# What is the problem?

There are thousands of school districts in the United States. Many school districts are struggling with performance or financial issues, and unfortunately, some school districts may close within five years. In this project, I will identify school districts that are at risk (close within five years).

# Who is the client and why should they care?

The client would be organizations that work with educational institutions. These ma ynclude government (local, state, and federal) and non-profit organizations. Given the size of the educational system, it would be advantageous to identify school districts at risk and implement interventions tailored towards these school districts’ needs.

# What data is being used?

I will use the Common Core Dataset provided by the National Center for Education Statistics. This data contains financial, demographic, and general attributes for every school district within the United States for years after 1990. All data files are available to the public on the Common Core Dataset Website (<https://nces.ed.gov/ccd/ccddata.asp>) as zip files.

# What is the approach?

I will use both a supervised and unsupervised approach to the problem.

For the supervised approach, it will be a binary classification problem, and I will predict whether school districts in 2010 are still operational in 2015. There are hundreds of variables within the Common Core Dataset, and I will choose variables that: (1) have minimal missing values, (2) are also available in the most current dataset, and (3) do not contain future information (no data leakage). In addition, the data in 2010 will be split into training and test sets to prevent overfitting.

For the unsupervised approach, I will apply clustering to group school districts with similar characteristics. I will then apply statistical tests to determine whether certain clusters have more school districts at risk than others.

# What are the deliverables?

The final deliverables of this project are:

* Trained and tuned machine learning (classification) model
* Cluster labels after applying clustering algorithm
* Final report detailing processes, results, and opportunities.