A Probabilistic Approach to the Positive and Unlabeled Learning Problem
Kristen Jaskie, Andreas Spanias
SenSIP Center, School of ECEE, Arizona State University.

Traditional binary classification requires well-labeled data.
- Both positive and negative labels:
  - y = 1
  - y = 0

Negative data is EXPENSIVE in many interesting problems.
- Ex: Cancer Detection
  - Known positive set: People who have cancer
  - Unlabeled set: Everyone else
  - Finding true negatives – people who ABSOLUTELY do not have cancer – is expensive or impossible.

This leaves us with some positive and no negative labels.

Other examples:
- Fraud detection
- Terrorist detection
- Threat detection

**GOAL**
- Given data samples \( x \) and data labels \( y \)
- We want to learn a probabilistic classifier \( p(y = 1| x) \)

**PREVIOUS SOLUTION [1]**
- Include a new random variable \( s \):
  - If sample is labelled, \( s = 1 \), if unlabelled, \( s = 0 \)
- It can be shown that
  \[
  p(y = 1| x) = \frac{p(s = 1| x)}{c}
  \]
- Used Standard Logistic Regression (SLR) to learn non-traditional classifier \( p(s = 1| x) \)
- Constructed estimators for \( c \) using a validation data.
  - Found to be INEFFECTIVE in practice.

**OUR SOLUTION**
- Created a Modified Logistic Regression (MLR) to learn non-traditional classifier \( p(s = 1| x) \)
  
**REFERENCE**