Towards A General Theory of Learning with Models & Data

"An Information Theory Perspective on Uncertainty & Learning"

Hoshin V Gupta (University of Arizona)

2017 RE Horton Lecture, Annual Meeting of the American Meteorological Society

Seattle, Washington, Jan 25

Thank You

Information Theory Perspective on System Identification

AGU PUBLICATIONS

Water Resources Research

COMMENTARY
10.1002/2013WR015096

Debates—The future of hydrological sciences: A (common) path forward? Using models and data to learn: A systems theoretic perspective on the future of hydrological science

Hoshin V. Gupta[1] and Grey S. Nearing[2,3]

AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2014WR015895

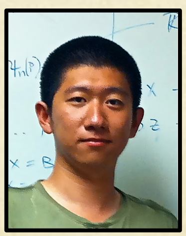
The quantity and quality of information in hydrologic models

Grey S. Nearing[1] and Hoshin V. Gupta[2]

Some Main Collaborators

Grey Nearing   Steven Weijs   Uwe Ehret

Beth Jackson



Gong Wei

Jasper Vrugt

Martyn Clark

Gab Abramowitz

Ben Ruddell

Shervan Gharari

In Particular

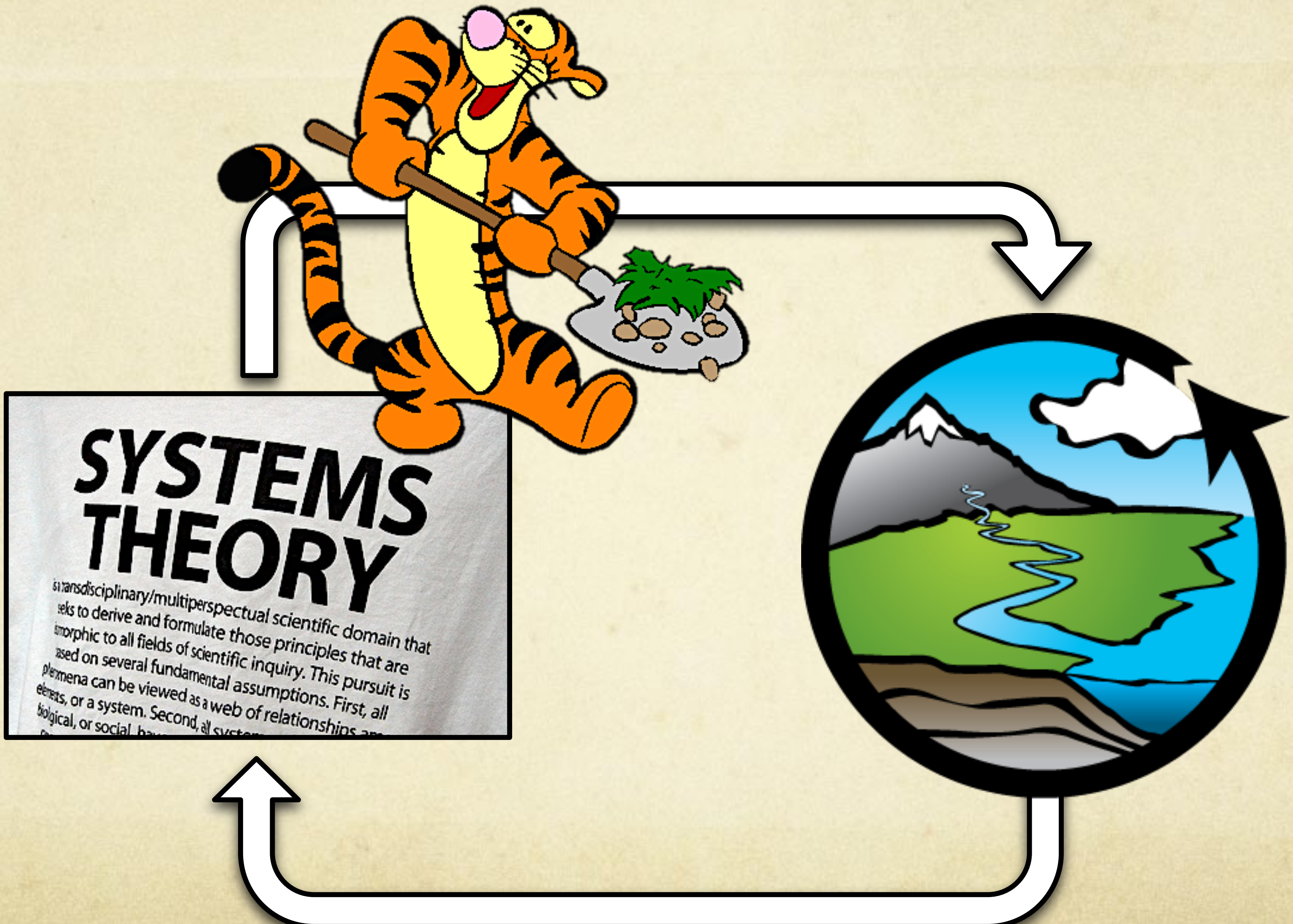
Grey Nearing

HYDROLOGICAL SCIENCES JOURNAL – JOURNAL DES SCIENCES HYDROLOGIQUES, 2016
VOL. 61, NO. 9, 1666–1678
http://dx.doi.org/10.1080/02626667.2016.1183009
Special issue: *Facets of Uncertainty*

A philosophical basis for hydrological uncertainty

Grey S. Nearing[a,b], Yudong Tian[b,c], Hoshin V. Gupta[d], Martyn P. Clark[a], K

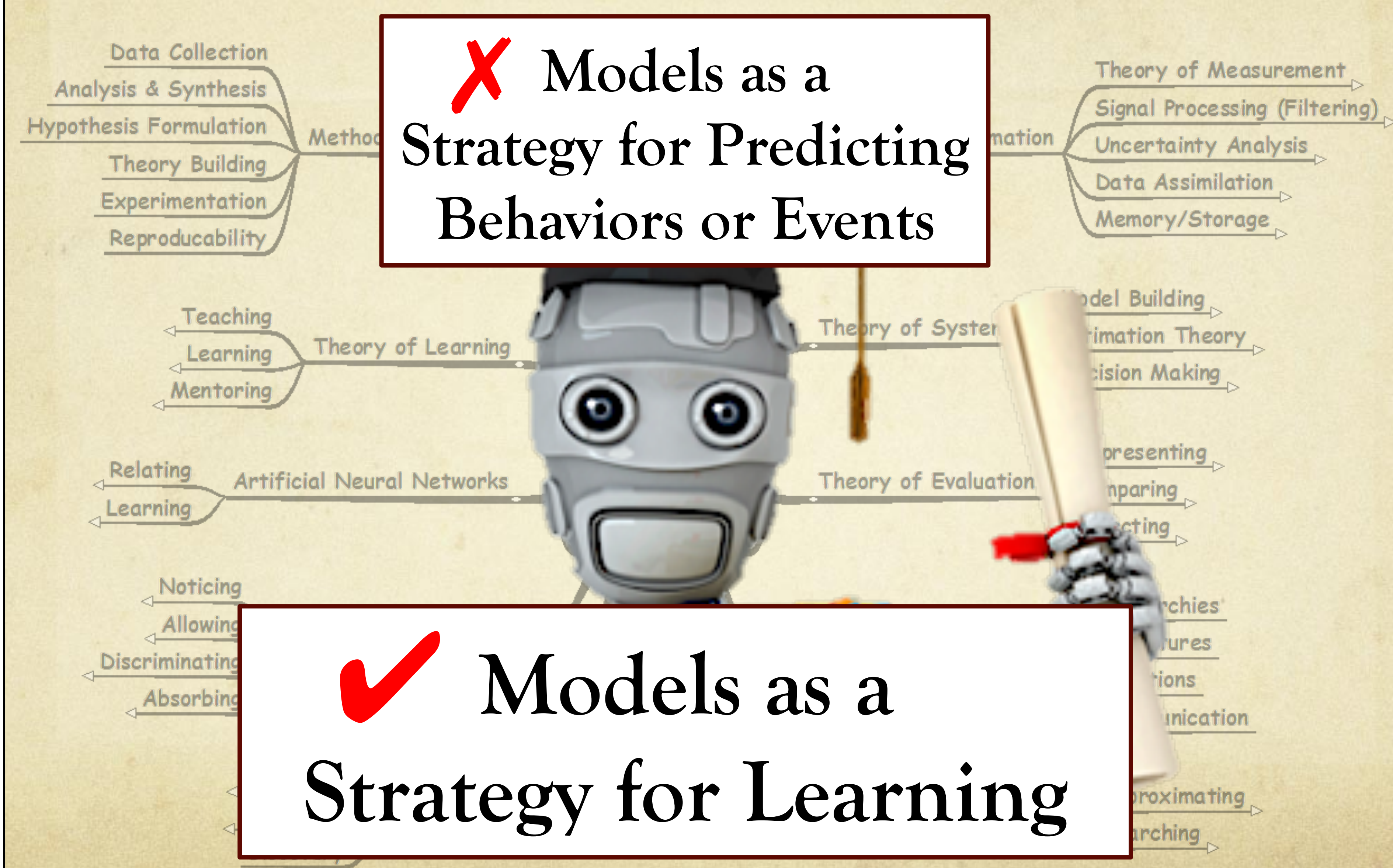
My Background

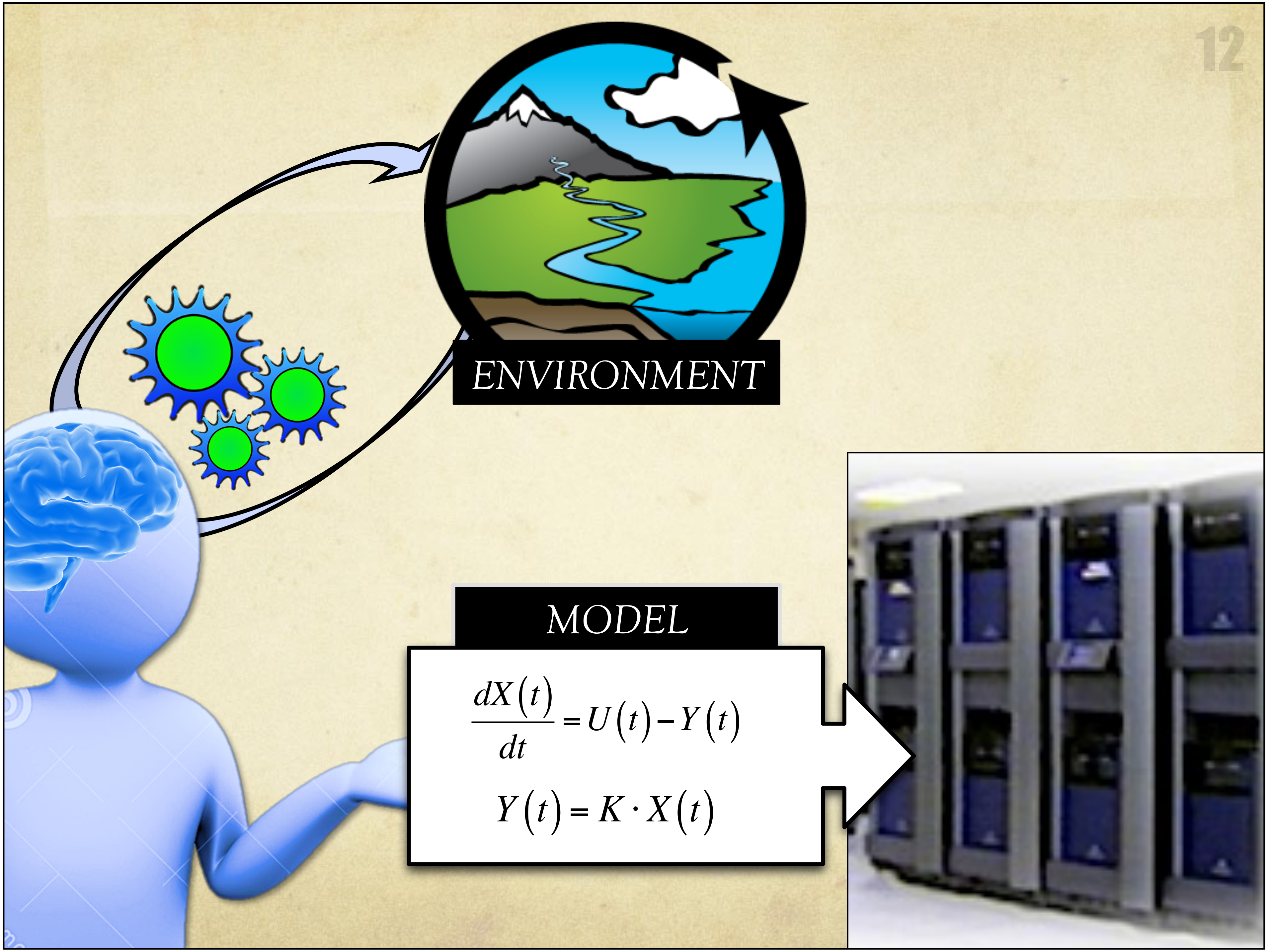


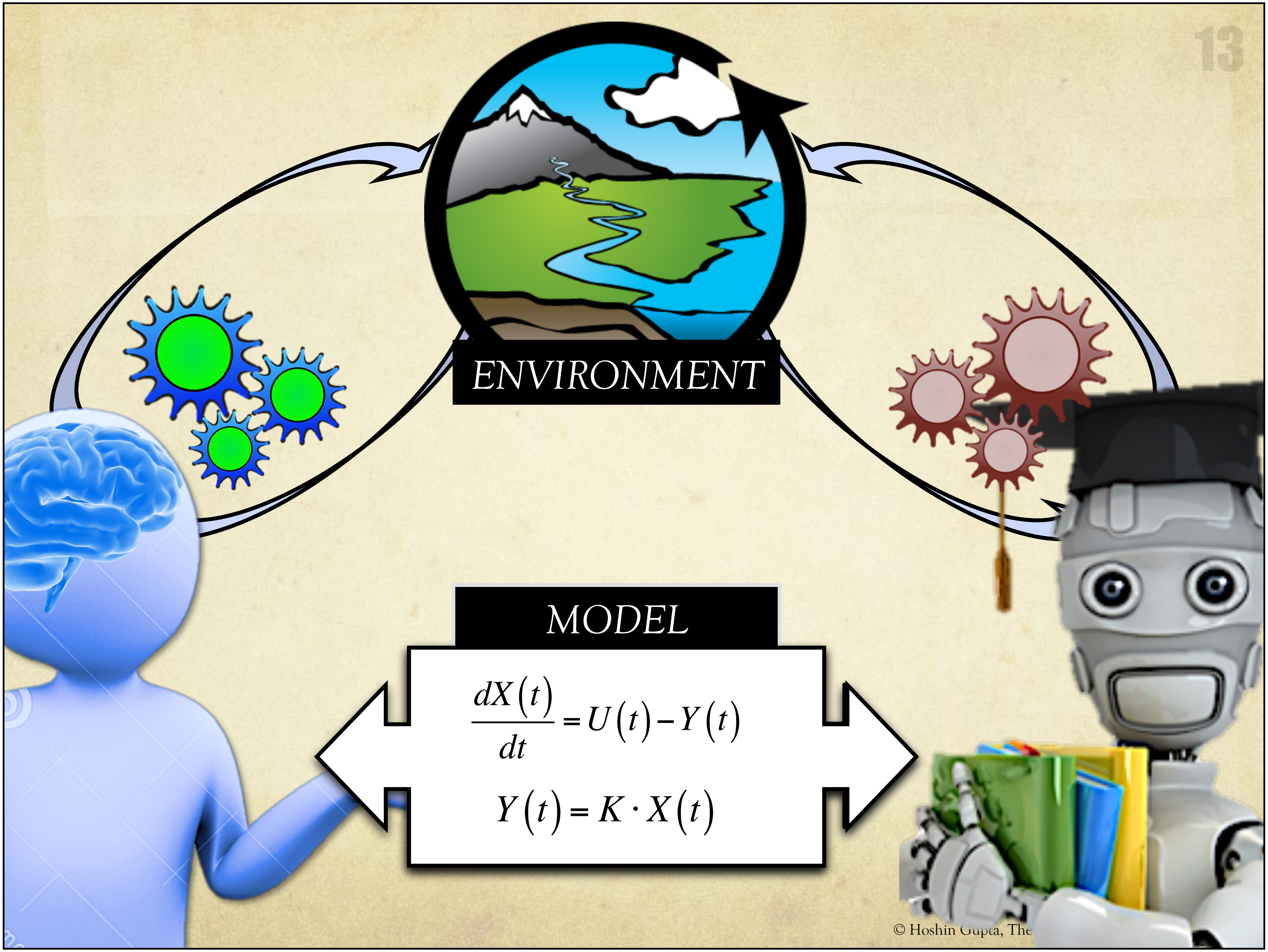
My Interest is the "Learning" Problem

✗ Models as a Strategy for Predicting Behaviors or Events

✔ Models as a Strategy for Learning




12
ENVIRONMENT
MODEL
$$\frac{dX(t)}{dt} = U(t) - Y(t)$$
$$Y(t) = K \cdot X(t)$$


13
ENVIRONMENT
MODEL
$\frac{dX(t)}{dt} = U(t) - Y(t)$
$Y(t) = K \cdot X(t)$
© Hoshin Gupta, The

OUTLINE

1. Models - Dynamical Environmental Systems (DES)
    - *DES models as Tools for Scientific Investigation*
2. Information - Its Fundamental Nature & Different Kinds
    - *Data, Models and Assumptions are different, but related, kinds of (Uncertain) Information*
3. Learning - How Information & Uncertainty are Related
    - *Learning involves "Change". Can Information be "Bad" ?*
4. Structure of Information - How DES Models Encode Knowledge
    - *Information is encoded as a Structured Hierarchy of Hypotheses*
5. Model "Failure" - What Could Possibly Go Wrong ?
    - *Bias and Overconfidence*
6. Inference – Learning From Our Mistakes
    - *Different kinds of model-based Learning*
7. Maximum Entropy Approach – Injecting Rigor into Learning
    - *Dealing with Uncertainty in System Architecture & Process Parameterization*
8. What Next? – Improving The Way We Use Models in Science

I

Models

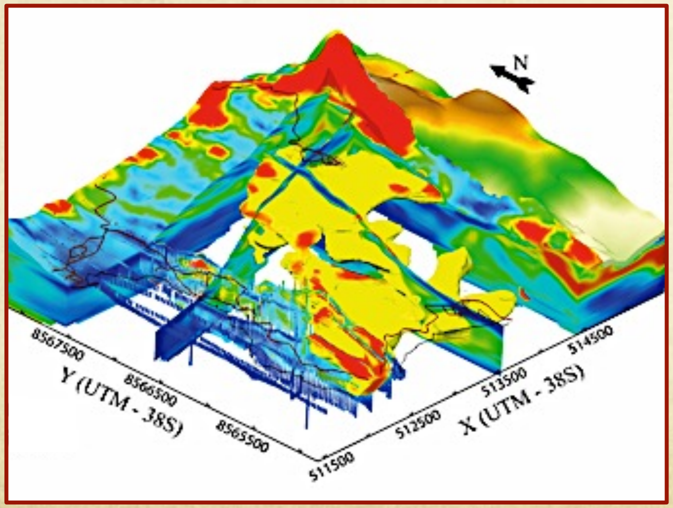
Dynamical Environmental System Models (DESMs)

Working Definition:

A **DESM** *is a* **simplified representation** *of the* **structure & function** *of a dynamical system that:*

1. Enables
(a) **Simulations** *that are acceptably accurate*
(b) **Testable Predictions** *under new circumstances*
(c) **Reasoning** *within an idealized framework*

2. By Encoding Knowledge about
(a) **Physics** *(conservation, thermodynamics, etc.)*
(b) **System Properties** *(Geometry & Materials etc.)*
(c) **Uncertainty** *(What we know that we don't know)*



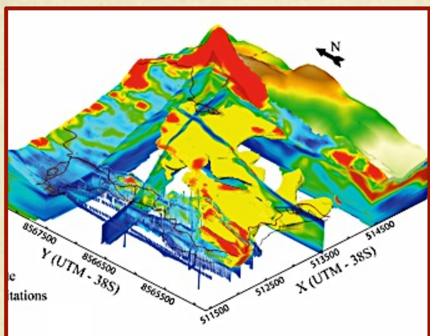
Why Simplified?

a) *Knowledge is* **Incomplete & Uncertain**
b) *Real system is* **Infinite Dimensional**
c) *Need to "compute" in* **Finite Time using Finite Resources**

We Use DESMs for Scientific Investigation

<u>Intuitively</u>

We understand that "**Models**" & "**Data**" codify **Knowledge** about the World … in the form of <u>*Information*</u>



How ?

How can I use this insight to improve my models ?

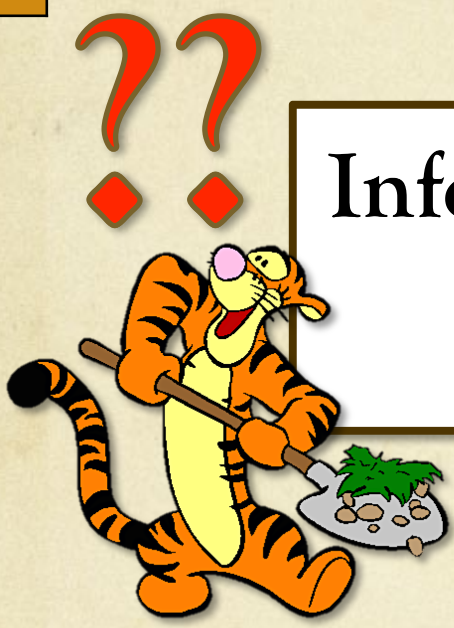
How does "Learning" happen ?

1

2

Information

What is The Nature of Information ?



Info is the answer to questions such as:

When, What, Where, How, Why ...

& How Much ... etc.

Information is Relational / Contextual

Info is always "<u>about</u>" something

Context Matters

DATA is not Info … until viewed in context

3.14

Information is Relational / Contextual

Info is always "<u>about</u>" something

Context Matters

DATA is not Info ... until viewed in context

3.14

Streamflow

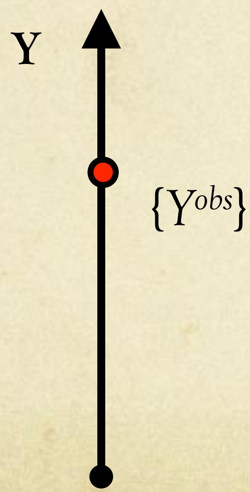
(mt^3/sec)

So Information is always "ABOUT" Something …

In the Context of Model-Data Learning

Info is always *about*

- Values (Y)
- Relationships (R: X ➔ Y)
- Constraints (C) … *Assumptions are a kind of Constraint*



A single DATA Point encodes Info about:

A Value of "Something"

1

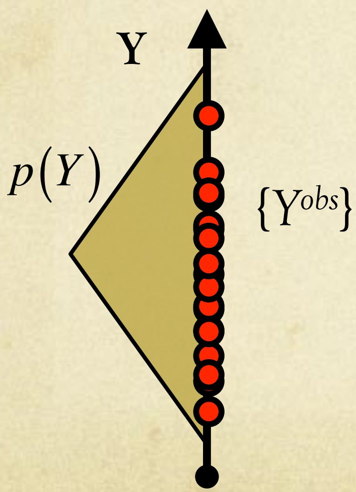
2

So Information is always "ABOUT" Something …

In the Context of Model-Data Learning

Info is always *about*

- Values (Y)
- Relationships (R: X➔Y)
- Constraints (C) … *Assumptions are a kind of Constraint*



A set of DATA Points encodes Info about:

The Distribution of Values

1

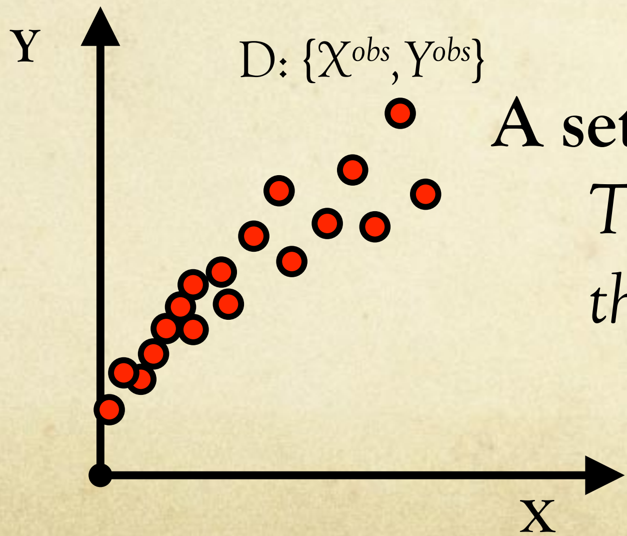
2

So Information is always "ABOUT" Something …

In the Context of Model-Data Learning

Info is always *about*

- Values (Y)
- Relationships (R: X➔Y)
- Constraints (C) … *Assumptions are a kind of Constraint*



A set of DATA Points encodes Info about:

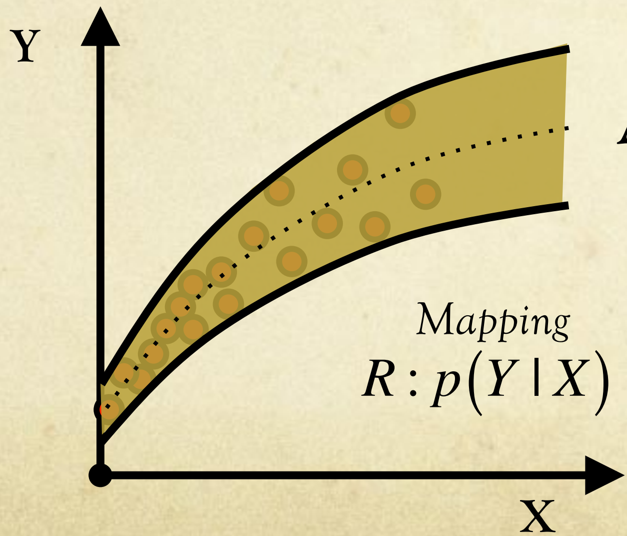
The space-time-ordered Relationships among those values

So Information is always "ABOUT" Something …

In the Context of Model-Data Learning

Info is always *about*

- Values (Y)
- Relationships (R: X➔Y)
- Constraints (C) … *Assumptions are a kind of Constraint*



A MODEL encodes Info about:

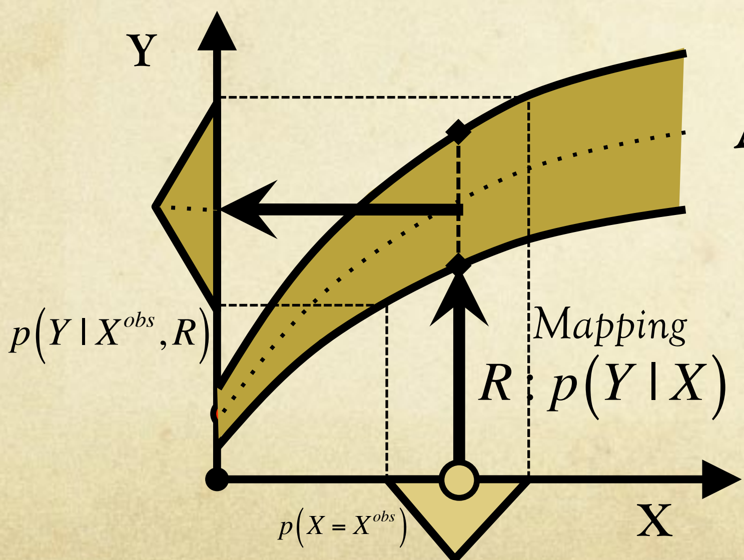
The space-time-ordered Relationships between variables 'behind' those values

So Information is always "ABOUT" Something …

In the Context of Model-Data Learning

Info is always *about*

- Values (Y)
- Relationships (R: X➔Y)
- Constraints (C) … *Assumptions are a kind of Constraint*



A MODEL encodes Info about:

Values p(Y|X,M) conditional on Relationships and Data

1

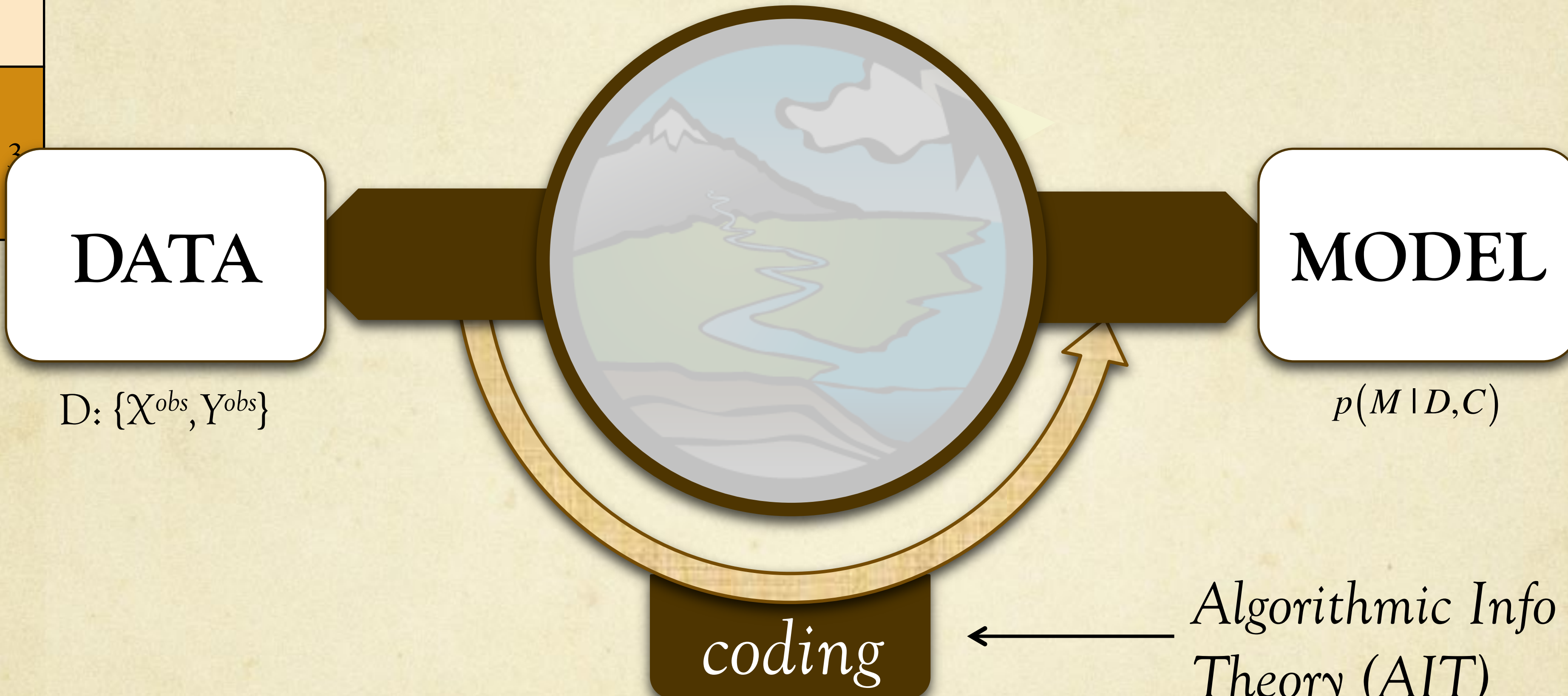
2

3

Learning

Simply put … Learning Involves

Converting Data-Info → Model-Info



D: $\{X^{obs}, Y^{obs}\}$

$p(M \mid D, C)$

Algorithmic Info Theory (AIT)

** *Info can be added in/by the "Conversion Process" (have to be very careful here)*

** *The conditioning role of Assumptions (A) is very strong*

1

2

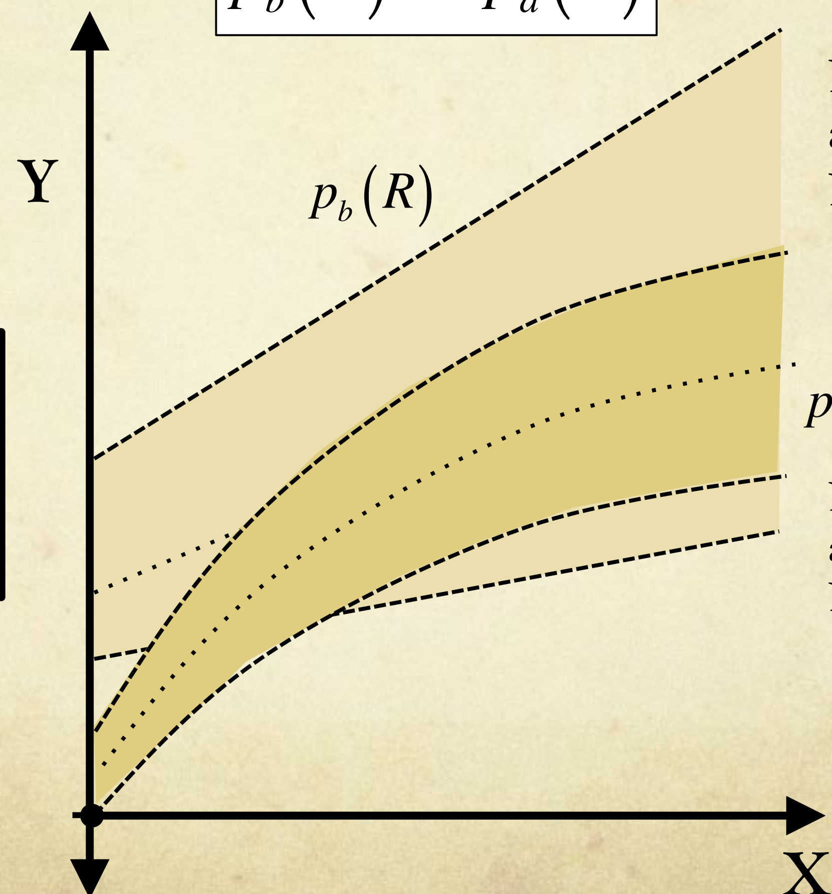
3

And Occurs When …

Our <u>Prior Uncertainty is Changed</u>

due to the Assimilation of New Information

$$\boxed{p_b(R) \Rightarrow p_a(R)}$$

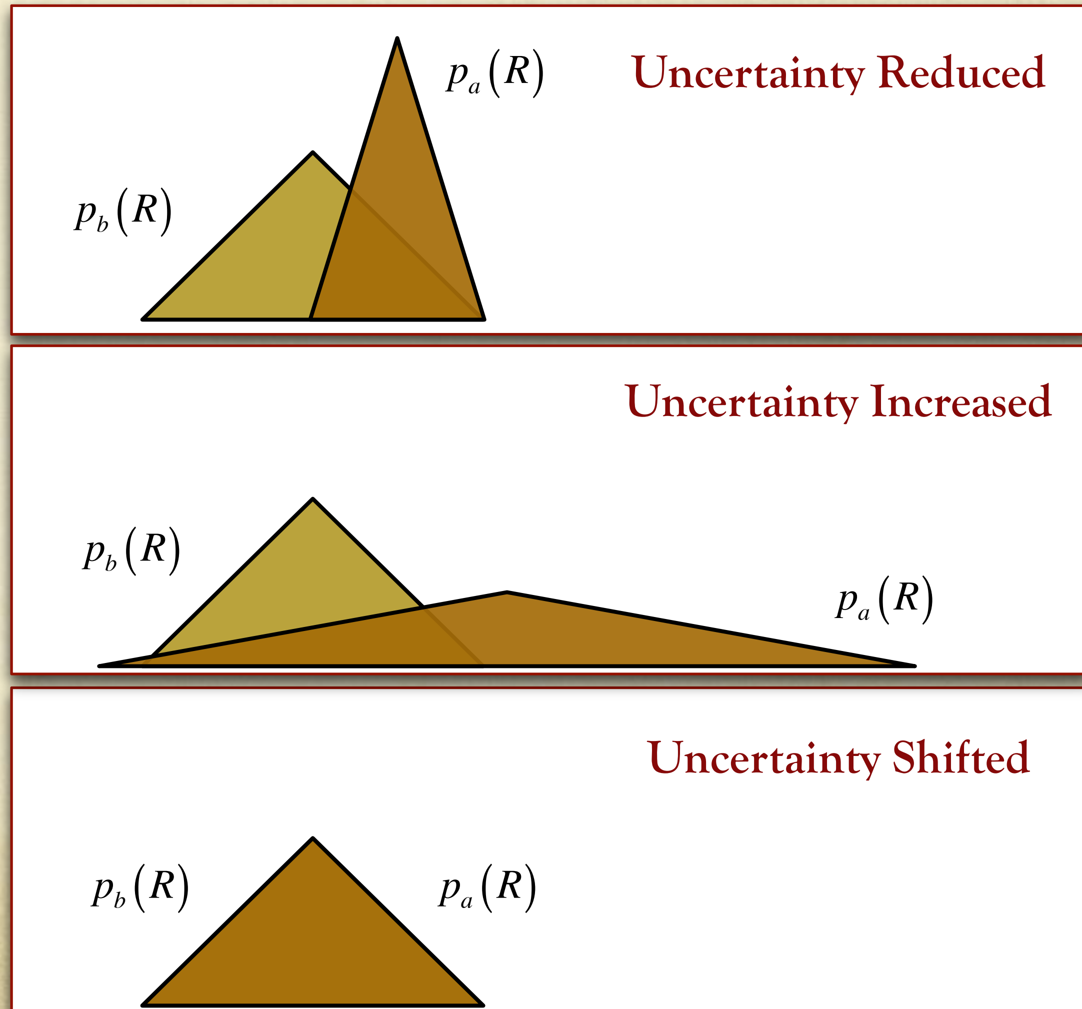


"<u>Changed</u>" …
not Reduced

1
2
3

Kinds of Learning …

Learning occurs when our **<u>Prior Uncertainty is Changed</u>**
due to the Assimilation of New Information

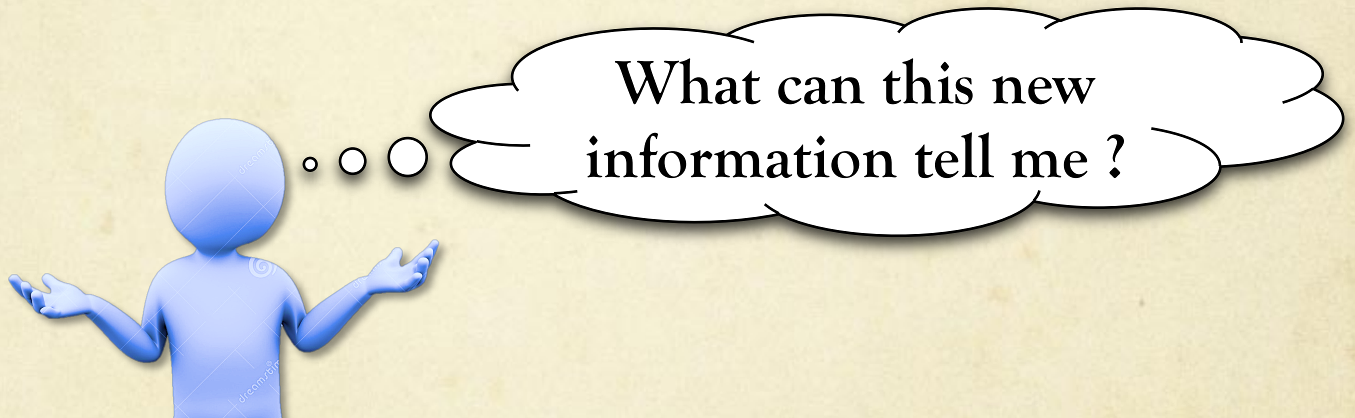


"Changed" …
not Reduced

Can Information Can Be **Bad** ?

There is a view (advanced in the literature) that Information can be "Bad" (so called "Mis-Information")

But Information is Simply What it Is ...
<u>*What Can be Suspect is* **Our Interpretation**</u>



How to Handle Different Kinds of Information is the focus of ESTIMATION THEORY

1

2

3

4

The Structure of Information

(in DES Models)

DESM's Encode Information As a Hierarchical Sequence of Decisions

1. *Control Volume, Physics, Processes to Include, System Geometry & Material Properties*

2. *Scale, Dimension & 3D Spatial Structure*

3. *Process Relationships*

4. *Uncertainty*

5. *Solution Methodology*

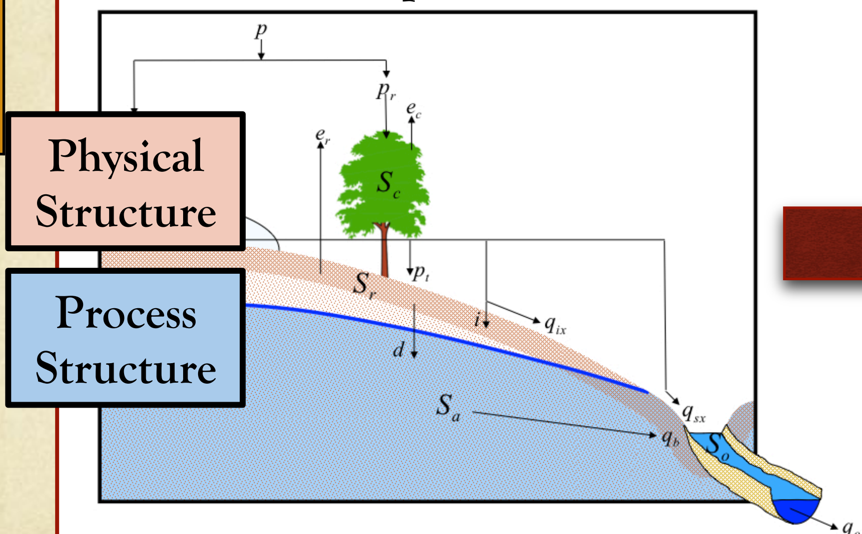
Step One – Conservation Laws

Information About:

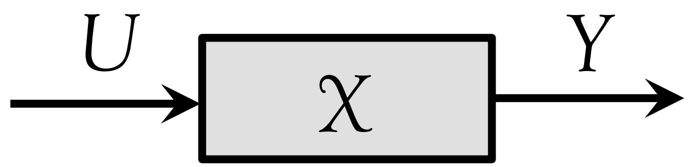
1. Physics (*conservation*, *thermodynamics*), Physical Processes, System Geometry & Material Properties to include

Question: *What knowledge to encode and what to ignore?*

Conceptual Model



System Diagram



$$\frac{dX_t}{dt} = U_t - Y_t$$

Conservation

$$\{U_t, X_t, Y_t\} \geq 0$$

$$X_t \geq Y_t$$

Constraints

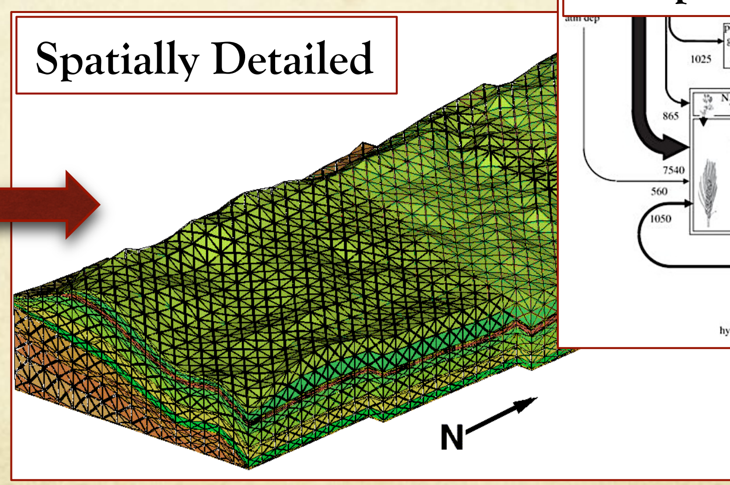
Result: *A System Diagram and Conservation Law Hypothesis - H^{CL}*

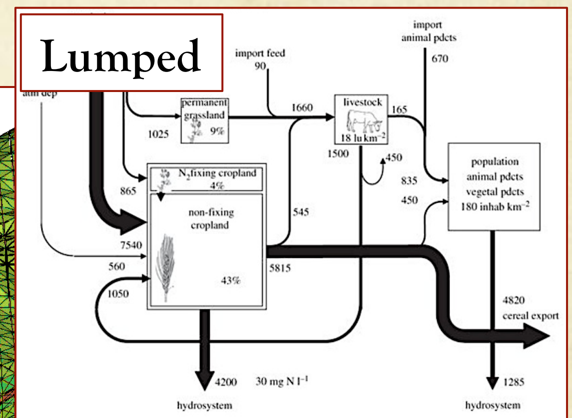
Step Two – System Architecture

<u>Information About:</u>

2. Scale, Dimension and *3D* Spatial Structure of the *State-Space* (elements), to enable finite computation

Question: *What is a sufficiently complex, <u>finite dimensional</u>, spatially organized representation of sub-system architecture?*





Result: *A System Architecture Hypothesis - H^{SA}*

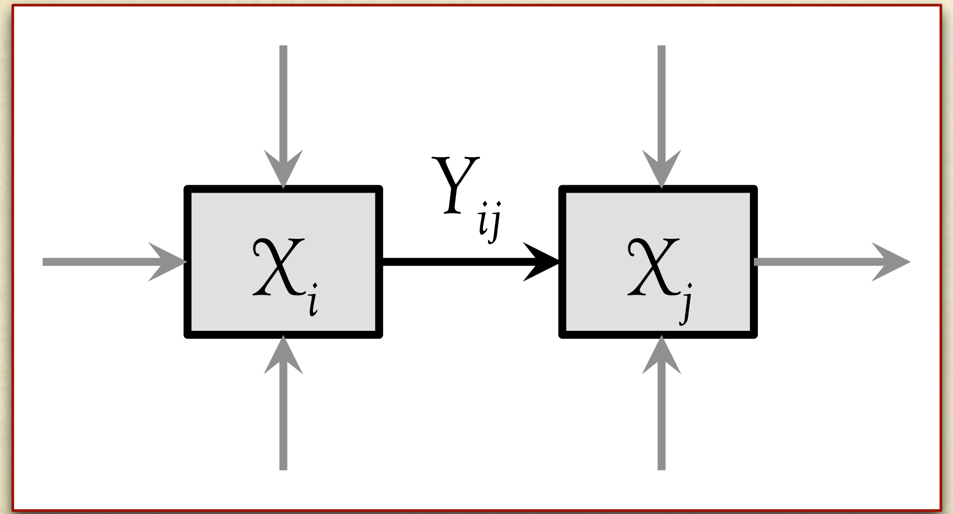
Step Three – Process Parameterization

1
2
3
4

Information About:

3. Process Relationships via *Equations*, that account for Sub-Element Process & Material Heterogeneity

Question: *What mathematical forms to use for the Process Parameterization equations, at the architectural scale of interest?*



$$Y_{ij} = \overset{?}{f}_{xy}\left(\Delta X_{ij} \mid \theta_{xy}\right)$$

Parameters are artifacts of equation choice

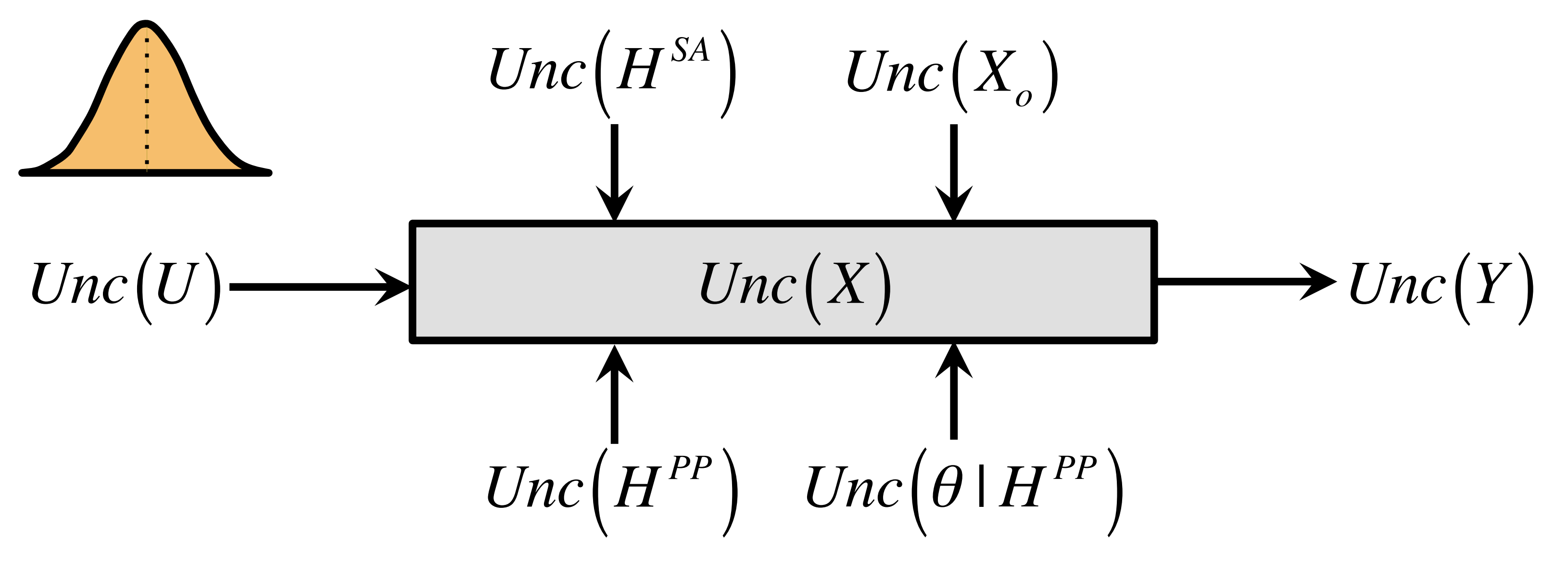
Result: *A Process Parameterization Hypothesis - H^{PP}*

Step Four – Uncertainty

Information About:

4. What we KNOW that we do not know precisely (or at all)

Question: *What uncertainties are important, and how to represent them mathematically?*



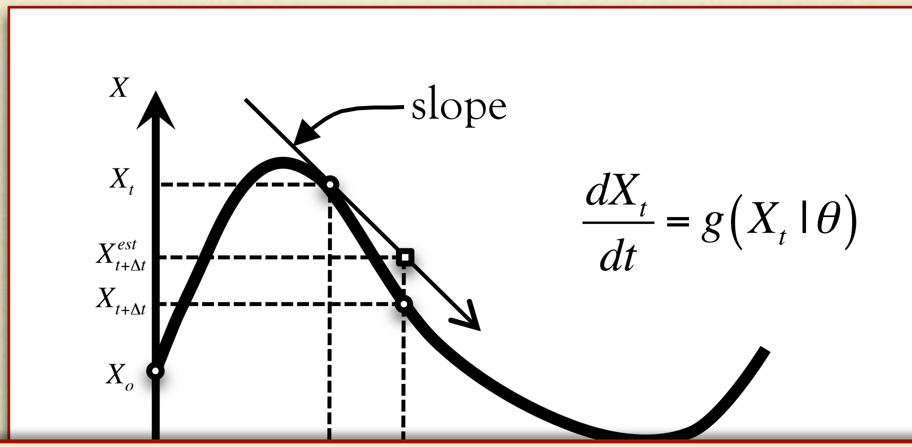
Result: *An Uncertainty Hypothesis - H^{UN}*

Step Five – Solution Procedure

Information About:

5. Procedure for 'Solving' the resulting *Mathematical Model*

Question: *How to Integrate (in space & time) the resulting system of (stochastic) differential equations?*



Result: *A Computational Model*

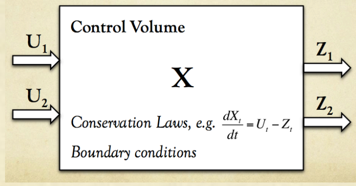
→ Practical manifestation of the Overall System Hypothesis - H^{OS}

→ Structured hierarchy of Conservation Law, System Architecture, Process Parameterization, and Uncertainty Hypotheses - $H^{OS} = \{H^{UN} \mid H^{PP} \mid H^{SA} \mid H^{CL}\}$

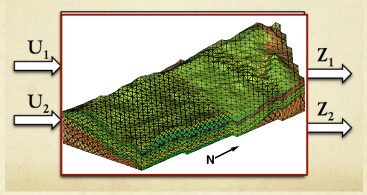
Information is Added at Each Step

(Uncertainty is Changed)

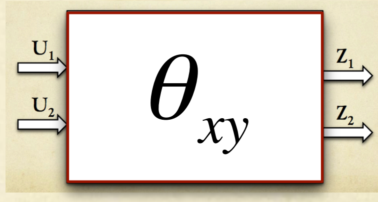
1. Conservation Laws *restrict possible* **U-X-Y** *trajectories*



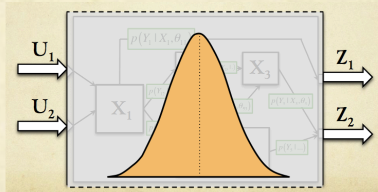
2. System Architecture (a) *further restricts trajectories* & (b) *determines spatial variability*



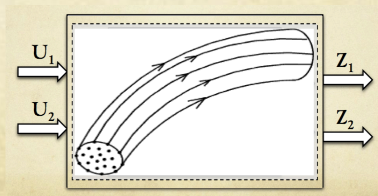
3. Process Parameterization (a) *further restricts trajectories* & (b) *introduces "tunable" parameters*



4. Specification of Uncertainty *characterizes and quantifies "known unknowns"*



5. Solution Procedure *converts Model Info & Input Info into specific (uncertain)* **X-Y** *trajectories*



1
2
3
4
5

Model "Failure"

What Could Possibly Go Wrong ?

1
2
3
4
5

What Could Go Possibly Wrong?

1. Problem Becomes Over-Constrained

Due to <u>Hypotheses that are Unjustifiably Strong</u>

a) Neglect Heterogeneity that is important

b) Over-simplify the System Architecture

c) Incorrect Process Equations forms

d) Deterministic Process Parameterizations (instead of Stochastic)

2. Problem Remains Under-Constrained

Due to <u>Lack of Knowledge</u>

a) Do not know Process Physics at the scale of system elements

b) Do not know Heterogeneity of Material Properties and Geometry at scale of system elements

c) Do not know (& account for) Heterogeneity of Material Properties and Geometry *<u>at scales smaller than the system elements</u>*

1
2
3
4
5

What Can Go Wrong?

1. **Problem Becomes Over-Constrained**

   Due to <u>Hypotheses that are Unjustifiably Strong</u>

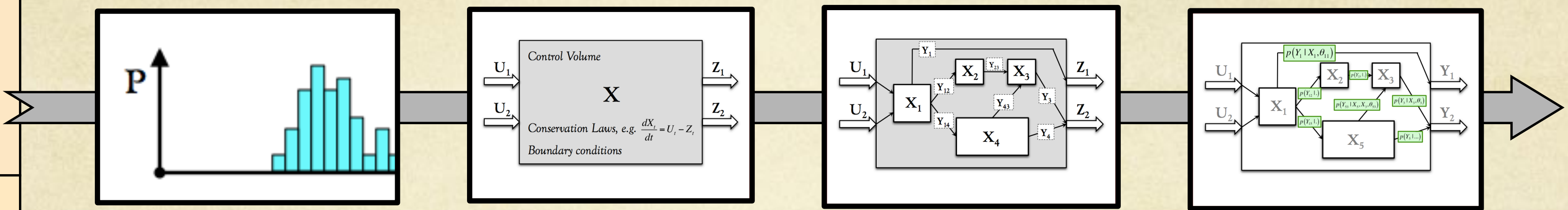
Usually some Combination of Both

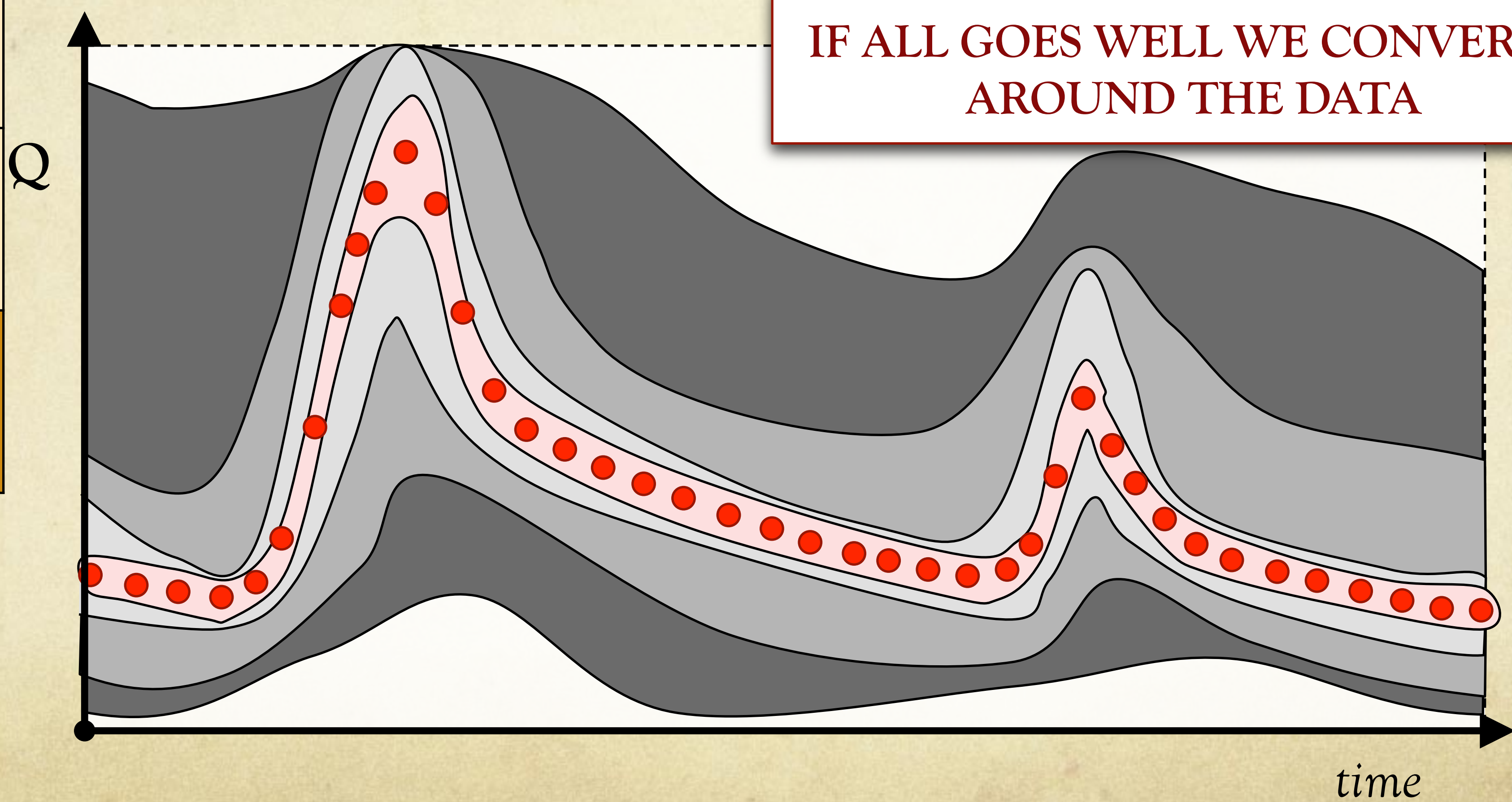
2. **Problem Becomes Under-Constrained**

   Due to <u>Lack of Knowledge</u>

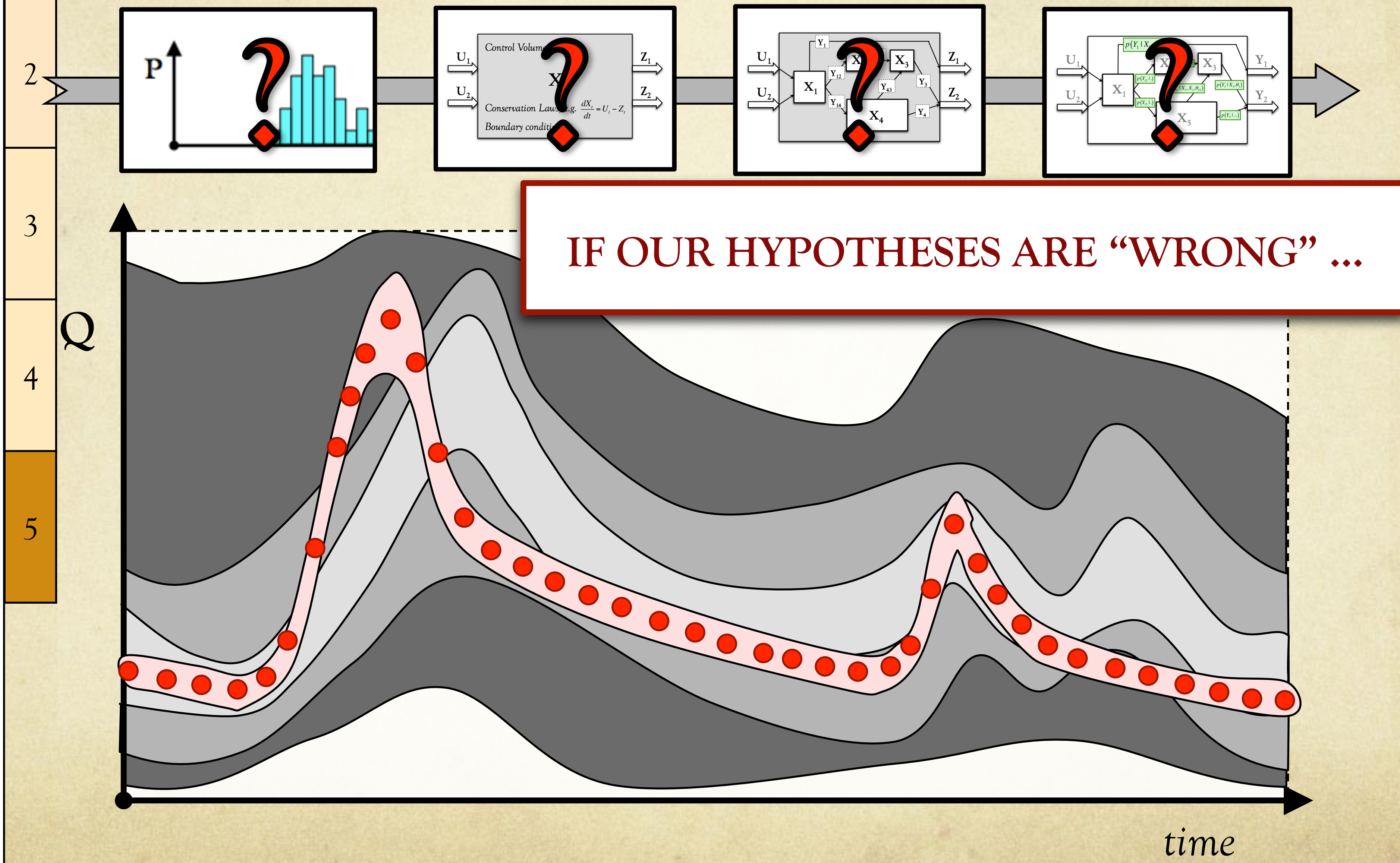
1
2
3
4
5

Conceptual Illustration





Conceptual Illustration



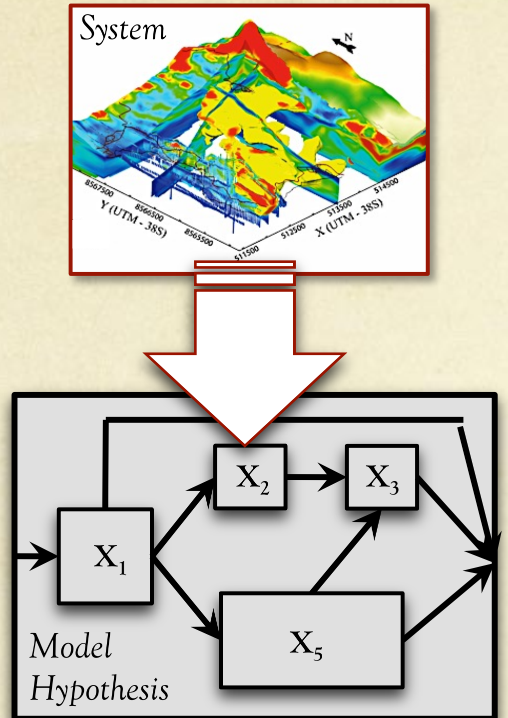
Inference …

Learning by

“Try … Assess … & Try Again”

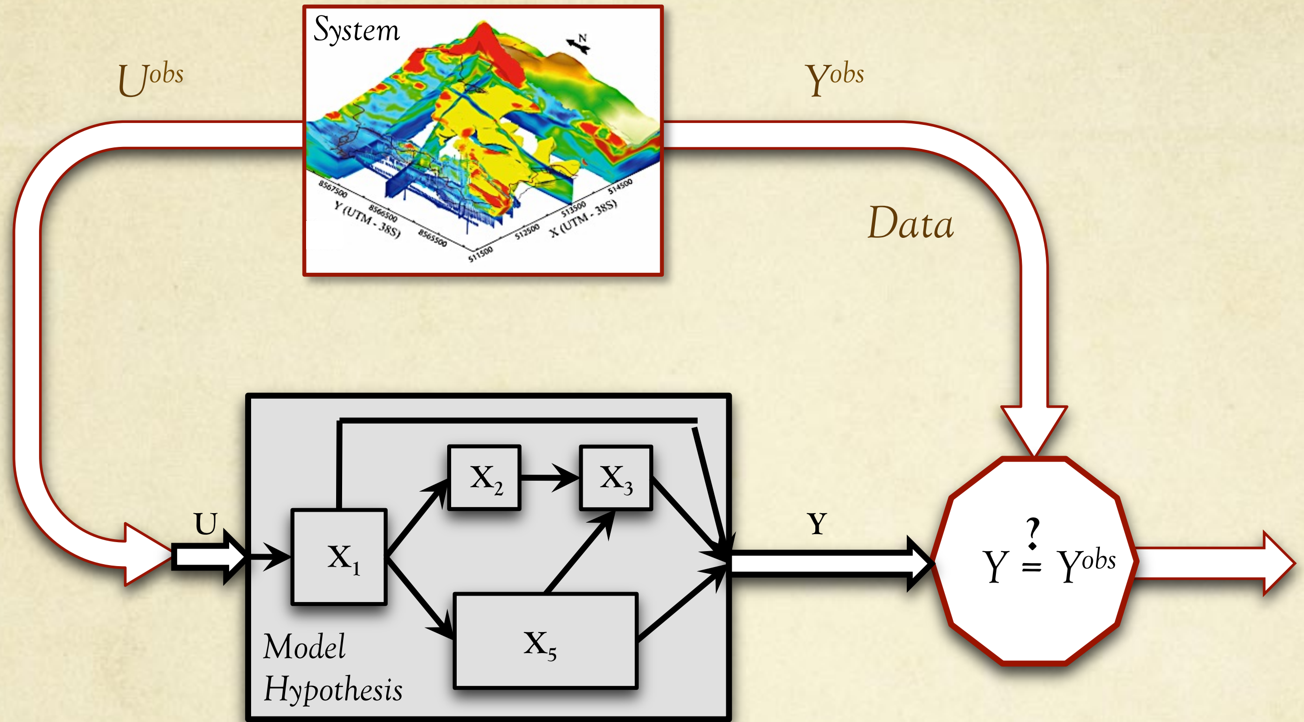
1
2
3
4
5
6

Make Hypothesis

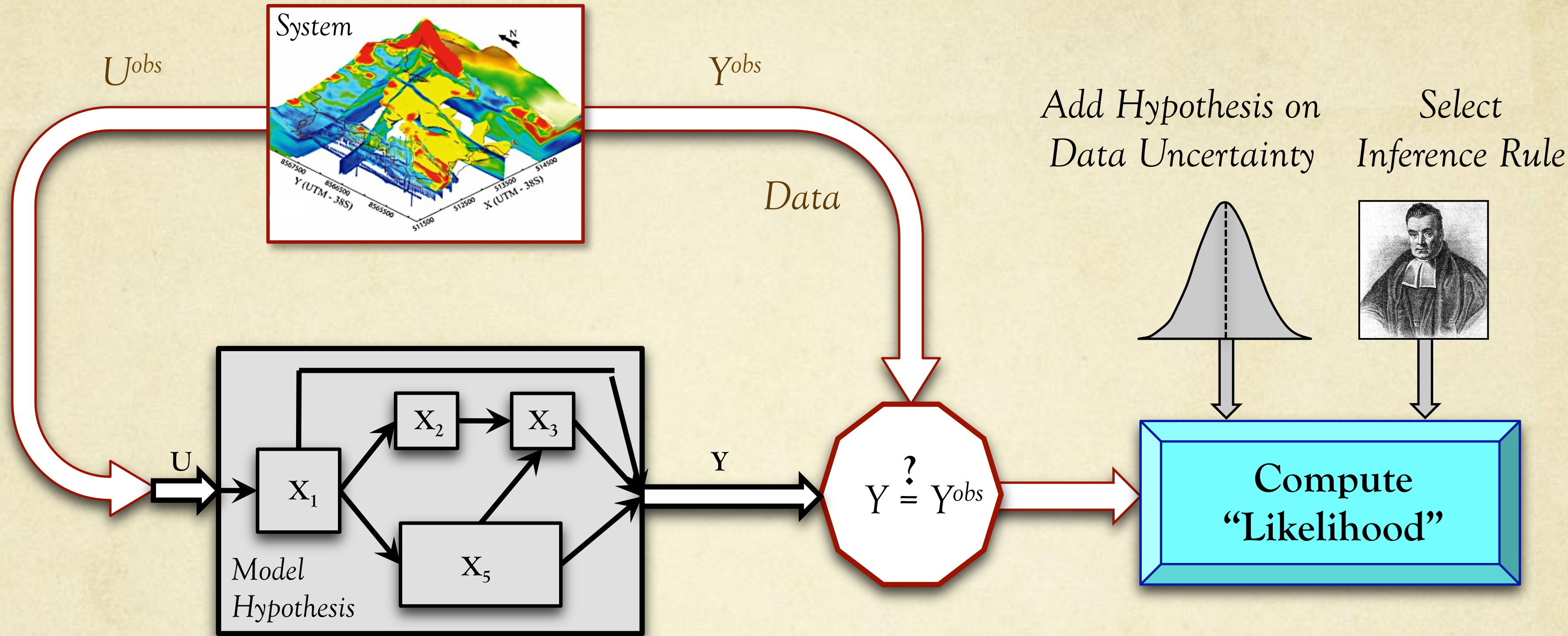


1
2
3
4
5
6

Make Hypothesis → Collect System Data

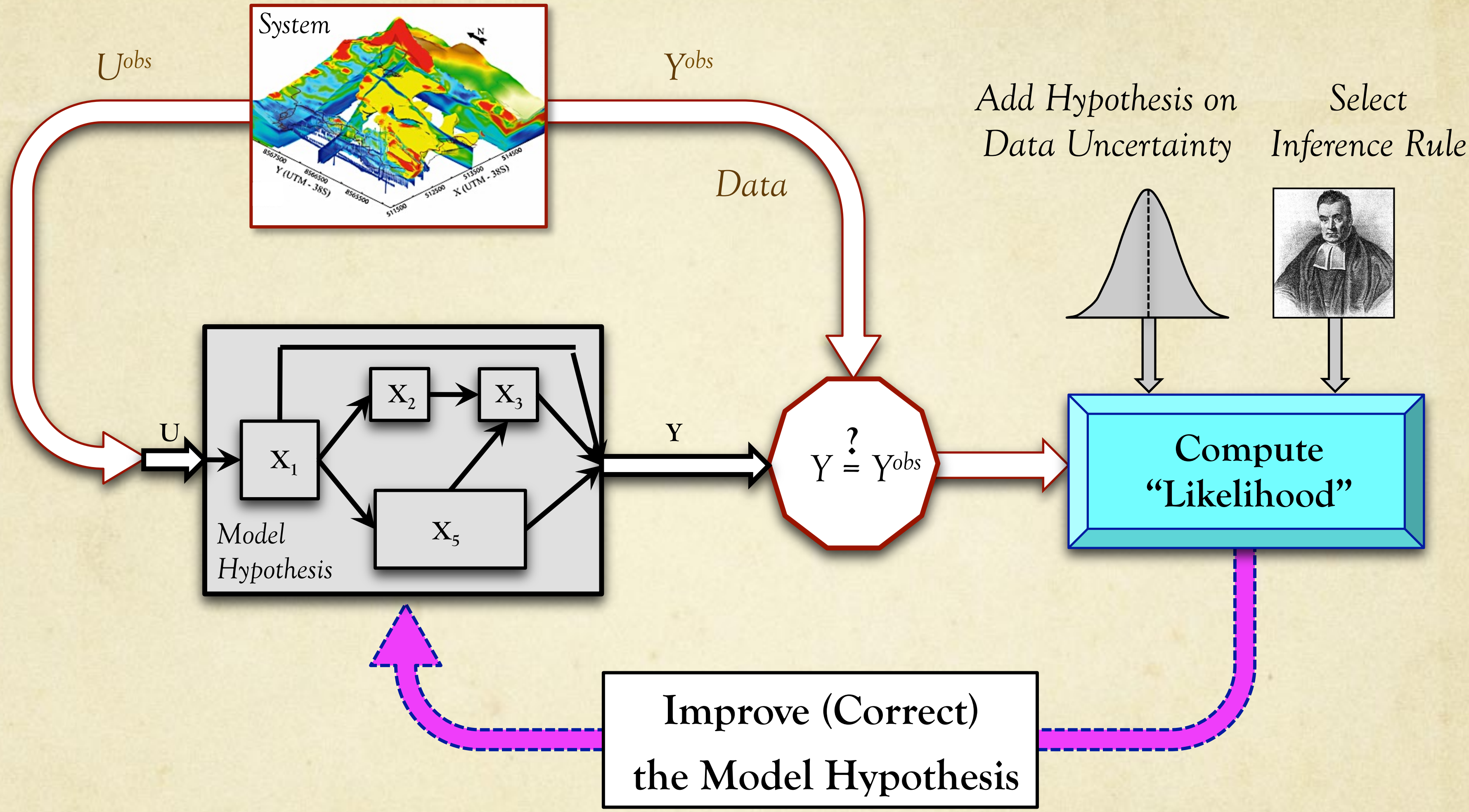


Make Hypothesis → Collect System Data → Assess Likelihood



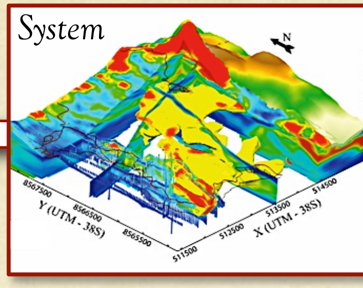
1
2
3
4
5
6

Try to Improve ("*Fix*") the Hypothesis



1. Basic → Check Assumptions about Data

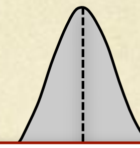
U^{obs}



Y^{obs}

Data

Add Hypothesis on Data Uncertainty



Select Inference Rule

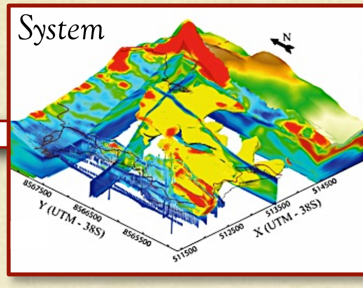
Make Sure that the "*Likelihood*" Metric Accurately Reflects the Stochastic Nature of the Information Provided by the Data

Maximize L (Model | Data)

1
2
3
4
5
6

2. Easiest → Parameter/State Estimation

U^{obs}



Y^{obs}

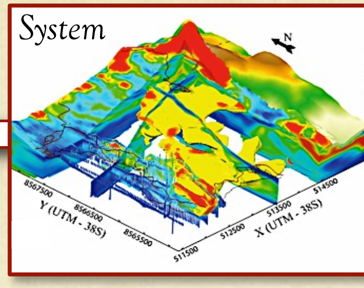
Data

Maximize $L(\chi, \boldsymbol{\theta} \mid \underbrace{H^{UN}, H^{PP}, H^{SA}, H^{CL}, \textbf{Data}}_{\text{Assumed Correct}})$

Search over all feasible values

3. Harder → Process Parameterization Estimation

U^{obs}



Y^{obs}

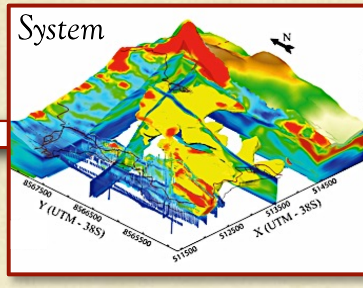
Data

$$\textbf{Maximize} \;\; L\left(H^{PP} \,\middle|\, \underbrace{H^{UN}, H^{SA}, H^{CL}, \textbf{Data}}_{\text{Assumed Correct}}\right)$$

Re-Read Literature
More Field Work
Try other Guesses

4. Harder Still → System Architecture Estimation

U^{obs}



Y^{obs}

Data

$$\text{Maximize} \;\; L\left(H^{SA}, H^{PP} \;\middle|\; \underbrace{H^{UN}, H^{CL}, \textbf{Data}}_{\text{Assumed Correct}}\right)$$

How to Distinguish the effects of *Architecture* from *Parameterization*?

"Maximum Entropy"

The "ME" Approach to DESM Learning

Grey Nearing

Uwe Ehret

Shervan Gharari

What is a Maximum-Entropy (ME) Approach ?

DESM's Code Information as a Hierarchical Sequence of Decisions

In ME → At each stage, we adopt an *Informationally Justifiable* approach

In other words …

Try to not build Overly Strong Assumptions into either the *Model* or the *Inference Procedure*

Use "Maximal Entropy" Assumptions

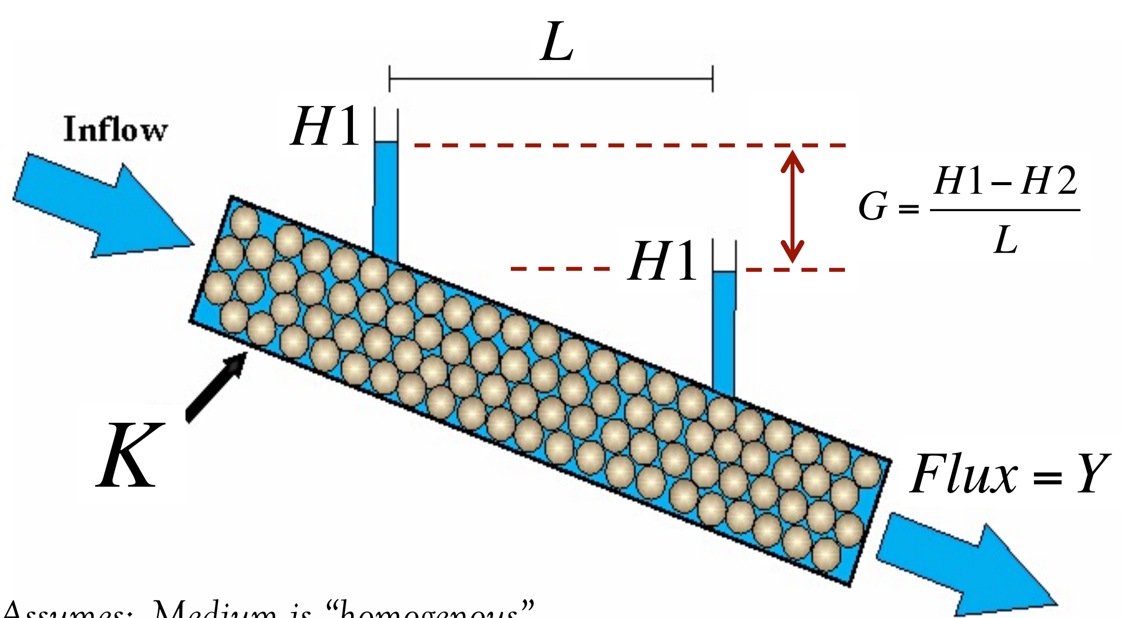
That add only as much Info to the Model as is justified by available evidence at the model relevant space-time scale

Example → ME Process Parameterization

1. Any Flux Parameterization must have the general form

$$\boxed{Y_t = K_t^{xy} \cdot X_t}$$

where X_t is the *gradient* to be dispersed and K_t^{xy} is the *conductivity* of the medium



Assumes: Medium is "homogenous"
Gradient established "instantaneously"

Basic Principle of Thermodynamics

1
2
3
4
5
6
7

Example → ME Process Parameterization

1. A Flux Parameterization must have the general form

$$\boxed{Y_t = K_t^{xy} \cdot X_t}$$

where X_t is the *gradient* to be dispersed and K_t^{xy} is the *conductivity* of the medium

2. Condition $0 \le Y_t \le X_t$ must hold to preserve *mass balance*, implying that

$$\boxed{0 \le K_t^{xy} \le 1}$$

1
2
3
4
5
6
7

Example → ME Process Parameterization

1. A Flux Parameterization must have the general form

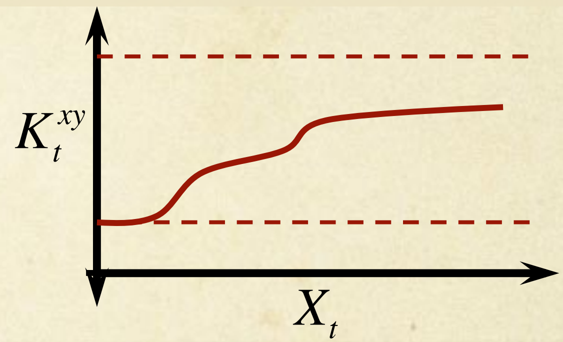
$$\boxed{Y_t = K_t^{xy} \cdot X_t}$$

where X_t is the *gradient* to be dispersed and K_t^{xy} is the *conductivity* of the medium

2. Condition $0 \le Y_t \le X_t$ must hold to preserve *mass balance*, implying that

$$\boxed{0 \le K_t^{xy} \le 1}$$

3. K_t^{xy} is a *monotonic non-decreasing or constant* function of X_t



Consistent with physical principle that

Larger gradients → Larger fluxes

Example → ME Process Parameterization

1. A Flux Parameterization must have the general form

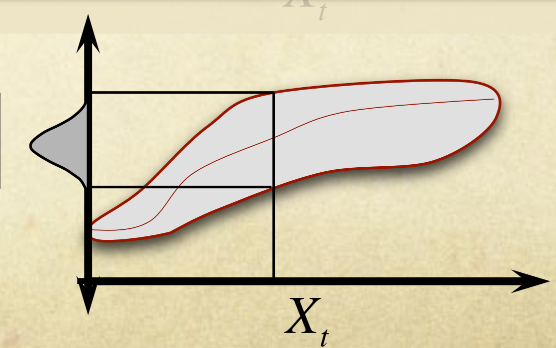
$$Y_t = K_t^{xy} \cdot X_t$$

where X_t is the <u>*gradient*</u> to be dispersed and K_t^{xy} is the <u>*conductivity*</u> of the medium

Because sub-element conditions will generally be different (in an unknown manner) each time the gradient ΔX is applied

4. K_t^{xy} is a Probabilistic function of X_t

$$K_t^{xy} \sim p\left(K_t^{xy} \mid X_t\right)$$



1 2 3 4 5 6 7

Example → ME Process Parameterization

1. A Flux Parameterization must have the general form

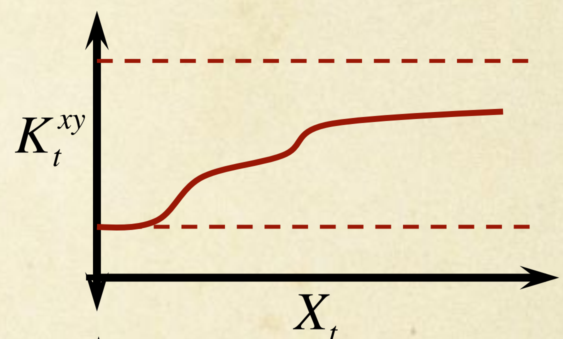
$$\boxed{Y_t = K_t^{xy} \cdot X_t}$$

where X_t is the *gradient* to be dispersed and K_t^{xy} is the *conductivity* of the medium

2. Condition $0 \leq Y_t \leq X_t$ must hold to preserve *mass balance*, implying that

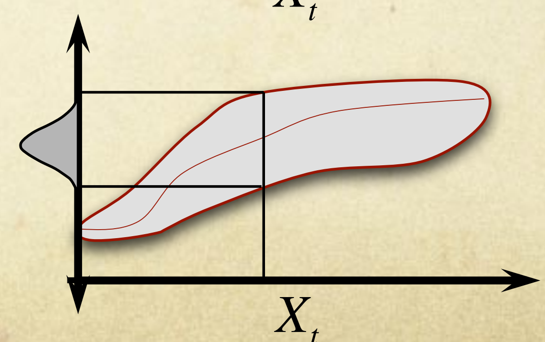
$$\boxed{0 \leq K_t^{xy} \leq 1}$$

3. K_t^{xy} is a *monotonic non-decreasing or constant* function of X_t



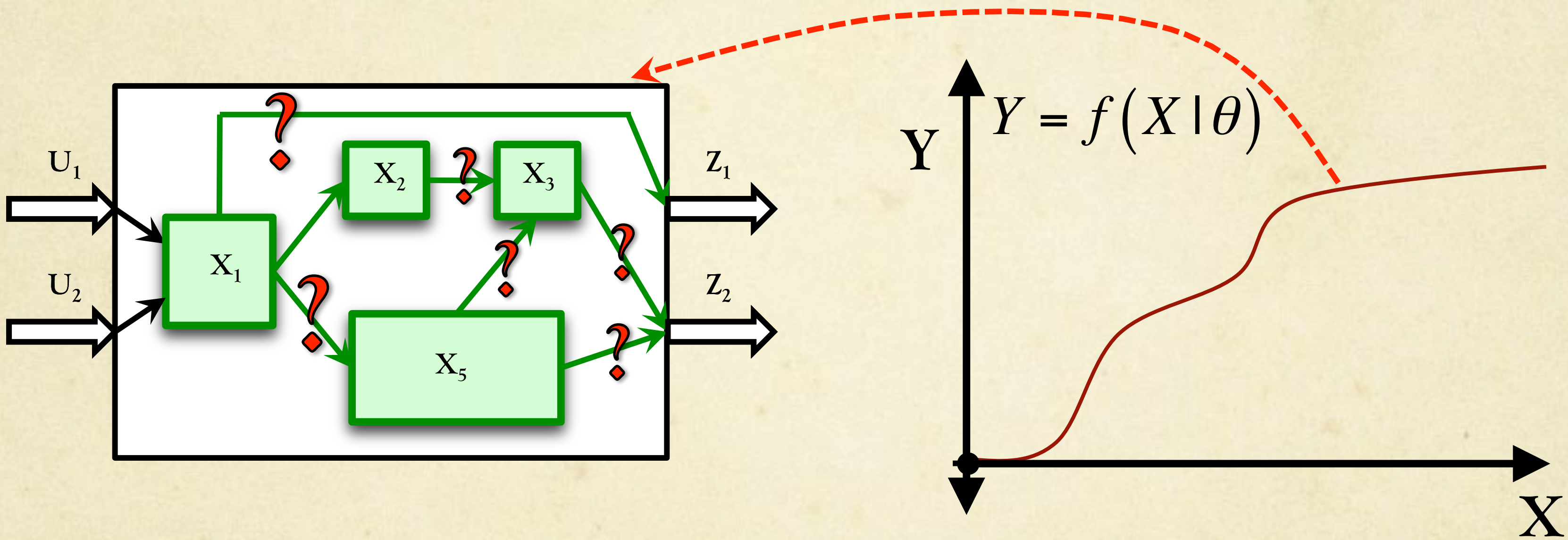
4. K_t^{xy} is a Probabilistic function of X_t

$$\boxed{K_t^{xy} \sim p\left(K_t^{xy} \mid X_t\right)}$$



In the Common Modeling Approach ...

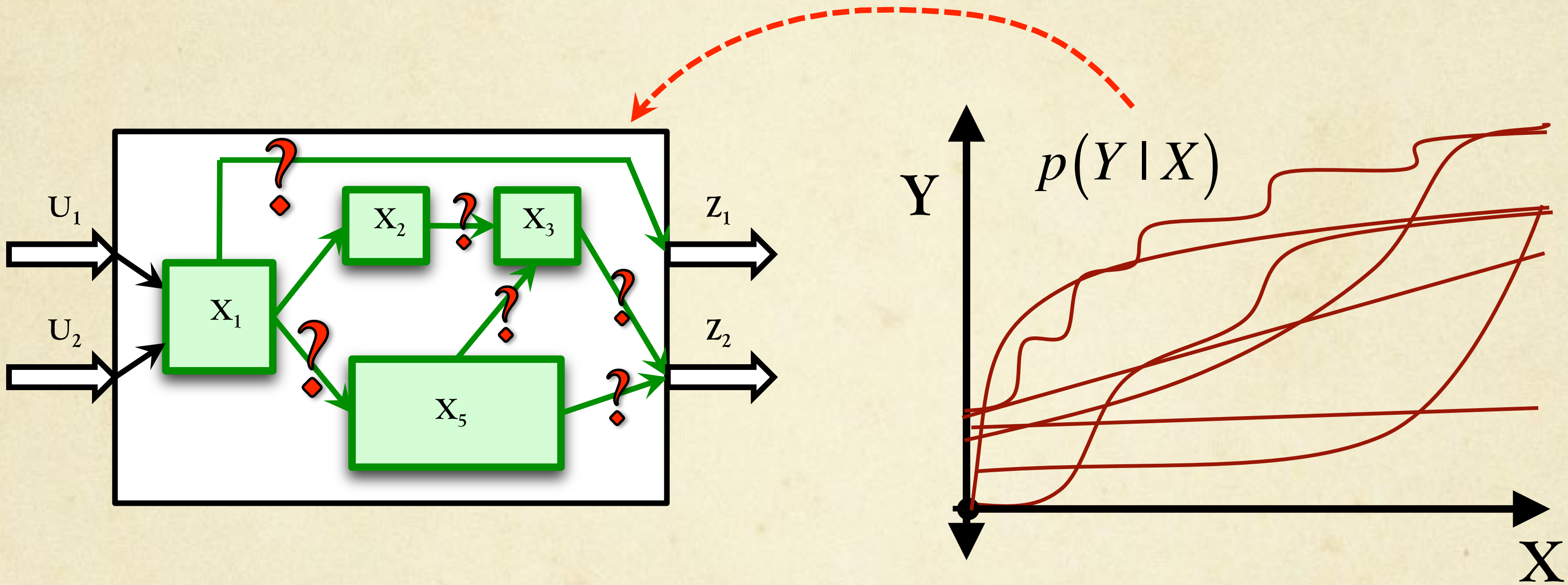
<u>Deterministic Representations</u> *of the Process Parameterization Equations impose* **<u>Overly Strong Assumptions</u>** *about what we really know regarding the actual nature of the process relationships* **<u>at the modeling scale</u>**



Invariably based on small scale field or lab studies

Maximum-Entropy Parameterization

Imposes only the minimal information required by Physics

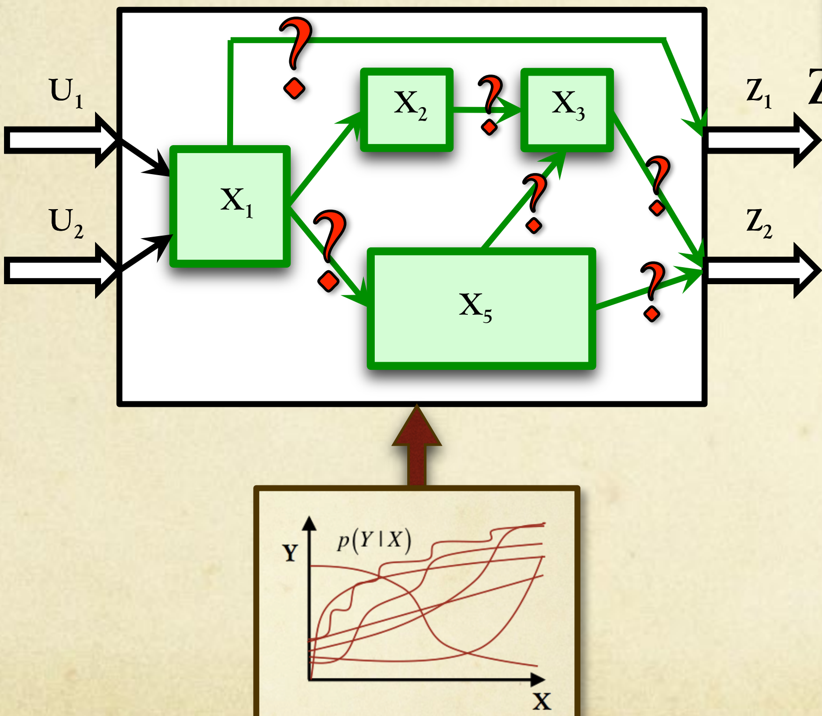


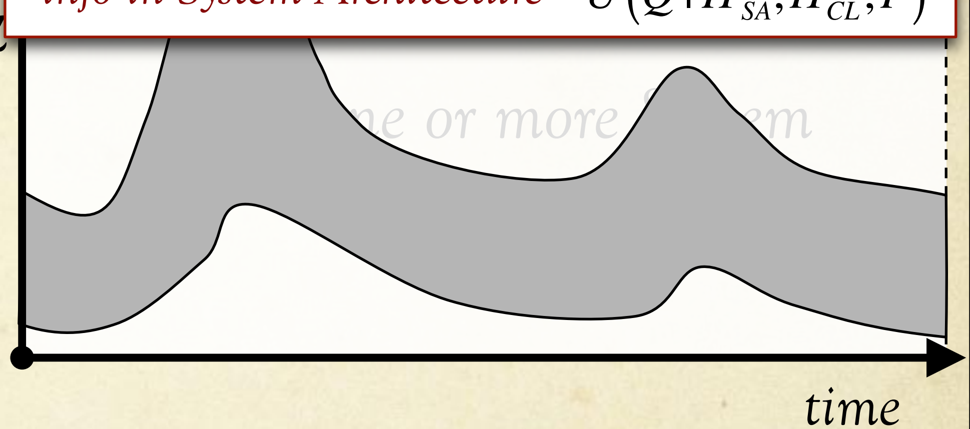
PPE's are randomly selected from a Maximum-Entropy Distribution of Random Functions that is Constrained to obey Physical Principles

This Actually Enables Better Inference of System Architecture

By minimizing the confounding effect of incorrect assumptions regarding the PPE's

Simulation Uncertainty constrained mainly by info in System Architecture $U(Q \mid H_{SA}, H_{CL}, P)$

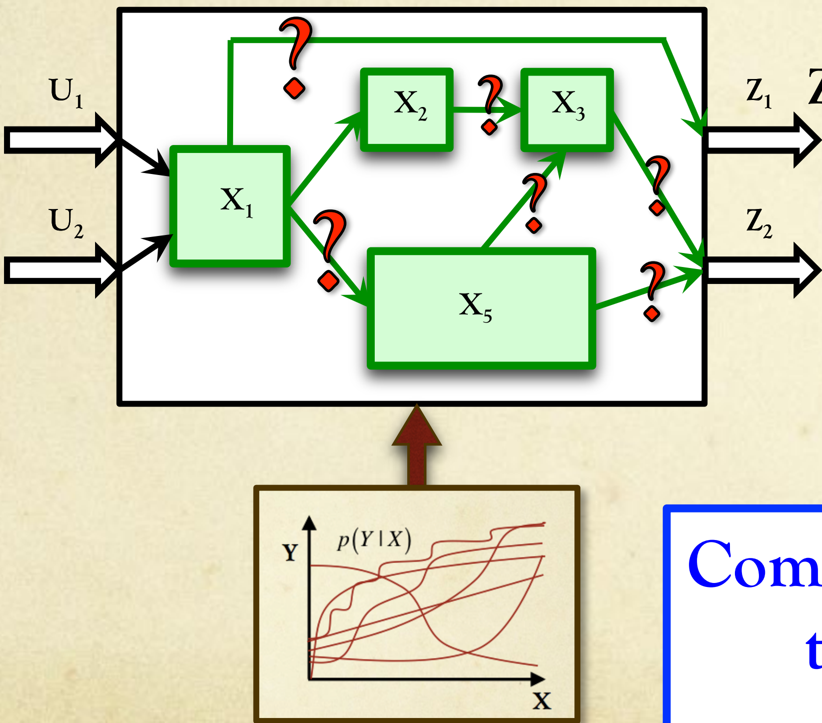


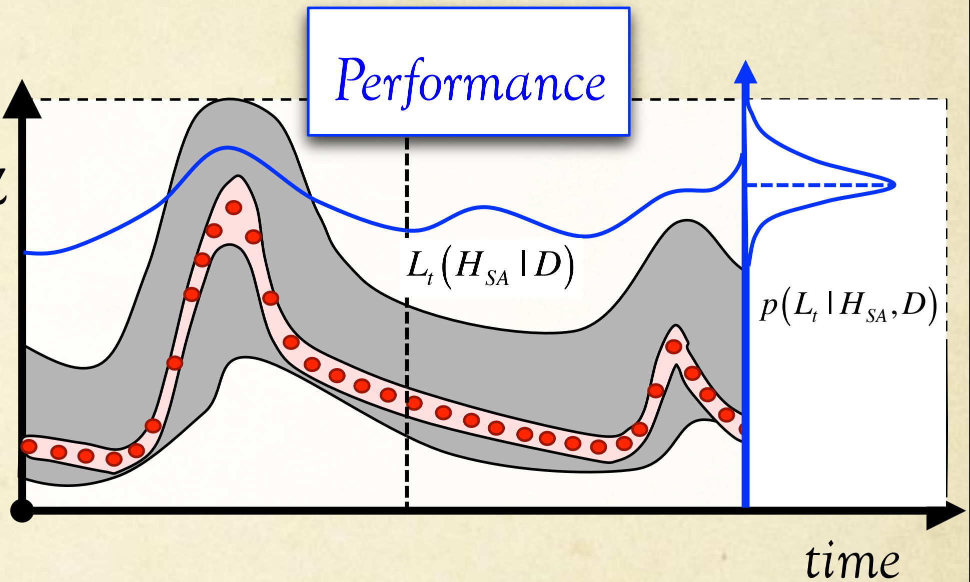


Run e.g., 10,000+ random cases (Monte Carlo on "Functions")

Inferring System Architecture

Evaluate Performance using Data





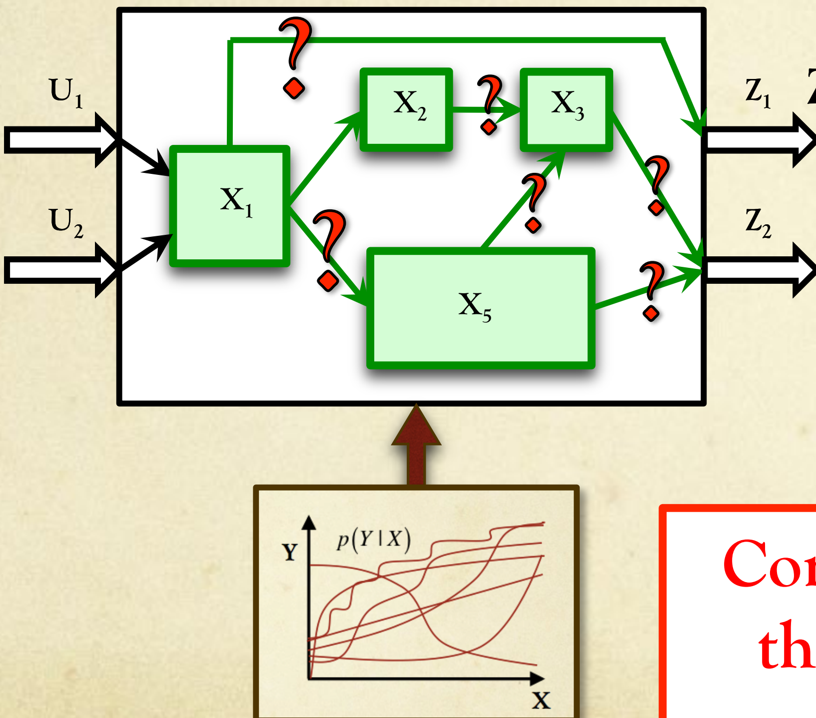
Run e.g., 10,000+ random

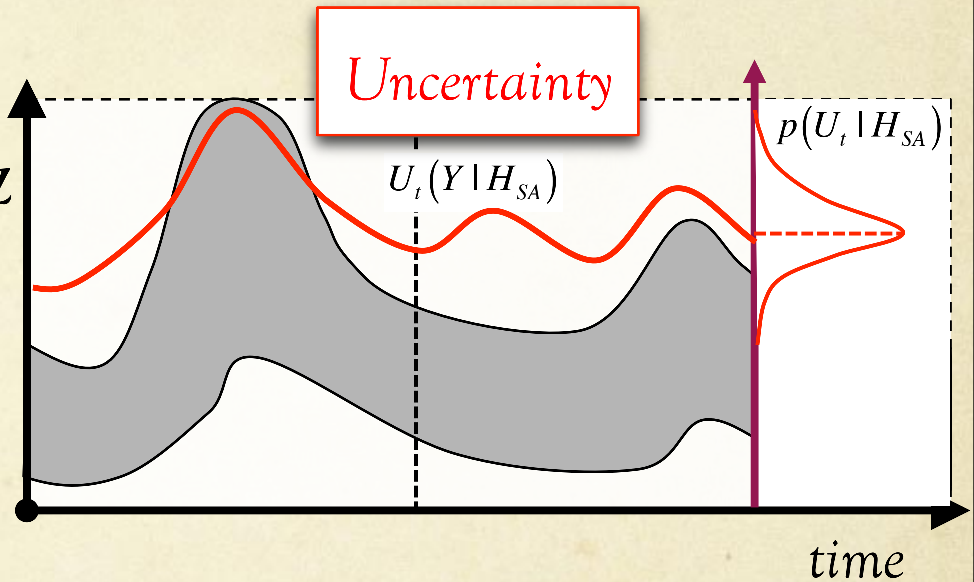
Compute the (temporal) Likelihood of the Hypothesis given the Data

$$L(H_{SA} \mid D)$$

Inferring System Architecture

Evaluate Remaining Uncertainty





Run e.g., 10,000+ random

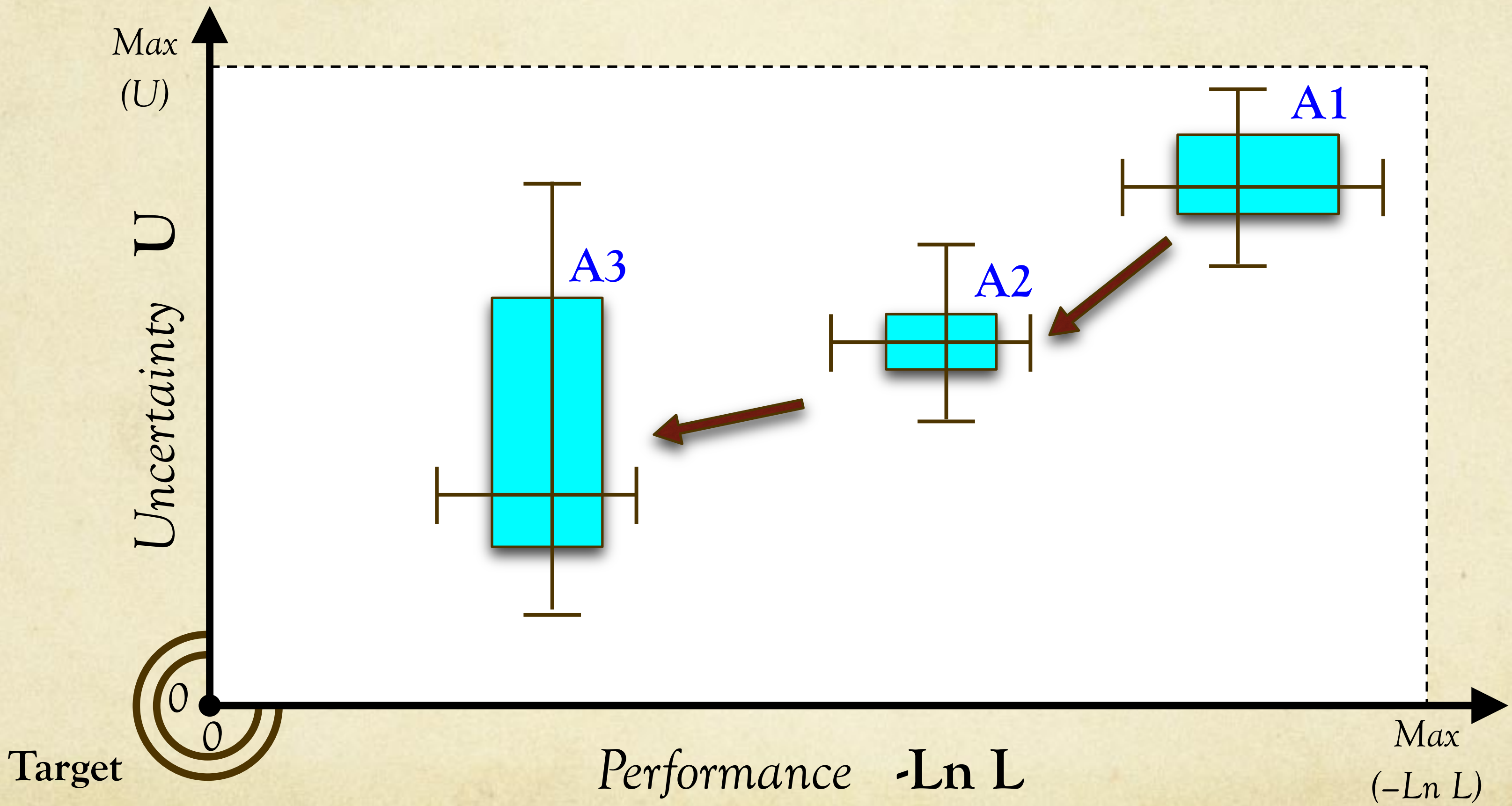
Compute the (temporal) Entropy of the Model Ensemble Simulations

$$U\left(Y \mid H_{SA}\right)$$

1
2
3
4
5
6
7

Inferring System Architecture

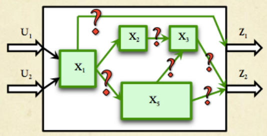
Bootstrap & Plot the Results
For different System Architectures

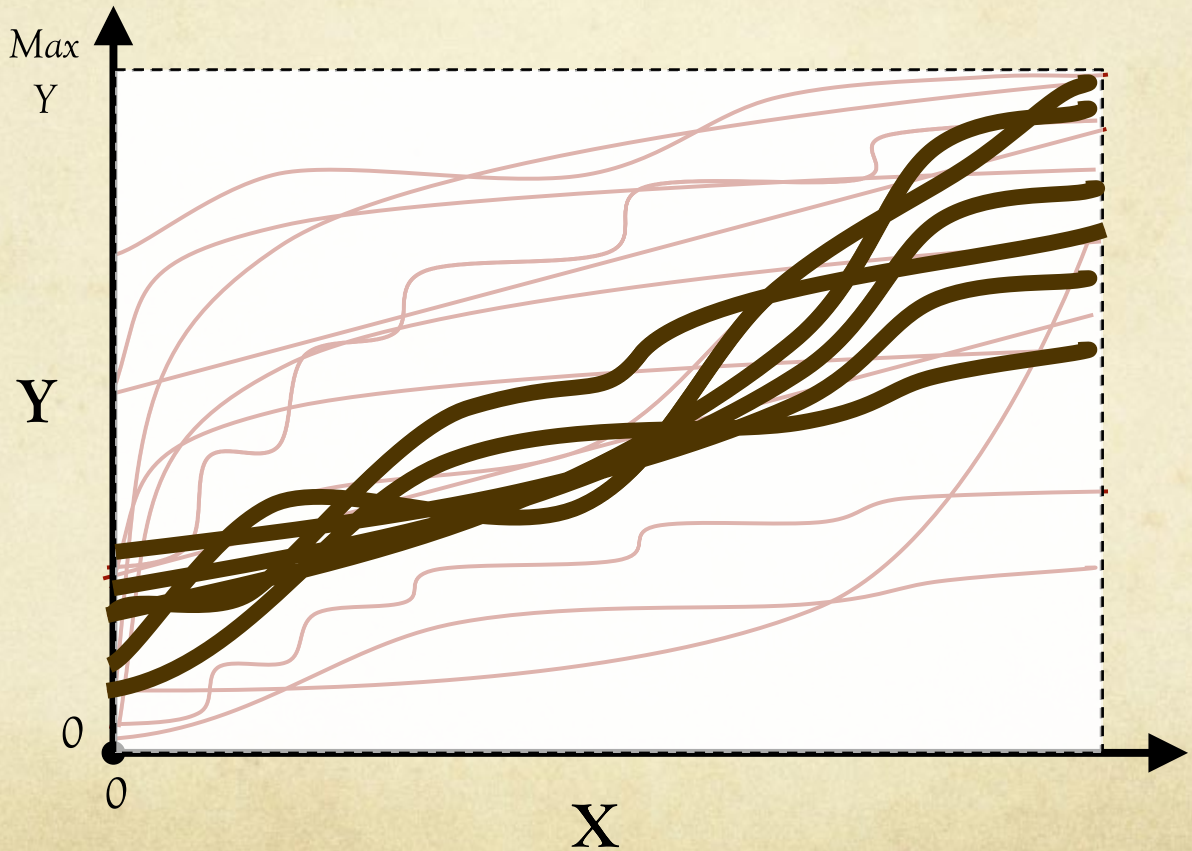


Inferring PPE Form

Use Bayesian Updating to Select Equation Forms with Highest Performance

(conditional on a selected Architecture)

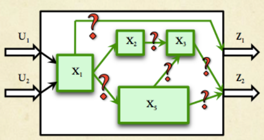


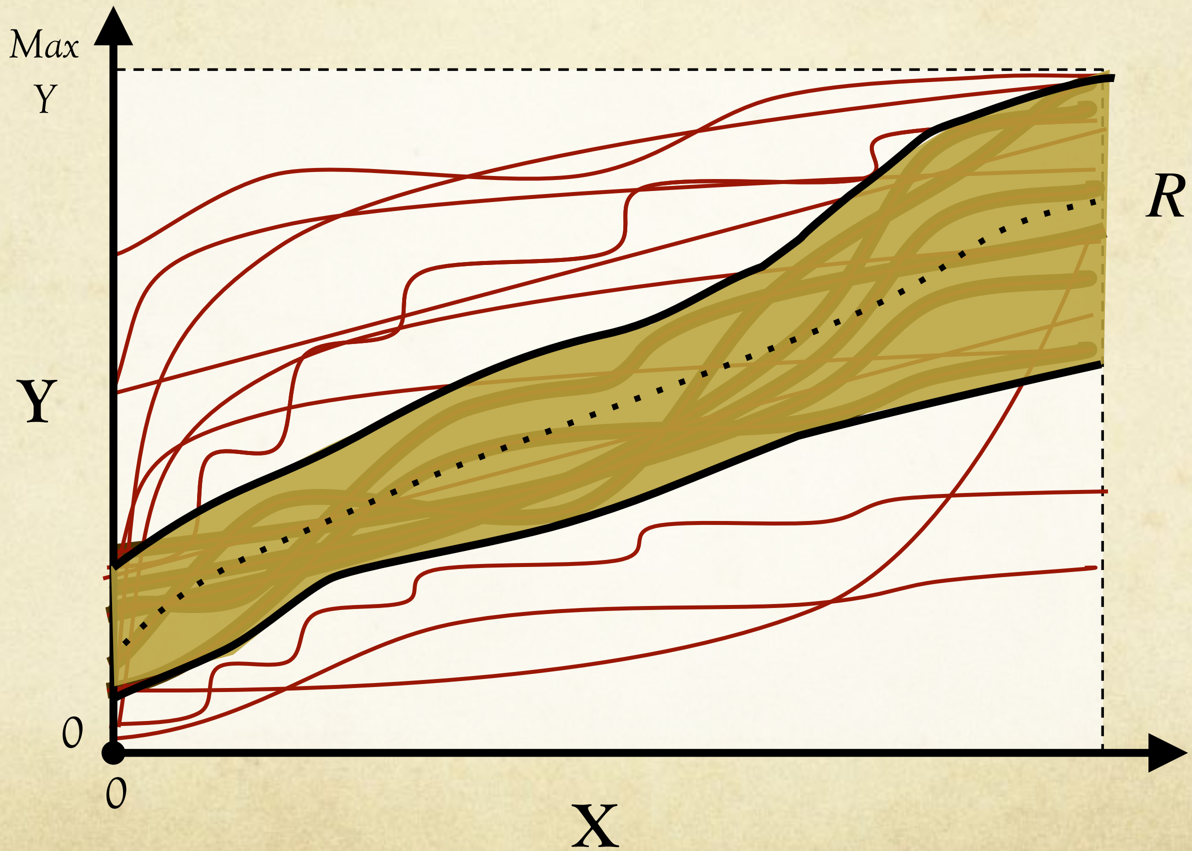


Inferring PPE Form

Propose Mathematical Forms for Equations

(conditional on a selected Architecture)





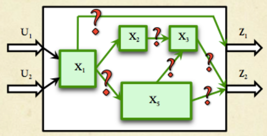
$$R : p\left(Y \mid X, \theta\right)$$

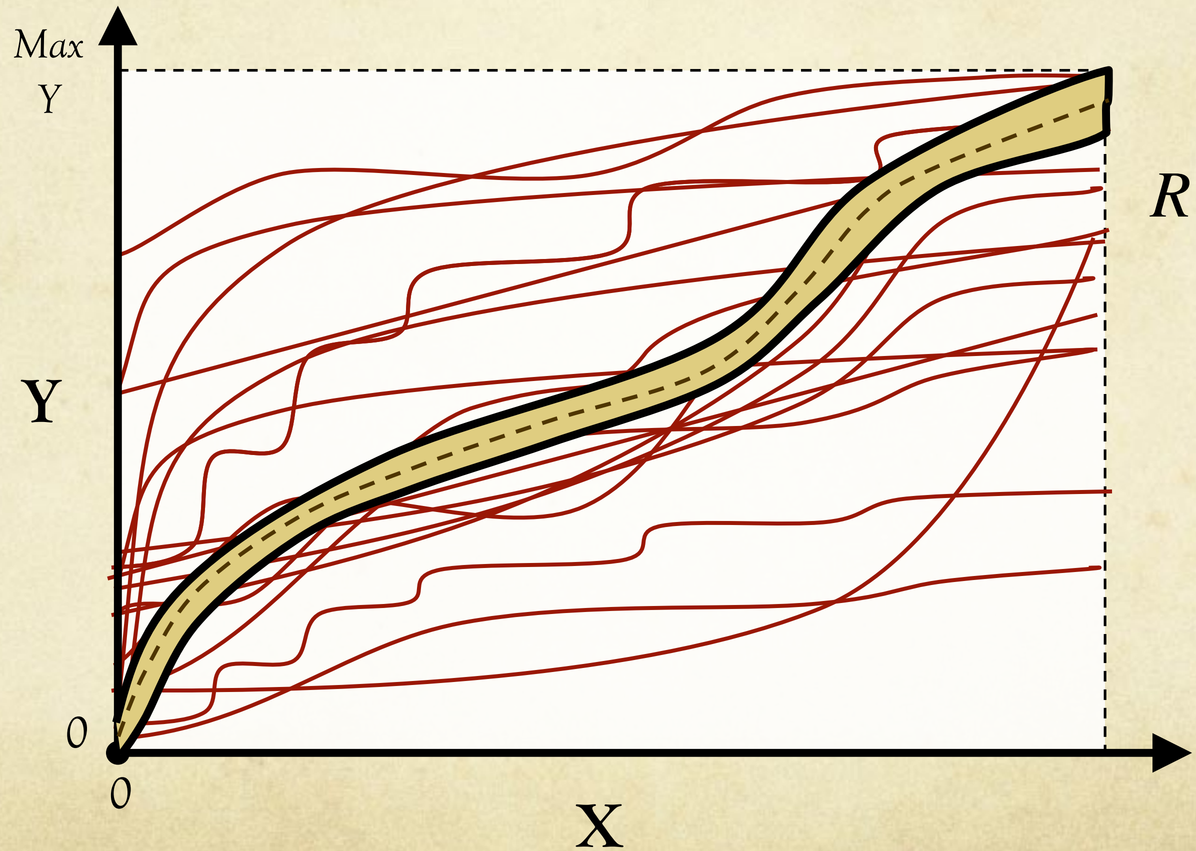
Mapping

Inferring PPE Form

Proceed with Parameter Estimation

(conditional on a selected Architecture & Parameterization Form)





$$R : p\left(Y \mid X, \theta\right)$$

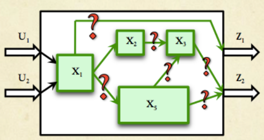
Mapping

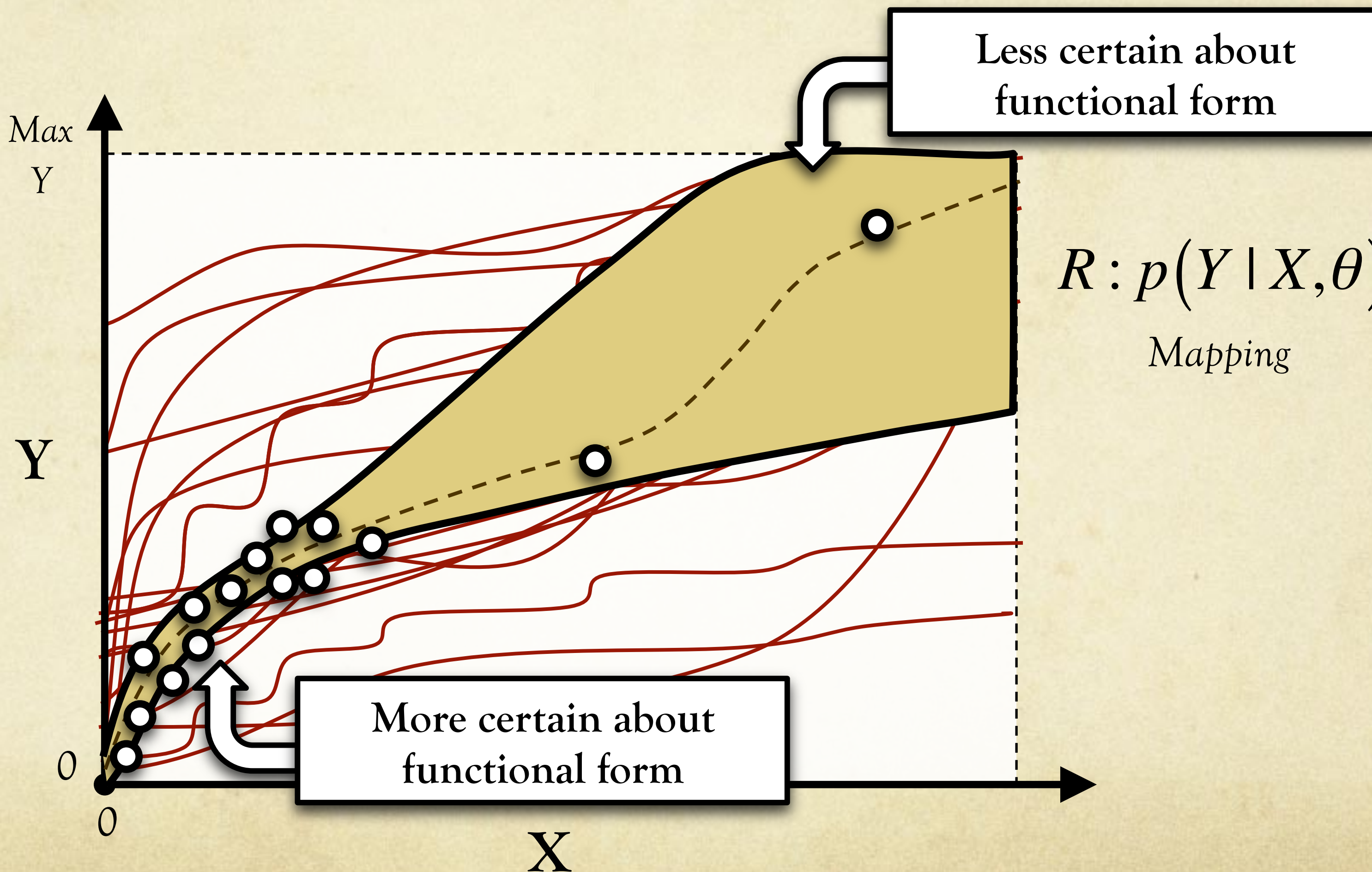
1
2
3
4
5
6
7

Inferring PPE Form

Evaluate Remaining Uncertainty

(conditional on a selected Architecture & Parameterization Form)





$$R : p(Y \mid X, \theta)$$

Mapping

1
2
3
4
5
6
7

The Result

Strategy to investigate
Model Structural Hypotheses

Without the need to make Strong Assumptions
Regarding Process Parameterizations (Equations)

In Principle a similar approach could be
used to investigate value of different
Conservation Laws

1
2
3
4
5
6
7

More Generally

Bring more "Honesty/Rigor" into the Model Building Process

1) *Build Into the Model Clarity Regarding What We Feel Certain/Uncertain About*

2) *Be Clear about "What is Known" versus "What is Hypothesis / Assumption"*

"Maximum Entropy Approach" To Model Building

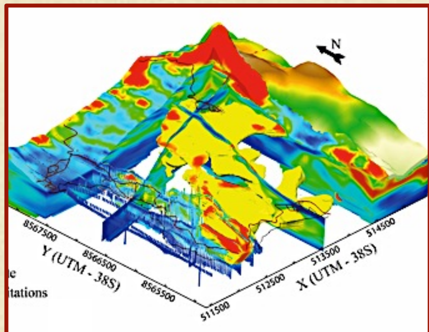
1 2 3 4 5 6 7

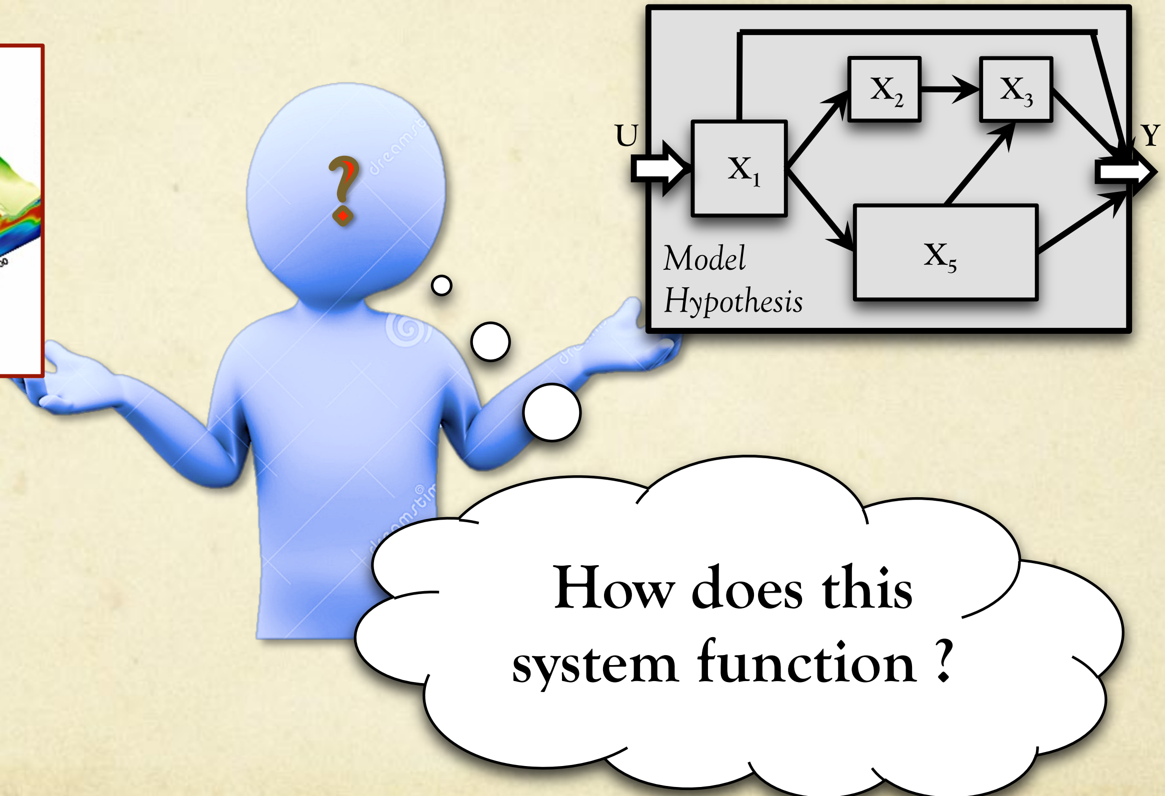
<u>Some Comments in Conclusion</u>

1. **Models & Data codify** ***<u>Information</u>*** **about the world**
2. **Information implies** ***<u>Change in Uncertainty</u>*** **about Something**
3. **Models are** ***<u>Hierarchical Assemblages of Hypotheses</u>***
   1. *Conservation Laws*
   2. *System Architecture*
   3. *Process Parameterization*
   4. *Uncertainty*
4. **Model Hypotheses can be:**
   1. *<u>Over-Constrained</u>* **by un-justifiably strong hypotheses**
   2. *<u>Under-Constrained</u>* **by lack of knowledge about**
      - a) **Scale-dependence of process relationships**
      - b) **Sub-element heterogeneity**
5. **Model Structural Inference can be done using** ***<u>Max-Entropy</u>*** **PP's**
6. **Process Equation Inference can be done using Bayes' Law**

And Finally …

An *Information Theory* perspective can help improve the way we use models as Hypotheses for *Scientific Investigation*





Thank You