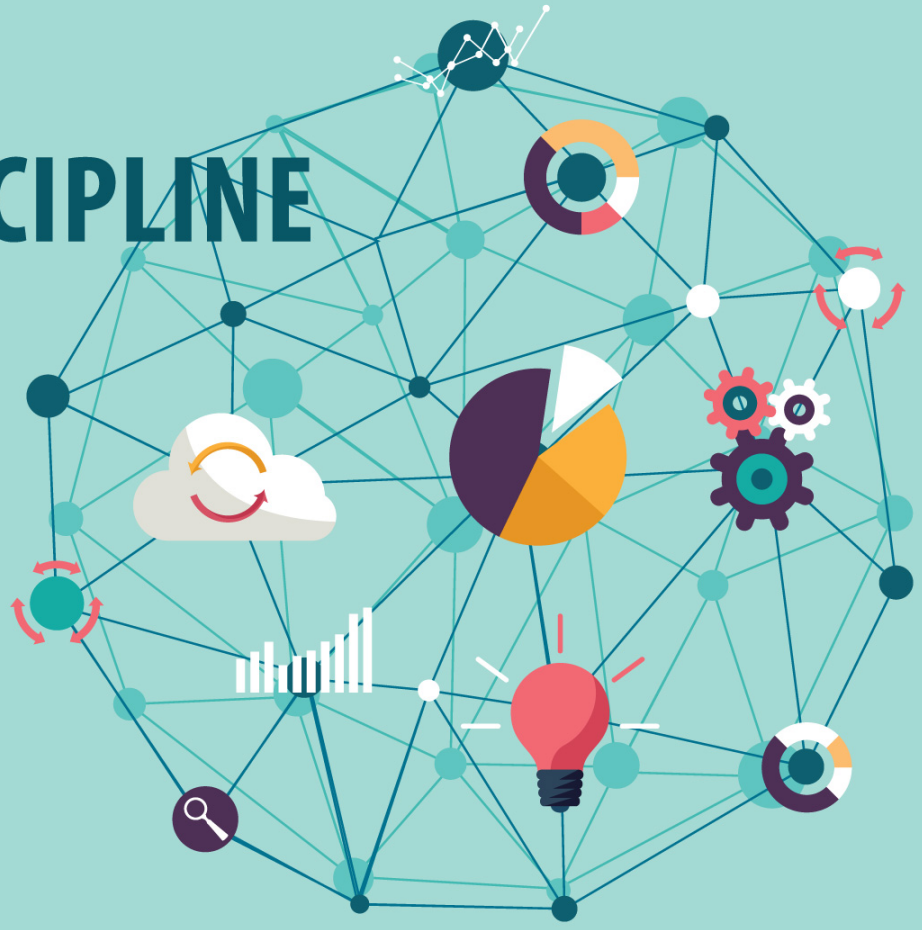
Envisioning the

DATA SCIENCE DISCIPLINE

The Undergraduate Perspective



Webinar Series
Fall 2017

The National Academies of | SCIENCES ENGINEERING MEDICINE

nas.edu/EnvisioningDS

Envisioning the DATA SCIENCE DISCIPLINE

The Undergraduate Perspective

Building Data Acumen

Nicole Lazar, University of Georgia
Professor, Department of Statistics

Mladen Vouk, North Carolina State University
Distinguished Professor of Computer Science, Associate Vice Chancellor for Research Development and Administration

Envisioning the DATA SCIENCE DISCIPLINE

The Undergraduate Perspective

Building Data Acumen

Nicole Lazar, University of Georgia
Professor, Department of Statistics

Capstone Courses

History of Capstone at the University of Georgia

- First offered in AY 2007–2008, 10 students
- Part of UGA's Writing Intensive Program since AY 2008–2009
- Required from AY 2010–2011 for all statistics majors
- Poster session introduced in AY 2010–2011
- Enrollment has grown steadily, tracking growth of major;

  46 students in current offering

Goals of Capstone at the University of Georgia

- Exposure to advanced statistical techniques
- Practice communication of statistical ideas, in writing and orally
- Group work
- Vertical integration of learning

  (faculty → graduate assistants → students)
- Consulting/work with client
- Work with real data
- Professional development

With growth, challenging to maintain all of these.
Current offering: relaxing client and group work.

Capstone at the University of Georgia: Project Formats

- Typically, find projects from researchers around campus; scope and scale have increased dramatically over time
- Also projects from government offices (e.g., USDA, CDC)
- When class size was smaller, community-based survey projects incorporated (e.g., Nuci's Space, Campus Transit)
- Current offering:
  - experimental individual and group ("data repository") options
  - more public outreach projects (Humane Society, United Way)

Essential Components for Building Data Acumen: The Data

- Moving beyond t-tests and ANOVA (topics covered have included bootstrap, classification and regression trees, survival analysis, multiple testing, issues of reproducibility in science ...)
- Hands-on practice beyond the project
- Real data (all projects involve real, often large, data sets)
- Messy data (students are not guaranteed to receive clean data sets from clients)

Essential Components for Building Data Acumen: The Student

- Building awareness of a professional identity as statisticians and (more recently) data scientists
- Communication! Much resistance on this at the time; benefits are seen later (in graduate school, in the work force)
- Building a sense of community among students and instructors, TAs (how to scale up from 10 to ~50 students?)
- Challenge students to go above and beyond their intellectual comfort zones

Looking to the Future

- Conflicting directions
  - Can expect continued growth in statistics and data science programs
  - Hands-on practical experience with real data is essential
- How to scale and still provide a useful experience to students?
- Need for innovative approaches (e.g. team "competitions", smaller workshop groups . . . others?)
- Any such course will be a constant work in progress: don't fall into a routine of "We've always done it this way" – that is no longer sufficient

How Do We (Can We) Know It's Working?

- Internal assessments: Quality of projects; client satisfaction (repeat participants); poster session checklist (completed by attendees – statistics faculty and graduate students, clients)
- Feedback from graduates: What's useful when they get to the next stage?
- What are employers and graduate schools looking for?

Envisioning the DATA SCIENCE DISCIPLINE

The Undergraduate Perspective

Building Data Acumen

Nicole Lazar, University of Georgia
Professor, Department of Statistics

Capstone Courses

Q&A

Envisioning the DATA SCIENCE DISCIPLINE

The Undergraduate Perspective

Building Data Acumen

NC State University Data Science Initiative (dis.ncsu.edu)

Mladen Vouk, North Carolina State University
Distinguished Professor of Computer Science, Associate Vice Chancellor for Research Development and Administration

Context

- Understanding, managing, and using data — often large amounts of unstructured data — is becoming increasingly important in nearly every industry, government sector, and academic domain.
- Not having the skills and infrastructure to apply data science and analytics reliably and correctly has become a major risk for all sectors.

Data (and) Science Literacy

A good data scientist does not have to be a computer scientist, a mathematician, or a statistician.

But ….

Education?

- What should be included in data science curriculum, both now and in the future?
- How to prioritize or best convey for differing types of data (science programs)?
- How can opportunities to enhance data acumen (i.e., the ability to make good judgments and decisions with data) be integrated into data science educational programs?
- How can data acumen be measured or evaluated?
- etc.

Disruptive?

Largest taxi companies may not own taxis? (e.g., Uber)

Largest accommodation providers may not own any accomodations? (AirBnB)

etc.

The Five V-s of (Big)Data*

Kb/sec
Mb/sec
Gb/sec
Tb/sec

Velocity
(How fast is data arriving?)

How much is needed?

Volume
(How much data is arriving?)

MB, 6
GB, 9
TB, 12
PB, 15
XB, 18
ZB, 21
YB, 24

Value
(Economics, Health, Security, …)

Veracity
(Uncertainty, Trust, Source, Information Density, Intent, …)

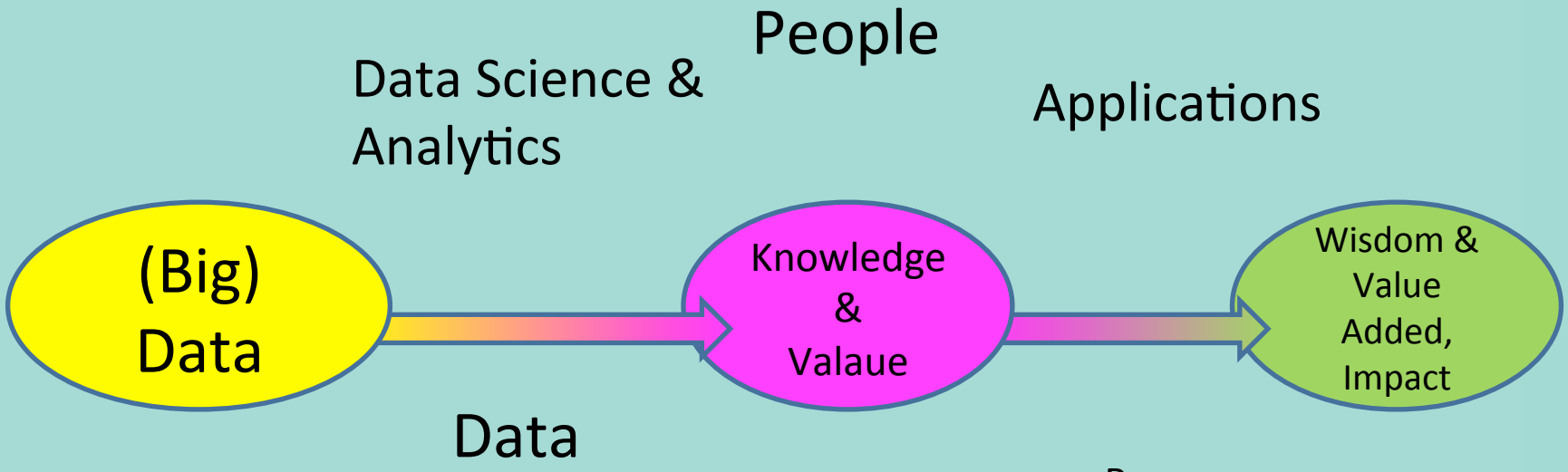
Variety
(Range of Data - text, visual, numerical, location…)

Real Life Data Types

Errors:
Epistemic
Aleatoric
Working with an AI

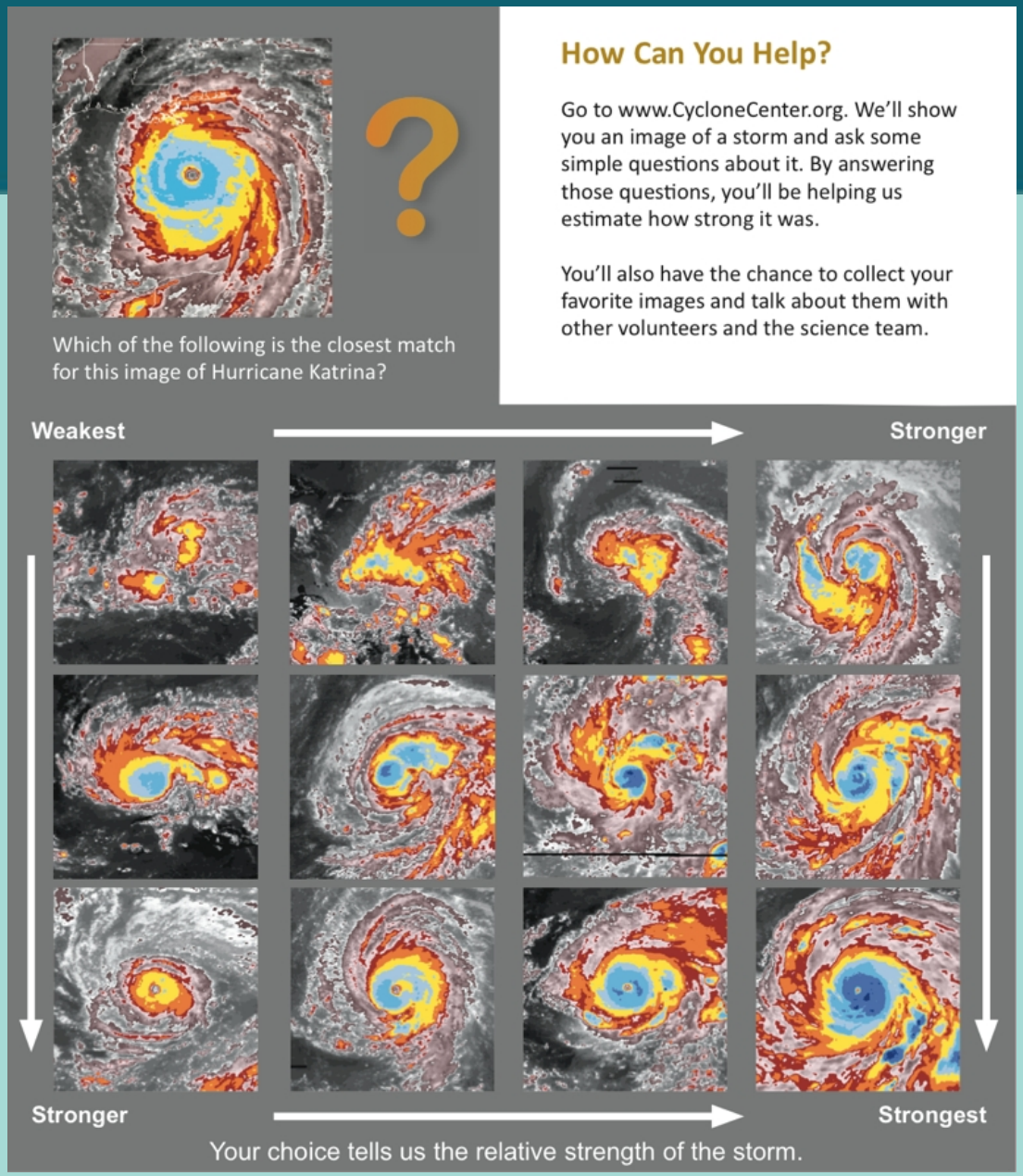
(*) *www.ibmbigdatahub.com/infographic/four-vs-big-data*

Value Path


People
Data Science & Analytics
Applications
(Big) Data
Knowledge & Valaue
Wisdom & Value Added, Impact
Data
e.g., Extract, Verify, Transform, Load, Organize, …
e.g. curation
Technology
Provenance,
Compliance,
Security,
Usability,
Privacy, Ethics
Reliability, Trustworthiness,
Sense of Impact
…

Data Science Work That Matters

Imagine being able to hold a geographic information system (GIS) in your hands, feel the shape of the earth, sculpt its topography, and direct the flow of water. Researchers at NC State's Center for Geospatial Analytics have made this novel idea a reality with Tangible Landscape, an open source tangible interface powered by GRASS GIS that physically and interactively manifests geospatial data so that users can naturally feel it, shape it, and immediately see results projected onto the 3-D model. This makes GIS far more intuitive and accessible for beginners, empowers geospatial experts, and creates exciting new opportunities for data scientists and developers alike – like gaming with GIS. Tangible Landscape is now being applied to tackle complex real world problems, from controlling the spread of wildfire or emerging infectious diseases to understanding the impacts of storm hazards.



Citizen Scientists are comparing nearly 300,000 satellite images with these examples to help improve the global record of hurricanes and tropical cyclones. See how you can help improve our understanding of tropical cyclones.

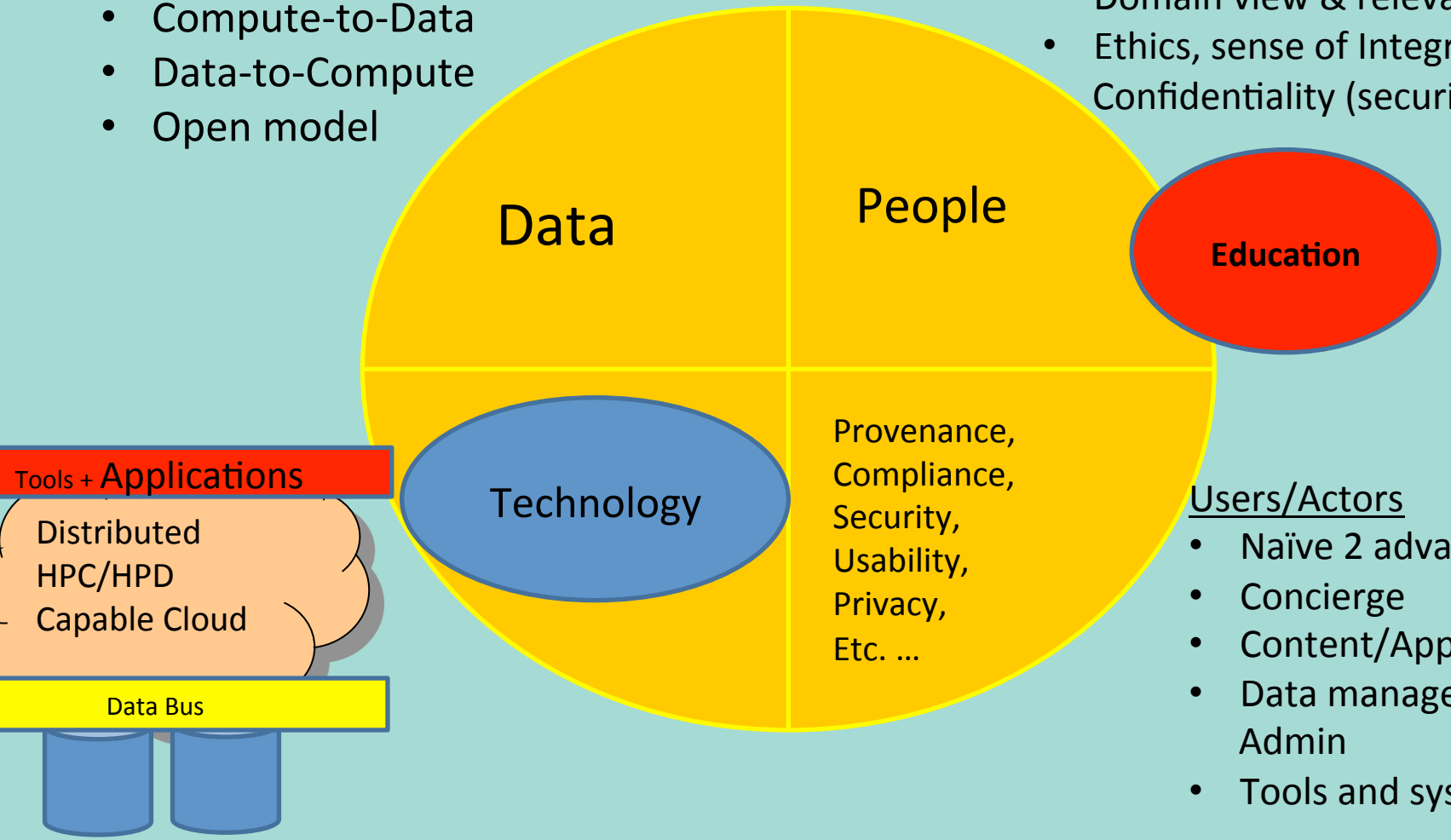
Crowd-sourcing.

https://ncics.org/events/cics-nc-leads-the-launch-of-cyclonecenter-org/

Framework

Data and Analytics Literacy
- Computational Thinking
- Math, modeling, software, data management, methods
- Communications
- Domain view & relevancy
- Ethics, sense of Integrity & Confidentiality (security)

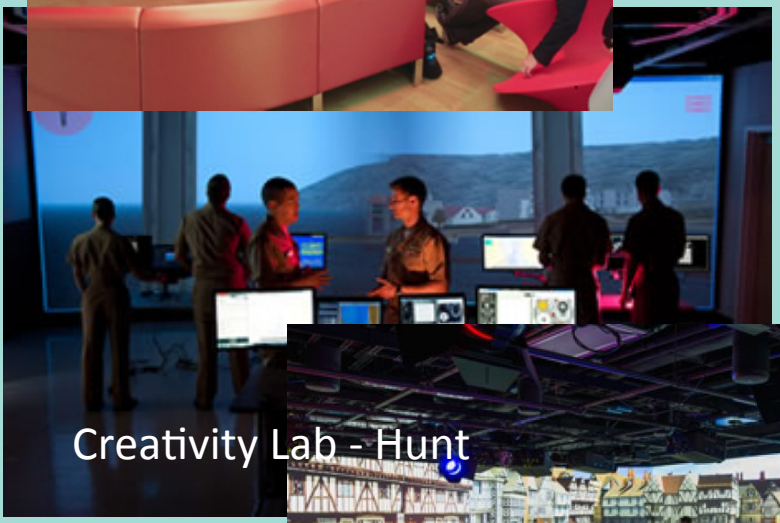
Data Interaction Models:
- Total Isolation
- Compute-to-Data
- Data-to-Compute
- Open model



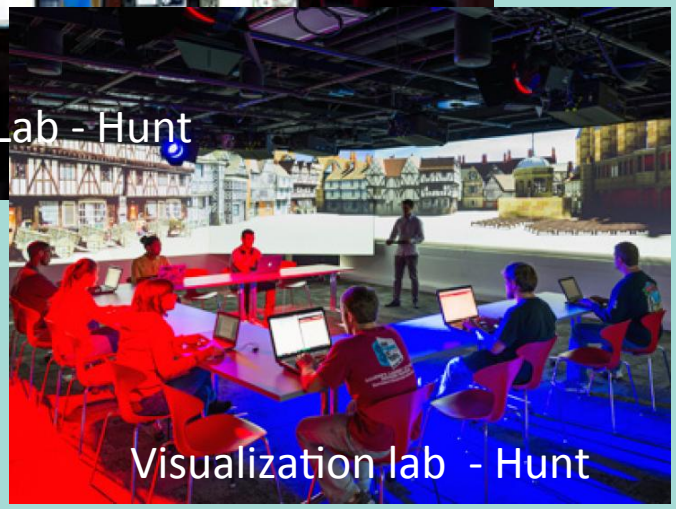
Users/Actors
- Naïve 2 advance user
- Concierge
- Content/App developer
- Data manager and Admin
- Tools and system dev.

NC State has Extensive Visualization Facilities

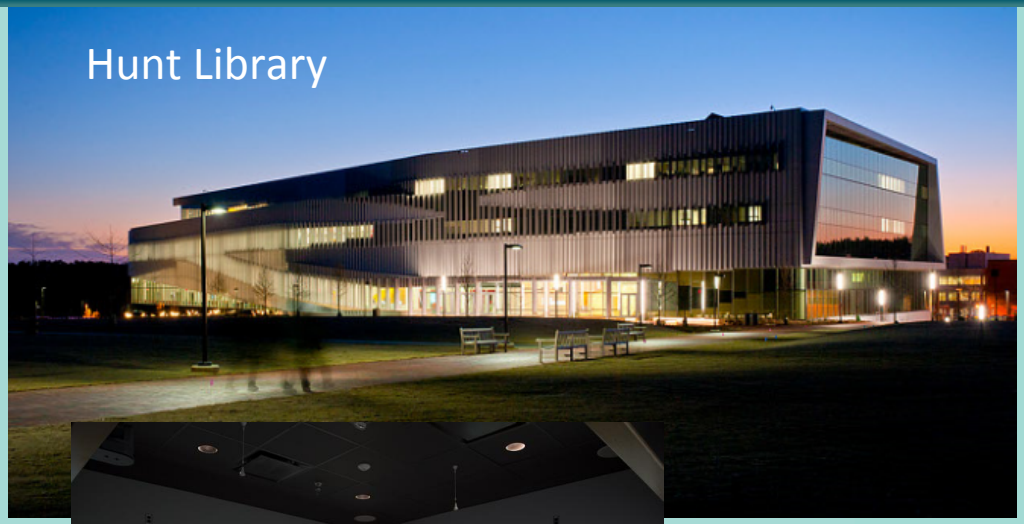
Games Room - Hunt



Creativity Lab - Hunt



Visualization lab - Hunt



Hunt Library

Viz Lab – D.H. Hill

LAS Mersive Experience EB2

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

NC State University Data Science Initiative

Goals

- Coordinate data science activities
- Establish an interdisciplinary data science curriculum to further data science education
- Foster research collaboration, both internal and external
- Increase research funding and competitiveness
- Build industry partnerships
- Provide services & infrastructure to faculty
- Raise visibility & increase reputation
- …

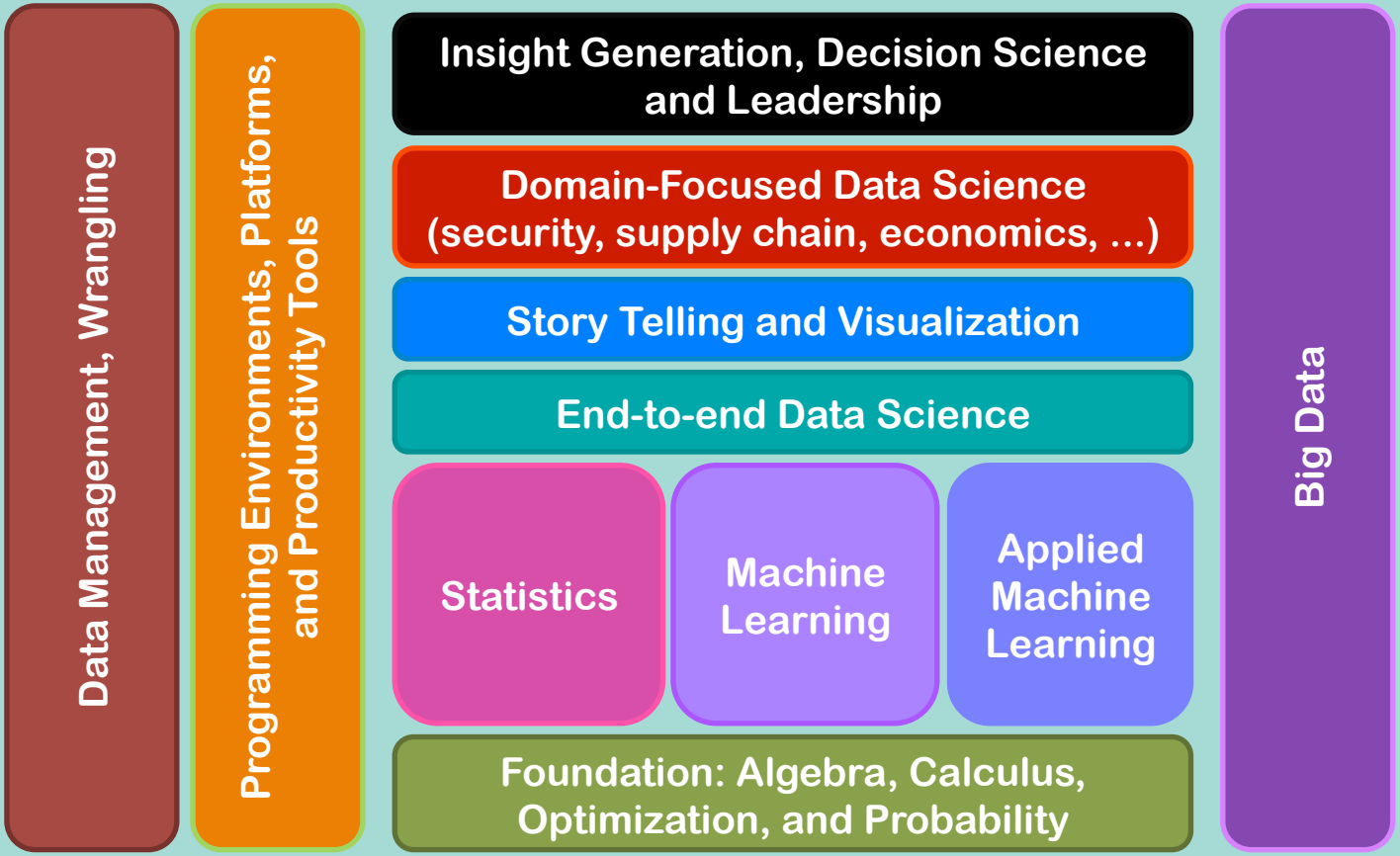
Institutionalize Data Science

Analytics & Data Science Programs

- Undergraduate
  - General Education thematic track
  - Co-taught Computer Science/Statistics undergraduate electives
  - Poole College of Management (PCOM) Undergraduate 15-hour Data Analytics Honors Program
  - Some of the Executive Education and "Data Matters" offerings
- Graduate
  - Institute for Advanced Analytics (IAA) – Professional MS (analytics.ncsu.edu)
  - Data Science Graduate Certificate and MS (CSC/Stat)
  - CSC and Stat graduate tracks in Data Science
  - Poole College of Management – Digital Analytics Certificate within MBA program, CSC/Stat/PCOM executive education for companies
  - Data Matters courses (https://research.ncsu.edu/dsi/data-matters/)
- Other … (e.g., Library offerings)

For Example:

NCSU End-to-End Data Science Curriculum*



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

Data Science (DS) Mastery Levels

- Core (C):
  - Able to master individual core concepts within the Bloom's taxonomy*: Knowledge, Comprehension, Application, Analysis, Evaluation, and Synthesis
  - Able to adapt previously seen solutions to data science problems for target domain-focused applications utilizing these core concepts
- Intermediate Electives (I):
  - Able to synthesize multiple concepts to solve, evaluate, and validate the proposed data science problem from the end-to-end perspective
  - Able to identify and properly apply the textbook-level techniques suitable for solving each part of the complex data science problem pipeline
- Advanced Electives (A):
  - Able to formulate new domain-targeted data science problems, justify their business value, and make data-guided actionable decisions
  - Able to research the cutting edge technologies, compare them and create the optimal ones for solving DS the problems at hand
  - Able to lead a small team working on the end-to-end execution of the DS project

(*) e.g., https://www.csun.edu/science/ref/reasoning/questions_blooms/blooms.html

Core Curriculum Courses

Foundations:

- Matrix Algebra
- Calculus
- Optimization
- Probability

Data Science Methods:

- Statistics
- Machine Learning and Data Mining
- Algorithms and Data Structures

Decision Making

- Data-guided Decision Making
- Visualization, Visual Data Exploration, and Story Telling

Infrastructure and Programming Environments

- Scripting Languages for DS: e.g., R, Python
- Database Management and Optimization

Advanced Elective Curriculum Courses

Data Science Methods

- Deep Learning
- Bayesian Reasoning and Probabilistic Graphical Models
- Process Mining

Data Science Applications:

- Natural Language Processing and Text Analytics
- Graph Data Mining
- Social Network Analytics
- Data Stream Analytics
- Sentiment Analytics and Recommendation Systems
- Supply Chain Data Analytics
- Marketing and Finance Data Analytics

Decision Making

- Discrete Event Simulation and Process Modeling and Control Infrastructure and Programming Environments
- Big Data Middleware: Hadoop, Spark, Graph DB, etc
- Parallel Programming (e.g., with Spark, MPI, etc.)
- IoT

Closing Thoughts

How could these components be prioritized or best conveyed for differing types of data science programs?

- Conference on Undergraduate Research in Data Science
- Senior Design Projects on Data Science
- Internship/job Opportunities in partnership with Data Science and other Industries
- Annual Data Science Hackathons
- Data Science Competitions

How can opportunities to enhance data acumen (i.e., the ability to make good judgments & decisions with data) be integrated into DS educational programs?

- Industry Surveys
- Foundations and Methods: High Priority
- Is data science introduced as an academic enrichment to the existing university curriculum?
  - e.g., Data Science Concentration within the Computer Science Curriculum

How can data acumen be measured or evaluated?

- Individual Course Level Capstone Projects
- Senior Design Project in collaboration with Industry Partners
- Standardized Placement Tests for different Levels of the DS ladder
- Data Science Competition and Contests: perhaps developed in collaboration with industry
- Data Science and DS-related Job Placement Statistics: collected and monitored

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

Acknowledgments

- I would like to thank a number of my colleagues at NC State, UNC-Charlotte, UNC-Chapel Hill and RENCI for discussion and direct or indirect input and contributions as summarized in this presentation. This includes:

  Alyson Wilson, Nagiza Samatova, Rada Chirkova, Patrick Dreher, Jamie Roseborough, Michael Rappa, Christopher Healey, Dan McGurrin, Trey Overman, Stan Ahalt, Andrew Wilson, Mirsad Hadjikadic, Ashok Krishnamurthy, Raju Varsavai, Otis Brown, Mike Kowolenko, Andy Rindos.

- Support for some of the described activities comes in part from NC State University, UNC General Administration, State of North Carolina, a number of USA federal agencies and a number of industrial partners.

Envisioning the DATA SCIENCE DISCIPLINE

The Undergraduate Perspective

Building Data Acumen – Q&A

Nicole Lazar, University of Georgia
Professor, Department of Statistics

Mladen Vouk, North Carolina State University
Distinguished Professor of Computer Science, Associate Vice Chancellor for Research Development and Administration

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

Envisioning the DATA SCIENCE DISCIPLINE

The Undergraduate Perspective

9/12/17 – Building Data Acumen

9/19/17 – Incorporating Real-World Applications

9/26/17 – Faculty Training and Curriculum Development

10/3/17 – Communication Skills and Teamwork

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 and learn more about the study at www.nas.edu/EnvisioningDS