# New Formulations for Predictive Learning

Vladimir Cherkassky
University of Minnesota
cherk001@umn.edu

Tutorial at IJCNN-05

July 31, 2005

# Outline

- Motivation and Background
- Standard Inductive Learning Formulation
- Alternative Formulations

  - non-inductive types of inference

  - non-standard inductive formulations
- Predictive models for interpretation
- Conclusions

# Motivation:
## Importance of Problem Formulation

- Traditional (Simplistic) View

Data + Learning Method → Useful Model

  - 'Useful' ='Predictive'
  - May lead to misconceptions:
    - Inductive models are completely data-driven
    - The goal is to design better algorithms

# Motivation: philosophical

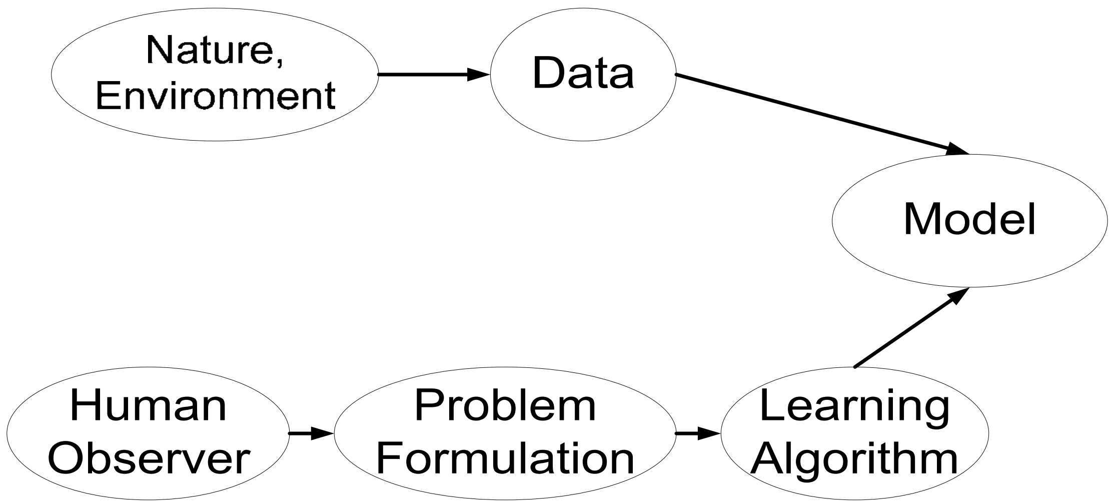
- Karl Popper: *Science starts from problems, and not from observations*

- Confucius: *Learning without thought is useless, thought without learning is dangerous*

- What to do vs how to do

# Motivation

- Another view of Predictive Learning




- Importance of problem formulation (vs algorithm)
- Just a few known formulations
- Thousands of algorithms

# Background: historical

- The problem of predictive learning

  *Given* past data + reasonable assumptions

  *Estimate* unknown dependency for future predictions

- Driven by applications (not theory)

# Historical Development

- Statistics (mathematical science)

  **Goal:** model identification, density estimation

- Neural Networks (empirical science)

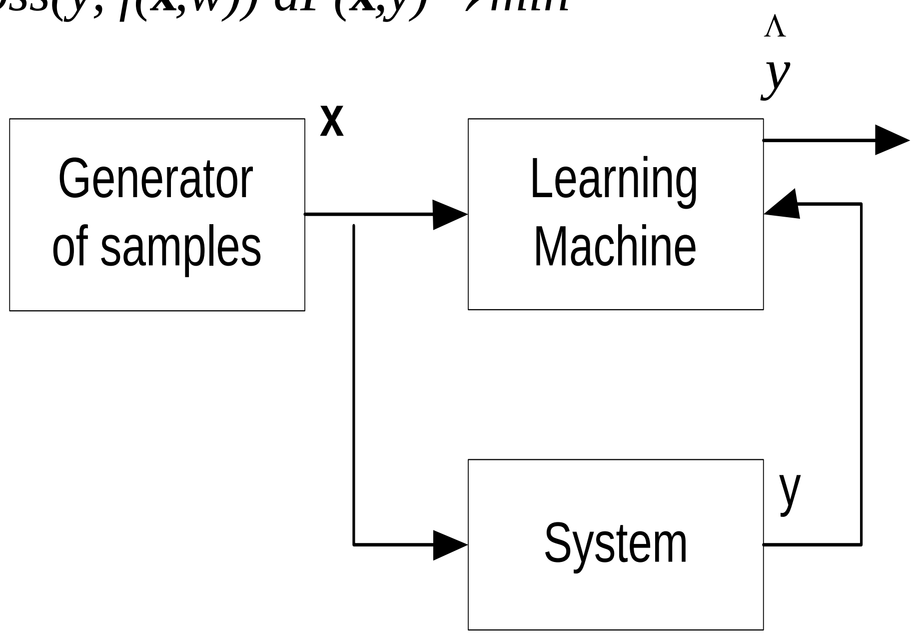
  **Goal of learning:** generalization, risk minimization

- Statistical Learning (VC theory) (natural science)

  **Goal of learning:** generalization for distinct learning problem formulations

# Standard Inductive Learning

- The learning machine observes samples (**x** ,*y*), and returns an estimated response $\hat{y} = f(\mathbf{x}, w)$
- Two modes of inference: identification vs imitation
- Risk $\int Loss(y, f(\mathbf{x},w))\, dP(\mathbf{x},y) \to min$



# Two Learning Problems

- Learning ~ estimating mapping $\mathbf{x} \to y$ (in the sense of risk minimization)
- Binary Classification: estimating an indicator function (with 0/1 loss)
- Regression: estimating a real-valued function (with squared loss)
- Assumptions: iid, training/test, loss fct

# Contributions of VC-theory

- The Goal of Learning

  system imitation vs system identification
- Two factors responsible for generalization
- Keep-It-Direct Principle (Vapnik, 1995)

  Do not solve a problem of interest by solving a more general (harder) problem as an intermediate step
- Clear Distinction between

  - problem setting
  - solution approach (inductive principle)
  - learning algorithm

# Alternative Formulations

- Re-examine assumptions behind standard inductive learning

1 Finite training + large unknown test set

→ non-inductive inference (transduction, ...)

2 Particular loss functions

→ new inductive formulations (application-driven)

3 Single model

→ multiple model estimation

# 1.Transduction

- How to incorporate unlabeled test data into the learning process
- Estimating function at given points

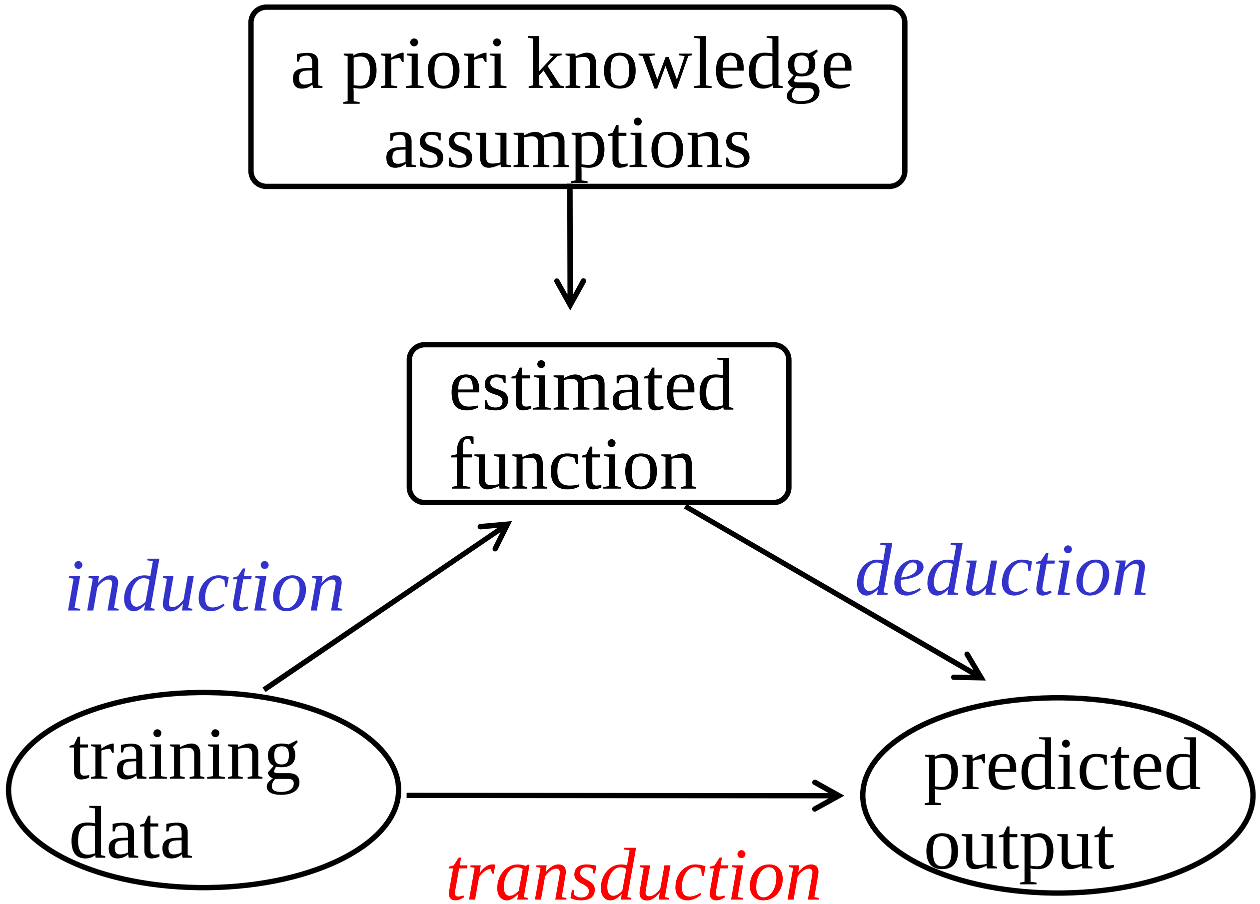
*Given:* training data ($\mathbf{X}$i, yi) , i = 1,...n

and unlabeled test points $\mathbf{X}$n+j , j = 1,...k

*Estimate:* class labels at these test points

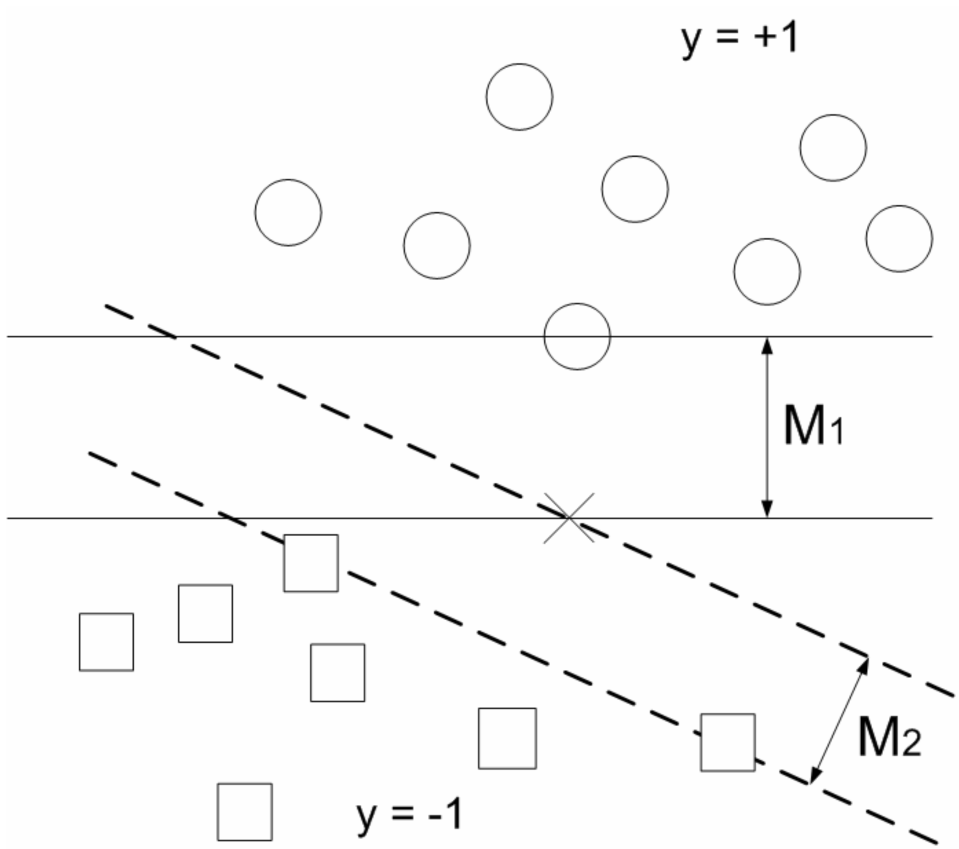
*Note:* need to predict only at given test points $\mathbf{X}$n+j, not for every possible input $\mathbf{X}$

# Transduction vs Induction



# Transduction based on size of margin

*The problem:* Find class label of test input X



# Many potential applications

- Prediction of molecular bioactivity for drug discovery
- Training data~1,909; test~634 samples
- Input space ~ 139,351-dimensional
- Prediction accuracy:

SVM induction~74.5%; transduction ~ 82.3%

*Ref:* J. Weston et al, KDD cup 2001 data analysis: prediction of molecular bioactivity for drug design – binding to thrombin, *Bioinformatics 2003*

# Beyond Transduction: Selection

- Selection Problem

*Given:* training data ($\mathbf{X}$i, yi) , i = *1,…n*

and unlabeled test points $\mathbf{X}$n+j , j = *1,…k*

*Select:* a subset of *m* test points with the highest probability of belonging to one class

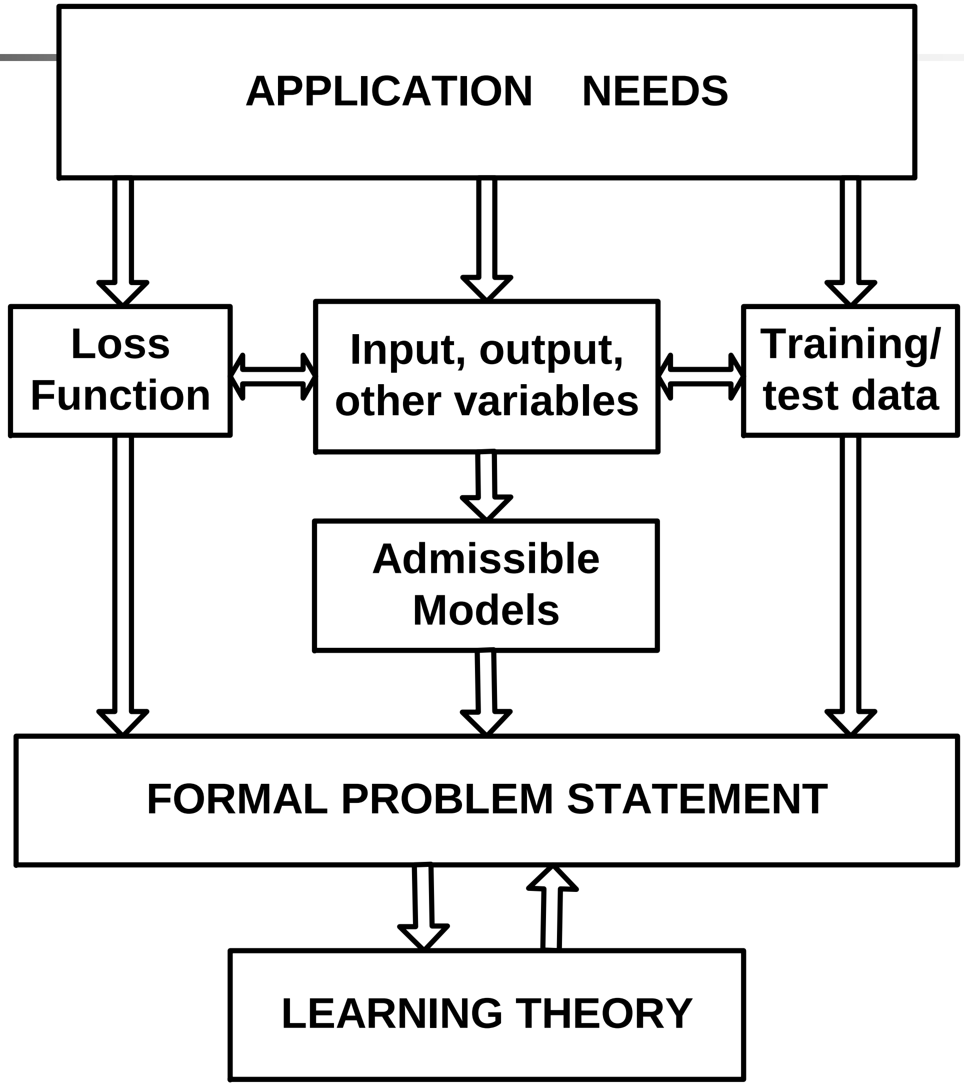
*Note:* selective inference needs only to select a subset of *m* test points, rather than assign class labels to all test points.

# Hierarchy of Types of Inference

- Identification
- Imitation
- Transduction
- Selection
- .....

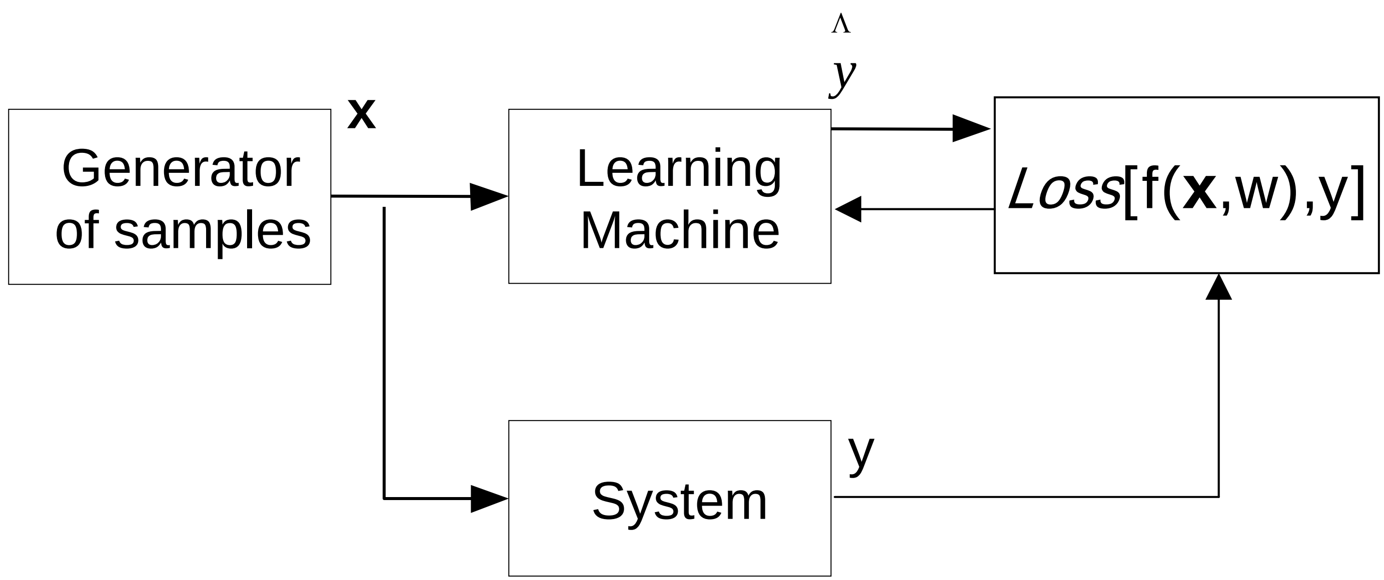
*Implications:* philosophical, human learning

# 2. Application-driven formulations



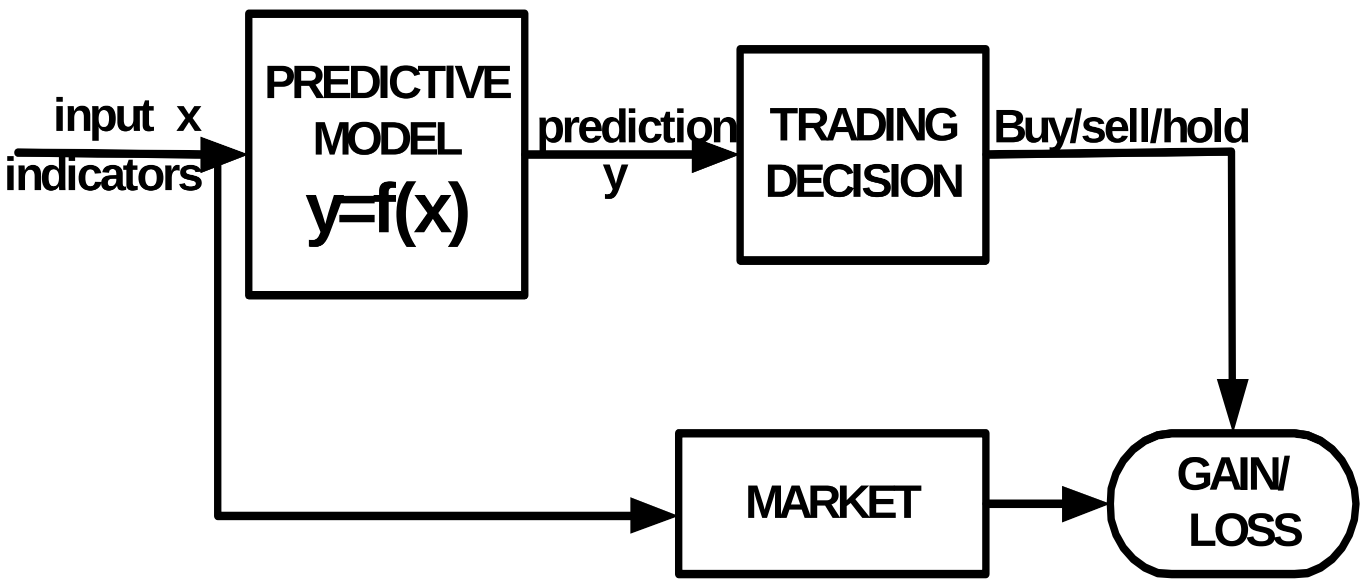
# Inductive Learning System (revised)

- The learning machine observes samples (**x** ,$y$), and returns an estimated response $\hat{y}$ to minimize application-specific *Loss* [f(**x**,w), y]



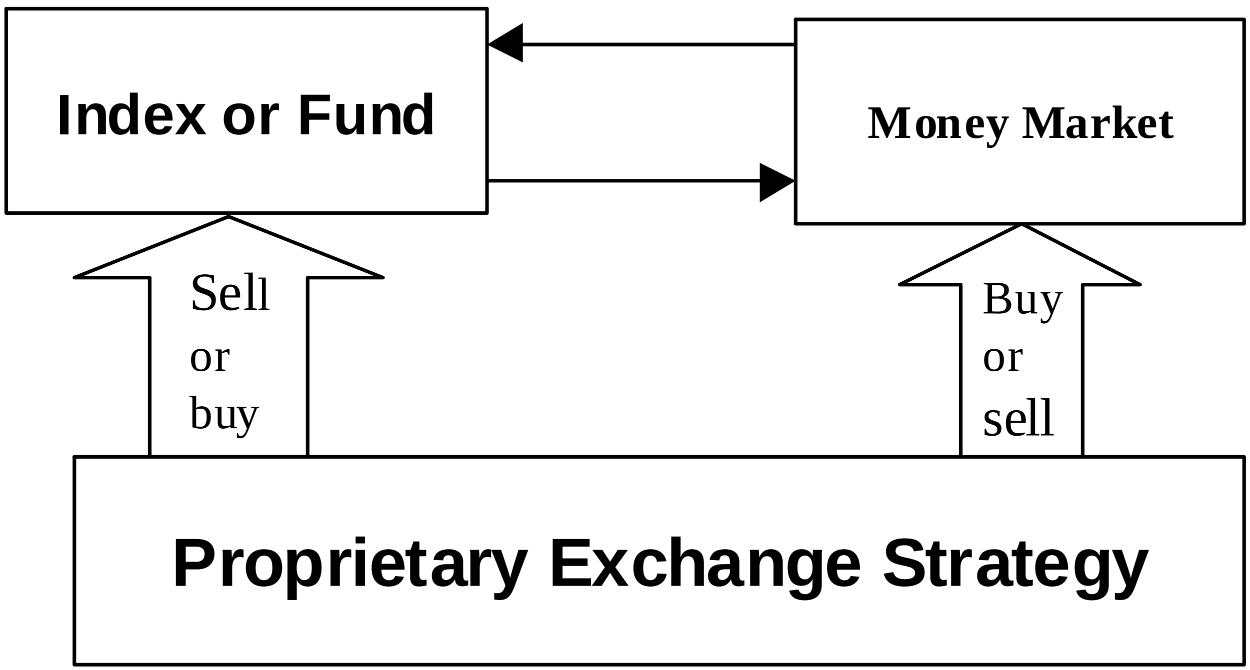
# Application: financial engineering

- Asset management via daily trading:
  non-standard learning formulation



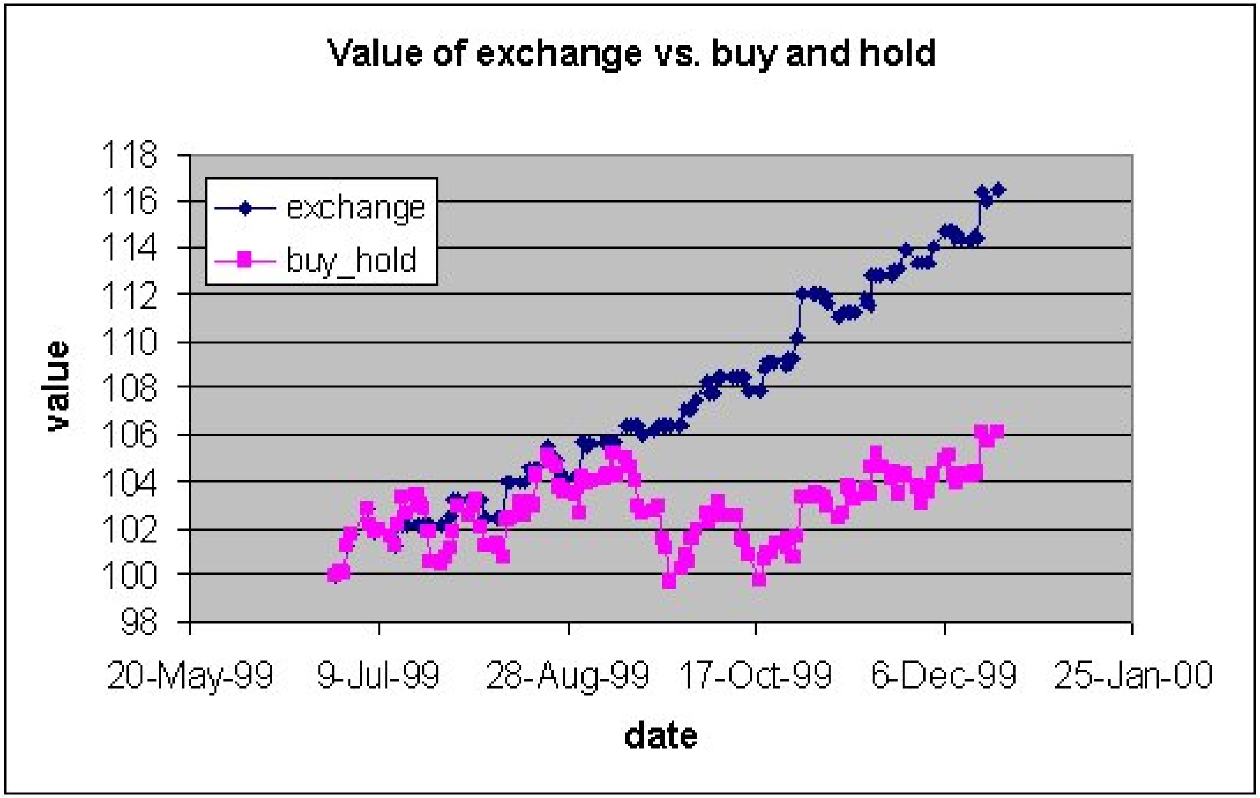
# Example: timing of mutual funds

- Background: buy-and-hold vs trading
- Recent scandals in mutual fund industry
- Daily trading scenario



# Example of Actual Trading

- Improved return + Reduced risk/ volatility:



# Learning formulation for fund trading

Given

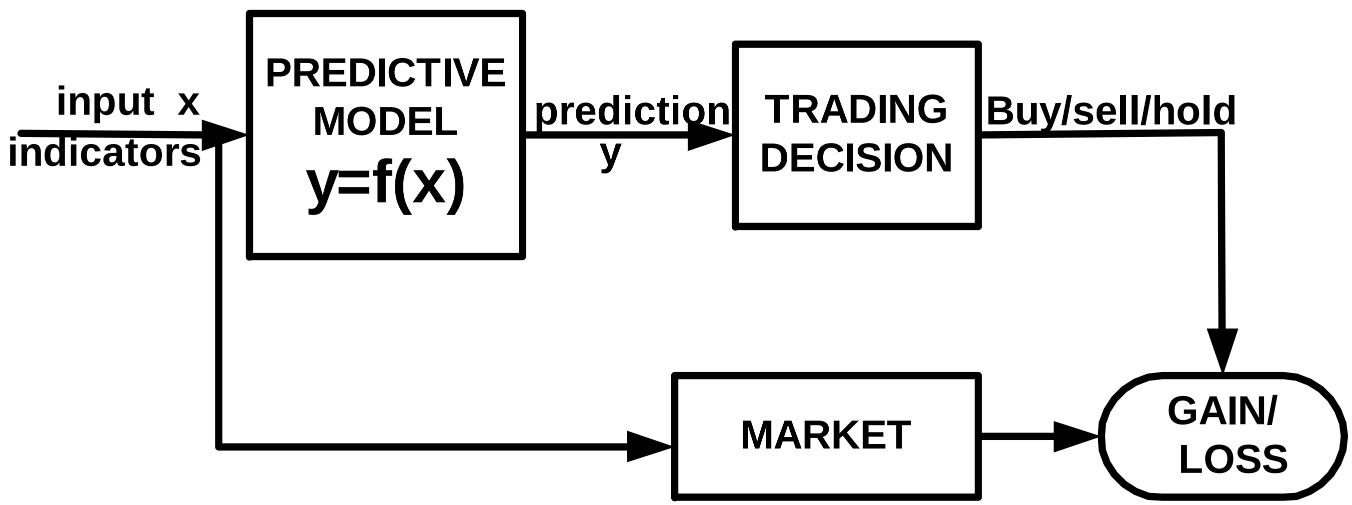
- Daily % price changes of a fund $q_i = \left(p_i - p_{i-1}\right) / p_i$
- Time series of daily values of input variables $X_i$
- Indicator decision function (1/0 ~ Buy/Sell) $y_i = f(x_i, w)$

Objective: maximize total return over n-day period

$$Q(w) = \sum_{i=1}^{n} f(x_i, w) q_i$$

# Non-standard inductive formulation

- Maximize total account value $Q(w) = \sum_{i=1}^{n} f(x_i, w) q_i$
- Neither classification, nor regression

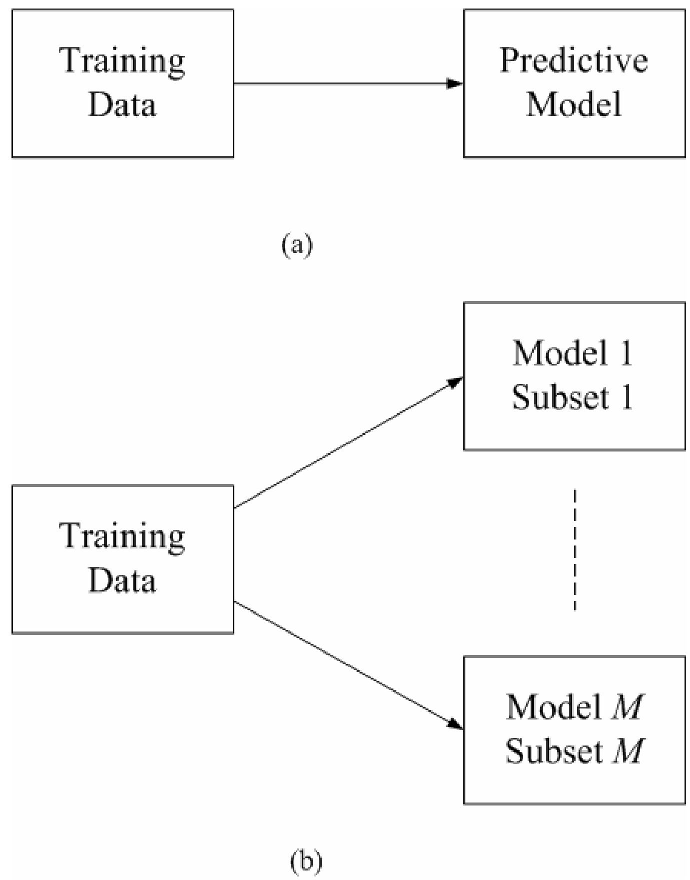


# 3. Multiple Model Estimation

- Single-model formulation
  - Estimate unknown dependency

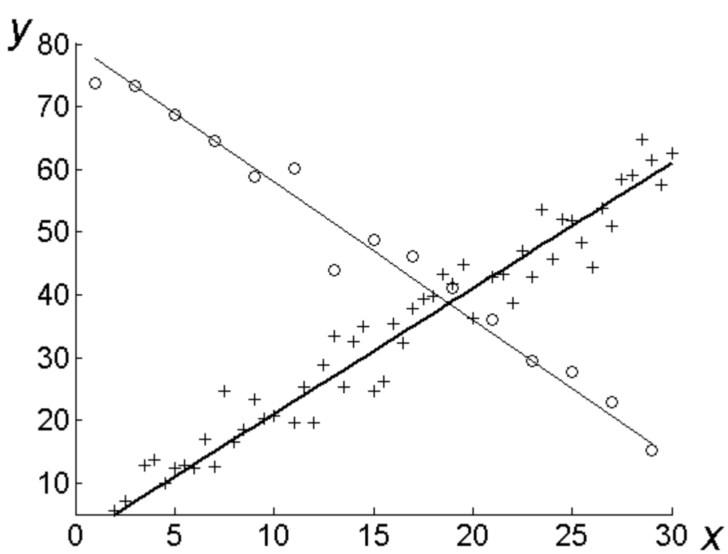
$$\mathbf{x} \rightarrow y$$

- Multiple-model approach:
  - Available data can be 'explained' using several models

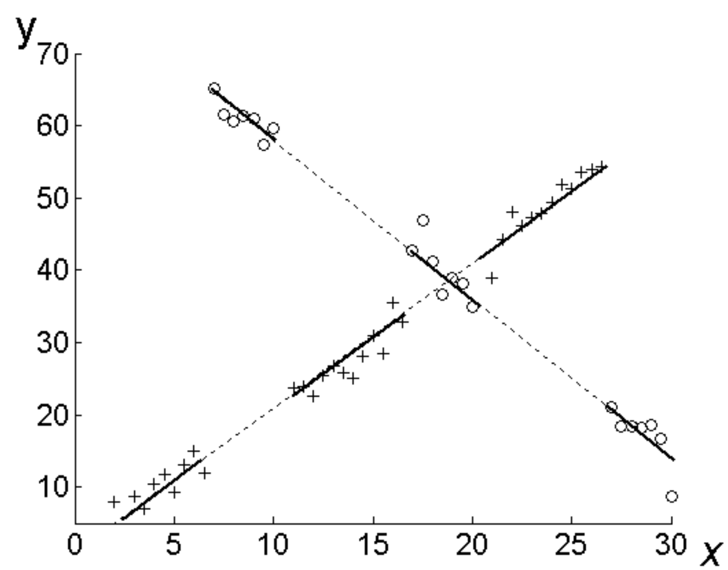


# Example data sets: Regression

- Two regression models




- Single complex model



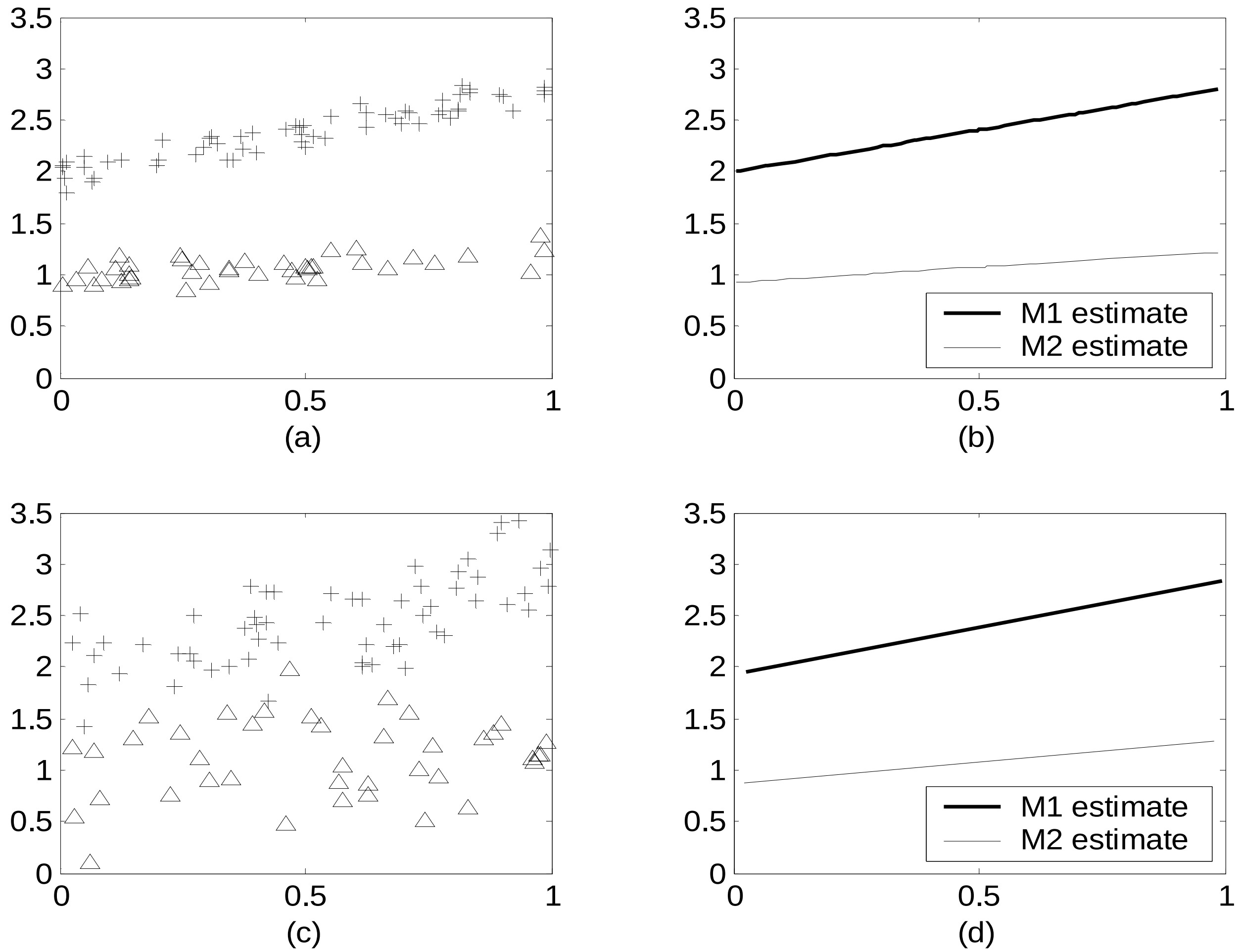
# Multiple Model Formulation

- Available (training) data are generated by several (unknown) regression models,

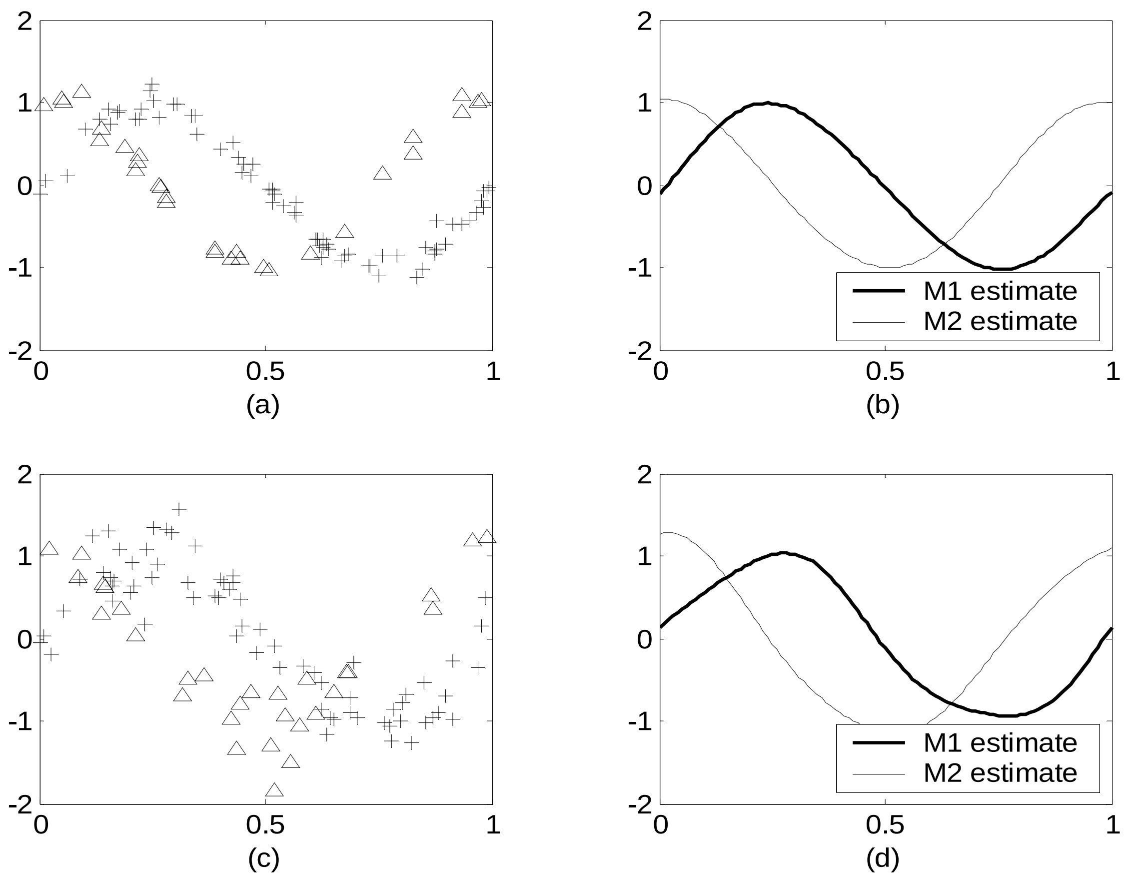
$$y = t_m(\mathbf{x}) + \xi_m \qquad\qquad \mathbf{x} \in X_m$$

- Goals of learning:
  - Partition available data (clustering, segmentation)
  - Estimate a model for each subset of data (supervised learning)
- Assumption:
  - Majority of the data samples can be explained (described) by a single model.

# Experimental Results: Linear

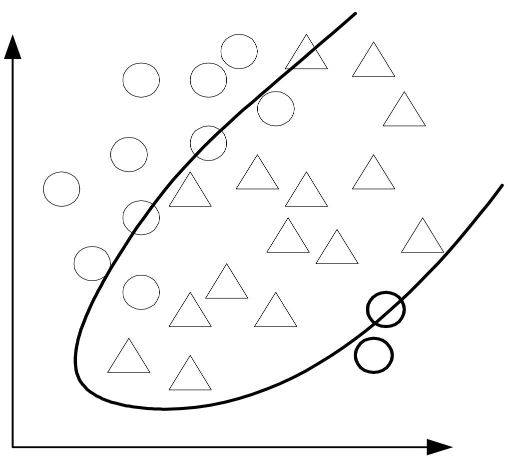


# Experimental Results: Non-Linear

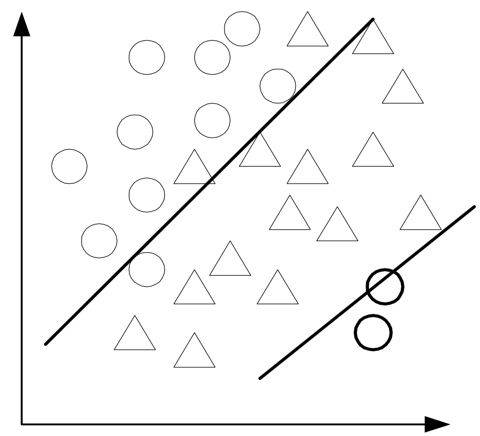


# Multiple Model Classification

- Single-model approach
  - →complex model



- Multiple-model approach
  - → two simple models

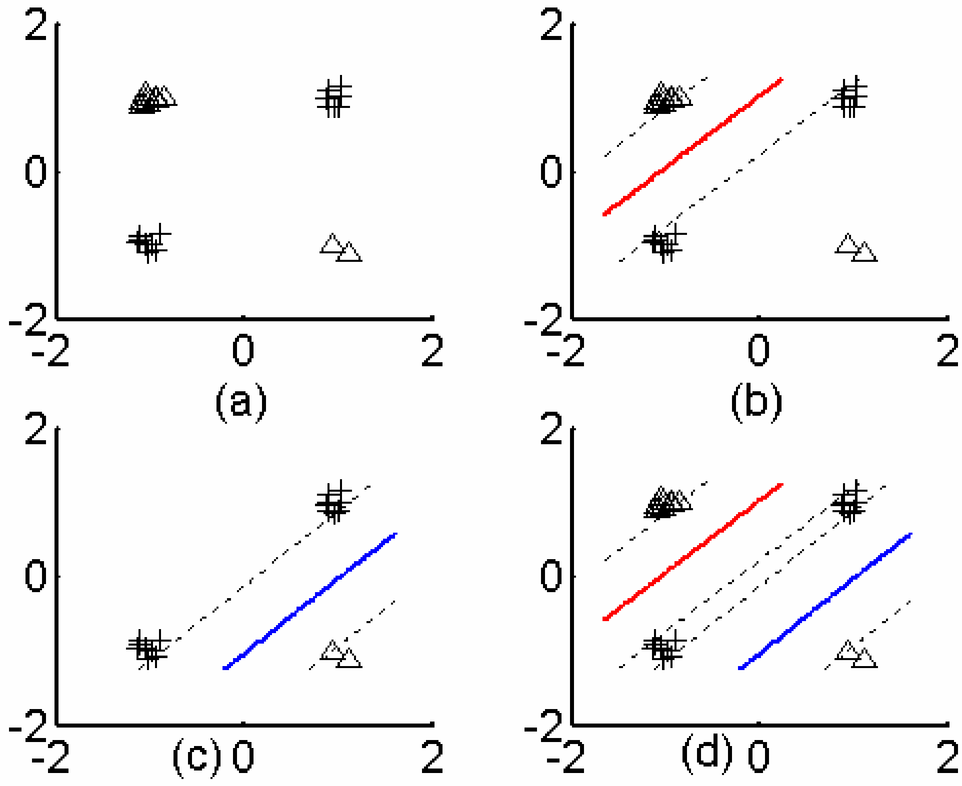


# Procedure for MMC

- *Initialization:* Available data = all training samples.
- Step 1: **Estimate major model**, i.e. apply robust classification to available data
  - Here, 'Robustness' wrt variations of data generated by minor model (s)
- Step 2: **Partition available data** (from one class) into two subsets
- Step 3: **Remove subset of data** (from one class) classified by the major model from available data.
- *Iterate*

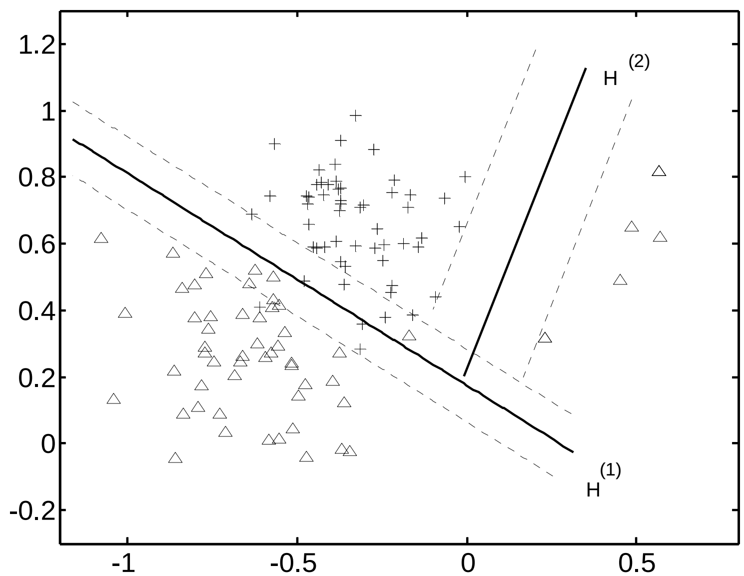
# Example of MMC: XOR data set

- Training phase

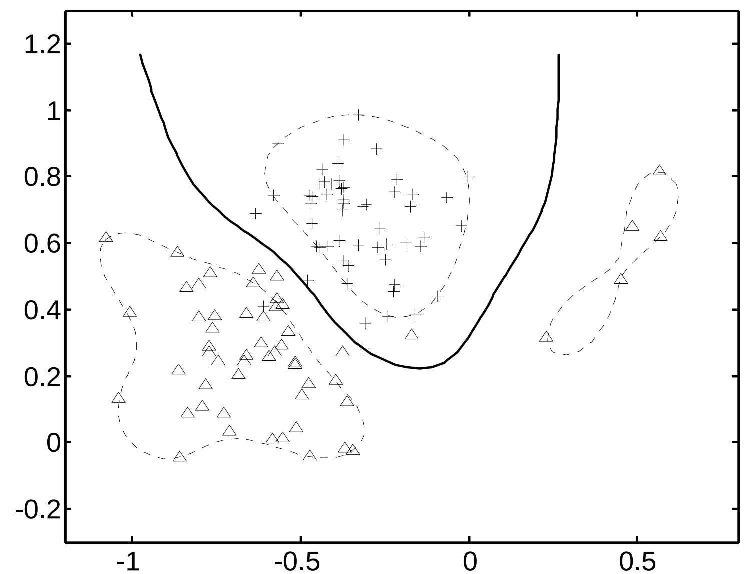


# Comparison for toy data set

- MMC hyperplanes
- RBF-SVM



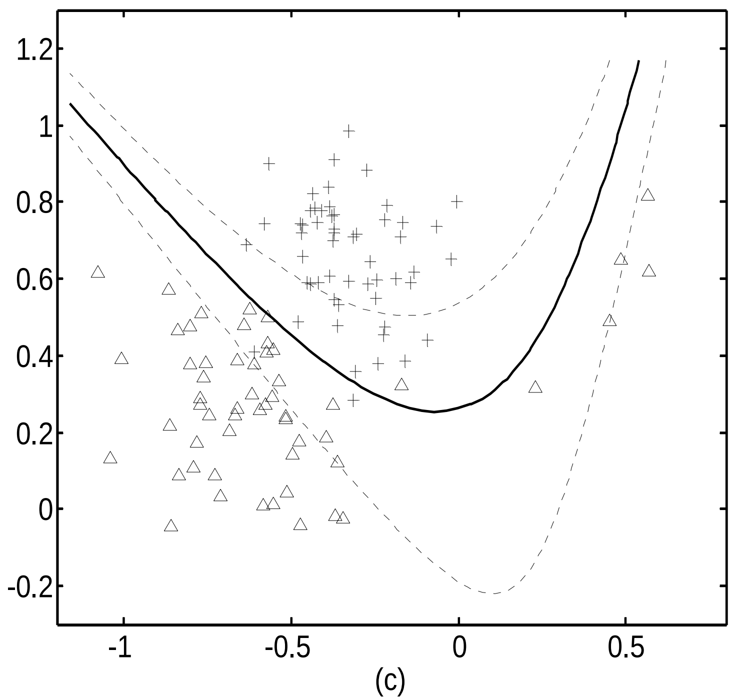

(a)




(b)

# Comparison continued

- SVM polynomial kernel




- Prediction Accuracy

Error (%SV)

- RBF 0.058 (25.5%)
- Poly 0.067 (26.4%)
- MMC 0.055 (14.5%)

# Summary for Multiple Model Estimation

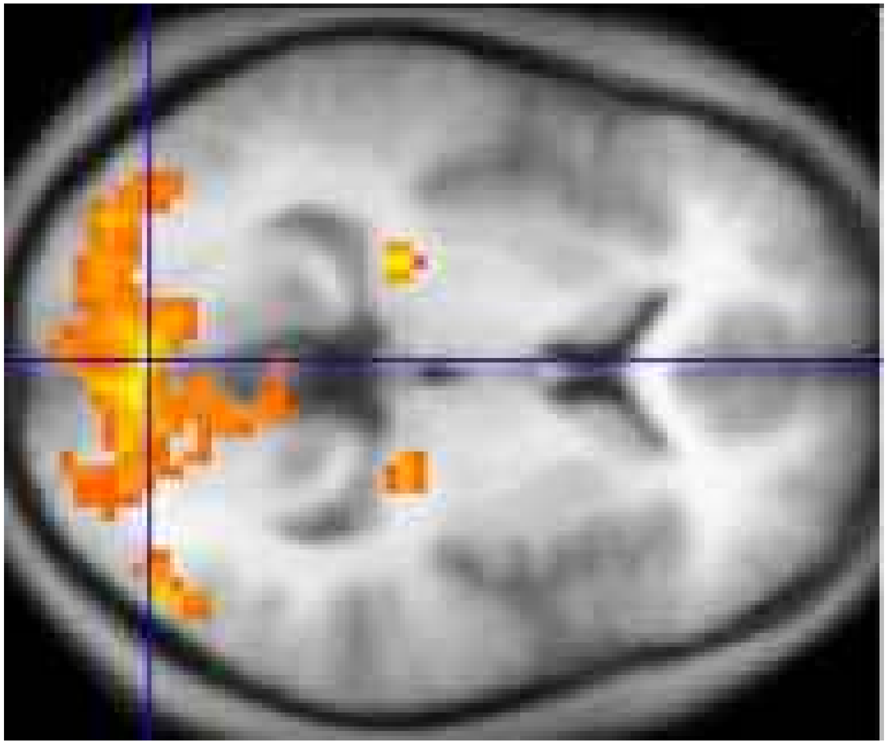
- Improvements due to novel problem formulation, not sophisticated algorithms
- Practical learning algorithm using based on (linear) SVM
- Resulting model has hierarchical structure
- Advantages:
  - Interpretation
  - No Kernel Selection

# Prediction and interpretation

- Many, many applications intrinsically difficult to formalize
- Two practical goals of learning:
  - prediction (objective loss function)
  - interpretation, understanding (subjective)
- Most algorithms developed for *predictive* settings, but used for *interpretation* and human *decision making*
- *Rationale:* good predictive model ~ true

# Example:functional neuroimaging

- Understanding fMRI image data:
  - estimate ‘good’ Brain Activation Maps showing brain activity (colored patches) in response to specific tasks
- Measure of goodness: predictability, reproducibility



# Predictive models for understanding

- Always assume inductive formulation
- What if transduction yields much better prediction?
- Fundamental problem (classical view):

  - human reasoning ~ *logic + induction*

  - transduction does not fit this paradigm
- Goal of science: understanding
- Goal of science: perform/act well

# Conclusions

- **Methodological shift:**
  *think first* about the problem formulation, rather than learning algorithms
- **Importance of problem formulation**
  - for empirical comparisons
  - the limits of predictive models
- **Philosophical impact** of Vapnik's new types of (non-inductive) inference

# References


- VC-Theory: V. Vapnik, Statistical Learning Theory, Wiley, NY
- Transduction: V. Vapnik (1998), Statistical Learning Theory, Wiley, + many recent papers
- Timing of Mutual Funds: E. Zitzewitz (2002), Who cares about shareholders: arbitrage-proofing mutual funds. Journal of Law, Economics and Organization, 19, 2, pp. 245-280
- Multiple Model Estimation:

  Y. Ma and V. Cherkassky (2003), Multiple model classification using SVM-based approach, in Proc. IJCNN

  V. Cherkassky and Y. Ma (2005), Multiple model regression estimation, IEEE TNN, 14, 4, pp. 785-798