CS341: Project in Mining Massive Datasets
Info session

Jure Leskovec
Anand Rajaraman
Jeff Ullman
Data mining research project on real data
- Teams of 3 students *(Use Piazza to form teams)*
- We have room for 10-15 teams

**We provide:**
- Data
- **Computers** (Amazon EC2, 3k$ per team)
- **Mentoring:** Each group will have an assigned mentor that they meet on a weekly basis

**You provide:**
- Project proposals
- Work
CS341: Schedule

- Today (3/18): Info session
- Tue 3/25: Project proposals due
- Mon 3/31: Admission results
  - 10 to 12 projects will be admitted
- Tue 4/1: First class meeting
- Tue 4/29, Thu 5/1: Midterm presentations
- Tue 6/3, Thu 6/5: Presentations, poster session

More info: http://cs341.stanford.edu
Project Types

- **Data analysis/Modeling project:**
  - Discovers *interesting relationships* within a significant amount of data

- **Algorithmic project that extends/builds on what we learned in class**
  - Extend/Improve/Speed-up some existing algorithm
  - Define a *new problem* and *solve it*
(1) **What is the problem/question your team is solving?**

- Give a brief but precise description or definition of the problem or question
  - **Examples:**
    - (a) Analyze the data to understand why editors are leaving Wikipedia
    - (b) Build a social recommender engine for movies
    - (c) Design a MapReduce algorithm for finding clusters in graphs

(2) **What data will you use?**

- Why is the data you plan to use appropriate? Does it have the right labels/information?
- It is ok to use your own data (give detailed description)!
  - **Examples:**
    - (a) Wikipedia edit history where every action of every user is recorded
    - (b) We crawled Yelp and obtained X million reviews from Y million users
    - (c) We will use the Altavista web graph on X million nodes.
(3) How will you solve the problem? What is your plan of action?

- Describe and think about your approach!
  - What method, algorithm, technique? How will you scale it up?
  - Be as specific as you can!
  - Examples:
    - (a) We will create edit histories of every article. We will then compare article edit histories and argue that users are leaving since all the “easy/obvious” articles have already been written.
    - (b) Our hypothesis is that friends have similar tastes. We will include a regularization term to a Latent Factor Rec. Sys. which will encourage neighboring users to have similar parameters.
    - (c) We will implement a scalable Frequent itemset based approach to identify cluster seeds (complete bipartite subgraphs). In the second pass we will then use a random walk based approach to expand around the seed and extract the clusters.
(4) How will you evaluate your method?

How will you measure performance or success of your method? What baselines will you use?

Examples:

(a) Using insights from our analysis we will build a model that will predict how complete is the article (much the article will change in the future). We will evaluate predictive accuracy of the model.

(b) We will measure RMSE of our system. As a baseline for comparison will use traditional latent factor recommender.

(c) We will measure resource usage and execution time of our algorithm and compare it to open source algs. Metis and Graclus.

(5) What do you expect to submit/accomplish by the end of the quarter?
Projects: Proposals

- **Submit** to cs341-spr1314-staff@lists.stanford.edu
  - **PDF should include**
    - Project title
    - Project narrative addressing the 5 Qs
    - Information about team members:
      - For each team member: 5 line CV/Bio about prior experience, and why are you suitable to take this course
  - **No page limit** (we don’t promise to read past page 3)
  - **Due Tuesday 3/24 11:59pm Pacific time**
  - **We will let you know whether you got in by Monday March 31**
  - **More info at:** [http://cs341.stanford.edu/info.html](http://cs341.stanford.edu/info.html)
Project ideas / datasets
(1) Movie Scenes (Jure)

- **Dataset of 10,000 movies**
- **Every scene is annotated:**
  - **Actions:**
    - Attacking, Fighting, Flying, ...
  - **Locations:**
    - Airport, Garage, Gym, ...
  - **Objects:**
    - Animal, Books, Drink, Drugs, ...
  - **Appearance:**
    - Actor, Character, Nickname, Type
  - **Genre of the scene:**
    - Action, Family, History, War,

```
"netflix": {
  "id": "70213514",
  "genres": [
    "Action & Adventure",
    "Action Sci-Fi & Fantasy",
    "Sci-Fi Thrillers",
    "Action Thrillers",
    "Blockbusters"
  ]
},
"rotten_tomatoes": {
  "id": "771041731",
  "critics_score": 87,
  "audience_score": 92
},
{ "hitType": "tag",
  "subTrack": "Vehicle",
  "startTime": 2682.9823,
},
{ "hitType": "tag",
  "subTrack": "Driving",
  "track": "Action",
  "startTime": 2685.8498,
}"
```
Some ideas:

(1) How films are similar to each other
(2) Trending scene types in popular films
(3) Predicting scene based Genre
(4) Learn models that automatically classify edges (e.g., adversary, friend, lover) based on the patterns of interaction between people
(5) A qualitative study of how interaction patterns etc. differ between genres, older films etc.

Can be combined with IMDB, Rotten Tomatoes,...

If you are interested send us email and we’ll share data for 1 movie with you
Online media

- Collection of 6B news documents and 300M short textual phrases that appear in them
  - Think of this as a complete trace of U.S. news media space for the last 5 years!

Goal:

- Detect trending topics and explores the dynamics of online news
  - Based on time, named entities, mutation of information
- If interested send Jure email and we can talk more
Distributed processing of large graphs
- System and algorithms for processing graphs
- We have some ideas here, talk to us

Graph anomaly detection
- Imagine a communication network
  - Email, phone calls (we have both)
- Can you spot and identify anomalies in communication patterns

If interested send us email and we can talk more
Ratings and reviews of Beer, Wine, Movies
- Plus user data, temporal information
- [http://snap.stanford.edu/data/](http://snap.stanford.edu/data/)

Questions:
- Evolution of user tastes and opinions
- User contribution to the community
- Beer is reviewed along 5 distinct dimensions
  - Taste, smell, feel, color and overall
Tweets, videos, or apps that come from nowhere and become huge hits

How soon can we identify such hits?
- Can we say a tweet will be re-tweeted a million times in the next 24 hours a few minutes after it’s tweeted?
- An app will make the Top 10 in the app store a few days after its launch?

Two datasets: Twitter and mobile apps
10% sample of Twitter firehose for March 2014

Each tweet contains many fields
- Tweet text, user, #followers, location, link(s), hashtag(s), ...

Predict the number of retweets for a tweet within a given time interval (e.g., 24 hours)
- Predict well in advance which tweets will be the most re-tweeted
- Normalize by number of followers or other means
(5.2) Mobile App downloads (Anand)

- Download ranks of iOS apps over several months
  - Category (gaming, productivity, ...)
  - Country (6 English-speaking countries)
  - Device type (iPhone, iPad)
  - Publisher

- Can we separate the marketing hype from the true winners?

- Can also use Twitter data for additional boost
Data set from Synapse.org
- Brief voice recordings (3-30 seconds long)
- 39 features extracted from each recording
- Self-reported Parkinson’s symptoms assessment
- Each patient recorded multiple times

Can you build a model to detect the level and severity of Parkinson’s using the voice recordings?
- Can extract your own features from the voice data in addition to the 39 provided
(7) Relationship Affinity and Type (Anand)

- 4 kinds of people data from Refresh.io
  - Facebook relationships
  - LinkedIn connections
  - Twitter followers
  - Email social network

- Email data includes
  - From, To, Time, Subject

- User Id obfuscated but same user maps to same id on all networks
Given two people, we’d like to know the affinity of their relationship
- “Strong” or “weak”
- Like Facebook’s “edge rank” but with more data!

Also categorize relationships
- Personal
- Business
- Asymmetrical (one node is dominant)
Data from Yume, a leading Video Ad Network

Goal: identify the same user across multiple sessions and devices (phone, tablet, PC)
  - Can’t rely on cookies any more!

Data from billions of sessions for millions of users
  - Rich data with many fields, see separate slides for details
Send in your proposals

- Proposals due Tuesday 3/25 midnight

- For more detail on a dataset or problem, please contact the appropriate instructor
  - Jure Leskovec (jure@cs)
  - Anand Rajaraman (datawocky@gmail.com)
  - Jeff Ullman (ullman@gmail.com)