Introduction to Data Mining II

Statistical Data Mining II

Spring 2016

Rachael Hageman Blair

What is Data Mining?

Data mining: Tools, methodologies and theories for revealing patterns in the data – a critical step in knowledge discovery.

Driving forces (BIG DATA):

- Explosive growth of data in a variety of fields
  - Cheaper storage devices with high capacity
  - Faster Communications
  - Better database management systems
- Better Computing Power
- More emphasis on 'making the data work for us'.

Also known as … Multivariate Analysis

Multivariate Analysis: the simultaneous statistical analysis of a collection of random variables.

Origins: social, behavioral sciences, agriculture, biology, astrology.

- Factor Analysis -> explain psychological theories of human behavior.
- PCA -> analyze student scores over a battery of tests concerning psychological measurement.
- Discriminant analysis -> classification based on botanical measurements.
- Regression and correlation -> heredity and the orbits of the planets.

Research Fields

- Statistics
- Bioinformatics
- Signal Processing
- Machine Learning
- Pattern Recognition
- Computer Science
- Databases
- Geography

Two flavors of data mining

- **Descriptive data mining:** Search massive data sets and discover the locations of unexpected structures or relationships, patterns, trends, clusters, and outliers in the data.

- **Predictive data mining:** Build models and procedures for regression, classification, pattern recognition, or machine learning tasks, and assess the predictive accuracy or those models and procedures when applied to fresh data.

Two flavors of data mining

- **Descriptive data mining:** Search massive data sets and discover the locations of unexpected structures or relationships, patterns, trends, clusters, and outliers in the data.

- **Predictive data mining:** Build models and procedures for regression, classification, pattern recognition, or machine learning tasks, and assess the predictive accuracy or those models and procedures when applied to fresh data.

Terminology

Notation:

- **Input X:** X is often multi-dimensional and in matrix form. Each dimension is denoted by X_j and is referred to as a feature, predictor, or independent variable.

  - **Output Y:** response, dependent variable.

  Input and Output may be of different variable type:

  - **Quantitative variable:** cholesterol levels, height.
  - **Qualitative variable:** flower species, low/high risk.
  - **Ordered categorical:** small, medium, and large.

  Why does it matter?

  - Supervised Learning vs. Unsupervised learning

    Is Y available?
  - Regression vs. Classification

    Is Y qualitative or quantitative?

Where we have been, and where we are going?

Supervised Learning

Unsupervised Learning

In reality….

We often re-cast unsupervised problems as supervised.
Why?

- We have a well formulated problem.
- We can estimate loss.
- We can leverage more robust techniques, less heuristics.

In reality….

We often re-cast unsupervised problems as supervised.
Why?

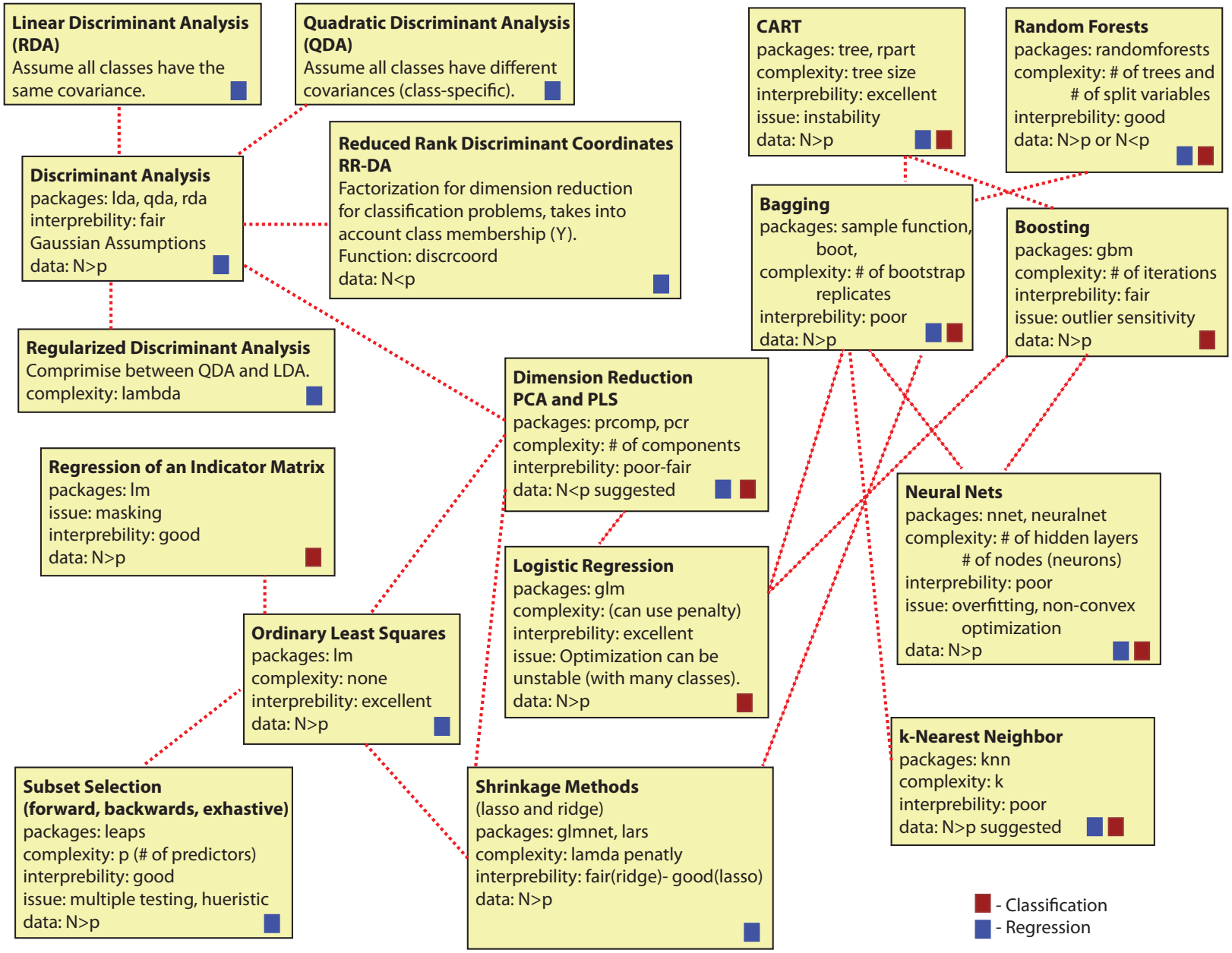
- We have a well formulated problem.
- We can estimate loss.
- We can leverage more robust techniques, less heuristics.

A wolf in sheep's clothing….

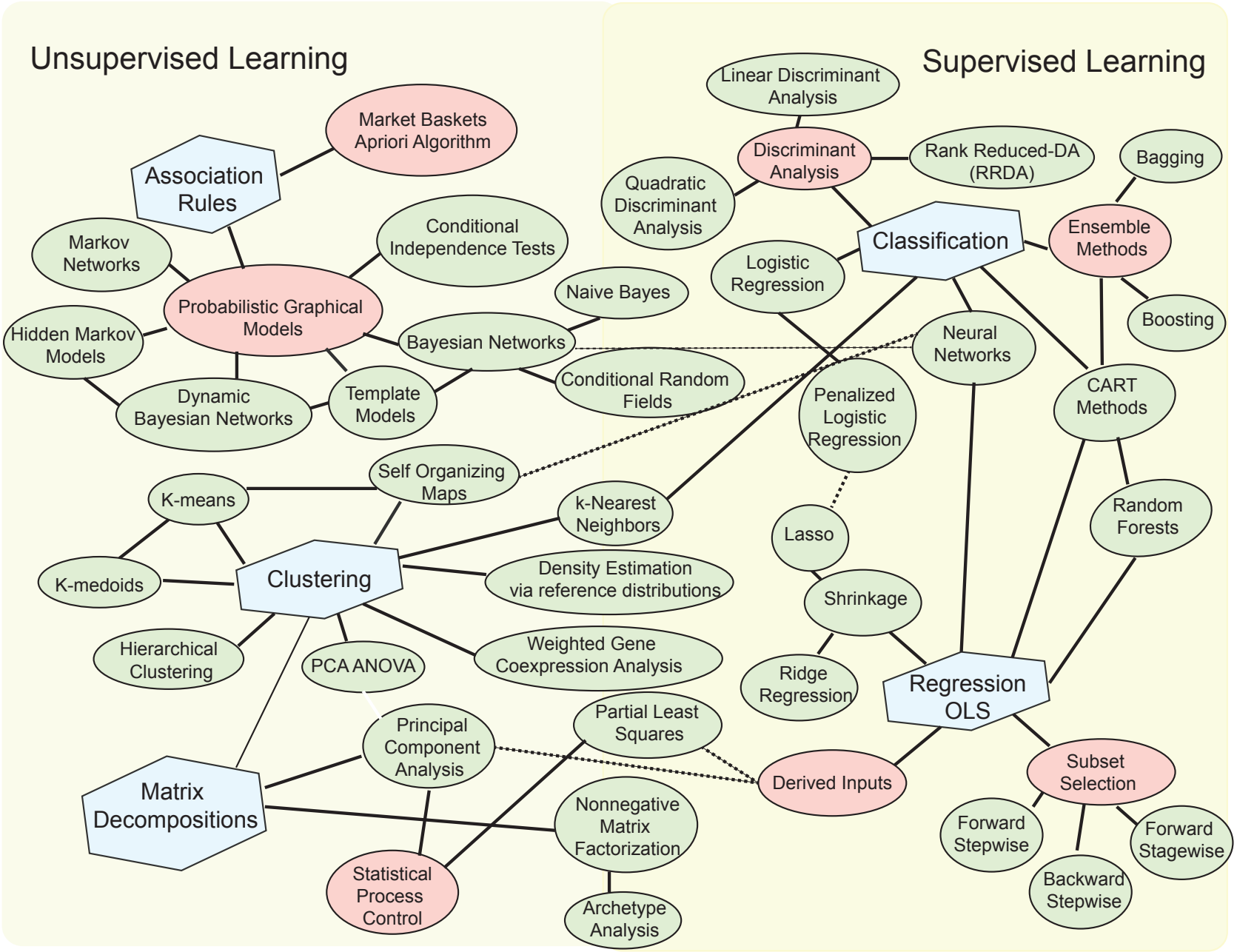
“Sheep”
Supervised learning

“Wolf”
Unsupervised learning

Data Mining I



Data Mining I & II



A Few Applications …..

- Business
  Insurance Companies, Banks
  Internet Marketing ~ suggested buys (e.g., Amazon).

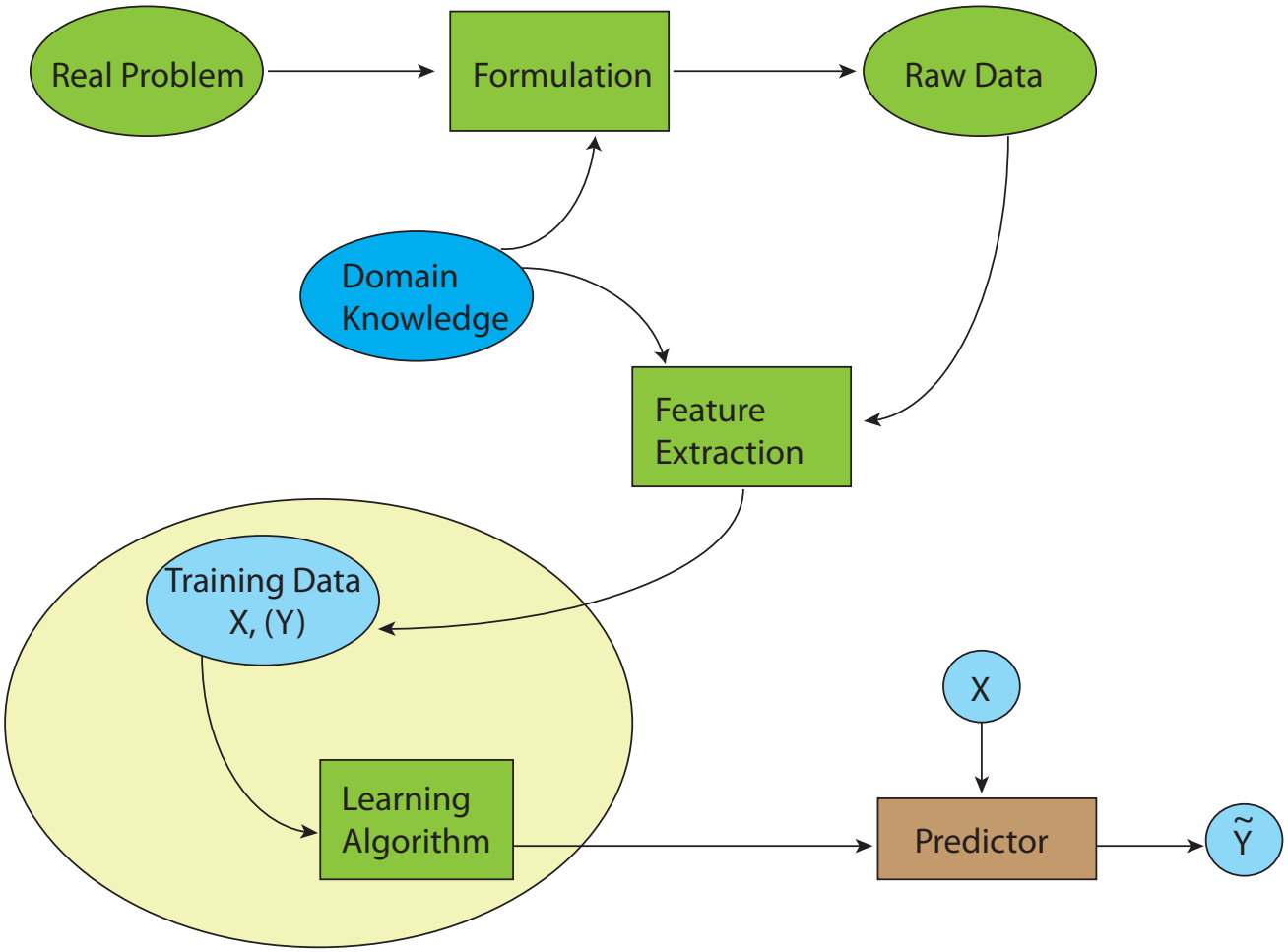
- Life Sciences
  High-throughput Analysis
  Medical Imaging

- Communication Systems
  Speech recognition
  Image Analysis

- Social Networks
  Friend and group recommendations.
  Terrorist Links

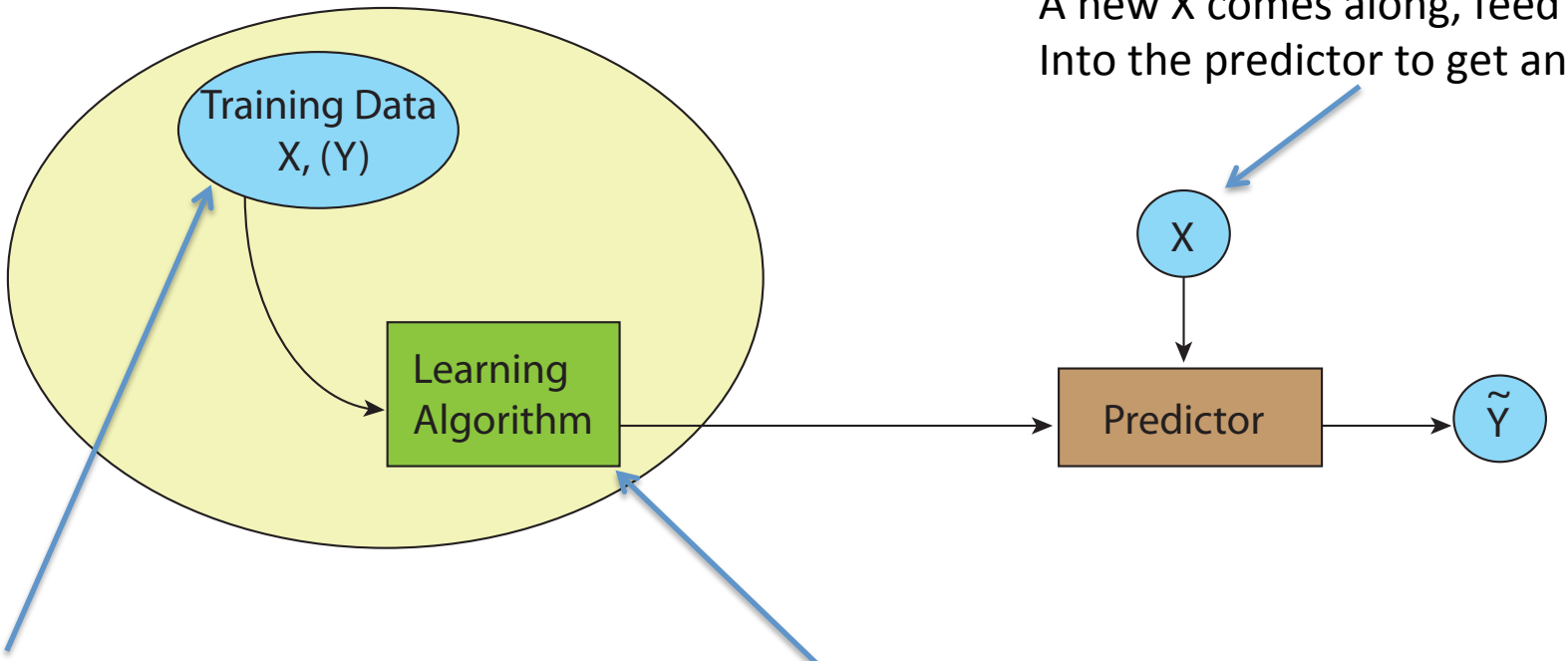
- **Marketing:** Predict new purchasing trends. Predict what types of customers will respond to direct mailings, telemarketing, promotions. Given customers that have purchased product A, B, or C, identify those who are likely to purchase product D in general. Which products are likely to sell together (market basket analysis).
- **Banking**: Predict which customers will likely switch credit card companies.
- **Financial models**: Bankruptcy prediction for small businesses.
- **Insurance:** Identify characteristics of buyers of new policies. Find unusual claims patterns Identify risky customers.
- **Healthcare:** Identify successful medical treatments and procedures by examining insurance claims and billing data. Identify people "at risk" for illness. Predict doctor visits from patient characteristics.
- **Biology:** Collect, organize and integrate massive amounts of bioinformatics, functional genomics, proteomics, gene expression and microarrays.
- **Sports:** Identify in realtime which players and which designed plays are most effective at specific points in the game and in relation to combinations of opposing players. Discover game patterns hidden behind summary statistics.

Birds Eye View: Prediction



Functional Approximation

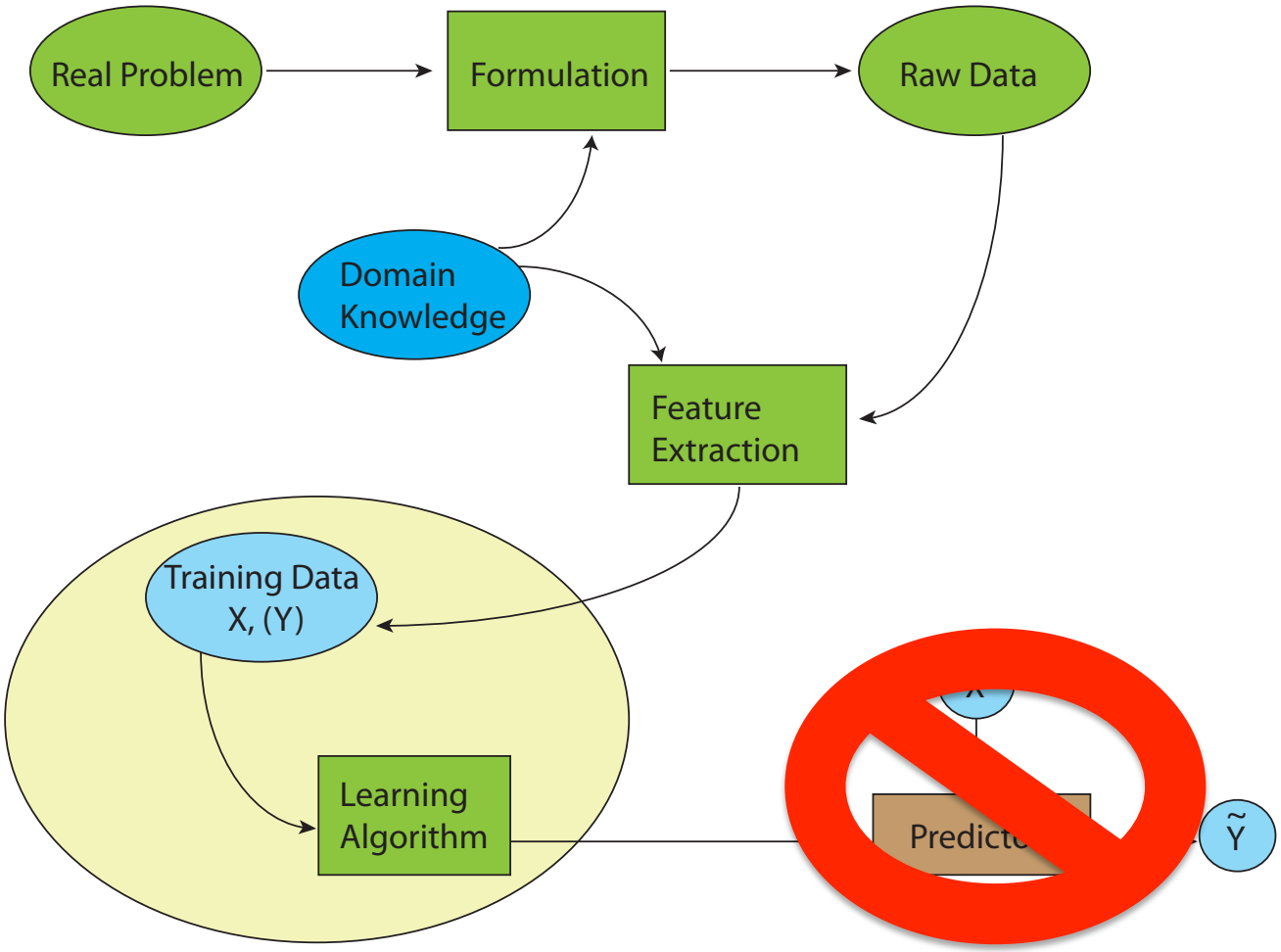
Supervised Learning with the additive error model $Y = f(x) + \varepsilon$



f(x) may or may not be deterministic, we only observe X and Y. Assume we have no prior knowledge

Feed X and Y to the machine, it will tell you the $\hat{f}(x)$ that minimizes the residual sum of squares. We "learn" the optimal predictor.

Birds Eye View: Descriptive



Where we are going ….

- Association
- Hierarchical and Agglomerative clustering
- Matrix decompositions and factorizations (PCA, ICA, NMF)
- Self organizing maps
- Image Processing (Face recognition)
- Undirected graphical models
- Directed graphical models
- High dimensional considerations

Data Mining Methods

Scalability: The capability of an algorithm to remain efficient and accurate as we increase the complexity of the problem.

- One objective of data mining is to create a library of scalable algorithms for the statistical analysis of large data sets. Purpose is prediction via inference.

- Classical statistical inference may have no meaning or dubious validity – e.g., the entire population is searched for answers, or poor sampling.

Data Mining Methods

Other ingredients of successful techniques:

- Computational Efficiency
- Automation of technique and data processing
- Dynamic and interactive data visualization
- Algorithmic development

- Prediction Accuracy
- Avoiding overfitting
- Sensitivity
- Generalization

Data Mining Methods

Knowledge, Discovery, and Databases (KDD)

Consists of the following activities:

1. Selection of the target set (which data and/or variables are to be used for data mining).
2. Data Cleaning and Preprocessing (noise removal, outlier identification, imputation for missing data).
3. Preprocessing the data (transformations, tracking time-dependent information and relevant covariates).
4. Deciding which data mining techniques are appropriate (regression, classification, clustering, graphical modeling).
5. Analyze the clean data using data-mining software (algorithms for data reduction, dimensionality reduction, model fitting, prediction extraction).
6. Interpreting and assessing the knowledge derived from data-mining results.

Machine Learning

Machine learning – evolved out of artificial intelligence (AI), a branch of computer science, which aims to make "machines intelligent", in other words to think rationally like humans and solve problems.

- A machine learns through experience, it accumulates experience through data, and develops knowledge to complete tasks, with improved performance over time.

- Intelligence cannot be achieved without learning, therefore, machine learning plays a dominate role in AI.