**Capstone 2: Project Ideas**

1) Predict Chief Executive Officer (CEO) compensation.

- The dataset includes the following variables:

- Target variable: Total CEO compensation in 2006 fiscal year.

- The features: total compensation in 2005 fiscal year, Sales revenues in 2005 and 2006 fiscal year, company accounting performance (Return on Asset (ROA)) in 2005 and 2006, company stock return in 2005 and 2006, stock price volatility, and median CEO compensation in the company chosen peer groups.

- I created this dataset by combining several (paid) datasets:

- Execucomp – dataset about compensation of executives (covers SP1500 companies).

- Compustat – dataset about accounting variables of U.S. public companies.

- IncentiveLab – dataset about compensation peer groups.

- Dataset includes 707 firms (SP1500 firms) that use compensation peer groups (companies usually benchmark their CEO compensation against a group of firms – i.e. peer group).

- For example, if Amazon wants to determine the fair total compensation for its CEO, the board of directors can choose a group of firms (compensation peer group) and benchmark the compensation against this peer group. Amazon peer group probably includes big technology firms such as Apple, Walmart, Target and so.

2) Stroke prediction

- Project goal: predict whether a person would have stroke.

- Stroke Prediction Dataset from Kaggle [Stroke Prediction Dataset | Kaggle](https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset)

- The data contains 5110 observations with 12 attributes.

3) Personality Traits vs Drug Consumption

- Project goal: choose one of the 18 drugs; and make prediction of a person being “Non-user” or “User” of this drug.

- Each of outcome variables in this dataset contains seven classes: "Never Used", "Used over a Decade Ago", "Used in Last Decade", "Used in Last Year", "Used in Last Month", "Used in Last Week", and "Used in Last Day". I will transform these outcome variables into binary classification. For example, "Never Used", "Used over a Decade Ago" form class "Non-user" and all other classes form class "User".

- The features are 12 personality attributes.

- The dataset has 1885 rows.

- Dataset [UCI Machine Learning Repository: Drug consumption (quantified) Data Set](https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29)

# 4) Give Me Some Credit

- Project goal: predicting the probability that somebody will experience financial distress in the next two years.

- Dataset <https://www.kaggle.com/c/GiveMeSomeCredit/data>

- The dataset has 150000 rows with 10 predicting features.

5) Heart Disease

- Project goal: predict weather a person has heart disease.

- Dataset <https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset>.

- The dataset has 1025 rows, one target variable (binary) and 13 features.