Data Analysis for the New Millennium: Data Mining, Genetic Algorithms, and Visualization

by

Bruce L. Golden
RH Smith School of Business
University of Maryland

IMC Knowledge Management Seminar
April 15, 1999

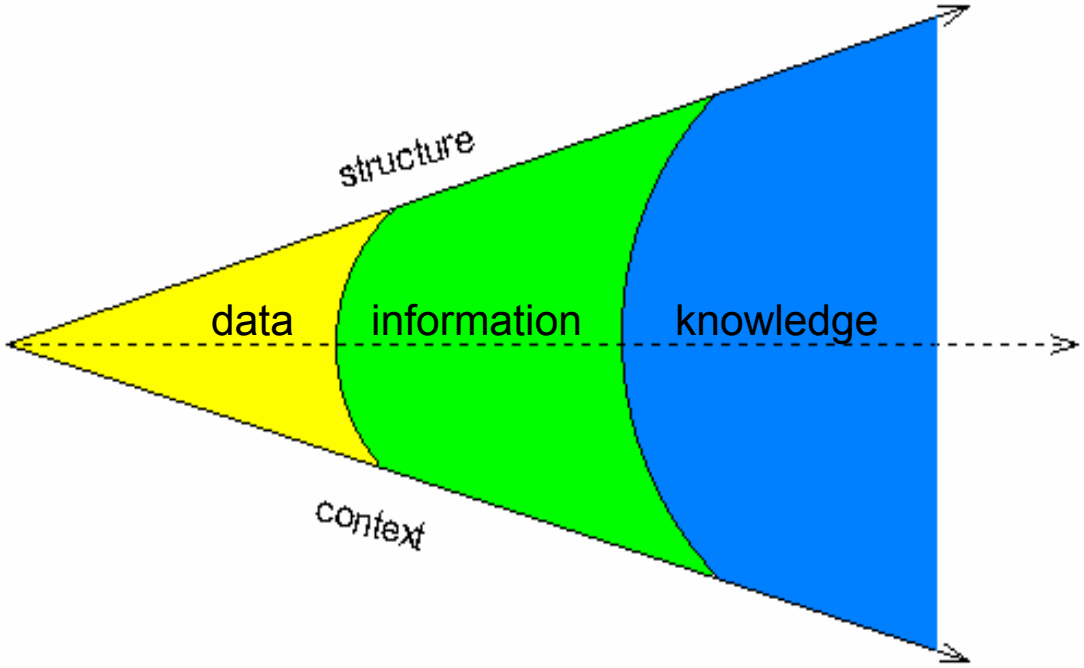
Focus

- Connection between data and knowledge
- Examples of data analysis from late 1980s
- Contrast with data analysis in late 1990s
- Introduce techniques
  - MDS and Sammon maps
  - neural networks and SOMs
  - decision trees
  - genetic algorithms
- Illustrate the power of visualization
- Data analysis as a strategic asset
- Conclusion

Setting the Stage

- "Data is information devoid of context, information is data in context and knowledge is information with causal links. The more structure is added to a pool of information, the more we can talk about knowledge."

  (California Management Review, Winter 1999)

- 

- How can we systematically discover knowledge from data?

Modeling Salinity Dynamics in the Chesapeake Bay

- The Goal: Construct multiple regression models that accurately describe the dynamics of salinity in the Maryland portion of the Bay
- The work was done in late 1989/early 1990
- Motivation:
  - Salinity exerts a major influence over the survival and distribution of many fish species in the Chesapeake Bay
  - Maryland was the nation's leader in oyster production several decades ago
  - MDNR's oyster production rebuilding program relied on predicting salinity levels for various areas in the Bay

Model Building

- Data collected at 34 stations by USEPA from 1984 to 1989
- 36,258 water samples collected at different depths (9 variables per sample)
- Constructed 10 models in all (bottom data / total data)
    - Upper, Middle, Lower, Entire Bay, Lower Tributaries
- Four key independent variables
    - Day, Depth, Latitude, Longitude
- Transformation of key variables
- Extensive screening for independent variables
- Used stepwise regression in SPSS/PC

Model Results

- Six independent variables in each model
- Model assumptions are not violated
- R^2 values range from 0.56 to 0.81
- Entire Bay Model

  $R^2 = 0.649$

  depth increases, salinity increases

  Salinity = 199.839 - 1.151 Day1 + 1.161 Day2
  + 0.283 Depth - 4.863 Latitude
  - 1.543 Longitude - 13.402 Longitude1

Salinity Modeling Summary

- The regression models were validated using new (1990) data involving 7,000 observations

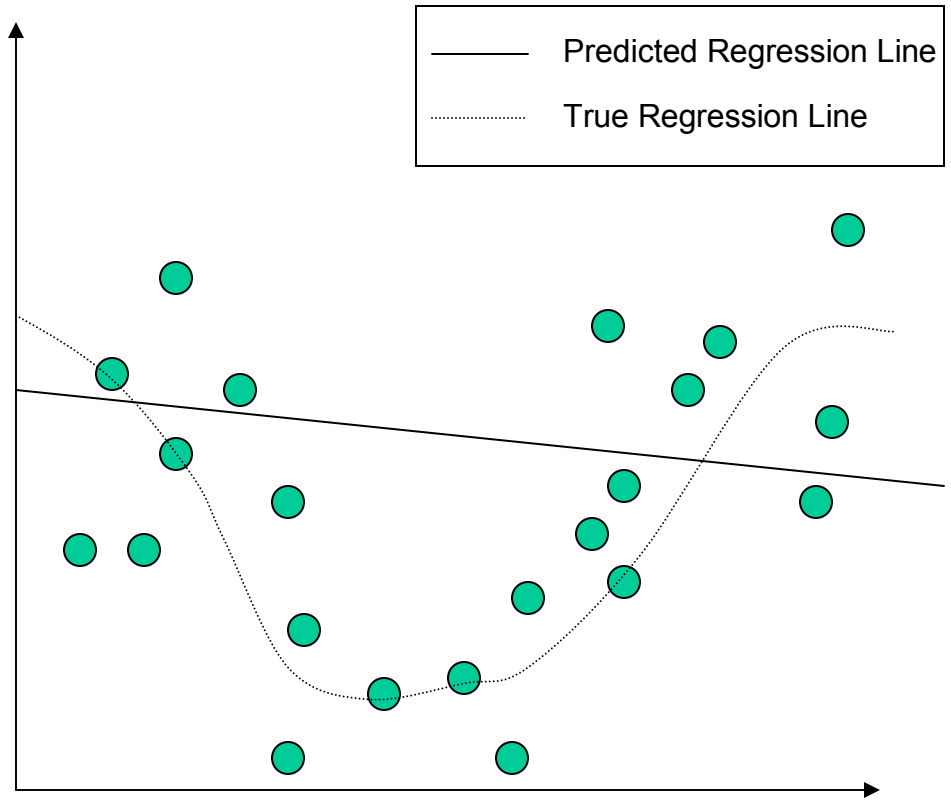
- These regression models can be used to predict salinity for a location on the Bay at a specified depth and date

- In 1991, we applied neural networks to the same problem

- To our surprise, the neural network models predicted salinity levels more accurately than the regression models in 90% of the cases

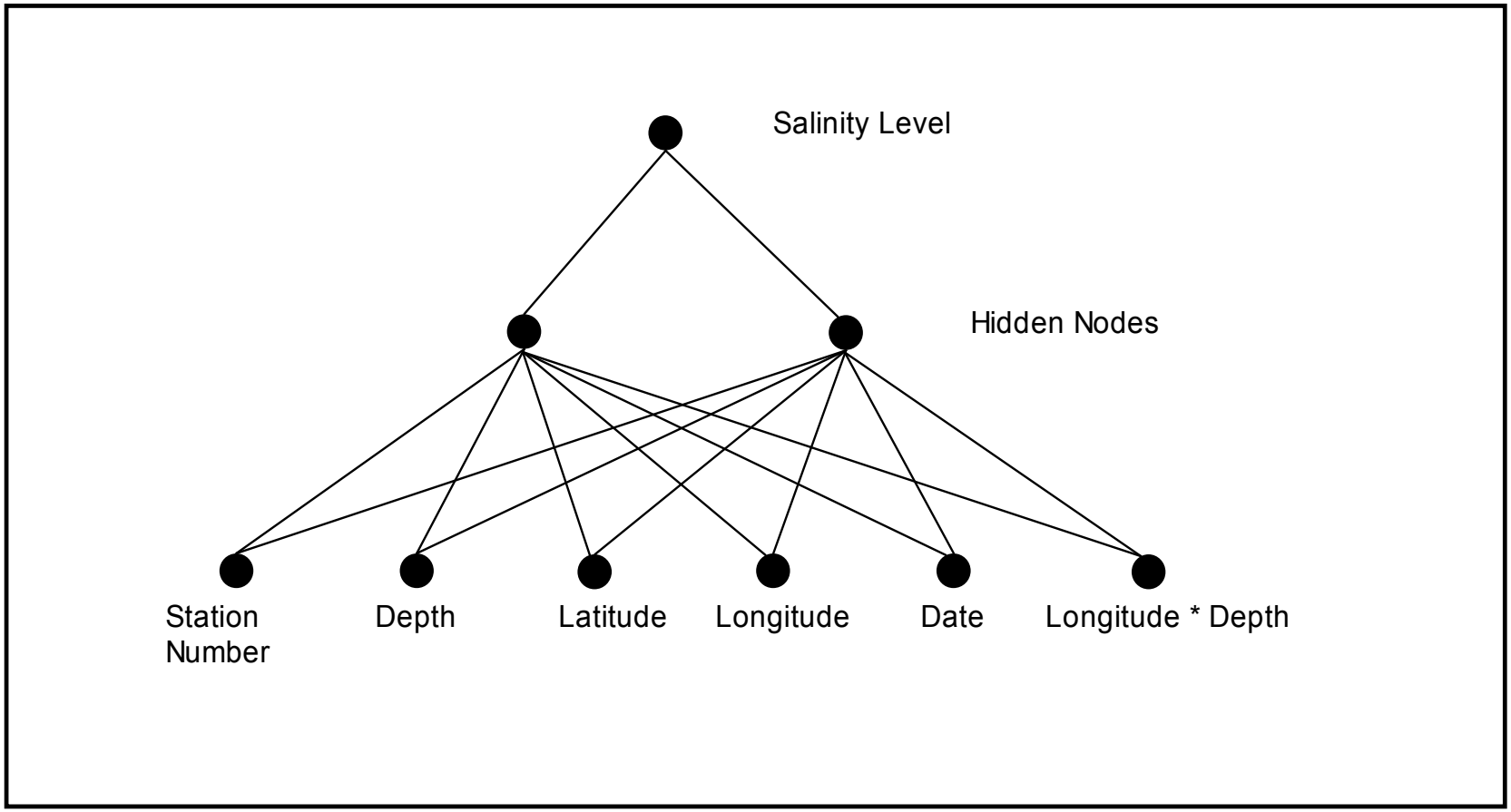
The Problem with Linear Regression

- "But we all know the world is nonlinear." (Harold Hotelling, 1948)



Linear Regression Shortcomings: Nonlinear Data (Cabena et al., 1998)

Neural Network Configuration



Neural Networks

- Neural networks are computer programs designed to recognize patterns and learn “like” the human brain

- They are versatile and have been used to perform *prediction* and *classification*

- The key is to iteratively determine the “best” weights for the links connecting the nodes

- Drawback: It is difficult to explain/interpret the results (same is true for regression)

Visualization

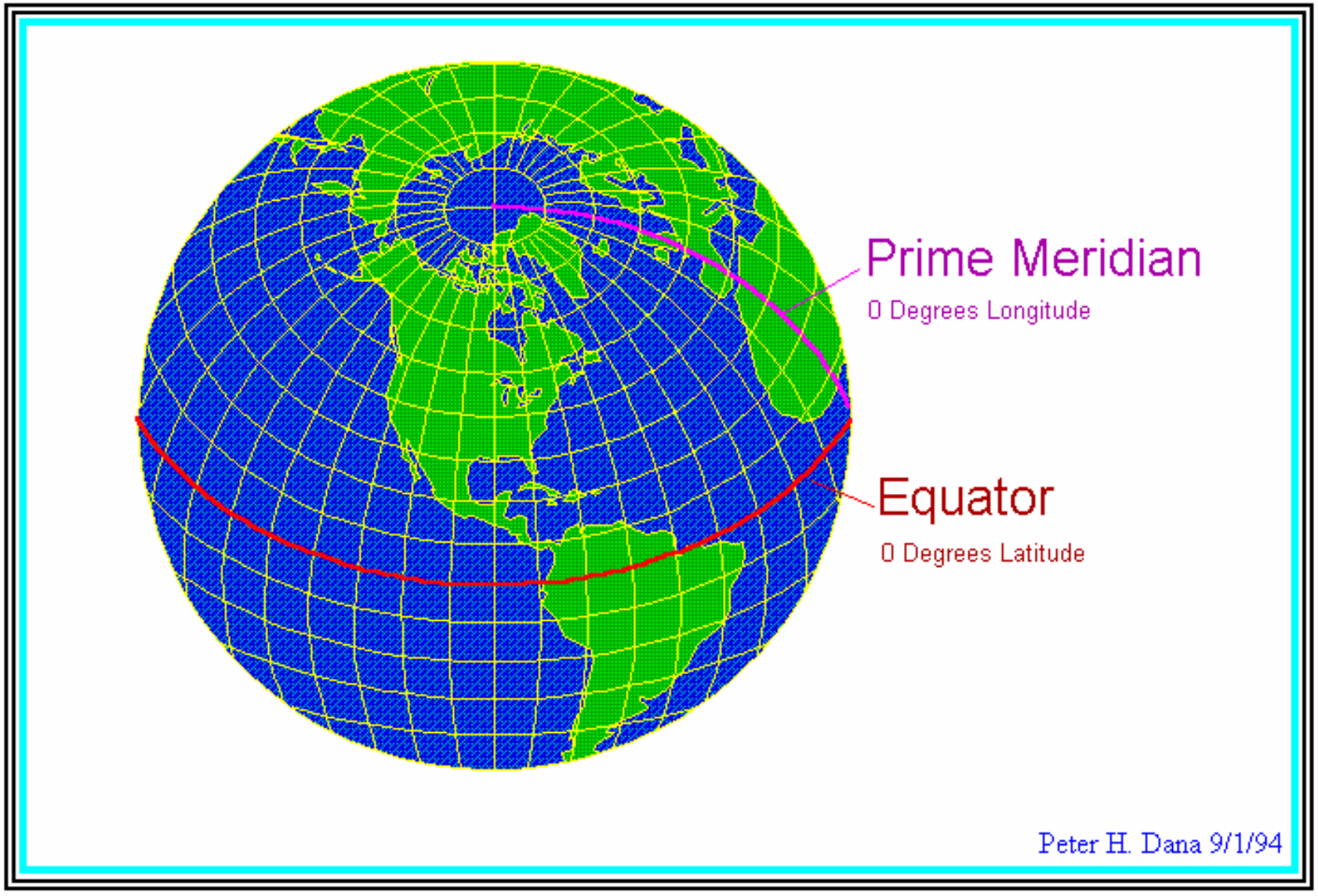
- Psychologists claim that more than 80% of the information we absorb is received visually (Cabena et al., 1997)
- Data is often highly multidimensional
- Mapping from three or more dimensions to two dimensions is not easy

Flattening the Earth

- “Would you tell me, please, which way I ought to go from here?” asked Alice.

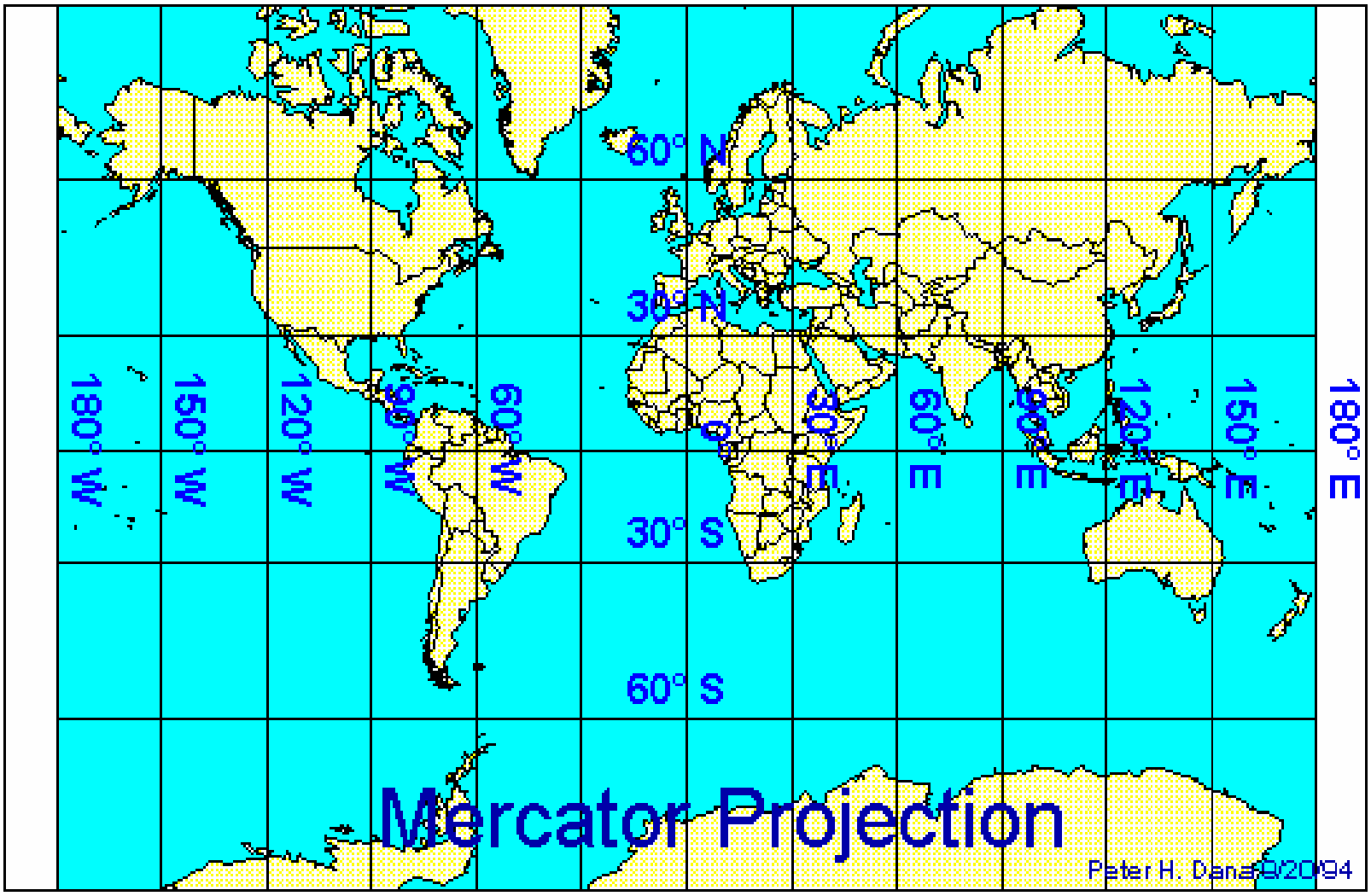
  “That depends a good deal on where you want to get to,” said the Cat.

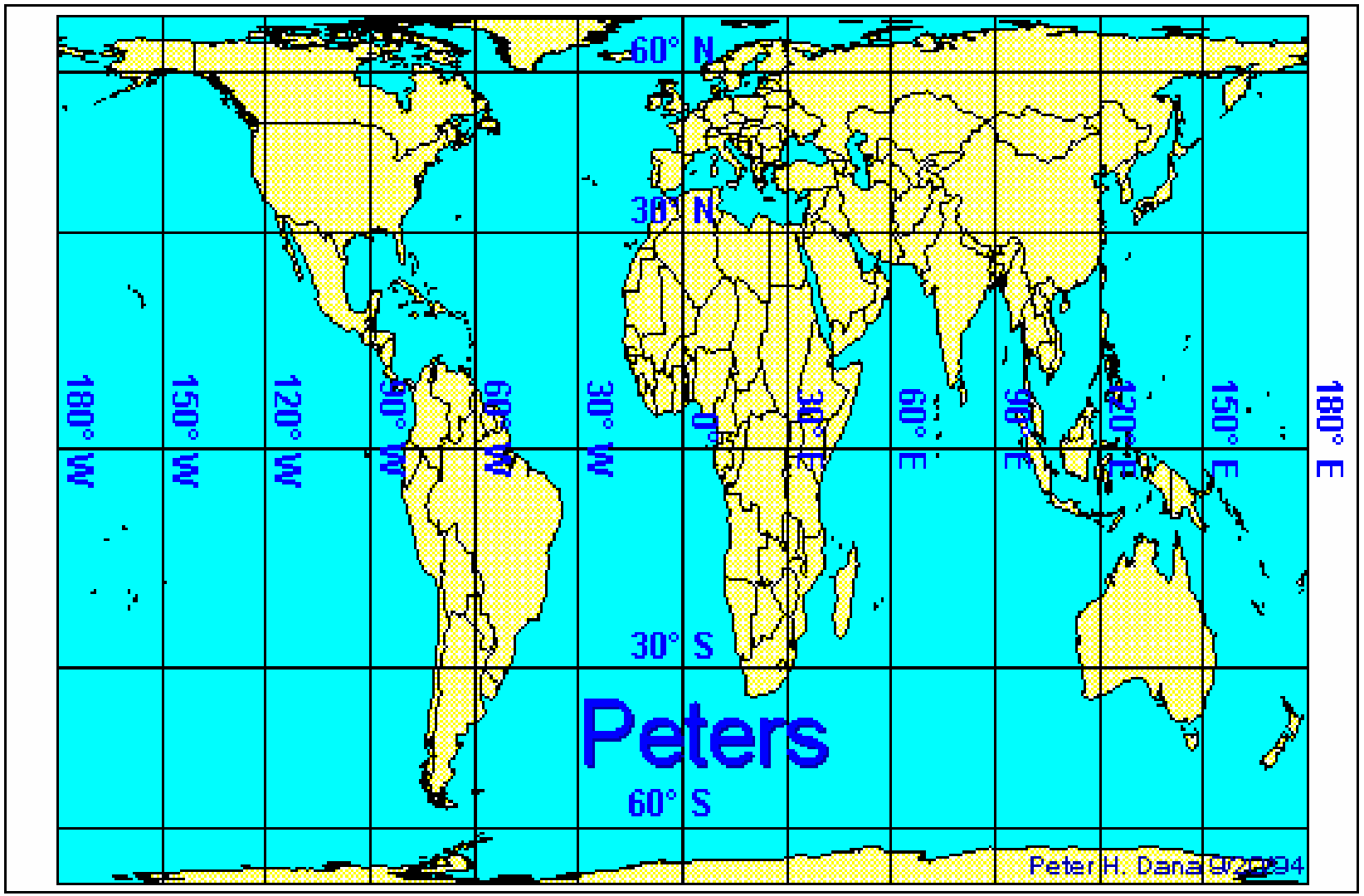
(Lewis Carroll)

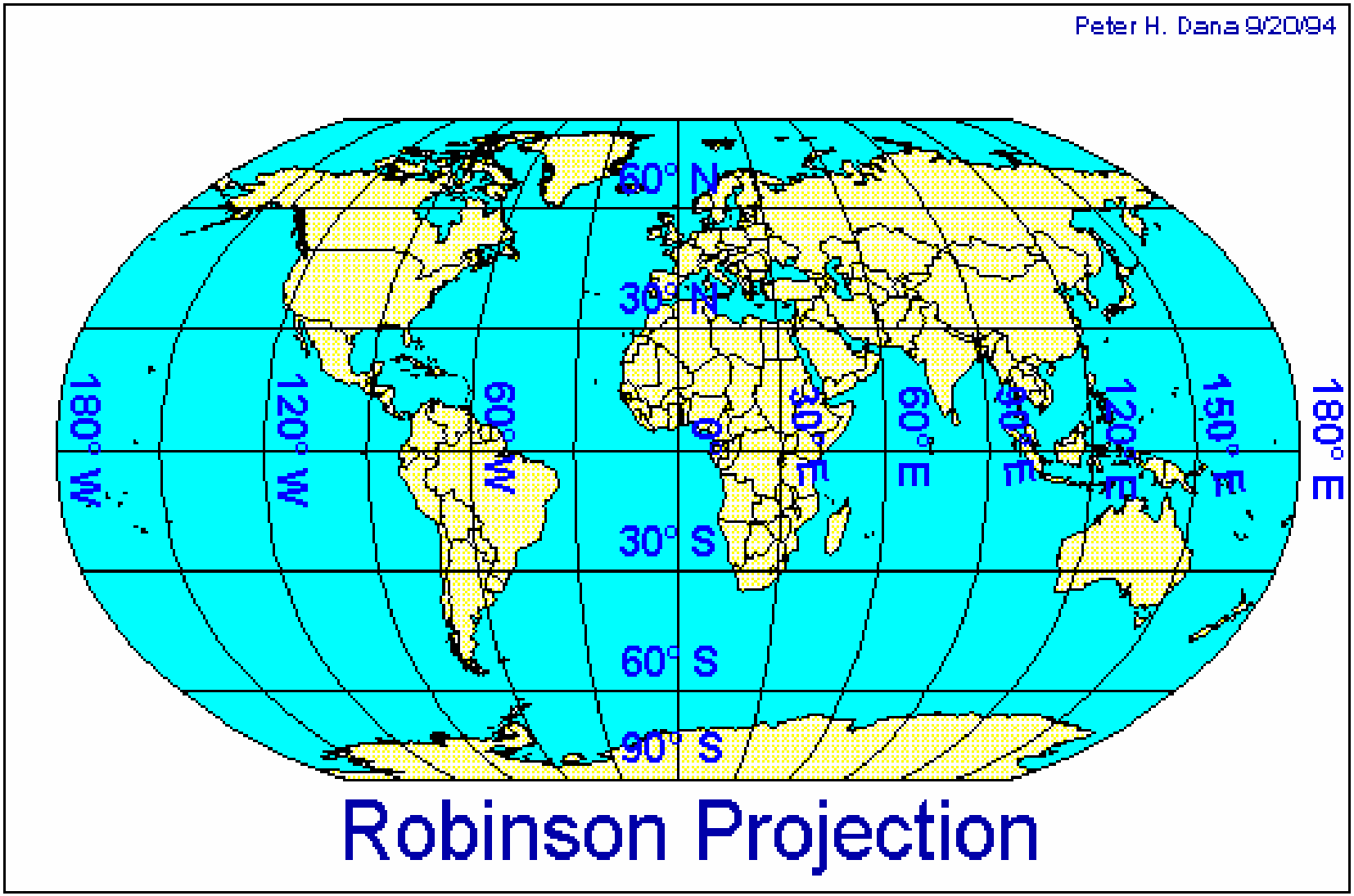

Prime Meridian
0 Degrees Longitude
Equator
0 Degrees Latitude
Peter H. Dana 9/1/94

Map Projections

- We use map projections to represent a spherical Earth on a flat surface
- Two map projections of the world can look quite different
- All map projections distort reality in some ways -- shape, area, distance, angles, etc.
- *Equivalent* projections preserve area
- *Conformal* projections preserve angles
- No projection can be both *conformal* and *equivalent*
- Bottom line: map projections are extremely useful, but offer compromise solutions

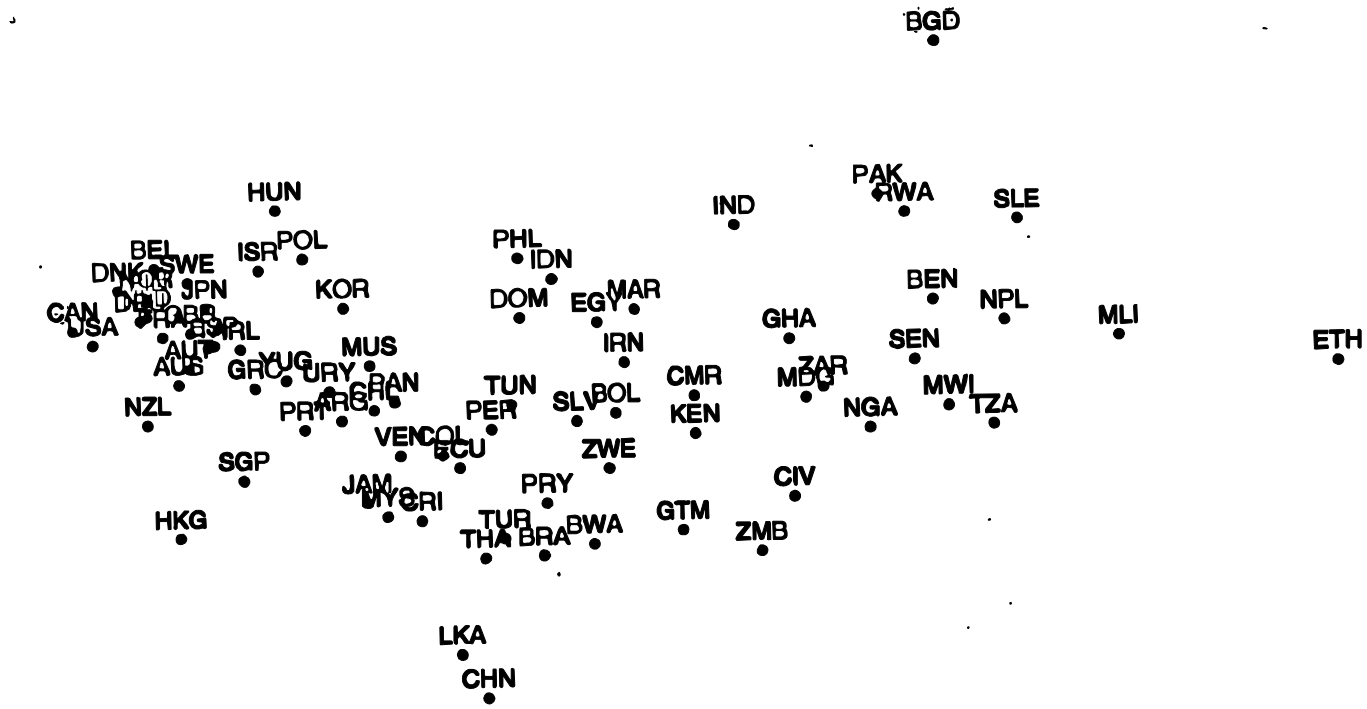

60° N
30° N
30° S
60° S
180° W
150° W
120° W
90° W
60° W
0°
30° E
60° E
90° E
120° E
150° E
180° E
Mercator Projection
Peter H. Dana 9/2004


60° N
30° N
0°
30° S
60° S
180° W
150° W
120° W
90° W
60° W
30° W
30° E
60° E
90° E
120° E
150° E
180° E
Peters
Peter H. Dana 9/2024

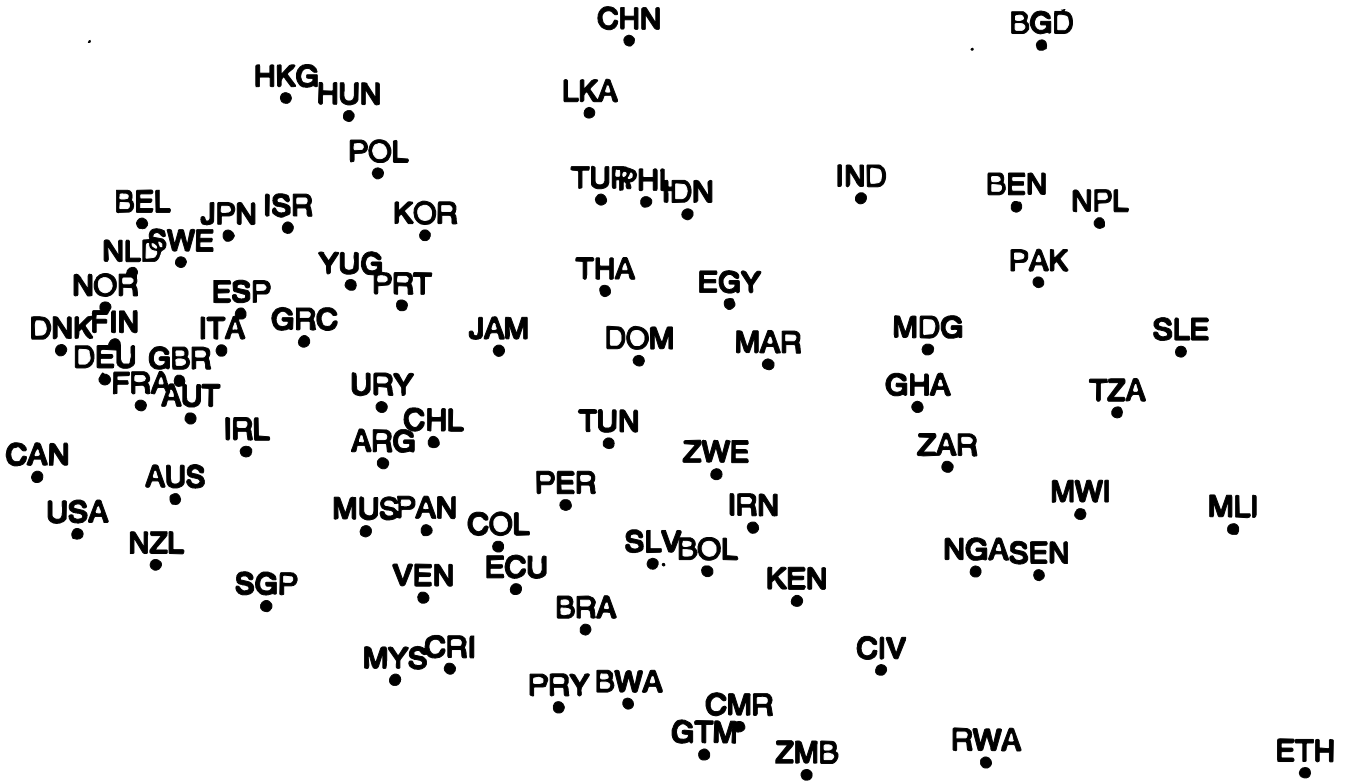

Peter H. Dana 9/20/94
60° N
30° N
0°
30° W
60° W
120° W
180° W
30° E
60° E
90° E
120° E
150° E
180° E
30° S
60° S
90° S
Robinson Projection

Visual Clustering (Segmentation) Methods

- Multi-Dimensional Scaling (MDS)
- Sammon Mapping
- Self-Organizing Maps
- Euclidean distance (more or less) is used as a similarity measure



MDS Applied to a World Poverty Data Set (World Bank, 1992)



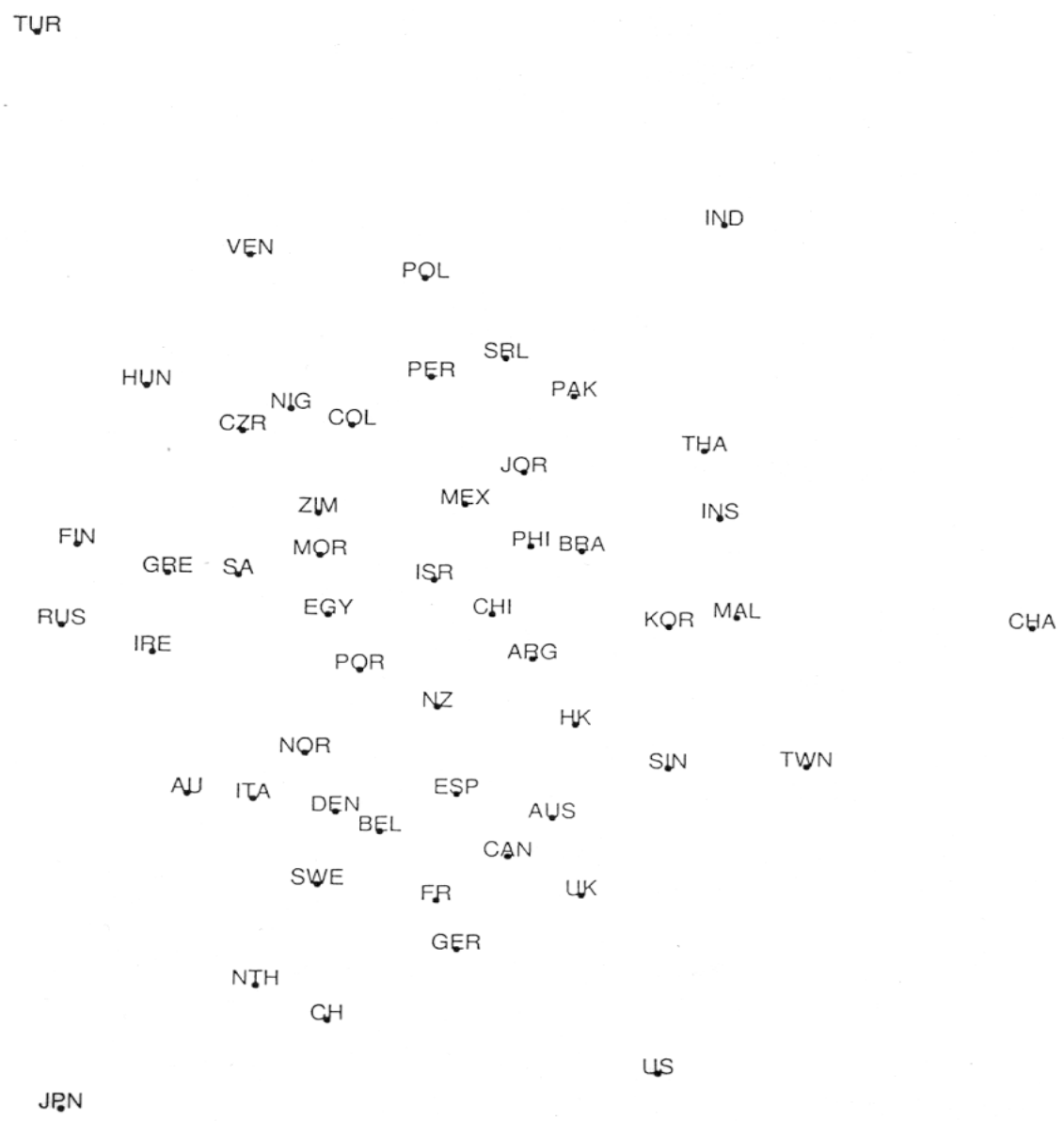
Sammon Mapping of World Poverty Data Set

Self-Organizing Maps (SOMs)

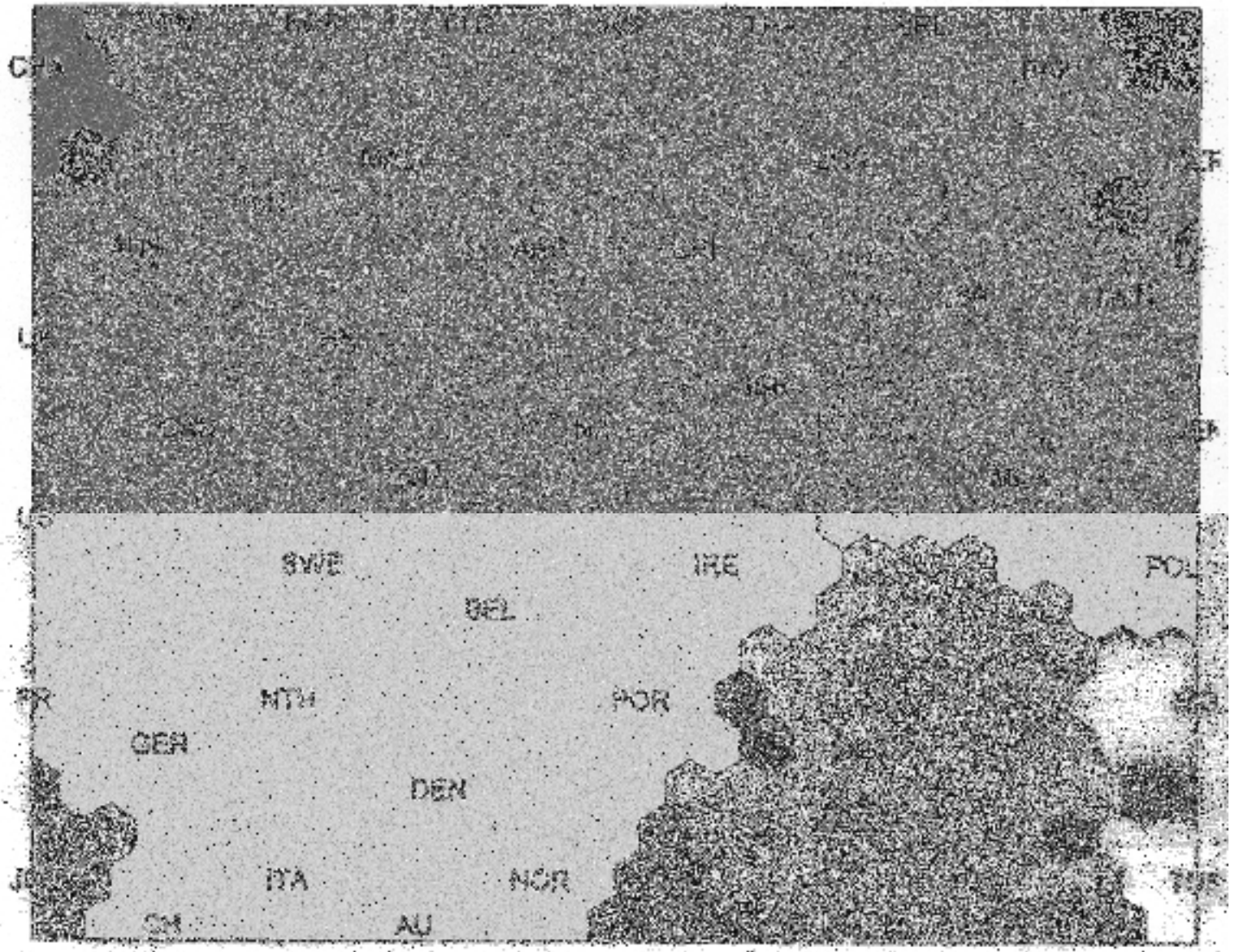
- Developed by Teuvo Kohonen in early 1980s
- Observations are mapped onto a two-dimensional hexagonal grid
- Related to MDS and Sammon maps, but ensures better spacing
- Colors are used to indicate clusters
- Software: SOM_PAK (Public domain, WWW), Viscovery (Eudaptics, Austria)

Country Risk Data

- Goal: Look at risks involved in investing in stock markets around the world
- Source: *Wall Street Journal* of June 26, 1997
- 52 countries, 20 variables
- The article clusters countries into five groups of approximately equal size
  - those most similar to the U.S.
  - other developed countries
  - mature and emerging markets
  - newly emerging markets
  - frontier markets



Sammon Mapping of Country Risk Data

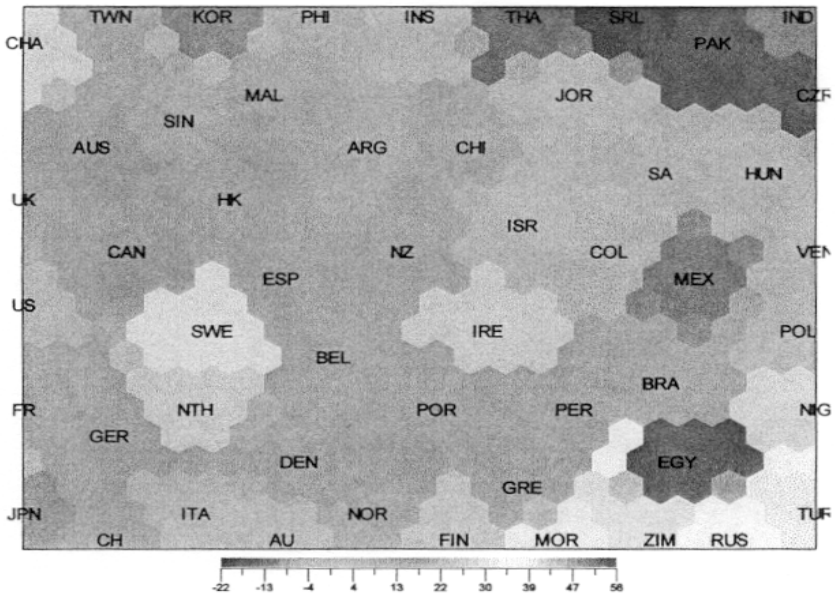


Clustering of countries based on country risk measures using Ward's

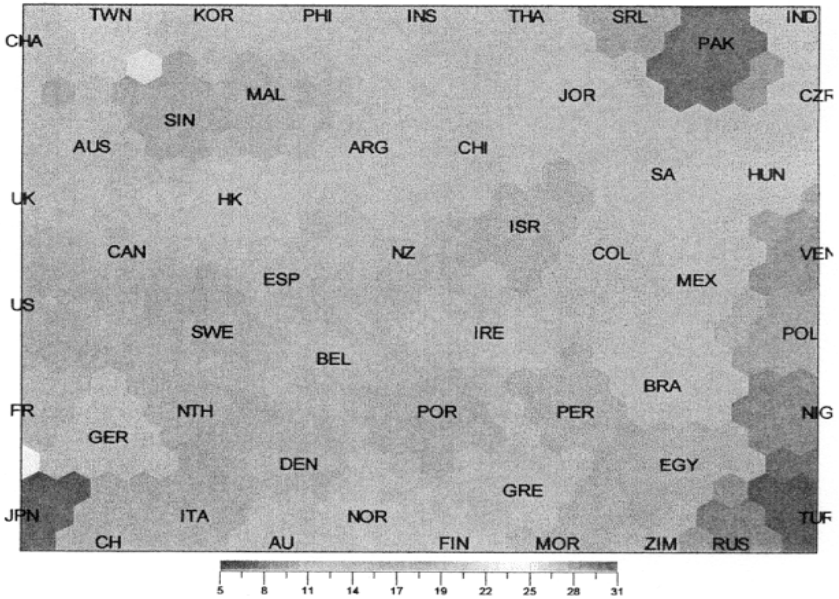
Country Risk Data (continued)

- Nine clusters is a better representation of the data than five clusters
- Component maps and cluster summary statistics help explain why
- Numerous other applications in finance and economics

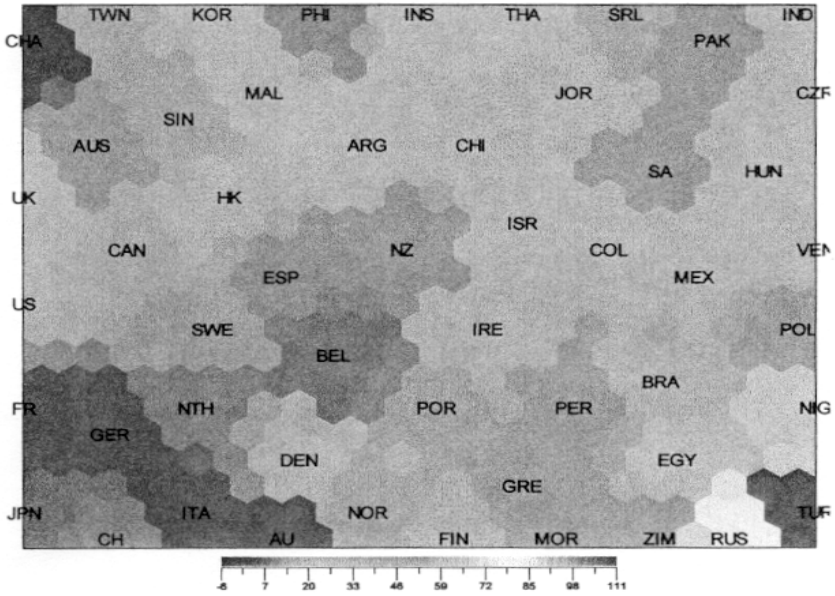
[Comp.ROR3yrsUSS] - AllTrMap1.som



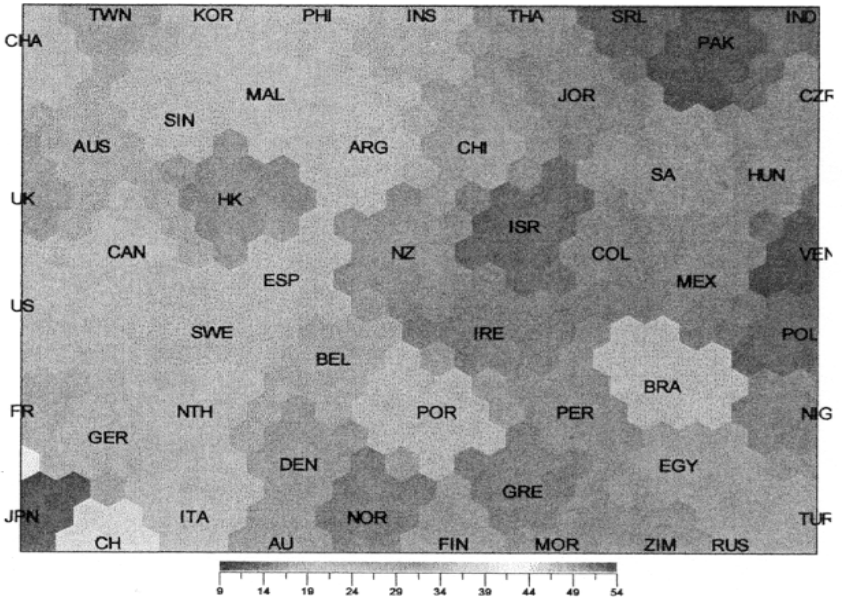
[PEForward] - AllTrMap1.som



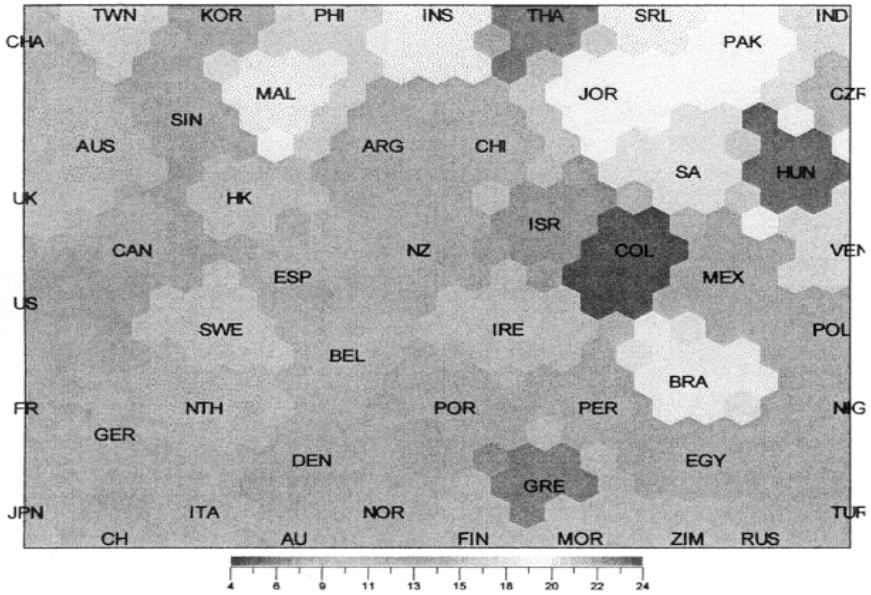
[Hist.Earn.Growth] - AllTrMap1.som



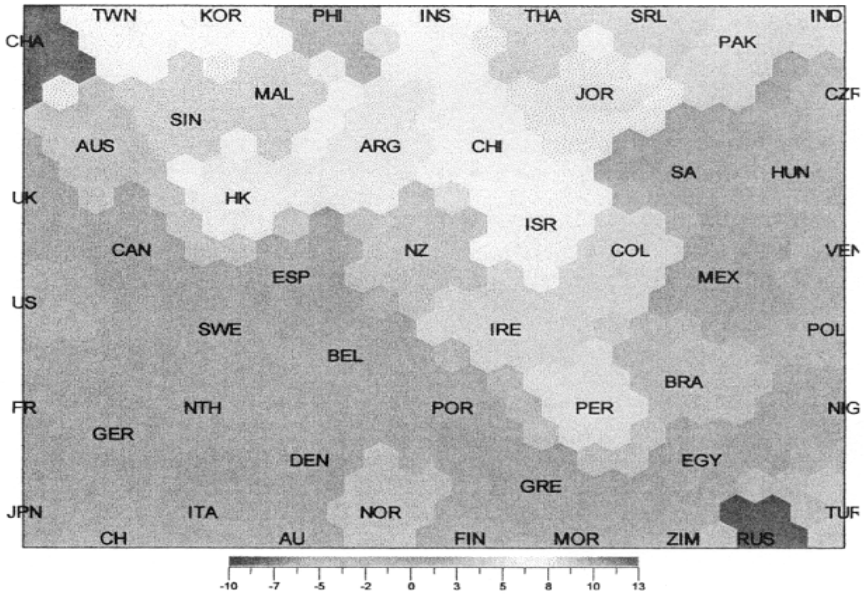
[PEvs1996E] - AllTrMap1.som



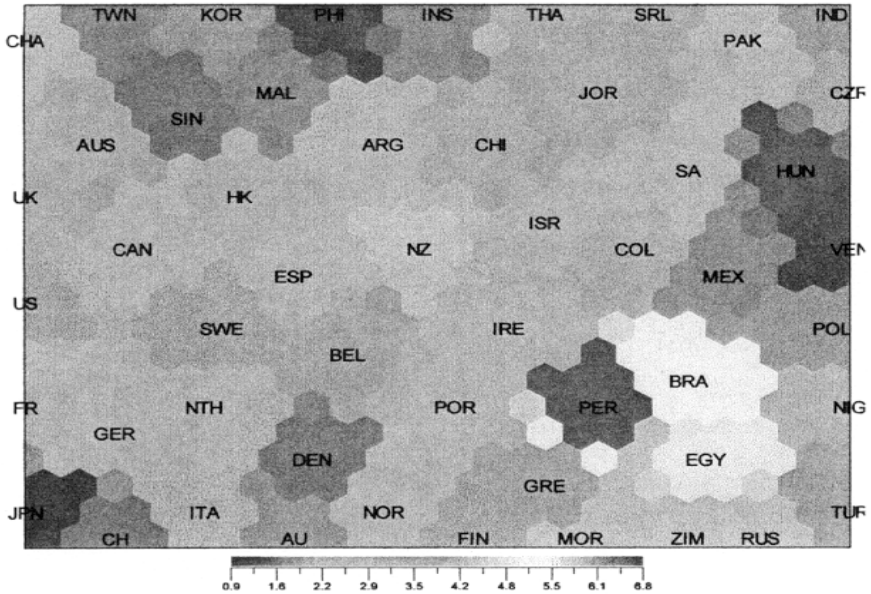
[Proj.Earn.Growth] - AllTrMap1.som



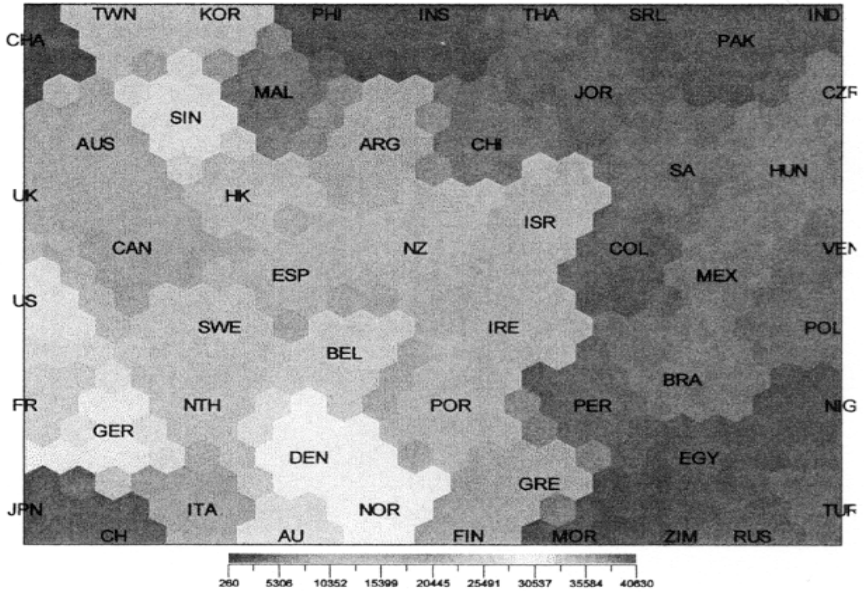
[GDP Growth 90-95] - AllTrMap1.som



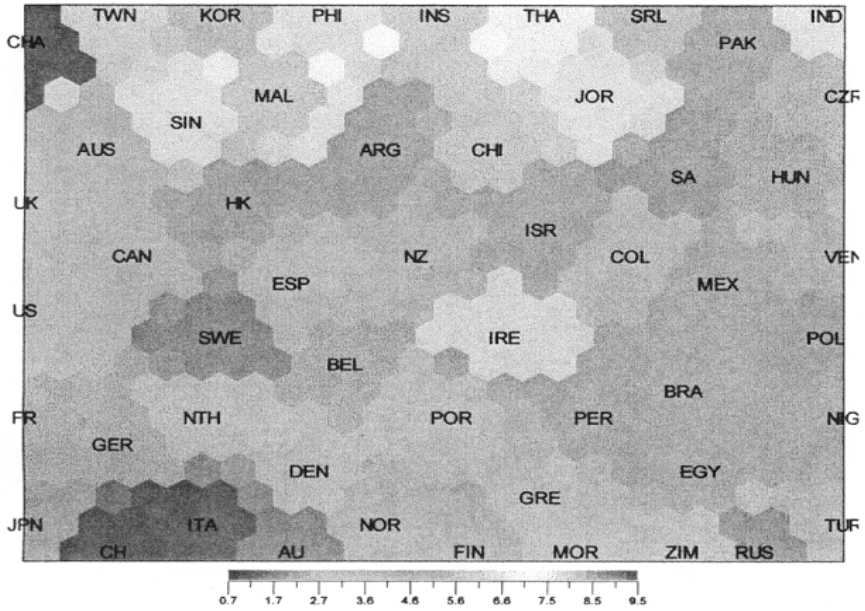
[Did Yld] - AllTrMap1.som



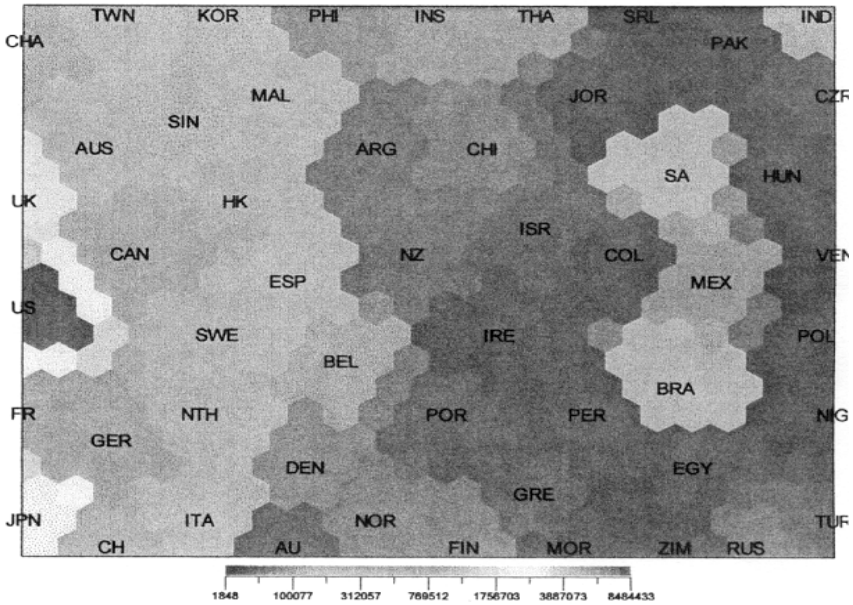
[GNP/capita 1995] - AllTrMap1.som



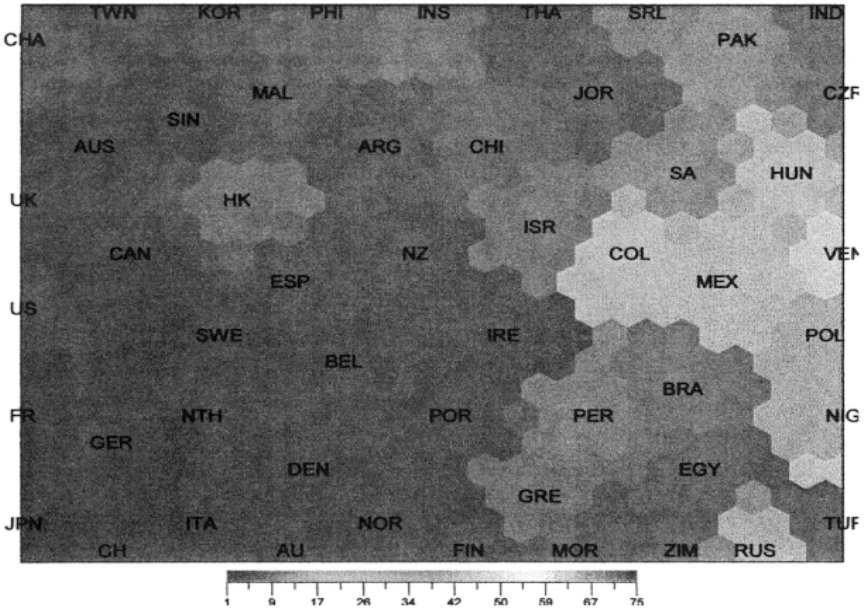
[Proj GDP Growth 97] - AllTrMap1.som



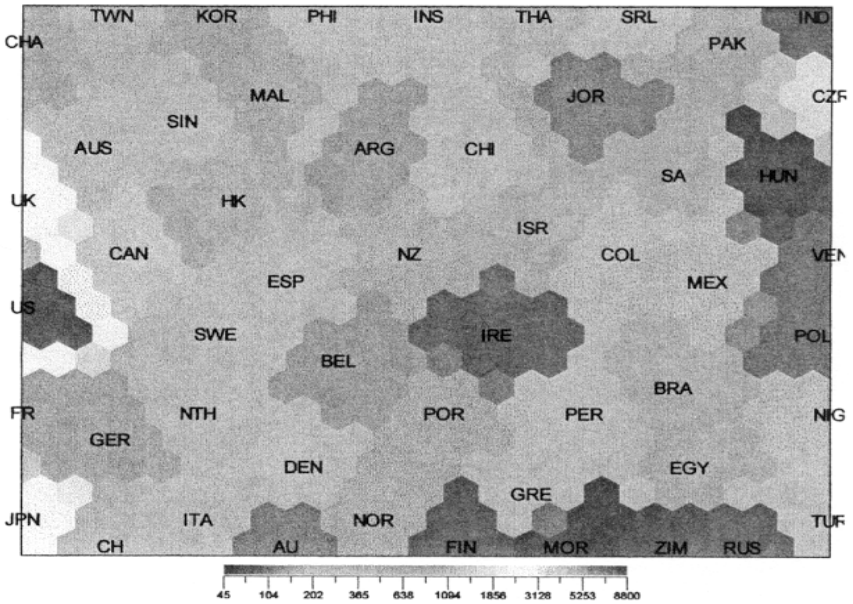
[Market Cap(m $)] - AllTrMap1.som



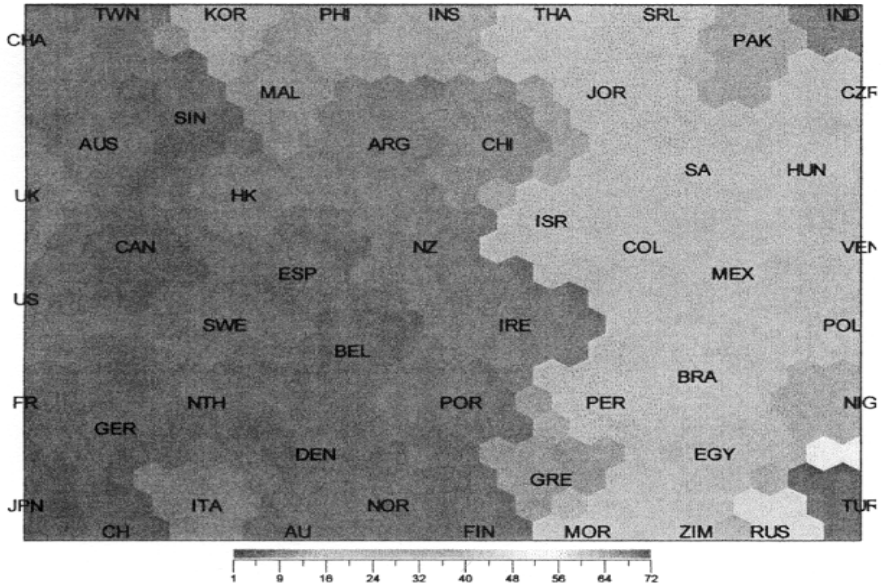
[Proj. Inflation] - AllTrMap1.som



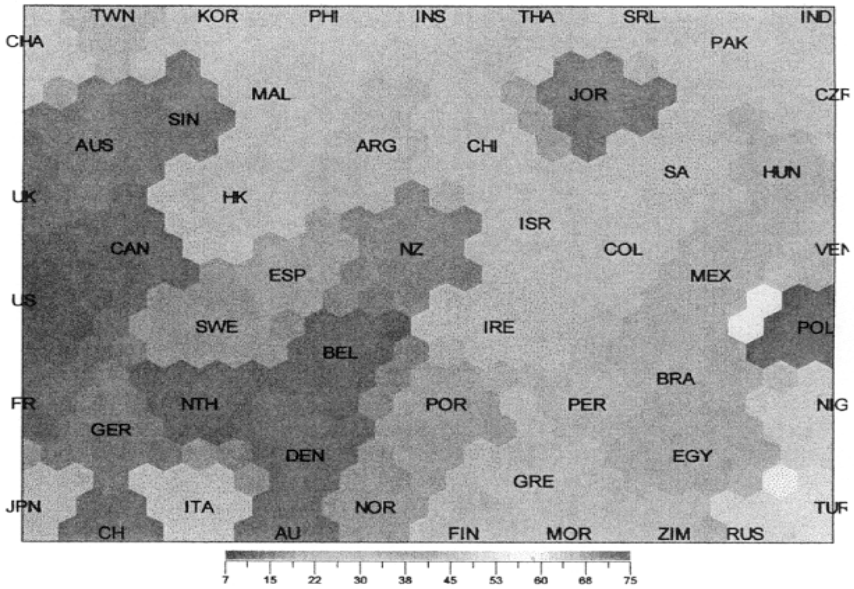
[No Companies] - AllTrMap1.som



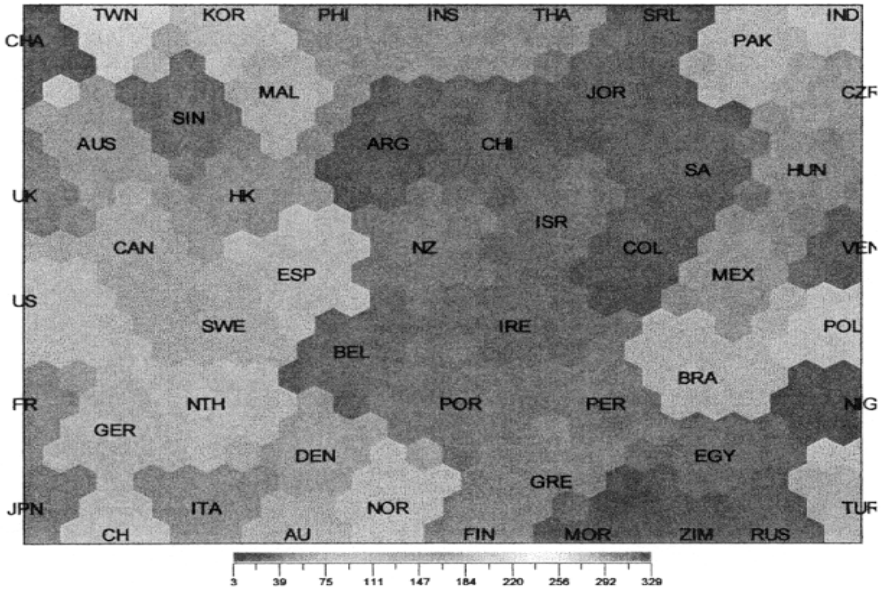
[Short Interest rate] - AllTrMap1.som



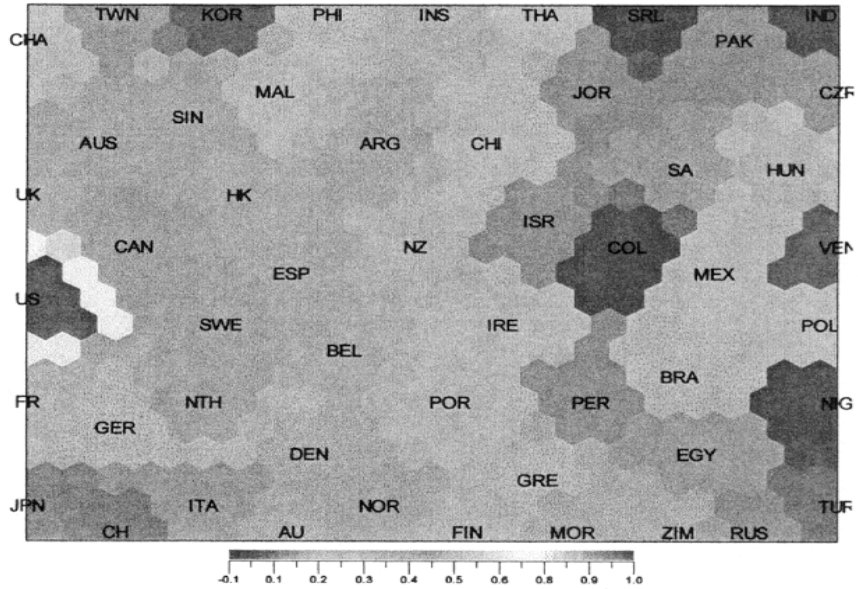
[Volatility] - AllTrMap1.som



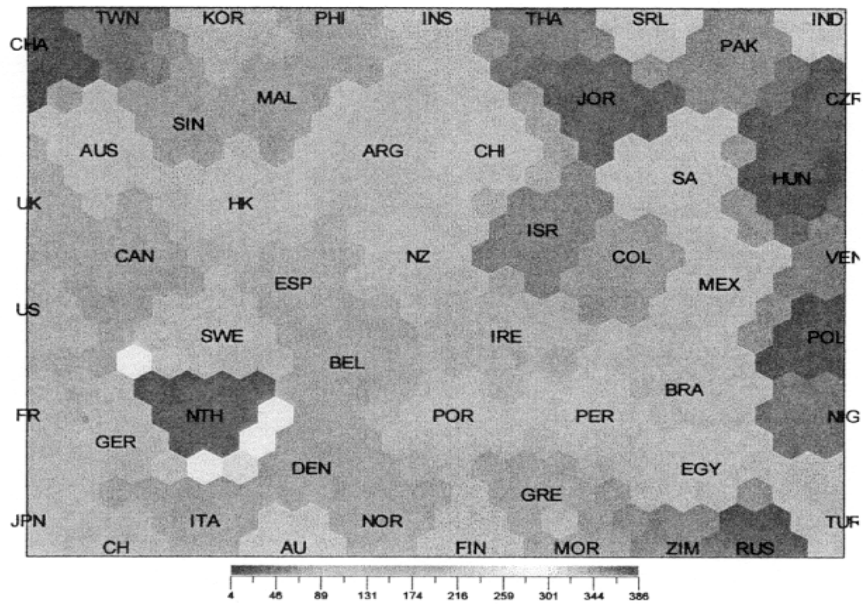
[Turnover %] - AllTrMap1.som



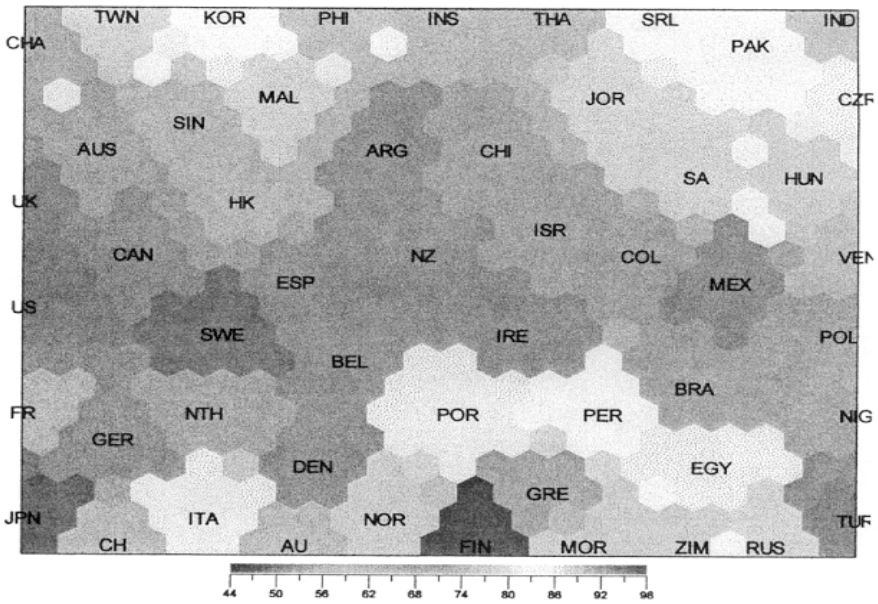
[Correlation_vs_US] - AllTrMap1.som



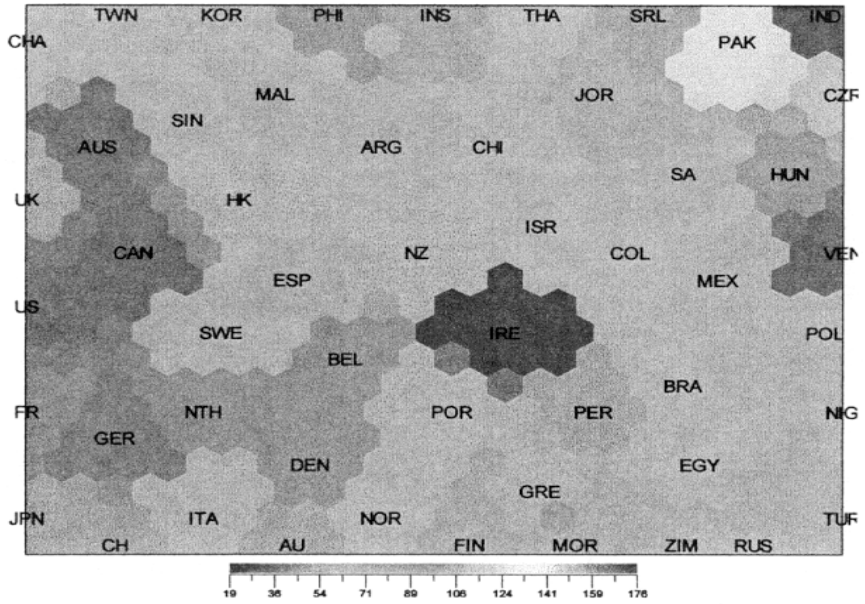
[Age of Market] - AllTrMap1.som



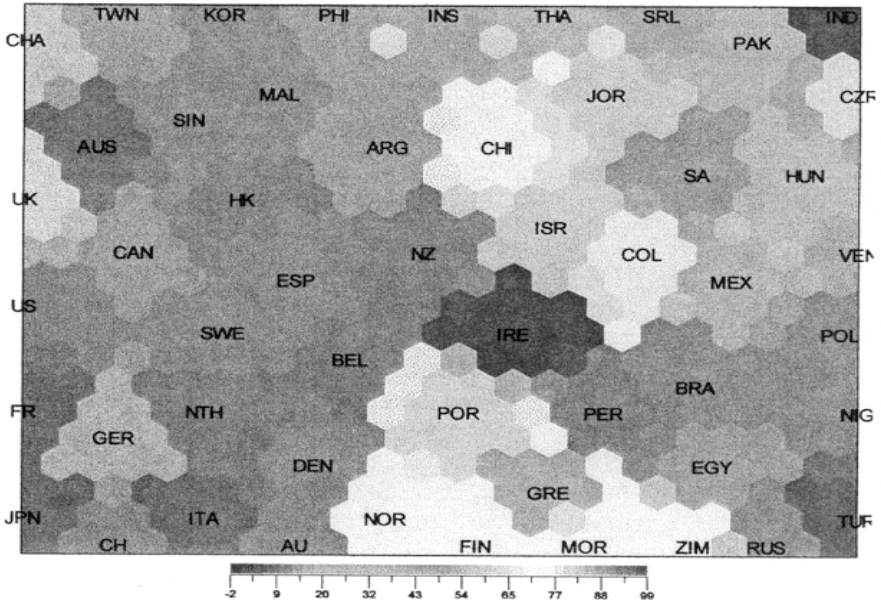
[Safekeeping Efficiency] - AllTrMap1.som



[Operational Costs] - AllTrMap1.som



[Settlement Efficiency] - AllTrMap1.som



Picking Mutual Funds with SOMs

- Source: Morningstar 1997 data on 500 mutual funds
- Among the most successful funds, historically
- Approximately 15 variables
- Categories: World Stocks, International Bonds, Large & Midsize Stocks, Small Cap Stocks, Emerging Markets, All Funds
- Diversification $\Rightarrow$ invest in funds that are in different clusters

Current SOM Projects

- Direct Mail Response
  - observations -- hundreds of thousands of customers
  - variables -- customer history with firm, age, zip code
  - goal -- identify clusters of customers for direct mail promotion

- Profit Opportunities in Telecommunications Worldwide
  - observations -- approximately 200 countries
  - variables -- socio-economic measures, teledensity measures
  - goal -- identify clusters of countries in which demand for wireless services may be high

Flattening the Earth and SOMs: Connections

- There is an art and science to each
- Each is based on sophisticated mathematics
- When you move from many dimensions to two dimensions, you lose important details
- On the other hand, visualization generates insights and impact

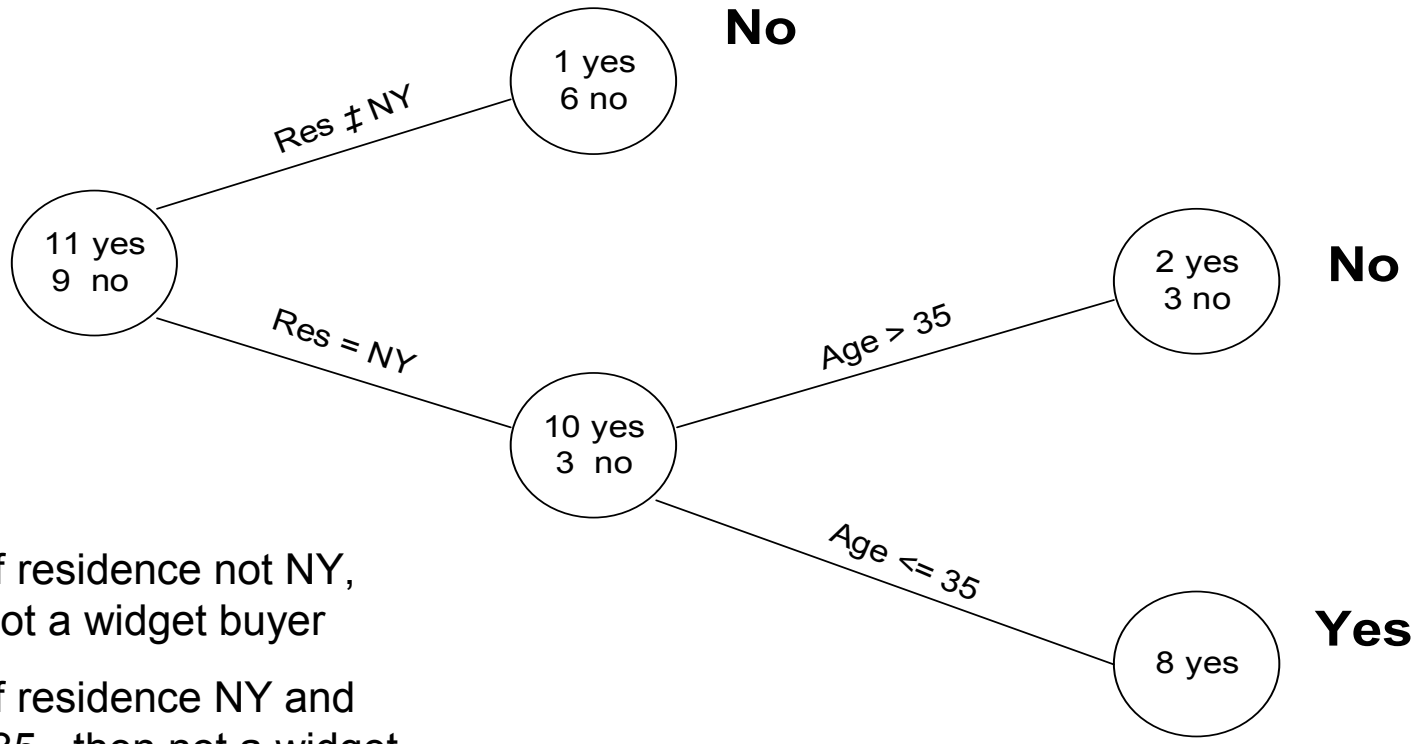
Data Mining and Knowledge Management

- Two types of organizational knowledge
  - Explicit Knowledge
    - databases
    - reports
    - manuals
  - Tacit Knowledge
    - in employees' heads
    - learned from experience
    - not yet codified
- Data mining attempts to convert some of this tacit knowledge into explicit knowledge

Decision Trees

- Given a table of data
  - potential customers are the rows
  - independent variables and dependent variable are the columns
- Decision trees are used for classification, prediction, or estimation of the dependent variable
- Accuracy is typically less than 100%
- One popular approach -- information theory
  - maximize information gain at each split
  - limit the number of splits
  - software: C4.5
- Another popular approach -- statistics
  - software: CART, CHAID

A Decision Tree for Widget Buyers



Rule 1. If residence not NY, then not a widget buyer

Rule 2. If residence NY and age > 35, then not a widget buyer

Rule 3. If residence NY and age <= 35, then a widget buyer

Adapted from (Dhar & Stein, 1997)

$$Accuracy = \frac{\#\ correctly\ \ classified}{total\ \#}$$

$$= \frac{6+3+8}{20} = 0.85$$

Evolutionary Algorithms / Genetic Algorithms

- Developed by John Holland in the late 1960s / early 1970s
- Speed up evolution a millionfold or so on the computer
- Simple, elegant, powerful idea

A Simple Genetic Algorithm

1. Start with a randomly generated population of n chromosomes (candidate solutions to a problem)
2. Calculate the fitness f(x) of each chromosome x in the population
3. Repeat the following steps until n offspring have been created
    a. Randomly select a pair of parent chromosomes from the current population
    b. Crossover (mate) the pair at a randomly chosen point to form two offspring
    c. Randomly mutate the two offspring and add the resulting chromosomes to the population
    d. Calculate the fitness of the resulting chromosomes

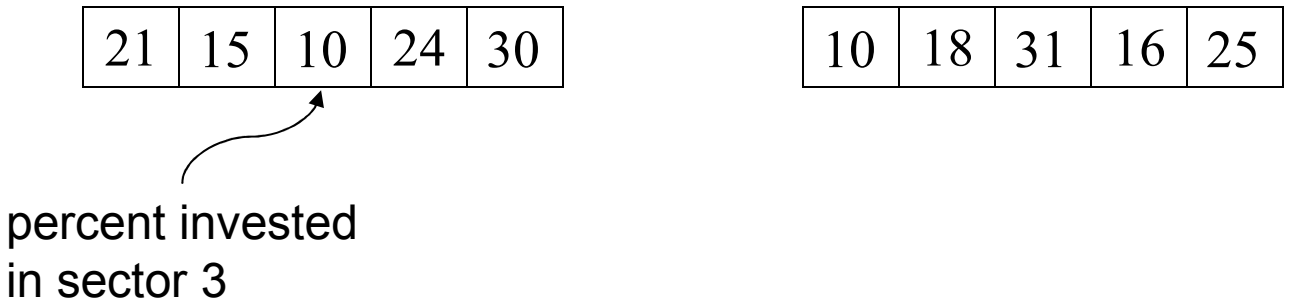
A Simple Genetic Algorithm (continued)

4. Let the n fittest chromosomes survive to the next generation
5. Go to Step 3 (repeat for 50 generations)

Financial Investment Example

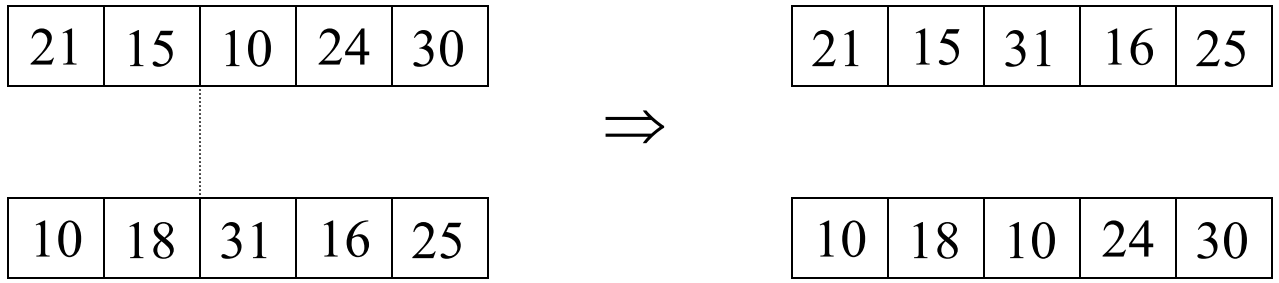
- Five sectors
  1. Financial services
  2. Health care
  3. Utilities
  4. Technology
  5. Consumer

- Two parents



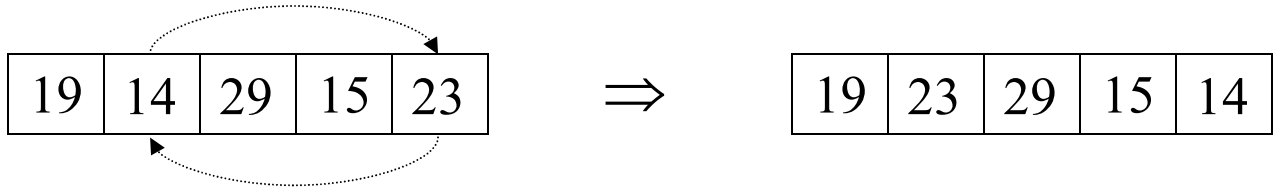
Financial Investment Example (continued)

- Crossover

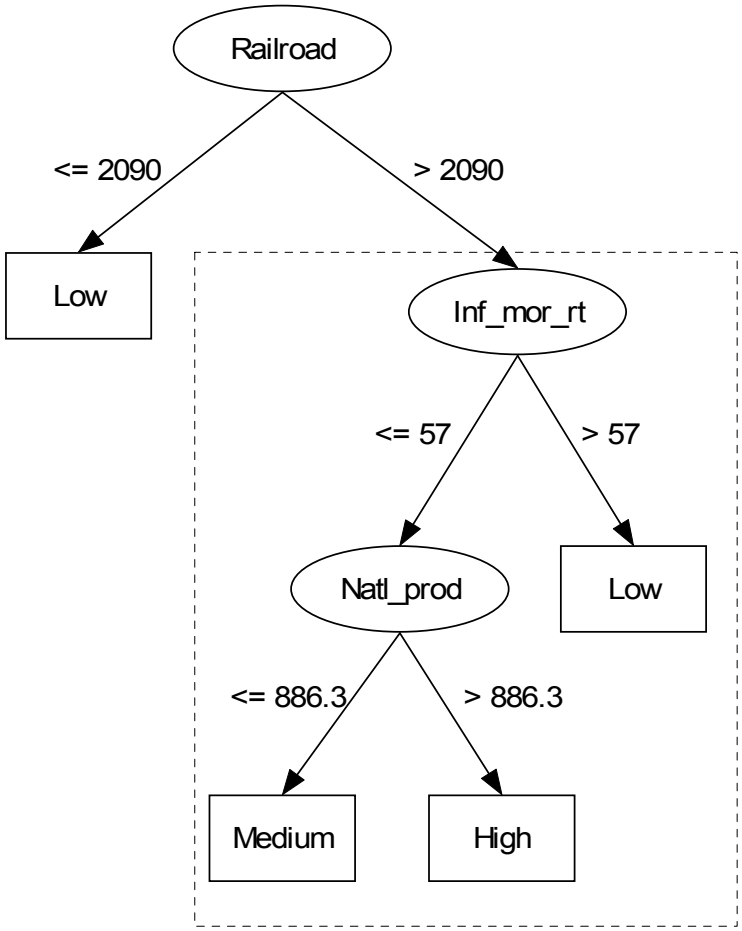


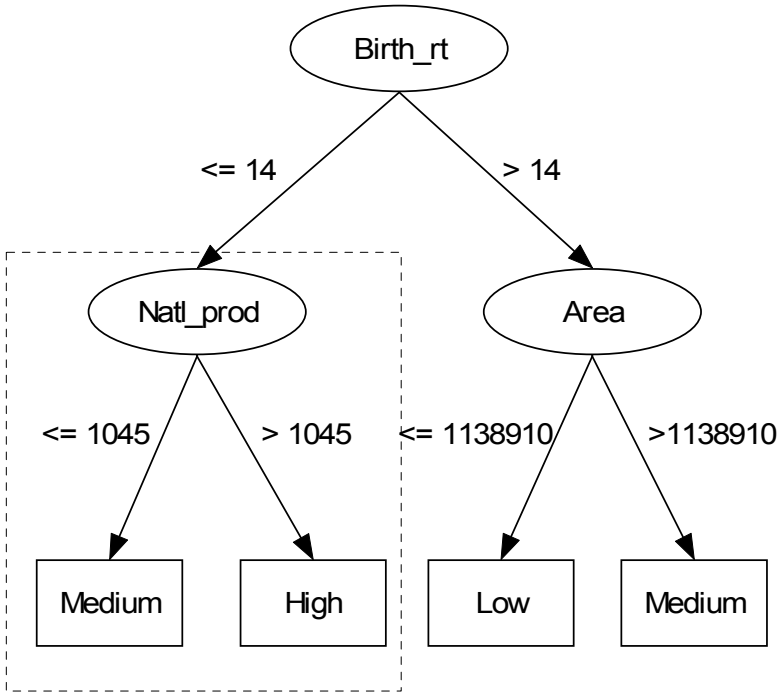
- After normalization and rounding, we obtain two offspring

- Mutation



Crossover Illustration for Decision Trees

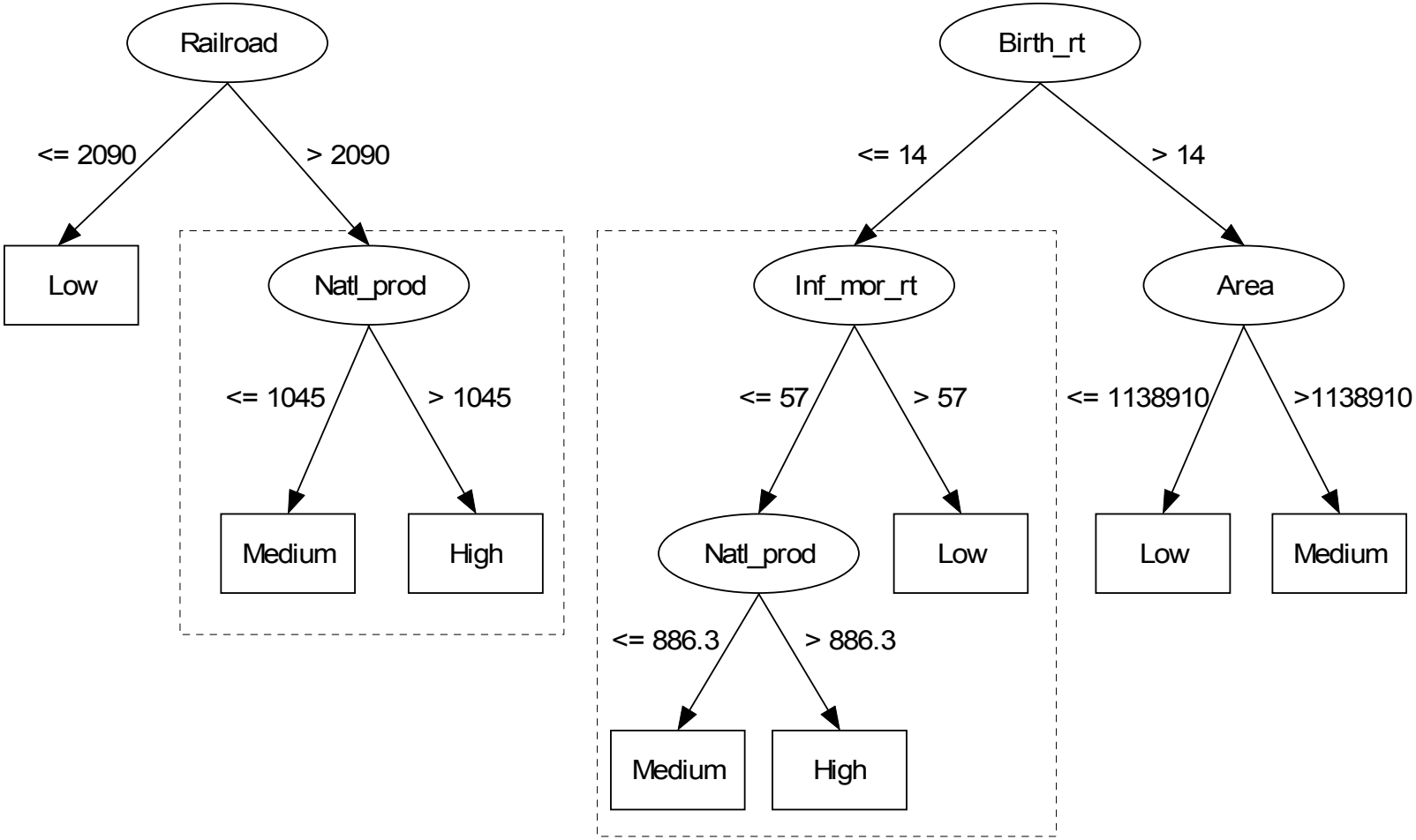




Parent 1

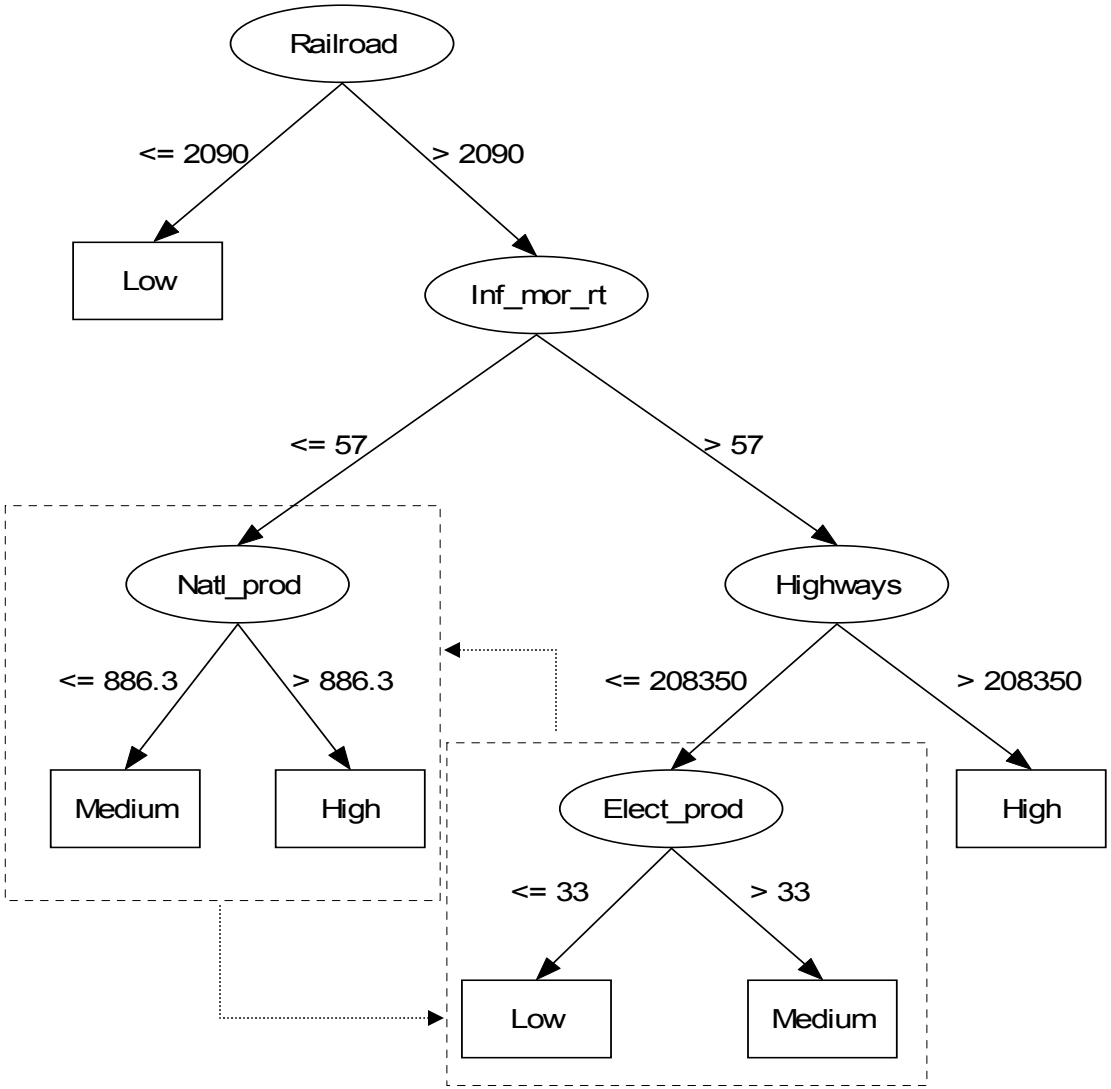
Parent 2

Crossover Illustration (continued)



Child 1 **Child 2**

Mutation Illustration



GA Applications

- Financial Data Analysis
  - State Street Global Advisors
  - Advanced Investment Technologies
  - Barclays Global Investors
  - PanAgora Asset Management
  - Fidelity Funds
- Operations and Supply Chain Management
  - General Motors
  - Volvo
  - Cemex
- Engineering Design
  - General Electric
  - Boeing

Data Analysis Then and Now

Late 1980s	Late 1990s
Linear methods	Nonlinear methods
Large data sets	Massive data sets
Few dimensions	Highly multi-dimensional
Ask specific questions	What can we infer?
Search for information	Search for knowledge

Data Analysis as a Strategic Asset

- Competitive advantage
- Sustainable over a period of time

Examples

- BT Labs
- Enterprise Rent-A-Car
- Dupont

Concluding Remarks

- Corporate data is more plentiful than ever before
- Companies are becoming more serious about mining that data
- Powerful software tools are widely available
- Many companies already view data analysis/data mining as a strategic asset
- Data analysis/data mining is a key area within knowledge management
- Exciting opportunities exist for collaborative research

Recommended Books

- Berry & Linoff, Data Mining Techniques for Marketing, Sales, and Customer Support, Wiley (1997)
- Deboeck & Kohonen, Visual Explorations in Finance with SOMs, Springer (1998)
- Dhar & Stein, Seven Methods for Transforming Corporate Data into Business Intelligence, Prentice Hall (1997)
- Mitchell, An Introduction to Genetic Algorithms, MIT Press (1996)
- Monmonier, How to Lie with Maps (second edition), University of Chicago Press (1996)