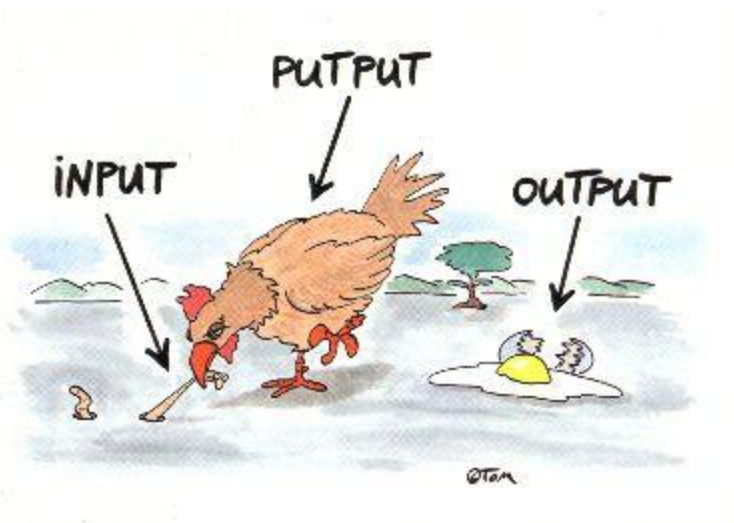# DATA MINING IN A NUTSHELL

Statistics   vs.   Data Mining

- Macro-decisioning

- Explain/describe population relationships

- Small sample, few variables

- Find good fitting statistical model

- Confidence intervals, hypothesis test, p-value

- Micro-decisioning

- Predict values of new records

- Large sample, many variables

- Models/algorithms with high predictive power

- Predictive power metrics and costs

Types of Methods

"Supervised Learning"

Prediction (numerical Y)

Classification (categorical Y)

"Unsupervised Learning"

Dimension Reduction

Segmentation

"What goes with what?"

The Data Mining Process

Define purpose	Obtain data	Explore & clean data	Determine DM task	Choose DM Methods	Apply methods, select final model	Evaluate performance	Deploy

Data Preparation & Exploration (2-3)

- Sampling
- Cleaning
- Summaries
- Visualization
- Partitioning
- Dimension reduction

Prediction
- MLR (5)
- K-nearest neighbor (6)
- Regression trees (7)
- Neural Nets (9)

Classification
- K-nearest neighbor (6)
- Naive Bayes (6)
- Logistic regression (8)
- Classification trees (7)
- Neural nets (9)
- Discriminant analysis (10)

Segmentation/ clustering (12)

Affinity Analysis / Association Rules (11)

Model evaluation & selection (4)

Scoring new data

Deriving insight