UNIT 3

DATA PRE-PROCESSING - Part 1/4

DIMENSIONALITY REDUCTION - The What and Why?

Dimensionality Reduction!

IRIS dataset

Iris Versicolor

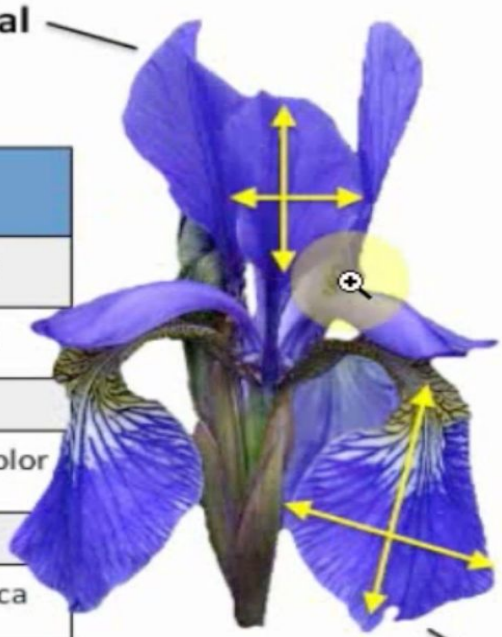
Iris Setosa

Iris Virginica

What is Dimensionality??

Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
			...		
50	6.4	3.5	4.5	1.2	Versicolor
			...		
150	5.9	3.0	5.0	1.8	Virginica



Here, Dimension = 4 not 5

Features
(attributes, measurements, dimensions)

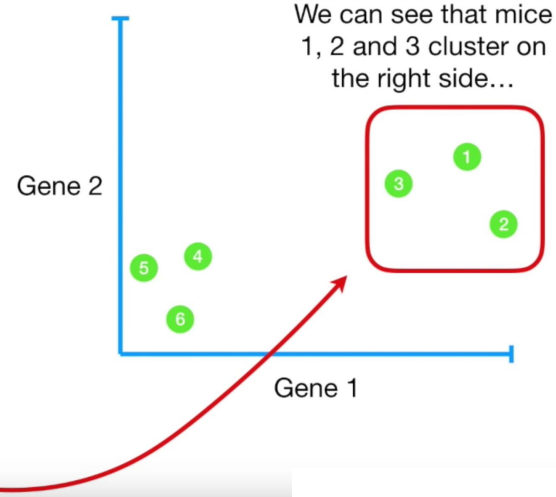
Class labels
(targets)

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

Even though it's a simple graph, it shows us that mice 1, 2 and 3 are more similar to each other than they are to mice 4, 5 6.

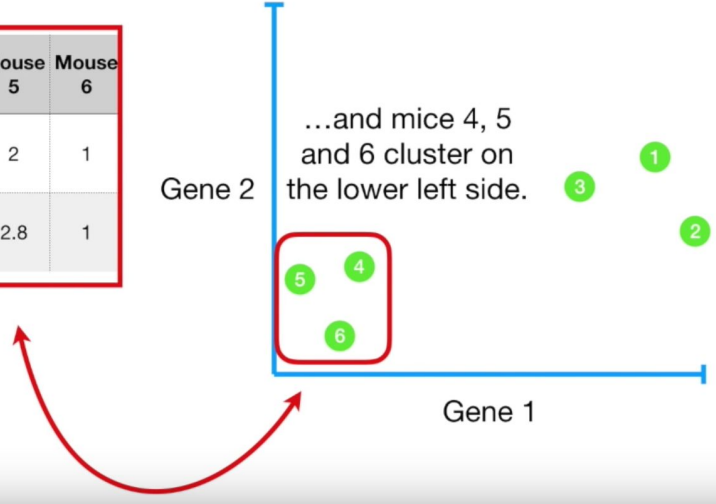
Data Values Plotted in 1 Dimension

	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

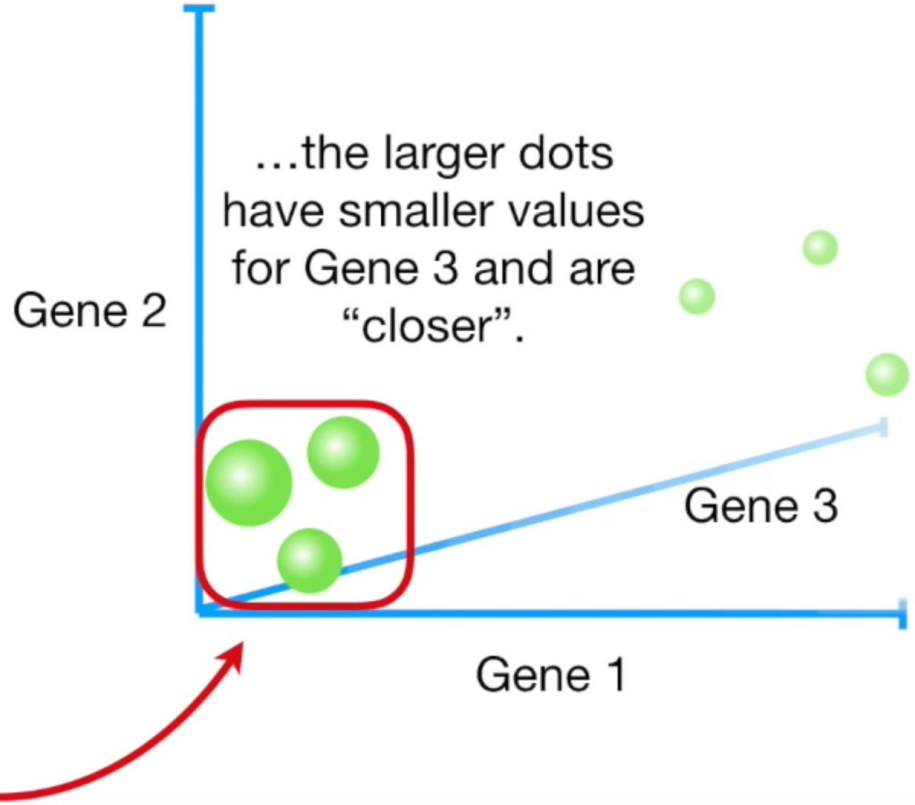


Data Values Plotted in 2 Dimension

	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



	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
Gene 3	12	9	10	2.5	1.3	2



Data Values Plotted in 3 Dimensions

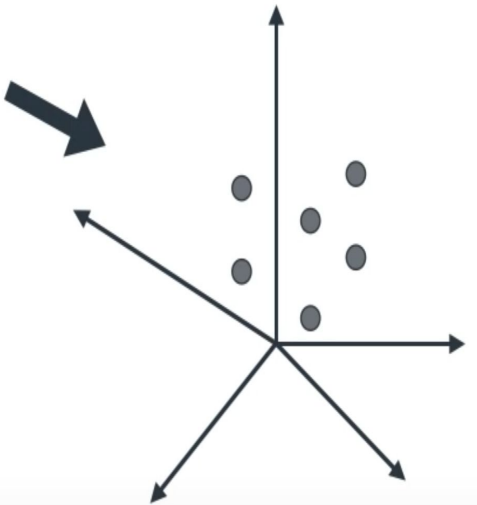
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2
Gene 4	5	7	6	2	4	7

If we measured 4 genes, however, we can no longer plot the data - 4 genes require 4 dimensions.

Beyond 3D Human Capability to do Data Visualization is NIL!!!

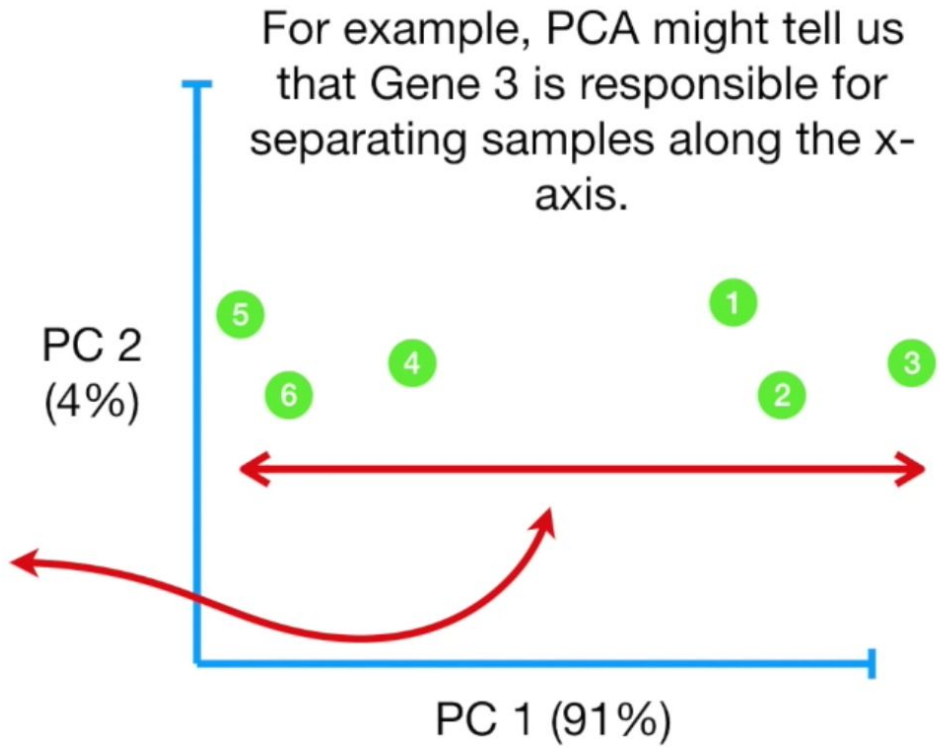
Large Table

X1	X2	X3	X4	X5
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*



5D Plot

	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
Gene 3	12	9	10	2.5	1.3	2
Gene 4	5	7	6	2	4	7

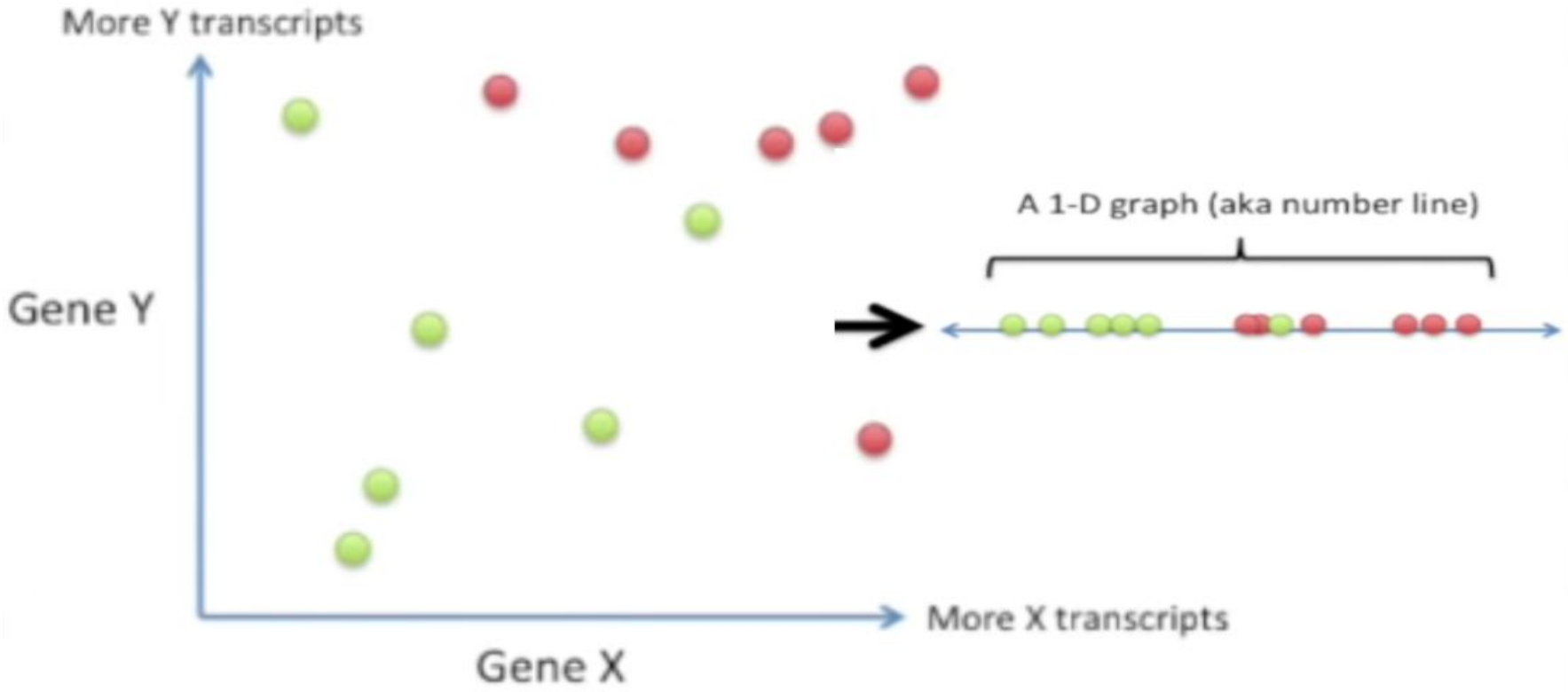


Need for PCA?

2. Which Variable is most useful for clustering the data correctly?

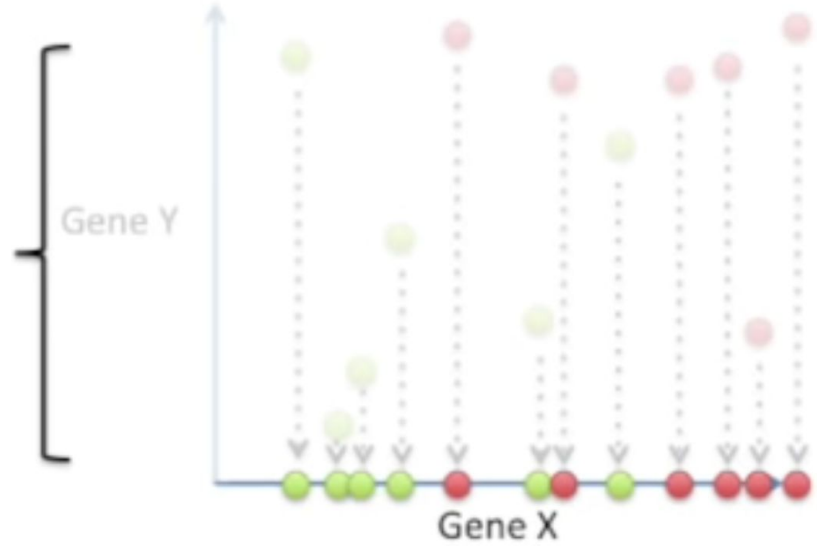
3. PCA can help us determine how accurate the 2D graph is.

Reducing a 2-D graph to a 1-D graph



Reducing a 2-D graph to a 1-D graph

One bad option would be to ignore Gene Y



This way is bad because it ignores the useful information that Gene Y provides...

Projecting the genes onto the Y-axis (i.e. ignoring Gene X) isn't any better

Solutions?

1. **PCA** - Principal Component Analysis (Un-Supervised Technique)
2. **LDA** - Linear Discriminant Analysis (Supervised Technique)

NOTE: Neither PCA nor PDA are “Machine Learning Algorithms”.They are used much before ML Algorithms comes into effect - in the Data Pre-Processing Stage. Both are used for Dimensionality Reduction.

SOURCES:

1. **Stat Quest:** https://www.youtube.com/watch?v=FgakZw6K1QQ
2. **IRIS Dataset:** https://www.youtube.com/watch?v=S8YSqrzqERI