Capstone Project

# Goal:

* Bigger impact
  + How to look at *in vitro* and *in vivo* relationships in cheminformatics ways
  + Setup workflow for future analysis
* Identify actual chemical features (feature development) that are responsible for the relationships
  + Features that could be impacted by metabolism

# Data Descriptions:

* Dependent Variable
  + An Excel file containing *in vivo* outcomes (genetoxicity endpoint)
    - Row = a list of chemical identifiers
    - Column = a binary value indicating whether the chemical is active (1) or inactive (0) in the assay
* Independent Variable
  + In Vitro assays results
    - An Excel file containing ~ 800 *in vitro* assay results
      * Row = a list of chemical identifiers
      * Column = a binary value indicating whether the chemical is active (1) or inactive (0) in the assays
  + Descriptors
    - An Excel file containing ~ 779 chemical features
      * Row = a list of chemical identifiers
      * Column = a list of chemical features
      * Value = 1 (present of feature in compounds) and 0 (absent of feature in compounds)
    - An Excel file containing chemical property data
      * Row = a list of chemical identifiers
      * Column = a list of chemical properties
    - Value = continuous values
* Validation dataset: Metabolism data
  + Everywhere. Need to organize them in some ways.
  + Can we use R to organize this dataset?

# Modeling Method:

* Classification methods

# Steps:

1. Determine chemical coverage in the *in vitro* active and inactive chemicals (chemotype for visualizations)
2. Determine chemical coverage in the *in vivo* active and inactive chemicals (chemotype for visualizations)
3. Determine the feature relationships between the two datasets (visualizations, feature selection, statistics analysis, etc.)
4. Look at the features that could be affecting the parent compounds
5. Look at the features that could be affecting the metabolites (compare to Meteor/Derek features – CYP450)
6. Reverse engineer metabolite features to chemotypes
7. Build model using classification methods and determine which one works better