

Raken Putra Athallah

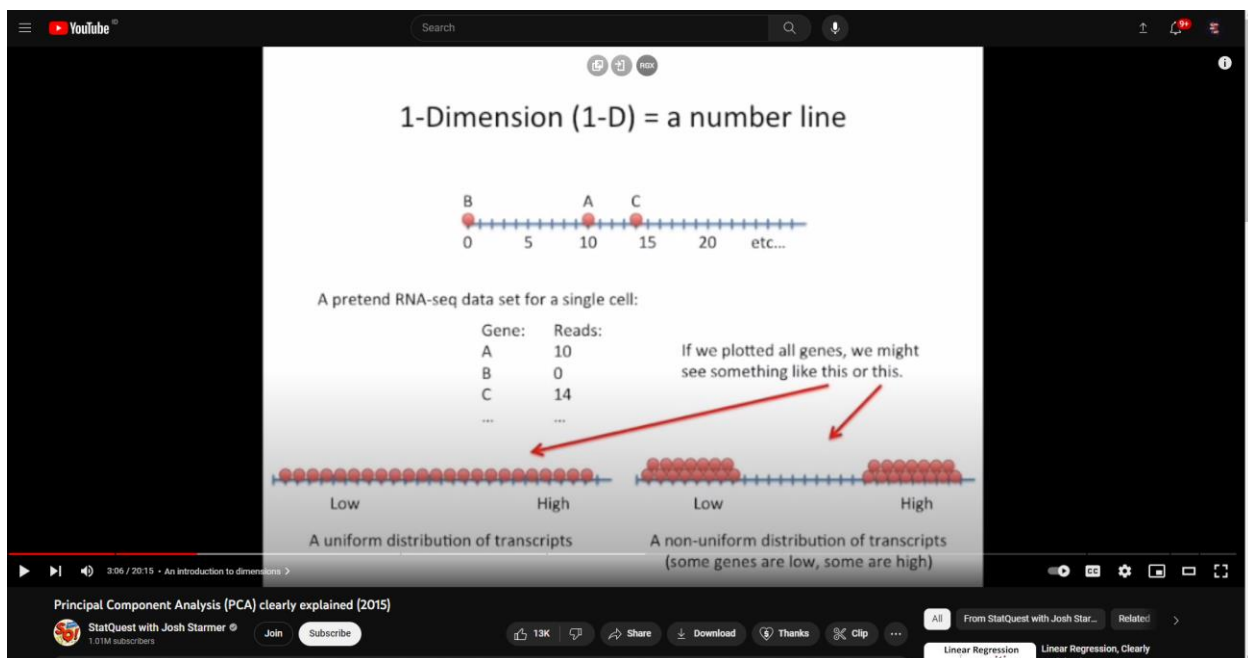
1103204186

Principal component analysis (PCA) adalah teknik reduksi dimensi yang digunakan untuk mengurangi jumlah variabel dalam data set tanpa kehilangan informasi yang signifikan.

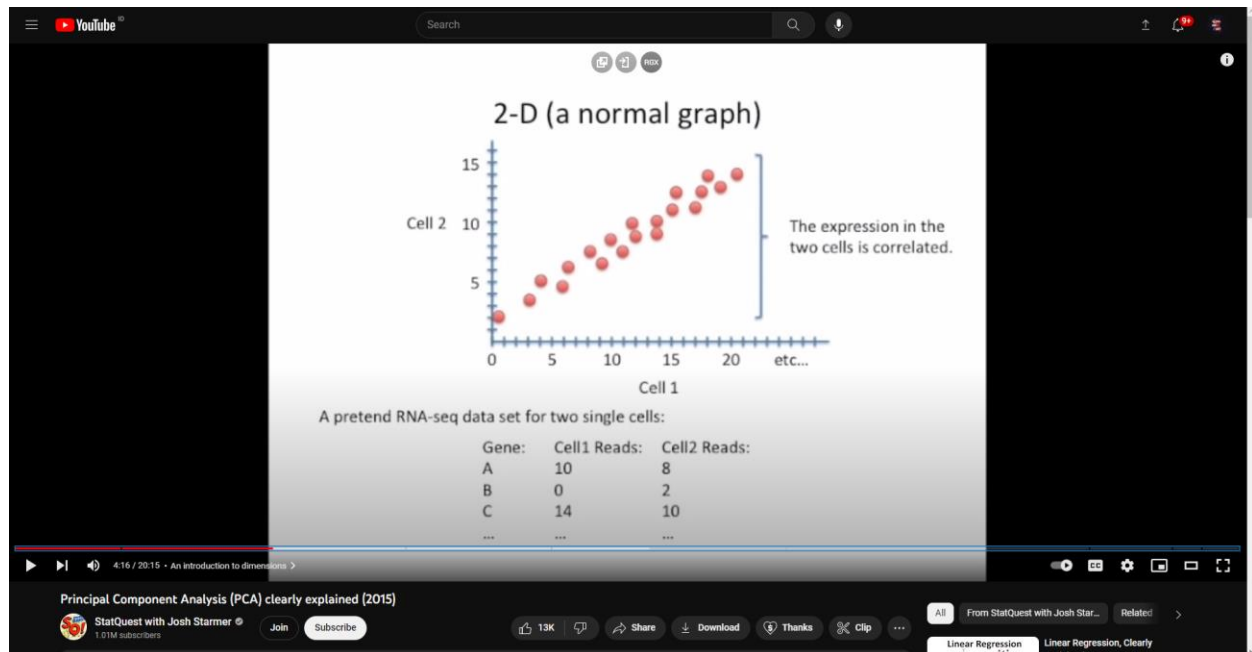
PCA bekerja dengan memproyeksikan data ke dalam ruang baru dengan dimensi yang lebih sedikit.

Dimensi-dimensi ini disebut komponen utama. Komponen utama adalah kombinasi linear dari variabel asli dalam data set.

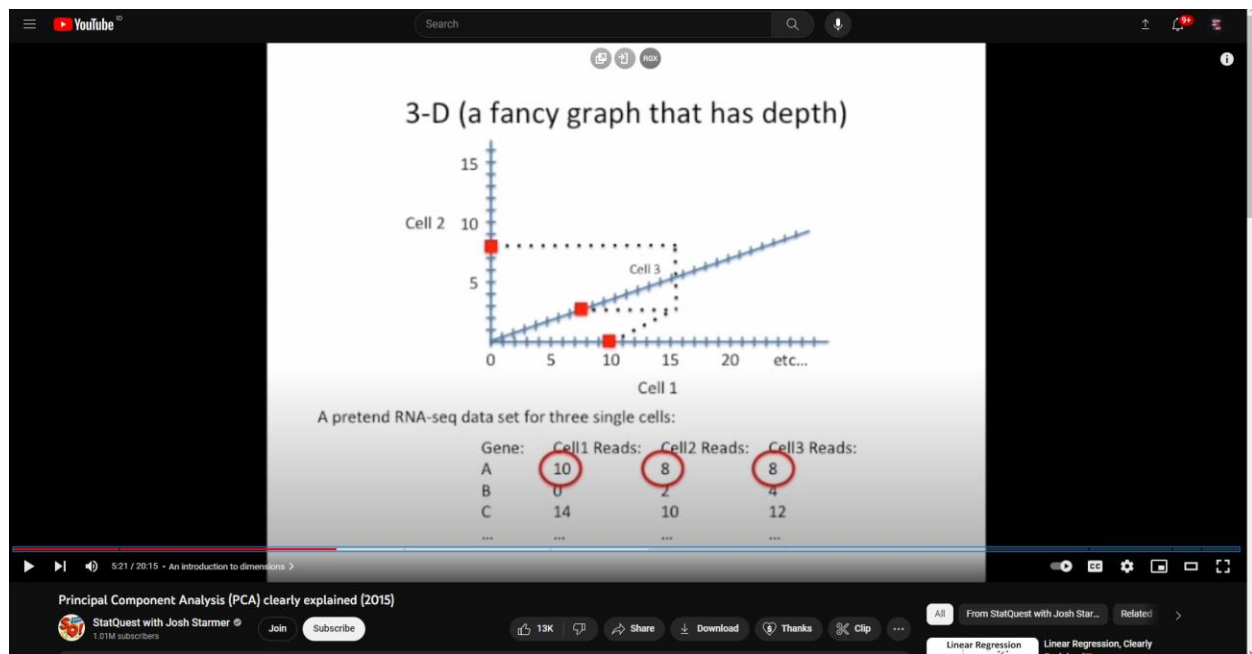
PCA 1 dimensi adalah representasi data yang paling sederhana. Ini hanya mewakili satu variabel, yang merupakan kombinasi linear dari semua variabel asli dalam data set.



PCA 2 dimensi adalah representasi data yang lebih kompleks. Ini mewakili dua variabel, yang merupakan kombinasi linear dari semua variabel asli dalam data set.



PCA 3 dimensi adalah representasi data yang paling kompleks. Ini mewakili tiga variabel, yang merupakan kombinasi linear dari semua variabel asli dalam data set.



What does all of this have to do with PCA?

- PCA takes a dataset with a lot of dimensions (i.e. lots of cells) and flattens it to 2 or 3 dimensions so we can look at it.
  - It tries to find a meaningful way to flatten the data by focusing on the things that are different between cells. (much, much more on this later)
- This is sort of like flattening a Z-stack of microscope images to make a single 2-D image for publication.

Principal Component Analysis (PCA) clearly explained [2015]

StatQuest with Josh Starmer 1.61M subscribers

Join Subscribe

13K 1 Share Download Thanks Clip

All From StatQuest with Josh Starmer Related

Linear Regression Linear Regression, Clearly

Hooray! We know what the X and Y axis are in this figure!!!

The figure is a scatter plot titled 'Hooray! We know what the X and Y axis are in this figure!!!'. The x-axis is labeled 'PC1' and the y-axis is labeled 'PC2'. The plot shows four distinct clusters of data points, each enclosed by a dashed ellipse and labeled with a cell type: 'Blood cells' (top), 'Pluripotent cells' (middle), 'Neural cells' (bottom right), and 'Dermal or epidermal cells' (bottom left). The legend on the right lists the following cell types and their corresponding symbols and colors: K562 (yellow circle), HL60 (red circle), 2339 (purple circle), NPC (green 'x'), GW16 (green circle), GW21 (green 'x'), GW21+3 (green 'x'), Kera (blue circle), BJ (blue 'x'), hiPSC (black circle), and 2338 (light blue circle). The plot is labeled 'C' in the top left corner.

Tergantung pada banyak dimensi yang digunakan setiap dimensi akan mewakili 1 principal komponen berdasarkan pada axisnya masing2 masing,seperti contoh pada gambar pc1 mewakili most variation pada data dari kiri ke kana dan pc2 mewakili most 2<sup>nd</sup> variation dari atas sampai bawah

## 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
...	...	...

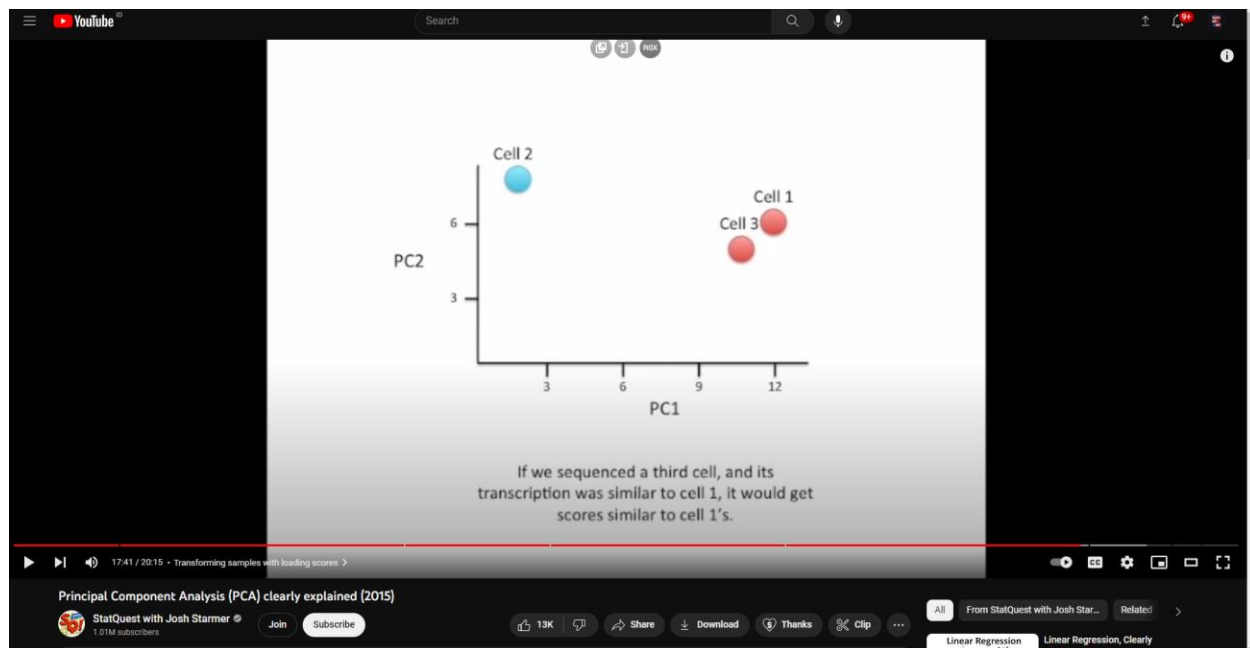
PC2

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 =  $(10 * 10) + (0 * 0.5) + \dots \text{etc} \dots = 12$

Cell1 PC2 score =  $(10 * 3) + (0 * 10) + \dots \text{etc} \dots = 6$

Untuk membuat polt cells dengan principal component dapat melakukannya dengan menentukan score atau seberapa mempengaruhi suatu gene pada suatu principal components(pc) pada contoh pc1 di dalam video genes yang paling berpengaruh adalah genes pada bagian paling kiri atau kanan yang disebut dengan extreme genes(?). Setelah menentukan dan tingkat influence suatu genes dan memberikannya nilai atau dalam bentuk angka kita dapat menghitung seluruh principal score untuk dibuat menjadi plot cells

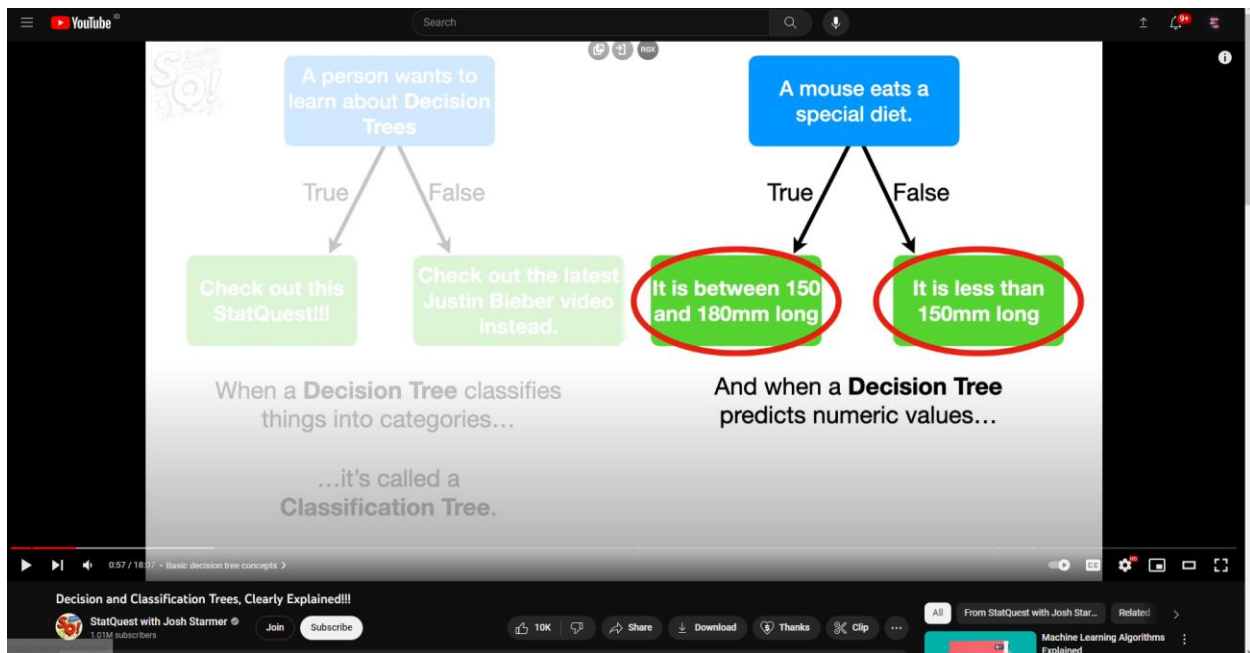


## KNN

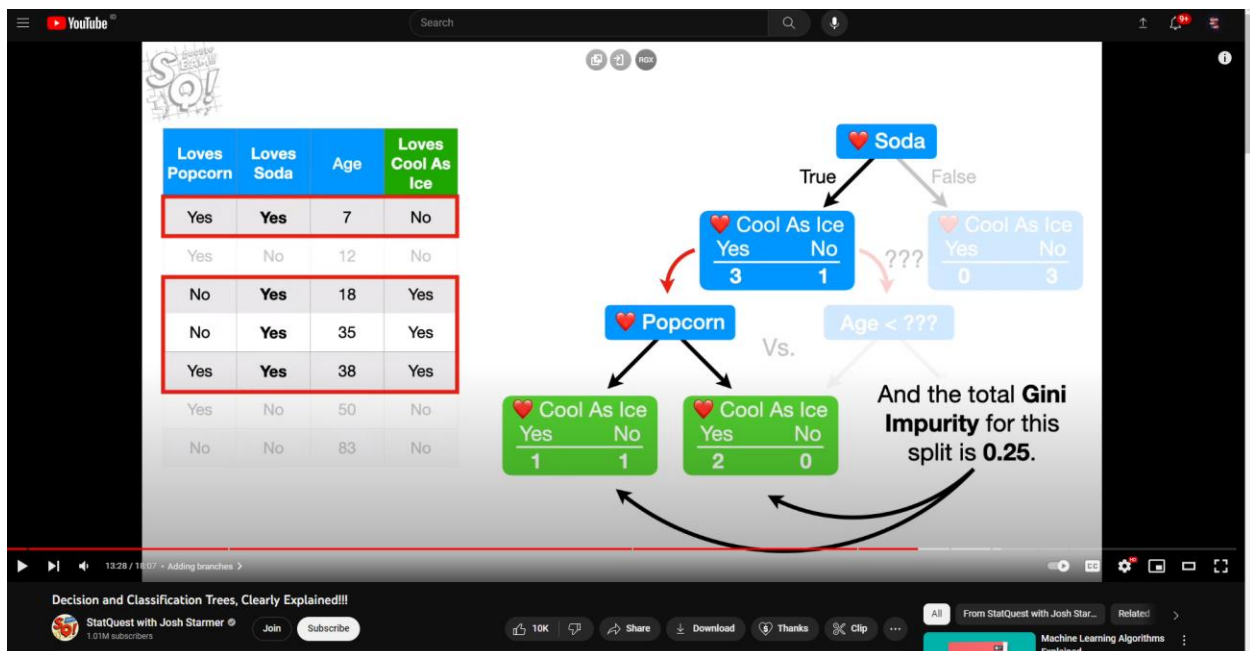


Pada contoh ini dia menjelaskan bahwa kita dapat menentukan suatu data jika kita sudah mempunyai banyak data yang memenuhi aspek-aspek yang diinginkan. Pada contoh ini dengan menggunakan scatterplot sudah ada cells yang sudah diketahui dan ada 1 cell yang belum diketahui, dia menjelaskan kita dapat mengklasifikasikan cell tersebut dengan cell terdekat yang sudah terklasifikasi jika termasuk lebih dari 1 cell yang terklasifikasi kita dapat memilih dengan mencari cell terdekat paling banyak. Prinsip yang sama juga dapat diterapkan jika kita menggunakan heatmap.





Pada dasarnya Decision tree adalah seperti memilih pilihan berdasarkan suatu statement berdasarkan pada jawabannya benar atau salah. Jika statement atau decision decision tree menggunakan kategori maka termasuk classification tree sedangkan jika menggunakan suatu prediksi maka termasuk regression tree



Untuk membuat decision tree kita haru menentukan apa yang menjadi statement pertama dengan mengambil statement yang memiliki gini Impurity paling sedikit. Impure sendiri adalah ketika suatu statement mendapatkan hasil yang cukup tidak konsisten atau mendapatkan jawaban yang cukup

acak. Gini impurity juga dapat digunakan untuk membuat branch dengan metode yang hampir mirip dengan metode untuk menentukan statement pertama, kita dapat melakukannya hingga mendapatkan prediksi dari jawaban yang kita inginkan