GEOG 432/832: Programming, Scripting, and Automation for GIS

Unit 13.01: Clustering

Dr. Bitterman

Today's schedule

- Open discussion
- Slides, discussion, and exercises
- For next class

Open discussion

Today's prep

Download *unit13_inclass.ipynb* and *unit13data.zip* from GitHub repository

Today's prep:

```
%matplotlib inline

import seaborn as sns
import pandas as pd
import geopandas as gpd
import numpy as np
import matplotlib.pyplot as plt
# new ones below
from sklearn import cluster # note the difference
import contextily as cx
```

Verify all packages are in your environment

...and that you're in the correct environment

...can you find them all???

Alternative sources

Not all packages are in the default channel (like *contextily*, depending on your version of Anaconda)

Let's add one

1. While selecting the correct environment, click on "Channels", then "Add..."
2. in the new box, type the URL of conda-forge: `https://conda.anaconda.org/conda-forge/`
3. Press "Enter" to add
4. Update index (or now, indices)
5. Add *contexily* to your environment

Today we'll be tackling clustering

What's a cluster???

Background

Real world is complex

- non-linear
- uncertain
- multivariate (as opposed to univariate...)

Tackling mutivariate data

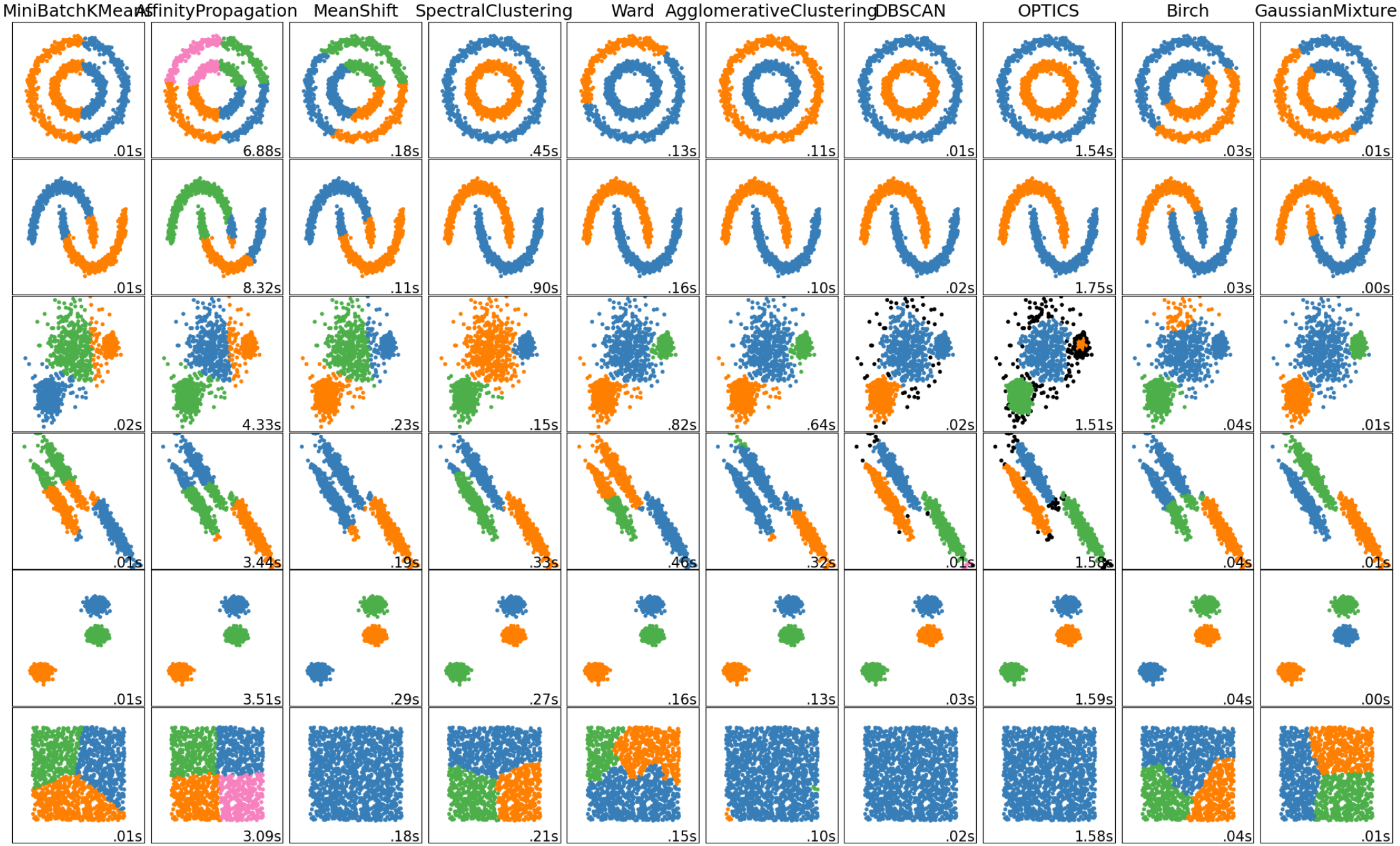
- Can be difficult to "look at" more than a few variables simultaneously
- Often necessary to *reduce dimensionality*
- Many techniques don't require preliminary assumptions about data structure
- Also useful as an exploratory tool

Statistical clustering

- **Basic idea:** summarize the multivariate data by creating a (relatively) small number of classes
- Each observation is assigned to one (and only one) class, depending on its position in multidimensional space
- MANY, MANY, MANY techniques to statistically group observations
- Common thread: define classes or categories of observations that are similar within each of them, but differ between groups

But implmentation details differ

- How "similarity" and "dissimilarity" are defined differs among algorithms
- Applications and utility depend on problem set


MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
AgglomerativeClustering
DBSCAN
OPTICS
Birch
GaussianMixture
.01s
6.88s
.18s
.45s
.13s
.11s
.01s
1.54s
.03s
.01s
.01s
8.32s
.11s
.90s
.16s
.10s
.02s
1.75s
.03s
.00s
.02s
4.33s
.23s
.15s
.82s
.64s
.02s
1.51s
.04s
.01s
.01s
3.44s
.19s
.33s
.46s
.32s
.01s
1.58s
.04s
.01s
.01s
3.51s
.29s
.27s
.16s
.13s
.03s
1.59s
.04s
.00s
.01s
3.09s
.18s
.21s
.15s
.10s
.02s
1.58s
.04s
.01s

K-means

- A (very) common statistical clustering technique
- anyone familiar???

Whiteboard demo

(also: https://stanford.edu/class/engr108/visualizations/kmeans/kmeans.html)

What are the big assumptions?

(and/or required parameters)

Interactive session (Jupyter notebook)

For next class

- Lab 7 due Thursday
- Lab 8 starts Friday
- Readings are linked/posted on Canvas...