Introduction: Study Team, to study sponsor.

Purpose and agenda.

Problem statement and proposed study issues.

Constraints, limitations, and assumptions.

Methodology?

1. What type of data are we dealing with?

* P >> N.
* Mostly sparse.

1. What is the shape of the data?

* Not normal initially.
* Center, scale, and Box-Cox transform.
* Mardia test for multivariate Gaussian; p-value=\_\_\_.
* Inferential statistics simply aren’t possible without normality.

1. How do our variables correlate in our potent dataset?

* How many components of PCA contribute to about the total variance? (At least this many significant interactions exist)
* What features contain over 50% correlation? (I.e., more than half of their occurrences relate)

1. How do our variables correlate in our impotent datasets? (Select exemplar)
   * How many components of PCA contribute to about the total variance? (At least this many significant interactions exist)
   * What features contain over 50% correlation? (I.e., more than half of their occurrences relate)
2. Examine (3) and (4):
   * Are there features found in both?
   * Are they completely different? (Suspect)
3. How does PLS drop dimensions for potent data?
4. How does PLS drop dimensions for non-potent data?
5. What are the top interaction terms from a simple linear regression? (Top 10%)
6. What are the top feature importance’s from a random forest? (Top 10%)