

## Syllabus: Math 561 Advanced Data Analytics

**Instructor:** Brian Steele      Office: Math 314  
Phone: 243-5396      email: brian.steele@umontana.edu  
Office hours      MW 2 pm, T 1 pm & by appointment

**Course Format:** Meetings: Monday, Wednesday, Friday 9:00-9:50 a.m., Math 306. Class time is split (75/25) between lecture and individual and small group work on programming and algorithm development.

### **Learning Outcomes**

1. Understand the theory and foundations of data reduction and information extraction (e.g., associative statistics and data mapping).
2. Develop understanding and practical experience regarding reduction of massive data sets and data streams.
3. Understand the theory and mechanics of distributed computing.
4. Ability to formulate and implement algorithms for processing massively large data sets. Ability to compute histograms, correlation matrices, and linear regression estimators using massively large data sets.
5. Understand the objectives of multiple regression, methods for fitting models, and examining model assumptions. Derive estimators from first principles. Ability to carry out and interpret hypothesis tests. Understand the details of fitting qualitative and quantitative predictors.
6. Proficiency using Python, Hadoop/MapReduce and R.
7. Proficiency in data analytic algorithm design. Hadoop/MapReduce

**Course Content:** Algorithms are the machinery behind the data analytics (the subject matter of the course). To be good at data analytics, one must be competent at programming and have experience with the data and the algorithms of data science. To be expert at data analytics requires an understanding of the foundations and principles behind the algorithms. Why? Applying the algorithms to real problems often requires adapting existing algorithms and creating new ones. To get to the point where innovation is not intimidating but instead an opportunity for creativity, students will work with a set of prototypical algorithms that span much of data analytics. Graduate students will derive and develop key components of the algorithms. Algorithms will be implemented in Python or R and applied to practical problems involving publicly available data sets.

M 561(G) co-convenes with M 461(U). Course content differs however. In particular, M 561 involves theoretical and foundational learning outcomes absent from M 461. Principally, graduate students will derive estimators from theoretical principles and create innovations to core algorithms.

We shall cover most, but not all of the material in chapters 1 through 10 of the textbook. The main topics are

1. Data mappings and the concepts of data reduction. Similarity measures and distance metrics. List, set, and dictionary comprehension.
2. Scalable algorithms and associative statistics. Computing univariate and multivariate statistics from massively large data sets.
3. Distributed computing using MapReduce algorithms and the Hadoop environment. Basics of the command line. Utilizing Elastic MapReduce.

4. Linear regression for prediction (using R).
5. Data visualization.
6. Cluster analysis. Hierarchical and  $k$ -means methods.
7.  $k$ -nearest neighbor prediction
8. Multinomial Naive Bayes prediction

**Textbook: Algorithms for Data Science (ISBN-10: 3319457950)**

**Prerequisites:** Stat 341 or M 421 and at least 2 upper division mathematics courses

**Homework:** Homework exercises emphasizing applications of the algorithms will be assigned most weeks. Occasionally, a quiz will replace the weekly homework assignment. Students are to complete approximately 4 tutorials per month. Tutorials are oriented toward gaining proficiency in programming and algorithm design.

Homework exercises tend to be more difficult than those assigned to M 461 students as expectations and grading are commensurate with graduate standing (expectations for undergraduates are less). Graduate students that are not in Mathematics will be responsible for problems commensurate with their major and interests. Hence, their problems usually are data analysis-type exercises.

In addition students are required complete 4 tutorials per month. Tutorials are oriented toward gaining proficiency in programming and algorithm design.

**Project:** Graduates are responsible for an analysis of a complex data set. You should work on it throughout the semester. We'll determine the topic by the end of the second week. Students may work in groups of no more than three individuals. Graduate students are responsible for a written paper describing the project and oral presentation. (Each student is to write their own paper). The final meeting time and the presentation of projects) is Wednesday, December 11, 12:10 p.m. The presentation and paper should cover objectives, data origins, methods, and results.

**Grading:** Your course grade will be based on homework, tutorials, and the project (no final exam). Students are responsible for completing 4 tutorials (usually) per month (due on the last day of each month except December). Homework assignments and tutorials are worth 25% and 50% of the course grade, respectively. The project is worth 25%. Expectations and grading are commensurate with graduate standing (undergraduates generally are not asked to prove theorems).