RUTGERS
THE STATE UNIVERSITY
OF NEW JERSEY

01:198:443
More On Class Projects

Tina Eliassi-Rad

tina@eliassi.org

Class Project
(40% of Your Grade)

- You will solve a data-science problem from data preparation to data product
- The class project can be done either individually or in groups of two
- The class project has three parts:
  - Proposal report and pitch (10%): 2 pages maximum and 4-minute in-class pitch (no slides are needed)
  - Class presentation (12%): 6-minute in-class presentation and preparing slides for the presentation
  - Final report (18%): 6 pages maximum

Proposal Report & In-class Pitch (10% of Your Grade)

- Due on Thu 3/12
- Proposal report: 2 pages maximum
- In-class pitch: 4 minutes maximum
- Your proposal report and pitch should include answers to the following questions:
  1. What is the problem?
  2. Why is it interesting and important?
  3. Why is it hard? Why have previous approaches failed?
  4. What are the key components of your approach?
  5. What data sets and metrics will be used to validate the approach?
- By the due date (Thu 3/12), you should have identified and gathered the data.

In-class Project Presentation (12% of Your Grade)

- You will have 6 minutes to present your project in class.
- Your presentation will be on either Mon 4/27 or Mon 5/4.
  - I will randomly assign projects to these dates.
  - You will be notified two weeks in advance as to which date your presentation will be.
- Use the following guideline when preparing your slides
  - Title & Author(s) (1 slide)
  - Motivation & Problem Statement (1 slide)
  - Related Work (1 slides)
  - Methods/Algorithms (1-2 slides)
  - Data Characterization & Results (2-3 slides)
  - Summary & future work (1 slide)
- See sample presentation uploaded to the Sakai site (under Resources, week 1).

Course Project – Final Report (18% of Your Grade)

- 6 pages maximum
- Due on Mon 5/11
- For guidance on writing the final report, see slide 70 of Eamonn Keogh's KDD'09 Tutorial on **How to do good research, get it published in SIGKDD and get it cited!**
  - http://www.cs.ucr.edu/~eamonn/Keogh_SIGKDD09_tutorial.pdf
- Follow **ACM formatting guidelines**
  - http://www.acm.org/sigs/publications/proceedings-templates

Class Project (40%)
Scrap & Analyze a Large Data Set

- What it should be
  - More than just measurements
  - An interpretation of measurement results
  - Discovery of new information via classification, clustering, dimensional reduction, frequent itemsets, etc
  - Visualizations of all or part of the data set that points out a particular feature
  - Qualitative comparison with other data sets
- What it should not be
  - A literature review

More on Class Projects

- Discovering interesting relationship within a significant amount of data
- Having some original ideas that extend/build on what we learn in class
  - Extend/improve/speed-up some existing algorithm
  - Define a new problem and solve it

When deciding on a project, think about the following?

- What is the problem that you are solving?
- What data will you use (where will you get it)?
- How will you solve the problem?
  - Which algorithms/techniques will you use?
  - Be as specific as you can.
- How will you evaluate your solution / measure success?
- What do you expect to submit?

Some Data Sets

- You can find a bunch of datasets at these popular sites:
  - http://www.kaggle.com/competitions
  - http://datamob.org/datasets
  - http://archive.ics.uci.edu/ml/
  - http://kdd.ics.uci.edu/
  - http://lib.stat.cmu.edu/datasets/
  - http://snap.stanford.edu/data/
  - http://data.gov
- Obviously, this is not an exhaustive list.

Analyzing Social Media Data

- If you are interested in analyzing social media data, I recommending reviewing these
  - http://snap.stanford.edu/proj/socmedia-kdd
  - http://www.umiacs.umd.edu/conferences/sbp2011/nitin_agarwal_tutorial.pdf
    - (see slides 44 to 72 for data collection APIs)

Each Web 2.0 site has its own API.

- Here are some examples:
  - Twitter Archivist: http://archivist.visitmix.com
  - Delicious Developer Page: http://delicious.com/developers
  - YouTube's Developer Page: https://developers.google.com/youtube/
  - Flickr's API Gardens: http://www.flickr.com/services/api/

Some Resources on the Web

- Data & ideas:
  - http://web.stanford.edu/class/cs341/data.html
- Example of projects:
  - http://web.stanford.edu/class/cs341/projects.html

More Data

- Wikipedia
- IM buddy graph
- Yahoo Altavista web graph
- Stanford WebBase
- Twitter Data
- Blogs and news data
- Netflix
- Restaurant reviews
- Yahoo Music Ratings

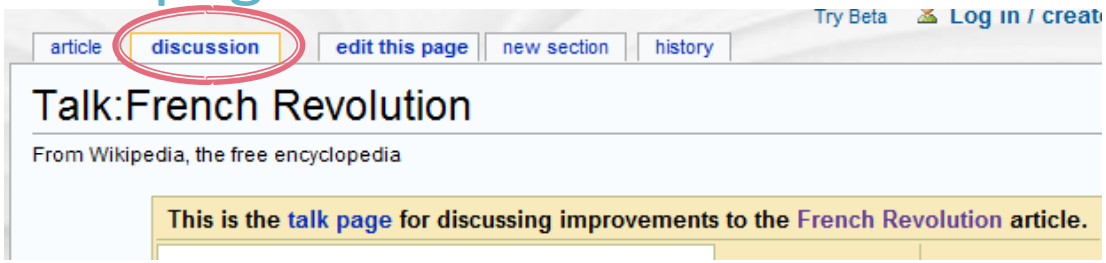
Wikipedia (1)

- Complete edit history of Wikipedia until January 2008
- For every single edit the complete snapshot of the article is saved
- Each **page** has a **talk** page:

Slide from Jure Leskovec and Anand Rajaraman

Wikipedia (2)

- Talk page:

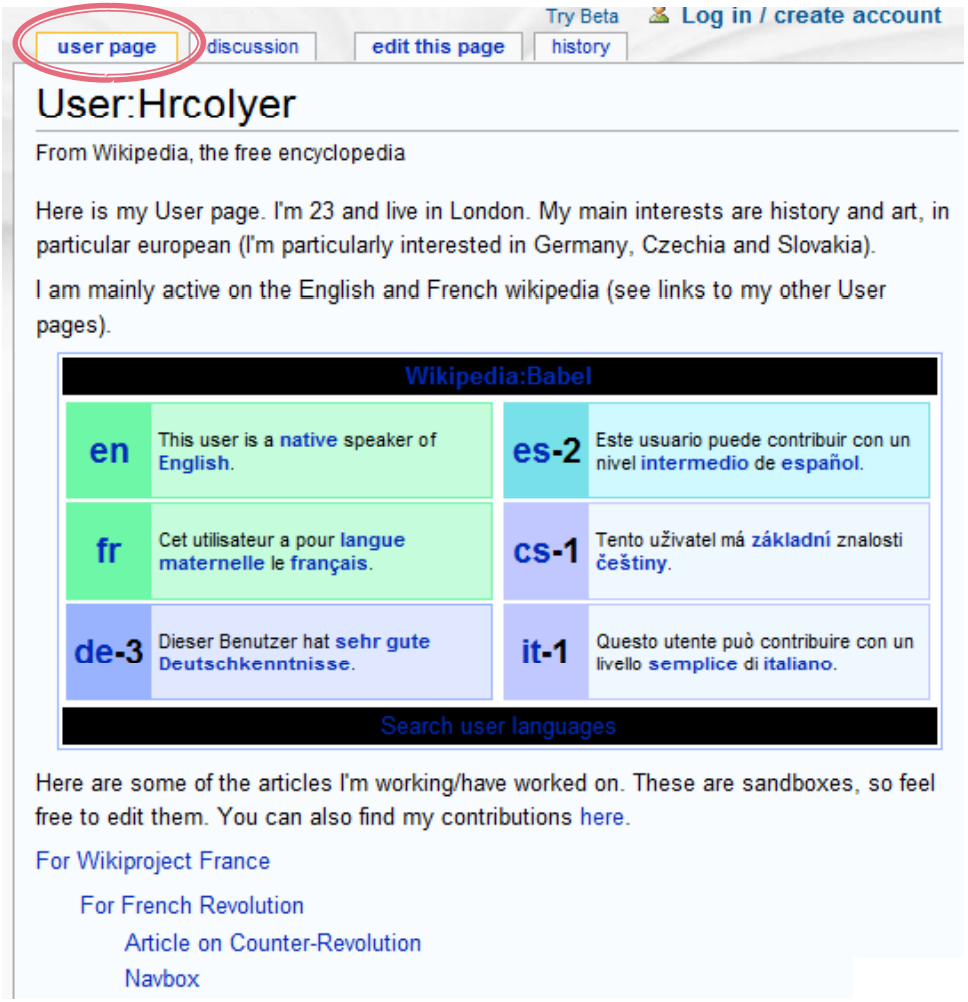


- Editors discuss things like:

Slide from Jure Leskovec and Anand Rajaraman

Wikipedia (3): User Pages

- Every registered user has a page:

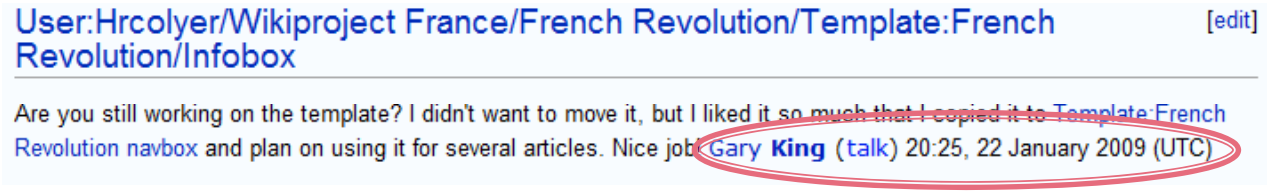


Slide from Jure Leskovec and Anand Rajaraman

Wikipedia (4): User Talk Pages

- Every user's page has a talk page:

- Users discuss things:



Slide from Jure Leskovec and Anand Rajaraman

Wikipedia (5): User Pages

Slide from Jure Leskovec and Anand Rajaraman

Wikipedia (6): Data Format

```
<page>
    <title>Anarchism</title>
    <id>12</id>
    <revision>
      <id>18201</id>
      <timestamp>2002-02-25T15:00:22Z</timestamp>
      <contributor>
        <ip>Conversion script</ip>
      </contributor>
      <minor />
      <comment>Automated conversion</comment>
      <text xml:space="preserve">''Anarchism'' is the political
        theory that advocates the abolition of all forms of
        government.
        ...
      </text>
    </revision>
    <revision>
      <id>19746</id>
      <timestamp>2002-02-25T15:43:11Z</timestamp>
      <contributor>
        <ip>140.232.153.45</ip>
      </contributor>
      <comment>*</comment>
      <text xml:space="preserve">''Anarchism'' is the political
        theory that advocates the abolition of all forms of government.
        ...
```


Slide from Jure Leskovec and Anand Rajaraman

Wikipedia (7): Ideas

- Complete edit and talk history of Wikipedia:
  - How do articles evolve?
    - Use string edit distance like approach to measure differences between versions of the article
    - Model the evolution of the content
  - Which users make what types of edits?
    - Big vs. small changes, reorganization?
    - Suggest to a which user should edit the page?
  - How do users talk and then edit same pages?
    - Do users first talk and then edit?
    - Is it the other way around?
  - Suggest users which pages to edit

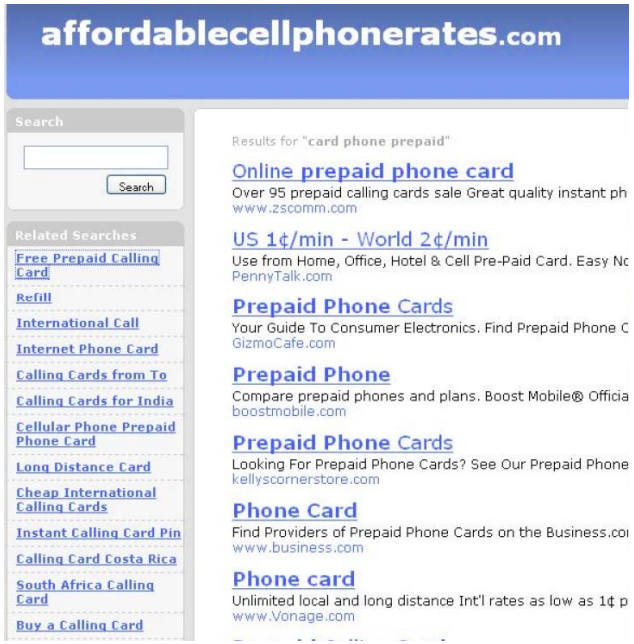
Slide from Jure Leskovec and Anand Rajaraman

Yahoo Altavista Web Graph

- Altavista web graph from 2002:
  - Nodes are webpages
  - Directed edges are hyperlinks
  - 1.4 billion public webpages
  - Several billion edges
  - For each node we also know the page URL

Altavista: Ideas (1)

- SPAM:
  - Use the web-graph structure to more efficiently extract spam webpages
  - Link farms
  - Spider traps
- Personalized and topic-sensitive PageRank



Slide from Jure Leskovec and Anand Rajaraman

Altavista: Ideas (2)

- Website structure identification:
  - From the webgraph extract “websites”
  - What are common navigational structures of websites?
    - Cluster website graphs
    - Identify common subgraphs and patterns
  - What are roles pages/links play in the graph:
    - Content pages
    - Navigational pages
    - Index pages
  - Build a summary/map of the website

Slide from Jure Leskovec and Anand Rajaraman

Twitter: Data

- 50 million tweets per month starting June 2009 (6 months)
- Format:

```
T       2009-06-07 02:07:42
U       http://twitter.com/redsoxtweets
W       #redsox Extra Bases: Sox win, 8-1: The Rangers
spoiled Jon Lester's perfecto and his shutout..
http://tinyurl.com/pyhgwy
```

- Two important things:
  - URLs
  - Hash-tags

Slide from Jure Leskovec and Anand Rajaraman

Twitter: Ideas

- Trending topics: raising, falling
- Inferring links of the who-follows-whom network
- What is the lifecycles of URLs and hash-tags?
- Finding early/influential users?
- Clustering tweets by topic or category
- Sentiment analysis – are people positive/negative about something (a product?)

Memetracker Data

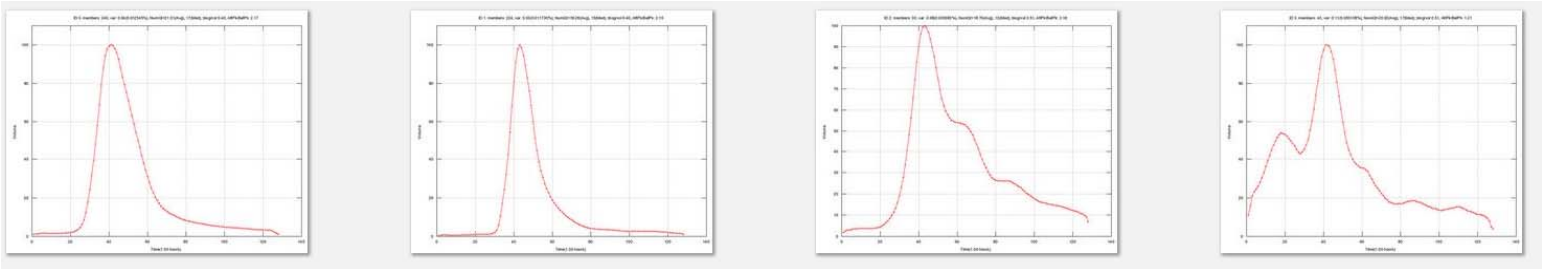
- More than 1 million newsmedia and blog articles per day since August 2008
- Extract phrases (quotes) and links
- http://memetracker.org
- Format:

```
P       http://cnnpoliticalticker.wordpress.com/2008/08/31/mccain-defends-
palins-experience-level
T       2008-09-01 00:00:13
Q       dangerously unprepared to be president
Q       even more dangerously unprepared
Q       understands the challenges that we face
Q       worked and succeeded
Q       still to this day refuses to acknowledge that the surge has
succeeded
L       http://www.cnn.com
```


Slide from Jure Leskovec and Anand Rajaraman

Memetracker: Ideas (1)

- Find all variants (mutations) of the same phrase – cluster phrases based on edit distance and time:
  - lipstick on a pig
  - you can put lipstick on a pig
  - you can put lipstick on a pig but it's still a pig
  - i think they put some lipstick on a pig but it's still a pig
  - putting lipstick on a pig
- Temporal variations of the phrase volume



Slide from Jure Leskovec and Anand Rajaraman

Memetracker: Ideas (2)

- Predict the popularity of a phrase over time
- How does information mutate/change over time?
- Which media sites are the most influential? Build a predictive model of site influence
- Which nodes are early mentioners, late comers, summarizers?
- Sentiment analysis – are people positive/negative about something (news, a product)
- Create a model of political bias (liberal vs. conservative)
- What is genuine news, what are genuine phrases and what is spam?

Slide from Jure Leskovec and Anand Rajaraman

IM Buddy Graph: Data

- A large IM buddy graph from March 2005
- 230 million nodes
- 7,340 million undirected edges

- Limitations:
  - Only have the buddy graph with random node ids
  - No communication or edge strength

IM Buddy Graph: Ideas

- Find communities, clusters in such a big graph
- Count frequent subgraphs
- Design algorithms to characterize the structure of the network as a whole

Slide from Jure Leskovec and Anand Rajaraman

Recommendations: Data

- Movie ratings:
  - Netflix prize dataset:
  - http://www.netflixprize.com/
- Yahoo Music ratings:
  - Yahoo Music user ratings of songs with artist, album and genre information
  - 717 million ratings
  - 136,000 songs
  - 1.8 users
- Restaurant reviews

Slide from Jure Leskovec and Anand Rajaraman

Recommendations: Ideas

- Collaborative filtering:
  - Predict what ratings will user give to particular songs/movies, i.e., which sons will he/she like?
- Supplement the data with additional data sources:
  - Movies -- IMDB
  - Playlists from the web
  - Lyric (text of the song)
- Include taste, temporal component, diversity into the model

Many Other Ideas and Datasets

- New York Times articles since 1987
  - Article are manually annotated by subject categories and keywords
  - Entity or relation extraction
  - Extract keywords, predict article category

- Don't feel limited by these
- You can collect the dataset yourself
- And define the project/question yourself

Slide from Jure Leskovec and Anand Rajaraman

Project Titles from a Similar Course Taught at Stanford

1. Frequency-Domain Characterization of Trending Topics
2. A Music Recommendation System based on Yahoo! Data Corpus
3. Identifying Trending Topics on Twitter
4. Wikipedia Vandalism
5. Product Offer Comparison across Different Merchants
6. Extracting Information from Yelp Reviews
7. Exploring Methods of De-Novo Short Read Assembly Using MapReduce
8. Topic Chaining and Phrase Linking
9. Understanding Correlations between Product Reviews and Ratings
10. Finding the Social Roots of Controversy in Wikipedia
11. Techniques to Improve Detection of Trending Topics on Twitter
12. Mining Hospital Records for Predicting Patient Drop-off
13. Social Information Engine: Data Mining Twitter for Product Recommendations
14. Comparing the Impact of Cross-disciplinary and Cross-institutional Academic Research: An Exploration of the ISI Web of Science Database

http://www.stanford.edu/class/cs345a/project.html

Project Titles from a Similar Course Taught at Stanford

15. Woodstock: Using Twitter tweets' sentiments to predict stock price change
16. Book Recommendation System
17. Seven years of Wikipedia's Revision History as a Time dependent Graph: A Love Story
18. Adaptive Locality Sensitive Hashing for Recommending Twitter Followers
19. Combining Content Filtering and Collaborative Filtering for the Netflix Prize
20. Twitter #Hashtags
21. Collaborative Filtering on Netflix Challenge
22. A Music Recommendation System
23. Content Based Auto-tagging of Flickr Images using ImageWebs
24. A Data Mining Based Approach to Determining Causal Associations Between Drugs and Condition
25. Twitter Personal Newspaper
26. WikiSuggest: A Suggestion Engine for Editors on Wikipedia
27. Hashtags on Twitter

http://www.stanford.edu/class/cs345a/project.html

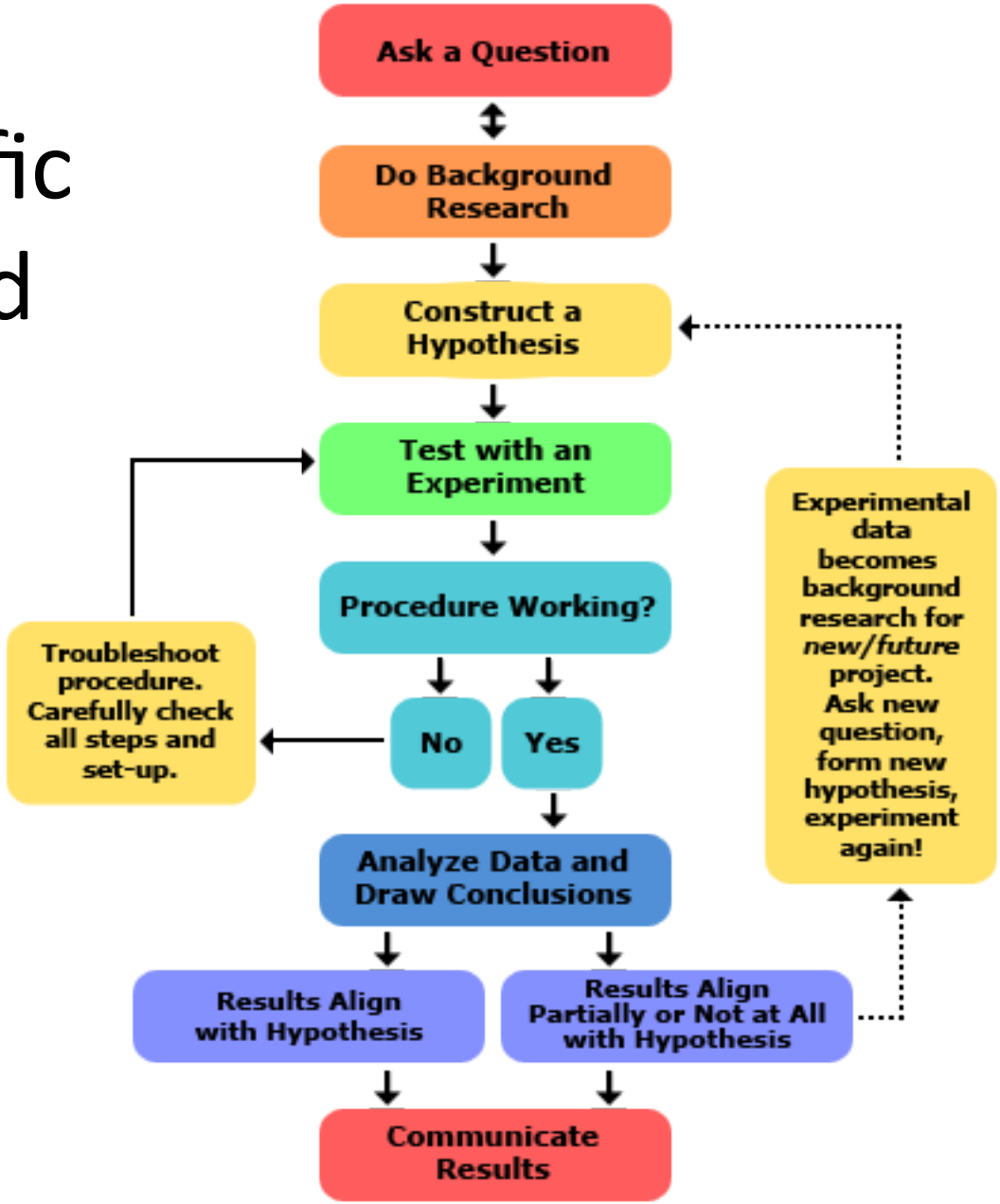
Mining Mobile Data

- Mining mobile data:
  - Abstract: http://mobilemining.clusterhack.net/
  - Slides: http://mobilemining.clusterhack.net/dl
  - Resources (including data): http://mobilemining.clusterhack.net/info

More on Course Project

- Start thinking about your project early
- Come talk to me if you have questions
- Don’t forget the scientific method

The Scientific Method



http://www.sciencebuddies.org/science-fair-projects/project_scientific_method.shtml#overviewofthescientificmethod