TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA HỆ THỐNG THÔNG TIN

Tài liệu bài giảng:

KHAI THÁC DỮ LIỆU – IS252

Chương 8:

Các xu hướng nghiên cứu

ThS. Dương Phi Long – Email: longdp@uit.edu.vn

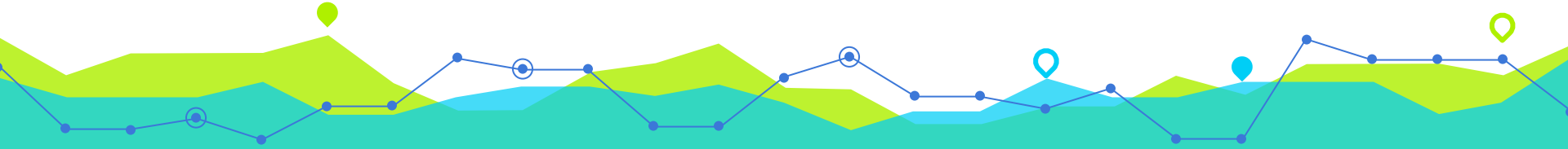
NỘI DUNG BÀI HỌC

01 Khai thác các kiểu dữ liệu phức tạp

02 Các vấn đề khác của KTDL

03 Ứng dụng KTDL

04 Các xu hướng nghiên cứu

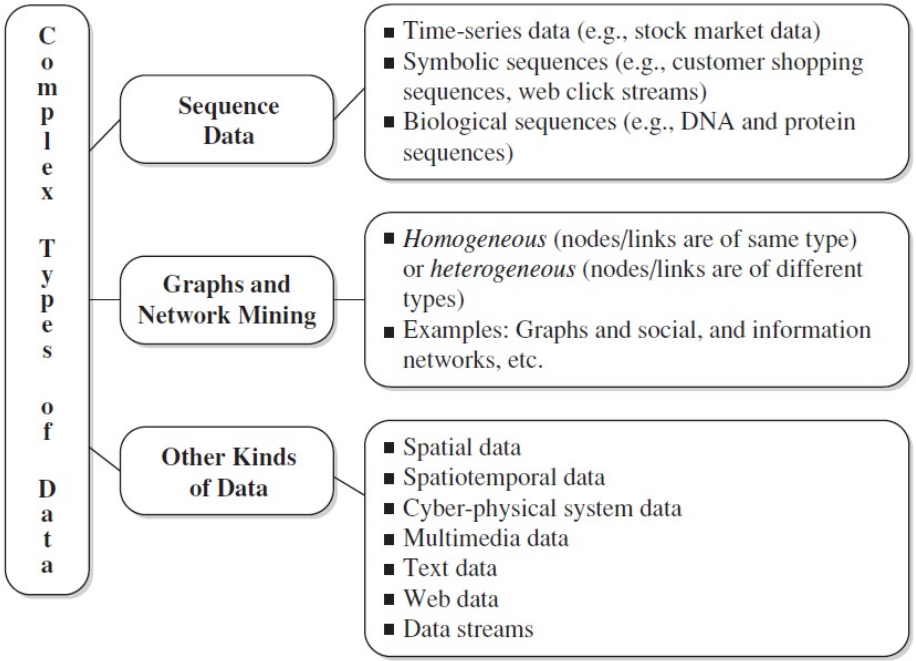


1 Khai thác các kiểu dữ liệu phức tạp

1. Mining Sequence Data
2. Mining Graphs and Networks
3. Mining Other Kinds of Data

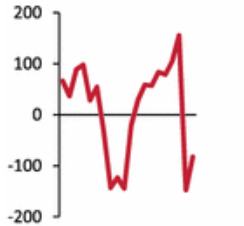
Khai thác các KDL phức tạp

- **Khai thác dữ liệu tuần tự**
  - Time series
  - Symbolic Sequences
  - Biological Sequences
- **Khai thác đồ thị và mạng**
- **Khai thác các KDL khác**

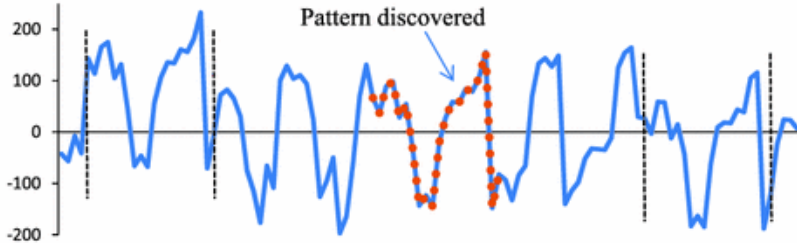


1. Khai thác dữ liệu tuần tự

- **Similarity Search in Time Series Data**
  - Subsequence match
  - Dimensionality reduction
    - Principle components analysis (PCA)
    - Discrete Fourier transform (DFT)
    - Discrete wavelet transforms (DWT)
    - Singular value decomposition (SVD)
  - Query-based similarity search
  - Motif-based similarity search



(a) Pattern

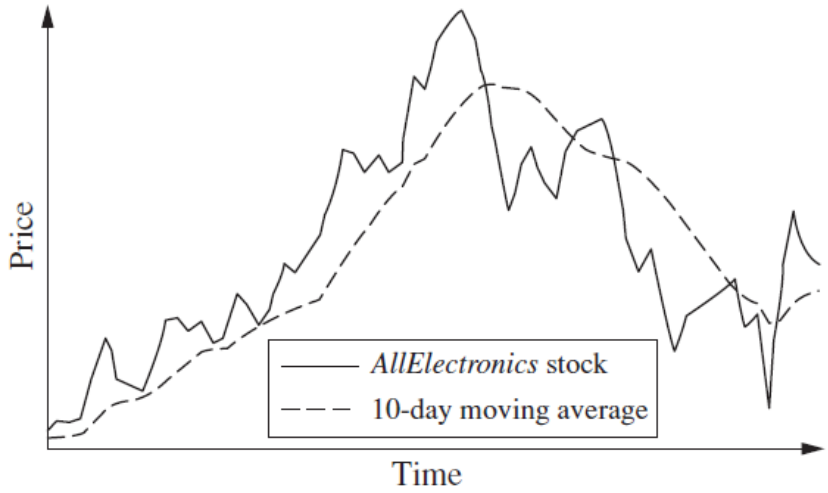


(b) EEG time series

1. Khai thác dữ liệu tuần tự

- **Regression and Trend Analysis in Time-Series Data**

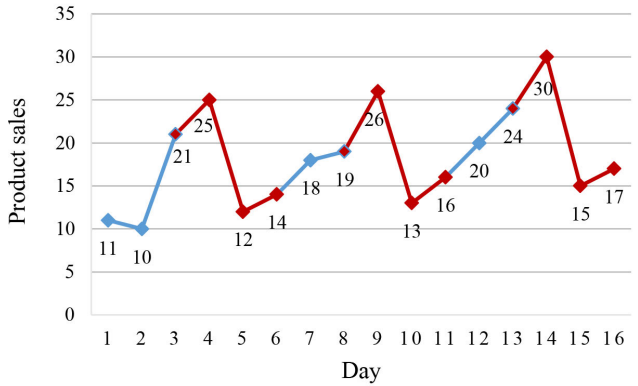
  - Trend or long-term movements: Sử dụng các phương pháp trung bình động có trọng số và bình phương nhỏ nhất để tìm các đường cong xu hướng.
  - Cyclic movements: dao động dài hạn
  - Seasonal variations: VD: mùa lễ hội mua sắm
  - Random movements



1. Khai thác dữ liệu tuần tự

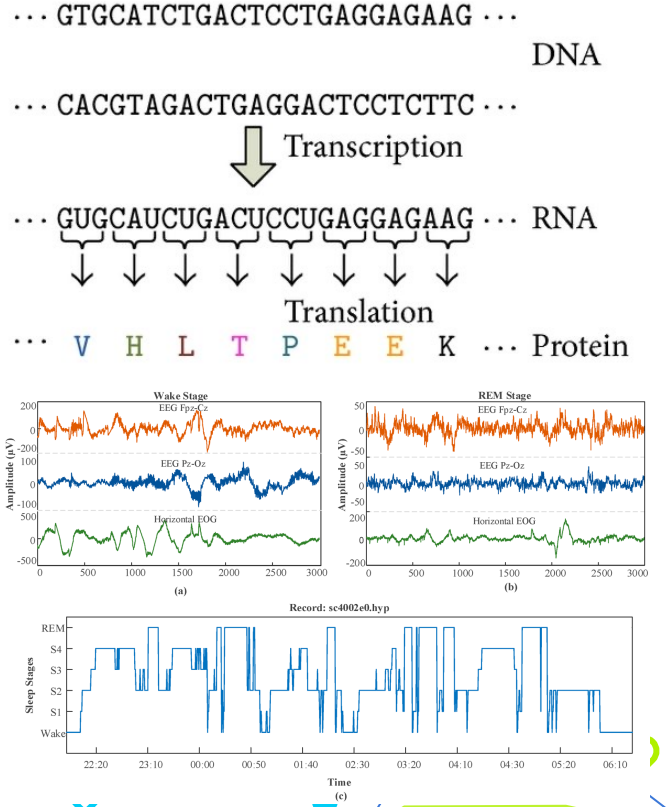
- **Sequential Pattern Mining in Symbolic Sequences**
  - Mining symbolic sequences.
  - Constraint-based sequential pattern mining: sử dụng các ràng buộc do người dùng chỉ định để giảm không gian tìm kiếm trong khai thác mẫu tuần tự và chỉ lấy được các mẫu mà người dùng quan tâm.
  - Relax constraints:

    Đưa các sự kiện vào các cửa sổ có kích thước phù hợp và tìm các chuỗi con định kỳ trong các cửa sổ này



1. Khai thác dữ liệu tuần tự

- **Sequence Classification**
  - Feature-based classification: Chuyển đổi một chuỗi thành một vector đặc trưng và sau đó áp dụng các phương pháp phân lớp thông thường
  - Sequence distance–based classification: Đo lường mức độ tương tự giữa các sequence.
  - Model-based classification: Mô hình Hidden Markov

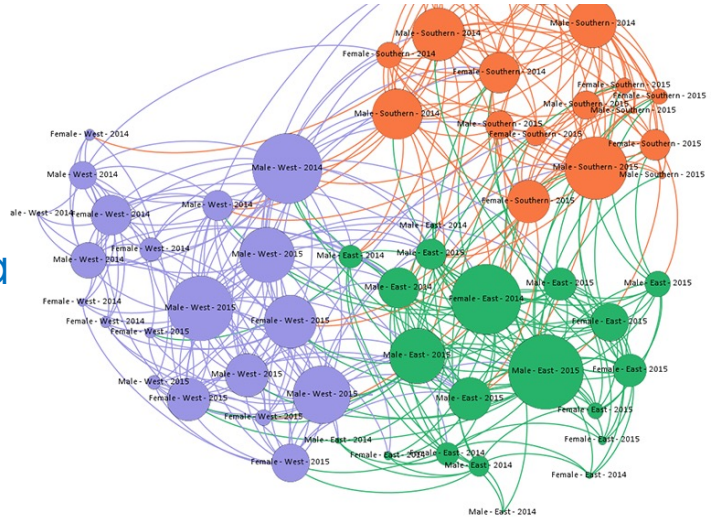


1. Khai thác dữ liệu tuần tự

- **Alignment of Biological Sequences**
  - Sequence alignment:
    - Sắp xếp các sequence để đạt được mức nhận dạng tối đa
    - Sắp xếp cục bộ và toàn cục
  - Substitution matrices:
    - Thể hiện xác suất thay thế/ chèn/ xóa nucleotide hoặc axit amin.

2. Khai thác đồ thị và mạng

- **Graph Pattern Mining**
  - Khai thác mẫu đồ thị con phổ biến
  - Tìm kiếm sự tương đồng về cấu trúc
- **Statistical Modeling of Networks**
  - Homogeneous (đồng nhất): Các node và liên kết cùng loại.
  - Heterogeneous (không đồng nhất): Các node và liên kết khác loại.
  - Scale-free model: power law distribution
- **Data Cleaning, Integration, and Validation by Information Network Analysis**

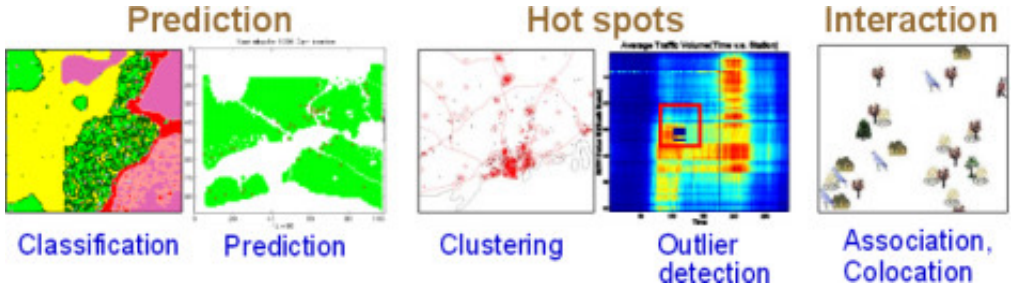


2. Khai thác đồ thị và mạng

- **Clustering and Classification of Graphs and Homogeneous Networks**
  - Khám phá các cộng đồng, trung tâm và ngoại lệ ẩn
- **Clustering, Ranking, and Classification of Heterogeneous Networks**
- **Role Discovery and Link Prediction in Information Networks**
  - Link prediction: Đánh giá các mối quan hệ dự kiến giữa các node/ liên kết ứng cử viên.
- **Similarity Search and OLAP in Information Networks**
  - OLAP: Xử lý phân tích trực tuyến
  - Path-based similarity
- **Evolution of Social and Information Networks**

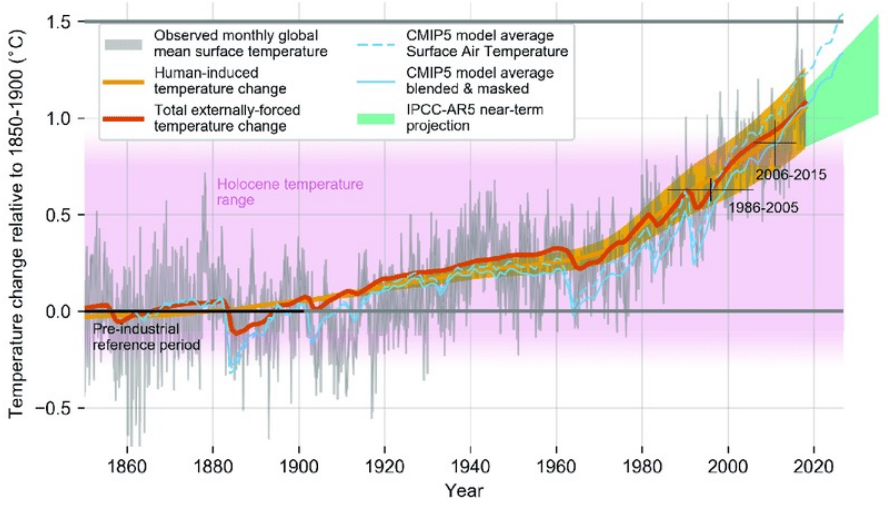
3. Khai thác các KDL khác

- **Mining Spatial Data**
  - Khám phá các mẫu và kiến thức từ dữ liệu không gian, như dữ liệu liên quan đến không gian địa lý
  - Các chủ đề:
    - Mining spatial associations and co-location patterns
    - Spatial clustering
    - Spatial classification
    - Spatial modeling
    - Spatial trend and outlier analysis
    - Discover hidden communities, hubs, and outliers



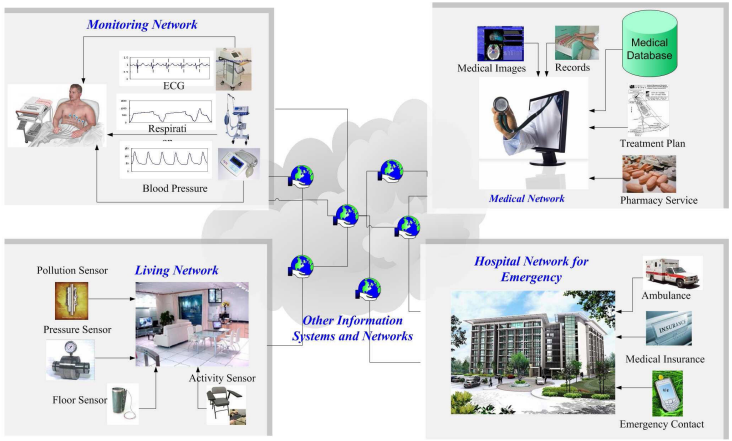
3. Khai thác các KDL khác

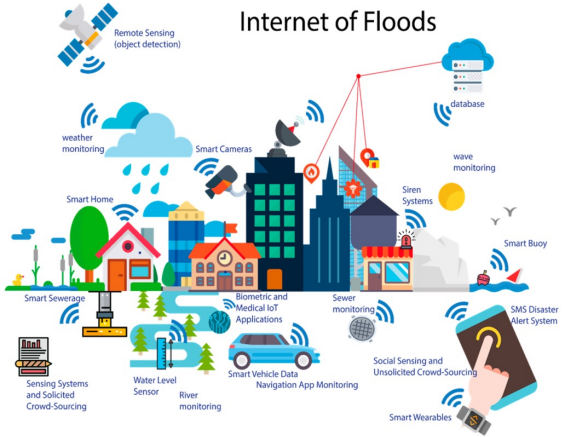
- **Mining Spatiotemporal Data and Moving Objects**
  - Spatiotemporal Data: Liên quan đến cả không gian và thời gian, như lịch sử tiến hóa của các thành phố và vùng đất, xu hướng nóng lên toàn cầu
  - Moving-object data: Khai thác các mẫu chuyển động của nhiều đối tượng chuyển động



3. Khai thác các KDL khác

- **Mining Cyber-Physical System Data**
  - Hệ thống giao thông kết nối mạng lưới giám sát giao thông
  - Chăm sóc sức khỏe, kiểm soát không lưu, mô phỏng lũ lụt
  - Cần tính toán thời gian thực và phản hồi nhanh chóng



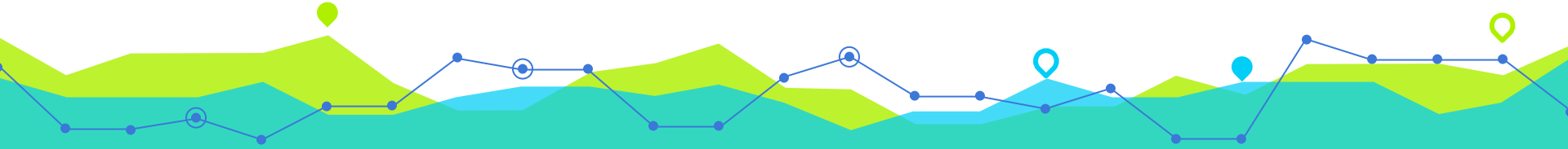


3. Khai thác các KDL khác

- **Mining Multimedia Data**
  - Dữ liệu hình ảnh, dữ liệu video, dữ liệu âm thanh, dữ liệu tuần tự và dữ liệu siêu văn bản
- **Mining Text Data**
  - Khám phá các mẫu và xu hướng bằng cách sử dụng học mẫu thống kê, mô hình chủ đề và mô hình ngôn ngữ thống kê, ….

3. Khai thác các KDL khác

- **Mining Web Data**
  - Web content mining: văn bản, dữ liệu đa phương tiện và dữ liệu có cấu trúc
  - Web structure mining: hyperlinks
    - Sử dụng các phương pháp khai thác đồ thị và mạng để phân tích các node và cấu trúc kết nối trên Web.
  - Web usage mining: Server logs
    - Hiểu các mẫu, xu hướng và liên kết tìm kiếm của người dùng
    - Dự đoán những gì người dùng đang tìm kiếm trên Internet

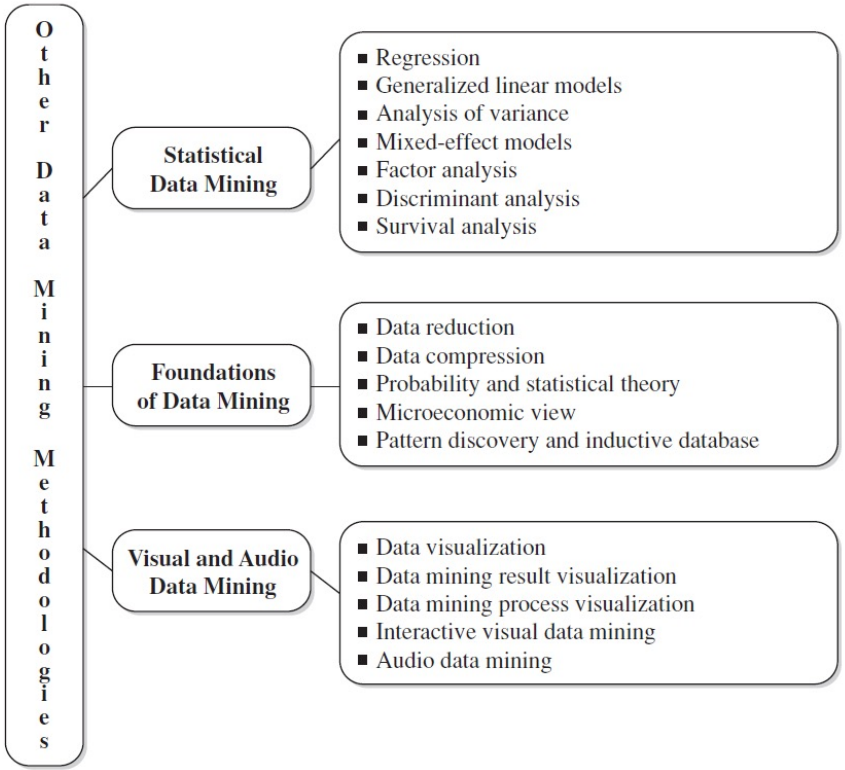


2 Các vấn đề khác của KTDL

1. KTDL thống kê
2. Quan điểm về KTDL
3. Trực quan KTDL

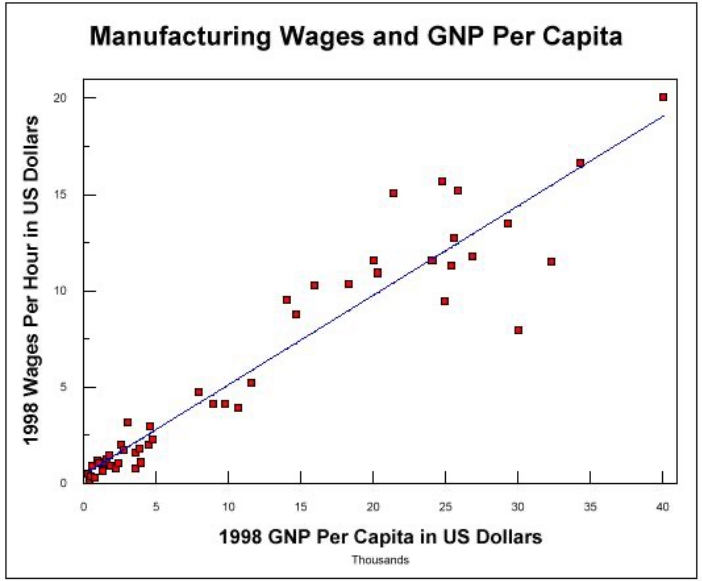
Các vấn đề khác của KTDL

- Khai thác dữ liệu thống kê
- Quan điểm về nền tảng khai thác dữ liệu
- Trực quan KTDL



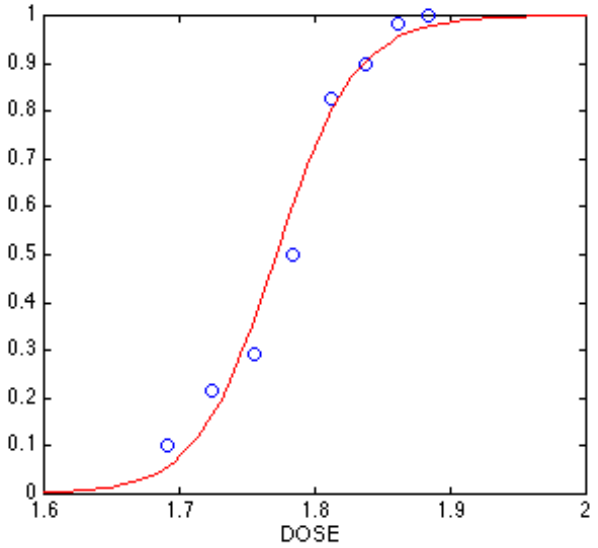
1. Khai thác dữ liệu thống kê

- **Regression:**
  - Dự đoán giá trị của biến phản hồi (phụ thuộc) từ một hoặc nhiều biến dự đoán (độc lập) trong đó các biến là số
  - Các dạng regression: linear, multiple, weighted, polynomial, nonparametric, and robust



1. Khai thác dữ liệu thống kê

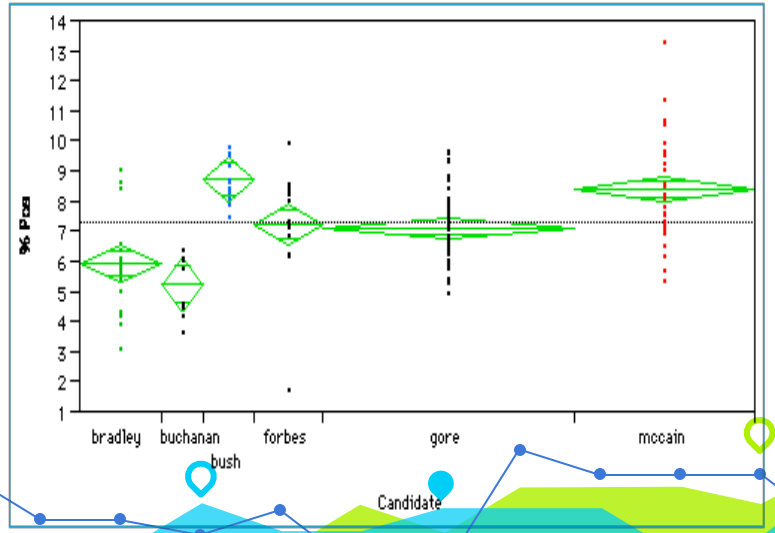
- **Generalized linear models**
  - Cho phép một biến phản hồi phân loại được liên kết với một tập hợp các biến dự đoán
  - Tương tự như mô hình hóa một biến phản hồi số bằng hồi quy tuyến tính
  - Bao gồm: logistic regression and Poisson regression



1. Khai thác dữ liệu thống kê

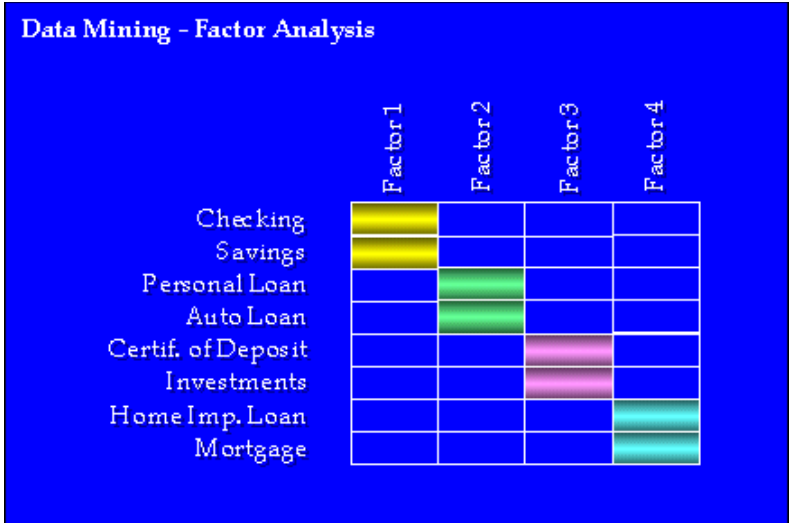
- **Mixed-effect models**
  - Phân tích dữ liệu được nhóm, dữ liệu có thể được phân loại theo một hoặc nhiều biến nhóm
  - Mô tả điển hình các mối quan hệ giữa một biến phản hồi và một số đồng biến trong dữ liệu được nhóm theo một hoặc nhiều yếu tố

- **Analysis of variance**
  - Phân tích dữ liệu thử nghiệm cho hai hoặc nhiều nhóm được mô tả bằng một biến phản hồi số và một hoặc nhiều biến phân loại (yếu tố)



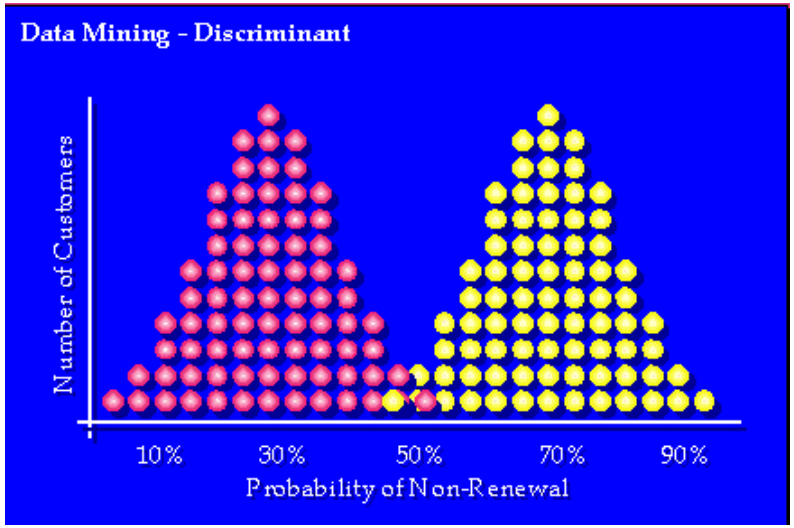
1. Khai thác dữ liệu thống kê

- **Factor analysis**
  - Xác định biến nào được kết hợp để tạo ra một yếu tố nhất định
  - VD: đối với nhiều dữ liệu tâm thần, có thể gián tiếp đo lường các đại lượng khác phản ánh factor được quan tâm



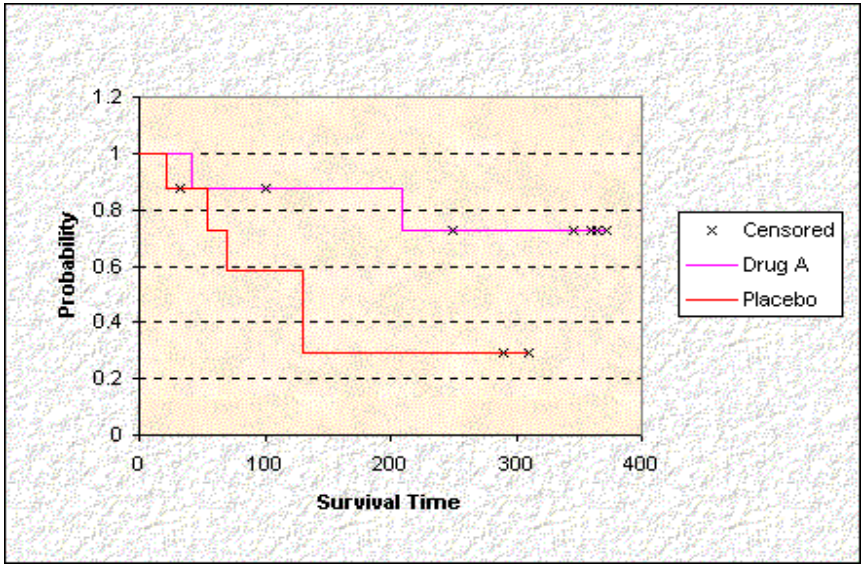
1. Khai thác dữ liệu thống kê

- **Discriminant analysis**
  - Dự đoán một biến phản ứng phân loại, thường được sử dụng trong khoa học xã hội
  - Cố gắng xác định một số hàm phân biệt (tổ hợp tuyến tính của các biến độc lập) phân biệt giữa các nhóm được xác định bởi biến phản hồi.



1. Khai thác dữ liệu thống kê

- **Quality control:**
  - Thể hiện biểu đồ tóm tắt nhóm: Shewhart charts, CUSUM charts
- **Survival analysis**
  - Dự đoán khả năng một bệnh nhân đang được điều trị y tế sẽ sống sót ít nhất đến thời điểm t (dự đoán tuổi thọ)



2. Quan điểm về nền tảng khai thác dữ liệu

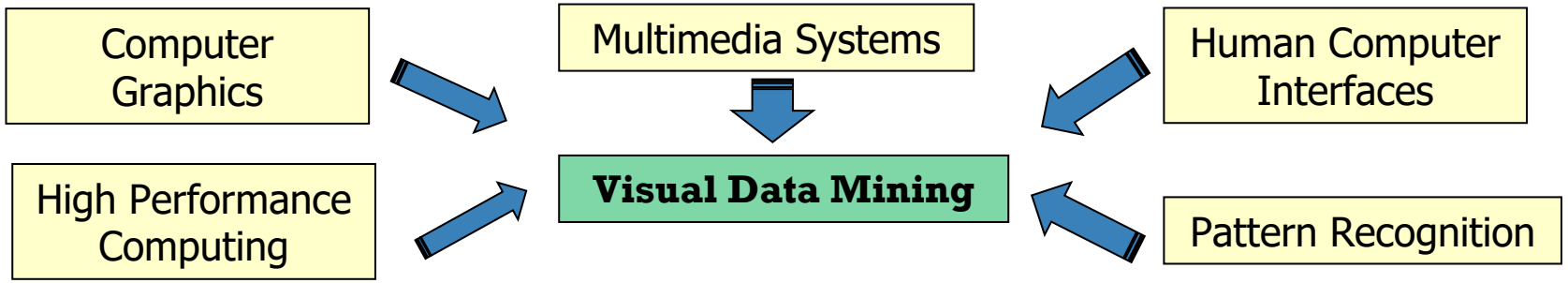
- **Data reduction**
  - Reduce data representation
  - Đánh đổi độ chính xác để lấy tốc độ đáp ứng
  - Bao gồm: singular value decomposition, wavelets, regression, log-linear models, histograms, clustering, sampling, the construction of index trees
- **Data compression**
  - Nén dữ liệu đã cho bằng cách mã hóa theo bit, quy tắc kết hợp, cây quyết định, cụm, ….
- **Probability and statistical theory**
  - Khám phá các phân phối xác suất chung của các biến ngẫu nhiên

2. Quan điểm về nền tảng khai thác dữ liệu

- **Microeconomic view (quan điểm kinh tế vi mô)**
  - Tìm kiếm các mẫu thú vị có thể được sử dụng trong quá trình ra quyết định của một số doanh nghiệp
- **Pattern Discovery and Inductive databases (CSDL quy nạp và khám phá mẫu)**
  - Khám phá các mẫu xuất hiện trong CDSL: các liên kết, mô hình phân loại, mẫu tuần tự, ….
  - KTDL là vấn đề thực hiện logic quy nạp trên cơ sở dữ liệu, truy vấn dữ liệu và các mẫu của CSDL

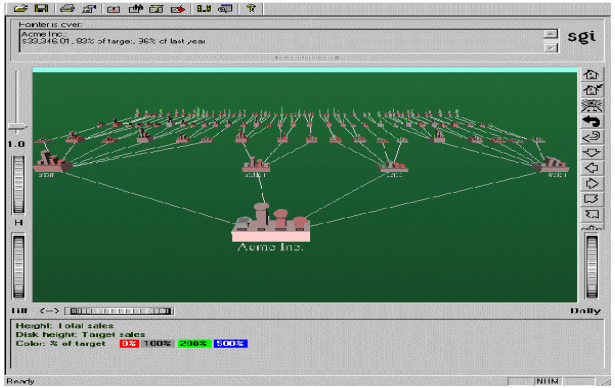
3. Trực quan KTDL

- **Visual Data Mining**
  - Data visualization: Sử dụng đồ họa máy tính để tạo hình ảnh trực quan giúp hiểu được các biểu diễn dữ liệu phức tạp, thường có khối lượng lớn.
  - Visual Data Mining: khám phá kiến thức tiềm ẩn nhưng hữu ích từ các tập dữ liệu lớn bằng cách sử dụng các kỹ thuật trực quan



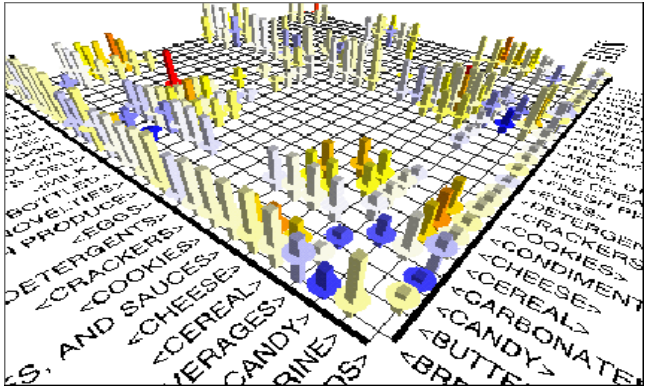
3. Trực quan KTDL

- **Visual Data Mining**



Visualization of a
Decision Tree
in SGI/MineSet 3.0

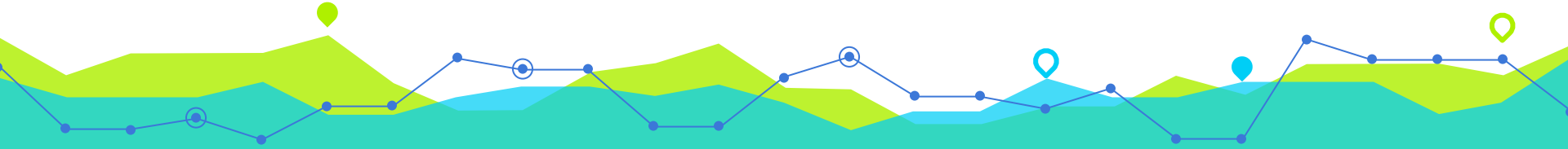
Visualization of
Cluster Grouping
in IBM Intelligent Miner



Visualization of
Association Rules
in SGI/MineSet 3.0

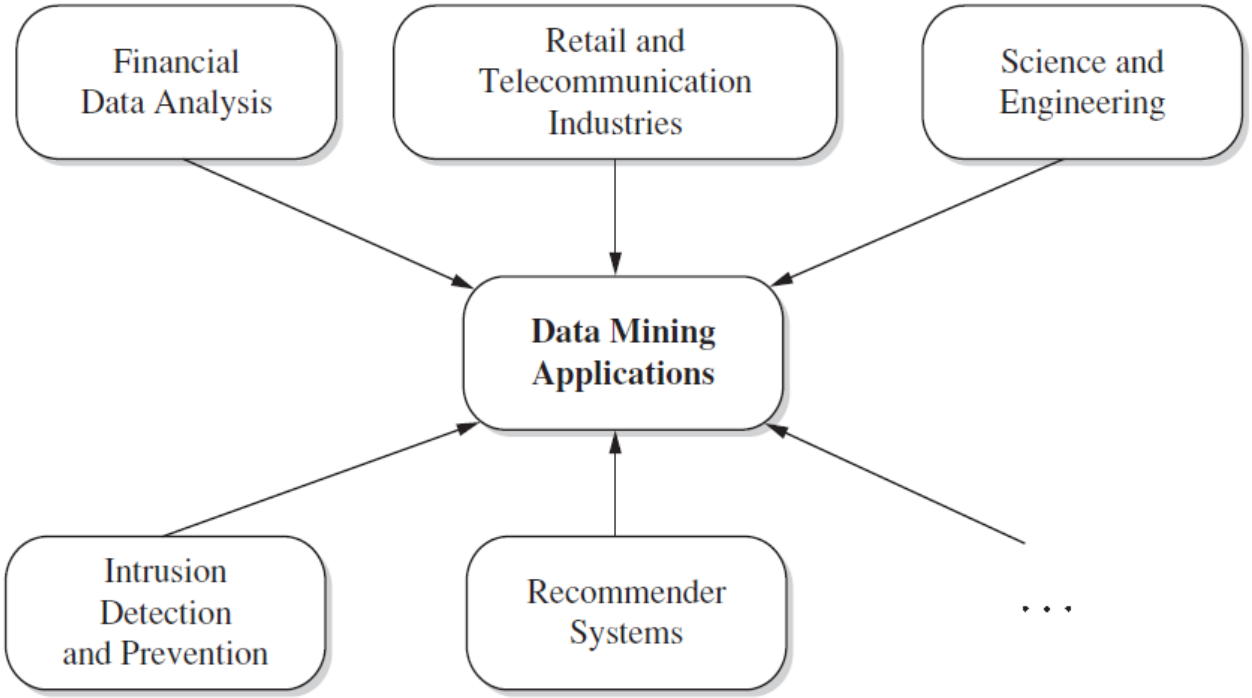
3. Trực quan KTDL

- **Audio Data Mining**
  - Sử dụng tín hiệu âm thanh để biểu thị các mẫu dữ liệu hoặc các đặc trưng của kết quả khai thác dữ liệu
  - Một giải pháp thay thế thú vị cho visual mining
  - Ngược lại với nhiệm vụ KTDL âm thanh (âm nhạc) là tìm các mẫu từ dữ liệu âm thanh.
  - Visual data mining có thể tìm thấy các mẫu thú vị bằng cách sử dụng màn hình đồ họa, nhưng yêu cầu người dùng tập trung vào việc xem các mẫu. Thay vào đó, Audio data mining biến các mẫu thành âm thanh và âm nhạc, đồng thời lắng nghe cao độ, nhịp điệu, giai điệu và giai điệu để xác định điều thú vị hoặc bất thường.



3 Ứng dụng KTDL

Ứng dụng KTDL



1. KTDL trong phân tích dữ liệu tài chính

- Thiết kế, xây dựng kho dữ liệu phục vụ phân tích dữ liệu đa chiều và khai thác dữ liệu
- Dự đoán thanh toán khoản vay/phân tích chính sách tín dụng tiêu dùng
- Phân lớp và gom nhóm khách hàng để tiếp thị mục tiêu
- Phát hiện rửa tiền và các tội phạm tài chính khác

2. KTDL trong ngành bán lẻ và viễn thông

- Ngành bán lẻ: lượng dữ liệu khổng lồ về doanh số bán hàng, lịch sử mua sắm của khách hàng, thương mại điện tử, …
- Các ứng dụng khai thác dữ liệu bán lẻ
  - Xác định hành vi mua của khách hàng
  - Khám phá các mô hình và xu hướng mua sắm của khách hàng
  - Nâng cao chất lượng dịch vụ khách hàng
  - Đạt được sự duy trì và hài lòng của khách hàng tốt hơn
  - Nâng cao tỷ lệ tiêu thụ hàng hóa
  - Thiết kế chính sách vận chuyển và phân phối hàng hóa hiệu quả
  - Phân tích hiệu quả của các chiến dịch bán hàng
  - Giữ chân khách hàng: Phân tích lòng trung thành của khách hàng
  - Phân tích gian lận

3. KTDL trong khoa học và kỹ thuật

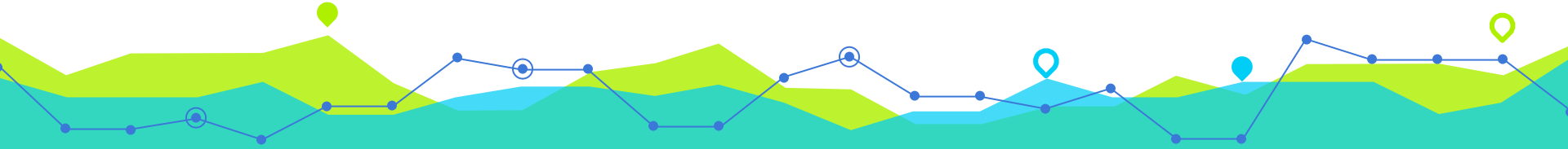
- Kho dữ liệu và tiền xử lý dữ liệu
  - Giải quyết sự không nhất quán hoặc dữ liệu không tương thích được thu thập trong các môi trường đa dạng và các thời kỳ khác nhau (ví dụ: nghiên cứu hệ sinh thái)
- Khai thác các loại dữ liệu phức tạp
  - Không gian, thời gian, sinh học, ngữ nghĩa đa dạng,…
- Khai thác dựa trên đồ thị và dựa trên mạng: Liên kết, mối quan hệ, luồng dữ liệu,..,.
- Khai thác dữ liệu trong khoa học xã hội và nghiên cứu xã hội: văn bản và phương tiện truyền thông xã hội
- Khai thác dữ liệu trong khoa học máy tính: hệ thống giám sát, lỗi phần mềm, xâm nhập mạng

4. KTDL và Hệ thống khuyến nghị

- Hệ khuyến nghị: Cá nhân hóa, đưa ra các đề xuất sản phẩm có khả năng được người dùng quan tâm
- Phương pháp tiếp cận:
  - Content-based: Đề xuất các mục tương tự với các mục mà người dùng ưa thích hoặc truy vấn trong quá khứ
  - Collaborative filtering: Xem xét môi trường xã hội của người dùng, ý kiến của những khách hàng khác có sở thích hoặc sở thích tương tự
  - Kết hợp Content-based và Collaborative filtering

4. KTDL và Hệ thống khuyến nghị

- Khai phá dữ liệu và các hệ khuyến nghị
  - Users C × Item S: trích xuất từ xếp hạng đã biết đến chưa biết để dự đoán kết hợp item - user
  - Memory-based method: sử dụng cách tiếp cận k-nearest neighbor
  - Model-based method: sử dụng tập hợp các xếp hạng để tìm hiểu một mô hình (ví dụ: mô hình xác suất, phân cụm, mạng Bayes, …)
  - Các phương pháp kết hợp tích hợp cả hai để cải thiện hiệu suất



4 Các xu hướng nghiên cứu

Các xu hướng nghiên cứu

- Application exploration: Xử lý các vấn đề dành riêng cho ứng dụng
- Tích hợp khai thác dữ liệu với các công cụ tìm kiếm Web, database, data warehouse, hệ thống cloud computing
- Khai thác mạng xã hội
- Khai thác không gian thời gian, moving objects, cyber-physical systems
- Khai thác dữ liệu multimedia, text, web
- Khai thác dữ liệu biological, biomedical
- Khai thác dữ liệu trong kỹ thuật phần mềm, kỹ thuật hệ thống
- Khai thác dữ liệu âm thanh,
- Khai thác dữ liệu phân tán, dữ liệu thời gian thực
- Khai thác dữ liệu và bảo mật thông tin, quyền riêng tư

THANKS!

Any questions?