The Role of Artificial Neural Networks in Understanding Complex Systems Behavior

Gary R. Weckman

Ohio University

Palm Island Enviro-Informatics & Business Solutions, LLC

David F. Millie

Palm Island Enviro-Informatics & Business Solutions, LLC

Loyola University, New Orleans

Andrew P. Snow

Ohio University

School of Information and Telecommunication Systems

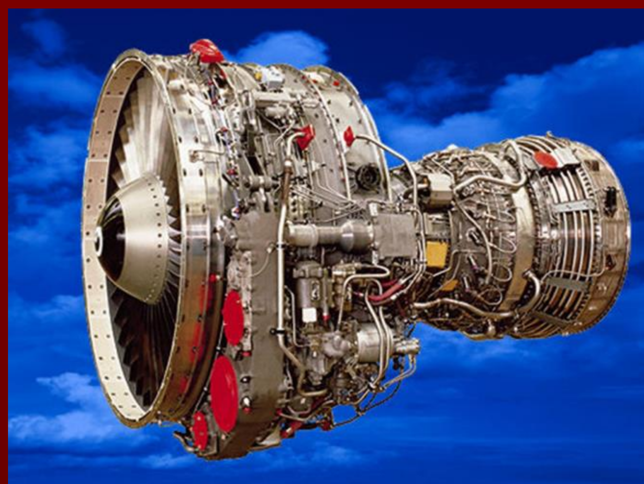
What are complex systems?

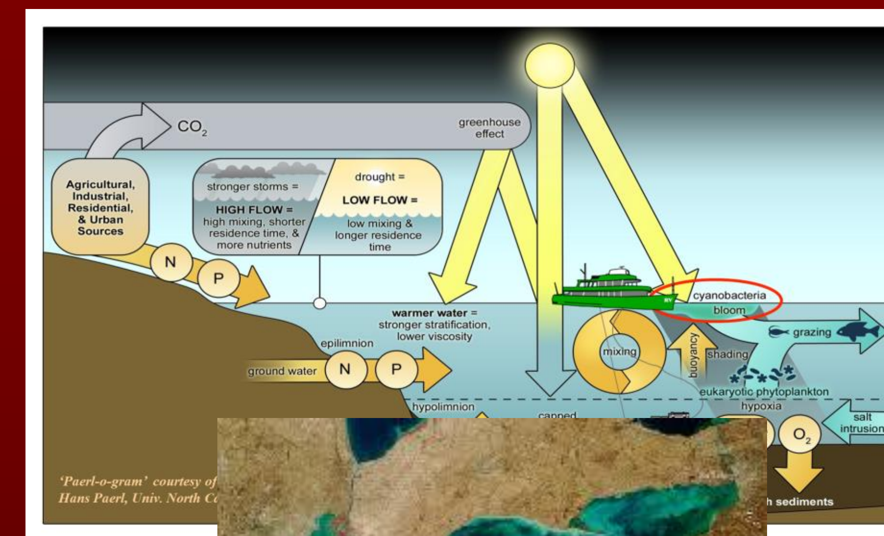
- **"A system comprised of a (usually large) number of (usually strongly) interacting entities, processes, or agents, the understanding of which requires the development, or the use of, new scientific tools, nonlinear models, out-of equilibrium descriptions and computer simulations." [Advances in Complex Systems Journal]**

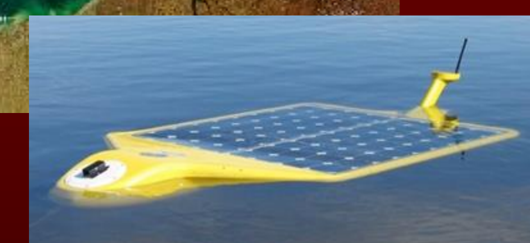
- **"A system that can be analyzed into many components having relatively many relations among them, so that the behavior of each component depends on the behavior of others. [Herbert Simon]"**

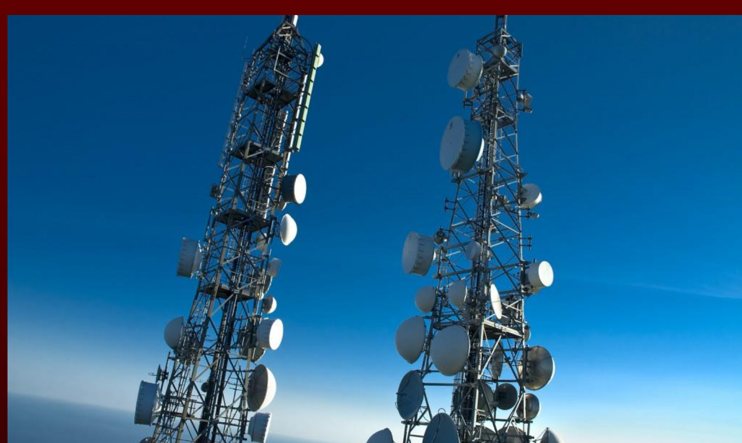
- **"A system that involves numerous interacting agents whose aggregate behaviors are to be understood. Such aggregate activity is nonlinear, hence it cannot simply be derived from summation of individual components behavior." [Jerome Singer]**

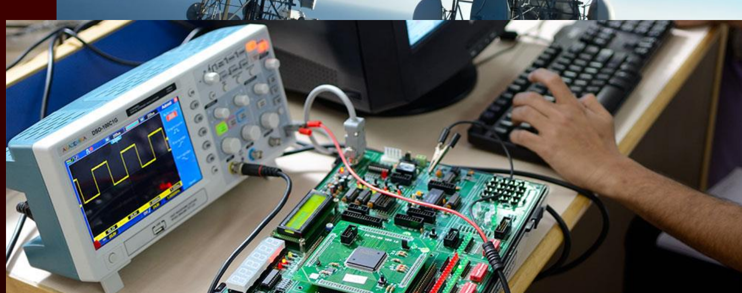
Our Research: Complex Systems











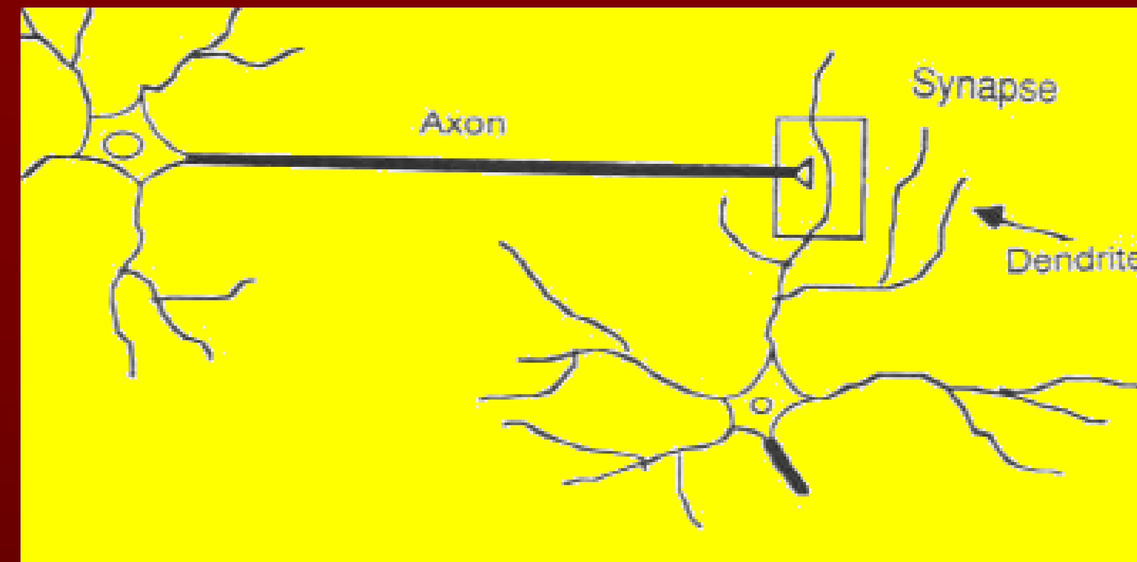
Artificial Neural Networks

Machine-learning algorithms that identify data patterns and perform decision making in a manner imitating cognitive functionality

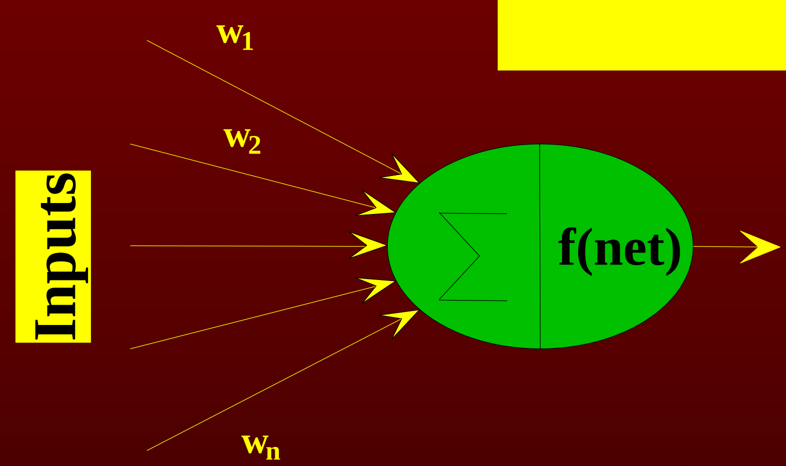
- **'*Learning*' (analogous to problem solving) is:**
  - ✓ **adaptive - knowledge is altered, updated, & stored (via weights)**
  - ✓ **iterative - examples to generalizations**
- **'*Universal approximators*' – can discover & reproduce any *(linear / non-linear)* trend given enough data & computational (processing) capability**
  - ✓ **No expert knowledge required**
  - ✓ **Few (if any)'formal' assumptions - i.e. Gaussian requirements, etc.**
- **Disadvantage - *(superficially ? ?)* lack a declarative knowledge structure**
  - ✓ **a *'Black Box'* (i.e. no global equation)**

Biological Analogy

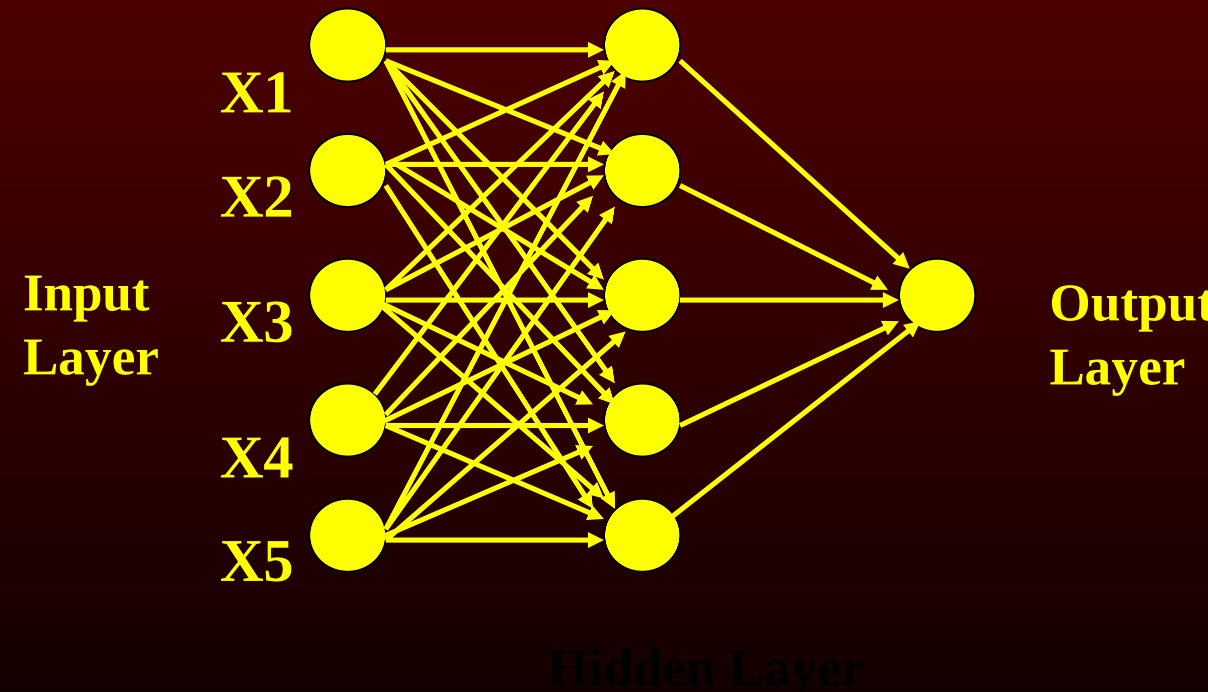
- Brain Neuron



- Artificial neuron

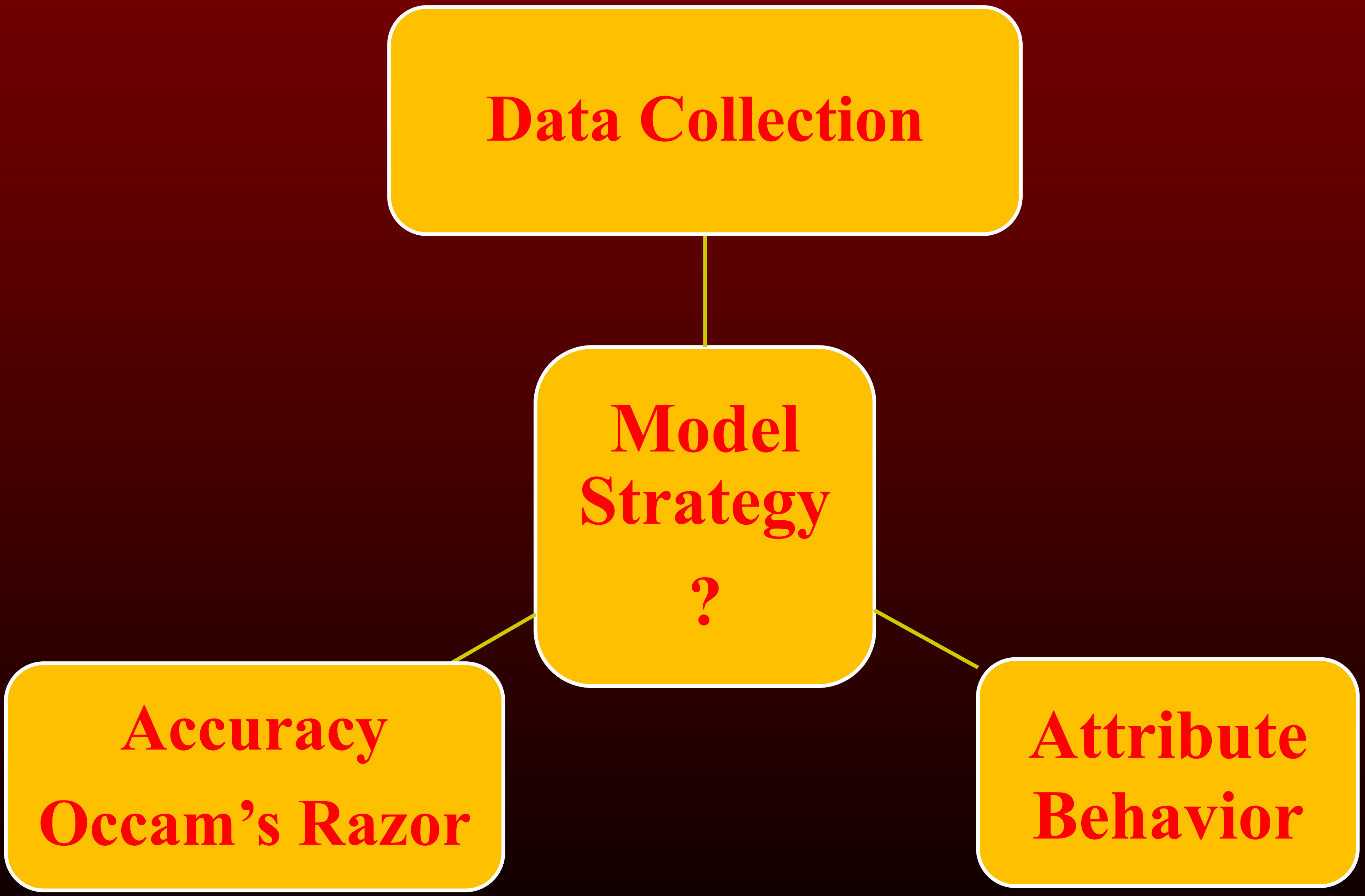


- Set of processing elements (PEs) and connections (weights) with adjustable strengths

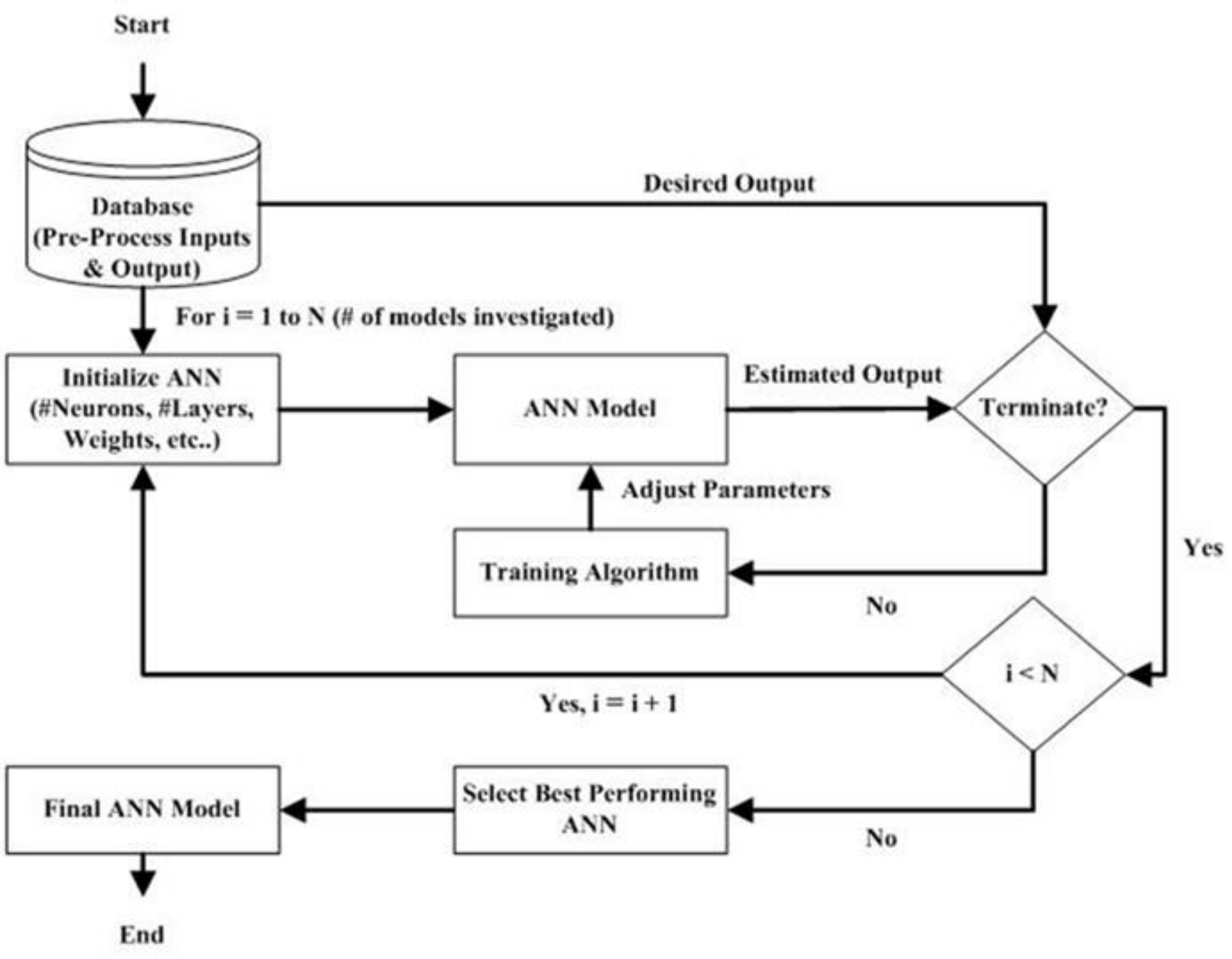


Modeling Approach

Early Days: Interested in "Model Accuracy"

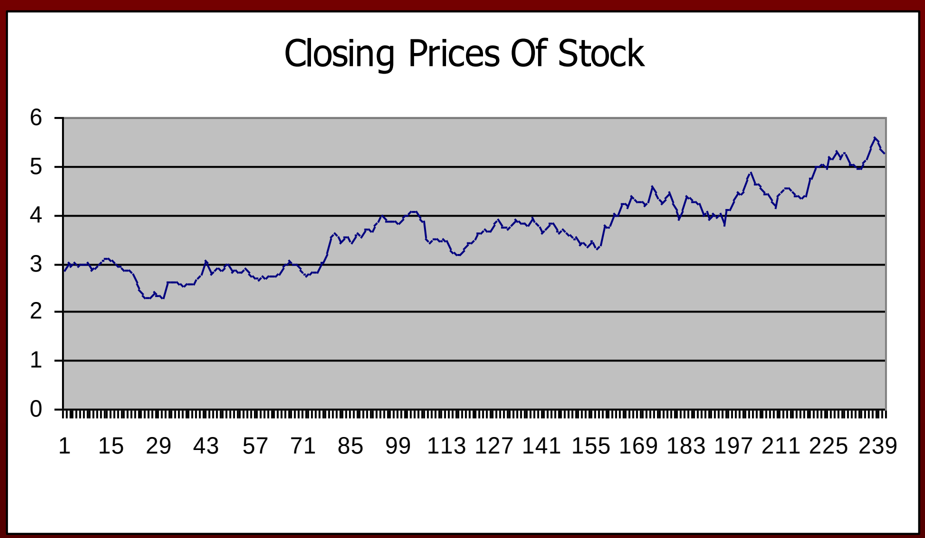


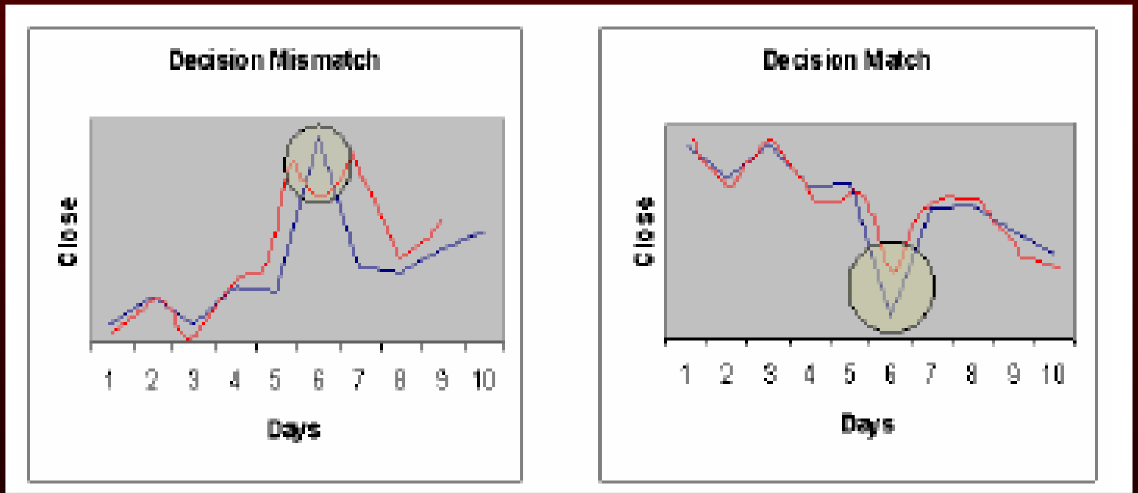
Modeling Approach

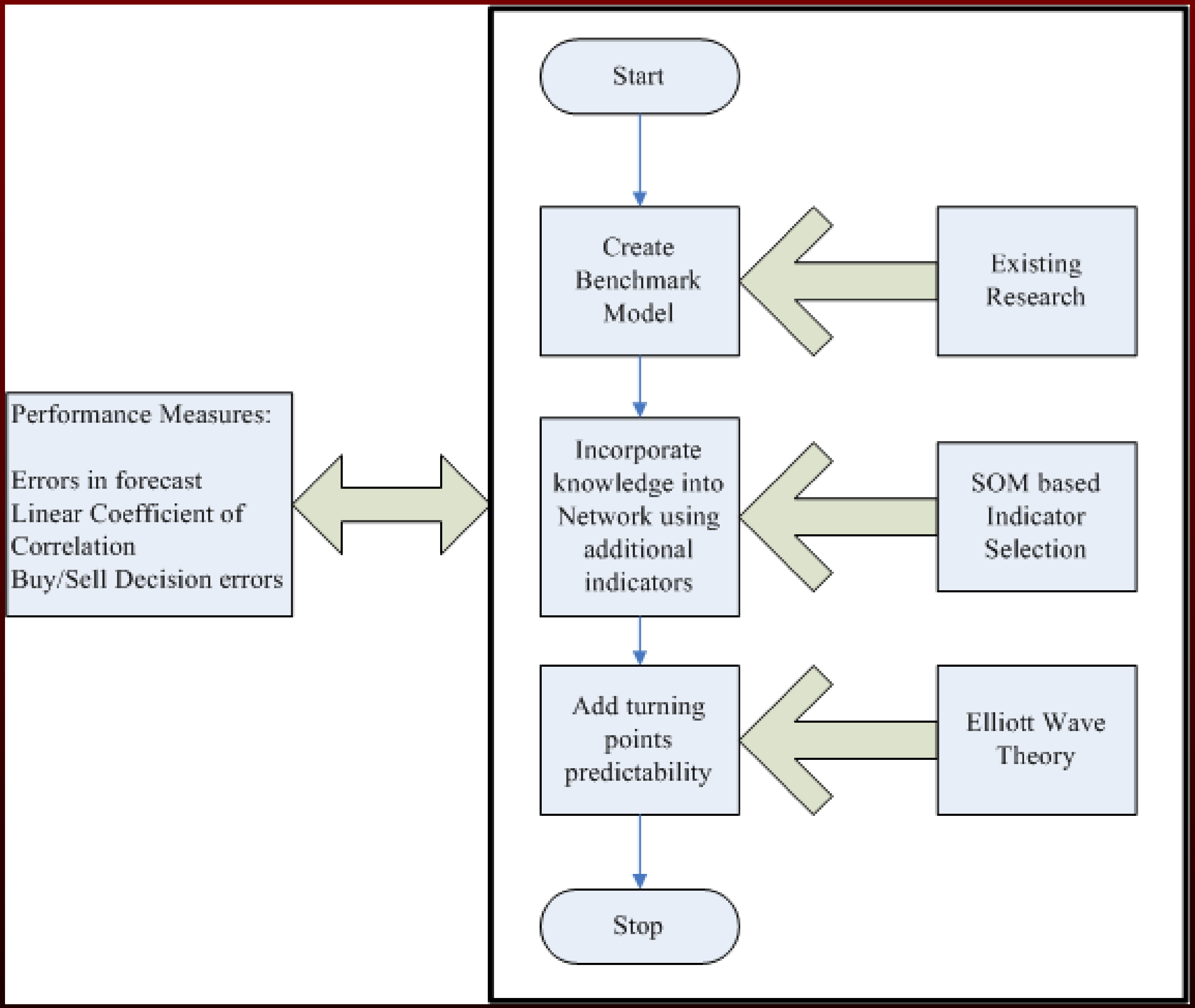


Early Project: Stock Market Model

Accuracy of predicting market turns – not necessarily why





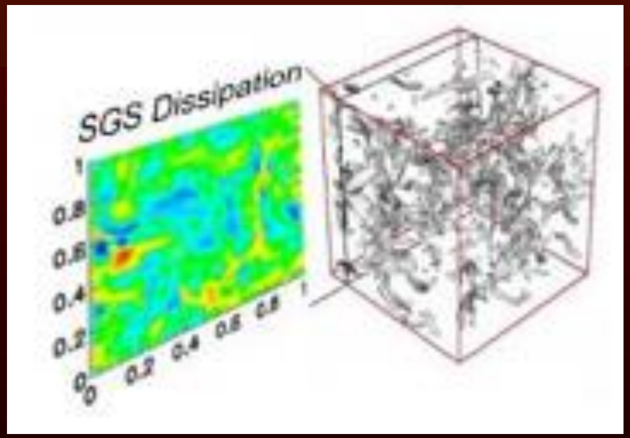


*‘Paradigms’ of Scientific Discovery **

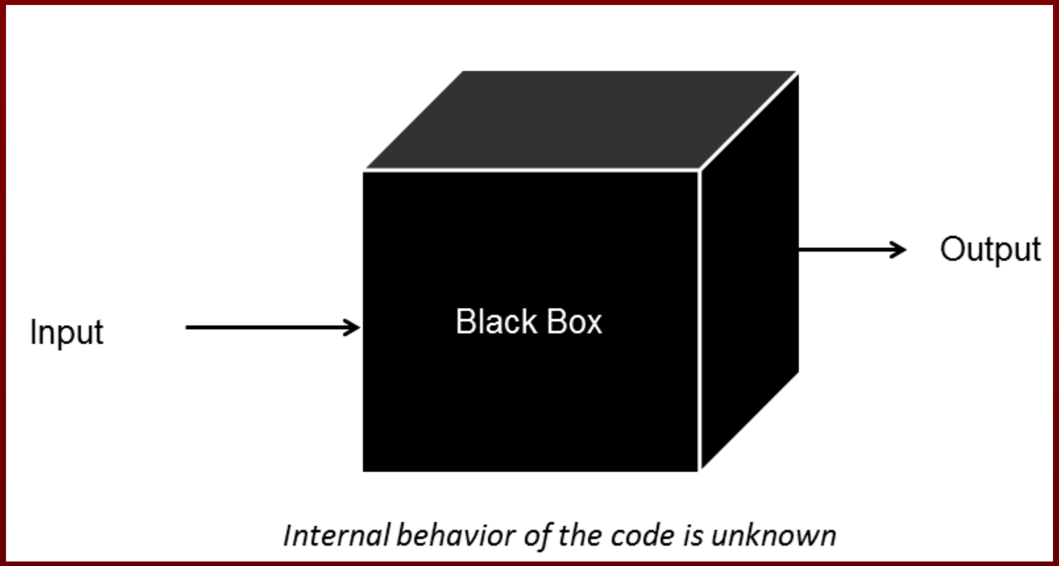
- ***Empirical** - describing natural phenomena*
  - **initiated, a thousand years ago**
- ***Theoretical** - models, ‘laws’ & generalizations*
  - **initiated, the last few hundred years**

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$

- ***Computational** - simulating complex phenomena*
  - **initiated, the last few decades**



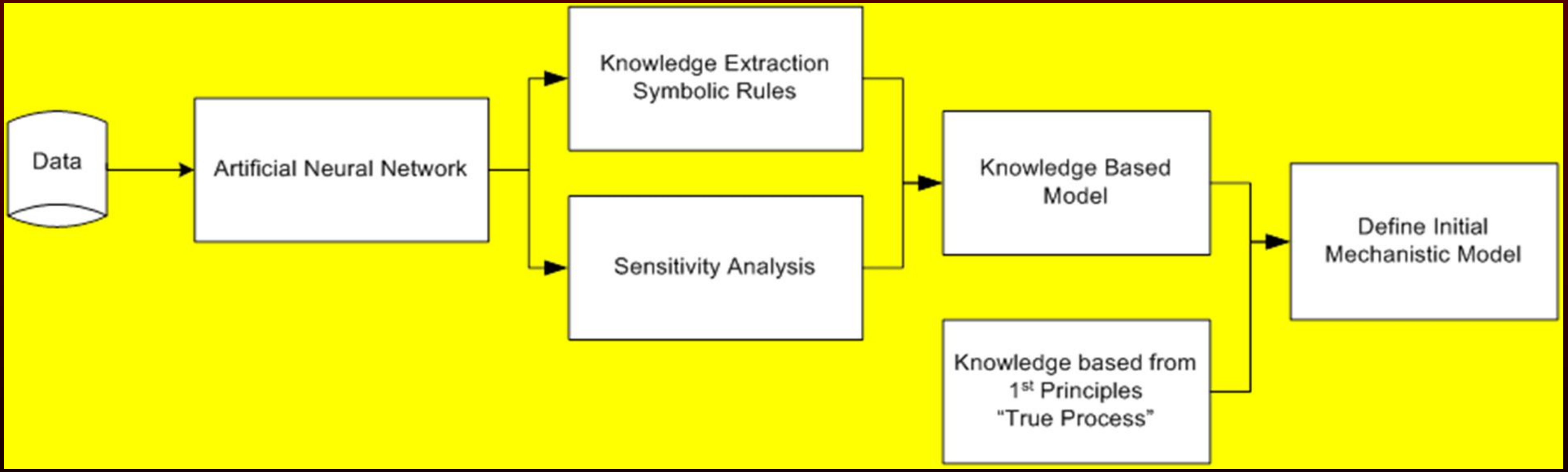
ANN: BLACK BOX



KNOWLEDGE EXTRACTION defined:

is the creation of knowledge from structured (relational databases, XML) and unstructured (text, documents, images) sources [https://en.wikipedia.org/wiki/]

Is there a way illuminate the black box?



Environmental Modeling & Knowledge Extraction

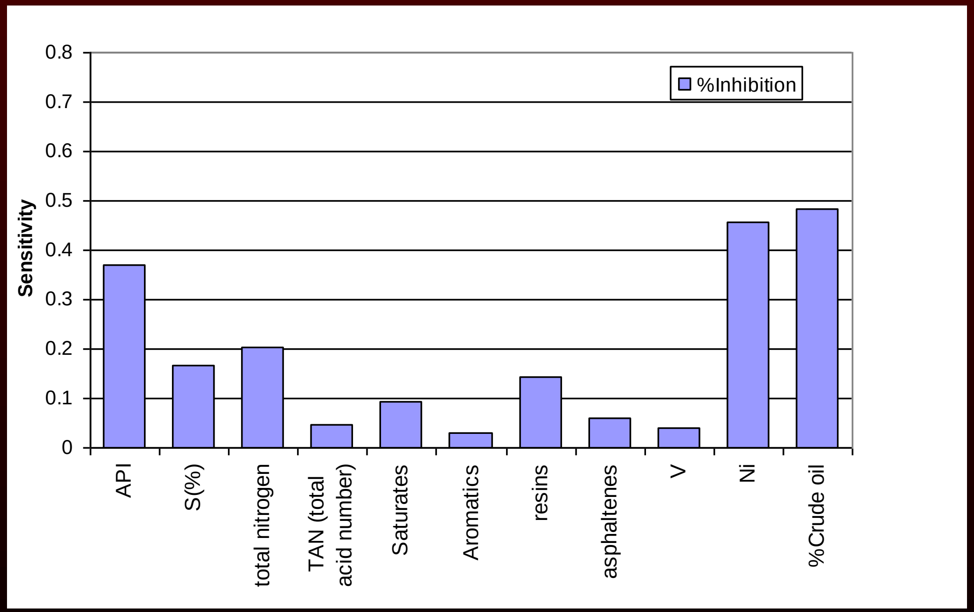
1st ATTEMPT:

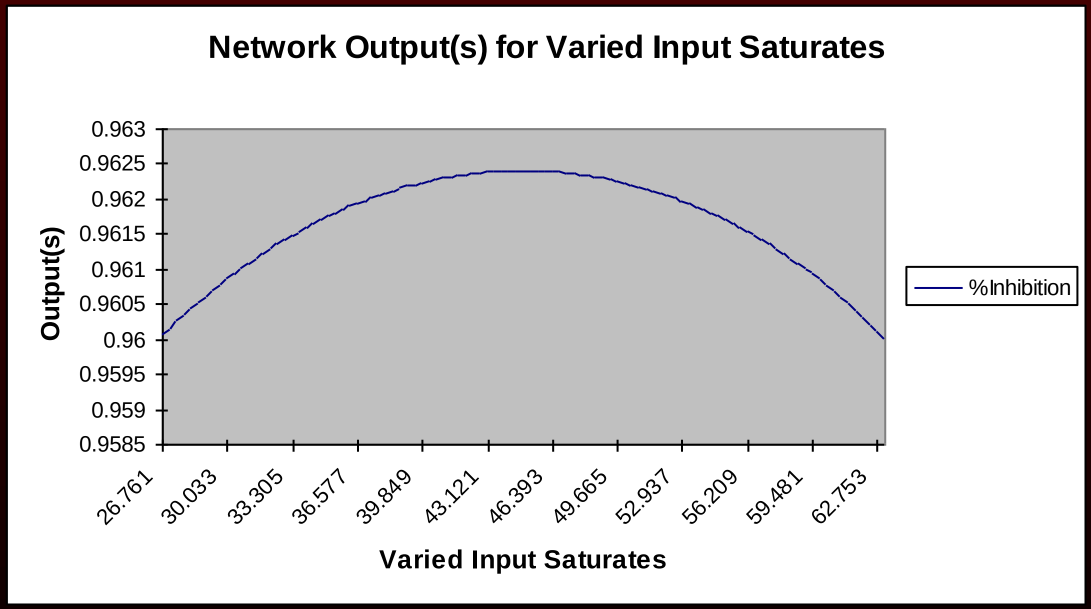
- **Included all attributes collected**
- **Sensitivity about the means**

- **Found many limitations to current method**

How are we to explain a more complex situation?

Variable Behavior



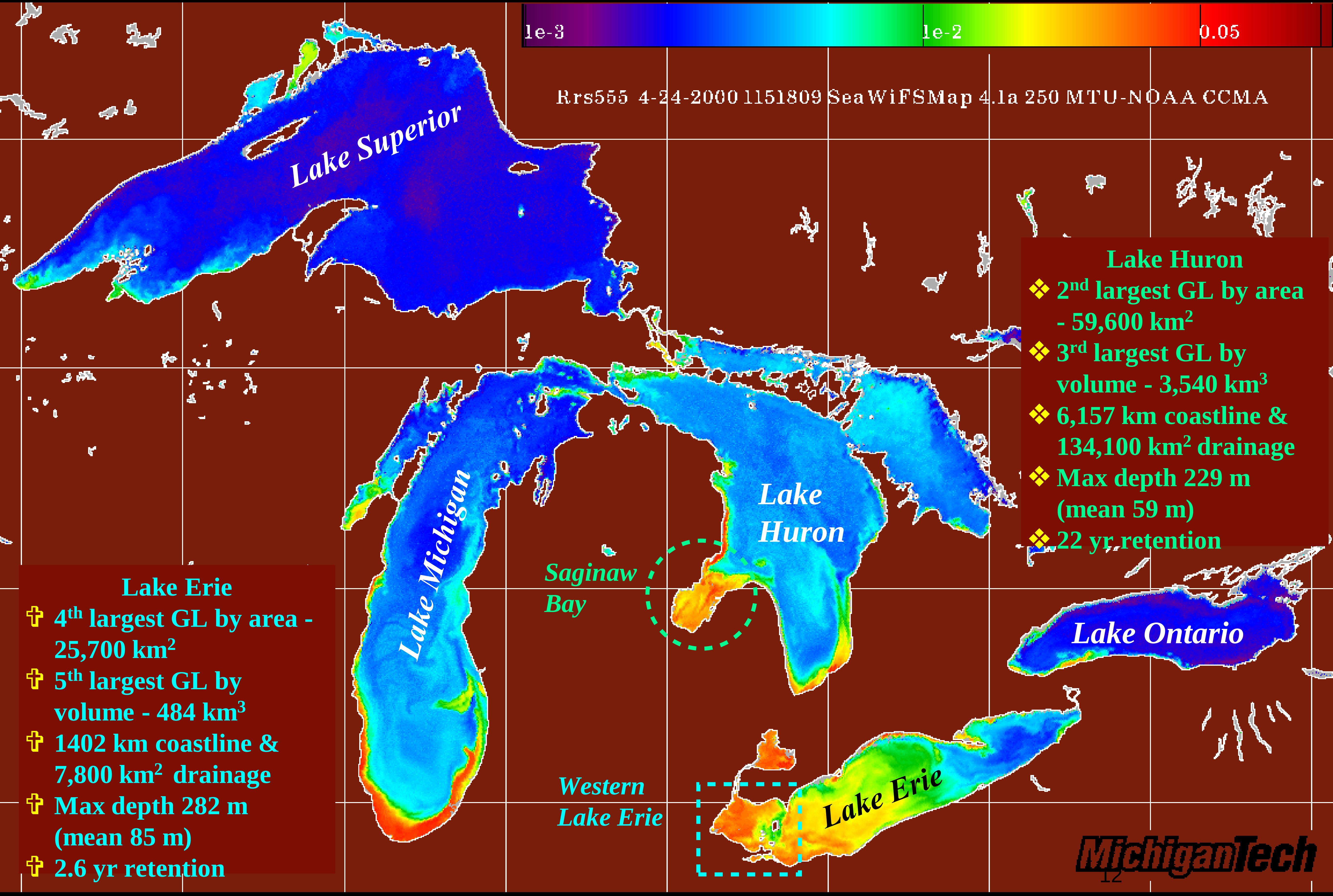


Lake Huron

- 2^{nd} largest GL by area - 59,600 km^2
- 3^{rd} largest GL by volume - 3,540 km^3
- 6,157 km coastline & 134,100 km^2 drainage
- Max depth 229 m (mean 59 m)
- 22 yr retention

Lake Erie

- 4^{th} largest GL by area - 25,700 km^2
- 5^{th} largest GL by volume - 484 km^3
- 1402 km coastline & 7,800 km^2 drainage
- Max depth 282 m (mean 85 m)
- 2.6 yr retention



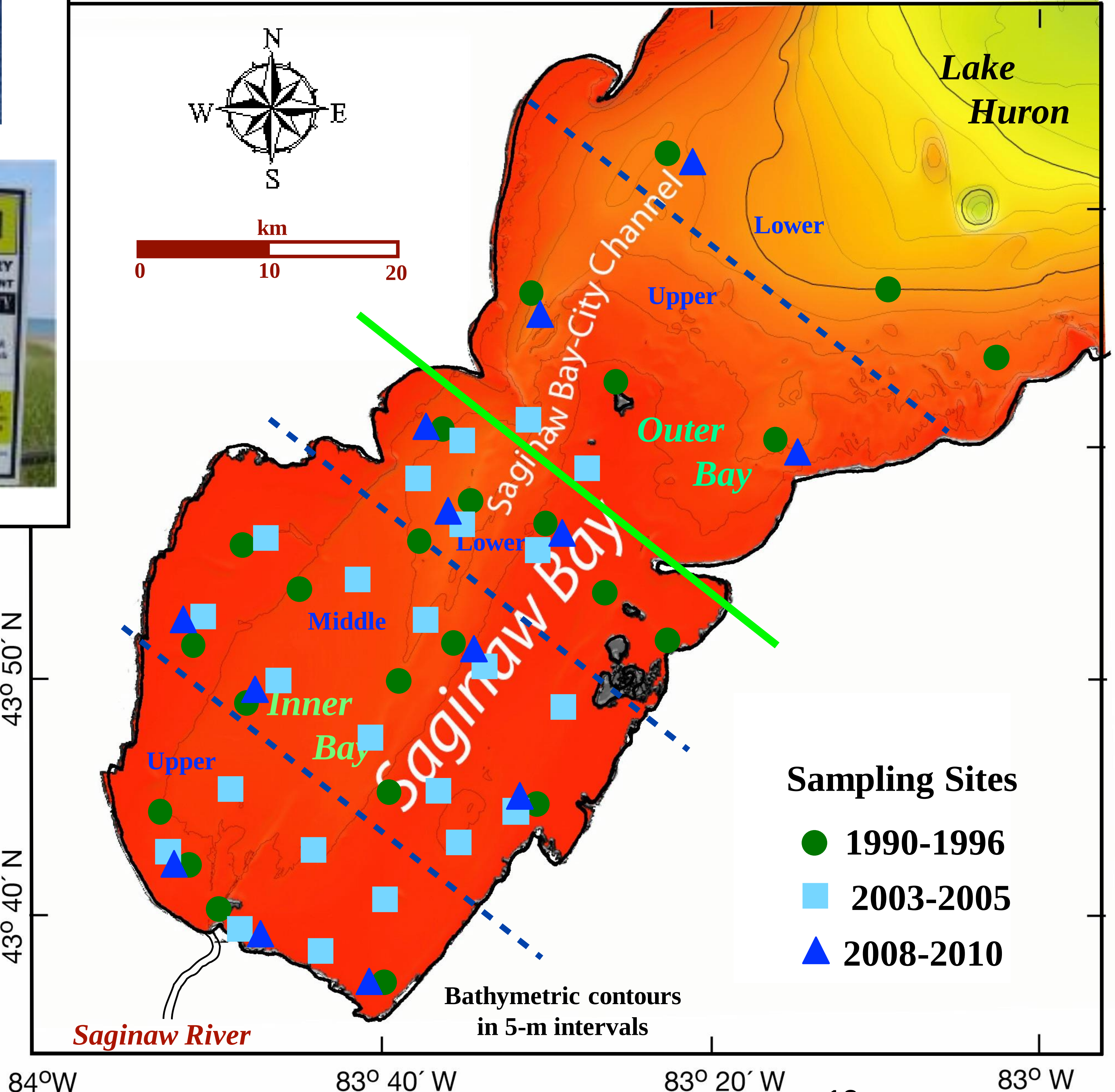
MichiganTech

Zebra Mussels & Water Quality Assessment (1990 - 1996)

Oceans & Human Health Initiative (2003 - 2005)

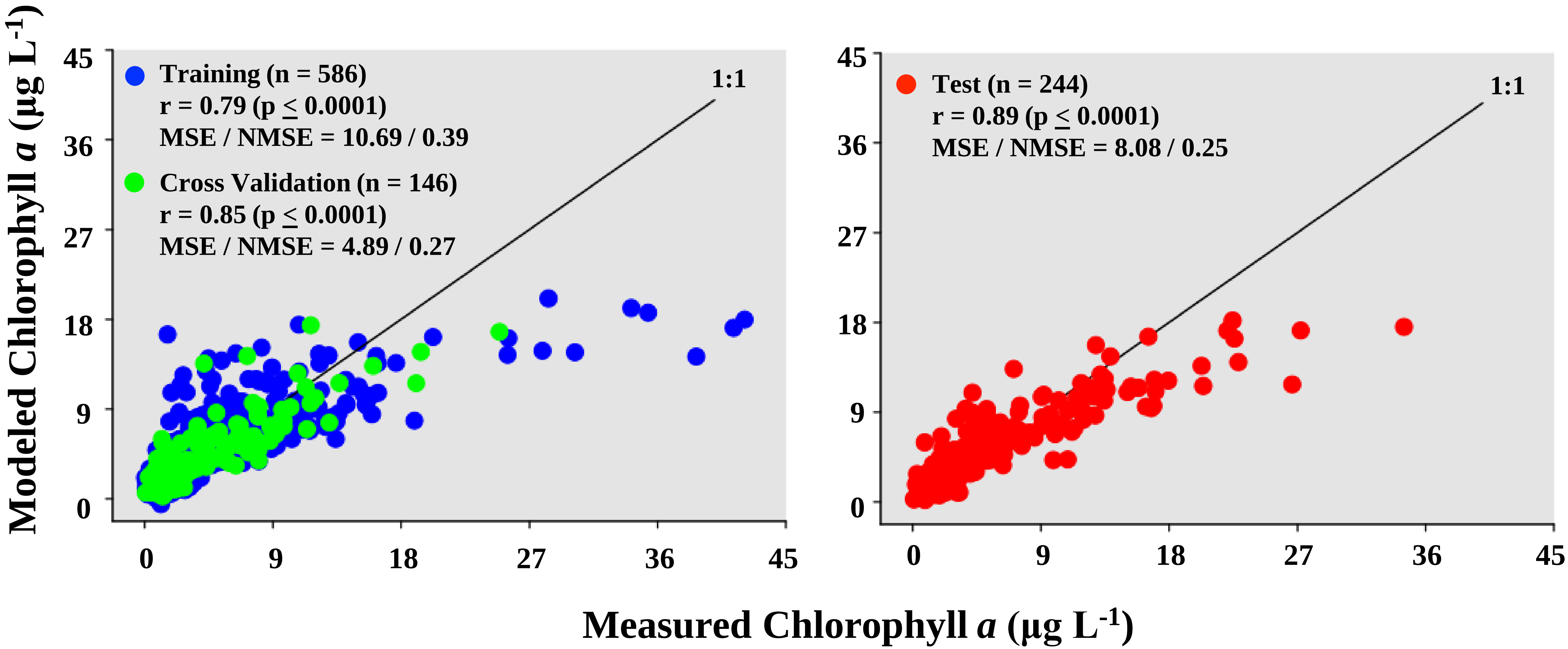
Multiple Stressor Program (2008 - 2010)

	Inner	Outer
Max / Mean Depth (m)	14.0 / 5.09	40.5 / 13.66
Surface Area (km^2)	1,554	1,217
Volume (km^3)	7.91	16.63
Retention time (d)	58 - River mouth to Lake proper	



Predicting Saginaw Bay Chl a *(1991-1996)*
MLP - 1 Hidden Layer of 4 Processing Elements

Hydrological Predictors: **^oC, Sechhi, K_d, Cl, NO_3, NH_4, SRP, TP, SiO_2, $PSiO_2$, DOC, POC**



Existing Knowledge Extraction Tools

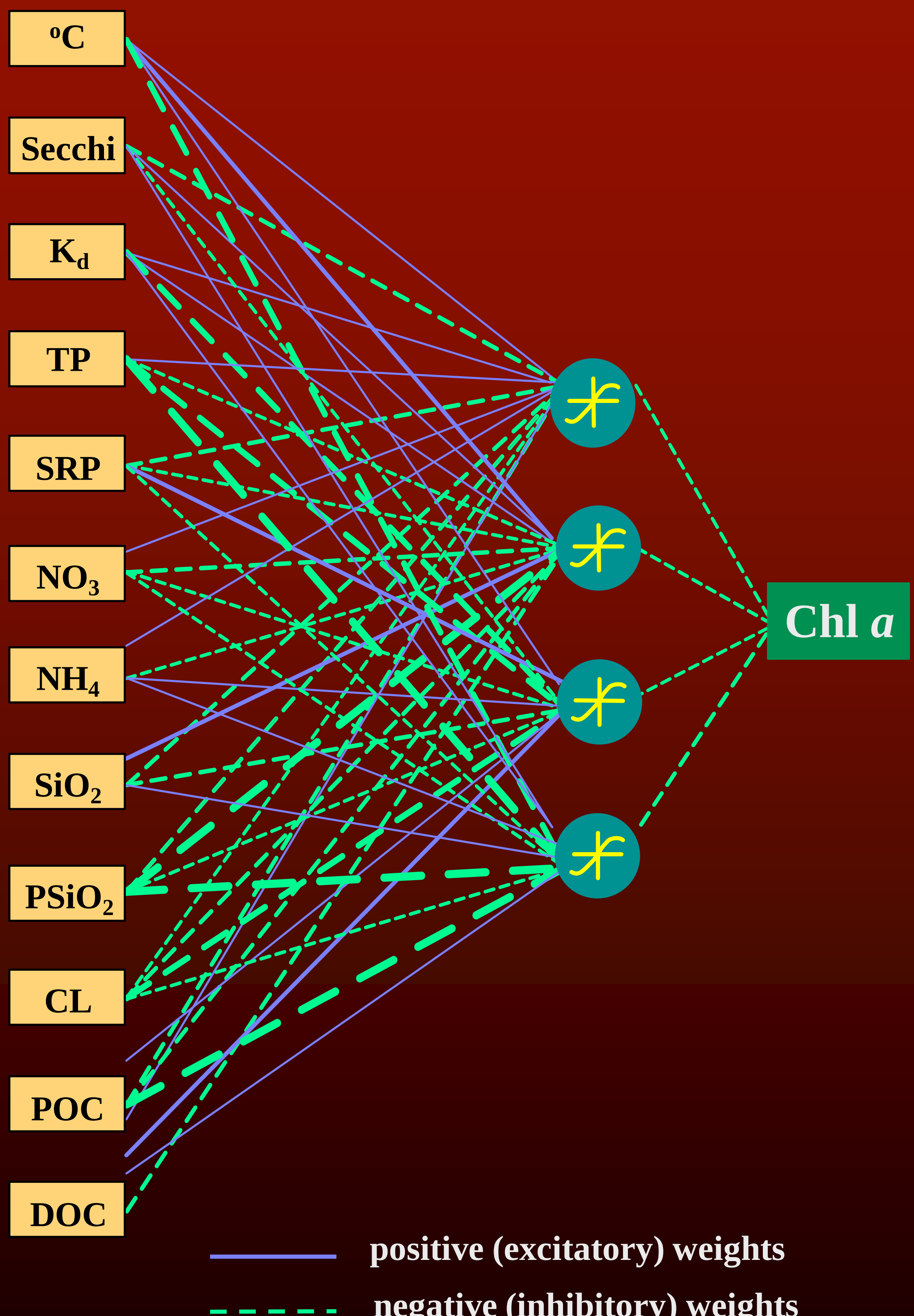
Neural Interpretation Diagram

- Decomposition method to visual
  - Determine significance of input variables
  - Based on the magnitude of interconnecting weights

Connected Weights

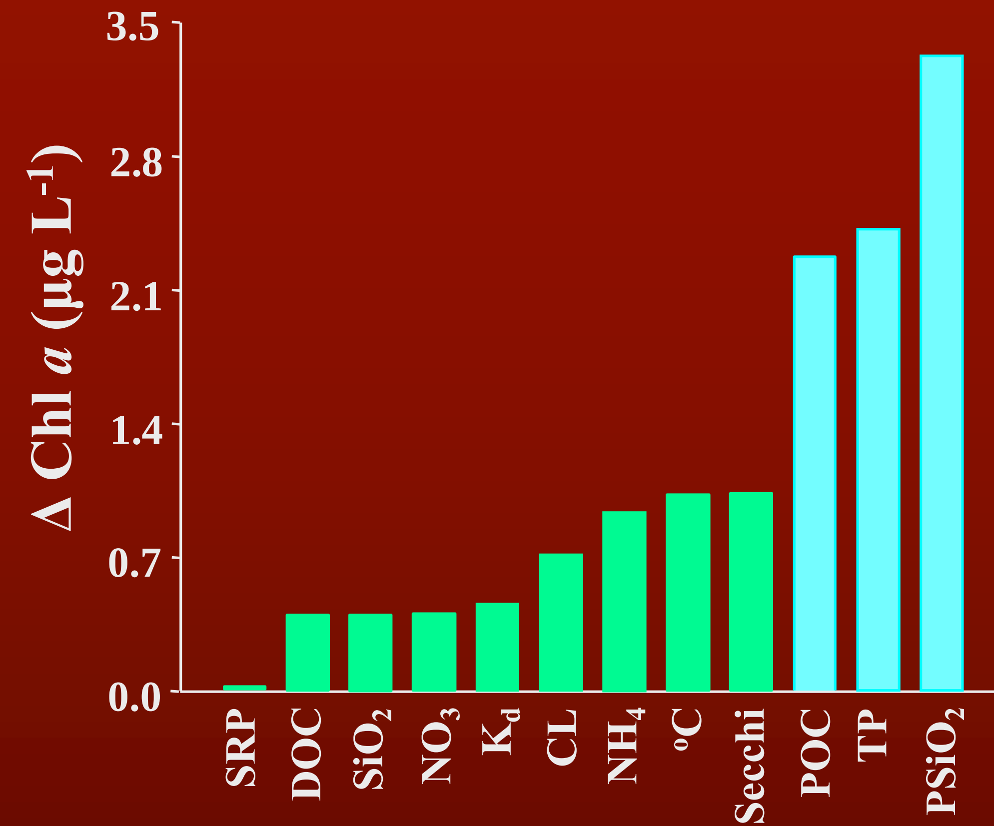
- Decomposition method that uses weights of an ANN to determine:
  - Input Significance to model
  - Nodes Significance to ANN
- Procedure
  - Calculate "connected weights" for all possible paths of the network

*Network Interpretive Diagram**
(of a trained network)

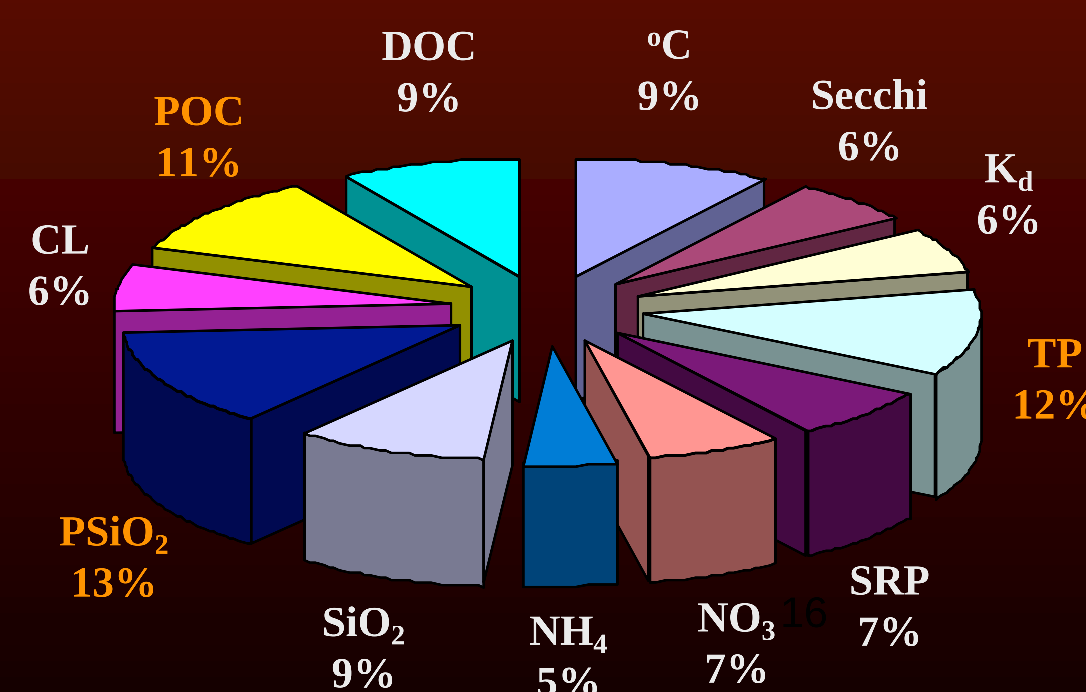


* Line thickness portrays the relative magnitude of the weight

Single Parameter Sensitivity Analysis (± 1 SD)

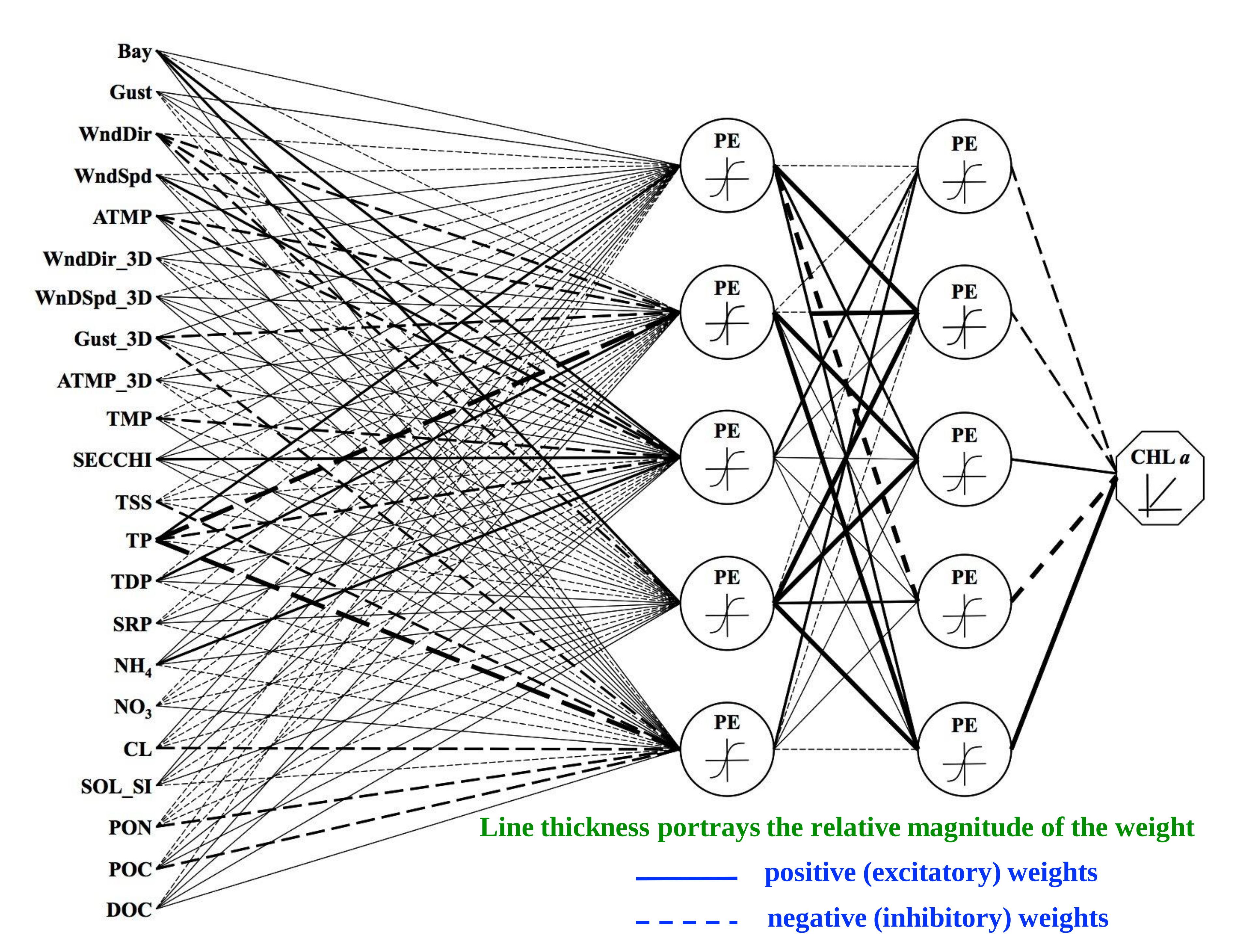


Garson's Algoritthm
Relative Share of Prediction



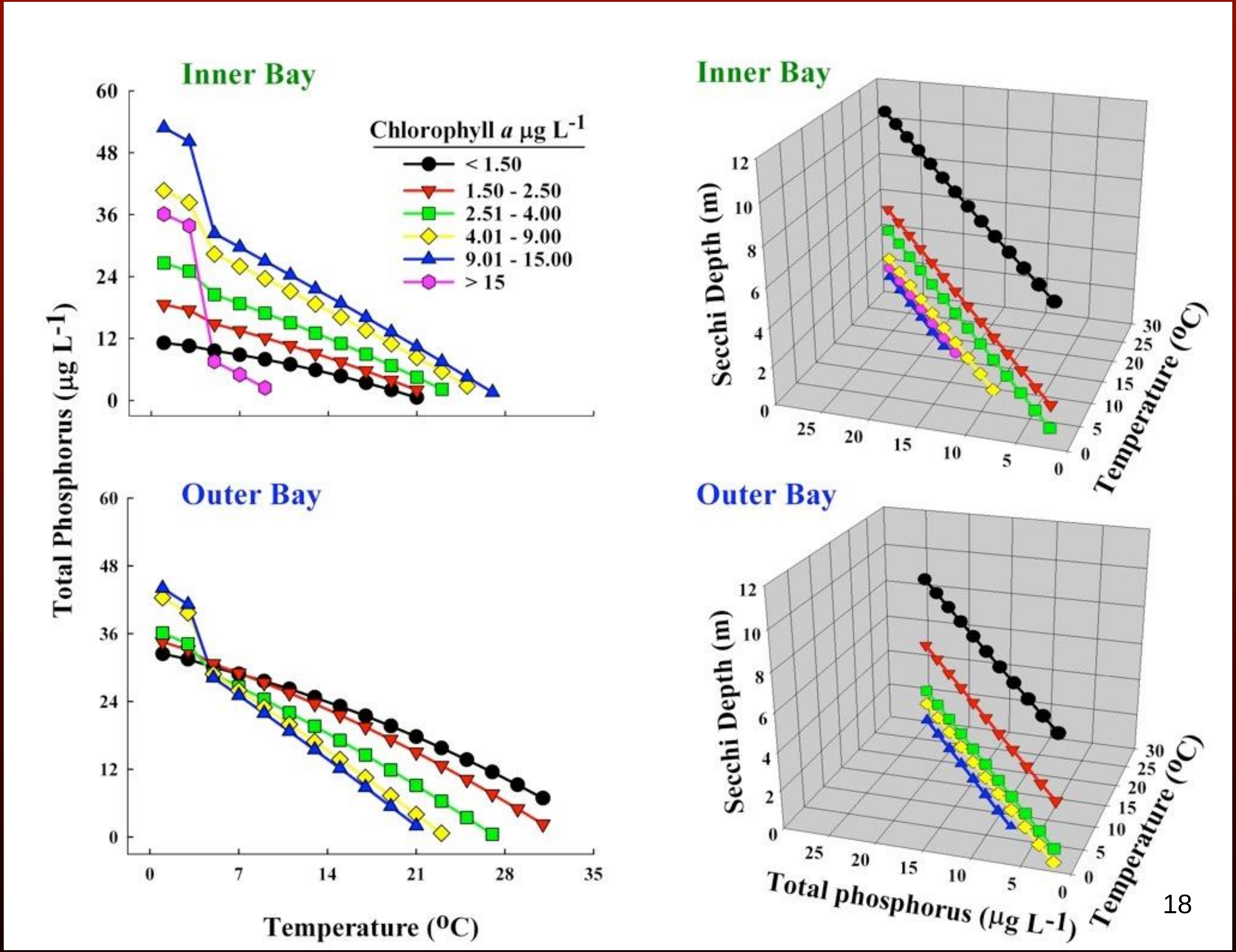
Developed More Complex Networks

Saginaw Bay CHL a (2008-2010) - Hydrological & Meteorological Predictors

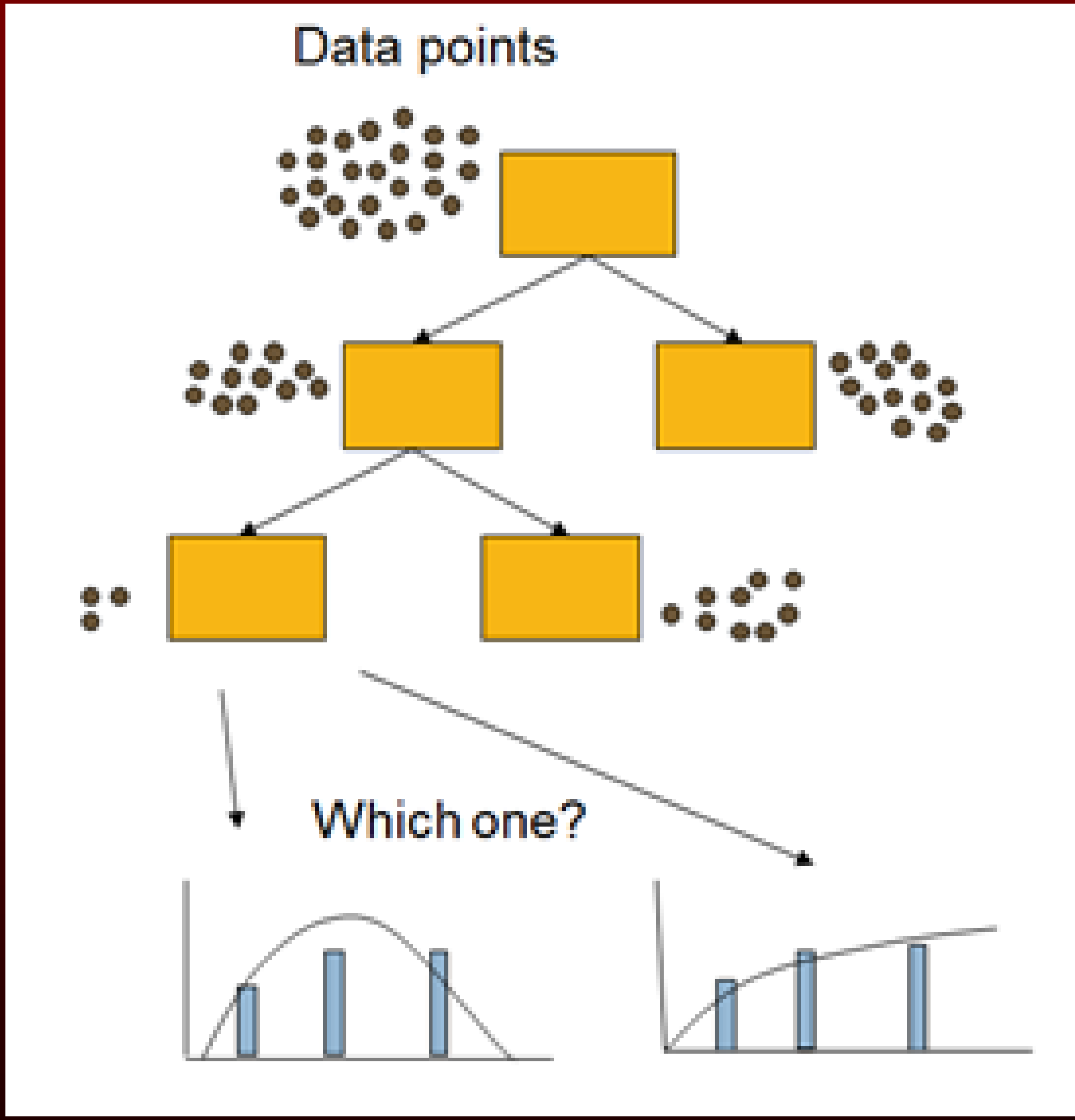


Developed New Approaches to Observe Interactions

Multi-Variable Sensitivity Analysis (circa 2006 !)

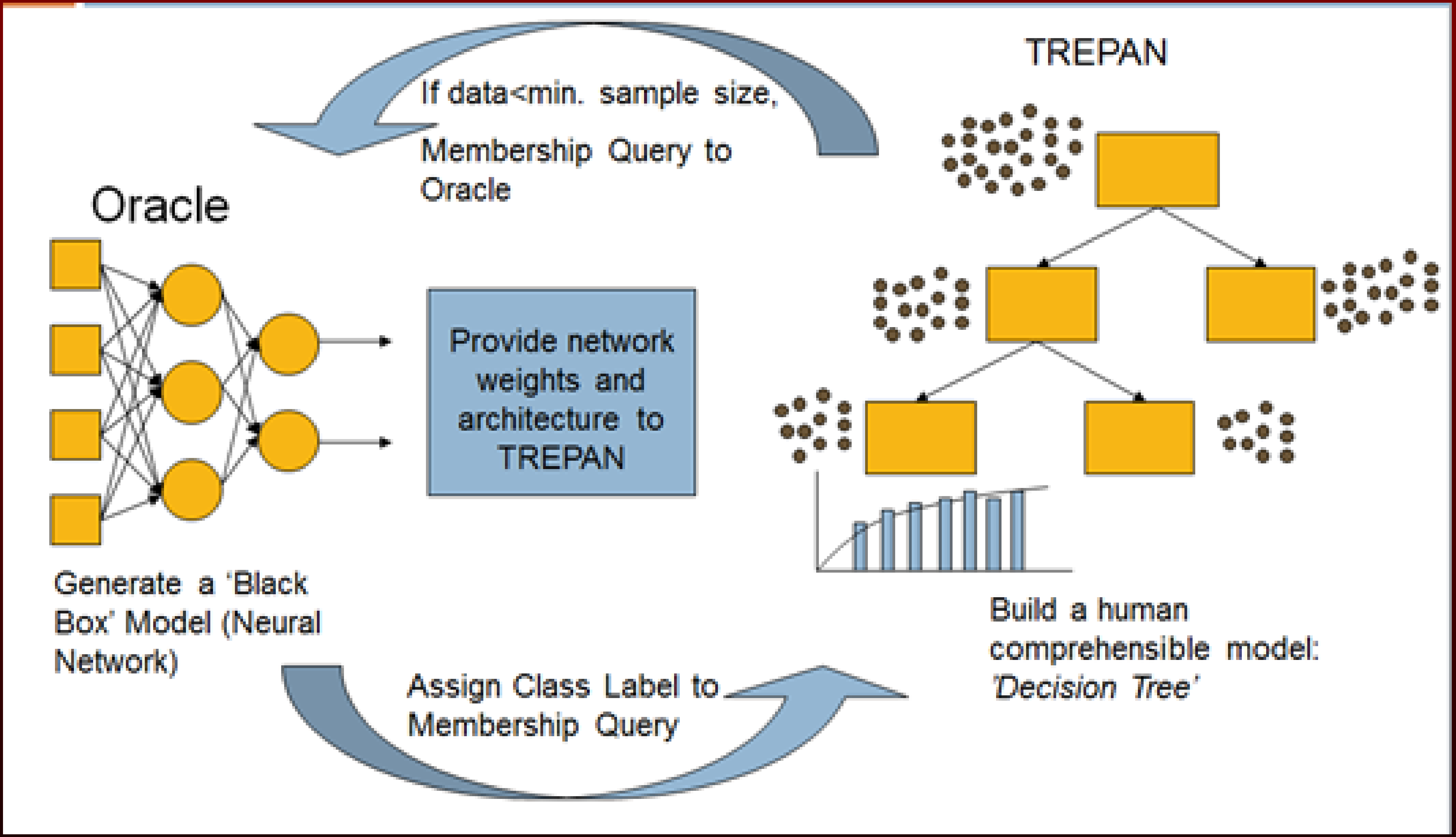


Decision Trees



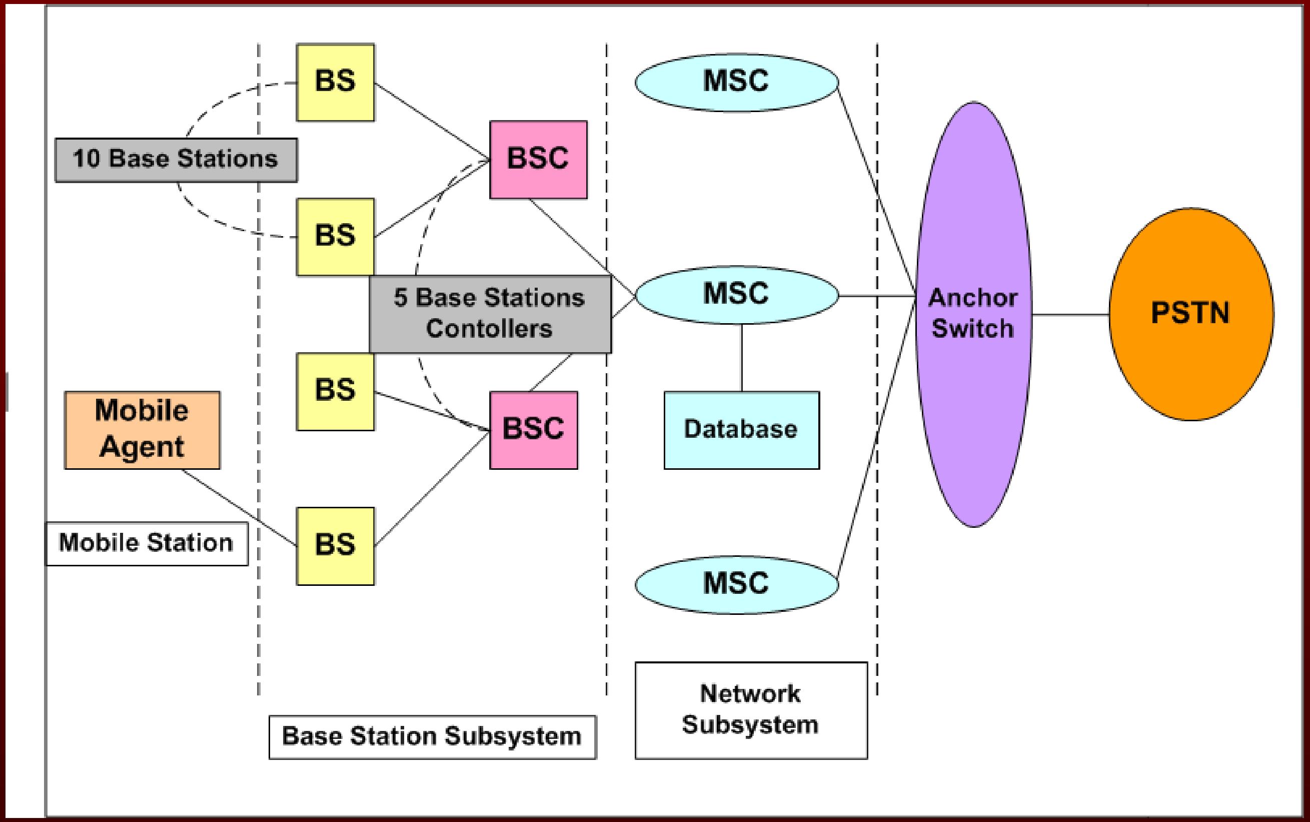
- Symbolic Knowledge Extraction Technique
- Most commonly used decision tree induction algorithm – C4.5 (Quinlan)
- Recursive partitioning of the data
- Drawback: Amount of data reaching each node decreases with the depth of the tree
- Alternative: TREPAN

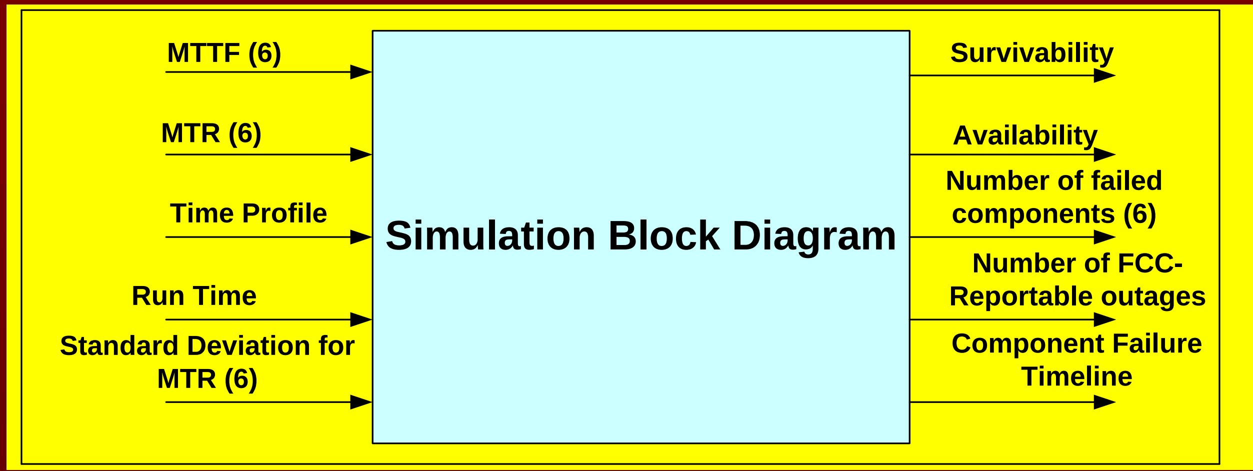
TREPAN+ Methodologies

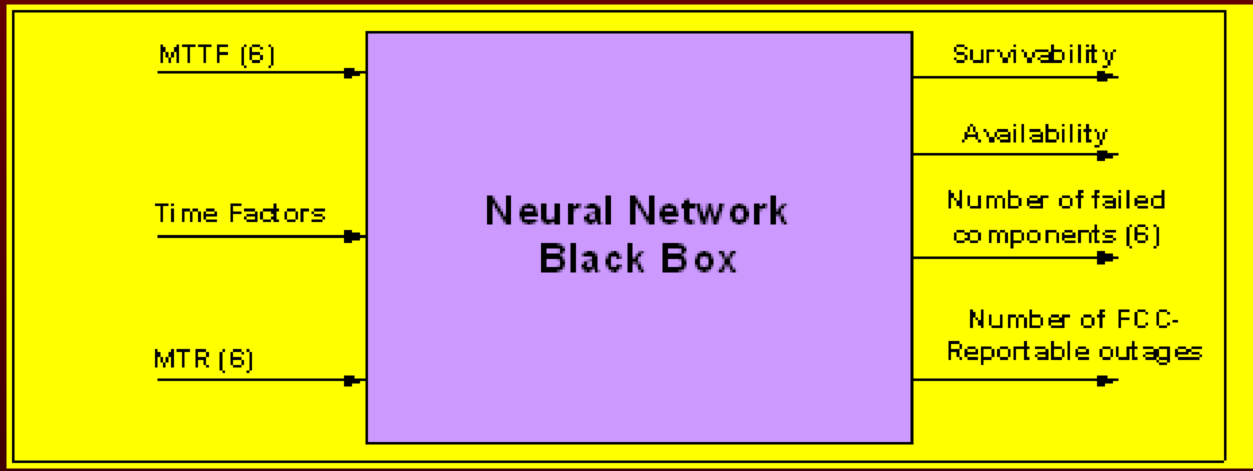


Simulation Based Neural Network Modeling

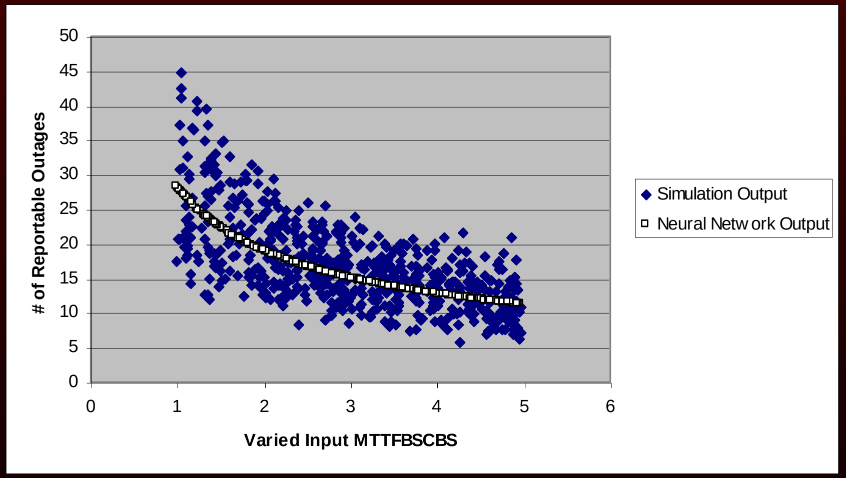
Investigate training a NN network with results from wireless simulation



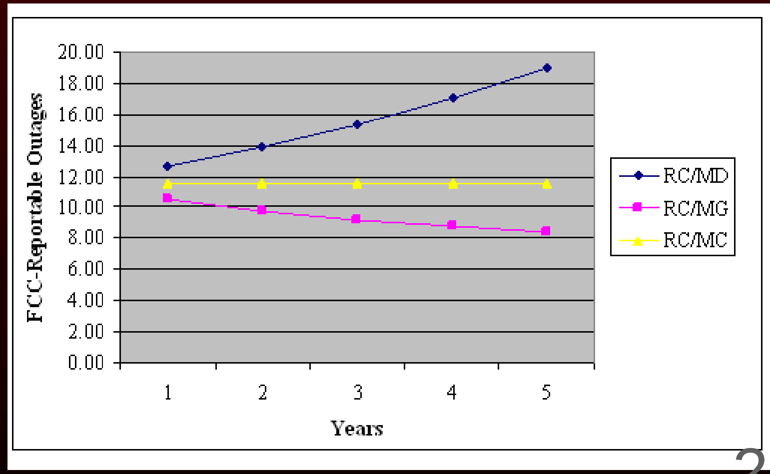




Input & Output Behavior



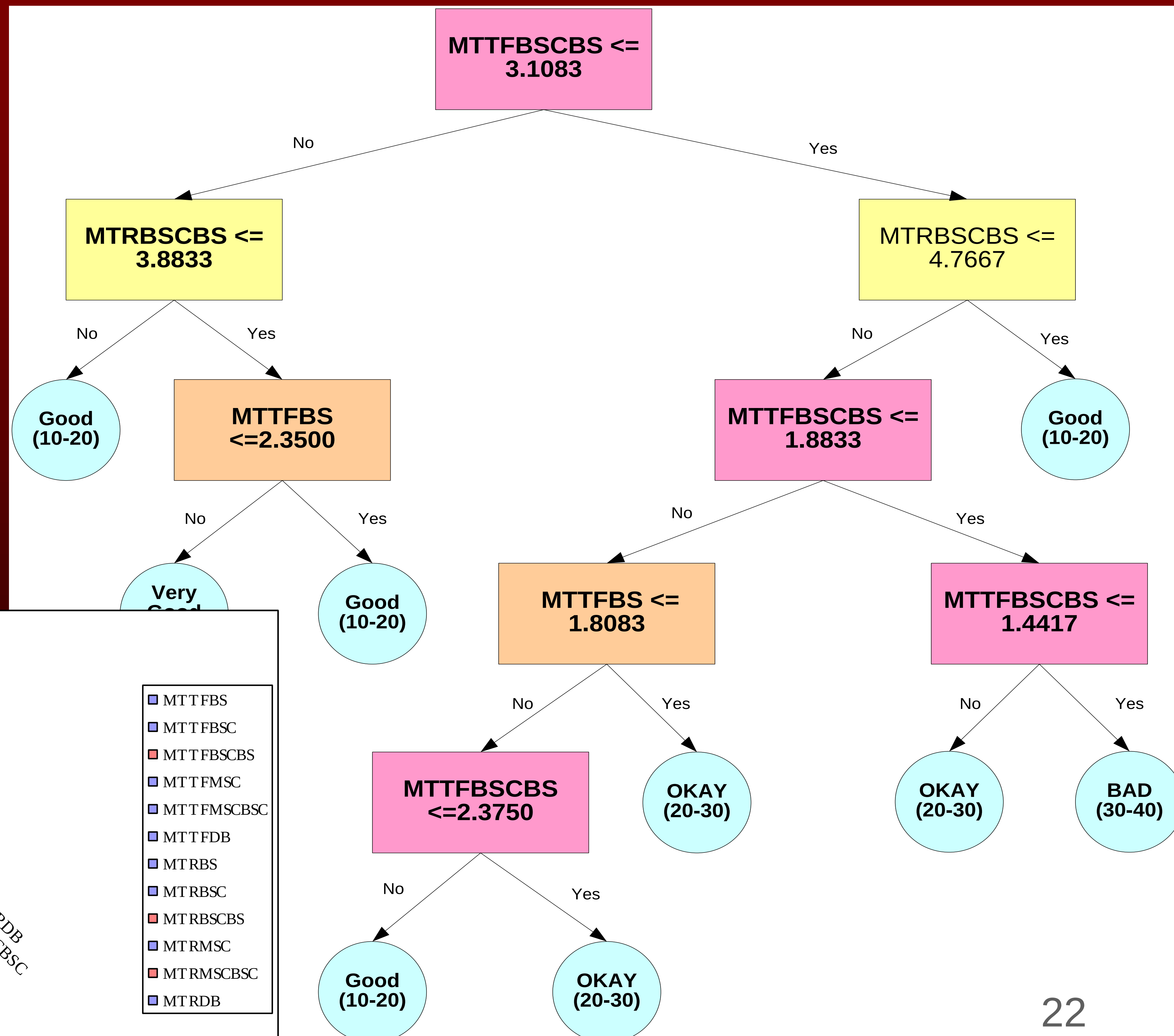
Scenario Testing

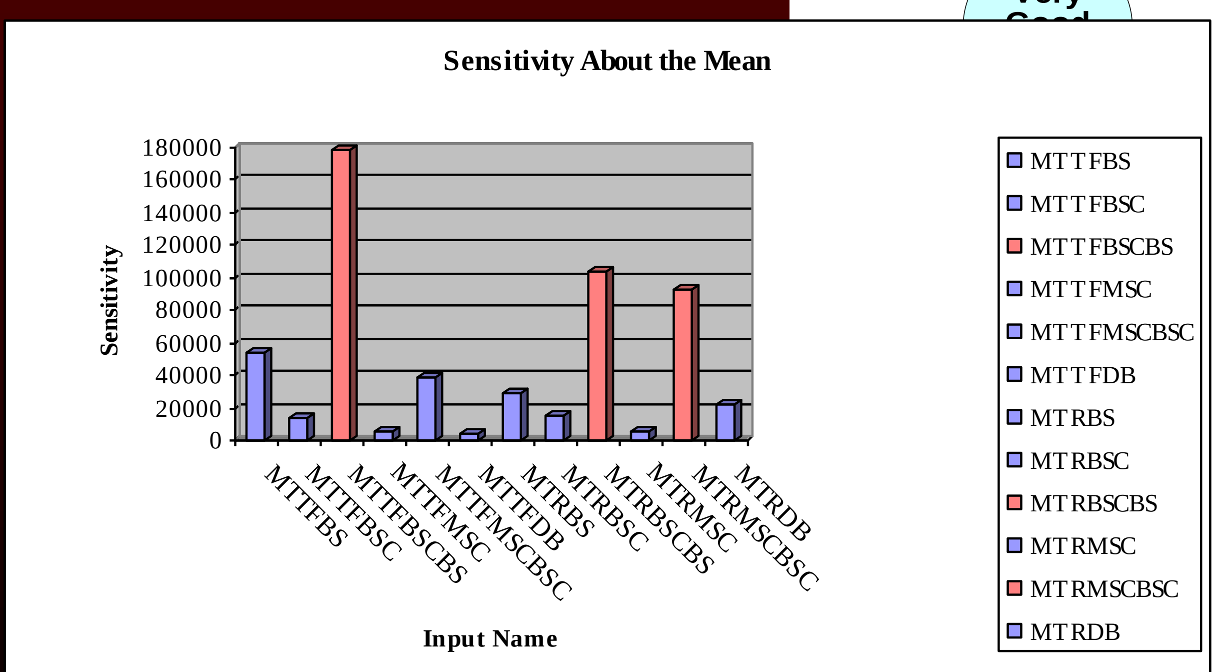


Knowledge Extraction for Wi-Fi

Utilized Techniques:

- TREPAN+
- Sensitivity

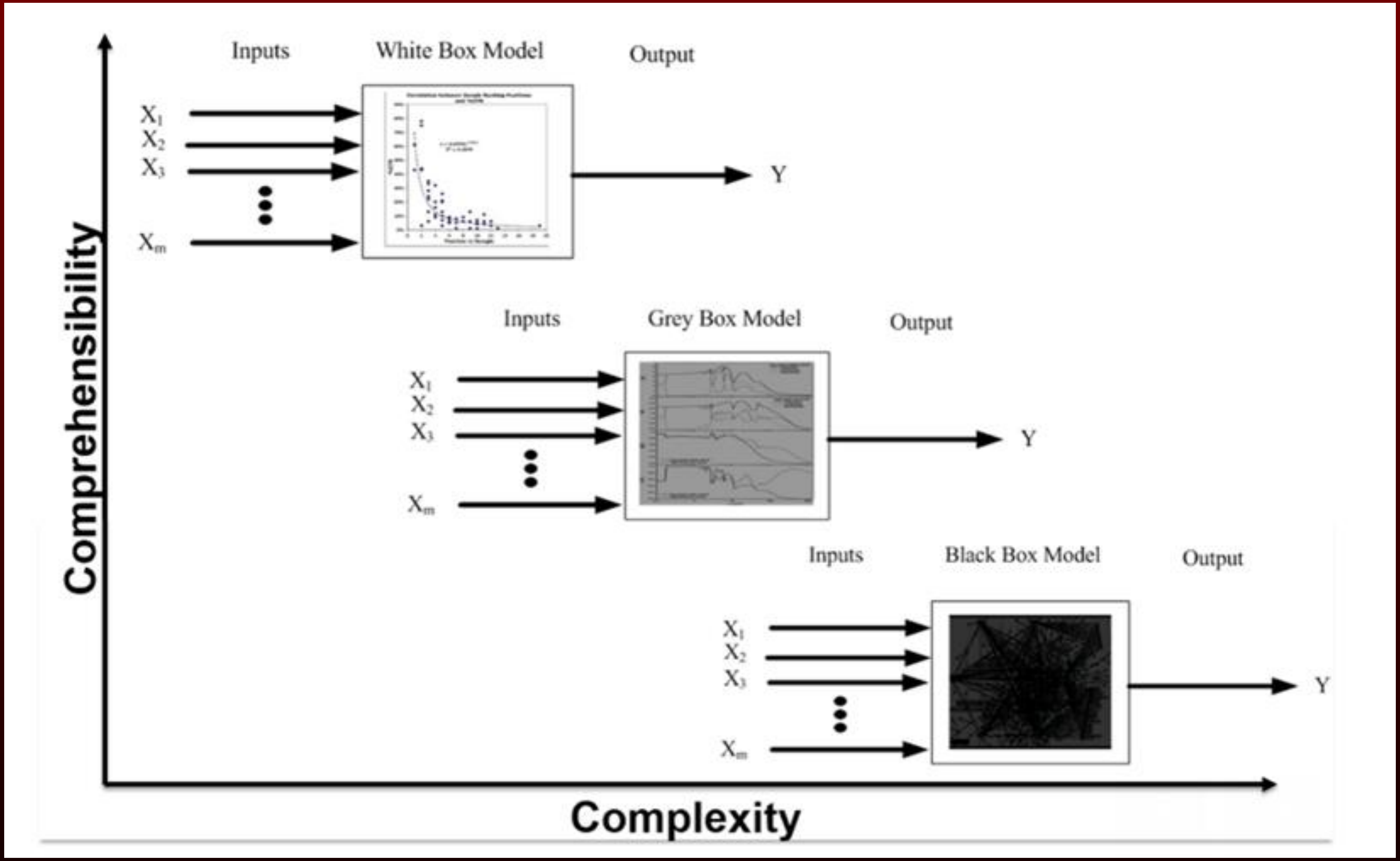




Needed More Understanding: Variable Interactions

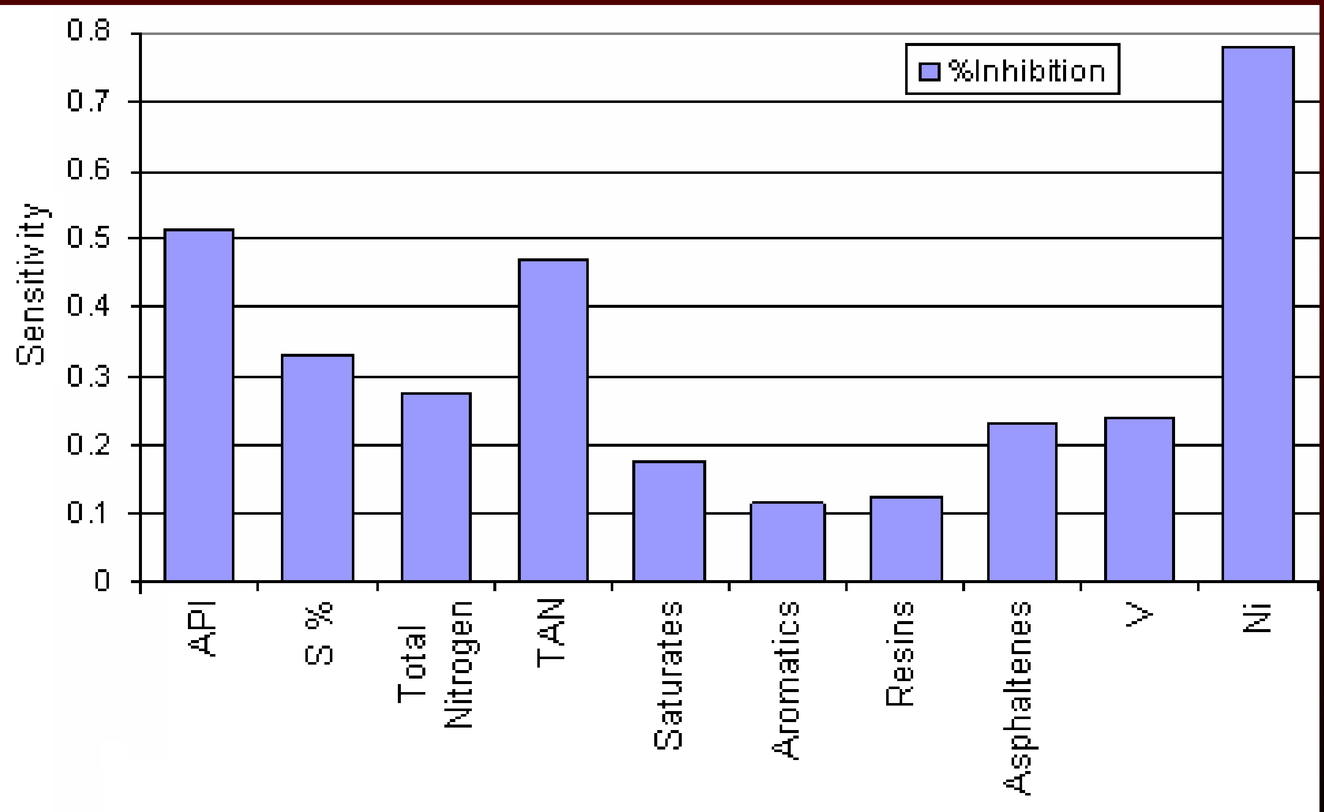
Multiple Variable Interactions while looking at various states!

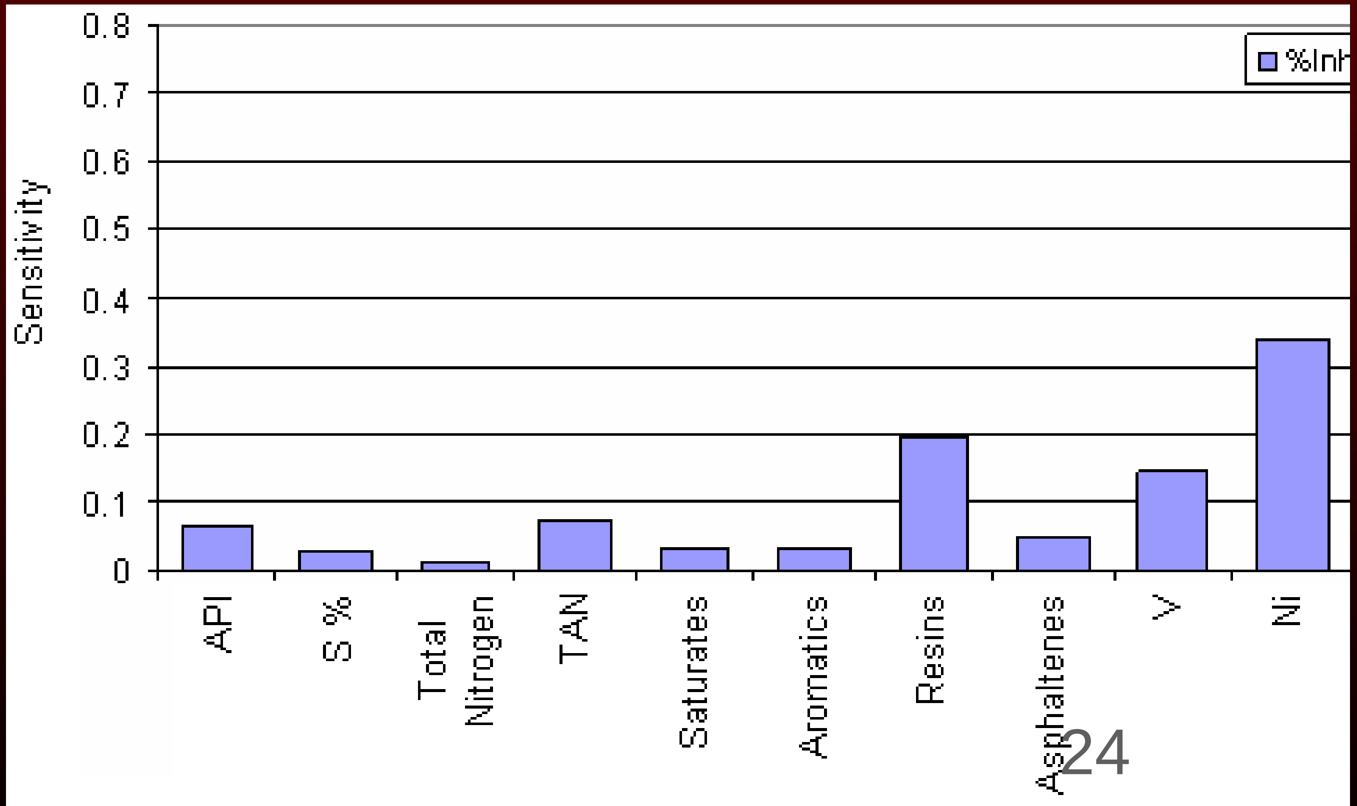
Our drive to Mechanistic Model: Grey Box => WHITE BOX



Different Project: Crude Oil Impact

- **Used New Set of Tools:**
  - **Limitations to Sensitivity:**
    - **2 ANNs were created for “high” and “low” %Crude Oils**
    - **Sensitive results were very different**
    - **%Crude Oil <=20%** **Crude Oil >=50%**

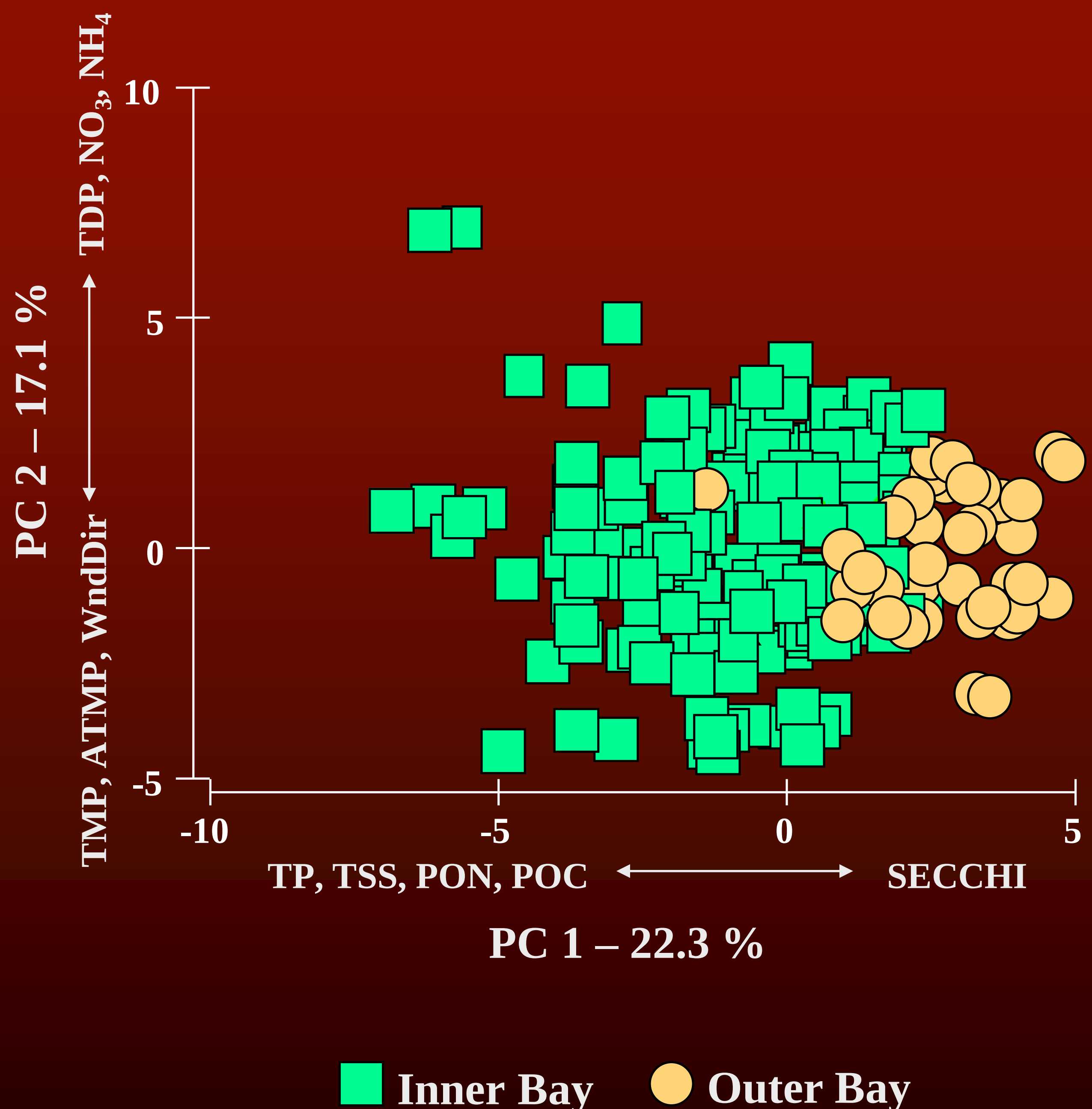




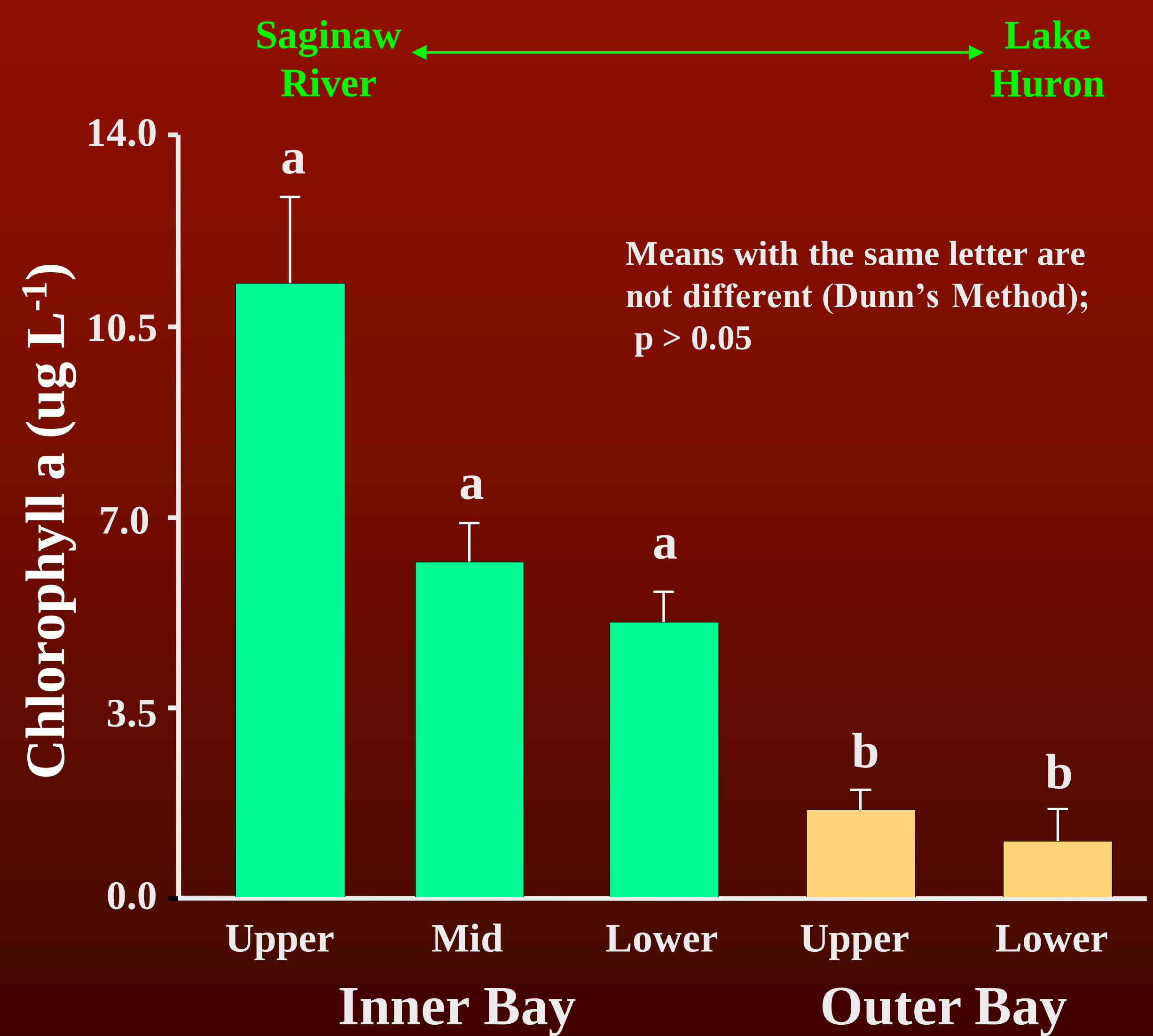
Revised Look: Saginaw Bay 2008 - 2010

PCA – Hydrological / Meteorological Variables

Monthly Means



Mean [Chlorophyll *a*]

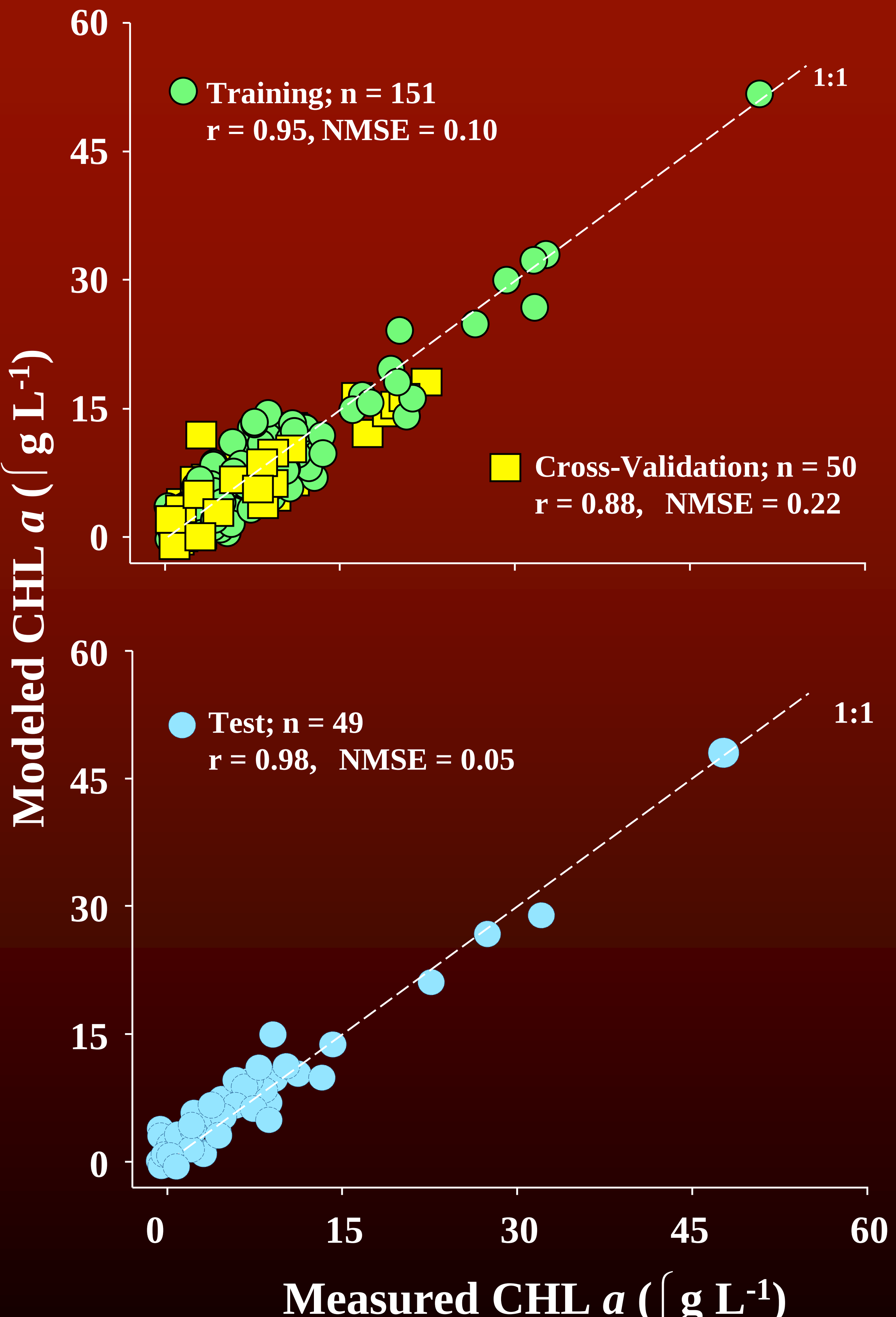


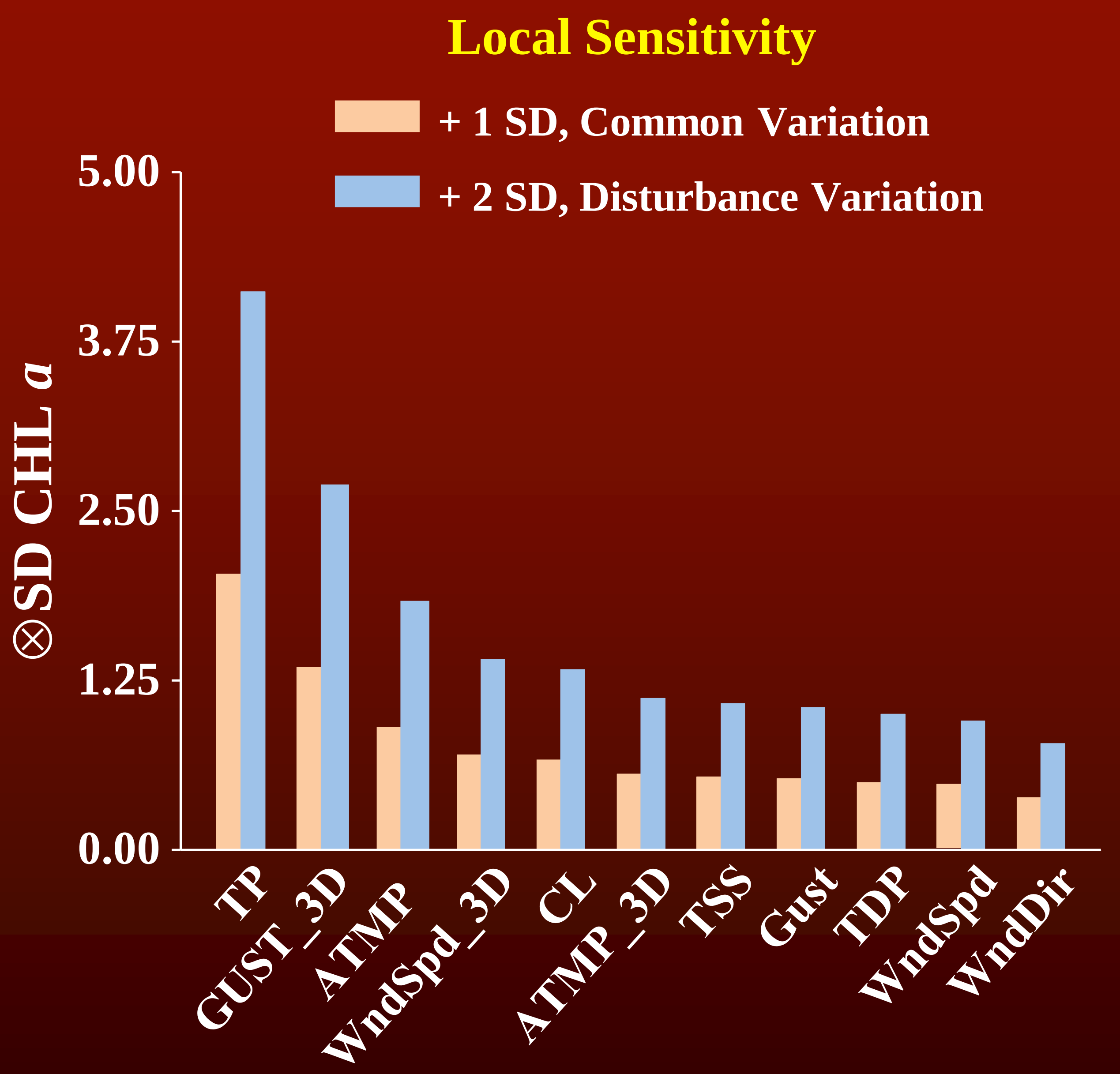
Inner vs Outer Bay:
Mann-Whitney U Statistic = 559
T = 1024; p = < 0.001

Distinct difference in Inner and Outer Bays

Bay Reach:
Kruskal Wallis ANOVA on Ranks = 559
H = 50.65, df = 4; p = < 0.001

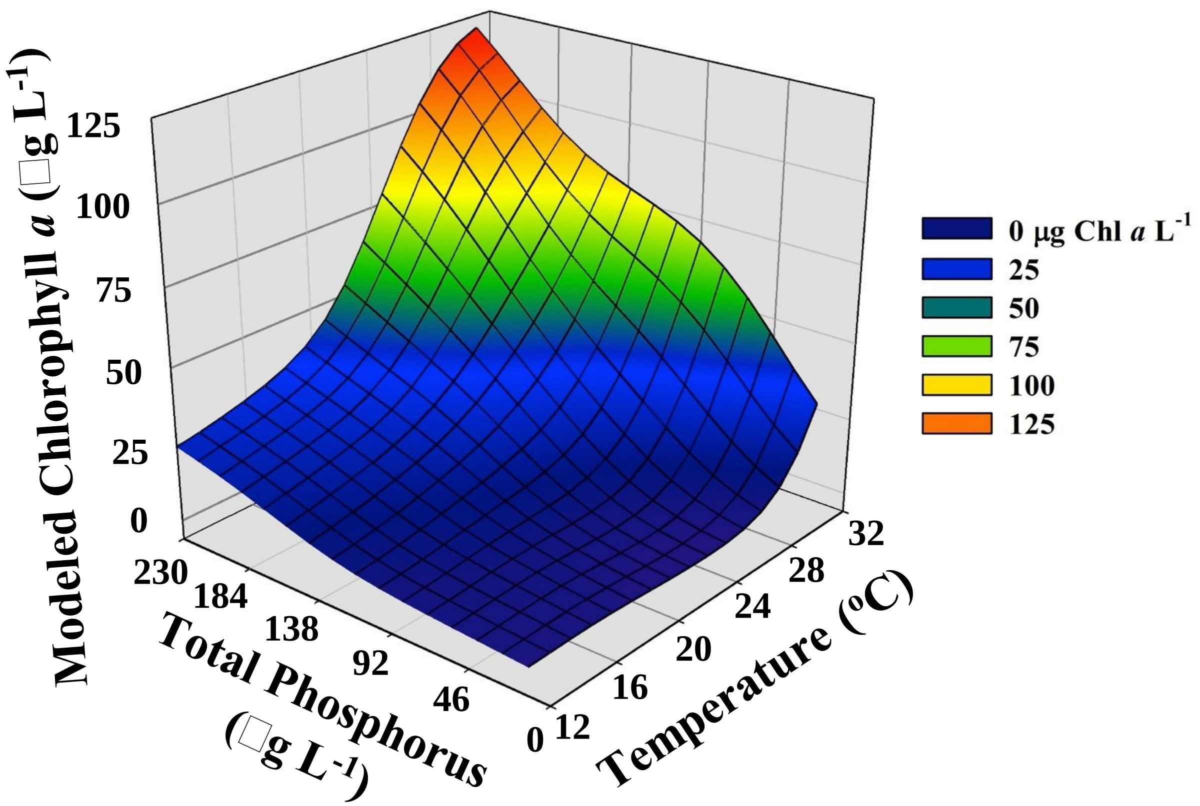
Distinct differences based on range of Sensitivity

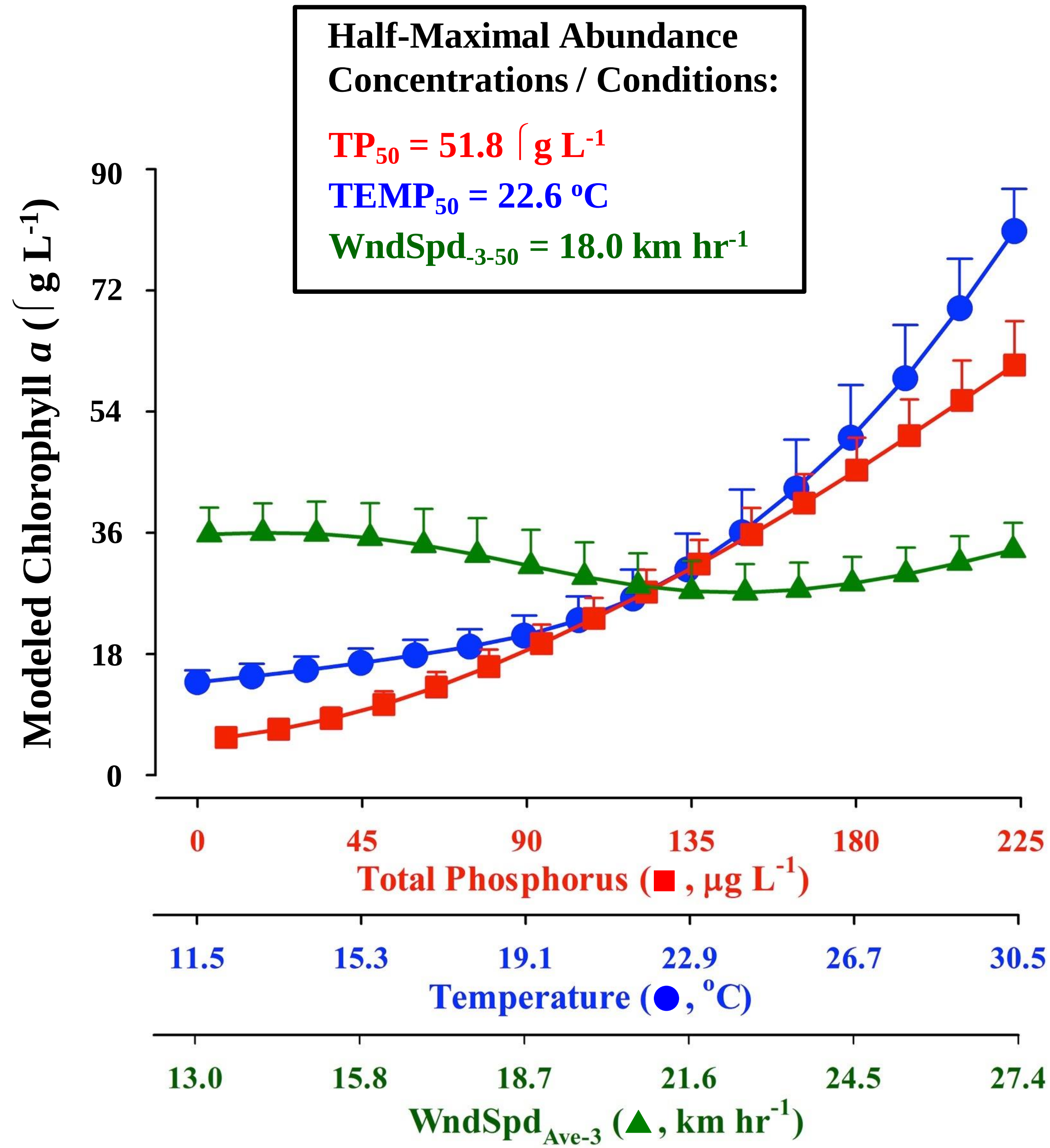




Introduced New Visualizations: Multi-variable Impact on *Chlorophyll a*

CHL as a function of TP & TEMP





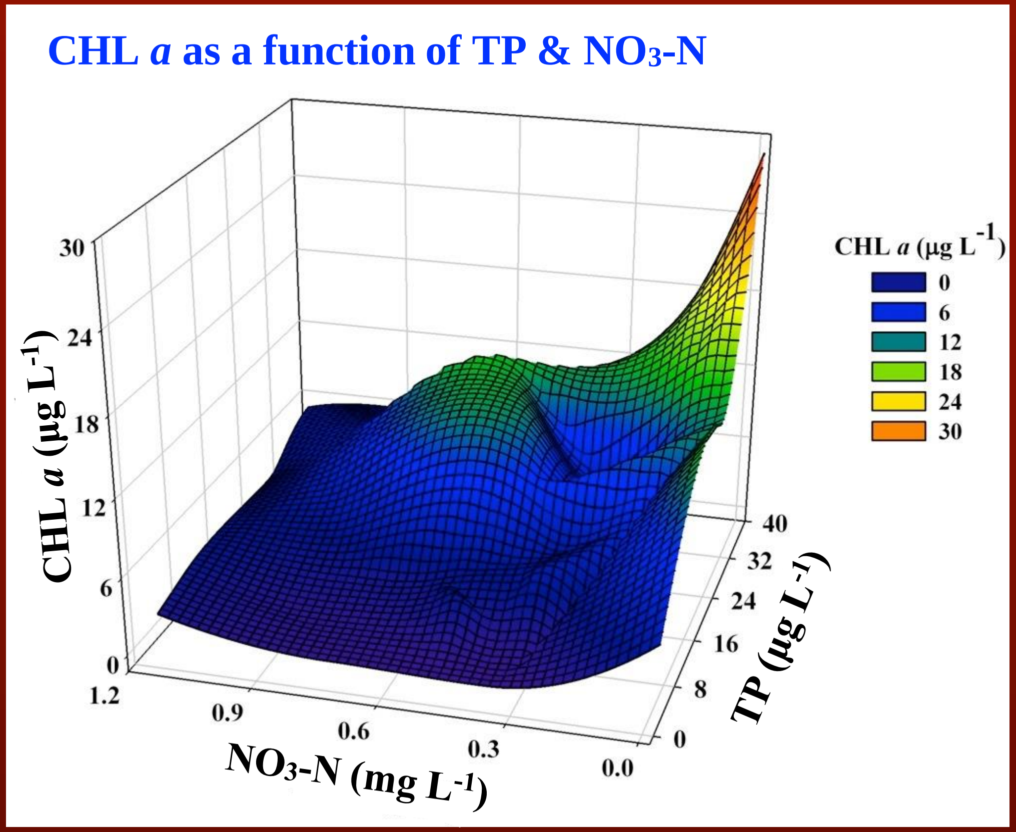
CHL $a^{0.5}$ = 1.98 + (0.03*TP)
adj r^2 = 0.99, Fit SE = 0.41, Fstat = 29857.36

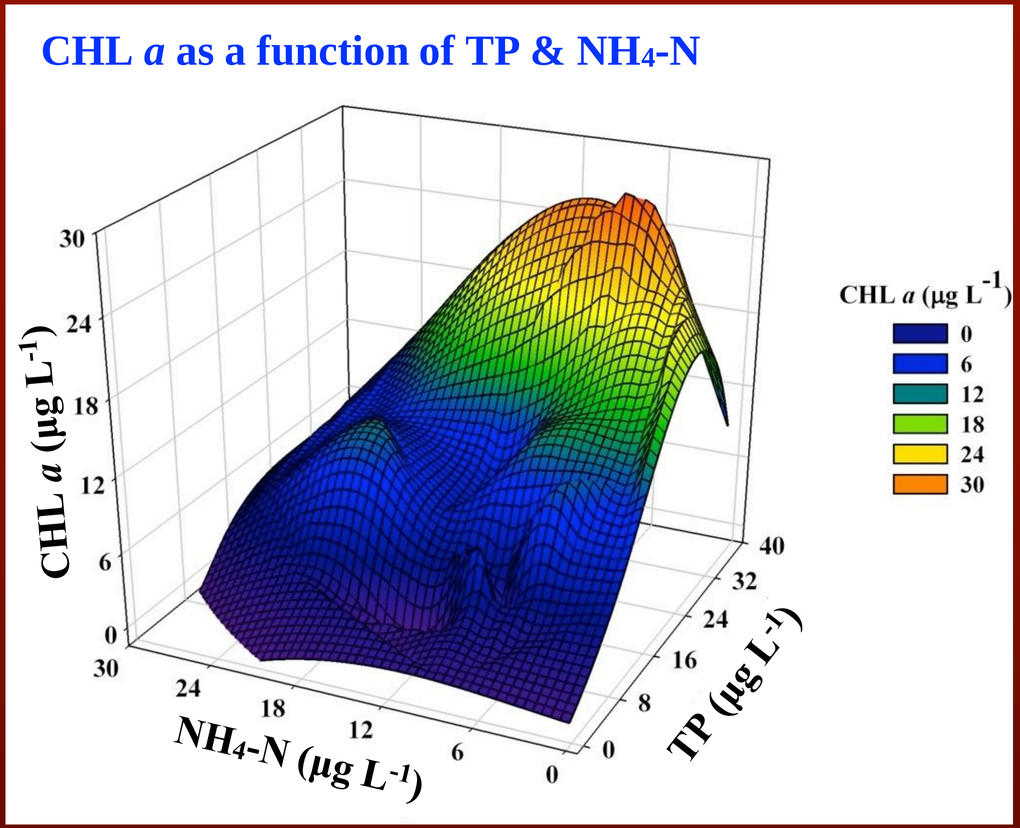
ln CHL a = 2.23 + (0.002*TEMP^2)
adj r^2 = 0.99, Fit SE = 1.03, Fstat = 6323.88

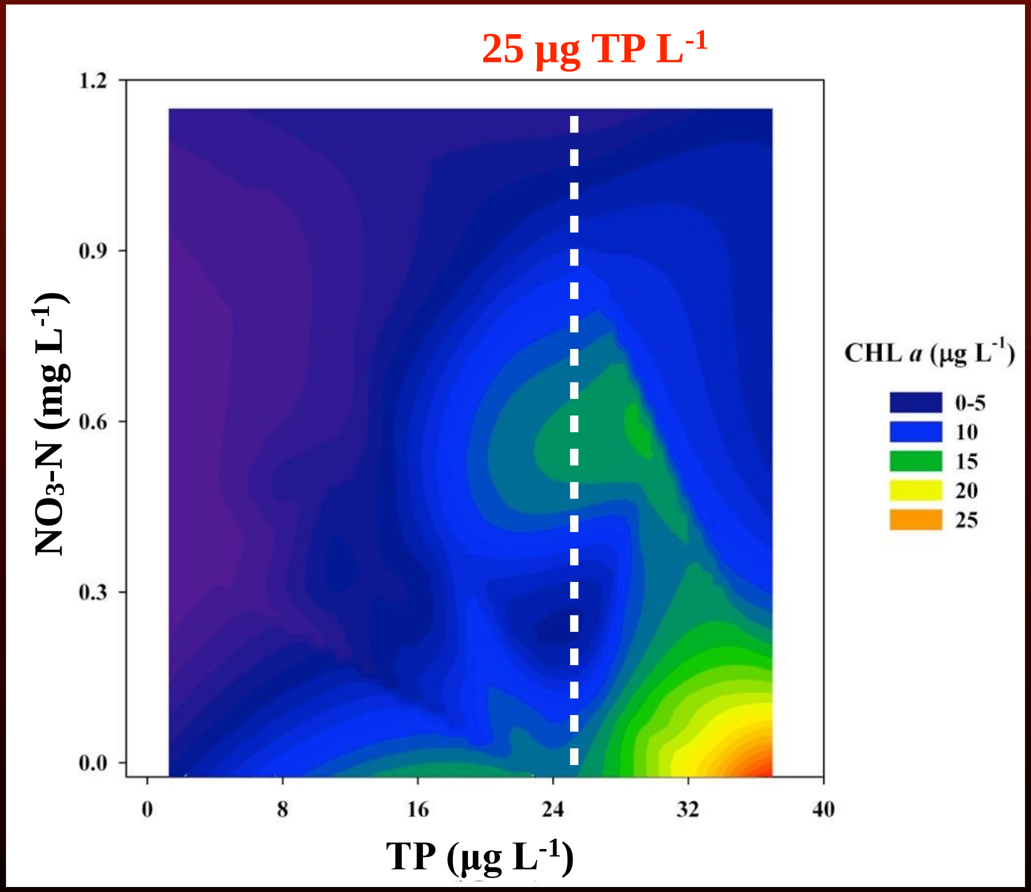
CHL a = -862.16 + (473.88*$WndSpd_{Ave-3}$) - (103.65*$WndSpd_{Ave-3}^2$) + (12.14*$WndSpd_{Ave-3}^3$) - (0.82*$WndSpd_{Ave-3}^4$) + (0.03*$WndSpd_{Ave-3}^5$) - (0.001*$WndSpd_{Ave-3}^6$) + (5.80e-6*$WndSpd_{Ave-3}^7$)
adj r^2 = 0.99, Fit SE = 0.13, Fstat = 13,127.67

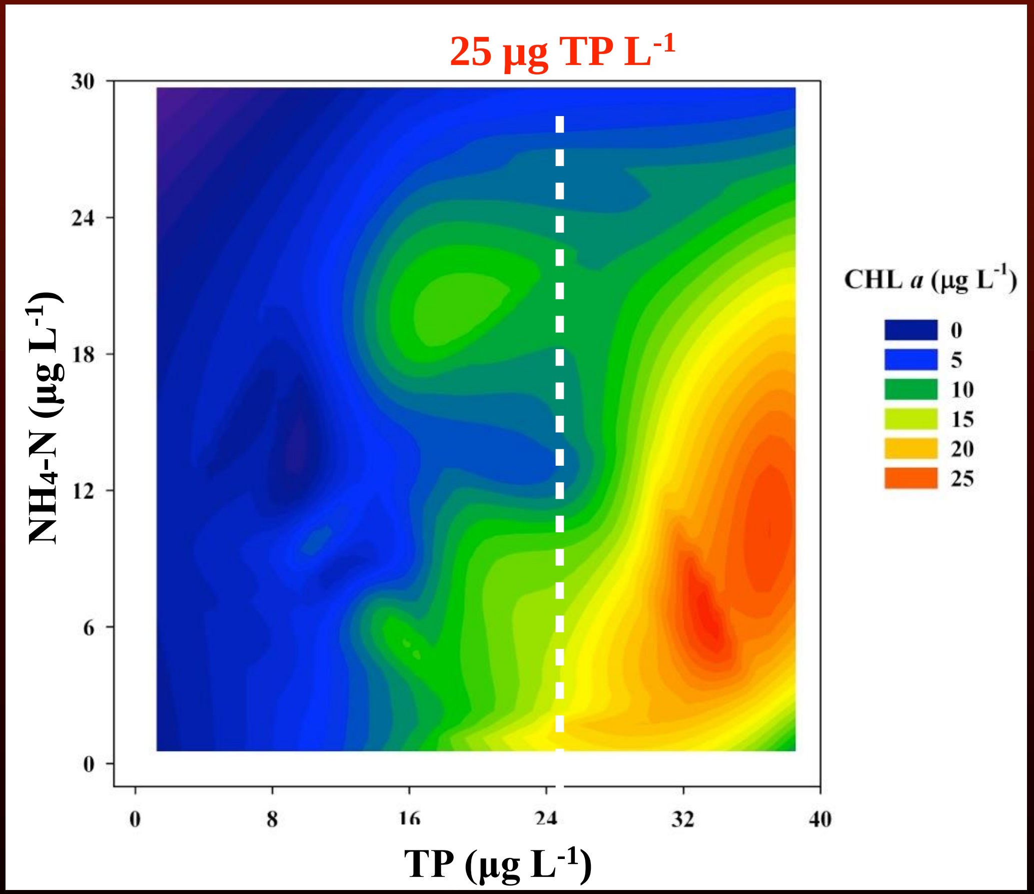
Delineating TP Thresholds for Saginaw Bay CHL a (2008-2010)

(Taking Into Account the Interactions and/or Synergisms of Co-Limiting Nutrients)





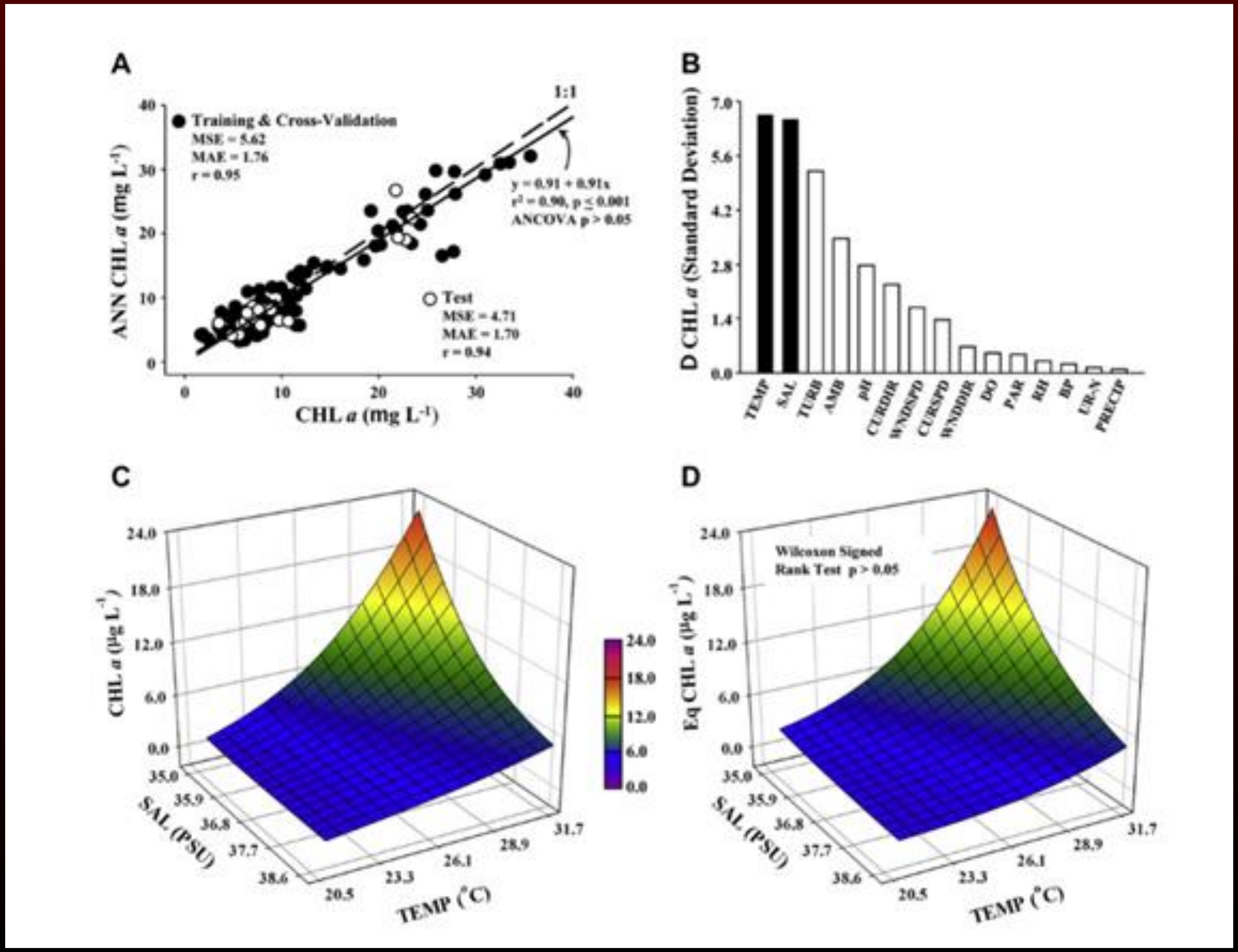




Development of Grey Box Technique

$$[\mathrm{CHL}\ a] = w_1 \cdot f(x_1, y_1) + r_1,\ \ r_1 = w_2 \cdot f(x_2, y_2) + r_2,$$
$$r_2 = w_3 \cdot f(x_3, y_3) + r_3,\ \text{and}\ r_{n-1} = w_n \cdot f(x_n, y_n) + r_n$$

Generalized Equation for 2 variable interaction with output (CHL a)

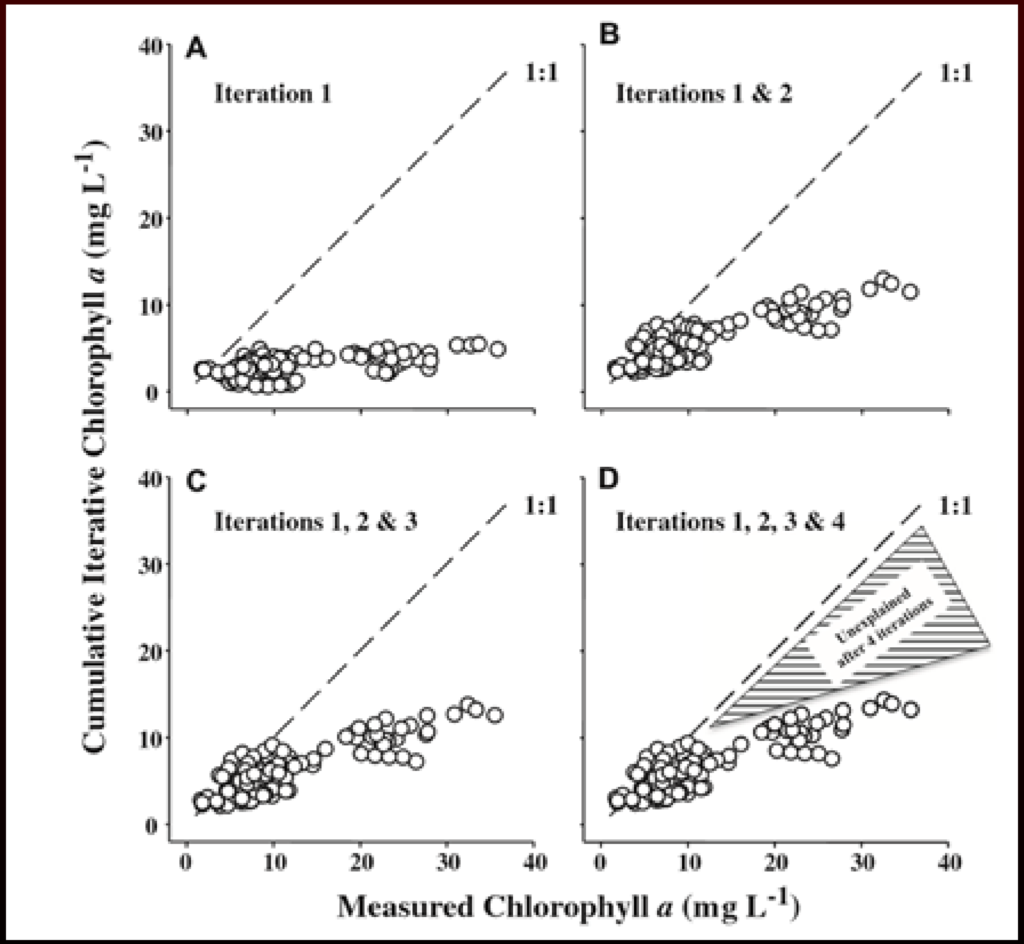


Iterations : ANNs Models

$$[\mathrm{CHL}\ a]_{\text{Grey-Box}} = [\mathrm{CHL}\ a]_{\text{1st iteration}} + [\mathrm{CHL}\ a]_{\text{2nd iteration}\ldots} + [\mathrm{CHL}\ a]_{\text{nth iteration}} + r_n$$

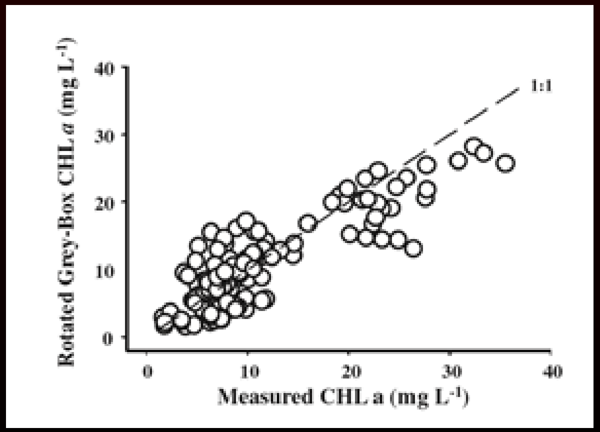
Multiple ANN models utilizing 2 variables at a time to predict Output

Iterations: Additive Models



Finalized Combined Model

$$[\mathrm{CHL}\ a_{\text{Grey-Box}} = -13.3 + 2.25\left(\begin{array}{c} (0.41*1.16 + 5.0\mathrm{E}17*(\mathrm{TEMP}^{6.592}*\mathrm{SAL}^{-17.01})) \\ +\left(0.35*\dfrac{(0.72 - 0.01*\mathrm{pH} + 0.115*\mathrm{TURB})}{(1 - 1.08*\log_{10}\mathrm{pH} + e*\log_{10}\mathrm{TURB})}\right) \\ +\left(0.12*\dfrac{(70.89 + 12.69*\log_{10}\mathrm{CURSPD} - 12.37*\log_{10}\mathrm{PAR})}{(1 - 13.43*\mathrm{CURSPD} + 0.01*\mathrm{PAR})}\right) \\ +\left(0.06*\dfrac{(8.01 - 1.98*\log_{10}\mathrm{CURDIR} - 2.69*\log_{10}\mathrm{WNDSPD})}{(1 - 0.01*\mathrm{CURDIR} + 0.03*\mathrm{WNDSPD})}\right) \end{array}\right)$$

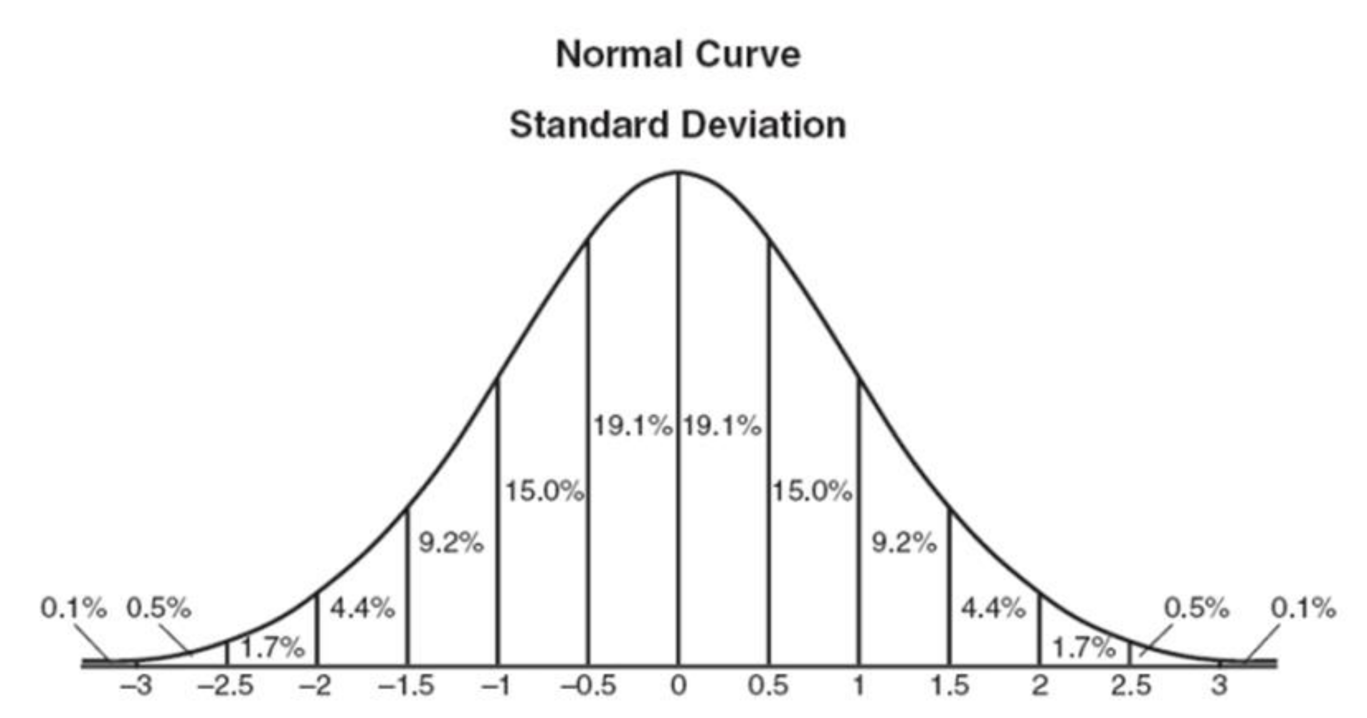


Global Sensitivity

- Sensitivity about Means
  - Local Sensitivity
  - Does not consider variable interactions as states change

- Developed Global Sensitivity
  - Looks at how variables interact as their states change!

Global Sensitivity

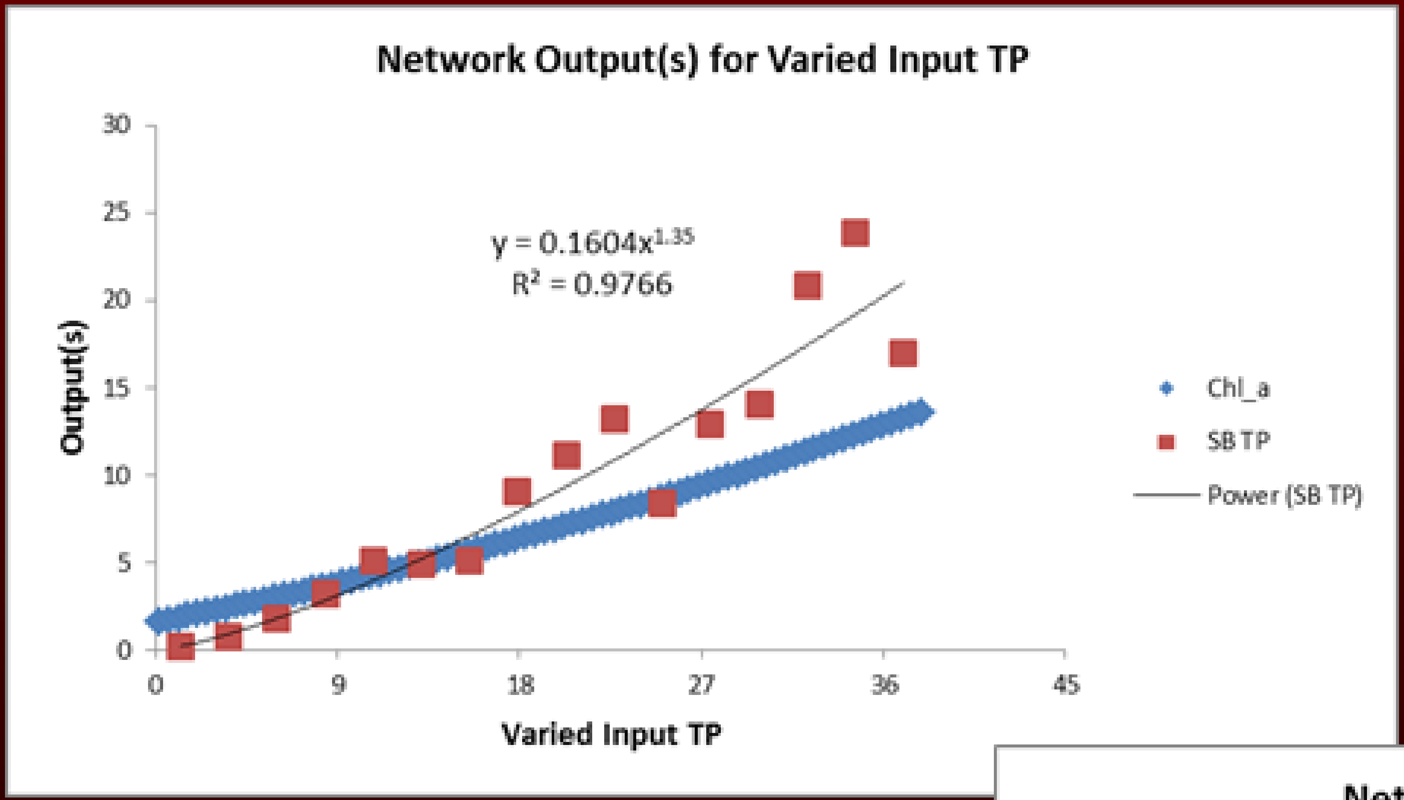


Each Variable has its own distribution of values (States)

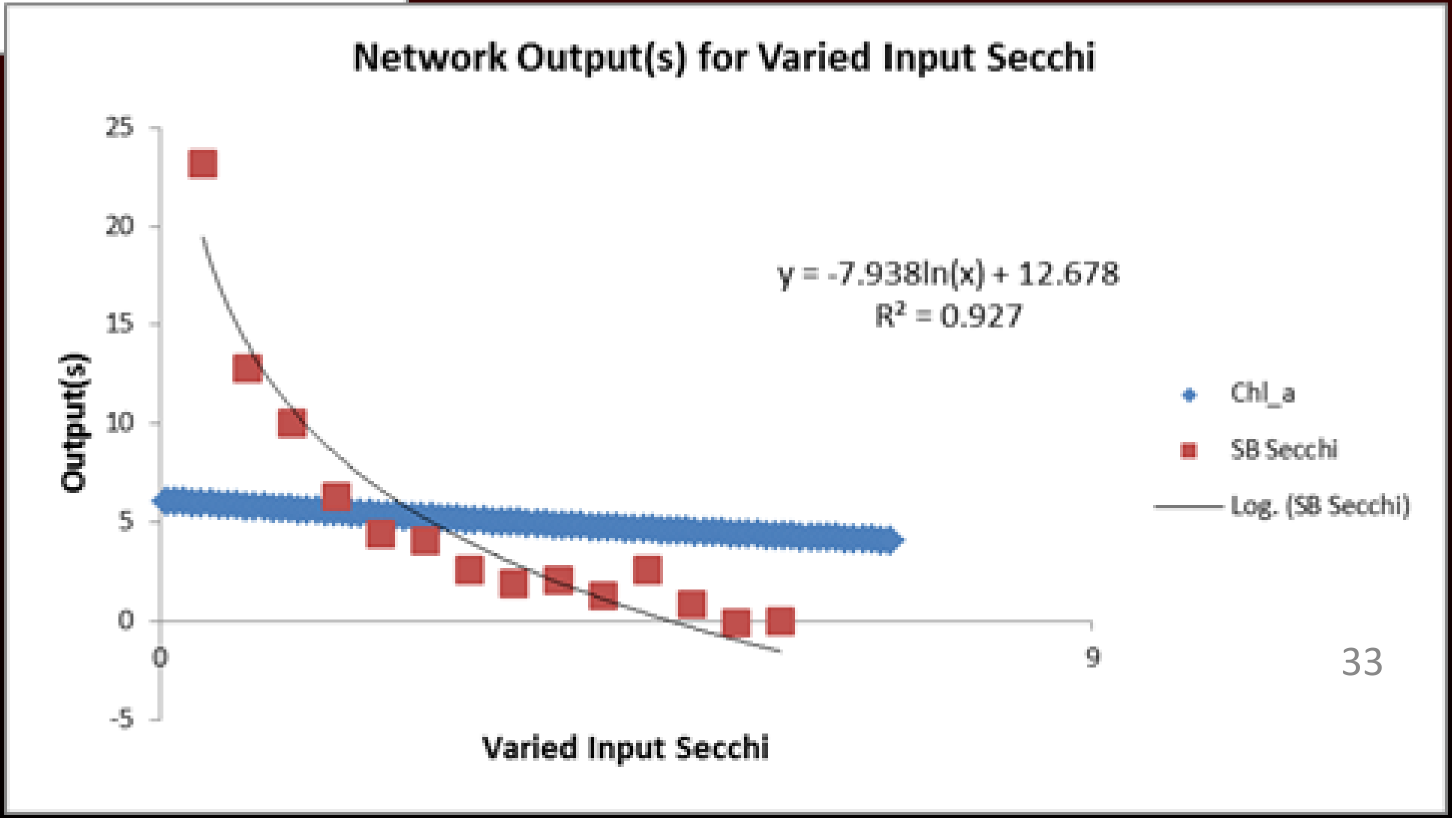
Impact of Correlation on State Behavior

PON	Secchi	TSS	TP	TDP	SRP	NH4	NO3	CL	Sol_Si	POC	DOC
-1.25 σ	1.57	-0.70	-0.98	-0.48	-0.25	0.02	-0.02	-0.57	-0.16	-1.16	-0.80
-0.75 σ	0.53	-0.67	-0.59	-0.04	-0.14	0.09	0.41	-0.02	-0.40	-0.79	0.04
-0.25 σ	-0.17	-0.08	-0.16	-0.11	-0.09	-0.09	-0.04	-0.14	-0.09	-0.26	-0.14
0.25 σ	-0.40	-0.02	0.14	0.13	0.04	-0.26	-0.24	-0.16	0.39	0.35	-0.06
0.75 σ	-0.68	0.50	0.31	-0.37	-0.06	-0.49	-0.35	-0.06	0.20	0.87	0.15
1.25 σ	-0.75	0.64	1.58	0.97	0.72	-0.08	-0.64	0.89	0.31	1.42	0.42

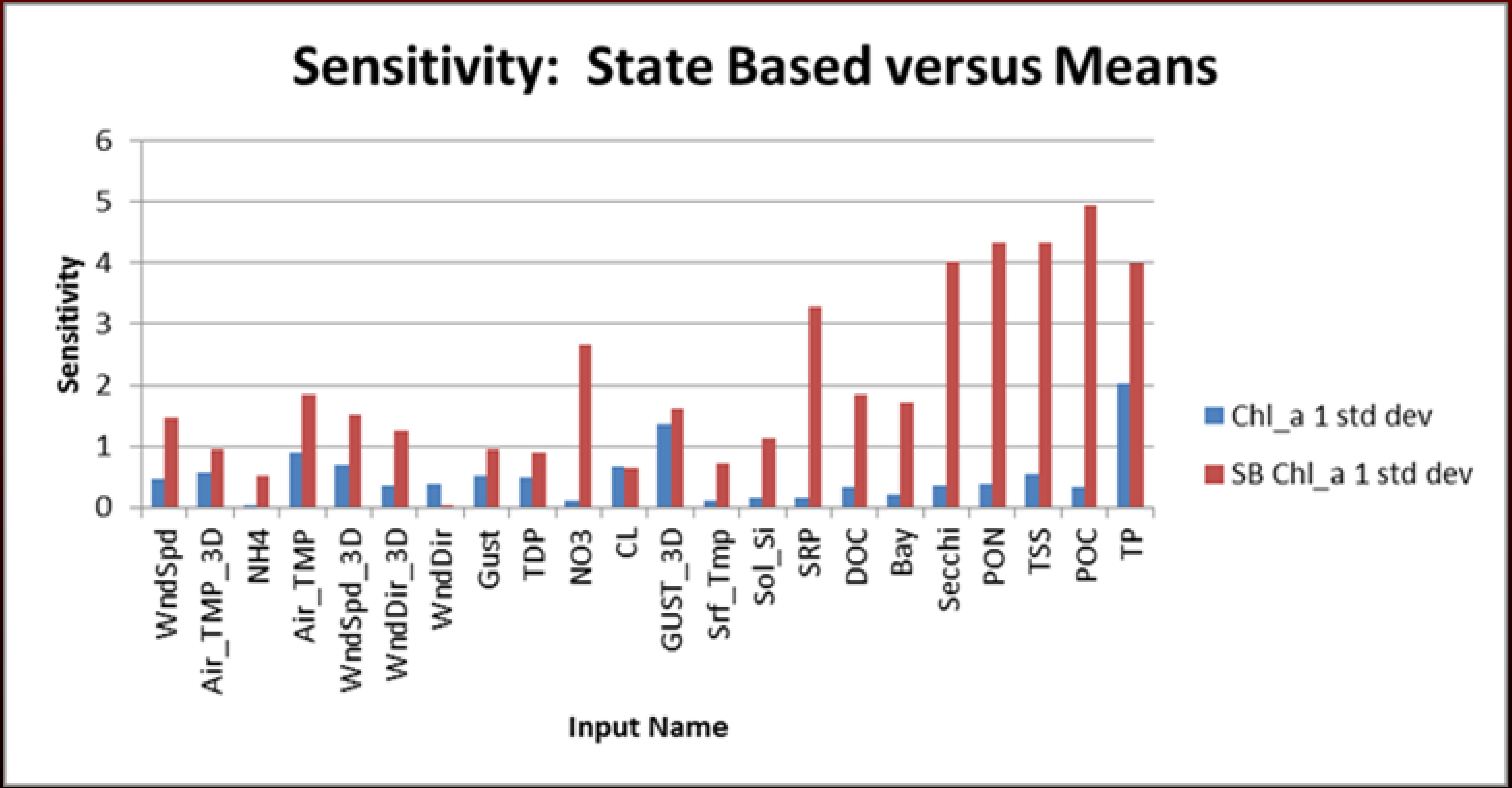
Global Variation Across States



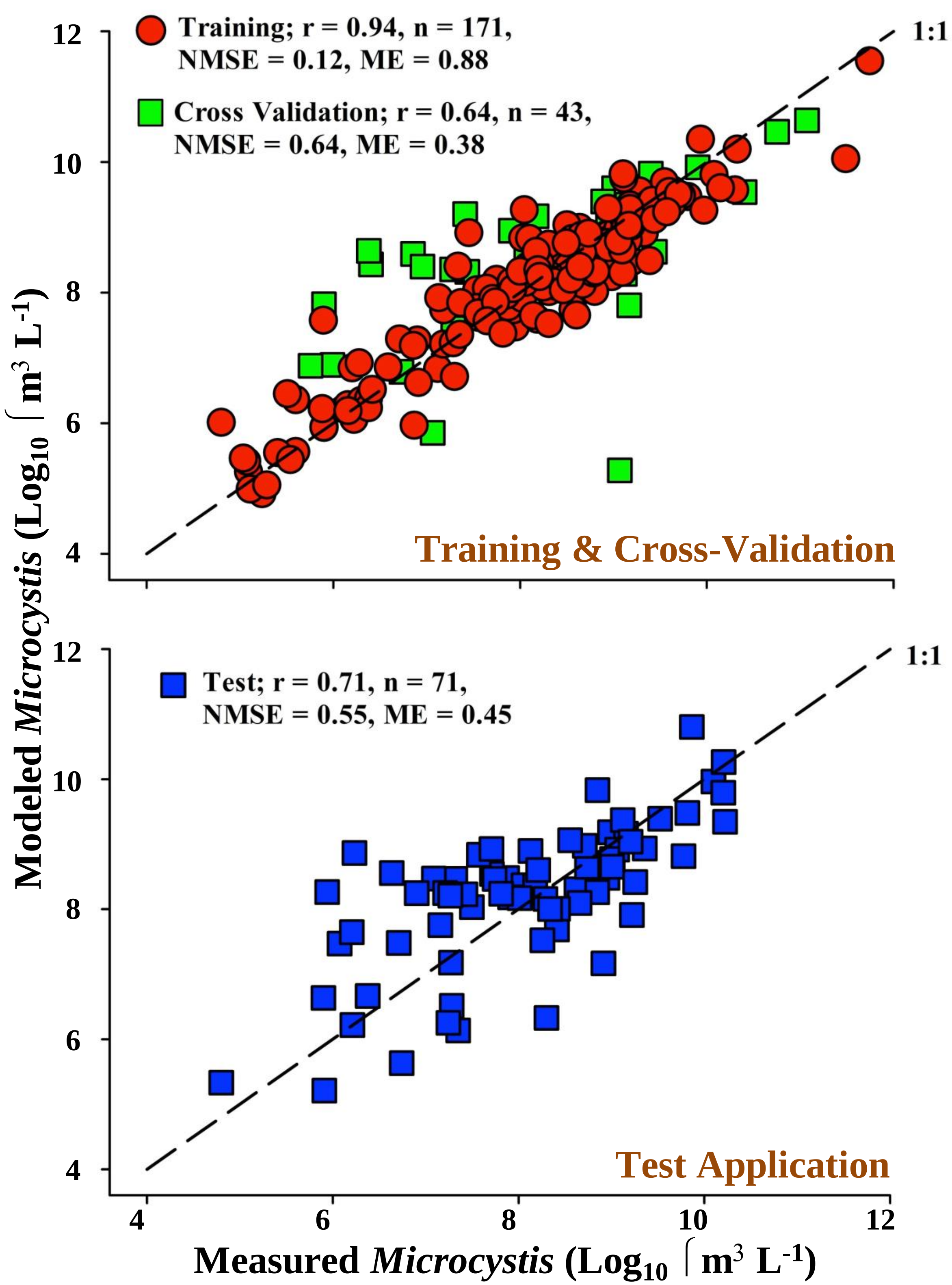
Significant difference in Global versus Local Sensitivity

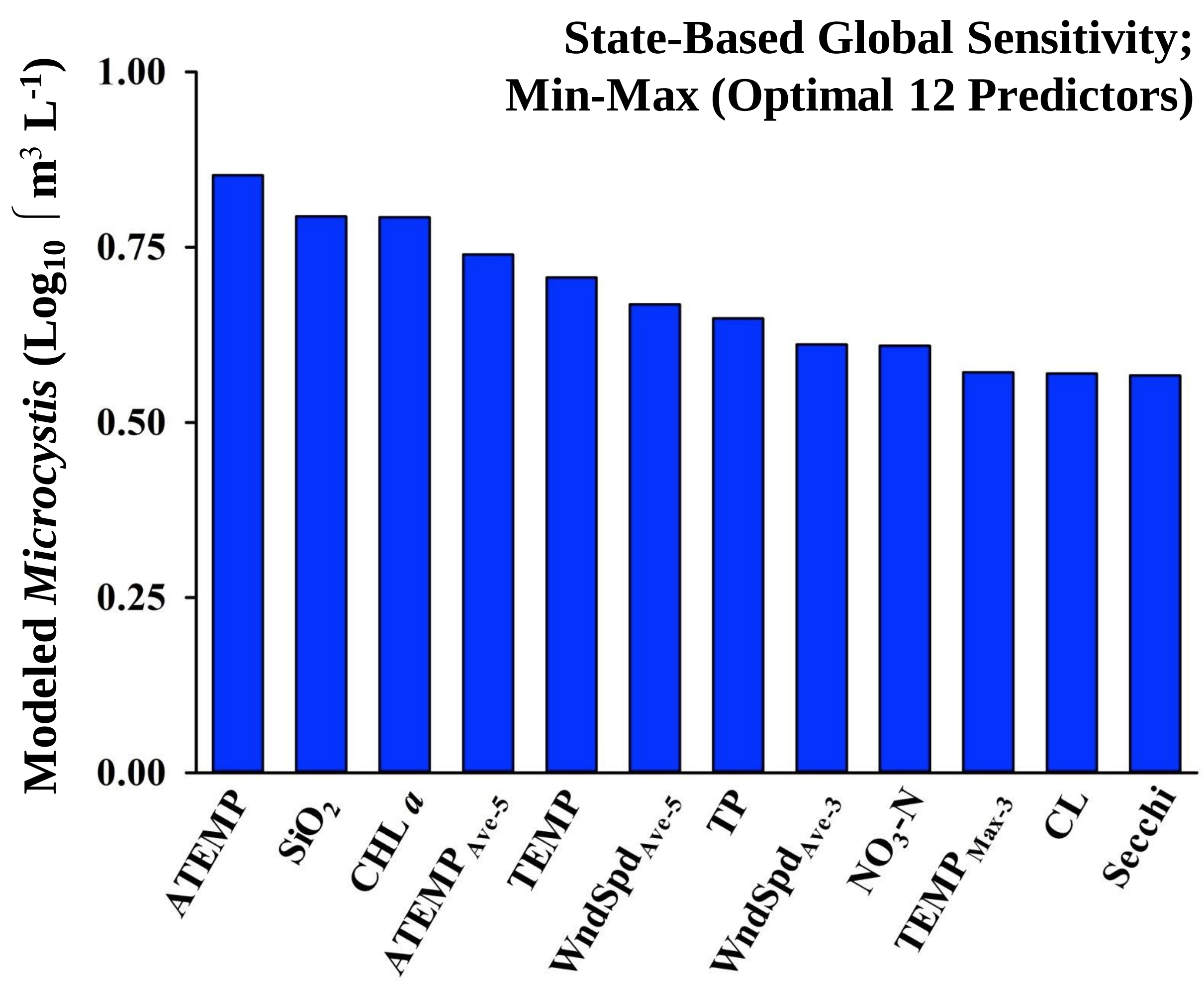


Global (State Based) versus Local (Means) Sensitivity



Lake Erie *Microcystis* (Continuous MLP); Hydrological & Meteorological
HLs: 32-15-14-10-1, TanH/Mom





Data Issues

- Big: Random reduction
- Little: Synthetic (SMOTE)
- Imbalance Data
- 0's

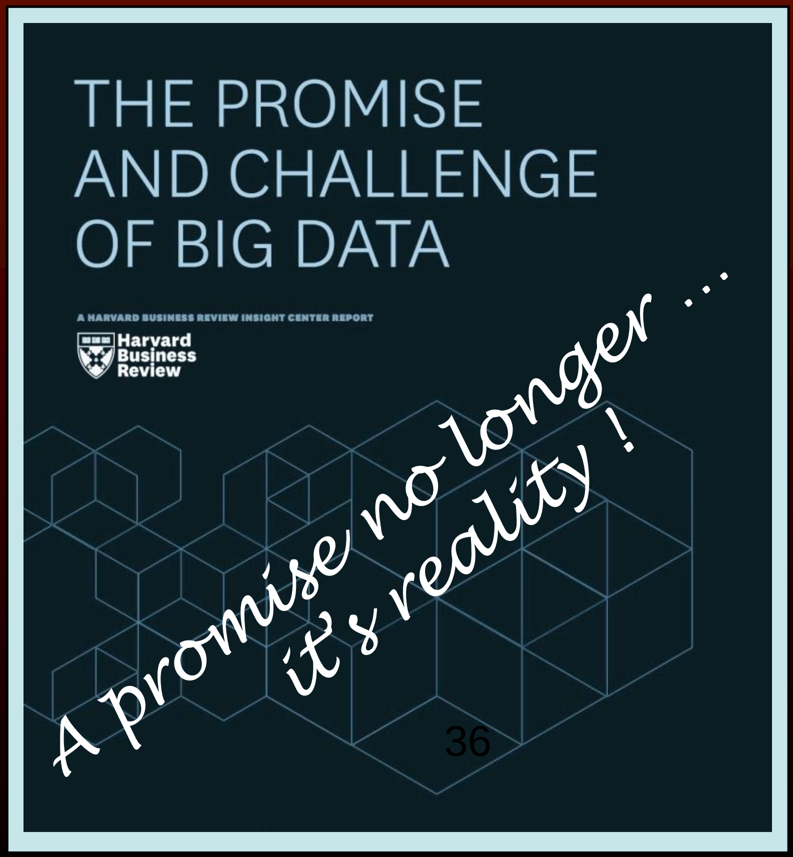
❖ ***Ecology & 'Big' Data:***

- **Not all 'Big Data' created equally:**

  volume, variety, velocity, volatility, veracity

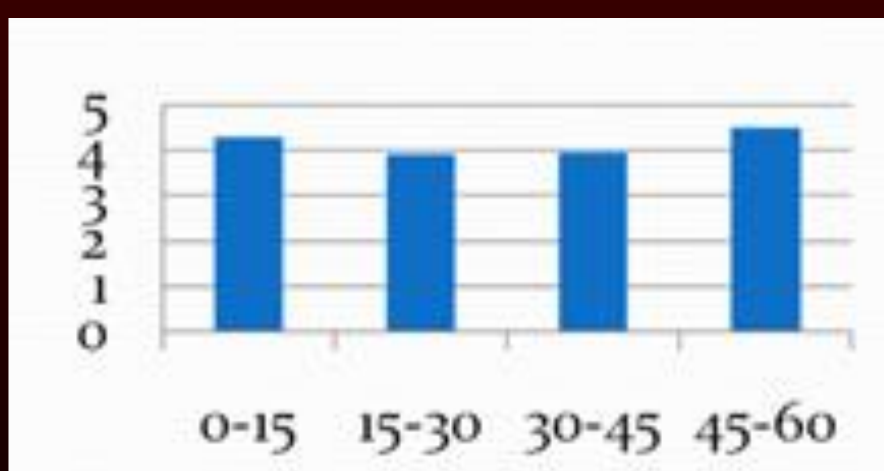
- **No longer '… *your daddy's database* …'**
- **'Big' Data = 'Big' Information = 'Big' Value**

Does 'Big' Data ensure 'Big' Science ?

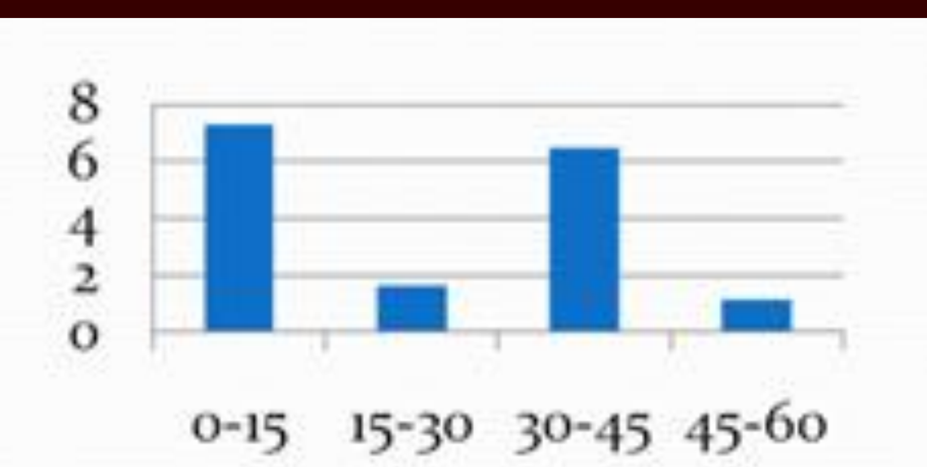


Imbalanced Datasets

- Definition: under or over representation of a class in a dataset is considered as an imbalance in a dataset.
- Ill-balanced, unbalanced, uneven

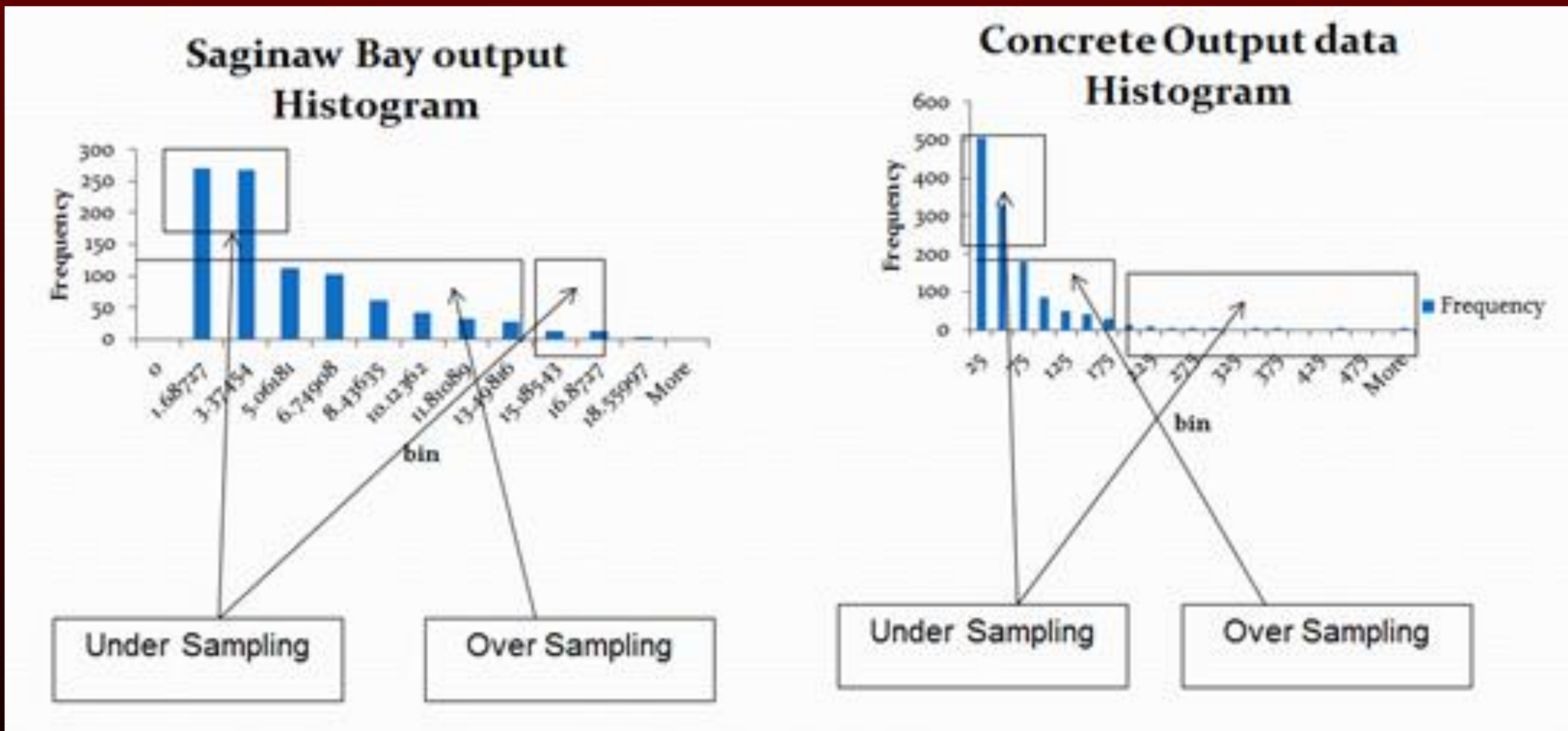


Balanced Dataset



Imbalanced Dataset

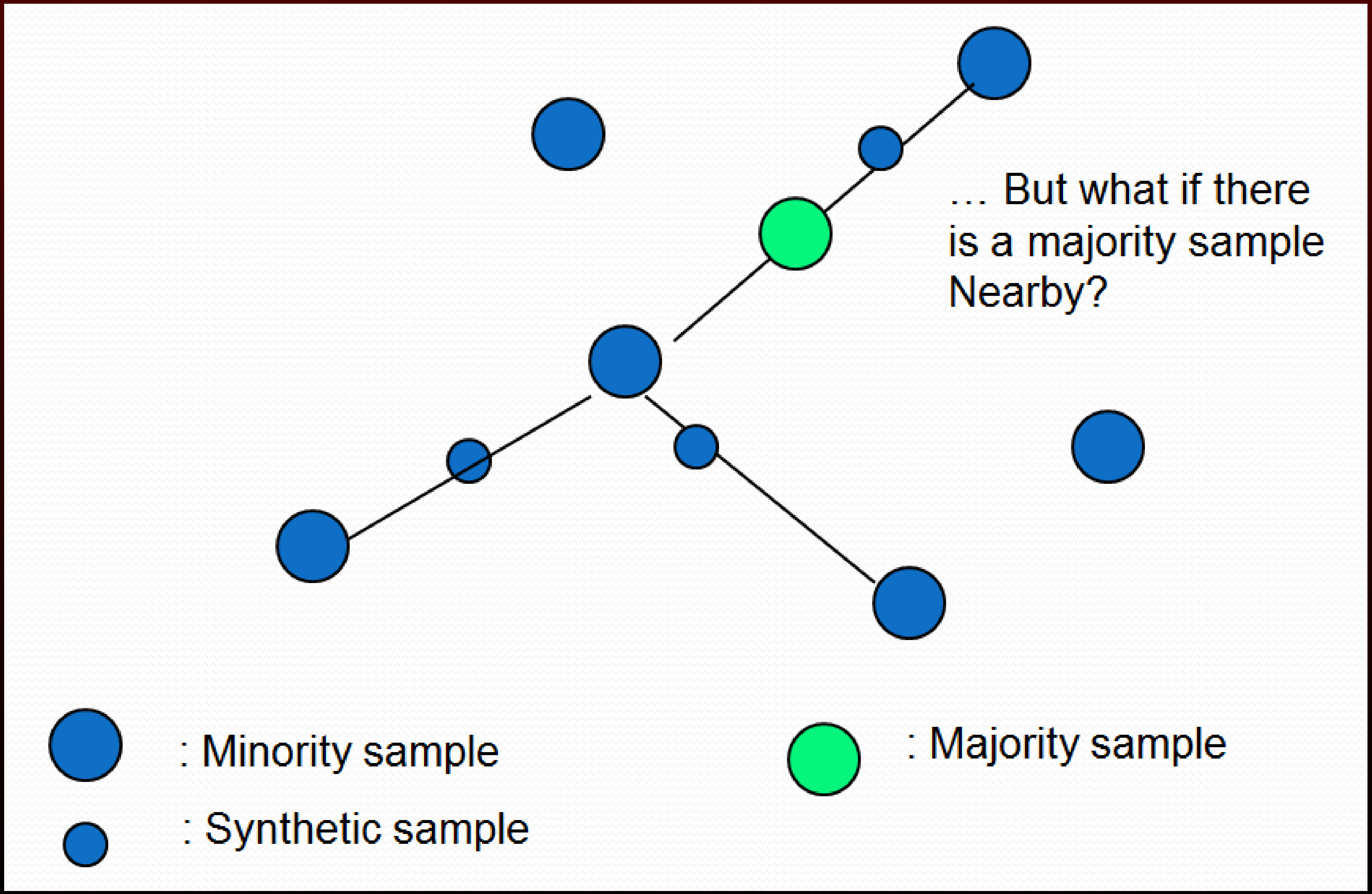
Graphic showing change under/over Sampling



Under Sampling
Over Sampling

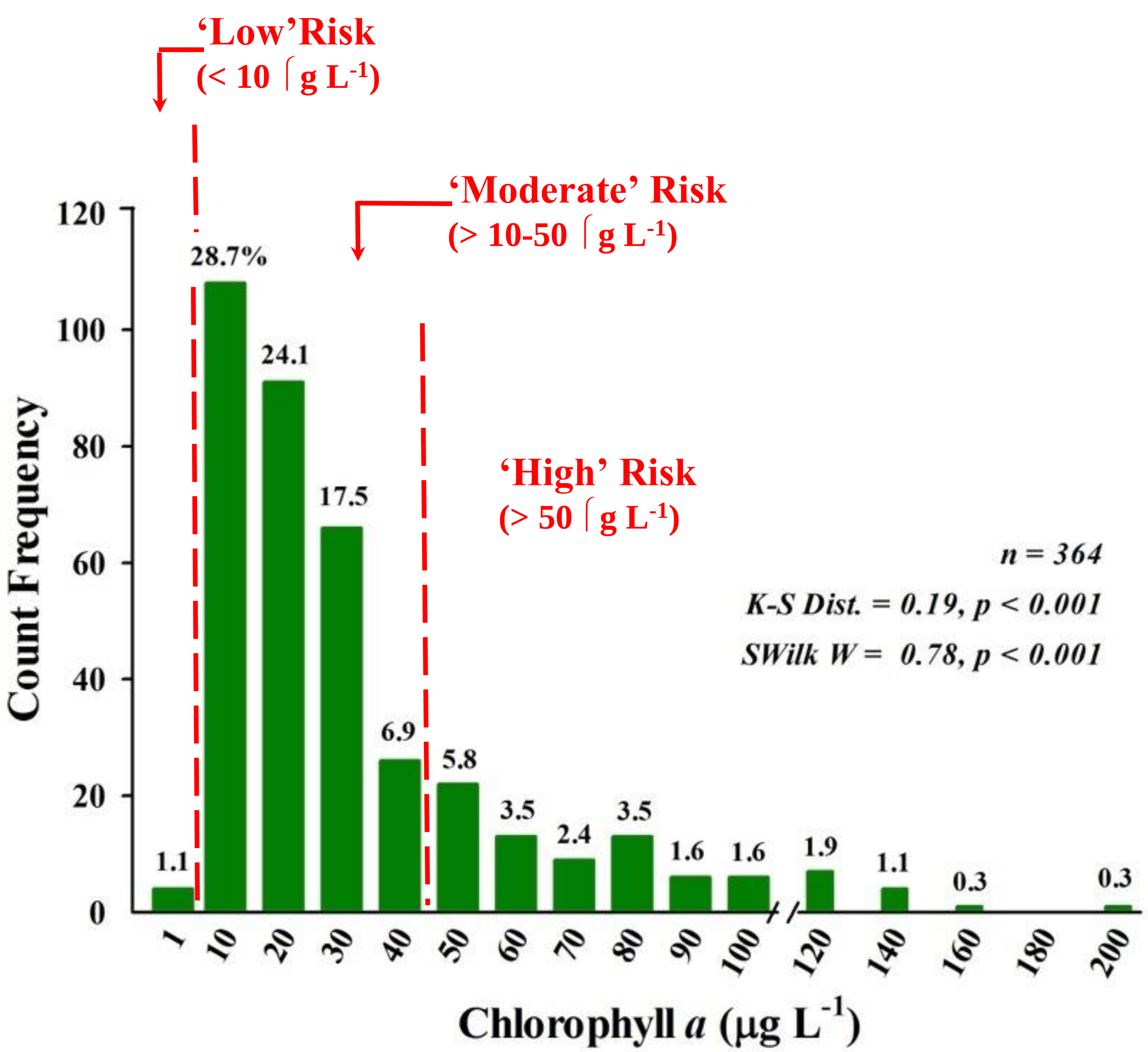
SMOTE's Informed Oversampling Procedure

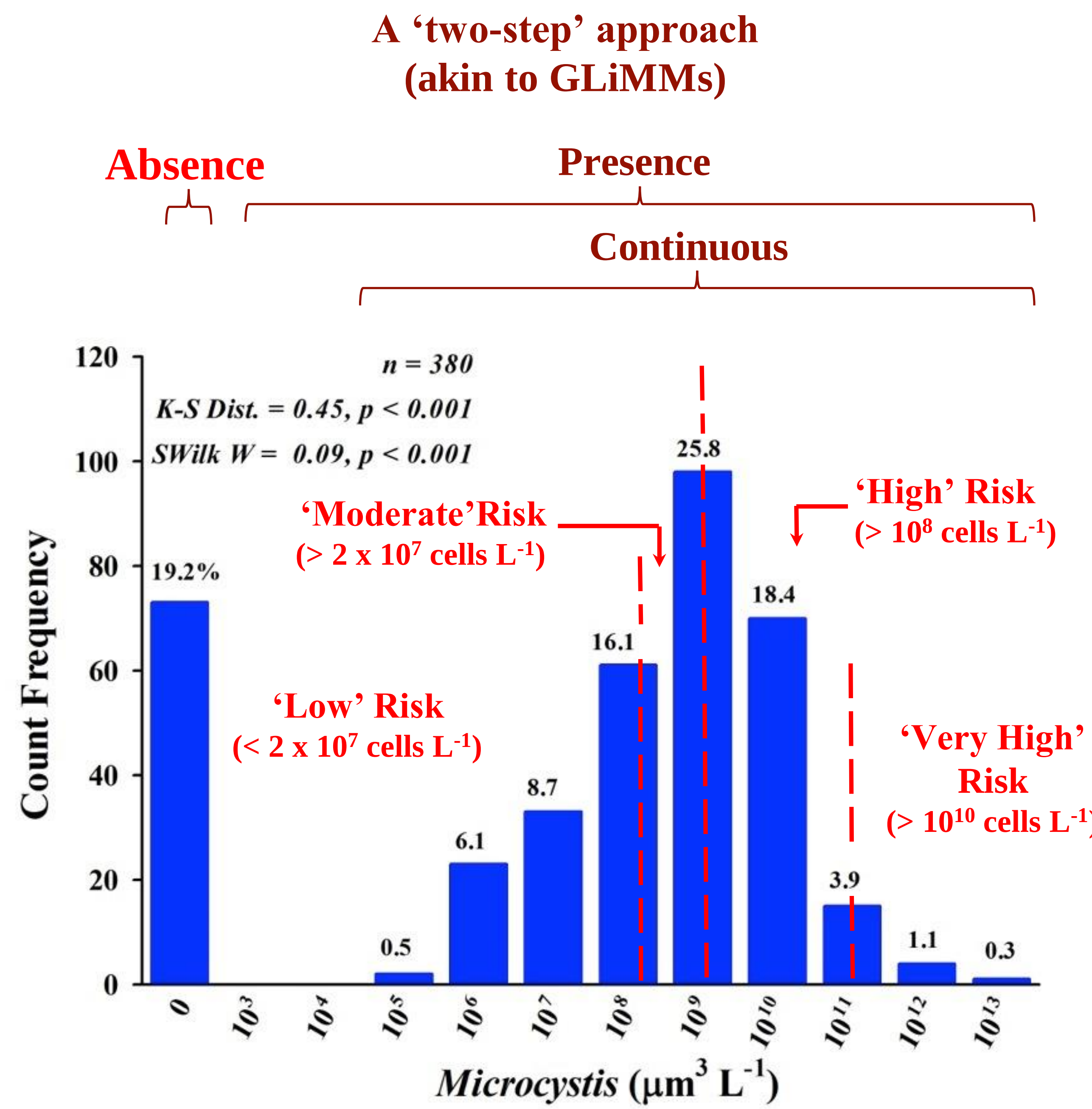
Smote: Synthetic Minority Over-sampling Technique



Lake Erie (2009-2011) Chlorophyll *a* & *Microcystis* Distributions

World Health Organization **Guidance Values for Acute Health Effects of Cyanobacteria-Dominated Waters ***





* after Chorus & Bartram 1999

Lake Erie *Microcystis (*Presence-Absence MLP); Hydrological & Meteorological HLs: 29-15-10-5-1

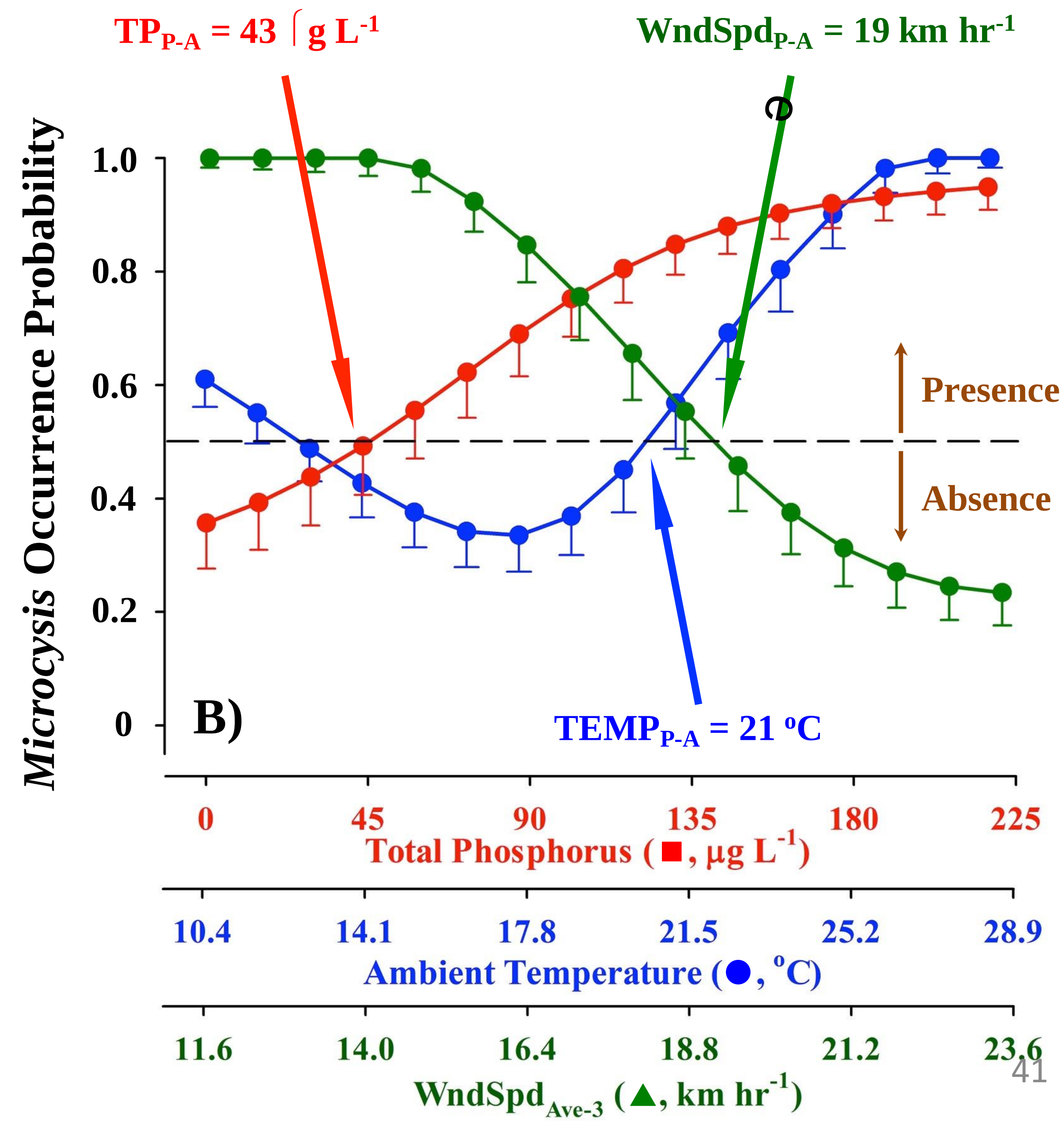
Training & Cross Validation *(class imbalance corrected via SMOTE)*	Absent	Present
Absent	130	10
Present	8	151
	Total	299

Accuracy (% correct) - 93.98
% Absent Correct – 94.20
% Present Correct – 93.79

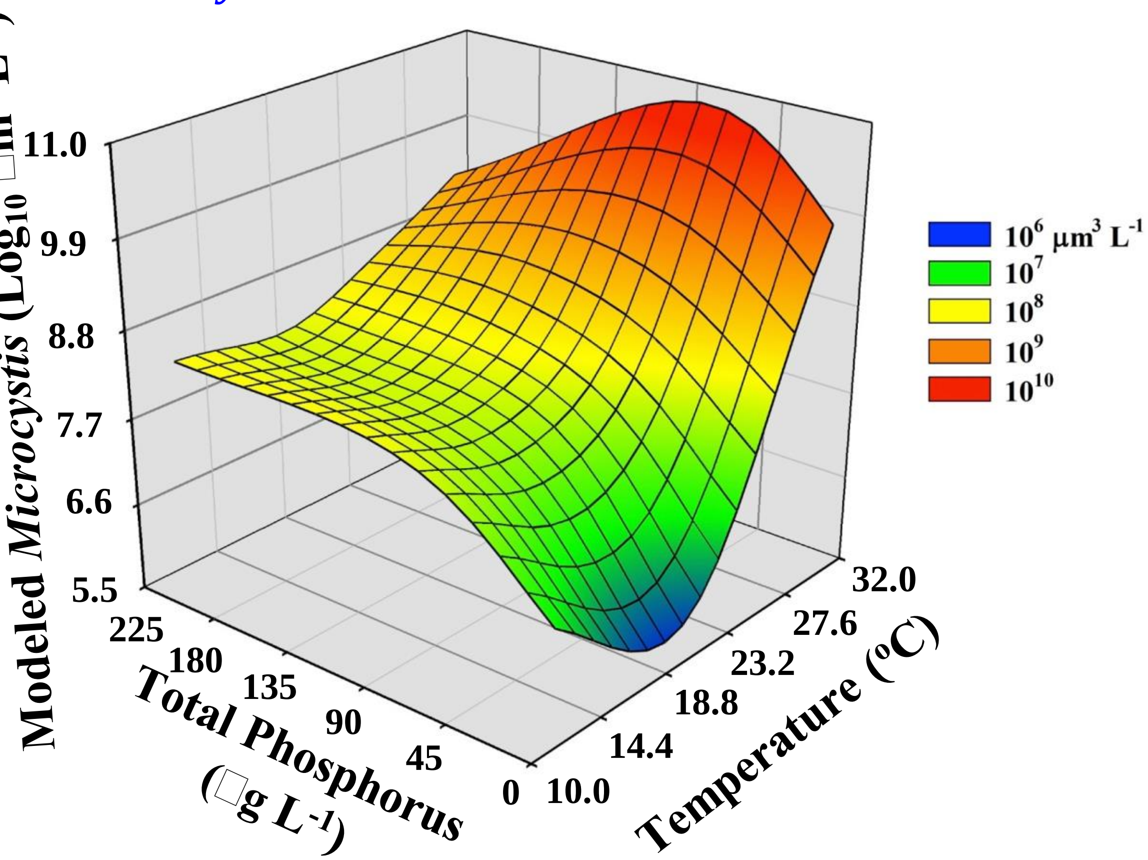
Test Application	Absent	Present
Absent	10	7
Present	4	65
	Total	86

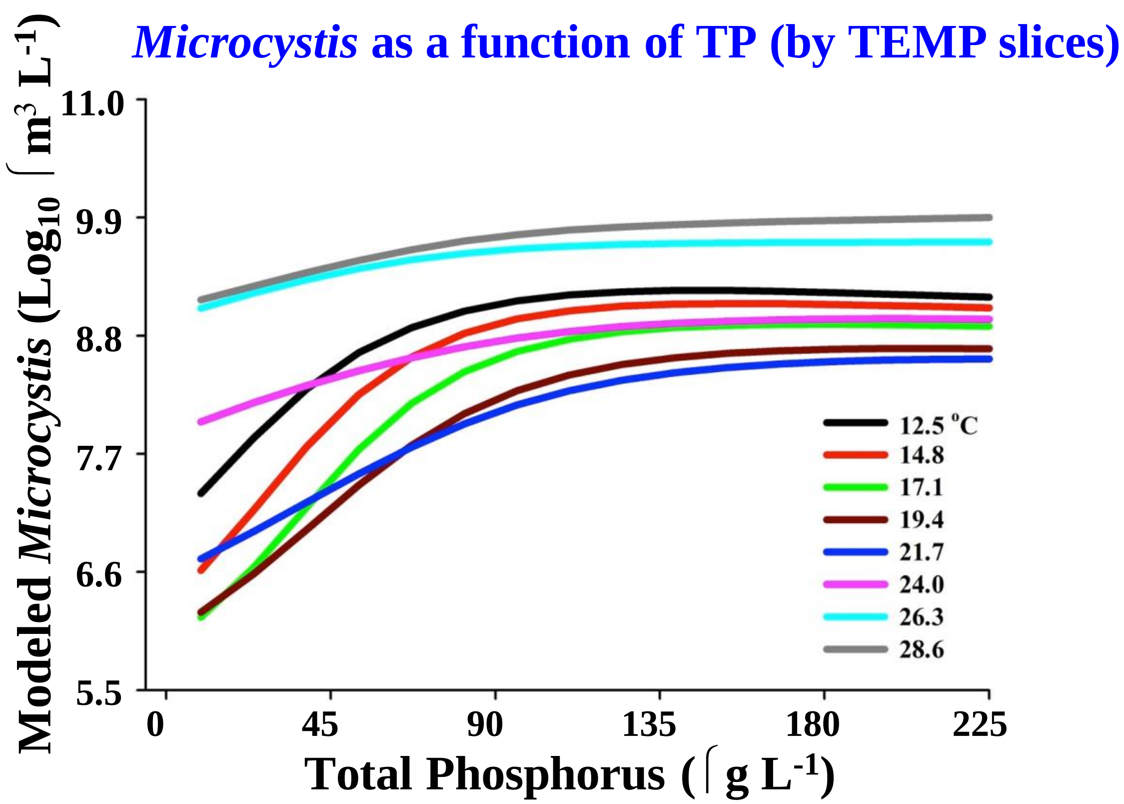
Accuracy (% correct) - 89.0
% Absent Correct – 71.43
% Present Correct – 90.23

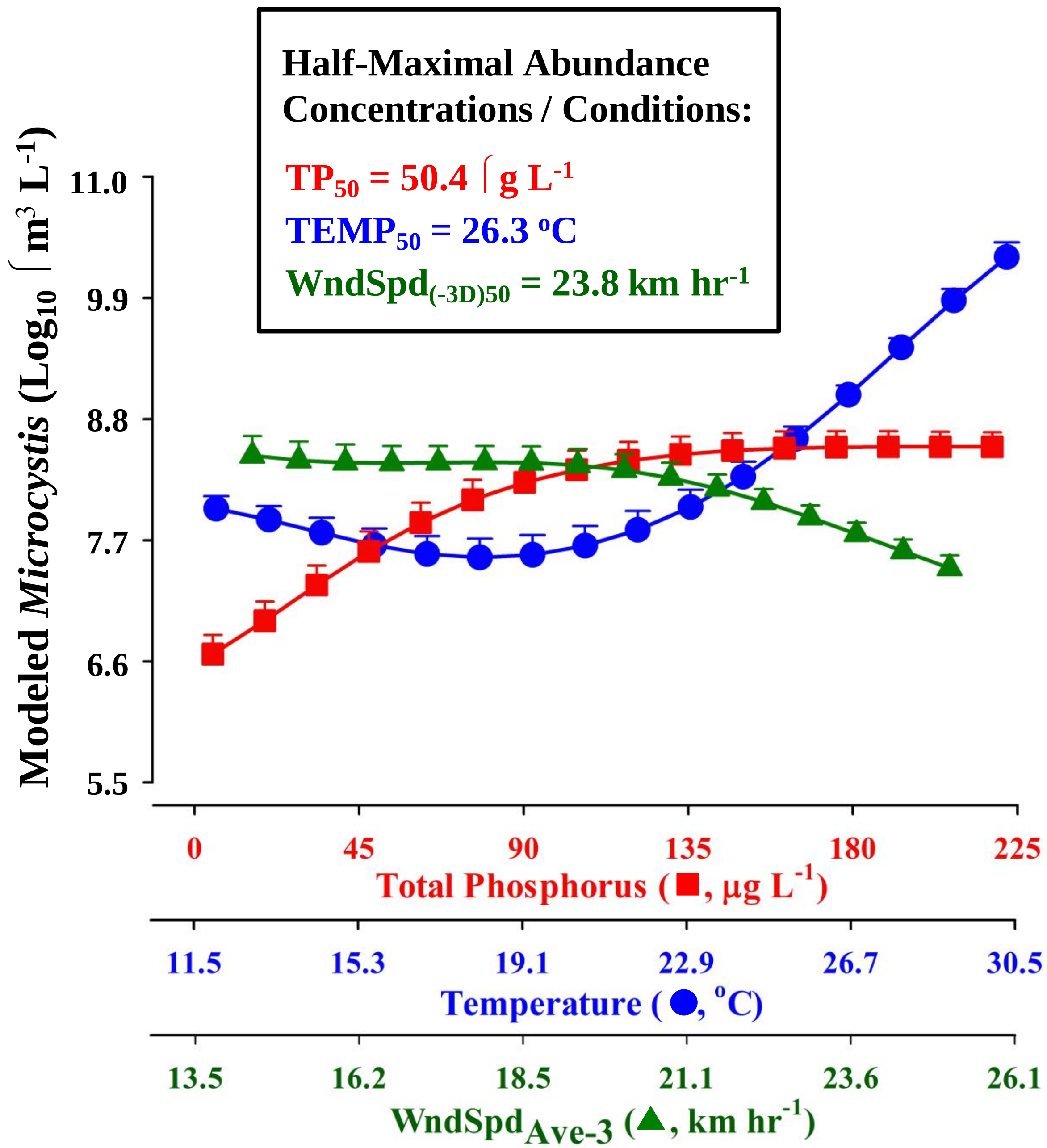
Concentrations / Conditions for Occurrence Likelihood of *Microcystis*:



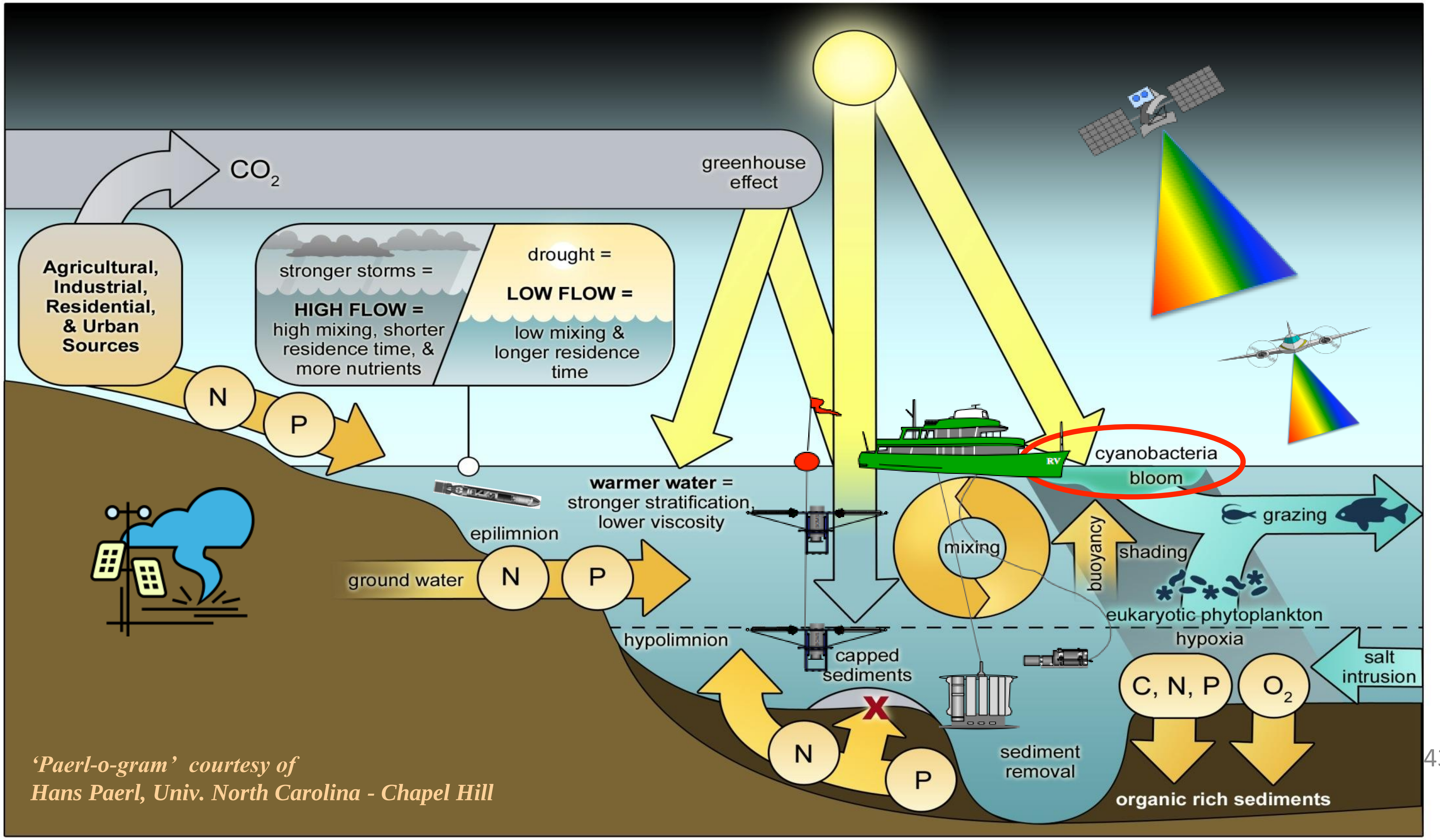
Visualizing Predictive Variances & Uncertainties for *Microcystis* (Continuous)





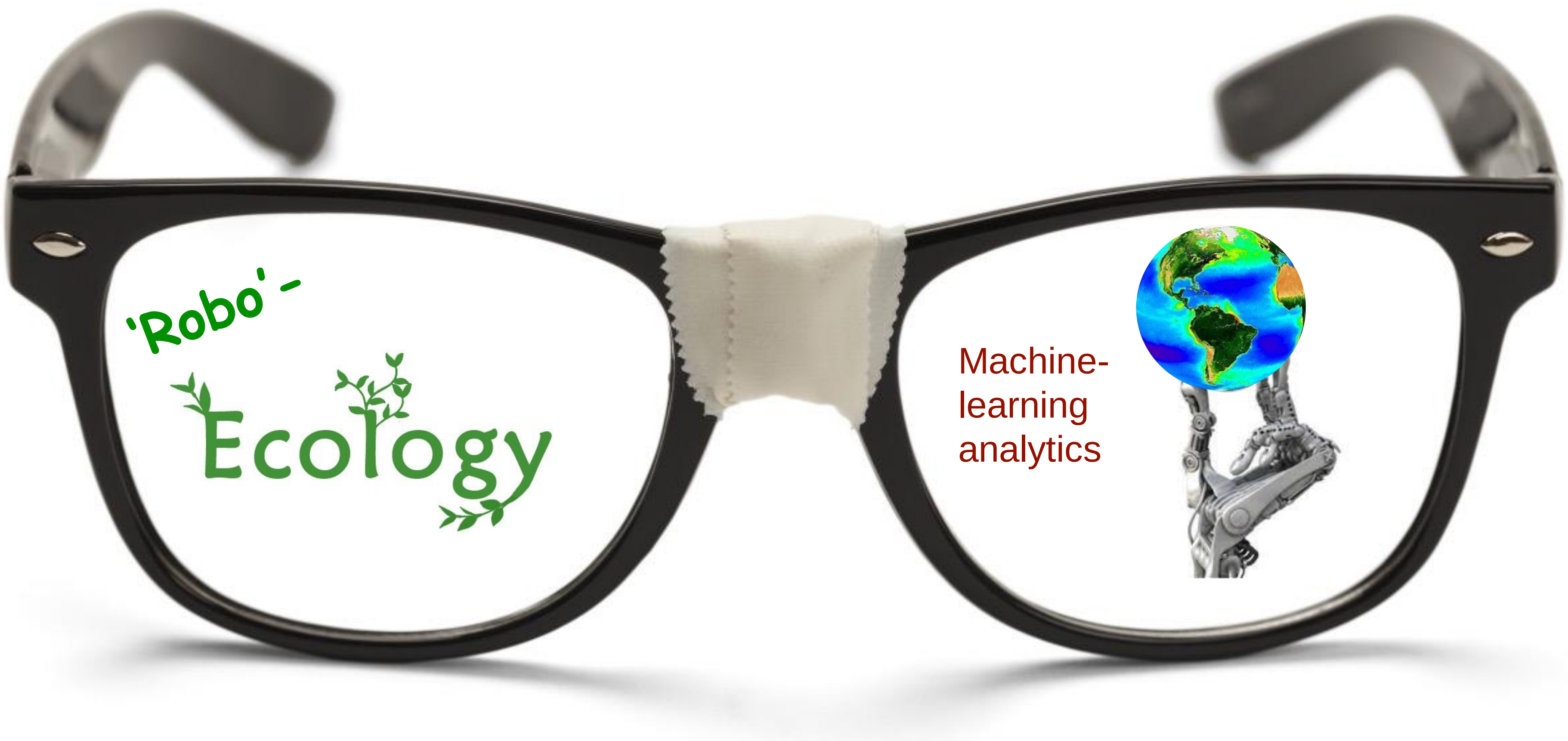


This is were we are TODAY!



'Paerl-o-gram' courtesy of
Hans Paerl, Univ. North Carolina - Chapel Hill

Still more effort to develop and investigate new ideas



Machine-learning algorithms capable of autonomously unearthing and reproducing complex patterns within sizeable data quantities afford great potential for fueling ecological hypothesis creation and 'intelligent' knowledge derivation (here, 'Robo-ecology').