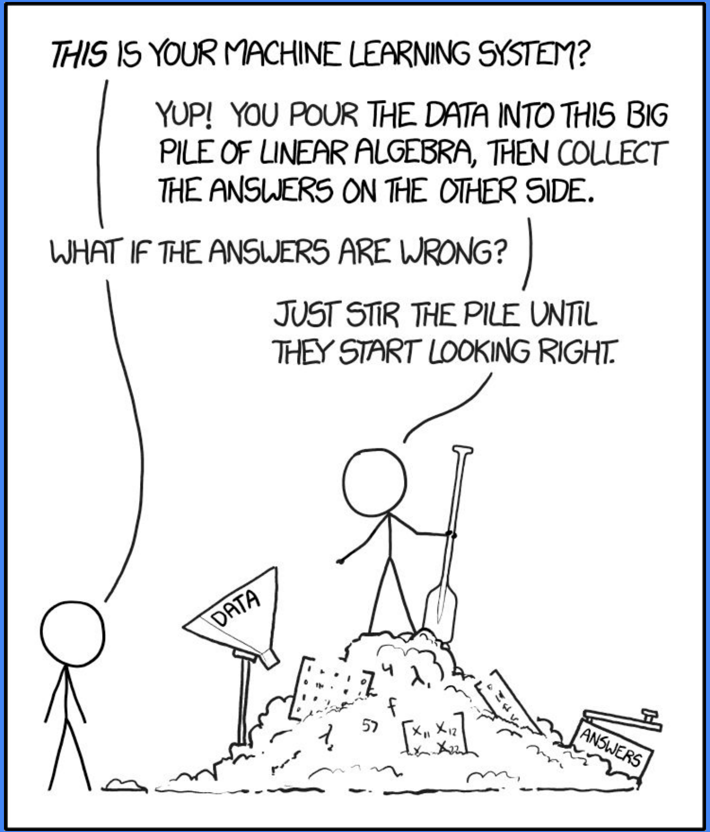
Data Science Seminar

Developing a toolkit to succeed in data science



Society of Physics Students (SPS)

Data Science Jargon

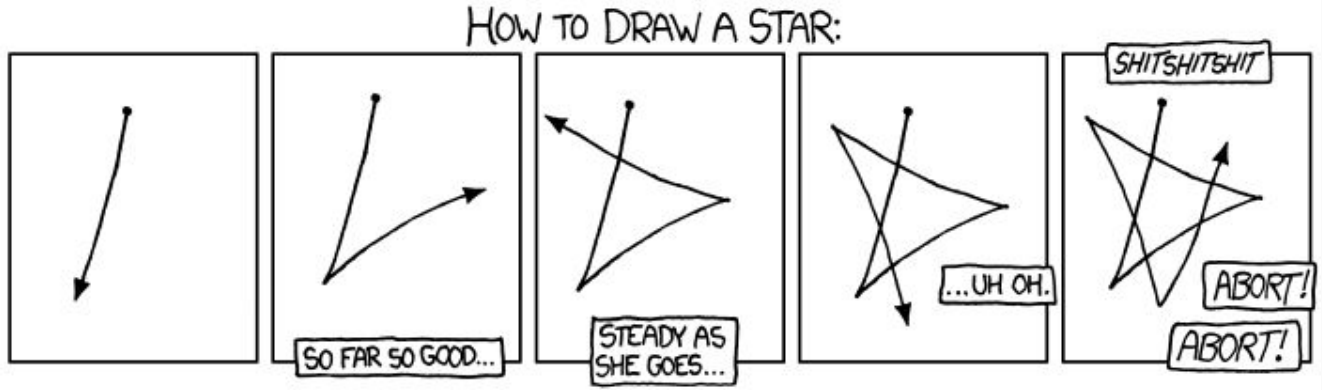
- Describe images (or features) by a “measurement”
  - Predefined function of the data

In statistics,

- **Sample statistic**: pre-defined function evaluated on a sample
  - Regarded as an **estimator** of a parameter of:
    - the underlying model
    - parameterization of the population from which samples are drawn
  - Also used for a **hypothesis test** about the sample.

Example

If we have a noisy image of star, *I*, we can form the statistic:



$$F = \sum_{ij} I_{ij}$$

- **Estimator** of the true flux of the star, **ν**
- *F* can be used to test how likely it is that *I* is the image of a star with true flux **ν**

Goals of Data Science

- Data science happens at the level of sample statistics
  - Jargon: at the catalog level
  - I.e. once the measurement is made.
- It does NOT deal with defining the statistics, instead:
  - Characterize the underlying populations
    - Based on multiple statistics of typically large number of samples

Two large groups:

- Supervised learning
- Unsupervised learning

Analysis Approaches | Supervised Learning

Supervised Learning

- Techniques are based on data, for which
    - Independant and dependant variables/features are known
    - I.e. there are samples (x_i, y_i) of the mapping $X \rightarrow Y$

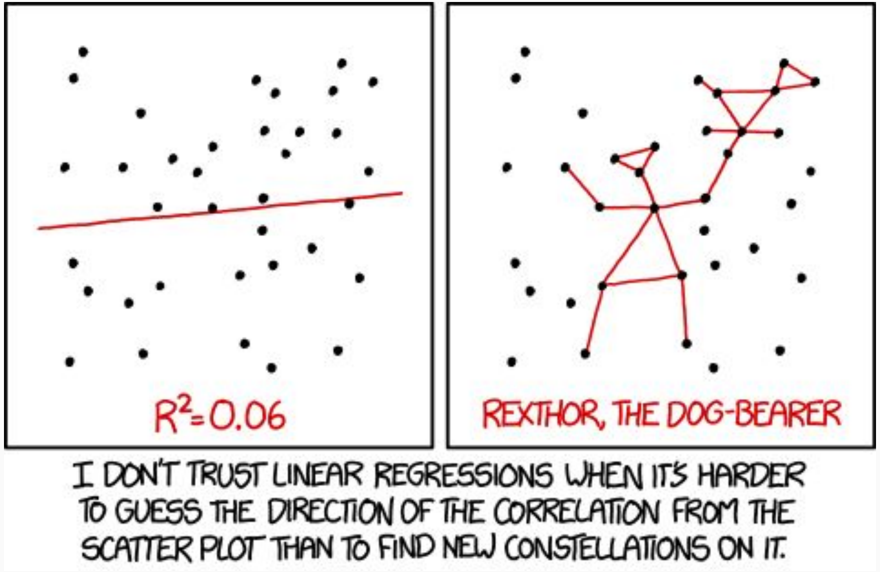
The task is to learn the mapping. Again, two types arise:

- Regression
- Classification

Supervised Learning | Regression

Y is the space of continuous variables, typically $\mathbb{R}$ or $\mathbb{R}^d$

- Colloquially known as "fitting a model to data"
- Task: **Regression**



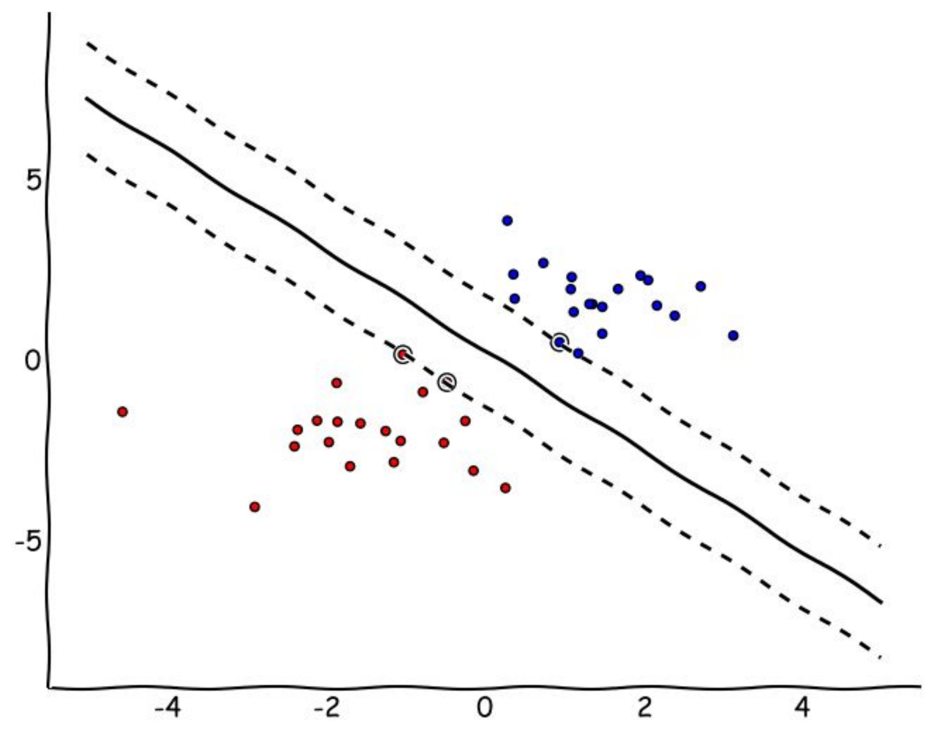
Supervised Learning | Classification

$$Y = \{L_1, ..., L_K\}$$

Is a finite set of class labels

where $x_i \sim X$ is assumed to belong to one class

Task: **Classification**

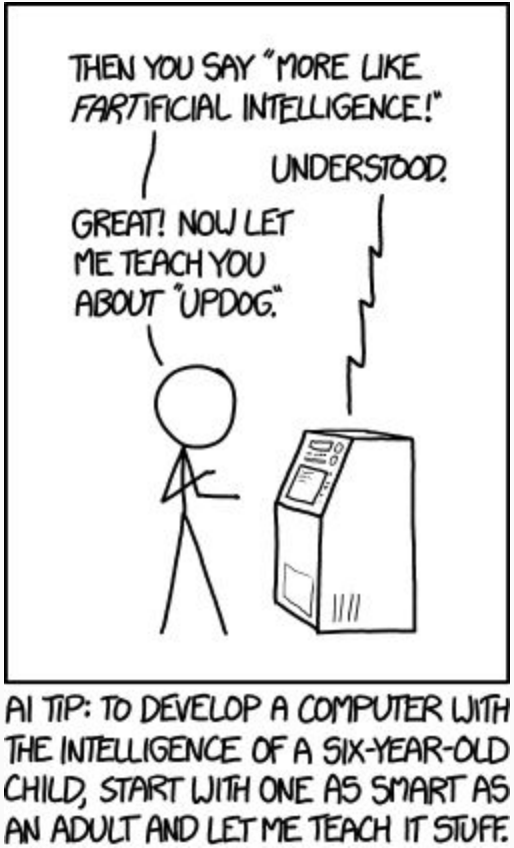


Analysis Approaches | Unsupervised Learning

For which only $x_i \sim X$ are available.

Goals are to describe their:

- distribution in X (density estimation)
- Labeling (clustering)

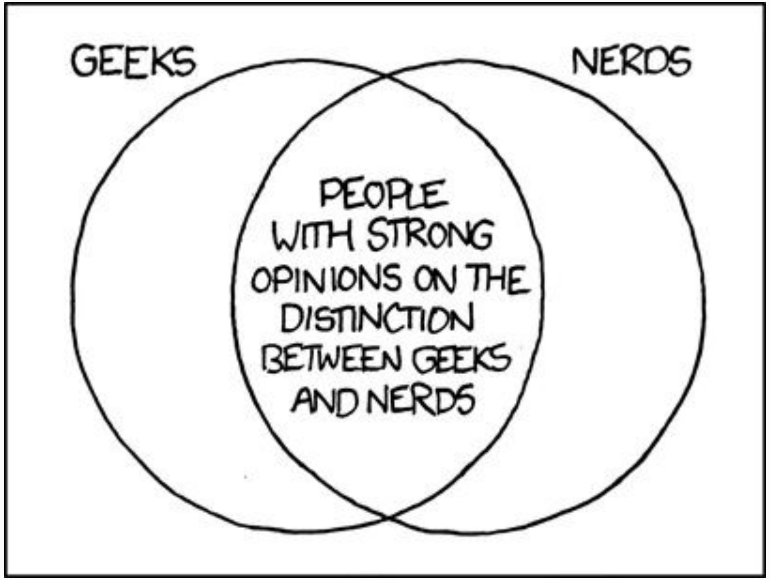


Analysis Approaches | Unsupervised Learning

Tasks are essentially equivalent to the supervised learning ones (regression, classification), except:

- The **output** of the mapping is **never observed**.

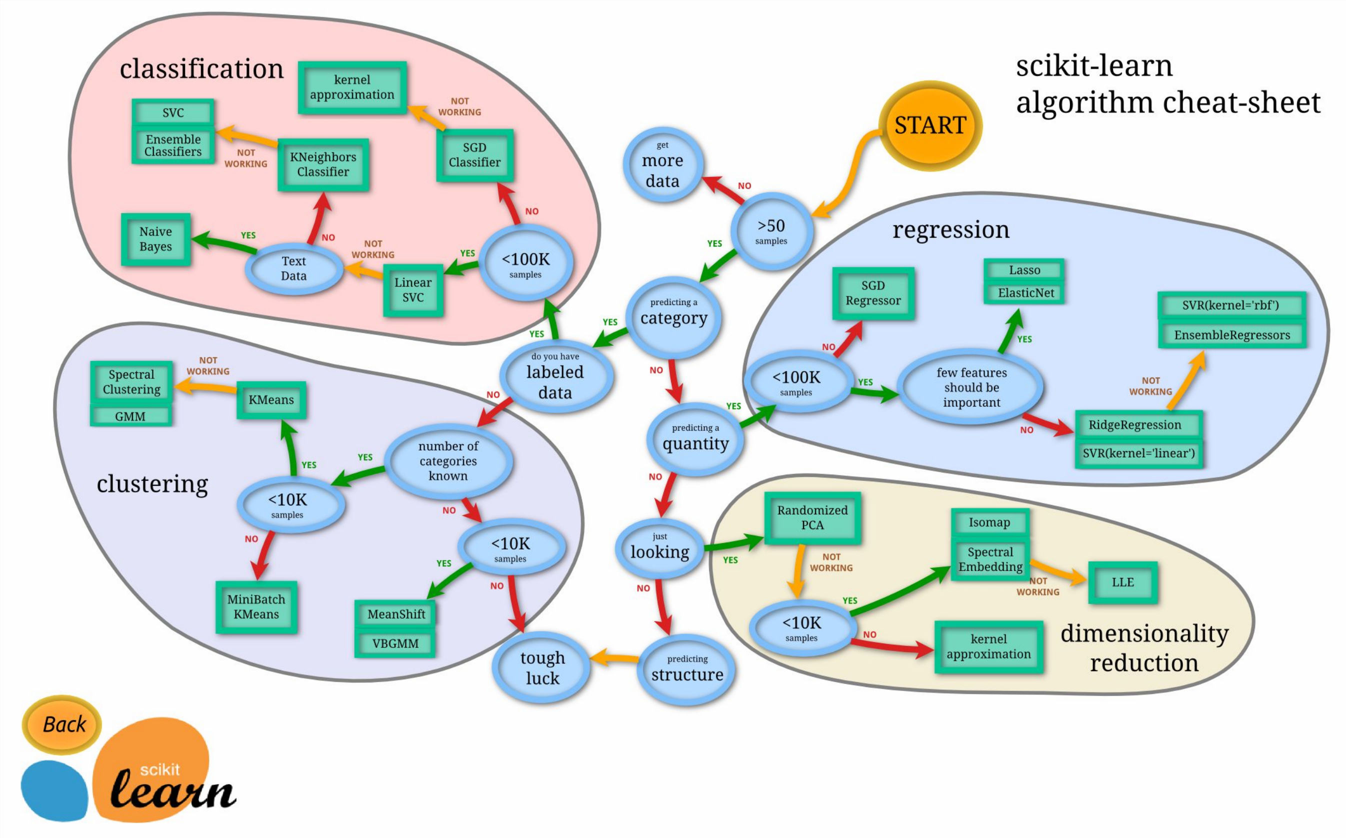
For example, clustering seeks to divide populations into similar groups:



Machine Learning in Python w/ Scikit-Learn

- Single python package that implements all of the above.
    - Some specific methods are missing
- Data can be handled as numpy arrays and pandas DataFrames.
- Powerful for smaller analyses and fast exploration.
- Very well documented with a neat tutorial.

Scikit-Learn Algorithm Cheat Sheet



Introductory Example

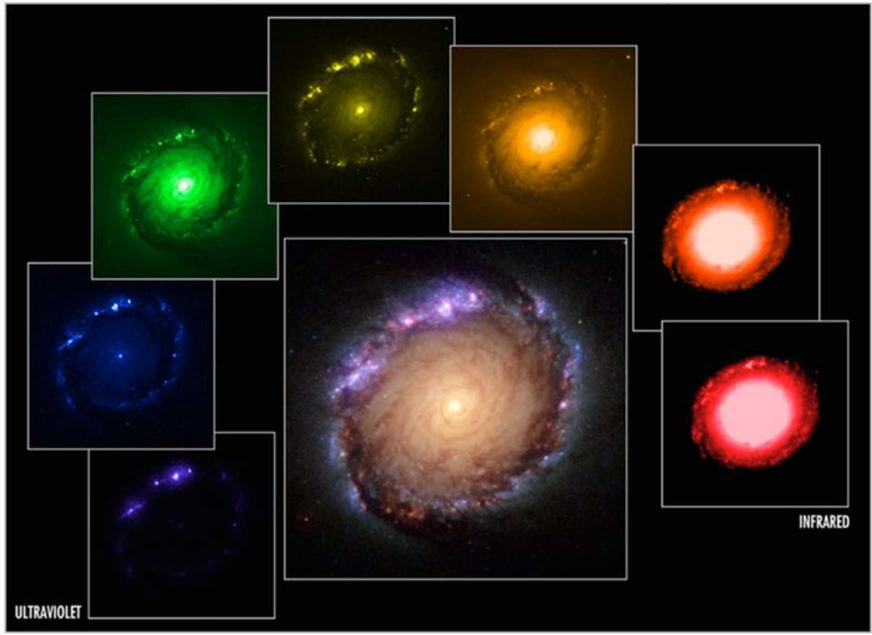
Multi-band detection by pixel-level clustering

Disclaimer: This is **not** a thorough application, more a quick test if the idea has merit. And we'll do some serious harm to error propagation…

IT'S IMPORTANT TO KNOW THE INTERNATIONAL WARNING SYMBOL FOR RADIOACTIVE HIGH-VOLTAGE LASER-EMITTING BIOHAZARDS THAT COAT THE FLOOR AND MAKE IT SLIPPERY.

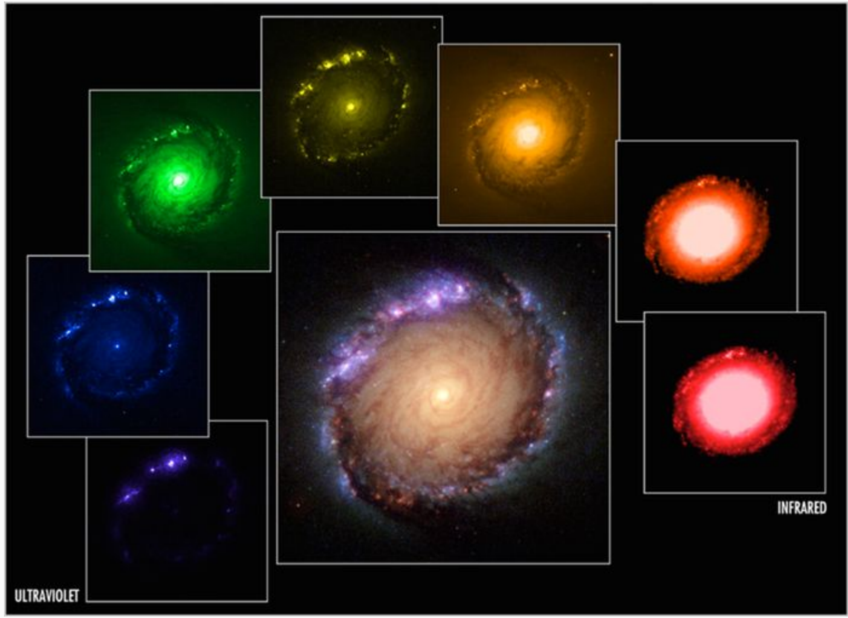
Multi-Band Detection by Pixel-Level Clustering

- Traditionally, astronomical image analysis is performed in individual "bands"
  - I.e. "red" band
  - Detect objects, determine properties, etc.
- Multiple bands
  - I.e. different "colors"
  - Repeat the process for each band
- Weird… we naturally expect images to have colors.



Multi-Band Detection by Pixel-Level Clustering

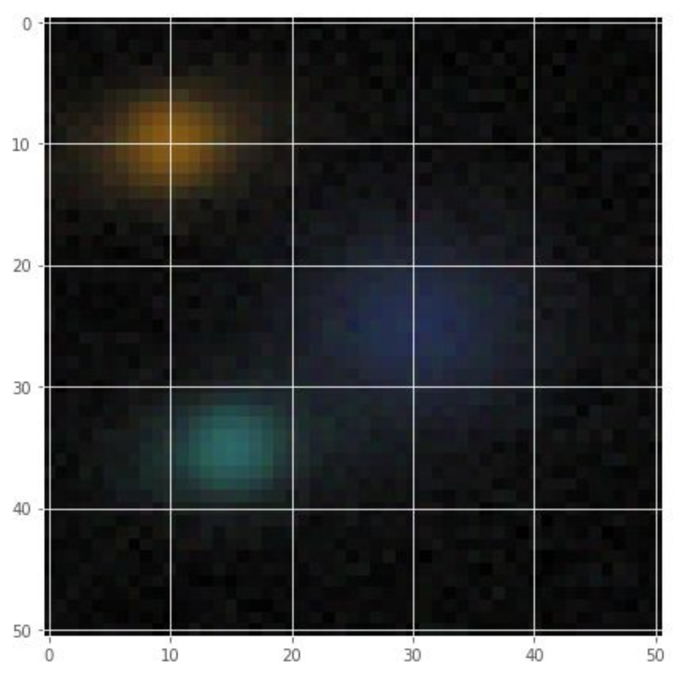
- What if we did our analyses on multi-band image cubes?
  - More information
  - Access to alternative ways of looking at the data.
  - Play with analysis methods that are often associated with catalog-level data, but this time we use them **directly on pixels**.



Multi-Band Detection by Pixel-Level Clustering

Attempt to do:

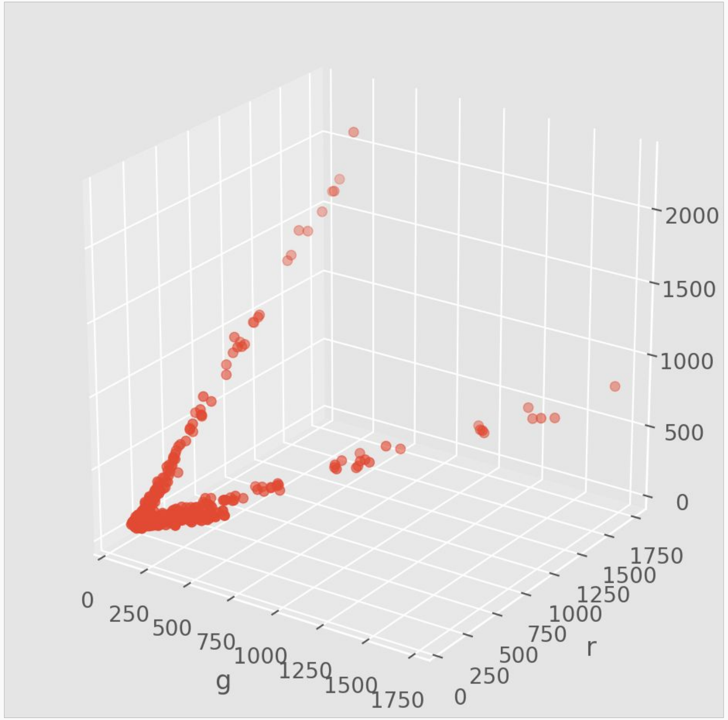
- Detection and Segmentation (finding out which pixel belong to which object)
    - Looking for pixels with similar colors
    - Clustering in a suitable color space.



Example image with 3 objects in a 3-band image cube:

Multi-Band Detection by Pixel-Level Clustering

- Traditional view of false-color image, ordered by pixel position.
- But what about flipping this:
  - the same data, but now ordered by intensity in *gri*, each dot is one pixel.

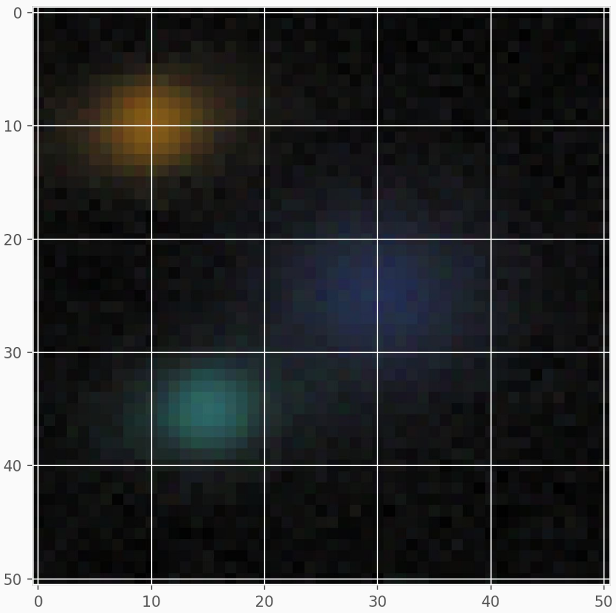


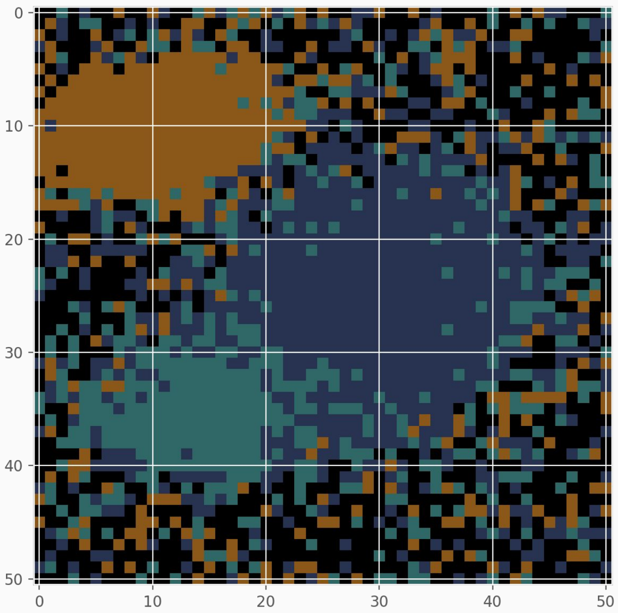
Multi-Band Detection by Pixel-Level Clustering

- Series of image transformations later
  - Filtering noise, etc.
- Using very little code:

```
from sklearn import cluster
kmeans = cluster.KMeans(n_clusters=3)
labels_ = kmeans.fit_predict(v_)
```

Multi-Band Detection by Pixel-Level Clustering





k-means found a clustering in color space that looks plausible!

Bright pixels : single object

Black pixels : outliers (regions of mostly sky)

That’s it! We got something that makes sense with NO spatial information

Structure of Future Seminars

- Where: Watanabe 415
- When: Every other Tuesday evening ~1 hour (TBD)
- Topics
  - **Algorithms:** Regression, Classification, Density Estimation, Clustering, Dimensionality Reduction
    - Neural Networks, Deep learning, Markov Chain Monte Carlo, etc.
  - **Tools:** GitHub, Jupyter Notebooks, TensorFlow, Code documentation, Writing good (readable) code.