# Problem Statement

Indix deals with collecting, structuring & analyzing product data. Majority of our sources for data happens to be the Web pages from the Internet. Collecting information from webpages have a wide variety and unique set of challenges. It's a space where there are standards to some extent, but, not followed rigorously. Hence, in the industry, HTML content is usually considered to be semi-structured.

Given that Indix deals with collecting product data, most of the HTML pages that we deal with tend to be product or listing pages. For the purposes of this event, we will mostly be dealing with Product pages. One of the hardest challenges in this domain is to mine and extract meaningful product attribute data. And, if you've been to any e-commerce portal, you'd know rich source of product attribute data is available in the form of Product specification, usually a tabular section in the page. Our problem for this day is to mine HTML content and identify such tabular sections of the page. We are looking for solutions that cater to a wide variety of sites or patterns - more technically, where the generalization error for the model is low.

You're given a dataset which contains html content, url(encoded) and a field(to be predicted) - yes / no, indicative of whether the HTML content, represents a "visual table" entity. Note that, presence of table or tr or td need not necessarily mean that the visual representation is going to be tabular.

The task is to train a model based on the html features to make predictions on whether a particular html blob has "visual tables" or not.

F-score is used to measure the perfomance of your predictor

## Problem Statement PDF

- problem_statement_table_detection.pdf

## Training Dataset

Header ['label', 'table-text', 'url']

- table_train.csv.gz

## Blind Dataset

Header ["id", "table-text", "url", "site"]

- blindset_table_out.csv.tar.gz

## Sample Random Submission

- sample_random_submission.csv

## Learning Resources

- Featurizing Text - Wharton Statistics Department
- Text classification and Naive Bayes
- Large-Scale Bayesian Logistic Regression for Text Categorization
- The Semantic Web
- Semantic framework for web scraping
- Schema.org
- Product - schema.org
- Comparison of HTML parsers