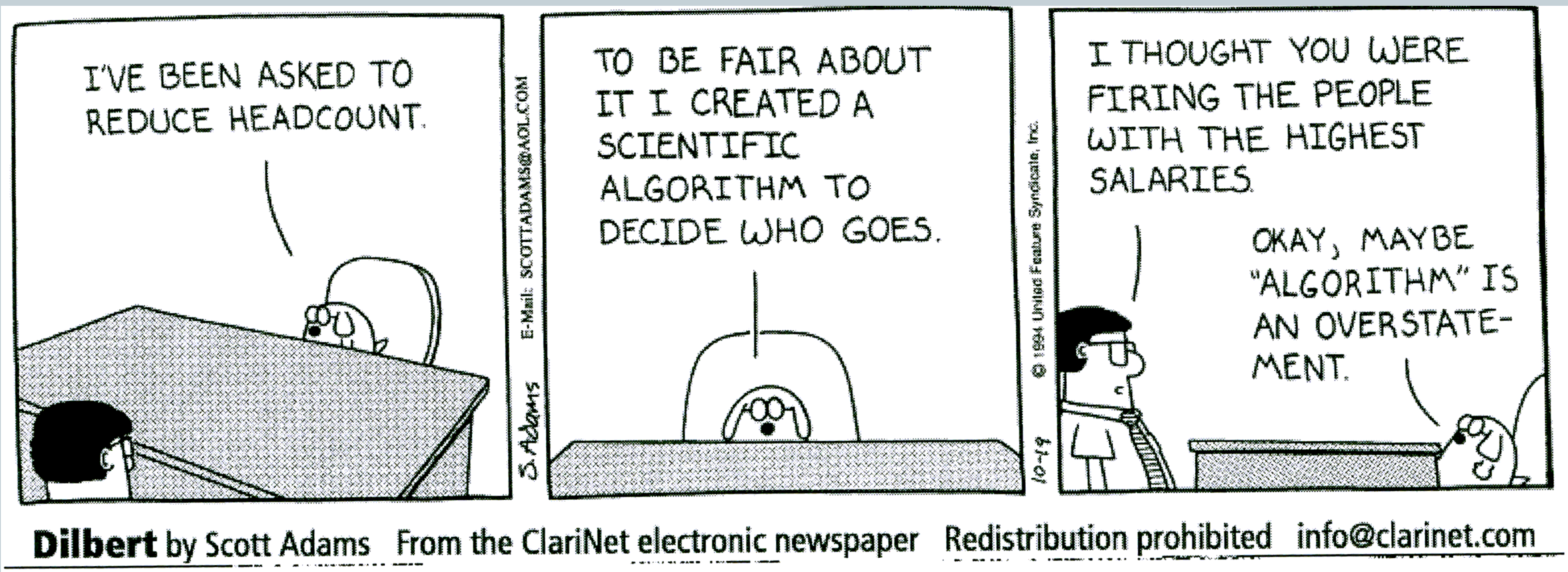
Introduction to Algorithms

SU 5050
LECTURE 5
JESSICA L. MCCARTY, PH.D.

Algorithm Definition 1

- An algorithm is a description of a procedure which terminates with a result.



Algorithm Definition 2

- A step-by-step problem solving procedure, especially an established, recursive computational procedure for solving a problem in a finite number of steps.

Algorithm 1 Compute sum of integers in array

```
procedure ArraySum(A)
    sum = 0
    for each integer i in A do
        sum = sum + i
    end for
    Return sum
end procedure
```

Algorithm Definition 3

- Sequence of unambiguous instructions for solving a problem by obtaining a required output for any legitimate input in a finite amount of time.

Algorithm MonteCarloEuropeanCallOption

Input:

S_0 = Security Price
K = Strike Price
r = Risk free Interest Rate
σ = Security Volatility
t = Time to expiration (Years)
n = number of simulations

Output: Average call option payoff

$\text{vsqrdt} = \sigma\sqrt{t}$

$\text{drift} = \left(r - \frac{\sigma^2}{2}\right)t$

$\text{expRT} = e^{-rt}$

sum = 0

sum = 0

for i = 1 to n, **do**

$St = S0 \times e^{(drift + vsqrdt \times NextGaussian())}$

$if\ (St - K > 0)$

$sum{+}{=}\ (St - K) \times expRT$

return sum / n

Why do we study algorithms?

- The basis of computer programs and the computations generated by them permeate society:

  - Airplane wings design
  - Climate change modeling
  - Groundwater contamination simulations

Continuous nature of algorithms

- Approximating an integral
- Solving a system of linear equations
- Finding the roots of a function
- Solving a differential equation

1. $\left(\frac{d^3y}{dx^3}\right)^4 + 2\frac{dy}{dx} = \sin x$

2. $\frac{dy}{dx} - 2xy = x^2 - x$

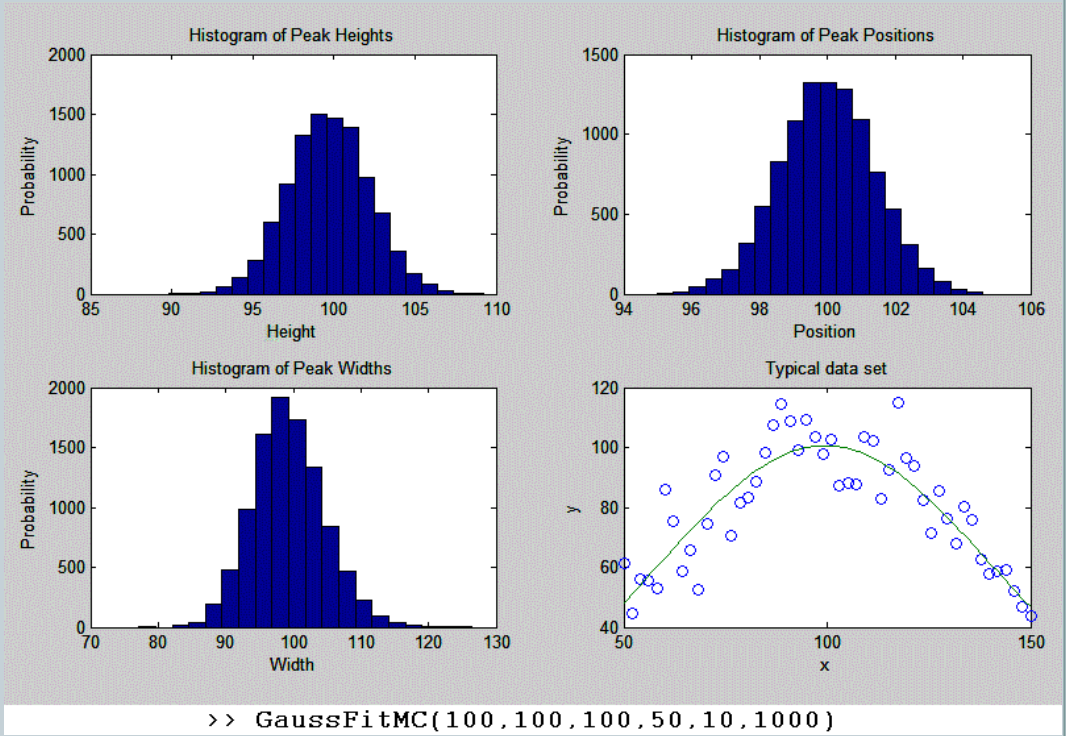
3. $\frac{dy}{dx} - \sin y = -x$

4. $\frac{d^2y}{dx} = 2xy$

www.analyzemath.com

Discrete nature of algorithms

- Text string
- Sorting a list of objects
- Optimal path between cities
- Simulating a random process
- Finding a point from a list which is closest to a given point



Pseudocode

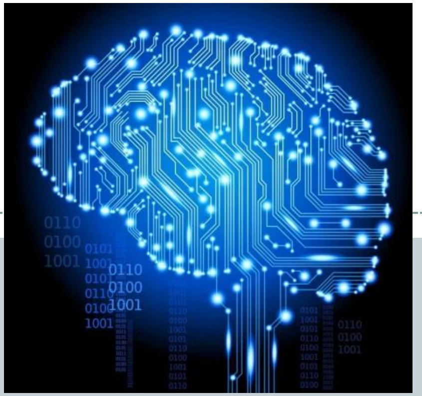
- Developed out of a clear and concise way to describe an algorithm which is not language dependent
- Combination of common language and terminology (i.e., loops and conditionals)
- Does NOT contain correct syntax for any language because it is language independent

Pseudocode

- The goal of writing an algorithm in psuedocode is to allow you to understand precisely what the steps of the algorithm are.

```
for i=1, n-1
    for j=1,n-i
      if ( a(j+1) > a(j ) ) swap a(j) and a(j+1)
    end for loop over j
end for loop over i
```

More Psuedocode



- Human reading vs. Machine reading

Algorithm 1: computeTRI(DEM, TRI)

Input: DEM is the input matrix that contains the elevation values
Output: TRI contains the concepts that describe the ruggedness
for each *cell* $DEM_{ij} \in DEM$ **do**
  $\mathrm{aux}_{ij} \leftarrow 0$;
  for each *cell* $v_{ij} \in N_8(DEM_{ij})$ **do**
    $\mathrm{aux}_{ij} \leftarrow \sqrt{\mathrm{aux}_{ij} + (v_i - \mathrm{DEM}_{ij})^2}$
  for each *cell* $aux_{ij} \in aux$ **do**
    if $aux_{ij} \geq 0$ *and* $aux_{ij} \leq 80$ **then** $\mathrm{TRI}_{ij} \leftarrow \mathrm{LTS}$;
    if $aux_{ij} > 80$ *and* $aux_{ij} \leq 116$ **then** $\mathrm{TRI}_{ij} \leftarrow \mathrm{NLS}$;
    if $aux_{ij} > 116$ *and* $aux_{ij} \leq 161$ **then** $\mathrm{TRI}_{ij} \leftarrow \mathrm{SRS}$;
    if $aux_{ij} > 161$ *and* $aux_{ij} \leq 239$ **then** $\mathrm{TRI}_{ij} \leftarrow \mathrm{IRS}$;
    if $aux_{ij} > 239$ *and* $aux_{ij} \leq 497$ **then** $\mathrm{TRI}_{ij} \leftarrow \mathrm{MRS}$;
    if $aux_{ij} > 497$ *and* $aux_{ij} \leq 958$ **then** $\mathrm{TRI}_{ij} \leftarrow \mathrm{HRS}$;
    if $aux_{ij} > 958$ *and* $aux_{ij} \leq 4367$ **then** $\mathrm{TRI}_{ij} \leftarrow \mathrm{ERS}$;

Figure 2 Pseudo code of the TRI algorithm

Some Important Types of Problems

1. Sorting
2. Searching
3. Randomness
4. Graph problems
5. Optimization problems
6. Knapsack or rucksack problem (mass and volume)
7. Clustering
8. Computation Geometry
9. Image Processing
10. Numerical

Data Mining Tasks

- Classification *[Predictive]*
- Clustering *[Predictive]*
- Association Rule Discovery *[Descriptive]*
- Sequential Pattern Discovery *[Descriptive]*
- Natural Language Processing *[Descriptive]*
- Regression *[Predictive]*
- Deviation Detection *[Predictive]*

Common Approaches for Designing Algorithms

1. *Brute Force* – straightforward
2. *Divide and conquer* – divide into smaller problems
3. *Decrease and conquer* – $\pi^8 = \pi^4 \pi^4$
4. *Transform and conquer* – make more amenable to a solution
5. *Greedy algorithms* – solution through sequence of steps, each step choice made based on criteria

Data Mining Tasks

- Classification *[Predictive]*
- Clustering *[Predictive]*
- Association Rule Discovery *[Descriptive]*
- Sequential Pattern Discovery *[Descriptive]*
- Natural Language Processing *[Descriptive]*
- Regression *[Predictive]*
- Deviation Detection *[Predictive]*

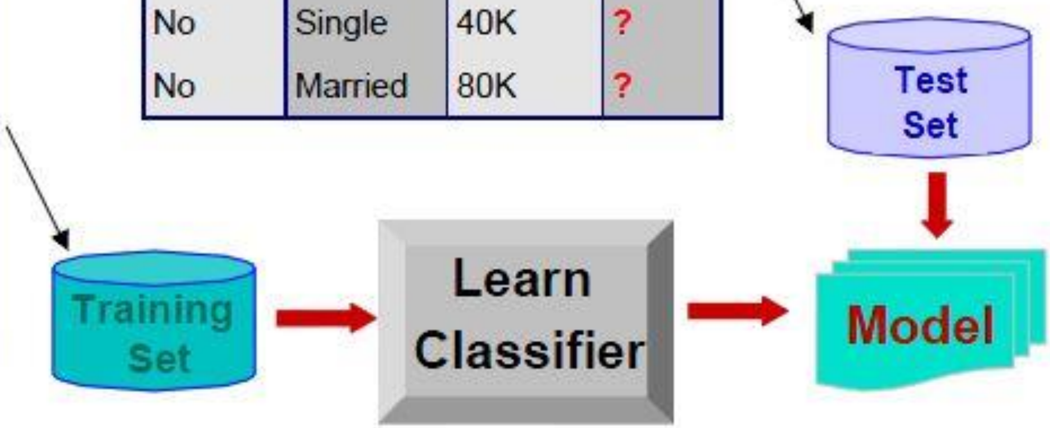
Classification

- 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

	categorical	categorical	continuous	class
Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



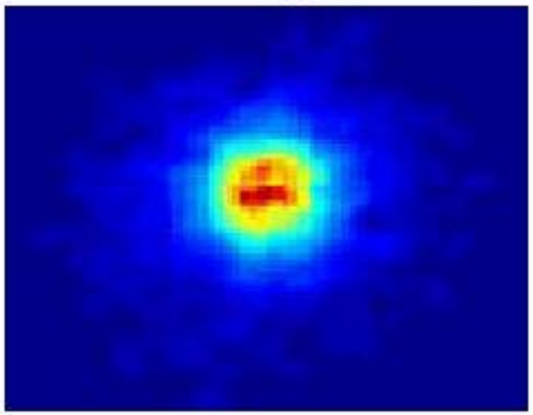
Classification: Application

- Sky Survey Cataloging
  - Goal: To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
    - 3000 images with 23,040 x 23,040 pixels per image.
  - Approach:
    - Segment the image.
    - Measure image attributes (features) - 40 of them per object.
    - Model the class based on these features.
    - Success Story: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find!

Classifying Galaxies

Courtesy: http://aps.umn.edu

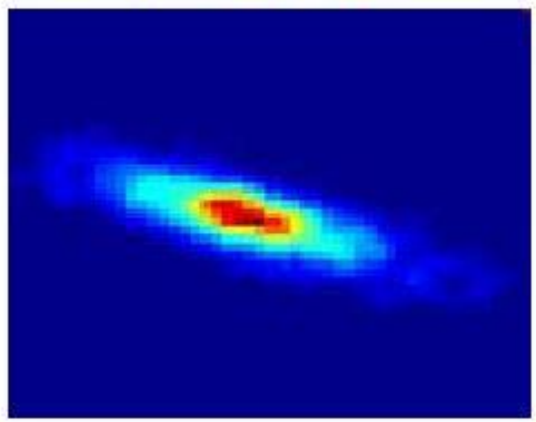
Early



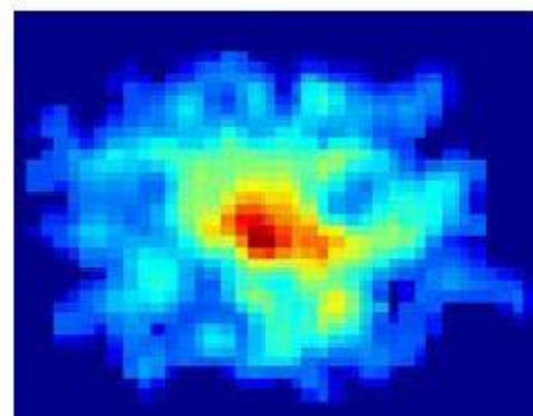
Class:
- Stages of Formation

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

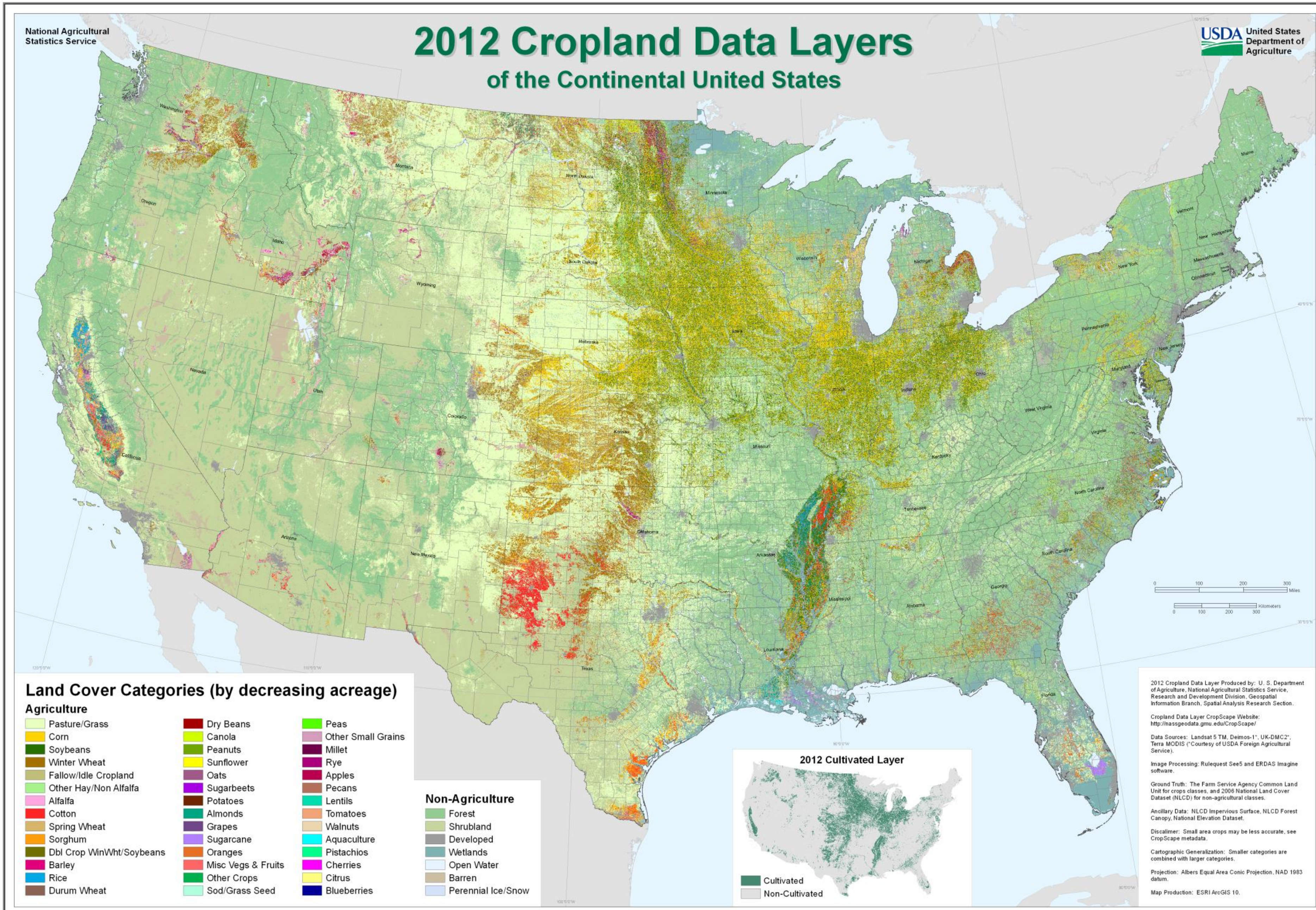
Intermediate

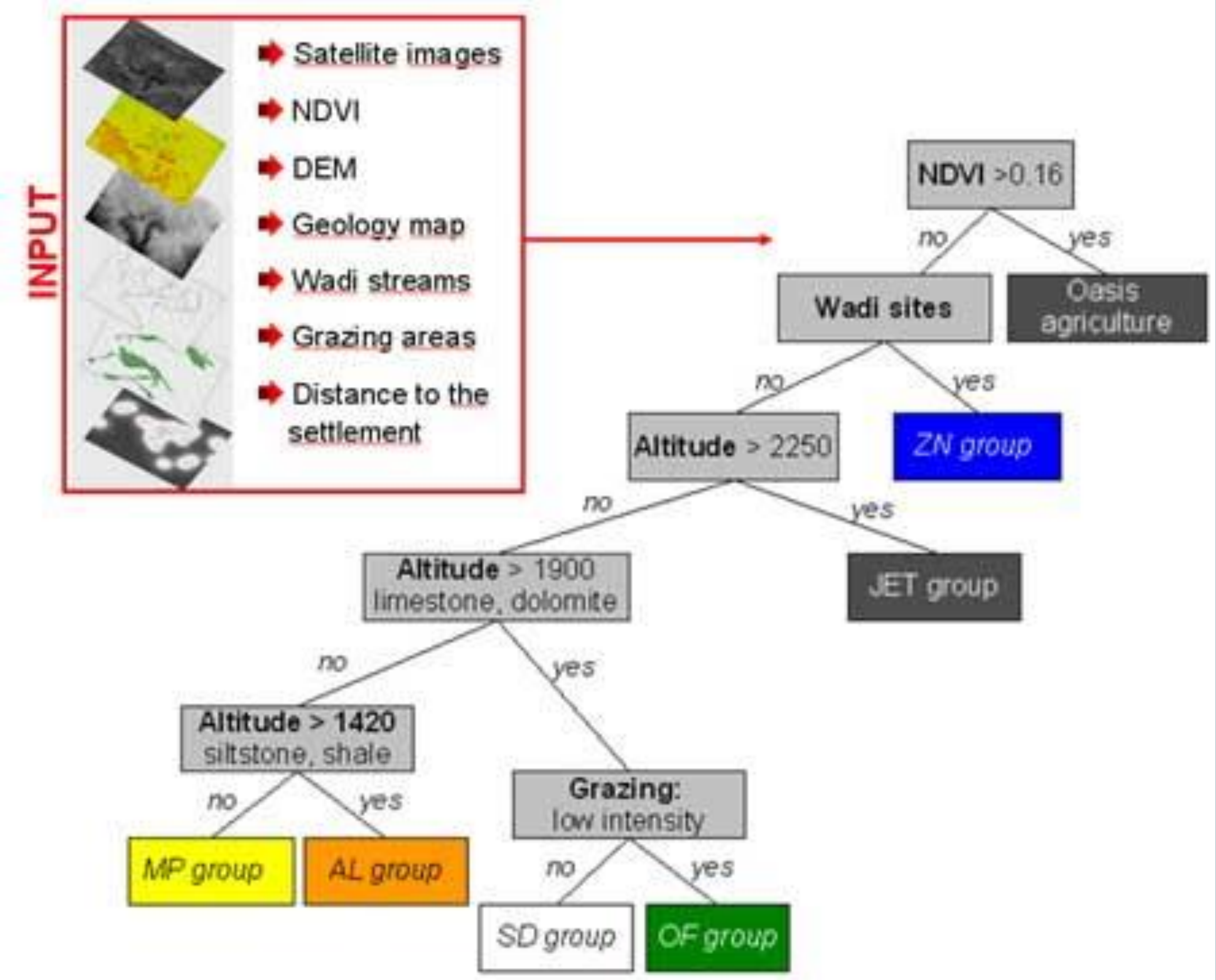


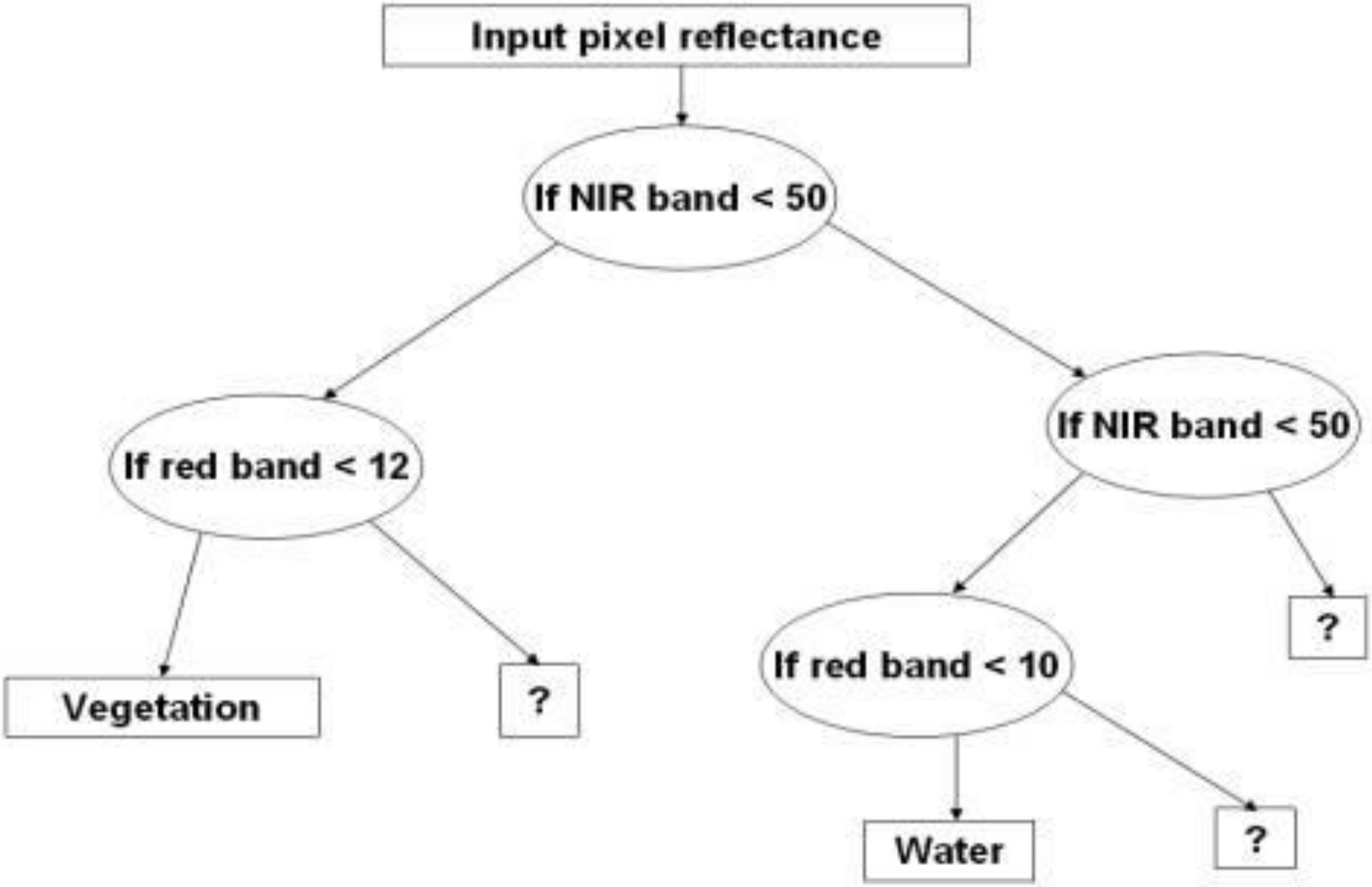
Late



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


National Agricultural Statistics Service
2012 Cropland Data Layers
of the Continental United States
USDA United States Department of Agriculture
Washington
Oregon
Idaho
Montana
North Dakota
Minnesota
South Dakota
Wyoming
Wisconsin
Michigan
Nevada
Utah
Colorado
Nebraska
Iowa
Illinois
Indiana
Ohio
Pennsylvania
New York
Vermont
New Hampshire
Maine
Massachusetts
Connecticut
New Jersey
Maryland
West Virginia
Virginia
Kentucky
Kansas
Missouri
California
Arizona
New Mexico
Oklahoma
Arkansas
Tennessee
North Carolina
South Carolina
Georgia
Alabama
Mississippi
Louisiana
Texas
Florida
0 100 200 300 Miles
0 100 200 300 Kilometers
Land Cover Categories (by decreasing acreage)
Agriculture
Pasture/Grass
Corn
Soybeans
Winter Wheat
Fallow/Idle Cropland
Other Hay/Non Alfalfa
Alfalfa
Cotton
Spring Wheat
Sorghum
Dbl Crop WinWht/Soybeans
Barley
Rice
Durum Wheat
Dry Beans
Canola
Peanuts
Sunflower
Oats
Sugarbeets
Potatoes
Almonds
Grapes
Sugarcane
Oranges
Misc Vegs & Fruits
Other Crops
Sod/Grass Seed
Peas
Other Small Grains
Millet
Rye
Apples
Pecans
Lentils
Tomatoes
Walnuts
Aquaculture
Pistachios
Cherries
Citrus
Blueberries
Non-Agriculture
Forest
Shrubland
Developed
Wetlands
Open Water
Barren
Perennial Ice/Snow
2012 Cultivated Layer
Cultivated
Non-Cultivated
2012 Cropland Data Layer Produced by: U. S. Department of Agriculture, National Agricultural Statistics Service, Research and Development Division, Geospatial Information Branch, Spatial Analysis Research Section.
Cropland Data Layer CropScape Website: http://nassgeodata.gmu.edu/CropScape/
Data Sources: Landsat 5 TM, Deimos-1*, UK-DMC2*, Terra MODIS (*Courtesy of USDA Foreign Agricultural Service).
Image Processing: Rulequest See5 and ERDAS Imagine software.
Ground Truth: The Farm Service Agency Common Land Unit for crops classes, and 2006 National Land Cover Dataset (NLCD) for non-agricultural classes.
Ancillary Data: NLCD Impervious Surface, NLCD Forest Canopy, National Elevation Dataset.
Discalimer: Small area crops may be less accurate, see CropScape metadata.
Cartographic Generalization: Smaller categories are combined with larger categories.
Projection: Albers Equal Area Conic Projection, NAD 1983 datum.
Map Production: ESRI ArcGIS 10.


INPUT
Satellite images
NDVI
DEM
Geology map
Wadi streams
Grazing areas
Distance to the settlement
NDVI >0.16
no
yes
Wadi sites
Oasis agriculture
no
yes
Altitude > 2250
ZN group
no
yes
Altitude > 1900 limestone, dolomite
JET group
no
yes
Altitude > 1420 siltstone, shale
Grazing: low intensity
no
yes
MP group
AL group
no
yes
SD group
OF group


Input pixel reflectance
If NIR band < 50
If red band < 12
If NIR band < 50
Vegetation
?
If red band < 10
?
Water
?

Remote Sensing

- **Data Mining Tools See5 and C5.0:** https://www.rulequest.com/see5-info.html

- **Data Mining In Earth System Science:** http://www.northeastern.edu/sds/DataminingInEarth.pdf

- **Fresh Approach to Agricultural Statistics: Data Mining and Remote Sensing:** http://1.usa.gov/10LyiRu

Association Rule Discovery

- 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

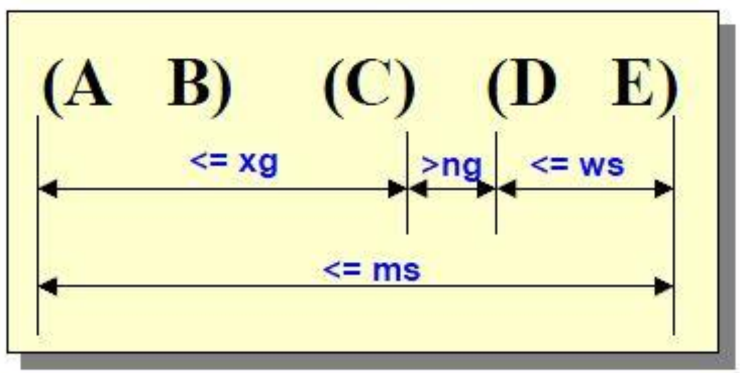
- Supermarket shelf management.
  - Goal: To identify items that are bought together by sufficiently many customers.
  - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
  - A classic rule --
    - If a customer buys diaper and milk, then he is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

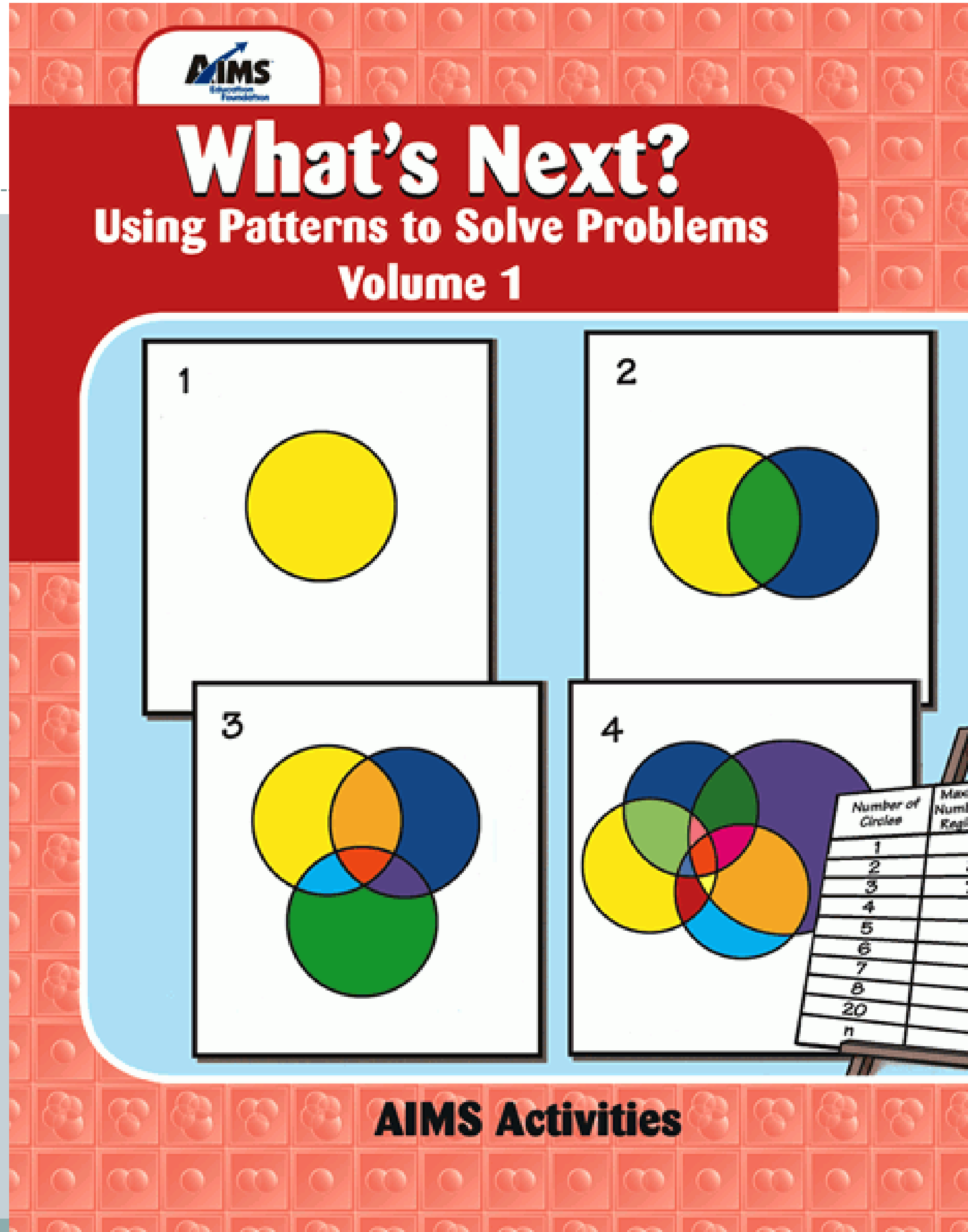
Sequential Pattern Discovery

- 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.

$$(A \quad B) \quad (C) \longrightarrow (D \quad E)$$

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




AIMS
What's Next?
Using Patterns to Solve Problems
Volume 1
1
2
3
4
Number of Circles
1
2
3
4
5
6
7
8
20
n
1
3
7
AIMS Activities

Sequential Pattern Discovery

- In telecommunications alarm logs,
  - (Inverter_Problem Excessive_Line_Current)

    (Rectifier_Alarm) --> (Fire_Alarm)
- In point-of-sale transaction sequences,
  - Computer Bookstore:

    (Intro_To_Visual_C) (C++_Primer) -->
    (Perl_for_dummies,Tcl_Tk)
  - Athletic Apparel Store:

    (Shoes) (Racket, Racketball) --> (Sports_Jacket)

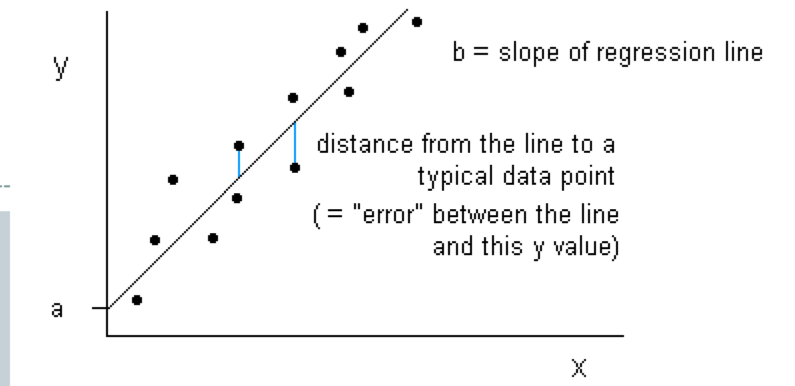
Regression

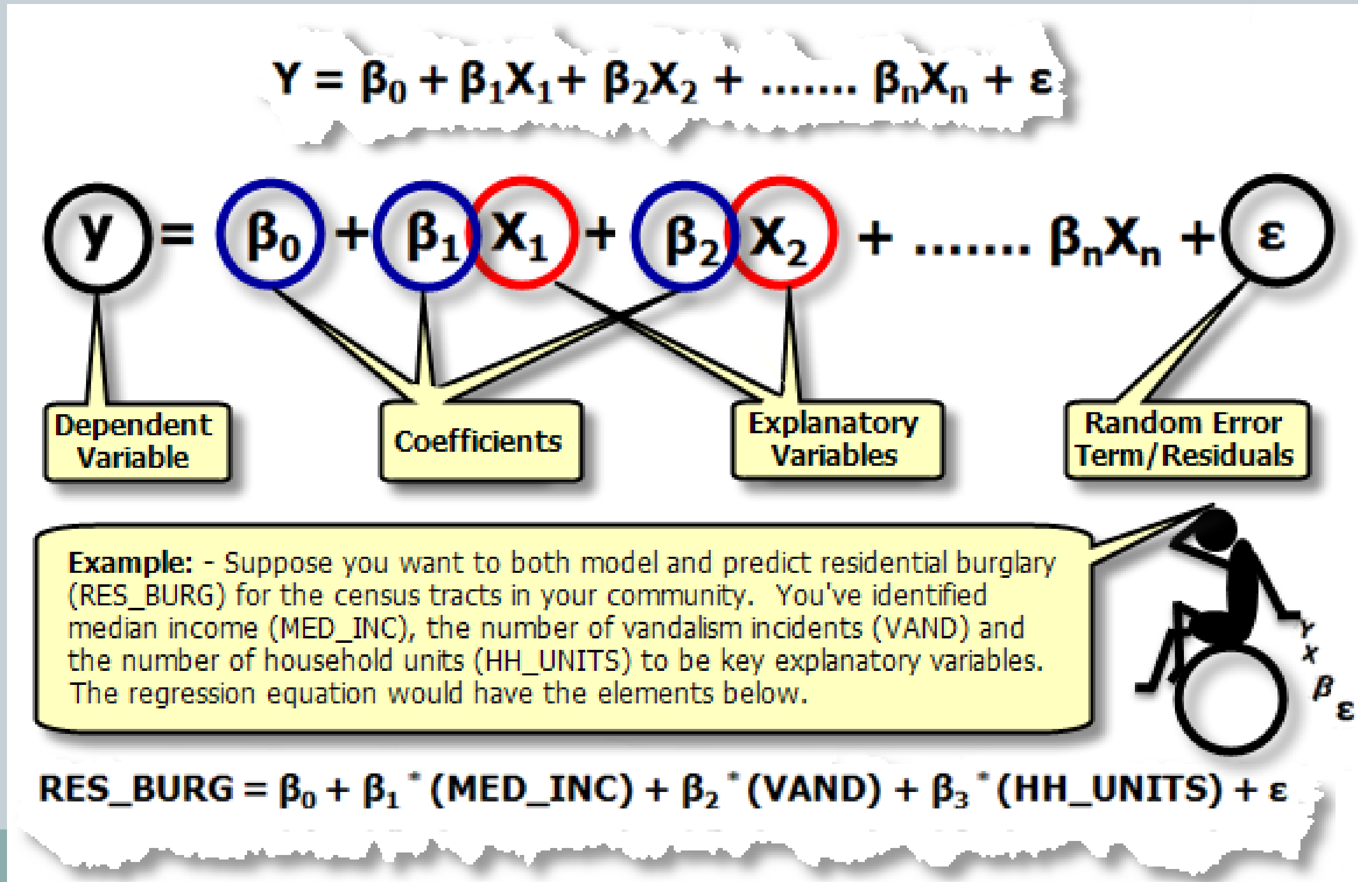
- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Greatly studied in statistics, neural network fields.

- Examples:
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

Regression

- The regression functions are used to determine the relationship between the dependent variable (target field) and one or more independent variables.
- The dependent variable is the one whose values you want to predict, whereas the independent variables are the variables that you base your prediction on.


y
a
x
b = slope of regression line
distance from the line to a typical data point
(= "error" between the line and this y value)


$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots \beta_n X_n + \varepsilon$
$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots \beta_n X_n + \varepsilon$
Dependent Variable
Coefficients
Explanatory Variables
Random Error Term/Residuals
Example: - Suppose you want to both model and predict residential burglary (RES_BURG) for the census tracts in your community. You've identified median income (MED_INC), the number of vandalism incidents (VAND) and the number of household units (HH_UNITS) to be key explanatory variables. The regression equation would have the elements below.
$RES\_BURG = \beta_0 + \beta_1 * (MED\_INC) + \beta_2 * (VAND) + \beta_3 * (HH\_UNITS) + \varepsilon$

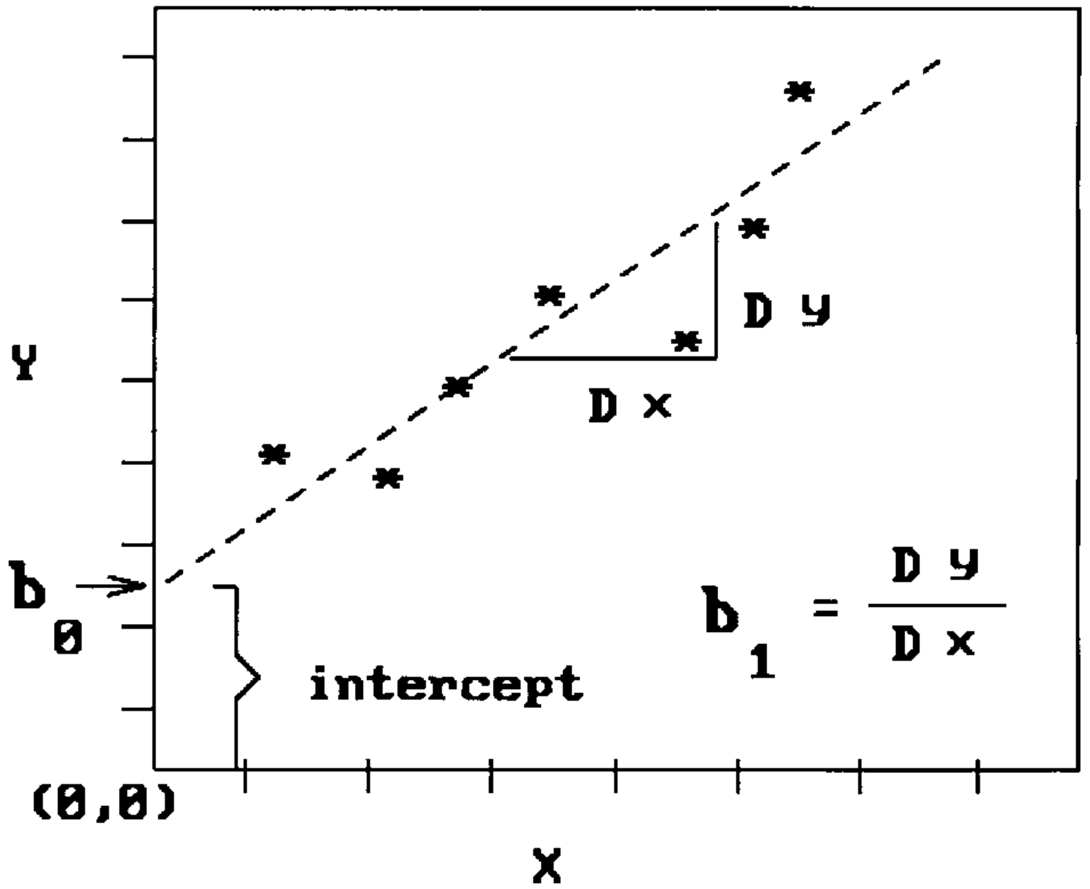
Regression

- Generally three types of regression: linear, polynomial, logistic

Example: Linear relationship (e.g. Y=cholesterol versus X=age)

$$Y = b_0 + b_1 X$$

b_0 is the intercept,
b_1 is the slope.



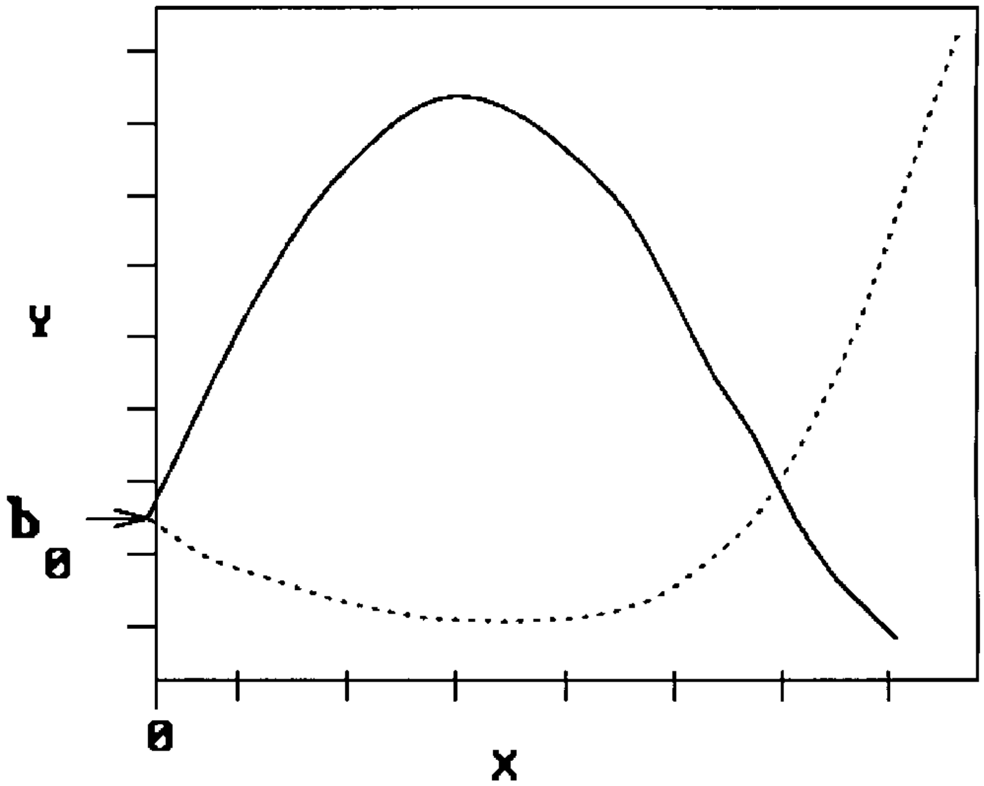
Example: Polynomial relationship (e.g. Y=crop yield vs. X=pH)

$$\boxed{Y = b_0 + b_1 X + b_2 X^2}$$

b_0: intercept,

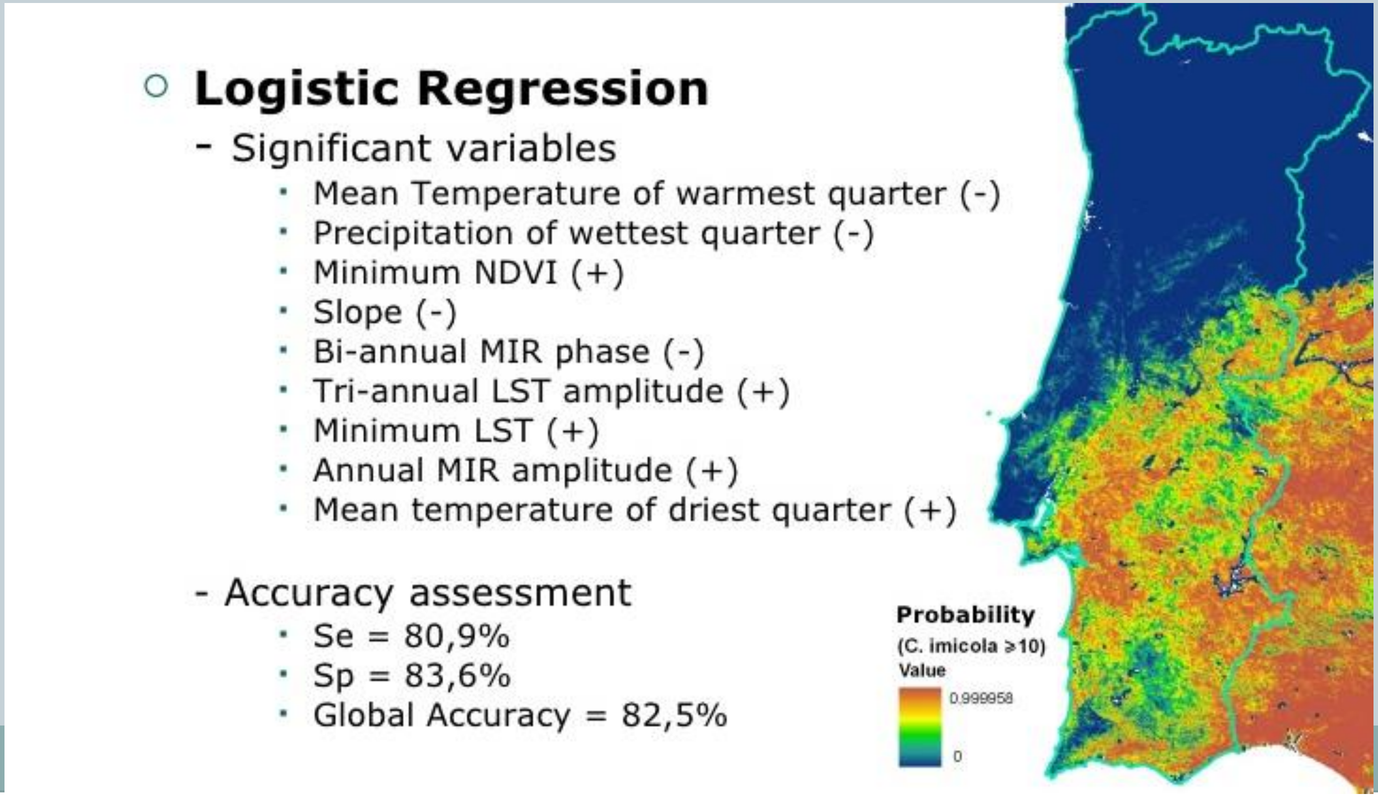
b_1: linear coefficient,

b_2: quadratic coefficient.



Logistic Regression

- Or logit, is a regression model where the dependent variable (DV) is categorical.
- Used a lot in remote sensing and GIS.



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