Model Identification: A Key Challenge in Computational Systems Biology

Eberhard O.Voit

Department of Biomedical Engineering
Georgia Institute of Technology and Emory University
Atlanta, Georgia

The 2nd International Symposium
on Optimization and Systems Biology (OSB'08)
Lijiang, China, 31 October – 3 November 2008

Overview

Systems Biology and Optimization

Choice of a Suitable Model

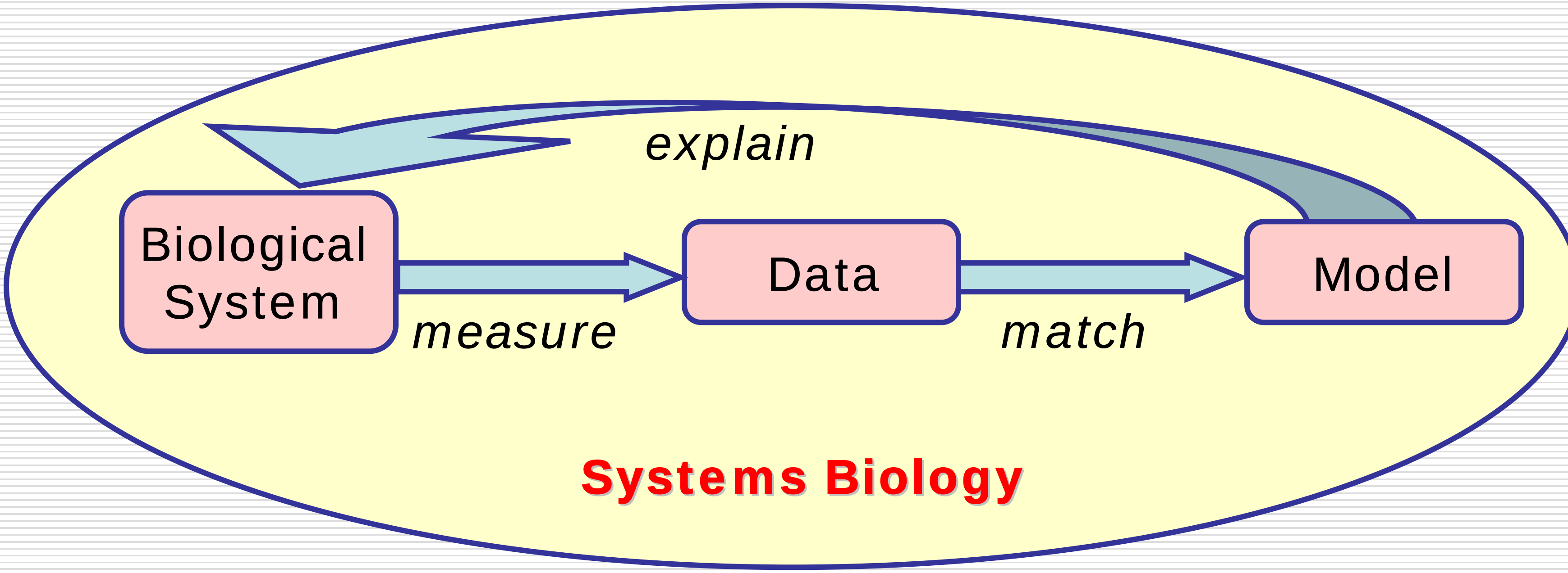
Bottom-up and Top-down Model Estimation

Technical Issues

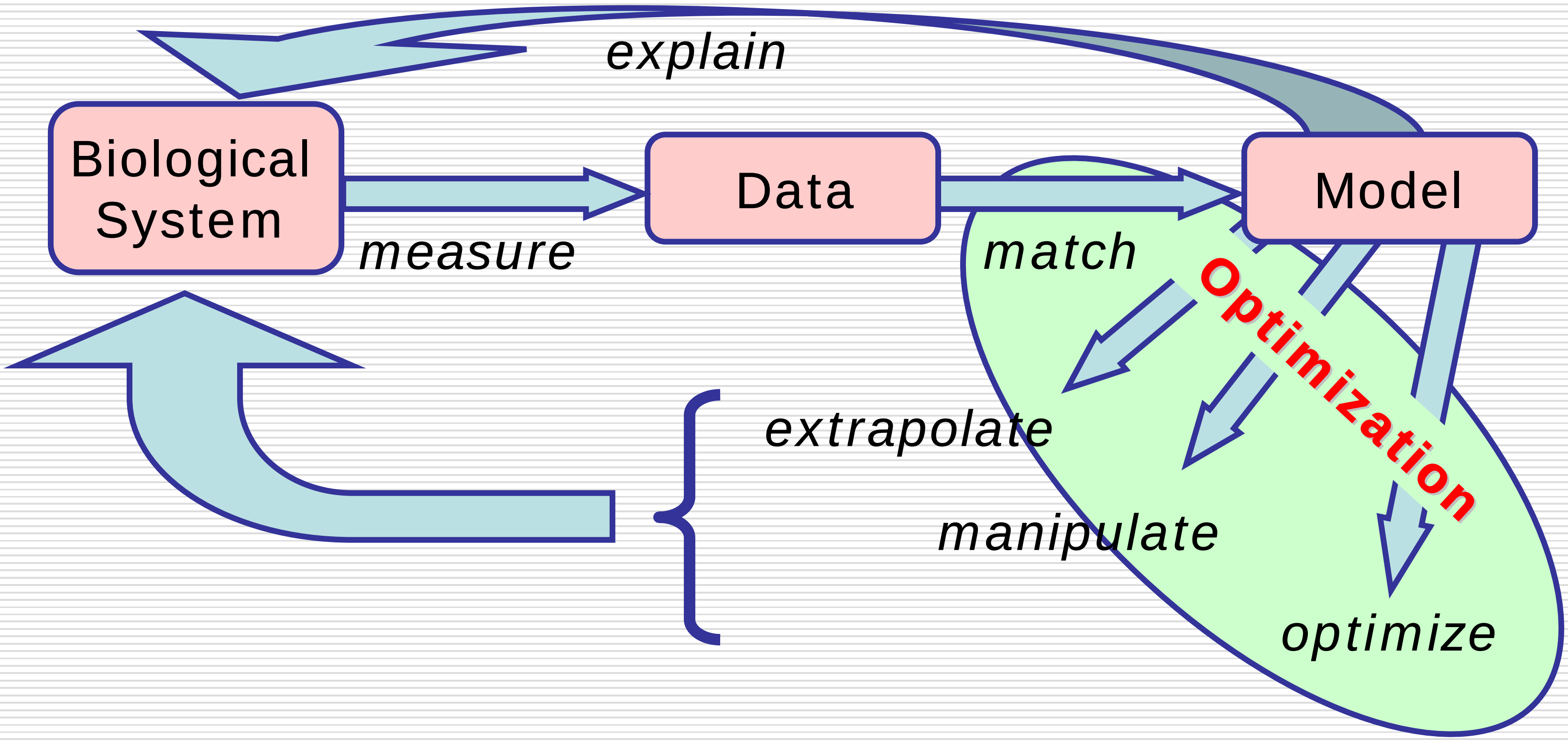
Dynamic Flux Estimation

Open Problems

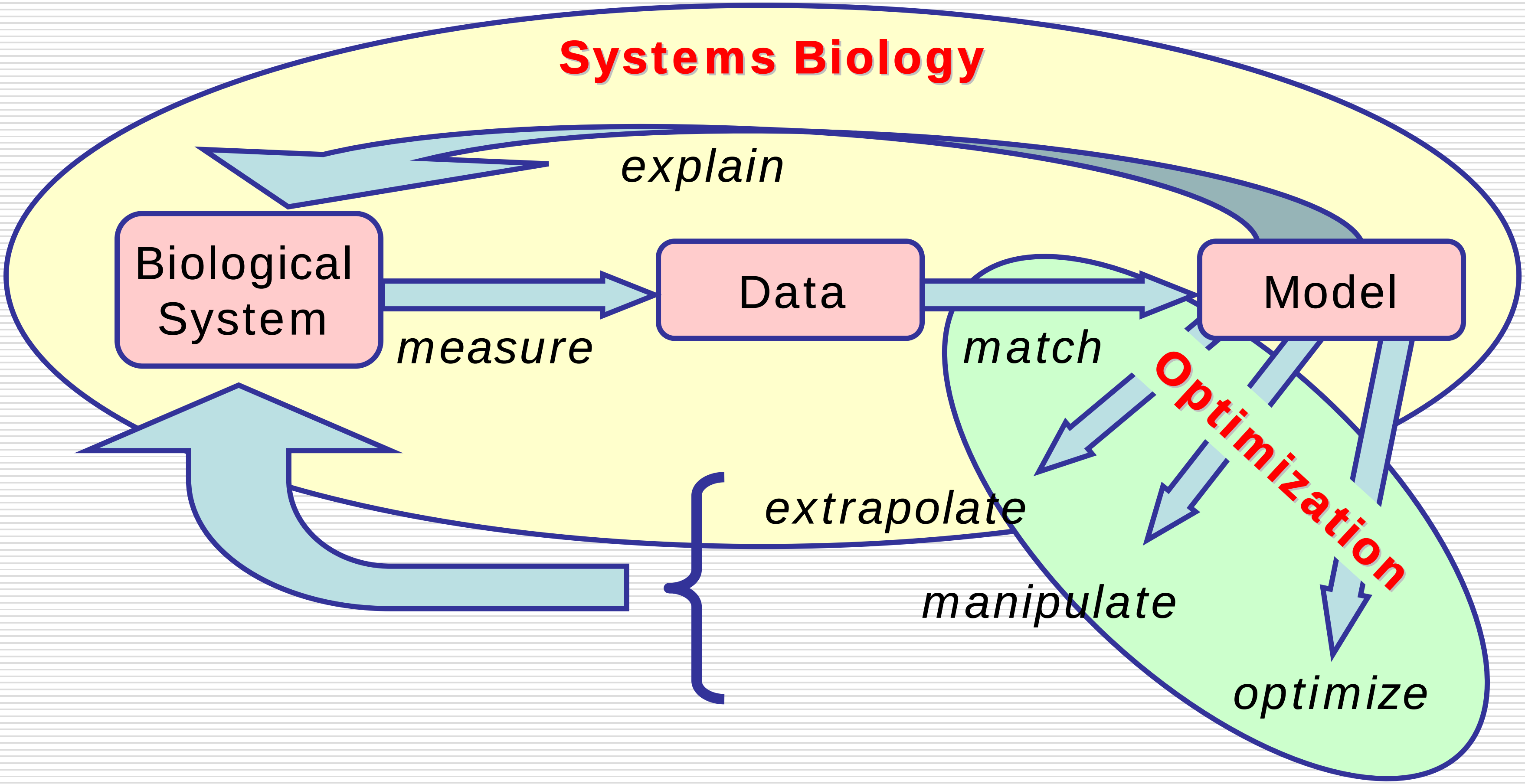

International Symposium on
Optimization and Systems Biology
Oct. 31 - Nov. 3, 2008, Lijiang, China


explain
Biological System
Data
Model
measure
match
Systems Biology

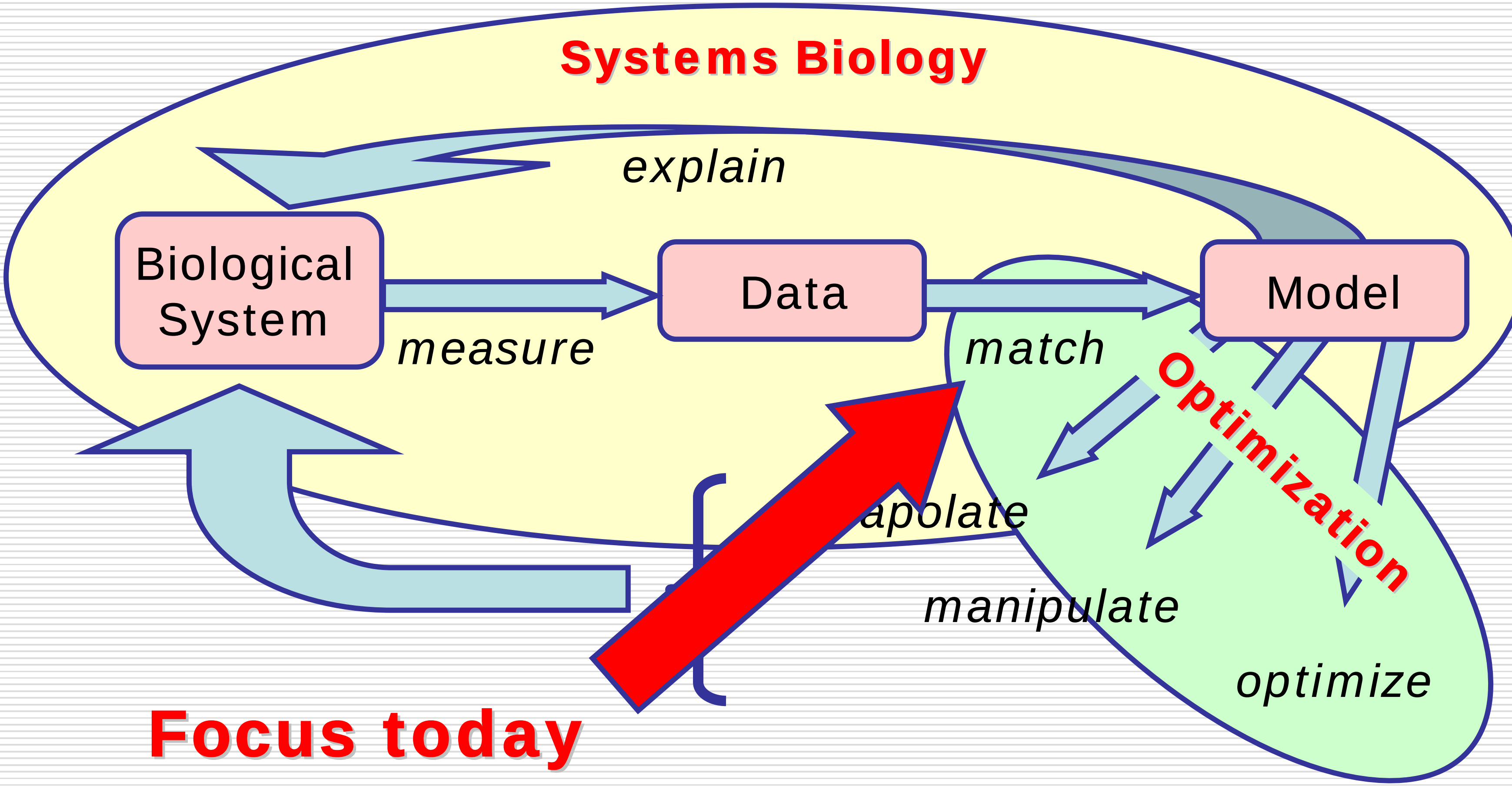

International Symposium on
Optimization and Systems Biology
Oct. 31 - Nov. 3, 2008, Lijiang, China


explain
Biological System
measure
Data
match
Model
Optimization
extrapolate
manipulate
optimize

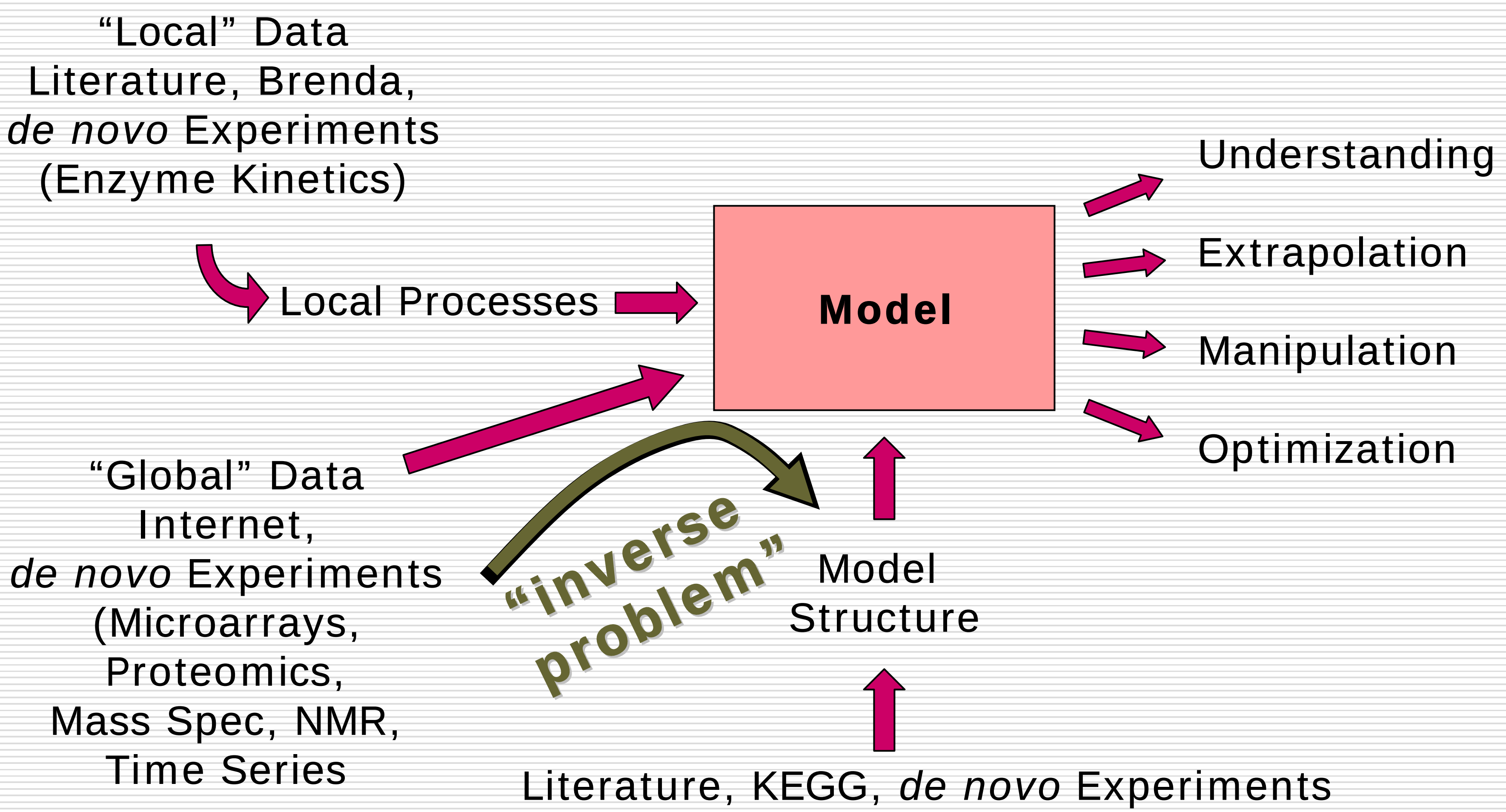

International Symposium on
Optimization and Systems Biology
Oct. 31 - Nov. 3, 2008, Lijiang, China


Systems Biology
explain
Biological System
Data
Model
measure
match
Optimization
extrapolate
manipulate
optimize

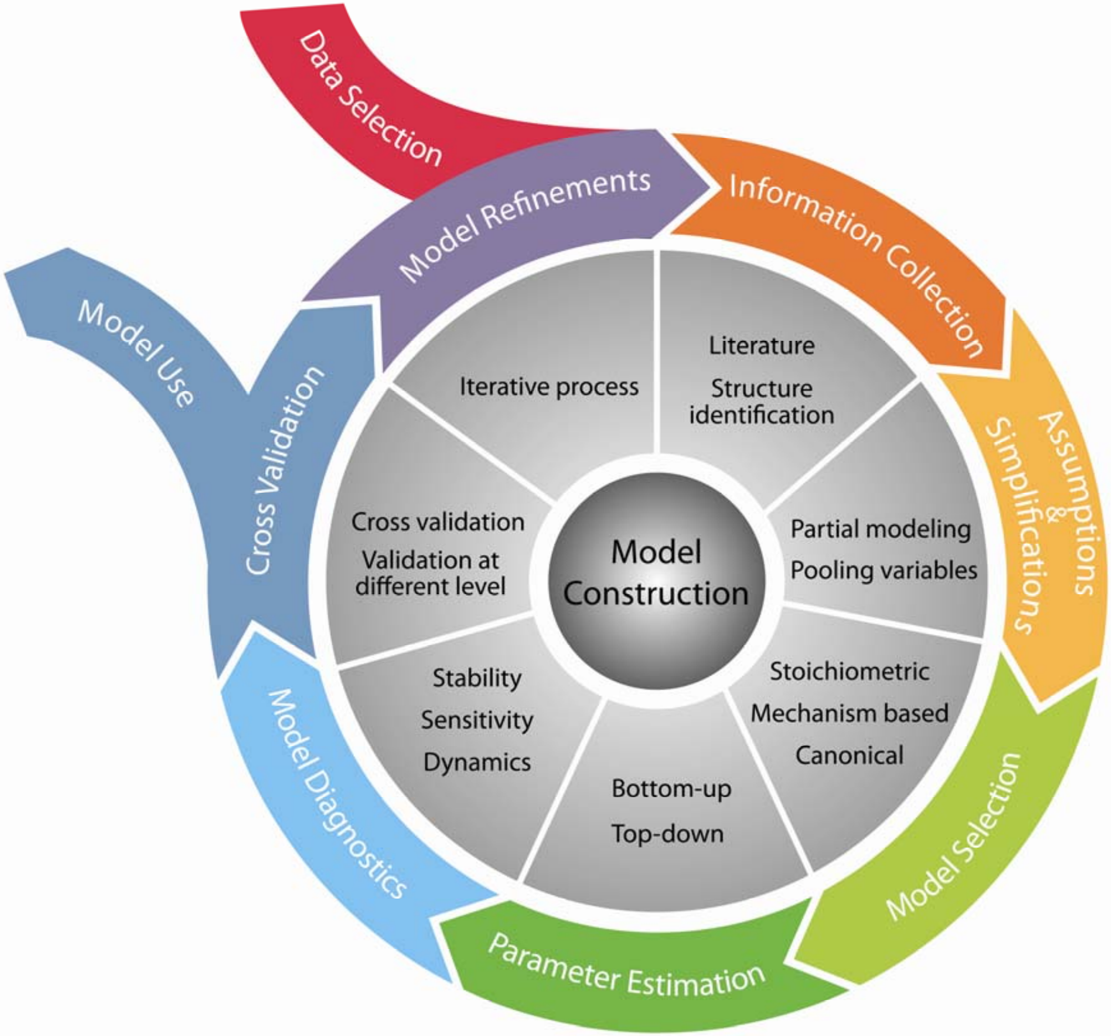

International Symposium on
Optimization and Systems Biology
Oct. 31 - Nov. 3, 2008, Lijiang, China


Systems Biology
explain
Biological System
Data
Model
measure
match
Optimization
apolate
manipulate
optimize
Focus today

Application: Pathway Modeling



Overview of Modeling Process



Formulation of a Dynamical Systems Model

$$\xrightarrow{\;V_i^+\;} \boxed{X_i} \xrightarrow{\;V_i^-\;}$$

$$\dot{X}_i = \frac{dX_i}{dt} = V_i^+ - V_i^-$$

$$V_i^+ = V_i^+(\underbrace{X_1, X_2, \ldots, X_n}_{\textit{inside}}, \underbrace{X_{n+1}, \ldots, X_{n+m}}_{\textit{outside}})$$

complicated

Big Problem: Where do we get functions from?

Sources of Functions for Complex Systems Models

Physics: Functions come from theory

Biology: No theory available

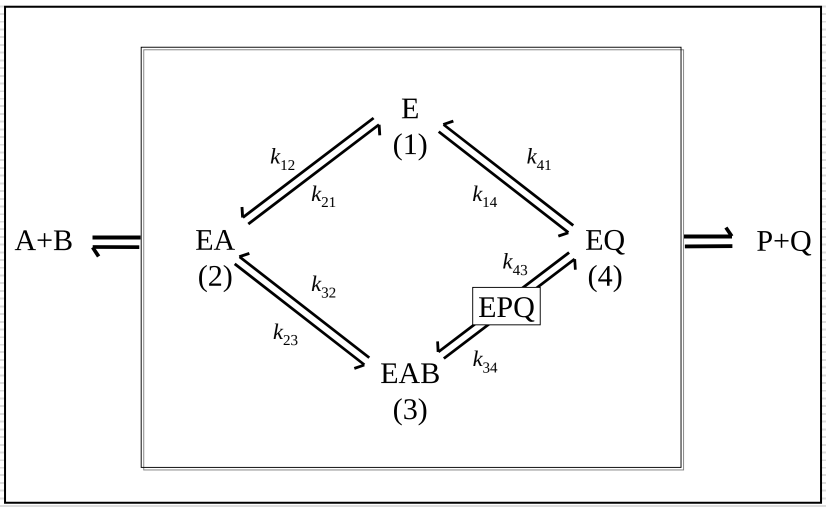
Solution 1: Educated guesses: growth functions

Solution 2: “Partial” theory: Enzyme kinetics

Solution 3: Generic approximation

Why not Use "True" Functions?

$$A+B \rightleftharpoons P+Q$$

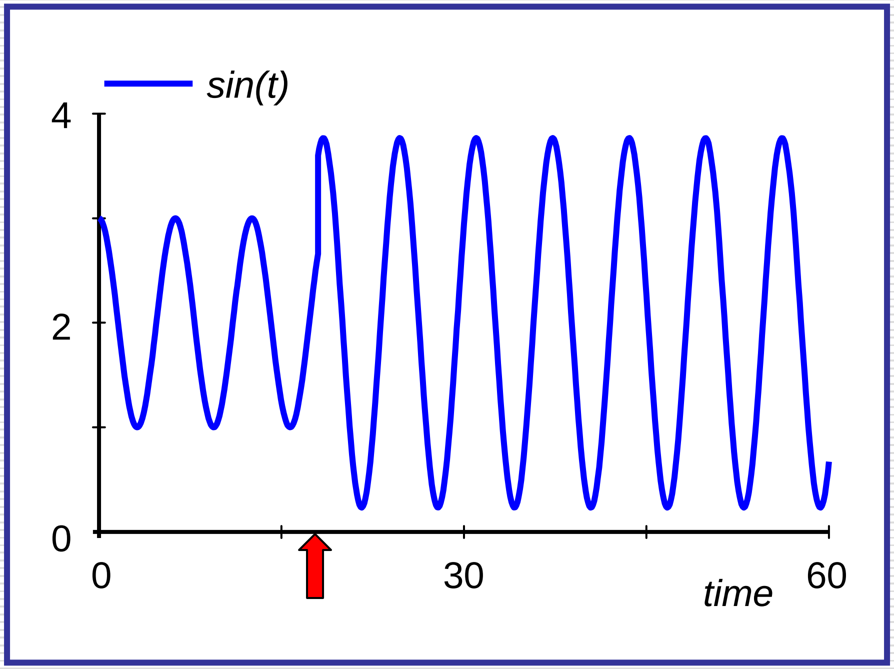


$$v = \frac{\left(\dfrac{\text{num.1}}{\text{coef. AB}}\right)(A)(B) - \left(\dfrac{\text{num.1}}{\text{coef. AB}} \times \dfrac{\text{num.2}}{\text{num.1}}\right)(P)(Q)}{\begin{aligned} &\left(\dfrac{\text{constant}}{\text{coef. A}} \times \dfrac{\text{coef. A}}{\text{coef. AB}}\right) + \left(\dfrac{\text{coef. A}}{\text{coef. AB}}\right)(A) + \left(\dfrac{\text{coef. B}}{\text{coef. AB}}\right)(B) \\ &+ \left(\dfrac{\text{coef. AB}}{\text{coef. AB}}\right)(A)(B) + \left(\dfrac{\text{coef.P}}{\text{coef. AP}} \times \dfrac{\text{coef. AP}}{\text{coef. A}} \times \dfrac{\text{coef. A}}{\text{coef. AB}}\right)(P) \\ &+ \left(\dfrac{\text{coef.Q}}{\text{constant}} \times \dfrac{\text{constant}}{\text{coef. A}} \times \dfrac{\text{coef. A}}{\text{coef. AB}}\right)(Q) \\ &+ \left(\dfrac{\text{coef.AP}}{\text{coef. A}} \times \dfrac{\text{coef.A}}{\text{coef. AB}}\right)(A)(P) + \left(\dfrac{\text{coef.BQ}}{\text{coef. B}} \times \dfrac{\text{coef.B}}{\text{coef. AB}}\right)(B)(Q) \\ &+ \left(\dfrac{\text{coef.PQ}}{\text{coef.Q}} \times \dfrac{\text{coef.Q}}{\text{constant}} \times \dfrac{\text{constant}}{\text{coef. A}} \times \dfrac{\text{coef. A}}{\text{coef. AB}}\right)(P)(Q) \\ &+ \left(\dfrac{\text{coef.ABP}}{\text{coef. AB}}\right)(A)(B)(P) \\ &+ \left(\dfrac{\text{coef.BPQ}}{\text{coef. BQ}} \times \dfrac{\text{coef.BQ}}{\text{coef. B}} \times \dfrac{\text{coef.B}}{\text{coef. AB}}\right)(B)(P)(Q) \end{aligned}}$$

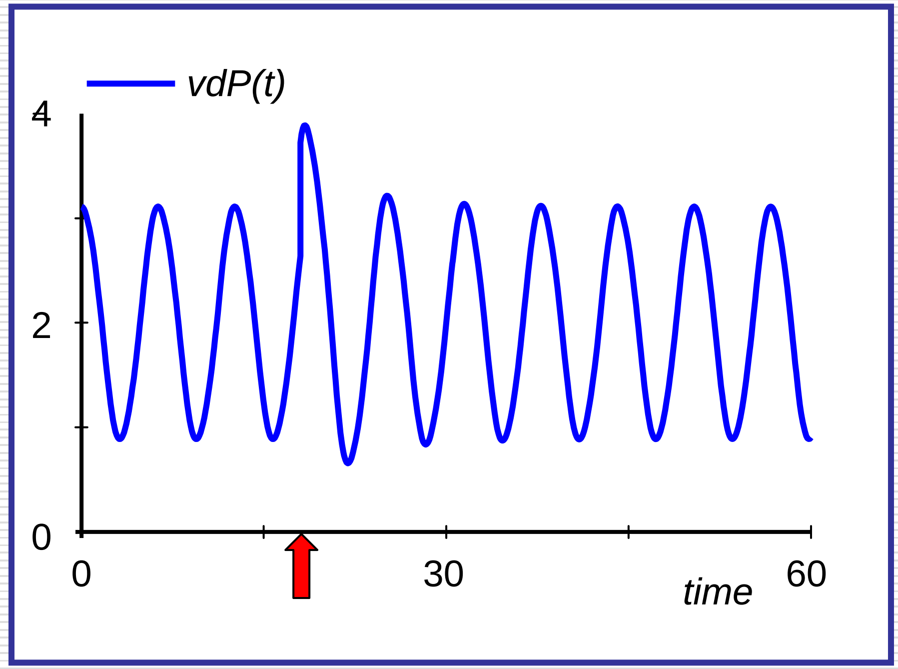
from Schultz (1994)

Why not Use Linear Functions?

Example: Heartbeat modeled as stable limit cycle



System of linear
differential equations



System of non-linear
differential equations

Formulation of a Nonlinear Model for Complex Systems

Challenge:

Linear approximation unsuited

Infinitely many nonlinear functions

Solution with Potential:

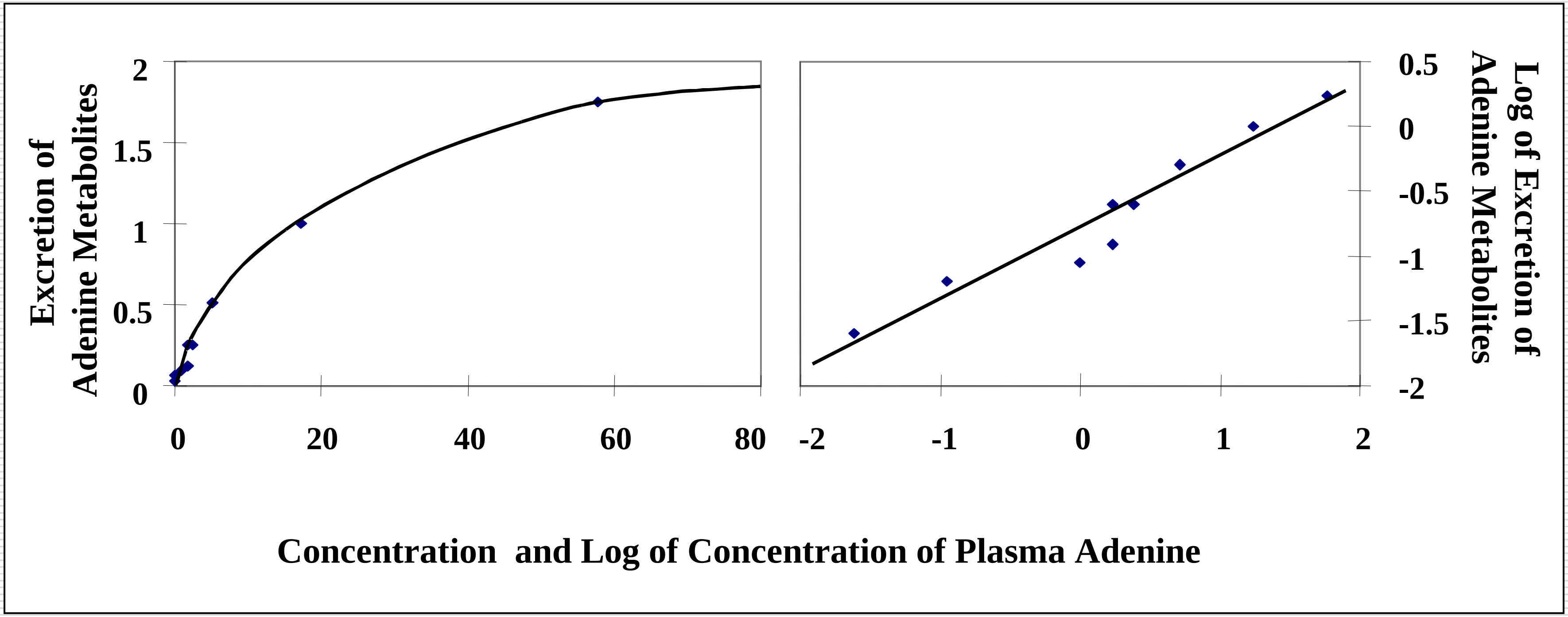
$$\dot{X}_i = \frac{dX_i}{dt} = V_i^+ - V_i^-$$

Savageau (1969): Approximate V_i^+ and V_i^- in a logarithmic coordinate system, using Taylor theory.

Result: *Canonical Modeling; Biochemical Systems Theory*.

Example

Adenine Excretion as a Function of Plasma Adenine Concentration



Result: S-system

$$\dot{X}_i = \alpha_i X_1^{g_{i1}} X_2^{g_{i2}} \ldots X_{n+m}^{g_{i,n+m}} - \beta_i X_1^{h_{i1}} X_2^{h_{i2}} \ldots X_{n+m}^{h_{i,n+m}}$$

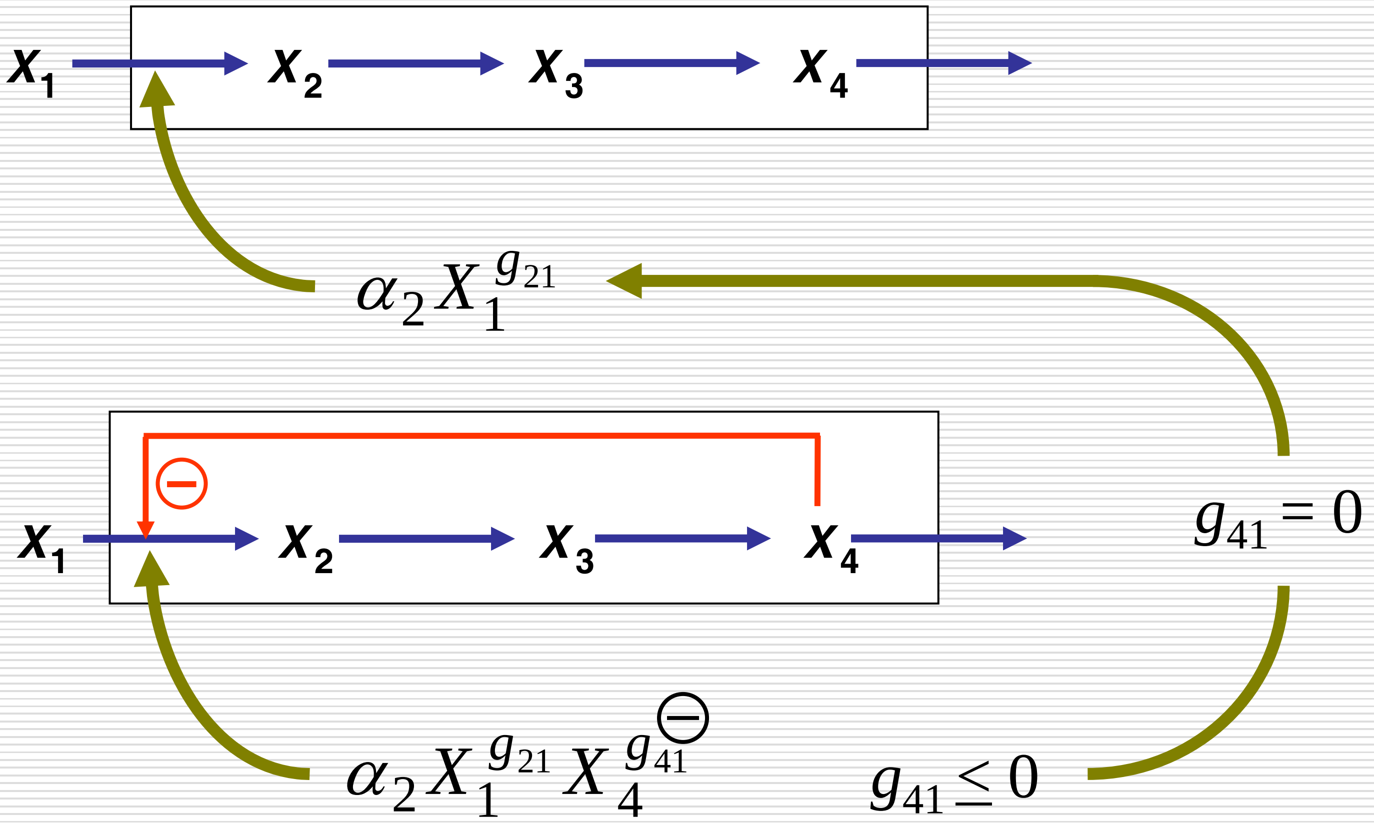
Each term is represented as a product of power-functions.

Each term contains and only those variables that have a direct effect; others have exponents of 0 and drop out.

α's and β's are *rate constants*, *g*'s and *h*'s *kinetic orders*.

Important:
Each term contains exactly those variables that have a direct effect; others have exponents of 0 and drop out.

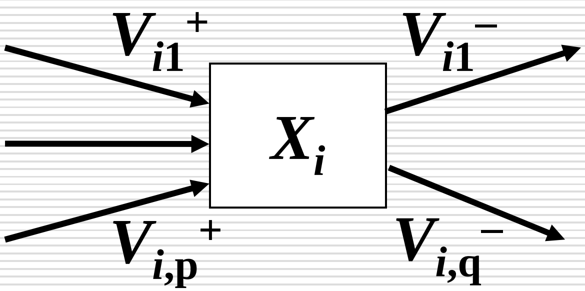
Mapping
Structure ⟺ Parameters



Alternative Formulations Within BST

S-system Form:

$$\dot{X}_i = \alpha_i X_1^{g_{i1}} X_2^{g_{i2}} \ldots X_{n+m}^{g_{i,n+m}} - \beta_i X_1^{h_{i1}} X_2^{h_{i2}} \ldots X_{n+m}^{h_{i,n+m}}$$

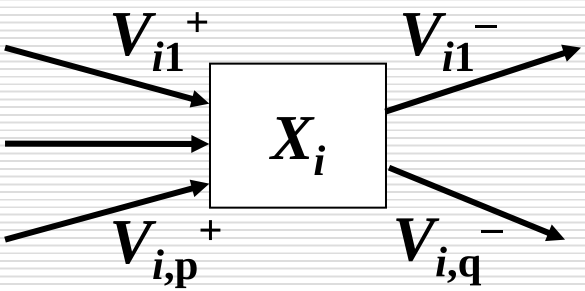


$$\dot{X}_i = \frac{dX_i}{dt} = \sum V_{ij}^{+} - \sum V_{ij}^{-}$$

Alternative Formulations

S-system Form:

$$\dot{X}_i = \alpha_i X_1^{g_{i1}} X_2^{g_{i2}} \ldots X_{n+m}^{g_{i,n+m}} - \beta_i X_1^{h_{i1}} X_2^{h_{i2}} \ldots X_{n+m}^{h_{i,n+m}}$$

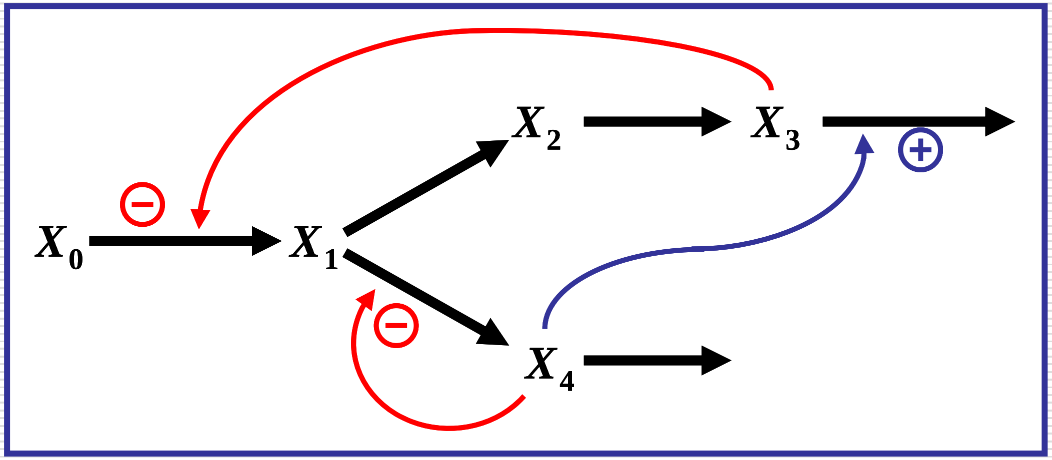


$$\dot{X}_i = \frac{dX_i}{dt} = \sum V_{ij}^{+} - \sum V_{ij}^{-}$$

Generalized Mass Action Form:

$$\dot{X}_i = \sum \pm \gamma_{ik} \prod X_j^{f_{ijk}}$$

Example of Canonical Model Design



GMA / S: $\dot{X}_2 = 8X_1^{0.75} - 5X_2^{0.3}$ $X_2(t_0) = 1$

GMA / S: $\dot{X}_3 = 5X_2^{0.3} - 5X_3^{0.5}X_4^{0.2}$ $X_3(t_0) = 0.5$

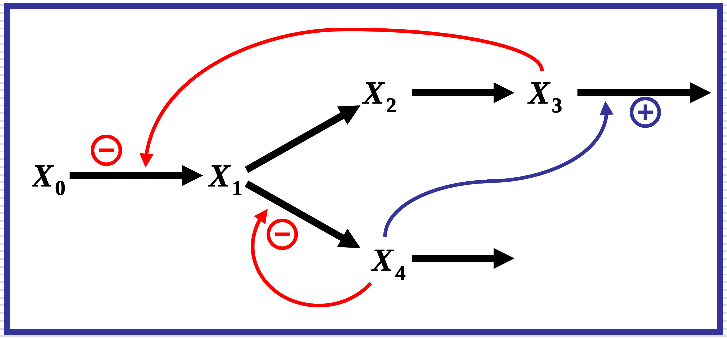
GMA / S: $\dot{X}_4 = 12X_1^{0.5}X_4^{-1} - 4X_4^{0.8}$ $X_4(t_0) = 6$

GMA / S: $X_0 = 1.1$ (constant)

GMA: $\dot{X}_1 = 20X_0X_3^{-0.9} - 8X_1^{0.75} - 12X_1^{0.5}X_4^{-1}$ $X_1(t_0) = 0.8$

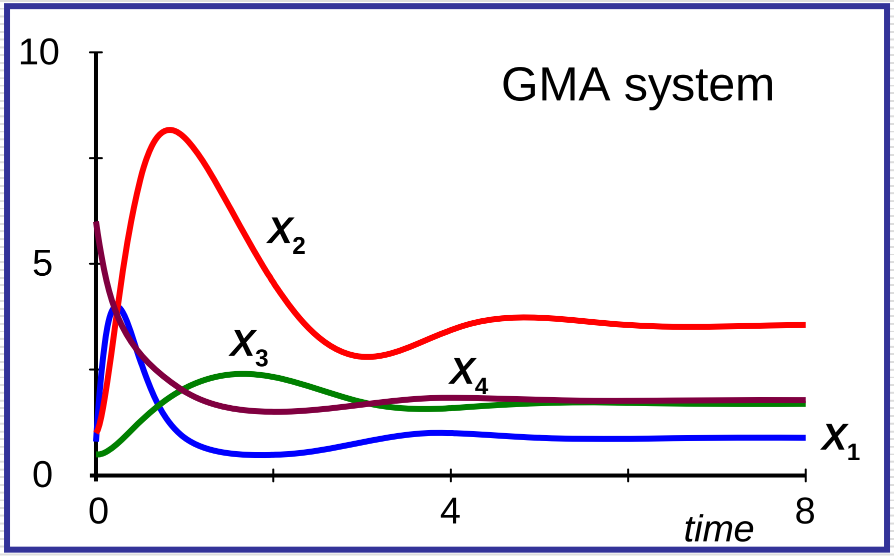
S-system: $\dot{X}_1 = 20X_0X_3^{-0.9} - 19X_1^{0.64}X_4^{-0.45}$ $X_1(t_0) = 0.8$

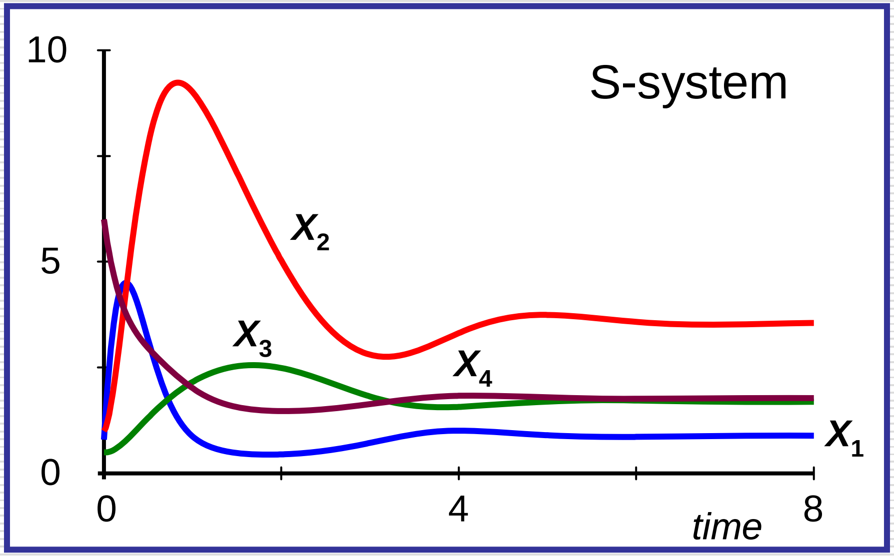
Example of Canonical Model Design



GMA: $\dot{X}_1 = 20 X_0 X_3^{-0.9} - 8 X_1^{0.75} - 12 X_1^{0.5} X_4^{-1}$ $X_1(t_0) = 0.8$

S-system: $\dot{X}_1 = 20 X_0 X_3^{-0.9} - 19 X_1^{0.64} X_4^{-0.45}$ $X_1(t_0) = 0.8$





Doable Size

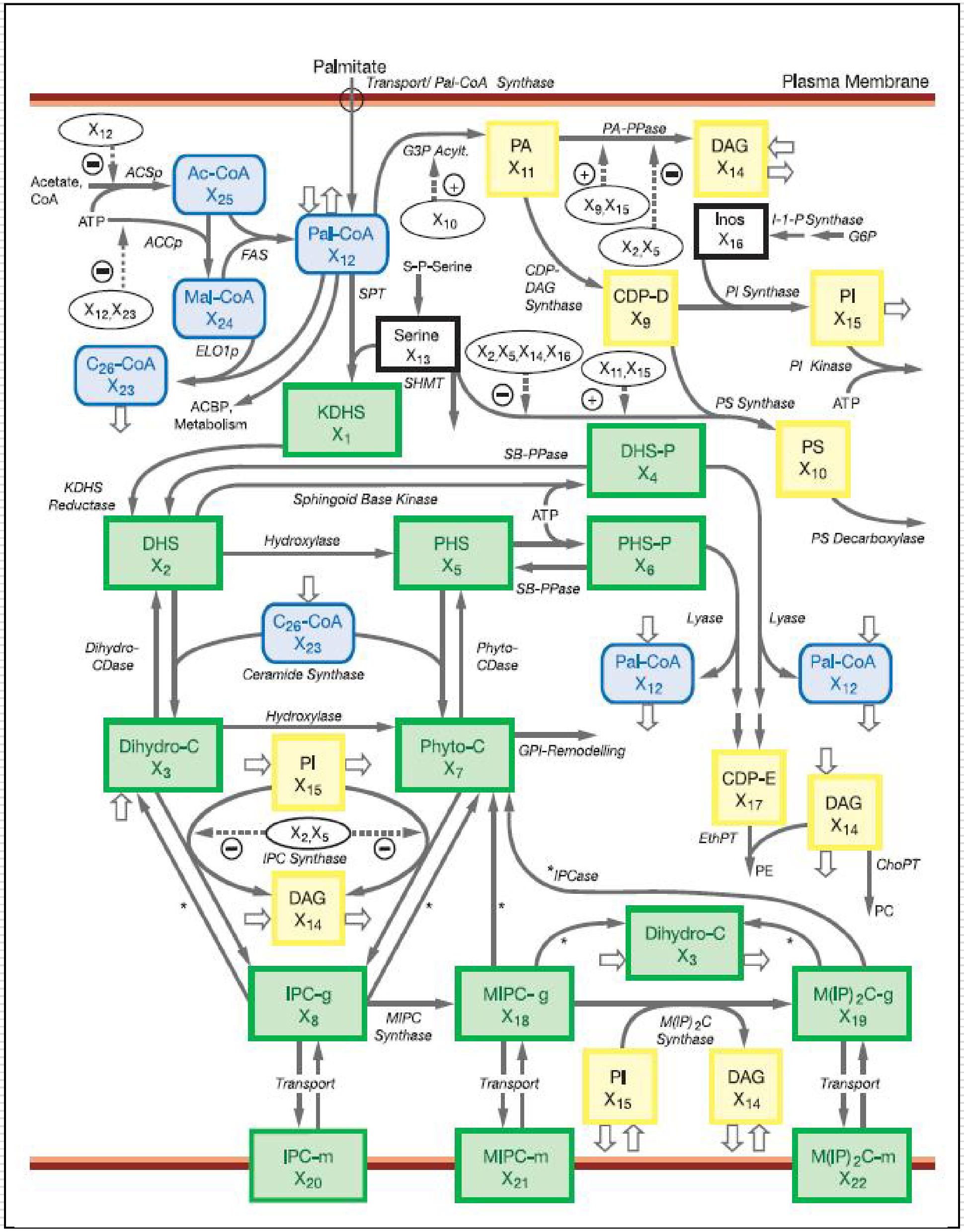
Sphingolipid pathway (purely metabolic)

1. Many metabolites

2. Many reactions

3. Many stimuli and agents regulate several enzymes of lipid metabolism

4. Some *in vivo* experiments



Alvarez, Sims, Hannun, Voit
JTB, 2004; *Nature*, 2005

Applications

Pathways: purines, glycolysis, citric acid, TCA, red blood cell, trehalose, sphingolipids, …

Genes: circuitry, regulation,…

Genome: explain expression patterns upon stimulus

Growth, immunology, pharmaceutical science, forestry, …

Metabolic engineering: optimize yield in microbial pathways

Dynamic labeling analyses possible

Math: recasting, function classification, bifurcation analysis,…

Statistics: S-system representation, S-distribution, trends; applied to seafood safety, marine mammals, health economics

Advantages of Canonical Models

Prescribed model design: Rules for translating diagrams into equations; rules can be automated

Direct interpretability of parameters and other features

One-to-one relationship between parameters and model structure simplifies parameter estimation and model identification

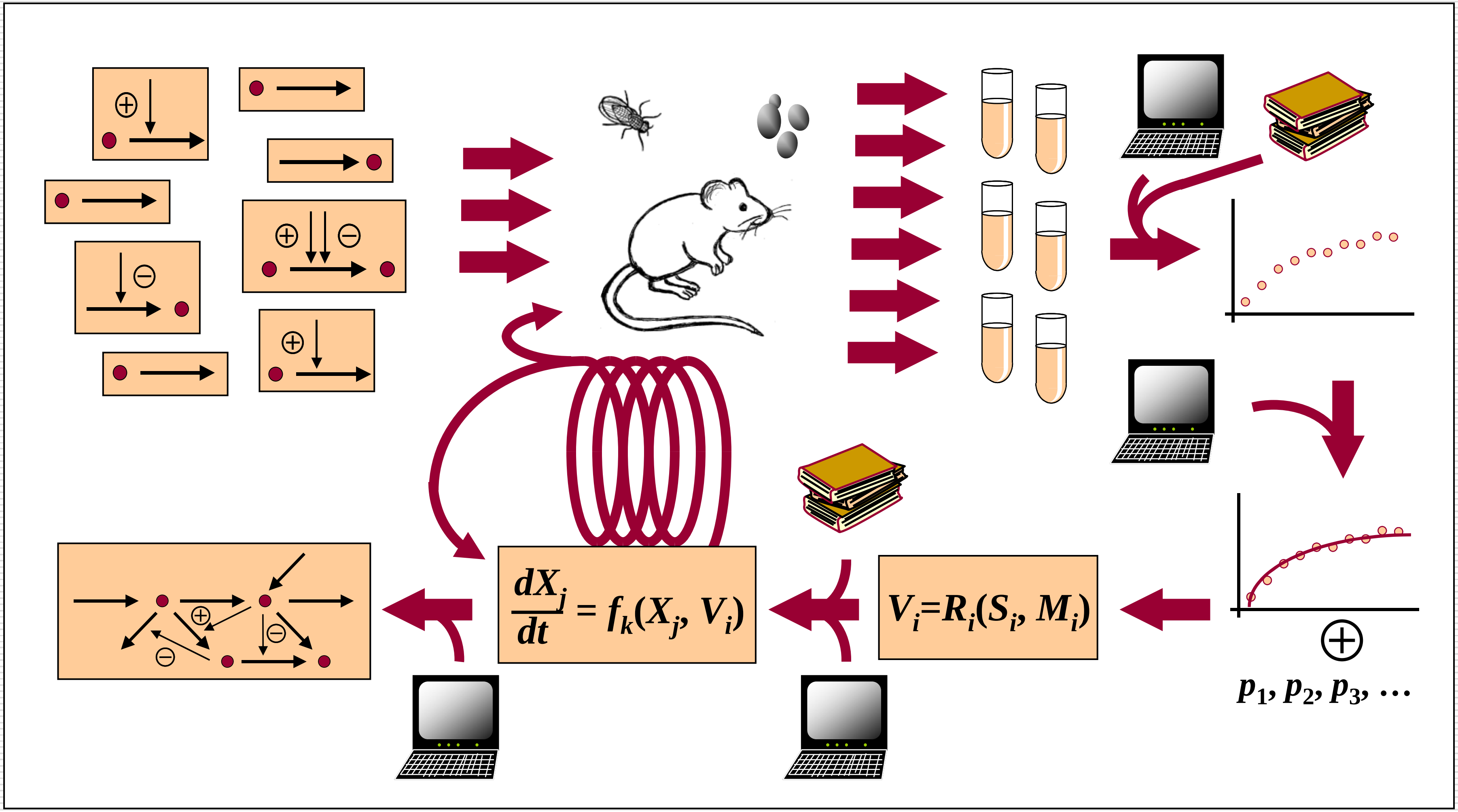
Simplified steady-state computations (for S-systems), including steady-state equations, stability, sensitivities, gains

Simplified optimization under steady-state conditions

Efficient numerical solutions and time-dependent sensitivities

In some sense minimal bias of model choice and minimal model size; easy scalability

Flow Chart of Systems Identification Strategy



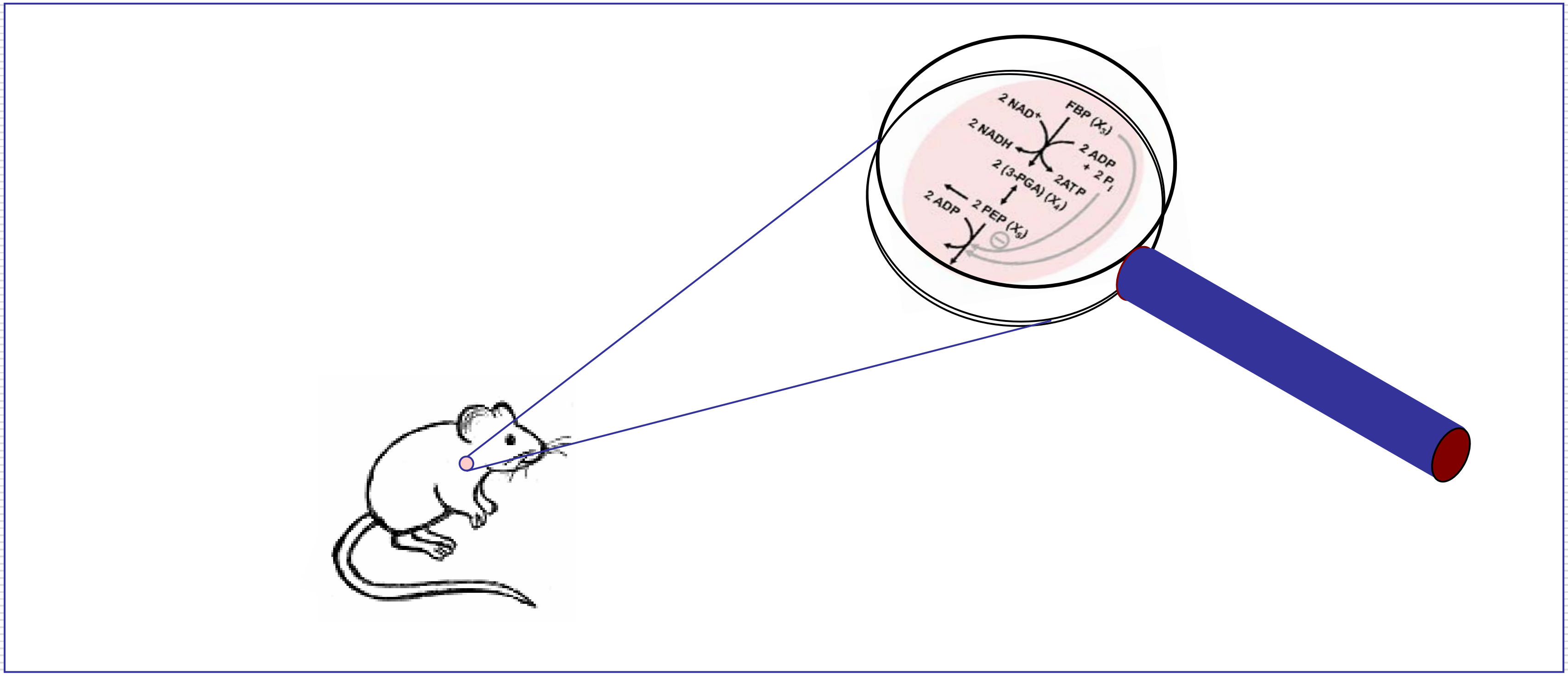
Voit, *Drug Discovery Today*, 2004

Problems with Traditional System Identification Strategy

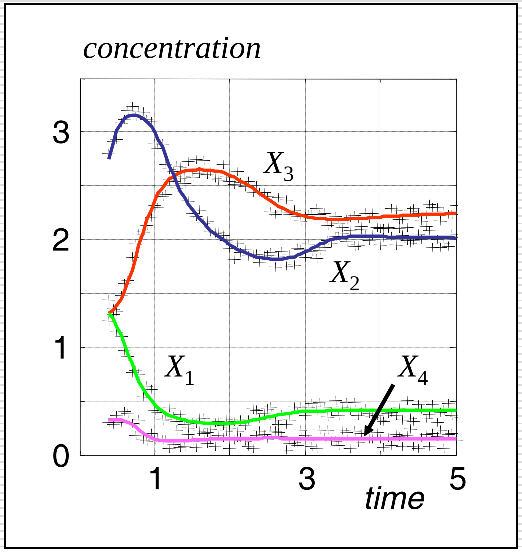
- Lots of time-consuming work and effort!
- Very many a priori assumptions
- What's important, what isn't?
- Topology
- Regulation
- Functional forms
- Seldom consistent experiments
- Mixing and matching of organisms, strains, conditions
- Paucity of data for comparisons with documented responses
- Iterative nature of process time consuming

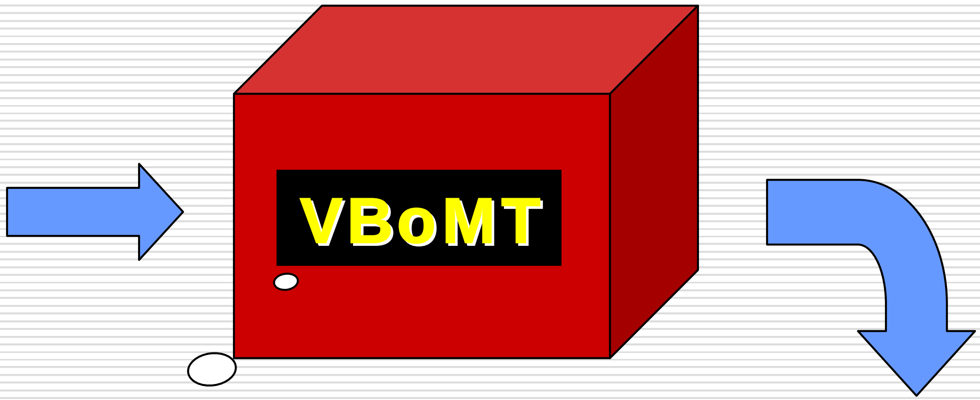
Alternative to Traditional Modeling: Top-Down Modeling

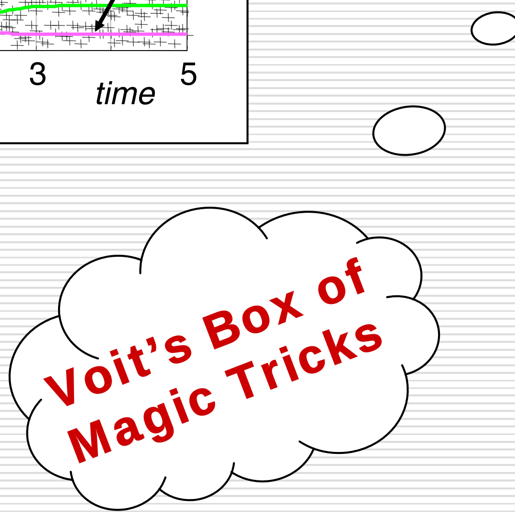
- Use information at the “global” level (*in vivo* time series data) to deduce (per model) structure and regulation at the “local” level (connectivity, signals,…)

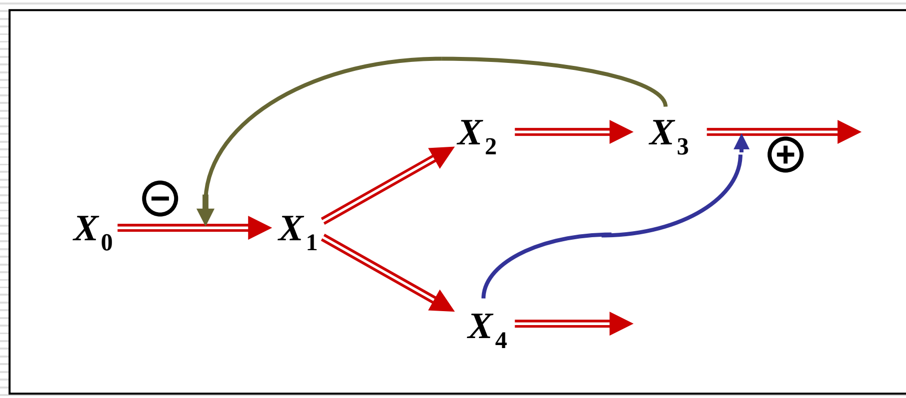


Inverse Problems: Sandbox Example

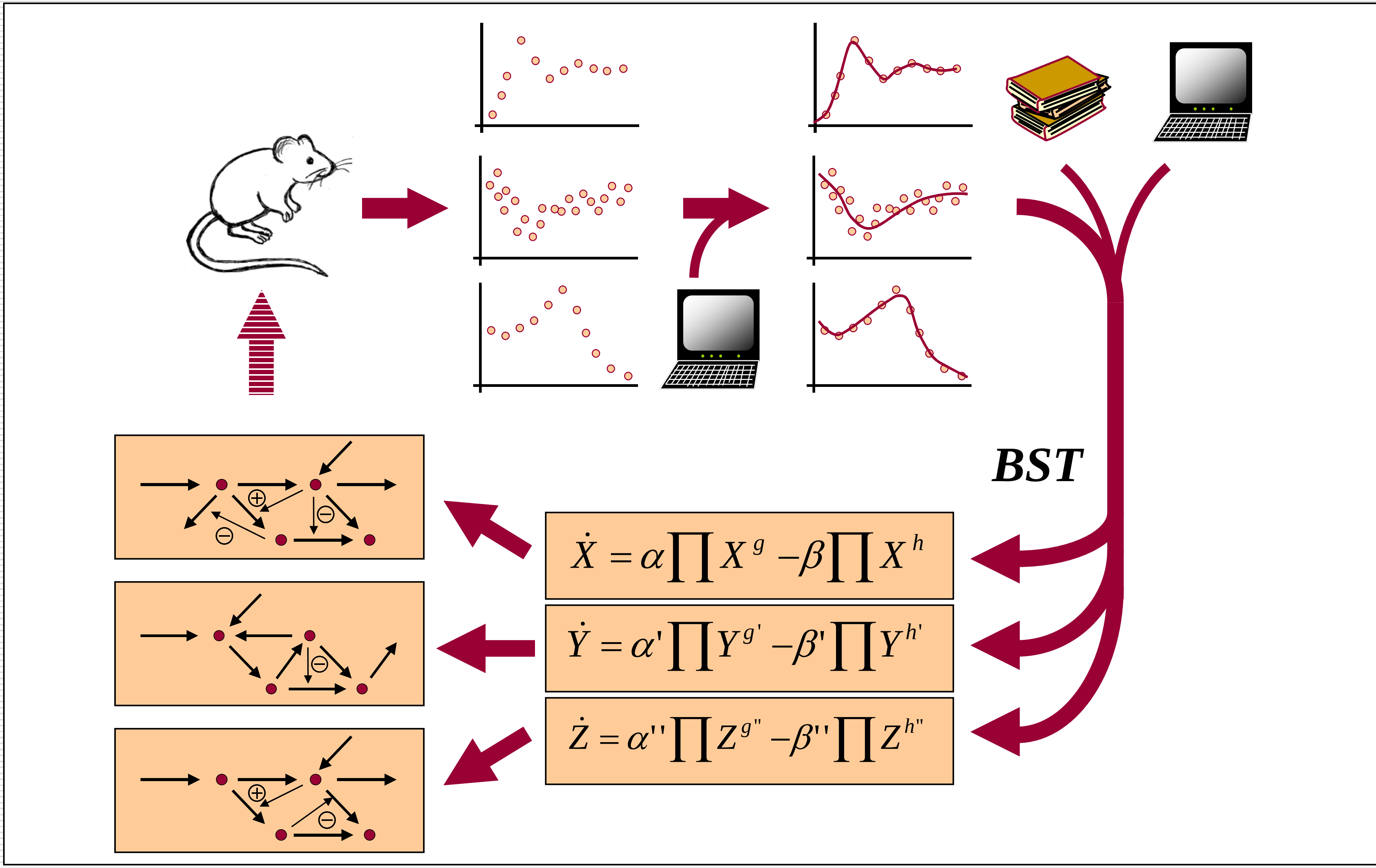








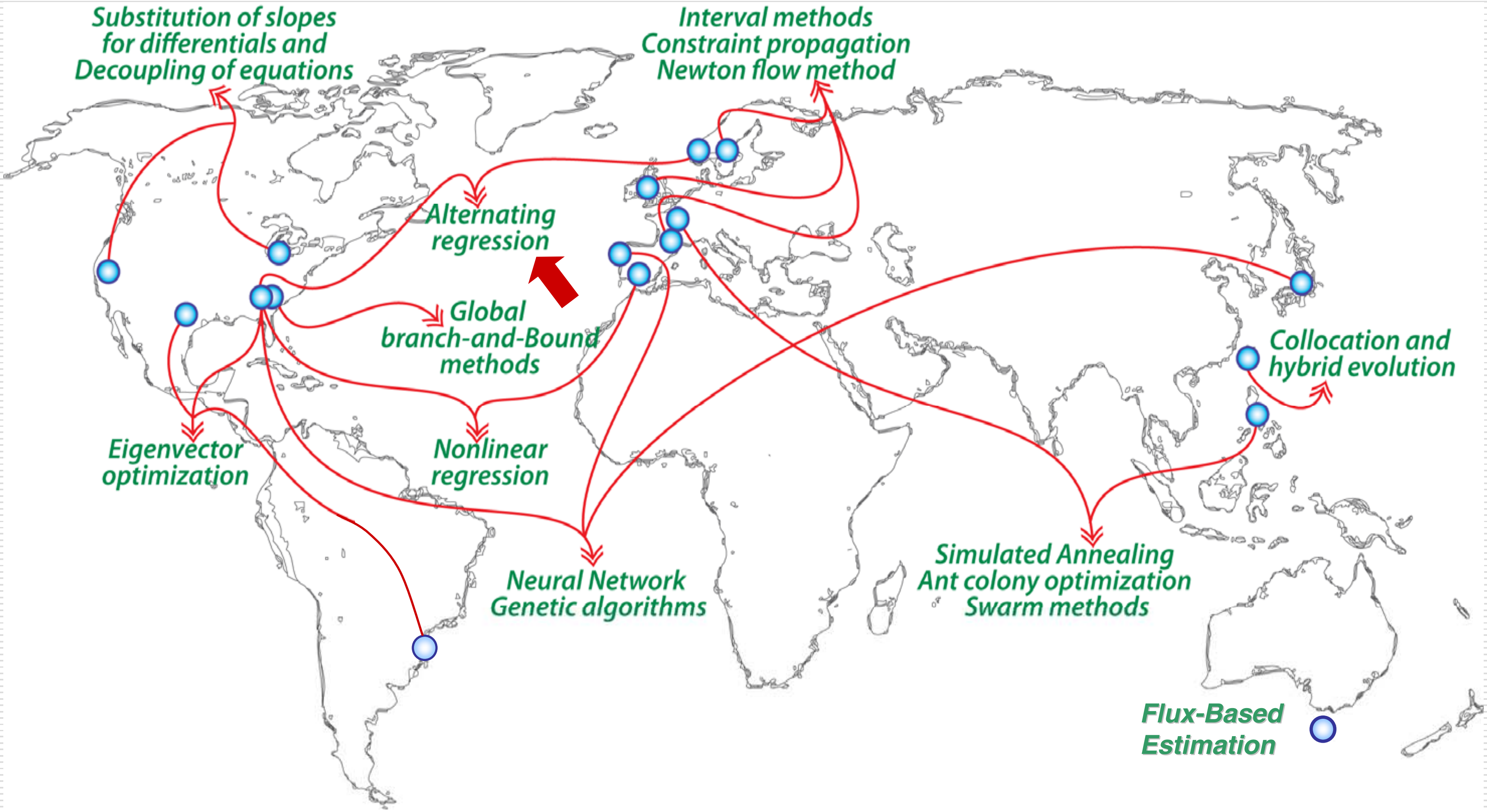
Top-Down “Inverse” Modeling



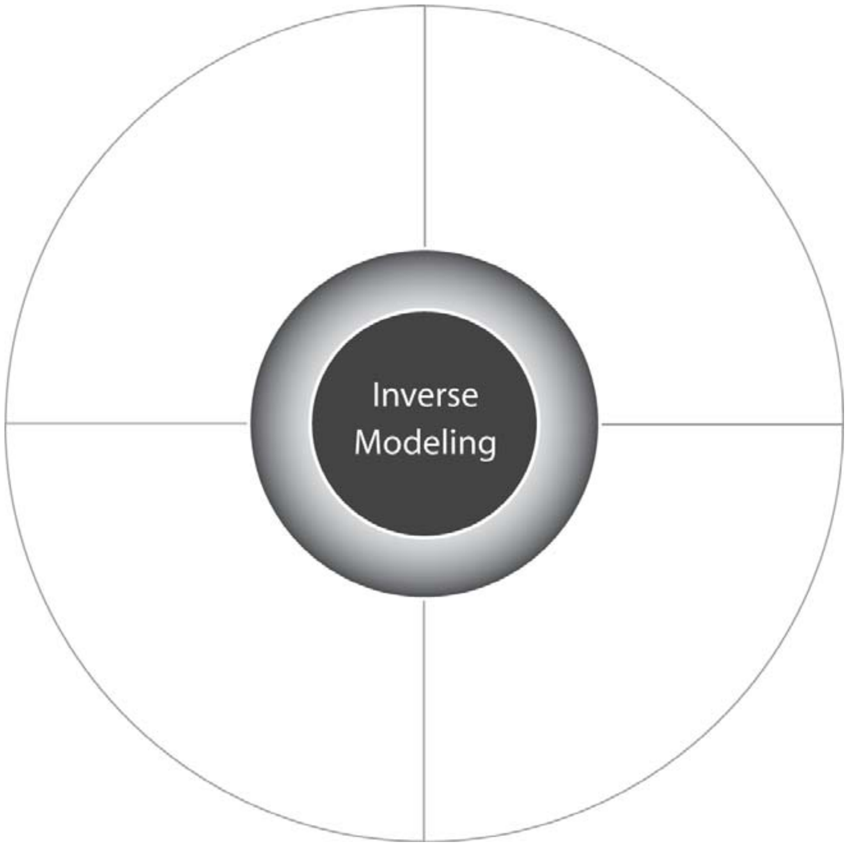
Key Step: Parameter Estimation from Time Series Data

- According to computer scientists: trivial, solved.
- Many methods
- Most work sometimes
- None works always
- Estimation remains to be a frustrating topic!
- Example: Kikuchi *et al.* 2003

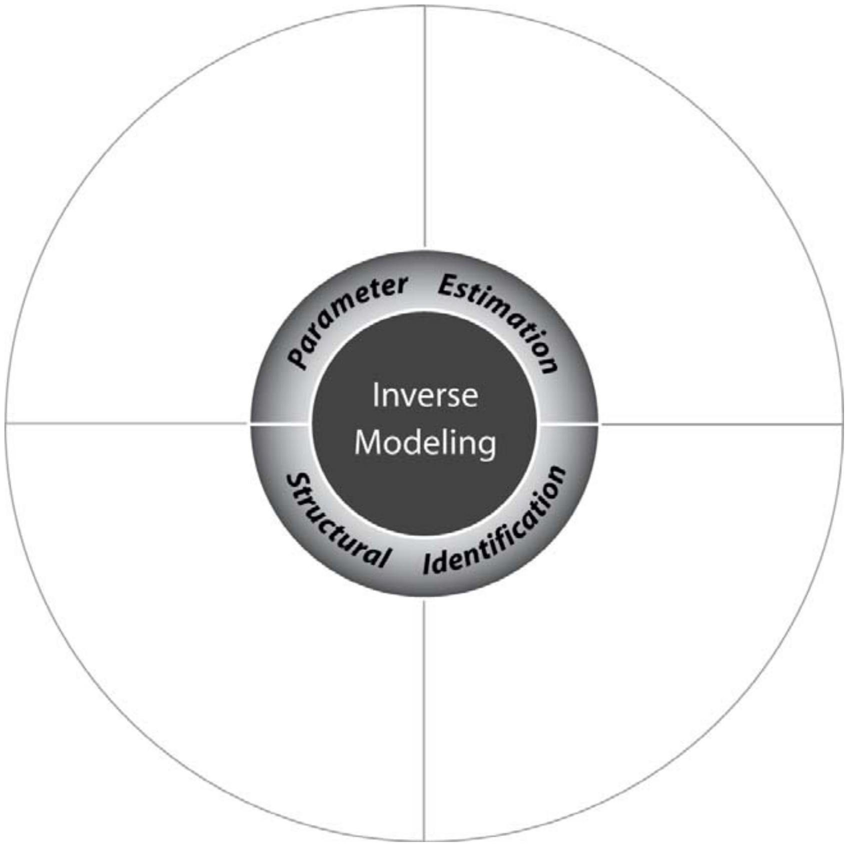
Recent Methods for Parameter Estimation in BST: ~ 100 papers; no method really good



Challenges of Inverse Modeling

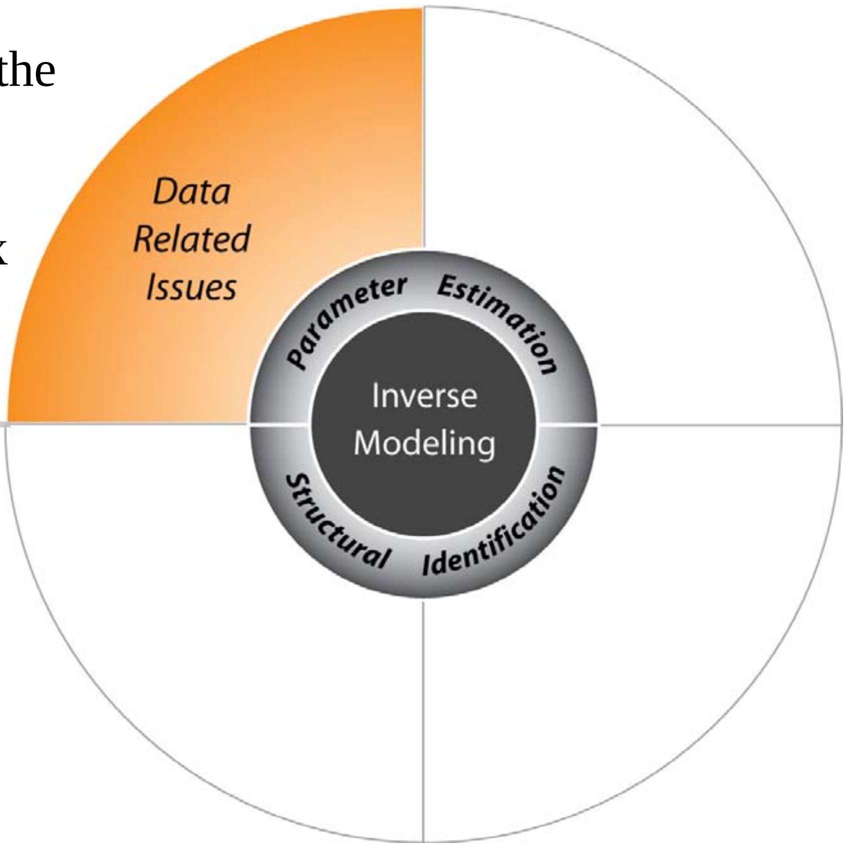


Challenges of Inverse Modeling



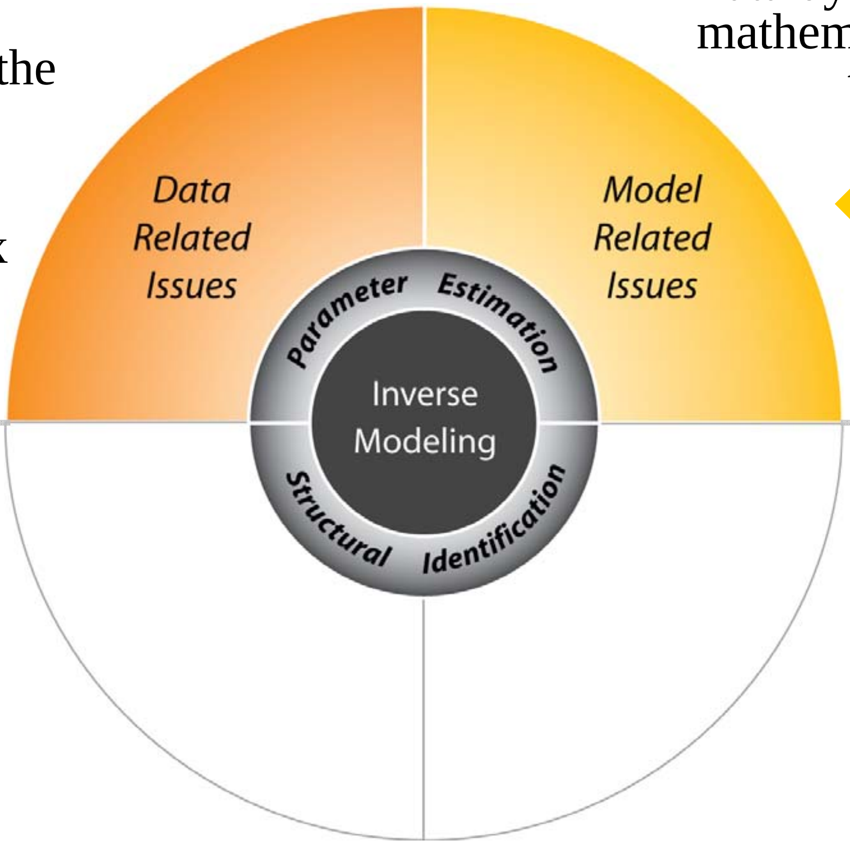
Challenges of Inverse Modeling

- Overly noisy data
- Missing data points
- Uncertainties about the measurements
- Non-informative
- Ill-posed data matrix



Challenges of Inverse Modeling

- Overly noisy data
- Missing data points
- Uncertainties about the measurements
- Non-informative
- Ill-posed data matrix

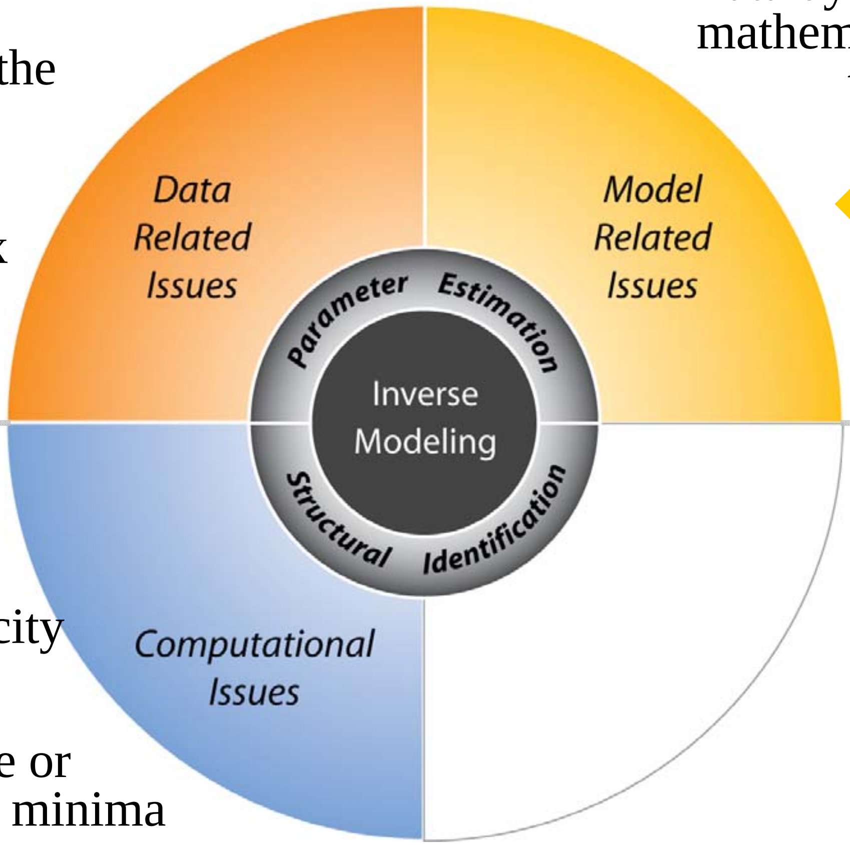


- Model selection criteria: Data dynamics capture ability, mathematical simplicity, tractability, results interpretability
- Infinite variety of formulations

Challenges of Inverse Modeling

- Overly noisy data
- Missing data points
- Uncertainties about the measurements
- Non-informative
- Ill-posed data matrix

- Model selection criteria: Data dynamics capture ability, mathematical simplicity, tractability, results interpretability
- Infinite variety of formulations

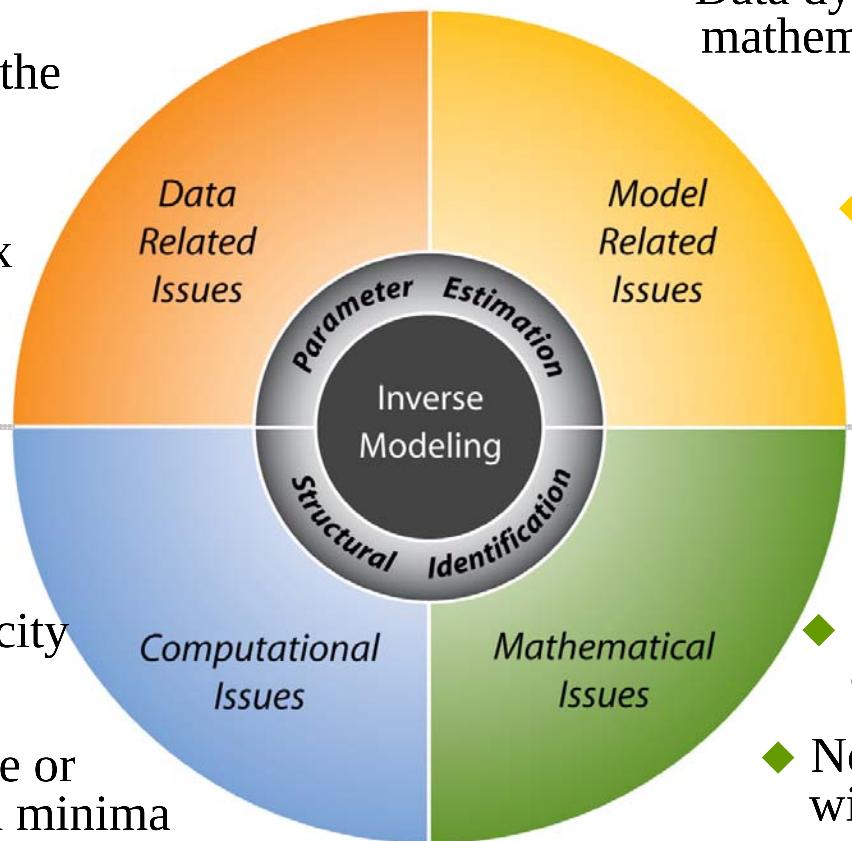


- Computational capacity
- Slow convergence
- Lacking convergence or convergence to local minima
- Time consuming for integration of differential equations

Challenges of Inverse Modeling

- Overly noisy data
- Missing data points
- Uncertainties about the measurements
- Non-informative
- Ill-posed data matrix

- Model selection criteria: Data dynamics capture ability, mathematical simplicity, tractability, results interpretability
- Infinite variety of formulations



- Computational capacity
- Slow convergence
- Lacking convergence or convergence to local minima
- Time consuming for integration of differential equations

- Distinctly different yet equivalent solutions
- Non-equivalent solutions with similar error
- Error compensation

Old Trick: Slope Estimation
(at least as old as Voit & Savageau, 1982)

$$S(t_k) \approx \dot{X}\Big|_{t_k} = f(X(t_k))$$

$$\vdots$$

$$S_i(t_j) \approx f_i(X_1(t_j), X_2(t_j), \ldots., X_n(t_j); p_{i1}, \ldots, p_{iM_i})$$

$$\vdots$$

S-System: $f_i \approx \alpha_i X_1^{g_{i1}} X_2^{g_{i2}} \ldots X_n^{g_{in}} - \beta_i X_1^{h_{i1}} X_2^{h_{i2}} \ldots X_n^{h_{in}}$

$$S_i \approx \alpha_i X_1^{g_{i1}} X_2^{g_{i2}} \ldots X_n^{g_{in}} - \beta_i X_1^{h_{i1}} X_2^{h_{i2}} \ldots X_n^{h_{in}} \quad at \quad t_k$$

Toward a New Trick

$$S_i \approx \alpha_i X_1^{g_{i1}} X_2^{g_{i2}} \ldots X_n^{g_{in}} - \beta_i X_1^{h_{i1}} X_2^{h_{i2}} \ldots X_n^{h_{in}} \quad at \quad t_k$$

estimated from data

...

measured

Terms become Numbers

Guess β_i and h_{ij}

New Trick: Alternating Regression

$$S_i \approx \alpha_i X_1^{g_{i1}} X_2^{g_{i2}} \ldots X_n^{g_{in}} - \beta_i X_1^{h_{i1}} X_2^{h_{i2}} \ldots X_n^{h_{in}} \qquad at \quad t_k$$

$$S_i - \beta_i X_1^{h_{i1}} X_2^{h_{i2}} \ldots X_n^{h_{in}} = \alpha_i X_1^{g_{i1}} X_2^{g_{i2}} \ldots X_n^{g_{in}} \qquad at \quad t_k$$

$$Number = \alpha_i X_1^{g_{i1}} X_2^{g_{i2}} \ldots X_n^{g_{in}} \qquad at \quad t_k$$

$$\log(Number) = \log(\alpha_i) + \sum g_{ij} \log(X_i) \qquad for \quad all \quad t_k$$

Linear regression yields $\hat{\alpha}_i$ and $\hat{g}_{ij}$

Alternating Regression (cont'd)

$$S_i \approx \alpha_i X_1^{g_{i1}} X_2^{g_{i2}} \ldots X_n^{g_{in}} - \beta_i X_1^{h_{i1}} X_2^{h_{i2}} \ldots X_n^{h_{in}} \quad at \quad t_k$$

Use $\hat{\alpha}_i$ and $\hat{g}_{ij}$ and compute "α-term"

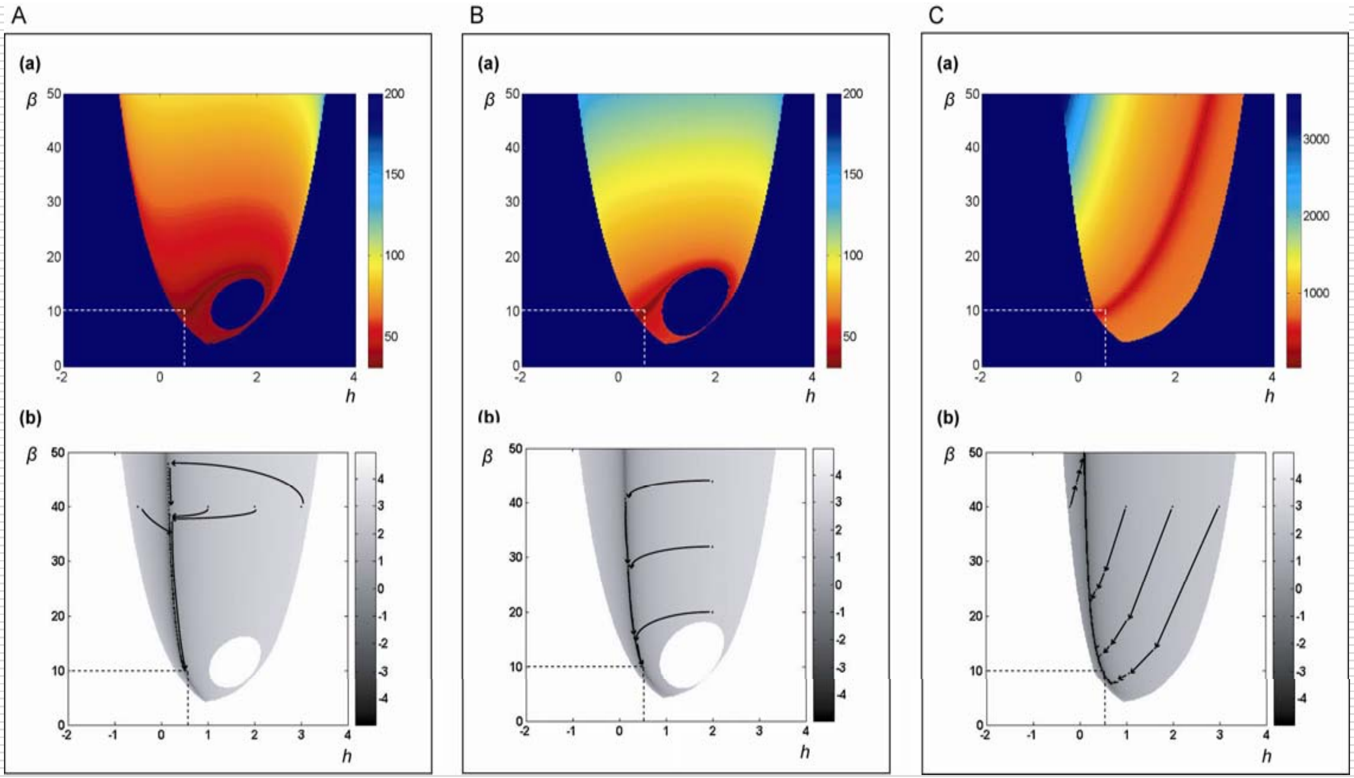
Merge the numerical value of the α-term with S_i and compute $\hat{\beta}_i$ and $\hat{h}_{ij}$ per linear regression for all time points.

Iterate between α - and β - terms until convergence

Alternating Regression (cont'd)

Results:

Extremely fast, if it converges.
Convergence issue very complex.



Problems with Traditional Methods

Time to (global) convergence

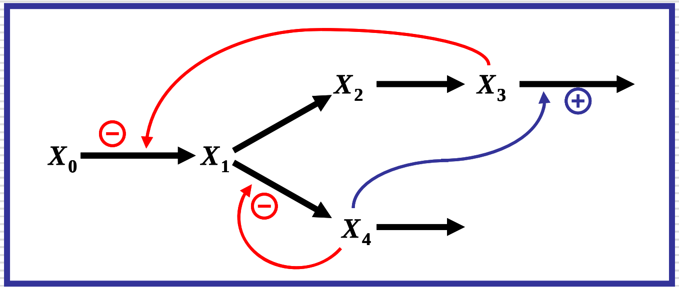
Problems with collinear data

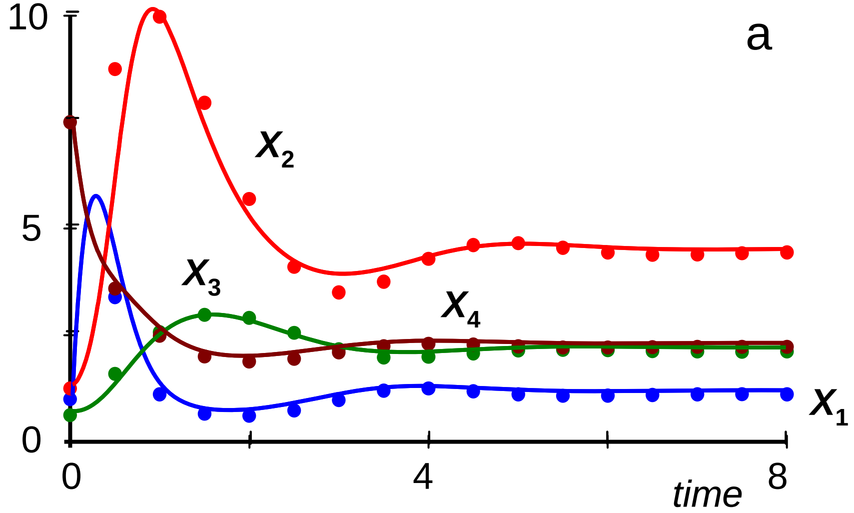
Problems with models permitting redundancies

Problems with compensation of error among terms

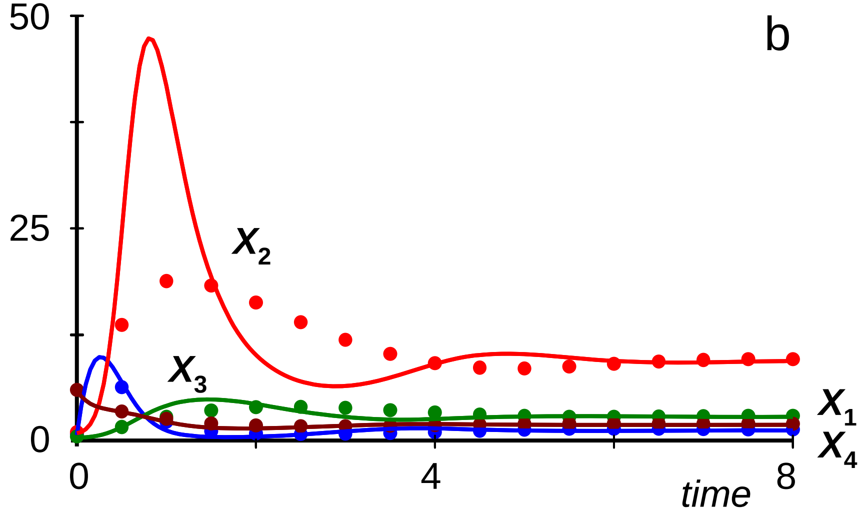
Problems with Traditional Methods: Extrapolation

Former model;
here using GMA form



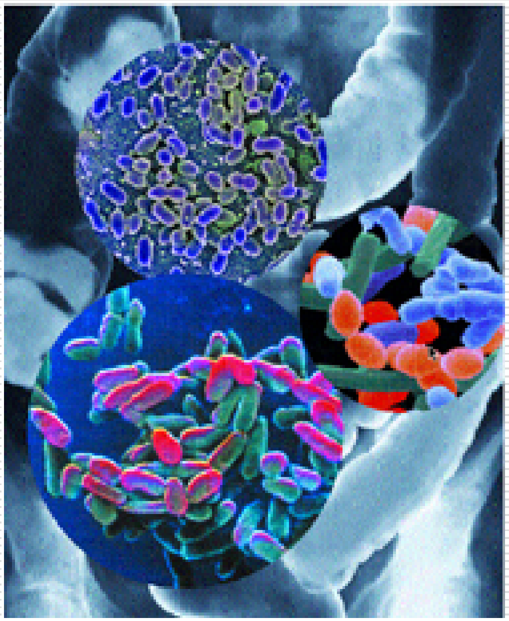


Bad parameters, but good fits because of error compensation



Problem with the “misestimated” system during extrapolation

Example: Regulation of Glycolysis in *Lactococcus lactis*



Bacteria found in yogurt and cheese:
Lactococcus lactis (top),
Lactobacillus bulgaricus (blue),
Streptococcus thermophilus (orange),
Bifidobacterium spec (magenta).

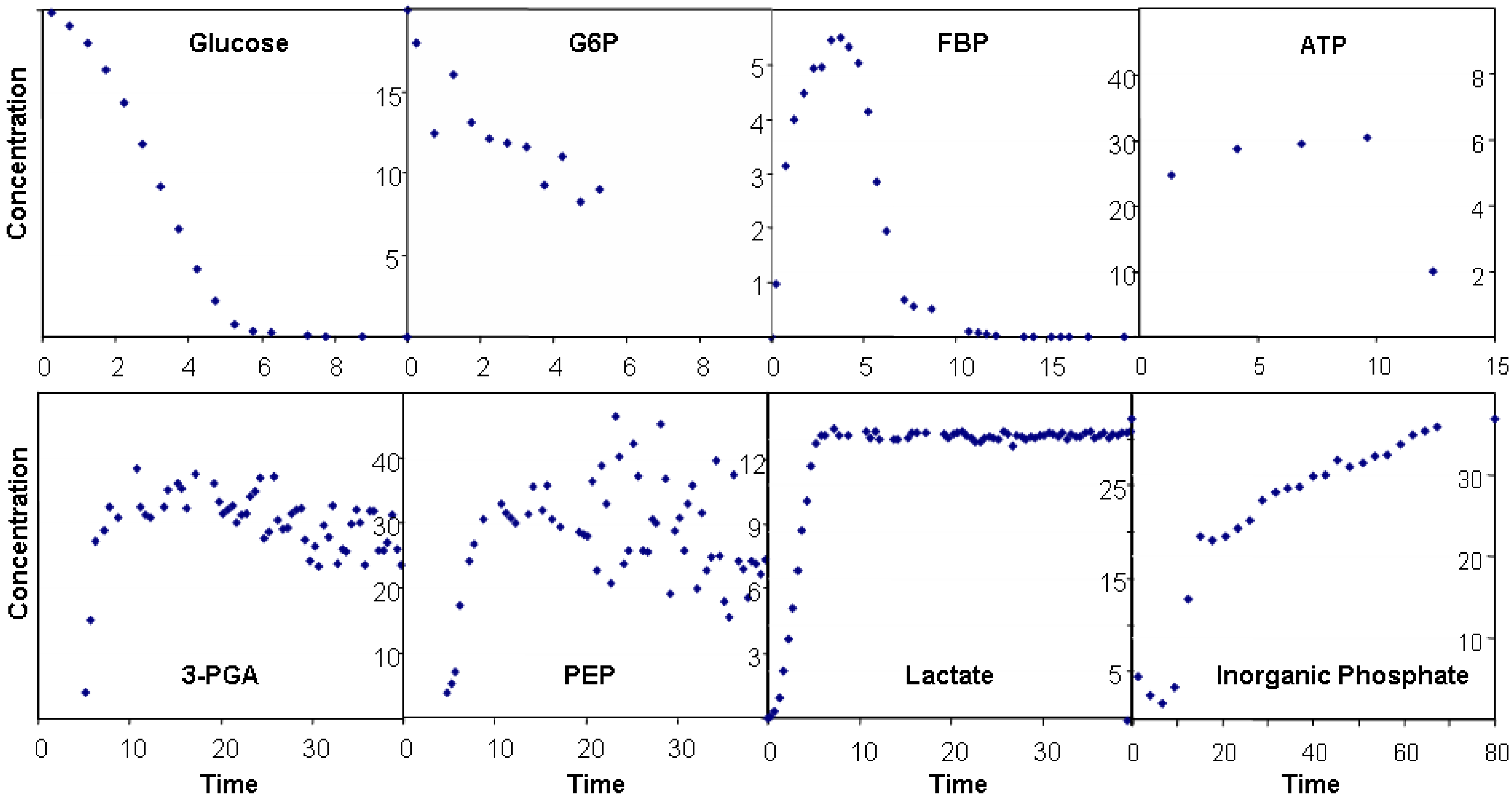
www.hhmi.org/bulletin/winter2005/images/bacteria5.jpg

Bacterium involved in dairy, wine, bread, pickle production.
Relatively simple organization. Here: study glucose regulation.

Goals of Modeling

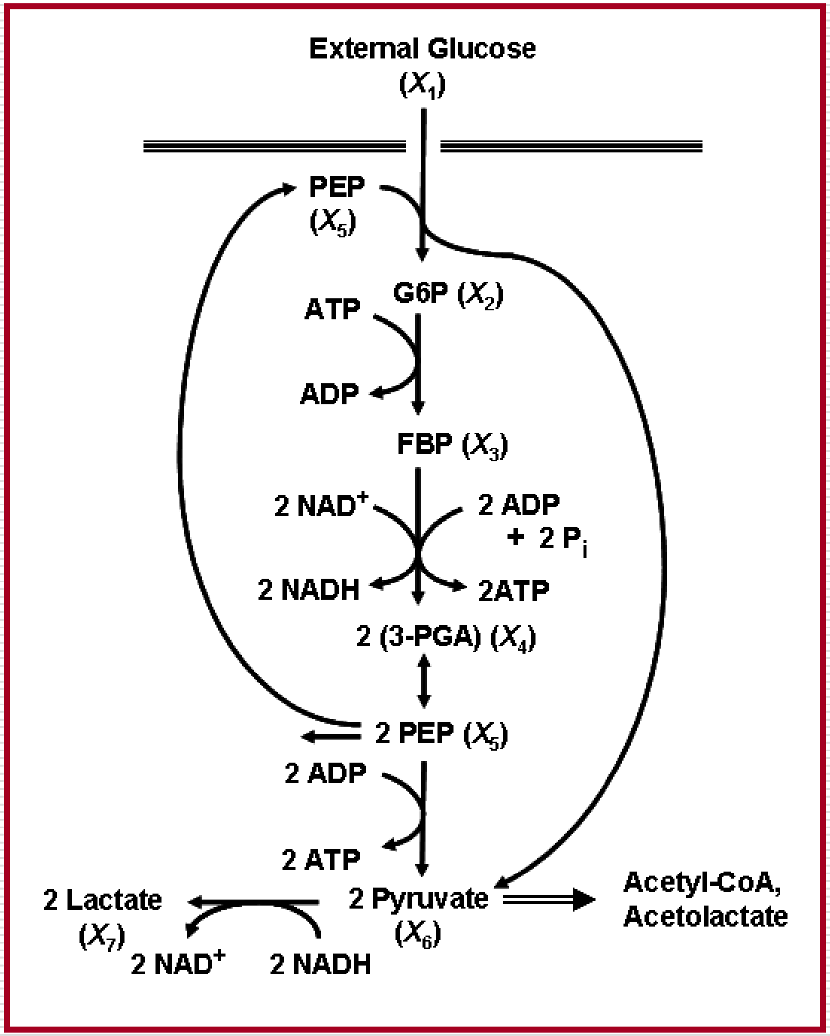
- Understand pathway; design, operation
- Allow extrapolation to new situations
- Allow prediction for manipulation
- Maximize yield of main product
- Optimize yield of secondary products
- Eventually develop a cell-wide model

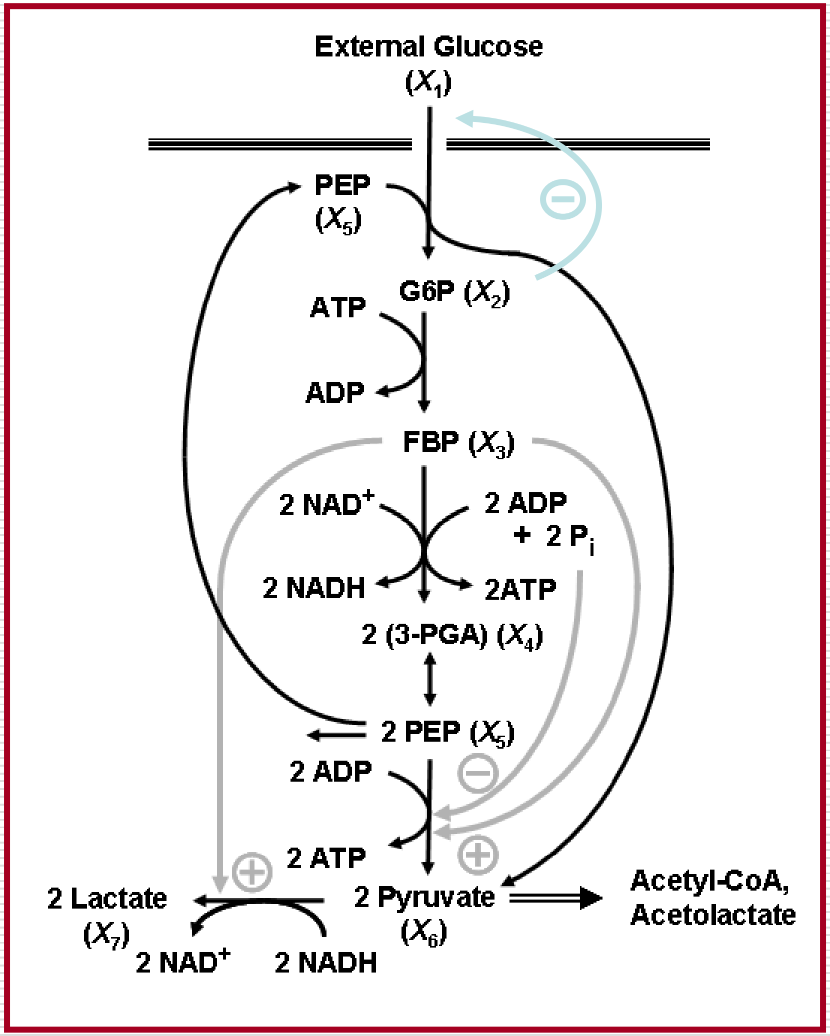
Experimental Time Series Data



E.O. Voit, J.S. Almeida, S. Marino, R. Lall, G. Goel, A.R. Neves, and H. Santos: *IEE Proc. Systems Biol. 2006*

Other Information





Lactococcus Data

Had modeled these data before

First, difficult to find any solutions

Combination of methods led to good fit

Later, many rather different solutions

Question: Is any of these solutions optimal?

Question: Is the BST model appropriate?

Problems with extrapolation

Dynamic Flux Estimation (DFE)

Inspired by Stoichiometric and Flux Balance Analysis

Extended to dynamic time courses

Study flux balance at each time point

Change in variable @ t = all influxes @ t – all effluxes @ t

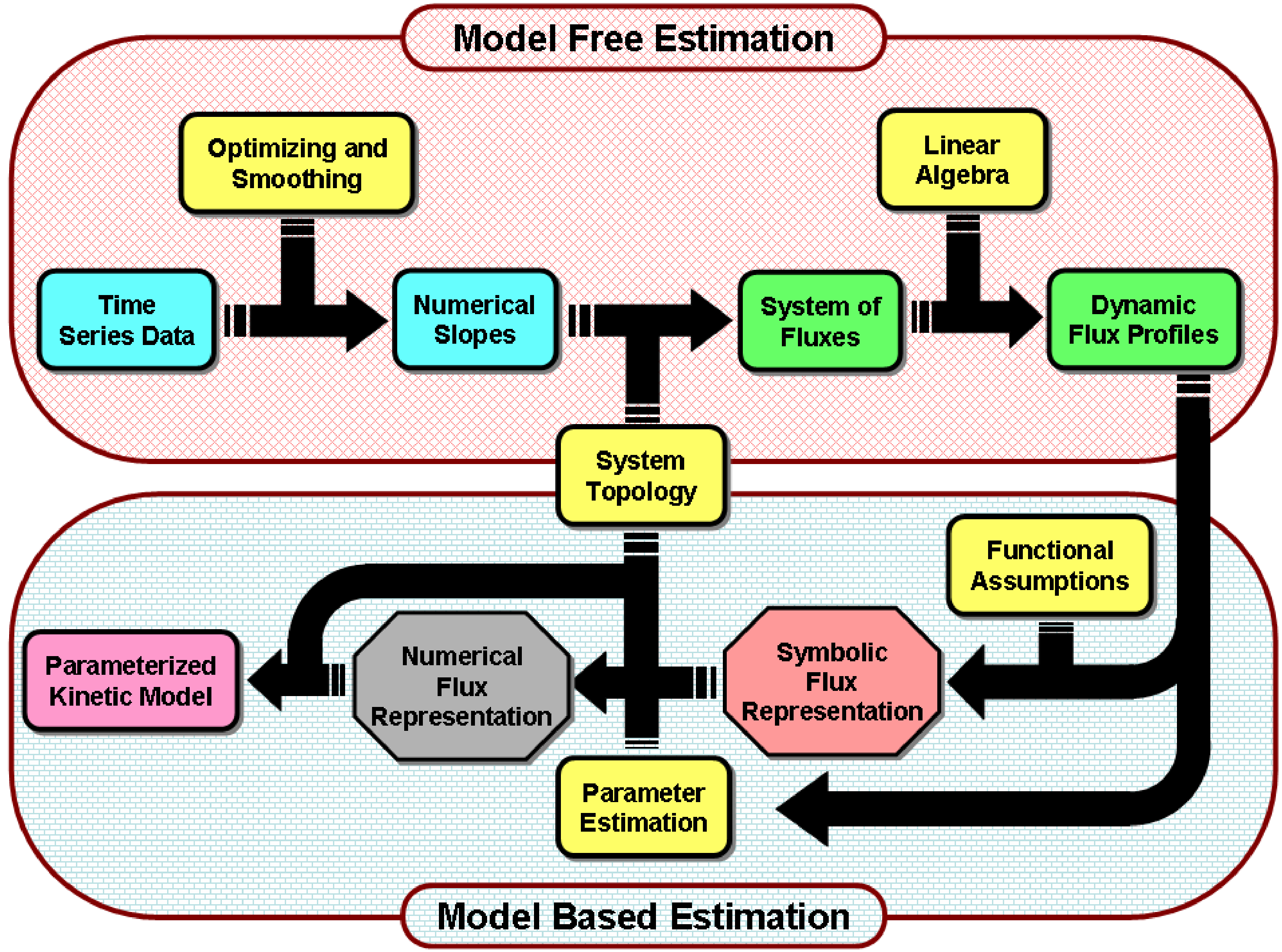
Linear system; solve as far as possible

Result: values of each flux @ t

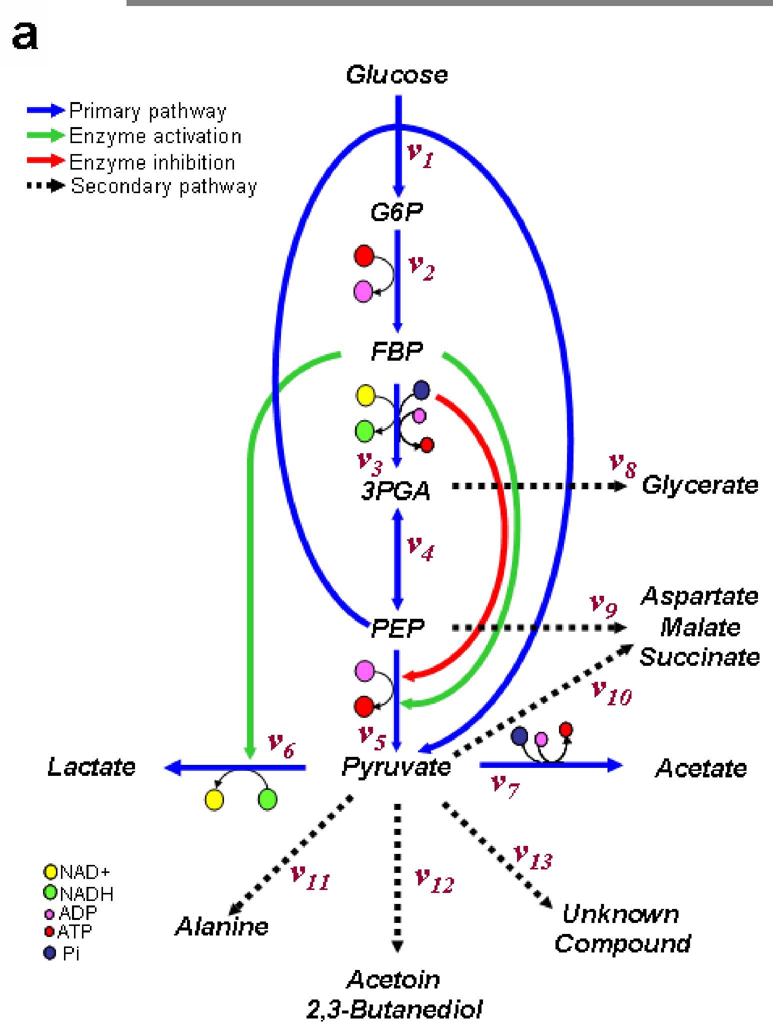
Represent fluxes with appropriate models

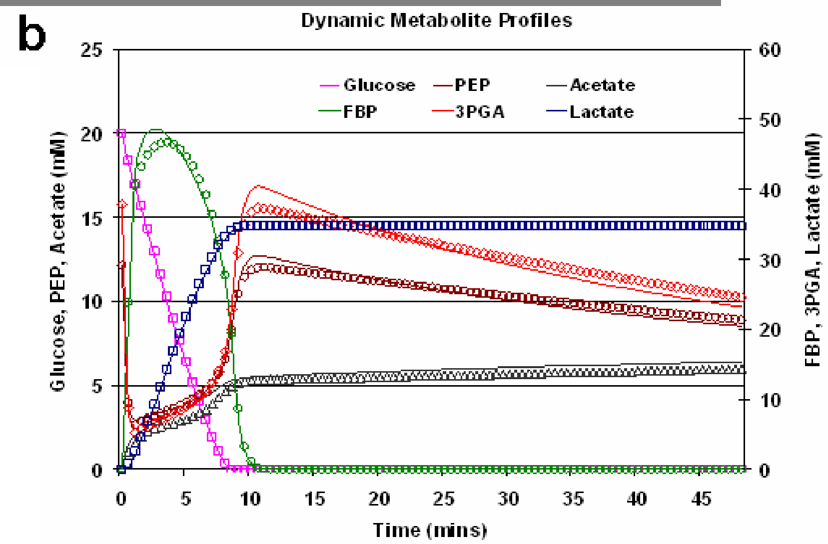
G. Goel *et al.*, *Bioinformatics 2008*

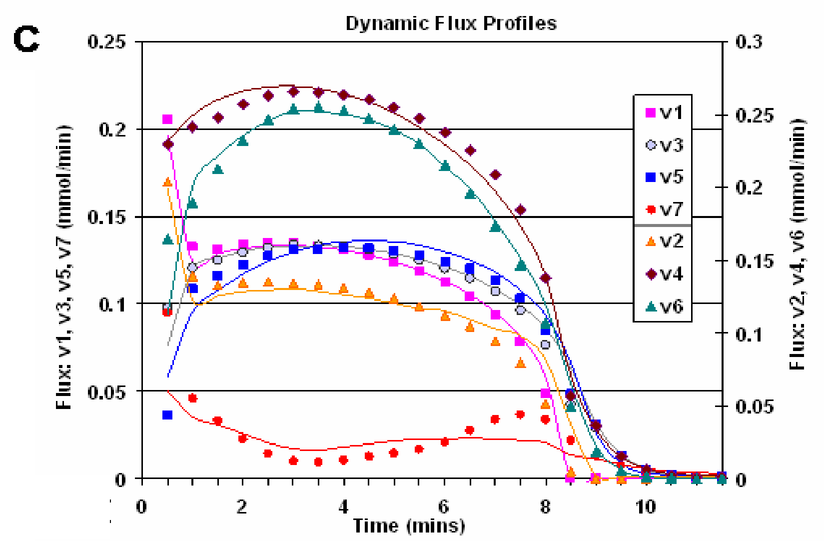
Dynamic Flux Estimation (DFE)



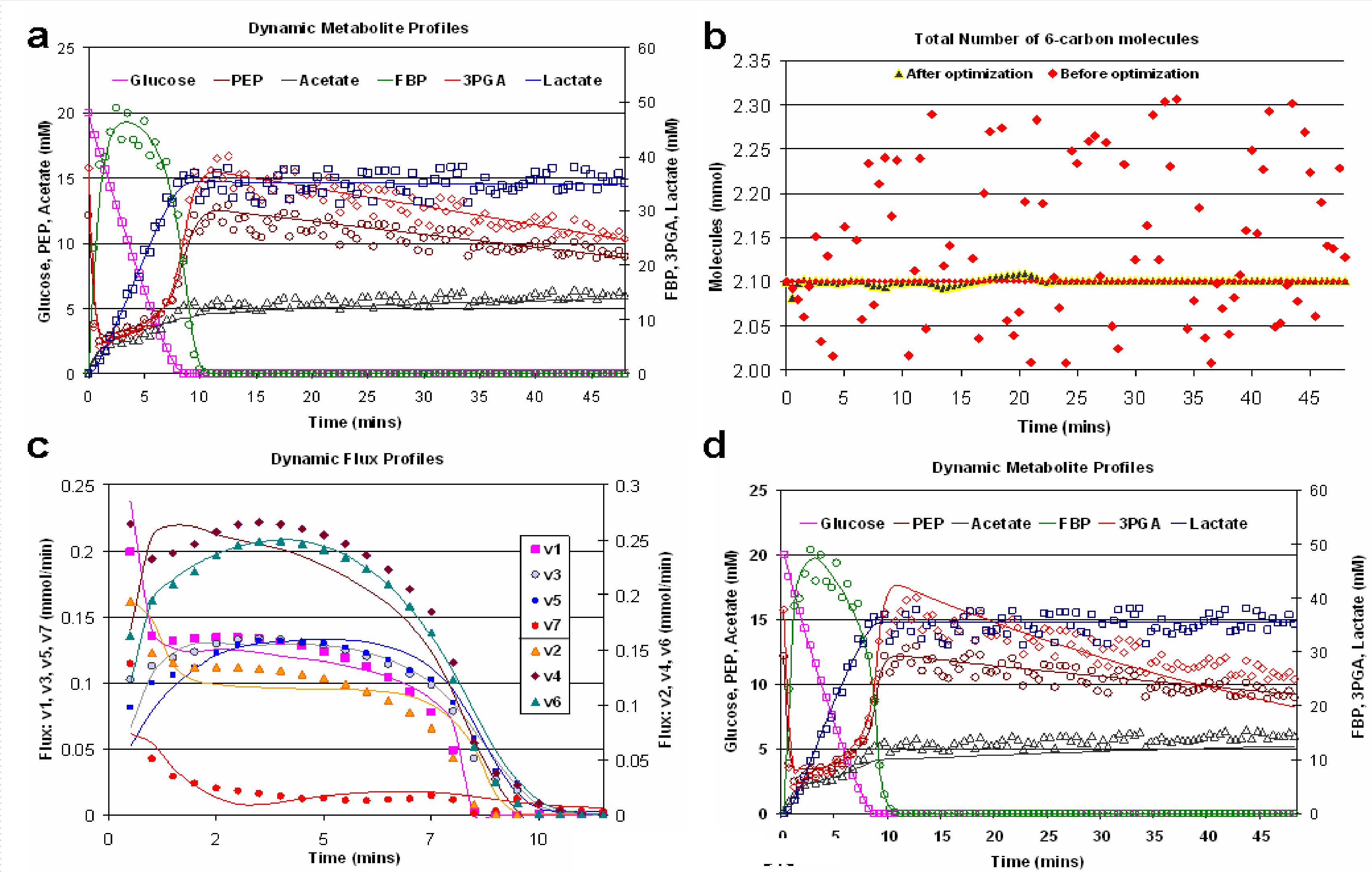
Dynamic Flux Estimation (DFE)



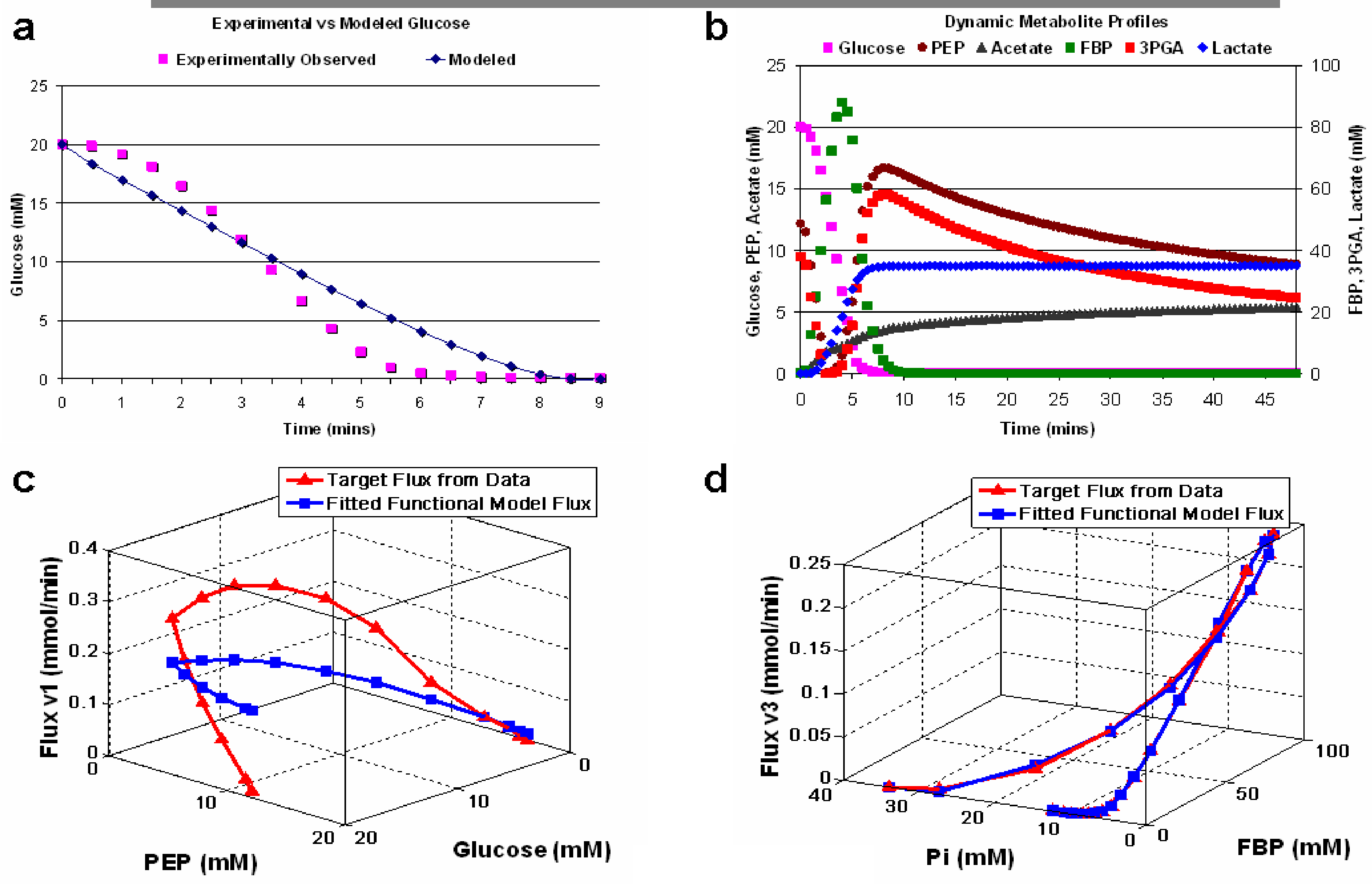




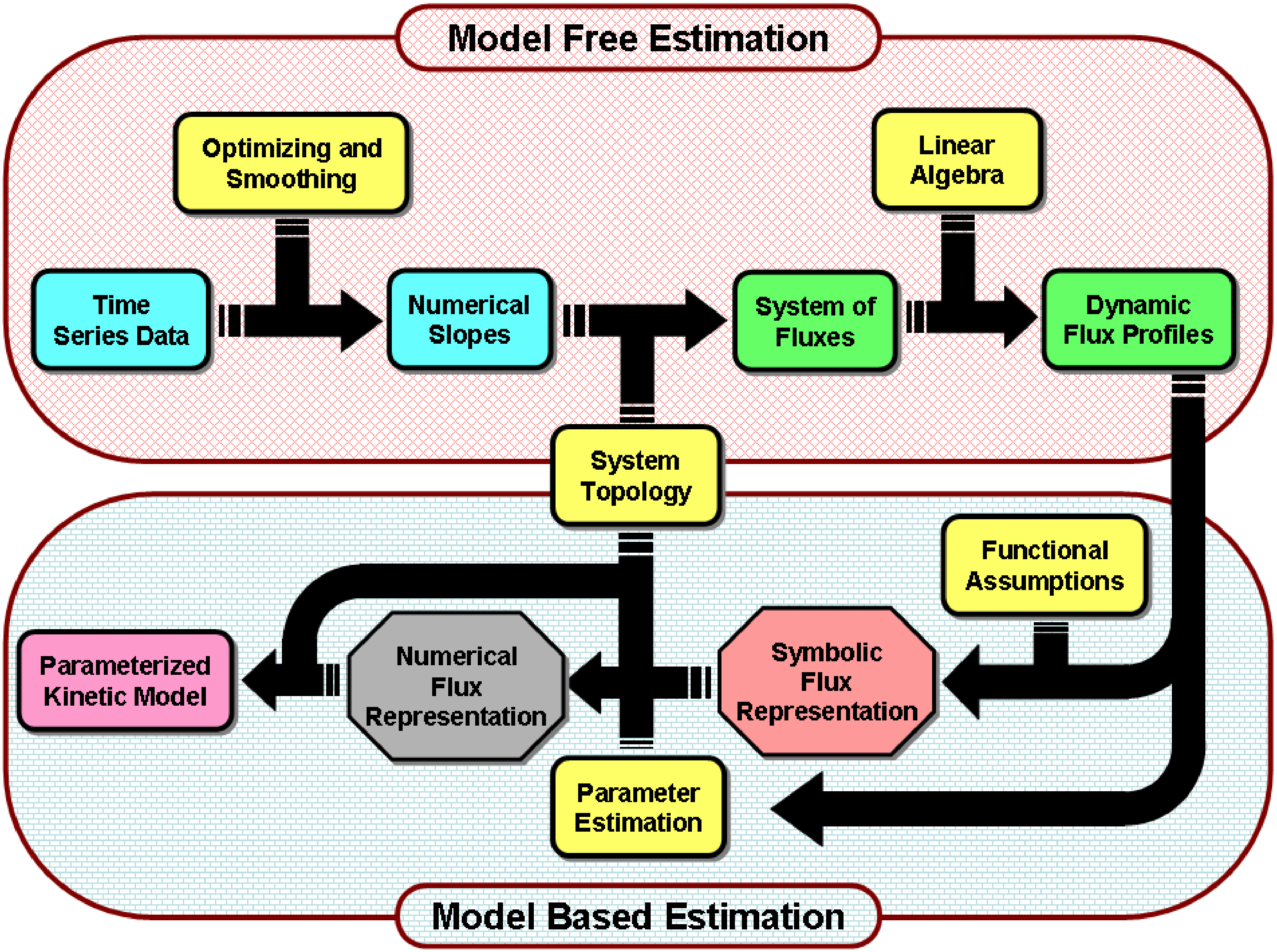
Dynamic Flux Estimation (DFE)



Dynamic Flux Estimation (DFE)



Dynamic Flux Estimation (DFE)



Open Problems

Smoothing and Mass conservation:

Noise in the data leads to loss or gain of mass

Underdetermined Flux Systems:

Linear system of flux often not of full rank

Augment DFE with other methods

(e.g., AR or bottom-up estimation)

Characterization of Redundancies:

Data collinear or non-informative (pooling?)

Model allows transformation groups (Lie analysis?)

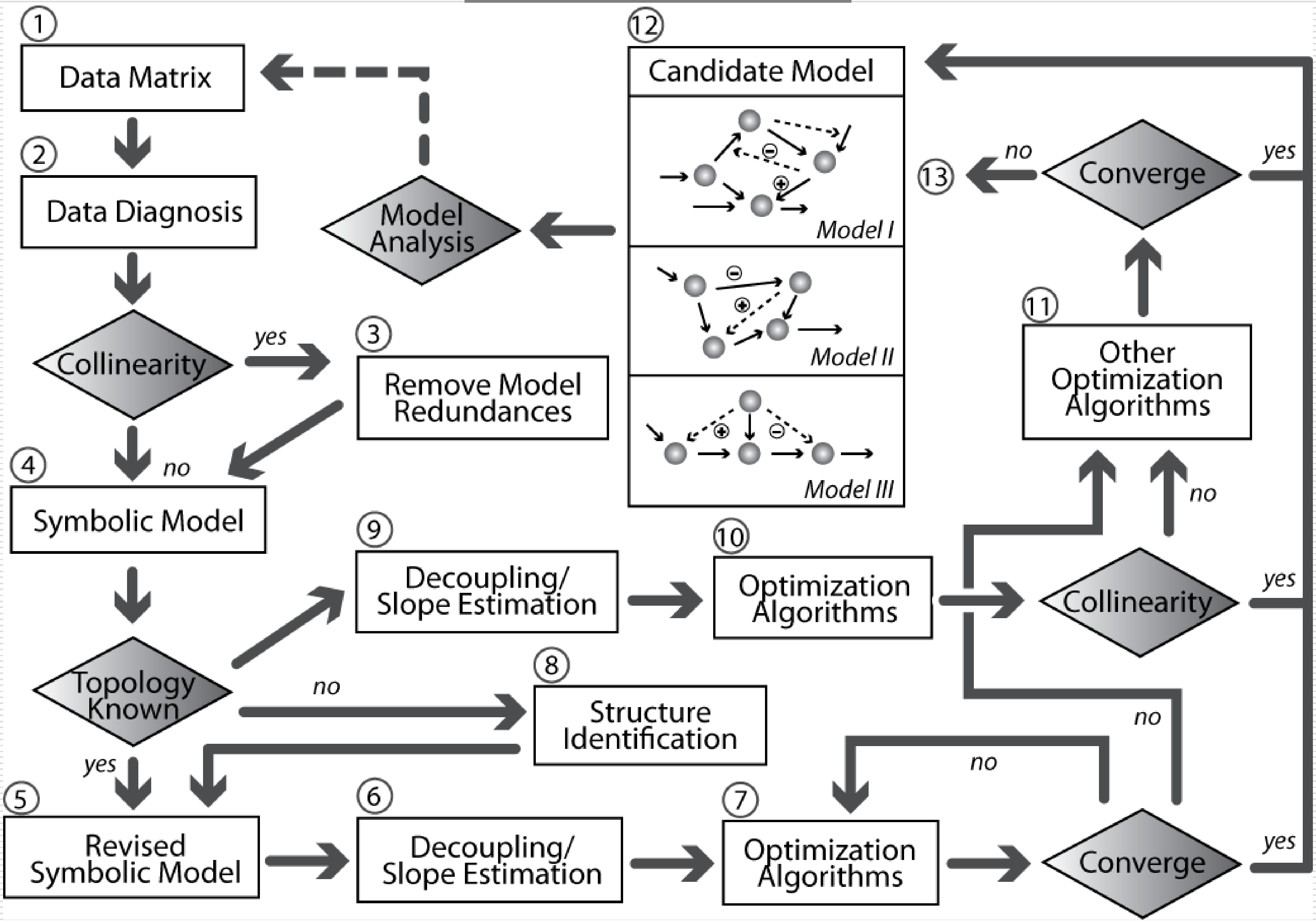
Overriding Challenge

Speed and Convenience

Algorithms for parameter estimation
from time series must become
much faster and more robust

They must run reliably and “semi-foolproof”
on ordinary PC’s without the need
of expensive software

Workflow



1
Data Matrix
2
Data Diagnosis
Collinearity
yes
3
Remove Model Redundances
no
4
Symbolic Model
Topology Known
no
8
Structure Identification
yes
5
Revised Symbolic Model
6
Decoupling/ Slope Estimation
7
Optimization Algorithms
Converge
yes
no
no
9
Decoupling/ Slope Estimation
10
Optimization Algorithms
Collinearity
yes
no
11
Other Optimization Algorithms
Converge
no
13
yes
12
Candidate Model
Model I
Model II
Model III
Model Analysis

Summary

Efficiently dealing with inverse problems presents new modeling opportunities:

1. Time series data are coming! They contain a lot of implicit information that must be extracted.

2. Technical challenges abound. Important: Efficient, robust, and fast solutions on PC's needed.

3. Important overlooked issue: Error compensation; extrapolation becomes unreliable. DFE promising

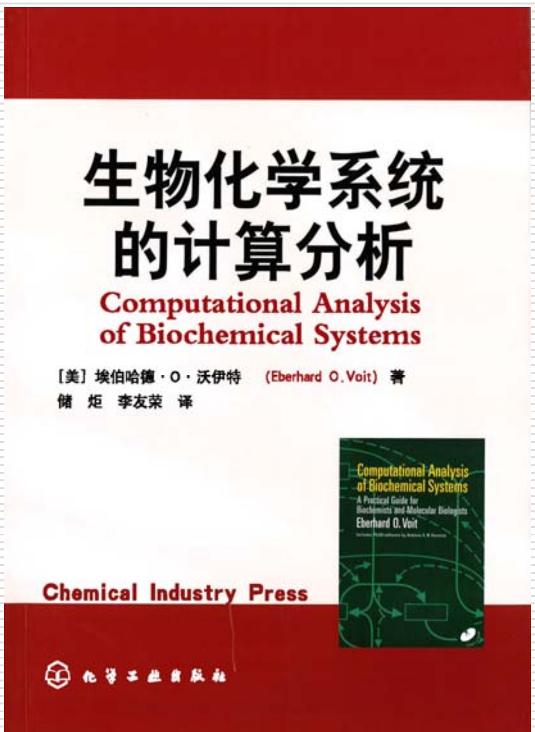
Acknowledgements

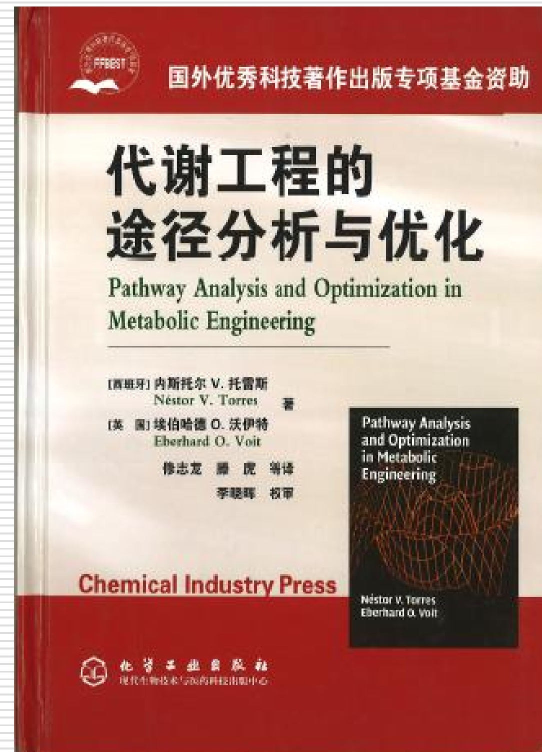
The Current Crew:

Funding: NIH, NSF, DOE, Woodruff Foundation, Georgia Research Alliance

Information: www.bst.bme.gatech.edu

Further Information





www.bst.bme.gatech.edu