IS467: Fundamentals of Data Science

Introduction

Dr. Eli T. Brown
eli.t.brown@depaul.edu
CDM 711 / 312-362-7115

School of Computing, CDM, DePaul University

Chapter 1. Introduction

- Motivation: Why data mining?
- What is data mining?
- Data Mining: On what kind of data?
- Data mining functionality
- Are all the patterns interesting?
- Classification of data mining systems
- Major issues in data mining

With thanks to D. Raicu, our textbook authors, and {Ramakrishnan, Gehrke, Garofalakis, Rastogi}

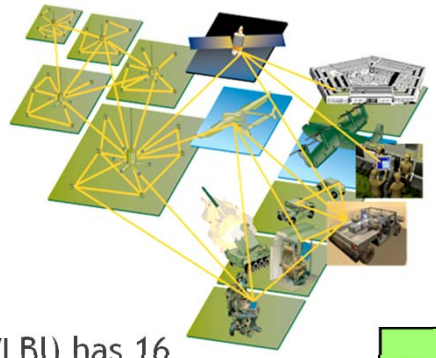
Why Data Mining

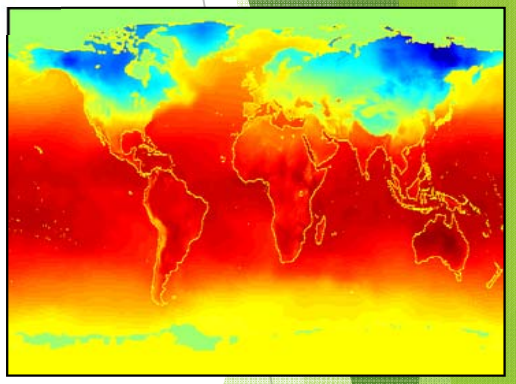
Necessity Is the Mother of Invention!

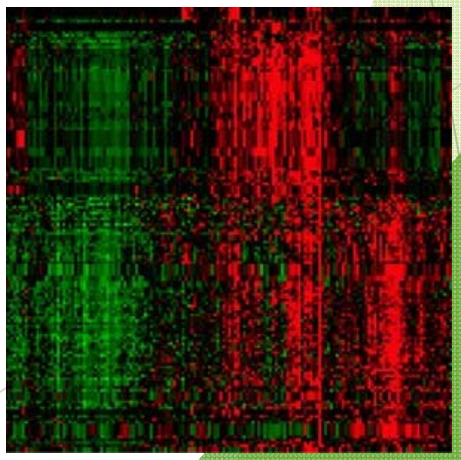
- Automated data collection tools
- Mature database technology
- Tremendous amounts of data accumulated
- No value without analysis

Why Data Mining? Scientific Viewpoint

- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
    - Europe's Very Long Baseline Interferometry (VLBI) has 16 telescopes, each of which produces **1 Gigabit/second** of astronomical data over a 25-day observation session
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data
- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in classifying and segmenting data
  - in Hypothesis Formation







Why Data Mining? Commercial Viewpoint

- Lots of data is being collected and warehoused
  - Web data, e-commerce
    - Alexa internet archive: 7 years of data, 500 TB
    - Google searches 4+ Billion pages, many hundreds TB
    - IBM WebFountain, 160 TB (2003)
    - Internet Archive (www.archive.org),~ 300 TB
  - Purchases at department/ grocery stores
  - Bank/Credit Card transactions

- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

5 million terabytes created in 2002

- UC Berkeley 2003 estimate: 5 exabytes (5 million terabytes) of new data was created in 2002. www.sims.berkeley.edu/research/projects/how-much-info-2003/
- US produces ~40% of new stored data worldwide
- Forbes 20 Facts: http://www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read/#5105cd1f6c1d
- Twice as much information was created in 2002 as in 1999 (~26% growth rate)
- Other growth rate estimates even higher
- Very little data will ever be looked at by a human and there is often information "hidden" in the data that is not readily evident
- Human analysts may take weeks to discover useful information
- Knowledge Discovery is **NEEDED** to make sense and use of data.

Nearly 45-Fold Annual Data Growth by 2020

- In 2009, amid the "Great Recession," the amount of digital information grew 62% over 2008 to 800 billion gigabytes (0.8 Zettabytes). One Zettabyte equals one trillion gigabytes. The amount of digital information created in 2010 (1.2 Zettabytes) will equal:

The digital information created by every man, woman and child on Earth "Tweeting" continuously for 100 years

75 billion fully-loaded 16 GB Apple iPads, which would fill the entire area of Wembley Stadium to the brim 41 times, the Mont Blanc Tunnel 84 times, CERN's Large Hadron Collider tunnel 151 times, Beijing National Stadium 15.5 times or the Taipei 101 Tower 23 times

A full-length episode of FOX TV's hit series "24" running continuously for 125 million years

Despite this growth, the number of IT professionals globally will grow only by a factor of 1.4. (by www.emc.com)

Storage growth

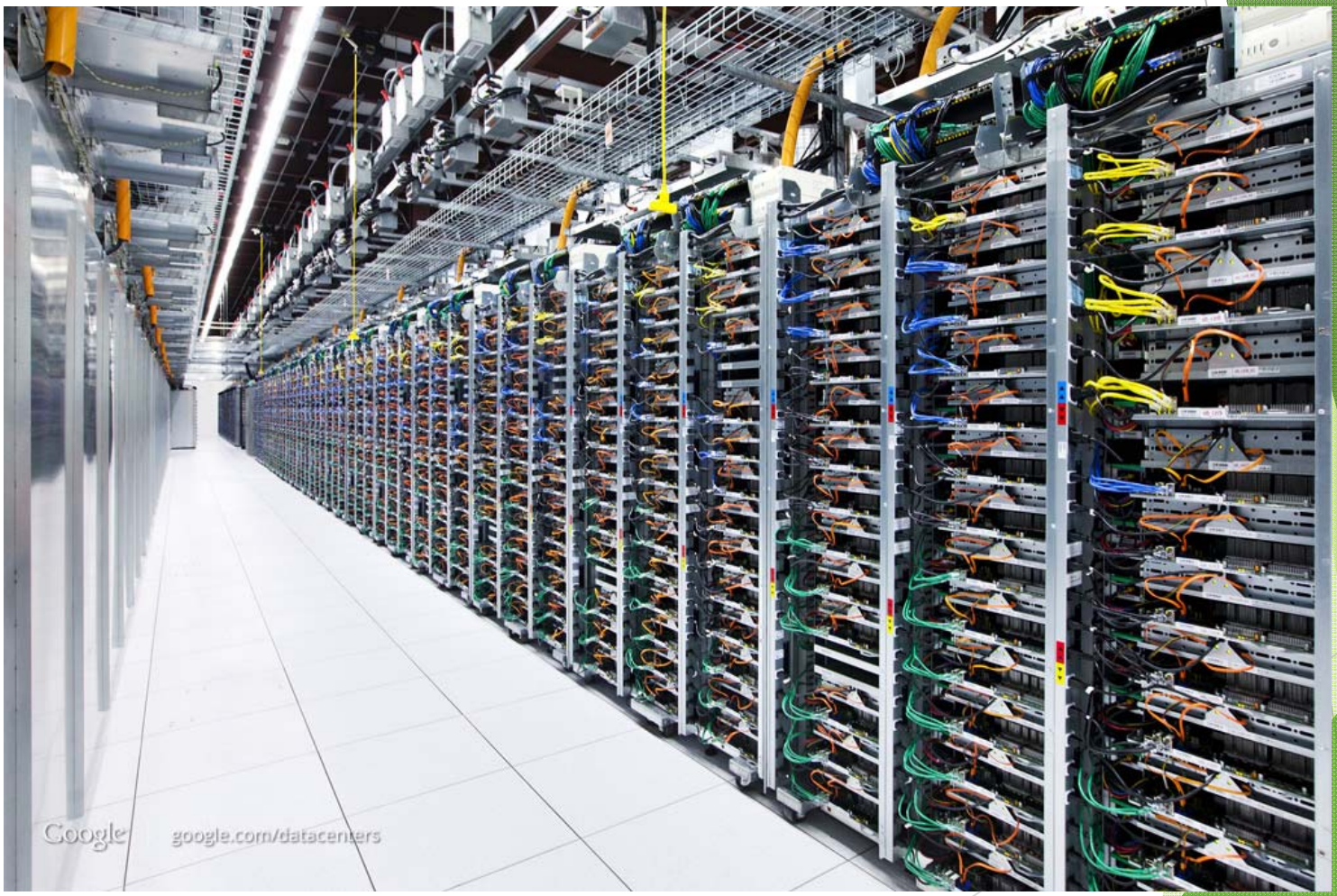
"The Petabyte Age is different because more is different. Kilobytes were stored on floppy disks. Megabytes were stored on hard disks. Terabytes were stored in disk arrays. Petabytes are stored in the cloud. As we moved along that progression, we went from the folder analogy to the file cabinet analogy to the library analogy to — well, at petabytes we ran out of organizational analogies." (Wired, The End of Theory…)

Data mining becomes critical because we cannot possibly learn from that data without it. No human can absorb it.

We see different types of patterns; not models like in physics but correlations and clusters, etc. that work because there is so much data.

Now we Need the Cloud…

Although… not normal computers

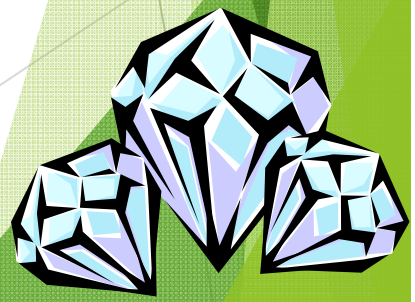


Necessity Is the Mother of Invention

- We are drowning in data, but starving for knowledge!
- Solution: Data warehousing and data mining
  - Data warehousing and on-line analytical processing
  - Mining interesting knowledge (rules, regularities, patterns, constraints) from data in large databases

What Is Data Mining?

- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
  - Data mining: a misnomer?
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Watch out: Is everything "data mining"?
  - (Deductive) query processing.
  - Expert systems or small ML/statistical programs



What is (not) Data Mining?

What is not Data Mining?

– Look up phone number in phone directory

– Query a Web search engine for information about "Amazon"

What is Data Mining?

– Certain names are more prevalent in certain US locations (O'Brien, O'Rurke, O'Reilly... in Boston area)

– Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com)

Why Data Mining?—Potential Applications

- Data analysis and decision support
  - Market analysis and management
    - Target marketing, customer relationship management (CRM), market basket analysis, cross selling, market segmentation
  - Risk analysis and management
    - Forecasting, customer retention, improved underwriting, quality control, competitive analysis
  - Fraud detection and detection of unusual patterns (outliers)
    - Auto insurance: ring of collisions
    - Money laundering: suspicious monetary transactions
    - Medical insurance
      - Professional patients, ring of doctors, and ring of references
      - Unnecessary or correlated screening tests
    - Telecommunications: phone-call fraud
      - Phone call model: destination of the call, duration, time of day or week. Analyze patterns that deviate from an expected norm
    - Retail industry
      - Analysts estimate that 38% of retail shrink is due to dishonest employees

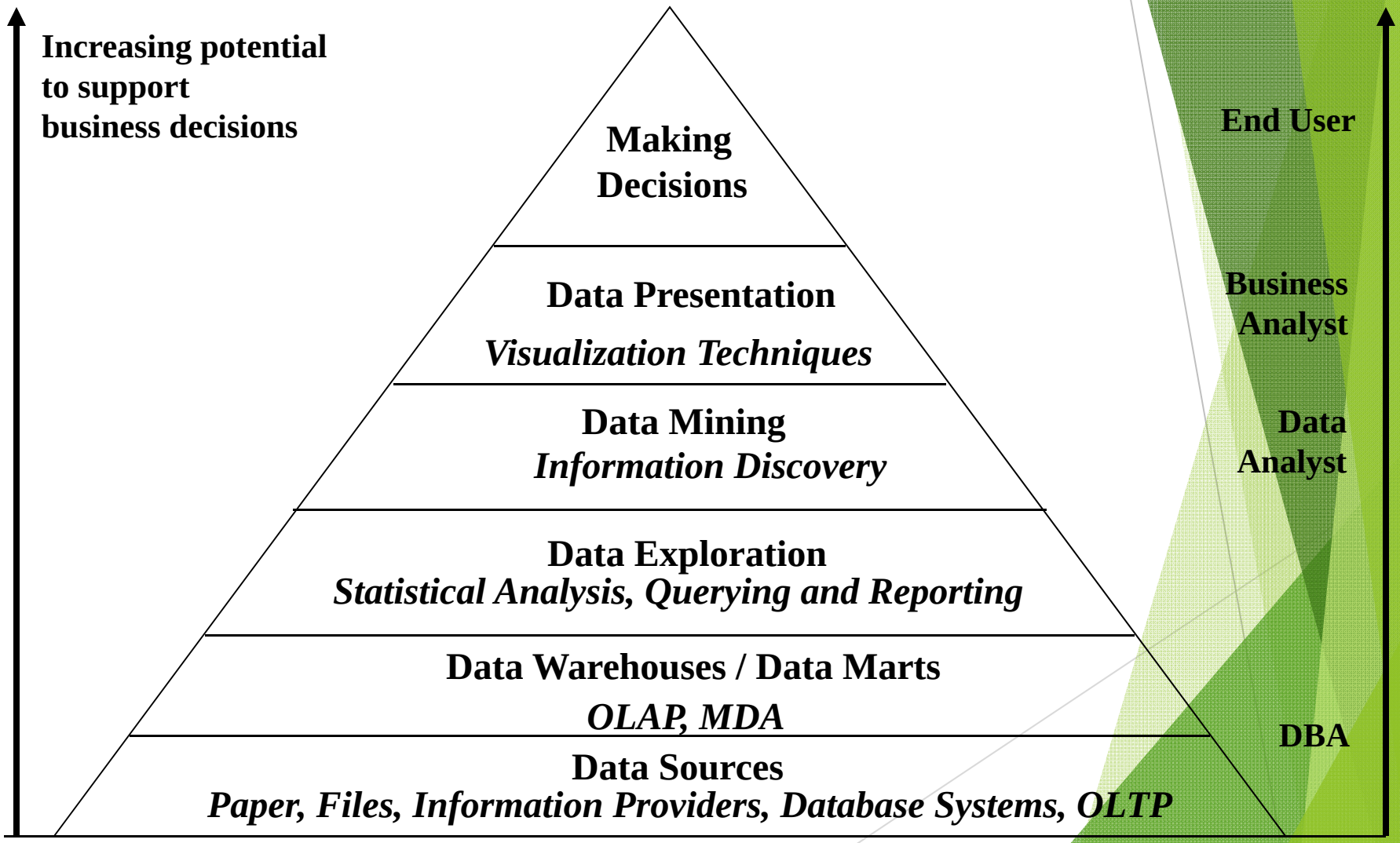
Why Data Mining?—Potential Applications

- Sports
  - IBM Advanced Scout analyzed NBA game statistics (shots blocked, assists, and fouls) to gain competitive advantage for New York Knicks and Miami Heat
- Astronomy
  - JPL and the Palomar Observatory discovered 22 quasars with the help of data mining
- Internet Web Surf-Aid
  - IBM Surf-Aid applies data mining algorithms to Web access logs for market-related pages to discover customer preference and behavior pages, analyzing effectiveness of Web marketing, improving Web site organization, etc.

Applications of Data Mining

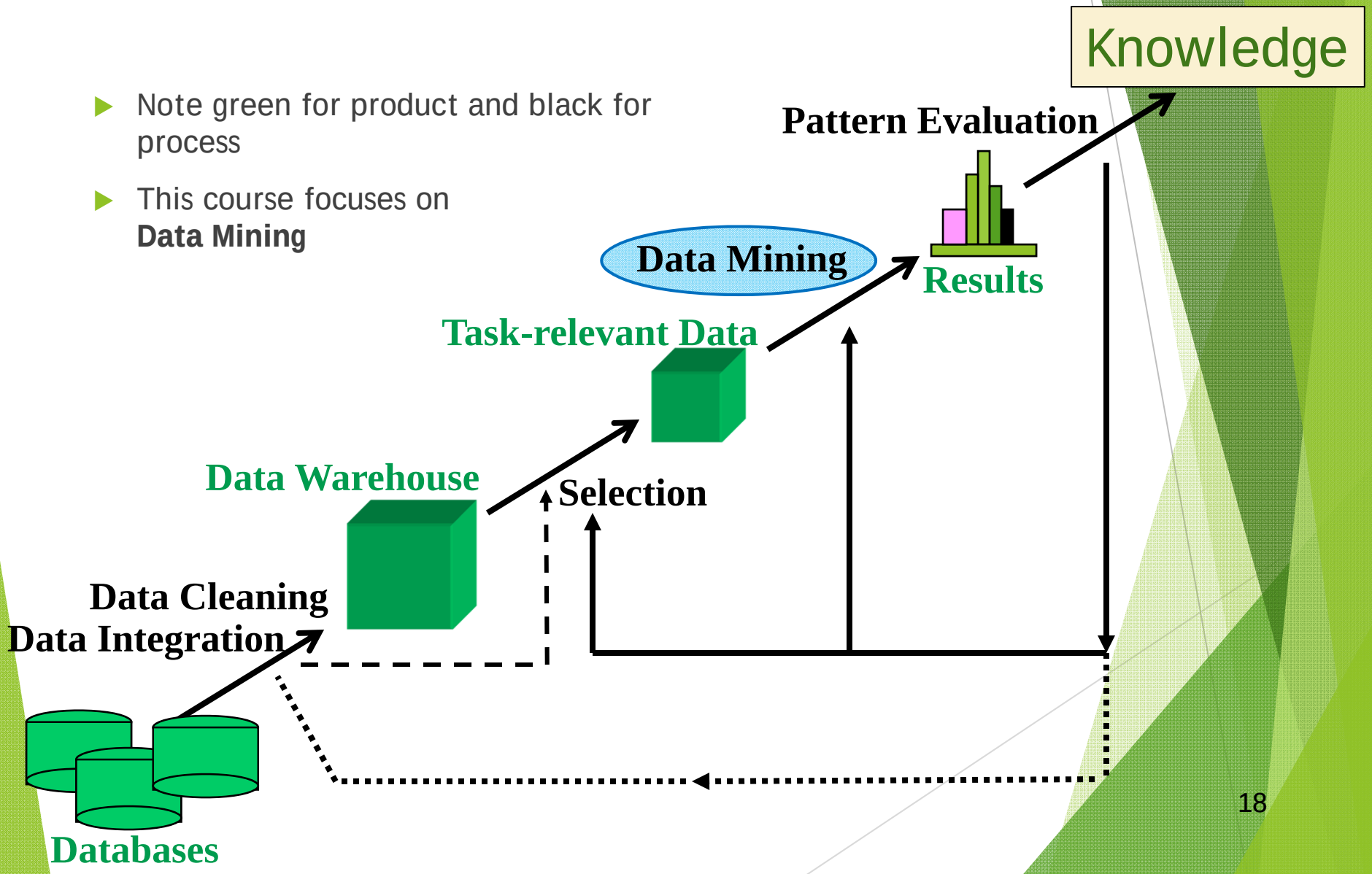
- Web page analysis: classification, clustering, ran
- Collaborative analysis & recommender systems
- Marketing funnel data analysis for targeted mark
- Biological and medical data analysis
- Data mining and software engineering
- Data mining and text analysis
- Data mining and social and information network analysis
- Built-in (invisible data mining) functions in Google, MS, Yahoo!, Linked, Facebook, …
- Major dedicated data mining systems/tools
  - SAS, MS SQL-Server Analysis Manager, Oracle Data Mining Tools)

Data Mining and Business Intelligence

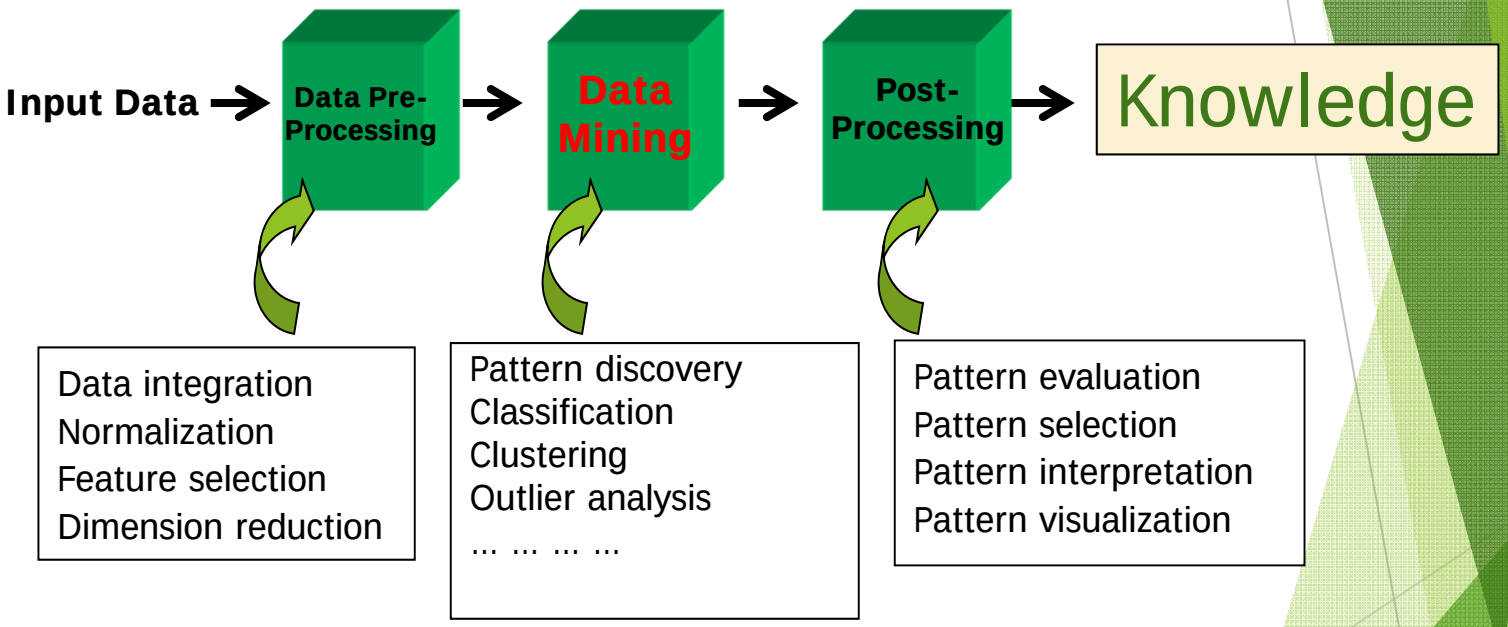


Data Mining: A KDD Process

- Note green for product and black for process
- This course focuses on **Data Mining**



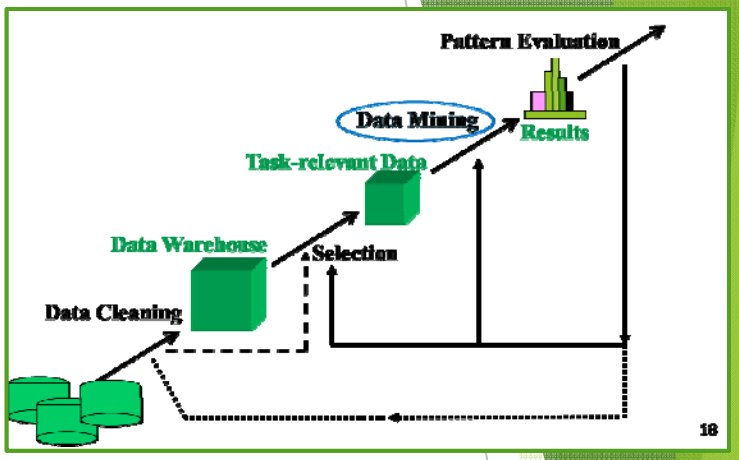
KDD Process: from ML and Statistics



- This is a view from typical machine learning and statistics communities

Steps of a KDD Process

- Learning the application domain
  - relevant prior knowledge and goals of application
- Creating a target data set: data collection
- Data cleaning and preprocessing **(may take 60% of effort!)**
- Data reduction and transformation
  - Find useful features, dimensionality/variable reduction
- Choosing functions of data mining
  - summarization, classification, regression, association, clustering.
- Choosing the mining algorithm(s)
- Data mining: search for patterns of interest
- Pattern evaluation and knowledge presentation
  - visualization, transformation, removing redundant patterns, etc.
- Use of discovered knowledge



Are results all interesting?

No.

What results are ‘interesting’?

Data mining is the exploration and analysis of large quantities of data in order to discover valid, novel, potentially useful, and ultimately understandable patterns in data.

Valid: The patterns hold in general (with some *certainty*).

Novel: We did not know the pattern beforehand.

Useful: We can devise actions from the patterns.

Understandable: We can interpret and comprehend the patterns.

Data Mining: On What Kinds of Data?

- Relational database
- Data warehouse
- Transactional database
- Advanced database and information repository
  - Object-relational database
  - Spatial and temporal data
  - Time-series data
  - Stream data
  - Multimedia database
  - Heterogeneous and legacy database
  - Text databases & WWW

Data Mining Functionalities

- Concept description: Characterization and discrimination
  - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet regions
  - Clustering and anomaly detection
- Association (correlation and causality)
  - Bought Diapers → Buys Beer [0.5%, 75%]
- Classification and Prediction
  - Construct models (functions) that describe and distinguish classes or concepts for future prediction
    - E.g., classify countries based on climate, or classify cars based on gas mileage
  - Presentation: decision-tree, classification rule, neural network
  - Predict some unknown or missing numerical values

Data Mining Functionalities (2)

- Cluster analysis
  - Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
  - Maximizing intra-class similarity & minimizing interclass similarity
- Outlier analysis
  - Outlier: a data object that does not comply with the general behavior of the data
  - Noise or exception? No! useful in fraud detection, rare events analysis
- Trend and evolution analysis
  - Trend and deviation: regression analysis
  - Sequential pattern mining, periodicity analysis
  - Similarity-based analysis
- Other pattern-directed or statistical analyses

Data Mining Functionalities

- Several well-studied tasks we'll cover
  - Classification
  - Clustering
  - Pattern mining
- Many methods (algorithms) proposed for each

Classification

Goal:

Learn a function that assigns a record to one of several predefined classes.

Requirements on the model:

- High accuracy
- Understandable by humans, interpretable
- Fast construction for very large training databases

Classification (Contd.)

Approaches include:

- *Decision trees*
- Linear Discriminant Analysis
- *k-nearest neighbor methods*
- Logistic regression
- Neural networks
- Support Vector Machines

Classification Example

- Two predictor attributes:
  Age and Car-type (**S**port, **M**inivan and **T**ruck)
- Age is ordered, Car-type is categorical attribute
- Class label indicates whether person bought product
- Dependent attribute is *categorical*
- *Goal: automatically learn function from attributes to class for prediction*

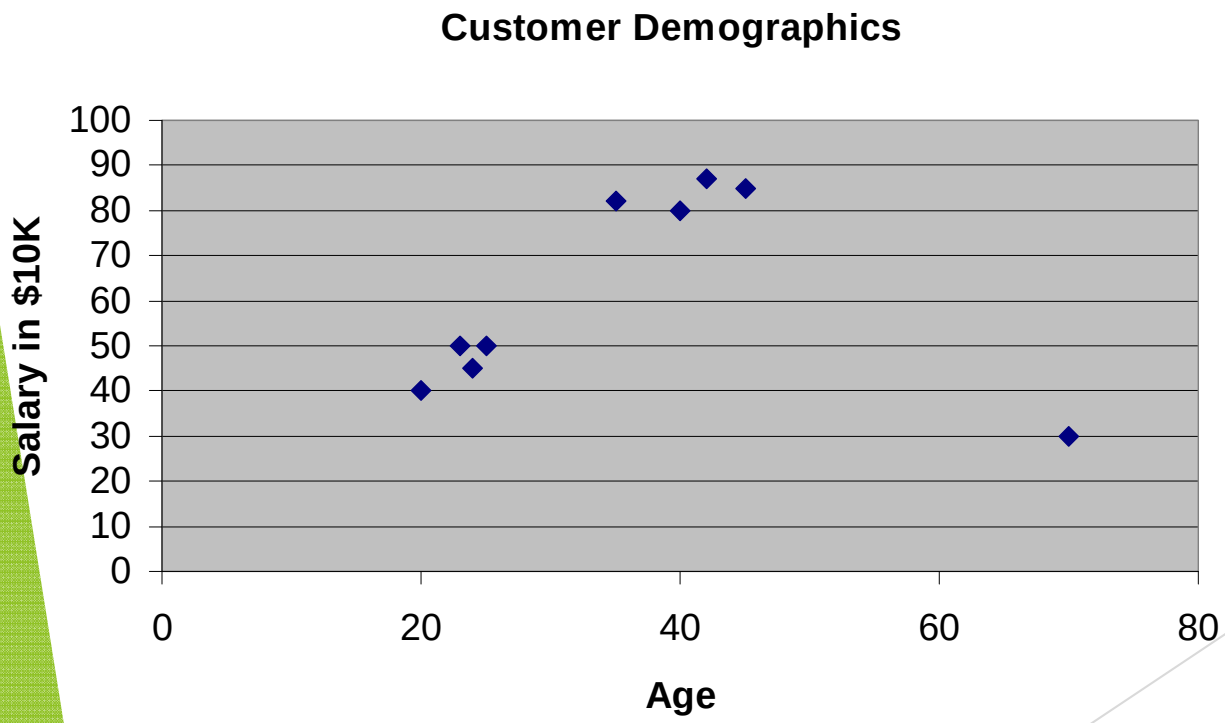
Age	Car	Class
20	M	Yes
30	M	Yes
25	T	No
30	S	Yes
40	S	Yes
20	T	No
30	M	Yes
25	M	Yes
40	M	Yes
20	S	No

Clustering

- Output: (k) groups of records called clusters, such that the records within a group are more similar to records in other groups
  - Representative points for each cluster
  - Labeling of each record with each cluster number
  - Other description of each cluster
- *This is unsupervised learning*: No record labels are given to learn from
- Usage:
  - Exploratory data mining
  - Preprocessing step (e.g., outlier detection)

Clustering (Contd.)

- Example input database: Two numerical variables
- How many groups are here?



Age	Salary
20	40
25	50
24	45
23	50
40	80
45	85
42	87
35	82
70	30

Cluster Representatives

A representative set of points:

- Small in number
- Distributed over the cluster
- Each point in cluster is close to one representative
- Can use representatives to get an easier to see picture of large data
  - Summary of data points by the representatives of each cluster
  - Difference (distance) between clusters can be distance just between their representatives; much faster to deal with

Example Application: Fraud Detection

- Industries: Health care, retail, credit card services, telecom, B2B relationships
- Approach:
  - Use historical data to build models of fraudulent behavior
  - Deploy models to identify fraudulent instances

Fraud Detection (Contd.)

- Examples:
  - Auto insurance: Detect groups of people who stage accidents to collect insurance
  - Medical insurance: Fraudulent claims
  - Money laundering: Detect suspicious money transactions (US Treasury's Financial Crimes Enforcement Network)
  - Telecom industry: Find calling patterns that deviate from a norm (origin and destination of the call, duration, time of day, day of week).

Fraud Detection (Contd.)

- Fraud Detection
  - Goal: Predict fraudulent cases in credit card transactions.
  - Approach:
    - Use credit card transactions and the information on its account-holder as attributes.
      - When does a customer buy, what does he buy, how often he pays on time, etc
    - Label past transactions as fraud or fair transactions. This forms the class attribute.
    - Learn a model for the class of the transactions.
    - Use this model to detect fraud by observing credit card transactions on an account.

Clustering of S&P 500 Stock Data

- Observe Stock Movements every day.
- Clustering points: Stock-{UP/DOWN}
- Similarity Measure: Two points are more similar if the events described by them frequently happen together on the same day.
  - We used association rules to quantify a similarity measure.

	Discovered Clusters	***Industry Group***
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP

Classification: Application 2

- Customer Attrition/Churn:
  - Goal: To predict whether a customer is likely to be lost to a competitor.
  - Approach:
    - Use detailed record of transactions with each of the past and present customers, to find attributes.
      - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
    - Label the customers as loyal or disloyal.
    - Find a model for loyalty.

Example: Opinion Analysis

Word-of-mouth on the Web

- The Web has dramatically changed the way that consumers express their opinions.
- One can post reviews of products at merchant sites, Web forums, discussion groups, blogs
- Techniques are being developed to exploit these sources.
- Benefits of Review Analysis
  - Potential Customer: No need to read many reviews
  - Product manufacturer: market intelligence, product benchmarking

Feature Based Analysis & Summarization

- Extracting product features (called Opinion Features) that have been commented on by customers.
- Identifying opinion sentences in each review and deciding whether each opinion sentence is positive or negative.
- Summarizing and comparing results.

An example

GREAT Camera., Jun 3, 2004

Reviewer: **jprice174** from Atlanta, Ga.

I did a lot of research last year before I bought this camera… It kinda hurt to leave behind my beloved nikon 35mm SLR, but I was going to Italy, and I needed something smaller, and digital.

The pictures coming out of this camera are amazing. The 'auto' feature takes great pictures most of the time. And with digital, you're not wasting film if the picture doesn't come out. …

….

Summary:

Feature1: **picture**

Positive: 12

- The pictures coming out of this camera are amazing.
- Overall this is a good camera with a really good picture clarity.

…

Negative: 2

- The pictures come out hazy if your hands shake even for a moment during the entire process of taking a picture.
- Focusing on a display rack about 20 feet away in a brightly lit room during day time, pictures produced by this camera were blurry and in a shade of orange.

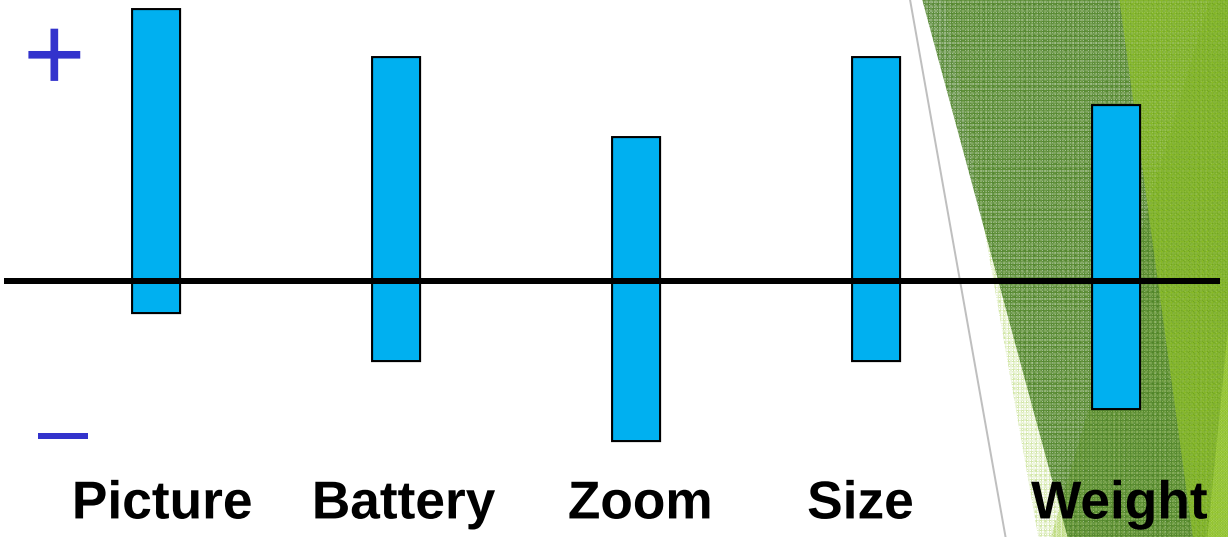
Feature2: **battery life**

…

Visual Comparison

- Summary of reviews of

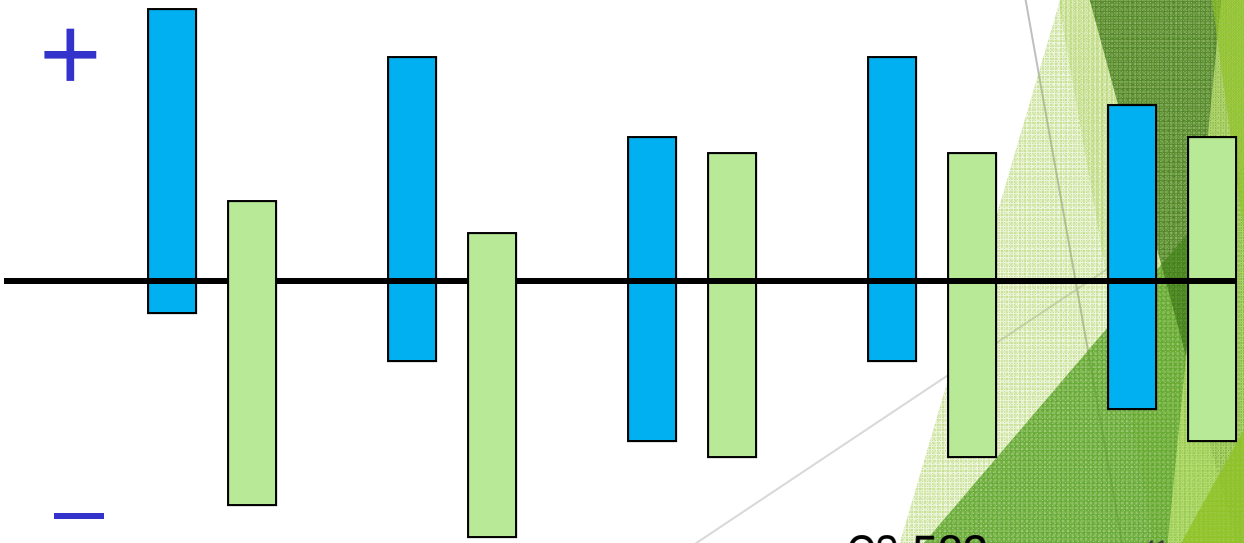
Digital camera 1



- Comparison of reviews of

Digital camera 1

Digital camera 2



Sequential Pattern Discovery: Examples

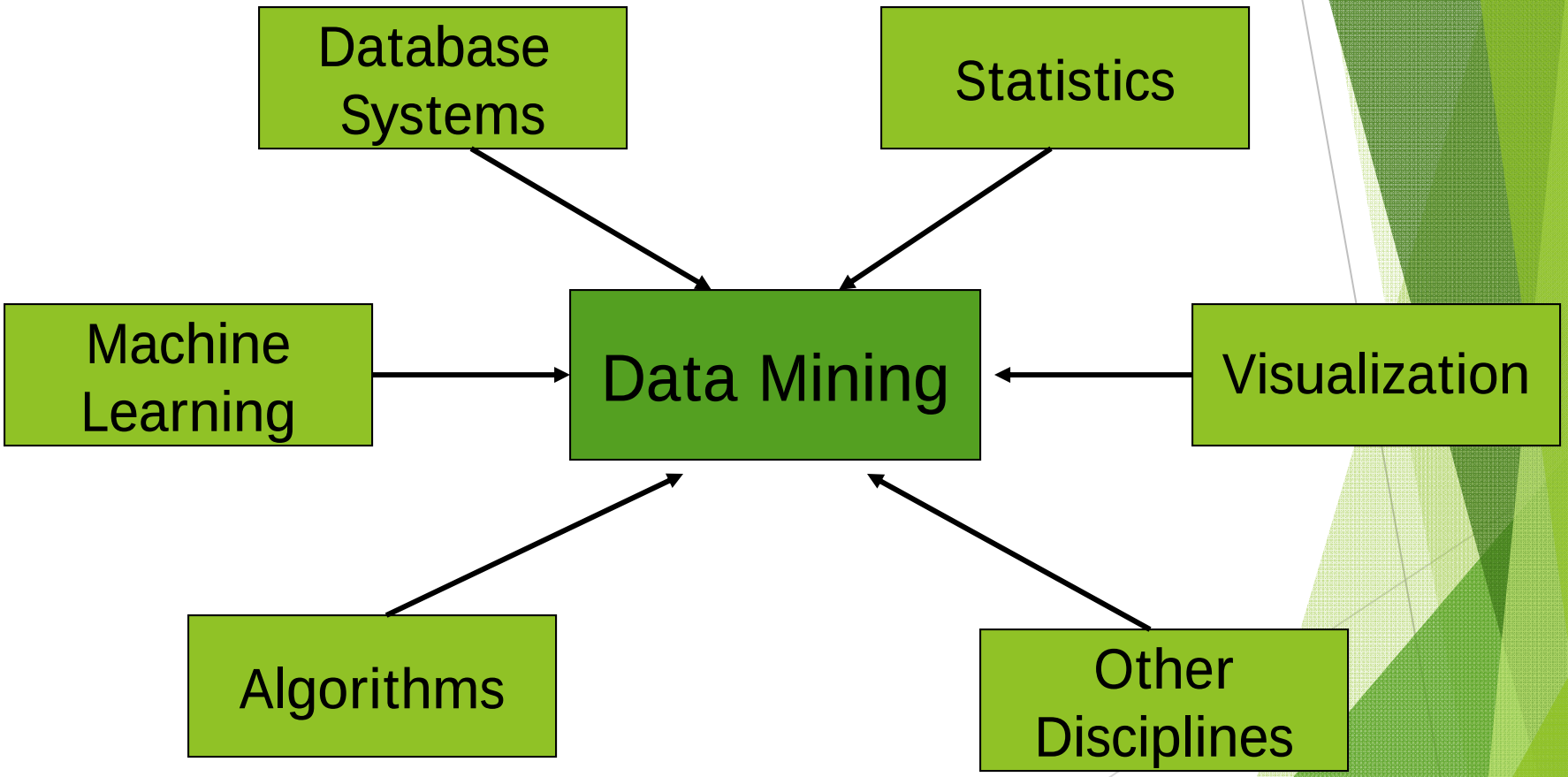
- In telecommunications alarm logs,
  - (Inverter_Problem Excessive_Line_Current)

    (Rectifier_Alarm) --> (Fire_Alarm)
- In point-of-sale transaction sequences,
  - Computer Bookstore:

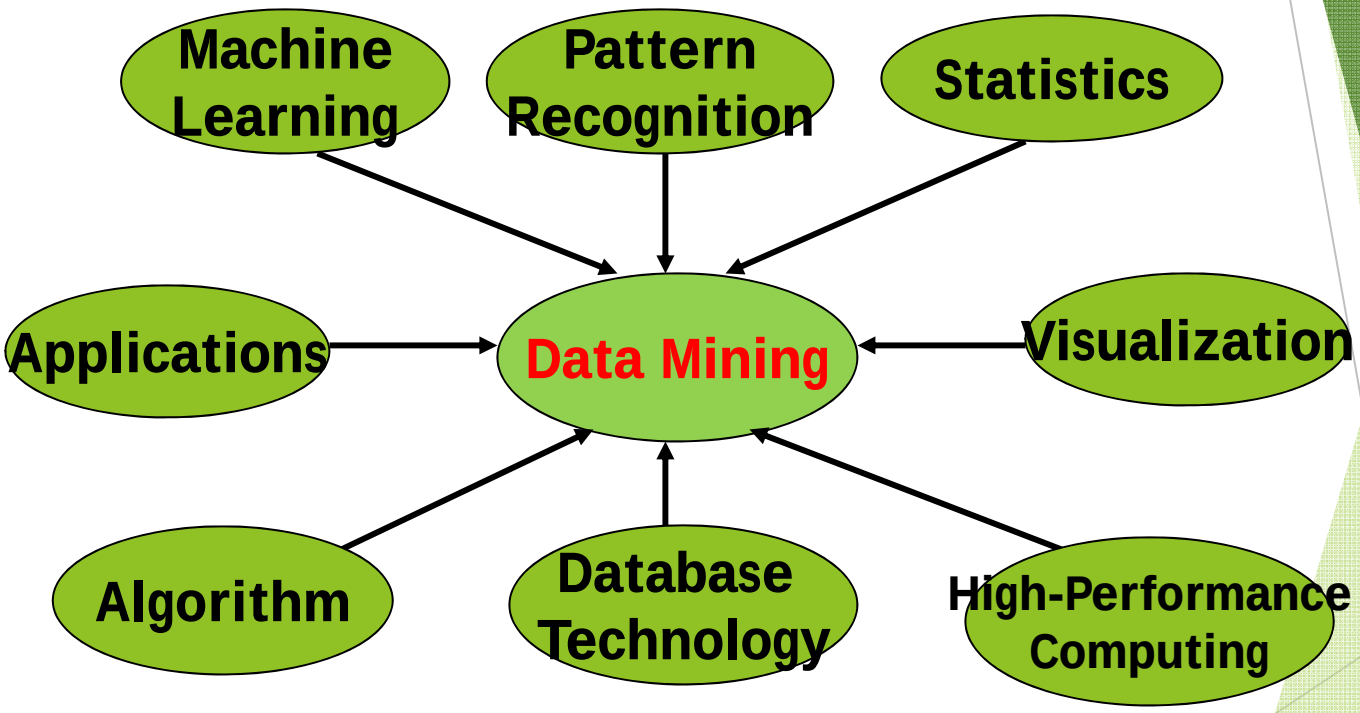
    (Intro_To_Visual_C) (C++_Primer) -->
    (Perl_for_dummies,Tcl_Tk)
  - Athletic Apparel Store:

    (Shoes) (Racket, Racketball) --> (Sports_Jacket)

Data Mining: Confluence of Multiple Disciplines



Data Mining: Confluence of Multiple Disciplines



Statistics Machine Learning and Data Mining

- Statistics
  - More theory-based
  - More focused on testing hypotheses, building models
  - Mathematical soundness behind mining methods
- Machine Learning
  - More heuristic
  - Focused on improving predictive performance (effectiveness)
  - Also looks at real-time learning and robotics – not part of data mining
  - Developed from Artificial Intelligence, but the two are different now
- Data Mining and Knowledge Discovery
  - Integrates theory and heuristics
  - Focus on the entire process of knowledge discovery (KDD), including data integration and cleaning, machine learning and pattern recognition, and visualization of results

Data Mining: Classification Schemes

- General functionality
  - Descriptive data mining
  - Predictive data mining
- Different views, different classifications
  - Kinds of data to be mined
  - Kinds of knowledge to be discovered
  - Kinds of techniques utilized
  - Kinds of applications adapted

Major Issues in Data Mining

- Mining methodology
  - Mining different kinds of knowledge from diverse data types, e.g., bio, stream, Web
  - Performance: efficiency, effectiveness, and scalability
  - Pattern evaluation: the interestingness problem
  - Incorporation of background knowledge
  - Handling noise and incomplete data
  - Size! - Parallel, distributed and incremental mining methods
  - Integration of the discovered knowledge with existing one: knowledge fusion
- User interaction
  - Data mining query languages and ad-hoc mining
  - Expression and visualization of data mining results
  - Interactive mining of knowledge at multiple levels of abstraction
- Applications and social impacts
  - Domain-specific data mining & invisible data mining
  - Protection of data security, integrity, and privacy

Pitfalls in Data Mining

- With enough data and time can find anything
- Working with a United Nations data set: butter production in Bangladesh is the single best predictor of the Standard & Poor's 500-stock index
- Correlation: a win by the NFC team in the Super Bowl implies a rise in stock prices.
- Four Pitfalls
  - Develop theory to fit an oddity, though it might be pure chance
  - Find evidence to support any preconception
  - Storytelling - finding would make more sense if plausible theory for it but beguiling story can disguise weakness in data
  - Too many variables => too likely to find relationships that aren't real

Summary

- Data mining: discovering interesting patterns from large amounts of data
- A natural evolution of database technology, in great demand, with wide applications
- A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation
- Mining can be performed in a variety of information repositories
- Data mining functionalities: characterization, discrimination, association, classification, clustering, outlier and trend analysis, etc.
- Data mining systems and architectures
- Major issues in data mining

Advice to new user of data mining

- Planning and prep are essential; not just having lots of data
- Business need over technical excitement. Make sure to gather data needed for specific, unambiguous problem and know the end-user who needs result and possible sources
- Data Preparation is crucial - up to 80% of time spent
- Collection of methodologies; only way to know is try them
- Keep end users involved (so final results will be properly targeted)
- Iterative process

A Brief History of Data Mining Society

- 1989 IJCAI Workshop on Knowledge Discovery in Databases
  - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
  - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
  - Journal of Data Mining and Knowledge Discovery (1997)
- ACM SIGKDD conferences since 1998 and SIGKDD Explorations
- More conferences on data mining
  - PAKDD (1997), PKDD (1997), SIAM-Data Mining (2001), (IEEE) ICDM (2001), WSDM (2008), etc.
- ACM Transactions on KDD (2007)

Conferences and Journals on Data Mining

- KDD Conferences
  - ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining (KDD)
  - SIAM Data Mining Conf. (SDM)
  - (IEEE) Int. Conf. on Data Mining (ICDM)
  - European Conf. on Machine Learning and Principles and practices of Knowledge Discovery and Data Mining (ECML-PKDD)
  - Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)
  - Int. Conf. on Web Search and Data Mining (WSDM)
- Other related conferences
  - DB conferences: ACM SIGMOD, VLDB, ICDE, EDBT, ICDT, …
  - Web and IR conferences: WWW, SIGIR, WSDM
  - ML conferences: ICML, NIPS
  - PR conferences: CVPR,
- Journals
  - Data Mining and Knowledge Discovery (DAMI or DMKD)
  - IEEE Trans. On Knowledge and Data Eng. (TKDE)
  - KDD Explorations
  - ACM Trans. on KDD

Where to Find References? DBLP, CiteSeer, Google

- Data mining and KDD (SIGKDD)
  - Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
  - Journal: Data Mining and Knowledge Discovery, KDD Explorations, ACM TKDD
- Database systems (SIGMOD)
  - Conferences: ACM-SIGMOD, ACM-PODS, VLDB, IEEE-ICDE, EDBT, ICDT, DASFAA
  - Journals: IEEE-TKDE, ACM-TODS/TOIS, JIIS, J. ACM, VLDB J., Info. Sys., etc.
- AI & Machine Learning
  - Conferences: Machine learning (ML), AAAI, IJCAI, COLT (Learning Theory), CVPR, NIPS, etc.
  - Journals: Machine Learning, Artificial Intelligence, Knowledge and Information Systems, IEEE-PAMI, etc.
- Web and IR
  - Conferences: SIGIR, WWW, CIKM, etc.
  - Journals: WWW: Internet and Web Information Systems,
- Statistics
  - Conferences: Joint Stat. Meeting, etc.
  - Journals: Annals of statistics, etc.
- Visualization
  - Conference proceedings: CHI, ACM-SIGGraph, etc.
  - Journals: IEEE Trans. visualization and computer graphics, etc.

Recommended Reference Books

- **Foster Provost & Tom Fawcett, Data Science for Business: What you need to know about data mining and data-analytic thinking, 2013.**
- R. Agrawal, J. Han, and H. Mannila, Readings in Data Mining: A Database Perspective, Morgan Kaufmann
- U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996
- U. Fayyad, G. Grinstein, and A. Wierse, Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2001
- D. J. Hand, H. Mannila, and P. Smyth, Principles of Data Mining, MIT Press, 2001
- T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer-Verlag, 2001
- T. M. Mitchell, Machine Learning, McGraw Hill, 1997
- G. Piatetsky-Shapiro and W. J. Frawley. Knowledge Discovery in Databases. AAAI/MIT Press, 1991
- S. M. Weiss and N. Indurkhya, Predictive Data Mining, Morgan Kaufmann, 1998
- I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2001

Case Study: Bank

- Business goal: Sell more home equity loans
- Current models:
  - Customers with college-age children use home equity loans to pay for tuition
  - Customers with variable income use home equity loans to even out stream of income
- Data:
  - Large data warehouse
  - Consolidates data from 42 operational data sources

Case Study: Bank (Contd.)

1. Select subset of customer records who have received home equity loan offer
   - Customers who declined
   - Customers who signed up

Income	Number of Children	Average Checking Account Balance	…	Reponse
$40,000	2	$1500		Yes
$75,000	0	$5000		No
$50,000	1	$3000		No
…	…	…	…	…

Case Study: Bank (Contd.)

2. Find rules to predict whether a customer would respond to home equity loan offer

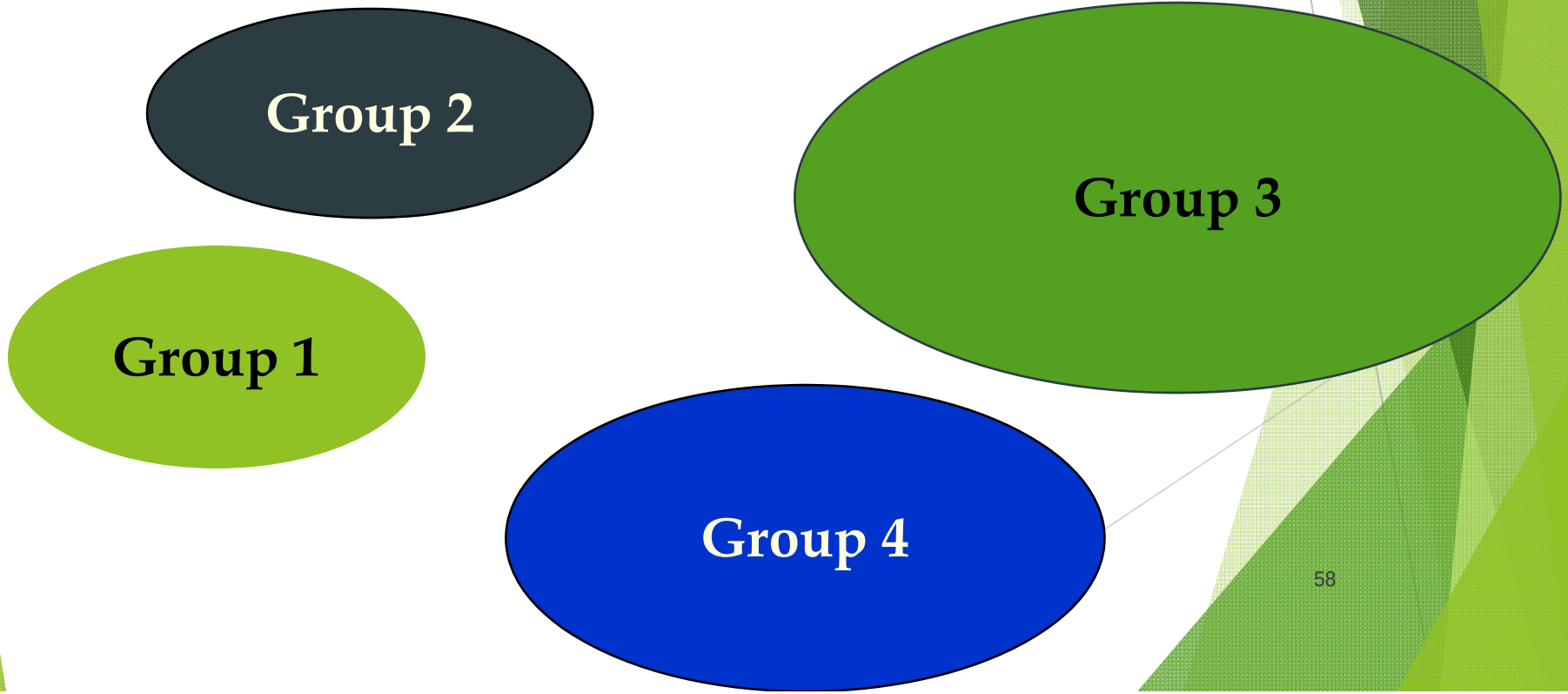
IF (Salary < 40k) and
(numChildren > 0) and
(ageChild1 > 18 and ageChild1 < 22)

THEN YES

...

Case Study: Bank (Contd.)

3. Group customers into clusters and investigate clusters



Case Study: Bank (Contd.)

4. Evaluate results:
   - Many “uninteresting” clusters
   - One interesting cluster! Customers with both business and personal accounts; unusually high percentage of likely respondents

Example: Bank (Contd.)

Action:

- New marketing campaign

Result:

- Acceptance rate for home equity offers more than doubled

Market Basket Analysis

- Consider shopping cart filled with several items
- Market basket analysis tries to answer the following questions:
  - Who makes purchases
  - What do customers buy

Market Basket Analysis

- Given:
  - A database of customer transactions
  - Each transaction is a set of items
- Goal:
  - Extract rules

TID	CID	Date	Item	Qty
111	201	5/1/99	Pen	2
111	201	5/1/99	Ink	1
111	201	5/1/99	Milk	3
111	201	5/1/99	Juice	6
112	105	6/3/99	Pen	1
112	105	6/3/99	Ink	1
112	105	6/3/99	Milk	1
113	106	6/5/99	Pen	1
113	106	6/5/99	Milk	1
114	201	7/1/99	Pen	2
114	201	7/1/99	Ink	2
114	201	7/1/99	Juice	4

Market Basket Analysis (Contd.)

- Co-occurrences
  - 80% of all customers purchase items X, Y and Z together.
- Association rules
  - 60% of all customers who purchase X and Y also buy Z.
- Sequential patterns
  - 60% of customers who first buy X also purchase Y within three weeks.

Confidence and Support

We prune the set of all possible association rules using two interestingness measures:

- Confidence of a rule:
  - X => Y has confidence c if P(Y|X) = c
- Support of a rule:
  - X => Y has support s if P(XY) = s

We can also define

- Support of a co-ocurrence XY:
  - XY has support s if P(XY) = s

Example

- Example rule:
  {Pen} => {Milk}
  Support: 75%
  Confidence: 75%

- Another example:
  {Ink} => {Pen}
  Support: 100%
  Confidence: 100%

TID	CID	Date	Item	Qty
111	201	5/1/99	Pen	2
111	201	5/1/99	Ink	1
111	201	5/1/99	Milk	3
111	201	5/1/99	Juice	6
112	105	6/3/99	Pen	1
112	105	6/3/99	Ink	1
112	105	6/3/99	Milk	1
113	106	6/5/99	Pen	1
113	106	6/5/99	Milk	1
114	201	7/1/99	Pen	2
114	201	7/1/99	Ink	2
114	201	7/1/99	Juice	4

Extensions

- Imposing constraints
  - Only find rules involving the dairy department
  - Only find rules involving expensive products
  - Only find rules with “whiskey” on the right hand side
  - Only find rules with “milk” on the left hand side
  - Hierarchies on the items
  - Calendars (every Sunday, every 1st of the month)

Market Basket Analysis: Applications

- Sample Applications
  - Direct marketing
  - Fraud detection for medical insurance
  - Floor/shelf planning
  - Web site layout
  - Cross-selling

Motivating Example

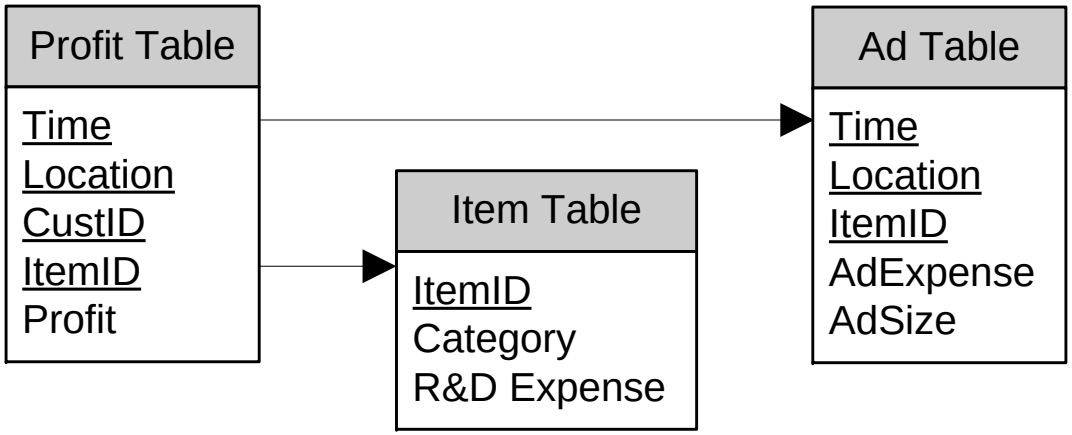
- A company wants to predict the first year worldwide profit of a new item (e.g., a new movie)
  - By looking at features and profits of previous (similar) movies, we predict expected total profit (1-year US sales) for new movie
    - Wait a year and write a query! If you can't wait, stay awake …
  - The most predictive "features" may be based on sales data gathered by releasing the new movie in many "regions" (different locations over different time periods).
    - Example "region-based" features: 1st week sales in Peoria, week-to-week sales growth in Wisconsin, etc.
    - Gathering this data has a cost (e.g., marketing expenses, waiting time)
- **Problem statement:** Find the most predictive region features that can be obtained within a given "cost budget"

Key Ideas

- Large datasets are rarely labeled with the targets that we wish to learn to predict
  - But for the tasks we address, we can readily use database queries to generate features (e.g., 1st week sales in Peoria) and even **targets** (e.g., profit) for mining
- We use data-mining models as building blocks in the mining process, rather than thinking of them as the end result
  - The central problem is to find data subsets **("bellwether regions")** that lead to predictive features which can be gathered at low cost for a new case

Example

- A company wants to predict the first year's worldwide profit for a new item, by using its historical database
- Database Schema:

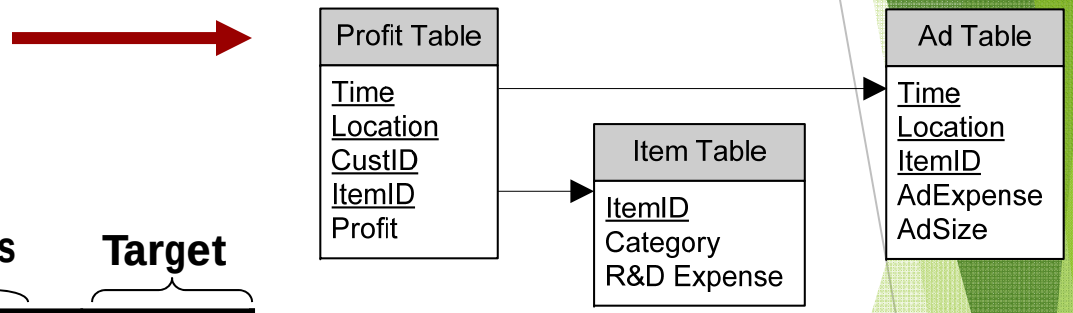


- The combination of the underlined attributes forms a key

A Straightforward Approach

By joining and aggregating tables
in the historical database
we can create a training set:

- Build a regression model to predict item profit



Item-table features **Target**

ItemID	Category	R&D Expense	Profit
1	Laptop	500K	12,000K
2	Desktop	100K	8,000K
…	…	…	…

An Example regression model:

$$\text{Profit} = \beta_0 + \beta_1 \text{ Laptop} + \beta_2 \text{ Desktop} + \beta_3 \text{ RdExpense}$$

- There is much room for accuracy improvement!