Automated Learning and Data Analysis
CSC 422 / 522
Spring 2013

Jon Doyle
Department of Computer Science

Subject introduction

Topic

- Obtaining knowledge by examining data
- Learning how to do it
- Learning how to think about it

``**The purpose of computing is insight, not numbers**."
- R. W. Hamming

The Problem

- Too much data
  - Commercial data
  - Scientific data
  - Governmental data
  - Health data
  - Personal data
- Too few people to understand it all
  - Reduce need to managable numbers
- Too limited human minds
  - Even obvious facts are invisible at this scale

Exercise

- Task?
- Data sources?
- Hypotheses or questions?
- Difficulties?

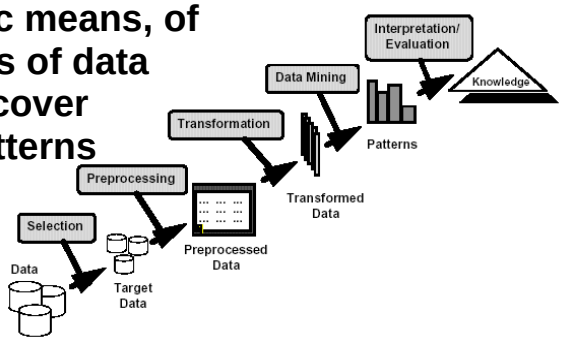
- How much data is there?
- How much is enough?
- How much is too much?

Solutions?

- Stop making data
  - But data can improve our lives
- Enlist more people
  - But who will tend the farms, etc.?
- Make smarter people
  - Computer-aided data analysis
    - Data mining
    - Machine learning
    - Knowledge discovery in databases, etc.
- Why not try all?

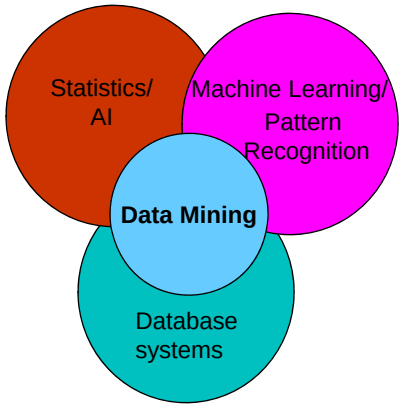
What is Data Mining?

- **Many Definitions**
  - **Non-trivial extraction of implicit, previously unknown and potentially useful information from data**
  - **Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns**



Origins of Data Mining

- **Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems**
- **Traditional Techniques may be unsuitable due to**
  - **Enormity of data**
  - **High dimensionality of data**
  - **Heterogeneous, distributed nature of data**



Where does this come from?

- Data analysis goes back millennia
  - Invention of history
  - Planetary and lunar astronomy
  - Detecting cheating satraps
- Formalization
  - Natural philosophy
  - Scientific method
  - Statistics
  - Artificial intelligence

Where does this come from?

- Natural philosophy and the scientific method
  - **Non-trivial extraction of implicit, previously unknown and potentially useful information from data**
  - Logical
    - Consistent
    - Implicit
  - Falsifiable
    - Nontrivial
  - Scholarly
    - Previously unknown
  - Experimentally verifiable
    - Robust, unchanging

Where does this come from?

- Taxonomy, cladistics, linguistics, info retrieval
  - Similarity and metric methods
  - Focused on comparison of features
- Statistics
  - Sampling
  - Regression: linear and nonlinear
  - Decision trees
  - Focused on evaluation, not construction, of hypotheses
- Artificial intelligence
  - Logical, probabilistic, prototypical representations
  - Knowledge acquisition and learning
  - Automated knowledge-based analysis, discovery, and exploration
  - Focused on search and construction, not evaluation

Data Mining Tasks

- Prediction Methods
  - Use some variables to predict unknown or future values of other variables.
- Description Methods
  - Find human-interpretable patterns that describe the data.

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

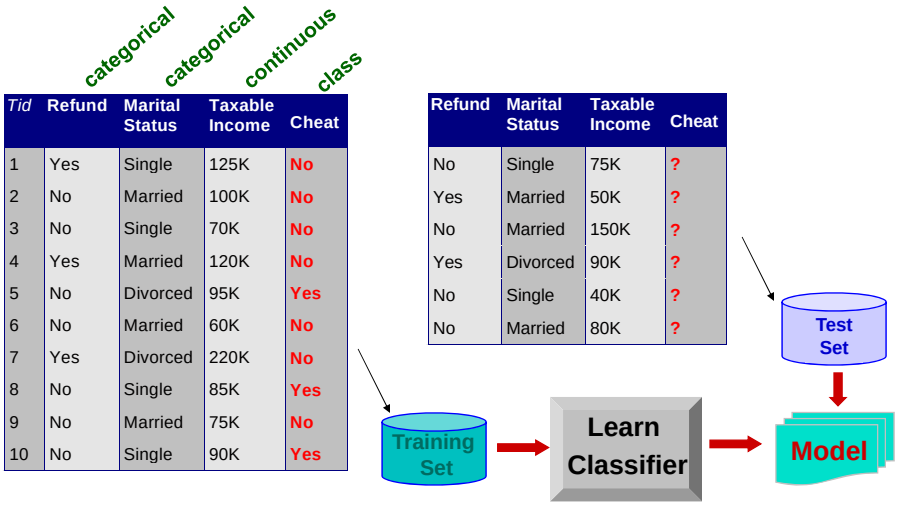
Data Mining Tasks...

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]

Classification: Definition

- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Classification Example



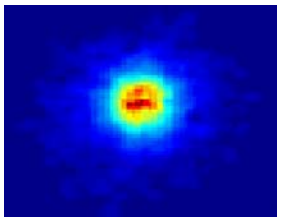
Classification Applications

- Direct marketing
  - What attributes make someone a likely customer?
- Credit card fraud detection
  - What attributes make some transaction suspect?
- Dropout prevention
  - What attributes signal impending student failure?
- Stellar object classification
  - What attributes identify interesting objects?

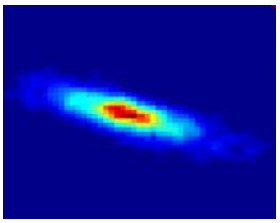
Classifying Galaxies

Courtesy: http://aps.umn.edu

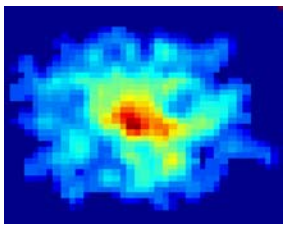
Early



Intermediate



Late



Class:
- **Stages of Formation**

Attributes:
- **Image features,**
- **Characteristics of light waves received, etc.**

Data Size:
- **72 million stars, 20 million galaxies**
- **Object Catalog: 9 GB**
- **Image Database: 150 GB**

Clustering Definition

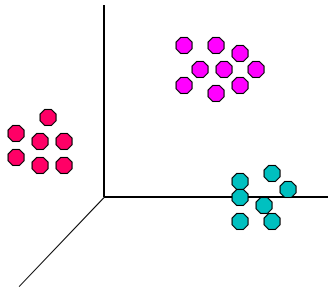
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance if attributes are continuous.
  - Other Problem-specific Measures.

Illustrating Clustering

☒Euclidean Distance Based Clustering in 3-D space.

Intracluster distances are minimized

Intercluster distances are maximized



Clustering: Application 1

- Market Segmentation:
  - Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
  - Approach:
    - Collect different attributes of customers based on their geographical and lifestyle related information.
    - Find clusters of similar customers.
    - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Clustering: Application 2

- Document Clustering:
  - Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
  - Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
  - Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

Illustrating Document Clustering

- Clustering Points: 3204 Articles of Los Angeles Times.
- Similarity Measure: How many words are common in these documents (after some word filtering).

Category	Total Articles	Correctly Placed
Financial	555	364
Foreign	341	260
National	273	36
Metro	943	746
Sports	738	573
Entertainment	354	278

Clustering of S&P 500 Stock Data

- Observe Stock Movements every day.
- Clustering points: Stock-{UP/DOWN}
- Similarity Measure: Two points are more similar if the events described by them frequently happen together on the same day.
  - We used association rules to quantify a similarity measure.

	Discovered Clusters	Industry Group
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP

Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection;
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:
{Milk} --> {Coke}
{Diaper, Milk} --> {Beer}

Association Rule Discovery: Application 2

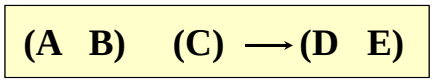
- Supermarket shelf management.
  - Goal: To identify items that are bought together by sufficiently many customers.
  - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
  - A classic rule --
    - If a customer buys diaper and milk, then he is very likely to buy beer.
    - So, don't be surprised if you find six-packs stacked next to diapers!

Association Rule Discovery: Application 3

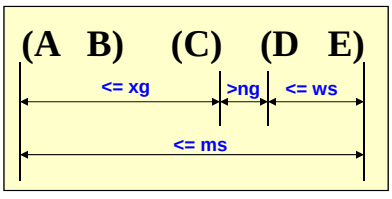
- Inventory Management:
  - Goal: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households.
  - Approach: Process the data on tools and parts required in previous repairs at different consumer locations and discover the co-occurrence patterns.

Sequential Pattern Discovery: Definition

- Given is a set of *objects*, with each object associated with its own *timeline of events*, find rules that predict strong sequential dependencies among different events.



- Rules are formed by first discovering patterns. Event occurrences in the patterns are governed by timing constraints.



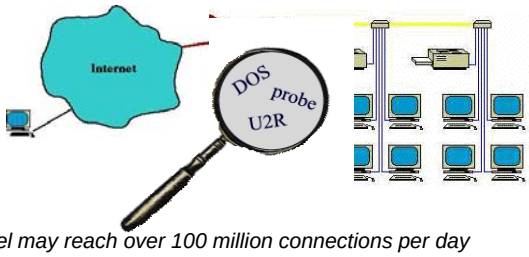
Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Greatly studied in statistics, neural network fields.
- Examples:
  - Predicting sales amounts of new product based on advertising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

Deviation / Anomaly Detection

- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection

  - Network Intrusion Detection



Typical network traffic at University level may reach over 100 million connections per day

The Big Picture

Seek Ye

Truth

Goodness

Beauty

Perfection

Categories of Qualities

- Commence to continuously correct the conception via the critique categories until convergence

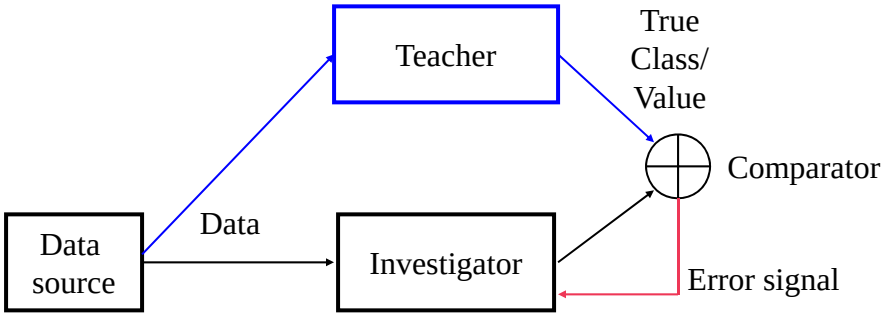
Truth
- Correctness
- Consistency
- Completeness
- Categoricity
- Contingency
- Chance
- Coverage

Goodness
- Computability
- Complexity
- Cardinality
- Compromises
- Convenience
- Compactness
- Cost

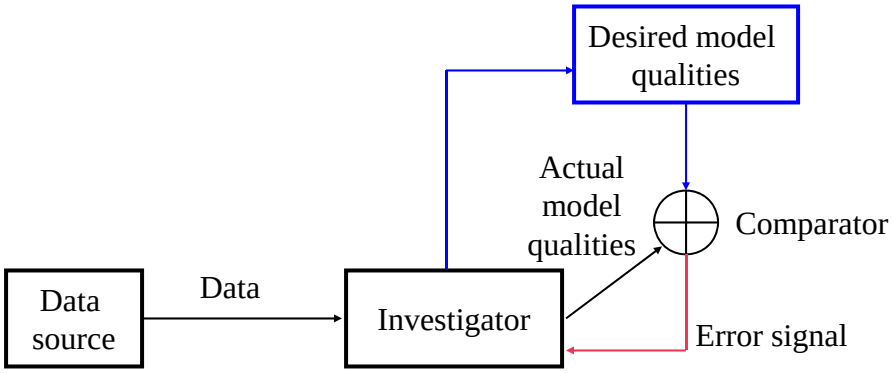
Beauty
- Clarity
- Comprehensibility
- Cleavage
- Cogency
- Commonsensicality
- Continuity

Perfection
- Closeness
- Cumulativity
- Convergence
- Constancy

Supervised learning



Unsupervised learning



Punchlines

- Representations are the key issue
  - Approximating data is easy
  - Approximating data closely is easy
  - Tradeoffs abound
- Analysis assumptions can change everything
  - Reference class problem
  - There is no free lunch
  - Data comes without meaning
  - Random in, random out
- Bulk of work is in preparation
  - Garbage in, garbage out

The Science of Learning

- What can be learned from data? Feasibly? Reliably?
- How can one express what one learns?
- What knowledge can be expressed? Approximated? Learned?
- How hard is it to learn knowledge type T from data type D?
- Is learning method M_1 better than method M_2?
- How can one distinguish more useful from less useful knowledge?

Knowledge ≠ formal expression

- Classification rules
- Association rules
- Predictive functions
- Cluster identifications

- Formal expressions or structural descriptions can represent knowledge, but are not themselves knowledge

Feeling cheated

- The data falls into two groups
  - One with "SEX" = "M"
  - One with "SEX" = "F"
- The data exhibits the association
  - If "CITY" = "MISSING", then "STREET" = "MISSING"

Feeling cheated

IF There is pressure on project leader,
Respondent is R&D manager,
No increase in growth stage of life cycle,
Increased probability of commercial success,
Project champion did not appear at end,
Product not in infancy stage of life cycle,
Top management support,
Association between commercial and technological aspects,
R&D perceives project mgmt commitment as high
There is a project champion
Don't know about newly enacted favorable international regulations
Project champion appeared in the middle,
Respondent is not the VP,
Project Champion didn't appear at beginning,
Respondent is not marketing manager,

THEN the Project is likely to succeed.

From Gallant and Balachandra, Using automated techniques to generate an expert system for R&D project monitoring, *First IFAC Intl Symposium on Economics and AI*, Sept. 1986.

(Thanks to Randall Davis)

There is no free lunch

- Knowledge is *not* mere formal expression
- The more knowledge you want *out* of data, the more knowledge you have to put *in*
- Knowledge is closely linked to rationality and purpose

Character of knowledge

- Isolated empirical knowledge
  - "Facts", rules of thumb, heuristics
- Coherent design knowledge
  - "Theory", model, causal understanding

Empirical knowledge

- Classification
  - Wings, feathers, two legs mean bird
- Association
  - Customers who buy tortilla chips also buy salsa
- Prediction
  - House price = A*area + B*bathrooms + C*traffic volume
- Clustering
  - Urban vs rural buyers

Design knowledge

- Coherence
  - What facts hang together?
    - Kepler's laws vs Newton's theory
    - People buy chips to scoop up salsa
  - Deductive classification, association, and prediction
- Intent
  - Not triviality or chance
  - Purpose and utility

Challenges of Data Mining

- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation
- Streaming Data

The Dark Side

- Knowledge is power
- Data collection sites as sources of power
  - *One place to rule them all*
  - *One place to find them,*
  - *One place to bring them all*
  - *And in the darkness bind them*

Some of the worries

- Tyrants
  - Hitler: detect Jews, communists, gypsies, homosexuals, spies
  - Stalin: detect Trotskyites, revanchists, bourgeoisie, nationalists, spies
- Criminals
  - Identity theft
  - Extortion
- Discrimination
  - Insurance, credit, job

For Good and Ill

- Correlating disparate data sources
  - Bringing civil defense data from FBI, CIA, ICE, CDC, and police into one place
  - Bringing together medical records from all sources
- Collecting data just because it can be collected
  - We don't know what will be useful yet

Questions to keep in mind

- Who wants to know?
- Who has the right to know?
- Who owns data?
  - The subject?
  - The collector?
  - The consumer?
- Who decides access to data?
- Can one guarantee anonymity?
  - US census data