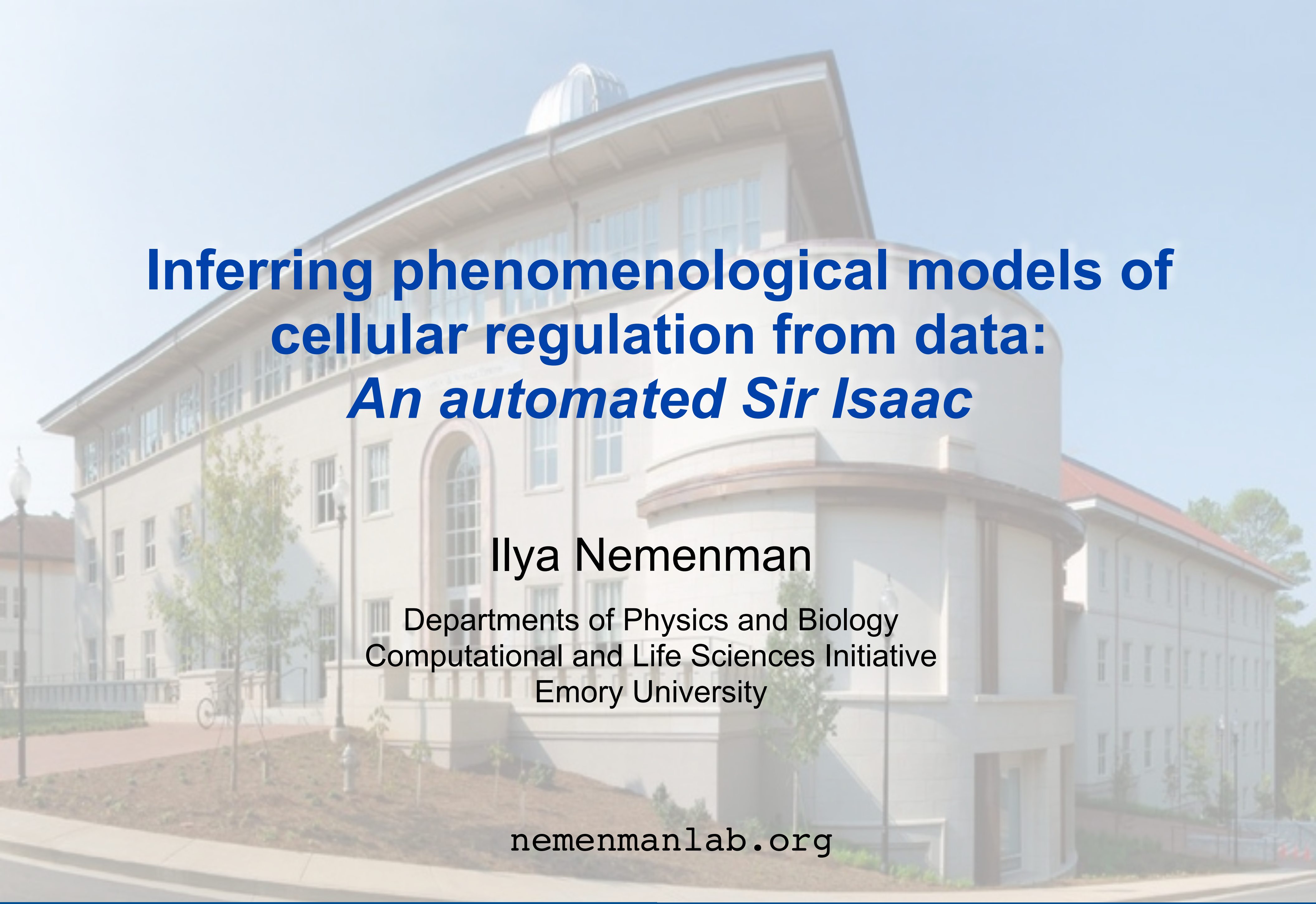


Inferring phenomenological models of cellular regulation from data: *An automated Sir Isaac*

Ilya Nemenman

Departments of Physics and Biology
Computational and Life Sciences Initiative
Emory University

nemenmanlab.org

EMORY
UNIVERSITY

Thanks

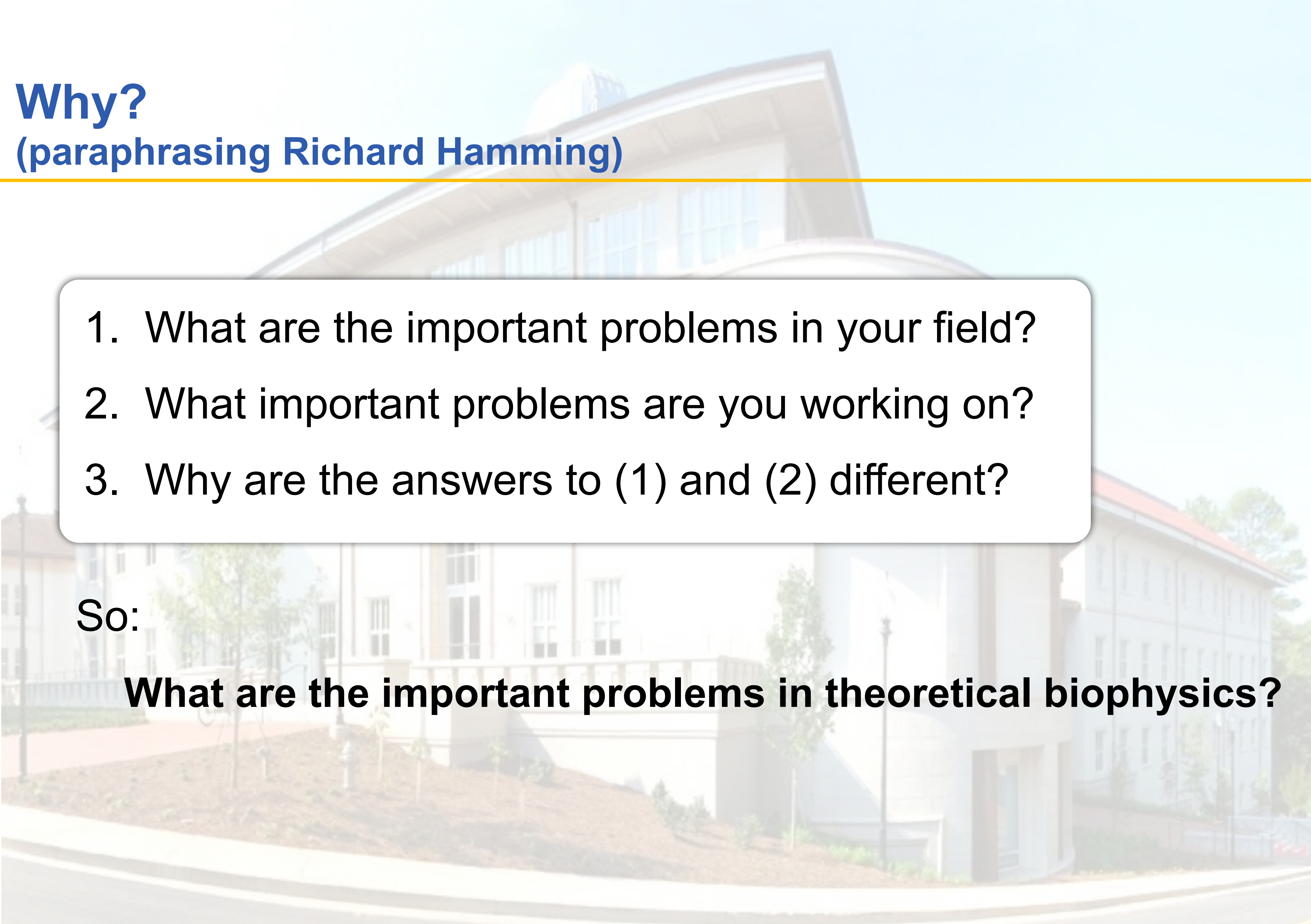
- Bryan Daniels (U Wisconsin, Madison)

JAMES S. MCDONNELL FOUNDATION

JOHN TEMPLETON FOUNDATION

Why?

(paraphrasing Richard Hamming)



1. What are the important problems in your field?
2. What important problems are you working on?
3. Why are the answers to (1) and (2) different?

So:

What are the important problems in theoretical biophysics?

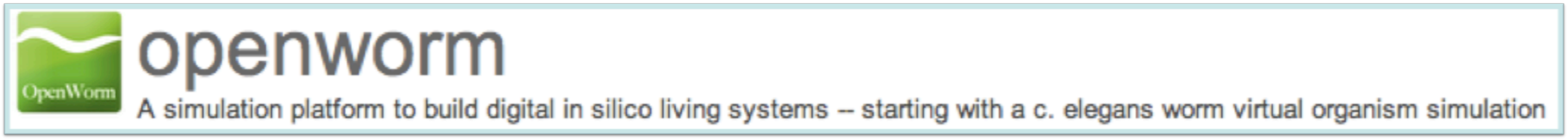
Of exactitude in science

…In that Empire, the craft of Cartography attained such Perfection that the Map of a Single province covered the space of an entire City, and the Map of the Empire itself an entire Province. In the course of Time, these Extensive maps were found somehow wanting, and so the College of Cartographers evolved a Map of the Empire that was of the same Scale as the Empire and that coincided with it point for point. Less attentive to the Study of Cartography, succeeding Generations came to judge a map of such Magnitude cumbersome, and, not without Irreverence, they abandoned it to the Rigours of sun and Rain. In the western Deserts, tattered Fragments of the Map are still to be found, Sheltering an occasional Beast or beggar; in the whole Nation, no other relic is left of the Discipline of Geography.

From Travels of Praiseworthy Men (1658) by J. A. Suarez Miranda (a fictional reference).
By Jorge Luis Borges and Adolfo Bioy Casares.
English translation quoted from J. L. Borges, A Universal History of Infamy,
Penguin Books, London, 1975.

At a recent meeting...

- Many have expressed an opinion that: “**The final** theory of biological systems will be a large multiscale computational model. We need more and more experimental data to specify details of these models.”



OpenWorm

openworm

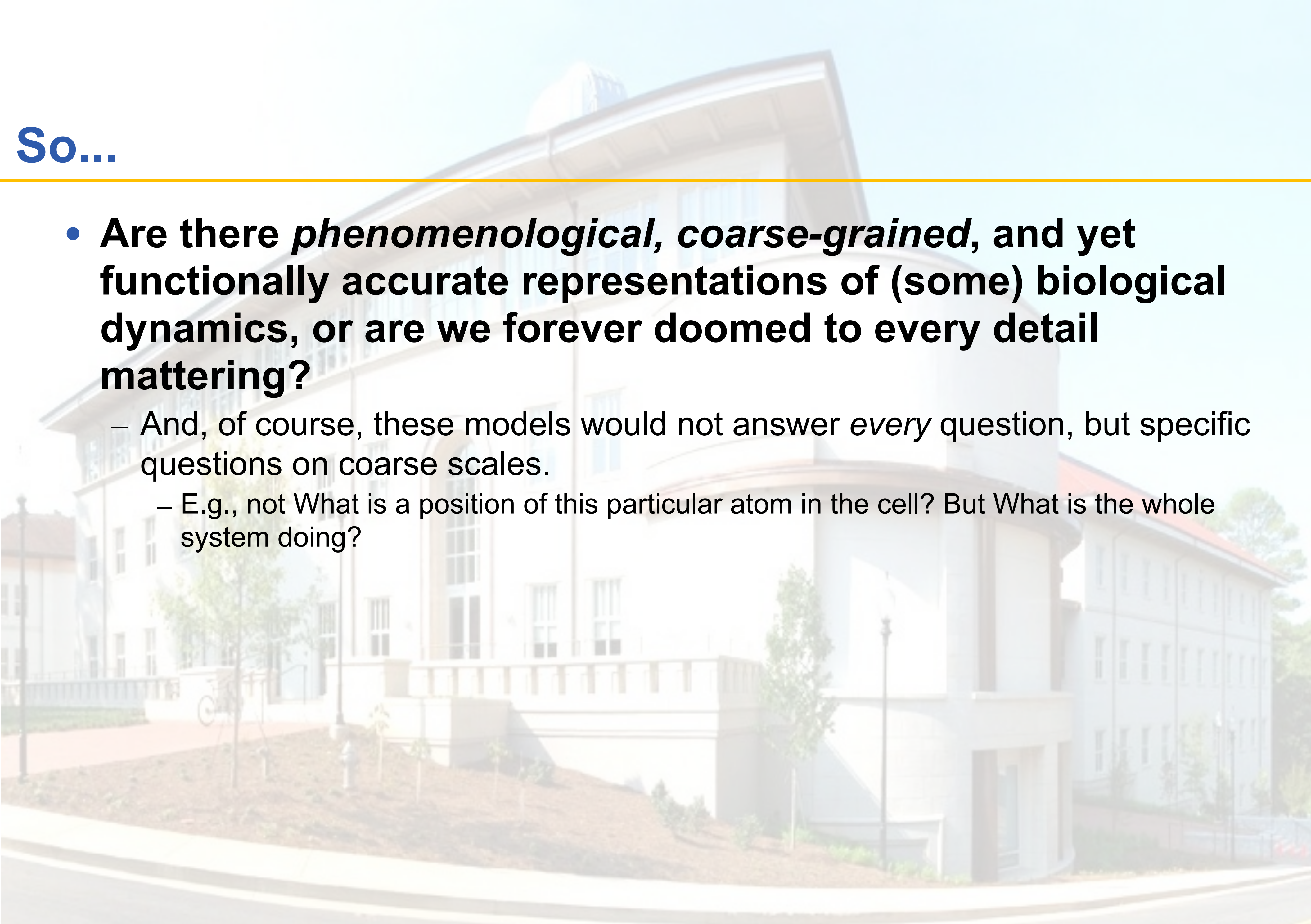
A simulation platform to build digital in silico living systems -- starting with a c. elegans worm virtual organism simulation

- There’s something wrong with this statement.
  - The “**final**” theory?
  - Do we need the theory of “**everything**” in any biological (or physical) system?

- The best material model of a cat is another, or preferably the same, cat.
(*Philosophy of Science*, Wiener and Rosenblueth, 1945)

Physics analogy

- What is the final, complete theory of the chair you are sitting in?
  - How does it fall from the second floor?
  - How does the cloth seat age and tear?
  - How much weight would the chair hold before it breaks?
  - How does it conduct electricity?
  - How much food can I cook when I burn it?
  - …
- There’s no such thing as “the full theory of the chair”.
  - We build models tailored to answer **specific questions**.
  - The complete theory that answers **every** question would need to include quarks, superstrings…
  - Each modeling level needs its own **effective** degrees of freedom
    - “Don’t model bulldozers with quarks.” (Goldenfeld and Kadanoff, 1999)
- Models must loose details. Otherwise, just use the same cat…

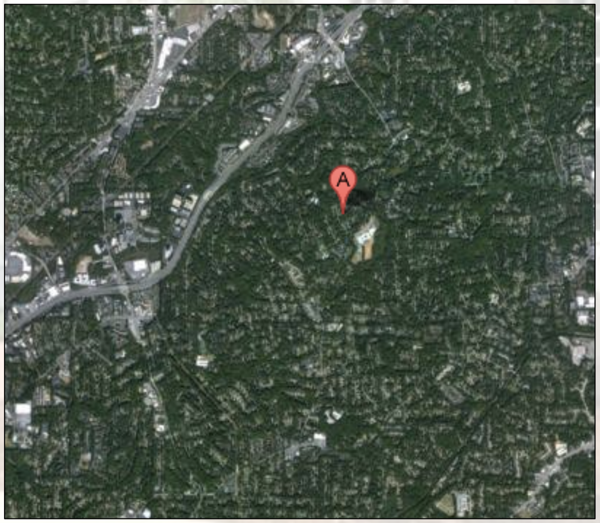


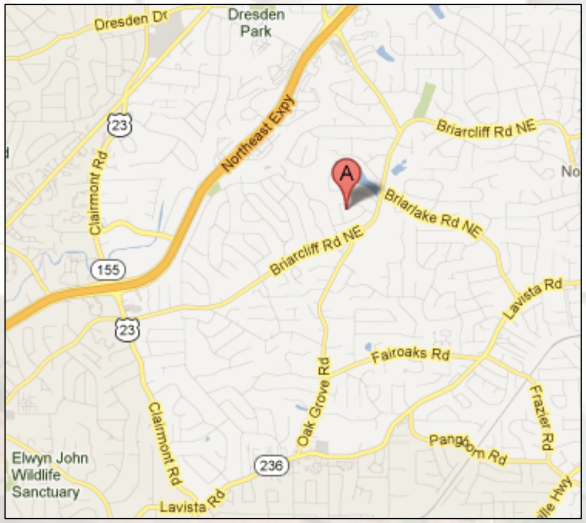
So...

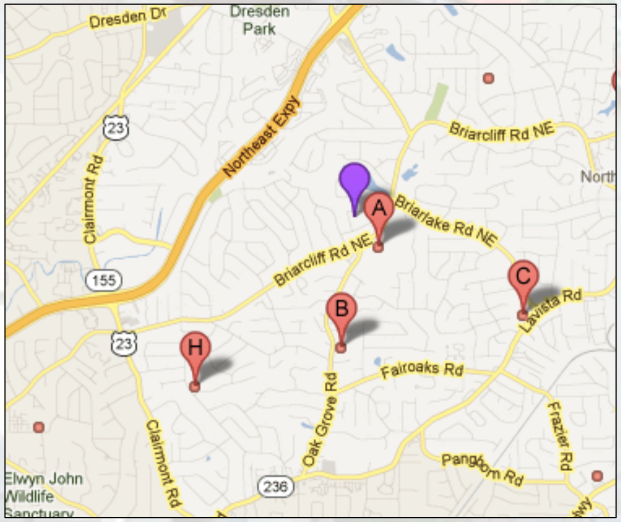
- **Are there *phenomenological, coarse-grained*, and yet functionally accurate representations of (some) biological dynamics, or are we forever doomed to every detail mattering?**
  - And, of course, these models would not answer *every* question, but specific questions on coarse scales.
    - E.g., not What is a position of this particular atom in the cell? But What is the whole system doing?

How’s this done in physics?

- Many degrees of freedom + symmetries (either exact, or emerging on average)
  - Essentially, the **law of large numbers** produces universalities on large scales if the right questions are asked (not about a position of a certain atom, but about large-scale quantities).
- Unclear if such properties would emerge in many other biological systems.
- Still, asking the **right** questions may help even without symmetries
  - By not merely scaling, but throwing away details (different, depending on the specific question asked). **Which of the pictures is more useful for driving? for driving to a school?**







- **Many have worked on these, but I will focus on our approach.**

Cellular networks: complex beasts

A culture's icons are a window onto its soul. Few would disagree that, in the culture of molecular biology that dominated much of the life sciences for the last third of the 20th century, the dominant icon was the double helix. In the present, post-modern, 'systems biology' era, however, it is, arguably, the hairball.

A.D. Lander. BMC Biology 2010, 8:40

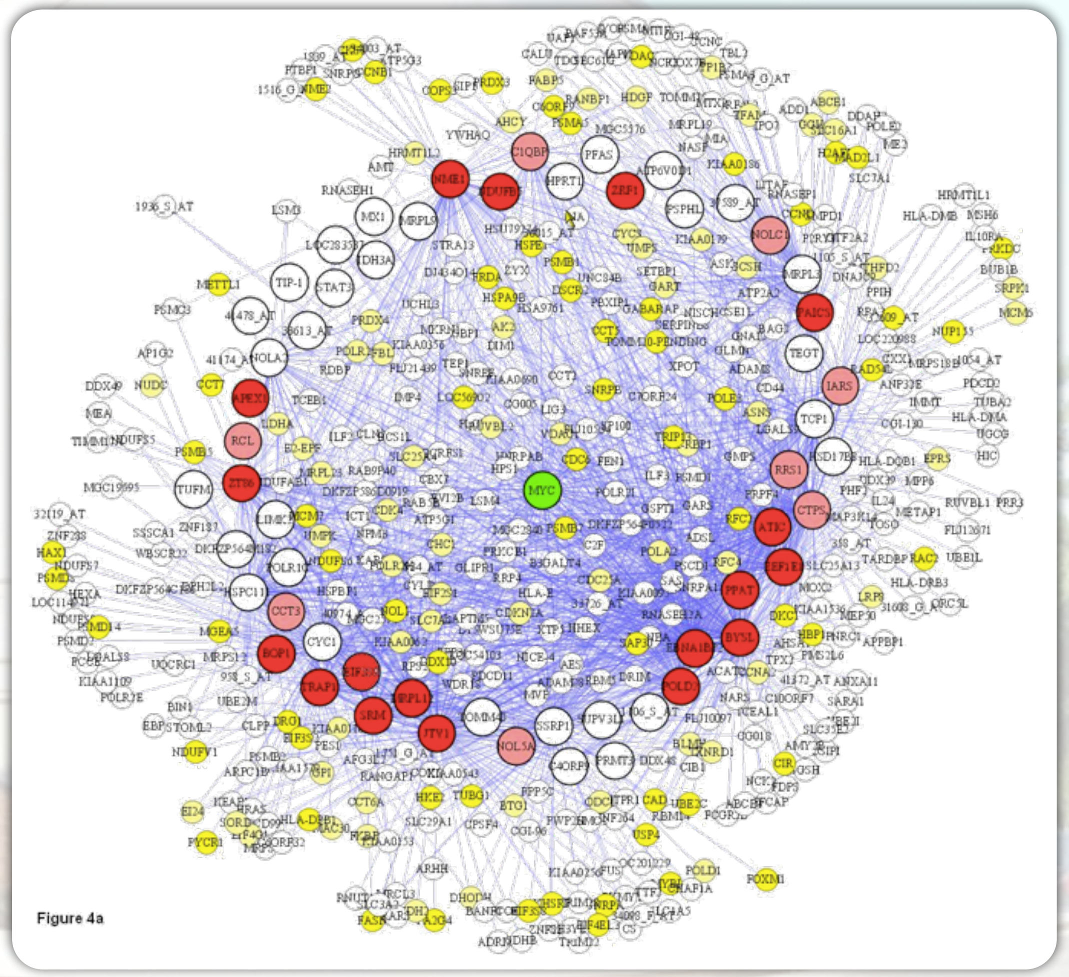
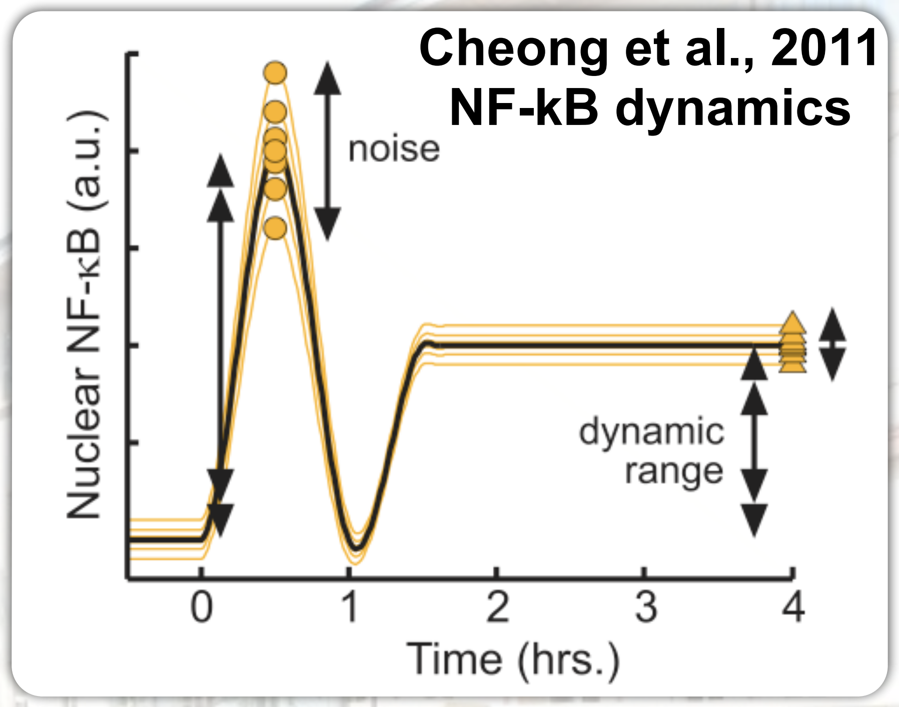


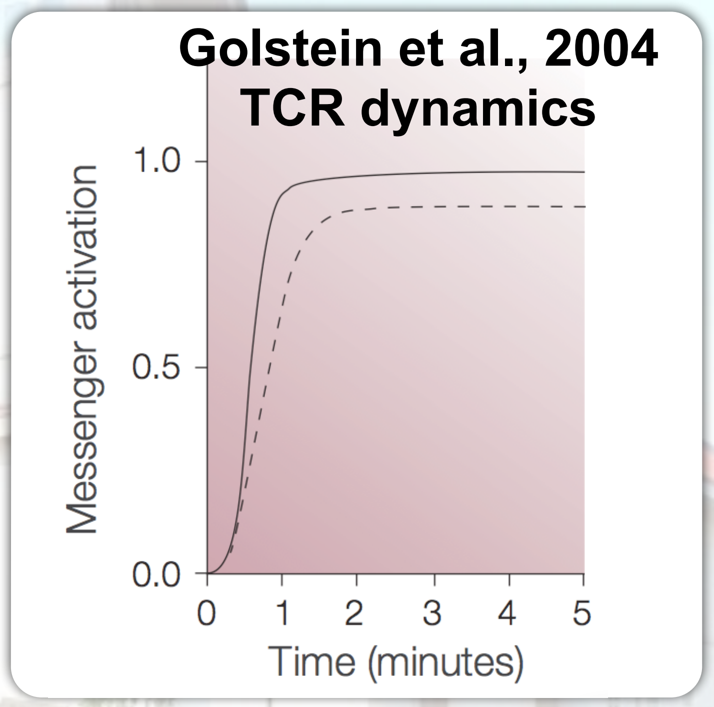
Figure 4a

Margolin et al., 2006

And yet, for typical inputs, their dynamics are rather simple



Cheong et al., 2011
NF-kB dynamics



Golstein et al., 2004
TCR dynamics

- A handful of parameters (time scales, amplitudes) describe responses of networks to most experimentally accessible perturbations.
- **Do we need complex networks to describe simple dynamics?**

We have worked on this problem for a while

- Bottom-up methods, reducing a known microscopic, mechanistic network.
  - with Sinitsyn et al., 2006-2010; with Munsky, Bel et al., 2009-2013; with Merchan et al., in prep., etc.
- Can we instead build **phenomenological models** top-down, from data directly, and without reconstructing a mechanistic network as an intermediate step?
  - Purpose: **predict responses** to exogenous signals.
  - Purpose: drive all of us, modelers, out of work?

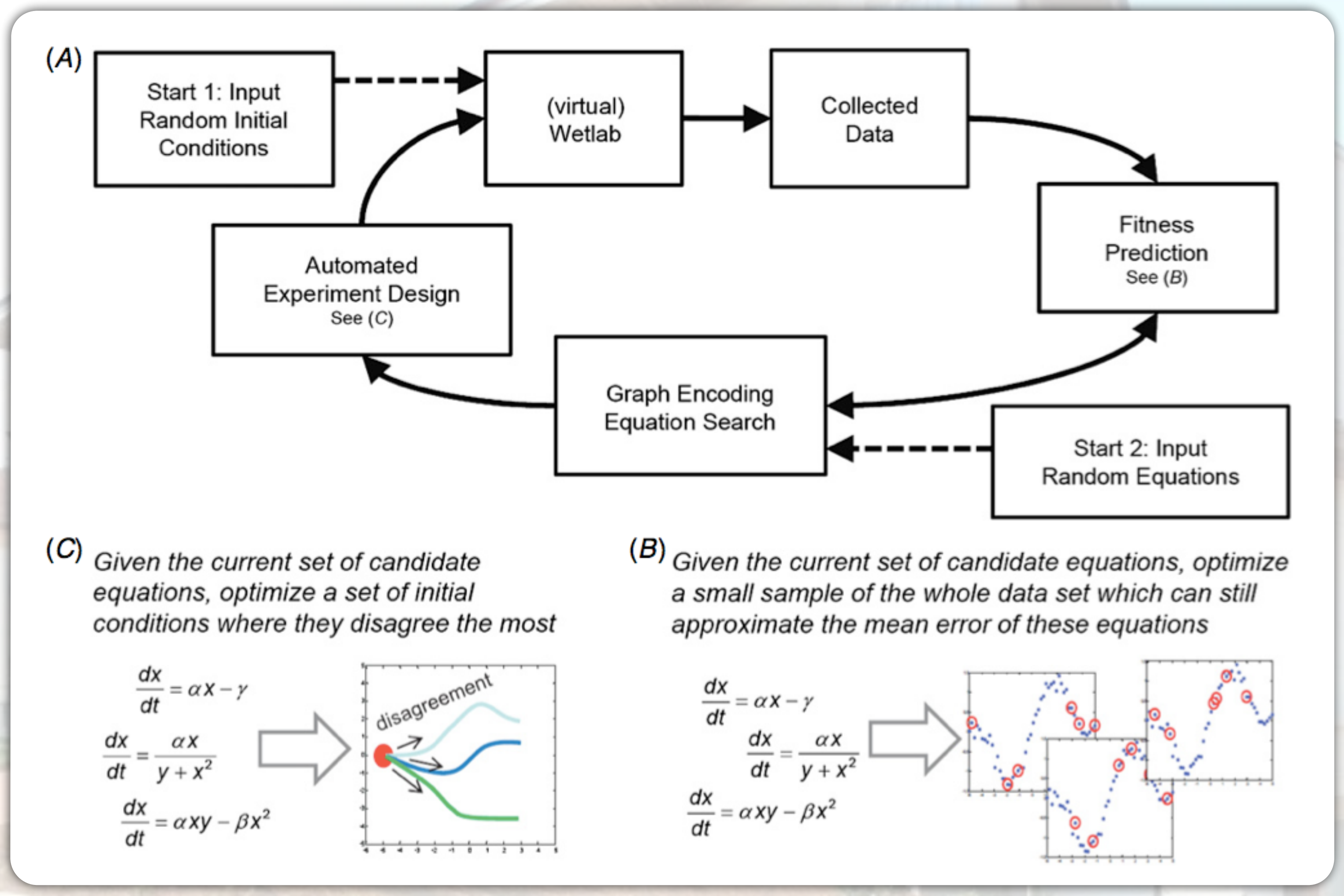
Can we fit simple, phenomenological models to biological data?

- We will assume that dynamics of cellular networks is given by local **ordinary differential equations**.
  - Do not fit curves; **fit dynamics**.
- We will neglect stochasticity, and spatial structure for now

$$\begin{cases} \dfrac{dx_1}{dt} = f_1(x_1, x_2, \ldots, x_n) \\ \cdots \\ \dfrac{dx_n}{dt} = f_n(x_1, x_2, \ldots, x_n) \end{cases}$$

- Can we automatically fit these functions f_i from data?
  - How do we enumerate the set of all possible multivariate functions?
  - How do we search through this list?
  - How do we not overfit?

Prior art:
EUREQA: full search approach



(A)

Start 1: Input Random Initial Conditions

(virtual) Wetlab

Collected Data

Fitness Prediction See (B)

Graph Encoding Equation Search

Start 2: Input Random Equations

Automated Experiment Design See (C)

(C) *Given the current set of candidate equations, optimize a set of initial conditions where they disagree the most*

$$\frac{dx}{dt} = \alpha x - \gamma$$

$$\frac{dx}{dt} = \frac{\alpha x}{y + x^2}$$

$$\frac{dx}{dt} = \alpha xy - \beta x^2$$

(B) *Given the current set of candidate equations, optimize a small sample of the whole data set which can still approximate the mean error of these equations*

$$\frac{dx}{dt} = \alpha x - \gamma$$

$$\frac{dx}{dt} = \frac{\alpha x}{y + x^2}$$

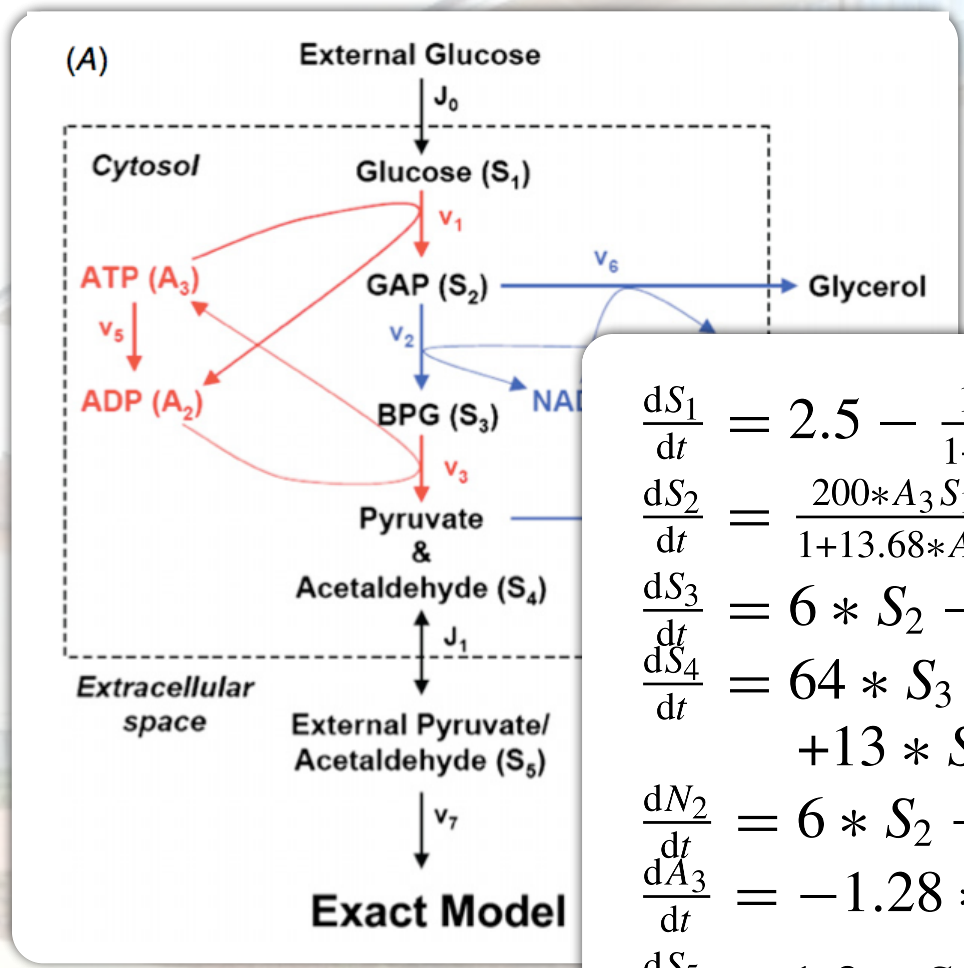
$$\frac{dx}{dt} = \alpha xy - \beta x^2$$

Exhaustive search through all possible elementary functions with a smart choice of experiments.

Schmidt et al., 2011

Testing Model: Yeast Glycolytic Oscillator

Ruoff et al., 2003



- 7 species, 28 variables
- Complex rational dynamical laws

$$\frac{dS_1}{dt} = 2.5 - \frac{100{*}A_3S_1}{1{+}13.68{*}A_3^4}$$

$$\frac{dS_2}{dt} = \frac{200{*}A_3S_1}{1{+}13.68{*}A_3^4} - 6*S_2 - 6*S_2N_2$$

$$\frac{dS_3}{dt} = 6*S_2 - 6*N_2S_2 - 64*S_3 + 16*A_3S_3$$

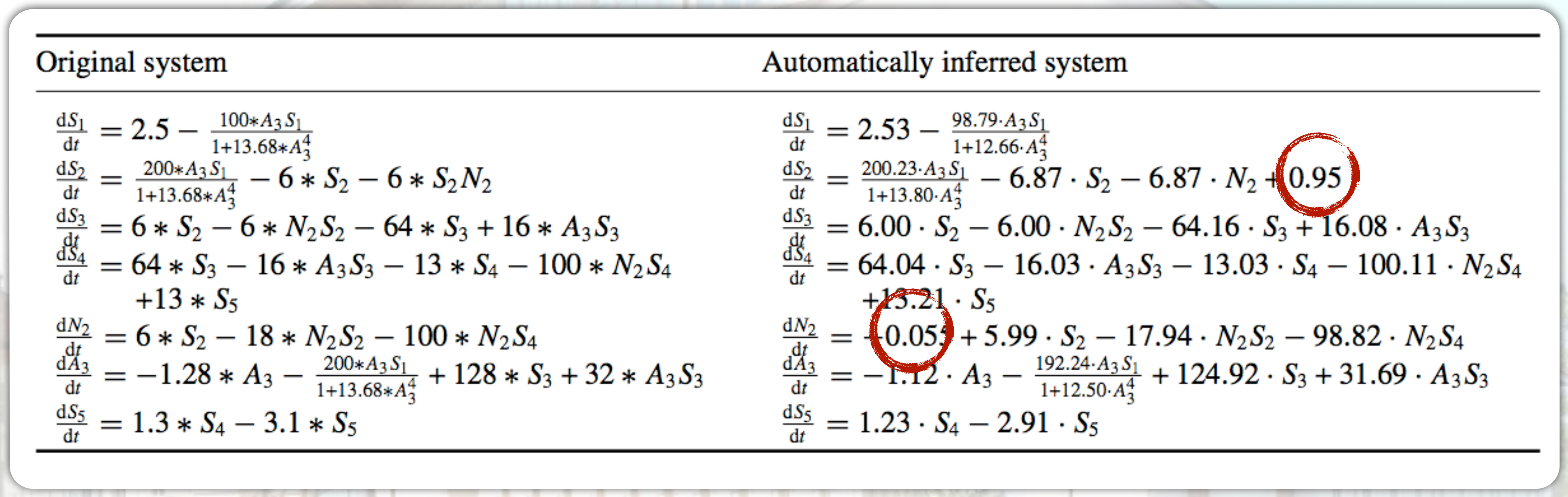
$$\frac{dS_4}{dt} = 64*S_3 - 16*A_3S_3 - 13*S_4 - 100*N_2S_4 \\ +13*S_5$$

$$\frac{dN_2}{dt} = 6*S_2 - 18*N_2S_2 - 100*N_2S_4$$

$$\frac{dA_3}{dt} = -1.28*A_3 - \frac{200{*}A_3S_1}{1{+}13.68{*}A_3^4} + 128*S_3 + 32*A_3S_3$$

$$\frac{dS_5}{dt} = 1.3*S_4 - 3.1*S_5$$

Results of Schmidt et al.

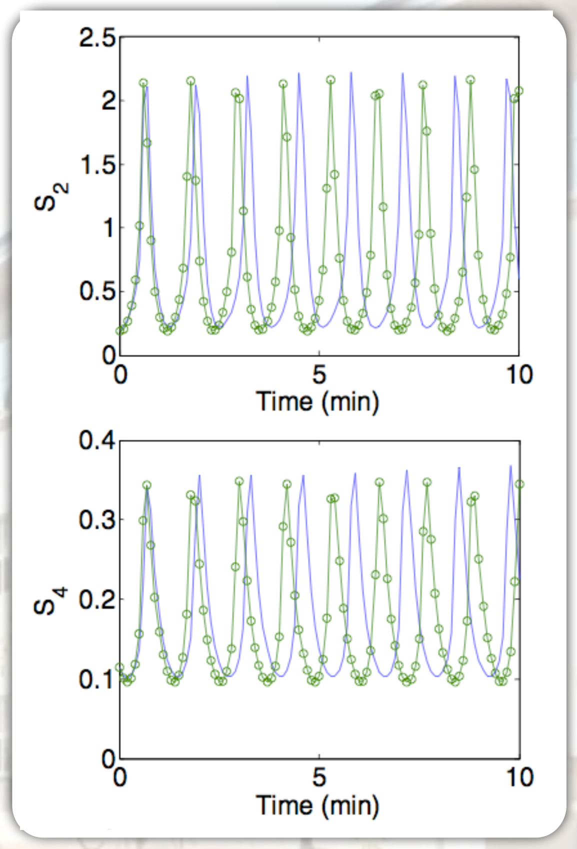


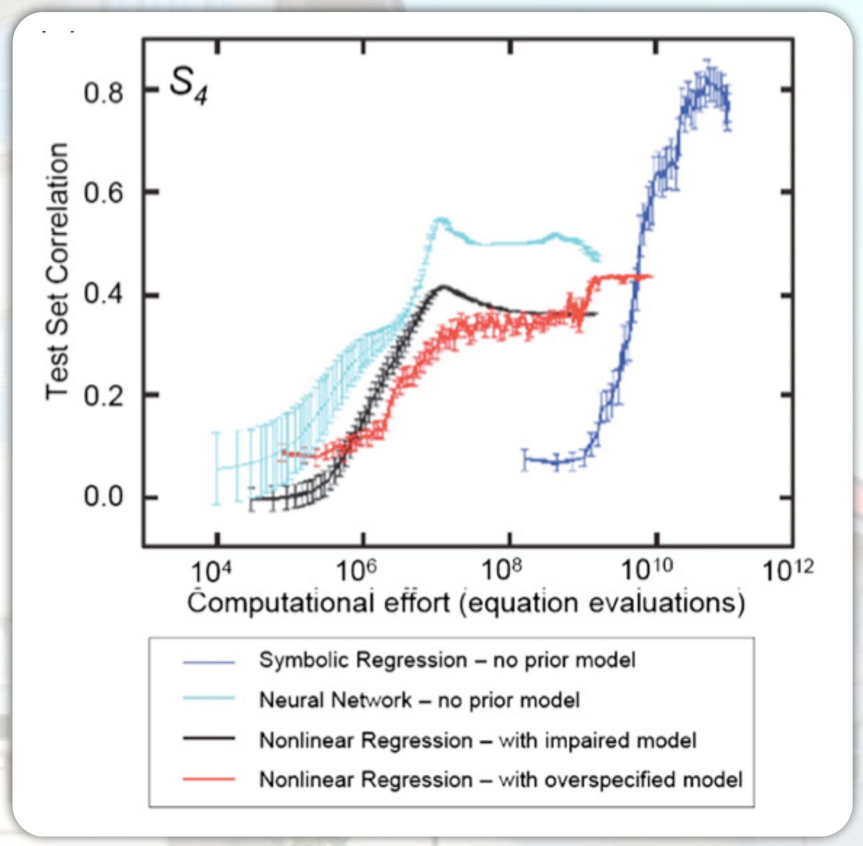
Original system	Automatically inferred system
$\frac{dS_1}{dt} = 2.5 - \frac{100 * A_3 S_1}{1+13.68 * A_3^4}$	$\frac{dS_1}{dt} = 2.53 - \frac{98.79 \cdot A_3 S_1}{1+12.66 \cdot A_3^4}$
$\frac{dS_2}{dt} = \frac{200 * A_3 S_1}{1+13.68 * A_3^4} - 6 * S_2 - 6 * S_2 N_2$	$\frac{dS_2}{dt} = \frac{200.23 \cdot A_3 S_1}{1+13.80 \cdot A_3^4} - 6.87 \cdot S_2 - 6.87 \cdot N_2 + 0.95$
$\frac{dS_3}{dt} = 6 * S_2 - 6 * N_2 S_2 - 64 * S_3 + 16 * A_3 S_3$	$\frac{dS_3}{dt} = 6.00 \cdot S_2 - 6.00 \cdot N_2 S_2 - 64.16 \cdot S_3 + 16.08 \cdot A_3 S_3$
$\frac{dS_4}{dt} = 64 * S_3 - 16 * A_3 S_3 - 13 * S_4 - 100 * N_2 S_4$ $+13 * S_5$	$\frac{dS_4}{dt} = 64.04 \cdot S_3 - 16.03 \cdot A_3 S_3 - 13.03 \cdot S_4 - 100.11 \cdot N_2 S_4$ $+13.21 \cdot S_5$
$\frac{dN_2}{dt} = 6 * S_2 - 18 * N_2 S_2 - 100 * N_2 S_4$	$\frac{dN_2}{dt} = -0.055 + 5.99 \cdot S_2 - 17.94 \cdot N_2 S_2 - 98.82 \cdot N_2 S_4$
$\frac{dA_3}{dt} = -1.28 * A_3 - \frac{200 * A_3 S_1}{1+13.68 * A_3^4} + 128 * S_3 + 32 * A_3 S_3$	$\frac{dA_3}{dt} = -1.12 \cdot A_3 - \frac{192.24 \cdot A_3 S_1}{1+12.50 \cdot A_3^4} + 124.92 \cdot S_3 + 31.69 \cdot A_3 S_3$
$\frac{dS_5}{dt} = 1.3 * S_4 - 3.1 * S_5$	$\frac{dS_5}{dt} = 1.23 \cdot S_4 - 2.91 \cdot S_5$

Pretty amazing!

Schmidt et al., 2011

Results of Schmidt et al.

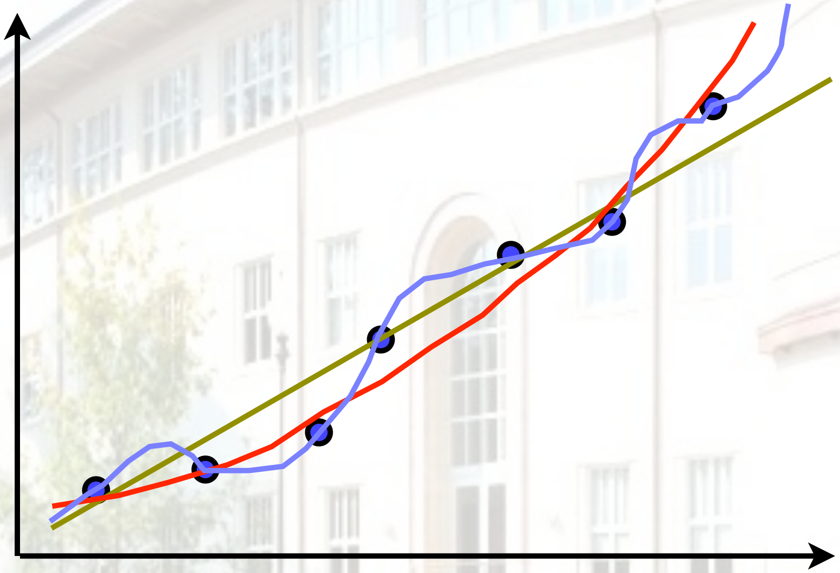




- Astronomical computation times -- **exhaustive search.**
  - Super-exponential scaling with the number of fitted species.
  - Have been unable to go beyond 7 species, or consider hidden species.
  - **Overfitting** -- need astronomical sample sizes.
- Two exponential costs: **selecting** the best model family, **fitting** the best family with the model.

Can we avoid exhaustive search?

- We don't need to do an exhaustive search when fitting 1-dimensional curves



$$y_K(x) = \sum_{k=1}^{K} A_k x^k + \text{noise}$$

  - Taylor or Fourier representations are some of many (nested) ways to represent any function.
  - Don't need to search through all functions as we can do, for example, Bayesian model selection to limit the complexity of the search space (the value of maximum *K*).

Bayesian Model selection

$$P(K|\{x_i\}) = \int d^K\vec{\alpha} P(\vec{\alpha}|\{x_i\}) = \int d^K\vec{\alpha}\frac{P(\{x_i\}|\vec{\alpha})\mathcal{P}(\alpha)}{P(\{x_i\})}$$
$$= \int d^k\vec{\alpha}\exp(-N\mathcal{L})$$

$$\log P(K|\{x_i\}) = \log P(\{x_i\}|\vec{\alpha}_{\rm ML}) - \frac{1}{2}\log\det N\mathcal{F} + O(N^0)$$

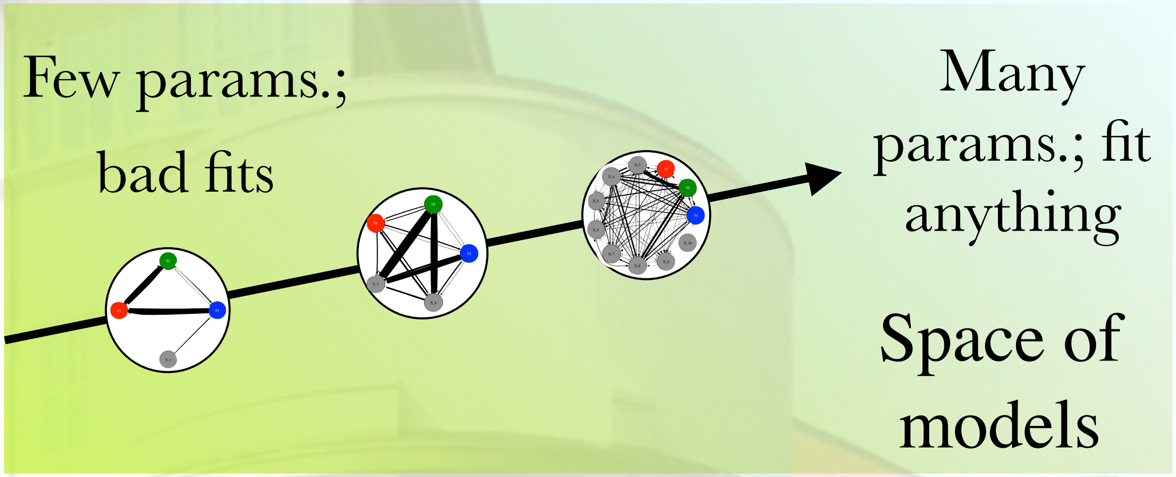
- For large sample size *N*, averages done in the Laplace (saddle point) limit.
- Penalty for model complexity (the log term) “selects” the best model family.
- Not that simple in detail, but this description is roughly accurate.
- Beautiful consistency properties for **nested, complete** model families.

MacKay 1992, Balasubramanian 1996,
Nemenman 2005

Dynamical fits: no nested structures

$$\begin{aligned}\frac{d\vec{x}}{dt} = \quad & A_{\{xx\}}\vec{x}\\ & + A^{(2)}_{\{xx\}}\vec{x}\odot\vec{x}\\ & + \dots\\ & + A^{(K)}_{\{xx\}}\vec{x}\odot\cdots\odot\vec{x}\end{aligned}$$

More nonlinearities



$$\frac{d\vec{x}}{dt} = A_{\{xx\}}\vec{x}$$

More hidden variables

$$\begin{cases}\frac{d\vec{x}}{dt} = A_{\{xx\}}\vec{x} + B_{\{x\}1}\xi_1 + \dots + B_{\{x\}K}\xi_K\\ \frac{d\xi_1}{dt} = A_{1\{x\}}\vec{x} + B_{11}\xi_1 + \dots + B_{1K}\xi_K\\ \dots\\ \frac{d\xi_K}{dt} = A_{K\{x\}}\vec{x} + B_{1K}\xi_1 + \dots + B_{KK}\xi_K\end{cases}$$

- Existence of hidden degrees of freedom and nonlinearities breaks nestedness -- no consistency guarantees.
- Choosing any (reasonable, **complete**) ordered structure through the model space is better than not choosing.
  - For a good choice, we will fit well with few data; for a bad choice we will not do any worse than exhaustive search, which is astronomically slow.

Two types of model families

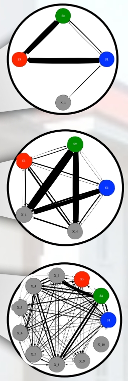
- Both nested and complete.
- Account for nonlinearities **and** hidden variables as more variables are added.
- BIochemically reasonable.

Sigmoidal recurrent networks
(Daniels and Beer)

Degradation Interactions Input

$$\frac{dx_i}{dt} = -x_i/\tau_i + \sum_{j=1}^{J} W_{ij}\ \xi(x_j + \theta_j) + \sum_{k=1}^{K} V_{ik} I_k$$

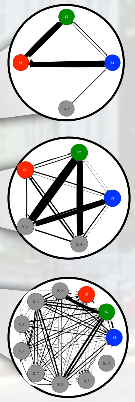
with $\xi(y) = 1/(1 + e^{-y})$



S-systems
(Savageau et. al)

Interactions and input dependence

$$\frac{dx_i}{dt} = A_i \prod x_j^{\alpha_{ij}} \prod_k I_k^{a_{ik}} - B_i \prod x_j^{\beta_{ij}} \prod_k I_k^{b_{ik}}$$

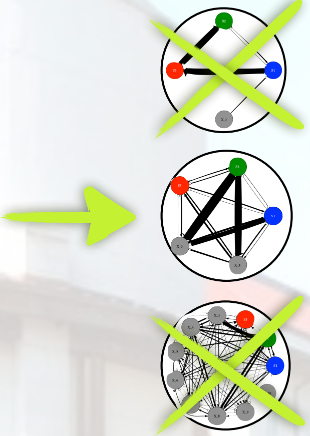


Use Bayesian model selection to choose best models

Chi-squared cost Bayesian penalty

$$\mathcal{L} = \frac{1}{2}\chi^2(\alpha_{\text{best}}) + \frac{1}{2}\sum \log \lambda_\mu$$

Eigenvalues of the Fisher matrix
[Some parameter directions
are not well-defined]



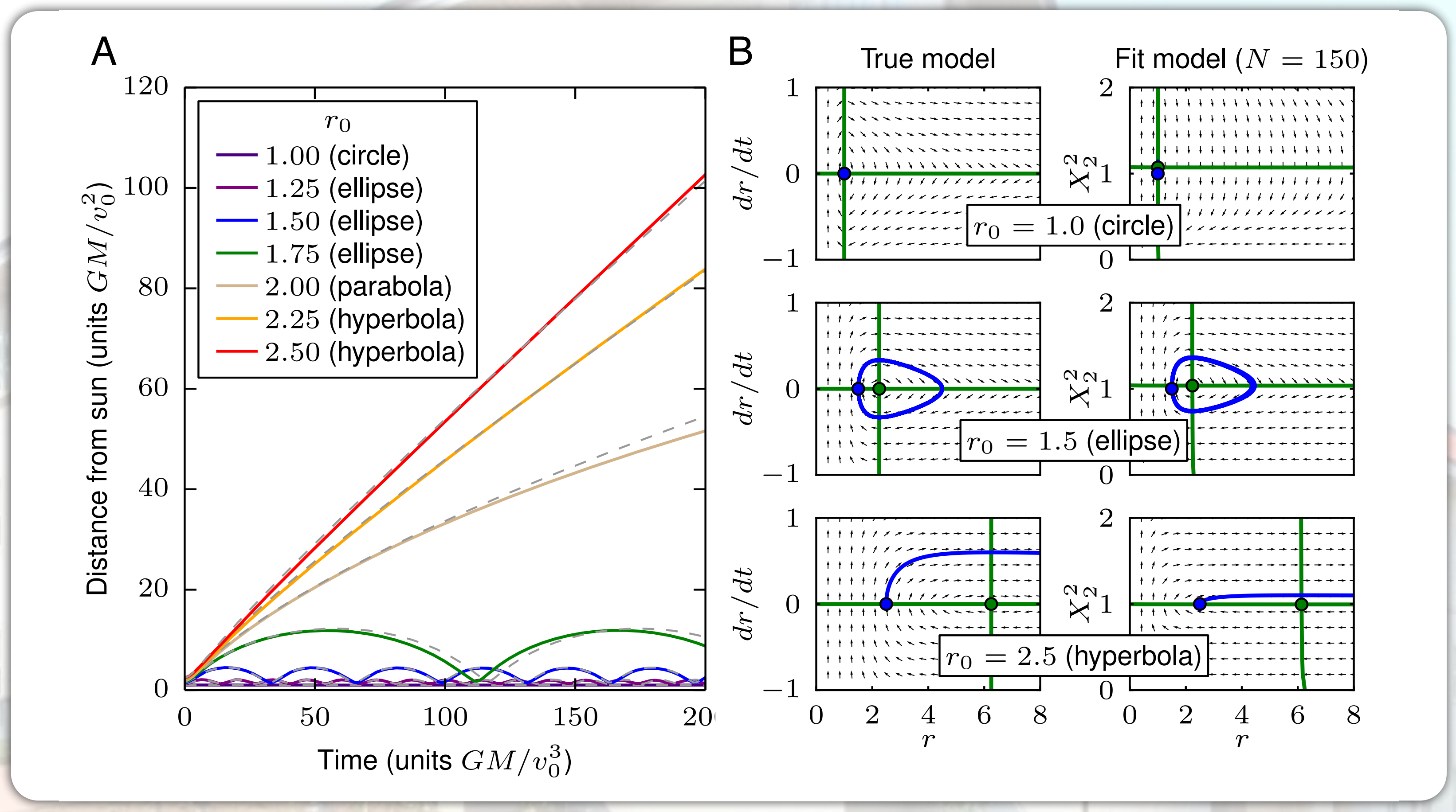
- Bayesian model selection modified for parameter sloppiness

Gutenkunst et al., 2007

Algorithm

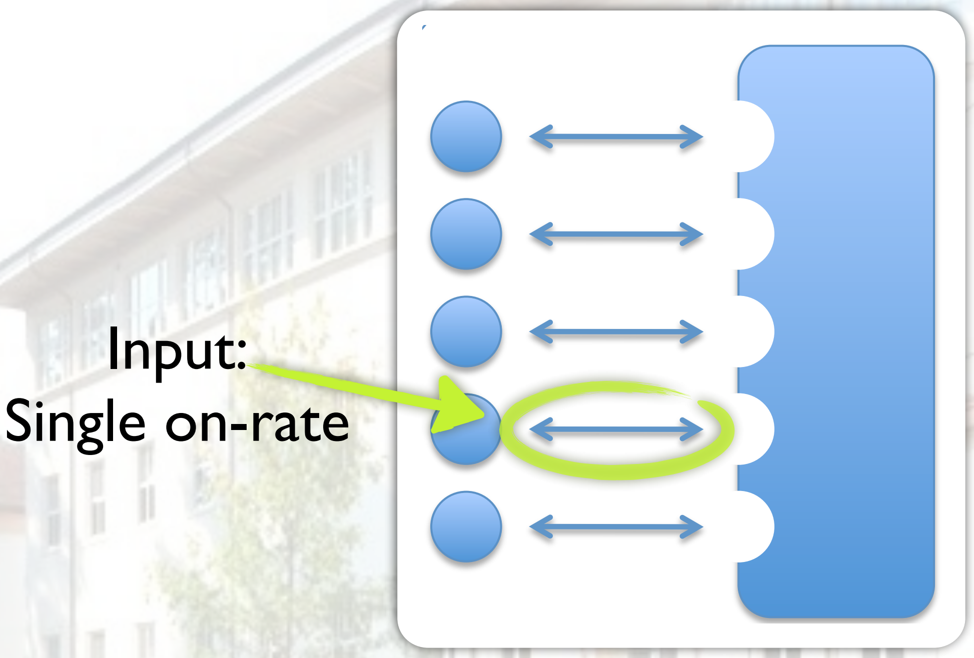
- Specify a particular hierarchy of model families.
- For given data:
  - Choose a model family within the hierarchy.
  - Fit for the best model within the family.
  - Calculate the posterior likelihood of the family using modified Bayesian criterion.
  - Choose more complex family and terminate when the modified likelihood starts to decrease.
- Algorithmic improvements to ensure that no complete re-fitting is done when move to the next family, or increase data set size.
- Two exponential complexities: **search of a model family**, and fitting a model within the family.
  - This only solves the **first**.
  - In practice works OK for both.

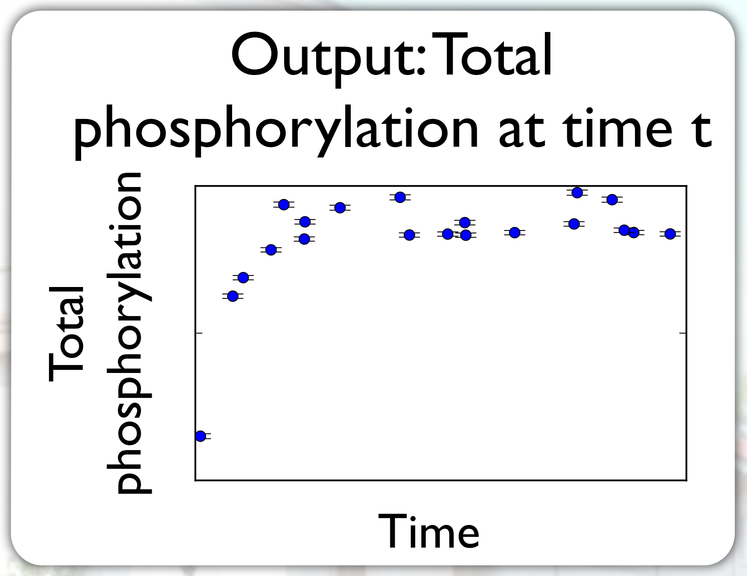
Finding laws that we already know: An automated Sir Isaac *(Sir Isaac on GitHub)*



- Finds the simplest structure that can account for Newton's laws.
- Accounts for different classes of trajectories.

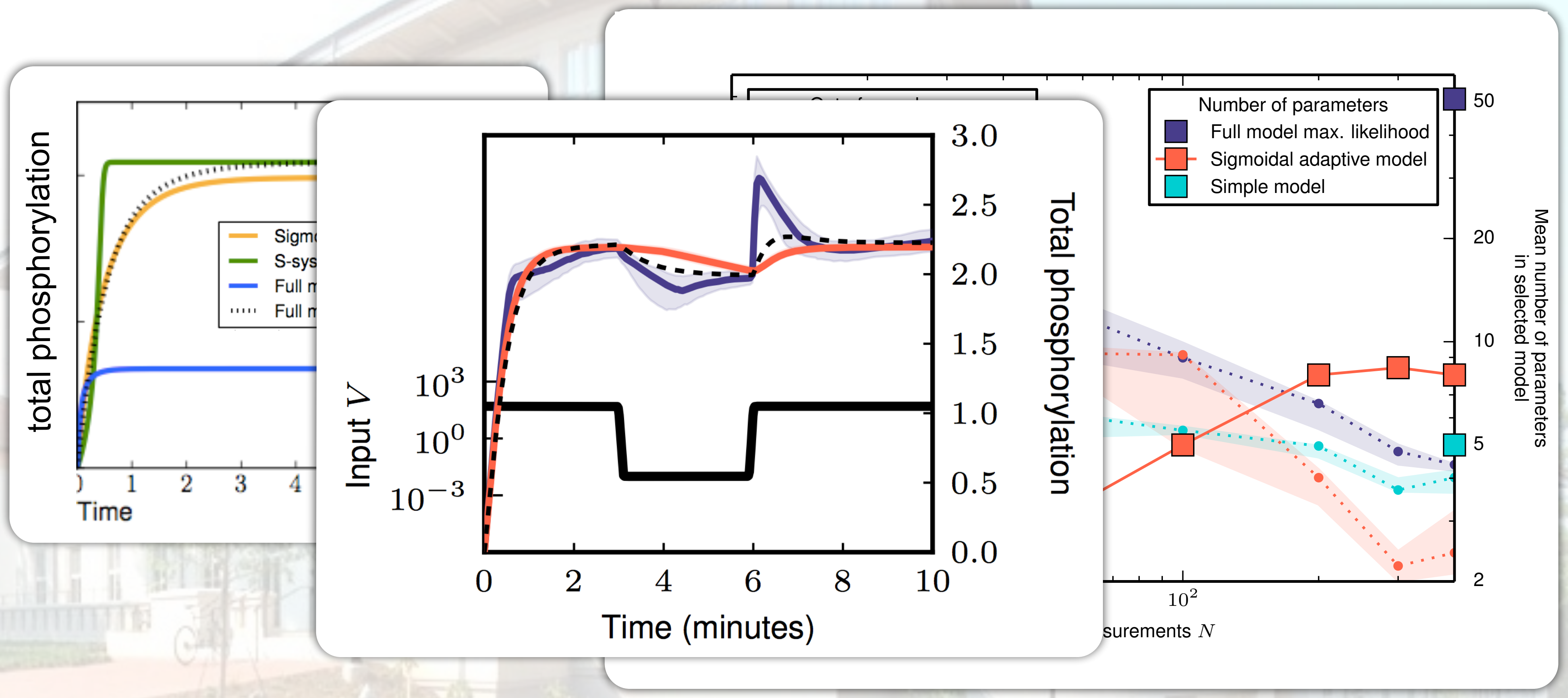
Simple dynamics from a complex network: Combinatorial multisite phosphorylation





- Rates depend on occupancy of the nearby sites, about 50 parameters total.
- Caricature of some of the most combinatorially complex signaling models.
- Typically more parameters than data.

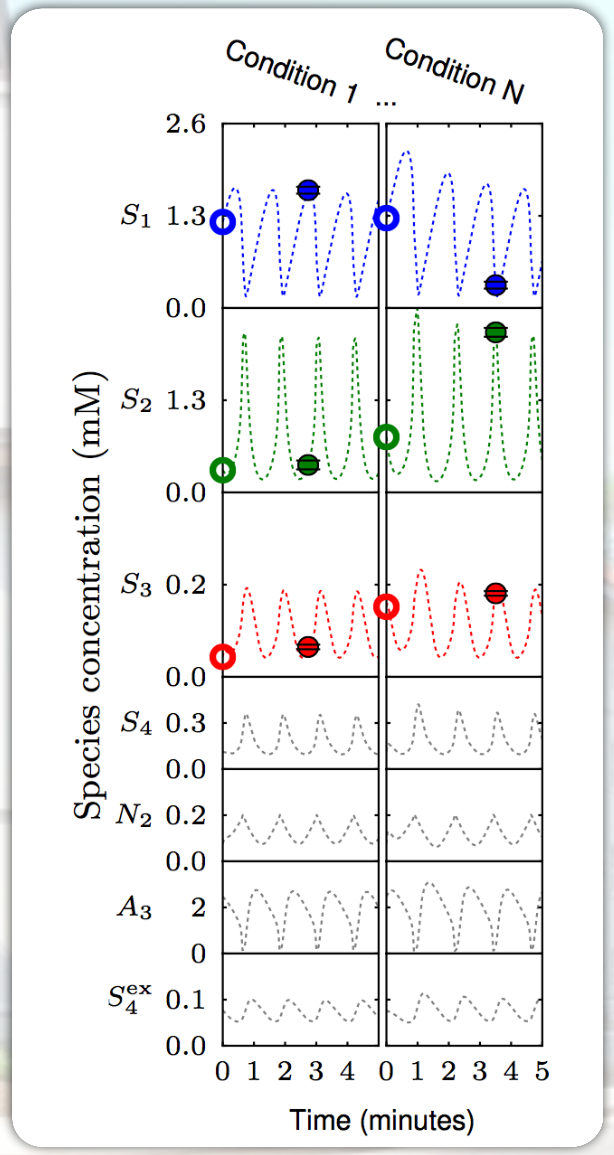
Effective, reduced model of multi-site phosphorylation

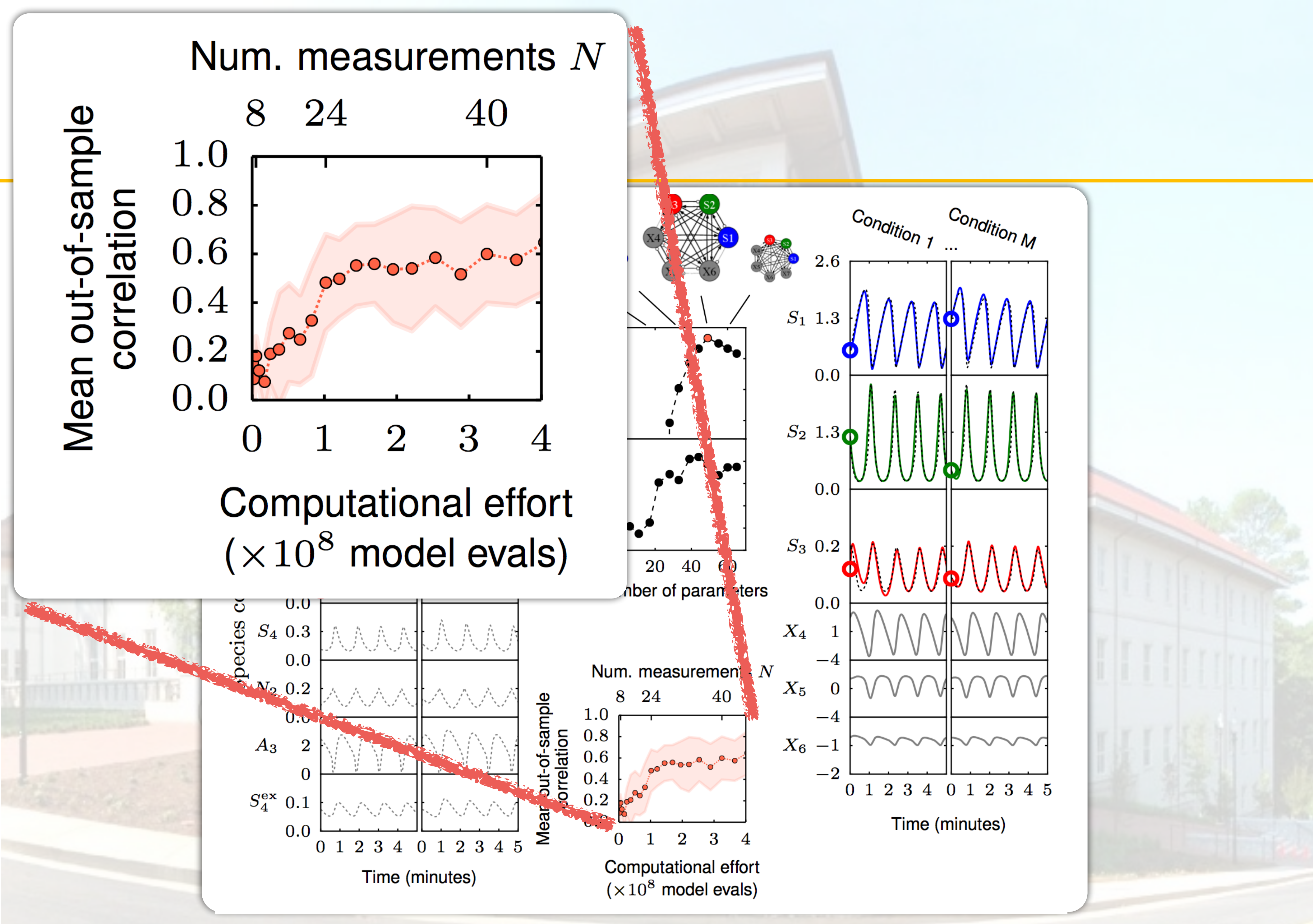


- Effective models (especially sigmoidal) fit better than the true, full model for small data sets!
- Can even *extra*polate to new signal classes, and not just *inter*polate.
- (Of course eventually the full, true model would win).

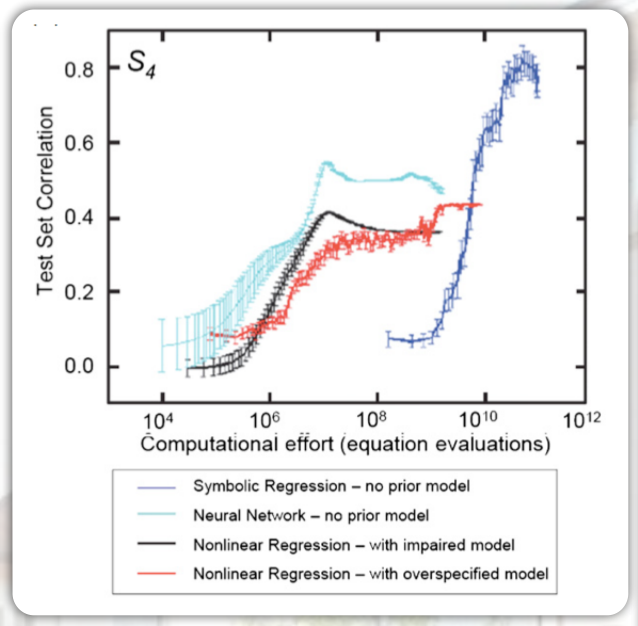
The yeast glycolytic oscillations: Complex dynamics needing complex structure

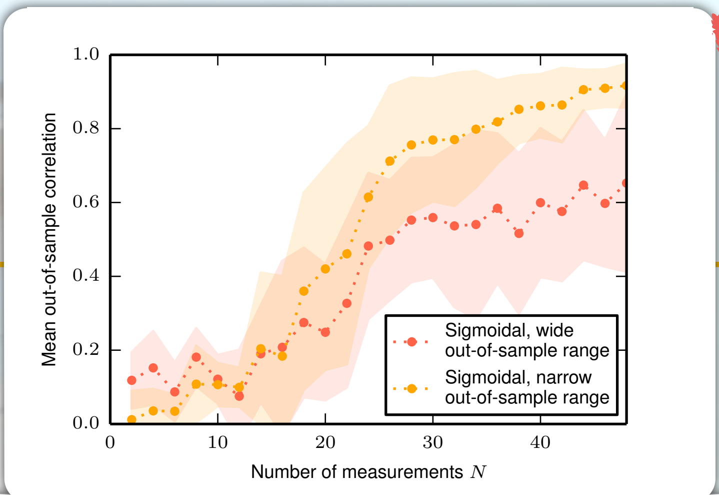
- Observe only 3/7 of variables; add 10% noise.
- Data: *N* samples of structure
  - Initial condition of the 3 species;
  - Some random time later;
  - The value of these 3 species at that time.

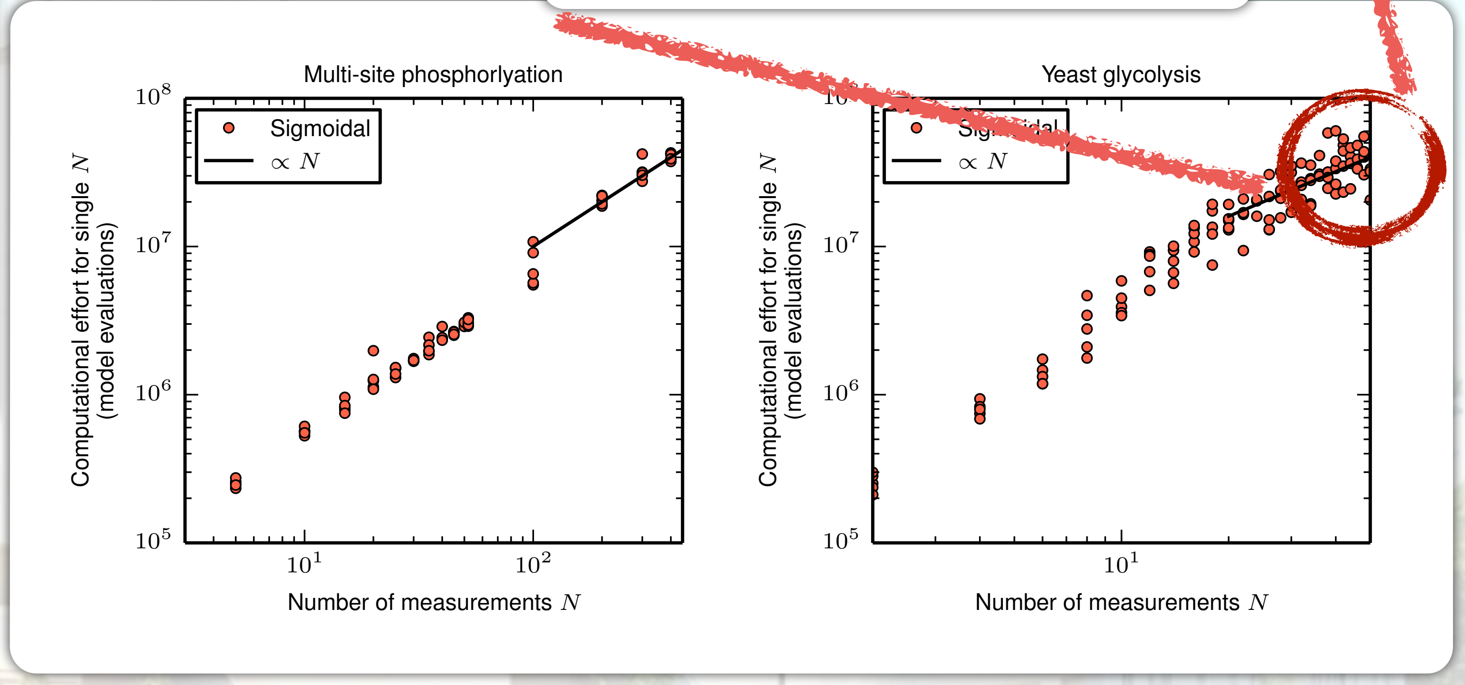



Num. measurements N
8 24 40
Mean out-of-sample correlation
1.0 0.8 0.6 0.4 0.2 0.0
0 1 2 3 4
Computational effort ($\times 10^8$ model evals)
S3 S2 S1 X4 X5 X6
20 40 60
Number of parameters
Condition 1 … Condition M
S_1 2.6 1.3 0.0
S_2 1.3 0.0
S_3 0.2 0.0
X_4 1 −4
X_5 0 −4
X_6 −1 −2
0 1 2 3 4 0 1 2 3 4 5
Time (minutes)
Species c
S_4 0.3 0.0
N_2 0.2 0.0
A_3 2 0
S_4^{ex} 0.1 0.0
0 1 2 3 4 0 1 2 3 4 5
Time (minutes)

Computational effort







- Two orders of magnitude fewer evaluations for the same accuracy compared to EUREQA.
- Only 40 data points (three orders of magnitude fewer).
- Better accuracy than curve (rather than dynamics) fitting.

Conclusions

- Fitting **dynamics,** not curves.
- **Complete, nested** model families of dynamics allow to use Bayesian model selection to adapt effective model complexity to the available data.
  - To the zeroth order, selecting any structure is better than having no structure at all.
  - To the first order, the models one uses (the primitives and the structure) start mattering.
- Such effective models make accurate predictions in the undersampled regime, where true models overfit.
- Why do this?
  - **The cat test:** If it purrs like a cat and has whiskers like a cat, then it probably is a cat
    - Indeed, **can predict response to yet-unseen perturbations!**
  - **Find new laws of nature:** build effective models of many similar systems and look for similarities.