Memory-Based Reasoning

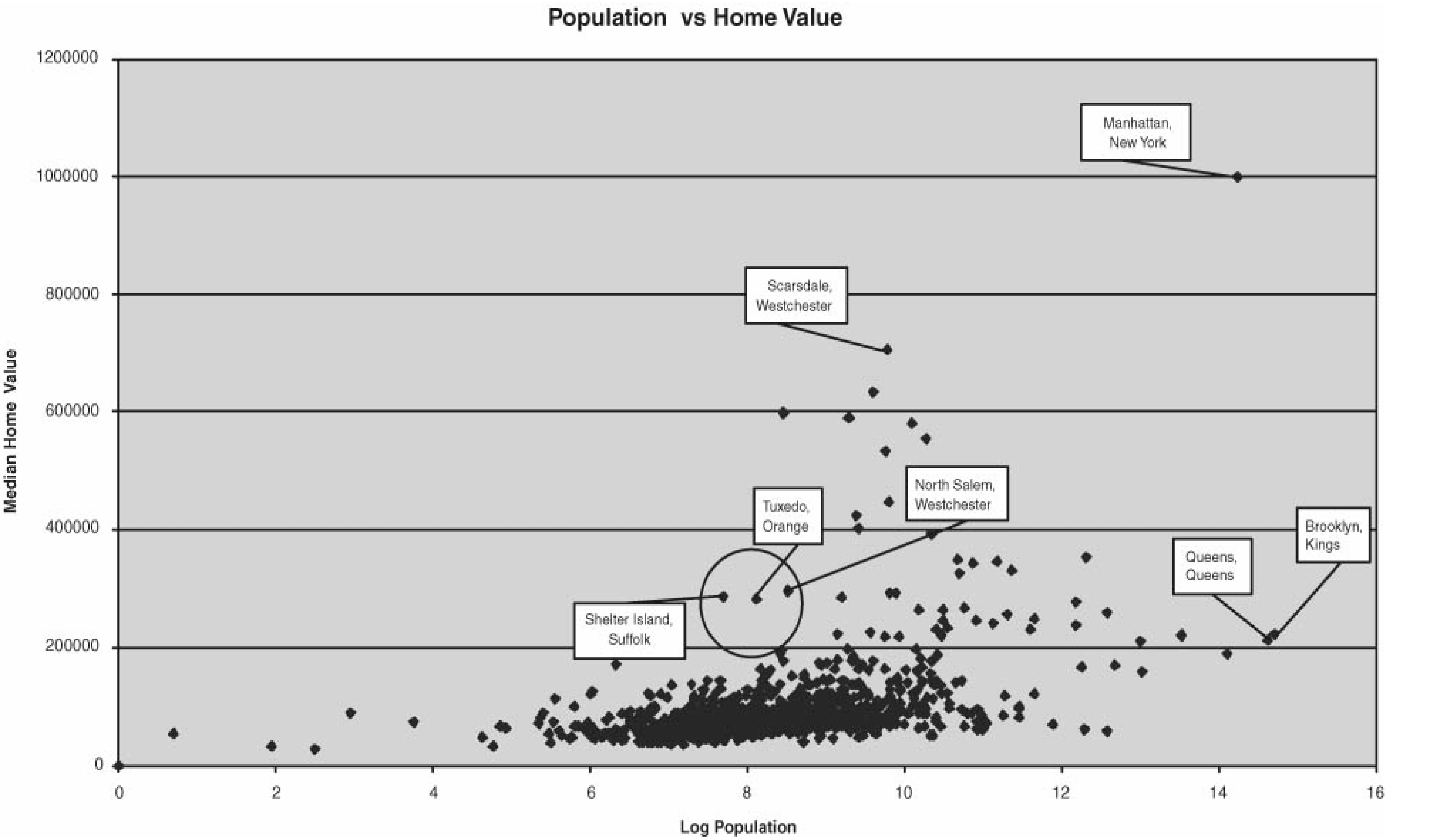
Data Mining Techniques, by M.J.A. Berry and G.S Linoff, 2004

Memory-based reasoning

- Nearest neighbors methods
  - K-nearest neighbors
  - Predict unknown values for a case based on similarity with K most similar cases

    Case based reasoning
    Reasoning by analogy
- Collaborative Filtering
  Use preferences in addition to similarity with past cases

Rents in Tuxedo, NY ?
Similarity based on Population and Home-value



Rents in Tuxedo

- Finding neighbors
  - “distance” metric
  - Consider K nearest neighbors

- Using data from neighbors to estimate rent in Tuxedo
  - Combining data from neighbors (Combination function)
  - Average of median rent values in neighbors?
  - Other functions?
    - weighting by distance (closer neighbors get larger weight)

Obtaining MBR scores

- Training data is the model

- Find k nearest neighbors (distance measure)
  - use with different data types

- Use data on neighbors to determine value (score) for new case (combining function)

Distance

- Non-negative d (A, B) >= 0
- Identity d (A,A) = 0
- Commutative d (A, B) = d (B, A)
- Triangle inequality d (A, C) = d (A, B) + d (B, C)

- Some similarity measures are not true distance measures (Eg. “distance” between pages of text)

- Absolute difference |A – B|
- Sum of sq. of differences $(A – B)^2$
- Normalized absolute difference |A – B| / (max. difference)
- Absolute value of difference of standardized values

  | (A – mean)/stdev – (B-mean)/stdev |

Distance

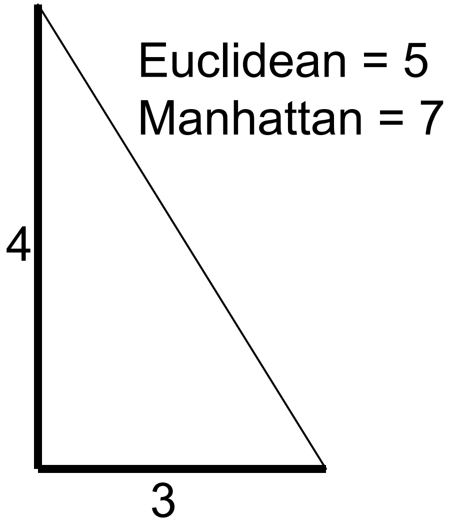
- Categorical fields
  - d (male, male) = 0, d (male, female) = 1
  - Hierarchical categories
    difference in hierarchical levels of values
  - Replace categories with numeric values (!)

- Combine individual field distances

  Euclidean distance: sq. root of sum of sq. of distances on different fields

  Manhattan distance: sum of distances on different fields

  Weighted sum



Combination function

- Averaging
- Majority voting
- Use Weights
  - Weights inversely proportional to distance
  - Weighted averaging, weighted voting

Example: Classifying news stories

- Assigning classification codes

  Types: Government, Industry, Market sector, Product, Region, Subject
  I/INS – insurance industry, S/IPO - related to IPOs

  Multiple codes for articles

- Assign codes to new articles based on codes for most similar articles

  Distance between two documents (word-sets)

  - How many neighbors?

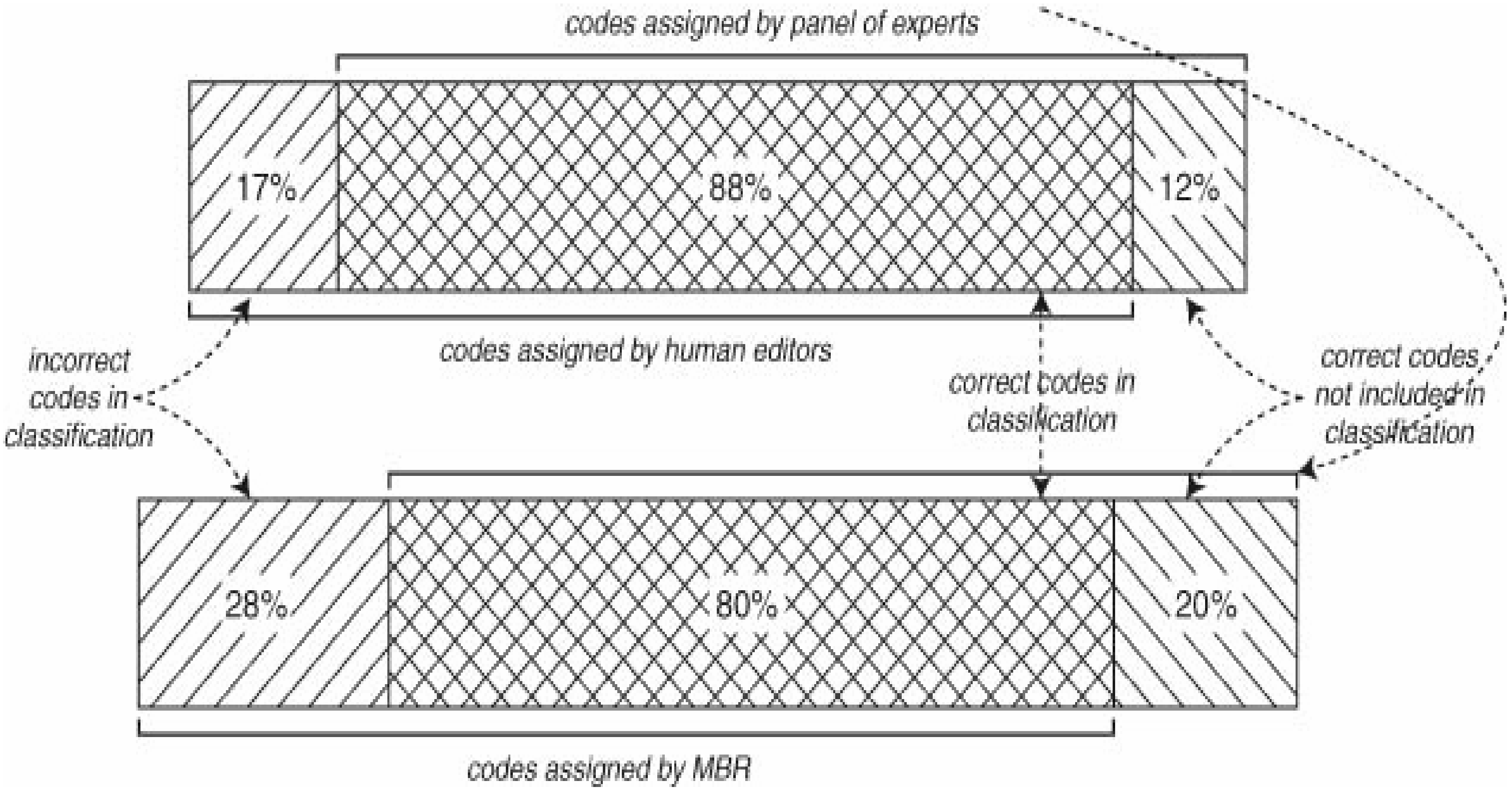
- Combination function
  - Majority voting
    - Assign codes that are in a majority of neighbors
  - Weighted majority

- Distance -- text data
  - Remove non-content words (“it”, “and”, “for”, etc)
  - Remove most common words – they convey little information to distinguish between articles
  - Collect remaining words into a dictionary of searchable terms
  - Distance function: score based on number of common words in two documents

    d (A, B) = 1 – score (A, B) / score (A, A)

    Note: d (A, B) is not same as d (B, A)

- Training data – around 50K news stories, over 300 codes
- Test data - 200 cases to assign codes to
  - compare with human assigned codes


codes assigned by panel of experts
17%
88%
12%
codes assigned by human editors
incorrect codes in classification
correct codes in classification
correct codes not included in classification
28%
80%
20%
codes assigned by MBR

Training data is critical

- Model is the training data

- Similar number of cases from each category
- Many examples from each category

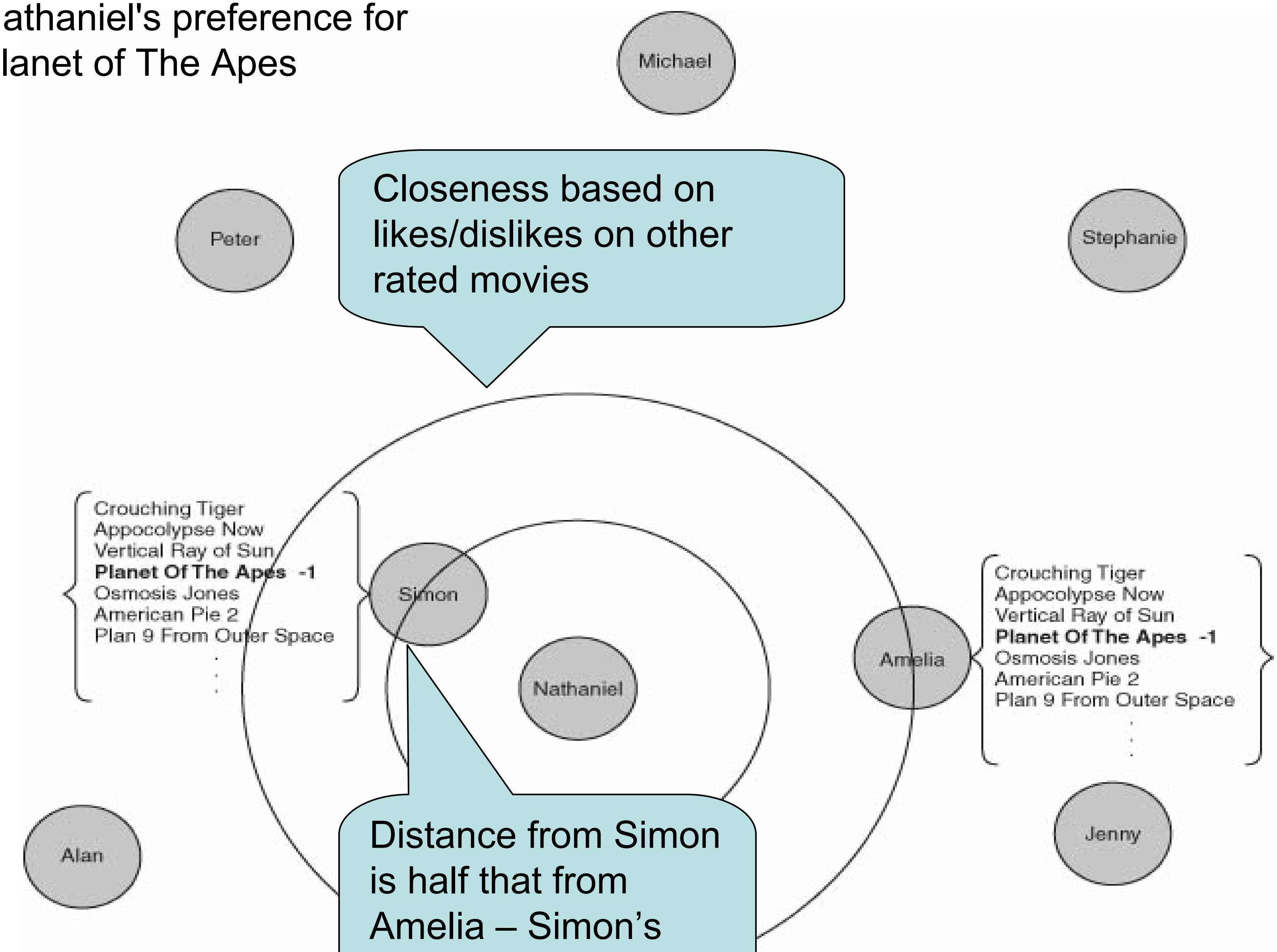
- Comparable distance measure for different fields
  - Numeric data should be scaled to same range (standardized data)
  - Substitute numeric data for categories where possible (response frequency, sales in zip rather than zip codes)

Collaborative filtering

- Variant of MBR suited for making personalized recommendations

- Stored preferences
- Similarity based on overlap in preferences

Nathaniel's preference for Planet of The Apes



Closeness based on likes/dislikes on other rated movies

Distance from Simon is half that from Amelia – Simon’s ratings are weighted twice that of Amelia