PRINCIPAL COMPONENT ANALYSIS

Partha Sarathi Kar
IVSM 166777

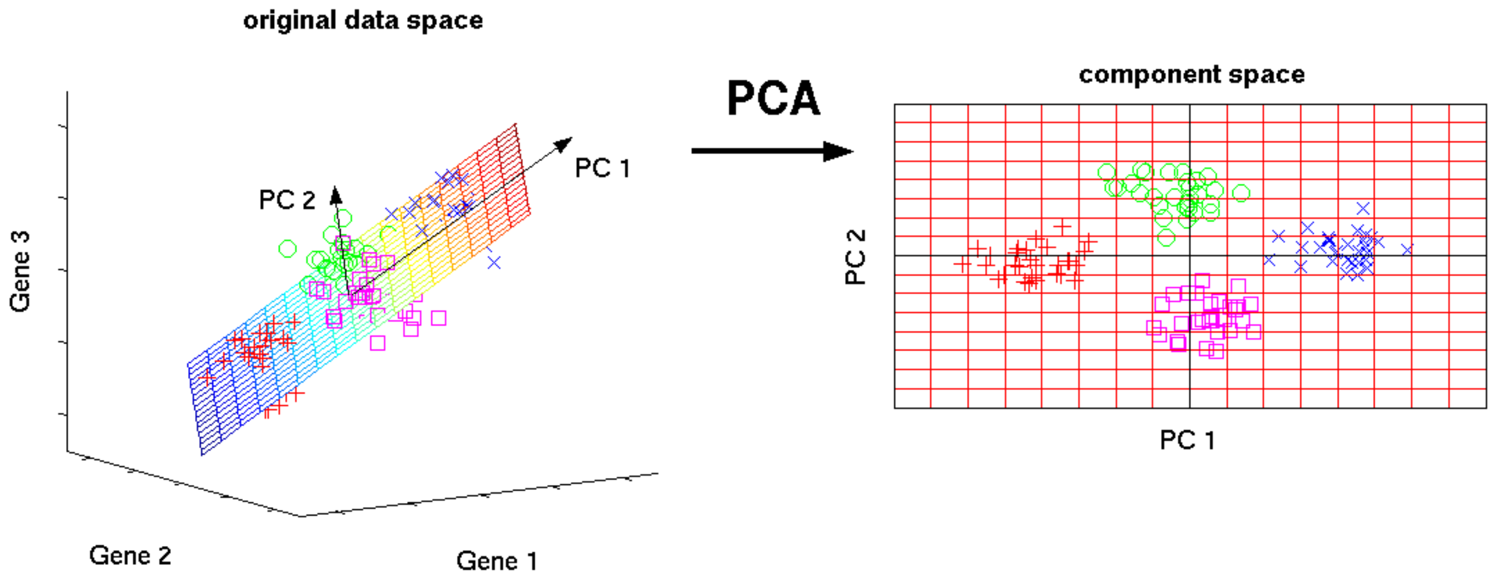
CONTENTS

WHAT IS PCA

Principal component analysis (PCA) is a technique used to **emphasize variation** and **bring out strong patterns** in a dataset.

It's often used to make data easy to explore and visualize.

PCA takes a dataset with a lots of dimension (i.e. Lots of Cells) and flattens it to **2** or **3** dimensions so we can look on it.

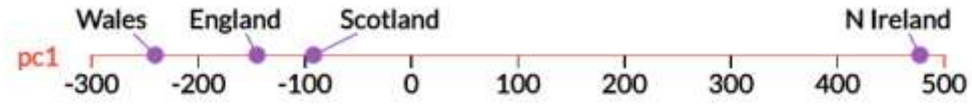


HOW IT WORKS

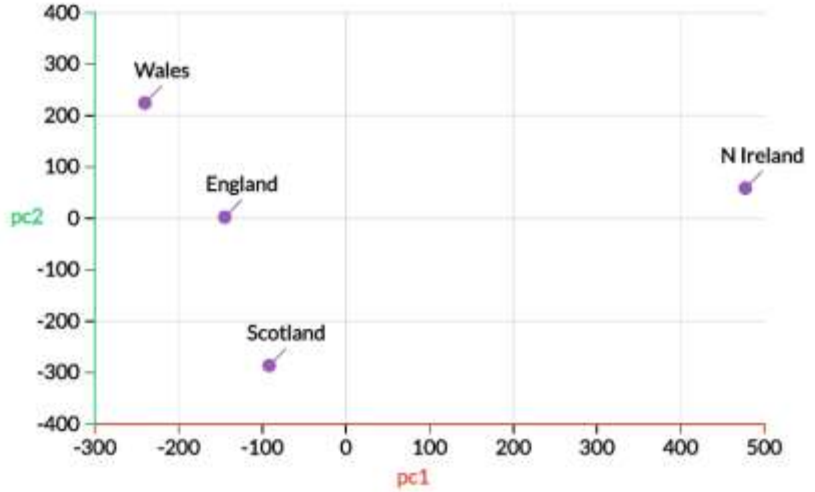
Eating in the UK (a 17D example)

	England	N Ireland	Scotland	Wales
Alcoholic drinks	375	135	458	475
Beverages	57	47	53	73
Carcase meat	245	267	242	227
Cereals	1472	1494	1462	1582
Cheese	105	66	103	103
Confectionery	54	41	62	64
Fats and oils	193	209	184	235
Fish	147	93	122	160
Fresh fruit	1102	674	957	1137
Fresh potatoes	720	1033	566	874
Fresh Veg	253	143	171	265
Other meat	685	586	750	803
Other Veg	488	355	418	570
Processed potatoes	198	187	220	203
Processed Veg	360	334	337	365
Soft drinks	1374	1506	1572	1256
Sugars	156	139	147	175

Here's the plot of the data along the first principal component. Already we can see something is different about Northern Ir

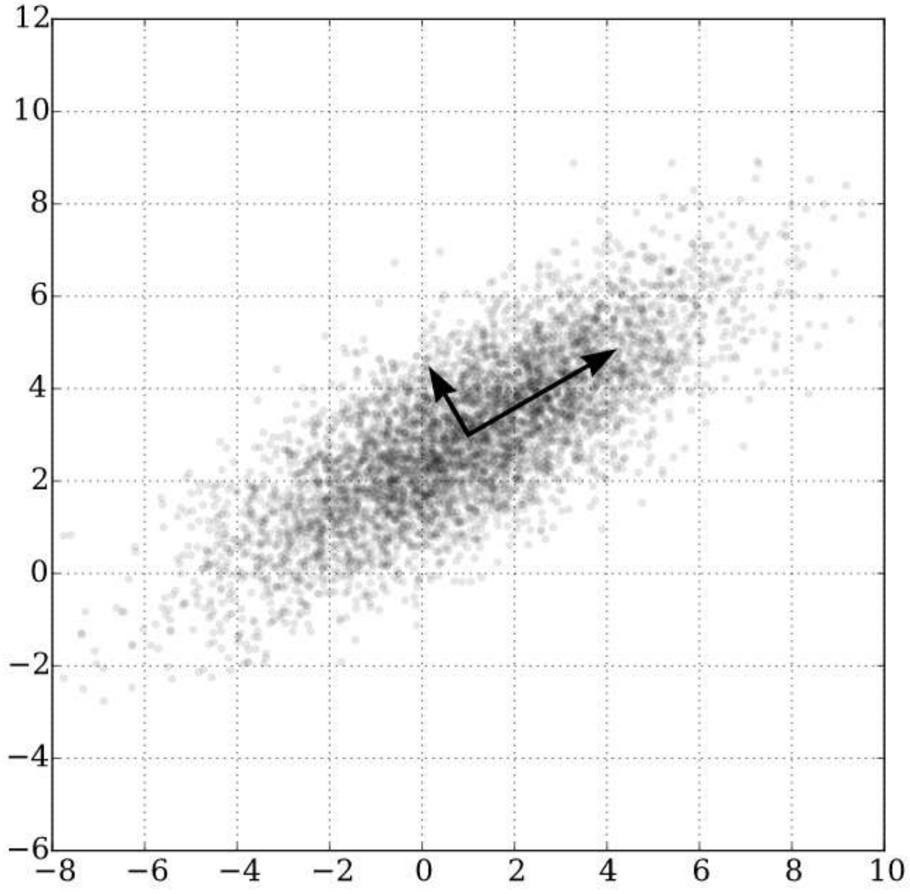


Northern Irish eat way more grams of fresh potatoes and way fewer of fresh fruits, cheese, fish and alcoholic drinks



HISTORY OF PCA

- PCA was invented in 1901 by *Karl Pearson*
- as an analogue of the *principal axis theorem* in mechanics



src: https://commons.wikimedia.org/wiki/File:GaussianScatterPCA.svg

HISTORY OF PCA

Depending on the field of application, it is also named:

- **discrete Kosambi-Karhunen–Loève transform** (KLT) in signal processing,
- the **Hotelling transform** in multivariate quality control,
- **proper orthogonal decomposition** (POD) in mechanical engineering,
- **singular value decomposition** (SVD) of X (Golub and Van Loan, 1983),
- **eigenvalue decomposition** (EVD) of XTX in linear algebra,
- **Eckart–Young theorem** (Harman, 1960), or **Schmidt–Mirsky theorem** in psychometrics,
- **empirical orthogonal functions** (EOF) in meteorological science,
- **empirical eigenfunction decomposition** (Sirovich, 1987) etc

PCA IMPLEMENTATION

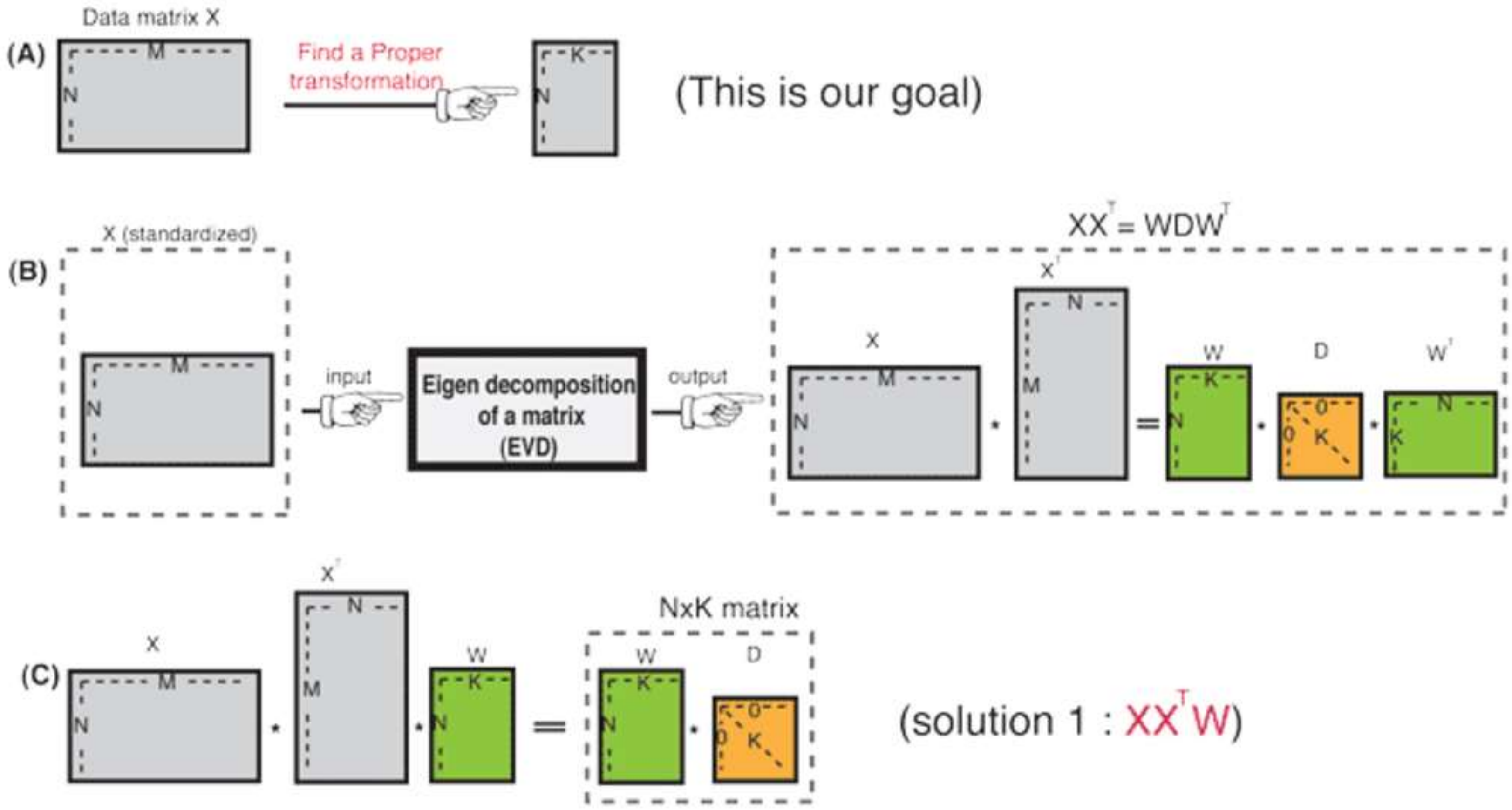
PCA could have different implementations.

But most popular ones are

- **eigenvalue decomposition** (EVD) and
- **singular value decomposition** (SVD).

PCA IMPLEMENTATION

Eigenvalue decomposition



NxM > NxK (K<=M)
X = original data matrix

W and D new Matrix from

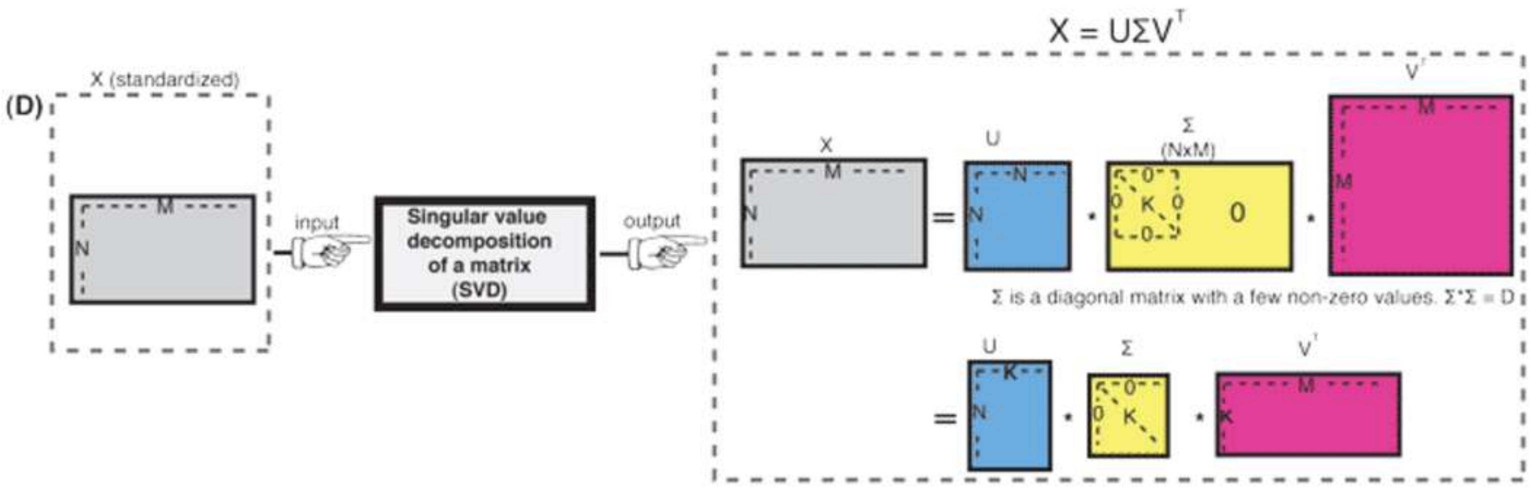
W contains all principal component vectors, while **D** contains all ranks of those vectors (ordered from the largest variance to the least one

X^T and W^T are transposes of X and W

XXTW=WD

PCA IMPLEMENTATION

Singular Value decomposition

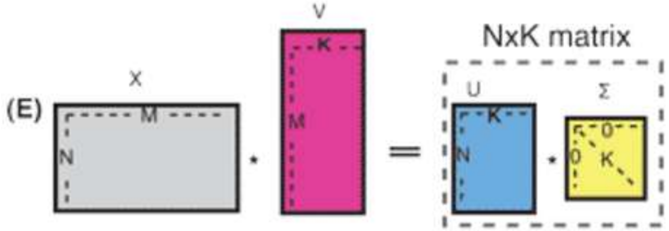


NxM > NxK (K<=M)

X = original data matrix

X=UΣVT

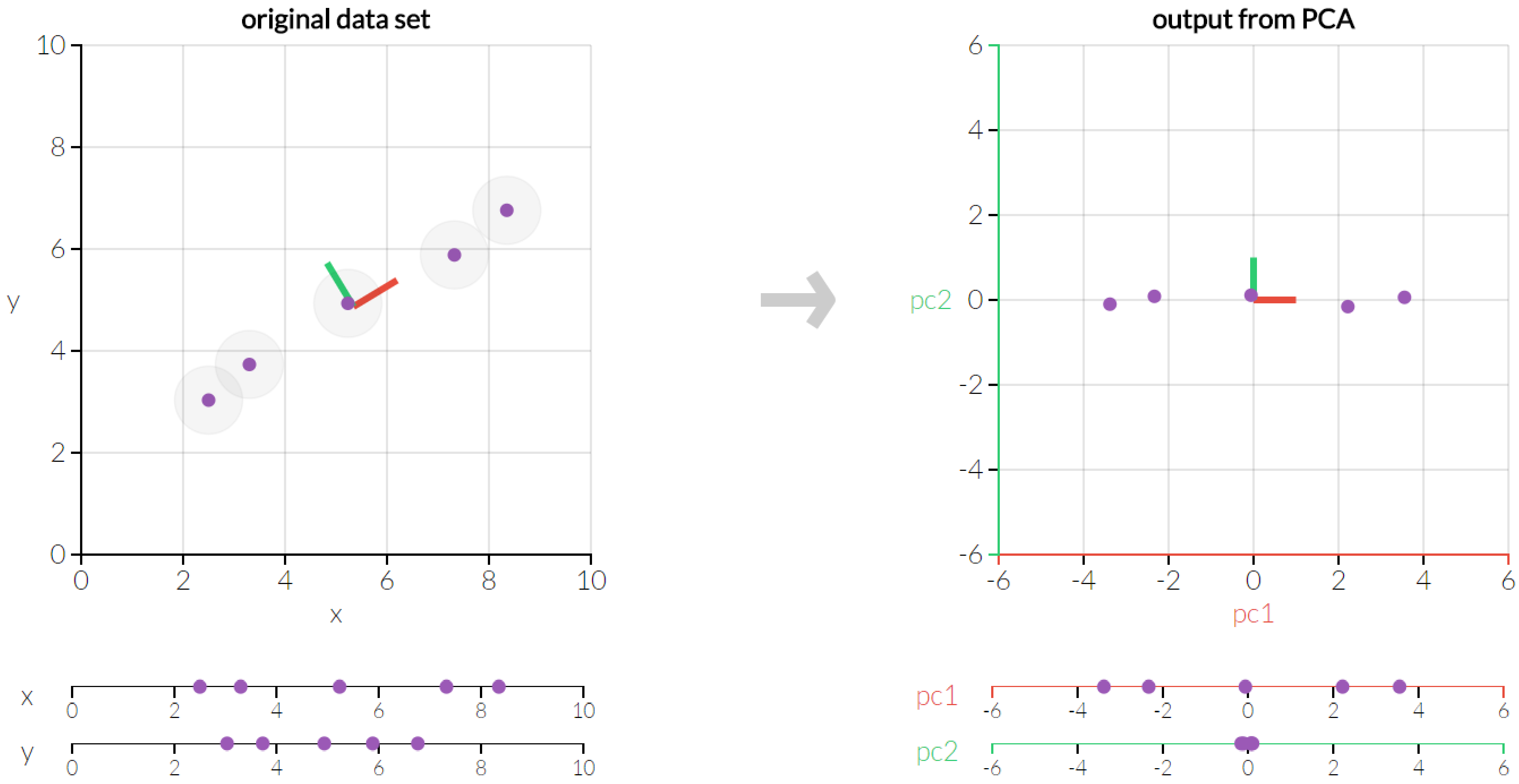
XV=UΣ



(solution 2 : XV)

Three new matrix U, Σ and V^T. U and V^T contain principal component vectors for two directions (column and row of raw data) accordingly. Σ contains ordered ranks of those principal components.

PCA IMPLEMENTATION



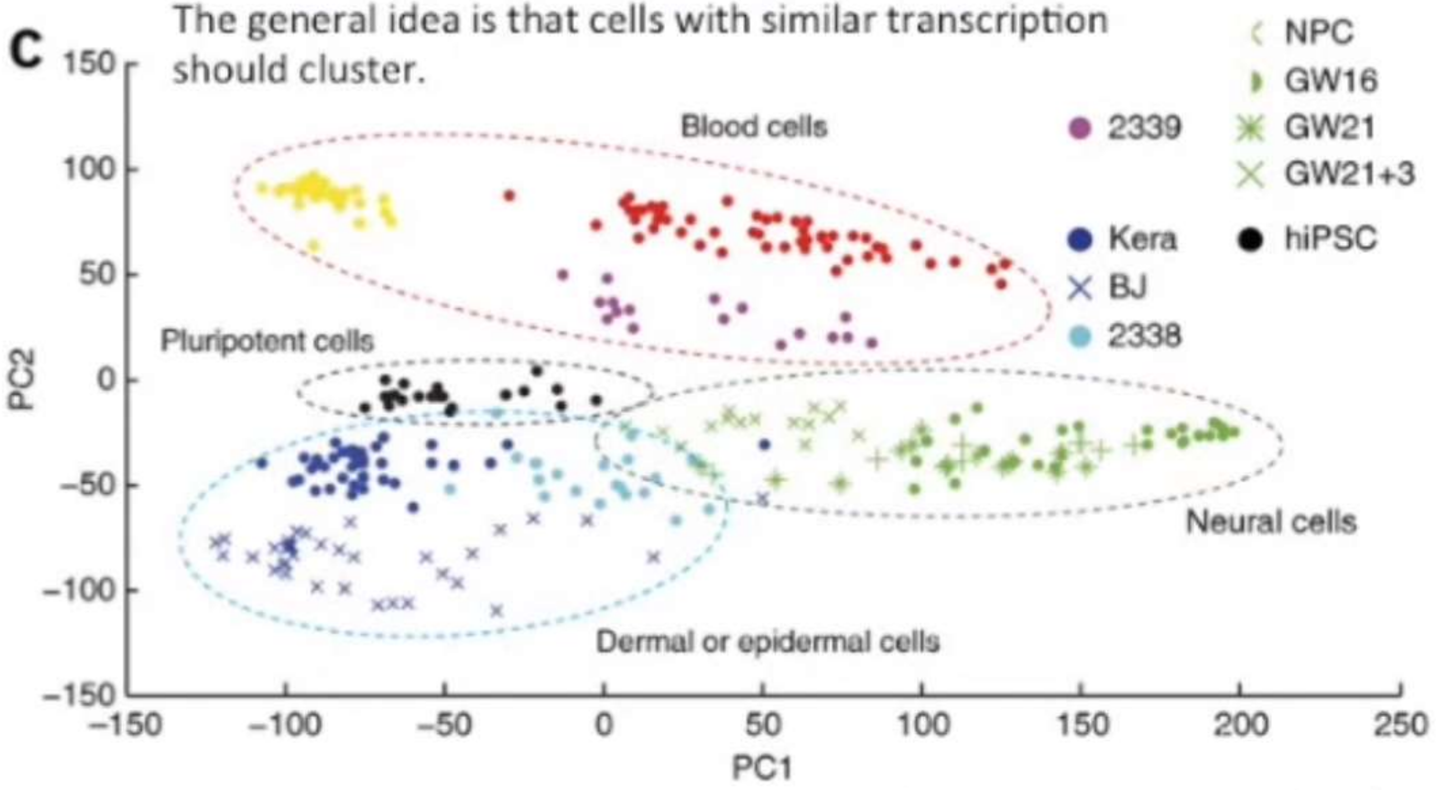
PCA IMPLEMENTATION

EXAMPLE

This PCA plot shows clusters of cell types.

This graph was drawn from single-cell RNA-seq.
There were about 10,000 transcribed genes in each cell.

Each dot represents a single-cell and its transcription profile
The general idea is that cells with similar transcription should cluster.



Pollen et al. Nature Biotechnology 2014

PCA IMPLEMENTATION

EXAMPLE

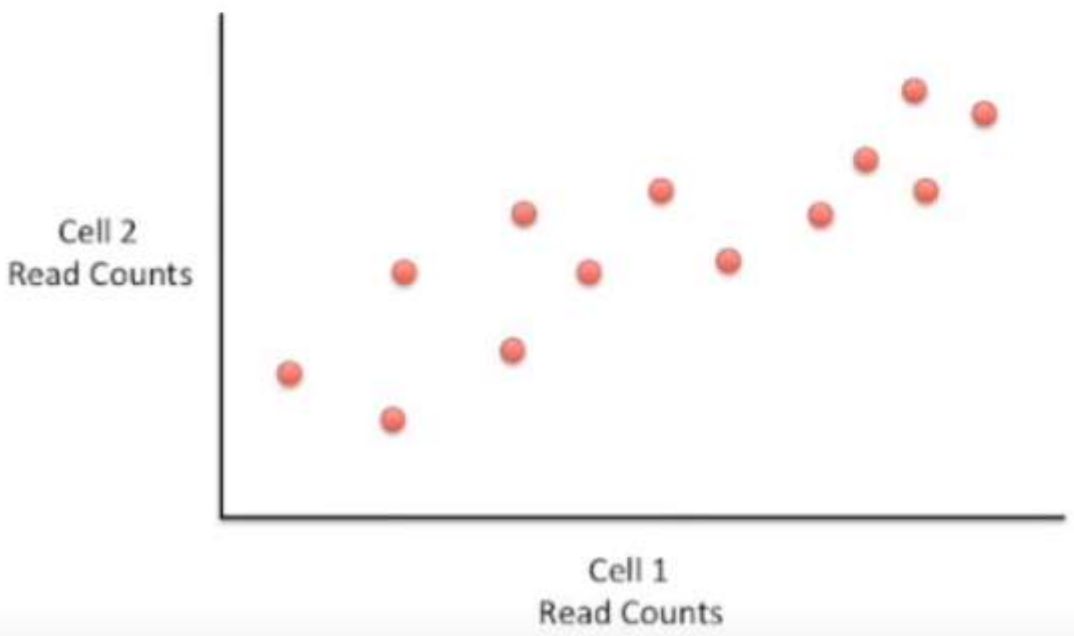
Gene	Cell1 reads	Cell2 reads
a	10	8
b	0	2
c	14	10
d	33	45
e	50	42
f	80	72
g	95	90
h	44	50
i	60	50
... (etc)	... (etc)	... (etc)

PCA IMPLEMENTATION

EXAMPLE

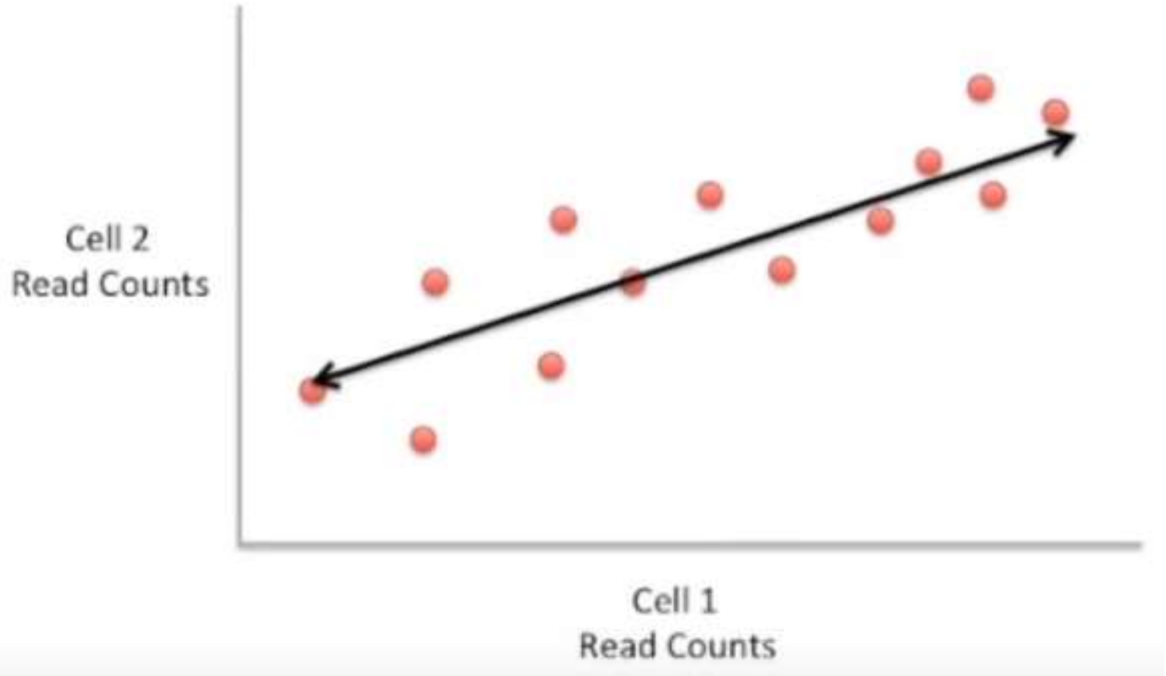
Gene	Cell1 reads	Cell2 reads
a	10	8
b	0	2
c	14	10
d	33	45
e	50	42
f	80	72
g	95	90
h	44	50
i	60	50
... (etc)	... (etc)	... (etc)

Here is a 2-D plot of the data from 2 cells.



PCA IMPLEMENTATION

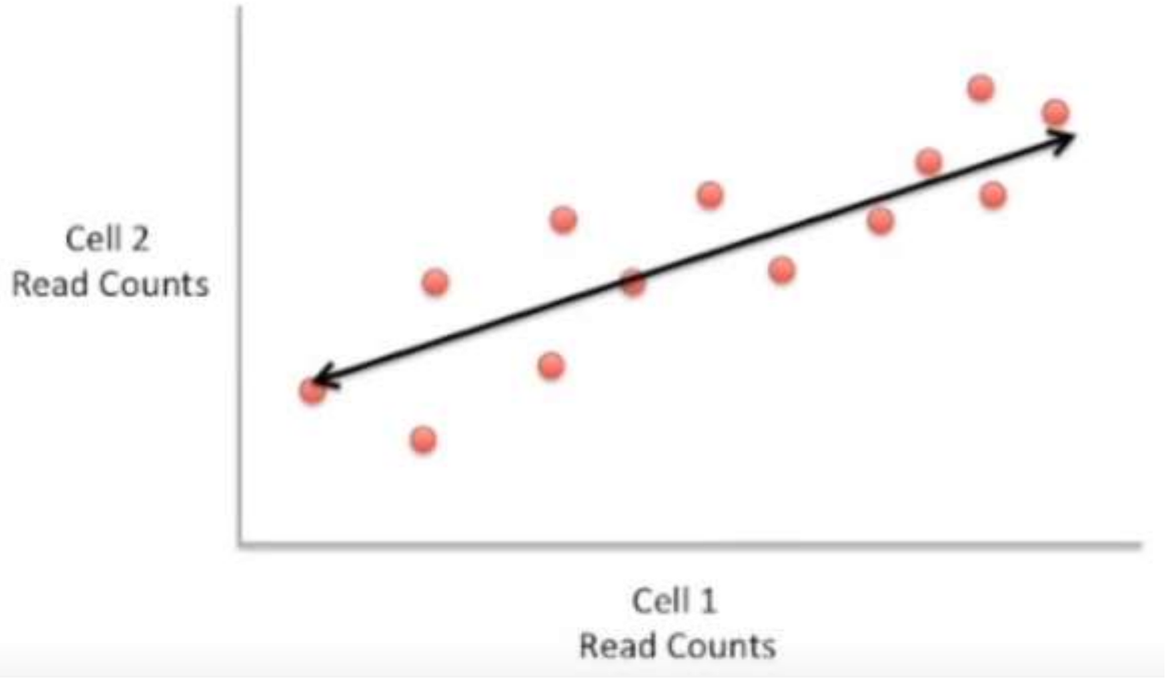
EXAMPLE

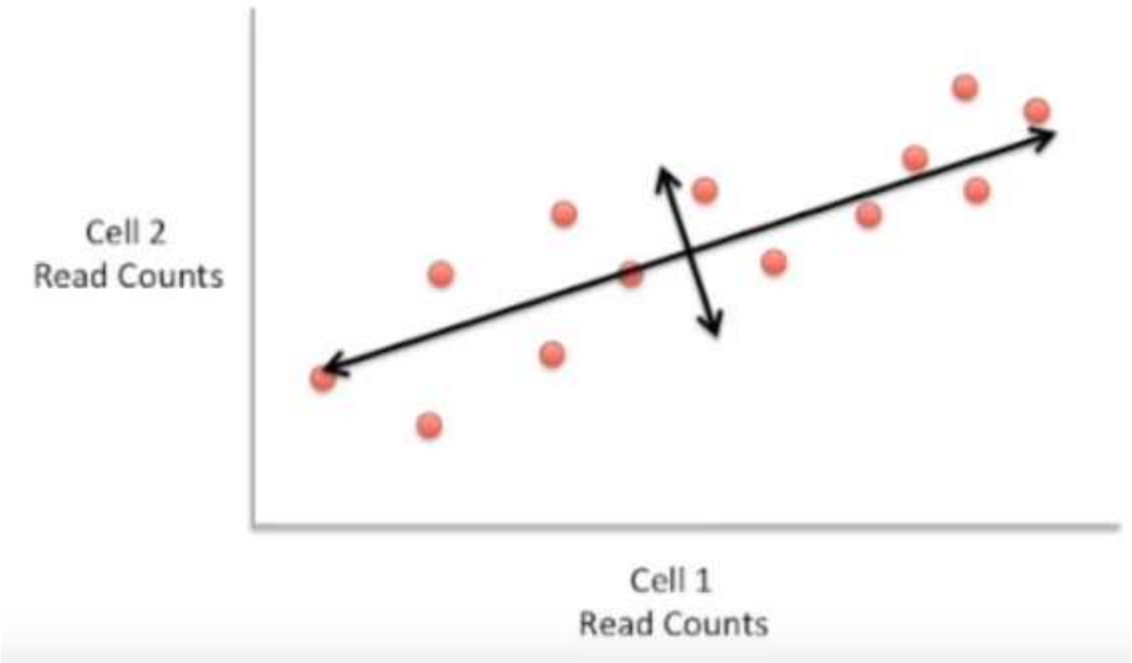


Dot are spread out along a diagonal line and maximum variation of data is between the two end points of line

PCA IMPLEMENTATION

EXAMPLE



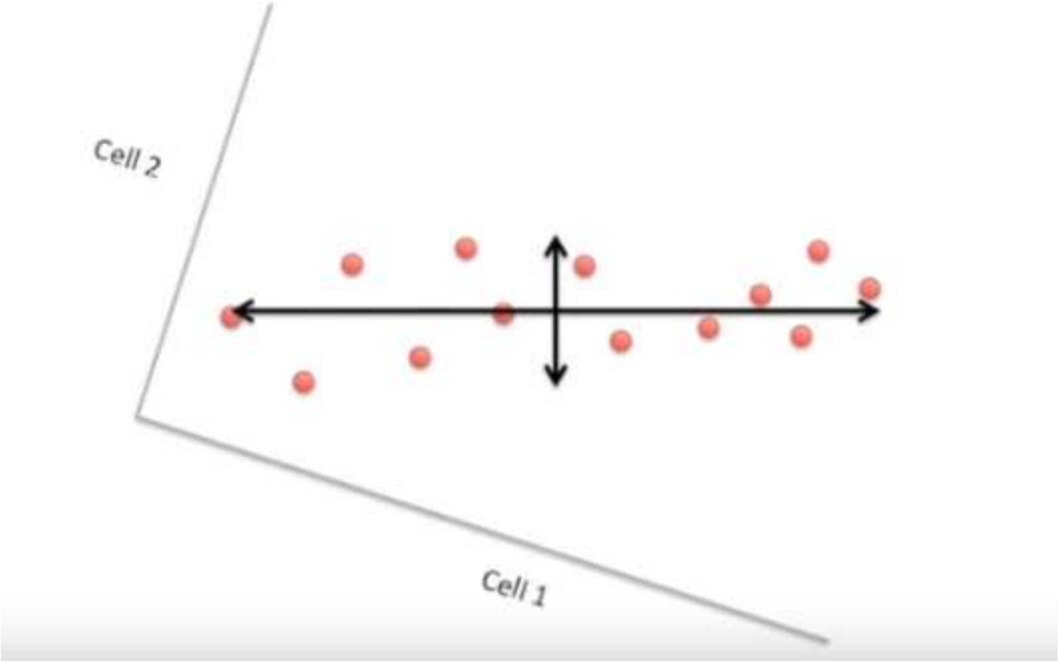


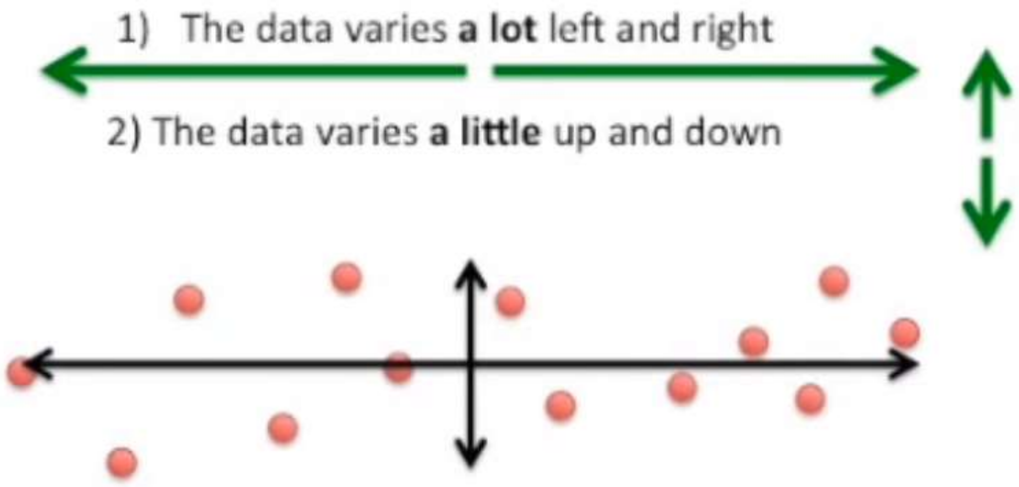
Dot are spread out along a diagonal line and maximum variation of data is between the two end points of line

Dots are also spread out a little above and below the first line and 2nd largest amount of variation is at the endpoints of the new line

PCA IMPLEMENTATION

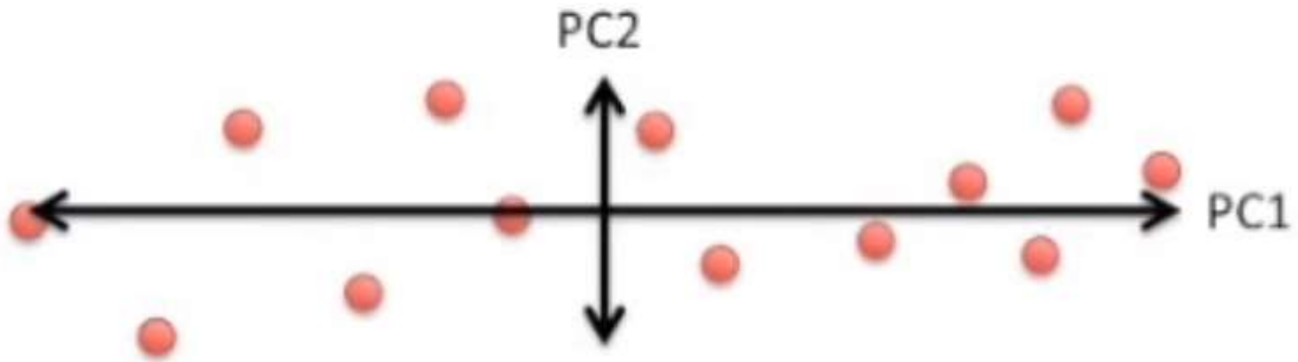
EXAMPLE





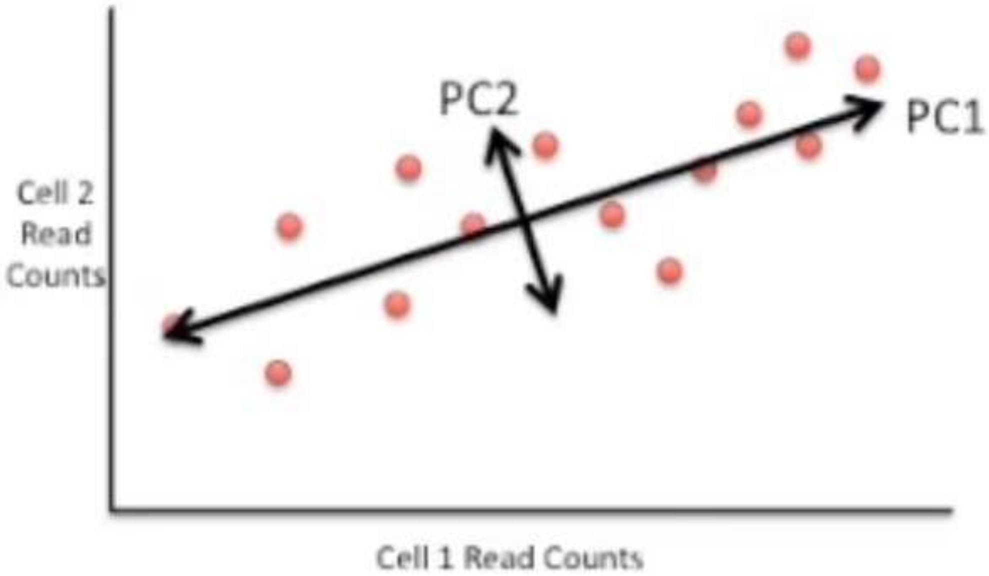
EXAMPLE

These two new axes that describe the variation in the data are “Principal Components”



PCA IMPLEMENTATION

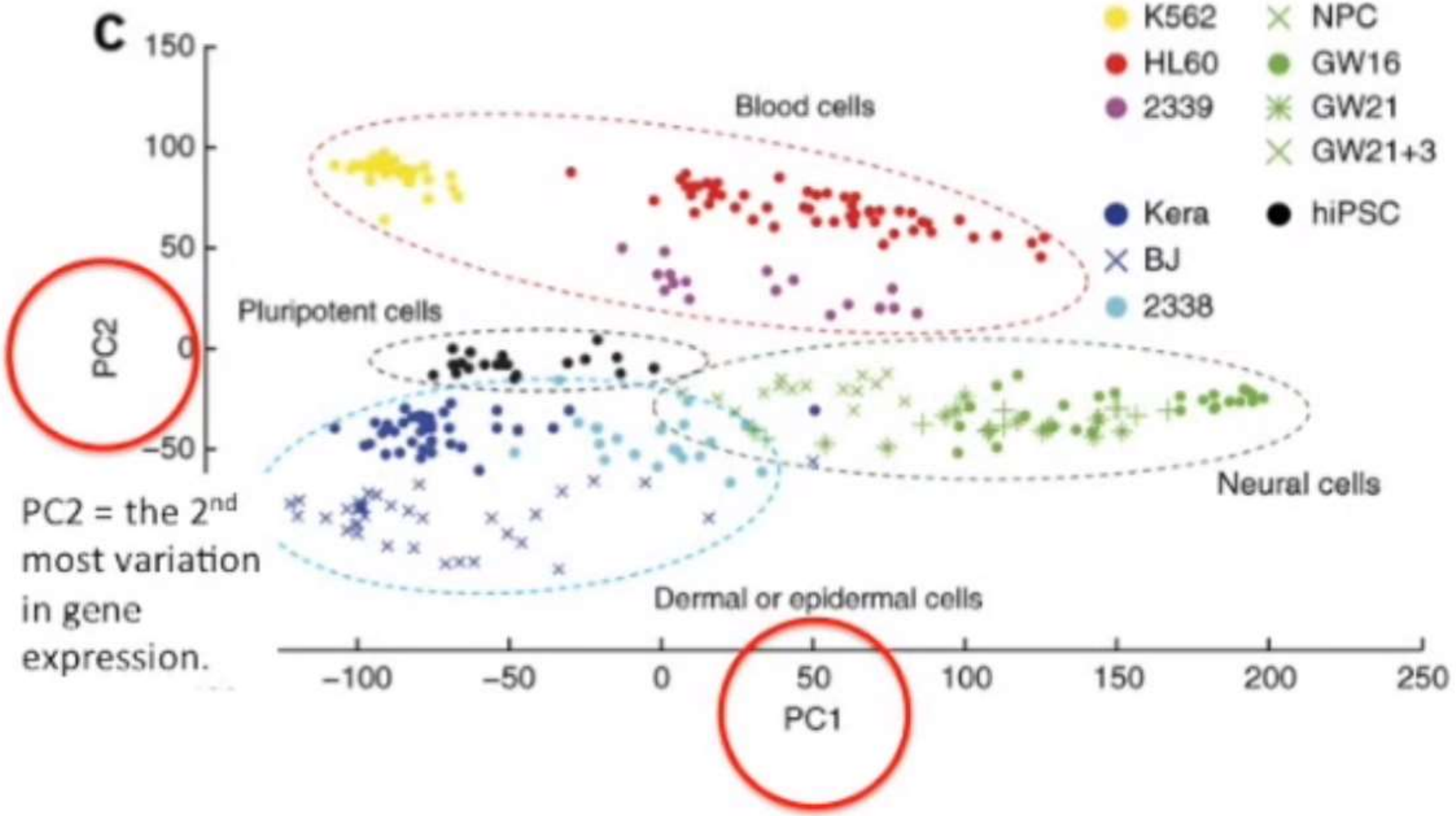
EXAMPLE



- PC1 captures the direction where most of the variation is.
- PC2 captures the direction with the 2nd most variation.

PCA IMPLEMENTATION

EXAMPLE



PC2 = the 2nd most variation in gene expression.

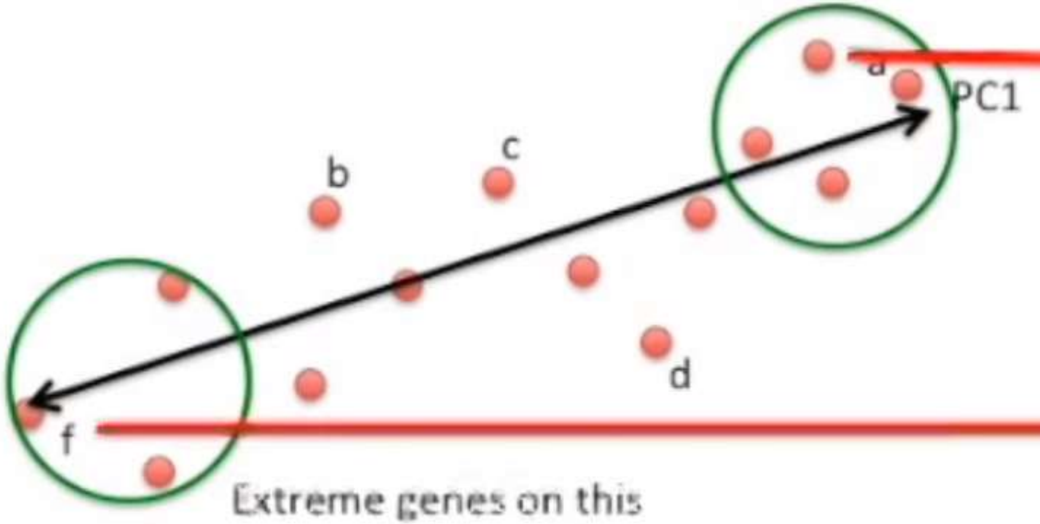
PC1 – the direction of the most variation in gene expression.

PCA IMPLEMENTATION

EXAMPLE

Some genes have more influence on PC1 than others.

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



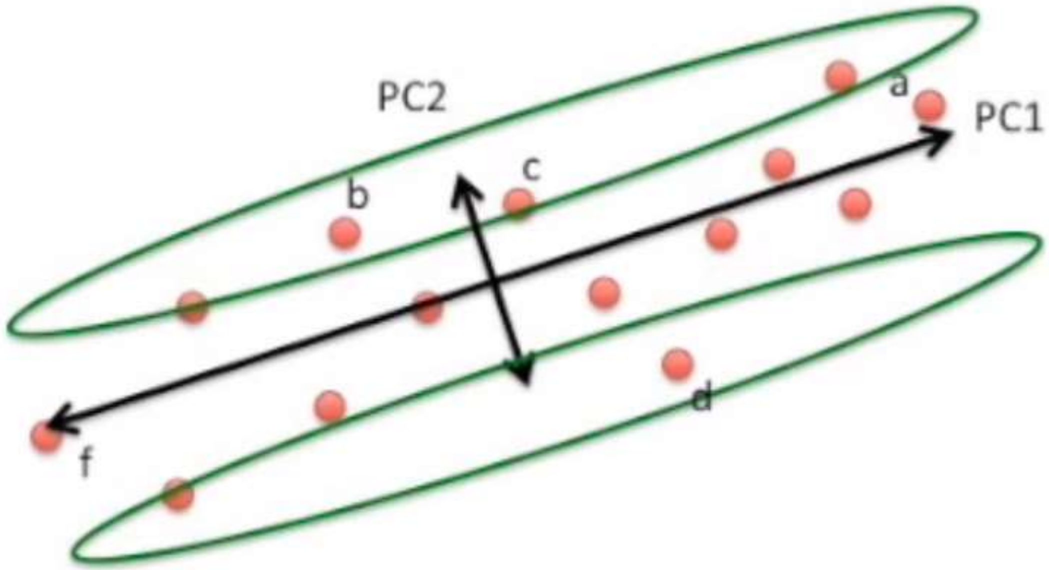
Gene	Influence on PC1	In numbers
a	high	10
b	low	0.5
c	medium	3
d	low	-0.2
e	high	13
f	high	-14
...	...	

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.

PCA IMPLEMENTATION

EXAMPLE



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
...	...	

PCA IMPLEMENTATION

EXAMPLE

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 = (read count * influence) + ... for all genes

PCA IMPLEMENTATION

EXAMPLE

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

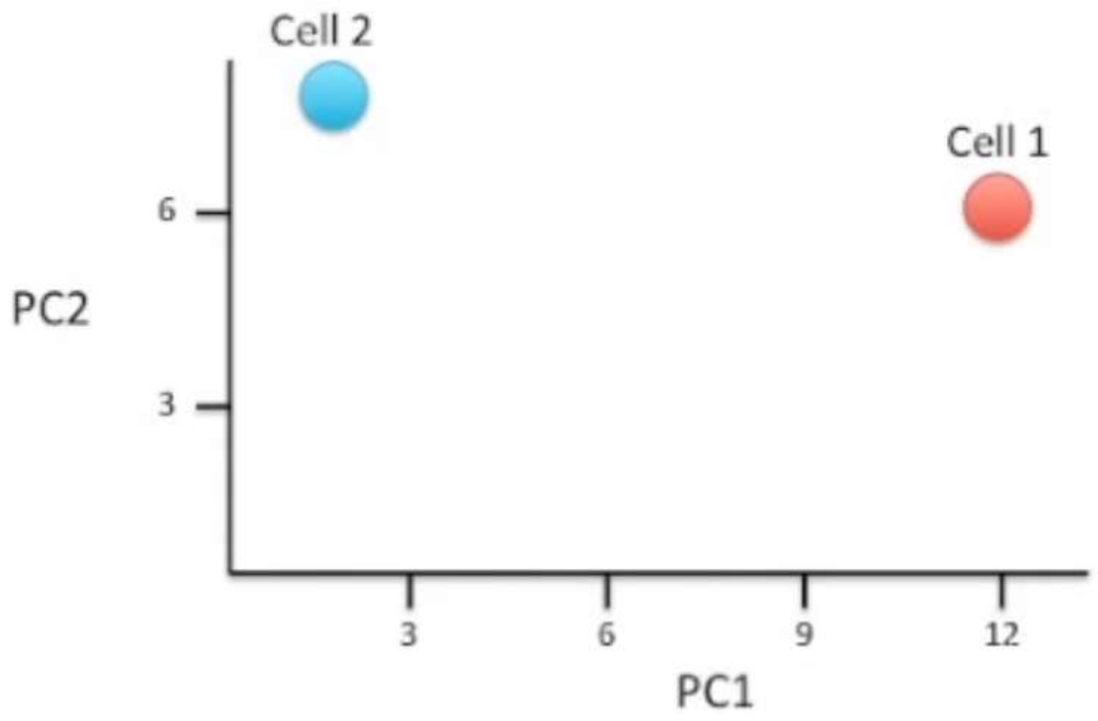
Gene	Influence on PC2	In numbers
a	medium	3
b	high	10
c	high	8
d	high	-12
e	low	0.2
f	low	-0.1
...	...	

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

Cell1 PC2 score = (10 * 3) + (0 * 10) + ... etc... = 6

PCA IMPLEMENTATION

EXAMPLE



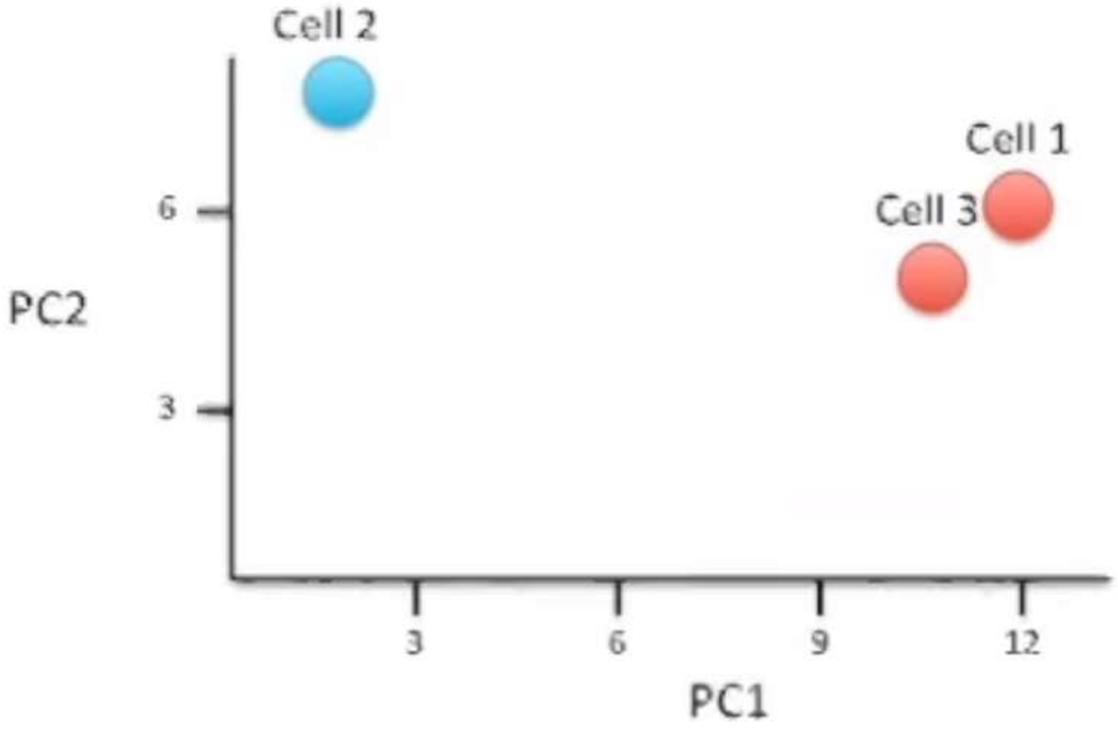
Now calculate scores for Cell2

Cell2 PC1 score = (8 * 10) + (2 * 0.5) + ... etc... = 2

Cell2 PC2 score = (8 * 3) + (2 * 10) + ... etc... = 8

PCA IMPLEMENTATION

EXAMPLE



If we sequenced a third cell, and its transcription was similar to cell 1, it would get scores similar to cell 1's.

USES OF PCA

PCA is mostly used as a tool for **Compression** and **Simplifying** data for **easier learning** in exploratory data analysis and for making predictive models.

1- Better Perspective and less Complexity

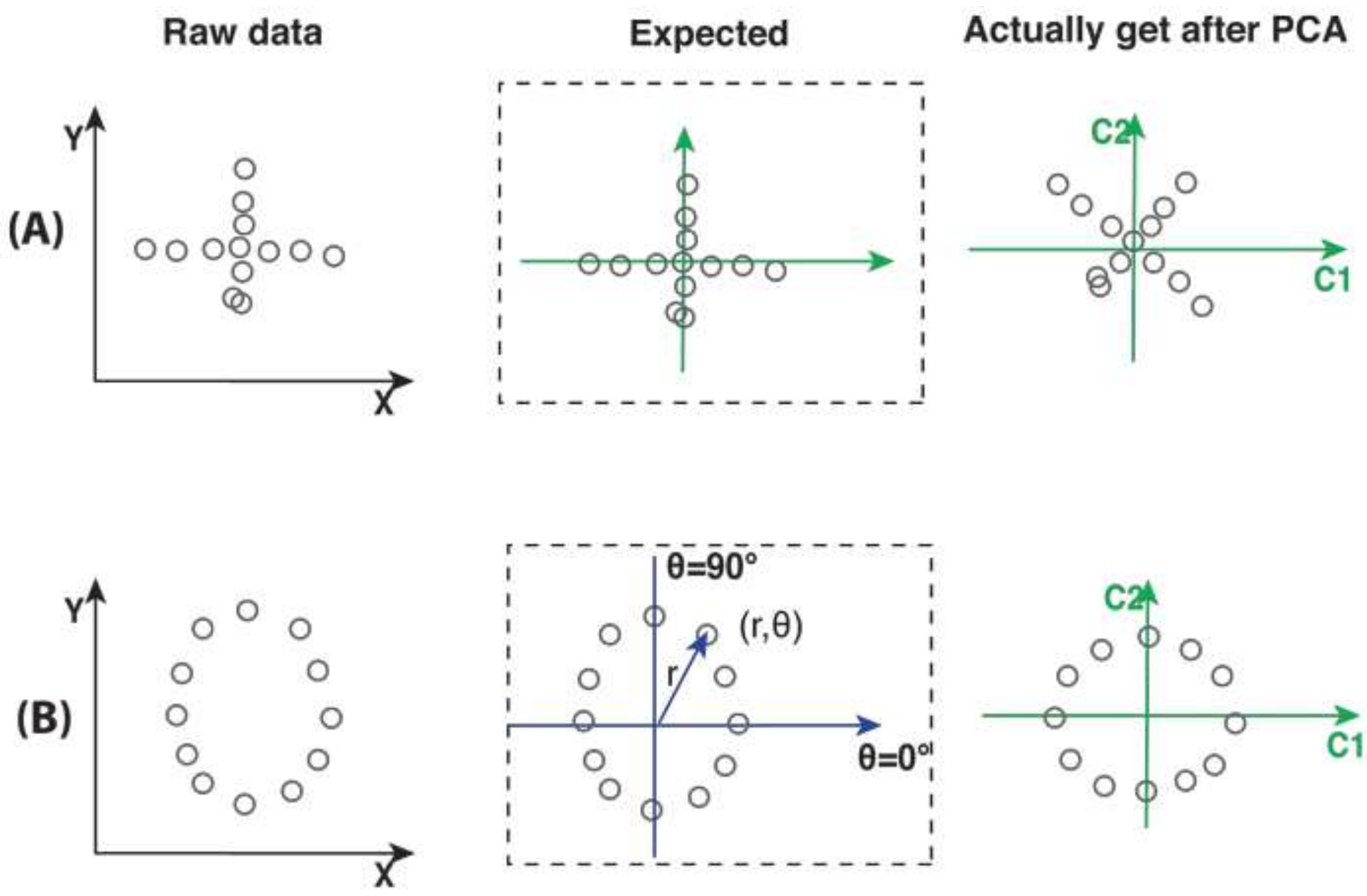
2 - Better visualization

3- Reduce size

4- Different perspective:

LIMITATION OF PCA

If the data does not follow a multidimensional normal (Gaussian) distribution, PCA may not give the best principal components



REFERENCES

Information and Image Credit :

- http://www.mit.edu/~gari/teaching/6.555/LECTURE_NOTES/ch28_bss.pdf
- https://www.quora.com/What-are-some-of-the-limitations-of-principal-component-analysis
- http://mengnote.blogspot.com.ee/2013/05/an-intuitive-explanation-of-pca.html
- https://www.youtube.com/watch?v=_UVHneBUBW0&t=2s
- http://setosa.io/ev/principal-component-analysis/

THANKS