Special Topics

PHYS 453

Dr Daugherity

ACU
ABILENE CHRISTIAN UNIVERSITY


ML-309:
UNSUPERVISED LEARNING
THEJENKINSCOMIC

Unsupervised Learning

Problem #1

What can we do without training data?

Can we **learn** anything?

One goal is to look for natural clusters in unlabeled data. We can then group similar data, and even (kinda) classify new data as belonging to a certain group.

Clustering

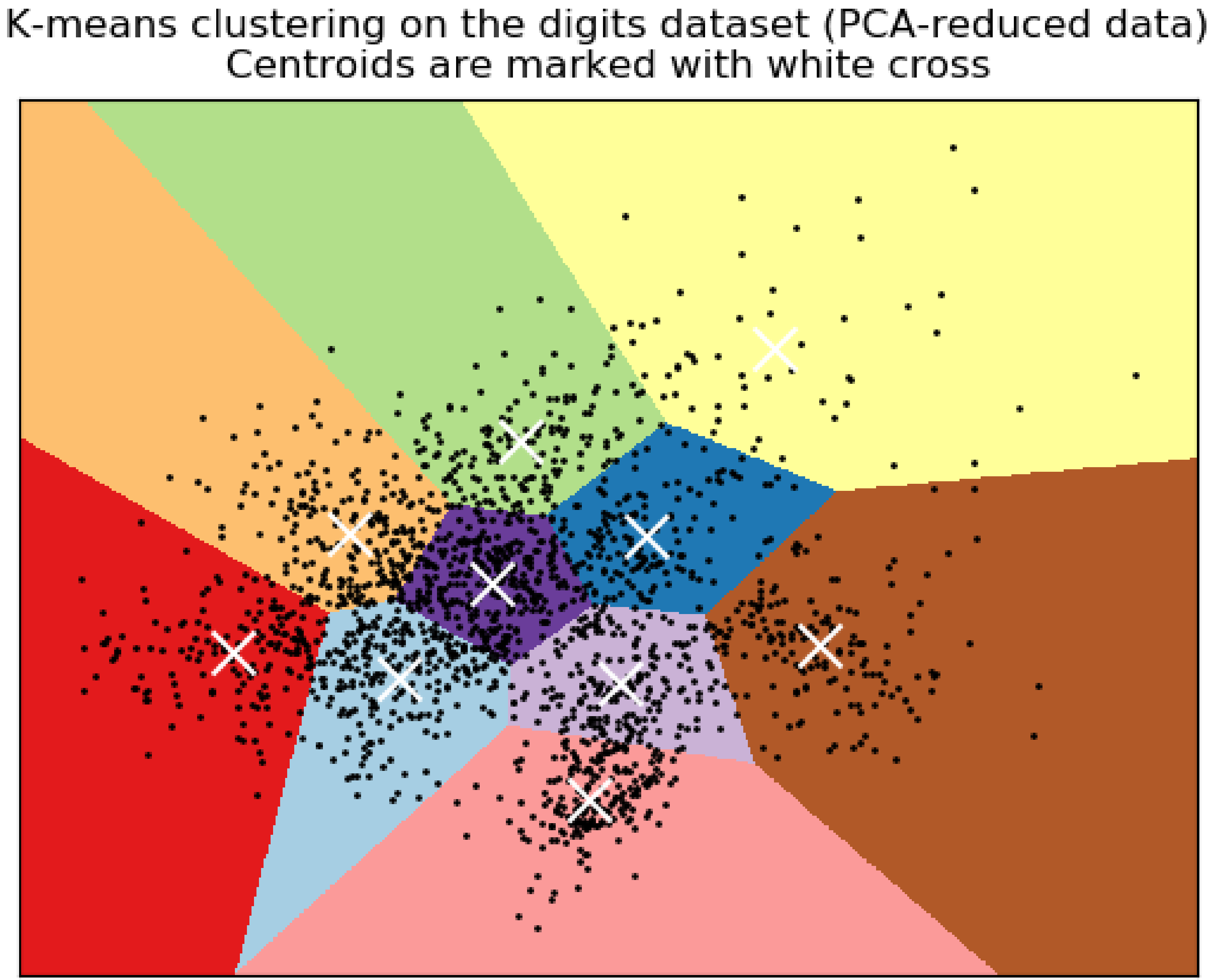
Clustering code

References:

- User's Guide http://scikit-learn.org/stable/modules/clustering.html
- Intro to ML Chapter 3: https://github.com/amueller/introduction_to_ml_with_python/blob/master/03-unsupervised-learning.ipynb
- Thoughtful ML: https://github.com/thoughtfulml/examples-in-python/tree/master/em-clustering

k Means Clustering

http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_digits.html



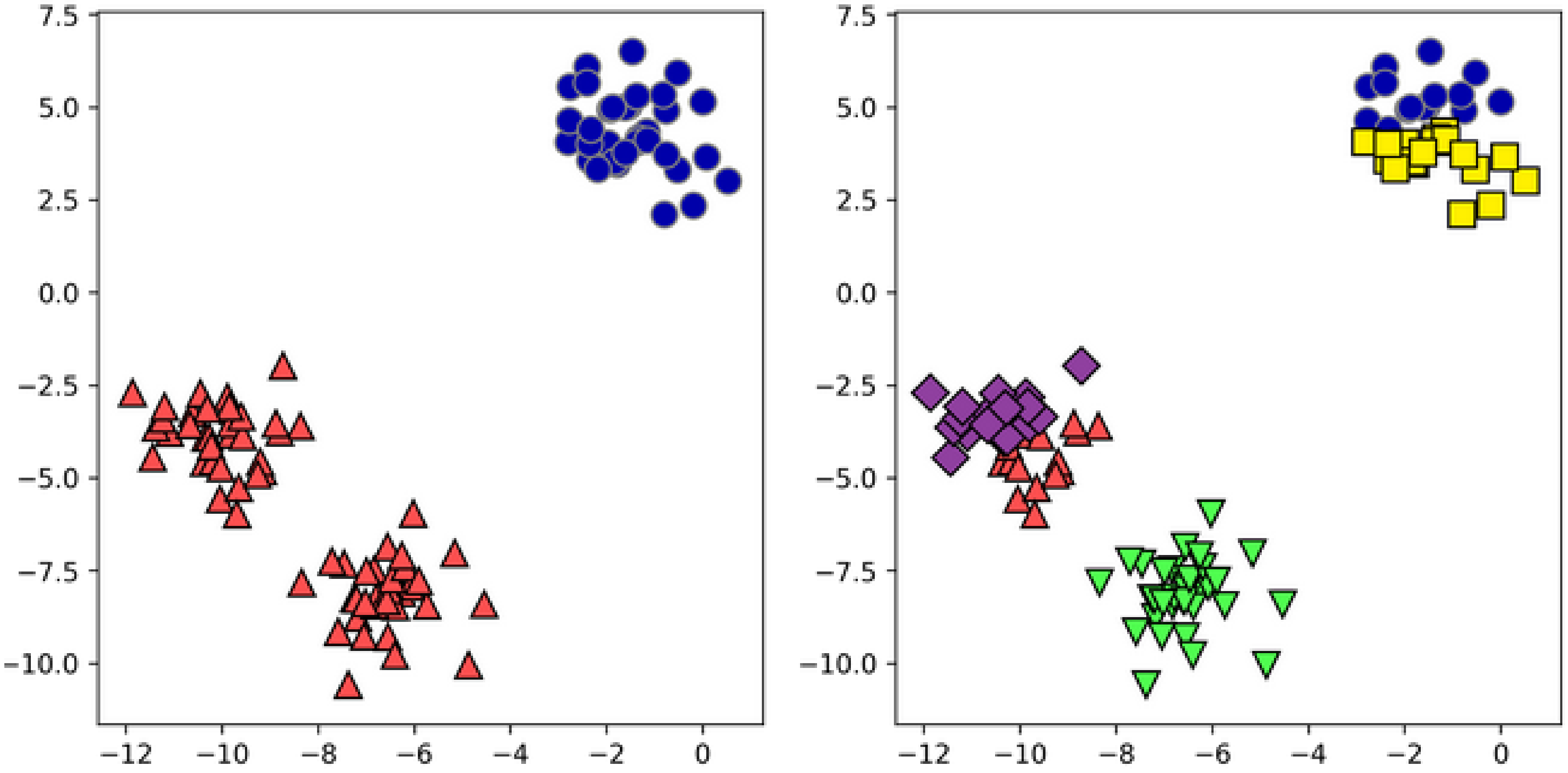
k Means Clustering

Algorithm:
choose k
randomly choose k data points for each cluster

    REPEAT
    find mean of each cluster
    assign every data point to nearest cluster
    UNTIL DONE

k Means Clustering

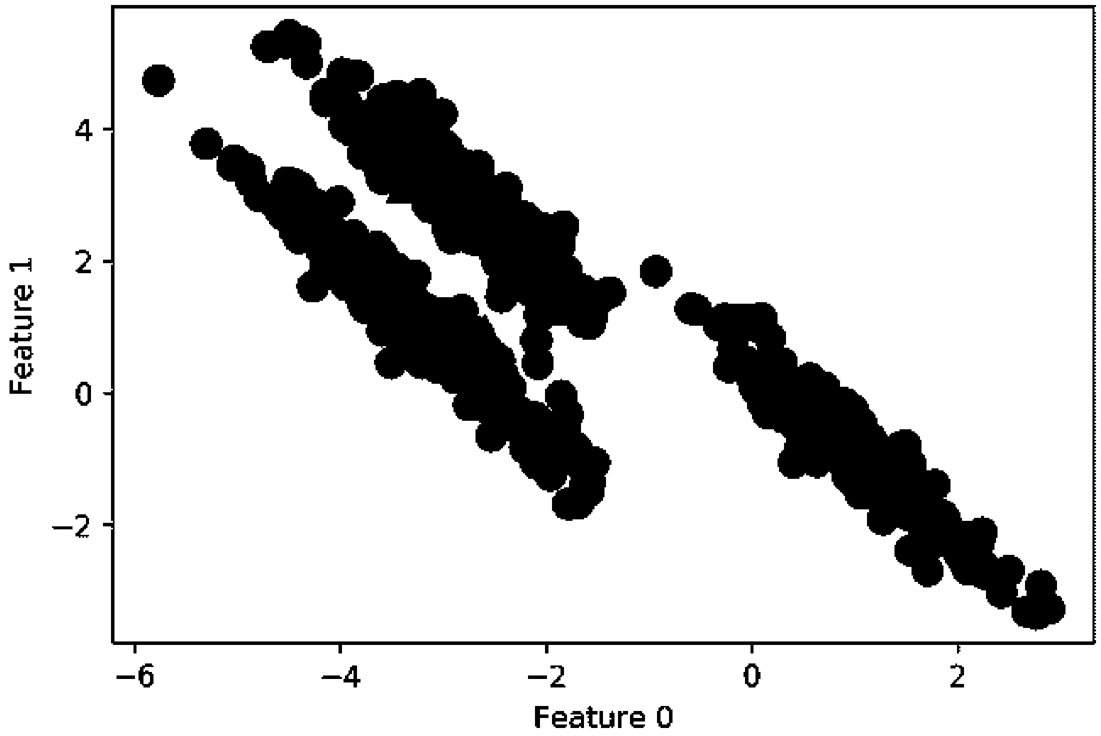
https://github.com/amueller/introduction_to_ml_with_python/blob/master/03-unsupervised-learning.ipynb

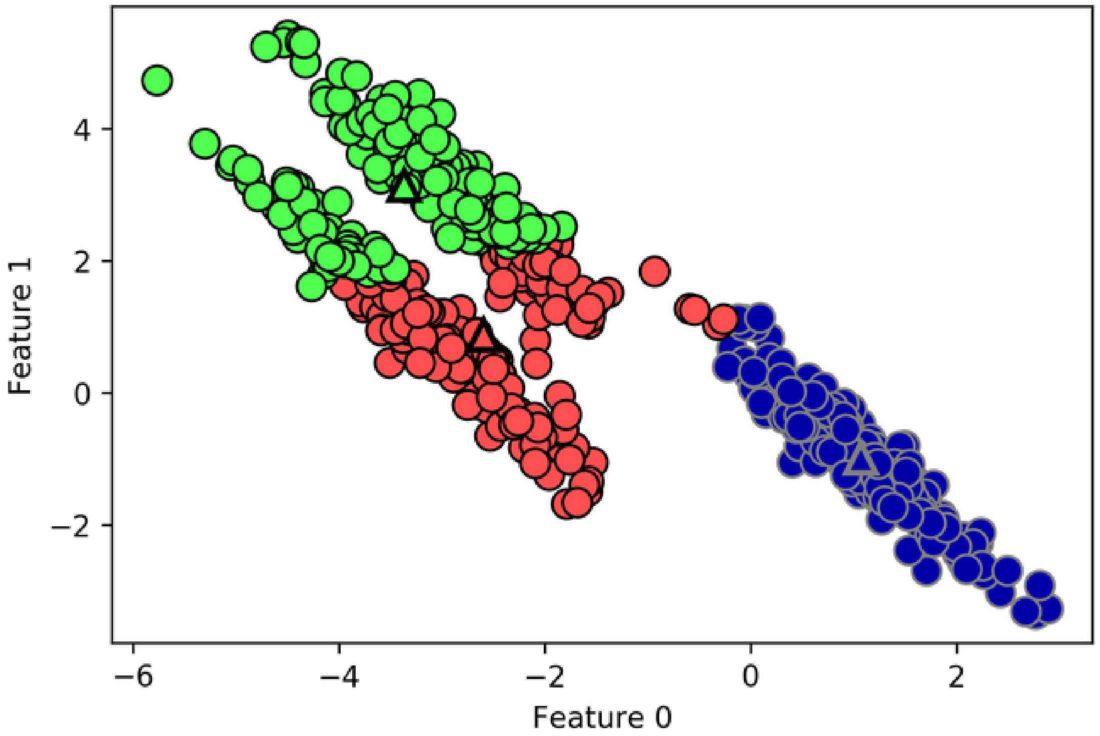


Wrong k values!

k Means Clustering

https://github.com/amueller/introduction_to_ml_with_python/blob/master/03-unsupervised-learning.ipynb

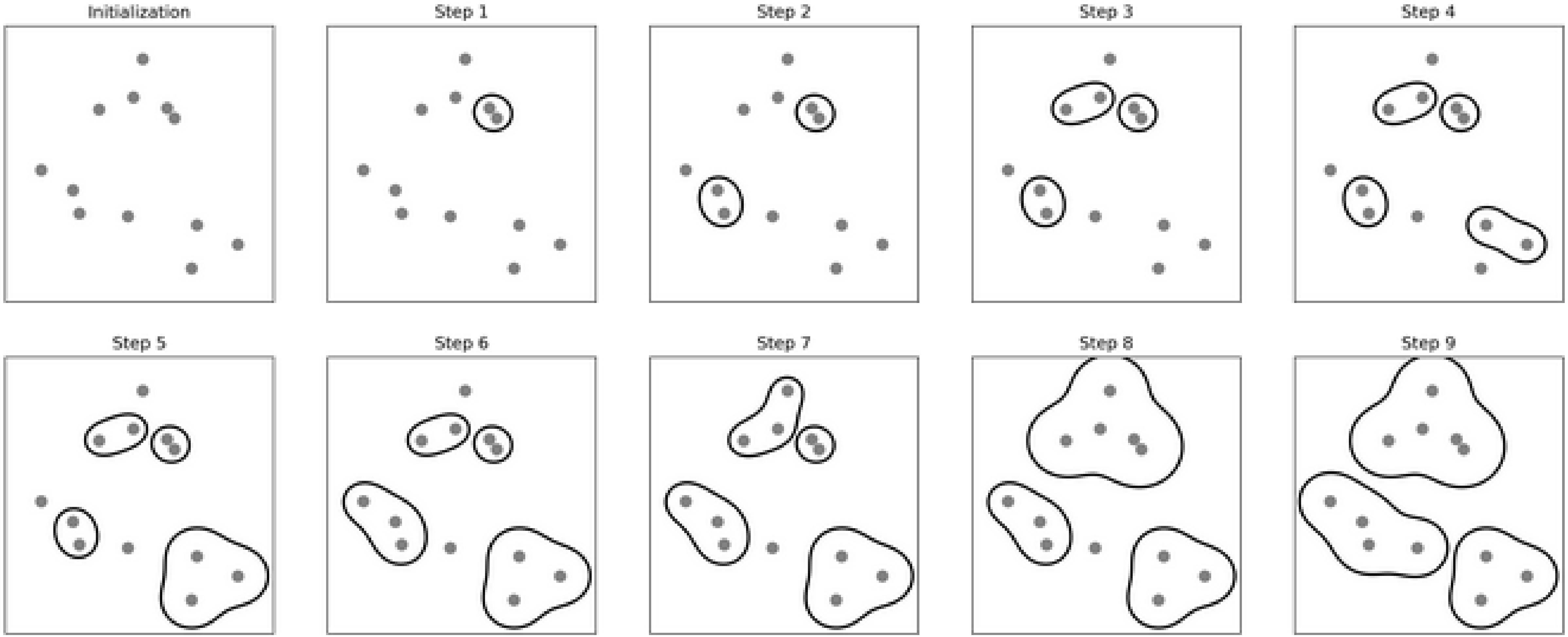




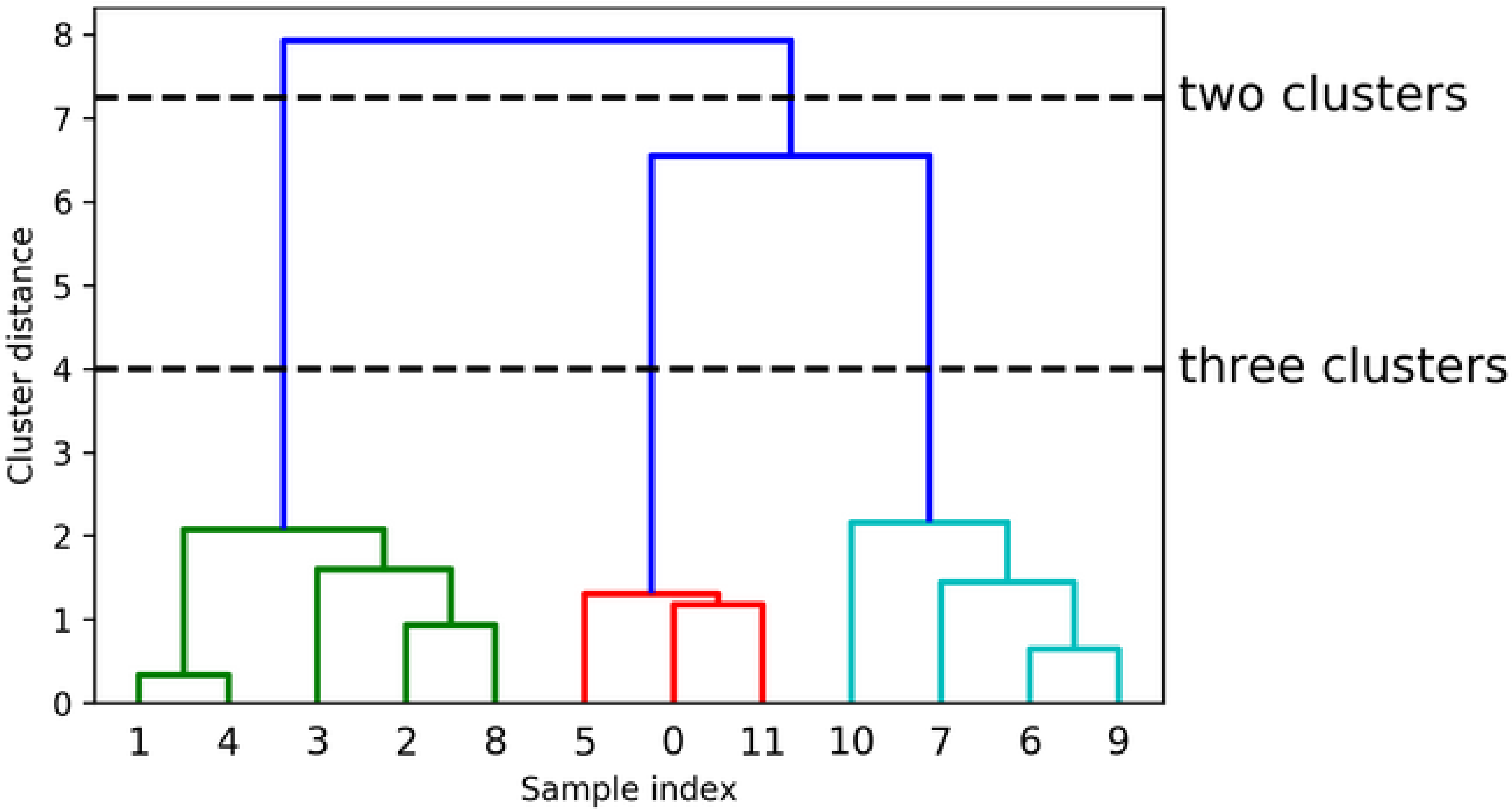
fails with non-spherical clusters

Agglomerative Clustering

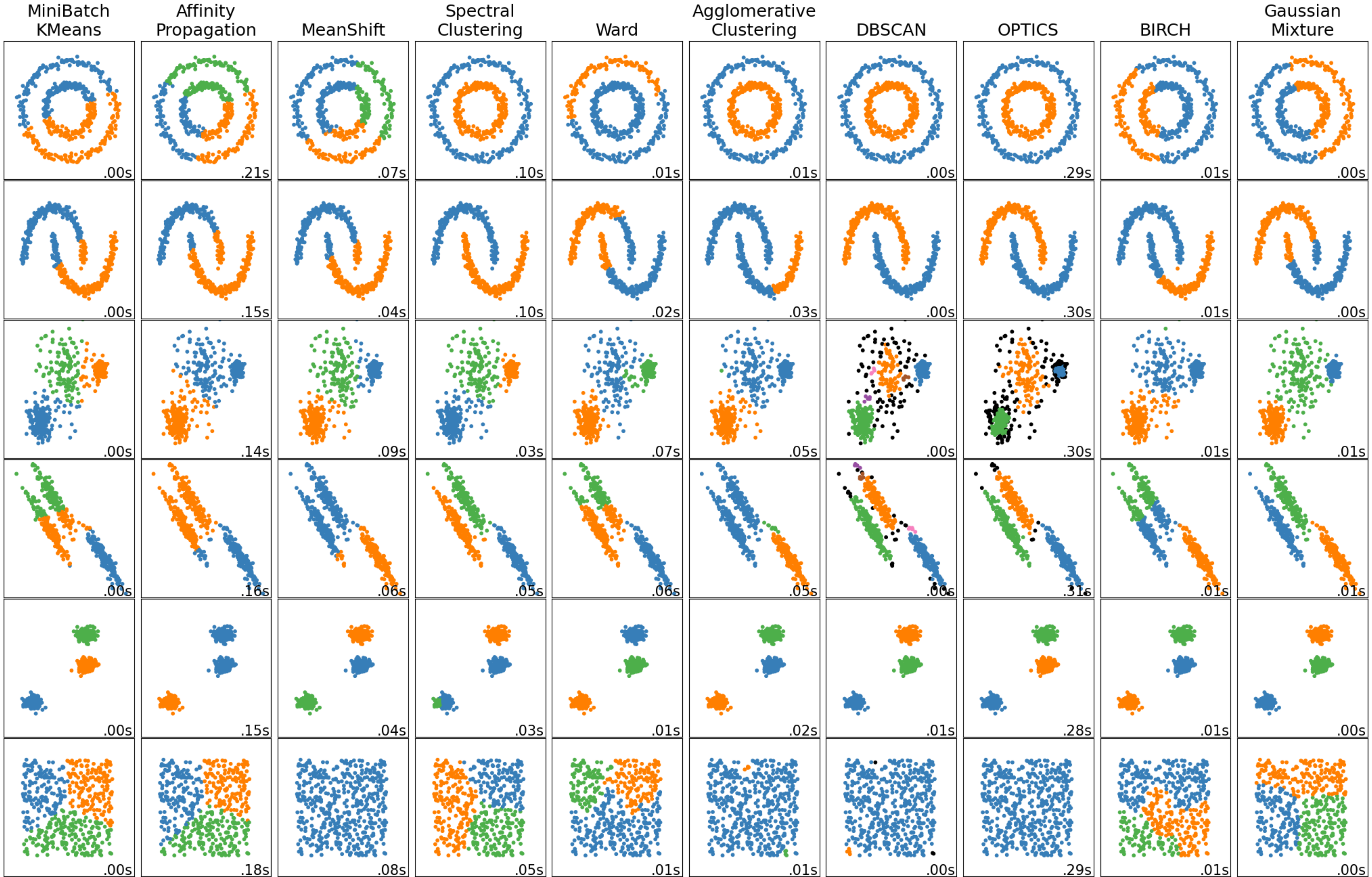
- Every point starts as its own cluster
- Merge nearby clusters until done.



Agglomerative Clustering



https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html



https://scikit-learn.org/stable/modules/clustering.html

Method name	Parameters	Scalability	Usecase	Geometry (metric used)
K-Means	number of clusters	Very large `n_samples`, medium `n_clusters` with MiniBatch code	General-purpose, even cluster size, flat geometry, not too many clusters, inductive	Distances between points
Affinity propagation	damping, sample preference	Not scalable with n_samples	Many clusters, uneven cluster size, non-flat geometry, inductive	Graph distance (e.g. nearest-neighbor graph)
Mean-shift	bandwidth	Not scalable with `n_samples`	Many clusters, uneven cluster size, non-flat geometry, inductive	Distances between points
Spectral clustering	number of clusters	Medium `n_samples`, small `n_clusters`	Few clusters, even cluster size, non-flat geometry, transductive	Graph distance (e.g. nearest-neighbor graph)
Ward hierarchical clustering	number of clusters or distance threshold	Large `n_samples` and `n_clusters`	Many clusters, possibly connectivity constraints, transductive	Distances between points
Agglomerative clustering	number of clusters or distance threshold, linkage type, distance	Large `n_samples` and `n_clusters`	Many clusters, possibly connectivity constraints, non Euclidean distances, transductive	Any pairwise distance
DBSCAN	neighborhood size	Very large `n_samples`, medium `n_clusters`	Non-flat geometry, uneven cluster sizes, outlier removal, transductive	Distances between nearest points
OPTICS	minimum cluster membership	Very large `n_samples`, large `n_clusters`	Non-flat geometry, uneven cluster sizes, variable cluster density, outlier removal, transductive	Distances between points
Gaussian mixtures	many	Not scalable	Flat geometry, good for density estimation, inductive	Mahalanobis distances to centers
BIRCH	branching factor, threshold, optional global clusterer.	Large `n_clusters` and `n_samples`	Large dataset, outlier removal, data reduction, inductive	Euclidean distance between points
Bisecting K-Means	number of clusters	Very large `n_samples`, medium `n_clusters`	General-purpose, even cluster size, flat geometry, no empty clusters, inductive, hierarchical	Distances between points

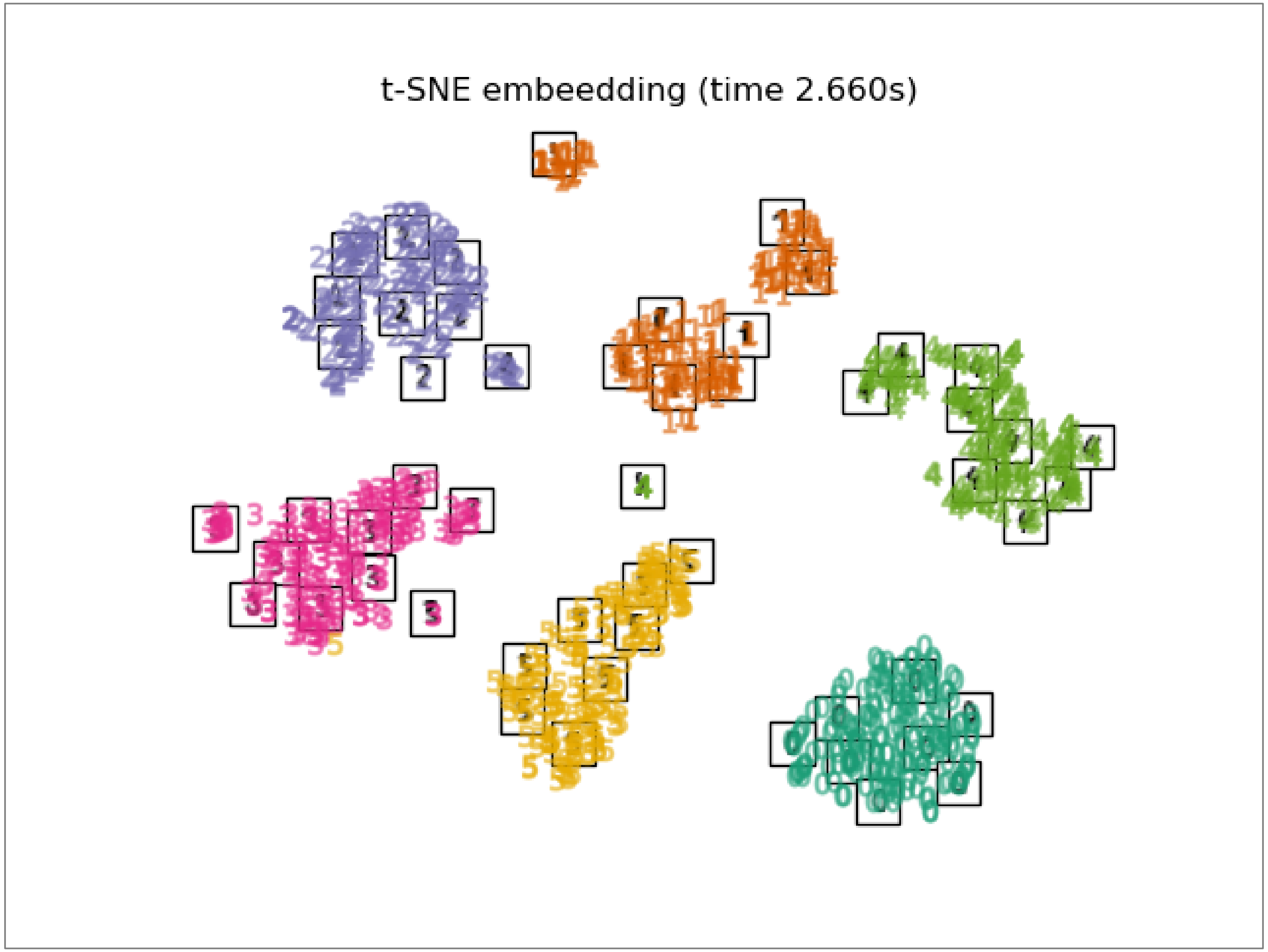
tSNE

t-distributed Stochastic Neighbor Embedding

Imagine we want to construct a 2D map of high-dimensional points. We can do this by attaching "springs" to each point in 2D space where the springs get smaller for points that are close neighbors

- https://distill.pub/2016/misread-tsne/
- https://www.youtube.com/watch?v=RJVL80Gg3lA&list=UUtXKDgv1AVoG88PLl8nGXmw (algorithm starts at 10:30)

https://scikit-learn.org/stable/auto_examples/manifold/plot_lle_digits.html



T-SNE on the digits dataset (only using digits 0-5) makes a 2D plot showing clusters of “similar” without knowing the right answers!

Stochastic Search

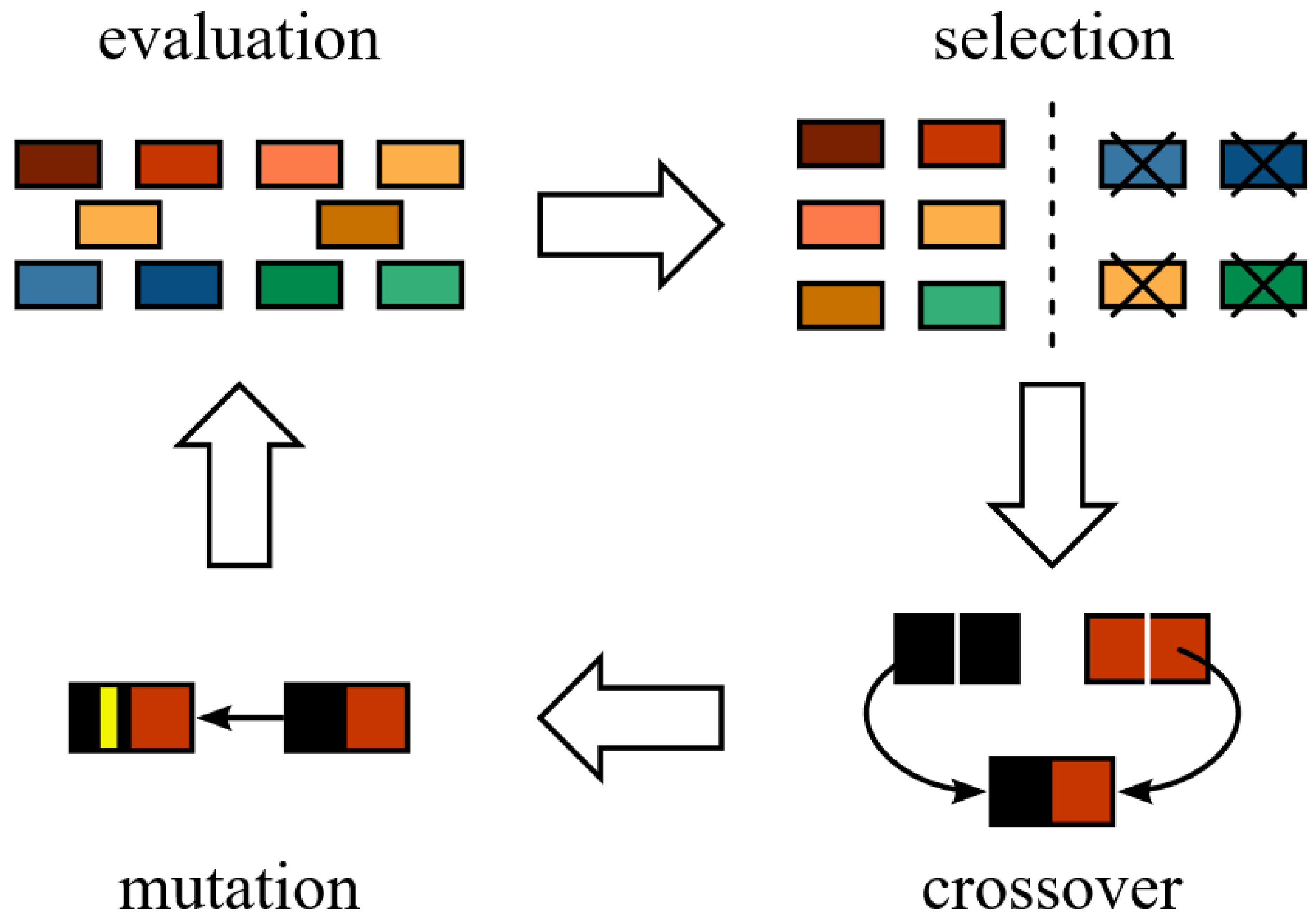
Problem #2

Can we do any training when we don't know the "right" answer?

If we have a given set of parameters (e.g. neural network weights, decision tree thresholds, etc.) we can try to find the optimal set that maximizes some metric

CGP Grey: How Machines Learn
https://www.youtube.com/watch?v=R9OHn5ZF4Uo

Genetic Breeding Algorithms



A neural network car with genetic weights (no backprop!)
https://www.youtube.com/watch?v=0Str0Rdkxxo

Genetic Breeding Algorithms

https://gplearn.readthedocs.io/en/stable/intro.html

Gplearn (a package that plays nicely with sklearn) represents arbitrary equations as trees:

$$y = X_0^2 - 3 \times X_1 + 0.5$$

This could be re-written as:

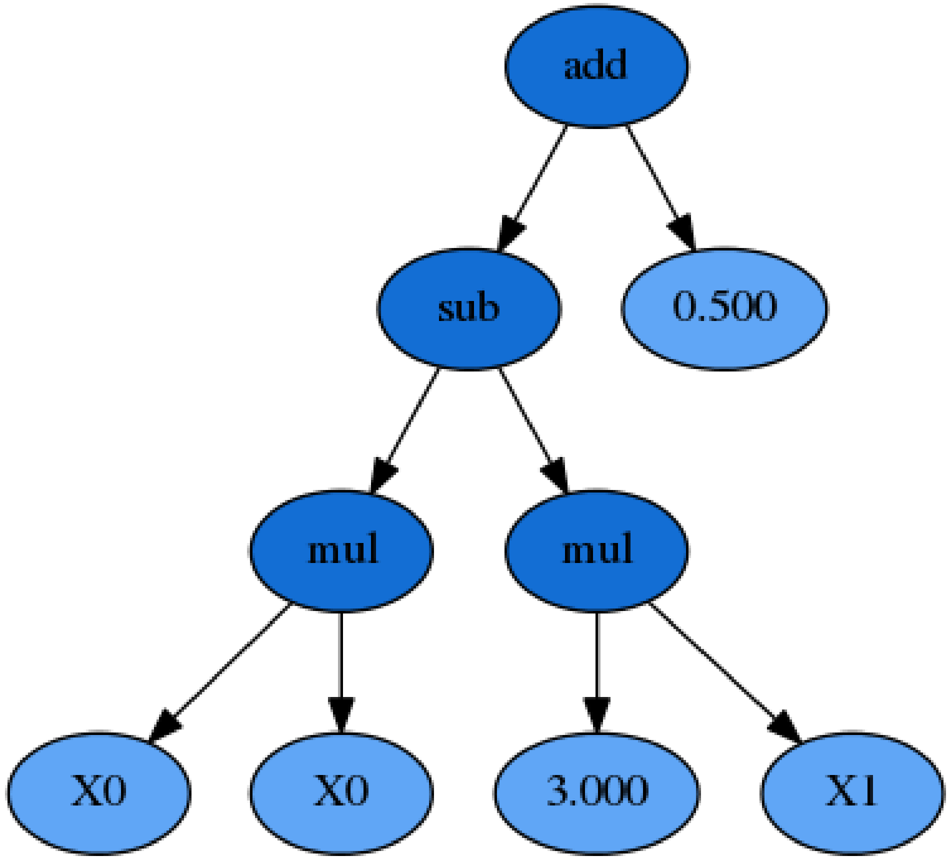
$$y = X_0 \times X_0 - 3 \times X_1 + 0.5$$

Or as a LISP symbolic expression (S-expression) representation which uses prefix-notation, and happens to be very common in GP, as:

$$y = (+(-(\times X_0 X_0)(\times 3 X_1))0.5)$$

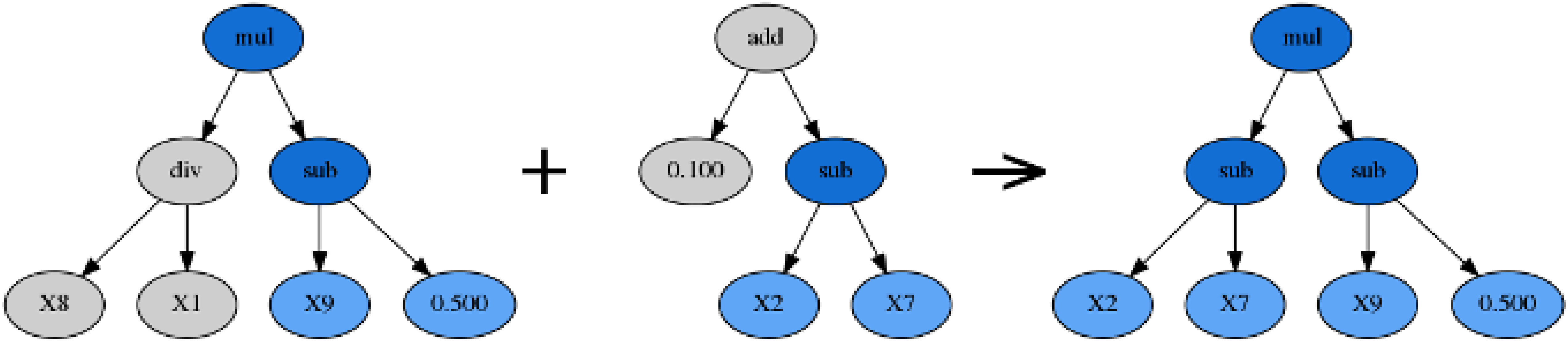
Or, since we're working in python here, let's express this as a numpy formula:

```
y = np.add(np.subtract(np.multiply(X0, X0), np.multiply(3., X1)), 0.5)
```



Genetic Breeding Algorithms

https://gplearn.readthedocs.io/en/stable/intro.html



Then it randomly evolves the trees to discover an equation that represents the data

Simulated Annealing

Another stochastic search uses the **METROPOLIS ALGORITHM** to look for an optimal state

AIP The Journal of Chemical Physics

HOME BROWSE INFO FOR AUTHORS COLLECTIONS

Home > The Journal of Chemical Physics > Volume 21, Issue 6 > 10.1063/1.1699114

Published Online: December 2004

Equation of State Calculations by Fast Computing Machines

The Journal of Chemical Physics **21**, 1087 (1953); https://doi.org/10.1063/1.1699114

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, *and* Augusta H. Teller
Los Alamos Scientific Laboratory, Los Alamos, New Mexico
Edward Teller
Department of Physics, University of Chicago, Chicago, Illinois

less

PDF ABSTRACT CITED BY TOOLS

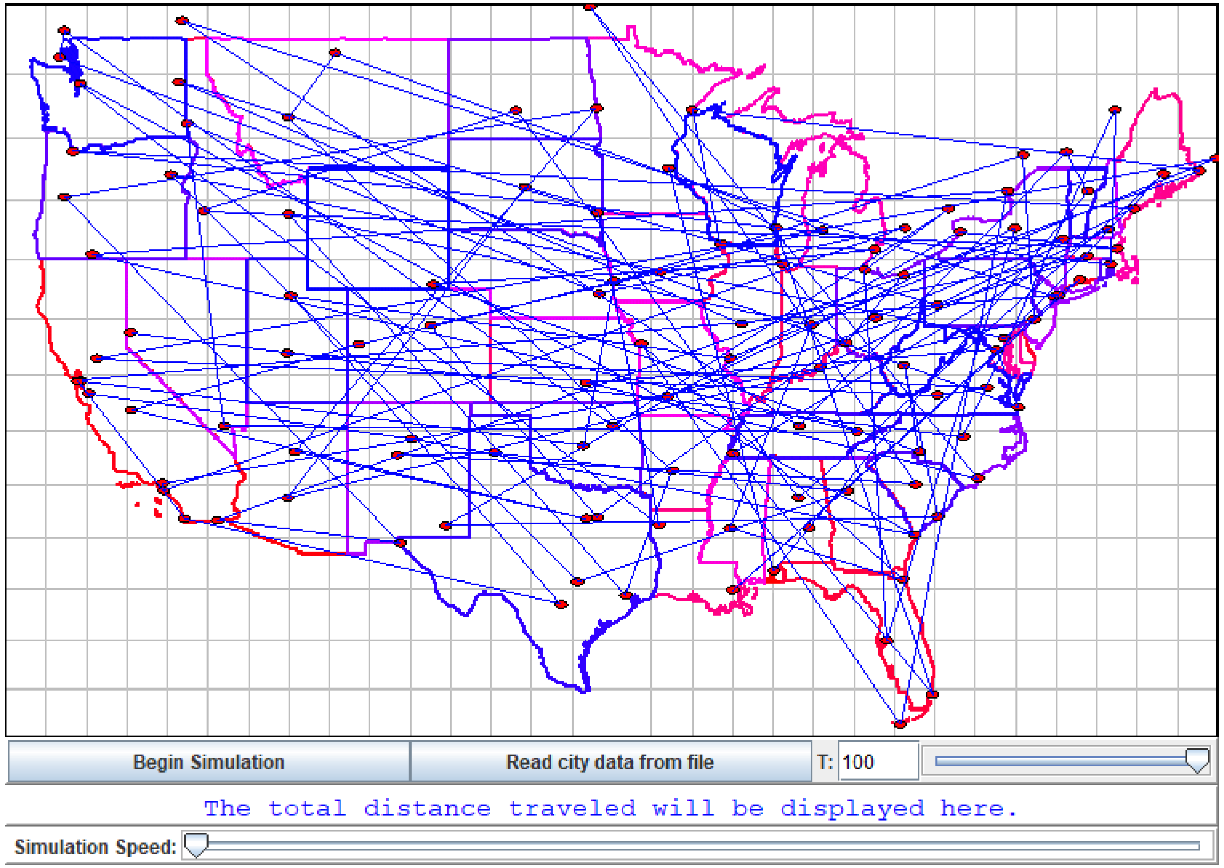
TOPICS
- Monte Carlo methods
- Equations of state
- Atomic and molecular interactions

ABSTRACT

A general method, suitable for fast computing machines, for investigating such properties as equations of state for substances consisting of interacting individual molecules is described. The method consists of a modified Monte

Simulated Annealing

https://www.compadre.org/osp/items/detail.cfm?ID=11538



#	City Name	Long. (deg)	Lat. (deg)
1	"Columbia, ...	-81.03333333	34
2	"Albuquerqu...	-106.65	35.08333333
3	"Amarillo, ...	-101.8333333	35.18333333
4	"Atlanta, G...	-84.38333333	33.75
5	"Austin, Te...	-97.73333333	30.26666667
6	"Baker, Ore. "	-117.8333333	44.78333333
7	"Baltimore,...	-76.63333333	39.3
8	"Bangor, Ma...	-68.78333333	44.8
9	"Birmingham...	-86.83333333	33.5
10	"Bismarck, ...	-100.7833333	46.8
11	"Boise, Ida...	-116.2166667	43.6
12	"Boston, Ma...	-71.08333333	42.35
13	"Buffalo, N...	-78.83333333	42.91666667
14	"Carlsbad, ...	-104.25	32.43333333
15	"Charleston...	-79.93333333	32.78333333
16	"Charleston...	-81.63333333	38.35
17	"Charlotte,...	-80.83333333	35.23333333
18	"Cheyenne, ...	-104.8666667	41.15
19	"Chicago, I...	-87.61666667	41.83333333
20	"Cincinnati...	-84.5	39.13333333
21	"Cleveland,...	-81.61666667	41.46666667
22	"Albany, N....	-73.75	42.66666667
23	"Columbus, ...	-83.01666667	40
24	"Dallas, Te...	-96.76666667	32.76666667
25	"Denver, Co...	-105	39.75
26	"Des Moines...	-93.61666667	41.58333333
27	"Detroit, M...	-83.05	42.33333333
28	"Dubuque, I...	-90.66666667	42.51666667
29	"Duluth, Mi...	-92.08333333	46.81666667
30	"Eastport, ...	-67	44.9
31	"El Centro,...	-115.55	32.63333333
32	"El Paso, T...	-106.4833333	31.76666667
33	"Eugene, Or...	-123.0833333	44.05
34	"Fargo, N.D. "	-96.8	46.86666667
35	"Flagstaff,...	-111.6833333	35.21666667
36	"Fort Worth...	-97.31666667	32.71666667
37	"Fresno, Ca...	-119.8	36.73333333
38	"Grand Junc...	-108.55	39.08333333
39	"Grand Rapi...	-85.66666667	42.96666667
40	"Havre, Mon...	-109.7166667	48.55
41	"Helena, Mo...	-112.0333333	46.58333333
42	"Hot Spring...	-93.05	34.51666667
43	"Houston, T...	-95.35	29.75