Individual Project (Lung SBRT Outcome)

- Goal: Develop models for predicting distant metastasis after stereotactic body radiation therapy (SBRT) in early stage non-small cell lung cancer (NSCLC).

  - Binary outcome: Prediction (Distant metastasis or not)
  - 98 early stage NSCLC patients with at least 6 months of follow-up who underwent SBRT between 2006 and 2012.
  - 18 variables (All the variables are clinical parameters)
  - No missing values

  - Data Source: Dr. Jing Wang

Literature:

1. Z. Zhou, M. Folkert, N. Cannon, P. Iyengar, K. Westover, H. Choy, R. Timmerman, S. Jiang, and J. Wang, "Predicting distant failure in early stage NSCLC treated with SBRT using clinical parameters", Radiotherapy & Oncology, 119 (3), 501-504, 2016.

Individual Project (Credit Card Fraud Detection)

- Goal: Develop models to detect fraudulent credit card transactions so that customers are not charged for items that they did not purchase.
  - Binary outcome: Fraudulent credit card transactions (Yes or No)?
  - It contains only numerical input variables which are the result of a PCA transformation. Features V1, V2, ... V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Class' is the response variable (1 in case of fraud and 0 otherwise).
  - The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.
  - Data for two days. First day as training, second day as testing
  - Data Source: Kaggle Dataset

Literature: Andrea Dal Pozzolo, et al. Calibrating Probability with Undersampling for Unbalanced Classification. In Symposium on Computational Intelligence and Data Mining (CIDM), IEEE, 2015

Individual Project (Prostate Cancer patient survival)

- Goal: Prediction of overall survival for patients with metastatic castration-resistant prostate cancer

- Outcome: Overall survival
- Training set: 1600, testing set: 470
- About 120 variables including basic demographics, blood work, metastasis, medicine use, and medical history

- Data Source: Past DREAM Challenge, Dr. Tao Wang

References:

https://www.thelancet.com/journals/lanonc/article/PIIS1470-2045(16)30560-5/abstract
https://www.projectdatasphere.org/projectdatasphere/html/pcdc

Individual Project (Genomic Data Analysis Project)

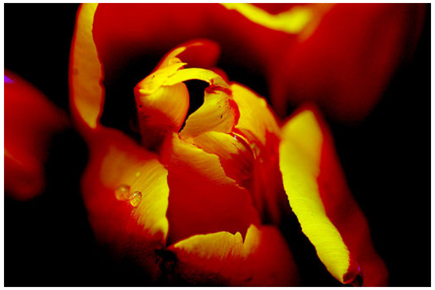
- Goal: Develop models for transcription factor binding site prediction
  - Binary outcome: Does a piece of DNA contain binding site of a new cancer driver gene PAX3-FOXO1 (YES or NO)?
  - 5881 known PAX3-FOXO1 binding sites on a genome scale based on ChIP-seq data of PAX3-FOXO1 in RH4 cell line
  - 3,000 binding sites as train set, and the rest 2,881 as validation set
  - 3,000 sequences without binding sites as negative controls
  - No missing values

  - Data Source: Dr. Lin Xu

Individual Project (Flower recognition)

- Goal: Develop models to images of flowers: chamomile, tulip, rose, sunflower, dandelion.
  - ▪ Categorical outcome: Type of flowers: tulip, rose, sunflower, dandelion, chamomile.
  - ▪ 4242 images of flowers. The pictures are divided into five classes: chamomile, tulip, rose, sunflower, dandelion.
  - Half as training, half as testing
  - Data Source: Kaggle dataset

Tulip



Rose

Individual Project (Housing price)

- Goal: Develop models to predict sale price of each house

  - Continuous outcome: SalePrice
  - 79 variables describing different aspects of residential homes in one city.
  - Sale record of 1461 houses as training
  - 1460 houses as testing

  - Data Source: ongoing Kaggle Competition

Individual project (Low dose CT image processing)

- Goal: Predict a high quality CT image given the low-quality CT acquired at a low-radiation dose level
  - Outcome: High quality CT images
  - Training: 80 pairs of volumetric CT images with high quality and low quality.
  - Testing: new low quality CT images that are not in the training set
  - Evaluation: against known ground truth, evaluated using CT evaluation metrics, such as contrast to noise ratio, mean error etc.

- Data Source: Dr. Xun Jia

Medical literature:

Hu Chen, et. al., Low-dose CT denoising with convolutional neural network. arXiv:1610/00321, (2016).

Group project (Identify similar patient)

- Goal: For a given CT image, identify an image in the database of CT images that is geometrically most similar to the given image.
  - Outcome: ID of the identified image in database
  - Training: ~2000 abdomen CT images of patients
  - Testing: abdomen images not included in the database
  - Evaluation: metrices quantifying geometrical similarity between two images.

  - Data Source: Dr. Xun Jia

Medical literature:

Vorakarn Chanyavanich Shiva K. Das William R. Lee Joseph Y. Lo. Knowledge-based IMRT treatment planning for prostate cancer. Medical Physics 38, 2515, (2011).

Group Project (Housing marker trend)

- Goal: Develop models for predicting housing price at different zip code areas.

  - continuous outcome: monthly median sale price for each zip code
  - ~15,000 different locations in US
  - median sale price for each month from 1996/04-2018/06 at zip code level
  - Training 1996/04 – 2016, testing 2017-2018/06

  - Data Source: Zillow research data

Group Project (Breast Histopathology Images)

- Goal: Develop models for classify Invasive Ductal Carcinoma (IDC) positive and negative image patches.

  - Binary outcome: IDC Positive or negative
  - The original dataset consisted of 162 whole mount slide images of Breast Cancer specimens scanned at 40x. From that, 277,524 patches of size 50 x 50 were extracted (198,738 IDC negative and 78,786 IDC positive).
  - Half has training and half has testing

  - Data Source: Kaggle Dataset

- For more information about the data, see
  https://www.ncbi.nlm.nih.gov/pubmed/27563488
  http://spie.org/Publications/Proceedings/Paper/10.1117/12.2043872

Group Project (Chest X-Ray Images)

- Goal: Develop models for identify Pneumonia patients.

  - Binary outcome: Pneumonia patient or Normal control
  - 5,863 images, 2 categories
  - Half has training and half has testing

  - Data Source: Kaggle Dataset

Open to new ideas

- Bring your own data and project
- Ongoing Kaggle competition

  https://www.kaggle.com/competitions

- Ongoing DREAM Challenges

  http://dreamchallenges.org/