BIMM 194

Genomics, Big Data & Human Health

Barry Grant

UC San Diego

http://thegrantlab.org/bimm194

Today's Menu

Methods Recap:

- **Principal component analysis** (PCA), Network analysis, and Heatmaps

Missing Discussion Topics:

- How much data can be gathered and how can this data help?
- Who controls your 'omic' information?
- Can your omics data be used against you?
- Who pays for genome sequencing in health care?
- Will sequencing make health care cheaper?
- Will people act on genomic information?

Today's Menu

Methods Recap:

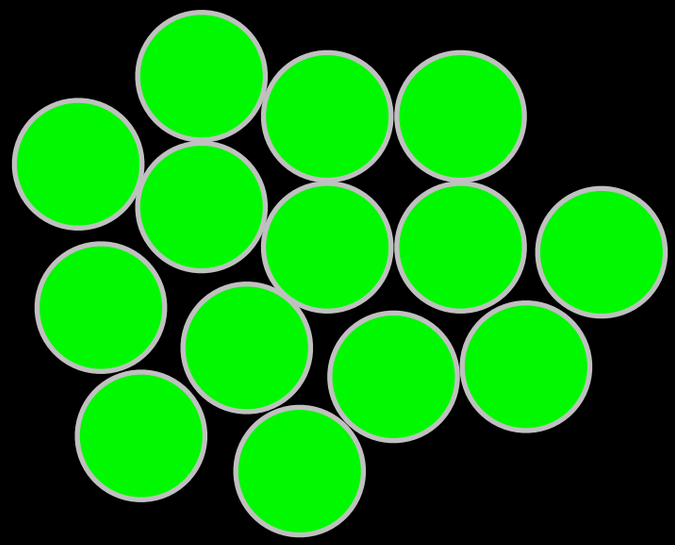
- **Principal component analysis** (PCA), Network analysis, and Heatmaps

Missing Discussion Topics:

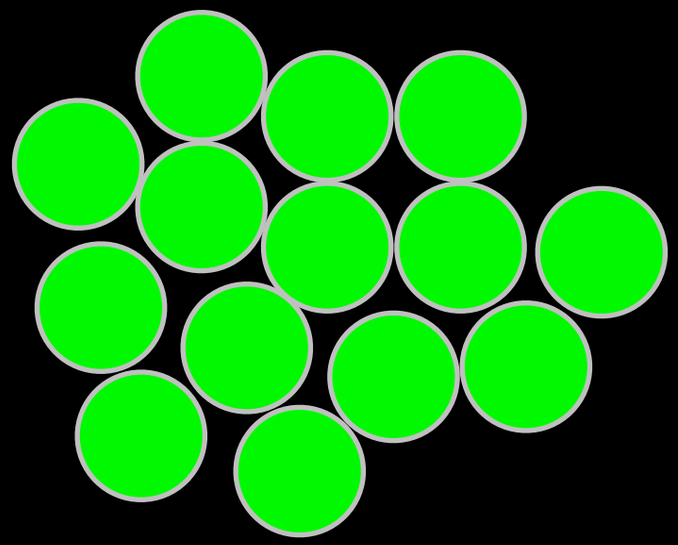
- How much data can be gathered and how can this data help?
- Who controls your 'omic' information?
- Can your omics data be used against you?
- Who pays for genome sequencing in health care?
- Will sequencing make health care cheaper?
- Will people act on genomic information?

PCA: The absolute basics

Bunch of “normal” cells

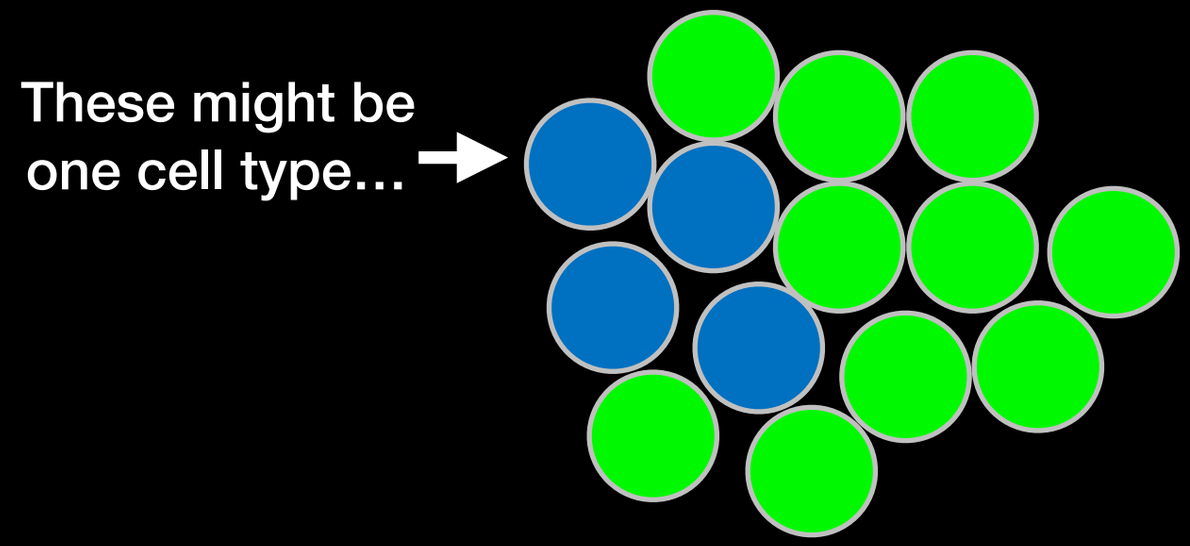


Bunch of “normal” cells



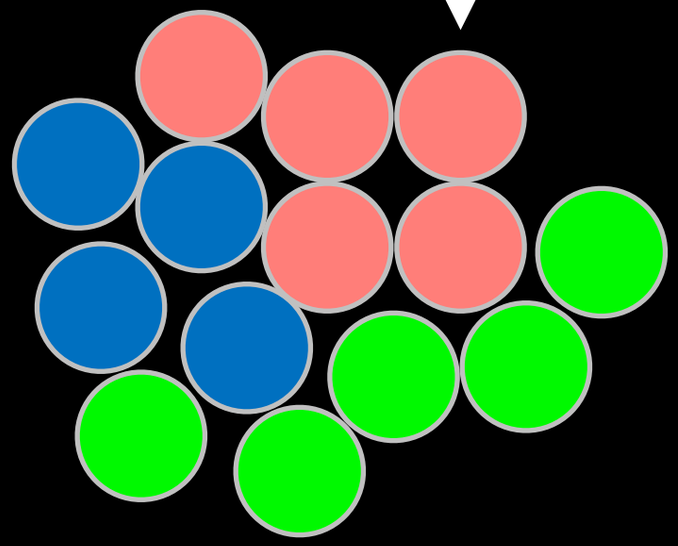
Even though they look the same we suspect that there are differences…

These might be one cell type…



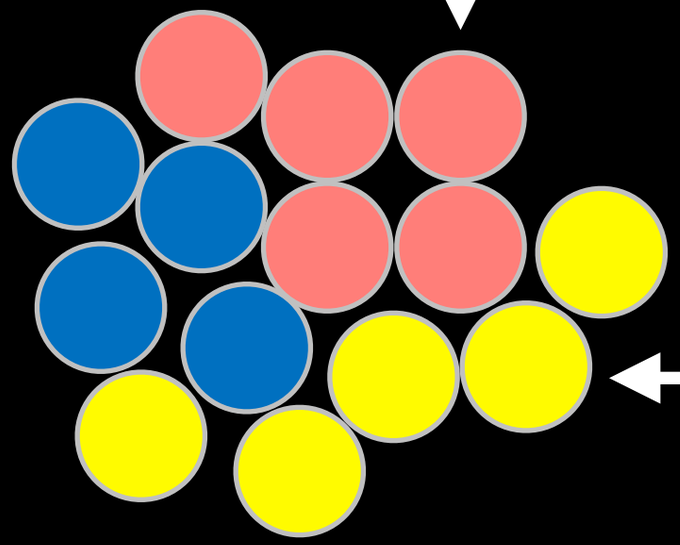
These another
cell type…

These might be
one cell type…

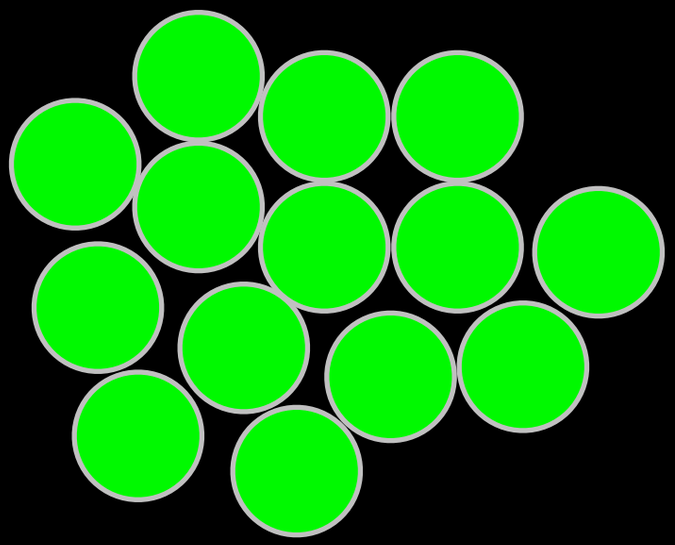


These another cell type…

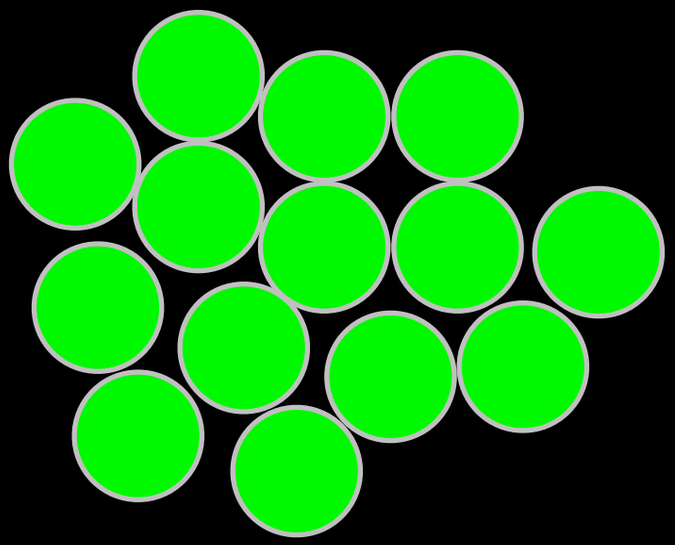
These might be one cell type…



These might be a third cell type…



Unfortunately we can’t observe
the differences visually



Unfortunately we can’t observe
the differences visually

So we sequence the mRNA in each
cell to identify which genes are
active and at what levels.

Here is the data…

	Cell1	Cell2	Cell3	Cell4	…
Gene1	3	0.25	2.8	0.1	…
Gene2	2.9	0.8	2.2	1.8	…
Gene3	2.2	1	1.5	3.2	…
Gene4	2	1.4	2	0.3	…
Gene5	1.3	1.6	1.6	0	…
Gene6	1.5	2	2.1	3	…
Gene7	1.1	2.2	1.2	2.8	…
Gene8	1	2.7	0.9	0.3	…
Gene9	0.4	3	0.6	0.1	…

Each column shows how much each gene is transcribed in each cell

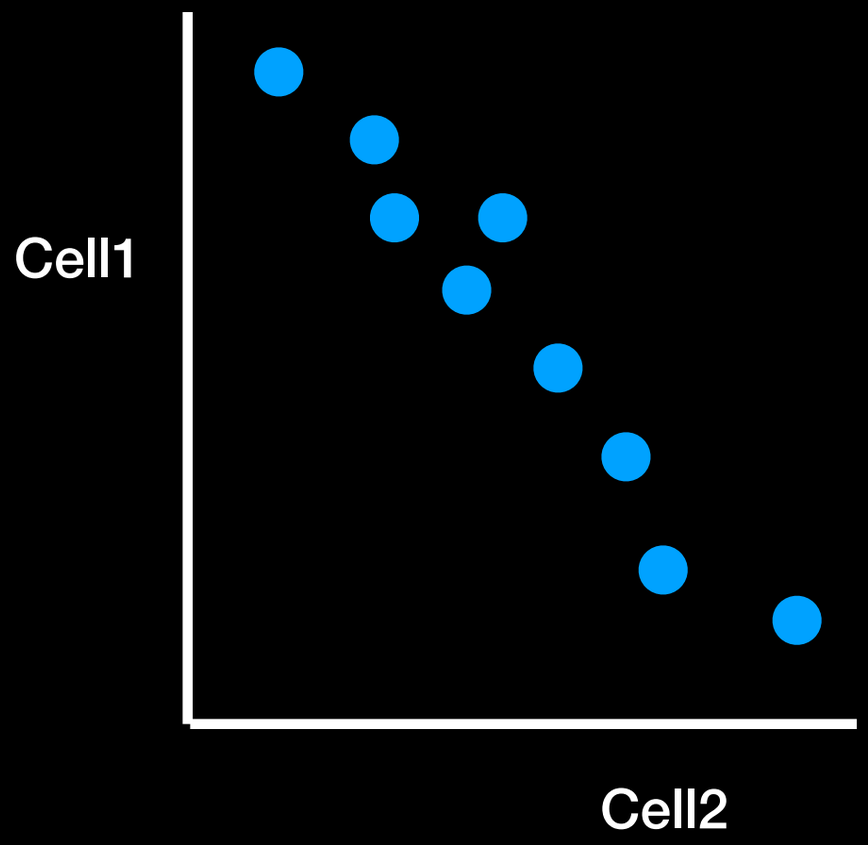
Here is the data…

	Cell1	Cell2	Cell3	Cell4	…
Gene1	3	0.25	2.8	0.1	…
Gene2	2.9	0.8	2.2	1.8	…
Gene3	2.2	1	1.5	3.2	…
Gene4	2	1.4	2	0.3	…
Gene5	1.3	1.6	1.6	0	…
Gene6	1.5	2	2.1	3	…
Gene7	1.1	2.2	1.2	2.8	…
Gene8	1	2.7	0.9	0.3	…
Gene9	0.4	3	0.6	0.1	…

For now lets consider only two cells

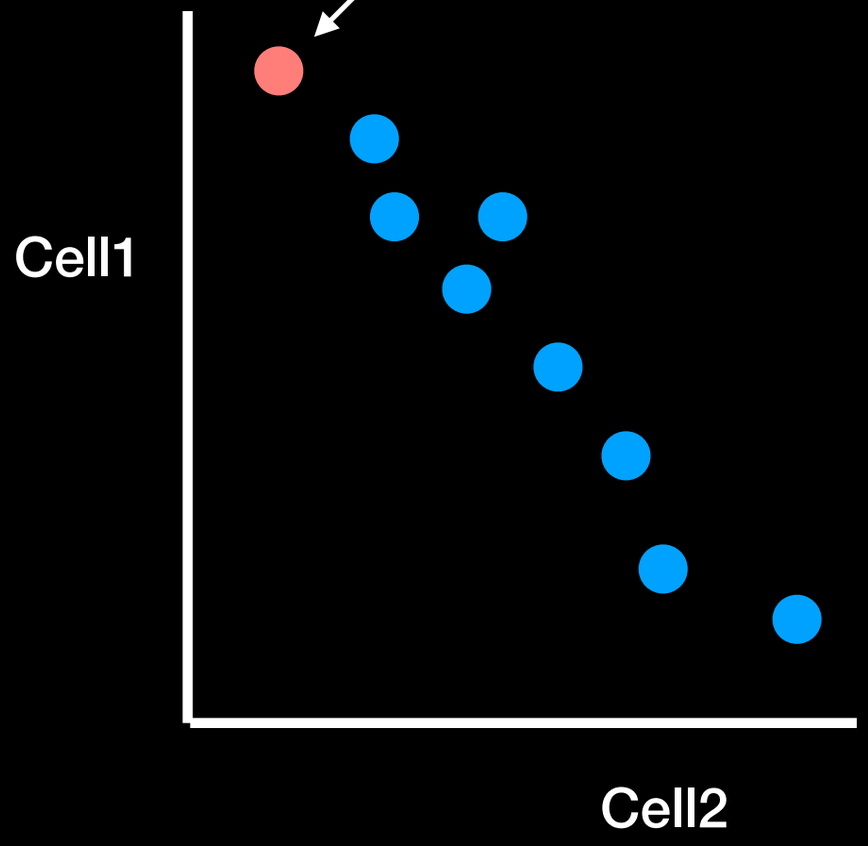
	Cell1	Cell2
Gene1	3	0.25
Gene2	2.9	0.8
Gene3	2.2	1
Gene4	2	1.4
Gene5	1.3	1.6
Gene6	1.5	2
Gene7	1.1	2.2
Gene8	1	2.7
Gene9	0.4	3

We have just 2 cells so we can plot the measurements for each gene



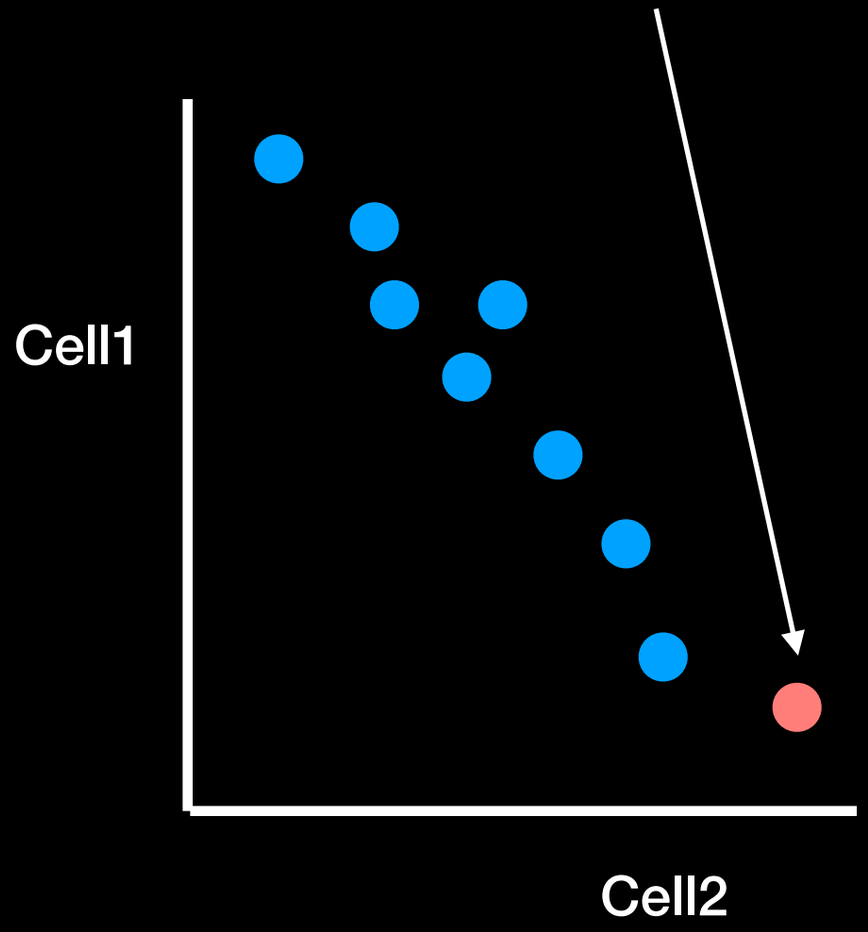
	Cell1	Cell2
Gene1	3	0.25
Gene2	2.9	0.8
Gene3	2.2	1
Gene4	2	1.4
Gene5	1.3	1.6
Gene6	1.5	2
Gene7	1.1	2.2
Gene8	1	2.7
Gene9	0.4	3

This gene (Gene1) is highly transcribed in Cell1 and lowly transcribed in Cell2…



	Cell1	Cell2
Gene1	3	0.25
Gene2	2.9	0.8
Gene3	2.2	1
Gene4	2	1.4
Gene5	1.3	1.6
Gene6	1.5	2
Gene7	1.1	2.2
Gene8	1	2.7
Gene9	0.4	3

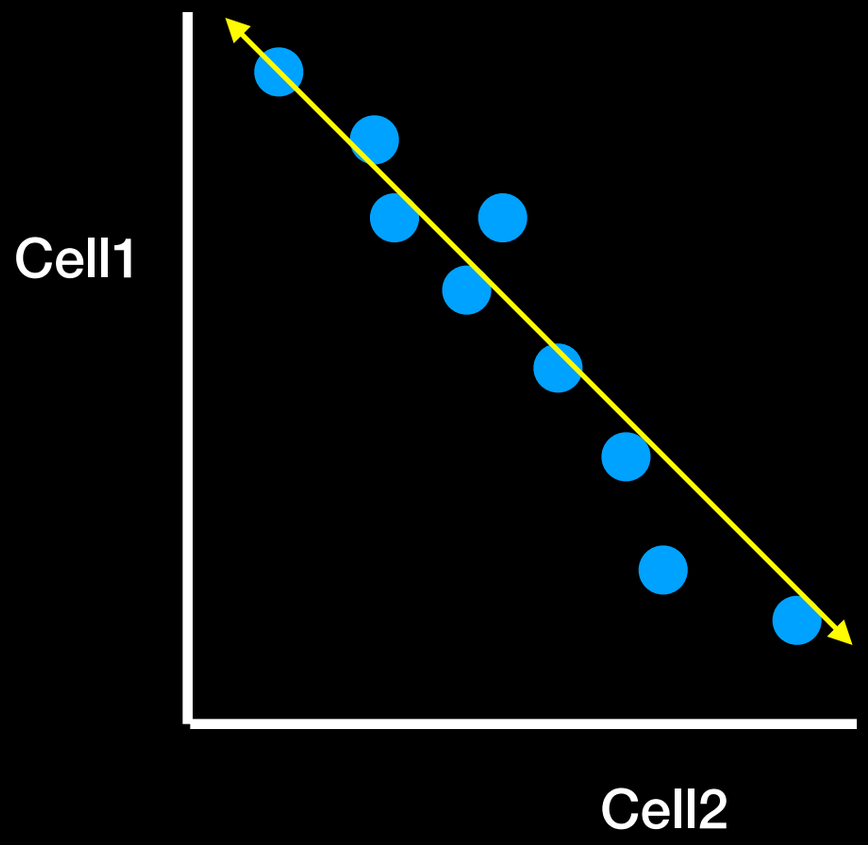
This gene (Gene9) is lowly transcribed in Cell1 and highly transcribed in Cell2…



	Cell1	Cell2
Gene1	3	0.25
Gene2	2.9	0.8
Gene3	2.2	1
Gene4	2	1.4
Gene5	1.3	1.6
Gene6	1.5	2
Gene7	1.1	2.2
Gene8	1	2.7
Gene9	0.4	3

In generel, Cell1 and Cell2 have an inverse correlation.

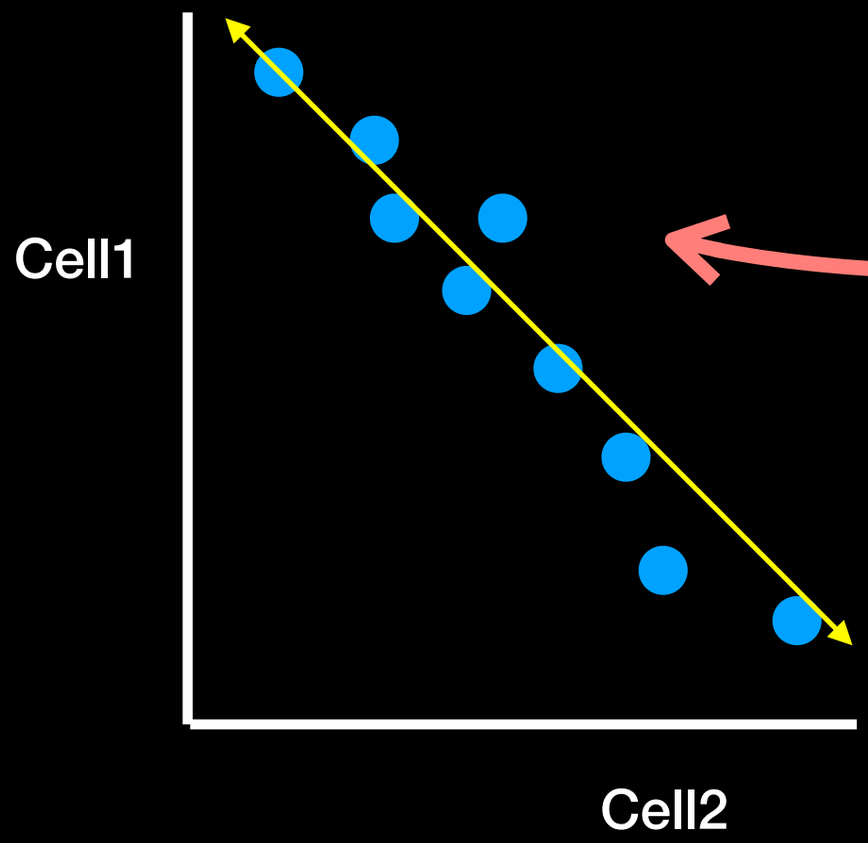
This suggests that they may be two different types of cells as they are using different genes



	Cell1	Cell2
Gene1	3	0.25
Gene2	2.9	0.8
Gene3	2.2	1
Gene4	2	1.4
Gene5	1.3	1.6
Gene6	1.5	2
Gene7	1.1	2.2
Gene8	1	2.7
Gene9	0.4	3

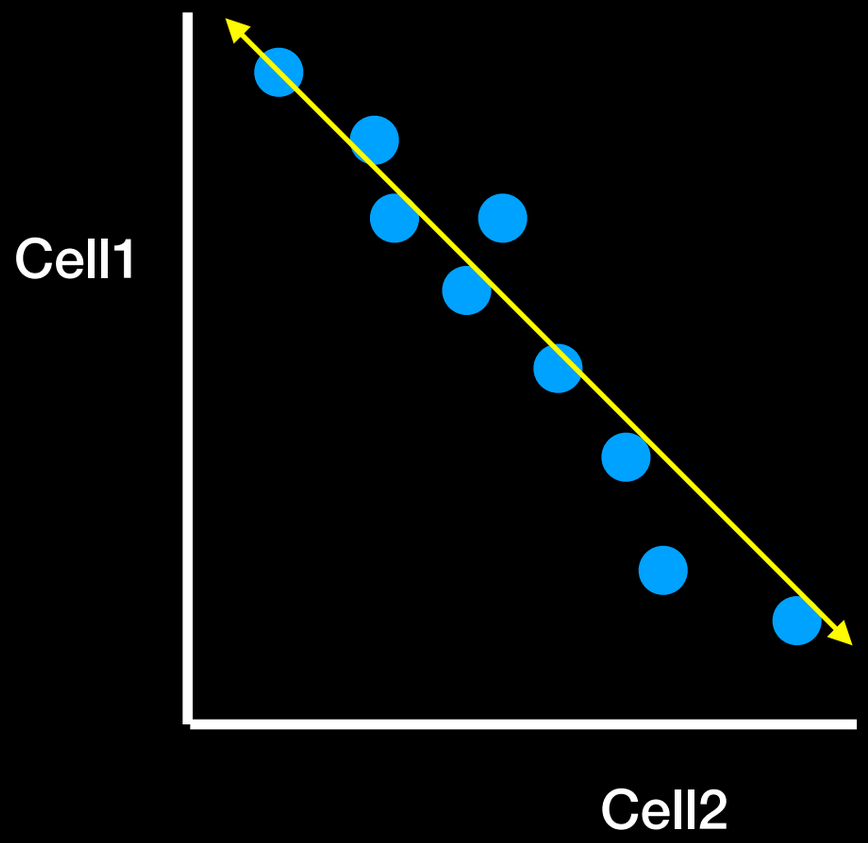
Now lets imagine there are three cells

	Cell1	Cell2	Cell3
Gene1	3	0.25	2.8
Gene2	2.9	0.8	2.2
Gene3	2.2	1	1.5
Gene4	2	1.4	2
Gene5	1.3	1.6	1.6
Gene6	1.5	2	2.1
Gene7	1.1	2.2	1.2
Gene8	1	2.7	0.9
Gene9	0.4	3	0.6



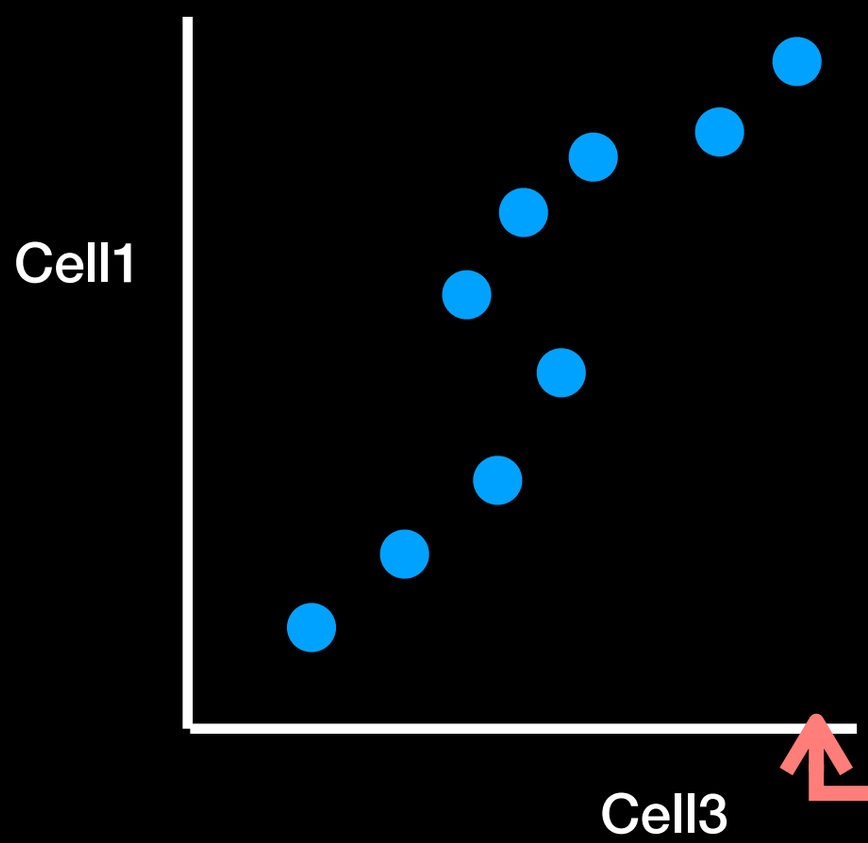
We have already seen how we can plot the first two cells to see how closely related they are

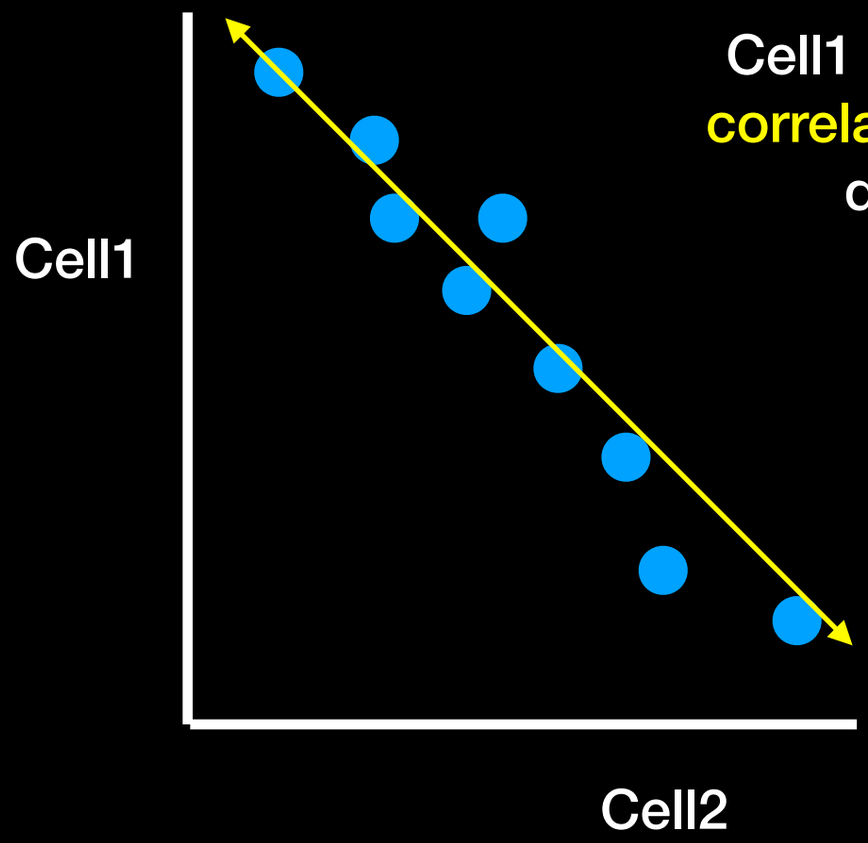
	Cell1	Cell2	Cell3
Gene1	3	0.25	2.8
Gene2	2.9	0.8	2.2
Gene3	2.2	1	1.5
Gene4	2	1.4	2
Gene5	1.3	1.6	1.6
Gene6	1.5	2	2.1
Gene7	1.1	2.2	1.2
Gene8	1	2.7	0.9
Gene9	0.4	3	0.6



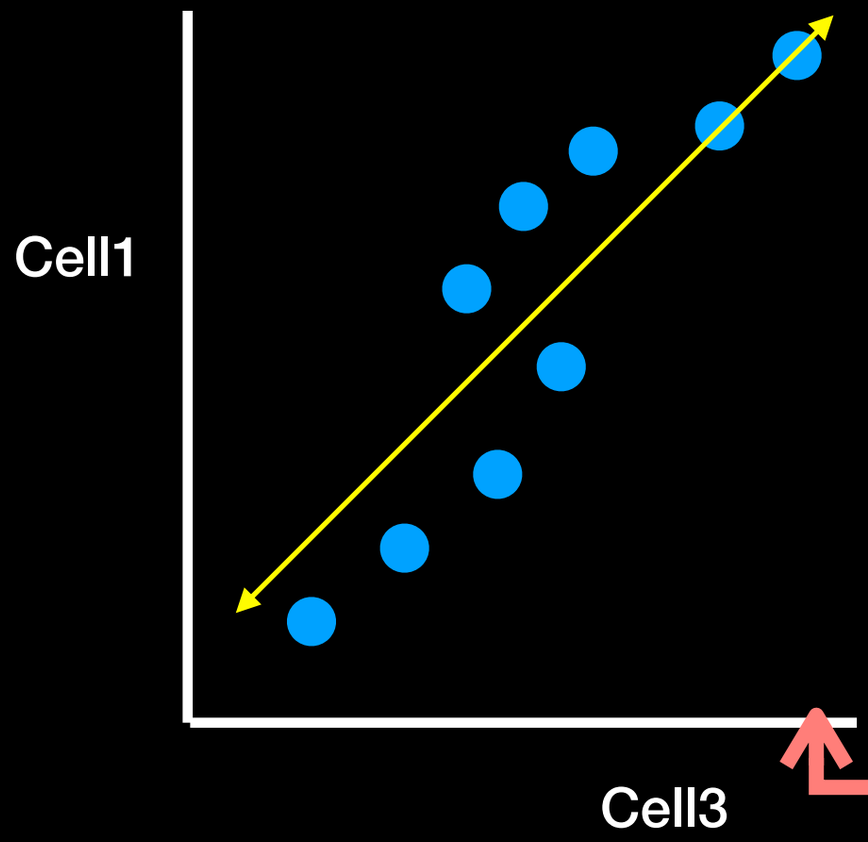
Now we can also compare
Cell1 to Cell3

	Cell1	Cell2	Cell3
Gene1	3	0.25	2.8
Gene2	2.9	0.8	2.2
Gene3	2.2	1	1.5
Gene4	2	1.4	2
Gene5	1.3	1.6	1.6
Gene6	1.5	2	2.1
Gene7	1.1	2.2	1.2
Gene8	1	2.7	0.9
Gene9	0.4	3	0.6

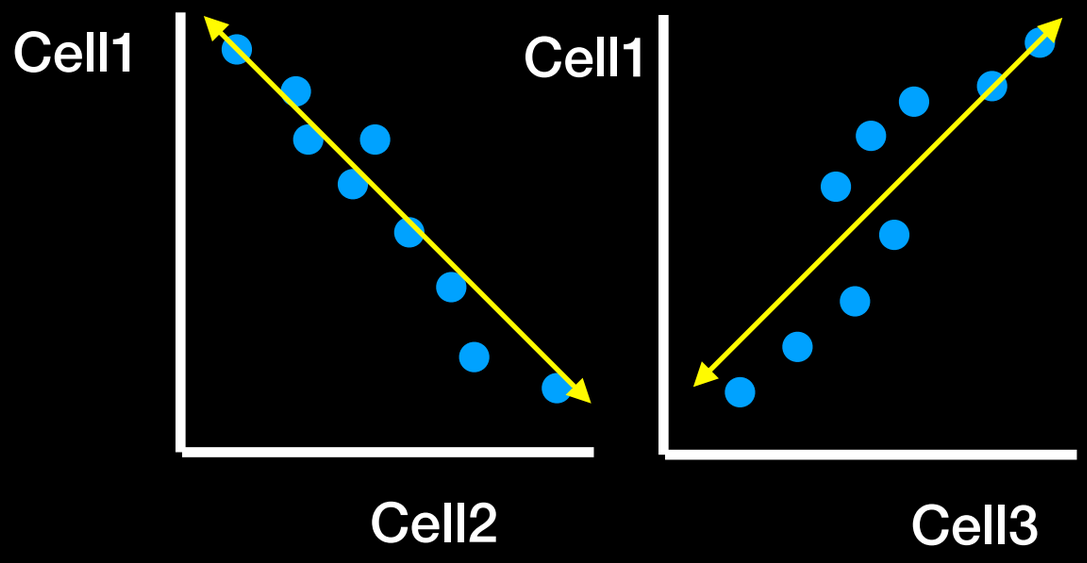




Cell1 and Cell3 are positively correlated suggesting they are doing similar things



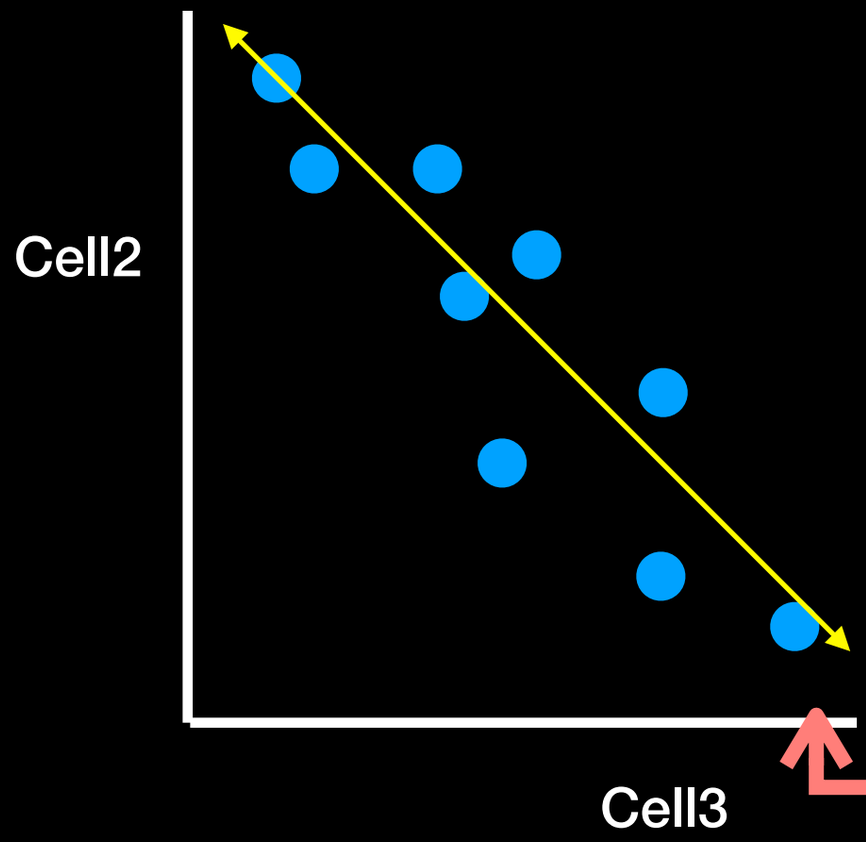
	Cell1	Cell2	Cell3
Gene1	3	0.25	2.8
Gene2	2.9	0.8	2.2
Gene3	2.2	1	1.5
Gene4	2	1.4	2
Gene5	1.3	1.6	1.6
Gene6	1.5	2	2.1
Gene7	1.1	2.2	1.2
Gene8	1	2.7	0.9
Gene9	0.4	3	0.6



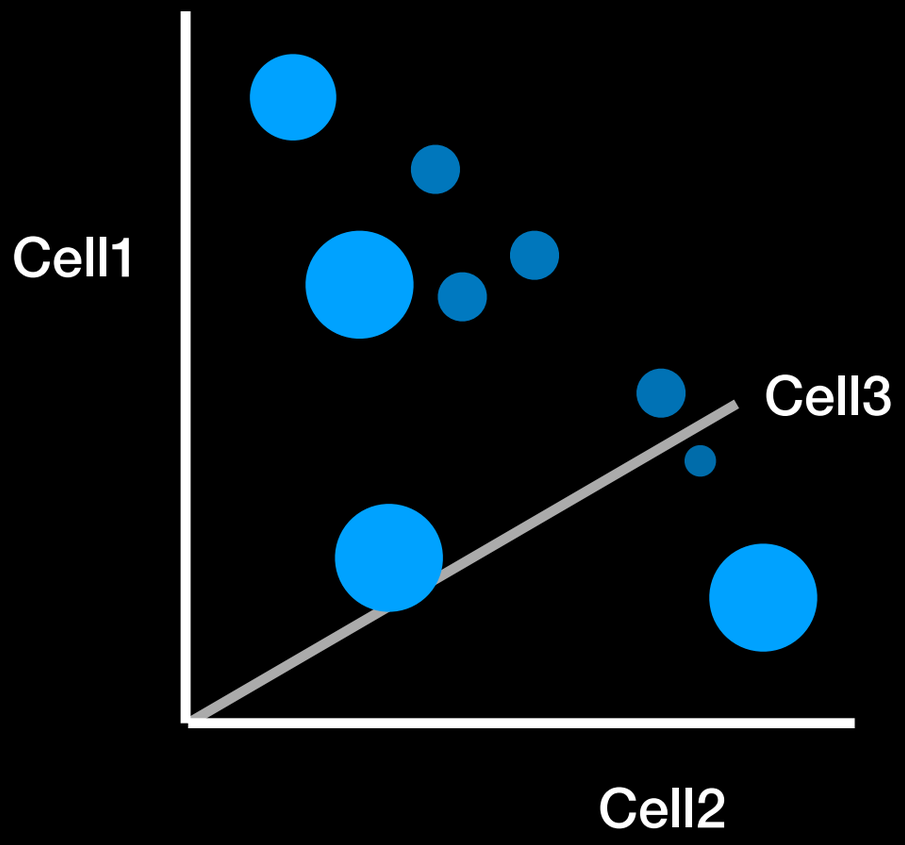
We can also compare Cell2 to Cell3…

	Cell1	Cell2	Cell3
Gene1	3	0.25	2.8
Gene2	2.9	0.8	2.2
Gene3	2.2	1	1.5
Gene4	2	1.4	2
Gene5	1.3	1.6	1.6
Gene6	1.5	2	2.1
Gene7	1.1	2.2	1.2
Gene8	1	2.7	0.9
Gene9	0.4	3	0.6

The inverse correlation suggests that Cell2 is doing something different from Cell3



Alternatively, we could try to plot all 3 cells at once on a 3-dimensional graph.



	Cell1	Cell2	Cell3
Gene1	3	0.25	2.8
Gene2	2.9	0.8	2.2
Gene3	2.2	1	1.5
Gene4	2	1.4	2
Gene5	1.3	1.6	1.6
Gene6	1.5	2	2.1
Gene7	1.1	2.2	1.2
Gene8	1	2.7	0.9
Gene9	0.4	3	0.6

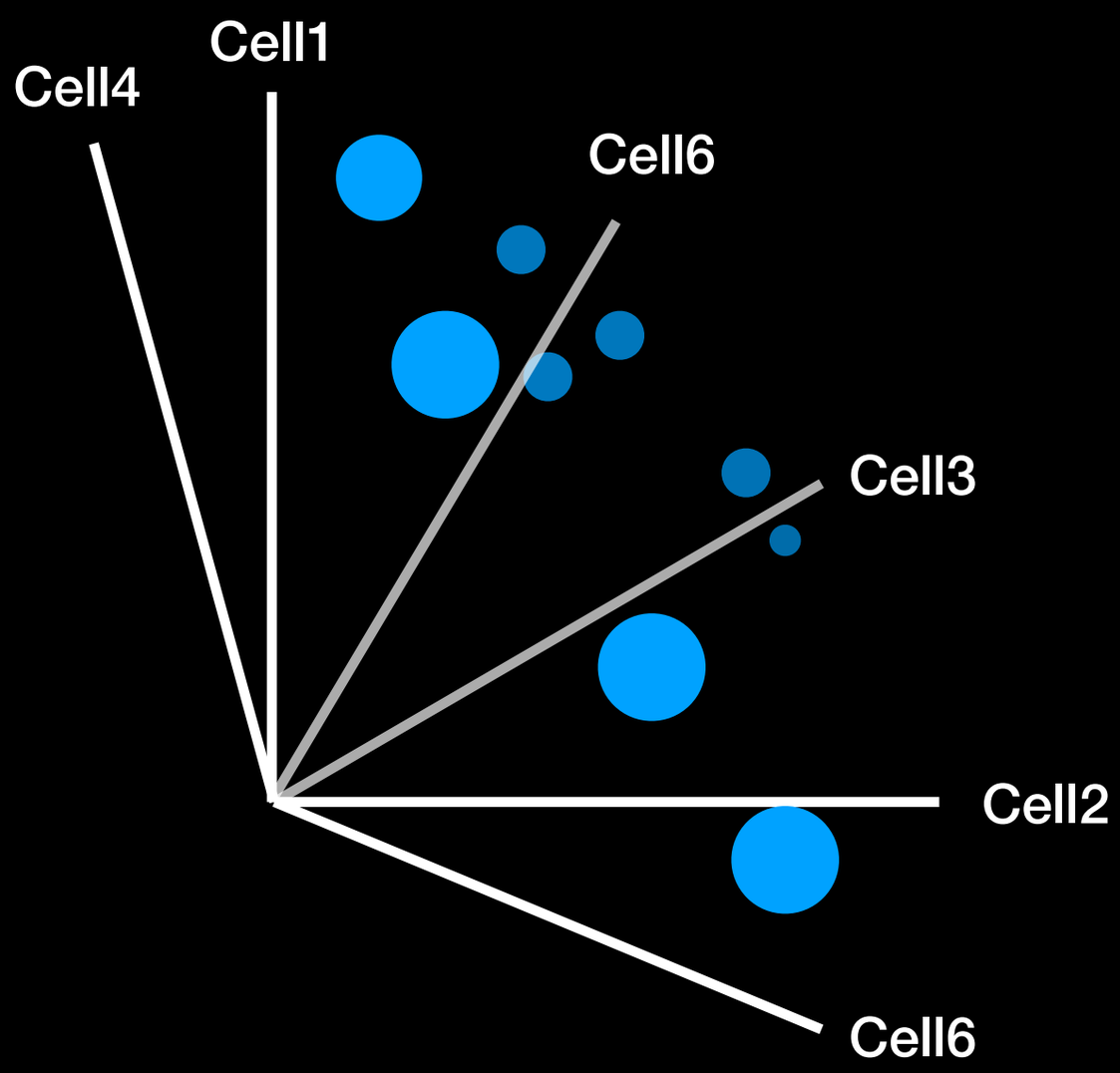
But what if we have 4 or more Cells?

	Cell1	Cell2	Cell3	Cell4	…
Gene1	3	0.25	2.8	0.1	…
Gene2	2.9	0.8	2.2	1.8	…
Gene3	2.2	1	1.5	3.2	…
Gene4	2	1.4	2	0.3	…
Gene5	1.3	1.6	1.6	0	…
Gene6	1.5	2	2.1	3	…
Gene7	1.1	2.2	1.2	2.8	…
Gene8	1	2.7	0.9	0.3	…
Gene9	0.4	3	0.6	0.1	…

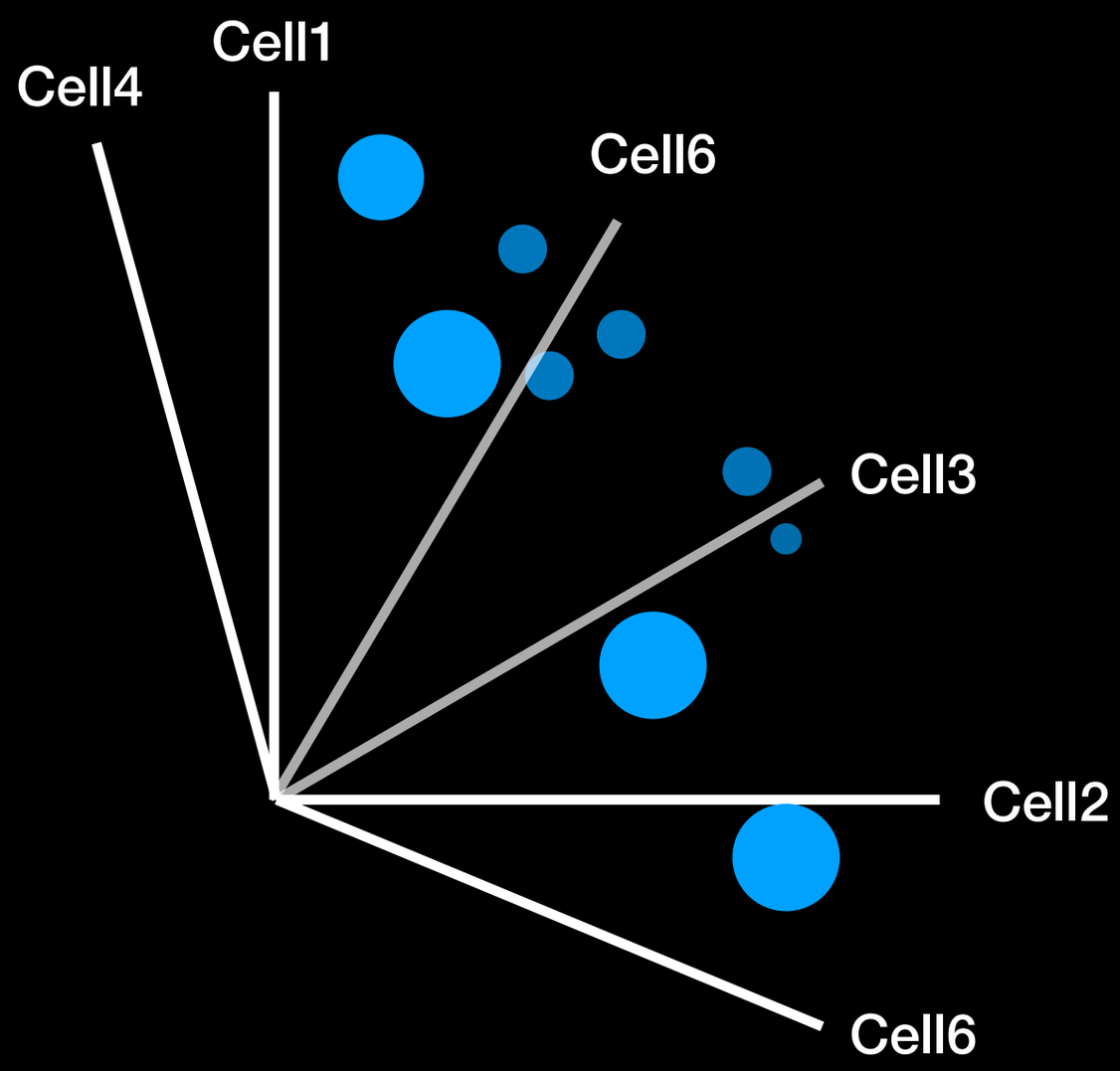
Draw lots of 2 cell plots and try to make sense of them all?

	Cell1	Cell2	Cell3	Cell4	…
Gene1	3	0.25	2.8	0.1	…
Gene2	2.9	0.8	2.2	1.8	…
Gene3	2.2	1	1.5	3.2	…
Gene4	2	1.4	2	0.3	…
Gene5	1.3	1.6	1.6	0	…
Gene6	1.5	2	2.1	3	…
Gene7	1.1	2.2	1.2	2.8	…
Gene8	1	2.7	0.9	0.3	…
Gene9	0.4	3	0.6	0.1	…

Or draw some crazy graph that has an axis for each cell and makes or brains hurt!

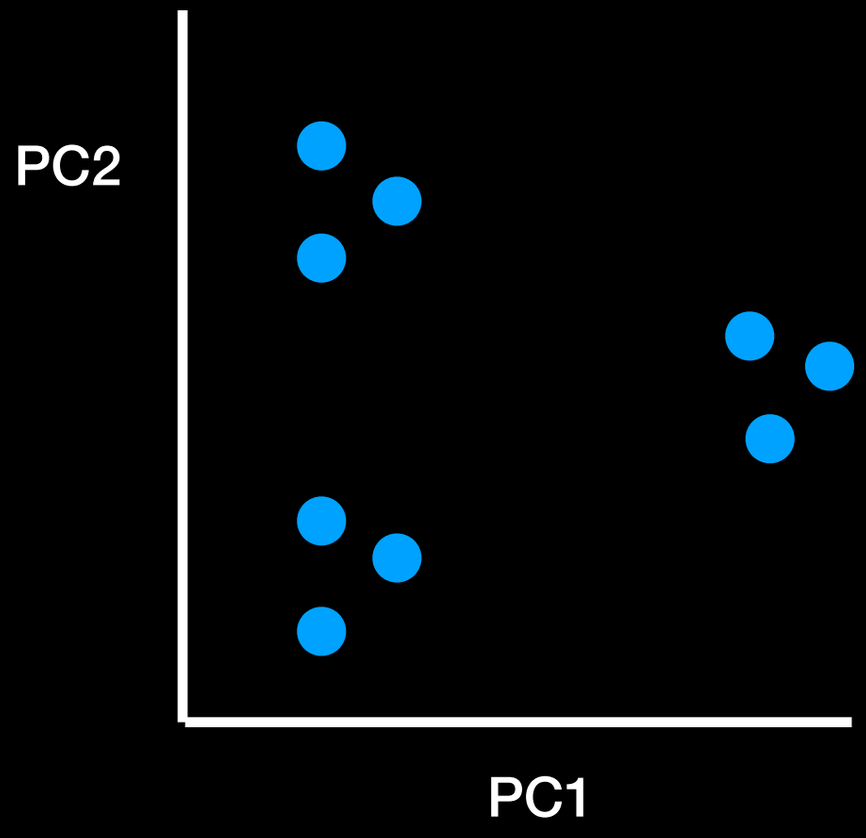


	Cell1	Cell2	Cell3	Cell4	…
Gene1	3	0.25	2.8	0.1	…
Gene2	2.9	0.8	2.2	1.8	…
Gene3	2.2	1	1.5	3.2	…
Gene4	2	1.4	2	0.3	…
Gene5	1.3	1.6	1.6	0	…
Gene6	1.5	2	2.1	3	…
Gene7	1.1	2.2	1.2	2.8	…
Gene8	1	2.7	0.9	0.3	…
Gene9	0.4	3	0.6	0.1	…



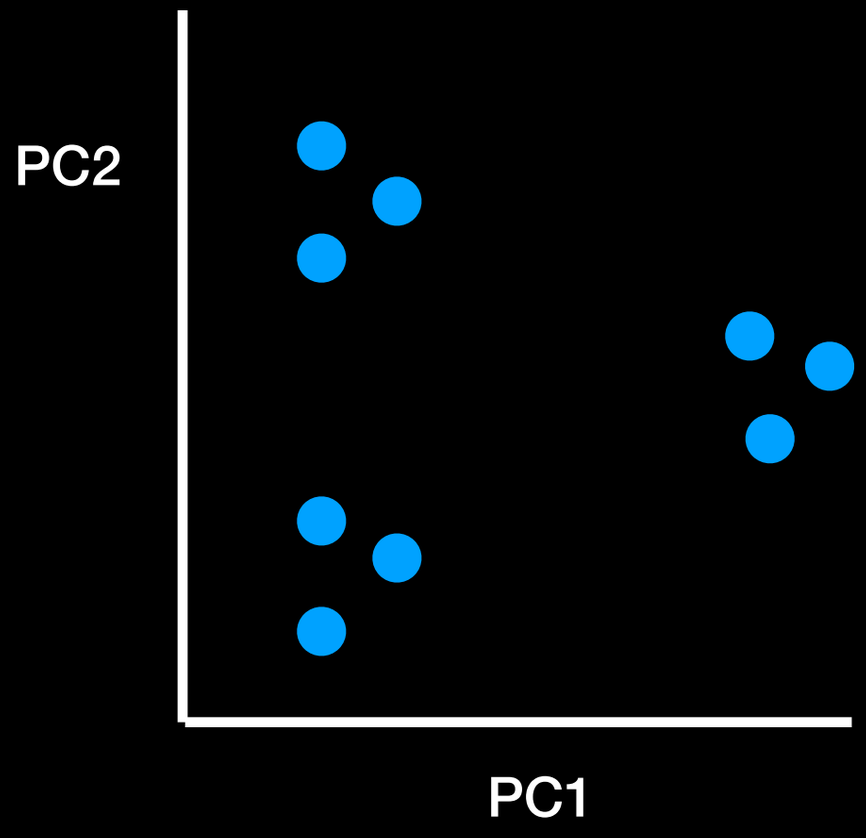
	Cell1	Cell2	Cell3	Cell4	…
Gene1	3	0.25	2.8	0.1	…
Gene2	2.9	0.8	2.2	1.8	…
Gene3	2.2	1	1.5	3.2	…
Gene4	2	1.4	2	0.3	…
Gene5	1.3	1.6	1.6	0	…
Gene6	1.5	2	2.1	3	…
Gene7	1.1	2.2	1.2	2.8	…
Gene8	1	2.7	0.9	0.3	…
Gene9	0.4	3	0.6	0.1	…

Enter Principal Component Analysis (PCA)



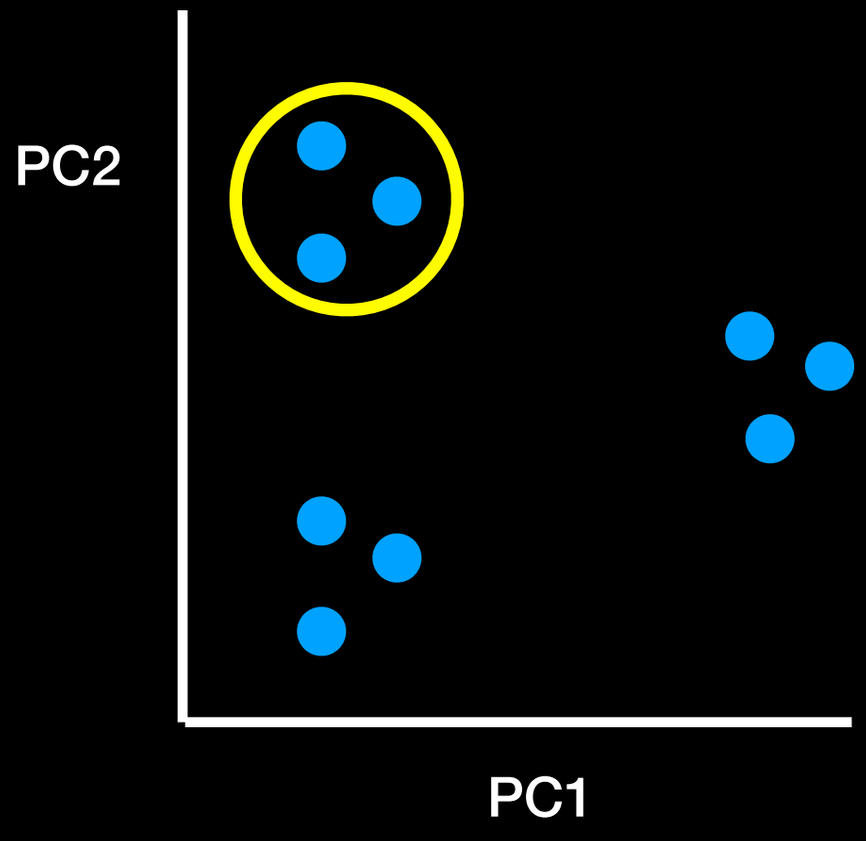
	Cell1	Cell2	Cell3	Cell4	…
Gene1	3	0.25	2.8	0.1	…
Gene2	2.9	0.8	2.2	1.8	…
Gene3	2.2	1	1.5	3.2	…
Gene4	2	1.4	2	0.3	…
Gene5	1.3	1.6	1.6	0	…
Gene6	1.5	2	2.1	3	…
Gene7	1.1	2.2	1.2	2.8	…
Gene8	1	2.7	0.9	0.3	…
Gene9	0.4	3	0.6	0.1	…

PCA converts the correlations (or lack there of) among all cells into a representation we can more readily interpret (e.g. a 2D graph!)



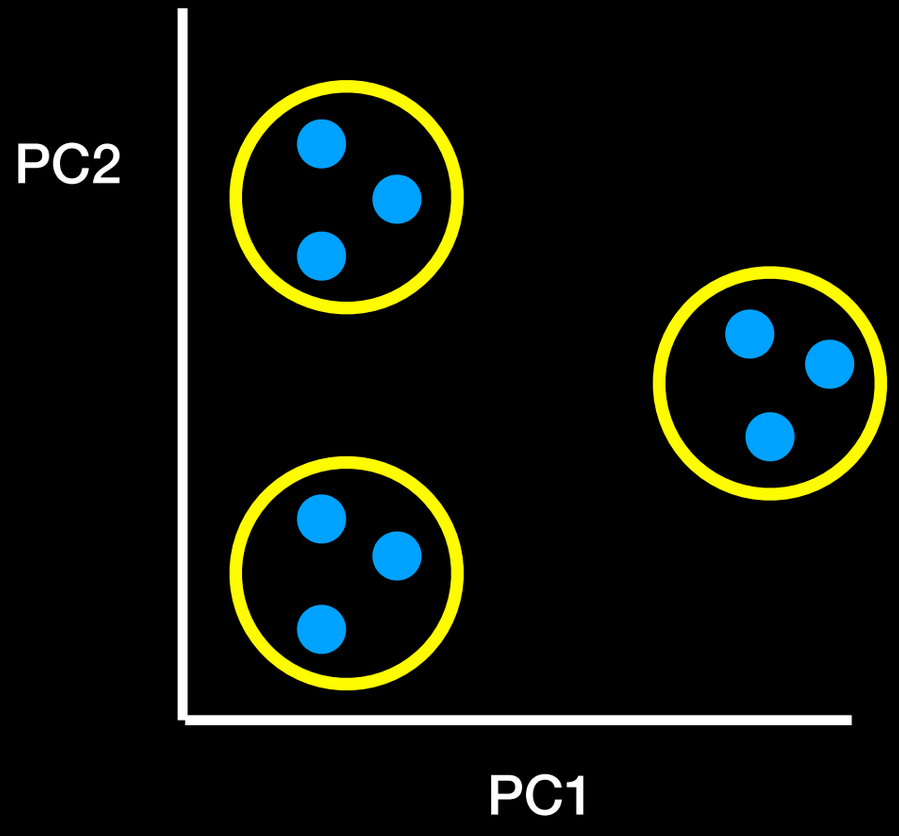
	Cell1	Cell2	Cell3	Cell4	…
Gene1	3	0.25	2.8	0.1	…
Gene2	2.9	0.8	2.2	1.8	…
Gene3	2.2	1	1.5	3.2	…
Gene4	2	1.4	2	0.3	…
Gene5	1.3	1.6	1.6	0	…
Gene6	1.5	2	2.1	3	…
Gene7	1.1	2.2	1.2	2.8	…
Gene8	1	2.7	0.9	0.3	…
Gene9	0.4	3	0.6	0.1	…

Cells that are highly correlated cluster together



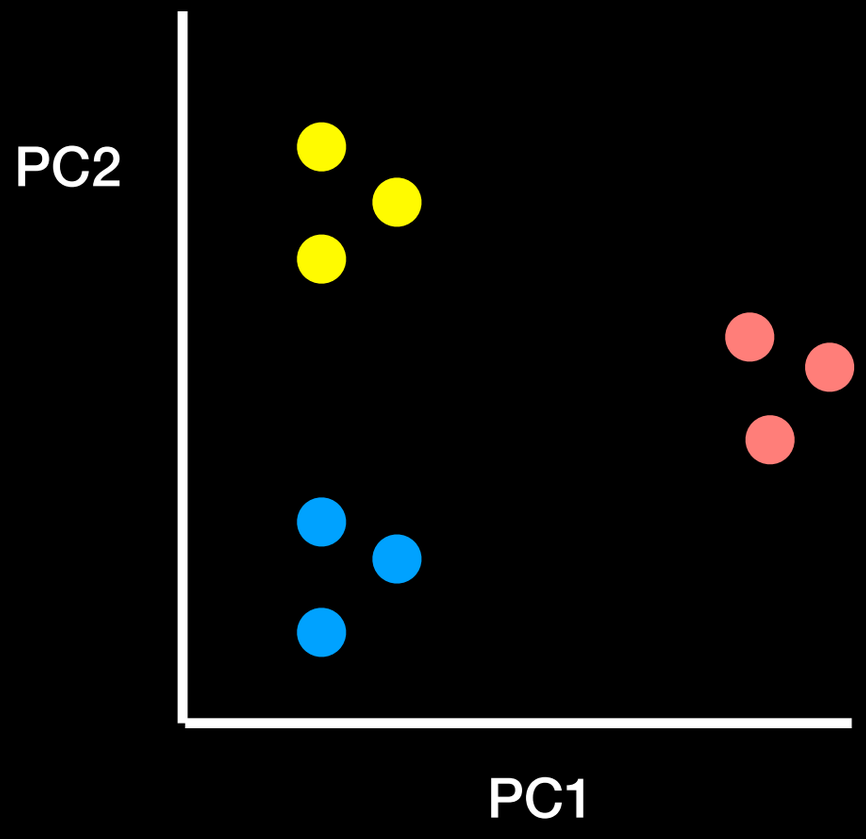
	Cell1	Cell2	Cell3	Cell4	…
Gene1	3	0.25	2.8	0.1	…
Gene2	2.9	0.8	2.2	1.8	…
Gene3	2.2	1	1.5	3.2	…
Gene4	2	1.4	2	0.3	…
Gene5	1.3	1.6	1.6	0	…
Gene6	1.5	2	2.1	3	…
Gene7	1.1	2.2	1.2	2.8	…
Gene8	1	2.7	0.9	0.3	…
Gene9	0.4	3	0.6	0.1	…

Cells that are highly correlated cluster together



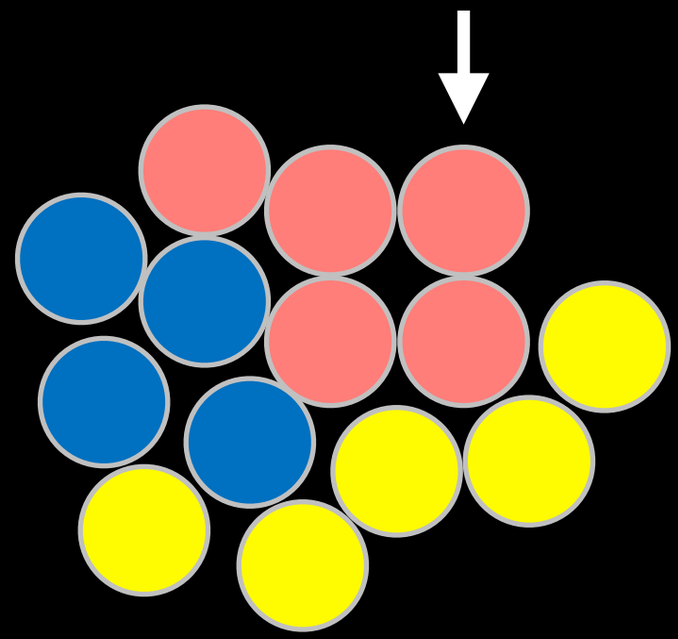
	Cell1	Cell2	Cell3	Cell4	…
Gene1	3	0.25	2.8	0.1	…
Gene2	2.9	0.8	2.2	1.8	…
Gene3	2.2	1	1.5	3.2	…
Gene4	2	1.4	2	0.3	…
Gene5	1.3	1.6	1.6	0	…
Gene6	1.5	2	2.1	3	…
Gene7	1.1	2.2	1.2	2.8	…
Gene8	1	2.7	0.9	0.3	…
Gene9	0.4	3	0.6	0.1	…

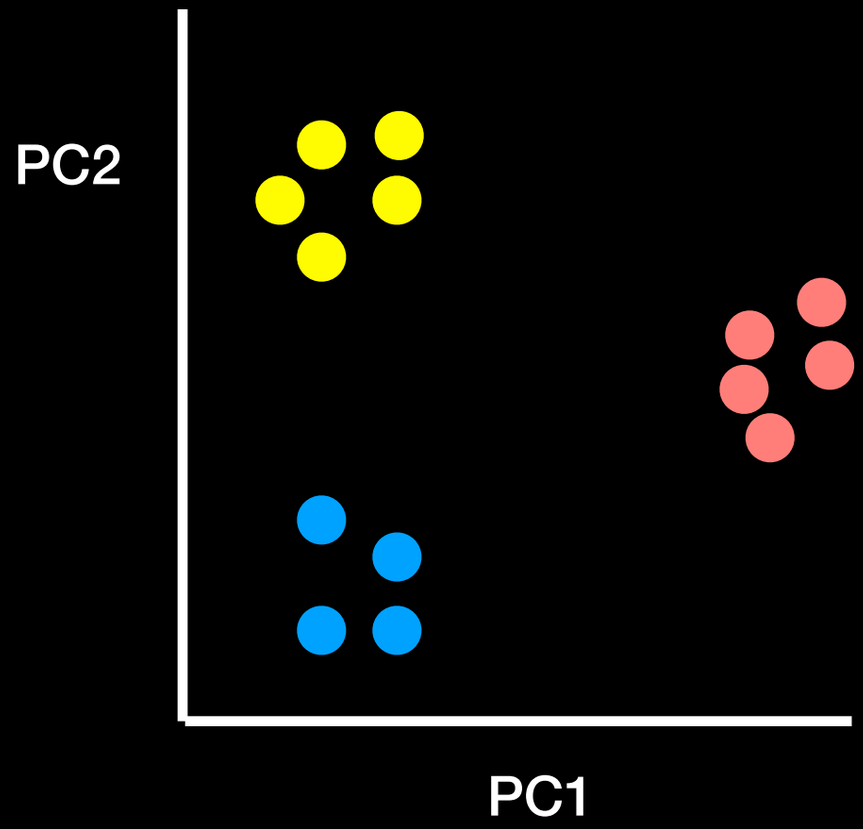
To make the clusters easier to see
we can color code them…

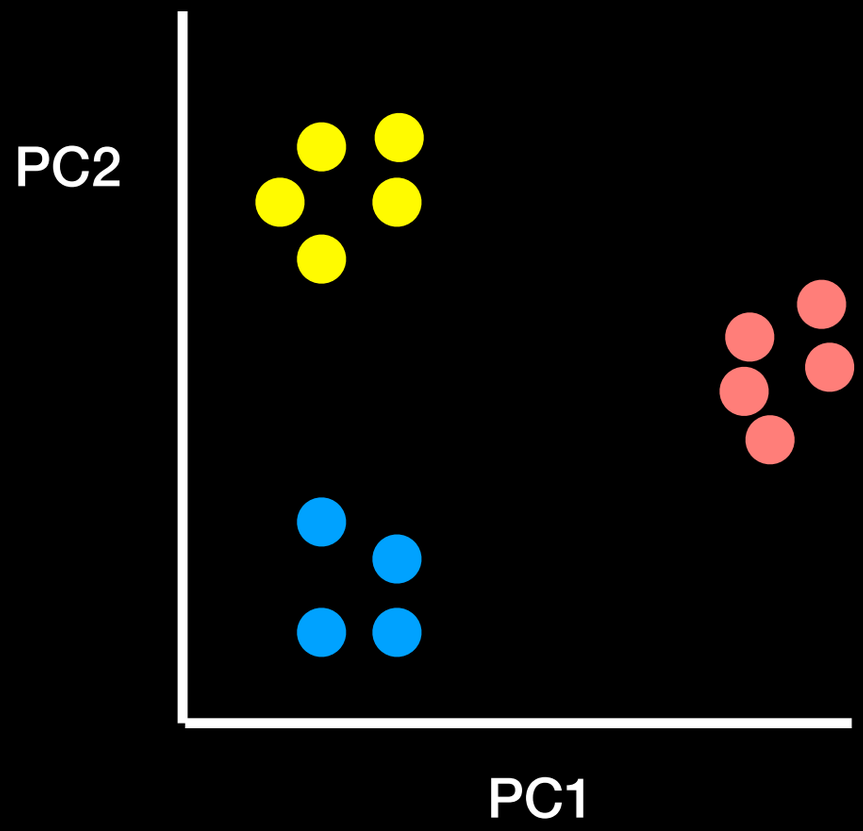


	Cell1	Cell2	Cell3	Cell4	…
Gene1	3	0.25	2.8	0.1	…
Gene2	2.9	0.8	2.2	1.8	…
Gene3	2.2	1	1.5	3.2	…
Gene4	2	1.4	2	0.3	…
Gene5	1.3	1.6	1.6	0	…
Gene6	1.5	2	2.1	3	…
Gene7	1.1	2.2	1.2	2.8	…
Gene8	1	2.7	0.9	0.3	…
Gene9	0.4	3	0.6	0.1	…

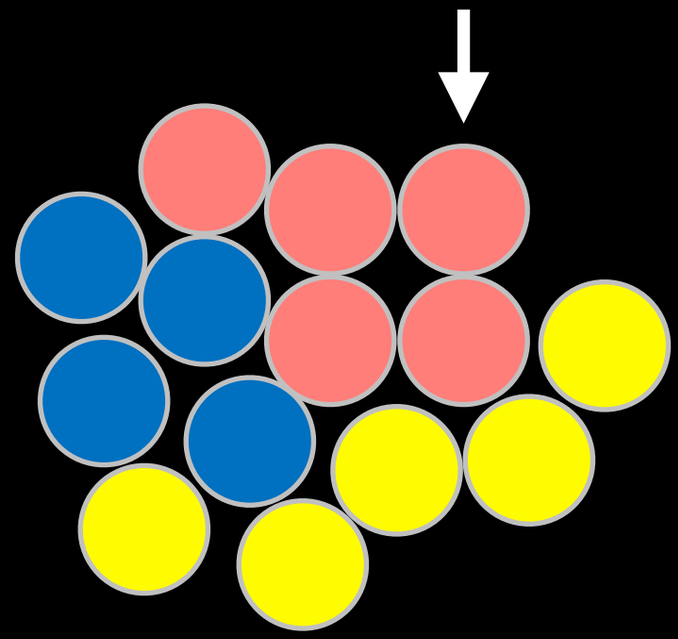
Once we have identified the clusters from our PCA results, we can go back to or original cells…



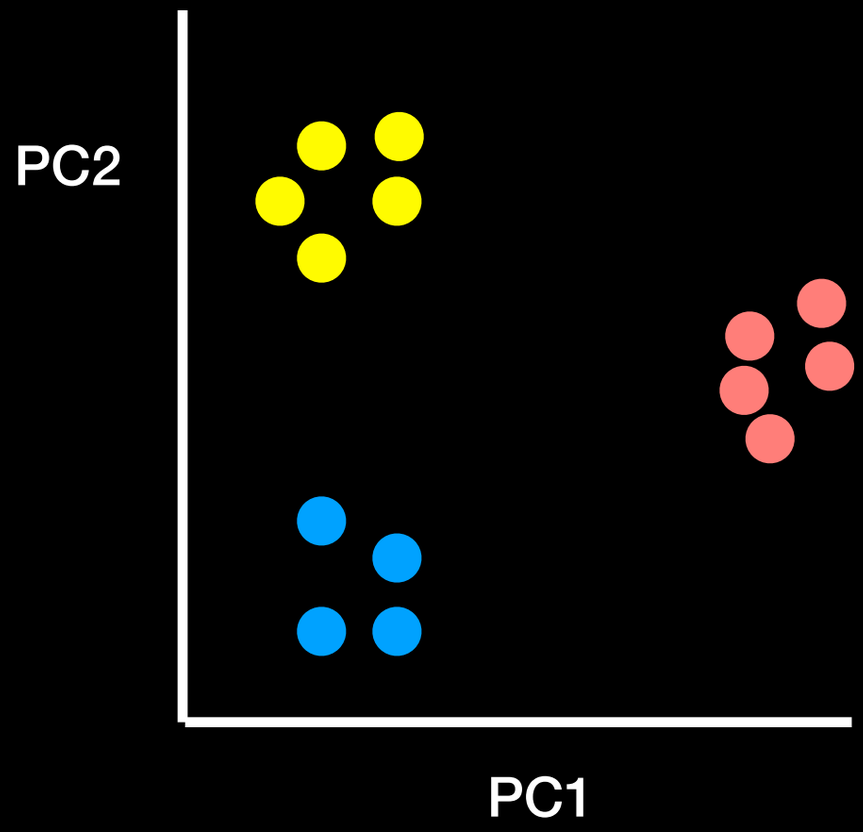




Once we have identified the clusters from our PCA results, we can go back to or original cells…



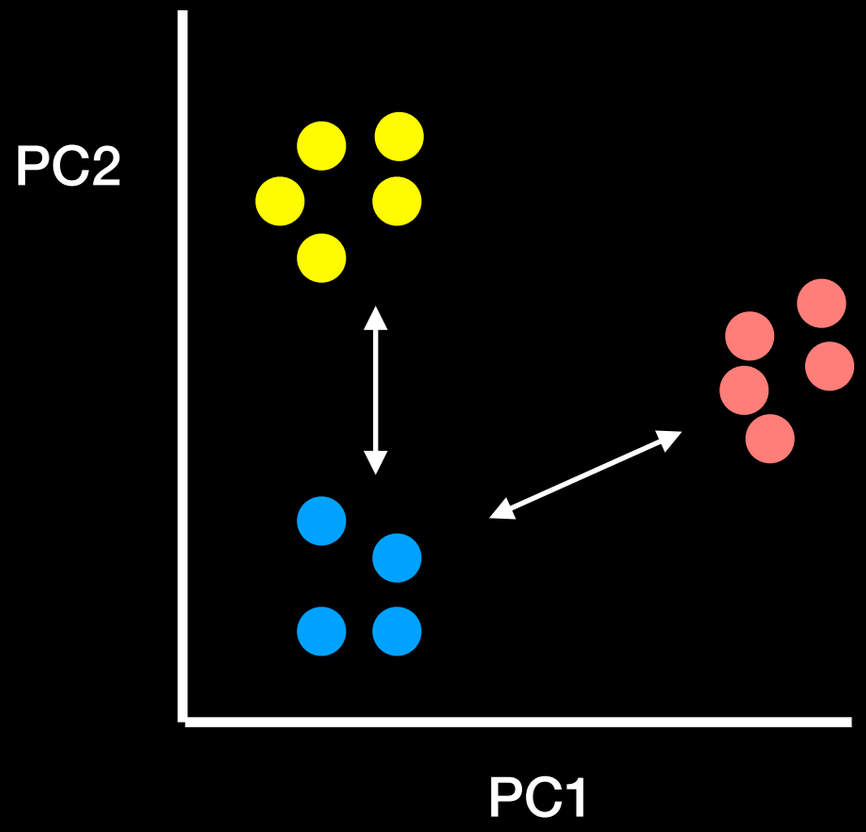
…and see they represent three different types of cells doing three different things with their genes!



<u>Some key points</u>:

The PCs (i.e. new plot axis) are ranked by their importance

So PC1 is more important than PC2 which in turn is more important than PC3 etc.

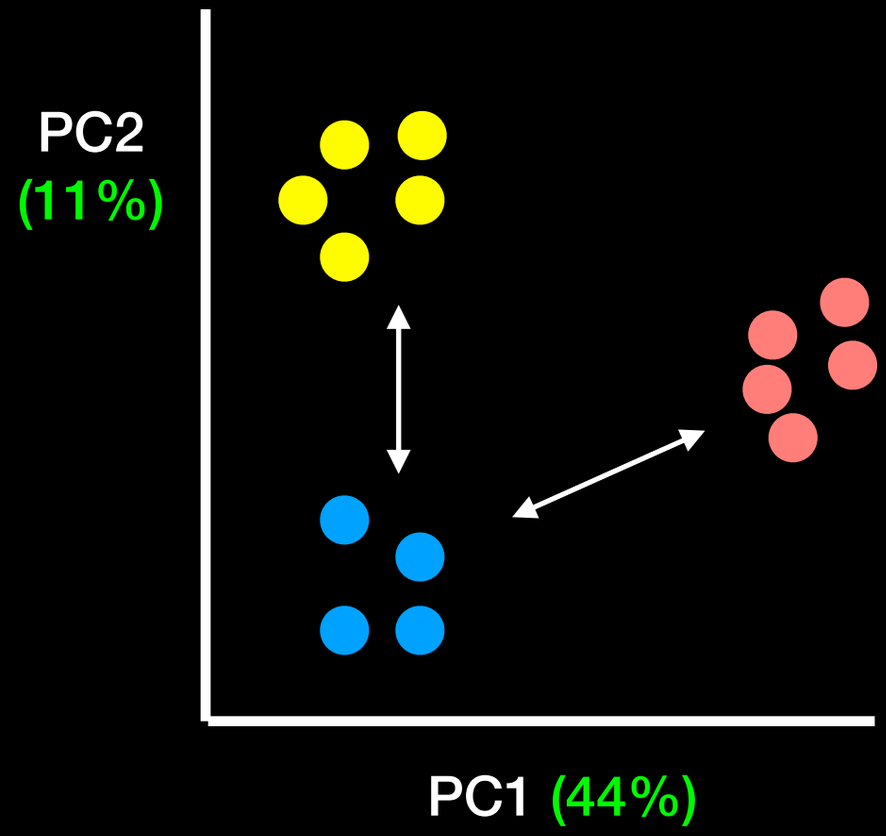


<u>Some key points</u>:

The PCs (i.e. new plot axis) are ranked by their importance

So PC1 is more important than PC2 which in turn is more important than PC3 etc.

So the red and blue cluster are more dissimilar than the yellow and blue clusters



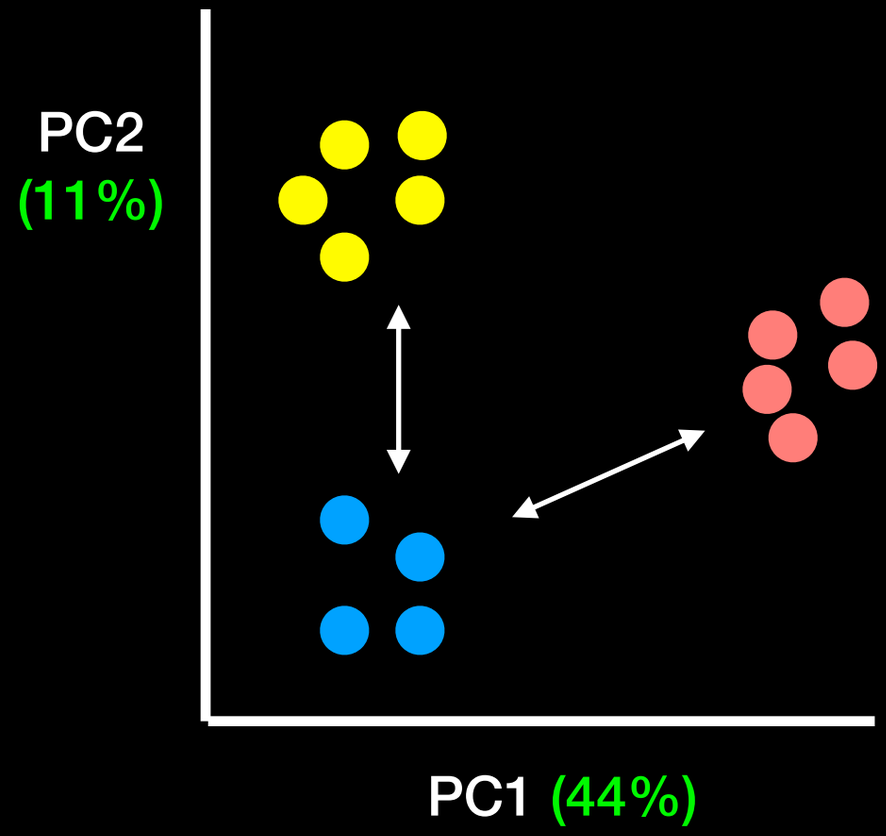
<u>Some key points</u>:

The PCs (i.e. new plot axis) are ranked by their importance

So PC1 is more important than PC2 which in turn is more important than PC3 etc.

So the red and blue cluster are more dissimilar than the yellow and blue clusters

The PCs (i.e. new plot axis) are ranked by the amount of variance in the original data (i.e. gene expression values) that they “capture”



<u>Some key points</u>:

The PCs (i.e. new plot axis) are ranked by their importance

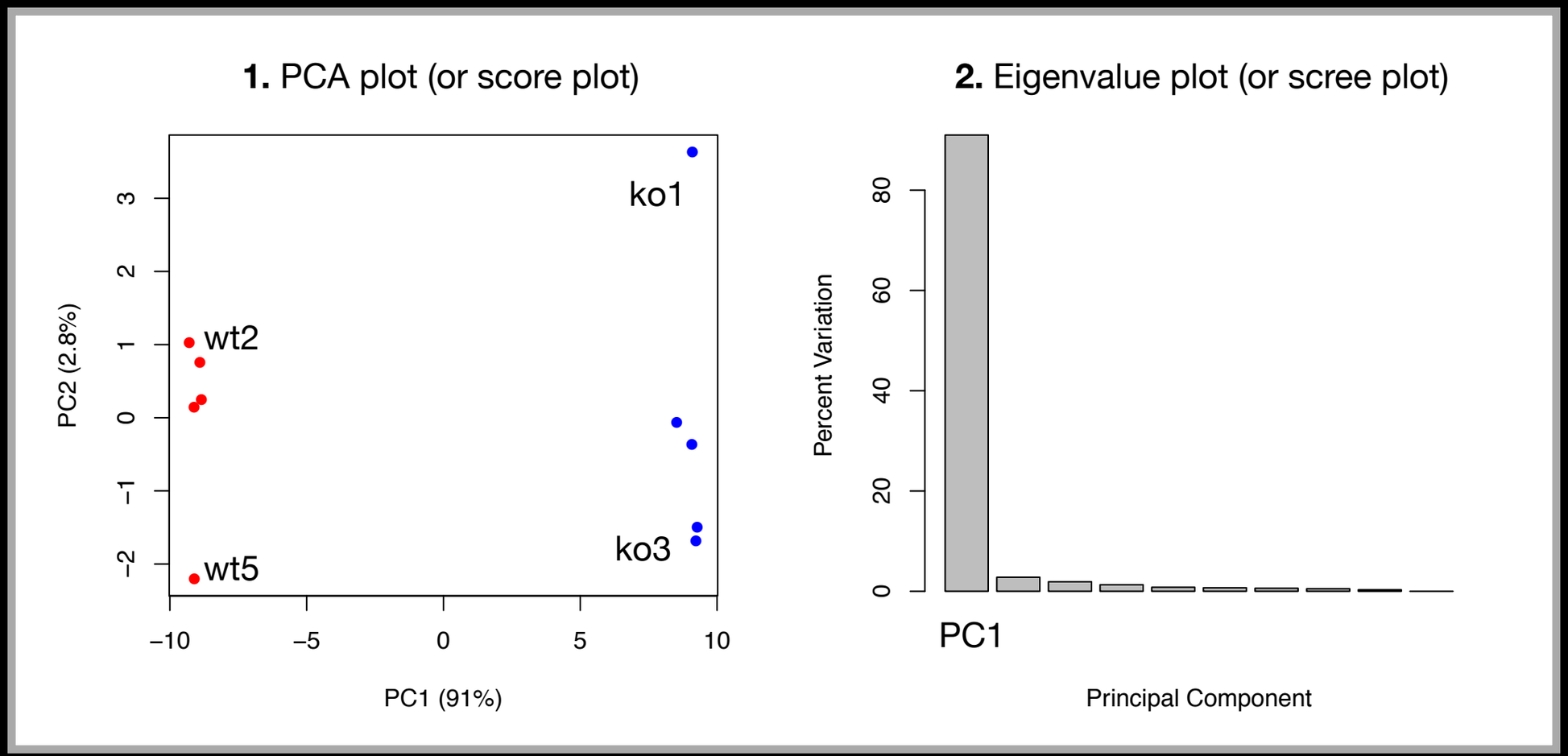
So PC1 is more important than PC2 which in turn is more important than PC3 etc.

So the red and blue cluster are more dissimilar than the yellow and blue clusters

The PCs (i.e. new plot axis) are ranked by the amount of variance in the original data (i.e. gene expression values) that they “capture”

In this example PC1 ‘captures’ 4x more of the original variance than PC2 (44/11 = 4)

- We actually get two main things out of a typical PCA
  - The new axis (called PCs or Eigenvectors) and
  - Eigenvalues that detail the amount of variance captured by each PC



- Another cool thing we can get out of PCA is a quantitive report on how the original variables contributed to each PC

    - In other words, which were the most important genes that lead to the observed clustering in PC-space

- These are often called the loadings and we can plot them to see which are the most important genes for the observed separation as well as outputting ranked lists of genes that act to discriminate the samples

```
    gene64     gene39
0.1047968  0.1047629
```

```
     gene7      gene65
-0.1047629 -0.1047443
```

Do it Yourself!

Hands-on time!

http://setosa.io/ev/principal-component-analysis/

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

New Course

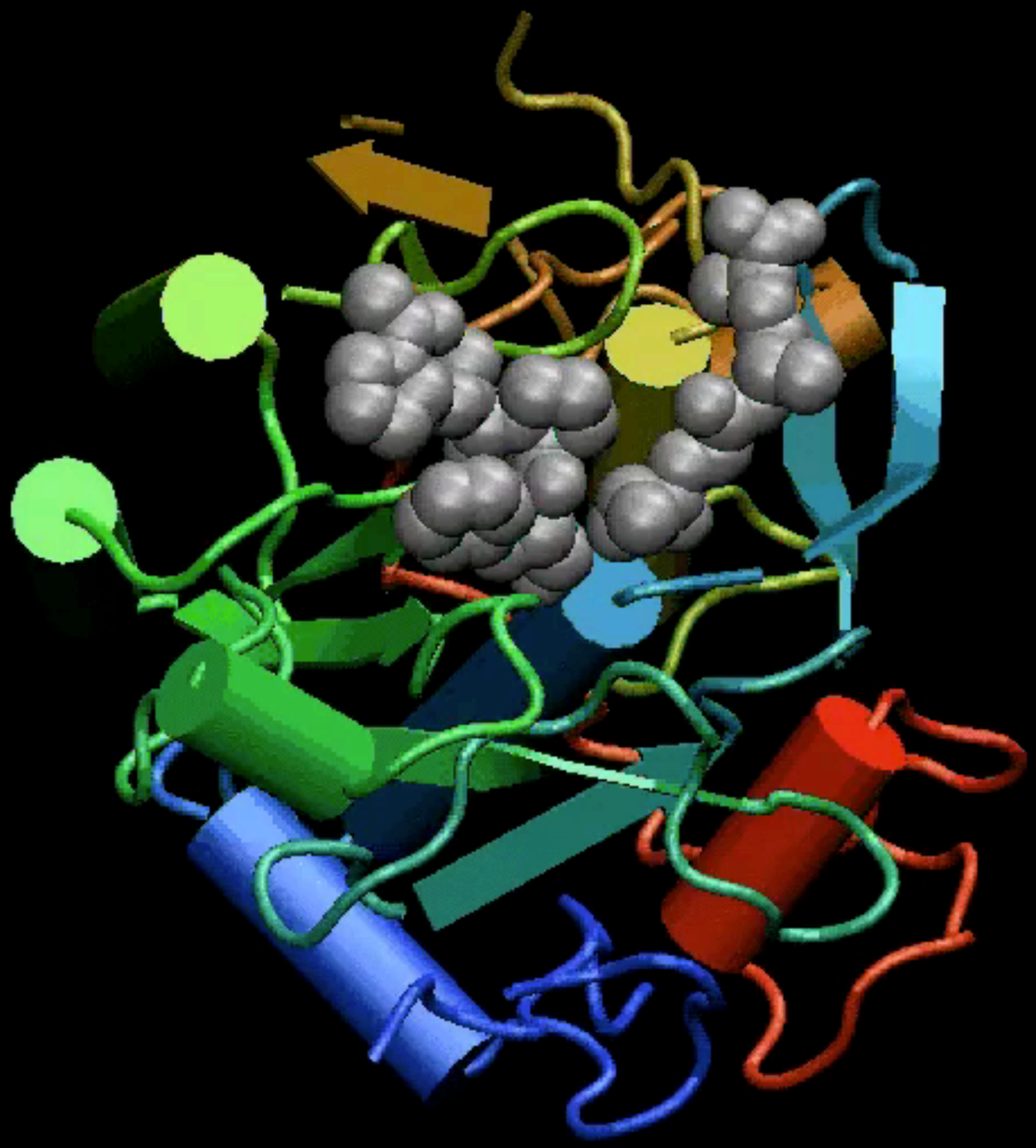
To find out more:

Register for **BIMM-143** to learn
Essential Bioinformatics

http://thegrantlab.org/bimm143

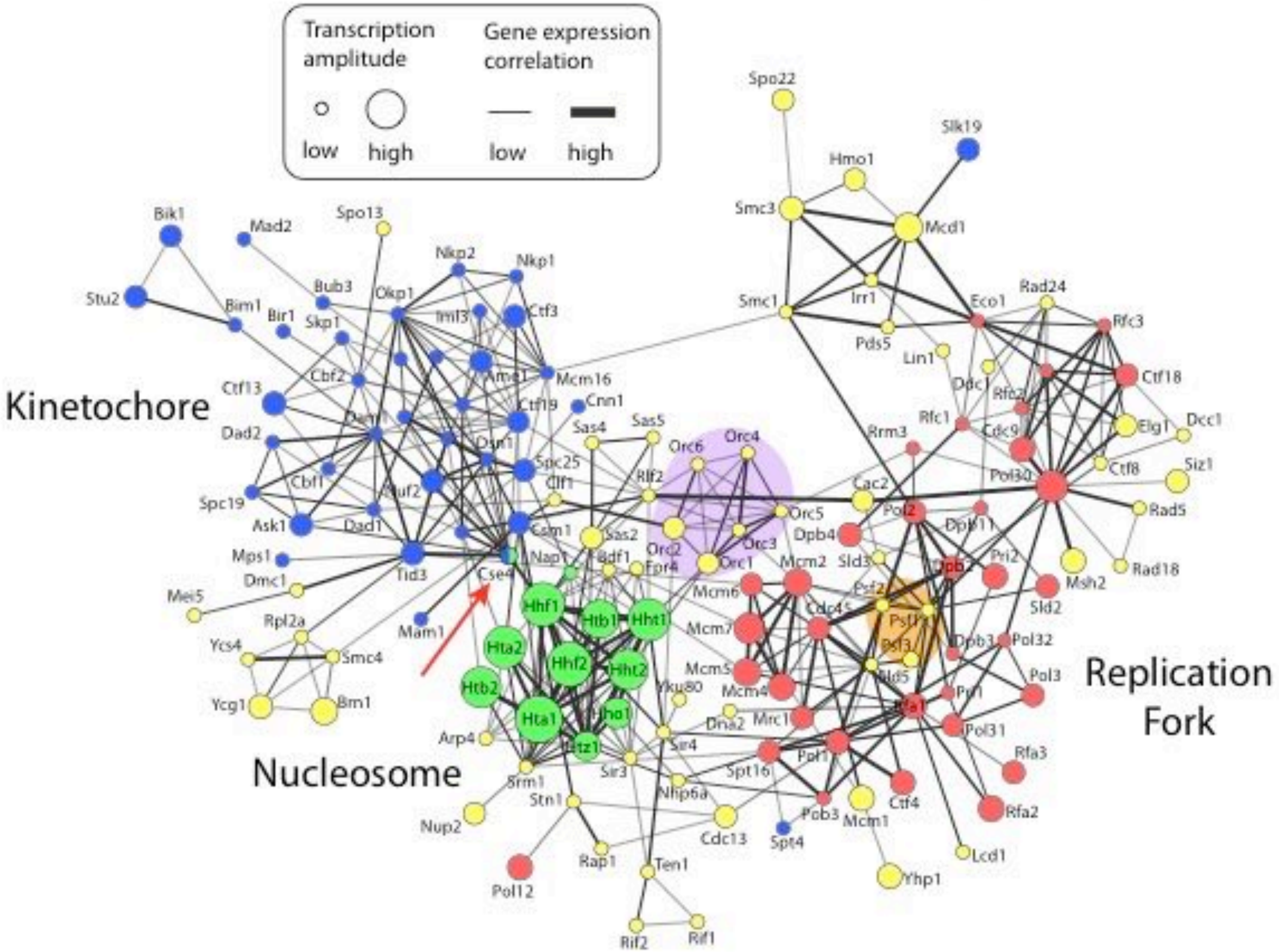
Hands on analysis of large datasets from RNA-Seq, cancer cells, metagenomics, drug design, etc…

PCA models can tell us about protein function and drugability



Proteinase K

[Yeast, cell-cycle PPIs]



Today's Menu

Methods Recap:

- **Principal component analysis** (PCA), Network analysis, and Heatmaps

Missing Discussion Topics:

- How much data can be gathered and how can this data help?
- Who controls your 'omic' information?
- Can your omics data be used against you?
- Who pays for genome sequencing in health care?
- Will sequencing make health care cheaper?
- Will people act on genomic information?

Question:

How much data can be gathered about a single person?

Genome

GGTTCCAAAAGTTTATTGGATGCCGT
TTCAGTACATTTATCGTTTGCTTTGG
ATGCCCTAATTAAAAGTGACCCTTTC
AAACTGAAATTCATGATACACCAATG
GATATCCTTAGTCGATAAAATTTGCG
AGTACTTTCAAAGCCAAATGAAATTA
TCTATGGTAGACAAAACATTGACCAA
TTTCATATCGATCCTCCTGAATTTAT
TGGCGTTAGACACAGTTGGTATATTT
A....

??

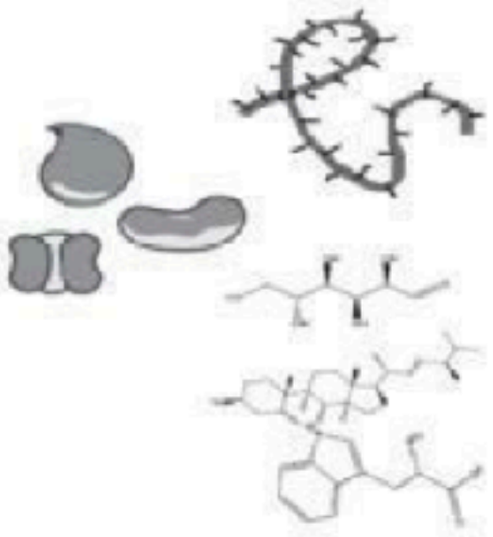
??

Microbiome

Sensors/Activity

??

Other omes ??



??

DNA methylation

GGTTCCAAAAGTTTATTGGATGCC
GTTTCAGTACATTTATCmGTTTG
CTTTGGATGCCCmGTAATTAAAA
GTGACCCTTTCATCATATCGATCC
TCCTGAATGTTAGACACAGTTGGT
ATATTTA....

- **Personal 'omics' data**
  - ➡ Genome, epigenome, transcriptome, proteome, metabolome, microbiome, etc.
  - ➡ A single genome sequence takes ~0.5 Tb.
  - ➡ For others we often want multiple time-points yielding 100s Tbs

- **Imaging data**
  - ➡ MRIs, radiographs, PRT scans, etc.
  - ➡ Several Tb but again we likely want multiple time-points

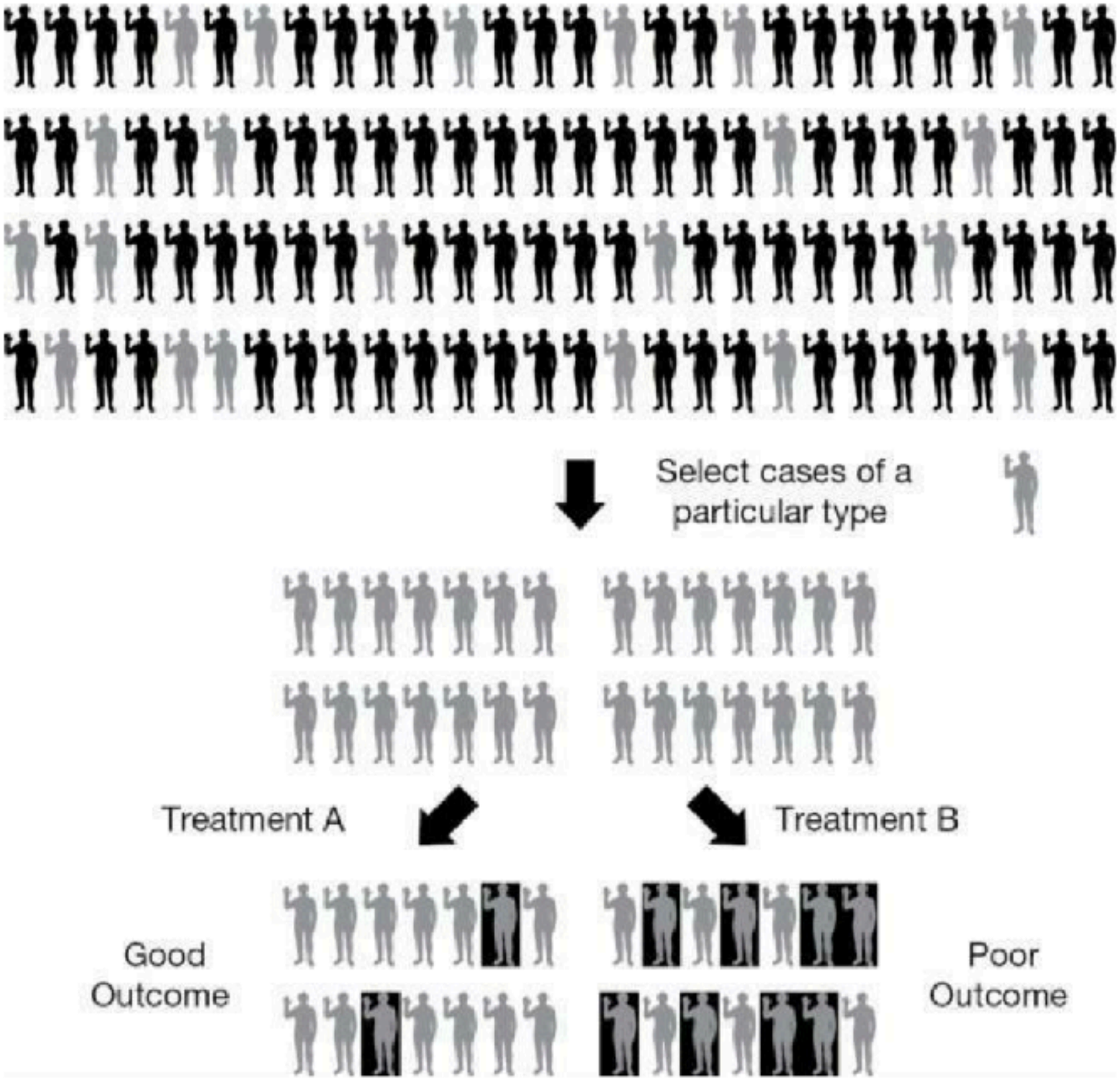
- **Other information**
  - ➡ Environmental exposure, exercise, diet, etc.
  - ➡ Collectively we are approaching petabytes for one person - **we want whole populations!**

How can this data help?

- Currently for most diseases doctors prescribe therapies based on their experience and best guess.

- They do not have precise information about what treatments lead to what outcomes
  - ➡ Standard guidelines exist for only a limited number of cases.

The **big data revolution** will change this paradigm!


Select cases of a particular type
Treatment A
Treatment B
Good Outcome
Poor Outcome

- **If a particular person becomes ill with a particular disease (e.g. cancer associated with a particular genetic mutation and molecular profile) then algorithms can search all other patients with the same disease and mutation, how they were treated, and what the outcome was.**

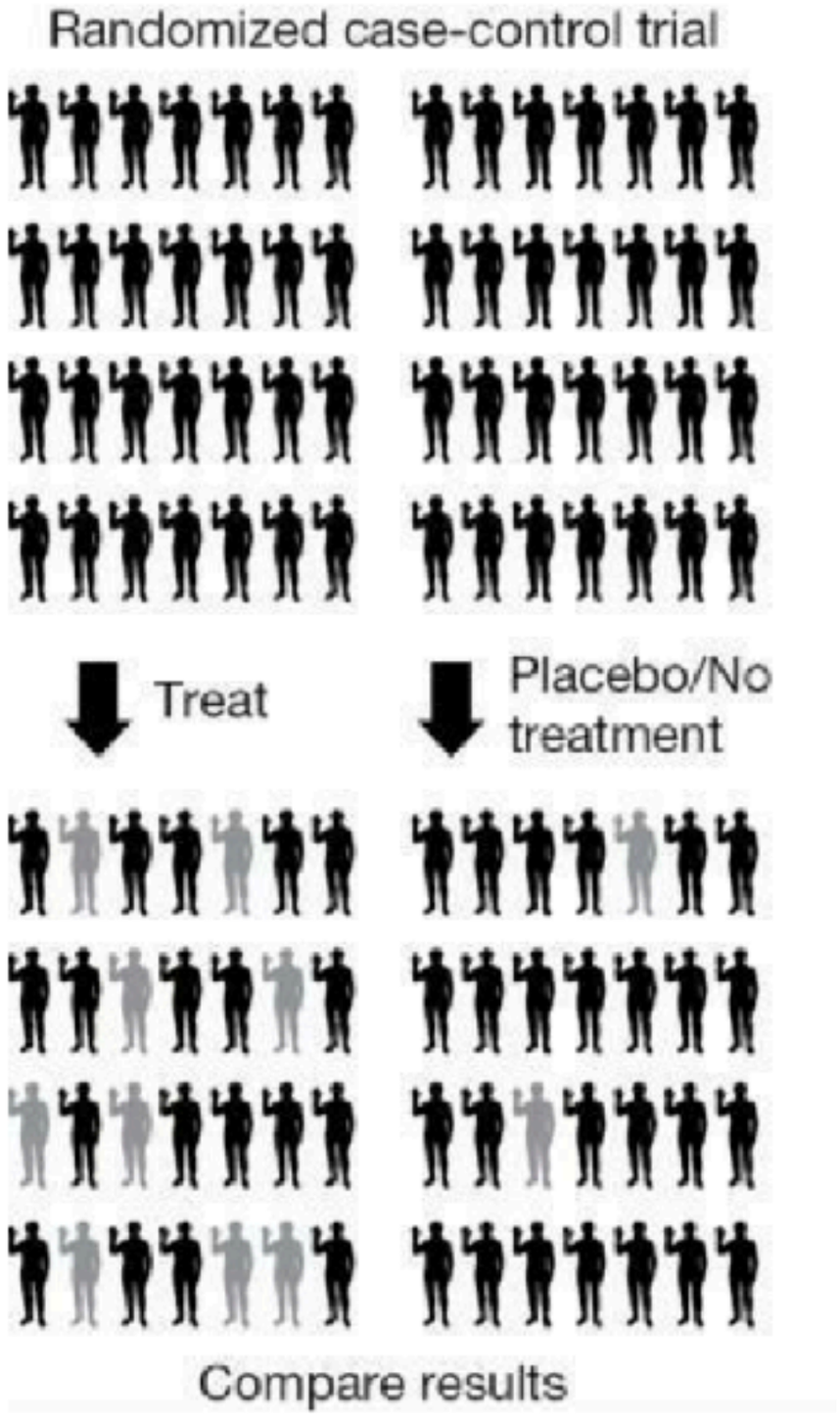
- **Each individual who is treated can be fed back into the "living" database to improve accuracy for future patients.**

This is new level of data-driven medicine!

- It is important to note that this "**living database**" model (in which observational data is collected and utilized clinically), is different from the manner in which drugs are currently vetted for clinical use.

- Existing **randomized case-control trials** are expensive and slow.

- In contrast, collecting pooled data that already exists will be less expensive and more rapid.



Randomized case-control trial

Treat

Placebo/No treatment

Compare results

Observational trial

Identify already treated and untreated

Look for enrichment

Data guided lifestyle decisions

- This big data can be used to manage one's health prior to the onset of disease.
- As more associations are made between genetic changes and human disease, this information will become more commonplace.
- E.g. individuals with mutations in the breast cancer genes will be screened at higher frequency; those with alterations in MODY and cardiomyopathy genes will be monitored for insulin and heart defects respectively, and managed accordingly.

What are the opportunities for industry & jobs?

- Industries that can manage large data and extract meaningful information from massive amounts of data will gain enormous opportunities

- This will require **data scientists** with biology and genetics backgrounds who can handle, integrate and visualize complex health data!

Question:

Who controls your genetic information?

- The answer should be obvious - **You do!**
- However, the implementation of this concept is not as straightforward as you might think.
  - ➡ What information you want returned from omics datasets requires serious thought.
  - ➡ Do you want all of your information returned to you or just the actionable information?
  - ➡ What its your definition of actionable?
  - ➡ What if there is uncertainty about the accuracy of the prediction?

- An example of non actionable information might be whether you have the Huntington's risk factor change?
  - ➡ If you have the high risk change for this gene, your change of getting this progressive neurological (and fatal) disease is extremely high.
  - ➡ There is no cure so the information will not help you medically.

- An example of non actionable information might be whether you have the Huntington's risk factor change?
  - ➡ If you have the high risk change for this gene, your change of getting this progressive neurological (and fatal) disease is extremely high.
  - ➡ There is no cure so the information will not help you medically.

- However, some might argue that you can chose to live your life differently based on this knowledge.

Do these fundamental decisions lie with the **patient**, the **genetic counselor**, **physician**, or **government**?

- Recall that possible scenarios include learning that
  - (1) high risk for a certain disease that may or may not be actionable;
  - (2) that are likely to be a carrier for certain genetic diseases thus their children may be at risk; and
  - (3) their ancestry or parentage is not what they thought.
  - (4) Different types of results may be reported by different providers, some of whom are willing to report only exact matches to established disease-causing variants, whereas others report every variant that might be predicted to be damaging (but yet to be firmly established).

Who would you want to deliver genomic information to you and how? From a direct-to-consumer test?

Direct-to-consumer testing

- If patients are entitled to their own DNA, shouldn't they have direct and easy access to the data? For example via direct-to-consumer testing?
- Many experts feel that this data should only be communicated via live consultation with an expert. Do you agree?
- How can we better educate patients (and others) on genomic test interpretation?

Question:

Can your ‘omics data’ be used against you?

- In the US there is a relatively recent law called **GINA** (Genetic Information Nondiscrimination Act) that prevents discrimination on the basis of genetic information for health insurance and employment.

- **Limitations**: GINA does not protect from discrimination for life insurance or long-term disability insurance!

- In other countries ‘socialized health care’ discrimination may be less of a concern with regards to health care costs.

- Unfortunately, personal discrimination presently occurs on the basis of race, ethnicity, gender, sexual orientation, etc. Perhaps DNA will be added to this list…

Concerns surrounding genetic screening

- If performed at an early stage of pregnancy, this information may - depending on the wishes of the parents - lead to an early termination.
  - ➡ Clearly there are ethical issues surrounding termination for non disease situations (gender, intellectual ability, etc.),
  - ➡ Will this lead to "prescreened children" selected to display a limited number of traits
- What about variants deleterious in some conditions but beneficial in others?
  - ➡ Sickle cell anemia variant protective for malaria
  - ➡ Cystic fibrosis variants that are protective for TB
  - ➡ Mutation in CCR5 gene protective for HIV
- Species perspective: less natural variation will result in a more homogenous population that may be more susceptible to pathogens…

What would our world be like without these people

- **Woody Guthrie:** Huntington's disease
- **Frederic Chopin**: Cystic Fibrosis
- **Miles Davis**: Sickle Cell Anemia
- **John F. Kennedy**: Addison's disease
- **Maurice Ravel**: Frontotemporal dementia
- **Lou Gehrig**: ALS
- **Ronald Reagan**: Alzheimer's disease
- **Charles K. Kao** (Nobel prize in physics, father of fiber optics and broad band): Alzheimer's
- **Stephen Hawking**: ALS

etc…

Side-Note:

- It is now possible to directly **identify people based on their DNA** sequence when correlated with other publicly available information.
  - ➡ A recent study investigating a number of released genome sequences showed that the personal identity of individuals could be identified.
- This raises privacy concerns and the possibility that we you could be identified from samples of hair, skin etc. much like fingerprinting from crime scenes.
  - ➡ These samples could be matched against disease databases (e.g. Alzheimer's) and reveal whether they, or a relative, are in the database.
  - ➡ Contributing your sequence to disease databases has ramifications for your whole family that are rarely discussed.

Question:

Who pays for genome sequencing in treating disease?

Who pays for genome sequencing in treating disease

- Arguably the biggest detriment to implementing genomic medicine is: **Who pays?**

- Some insurance companies will reimburse for sequencing tumor genomes of cancer patients - particularly if they can result in targeted treatment and cost savings.
  - ➡ Many new drugs can cost $100,000 or more

	Drug	Treats	Cost
1	Factor viii Recombinant	Hemophilia A	$216,833
2	Remodulin	Pulmonary arterial hypertension	$130,772
3	Ventavis	Pulmonary arterial hypertension	$84,205
4	Primacor	Acute decompensated heart failure	$62,790
5	Erbitux	Cancer	$25,898
6	Dacogen	Myelodysplastic syndrome	$25,858
7	Herceptin	Cancer	$25,797
8	Vidaza	Myelodysplastic syndrome	$22,957
9	Sandostatin LAR Depot	Acromegaly, diarrhea, and flushing caused by cancerous tumors and vasoactive intestinal peptide secreting adenomas	$22,748

Medicare drug expenses per year in 2010

- For cases not related to cancer (or undiagnosed childhood disease), genome sequencing is not covered by insurance and usually not available in the clinic.

- In the US there is no incentive from the insurers perspective to pay for for sequencing even if it will lead to cost savings in the long term.
  - ➡ Insurance is usually obtained from employers and individuals change jobs and health plans. It may not benefit one insurance company to invest large sums to prevent future disease when the insured individual will not be in their plan in the future.

- Preventative healthcare genome sequencing is gaining traction in Europe and Japan (e.g. UK's 100K genome project). Will the US be left behind?

Question:

Will genome sequencing make health care cheaper?

- In the cancer area the answer is clearly yes!
  - ➡ Giving expensive drugs to patients that are not likely to respond is already realizing cost savings

- In the future, catching individuals with heart or coronary artery problems prior to heart attack or stroke is likely to reduce these adverse events and lead to cost savings.

- Solving undiagnosed diseases will cut down on many futile tests as well as on considerable anxiety.

- In other areas, it is not yet clear if genome sequencing will result in overall cost savings.

- In the case of preventative medicine will it just postpone adverse events but not ultimately save funds in the long term?

- Arguably, the real savings here are in the quality of health care!

Question:

Will people act on genomic information?

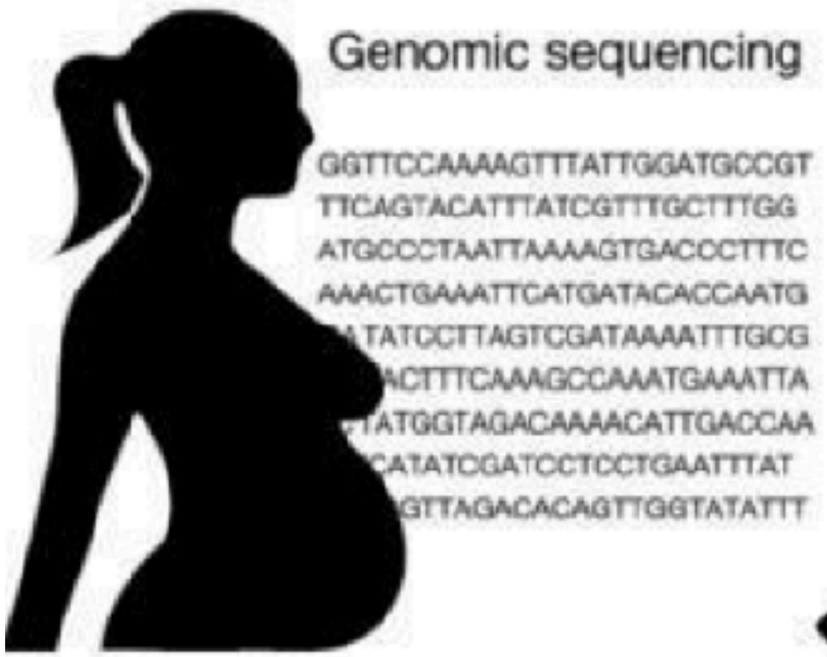
- In order for genome sequencing to be useful for preventative medicine, individuals who have been sequencing will need to act on the information.

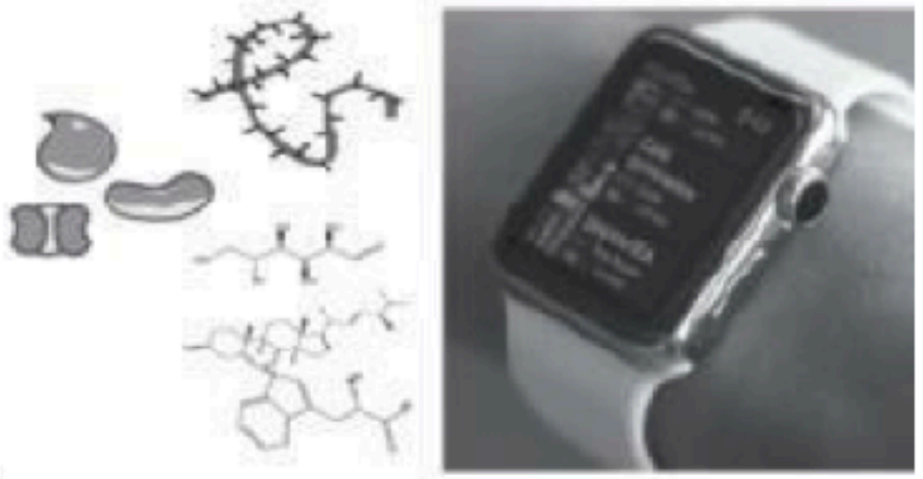
- In cases such as *BRCA* mutation, this is likely. For others it is perhaps not so clear.
  - ➡ E.g. Obesity is a major risk factor for many diseases (including heart attack, diabetes and cancer)
  - ➡ The solution is obvious - controlled diet and exercise. Yet many overweight and obese people do not do either.
  - ➡ Why would genome sequencing have a better outcome?

- It is likely that sequencing will result in early detection and more accurate diagnosis.
- People may begin to use medicine before the disease arises:
  - ➡ E.g. Statins for heart problems, metformin diabetes drugs, and new Alzheimer's drugs that could be more effective prior to disease onset.

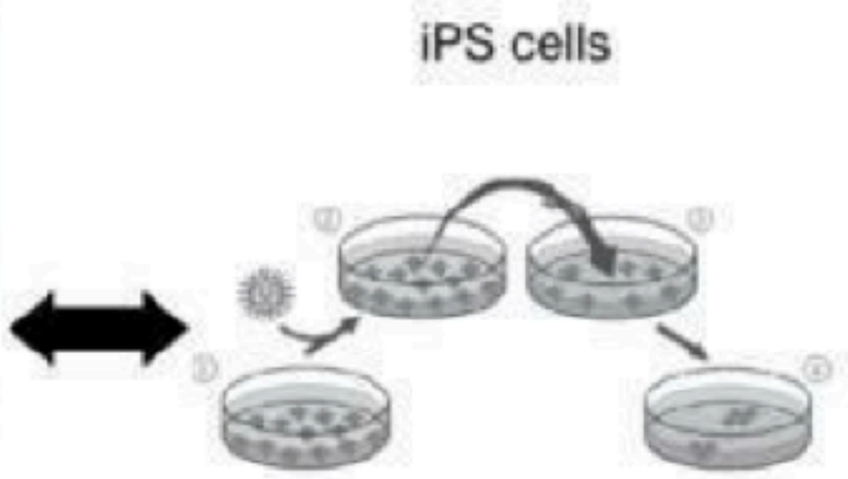
Question:

What might the future look like?





1. Predict risk
2. Early diagnoses
3. Monitor
4. Treat



- **Many people will have their genome sequenced, likely before birth, and epigenomes and a wealth of other information will be commonly used to predict, diagnose and treat disease.**

- **This will be increasingly coupled with tracking molecular information and diet etc. into a system for managing and maintaining each person's health.**

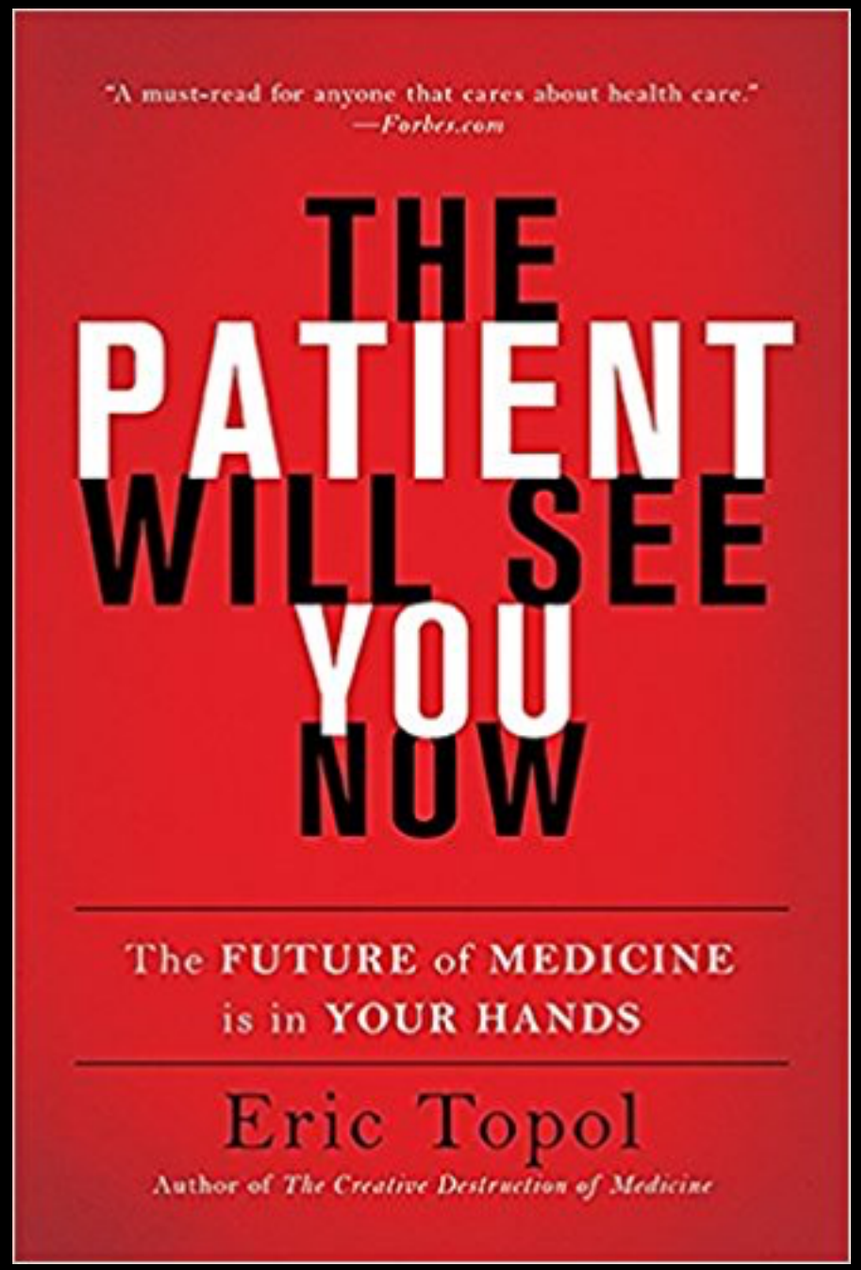
Implications

- Physicians of the future need to ensure they are well-educated in genomics and bioinformatics so that they can optimally apply the information generated to patient care.

- Medicine will become more information rich, more accurate and more predictive.

As patients we will have more control of our medical destiny and that of our children.

Book recommendation



- Eric Topol argues that power will shit from doctors to patients, and he emphasizes the comping age of **patient-centered medicine**, in which patients generate medical data using their own digital devices and communicating and accessing their data via their smartphones.

Thats all folks!

Happy St Patrick's Weekend!

Bonus Methods Slides

For Reference

Types of machine learning

- Unsupervised learning
  - ➡ Finding structure in unlabeled data

- Supervised learning
  - ➡ Making predictions based on labeled data
  - ➡ Predictions like regression or classification

- Reinforcement learning
  - ➡ Making decisions based on past experience

Types of machine learning

- Unsupervised learning
  - ➡ Finding structure in unlabeled data

- Supervised learning
  - ➡ Making predictions based on labeled data
  - ➡ Predictions like regression or classification

- Reinforcement learning
  - ➡ Making decisions based on past experience

- Introduction to machine learning
  - Unsupervised, supervised and reinforcement learning

- Clustering
  - K-means clustering
  - Hierarchical clustering
  - Heatmap representations

- Dimensionality reduction, visualization and ‘structure’ analysis
  - Principal Component Analysis (PCA)

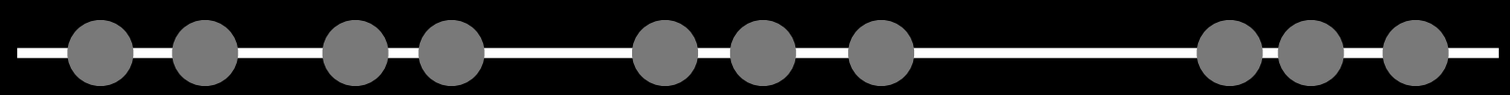
- Network analysis

k-means clustering algorithm

- Breaks observations into ***k*** pre-defined number of clusters
- You define ***k*** the number of clusters!

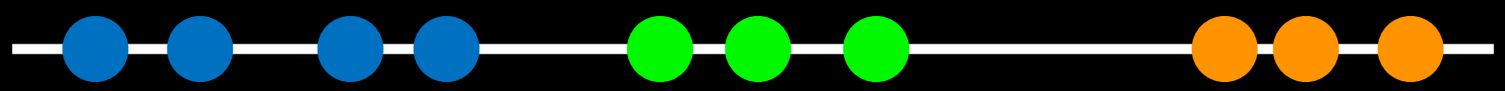
k-means clustering algorithm

- Breaks observations into ***k*** pre-defined number of clusters
- You define ***k*** the number of clusters!
  - ➡ Imagine you had data that you could plot along a line and you knew you had to put them into *k*=3 “clusters” (e.g. data from three types of tumor cells)



k-means clustering algorithm

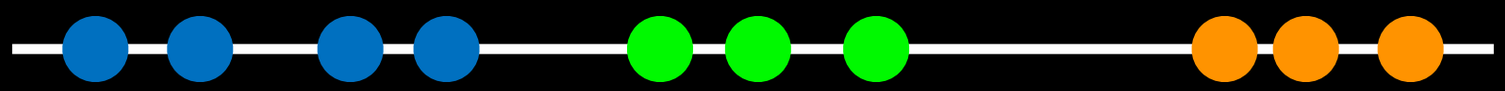
- Breaks observations into ***k*** pre-defined number of clusters
- You define ***k*** the number of clusters!
  - ➡ Imagine you had data that you could plot along a line and you knew you had to put them into *k*=3 “clusters” (e.g. data from three types of tumor cells)



Here your eyes can clearly see 3 natural groupings

k-means clustering algorithm

- Breaks observations into ***k*** pre-defined number of clusters
- You define ***k*** the number of clusters!
  - ➡ Imagine you had data that you could plot along a line and you knew you had to put them into *k*=3 “clusters” (e.g. data from three types of tumor cells)



Here your eyes can clearly see 3 natural groupings
How does k-means attempt to define this grouping?

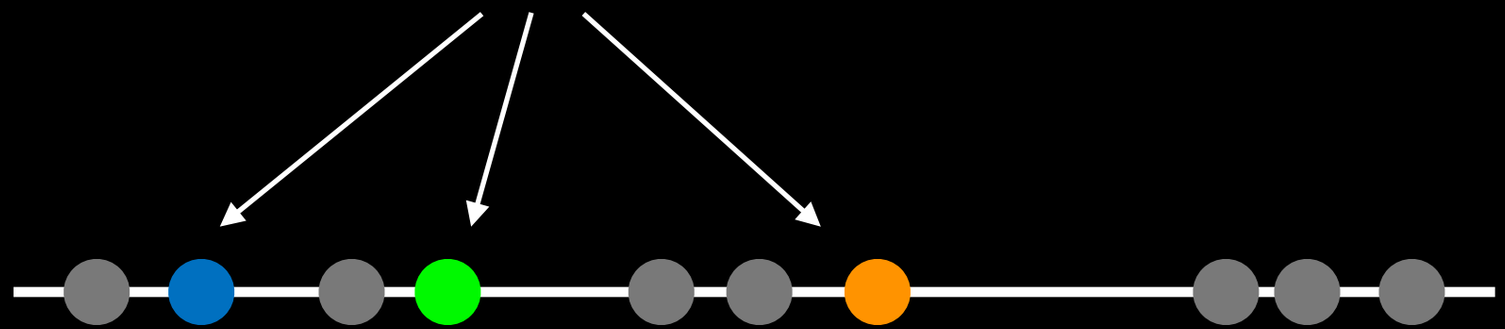
Step 1.

Select ***k*** (the number of clusters)

Step 2.

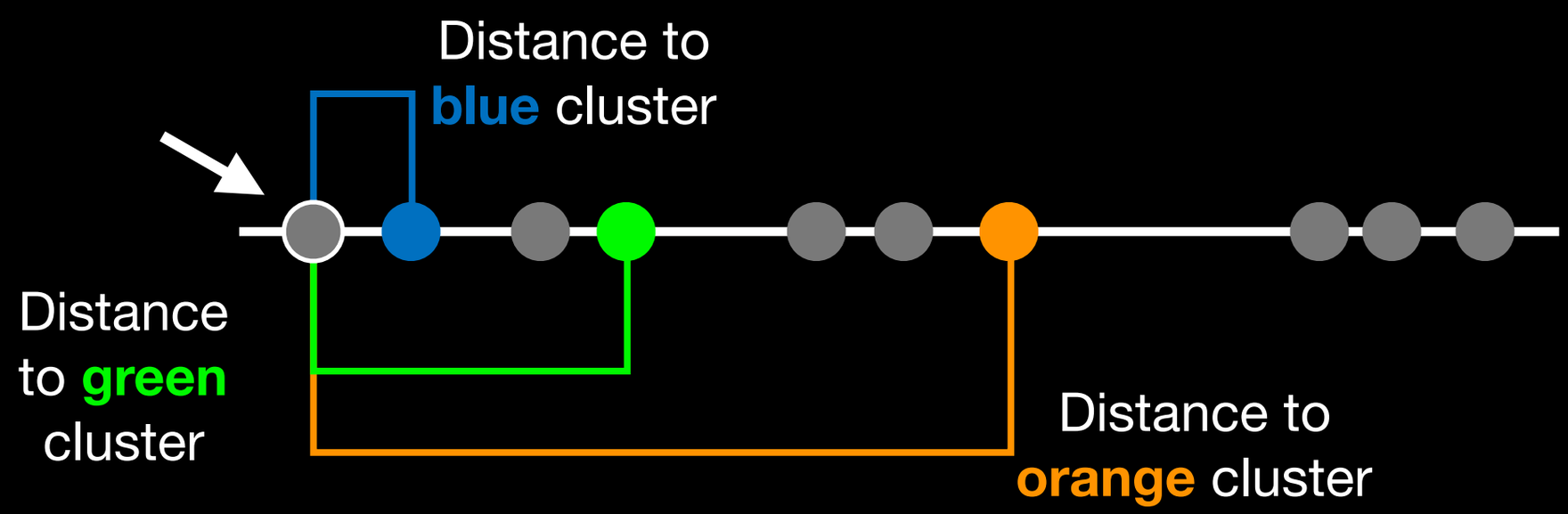
Select ***k=3*** distant data points at random

These are the initial clusters



Step 3.

Measure distance between the 1st point and the ***k=3*** initial clusters



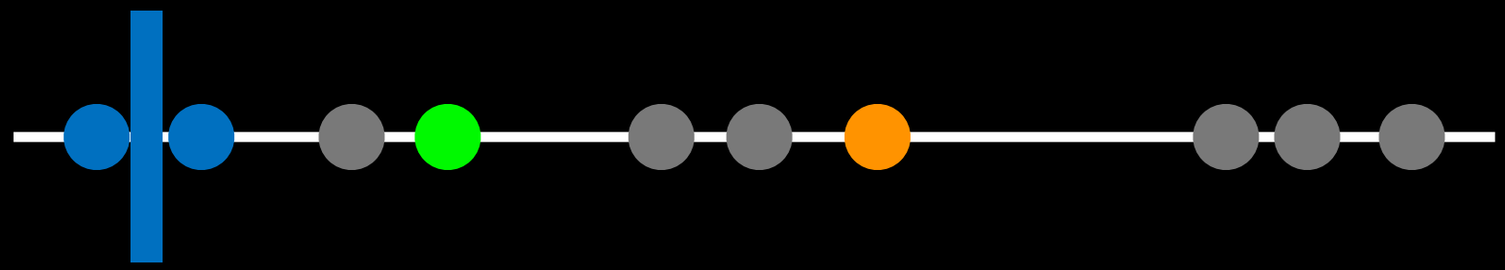
Step 4.

Assign the 1st point to the nearest cluster

Step 5.

Update cluster centers

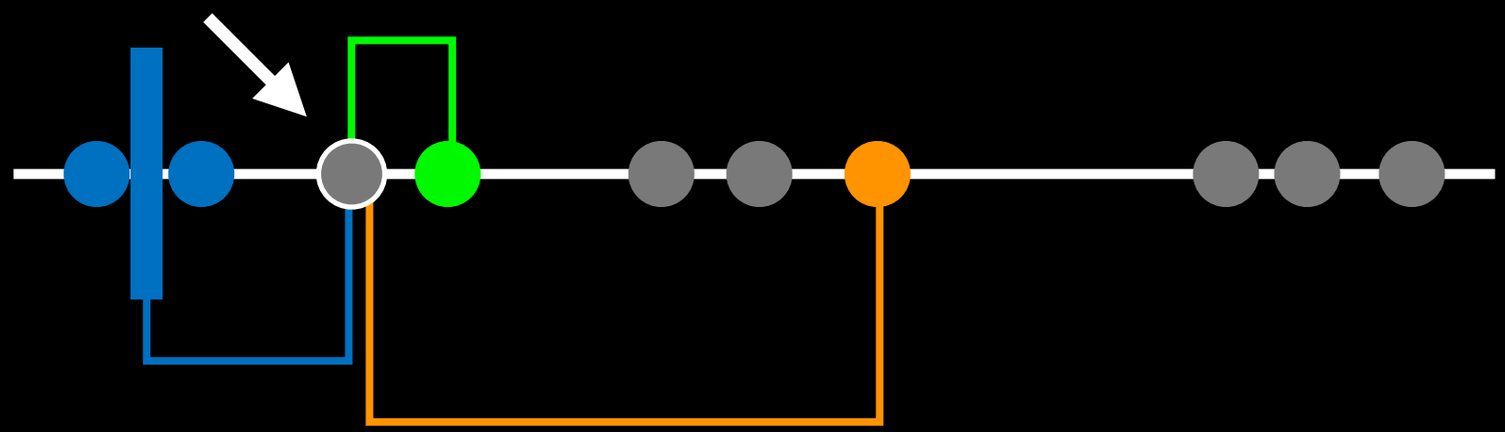
Calculate the mean value for the blue cluster including the new point



Step 6.

Assign next point to closest cluster

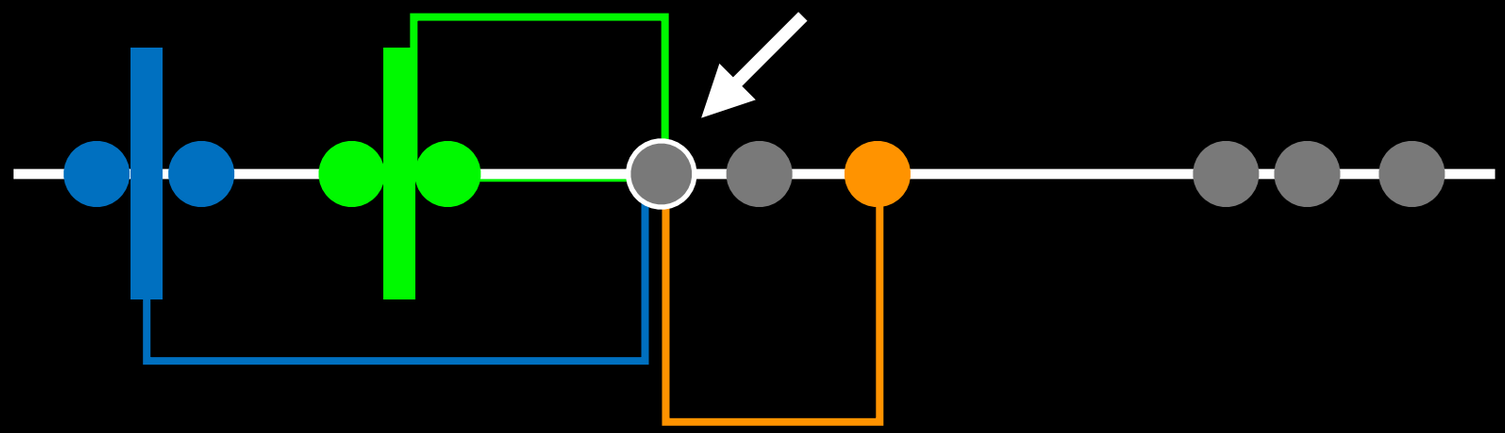
Use updated cluster centers for distance calculation



Step 7.

Update cluster centers and move to next point

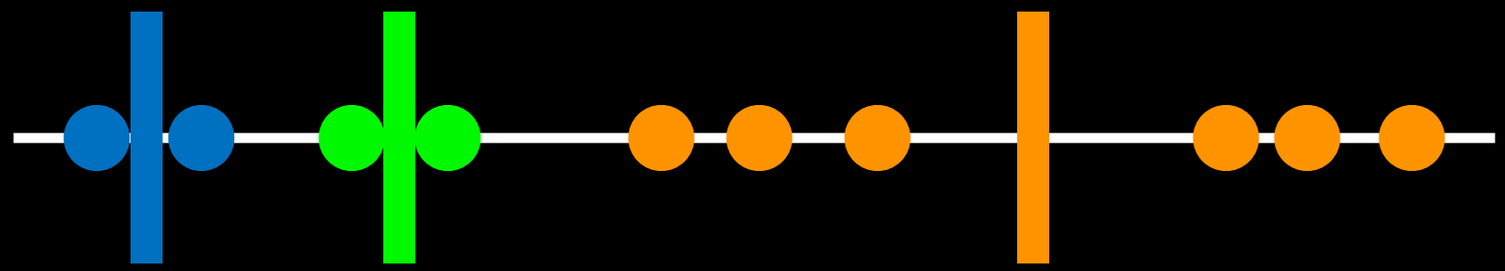
Use updated cluster centers for distance calculation



Step 8.

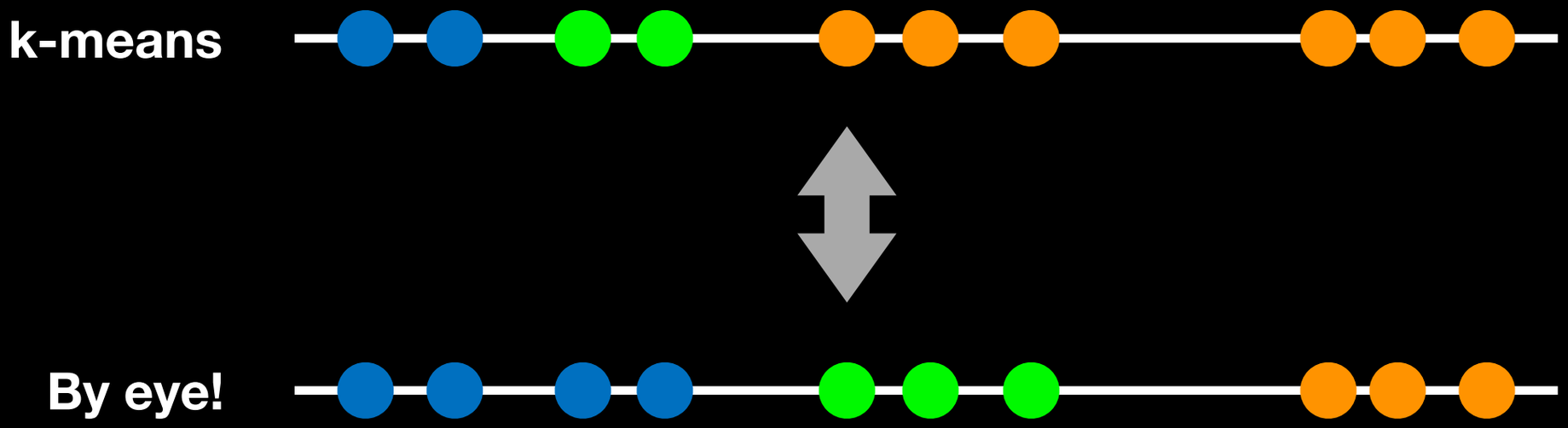
Repeat for each point

Each time updating cluster centers



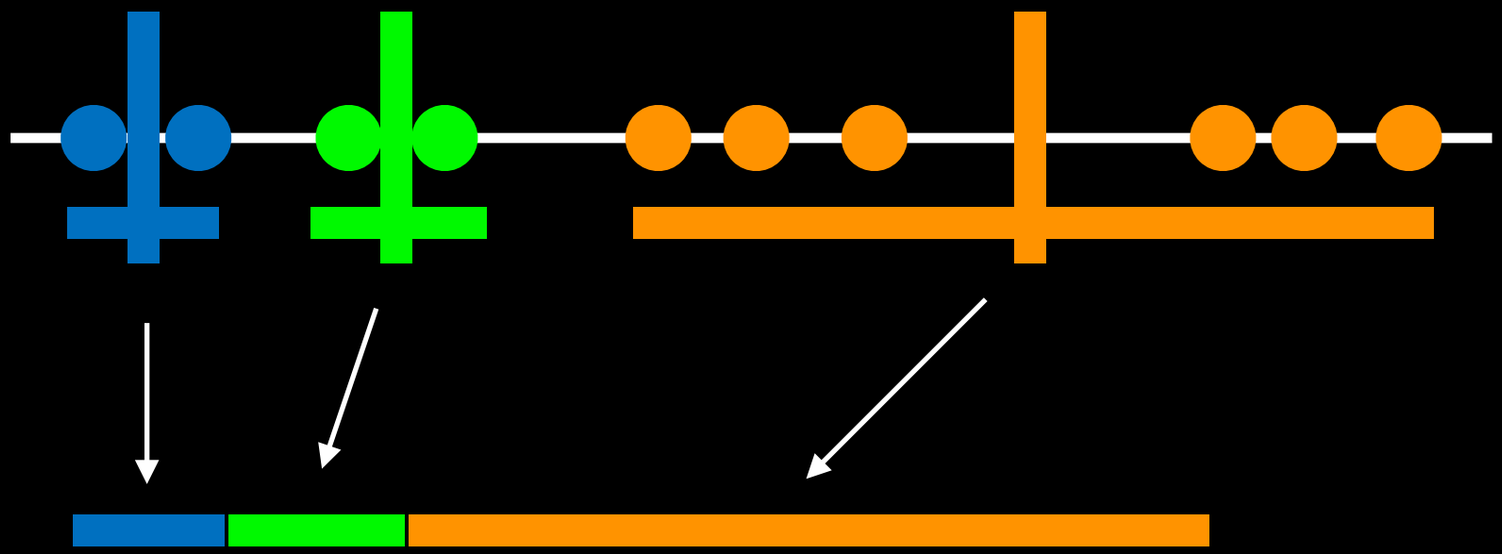
Hmm….

Here the k-means result does not look as good as what we were able to do by eye!



Step 9.

Assess the quality of the clustering by adding up the variation within each cluster



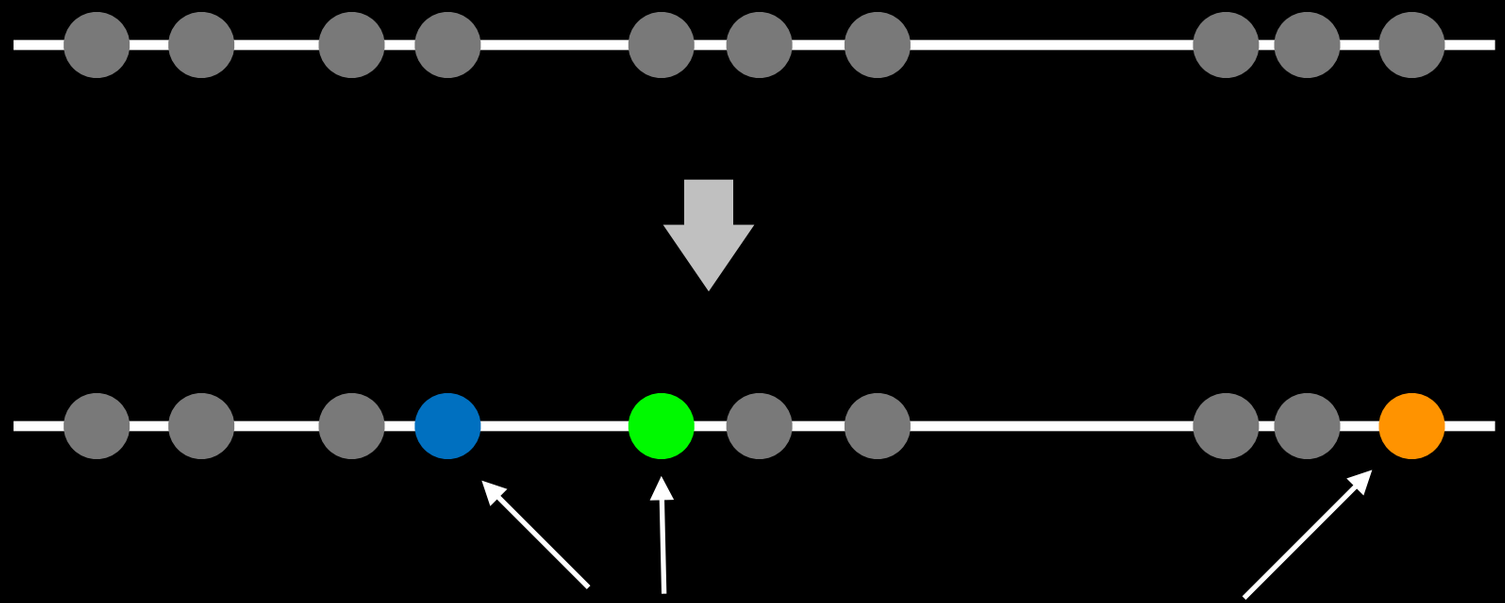
The total variation within clusters

K-means keeps track of these clusters and their total **variance** and then does the whole thing over again with different starting points

Step 10.

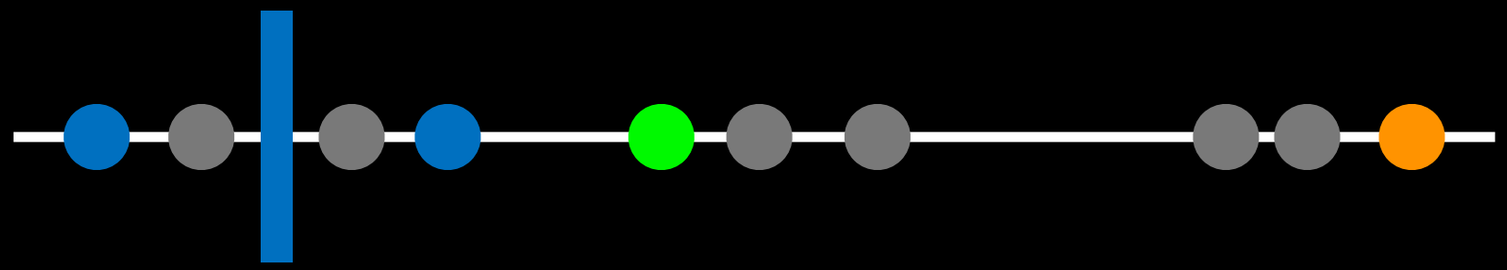
Repeat with different starting points

Back to the beginning and do all steps over again…

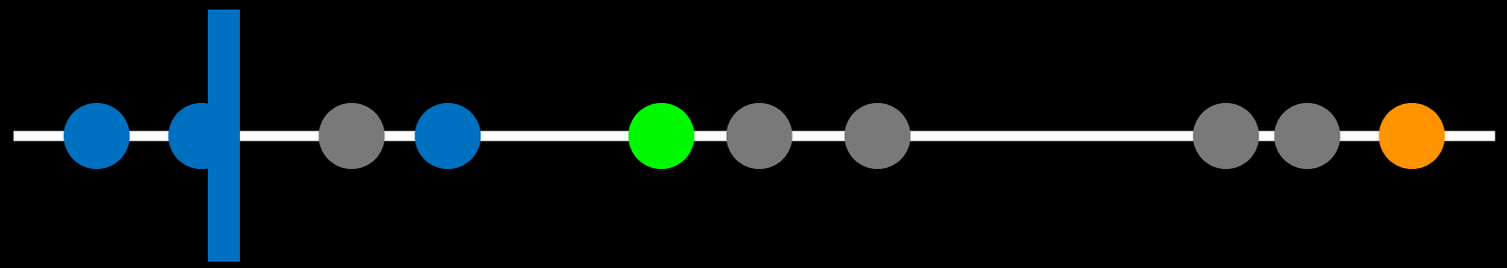


…Pick *k*=3 initial clusters and add the remaining points to the cluster with the nearest mean, recalculating the mean each time a new point is added…

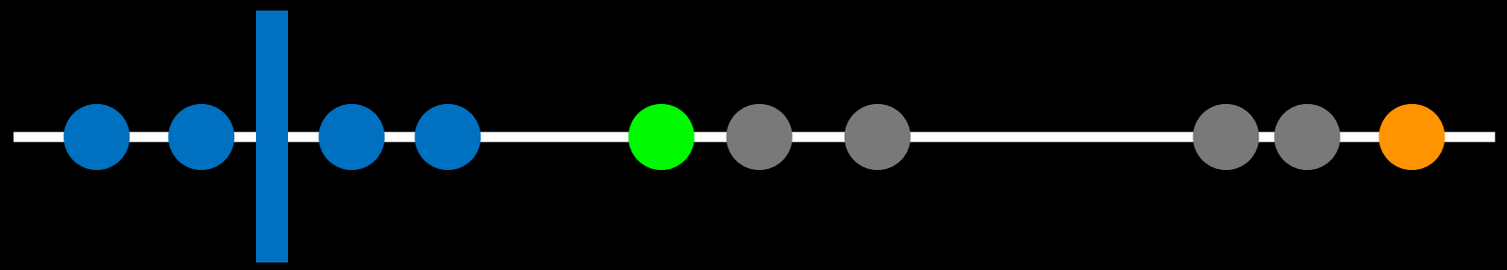
…Pick *k*=3 initial clusters and add the remaining points to the cluster with the nearest mean, recalculating the mean each time a new point is added…



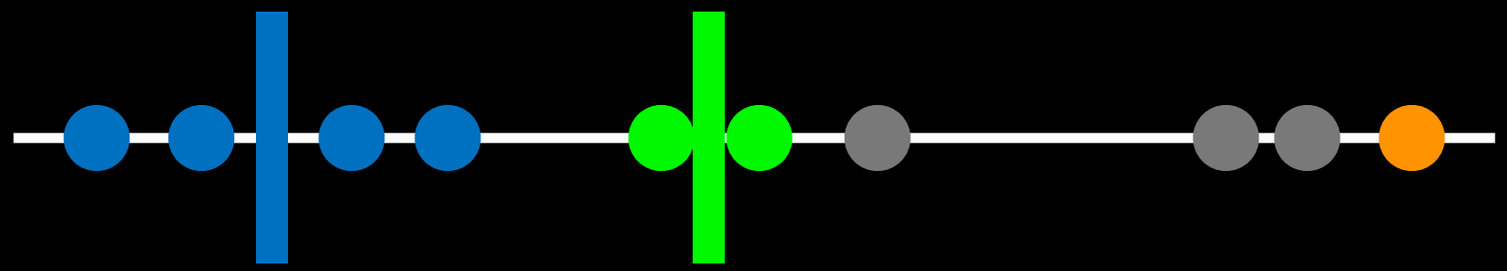
…Pick *k*=3 initial clusters and add the remaining points to the cluster with the nearest mean, recalculating the mean each time a new point is added…



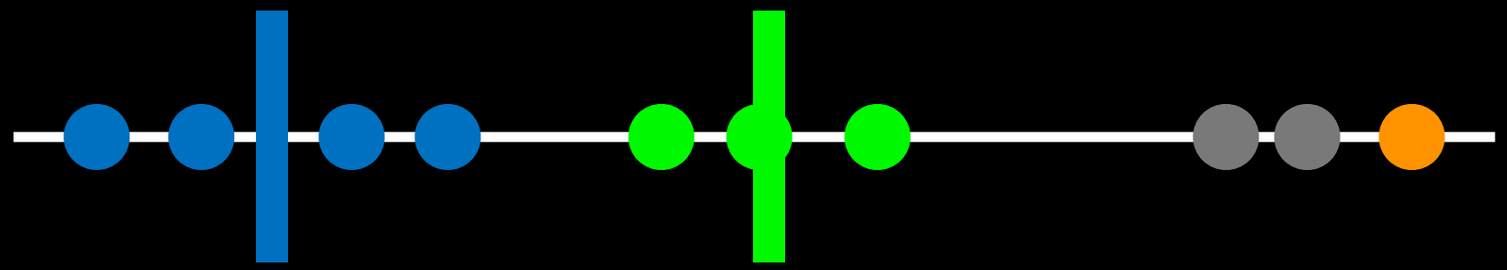
…Pick *k*=3 initial clusters and add the remaining points to the cluster with the nearest mean, recalculating the mean each time a new point is added…



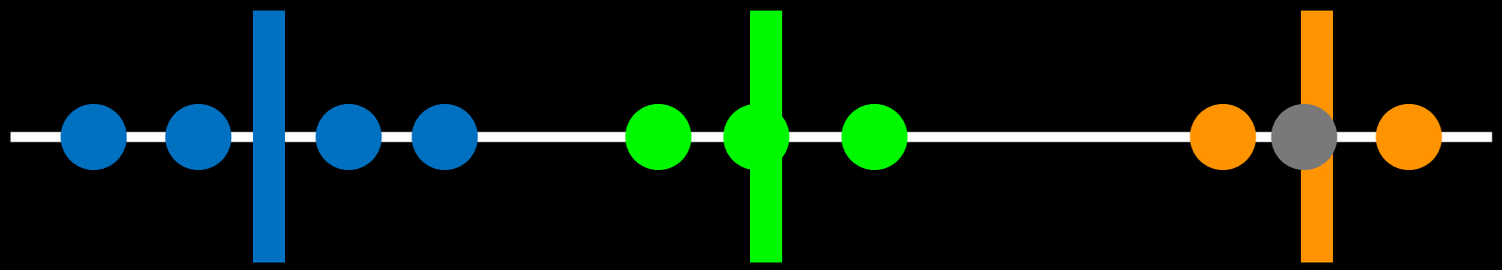
…Pick *k*=3 initial clusters and add the remaining points to the cluster with the nearest mean, recalculating the mean each time a new point is added…



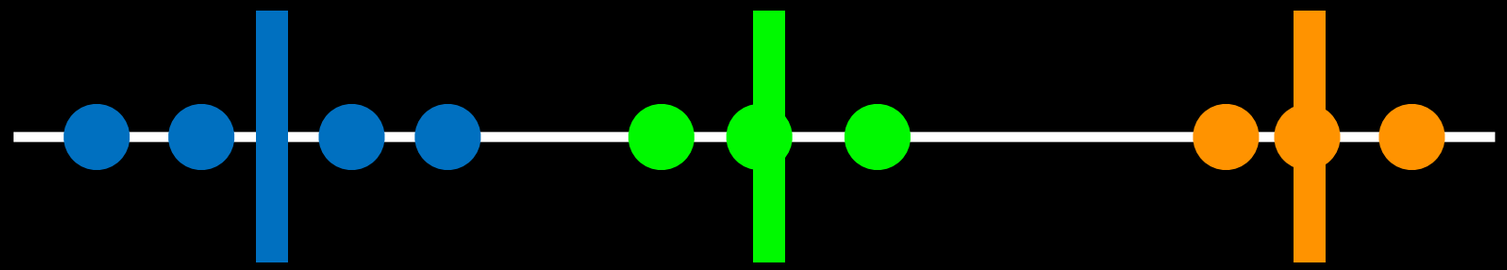
…Pick *k*=3 initial clusters and add the remaining points to the cluster with the nearest mean, recalculating the mean each time a new point is added…



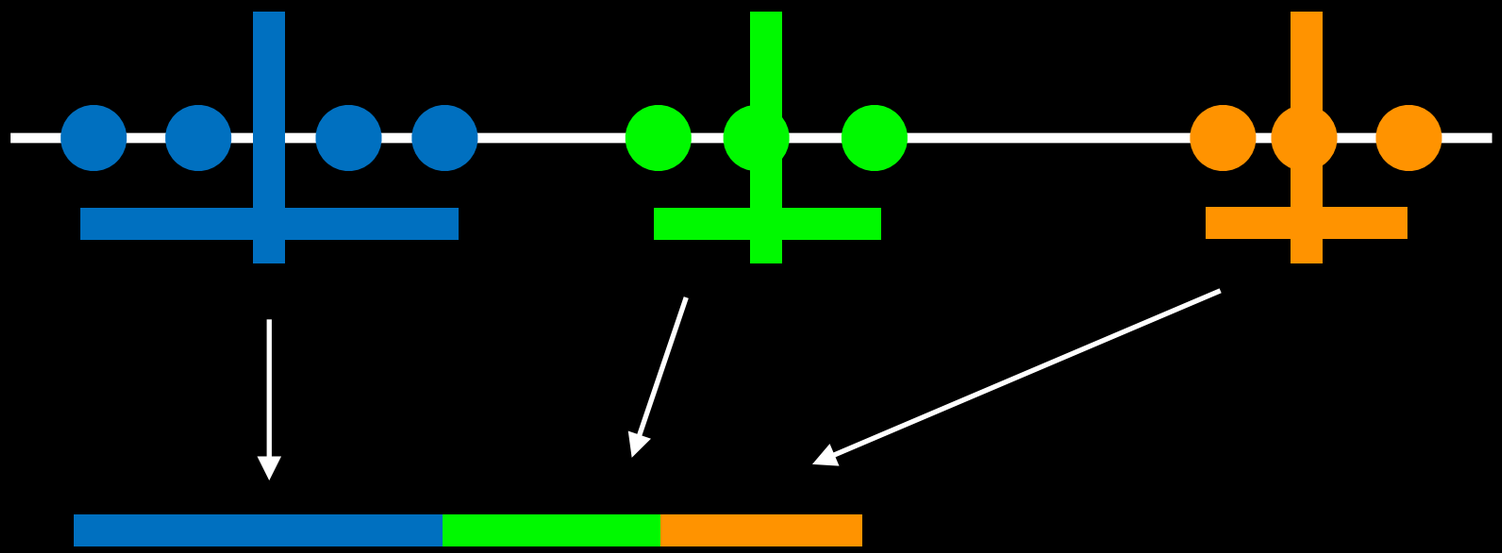
…Pick *k*=3 initial clusters and add the remaining points to the cluster with the nearest mean, recalculating the mean each time a new point is added…



…Pick *k*=3 initial clusters and add the remaining points to the cluster with the nearest mean, recalculating the mean each time a new point is added…



Now the data are all assigned to clusters, we again sum the variation within each cluster



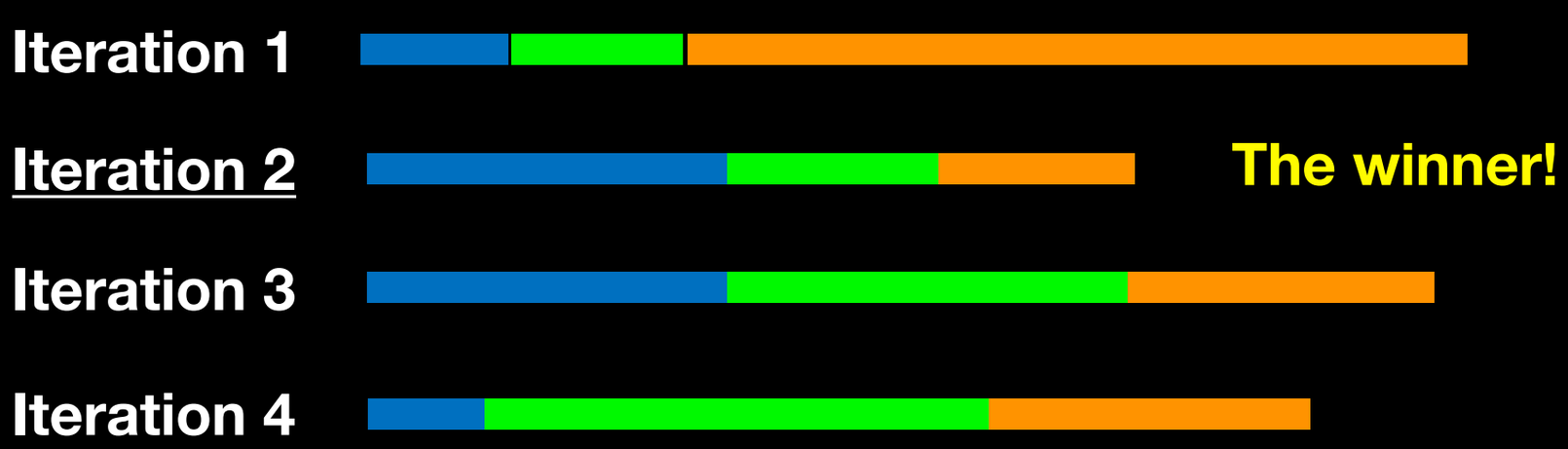
The total variation within clusters

Step 10.

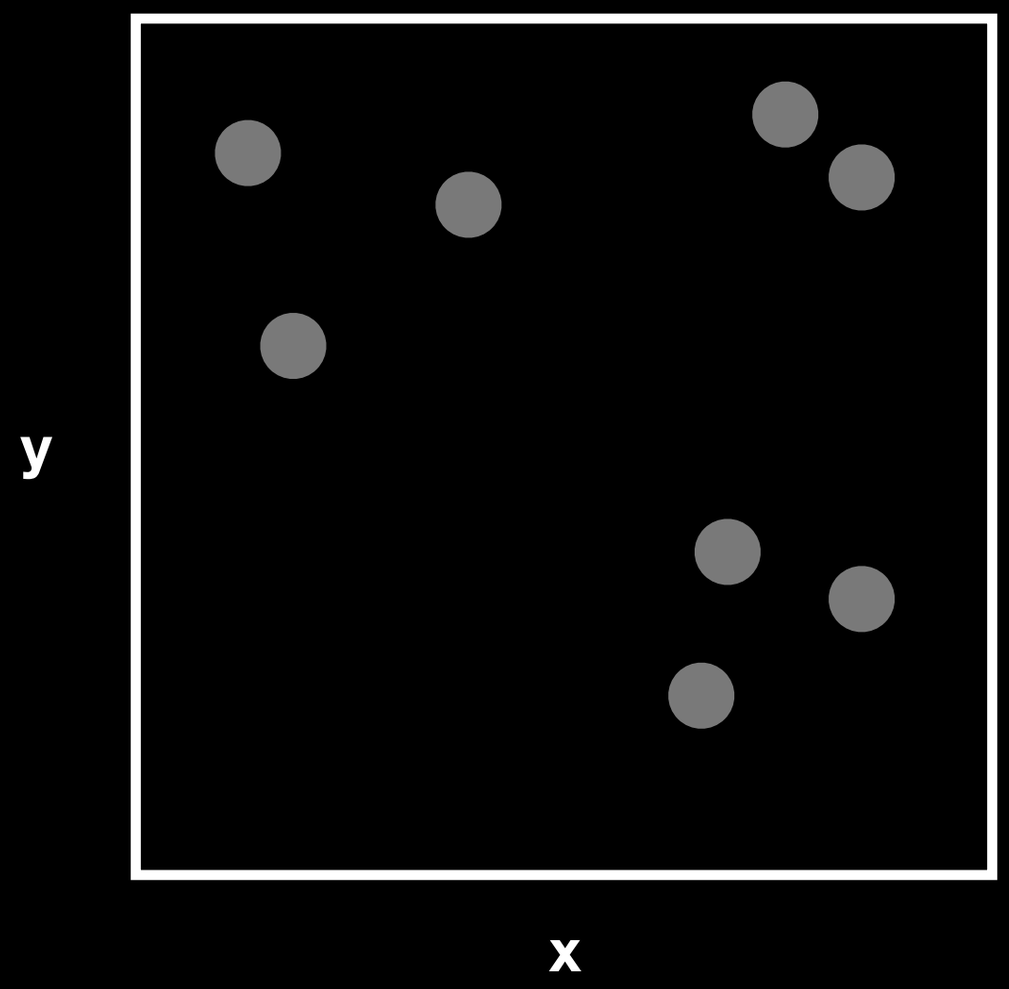
Repeat again with different starting points

After several iterations k-means clustering knows it has the *best clustering so-far* based on the smallest total variation with clusters.

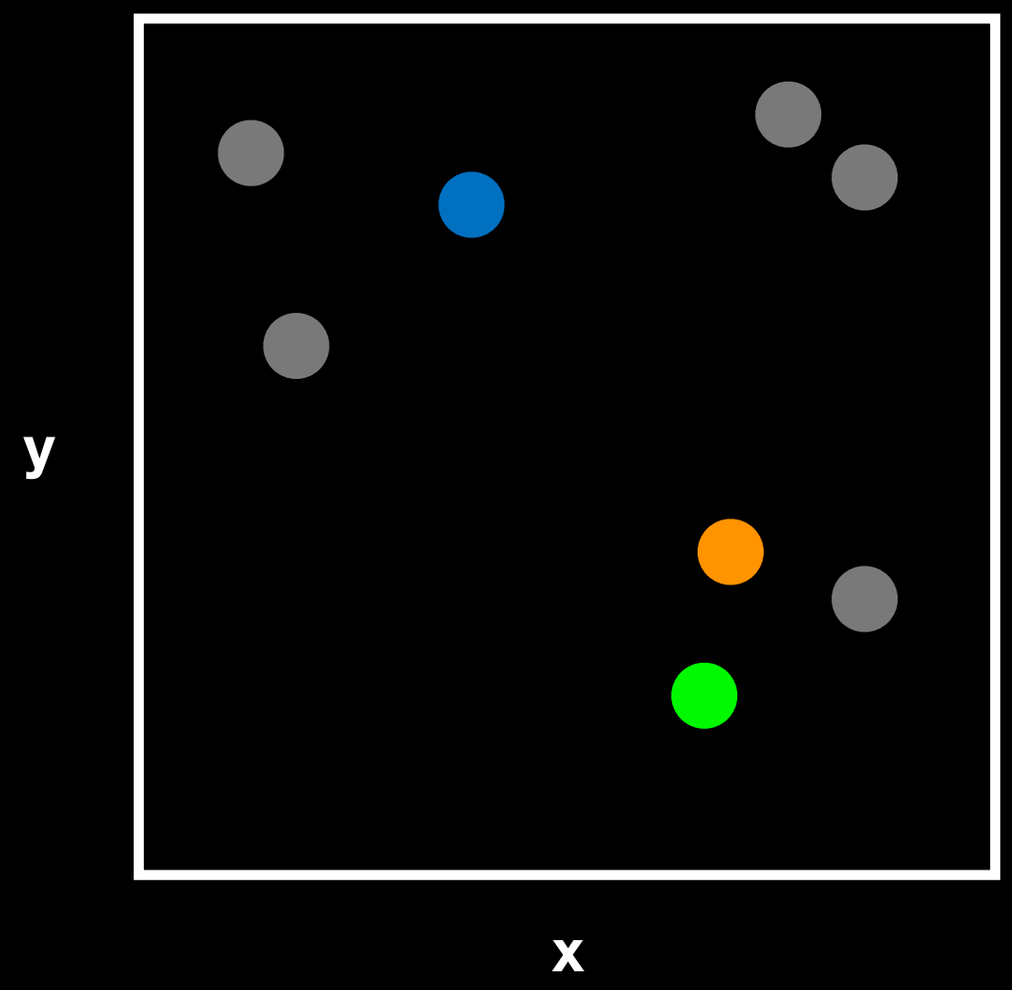
However, it does not know if it has found the *best overall*. So it will perform several more iterations with different starting points…



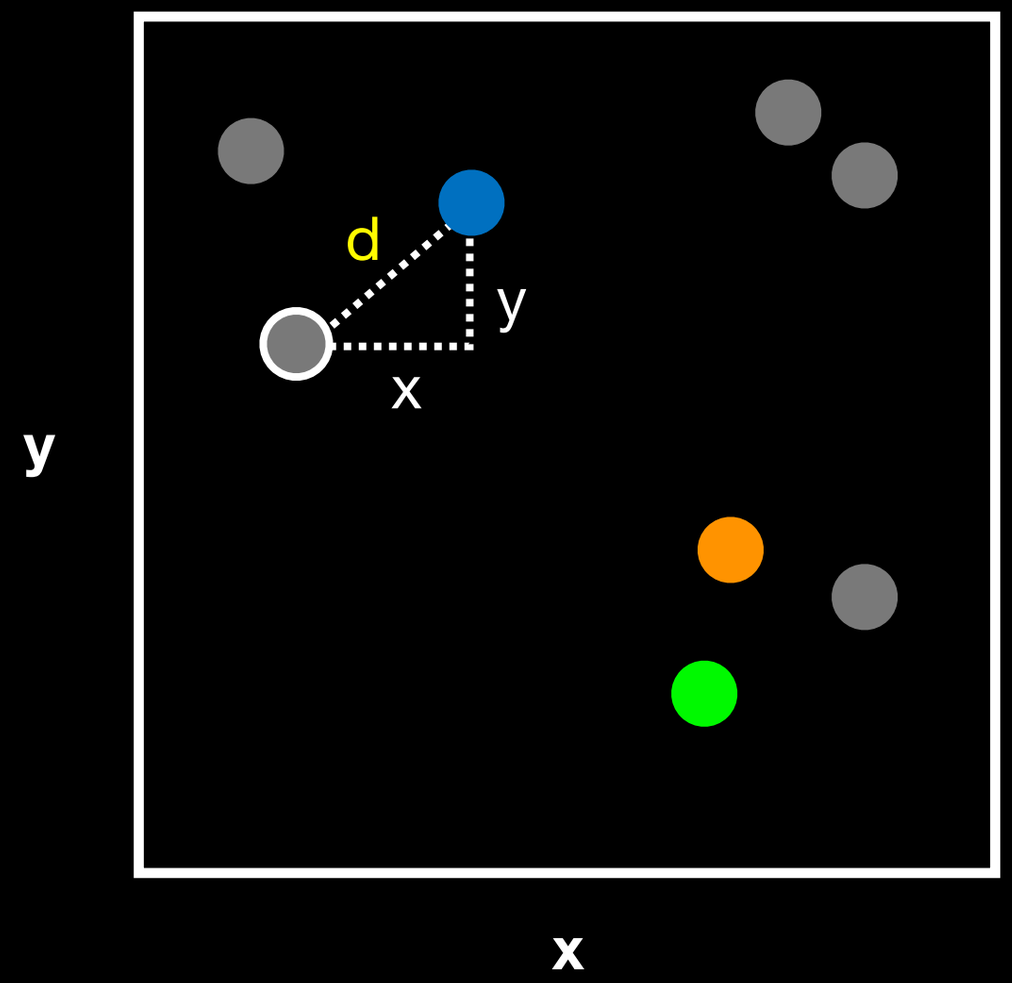
What if we have more dimensions?



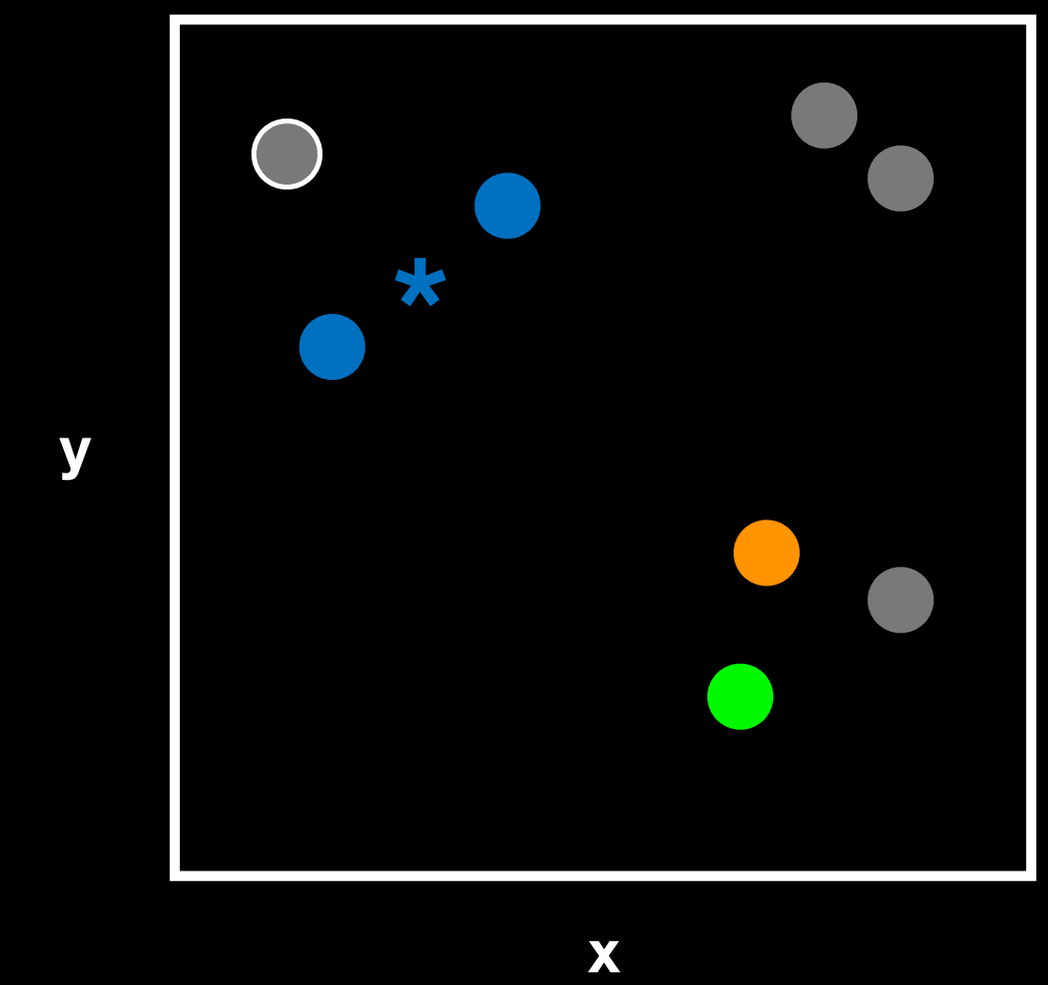
Just like before, we pick 3 random points…



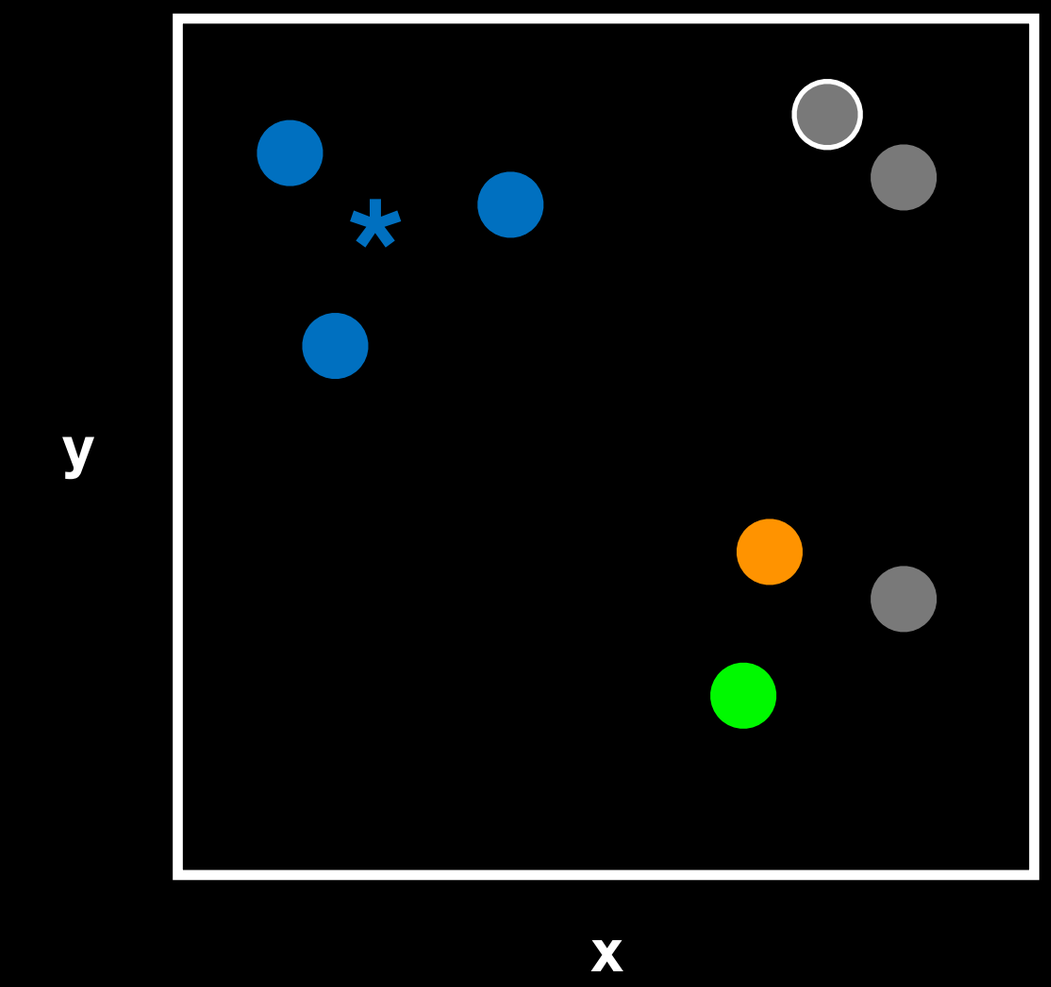
…and use the Euclidean distance.
In 2 dimensions the Euclidean distance is the same as the Pythagorean theorem d = sqrt(x^2 + y^2)



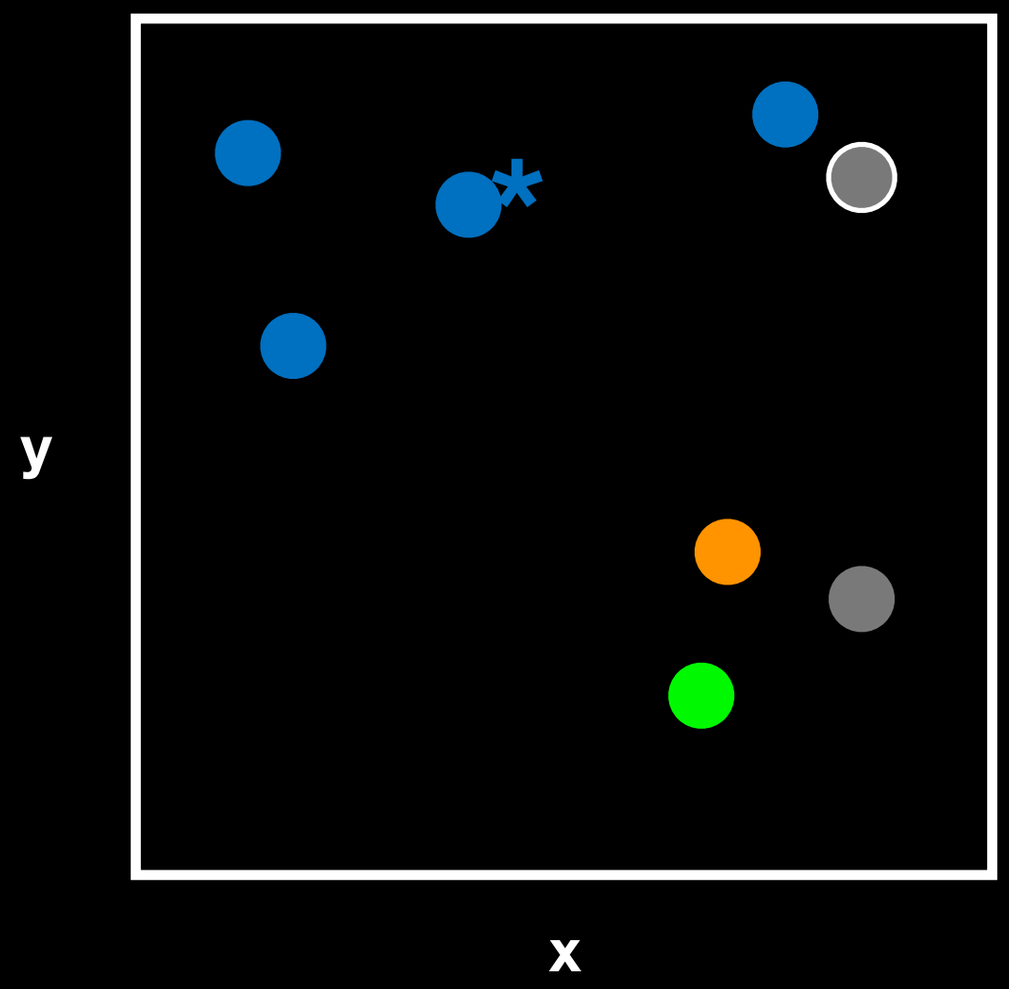
…assign point to nearest cluster and update cluster center *



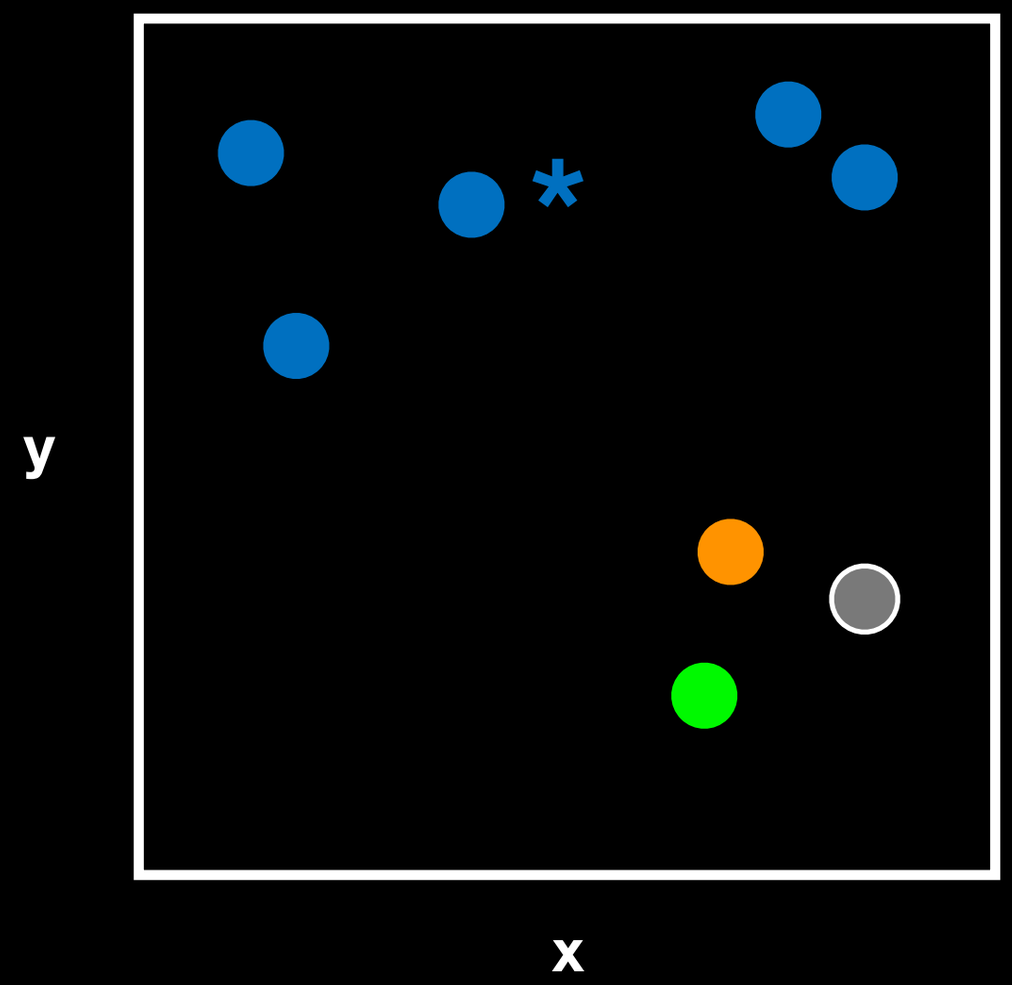
…and continue



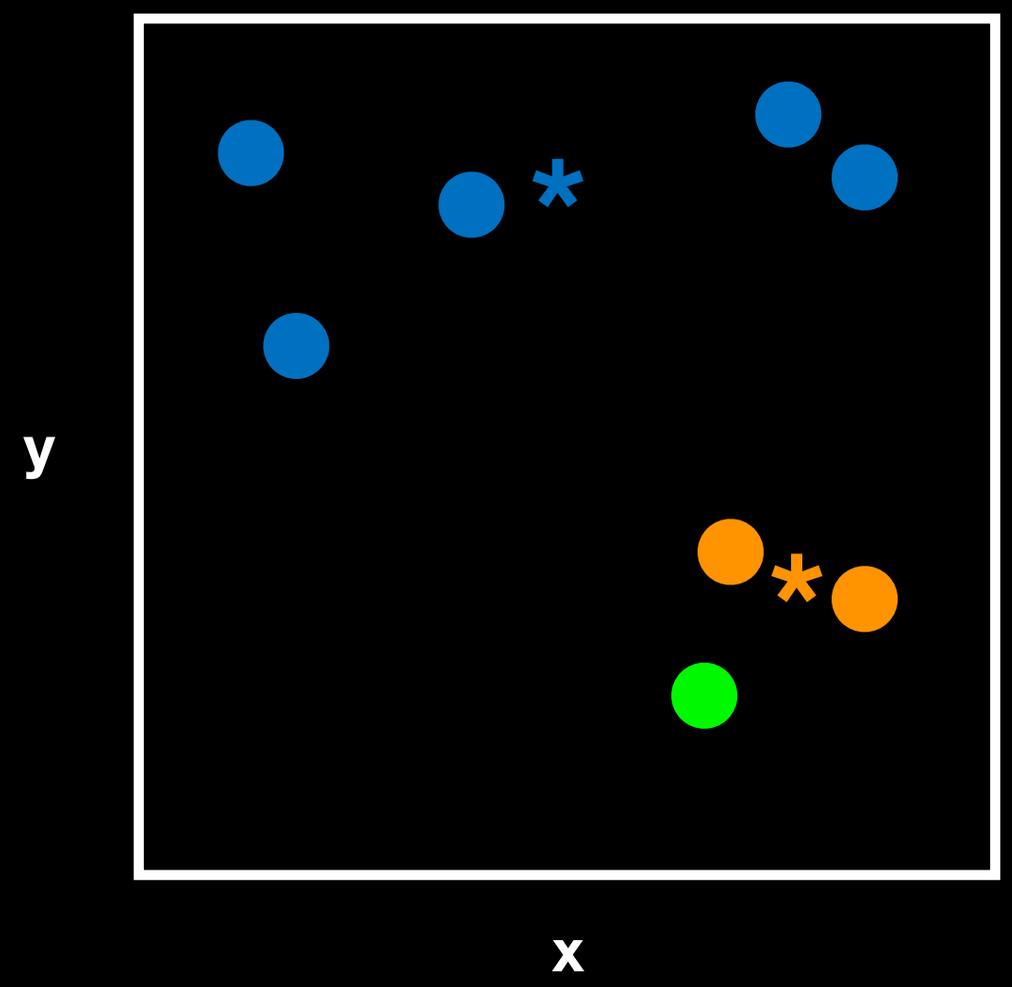
…and continue



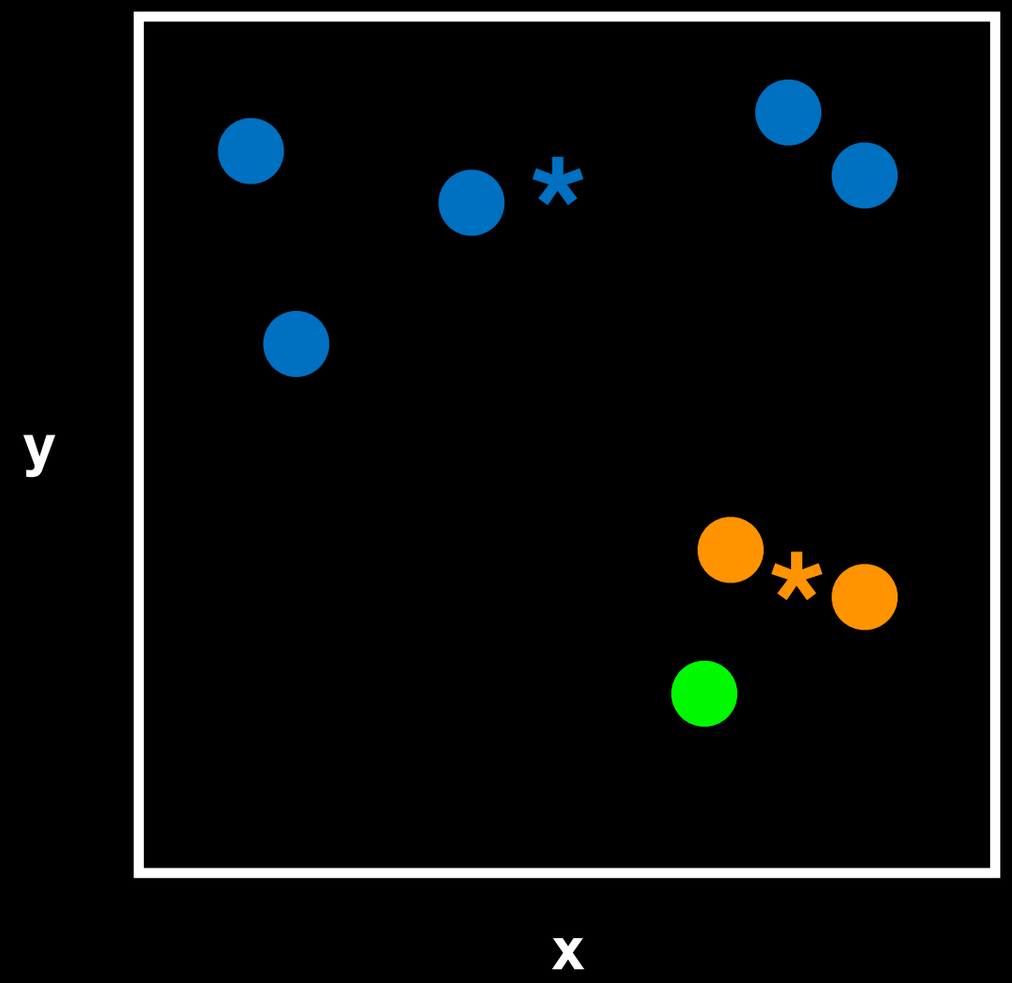
…and continue



…and continue



Again we have to use a number of different starting conditions before deciding on a good clustering!



What if we have even more dimensions?

Cell Samples

	#1	#2	#3
Gene 1	12	6	-13
Gene 2	-7	13	10
Gene 3	8	6	-9
Gene 4	9	5	-11
Gene 5	-3	1	6
Gene 6	10	4	-8

What if we have even more dimensions?

Cell Samples

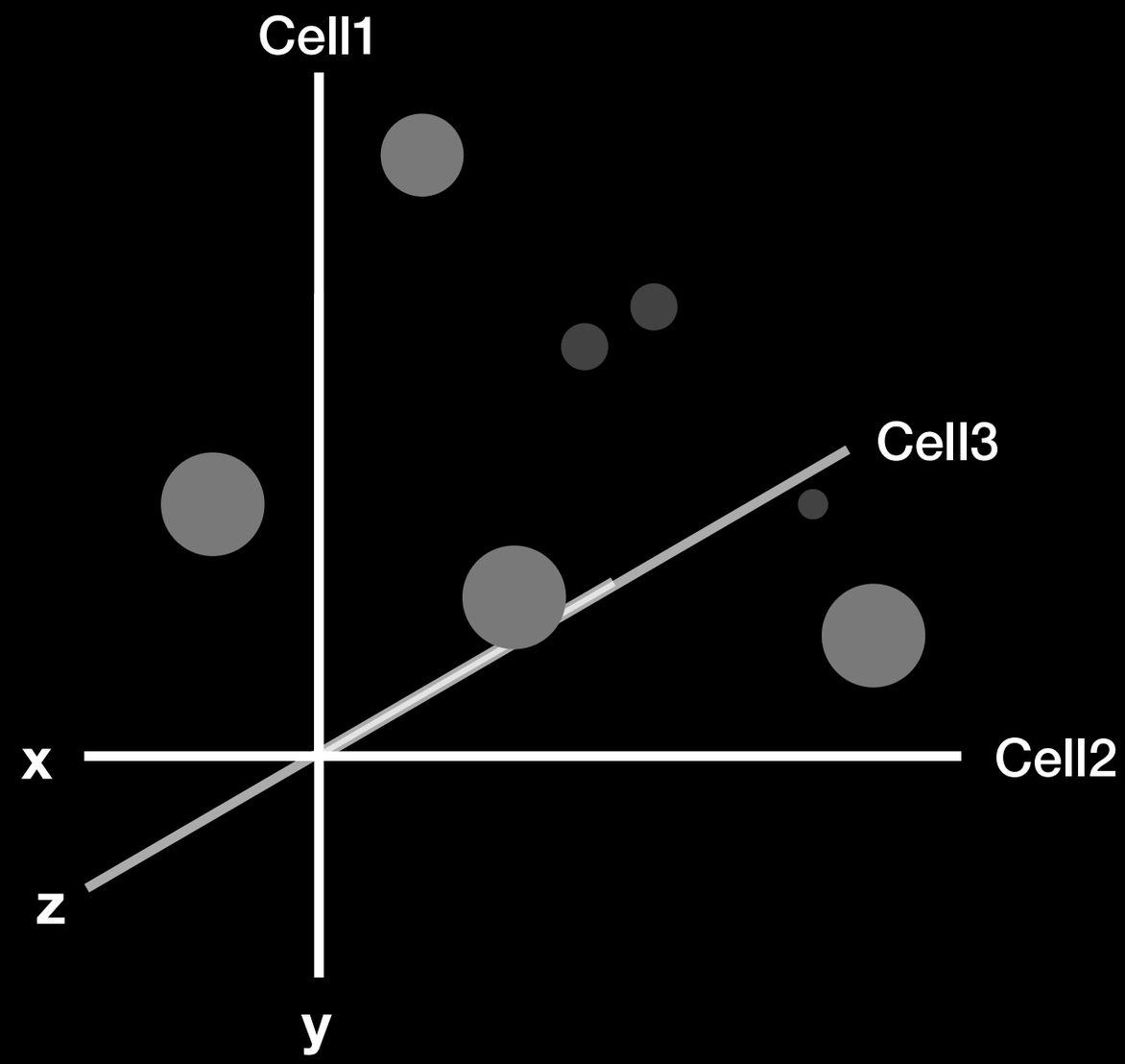
	#1	#2	#3
Gene 1	12	6	-13
Gene 2	-7	13	10
Gene 3	8	6	-9
Gene 4	9	5	-11
Gene 5	-3	1	6
Gene 6	10	4	-8
	x	**y**	**z**

We could simply plot them by relabeling the cell samples as **x**, **y**, and **z** (i.e. a 3D plot)

What if we have even more dimensions?

Cell Samples

	#1	#2	#3
Gene 1	12	6	-13
Gene 2	-7	13	10
Gene 3	8	6	-9
Gene 4	9	5	-11
Gene 5	-3	1	6
Gene 6	10	4	-8
	x	**y**	**z**

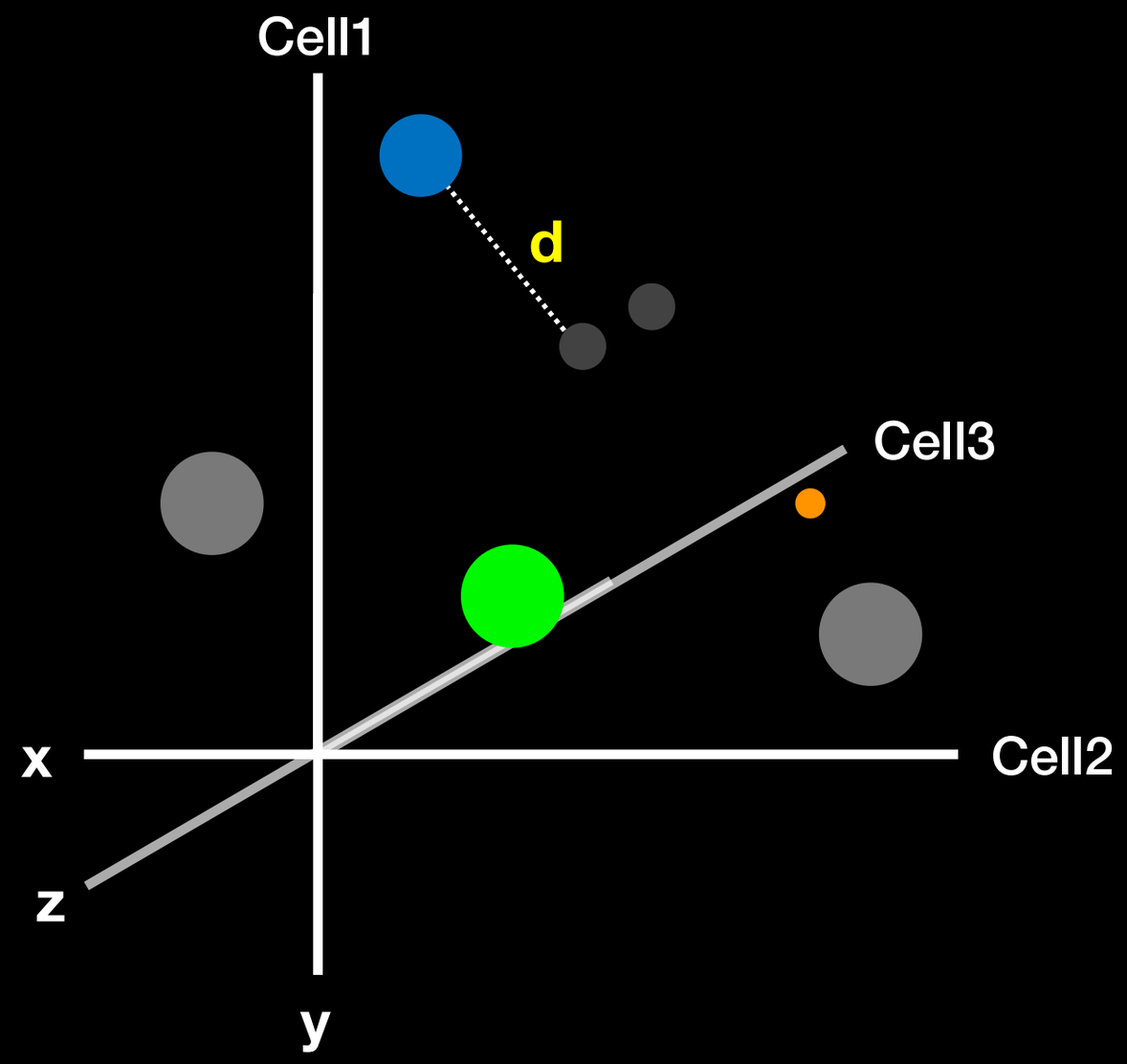


…and go through exactly the same procedure with initial cluster assignment followed by distance calculation etc…

$$d = \sqrt{x^2 + y^2 + z^2}$$

Cell Samples

	#1	#2	#3
Gene 1	12	6	-13
Gene 2	-7	13	10
Gene 3	8	6	-9
Gene 4	9	5	-11
Gene 5	-3	1	6
Gene 6	10	4	-8
	x	**y**	**z**



…and go through exactly the same procedure with initial cluster assignment followed by distance calculation etc…

d = sqrt(x^2 + y^2 + z^2)

Cell Samples

	#1	#2	#3
Gene 1	12	6	-13
Gene 2	-7	13	10
Gene 3	8	6	-9
Gene 4	9	5	-11
Gene 5	-3	1	6
Gene 6	10	4	-8

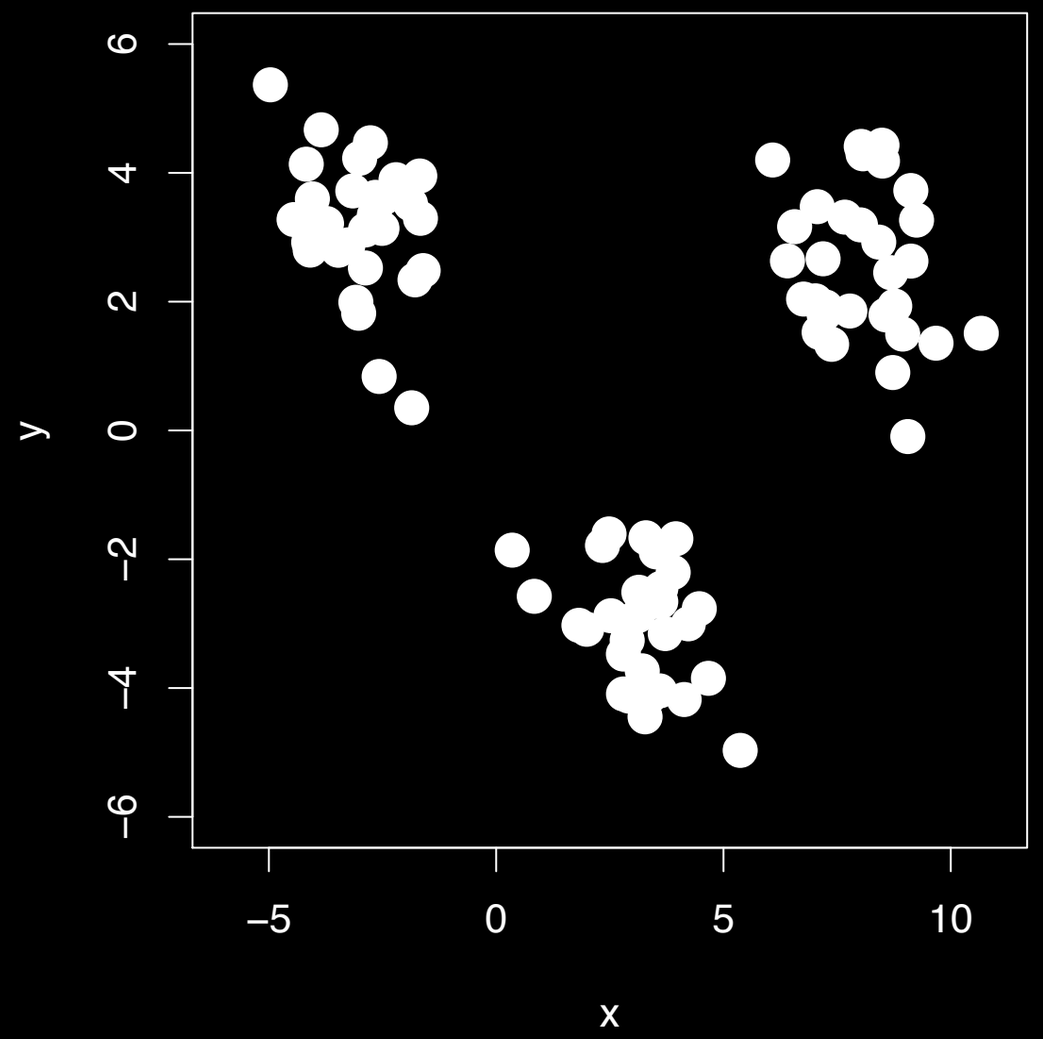
Of course we don't actually need to plot anything.

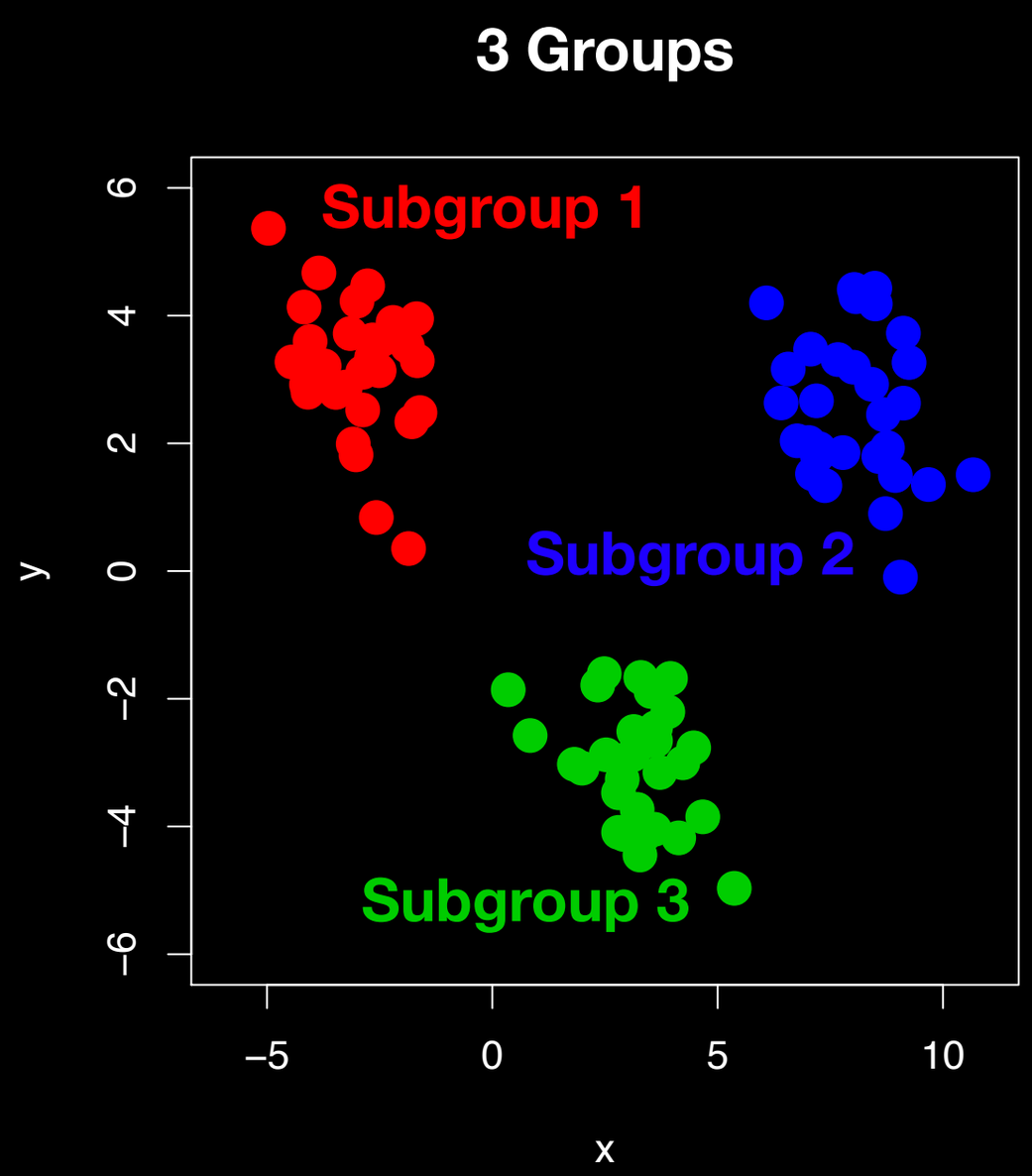
We can just calculate the Euclidean distance along any number of dimensions and perform our k-means clustering in the same way.

k-means in R

```
# k-means algorithm with 3 centers, run 20 times
kmeans(x, centers= 3, nstart= 20)
```

- Input **x** is a numeric matrix, or data.frame, with one observation per row, one feature per column
- k-means has a random component
- Run algorithm multiple times to improve odds of the best model


y
6
4
2
0
−2
−4
−6
−5
0
5
10
x


3 Groups
Subgroup 1
Subgroup 2
Subgroup 3
y
x
6
4
2
0
−2
−4
−6
−5
0
5
10

Model selection

- Recall k-means has a random component
- Best outcome is based on total within cluster sum of squares:
  - ➡ For each cluster
    - For each observation in the cluster
      - Determine squared distance from observation to cluster center
  - ➡ Sum all of them together

Model selection

```
# k-means algorithm with 5 centers, run 20 times
kmeans(x, centers=5, nstart=20)
```

- Running algorithm multiple times (i.e. setting `nstart`) helps find the global minimum total within cluster sum of squares

- Increasing the default value of `nstart` is often sensible

- Introduction to machine learning
  - Unsupervised, supervised and reinforcement learning

- Clustering
  - K-means clustering
  - Hierarchical clustering
  - Heatmap representations

- Dimensionality reduction, visualization and ‘structure’ analysis
  - Principal Component Analysis (PCA)

- Network analysis

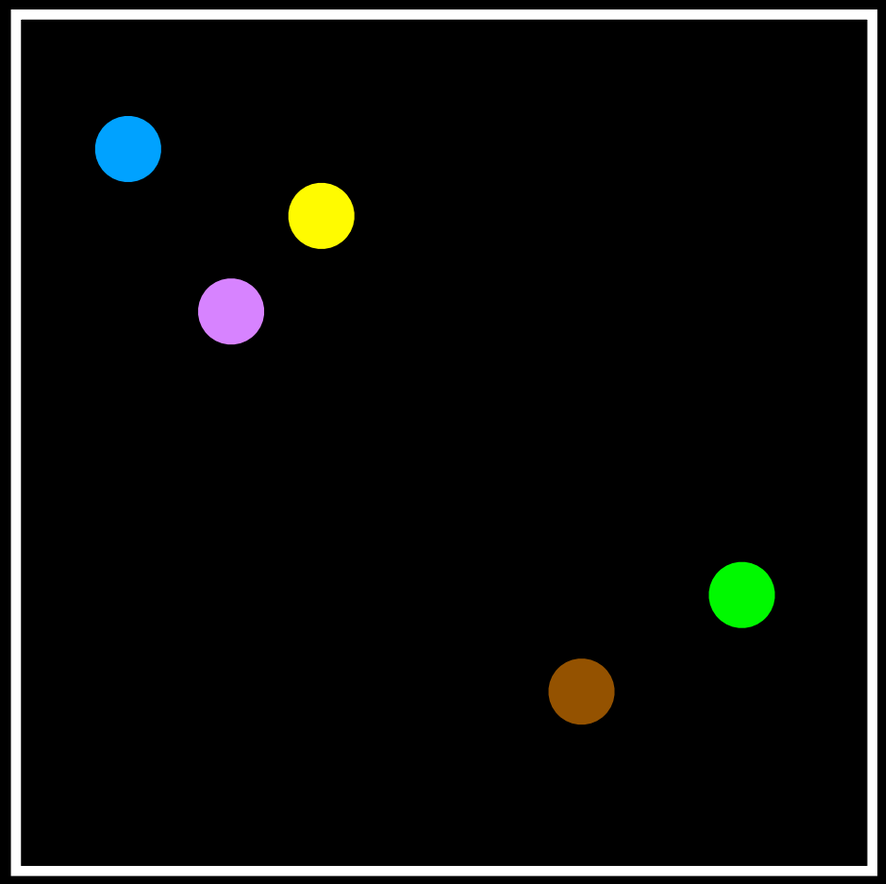
Hierarchical clustering

- Number of clusters is not known ahead of time
- Two kinds of hierarchical clustering:
  - ➡ bottom-up
  - ➡ top-down

Hierarchical clustering

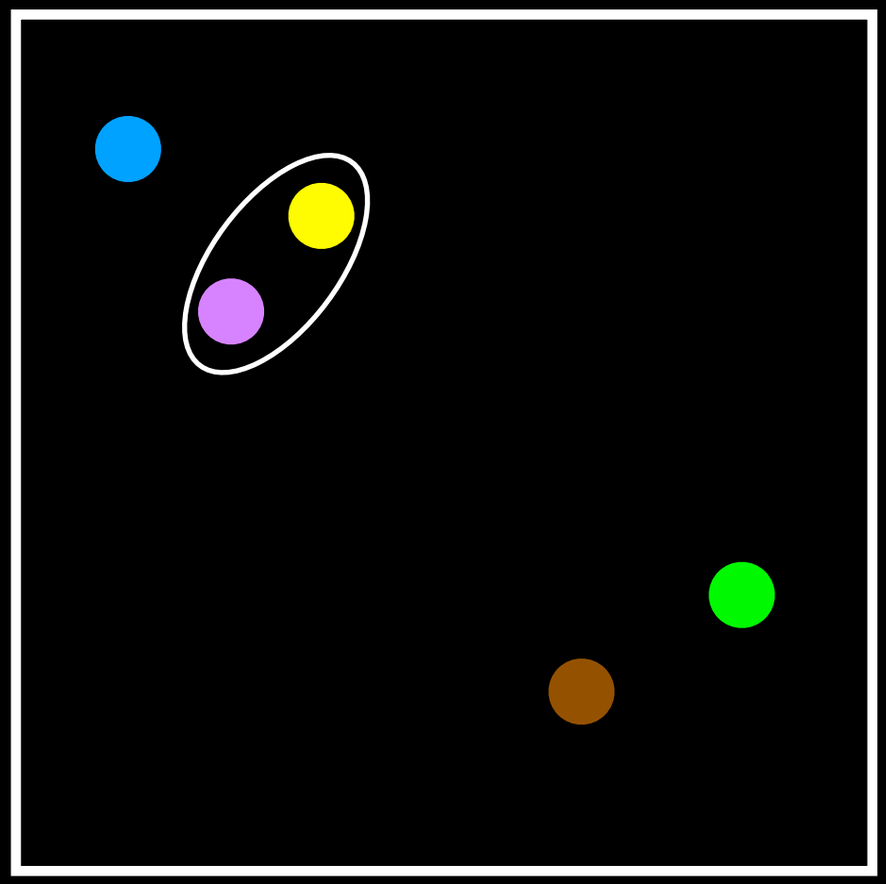
Simple example:

5 clusters: Each point starts as it's own "cluster"!



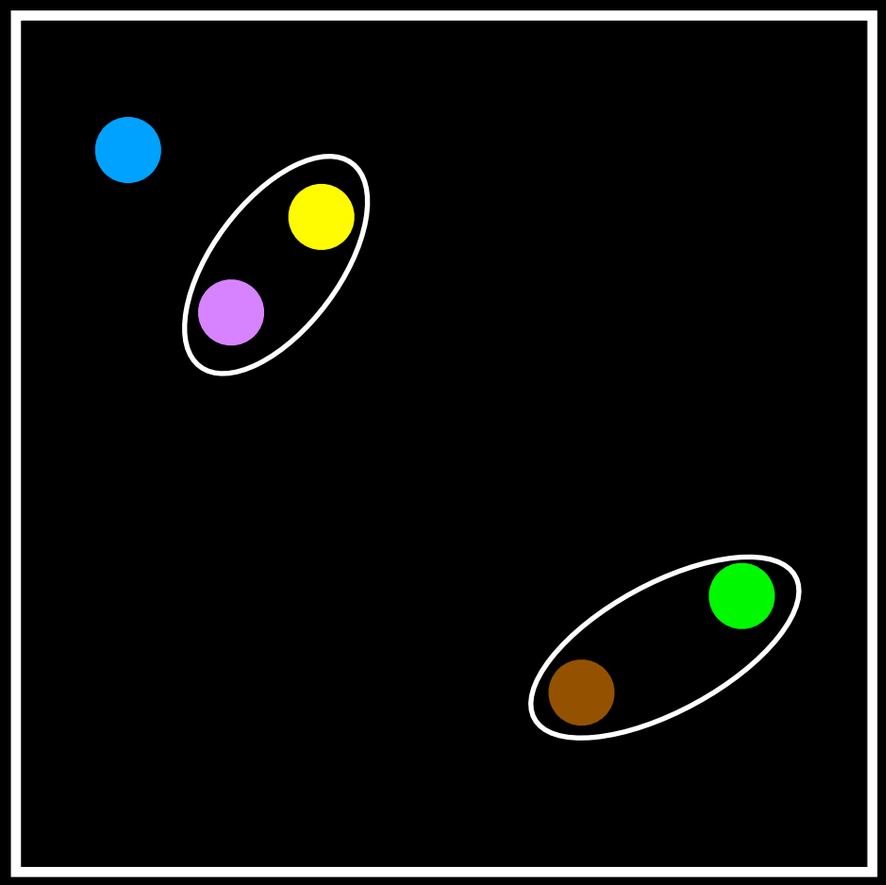
Hierarchical clustering

4 clusters



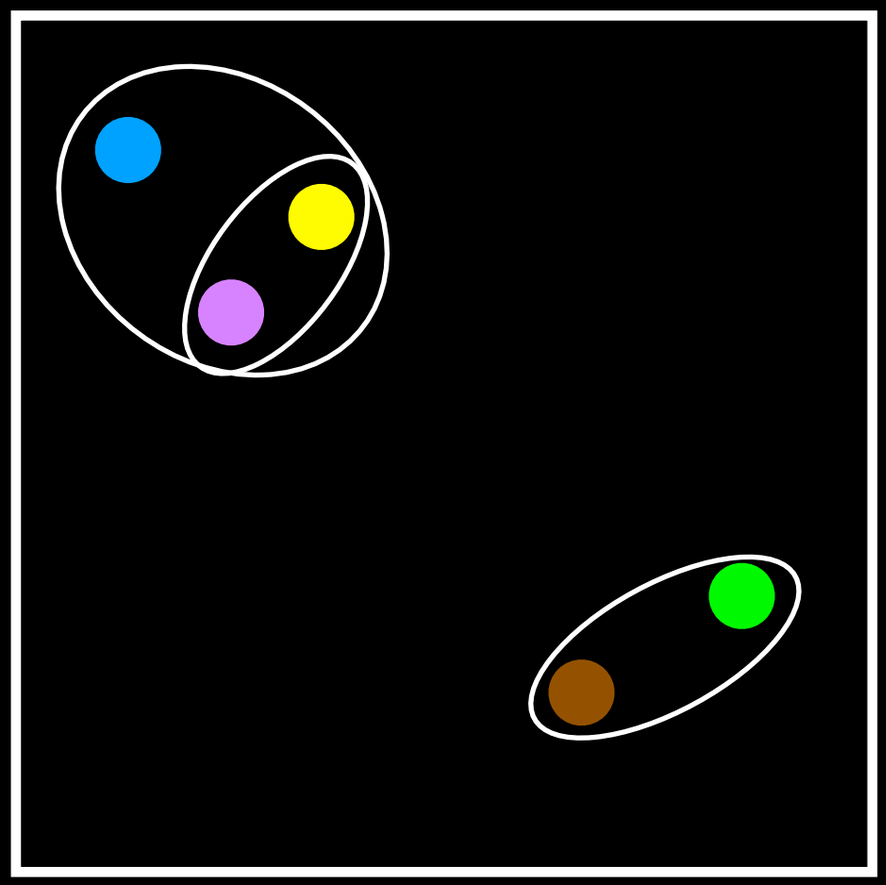
Hierarchical clustering

3 clusters



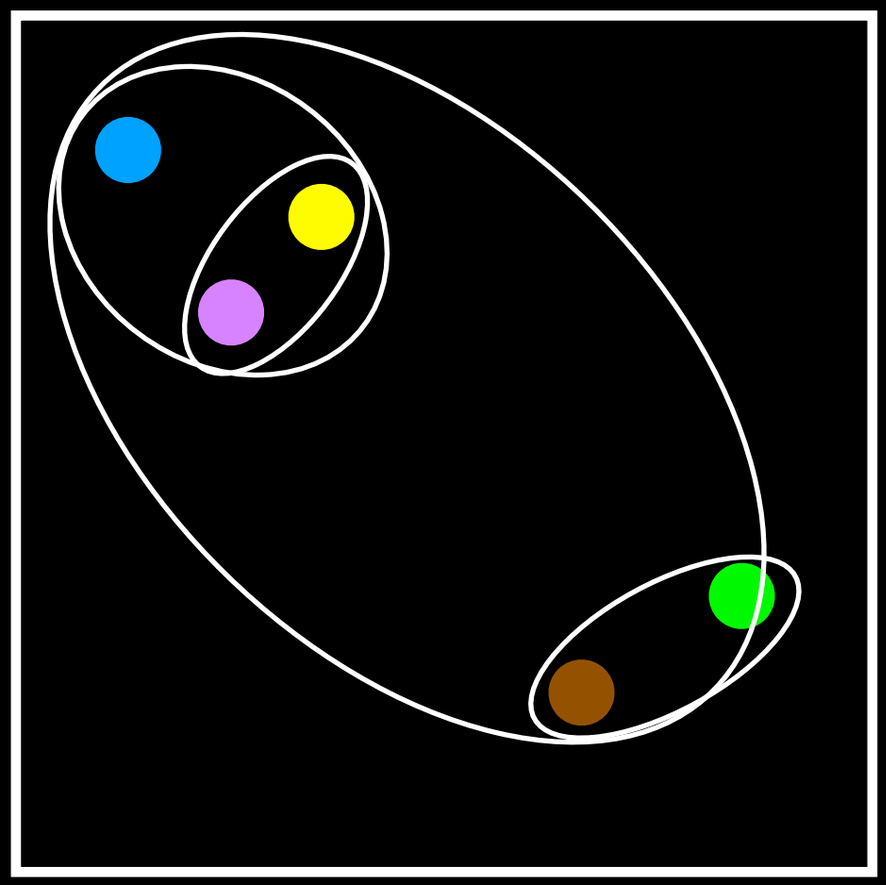
Hierarchical clustering

2 clusters



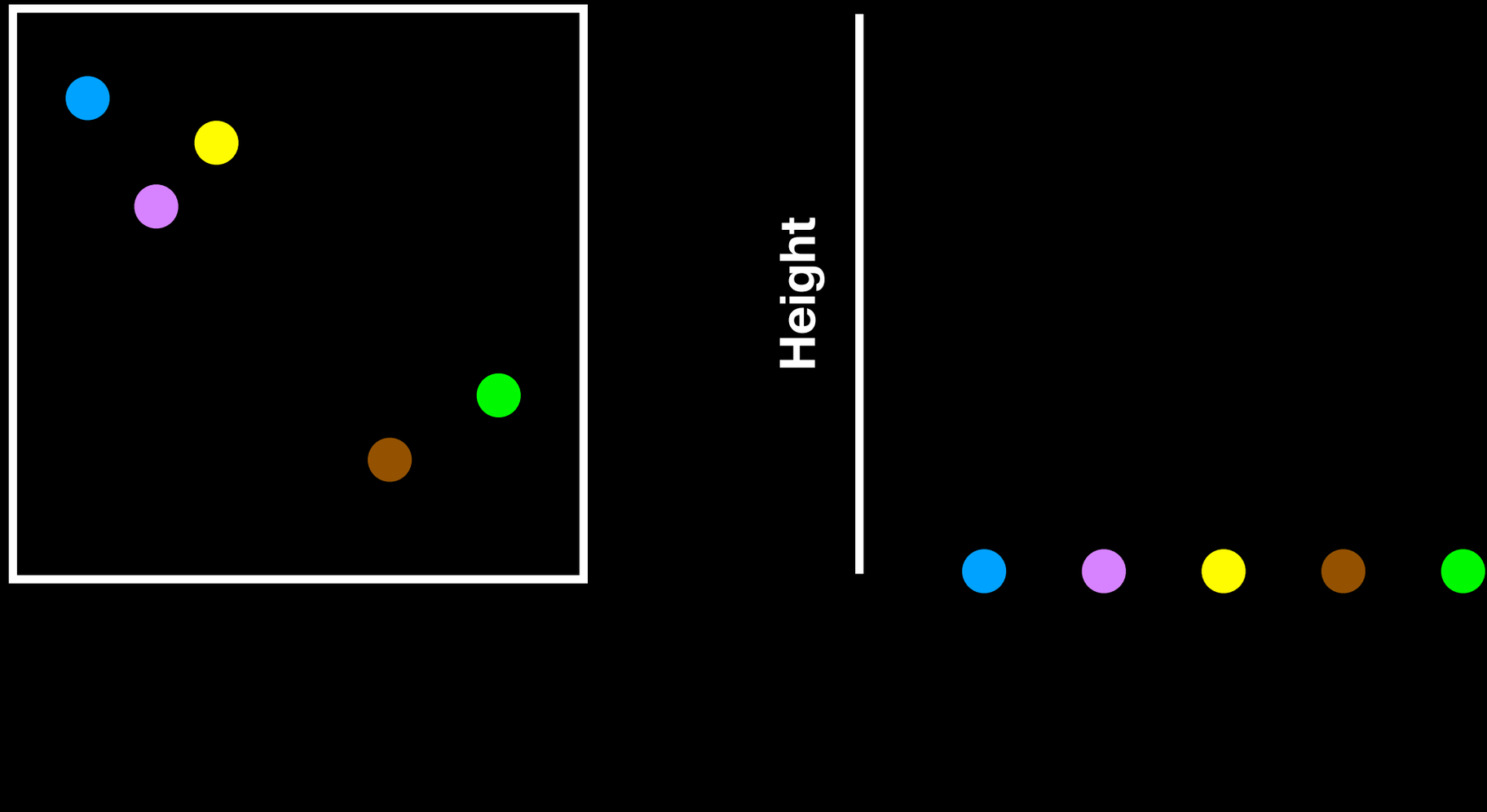
Hierarchical clustering

End: 1 cluster



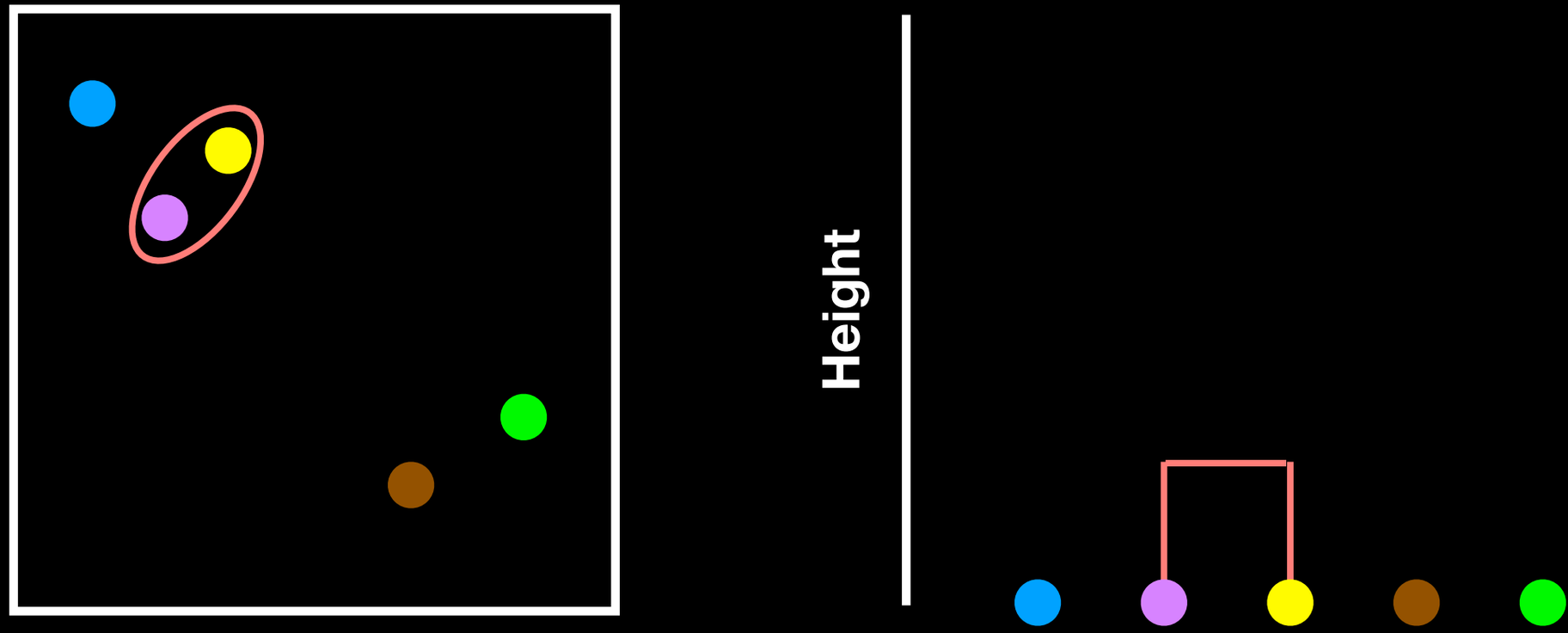
Dendrogram

- Tree shaped structure used to interpret hierarchical clustering models



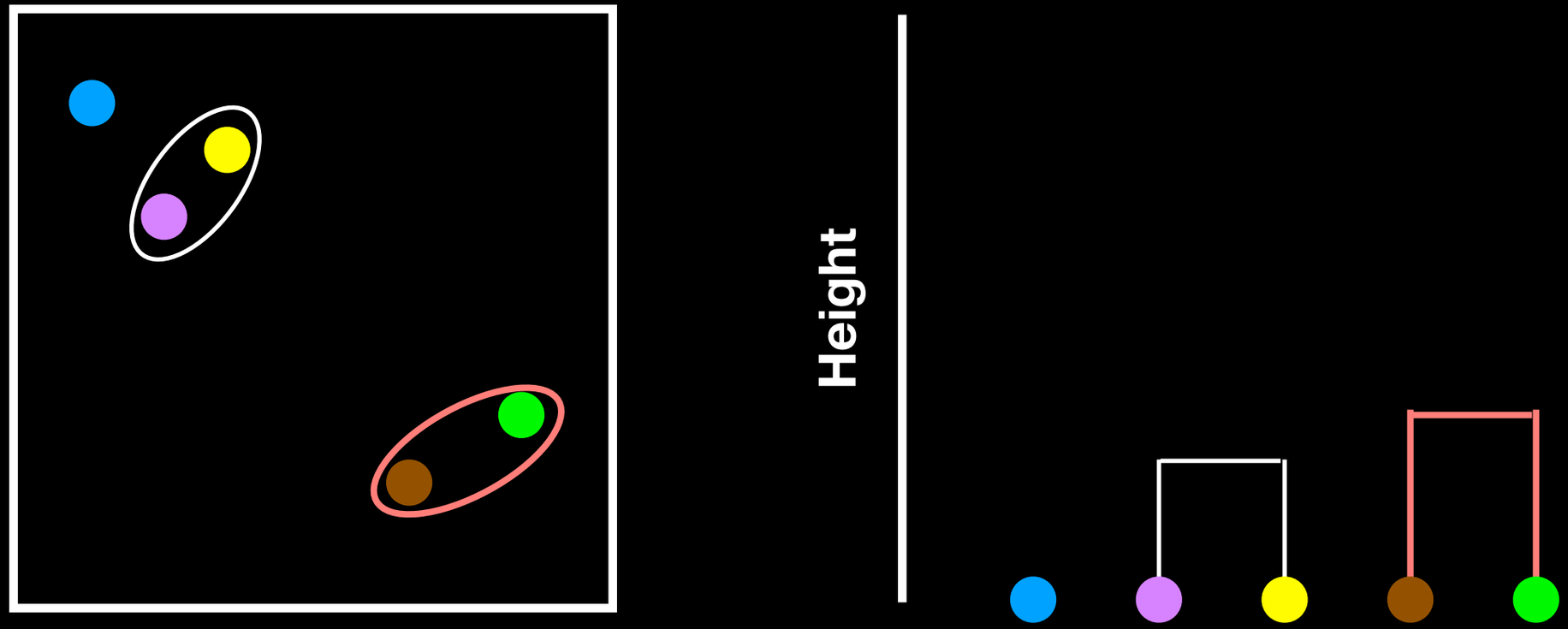
Dendrogram

- Tree shaped structure used to interpret hierarchical clustering models



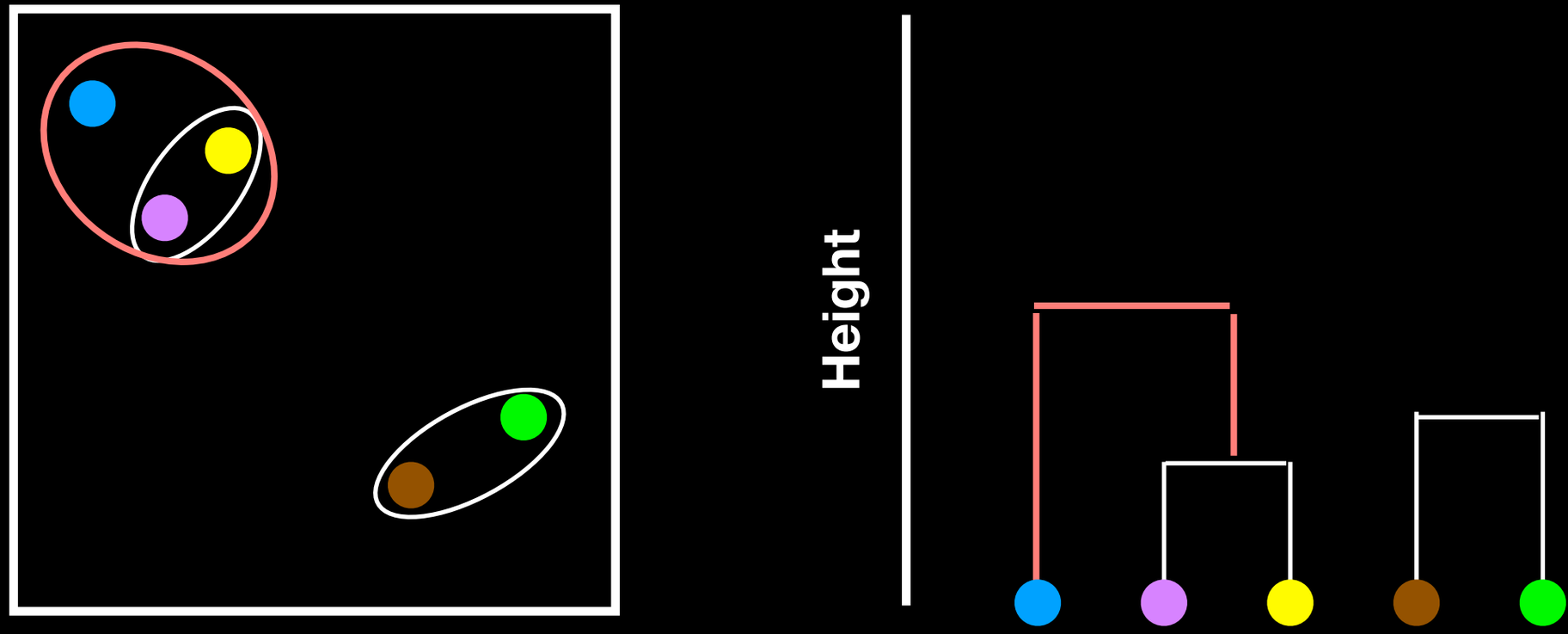
Dendrogram

- Tree shaped structure used to interpret hierarchical clustering models



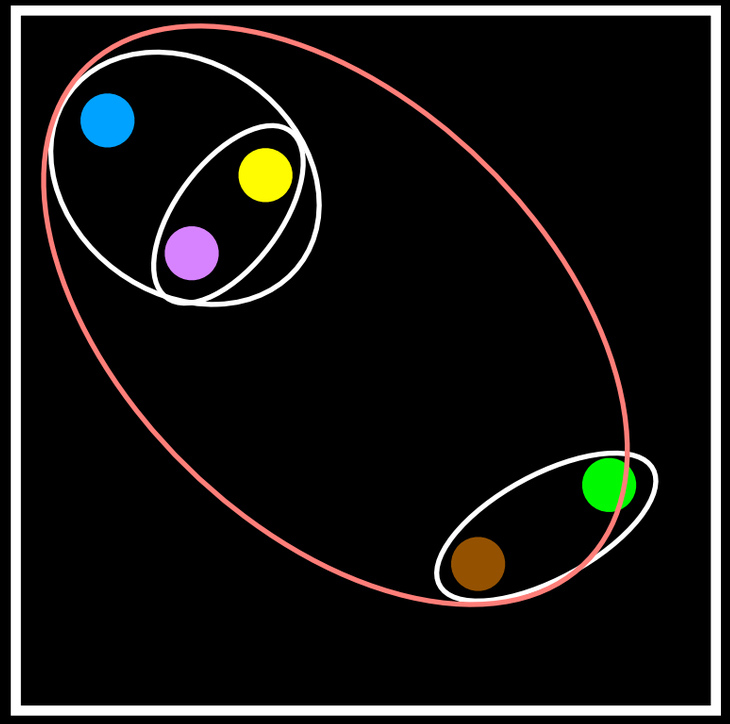
Dendrogram

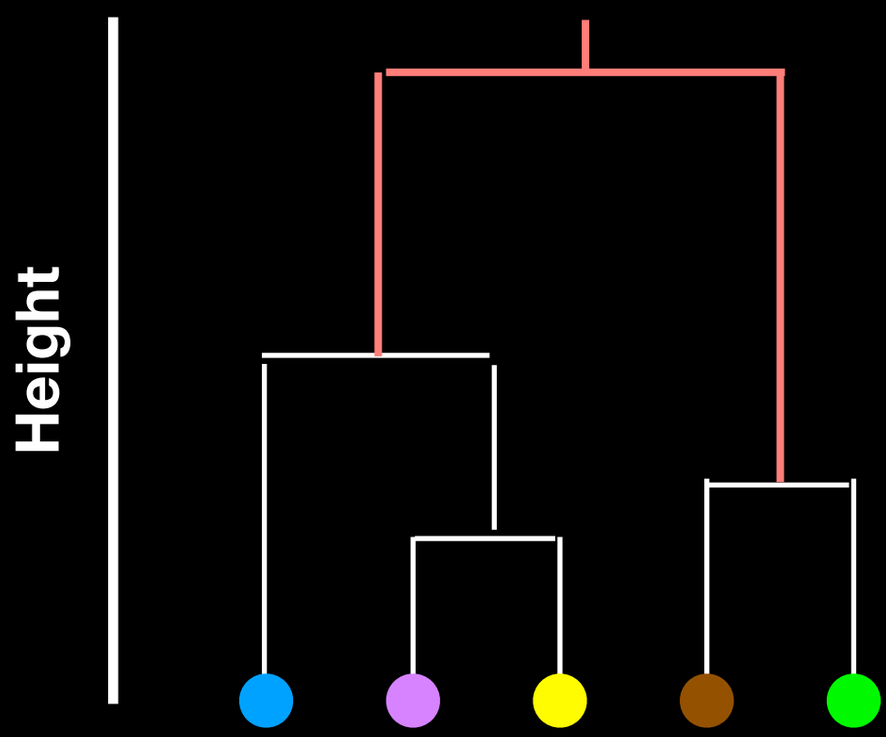
- Tree shaped structure used to interpret hierarchical clustering models



Dendrogram

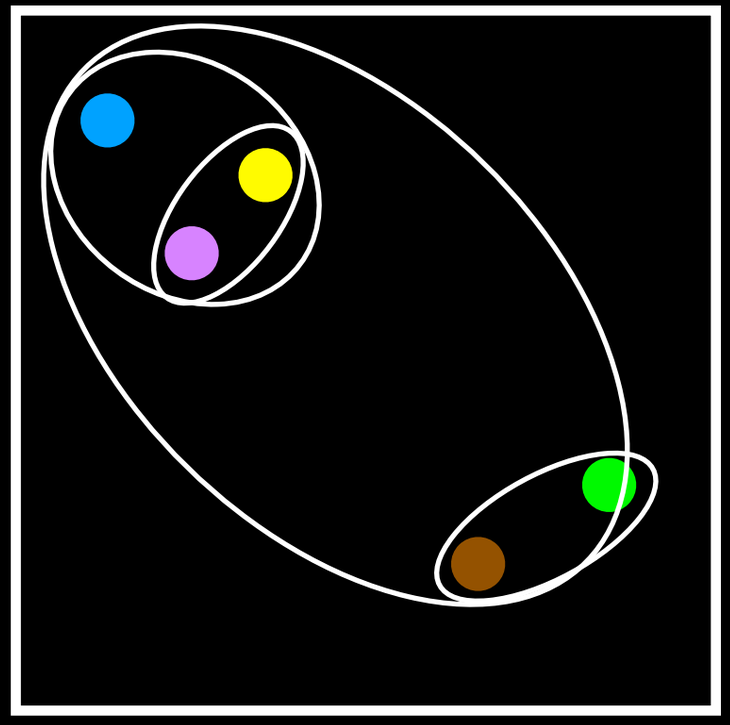
- Tree shaped structure used to interpret hierarchical clustering models

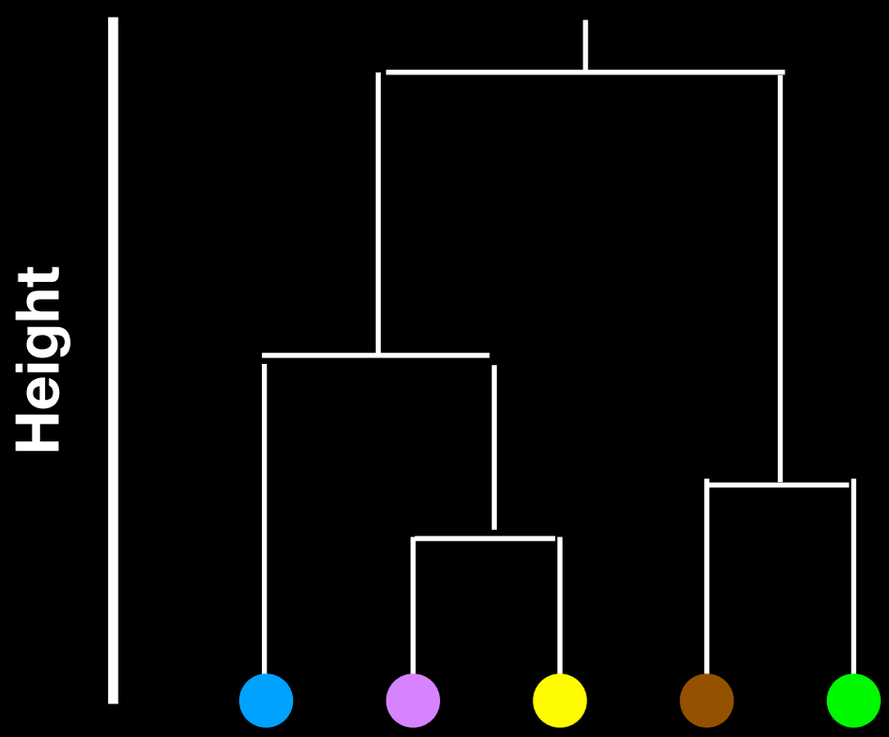




Dendrogram

- Tree shaped structure used to interpret hierarchical clustering models





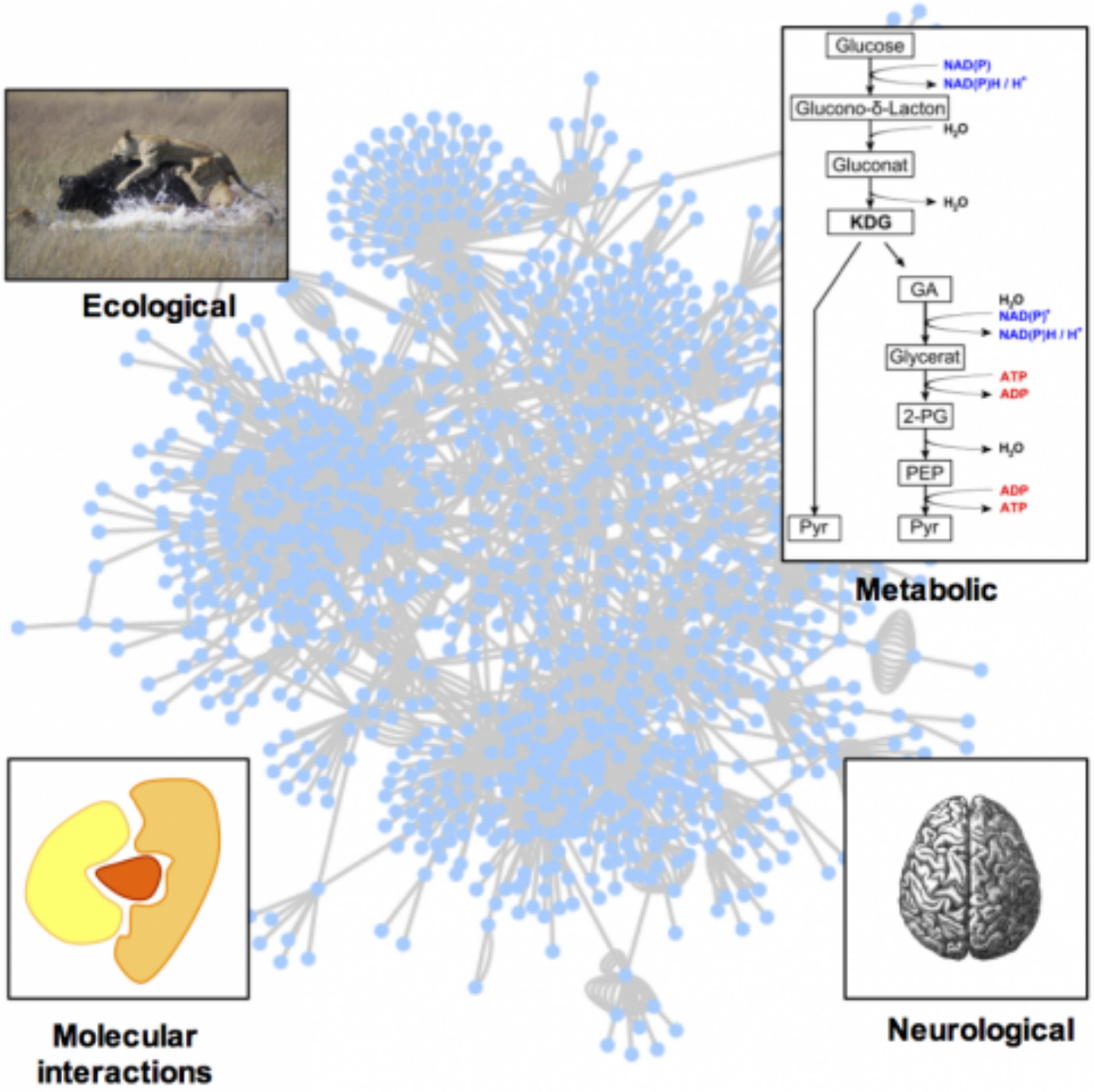
- Introduction to machine learning
  - Unsupervised, supervised and reinforcement learning

- Clustering
  - K-means clustering
  - Hierarchical clustering
  - Heatmap representations

- Dimensionality reduction, visualization and ‘structure’ analysis
  - Principal Component Analysis (PCA)

- Network analysis

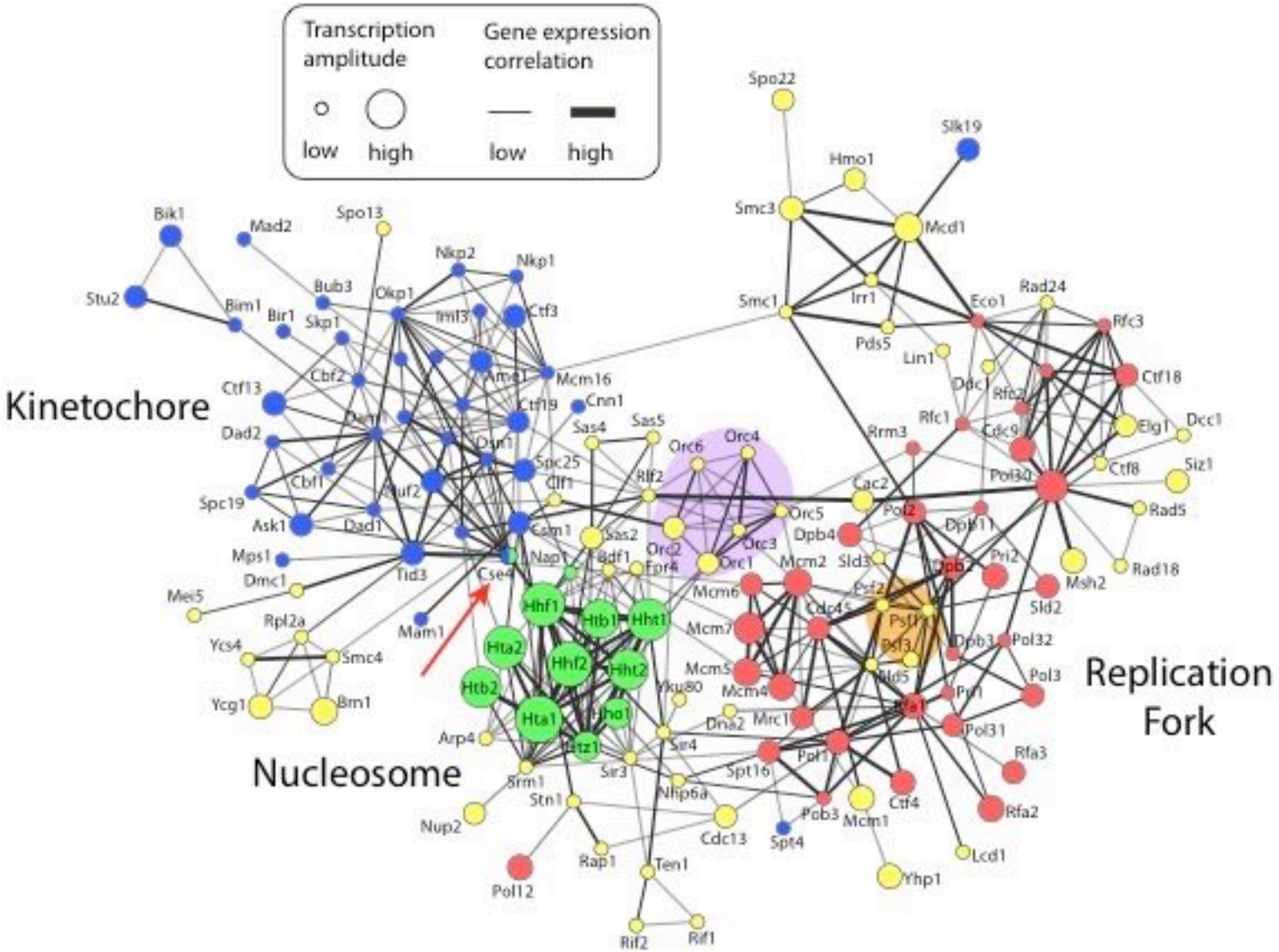
Networks can be used to model many types of biological data



Biological Networks

- **Represent biological interactions**
  - ➡ Physical, regulatory, genetic, functional, etc.
- **Useful for discovering relationships in big data**
  - ➡ Better than tables in Excel
- **Visualize multiple heterogenous data types together**
  - ➡ See interesting patterns
- **Network analysis**
  - ➡ Well established quantitive metrics from graph theory

[Yeast, cell-cycle PPIs]



TODAYS MENU:

- Network introduction
- **Network visualization**
- **Network analysis**
- **Hands-on:**

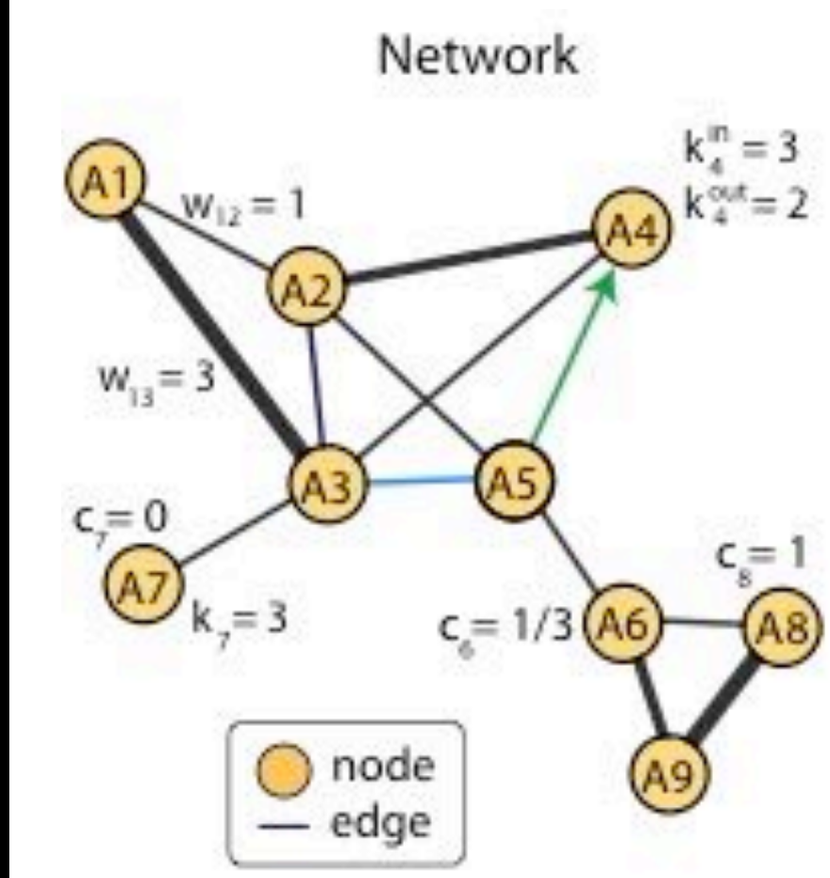
  Cytoscape and R (igraph) software tools for network visualization and analysis

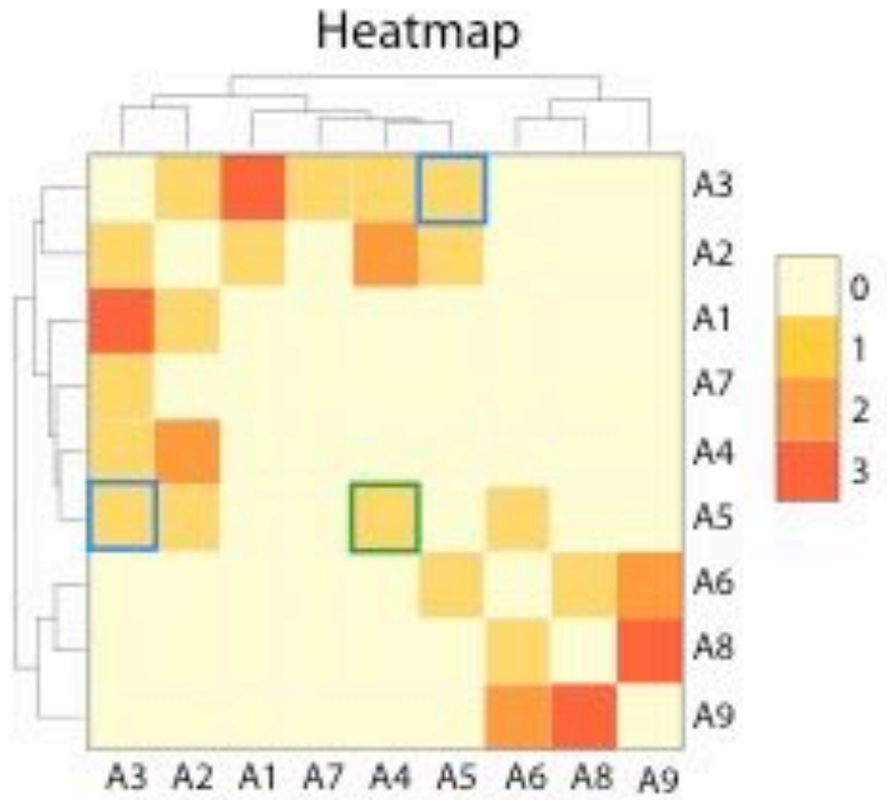
Network Visualization Outline

- Network representations
- Automatic network layout
- Visual features
- Visually interpreting a network

Network representations

Relationships	Optional weight
A1 ↔ A2	1
A1 ↔ A3	3
A2 ↔ A3	1
A2 ↔ A4	2
A2 ↔ A5	1
A3 ↔ A4	1
A3 ↔ A5	1
A3 ↔ A7	1
A5 → A4	1
A5 ↔ A6	1
A6 ↔ A8	1
A6 ↔ A9	2
A8 ↔ A9	3





1 **List of relationships**

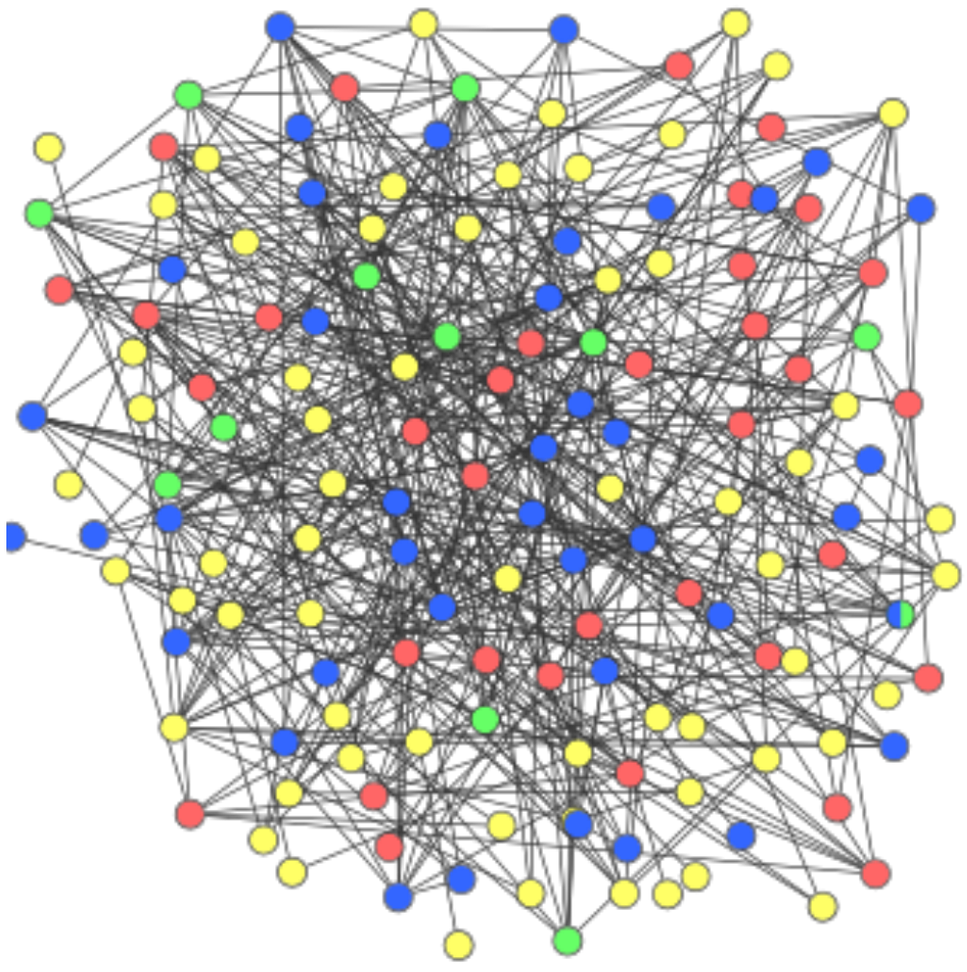
2 **Network view**

3 **Adjacency matrix view**

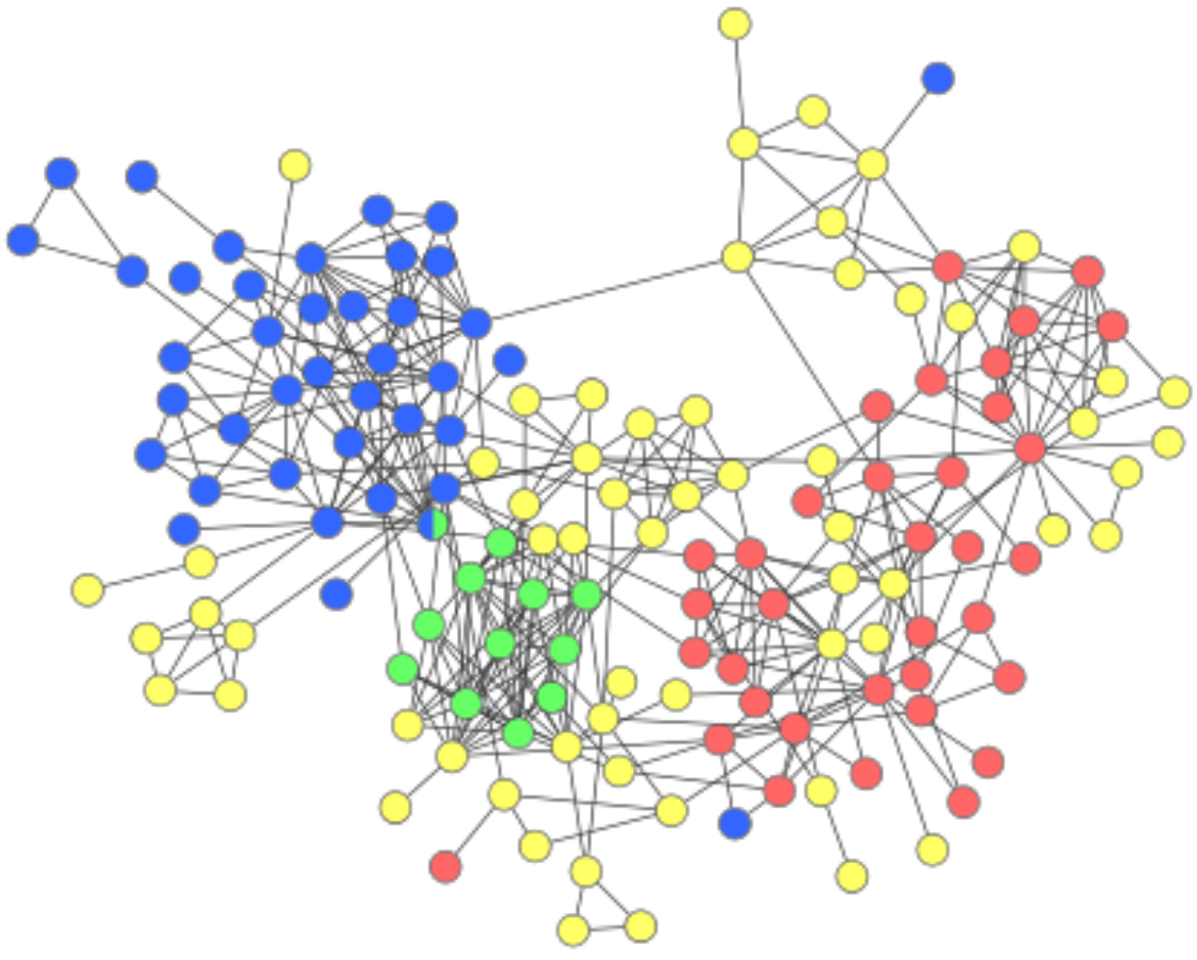
Network view is most useful when network is sparse!

Automatic network layout

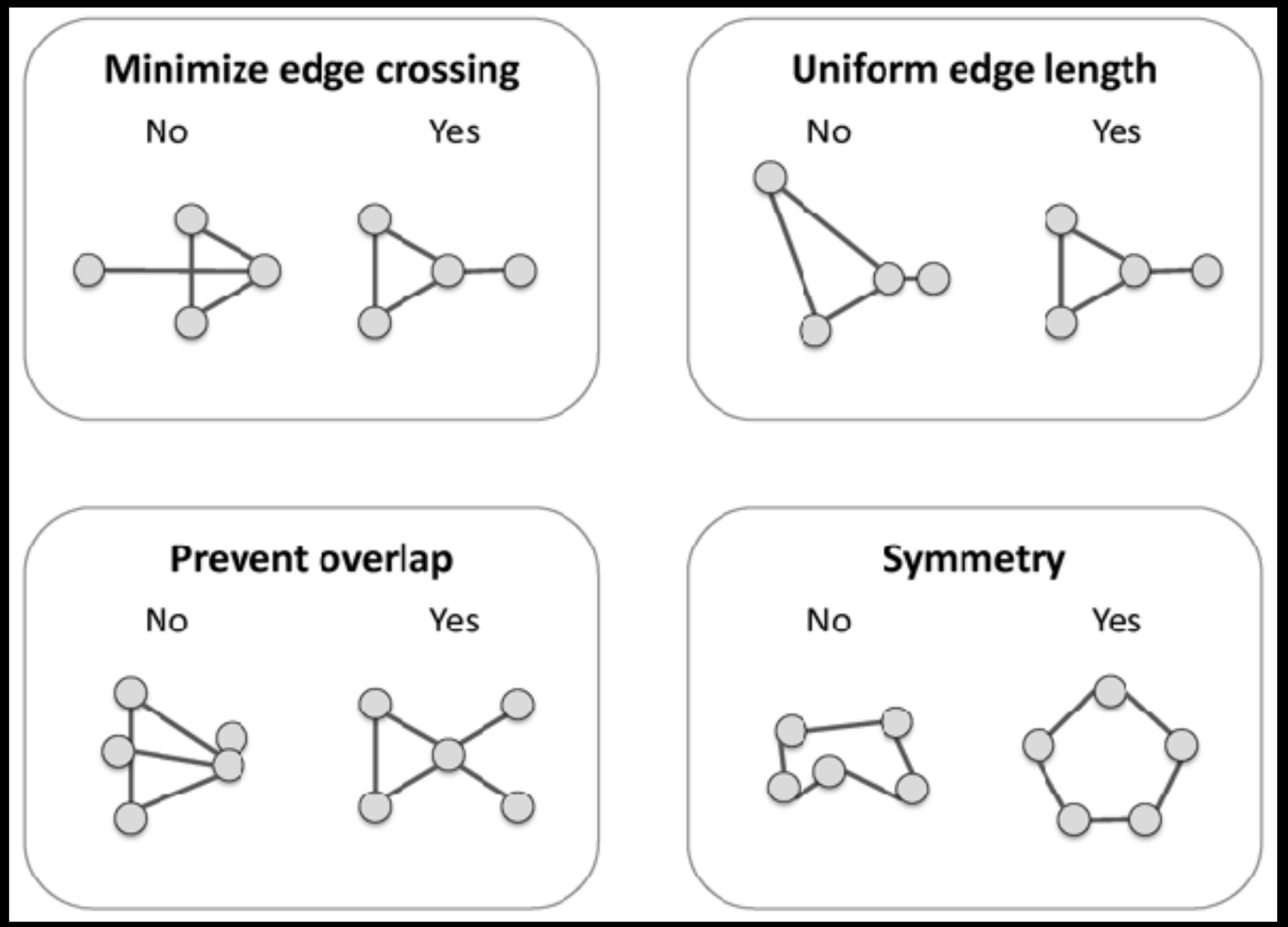
Before layout



After layout



- Modern **graph layouts** are optimized for speed and aesthetics. In particular, they seek to minimize overlaps and edge crossing, and ensure similar edge length across the graph.



Force-directed layout:

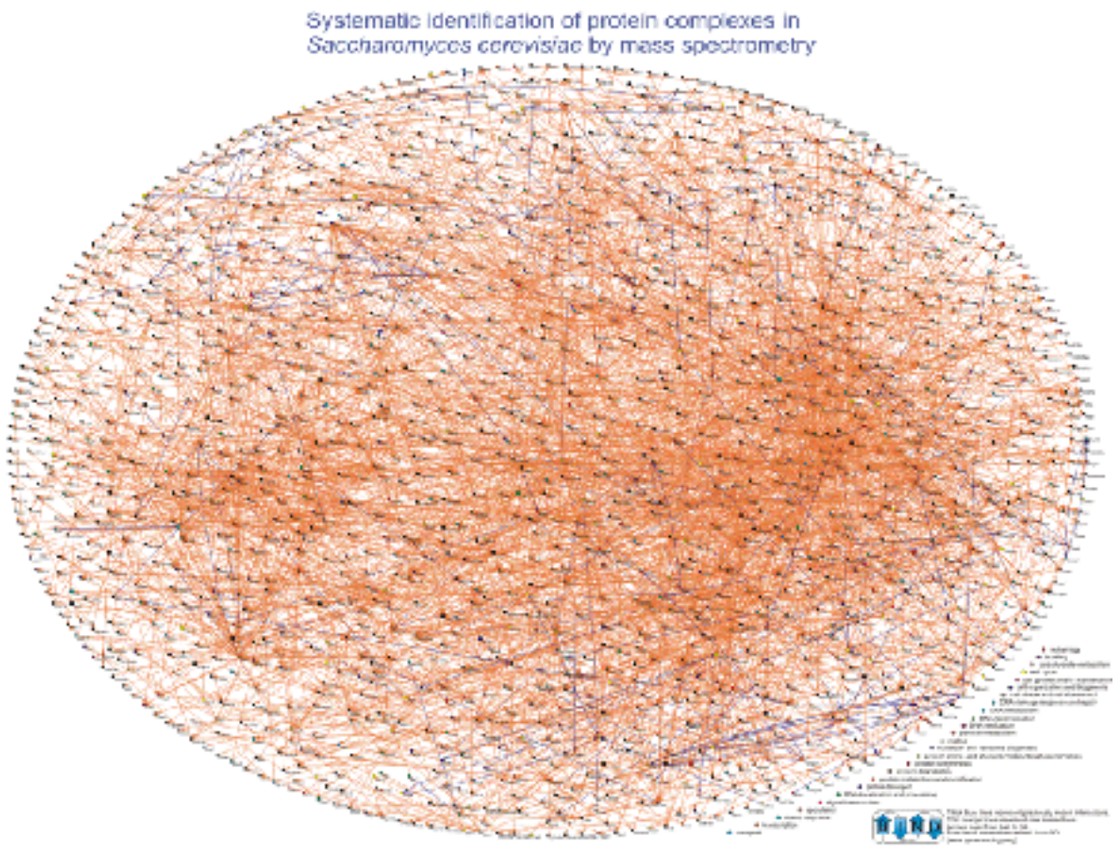
Nodes repel and edges pull

- Good for up to 500 nodes
  - ➡ Bigger networks give hairballs
  - ➡ Reduce number of edges
  - ➡ Or just use a heatmap for dense networks

- Advice: try force directed first, or hierarchical for tree-like networks

- Tips for better looking networks
  - ➡ Manually adjust layout
  - ➡ Load network into a drawing program (e.g. Illustrator) and adjust labels

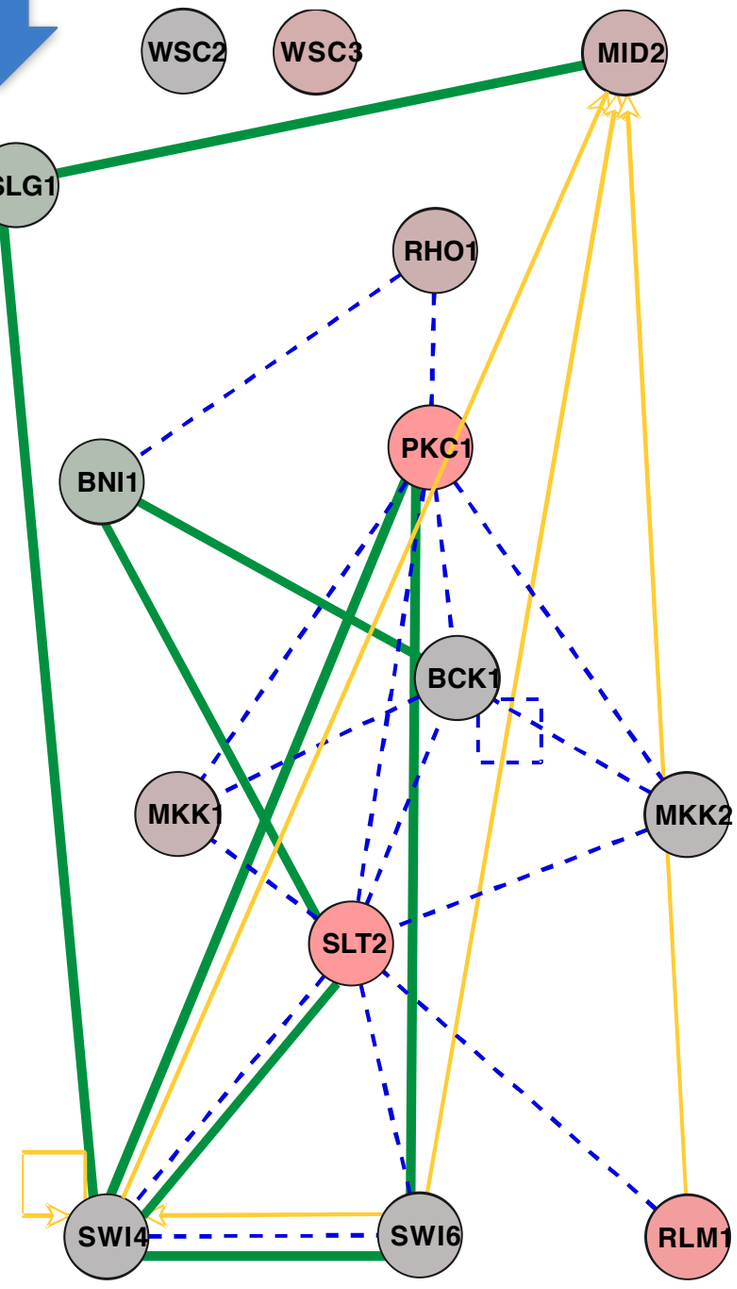
Dealing with ‘hairballs’: zoom or filter



Zoom

Focus

PKC Cell Wall Integrity



Synthetic Lethal

Transcription Factor Regulation

Protein-Protein Interaction

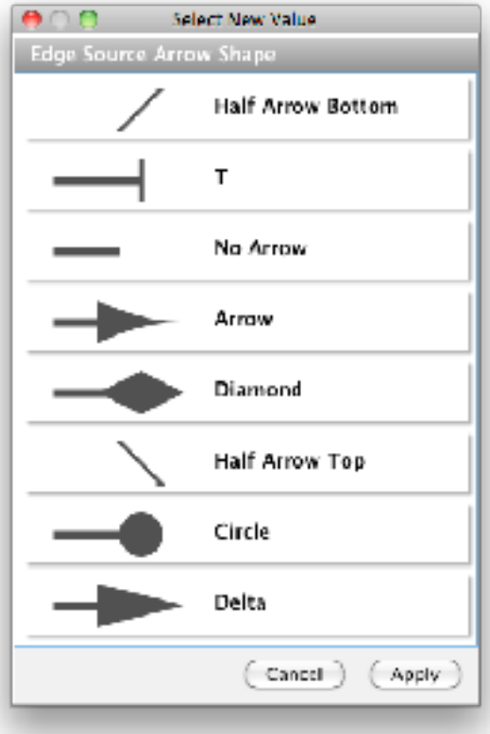
Up Regulated Gene Expression

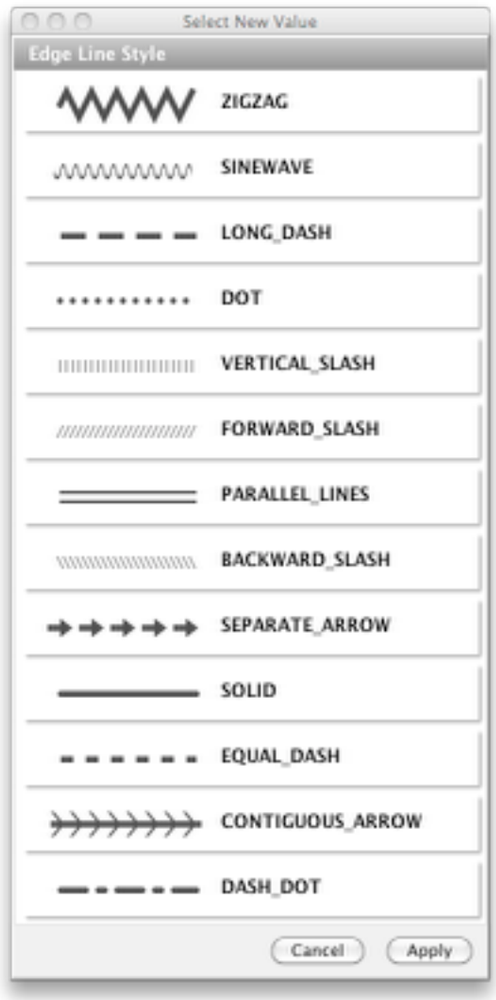
Down Regulated Gene Expression

Visual Features

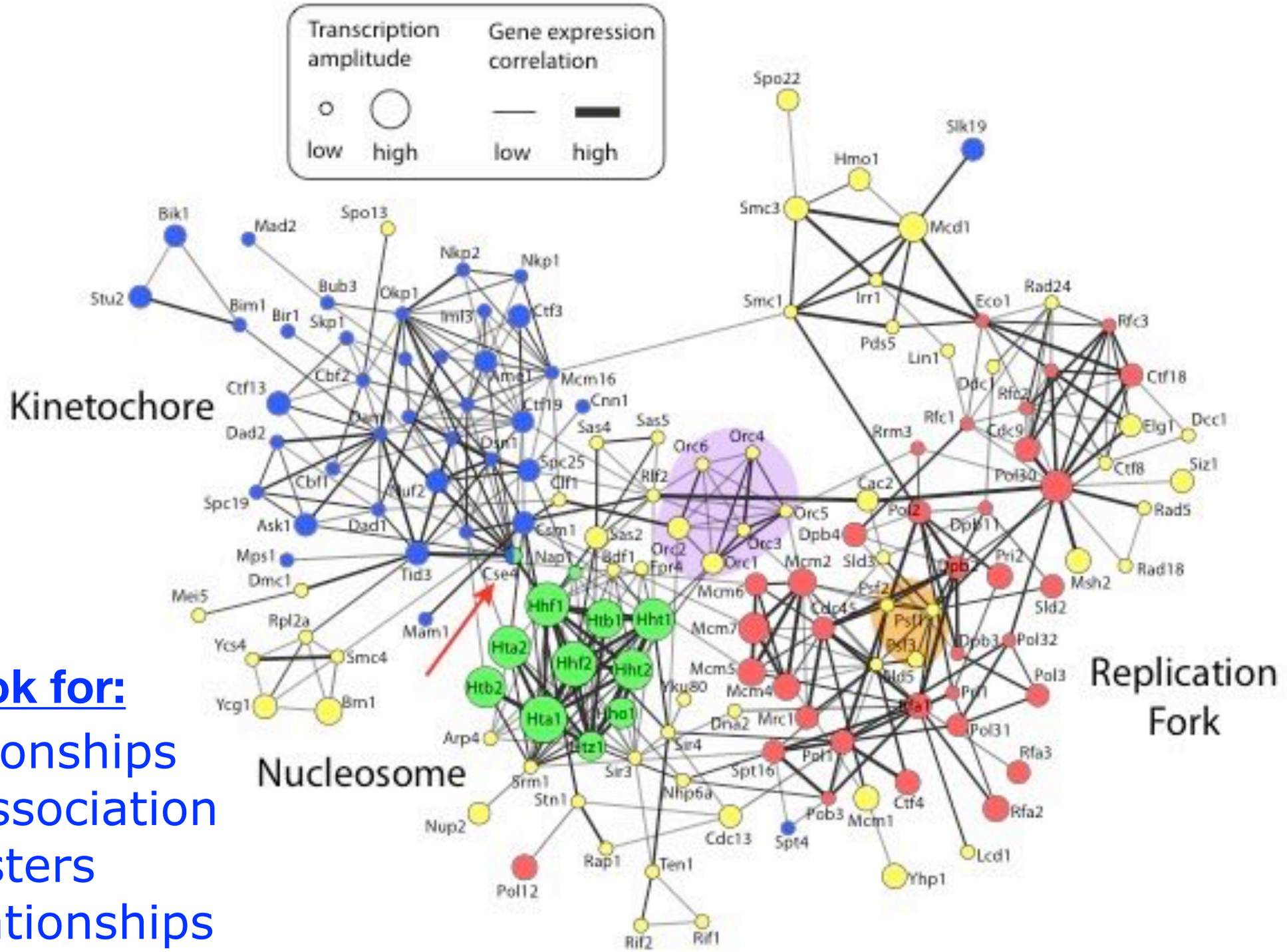
- Node and edge attributes
  - Text (string), integer, float, Boolean, list
  - E.g. represent gene, interaction attributes

- Visual attributes
  - Node, edge visual properties
  - Color, shape, size, borders, opacity...





Visually Interpreting a Network



What to look for:

Data relationships

Guilt-by-association

Dense clusters

Global relationships

What have we learned so far…

- Automatic layout is required to visualize networks
- Networks help you visualize interesting relationships in your data
- Avoid hairballs by focusing analysis
- Visual attributes enable multiple types of data to be shown at once – useful to see their relationships

TODAYS MENU:

- **Network introduction**
- **Network visualization**
- **Network analysis**
- **Hands-on:**
  - Cytoscape and R (igraph) software tools for network visualization and analysis

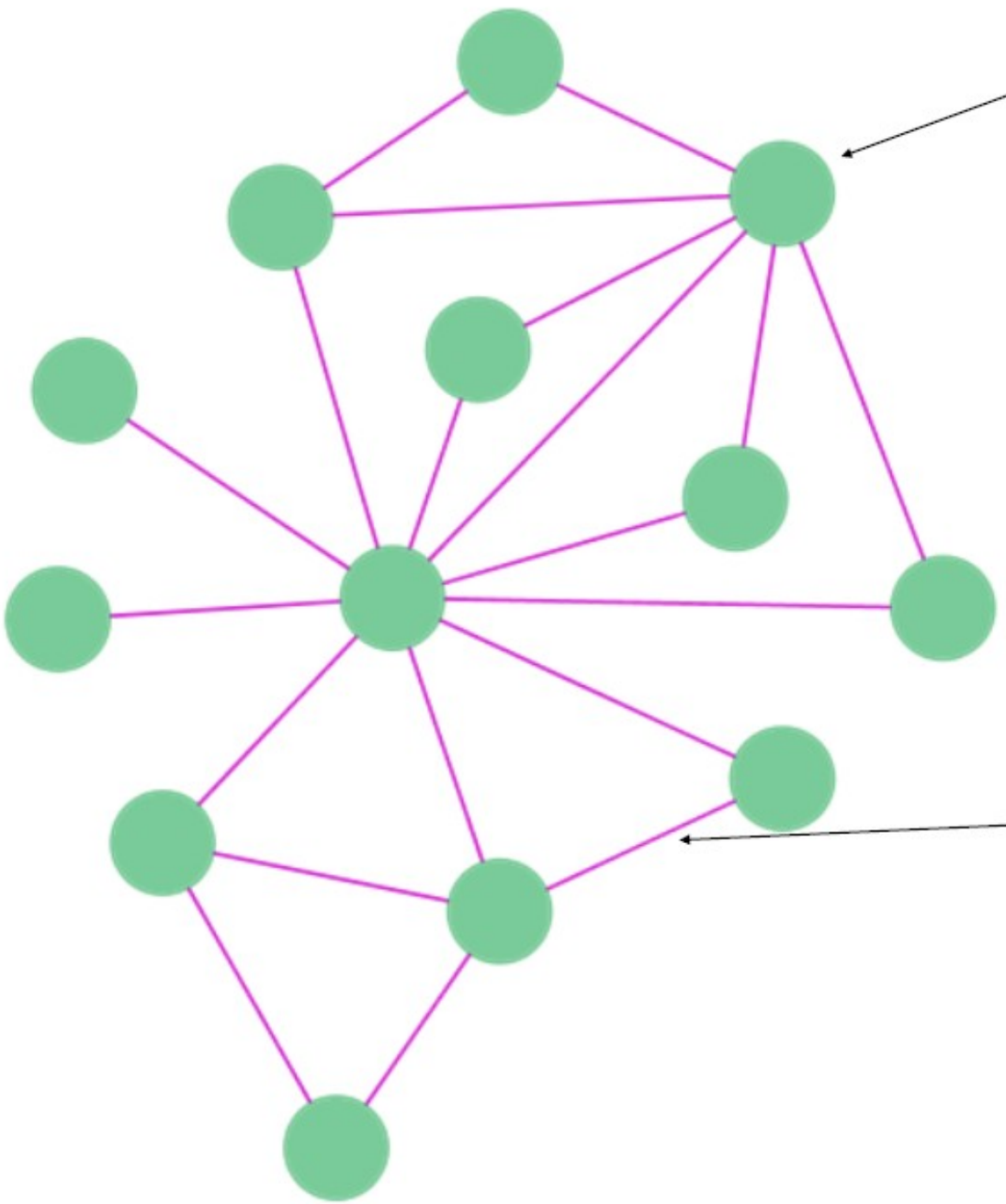
Introduction to graph theory

- Biological network analysis historically originated from the tools and concepts of social network analysis and the application of graph theory to the social sciences.

- Wikipedia defines graph theory as:
  - ➡ “[…] the study of graphs used to model pairwise relations between objects. A graph in this context is made up of **vertices** connected by **edges**”.

- In practical terms, it is the set of concepts and methods that can be used to visualize and analyze networks

Network or graph

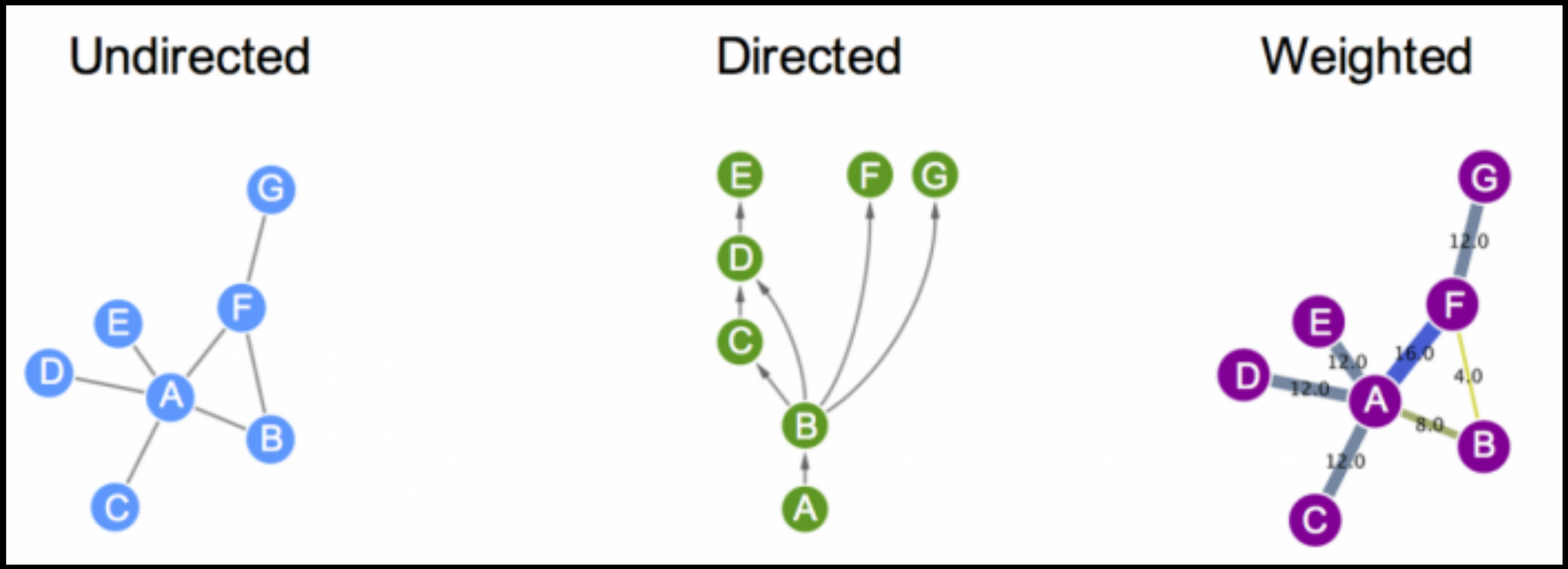


Node or vertex: protein, gene, drug, disease

Edge or link: relation between nodes

- Binary or continuous
- Directed or indirected
- Edge types

Types of network edges



Connection, without a given ‘flow’ implied

(e.g. protein A binds protein B)

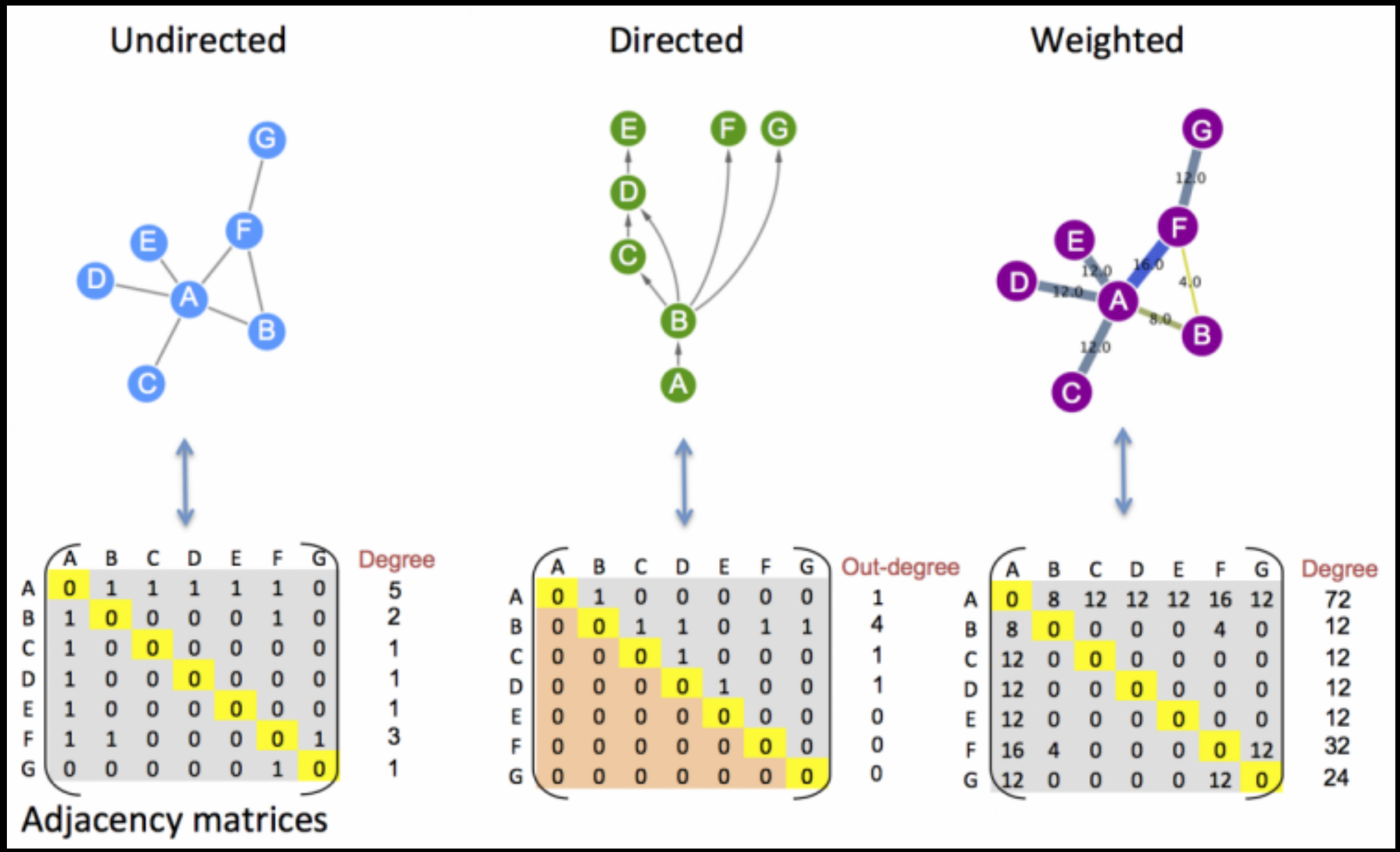
There is directional flow/signal implied

(e.g. metabolic or gene networks)

Edges can also have weight

(i.e. a ‘strength’ of interaction).

- Every network can be expressed mathematically in the form of an adjacency matrix



	A	B	C	D	E	F	G	Degree
A	0	1	1	1	1	1	0	5
B	1	0	0	0	0	1	0	2
C	1	0	0	0	0	0	0	1
D	1	0	0	0	0	0	0	1
E	1	0	0	0	0	0	0	1
F	1	1	0	0	0	0	1	3
G	0	0	0	0	0	1	0	1

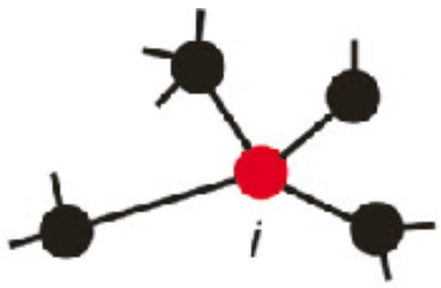
	A	B	C	D	E	F	G	Out-degree
A	0	1	0	0	0	0	0	1
B	0	0	1	1	0	1	1	4
C	0	0	0	1	0	0	0	1
D	0	0	0	0	1	0	0	1
E	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0

	A	B	C	D	E	F	G	Degree
A	0	8	12	12	12	16	12	72
B	8	0	0	0	0	4	0	12
C	12	0	0	0	0	0	0	12
D	12	0	0	0	0	0	0	12
E	12	0	0	0	0	0	0	12
F	16	4	0	0	0	0	12	32
G	12	0	0	0	0	12	0	24

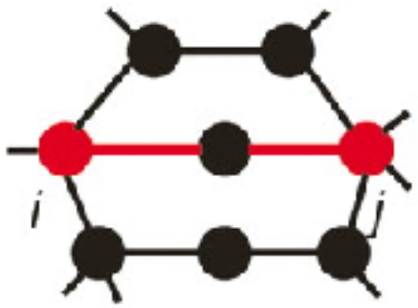
Network topology

- Topology is the way in which the nodes and edges are arranged within a network.

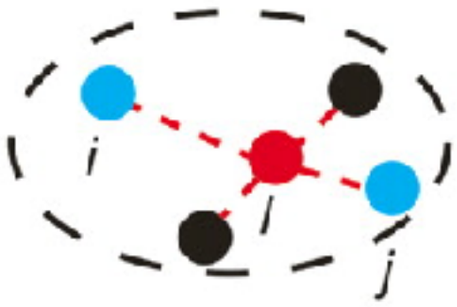
- The most used topological properties and concepts include:
  - **Degree** (i.e. how may node neighbors)
  - **Communities** (i.e. clusters of well connected nodes)
  - **Shortest Paths** (i.e. shortest distance between 2 nodes)
  - **Centralities** (i.e. how ‘central’ is a given node?)
  - **Betweenness** (a measure of centrality based on shortest paths)



Degree k_i = number of links connected to node i



Distance d_{ij} = shortest path length between node i and j

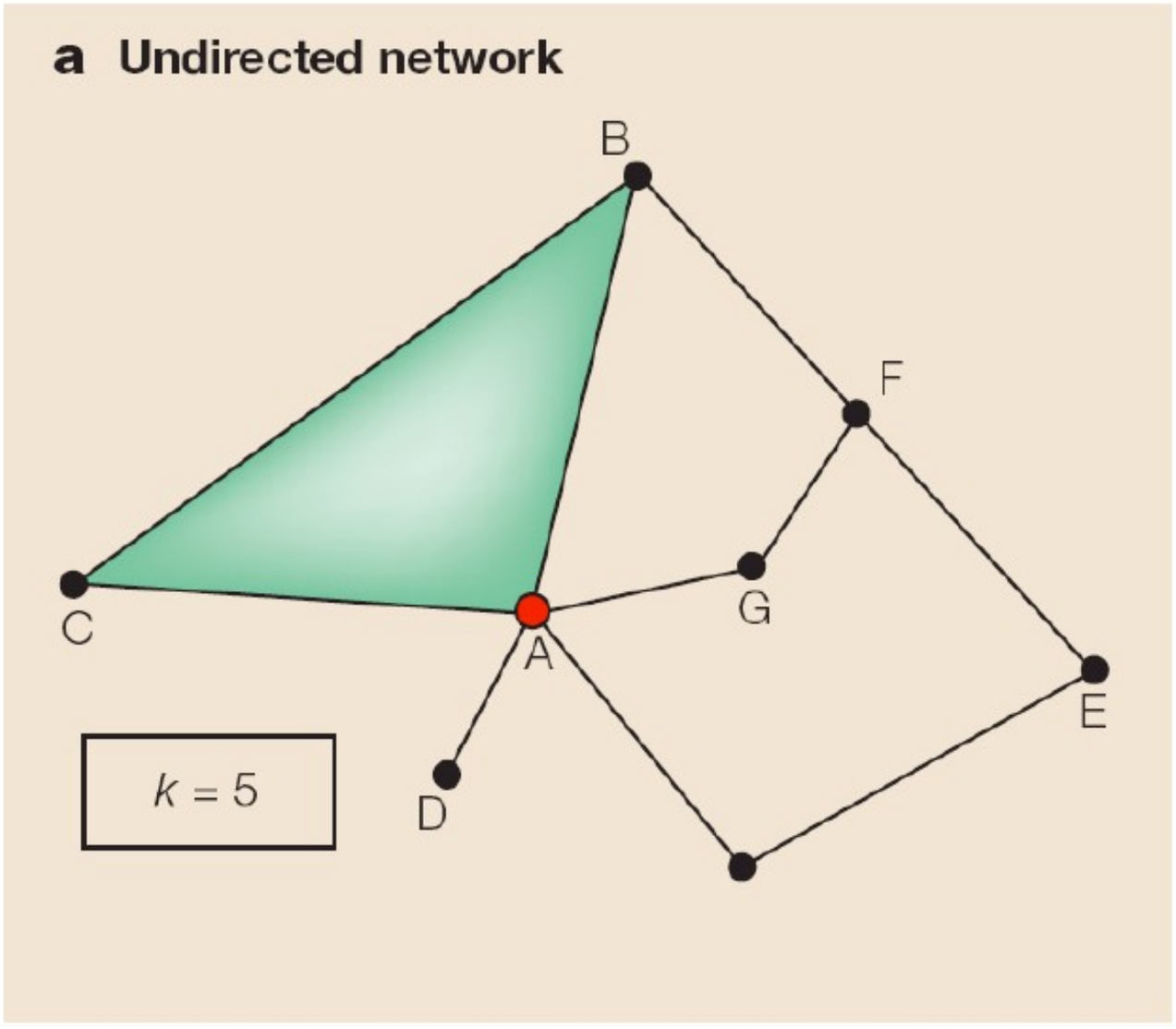


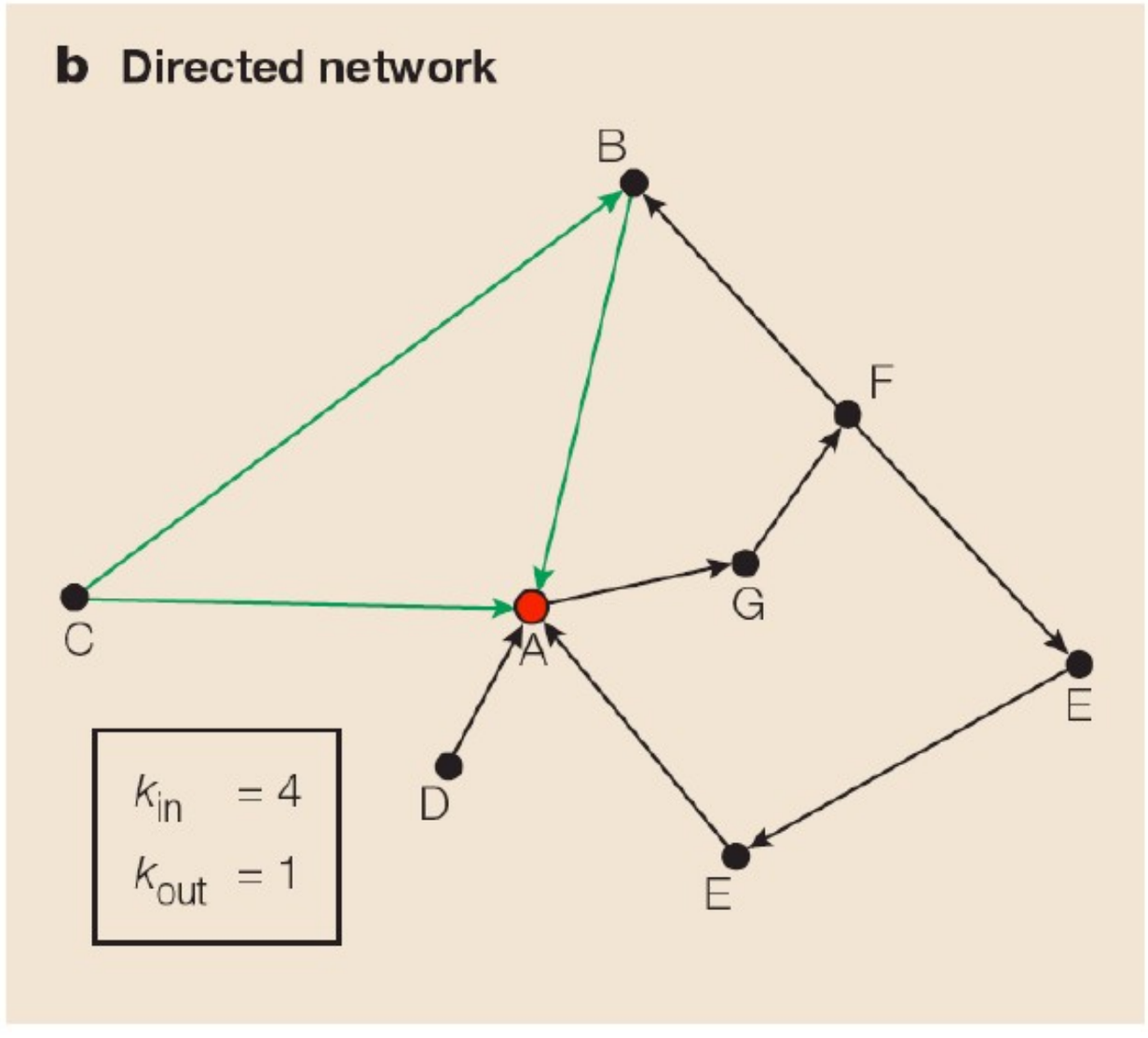
Betweenness $b_l = \sum_{ij} p_{ij}(l)/p_{ij}$

p_{ij} : number of shortest paths between i and j

$p_{ij}(l)$: number of shortest paths between i and j going through node l

Network Measures: Degree

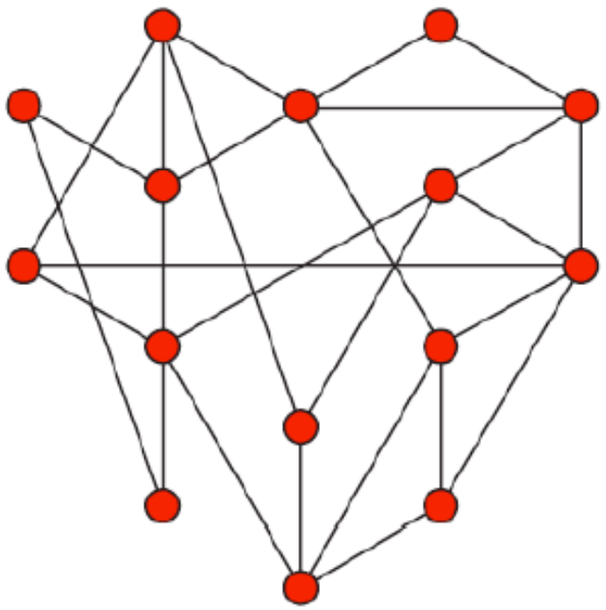




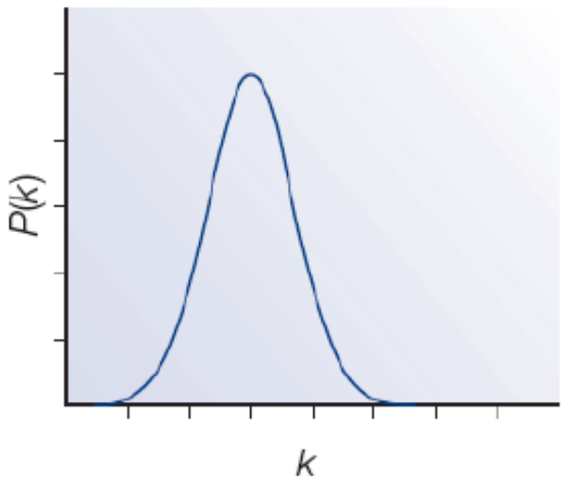
Degree Distribution

A Random network

Aa

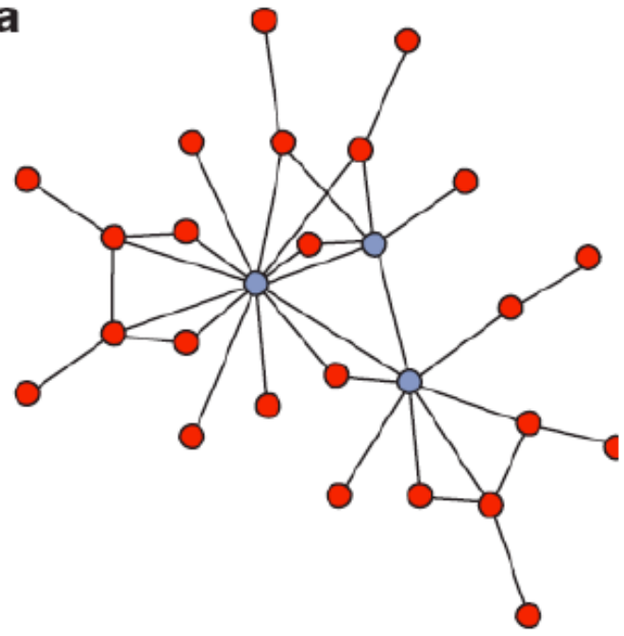


Ab

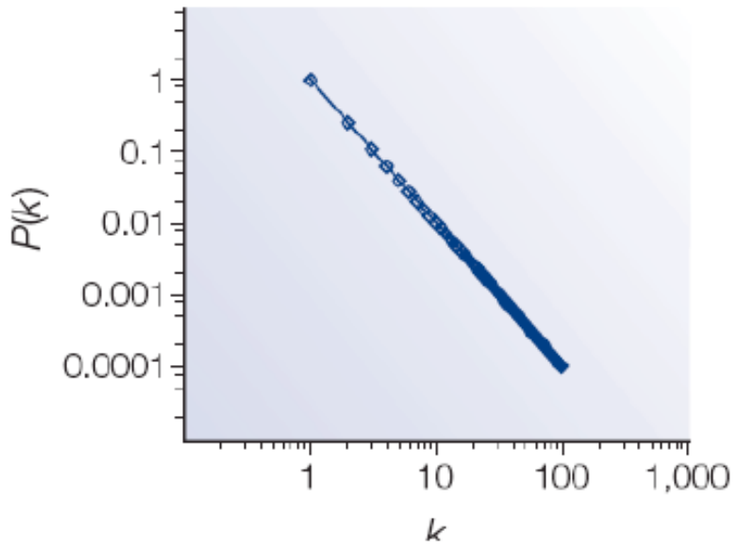


B Scale-free network

Ba



Bb



P(k) is probability of each degree k, i.e fraction of nodes having that degree.

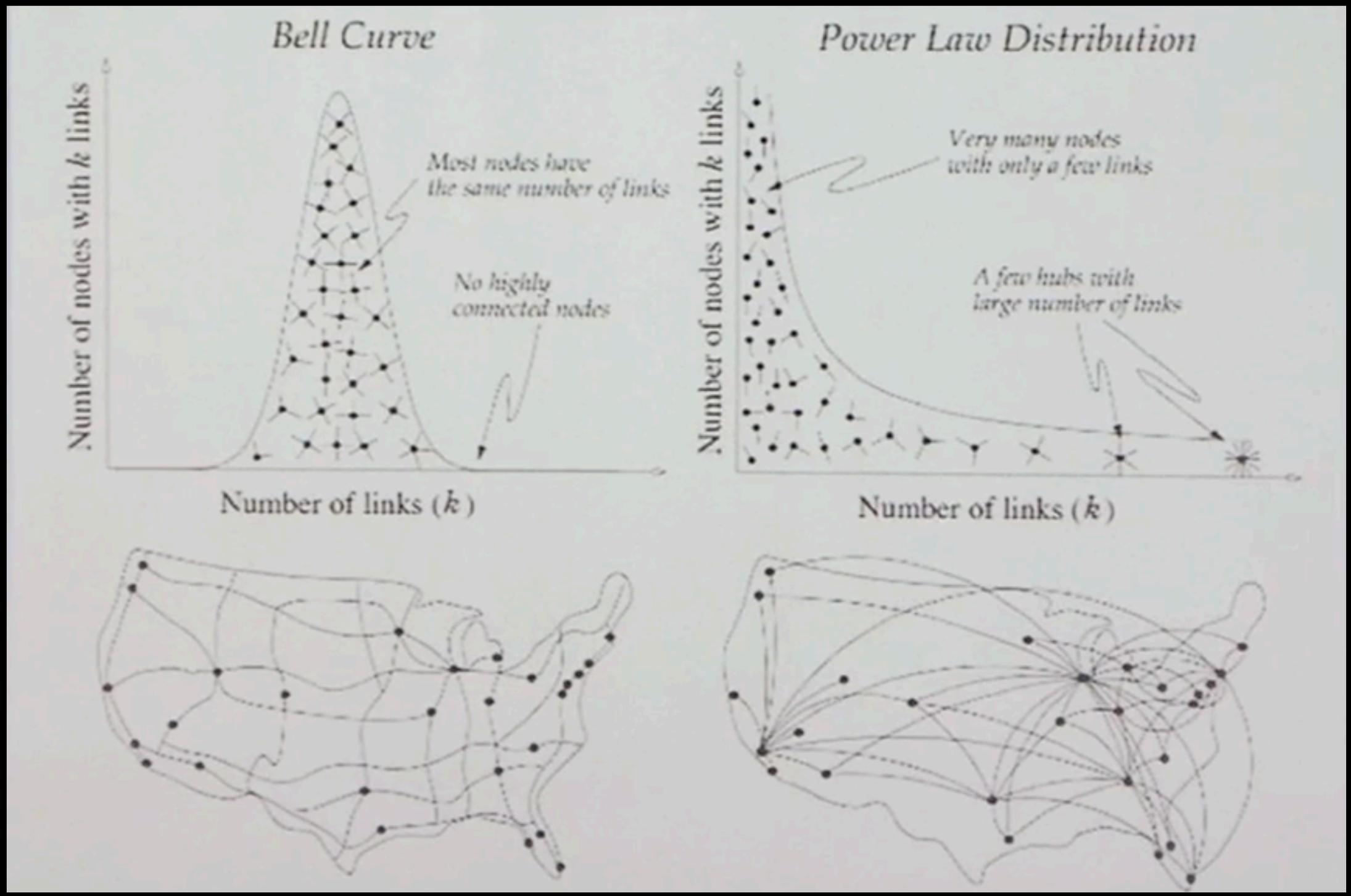
For random networks, P(k) is normally distributed.

For real networks the distribution is often a power-law:

$P(k) \sim k^{-\gamma}$

Such networks are said to be scale-free

Random graphs vs scale free



Scale-Free Networks are Robust

- Complex systems (cell, internet, social networks), are resilient to component failure

- Network topology plays an important role in this robustness
  - Even if ~80% of nodes fail, the remaining ~20% still maintain network connectivity

- *Attack vulnerability* if hubs are selectively targeted

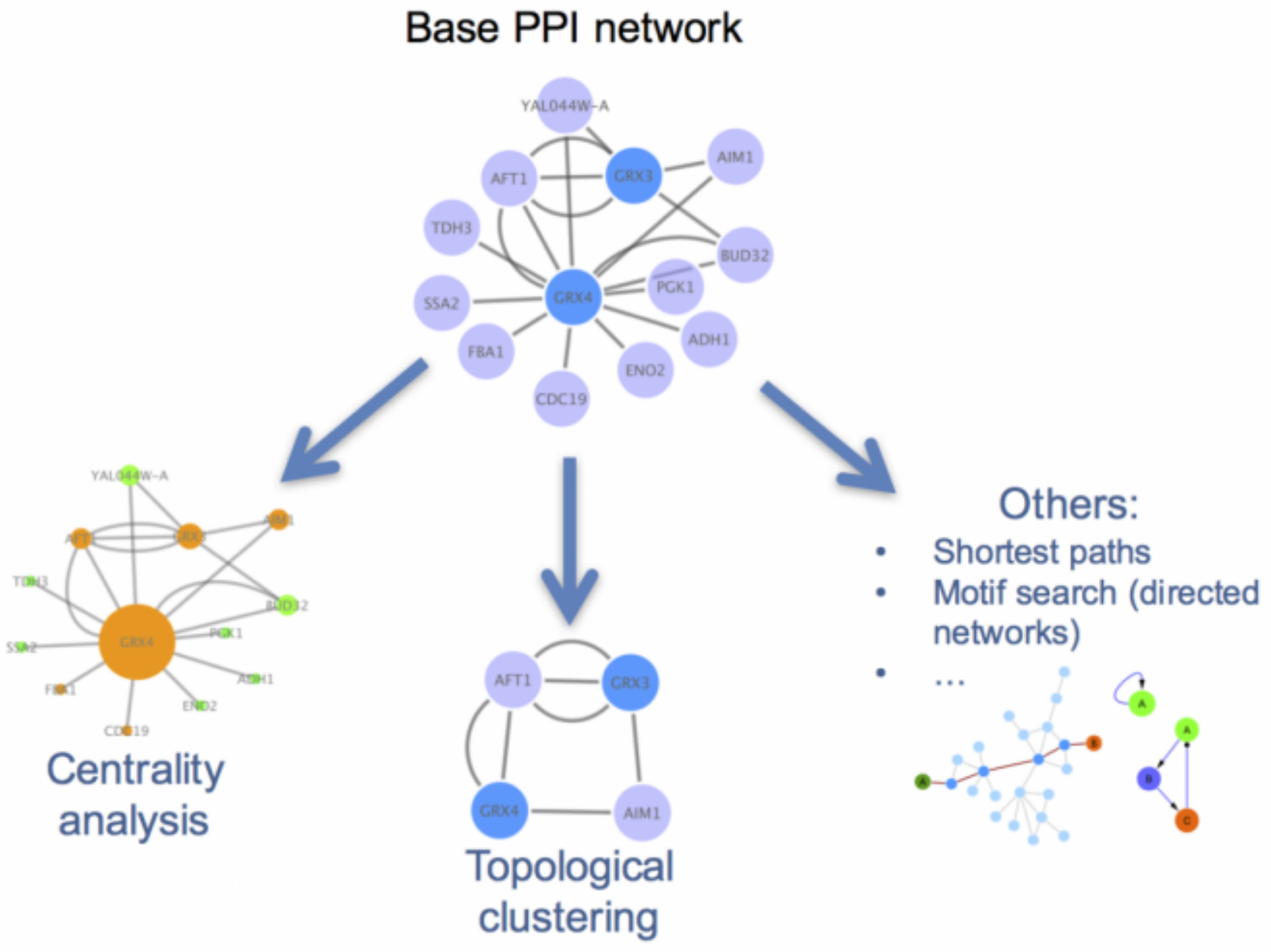
- In yeast, only ~20% of proteins are lethal when deleted, and are 5 times more likely to have degree $k>15$ than $k<5$.

Implications

- Many biological networks (protein-protein interaction networks regulatory networks, etc…) are thought to have hubs, or nodes with high degree.

- For protein-protein interaction networks (PPIs) these hubs have been shown to be older [1] and more essential than random proteins [2]
  - [1] Fraser et al. *Science* (2002) 296:750
  - [2] Jeoung et al. *Nature* (2001) 411:41

Analyzing the topological features of a network is a useful way of identifying relevant participants and substructures that may be of biological significance.



Centrality analysis

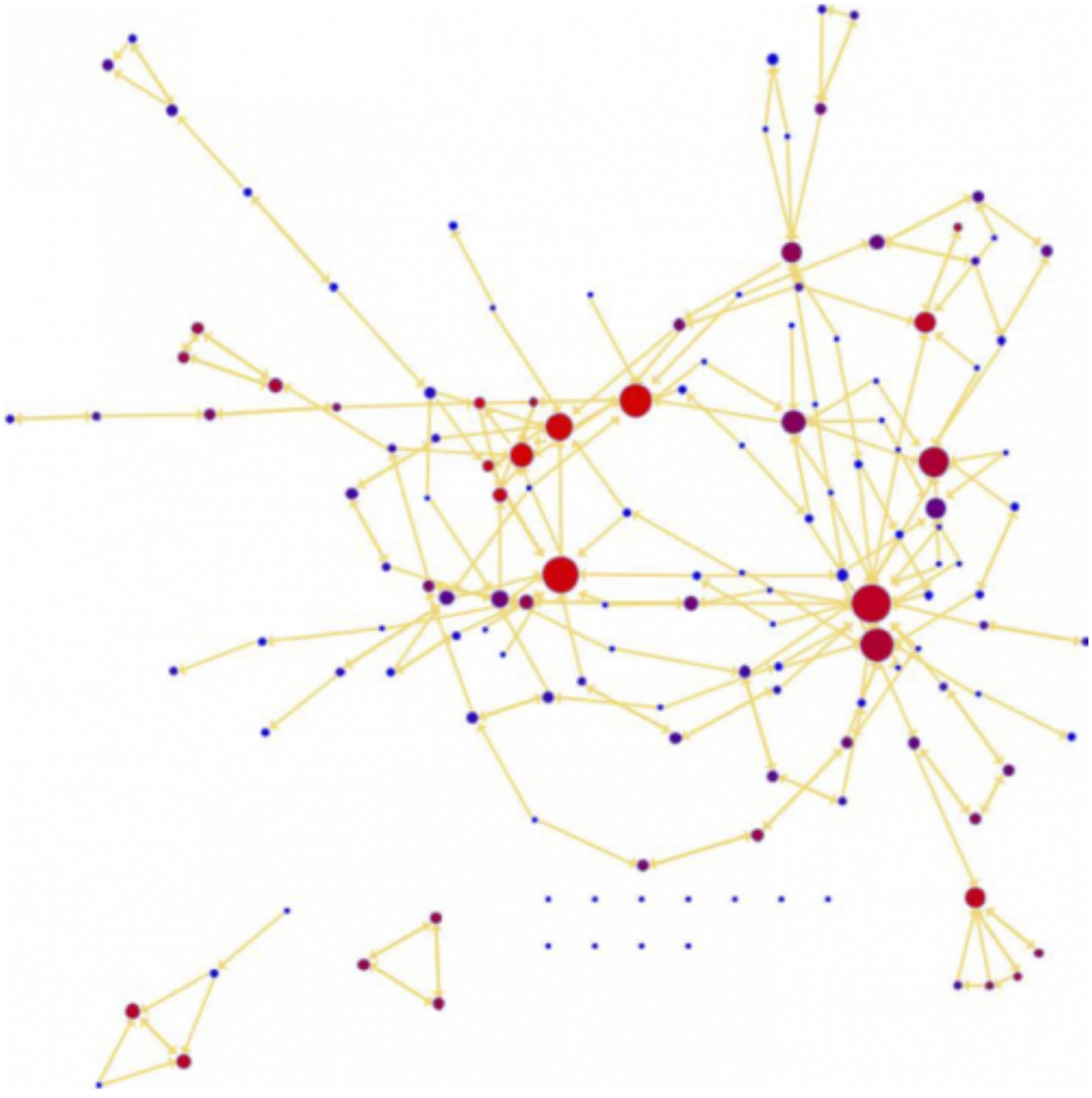
- Centrality gives an estimation on how important a node or edge is for the connectivity or the information flow of the network

- It is a useful parameter in signalling networks and it is often used when trying to find drug targets.

- Centrality analysis in PPINs usually aims to answer the following question:

  ➡ Which protein is the most important and why?

Bigger, redder nodes have higher **centrality values** in this representation.

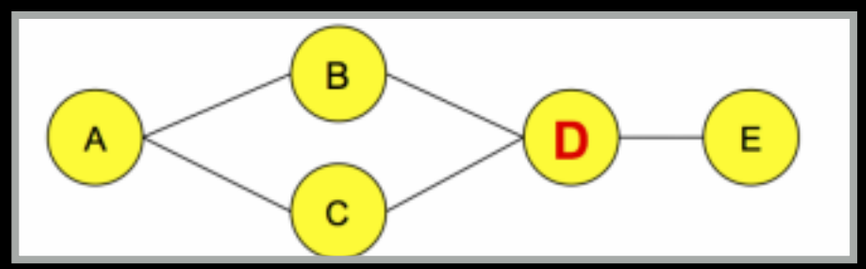


Betweenness centrality

- Nodes with a high betweenness centrality are interesting because they lie on communication paths and can control information flow.

- The number of shortest paths in the graph that pass through the node divided by the total number of shortest paths.

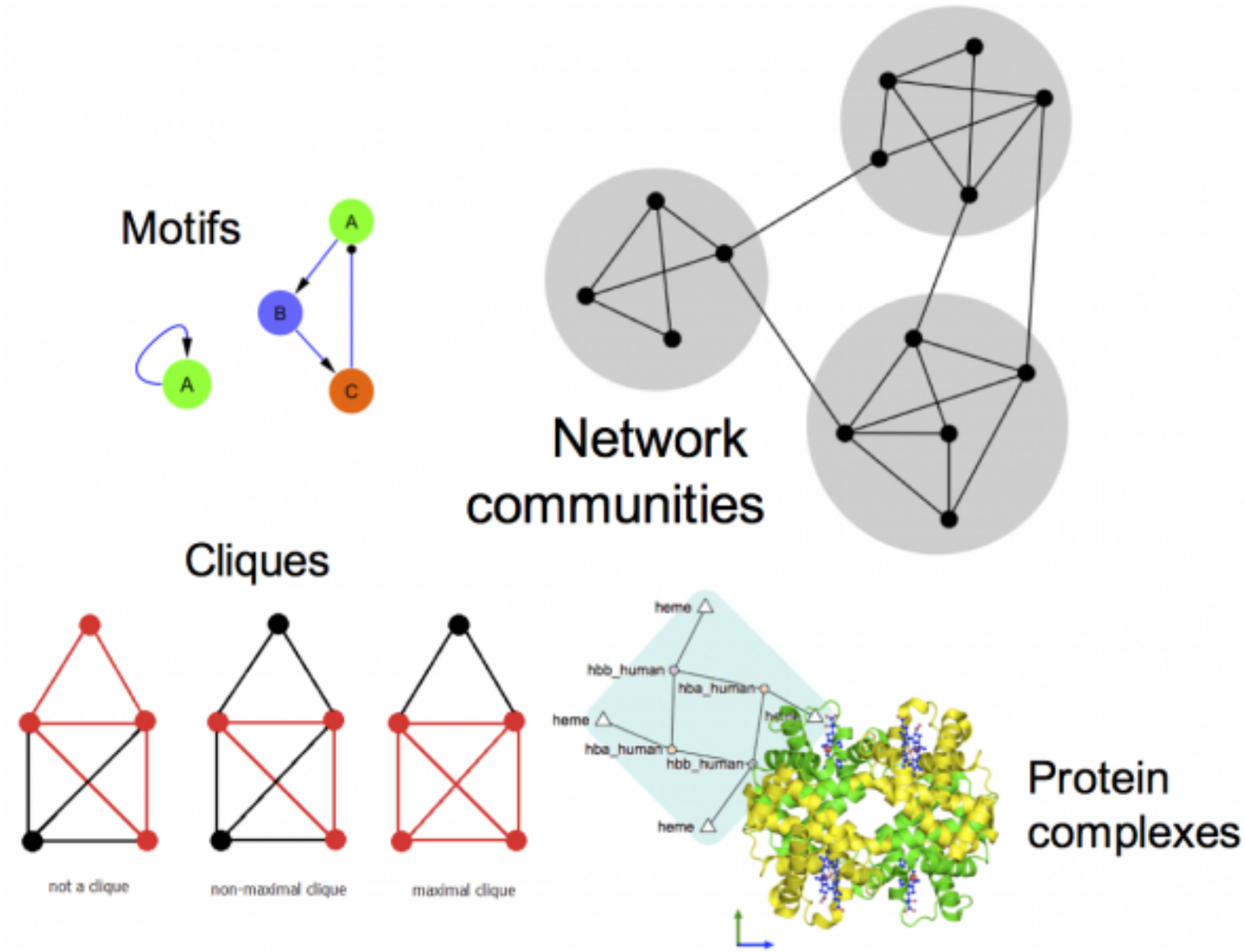
- Betweenness centrality measures how often a node occurs on all shortest paths between two nodes.



Community analysis

- **Community**: A general, catch-all term that can be defined as a group (i.e. *cluster*) of nodes that are more connected within themselves than with the rest of the network. The precise definition for a community will depend on the method or algorithm used to define it.

Looking for communities in a network is a nice strategy for reducing network complexity and extracting functional modules (e.g. protein complexes) that reflect the biology of the network.



Major impact areas for genomic medicine

- Cancer: Identification of driver mutations and drugable variants, Molecular stratification to guide and monitor treatment, Identification of tumor specific variants for personalized immunotherapy approaches (precision medicine).

- Genetic disease diagnose: Rare, inherited and so-called ‘mystery’ disease diagnose.

- Health management: Predisposition testing for complex diseases (e.g. cardiac disease, diabetes and others), optimization and avoidance of adverse drug reactions.

- Health data analytics: Incorporating genomic data with additional health data for improved healthcare delivery.

Major impact areas for genomic medicine

- Cancer: Identification of driver mutations and drugable variants, Molecular stratification to guide and monitor treatment, Identification of tumor specific variants for personalized immunotherapy approaches (precision medicine).

- Genetic disease diagnose: Rare, inherited and so-called ‘mystery’ disease diagnose.

- Health management: Predisposition testing for complex diseases (e.g. cardiac disease, diabetes and others), optimization and avoidance of adverse drug reactions.

- Health data analytics: Incorporating genomic data with additional health data for improved healthcare delivery.

Genomics in the whole of life healthcare