# Course Information

**Instructor:** Jure Leskovec
Office Hours: Wednesdays 9-10am, Gates 418

**Lectures:** 9:30AM - 10:45AM Tuesday and Thursday in NVidia, Huang Engineering Center

**Course website:** http://cs246.stanford.edu

**Contact:**
- E-mail us at cs246-win1415-staff@lists.stanford.edu
- Use Piazza to post questions: http://piazza.com/stanford/winter2015/cs246
- SCPD students can attend office hours remotely via a Google Hangout; the link will be posted on Piazza just before the office hours start.

**TAs and office hours:** See the course website for times and locations.

## Topics

- MapReduce and Hadoop
- Frequent itemsets and Association rules
- Near Neighbor Search in High Dimensions
- Locality Sensitive Hashing (LSH)
- Dimensionality reduction: SVD and CUR
- Recommender systems
- Clustering
- Analysis of massive graphs

- Link Analysis: PageRank, HITS
- Web spam and TrustRank
- Proximity search on graphs
- Large-scale supervised machine learning
- Mining data streams
- Learning through experimentation
- Web advertising
- Optimizing submodular functions

## Assignments and grading

- Four problem sets requiring coding and theory (40%)
- Final exam (40%)
- Gradiance quizzes (20%)
- Piazza and course participation (up to 1% extra credit)

## Homework policy

**Questions** We try very hard to make questions unambiguous, but some ambiguities may remain. Ask (i.e., post a question on Piazza) if confused or state your assumptions explicitly. Reasonable assumptions will be accepted in case of ambiguous questions.

**Honor code** We take honor code extremely seriously (http://stanford.io/1F3TWN0). The standard penalty includes a one-quarter suspension from the University and 40 hours of community service.
We strongly encourage students to form study groups. Students may discuss and work on homework problems in groups. However, each student must write down the solutions and the code independently. In addition, each student should write down the set of people whom s/he interacted.

**Late assignments** Each student will have a total of 2 late periods to use for homeworks. One late period expires at 5pm. (If the assignment is due on Thursday 5pm then the late period expires next Tuesday 5pm.) No assignment will be accepted more than one late period after its due date.

**Assignment submission** All students (SCPD and non-SCPD) submit their homeworks via Scoryst (http://www.scoryst.com). Students can typeset or scan their homeworks.

Students also need to upload their code at http://snap.stanford.edu/submit. Put all the code for a single question into a single file and upload it. Refer to the course FAQ for more info.

**Regrade requests** We take great care to ensure that grading is fair and consistent. Since we will always use the same grading procedure, any grades you receive are unlikely to change significantly. However, if you feel that your work deserves a regrade, email the course staff mailing list within a week of receiving your grade. However, note that we reserve the right to regrade the entire assignment. Moreover, if the regrade request is unjustified and thus not honored, then every future unsuccessful regrade request will be penalized 5 points.

**Gradiance** Quizzes are posted on Friday afternoon and due exactly a week later (hard deadline Friday 11:59pm Pacific time). Once the deadline has passed students will not be able to submit the quiz.

# Prerequisites

Students are expected to have the following background (recitation sessions will refresh these topics):

- The ability to write very non-trivial computer programs (at a minimum, at the level of CS107). Good knowledge of Java will be extremely helpful since most assignments will require the use of Hadoop/Java.
- Familiarity with basic probability theory is essential (at a minimum, at the level of CS109 or Stat116).
- Familiarity with writing rigorous proofs (at a minimum at the level of CS 103).
- Familiarity with basic linear algebra (e.g., any of Math 51, Math 103, Math 113, CS 205, or EE 263).
- Familiarity with algorithmic analysis (e.g., CS 161).

# Materials

Notes and reading assignments will be posted on the course web site. Reading for the class will be from:

- Mining Massive Datasets by J. Leskovec, A. Rajaraman, J. Ullman (PDFs at http://mmds.org).

# Important dates

| Assignment | Out Date | Due Date (all 5pm) |
|---|---|---|
| Hadoop tutorial | now | Jan 13 |
| Assignment 1 | Jan 8 | Jan 22 |
| Assignment 2 | Jan 22 | Feb 5 |
| Assignment 3 | Feb 5 | Feb 19 |
| Assignment 4 | Feb 19 | Mar 5 |
| Alternate final | | Mar 16, 7-10pm |
| Final exam | | Mar 20, 12:15-3:15pm |

We will also hold two review sessions in the first two weeks of the course (sessions will be video recorded):

- Review of basic probability. Friday, January 9, at 4:15-5:30pm in Gates B01.
- Review of basic linear algebra. Friday, January 16, at 4:15-5:30pm in Gates B01.

# Next steps for students

- Register for Piazza: http://piazza.com/stanford/winter2015/cs246
- Register for Gradiance: http://www.newgradiance.com/services class token B343F7F0
- Register for Scoryst: https://scoryst.com/enroll/suyVLAjbns
- Download Hadoop VM, start the tutorial: http://cs246.stanford.edu/homeworks/tutorial.pdf