The Data Analysis Minor at Georgetown College

Definition of Terms

We define *Data Analysis* as the combination of all phases of the process of dealing with data: from its collection, storage and management through the extraction of knowledge from it and finally to the communication of that knowledge. It is understood that the data is often so large or messy that the assistance of computers is required at various points in the process of dealing with it.

The following Venn diagram identifies Data Analysis as the intersection of three different areas of knowledge: Statistics, Programming/Hacking Skills, and a specific field, known as the "domain of application", from which the data itself derives.

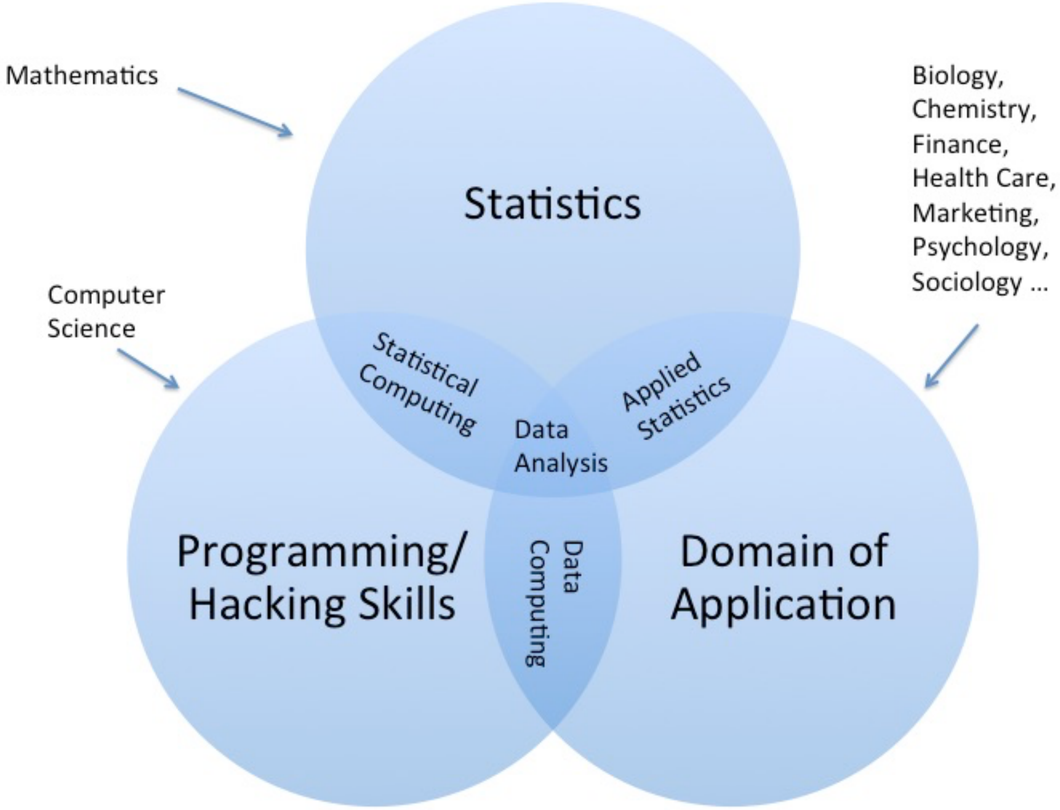

Figure 1: Venn Diagram showing Data Analysis in terms of intersections of fields.

The field of Statistics emerges from and is undergirded by the liberal arts discipline known as Mathematics. The mathematical background for this field is provided in MAT 125, MAT 225 and MAT 331, and Statistics itself is treated in MAT 111 and MAT 337 (see the section on minor requirements below).

Computer Science is the liberal arts discipline that provides the context for computer programming and the “hacking” skills associated with computer use. Programming/kacking are addressed in CSC 303, CSC 400, and MAT 337 (see the section on minor requirements below).

Any number of other liberal arts disciplines or areas of human activity—Biology, Chmistry, Finance, Sociology, Marketing, Health Care, etc.—could constitute the specific domain of application.

At the intersection of Programming/Hacking and Statistics is the field known as Statistical Computation. This areas is addressed primarily in CSC and MAT 337.

Data Computing comprises the intersection of Programming and a domain of application. The Data Computing option within the Data Analysis minor corresponds to this particular intersection. The minor includes enough instruction in statistics to permit the student to conduct an analysis of data, but the focus of Data Computing option is on the computing skills required to gather, store, manage and access large quantities of data, and to develop web-based applications for the collection of data and the dissemination of the results of data analysis. This area is addressed primarily in CSC 303 and CSC 400.

Applied Statistics comprises the intersection of Statistics and a domain of application. The emphasis here is on the application of statistical methods, derived from mathematics, to the extraction of knowledge from data. MAT 337 is the primary courses in this area.

Courses in the Minor

The minor has four required courses totalling 12 credit hours, plus a 3-credit course selected from a small list of possibilities.

The four required courses are as follows:

- **MAT 111 Elementary Probability and Statistics.** This is the College’s basic course in Statistics. It has no formal prerequisites and it is offered every semester.
- **CSC 303 Fundamentals of Data Computing.** This course focuses on data analysis in settings where the data is so large, dispersed or messy that machine-processing is required to gather, clean and transform it into forms suitable for analysis. We also study computer-based techniques for the analysis of such data, including machine data visualization and modeling with data. Principles of reproducible research are studied and put into practice throughout the course. The course is offered every Fall semester. There are some prerequisites: the student should have already taken either

MAT 111 or CSC 115 Computer Science I, or should obtain permission from the instructor prior to enrolling. The course teaches the required statistics and programming from scratch, so we have found that motivated students do well in the course without any prior statistics or programming courses.

- **MAT 125 Calculus I.** This is one of threeseriously mathematical courses in the minor. It is offered every semester. As far as prerequisites are concerned, students should have a solid background in pre-calculus before attempting this course. If your Math ACT score is below 26, you might need to take MAT 123 Precalculus first, but note that MAT 123 is offered in Fall semesters only.
- **MAT 225 Calculus II.** This is the second of the three seriously mathematical courses in the major. It is offered every semester, and the prerequisite is MAT 125.
- **MAT 331 Probability Theory.** This is a calculus-based study chance phenomena and probability distributions, with selected applications. Topics include probability laws and elementary combinatorics, random variables, discrete and continuous probability distributions, joint distributions, conditional probability, and the Central Limit Theorem. It is offered in the Fall semester of even-numbered years, so only once every two years! The prerequisite is MAT 225.

In addition, the student must take one of the following two courses:

- **CSC 400 Modern Data Science**. This course continues the work of CSC 303. Topics include supervised machine-earning, unsupervised machine-learning, interactive graphics, database query, web-app frameworks for data exploration and reporting, and workflow-tools such as version-control systems. Additional topics such as network analysis or text-based analysis may be covered as time permits. It is taught only once every two years, in the Spring semester of even-numbered years. The prerequisite, obviously, is CSC 303.
- **MAT 337 Applied Statistical Models.** This is a course on modeling in statistics, with a focus on linear models and their applications. Topics include: basic model designs, geometric understanding of models and random vectors, interpretation of models and inference from them (confidence intervals and hypothesis testing), investigating causation, experiments. The course is offered only once every two years, in the Spring semester rof off-numbered years. It can be quite mathematical, so MAT 225 is a prerequisite. Our work is computationally-intensive, but the necessary computer programming is taught from scratch.

Possible Completion Sequences

If you begin taking courses in an even-numbered Fall, you might proceed as follows:

- Fall of Year One:
    - MAT 111
    - MAT 123 (if needed to prepare for MAT 125)
- Spring of Year One
    - MAT 125
- Fall of Year Two
    - CSC 303
    - MAT 225
- Spring of Year Two
    - CSC 400
- Fall of Year Three
    - MAT 331

If you begin taking courses in an odd-numbered Fall, you might proceed as follows:

- Fall of Year One:
    - MAT 111
    - MAT 123 (if needed to prepare for MAT 125)
- Spring of Year One
    - MAT 125
- Fall of Year Two
    - MAT 225
- Spring of Year Two
    - Free semester
- Fall of Year Three
    - CSC 303
- Spring of Year Three
    - CSC 400 (or free semester)
- Fall of Year Four
    - MAT 331
- Spring of Year Four
    - Free semester (or MAT 337 if you did not previously take CSC 400)

You can see that it is important to begin your mathematics courses early, especially if you have to take pre-calculus before taking Mat 125.

Limitation

Students who wish major in Mathematics and also minor in Data Analysis must be bery careful in their course-selections, because at Georgetown College the minor must include at least 9 credit-hours of courses that are NOT used to satisfy the requirements of the major.

Because of the above 9-hour rule, it is not possible to major in Engineering Mathematics and also minor in Data Analysis.