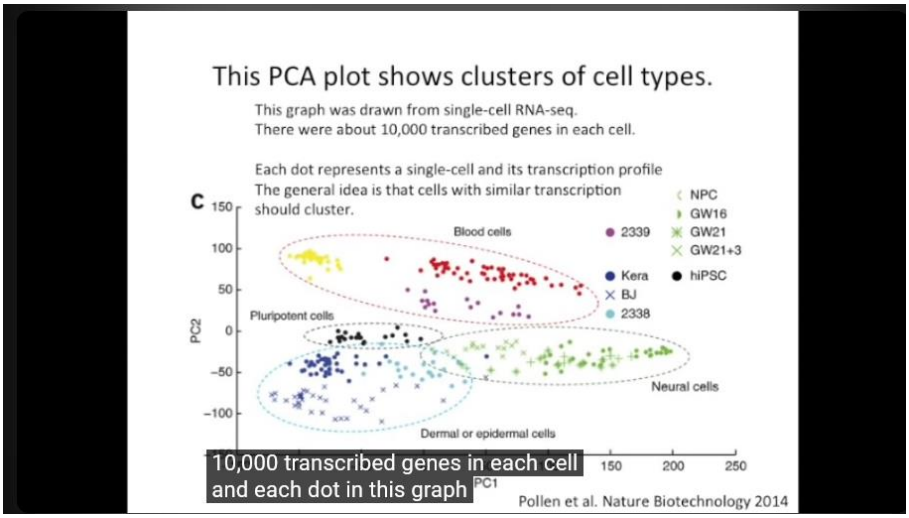
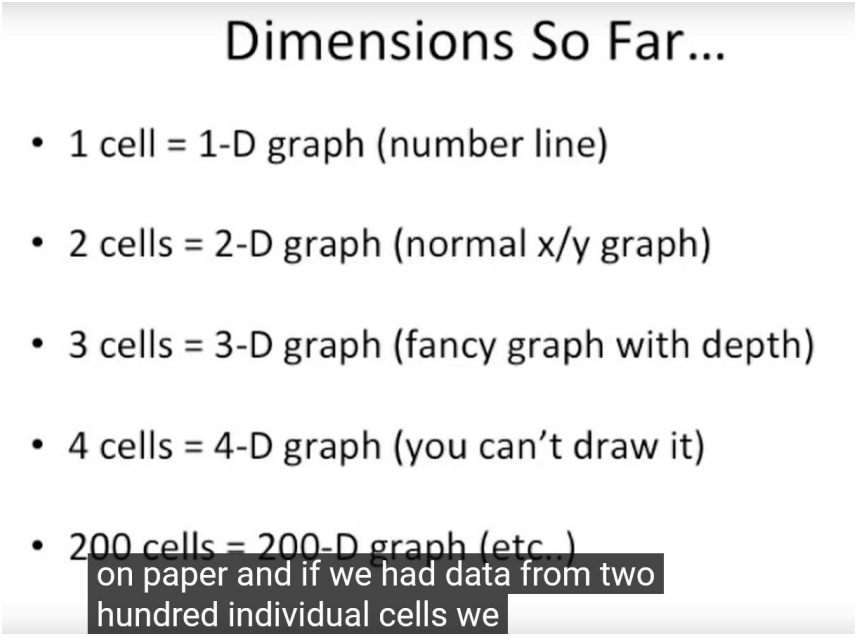
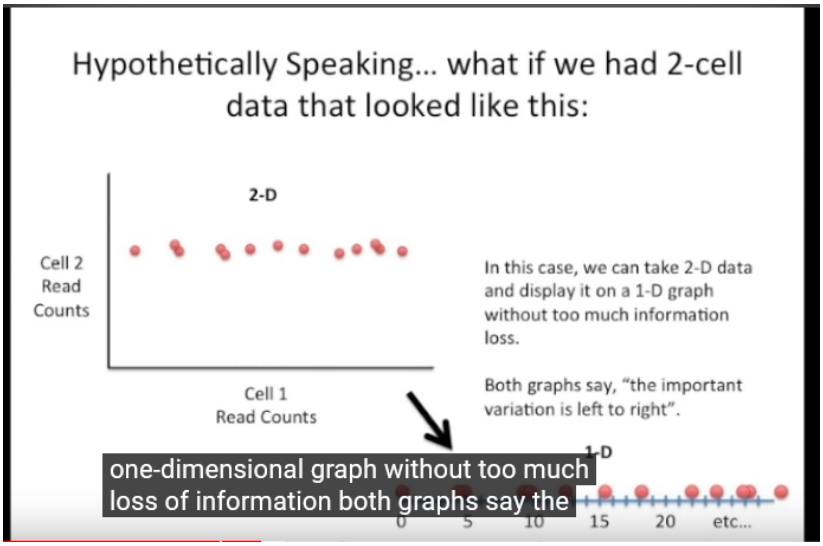
1. Principal Component Analysis (PCA) clearly explained (2015)



Pada video ini, Grafik diambil dari RNA-sel Tunggal. Ada sekitar 10.000 gen yang ditranskripsi di setiap sel. Setiap titik mewakili satu sel dan profil transkipnya. Ide umumnya adalah bahwa sel-sel dengan transkip serupa harus berkelompok. Metode PCA berfungsi untuk mengompresi banyak data menjadi sesuatu yang menangkap esensi data asli.



Dalam beberapa dimensi PCA adalah sebuah metode statistik dan matematika yang digunakan untuk mengurangi dimensi (fitur) dalam dataset multidimensi dengan tujuan untuk mengurangi kompleksitas data sambil mempertahankan informasi yang paling penting.

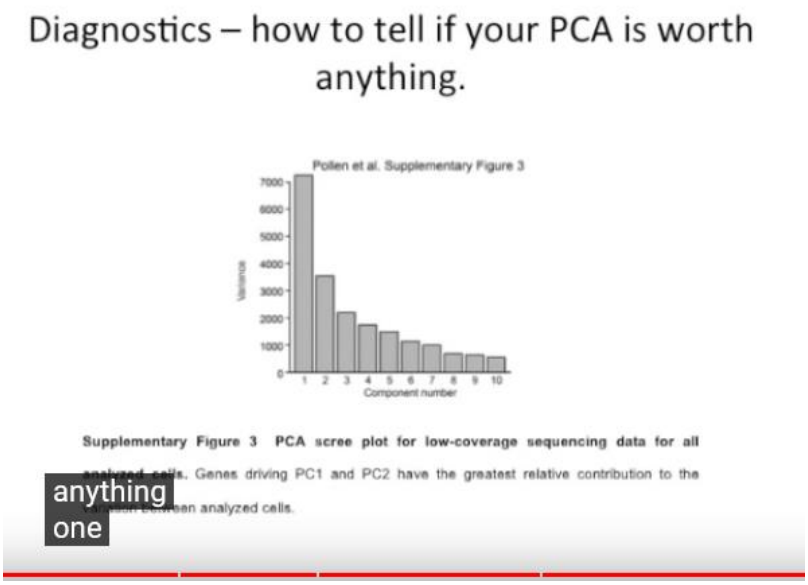


Gambar diatas memiliki satu contoh kasus yaitu TV dan Film. Tv dan film menghasilkan 2d, meskipun subjek 3d. Pembuatan film selalu direkam menggunakan kamera, dari objek 3-d yang tertangkap kamera akan direkam, rekaman tersebut akan menghasilkan video yang menjadikan objek tersebut berupa 2d.

Di bawah ini adalah contoh data tabel yang menggambarkan bagaimana PCA bisa diterapkan pada data ekspresi gen dengan dua sel (cell 1 dan cell 2) dan sejumlah gen sebagai fitur. Data ekspresi gen ini digunakan untuk menjelaskan bagaimana PCA dapat digunakan untuk mengurangi dimensi data genetik:

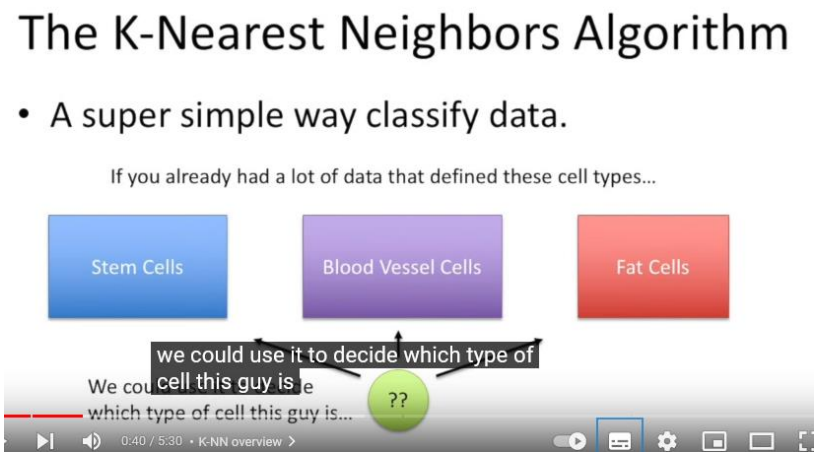
Gene	Cell 1 Reads	Cell 2 Reads
Gene A	100	120
Gene B	80	95
Gene C	150	140
Gene D	75	90
...	...	...

Dalam tabel ini, setiap baris mewakili gen yang berbeda, dan setiap kolom (Cell 1 Reads dan Cell 2 Reads) berisi data ekspresi gen untuk dua sel yang berbeda (cell 1 dan cell 2). Data ini biasanya berupa jumlah pembacaan (reads) yang mewakili tingkat ekspresi relatif gen tersebut dalam masing-masing sel.



Gambar diatas merupakan 3 plot PCA untuk data pengurutan cakupan rendah untuk semua panggilan yang dianalisis. Gen yang menggunakan PC1 dan PC2 memiliki kontribusi relative terbesar terhadap variasi antar sel yang didiagnosis.

2. StatQuest: K-nearest neighbors, Clearly Explained

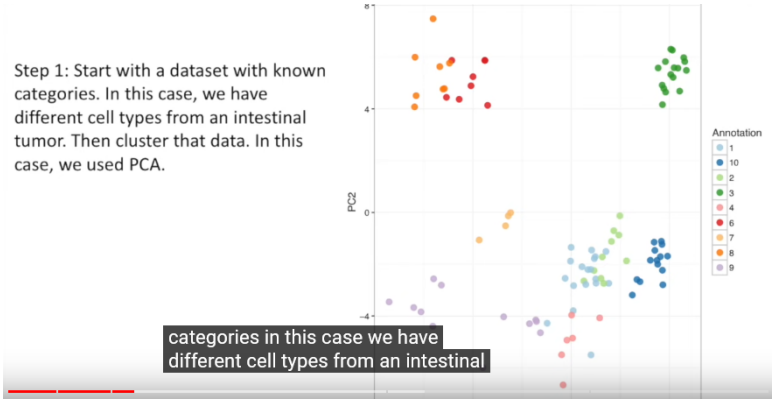


Dalam K-Nearest Neighbors Algorithm terdapat 3 type yaitu stem cells, blood vessel cells, fat cells yang digunakan untuk klasifikasi.

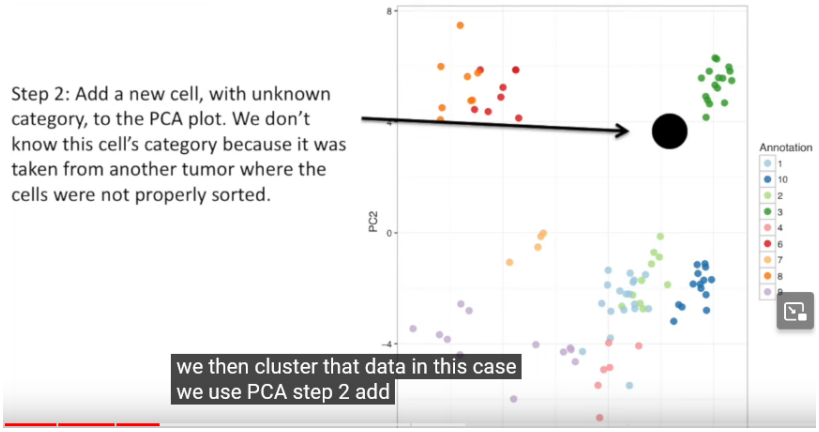
Stem cells adalah adalah salah satu kategori atau kelas yang mungkin digunakan dalam masalah klasifikasi. Masing-masing data atau sampel dapat diklasifikasikan sebagai "stem cells" jika memiliki atribut atau fitur yang sesuai.

Blood Cells adalah kategori atau kelas lain yang mungkin digunakan dalam masalah klasifikasi. Data atau sampel yang memiliki atribut yang sesuai akan diklasifikasikan sebagai "blood vessel."

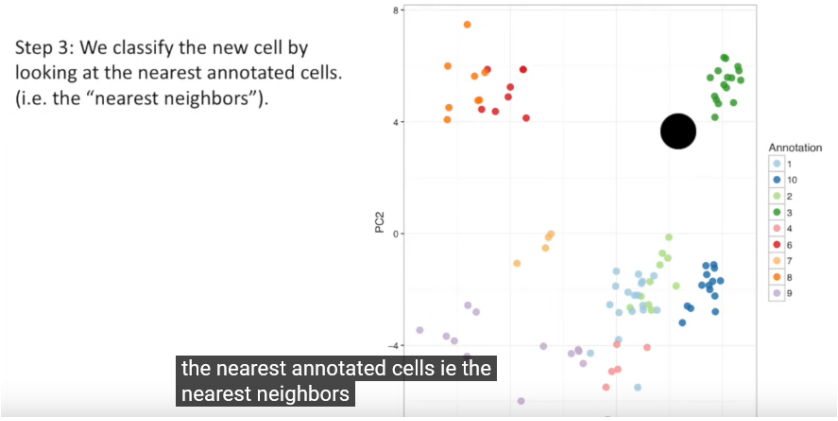
Fat cells adalah kategori atau kelas yang ketiga yang mungkin digunakan dalam masalah klasifikasi. Data atau sampel yang memiliki atribut yang sesuai akan diklasifikasikan sebagai "fat cells."



Pada video diatas dimulai dengan mengumpulkan data berdasarkan kategori. Dalam hal tersebut harus memiliki sell yang berbeda, kemudian data tersebut dikelompokan menggunakan PCA.



Tambahkan sell yang baru dengan kategori yang belum diketahui. Kategori tersebut belum diketahui karena diambil dari tumor yang lain, dimana sell tersebut belum terurut.



Kemudian mengklasifikasikan sell baru dengan melihat anotasi sell yang paling terdekat. Jika K dalam KNN setara dengan 1, maka bisa menggunakan pendekatan dengan tetangga yang terdekat untuk mendifinisikan kategori. Titik hitam dekat dengan titik hijau, maka masuk kedalam kategori hijau.

Beberapa pemikiran tentang memilih nilai untuk "K"

Tidak ada cara fisik atau biologis untuk menentukan nilai terbaik untuk "K", jadi Anda mungkin harus mencoba beberapa nilai sebelum menentukan salah satunya. Lakukan ini dengan menganggap sebagian data pelatihan "tidak diketahui".

Nilai K yang rendah (seperti K=1 atau K=2) dapat menimbulkan gangguan dan terkena efek outlier.

Nilai K yang besar akan memuluskan segalanya, namun Anda tentu tidak ingin K menjadi terlalu besar sehingga kategori yang hanya memiliki sedikit sampel di dalamnya akan selalu dikeluarkan dari kategori lain.