**Arkansas AI-Campus: Project Highlights**
**ASU System Office, Feb. 15, 2019**

**Agriculture Project: Deep Learning in Plant Genomic Selection.**
In this study, we explore the application of artificial intelligence in plant genomic selection, with the intention of increasing phenotype trait prediction accuracy. Plant breeders face challenges in developing new varieties, since it is difficult to predict specific traits of a new strain until it is grown and harvested. The ability to predict which plants are more likely to express desirable traits with a simple genetic test that can be administered in the lab prior to planting test plots would increase the breeder's ability to quickly perfect new varieties. Participants in this project will apply a variety of state-of-the-art machine learning models to the phenotype prediction problem in soybeans. We target traits that are both economically important and difficult to measure without planting test plot, such as yield, branching, plant height, and oil content.

**Self-Driving Car Project: Cityscapes Segmentation.**
In this project, we will focus on *scene understanding*. Scene understanding is an important component for self-driving cars, and requires the ability to perform natural image segmentation and utilize sequence information available from the sequential frames of video. Image segmentation is a core problem that has driven many of this decade's AI breakthroughs. We will focus on the following aspects in this project:
1. Multi-class object segmentation in natural images. For example, using a standard video camera's output to identify the difference between cars, people, bicycles, and the road itself.
2. Utilizing time series information from video as well as other techniques to improve image segmentation and scene understanding. Different kinds of objects behave differently over time, and the behavior of the car in which the camera is mounted must also be considered. This adds to the challenge of this project.

**Self-Driving Car Project: Point-Cloud Segmentation Project.**
In this project, we will focus on a type of data known as a *point cloud*. A point cloud provides depth information and complements images in scene understanding; it is usually generated from LIDAR hardware, or similar instruments. Recent results show that when well used, methods using point-cloud alone can beat methods using both point-cloud and camera images. Kitti is an ongoing competition that provides the point cloud data we will utilize in this project. We focus on the following questions:
1. How can we accurately classify objects using point cloud data?
2. How does detection difficulty increase with distance and what algorithms might help improve accuracy on far-away objects?

**Natural Language Processing: Text-to-Image Project: AI algorithm to automatically find images that match an article.**
In this project, we will examine techniques for understanding natural language in written form, extracting semantic "meaning" from that text, extracting information from images, and matching the information extraction from both data modalities in order to match text descriptions to appropriate images. The three tasks of interest are:
1. Natural language processing for the purpose of extracting semantic information from text.
2. Image processing with deep learning techniques for the purpose of extracting feature information from natural images.
3. Predicting "matching" image/text pairs by applying AI models to the information extracted from the text and the images.

**Lung Cancer CT Image Project**.
Lung cancer is the leading cause of cancer deaths for both sexes in industrialized countries and the second most common cancer in both men and women. Early detection is key to improve patient treatment and survival. Low-dose computed tomography (CT) is shown to reduce mortality from lung cancer by at least 20% and has been recommended in the US for lung cancer early screening. Radiologist assessment of CT images is a tedious process and highly subjective, which is an impediment of clinical throughput. This project is to develop and apply machine learning algorithms that will capture nodule features as well as other clinical features. This work will help improve the accuracy of lung cancer early screening. Data for training the models and evaluating the performance is from NIH NCI the Cancer Imaging Archive (TCIA). We will follow a rigorous statistical evaluation of the algorithm, including using criteria of log-loss, accuracy, sensitivity, specificity and the area under the receiver operating characteristic curve.

**Natural Language Processing Project: Fraud Detection.**
In this project, we will examine techniques for extracting meaningful features and information from natural language texts. Our application is to detect companies that are likely to be sued for fraud by the Securities and Exchange Commission, based on AI models using publicly available financial disclosure documents filed by the companies. We will examine techniques to extract information from financial disclosure documents. Then we will match features extracted from the filings to lawsuits filed by the SEC and announced on its "litigation releases" page, with the goal of building a model that can predict future litigation based on information in a company's filings. The skills learned in this project could easily be adapted to other business uses where analysis of written documentation, legal, or financial filings is of interest.

**Project on Production Ratings by Customers.**
In this project, we will examine the dataset originally made available for The Netflix Prize - an open competition for the best collaborative filtering algorithm to predict user ratings for films, based on previous ratings without any other information about the users or films. The dataset contains 100,480,507 ratings of 17,770 movies provided by 480,189 users. We will first study approaches for imputing (estimating) missing values within such a large dataset. Then, we will apply modern AI techniques for making predictions related to the users and their movie ratings.

**Genomics Project.**
In this project, we will be looking at gene expression data first made available by the TCGA Pan Cancer analysis project and collected by the UCI Machine Learning Repository. This dataset has been sampled so that there is enough data for us to train models, without being too bulky to easily store locally or in the cloud. We will focus on unsupervised learning to explore relationships within the data and supervised learning to classify cancer type from the gene expression. Then, we will investigate feature selection to identify the best subset of genes that are able to accurately discriminate between the cancer types.

**Medical Imaging Project: chest X-Ray.**
In this project, we will be looking at techniques for classification and segmentation of medical images, specifically chest X-ray images. The dataset we will use is provided by the NIH Clinical Center. It consists of over 112,000 chest X-ray images from more than 30,000 unique patients. The dataset has been annotated with 14 abnormal chest classifications, as well as "no finding" (normal X-ray) images. We will focus on two tasks:
1. Classify X-ray images according to the disease or abnormality present.
2. Determine regions within the X-ray image that are associated with the disease or abnormality, and provide a visualization of these regions to guide a diagnostic review by a human viewer.