Inducing Process Models in Systems Biology

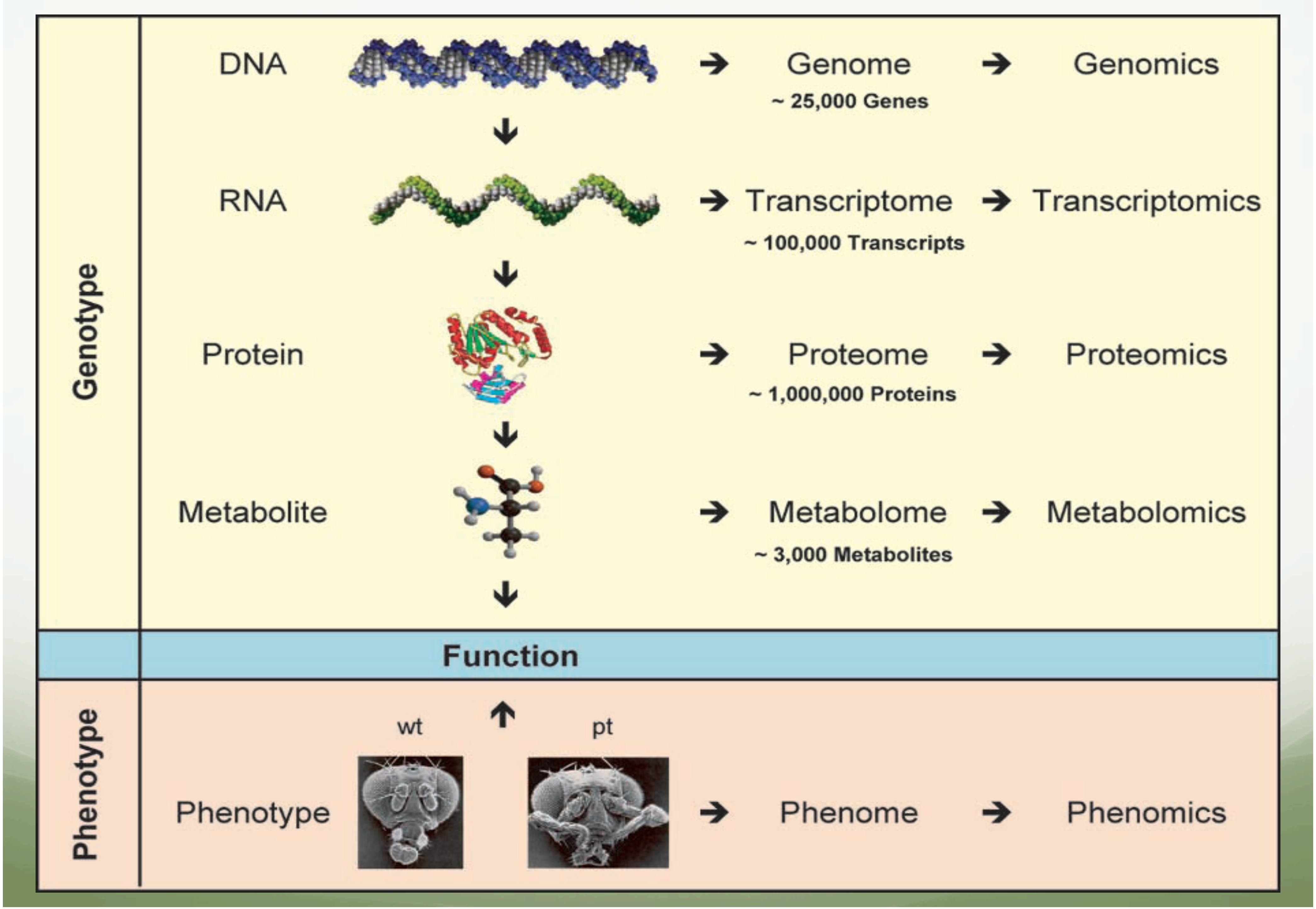
Ljupco Todorovski
University of Ljubljana, Faculty of Administration

Outline

- Systems Biology and Biological Networks
- Process Models and Inductive Process Modeling
- Applications of IPM
  - In population ecology
  - In systems biology
- Summary and Outlook

Systems Biology (SB)

- New branch of the life sciences
  - Tries to understand organisms as a whole
- Need to have an integrated picture of the processes
  - That happen in the system at all levels
  - And their dynamics
- From the genome to the phenome
  - Integrating high-throughput data from –omics
  - genomics, transcriptomics, proteomics, metabolomics, and phenomics


Genotype
DNA
Genome
~ 25,000 Genes
Genomics
RNA
Transcriptome
~ 100,000 Transcripts
Transcriptomics
Protein
Proteome
~ 1,000,000 Proteins
Proteomics
Metabolite
Metabolome
~ 3,000 Metabolites
Metabolomics
Function
Phenotype
wt
pt
Phenotype
Phenome
Phenomics

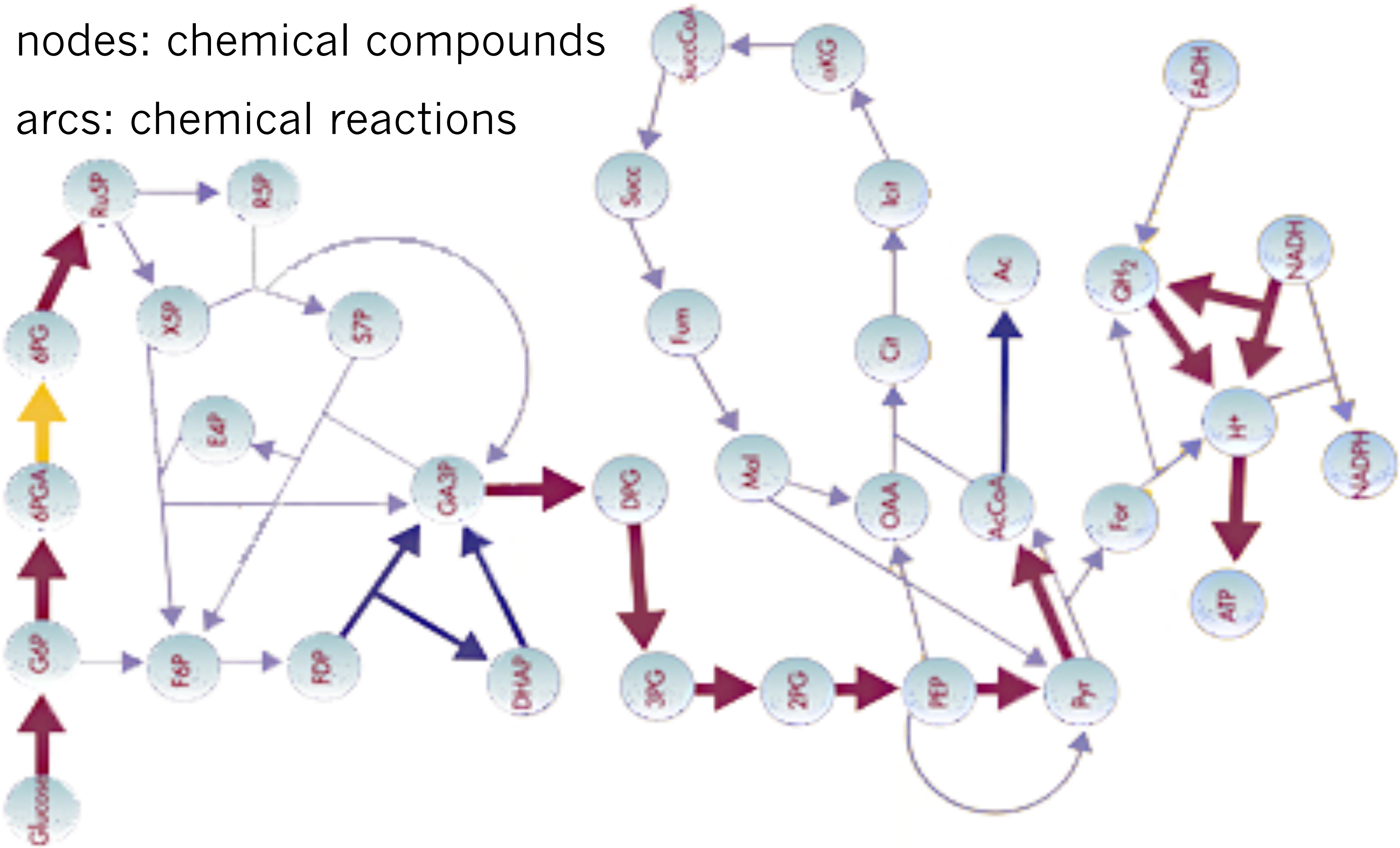
Biological Networks

- A crucial distinction for SB is the study of
  - connections/interactions among components of the system, rather than just the individual components
- Thus, SB focuses on interaction networks
  - Nodes are chemical compounds, metabolits, proteines, receptors, kinases, genes, etc.
  - Arcs are influences, interactions, processes
- Types of biological networks in SB
  - Metabolic networks
  - Gene regulatory networks

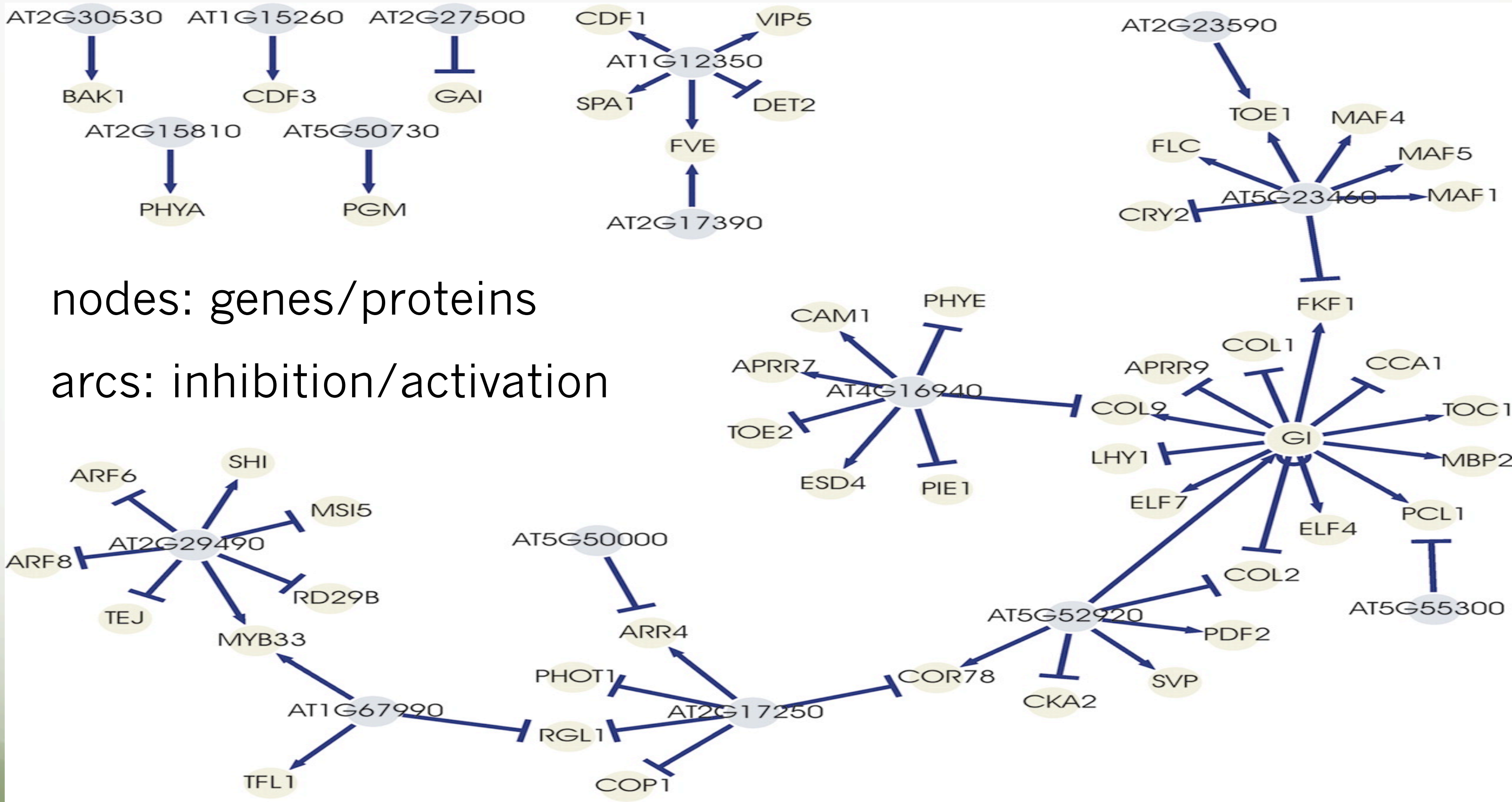
Metabolic Networks

nodes: chemical compounds

arcs: chemical reactions



Gene Regulatory Network



nodes: genes/proteins

arcs: inhibition/activation

Models of Bio Nets

- Boolean networks: Discrete time, deterministic models
  - Nodes are Boolean variables
  - (Hyper-)Arcs are Boolean functions
- Dynamic Bayesian networks: Discrete time, stochastic
  - Nodes are Boolean variables
  - (Hyper-)Arcs are probabilistic influences
- Process models: Continuous time, deterministic
  - Nodes are continuous variables
  - (Hyper-)Arcs are modeled using equations: algebraic or ordinary differential equations

Reconstructing Bio Nets

- A task of central interest in system biology
  - Formulating models of biological networks that capture the dynamics of the studied systems
  - From time series data about the measured concentrations of compounds, metabolits, etc.
- Need to determine structure and parameters
  - Networks structure
  - Functional form of the equations
  - Values of the constant parameters in the equations (e.g., reaction rates constants)

Process Models

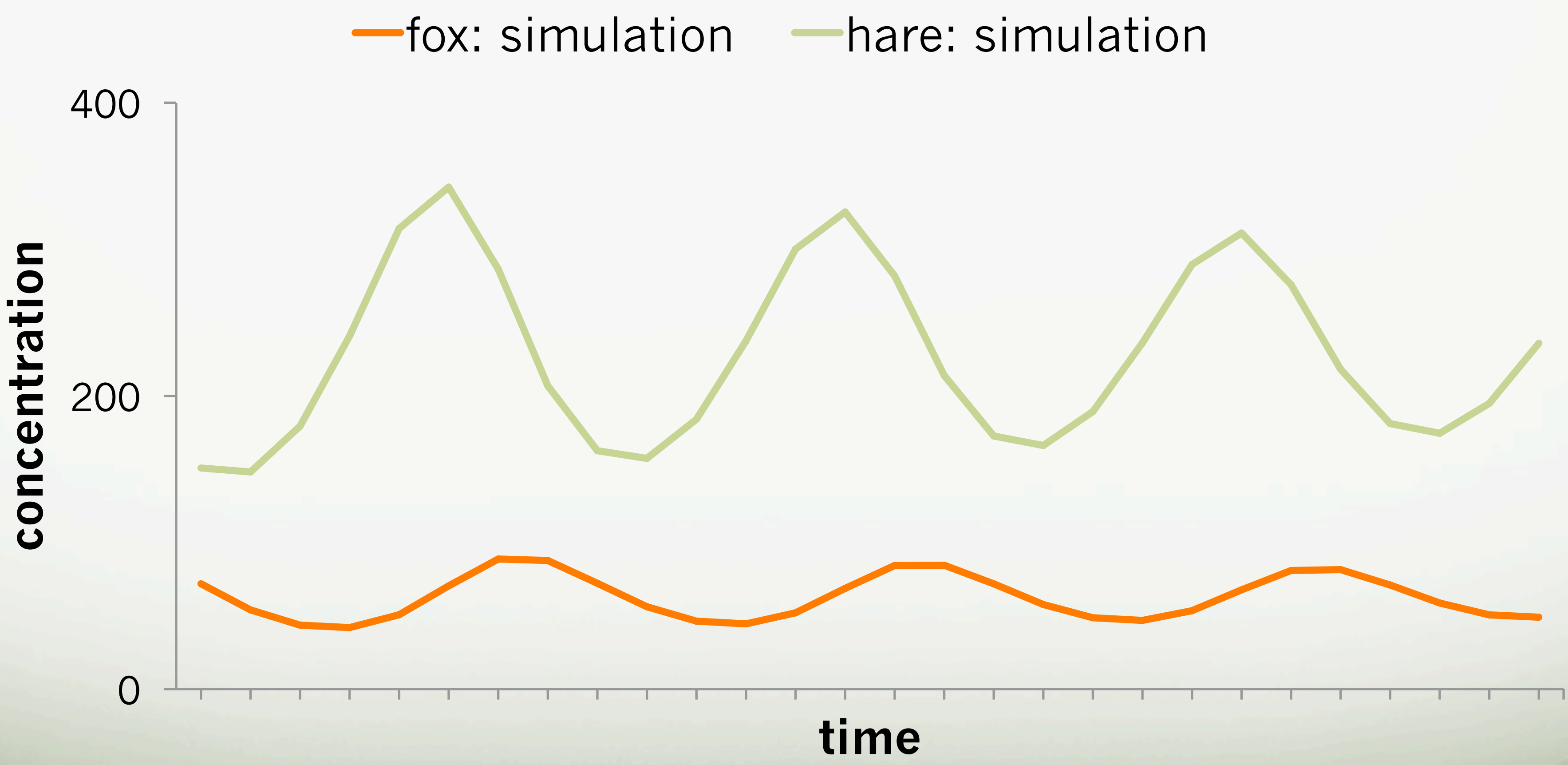
Process Models (PM)

- Integrate two aspects of equation-based models
  - Quantitative aspect: important for simulation
  - Qualitative aspect: important for explanation
- Predator-prey interaction example
  - Two system variables: fox (predator) and hare (prey)
  - Quantitative aspect: two ordinary differential equations that allows simulation
  - Qualitative aspect: three processes that explains the structure of the model

PM: Quantitative Aspect

- System of two differential equations
  - $d\, hare / dt = 2.5 \times hare - 0.35 \times hare \times fox$
  - $d\, fox / dt = 0.03 \times hare \times fox - 1.2 \times fox$
- Models temporal change of the system variables
  - dX / dt denotes first-order time derivative of X
  - First-order time derivative = temporal change
- Allows simulation of the model
  - Given the initial values of the system variables
  - See the next slide

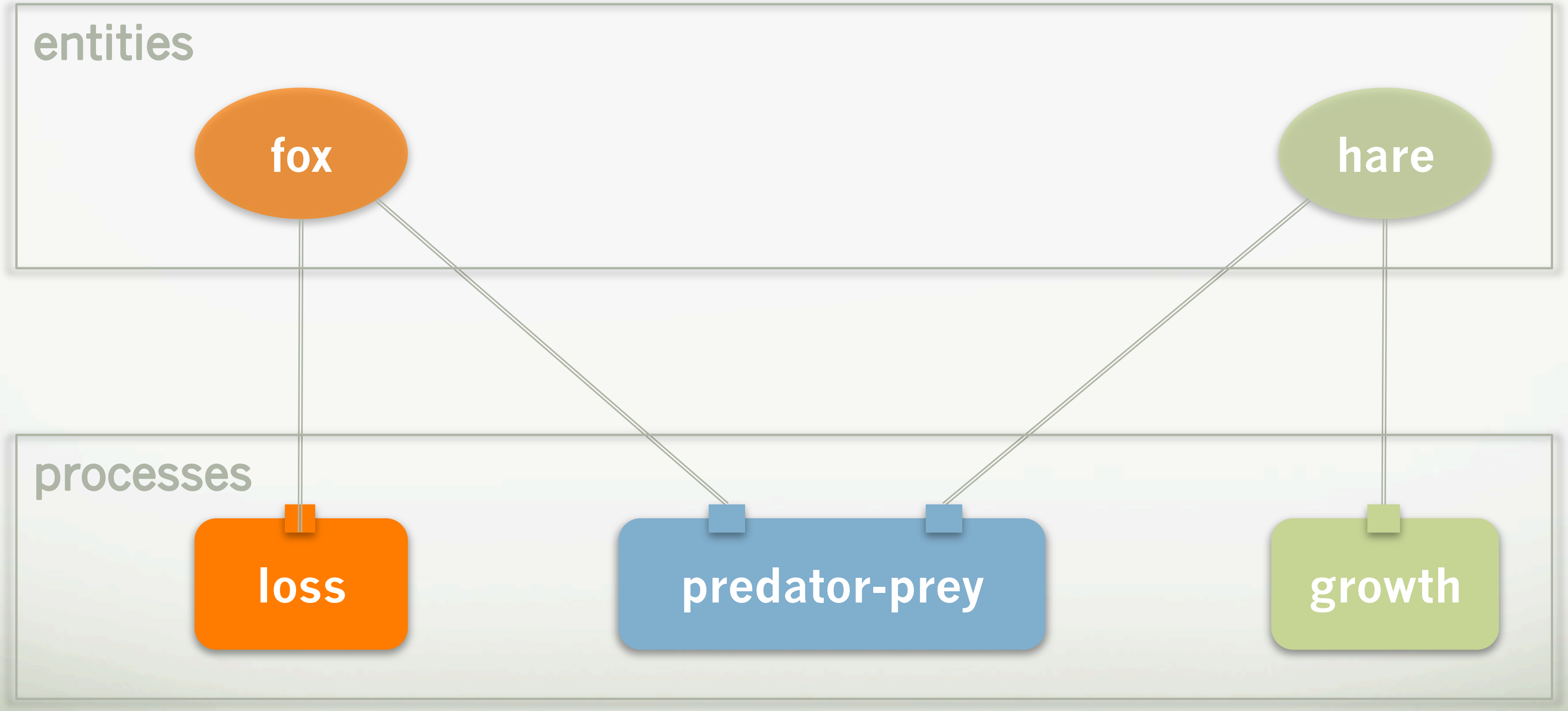
PM: Simulation



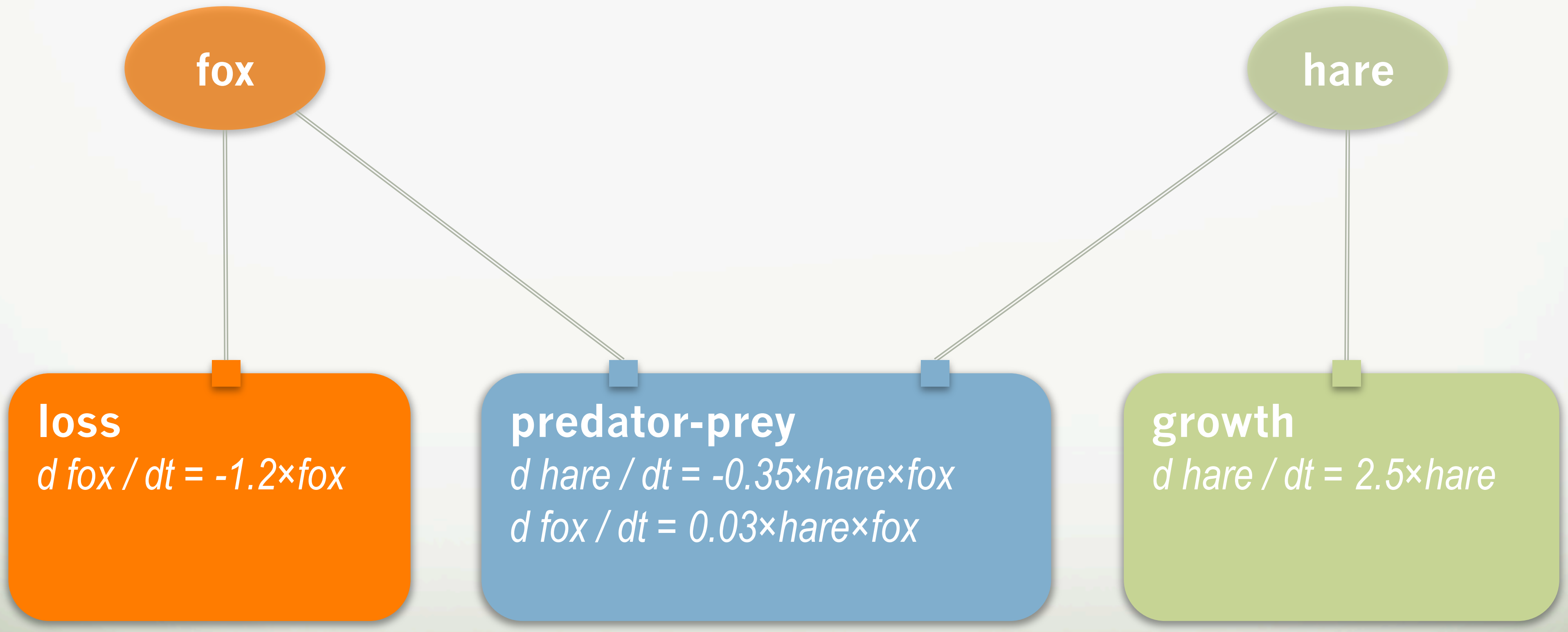
PM: Explanatory Aspect

- What is the semantics of the equation terms?
- The semantics reveals three processes
  - *d hare / dt = 2.5 × hare – 0.35 × hare × fox*
  - *d fox / dt = 0.03 × hare × fox – 1.2 × fox*
- The three processes correspond to
  - Growth of hare population
  - Loss of fox population
  - Predator-prey interaction between the two species

PM: Qualitative Aspect



PM: Integration

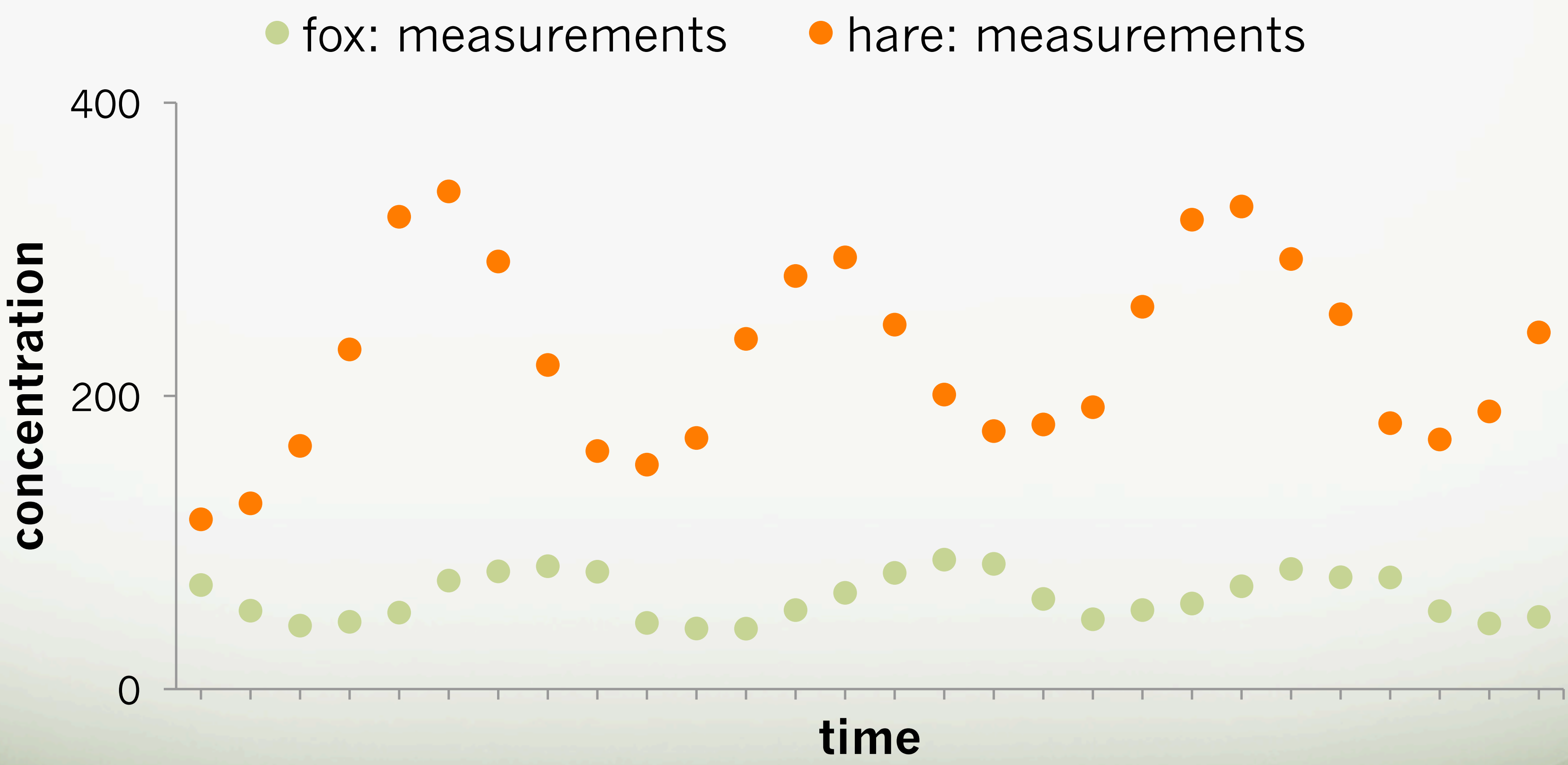


Inductive Process Modeling

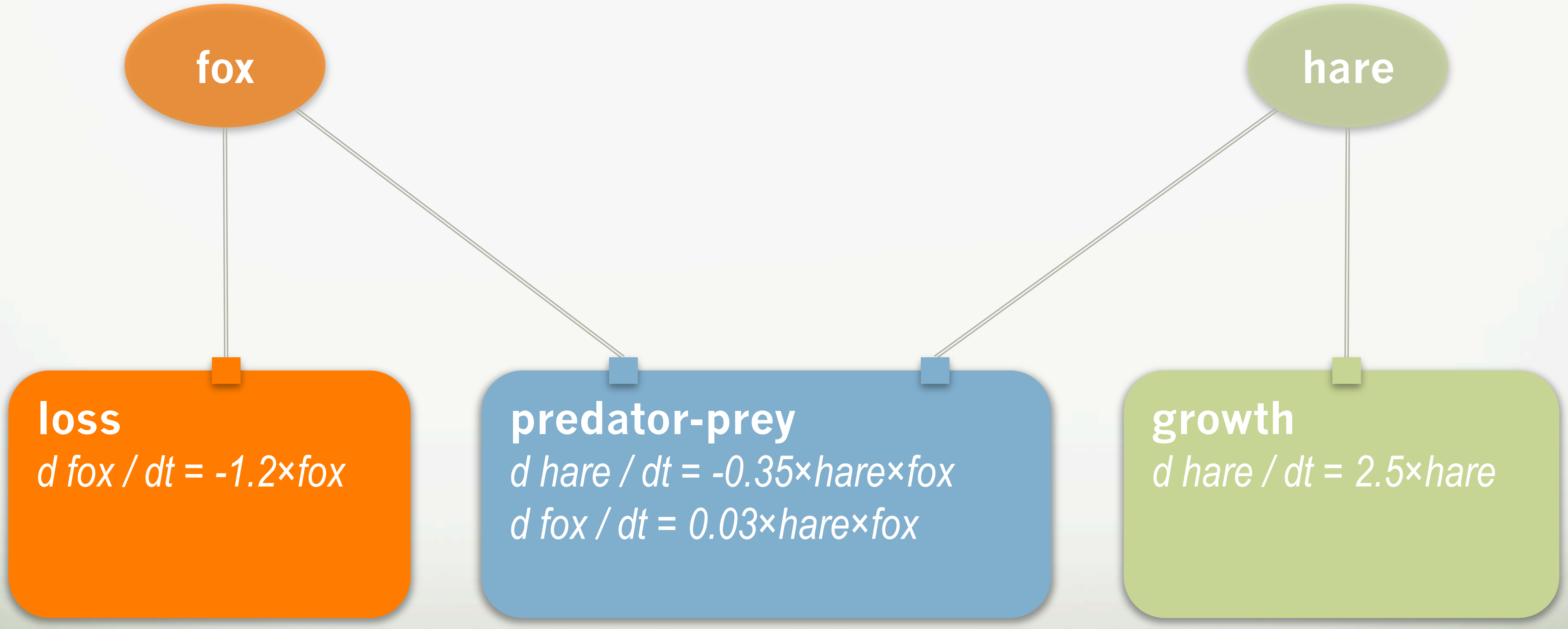
Inductive Process Modeling

- The task of inductive process modeling
  - Is the task of formulating a model
  - That explains observed time series data
- Given
  - Measured behavior of an observed system
  - And library of modeling knowledge
- Find process model
  - Composed of the given generic processes
  - That fit the measured behavior of the system

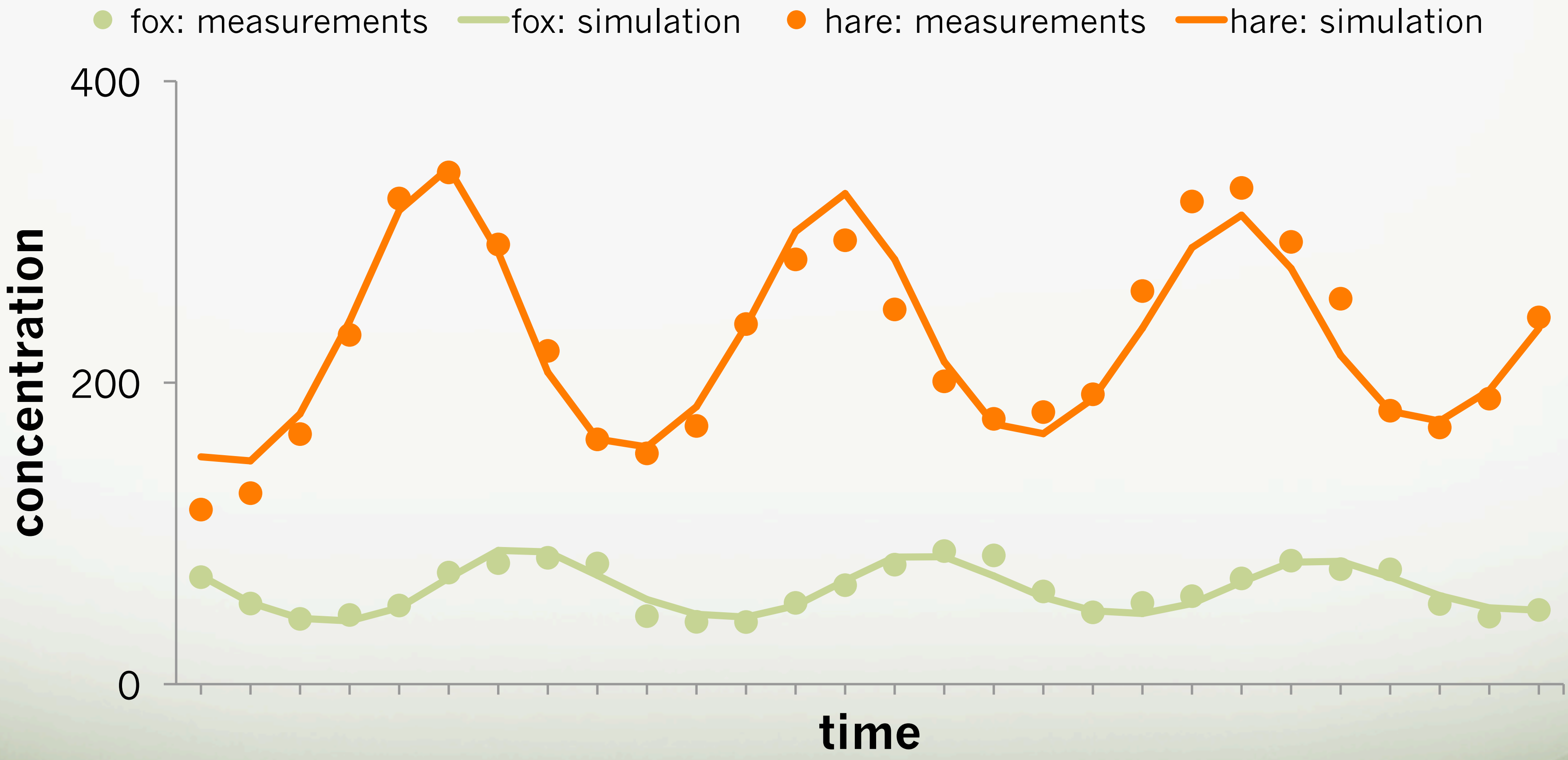
IPM Task Example: Given...



… Find a Process Model…

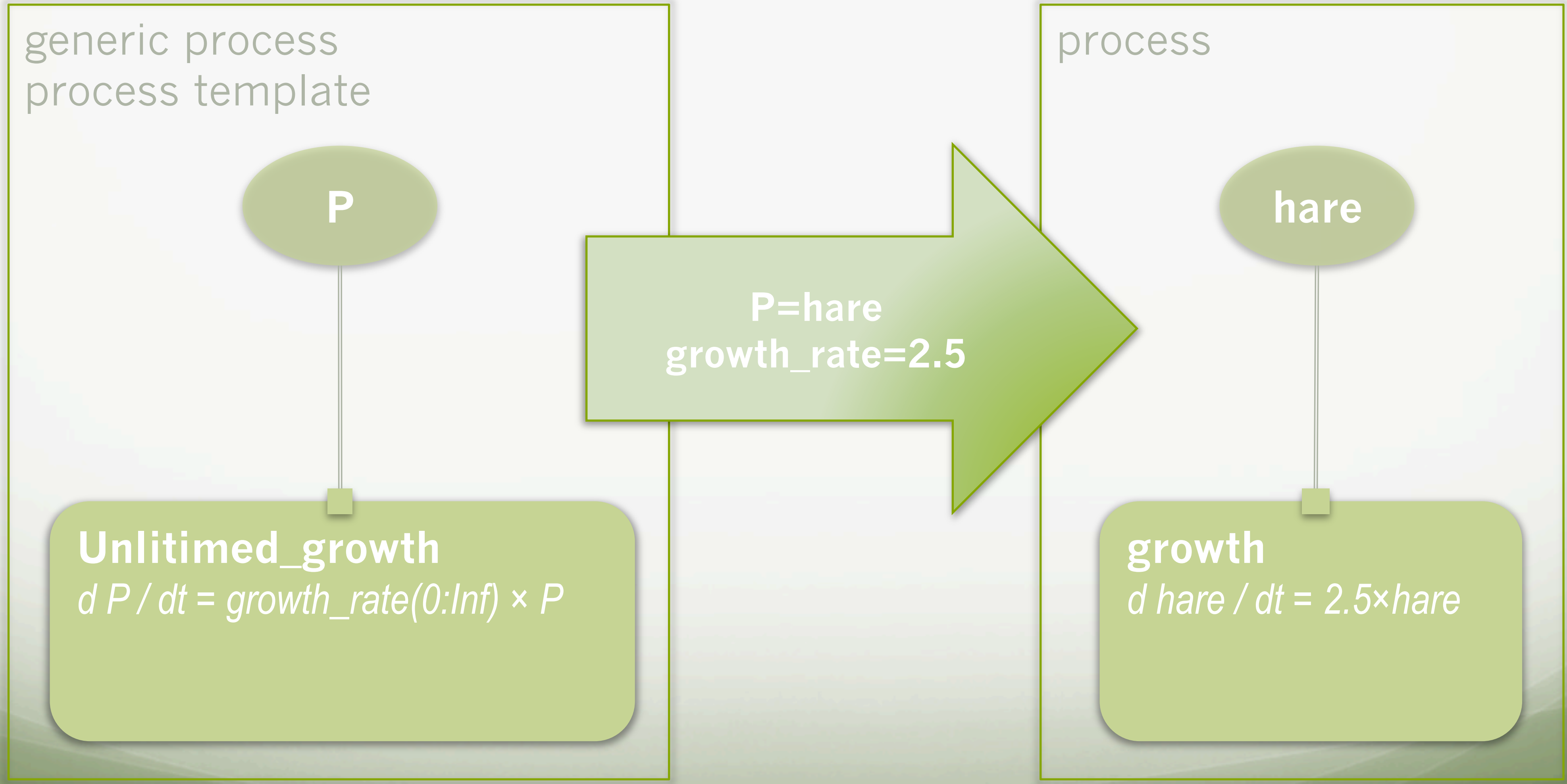


... Such That



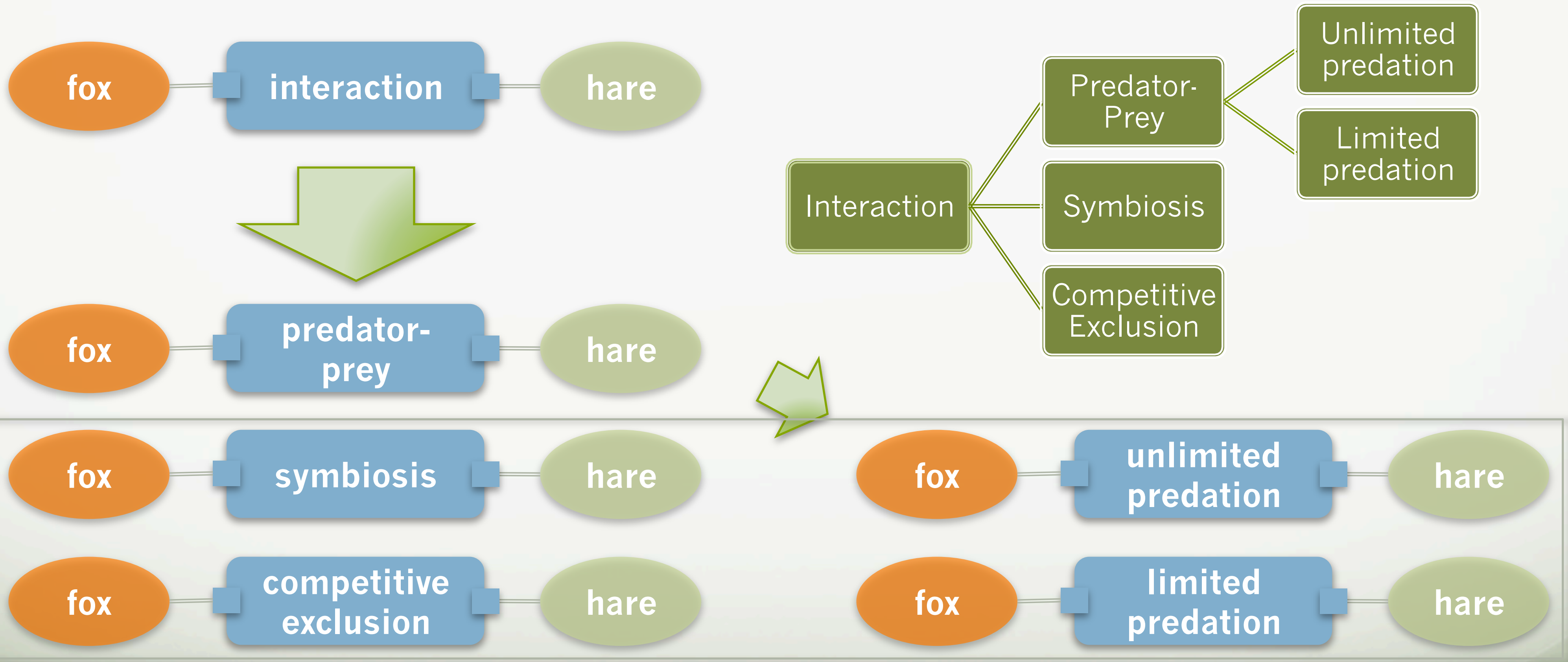
fox: measurements
fox: simulation
hare: measurements
hare: simulation
400
200
0
concentration
time

Modeling Knowledge: Generic Processes

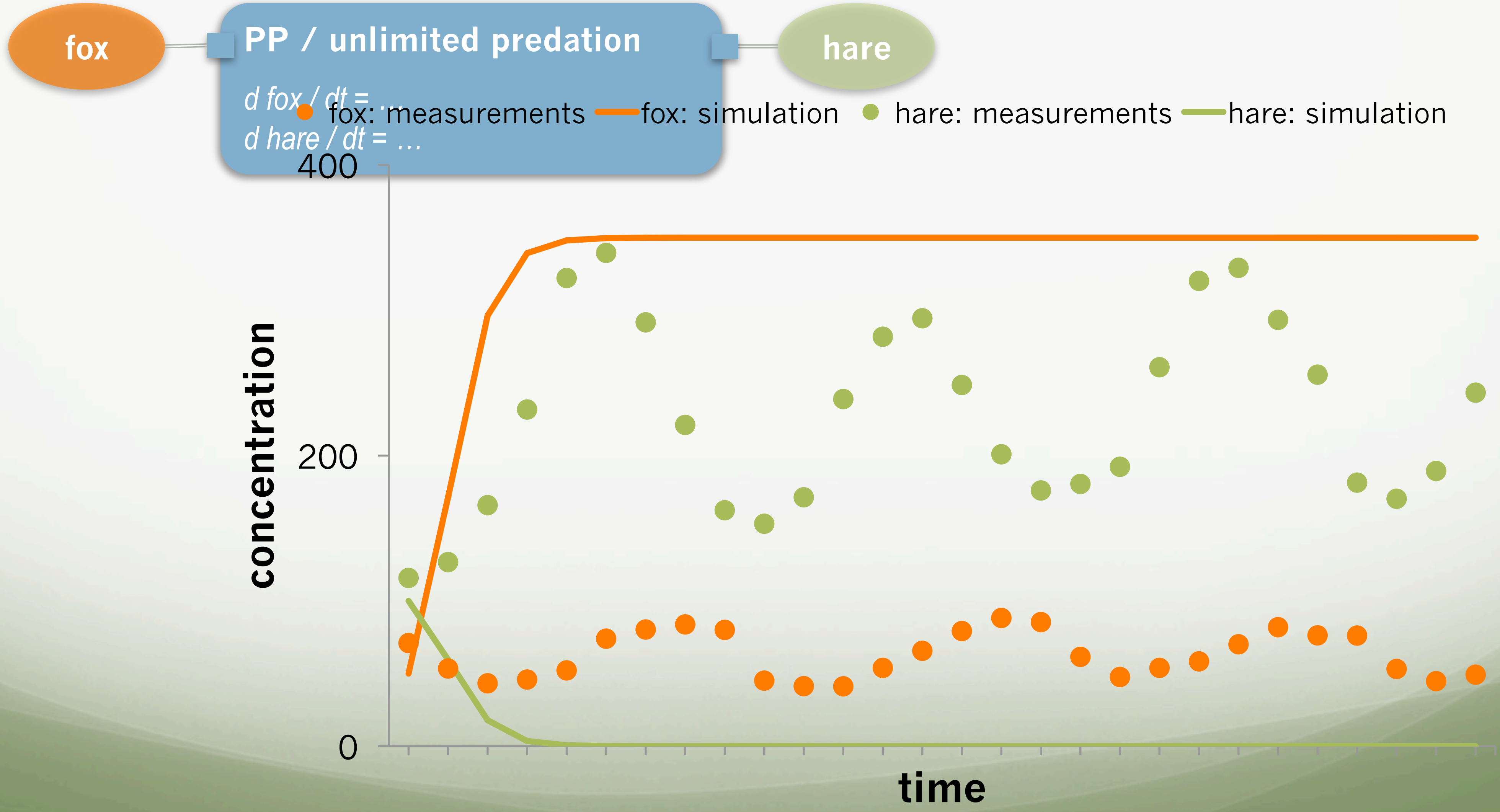


Hierarchy of Generic Processes

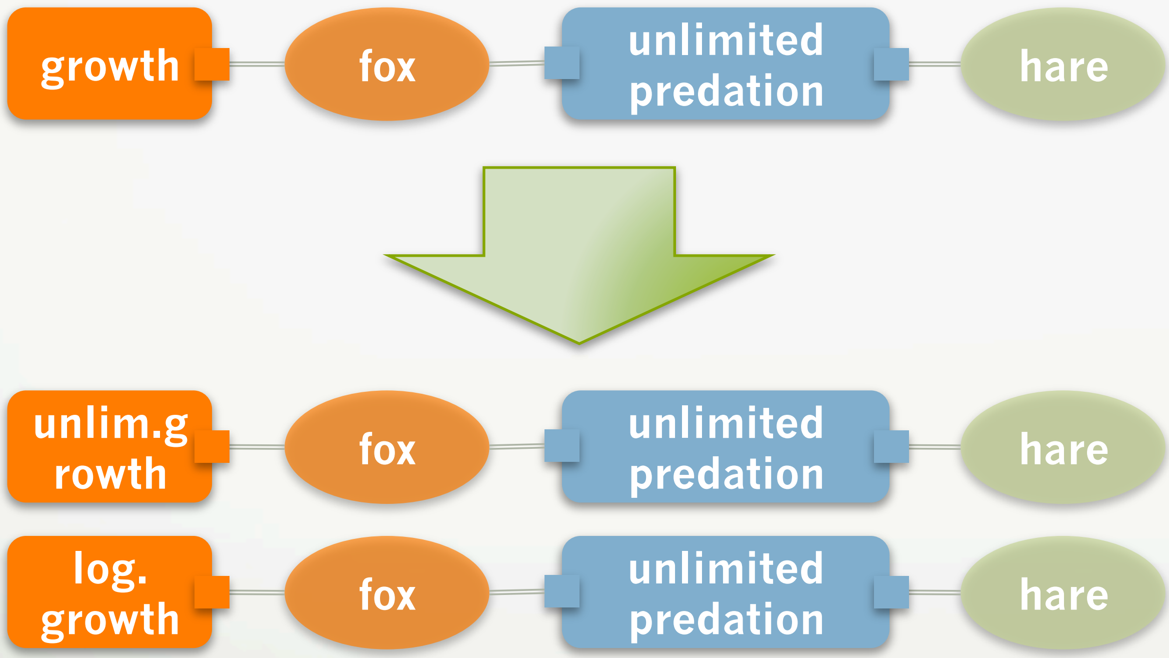
IPM: Generate Models

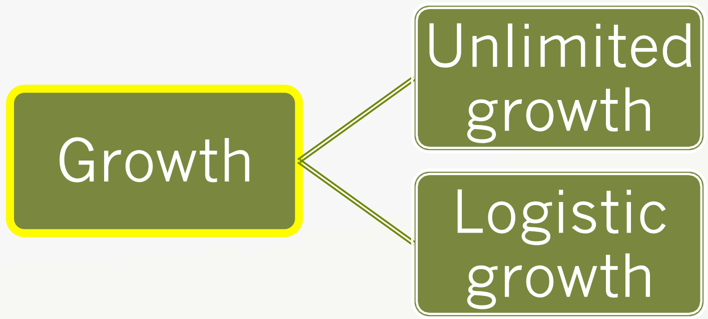


IPM: Evaluate Each Model



IPM: Generate Models (2)





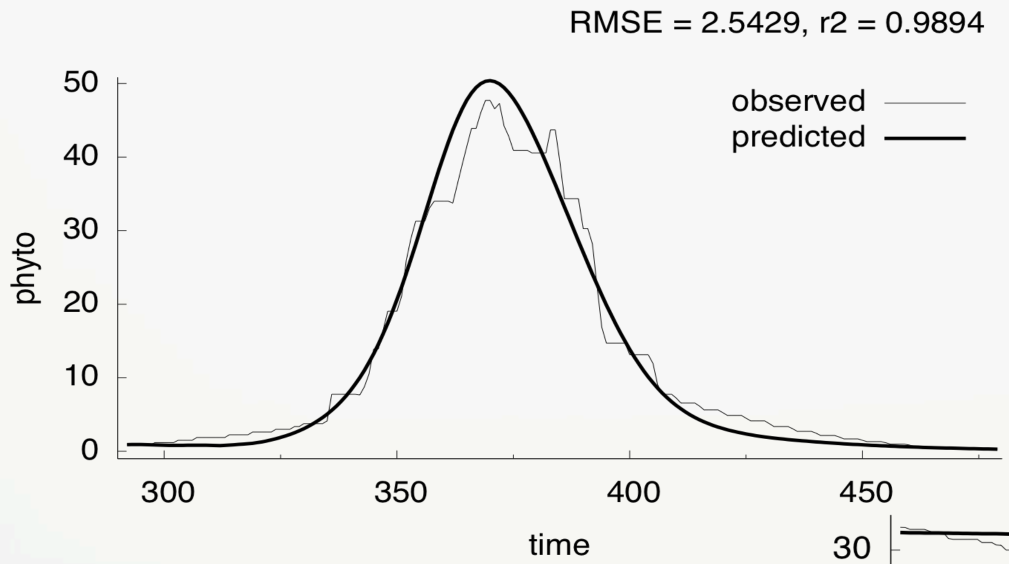
etc.

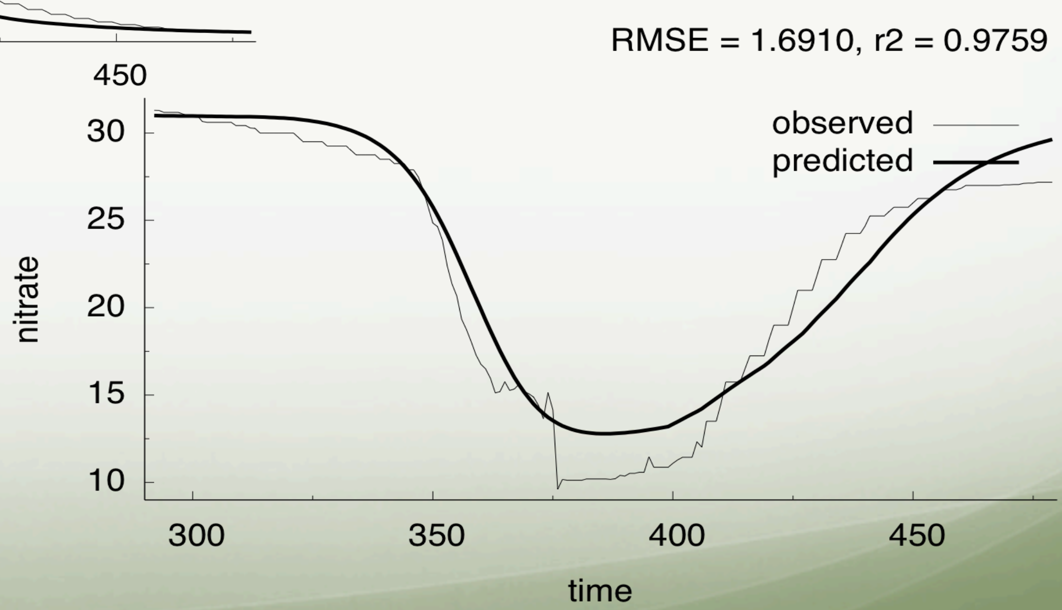
Process Models of Population Dynamics

Modeling Food Chains in Aquatic Ecosystems

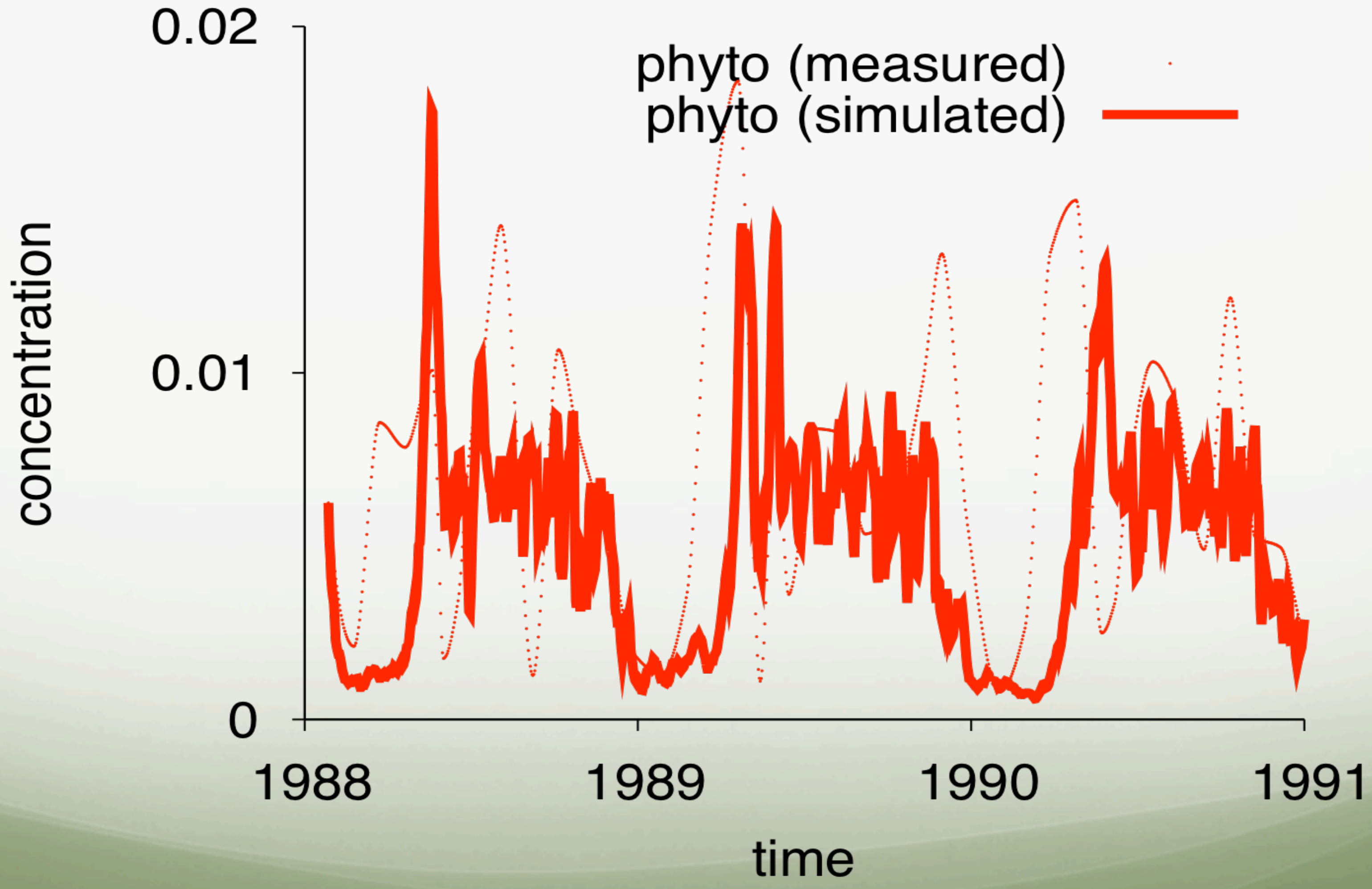
- Modeling library
  - Tens of generic processes related to growth, loss, food limitations of growth, temperature/light influence on growth, mineralization, etc.
  - Variety of different generic entities corresponding to primary producers, animals, inorganic nutrients, etc.
- Variety of data sets modeled
  - Ross Sea, Antarctica
  - Bled Lake, Slovenia
  - Greifensee Lake, Switzerland
  - etc.

Ross Sea, Antarctica

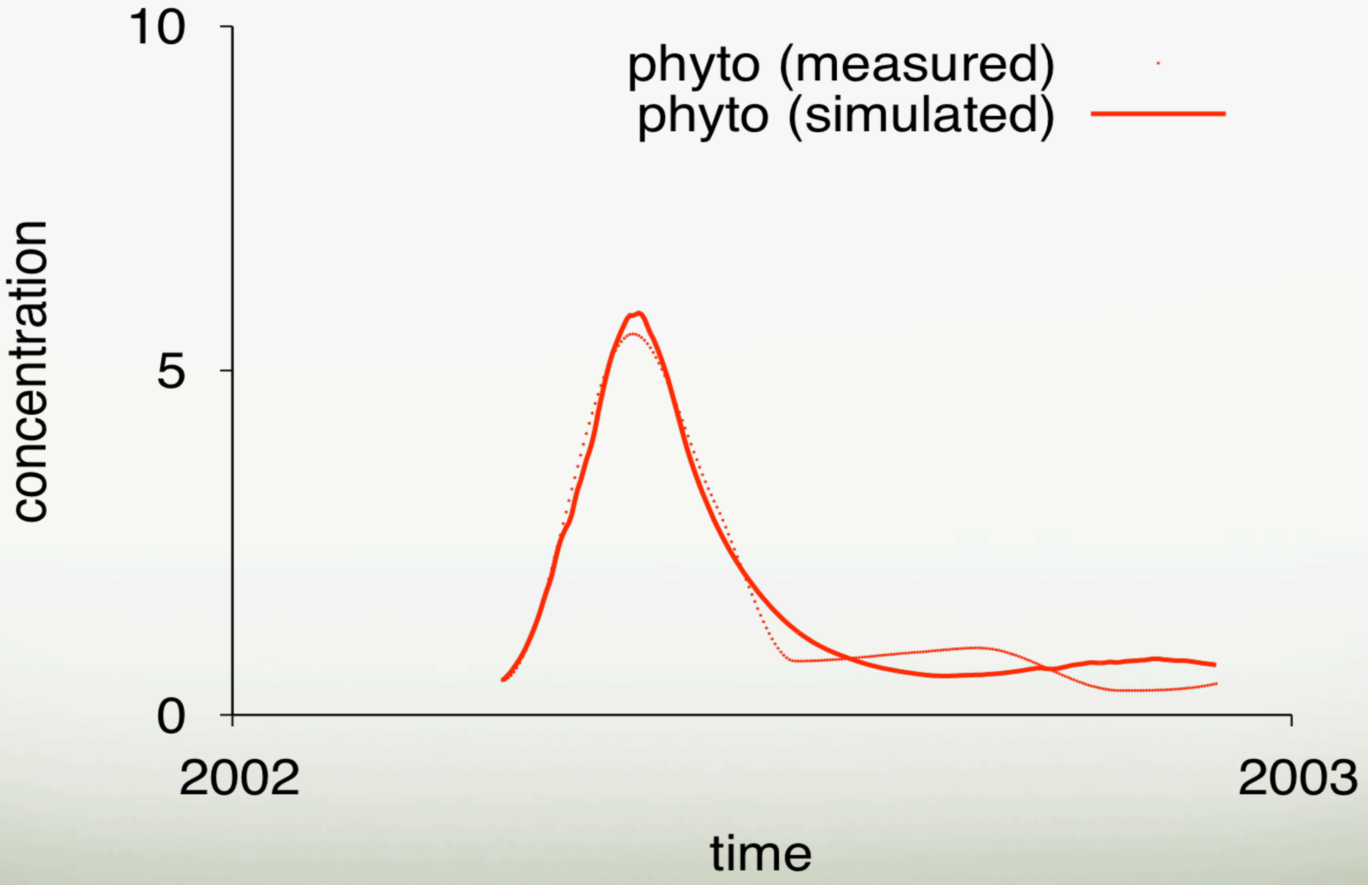




Greifensee Lake, Switzerland



Bled Lake, Slovenia



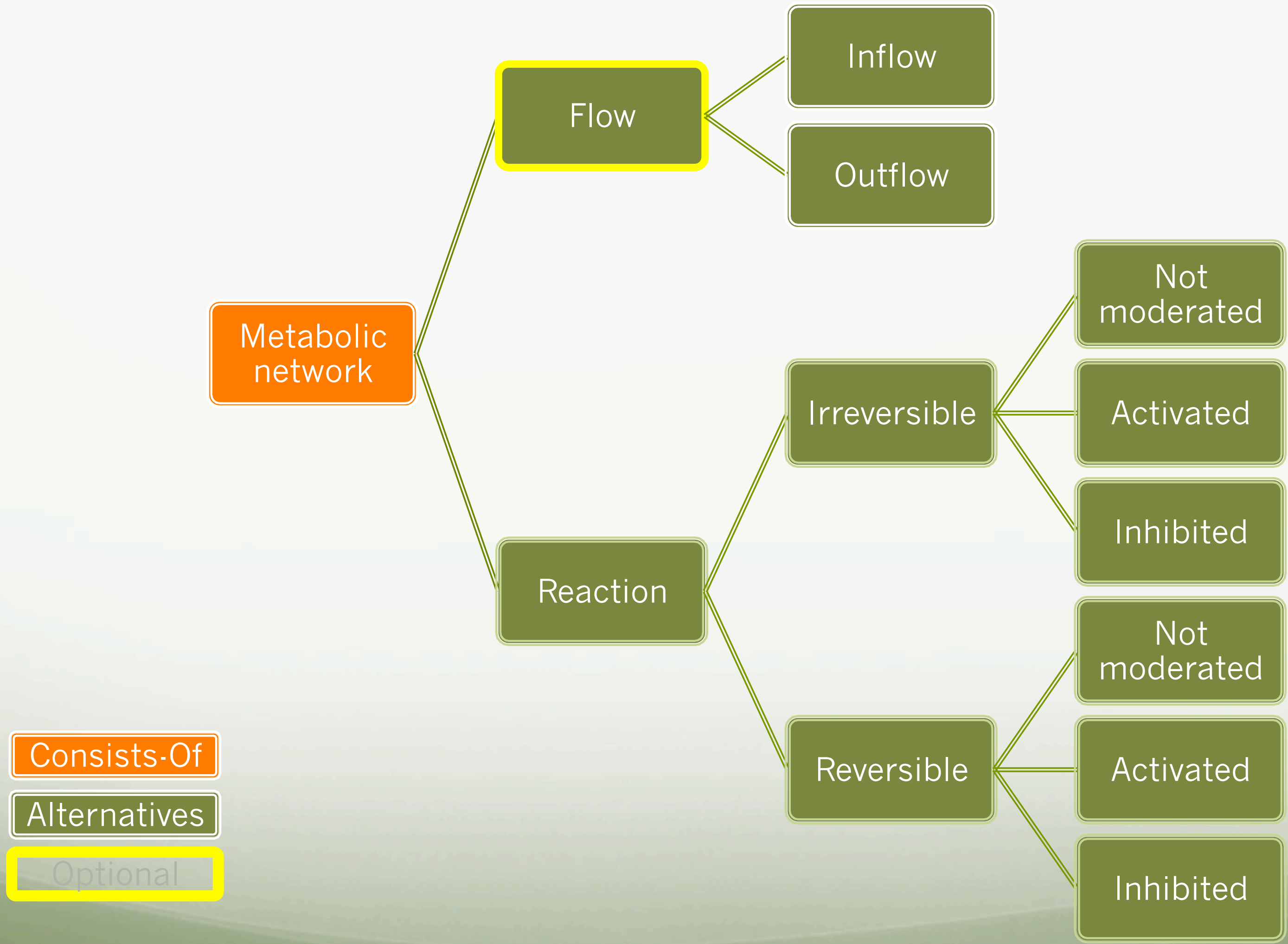
Process Models of Biological Networks

Metabolic Networks and PMs

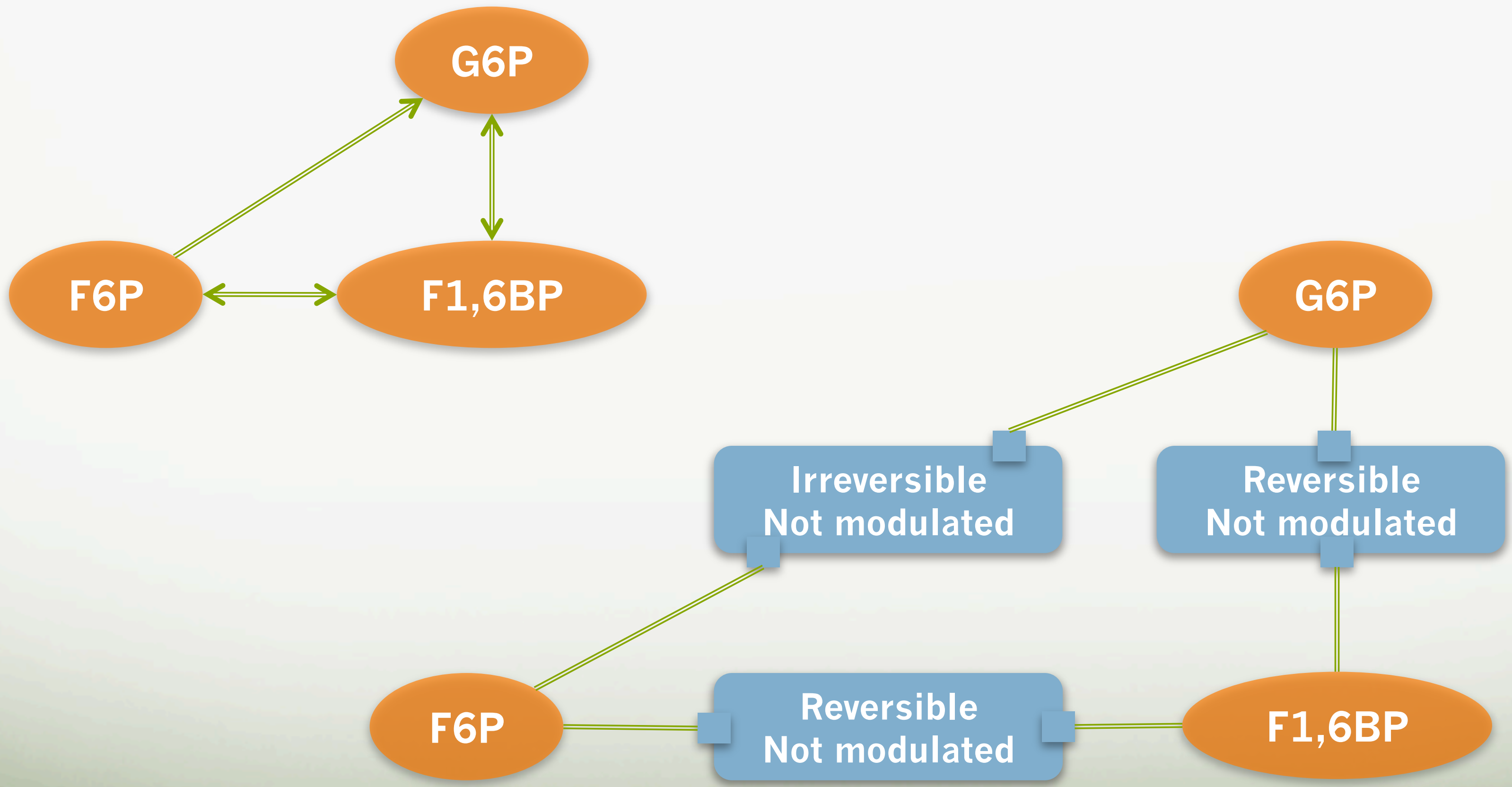
- Entities = chemical compounds
- Processes = chemical reactions

- Entities can have different roles in reactions
  - Substrates are input compounds
  - Products are output constraints
  - Modulates are enzymes that activate/inhibit the reaction

Modeling Knowledge



Simple Example Network



Generic Processes: Irreversible

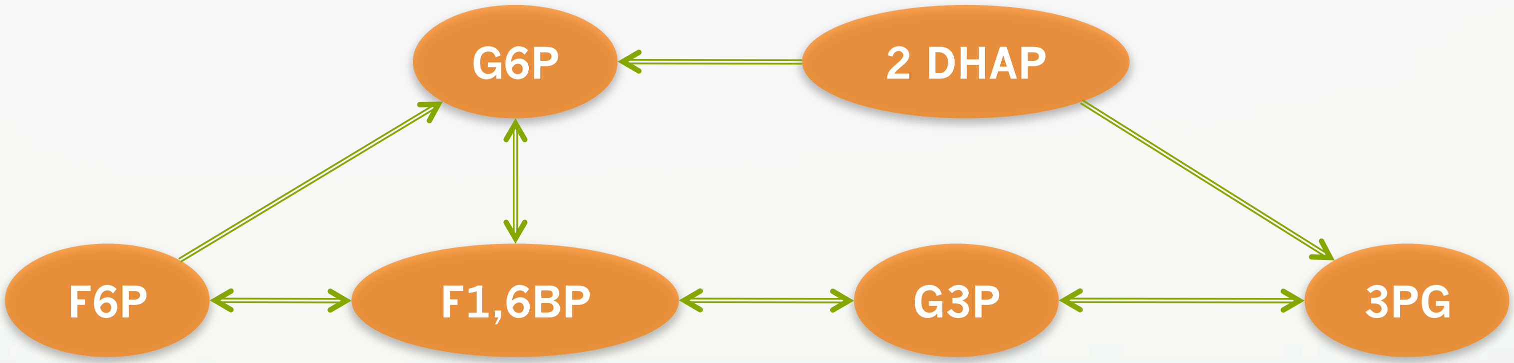
- generic process Irreversible_not_modulated
  - variables *S{compound}*, *P{compound}*
  - constants *reaction_rate(0, Inf)*
  - equations
    - *dS / dt = –1 × reaction_rate × S*
    - *dP / st = reaction_rate × S*

- generic process Irreversible_activated
  - variables *S{compound}*, *P{compound}*, *M{compound}*
  - constants *reaction_rate(0, Inf)*, *modulation_rate(0, Inf)*
  - equations
    - *dS / dt = –1 × reaction_rate × S × S / (S + modulation_rate)*
    - *dP / st = reaction_rate × S × S / (S + modulation_rate)*

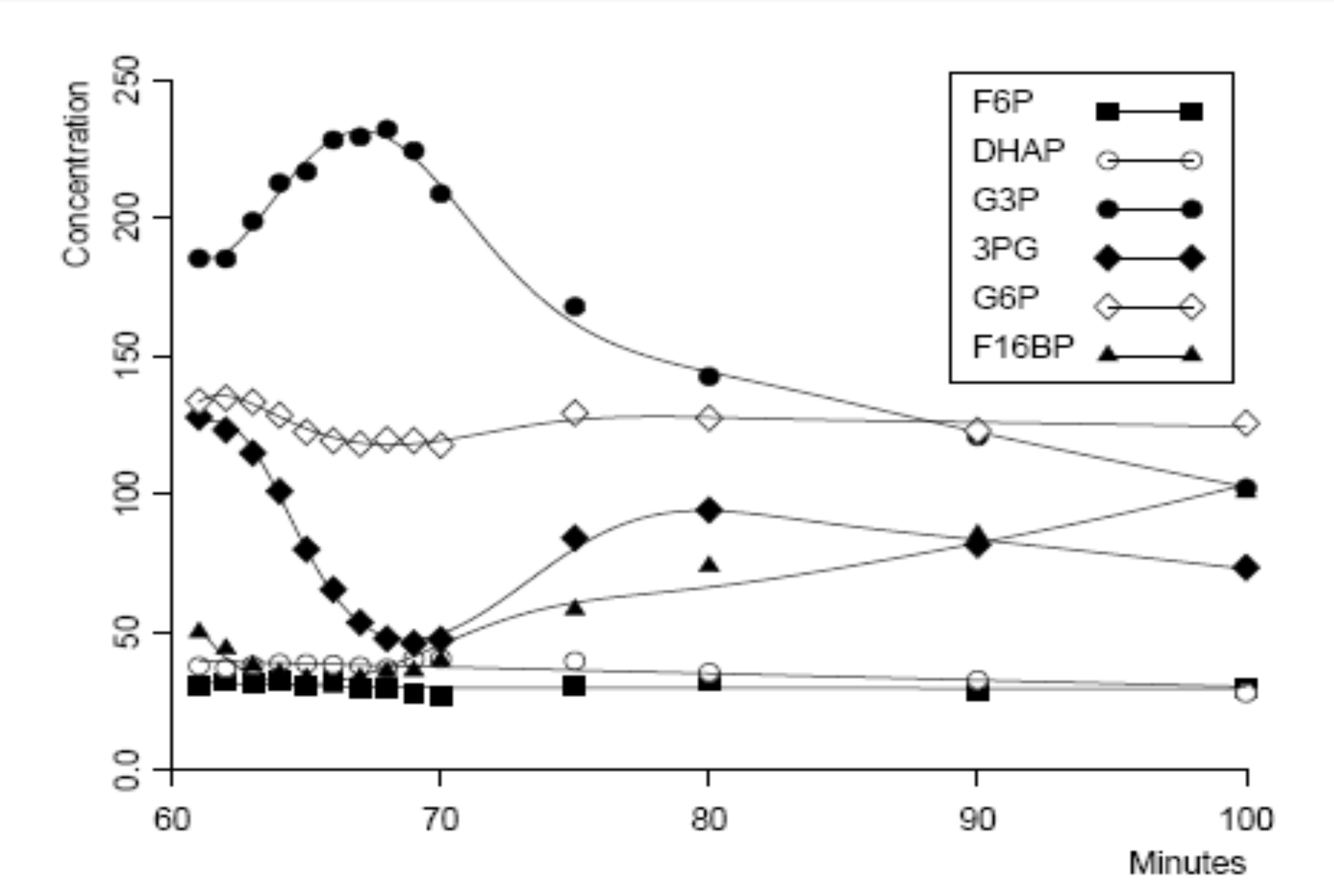
Example Application: Glycosis

- Inducing (partial) chemical network of glycosis
  - Data: temporal responses of species to pulse changes (14 time points)
  - From: Torralba et al. (2003) PNAS 100(4): 1494-1498
- Responses of six chemical compounds:
  - G6P (glucose 6-phosphate)
  - F6P (fructose 6-phosphate)
  - F1,6BP (fructose 1,6-bisphosphate)
  - G3P (glycerol 3-phosphate)
  - 3PG (3-phosphoglycerate)
  - DHAP (dihydroxyacetone 3-phosphate)

Induced Glycosis Network



Glycolisys Model Simulation



Summary and Outlook

Summary

- We have developed a variety of approaches to equation discovery, culminating in IPM ones
  - Learn both parameters and structure of ODE models
  - Integrate data-driven and knowledge-driven approaches to modeling through IPM
- We have applied these to many applications in the domain of population biology/ecology
  - Initial applications to problems in systems biology
  - But much work remains to be done

PM Advantages

- Three important properties of process models
  - data integration (generality): integrating large data sets in a single entity
  - prediction: predicting future system behavior
  - explanation: revealing the processes that govern the system behavior
- In comparison with other modeling approaches
  - Other focus on generality and prediction accuracy
  - IPM takes care about explanatory power of models

The Promise of IPM for SB: Support for Scenarios Like

- Scientific task: Construct model of a system, for which some measurements are available
- 1. First, find a model from the existing literature that has been constructed for a similar system
- 2. Apply this model to the dataset at hand
- 3. Say the fit of the model to the data is bad
  - Revise the model by using the data or construct
  - Induce new model by using data and domain knowledge
- Currently support for individual steps, not process

Challenges/To Do List

- Inducing/Learning models of large networks
- Learning from little data
  - Short time courses
  - Parameter fitting therefrom
- Learning models with unobserved variables
- Developing libraries of modeling knowledge for biological networks
- Full support for scenarios as above

New Project: PHAGOSYS

- Systems biology of phagosome formation and maturation: modulation by intracellular pathogens
- Coordinator: Imperial College London
- Partners:
  - Leiden University Medical Centre
  - Netherlands Cancer Institute, Amsterdam
  - MPI Infection Biology, Berlin
  - MPI Cell Biology and Genetics, Dresden
  - Jozef Stefan Institute, Ljubljana
- Study Mycobacteria and Salmonella
- Construct models (also PBMs) of phagocytosis

References

- Dzeroski S and Todorovski L. (2008) Equation discovery for systems biology: finding the structure and dynamics of biological networks from time course data. *Current Opinion in Biotechnology, 19*: 360–368.
- Bridewell W, Langley P, Todorovski L, and Dzeroski S. (2008) Inductive process modeling. *Machine Learning, 71*: 1–32.
- Langley P, Shiran O, Shrager J, Todorovski L, and Pohorille A. (2006) Constructing explanatory process models from biological data and knowledge. *Artificial Intelligence in Medicine, 37*: 191–201.
- Dzeroski S and Todorovski L. (2007) *Computational Discovery of Scientific Knowledge*. Springer.
- Todorovski L, Bridewell W, Shiran O, & Langley P. (2005) Inducing hierarchical process models in dynamic domains. Proceedings of the Twentieth National Conference on Artificial Intelligence, 892–897.

Thanks

- For your attention and the questions/feedback

- To my collaborators
  - **Saso Dzeroski**, Jozef Stefan Institute
  - **Pat Langley and Will Bridewell**, Stanford University
  - **Boris Kompare and Natasa Atanasova**, University of Ljubljana, Faculty of Civil and Geodetic Engineering