美国大学数据科学本科课程设置指南介绍

报告人：胡国清

单　位：中南大学湘雅公共卫生学院

广州，2019年7月

汇报内容

1. 《Curriculum Guidelines for Undergraduate Programs in Data Science》文章简介

2. 数据科学发展对国内医学统计学教学和科研的启示

1 数据科学本科课程设置指南

参考文献：

1. De Veaux RD, Agarwal M, Averett M et al. Curriculum Guidelines for Undergraduate Programs in Data Science. Annu. Rev. Stat. Appl. 2017.4:15-30.

2. Cady F . The Data Science Handbook[M]. 2017.

1.1 Introduction

- A need for **hundreds of thousands of Data Science jobs** in the next decade (McKinsey report).
- **530 programs** in Data Science, analytics and related fields at over 200 universities around the world. (*http://datascience.community/colleges*)
- Rapid growth of undergraduate programs at both **research institutions and liberal arts colleges**.

1.1 Introduction

The 2016 Park City Mathematics Institute (PCMI)

- For the purpose of **composing guidelines for undergraduate programs in Data Science**.
- **25 faculties** (computer scientists, statisticians and mathematicians from liberal arts colleges and research universities), **three weeks**.
- Vision for Data Science in an **undergraduate context**.

1.2 Background and Guiding Principles

1.2.1 Data Science as Science

➢ The StatSNSF committee statement that Data Science comprises the "science of **planning for, acquisition, management, analysis of, and inference from data**." (See *http://www.nsf.gov/attachments/130849/public/Stodden-StatsNSF.pdf*.)

1.2.2 Interdisciplinary Nature of Data Science

➢ Data science is **inherently interdisciplinary**.

➢ The integration of courses is a fundamental feature of an effective Data Science program and results in a synergistic approach to problem solving.

1.2 Background and Guiding Principles

1.2.3 Data at the Core

- The recursive data cycle of **obtaining, wrangling, curating, managing and processing data, exploring data, defining questions, performing analyses and communicating the results** lies at the core of the Data Science experience.

1.2.4 Analytical (Computational and Statistical) Thinking

- The two pillars of **computational** and **statistical thinking** should not be taught separately.
- The balance between them may change from one course to another, but both should be present for the most effective and efficient teaching.

1.2 Background and Guiding Principles

1.2.5 Mathematical Foundations

- A working data scientist requires a firm foundation in mathematics.
- An efficient Data Science major should present these mathematical concepts in two courses, in the context of modeling for data-driven problems.
- Propose modeling (both algorithmic and statistical) as a motivator for mathematical tool development, introducing concepts in order to solve our real-world problems.

1.2.6 Flexibility

- Prepare students to learn new techniques and methods that may not exist today.
- Pay attention to the core foundations of mathematics, computational and statistical thinking and practice while incorporating the practical and important Data Science skills.

1.3 Key Competencies and Features of a Data Science Major

- Computational and Statistical Thinking
- Mathematical Foundations
- Model Building and Assessment
- Algorithms and Software Foundation
- Data Curation
- Knowledge Transference – Communication and Responsibility

1.3 Key Competencies and Features of a Data Science Major

1.3.1 Analytical (Computational and Statistical) Thinking

1.3.1.1 Statistical Thinking in a Data-Rich Environment

- **The data scientist:** needs an understanding of basic statistical theory.
- **Students**: understand the basic statistical concepts of data analysis, data collection, modeling, and inference.
- **Graduates**: apply statistical understandings and computational skills to formulate problems, plan data collection campaigns or identify and gather relevant existing data, and then analyze the data to provide insights.

1.3 Key Competencies and Features of a Data Science Major

1.3.1 Analytical (Computational and Statistical) Thinking

1.3.1.2 Computational Thinking

- **Students**: prepared to work with data commonly found in the workplace and research labs; professional statistical analysis software packages, and the underlying principles of programming and algorithmic problem-solving.
- **Graduates**: proficient in many foundational software skills and the associated algorithmic, computational problem-solving of the discipline of computer science.

1.3 Key Competencies and Features of a Data Science Major

1.3.1 Analytical (Computational and Statistical) Thinking

1.3.1.3 Integration of Approaches

- **Graduates**:
  - an understanding of the connections between these two knowledge domains;
  - bring different skills and problem-solving approaches to bear on any particular problem;
  - make informed choices about which skills are appropriate in a given setting;
  - work with various tools, learn new tools and even develop new tools themselves.
- **Data scientists:**
  - must be capable of adapting smoothly to changes in the computing environments.
  - should understand both the computational and modeling challenges in their work, and how they might be intertwined.

1.3 Key Competencies and Features of a Data Science Major

1.3.2 Mathematical Foundations

- Students should be able to impose mathematical structure on data-driven problems by developing structured mathematical problem solving skills.
- Students should have enough mathematics to understand the underlying structure of common models used in statistical and machine learning as well as the issues of optimization and convergence of the associated algorithms.

1.3 Key Competencies and Features of a Data Science Major

1.3.3 Model Building and Assessment

1.3.3.1 Informal Modeling

- Graduates must also be adept at data visualization (an important tool in informal modeling, it can communicate with others and identify weaknesses in proposed models).

Informal modeling involves identifying potential sources of variation, discerning between stochastic and deterministic variation, and understanding how these might be modeled mathematically and computationally.

1.3 Key Competencies and Features of a Data Science Major

1.3.3 Model Building and Assessment

1.3.3.2 Formal Modeling

- Graduates:
  - can build and assess statistical and machine learning models, employ various formal inference procedures, and draw appropriate scope of conclusions from the analysis.
  - be able to bring computational considerations to bear in the analysis of data, including issues of scale.

1.3 Key Competencies and Features of a Data Science Major

1.3.4 Algorithms and Software Foundation

- Graduates
  - be able to employ algorithmic problem solving skills to the task at hand.
  - should understand the memory and execution performance of the structures and software, and that of the libraries and packages.
  - should know and utilize good practices in documentation and structure and be able to use appropriate tools for maintaining their software.
  - should be able to leverage existing packages and tools to solve their computational problems.

1.3. Key Competencies and Features of a Data Science Major

1.3.5 Data Curation

1.3.5.1 Data Preparation

- Graduates should be able to work with data from various sources and formats.
- Graduates should be able to prepare the data for use with various statistical methods and models; and should recognize how the quality of the data and the means of data collection may affect conclusions.

1.3.5.2 Data Management

- Data scientists must ensure the integrity of the data.
- This requires working with relational databases (such as a SQL database), maintaining version control and tracking data provenance.

1.3 Key Competencies and Features of a Data Science Major

1.3.6 Knowledge Transference

1.3.6.1 Communication

- Programs in Data Science should feature exposure to and ethical training in areas such as:
  - citation and data ownership
  - security and sensitivity of data
  - consequences and privacy concerns of data analysis
  - the professionalism of transparency and reproducibility

1.3.6.2 Ethics and Reproducibility

- Data scientists must communicate to teammates and those with less intimate knowledge of the project particulars.
- Students should gain experience using oral, written and visual modes to communicate effectively to various audiences.

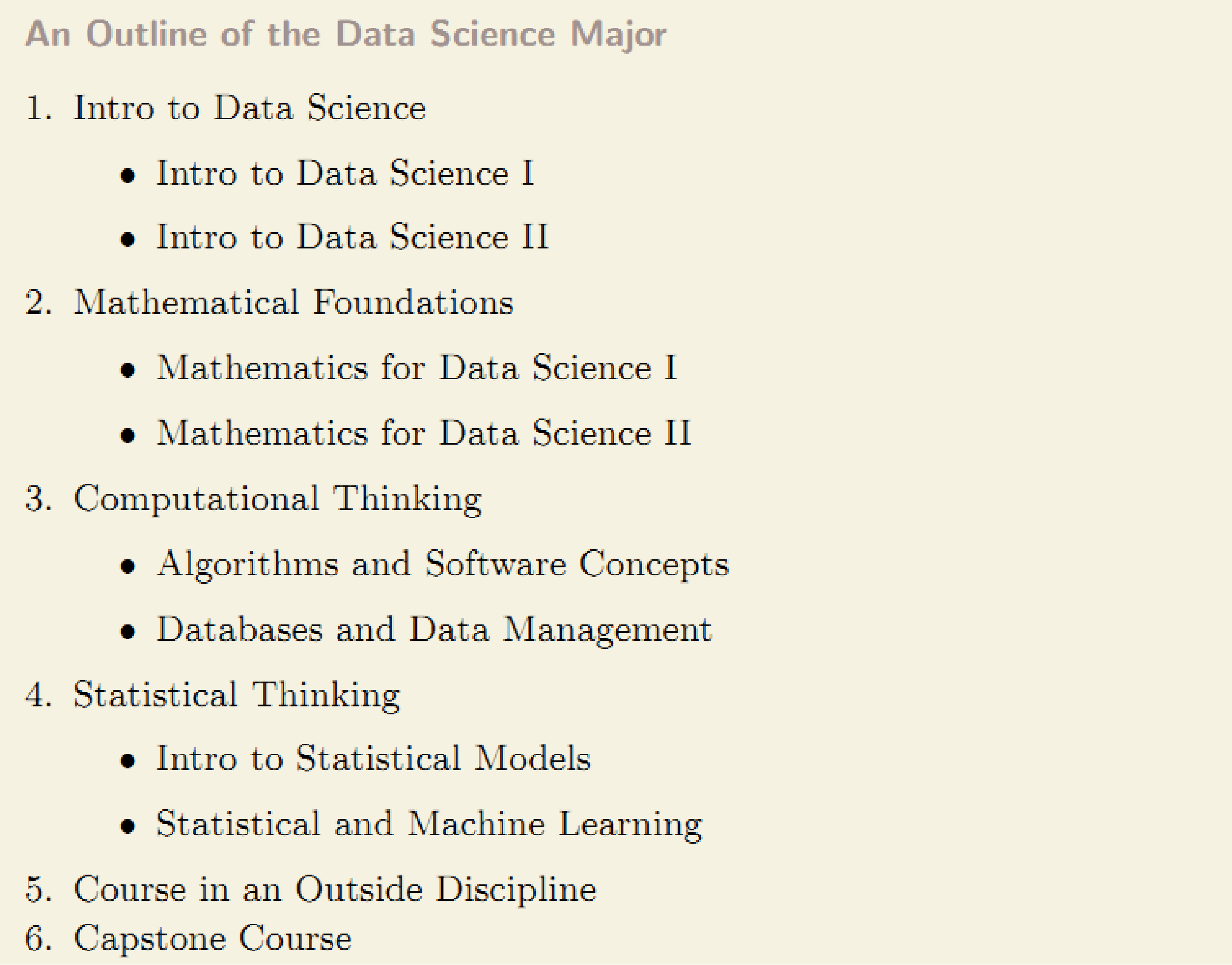
1.4 Curricular Content for Data Science Majors

Six Main Subject Areas of a Data Science Major

- Data Description and Curation
- Mathematical Foundations
- Computational Thinking
- Statistical Thinking
- Data modeling
- Communication, Reproducibility and Ethics

1.4 Curricular Content for Data Science Majors

A summary of the courses designed for these subject areas is found in the following:



An Outline of the Data Science Major

1. Intro to Data Science
   - Intro to Data Science I
   - Intro to Data Science II
2. Mathematical Foundations
   - Mathematics for Data Science I
   - Mathematics for Data Science II
3. Computational Thinking
   - Algorithms and Software Concepts
   - Databases and Data Management
4. Statistical Thinking
   - Intro to Statistical Models
   - Statistical and Machine Learning
5. Course in an Outside Discipline
6. Capstone Course

1.4 Curricular Content for Data Science Majors

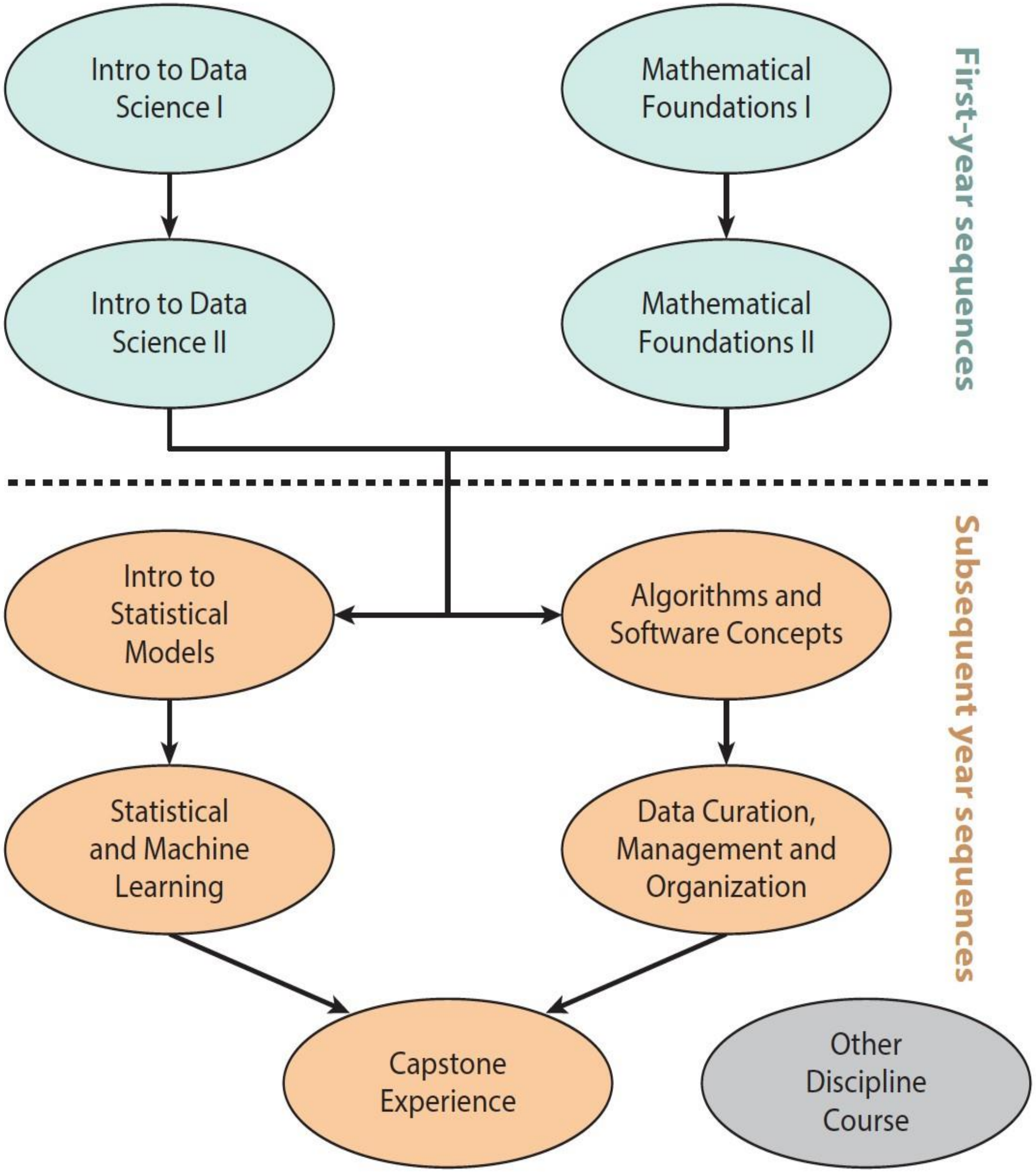

Intro to Data Science I
Mathematical Foundations I
Intro to Data Science II
Mathematical Foundations II
First-year sequences
Intro to Statistical Models
Algorithms and Software Concepts
Statistical and Machine Learning
Data Curation, Management and Organization
Subsequent year sequences
Capstone Experience
Other Discipline Course

Figure 1 A Flow Chart Displaying a possible path through the major

1.5 Additional Considerations

1.5.1 Graduate Study

- Students interested in graduate study in mathematics, statistics or computer science may consider taking more advanced courses in theoretical foundations.
- The courses in mathematics for Data Science will not likely prepare a student for immediate acceptance into a PhD program in one of the three disciplines.

1.5 Additional Considerations

1.5.2 Articulation with community colleges

- Students can prepare by taking Calculus 1 and 2 as well as an Introduction to Computer Science course.
- Institutions should encourage collaboration between departments of mathematics and computer.
- The Statway (*http://www.carnegiefoundation.org/resources/videos/introducing- statway/*) and the New Mathways (*http://www.utdanacenter.org/higher- education/new-mathways-project/*) course sequences

1.5 Additional Considerations

1.5.3 Prerequisites and preparation in high school

1.5.4 Internship and applied experiences

- Practical projects should be implemented often throughout the curriculum and provide the central experience of a capstone course.

1.6 Transitioning to a Data Science Major Using Typical Existing Courses

1.6.1 Courses in Mathematics

- Calculus 1
- Calculus 2
- Calculus 3
- Linear Algebra
- Probability Theory
- Discrete Math

1.6.2 Courses in Computer

- Intro to Computer Science
- CS2: Data Structures/Algorithms
- Computer Systems and Architecture
- Advanced Algorithms
- Databases
- Software Engineering

1.6 Transitioning to a Data Science Major Using Typical Existing Courses

1.6.3 Courses in Computer

- Intro to Statistics
- Statistical Modeling/Regression
- Machine Learning/Data Mining
- Theory of Statistics (requires Probability Theory)

1.6.4 Related Courses

- Introduction to [Partner Discipline]
- Intermediate course in Discipline
- Capstone Course with Data Experience and Projects
- Two courses in writing, preferably one in technical writing
- Public Speaking
- Ethics

2 数据科学发展对国内医学统计学教学和科研的启示

个人不成熟思考

a) 多学科跨学院配合 vs.医学院校校师资单一、教学组织难

b) 能力要求高 vs. 有限课时

c) 数据科学在本科生与研究生之间的差异?

d) 国内对此类毕业生数的真实需求?

e) **医学统计学**是否应主动转型为数据科学?

请批评指正！