CS 484
Data Mining

Introduction

9/1/2015

Basics

Instructor: Dr. Jessica Lin

Contact Info:

Email: jessica@cs.gmu.edu
Homepage: http://www.cs.gmu.edu/~jessica
Office: Engineering Building Room 4419
Phone: (703)993-4693
Office Hours: Wednesday 2-3pm
                     Thursday 1:30-2:30pm

Class Meeting: Tuesday/Thursday 12:00-1:15PM
                        Art & Design Building 2026

Pre-requisites: C or better in CS 310 and STAT 344

TA: Jatin Mistry (jmistry2@gmu.edu)
TA Office Hours: TBA

Outline

- Course syllabus
- Introduction to Data Mining

Administration Trivia

- Class webpage:
  http://www.cs.gmu.edu/~jessica/cs484_f15.html
- In most cases, I will put the slides online the night before the lecture.
- You are 100% responsible for any announcements and updates on the class webpage, so visit the page frequently.

Textbook

- Required:
  - Introduction to Data Mining by Tan, Steinbach and Kumar

- Recommended:
  - Data Mining and Analysis by Mohammed Zaki
  - online pdf version: http://www.dataminingbook.info/pmwiki.php/Main/BookDownload

Grading

- Midterm Exams: 30%
- Final Exam: 30%
- Project: 20%
- Assignments: 20%

- There will be two midterms and one final exam.
- Final exam is comprehensive.
- All exams are closed-book, closed-notes.

Honor Code System

- GMU honor Code
  http://www.gmu.edu/academics/catalog/9798/honorcod.html

- In addition, the CS Department has specific honor code policies for programming projects, etc.:
  http://cs.gmu.edu/wiki/pmwiki.php/HonorCode/CSHonorCodePolicies

- For this class
  - You may work in a team of 2 for the project.
  - Homework: individual effort
  - Exams: individual effort, closed books/notes

Tools we will use for the class

- Weka (more on this later)
- Piazza
  - A free online class Q&A platform. You should have received an invitation to sign up.
  - Think before posting the question (e.g. is the answer in the book or the lecture slides?). You are encouraged to answer each other's questions but do not give out answers for assignments, obviously.
- Blackboard

Homework & Project Submission

- https://mymasonportal.gmu.edu/webapps/portal/frameset.jsp
- Login with your GMU student account
- Click on the Courses tab on the upper right hand corner
- Choose CS484
- Use Blackboard for:
  - Electronic submission of assignments and project
  - Checking grades
  - Getting course materials such as homework solutions
- Please bring a hardcopy to class.

Contact Policy

- Contact the TA (or post on Piazza) if you
  - Have questions about the course materials
  - Have questions about homework or project
- Contact me if you
  - Have questions about the exams
  - Have general questions/concerns about the course

Email Policy

- Please email me from your official GMU email. If you must email me from another account, you must state your full name and your official GMU email address.

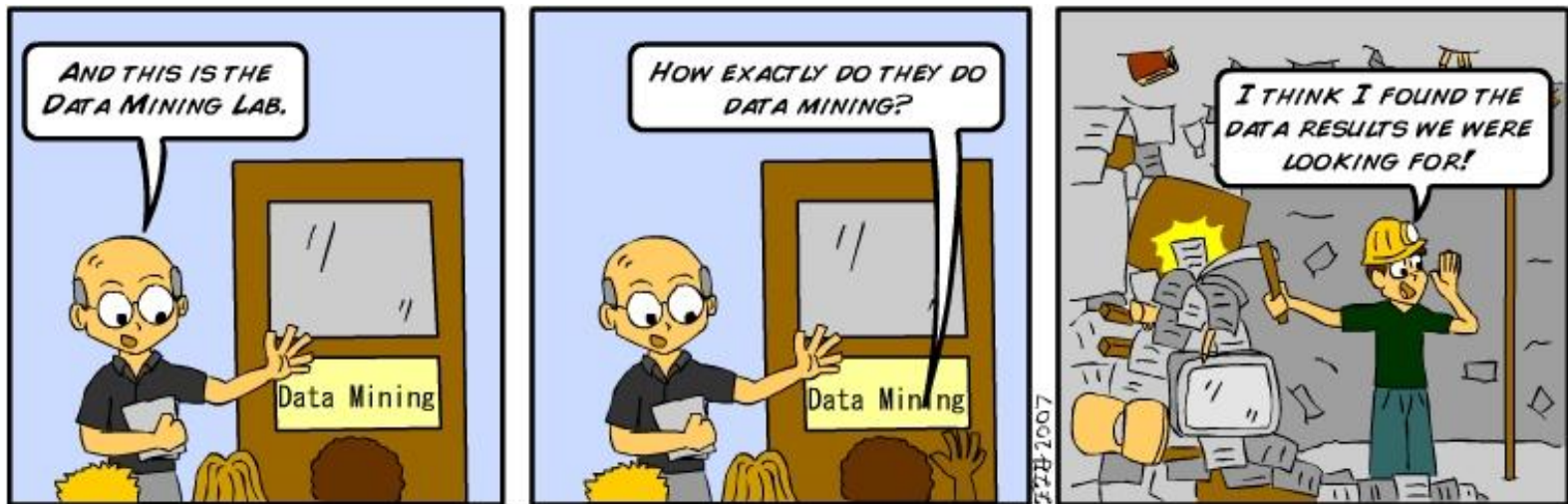- Please put [CS484] in the beginning of the subject line

Class Schedule

- See class website for the tentative class schedule (will be available tonight): http://www.cs.gmu.edu/~jessica/ cs484_f15.html

Any Questions?

What do you think of data mining?
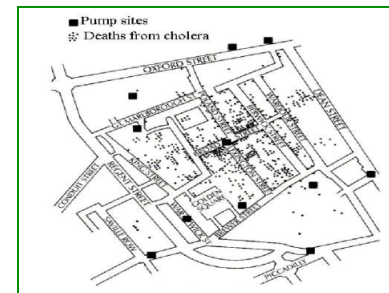
Large-scale Data is Everywhere!

- Tremendous data growth
- New mantra
  - Gather whatever data you can whenever and wherever possible.
- Expectations
  - Gathered data will have value either for the purpose collected or for a purpose not envisioned.

Homeland Security

Geo-spatial data

Business Data

Sensor Networks

Social Media & the Internet

An example business problem

- TelCo, a major telecommunications firm, wants to investigate its problem with customer attrition, or "churn"

- Lets consider this for now as a marketing problem only

How would you go about targeting some customers with a special offer, prior to contract expiration? Think about what data should be available for your use.

From data & business to strategy

HURRICANE FRANCES was on its way, barreling across the Caribbean, threatening a direct hit on Florida's Atlantic coast. Residents made for higher ground, but far away, in Bentonville, Ark., executives at Wal-Mart Stores decided that the situation offered a great opportunity for one of their newest data-driven weapons, something that the company calls predictive technology.

A week ahead of the storm's landfall, Linda M. Dillman, Wal-Mart's chief information officer, pressed her staff to come up with forecasts based on what had happened when Hurricane Charley struck several weeks earlier. Backed by the trillions of bytes' worth of shopper history that is stored in Wal-Mart's data warehouse, she felt that the company could "start predicting what's going to happen, instead of waiting for it to happen," as she put it.

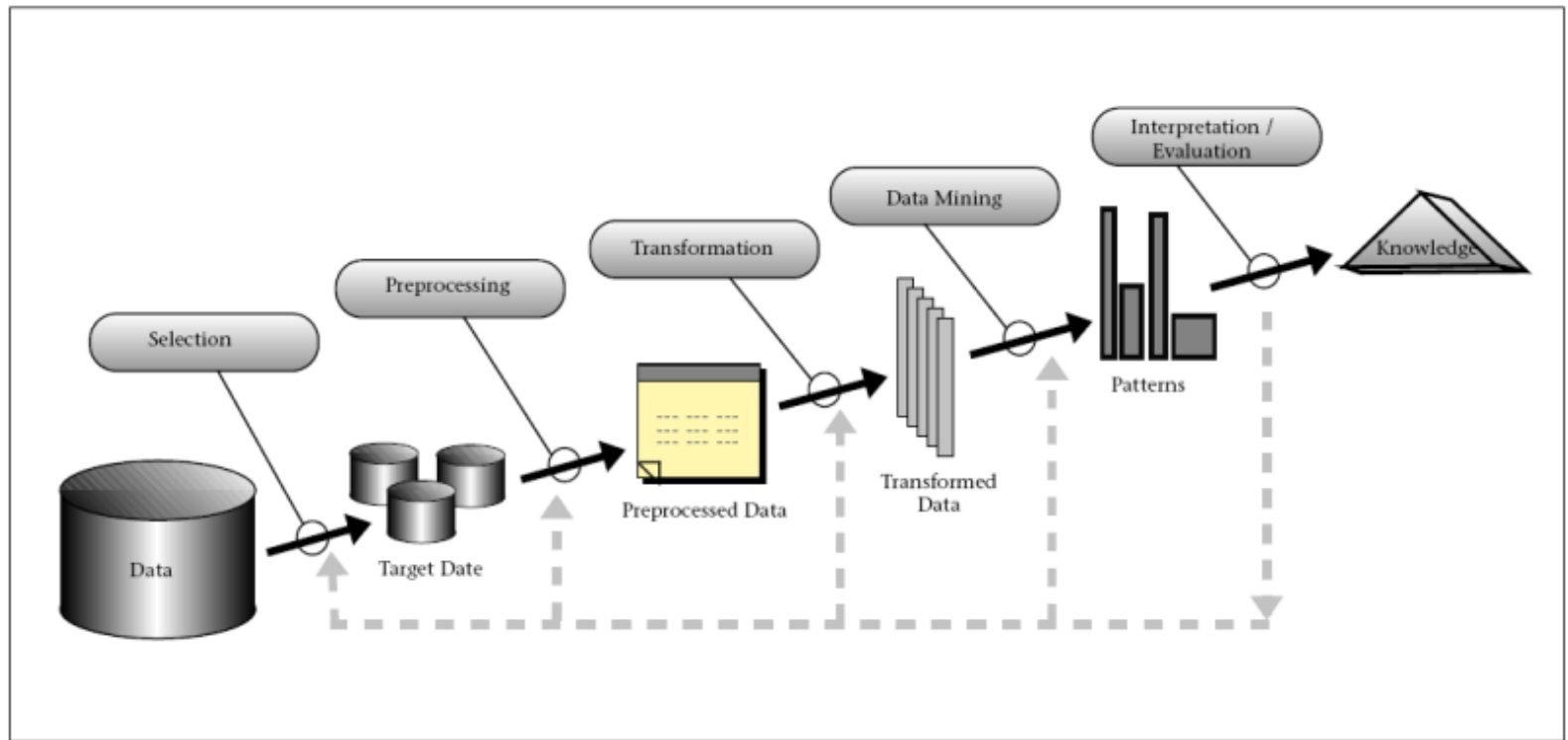From NY Times

What is Data Mining?

What really *is* data mining?

- *A process for using information technology to extract useful (non-trivial, hopefully actionable) knowledge from large bodies of data*

KDD Process

- KDD: (Knowledge Discovery in Databases)

**CONVERTING RAW DATA TO USEFUL INFORMATION.**

Why Data Mining? Commercial Viewpoint

- Lots of data is being collected and warehoused
  - Web data
    - Google processes 20 PB/day
    - Facebook has 955M active users
    - Twitter has more than 400M tweets/day
  - purchases at department/ grocery stores, e-commerce
    - Amazon has 42 TB of data
  - 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)

Why Data Mining? Scientific Viewpoint

- Data collected and stored at enormous speeds
  - remote sensors on a satellite
    - NASA EOSDIS archives over 1-petabytes of earth science data / year
  - telescopes scanning the skies
    - Sky survey data
    - LSST (Large Synoptic Survey Telescope) project: 20 PB science data & 100 PB image archive
  - High-throughput biological data
  - scientific simulations
    - terabytes of data generated in a few hours
- Data mining helps scientists
  - in automated analysis of massive datasets
  - in hypothesis formation
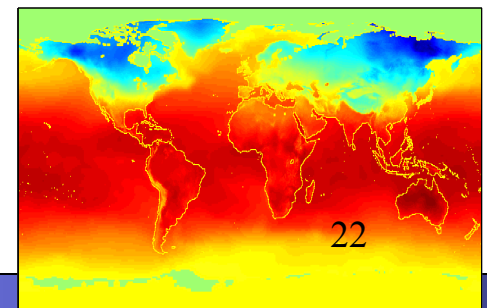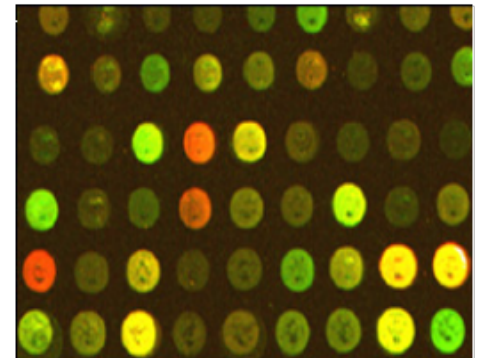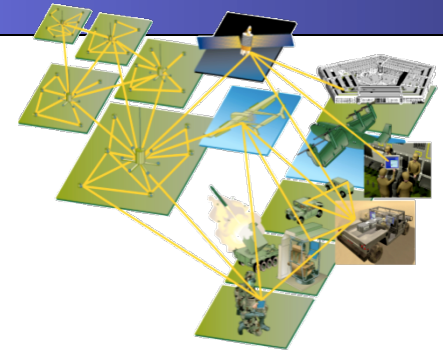
What is (not) Data Mining?

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)

23

Some Data Mining Examples

- Amazon.com, Google, Netflix
 - Personal Recommendations
  - Profile-based advertisements

- Spam Filters/Priority Inbox
  - 94B spam messages sent daily; cost society $20B annually

- Scientific Discovery
  - Finding potential causes of cancer
  - Prediction of weather and natural disasters

- Security
  - Phone Conversations, Network Traffic

- Politics
  - Obama campaign: how data mining helped Obama win the election

My Data Mining Projects

- CPU manufacturing data
  - Compare "good" chips and "bad" chips
  - What are the potential causes of "bad" chips?
- ICU medical alarms
  - "Alarm fatigue"
  - How to reduce false alarms
- Trajectory data mining
  - Find repeated routes
  - Anomaly detection, route recommendation, etc.
- Power load prediction

Mining Large Data Sets - Motivation

- There is often information "hidden" in the data that is not readily evident
- Human analysts may take weeks to discover useful information
- Much of the data is never analyzed at all
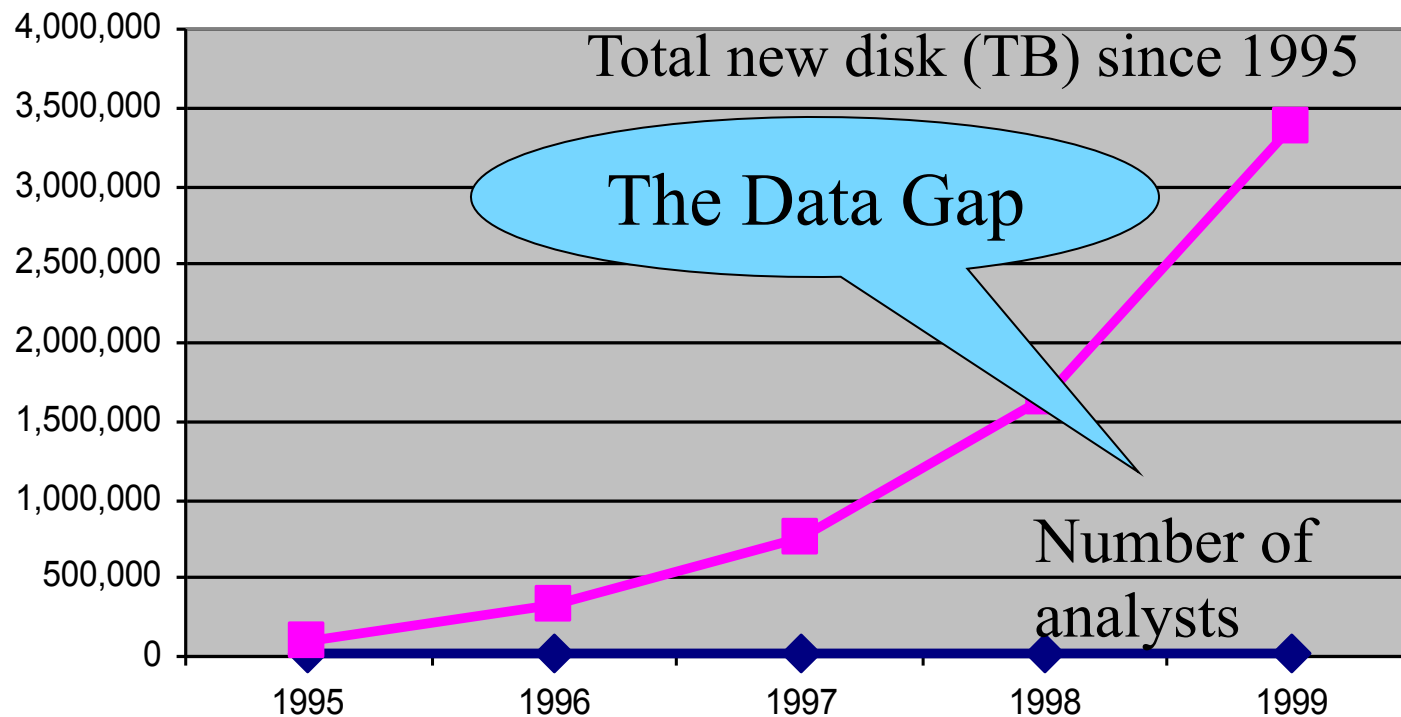
From: R. Grossman, C. Kamath, V. Kumar, "Data Mining for Scientific and Engineering Applications"

Data Mining Tasks

- Prediction Methods

  – Use some variables to predict unknown or future values of other variables.

- Description Methods

  – Find human-interpretable patterns that describe the data. Make good inferences from the data.

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Applications of Data Mining

- **Prediction** based on past history
  - Predict if a credit card applicant poses a good credit risk, based on some attributes (income, job type, age, ..) and past history
  - Predict if a customer is likely to switch brand loyalty (churn)
  - Predict if a customer is likely to respond to "junk mail"
  - Predict if a pattern of phone calling card usage is likely to be fraudulent
- Example of prediction mechanisms:
  - **Classification -** Given a training set consisting of items belonging to different classes, and a new item whose class is unknown, predict which class it belongs to
  - **Regression** – Predicting the value of one variable from one or more variables.

Applications of Data Mining (Cont.)

- **Descriptive Patterns**
  - **Associations** – Find items that are often bought by the same customers. If a new customer buys one such item, suggest that he buys the others too.
    - Associations may also be used as a first step in detecting **causation**
    - E.g. association between exposure to chemical X and cancer, or new medicine and cardiac problems
  - **Clusters** – Finding natural grouping in data
    - E.g. Do my customers form natural groups?
    - E.g. typhoid cases were clustered in an area surrounding a contaminated well

29

Classification Example

categorical   categorical   continuous   class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?

Training Set → Learn Classifier → Model

Test Set → Model
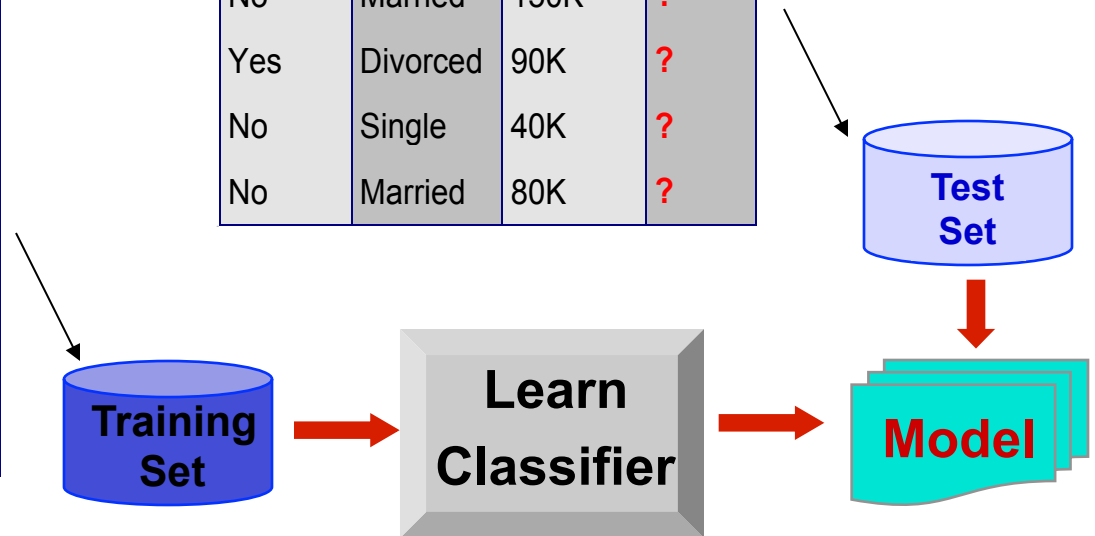
Classification: Definition

- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: <u>previously unseen</u> records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.
- ***Supervised*** learning

Classification: Direct Marketing

- Direct Marketing
  - Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
  - Approach:
    - Use the data for a similar product introduced before.
    - We know which customers decided to buy and which decided otherwise. This *{buy, don't buy}* decision forms the *class attribute*.
    - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
      - Type of business, where they stay, how much they earn, etc.
    - Use this information as input attributes to learn a classifier model.

From [Berry & Linoff] Data Mining Techniques, 1997

Classification: Fraud Detection

- Fraud Detection
  - Goal: Predict fraudulent cases in credit card transactions.
  - Approach:

Classification: Your Turn

- Churn prediction
  - Goal: Predict which customers will terminate contracts soon after they expire
  - Approach:

- Can you think of 3 more applications for classification?