# Lecture 0

# Overview of Data Mining and Knowledge Discovery in Databases (KDD)

**Monday, May 15, 2000**

**William H. Hsu**

**Department of Computing and Information Sciences, KSU**

http://www.cis.ksu.edu/~bhsu

Recommended Reading:

KDD Intro, U. Fayyad

Chapter 1, *Machine Learning*, T. M. Mitchell

*MLC++ Tutorial*, R. Kohavi and D. Sommerfield

# Course Outline

- **Overview: Knowledge Discovery in Databases (KDD) and Applications**
- **Artificial Intelligence (AI) Software Development Topics**
  - **Data mining and machine learning**
  - **Simple, common data mining models**
    - **Association rules**
    - **Simple Bayes**
  - **Intermediate and advanced models**
    - **Artificial neural networks (ANNs) for KDD**
    - **Simple genetic algorithms (GAs) for KDD**
- **Practicum (Short Software Implementation Project)**
  - **High-performance data mining systems ("HPC for KDD")**
    - **HPC platform: Beowulf**
    - **Codes: *NCSA D2K, MLC++,* other (*MineSet, JavaBayes, GPSys, SNNS*)**
  - **<u>Stages of KDD</u> and practical software engineering issues**
  - **Implementing learning and visualization modules**

# Questions Addressed

- Problem Area
  - What *are* data mining (DM) and knowledge discovery in databases (KDD)?
  - Why are we doing DM?
- Methodologies
  - What kind of software is involved? What kind of math?
  - How do we develop it (software, repertoire of statistical models)?
  - Who does DM? (Who are practitioners in academia, industry, government?)
- Machine Learning as Model-Building Stage of DM
  - What is machine learning (ML) and what does it have to do with DM?
  - What are some interesting problems in DM, KDD?
  - *Should I be interested in ML (and if so, why)?*
- Brief Tour of Knowledge-Based Systems (KBS) Topics
  - *Knowledge and data engineering* (KDE) for KDD
  - *Knowledge-based software engineering* (KBSE)
  - Expert systems and human-computer intelligent interaction (HCII)

# Why Knowledge Discovery in Databases?

- **New Computational Capability**
  - **Database mining: converting (technical) records into knowledge**
  - **Self-customizing programs: learning news filters, adaptive monitors**
  - **Learning to act: robot planning, control optimization, decision support**
  - **Applications that are hard to program: automated driving, speech recognition**
- **Better Understanding of Human Learning and Teaching**
  - **Cognitive science: theories of knowledge acquisition (e.g., through practice)**
  - **Performance elements: reasoning (inference) and *recommender* systems**
- **Time is Right**
  - **Recent progress in algorithms and theory**
  - **Rapidly growing volume of online data from various sources**
  - **Available computational power**
  - **Growth and interest of learning-based industries (e.g., data mining/KDD)**

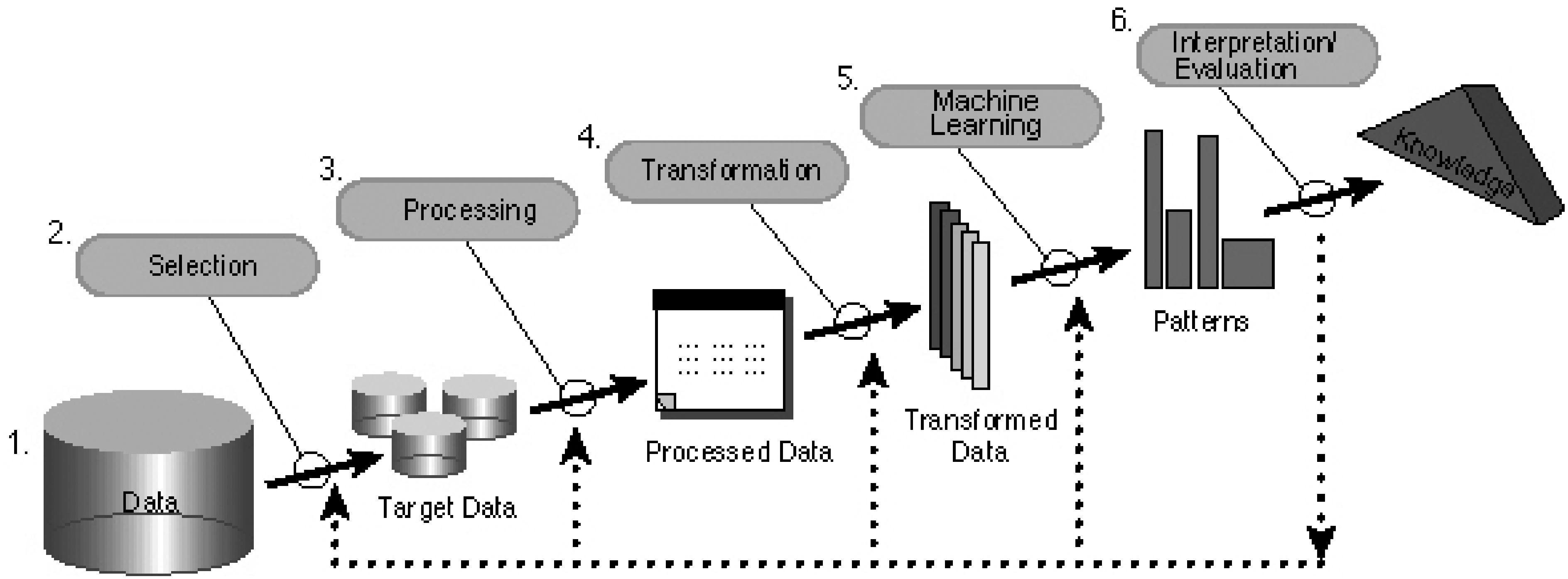
# What Are KDD and Data Mining?

- **Two Definitions (FAQ List)**
  - **The process of automatically extracting valid, useful, previously unknown, and ultimately comprehensible information from large databases and using it to make crucial business decisions**
  - ***"Torturing the data until they confess"***
- **KDD / Data Mining: An Application of Machine Learning**
  - **Guides and integrates learning (model-building) processes**
    - **Learning methodologies: supervised, unsupervised, reinforcement**
    - **Includes preprocessing (data cleansing) tasks**
    - **Extends to pattern recognition (inference or *automated reasoning*) tasks**
  - **Geared toward such applications as:**
    - **Anomaly detection (fraud, inappropriate practices, intrusions)**
    - **Crisis monitoring (drought, fire, resource demand)**
    - **Decision support**
- **What Data Mining Is *Not***
  - **<u>D</u>ata <u>B</u>ase <u>M</u>anagement <u>S</u>ystems: *related but not identical field***
  - **"Discovering objectives": still need to *understand performance element***



**Kansas State University**
**Department of Computing and Information Sciences**

# Stages of KDD




An Overview of the Steps That Compose the KDD Process

# Rule and Decision Tree Learning

- **Example: Rule Acquisition from Historical Data**
- **Data**
  - **Customer 103 (visit = 1): Age 23, Previous-Purchase: no, Marital-Status: single, Children: none, Annual-Income: 20000, Purchase-Interests: *unknown*, Store-Credit-Card: no, Homeowner: *unknown***
  - **Customer 103 (visit = 2): Age 23, Previous-Purchase: no, Marital-Status: married, Children: none, Annual-Income: 20000: Purchase-Interests: <u>car</u>, Store-Credit-Card: yes, Homeowner: no**
  - **Customer 103 (visit = n): Age 24, Previous-Purchase: <u>yes</u>, Marital-Status: married, Children: yes, Annual-Income: <u>75000</u>, Purchase-Interests: <u>television</u>, Store-Credit-Card: yes, Homeowner: no, Computer-Sales-Target: YES**
- **Learned Rule**
  - **IF *customer has made a previous purchase*, AND *customer has an annual income over $25000*, AND *customer is interested in buying home electronics***

    **THEN *probability of computer sale is 0.5***
  - **Training set: 26/41 = 0.634, test set: 12/20 = 0.600**
  - **Typical application: <u>target marketing</u>**



Kansas State University
Department of Computing and Information Sciences

# Text Mining: Information Retrieval and Filtering

- **20 *USENET* Newsgroups**
  - comp.graphics
  - comp.os.ms-windows.misc
  - comp.sys.ibm.pc.hardware
  - comp.sys.mac.hardware
  - comp.windows.x
  - misc.forsale
  - rec.autos
  - rec.motorcycles
  - rec.sports.baseball
  - rec.sports.hockey
  - soc.religion.christian
  - talk.politics.guns
  - talk.politics.mideast
  - talk.politics.misc
  - talk.religion.misc
  - alt.atheism
  - sci.space
  - sci.crypt
  - sci.electronics
  - sci.med
- **Problem Definition [Joachims, 1996]**
  - Given: 1000 training documents (posts) from each group
  - Return: classifier for new documents that identifies the group it belongs to
- **Example: Recent Article from *comp.graphics.algorithms***

Hi all

I'm writing an adaptive marching cube algorithm, which must deal with cracks. I got the vertices of the cracks in a list (one list per crack).

Does there exist an algorithm to triangulate a concave polygon ? Or how can I bisect the polygon so, that I get a set of connected convex polygons.

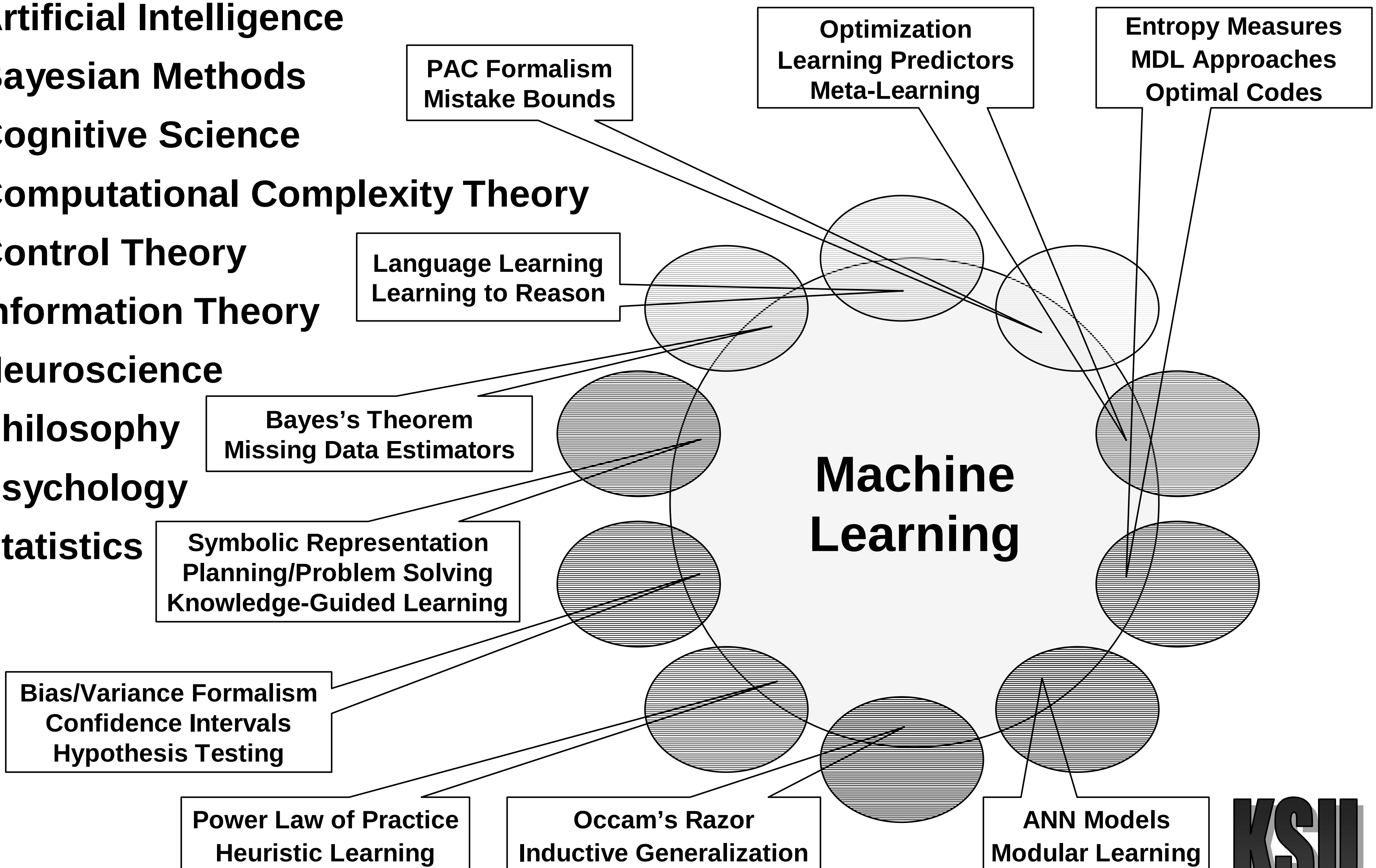
The cases of occuring polygons are these:

...

- **Performance of *Newsweeder* (Naïve Bayes): 89% Accuracy**

# Relevant Disciplines

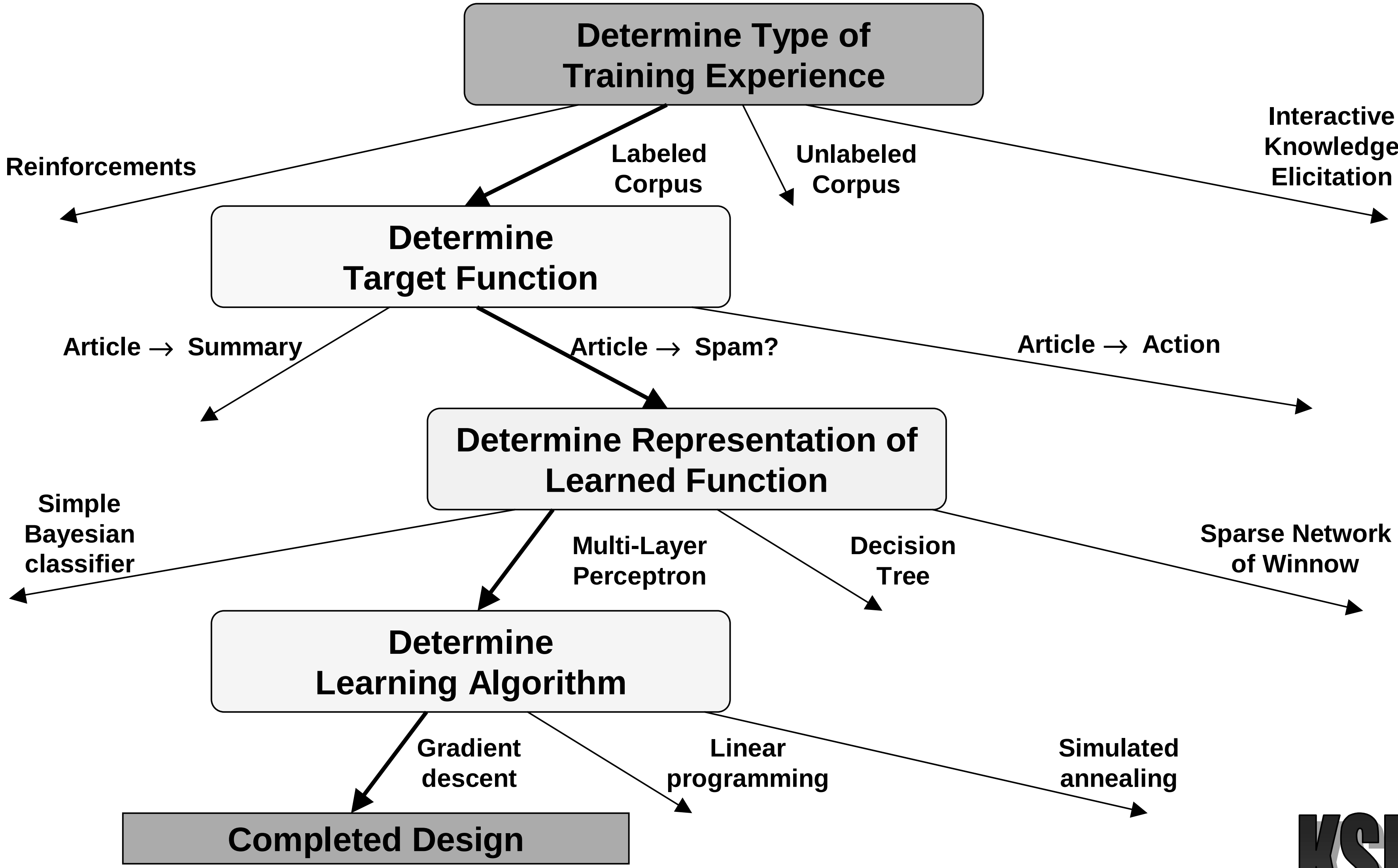
- Artificial Intelligence
- Bayesian Methods
- Cognitive Science
- Computational Complexity Theory
- Control Theory
- Information Theory
- Neuroscience
- Philosophy
- Psychology
- Statistics



# Specifying A Learning Problem

- **Learning = Improving with Experience at Some Task**
  - **Improve over task *T*,**
  - **with respect to performance measure *P*,**
  - **based on experience *E*.**
- **Example: Learning to Filter Spam Articles**
  - ***T*: analyze USENET newsgroup posts**
  - ***P*: function of classification accuracy (discounted error function)**
  - ***E*: <u>training corpus</u> of labeled news files (e.g., annotated from Deja.com)**
- **Refining the Problem Specification: Issues**
  - **What experience?**
  - **What *exactly* should be learned?**
  - **How shall it be *represented*?**
  - **What specific algorithm to learn it?**
- **Defining the Problem Milieu**
  - **Performance element: How shall the results of learning be applied?**
  - **How shall the performance element be evaluated? The learning system?**

# Design Choices and Issues in KDD



# Survey of Machine Learning Methodologies

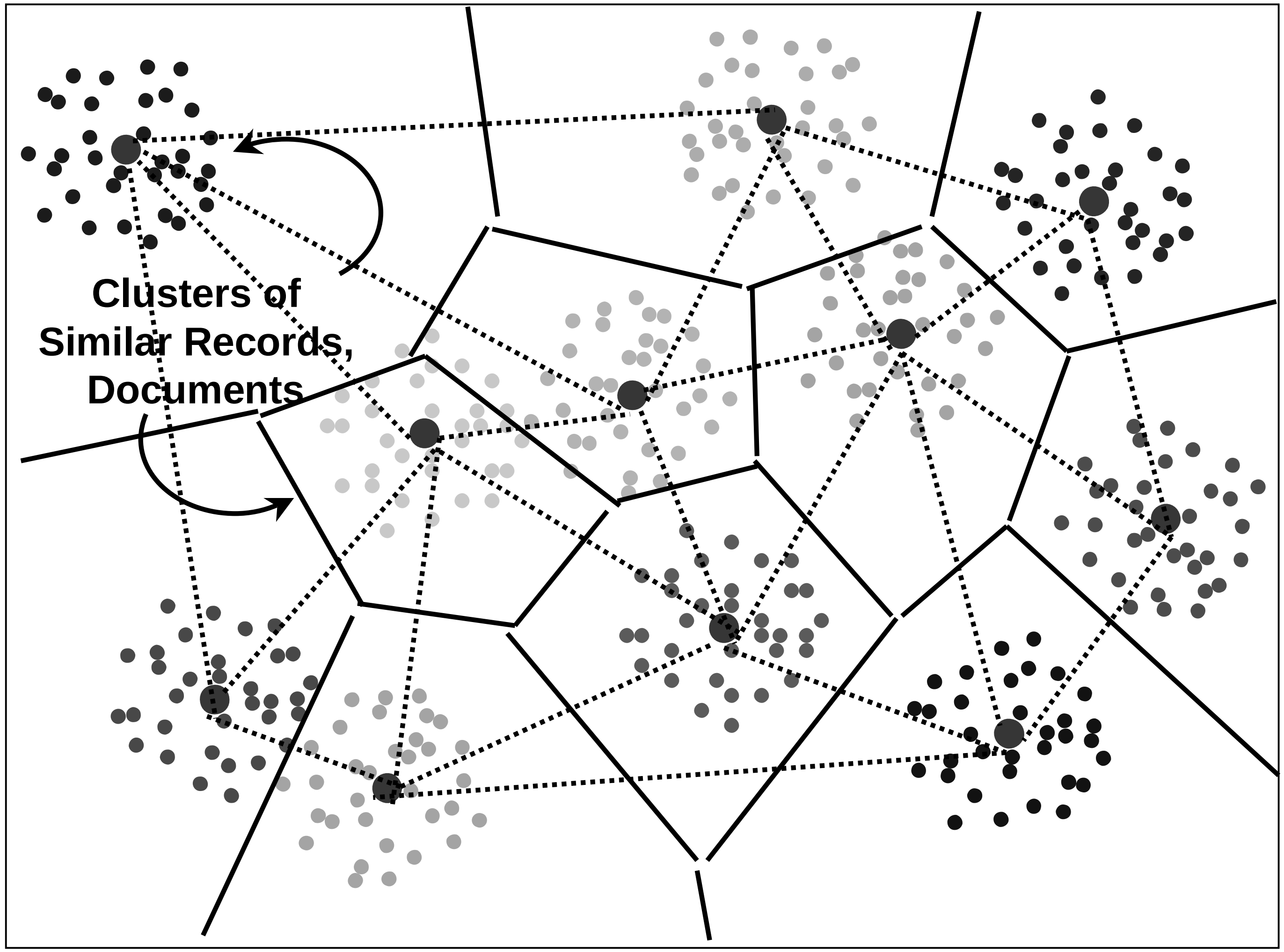
- **Supervised (Focus of CIS690)**
  - **What is learned? Classification function; other models**
  - **Inputs and outputs? Learning:** examples $\langle x, f(x) \rangle \rightarrow$ approximation $\hat{f}(x)$
  - **How is it learned? Presentation of examples to learner (by teacher)**
  - **<u>Projects: *MLC++* and *NCSA D2K*; *wrapper, clickstream* mining applications</u>**
- **Unsupervised (Surveyed in CIS690)**
  - **Cluster definition, or *vector quantization* function (*codebook*)**
  - **Learning:** observations $x \times$ distance metric $d(x_1, x_2) \rightarrow$ discrete codebook $f(x)$
  - **Formation, segmentation, labeling of clusters based on observations, metric**
  - **<u>Projects: *NCSA D2K*; *info retrieval (IR), Bayesian network* learning applications</u>**
- **Reinforcement (Not Emphasized in CIS690)**
  - **Control policy (function from states of the world to actions)**
  - **Learning:** state/reward sequence $\{\langle s_i, r_i \rangle : 1 \le i \le n\} \rightarrow$ policy $p : s \rightarrow a$
  - **(Delayed) feedback of reward values to agent based on actions selected; model updated based on reward, (partially) observable state**

# Unsupervised Learning: Data Clustering for Information Retrieval

Dimensionality-Reducing Projection ($x'$)

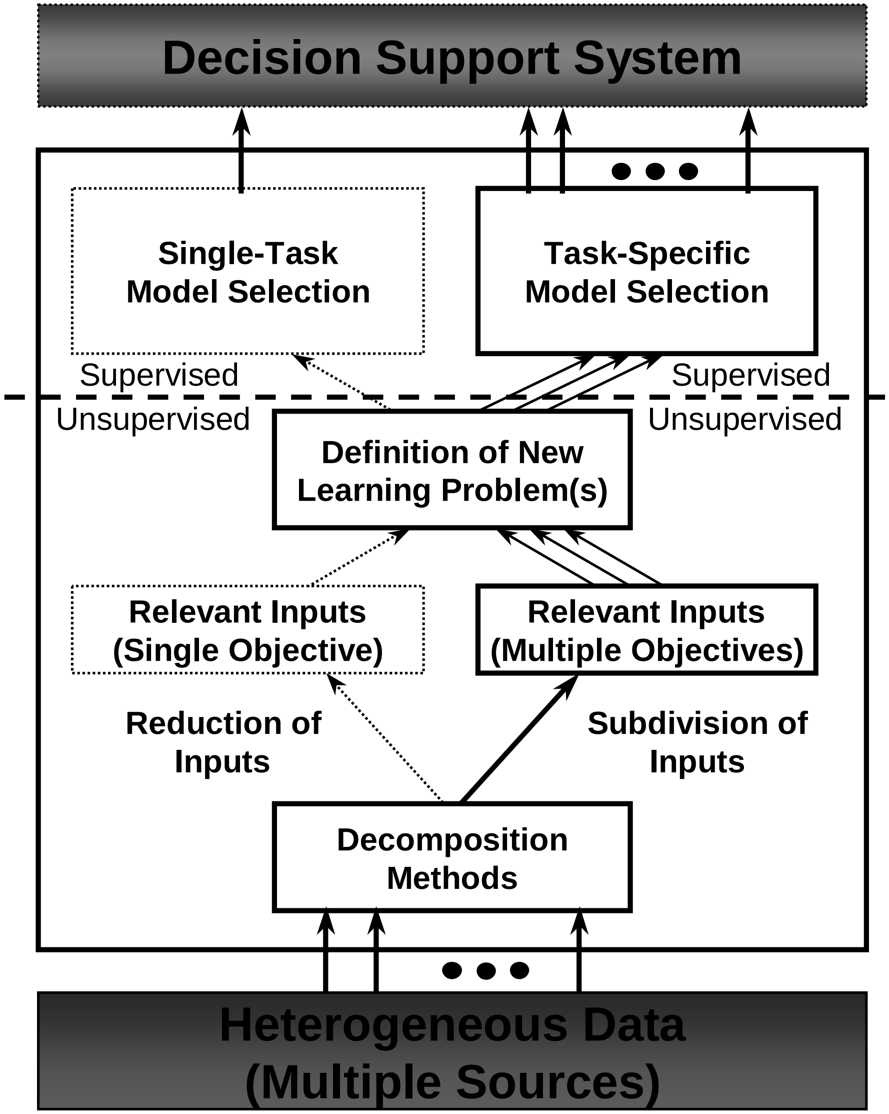
Delaunay Triangulation

Voronoi (Nearest Neighbor) Diagram ($y$)




Cluster Formation and Segmentation Algorithm (Sketch)

# High-Performance Computing and KDD: Wrappers for Performance Enhancement




- Wrappers
  - "Outer loops" for improving inducers
  - Use inducer performance to optimize
- Applications of Wrappers
  - Combining knowledge sources
    - Statistical methods: bagging, stacking, boosting
    - Other sensor and data fusion
  - Tuning hyperparameters
    - Number of ANN hidden units
    - GA control parameters
    - Priors in Bayesian learning
  - Constructive induction
    - Attribute (feature) subset selection
    - Feature construction
- Implementing Optimization Wrappers
  - Parallel, distributed (e.g., GA)
  - HPC application (e.g., Beowulf)

# AI and Machine Learning: Some Basic Topics

- Analytical Learning: Combining Symbolic and Numerical AI
  - Inductive learning
  - Role of knowledge and deduction in integrated inductive and analytical learning
- Artificial Neural Networks (ANNs) for KDD
  - Common neural representations: current limitations
  - Incorporating knowledge into ANN learning
- Uncertain Reasoning in Decision Support
  - Probabilistic knowledge representation
  - Bayesian knowledge and data engineering (KDE): elicitation, causality
- Data Mining: KDD Applications
  - Role of <u>causality</u> and explanations in KDD
  - Framework for data mining: wrappers for performance enhancement
- Genetic Algorithms (GAs) for KDD
  - Evolutionary algorithms (GAs, GP) as optimization wrappers
  - Introduction to classifier systems

# Online Resources

- Research
  - KSU Laboratory for Knowledge Discovery in Databases
    http://ringil.cis.ksu.edu/KDD (see especially Group Info, Web Resources)
  - KD Nuggets: http://www.kdnuggets.com
- Courses and Tutorials Online
  - At KSU
    - CIS798 *Machine Learning and Pattern Recognition*
      http://ringil.cis.ksu.edu/Courses/Fall-1999/CIS798
    - CIS830 *Advanced Topics in Artificial Intelligence*
      http://ringil.cis.ksu.edu/Courses/Spring-2000/CIS830
    - CIS690 *Implementation of High-Performance Data Mining Systems*
      http://ringil.cis.ksu.edu/Courses/Summer-2000/CIS690
  - Other courses: see KD Nuggets, www.aaai.org, www.auai.org
- Discussion Forums
  - Newsgroups: comp.ai.*
  - Recommended mailing lists: *Data Mining*, *Uncertainty in AI*
  - KSU KDD Lab Discussion Board: http://ringil.cis.ksu.edu/KDD/Board

# Terminology

- **Data Mining**
  - **<u>Operational definition</u>: automatically extracting *valid*, *useful*, *novel*, *comprehensible* information from large databases and *using it to make decisions***
  - **<u>Constructive definition</u>: expressed in stages of data mining**
- **Databases and Data Mining**
  - **<u>D</u>ata <u>B</u>ase <u>M</u>anagement <u>S</u>ystem (<u>DBMS</u>): data *organization, retrieval, processing***
  - **<u>Data warehouse</u>: repository of integrated information for queries, analysis**
  - **<u>O</u>nline <u>A</u>nalytical <u>P</u>rocessing (<u>OLAP</u>): storage/CPU-efficient manipulation of data for summarization (descriptive statistics), inductive learning and inference**
- **Stages of Data Mining**
  - **<u>Data selection</u> (*aka* <u>filtering</u>): sampling original (<u>raw</u>) data**
  - **<u>Data preprocessing</u>: sorting, segmenting, aggregating**
  - **<u>Data transformation</u>: change of representation; feature construction, selection, extraction; <u>quantization</u> (<u>scalar</u>, e.g., <u>histogramming</u>, <u>vector</u>, *aka* <u>clustering</u>)**
  - **<u>Machine learning</u>: unsupervised, supervised, reinforcement for model building**
  - **<u>Inference</u>: application of performance element (pattern recognition, *etc.*); evaluation, assimilation of results**

# Summary Points

- **Knowledge Discovery in Databases (KDD) and Data Mining**
  - **<u>Stages</u>: selection (filtering), processing, transformation, learning, inference**
  - **Design and implementation issues**
- **Role of Machine Learning and Inference in Data Mining**
  - **Roles of unsupervised, supervised learning in KDD**
  - **Decision support (information retrieval, prediction, policy optimization)**
- **Case Studies**
  - **Risk analysis, transaction monitoring (filtering), prognostic monitoring**
  - **Applications: business decision support (pricing, fraud detection), automation**
- **More Resources Online**
  - **Microsoft DMX Group (Fayyad): http://research.microsoft.com/research/DMX/**
  - **KSU KDD Lab (Hsu): http://ringil.cis.ksu.edu/KDD/**
  - **CMU KDD Lab (Mitchell): http://www.cs.cmu.edu/~cald**
  - **KD Nuggets (Piatetsky-Shapiro): http://www.kdnuggets.com**

  **NCSA Automated Learning Group (Welge)**
    - **ALG home page: http://www.ncsa.uiuc.edu/STI/ALG**
    - **NCSA *D2K*: http://chili.ncsa.uiuc.edu**