**Capstone 3 is an example of Unsupervised Learning Clustering**

**Goal**: Group or cluster soybean genotypes based on genetic similarity (Genetic diversity Analysis)

Soybean Germplasm includes 18,480 domesticated and 1168 wild soybean accessions introduced from 84 countries or developed in the United States, consists of wild, landrace and North American cultivars. This collection was genotyped with the SoySNP50K Bead Chip consists of >50K single-nucleotide polymorphisms (SNPs).

**Data source:**

1. The SoySNP50K iSelect BeadChip has been used to genotype the USDA soybean germplasm (Song et al., 2015). Complete data set for 20,087 G. max and G. soja accessions genotyped with 42,509 SNPs is available for Wm82.a2 in vcf format <https://soybase.org/snps/soysnp50k_wm82.a2_41317.vcf.gz>.
2. Minor Allele Frequencies (MAF) of SoySNP50K SNPs. Contains 40,841 of the 47,337 SNPs (86%) had minor allele frequencies >10% among the landraces, elite cultivars and the wild soybean accessions <https://soybase.org/snps/snp50k_maf.txt>
3. Table S1, from Song et al., 2015. Fingerprinting soybean germplasm and its utility in genomic research. G3: Genes| Genomes| Genetics 50(10):1999-2006.)

Table S1. Description of 19,648 Glycine max and G. soja accessions genotyped with the SoySNP50K BeadChip. Data include the PI (Plant Introduction) number, Species (G. max vs. G. soja), Country of origin, Cultivar name if applicable, Maturity Group, 99.9% similar to another accession (Y/N), Landrace or North American Elite cultivars used for analysis and seed weight of accessions used in the association analysis of seed weight.

[Soy\_USDA\_Lines\_descritpion.xlsx](file:///C:\Users\gellima\Desktop\Springboard_DScience\CAPSTONE_3\'Soy_USDA_Lines_descritpion.xlsx')

**Data Merging:**

1. vcf file contains data related to SNPs, chromosomes, positions, genotype scores for samples. Read\_vcf file, convert the numpy arrays to dataframe.
2. Add MAF values of SNPs from \_maf.txt to dataframe.
3. Add samples information i.e Species, Country, maturity group, GE’s with 99.9% similarity to other lines, Landrace or N. Am. Cultivar and Seed weight) from Soy\_USDA\_Lines\_description.xslx to dataframe

**Data QC/Filtering**

1. SNP genotyping: Drop any polymorphic SNP with rate of missing & het alleles >0.1 among the 19,648 soybean and wild soybean accessions. The het allele calls in the remaining loci were set as missing in the subsequent analysis
2. Also, drop/filter SNPs based on MAF values listed in data source 2
3. Similarly, drop samples based on high % missing, outliers.

# Potential outcomes of Analysis

# The geographic distribution of the 120 wheat accessions used in this study. The circles represent landraces, while triangles represent cultivars. Different colors represent populations from different geographic regions. EA, East Asia; WA, West Asia; EU, Europe; AM, America; SCA, South and Central Asia.

# Similarity analysis

Genetic similarity between pairs of genotypes among the 18,480 cultivated and among the 1168 wild accessions calculated as the ratio of the number of identical SNP allele calls and the total number of SNPs for which allele calls were made for the pair

# Cluster analysis (KMeans, Hierarchial)

Pair-wise distance among the accessions of wild and landrace soybeans will be obtained based on the allelic dissimilarity of the 42,509 SNPs; *the neighbor-joining tree* will be constructed

1. Dendrogam of wild and landrace genotypes from different countries will be constructed.