Columbia University
October 10, 2014

Uncertainty Quantification Framework for Modeling Prediction

Michael Frenklach

Collaborators:

- Andy Packard
- W. A. Lester, Jr. (DFT)
- P. Westmoreland (ALS)
- N. Slavinskaya (SynGas)
- R. Feely, P. Seiler, T. Russi, D. Yeates, X. You, F. Lei, D. Edwards, D. Zubarev, W. Speight

Supported by: NSF, AFOSR, DOE-NNSA (PSAAP II)

Berkeley
University of California

OUTLINE

- Introduction: UQ-predictive modeling
- Bound-To-Bound Data Collaboration
- Introductory case: Energetics of water clusters
- Full-blown case: Combustion of natural gas

THE KEY CHALLENGE:

PREDICTION

"Model predicts reasonably well the experimental behavior"

"Model matches the experimental data"

"…excellent agreement between model and data."

"The model predictions match reasonably well the experimental data"

"Model predicts data" ?

"Model falls short in predicting experimental data"

"The prediction matches very well with experimental data"

"Simulation agrees well with the data"

"The model well predicts the data"

"Good agreement was found between the model and the data"

ANDREA SALTELLI
SILVIO FUNTOWICZ

When All Models Are Wrong

More stringent quality criteria are needed for models used at the science/policy interface, and here is a checklist to aid in the responsible development and use of models.

Modeling and Predicting Behavioral Dynamics on the Web

Kira Radinsky[†]; Krysta Svore[†], Susan Dumais[†],
Jaime Teevan[†], Alex Bocharov[†], Eric Horvitz[†]

WWW 2012
Lyon, France

ABSTRACT

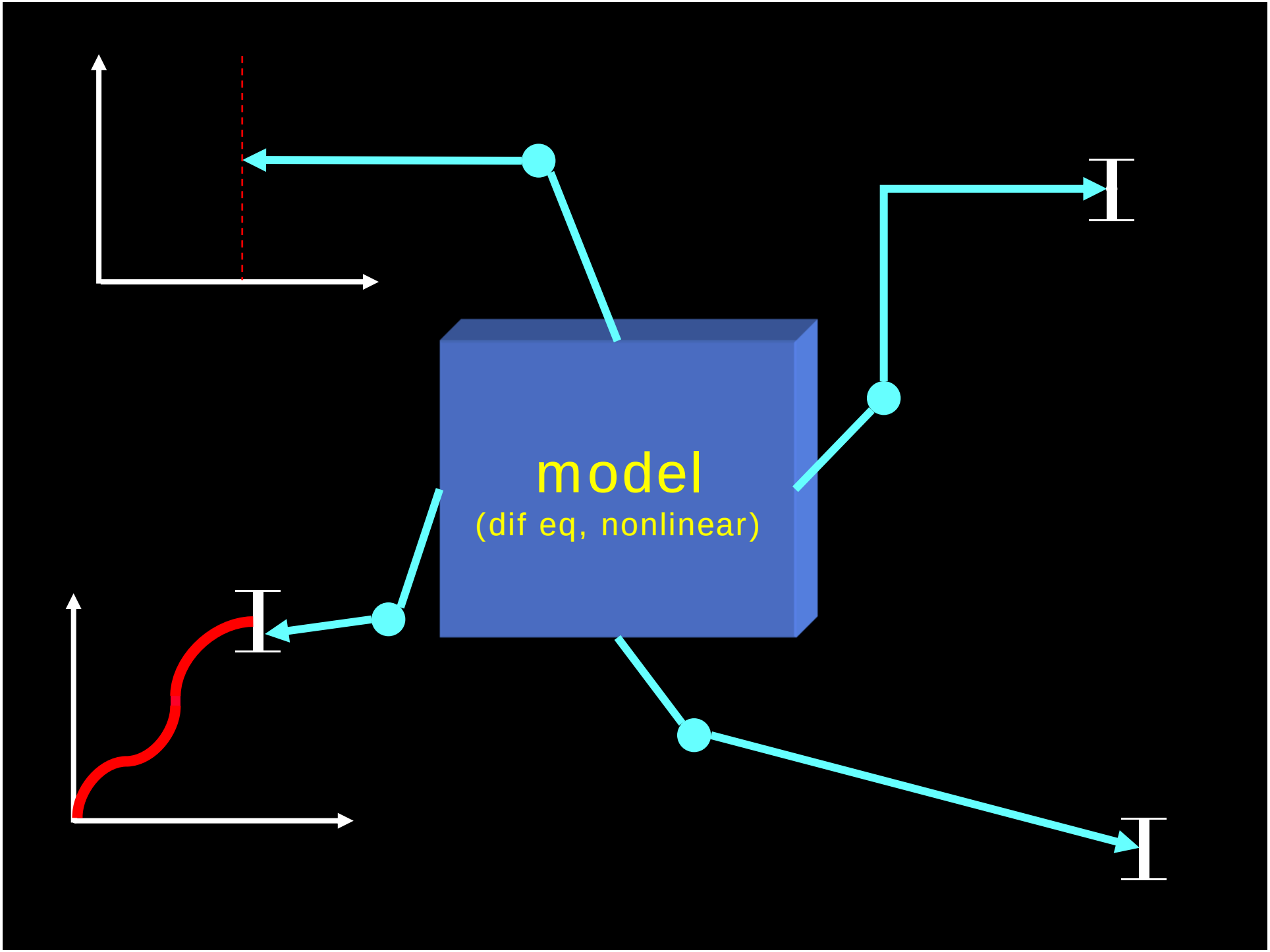
User behavior on the Web changes over time. For example, the queries that people issue to search engines, and the underlying informational goals behind the queries vary over time. In this paper, we examine how to model and predict this temporal user behavior. We develop a temporal modeling framework adapted from physics and signal processing that can be used to predict time-varying user behavior using smoothing and trends. We also explore other dynamics of Web behaviors, such as the detection of periodicities and surprises. We develop a learning procedure that can be used to construct models of users' activities based on features of current and historical behaviors. The results of experiments indicate that by using our framework to predict user behavior, we can achieve significant improvements in prediction compared to baseline models that weight historical evidence the same for all queries. We also develop a novel learning algorithm that explicitly learns when to apply a given prediction model among a set of such models. Our improved temporal modeling of user behavior can be used to enhance query suggestions, crawling policies, and result ranking.

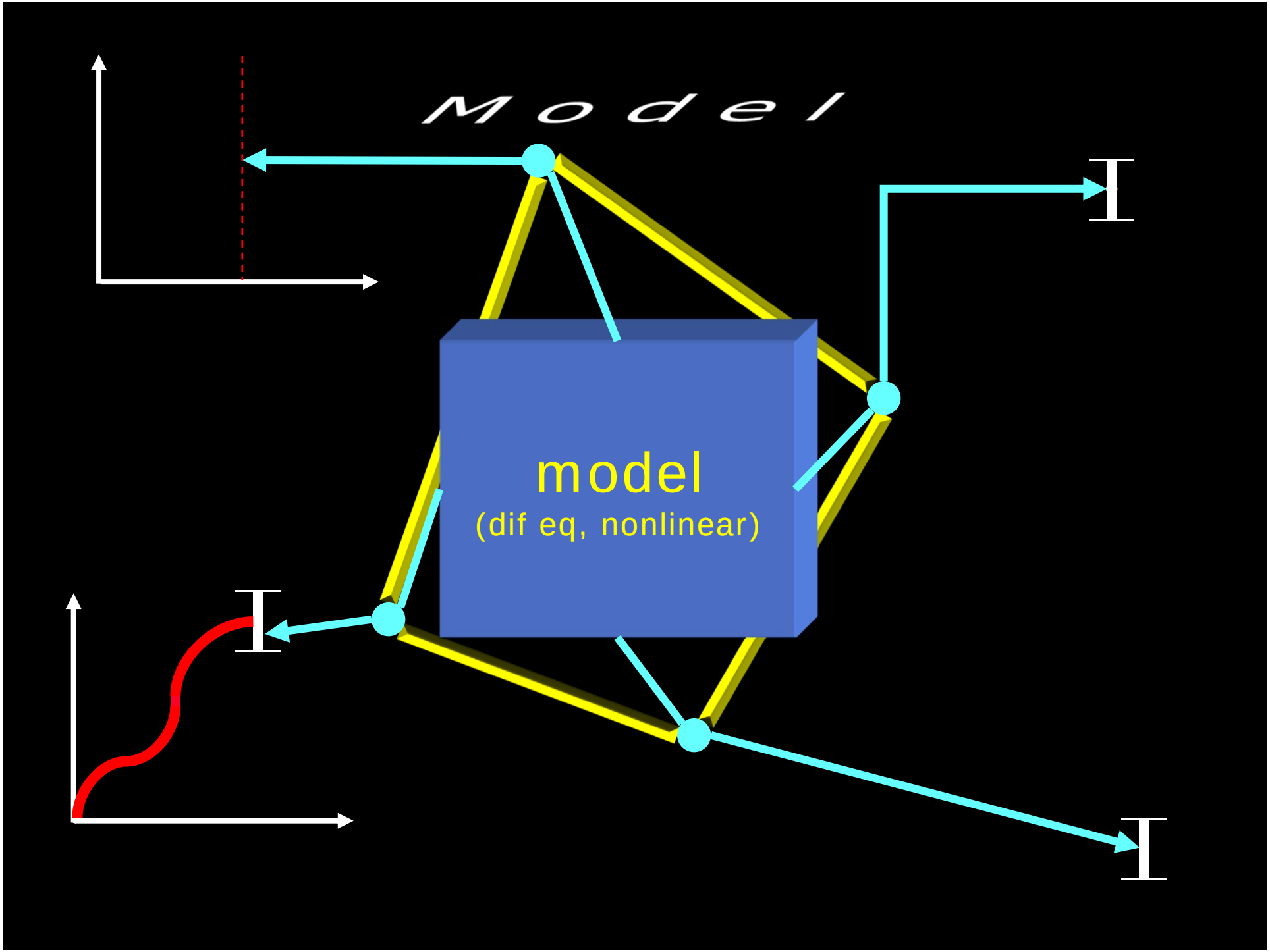
We develop a temporal modeling framework adapted from physics and signal processing …

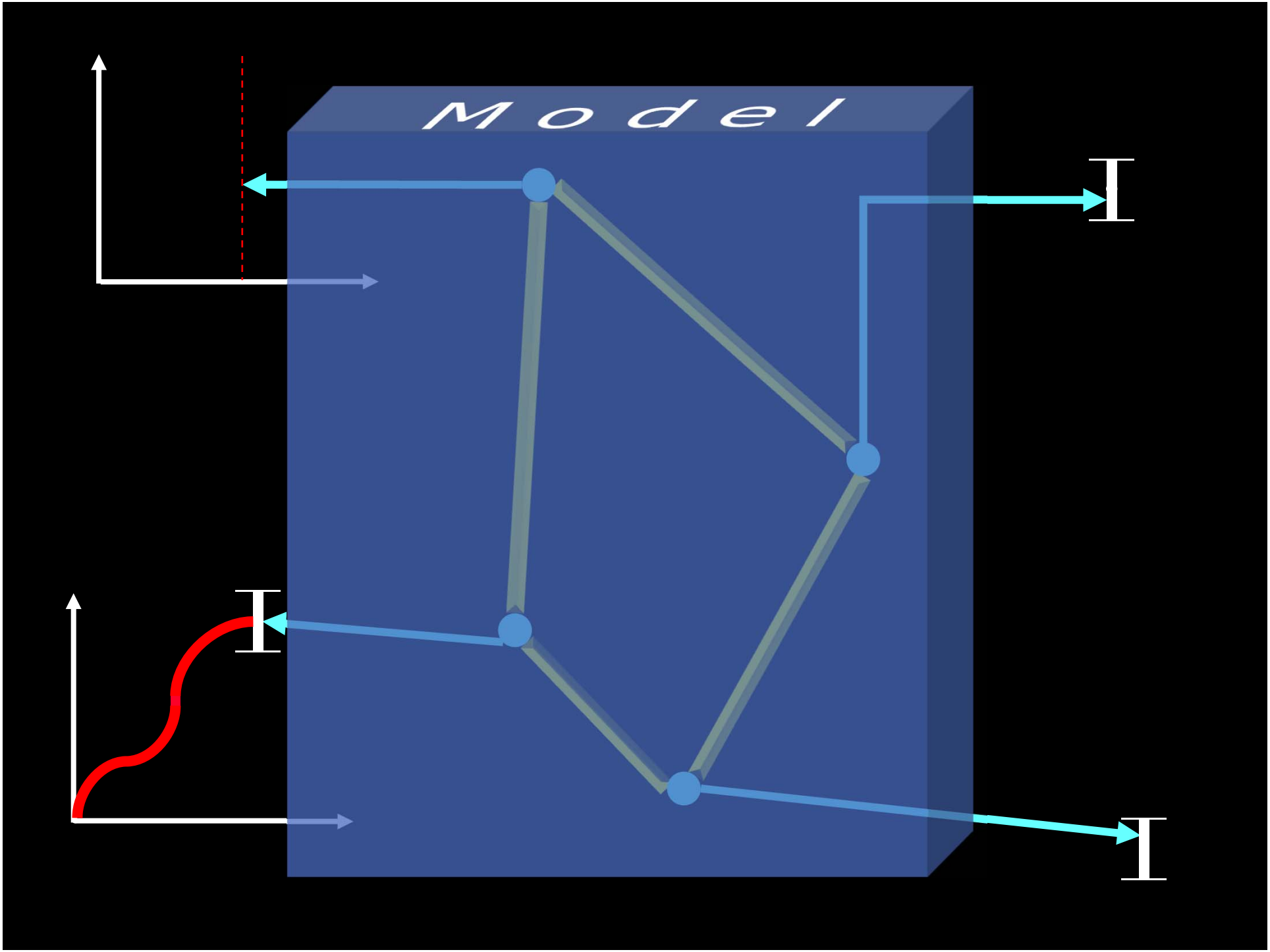
The results … indicate that by using our framework …. we can achieve significant improvements in prediction compare to baseline models …

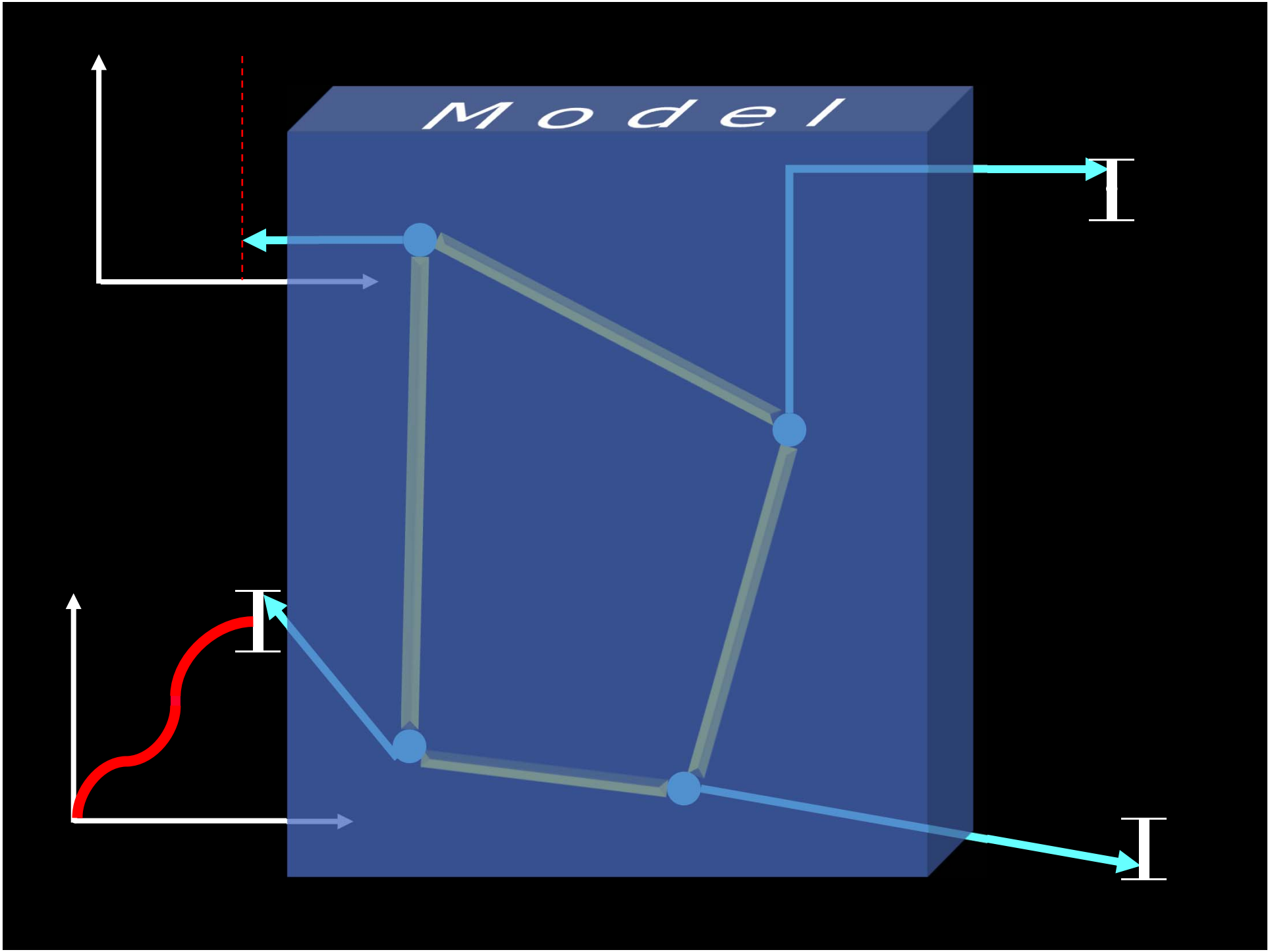
PREAMBLE

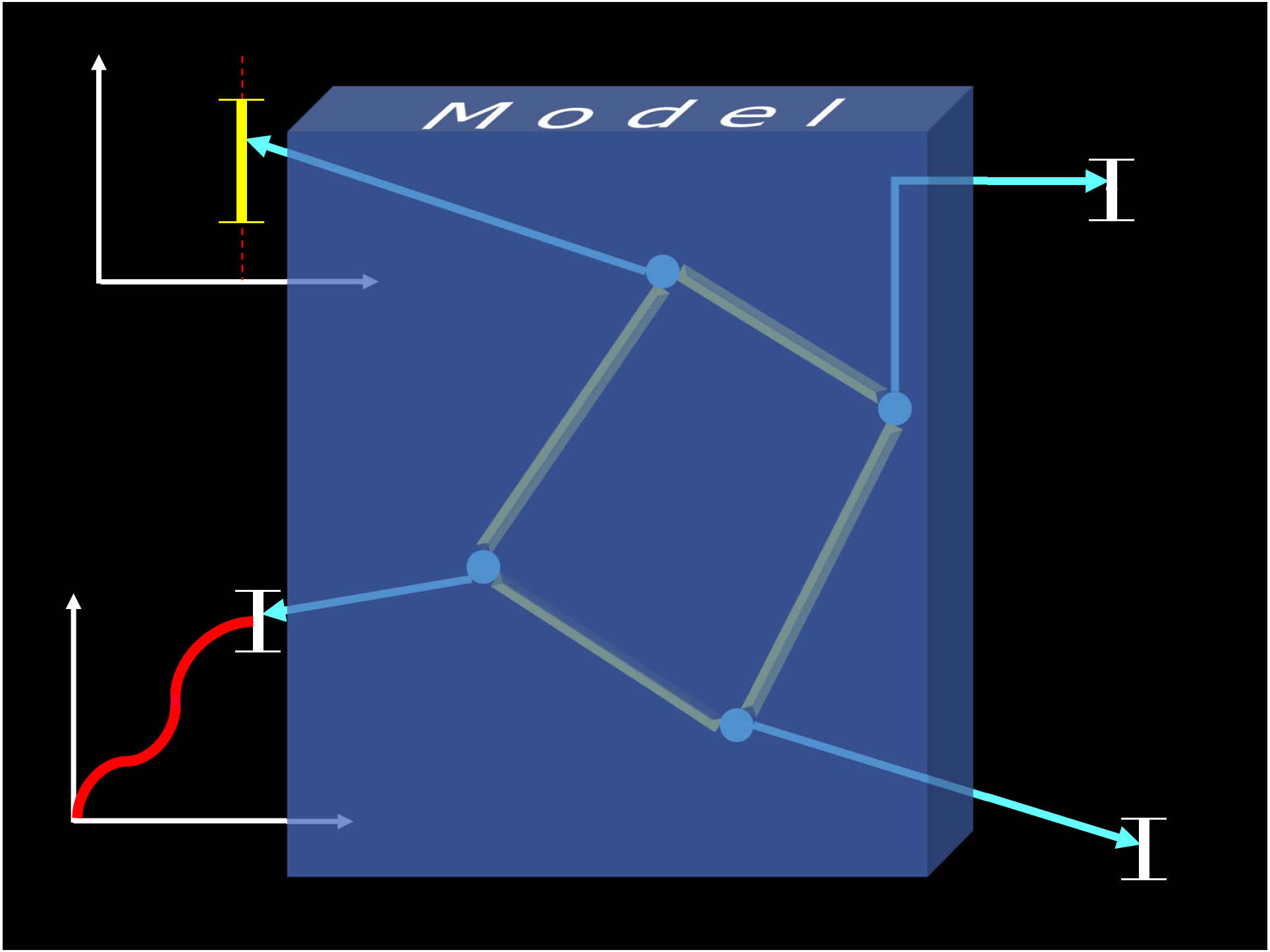
- Predictive $\rightarrow$ UQ-Predictive
- Physics-based models with the focus on data
- Validation is part of the process
- Dimensionality reduction is part of the process
- Practicality $\rightarrow$ use of surrogate models (Emulators)
- Data/Models
  - Access, sharing, documentation, …
  - Reproducibility


model
(dif eq, nonlinear)


Model
model
(dif eq, nonlinear)

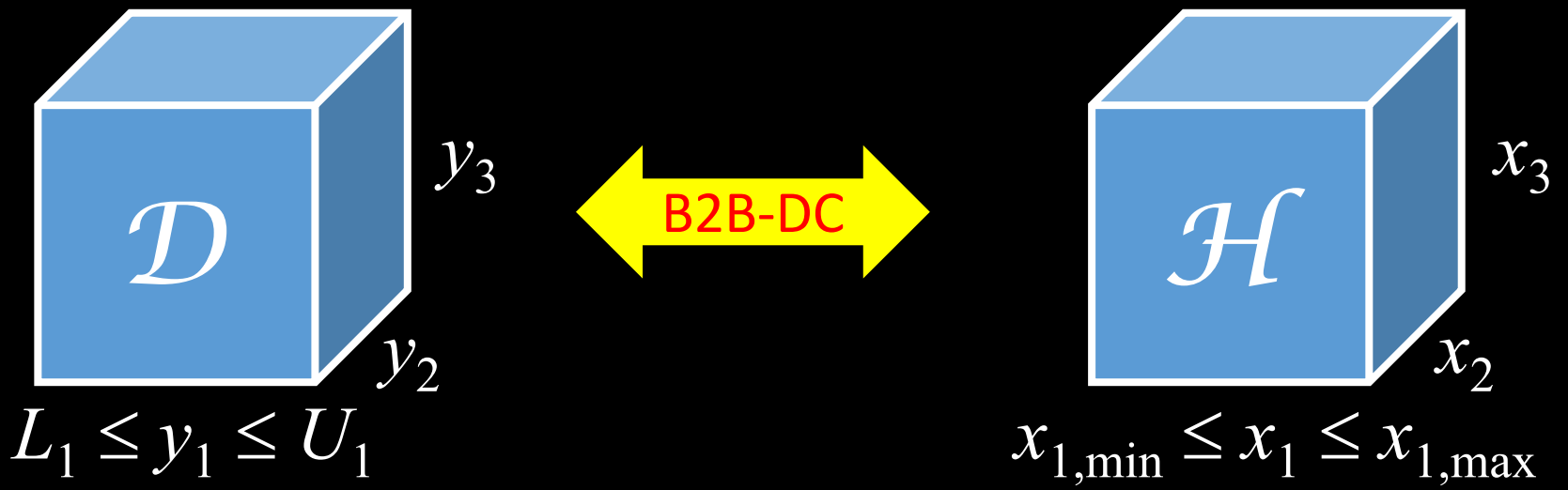

Model


Model


Model

Bound-to-Bound Data Collaboration (B2B-DC)

– an optimization-based framework for combining models and data to ascertain the collective information content

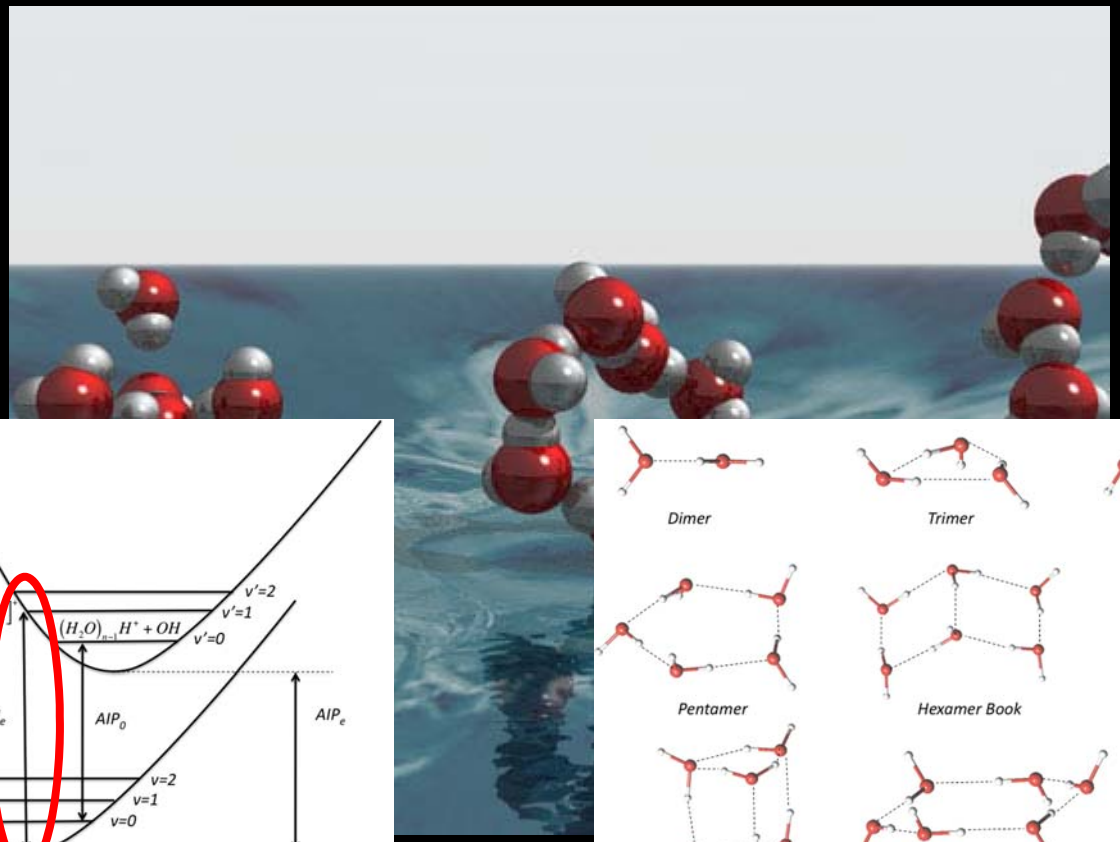


experimental uncertainties

prior knowledge on parameters

INTRODUCTORY CASE:

PREDICT IONIZATION POTENTIAL OF WATER CLUSTERS



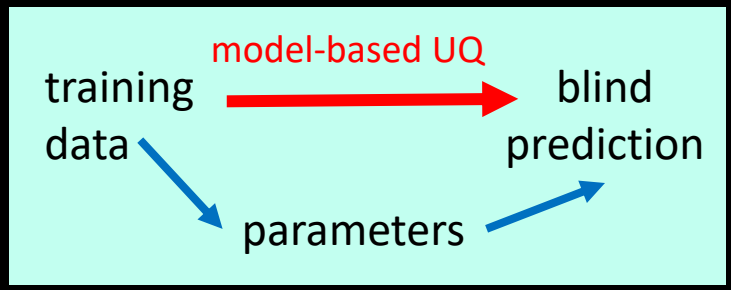
Phys Rev Lett 112: 253003 (2014)

QUANTUM-CHEMISTRY APPROACHES

- empirical: force-field – guessed potential, empirically fitted; …
- semi-empirical HF – quantum "core" with some terms replaced by parameters fitted to data (AM1, RM1, PM3, PM6, ZINDO, …)
- DFT with fitted parameters: meta-GGA (Truhlar, M05,M06,M11,…), double-hybrid DFT (Grimme), ...

$$E_{\mathrm{XC}} = \left(1-\alpha_{\mathrm{X}}\right) E_{\mathrm{X}}^{\mathrm{DFT}} + \alpha_{\mathrm{X}} E_{\mathrm{X}}^{\mathrm{HF}} + \left(1-\alpha_{\mathrm{C}}\right) E_{\mathrm{C}}^{\mathrm{DFT}} + \alpha_{\mathrm{C}} E_{\mathrm{C}}^{\mathrm{MP2}}$$

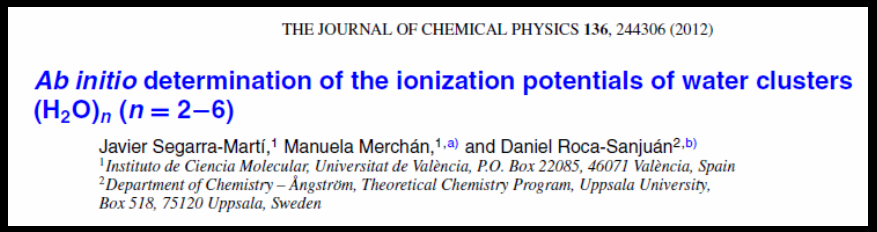
- "static" outcome: the optimized model needs (constant) retuning
- *the optimum is not unique!*
- partial loss of information (two-step process)



B2B-DC

Model: $E_{XC} = (1-\alpha) E_X^{GGA} + \alpha E_X^{HF} + (1-\beta) E_C^{GGA} + \beta E_C^{MP2}$

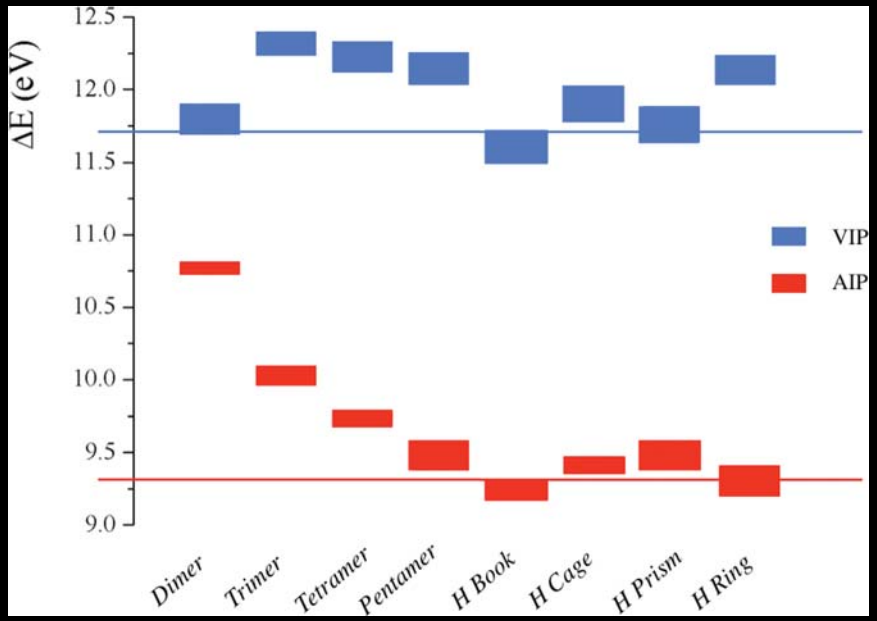
Data



THE JOURNAL OF CHEMICAL PHYSICS **136**, 244306 (2012)

***Ab initio* determination of the ionization potentials of water clusters $(H_2O)_n$ ($n = 2-6$)**

Javier Segarra-Martí,[1] Manuela Merchán,[1,a)] and Daniel Roca-Sanjuán[2,b)]
[1]*Instituto de Ciencia Molecular, Universitat de València, P.O. Box 22085, 46071 València, Spain*
[2]*Department of Chemistry – Ångström, Theoretical Chemistry Program, Uppsala University, Box 518, 75120 Uppsala, Sweden*

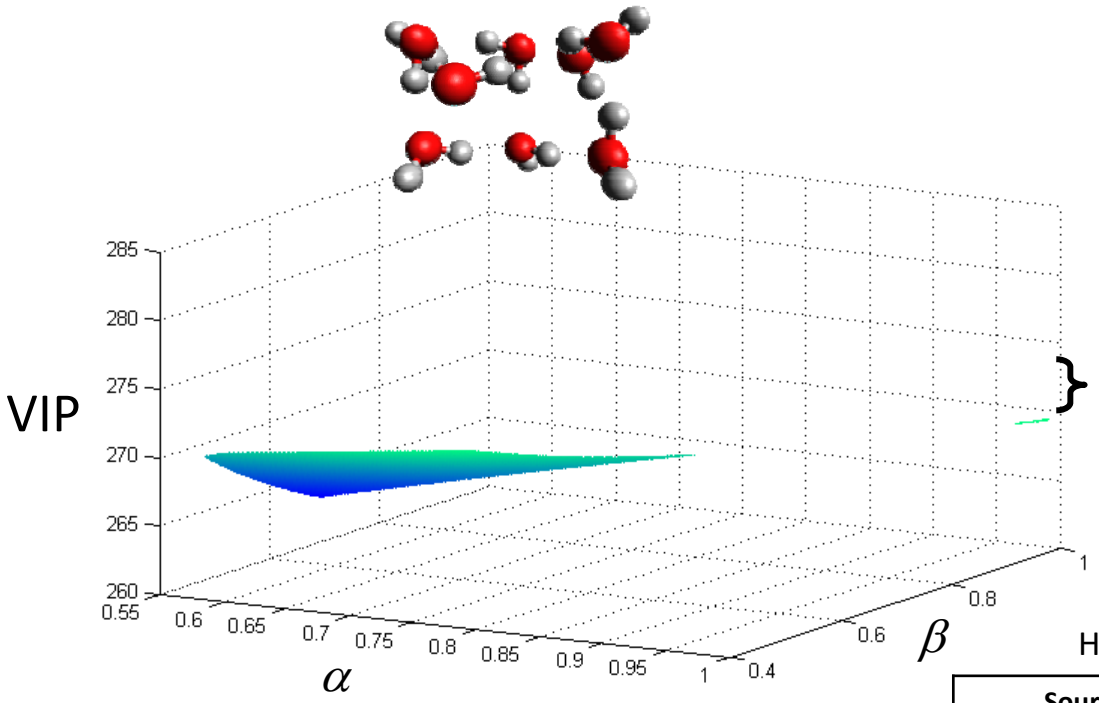


Solve for:

use ΔE intervals computed for dimer, trimer, tetramer, and pentamer

to predict ΔE interval of hexamer

$$E_{XC} = (1-\alpha)E_X^{GGA} + \alpha E_X^{HF} + (1-\beta)E_C^{GGA} + \beta E_C^{MP2}$$



high-level theory result (or experiment)

Combined Feasible Set

Hexamer prediction (kcal/mol):

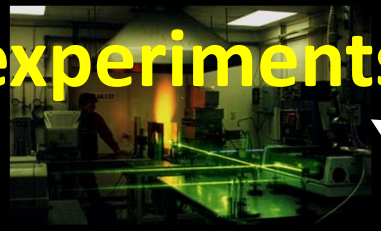
Source	Min	Max	Range
Over Feasible Set	**267.7**	**269.7**	**2.0**
Segarra-Marti et al.	265.9	270.0	4.1

FULL-BLOWN CASE:

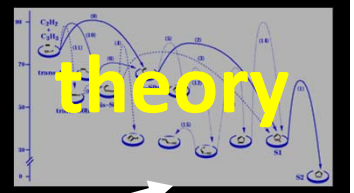
COMBUSTION CHEMISTRY OF NATURAL GAS

- mixture of mostly methane with other light gases
- lowest emissions among fossil fuels; no soot; smallest carbon footprint
- various, expanding sources (biofuels, artificial synthesis,…)
- plenty and cheap; booming US (and world) economy
- technology issues/needs
  - varying compositions – hard to categorize empirically
  - prediction needs: emissions, combustion efficiency, …

Methane Combustion: $CH_4 + 2\ O_2 \Rightarrow CO_2 + 2\ H_2O$



experiments

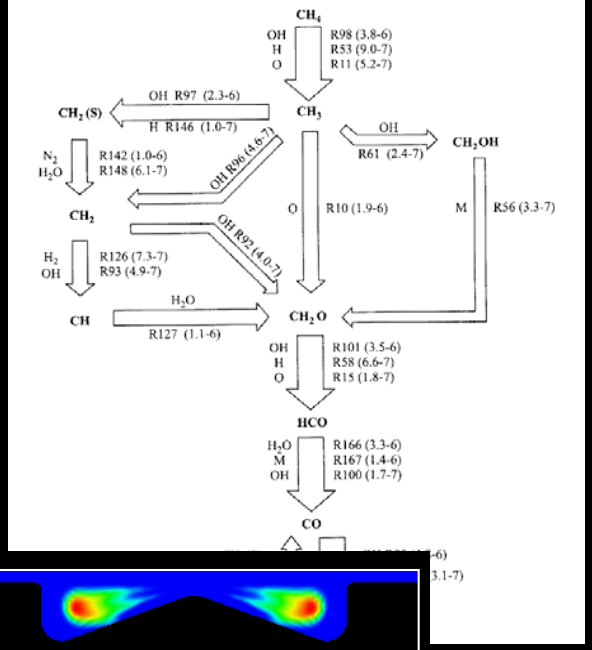


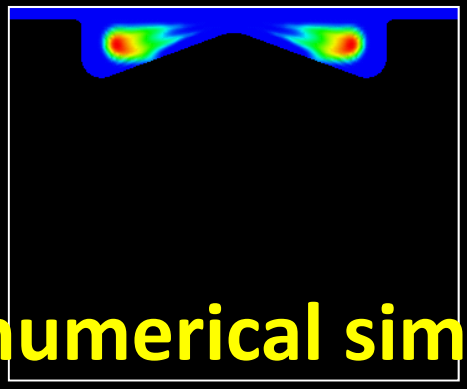
theory

Foundation

- A physically-based model
- The network is complex, but the governing equations (rate laws) are known
- Uncertainty exists, but much is known where the uncertainty lies (rate parameters)
- Numerical simulations with parameters fixed to certain values may be performed “reliably”
- There is an accumulating experimental portfolio on the system
- The model is reduced in size for applications

300+ reactions, 50+ species





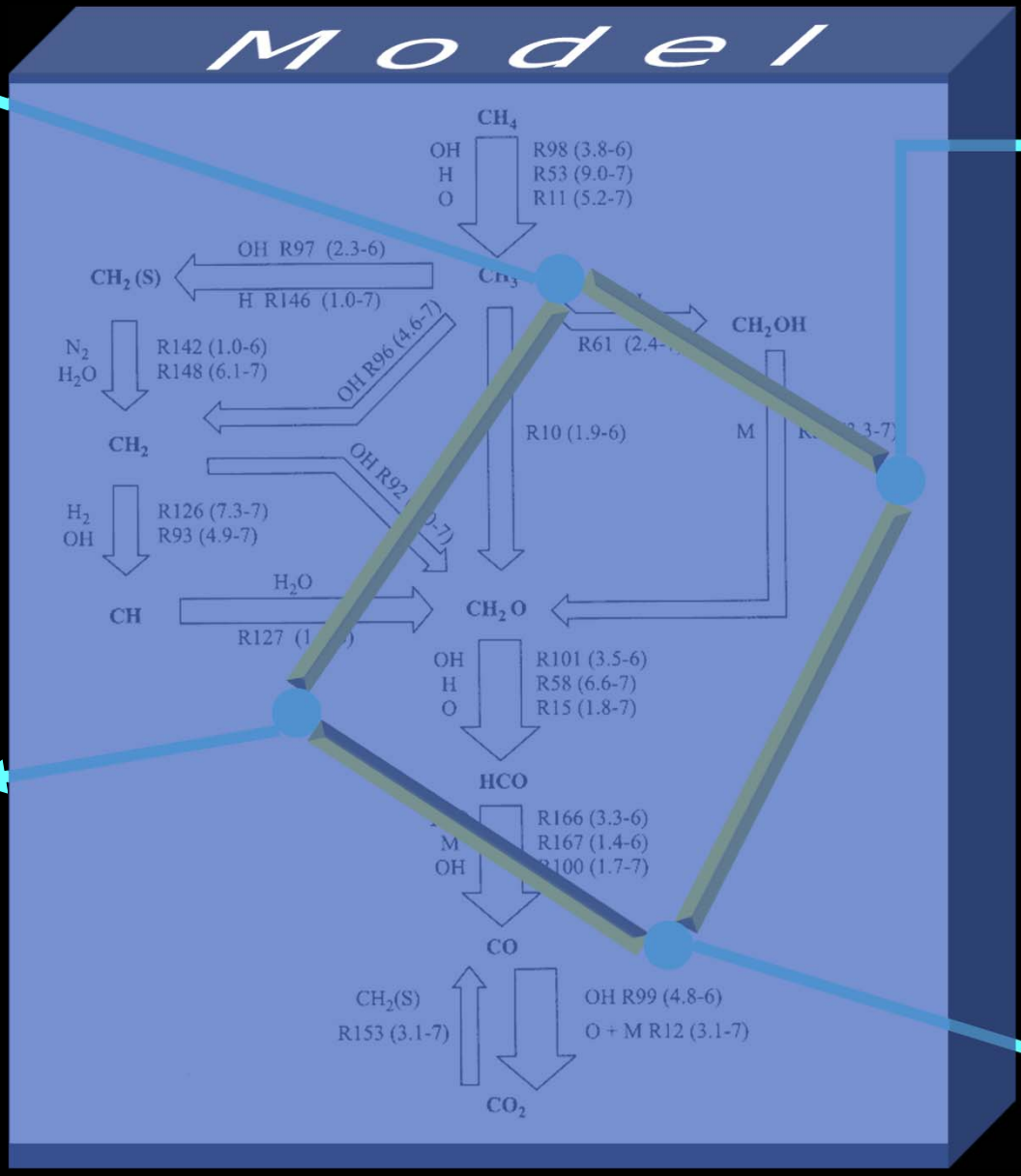
numerical simulations

PREDICTION:
ignition delay
in HCCI engine

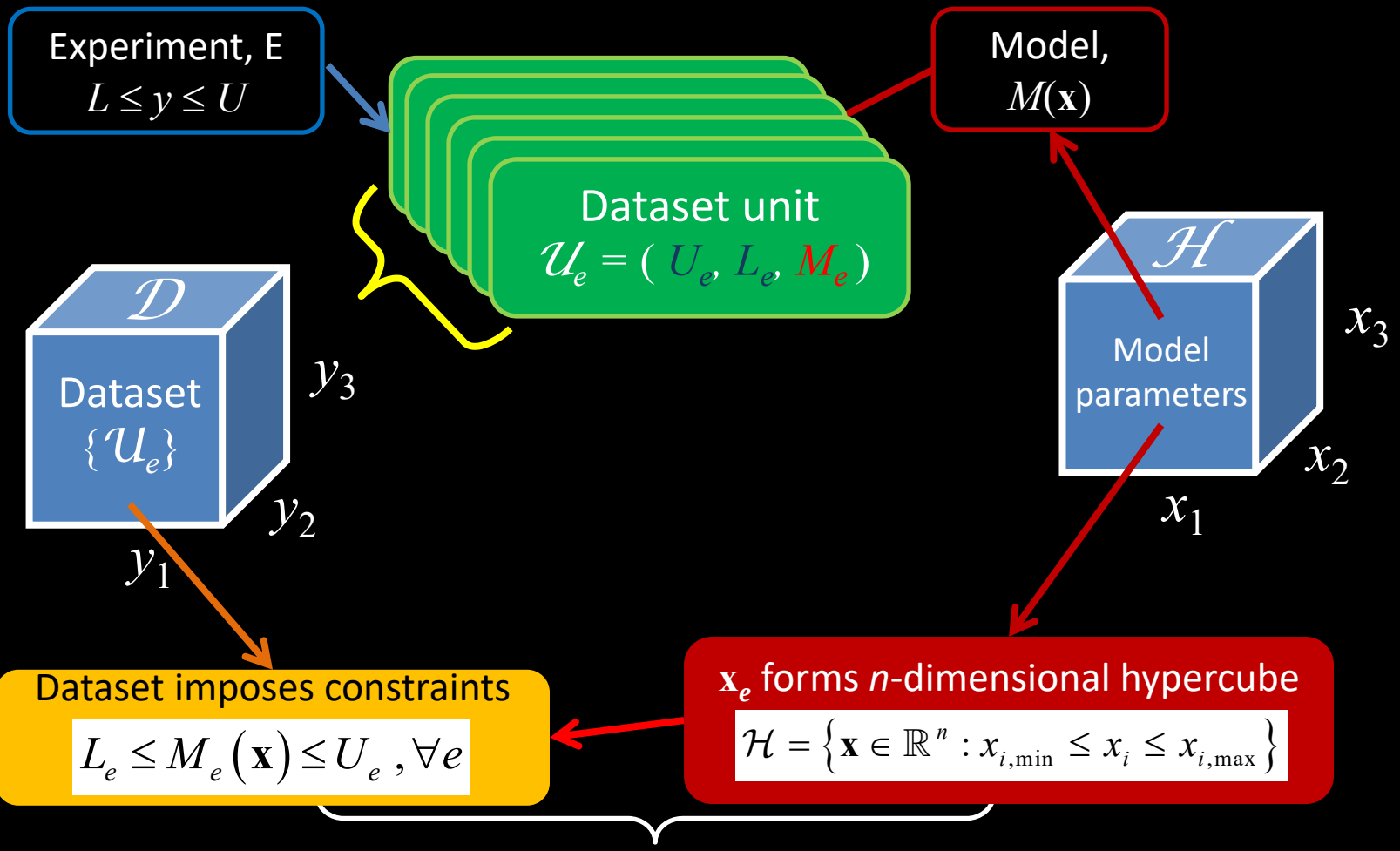
theoretical
rate constants

flow-reactor
measurements

laboratory flame
measurements



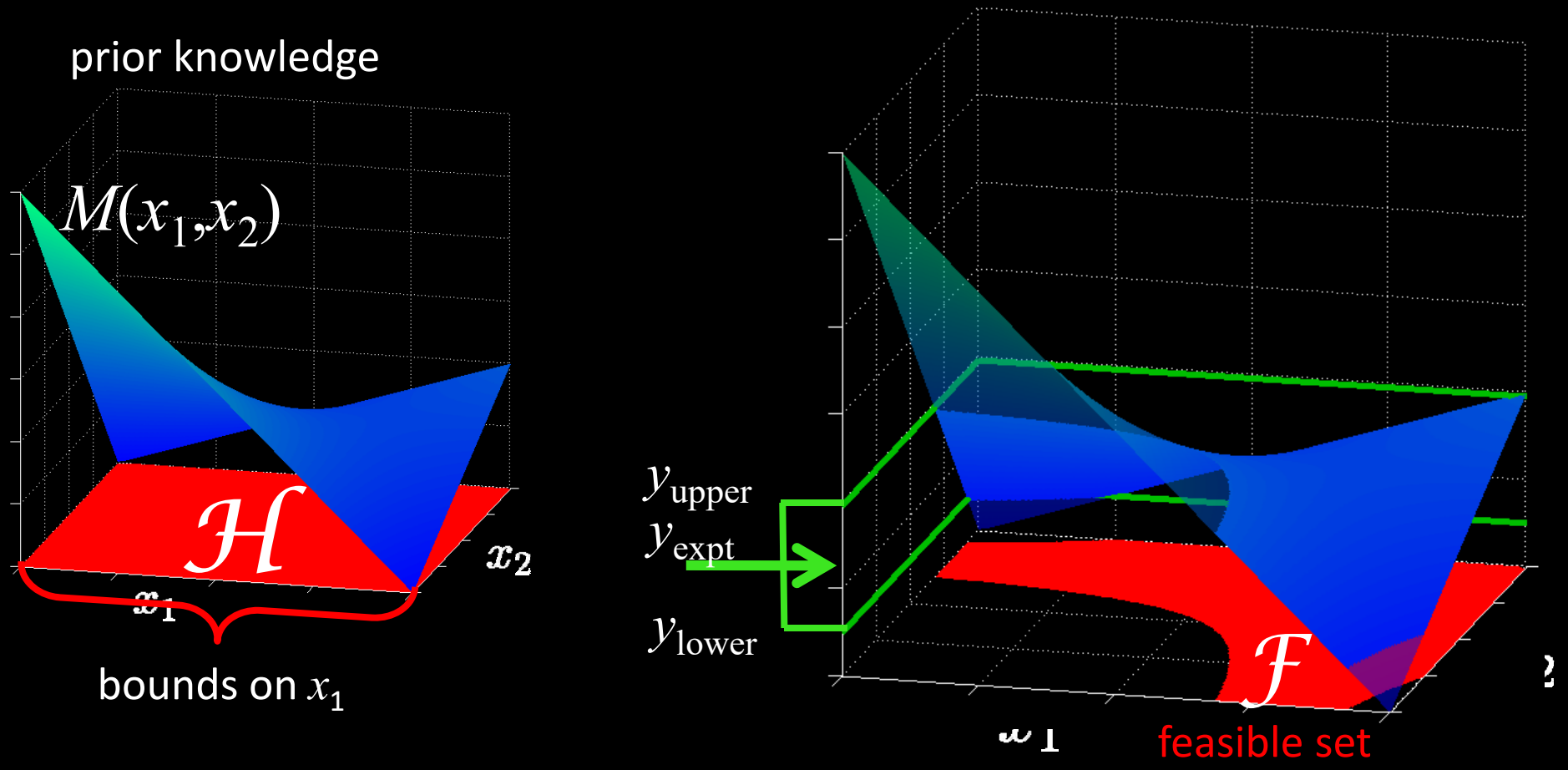
B2B-DATA COLLABORATION



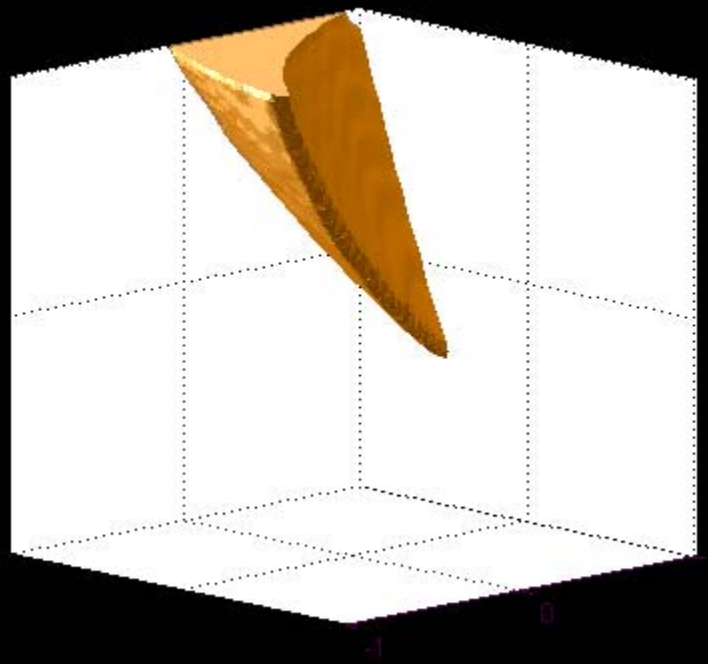
Feasible set of $\mathbf{x}$, $\mathcal{F}$

If empty, *inconsistent*, otherwise, *consistent*

experiment/theory constrain feasible set



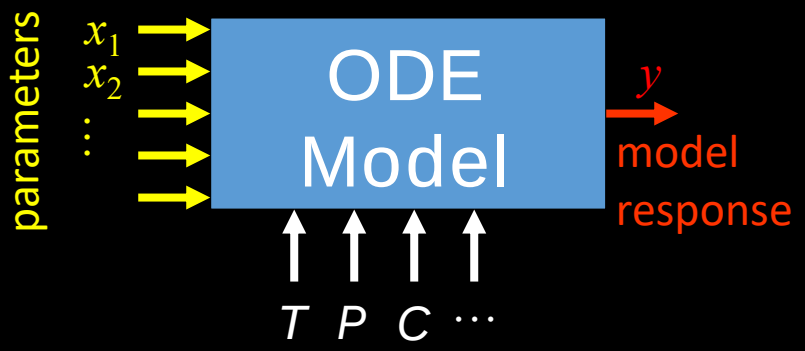
a realistic feasible set:



a set of individual uncertainties does not represent the true compound uncertainty

SURROGATE MODELS (EMULATORS)

- build surrogate models for *individual responses* y (rather than for overall objective)
- construct global objective from individual responses (higher fidelity)



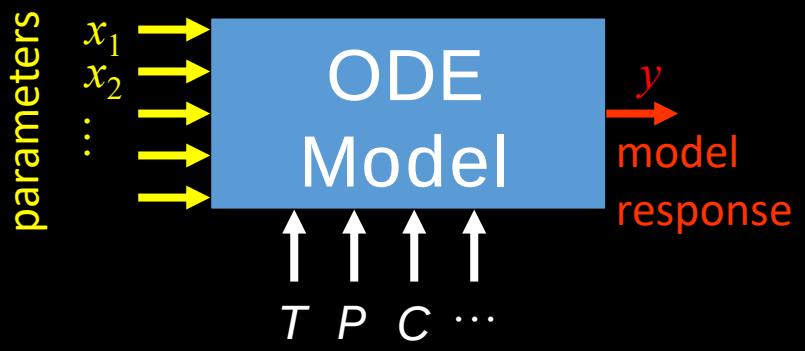
$$\Phi = \sum_{all\ responses} w\left(y_{computed} - y_{observed}\right)^2 \rightarrow \min_x$$

$$y(\{x\}) \approx a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_{1,2} x_1 x_2 + \ldots + a_{1,1} x_1^2 + a_{2,2} x_2^2 + \ldots$$

surrogate model

Surrogate Models (Emulators)

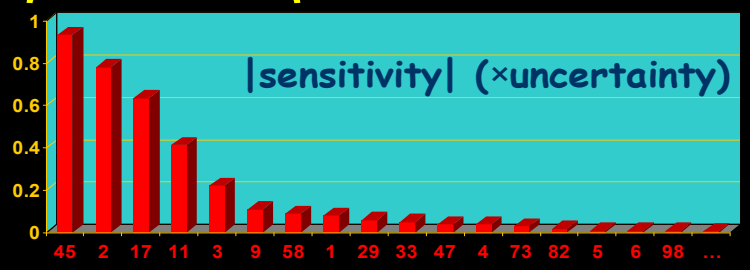
- build surrogate models for individual responses
  (rather than for overall objective)
- construct global objective from individual responses (higher fidelity)



$$\Phi = \sum_{\text{all responses}} w\left(y_{computed} - y_{observed}\right)^2 \rightarrow \min_x$$

dimensionality reduction

active variables

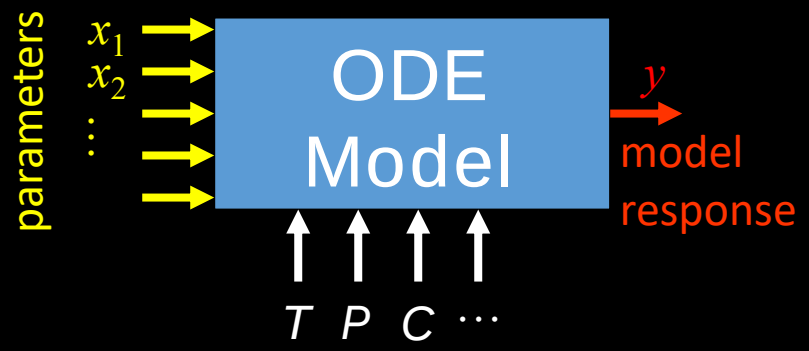


$$y(\{x\}) \approx a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_{1,2} x_1 x_2 + \ldots + a_{1,1} x_1^2 + a_{2,2} x_2^2 + \ldots$$

surrogate model

Surrogate Models (Emulators)

- build surrogate models for individual responses
  (rather than for overall objective)
- construct global objective from individual responses (higher fidelity)



$$\Phi = \sum_{all\ responses} w\left(y_{computed} - y_{observed}\right)^2 \rightarrow \min_x$$

dimensionality reduction

dimensionality of individual response

flame speed $= P_2(x_1, x_2, x_7, x_{23}, \ldots)$

• • •

ignition delay $= P_2(x_1, x_4, x_5, x_{17}, \ldots)$

• • •

species conc $= P_2(x_3, x_4, x_7, x_{12}, \ldots)$

• • •

dimensionality of optimization

Bound-to-Bound UQ Methodology

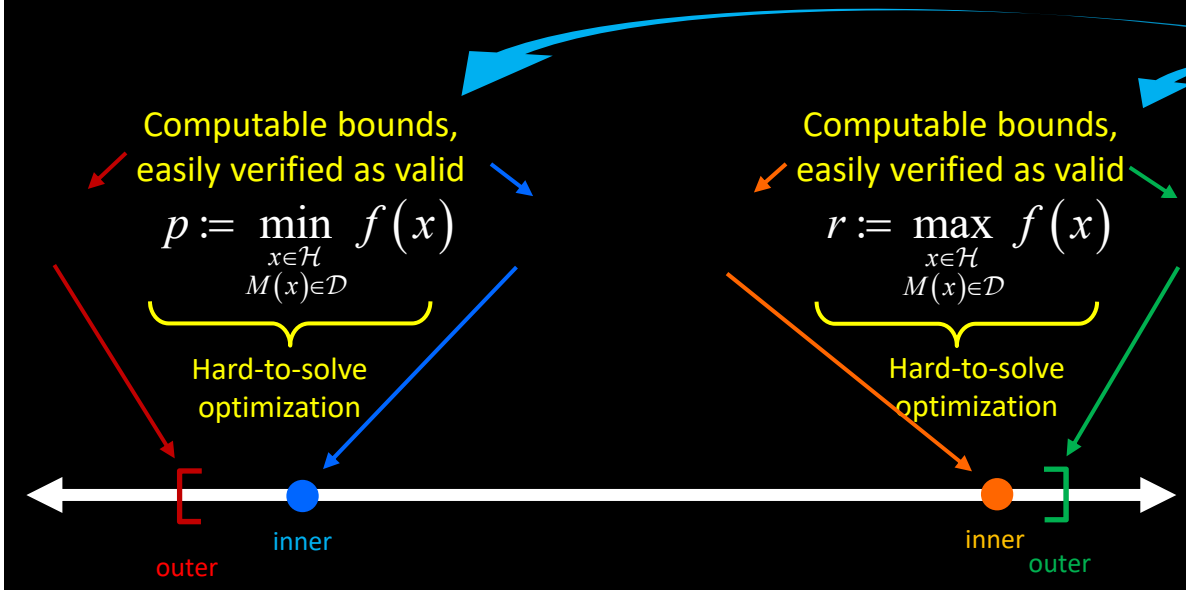
Uncertainty is constrained by:

- prior knowledge of parameters, $x \in \mathcal{H}$, the "$\mathcal{H}$ cube"
- observed data/models, $M(x) \in \mathcal{D}$, the "$\mathcal{D}$ cube"

Prediction model: $f(x)$

–establish possible range of $f(x)$, constrained by

$$\left[\min_{\substack{x \in \mathcal{H} \\ M(x) \in \mathcal{D}}} f(x) \quad \max_{\substack{x \in \mathcal{H} \\ M(x) \in \mathcal{D}}} f(x) \right]$$



If f and M are *quadratic*, then the **min** and **max** problems → SDP and p's and r's bounds are

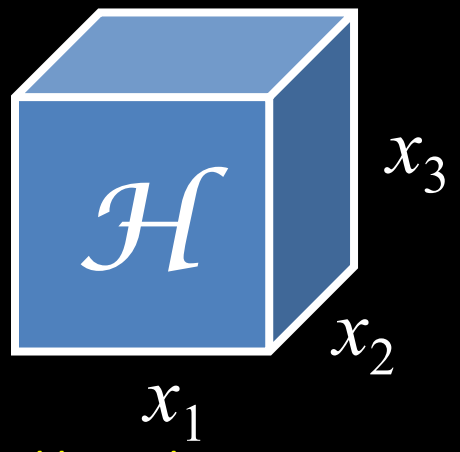
- computable
- easily verified as valid
- same for their global sensitivities

A dataset is ***consistent*** if the *Feasible Set* is nonempty;

i.e., there exists a parameter vector that satisfies:

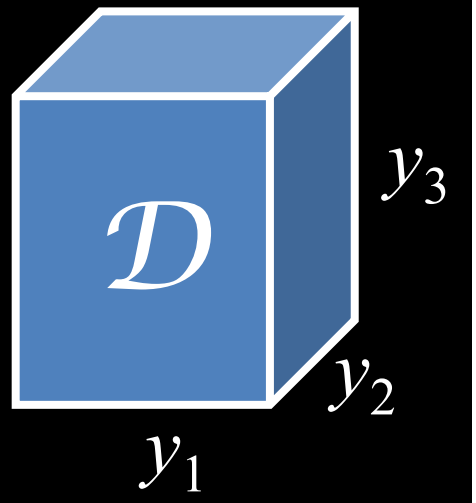
- all parameters are within *prior bounds*, $\mathcal{H}$

$$x_{1,\min} \le x_1 \le x_{1,\max}$$
$$x_{2,\min} \le x_2 \le x_{2,\max}$$
$$\cdots$$



- all model predictions are within *experimental bounds*

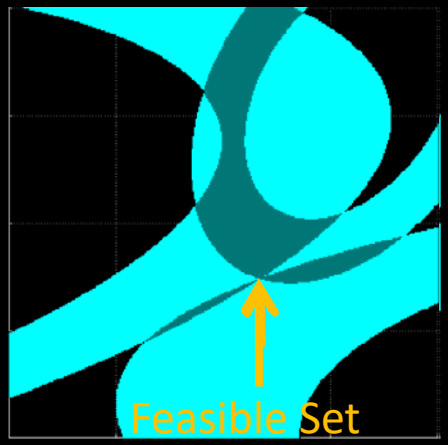
$$L_e \le M_e(\mathbf{x}) \le U_e$$

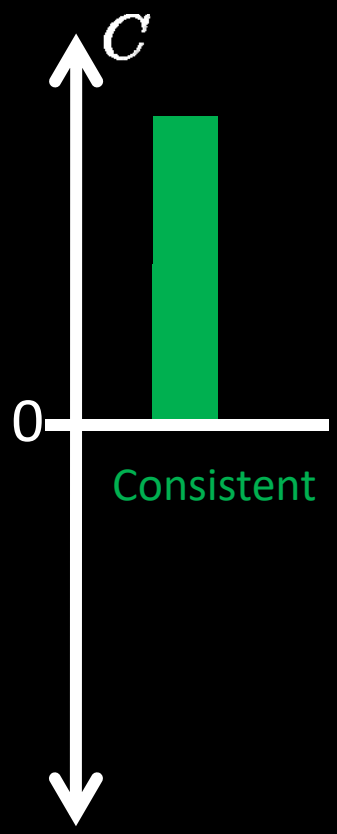


- numerical measure of consistency

$$C_D = \max_{\mathbf{x} \in H} \gamma$$
$$L_e(1-\gamma) \le M_e(\mathbf{x}) \le U_e(1-\gamma), \quad \forall e$$

Consistency Measure

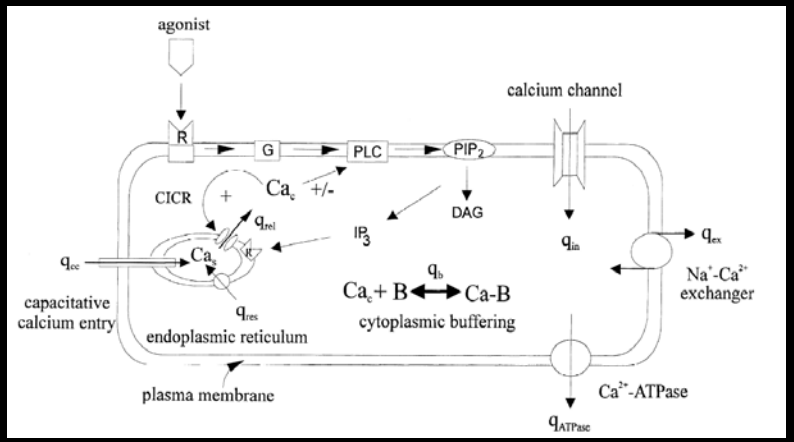




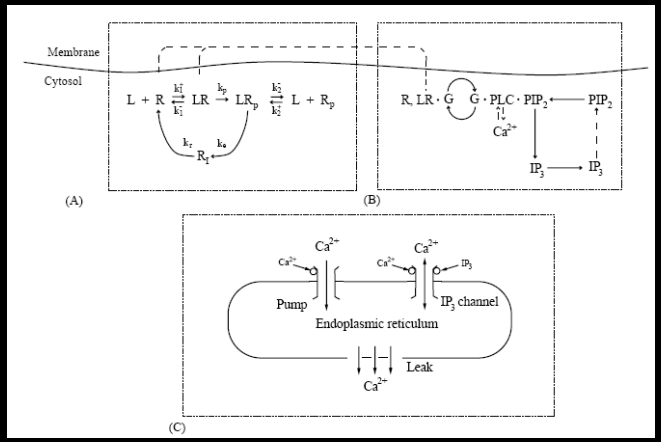
J. Phys. Chem. A 108:9573 (2004)

DISCRIMINATION AMONG MODELS

Wiesner et al. 1996
27 active variables



Lemon et al. 2003
34 active variables



consistency measure

0

+0.24

–0.03

inconsistent

consistent

J. Phys. Chem. A 110:6803 (2006)

SENSITIVITY COEFFICIENTS

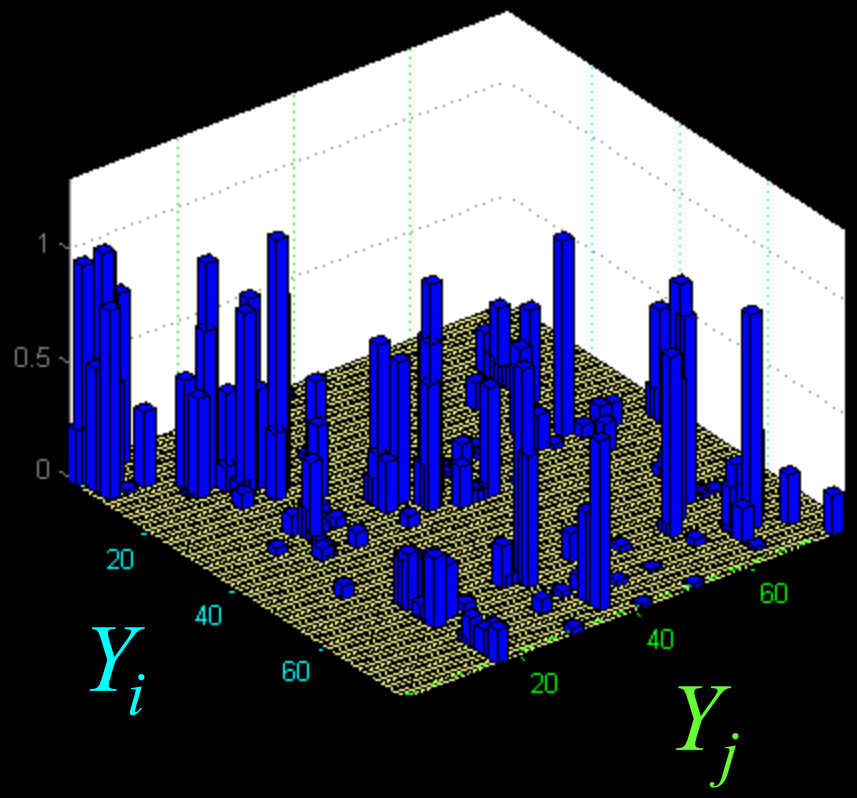
$$\frac{\partial\left(\dfrac{prediction}{prediction}\right)}{\partial\left(\dfrac{parameter}{parameter}\right)}$$

$$\frac{\partial\left(\dfrac{prediction}{prediction}\right)}{\partial\left(\dfrac{experiment}{uncertainty}\right)}$$

$$\lambda := \frac{\partial\left(\begin{array}{c}\textit{prediction}\\ \textit{interval}\end{array}\right)}{\partial\left(\begin{array}{c}\textit{experiment}\\ \textit{uncertainty}\end{array}\right)} = \frac{1}{2}\left(\frac{\partial \overline{M}}{\partial U} - \frac{\partial \overline{M}}{\partial L} + \frac{\partial \underline{M}}{\partial U} - \frac{\partial \underline{M}}{\partial L}\right)$$

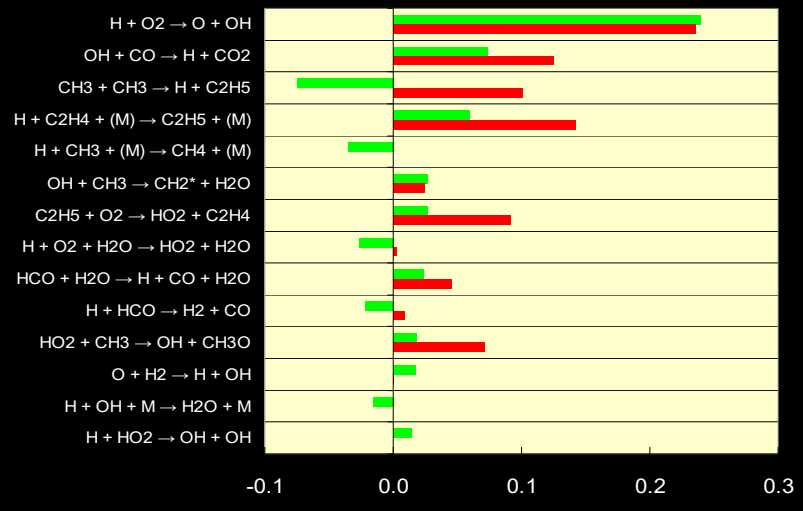
$$\nu := \frac{\partial\left(\begin{array}{c}\textit{prediction}\\ \textit{interval}\end{array}\right)}{\partial\left(\begin{array}{c}\textit{parameter}\\ \textit{uncertainty}\end{array}\right)} = \frac{1}{2}\left(\frac{\partial \overline{M}}{\partial x_{\max}} - \frac{\partial \overline{M}}{\partial x_{\min}} + \frac{\partial \underline{M}}{\partial x_{\max}} - \frac{\partial \underline{M}}{\partial x_{\min}}\right)$$

Sensitivity $\dfrac{\text{of uncertainty in predicting } Y_i}{\text{to uncertainty in observing } Y_j}$



sensitivity

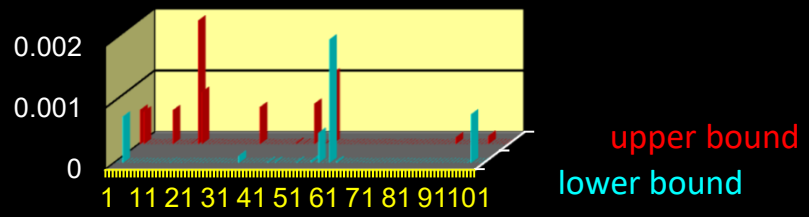
"w.r.t. value" vs "w.r.t. uncertainty"



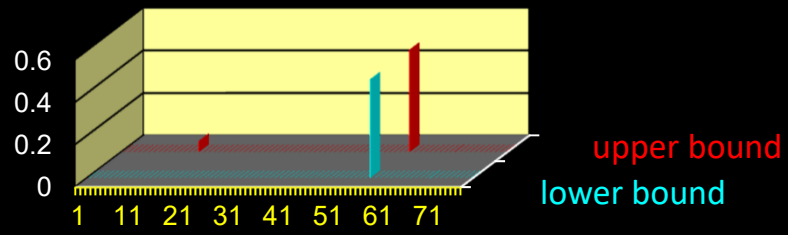
laminar flame speed in a stoichiometric atmospheric C_2H_6-air mixture

sensitivity of methane dataset *consistency*

to *uncertainty* in model parameters

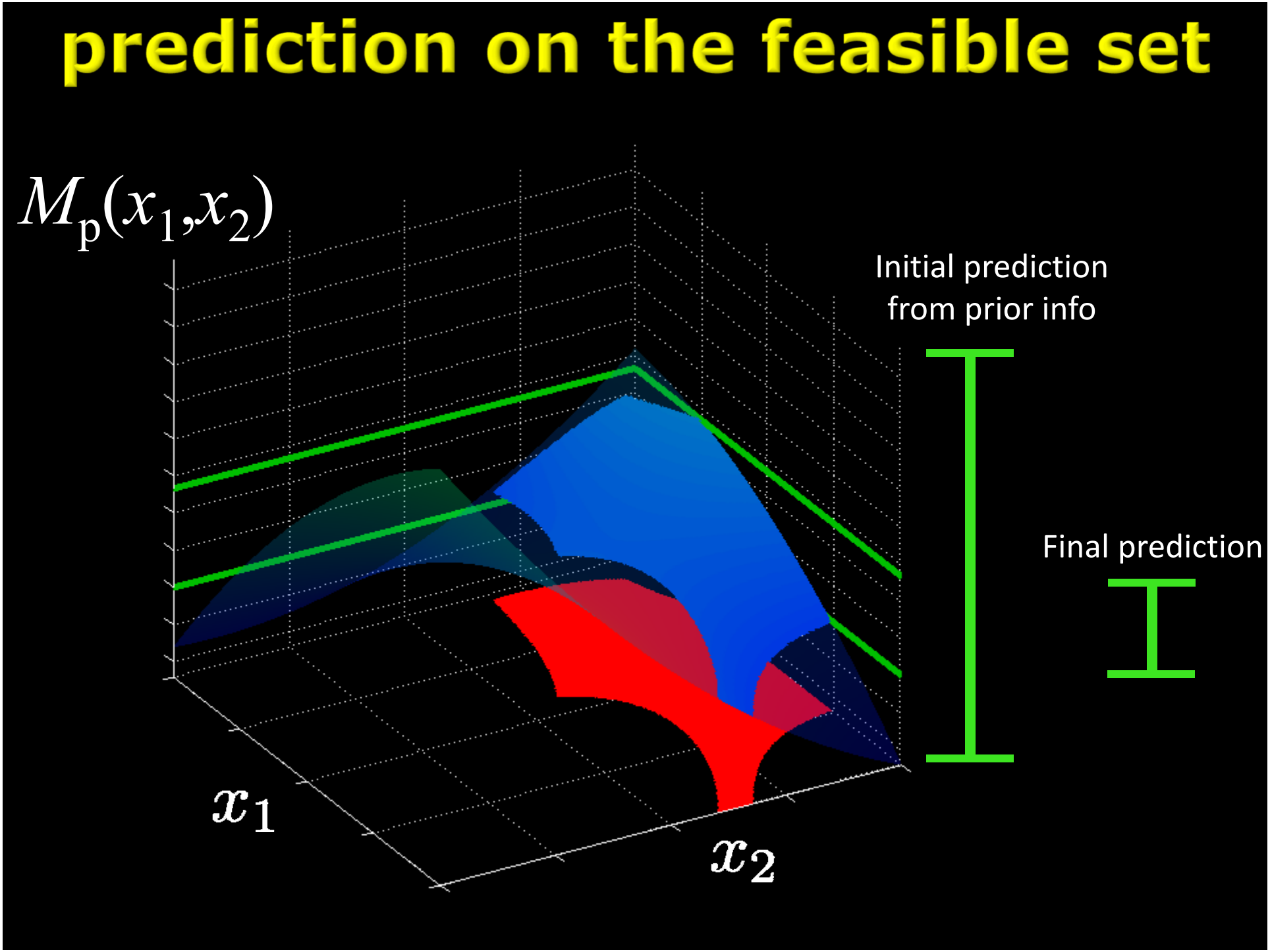


to *uncertainty* in experimental observations



J. Phys. Chem. A 112:2579 (2008)

prediction on the feasible set



prediction interval

is the range of values M_p takes over the set of feasible values of parameters

$$= \overline{M}_p(\mathbf{x}) - \underline{M}_p(\mathbf{x})$$

$$\overline{M}_p(\mathbf{x}) = \max_{\mathbf{x}} M_p(\mathbf{x})$$

$$\underline{M}_p(\mathbf{x}) = \min_{\mathbf{x}} M_p(\mathbf{x})$$

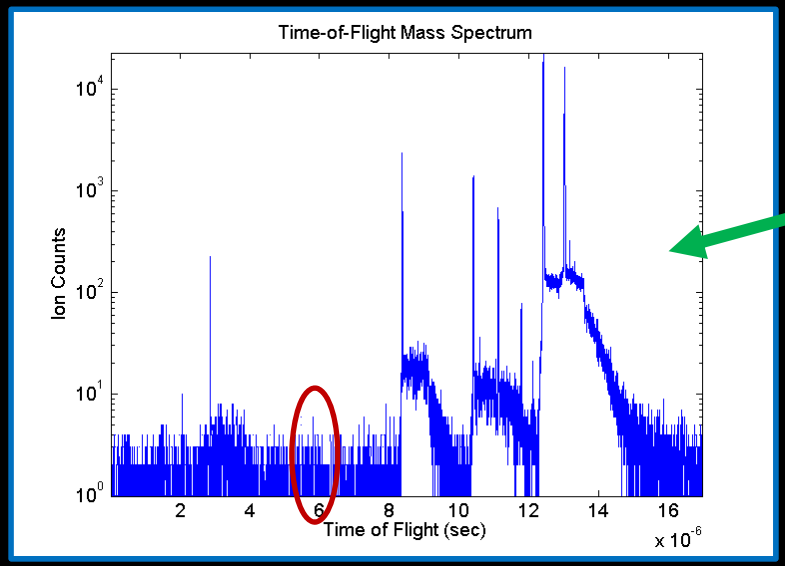
subject to : $\mathbf{x} \in \mathcal{H}$

$L_e \le M_e(\mathbf{x}_e) - a_e \le U_e, \quad \forall e$

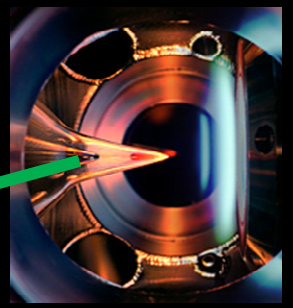
feasible set

Combining kinetic and instrumental models, B2B-DC predicts noisy/weak signals

Prediction Feature	Prediction Interval
O Peak Value	$[2.7, 4.3] \times 10^{-2}$
O Peak Location	[1.9, 2.2] cm
OH Peak Value	$[3.0, 3.6] \times 10^{-2}$
OH Peak Location	[1.60, 1.67] cm
C_2H_3 Peak Value	$[0.09, 1.15] \times 10^{-4}$
C_2H_3 Peak Location	$[0.6, \ 3.9 \ ] \times 10^{-2}$ cm



Advance Light Source Flame Experiments

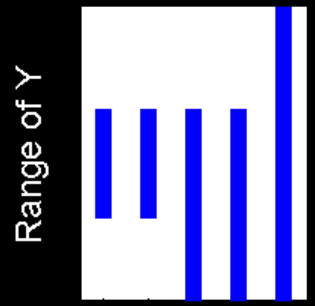


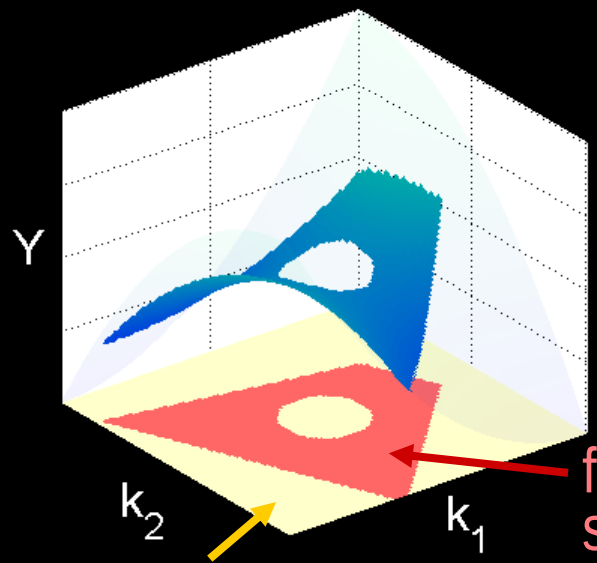
Information Content of an Experiment

Int. J. Chem. Kinet. 36:57 (2004)

Including Experimental Observations

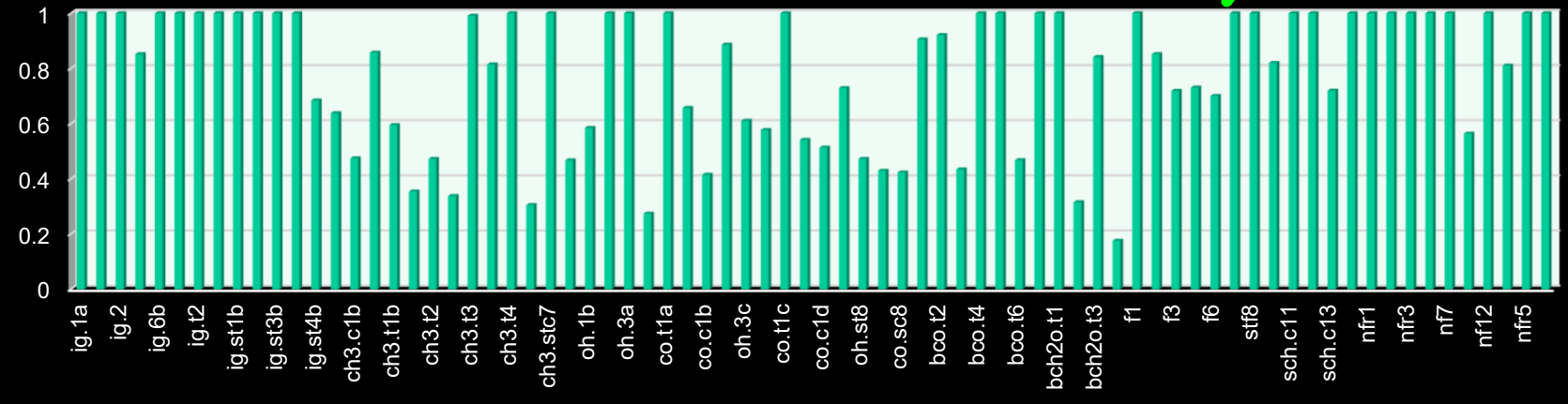
Only prior knowledge



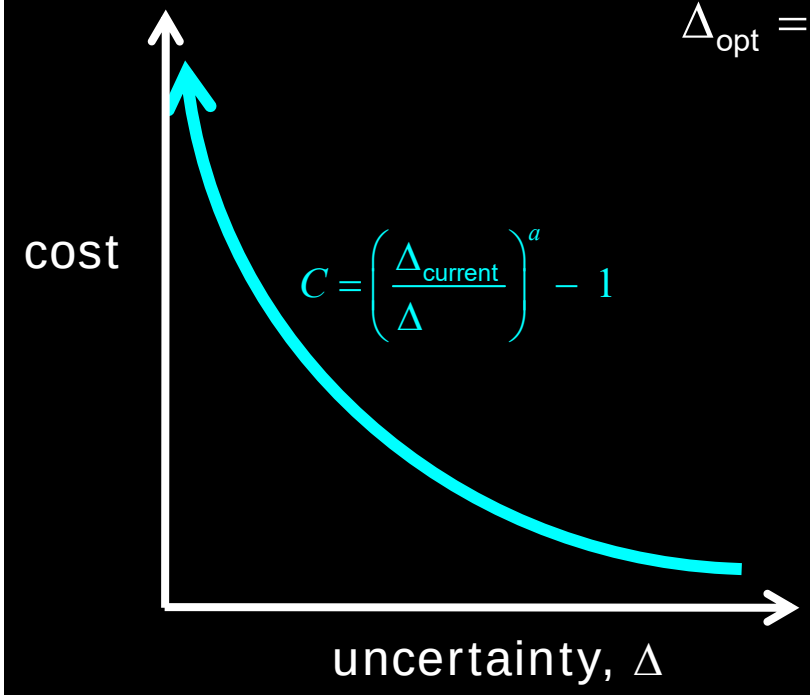


Information Gain:

$$I = 1 - \frac{\text{Posterior Range}}{\text{Prior Range}}$$

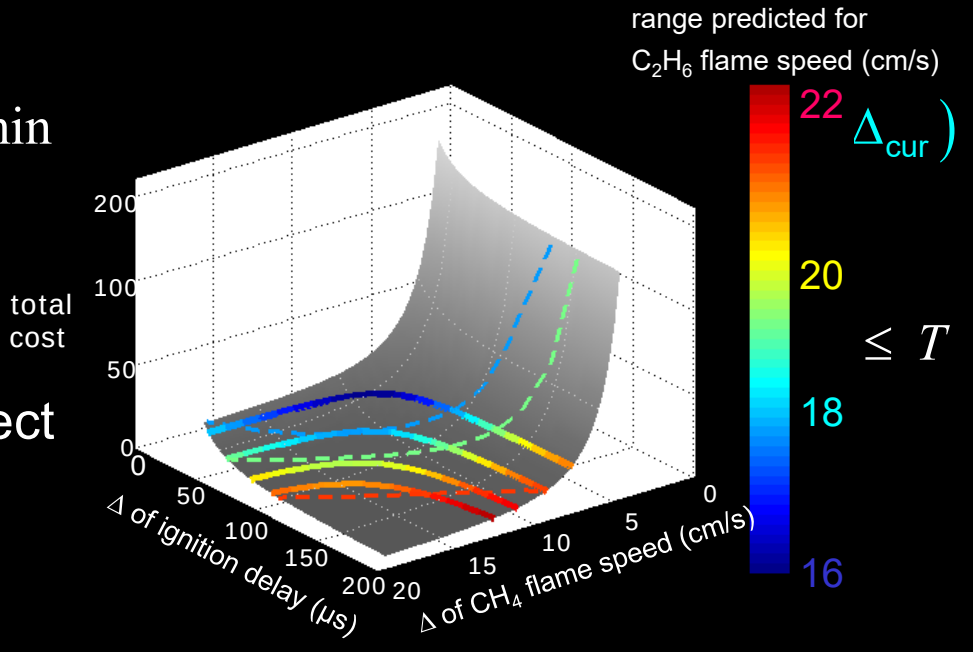


Given a budget T, determine the best strategy for reducing the uncertainty in model prediction



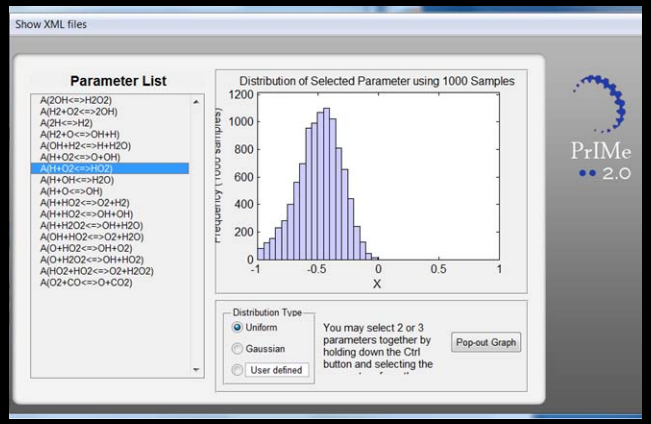
$$\Delta_{\text{opt}} = \arg\min_{\Delta} \left(\Delta_{\text{cur}} \right)$$

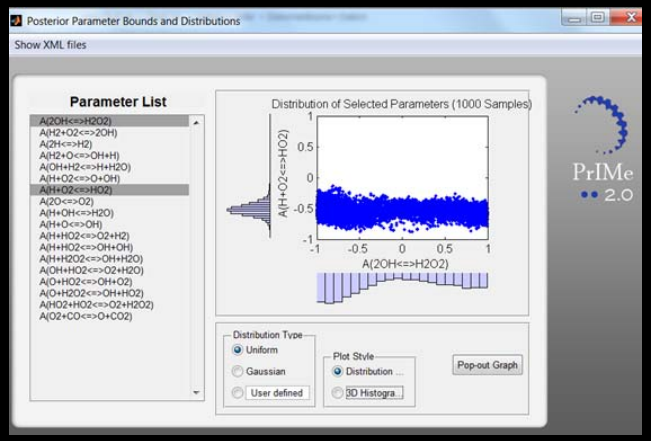
subject to $\leq T$

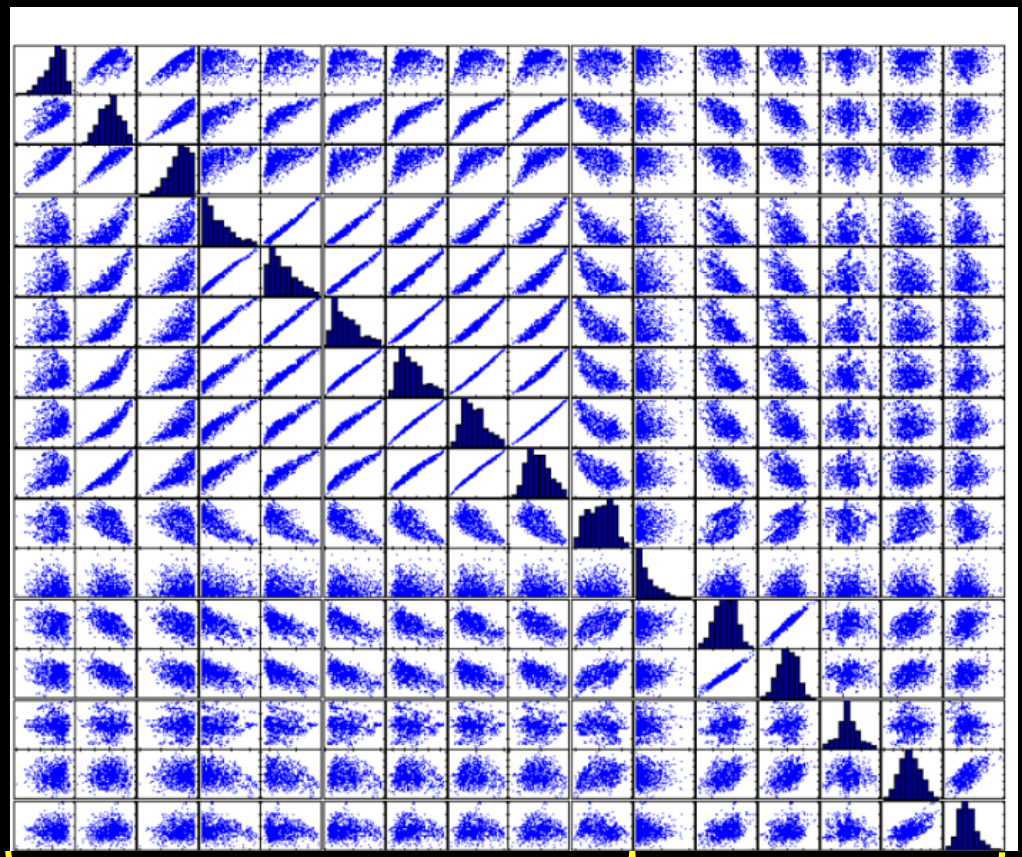


J. Phys. Chem. A 112:2579 (2008)

Analysis of parameter distributions, predictions, and uncertainty correlations by sampling the feasible set





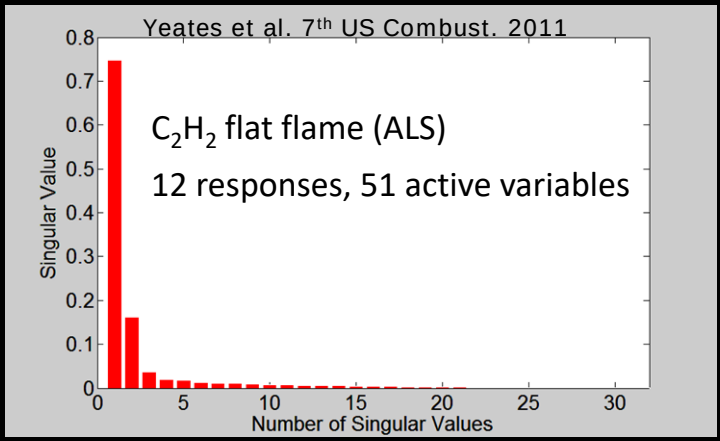


Slavinskaya, et al., 2013, 2014

Discovery of Active Subspace of Active Variables through Sampling of the Feasible Set

While a $M(\mathbf{x})$ formally depends on all n active variables, in reality it mostly vary in $r \ll n$ linear combinations of the variables.

For creating a surrogate of $M(\mathbf{x})$ we would like to do the design in the r-dimensional space.



Factor $f(\mathbf{x}) = g\left(S^T\mathbf{x}\right)$

Compute gradients of $f(x)$ at points of $\mathcal{H}$

$$\underbrace{\begin{bmatrix} \nabla f(\mathbf{x}^{(1)}) \\ \vdots \\ \nabla f(\mathbf{x}^{(N)}) \end{bmatrix}}_{F} = \underbrace{\begin{bmatrix} \nabla g\left(S^{\mathrm{T}}\mathbf{x}^{(1)}\right) \\ \vdots \\ \nabla g\left(S^{\mathrm{T}}\mathbf{x}^{(N)}\right) \end{bmatrix}}_{G} \cdot S^{\mathrm{T}}$$

F: $(N \times n)$; G: $(N \times r)$; S^{T}: $(r \times n)$

Perform SVD of F; this gives S

Sample r-subspace of $\mathcal{H}$ to build surrogate design

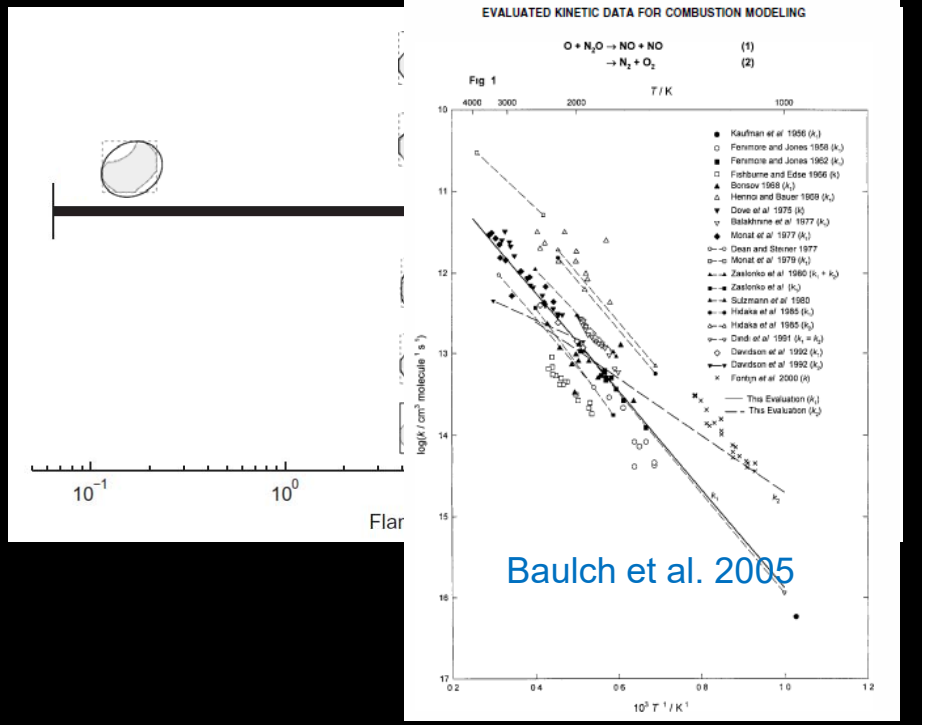
How B2B-DC compares to other methods, as far as approximations are concerned

- Even in case of rigorous Bayesian, use of a prescribed prior (e.g., Gaussian) underestimates the uncertainty in prediction (Phillip Stark, "Constraints versus Priors", 2012)

  AND we unlikely to have Gaussian priors!

- Approximations, even seemingly "harmless", may lead to substantial differences in prediction of uncertainty (Russi et al, *Chem. Phys. Lett.* 2010)

- Optimization-based methods, transferring uncertainty in two-steps — from data to parameters and then from parameters to prediction — necessarily overestimate the predicted uncertainty



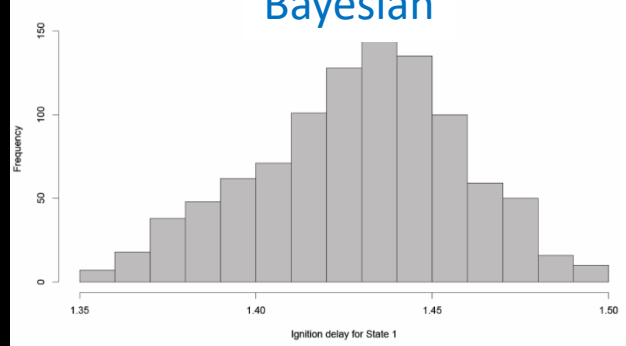
B2B-DC and rigorous Bayesian produce consistent results

An ongoing collaborative study with

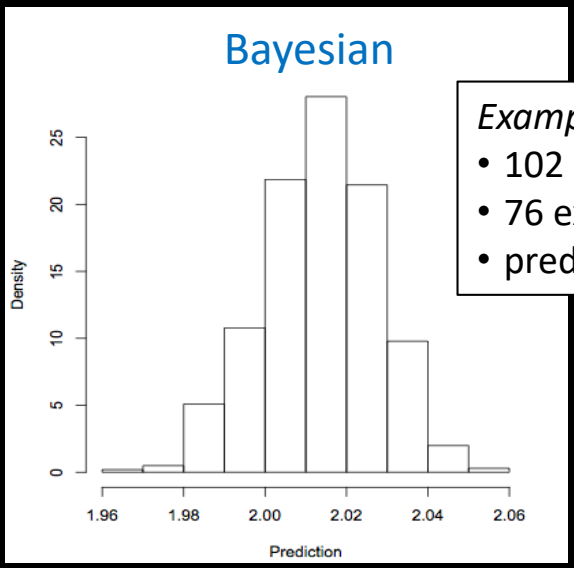
Jerome Sacks, *National Institute of Statistical Sciences*
Rui Paulo, *ISEG Technical University of Lisbon*
Gonzalo Garcia-Donato, *Universidad de Castilla-La Mancha*, Spain

Example: H2/O2
- 21 active variables
- 12 experimental targets
- predicting one "blind"

Bayesian



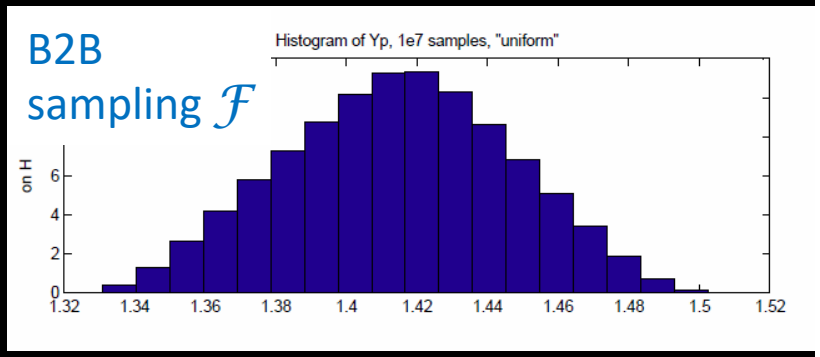
Bayesian



Example: GRI-Mech 3.0
- 102 active variables
- 76 experimental targets
- predicting one "blind"

B2B-DC prediction for this blind target is [1.89 2.12]

B2B sampling $\mathcal{F}$

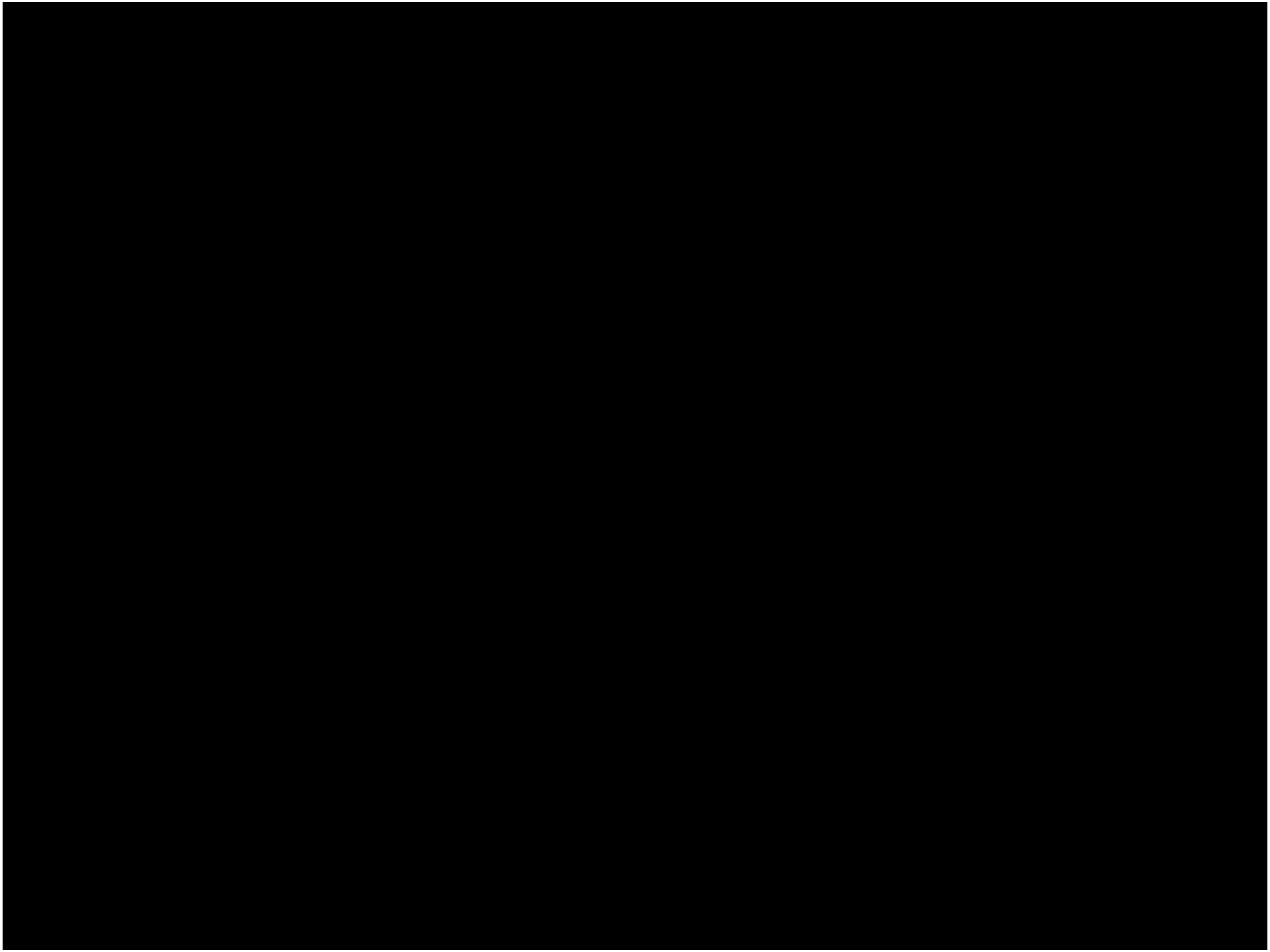


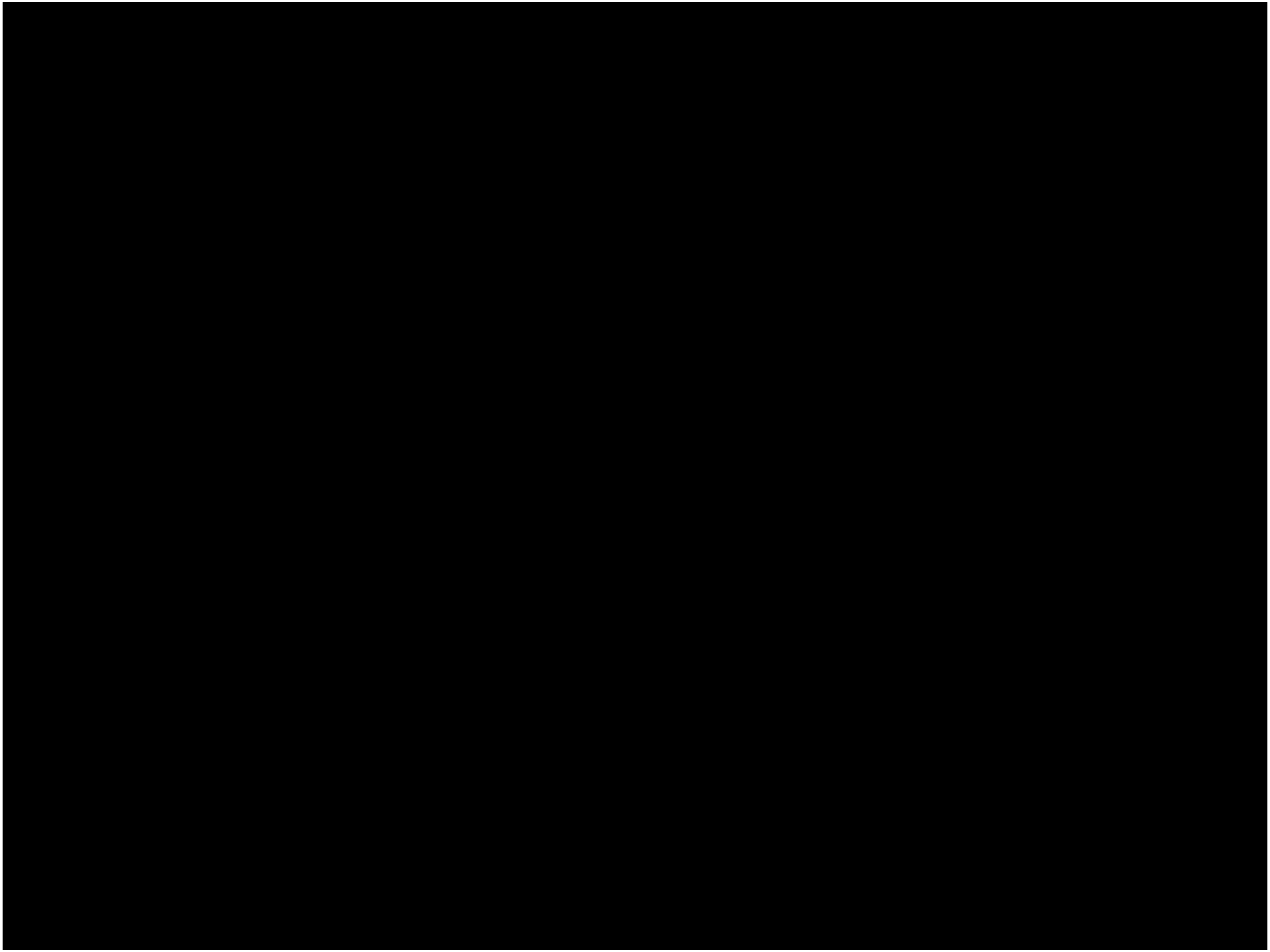
Personal Observations

- current inability of truly predictive modeling
  - conflicting data in/among sources
  - poor documentation of data/models
  - no uncertainty reporting or analysis
  - not much focus on integration of data
- resistance to data sharing
  - no personal incentives
  - no easy-to-use technology
- no recognition of the problem

SUMMARY: B2B-DC

- is mathematically rigorous, numerically efficient, and UQ-rich approach to analysis of practical systems
- is data-centric, handles heterogeneous data, and is easily scalable to a large number of data sets
- is scalable to a large number of parameters through Solution Mapping features, combined with the Active Space Discovery
- establishes a clear measure of consistency among data and models, and identifies the cause of inconsistency if detected
- "measures" information content of an experiment
  - assess an impact of a given or planned experiment ("what if")
  - design new experiments/theory that impact the most
- reduces uncertainties of known and predicts correctly uncertainty of unknown





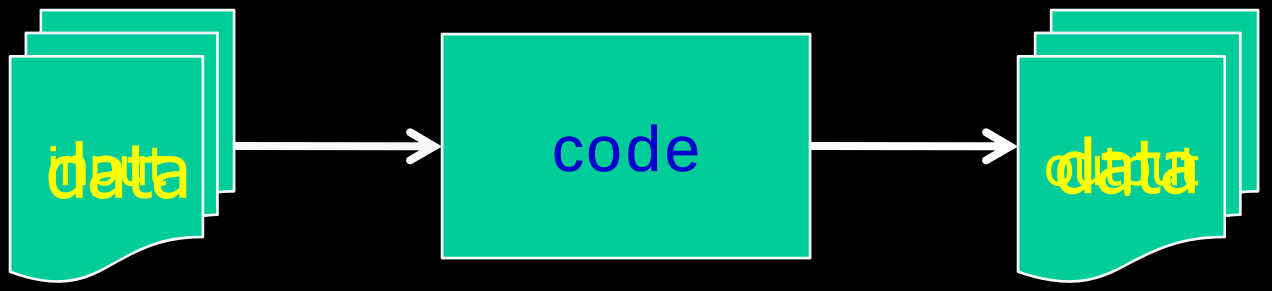
ANSWER QUESTIONS

- What causes/skews model predictiveness?
- Are there new experiments to be performed, old repeated, theoretical studies to be carried out?
- What impact could a planned experiment have?
- What is the information content of the data?
- What would it take to bring a given model to a desired level of accuracy?

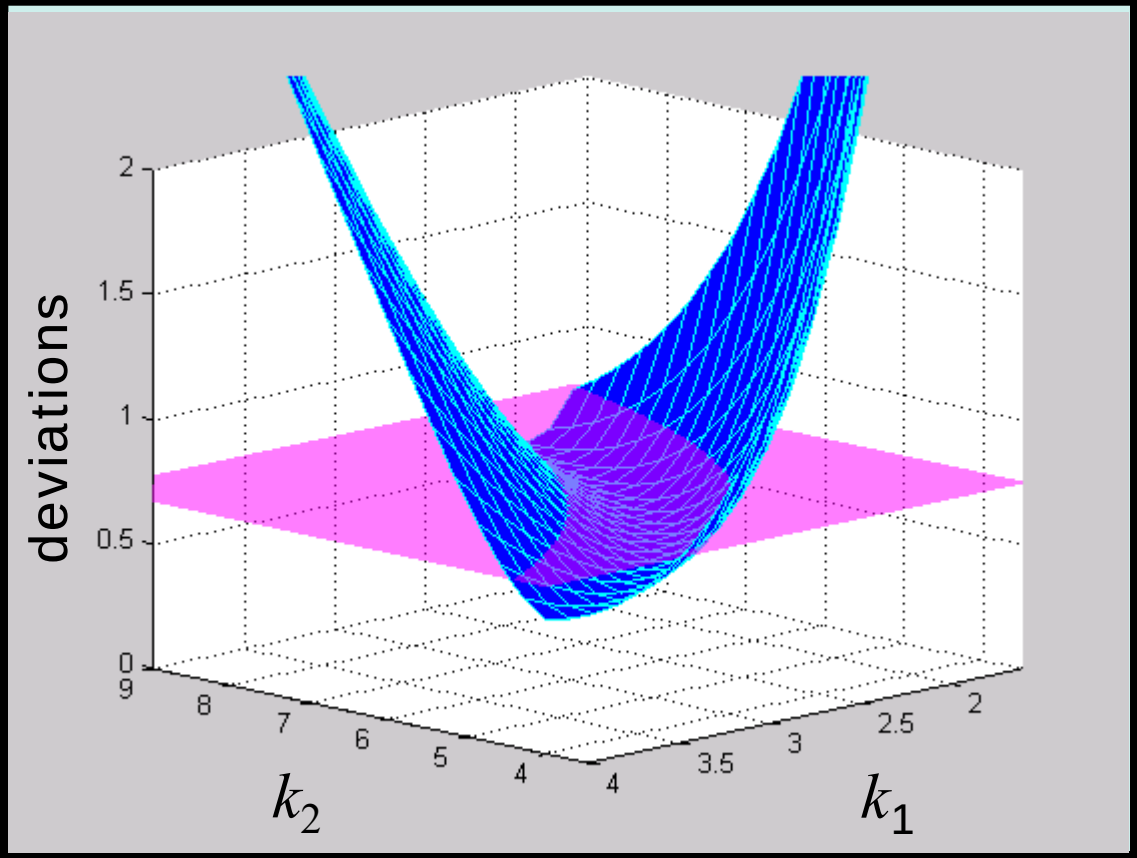
NEED A PARADIGM SHIFT

from *algorithm-centric* view

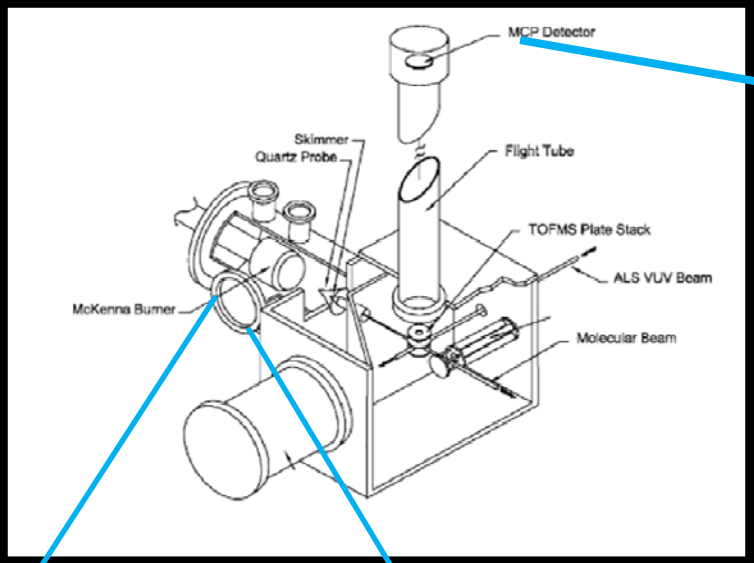
to *data-centric* view

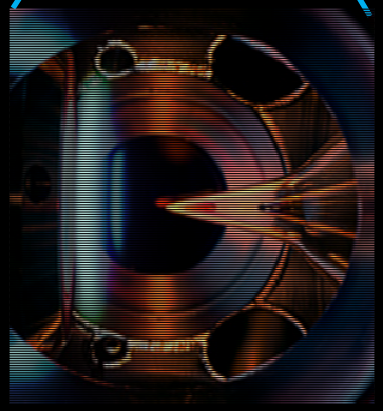


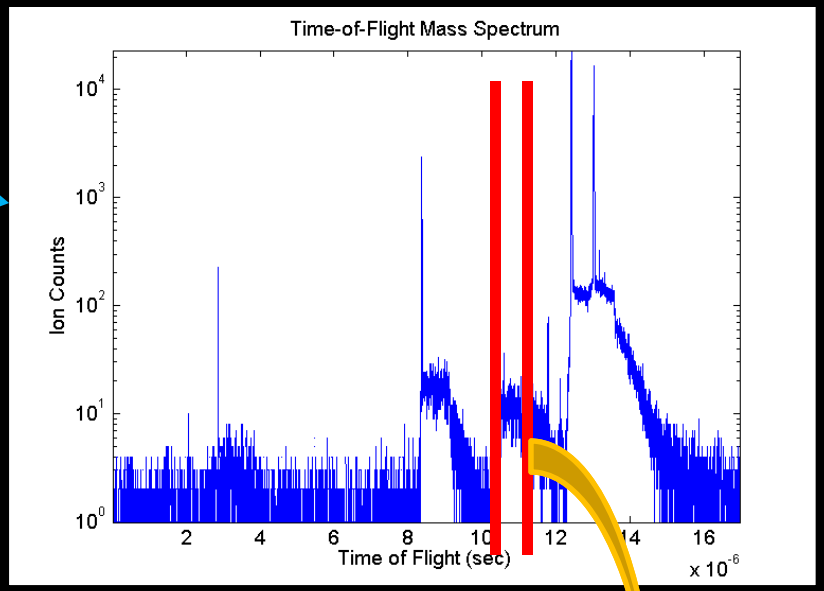
Valley Character of Objective

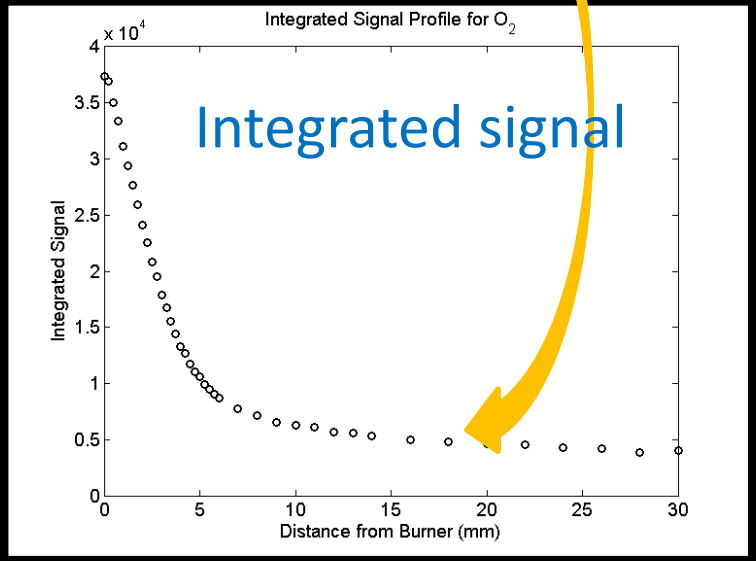


Advance Light Source Flame Experiments



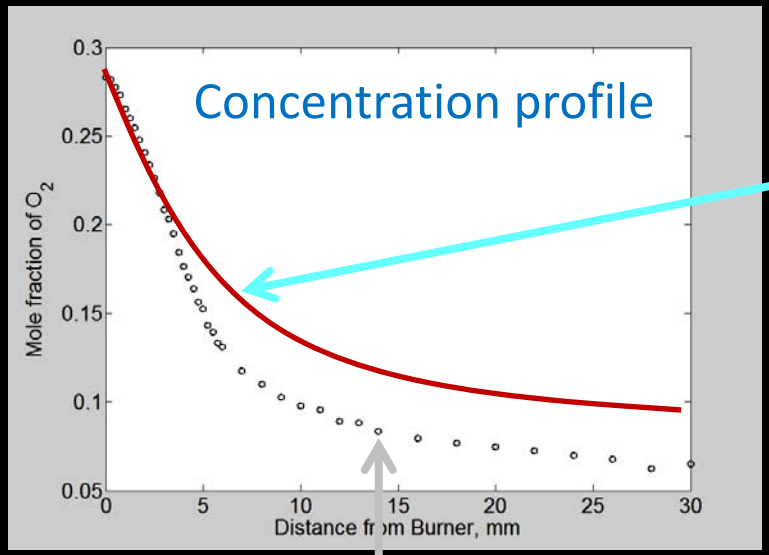






(top) Cool et al., Rev. Sci. Instrum. **76**, 094102 (2005)
(bottom) Courtesy of Sandia CRF - http://www.sandia.gov/ERN/images/CRF-Science.jpg

"Standard" Data Analysis



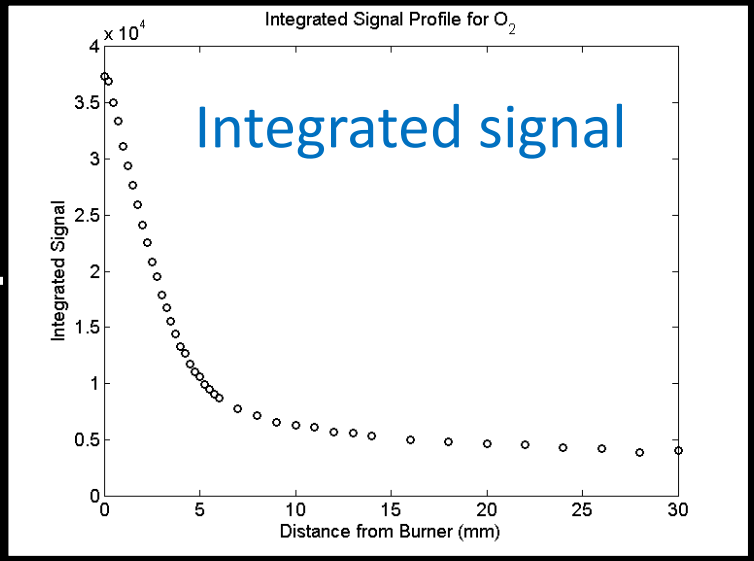
Model

Data Analysis

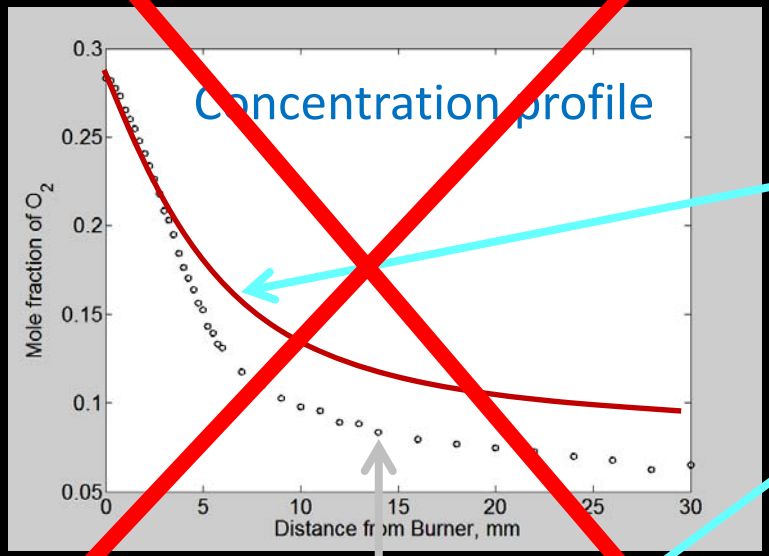
$y = f(x, c)$

Calibration, c

Assumptions



Using Instrumental Model



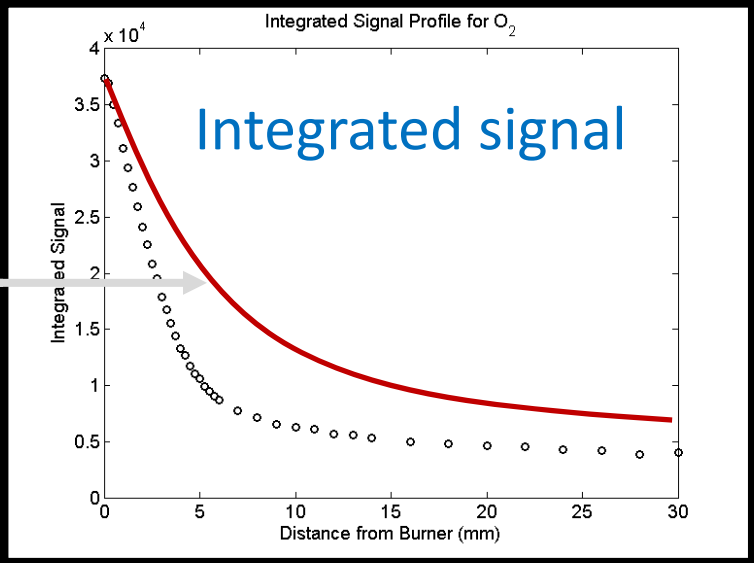
Model

Data Analysis

$y = f(x, c)$

Calibration, c

Assumptions

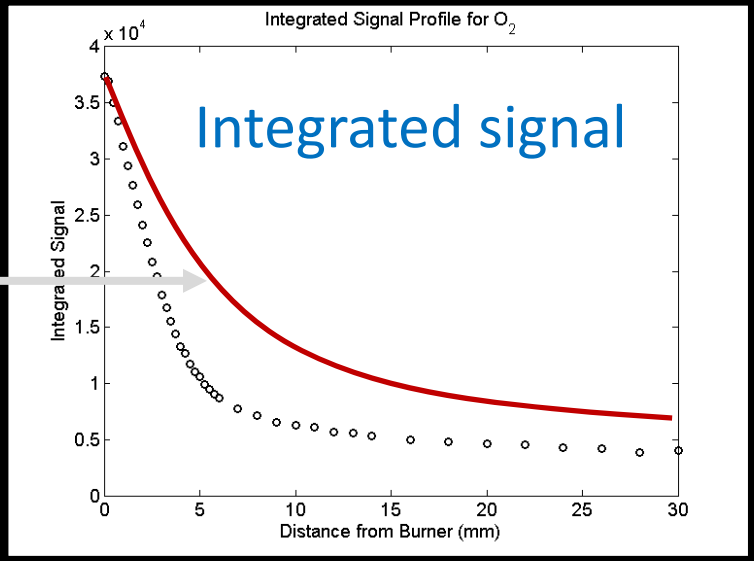


Using Instrumental Model

Model

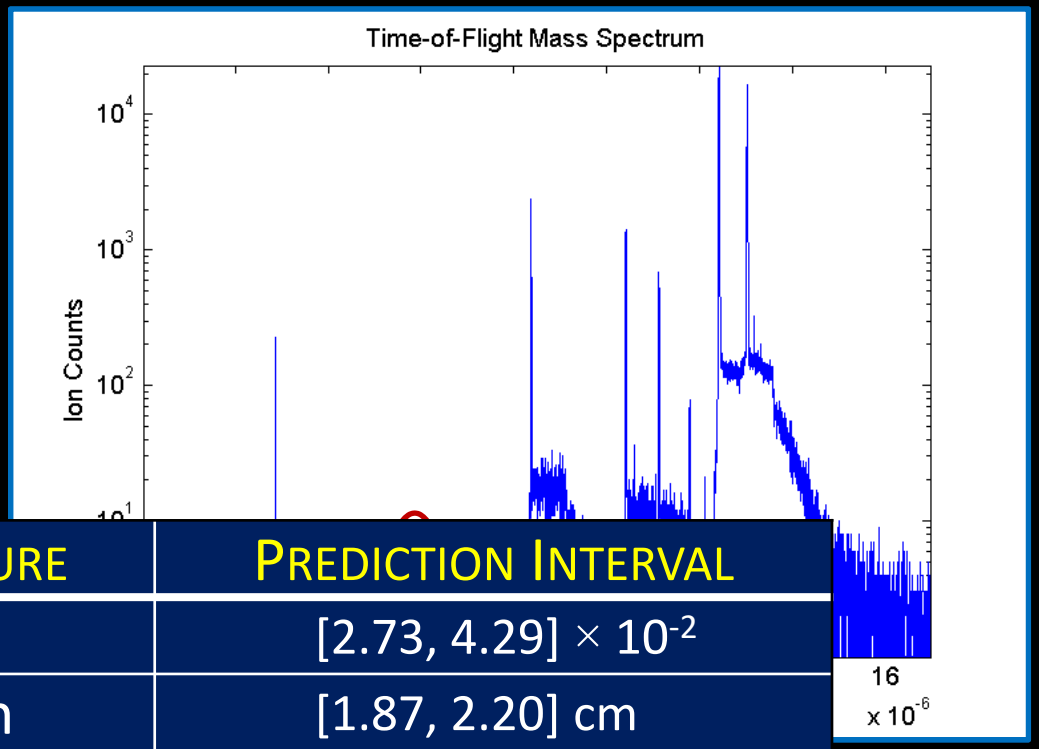
Instrumental Model

Calibration, c



Predicting Weak Signals

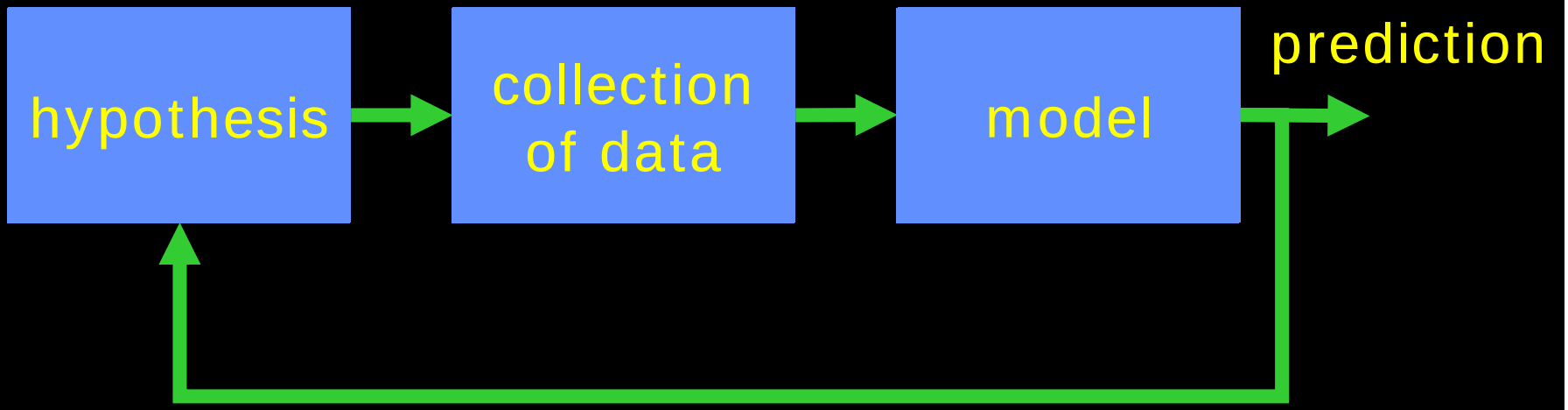
- O, OH, C_2H_3
  - Peak Value
  - Peak Location



Prediction Feature	Prediction Interval
O Peak Value	[2.73, 4.29] × 10^{-2}
O Peak Location	[1.87, 2.20] cm
OH Peak Value	[2.97, 3.59] × 10^{-2}
OH Peak Location	[1.60, 1.67] cm
C_2H_3 Peak Value	[0.09, 1.15] × 10^{-4}
C_2H_3 Peak Location	[0.60, 3.90] × 10^{-2} cm

SCIENTIFIC METHOD

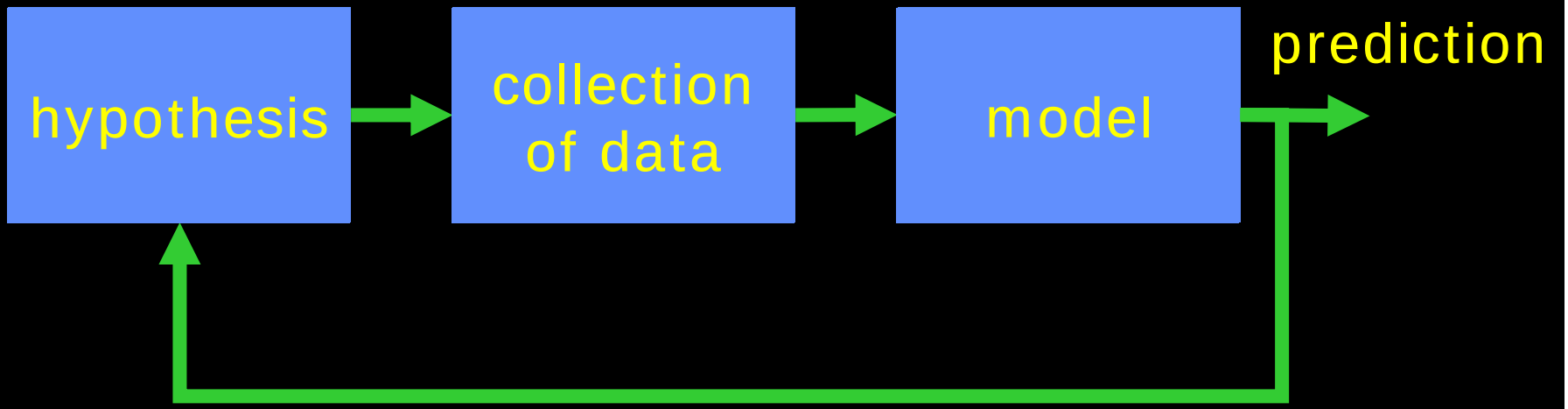
SCIENTIFIC METHOD



SENSITIVITY ANALYSIS:

What conditions will maximize sens to k?

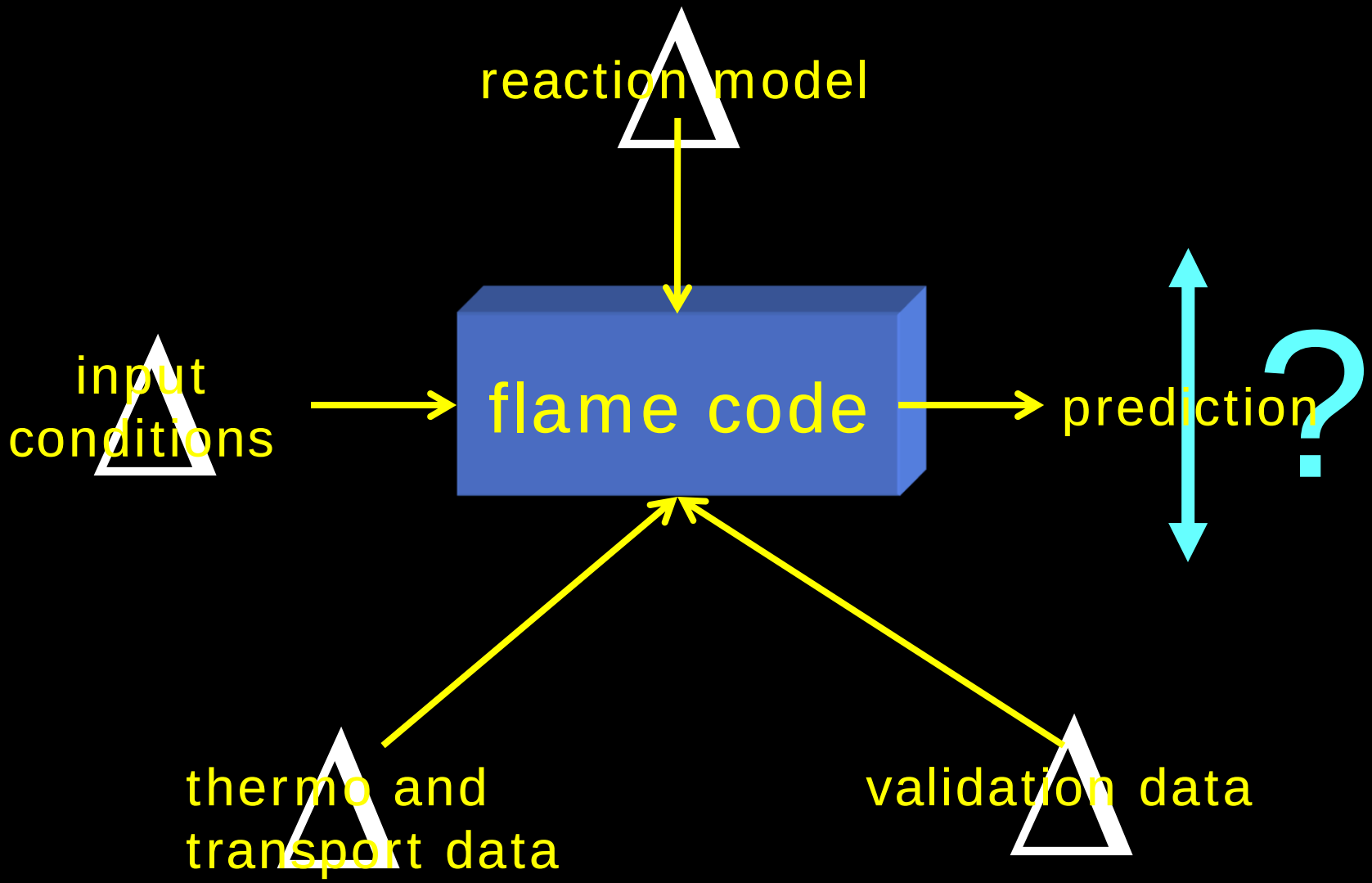
SCIENTIFIC METHOD



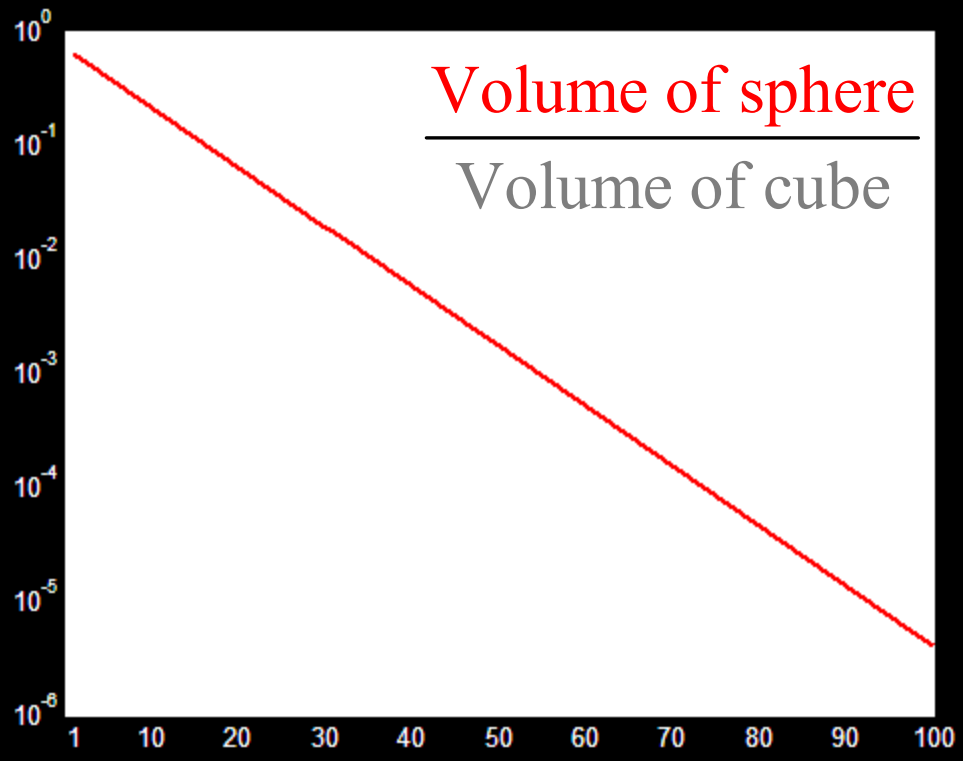
UNCERTAINTY QUANTIFICATION:

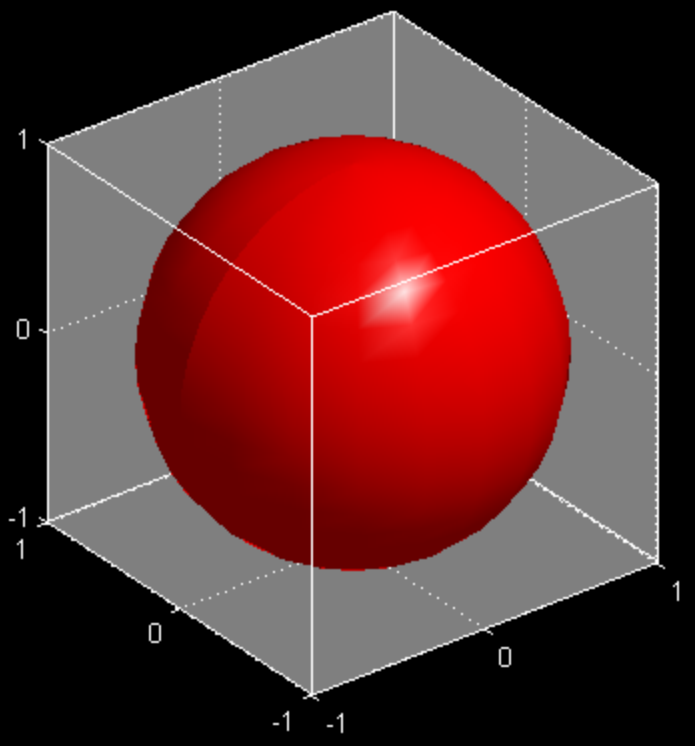
What experiment will be most informative?

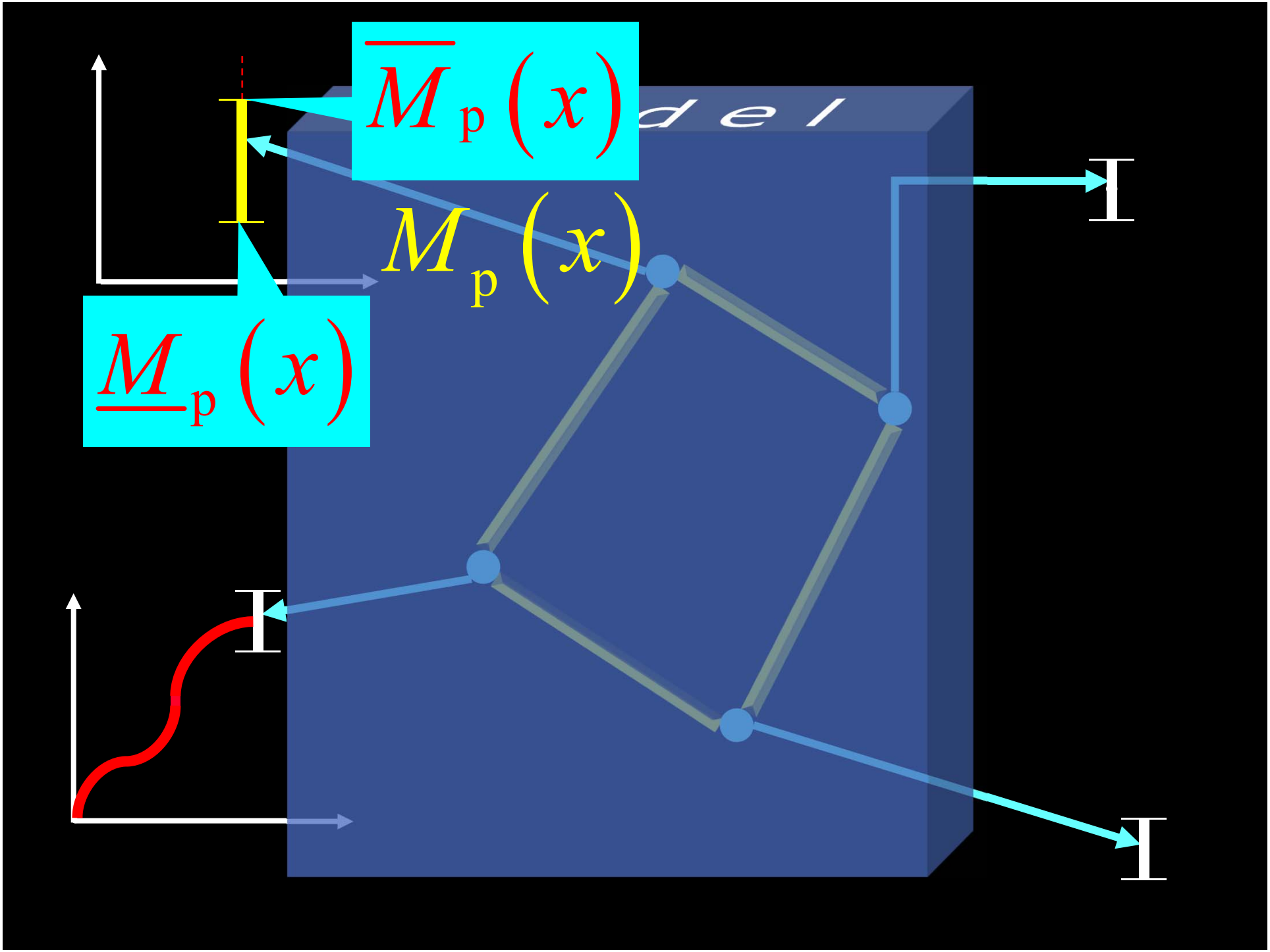
challenge: prediction



curse of dimensionality





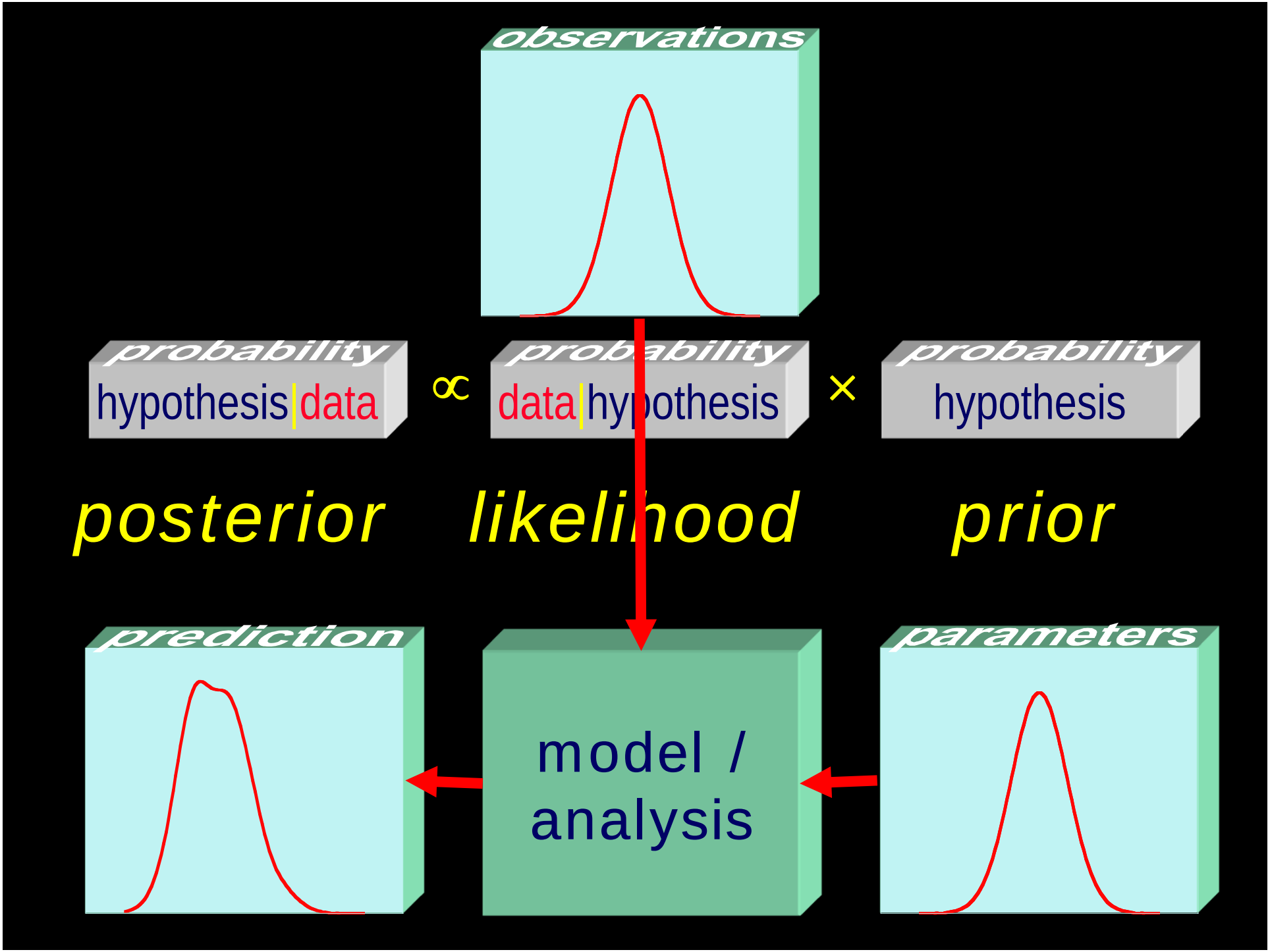

$\overline{M}_p(x)$
$M_p(x)$
$\underline{M}_p(x)$

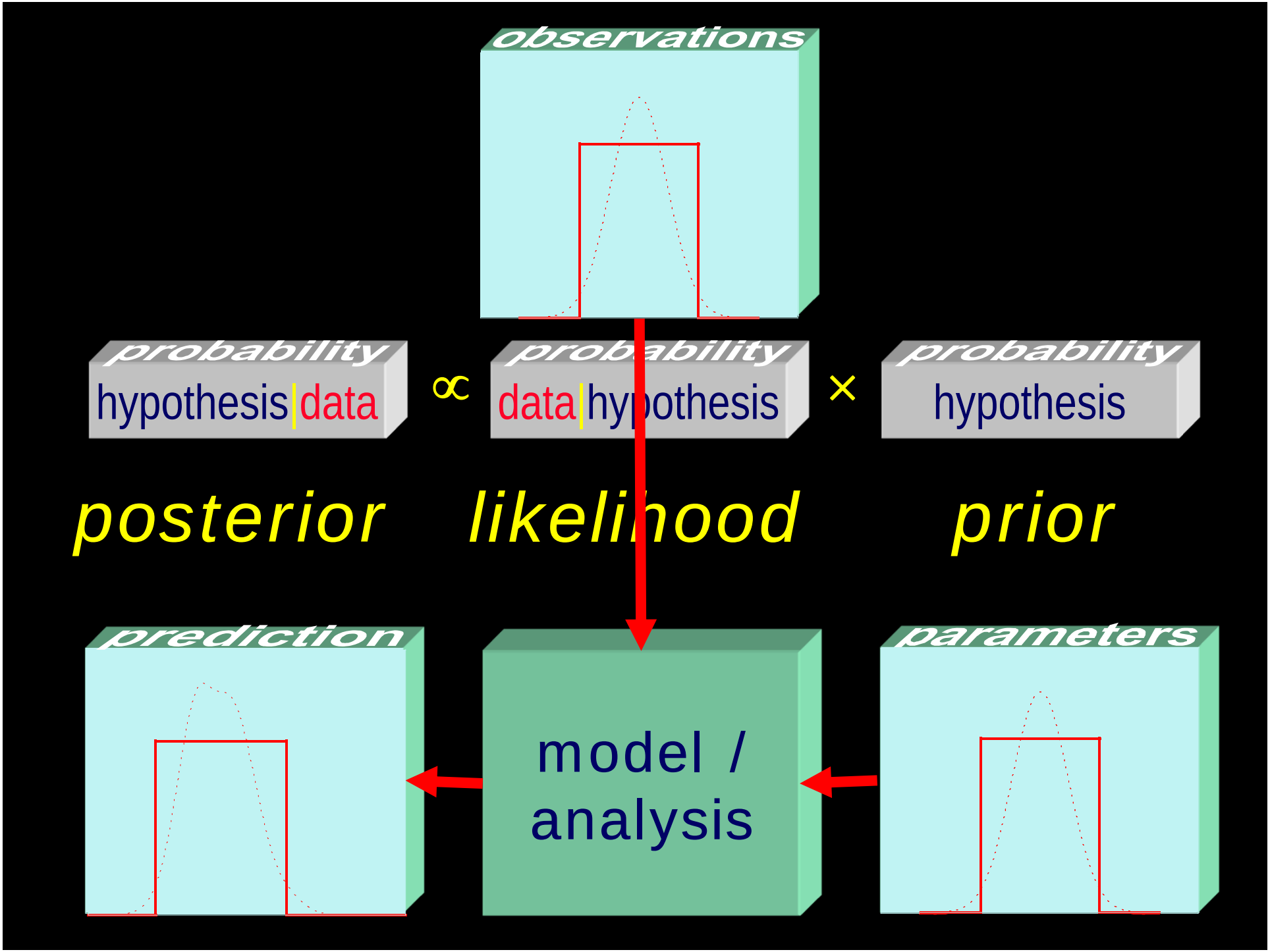
Data analysis performed

in isolation

leads to loss of information

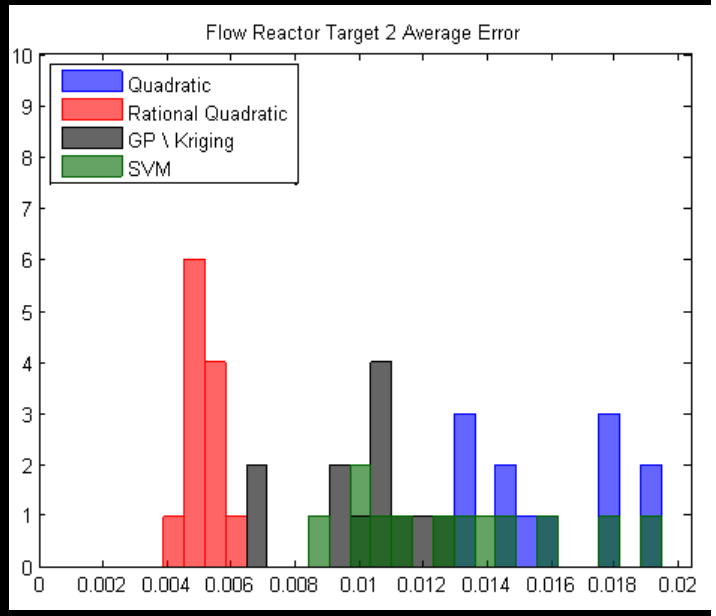
Bayes theorem:


observations
probability
hypothesis|data
∝
probability
data|hypothesis
×
probability
hypothesis
posterior
likelihood
prior
prediction
model /
analysis
parameters


observations
probability
hypothesis|data
∝
probability
data|hypothesis
×
probability
hypothesis
posterior
likelihood
prior
prediction
model /
analysis
parameters

On Quadratic Surrogates of B2B-DC

- Quadratic surrogates enable mathematically rigorous, numerically efficient, and UQ-rich approach of B2B-DC to practical systems
- Quadratics work in practice because model parameters are limited
  - by physical constraints; e.g., 0 < k < collision limit
  - by reaction theory / chemical analogy
  - by prior experimental / theoretical studies
  - and can be linearized; e.g., by the **log** transformation
- And if they do not work, then
  - rational quadratics ("native" with B2B framework)
  - a two-level surrogate approach
    - ❑ first, use machine learning to build "high-order surrogates", e.g., Gaussian Process, Kriging, ε-SVM, Polynomial Chaos
    - ❑ then, build/use on-demand piece-wise quadratics from the high-order surrogates



J. Phys. Chem. A 110:6803 (2006)

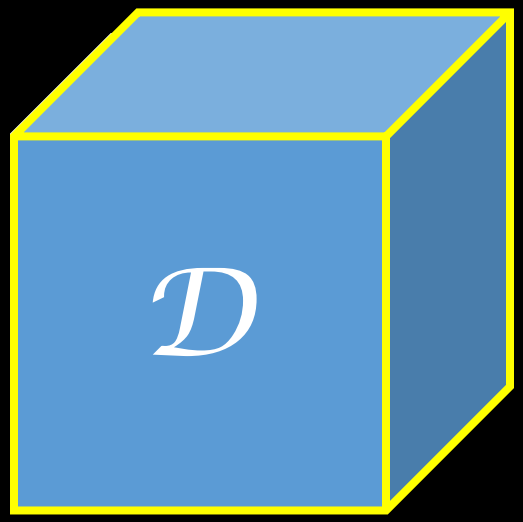
B2B-DC CAN ACCOUNTS FOR EMULATOR ERRORS

$$L \le y \le U$$

surrogate model

$$y = M(\mathbf{x}) + \varepsilon$$

fitting error



$$L - \varepsilon \le M(\mathbf{x}) \le U + \varepsilon$$

$$L' \le M(\mathbf{x}) \le U'$$