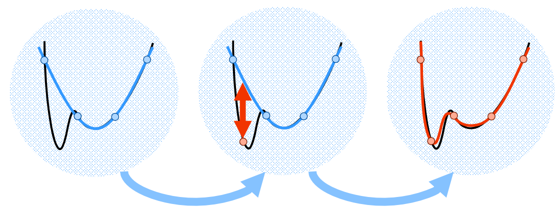
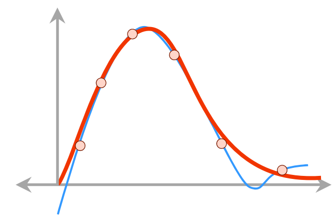




The ALAMO approach to machine learning: Best subset selection, adaptive sampling, and constrained regression

Nick Sahinidis

Acknowledgments:

Alison Cozad, David Miller, Zach Wilson

MACHINE LEARNING PROBLEM

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

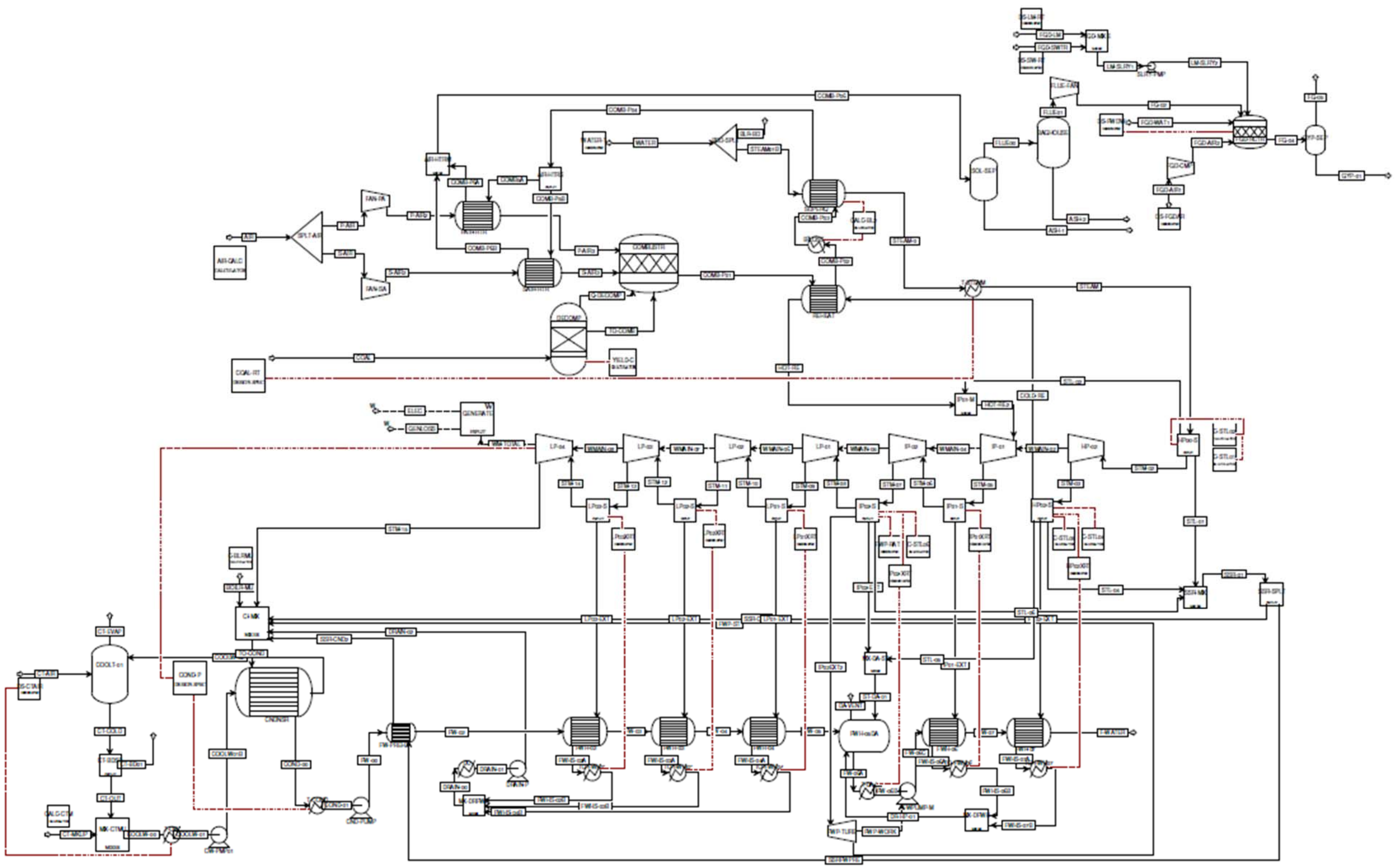
$$x \in \mathbb{R}^k \qquad x^l \leq x \leq x^u \qquad \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_j \\ \vdots \\ x_k \end{pmatrix}$$

$$\begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_l \\ \vdots \\ z_m \end{pmatrix} \qquad z \in \mathbb{R}^m \qquad z = f(x)$$

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

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

SIMULATION OPTIMIZATION

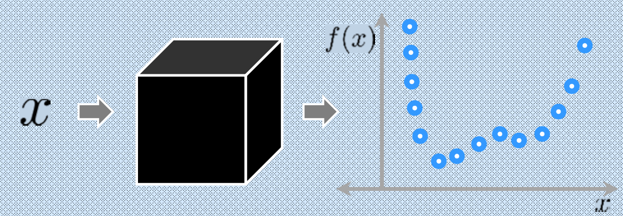


Pulverized coal plant Aspen Plus® simulation provided by the National Energy Technology Laboratory

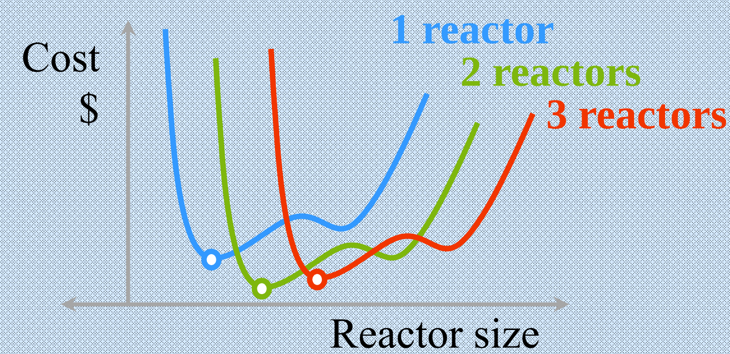
CHALLENGES

OPTIMIZER

No algebraic model

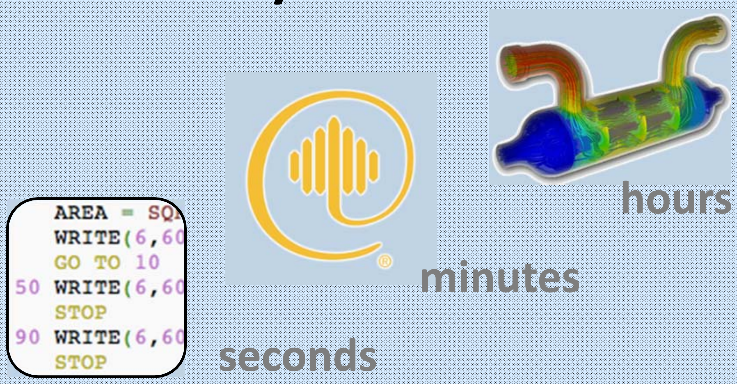


Complex process alternatives



SIMULATOR

Costly simulations

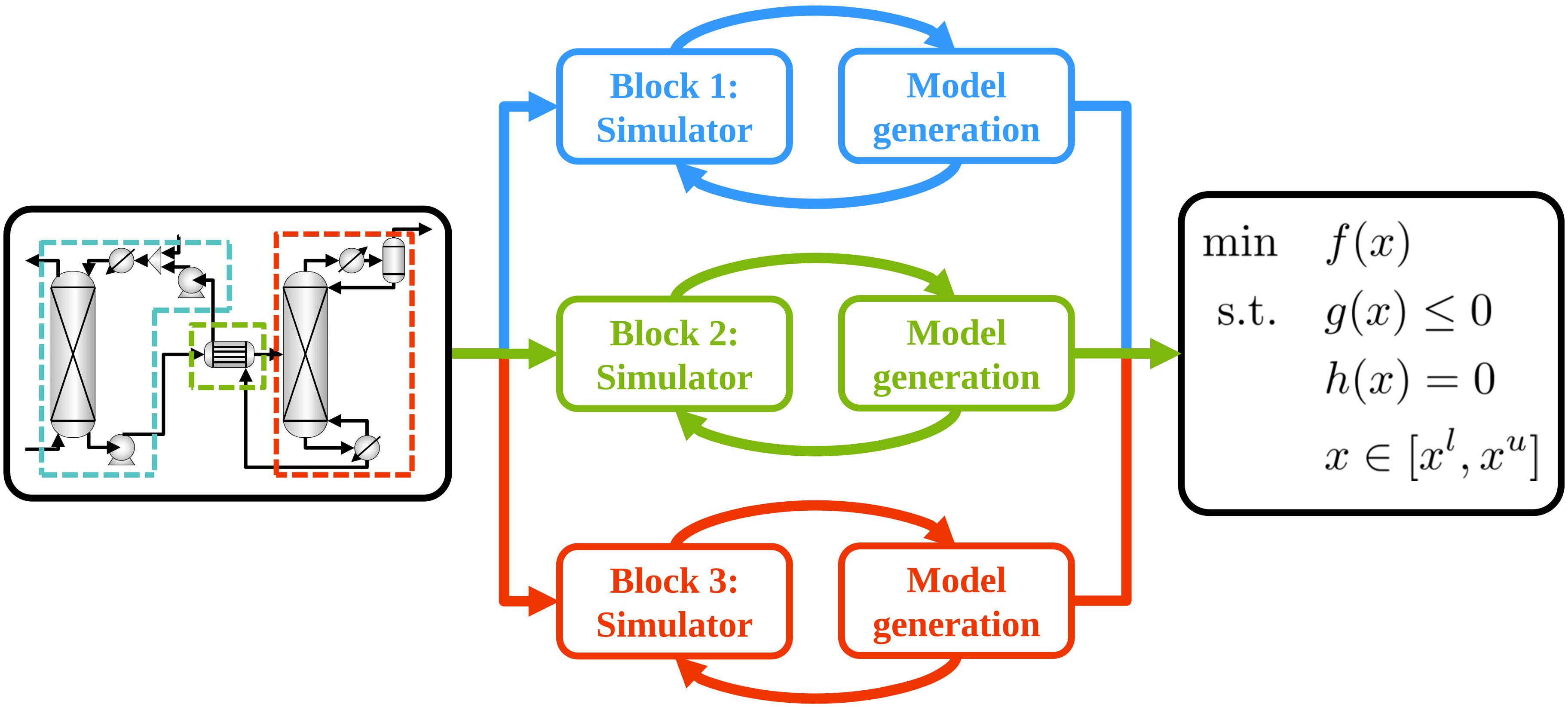


Scarcity of fully robust simulations

✗ **Gradient-based methods**

✗ **Derivative-free methods**

PROCESS DISAGGREGATION



Process Simulation

Disaggregate process into process blocks

Surrogate Models

Build simple and accurate models with a functional form tailored for an optimization framework

Optimization Model

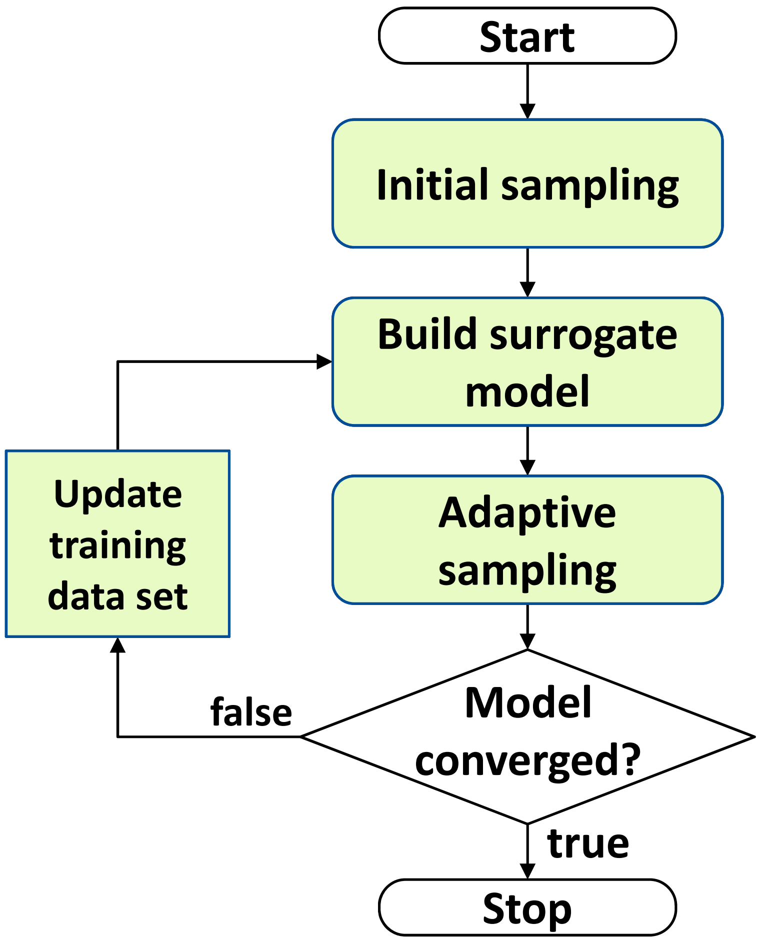
Add algebraic constraints design specs, heat/mass balances, and logic constraints

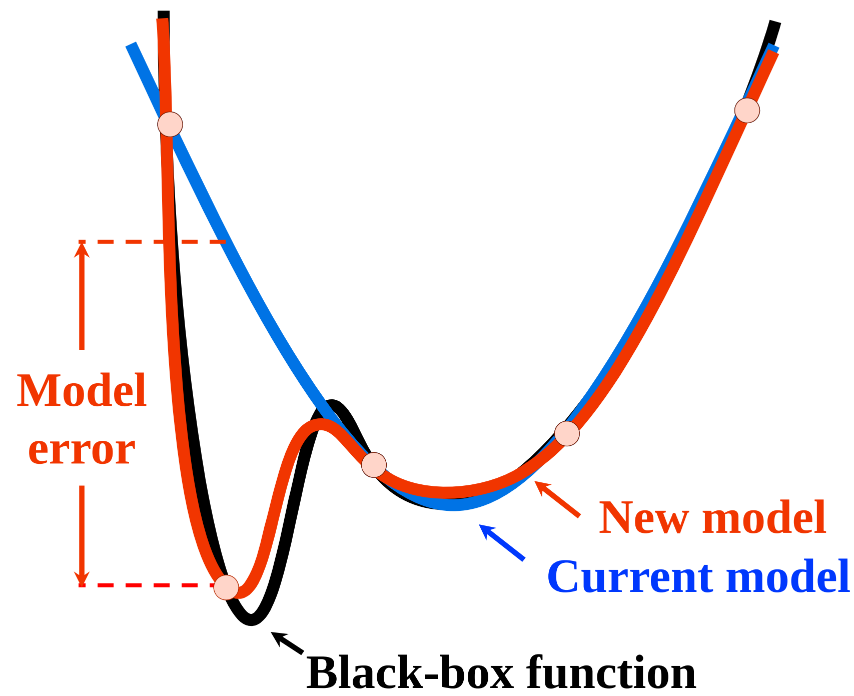
DESIRED MODEL ATTRIBUTES

1. **Accurate**
   - **We want to reflect the true nature of the system**
2. **Simple**
   - **Usable for algebraic optimization**
   - **Interpretable**
3. **Generated from a minimal data set**
   - **Reduce experimental and simulation requirements**
4. **Obeys physics and user insights**
   - **Increase fidelity and validity in regions with no measurements**

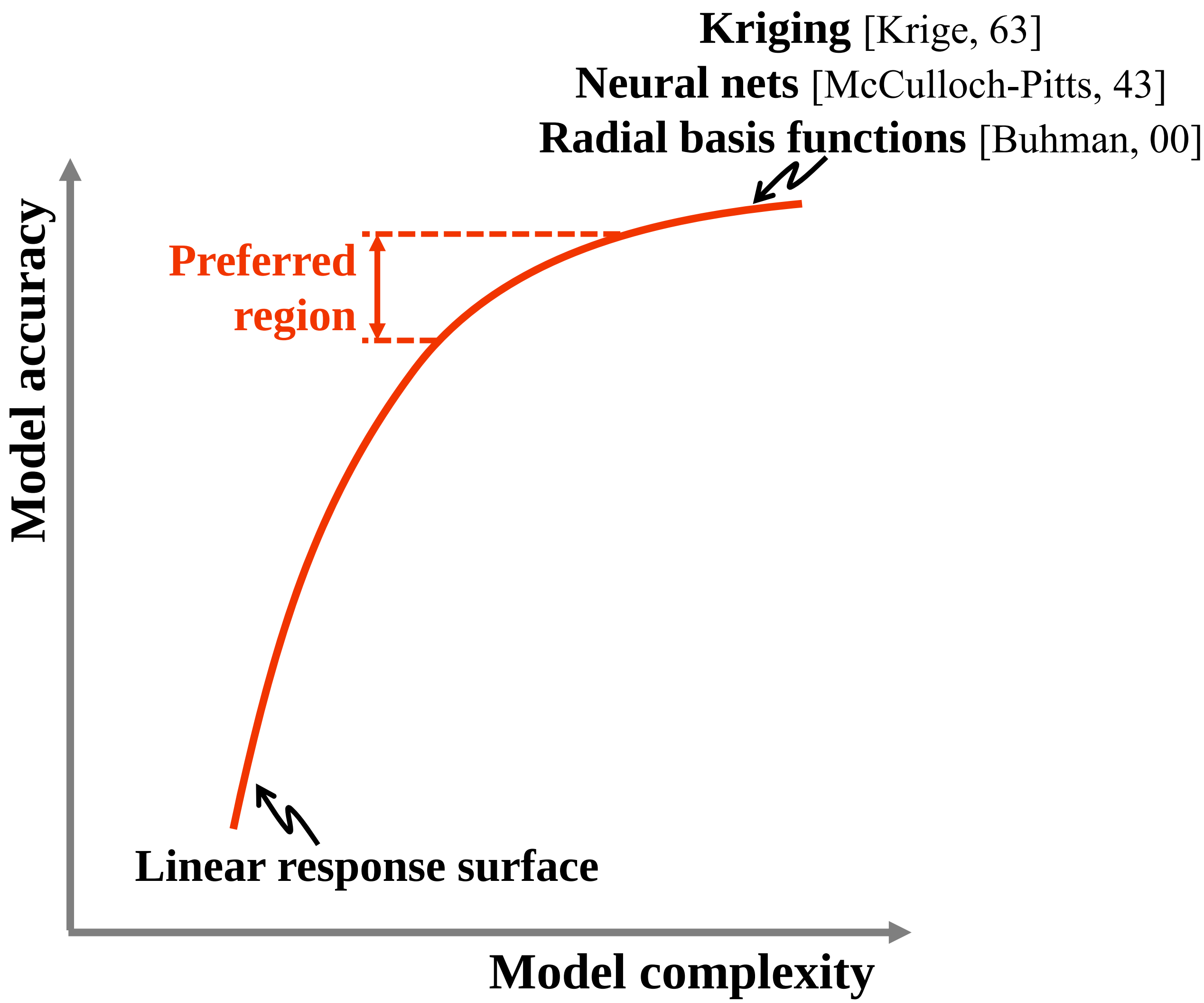
ALAMO

Automated Learning of Algebraic MOdels





MODEL COMPLEXITY TRADEOFF



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

OVERFITTING AND TRUE ERROR

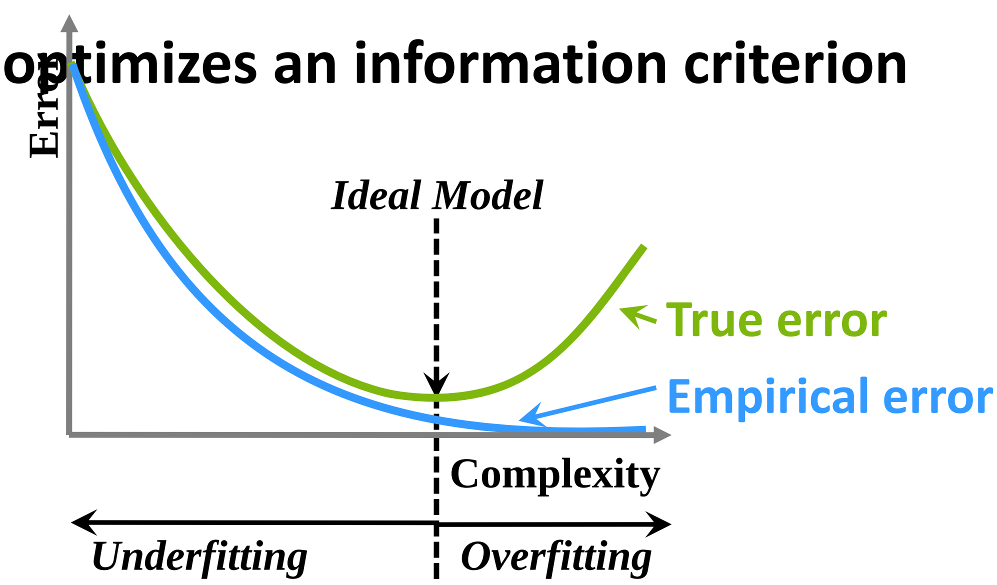
- **Step 1: Define a large set of potential basis functions**

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

- **Step 2: Model reduction**

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

Select subset that optimizes an information criterion



MODEL REDUCTION TECHNIQUES

- **Qualitative tradeoffs of model reduction methods**

CPU modeling cost

Efficacy

Best subset methods
- *Enumerate all possible subsets*

Regularized regression techniques
- *Penalize the least squares objective using the magnitude of the regressors*

Stepwise regression [Efroymson, 60]

Backward elimination [Oosterhof, 63]

Forward selection [Hamaker, 62]

MODEL SELECTION CRITERIA

- **Balance fit (sum of square errors) with model complexity (number of terms in the model; denoted by *p*)**

Corrected Akaike information criterion

$$AIC_c = N\log\left(\frac{1}{N}\sum_{i=1}^{N}(z_i - X_i\beta)^2\right) + 2\boldsymbol{p} + \frac{2\boldsymbol{p}(\boldsymbol{p}+1)}{N-\boldsymbol{p}-1}$$

Mallows' Cp

$$C_p = \frac{\sum_{i=1}^{N}(z_i - X_i\beta)^2}{\widehat{\sigma^2}} + 2\boldsymbol{p} - N$$

Hannan-Quinn information criterion

$$HQC = N\log\left(\frac{1}{N}\sum_{i=1}^{N}(z_i - X_i\beta)^2\right) + 2\boldsymbol{p}\log(\log(N))$$

Bayes information criterion

$$BIC = \frac{\sum_{i=1}^{N}(z_i - X_i\beta)^2}{\widehat{\sigma^2}} + \boldsymbol{p}\log(N)$$

Mean squared error

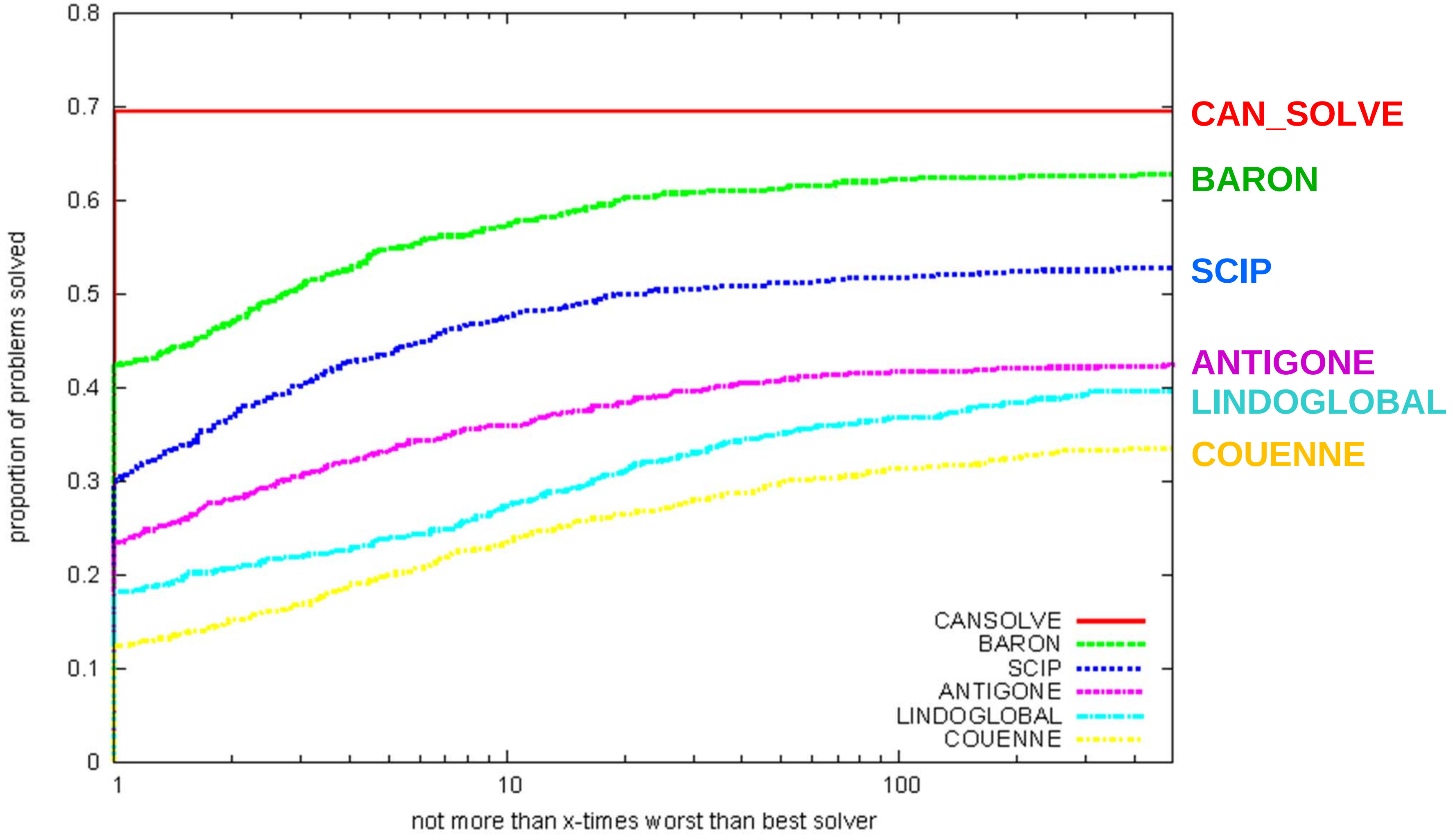
$$MSE = \frac{\sum_{i=1}^{N}(z_i - X_i\beta)^2}{N-\boldsymbol{p}-1}$$

- **Mixed-integer nonlinear programming models**

BRANCH-AND-REDUCE

- **Constraint propagation and duality-based bounds tightening**
  - Ryoo and Sahinidis, 1995, 1996
  - Tawarmalani and Sahinidis, 2004
- **Finite branching rules**
  - Shectman and Sahinidis, 1998
  - Ahmed, Tawarmalani and Sahinidis, 2004
- **Convexification**
  - Tawarmalani and Sahinidis, 2001, 2002, 2004, 2005
  - Khajavirad and Sahinidis, 2012, 2013, 2014, 2016
  - Zorn and Sahinidis, 2013, 2013, 2014
- **Implemented in BARON**
  - First deterministic global optimization solver for NLP and MINLP

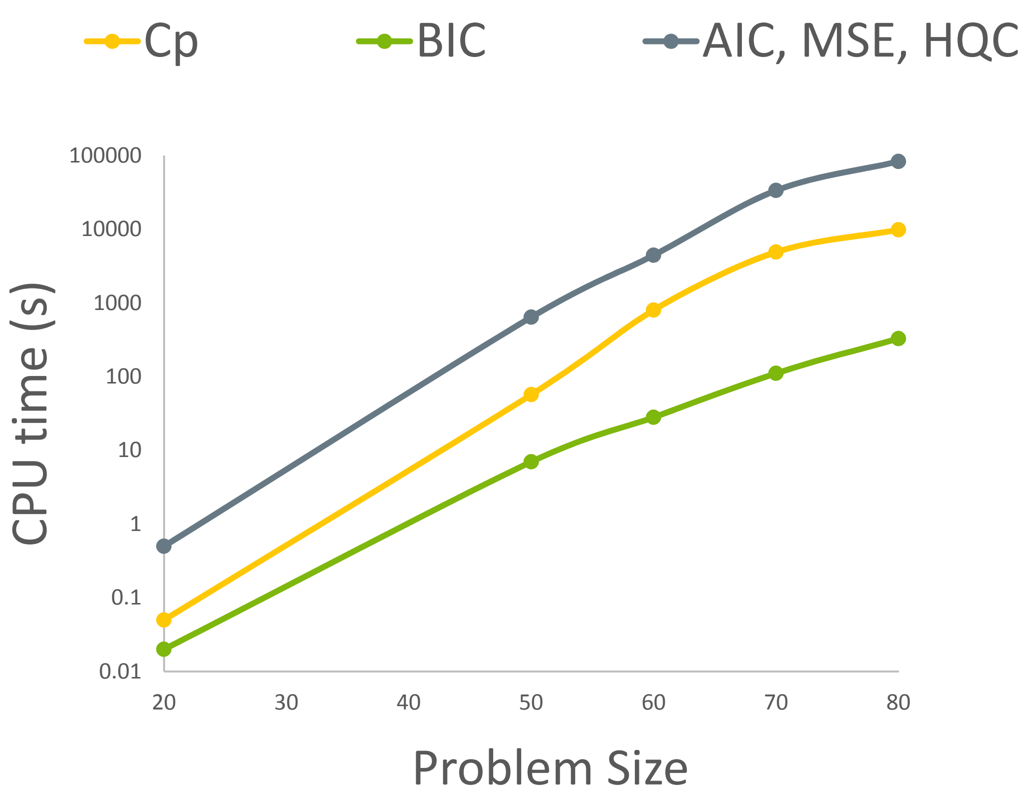
GLOBAL MINLP SOLVERS ON MINLPLIB2



Con: 1893 (1—164,321), Var: 1027 (3—107,223), Disc: 137 (1—31,824)

CPU TIME COMPARISON OF METRICS

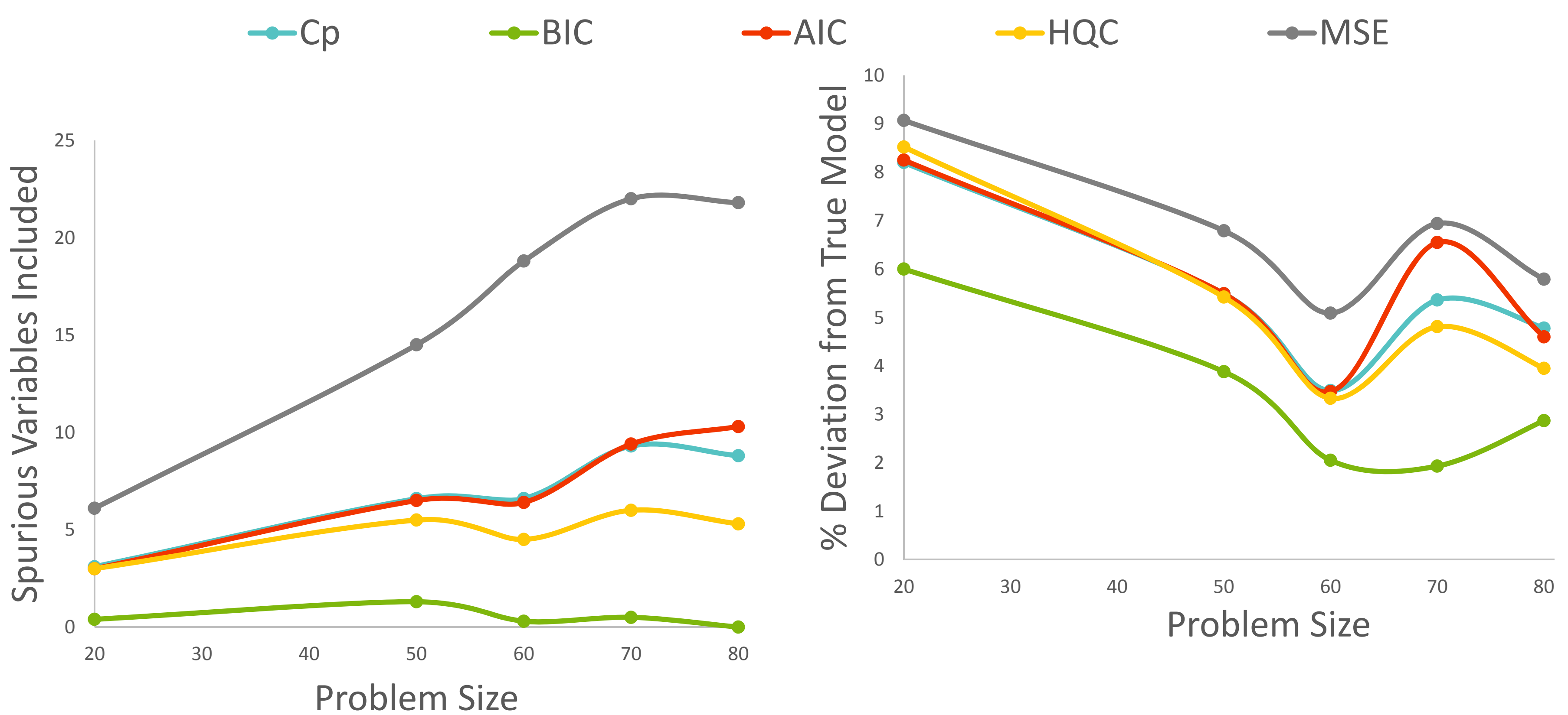
- Eight benchmarks from the UCI and CMU data sets
- Seventy noisy data sets were generated with multicolinearity and increasing problem size (number of bases)



BIC solves more than two orders of magnitude faster than AIC, MSE and HQC

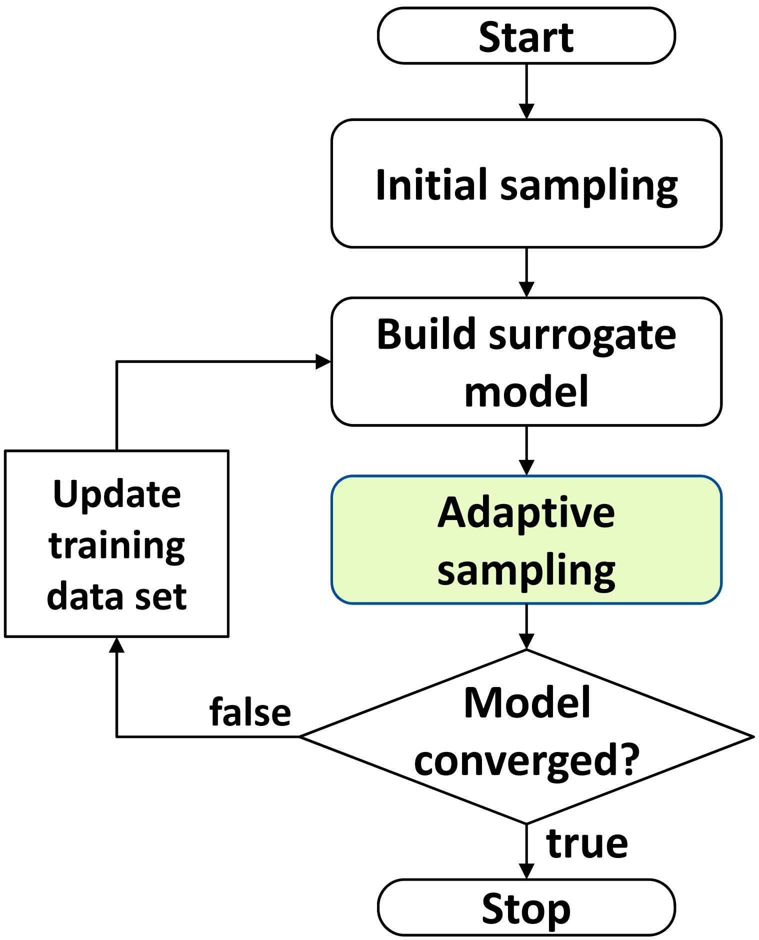
MODEL QUALITY COMPARISON

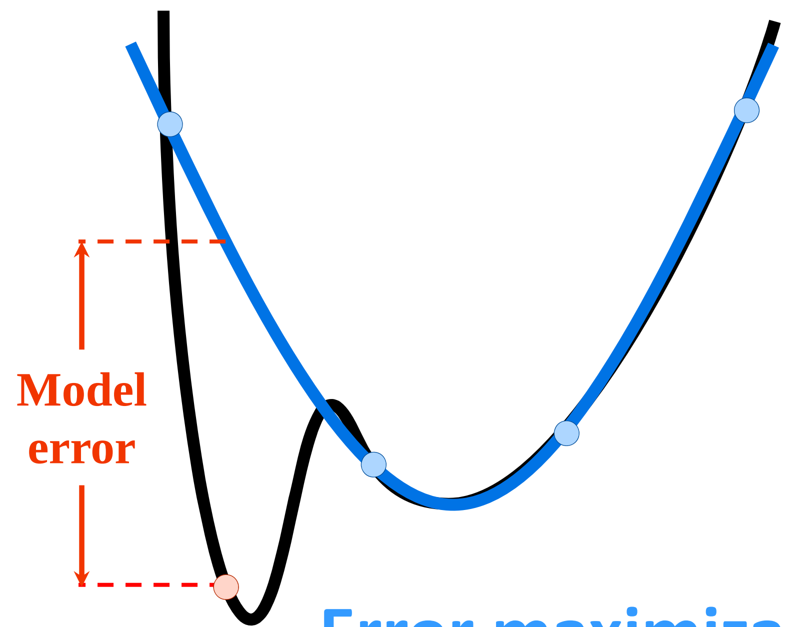
- **BIC leads to smaller, more accurate models**
  - Larger penalty for model complexity



ALAMO

Automated Learning of Algebraic MOdels





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**

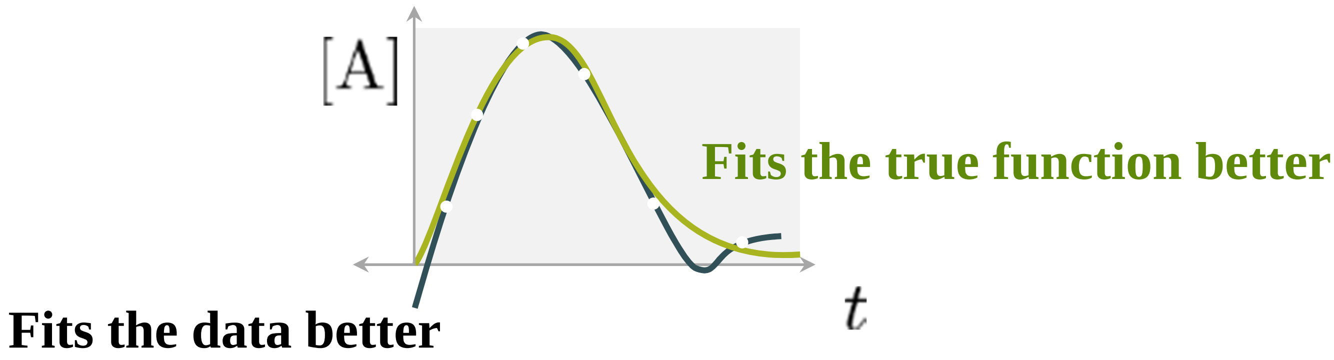
Surrogate model

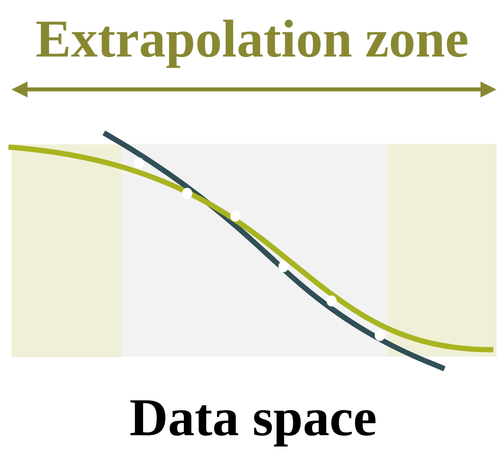
$$\max_x \left(\frac{z(x) - \hat{z}(x)}{z(x)} \right)^2$$

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

CONSTRAINED REGRESSION

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





Safe extrapolation

CONSTRAINED REGRESSION

Standard regression

$$\min_{\beta_1,\beta_2} \sum_{i=1}^{4} (z_i - \hat{z}(x_i;\beta_1,\beta_2))^2$$

Surrogate model

easy

tough

Constrained regression

$$\min_{\beta_1,\beta_2} \sum_{i=1}^{4} (z_i - \hat{z}(x_i;\beta_1,\beta_2))^2$$
$$\text{s.t.} \quad \beta_1 \geq \beta_2$$

$$\min_{\beta_1,\beta_2} \sum_{i=1}^{4} (z_i - \hat{z}(x_i;\beta_1,\beta_2))^2$$
$$\text{s.t.} \quad \hat{z}(x_i;\beta_1,\beta_2) \geq 0 \quad \forall x$$

- **Challenging due to the semi-infinite nature of the regression constraints**
- **Use intuitive restrictions among predictor and response variables to infer nonintuitive relationships between regression parameters**

IMPLIED PARAMETER RESTRICTIONS

Find a model $\hat{z}$ such that $\hat{z}(x) \geq 0$ with a fixed model form:

$$\hat{z}(x) = \beta_1 x + \beta_2 x^3$$

Step 1: Formulate constraint in z- and x-space

$$\min_{\beta_1, \beta_2} \sum_{i=1}^{4} \left(z_i - [\beta_1 x + \beta_2 x^3]\right)^2$$

$$\text{s.t.} \quad \beta_1 x + \beta_2 x^3 \geq 0 \quad x \in [0, 1]$$

Step 2: Identify a sufficient set of β-space constraints

$$\min_{\beta_1, \beta_2} \sum_{i=1}^{4} \left(z_i - [\beta_1 x + \beta_2 x^3]\right)^2$$

$$\text{s.t.} \quad \begin{cases} 0.240\, \beta_1 + 0.0138\, \beta_2 \geq 0 \\ 0.281\, \beta_1 + 0.0223\, \beta_2 \geq 0 \\ 0.120\, \beta_1 + 0.00173\, \beta_2 \geq 0 \\ 0.138\, \beta_1 + 0.00263\, \beta_2 \geq 0 \end{cases}$$

1 parametric constraint

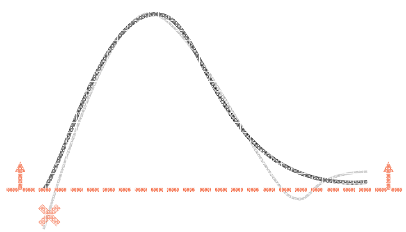
4 β-constraints

Global optimization problems solved with BARON

TYPES OF RESTRICTIONS

Response bounds

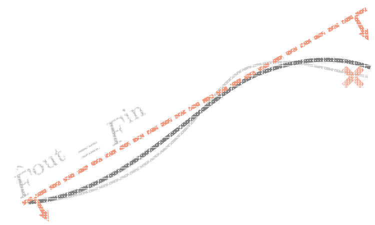
$$[\hat{A}]_t \geq 0$$



pressure, temperature, compositions

Individual responses

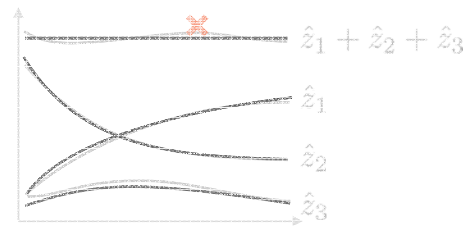
$$\hat{F}^{\text{out}}(x) \leq F^{\text{in}}$$



mass and energy balances, physical limitations

Multiple responses

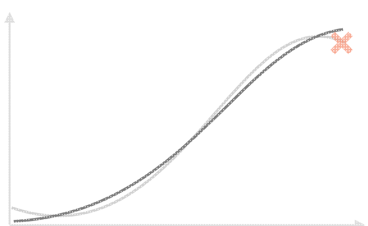
$$\hat{z}_1 + \hat{z}_2 + \hat{z}_3 = 1$$



mass balances, sum-to-one, state variables

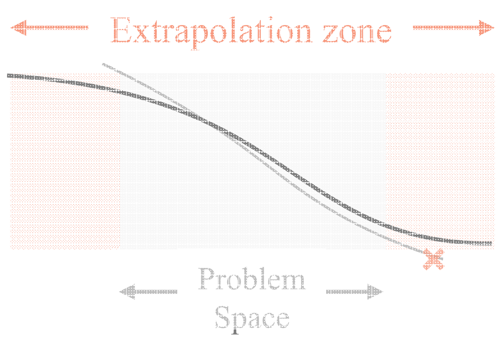
Response derivatives

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



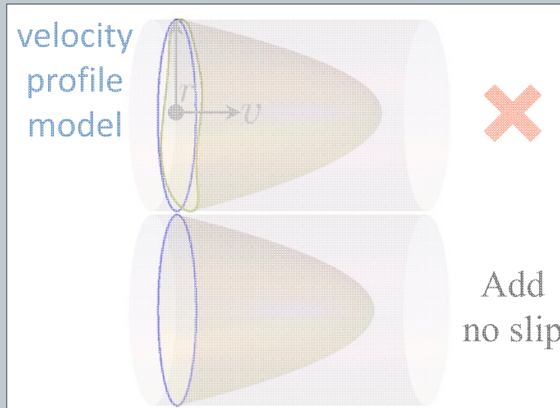
monotonicity, numerical properties, convexity

Alternative domains



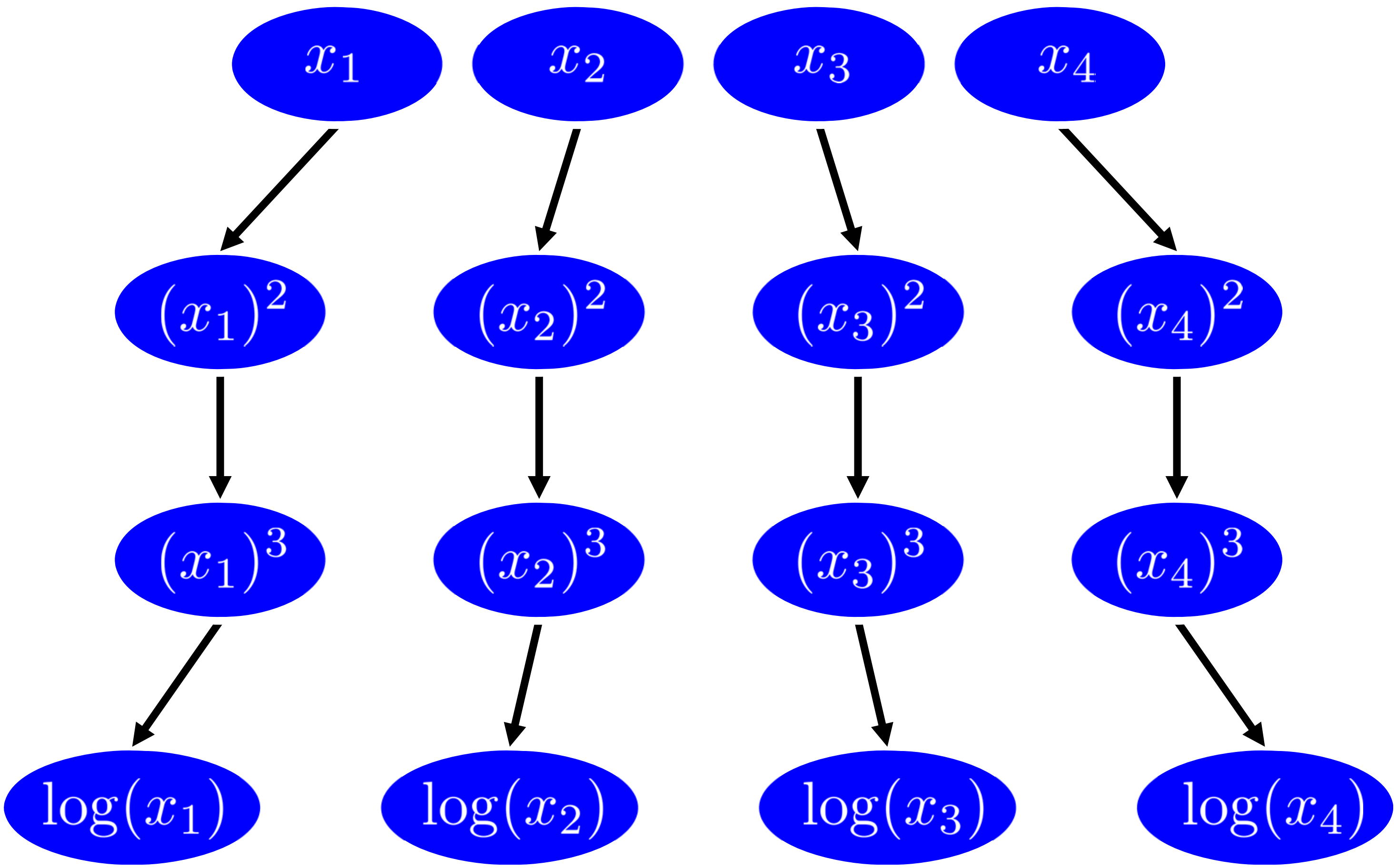
safe extrapolation, boundary conditions

Boundary conditions



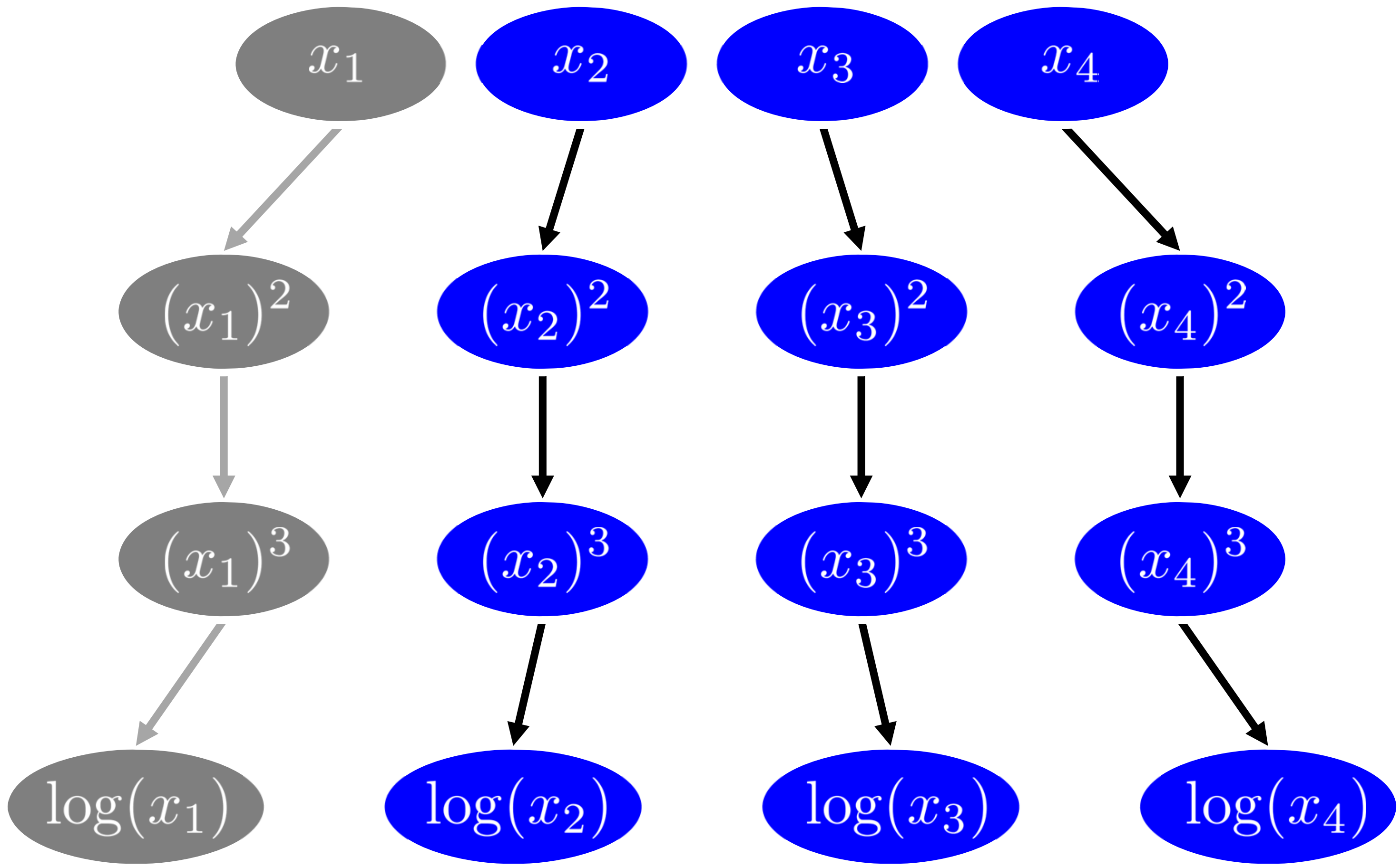
$$\hat{v}(R, \theta) = 0 \qquad \forall \theta$$

CATEGORICAL CORRELATED



CATEGORICAL CORRELATED

x_1 is shown not to be a significant predictor



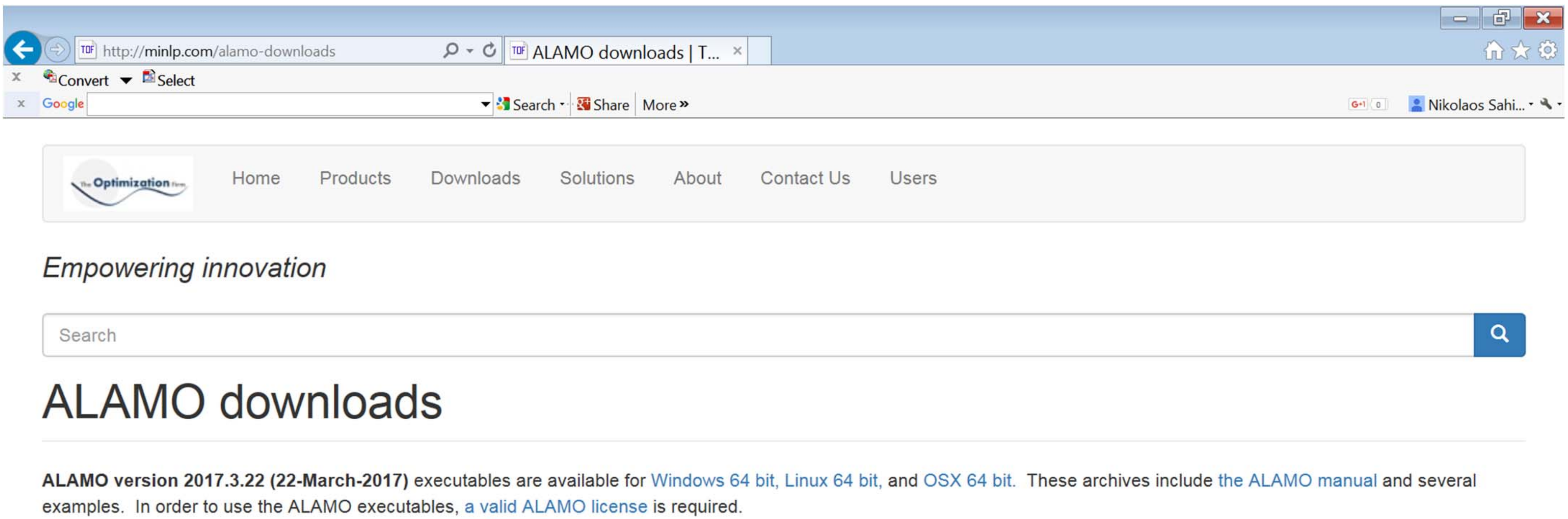
GROUP CONSTRAINTS

- **A collection of basis functions or groups of basis functions is referred to as a *group***
- **A group is selected iff any of its components is selected**
- **Group constraints**
  - **NMT:** no more than a specified number of components is selected from a group
  - **ATL:** at least a specified number of components is selected from a group
  - **REQ:** if a group is selected, another group must also be selected
  - **XCL:** if a group is selected, another group must not be selected

ALAMO
TUTORIAL AND APPLICATIONS

DOWNLOAD

- From The Optimization Firm at http://minlp.com
- ZIP archives of Windows, Linux, and OSX versions
- Free licenses for academics and CAPD companies



WORKING WITH PRE-EXISTING DATA SETS

EXAMPLE ALAMO INPUT FILE

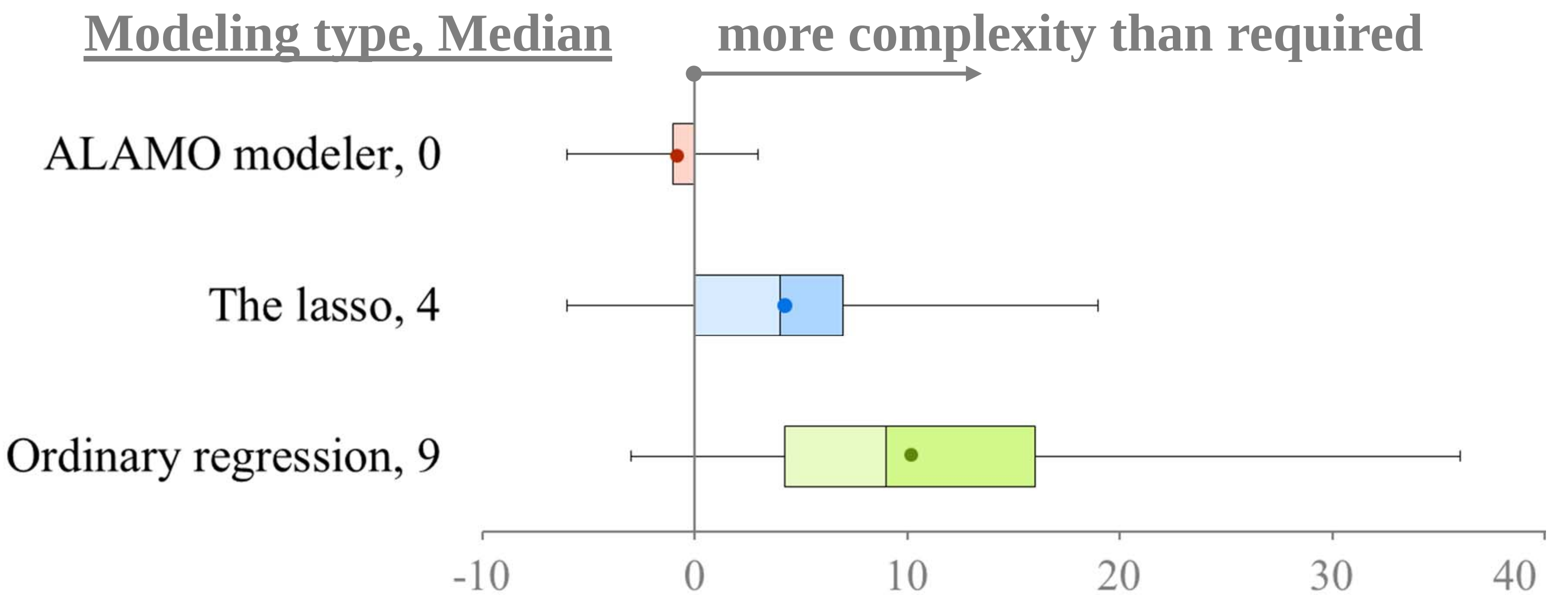
```
ninputs 1
noutputs 1

xmin -5
xmax 5

ndata 11

BEGIN_DATA
-5       25
-4       16
-3       9
-2       4
-1       1
0        0
1        1
2        4
3        9
4        16
5        25
END_DATA

logfcns 1
expfcns 1
sinfcns 1
cosfcns 1
monomialpower 1 2 3
```

128 alternative models

ALAMO OUTPUT

```
Step 1: Model building using BIC

Model building for variable z
----
BIC = -0.100E+31 with z = x**2.0

Calculating quality metrics on observed data set.

Quality metrics for output z
----------------------------
SSE OLR:                 0.00
SSE:                     0.00
RMSE:                    0.00
R2:                      1.00
Model size:              1
BIC:                     -0.100E+31
Cp:                      -9.00
AICc:                    -0.100E+31
HQC:                     -0.100E+31
MSE:                     0.00
SSEp:                    0.00
RIC:                     3.89

Total execution time 0.30E-02 s
```

SIMPLE AND ACCURATE MODELS

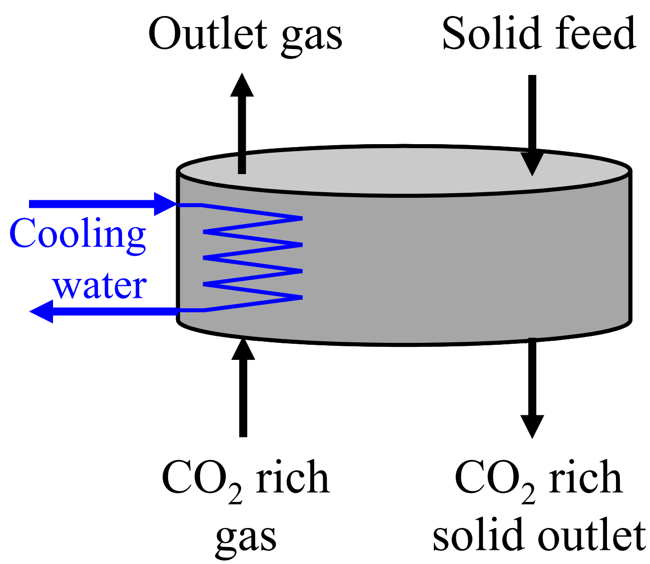


[Number of terms in the surrogate model] – [Number of terms in the true function]

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

WORKING WITH A SIMULATOR

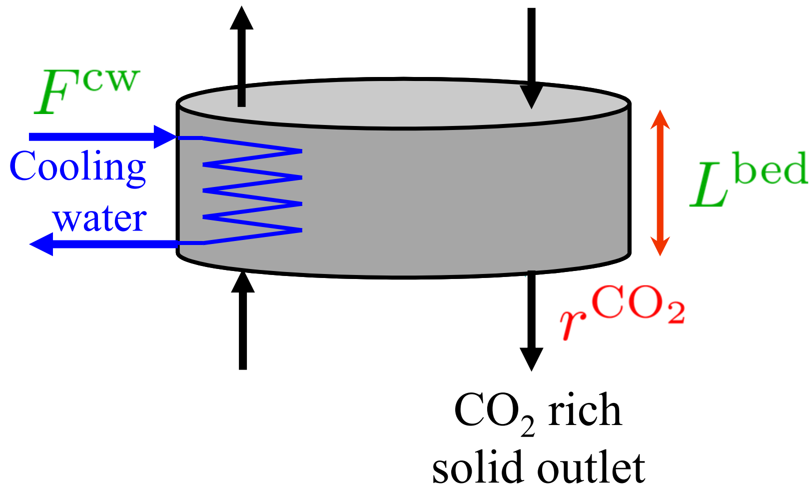
BUBBLING FLUIDIZED BED ADSORBER



Goal: Optimize a bubbling fluidized bed reactor by

- **Minimizing the increased cost of electricity**
- **Maximizing CO_2 removal**

BUBBLING FLUIDIZED BED ADSORBER



Generate model of % CO_2 removal:

$$r^{\mathrm{CO_2}}(L^{\mathrm{bed}}, F^{\mathrm{cw}}) = f_1(L^{\mathrm{bed}}, F^{\mathrm{cw}})$$

Over the Range:

$$\begin{aligned} 1\ \mathrm{m} &\le L^{\mathrm{bed}} \le 10\ \mathrm{m} \\ 1 \cdot 10^4\ \frac{\mathrm{kmol}}{\mathrm{h}} &\le F^{\mathrm{cw}} \le 20 \cdot 10^4\ \frac{\mathrm{kmol}}{\mathrm{h}} \end{aligned}$$

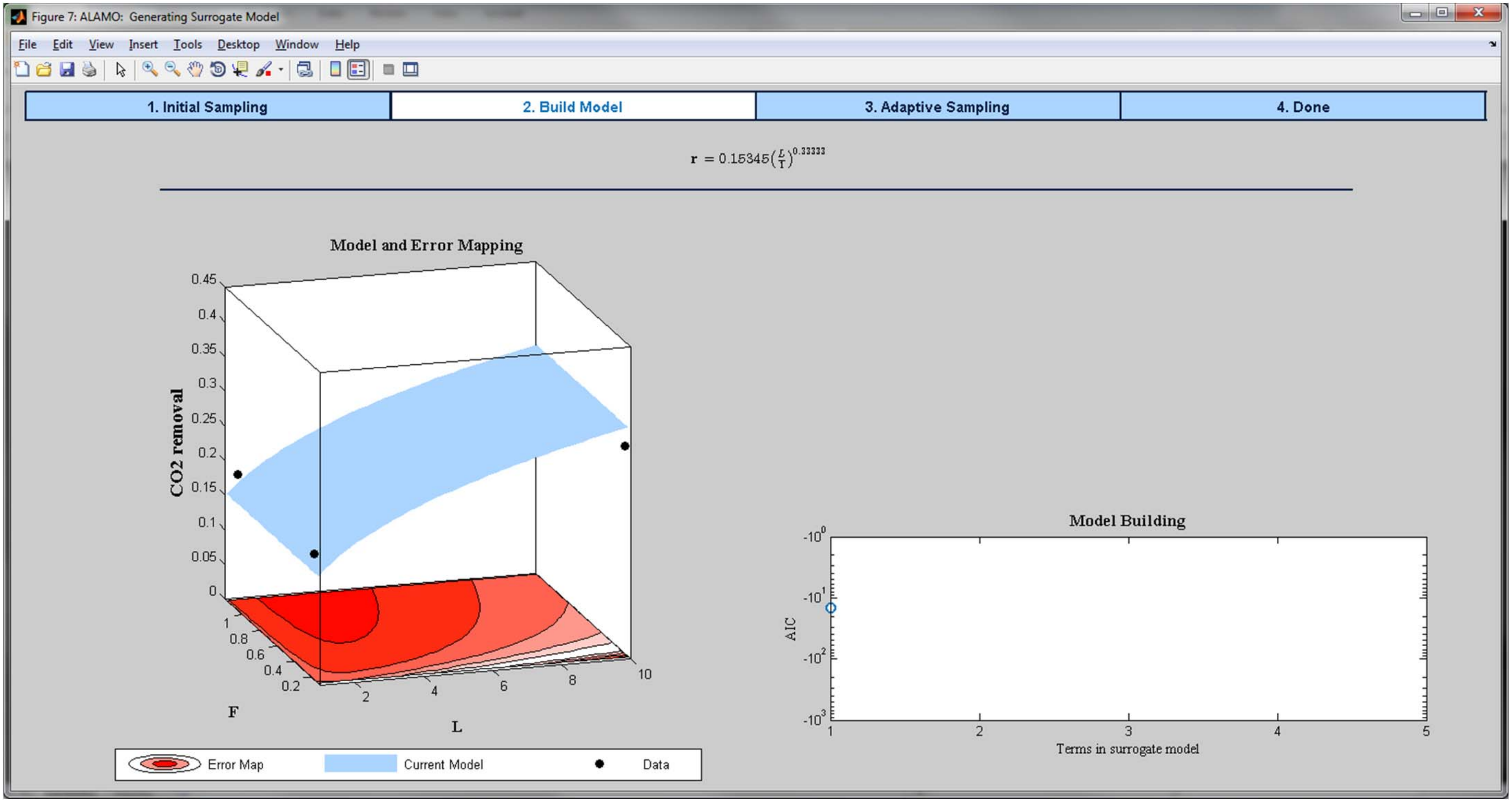
POTENTIAL MODEL

$$r^{CO_2} = \beta_0 + \beta_1 F + \beta_2 L + \beta_3 e^F + \beta_4 e^L + \beta_5 \ln(F) + \beta_6 \ln(L) + \beta_7 \frac{F}{L} + \beta_8 \frac{2F}{L^2} + \beta_9 F L^2 + \beta_{10} F^2 L + \beta_{11} \frac{F}{\sqrt{L}} + \beta_{12} F \sqrt{L} + \beta_{13} \frac{L}{\sqrt{F}} + \beta_{14} \sqrt{F} L + \beta_{15} \frac{F}{L^{\frac{1}{3}}} + \beta_{16} F L^{\frac{1}{3}} + \beta_{17} \frac{L}{F^{\frac{1}{3}}} + \beta_{18} F^{\frac{1}{3}} L + \beta_{19} 2 F \ln(F) + \beta_{20} F \ln(L) + \beta_{21} L \ln(F) + \beta_{22} L \ln(L) + \beta_{23} \frac{1}{F} + \beta_{24} \frac{1}{F^2} + \beta_{25} F^2 + \beta_{26} \frac{1}{F^3} + \beta_{27} \frac{1}{\sqrt{F}} + \beta_{28} \sqrt{F} + \beta_{29} F^3 + \beta_{30} \frac{1}{F^{\frac{1}{3}}} + \beta_{31} F^{\frac{1}{3}} + \beta_{32} \frac{1}{L} + \beta_{33} \frac{1}{L^2} + \beta_{34} L^2 + \beta_{35} \frac{1}{L^3} + \beta_{36} \frac{1}{\sqrt{L}} + \beta_{37} \sqrt{L} + \beta_{38} L^3 + \beta_{39} \frac{1}{L^{\frac{1}{3}}} + \beta_{40} L^{\frac{1}{3}} + \beta_{41} \frac{1}{F L} + \beta_{42} \frac{2 L^2}{F} + \beta_{43} \frac{\sqrt{L}}{F} + \beta_{44} \frac{F^2}{L^2} + \beta_{45} F^2 L^2 + \beta_{46} \frac{\sqrt{F}}{L} + \beta_{47} \frac{F^3}{L} + \beta_{48} \frac{1}{F^3 L^3} + \beta_{49} \frac{1}{\sqrt{F}\sqrt{L}} + \beta_{50} \frac{\sqrt{F}}{\sqrt{L}} + \beta_{51} \sqrt{F}\sqrt{L} + \beta_{52} \frac{F^3}{L^3} + \beta_{53} F^3 L^3 + \beta_{54} \frac{L^{\frac{1}{3}}}{F} + \beta_{55} \frac{F^{\frac{1}{3}}}{L} + \beta_{56} \frac{1}{F^{\frac{1}{3}} L^{\frac{1}{3}}} + \beta_{57} \frac{F^{\frac{1}{3}}}{L^{\frac{1}{3}}} + \beta_{58} F^{\frac{1}{3}} L^{\frac{1}{3}} + \beta_{59} F^{\frac{1}{3}} \ln(L) + \beta_{60} L^{\frac{1}{3}} \ln(F) + \beta_{61} L^{\frac{1}{3}} \ln(L) + \beta_{62} F L$$

66 basis functions

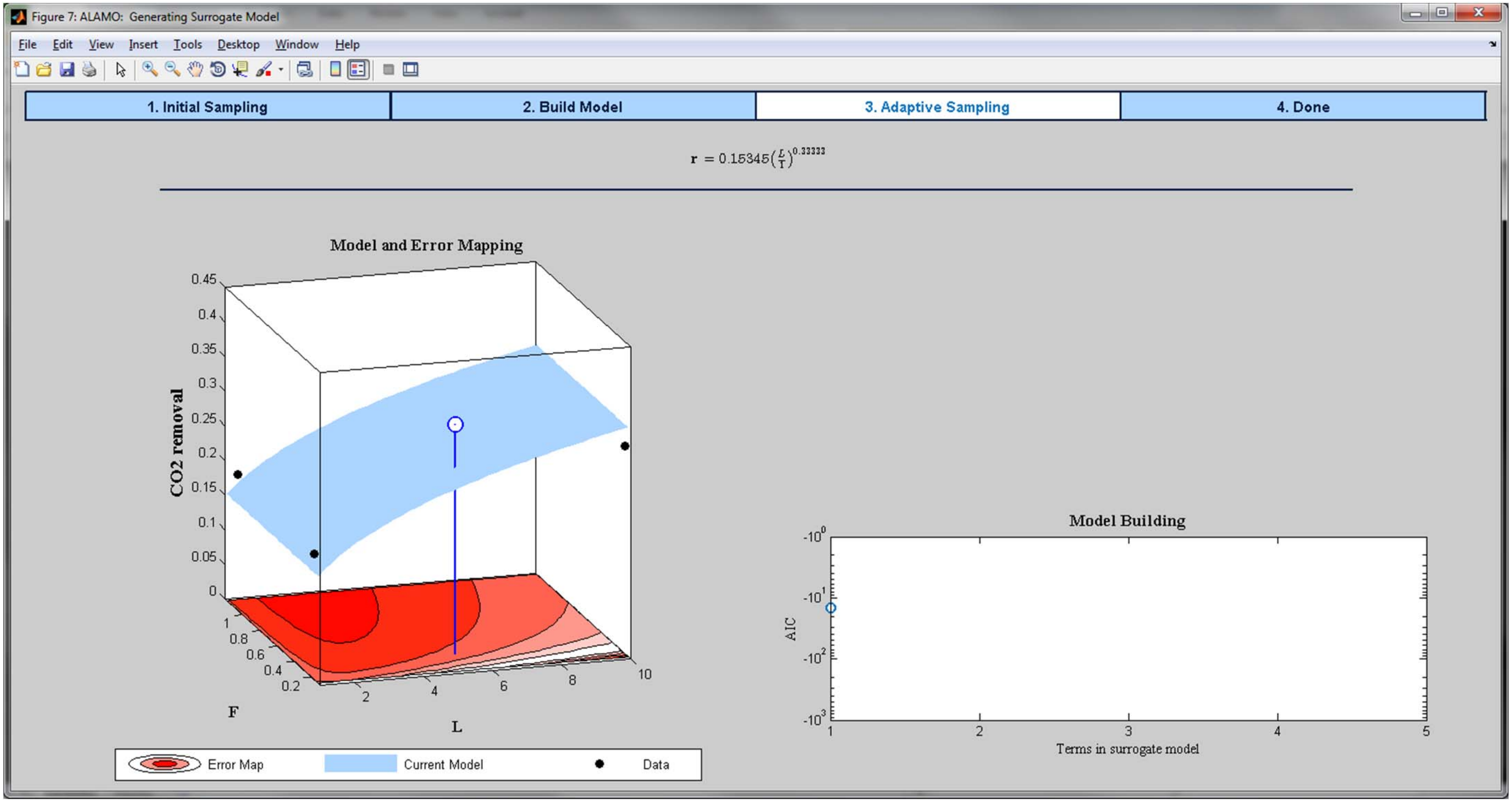
Not tractable in an algebraic superstructure formulation!

BUILD MODEL



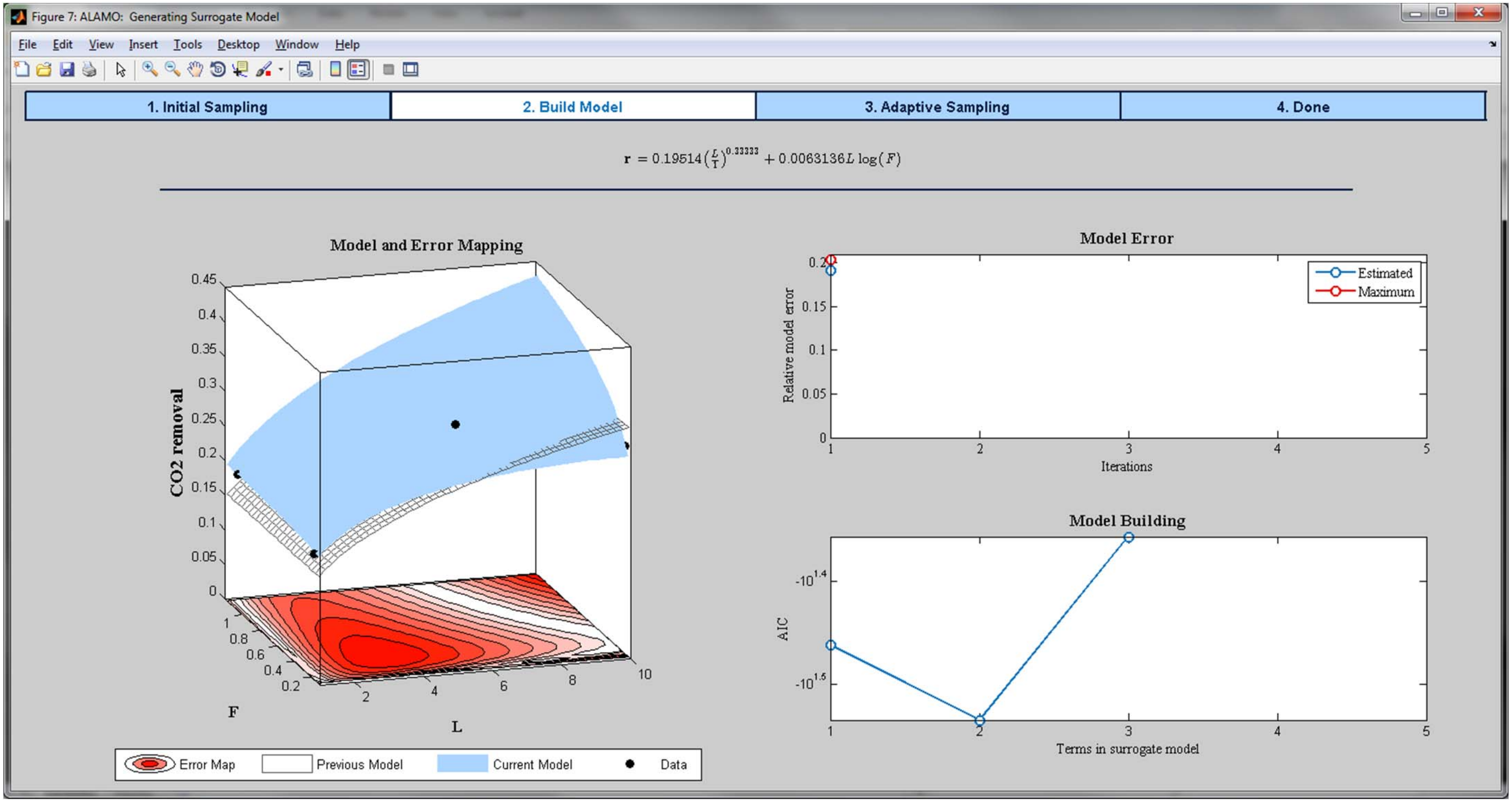
- **Current model form, model surface, and data points**

ADAPTIVE SAMPLING



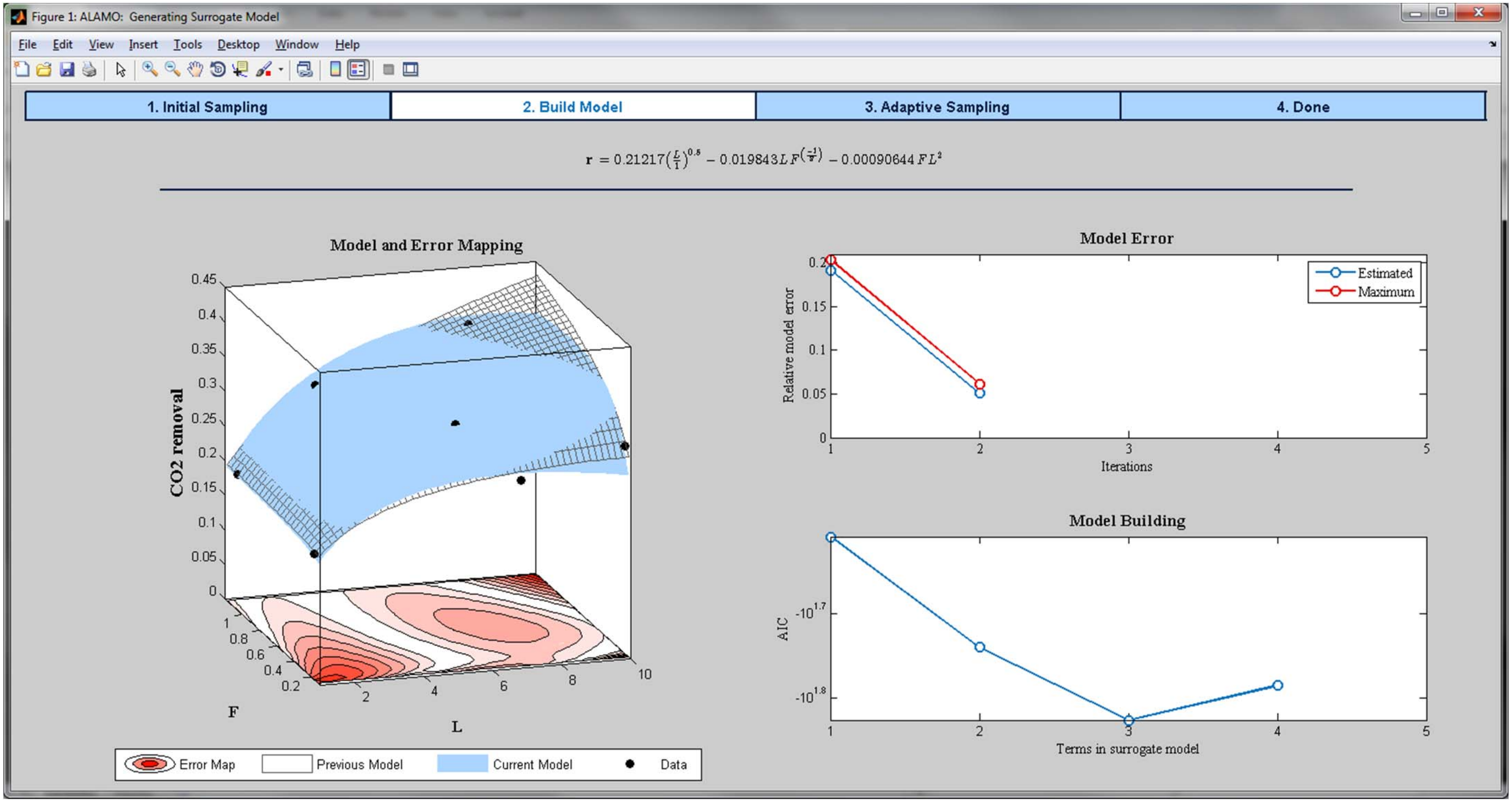
- **New data point(s) shown**

REBUILD MODEL



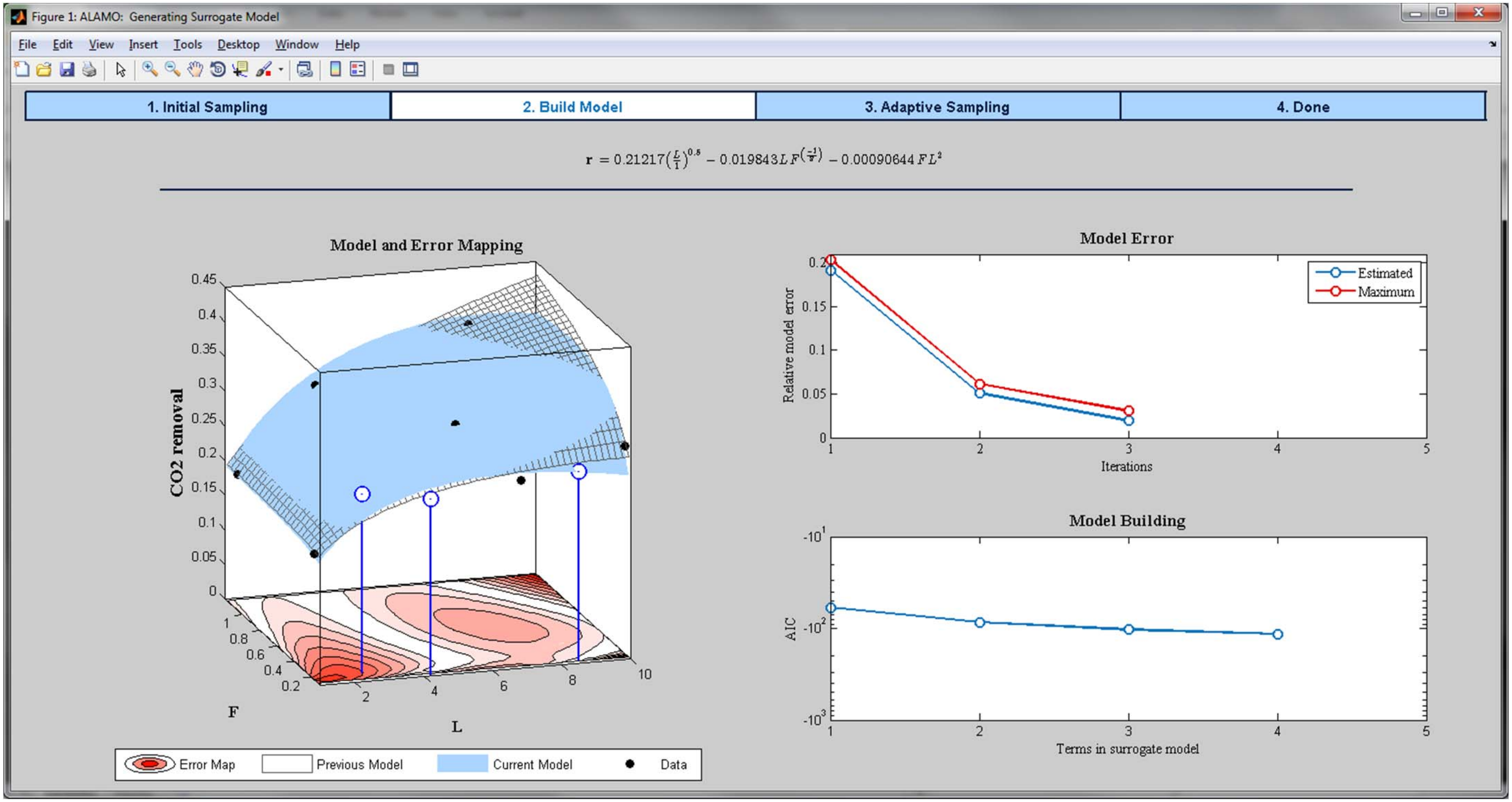
- **Previous model in mesh**

REBUILD MODEL

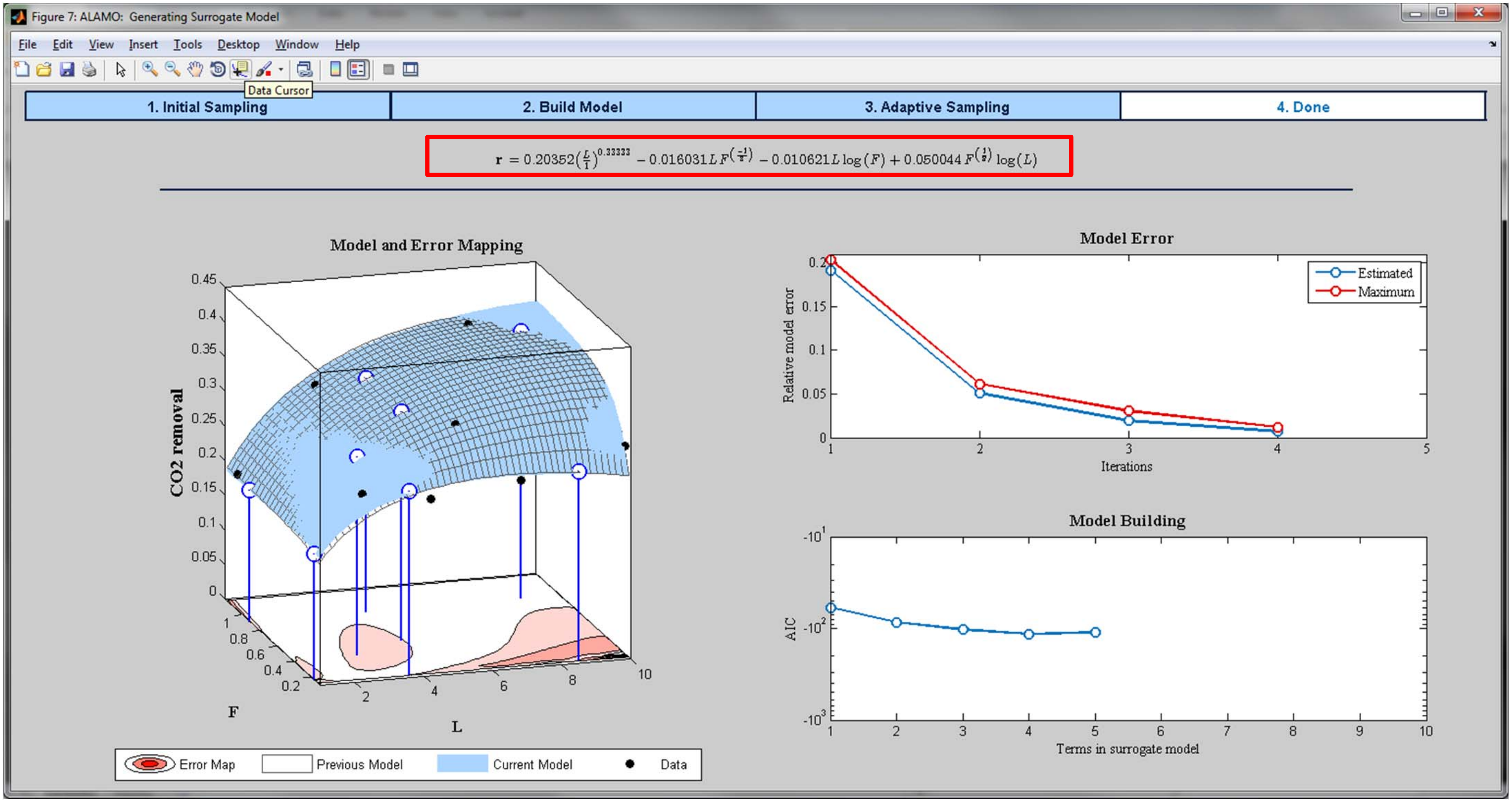


- **Model error(iterations) and AIC(T) shown**

ADAPTIVE SAMPLING

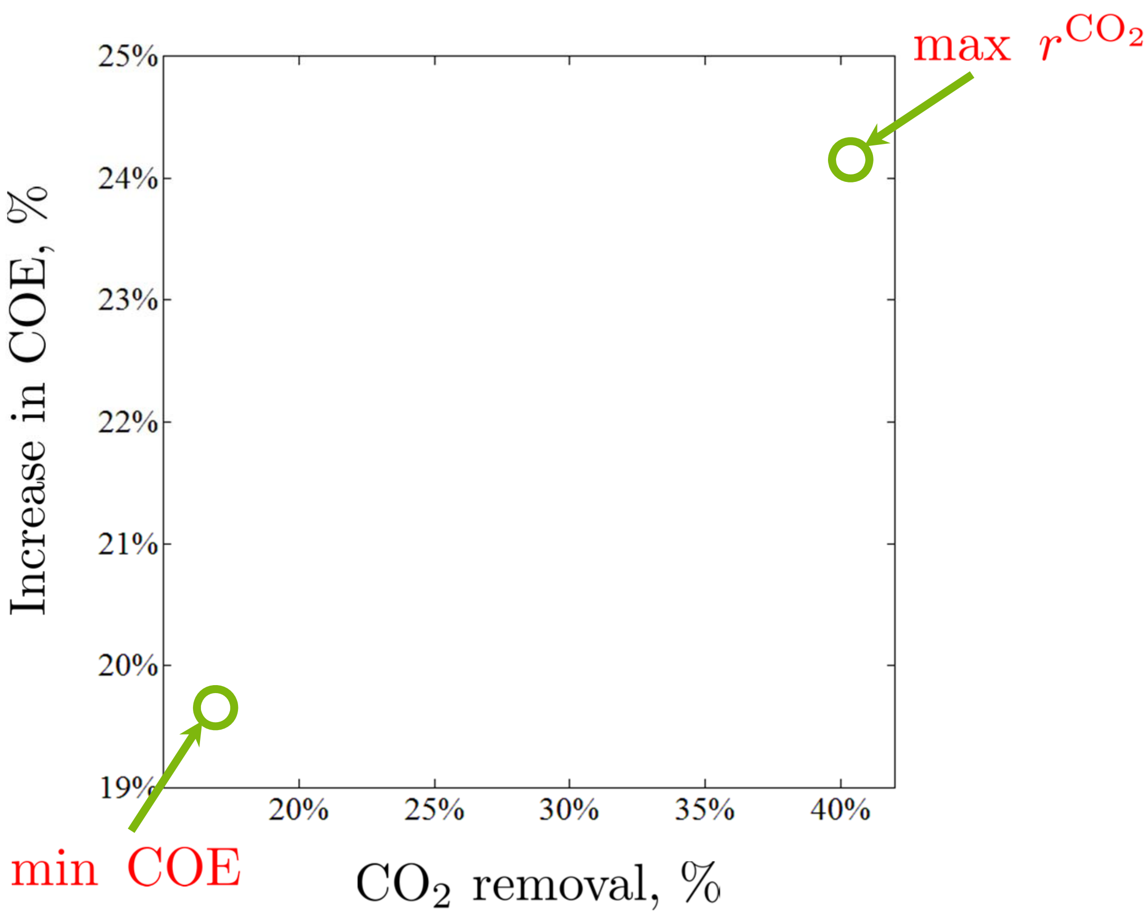


REBUILD MODEL

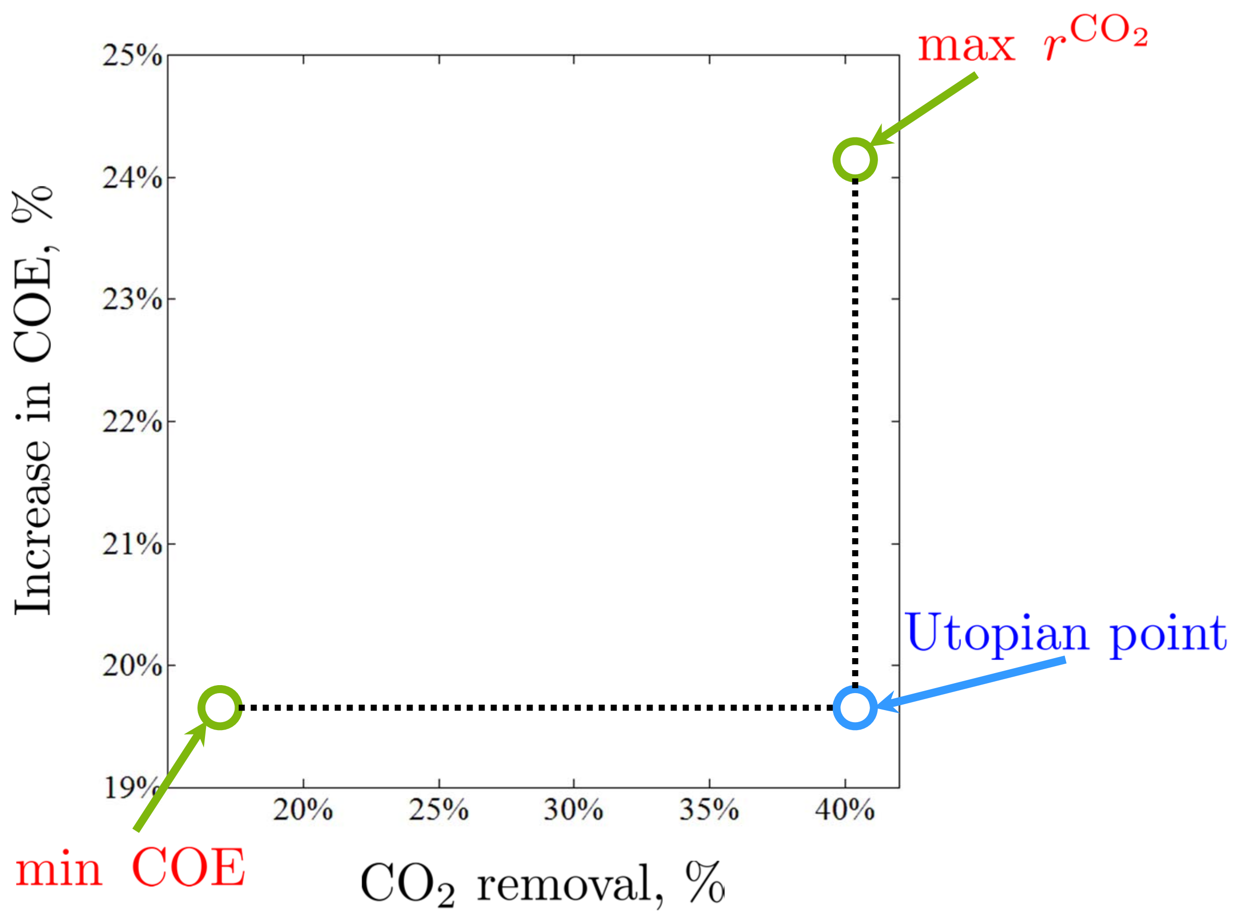


- **Sufficiently low error, model converges**

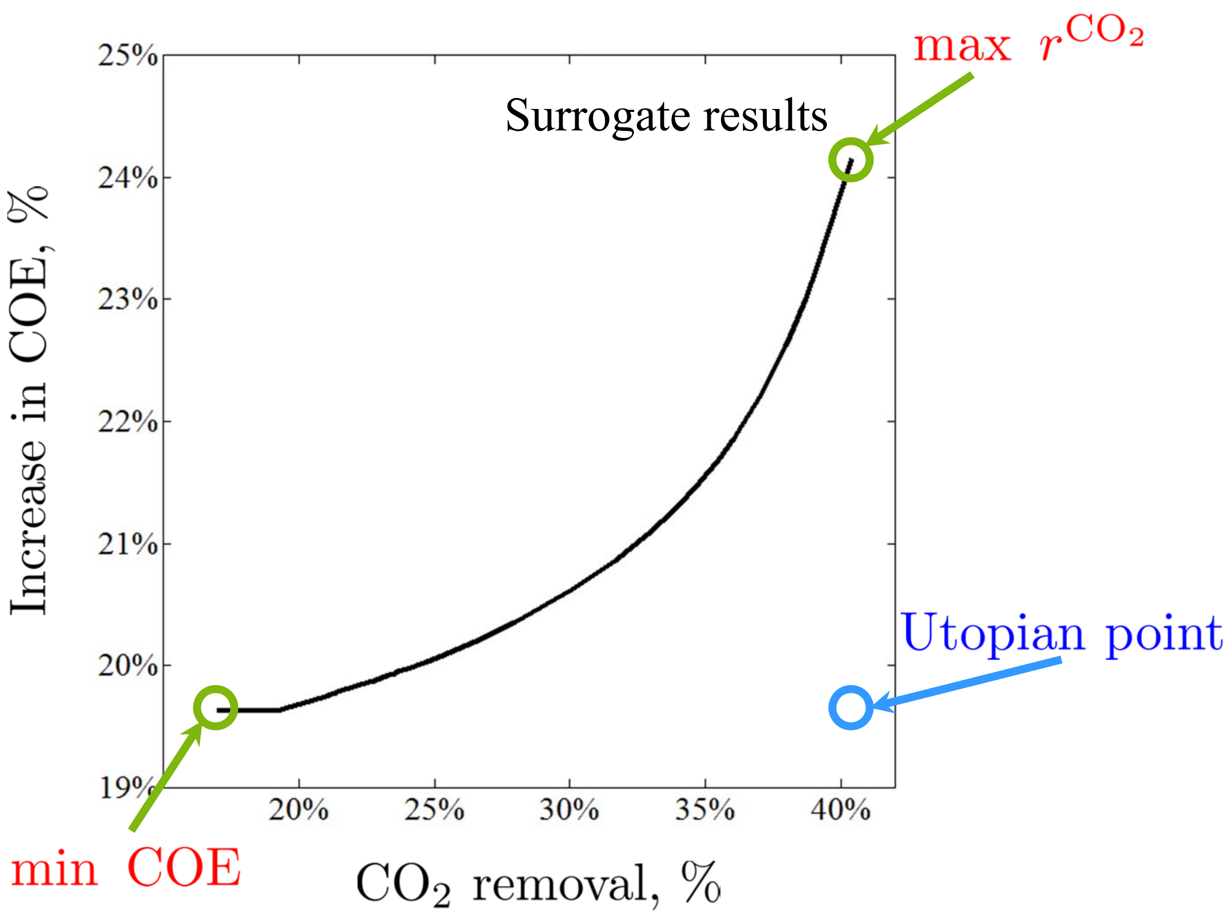
OPTIMAL PARETO CURVE



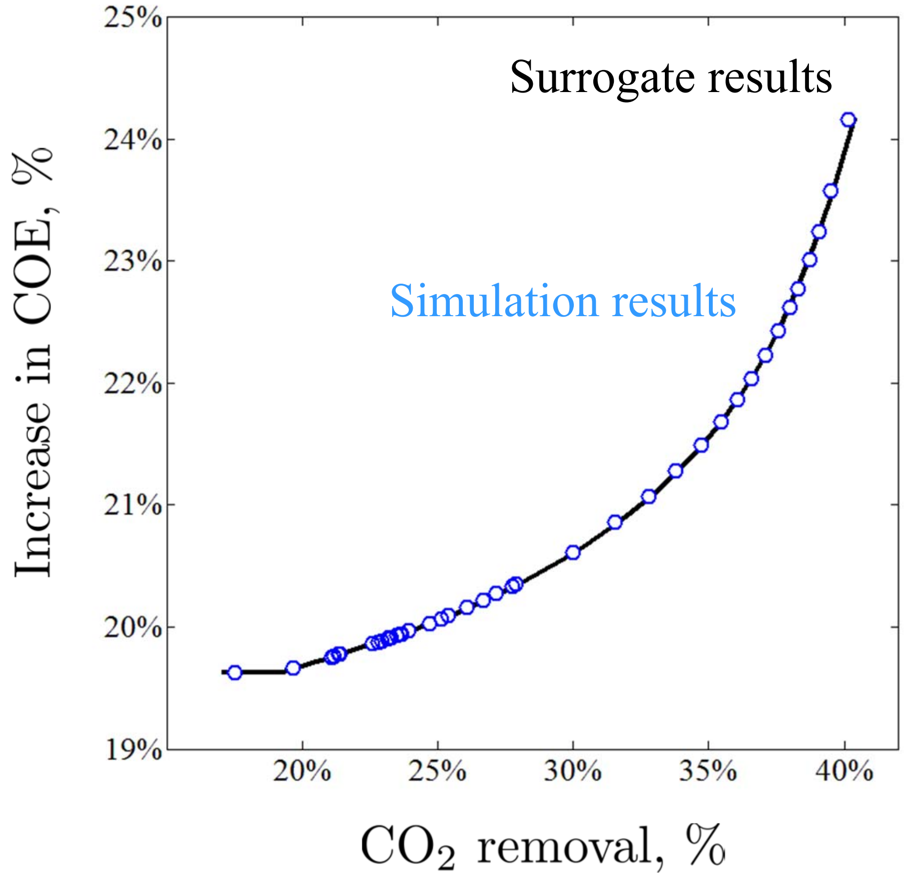
OPTIMAL PARETO CURVE



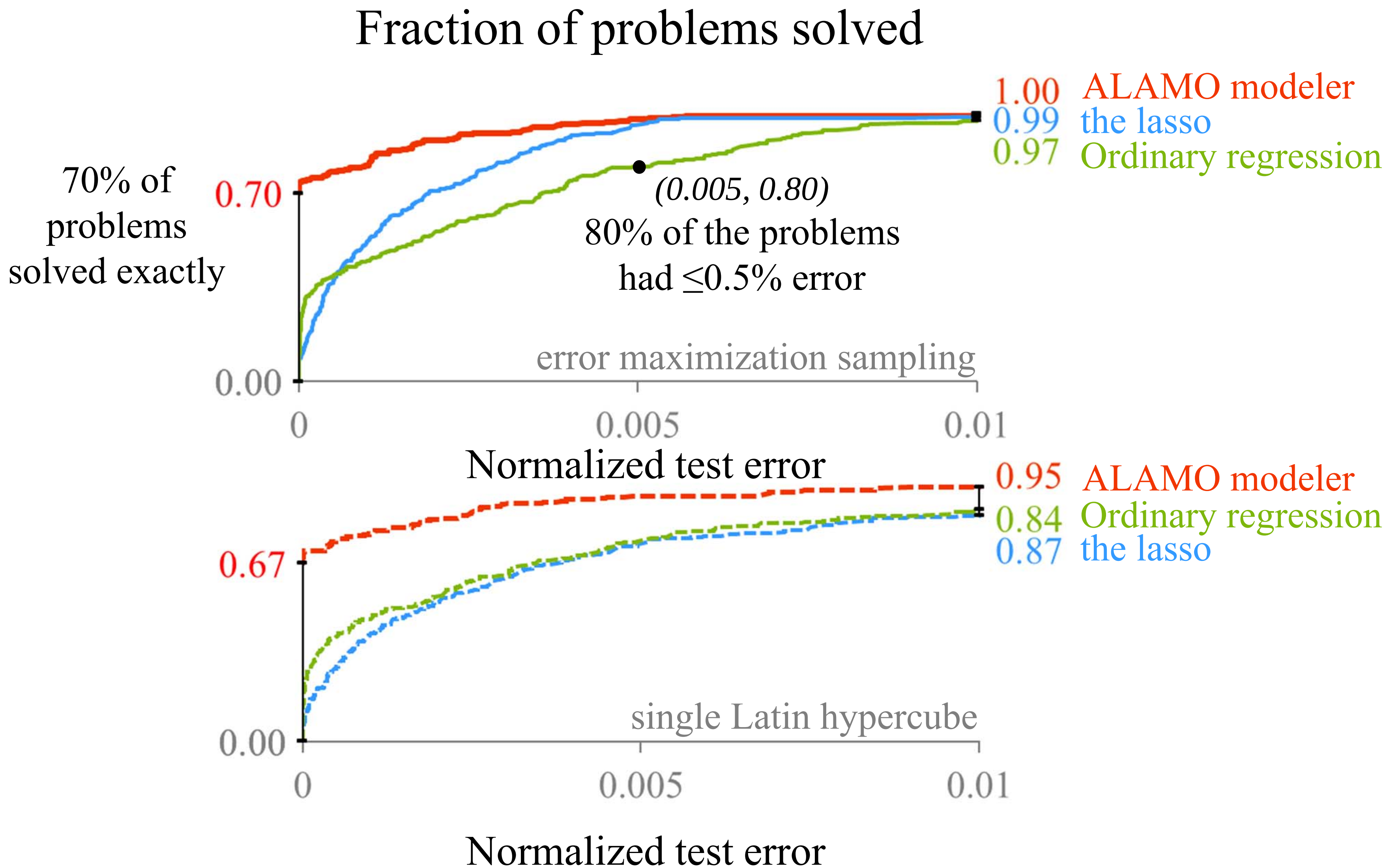
OPTIMAL PARETO CURVE



OPTIMAL PARETO CURVE



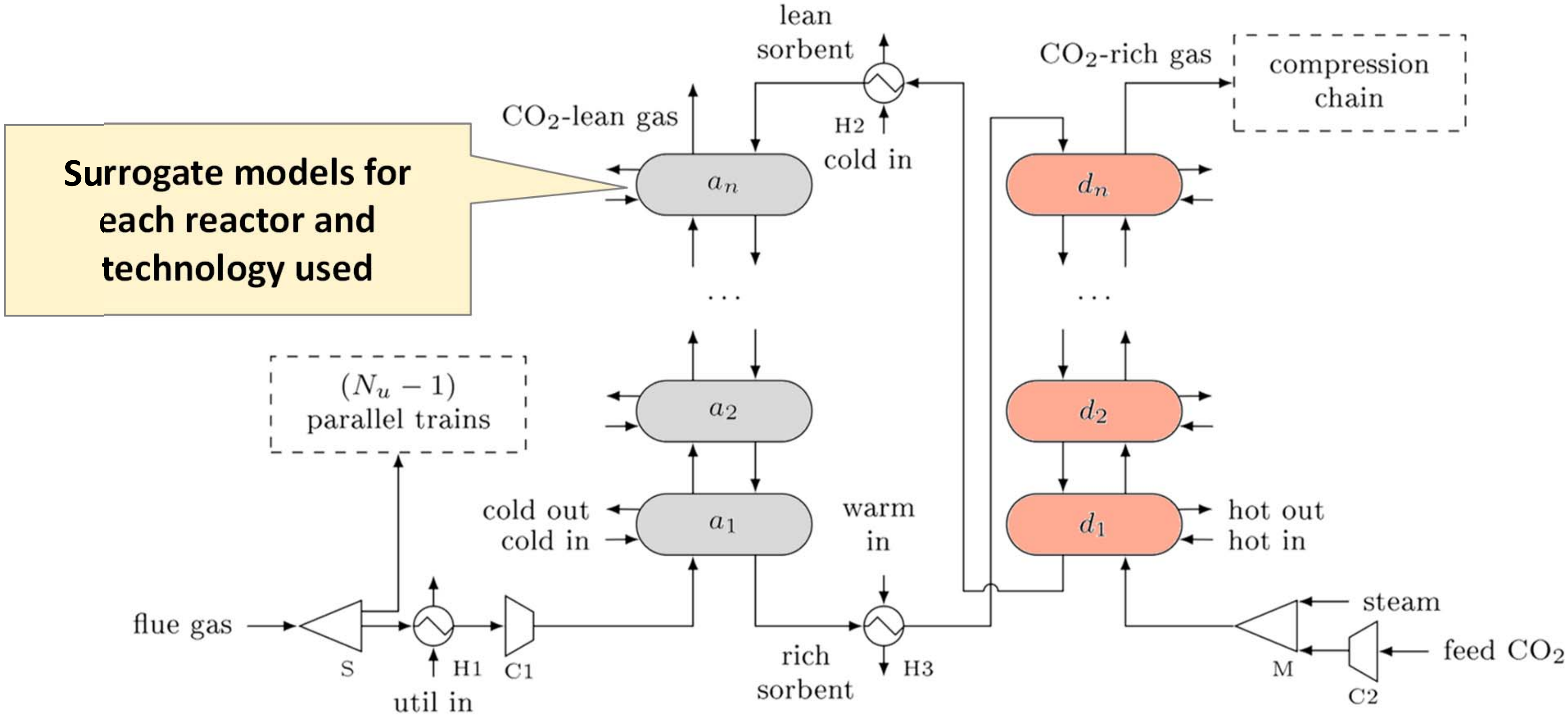
SAMPLING EFFICIENCY



Results with 45 known functions with bases that are available to all modeling methods

SUPERSTRUCTURE OPTIMIZATION

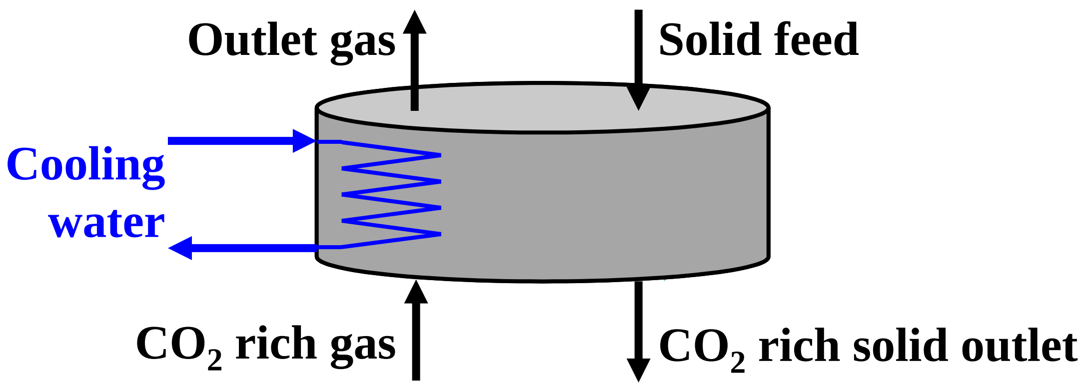
CARBON CAPTURE SYSTEM DESIGN



- **Discrete decisions:** How many units? Parallel trains? What technology used for each reactor?
- **Continuous decisions:** Unit geometries
- **Operating conditions:** Vessel temperature and pressure, flow rates, compositions

BUBBLING FLUIDIZED BED

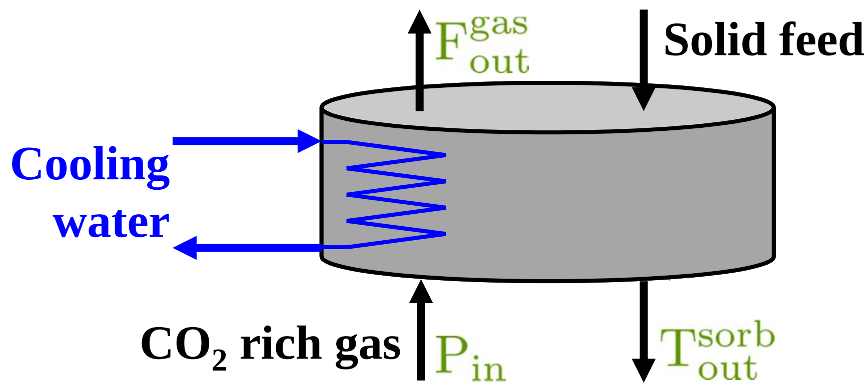
Bubbling fluidized bed adsorber diagram



- **Model inputs (16 total)**
  - Geometry (3)
  - Operating conditions (5)
  - Gas mole fractions (2)
  - Solid compositions (2)
  - Flow rates (4)
- **Model outputs (14 total)**
  - Geometry required (2)
  - Operating condition required (1)
  - Gas mole fractions (3)
  - Solid compositions (3)
  - Flow rates (2)
  - Outlet temperatures (3)

Model created by Andrew Lee at the National Energy Technology Laboratory

EXAMPLE MODELS - ADSORBER



$$P_{in} = 1.0\,P_{out} + 0.0231\,L_b - 0.0187\,\ln(0.167\,L_b) - 0.00626\,\ln(0.667\,v_{gi}) - \frac{51.1\,xHCO3_{in}^{ads}}{F_{in}^{gas}}$$

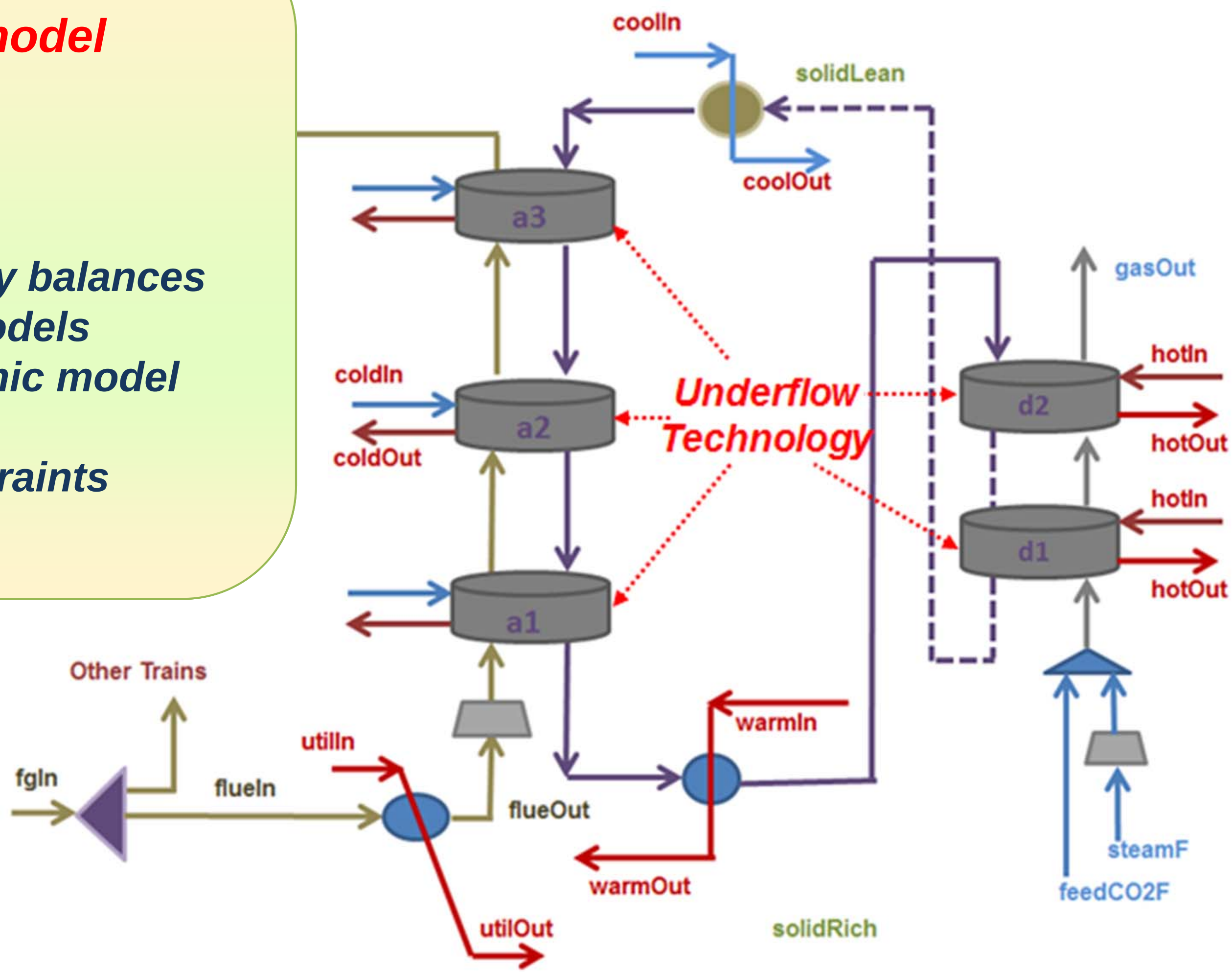
$$T_{out}^{sorb} = 1.0\,T_{in}^{gas} - \frac{\left(1.77 \cdot 10^{-10}\right)\,NX^2}{\gamma^2} - \frac{3.46}{NX\,T_{in}^{gas}\,T_{in}^{sorb}} + \frac{1.17 \cdot 10^4}{F^{sorb}\,NX\,xH2O_{in}^{ads}}$$

$$F_{out}^{gas} = 0.797\,F_{in}^{gas} - \frac{9.75\,T_{in}^{sorb}}{\gamma} - 0.77\,F_{in}^{gas}\,xCO2_{in}^{gas} + 0.00465\,F_{in}^{gas}\,T_{in}^{sorb} - 0.0181\,F_{in}^{gas}\,T_{in}^{sorb}\,xH2O_{in}^{gas}$$

SUPERSTRUCTURE OPTIMIZATION

Mixed-integer nonlinear programming model

- ***Economic model***
- ***Process model***
- ***Material balances***
- ***Hydrodynamic/Energy balances***
- ***Reactor surrogate models***
- ***Link between economic model and process model***
- ***Binary variable constraints***
- ***Bounds for variables***



MINLP solved with BARON

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
  - ✓ Variable groups
  - ✓ Forthcoming: constraints on gradient of response variables

- **Built on top of state-of-the-art optimization solvers**

- **Available from www.minlp.com**