Envisioning the DATA SCIENCE DISCIPLINE
The Undergraduate Perspective

9/12/17 – Building Data Acumen
(recording posted)

9/19/17 – Incorporating Real-World Applications
(recording posted)

9/26/17 – Faculty Training and Curriculum Development
(recording posted)

10/3/17 – Communication Skills and Teamwork
(recording posted)

10/10/17 – Inter-Departmental Collaboration and Institutional Organization

10/17/17 – Ethics

10/24/17 – Assessment and Evaluation for Data Science Programs

11/7/17 – Diversity, Inclusion, and Increasing Participation

11/14/17 – Two-Year Colleges and Institutional Partnerships

Provide input, download the interim report, and learn more about the study at www.nas.edu/EnvisioningDS
Envisioning the
DATA SCIENCE DISCIPLINE
The Undergraduate Perspective
Inter-Departmental Collaboration & Institutional Organization

Mark Embree, Virginia Tech
Professor, Department of Mathematics
Leader, Computational Modeling and Data Analytics (CMDA) division
Associate Director, Virginia Tech Smart Infrastructure Laboratory

Michael Franklin, University of Chicago
Liew Family Chair of Computer Science
Senior Advisor to Provost on Computation and Data Science
Chairman, Department of Computer Science

Provide input and learn more about the study at www.nas.edu/EnvisioningDS
Envisioning the DATA SCIENCE DISCIPLINE

The Undergraduate Perspective

Inter-Departmental Collaboration & Institutional Organization

Forging Virginia Tech’s CMDA Major Across Departments

Mark Embree, Virginia Tech
Professor, Department of Mathematics
Leader, Computational Modeling and Data Analytics (CMDA) division
Associate Director, Virginia Tech Smart Infrastructure Laboratory

Provide input and learn more about the study at www.nas.edu/EnvisioningDS
Virginia Tech’s CMDA Major

CMDA = Computational Modeling and Data Analytics

The CMDA undergraduate major was founded in 2015 as a collaboration between CS, Math, and Statistics, via the leadership of Dean of Science Lay Nam Chang.

In addition to existing faculty who shaped the program, VT has hired
• Five tenure track faculty in Math (including two full professors);
• Two tenure track faculty and one collegiate faculty in Statistics.

This year (2017–2018), CMDA will hire four faculty:
• Tenure track in Math
• Tenure track and collegiate faculty in Statistics
• Tenure track in Economics

Provide input and learn more about the study at www.nas.edu/EnvisioningDS
Ingredients of CMDA Curriculum

STATISTICS FOR BIG DATA
   Data mining, machine learning, visualization

APPLIED MATHEMATICS FOR MODELING
   Linear algebra, differential equations, numerical analysis

HIGH-PERFORMANCE COMPUTING
   Parallel/GPU programming for data/science/engineering apps

ACCESS TO RELEVANT APPLICATIONS
   Natural and social sciences, engineering, humanities, internet
   Specialized degree options in Economics, Physics, more coming.

PRACTICAL SKILLS FOR PROBLEM SOLVING (CAPSTONE)
   Ethics, collaboration, leadership, presentation skills

Provide input and learn more about the study at www.nas.edu/EnvisioningDS
CMDA Curriculum

CALCULUS

LINEAR ALGEBRA

SOFTWARE DESIGN & PROGRAMMING

INTEGRATED QUANTITATIVE SCIENCE

HIGH PERFORMANCE COMPUTING

MATHEMATICAL MODELING

DATA ANALYTICS & VISUALIZATION

CAPSTONE PROJECT

Four electives drawn primarily from

CMDA

COMPUTER SCIENCE

MATHEMATICS

STATISTICS

Provide input and learn more about the study at www.nas.edu/EnvisioningDS
Students take either the Integrated Path or the Traditional Path.
Planning for the CMDA major began around 2012, a collaboration of a dozen faculty in Computer Science, Mathematics, Physics, and Statistics.

The curriculum builds on faculty research interests in applied math, high performance scientific computing and Bayesian analytics.

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>CMDA proposal finalized</td>
</tr>
<tr>
<td>SPRING 2014</td>
<td>CMDA approved by Virginia Tech’s Board of Visitors</td>
</tr>
<tr>
<td>JULY 2014</td>
<td>CMDA approved by State Council on Higher Ed. (SCHEV)</td>
</tr>
<tr>
<td>SPRING 2015</td>
<td>Students can first declare the CMDA major</td>
</tr>
<tr>
<td>FALL 2015</td>
<td>First freshman class arrives (45 students)</td>
</tr>
<tr>
<td>FALL 2016</td>
<td>Second freshman class arrives (52 students)</td>
</tr>
<tr>
<td>MAY 2017</td>
<td>First graduation (22 students)</td>
</tr>
</tbody>
</table>
CMDA Enrollment

Class of 2017
Class of 2018
Class of 2019
Class of 2020
Class of 2021

0 20 40 60 80 100 120 140

incoming freshmen  added/transferred into CMDA

Provide input and learn more about the study at www.nas.edu/EnvisioningDS
Capstone Project Course

- Teams of 3–4 students work on one project for the entire semester.

- Projects come from external clients (companies or within VT).

- Teams are guided through a methodical problem-solving process.

- Focus on teamwork, leadership, collaboration skills, ethics, project management, and communications (written, visual, oral).
Capstone Projects for Fall 2017

Project sponsors from Virginia Tech:

• **Math Emporium** (tutor response speed; quiz analytics)
• **Economics Department** (infrastructure failures)
• **Social and Decision Analytics Lab** (open source software)
• **Biocomplexity Institute** (sick tweeting; disease dynamics)
• **VT Center for Autism Research** (geographic disparities)
• **VT Athletics Department** (press release effectiveness)
Reflections on CMDA Curriculum

MATH MATTERS
The foundational math curriculum is demanding compared to some data science programs, but the foundation unlocks the ability to dig deeply into modern algorithms.

DATA SCRAPING IS AN EMPOWERING TECHNOLOGY
Once students learn to scrape data, they are empowered to pursue their own favorite applications as side projects.

HIGH PERFORMANCE COMPUTING FOR THE MASSES
Many CMDA majors arrive with little programming experience, but they all end up learning (and usually enjoying) HPC.

A MISSING INGREDIENT: HIGH PERFORMANCE DATABASES
We would like to add a course in high performance data (as well as conventional databases); cf. [De Veaux et al. 2017].
Administrative Structure

CMDA is administratively housed in the College of Science, though the key departments span two colleges:

- Computer Science (College of Engineering)
- Mathematics (College of Science)
- Statistics (College of Science)

Current CMDA degree options engage with

- Economics (College of Science)
- Physics (College of Science)

but could easily expand into departments in other colleges.
The College of Science set up the *Academy of Integrated Science (AIS)*, a department-level unit that administers the College’s interdisciplinary programs:

- B.S. CMDA
- B.S. Nanoscience
- B.S. Systems Biology
- Minor in Science, Technology, and Law
- Integrated Science Curriculum (freshmen/sophomores)

The AIS manages budgets, undergrad advising, student recruiting, and assessment for CMDA.

The CMDA faculty director reports to the AIS director, Prof. Michel Pleimling (Physics).
CMDA Faculty Expectations

Faculty are hired into a home department (e.g. Math, Stats), governed by a Memorandum of Understanding with AIS.

• Each CMDA hire obliges the home department to teach two CMDA classes per year.
• These courses need not be taught by the CMDA faculty member (though they usually are).
• CMDA hires devote much of their service to CMDA, rather than their departments.
• The AIS and CMDA directors contribute a letter to tenure/promotion dossiers for CMDA faculty.
Challenge 1: Hiring

How can we best hire into an interdisciplinary program?
We have learned a few lessons over the past few years.

• The home department should lead the search. Candidates need to understand clearly that the tenure home is in the department, not the AIS. The search should look like a departmental search, but with CMDA faculty on the hiring committee, and a meeting with the AIS director.

• The candidate’s role in CMDA must be clearly articulated. The role must be clear and understood by all interviews.

• The interdisciplinary program should be an attractor. Rather than teaching conventional courses, candidate can teach more innovative curriculum that aligns well with research interests.

• Reinforce these messages with current CMDA faculty.

• Introduce the candidate to CMDA students.
Challenge 2: Teaching

CMDA teaching needs good collaboration with departments.

• During our boot-up, CMDA teaching needs outstrip departmental teaching obligation from CMDA hires.

• Innovative new courses need creative teachers – often strong faculty who are popular with students.

• Fast growing program demands extra sections beyond initial projections: flexibility is needed.

• GTA resources come from departments (CMDA does not have a graduate degree).

• Good communication between CMDA leader and department chairs is key!
Challenge 3: Number of Majors

CMDA attracts students to VT (often from out of state). More students transfer into CMDA once they are at VT.

Number of CMDA majors:
- Smaller than CS
- Comparable to Math
- Much larger than Statistics

*
CMDA draws students away from CS and Math.
*  
- Pro: CMDA is a better fit for some students
- Con: could spark rivalry with departments, depending on how university budget is allocated.
Summary

• Faculty are excited about this interdisciplinary project.
• Students are responding; so are employers.
• Deans of Science (Lay Nam Chang and Sally Morton) have been vital boosters for CMDA.
• Generous department chairs are essential:
  – Cal Ribbens (Computer Science)
  – Peter Haskell (Mathematics)
  – Eric Smith and Ron Fricker (Statistics)
• Good communication is vital.
Envisioning the
DATA SCIENCE DISCIPLINE
The Undergraduate Perspective
Inter-Departmental Collaboration & Institutional Organization

Forging Virginia Tech’s CMDA Major Across Departments

Mark Embree, Virginia Tech
Professor, Department of Mathematics
Leader, Computational Modeling and Data Analytics (CMDA) division
Associate Director, Virginia Tech Smart Infrastructure Laboratory

Q&A
embree@vt.edu

Provide input and learn more about the study at www.nas.edu/EnvisioningDS
Envisioning the
DATA SCIENCE DISCIPLINE

The Undergraduate Perspective

Inter-Departmental Collaboration & Institutional Organization

Some Thoughts on Data Science Education for Undergraduates

Michael Franklin, University of Chicago
Liew Family Chair of Computer Science
Senior Advisor to Provost on Computation and Data Science
Chairman, Department of Computer Science

Provide input and learn more about the study at www.nas.edu/EnvisioningDS
Computing Research and the Emerging Field of Data Science

October 7, 2016 / In: Featured Announcements, For Researchers, Research Issues, Resources /

By CRA’s Committee on Data Science: Lise Getoor (Chair), David Culler, Eric de Sturler, David Ebert, Mike Franklin, and H.V. Jagadish on behalf of the CRA Board

“... data science provides new opportunities for creative collaborations between industry, academia and government for pure and applied research.”

Provide input and learn more about the study at www.nas.edu/EnvisioningDS
If NSF can help foster the evolution and development of both Data Science and Data Scientists over the next decade, we can begin to meet the potential of Data Science to drive new discovery and innovation...

This should include not only a focus on fundamental Data Science, but also on translational efforts to move ideas from research to practice across the broadest landscape of commercial applications.
Data Science Programs at U Chicago

**MS** in Computational Analysis and Public Policy (CAPP)

- Yr 1: Stats, Econ, CS w/apps, DB, ML for Policy
- Yr 2: Analytical Politics, Program Eval, Capstone...

**MA** in Computational Social Science

- Yr 1: CS w/apps; Perspectives on Analysis, Modeling, Computing; Math & Stats
- Yr 2: Computational Methods, Social Sci, Capstone

Joint **MBA/MS in CS**

- Students get both degrees

Provide input and learn more about the study at [www.nas.edu/EnvisioningDS](http://www.nas.edu/EnvisioningDS)
• Like most places – we’ve experienced dramatic increases in undergraduate enrollments in many CS and Stats classes (esp. Machine Learning)

• Initiatives are arising in Biological Sciences, Computational Social Sciences, Digital Humanities,…

• U Chicago’s ”Core” approach could provide an opportunity for curriculum development

• A campus-wide faculty committee is assessing and will make recommendations

Provide input and learn more about the study at www.nas.edu/EnvisioningDS
WHERE DOES DATA SCIENCE LIVE ON A MODERN CAMPUS?
Nearly Everywhere! (e.g., Berkeley circa 2014)
Some Big Issues in DS Education

- Establishing Data Science as a Discipline
  - Came from industry – not driven by academia
  - Unique intellectual foundations of Data Science?
  - Need scientific culture: e.g., Journals, Conferences
  - Training vs. undergraduate education

- Where on campus should DS be taught?
  - Department of Data Science? School of Data Science? Everywhere?
  - What departments should contribute? drive it?

- To whom
  - All Undergrads? (see Berkeley’s “Data Science 8”)
  - Certificates?, Minors?, Majors?

- How to manage the “Hype Cycle”?
  - Everyone wants a piece of it
  - Also, some skepticism

Provide input and learn more about the study at www.nas.edu/EnvisioningDS
A Lifecycle View of Data Science

{Ethics, Policy, Regulatory, Stewardship, Platform, Domain} Environment

Acquire
Create, capture, gather from:
- Lab
- Fieldwork
- Surveys
- Devices
- Simulations
- etc

Clean
- Organize
- Filter
- Annotate
- Clean

Use / Reuse
- Analyze
- Mine
- Model
- Derive ++data
- Visualize
- Decide
- Act
- Drive:
  - Devices
  - Instruments
  - Computers

Publish
- Share
  - Data
  - Code
  - Workflows
  - Disseminate
  - Aggregate
  - Collect
  - Create portals, databases, etc
  - Couple with literature

Preserve / Destroy
- Store to:
  - Preserve
  - Replicate
  - Ignore
  - Subset, compress
  - Index
  - Curate
  - Destroy

from the National Science Foundation CISE AC Data Science Report

Provide input and learn more about the study at www.nas.edu/EnvisioningDS
Data Science $\neq$ CS + Statistics

- In general – Data as a First-Class Concept
- Structure: Schema-on-read and Data Lakes (DataSpaces)
- Data Science Lifecycle
- Safe Data Science
  - “end-to-end” Bias Mitigation
  - Ethics and Data Privacy
  - Communicating results and influencing decisions
- Foundations & Methodologies vs. current tool set
- Note: DATA SCIENCE $\neq$ BIG DATA
  - Much can be taught on laptops
  - Scalability adds further issues and tradeoffs

Provide input and learn more about the study at www.nas.edu/EnvisioningDS
Some Declared Data Science Majors

- Michigan: joint EECS (CoE) and Stats in LSA (Literature Science and Arts)
- Ohio State: Data Analytics Major – joint CS/Stats – both in A&S with “Curricular partnerships”: Eng, Med, Business
- Penn State: “Data Sciences” – Colleges of Info Sys, Eng and Science – core, 1 of 3 concentrations, capstone project
- Purdue (fall 2017): Joint CS+Stats – (Eng and Coll of Sciences)
- U Rochester: CS+Stats+advanced coursework in an application area
- Yale: “Department of Stats and DS” – major approved (March 2017)
My Personal Take

• Data Science by necessity must span existing academic boundaries
• A “one size fits all” approach will not work
  – Some students need training and tools
  – Other students will drive the discipline forward
• Modern university structures are not optimized for such fields
• Widespread enthusiasm and interest provides an opportunity to innovate and collaborate across campus

Provide input and learn more about the study at www.nas.edu/EnvisioningDS
Envisioning the DATA SCIENCE DISCIPLINE

The Undergraduate Perspective

Inter-Departmental Collaboration & Institutional Organization

Mark Embree, Virginia Tech
Professor, Department of Mathematics
Leader, Computational Modeling and Data Analytics (CMDA) division
Associate Director, Virginia Tech Smart Infrastructure Laboratory

Michael Franklin, University of Chicago
Liew Family Chair of Computer Science
Senior Advisor to Provost on Computation and Data Science
Chairman, Department of Computer Science

Provide input and learn more about the study at www.nas.edu/EnvisioningDS
<table>
<thead>
<tr>
<th>Date</th>
<th>Event</th>
<th>Recording Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>9/12/17</td>
<td>Building Data Acumen</td>
<td>posted</td>
</tr>
<tr>
<td>9/19/17</td>
<td>Incorporating Real-World Applications</td>
<td>posted</td>
</tr>
<tr>
<td>9/26/17</td>
<td>Faculty Training and Curriculum Development</td>
<td>posted</td>
</tr>
<tr>
<td>10/3/17</td>
<td>Communication Skills and Teamwork</td>
<td>posted</td>
</tr>
<tr>
<td>10/10/17</td>
<td>Inter-Departmental Collaboration and Institutional Organization</td>
<td>posted</td>
</tr>
<tr>
<td>10/17/17</td>
<td>Ethics</td>
<td></td>
</tr>
<tr>
<td>10/24/17</td>
<td>Assessment and Evaluation for Data Science Programs</td>
<td></td>
</tr>
<tr>
<td>11/7/17</td>
<td>Diversity, Inclusion, and Increasing Participation</td>
<td></td>
</tr>
<tr>
<td>11/14/17</td>
<td>Two-Year Colleges and Institutional Partnerships</td>
<td></td>
</tr>
</tbody>
</table>

Provide input, download the interim report, and learn more about the study at [www.nas.edu/EnvisioningDS](http://www.nas.edu/EnvisioningDS)