Course Staff

- **Instructor:**
  - Jure Leskovec

- **TAs:**
  - Aditya Parameswaran
  - Bahman Bahmani
  - Peyman Kazemian
Course Logistics

- Course website:
  http://cs246.stanford.edu
  - Lecture slides (~30min before the lecture)
  - Announcements, homeworks, solutions
  - Readings!

- **Readings**: Book *Mining of Massive Datasets* by Anand Rajaraman and Jeffrey D. Ullman

Fee online:
http://i.stanford.edu/~ullman/mmds.html
Logistics: Communication

- Send questions/clarifications to: cs246-win1011-staff@lists.stanford.edu

- Course mailing list:
  cs246-win1011-all@lists.stanford.edu
  - If you are auditing send us email and we will subscribe you!

- Office hours:
  - Jure: Tuesdays 9-10am, Gates 418
  - See course website for TA office hours
Work for the Course

- **4 Longer homeworks**: 30%
  - theoretical and programming/data analysis questions
  - All homeworks (even if empty) must be handed in
  - **Start early!!!!**
- **Short weekly quizzes**: 20%
  - Short e-quizzes on Gradiance
  - **No late days!**
- **Final Exam**: 50%

- **It’s going to be fun and hard work 😊**
## Course Calendar

<table>
<thead>
<tr>
<th>Date</th>
<th>Out</th>
<th>In</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/5</td>
<td>HW1</td>
<td></td>
</tr>
<tr>
<td>1/19</td>
<td>HW2</td>
<td>HW1</td>
</tr>
<tr>
<td>2/2</td>
<td>HW3</td>
<td>HW2</td>
</tr>
<tr>
<td>2/16</td>
<td>HW4</td>
<td>HW3</td>
</tr>
<tr>
<td>3/2</td>
<td></td>
<td>HW4</td>
</tr>
</tbody>
</table>

- **No class:** 1/17: Martin Luther King Jr.
  2/21: President’s day

- **2 recitations:**
  - Review of basic concepts
  - Installing and working with Hadoop
**What is Data Mining?**

- Discovery of useful, possibly unexpected, patterns in data
- Subsidiary issues:
  - **Data cleansing**: detection of bogus data
    - E.g., age = 150
    - Entity resolution
  - **Visualization**: something better than megabyte files of output
  - **Warehousing** of data (for retrieval)
Different Cultures

- **Databases:**
  - concentrate on large-scale (non-main-memory) data

- **AI (machine-learning):**
  - concentrate on complex methods, usually small data

- **Statistics:**
  - concentrate on models
Models vs. Analytic Processing

- To a database person, data-mining is an extreme form of **analytic processing** – queries that examine large amounts of data:
  - Result is the data that answers the query.

- To a statistician, data-mining is the inference of models:
  - Result is the parameters of the model.
Focus 1: Web Mining

- Much of the course will be devoted to ways to data mining on the Web:
  - Mining to discover things about the Web
    - E.g., PageRank, finding spam sites
  - Mining data from the Web itself
    - E.g., analysis of click streams, similar products at Amazon, making recommendations.
Much of the course will be devoted to Large scale computing for data mining

Challenges:
- How to distribute computation?
- Distributed/parallel programming is hard

Map-reduce addresses all of the above
- Google’s computational/data manipulation model
- Elegant way to work with big data
Course Outline (1)

- Association rules, frequent itemsets
- PageRank and related measures of importance on the Web (link analysis)
  - Spam detection
  - Topic-specific search Recommendation systems
  - E.g., what should Amazon suggest you buy?
- Large scale machine learning methods
  - SVMs, decision trees, ...
Min-hashing/Locality-Sensitive Hashing
  - Finding similar Web pages
- Clustering data
- Extracting structured data (relations) from the Web
- Managing Web advertisements
- Mining data streams
Prerequisites

- **Algorithms:**
  - Dynamic programming, basic data structures
- **Basic probability (CS109 or Stat116):**
  - Moments, typical distributions, regression, ...
- **Programming (CS107 or CS145):**
  - Your choice, but C++/Java will be very useful

- We provide some background, but the class will be fast paced
CS345a: Data mining got split into 2 course:

- **CS246: Mining massive datasets:**
  - Methods oriented course
  - Homeworks (theory & programming)
  - No massive class project

- **CS341: Advanced topics in data mining:**
  - Project oriented class
  - Lectures/readings related to the project
  - Unlimited access to Amazon EC2 cluster
  - We intend to keep the class to be small
    - Taking CS246 is basically essential
Why Mine Data? Industry

- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - Purchases at department/grocery stores
  - Bank/Credit Card transactions

- Computers are cheap and powerful

- Goal:
  - Provide better, customized services (e.g. in Customer Relationship Management)
Why Mine Data? Science

- 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 helps scientists
  - in classifying and segmenting data
  - in Hypothesis Formation
There is often information “hidden” in the data that is not readily evident

- Human analysts take weeks to discover useful information
- Much of the data is never analyzed at all

![Graph showing the Data Gap](image)

- **The Data Gap**
- **Total new disk (TB) since 1995**
- **Number of analysts**
What is Data Mining?

- Non-trivial extraction of implicit, previously unknown and useful information from data
A big data-mining risk is that you will “discover” patterns that are meaningless.

Bonferroni’s principle: (roughly) if you look in more places for interesting patterns than your amount of data will support, you are bound to find crap.
A parapsychologist in the 1950’s hypothesized that some people had Extra-Sensory Perception.

He devised an experiment where subjects were asked to guess 10 hidden cards – red or blue.

He discovered that almost 1 in 1000 had ESP – they were able to get all 10 right.
Rhine Paradox – (2)

- He told these people they had ESP and called them in for another test of the same type.

- Alas, he discovered that almost all of them had lost their ESP.

- What did he conclude?

- He concluded that you shouldn’t tell people they have ESP; it causes them to lose it.
Overlaps with machine learning, statistics, artificial intelligence, databases, visualization but more stress on

- **scalability** of number of features and instances
- stress on **algorithms** and architectures
- automation for handling large, heterogeneous 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.
Given database of user preferences, predict preference of new user

Example:
- Predict what new movies you will like based on
  - your past preferences
  - others with similar past preferences
  - their preferences for the new movies

Example:
- Predict what books/CDs a person may want to buy
  - (and suggest it, or give discounts to tempt customer)
Anomaly Detection

- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection
  - Network Intrusion Detection
Supermarket shelf management:

- **Goal**: To identify items that are bought together by sufficiently many customers.
- **Approach**: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
- **A classic rule**:
  - If a customer buys diaper and milk, then he is likely to buy beer.
  - So, don’t be surprised if you find six-packs stacked next to diapers!

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Coke, Milk</td>
</tr>
<tr>
<td>2</td>
<td>Beer, Bread</td>
</tr>
<tr>
<td>3</td>
<td>Beer, Coke, Diaper, Milk</td>
</tr>
<tr>
<td>4</td>
<td>Beer, Bread, Diaper, Milk</td>
</tr>
<tr>
<td>5</td>
<td>Coke, Diaper, Milk</td>
</tr>
</tbody>
</table>

Rules Discovered:

\[
\{\text{Milk}\} \rightarrow \{\text{Coke}\} \\
\{\text{Diaper, Milk}\} \rightarrow \{\text{Beer}\}
\]
Data Mining

- Process of semi-automatically analyzing large datasets to find patterns that are:
  - **valid**: hold on new data with some certainty
  - **novel**: non-obvious to the system
  - **useful**: should be possible to act on the item
  - **understandable**: humans should be able to interpret the pattern
Some Success Stories (1)

- Network intrusion detection using a combination of sequential rule discovery and classification tree on 4 GB DARPA data
  - Won over (manual) knowledge engineering approach
Some Success Stories (2)

- Major US bank: Customer attrition prediction
  - Segment customers based on financial behavior: 3 segments
  - Build attrition models for each of the 3 segments
  - 40-50% of attritions were predicted == factor of 18 increase
Some Success Stories (3)

- **Targeted credit marketing**: major US banks
  - find customer segments based on 13 months credit balances
  - build another response model based on surveys
  - increased response 4 times – 2%
Challenges of Data Mining

- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation
- Streaming Data
Applications

- **Banking: loan/credit card approval:**
  - predict good customers based on old customers

- **Customer relationship management:**
  - identify those who are likely to leave for a competitor

- **Targeted marketing:**
  - identify likely responders to promotions

- **Fraud detection:** telecommunications, finance
  - from an online stream of event identify fraudulent events

- **Manufacturing and production:**
  - automatically adjust knobs when process parameter changes
Applications (continued)

- **Medicine**: disease outcome, effectiveness of treatments
  - analyze patient disease history: find relationship between diseases

- **Molecular/Pharmaceutical**: identify new drugs

- **Scientific data analysis**: identify new galaxies by searching for sub clusters

- **Web site/store design and promotion**: find affinity of visitor to pages and modify layout