Data Mining

Brad Morantz PhD

bradscientist@machine-cognition.com

bradscientist@ieee.org

Apology

We could easily spend two or three semesters going through this topic.

We have one evening.

So, we will hit the highlights which will give you areas to go learn about.

Perspective

I am a decision scientist. I focus on the problem, not the computer.

The computer is a tool for problem solving as are the various programs.

An old saying: If your only tool is a hammer, then all of your problems look like nails

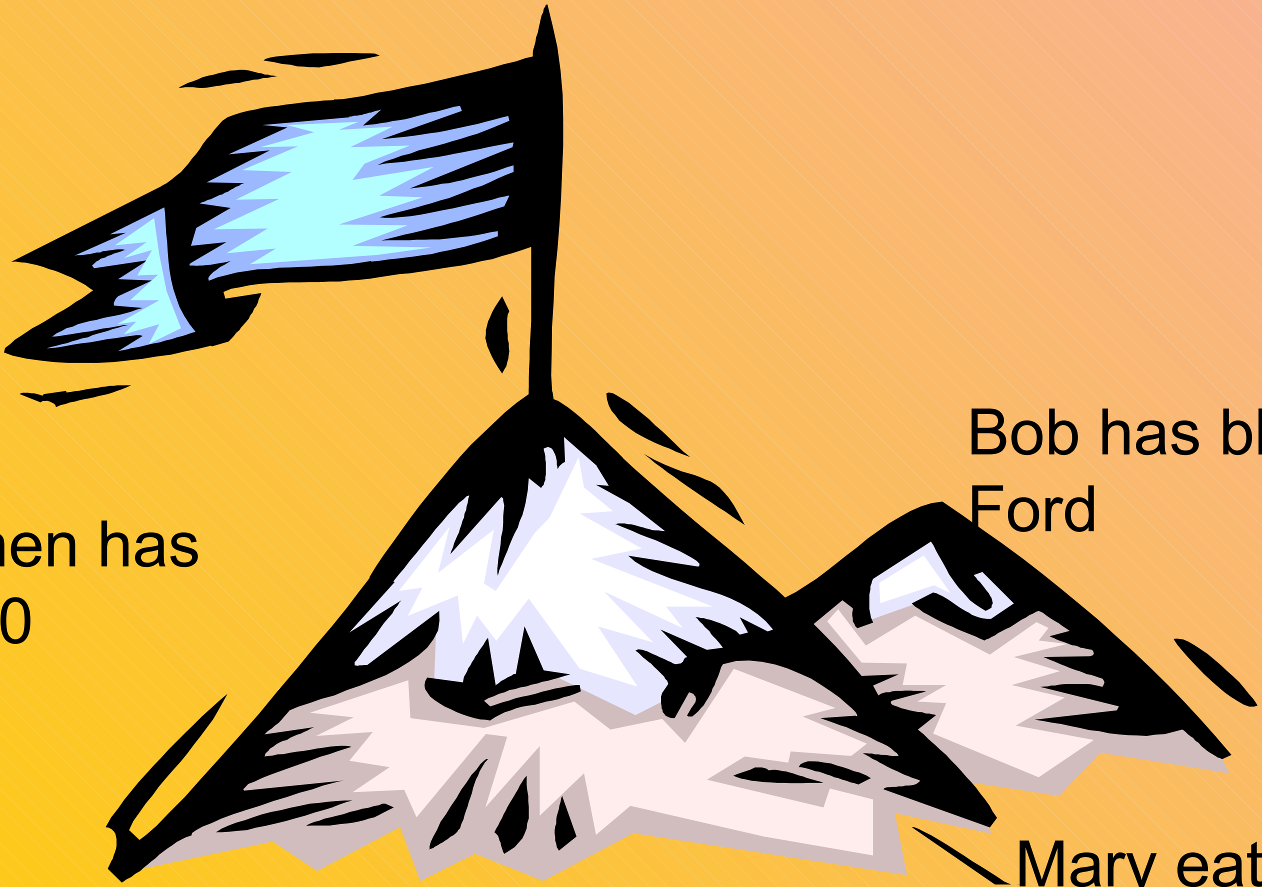
What is a Mountain?

- A big pile of dirt!

Data ?

- Some people would say that data on them is . . .
- Dirt
- A big pile of data is a . . .
- Mountain of data or a data warehouse

Mountain of data



Chen has
$10

Bob has blue
Ford

Mary eats rice krisp

Notice in Last Example . . .

- Different people
- Different facts
- Unconnected?
  - Maybe:
    - Bob bought the Ford from Chen (for $10?)
    - Mary bought the Rice Krisp from Chen
    - Maybe Bob delivers the Rice Krisp in his Ford
    - Maybe something else
    - Maybe nothing

Definitions

- Data Mining is the nontrivial extraction of implicit previously unknown and potentially useful information and data. (James Buckley)
- Machine Learning employs search heuristics to uncover interesting and systematic relationships in data. (Dhar & Stein)

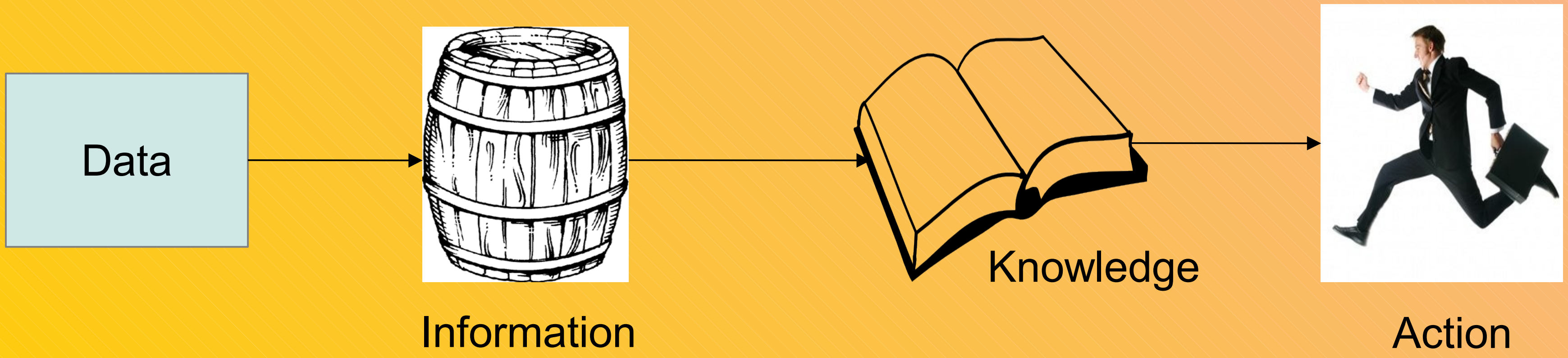
Data Mining

- Going into a data base and looking for patterns and relationships
- Use of historical data to find patterns and improve future decisions
- Also called KDD
  - Knowledge discovery from databases

Order of Transformation

Convert/transform:

1. Data to Information
2. Information to knowledge
3. Knowledge to action



Data

Information

Knowledge

Action

Uses

- Merchandising
  - Association rules
  - Customer profiling
- Problem solving & risk profiling
  - Medical
  - Financial
- Pattern Recognition
  - IRS
  - NSA
- Many more

We want information that is:

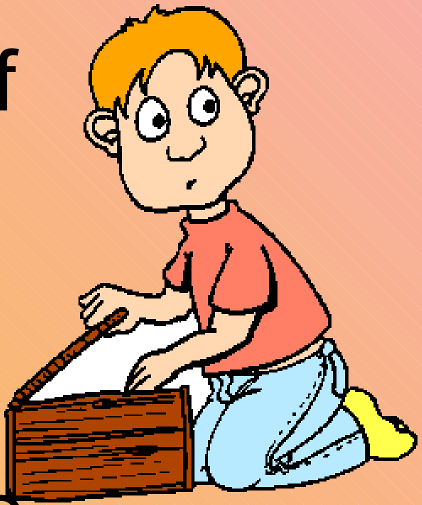
- Original
- Non-trivial
- Fundamental
- Simple
- Useful
- Currently needs to be numerical or categorical (changes under way)

Theoretical Definition of Data Mining

- Go into data
- No preconceived ideas
- See what is found
- A learning process

Analogy

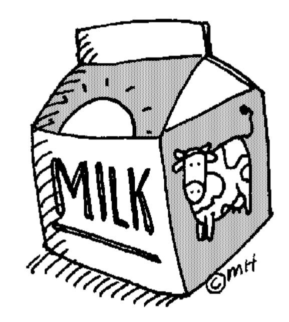
- Somebody gives you a box of stuff
- You have no idea what is in it
- You open it up
- You look through the contents
- You make note of all that you have found
- You may sort the stuff into piles based upon some similarities



Association Rule Example

- Check out at store
  - Each check out is a record
  - Called a “basket”

- Look at what was purchased
- Look for relationships or associated items
  - milk and cereal, beer and pizza



Grocery Example

Invoice	Milk	Cereal	Pizza	Beer	Eggs
1	X	X	X	X	
2			X	X	X
3	X	X			X
4	X	X			X
5			X	X	
6	X				X

Milk AND Cereal

- Support
  - Percent of baskets where rule is true
  - P(milk AND cereal) = 3/6 = 0.500
- Confidence
  - Percent of baskets that have cereal given milk is in the basket
  - P(cereal|milk) = ¾ = 0.75

Benefits to Retailer

- Store layout & increased sales
  - Puts items that sell together close to each other
  - Increase impulse purchases
- Help forecast inventory levels
- Learn customer preferences
- Increase profitability

Common Meaning of Data Mining

- Go into data
- Look for patterns
- Look for answers
- Look for similarities
- Look for rules
- Look for relationships
- Class of predictive analytics

More Analogy

- Someone gives you a box full of stuff
- You are looking for hit singles by Mick Jagger (very specific)
- Along the way, you might discover that they are round, plastic, about 7” diameter, with a 1” hole in the middle
- You might discover that they are on London label
- You might discard all other things in the box

Example

- High risk pregnancy
  - Tons of data
  - What are common factors?
- Myocardial Infarction (MI)
  - Tons of data
  - What factors common in non-occurrence?
- Loan application
- Many other situations

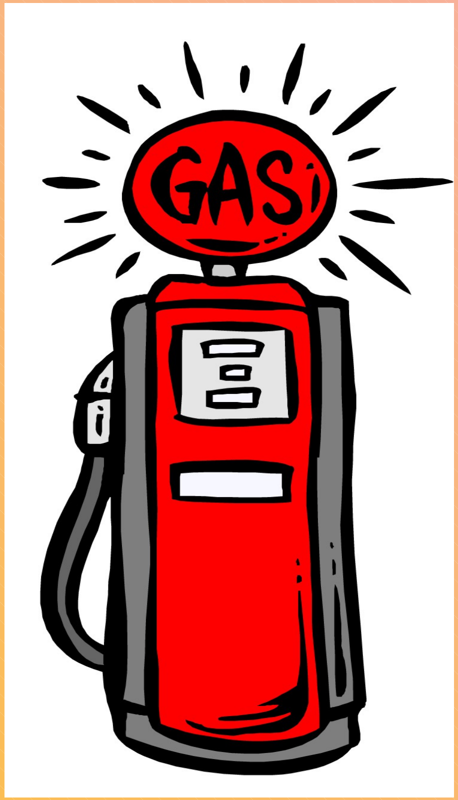
Simple Use of Common Factors

- Nurse/banker/etc can ask a few questions
  - The key factors
  - Identified by data mining
- Now has indication of risk or outcome

I just have 5 questions
Then I can tell you if you
get the loan

Interesting Finding

- Data mining stolen credit cards
  - Common pattern
  - Pay at pump authorization
    - To determine if card is good
  - Fast indication of stolen card



What is in the Mountain of data?

- Better yet, ask
- What is NOT in the mountain of data?
- Think about those shopper cards
  - Has Name
  - Phone, address, email, etc
  - What you buy
  - Brand
  - Size

Original

- US Navy
- Boilers on boats exploding
- This is predictive analytics
- Was very productive/helpful

Think about a trip to the store

- They scan your purchases
- You give them credit card & shopper card
    - They have your
        - name
        - Address
        - Telephone number
        - Maybe drivers license
        - Maybe SSN
        - List of items that you bought, including size

Grocery Store

- Single male
  - Pizza
  - Beer
  - Potato chips
- Mother with kids
  - Milk
  - Bread
  - Eggs
  - Cereal
  - If diapers, then infants

Ratio Rules

- An extension on data mining
- Gives ratios along with support and confidence
- Examples
  - Buy 1 lb peanut butter for each loaf of bread purchased
  - Buy 2 pizzas for each case of beer
- Requires singular value decomposition

Think About . .

- All of the “baskets” about you out there
  - Medical Information Bureau (MIB)
  - Charge Cards
  - Bank information
  - Public records
  - Store purchases
  - Travel records
  - Much more
- The combination of all this creates a pattern that describes you exactly (scary!)

Applications

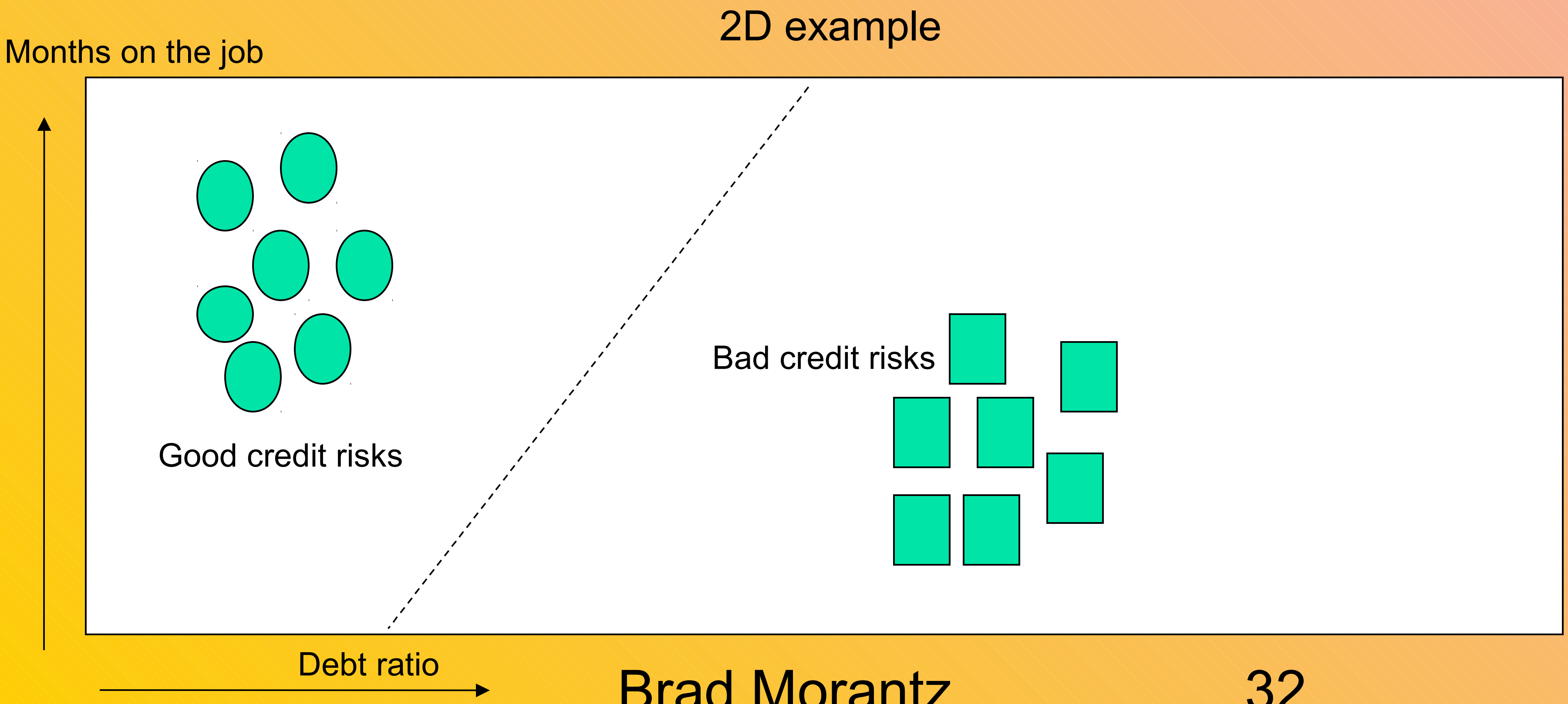
- Marketing
  - Target marketing
  - Store layout
  - Better understanding
- Medical
- Criminal
- Financial
- Science

Methods

- Rule Induction
- Similarity Engine
- OLS & logistic regression
- Data visualization
- Bayesian Classification & networks
- Clustering
- Specialized programs
  - Clementine
  - Enterprise Miner

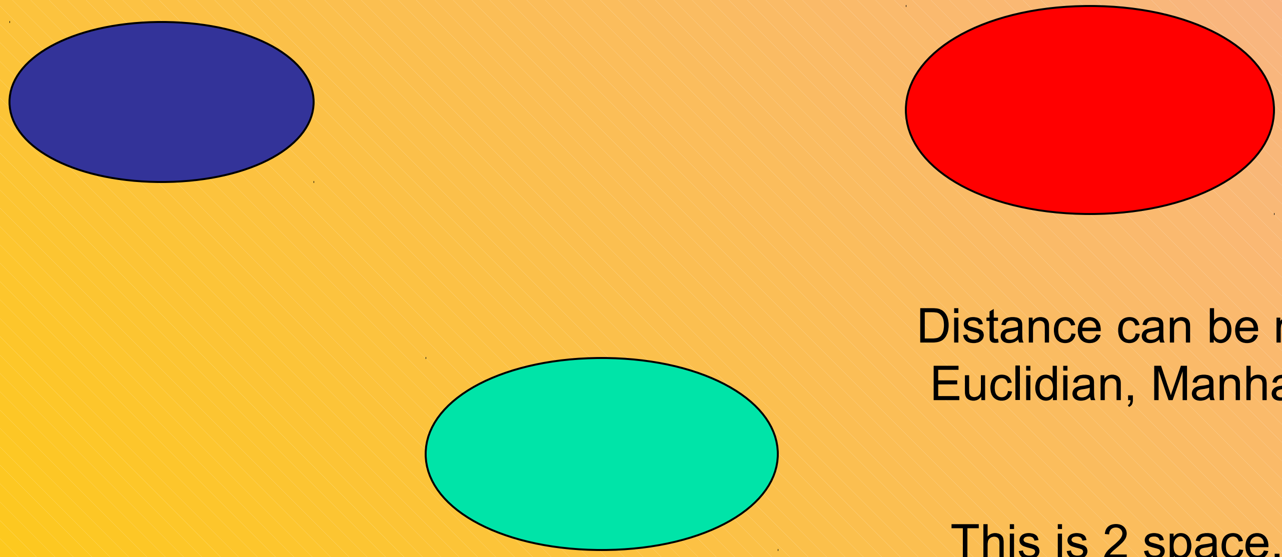
Plot

- Plot data and put in breaks to split it



Clustering

This can be in N-space or hyperspace
Each dimension is a variable



Distance can be mahalanobus,
Euclidian, Manhattan, or other metric

This is 2 space, we can picture
3 space, but in math we can have
N-Space

Each Cluster

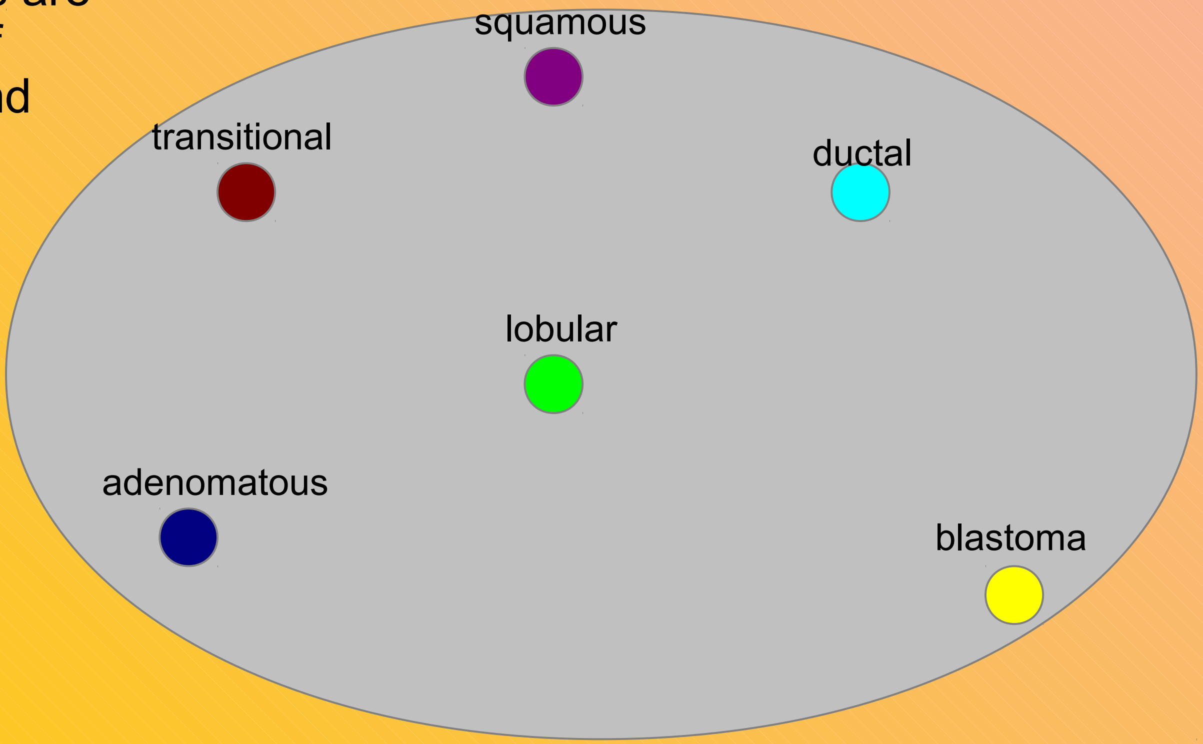
Recursive Clustering

- ✓ This is my work
- ✓ Take one of the clusters obtained and then do cluster analysis on it
- ✓ Ratio of radius:separation distance
- ✓ My cancer analysis work
  - ✓ First cluster between malignant and benign
  - ✓ Then cluster the malignant to separate out the various types

Recursive Clustering

The small clusters are individual types of malignant cells and occupy a small portion of the space

Cluster of malignant cells



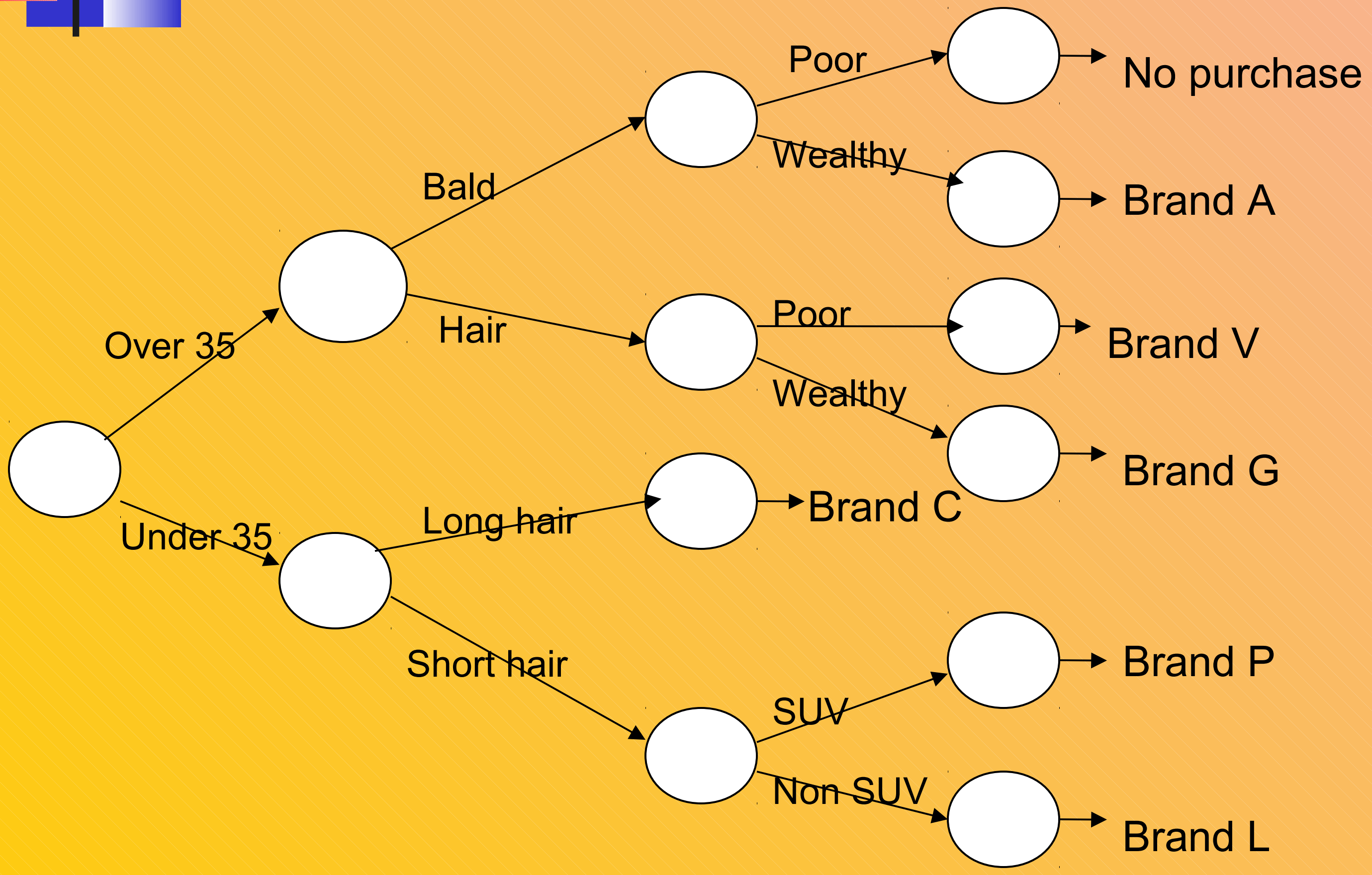
Accuracy

Accuracy is a function of the volume of the cluster divided by the volume of the variable space

In the picture on previous page, volume of each cell type was miniscule compared to total volume

Net result is the ability to accurately classify

Trees


Over 35
Under 35
Bald
Hair
Long hair
Short hair
Poor
Wealthy
Poor
Wealthy
SUV
Non SUV
No purchase
Brand A
Brand V
Brand G
Brand C
Brand P
Brand L

Recursive Partitioning Algorithms

- Recursively make choice at node purer
  - Adding another node = another variable
  - Add until desired accuracy is met
- CART
  - Classification And Regression Trees
  - Reducing variance at node makes cluster tighter
- ID3/C4.5
  - Bayes Classification
  - Reduce entropy/disorganization

Statistical Processes

- Plot it and look at plot
- Step wise regression
- Significant variables
  - T value
  - Eigenvalue
- Pearson correlation coefficient
- Cluster Analysis

Bayes Classification

- Conditional Probability
  - $P(A|B) = P(A \cap B) / P(B)$
- Independence
  - If $P(A|B) = P(A)$, then A independent B
- If not independent, then B contributes
- Compare $P(A|B)$ to $P(A|C)$ to $P(A|Z)$
  - – Can rank variables to see how much knowledge each adds

Required steps

- Build & maintain a database
- Data formatting
- Data cleansing & anomaly detection
- Data visualization
- Currently need human expert to evaluate

Data Preparation

- 75% of effort and time
- Can mean difference between success and failure
- Can greatly affect accuracy
- My results with abalones

Data Preparation Example

- Mathematical relationships
  - Weight of animal in shell < animal only
  - Ratios
  - More than possible quantity in container
- Impossible answers
- Out of range answers
- Data entry errors
- Etc

Why now?

- Rapidly decreasing cost of data storage
- Increased ease of collecting data
  - Networks
  - Internet
  - “Smart” phones
  - Store scanners
  - Computerized lifestyle
  - Heavy use of credit/debit cards
  - Store customer “value” cards
- More & cheaper computational power

Future

- Incorporation of background & associated knowledge
- More exacting data sources
- Hybrid systems
- Mixed media use
- Data Fusion

References

- IEEE Computational Intelligence Society
- My website: www.machine-cognition.com
- Tom M Mitchell & Bruce Buchanan
- Toshinori Munakata
- Quinlan
- Clark & Niblett
- Leo Breiman

Contact Info

Brad Morantz PhD
www.machine-cognition.com
brad@machine-cognition.com
480-348-5945

Questions

Any Questions?

Thank you

No need to applaud

I prefer chocolate chip cookies