KDD and Data Mining and More

Presented by...

Susan Imberman

Imberman@mail.csi.cuny.edu

What Is KDD?

- Knowledge discovery in databases
- Synonymous with large databases
- Automated discovery of patterns and relationships

Why KDD?

- Large databases are not uncommon
  - Point of sale info, government records, medical records, and credit card data
  - Scientific instruments can produce terabytes and petabytes at rates of gigs per hour
  - Storage capabilities better. Cheaper, larger
- Databases growing in field size (10^2 or 10^3)
- Databases growing in record size (10^9)
- Human limits

How big is big?

- Data Mining deals with terabytes plus amounts of data
- Today, petabytes are not unusual
- How big is a petabyte?
  - 250 billion pages of text
  - 20 million 4 drawer file cabinets
  - 2,000 mile high tower of 1 billion diskettes
    - Mount McKinley, the tallest mountain in North America, is about 4 miles high

The size of the Terror-Bite

KDD Is...

“ The non-trivial process of identifying valid, novel , potentially useful, and ultimately understandable patterns in data.”

Fayyad, U. M.; Piaetsky-Shapiro, G.; Smyth, P. 1996. From Data Mining to Knowledge Discovery: An Overview. In , *Advances In Knowledge Discovery and Data Mining*. AAAI/MIT press, Cambridge mass.

What Is a Pattern?

" The non-trivial process of identifying valid, novel , potentially useful, and ultimately understandable patterns in data."

- Let **F** be a set of facts
- **E** is an expression in some language **L**
- Given $F_g \subseteq \mathbf{F}$
- Then **E** is a pattern if it is simpler then F_g

What Is a Pattern?

" The non-trivial process of identifying valid, novel , potentially useful, and ultimately understandable patterns in data."

- Let **F** be a set of facts
- **E** is an expression in some language **L**
- Given $F_g \subseteq F$
- Then **E** is a pattern if it is simpler then F_g

What Is a Pattern?

" The non-trivial process of identifying valid, novel , potentially useful, and ultimately understandable patterns in data."

- Let **F** be a set of facts
- **E** is an expression in some language **L**
- Given $F_g \subseteq F$
- Then **E** is a pattern if it is simpler then F_g

What Is a Pattern?

" The non-trivial process of identifying valid, novel , potentially useful, and ultimately understandable patterns in data."

- Let **F** be a set of facts
- **E** is an expression in some language **L**
- Given $F_g \subseteq \mathbf{F}$
- Then **E** is a pattern if it is simpler then F_g

What Is Validity?

" The non-trivial process of identifying valid, novel , potentially useful, and ultimately understandable patterns in data."

- Patterns are valid if they fall within certain bounds of certainty
- Certainty is some function **C** that maps expressions **E** in **L** to a partially or totally ordered measure space $\mathbf{M}_{\mathbf{c}}$ where **c = C(E,F)**

What Is Validity?

" The non-trivial process of identifying valid, novel , potentially useful, and ultimately understandable patterns in data."

- Patterns are valid if they fall within certain bounds of certainty
- Certainty is some function **C** that maps expressions **E** in **L** to a partially or totally ordered measure space $\mathbf{M}_c$ where **c = C(E,F)**

What Makes a Pattern Novel?

" The non-trivial process of identifying valid, novel , potentially useful, and ultimately understandable patterns in data."

- Patterns are novel with respect to the system
- Define a function **N(E,F)**
- Novelty is some function **N** that maps expressions **E** in **L** to a partially or totally ordered measure space $\mathbf{M}_{\mathbf{n}}$ where **n = N(E,F)**

What Makes a Pattern Potentially Useful?

" The non-trivial process of identifying valid, novel , potentially useful, and ultimately understandable patterns in data."

- Potentially useful means that pattern has the potential of resulting in some useful action
- Let **U** be a utility function that measures this potential
- U maps expressions **E** in **L** to a partially or totally ordered measure space $\mathbf{M}_{\mathbf{u}}$ where **u = U(E,F)**

What Makes a Pattern Ultimately Understandable?

" The non-trivial process of identifying valid, novel , potentially useful, and ultimately understandable patterns in data."

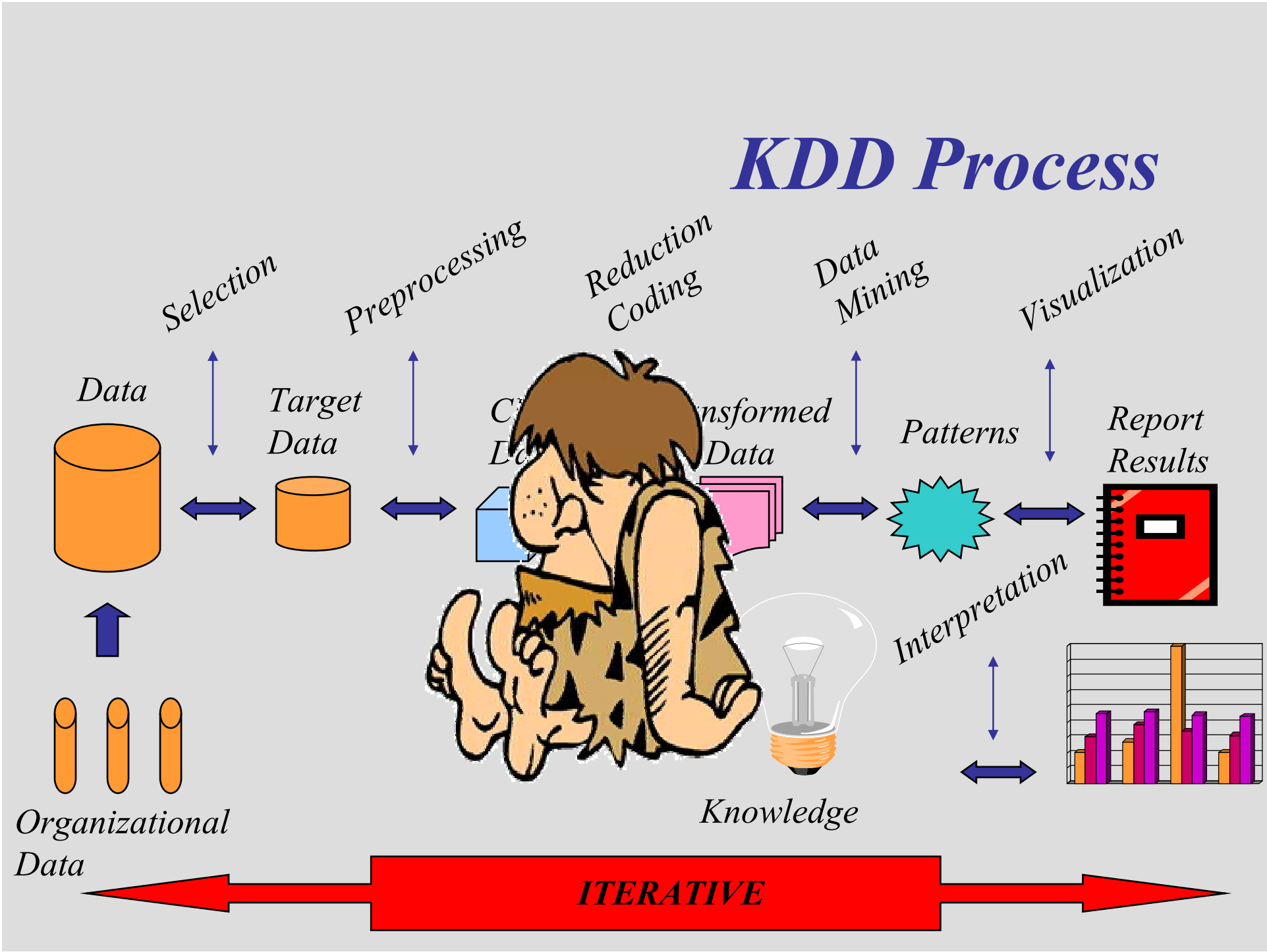
- To humans!! (End user)
- Hard parameter to measure - simplicity measure
- Let **S** be a simplicity measure
- **S** maps expressions **E** in **L** to a partially or totally ordered measure space M_s , where **s = S (E,F)**

What Else!! *(1 of 2)*

- Interestingness
  - Let **I** be some function that maps expressions **E** in **L** to a partially or totally ordered measure space $\mathbf{M}_{\mathbf{I}}$ where i = I(E,F,C,N,U,S)
  - Explicit or implicit measure

What Else!! *(2 of 2)*

- Knowledge
  - Given an expression $E \in L$, E is knowledge if for some user specified interestingness measure $i \in M_I$, $I(E,F,C,N,U,S) > i$
  - Not absolute measure, user defined
- Given some thresholds $c \in m_c$, $s \in m_s$ and $u \in m_u$, then a pattern E is knowledge iff $C(E,F) > c$ and $S(E,F) > s$ and $U(S,F) > u$


KDD Process
Selection
Preprocessing
Reduction
Coding
Data
Mining
Visualization
Data
Target
Data
Transformed
Data
Patterns
Report
Results
Interpretation
Organizational
Data
Knowledge
ITERATIVE

What Is Data Mining?

Data mining algorithms find patterns in large amounts of data by fitting models that are not necessarily statistical models.

- Fit models (may or may not be statistical)
- Determine patterns from large amounts of data
- Computational limits
  - Time
  - Hardware

Other Terminology for Data Mining In the Literature

- Knowledge extraction
- Knowledge mining from databases
- Information discovery
- Information harvesting
- Exploratory data analysis
- Data archeology
- Data dredging
- Data pattern analysis
- Intelligent Data Analysis

Data Mining – A search through a space of possibilities

More Formally:

- formally given a set of facts **F,** data mining results in an enumeration **Ej** of expressions, using a set of data mining algorithms

Data Mining Tasks

- Descriptive
  - find some human interpretable rules, relationships, and/or patterns
  - deviation detection, clustering, database segmentation, summarization and visualization, dependency modeling, cluster analysis
- Predictive
  - Infers from current data to make predictions
  - decision trees, neural networks, inductive logic programming (ILP), regression algorithms

Concept/Class Description

- Concept – poor student, good student
- Class – graduate student, undergraduate student, computer science student
- Data is associated with each class.
- Data characterization summarizes the data of the class under study or the target class
  - Summarize

Data Mining Algorithms – Three Components

- **model representation**
  - the language **L** use to represent the expressions (patterns) **E** in
  - is related to the type of information that is being discovered
  - language can also dictate the types of patterns discovered
  - need to choose the correct representation
  - If too descriptive a language is chosen there is a danger of over fitting the data.
  - the model has to be complex enough to explain the data but restrained enough to be able to generalize over new data
- **model evaluation**
  - the scoring methods used to see how well a pattern or model fits into the KDD process
- **search methodology**
  - greedy search, gradient descent

The Fear...

- Algorithms shouldn't be used ad hoc
- Might lead to discovery of patterns with no meaning
- If you look hard enough in a sufficiently large database, even a randomly generated one, can find statistically significant patterns

Men who buy diapers, buy beer!!

Gee Mom, we have to send Dad shopping more often !!!

The Truth About the Origins of Beer and Diapers

- K. Heath was trying to find groups of baby items were profitable
- Osco Drug Stores in Chicago (50 stores over a 90 day period)
- Association between diapers and beer was found using self joins in SQL

JARtool

- detected small volcanoes on Venus
- Magellan spacecraft - mapped surface for a period of 5 years
- 30,000 images of 1,000 X 1,000 pixels each
- Used a classifier that was trained on only 30 - 40 images

Health-Kefir

- analyzed healthcare transactions
  - control costs and improve quality
- Used deviation detection
- Interestingness measured by a deviation's impact
- Evaluation
  - field tested
  - no statistical corroboration
  - user feedback positive

Market Analysis

- predict buying patterns
- Customer databases are analyzed and searched for customer buying patterns and preferences.
- Techniques used
  - segmentation, interactive querying, predictive modeling
- Customers are selected in a more precise and targeted manner

Examples - Market Analysis (1 of 2)

- Coverstory and Spotlight
  - Analyzed supermarket sales data
  - patterns relating changes in product volume and share
    - Where do we ship those cans of beans?

Examples - Market Analysis (2 of 2)

- Opportunity explorer
  - Relationships for sales representatives of consumer packaged goods
  - The results are presented as advantages to retailers with regard to stocking of additional products or the running of special promotions

Market Basket Analysis

- point of sale info used to describe relationships between retail stock movement
  - shelf space allocation
  - store layout
  - product location and promotions
- IBM Data Miner, Lucent Technology's Niche Works, KEFIR (in the future)

Business Applications

- Coverstory and Spotlight - analyzed supermarket sales data to find relationships between product volume and share.
  - What type of beans do we ship and to which store do we ship it?
- IBM Data Miner, Lucent Technology's Niche Works - point of sale info used to describe relationships between retail stock movement
  - shelf space allocation
  - store layout
  - product location and promotions

Investments !!!

- manage stock portfolios
- proprietary
  - not usually described in the literature
  - Investment companies are competitive and don’t tend to publish their methods
- Use regression, neural networks.

Examples - Investment

- Fidelity Stock Selector
  - used a neural network to select investments
  - Results are presented to fund manager who makes final decision
  - Did well up to a point.
  - Uncertain whether the system was at fault or the human

- LBS Capital Management
  - manages funds worth $600 million
  - Uses a system of expert systems, neural networks and genetic algorithms
  - Since its inception in 1993, has outperformed the stock market.
- Carlberg & Associates
  - neural network used for predicting Standard and Poor's 500 Index
  - Used interest rates, oil prices, earnings, dividends, and the dollar index as inputs
  - Was able to explain 96% of the variation in S&P from 1986 to 1995

Even More Applications...

- PRISM and FALCON detect credit card fraud
- FAIS - detects money laundering
- AT&T uses a system to detect calling fraud
- Clonedetector by GTE - cellular phone clones
- IRS developing a pilot system for selecting returns for audit
- IBM’s ADVANCED SCOUT - analyzes data from NBA games to find patterns of play
- SKICAT - able to find faint sky objects

How to compare KDD systems (1 of 2)

- Who is the user of the KDDS?
- What types of tasks are supported by the KDDS?
- What tools are associated with each supported task?
- Are tools integrated with each other? Various steps in the process? User needs?

How to compare KDD systems (2 0f 2)

- In what manner does the system allow for incorporation of background knowledge?
- How is discovered information outputted?
- Are the results of the KDDS able to be used in some applicable way by a user other then the data engineer, i.e. some businessperson looking for a market trend, etc. ?

Ethical Issues (1 0f 2)

- invasion of privacy issues
- government and business databases contain a lot of personal information
- European Union Nations
  - The Organization for Economic Cooperation and Development (OECD)
    - data analysis on living individual's should not be done without their consent

Ethical Issues (2 of 2)

- Movements in this country for the same
- Most data mining deals with discerning patterns with regard to groups not individuals
- Problem in small datasets where combinations of group patterns may point to individuals.

Conclusions

- KDD is the process of finding patterns in large databases
- Data Mining is one step in the process
- Open areas of research exist in other steps of the process
- There are a wide breadth of successful applications with more to come

Knowledge Discovery in Databases

The answers are in there!!!