Dates:

Oct 26: project proposals due.
Nov 19: progress reports due.
Dec 10: Final report due. I will not start grading until 15th morning.

In class, we have learned the ABCs of big data, for example we will learn stochastic gradient descent methods, momentum methods, dual coordinate descent approaches, hashing and sketching, divide and conquer based methods. For the project, I expect you to use what you have learned in class to explore an interesting machine learning problem in the context of a real-world application. While you should explore different theoretical and algorithmic ideas, do not forget the big data component. Since this is a big data class, you must either have a dataset that is high dimensional, or that is very large, e.g. has a million or more datapoints, or both. Some experiments on toy data would result in a very low score.

The grading of the final project is split amongst three deliverables:

- A project proposal (20% of the grade).
- A progress report (20% of the grade).
- A final report (60% of the grade).

Your final report will be evaluated by the following criteria:

- Scientific merit: Did you explore your question with sound reasoning? Can you draw quantitative conclusions from you work? Are you taking a justifiably simple approach or, if you are choosing a more complicated method, do you have sound reasoning for doing this? If your project is on the creative axis, then you must still find a way to quantitatively judge your outcomes.
- Technical depth: How technically challenging was what you did? Did you use a package or write your own code? It is fine if you use a package, though this means other aspects of your project must be more ambitious. How challenging was dealing with the data set that you used? How challenging was your project in scope? How detailed was it? Did you compare with other baselines?
- Presentation and clarity: How well did you explain what you did, your results, and interpret the outcomes? Did you use good graphs and visualizations? How clear was the writing? Did you justify your approach? Did you present related ideas and related work clearly?

Format: Use the NIPS format for all submissions. This format should be identical to a NIPS paper (8 pages maximum in NIPS format, including references; this page limit is strict)

Regression:

We never really talked about graphical lasso or fused lasso in class. As it turns out there are fast algorithms with provable guarantees for fused lasso, read "The DFS fused lasso: Linear-time denoising over general graphs

Divide and conquer approaches:

Can you come up with methods that divide the data and combine the results to do clustering/regression on enormous datasets?

- Parallel kmeans/spectral clustering
  http://ntucsu.csie.ntu.edu.tw/~cjlin/papers/psc08.pdf
- Parallel Lasso
  http://iie.fing.edu.uy/~gmateos/pubs/dlasso/D_LASSO_TSP.pdf
- Parallel clustering with core-sets
  http://www.cs.princeton.edu/~yingyul/distributedClustering.pdf
- Divide and conquer Lasso
  http://www3.stat.sinica.edu.tw/sstest/oldpdf/A24n49.pdf

Spectral clustering of large image datasets.

- Fast approximate spectral clustering, Donghui Yan, Ling Huang, and Michael I. Jordan, KDD09 : https://people.eecs.berkeley.edu/~jordan/papers/yan-huang-jordan-kdd09.pdf
- Parallel spectral clustering in distributed systems, Wen-Yen Chen, Yangqiu Song, Hongjie Bai, Chih-Jen Lin, and E. Y.Chang, Pattern Analysis and Machine Intelligence 2011. https://ieeexplore.ieee.org/document/5444877/
- Spectral grouping using the nystrom method, Charless Fowlkes, Serge Belongie, Fan Chung, and Jitendra Malik, Pattern Anal. Mach. Intell. 2004 http://people.eecs.berkeley.edu/~malik/papers/FBCM-nystrom.pdf
- Fast spectral clustering with random projection and sampling, Tomoya Sakai and Atsushi Imiya, MLDM09.
  http://link.springer.com/chapter/10.1007/978-3-642-03070-3_28
- A Fast Incremental Spectral Clustering for Large Data Sets, Tengteng Kong, Ye Tian, Hong Shen, PDCAT11.

https://ieeexplore.ieee.org/document/6118531

Network models and inference on large scale

- Blockmodels survey:
  https://www.cs.umd.edu/class/spring2008/cmsc828g/Slides/block-models.pdf

- Mixed membership block models:
  http://jmlr.csail.mit.edu/papers/volume9/airoldi08a/airoldi08a.pdf

- Dynamic social network analysis using latest space models. P. Sarkar and A. W. Moore. https://dl.acm.org/citation.cfm?id=1117459

- Implement stochastic block model to analyze communities using Twitter data or web graph see "Gopalan, Prem. "Scalable Inference of Overlapping Communities." Neural Information Processing Systems, 2012."

- Semidefinite programs for network clustering. See "On semidefinite relaxations for the block model". https://arxiv.org/abs/1406.5647

- Fast nonconvex alternatives to semi-definite programs using Burer Monteiro methods. See "The non-convex Burer-Monteiro approach works on smooth semi-definite programs" by Nicolas Boumal, Vladislav Voroninski, Afonso S. Bandeira. https://arxiv.org/abs/1606.04970

Large-scale image classification or retrieval:

- Computing sift or other feature extraction for huge image datasets (data-parallel, ideal for Map-Reduce)
- Vector quantization and clustering (not data-parallel, may use GraphLab) Bag-of-words histogram representation + SVMs/logistic regression for classification or fast-approximate nearest neighbor methods for retrieval.

Parallel algorithm implementation and analysis:

- Parallelize Gibbs sampling in LDA (see distributed LDA paper) or mixture of Gaussian application using GraphLab. Amr Ahmed, Mohamed Aly, Joseph Gonzalez, Shravan Narayanamurthy, Alex Smola (2012). "Scalable Inference

in Latent Variable Models." Conference on Web Search and Data Mining (WSDM)

- Compare Shotgun, Hogwild! and distributed averaging on a range of datasets.

Datasets

- **Image Datasets**
  - ImageNet
  - 80 Million Tiny image dataset (from Antonio Torralba's group)
  - CIFAR-10
- **Text Datasets:**
  - Open web directory
  - Wikipedia

- **Network Datasets and collections:**
  - A collection of moderately-sized to larger datasets
  - Twitter 2010 graph with 1.4B edges
  - A collection of moderately sized datasets
  - Trec datasets
  - KDD cup datasets
  - Click through rate dataset
  - Lemur project
  - The Graph 500 Benchmark

Tools

- GraphLab: http://graphlab.org
- GraphChi: http://graphchi.org
- Hadoop: http://hadoop.apache.org/
- Spark: http://spark-project.org/