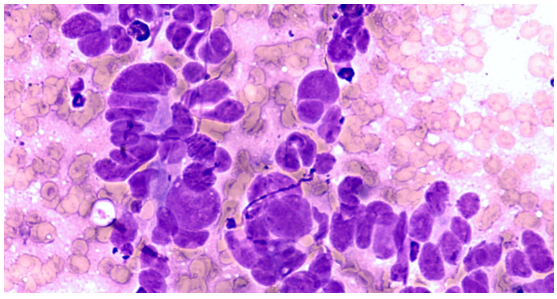
# Physics Dataset Practice Problems

Week of July 20, 2020

## 2 Breast Cancer

Can we correctly classify whether breast cancer masses are malignant or benign? The Wisconsin Breast Cancer Diagnostic Dataset contains 30 characteristics of cell nuclei computed after fine needle aspiration (FNA) of the masses. These values along with whether the mass was truly benign or malignant are tabulated for 569 patients in the dataset `BreastCancer.csv`.



### 2.1 Visualize the Dataset

Visualize a few dependencies within the dataset by plotting. Consider using the `pairplot` function from the `seaborn` library. `seaborn` is a wrapper for `matplotlib` which can be used to easily make informative graphics from `pandas` dataframes. You can use `pairplot` with the `hue=` and `vars=` keyword arguments to visualize malignancy's dependence on the dataset's features.

### 2.2 Build a Model

Use machine learning to model malignancy as a function of FNA features. Compare a few models' performance using area under the receiver operator characteristic curve (AUROC) and a 50% held out validation dataset. Plot ROCs for your models using the `sklearn.metrics` module. Note that you will have to use the method `.predict_probas()` rather than `.predict()` to use these metrics.