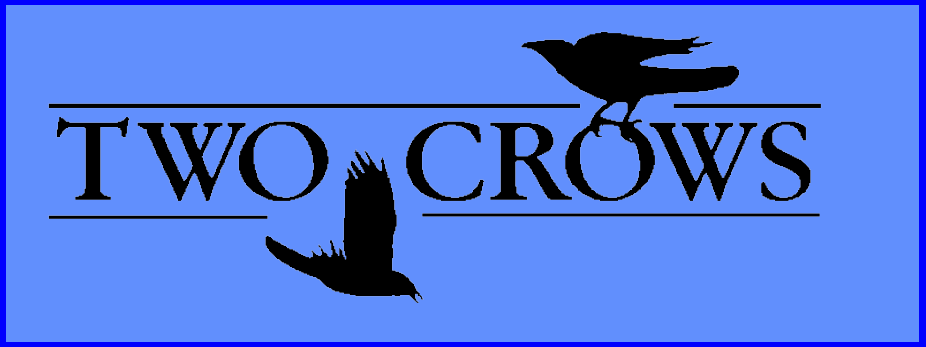
Assessing Data Mining: The State of the Practice



Herbert A. Edelstein
Two Crows Corporation
10500 Falls Road
Potomac, Maryland 20854
www.twocrows.com
(301) 983-3555

Objectives

- Separate myth from reality
- Interactive session: question driven! The slides are largely to ensure common background.

The Key to Value

- The utility of data increases as it spans the business value chain and is integrated
- Information increases as data are related
  - Consolidate similar databases
  - Consolidate different types of databases
- Without data and good analysis all you have are opinions.

Data Mining Definitions

- What IT departments call
  - OLAP
  - Query
  - Statistics

Data Mining Definitions

- ***Knowledge Discovery in Databases*** **is the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data. (Fayyad, Piatetsky-Shapiro, & Smyth)**
  - **KDD is the process, data mining is the application of algorithms**
  - **Includes description and prediction**
  - **Large databases often explicitly added to definition**

Data Mining Definitions

- *Data mining* is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions
- Exploration and description is required but not the goal

New Statistical Software

- Takes advantage of advances in hardware and software
- Provides new interfaces for a wider class of users
- Comes from statistics, machine learning and information systems

Why Data Mining is Taking Off

- Demand for information
- Availability of data
  - Enormous quantity of happenstance data
  - Spread of data warehouses
  - Data is easily accessible through the Web.
- Improved technology
  - Inexpensive, scalable processing
  - Inexpensive storage
  - High bandwidth

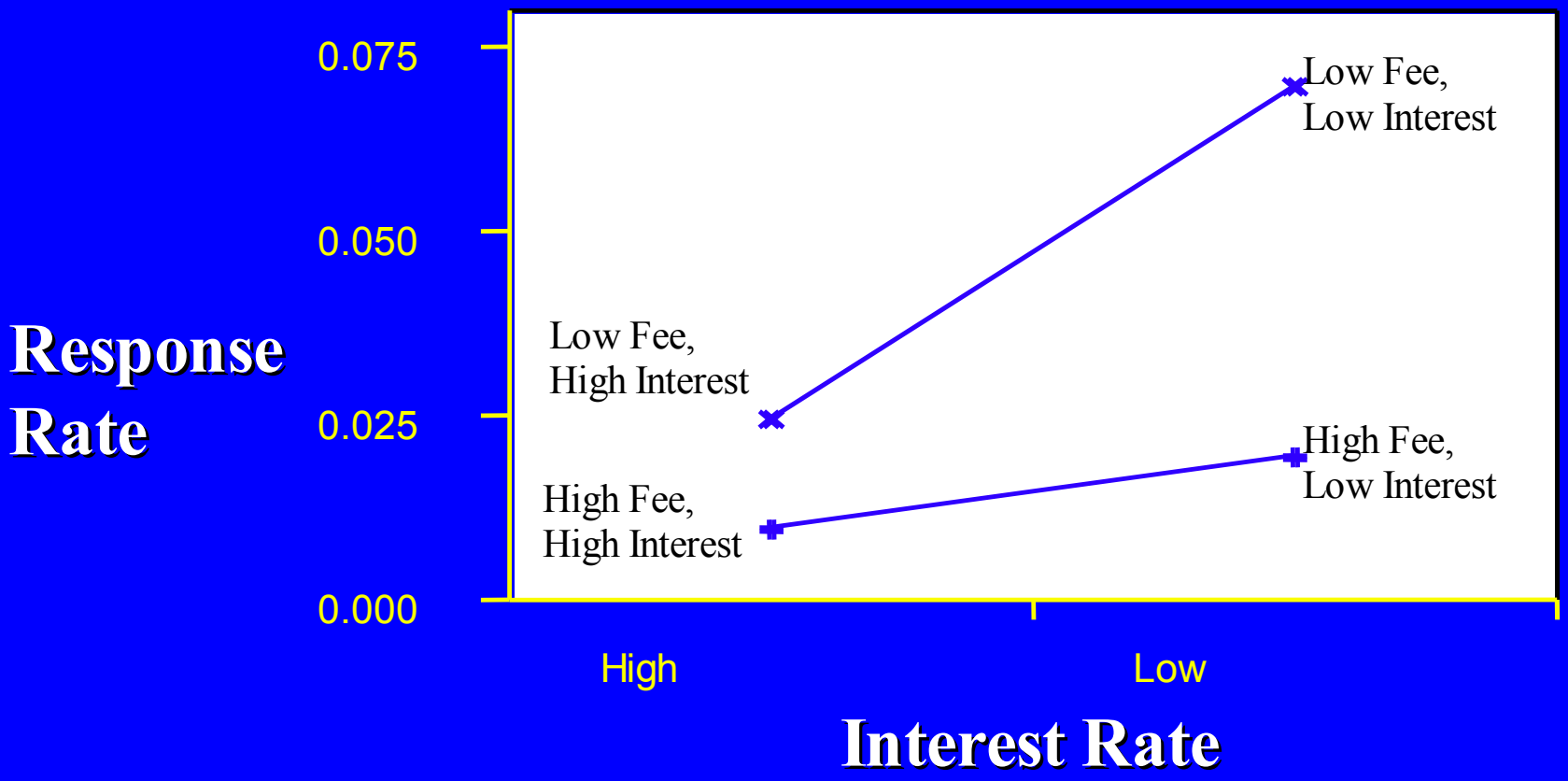
Why Data Mining Is Needed

- Massive amounts of data
  - Example:
    - 75 million customers
    - 3,000 columns for each customer
- Low signal to noise
  - Subtle relationships
  - Variation
- Allow domain experts to build predictive models

Data Mining Products

- Handle large volumes of data
- Reduce dependence on the modeler
  - Model specification
  - Knowing characteristics of variables
- Create hypotheses
- Emphasize prediction
- Simplify model deployment
- Data mining is a productivity tool even for the skilled statistician

Attribute Interactions



Data Mining Myths

- Data mining does NOT
  - Find answers to unasked questions
  - Explain behavior.
  - Continuously monitor data for new patterns
  - Eliminate the need to understand your business
  - Eliminate the need to collect good data
  - Eliminate the need to be a good data analyst

Commercial Applications

- Industry
  - Retail
  - Financial
  - Manufacturing
  - Insurance
  - Publishing
  - Health care
- Application
  - Marketing
  - Sales force management
  - Fraud detection
  - Risk management

Credit Risk Analysis

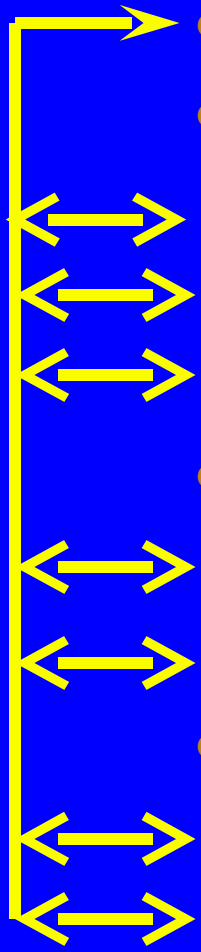
- Database checks
  - Data validation: Does an address exist, is the social security number consistent with date and place of birth, etc.
  - Where is Benford's Law applicable?
  - History checks: when was the last time a property sold?
- Profiling good and bad credit risks – Finding good risks within bad categories.
- Outlier detection – uni-dimensional and multi-dimensional
- Does not replace human follow up

Benford's Law

If the numbers under investigation are not entirely random but somehow *socially or naturally related*, the distribution of the first digit is not uniform. More accurately, digit D appears as the first digit with the frequency proportional to log10(1 + 1/D). In other words, one may expect 1 to be the first digit of a random number in about 30% of cases, 2 will come up in about 18% of cases, 3 in 12%, 4 in 9%, 5 in 8%, etc.

http://www.cut-the-knot.org/do_you_know/zipfLaw.shtml

Data Mining Process



- Define business problem
- Prepare the data
  - Build data mining database
  - Explore data
  - Prepare data for modeling
- Create models
  - Build models
  - Evaluate models
- Act on the results (implementation)
  - Apply models to new data
  - Integrate results with an application

Data Preparation

- Build data mining database
- Explore data
- Prepare data for modeling

60% to 95% of the time is spent preparing the data

The Tools

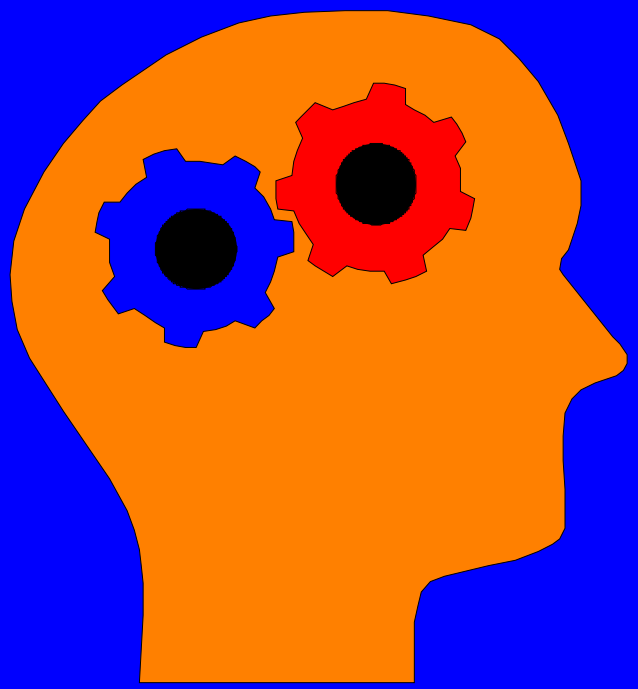
- Starting simply - linear regression
- Fancier regressions
- Projections
- Smoothing based regressions
- Survival analysis
- Nearest neighbor methods
- Collaborative filtering
- Trees
- MARS
- Neural networks
- Genetic algorithms

Decision Trees

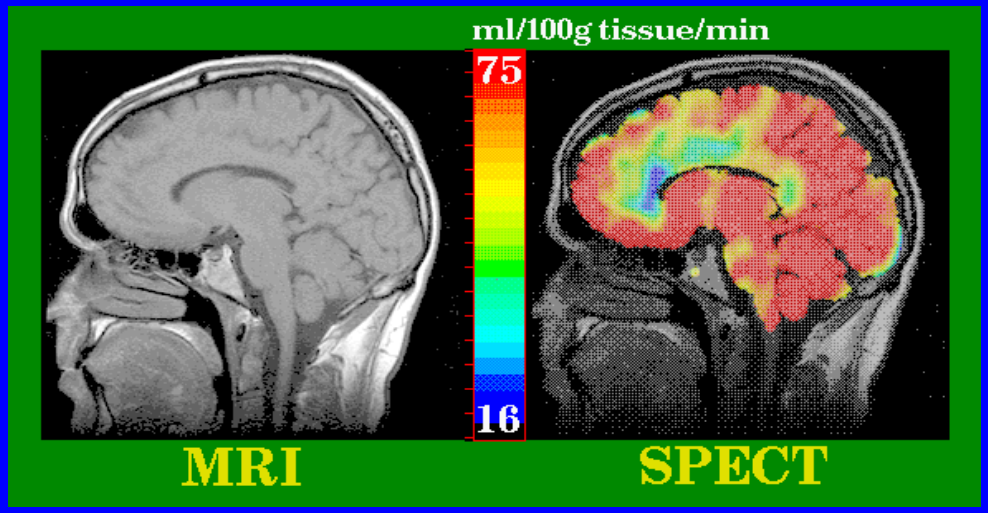
- Build a tree inductively that describes a set of data
- Classification tree: Hierarchical set of rules which classify data
- Regression tree: Hierarchical set of rules which predicts values

Neural Nets

- Don’t resemble the brain
  - A model of memory
  - A model of learning

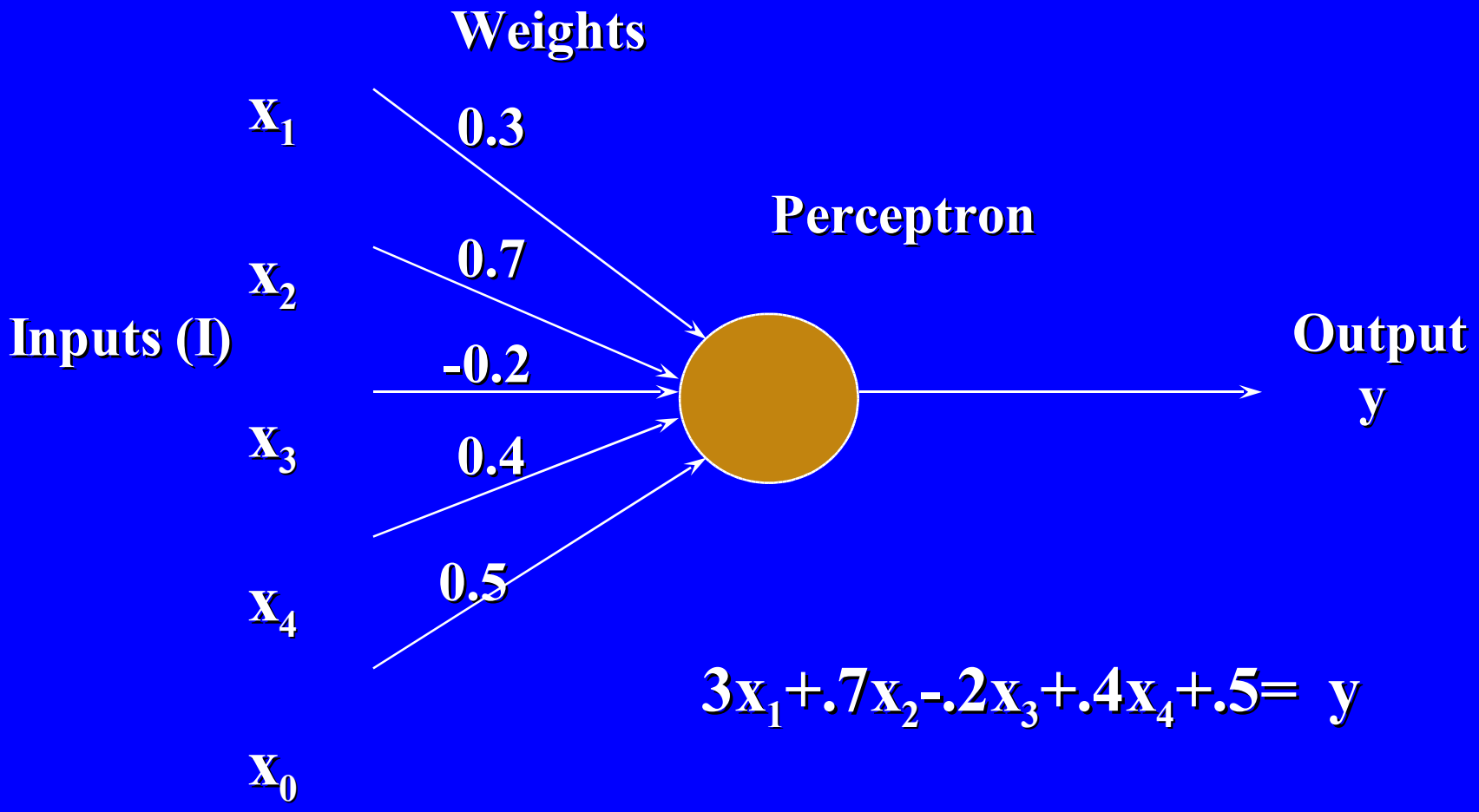


NN Don't Resemble the Brain

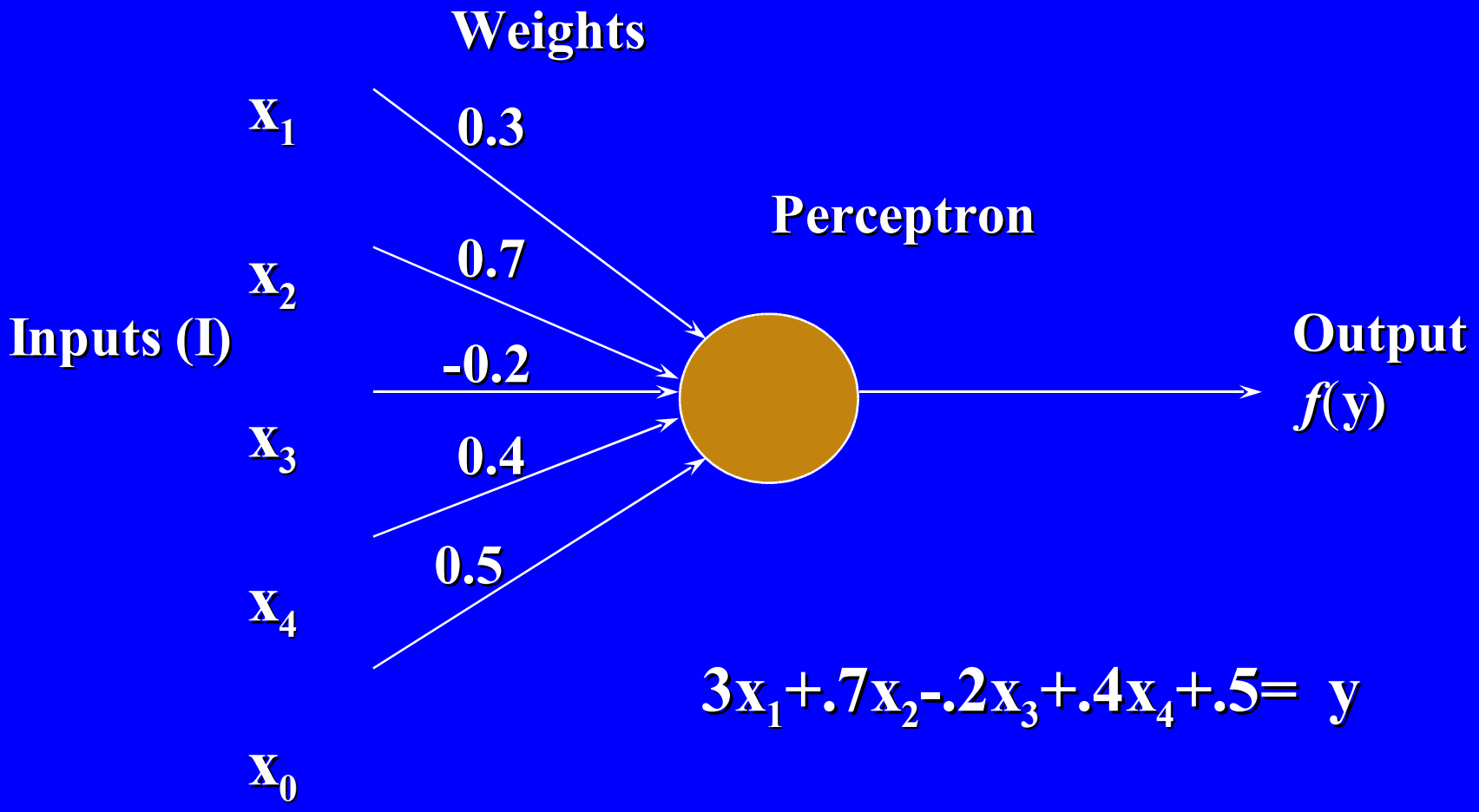


- Brain neurons can not only add signals, but they can subtract, multiply, divide, filter, average, etc.
- "The computational toolbox of individual neurons dwarfs the elements available to today's electronic circuit designers" Christof Koch, Professor, Computation and Neural Systems, Cal Tech

Linear Regression Example



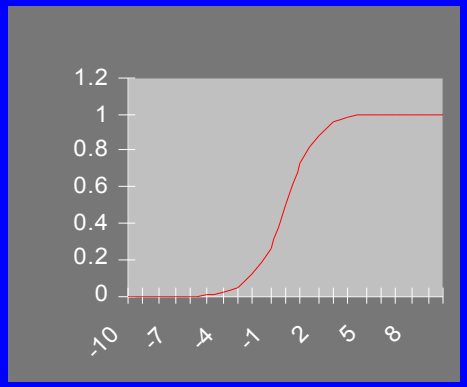
Logistic Regression Example



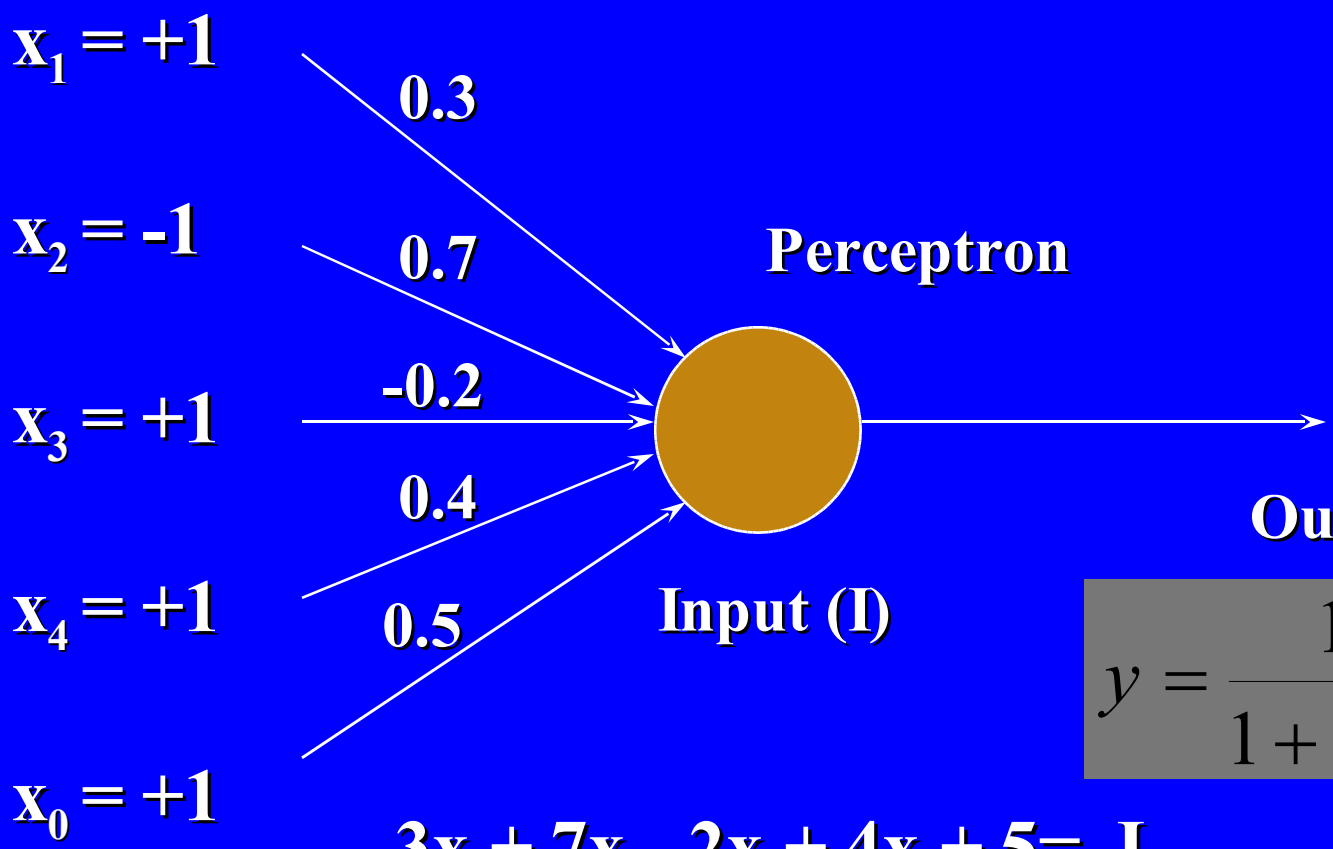
Sigmoid Activation Function

- S shaped
- Continuous approximation of threshold
- Has derivative

$$f(y) = \frac{1}{1 + e^{-y}}$$



Logistic Regression Example

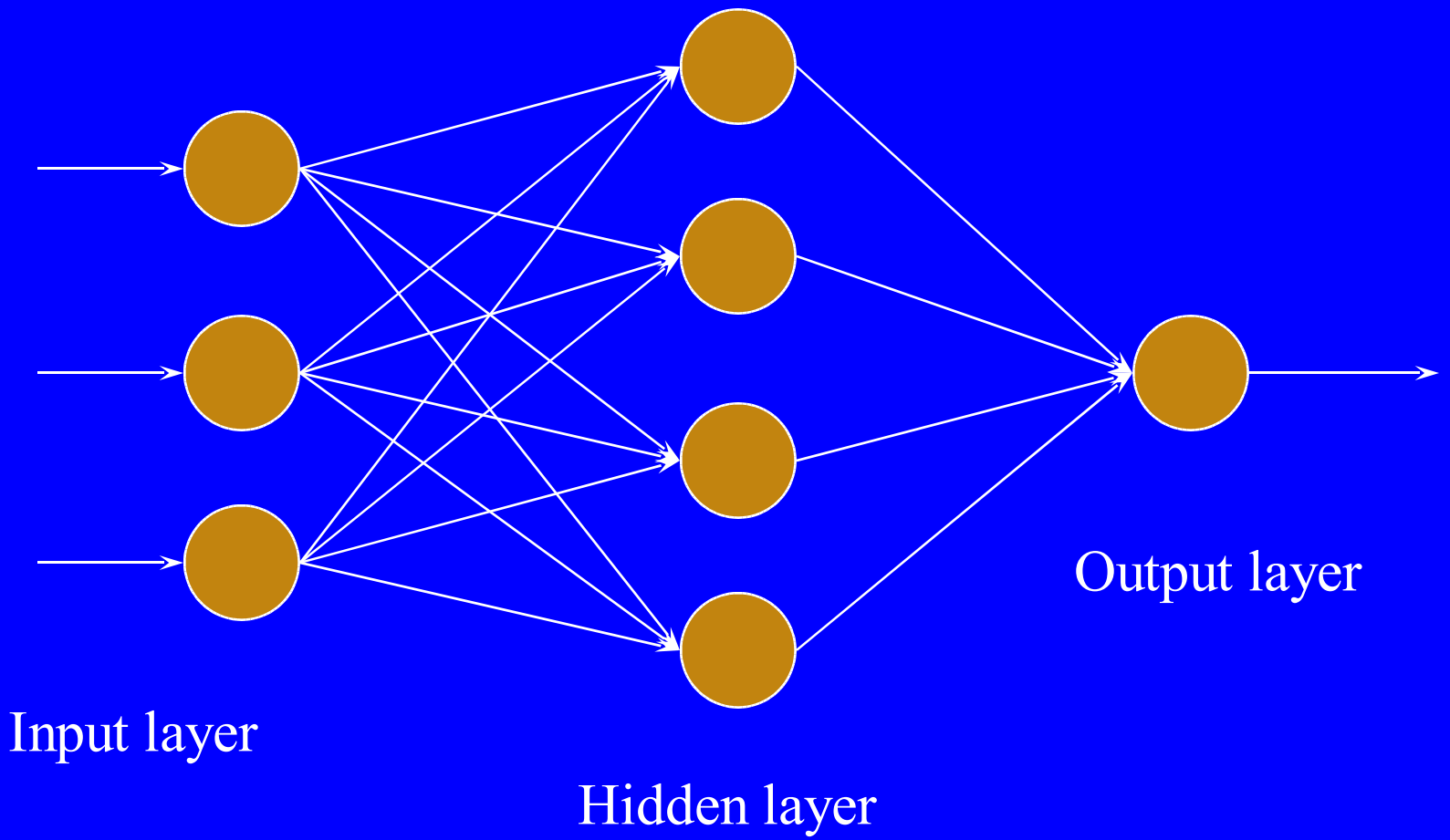


$$y = \frac{1}{1 + e^{-.3}} = .57$$

$$3x_1 + .7x_2 - .2x_3 + .4x_4 + .5 = I$$

$$= .3 - .7 - .2 + .4 + .5 = .3$$

Layered Architecture



State of the Industry Report Card

	1999	2001	2003
Products			
User interface	C	B -	B
Data preparation	D	C	C
Data exploration	C-	C	B
Algorithms	C	B	B+
Model deployment	D	C	B
Robustness	C	B-	B+
Adoption			
Organizational readiness	C	C	B
Successful applications	C+	B	A
Training	D	C	B
Available consulting	D	C	B

State of the Product Market

- Good products are available
  - Feature rich
  - Mature
  - Reasonably stable
  - Well supported

Tools and Technology

- Match tool to application and users.
  - Allow time for training and learning
- The tool is not the solution.
  - Model building is the fun and easy part.

Further Reading

Herb Edelstein, **Introduction to Data Mining and Knowledge Discovery**, 2003

Herb Edelstein, **Data Mining Technology Report**, 2003

M. Berry and G. Linoff, **Mastering Data Mining Techniques**, John Wiley, 1999

William S. Cleveland, **The Elements of Graphing Data**, revised, Hobart Press, 1994

Howard Wainer, **Visual Revelations, Copernicus, 1997**

T. Hastie, R. Tibshirani, J. H. Friedman, **The Elements of Statistical Learning : Data Mining, Inference, and Prediction,** Springer Verlag, 2001

Richard O. Duda, P. E. Hart, D. G. Stork, **Pattern Classification**, John Wiley, 2000

David W Hosmer Jr., S. Lemeshow, **Applied Logistic Regression**, John Wiley, 2000

David W Hosmer Jr., S. Lemeshow, **Applied Survival Analysis**, John Wiley, 1999

David J. Hand, H. Mannila, P. Smyth , **Principles of Data Mining** , MIT Press, 2001

Brieman, Freidman, Olshen, and Stone, **Classification and Regression Trees**, Wadsworth, 1984

J. R. Quinlan, C4.5: **Programs for Machine Learning**, Morgan Kaufmann, 1992

