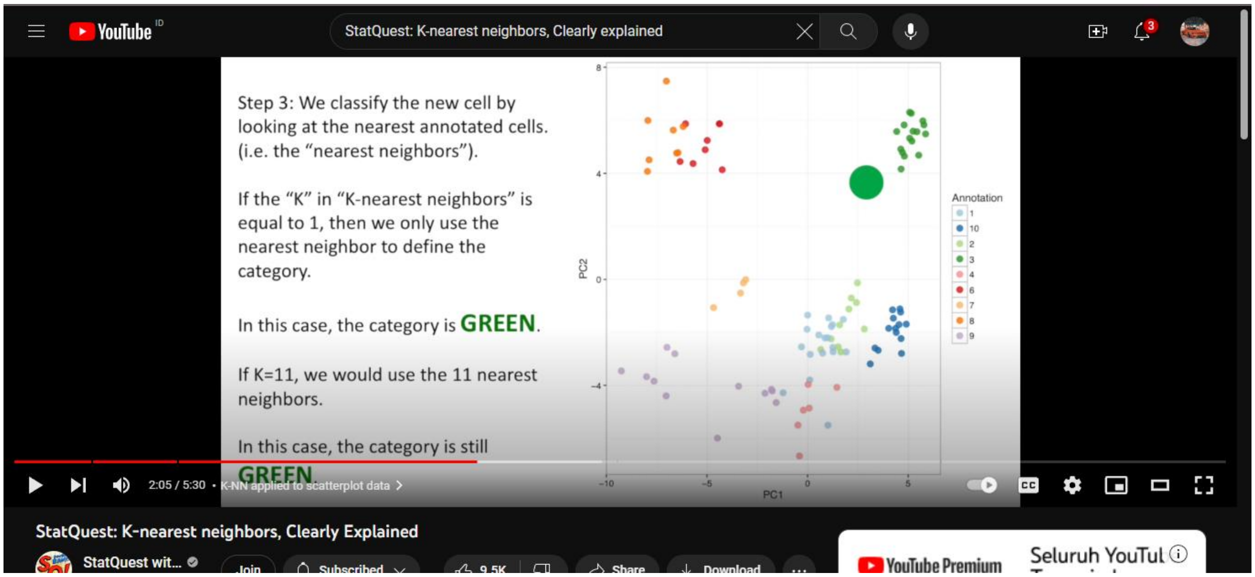
Step 1: Start with a dataset with known categories. In this case, we have different cell types from an intestinal tumor. Then cluster that data. In this case, we used PCA.

StatQuest: K-nearest neighbors, Clearly Explained



Step 2: Add a new cell, with unknown category, to the PCA plot. We don't know this cell's category because it was taken from another tumor where the cells were not properly sorted.

StatQuest: K-nearest neighbors, Clearly Explained



Step 3: We classify the new cell by looking at the nearest annotated cells. (i.e. the "nearest neighbors").

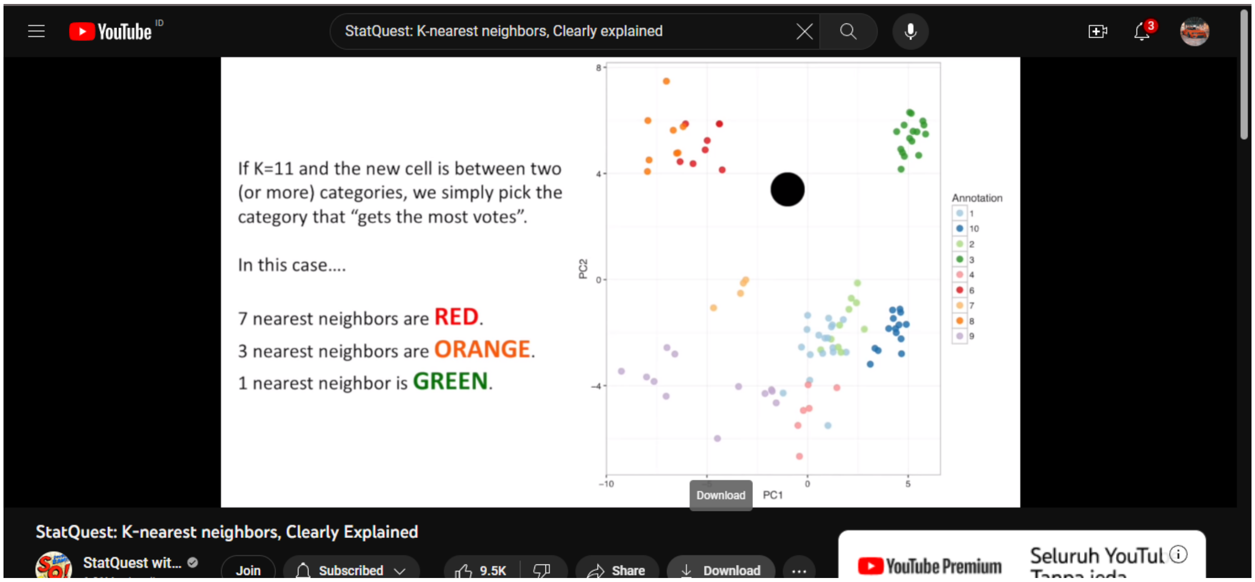
If the "K" in "K-nearest neighbors" is equal to 1, then we only use the nearest neighbor to define the category.

In this case, the category is **GREEN**.

If K=11, we would use the 11 nearest neighbors.

In this case, the category is still **GREEN**.

StatQuest: K-nearest neighbors, Clearly Explained



If K=11 and the new cell is between two (or more) categories, we simply pick the category that "gets the most votes".

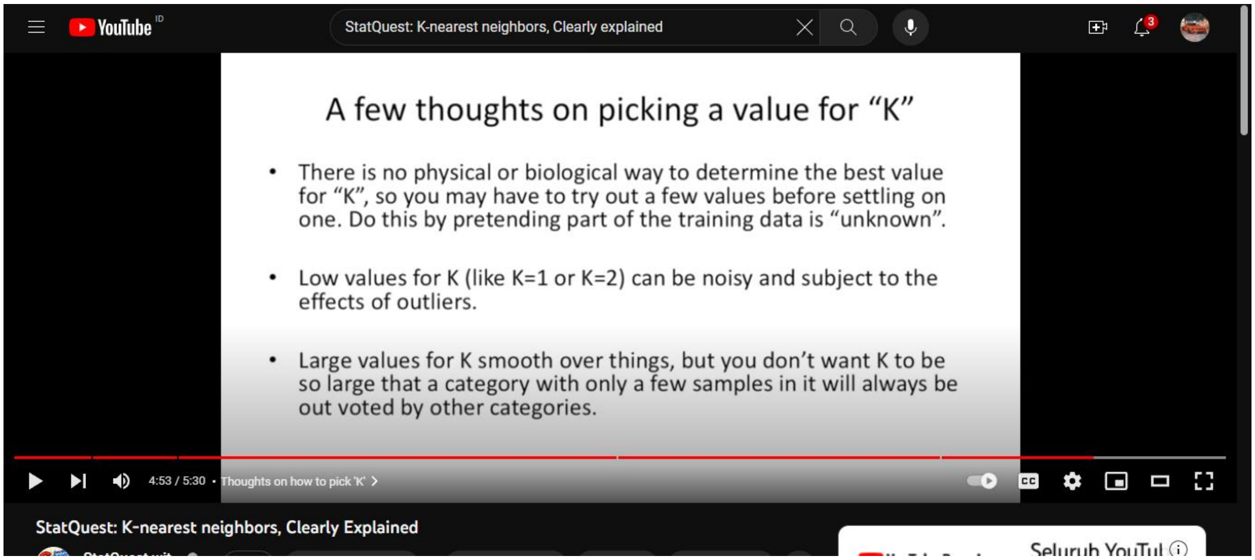
In this case....

7 nearest neighbors are **RED**.

3 nearest neighbors are **ORANGE**.

1 nearest neighbor is **GREEN**.

StatQuest: K-nearest neighbors, Clearly Explained



YouTube ID
StatQuest: K-nearest neighbors, Clearly explained
4:53 / 5:30 • Thoughts on how to pick 'K'
Seluruh YouTul

A few thoughts on picking a value for "K"

- There is no physical or biological way to determine the best value for "K", so you may have to try out a few values before settling on one. Do this by pretending part of the training data is "unknown".
- Low values for K (like K=1 or K=2) can be noisy and subject to the effects of outliers.
- Large values for K smooth over things, but you don't want K to be so large that a category with only a few samples in it will always be out voted by other categories.

StatQuest: K-nearest neighbors, Clearly Explained