Cell Image Data Processing:

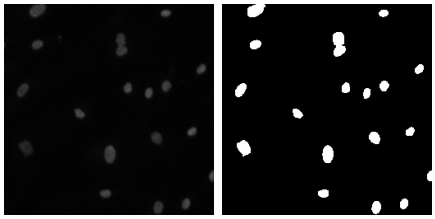
For standardization purposes, we've based all three algorithms of our project on the reprocessed data within the folders titled cell_imgs, mask_imgs and test_imgs. Only these data are required to fully reproduce our code. There are 670 files within each folder, and they are 1 to 1 matched with each other.

This document detaily describes how preprocessed data in cell_imgs and mask_imgs are obtained from original source to finish.

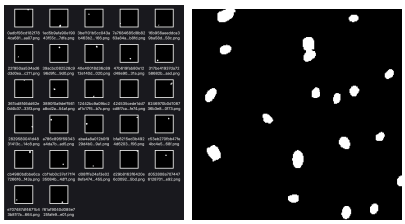
Data Source:

The original source of our cell and training mask data comes from https://www.kaggle.com/c/data-science-bowl-2018/data?select=stage1_train.zip. Specifically, we downloaded the data named stage1_train.zip.

For accuracy evaluation and UNet training purposes, each cell image on the right is paired with a cell mask on the left which is the ideal cell segmentation result.



However, after initial examination, we found that the number of raw masks for a single cell image is the number of cells in that image as shown below on the left (each cell had its own mask), what we need is something like the right:



Therefore, we've developed a short python script named **extract_cell_images.py** to "stack" the individual cell mask on the left together to form what we desire on the right. We've also cleaned up the naming of files and redundant folder structures using the data in "stage1_train".

Reproducing data cleaning results:

To reproduce the data we've been using, first, put **extract_cell_images.py** in the same directory as **stage1_train**, such that the path becomes something like `.../dir/stage1_train/...` for **stage1_train** and `.../dir/extract_cell_images.py` for the python script. Then run the python script by running `$ python extract_cell_images.py`

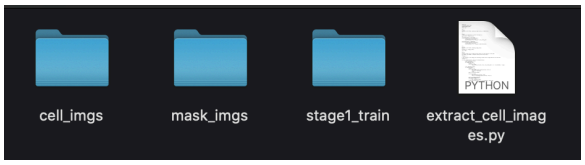
Specifically, the below are the packages imported for **extract_cell_images.py**

```
import os
from shutil import copy
from PIL import Image
import numpy as np
import cv2
```

In normal circumstances, shutil, PIL, os shouldn't need extra installation steps. If numpy isn't installed, then run `$ pip install numpy`. To successfully import cv2, run `$ pip install "opencv-python-headless<4.3"` for fast installation.

Ready to use data:

After running `$ python extract_cell_images.py`, we will have the below folder structure and two new folders named **cell_imgs** and **mask_imgs** that are created by our python script.



Entering the two new folders, we will see ready to use data with clean masks and new naming. The right screenshot below shows **cell_imgs** and the left screenshot shows **mask_imgs**. Each cell image **X.png** is matched with **Xmask.png**.