## Title: Predicting the presence of metastatic tissue in histopathologic scans of lymph node sections

**Problem:**

Cancer is among the leading causes of death worldwide. Approximately 38.4% of men and women will be diagnosed with cancer at some point during their lifetimes. However, periodic clinical checkups and self-tests help in early detection and thereby significantly increase the chances of survival. Cancer that’s diagnosed at an early stage, when it isn’t too large and hasn’t spread, is more likely to be treated successfully. If cancer spreads, effective treatment becomes more difficult, and generally a person’s chances of surviving are much lower. Invasive detection techniques cause rupture of the tumor, accelerating the spread of cancer to adjoining areas. Hence, there arises the need for a more robust, fast, accurate, and efficient noninvasive cancer detection system. In order to detect signs of cancer, breast tissue from biopsies is stained to enhance the nuclei and cytoplasm for microscopic examination. Then, pathologists evaluate the extent of any abnormal structural variation to determine whether there are tumors. Architectural Distortion (AD) is a very subtle contraction of the breast tissue and may represent the earliest sign of cancer. Pathologists require a lot of expertise and skill to perform the analysis. Hence, development of algorithms which rely on automated image analysis, is of great interest. We propose to use data and build a machine learning model that can help doctors find the cancer cells and ultimately save human lives.

**Client:** Doctors and pathologists in different Hospitalscan use such a model to identify metastatic tissue in histopathologic scans of lymph node sections, which as a result would help them make early clinical decisions with greater accuracy. Early detection can give patients more treatment options, which in turn would increase their survival.

**Data:** The data set will be acquired from the Kaggle Competition.  This data contains information about the ID of the image and the label (Positive or Negative) of the histopathological image. The data for this project  is a slightly modified version of the PatchCamelyon (PCam) [benchmark dataset](https://github.com/basveeling/pcam) (the original PCam dataset contains duplicate images due to its probabilistic sampling, however, the version presented on Kaggle does not contain duplicates).

**Modeling approach:  We take historical data in order to fit a better model. Since the data is already labeled** a supervised machine learning algorithm is a perfect choice to build a predictive model.

**Deliverables:**

1. Codes (notebooks) for data acquisition, data cleaning, data exploration analysis, and machine learning model development
2. Report on the capstone project
3. Presentation on the capstone project