

Statistical Learning and Data Mining

CS 363D/ SSC 358

Lecture: Introduction

Pradeep Ravikumar
pradeepr@cs.utexas.edu

What is this course about (in 1 minute)

- “Big Data”
- Data Mining, Statistical Learning
- Pre-reqs: Basic knowledge of **linear algebra**, **probability**, **programming**, data structures, and algorithms
 - ▶ We will review {linear algebra, probability} fundamentals at appropriate junctures

Class Webpage

- <https://piazza.com/utexas/spring2014/ssc358cs363d/>

The screenshot shows a web browser window displaying the Piazza class webpage. The browser's address bar shows the URL <https://piazza.com/utexas/spring2014/ssc358cs363d/home>. The page header includes the Piazza logo, the course name "SSC 358/CS 363D", and navigation links for "Q & A", "Course Page", and "Manage Class". The user's name, "Pradeep Ravikum", is visible in the top right corner. The main content area features the course title "SSC 358/CS 363D: Statistical Learning and Data Mining" and a "Syllabus" button. Below this, there are tabs for "Course Information", "Staff", and "Resources". The "Description" section provides a detailed overview of the course, mentioning the focus on data collection, storage, and mining. The "General Information" section lists the location (GDC 1.304) and the schedule (Mondays/Wednesdays 3:30 - 5:00 PM). The "Announcements" section includes an "Add" button and a message to "Add an Announcement". The footer of the page states "Copyright © 2013 Piazza Technologies, Inc. All Rights Reserved."

SSC 358/CS 363D | Class

[←](#) [→](#) [↻](#) <https://piazza.com/utexas/spring2014/ssc358cs363d/home>

piazza SSC 358/CS 363D Q & A Course Page Manage Class

The University of Texas at Austin - Spring 2014

SSC 358/CS 363D: Statistical Learning and Data Mining

Syllabus

Course Information Staff Resources

Description

In recent years, rapid developments in data collection and storage technologies have led to data sets that are "big" in many senses of the word. Data mining is the automatic discovery of interesting patterns and relationships in such "big data". This undergraduate course will provide an introduction to the topic of data mining, and some statistical principles underlying its key methods. Topics covered will include data preprocessing, regression, classification, clustering, dimensionality reduction, and association analysis.

General Information

Where
GDC 1.304

When
Mondays/Wednesdays 3:30 - 5:00 PM

Announcements

Add an Announcement
Click the Add button to add an announcement.

Copyright © 2013 Piazza Technologies, Inc. All Rights Reserved.

Class Discussion

- <https://piazza.com/utexas/spring2014/ssc358cs363d/>
- Piazza: a discussion board, where you can ask questions (anonymously if need be)
 - ▶ Any question has a single wiki-editable student answer; students can (and should) collectively add to this answer. I or TA will edit if need be, certify.
 - ▶ Statistics reg. the most helpful people will be visible; be a good citizen! :-)
 - ▶ Comment on or ask further questions about a post, by starting a followup discussion
 - ▶ Can format equations into your posts/questions/answers
 - ▶ Any small question, please add it to the class Piazza site

Textbooks and Materials

- Introduction to Data Mining. P. Tan, M. Steinbach, V. Kumar, Addison Wesley, 2006.
 - ▶ Textbook URL: <http://www-users.cs.umn.edu/~kumar/dmbook/index.php>
- Misc. materials will be posted on Piazza

Grading Policy

- 5 Homeworks: 25%
- 1 Midterm: 16%; 1 Final: 24%
- Final Project: 30%
- Class Attendance and Participation: 5%

Grading Policy: Homeworks

- Five Homeworks (25%)
 - ▶ Due beginning of class on the due date
 - ▶ Two “free” late days: use it all on one homework, or on two different homeworks
 - ▶ Homework will be worth 50% if one day late, and 0% if it is two or more days late. It is required to submit all homeworks even if after two days.

Grading Policy: Project

- Final Project (30 %)
 - ▶ a. Initial Project Milestone (5%) ; due Apr 02.
 - ▶ b. Final Project Presentation (25%) ; due May 02.
 - ▶ The list of candidate projects will be provided once the class gets underway. You have to work individually on your class project.

Why Data Mining: Commercial Viewpoint

- Lots of data is being collected and warehoused

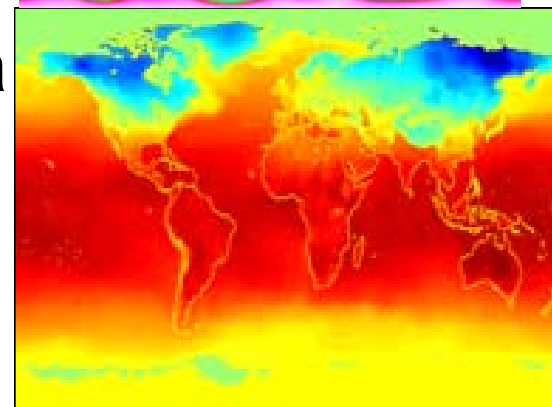
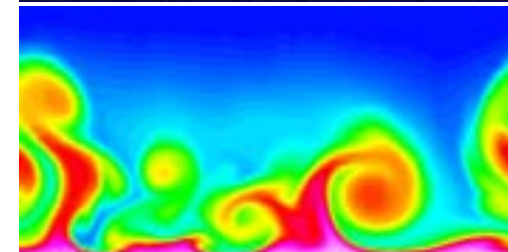
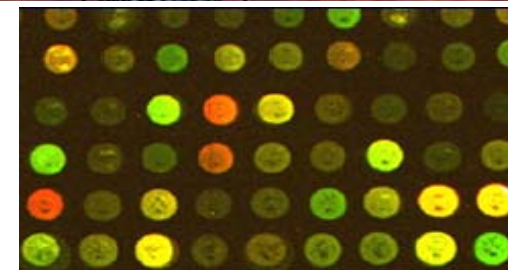
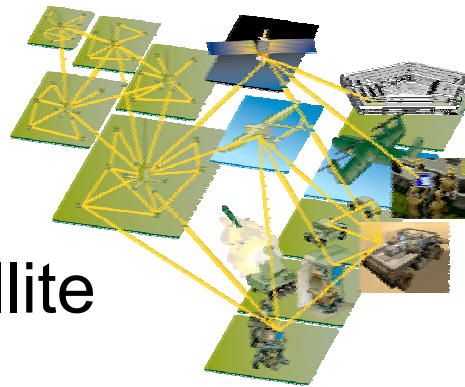
- Web data, e-commerce
- purchases at department/grocery stores
- Bank/Credit Card transactions



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

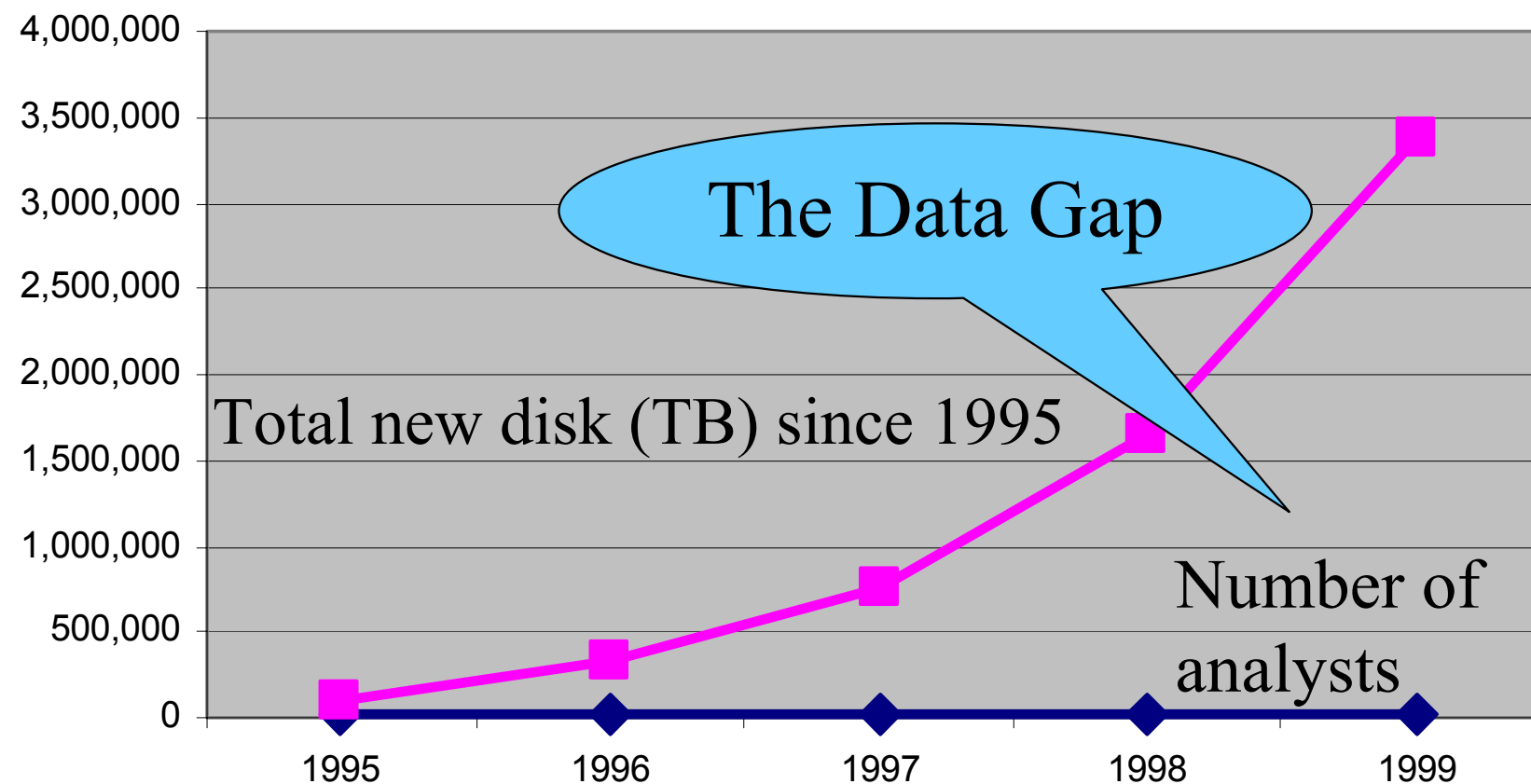
Why Data Mining: Scientific Viewpoint

- Data collected and stored at enormous speeds (GB/hour)
 - remote sensors on a satellite
 - telescopes scanning the skies
 - 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: Automation Viewpoint

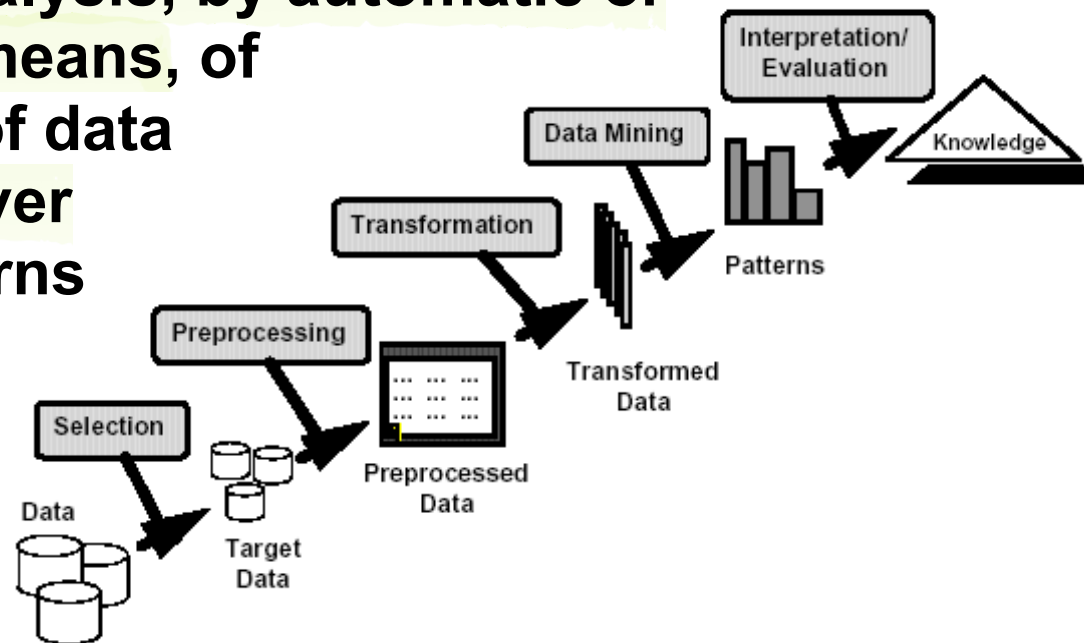
- There is often information “hidden” in the data that is not readily evident
- Human analysts may take weeks to discover useful information
- Much of the data is never analyzed at all



What is Data Mining?

● Many Definitions

- Non-trivial extraction of implicit, previously unknown and potentially useful information from data
- Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns



What is Data Mining?

● What is not Data Mining?

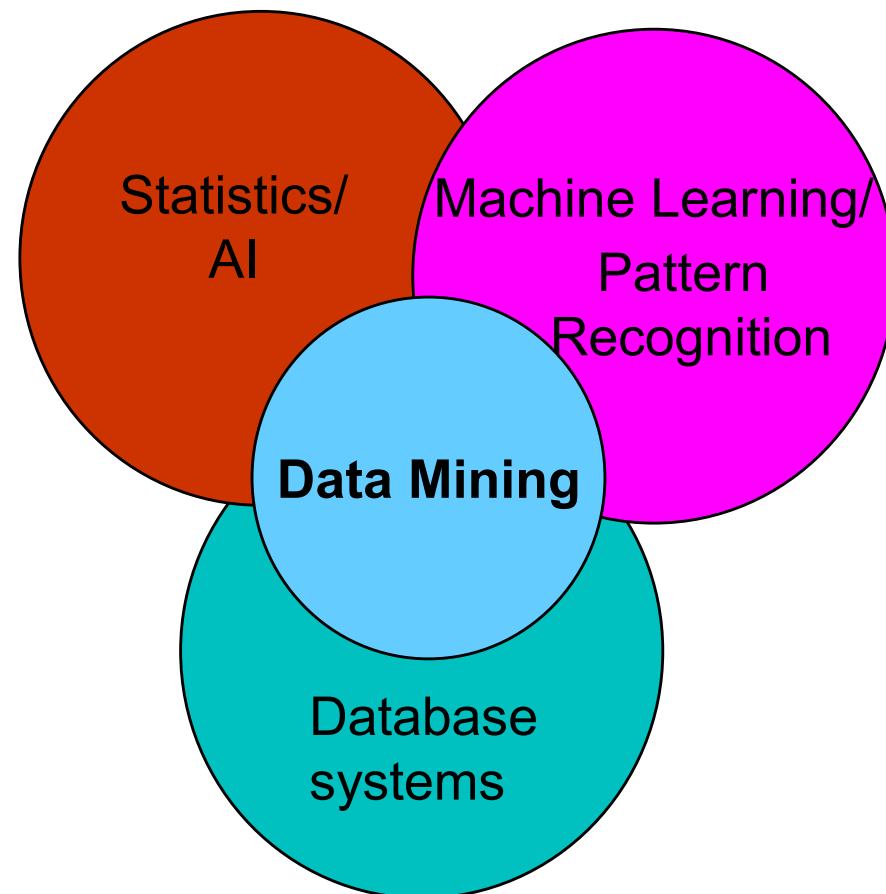
- Look up phone number in phone directory
- Query a Web search engine for information about “Amazon”

● What is Data Mining?

- Certain names are more prevalent in certain US locations (O’Brien, O’Rourke, O’Reilly... in Boston area)
- Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)

Origins of Data Mining

- **Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems**
- **Traditional Techniques may be unsuitable due to**
 - **Enormity of data**
 - **High dimensionality of data**
 - **Heterogeneous, distributed nature of data**



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.

Data Mining Tasks

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Regression [Predictive]
- Anomaly Detection [Predictive]
- SVD [Descriptive]