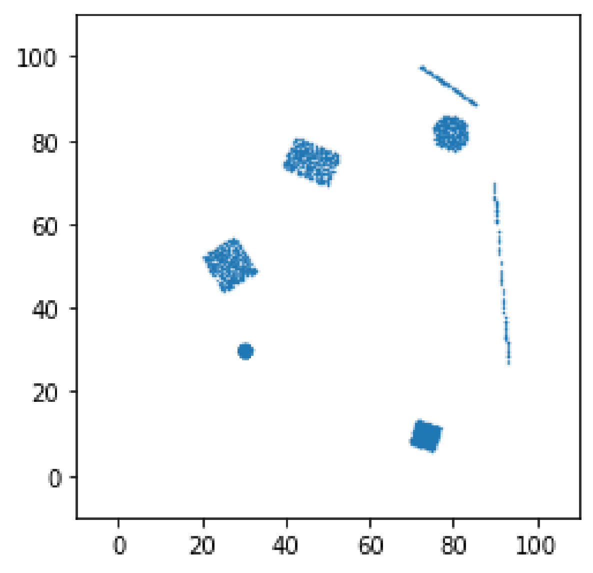
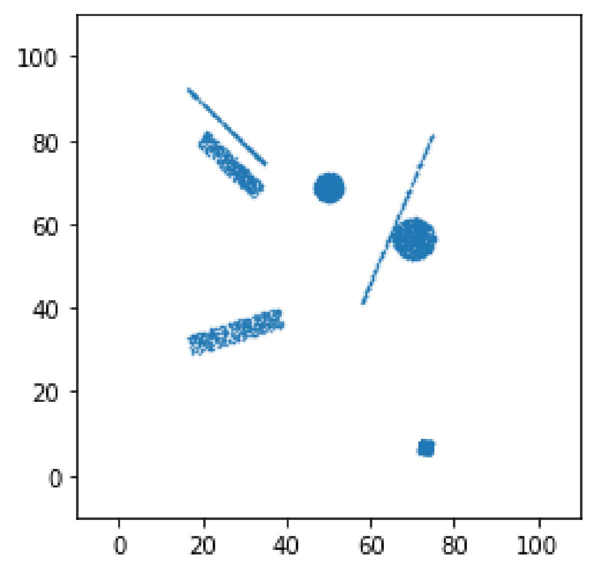
Assignment #4

Q: Using Spark, what features can we discover in a 6 dimensional data set?

In `~urbanic/LargeScaleComputing/Spark/space.dat` on Bridges you will find the data as a list of 6 dimensional coordinates within a cube of side length 100. *I can change this dataset to another variation at any time, so grab your own copy once!*

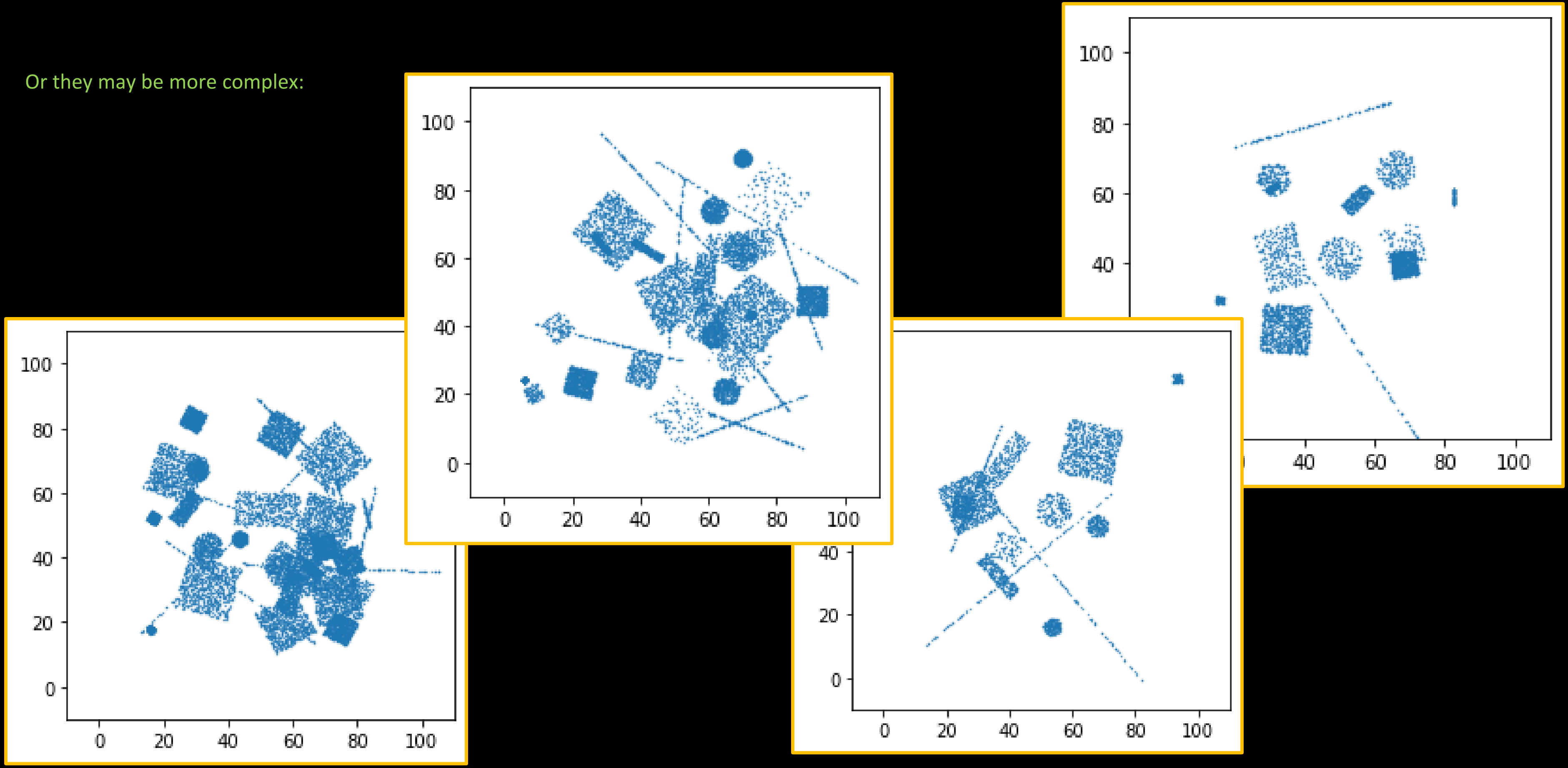
It contains some interesting distributions. They may be simple and few:





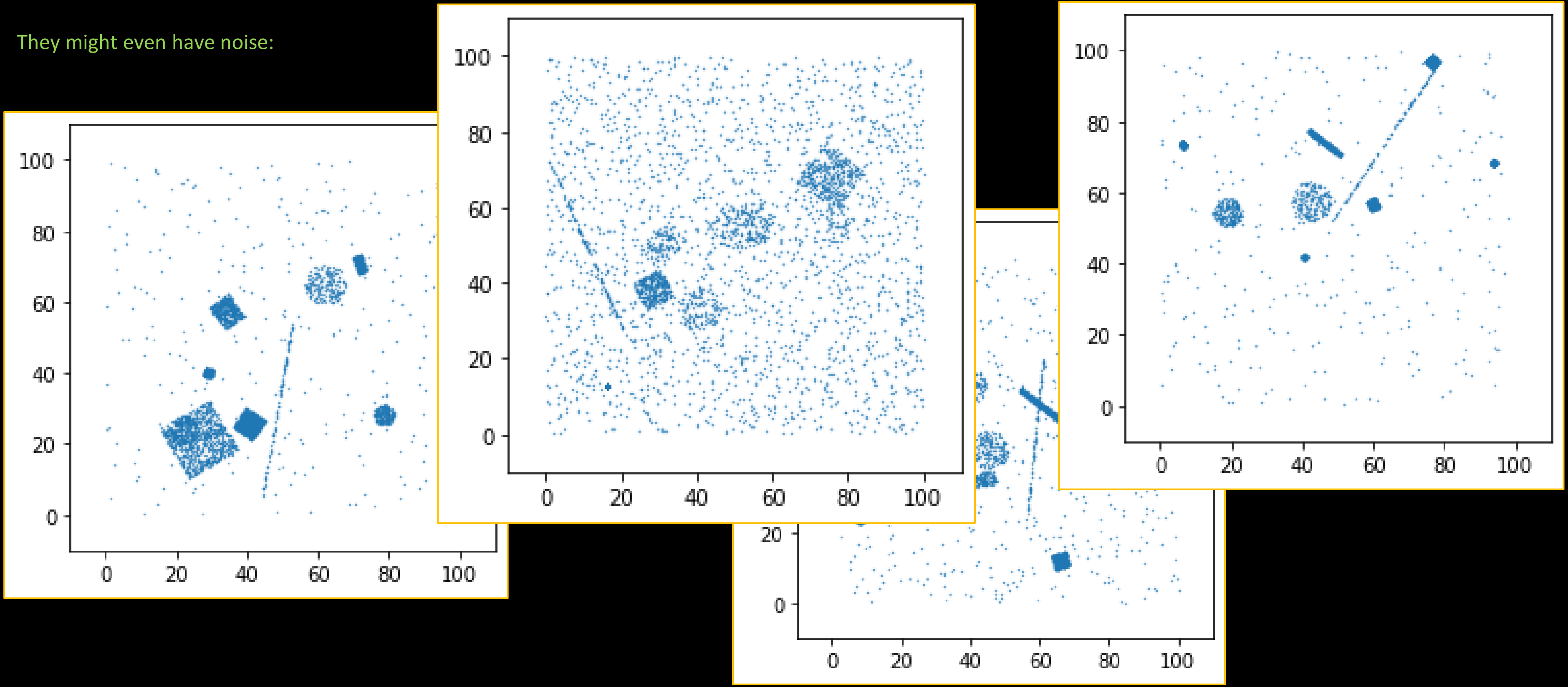
Homework #7:

Or they may be more complex:



Homework #7:

They might even have noise:



Homework #4

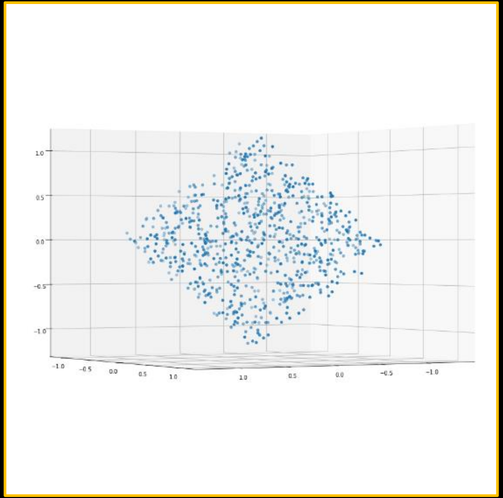
In 6 dimensions you will not be able to simply plot your way through this.

The distributions (objects) may each be any dimension up to 6. You may find lines, and you may find hypercubes or 6-balls.

They may be rotated in any direction. Make sure you appreciate this.

The good news is that the higher dimensionality lowers the chance that these objects will overlap significantly.

You can "extract" the 1 and 2 and 3 dimensional objects to plot:



Homework #4

You should do all of this within Spark. This dataset may or may not be too big for other resources you have access to, but I could certainly create one that would be intractable without a large cluster. So work as though it is.

You will be expected to navigate the Spark documentation yourself. It is complete and comprehensive. There are also plenty of other online Spark resources.

Here are a few hints:

- We have discussed that the Spark DataFrame and RDD APIs are not always the same. Some provide functionality that the other version does not. You should be prepared to switch between them. For example, if one version of PCA does not provide the capability that you require, look at the other.

- DataFrames will require that you pay close attention to the datatypes required. You may find yourself having to use some of the more formal machine learning types here. This may seem like painful meticulousness, but these types and terminology are common to machine learning everywhere, so this is generally useful experience.

- There is a 2D dataset in the directory called `simple_2D_example.dat` that you can use to test your methods. I have put an image of that dataset there for your convenience, `simple_2D_example.png` .

Homework #4

- When you save data using the convenient saveAsTextFile method with RDDs, you may find that your data has ended up not in the single file you hoped, but in several pieces in a directory. This is actually desirable behavior for enormous distributed files, but a little annoying here. If you use the rdd.repartition(1) transform on the RDD before you save it, it will keep your file in one piece (although still named something like "part-00000" in the named subdirectory).

- If you are using PCA to analyze the structure of data, it helps to normalize the data about 0. In the case of our spatial structures, you may want to recenter them to the origin before using any kind of PCA.

- As some of you are new to PCA, I will suggest that you will find more useful information applying it to any interesting objects you find _separately_. If you apply it to the whole dataset, you get results for all the pieces lumped together. And you may want to consider my previous hint with this as well.

Homework #4

This is a new, and hopefully exciting, exercise. It is possible to extract all of the meaningful structure using the tools you have. Perhaps some of you will. We will certainly grade on a kind curve, and be flexible about what you report.

- You should submit a summary of what you have found. It could be a short paragraph, or a brief table, maybe even some matplotlib. Perhaps the answer is "There was one 2D square, centered at 20,20,20,20,20,20 with edge length 5", and that is all you have to say. Or maybe you found:

Object	Location	Size	Orientation	Points
6D Sphere	12,12,23,12,11,16	6	NA	155
Square	23,33,12,78,55,33	22	?	2000

- You must also include the Spark script that you used. Use RDD's. Do not write an unscalable generic python algorithm.

- It should be one script that I can run myself. It should read in the datafile and spit out all of the data that you used for your summary, like a serious production code. Inline comments are welcome.

You may ask questions at any point before this is due. I will not give you spoilers, but I will clarify anything I can, and give you general suggestions. If I feel the need to disclose anything particularly useful, I will make it an Announcement.

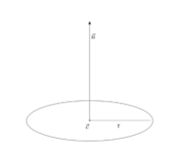
Homework #4

Orientation?

This is an "optional" detail that you might explore with the objects you find. It is not essential. But, as scientists you should at least be thinking about how this concept applies here.

In 2D, you might specify the orientation of a square by the angle of one edge to some axis. Does that work for a disc? So, does that mean that a disc has no orientation?

How about in 3D. Does a disc have an orientation now?



Of the various methods that can be used to describe 3D orientations (Euler angles, quaternions, or my favorite: rotors from geometric algebra), we might pick a normal to a surface (and maybe a rotation about that normal, if necessary, for something like a square).

Does this work for a 3D ball (which we call a "3-ball" in this context, and it's surface is a "2-sphere"; our disc is a "2-ball")?

How about our 3-ball in 4 dimensions? Does it have an orientation now?

Think about how these concepts generalize to higher dimensions. Once you get it, it all seems pretty elegant!

Homework #7

This is due anytime before we review answers in class on November 12th. Once we review, I can give no credit.

This is 15 points of your final grade.

The 15% of the final grade, and the due date, reflect the effort expected. This is not a good one to wait until the night before.

The open-ended and exploratory nature of this assignment is intended to be interesting. However, I recognize that is isn't fair to not know when you are "done". So I will bend my rule of determining grading only after I have reviewed the submissions to include this benchmark:

13 Distinguish *all* objects by some *structural* difference between them. Can be anything.

14 Describe all objects meaningfully. What are their shapes, etc.

15 TBD. There are some easter-eggs here.

There is all kinds of extra credit you might find, for example by noticing the orientation of some shape. So, even if you miss something you might make up for it by finding something else.