ELDER RESEARCH
DATA SCIENCE & PREDICTIVE ANALYTICS

Beyond Reliability: Advanced Analytics for Predicting Quality

William J. Goodrum, Jr., PhD
Elder Research, Inc.
william.goodrum@elderresearch.com

Headquarters
300 W. Main Street, Suite 301
Charlottesville, VA 22903
434.973.7673 | fax 434.973.7875

Office Locations
Arlington, VA
Linthicum, MD
Raleigh, NC

Who Should Attend?

Job Titles:

- Statisticians
- Researchers
- Managers
- Lean professionals
- Process improvement team members
- Continuous improvement managers
- Master Black Belts
- Six Sigma Black Belts

When Individual Entities Matter

- **Context**
  - Natural Gas Wells in Wyoming
- **Problem:** Well "Freezing"
  - Costly to send maintenance into the field to prevent
  - Lost production during downtime

When should they go?

Where should they go?

What do maintenance crews need to know?

- Which well has stopped producing?
- Why has this well stopped producing?
- What is the nature of the failure?
- What service has already been performed?

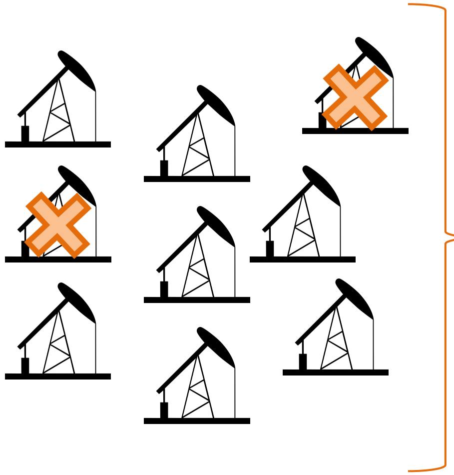
Reliability: A Question of Questions

Reliability: A Question of Questions

Which question is more important?

"Traditional"

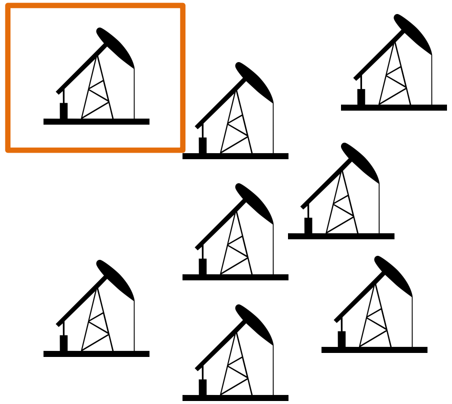
What proportion of wells will have an equipment failure in the next 180 days?



2/9

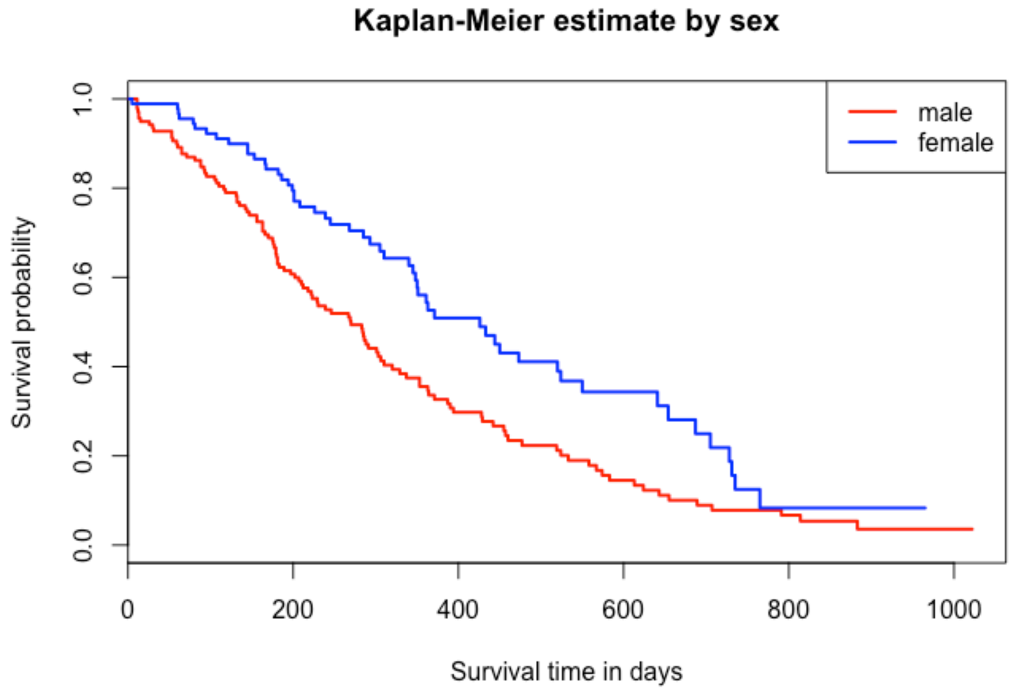
Predictive Analytics

What is the likelihood that *this particular* well will require maintenance in the next 180 days?

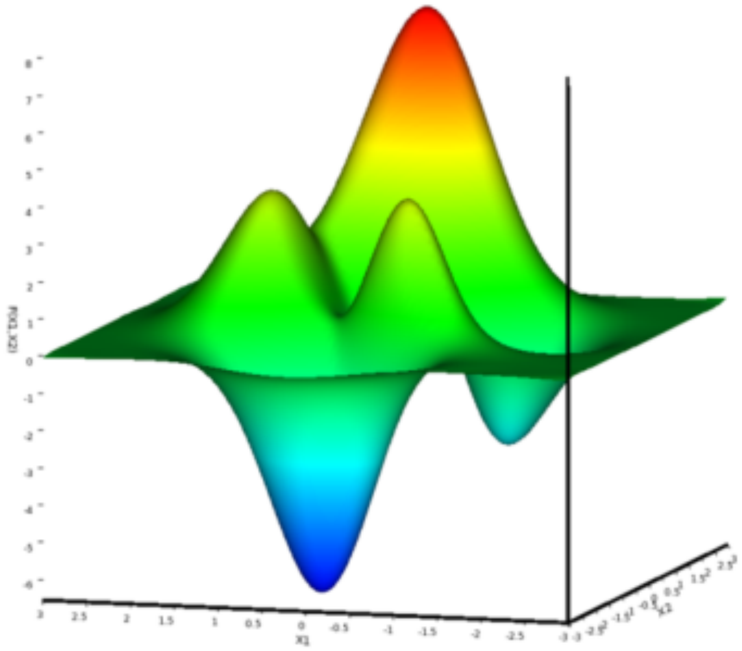


Traditional Paradigms

Kaplan-Meier/Cox Regression

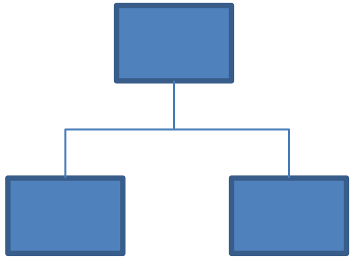


Design of Experiments

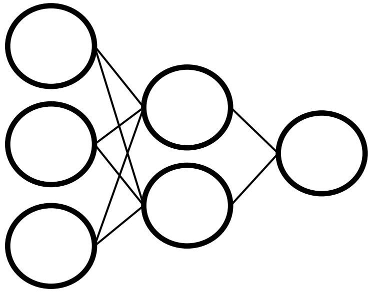


Predictive Analytics Methods

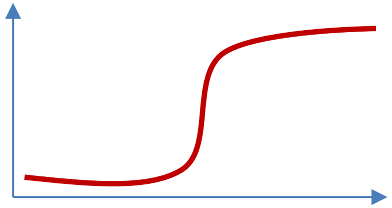
Decision Trees



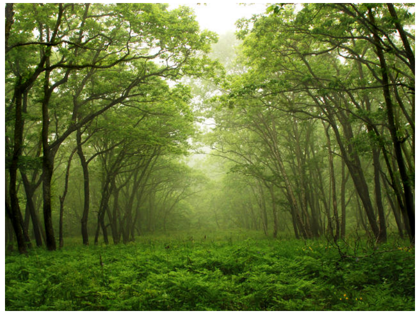
Neural Networks



Regression Analysis



Random Forests/ Ensembles



The Strengths of Traditional Methods

- We want to make a decision based on the reliability (or quality) of a ***population***
  - **How many or how much…?**
- Possible applications:
  - Estimating fleet overall performance
  - Budgeting for field maintenance
  - Population health management
  - Life/Long-Term Care portfolio analysis

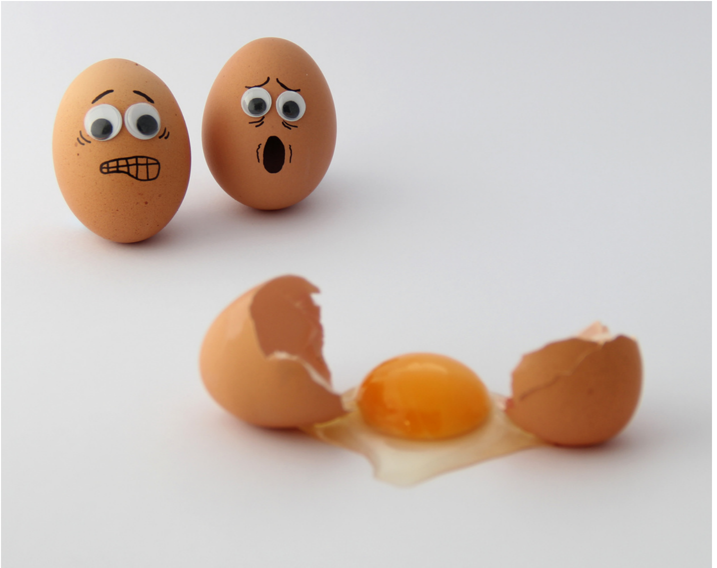
What do the traditional methods tell us about the wells?

Kaplan-Meier/Cox Regression

We might know what proportion of wells make it to a given life, but we may not know what contributes to the likelihood of failure for **a particular** well

Design of Experiments

We might know something about the wells in the center of our DoE, but we will know very little about the performance in extreme cases.

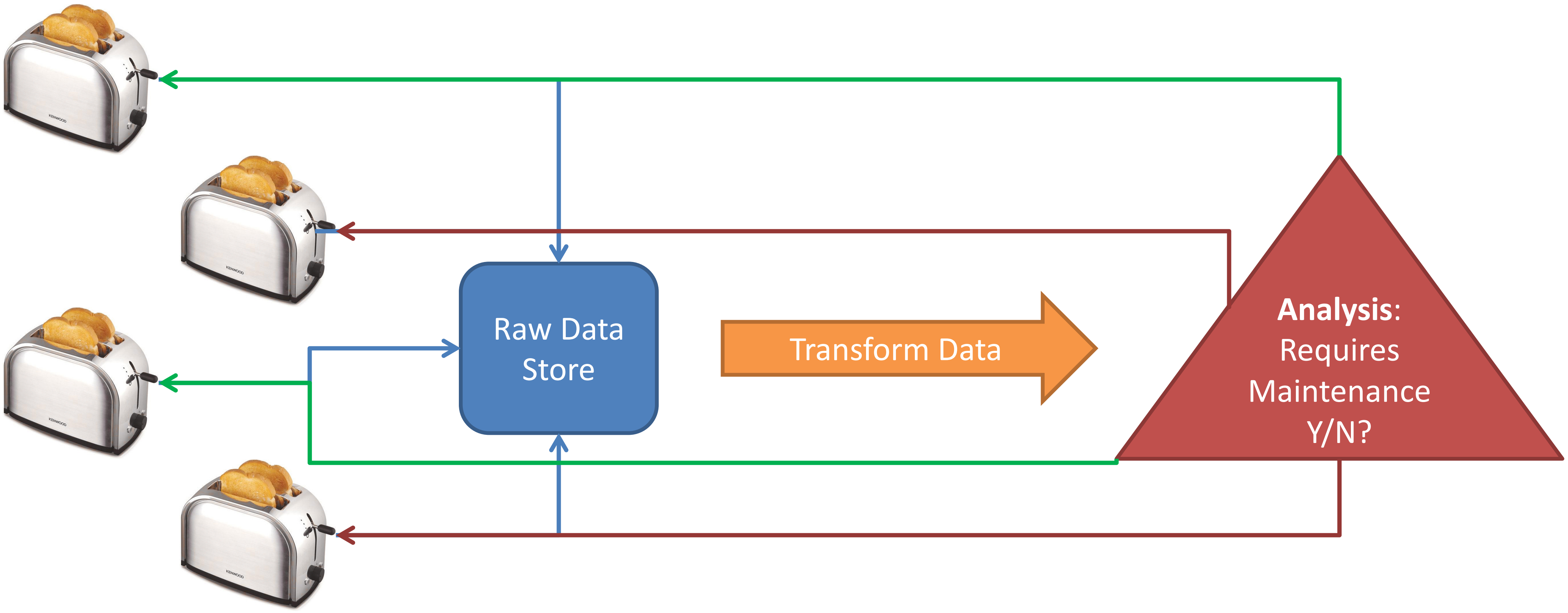


Predictive Analytics: A Complementary Paradigm

Case Study: Connected "Toasters"

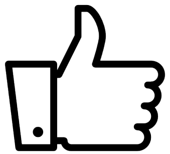
- **Client**: Connected "Toaster" Manufacturer
- **Goal:** use data as an asset for competitive advantage
- **Identified Opportunity**: preventative maintenance of "toasters"
- **Our Engagement:** Third-Party Validation

Unpacking the Opportunity



Elements of the “Toaster” Solution

- **Need:** Select/create variables related to maintenance and failures
- **Analysis Method:** Kaplan-Meier
  - Stratified by:
    - Date of Manufacture
    - Design
    - …
- **Goal:** Optimize for Consumer’s Risk of devices



Burnt "Toast": Limitations with Traditional Methods

- ***Average*** survival was correct, but ***particular*** survival probabilities did not match observations
- Some strata had no failures at all (e.g., newer "toasters")
- Excessive Stratification → Small samples!
- Assumptions for missing data grossly overestimated failures

Two Questions Again!

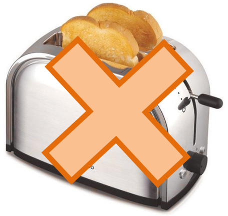
Our client's question

How many "toasters" can we afford to bring back for maintenance? How many failures in the field can we afford to have?

Their customer's question

*Does **my** toaster need maintenance?*

or



Applying Predictive Analytics to “Toaster” Data

- **Problem**: Classification
  - *Does* ***this*** *“toaster” require maintenance: Y/N?*
- **Method**: LASSO Regression
  - Combined variable selection and prediction
  - Mitigate overfit through regularization
- **Benefit**: also can estimate aggregate Consumer’s Risk

How did we compare?

Combining Traditional Methods with Predictive Analytics

- **Complementary Validation**
  - Identified similar feature space
  - Consumer’s Risk still matters!
- **Entity Failure Probabilities**
  - Likelihood of failure for individual “toasters”
  - Closer to user needs
- **Integration of Historical Performance**
  - Historical data for *each “toaster”* used to assess model performance

Back to Wyoming

The Old Challenges of Found Data

- >1 TB of ***ugly*** data

Some Challenges Included:

- Difficult integration
- Missing/sparse data
- Information unavailable until end of project (e.g., “freeze”)
- No information on well treatment (i.e., methanol pour-down)

More on “Freezing”

- **Initially:** well is “frozen”
- **Evolution 1:**
  - Subsurface Freezing (**more costly)**
  - Above-Ground Freezing
- **Evolution 2:** *any* downtime (including scheduled maintenance!)

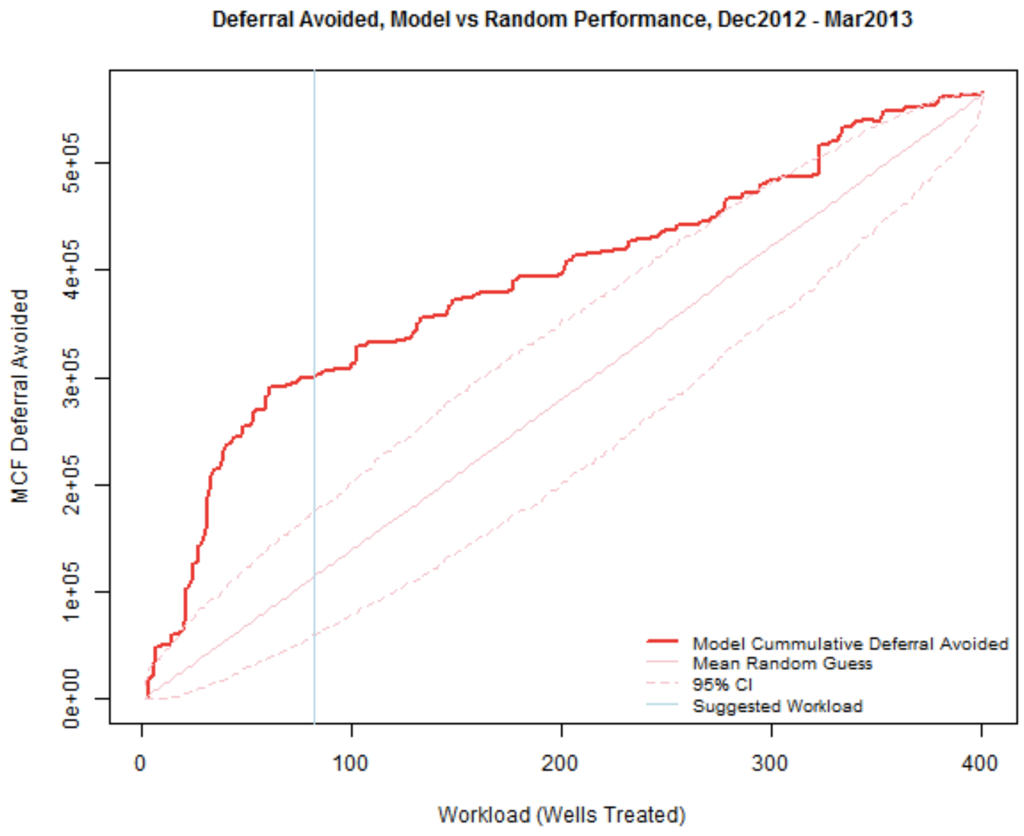
Starting Traditionally

- **Initial Analysis:** Kaplan-Meier
- **Key Insights:**
  - Field-level statistics still matter for resourcing/budgeting decisions
  - Fast and efficient statistics on aggregate well behavior
  - Repeat “freeze” 10x more likely after first freeze

Finishing with Predictive Analytics

- **Problem:** Classification
  - *How likely is **this** well to freeze in the next 180 days?*
- **Method:** Logistic Regression
- **Key Insights:**
  - Aggregated entity probabilities were more accurate than K-M
  - Significant additional effort

3x improvement over random baseline (at 20% workload)



ELDER RESEARCH
DATA SCIENCE & PREDICTIVE ANALYTICS

Success?

Summary

If the problem looks like this. . .

- Who. . .?
- Which. . .?
- Where. . .?
- When. . .?

. . . then predictive analytics may help like this

- **Who. . .?**
  - Prioritization of people for expert review
- **Which. . .?**
  - Highlight products of interest
- **Where. . .?**
  - Narrow geographical focus
- **When. . .?**
  - Likelihood of event in a given window of time

Reliability/Quality: Complementary Decisions

Traditional Methods

- **Decisions/Actions that Affect Groups**
  - High-level planning and cost analysis
  - Resource forecasting
  - “Portfolio” analysis

Predictive Analytics

- **Decisions/Actions that Affect Entities**
  - Maintenance Recommendations
  - Prioritization of Investigations/Audits
  - Resource scheduling

ELDER RESEARCH
DATA SCIENCE & PREDICTIVE ANALYTICS

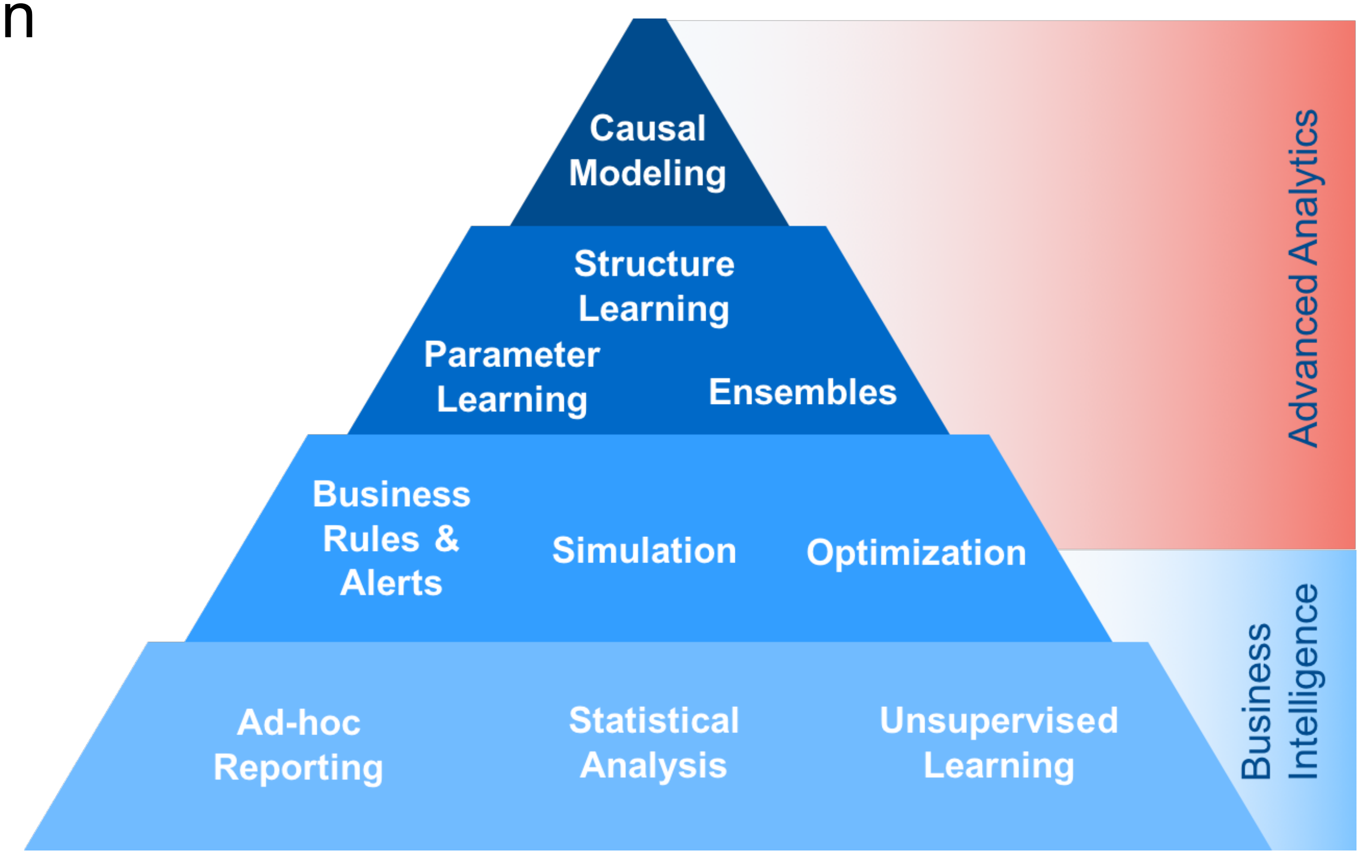
About Our Company

About Elder Research

- Founded in 1995 by Dr. John Elder

- **Offices**:
  - Charlottesville (HQ)
  - Arlington
  - Baltimore
  - Raleigh

- **Areas of Expertise:**
  - Data Science
  - Text Mining
  - Data Infrastructure
  - Data Visualization



Appendix

What is LASSO?

- **L**east **A**bsolute **S**hrinkage and **S**election **O**perator
- Generalized Linear Model (Logistic Regression is related)
- **Key Features**:
  - Budget on the sum of coefficients
  - “Regularization” term
- **Result:** prevents overfit, and helps select inputs!

About Me

Dr. William Goodrum is a Data Scientist with Elder Research; one of the oldest predictive analytics consultancies in North America. At Elder Research, Dr. Goodrum has led teams of Data Scientists and Software Engineers on a variety of different projects including analytics strategy development, predictive model validation, and predictive model building. These projects have been in industries as diverse as philanthropic development, maritime risk assessment, and connected device maintenance. He is also a frequent contributor to the company blog on analytics and analytics strategy.

Dr. Goodrum holds a B.S. in Mechanical Engineering from the University of Virginia, and a PhD in Engineering from Cambridge University.