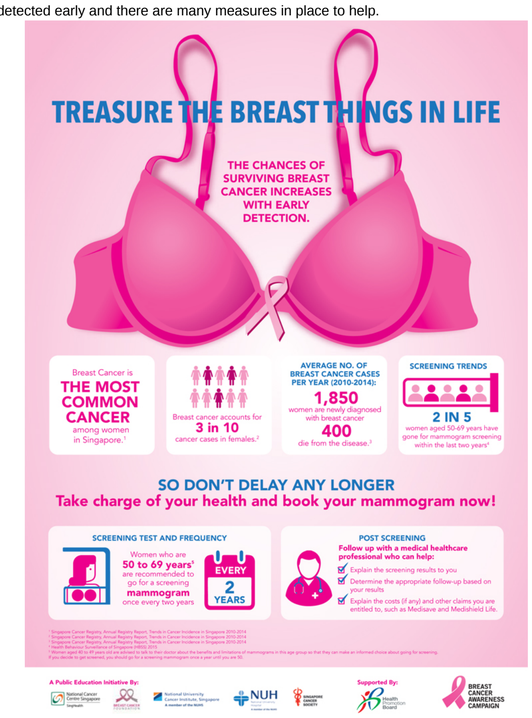
Introduction

From the infographics below, breast cancer accounts for 30% of the cancers in females, making it the most common cancer in women. Breast cancer is also one of the most curable disease if detected early and there are many measures in place to help.



Courtesy of [Singapore Cancer Society]

In our context, we are presented with a dataset that are taken from a biopsy procedure called **Fine Needle Aspiration (FNA)** performed on the breast. The tissue taken from the biopsy will then be sent to a lab and be examined by a pathologist, a report will be written if cancerous cells are spotted or not and be sent to the specialist to further explain the results to the patient. However, there may be disagreements whereby the pathologist report shows there are no signs of cancerous cells, while radiologist may disagree as he/she might find suspicious lesions from the mammogram/CT/MRI scans. This can happen if the biopsy taken is only on the benign cells and if there is dispute, a more thorough of biopsy may be performed again.

Although our aim in Machine Learning is to classify whether a tumor is benign or malignant, we should bear in mind that we are not trying to dispute the expertise of the doctors/pathologists/radiologists. Instead, we develop models to aid their understanding, and also to come up with a more systematic benchmark for one to refer to. More concretely, the dataset has features that are computed from a digitized image from **FNA**, and each observation describes statistics/characteristics of the cell nucleus. There are 10 base features, and 3 different measurements are taken for each feature, namely, the **mean, standard error and the "worst/largest"**. One thing to note is that **worst** means the **mean of the three largest values**.

Attribute Information:

- ID number
- Diagnosis (M = malignant, B = benign)

Ten real-valued features are computed for each cell nucleus:

- radius (mean of distances from center to points on the perimeter)
- texture (standard deviation of gray-scale values)
- perimeter
- area
- smoothness (local variation in radius lengths)
- compactness ($\mathrm{perimeter}^2/\mathrm{area} - 1.0$)
- concavity (severity of concave portions of the contour)
- concave points (number of concave portions of the contour)
- symmetry
- fractal dimension ("coastline approximation" - 1)

With these in mind, let us move on to defining the problem and state some initial assumptions.

Stage 0: Defining the problem and Assumptions

Problem Statement

Informal Description:

- To develop a Machine Learning Model that can classify whether a tumor is benign or malignant. We also note that we care more about whether a cancer patient is classified correctly.

Formal Description (Appendix for Notations):

- Given a dataset $\mathcal{D}$ describing characteristics of a tumor, the task $\mathcal{T}$ is a binary classification problem where we aim to find an optimal hypothesis $g \in \mathcal{H}$ using a learning algorithm $\mathcal{A}$. The optimal hypothesis g should generalize well, that is to say, has a low expected generalization error $\mathcal{E}$ over a performance measure M. We will choose the performance measure in the later sections (not accuracy).

Considerations

- **Size of Dataset**

The dataset is not too large, we need to be wary of an overly complex model which may easily overfit, but may not generalize well.

- **Model Interpretation**

There is a tradeoff between Model's complexity/flexibility and it's interpretability. If we need to explain our model to our business stakeholders, then it is a good idea to choose a model that can be interpreted well, models like Logistic Regression with Lasso may be a good choice as the model itself has better interpretation, and with lasso we can reduce the number of features. If we only care about our model's ability to predict, then interpretability may not be so important and we may choose a model that performs well, but the weights may be more difficult to understand.

- **Time and Space Complexity**

Practically speaking, we need to strike a balance between the speed of the training and the performance measure of the model.

- **Data Centric vs Model Centric**

From the one and only Andrew Ng, we understood that data plays a critical role in the Machine Learning world.