

# ***Data Mining using the Enterprise Miner***

J. Michael Hardin, Ph.D.  
Professor of Statistics

# ***Where Are We Going?***

## **Outline**

- **What is Data mining?**
- **Overview of the Enterprise Miner**
- **Transformations, Outliers, Missing Values, and Variable Selection**
- **Visualization**
- **Data Mining Technologies**
  - Decision Trees
  - Regression Analysis
  - Neural Networks
  - Cluster Analysis
  - Association Analysis



# **What is Data Mining?**

# What is Data Mining?

---

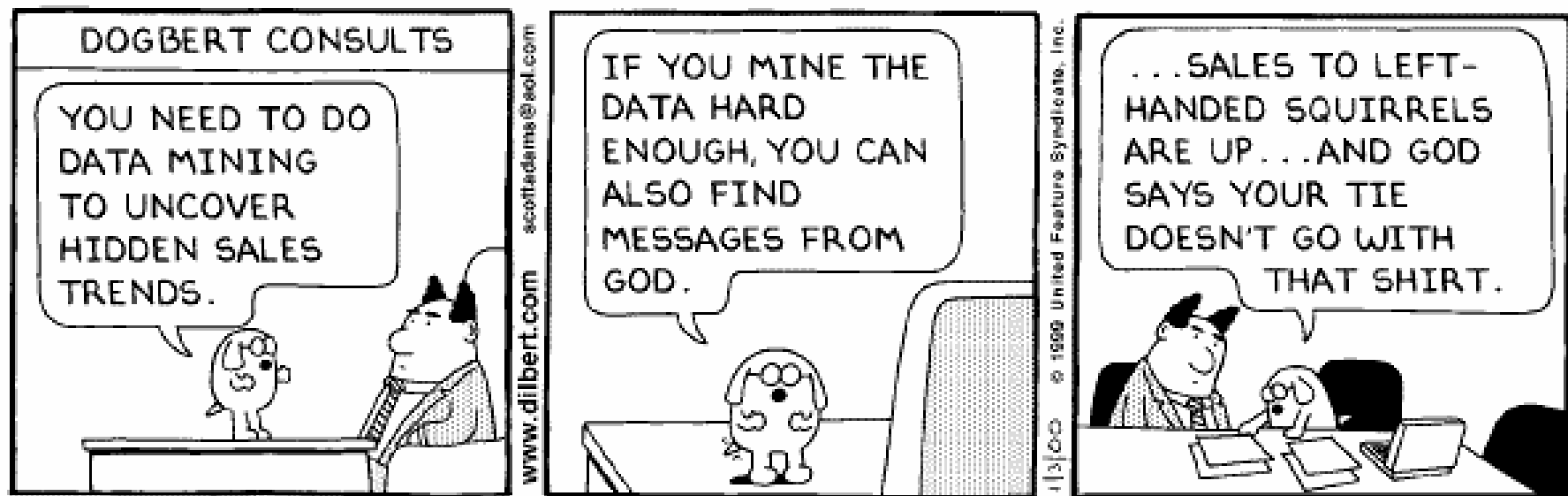
## Insights from Dilbert



Copyright © 2000 United Feature Syndicate, Inc.  
Redistribution in whole or in part prohibited

# *Further Insights form Dilbert*

---



Copyright © 2000 United Feature Syndicate, Inc.  
Redistribution in whole or in part prohibited

# *Data Mining*

---



# ***KDD Definition***

---

The non-trivial **process** of identifying **valid**, novel, potentially **useful**, and ultimately **understandable** patterns in the data

Ex. From Census Bureau data:

If Relationship=Husband then sex=male  
(prob=.996)

# *What is Data Mining?*

---

- Data Mining is the process of selecting, exploring, and modeling large amounts of data to uncover previously unknown patterns that can be exploited for business advantage
- A business process which uses a range of computer technologies to learn from the past, turning data into actionable knowledge

# *What is Data Mining?*

---

IT

Complicated database queries

ML

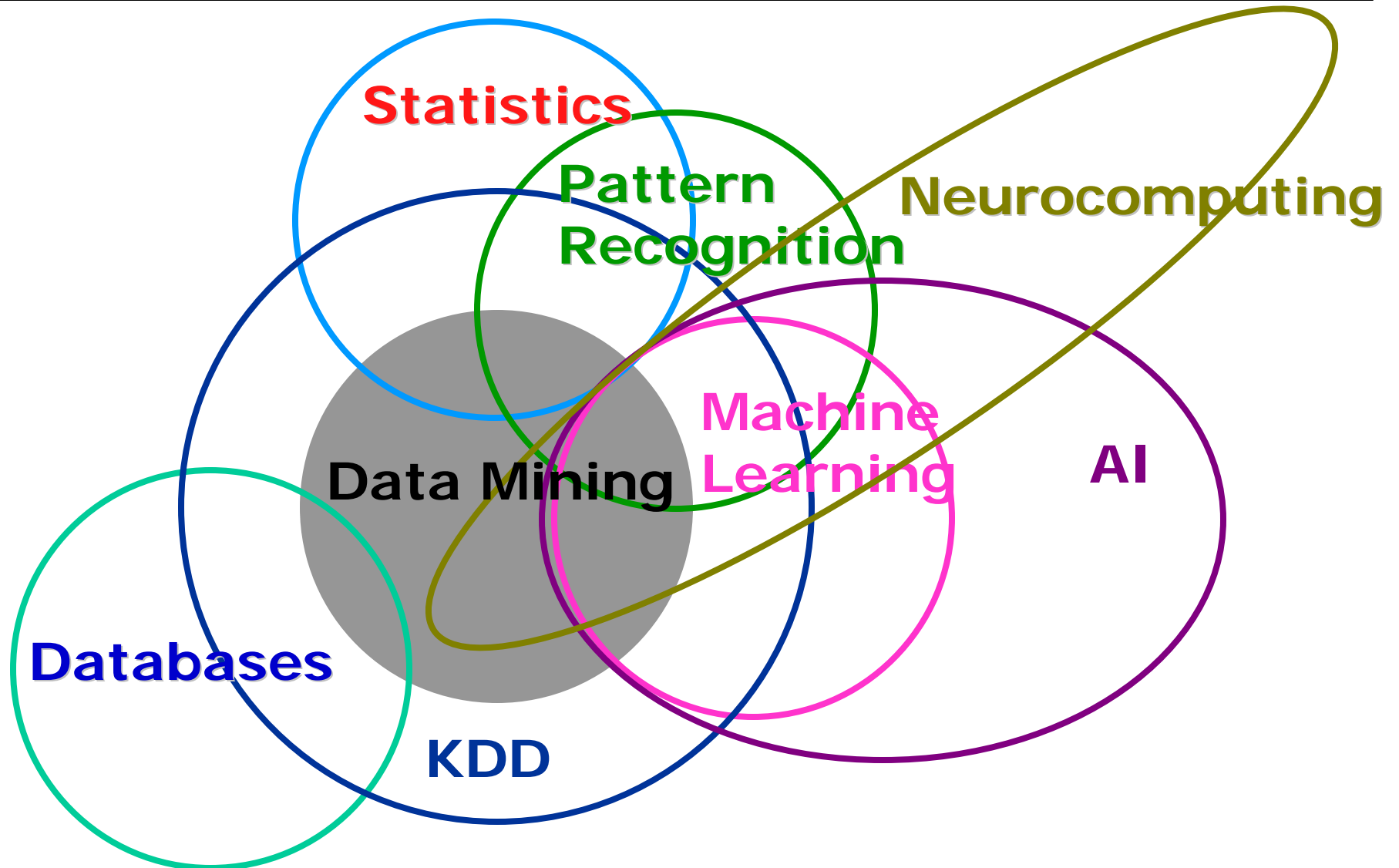
Inductive learning from examples

Stat

**What Statisticians were  
taught NOT to do!**

# *Data Mining has emerged from a Multidisciplinary Background*

---



# *Tower of Babel*

---

## "Bias"

STATISTICS: The expected difference between an estimator and what is being estimated.

NEUROCOMPUTING: The constant term in a linear combination.

MACHINE LEARNING: A reason for favoring any model that does not fit the data perfectly.



# Reference

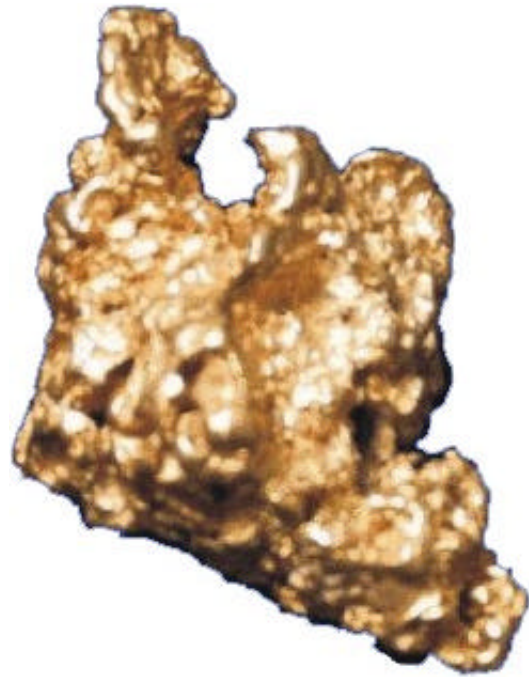
---

- **Authors:** James Myers and Edward Forgy
- **Title:** The Development of Numerical Credit Evaluation Systems
- **Publication:** *Journal of the American Statistical Association*
- **Date:** September,

1963

# *Nuggets*

---



“If you’ve got terabytes of data,  
and your relying on  
data mining to find  
interesting things  
in there for you,  
you’ve lost  
before you’ve even begun.”

— Herb Edelstein

# *Statistics and Data Mining*

---

Recent reflections on data mining and statistics:

David Hand

Jerome Friedman

Padhraic Smyth

Leo Breiman

# *Statistics and Data Mining* (cont)

---

Some key issues:

Data dredging, *fishing*, data snooping

Looking at the data, exploratory data analysis (EDA), and the scientific method

*Primary .vs. Secondary* data analysis

Large data sets, observational data, selection bias

Model selection, model uncertainty\*

# *Statistics and Data Mining* (cont)

---

Some key issues:

P-values, estimation .vs. prediction,  
classification, generalizability

“...classification error responds to error in ...probability estimates in a much different (and perhaps less intuitive) way than squared estimation error. This helps explain why improvements to the latter do not necessarily lead to improved classification performance, and why simple methods ... remain competitive, even though they usually provide poor estimates of the true probabilities (Friedman, 1997)

Single data analysis set .vs. data splitting  
(validation, test data sets) \*

Local .vs. global structure

# *Statistics and Data Mining* (cont)

---

Some key issues:

Two cultures in analysis of data:

Data modeling

Parameters are estimated

Model is validated via goodness-of-fit and residual examination

Algorithmic modeling

Construct algorithm that predicts response

Model validation by predictive accuracy

Brieman, L, (2001) "Statistical Modeling: The Two Cultures", Statistical Science, (16), 199-231.

# Overview of Data Mining/KDD Process

Creating a target set of data

Data cleaning and pre-processing

Data reduction and projection

Apply Data mining techniques

Evaluation and interpretation

Refinement of earlier steps based on evaluation and interpretation



# *Other Data Mining Process Names*

---

SEMMA (SAS)

**S**ample

**E**xplore

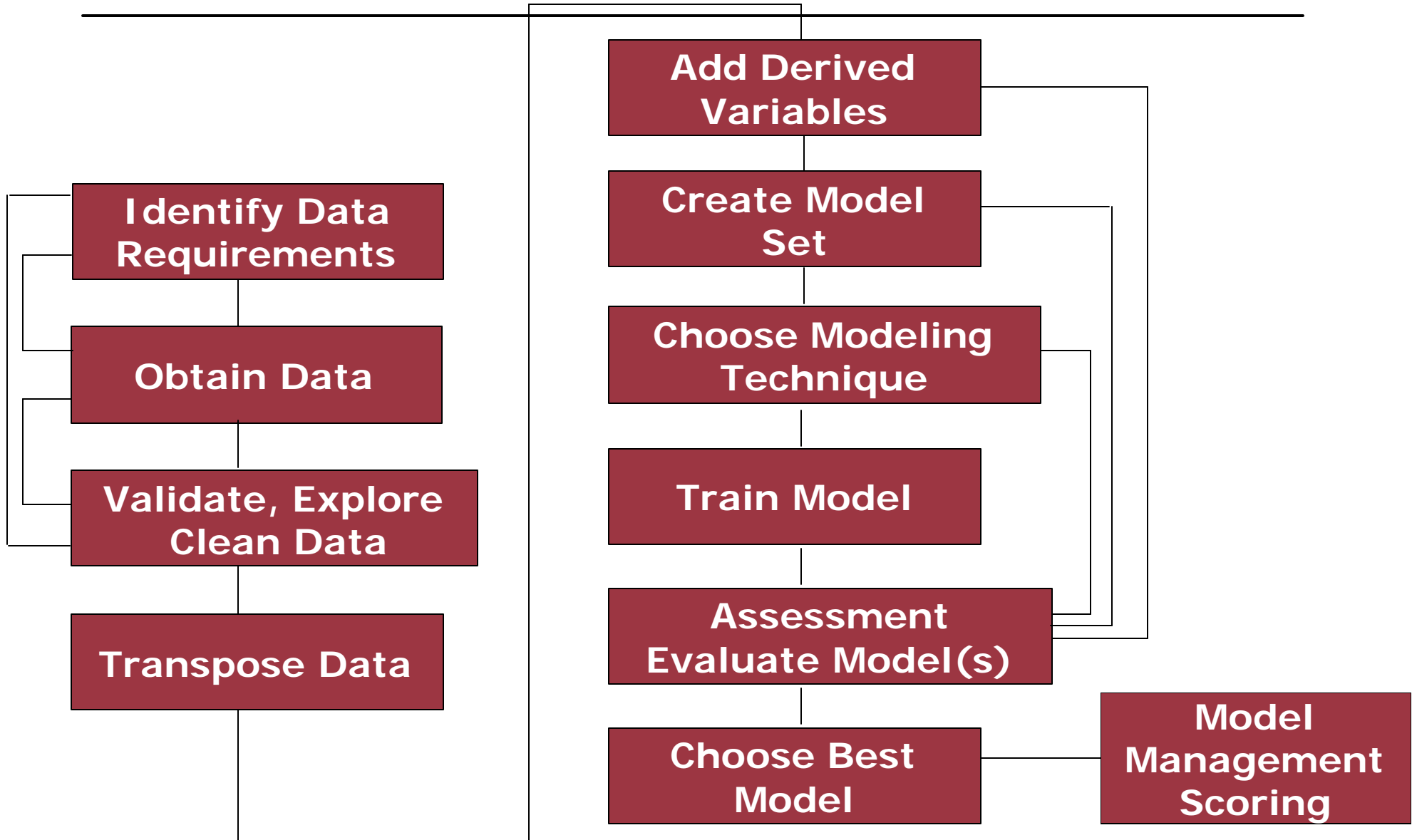
**M**odify

**M**odel

**A**ssess

CRISP-DM (**C**Ross-**I**ndustry **S**tandard  
**P**rocess for **D**ata **M**ining)

# *Data Mining Process*

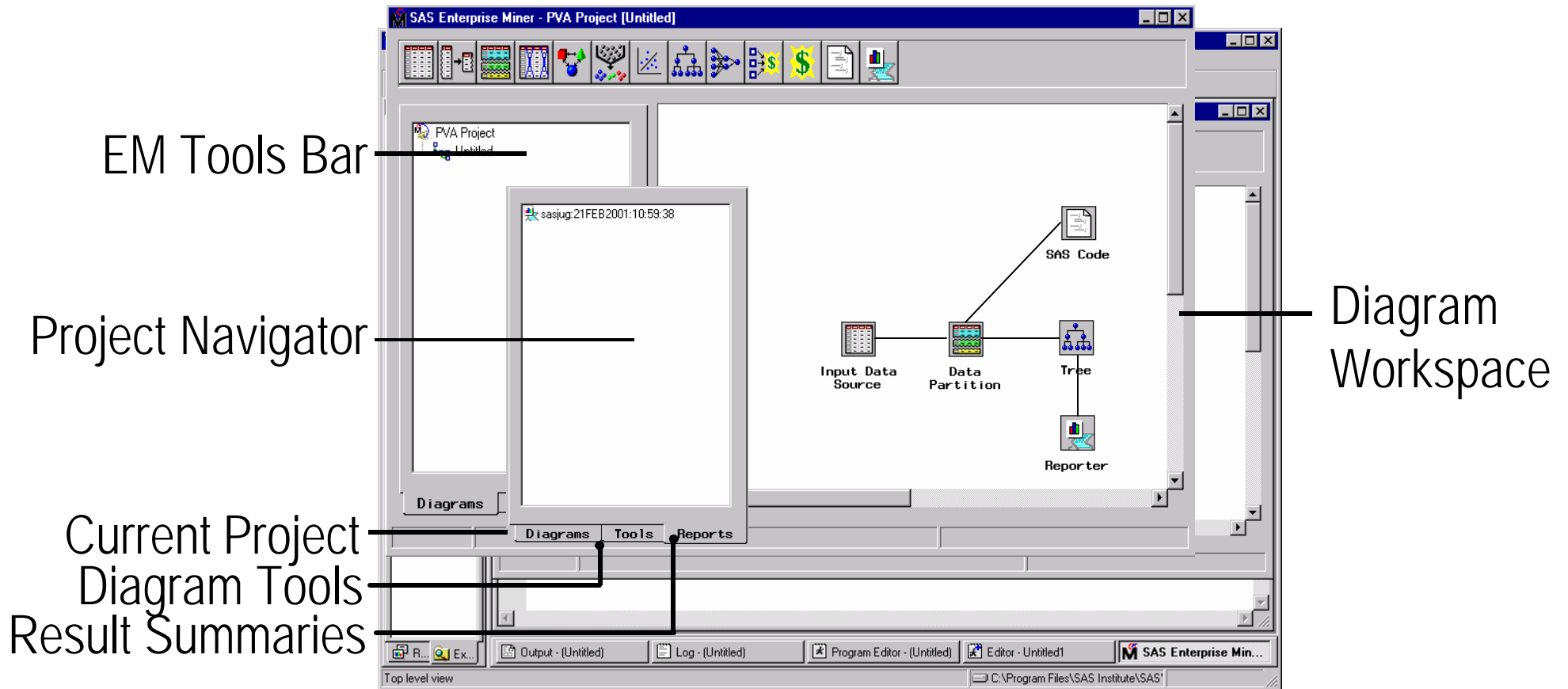




***Overview of the  
Enterprise Miner***

# Enterprise Miner Interface

---



# *Demonstration*

---

This demonstration illustrates:

Creating a client-only project

Accessing raw modeling data

Transformations

Outliers

Data replacement

Visualizations

# *Example Data Set 1 – Pima Indians Diabetes Database*

---

National Institute of Diabetes and Digestive  
Kidney Disease

Vincent Sigillito, John Hopkins

Summary:

The diagnostic, binary-valued variable investigated is whether the patient shows signs of diabetes according to World Health Organization criteria (i.e., if the 2 hour post-load plasma glucose was at least 200 mg/dl at any survey examination or if found during routine medical care). The population lives near Phoenix, Arizona, USA.

# *Example Data Set 1 – Pima Indians Diabetes Database*

---

Number of Case: 768

Number of Variables: 8 plus target variable

Variables:

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)<sup>2</sup>)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1) (*target variable*)

The image shows the exterior of the Amelia Gayle Gorgas Library. The building features a prominent portico with four tall, fluted columns supporting a pediment. The name 'AMELIA GAYLE GORGAS LIBRARY' is inscribed in capital letters across the top of the pediment. The building is surrounded by flowering trees with white blossoms. A red rectangular box is overlaid on the lower portion of the image, containing the text 'Data Mining Technologies' in a bold, italicized, red serif font.

AMELIA GAYLE GORGAS LIBRARY

***Data Mining Technologies***

# *Data Mining Technologies*

---

## Supervised Learning (Predictive Modeling)

Logistic Regression

Neural Networks

Decision Trees

## Unsupervised Learning

Cluster Analysis

Association Analysis





# *Mixed Measurement Scales*

---



sales, executive, homemaker, ...



88.60, 3.92, 34890.50, 45.01, ...



F, D, C, B, A



0, 1, 2, 3, 4, 5, 6, ...



M, F



27513, 21737, 92614, 10043, ...

# *Types of Targets*

---

## Supervised Classification

Event/no event (binary target)

Class label (multiclass problem)

## Regression

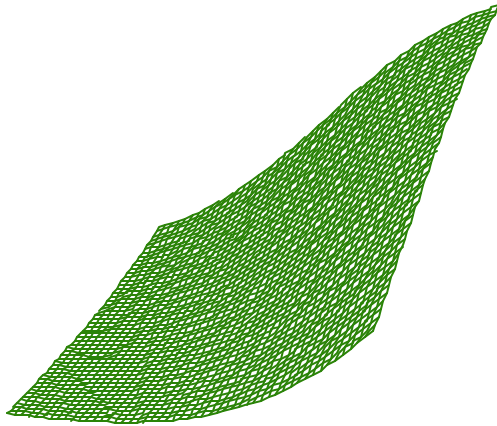
Continuous outcome

## Survival Analysis

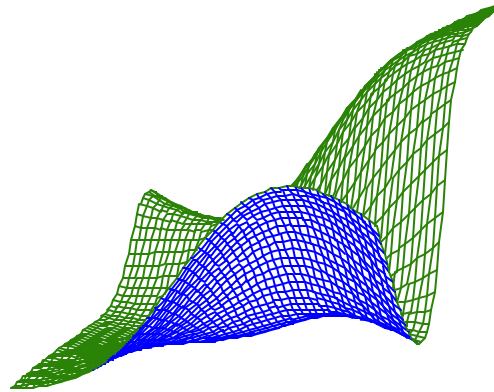
Time-to-event (possibly censored)

# *Modeling Methods*

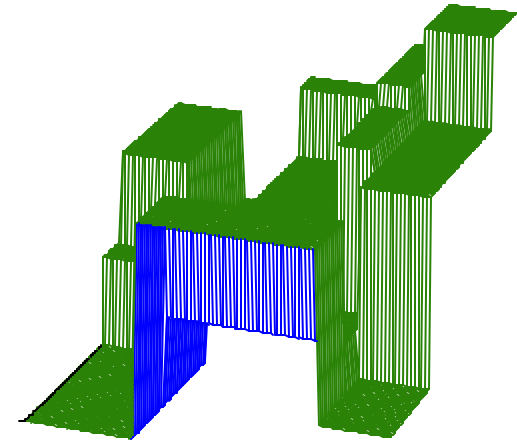
---



Generalized  
Linear Models



Neural  
Networks



Decision  
Trees

# *Logistic Regression*

---

# *Functional Form*

---

posterior probability

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}$$

parameter

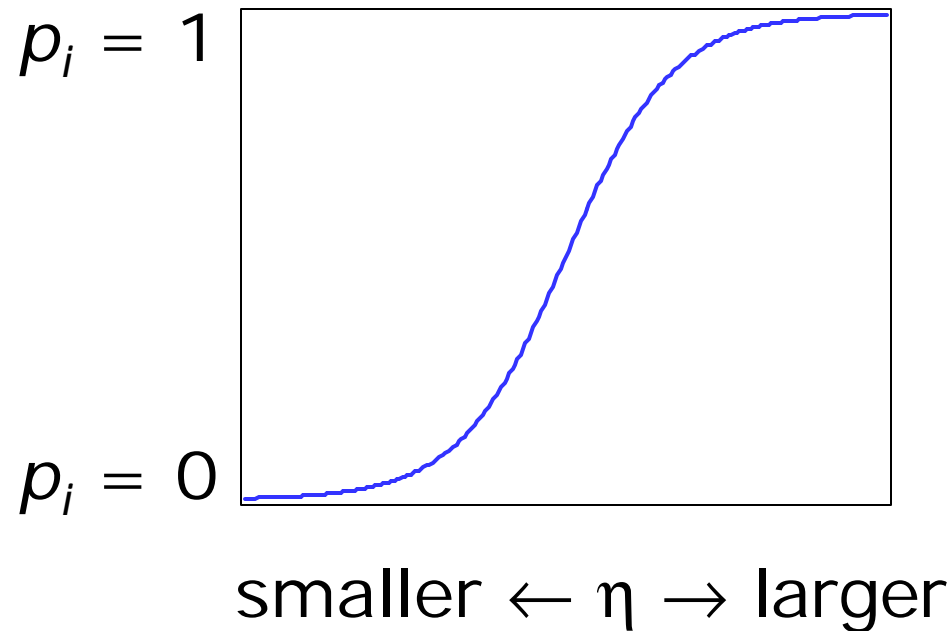
input

The diagram illustrates the functional form of a logit model. The central equation is  $\text{logit}(p_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}$ . A blue line points from the text 'posterior probability' to the  $p_i$  term in the logit function. Another blue line points from the text 'parameter' to the  $\beta_1$  coefficient. A third blue line points from the text 'input' to the  $x_{1i}$  variable.

# *The Logit Link Function*

---

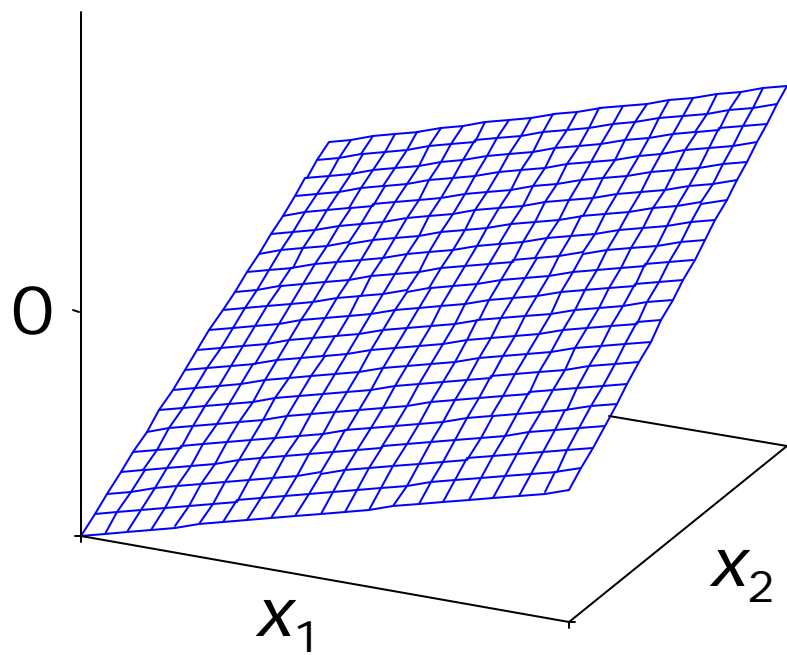
$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \eta \iff p_i = \frac{1}{1+e^{-\eta}}$$



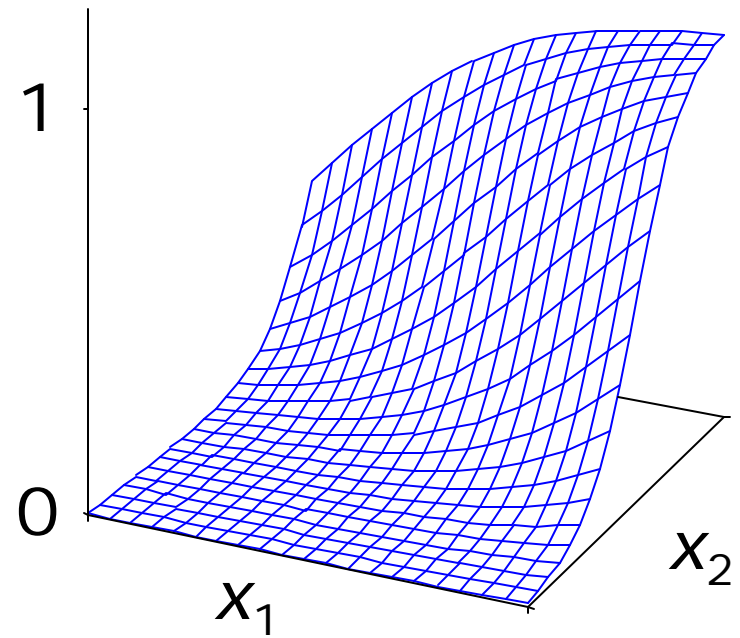
# *The Fitted Surface*

---

$\text{logit}(p)$

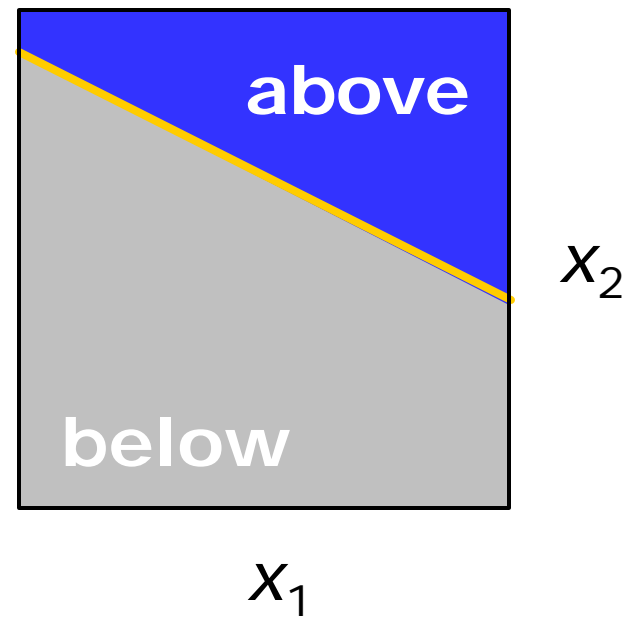
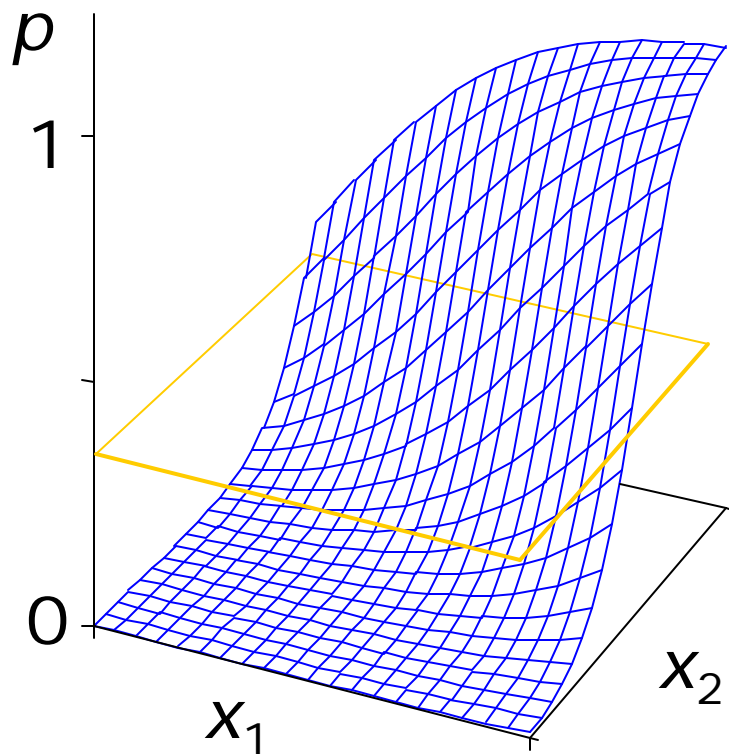


$p$



# *Logistic Discrimination*

---



# *Scoring New Cases*

---

$$\mathbf{x} = (1.1, 3.0)$$



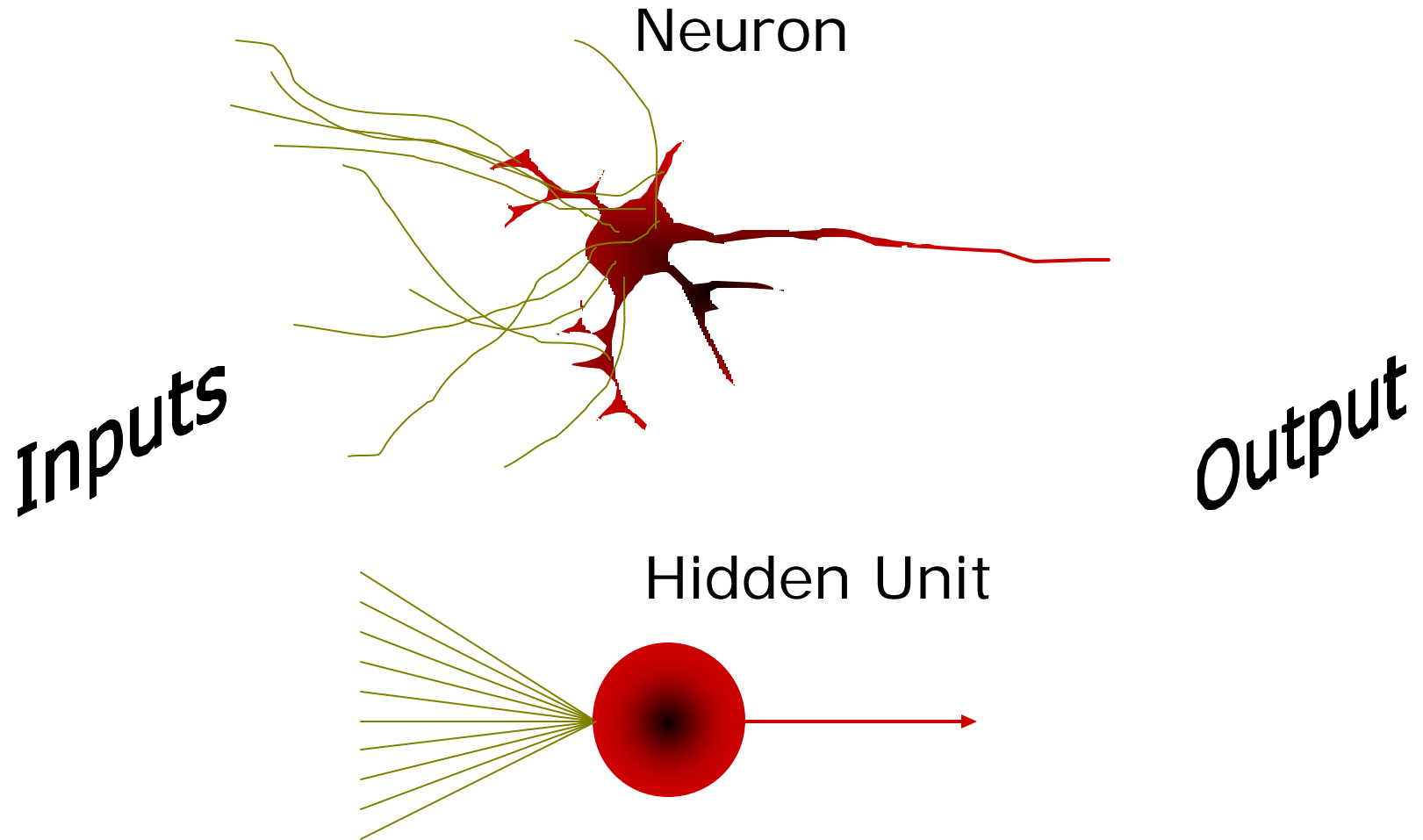
$$\hat{p} = .05$$

# *Demonstration*



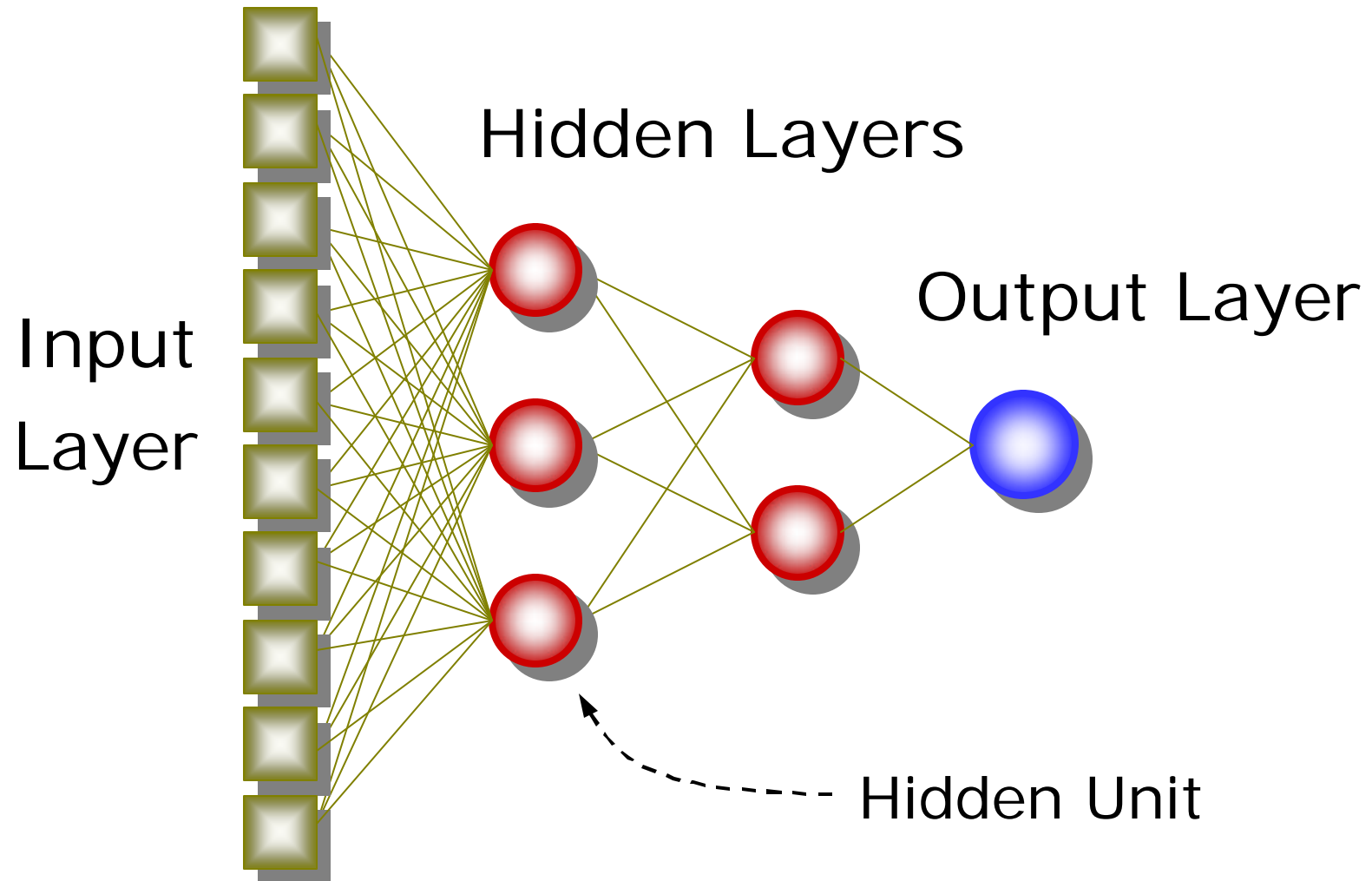
# *Artificial Neural Networks*

---



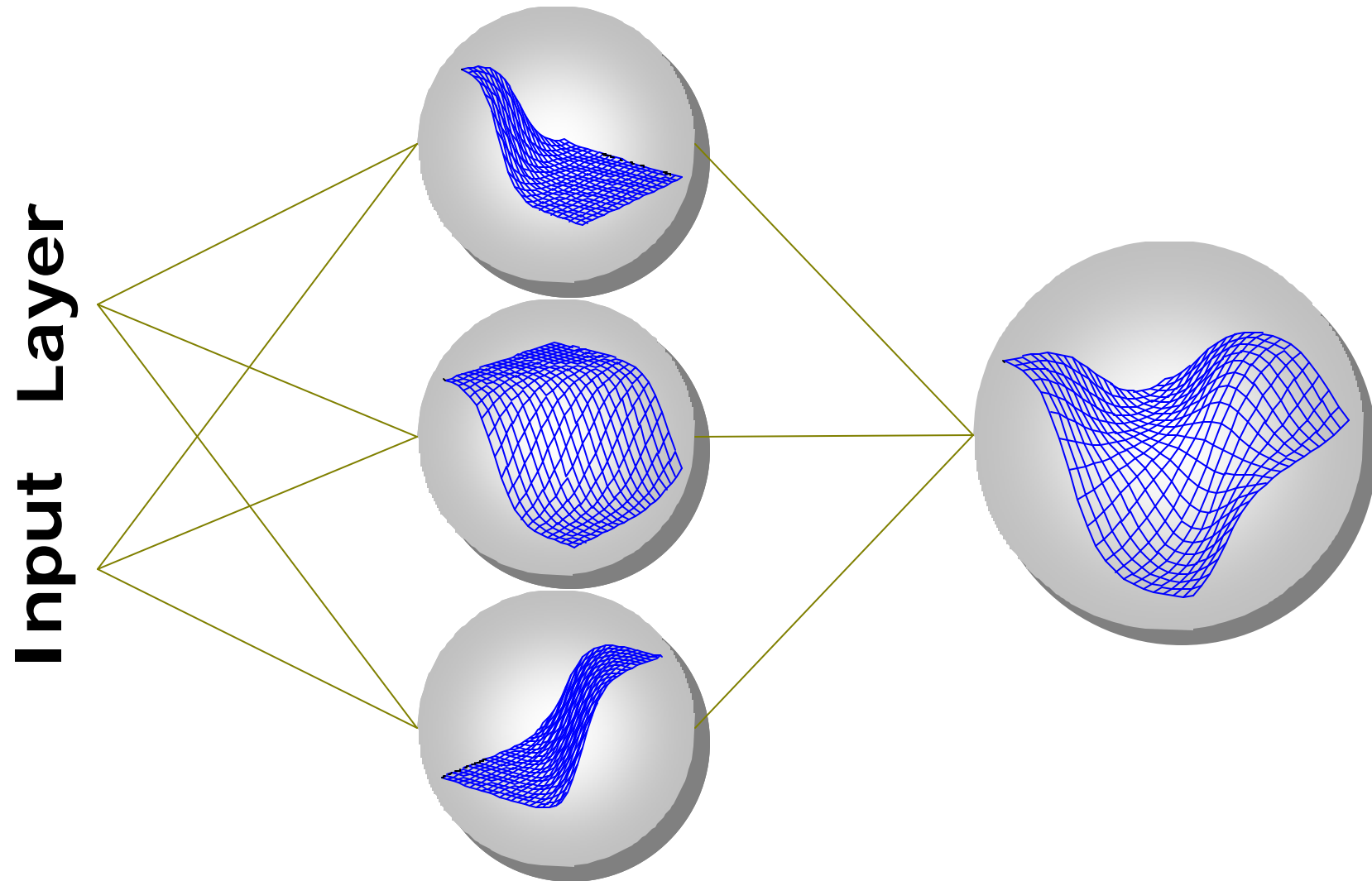
# *Multilayer Perceptron*

---



# *Activation Function*

---



## *Historical Background*

---

Rosenblatt, F. (1958), "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain", *Psychological Review*, (65), 1958.

## *Historical Background*

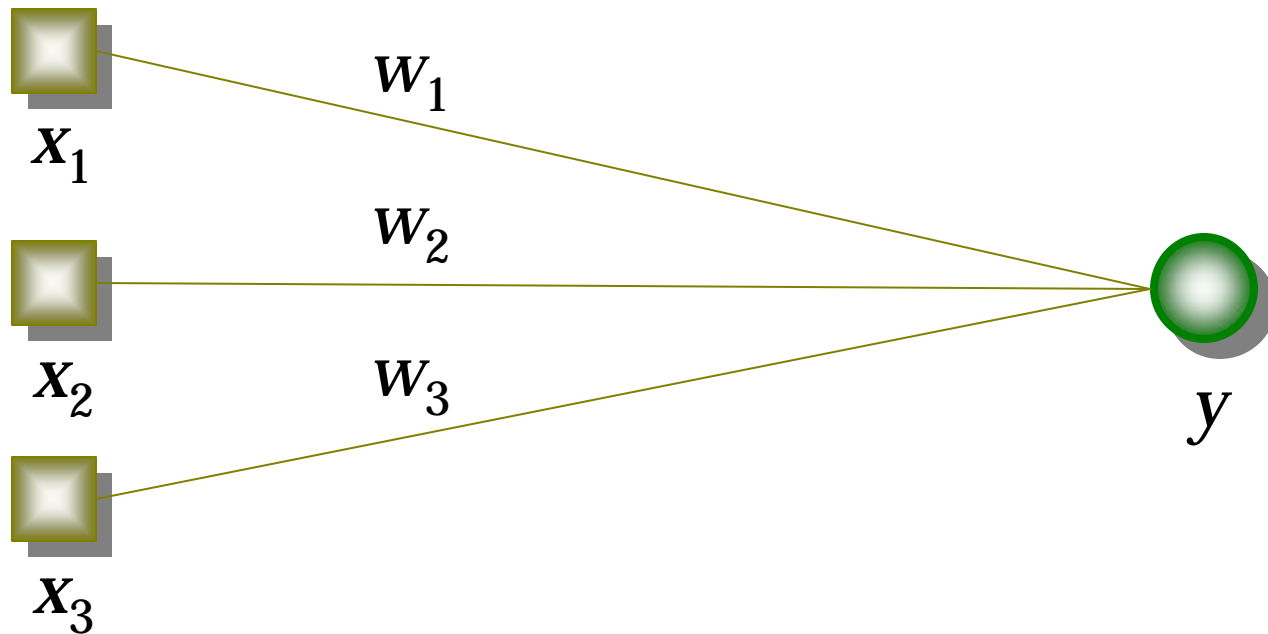
---

Ackerly, D.H., G.E. Hinton, and T.J. Sejnowski (1985), "A learning algorithm for Boltzmann Machines", *Cognitive Science*, (9), 147-169.

# *(Multiple) Linear Regression*

---

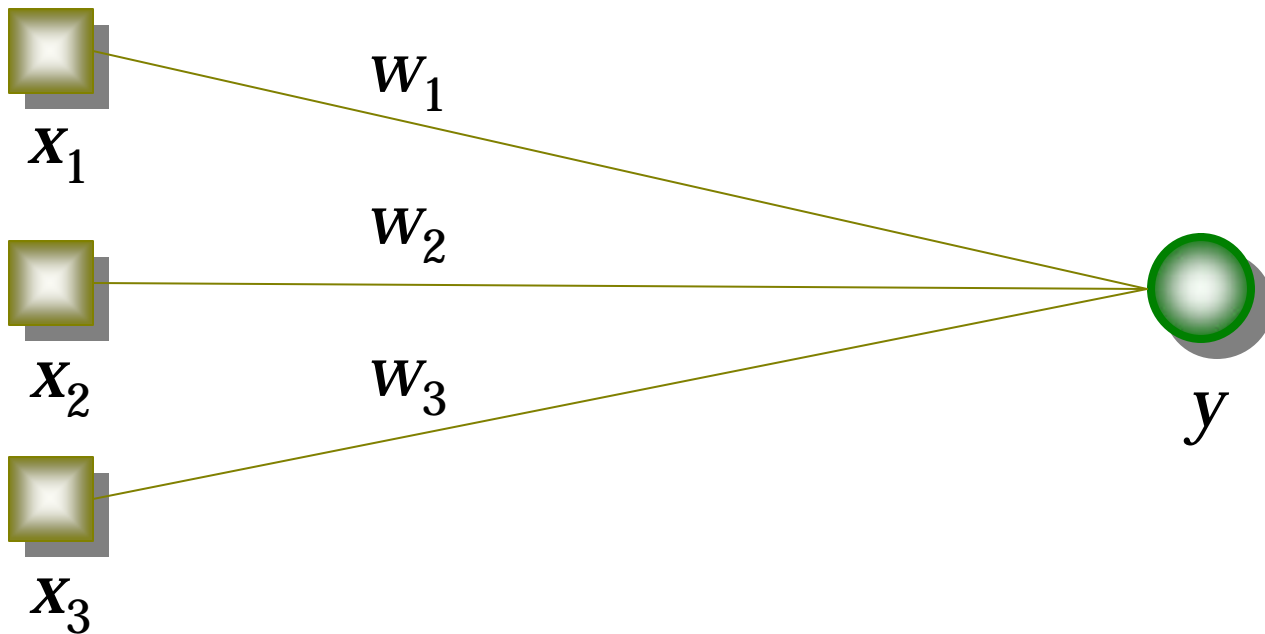
$$E(y) = w_0 + w_1x_1 + w_2x_2 + w_3x_3$$



# Logistic Regression

---

$$\ln\left(\frac{E(y)}{1 - E(y)}\right) = w_0 + w_1x_1 + w_2x_2 + w_3x_3$$



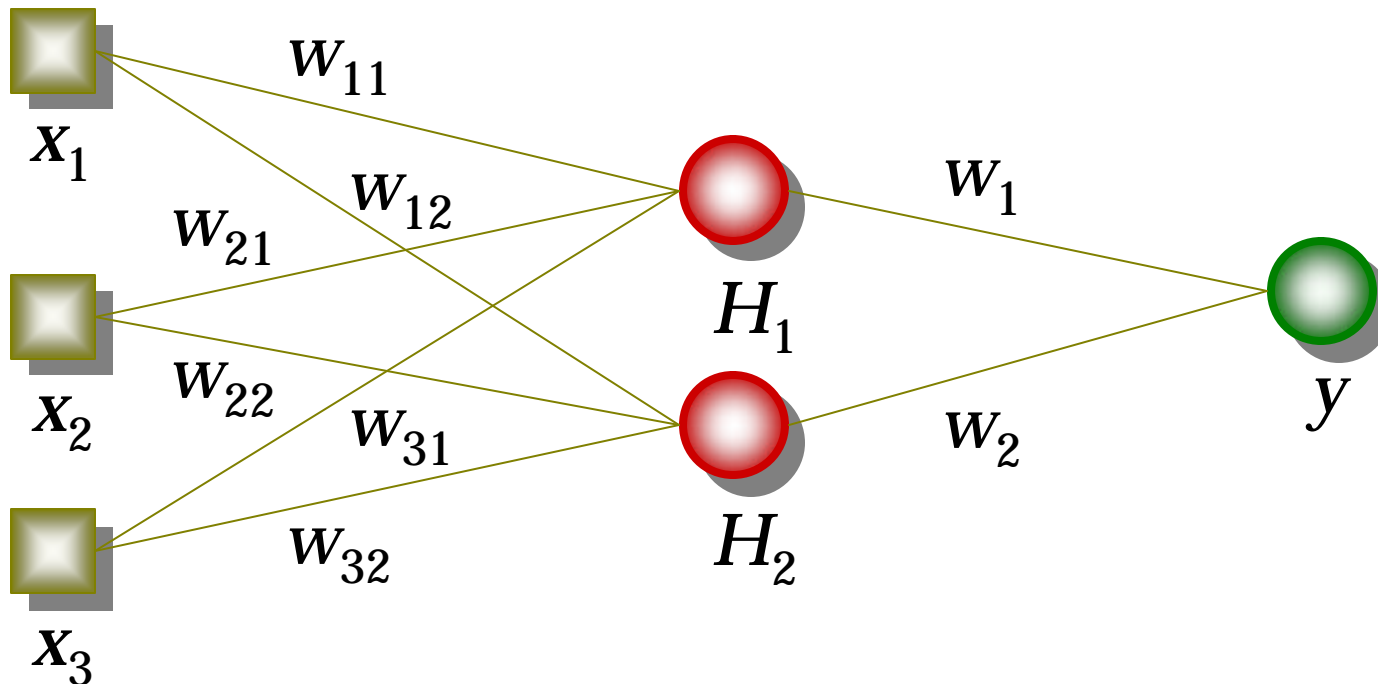
# Feed-Forward Neural Network

---

$$g_0^{-1}(E(y)) = w_0 + w_1 H_1 + w_2 H_2$$

$$H_1 = g_1(w_{01} + w_{11}x_1 + w_{21}x_2 + w_{31}x_3)$$

$$H_2 = g_2(w_{02} + w_{12}x_1 + w_{22}x_2 + w_{32}x_3)$$



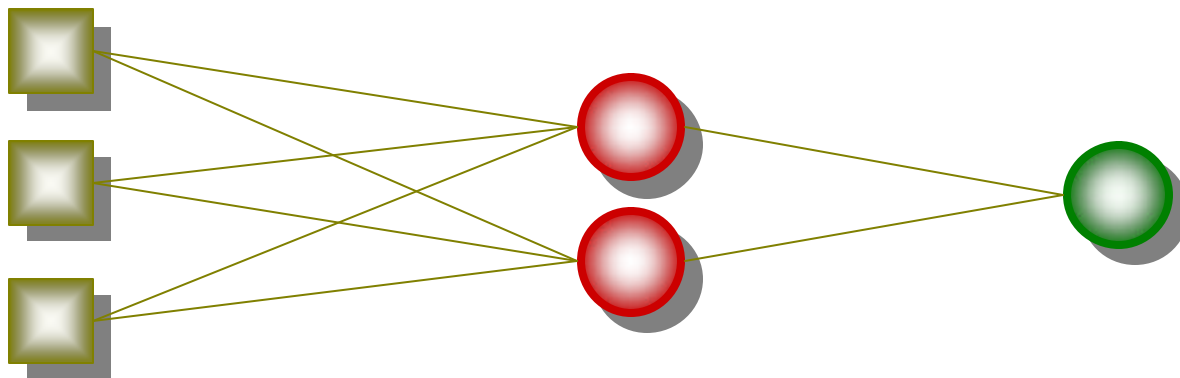
# Multilayer Perceptron

---

$$g_0^{-1}(E(y)) = w_0 + w_1 H_1 + w_2 H_2$$

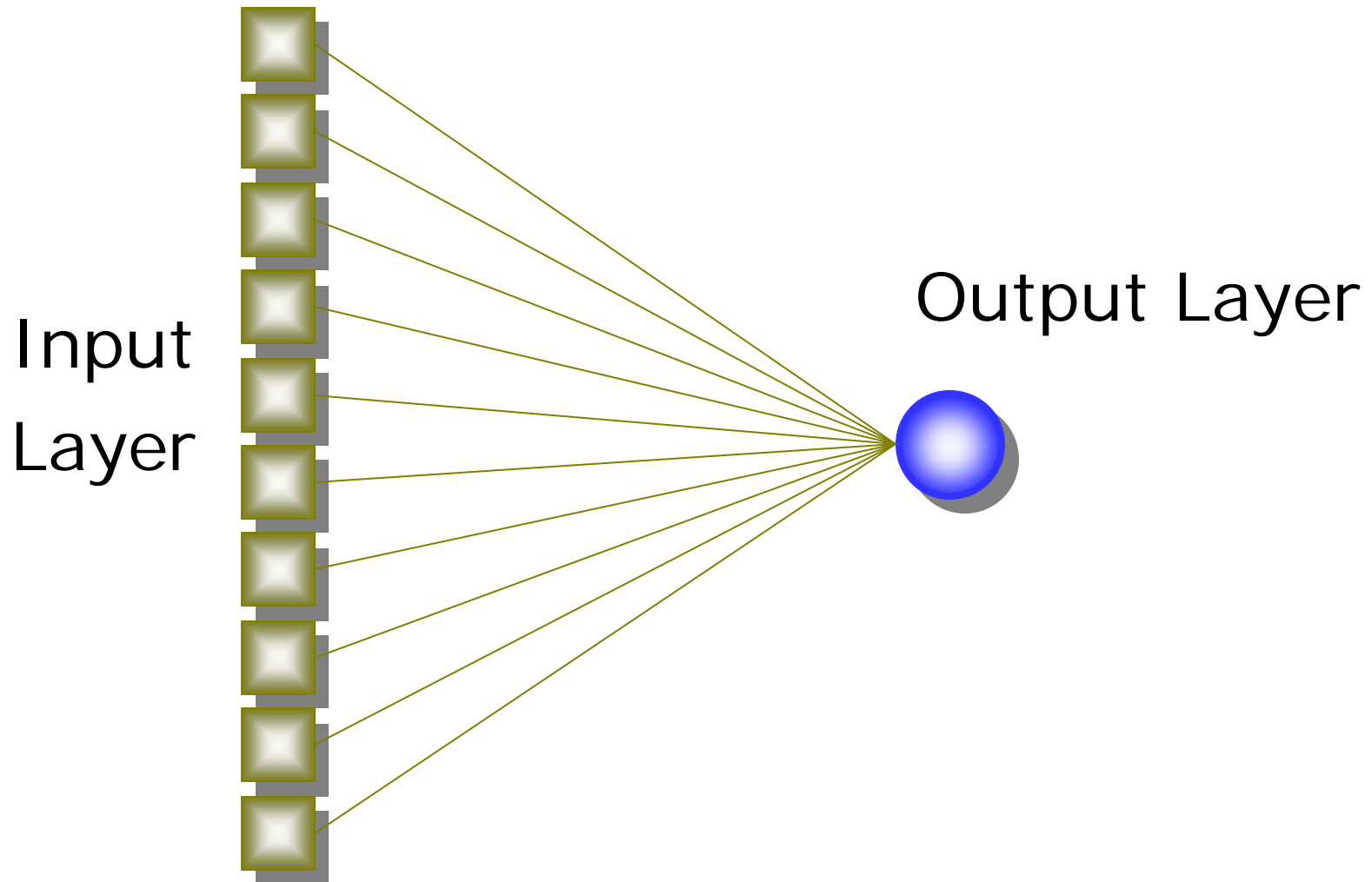
$$H_1 = \tanh(w_{01} + w_{11}x_1 + w_{21}x_2 + w_{31}x_3)$$

$$H_2 = \tanh(w_{02} + w_{12}x_1 + w_{22}x_2 + w_{32}x_3)$$



# *Generalized Linear Models*

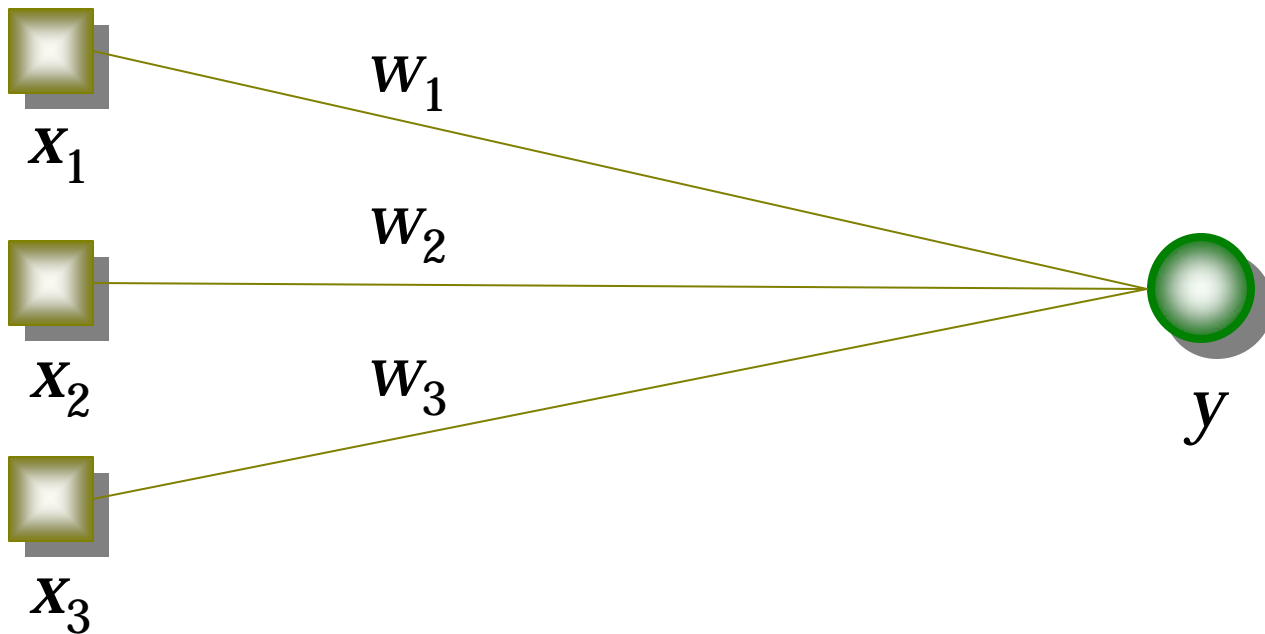
---



# Generalized Linear Model

---

$$g_0^{-1}(E(y)) = w_0 + w_1x_1 + w_2x_2 + w_3x_3$$



# *Output Activation Function*

---

$$g_0^{-1}(E(y)) = \mathbf{m}(\mathbf{x}, \mathbf{w}) \quad \Leftrightarrow \quad E(y) = g_0(\mathbf{m}(\mathbf{x}, \mathbf{w}))$$

Inverse output activation function  
= *link function*

# Link Functions

---

	$g_0^{-1}(E(y))$	$E(y)$	Range
Identity	$E(y)$	$m(\mathbf{x}, \mathbf{w})$	$(-\infty; +\infty)$
Logit	$\ln\left(\frac{E(y)}{1 - E(y)}\right)$	$\frac{1}{1 + e^{-m(\mathbf{x}, \mathbf{w})}}$	$(0; 1)$
Log	$\ln(E(y))$	$e^{m(\mathbf{x}, \mathbf{w})}$	$(0; +\infty)$

# *Link Function Inventory*

---

## Link

identity

log

logit

generalized logit

cumulative logit

## Output Act.

identity

exponential

logistic

softmax

logistic

## Scale

interval

nonnegative

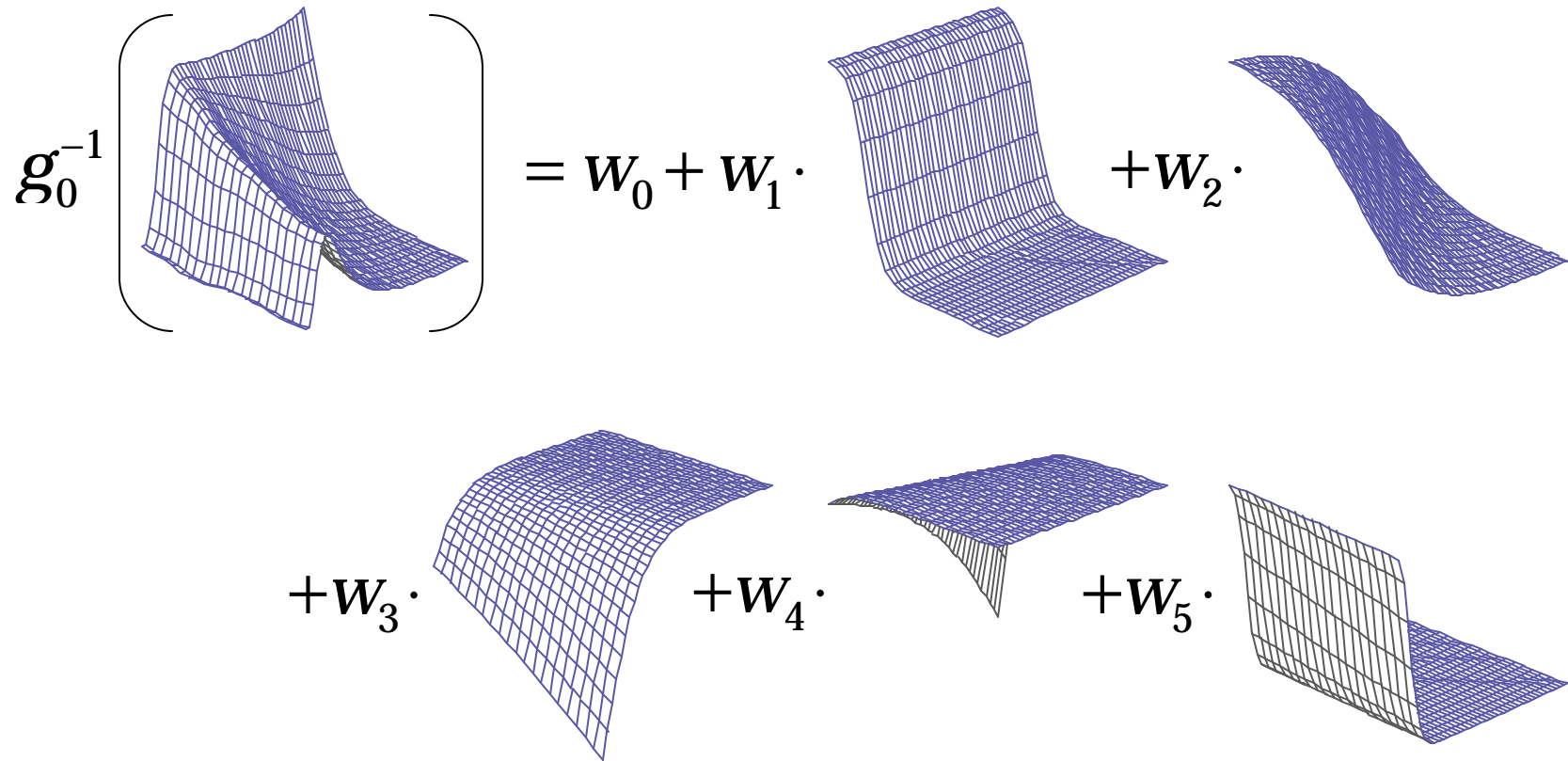
binary

polychotomous

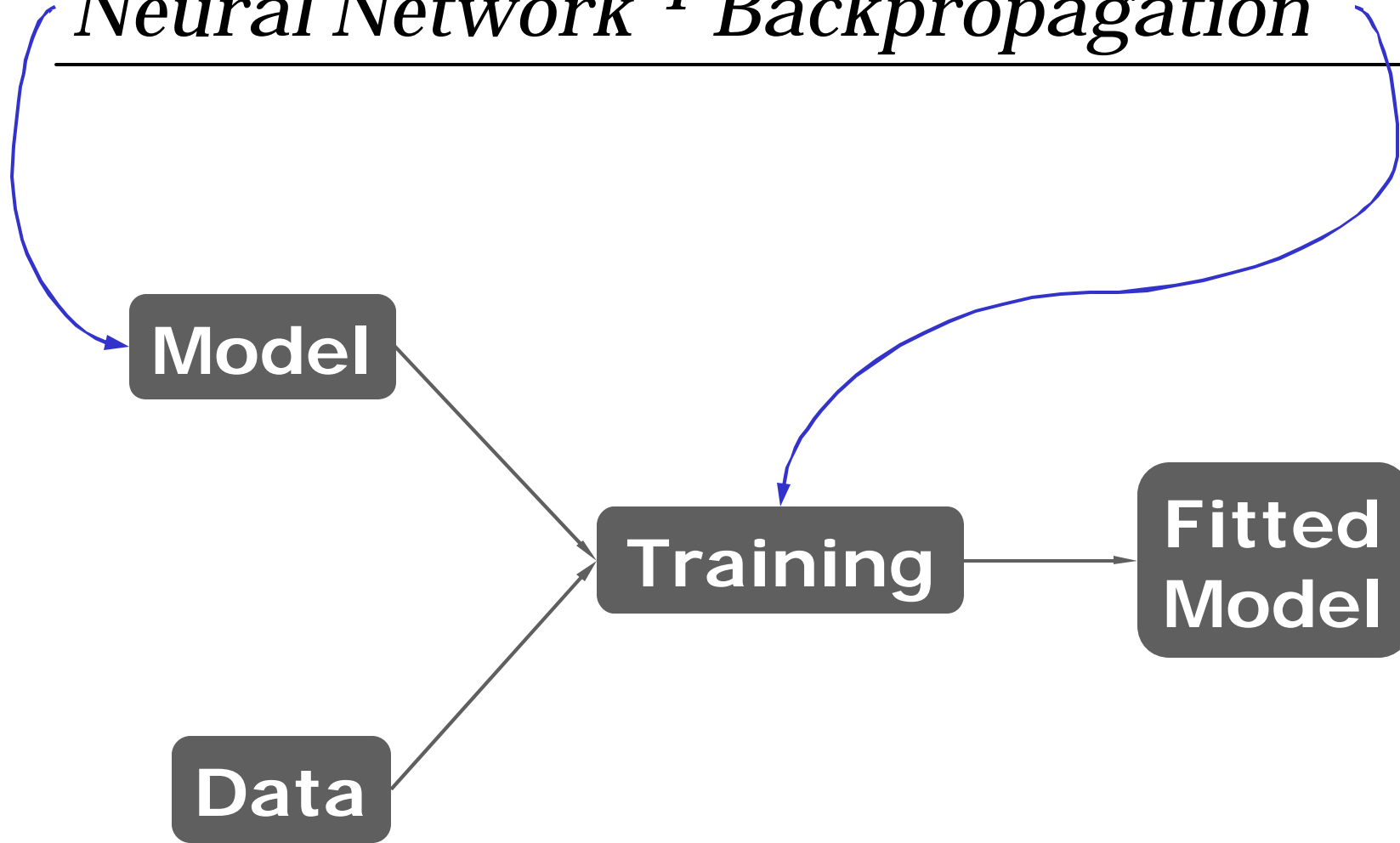
ordinal

# Universal Approximation

---



# Neural Network <sup>1</sup> Backpropagation



# *Practical Difficulties*

---

Troublesome Training

Model Complexity/Specification


Incomprehensibility  $(y, \mathbf{x}) \rightarrow$    $\rightarrow \hat{y}$

Unreasonable Expectations

Anthropomorphism

Noisy data

Data preparation



**“My CPU is a neural-net processor... a learning computer”**

**“My CPU fits regression models to data”**

A photograph of a red brick building with white columns and arched windows. The building is set against a clear blue sky. In the foreground, there is a large tree on the left and a row of green bushes. The word "Demonstration" is written in red cursive text across the middle of the image, enclosed in a thin red rectangular border.

*Demonstration*

# *The Cultivation of Trees*

---

## Split Search

Which splits are to be considered?

## Splitting Criterion

Which split is best?

## Stopping Rule

When should the splitting stop?

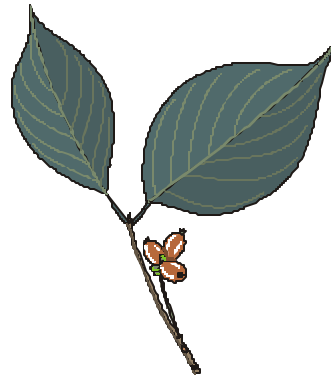
## Pruning Rule

Should some branches be lopped-off?

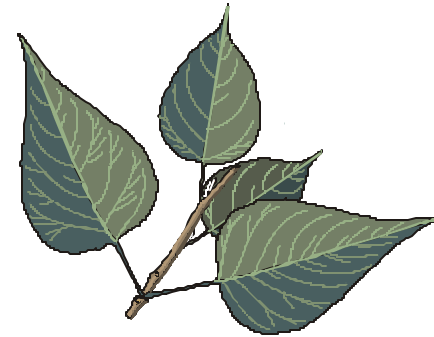
# *A Field Guide to Tree Algorithms*



**AID**  
**THAID**  
**CHAID**



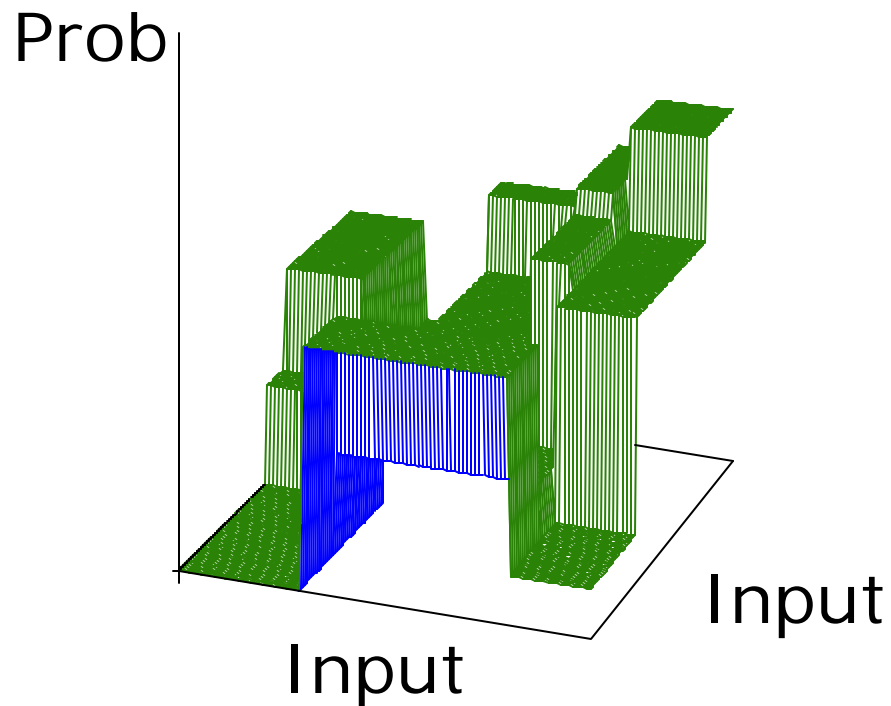
**CART**



**ID3**  
**C4.5**  
**C5.0**

# *...Benefits*

---



Multivariate  
Step Function

Automatically

Detects interactions  
(AID)

Accommodates  
nonlinearity

Selects input  
variables

Ease of  
interpretation

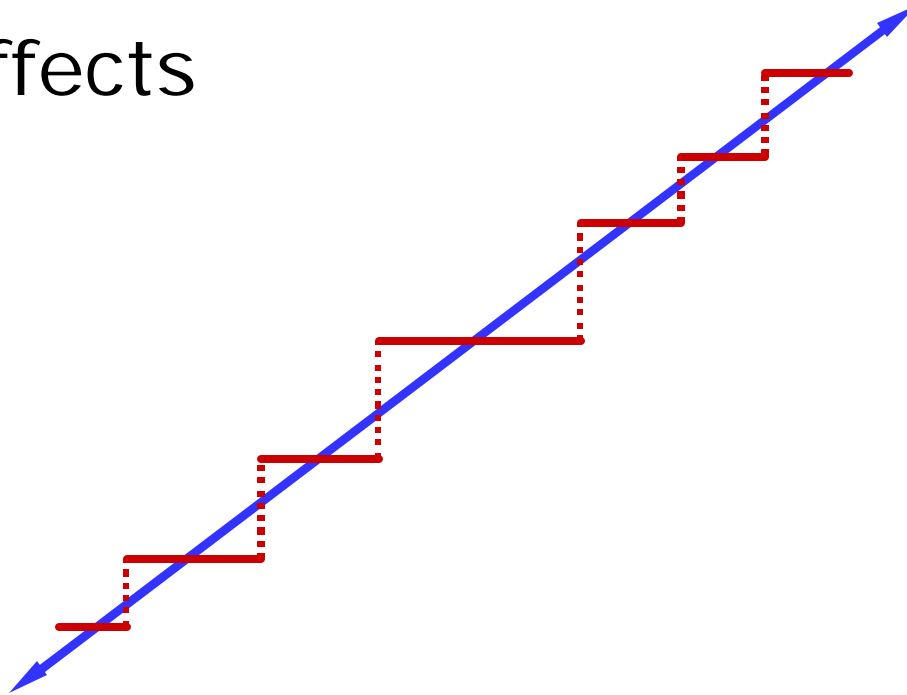
# *Drawbacks of Trees*

---

Roughness

Linear, Main Effects

Instability



# *Demonstration*



# Unsupervised Classification

---

## Training Data

case 1: inputs, ?  
case 2: inputs, ?  
case 3: inputs, ?  
case 4: inputs, ?  
case 5: inputs, ?

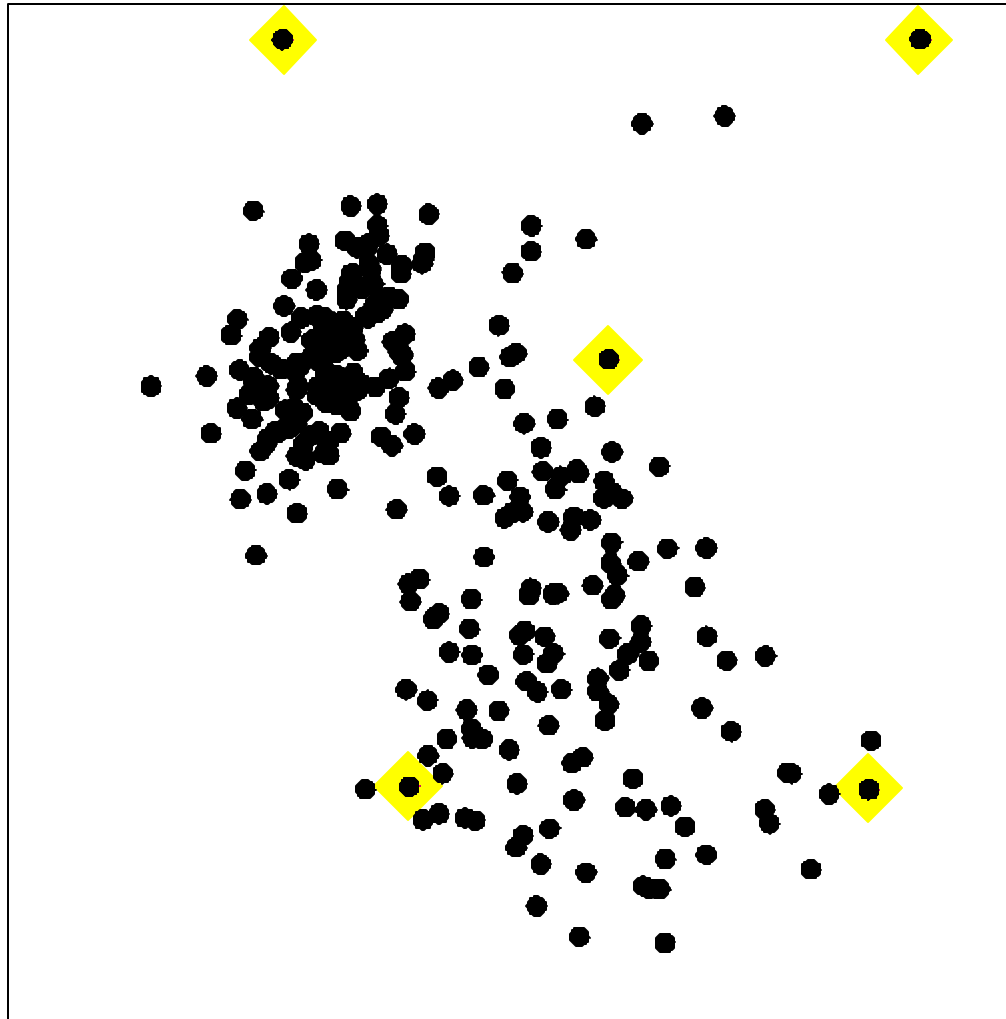
## Training Data

case 1: inputs, cluster 1  
case 2: inputs, cluster 3  
case 3: inputs, cluster 2  
case 4: inputs, cluster 1  
case 5: inputs, cluster 2



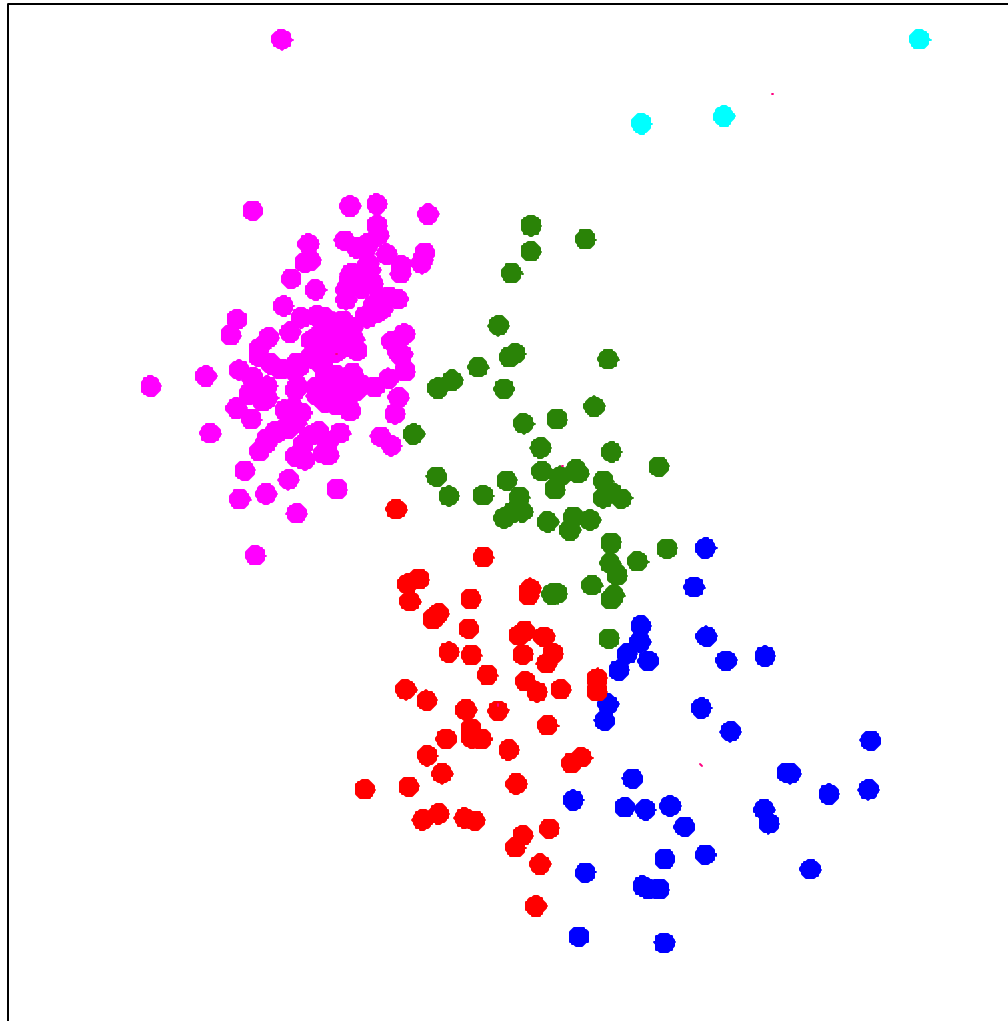
# *K-means Clustering*

---



# *Final Grouping*

---



# *Areas of Applications*

---

Genomics

Micro-Array

Others

Nursing Home Staff Management

Many others

# *Demonstration*



# Association Rules

---



<u>Rule</u>	<u>Support</u>	<u>Confidence</u>
$A \Rightarrow D$	2/5	2/3
$C \Rightarrow A$	2/5	2/4
$A \Rightarrow C$	2/5	2/3
$B \ \& \ C \Rightarrow D$	1/5	1/3

# *Occupational Epidemiology*

---

Identifying Risk patterns in Employment histories

Association Analysis

Employee is “basket”, events during tenure are “items”



# *UAB Data Mining and Knowledge Discovery Research Group*

---

Warren T. Jones<sup>1</sup>, J. Michael Hardin<sup>2, 3</sup>, Alan P. Spague<sup>1</sup>, Stephen E. Brossette<sup>1</sup>, and Stephen Moser<sup>4</sup>

<sup>1</sup>Department of Computer Science

<sup>2</sup>Department of Health Informatics

<sup>3</sup>Department of Biostatistics

<sup>4</sup>Department of Pathology



# *Data Mining Surveillance System* *(DMSS)*

---

A Knowledge Discovery System  
for Epidemiology

Stephen E. Brossette, J. Michael Hardin, Warren T.  
Jones, Alan P. Spague, and Stephen Moser

# ***A Strategy for Geomedical Surveillance Using the Hawkeye Knowledge Discovery System***

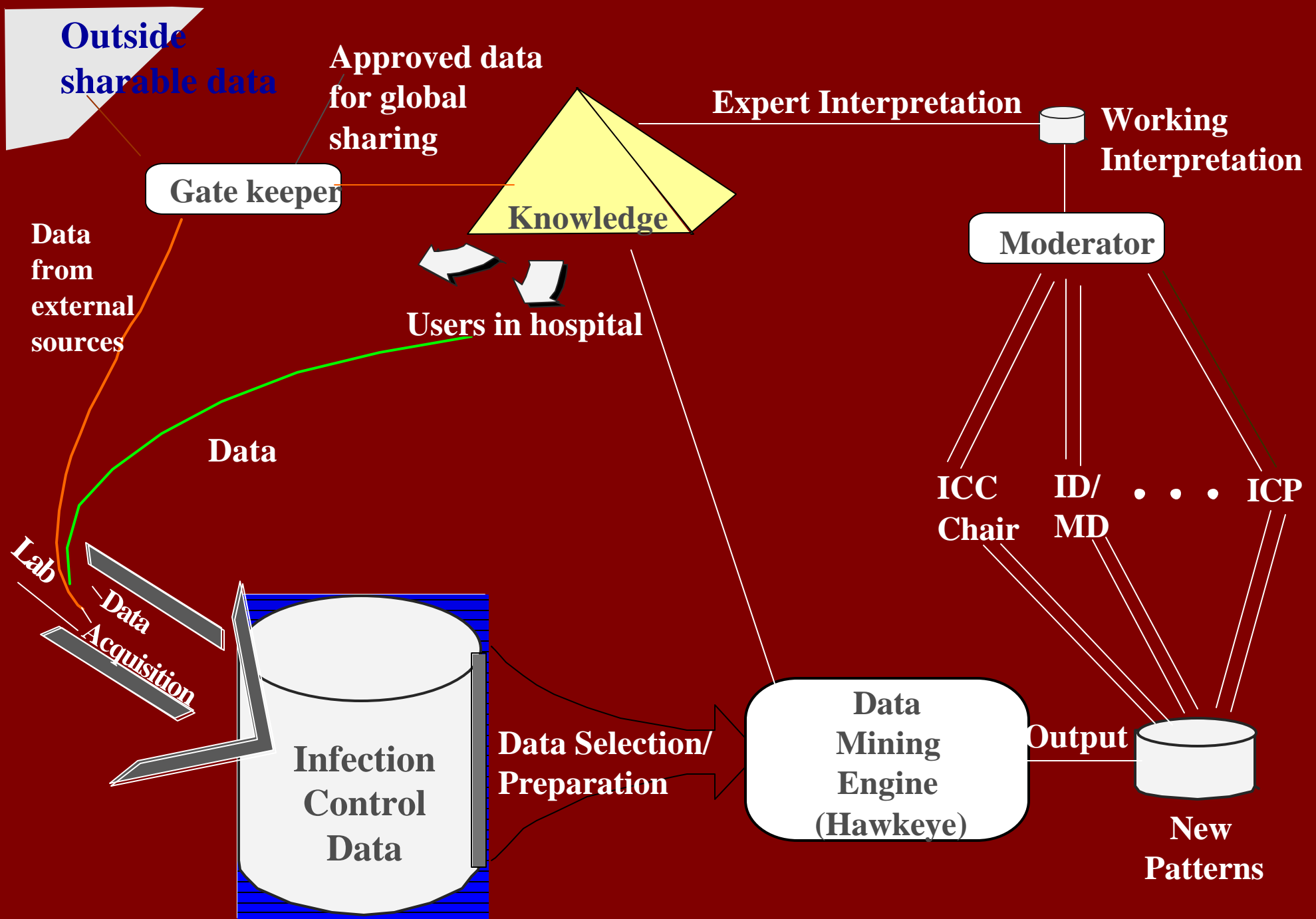
**Daisy Y. Wong <sup>3</sup>, Warren T. Jones <sup>3</sup>, Stephen E. Brossette <sup>3</sup>,  
J. Michael Hardin <sup>2</sup> and Stephen A. Moser <sup>1</sup>**

*Departments of Pathology <sup>1</sup>, Biostatistics <sup>2</sup>, Health Informatics<sup>2</sup>, Computer and Information Sciences <sup>3</sup>*

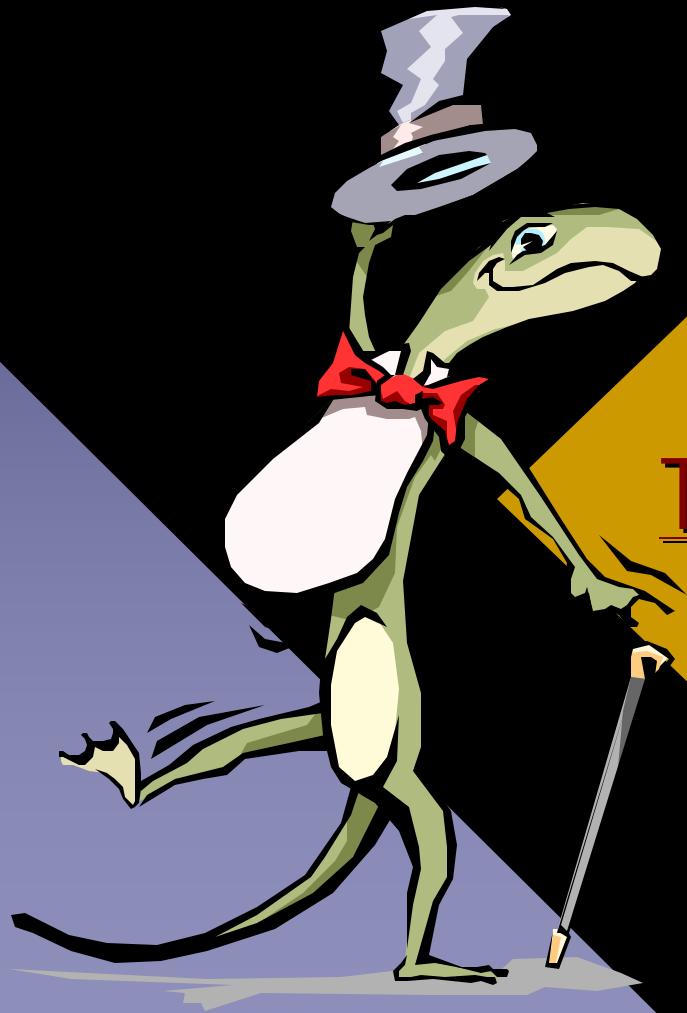
*University of Alabama at Birmingham*

*USA*

# A Local Site Model for Global Collaboration



Questions?



Thank You!