predictive modeling

the process by which a model is created or chosen to try to best predict the probability of an outcome

OR

the process of developing a mathematical tool or model that generates an accurate prediction

while predictive models guide us towards more satisfying products, better medical treatments, and more profitable investments, they regularly generate inaccurate predictions and provide the wrong answers.

our abilities to predict or make decisions are constrained by our present and past knowledge and are affected by factors that we have not considered

nonetheless this should not deter us from improving our process to build something better

common reasons why models fail

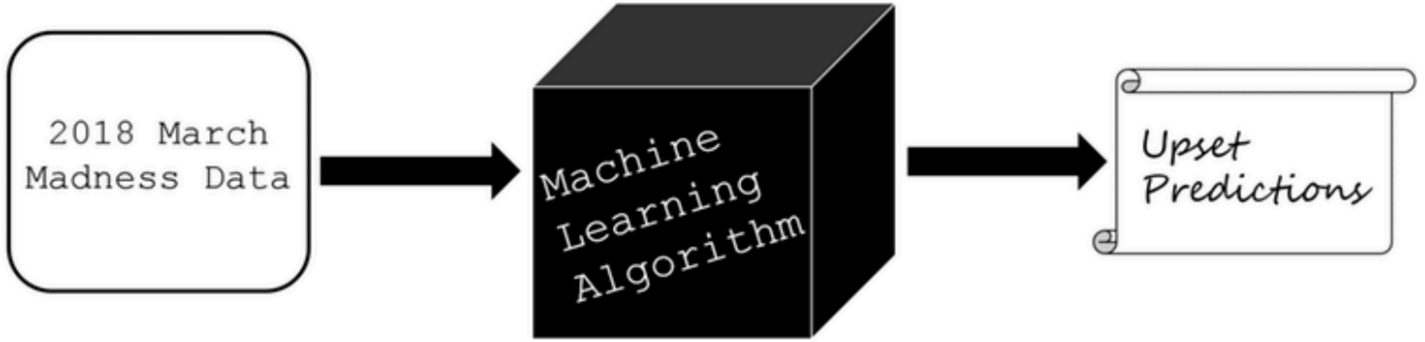
- poor data pre-processing
- inadequate model validation
- unreasonable extrapolation
- overfitting

prediction vs inference

unfortunate reality is that as we push towards higher accuracy, models become more complex and their interpretability becomes more difficult

but…it doesn't have to be like this

Machine learning techniques can be likened to a black box. First, you feed the algorithm past data, essentially setting the dials on the black box. Once the settings are calibrated, the algorithm can read in new data, compare it to past data and then spit out its predictions.



A black box view of machine learning algorithms. (Matthew Osborne, CC BY-SA)

In machine learning, there are a variety of black boxes available. For our March Madness project, the ones we wanted are known as classification algorithms. These help us determine whether or not a game should be classified as an upset, either by providing the probability of an upset or by explicitly classifying a game as one.


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

in the past decade R and Python as well as increased computing power make it fairly easy for anyone with some knowledge to begin to develop predictive models

BUT

the credibility of model building has weakened as this window to tools and data has grown

if there is a predictive signal in a set of data, MANY models will find some of that signal

the best, most predictive models are influenced by the modeler with expert knowledge and context of the problem

of course, this entails finding relevant data for desired research objectives

irrelevant information can often drive down predictive performance of many models — this is where domain-specific knowledge can play a large role

> try to separate potentially meaningful info from irrelevant info to separate the signal from the noise
>
> without expert knowledge identifying confounding signal may not be differentiated

effective model building comes as a result of

intuition and deep subject matter expertise

relevant data

a swiss army knife like set of skills including pre-processing, visualization, and computational abilities

it is important to keep in mind the famous words of George Box:

‘All models are wrong, but some are useful.’

Certainly, a useful model should fit the data well, and information criteria are helpful guides here, but other considerations, such as interpretability and scientific justification are also important

there are no explicit rules to follow when building a model – other than, perhaps, don’t build a model by blindly following rules – and different people could look at this same data and build different models

Characteristics of different methods.

Key: ▲ = good, ◆ = fair, and ▼ = poor

Characteristic	Lasso/ Elnet	Deep Nets	RF	Boosting	SVM
Predictive Power	◆	▲	▲	▲	◆
Interpretability	▲	▼	◆	◆	◆
Feature Selection	▲	▼	◆	◆	▼
Feature construction/exploit local structure in time & space	▼	▲	▼	▼	▼
Automatic	▲	▼	▲	◆	▲
Handling of ‘mixed’ type of data	▼	▼	▲	▲	▼
Handling of missing values	◆	◆	▲	▲	◆
Computational scalability	▲	◆	▲	▲	◆
‘Customizability’	◆	▲	◆	◆	◆

*taken from a R. Tibrashani talk. also an different version is published in The Elements of Statistical Learning

data pre-processing

additions, deletions, transformations, augmentations

this is vital to the predictive performance of a model

IT'S NOT ABOUT THE ALGORITHM

The data that transformed AI research—and possibly the world

data pre-processing

different models have different sensitivities to the type of predictors in the model; how the predictors enter the model is also important.

transformations of the data to reduce the impact of data skewness or outliers can lead to significant improvements in performance.

feature extraction is one empirical technique for creating surrogate variables that are combinations of multiple predictors.

additionally, simpler strategies such as removing predictors based on their lack of information content can also be effective.

type of model choice necessitates the need/amount of data pre-processing

tree-based models, are notably insensitive to the characteristics of the predictor data

linear regression is not

unsupervised data processing

the outcome variable is not considered by the techniques

we also describe strategies for removing predictors without considering how those variables might be related to the outcome

feature engineering

how the predictors are encoded can have a significant impact on model performance

for example, using combinations of predictors can sometimes be more effective than using the individual values: the ratio of two predictors may be more effective than using two independent predictors

often the most effective encoding of the data is informed by the modeler’s understanding of the problem not derived from any mathematical technique

data pre-processing

some encodings may be optimal for some models and poor for others.

tree-based models will partition the data into two or more bins. theoretically, if the month were important, the tree would split the numeric day of the year accordingly. in some models, multiple encodings of the same data may cause problems.

some models contain built-in feature selection only including predictors that help maximize accuracy — the model can pick and choose which representation of the data is best.

now this doesn't mean that the relationship between predictors and outcome is not important

if there are time dependencies and/or seasonalities then how you encode for instance a month/date/year may become important

as with many questions of statistics, the answer to “which feature engineering methods are the best?” is that it depends. specifically, it depends on the model being used and the true relationship with the outcome.

data transformations

transformations of predictor variables may be needed for several reasons.

some modeling techniques may have strict requirements, such as the predictors having a common scale

in other cases, creating a good model may be difficult due to specific characteristics of the data (e.g., outliers)

centering and scaling

most straightforward and common data transformation is to center - scale the predictor variables.

> to center a predictor variable, the average predictor value is subtracted from all the values as a result of the predictor has a zero mean
>
> to scale the data, each value of the predictor variable is divided by its standard deviation coercing the values to have a common standard deviation of one

only real downside to these transformations is a loss of interpretability

skewness

another common reason for transformations is to remove distributional skewness

a general rule of thumb to consider is that skewed data whose ratio of the highest value to the lowest value is greater than 20 have significant skewness.

also, the skewness statistic can be used as a diagnostic. The formula for the sample skewness statistic is

$$\text{skewness} = \frac{\sum(x_i - \overline{x})^3}{(n-1)v^{3/2}}$$
$$\text{where} \quad v = \frac{\sum(x_i - \overline{x})^2}{(n-1)},$$

as always look at distributions in exploratory data analysis

transformations for skewness

replacing the data with the log, square root, or inverse may help to remove the skew

a transformation doesn’t ensure symmetry but generally the data become better behaved than when they were in the natural units

alternatively, statistical methods can be used to empirically identify an appropriate transformation.

Box and Cox propose a family of transformationsthat are indexed by a parameter, denoted as λ:

$$x^* = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(x) & \text{if } \lambda = 0 \end{cases}$$

In addition to the log transformation, this family can identify square transformation (λ = 2), square root (λ = 0.5), inverse (λ = –1), and others in-between.

Using the training data, λ can be estimated.

this procedure can be applied independently to each predictor data that contain values greater than zero.

transformations for outliers

generally defined outliers are samples that are exceptionally far from the mainstream of the data.

under certain assumptions, there are formal statistical definitions of an outlier. even with a thorough understanding of the data, outliers can be hard to define. However, we can often identify an un-usual value by looking at a figure.

when one or more samples are suspected to be outliers, the first step is to make sure that the values are scientifically valid (e.g., positive blood pressure) and that no data recording errors have occurred

do NOT just delete outliers

with small sample sizes, apparent outliers might be a result of a skewed distribution where there are not yet enough data to see the skewness.

also, the outlying data may be an indication of a special part of the population under study that is just starting to be sampled.

depending on how the data were collected, a “cluster” of valid points that reside outside the mainstream of the data might belong to a different population than the other samples

there are several predictive models that are resistant to outliers.

tree-based classification models create splits of the training data and the prediction equation is a set of logical statements such as “if predictor A is greater than X, predict the class to be Y ,” so the outlier does not usually have an exceptional influence on the model.

also, support vector machines for classification generally disregard a portion of the training set samples when creating a prediction equation. the excluded samples may be far away from the decision boundary and outside of the data mainstream.

if a model is considered to be sensitive to outliers, one data transformation that can minimize the problem is the spatial sign.

this procedure projects the predictor values onto a multidimensional sphere with the effect of making all the samples the same distance from the center of the sphere.

mathematically, each sample is divided by its squared norm:

$$x_{ij}^{*} = \frac{x_{ij}}{\sqrt{\sum_{j=1}^{P} x_{ij}^{2}}}.$$

Note that this manipulation of the predictors transforms them as a group Removing predictor variables after applying the spatial sign transformation may be problematic.

dimension reduction

these methods reduce the data by generating a smaller set of predictors that seek to capture a majority of the information in the original variables.

fewer variables can be used that provide reasonable fidelity to the original data.

for most data reduction techniques, the new predictors are functions of the original predictors; so all the original predictors are still needed to create the surrogate variables

PCA is a commonly used data reduction technique

finds linear combinations of the predictors, known as principal components (PCs), which capture the most possible variance.

the first PC is defined as the linear combination of the predictors that captures the most variability of all possible linear combinations.

subsequent PCs are derived such that these linear combinations capture the most remaining variability while also being uncorrelated with all previous PCs

the primary advantage of PCA, and the reason that it has retained its popularity as a data reduction method, is that it creates components that are uncorrelated.

while PCA delivers new predictors with desirable characteristics, it must be used with understanding and care

understand that PCA seeks predictor-set variation without regard to any further understanding of the predictors (i.e., measurement scales or distributions) or to knowledge of the modeling objectives

PCA can generate components that summarize characteristics of the data that are irrelevant to the underlying structure of the data and also to the ultimate modeling objective

because PCA seeks linear combinations of predictors that maximize variability, it naturally will summarize predictors that have more variation

if the original predictors are on measurement scales that differ in orders of magnitude, then the first few components will focus on summarizing the higher magnitude predictors while latter components will summarize lower variance predictors

this means that the PC weights will be larger for the higher variability predictors on the first few components.

it also means that PCA will be focusing its efforts on identifying the data structure based on measurement scales rather than based on the important relationships within the data for the current problem

to help PCA avoid summarizing distributional differences and predictor scale information, it is best to first transform skewed predictors and then center and scale the predictors prior to performing PCA.

centering and scaling enables PCA to find the underlying relationships in the data without being influenced by the original measurement scales.

also one-hot encode categorical variables

the second caveat of PCA is that it does not consider the modeling objective or response variable when summarizing variability

because PCA is blind to the response, it is an unsupervised technique

if the predictive relationship between the predictors and response is not connected to the predictors' variability, then the derived PCs will not provide a suitable relationship with the response.

dealing with missing values

in many cases, some predictors have no values for a given sample. these missing data could be structurally missing, such as the number of children a man has given birth to. in other cases, the value cannot or was not determined at the time of model building. or they can be missing at random

customer ratings are a type of missing data; the Netflix Prize competition to predict which movies people will like based on their previous ratings is often conceived as a missing data problem

it is important to understand why values are missing

Aaron will cover this more in depth at a later date

there are potential advantages to removing predictors prior to modeling

fewer predictors means decreased computational time and complexity

if two predictors are highly correlated, this implies that they are measuring the same underlying information.

some models can be crippled by predictors with degenerate distributions.

adding predictors - one hot encoding

categories are re-encoded into smaller bits of information called “dummy variables.”

each category get its own dummy variable that is a zero/one indicator for that group.

really only n-1 dummy variables are needed; once you know the value of n-1 of the dummy variables, the nth can be inferred. however, the decision to include all of the dummy variables can depend on the choice of the model

however, the decision to include all of the dummy variables can depend on the choice of the model

models that include an intercept term, such as simple linear regression would have numerical issues if each dummy variable was included in the model

> the reason is that, for each sample, these variables all add up to one and this would provide the same information as the intercept.

if the model is insensitive to this type of issue, using the complete set of dummy variables would help improve interpretation of the model

binning predictors

one common approach to simplifying a data set is to take a numeric predictor and pre-categorize or “bin” it into two or more groups prior to data analysis

there are many issues with the manual binning of continuous data.

there can be a significant loss of performance in the model

many modeling techniques are very good at determining complex relationships between the predictors and outcomes; binning can limit this potential

research has shown that categorizing predictors can lead to a high rate of false positives (i.e., noise predictors determined to be informative).

how many people?

ShanghaiTech

1,198 images with a total of 330,165 people

Separated into Part A (482 images) and Part B (716 images)

Each Part separated into training/test set for evaluation

UCF 50

50 images with a total of 63,705 people

Images range from 94 to 4,543 people

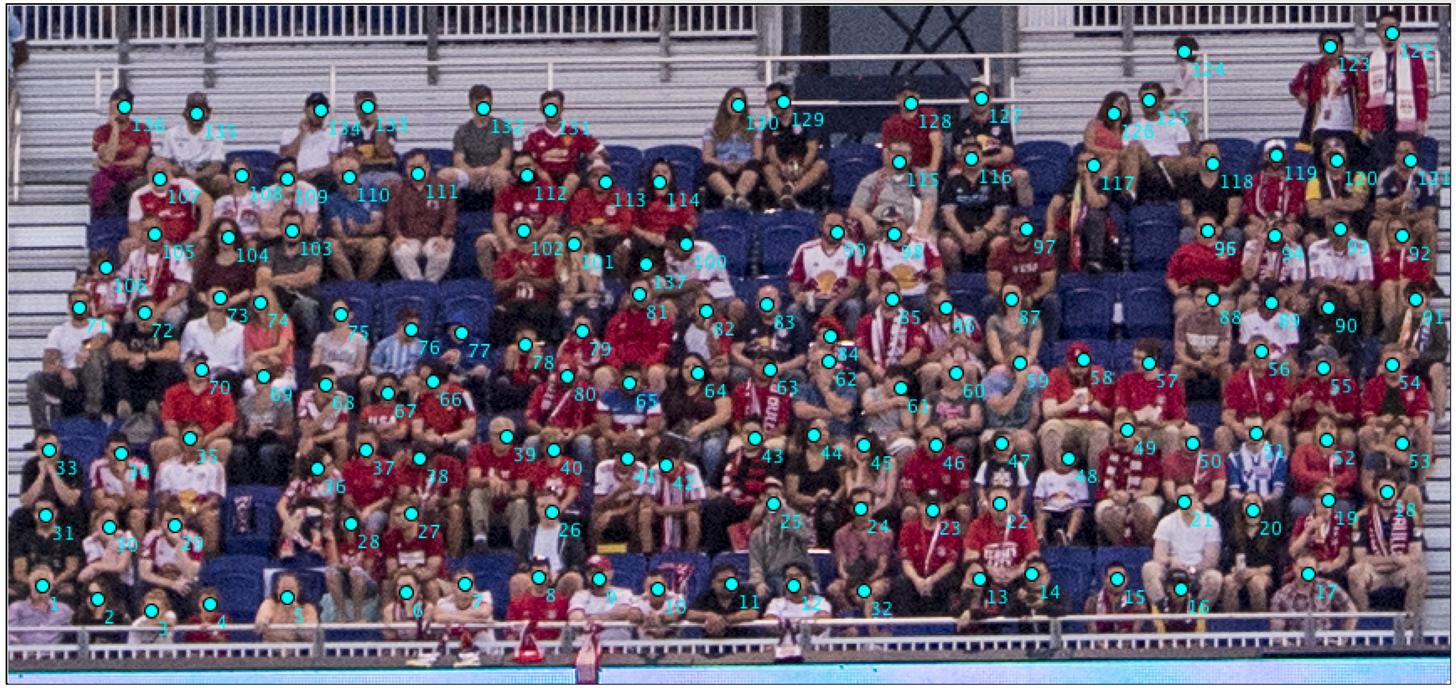
Ground Truth Annotation of 648 people

IMG_154_2.jpg

860x405 pixels; RGB; 1.3MB

ImageJ

ImageJ 2.0.0-rc-43/1.51p; Java 1.8.0_131 [64-bit];

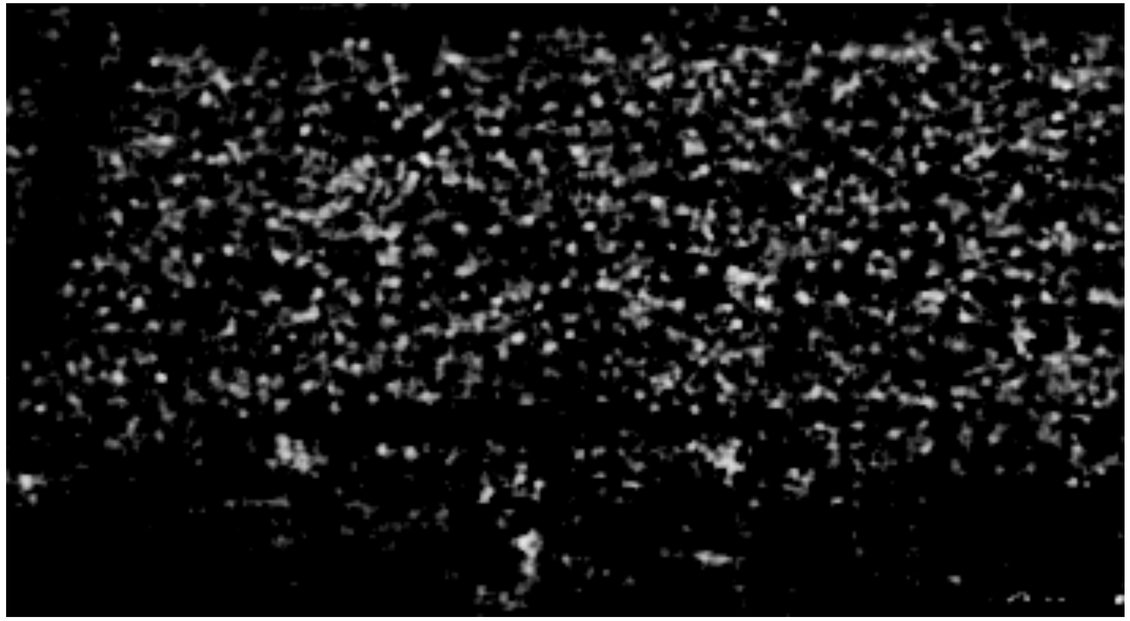


Counting isn’t particularly fun.

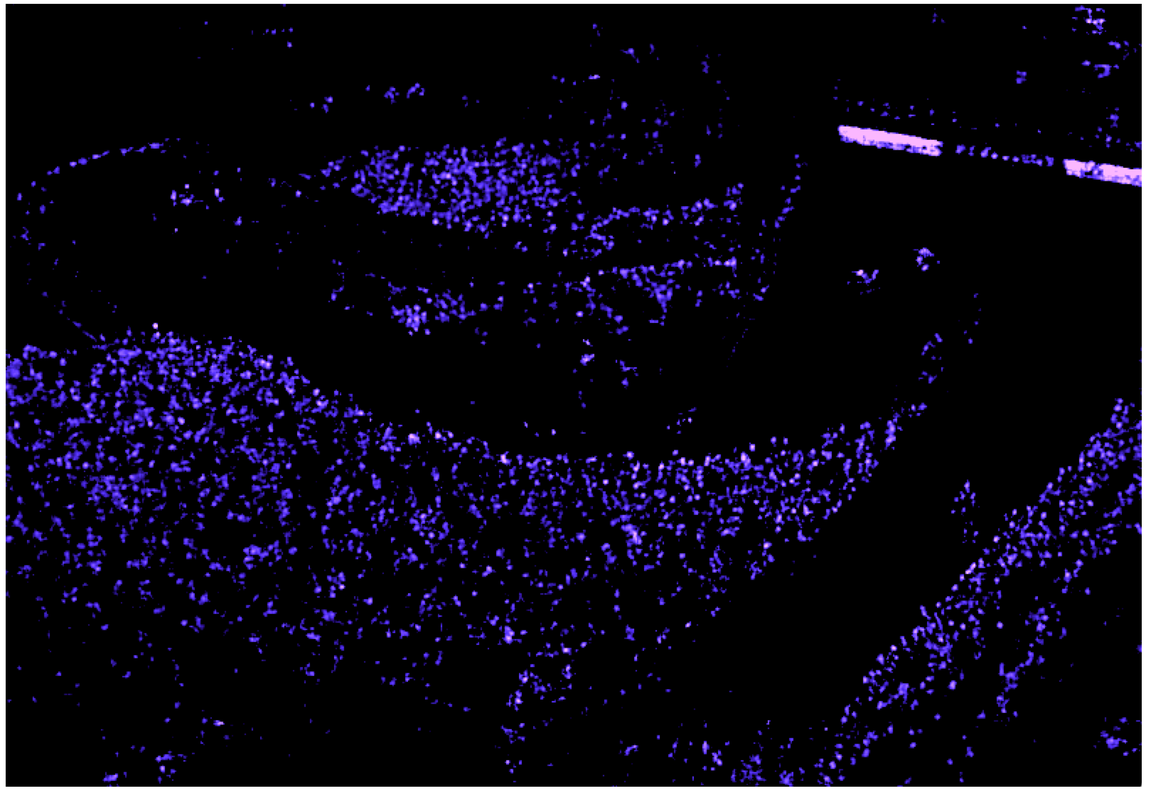
GROUND TRUTH = 136

MODEL ESTIMATE = 273

GROUND TRUTH = 452



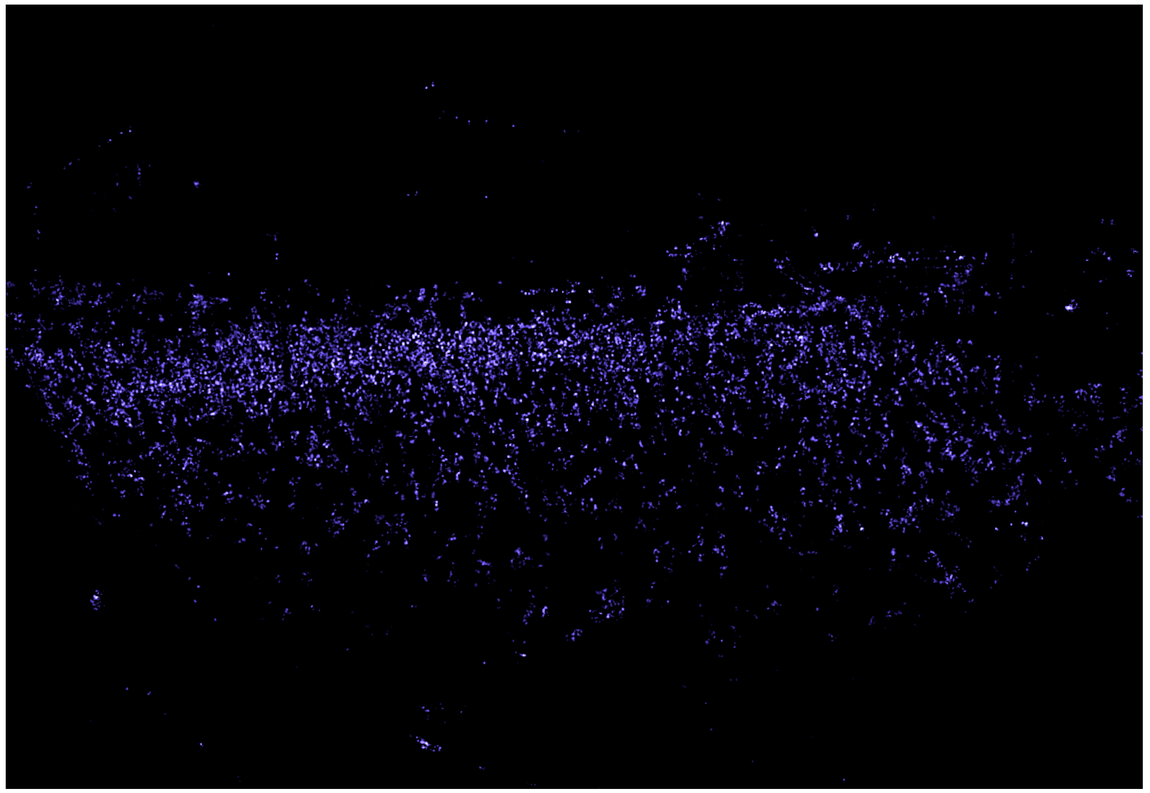
MODEL ESTIMATE = 612



GROUND TRUTH = ???

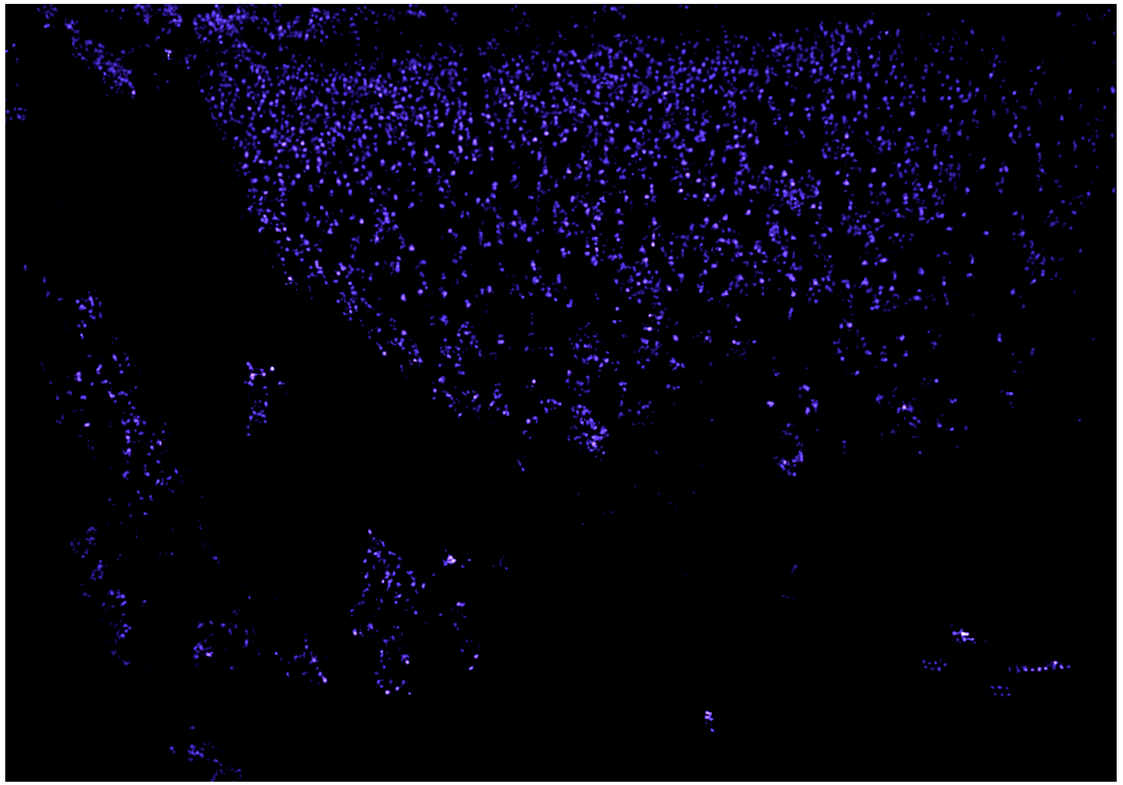
MODEL ESTIMATE = 1012

GROUND TRUTH = ???



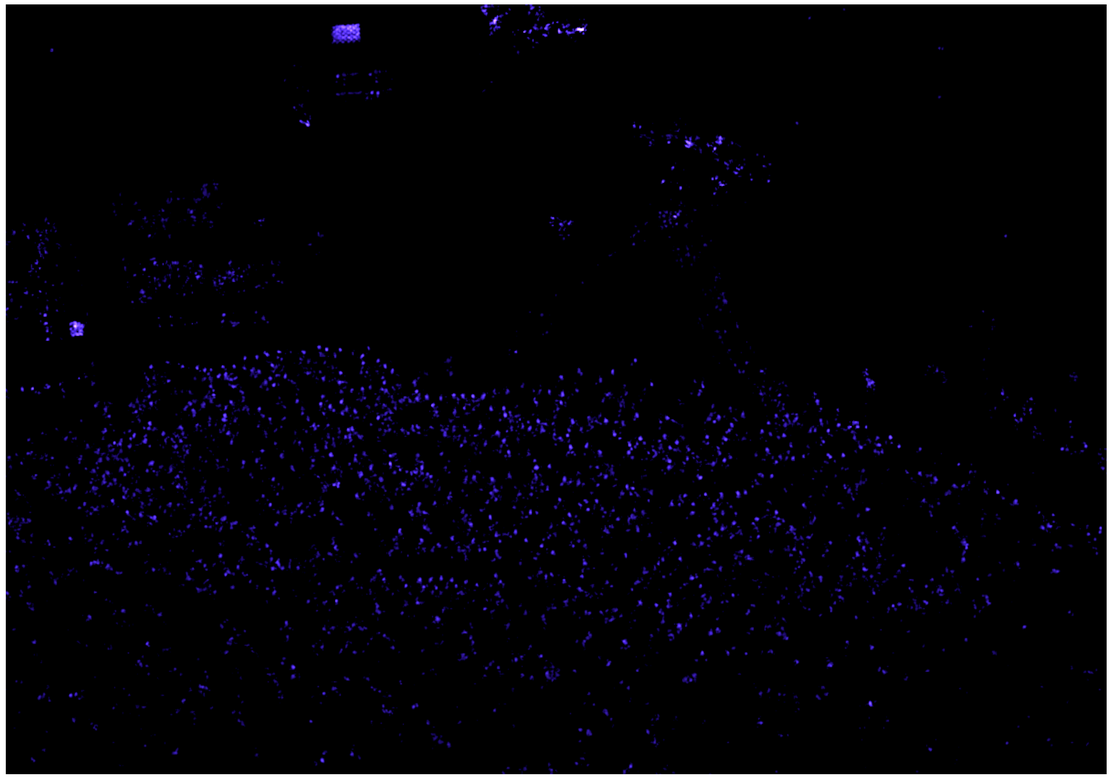
MODEL ESTIMATE = 1618

GROUND TRUTH = ???



MODEL ESTIMATE = 803

GROUND TRUTH = ???



MODEL ESTIMATE = 816