# DOE: A Critical Component in the Data Scientist's Toolbox

Abigael C. Nachtsheim
Statistical Sciences Group
Los Alamos National Laboratory

Christopher J. Nachtsheim
Carlson School of Management
University of Minnesota

~~DOE: A Critical Component in the Data Scientist's Toolbox~~

# From Agriculture to Artificial Intelligence: Innovation, Collaboration, and Rapid Growth in DOE


**Abigael C. Nachtsheim**
**Statistical Sciences Group**
**Los Alamos National Laboratory**

**Christopher J. Nachtsheim**
**Carlson School of Management**
**University of Minnesota**

# Claim: Use of designed experiments is exploding

## Anecdotal evidence from a 45-year career in DOE:

- Teaching executive MBAs in 1984 versus 2022
- Growth spurts in DOE over my career
  - 1960s/70s: Use mainly in large companies (duPont, 3M, P&G, Ford, GE)
  - 1980s: Taguchi method adopted by engineers for quality improvements
  - 1990s/2000s: Six sigma (“Improve” phase called for DOEs)
  - 2010s: A/B testing, computer experiments, definitive screening designs and more
- Experience at Los Alamos: 1979 versus 2023

# Our talk will focus on the following growth areas

1. Industrial experiments
2. Social media and online marketing experiments
3. AI modeling

# Industrial experimentation

- Designs are now easily accessible thanks to software (e.g., JMP)
  - classical designs
  - optimal designs
  - designs for computer experiments

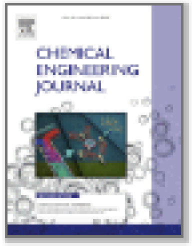
- Definitive screening designs allow for screening and optimization in one step, which has motivated a lot of experimentation

# Recent example

ELSEVIER

Chemical Engineering Journal

Volume 259, 1 January 2015, Pages 126-134



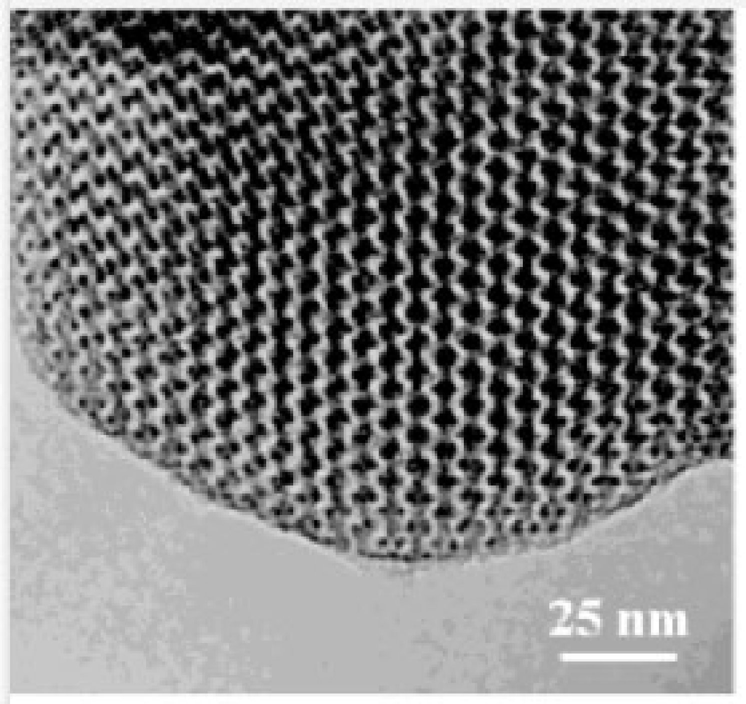

## Optimization of soft templated mesoporous carbon synthesis using Definitive Screening Design

Wannes Libbrecht [a, b, c], Frank Deruyck [d], Hilde Poelman [b], An Verberckmoes [a], Joris Thybaut [b], Jeriffa De Clercq [a], Pascal Van Der Voort [c]

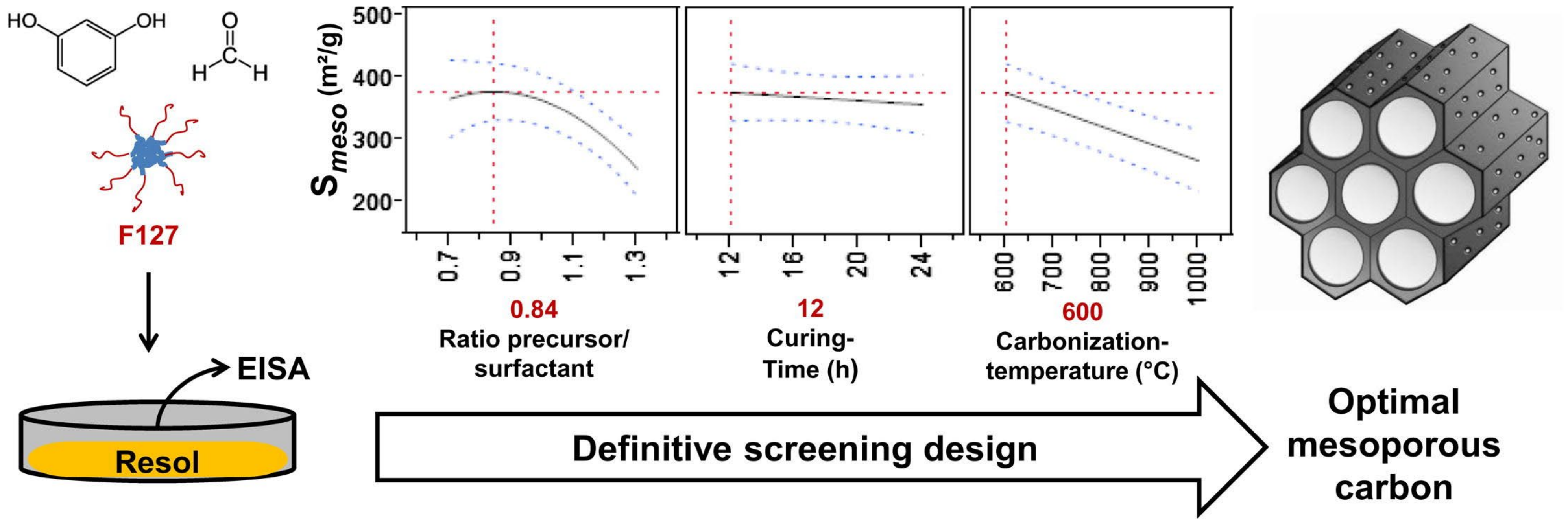
# What is mesoporous carbon?

**Promising new material containing pores with diameters between 2 and 50 nanometers**

- **Act as catalytic support in chemistry**
- **Use in energy storage devices**
- **Can control body's oral drug delivery system**
- **Can adsorb poisonous metal from water.**

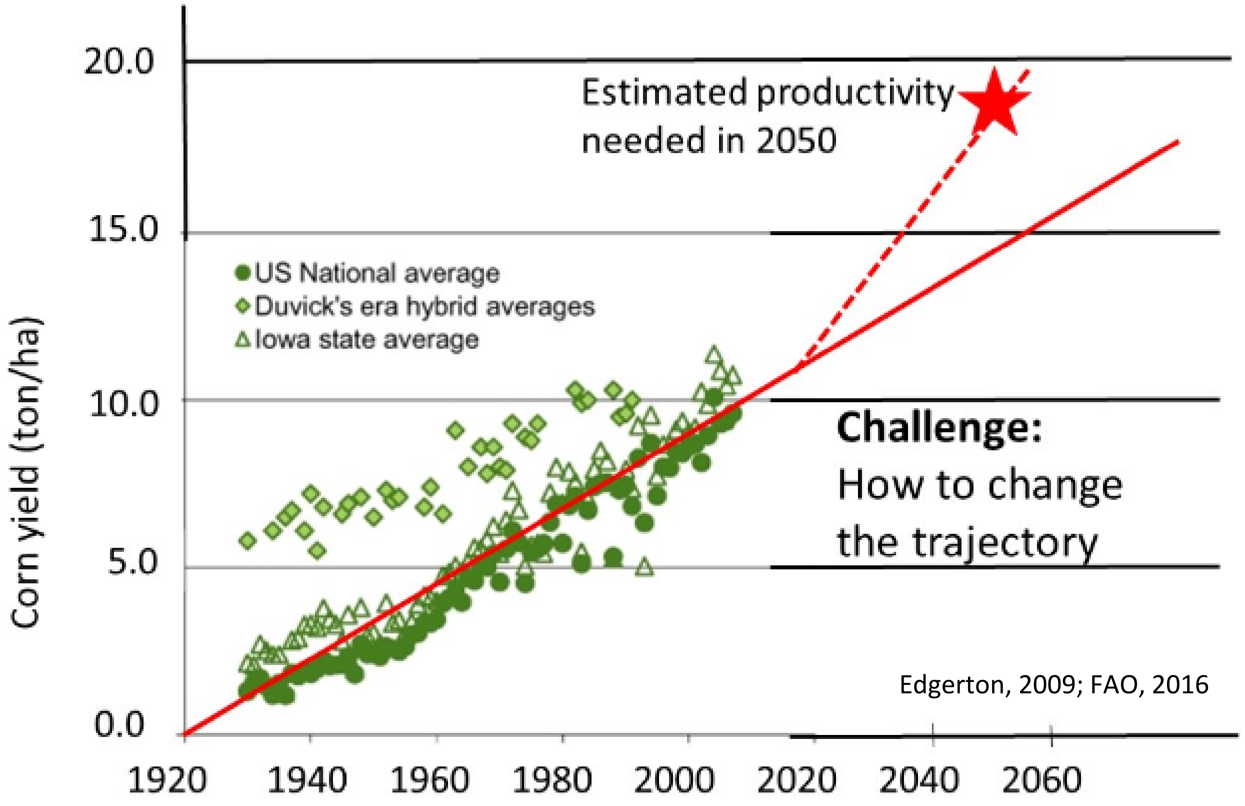


# “Graphical abstract” from the paper



# Two applications from Los Alamos National Laboratory

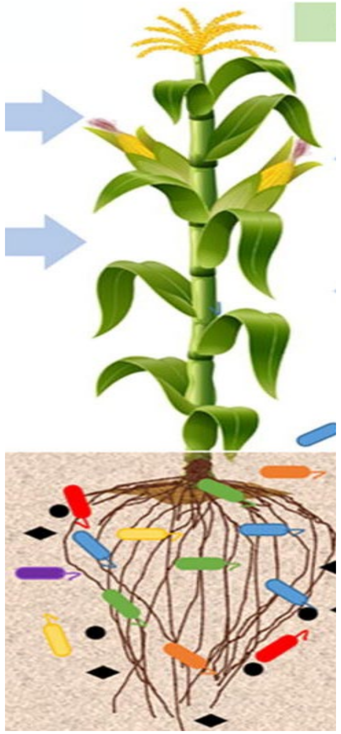
# Motivation: Current increase in crop yield is insufficient for meeting future needs




- **Droughts are most common cause of crop failures**
- **Becoming more frequent**
- **Agriculture: 70% of all freshwater use**
- **Drought resistant crops are essential**

# Developing drought resistant crops

- Targeting crop directly has drawbacks
  - Genome editing, traditional breeding methods, bioengineering
- Instead: Use directed evolution to evolve a microbiome that improves plant performance under drought



Adapted from:
Compant et al. 2019

# In principle, directed evolution is simple

- Select soil microbiome that leads to desirable plant function
- Propagate microbiome to the next generation
- Repeat


BUT:

- Which plant functions most affected?
- How many generations will it take?

# Strategic Approach is Needed

- Team included experts in
  - Plant physiology
  - Plant biochemistry and epigenetics
  - Metabolomics
  - Soil biogeochemistry

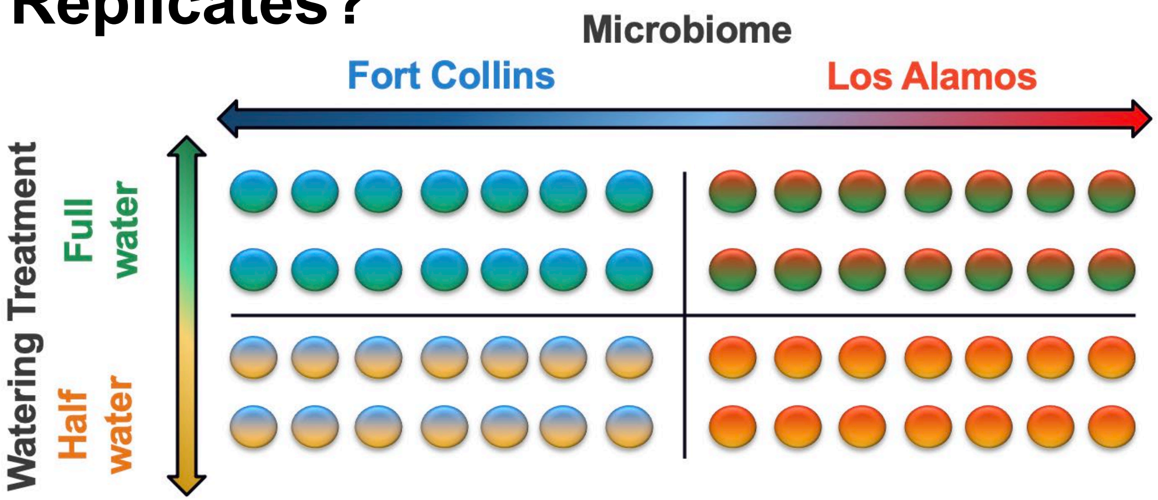
- Added statisticians with expertise in DoE

- Develop sequential design of experiments strategy

# DoE: Multistage approach

1) Pre-directed evolution experiment
   - Identified plant drought tolerance traits most affected

2) Strategy for directed evolution designed experiment
   - Sequential, multi-generational experiment
   - Full factorial structure
   - Which to propagate? Sample size? Replicates?



# Developed Successful Sequential DoE Strategy

- Analysis of preliminary data + subject matter expertise
- Result: Multi-generation study conducted over 3 years

**Results:**

- Directed evolution affects the microbiome
- Produced statistically significant differences in plant performance
- Publications, ongoing research

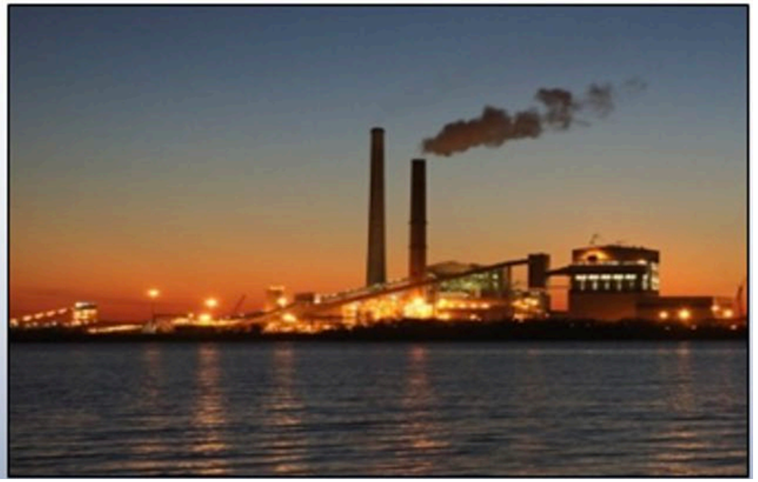
# Carbon Capture Simulation for Industry Impact

- **Partnership among national laboratories, industry and academic institutions**
- **Goal: Accelerate the commercialization of carbon capture technologies**

**National Labs** | **Academia** | **Industry**

# DoE Improves Efficiency in Scaling Up New Technologies

- **Industry partners developing carbon capture technology**
- **Test technology to further refine, improve**
- **Need DoE for efficient lab- bench- and pilot-scale testing**
- **Technical director's message: DoE can save years (and millions of dollars) off test schedule**

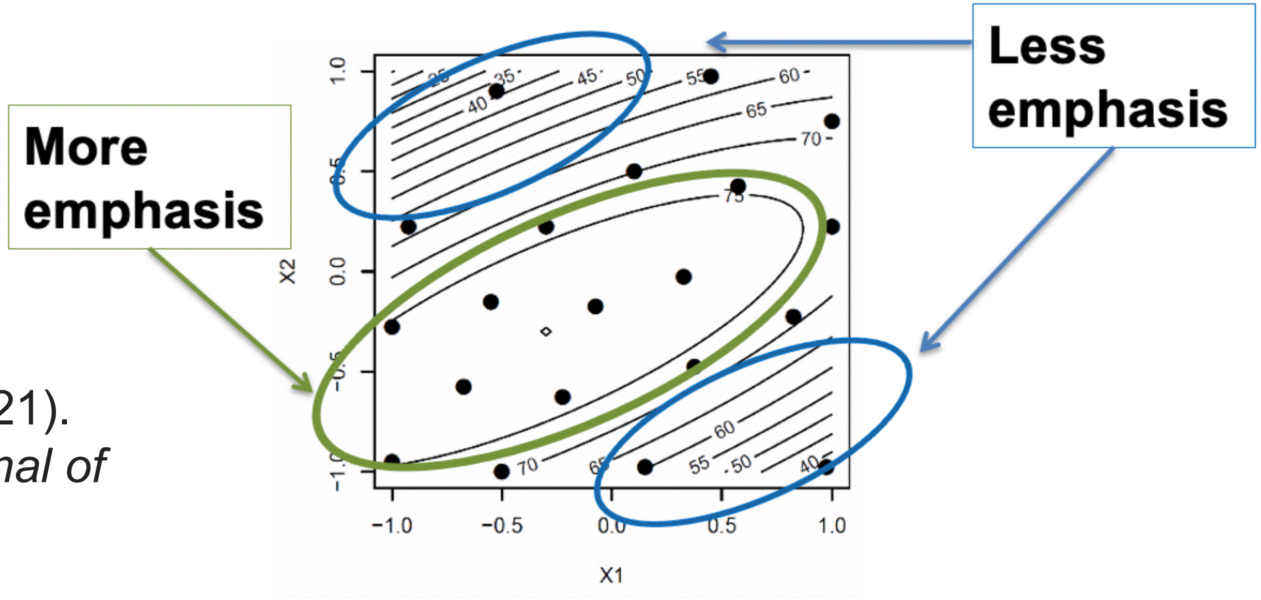


# Pilot Scale Testing at Technology Centre Mongstad (TCM)

- **TCM is world's largest test centers for carbon capture technology**
- **Testing time limited**
- **Need for strategic, cost-effective approach**

# Sequential DoE

- Partition experimental budget to collect data in stages
- Use a combination of Uniform and Non-Uniform Space-Filling Design*
  - Allows for more in-depth exploration of areas of interest
- Directly incorporate knowledge learned in previous stages
- Result: Strategic data collection across multiple stages




*Lu, L., Anderson-Cook, C. M., & Ahmed, T. (2021). Non-uniform space filling (NUSF) designs. *Journal of Quality Technology*, *53*(3), 309-330.

# Successful Collaboration with Industry Partners

- Industry partners are enthusiastic to collaborate on DoE-based testing
- Successfully complete several pilot-scale test campaigns at TCM
- Proven track record
  - Over 25% reduction in model uncertainty
  - Precisely predict $CO_2$ capture percentage; hit desired targets in testing

# Social media and on-line marketing

# Iconic example

- **In 2012, Bing employee suggested changing how ad headlines display.**
- **Simple idea: Lengthen the title line of ads by combining it with first line below title.**

Pic here

# Just try it!

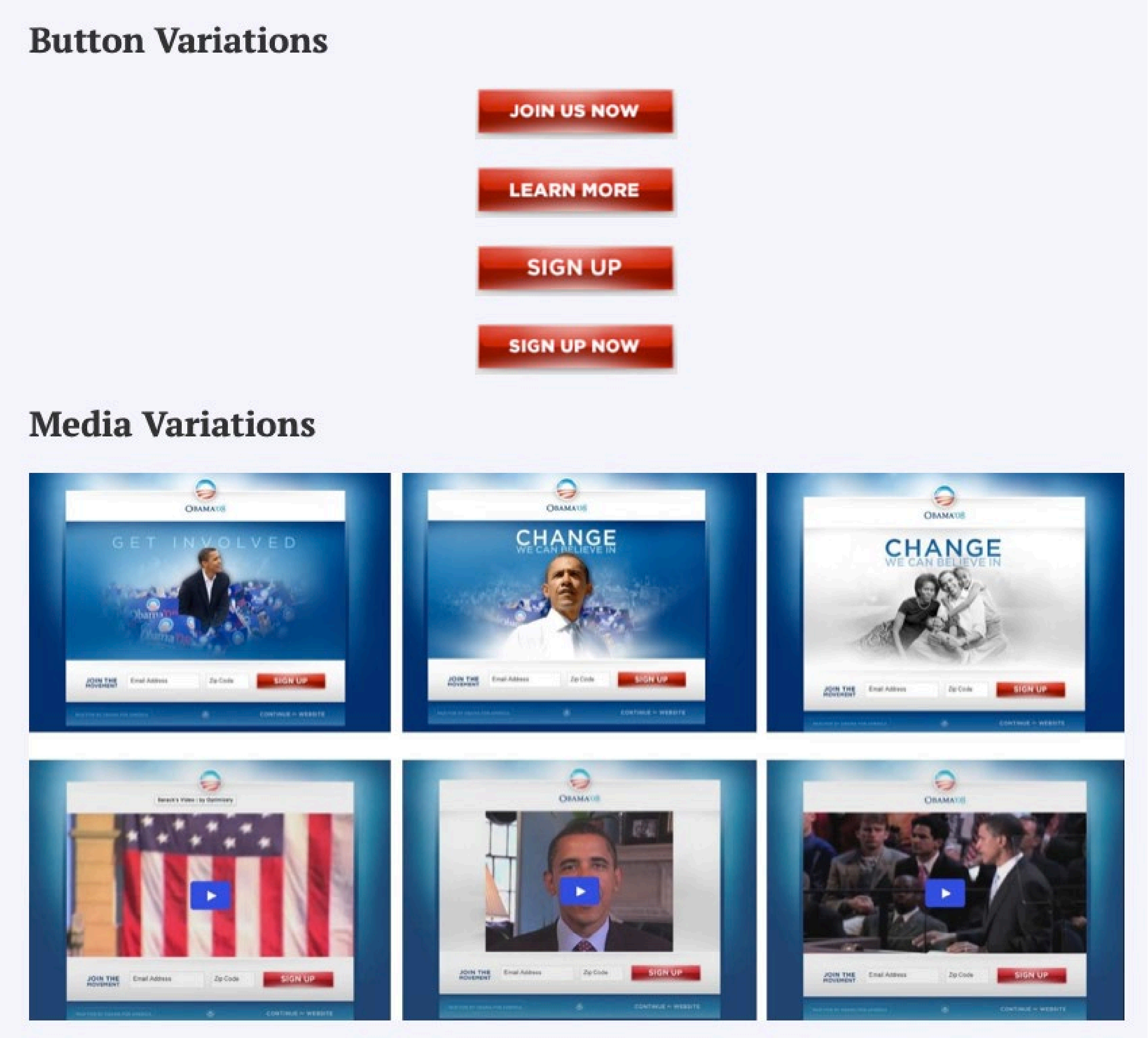
- **Feature was given low priority and languished for 6 months**
- **Finally, software developer decided to just try it.**
- **Showed random sample of users new title layout and another random sample the old version**
- **Recorded ad clicks and revenue**

# Single-factor experiment

- Compared treatment and control
- Referred to as A/B test
- Result: revenue increase estimated at 12%
- Translated to $100M in annual revenue for Bing
- Not bad for an afternoon's work

# Obama iconic A/B/n example

- 4 buttons x 6 videos
- 4x6 full factorial design
- Called an A/B/n test by computer scientists
- Increased contributions by $60M



# Other examples

- Google: 41 shades of blue
- Microsoft: Color tweaks improved user productivity, $10M annually
- Amazon: Moved credit card offer from home page to shopping cart

# A/B testing is everywhere

- Companies are creating departments of “Experimentation” or “Causal Modeling”
- Examples: Apple, Amazon, Google, Meta, Netflix, Atlassian, Intuit, Best Buy, Chewy, Etsy, TikTok
- So many experiments running simultaneously leads to problems:
  - Interference
  - Network effects

# Need to keep use of subjects to a minimum

- DOE can help---some recent work:

## Max Entropy Designs for Online Evaluation of Machine Learning Models

Gautham Sunder[a], Thomas A. Albrecht[b], Christopher J. Nachtsheim[a]

[a] *Carlson School of Management, University of Minnesota, Minneapolis, MN*
[b] *Atlassian Corporation*

# Two competing AI models: which is best?

- Ideal issue for a designed experiment (single factor design)
- Typical approach:
  - Random sample of users
  - Each model has a prediction for given user
  - Assign user to one of the models at random, observe outcome
  - Record the performance of the models

# Sample size reduction

Gautham Sunder

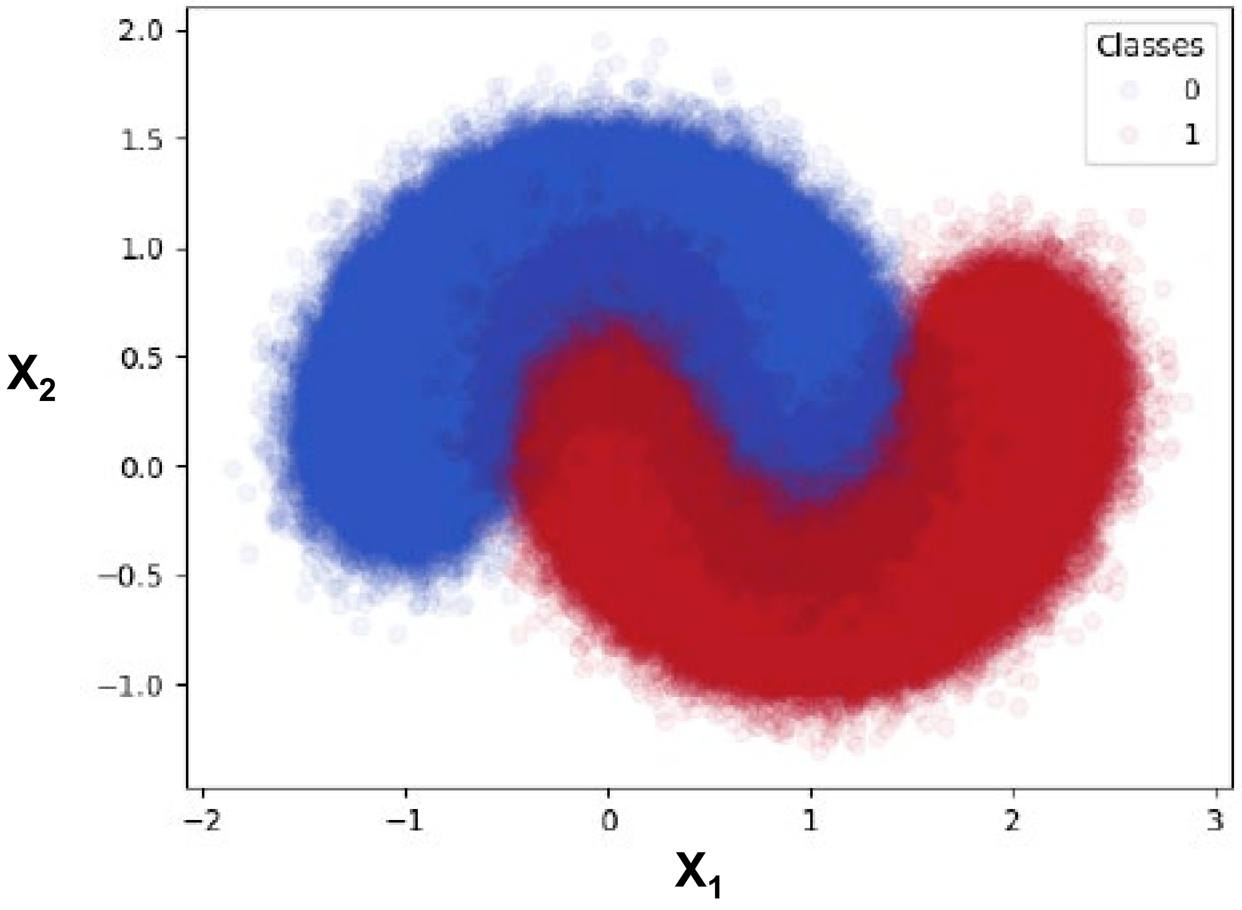
Gautham Sunder's approach:

- Only employ a user if the models differ in their predictions for the user
- Employ Box and Hill (1967) design criteria for model discrimination
- Points (users) chosen maximize the Jenson-Shannon Divergence for the models

# Example: Two classes of users in population

**Blue = Democrats**
**Red = Republicans**

**Predict class on basis of $X_1$ and $X_2$**

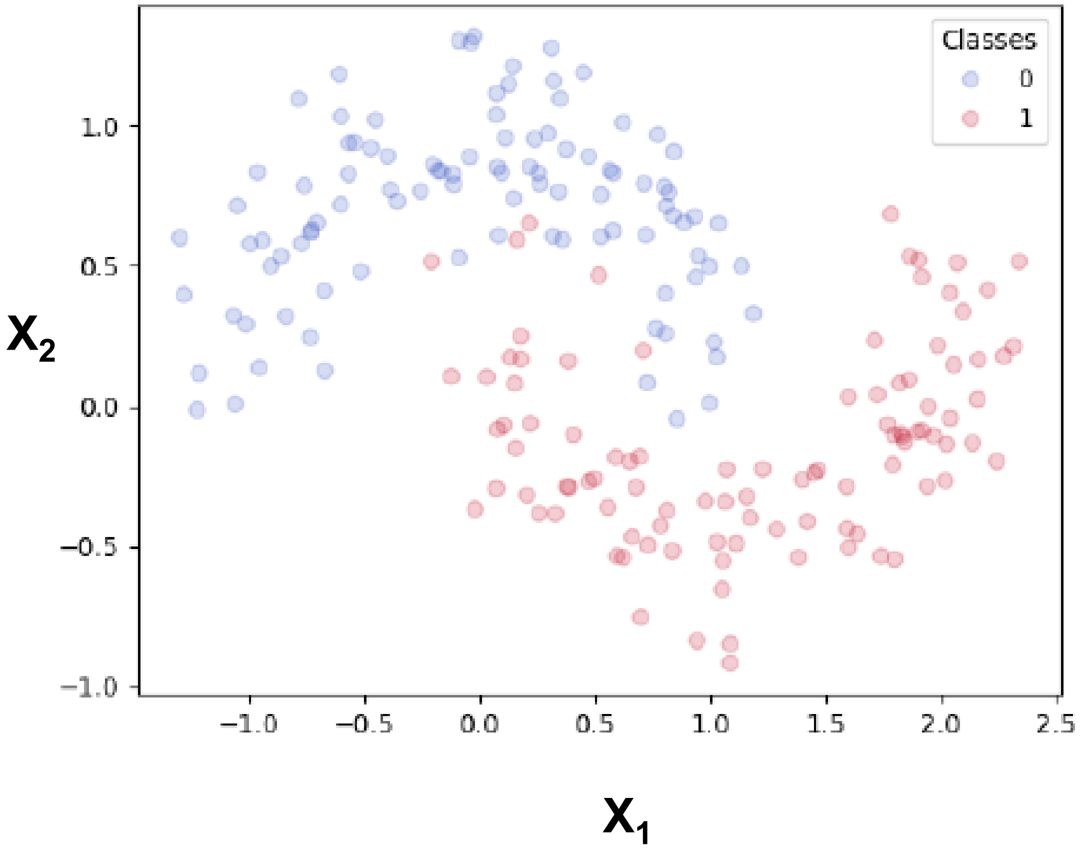


# Random sample is taken, two ML models fit

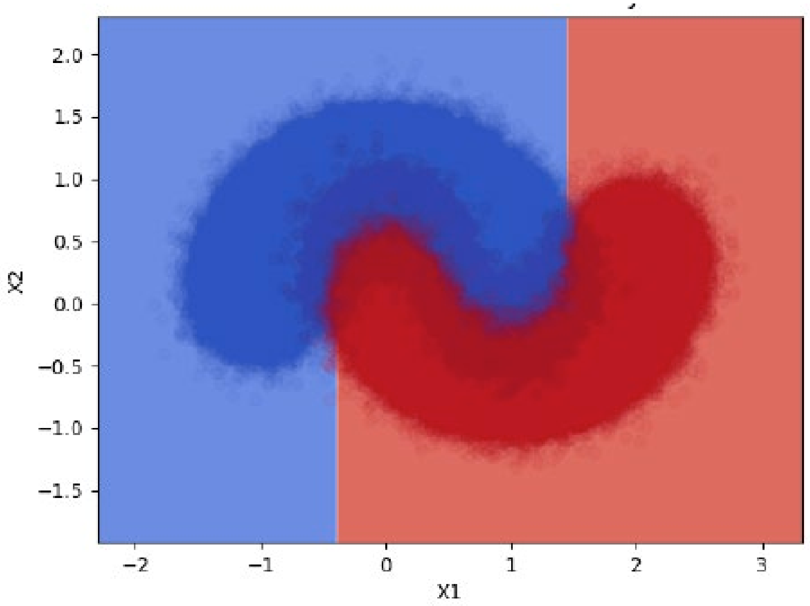
**ML Models:**

- **Decision Tree (DT)**
- **Support vector classifier (SVC)**

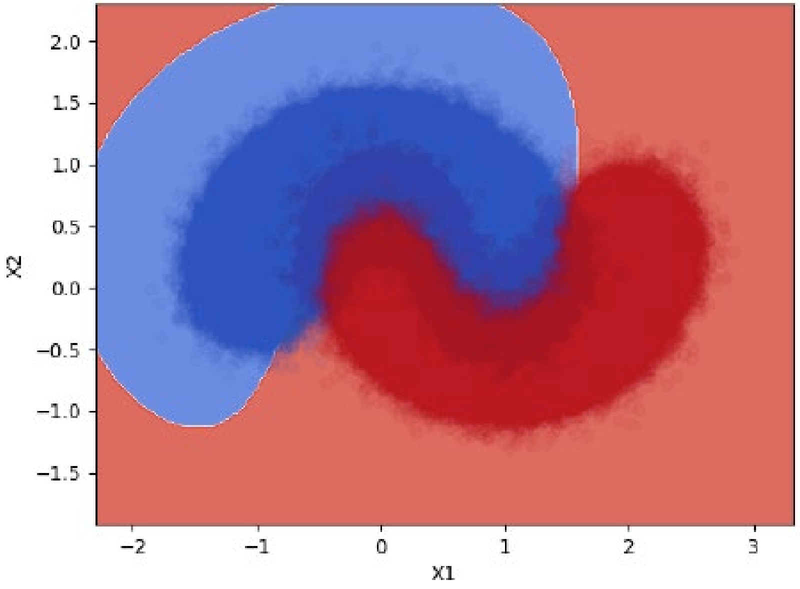
**(We know that SVC is the superior model when applied to the population)**



# Model (estimated) boundaries

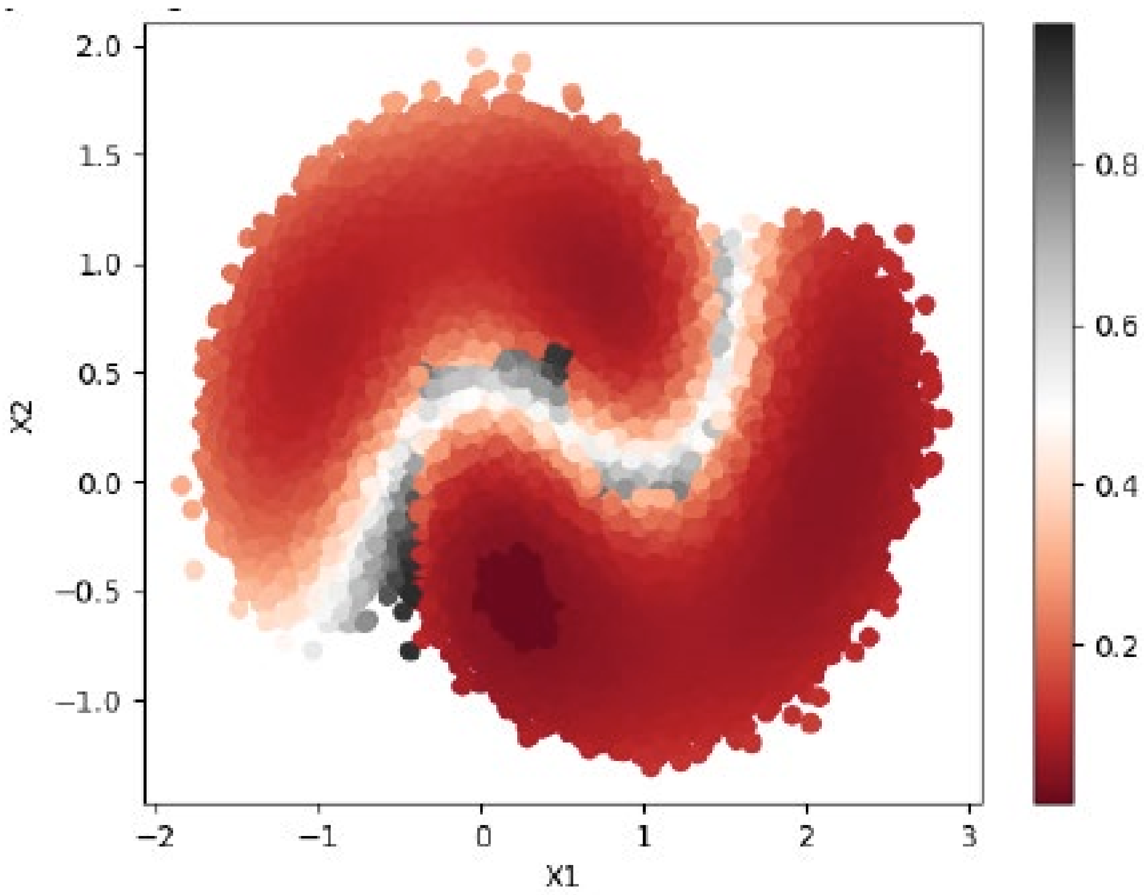



Decision Tree boundaries




SVC boundaries

# JS Divergence for the population




**Result: Significant sample size reduction**

# DOE and AI Modeling:
# Two halves of algorithm development

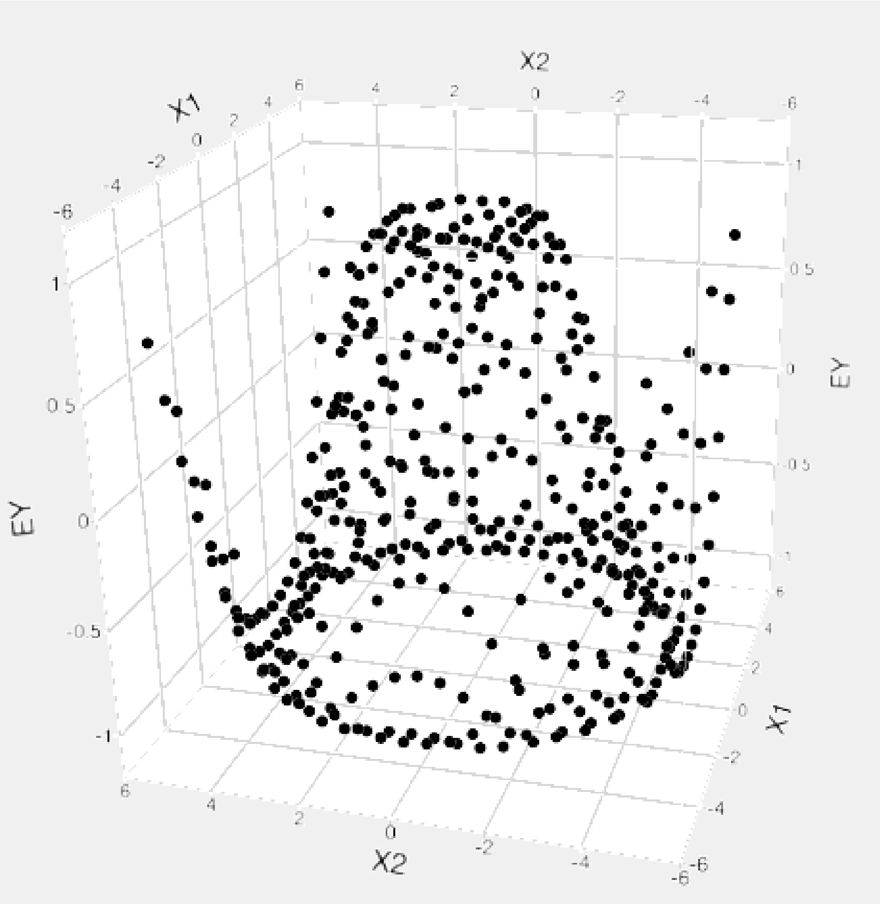
**Hyperparameter optimization (Chris)**
**Test set development (Abby)**

# What is hyperparameter optimization?

**Simple example:**

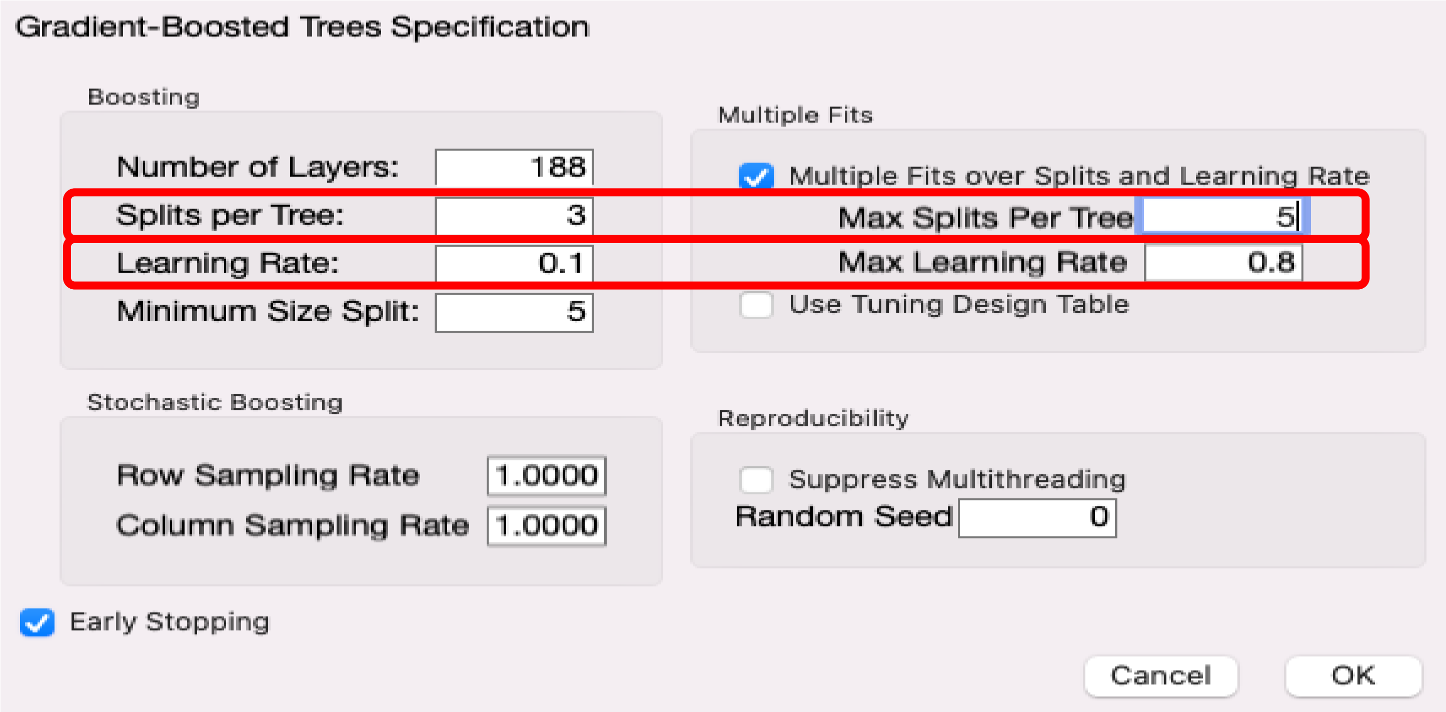
**Optimize a boosted regression tree using JMP**

# Two predictors: Cowboy Hat Data



# Fit a regression boosted regression tree

- **Two hyperparameters: learning rate and splits per tree**




**JMP allows the user to vary Splits per Tree and Learning Rate:**

**Splits: 3, 4, and 5**
**LRate: 0.1, 0.2, 0.4, 0.8**

**Aha: A 3x4 design!!**
**Pick the best result**

# JMP Result: For best $R^2$--Nsplits = 5, LR = 0.4

## Boosted Tree for Y

### Model Validation-Set Summaries

The fit below was the best of these models fit.

| N Splits | Learning Rate | Row Sampling Rate | Column Sampling Rate | N Layers | N Layers Specified | Minimum Size Split | RSquare | RASE |
|---|---|---|---|---|---|---|---|---|
| 3 | 0.1 | 1 | 1 | 188 | 188 | 5 | 0.8815 | 0.2398 |
| 4 | 0.1 | 1 | 1 | 188 | 188 | 5 | 0.9448 | 0.1469 |
| 5 | 0.1 | 1 | 1 | 188 | 188 | 5 | 0.9473 | 0.1633 |
| 3 | 0.2 | 1 | 1 | 188 | 188 | 5 | 0.9573 | 0.1424 |
| 4 | 0.2 | 1 | 1 | 188 | 188 | 5 | 0.9638 | 0.1297 |
| 5 | 0.2 | 1 | 1 | 188 | 188 | 5 | 0.9763 | 0.0972 |
| 3 | 0.4 | 1 | 1 | 188 | 188 | 5 | 0.9684 | 0.1207 |
| 4 | 0.4 | 1 | 1 | 188 | 188 | 5 | 0.9512 | 0.1428 |
| 5 | 0.4 | 1 | 1 | 188 | 188 | 5 | 0.9808 | 0.0952 |
| 3 | 0.8 | 1 | 1 | 188 | 188 | 5 | 0.9640 | 0.1237 |
| 4 | 0.8 | 1 | 1 | 188 | 188 | 5 | 0.9673 | 0.1205 |
| 5 | 0.8 | 1 | 1 | 188 | 188 | 5 | 0.9792 | 0.0947 |

# Hyperparameter optimization and AI Models

Gautham Sunder

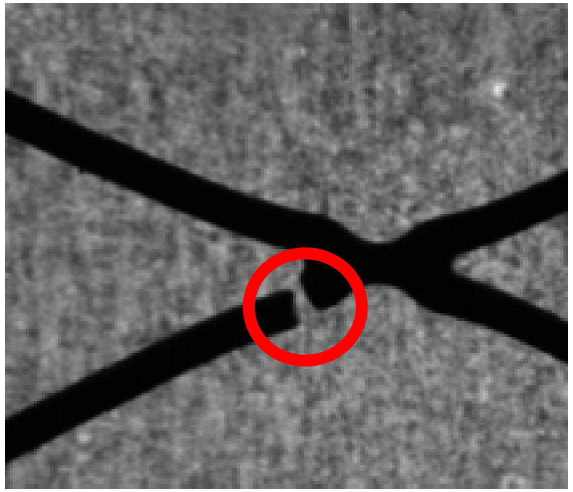
Tom Albrecht

**(2019) Tom Albrecht: "Chris, do you have a grad student to work on AI problem with Boston Scientific?"**

**Chris: I do! Gautham Sunder**

# Boston Scientific goal

- Identify defects on surface of stents using image recognition
- Example:



# Examples of hyperparameters for AI model

- Initial learning rate
- Fine tuning learning rate
- Dropouts
- L2 Regularization weight
- Architectures (e.g., layers, nodes)
- Input image transformations

# Boston Scientific approach

- Response surface optimization (RSO) using full quadratic models
- I-optimal saturated design for 12 hyperparameters (91 parameter regression model)
- Run experiment, fit model, optimize to find best hyperparameter values
- Tom Albrecht: ”RSO seems to work well”

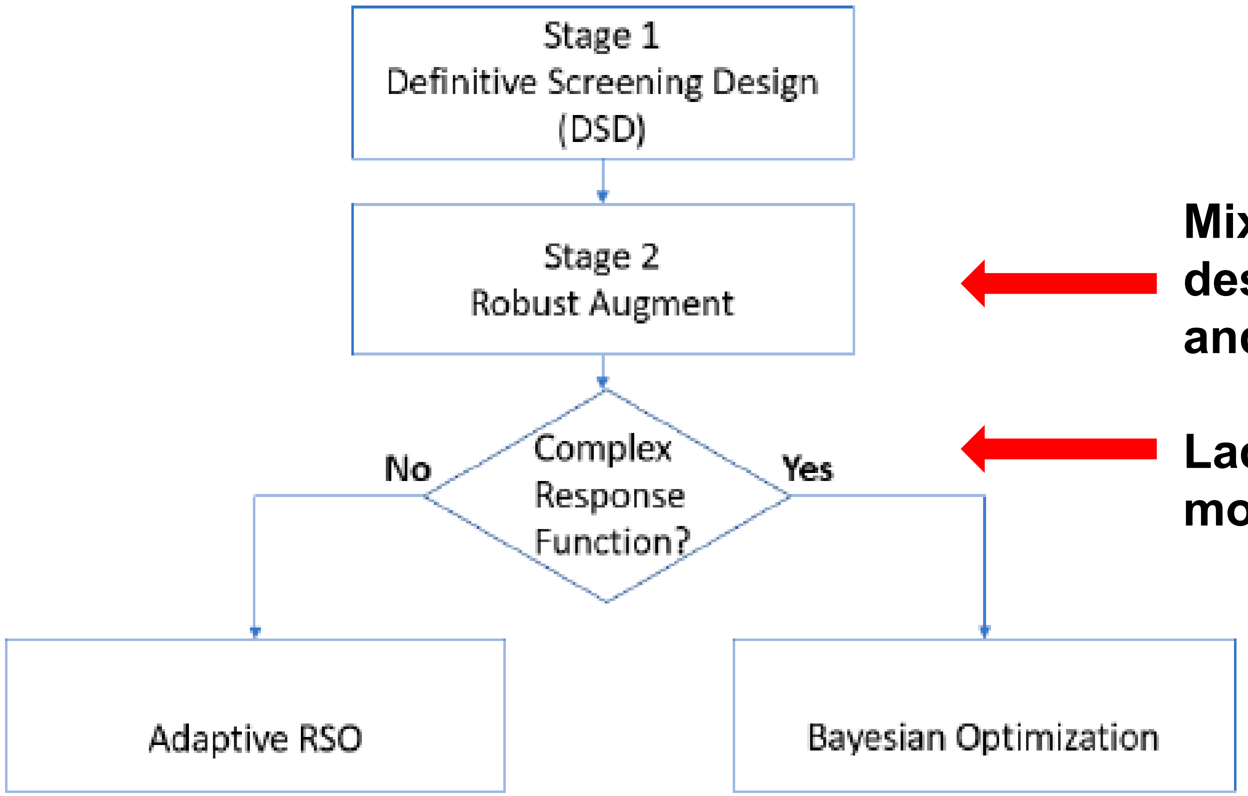
# Problems

- RSO works well only if response surface is locally quadratic
- Computer science literature
  - AI surfaces are always complex, never locally quadratic!!
  - Must use space-filling designs, Gaussian process models, Bayesian optimization. (Expensive and slow!)
- Statistics literature:
  - 60 years of success with response surface (locally quadratic) methods.
  - Why should this problem be different?

# So who is right: Statistics or computer science?

- Computer science approach is best for complex surfaces
- RSM approach is cheaper, more effective if locally quadratic
- Problem: We don’t know in advance!!
- Solution: Let the data decide

# Gautham's approach: 3-stage RSO

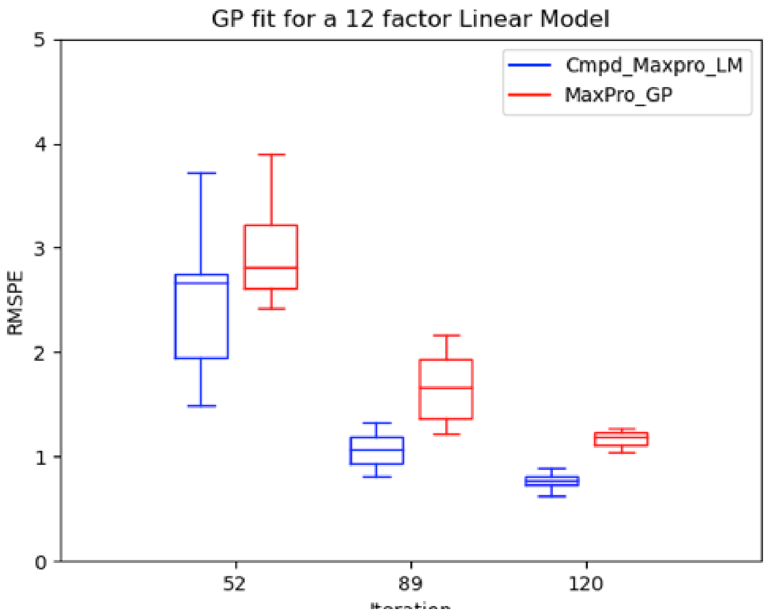



**Mixture of Bayesian D-optimal design for full quadratic model, and, space-filling points**

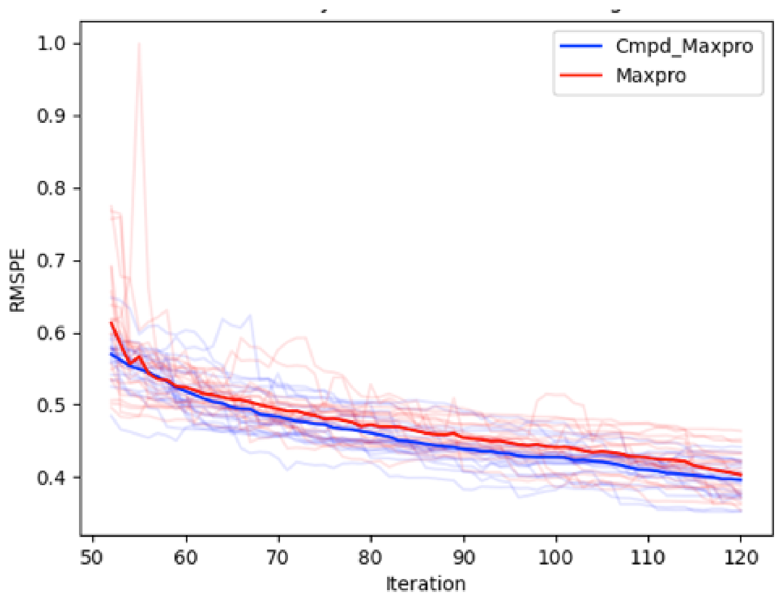
**Lack of fit test for quadratic model**

# Hint at results: Blue = RSO, Red = Space filling/BO

## Model is locally quadratic




## Model is complex

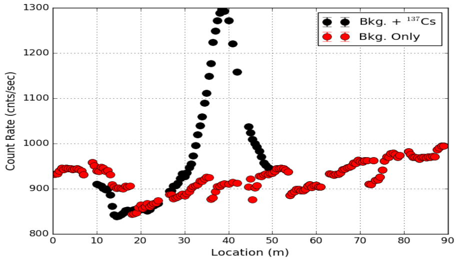



**(Smaller is better)**

# Developing test sets at Los Alamos National Laboratory

# DoE for Test and Evaluation of Algorithms

- Growing interest and work*
- Application area: urban radiological search
- Algorithms to detect, locate, and characterize radiological sources
  - Radiation detector measures energy spectrum
- Synthetic data is only option

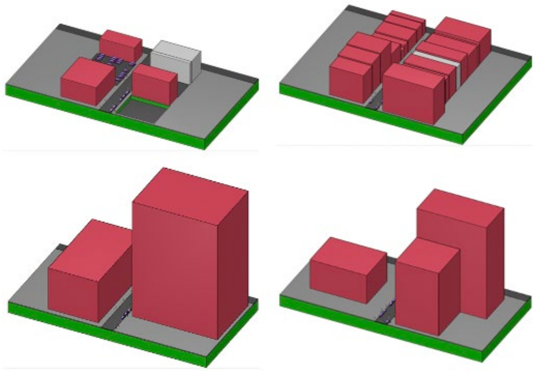

*Kary Myers, Christine Anderson-Cook, Lu Lu, others

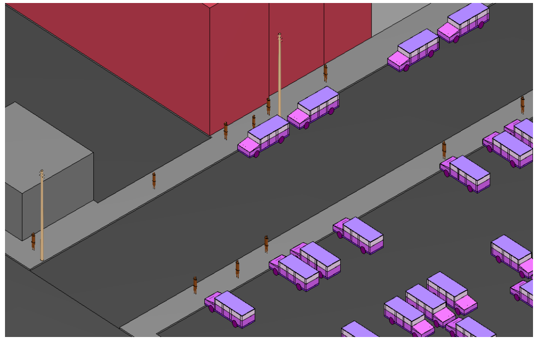
# Test Set Requirements

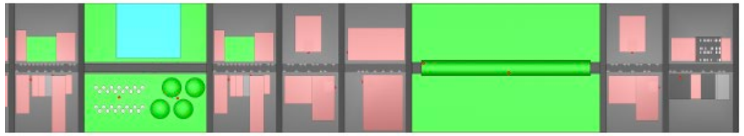
- Develop community standard benchmark dataset
  - Evaluate fieldable ML algorithms
- Complex backgrounds, source types
- Sufficient source encounters
  - Estimate performance
- Varied difficulty
  - Distinguish performance

# Develop Sequentially

- **Large scale Monte Carlo simulation (building blocks)**
- **Layers of probabilistic sampling (hour-long routes)**
- **Design of experiments (source encounters)**

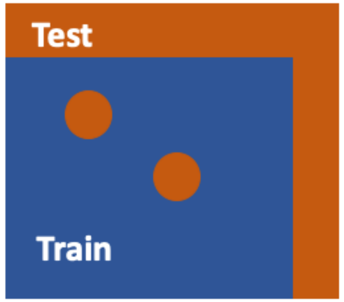






# Dataset Design Methodology

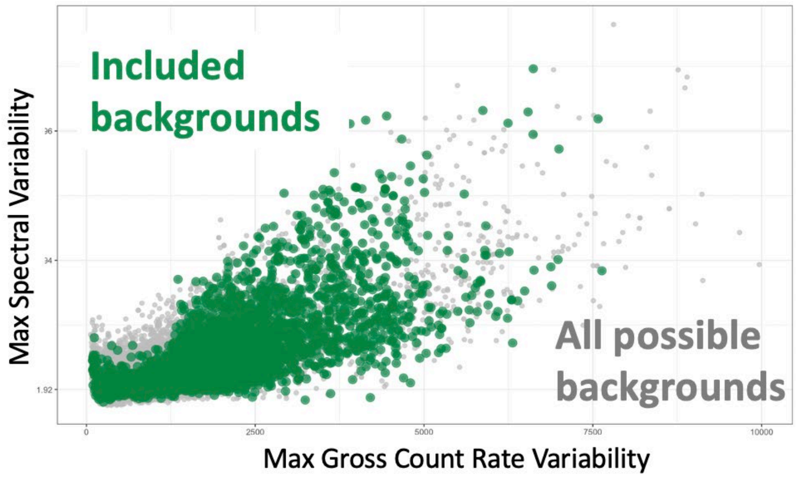
- To accommodate fixed routes: Nonparametric strategic subdata selection
- For source encounters that span space of interest: Fast Flexible space-filling designs*
- Emphasize interesting regions
  - Identified through baselining efforts
- Different strategies for training and testing sets




*Ryan Lekivetz & Brad Jones (2015), JMP

# Using DoE, Test Set Meets Objectives

- 400 hours of testing data
  - Complex backgrounds, source types
  - Performance across all source types
  - Difficulty varied to distinguish performance
- Will be made available to serve as community standard benchmark dataset



# Where have we been?

**We’ve talked about the application of DOE to subject areas including:**

- **Agriculture**
- **Global warming**
- **New materials development**
- **On-line marketing**
- **AI modeling**
- **Development of data sets for AI algorithm test and evaluation with application to nuclear nonproliferation.**

# Growth in DOE appears to be exponential

- Conclusion here is anecdotal, no data, no metrics
- DOE innovations and investments in the methodology in the marketplace suggest rapid growth
- As science and technology advance, so will the need for well-designed experiments!

We want to thank JMP and Anne Milley for the opportunity to participate in this series, and we look forward to the followup questions and discussion