9. Data mining

LAKSHMAN@OU.EDU

Timeline for rest of semester

- Mar 28: Data mining
- Apr 4: Guest lecture (active contours: real solution to something that several of you are working on)
  - Extra credit of 10% for incorporating snake into term project or exam
- Apr. 11: In-class exam lasting a maximum of 3 hours
  - Bring your laptop
  - I'll assign a dataset and a problem
    - Solve it however you want
  - Open book, open computer but no discussion allowed
    - Use any software you like as long as you can explain what it does
- Apr 18 start of class: term projects are due
- Apr 18: term project presentations (Adam, Heather, Madison)
- Apr 25: term project presentations (Diana, Jason, Nicholas)
- May 2: term project presentations (Gifford, John)
- Class ends (nothing in finals week)

Being lazy …

- Just repurposing an invited talk I gave a couple of years ago
  - Will fill information/details as we go along

Data Mining for Weather Nowcasting

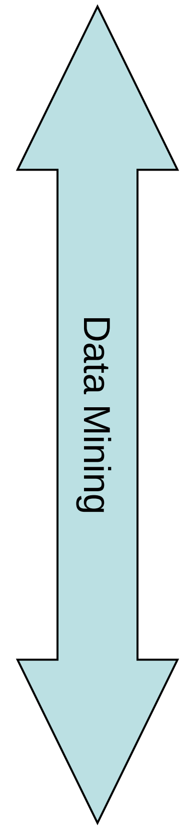
- What is Data Mining?
- Interactive and Exploratory Data Analysis
- Classification
- Retrieval and Approximation
- Unsupervised Learning
- Pre-and-post Processing
- Measures of Skill
- Data-driven Weather Applications

Data Mining

- Extracting useful information from large amounts of data
  - Closely related to applied statistics
    - Summarize information, identify outliers, etc.
  - Encompasses many research areas
    - Exploratory data analysis, visualization, signal/image processing, pattern recognition, information theory, machine intelligence
  - Many areas in common with artificial intelligence, knowledge discovery
- An engineering discipline
  - Concerned with how to perform operations on large collections of data
    - Computing on samples, and generalizing to larger collections
    - Calculating on data points one-by-one
- Theory (statistics) + Technique to work on practical data sets
  - Sub-optimal, non-provable approaches are common

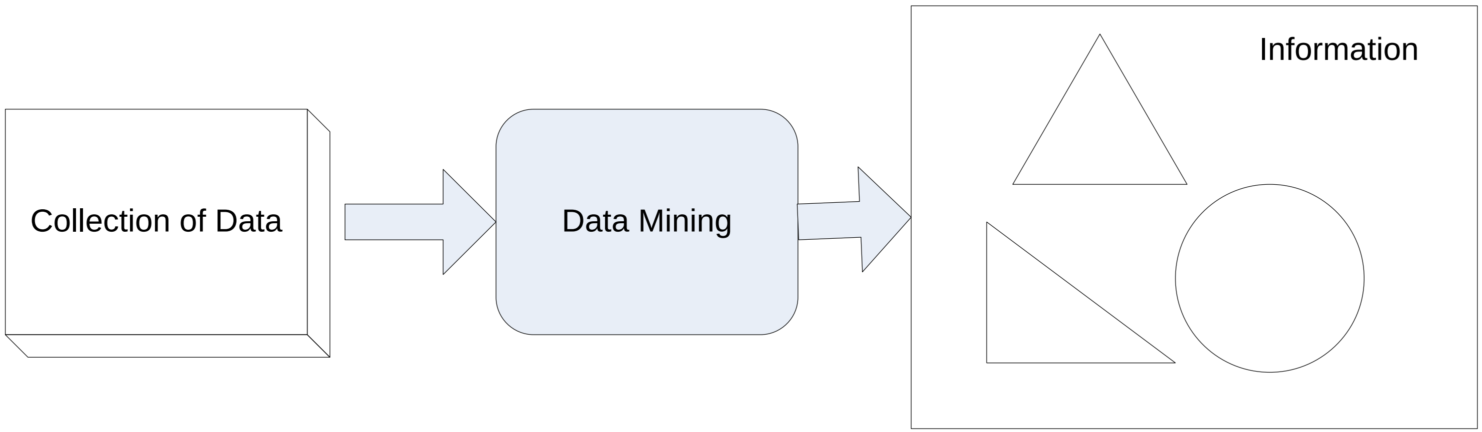
Example: Information Content

- Given a set of numbers, how can we determine how "interesting" the data collection is?
- Information theory
  - Higher the variance, the higher the information content
- Statistics $$\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$
- Computer Science
  - How can you compute the variance on a very large dataset?
  - How can you compute the variance in one pass?
  - Recast formula to use $\sum_{i=1}^{N} x_i$ and $\sum_{i=1}^{N} (x_i^2)$
    - Can both be updated during a single pass through data



Data Mining != Magic

- Data mining typically is a secondary concern
  - Techniques can work with whatever data are available



- However, data mining is not magic
  - Limited by the characteristics of the data
  - Limited by the questions that the users ask of the data

? Can I not just run through all the data looking for variables that are highly correlated?

Data Mining != Data Dredging

- Data mining requires care
  - Secondary analysis: so domain of data needs to be considered
  - Deals with large data sets: so significance tests need to be modified
- Should not automatically scan large amounts of data for any relationship
  - Due to chance, there will always be relationships between variables
    - Likelier to find statistically significant relationships in large data sets
    - Most likely, the relationship is spurious
  - Techniques exist to limit the potential for erroneous conclusions
    - Cross-validation
    - Set statistical significance threshold according to number of patterns expected based on size of data set
- Finding all possible correlations and constructing a plausible hypothesis to explain them is called “data dredging”

A Typical Data Mining Approach

- Given a collection of data:
  - Perform exploratory data analysis on it
    - Graph it, look at correlations, histograms, etc.
  - Visualize it in different ways
    - 3D, loops, projections, etc.
    - Gain gut-level feeling for how the data are organized
  - Build a conceptual model of parts of the data
    - Regression (fit model to data)
    - Pattern recognition (identify interesting combinations)
  - Test samples of the data against the model
    - Understand expected skill of model if used in automated manner
  - Use the model in automated manner
    - Replace costlier, more time-consuming method (Approximation)
    - Find patterns on new data (Prediction)

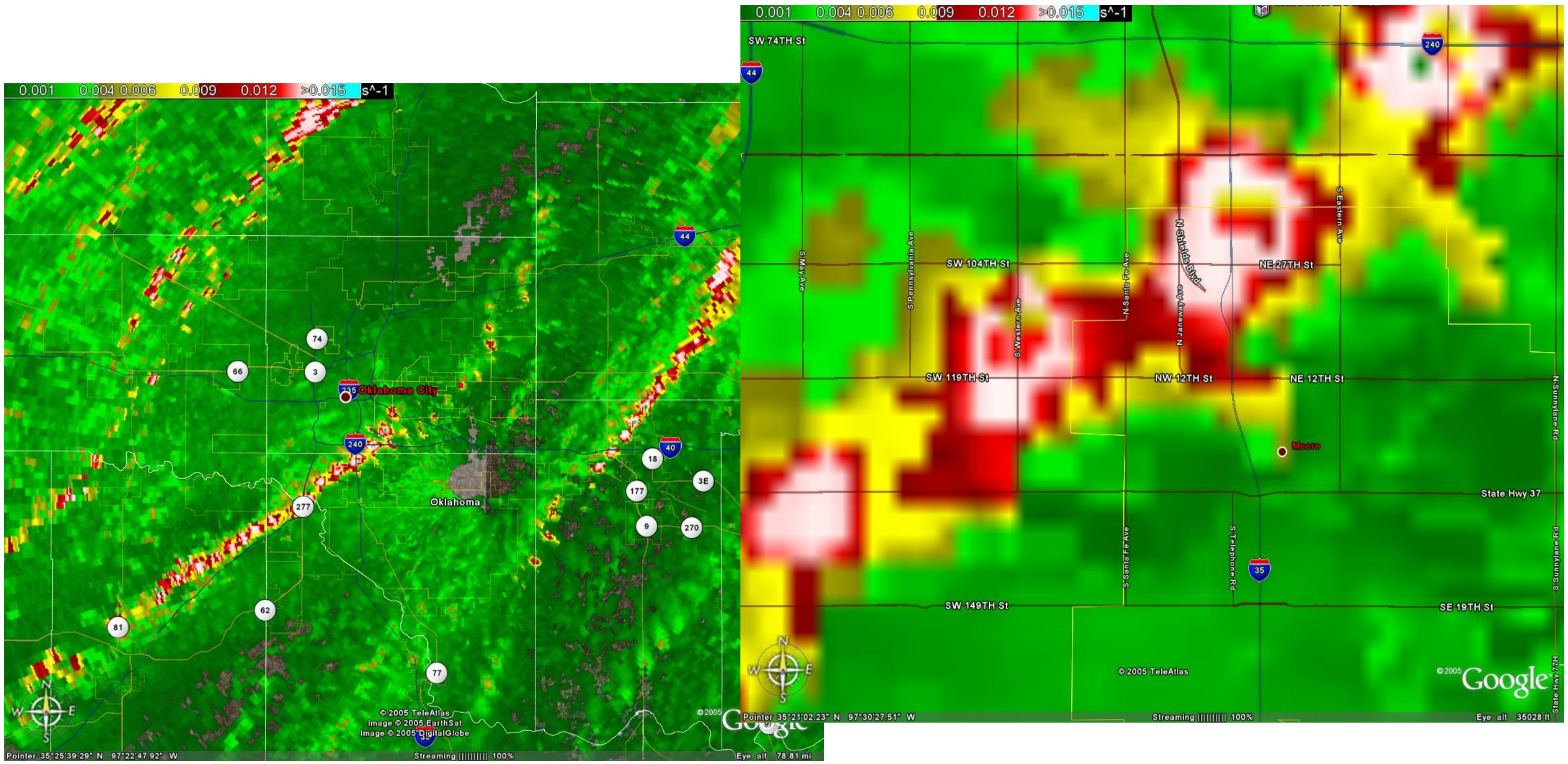
Data Mining for Weather Nowcasting

- What is Data Mining?
- ▶ Interactive and Exploratory Data Analysis
- Classification
- Retrieval and Approximation
- Unsupervised Learning
- Pre-and-post Processing
- Measures of Skill
- Data-driven Weather Applications

Role of Visualization

- Visualization of data is rarely the end in itself
  - Visualize data to extract information from it
  - Visualization needs to enable information retrieval
- Visualization needs to tie data into relevant conceptual framework
  - Geo-location, height, time, event, other sensors, numerical models

“Rotation Tracks” – path taken by intense low-altitude circulations

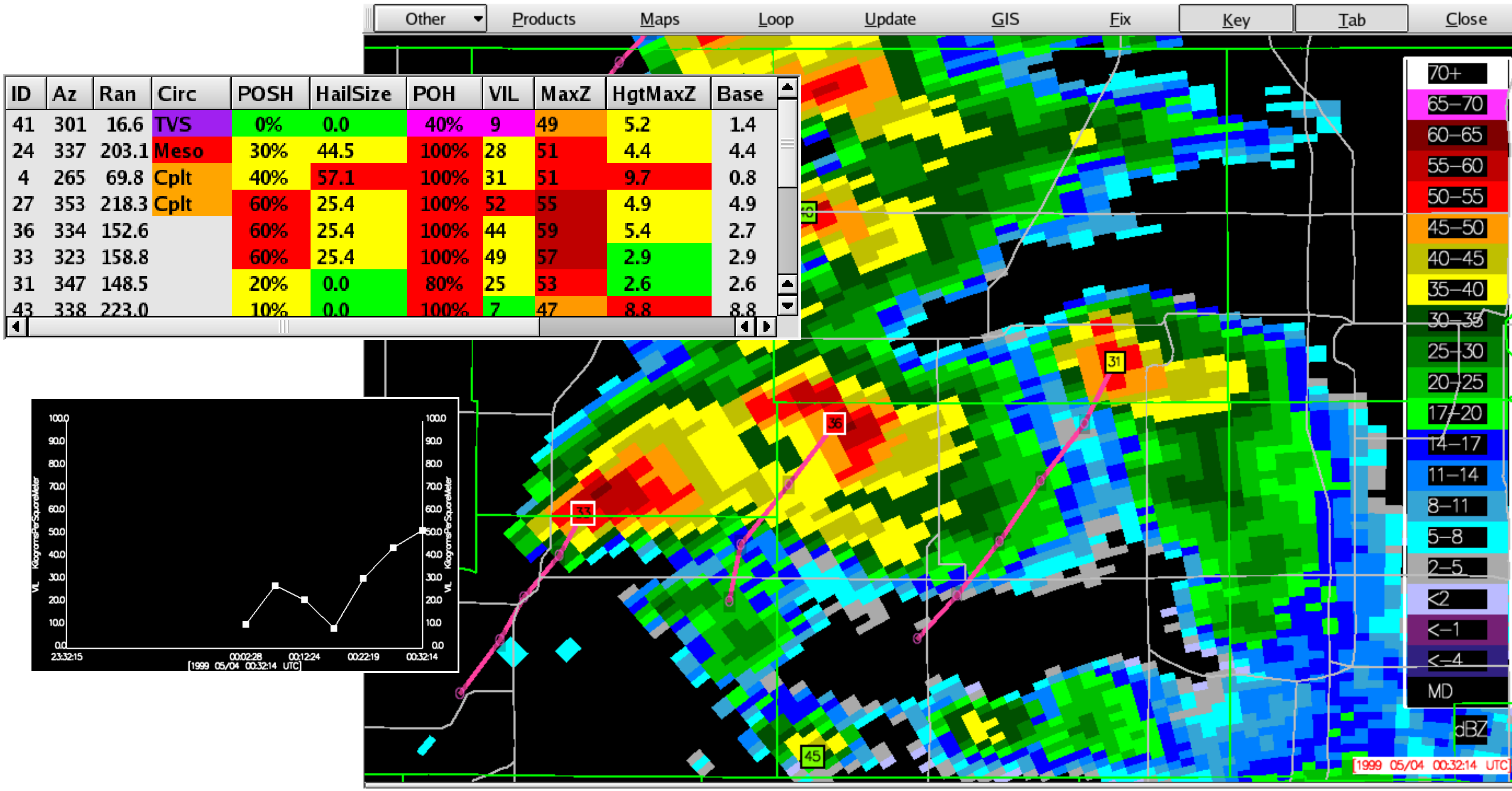


“Where should we send the damage survey team?” “Did it pass near Grandma Jones’ house?”

Decision Support Systems

- Humans can extract information from data sets
  - Make decisions by visualizing data
  - Humans do the data mining
  - Pattern recognition skills of humans far exceeds that of computers
- Human data mining may not always scale
  - Humans can not process large data sets
  - Problems of objectivity, fatigue
- Role of data mining algorithms when humans make decisions
  - Summarize the data using models, reductions, projections, metrics
  - Extract interesting sections of data for human analysis

Storm Cell Identification and Tracking (SCIT)



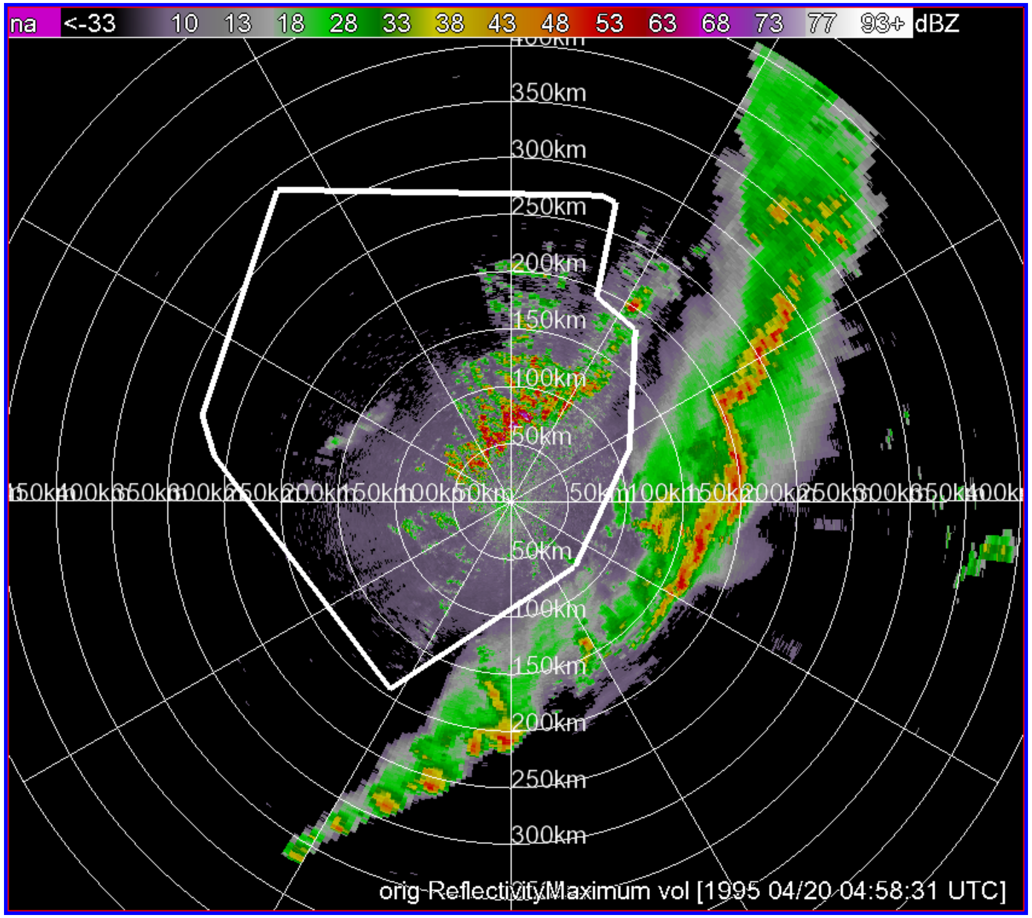
Visualization For Automated Analysis

- Visualization is essential for exploratory data analysis
  - Allow designer of data mining algorithm to get gut-level feel for data
  - Determine if there is something wrong with the data
- Humans can train automated algorithms on samples of data
  - Identify regions of interest
  - Then let the automated algorithm loose on full data set

Human Truthing for Quality Control

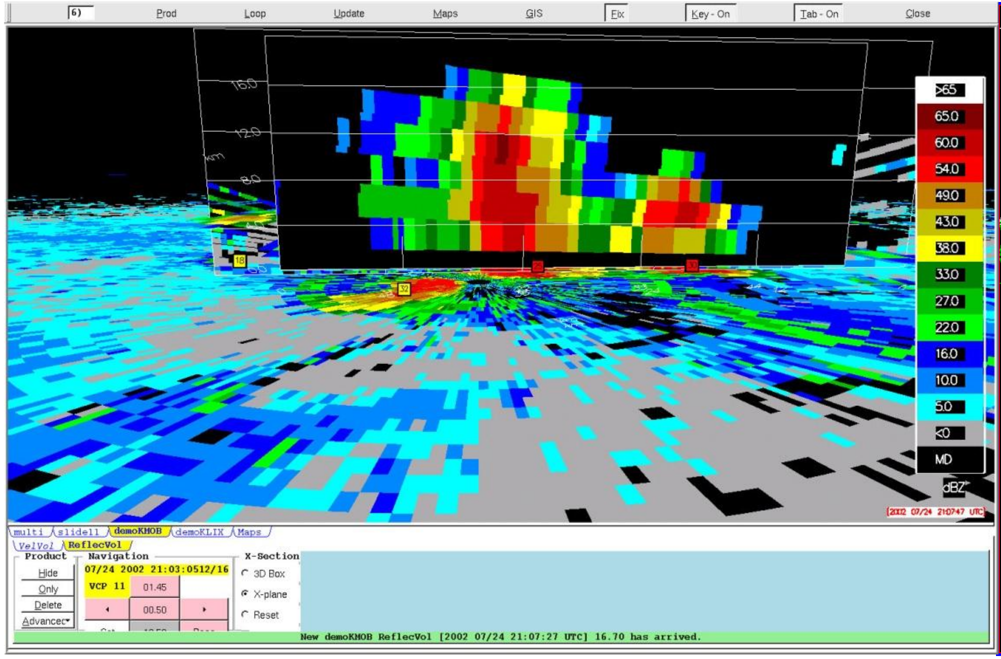
- Looked at loops
- Examined radar data
- Other sensors
- Considered terrain, time of day, etc.
- Identified bad echoes

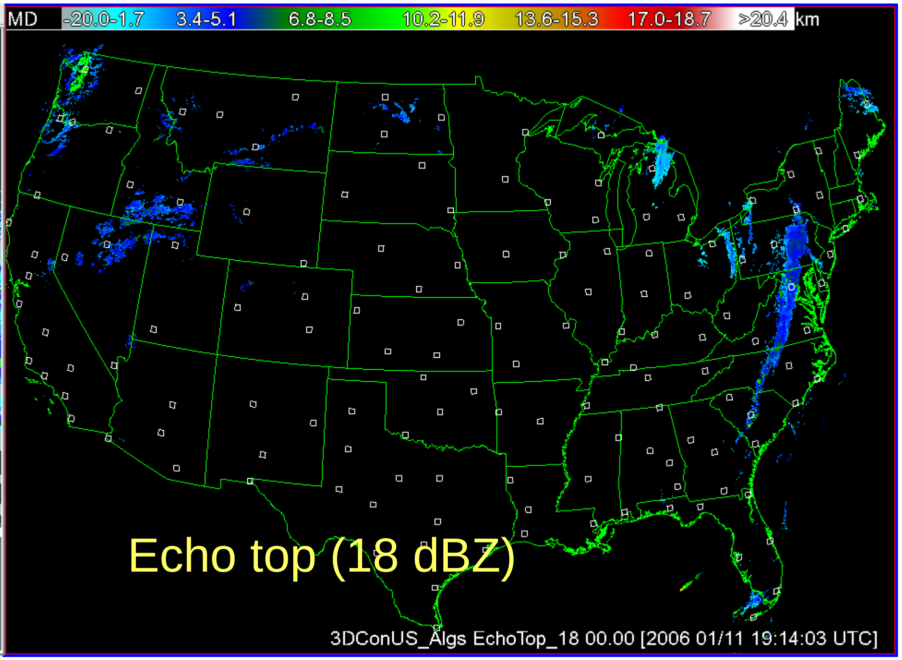
- Dataset was then used to train automated algorithm



Raw Data vs. Derived Products

- Visualizing the raw data may not always be the best choice
  - Tying the data into a framework may be more important
  - May want to visualize “derived” products





Data Mining for Weather Nowcasting

- What is Data Mining?
- Interactive and Exploratory Data Analysis
- ▶ Classification
- Retrieval and Approximation
- Unsupervised Learning
- Pre-and-post Processing
- Measures of Skill
- Data-driven Weather Applications

The Classification Problem

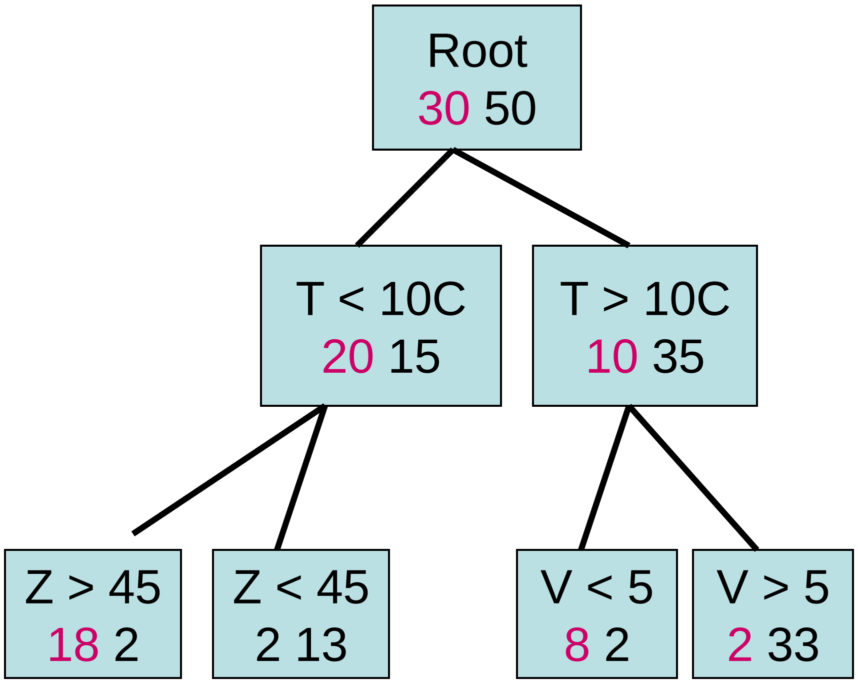
- A common requirement is to determine whether a particular data instance falls into a category
  - “If the radar reflectivity is x1, temperature is x2”, is it raining?
  - Can easily become a prediction
    - If shear is x1, reflectivity is x2, temperature is x3, is hail likely?
- Two broad ways to answer this question:
  - Physical reasoning
    - What types of storms produce reflectivity of x1 at a temperature of x2?
  - Data-driven
    - Use assortment of radar and temperature measurements with “truth” value
    - Train data mining classifier algorithm on dataset that has ground truth
    - Run data mining on new data

Data-driven Classifiers

- Possible to create automated classification algorithm from data
  - Decision Trees
    - Optimal if-then rules to separate data into classes
  - Fuzzy Logic
    - Represent human knowledge as rules
  - Genetic Algorithms
    - Fine-tune parameters to some set model using data
  - Neural Networks
    - Fit “black-box” model to create target result from data
- Each of these methods has its strengths and weaknesses
  - Use them in different situations

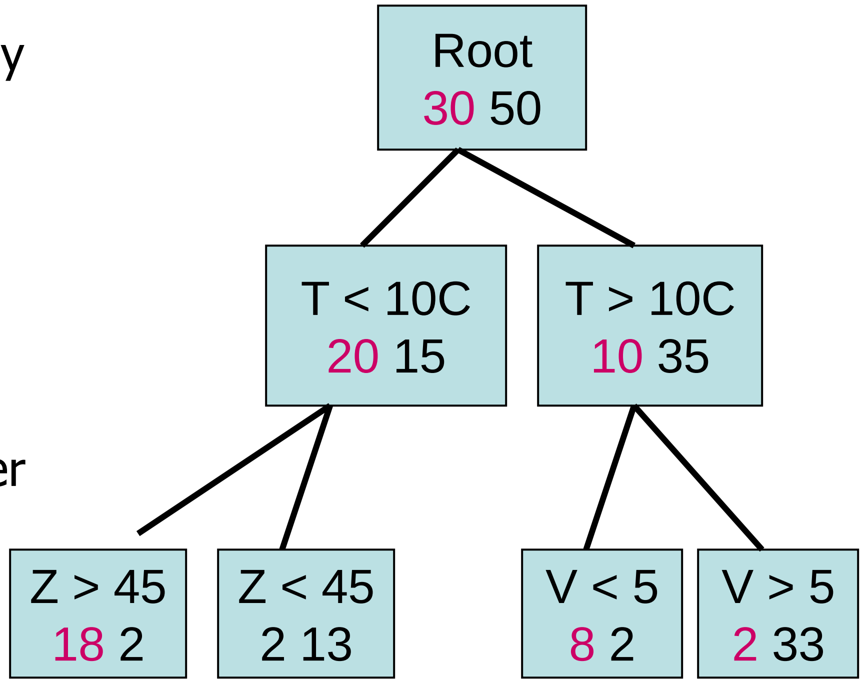
Decision trees

- Can automatically build decision trees from tagged data
- Attributes chosen based on "information gain"
  - How much entropy change if we divide up the classes based on this variable at this threshold?
- Information gain often biased towards too many splits on attributes with many values
  - Entropy = p log (p)
  - So when p is small, entropy can be larger
- Information gain ratio is better
  - The numerator is information gain
  - Denominator is entropy due to the split (not on the whole training case, just the part of this tree)



Decision trees

- Disadvantages
  - Piece-wise linear, so typically less skilled than neural networks
  - Large decision trees are effectively a blackbox
  - Can not do regression, only classification
- Advantages:
  - Fast to train
  - New advances: bagged, boosted decision trees approach skill of neural networks, but are no longer fast to train

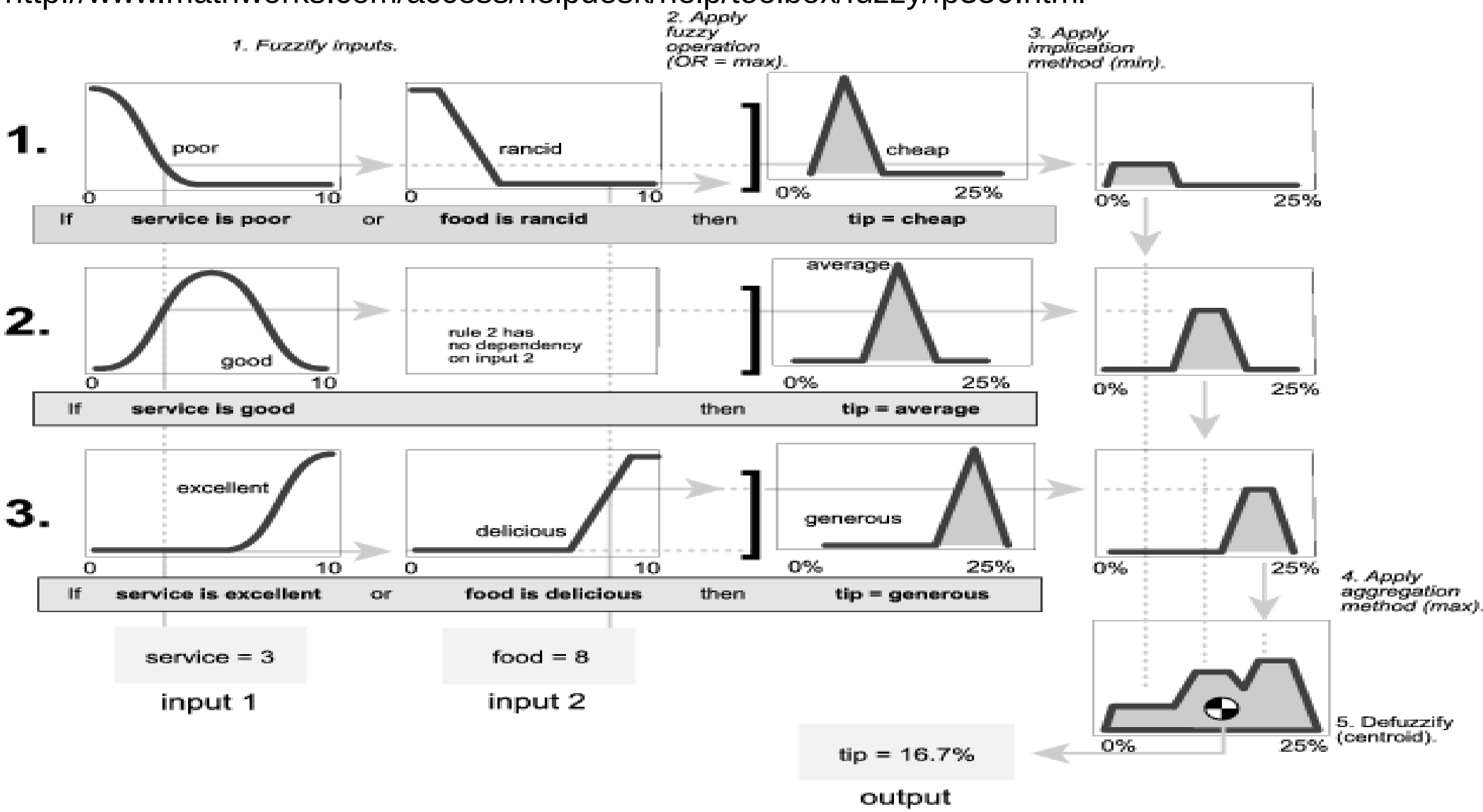


Fuzzy Logic

- Fuzzy logic addresses key problem in expert systems
  - How to represent domain knowledge
  - Humans use imprecisely calibrated terms
  - How to build decision trees on imprecise thresholds
- Fuzzy rules are easy to set up and troubleshoot
  - Humans suggest the rules
  - The encoding of the rules is easily understandable

Fuzzy logic example: How Much To Tip

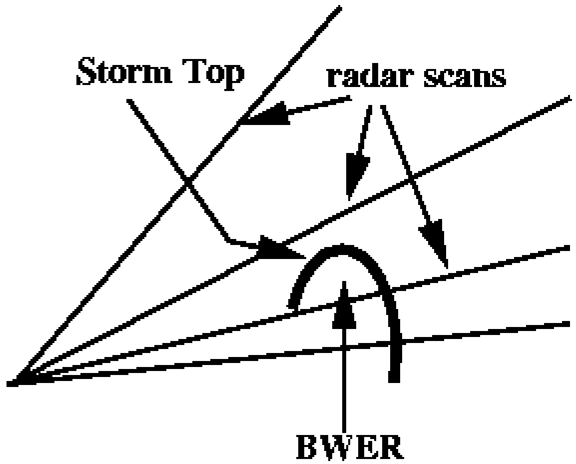
Source: Matlab fuzzy logic toolbox tutorial
http://www.mathworks.com/access/helpdesk/help/toolbox/fuzzy/fp350.html

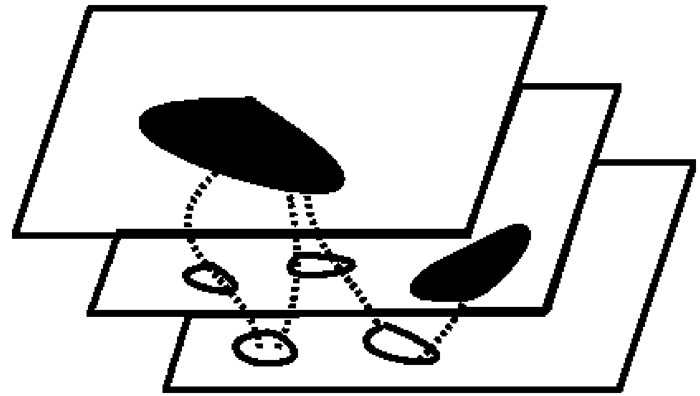


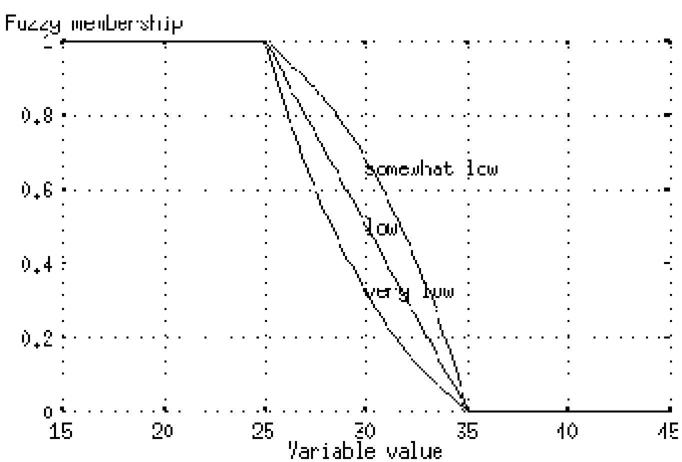
Fuzzy Logic: Pros and Cons

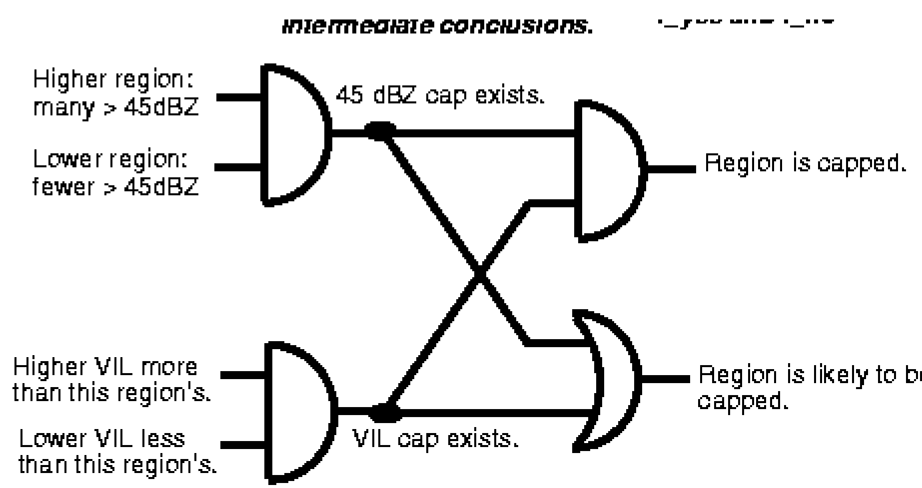
- Considerable skill for little investment
  - Fuzzy logic systems piggy bank on human analysis
    - Humans encode rules after intelligent analysis of lots of data
    - Verbal rules generated by humans are robust
  - Simple to create
    - Not much need for data or ground truth
    - Logic tends to be easy to program
- A fuzzy logic system is limited
  - Piece-wise linear approximation to a system
  - Non-linear systems can not be approximated
- Do not use fuzzy logic if humans do not understand the system
  - Different experts disagree
  - Knowledge can not be expressed with verbal rules
  - Gut instinct is involved
    - Not just objective analysis

Fuzzy Logic Algorithm for BWER Detection



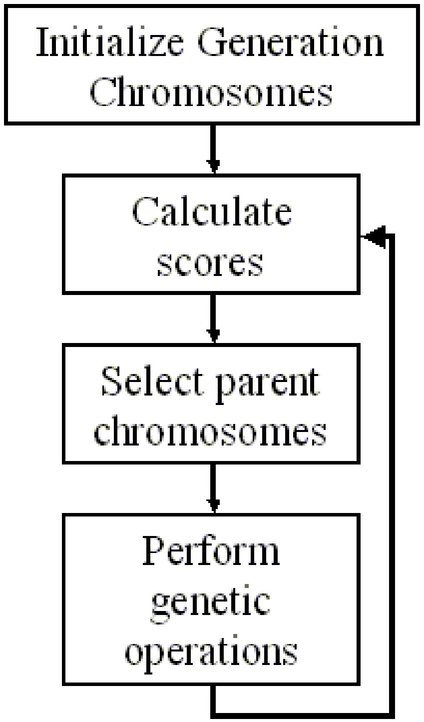






Genetic Algorithms

- Genetic algorithms are based around "survival of fittest"
  - Start with population of solutions
  - Breed the solutions together using crossover
    - Choosing parents who are more fit with higher probability
  - Mutate the solutions randomly
  - Over time, the population becomes more fit
- In genetic algorithms
  - One fixes the model
    - Rule base, equations, class of functions, etc.
  - Optimize the parameters to model on training data set
  - Use optimal set of parameters for unknown cases



Genetic Algorithms: Pros and Cons

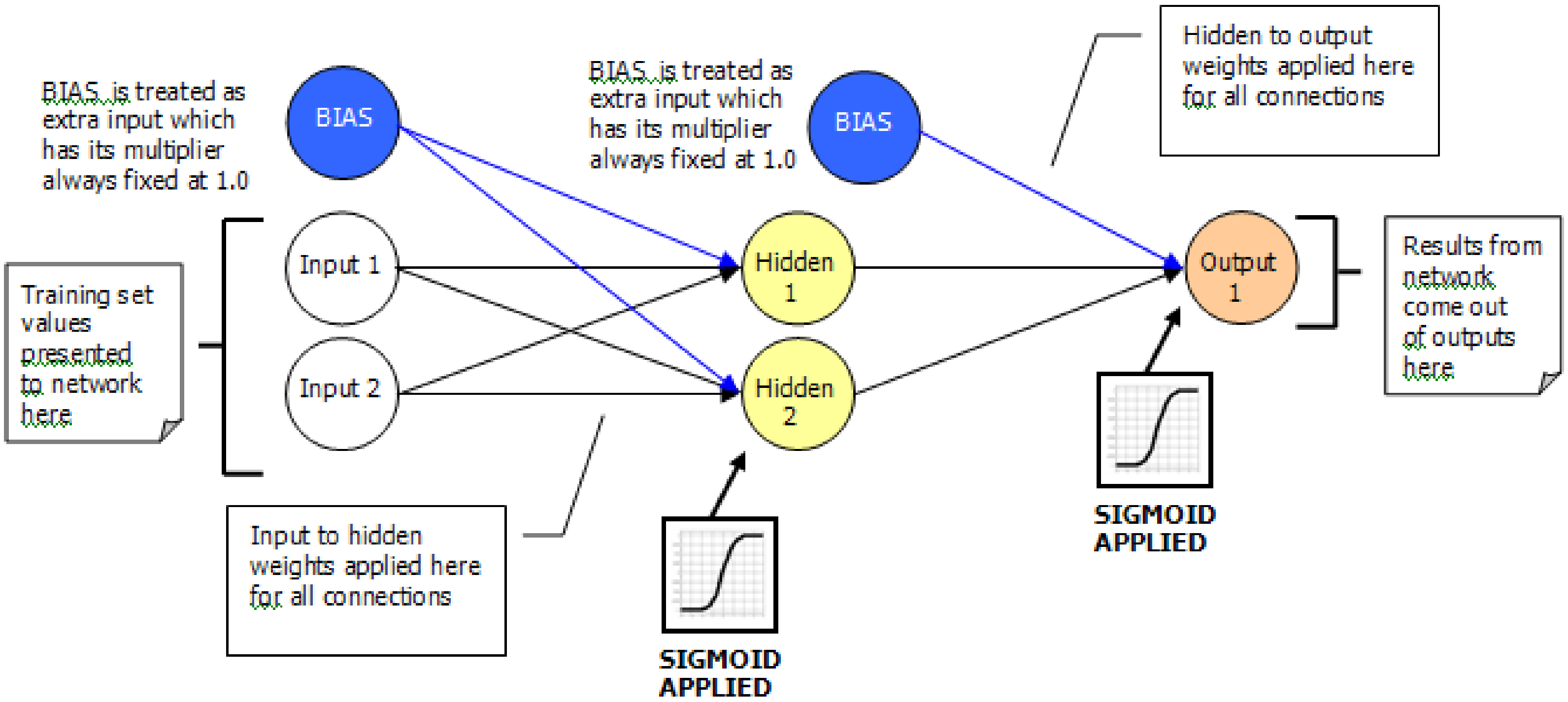
- Genetic algorithms provide near-optimal parameters for given model
  - Human-understandable rules, and best parameters for them
  - Cost function need not be differentiable
    - The process of training uses natural selection, not gradient descent
  - Requires less data than a neural network
    - Search space is more limited
- Performance is highly dependent on class of functions
  - If poor model is chosen, poor results
    - Optimization may not help at all
  - Known model does not always lead to better understanding
    - Magnitude of weights, etc. may not be meaningful if inputs are correlated
    - Problem may have multiple parametric solutions

Neural Networks

- Neural networks can approximate non-linear systems
  - Evidence-based
    - Weights chosen through optimization procedure on known dataset
    - Works even if experts can't verbalize their reasoning
    - Can be used as long as there is ground truth
- Typically, dataset split into three parts:
  - Training data set
    - Optimization procedure chooses weights to minimize error on this data set
  - Validation data set
    - Used to stop the training when optimization starts to overfit
    - Used to choose structure of neural network
  - Testing data set
    - Used to verify that the neural network generalizes to unseen data

How Neural Networks Work

Diagram from:
http://www.codeproject.com/useritems/GA_ANN_XOR.asp



Neural Networks: Pros and Cons

- Neural networks are general-purpose, easy to train and efficient at runtime
  - The three-layer neural network can approximate any smooth function
  - If output node is a sigmoid, can yield true probabilities
  - Training process (back propagation, ridge regression, etc.) are well understood optimization procedures
    - Heuristics to minimize problems of local minima, over-fitting, generalization
  - Efficient and easy to implement
    - Just a sum of exponential functions
    - Once trained, can calculate the output for a set of inputs quite fast
- Neural network training is an art form
  - Training set has to be complete and voluminous
  - Unpredictable output on data unlike training
  - Measure of skill needs to be continuous (e.g: entropy, RMS error)
- A working neural network yields no insights
  - Magnitude of weights doesn't mean much: "A black box"

Data Mining for Weather Nowcasting

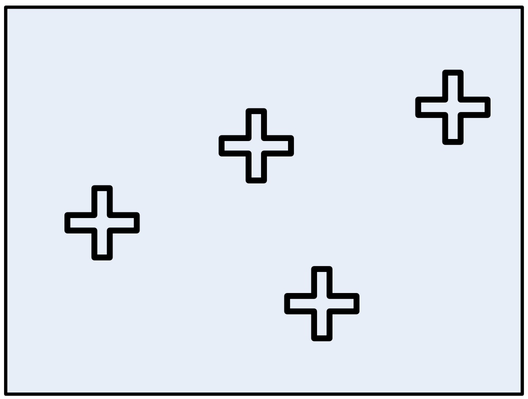
- What is Data Mining?
- Interactive and Exploratory Data Analysis
- Classification
- ▶ Retrieval and Approximation
- Unsupervised Learning
- Pre-and-post Processing
- Measures of Skill
- Data-driven Weather Applications

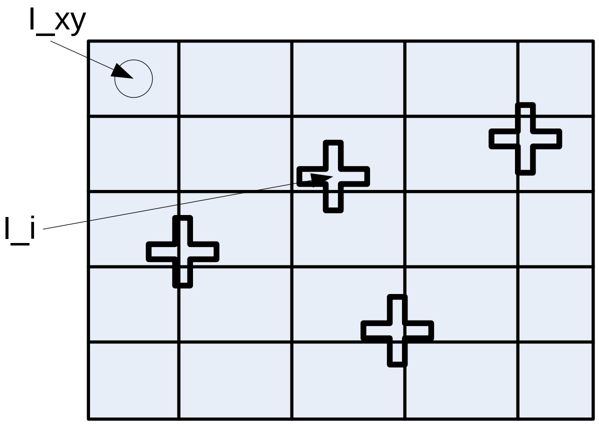
Approximation Techniques

- Several retrieval and approximation techniques:
  - Interpolation: the same data, just at higher resolution
    - Interpolation, objective analysis
  - Regression: create function to approximate data or process
    - Linear regression
    - Non-linear regression
      - Neural networks can do this

Gridding Observations

- Some environmental data are measured by in-situ instruments
  - Not remotely sensed
  - These measurements are at geographically dispersed points
  - Need to be converted into grids





Pixel Resolution

- If the chosen pixel resolution is too coarse, lose some of the observations
  - If the chosen pixel resolution is too fine, strong gradients may not be apparent
  - Choose pixel resolution to be half the mean distance between observation points
- Interpolation methods:
  - Cressman analysis
  - Barnes analysis
  - Kriging

$$I_{xy} = \frac{\sum_i I_i \frac{R^2-(x-x_i)^2-(y-y_i)^2}{R^2+(x-x_i)^2+(y-y_i)^2}}{\sum_i \frac{R^2-(x-x_i)^2-(y-y_i)^2}{R^2+(x-x_i)^2+(y-y_i)^2}}$$

Interpolation Techniques

- Cressman Analysis
  - Every observation gets weighted based on its distance from grid point
  - R is the "radius of influence"
  - Higher the R, the more distant points are considered
- The problem with a Cressman analysis:
  - Even if a pixel is collocated with an observation, the final value of pixel will be different
  - Due to (small) effect of pixels further away
- Barnes analysis: perform Cressman analysis on observation and errors
  - Compute errors at observation points
  - Perform Cressman analysis on errors, then add weighted error
  - N-pass Barnes analysis: closer and closer to the value of the observations at the observation points.
- Kriging is a principled method of interpolation using correlation between observations

Retrieval and Approximation

- Sometimes, a dataset is very hard to collect or very expensive to compute
  - Can we create an approximation of that data from data that are cheap and plentiful?

- For example:
  - Can I get the best estimate of rainfall given the radar reflectivity and satellite infrared temperature?
  - Train the system using rainfall collected by rain gauges
    - Rain gauges are not everywhere, so this data set is not enough
  - Use the trained system to evaluate rainfall everywhere else
    - Radar and satellite provide widespread coverages but don’t measure rainfall

Using NN To Solve Inverse Problem

- Equation of state to estimate salinity estimated at each time step
  - Involves solving (T=temperature, S=salinity, P=pressure) step by step

$$\rho(\boldsymbol{T}, \boldsymbol{S}, \boldsymbol{P}) = \frac{\rho(\boldsymbol{T}, \boldsymbol{S}, 0)}{1 - \dfrac{\boldsymbol{P}}{\boldsymbol{K}(\boldsymbol{T}, \boldsymbol{S}, \boldsymbol{P})}}$$

  - An inverse problem with 40 parameters that is very time consuming
- Instead, approximate it by a single-step neural network transfer function
  - With different weights at different locations

$$\boldsymbol{f} = \boldsymbol{b} + \boldsymbol{a} \tanh\{ \sum_{j=1}^{k} \omega_j [\tanh(\sum_{i=1}^{n} \Omega_{ji} \boldsymbol{x}_i + \boldsymbol{B}_j)] + \beta \}$$

  - Use simulated data set to train network (answer known at all points)
- Very, very close approximation that is significantly faster to compute

Source: Vladimir Kransopolksy, National Centers for Environmental Prediction

Data Mining for Weather Nowcasting

- What is Data Mining?
- Interactive and Exploratory Data Analysis
- Classification
- Retrieval and Approximation
- ▶ Unsupervised Learning
- Pre-and-post Processing
- Measures of Skill
- Data-driven Weather Applications

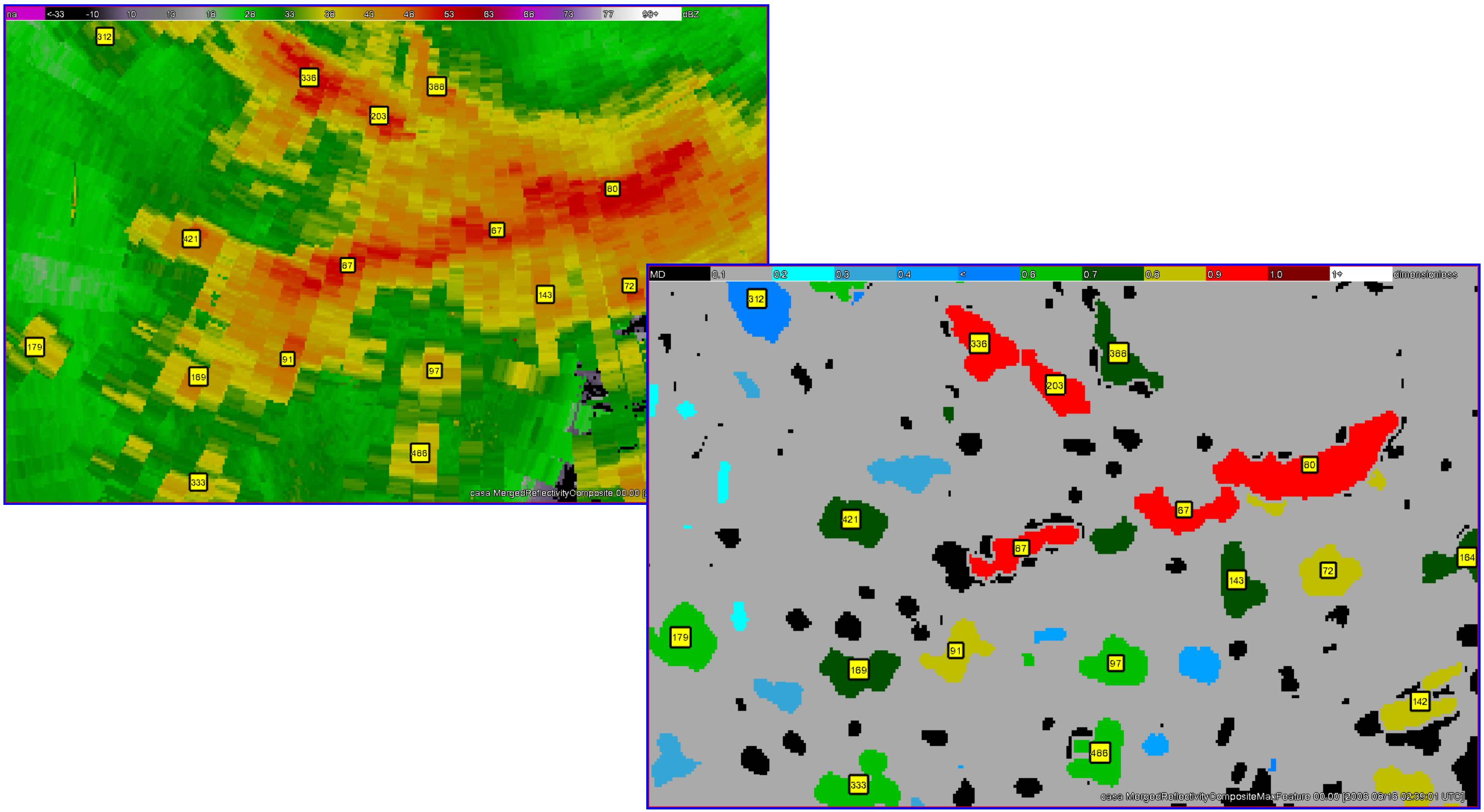
Unsupervised Learning

- In both classification and regression, a “truth” data set exists
  - What if you don’t know what categories are available in data?
  - Can a data mining algorithm provide “natural categories” in the data?
    - Called clustering
- Clustering techniques
  - Principal components analysis
    - Find linear combinations of parameters that explain most of the variance in data
  - K-Means
    - Find cluster centers so that inter-cluster variance is minimized
- Application of clustering: segmentation
  - Finding groups of pixels that together comprise an object

Region Growing

- Region growing can be applied only to binary images.
  - Apply to images that can be thresholded
  - Pixels below (above) a threshold are not of interest.
  - Your image now has only 0's (not interested) and 1's (interested)
- To void noisy regions when thresholding images, employ "hysterisis"
  - You are interested in areas > 30 dBZ (T_i – threshold of interest)
  - But if you see a 30 dBZ pixel, you lose interest only if pixel value < 20 dBZ (T_d – threshold of drop off)
  - Less chance of too many disconnected regions.
  - Choose T_i and T_d by experiment.

Example of Region Growing



Vector Segmentation

- Region growing:
  - Can lead to disconnected regions, even with hysterisis
  - Not hierarchical: can not always use hierarchy of thresholds
- Need a segmentation approach that:
  - Can avoid disconnected regions
  - Is hierarchical
- Contiguity-enhanced segmentation
  - Explicitly trade-off contiguity of region with the idea that a region’s pixels ought to be similar
  - Similarity of pixels based on a number of texture or wavelet features
  - Pixels are dissimilar if vector distance between features is large

Data Mining for Weather Nowcasting

- What is Data Mining?
- Interactive and Exploratory Data Analysis
- Classification
- Retrieval and Approximation
- Unsupervised Learning
- Pre-and-post Processing [skip]
- Measures of Skill
- Data-driven Weather Applications

Data Mining for Weather Nowcasting

- What is Data Mining?
- Interactive and Exploratory Data Analysis
- Classification
- Retrieval and Approximation
- Unsupervised Learning
- Pre-and-post Processing
- Measures of Skill
- Data-driven Weather Applications

Measuring Skill

- Once you have automated technique to extract information from data
  - Need to measure skill of technique
  - How often does it get it right?
  - What is the typical error?
- Regression skill measured based on root mean square error, etc.
- Classification skill measured based on contingency matrix

Number of 0's correctly identified as 0's (correct **nulls**)	Number of 0's misidentified as 1's (**false alarms**)
Number of 1's misidentified as 0's (**misses**)	Number of 1's correctly identified (**hits**)

- Also summarized by Accuracy, POD, FAR, CSI, HSS, TSS, etc.

Performance Assessment

Radar-only QCNN with no seasonal targeting

POD: Probability of detection of "good" echo (fraction of good echo retained)

FAR: Fraction of echoes in final product that are "bad"

CSI: Critical success index

HSS: Heidke skill score

Product	Data range	Measure	No QC	REC	QCNN
Composite	$> 0\ dBZ$	CSI	0.61 +/- 0.06	0.59 +/- 0.057	0.86 +/- 0.011
		FAR	0.39 +/- 0.06	0.4 +/- 0.06	0.02 +/- 0.0072
		POD	1 +/- 0	0.96 +/- 0.0031	0.88 +/- 0.0088
		HSS	0.89 +/- 0.02	0.88 +/- 0.019	0.98 +/- 0.0016
Composite	$> 10\ dBZ$	CSI	0.68 +/- 0.071	0.66 +/- 0.069	0.96 +/- 0.0083
		FAR	0.32 +/- 0.071	0.32 +/- 0.073	0.02 +/- 0.007
		POD	1 +/- 0	0.94 +/- 0.0023	0.92 +/- 0.0039
		HSS	0.93 +/- 0.017	0.93 +/- 0.016	0.99 +/- 0.0011
Composite	$> 30\ dBZ$	CSI	0.92 +/- 0.02	0.84 +/- 0.014	1 +/- 0.00072
		FAR	0.08 +/- 0.02	0.09 +/- 0.011	0 +/- 0.00057
		POD	1 +/- 0	0.92 +/- 0.0065	1 +/- 0.00029
		HSS	1 +/- 0.00064	0.99 +/- 0.00052	1 +/- 0
Composite	$> 40\ dBZ$	CSI	0.91 +/- 0.023	0.8 +/- 0.013	1 +/- 0.00038
		FAR	0.09 +/- 0.023	0.1 +/- 0.0074	0 +/- 0.00039
		POD	1 +/- 0	0.88 +/- 0.0088	1 +/- 0
		HSS	1 +/- 0.00016	1 +/- 0.00018	1 +/- 0
VIL	$> 0\ kg/m^3$	CSI	0.53 +/- 0.16	0.48 +/- 0.13	1 +/- 0.0011
		FAR	0.47 +/- 0.16	0.49 +/- 0.15	0 +/- 0.00053
		POD	1 +/- 0	0.9 +/- 0.0078	1 +/- 0.00084
		HSS	0.97 +/- 0.0091	0.97 +/- 0.0085	1 +/- 0
VIL	$> 25\ kg/m^3$	CSI	1 +/- 0.0022	0.65 +/- 0.033	0.99 +/- 0.0027
		FAR	0 +/- 0.0022	0.19 +/- 0.025	0 +/- 0.0022
		POD	1 +/- 0	0.76 +/- 0.026	1 +/- 0.00075
		HSS	1 +/- 0	1 +/- 0	1 +/- 0

Visual quality
95 to 97%

Effect on precip algorithms
99.9 to 100%

Effect on severe weather algorithms
99.9 to 100%

Data Mining for Weather Nowcasting

- What is Data Mining?
- Interactive and Exploratory Data Analysis
- Classification
- Retrieval and Approximation
- Unsupervised Learning
- Pre-and-post Processing
- Measures of Skill
- Data-driven Weather Applications [skip]