Homework 5

References

- Lectures 17-20 (inclusive).

Instructions

- Type your name and email in the "Student details" section below.
- Develop the code and generate the figures you need to solve the problems using this notebook.
- For the answers that require a mathematical proof or derivation you should type them using latex. If you have never written latex before and you find it exceedingly difficult, we will likely accept handwritten solutions.
- The total homework points are 100. Please note that the problems are not weighed equally.

In [1]:
```
import numpy as np
np.set_printoptions(precision=3)
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
sns.set(rc={"figure.dpi":100, "savefig.dpi":300})
sns.set_context("notebook")
sns.set_style("ticks")

import scipy
import scipy.stats as st
import urllib.request
import os

def download(
    url : str,
    local_filename : str = None
):
    """Download a file from a url.

    Arguments
    url            -- The url we want to download.
    local_filename -- The filemame to write on. If not
                      specified
    """
    if local_filename is None:
        local_filename = os.path.basename(url)
    urllib.request.urlretrieve(url, local_filename)
```

Student details

- **First Name: Shaunak**
- **Last Name: Mukherjee**
- **Email: mukher86@purdue.edu**

Problem 1 - Clustering Uber Pickup Data

In this problem you will analyze Uber pickup data collected during April 2014 around New York City. The complete data are freely on Kaggle. The data consist of a timestamp (which we are going to ignore), the latitude and longitude of the Uber pickup, and a base code (which we are also ignoring). The data file we are going to use is uber-raw-data-apr14.csv. As usual, you have to make it visible to this Jupyter notebook. On Google Colab, just run this:

In [2]:
```
url = "https://github.com/PredictiveScienceLab/data-analytics-se/raw/master/lecturebook/
download(url)
```

And you can load it using pandas:

In [3]:
```
import pandas as pd
p1_data = pd.read_csv('uber-raw-data-apr14.csv')
```

Here is how the data look like:

In [4]:
```
p1_data
```

Out[4]:

	Date/Time	Lat	Lon	Base
0	4/1/2014 0:11:00	40.7690	-73.9549	B02512
1	4/1/2014 0:17:00	40.7267	-74.0345	B02512
2	4/1/2014 0:21:00	40.7316	-73.9873	B02512
3	4/1/2014 0:28:00	40.7588	-73.9776	B02512
4	4/1/2014 0:33:00	40.7594	-73.9722	B02512
...	...	...	...	...
564511	4/30/2014 23:22:00	40.7640	-73.9744	B02764
564512	4/30/2014 23:26:00	40.7629	-73.9672	B02764
564513	4/30/2014 23:31:00	40.7443	-73.9889	B02764
564514	4/30/2014 23:32:00	40.6756	-73.9405	B02764
564515	4/30/2014 23:48:00	40.6880	-73.9608	B02764

564516 rows × 4 columns

If you have never played before with pandas, you can find a nice tutorial here.

We have half a million data points. Let's extract the latitude and longitude and put them in a numpy array:

In [5]:
```
loc_data = p1_data[['Lon', 'Lat']]
loc_data
```

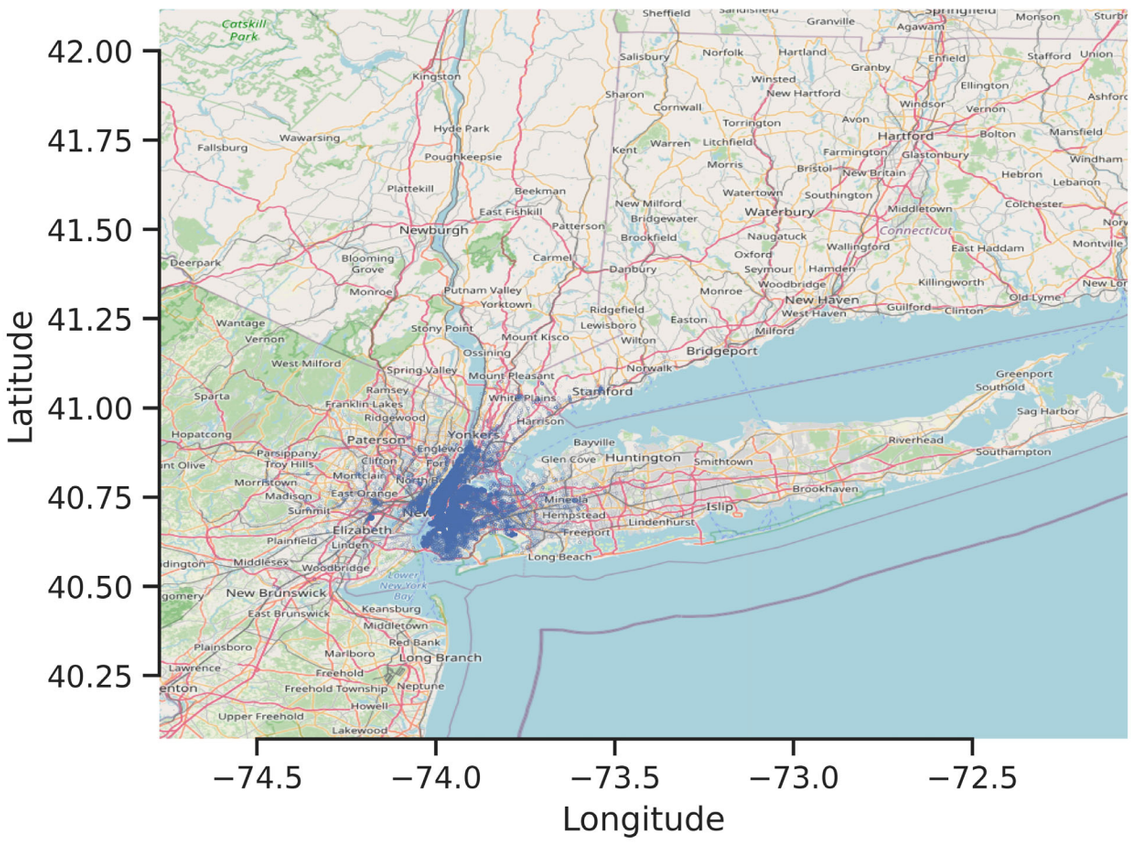
Out[5]:

	Lon	Lat
0	-73.9549	40.7690
1	-74.0345	40.7267
2	-73.9873	40.7316
3	-73.9776	40.7588
4	-73.9722	40.7594
...	...	...
564511	-73.9744	40.7640
564512	-73.9672	40.7629
564513	-73.9889	40.7443
564514	-73.9405	40.6756
564515	-73.9608	40.6880

564516 rows × 2 columns

Let's visualize these points on the map of New York City:

In [6]:
```
url = "https://github.com/PredictiveScienceLab/data-analytics-se/raw/master/lecturebook/
download(url)
ny_map = plt.imread('ny_map.png')
box = ((loc_data.Lon.min(), loc_data.Lon.max(),
        loc_data.Lat.min(), loc_data.Lat.max()))
fig, ax = plt.subplots(dpi=600)
ax.scatter(
    loc_data.Lon,
    loc_data.Lat,
    zorder=1,
    alpha= 1,
    c='b',
    s=0.001
)
ax.set_xlim(box[0],box[1])
ax.set_ylim(box[2],box[3])
ax.imshow(
    ny_map,
    zorder=0,
    extent=box,
    aspect= 'equal'
)
ax.set_xlabel('Longitude')
ax.set_ylabel('Latitude')
sns.despine(trim=True);
fig.show()
```



Machine learning algorithms will be a bit slow because we have over half a million data points. So, as you develop your code, use only 50K observations. Once you have a stable version of your code, modify the following code segment to use the entire dataset.

In [7]:
```
p1_train_data = loc_data[:50_000]
```

Part A - Splitting New York City into Subregions

Suppose you are assigned to split New York City into operating subregions with equal demand. When a pickup is requested in each subregion, only the drivers in that region are called. Note that this can become a challenging problem very quickly. We are not looking for the best possible answer here. We are looking for a data-informed heuristic solution that is good enough.

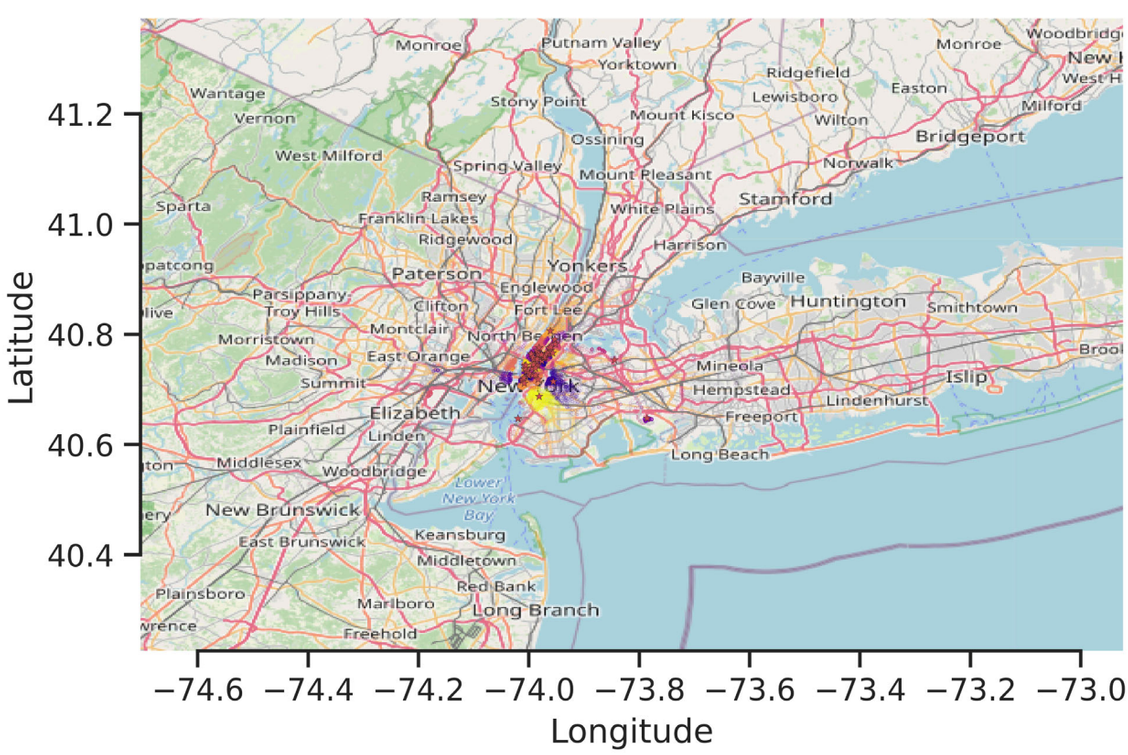
Do (at least) the following:

- Use Kmeans clustering on the pickup data with different numbers of clusters;
- Visualize the labels of the clusters on the map using different colors (see the hands-on activities);
- Visualize the centers of the discovered Kmeans clusters (in red color);
- Use common sense, e.g., ensure there are enough clusters so no region crosses the water. If it is impossible to get perfect results simply by Kmeans, feel free to ignore a small number of outliers as they could be handled manually;
- Use MiniBatchKMeans, which is a much faster version of Kmeans suitable for large datasets (>10K observations);

Answer with as many text and code blocks as you like below.

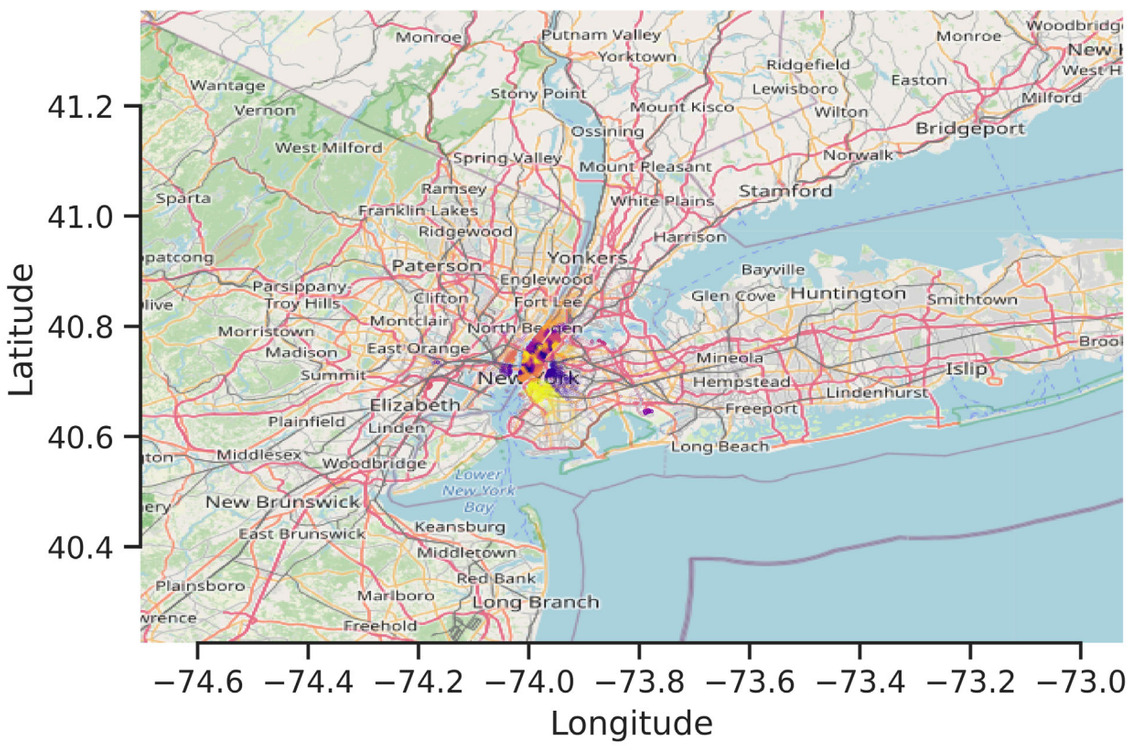
In [8]:

```
import sklearn.cluster
from matplotlib.colors import Normalize

n_clusters = 50
newbox = ((p1_train_data['Lon'].min(), p1_train_data['Lon'].max(),
           p1_train_data['Lat'].min(), p1_train_data['Lat'].max()))

kmeans = sklearn.cluster.MiniBatchKMeans(n_clusters=n_clusters, max_iter=1000, batch_si

fig, ax = plt.subplots(dpi=600)
scatter = ax.scatter(
    p1_train_data['Lon'],
    p1_train_data['Lat'],
    c=kmeans.labels_,
    cmap='plasma',
    zorder=1,
    alpha=0.7,
    s=0.001
)
ax.plot(kmeans.cluster_centers_[:, 0],kmeans.cluster_centers_[:, 1], '*r', markeredgeco

ax.set_xlim(newbox[0], newbox[1])
ax.set_ylim(newbox[2], newbox[3])

ax.imshow(
    ny_map,
    zorder=0,
    extent=box,
    aspect='equal'
)

ax.set_xlabel('Longitude')
ax.set_ylabel('Latitude')
sns.despine(trim=True);
```



In [9]:

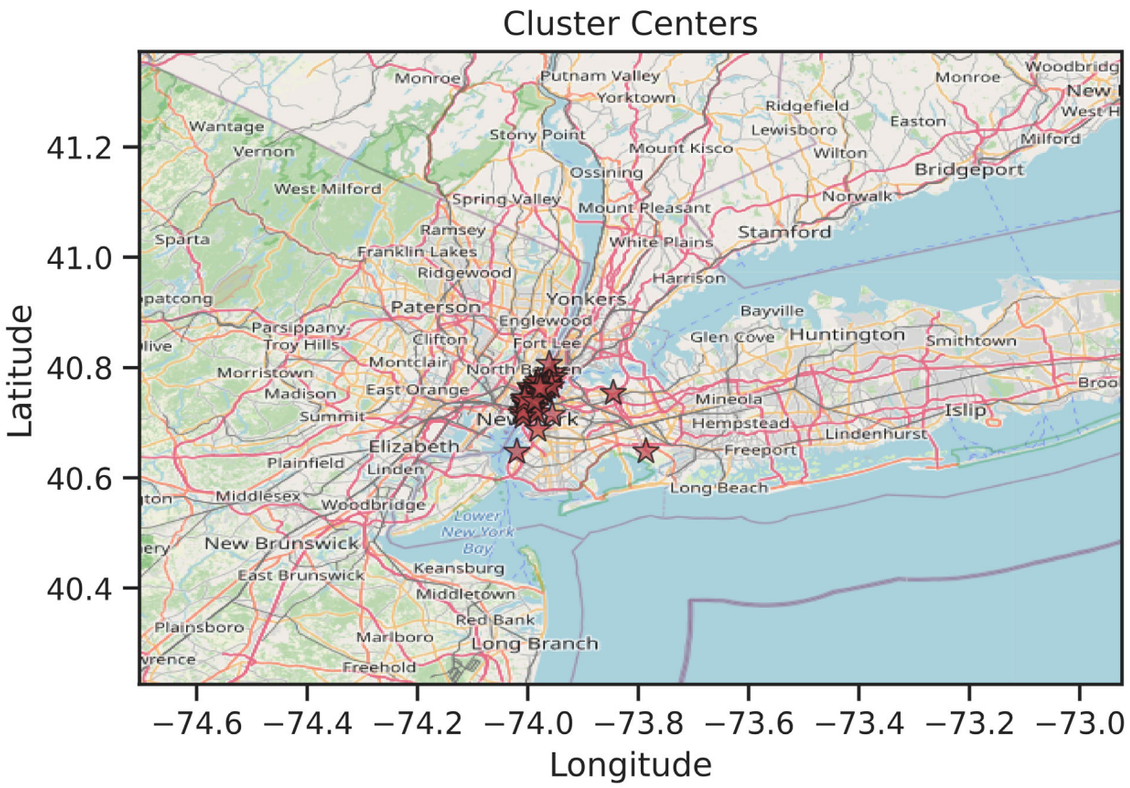
```
# Showing only clusters wihout the centers
import sklearn.cluster
from matplotlib.colors import Normalize

n_clusters = 50
newbox = ((p1_train_data['Lon'].min(), p1_train_data['Lon'].max(),
           p1_train_data['Lat'].min(), p1_train_data['Lat'].max()))

kmeans = sklearn.cluster.MiniBatchKMeans(n_clusters=n_clusters, max_iter=1000, batch_si

fig, ax = plt.subplots(dpi=600)
scatter = ax.scatter(
    p1_train_data['Lon'],
    p1_train_data['Lat'],
    c=kmeans.labels_,
    cmap='plasma',
    zorder=1,
    alpha=0.7,
    s=0.001
)
# ax.plot(kmeans.cluster_centers_[:, 0],kmeans.cluster_centers_[:, 1], '*r', markeredge

ax.set_xlim(newbox[0], newbox[1])
ax.set_ylim(newbox[2], newbox[3])

ax.imshow(
    ny_map,
    zorder=0,
    extent=box,
    aspect='equal'
)

ax.set_xlabel('Longitude')
ax.set_ylabel('Latitude')
sns.despine(trim=True);
```



In [10]:

```
# Plotting the cluster center only without the overlay for clarity

fig, ax = plt.subplots(dpi=600)

ax.imshow(
    ny_map,
    zorder=0,
    extent=box,
    aspect= 'equal'
)

ax.set_xlim(newbox[0],newbox[1])
ax.set_ylim(newbox[2],newbox[3])
```

```
ax.set_xlabel('Longitude')
ax.set_ylabel('Latitude')
ax.set_title('Cluster Centers')

ax.plot(kmeans.cluster_centers_[:, 0],kmeans.cluster_centers_[:, 1], '*r', markeredgeco
```



Part B - Create a Stochastic Model of Pickups

One of the key ingredients for a more sophisticated approach to optimizing the operations of Uber is the construction of a stochastic model of the demand for pickups. The ideal model for this problem is the Poisson Point Process. However, we will do something more straightforward, using the Gaussian mixture model and a Poisson random variable. The model will not have a time component, but it will allow us to sample the number and locations of pickups during a typical month. We will guide you through the process of constructing this model.

Subpart B.I - Random variable capturing the number of monthly pickups

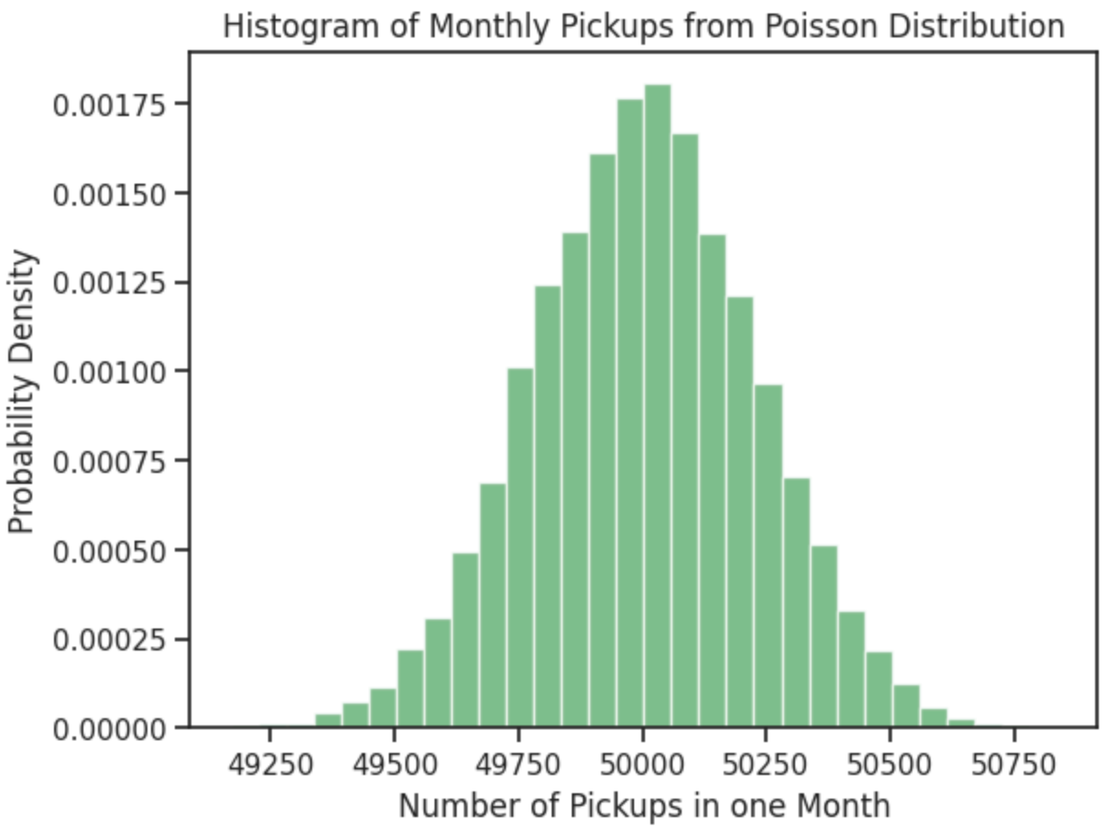
Find the rate of monthly pickups (ignore the fact that months may differ by a few days) and use it to define a Poisson random variable corresponding to the monthly number of pickups. Use `scipy.stats.poisson` to initialize this random variable. Sample from it 10,000 times and plot the histogram of the samples to get a feeling about the corresponding probability mass function.

In [11]:

```
# Calculate the monthly rate of pickups
monthly_rate = len(p1_train_data)
print(f"The rate of monthly pick up is {monthly_rate}")

# Define the Poisson random variable
monthly_pickups_dist = st.poisson(mu=monthly_rate)

# Sample from the Poisson distribution
monthly_samples = monthly_pickups_dist.rvs(size=10000)

# Plot the histogram of the samples
fig, ax = plt.subplots()

ax.hist(monthly_samples, bins=30, density=True, alpha=0.75, color='g')
ax.set_ylabel("Probability Density")
ax.set_xlabel("Number of Pickups in one Month")
ax.set_title("Histogram of Monthly Pickups from Poisson Distribution")
```

```
# Show the plot
plt.show()
```

```
The rate of monthly pick up is 50000
```

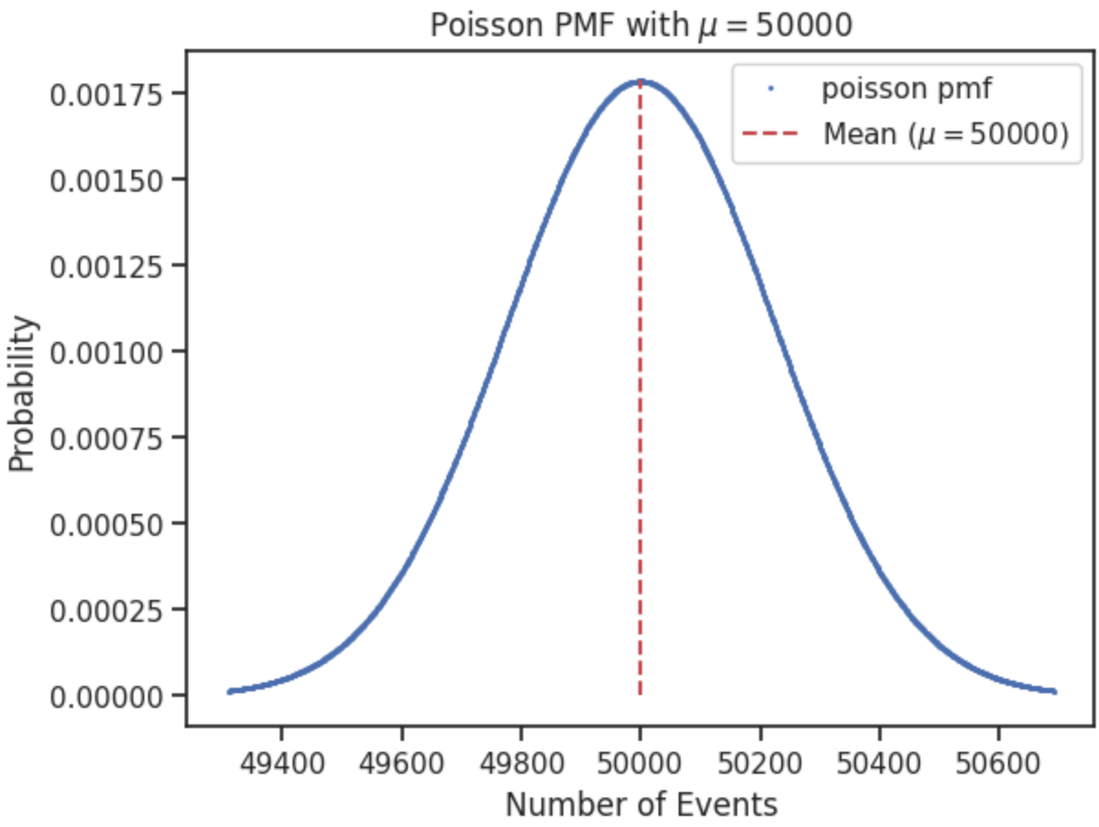


Lets also vizualize the PMF below to compare

In [12]:

```
#
# Mean of the Poisson distribution
mu = 50000

# Determine the range for the PMF using PPF to get the values at the 0.1th and 99.9th p
x1 = np.arange(st.poisson.ppf(0.001, mu), st.poisson.ppf(0.999, mu))

# Print the first and last values of the range
print(f"Distribution ranges from {x1[0]} to {x1[-1]} and centered at {mu}")

# Plot the PMF
fig, ax = plt.subplots()
ax.plot(x1, st.poisson.pmf(x1, mu), 'bo', ms=1, label='poisson pmf')
ax.vlines(mu, ymin=0, ymax=max(st.poisson.pmf(x1, mu)), colors='r', linestyles='dashed'

# Add labels and title
ax.set_xlabel("Number of Events")
ax.set_ylabel("Probability")
ax.set_title("Poisson PMF with $\mu=50000$")
ax.legend()

# Show the plot
plt.show()
```

```
Distribution ranges from 49310.0 to 50691.0 and centered at 50000
```



Subpart B.II - Sample some random monthly pickup numbers

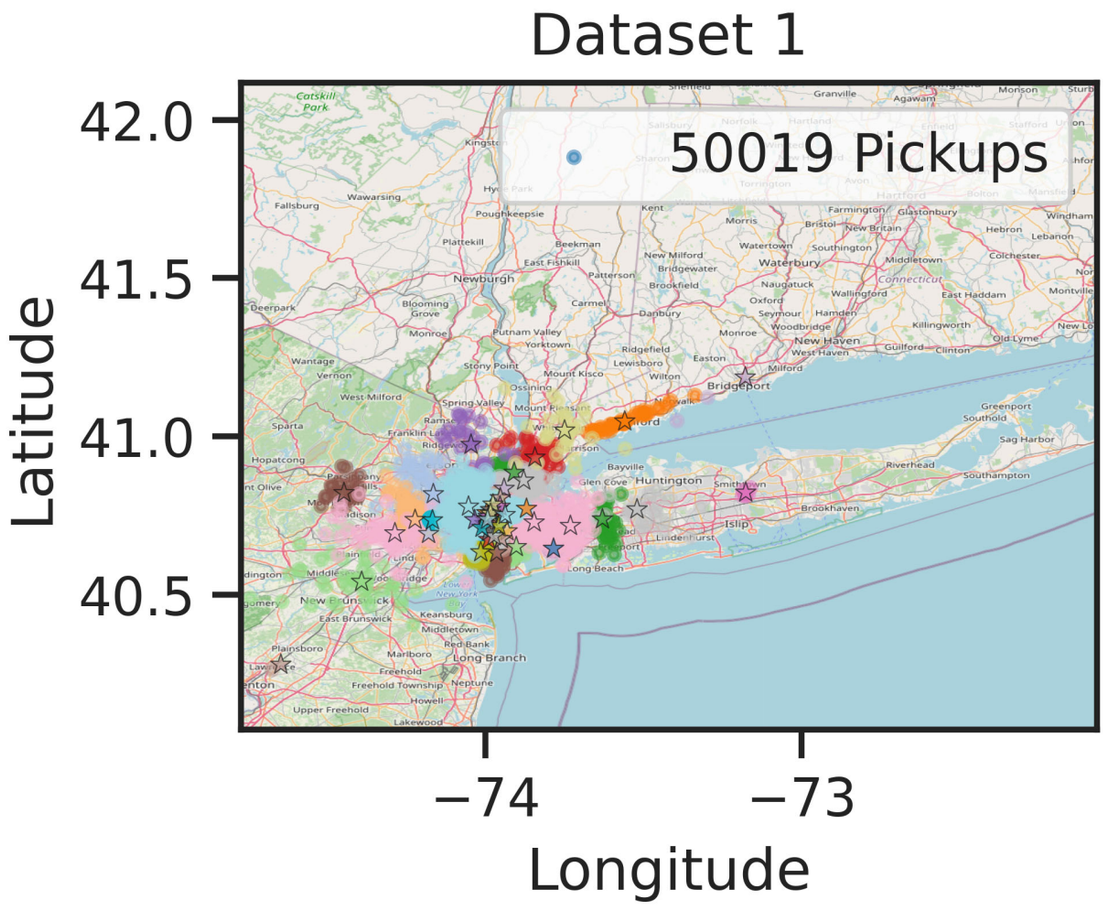
Now that you have a model that gives you the number of pickups and a model that allows you to sample a pickup location, sample five different datasets (number of pickups and location of each pick) from the combined model and visualize them on the New York map.

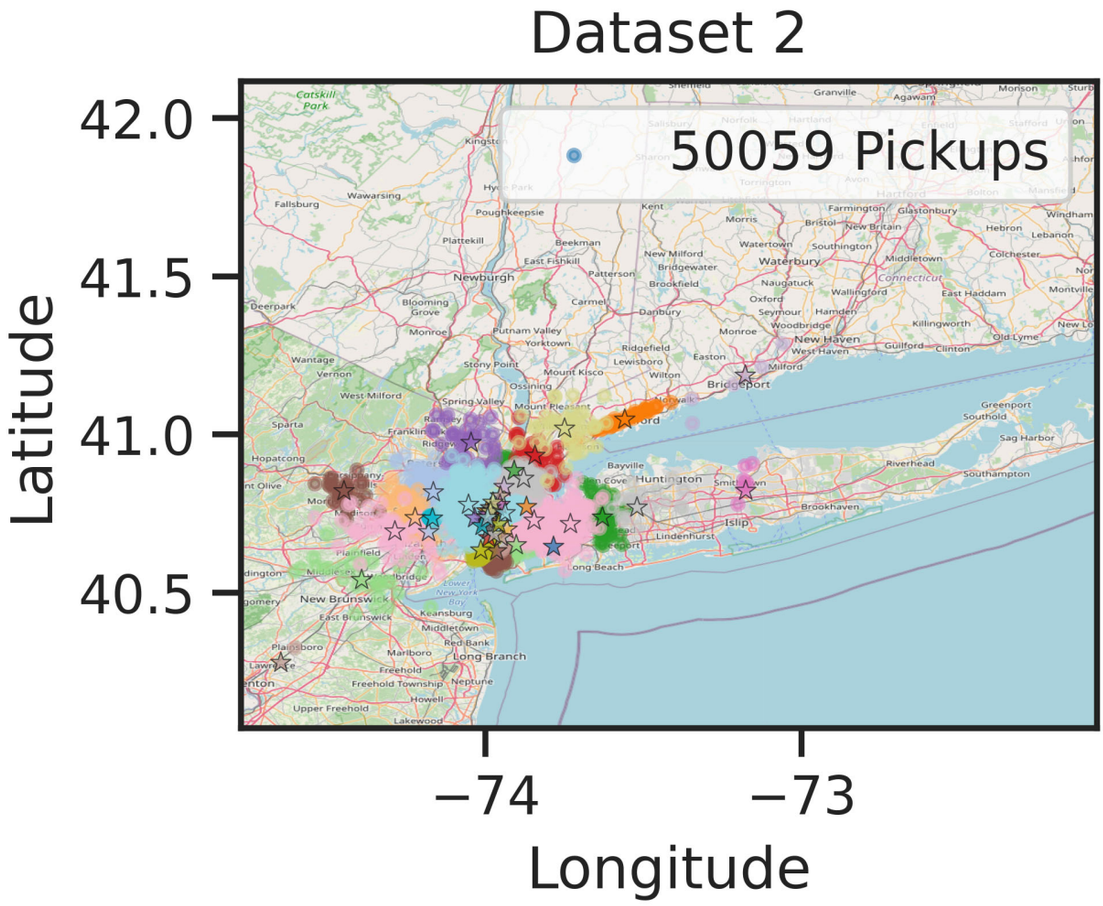
Hint: Don't get obsessed with making the model perfect. It's okay if a few of the pickups are on water.

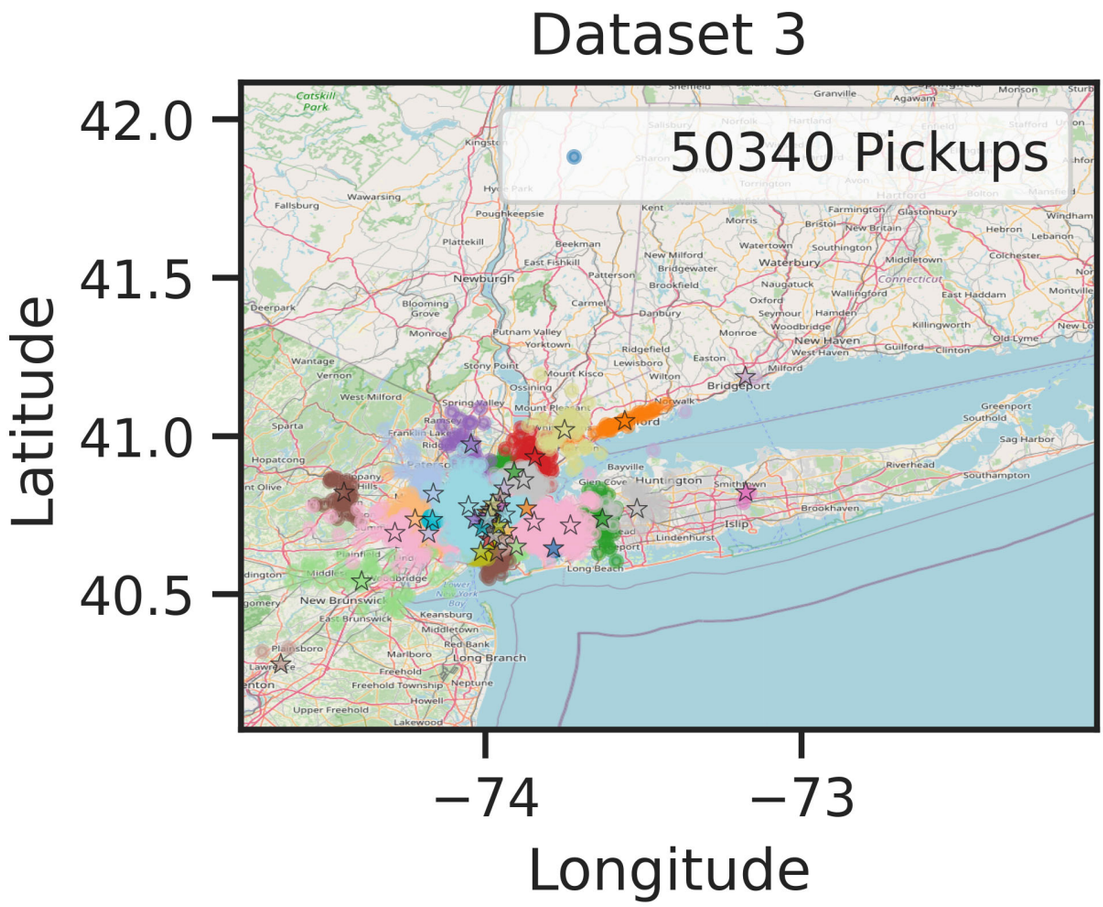
In [13]:

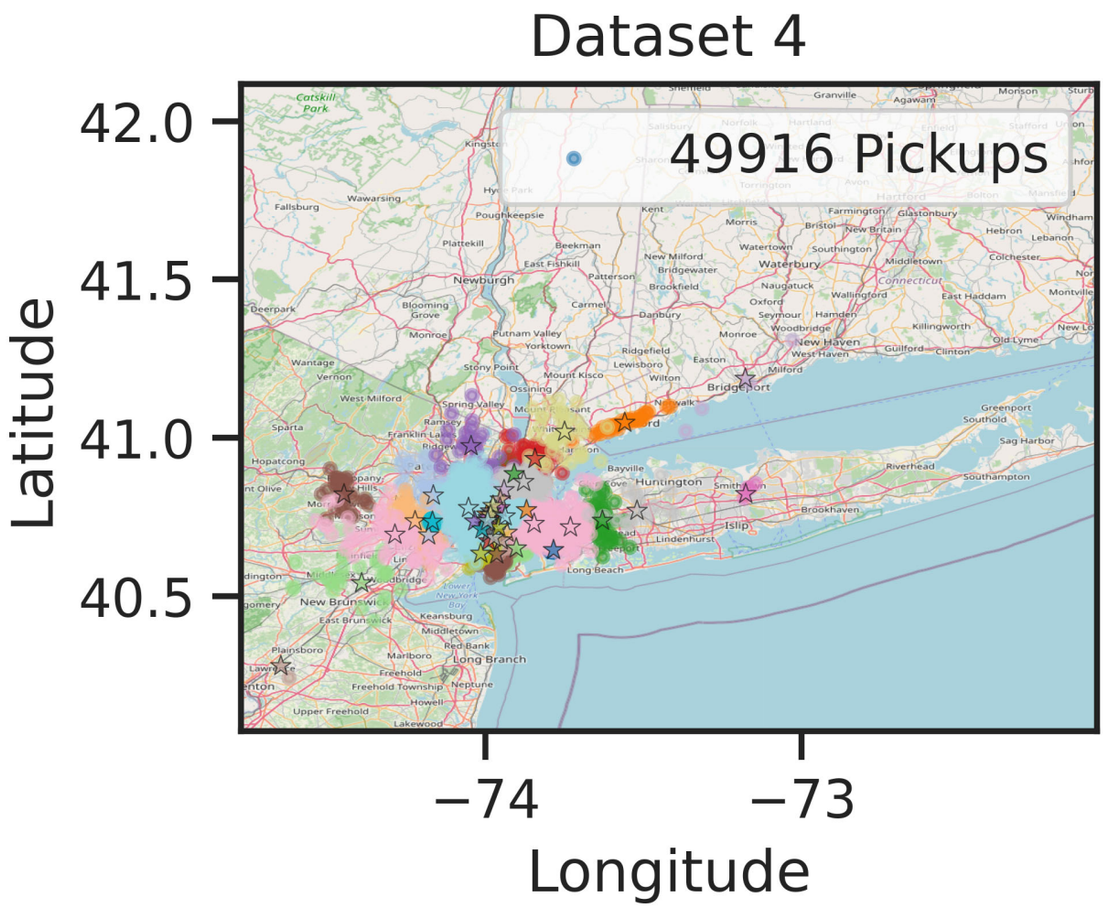
```
import sklearn.mixture

n_clusters = 50
# Fit Gaussian Mixture Model
gaussian_mixture = sklearn.mixture.GaussianMixture(n_components=n_clusters).fit(p1_trai

# Defining number of datasets
num_datasets = 5

# Sample number of pickups for each dataset
num_pickups_samples = monthly_pickups_dist.rvs(num_datasets)


# Figure setting
fig, axs = plt.subplots(num_datasets, 1, figsize = (8,20), dpi=600)

# Iterate over each dataset
for i, num_pickups in enumerate(num_pickups_samples):
    # Sample locations from the Gaussian Mixture Model
    location_samples, cluster_idx = gaussian_mixture.sample(num_pickups)

    # Plotting setup for each subplot
    ax = axs[i]
    ax.imshow(
        ny_map,
        zorder=0,
        extent=box,
        aspect='equal'
    )
    ax.scatter(
        location_samples[:, 0],
        location_samples[:, 1],
        s=5,
        alpha=0.5,
        label=f'{num_pickups} Pickups',
        c=cluster_idx,  # Color by cluster index
```

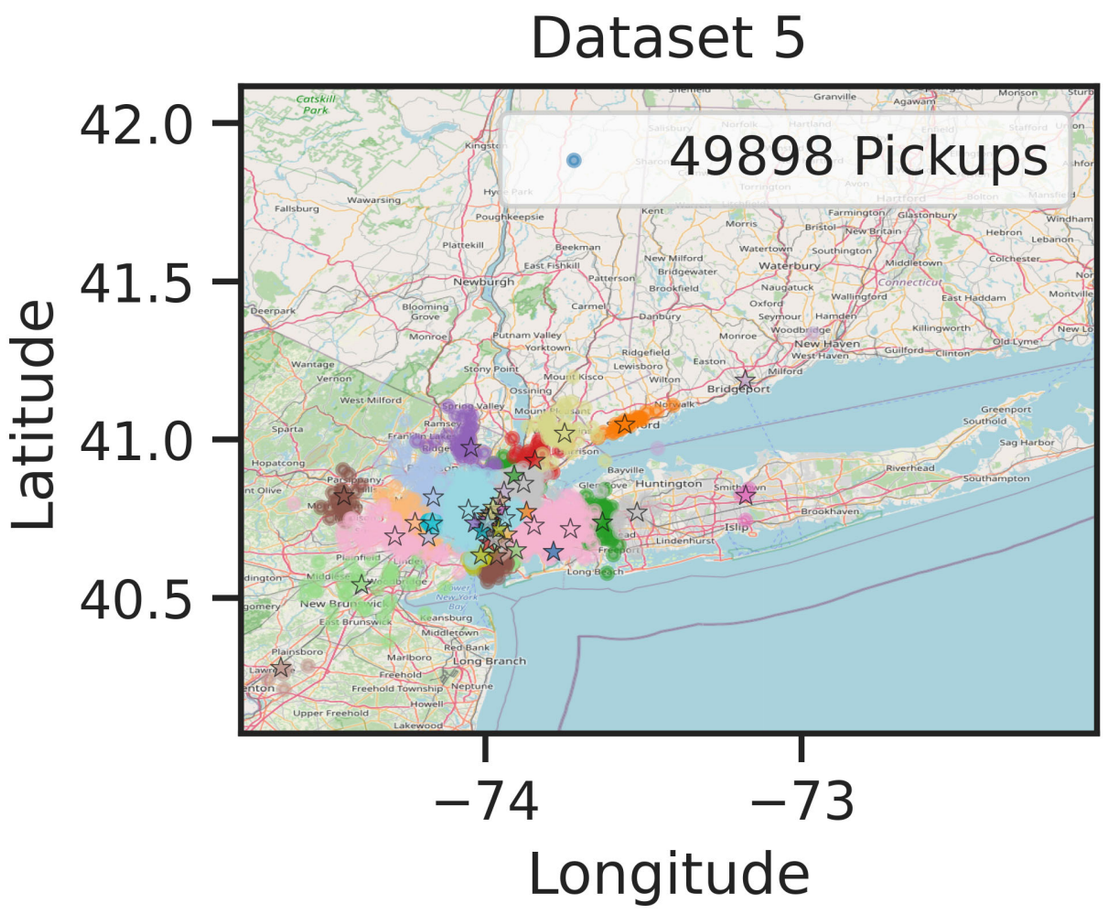
```
            cmap='tab20'  # Color map for visual distinction
        )
        ax.set_xlabel('Longitude')
        ax.set_ylabel('Latitude')
        ax.set_title(f'Dataset {i+1}')
        ax.legend()

        # Plotting means of Gaussian components
        ax.scatter(
            gaussian_mixture.means_[:, 0],
            gaussian_mixture.means_[:, 1],
            marker='*',
            c=np.arange(n_clusters),  # Color by cluster index
            edgecolor='k',
            s=20,  # Marker size
            linewidth=0.25,  # Edge width
            alpha=0.7,
            cmap='tab20'  # Color map for visual distinction
        )

    # Adjust layout and show plot
    # plt.tight_layout()
    plt.subplots_adjust(hspace=0.8) # Adjustment needed for pdf conversion
    plt.show()
```


Dataset 1
50019 Pickups
Latitude
42.0
41.5
41.0
40.5
−74
−73
Longitude


Dataset 2
50059 Pickups
Latitude
42.0
41.5
41.0
40.5
−74
−73
Longitude


Dataset 3
50340 Pickups
Latitude
42.0
41.5
41.0
40.5
−74
−73
Longitude


Dataset 4
49916 Pickups
Latitude
42.0
41.5
41.0
40.5
−74
−73
Longitude



Problem 2 - Counting Celestial Objects

Consider this picture of a patch of sky taken by the Hubble Space Telescope.

Let's download it so that you have it here:

In [14]:
```
url = 'https://raw.githubusercontent.com/PredictiveScienceLab/data-analytics-se/master/
download(url)
```

This picture includes many galaxies but also some stars. We will create a machine-learning model capable of counting the number of objects in such images. Our model will not be able to differentiate between the different types of objects and will not be very accurate. Still, it does form the basis of more sophisticated approaches. The idea is as follows:

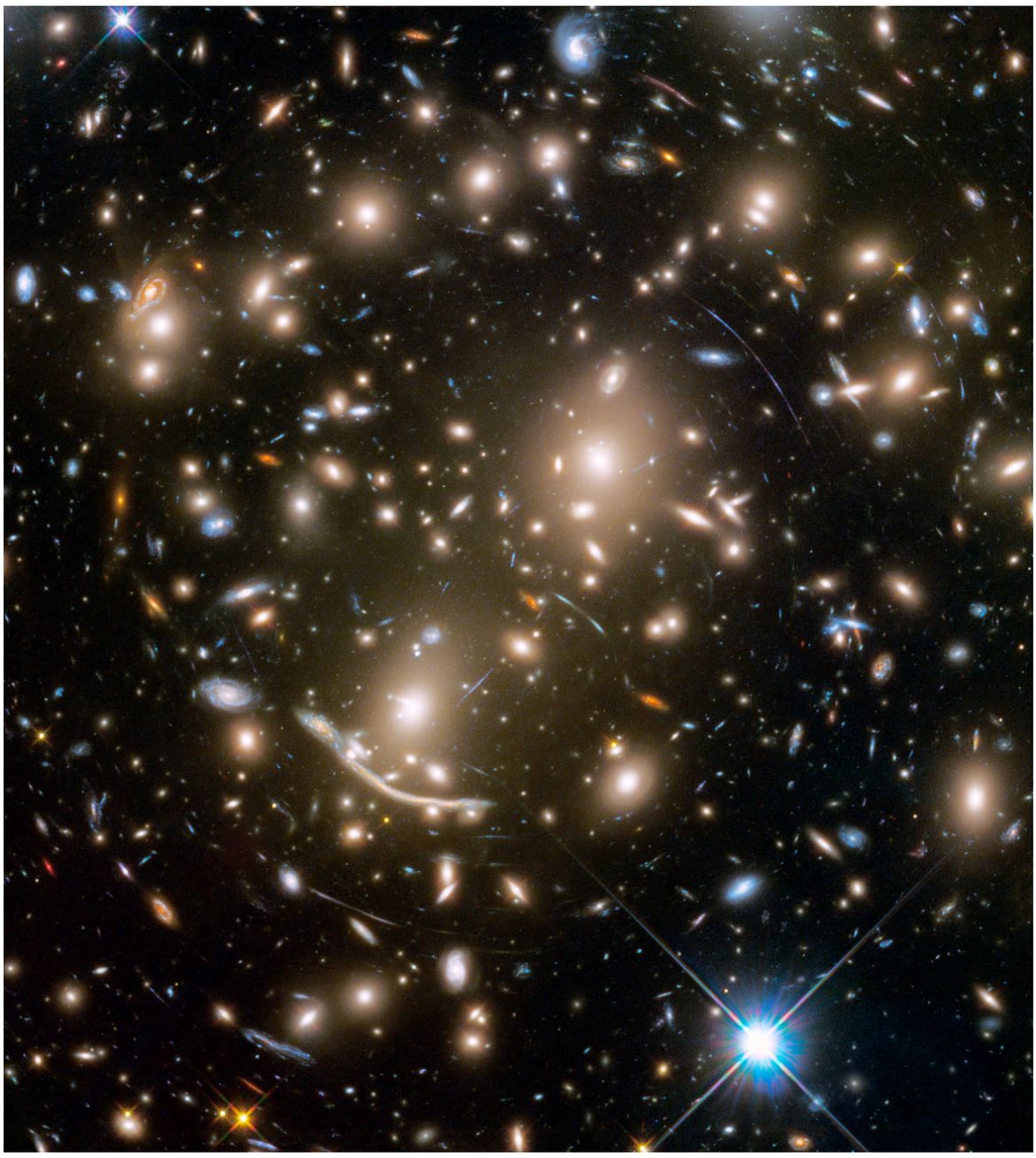
- Convert the picture to points sampled according to the intensity of light.
- Apply Gaussian mixture on the resulting points.
- Use the Bayesian Information Criterion to identify the number of components in the picture.
- Associate the number of components with the actual number of celestial objects.

I will set you up with the first step. You will have to do the last three.

We are going to load the image with the Python Imaging Library (PIL), which allows us to apply a few basic transformations to the image:

In [15]:
```
from PIL import Image
hubble_image = Image.open('galaxies.png')
# here is how to see the image
hubble_image
```

Out[15]:

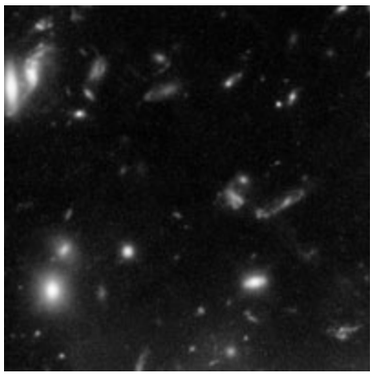


Now, we are going to convert it to grayscale and crop it to make the problem a little bit easier:

In [16]:

```
img = hubble_image.convert('L').crop((100, 100, 300, 300))
img
```

Out[16]:

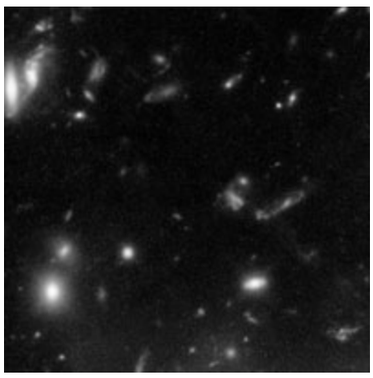


Remember that black-and white images are matrices:

In [17]:

```
img_ar = np.array(img)
img_ar
```

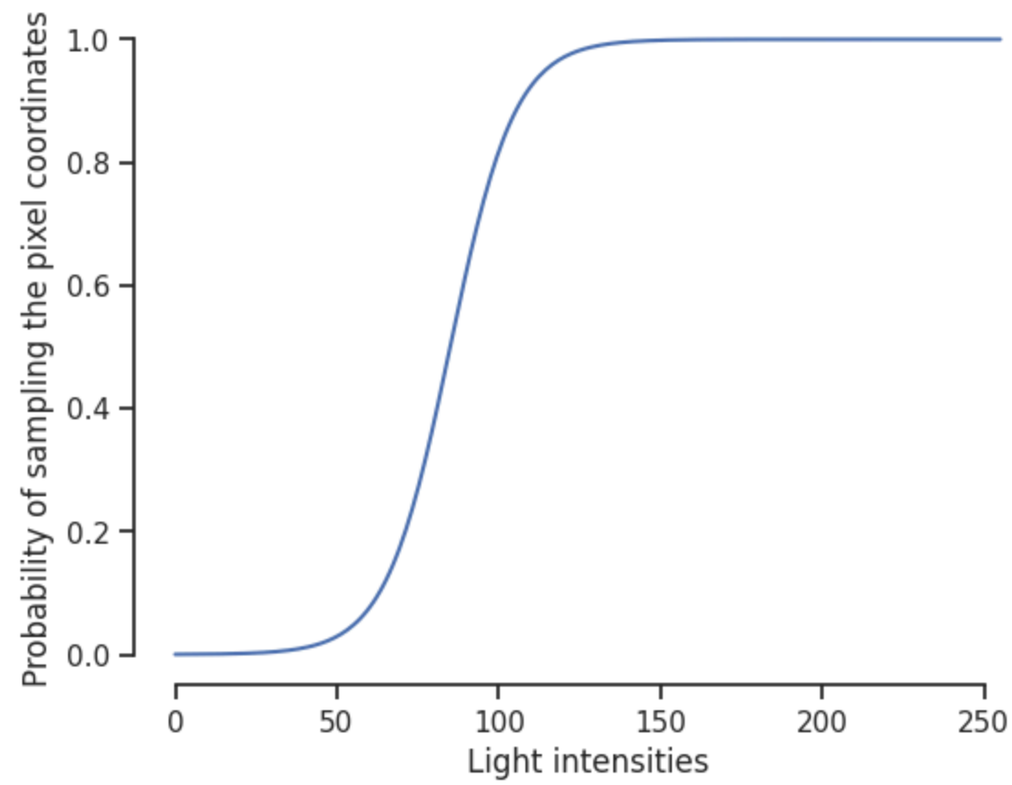
Out[17]: ndarray (200, 200) show data



The minimum number is 0, corresponding to black, and the maximum is 255, corresponding to white. Anything in between is some shade of gray.

Now, imagine that each pixel is associated with some coordinates. Without loss of generality, let's assume that each pixel is some coordinate in $[0, 1]^2$. We will loop over each pixel and sample its coordinates in a way that increases with increasing light intensity. To achieve this, we will pass the intensity values of each pixel through a sigmoid with parameters that can be tuned. Here is this sigmoid:

In [18]:
```
intensities = np.linspace(0, 255, 255)
fig, ax = plt.subplots()
alpha = 0.1
beta = 255 / 3
ax.plot(
    intensities,
    1.0 / (1.0 + np.exp(-alpha * (intensities - beta)))
);
ax.set_xlabel('Light intensities')
ax.set_ylabel('Probability of sampling the pixel coordinates')
sns.despine(trim=True);
```



And here is the code that samples the pixel coordinates. I am organizing it into a function because we may want to use it with different pictures:

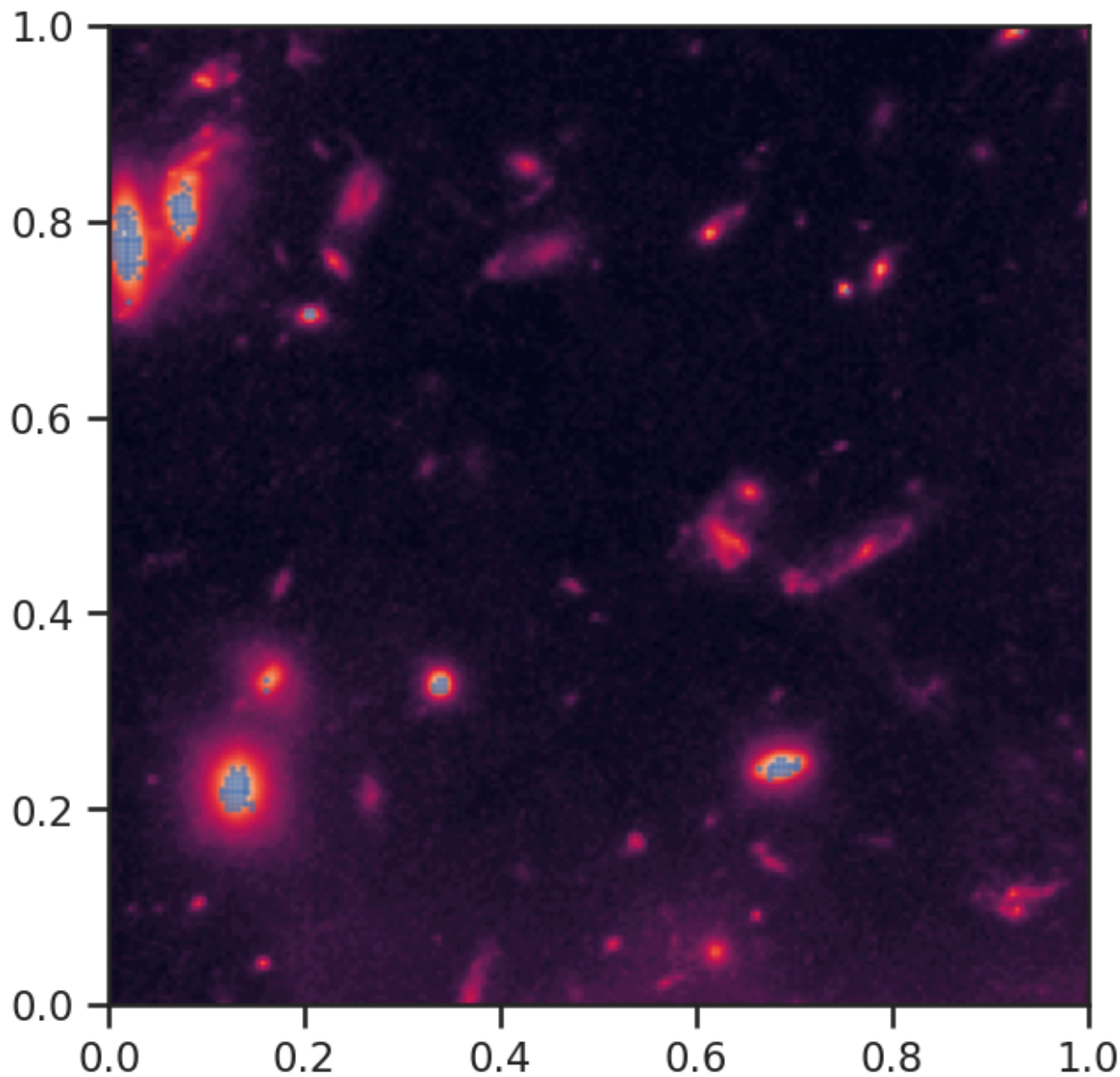
In [19]:
```
def sample_pixel_coords(img, alpha, beta):
    """
    Samples pixel coordinates based on a probability defined as the sigmoid of the inte

    Arguments:

        img    -     The gray scale pixture from which we sample as an array
        alpha     -     The scale of the sigmoid
        beta      -     The offset of the sigmoid
```

```
    """
    img_ar = np.array(img)
    x = np.linspace(0, 1, img_ar.shape[0])
    y = np.linspace(0, 1, img_ar.shape[1])
    X, Y = np.meshgrid(x, y)
    img_to_locs = []
    # Loop over pixels
    for i in range(img_ar.shape[1]):
        for j in range(img_ar.shape[0]):
            # Calculate the probability of the pixel by looking at each
            # light intensity
            prob = 1.0 / (1.0 + np.exp(-alpha * (img_ar[j, i] - beta)))
            # Pick a uniform random number
            u = np.random.rand()
            # If u is smaller than the desired probability,
            # the consider the coordinates of the pixel sampled
            if u <= prob:
                img_to_locs.append((Y[i, j], X[-i-1, -j-1]))
    # Turn img_to_locs into a numpy array
    img_to_locs = np.array(img_to_locs)
    return img_to_locs
```

Let's test it:

In [20]:

```
locs = sample_pixel_coords(img, alpha=0.1, beta=200)
fig, ax = plt.subplots(dpi=150)
ax.imshow(img, extent=((0, 1, 0, 1)), zorder=0)
ax.scatter(
    locs[:, 0],
    locs[:, 1],
    zorder=1,
    alpha=0.5,
    c='b',
    s=1
);
```



Note that playing with α and β makes the whole thing more or less sensitive to the light intensity.

Complete the following function:

In [21]:
```
from sklearn.mixture import GaussianMixture

def count_objs(img, alpha, beta, nc_min=1, nc_max=50):
    """Count objects in image.

    Arguments:
        img       -     The image
        alpha     -     The scale of the sigmoid
        beta      -     The offset of the sigmoid
        nc_min    -     The minimum number of components to consider
        nc_max    -     The maximum number of components to consider
    """
    locs = sample_pixel_coords(img, alpha, beta)
    bic = []
    models = []
    for n_components in range(nc_min, nc_max+1):
        model = GaussianMixture(n_components=n_components, random_state=0)
        model.fit(locs)
        bic.append(model.bic(locs))
        models.append(model)
    best_index = np.argmin(bic)
    best_nc = best_index + nc_min
    best_model = models[best_index]
    print(f"Best number of components: {best_nc} for alpha = {alpha} & beta = {beta}")
    return best_nc, best_model, locs
```

Once you have completed the code, try out the following images. Feel free to play with α and β to improve the performance. **Do not try to make a perfect model. We would have to go beyond the Gaussian mixture model to do so. This is just a homework problem.**

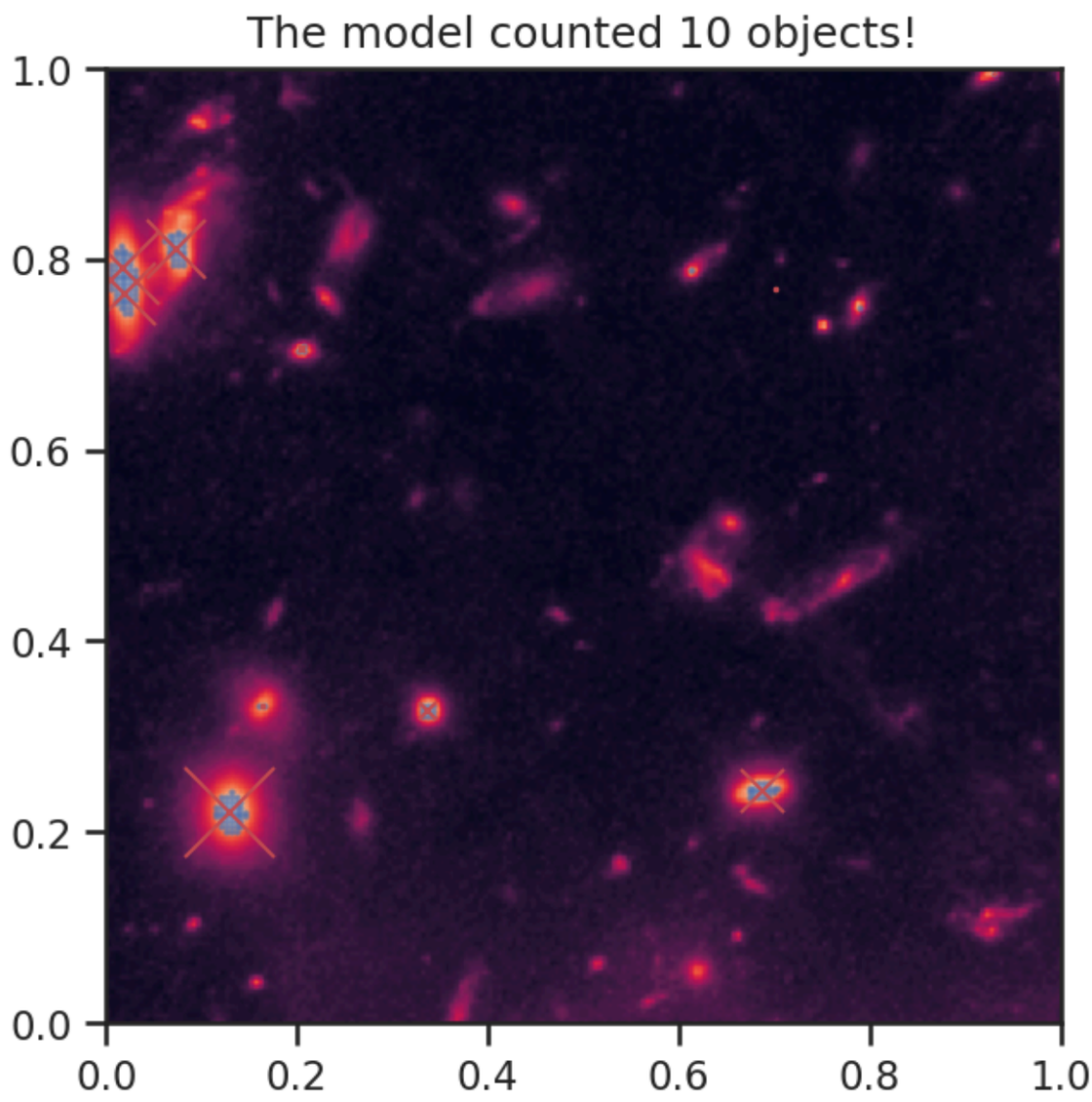
Here is a helpful function that you can use to visualize the results:

In [22]:
```
def visualize_counts(img, objs, model, locs):
    """Visualize the counts.

    Arguments
    img    --  The image.
    objs   --  Returned by count_objs()
    model  --  Returned by count_objs()
    locs   --  Returned by count_objs()
    """

    fig, ax = plt.subplots(dpi=150)
    ax.imshow(img, extent=((0, 1, 0, 1)))
    for i in range(model.means_.shape[0]):
        ax.plot(
            model.means_[i, 0],
            model.means_[i, 1],
            'rx',
            markersize=(
                10.0 * model.weights_.shape[0]
                * model.weights_[i]
            )
        )
    ax.scatter(
        locs[:, 0],
        locs[:, 1],
        zorder=1,
        alpha=0.5,
        c='b',
        s=1
    )


    ax.set_title(f'The model counted {objs} objects!')
```

Here is how to use it:

In [23]:
```
objs, model, locs = count_objs(img, alpha=1, beta=200)
visualize_counts(img, objs, model, locs)
```

```
Best number of components: 10 for alpha = 1 & beta = 200
```



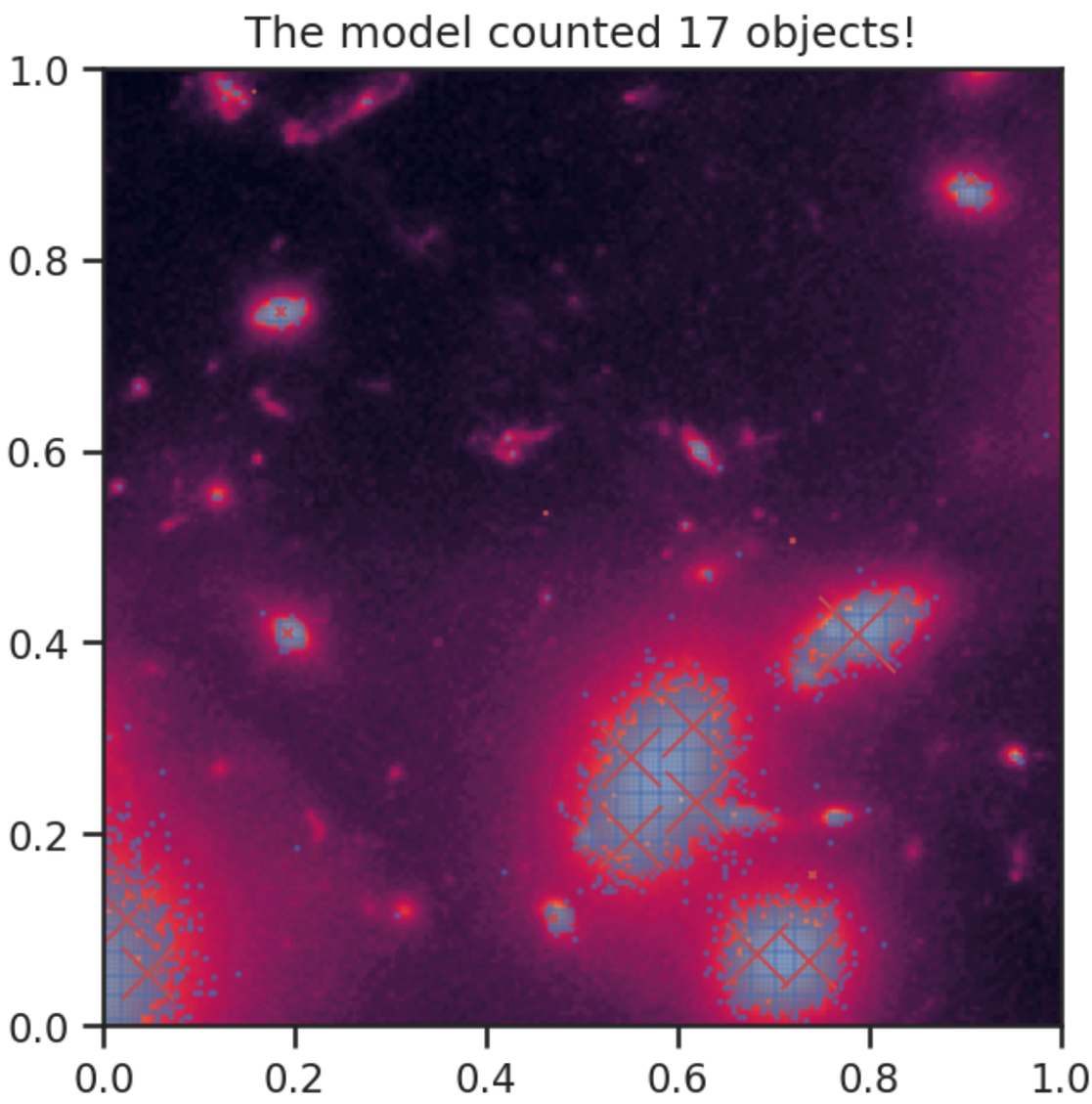
Try this image:

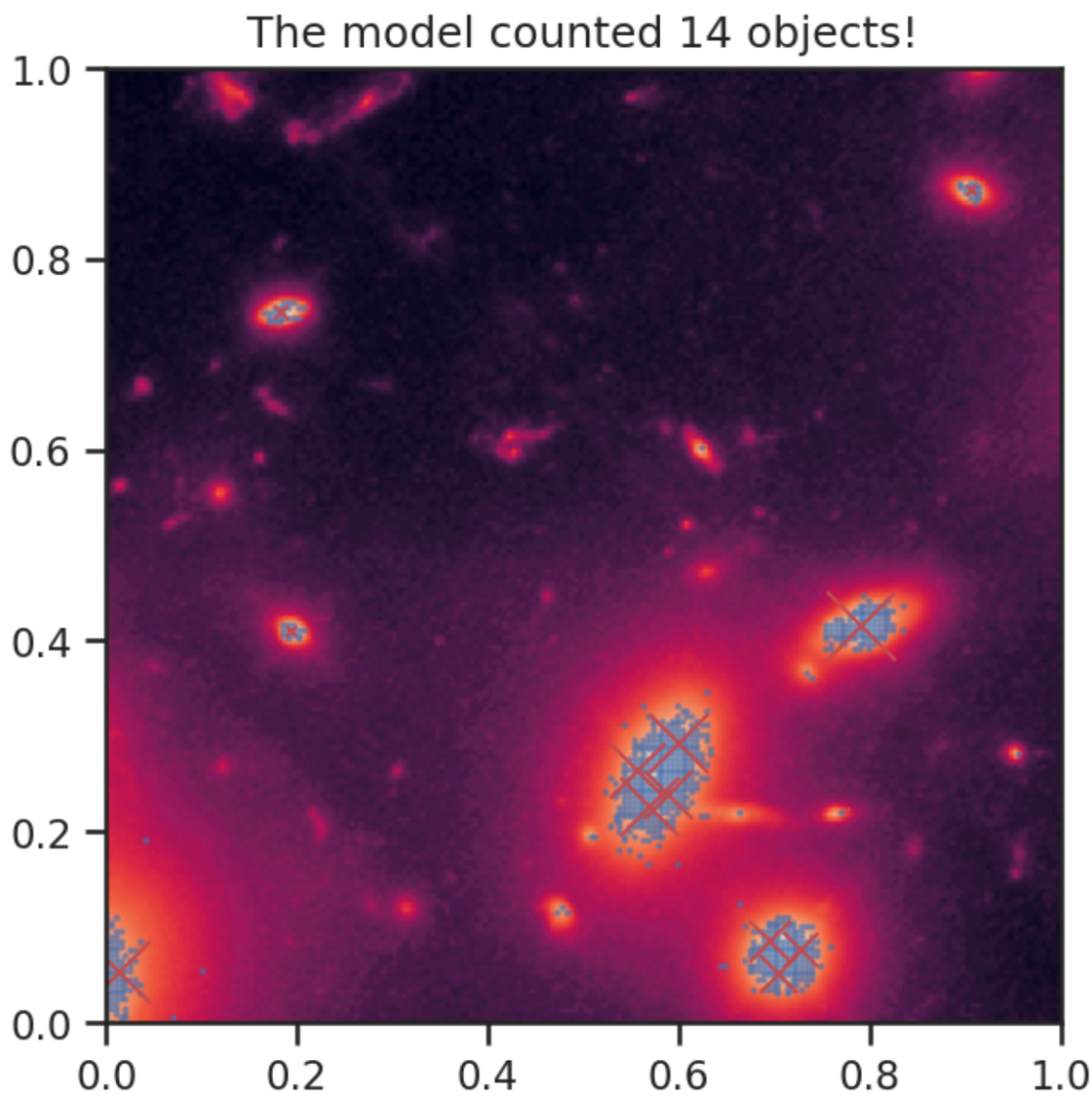
Since ask is to play with α and β without making a perfect model beyond gaussian mixture model, α and β values are systematically varied and their object detection perfoemance evaluated each case. Note the provided alpha and beta values are included in the single variable study.

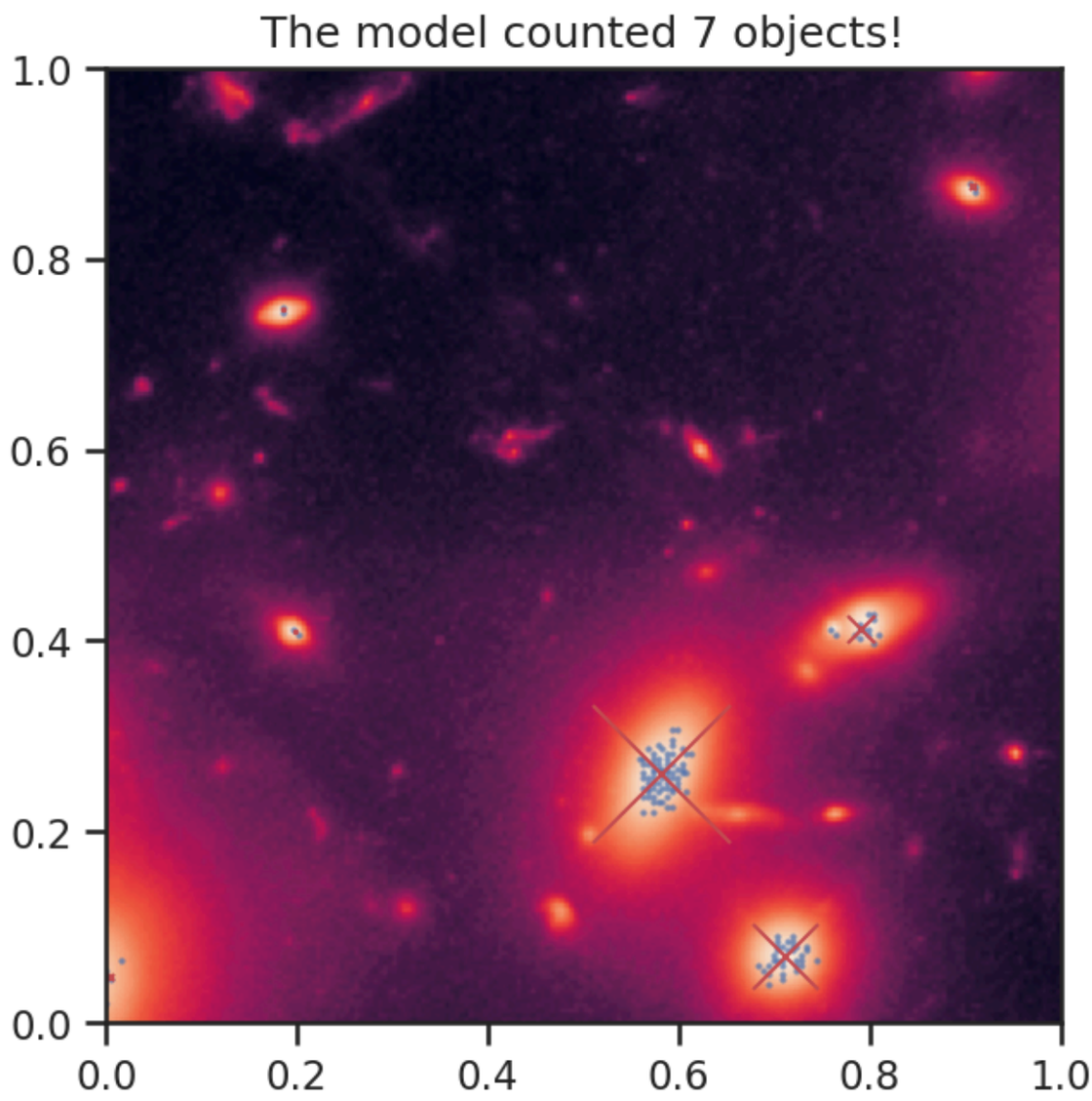
In [24]:

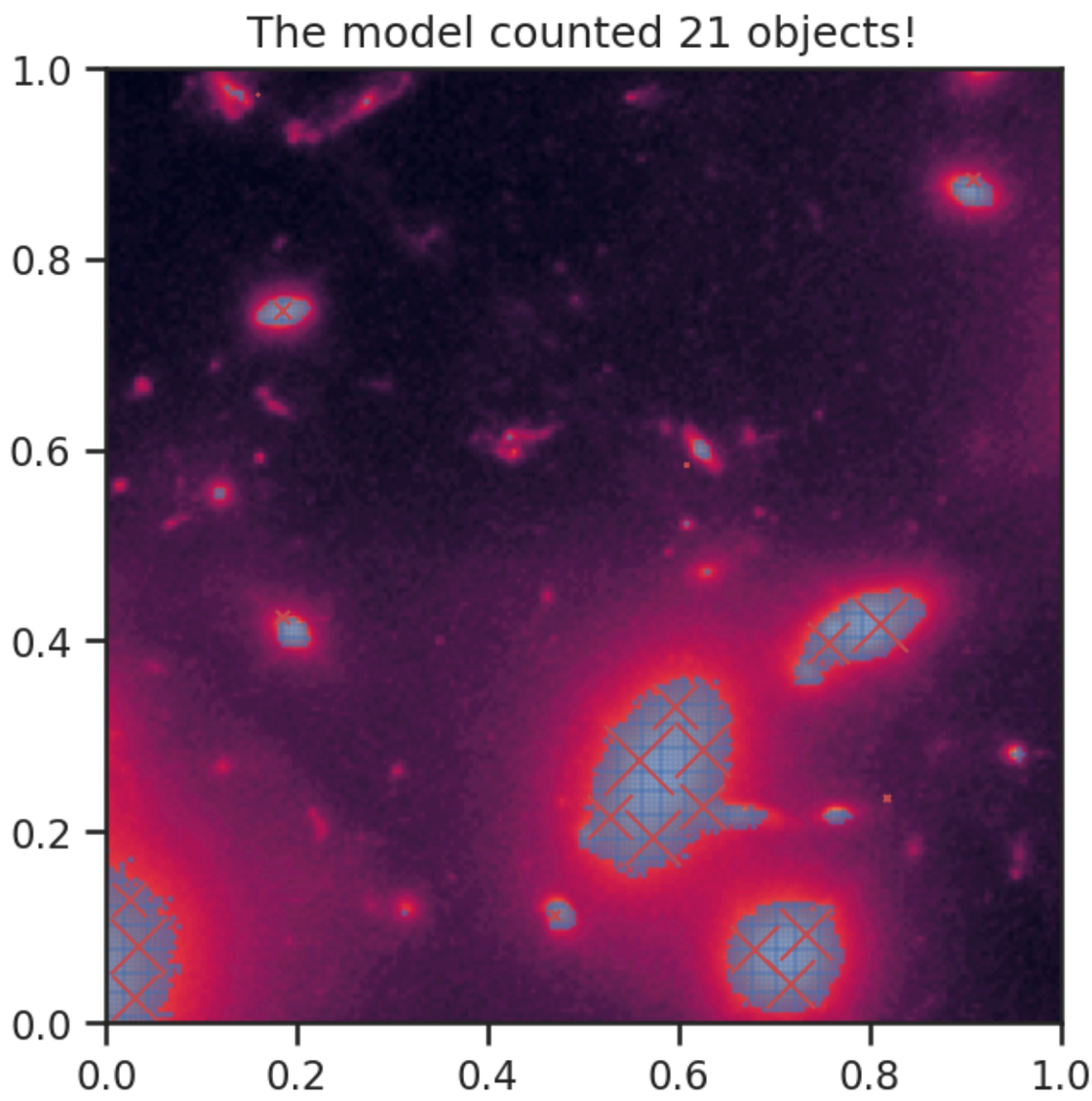
```
img = hubble_image.convert('L').crop((200, 200, 400, 400))
# objs, model, locs = count_objs(img, alpha=.1, beta=250)
# visualize_counts(img, objs, model, locs)

# Define parameter ranges including the parameter provided
alpha_all = [0.1, 0.5, 0.9]
beta_all = [150, 200, 250]

# Loop through alpha and beta values, perform object counting, and plot
for i, alpha_value in enumerate(alpha_all):
    for j, beta_value in enumerate(beta_all):
        objs, model, locs = count_objs(img, alpha_value, beta_value)
        visualize_counts(img, objs, model, locs)

# Adjust layout and display the stacked images
plt.tight_layout()
plt.show()
```

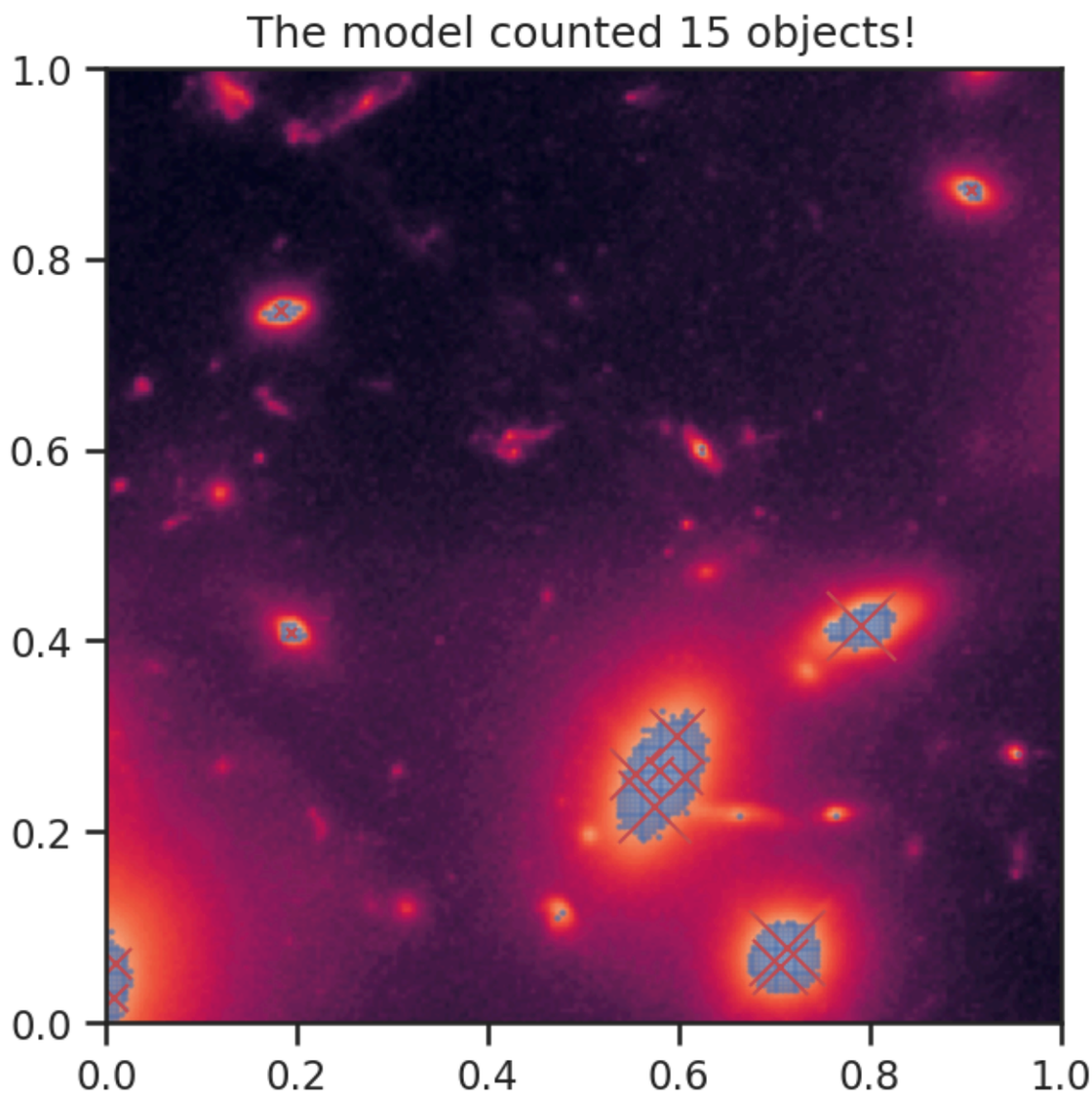
```
Best number of components: 17 for alpha = 0.1 & beta = 150
Best number of components: 14 for alpha = 0.1 & beta = 200
Best number of components: 7 for alpha = 0.1 & beta = 250
Best number of components: 21 for alpha = 0.5 & beta = 150
Best number of components: 15 for alpha = 0.5 & beta = 200
Best number of components: 4 for alpha = 0.5 & beta = 250
Best number of components: 25 for alpha = 0.9 & beta = 150
Best number of components: 15 for alpha = 0.9 & beta = 200
Best number of components: 3 for alpha = 0.9 & beta = 250
```

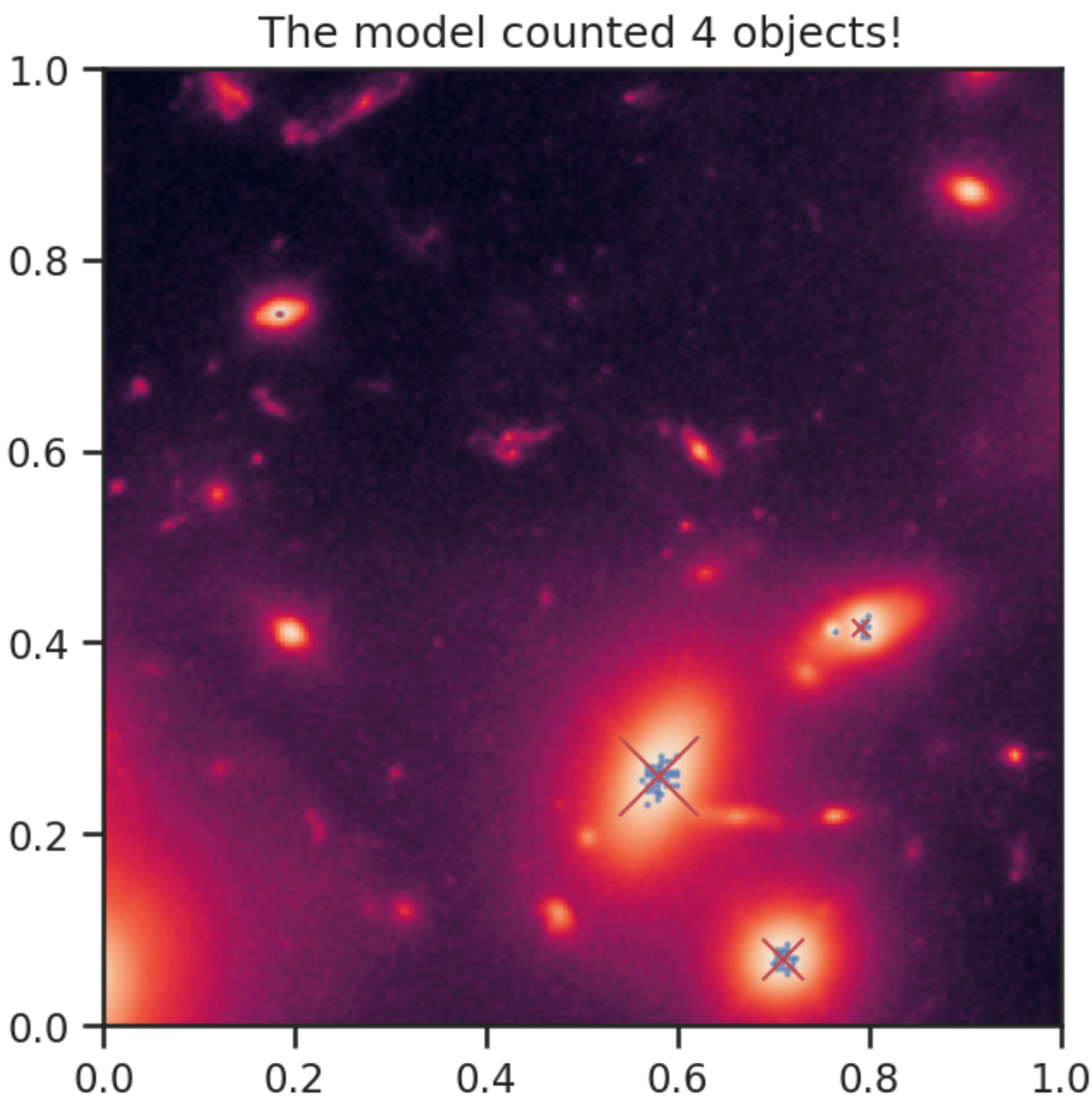

The model counted 17 objects!
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

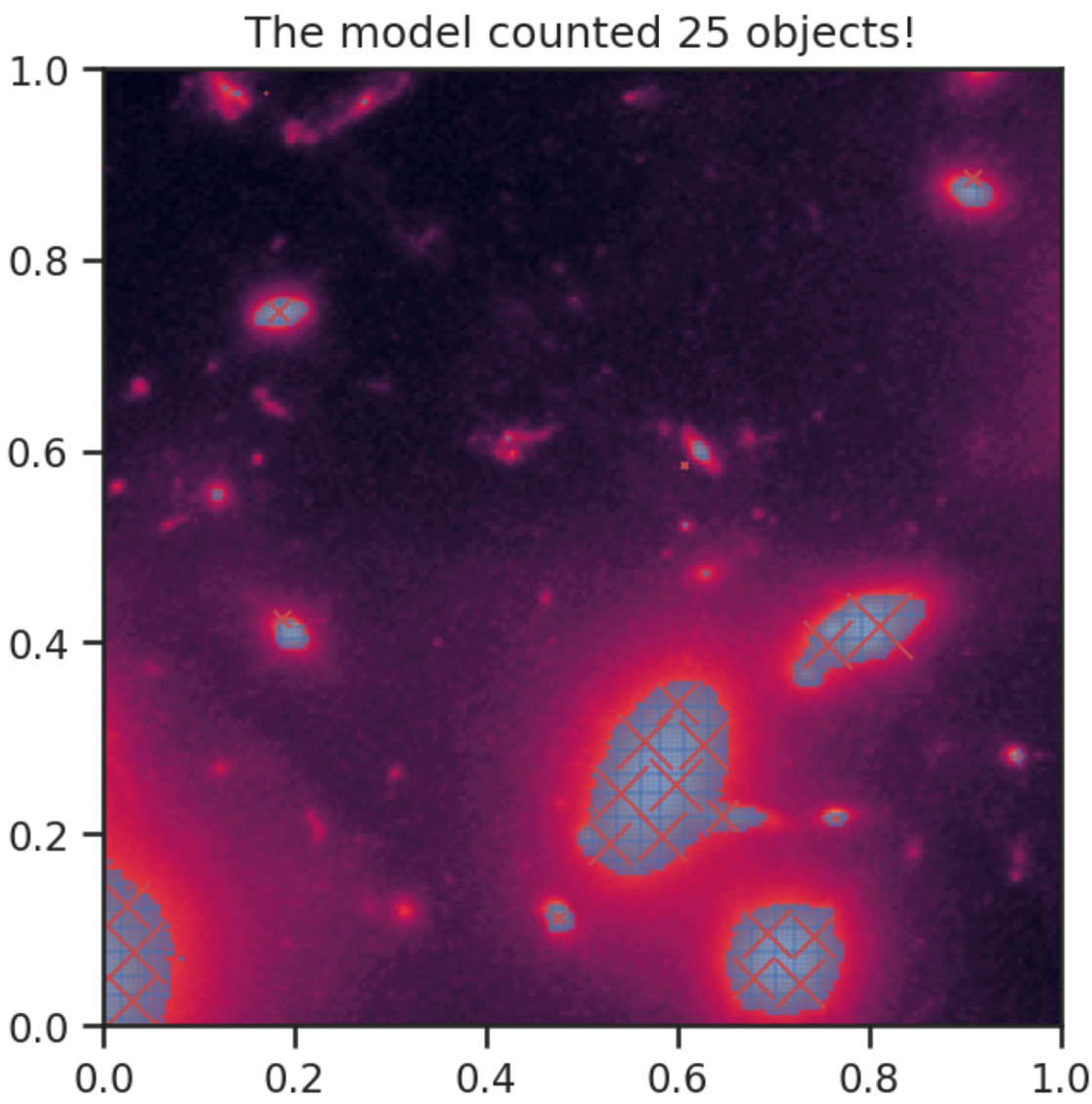

The model counted 14 objects!
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

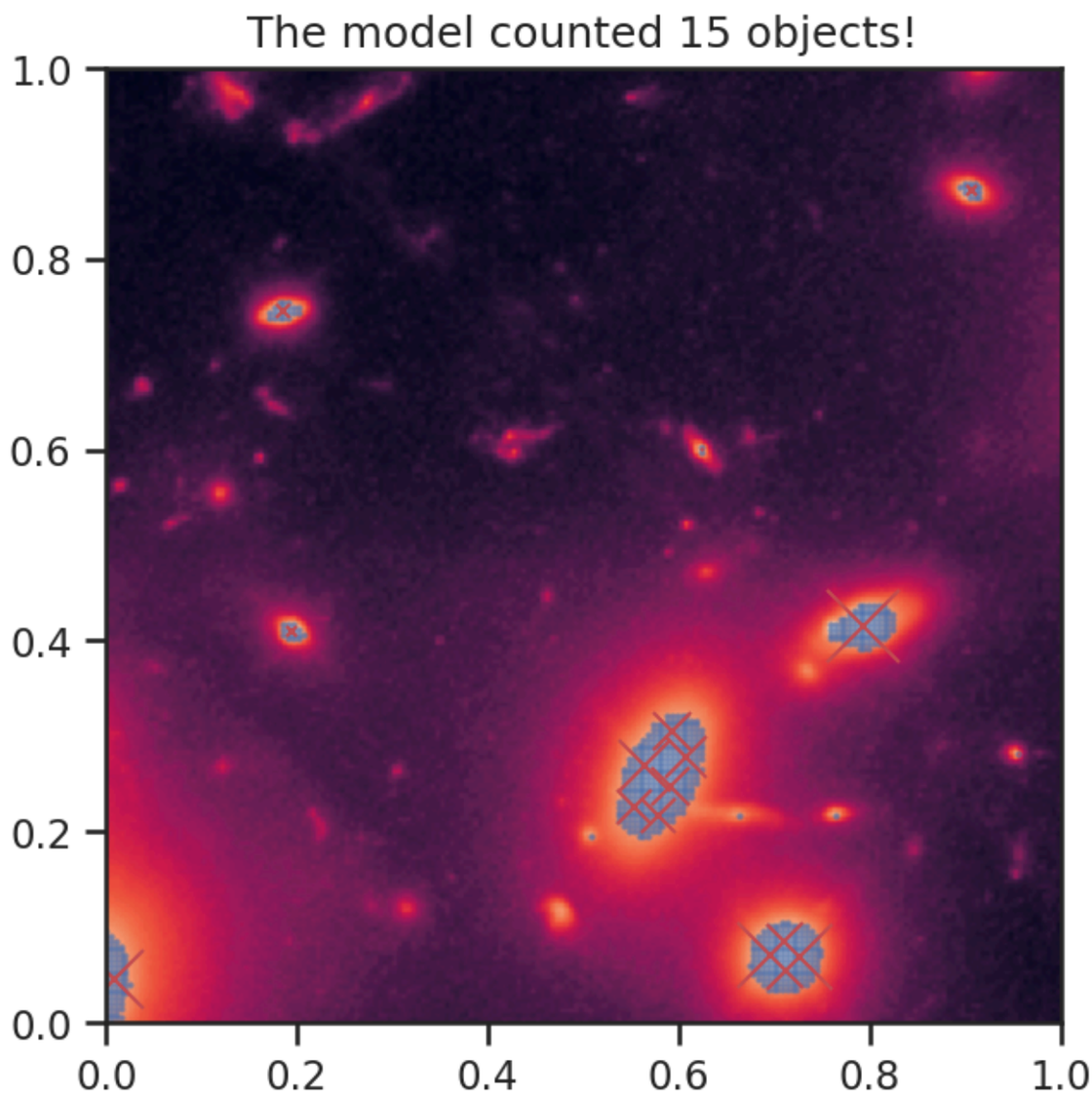

The model counted 7 objects!
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

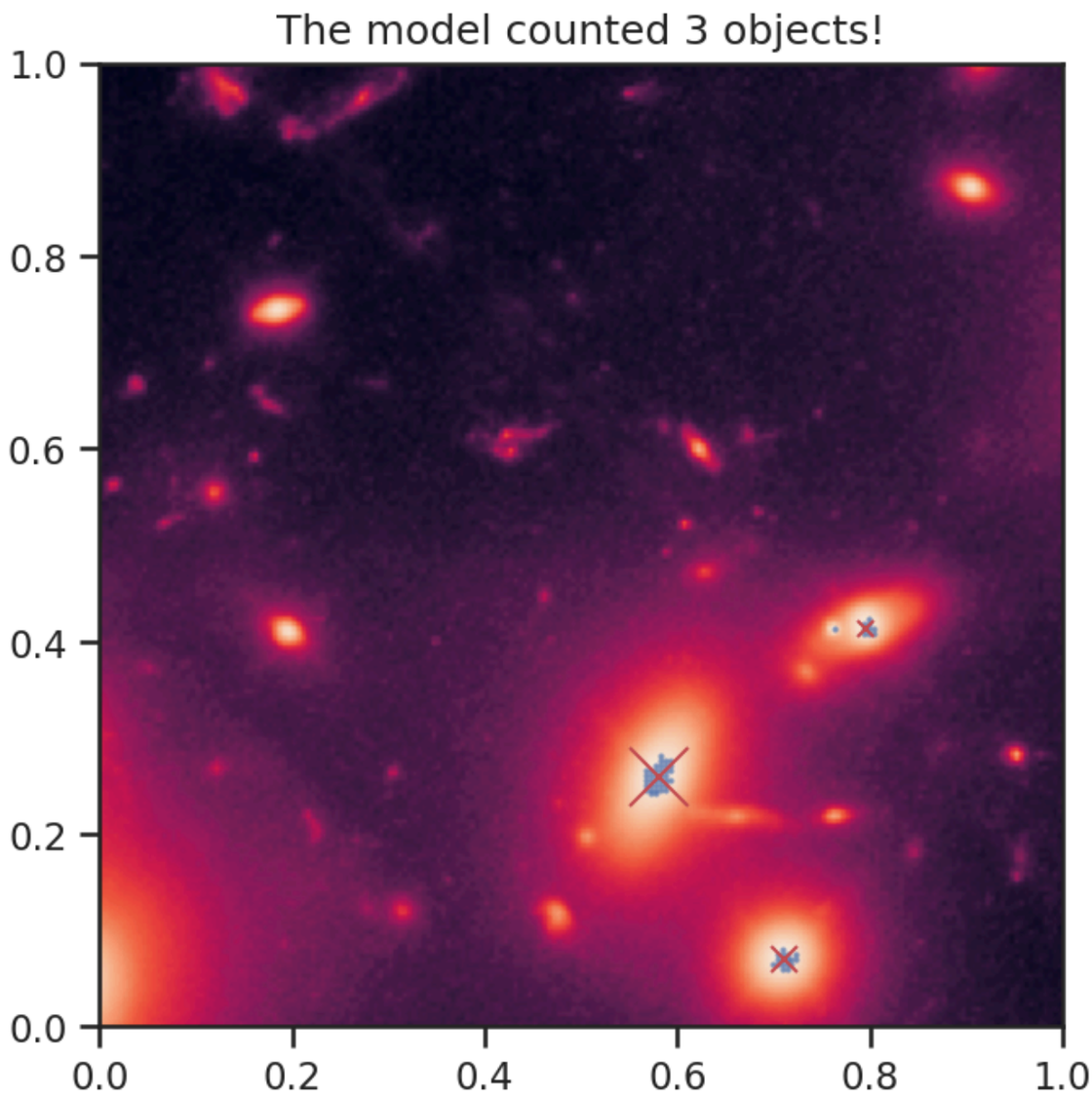

The model counted 21 objects!
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


The model counted 15 objects!
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


The model counted 4 objects!
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


The model counted 25 objects!
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


The model counted 15 objects!
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

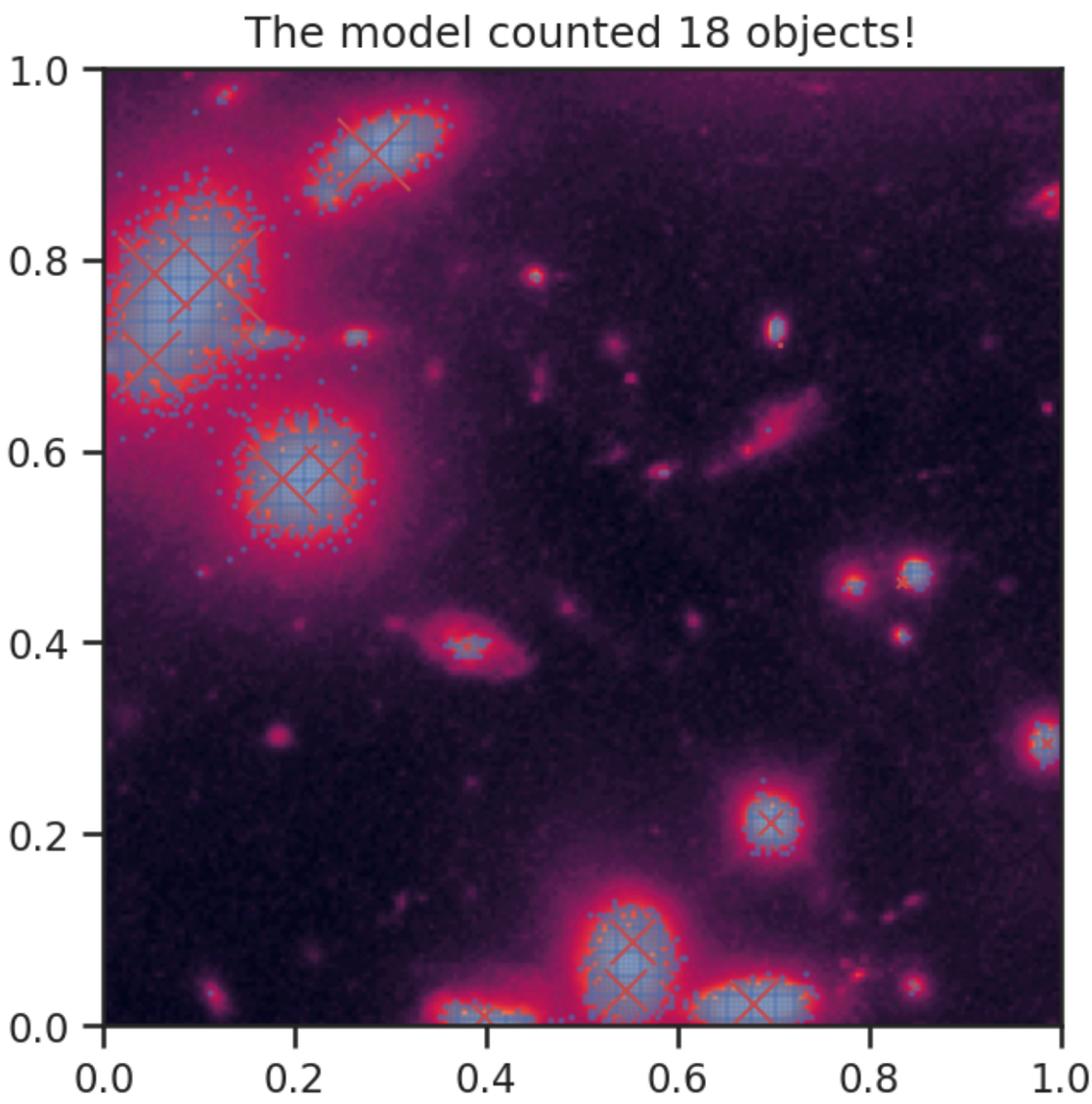


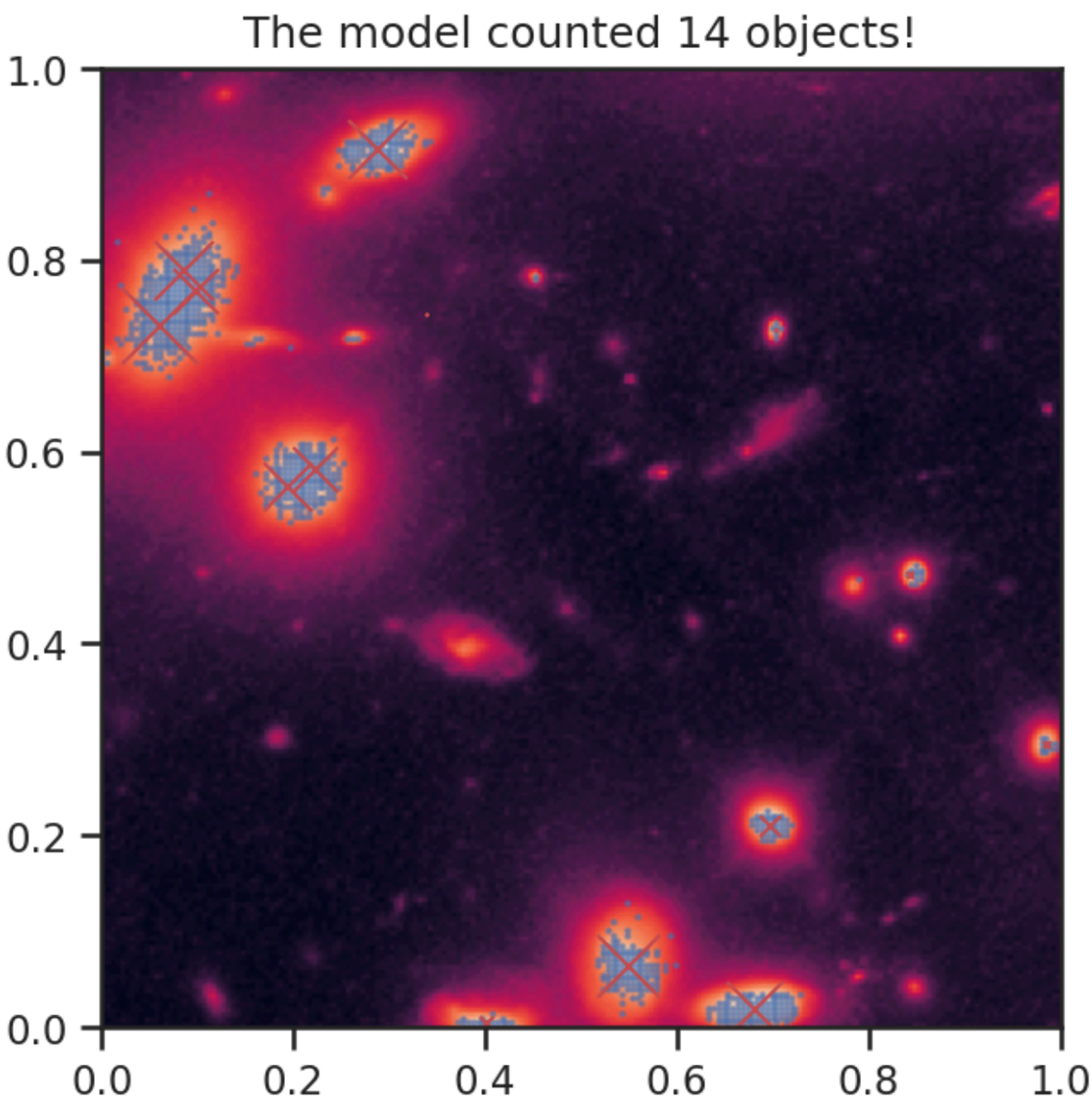
And this one:

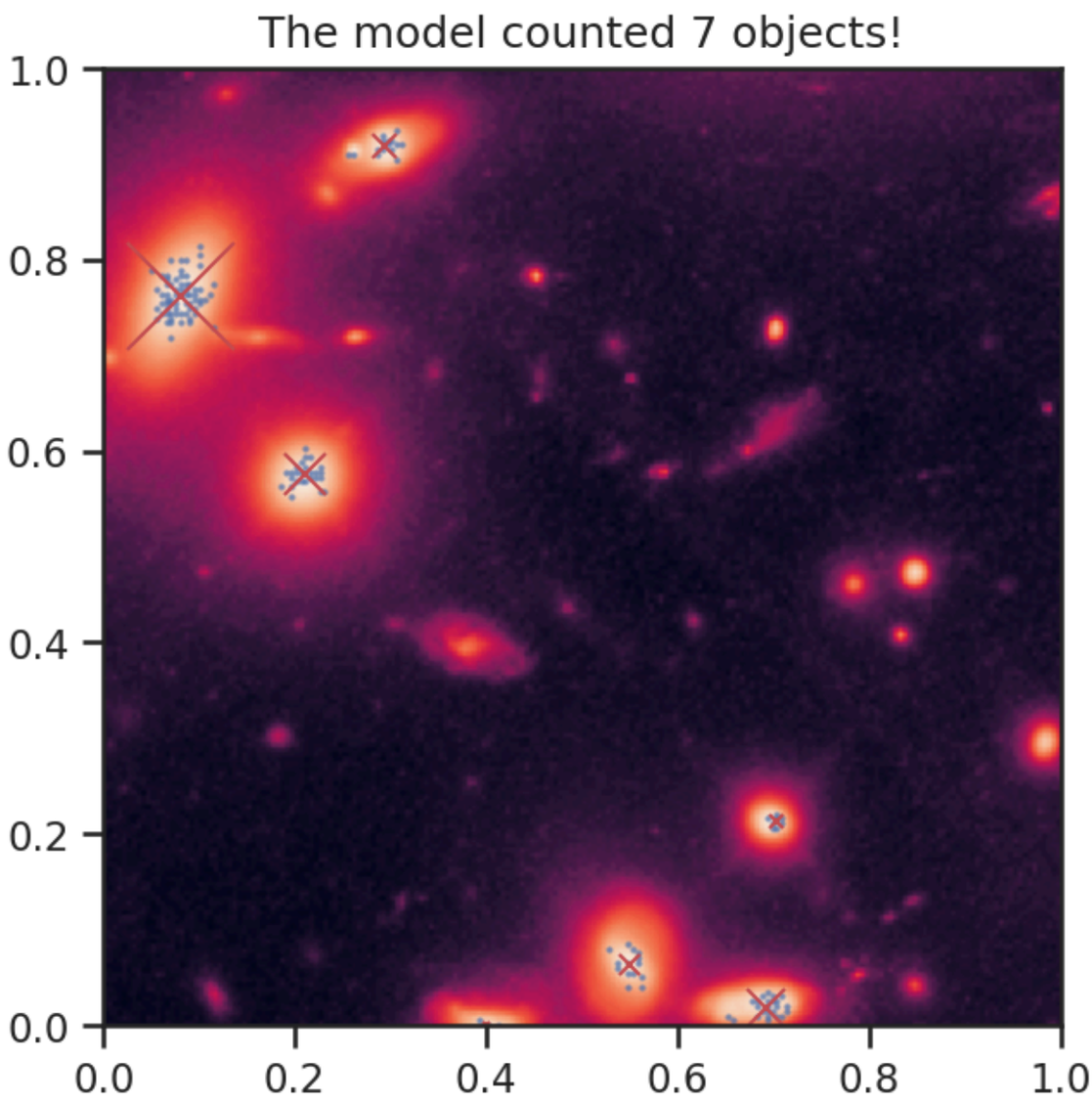
Since ask is to play with α and β without making a perfect model beyond gaussian mixture model, α and β values are systematically varied and their object detection perfoemance evaluated each case.

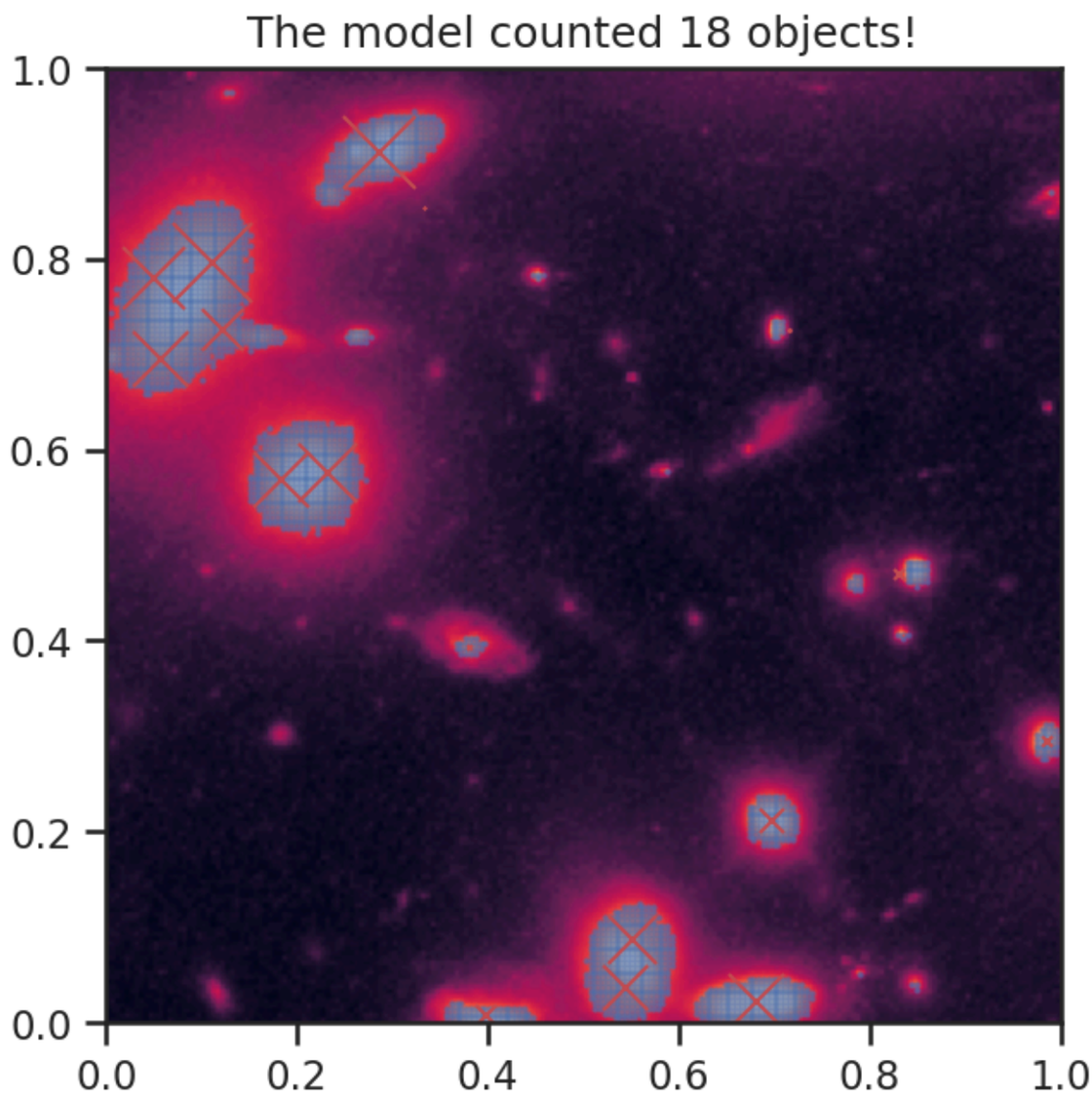
In [25]:
```
img = hubble_image.convert('L').crop((300, 300, 500, 500))
# objs, model, locs = count_objs(img, alpha=.1, beta=250)
# visualize_counts(img, objs, model, locs)

# Define parameter range including the parameter provided
alpha_all = [0.1, 0.5, 0.9]
beta_all = [150, 200, 250]

# Loop through alpha and beta values, perform object counting, and plot
for i, alpha_value in enumerate(alpha_all):
    for j, beta_value in enumerate(beta_all):
        objs, model, locs = count_objs(img, alpha_value, beta_value)
        visualize_counts(img, objs, model, locs)

# Adjust layout and display the stacked images
plt.tight_layout()
plt.show()
```

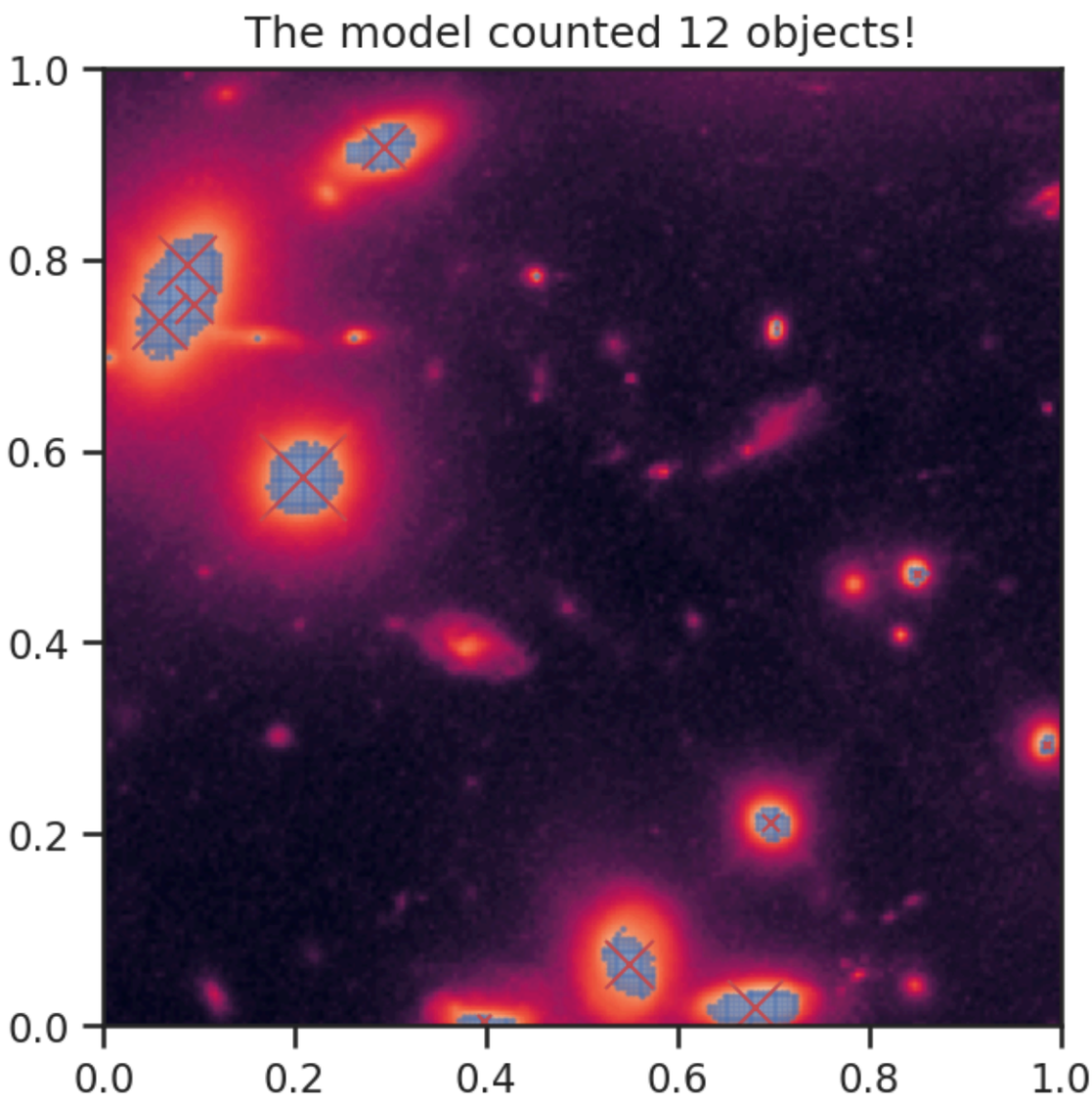
```
Best number of components: 18 for alpha = 0.1 & beta = 150
Best number of components: 14 for alpha = 0.1 & beta = 200
Best number of components: 7 for alpha = 0.1 & beta = 250
Best number of components: 18 for alpha = 0.5 & beta = 150
Best number of components: 12 for alpha = 0.5 & beta = 200
Best number of components: 6 for alpha = 0.5 & beta = 250
Best number of components: 15 for alpha = 0.9 & beta = 150
Best number of components: 14 for alpha = 0.9 & beta = 200
Best number of components: 5 for alpha = 0.9 & beta = 250
```

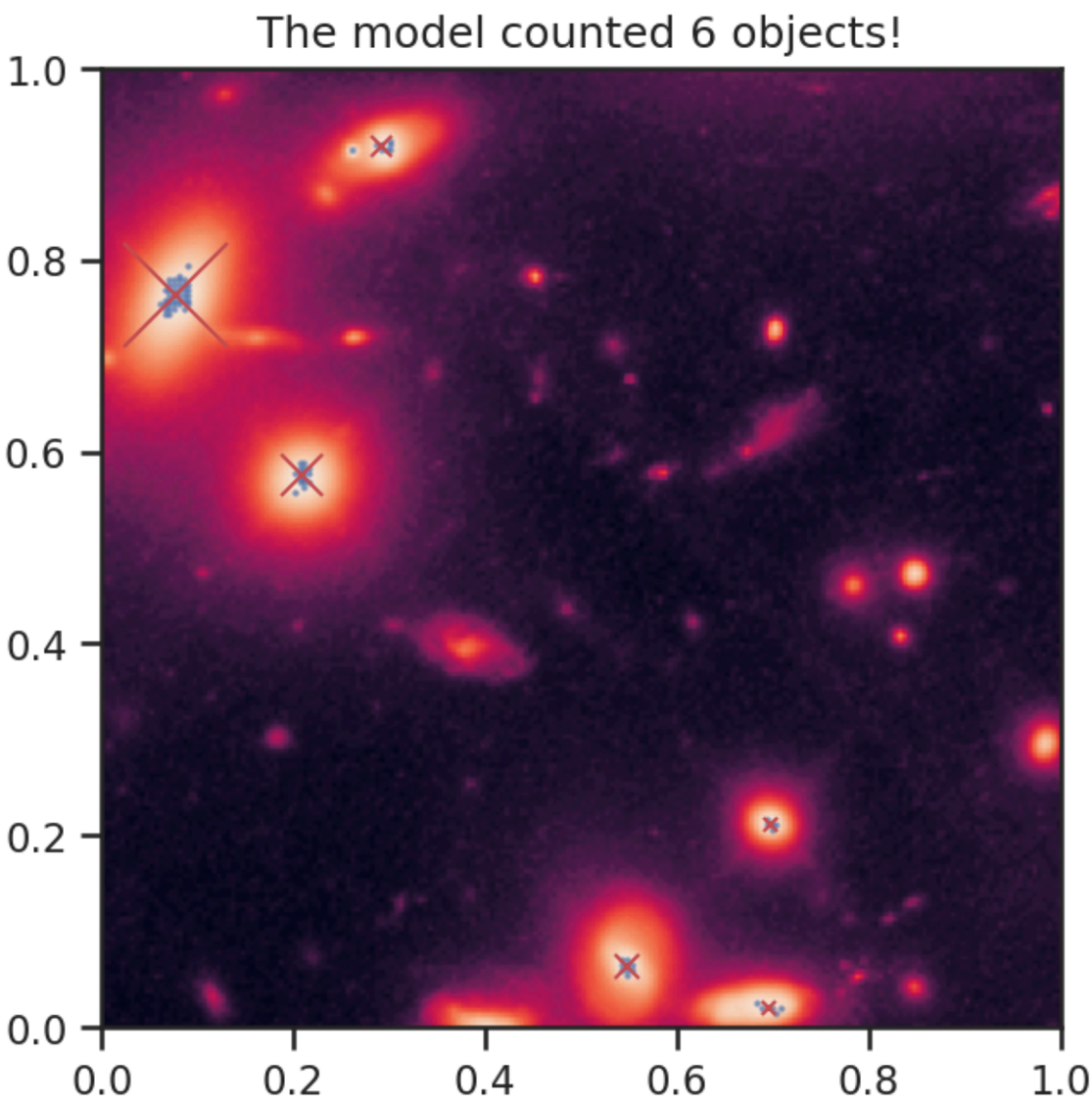

The model counted 18 objects!
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

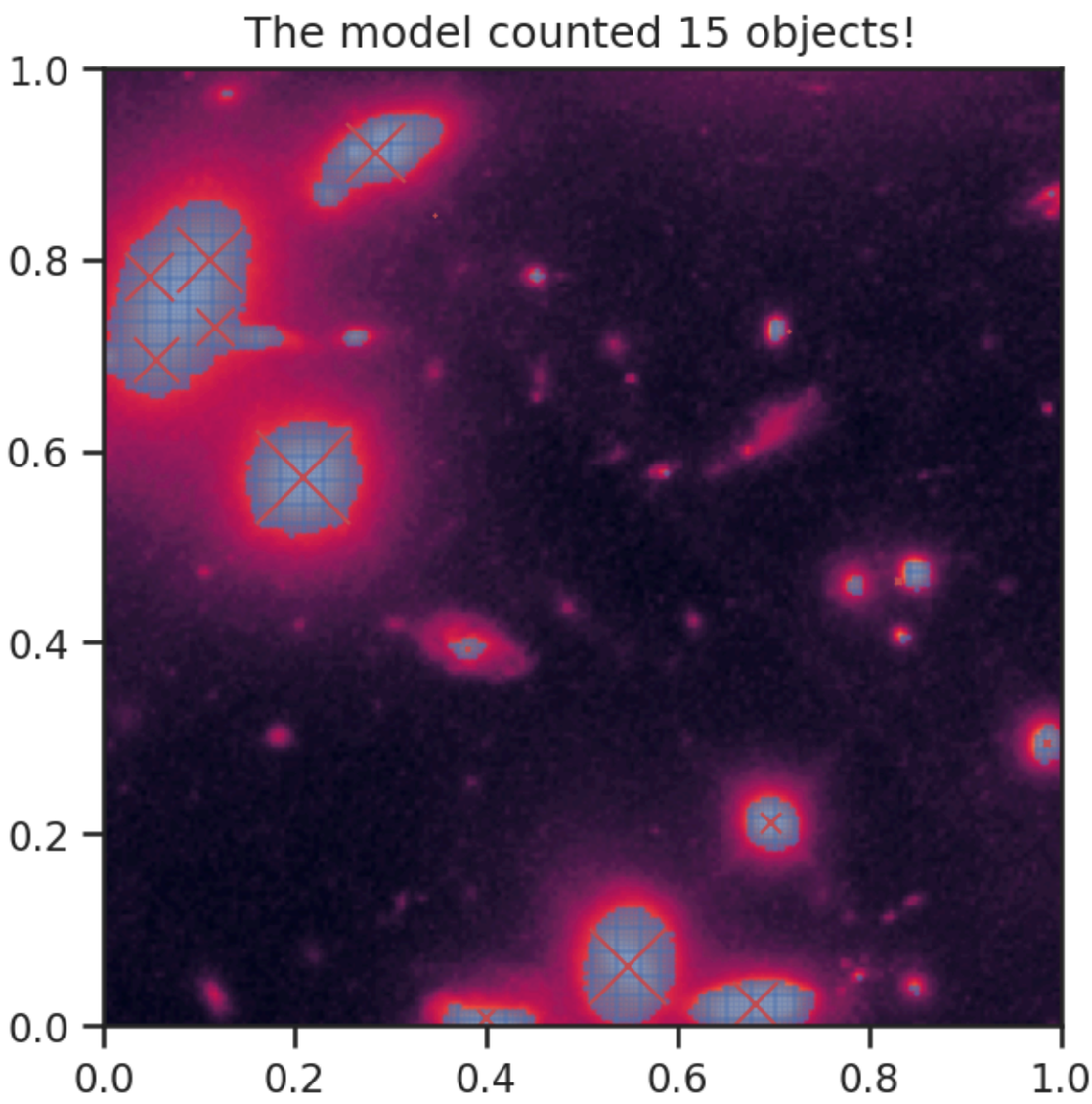

The model counted 14 objects!
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

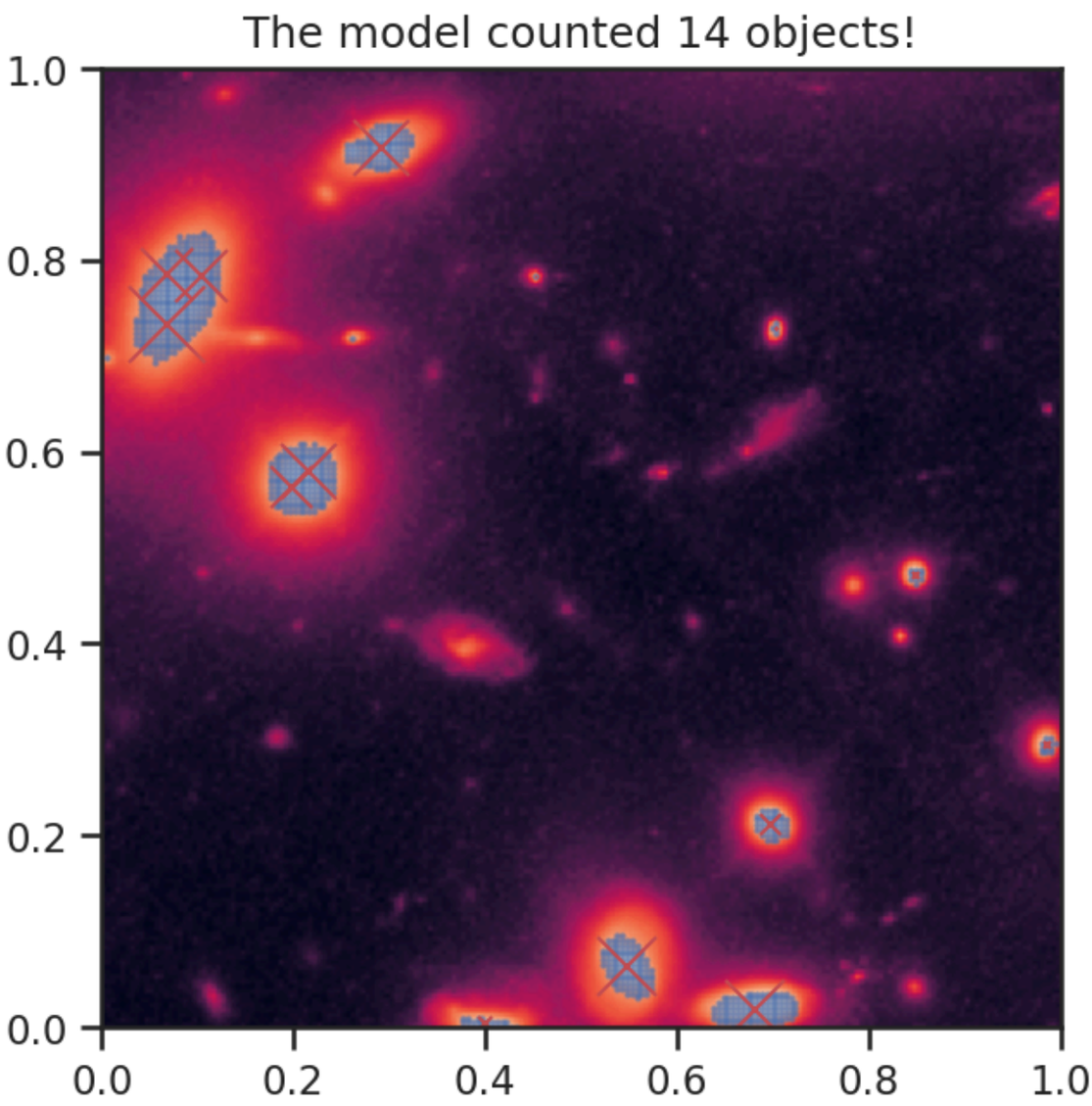

The model counted 7 objects!
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

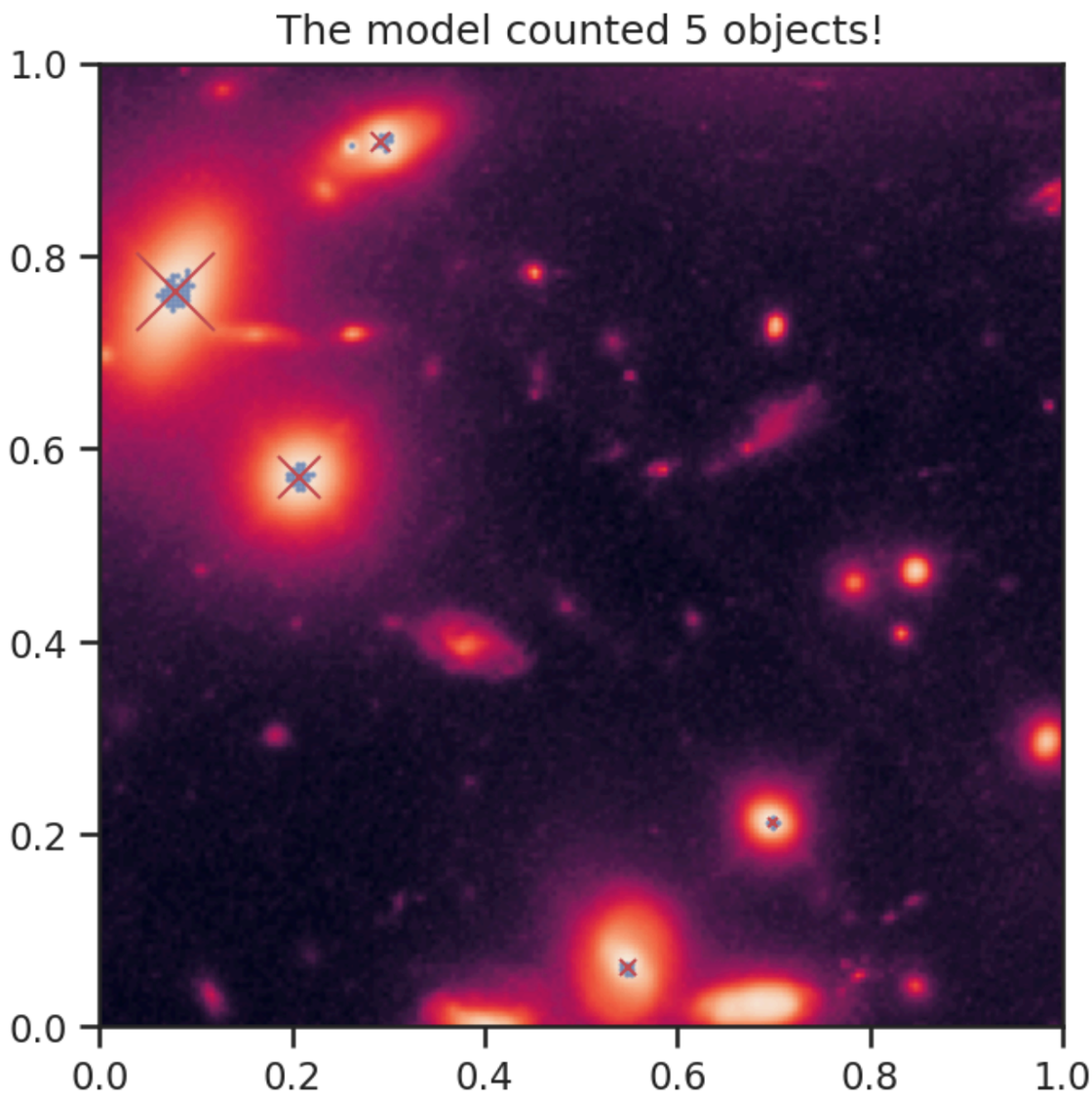

The model counted 18 objects!
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


The model counted 12 objects!
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


The model counted 6 objects!
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


The model counted 15 objects!
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


The model counted 14 objects!
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0



Problem 3 - Filtering of an Oscillator with Damping

Assume that you are dealing with a one-degree-of-freedom system which follows the equation:

$$\ddot{x} + 2\zeta\omega_0\dot{x} + \omega_0^2 x = u_0 \cos(\omega t),$$

where $x = x(t)$ is the generalized coordinate of the oscillator at time t, and the parameters ζ, ω_0, u_0, and ω are known to you (we will give them specific values later). Furthermore, assume that you are making noisy observations of the *absolute acceleration* at discrete timesteps Δt (also known):

$$y_j = \ddot{x}(j\Delta t) - u_0 \cos(\omega t) + w_j,$$

for $j = 1, \ldots, n$, where $w_j \sim \mathcal{N}(0, \sigma^2)$ with σ^2 also known. Finally, assume that the initial conditions for the position and the velocity (you need both to get a unique solution) are given by:

$$x_0 = x(0) \sim \mathcal{N}(0, \sigma_x^2),$$

and

$$v_0 = \dot{x} \sim \mathcal{N}(0, \sigma_v^2).$$

Of course, assume that σ_x^2 and σ_v^2 are specific numbers we will specify below.

Before we go over the questions, let's write code that generates the actual trajectory of the system at some random initial conditions and some observations. We will use the code to generate a synthetic dataset with known ground truth, which you will use in your filtering analysis.

The first step we need to do is to turn the problem into a first-order differential equation. We set:

$$\mathbf{x} = \begin{bmatrix} x \\ \dot{x} \end{bmatrix}.$$

Assuming $\mathbf{x} = (x_1, x_2)$, then the dynamics are described by:

$$\dot{\mathbf{x}} = \begin{bmatrix} \dot{x} \\ \ddot{x} \end{bmatrix} = \begin{bmatrix} x_2 \\ -2\zeta\omega_0\dot{x} - \omega_0^2 x + u_0\cos(\omega t) \end{bmatrix} = \begin{bmatrix} x_2 \\ -2\zeta\omega_0 x_2 - \omega_0^2 x_1 + u_0\cos(\omega t) \end{bmatrix}$$

The initial conditions are of course, just:

$$\mathbf{x}_0 = \begin{bmatrix} x_0 \\ v_0 \end{bmatrix}.$$

This first-order system can solved using scipy.integrate.solve_ivp. Here is how:

In [26]:
```
from scipy.integrate import solve_ivp

# You need to define the right hand side of the equation
def rhs(t, x, omega0, zeta, u0, omega):
    """Return the right hand side of the dynamical system.

    Arguments
    t       -    Time
    x       -    The state
    omega0  -    Natural frequency
    zeta    -    Dumping factor (0<=zeta)
    u0      -    External force amplitude
    omega   -    Excitation frequency
    """
    res = np.ndarray((2,))
    res[0] = x[1]
    res[1] = -2.0 * zeta * omega0 * x[1] - omega0 ** 2 * x[0] + u0 * np.cos(omega * t)
    return res
```

And here is how you solve it for given initial conditions and parameters:

In [27]:
```
# Initial conditions
x0 = np.array([0.0, 1.0])
# Natural frequency
omega0 = 2.0
# Dumping factor
zeta = 0.4
# External forcing amplitude
u0 = 0.5
# Excitation frequency
omega = 2.1
# Timestep
dt = 0.1
# The final time
final_time = 10.0
# The number of timesteps to get the final time
n_steps = int(final_time / dt)
# The times on which you want the solution
t_eval = np.linspace(0, final_time, n_steps)
# The solution
sol = solve_ivp(rhs, (0, final_time), x0, t_eval=t_eval, args=(omega0, zeta, u0, omega)
```

The solution is stored in the `sol` variable:

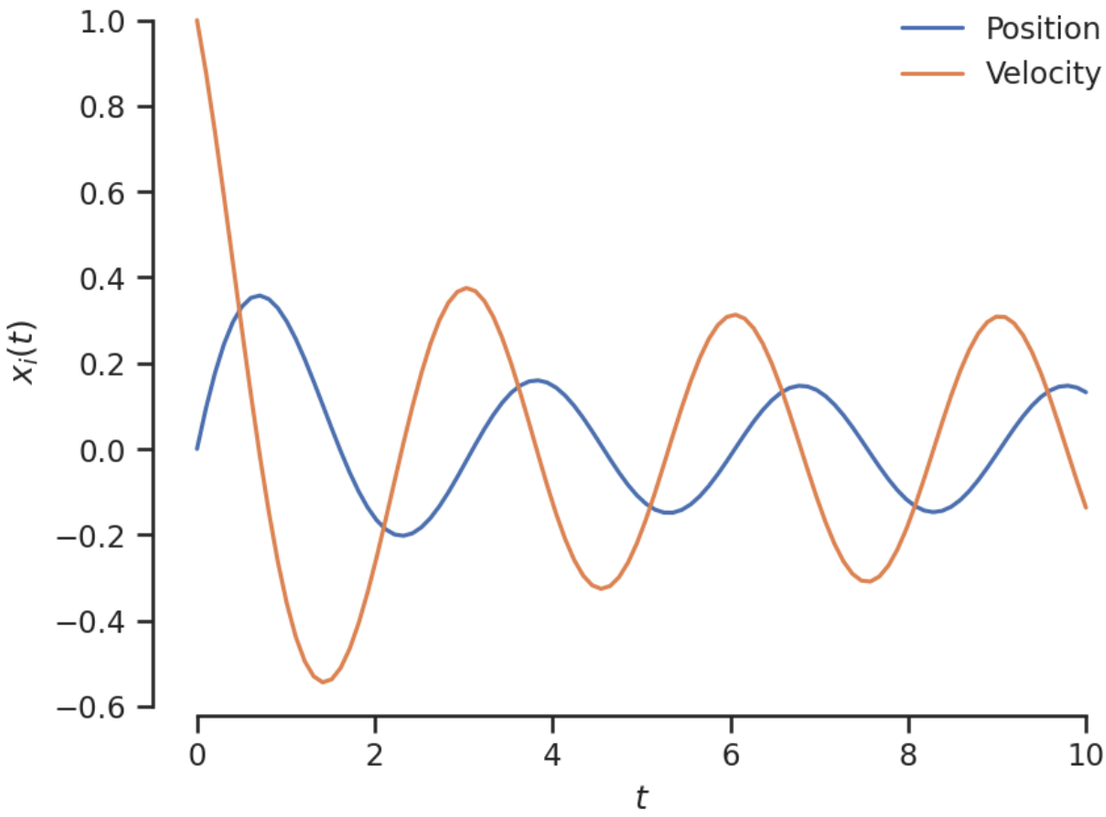
In [28]:
```
sol.y.shape
```

Out[28]:
```
(2, 100)
```

The shape of `sol.y` is (2, 100), which means that we have 100 timesteps and two variables (position and velocity). Let's plot the position and the velocity:

In [29]:
```
fig, ax = plt.subplots(dpi=150)
ax.plot(t_eval, sol.y[0, :], label='Position')
ax.plot(t_eval, sol.y[1, :], label='Velocity')
ax.set_xlabel('$t$')
ax.set_ylabel('$x_i(t)$')
```

```
plt.legend(loc='best', frameon=False)
sns.despine(trim=True);
```

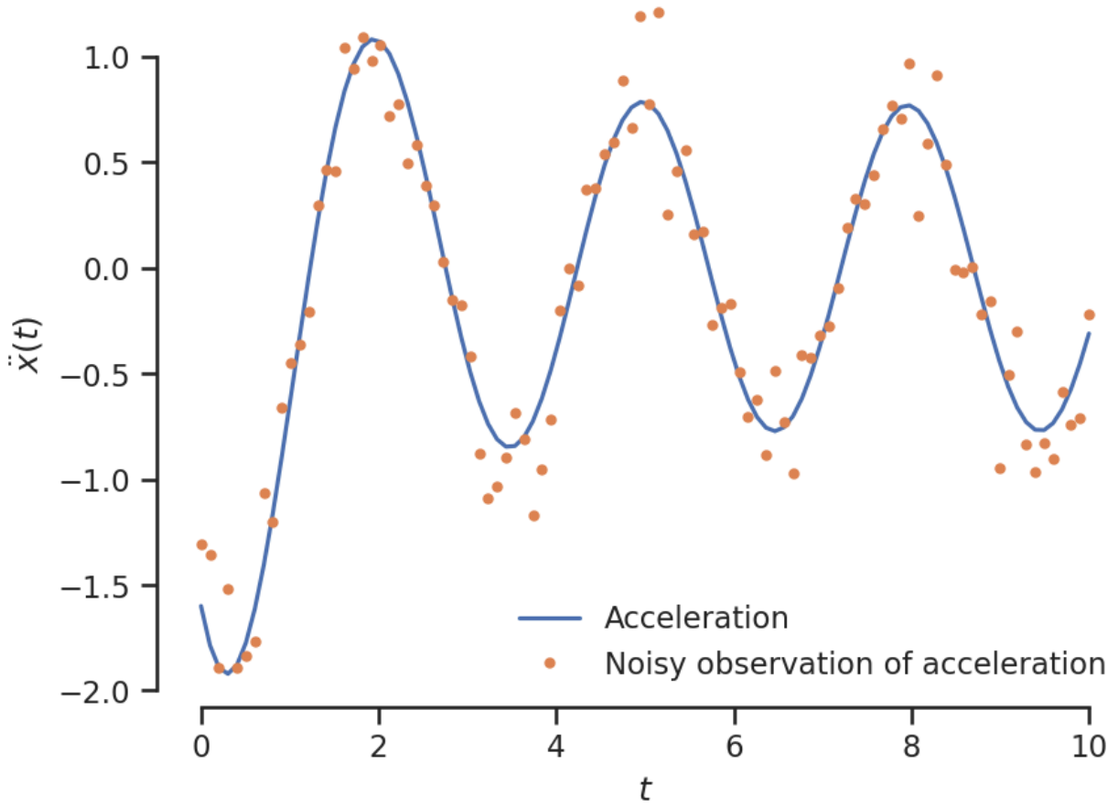


Let's now generate some synthetic observations of the acceleration with some given Gaussian noise. To get the acceleration, you can do this:

In [30]:
```
# Compute external excitation.
us = u0 * np.cos(omega * t_eval)
# Subtract us from \ddot{x}
true_acc = np.array([rhs(t, x, omega0, zeta, u0, omega)[1] for (t, x) in zip(t_eval, so
```

Let's add some noise:

In [31]:
```
sigma_r = 0.2
observations = true_acc + sigma_r * np.random.randn(true_acc.shape[0])

fig, ax = plt.subplots(dpi=150)
ax.plot(t_eval, true_acc, label='Acceleration')
ax.plot(
    t_eval,
    observations,
    '.',
    label='Noisy observation of acceleration'
)
ax.set_xlabel('$t$')
ax.set_ylabel(r'$\ddot{x}(t)$')
plt.legend(loc='best', frameon=False)
sns.despine(trim=True);
```



Okay. Now, imagine that you only see the noisy observations of the acceleration. The filtering goal is to recover the state of the underlying system (as well as its acceleration). I am going to guide you through the steps you need to follow.

Part A - Discretize time (Transitions)

Use the Euler time discretization scheme to turn the continuous dynamical system into a discrete-time dynamical system like this:

$$\mathbf{x}_{j+1} = \mathbf{A}\mathbf{x}_j + \mathbf{B}u_j + \mathbf{z}_j,$$

where

$$\mathbf{x}_j = \mathbf{x}(j\Delta t),$$

$$u_j = u(j\Delta t),$$

and $\mathbf{z}_j$ is properly chosen process noise term. You should derive and provide mathematical expressions for the following:

- The 2×2 transition matrix $\mathbf{A}$.
- The 2×1 control "matrix" $\mathbf{B}$.
- The process covariance $\mathbf{Q}$. For the process covariance, you may choose your values by hand.

Answer:

The mathematical derivation for the discrete-time transition matrix A, the control matrix B, and the process noise covariance Q is provided below. Provided already, the first-order system: The dynamic system equation provided already as:

$$\dot{\mathbf{x}} = \begin{bmatrix} \dot{x} \\ \ddot{x} \end{bmatrix} = \begin{bmatrix} x_2 \\ -2\zeta\omega_0\dot{x} - \omega_0^2 x + u_0\cos(\omega t) \end{bmatrix} = \begin{bmatrix} x_2 \\ -2\zeta\omega_0 x_2 - \omega_0^2 x_1 + u_0\cos(\omega t) \end{bmatrix}$$

Using the Euler time discretization scheme:

$$\mathbf{x}(t+\Delta t) \approx \mathbf{x}(t) + \Delta t \cdot \dot{\mathbf{x}}(t)$$

Transition Matrix A:

For the transition matrix A,

$$\dot{\mathbf{x}} = \mathbf{A}_{\text{cont}}\mathbf{x} + \mathbf{B}_{\text{cont}}u(t)$$

where

$$\mathbf{A}_{\text{cont}} = \begin{bmatrix} 0 & 1 \\ -\omega_0^2 & -2\zeta\omega_0 \end{bmatrix} \mathbf{B}_{\text{cont}} = \begin{bmatrix} 0 \\ u_0 \end{bmatrix}$$

Using the Euler discretization:

$$\mathbf{x}(t+\Delta t) \approx \mathbf{x}(t) + \Delta t \cdot (\mathbf{A}_{\text{cont}}\mathbf{x}(t) + \mathbf{B}_{\text{cont}}u(t))$$

Therefore, matrix A is:

$$\mathbf{A} = \mathbf{I} + \Delta t \cdot \mathbf{A}_{\text{cont}}$$

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \Delta t \cdot \begin{bmatrix} 0 & 1 \\ -\omega_0^2 & -2\zeta\omega_0 \end{bmatrix}$$

Final,

$$\mathbf{A} = \begin{bmatrix} 1 & \Delta t \\ -\Delta t \cdot \omega_0^2 & 1 - \Delta t \cdot 2\zeta\omega_0 \end{bmatrix}$$

Control Matrix B:

The control matrix B in the discrete-time system:

$$\mathbf{B} = \Delta t \cdot \mathbf{B}_{\text{cont}}$$

$$\mathbf{B} = \Delta t \cdot \begin{bmatrix} 0 \\ u_0 \end{bmatrix}$$

Final,

$$\mathbf{B} = \begin{bmatrix} 0 \\ \Delta t \cdot u_0 \end{bmatrix}$$

Process Covariance Q:

For simplicity, We choose a small diagonal matrix:

$$\mathbf{Q} = \begin{bmatrix} q_1 & 0 \\ 0 & q_2 \end{bmatrix}$$

example,

$$\mathbf{Q} = \begin{bmatrix} 1e-4 & 0 \\ 0 & 1e-4 \end{bmatrix}$$

In [32]:

```
# You should be using the parameters dt, omega0, zeta, etc.
# from above
A = np.array(
    [
        [1, dt],
        [-dt * omega0**2, 1 - dt * 2 * zeta * omega0]
    ]
)


B = np.array(
    [
        [0],
        [dt]
    ]
)


Q = np.array(
    [
        [1e-4, 0.0],
        [0.0, 1e-4]
    ]
)
```

In [33]:

```
print(f"The matrix A:\n{A}")
print(f"The matrix B:\n{B}")
print(f"The matrix Q:\n{Q}")
```

```
The matrix A:
[[ 1.    0.1 ]
 [-0.4   0.84]]
The matrix B:
[[0. ]
 [0.1]]
The matrix Q:
```

```
[[0. 0.]
 [0. 0.]]
```

Part B - Discretize time (Emissions)

Establish the map that takes you from the states to the accelerations at each timestep. That is, specify:

$$y_j = \mathbf{C}\mathbf{x}_j + w_j,$$

where

$$y_j = \ddot{x}(j\Delta t) - u_0 \cos(\omega t) + w_j,$$

and w_j is a measurement noise. You should derive and provide mathematical expressions for the following:

- The 1×2 emission matrix $\mathbf{C}$.
- The 1×1 covariance "matrix" R of the measurement noise.

Answer:

In [34]:
```
C = np.array(
    [
        [-omega0**2, -2 * zeta * omega0]
    ]
)

R = np.array(
    [
        [sigma_r**2]
    ]
)
```

In [35]:
```
print(f"The matrix C:\n{C}")
print(f"The matrix R:\n{R}")
```

```
The matrix C:
[[-4.  -1.6]]
The matrix R:
[[0.04]]
```

Part C - Apply the Kalman filter

Use `FilterPy` (see the hands-on activity of Lecture 20) to infer the unobserved states given the noisy observations of the accelerations. Plot time-evolving 95% credible intervals for the position and the velocity along with the true unobserved values of these quantities (in two separate plots).

In [36]:
```
!pip install filterpy
```

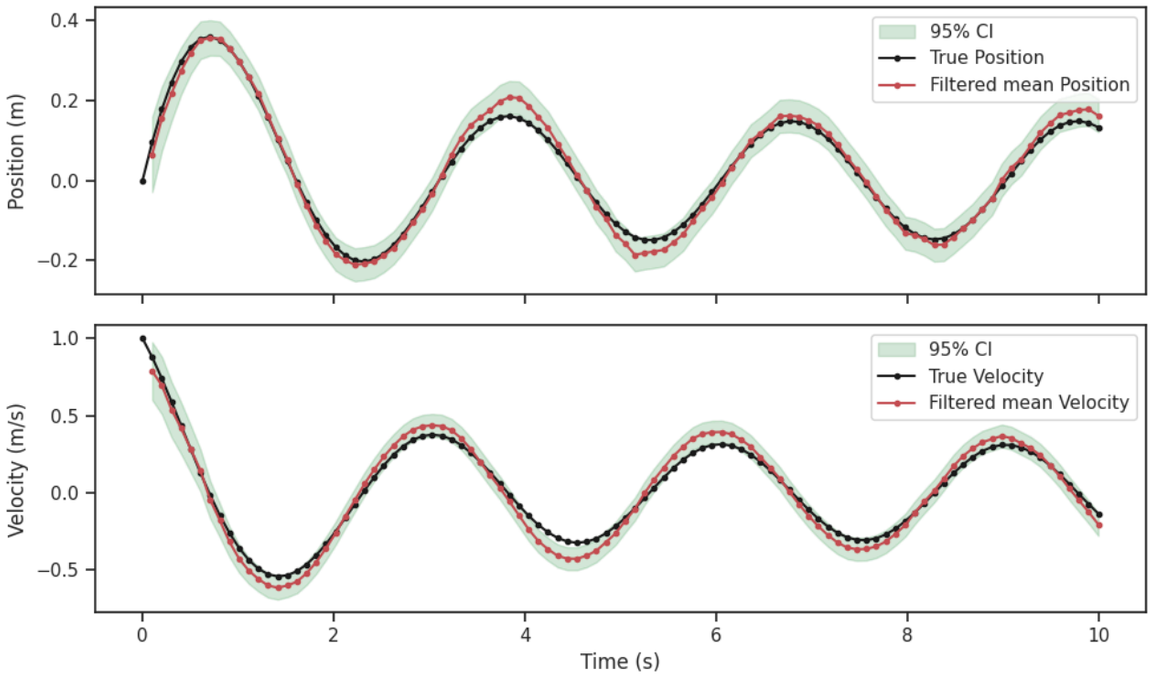
```
Requirement already satisfied: filterpy in /usr/local/lib/python3.10/dist-packages (1.4.
5)
Requirement already satisfied: numpy in /usr/local/lib/python3.10/dist-packages (from fi
lterpy) (1.25.2)
Requirement already satisfied: scipy in /usr/local/lib/python3.10/dist-packages (from fi
lterpy) (1.11.4)
Requirement already satisfied: matplotlib in /usr/local/lib/python3.10/dist-packages (fr
om filterpy) (3.7.1)
Requirement already satisfied: contourpy>=1.0.1 in /usr/local/lib/python3.10/dist-packag
es (from matplotlib->filterpy) (1.2.1)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.10/dist-packages
(from matplotlib->filterpy) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.10/dist-packa
ges (from matplotlib->filterpy) (4.53.1)
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.10/dist-packa
ges (from matplotlib->filterpy) (1.4.5)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.10/dist-package
s (from matplotlib->filterpy) (24.1)
Requirement already satisfied: pillow>=6.2.0 in /usr/local/lib/python3.10/dist-packages
```

```
(from matplotlib->filterpy) (9.4.0)
Requirement already satisfied: pyparsing>=2.3.1 in /usr/local/lib/python3.10/dist-packag
es (from matplotlib->filterpy) (3.1.2)
Requirement already satisfied: python-dateutil>=2.7 in /usr/local/lib/python3.10/dist-pa
ckages (from matplotlib->filterpy) (2.8.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-packages (from
python-dateutil>=2.7->matplotlib->filterpy) (1.16.0)
```

In [37]:

```
from filterpy.kalman import KalmanFilter

# Define Kalman filter
def run_kalman_filter(x0, Q, R, C, A, B, observations, us):
    """Run the Kalman filter with given parameters."""
    kf = KalmanFilter(dim_x=2, dim_z=1)

    # Initial state vector mean
    x = x0.reshape(2, 1)

    # Initial position and velocity stdevs
    sigma_x = 0.1
    sigma_v = 0.1

    # Initial state vector covariance
    P = np.array([[sigma_x**2, 0],
                  [0, sigma_v**2]])

    # Set Kalman filter parameters
    kf.x = x
    kf.P = P
    kf.Q = Q
    kf.R = R
    kf.H = C
    kf.F = A
    kf.B = B

    # Batch filtering using Kalman's filter
    means, covs, _, _ = kf.batch_filter(observations, us=us)

    return means, covs

# Define plotting confidence intervals
def plot_confidence_intervals(ax, t_eval, means, covs, index, color='red', alpha=0.25,
    """Plot confidence intervals for the given axis and data."""
    ax.fill_between(
        t_eval[1:],  # X-axis values
        means[1:, index, 0] - 1.96 * np.sqrt(covs[1:, index, index]),  # Lower bound of
        means[1:, index, 0] + 1.96 * np.sqrt(covs[1:, index, index]),  # Upper bound of
        color='g',
        alpha=alpha,
        label=label
    )

# Run Kalman filter
means, covs = run_kalman_filter(x0, Q, R, C, A, B, observations, us)

# Extracting position and velocity from the solution
position_index = 0
velocity_index = 1
true_pos = sol.y[position_index, :]
true_vel = sol.y[velocity_index, :]

# Create the figure and axes for subplots
fig, axes = plt.subplots(nrows=2, ncols=1, sharex=True, figsize=(10, 6))

# Indices and labels for the plots
indices = [position_index, velocity_index]
true_values = [true_pos, true_vel]
labels = ["Position (m)", "Velocity (m/s)"]

# Loop over indices to plot position and velocity
for i, index in enumerate(indices):
    plot_confidence_intervals(axes[i], t_eval, means, covs, index=index)
    axes[i].plot(t_eval, true_values[i], 'k.-', label=f"True {labels[i].split()[0]}")
    axes[i].plot(t_eval[1:], means[1:, index, 0], 'r.-', label=f"Filtered mean {labels[
    axes[i].set_ylabel(labels[i])
    axes[i].legend()
axes[-1].set_xlabel("Time (s)")
```

```
plt.tight_layout()
plt.show()
```



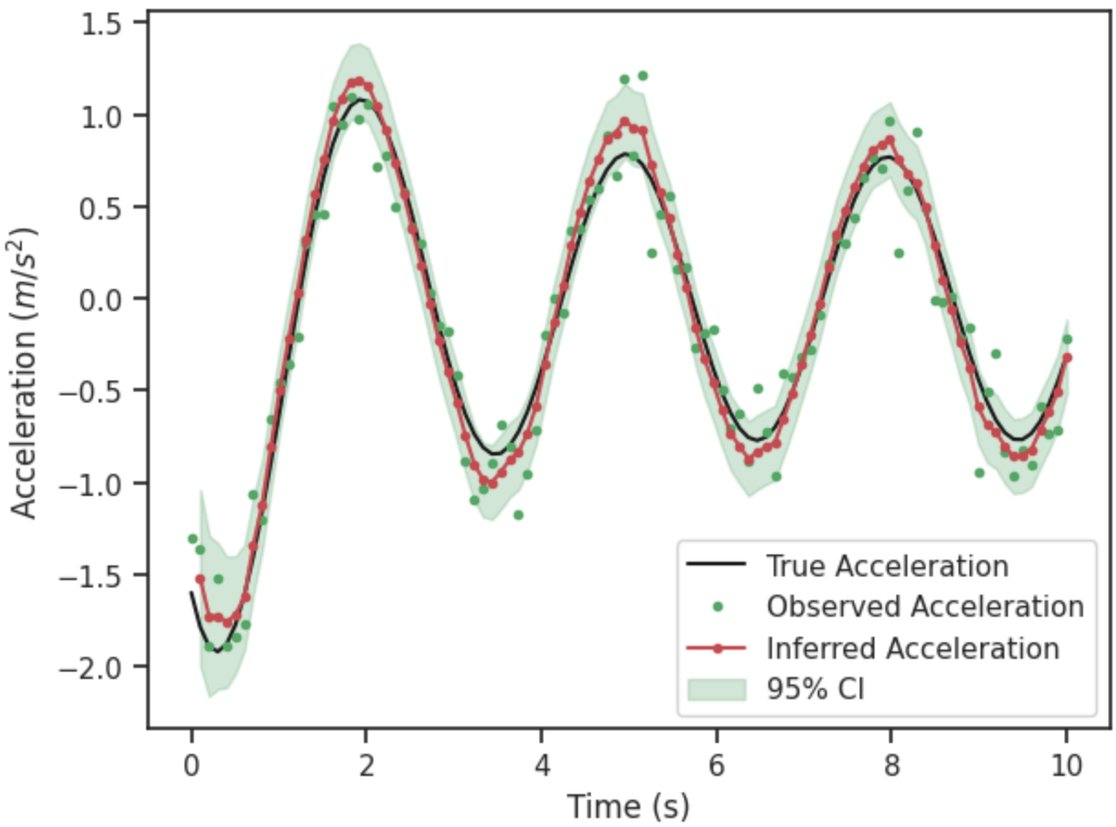
Part D - Quantify and visualize your uncertainty about the actual acceleration value

Use standard uncertainty propagation techniques to quantify your epistemic uncertainty about the true acceleration value. You will have to use the inferred states of the system and the dynamical model. This can be done either analytically or by Monte Carlo. It's your choice. In any case, plot time-evolving 95% credible intervals for the acceleration (epistemic only), the true unobserved values, and the noisy measurements.

In [38]:

```
# Solving Analytically
# Define Inferred acceleration
def compute_filtered_acceleration(C, means):

    # Apply the measurement matrix C to each state estimate in means
    inferred_acceleration = np.array([C @ mean for mean in means[1:]])
    return np.squeeze(inferred_acceleration)

# Define plotting confidence intervals
def plot_confidence_intervals(ax, t_eval, filtered_acceleration, accel_cov, alpha=0.25,

    ax.fill_between(
        t_eval[1:],  # X-axis values
        filtered_acceleration - 1.96 * np.sqrt(accel_cov[1:]),  # Lower bound of confid
        filtered_acceleration + 1.96 * np.sqrt(accel_cov[1:]),  # Upper bound of confid
        color='g',
        alpha=alpha,
        label=label,
    )

def plot_acceleration_data(t_eval, true_acc, observations, filtered_acceleration, accel
    fig, ax = plt.subplots()
    ax.plot(t_eval, true_acc, 'k', label="True Acceleration")
    ax.plot(t_eval, observations, 'g.', label="Observed Acceleration")
    ax.plot(t_eval[1:], filtered_acceleration, 'r.-', label="Inferred Acceleration")
    plot_confidence_intervals(ax, t_eval, filtered_acceleration, accel_cov)
    ax.set_ylabel("Acceleration ($m/s^2$)")
    ax.set_xlabel("Time (s)")
    ax.legend()
    plt.tight_layout()
    plt.show()

# Compute inferred acceleration and
inferred_acceleration = compute_filtered_acceleration(C, means)

# Linear combination of two Gaussian random variables
accel_cov = (-2 * zeta * omega0)**2 * covs[:, 1, 1]  + (-omega0**2)**2 * covs[:, 0, 0]
```

```
# Plot the acceleration data
plot_acceleration_data(t_eval, true_acc, observations, inferred_acceleration, accel_cov
```



End of Homework5!

In [39]:
```
# from google.colab import drive
# drive.mount('/content/drive')
```

In [40]:
```
# !apt-get update
# !sudo apt-get install inkscape texlive-xetex
```

In [41]:
```
# !sudo apt-get install pandoc
```

In [42]:
```
# !jupyter nbconvert --to pdf '/content/drive/MyDrive/Colab Notebooks/Shaunak_Mukherjee_
```