* Focus on exploratory data analysis and visualizations. Make sure you are diving into the data and looking at different distributions, checking for outliers, missing data, etc. At a minimum make sure you have conducted preliminary correlations (for more predictive projects) and visualized key variables from your datasets that you will ultimately have in your models.
* Put together a brief written proposal for what you modeling approach will be. It is important to define in advance what models you anticipate using (in writing) so you can discuss what worked and what didn’t. You can always explore other modeling strategies after, but it's important to be able to say "We thought X would work, but it gave us Y results. Later on, we tried A... etc."

Goal:

Model and exploit correlated losses to predict new losses in crop insurance

Data:

Primary: Summary of Business: County X Crop X Year X Plan : Liability, Indemnity, Premium, etc

Optional:

SoB Practices : “ “ X Irrigation : “ “

CoL : County X Crop X Year X Plan X Cause X Timing : Above, and Acres Lost

Y: Loss Ratio, IF ( Loss>0)

X: Insurance and Loss History for Y’s crop and other crops

Models to Use:

Loop over major crops, based on data from the others.

Debdeep : Loss Ratio = { Past Results, Past and Current Results for Other commodities, Other }

See Moving Average?

Victor: Binary Tree

Jon: KNN

Compare fit and notes on Wednesday

EXPERIMENT UNTIL SOMETHING WORKS

CROSS VALIDATE

Known Problem:

‘Missing’ Data, most county-years will not include all crops from the 7

Around The Missing Data Problem:

Tighten Scope: Fewer commodities, counties

Define Sample tautologically rather than look for a region or period

CORE crops with a big all-shared set of county-years: Corn, Wheat and Soybeans

Aggregate:

County+Surrounding Counties, or a State, aggregated in the same tuple

All counties in 1 year or State is 1 observation, redefine Y variable to the average or a % to calibrate to

Pairs/Tuples :

Multiple Analyses

For each good subset of the 7, analyze separately, then compare

‘Missing’ Is The Signal:

Binary Variable for each crop : 1 if Missing, separate from amount of liability or loss

To still include actual value, impute mean

If standardized, the mean is now 0, so works cleanly

End Product:

Goodness of fit, validation tests

Planned Output:

Which crops signal for which others?

How closely can we calibrate / predict with only past data?

Compare to naïve averages