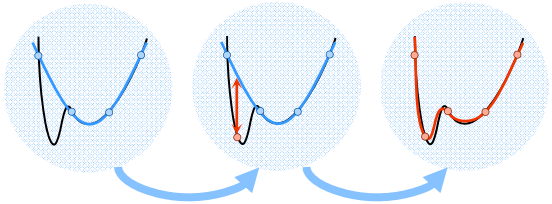
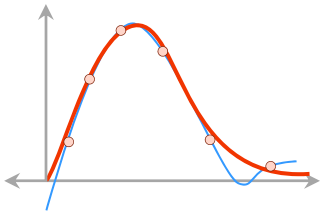




The ALAMO software for model building, constrained regression, and intelligent experimental design

Nick Sahinidis

Acknowledgments:

Alison Cozad, David Miller, Zach Wilson

CAPD CENTER

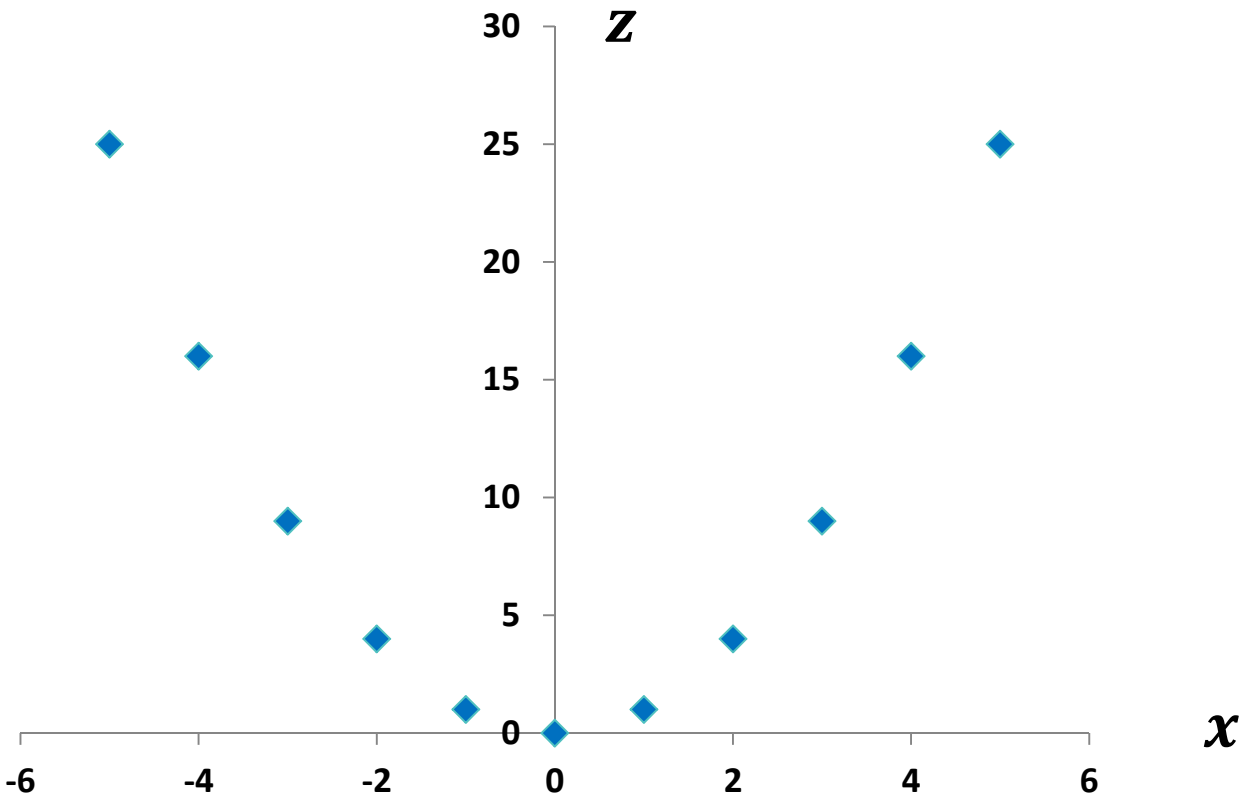
Carnegie Mellon University

NETL NATIONAL ENERGY TECHNOLOGY LABORATORY

CCSI™

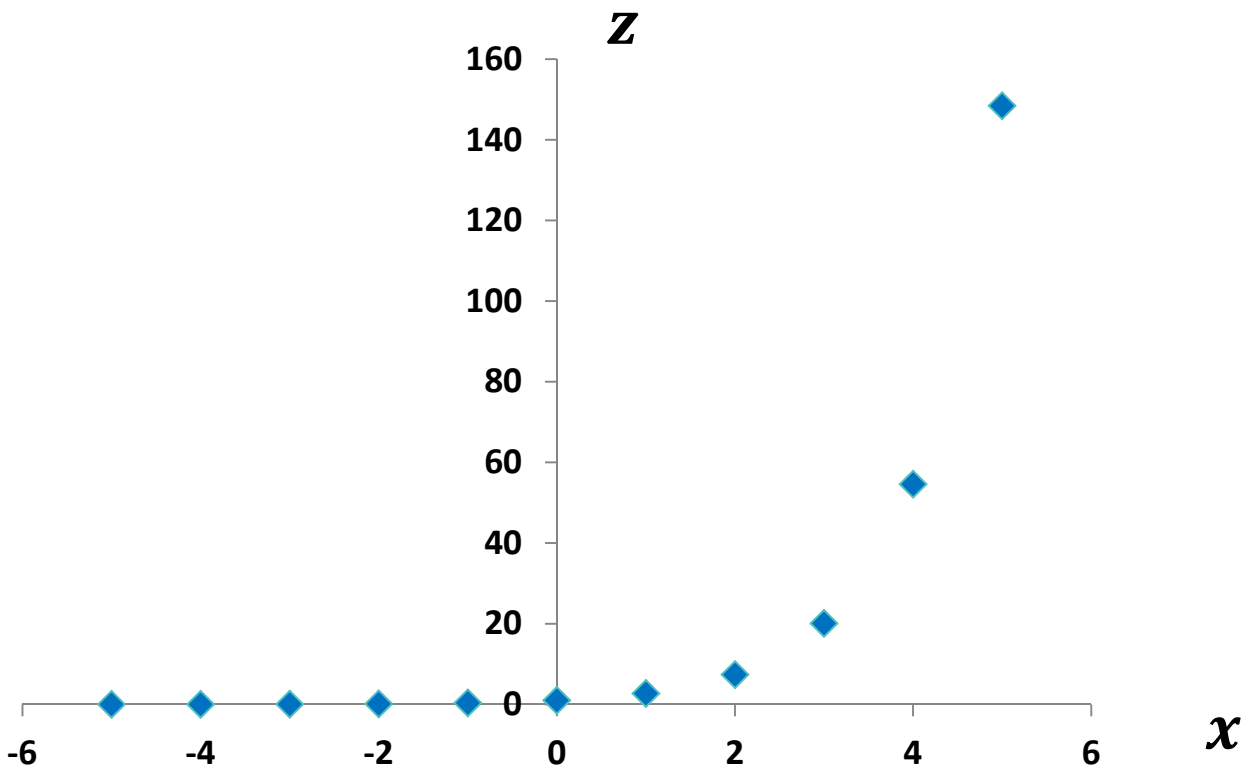
The Optimization Firm

FITTING MODELS TO DATA



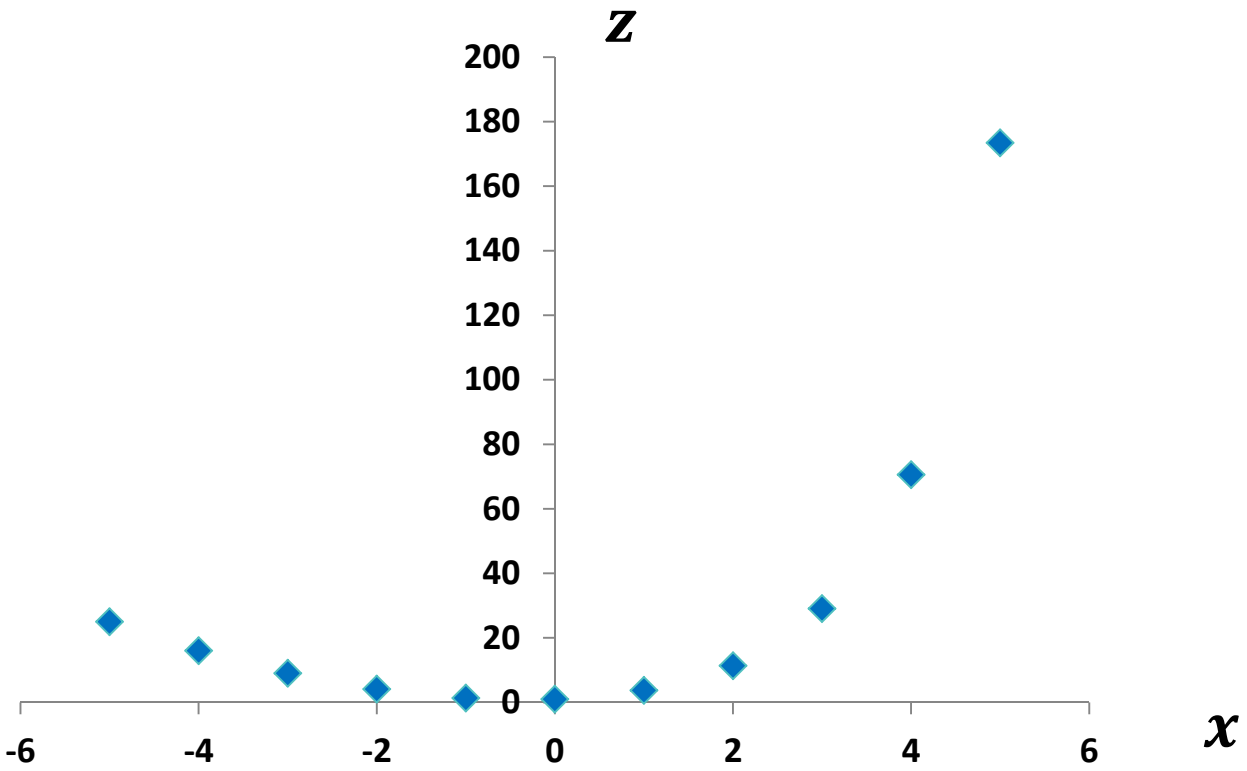
$$z = x^2$$

FITTING MODELS TO DATA



$$z = \exp(x)$$

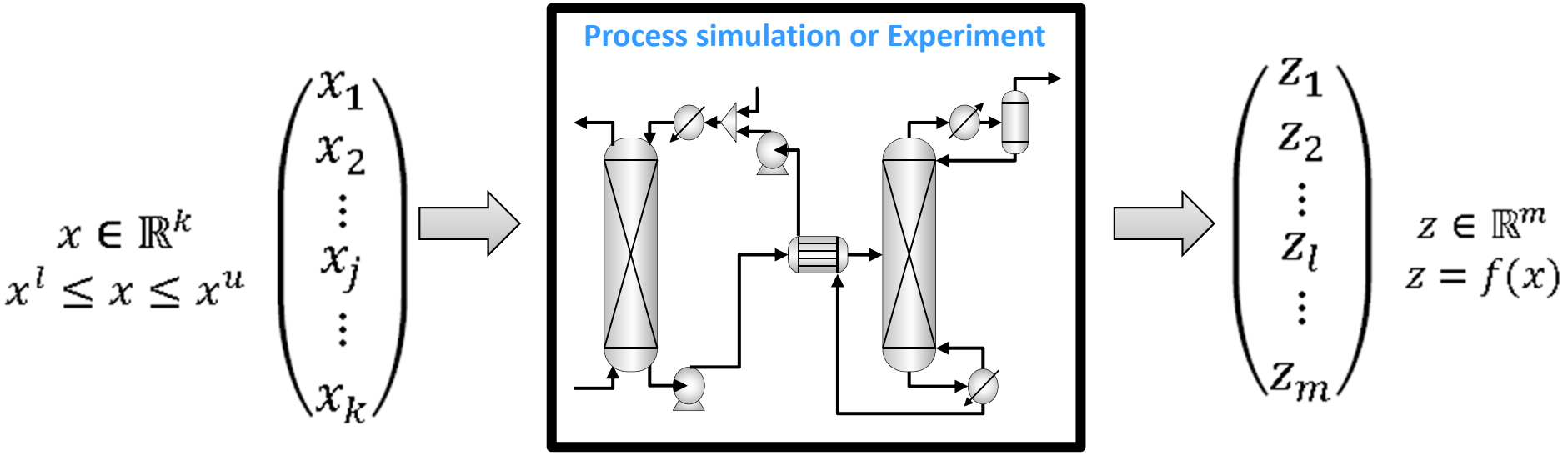
FITTING MODELS TO DATA



$$z = x^2 + \exp(x)$$

ALAMO SOFTWARE

Build a model of output variables z as a function of input variables x over a specified interval



Independent variables:
Operating conditions, inlet flow properties, unit geometry

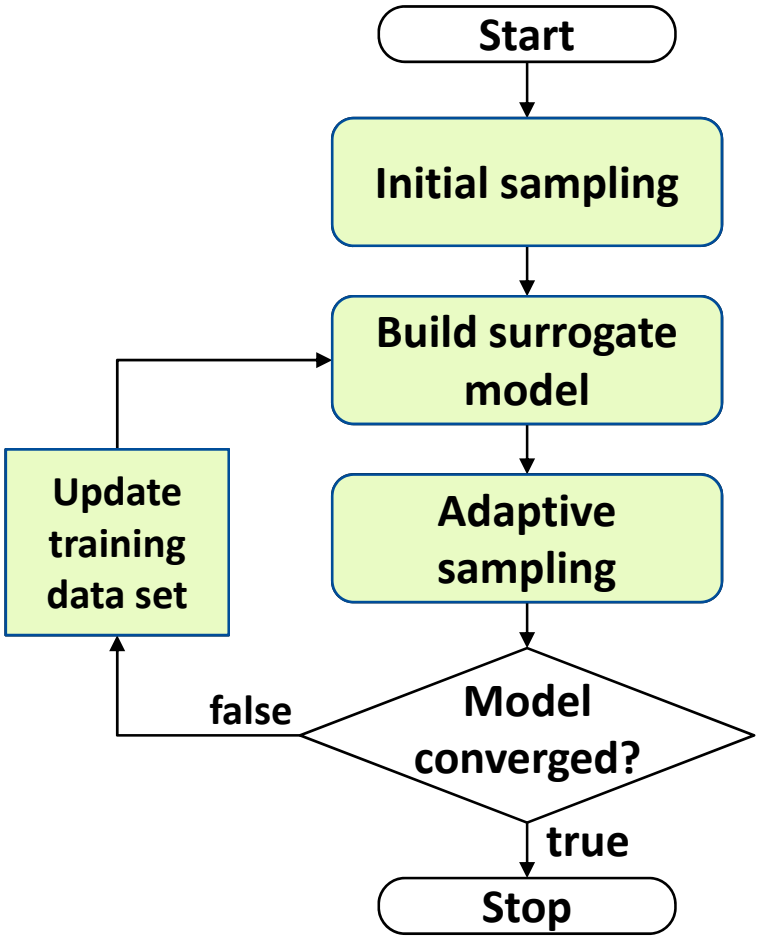
Dependent variables:
Efficiency, outlet flow conditions, conversions, heat flow, etc.

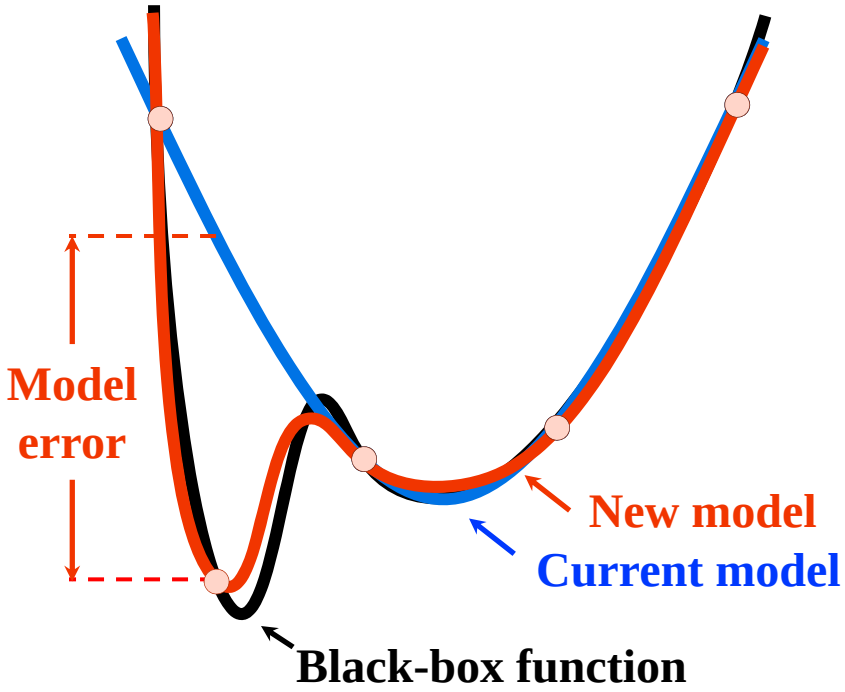
DESIRED MODEL ATTRIBUTES

- **ALAMO aims to build models that are**
  - **Accurate**
    - ***We want to reflect the true nature of the system***
  - **Simple**
    - ***Interpretable***
    - ***Tailored for algebraic optimization***
  - **Generated from a minimal data set**
    - ***Reduce experimental and simulation requirements***

ALAMO

Automated Learning of Algebraic Models using Optimization





MODEL IDENTIFICATION

- **Identify the functional form and complexity of the surrogate models** $z = f(x)$
- **Seek models that are combinations of basis functions**
  1. **Simple basis functions**

Category		$X_j(x)$
I.	Polynomial	$(x_d)^{\alpha}$
II.	Multinomial	$\prod_{d \in \mathcal{D}' \subseteq \mathcal{D}} (x_d)^{\alpha_d}$
III.	Exponential and logarithmic	$\exp\left(\frac{x_d}{\gamma}\right)^{\alpha}, \log\left(\frac{x_d}{\gamma}\right)^{\alpha}$

  2. **Radial basis functions for parametric regression**
  3. **User-specified basis functions for tailored regression**

OVERFITTING AND TRUE ERROR

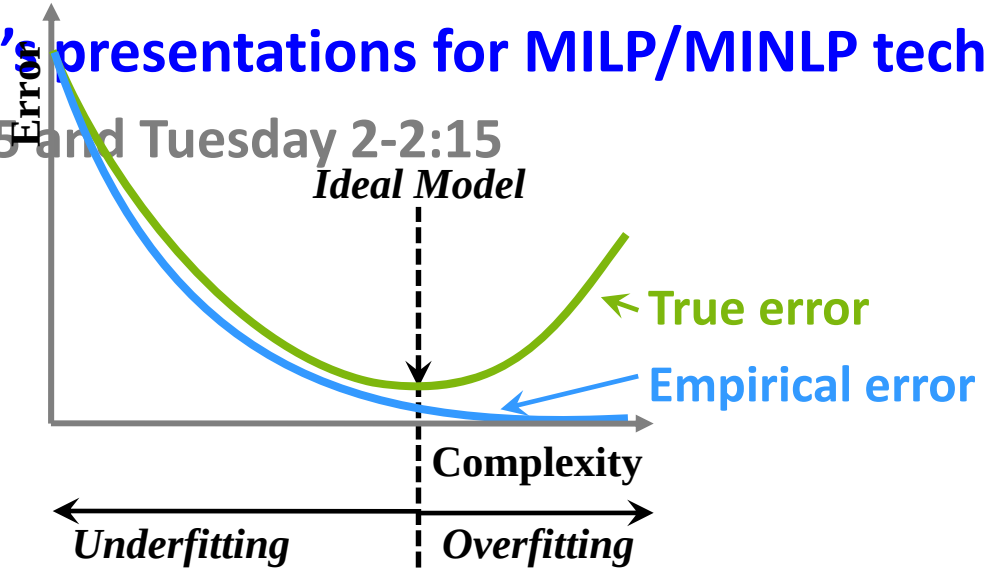
- **Step 1:** Define a large set of potential basis functions

$$\hat{z}(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 e^{x_1} + \beta_5 e^{x_2} + \dots$$

- **Step 2:** Model reduction

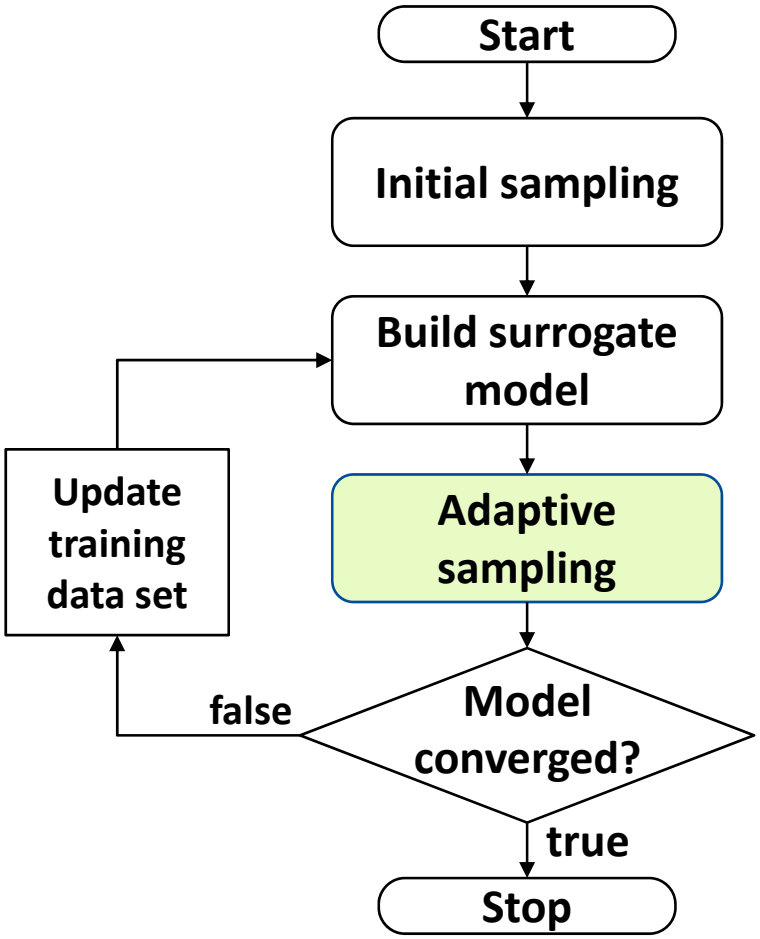
$$\hat{z}(x) = 2 + x_2 + 5\, e^{x_1}$$

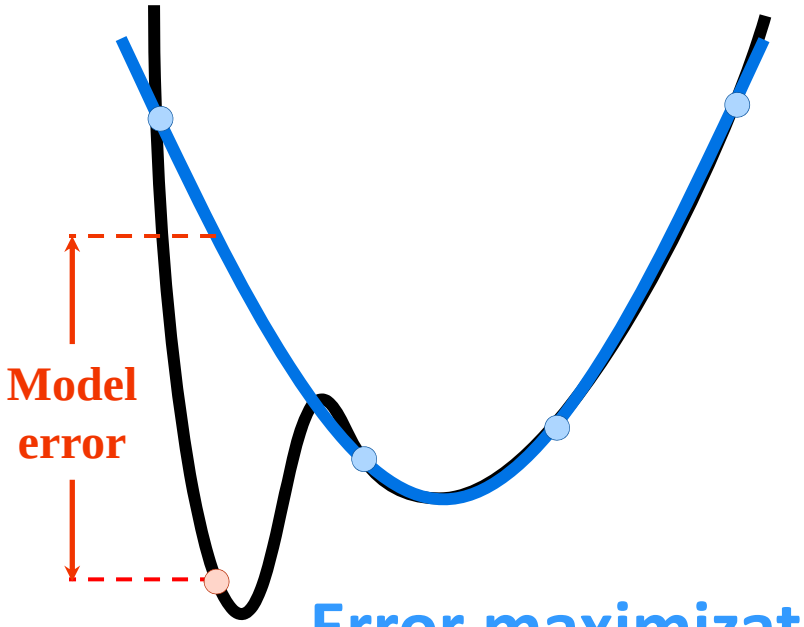
- **See Zach Wilson's presentations for MILP/MINLP techniques**
  - Monday 5-5:25 and Tuesday 2-2:15



ALAMO

Automated Learning of Algebraic Models using Optimization





Error maximization sampling

ERROR MAXIMIZATION SAMPLING

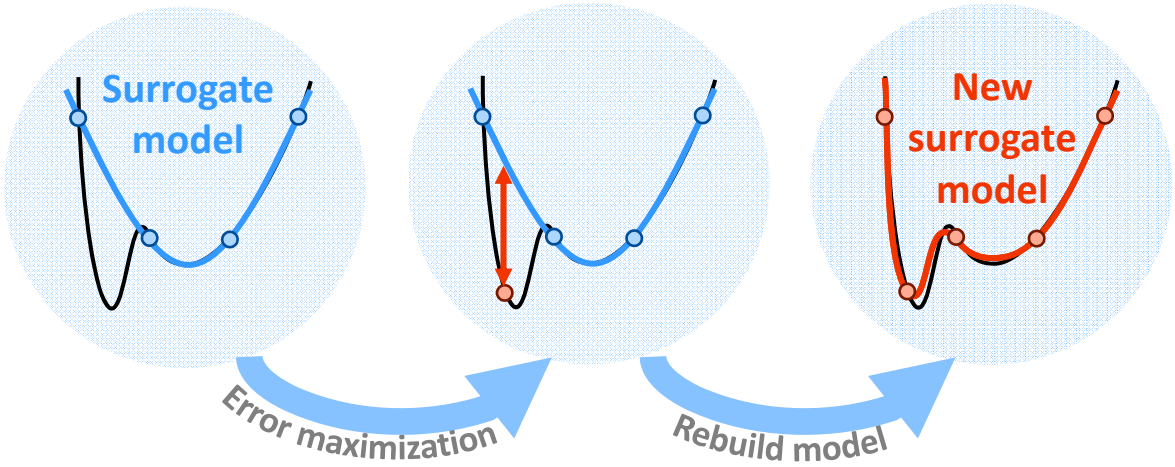
- **Search the problem space for areas of model inconsistency or model mismatch**
- **Find points that maximize the model error with respect to the independent variables**

Surrogate model

$$\max_x \left(\frac{z(x) - \hat{z}(x)}{z(x)} \right)^2$$

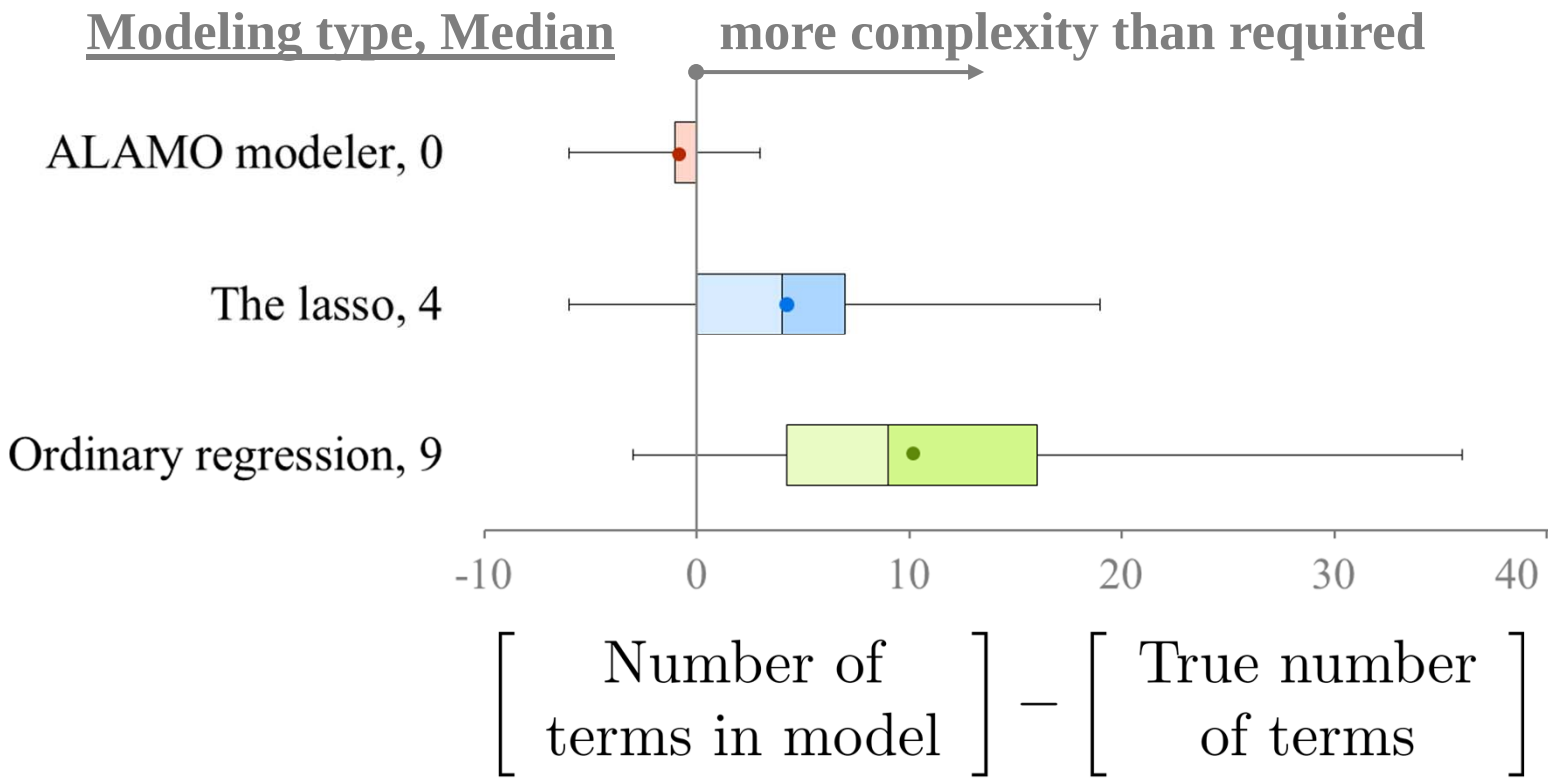
  - **Optimized using derivative-free solver SNOBFIT (Huyer and Neumaier, 2008)**
  - **SNOBFIT outperforms most derivative-free solvers (Rios and Sahinidis, 2013)**

KEY INGREDIENT: OPTIMIZATION



- **Surrogate model identification**
  - Simple, accurate model identification
  - Integer optimization
- **Error maximization sampling**
  - More information found per simulated data point
  - Derivative-free optimization

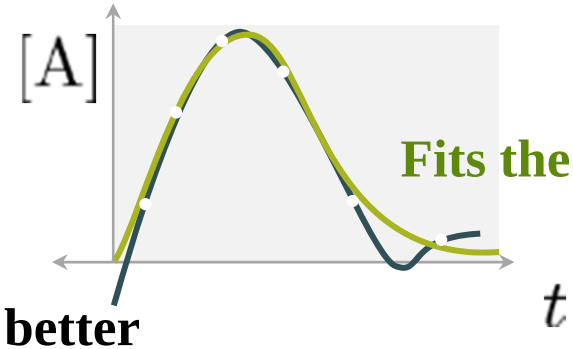
ALAMO PROVIDES SIMPLE MODELS



Results over a test set of 45 known functions treated as black boxes with bases that are available to all modeling methods

CONSTRAINED REGRESSION

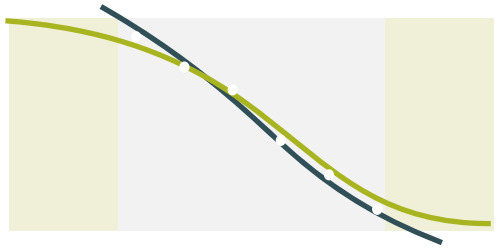
$$0 \leq [\mathrm{A}]_t \leq [\mathrm{A}]^{\max}$$



Fits the true function better

Fits the data better

Extrapolation zone



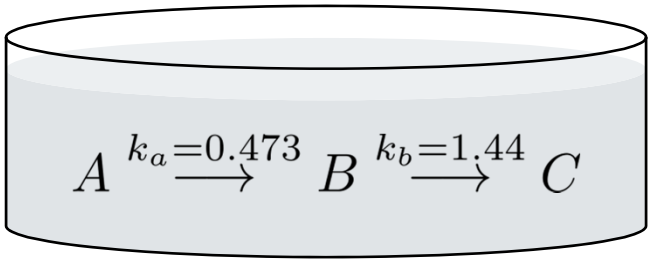
Data space

Safe extrapolation

BATCH REACTOR

Batch reactor problem:

First-order reactions in series



Find a model for the concentration of B such that

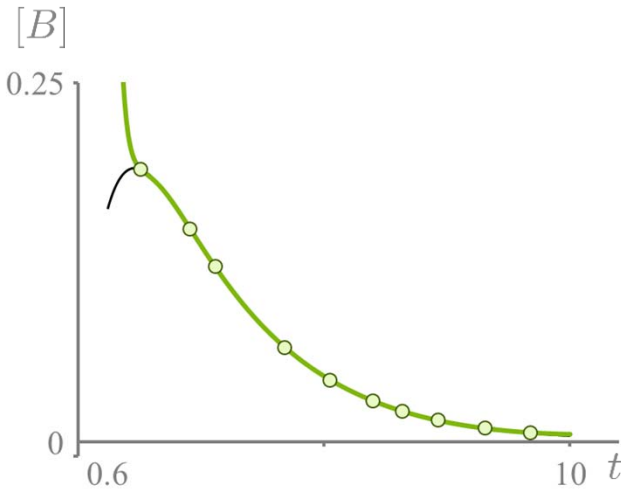
$$0 \leq [\hat{B}](t) \leq [A]_0$$

where $[\mathrm{A}]_0 = 1$, $[\mathrm{B}]_0 = 0$, and $[\mathrm{C}]_0 = 0$.

Problem selected from constrained regression test library

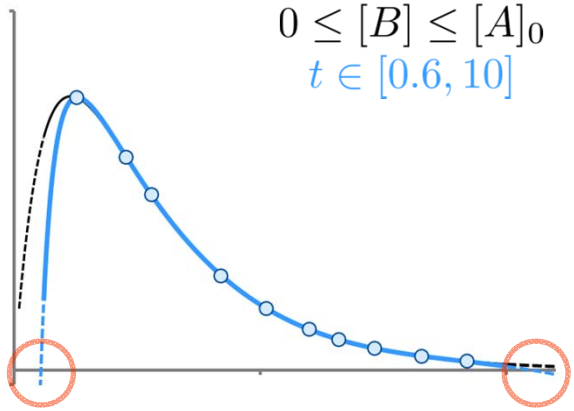
BATCH REACTOR MODELS

Unconstrained



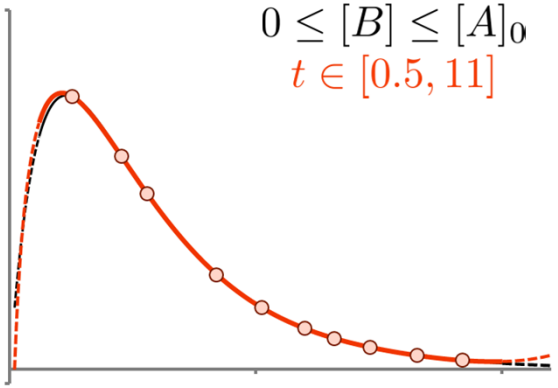
$$[\hat{B}](t) = 0.34 \log t + 2.3/\sqrt{t} \\ -0.91/t^2 + 0.32/t^4 - 1.5$$

Constrained

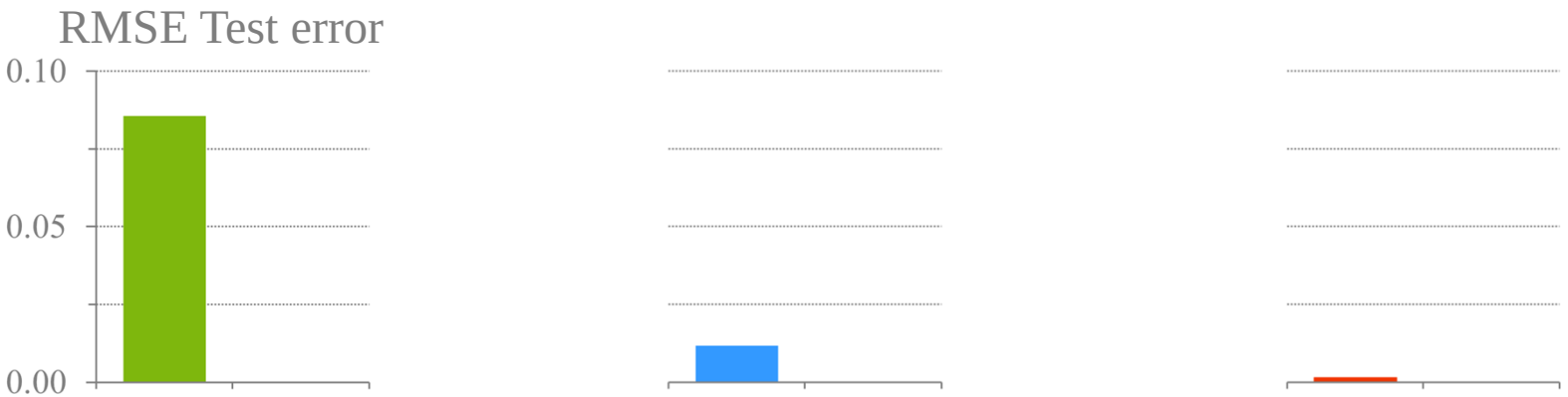


$$[\hat{B}](t) = -0.18 \log t + 0.89/\sqrt{t} \\ -0.71/t + 0.0040\, t^2 - 0.00020\, t^3$$

Safe extrapolation



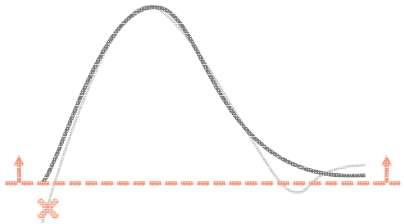
$$[\hat{B}](t) = 3.3 \cdot 10^{-6} \exp t + 0.46\sqrt{t} \\ -0.28\, t + 0.018\, t^2 - 5.2 \cdot 10^{-5}\, t^4$$



TYPES OF RESTRICTIONS

Response bounds

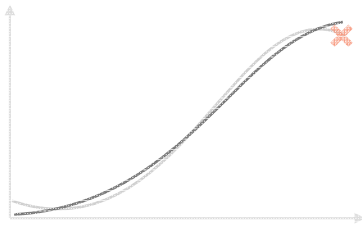
$$[\hat{A}]_t \geq 0$$



pressure, temperature, compositions

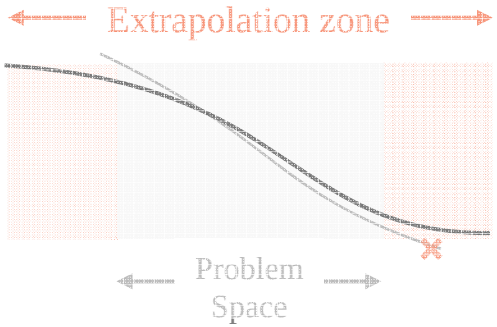
Response derivatives

$$\frac{dT}{dx} \geq 0$$



monotonicity, numerical properties, convexity

Alternative domains



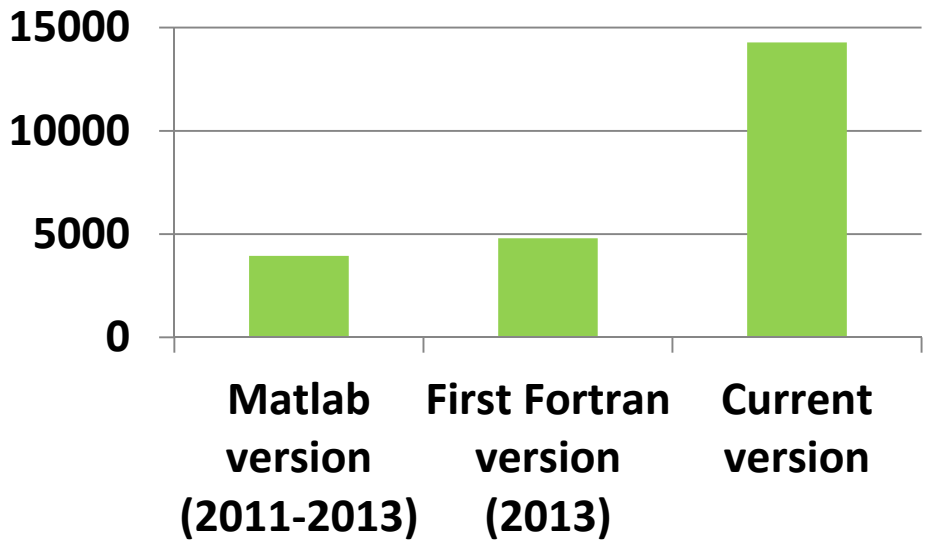
safe extrapolation, boundary conditions

Constrained regression relies on global optimization technology (BARON)

IMPLEMENTATION

$$\mu_k = \frac{\sum_{j \epsilon S} z_{jk}}{|S|}$$

mu = sum (z(:, k), S(:)) / count(S(:))



CONCLUSIONS

- **ALAMO provides algebraic models that are**
  - ✓ Accurate
  - ✓ Simple
  - ✓ Generated from a minimal number of data points
- **ALAMO's constrained regression facility allows modeling of**
  - ✓ Bounds on response variables
  - ✓ Convexity/monotonicity of response variables
- **Built on top of state-of-the-art optimization solvers**
- **ALAMO site: archimedes.cheme.cmu.edu/?q=alamo**