Data Mining Introduction

Instructor: Sharma Chakravarthy
sharmac@cse.uta.edu
The University of Texas at Arlington

A Few Tips and Suggestions

- My role as a teacher:
- Purpose of doing MS and its implications!
  - Get prepared for the next 30+ years of real-world job
- Is your Goal getting a GPA of 4.0 or getting a *good* job?
  - A good GPA does not imply you are prepared for job needs!
  - Answering questions by memorizing does not take you far
- You need to do BOTH: key is to understand concepts, not memorize them
  - Understanding goes a long way.
  - You can ace interviews as well
- You know how the interview process has changed over the years and what the employers are looking for!
  - Quick thinking on your feet! Problem solving approach! Not necessarily a complete solution
- How to Succeed in your Career!
- Understand the tradeoff between confidence and correctness in interviews!

A Few Tips and Suggestions

- Rather than thinking this a mining course course
- Think of what you are learning in this (and other courses as well) as a set of:
  - Technologies (skill sets)
  - Underpinnings of those technologies
  - How, why, and where they can be used
  - How to match tools to the job at hand!
- Most people get lost in theory, but never understand where it is appropriate to apply what you have learnt
- **Why, what, and how** are important in that order
  - **Why** do you want to use this particular approach/solution
  - **What** is the context and can you quantify the relevance
  - **How** to use it in the best possible way
- If you master the above three, you are in business

Data Mining

"The key in business is to know something that nobody else knows"

(Aristotle Onassis)

Motivation (Traditional mining)

Fraud division, some large telephone company:

"How do we find these guys? There are 10 billion records on 10 million customers in the main database. With all this information we have about our customers and all the calls they make, can't you just ask the database to figure out which lines have been set-up *temporarily* and exhibited *similar* calling patterns in the *same time periods*? The information is in there, I just know it …"

Problem

- "Find-similar" problem just described is hard
  - e.g., "Who should be given incentive?"
  - e.g., "How can we group usage patterns into k similar groups?"
  - E.g., "should we approve this purchase by this credit card by this user?"
- Why?
  - Massive amounts of data
  - More and more online data stores (e.g., Web, corporate databases, etc.)
- Challenge
  - Real-time vs. post-mortem analysis
- No easy way to describe what to look for
- Traditional, interactive approaches fail
  - Size of data, different purposes

Data Mining

- *Data Mining* (DM) is part of the knowledge discovery process carried out to extract valid patterns and relationships in very large data sets
  - Usually don't know what to look for, like a "voyage into the unknown"
- Regarded as knowledge discovery or learning from basic facts (axioms) and data
- Roots in AI and statistics
  - Uses techniques from machine learning, pattern recognition, statistics, database, visualization, etc.

What to do

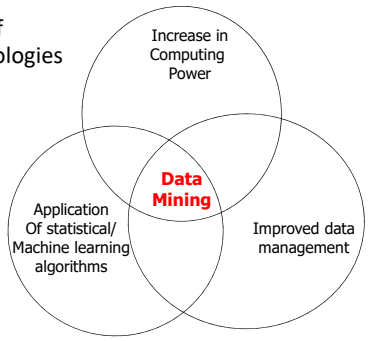
- A few months before the contract expires, if one can predict which customers are likely to quit,
  - Give incentive to those who are likely to quit
  - Don't do anything for those who are NOT likely to quit
- How do I predict future behavior?
  - Corporate Palm reading !
  - Human intuition !!
  - Data mining (DM) or knowledge discovery (KDD)

How Did the current Avatar of Data Mining come about?

- Mining has been around for quite a while!
- Enablers
  - Reduced cost of storage
  - Reduced cost of processing
  - Ability to store, process, and manage large volumes of data (e.g., DW, Internet)
  - New techniques such as association rules, sequence data processing, ontology, stream data processing, fusion
- However,
  - Scalability, visualization of results, filtering very large outputs are new and important issues!

How Did the current Avatar of Data Mining come about?

- Convergence of multiple technologies



Data Mining

- What is Causality?
  - "Cause and effect"
- What is correlation?
  - "mutual relationship or connection between two or more things and their strength"
- Do the above two mean the same thing?
  - NO!
- Which of the two we want to identify?
  - Causality!
- Which one does mining try to identify?
  - Correlation!
- Why?

check out spurious correlations

Characteristics: Data Mining

- Automated extraction of predictive/descriptive information from large data sets
- Key words:
  - Automated
  - Extraction
  - Predictive/descriptive
  - Large data sets
- A methodology is assumed (typically statistical)

AI and Statistics

- If DM is rooted in AI and statistics, what is the need for DM?
  - AI traditionally dealt with small samples
  - The emphasis was an learning, extrapolation, and generalization
  - The emphasis in DM is on processing actual data, not just samples!
  - DM tries to leverage the data collected, accumulated and derive tangible rules/conclusions (generalization is also possible)

DM Vs. Machine learning

- ML methods form the core of DM
- Amount of data makes a (big) difference
  - accessing examples can be a problem
  - missing values and incomplete data
- DM has more modest goals: automating the tedious discovery tasks

DM Vs. Statistics

- Similar goals; different methods
- Amount of data
- DM as a preliminary stage for statistical analysis
- Challenge to DM: better ties with statistics

DM vs. Big Data Analysis

- Complex (4V's), Large amounts of data
- Analysis requirements/expectations
- choice of modeling and computation using DM and any other techniques
- Vertical vs. Holistic analysis

What is NOT Data Mining?

- Data warehousing
- Ad hoc query (OLTP) /reporting
- Online Analytical Processing (OLAP), aggregation, summary
- Data Visualization
- Agents/mediators,
- Pervasive computing.
- Search / look up, …

What DM is not likely to do !

- Substitute for human intuition and discovery
- I don't think a DM system will (ever?) discover

  $e = mc^2$

  PV = RT

  Gravity, Newton's law's of motion, …
- It may discover new black holes !
- The value of pi is data-driven but its intuition is not!

DM Vs. DW

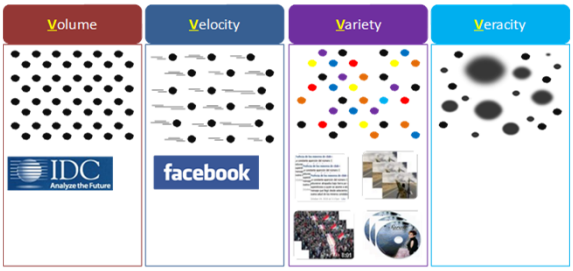
- DW makes DM a lot cheaper! Why?
- DM is one of the reasons for DW! Why?
- OLAP: verification-driven
  - sales in CA Vs. FL in Q1 of 2019
- DM: discovery-driven
  - Are there life forms outside of Earth?
  - Will company xx be a successful IPO?

OLAP Vs. Data Mining

- OLTP and OLAP
- OLAP is *user driven*
  - Analyst generates hypothesis, uses OLAP to *verify*
  - e.g., "people with high debt are bad credit risks"
- Data mining tool *generates* the hypothesis
  - Tool performs exploration
    - e.g., find risk factors for granting credit
  - Discover new patterns that analysts didn't think of
    - e.g., debt-to-income ratio
- OLAP and DM complement each other
- from Mining to Big Data Analytics

Characteristics of Big Data

- 4V: Volume, Velocity, Variety, Veracity

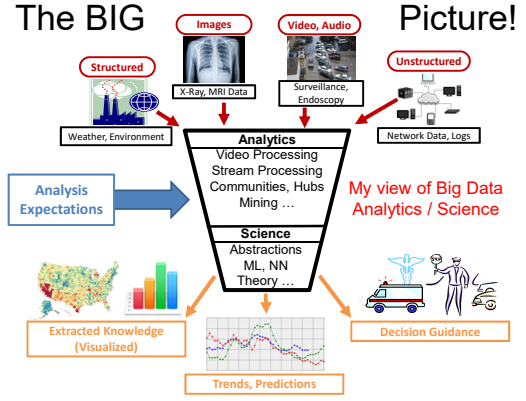


You will see many more V's in the trade magazines

DM vs. Big Data Analytics

- Definition: a process of inspecting, cleaning, transforming, and modeling big data with the goal of discovering useful information, suggesting conclusions, and supporting decision making **holistically**
- Connection to data mining
  - Analytics include both data analysis (mining) and communication (guide decision making)
  - Analytics is not so much concerned with individual analyses or analysis steps, but with the entire methodology

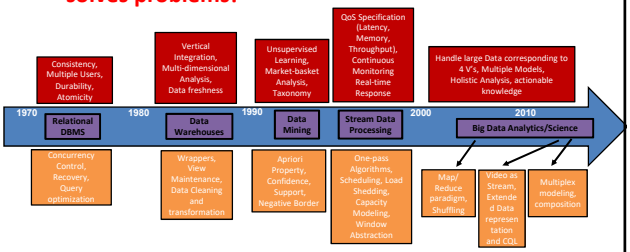
The BIG Picture!



Transforming **Disparate Data** into **Knowledge and Decisions**

Where are we headed?

- **Without understanding the past, it is very difficult to appreciate the present and plan for the future!**
- **Technology provides solutions; it does NOT mean it solves problems!**



Thank You !!!

For more information visit:
http://itlab.uta.edu