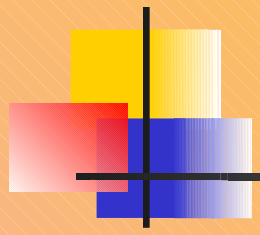
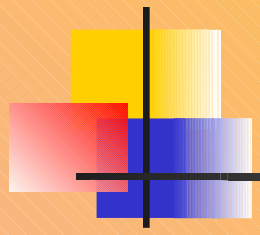


Data Mining

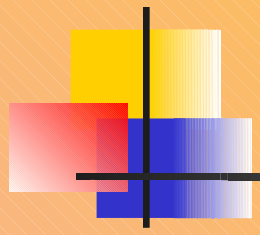
Brad Morantz PhD

bradscientist@machine-cognition.com



What is a Mountain?

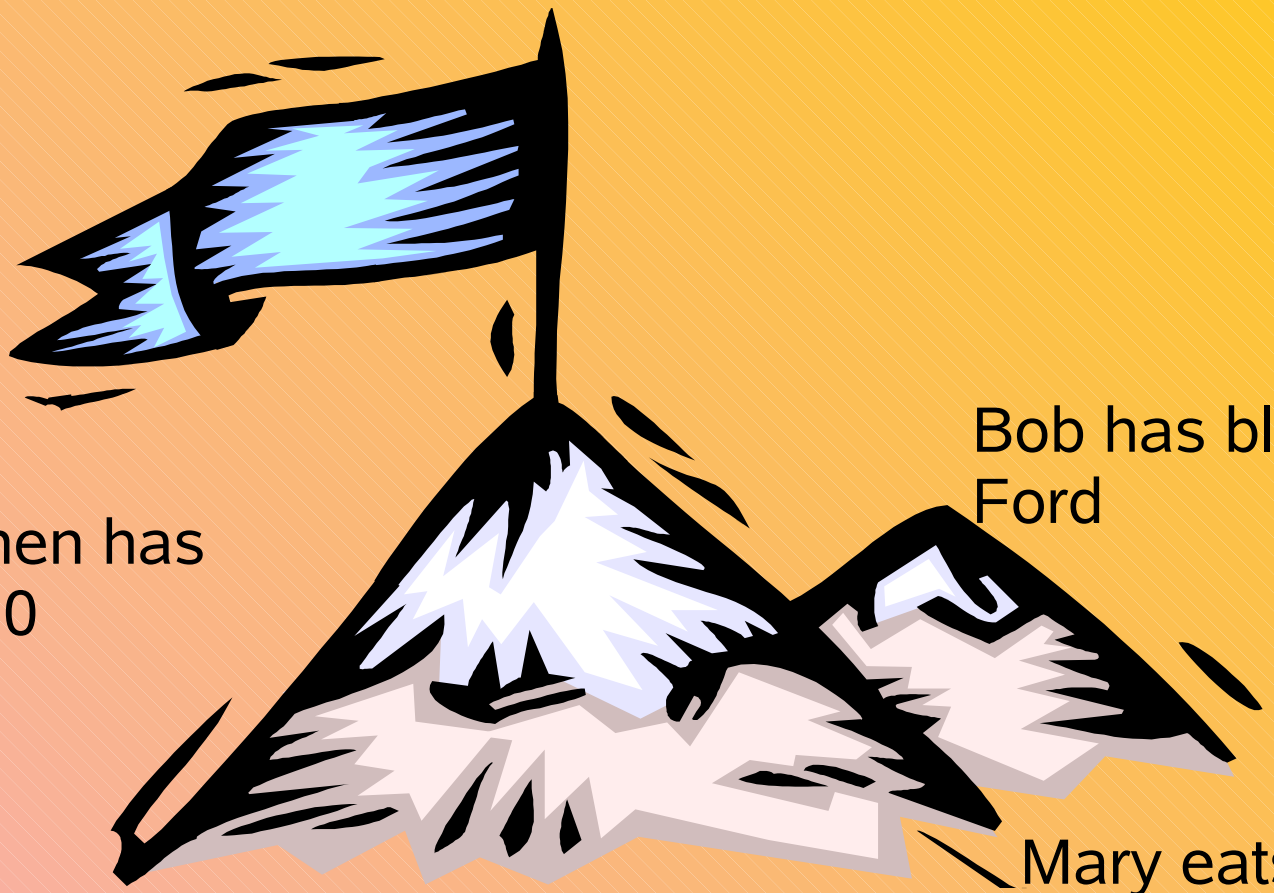
- A big pile of dirt!



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



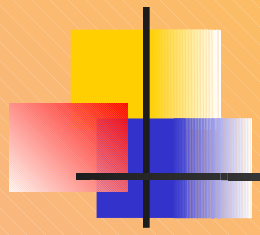
Chen has $10

Bob has blue Ford

Mary eats rice krisp

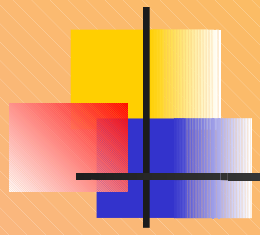
Notice in Last Example . .

- 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



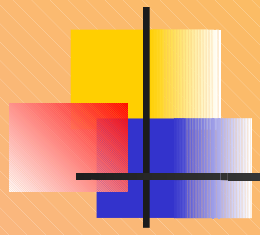
Data Mining

- 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)



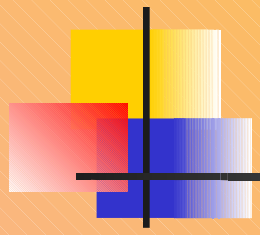
Data Mining

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



Uses

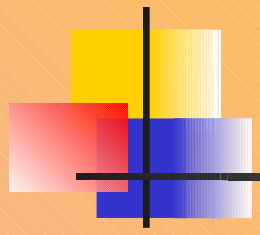
- Merchandising
  - Association rules
  - Customer profiling
- Problem solving & risk profiling
  - Medical
  - Financial
- Pattern Recognition
  - IRS
  - NSA
- Many more



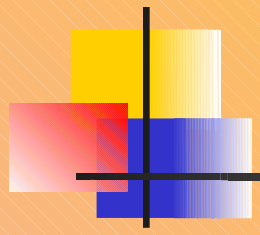
We want information that is:

- Original
- Nontrivial
- Fundamental
- Simple
- Useful
- Currently needs to be numerical or categorical

Theoretical Definition of Data Mining



- Go into data
- No preconceived ideas
- See what is found



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

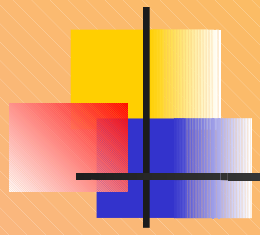
- Store layout & increased sales
  - Puts items that sell together close to each other
  - Increase impulse purchases
- Help forecast inventory levels
- Learn customer preferences
- Increase profitability

Common Meaning of Data Mining

- Go into data
- Look for patterns
- Look for answers
- Look for similarities
- Look for rules
- Look for relationships

More Analogy

- Someone gives you the 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

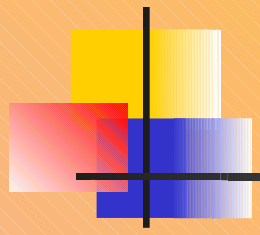


Example

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

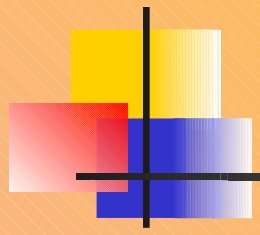
Simple Use of Common Factors

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



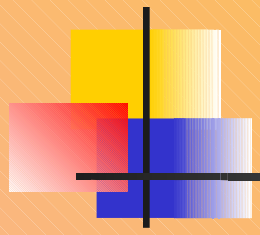
Interesting Finding

- 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?



Original

- US Navy
- Boilers on boats exploding

Think about a trip to the store

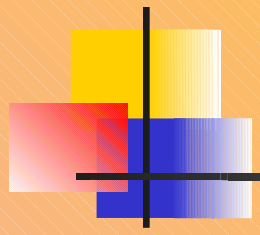
- They scan your purchases
- You give them credit card or check
  - They have your
    - name
    - Address
    - Telephone number
    - Maybe drivers license
    - Maybe SSN
    - List of items that you bought, including size

Grocery Store

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

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



Applications

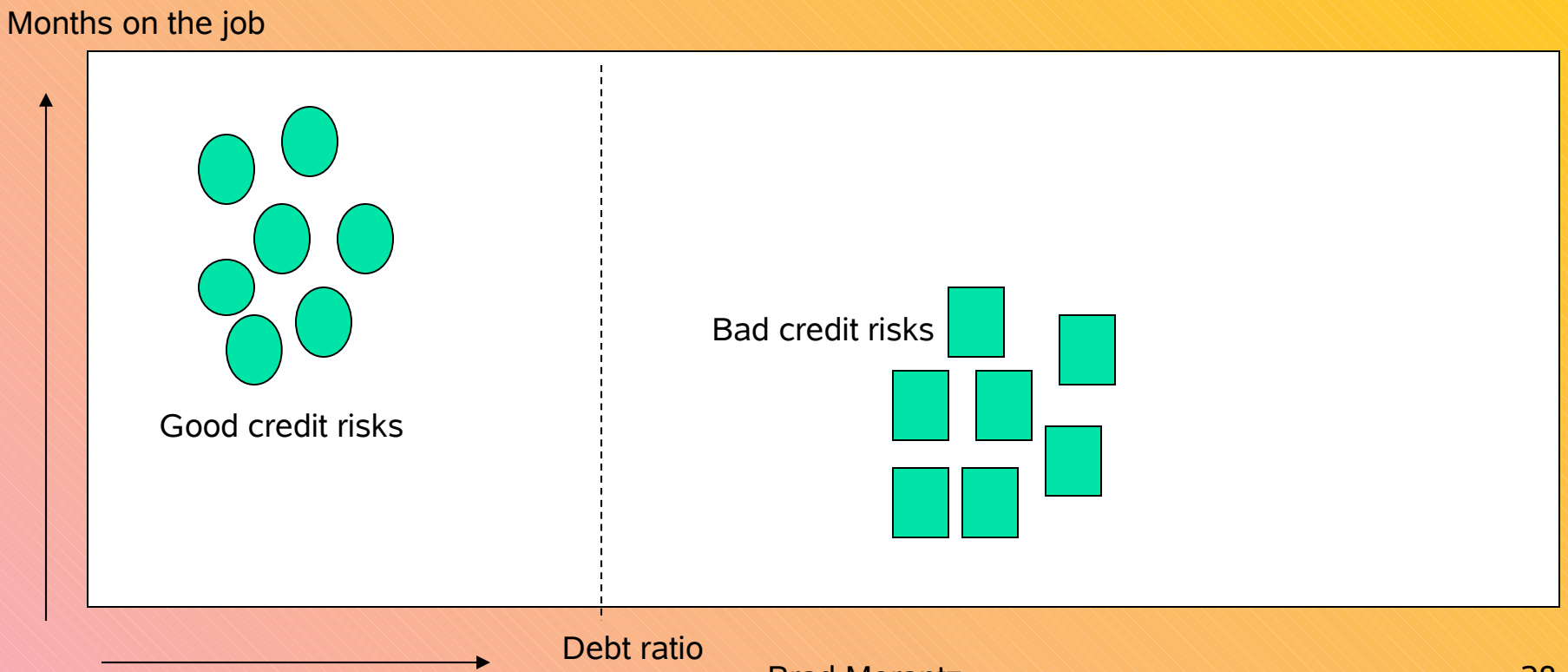
- Marketing
  - Target marketing
  - Store layout
  - Better understanding
- Medical
- Criminal
- Financial
- Science

Methods

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

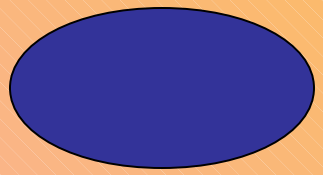
Plot

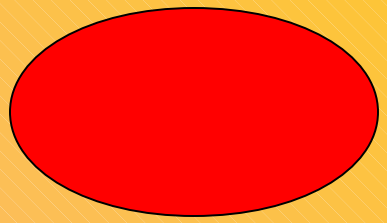
- Plot data and put in breaks to split it



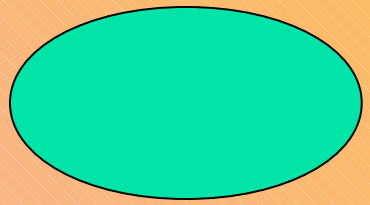
Clustering

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



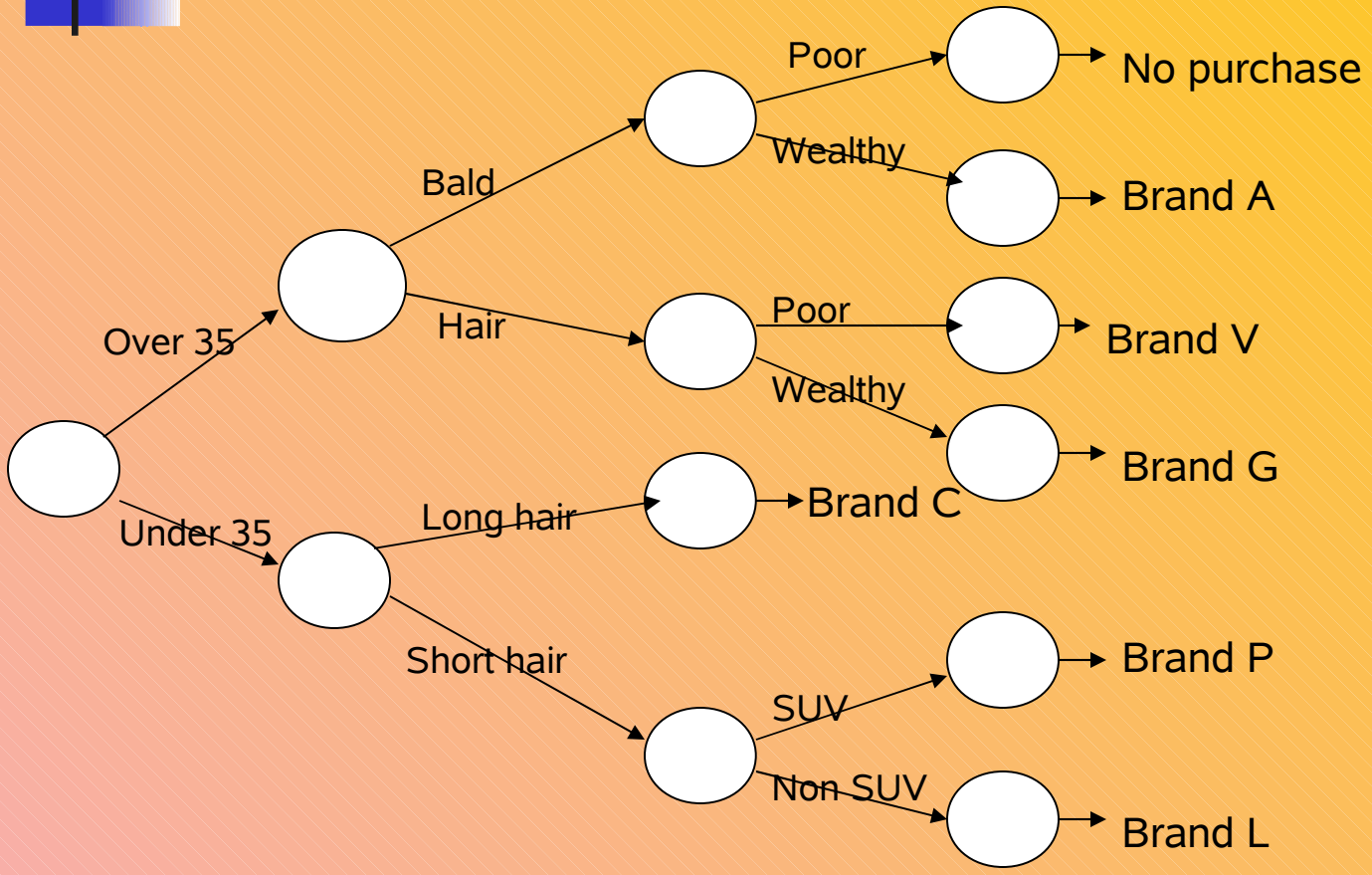


Distance can be mahalanobus,
Euclidian, or other metric

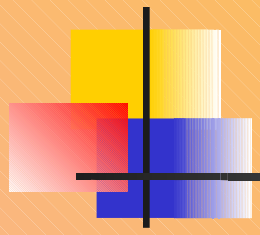


Each Cluster

Trees


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

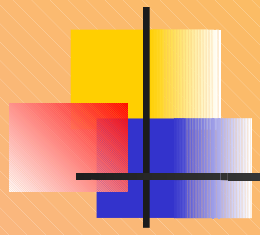
Recursive Partitioning Algorithms



- 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

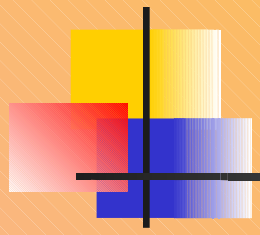
Statistical Processes

- 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 …….
  - Can rank variables to see how much knowledge each adds

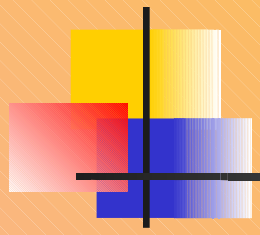


Required steps

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

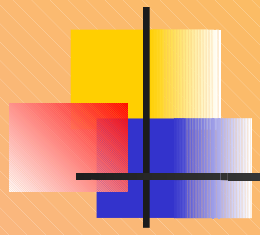
Data Preparation

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



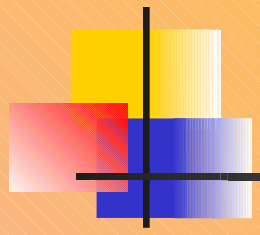
Data Preparation Example

- 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



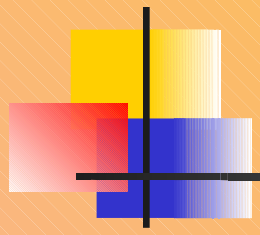
Why now?

- Rapidly decreasing cost of data storage
- Increased ease of collecting data
  - Networks
  - Internet
  - Intranet
  - 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



References

- Leo Breiman
- Tom M Mitchell & Bruce Buchanan
- Toshinori Munakata
- Quinlan
- Clark & Niblett
- www.kdnuggets.com
- www.machine-cognition.com