STAT4601/STAT5703 Data Mining I
Research Paper: due March 5, 2020

- You may work in teams of 3 (if all are grad students) or 4 (if group includes undergrad students). (No one is to work alone on a topic without my prior approval.)
- It is strongly suggested that you choose team members with different complementary strengths.
- Each group is to email me their first, second and third choice of topic; topics will be assigned by me by February 4, 2020 . I will email you your assigned topic. Every effort will be made to accommodate your stated interests but no topic will be assigned to more than one group.

Topics (a link may be provided to get you started on your research):

1. Document clustering/Text categorization for disputed authorship
Text categorization concerns the automatic assignment of documents to categories. The idea is learn a categorization algorithm from a set of hand-labelled documents. There are lots of interesting statistical/data mining ideas in this area. One application concerns disputed authorship: given samples of particular author's works, assign disputed samples to the right author. The classic work in this area is very old and updating it would make for a nice project. See:
https://www.jstor.org/stable/2283270?seq=1#metadata_info_tab_contents

2. Bayesian model averaging (BMA) versus Mixtures of Experts
Compare the predictive performance of Bayesian model averaging to so-called mixtures-of-experts models.
BMA software is at http://www.research.att.com/~volinsky/bma.html
ME software is at
https://www.researchgate.net/publication/260707711_Twenty_Years_of_Mixture_of_Experts

3. Big Bayesian Networks'scaling algorithms for learning Bayesian networks
Nice tutorial on David Heckerman's home page http://www.research.microsoft.com/~heckerman/

4. Social Network Analysis or Dynamic Network Analysis - see
http://www.slideshare.net/gcheliotis/social-network-analysis-3273045
or
http://www.casos.cs.cmu.edu/publications/protected/2000-2004/2003-2004/carley_2003_dynamicnetwork.pdf

5. Partial Least Squares - see
http://www.tandfonline.com/doi/abs/10.1207/s15328031us0304_4?journalCode=hzzk20

6. Text mining - see http://www.charuaggarwal.net/text-content.pdf

7. Support vector machines (SVM) - see
http://docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html

8. PCA and ICA applications - see
http://www.cc.gatech.edu/~isbell/reading/papers/draper_cviu03.pdf

http://www.cs.jhu.edu/~ayuille/courses/Stat161-261-Spring14/HyvO00-icatut.pdf

9. Weighted Association Rule Mining
https://eprints.soton.ac.uk/257986/1/331.tao.pdf
https://pdfs.semanticscholar.org/af64/b2df8f2a5567196681acbd3710d8e4cfdcbf.pdf

10. Steganalysis - see
https://pdfs.semanticscholar.org/2724/3ae662027ff79607c556c3127a10e79461b9.pdf

11. Random Forests - see https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf

12. Deep Learning
https://www.nature.com/articles/nature14539
https://www.mathworks.com/discovery/deep-learning.html

Your **e-paper** must contain:

- a statement of the problem you are researching,
- a short literature review,
- a discussion of the technical material behind the analysis you choose to undertake,
- an application of your methodology to an appropriate dataset,
- any code used must appear as an appendix to your paper and in a format that I can run,
- a list of references used,
- a statement outlining who is responsible for what part of the work done,
- slides to accompany your e-paper, so that you can present the results of your work in a class presentation. BE PREPARED TO GIVE A 10-15 MINUTE presentation of your work.