对包含3305个随sighting reports提交的文件的数据集分析得，数据集包含图像、视频和文本数据。对图像数据前期不做处理，视频数据提取关键帧转换成图像数据，打包图像数据，文本数据另行处理.

对包含4440个sighting reports的数据集分析得，数据集包含GlobalID，Detection Date，Notes，Lab Status ，Lab Comments，Submission Date，Latitude，Longitude。

GlobalID作为每个样本的唯一识别标签，

Detection Date和Latitude，Longitude组成时空数据集，

Notes、Lab Comments和前述文本数据组成文本数据集。

由于Submission Date受提交人主观因素影响，故认为不具有统计价值。

Lab Status作为数据的分类依据（Positive ID，Negative ID（二者统称为Identified，定义Negative ID为0，Positive ID为1），Unverified，Unprocessed（后两者统称为Unidentified）），Identified作为train set，Unidentified作为test set。

对图像数据集：

预处理：将原图像转化成64\*64大小的灰度图像。

**处理图像中的昆虫特征：**

由于图像中的昆虫特征可以看成高维空间中的classification问题，故使用支持向量机（Support Vector Machines）进行处理。

绝大部分图像中除了昆虫外均有非纯色的背景，故需要将图像中可能含有昆虫的部分提取出来，引入主成分分析PCA（Principal component analysis）。

经过PCA的图像数据输入SVC（C-Support Vector Classification）中进行学习，同时使用GridSearchCV对SVC的参数进行调整。

**对整幅图像进行考虑：**

我们认为，图像拍摄的环境也可能预示着hornet出现的规律，因此考虑整幅图像具有一定的必要性。为评估整张图像，提出一种基于神经网络的方法MLP（Multi-Layer Perceptron） Classifier。

对时空数据集：

预处理：数据集中的时间为日期格式，无法被程序读取，故提出一种方法，将日期转化为以日为单位的时间戳，时间戳原点为UNIX时间戳原点（1970-01-01），这样能直观的体现每个日期之间的距离。

调用StandardScaler()方法，将时空数据标准化，以消除内部构成不同所造成的对结果的影响。

假定观察位置与观察时间共同影响hornet出现的可能性，故已知的观测数据可以回归于一个函数，要判断一个Unidentified记录的真实性，可以通过判断这一记录输入函数的结果与已知函数值的距离大小。故此处使用带交叉验证的逻辑回归方法（Logistics Regression with cross-validation），此方法可以将新纪录映射到0到1之间的概率空间。

对文本数据集：

预处理：由于文本数据集包含提交者的notes和实验室的Lab comments，我们认为这两者在描述昆虫的信息上是等效的。

首先分别统计Positive ID与Negative ID的文本中各种词语出现的次数搭建N-gram模型，生成一个包含特征索引的字典。

但是尽管可能在描述同一个主题，长的文本相对于短的文本有更高的单词平均出现次数。为了避免这些潜在的差异，将各文本中每个单词的出现次数除以该文档中所有单词的总数：得到一个新的特征称之为词频 TF (Term Frequencies)。

在词频的基础上进行改良，通过降低在该训练文集中的很多文本中均出现的单词的权重，从而突出那些仅在该训练文集中在一小部分文档中出现的单词的信息量，得到IDF（Inverse Document Frequency）。得到一联合特征TF-IDF（Term Frequency times Inverse Document Frequency）。

由于文本数据的特征各自独立，为了判断样本的概率，我们采用多项分布朴素贝叶斯算法（Multinomial Naïve Bayes）

综合概率：

以上模型输出均为概率，故提出一个使用SGD（Stochastic Gradient Descent）计算综合概率的方法。