BIMM 143

Unsupervised Learning II

Lecture 9

Barry Grant

UC San Diego

http://thegrantlab.org/bimm143

PCA objectives in a nutshell

- to reduce dimensionality
- to visualize multidimensional data
- to choose the most useful variables (features)
- to identify groupings of objects (e.g. genes/samples)
- to identify outliers

Practical issues with PCA

- Scaling the data
- Missing values:
  - ➡ Drop observations with missing values
  - ➡ Impute / estimate missing values
- Categorical data:
  - ➡ Do not use categorical data features
  - ➡ Encode categorical features as numbers

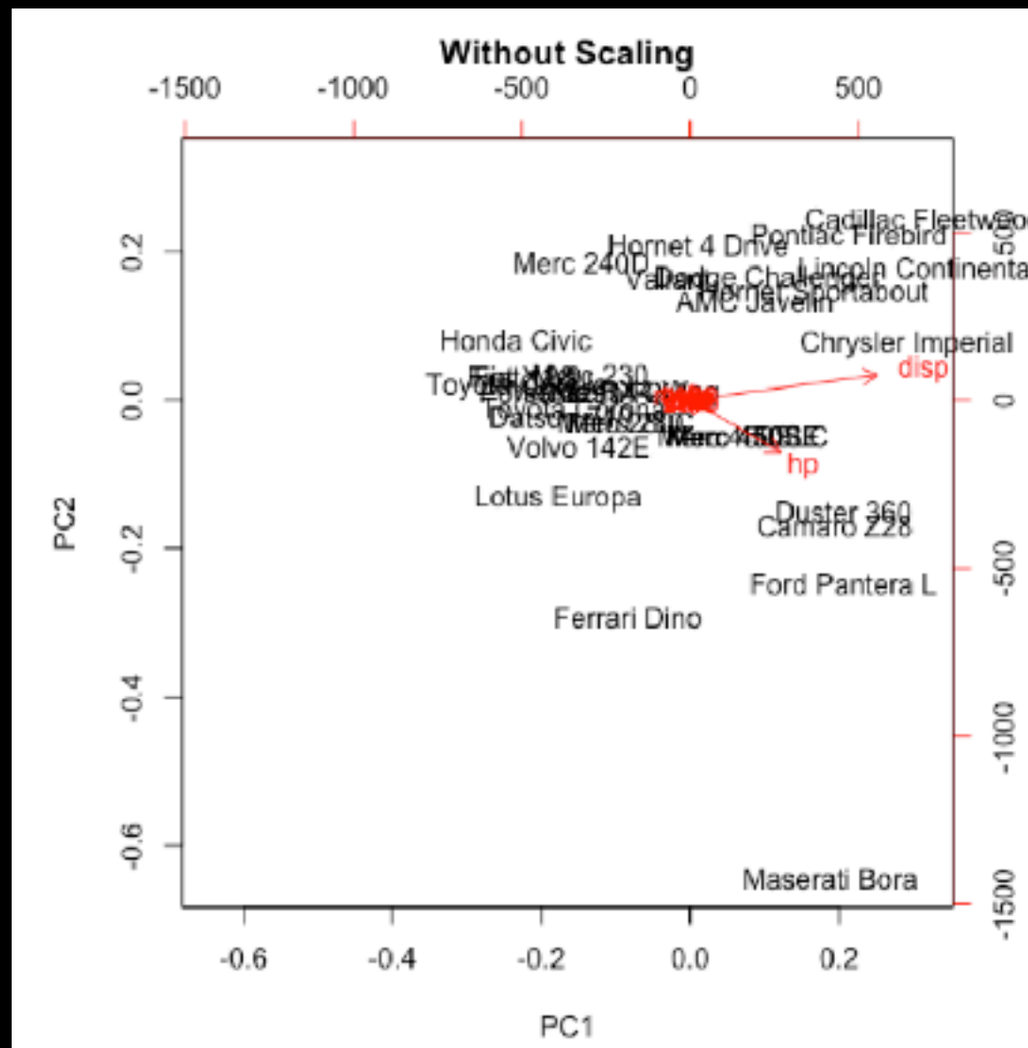
Scaling

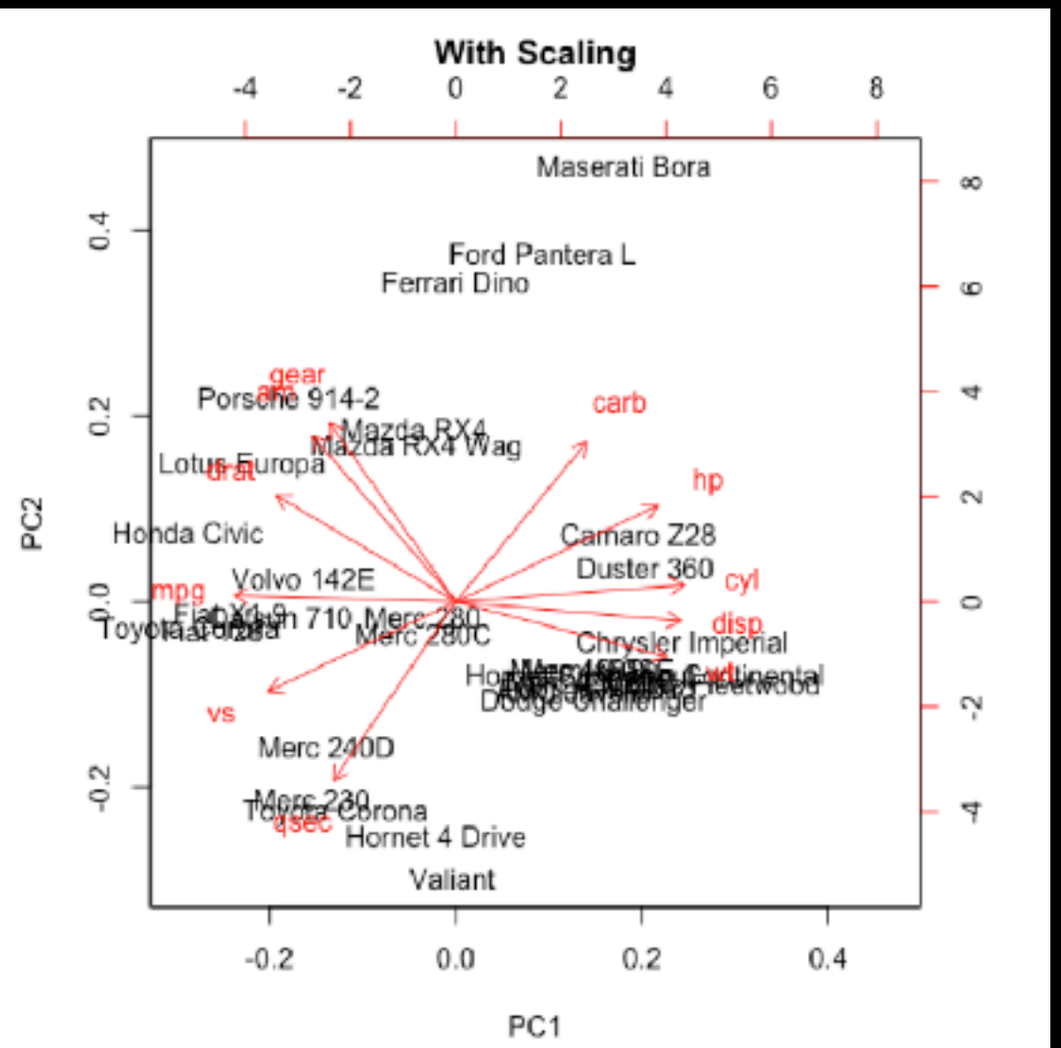
```
> data(mtcars)
> head(mtcars)
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1

# Means and standard deviations vary a lot
> round(colMeans(mtcars), 2)
   mpg    cyl   disp     hp   drat     wt   qsec     vs     am   gear   carb
 20.09   6.19 230.72 146.69   3.60   3.22  17.85   0.44   0.41   3.69   2.81
> round(apply(mtcars, 2, sd), 2)
   mpg    cyl   disp     hp   drat     wt   qsec     vs     am   gear   carb
  6.03   1.79 123.94  68.56   0.53   0.98   1.79   0.50   0.50   0.74   1.62
```

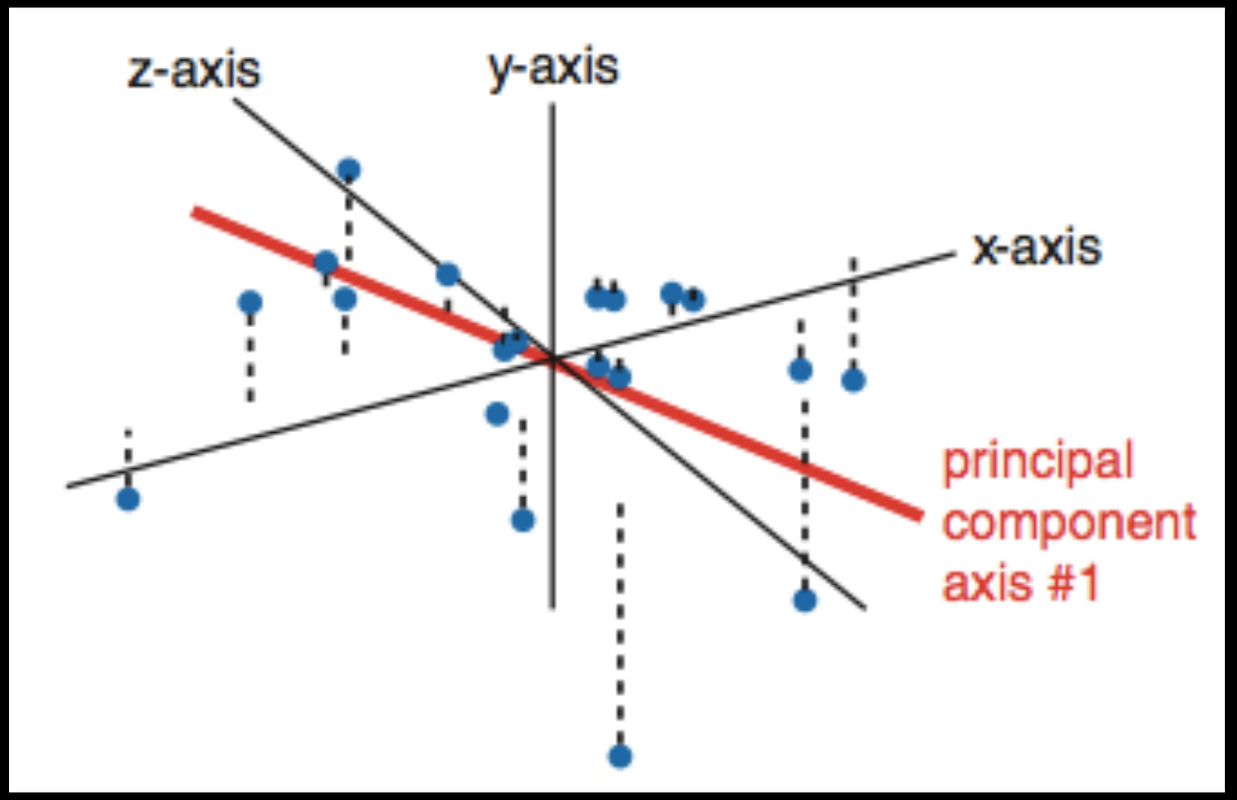
Scaling

```
prcomp(x, center=TRUE, scale=FALSE)
prcomp(x, center=TRUE, scale=TRUE)
```





Principal Components Analysis



- The first principal component (PC) follows a “best fit” through the data points. Other PCs must cross the origin of the plot, and must be orthogonal.

Do it Yourself!

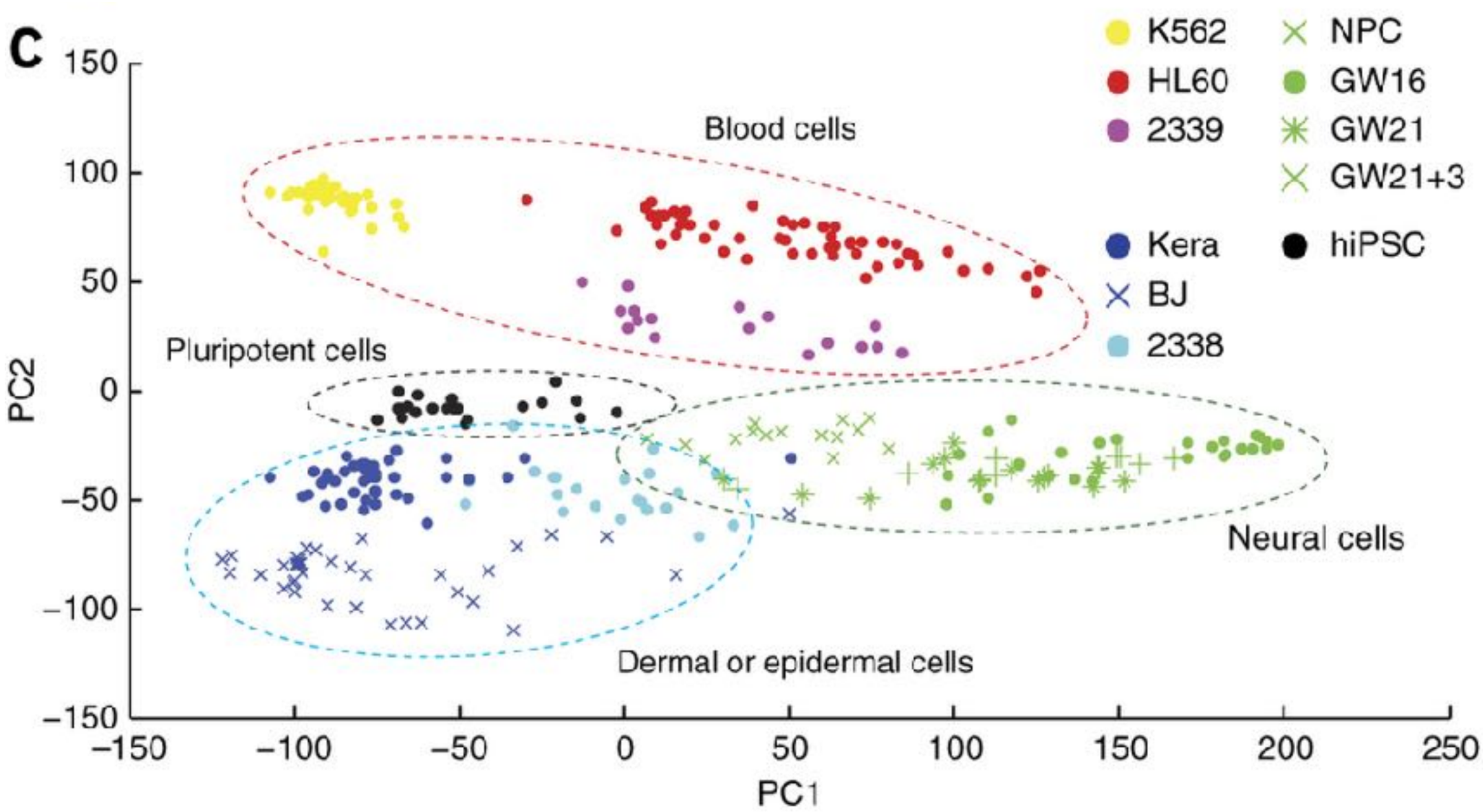
Your turn!

Perform a PCA on the UK foods dataset

Unsupervised Learning Mini-Project

Input: read, View/head,
PCA: prcomp,
Cluster: kmeans, hclust
Compare: plot, table, etc.

This PCA plot shows clusters of cell types.

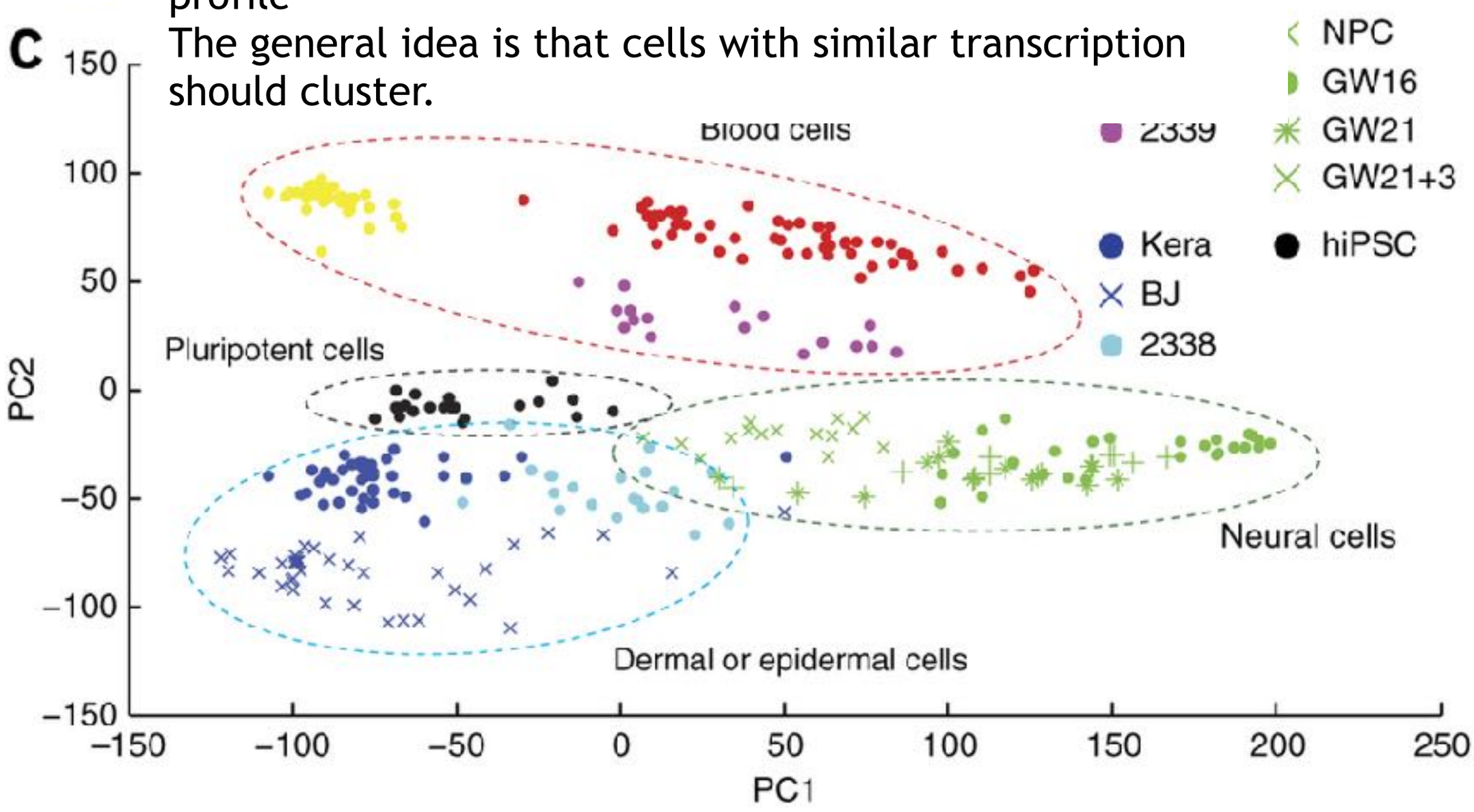


Pollen et al. Nature Biotechnology 2014

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.

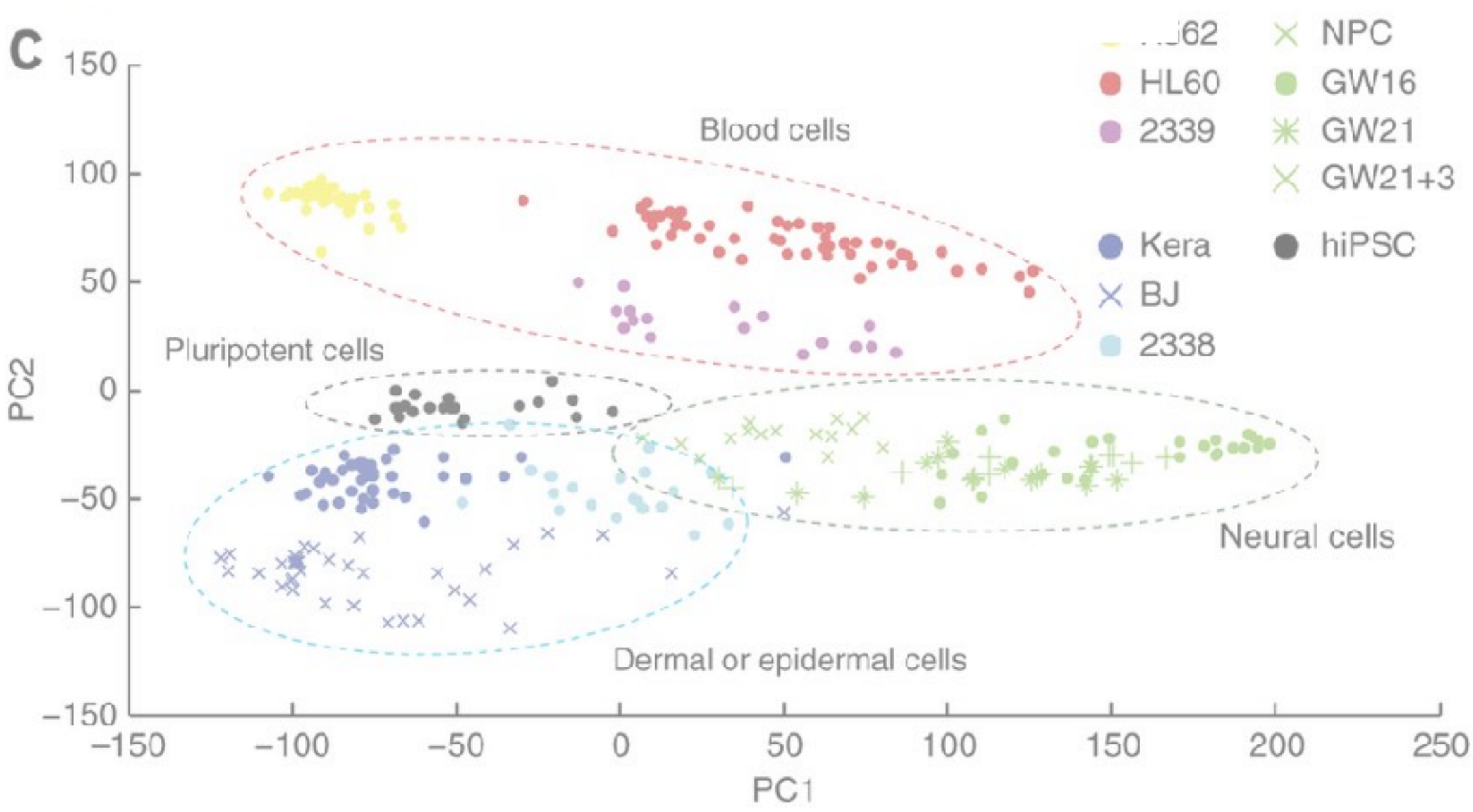
Each dot represents a single-cell and its transcription profile
The general idea is that cells with similar transcription should cluster.



Pollen et al. Nature Biotechnology 2014

This PCA plot shows clusters of cell types.

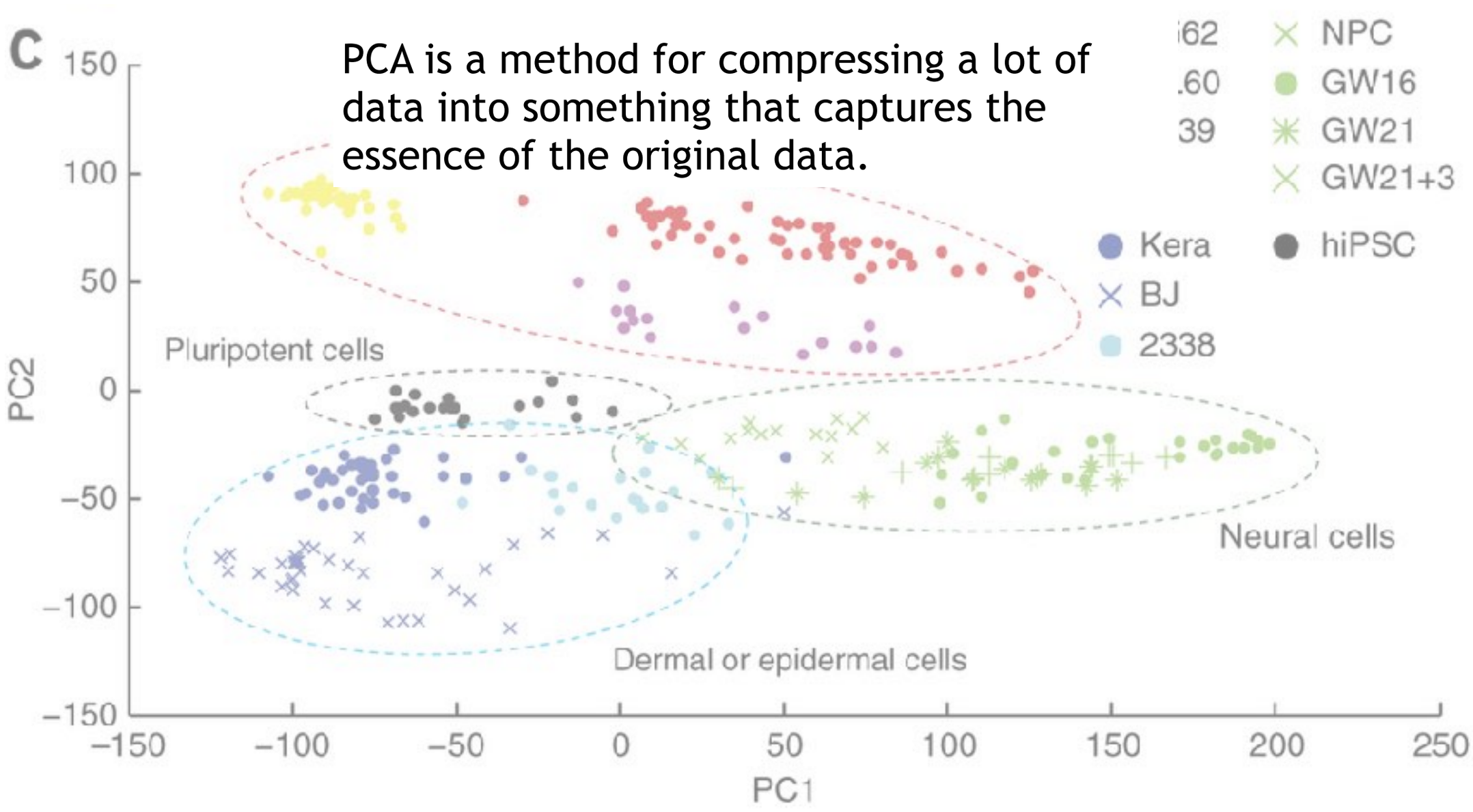
How does transcription from 10,000 genes
get compressed to a single dot on a graph?



This PCA plot shows clusters of cell types.

How does transcription from 10,000 genes get compressed to a single dot on a graph?

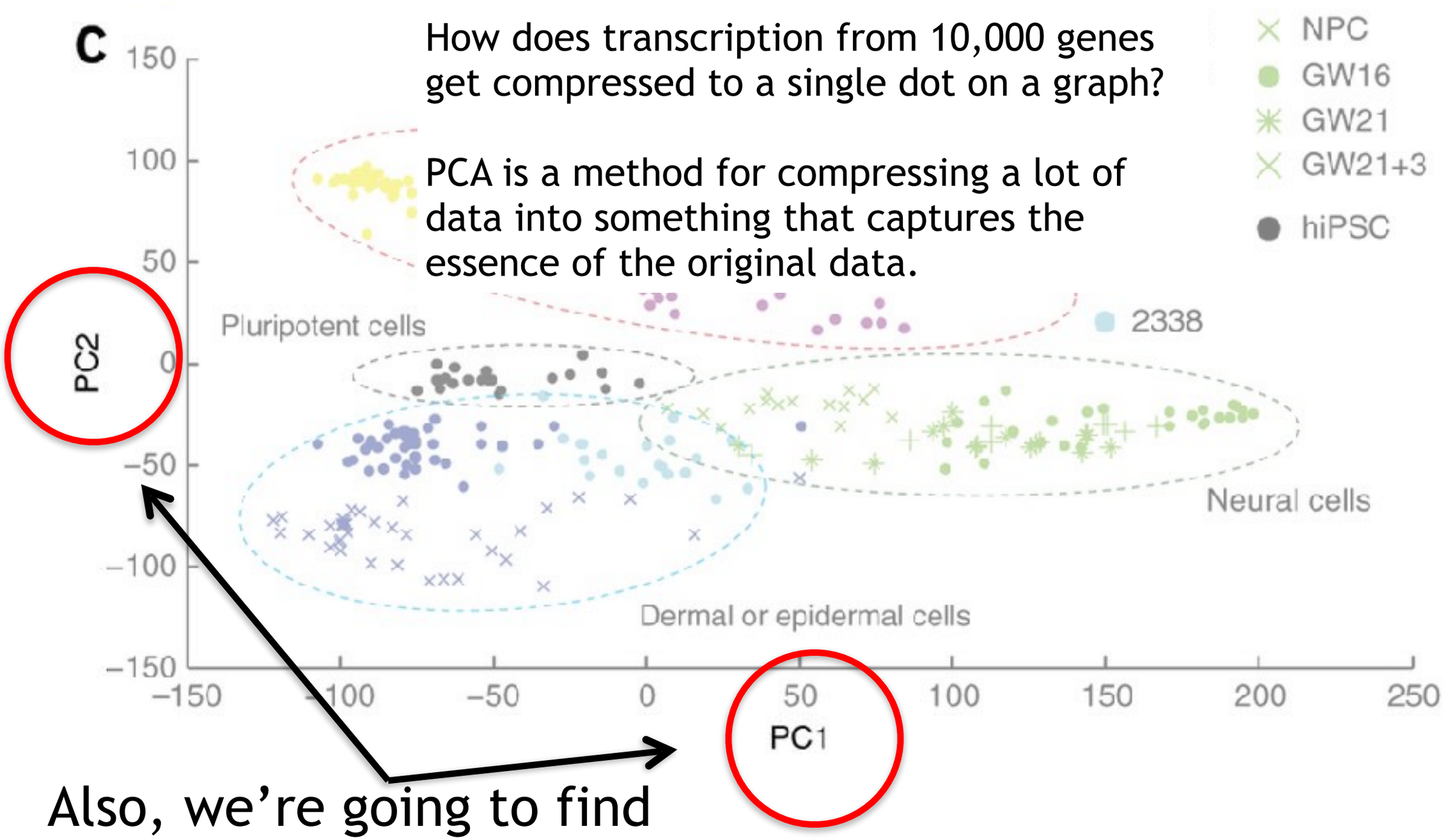
PCA is a method for compressing a lot of data into something that captures the essence of the original data.



This PCA plot shows clusters of cell types.

How does transcription from 10,000 genes get compressed to a single dot on a graph?

PCA is a method for compressing a lot of data into something that captures the essence of the original data.



Also, we're going to find out what these are.

What does PCA aim to do?

- 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)

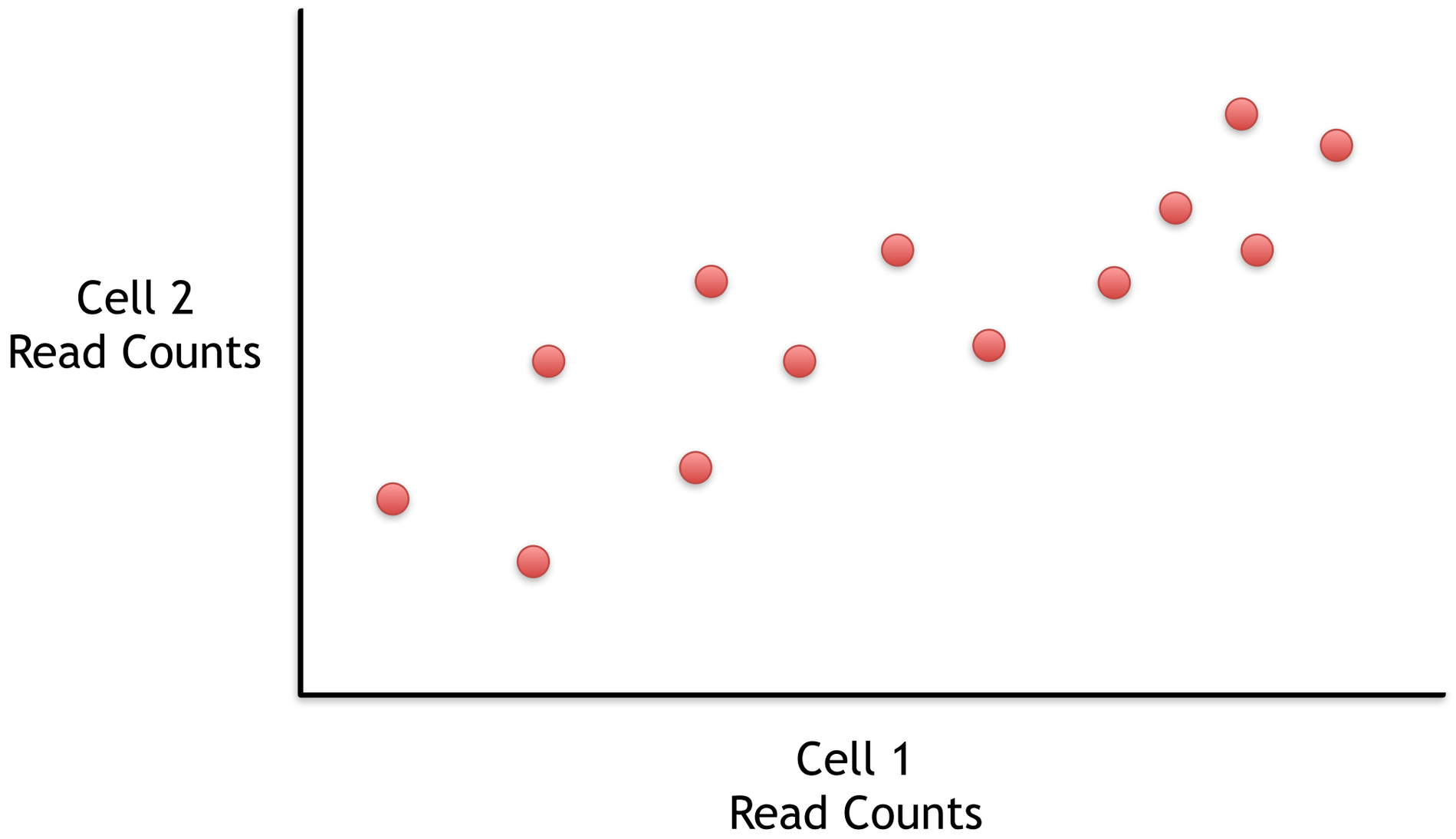
- This is sort of like flattening a Z-stack of microscope images to make a single 2-D image for publication.

A PCA example

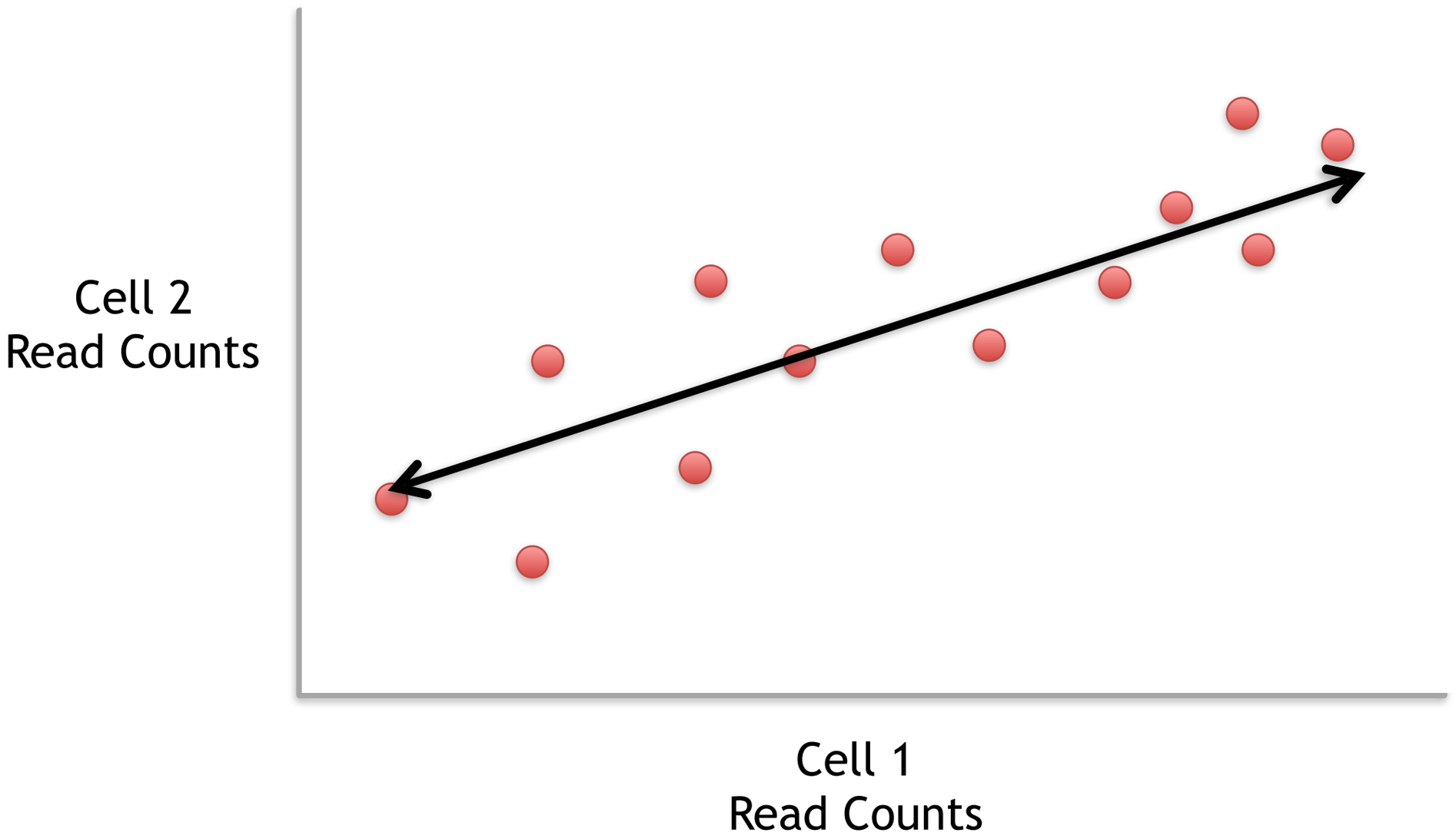
Again, we'll start with just two cells
Here's the data:

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.

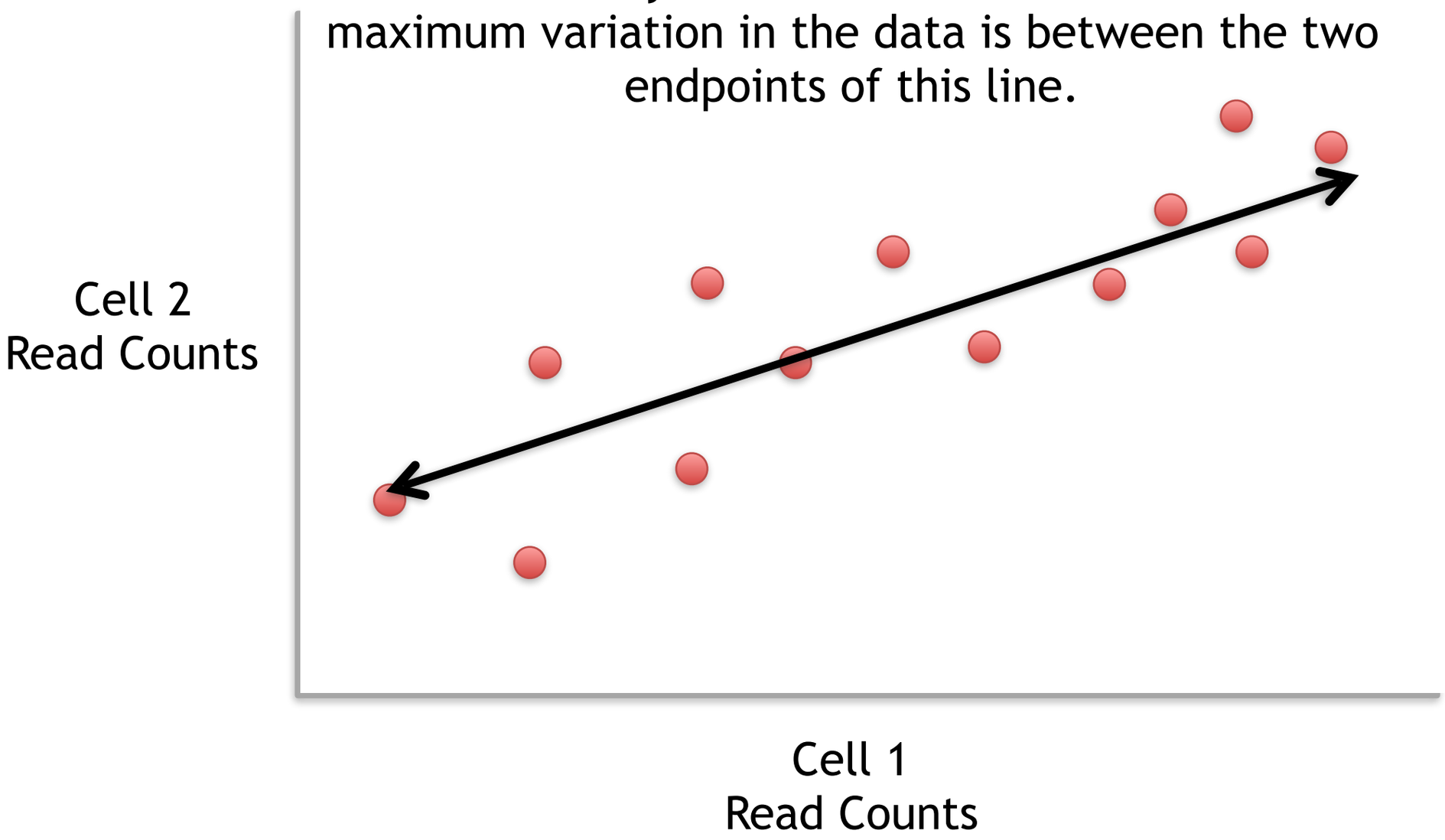


Generally speaking, the dots are spread out along a diagonal line.

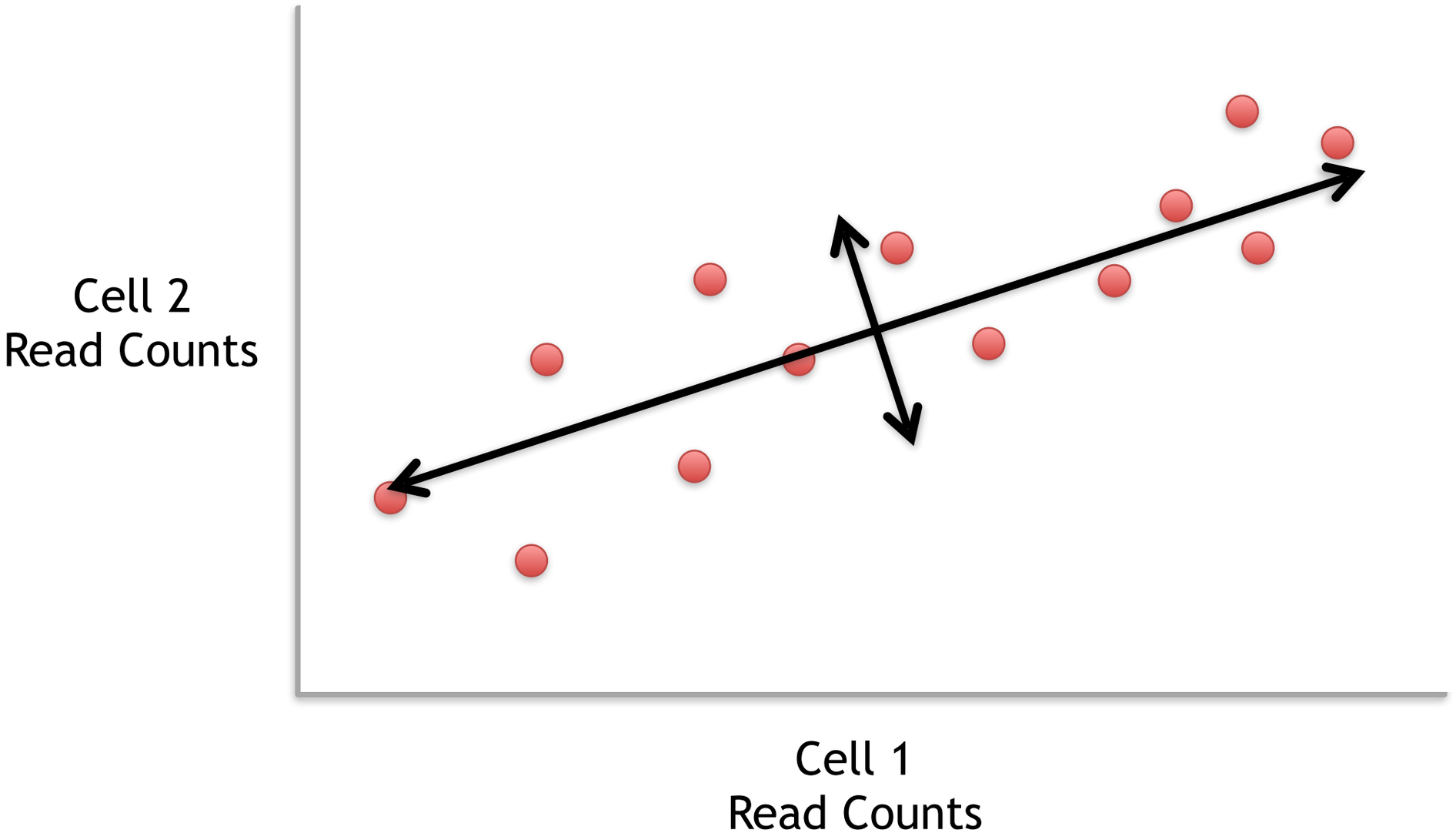


Generally speaking, the dots are spread out along a diagonal line.

Another way to think about this is that the maximum variation in the data is between the two endpoints of this line.

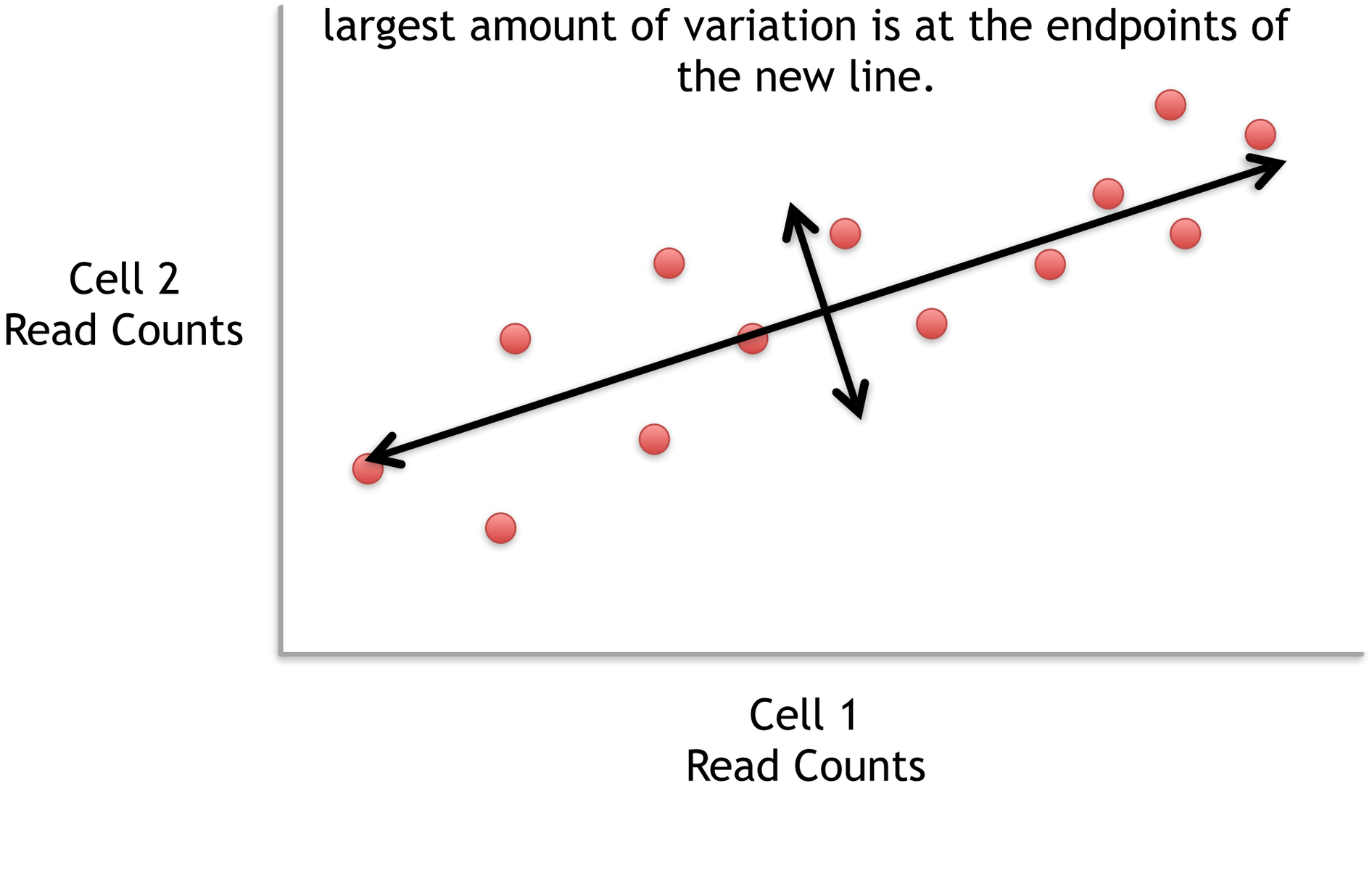


Generally speaking, the dots are also spread out a little above and below the first line.

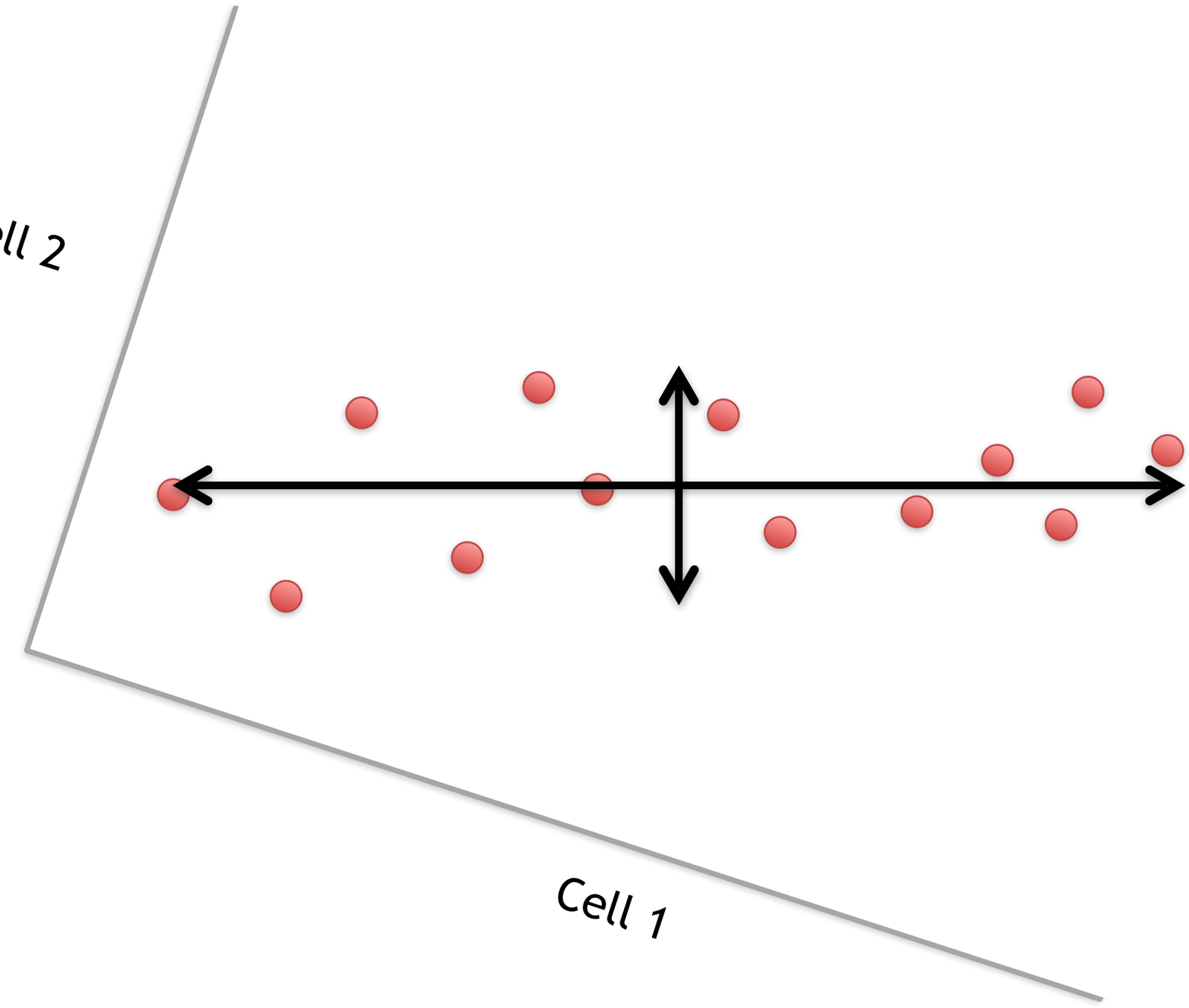


Generally speaking, the dots are also spread out a little above and below the first line.

Another way to think about this is that the 2nd largest amount of variation is at the endpoints of the new line.

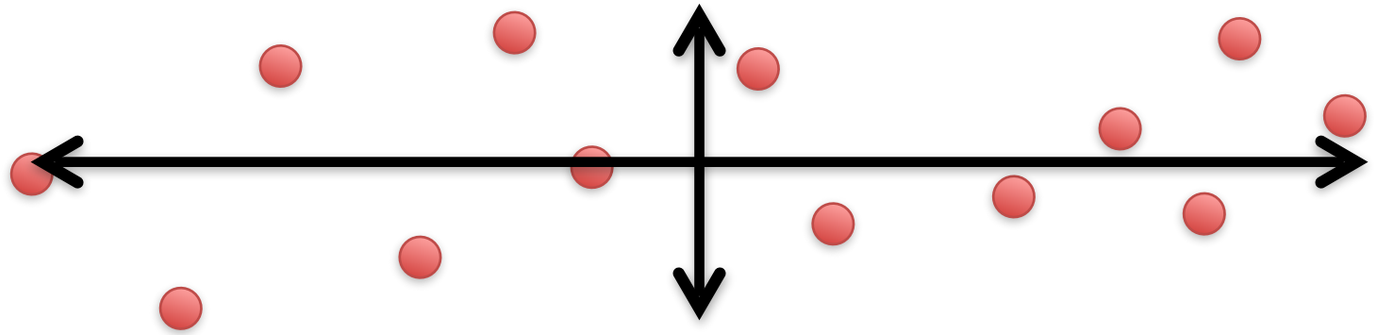


If we rotate the whole graph, the two lines that we drew make new X and Y axes.



If we rotate the whole graph, the two lines that we drew make new X and Y axes.

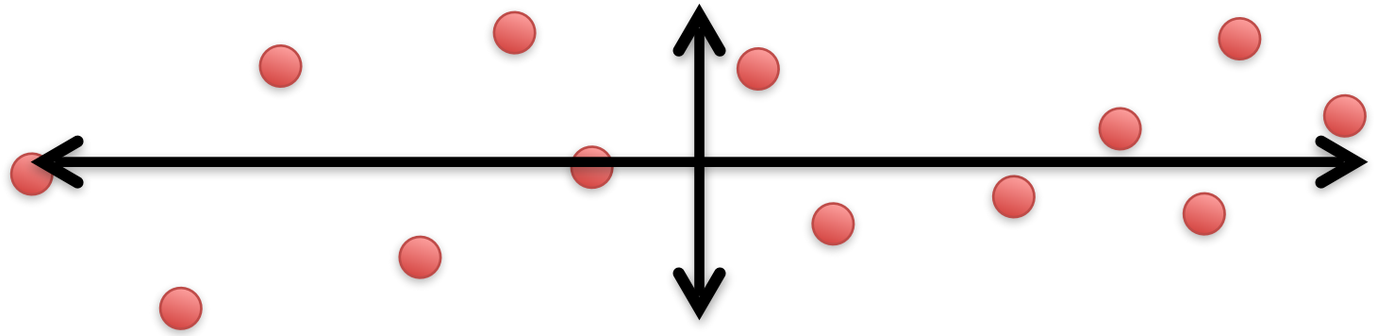
This makes the left/right, above/below variation easier to see.



If we rotate the whole graph, the two lines that we drew make new X and Y axes.

This makes the left/right, above/below variation easier to see.

1) The data varies **a lot** left and right

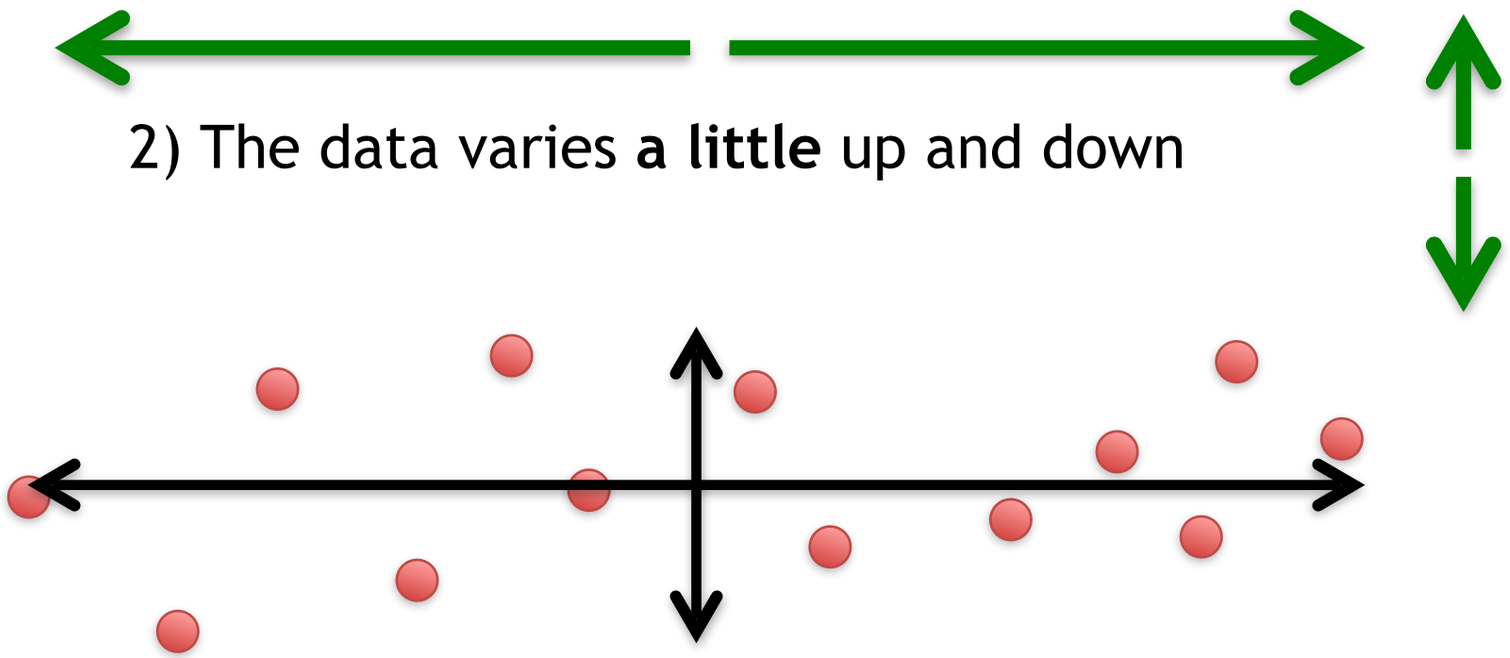


If we rotate the whole graph, the two lines that we drew make new X and Y axes.

This makes the left/right, above/below variation easier to see.

1) The data varies **a lot** left and right

2) The data varies **a little** up and down

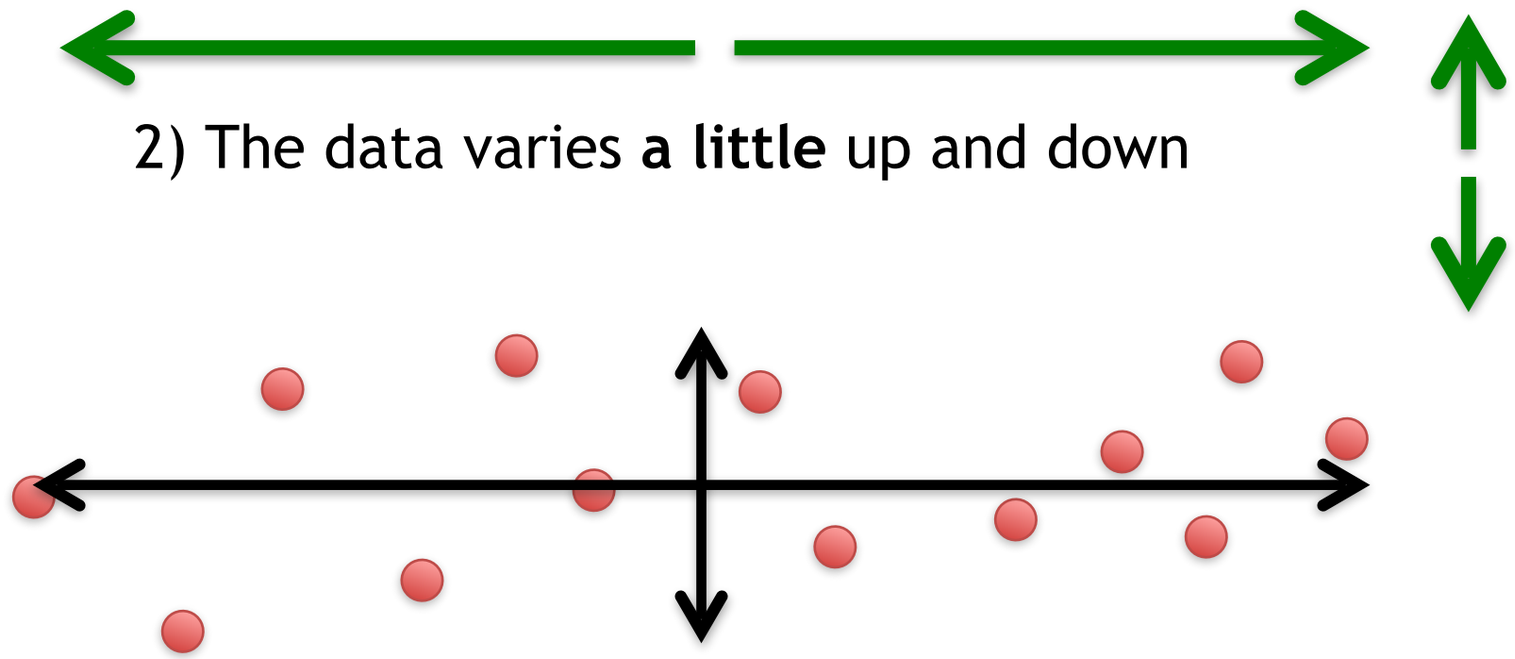


If we rotate the whole graph, the two lines that we drew make new X and Y axes.

This makes the left/right, above/below variation easier to see.

1) The data varies **a lot** left and right

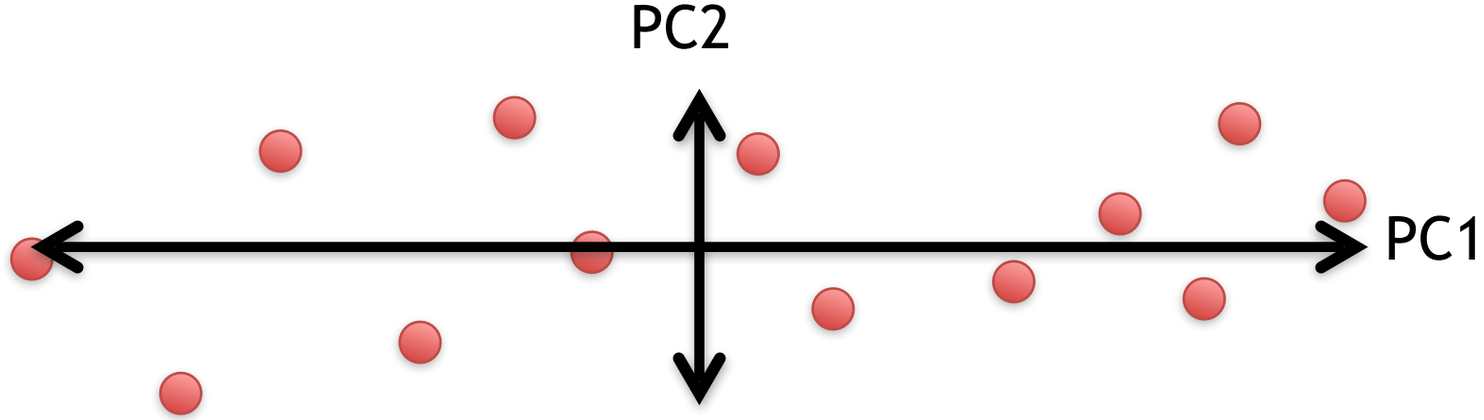
2) The data varies **a little** up and down



Note: All of the points can be drawn in terms of left/right + up/down, just like any other 2-D graph.

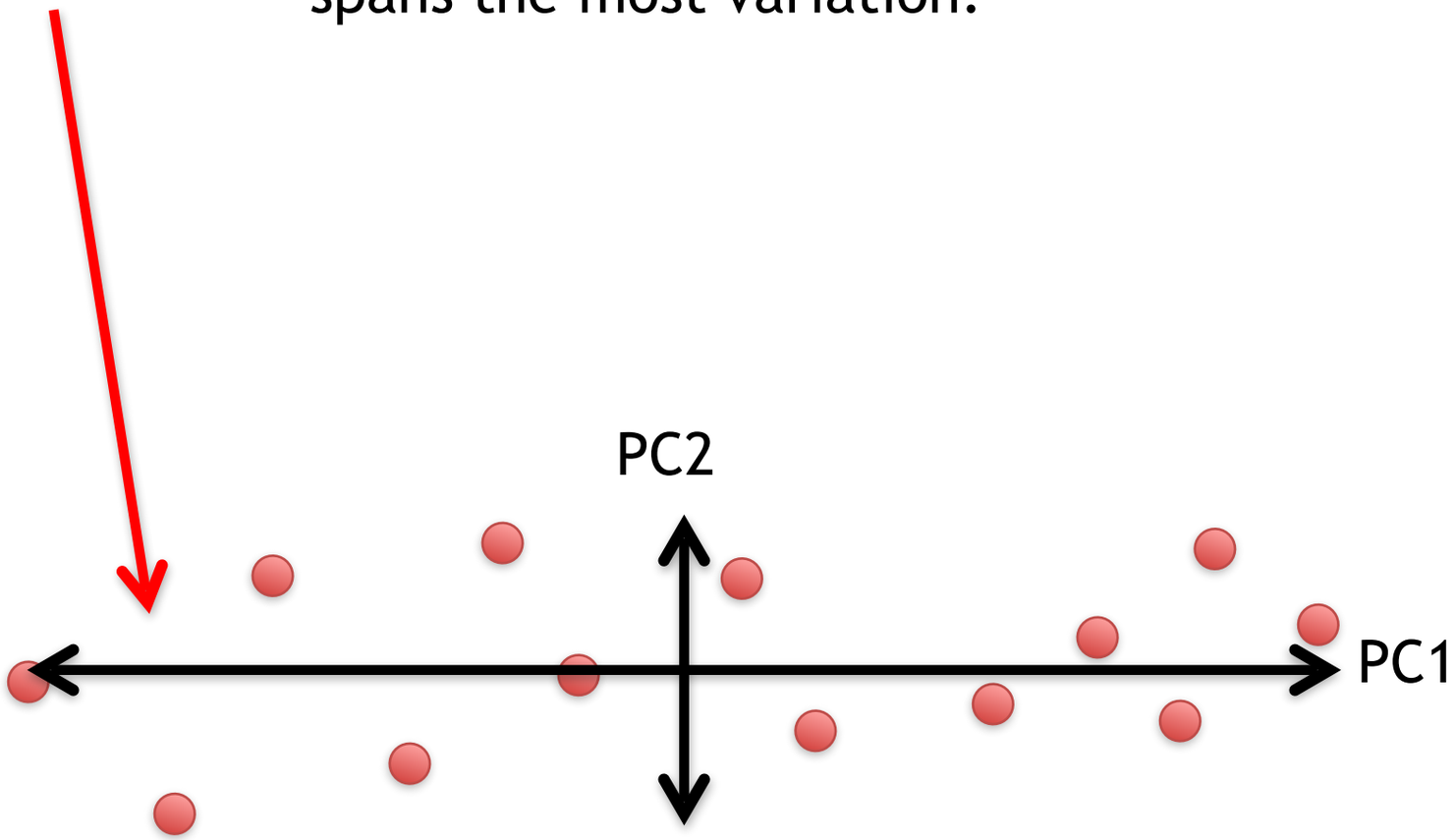
That is to say, we do not need another line to describe “diagonal” variation – we’ve already captured the two directions that can have variation.

These two “new” (or “rotated”) axes that describe the variation in the data are “Principal Components” (PCs)



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

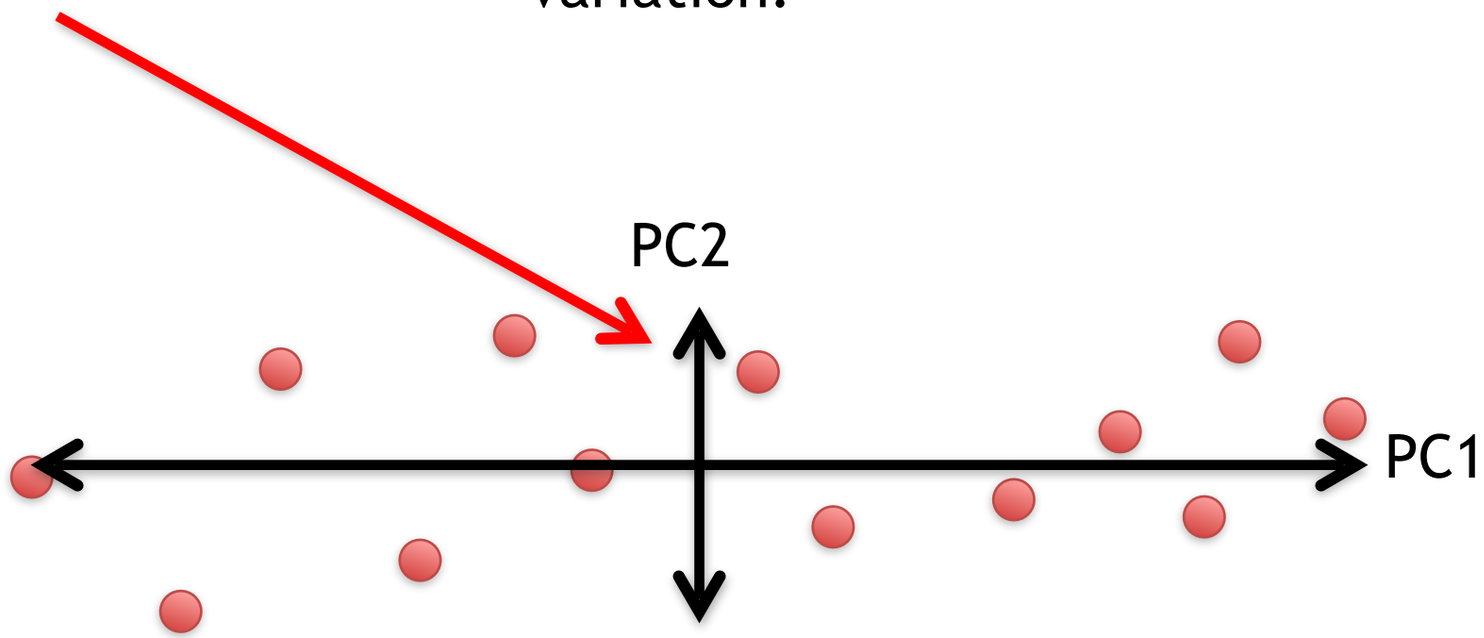
PC1 (the first principal component) is the axis that spans the most variation.



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

PC1 (the first principal component) is the axis that spans the most variation.

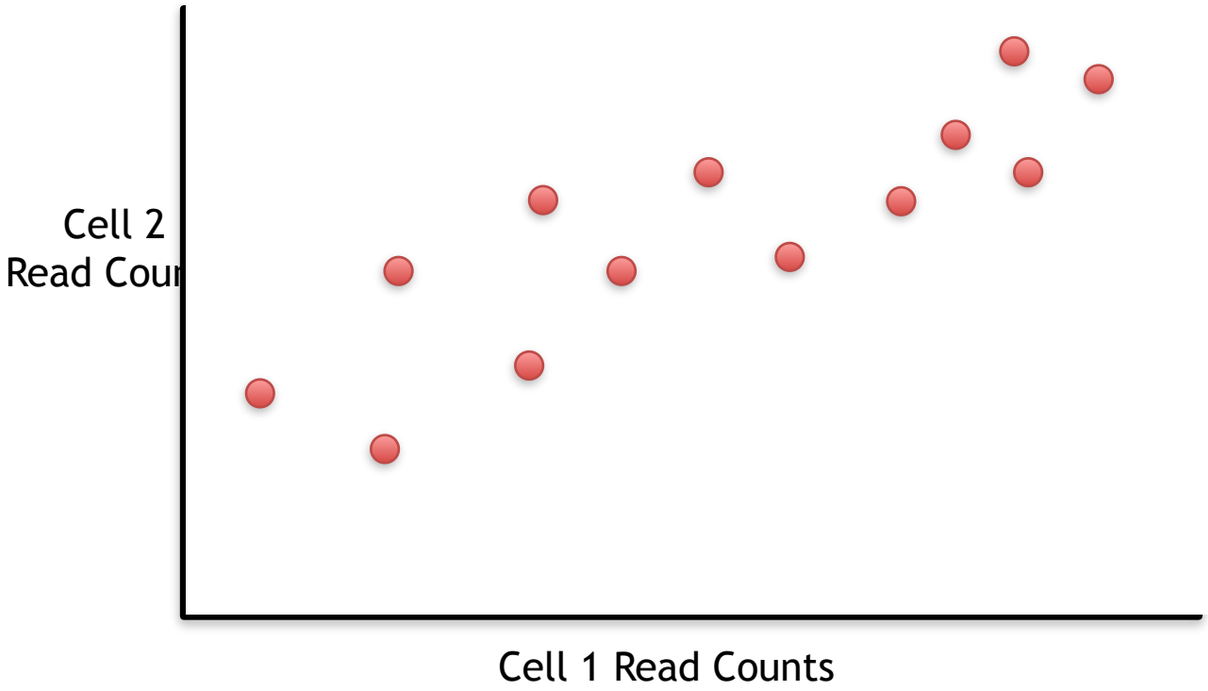
PC2 is the axis that spans the second most variation.



General ideas so far…

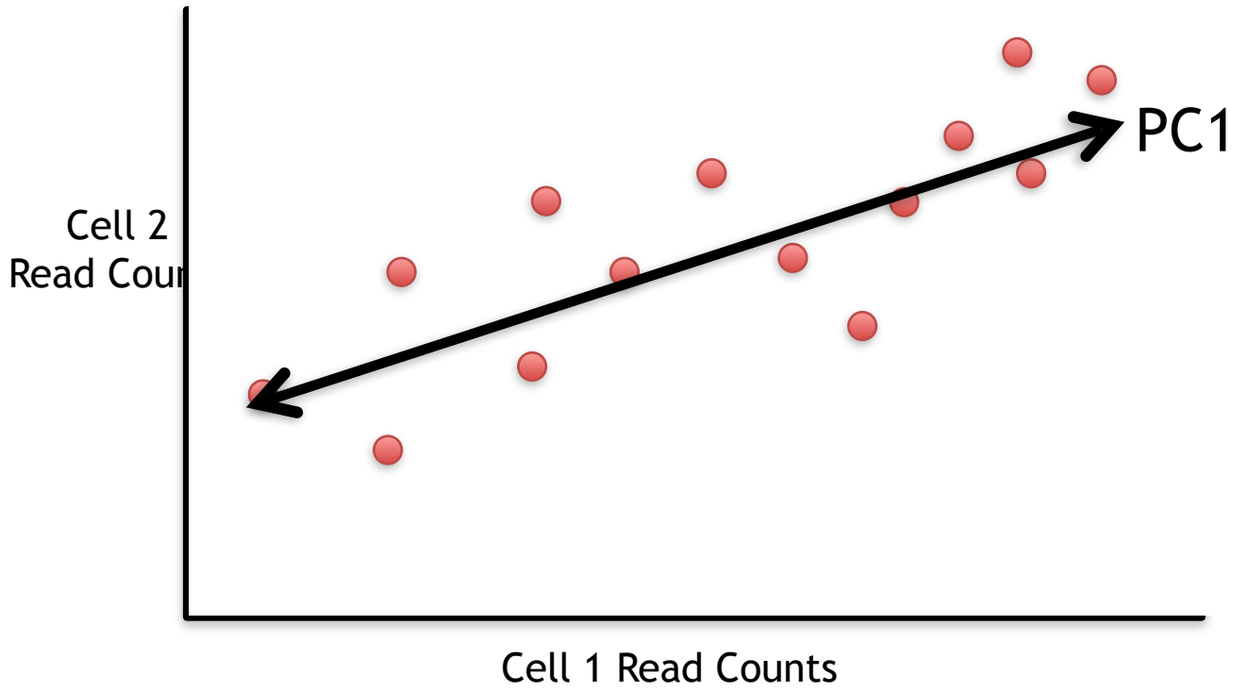
General ideas so far…

- For each gene, we plotted a point based on how many reads were from each cell.



General ideas so far…

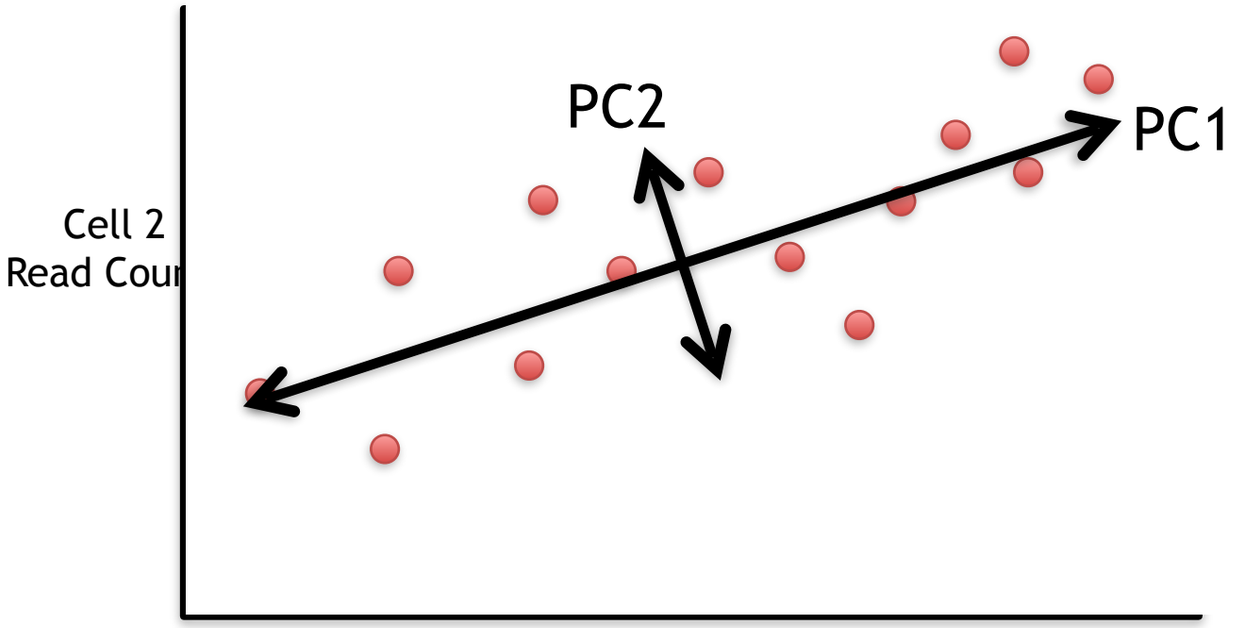
- For each gene, we plotted a point based on how many reads were from each cell.



- PC1 captures the direction where most of the variation is.

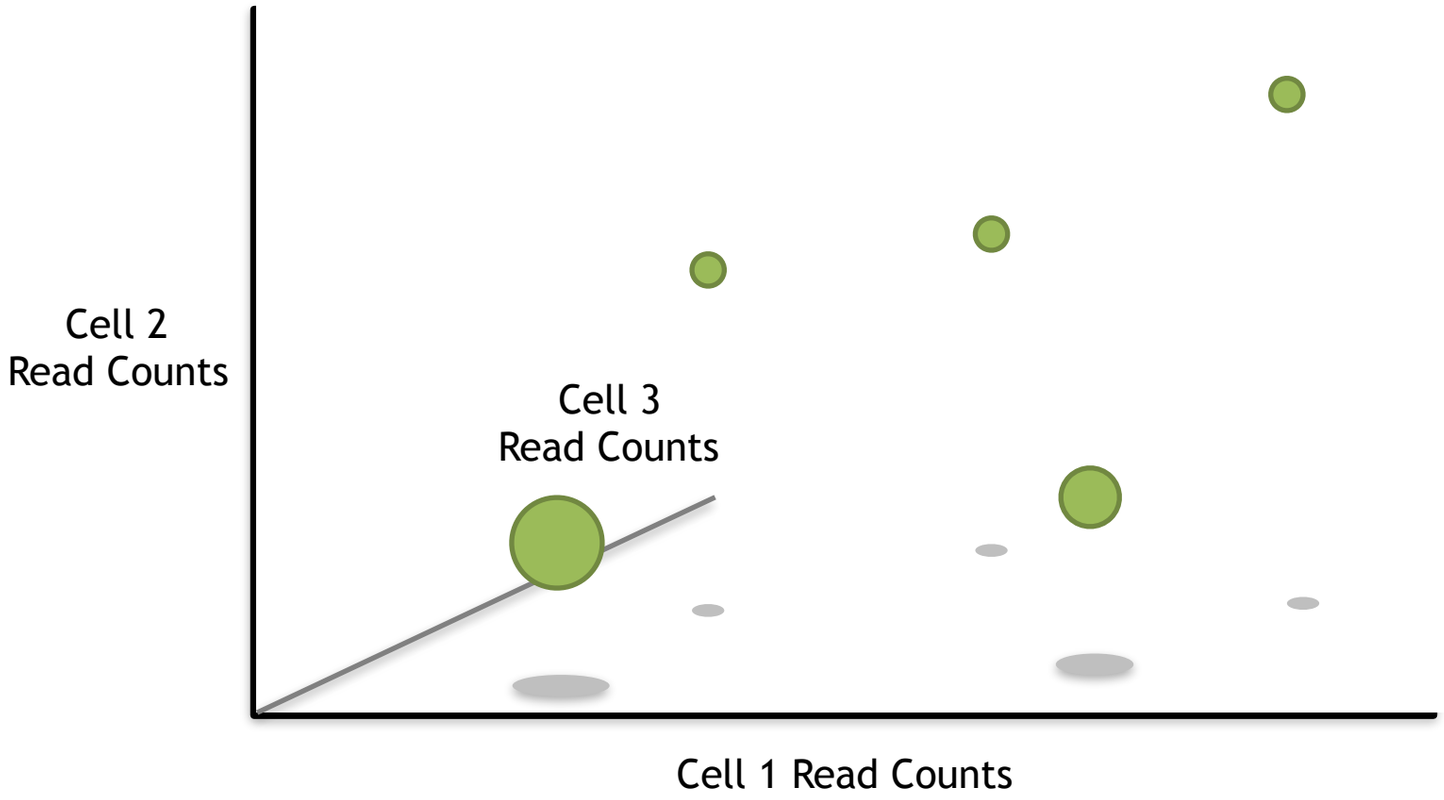
General ideas so far…

- For each gene, we plotted a point based on how many reads were from each cell.

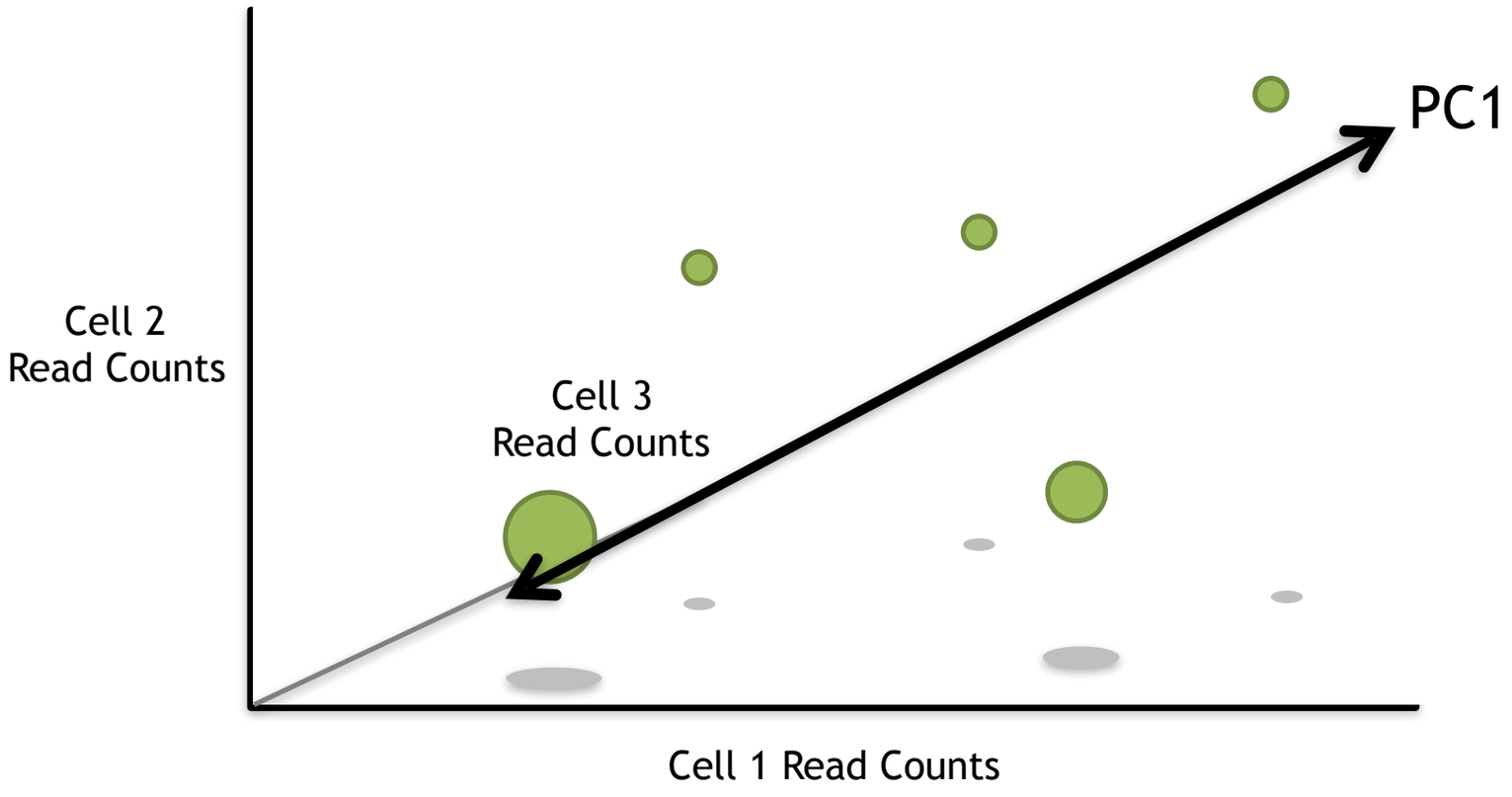


- PC1 captures the direction where most of the variation is.
- PC2 captures the direction with the 2nd most variation.

What if we had 3 cells?

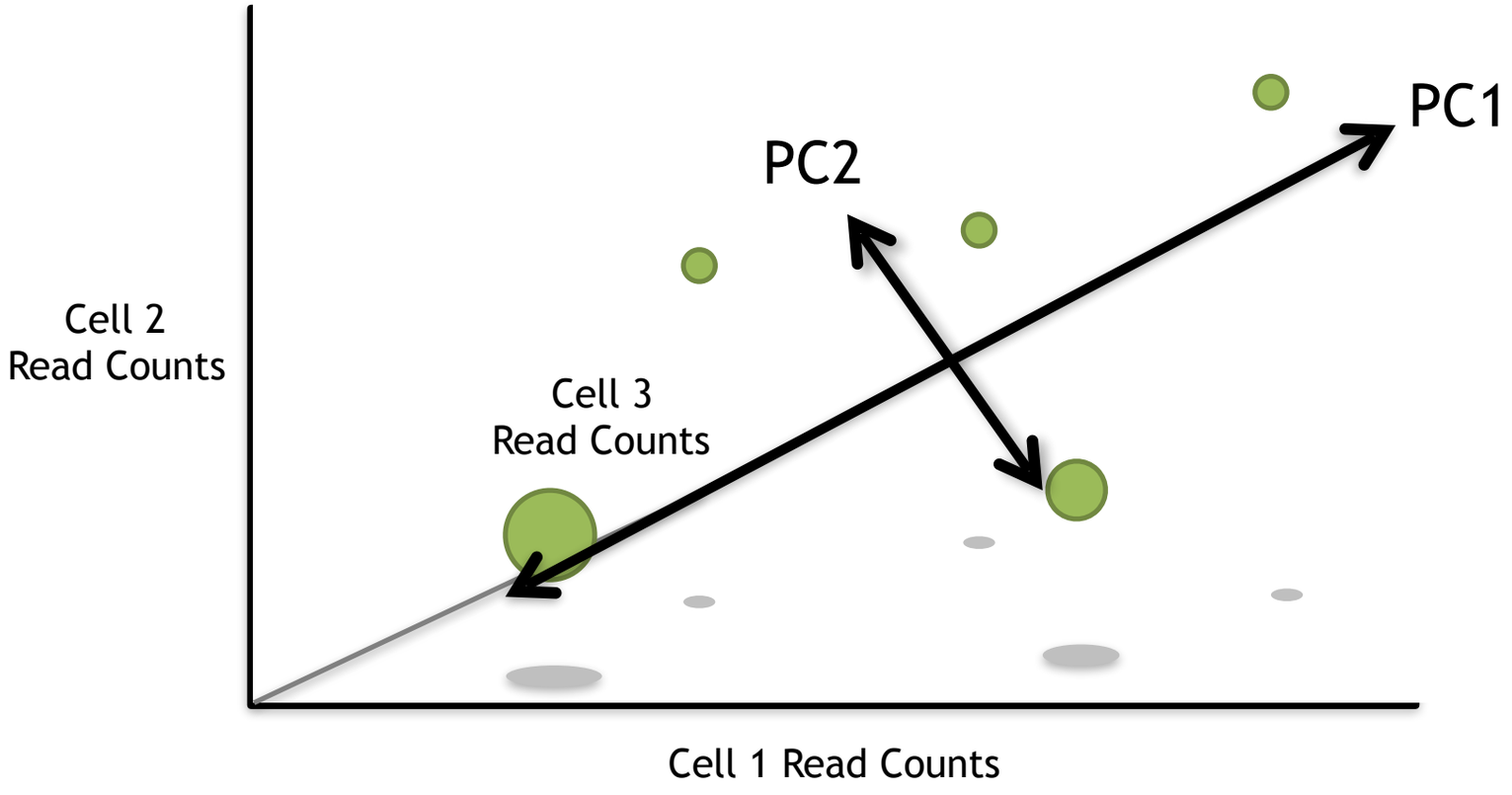


What if we had 3 cells?



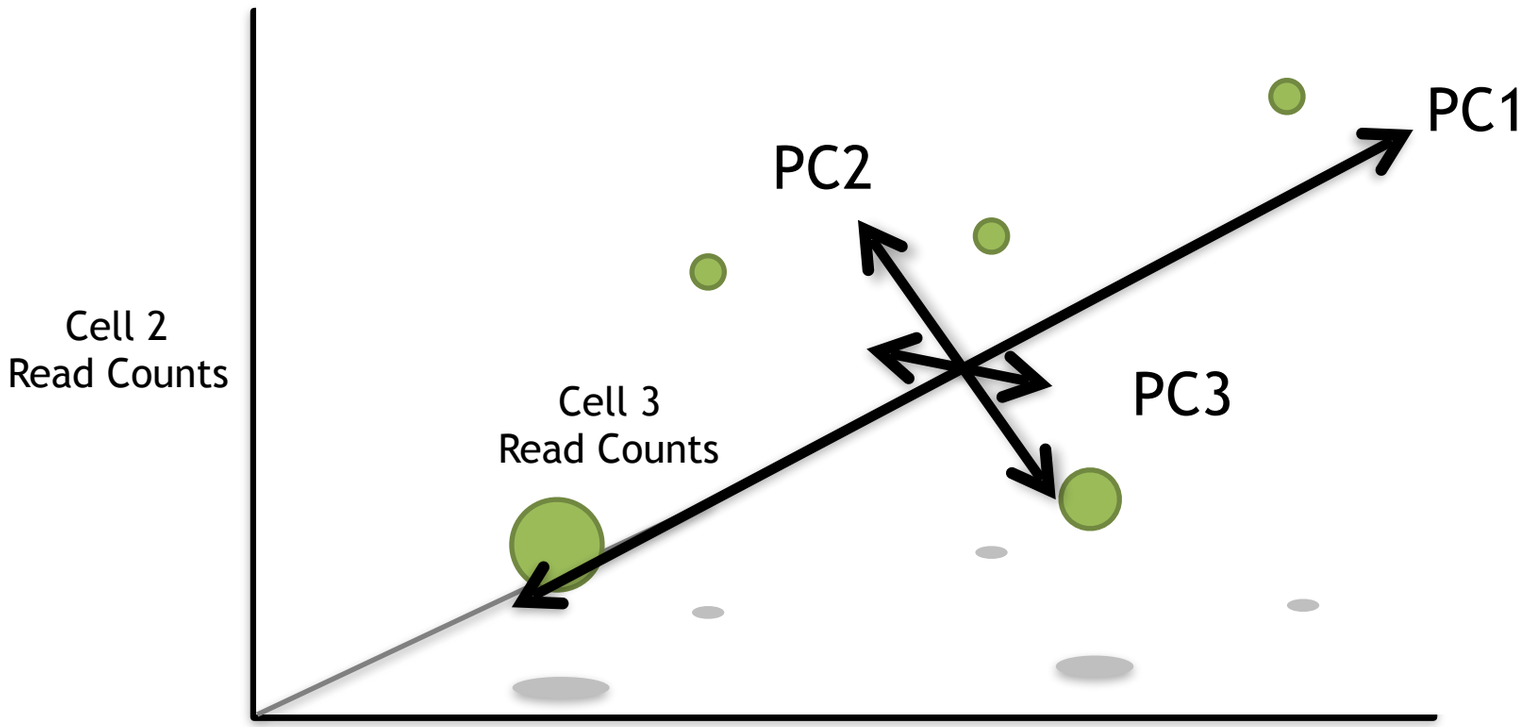
Just like before, PC1 would span the direction of the most variation.

What if we had 3 cells?



Just like before, PC1 would span the direction of the most variation.
PC2 would span the direction of the 2nd most variation.

What if we had 3 cells?



Just like before, PC1 would span the direction of the most variation.
PC2 would span the direction of the 2nd most variation.
However, since we have another direction we can have variation, we need another PC.

PC3 spans the direction of the 3rd most variation.

What if we had 4 cells?

What if we had 4 cells?

- PC1 would span the direction of the most variation.

What if we had 4 cells?

- PC1 would span the direction of the most variation.
- PC2 would span the direction of the 2nd most variation.

What if we had 4 cells?

- PC1 would span the direction of the most variation.
- PC2 would span the direction of the 2nd most variation.
- PC3 would span the direction of the 3rd most variation.

What if we had 4 cells?

- PC1 would span the direction of the most variation.
- PC2 would span the direction of the 2nd most variation.
- PC3 would span the direction of the 3rd most variation.
- PC4 would span the direction of the 4th most variation.

What if we had 4 cells?

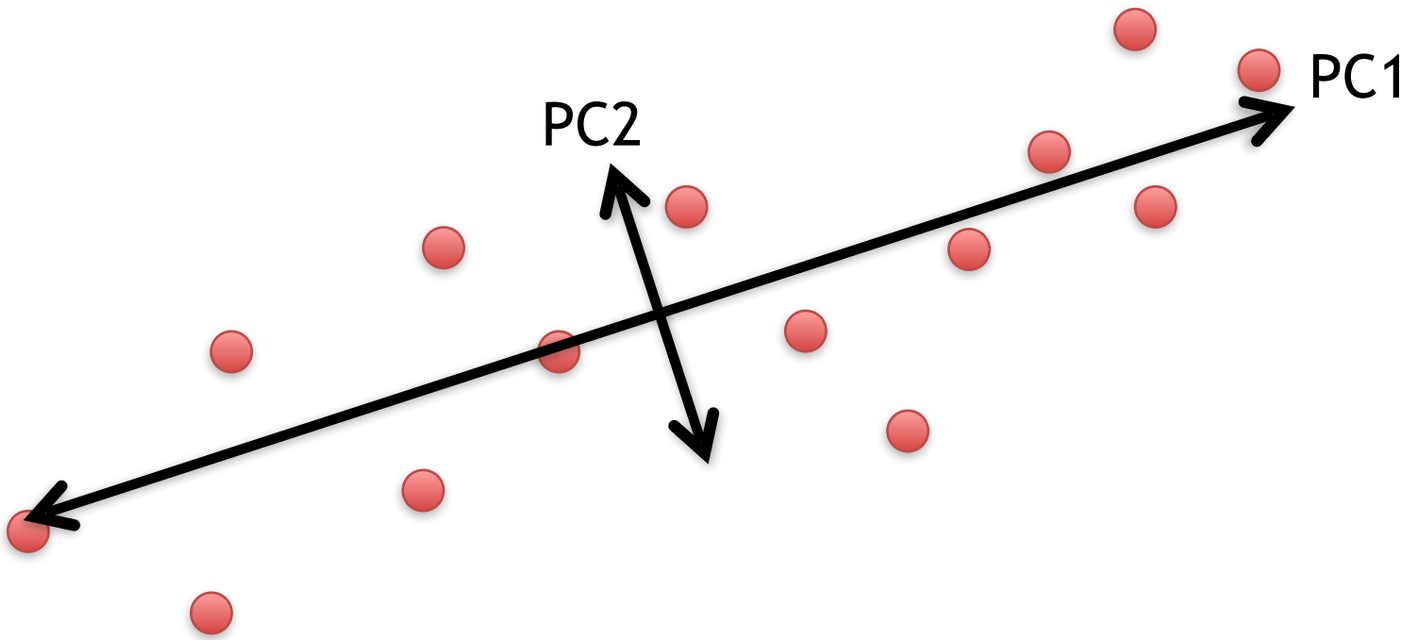
- PC1 would span the direction of the most variation.
- PC2 would span the direction of the 2nd most variation.
- PC3 would span the direction of the 3rd most variation.
- PC4 would span the direction of the 4th most variation.

There is a principal component for each dimension (cell).

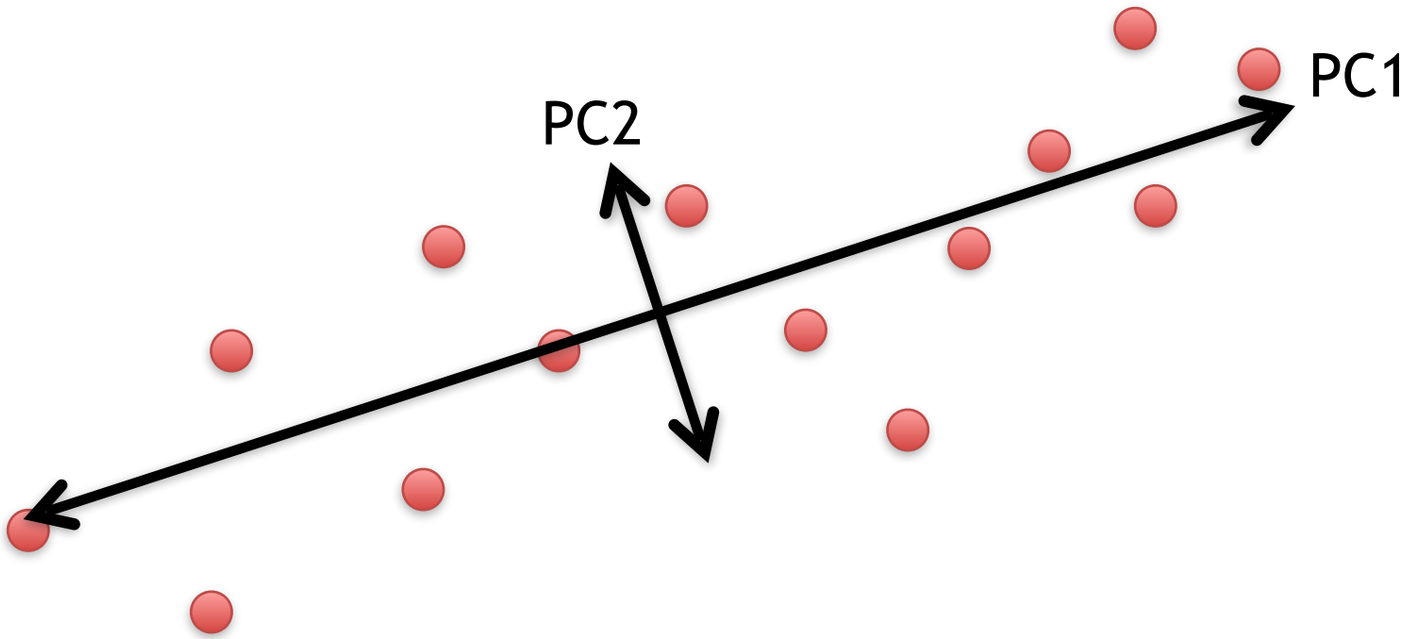
If we had 200 cells, we would have 200 principal components.

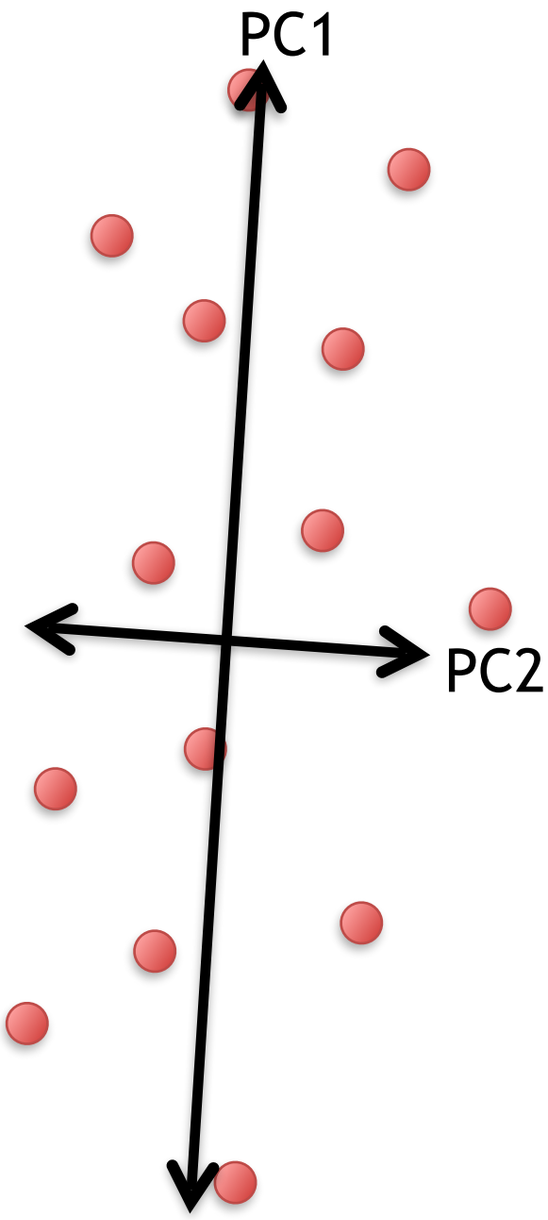
PC200 would span the direction of the 200th most variation.

Examples of PCs

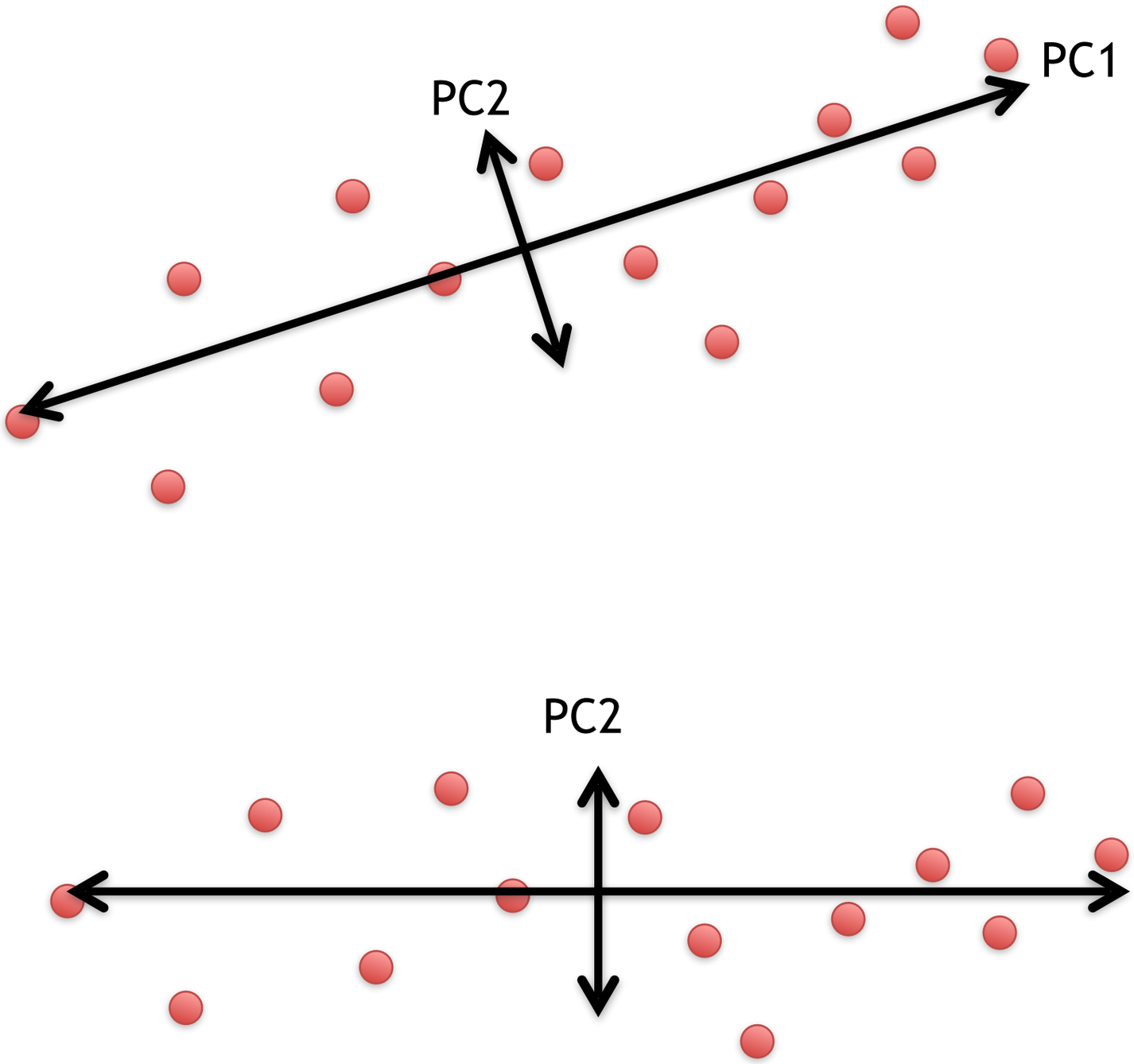


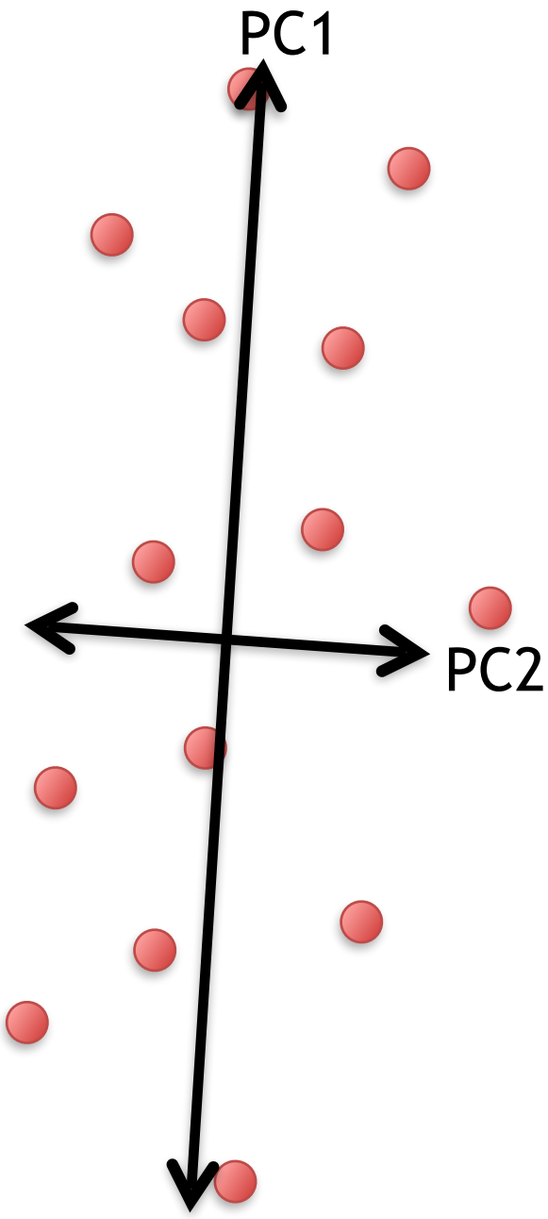
Examples of PCs



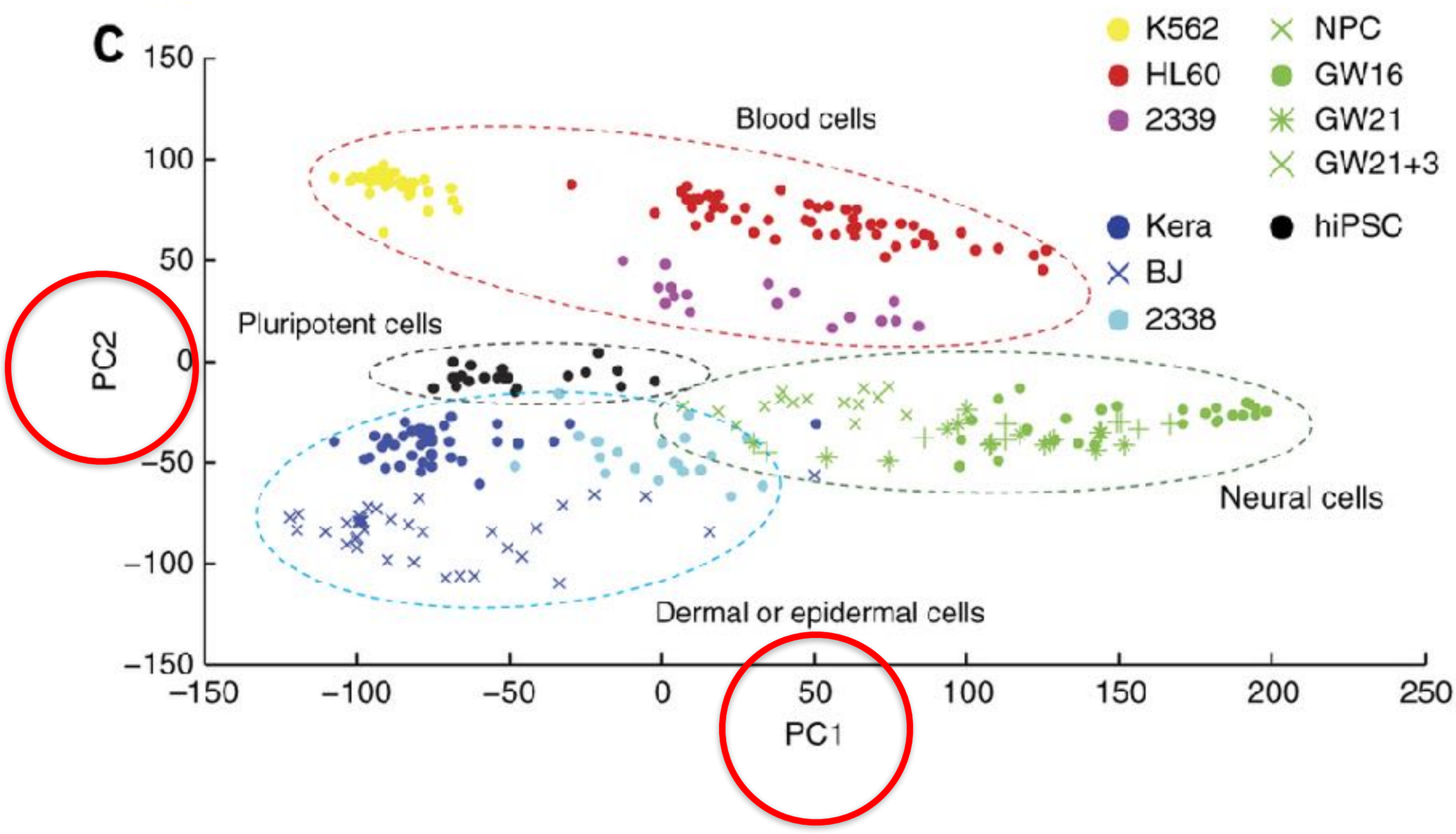


Examples of PCs

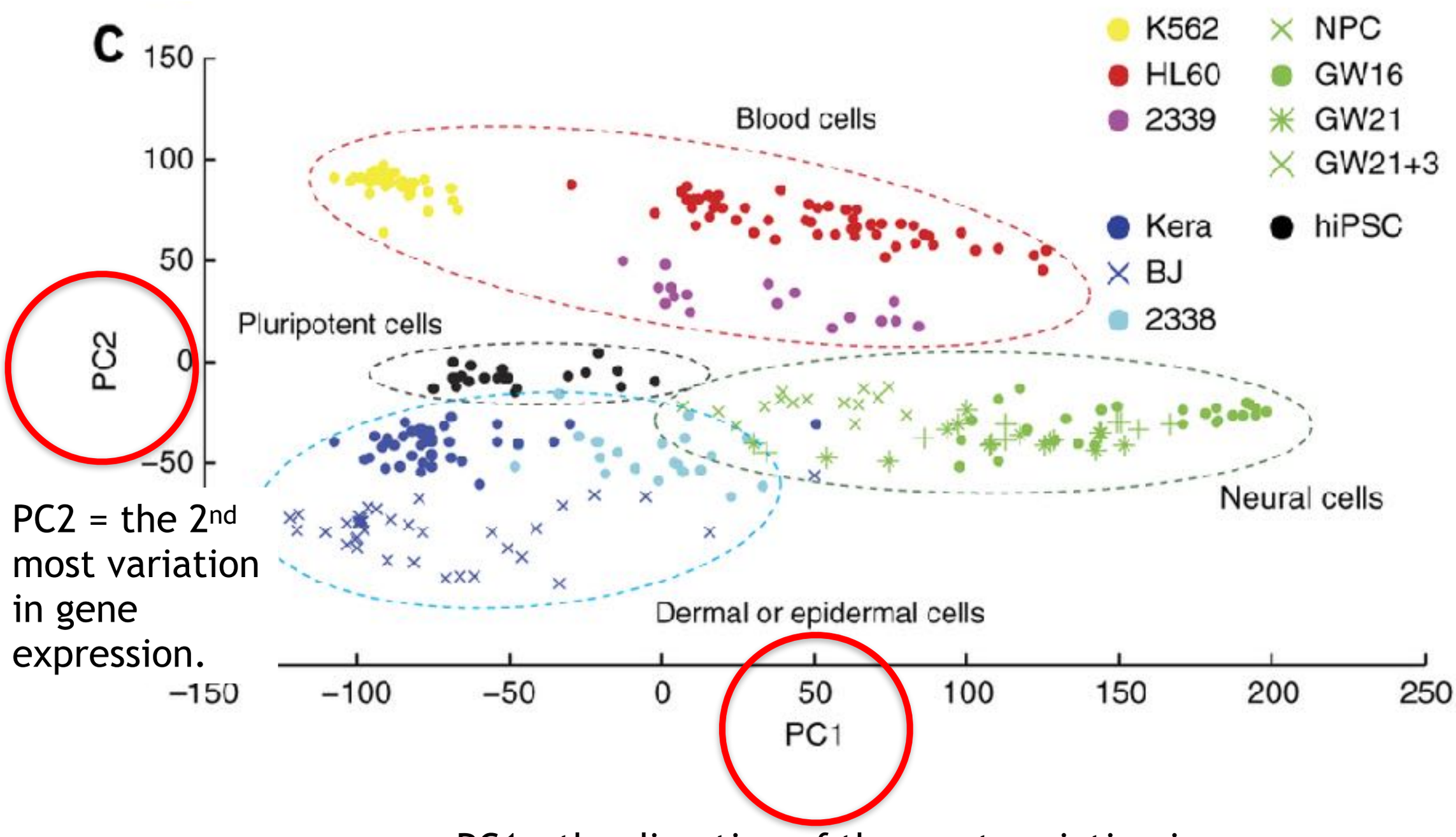




Hooray! We know what the X and Y axis are in this figure!!!



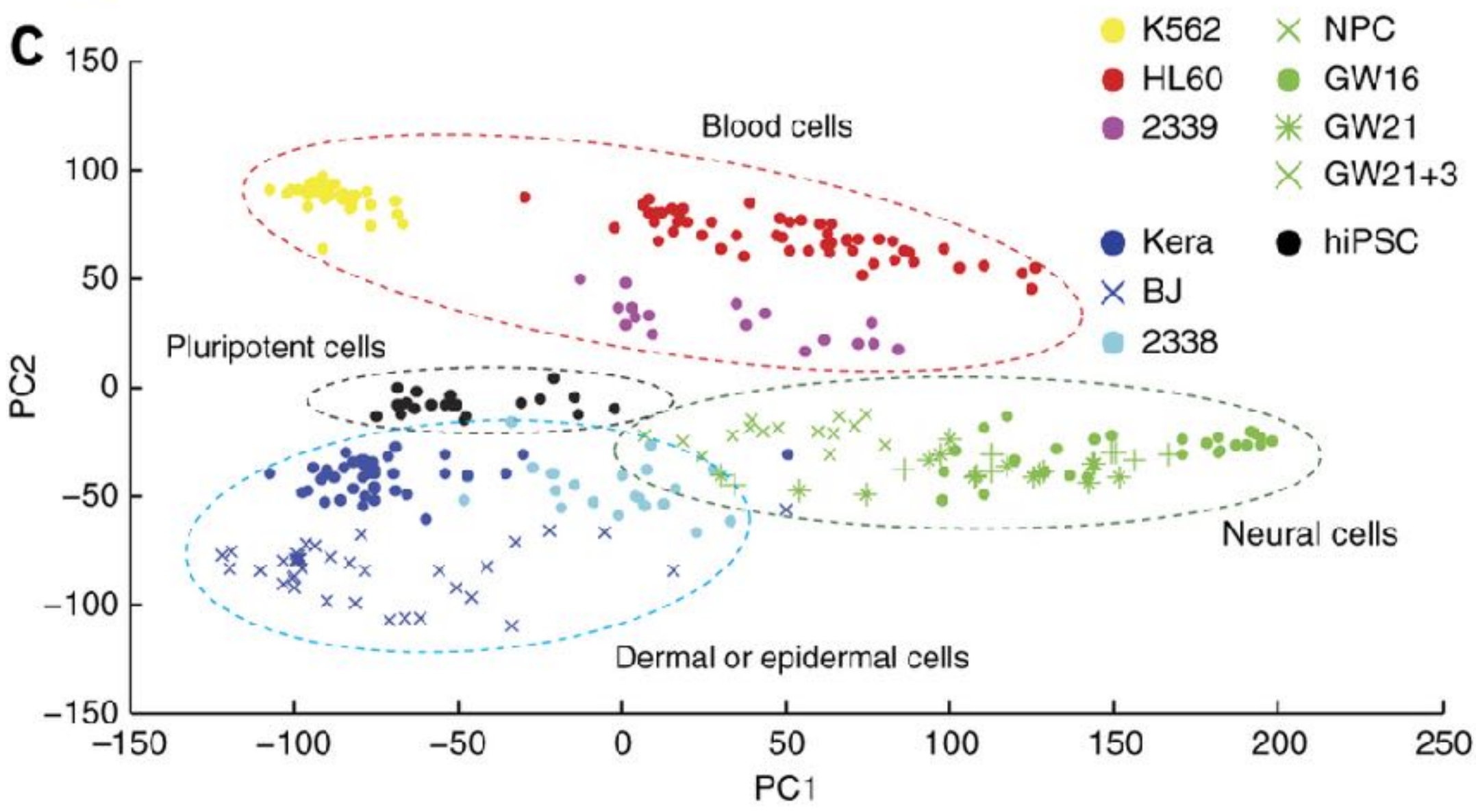
Hooray! We know what the X and Y axis are in this figure!!!



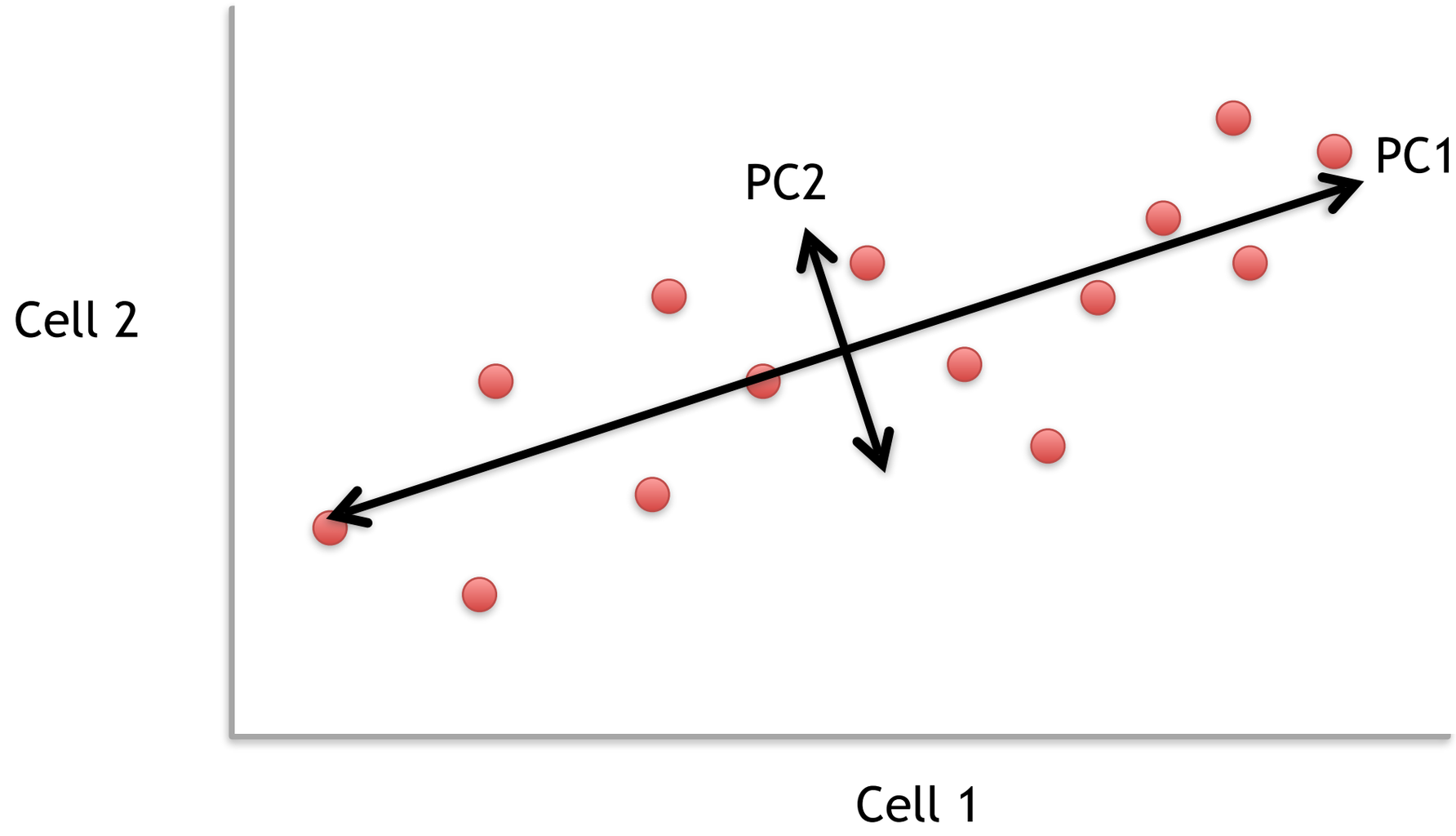
PC2 = the 2nd most variation in gene expression.

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

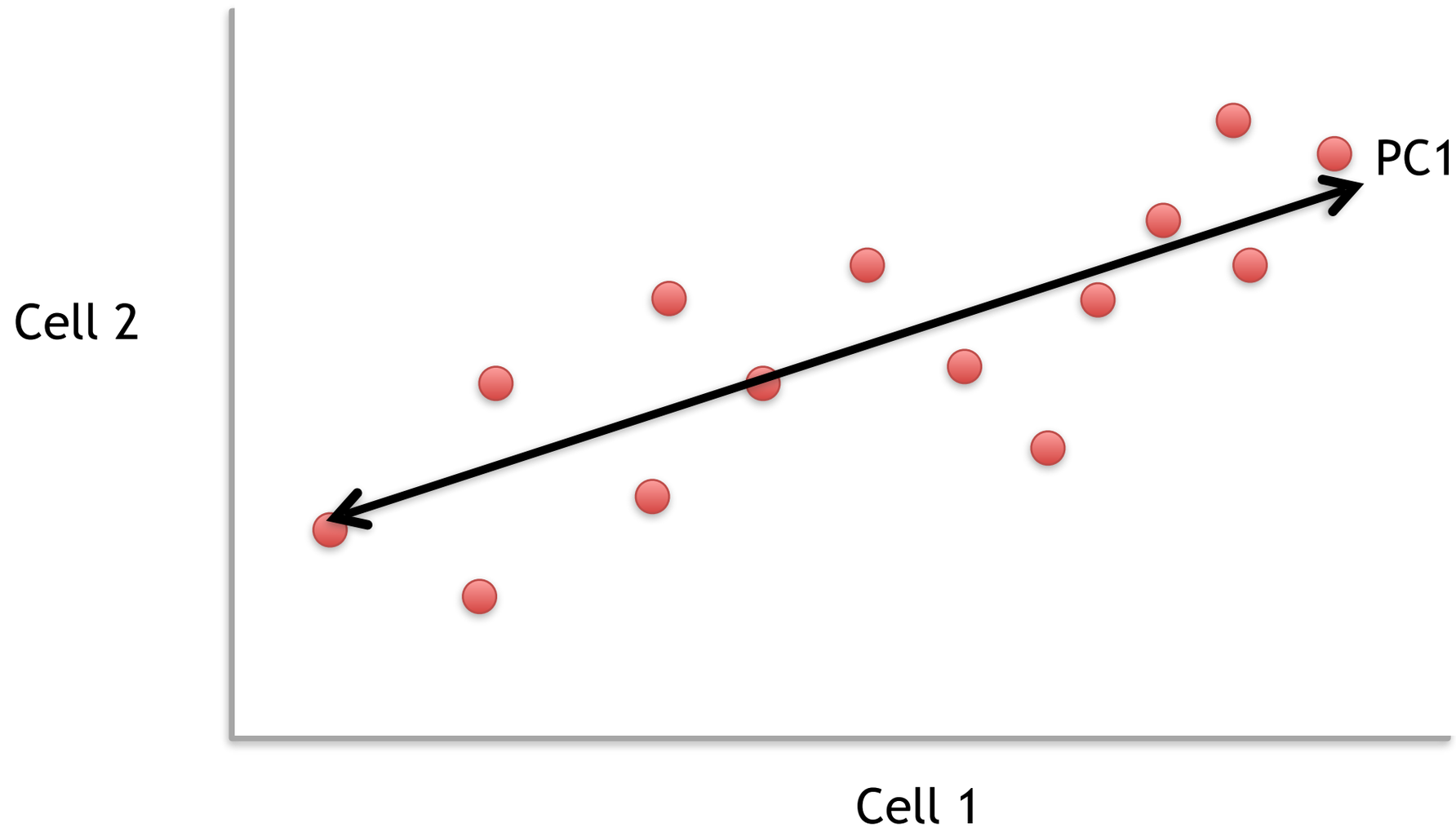
But this is a plot of cells, not genes? How do we plot cells?



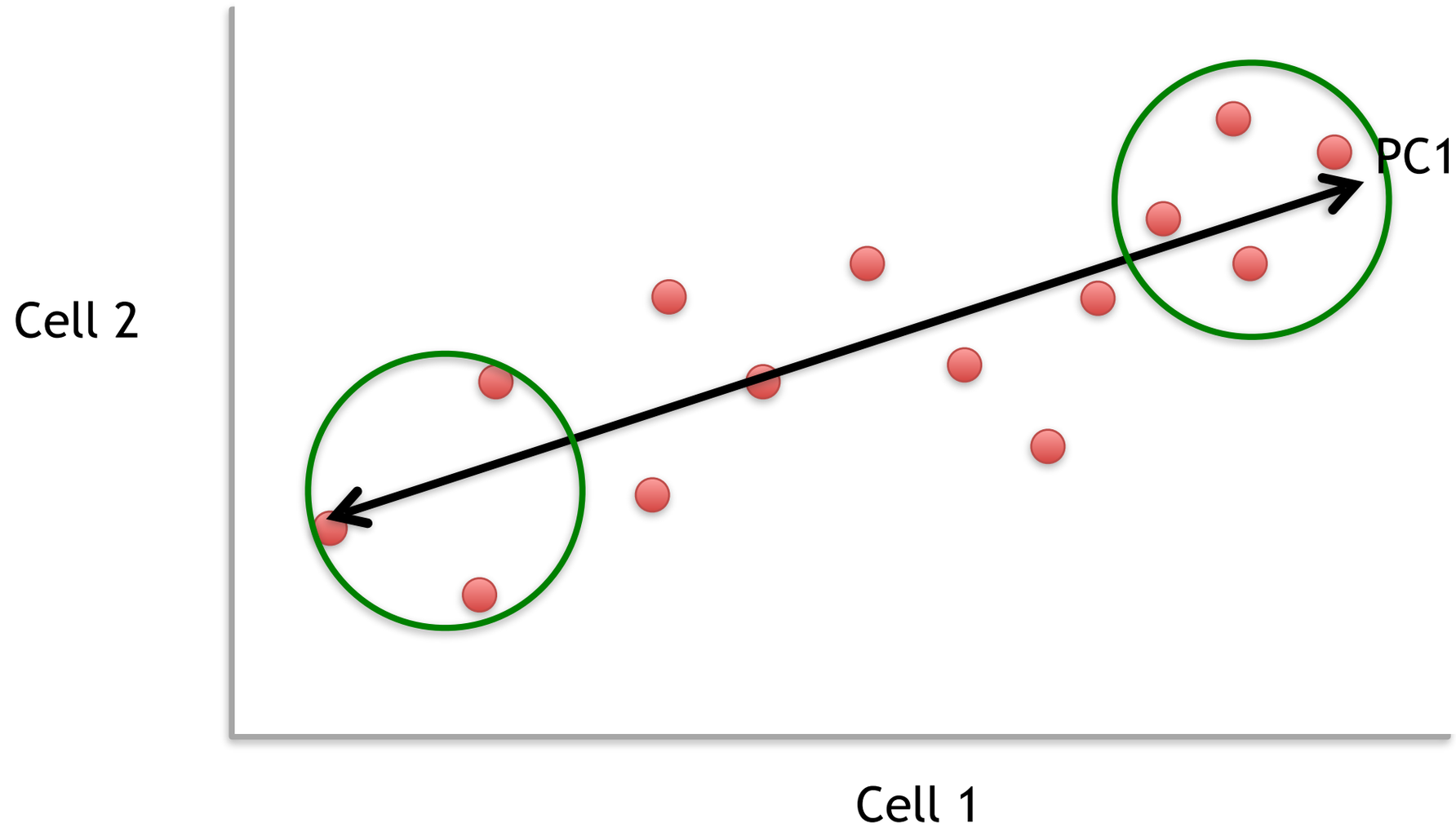
Back to the original scatter plot…



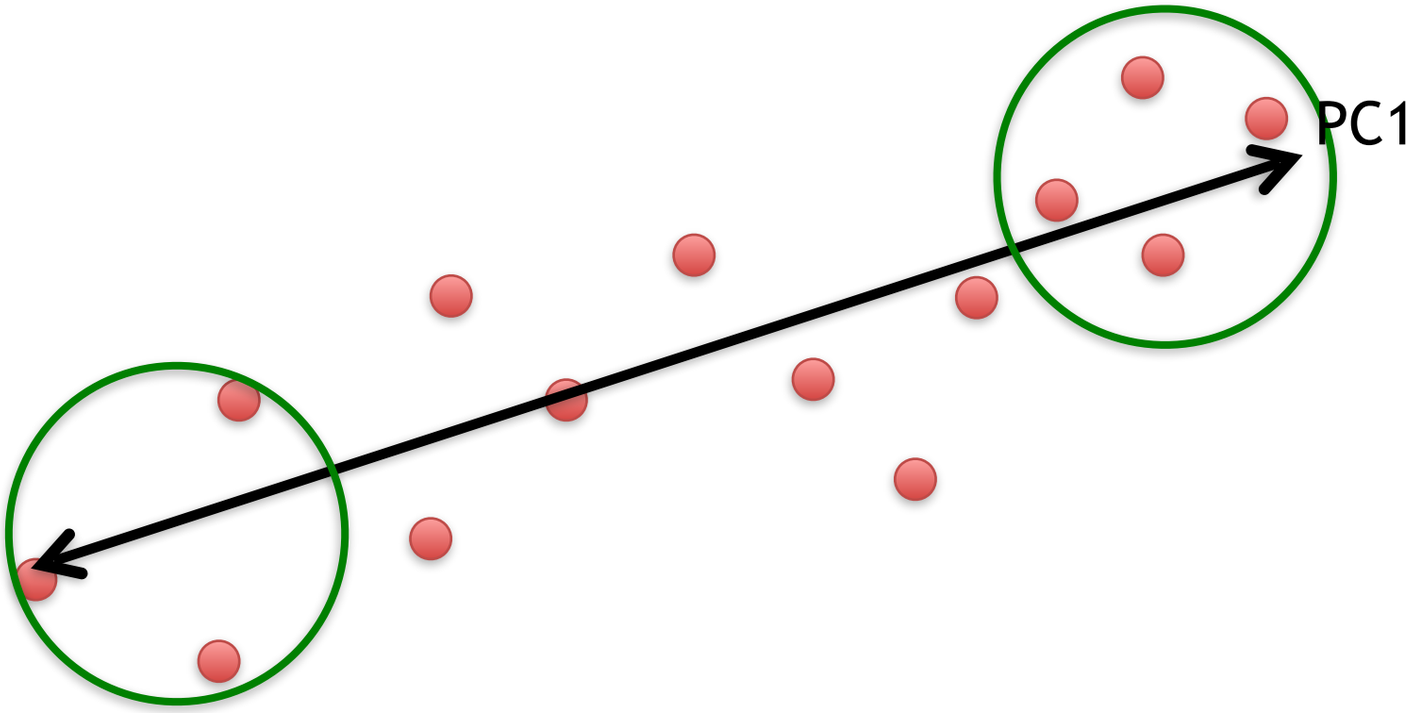
For now, let's focus on PC1



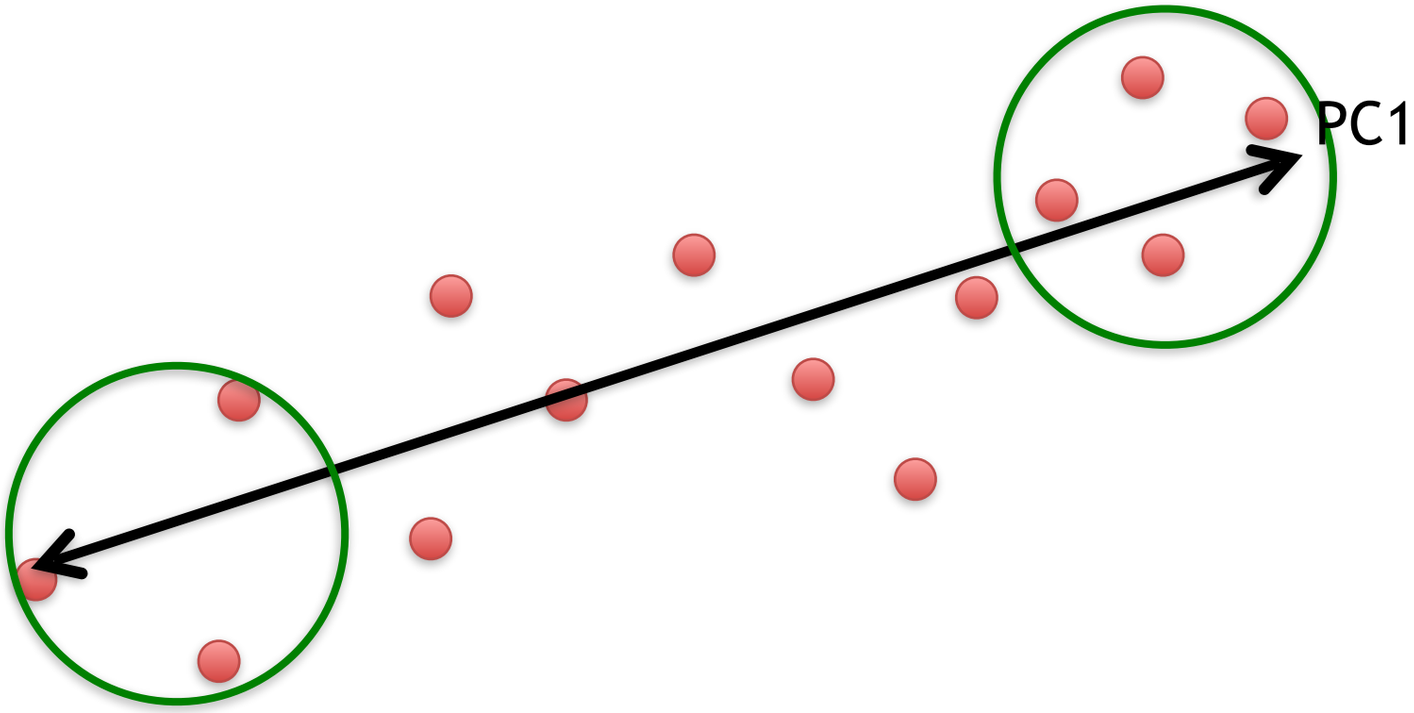
The length and direction of PC1 is mostly determined by the circled genes.



The length and direction of PC1 is mostly determined by the circled genes.

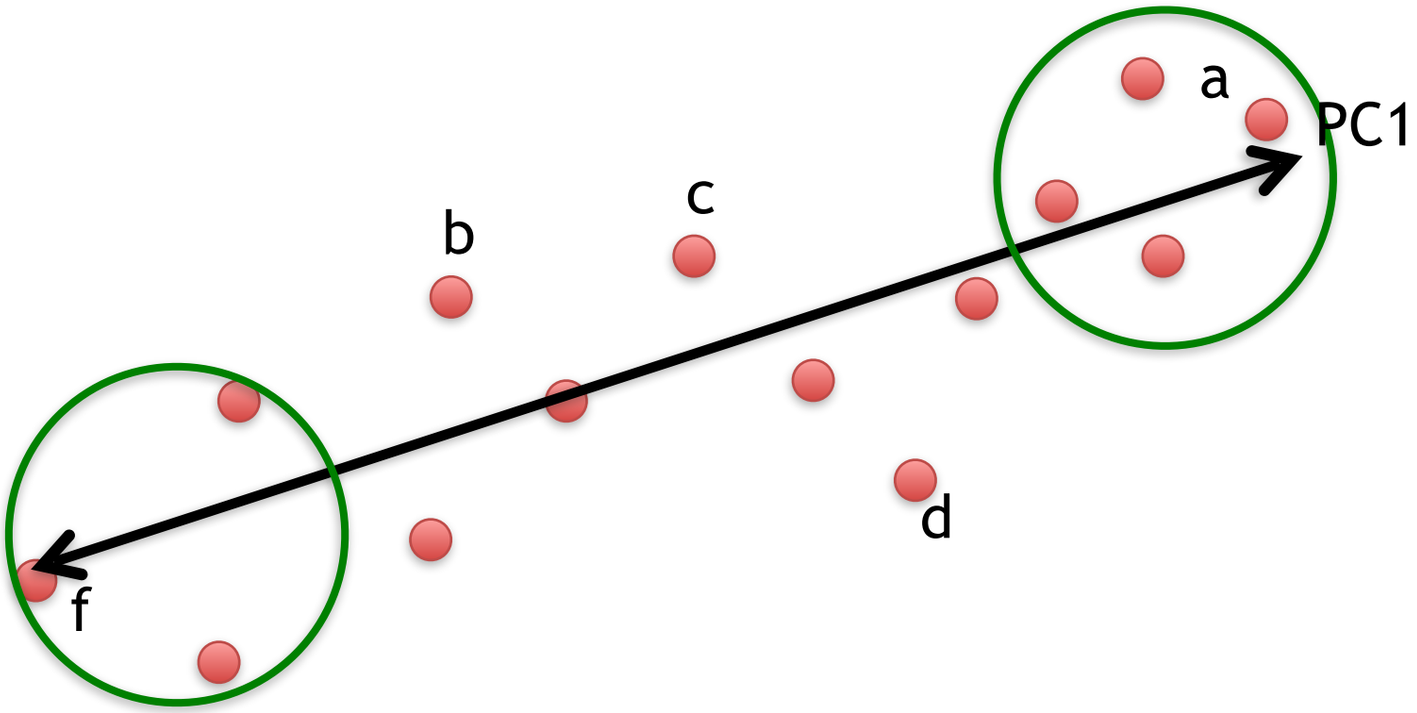


The length and direction of PC1 is mostly determined by the circled genes.



We can score genes based on how much they influence PC1.

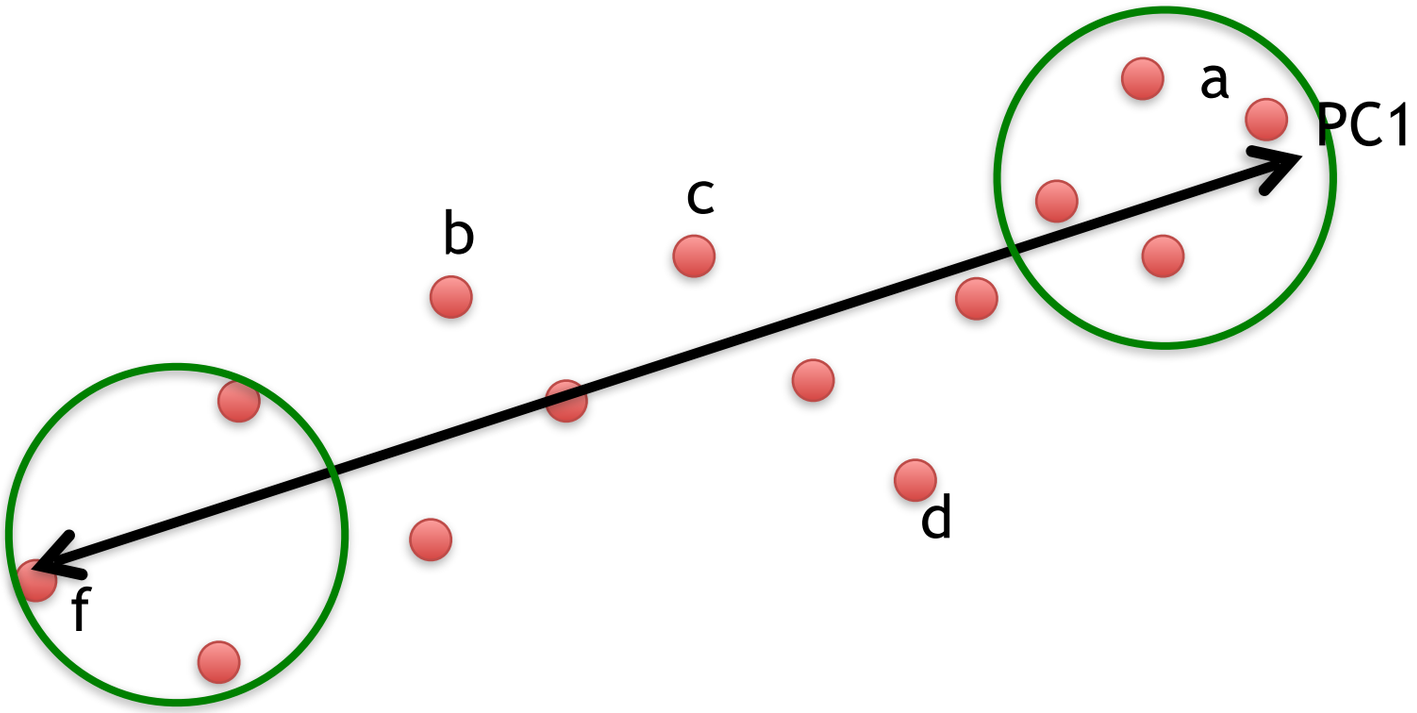
The length and direction of PC1 is mostly determined by the circled genes.



We can score genes based on how much they influence PC1.

Gene	Influence on PC1
a	high
b	low
c	low
d	low
e	high
f	high
…	…

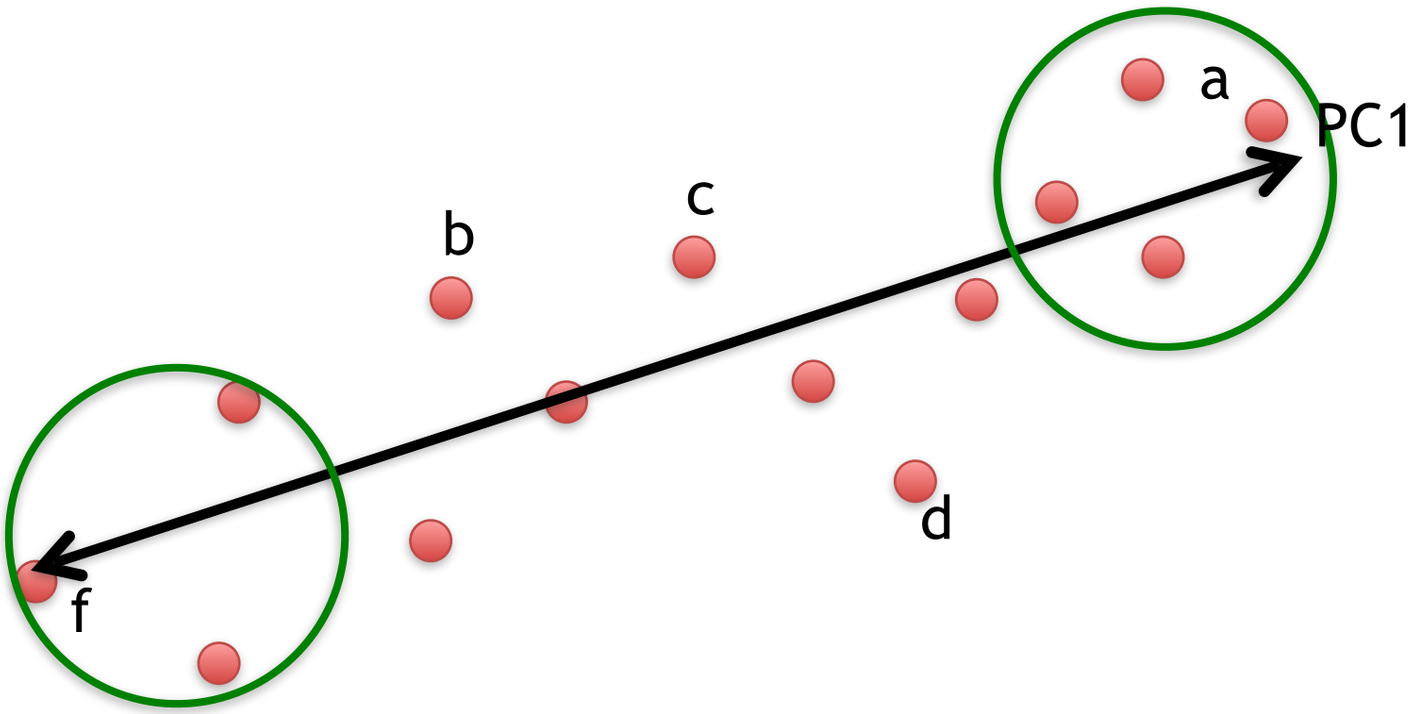
The length and direction of PC1 is mostly determined by the circled genes.



Some genes have more influence on PC1 than others.

Gene	Influence on PC1
a	high
b	low
c	low
d	low
e	high
f	high
…	…

The length and direction of PC1 is mostly determined by the circled genes.



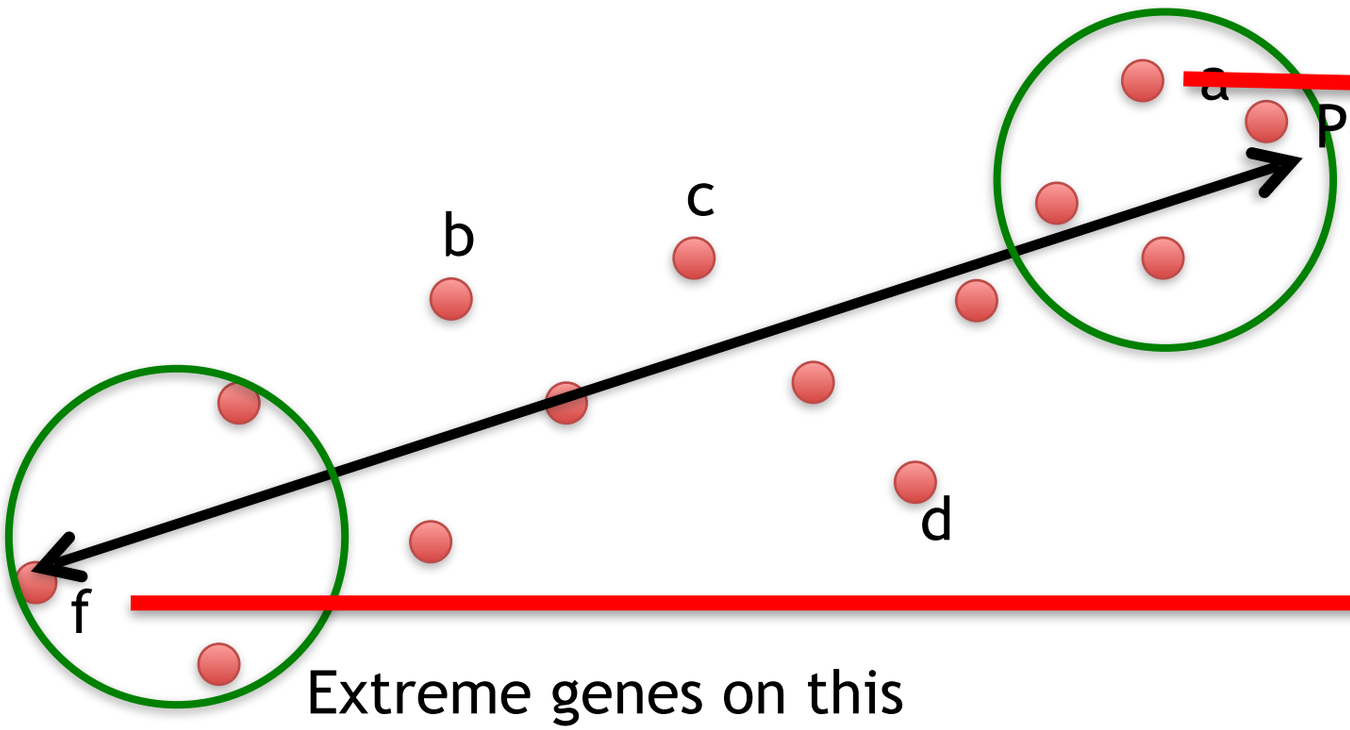
Some genes have more influence on PC1 than others.

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
…	…	

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

Some genes have more influence on PC1 than others.

Extreme genes on this end get large positive numbers…

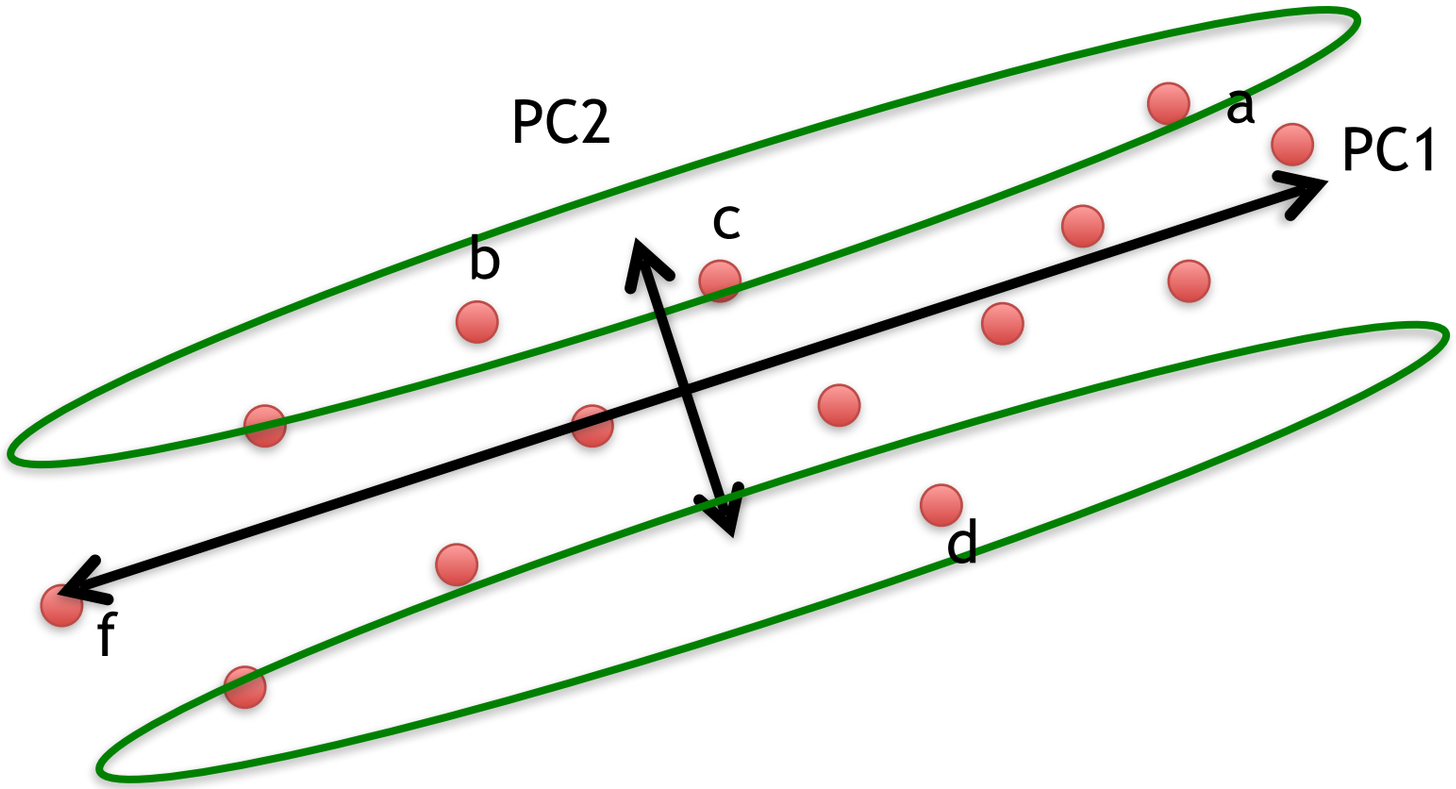


Gene	Influence on PC1	In numbers
a	high	10
b	low	0.5
c	medium	3
d	low	-0.2
e	high	13
f	high	-14
…	…	

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.

Genes that influence 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
…	…	

Our two Principle Components

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
…	…	

Using the two Principle Components to plot cells

Combining the read counts for all genes in a cell to get a single value.

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
…	…	

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
…	…	

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

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) + …

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) + …

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

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) + …

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) + …

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

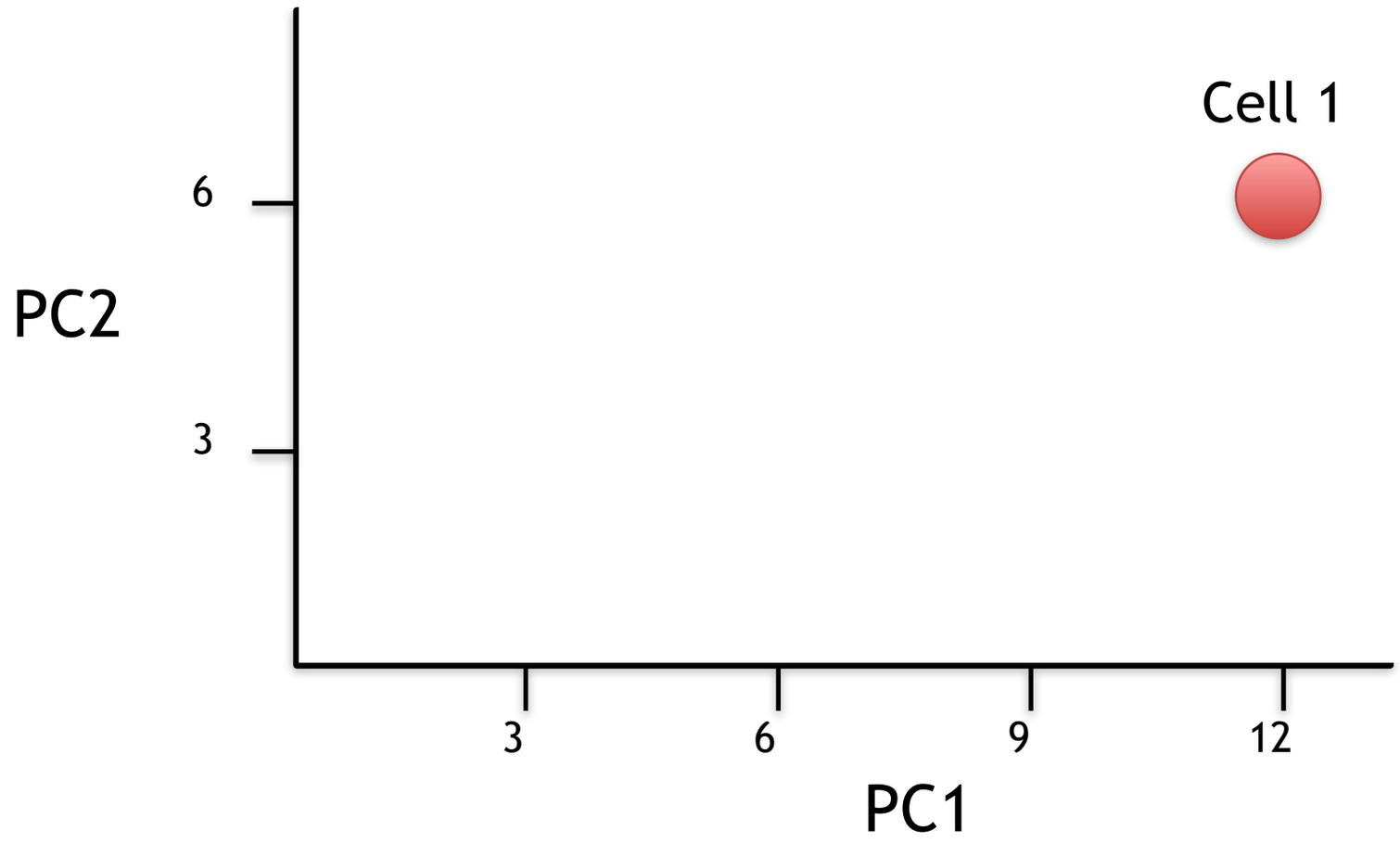
Gene	Influence on PC2	In numbers
a	medium	3
b	high	10
c	high	8
d	high	-12
e	low	0.2
f	low	-0.1
…	…	

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

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

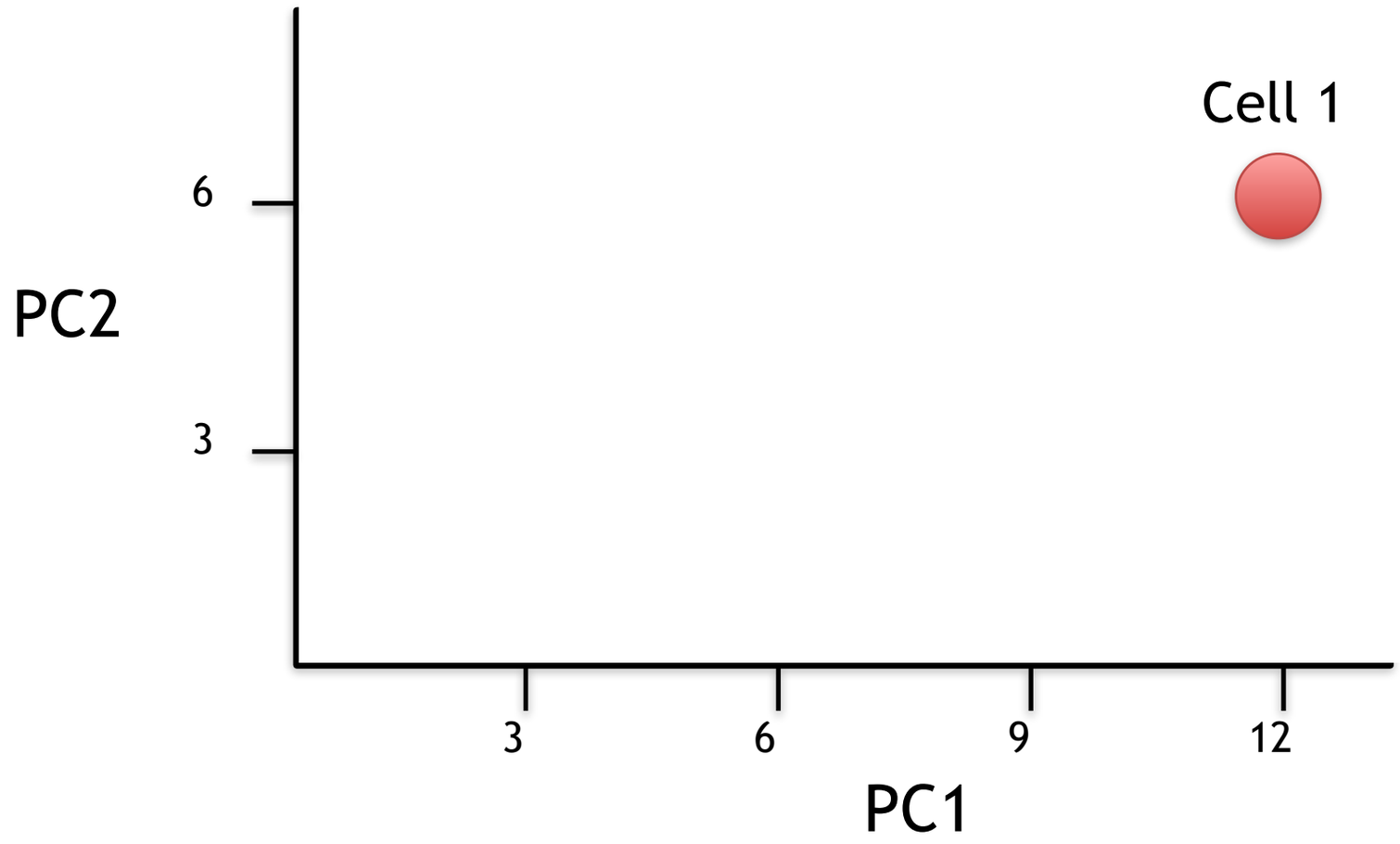
Cell1 PC1 score = (10 * 10) + (0 * 0.5) + … etc… = 12

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

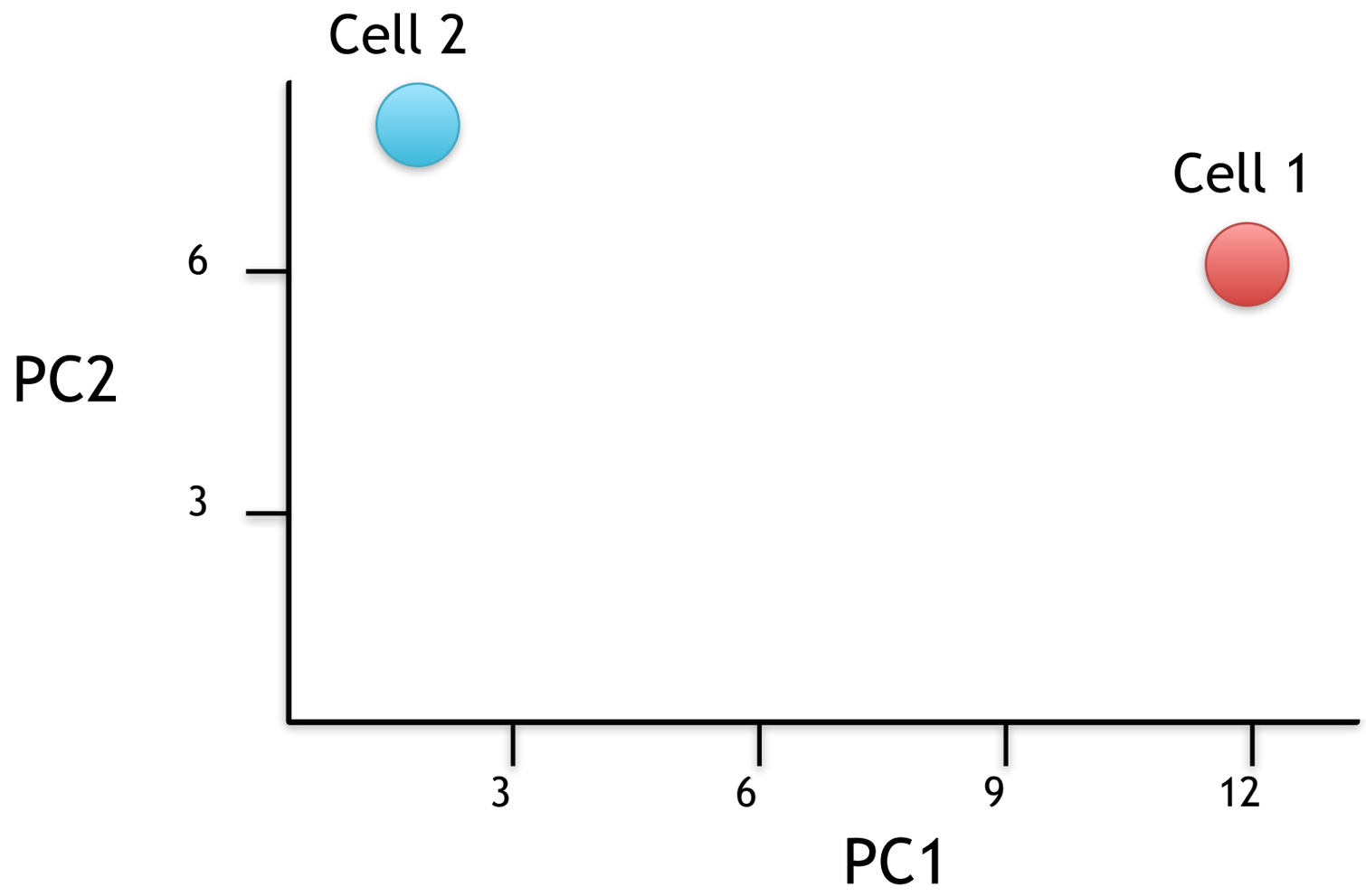


Cell1 PC1 score = (10 * 10) + (0 * 0.5) + … etc… = 12

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



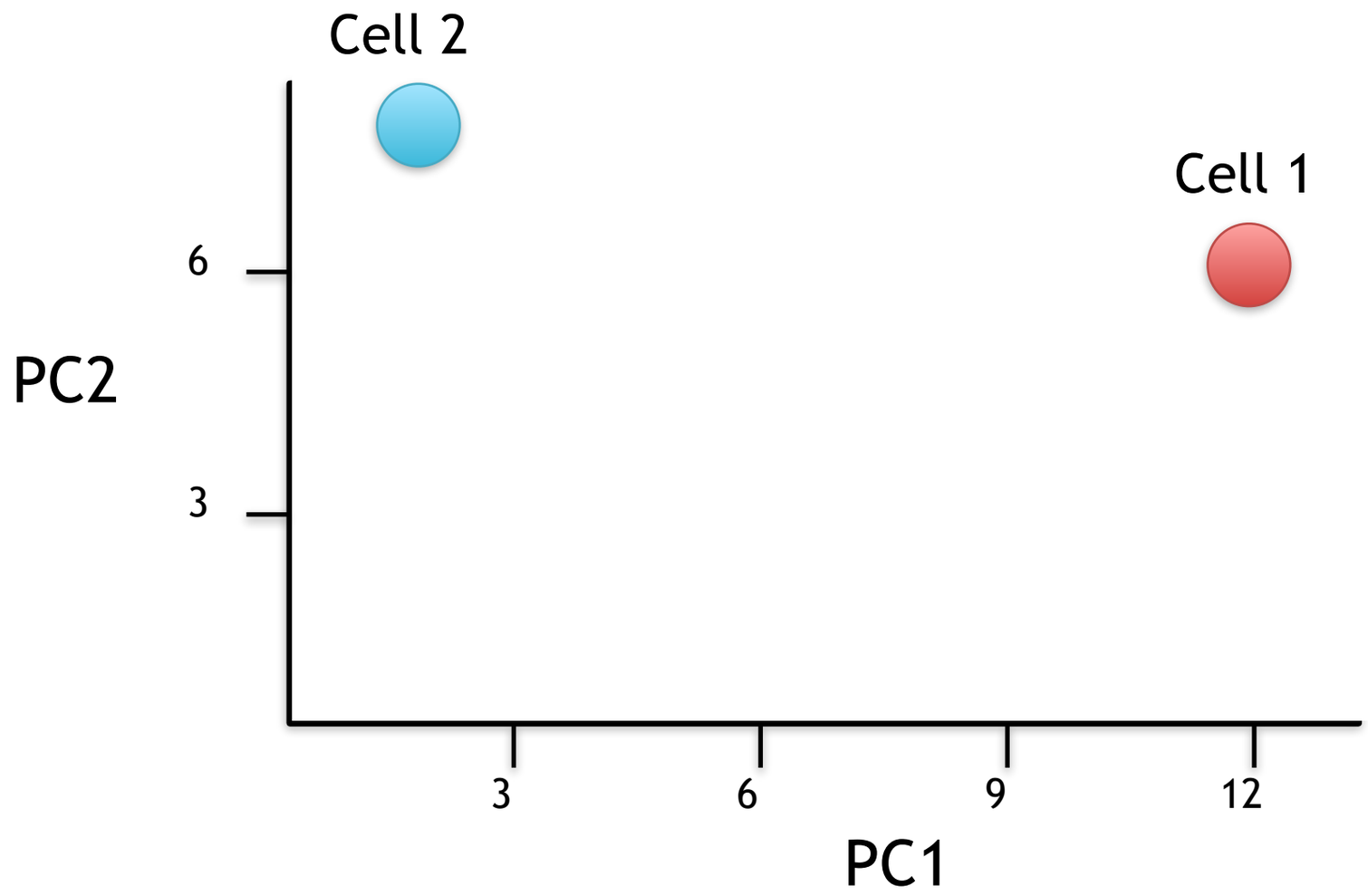
Now calculate scores for Cell2



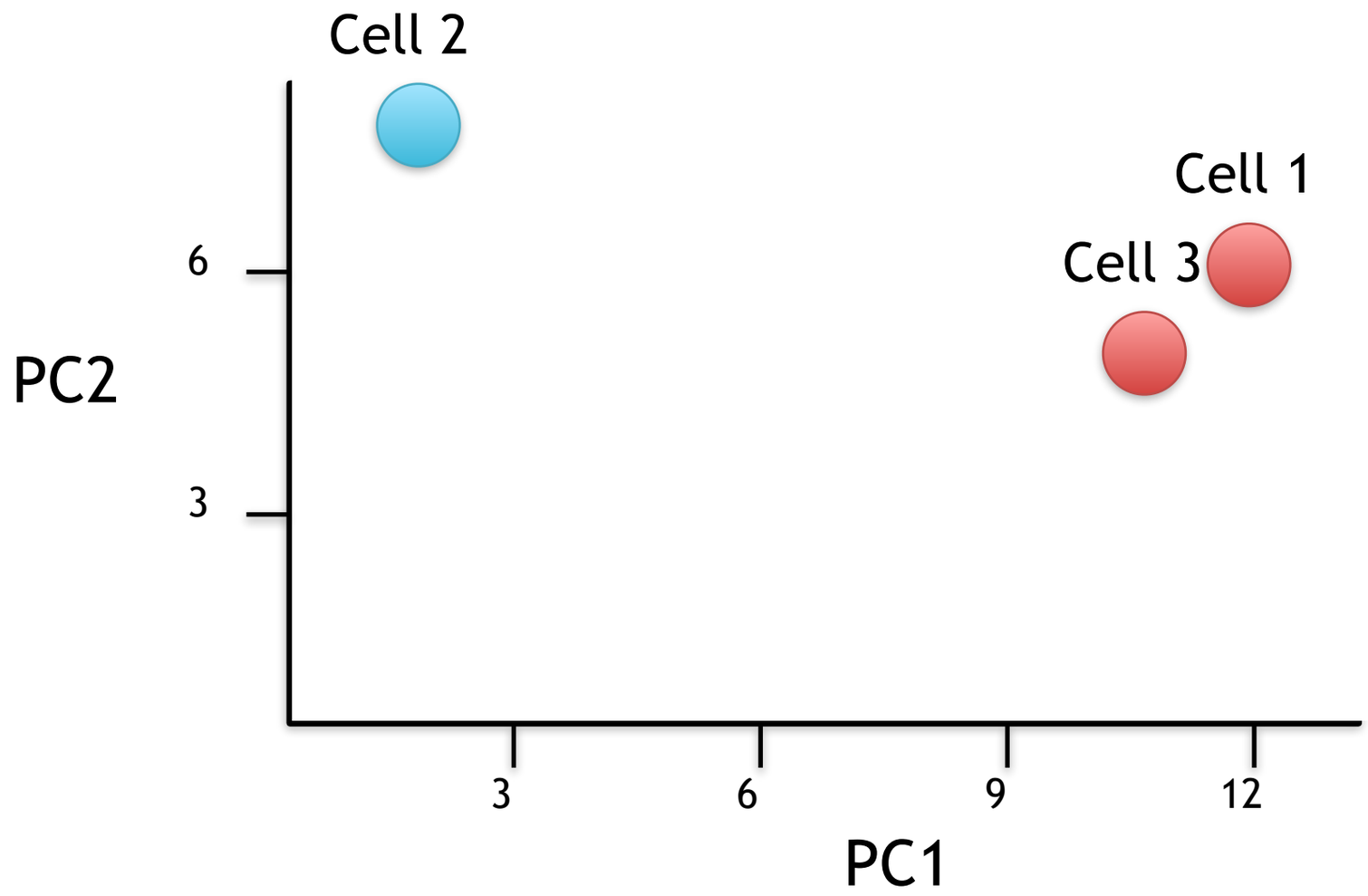
Now calculate scores for Cell2

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

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



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



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

Hooray! We know how they plotted all of the cells!!!