Data science and statistics pathways: workforce issues and implications

Nicholas J. Horton

Department of Mathematics and Statistics
Amherst College, Amherst, MA, USA

CBMS Pathways Forum, May 6, 2019

nhorton@amherst.edu

Overview and motivation

I believe it is the use of imagination and judgment that makes our subject appealing. **We owe it to our students not to keep that a secret** *(Cobb, 1982).*

ASA Undergrad Guidelines for Statistics Programs, 2014

We are concerned that many of our graduates do not have sufficient skills to be effective in the modern workforce. Thomas Lumley (personal communication) has stated that our students know how to deal with $n \to \infty$, but cannot deal with a million observations.

If statistics is the science of learning from data, then our students need to be able to "think with data" (as Diane Lambert of Google has so elegantly described).
- Horton and Hardin (TAS, 2015)

Changing landscape of K-12 statistics education (part 1)

Roxy Peck (JSM 2011) noted:

- statistics have been a recommended part of math curriculum for a long time
- recent developments: considerable more emphasis on statistics
- not just AP statistics: expectation for all students

She was correct: most high school students now see much of what was formerly just part of AP Statistics.

Changing landscape of K-12 statistics education (part 1)

Roxy Peck (JSM 2011) noted:

- statistics have been a recommended part of math curriculum for a long time
- recent developments: considerable more emphasis on statistics
- not just AP statistics: expectation for all students

She was correct: most high school students now see much of what was formerly just part of AP Statistics.

Can't teach multivariate thinking using calculators

Big Idea 3: Data and Information

Data and information facilitate the creation of knowledge. Computing enables and empowers new methods of information processing, driving monumental change across many disciplines — from art to business to science. Managing and interpreting an overwhelming amount of raw data is part of the foundation of our information society and economy. People use computers and computation to translate, process, and visualize raw data and to create information. Computation and computer science facilitate and enable new understanding of data and information that contributes knowledge to the world. Students in this course work with data using a variety of computational tools and techniques to better understand the many ways in which data is transformed into information and knowledge.

AP Computer Science Principles

Read the Introducing AP Computer Science Principles video transcript

Computer Science: The New Literacy

Whether it's 3-D animation, engineering, music, app development, medicine, visual design, robotics, or political analysis, computer science is the engine that powers the technology, productivity, and innovation that drive the world. Computer science experience has become an imperative for today's students and the workforce of tomorrow.

The AP Program designed AP Computer Science Principles with the goal of creating leaders in computer science fields and attracting and engaging those who are traditionally underrepresented with essential computing tools and multidisciplinary opportunities.

In 2017, largest first year AP exam ever (45,000 students)
In 2018 more than 83,000 students took the exam

AP Computer Science Principles

Enduring Understandings (Students will understand that ...)	**Learning Objectives** (Students will be able to ...)
EU 3.1 People use computer programs to process information to gain insight and knowledge.	**LO 3.1.1** Find patterns and test hypotheses about digitally processed information to gain insight and knowledge. [P4]

Changing landscape of college-level statistics education

- revised GAISE College report and the role of *multivariate thinking*
- dramatically improved open-source tools (R/RStudio and Python)
- simplified interfaces to decrease cognitive burden on students (and instructors, see Pruim et al, R Journal, 2017)
- cloud computing to faciliate workflow and reproducibility (Cetinkaya-Rundel and Rundel, TAS 2018)
- growth of data science at community colleges (see report from the Two Year College Data Science Summit)

Changing landscape of two year college data science education

Two Year College Data Science Summit (NSF funded, `https://www.amstat.org/ASA/Education/Two-Year-College-Data-Science-Summit.aspx`)

- certificate programs
- associates to workforce
- associates to transfer

Changing landscape of two year college data science education

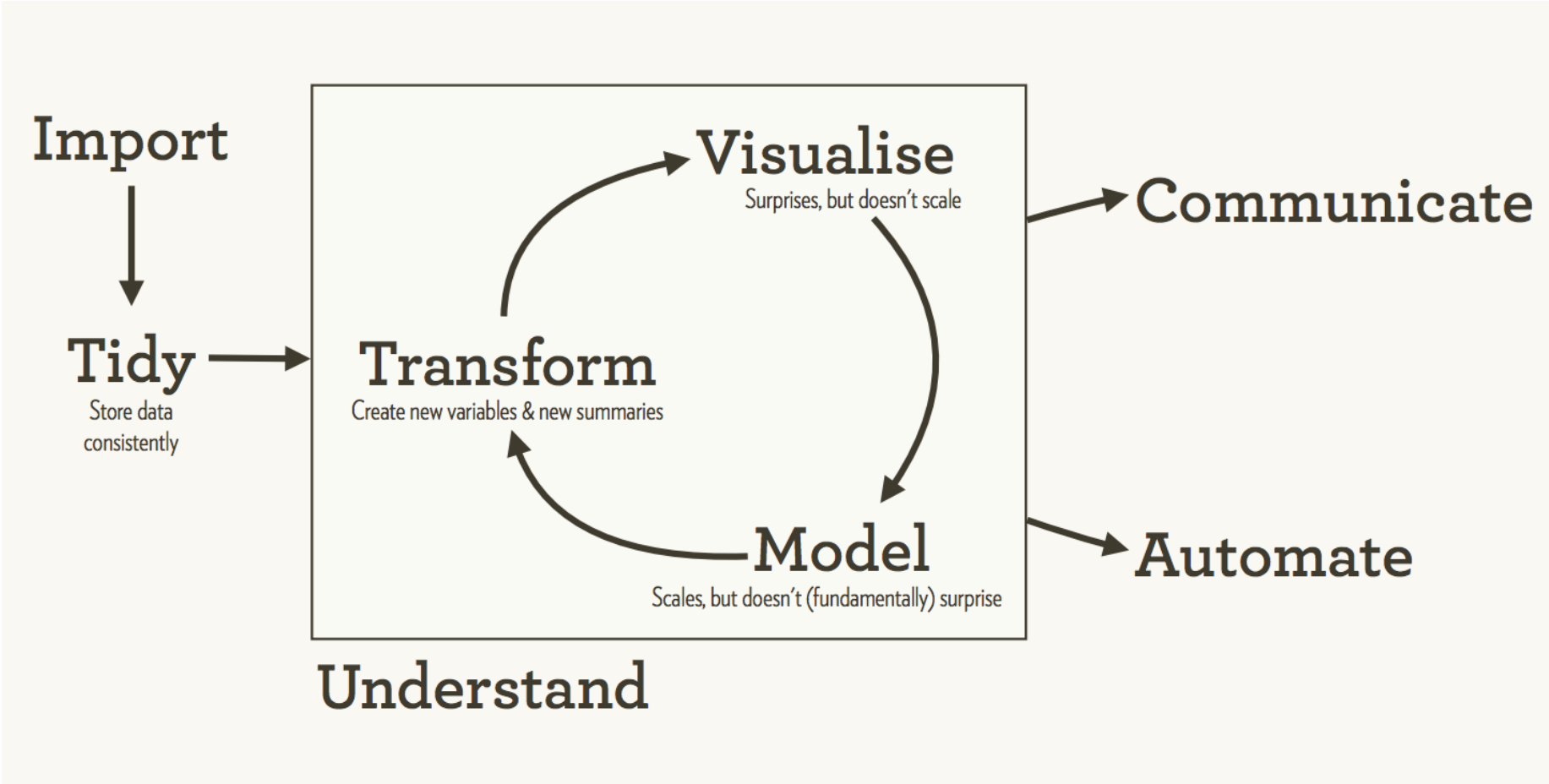
Two Year College Data Science Summit (NSF funded, `https://www.amstat.org/ASA/Education/Two-Year-College-Data-Science-Summit.aspx`)

- Computational Foundations
- Computational Thinking
- Statistical Foundations
- Statistical Thinking
- Statistical Modeling
- Data Management and Curation
- Mathematical Foundations
- Productivity Foundations

Workforce implications

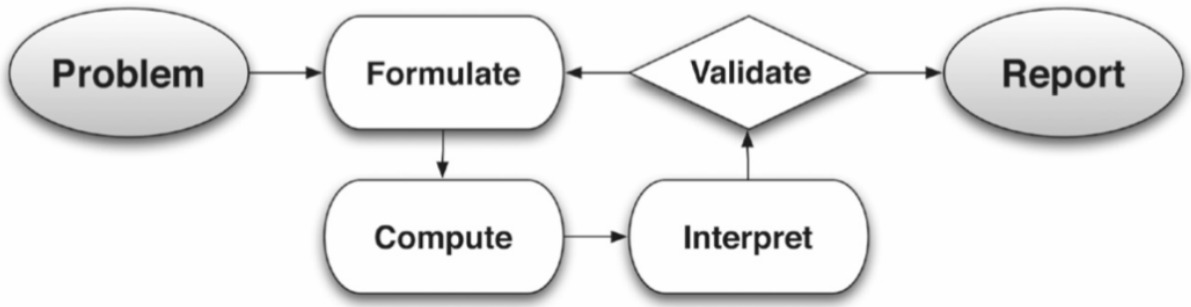
- NASEM Data Science report: "Recommendation 3.1: Four-year and two-year institutions should establish a forum for dialogue across institutions on all aspects of data science education, training, and workforce development."
- NASEM Data Science report: "Recommendation 4.1: As data science programs develop, they should focus on attracting students with varied backgrounds and degrees of preparation and preparing them for success in a variety of careers."
- Two Year College Summit: "Programs for students looking to go directly into the workforce will have more flexibility in the courses they can offer, but they will need to identify and collaborate with local industry partners to create programs that will ensure employability of their graduates."

Data analysis cycle (Wickham and Grolemund)



OVERVIEW

In Grade 8, students used functions for the first time to construct a function that models a linear relationship between two quantities (**8.F.B.4**) and to describe qualitatively the functional relationship between two quantities by analyzing a graph (**8.F.B.5**). In the first four modules of Algebra I, students learn to create and apply linear, quadratic, and exponential functions in addition to square and cube root functions (**F-IF.C.7**). In Module 5, they synthesize what they have learned during the year by selecting the correct function type in a series of modeling problems without the benefit of a module or lesson title that includes function type to guide them in their choices. This supports the CCLS requirement that student's use the modeling cycle, in the beginning of which they must formulate a strategy. Skills and knowledge from the previous modules support the requirements of this module, including writing, rewriting, comparing, and graphing functions (**F-IF.C.7**, **F-IF.C.8**, **F-IF.C.9**) and interpretation of the parameters of an equation (**F-LE.B.5**). Students also draw on their study of statistics in Module 2, using graphs and functions to model a context presented with data and tables of values (**S-ID.B.6**). In this module, we use the modeling cycle (see page 72 of the CCLS) as the organizing structure rather than function type.



Should all statistics students be programmers?

No!

July 2018

Hadley Wickham
@hadleywickham
Chief Scientist, **RStudio**

Should all statistics students program?

Yes!

July 2018

Hadley Wickham
@hadleywickham
Chief Scientist, **RStudio**

Key thoughts

- add more multivariate thinking (and multiple regression) in intro stats (multiple regression needs a prominent place)
- adopt a "Less Volume, More Creativity" approach to technology
- add causal inference learning outcomes to later courses
- use project-based learning to teach statistics and data science analysis cycle and reproducible workflows
- de-emphasize p-values (e.g., Allen Downey's "Inference in Three Hours, and More Time for the Good Stuff") to make room
- develop multiple mathematical pathways to address needs for "data acumen" (`https://nas.edu/envisioningds`)

Data science and statistics pathways: workforce issues and implications

Nicholas J. Horton

Department of Mathematics and Statistics
Amherst College, Amherst, MA, USA

CBMS Pathways Forum, May 6, 2019

nhorton@amherst.edu

AP Computer Science Principles

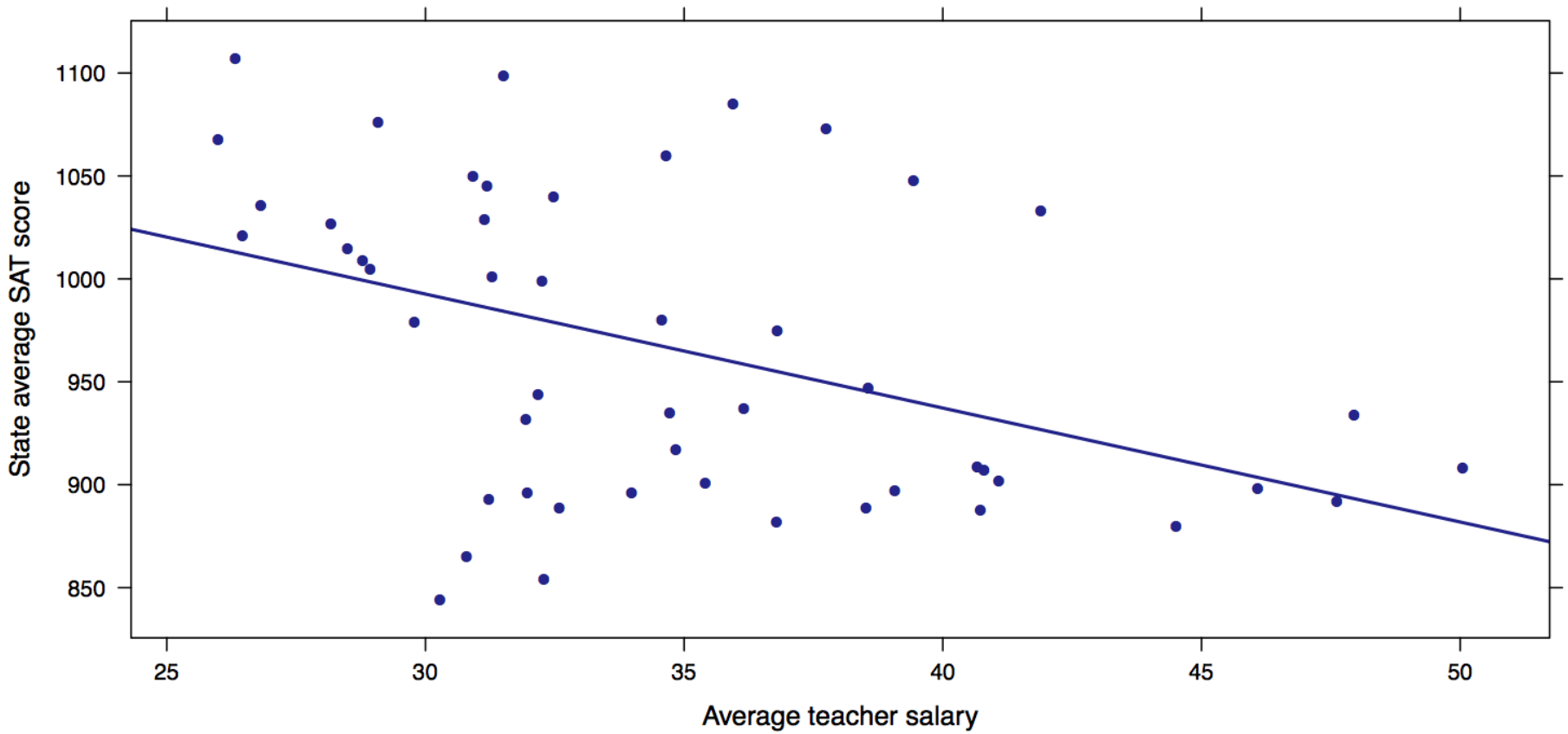
LO 3.1.3 Explain the insight and knowledge gained from digitally processed data by using appropriate visualizations, notations, and precise language. [P5]	**EK 3.1.3A** Visualization tools and software can communicate information about data.
	EK 3.1.3B Tables, diagrams, and textual displays can be used in communicating insight and knowledge gained from data.
	EK 3.1.3C Summaries of data analyzed computationally can be effective in communicating insight and knowledge gained from digitally represented information.
	EK 3.1.3D Transforming information can be effective in communicating knowledge gained from data.
	EK 3.1.3E Interactivity with data is an aspect of communicating.

AP Computer Science Principles

EU 3.2 Computing facilitates exploration and the discovery of connections in information.

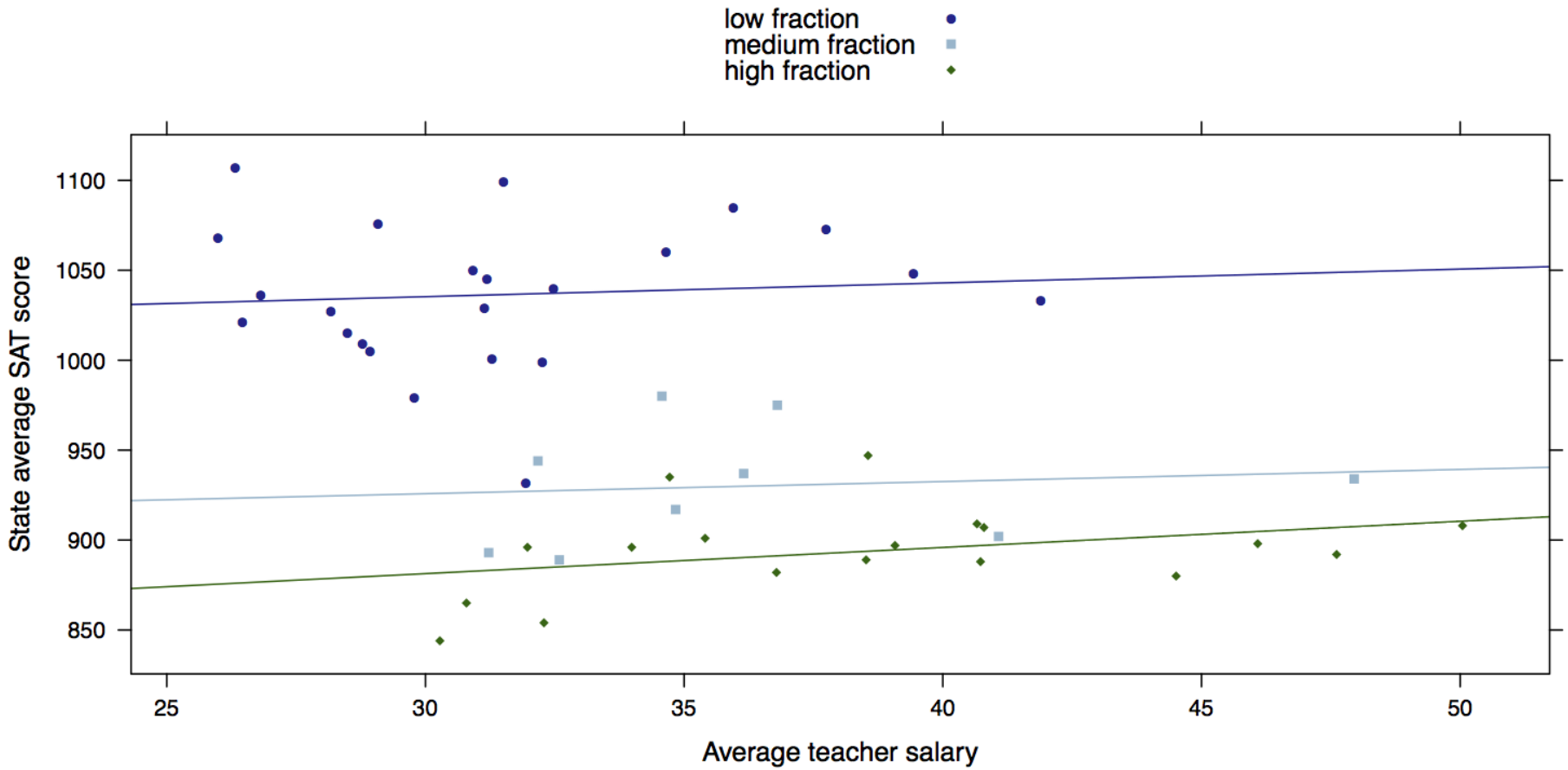
LO 3.2.1 Extract information from data to discover and explain connections or trends. [P1]

Multivariate thinking and confounding



College entrance scores and teacher salaries (US state data from 2010)

Multivariate thinking and confounding



Data Science for Undergraduates Mathematical Concepts/Skills

Key mathematical concepts/skills that would be important for all students in their data science programs and critical for their success in the workforce are the following:

- Set theory and basic logic
- Multivariate thinking via functions and graphical displays
- Basic probability theory and randomness
- Matrices and basic linear algebra
- Networks and graph theory and
- Optimization