Introduction to Clustering and Spectral Clustering

Karl Rohe

Originally a job talk at Williams College in January 2011.

1) **Intro to clustering**

2) **Spectral clustering**

Clustering divides a data set into sets of similar points.

Useful

Subjective

Computationally Challenging

Clustering has many applications

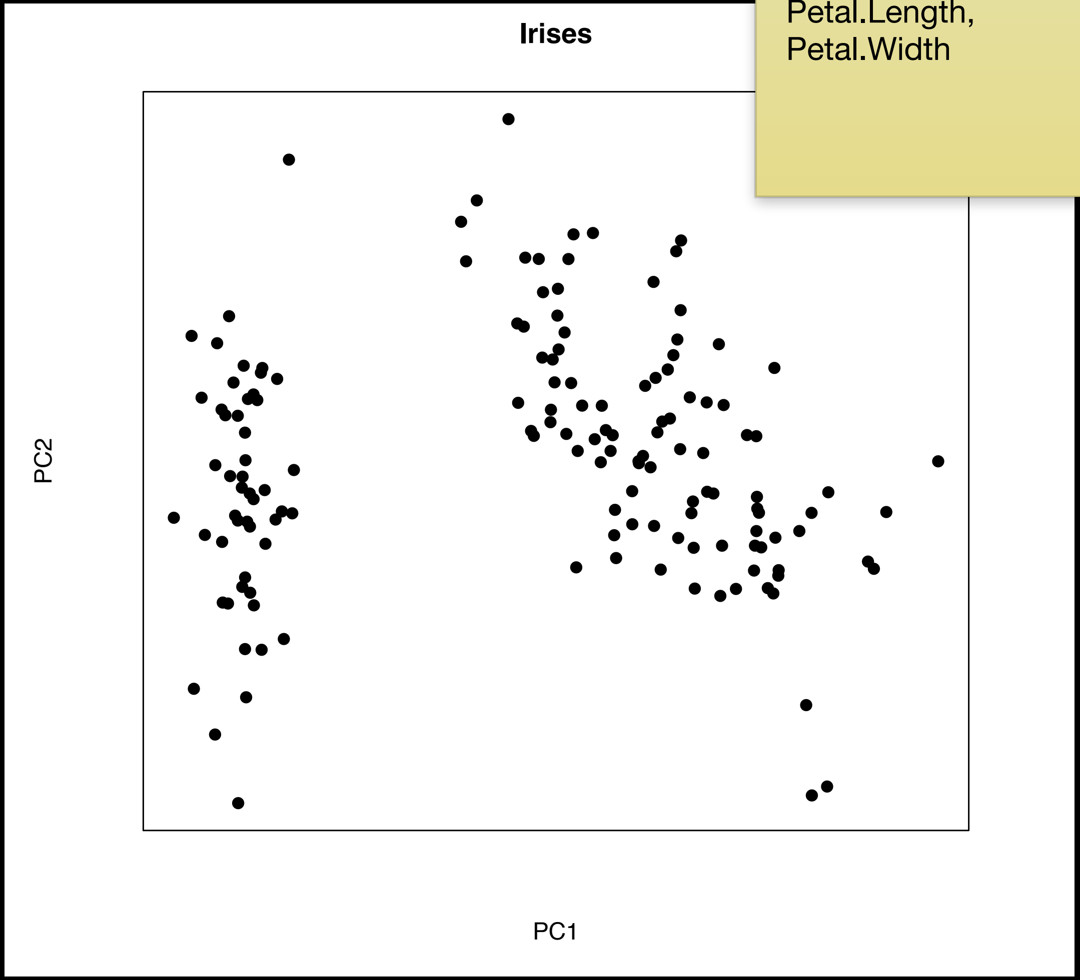
- Urbanization
- Irises
- Financial sectors
- Dolphin social network
- Natural image patches

Image from NASA

Clustering has many applications

Sepal.Length,
Sepal.Width,
Petal.Length,
Petal.Width

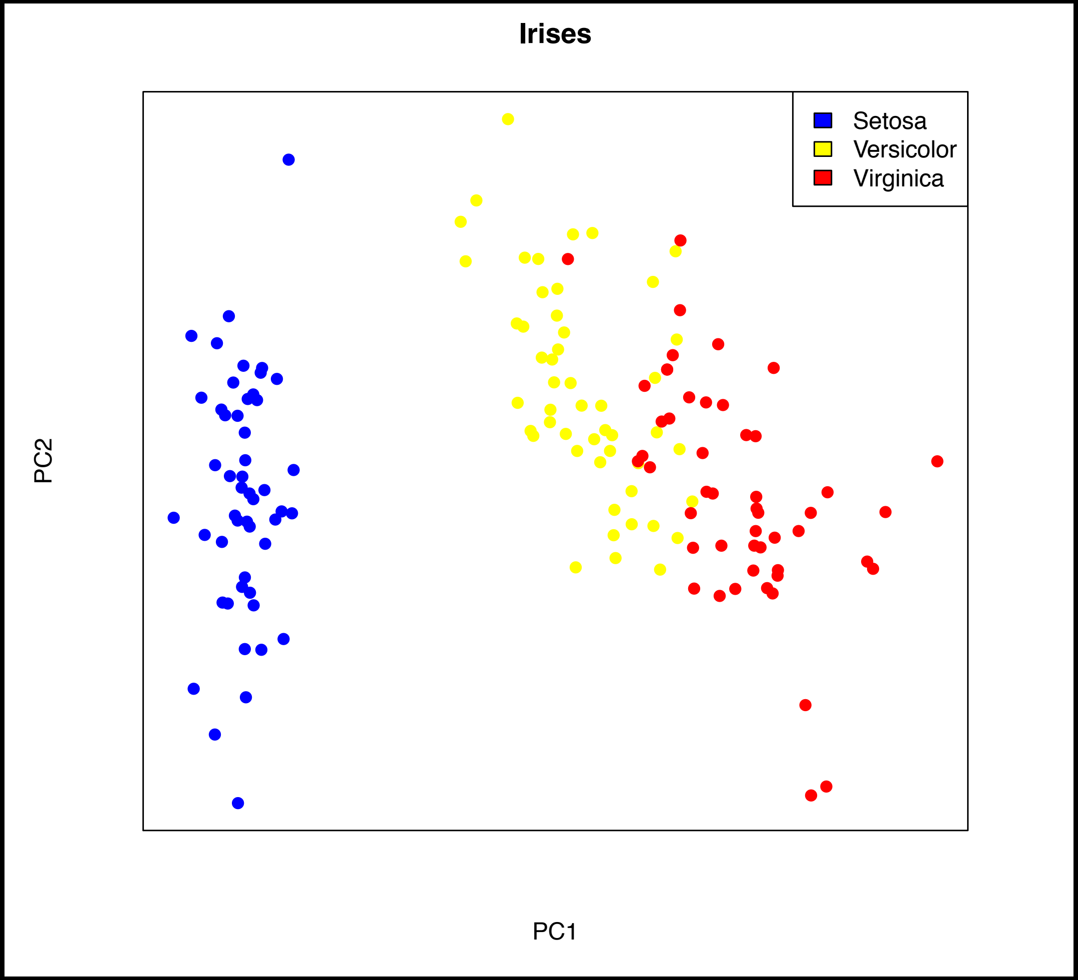
- Urbanization
- Irises
- Financial sectors
- Dolphin social network
- Natural image patches



Anderson, Edgar (1935). The irises of the Gaspe Peninsula, *Bulletin of the American Iris Society*, **59**, 2–5

Clustering has many applications

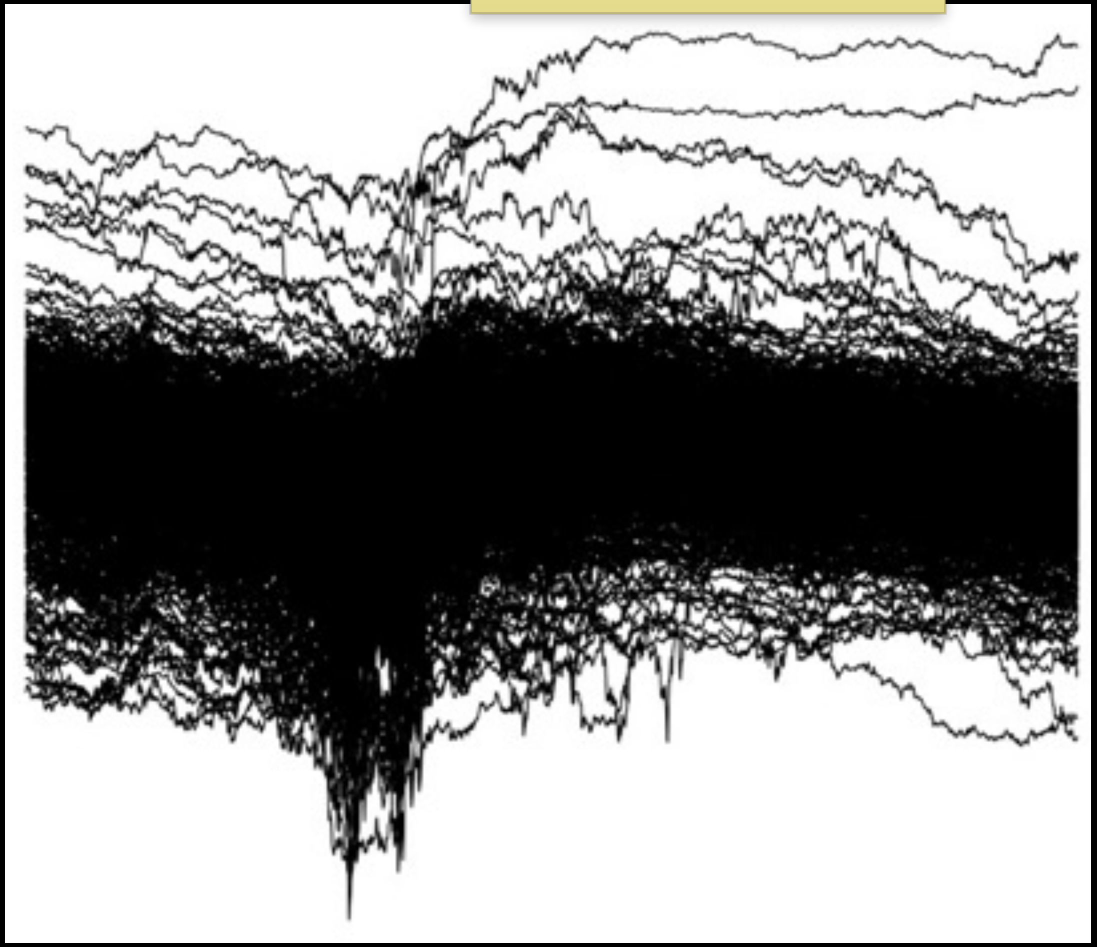
- Urbanization
- **Iris species**
- Financial sectors
- Dolphin social network
- Natural image patches



Anderson, Edgar (1935). The irises of the Gaspe Peninsula, *Bulletin of the American Iris Society*, **59**, 2–5

Clustering has many applications

- Urbanization
- Iris species
- **Financial sectors**
- Dolphin social network
- Natural image patches



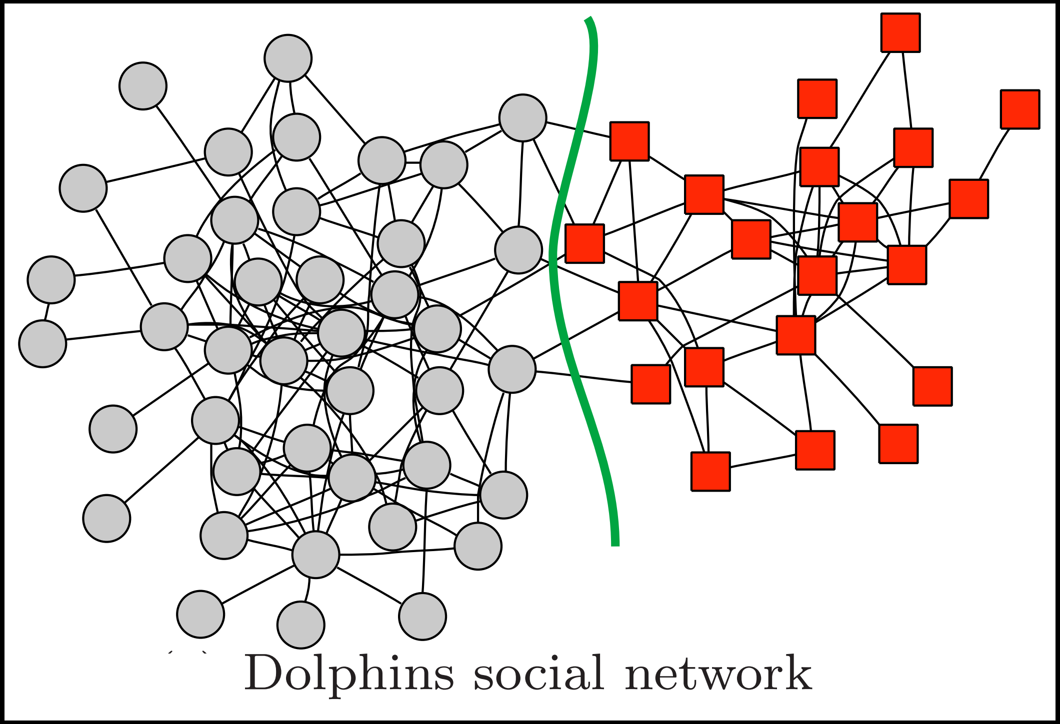
Daily closing stock prices for all S&P 500
Going back to 2005

Clustering has many applications

- Urbanization
- Iris species
- Financial sectors
- **Dolphin social network**
- Natural image patches

Clustering has many applications

- Urbanization
- Iris species
- Financial sectors
- **Dolphin social network**
- Natural image patches



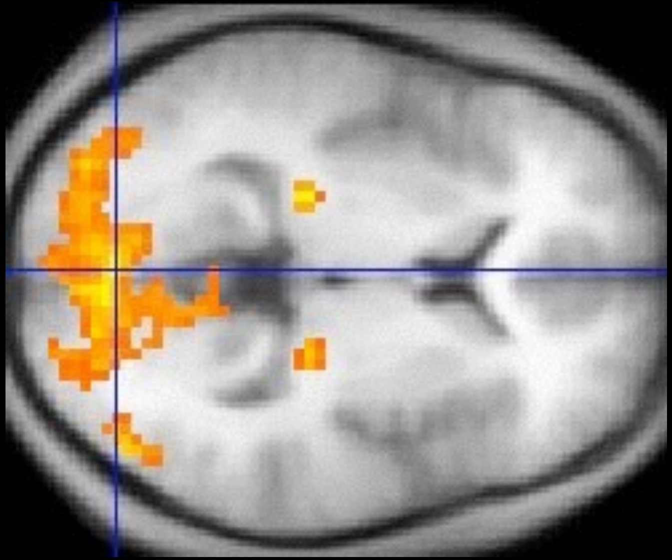
D. Lusseau, K. Schneider, O.J. Boisseau, P. Haase, E. Slooten, and S.M. Dawson. The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology*, **54**:396–405, 2003.

Clustering has many applications

- Urbanization
- Iris species
- Financial sectors
- Dolphin social network
- Natural image patches

what you are seeing changes what is going on in the back of your brain.

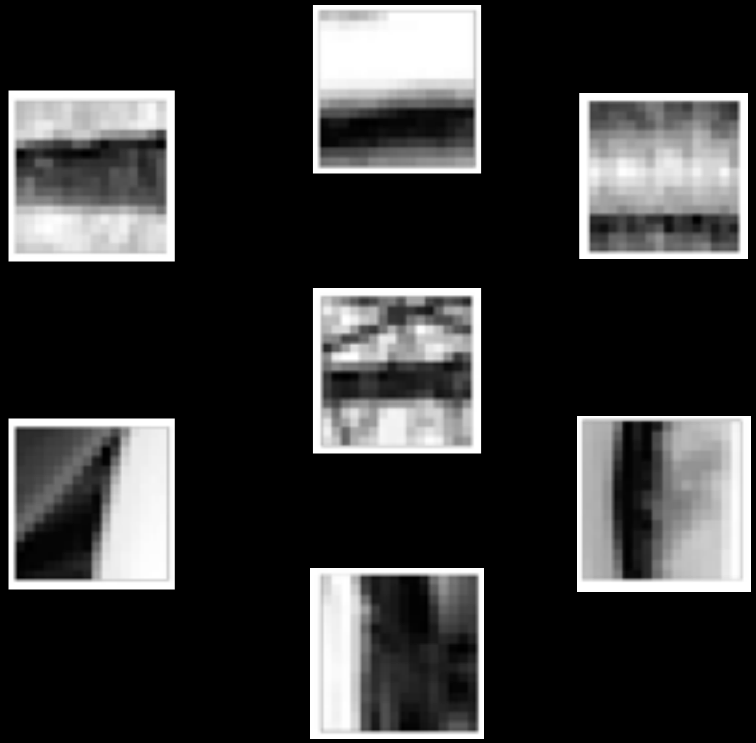
fMRI can measure this.



Gallant Neuroscience lab

Clustering has many applications

- Urbanization
- Iris species
- Financial sectors
- Dolphin social network
- Natural image patches



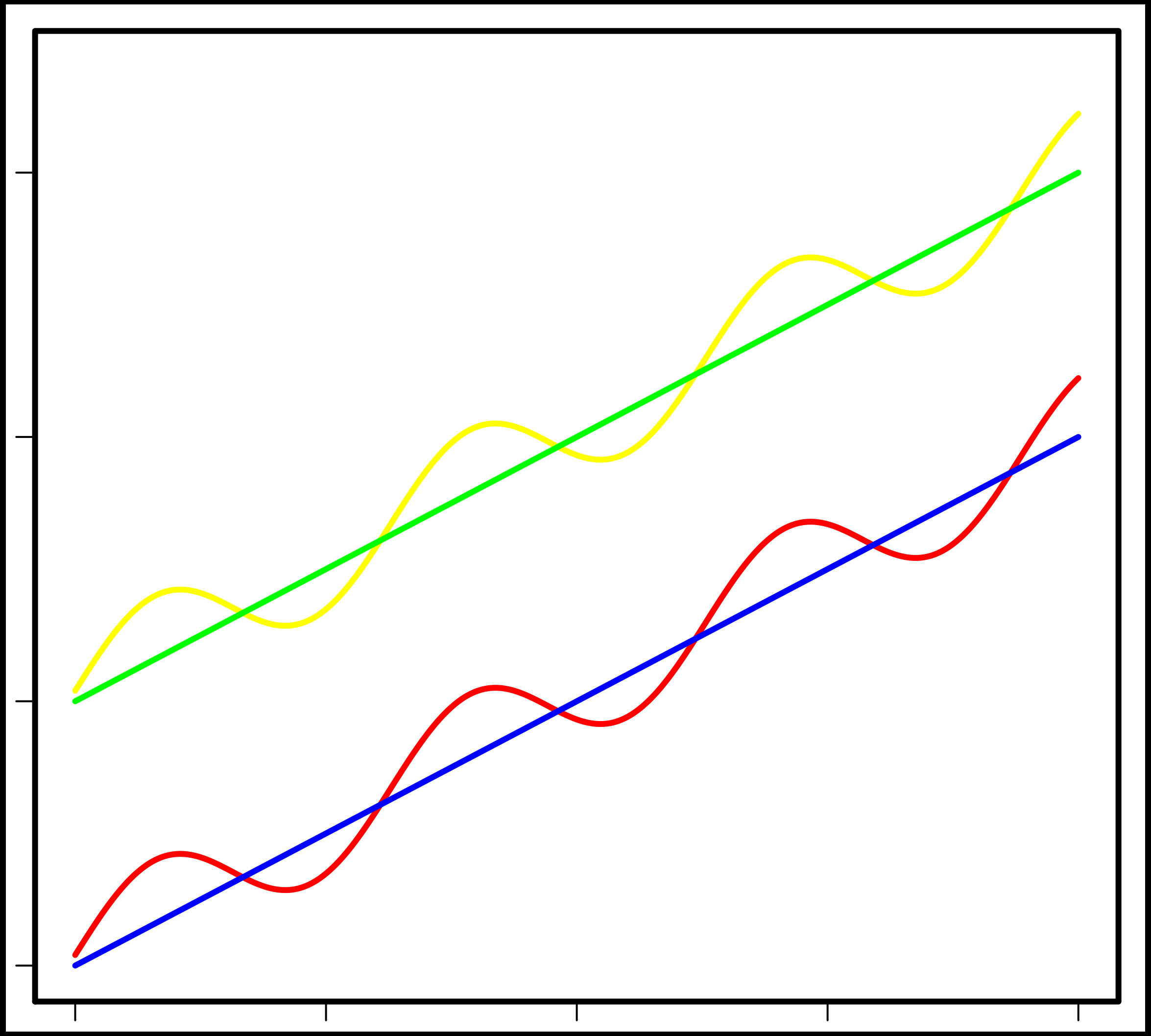
Clustering divides a data set into sets of similar points.

Useful

Subjective

Computationally Challenging

Grouping similar objects is natural. But, how do you define "similar"?



Grouping similar objects is natural. But, how do you define "similar"?

On those remote pages [of an ancient Chinese encyclopedia] it is written that animals are divided into (a) those that belong to the Emperor, (b) embalmed ones, (c) those that are trained, (d) suckling pigs, (e) mermaids, (f) fabulous ones, (g) stray dogs, (h) those that are included in this classification, (e) those that tremble as if they were mad, (j) innumerable ones, (k) those drawn with a very fine camel's hair brush, (l) other, (m) those that have just broken a flower vase, (n) those that resemble flies from a distance.

- Jorge Luis Borges, Other Inquisitions

Clustering divides a data set into sets of similar points.

Useful

Subjective

Computationally Challenging

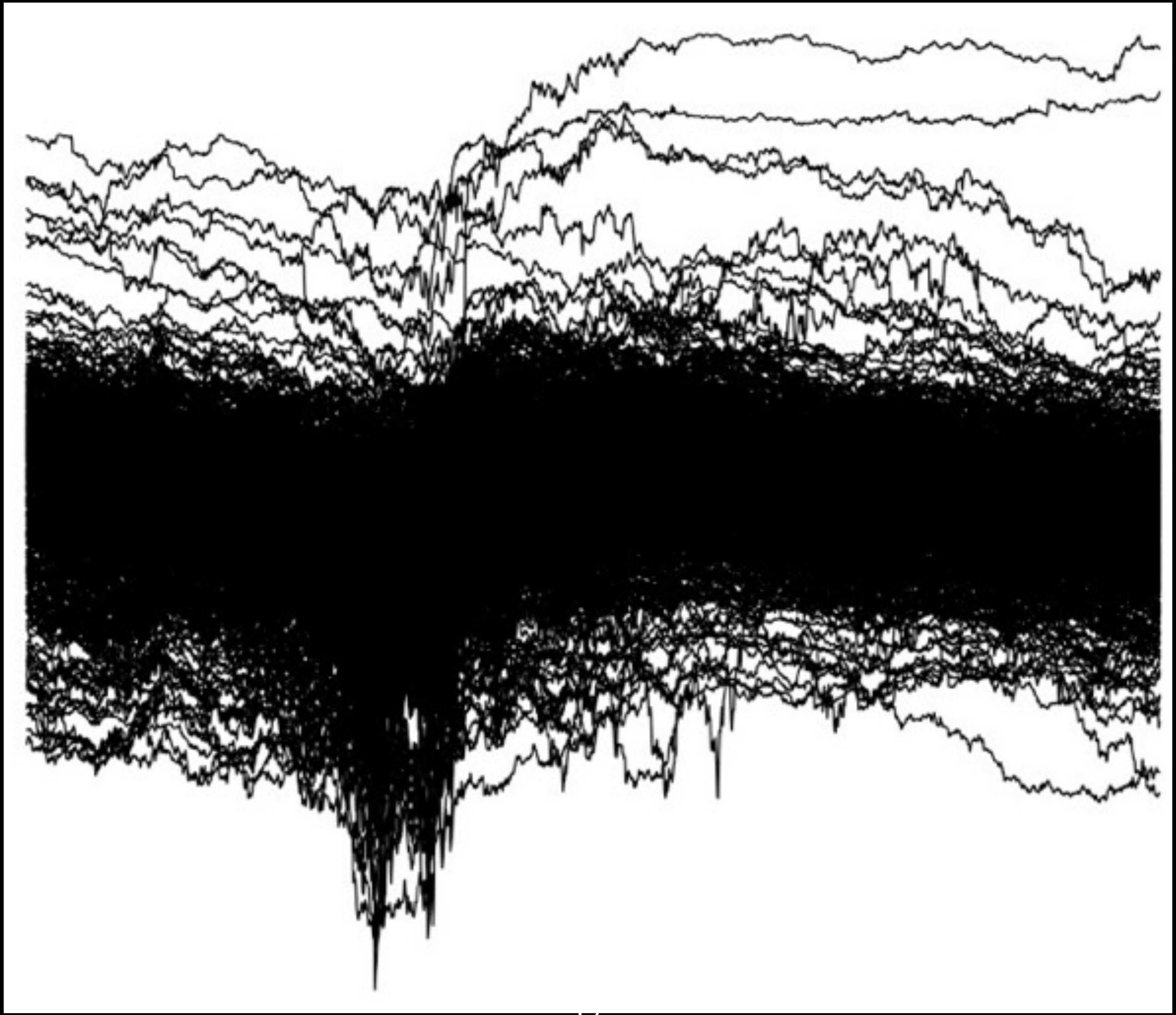








Log daily closing stock prices for all S&P 500, 2005 - Present.



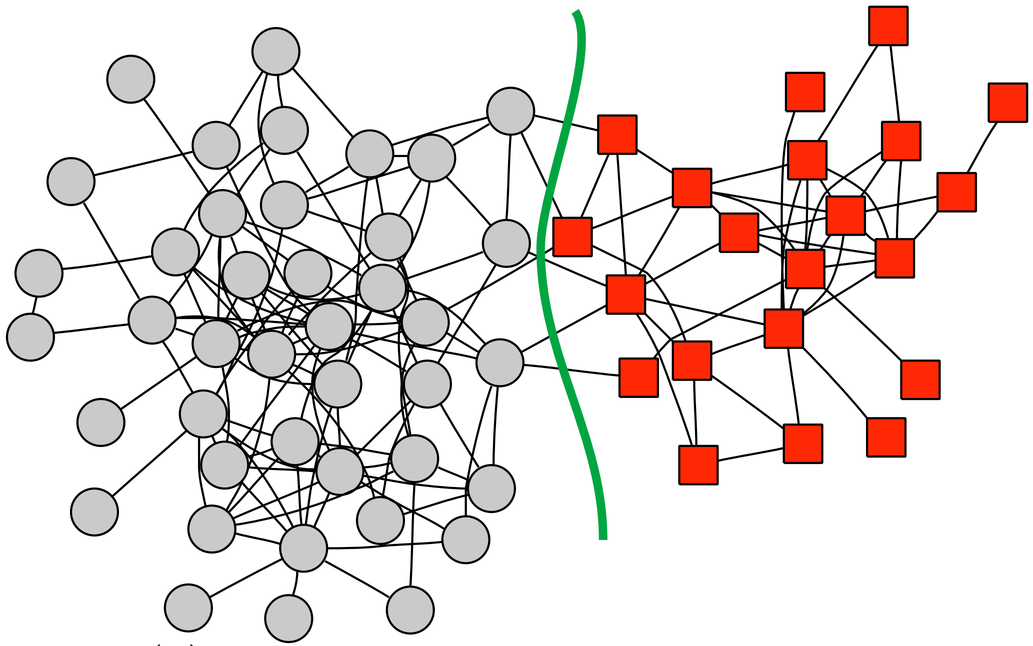
While clustering is natural for humans.
Clustering large data sets is difficult.
We need to teach computers how to cluster!

While clustering is natural for humans. Clustering large data sets is difficult. We need to teach computers how to cluster!

- Computers only understand algorithms.
- Often, clustering algorithm are motivated by optimization problems.

Partitions

Partitions



Dolphins social network

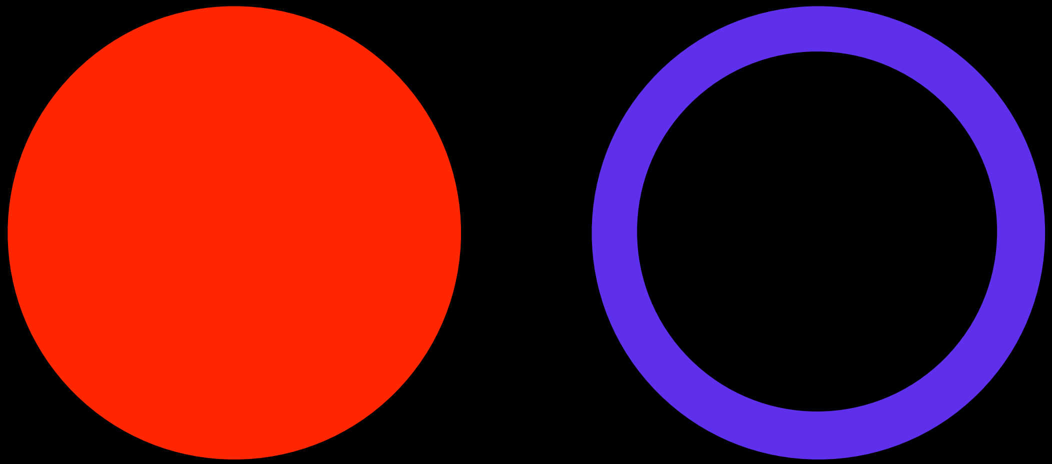
Clustering as an optimization problem

$$\min_{C} f(C)$$

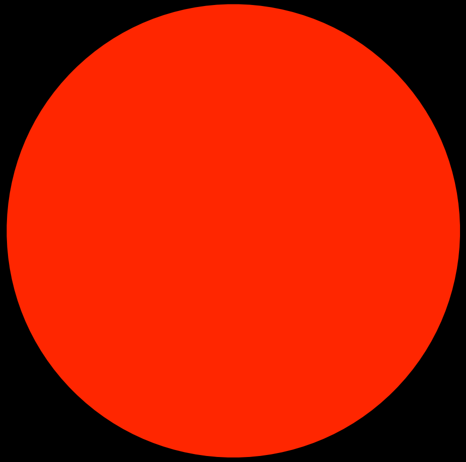
C: any partition
f: measures how bad C is

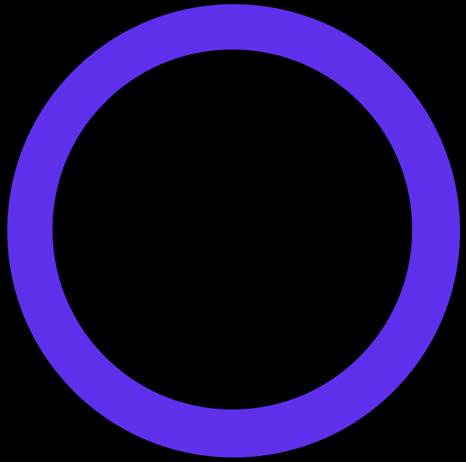
Convexity

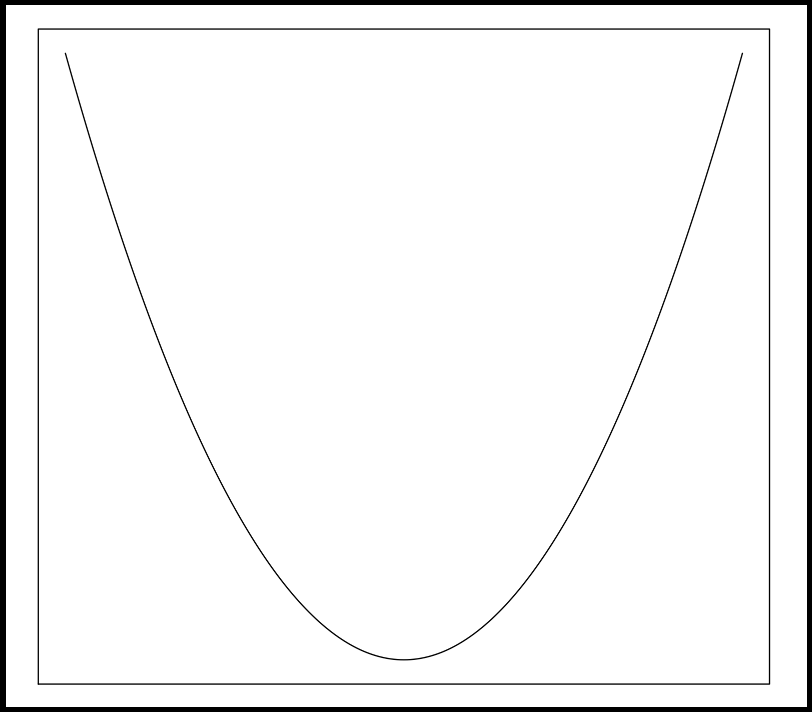
Convexity



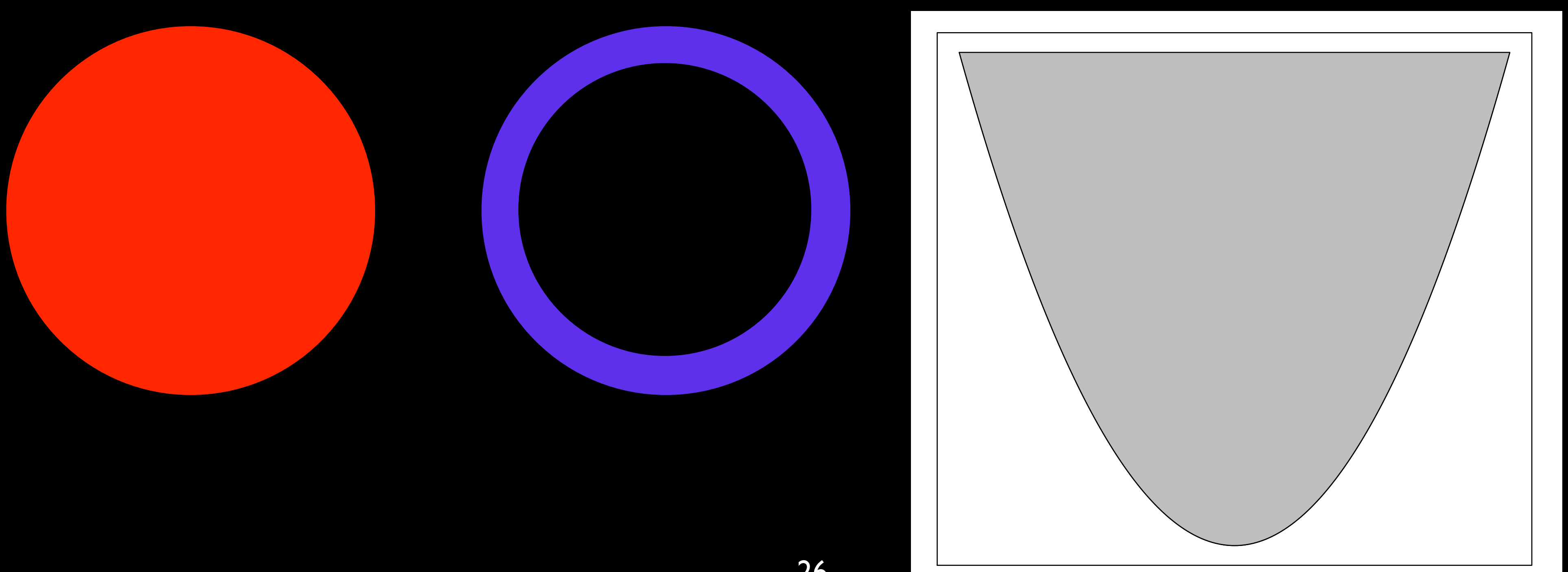
Convexity







Convexity

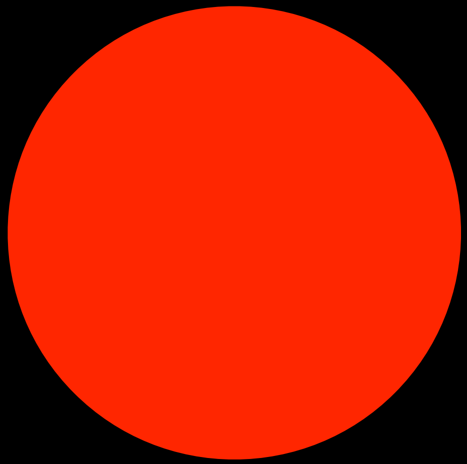


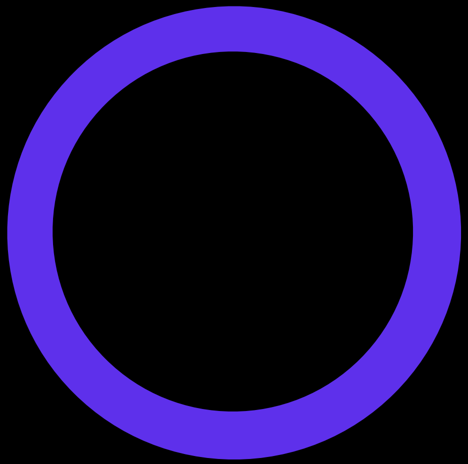
Convexity

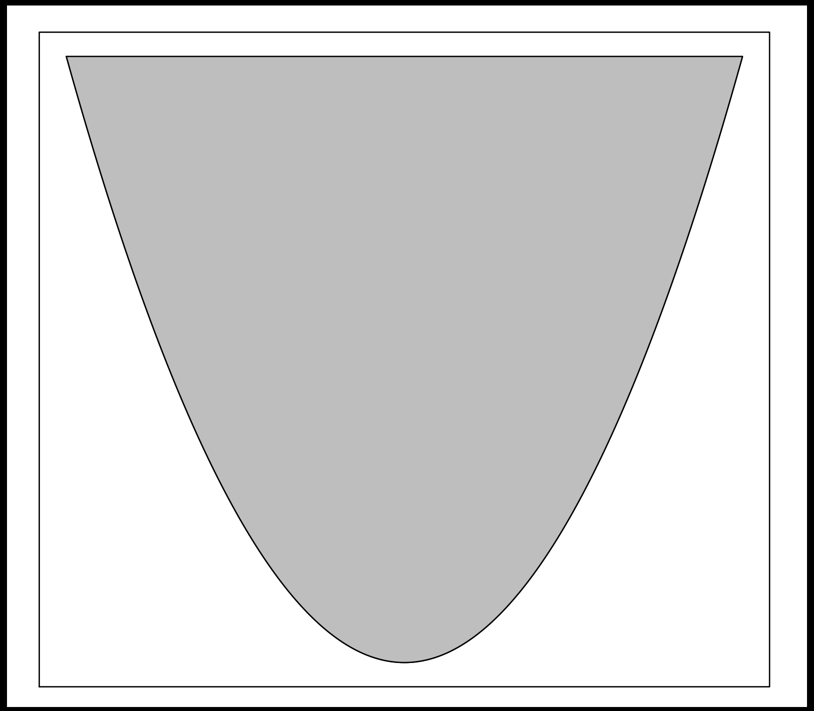
Minimizing *convex functions* on *convex sets* is "easy"

Set the derivative equal to zero or

Just go downhill!







Usually, clustering problems are not convex.

$$x_1, \ldots, x_n \in R^d$$

Usually, clustering problems are not convex.

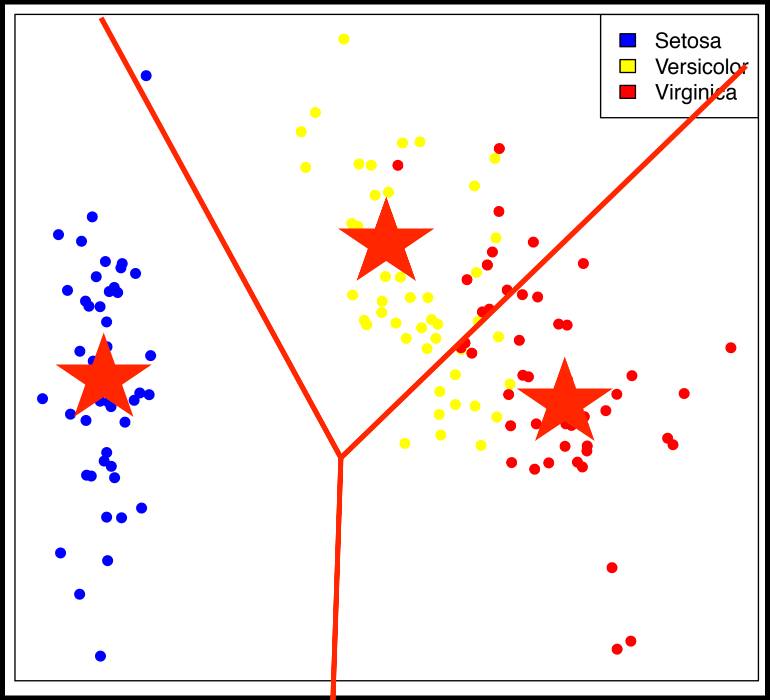
k-means $x_1, \ldots, x_n \in R^d$

$$f(c_1, \ldots, c_k) = \sum_i \min_j \|x_i - c_j\|_2^2$$

Usually, clustering problems are not convex.

k-means $x_1, \ldots, x_n \in R^d$

$$f(c_1, \ldots, c_k) = \sum_i \min_j \|x_i - c_j\|_2^2$$

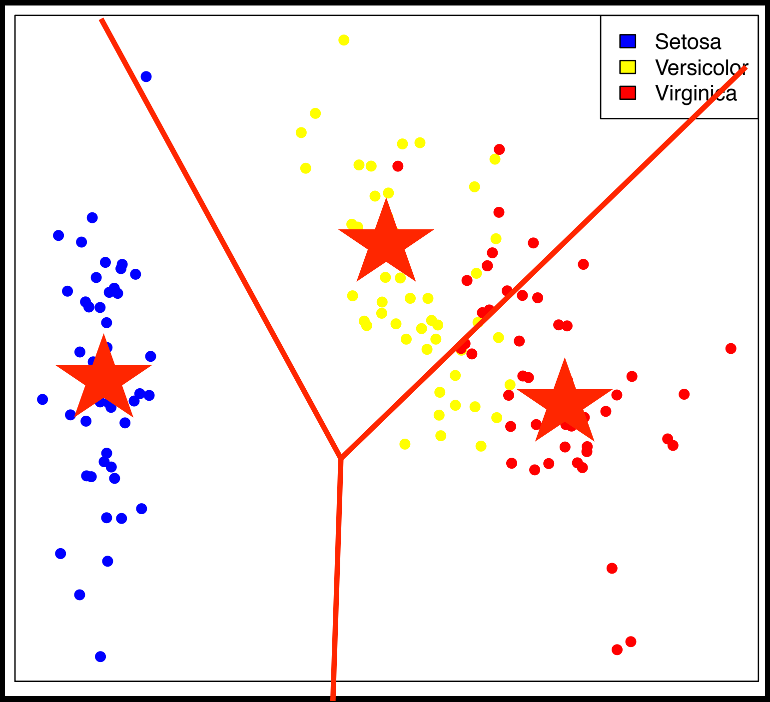


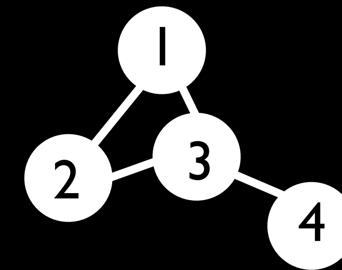
Usually, clustering problems are not convex.

k-means $\qquad x_1, \ldots, x_n \in R^d$

$$f(c_1, \ldots, c_k) = \sum_i \min_j \|x_i - c_j\|_2^2$$

$$\min_{c_1, \ldots c_k} f(c_1, \ldots, c_k)$$

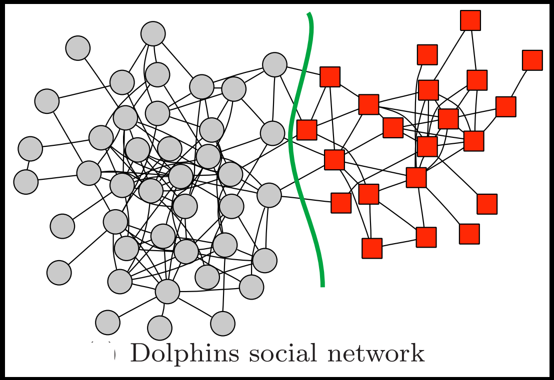




Symmetric adjacency matrix $A \in \{0,1\}^{n \times n}$

$$A_{ij} = \begin{cases} 1 & \text{if node } i \text{ is connected to node } j \\ 0 & \text{otherwise} \end{cases}$$

Normalized cuts:



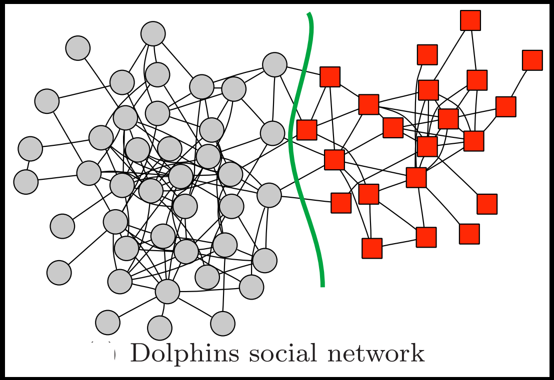
Symmetric adjacency matrix $A \in \{0,1\}^{n \times n}$

$$A_{ij} = \begin{cases} 1 & \text{if node } i \text{ is connected to node } j \\ 0 & \text{otherwise} \end{cases}$$

$$\min_{P,Q} \frac{\sum_{i \in P, j \in Q} A_{ij}}{\sum_{i \in P, j \in P \cup Q} A_{ij}} + \frac{\sum_{i \in P, j \in Q} A_{ij}}{\sum_{i \in Q, j \in P \cup Q} A_{ij}}$$

Shi, Malik (2000). Normalized Cuts and Image Segmentation, *IEEE Transactions Pattern Analysis and Machine Learning*, **22**, 888-905.

Normalized cuts:



It can be shown that normalized cuts is equivalent to the following problem:

$$\min_y \frac{y^T(D-A)y}{y^TDy} \quad \text{subject to} \quad \begin{array}{l} y_i \in \{-1/b, b\} \ \forall i \\ y^T D\mathbf{1} = 0 \end{array}$$

There are ___________ many partitions of n data points into 2 clusters.

There are very many partitions of the data points

There are $\underline{\;2^{n-1}-1\;}$ many partitions of n data points into 2 clusters.

(a,b) (b,a) (ab,) (,ab)

There are very many partitions of the data points

There are $\underline{\quad 2^{n-1} - 1 \quad}$ many partitions of n data points into 2 clusters.

(a,b) ~~(b,a)~~ ~~(ab,)~~ ~~(,ab)~~

There are very many partitions of the data points

There are $\underline{\;2^{n-1} - 1\;}$ many partitions of n data points into 2 clusters.

Champion has postulated that there are

2^{83} atoms in the universe.

Not possible to look at all partitions!

Matthew Champion, "Re: How many atoms make up the universe?", 1998

Clustering poses several challenges.

- The approach you choose is often subjective
- Difficult to optimize.
- Must rely on approximations: Local optima, “convex relaxation.”

Spectral Clustering

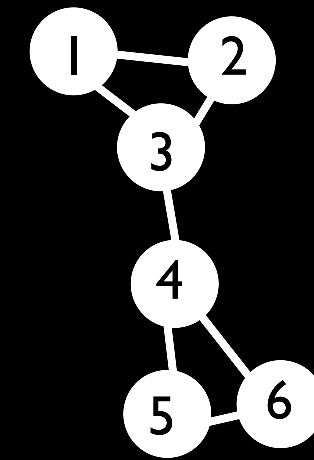
The algorithm

Relationship to normalized cuts

Euclidean version

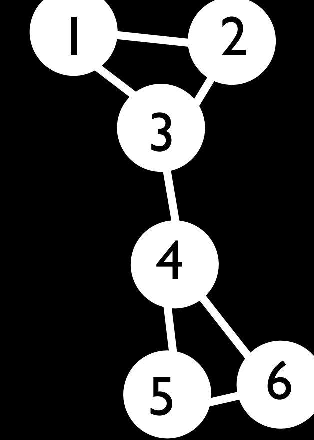
Advantages

The algorithm (with graph data)



$A \in \{0,1\}^{n \times n}$ Adjacency matrix

The algorithm (with graph data)

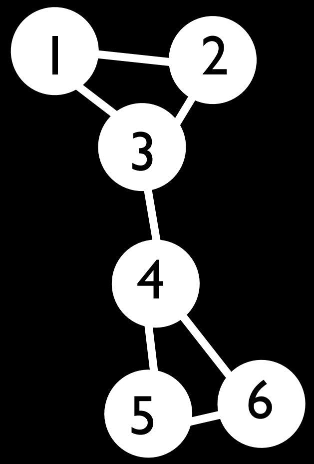


$A \in \{0,1\}^{n \times n}$ Adjacency matrix

$D \in R^{n \times n}$ $\quad D_{ii} = \sum_j A_{ij}$

$L = I - D^{-1} A$

The algorithm (with graph data)



$A \in \{0,1\}^{n \times n}$ Adjacency matrix

$D \in R^{n \times n}$ $\qquad D_{ii} = \sum_j A_{ij}$

$L = I - D^{-1}A$

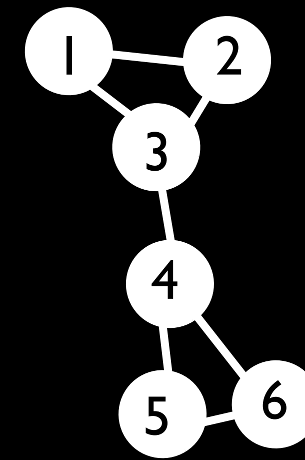
Find the eigenvector corresponding to the second smallest eigenvalue.

$y \in R^n$

Eigenvectors are beautiful AND elusive.

it is difficult to exactly define the vectors .

The algorithm
vith graph data)



$n \times n$ Adjacency matrix

$$D \in R^{n \times n} \qquad D_{ii} = \sum_j A_{ij}$$

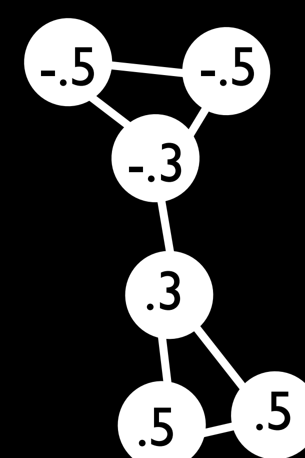
$$L = I - D^{-1} A$$

Find the eigenvector corresponding to the second smallest eigenvalue.

$$y \in R^n \qquad Ly = \lambda y$$

$$y_i < 0 \implies i \in A$$

$$y_i \geq 0 \implies i \in B$$



Spectral Clustering

- The algorithm
- **Relationship to normalized cuts**
- Euclidean version
- Advantages

Recall normalized cuts:

$$\min_{y} \frac{y^T(D-A)y}{y^T D y} \quad \text{subject to} \quad \begin{array}{l} y_i \in \{-1/b, b\} \ \forall i \\ y^T D\mathbf{1} = 0 \end{array}$$

$$\min_{P,Q} \frac{\sum_{i \in P, j \in Q} A_{ij}}{\sum_{i \in P, j \in P \cup Q} A_{ij}} + \frac{\sum_{i \in P, j \in Q} A_{ij}}{\sum_{i \in Q, j \in P \cup Q} A_{ij}}$$

Spectral clustering is a "convex relaxation" of normalized cuts

$$\min_y \frac{y^T(D-A)y}{y^T D y} \quad \text{subject to} \quad \begin{array}{l} y_i \in \{-1/b, b\} \ \forall i \\ y^T D \mathbf{1} = 0 \end{array}$$

Because of the restriction to a discrete set, this problem is not convex. "Relax" the problem. Optimize over

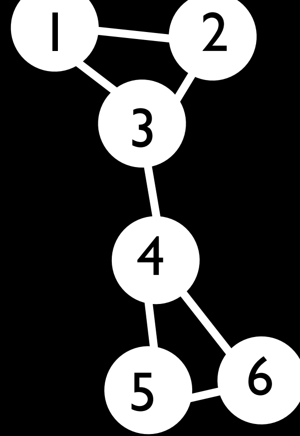
$$y \in R^n, \ y^T D \mathbf{1} = 0$$

The optimum is the second smallest eigenvector of L .

Spectral Clustering

- The algorithm
- Relationship to normalized cuts
- Euclidean version
- Advantages

The algorithm (with graph data)

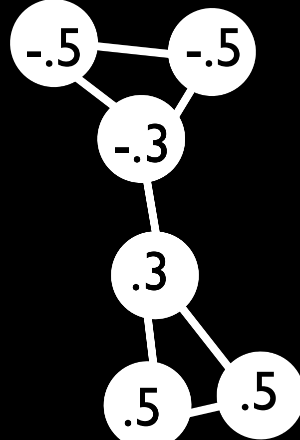


$A \in \{0,1\}^{n \times n}$ Adjacency matrix

$D \in R^{n \times n}$ $\quad D_{ii} = \sum_j A_{ij}$

$L = I - D^{-1}A$

Find the eigenvector corresponding to the second smallest eigenvalue.



$y \in R^n$ $\quad Ly = \lambda y$

$$y_i < 0 \implies i \in A$$

$$y_i \geq 0 \implies i \in B$$

The algorithm (with Euclidean data)

$K \in R^{n \times n}$ $\quad K_{ij} = \exp(-\|x_i - x_j\|_2^2 / \sigma^2)$

$D \in R^{n \times n}$ $\quad D_{ii} = \sum_j K_{ij}$

Here the subjectivity of the similarity function is obvious! Need to choose a function for K_{ij}

$L = I - D^{-1}K$

Find the eigenvector corresponding to the second smallest eigenvalue.

$y \in R^n$ $\quad Ly = \lambda y$

$$y_i < 0 \implies i \in A$$
$$y_i \geq 0 \implies i \in B$$

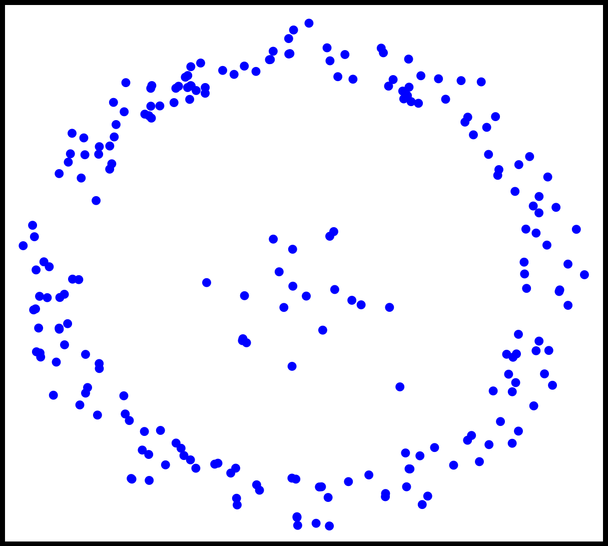
Spectral Clustering

The algorithm

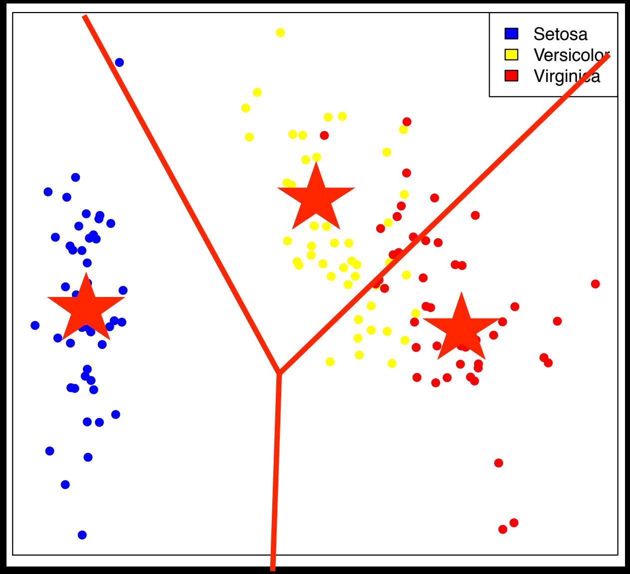
Relationship to normalized cuts

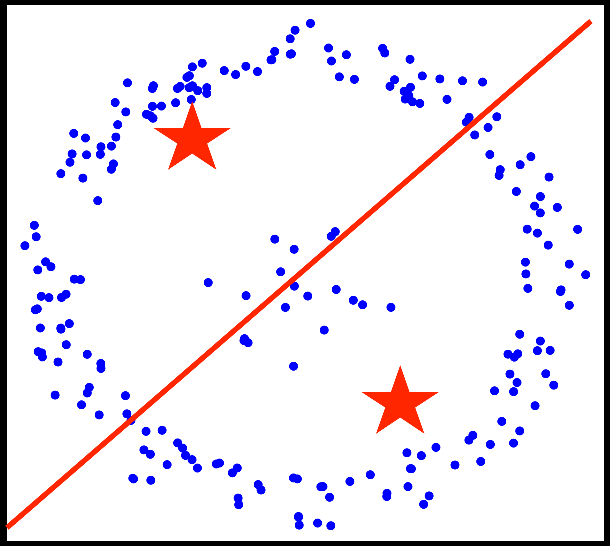
Euclidean version

Advantages

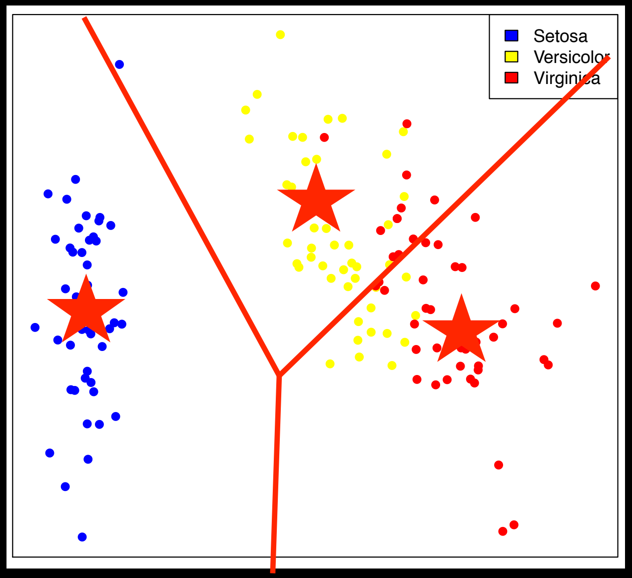


k-means

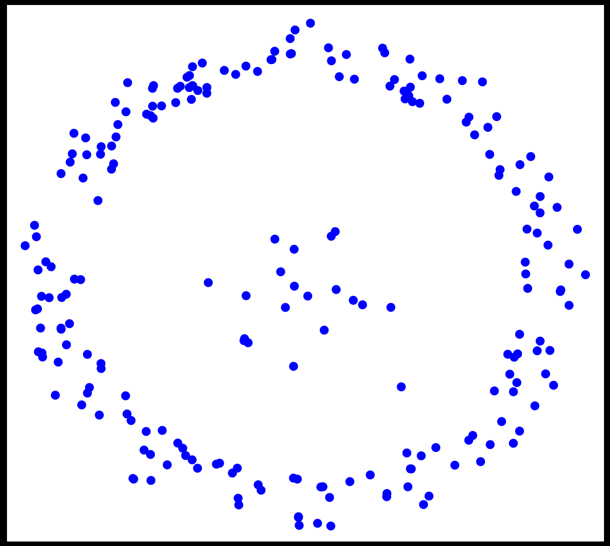




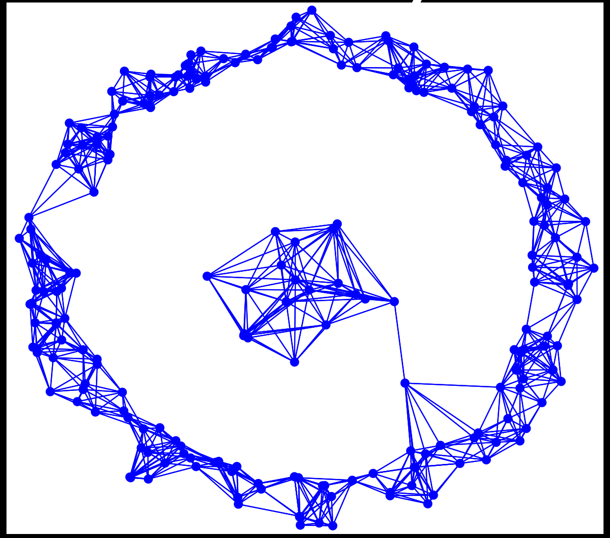
k-means



Because it relies on a "local" measure of similarity, spectral clustering can detect oddly shaped clusters.



Another similarity function



"k-nearest neighbors"

Spectral clusterin

In contrast to regression , clustering is much more subjective.

there are several different clustering algorithms?

What criteria should you use to decide?

- is computationally tractable.
- is a "convex relaxation" of normalized cuts.
- is able to find oddly shaped clusters.
- has interesting connections to
  - spectral graph theory,
  - random walks on graphs,
  - and electrical network theory.

1) **Intro to clustering**

2) **Spectral clustering**

3) **Research**

Why should you trust spectral clustering?

Statistical Estimation

Stochastic Blockmodel

Theorem

A statistical model to study an algorithm

For example, say you have the GPA of some students

$$y_1, \ldots, y_n \in R$$

and some predictors (height, SAT score, # roommates, etc.)

$$x_1, \ldots, x_n \in R^p$$

Say $Y_i = X_i\beta + \mathrm{error}_i$ with a *few conditions on the error distribution*, is a reasonable model for the data.

A statistical model to study an algorithm

$$Y_i = X_i\beta + \text{error}_i$$

What can be said about

$$\hat{\beta}_n = \text{argmin}_{b\in R^d} \sum_{i=1}^{n} (Y_i - X_i b)^2$$

One desirable result: $\hat{\beta}_n \to \beta$

This would suggest that least squares is reasonable.

To study the estimation performance of spectral clustering, we need a statistical model.

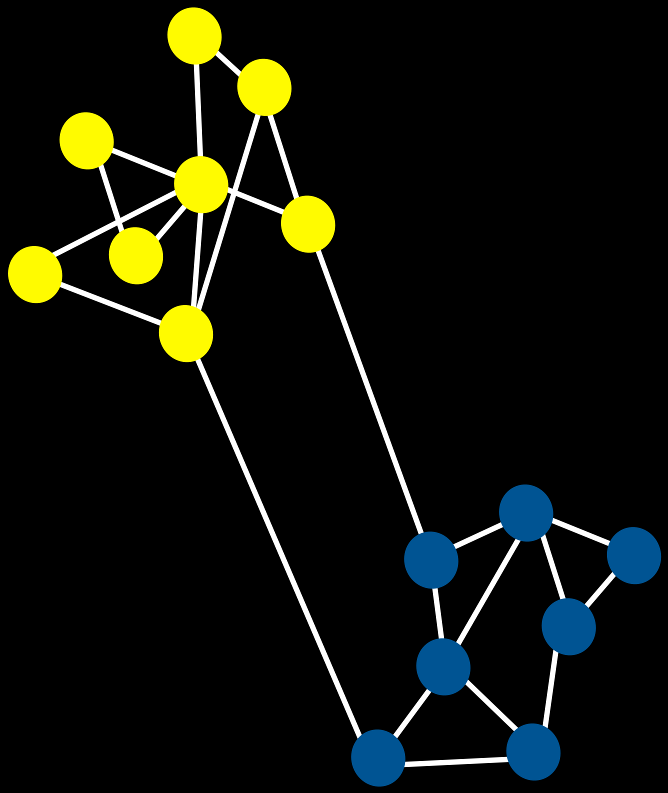
Why should you trust spectral clustering?

Statistical Estimation

Stochastic Blockmodel

Theorem

A simple example of the Stochastic Block Model



- Divide nodes into k blocks
- Let each block have an equal proportion of the nodes
- Edges: independent, Bernoulli with probability
  - p if nodes in same block
  - r otherwise

Clustering: estimate block membership for each node

Why should you trust spectral clustering?

Statistical Estimation

Stochastic Blockmodel

Theorem

Theorem

Under the previously described Stochastic Block Model, for $p \neq r$, the number of "misclustered" nodes is bounded

$$\text{number of misclustered nodes} = o(k^3 \log^2 n) \ a.s.$$

as the number of nodes $n \to \infty$ and the number of blocks $k = k(n)$

R., Chatterjee, Yu. Spectral clustering and the high-dimensional Stochastic Blockmodel. *Annals of Statistics, pending minor revisions.*

Conclusions

Clustering is useful, subjective, and computationally challenging.

Spectral clustering uses the eigenvectors of the graph Laplacian to relax a non-convex problem.

Spectral clustering can estimate the blocks in the Stochastic Blockmodel.

References

Anderson (1935). The irises of the Gaspe Peninsula, *Bulletin of the American Iris Society*, **59**, 2–5

Lusseau, Schneider, Boisseau, Haase, Slooten, and Dawson (2003). The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology*, **54**:396–405.

R., Chatterjee, Yu (2010). Spectral clustering and the high-dimensional Stochastic Blockmodel. *Annals of Statistics, pending minor revisions*.

Shi, Malik (2000). Normalized Cuts and Image Segmentation, *IEEE Transactions Pattern Analysis and Machine Learning*, **22**:888-905.