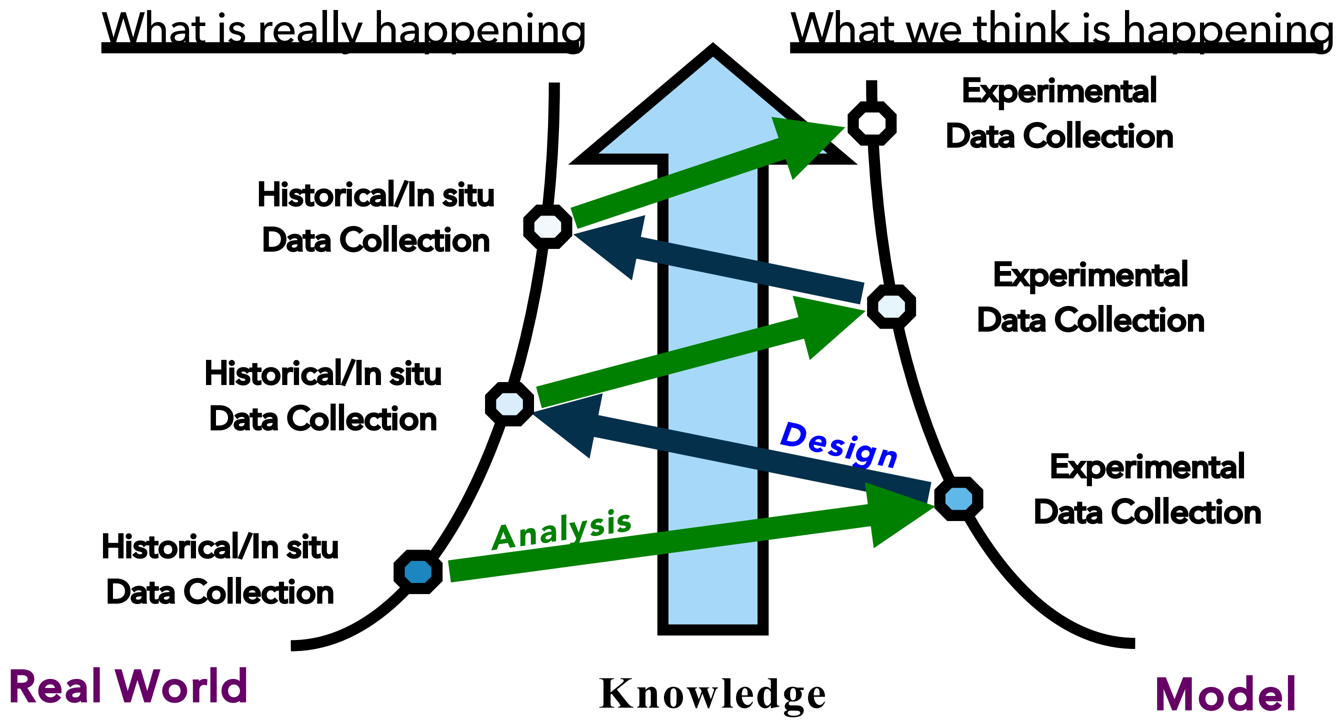
Combining Predictive Analytics and Experimental Design

Donald McCormack – Technical Enablement
don.mccormack@jmp.com

Statistical Discovery.™ From SAS.

Predictive Analytics/Experimental Design

Introduction



Adapted from Box, Hunter, & Hunter

Predictive Analytics/Experimental Design

Comparisons

- Predictive Analytics
  - Large number of inputs
  - Correlated factors
  - Complex relationship between inputs and outputs
  - Based on retrospective data
  - Flexible modeling techniques
  - Large and irregular factor space
  - Factor selection may be important
  - Prediction may be important
- Experimental Design
  - Small number of inputs
  - Little/no correlation among factors
  - Simple relationship between inputs and outputs.
  - Based on prospective data
  - Simple modeling techniques
  - Small and regular factor space
  - Factor selection may be important
  - Prediction may be important

Predictive Analytics/Experimental Design

Variable (Feature) Selection

- Experimental Design
  - Ad-Hoc
  - Stepwise Regression/All Subsets Regression
  - Design Specific (Fit DSD)
  - Model Selection for Designed Experiments Webcast
- Predictive Analytics
  - Variable Importance
  - Column Contributions (Trees)
  - Lasso
  - Other

Predictive Analytics/Experimental Design

Predictive Analytics

$$\mathcal{Y} = \mathcal{F}(\mathcal{X})$$

- Construct $\mathcal{X}$ is related to construct $\mathcal{Y}$ through function $\mathcal{F}$.
- Operationalized into mathematical function:

$$\boldsymbol{Y} = f(\boldsymbol{X}) + \boldsymbol{E}$$

- Where $\boldsymbol{Y}$ and $\boldsymbol{X}$ are quantifiable phenomena. $\boldsymbol{E}$ accounts for the disparity between the fitted model $f(\boldsymbol{X})$ and $\boldsymbol{Y}$.

Shmueli, G. (2010). To Explain or to Predict. *Statistical Science*, 25(**2**), 283 – 310

Predictive Analytics/Experimental Design

Predictive Analytics

$$\boldsymbol{Y} = f(\boldsymbol{X}) + \boldsymbol{E}$$

- An empirical representation that relates a set of inputs (predictors, $\boldsymbol{X}$) to one or more outcomes (responses, $\boldsymbol{Y}$)
  - $\boldsymbol{Y}$ is one or more continuous or categorical response outcomes
  - $\boldsymbol{X}$ is one or more continuous or categorical predictors
  - $f(\boldsymbol{X})$ describes predictable variation in $\boldsymbol{Y}$ (signal)
  - $\boldsymbol{E}$ describes non-predictable variation in $\boldsymbol{Y}$ (noise)
- The mathematical form of $f(\mathbf{X})$ can be based on domain knowledge or mathematical convenience.
- "Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful."

  – George Box

Predictive Analytics/Experimental Design

Predictive Analytics

- Common predictive modeling techniques:
  - Linear regression
  - Generalized regression
    - Also known as penalized regression or shrinkage methods. It is a technique applied to linear regression to account for correlated inputs. Ridge regression, LASSO, and Elastic Net are three examples.
  - Tree based methods: Partition (CART), Bootstrap Forrest (Random Forrest), and Boosted Tree.
  - Neural networks
  - Principal components regression (PCR) and partial least squares (PLS).

Predictive Analytics/Experimental Design

Predictive Analytics

- If the model is flexible what guards against overfitting (i.e., producing predictions that are too optimistic)?
  - Put another way, how do we protect from trying to model the noise variability as part of $f(\mathbf{X})$?
- Solution – Hold back part of the data, using it to check against overfitting. Break the data into two or three sets:
  - The model is ***built*** on the training set.
  - The validation set is used to ***select*** model by determining when the model is becoming too complex
  - The test set is often used to ***evaluate*** how well model predicts independent of training and validation sets.
  - Common methods include k-fold, random holdback and bootstrapping.

Predictive Analytics/Experimental Design

Experimental Design

- Designed experiments vary more than one input at a time.
- Inputs are varied independent of one another (or as close as possible).
- The experimental runs are determined in advance.
- The levels of the inputs are determined in advance.
- The order in which each experimental run is conducted is randomized.

Predictive Analytics/Experimental Design

Experimental Design

- All this is done to optimize the experimental goals established prior to running the experiment.
  - To find important input
  - To characterize a process or product
  - To find optimal settings
- DOE is the science of getting the most precise and accurate information from the experimental runs.

Predictive Analytics/Experimental Design

Experimental Design

More Factors – Fewer Runs/Factor – Less Detail

Fewer Factors – More Runs/Factor – Finer Detail

Add Other Effects — Add Q — Add 2 FI— ME Only — Some Confounding — Supersaturated

Custom Design
Full Factorial
RSD

Custom Design
DSD
Fractional Factorial
Plackett-Burman

Note: All things equal, fewer runs = more correlation

Questions?

jmp.com