Data and Computing in K12: implications for undergraduate data science education

Nicholas J. Horton, Amherst College

May 20, 2025, nhorton@amherst.edu

Image source: heylagostechie

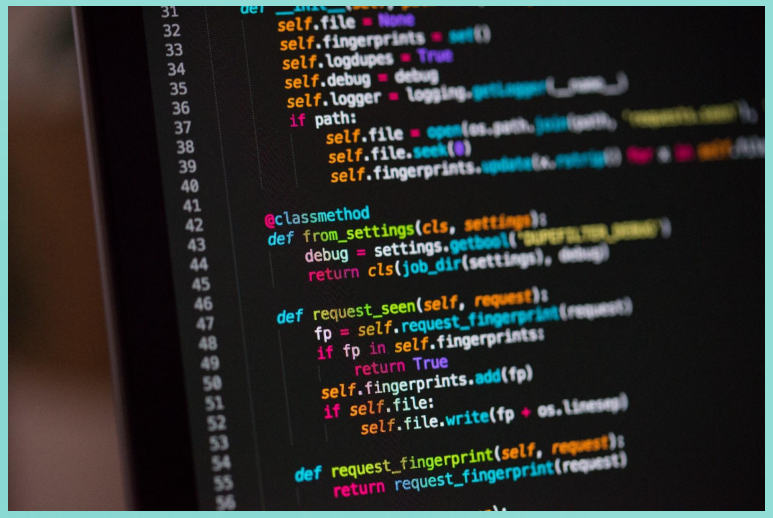


Image source: Wikicommons

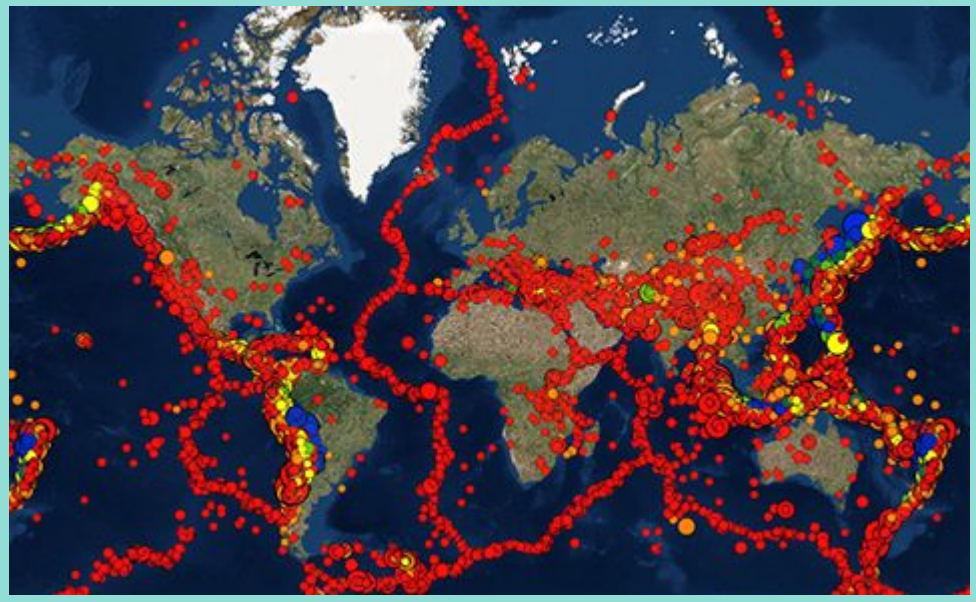


Image source: Concord Consortium

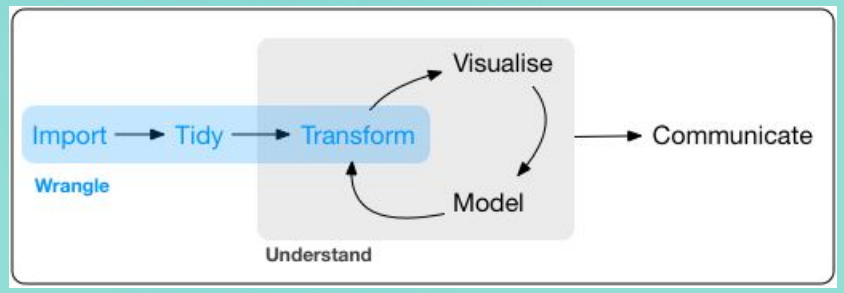


Image source: Hadley Wickham and Garrett Grolemund

Links at https://nicholasjhorton.github.io/K12-Data-Tools/icerm.html

Plan

- Insights about data acumen from the NASEM (2018) report
- Growth of K-12 data science
- What should we teach?
- Implications for undergraduate data science

DATA SCIENCE FOR UNDERGRADUATES

Opportunities and Options

consensus report published in 2018
free download from
https://nas.edu/envisioningds

Study funded by the National Science Foundation

The National Academies of SCIENCES ENGINEERING MEDICINE

nas.edu/EnvisioningDS

Key Insights NASEM (2018): Undergraduate Data Science

- There must be **multiple pathways** for undergraduates to study data science
- The undergraduate experience should cater to and **promote diversity** – demographic and intellectual – in the students it serves
- There are some core competencies that all data science students (and, ideally, all undergraduates) should have
  - They should develop **data acumen**
  - Ethical problem-solving is a key component of data acumen

A Central Finding

Finding 2.3 A critical task in the education of future data scientists is to instill data acumen. This requires exposure to key concepts in data science, real-world data and problems that can reinforce the limitations of tools, and ethical considerations that permeate many applications. Key concepts involved in developing data acumen include the following:

- Mathematical foundations
- Computational foundations
- Statistical foundations
- Data management and curation
- Data description and visualization
- Data modeling and assessment
- Workflow and reproducibility
- Communication and teamwork
- Domain-specific considerations
- Ethical problem solving.

Mathematical concepts

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.

Computational concepts

While it would be ideal for all data scientists to have extensive coursework in computer science, new pathways may be needed to establish appropriate depth in **algorithmic thinking and abstraction** in a streamlined manner. This might include the following:

- Basic abstractions,
- Algorithmic thinking,
- Programming concepts,
- Data structures, and
- Simulations.

Statistical concepts

Important **statistical foundations** might include the following:

- Variability, uncertainty, sampling error, and inference;
- Multivariate thinking;
- Nonsampling error, design, experiments (e.g., A/B testing), biases, confounding, and causal inference;
- Exploratory data analysis;
- Statistical modeling and model assessment; and
- Simulations and experiments

Data management concepts

Key **data management and curation** 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:

- Data provenance;
- Data preparation, especially data cleansing and data transformation;
- Data management (of a variety of data types);
- Record retention policies;
- Data subject privacy;
- Missing and conflicting data; and
- Modern databases.

Data visualization concepts

Key **data description and visualization** 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:

- Data consistency checking,
- Exploratory data analysis,
- Grammar of graphics,
- Attractive and sound static visualizations,
- Dynamic visualizations and dashboards.

Data modeling concepts

Key **data modeling and assessment** 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:

- Machine learning,
- Multivariate modeling and supervised learning,
- Dimension reduction techniques and unsupervised learning,
- Deep learning,
- Model assessment and sensitivity analysis, and
- Model interpretation (particularly for black box models).

Workflow and reproducibility concepts

Key **workflow and reproducibility** 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:

- Workflows and workflow systems,
- Reproducible analysis,
- Documentation and code standards,
- Source code (version) control systems, and
- Collaboration.

Communication and teamwork concepts

Key **communication and teamwork** 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:

- Ability to understand client needs,
- Clear and comprehensive reporting,
- Conflict resolution skills,
- Well-structured technical writing without jargon, and
- Effective presentation skills.

Ethical concepts

Key aspects of **ethics** needed for all data scientists (and for that matter, all educated citizens) include the following:

- Ethical precepts for data science and codes of conduct,
- Privacy and confidentiality,
- Responsible conduct of research,
- Ability to identify "junk" science, and
- Ability to detect algorithmic bias.

Developing data acumen is hard!

"Integrating Computing in the Statistics and Data Science Curriculum: Creative Structures, Novel Skills and Habits, and Ways to Teach Computational Thinking" (Horton and Hardin, *Journal of Statistics and Data Science Education*, 2021):

https://www.tandfonline.com/doi/full/10.1080/10691898.2020.1870416

- We need to think creatively about how to give undergraduate students repeated practice with the entire data science analysis cycle
- Requires new courses to fill in gaps
- Requires reformulation of other courses to develop data acumen
- Think back to yesterday's talks and what's needed for students to engage!

Problem-solving cycle

- **What are we hoping that students will learn?**

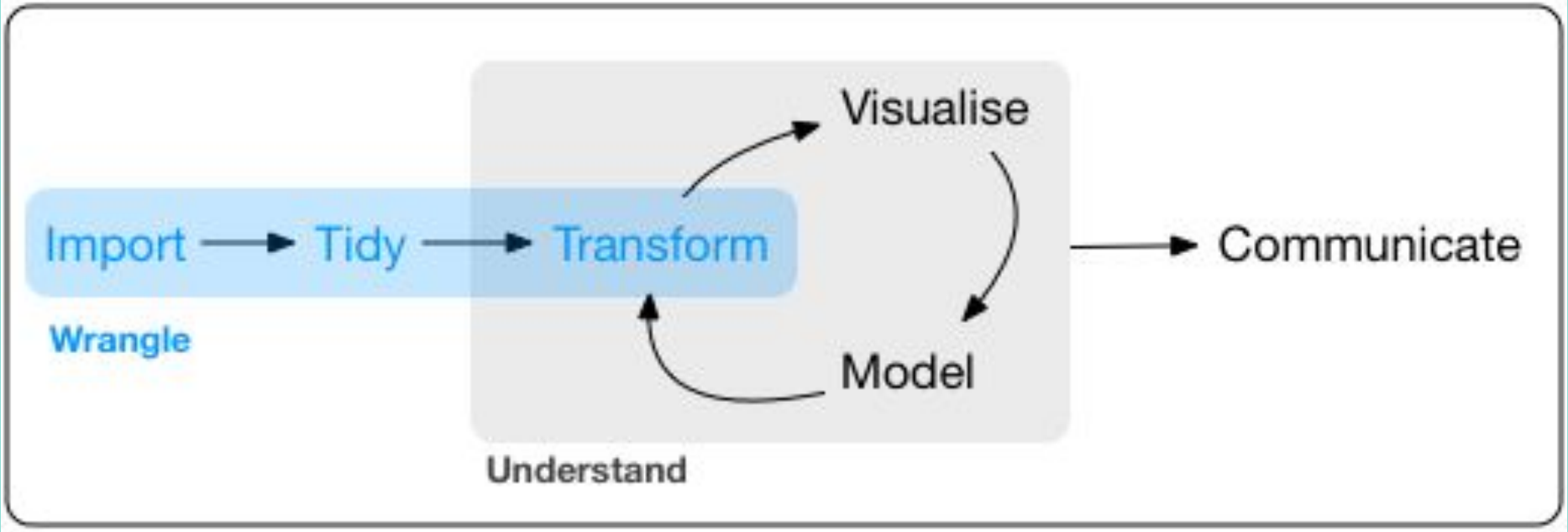
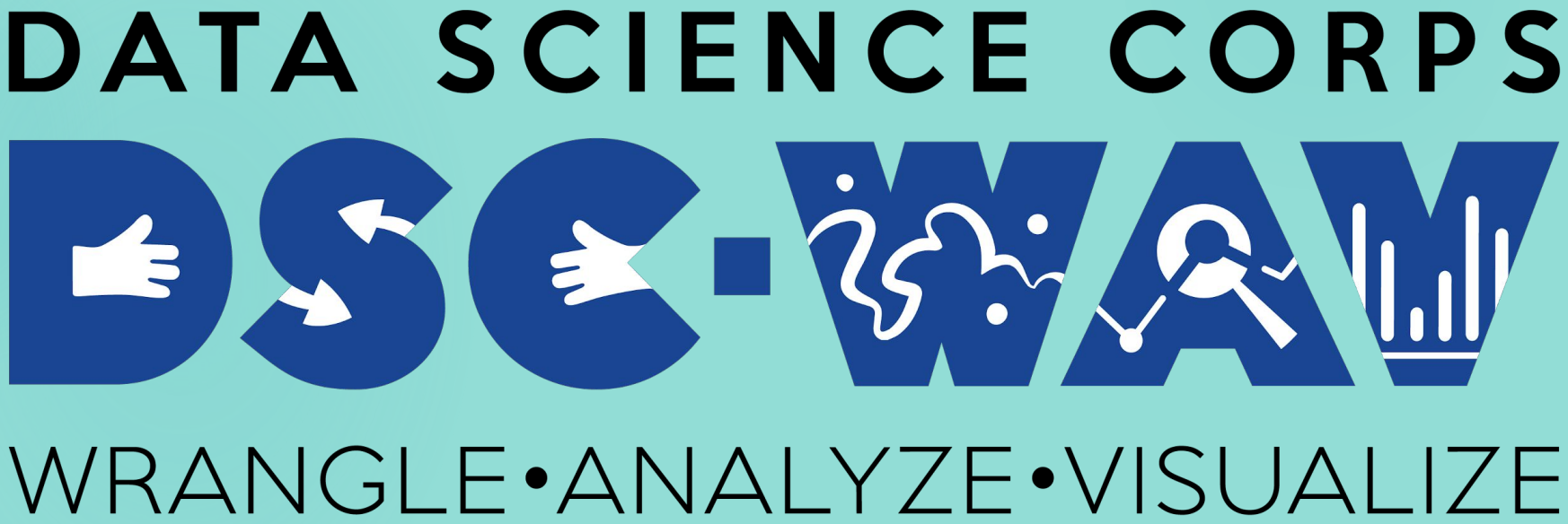


Image source: Hadley Wickham and Garrett Grolemund

DSC-WAV (Wrangle-Analyze-Visualize)

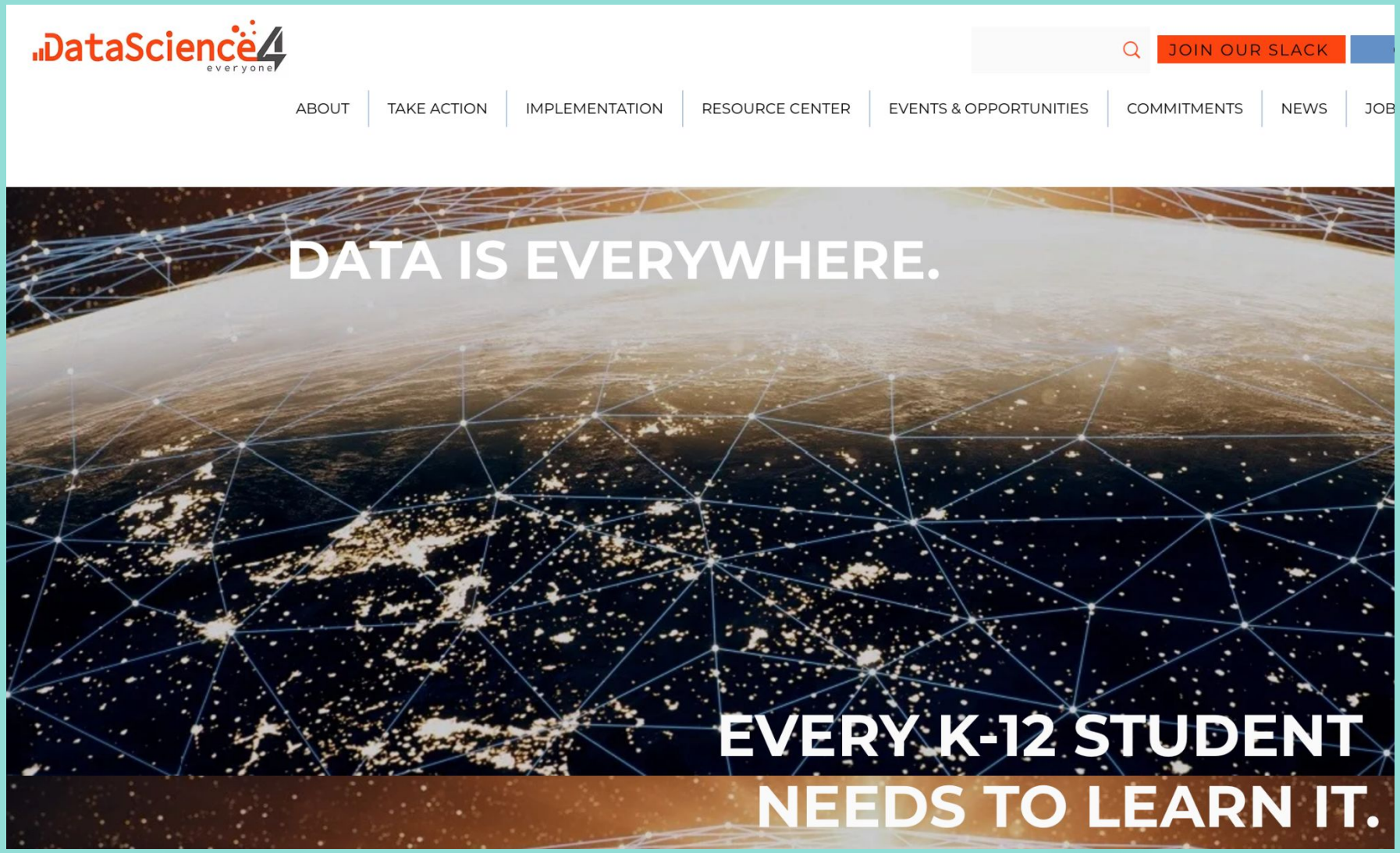
- NSF funded effort from the Harnessing the Data Revolution (HDR) Data Science Corps (DSC) initiative (Ben Baumer, PI): https://dsc-wav.github.io/www



Growth of K-12 Data Science

- In a world defined by data, we can't wait to introduce students in K-12 to the opportunities (and challenges) in making sense of it
- Increasing growth of K12 Data Science:
  - NASEM workshop, https://www.nationalacademies.org/our-work/foundations-of-data-science-for-students-in-grades-k-12-a-workshop
  - GAISE II, https://www.amstat.org/education/guidelines-for-assessment-and-instruction-in-statistics-education-(gaise)-reports
  - Next Generation Science Standards (NGSS), https://www.nextgenscience.org

Growth of K-12 Data Science

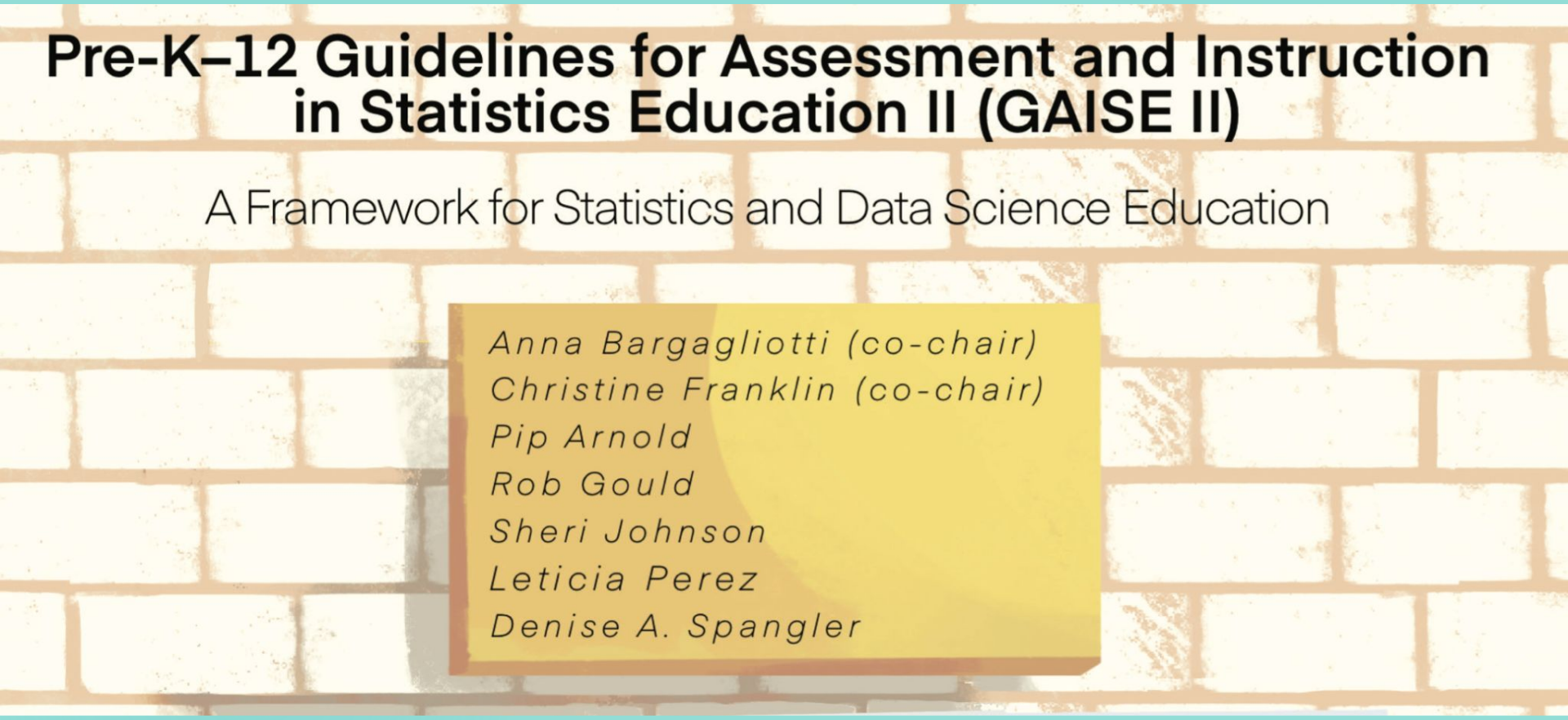


Revised Guidelines for Assessment and Instruction in Statistics Education (GAISE) COLLEGE report (2016)

- Teach statistical thinking.
  - Teach statistics as an investigative process of problem-solving and decision-making.
  - Give students experience with multivariable thinking.
- Focus on conceptual understanding.
- Integrate real data with a context and purpose.
- Foster active learning.
- Use technology to explore concepts and analyze data.
- Use assessments to improve and evaluate student learning.

https://www.amstat.org/education/guidelines-for-assessment-and-instruction-in-statistics-education-(gaise)-reports

Revised K-12 GAISE Guidelines



Original K-12 report written in 2005, published in 2007, revised (and renamed "GAISE II") in 2020

Revised Guidelines for Assessment and Instruction in Statistics Education PreK-12 [GAISE II] report (2020)

- Importance of questioning through the problem-solving cycle (see Lee et al, SERJ, 2022)
- Importance of design and considering different data types
- Inclusion of multivariate thinking
- Role of probabilistic thinking
- Shifts and deepening of technology
- Importance of communication

https://www.amstat.org/education/guidelines-for-assessment-and-instruction-in-statistics-education-(gaise)-reports

National Academies report

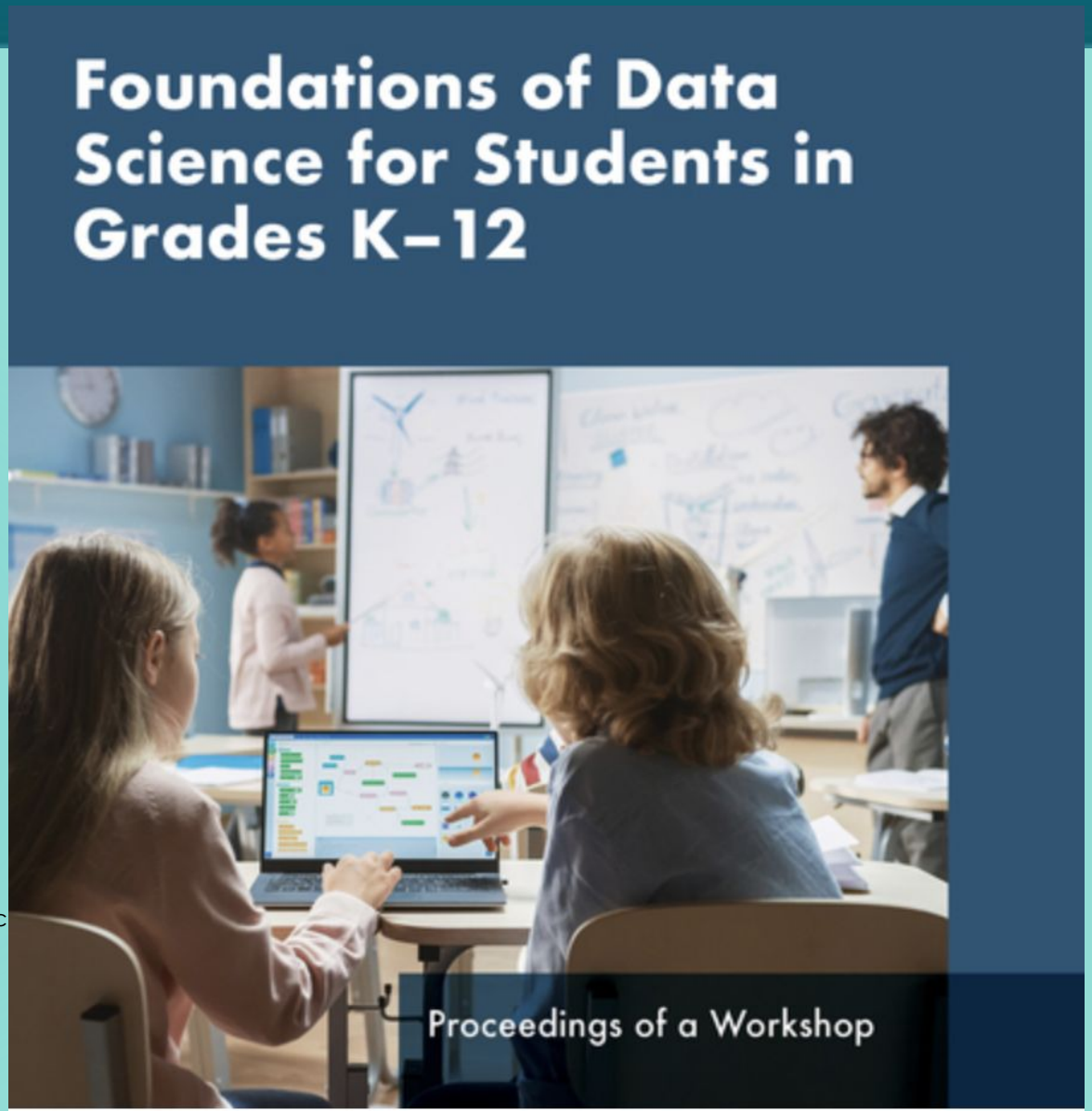


Image source: Hadley Wic

Next Generation Science Standards (NGSS, 2013)

See for example: MS-LS2-1 Ecosystems: Interactions, Energy, and Dynamics

Students who demonstrate understanding can:

MS-LS2-1. **Analyze and interpret data to provide evidence for the effects of resource availability on organisms and populations of organisms in an ecosystem.** [Clarification Statement: Emphasis is on cause and effect relationships between resources and growth of individual organisms and the numbers of organisms in ecosystems during periods of abundant and scarce resources.]

The performance expectation above was developed using the following elements from the NRC document *A Framework for K-12 Science Education*:

Science and Engineering Practices

Analyzing and Interpreting Data
Analyzing data in 6–8 builds on K–5 experiences and progresses to extending quantitative analysis to investigations, distinguishing between correlation and causation, and basic statistical techniques of data and error analysis.

- Analyze and interpret data to provide evidence for phenomena.

Disciplinary Core Ideas

LS2.A: Interdependent Relationships in Ecosystems

- Organisms, and populations of organisms, are dependent on their environmental interactions both with other living things and with nonliving factors.
- In any ecosystem, organisms and populations with similar requirements for food, water, oxygen, or other resources may compete with each other for limited resources, access to which consequently constrains their growth and reproduction.
- Growth of organisms and population increases are limited by access to resources.

Crosscutting Concepts

Cause and Effect

- Cause and effect relationships may be used to predict phenomena in natural or designed systems.

Connections to other DCIs in this grade-band:
MS.ESS3.A ; **MS.ESS3.C**

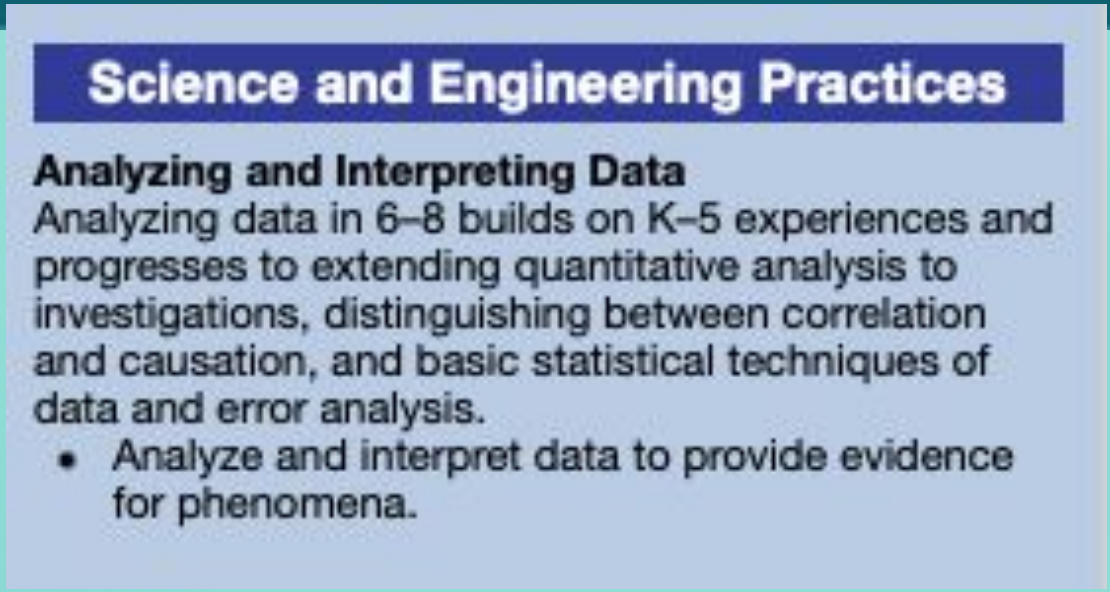
Articulation of DCIs across grade-bands:
3.LS2.C ; **3.LS4.D** ; **5.LS2.A** ; **HS.LS2.A** ; **HS.LS4.C** ; **HS.LS4.D** ; **HS.ESS3.A**

Common Core State Standards Connections:
ELA/Literacy -

RST.6-8.1 Cite specific textual evidence to support analysis of science and technical texts. (MS-LS2-1)
RST.6-8.7 Integrate quantitative or technical information expressed in words in a text with a version of that information expressed visually (e.g., in a flowchart, diagram, model, graph, or table). (MS-LS2-1)

https://www.nextgenscience.org

Next Generation Science Standards (NGSS, 2013)



Science and Engineering Practices

Analyzing and Interpreting Data

Analyzing data in 6–8 builds on K–5 experiences and progresses to extending quantitative analysis to investigations, distinguishing between correlation and causation, and basic statistical techniques of data and error analysis.

- Analyze and interpret data to provide evidence for phenomena.

Common Core State (Math) Standards Connections:

RST.6-8.1 Cite specific textual evidence to support analysis of science and technical texts. (MS-LS2-1)

RST.6-8.7 Integrate quantitative or technical information expressed in words in a text with a version of that information expressed visually (e.g., in a flowchart, diagram, model, graph, or table). (MS-LS2-1)

Computer Science Teachers Association K-12 Computing Standards (2017): 1A-DA-06

K-2	**Collect and present the same data in various visual formats.**  *The collection and use of data about the world around them is a routine part of life and influences how people live. Students could collect data on the weather, such as sunny days versus rainy days, the temperature at the beginning of the school day and end of the school day, or the inches of rain over the course of a storm. Students could count the number of pieces of each color of candy in a bag of candy, such as Skittles or M&Ms. Students could create surveys of things that interest them, such as favorite foods, pets, or TV shows, and collect answers to their surveys from their peers and others. The data collected could then be organized into two or more visualizations, such as a bar graph, pie chart, or pictograph.* **Practice(s):** Communicating About Computing, Developing and Using Abstractions: 7.1, 4.4	Data & Analysis

Computer Science Teachers Association K-12 Computing Standards (2017): 1A-DA-07

K-2	**Identify and describe patterns in data visualizations, such as charts or graphs, to make predictions.** *Data can be used to make inferences or predictions about the world. Students could analyze a graph or pie chart of the colors in a bag of candy or the averages for colors in multiple bags of candy, identify the patterns for which colors are most and least represented, and then make a prediction as to which colors will have most and least in a new bag of candy. Students could analyze graphs of temperatures taken at the beginning of the school day and end of the school day, identify the patterns of when temperatures rise and fall, and predict if they think the temperature will rise or fall at a particular time of the day, based on the pattern observed.* **Practice(s):** Developing and Using Abstractions: 4.1	Data & Analysis

Computer Science Teachers Association K-12 Computing Standards (2017): 1B-IC-18

3-5	**Discuss computing technologies that have changed the world, and express how those technologies influence, and are influenced by, cultural practices.** *New computing technology is created and existing technologies are modified for many reasons, including to increase their benefits, decrease their risks, and meet societal needs. Students, with guidance from their teacher, should discuss topics that relate to the history of technology and the changes in the world due to technology. Topics could be based on current news content, such as robotics, wireless Internet, mobile computing devices, GPS systems, wearable computing, or how social media has influenced social and political changes.* **Practice(s):** Recognizing and Defining Computational Problems: 3.1	Impacts of Computing

Mathematics Standards + Curriculum

- In 2011 Roxy Peck noted that there's now considerable statistics (EDA, informal inference, formal inference) in the K-12 Math curriculum

NATIONAL IMPACT

Eureka Math is the most widely used math curriculum in the United States, according to a **study** released by the RAND Corporation. It is also the only curriculum found by **EdReports.org** to align fully with the Common Core State Standards for all grades, K–8. Additionally, over a dozen lessons from *Eureka Math* were rated to be EQuIP exemplars by **Achieve**.

Eureka Math (K-12)

For a given sample, you can find the sample mean:

1. There is variability in the sample mean
2. A graph of the distribution of sample means from many random samples is a simulated sampling distribution
3. Sample means from random samples tend to cluster around the value of the population mean.
4. The variability in the sample mean decreases as the sample size increases.
5. Most sample means are within two standard deviations of the mean of the simulated sampling distributions.

Eureka Math (K-12)

When a single set of values is randomly divided into two groups

1. the two group means will tend to differ just by chance
2. The distribution of random groups' means will be centered at the single set's mean
3. the range of the distribution of the random groups' means will be smaller than the range of the original data
4. the shape of the distribution of the random groups' means will be symmetric

What foundation can we start to assume in college?

Next steps for data science education

- Focus on computational thinking early and often (key role of multivariate thinking and data acumen)
- Embrace simplified computational interfaces and approaches to minimize cognitive load and scaffold reproducibility
- Embrace cloud computing to minimize barriers to technology
- Integrate and adopt high impact practices and active learning techniques (e.g., pair programming, group- and project- based learning)
- Creatively scale up faculty development and training

Developing Competencies for the Future of Data and Computing: The Role of K-12

Back to NASEM (2018)

Recommendation 2.1: Academic institutions should embrace data science as a vital new field that requires specifically tailored instruction delivered through majors and minors in data science as well as the development of a cadre of faculty equipped to teach in this new field.

Recommendation 2.2: Academic institutions should provide and evolve a range of educational pathways to prepare students for an array of data science roles in the workplace.

Back to NASEM (2018)

Recommendation 2.3: To prepare their graduates for this new data-driven era, **academic institutions should encourage the development of a basic understanding of data science in all undergraduates**.

Good news: we can build on what's happening in K-12!

What should we teach?

Recommendation 2.3: To prepare their graduates for this new data-driven era, **academic institutions should encourage the development of a basic understanding of data science in all undergraduates**.

Good news: we can build on what's happening in K-12!

What should we teach?

OPINION

Ensuring That Mathematics is Relevant in a World of Data Science

Johanna S. Hardin and Nicholas J. Horton

Notices of the AMS (October 2017)

What should we teach (on the math side)?

- Proposed New Course 1—Mathematical Foundations I: Discrete Mathematics
- Proposed New Course 2—Mathematical Foundations II: Continuous Mathematics
- The Importance of Computing

See also the Park City Guidelines (Annual Reviews, 2017)

Shameless plug: JSDSE

- The *Journal of Statistics and Data Science Education* is a 33-year old open-access journal with no author publication fees published by the American Statistical Association and Taylor & Francis

- More information and content can be found at: https://www.tandfonline.com/toc/ujse21/current

- Submissions welcomed

Shameless plug: HDSR

- The *Harvard Data Science Review* is a 6-year old open-access journal with no author publication fees published by MIT Press
- More information and content can be found at: https://hdsr.mitpress.mit.edu
- Submissions (in the form of a two page proposal) welcomed

Data and Computing in K12: implications for undergraduate data science education

Nicholas J. Horton, Amherst College

May 20, 2025, nhorton@amherst.edu

Image source: heylagostechie

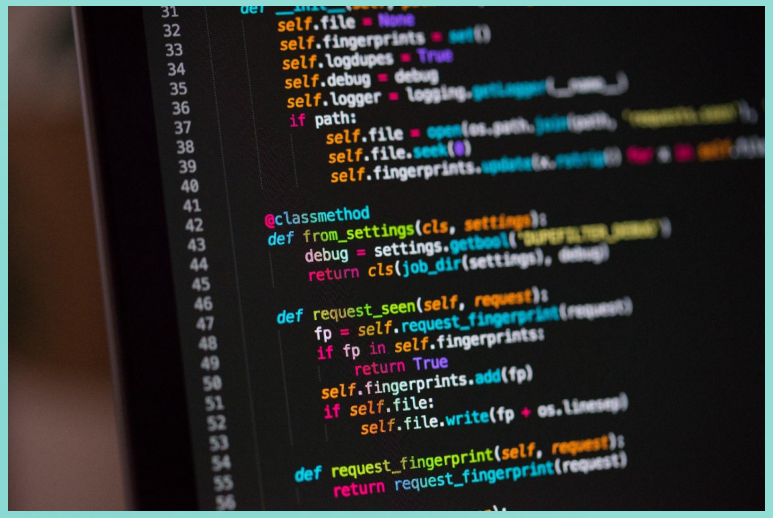


Image source: Wikicommons

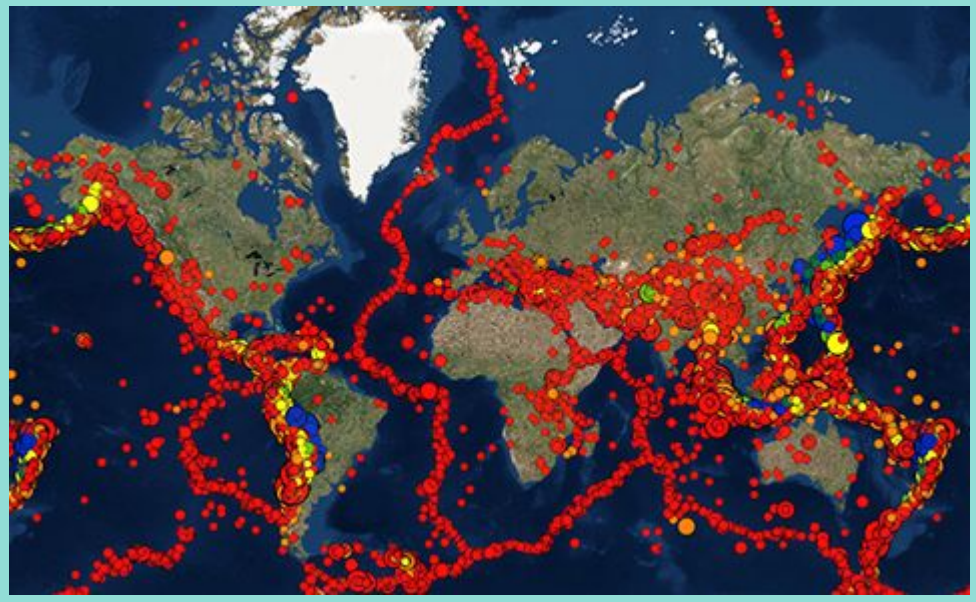


Image source: Concord Consortium

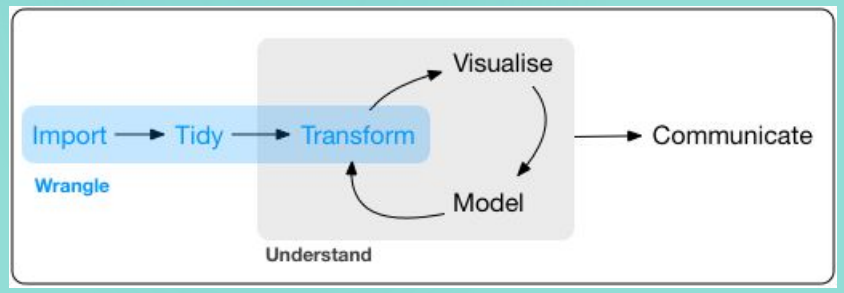


Image source: Hadley Wickham and Garrett Grolemund

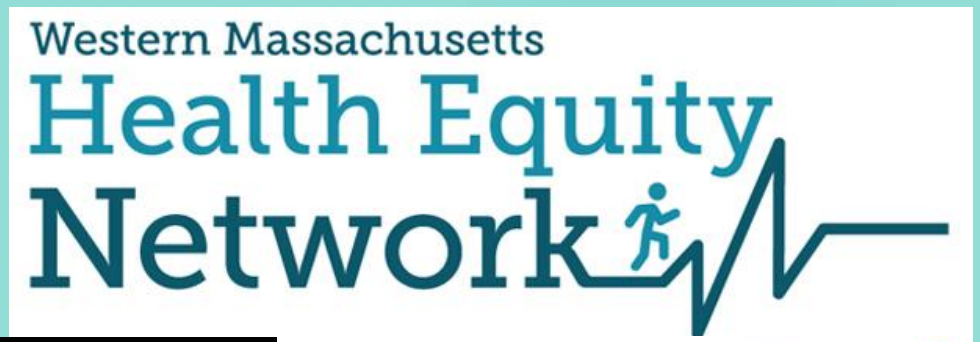
Links at https://nicholasjhorton.github.io/K12-Data-Tools/icerm.html

DSC-WAV (Wrangle-Analyze-Visualize)

- https://dsc-wav.github.io/www
- Collaborative project with Five Colleges (Amherst, Smith, Hampshire, Mount Holyoke, and UMass/Amherst), Greenfield Community College, Holyoke Community College, Springfield Technical Community College, and the University of Minnesota

DSC-WAV (Wrangle-Analyze-Visualize)

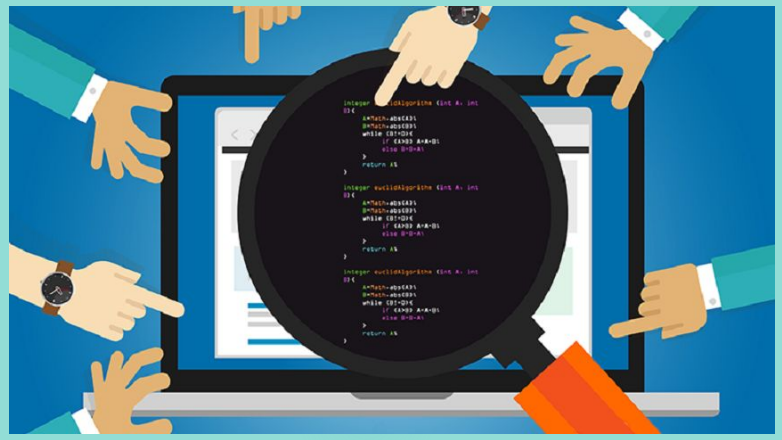
- Goal 1: create opportunities for undergraduate students to work on Data Science for Social Good projects for community organizations



Agile and scrum for undergraduates

- Horton et al (2021, HDSR)
  https://hdsr.mitpress.mit.edu/pub/nvflcexe/release/1

"While many of these courses and programs teach students relevant data science skills, we can expect coursework to develop students' data acumen only so far. It is unclear whether coursework alone is enough to provide students with the experiences with data and computing they need to be successful in tomorrow's workplace."



Source: techgig.com

Agile and scrum for undergraduates

- The work on the project is organized into a series of short sprints to break up tasks.
- Subtasks are organized into a **backlog** to identify priorities for that stage =
- The team and stakeholders (faculty and community organization liaison) meet regularly (**standups**) to share results and make adjustments
- **Kanban** project boards, implemented using Trello or GitHub Projects, are used to review the backlog and team progress.
- **Code review**, implemented using GitHub pull requests, is included as a regular part of the process.
- **Sprint demos** are places where current results are presented and discussed in the context of the broader goals of the project.
- **Sprint retrospectives** are used to identify issues with the process and ways that the team might improve their work.

Agile and scrum for undergraduates

- "Facilitating team-based data science: Lessons learned from the DSC-WAV project", *Foundations of Data Science* (Legacy et al, https://www.aimsciences.org/article/doi/10.3934/fods.2022003

The inspiration for the DSC-WAV program was a question of whether undergraduate students could tackle real-world data science problems utilizing the tools and approaches frequently seen in industry. Based on our experiences, the answer to this question is "yes."

Source: smartbear.com

Source:Esti Alvarez, see also https://teachdatascience.com/pairprogramming

DSC-WAV Lessons Learned

- Many challenges to helping undergraduate students develop the ability to "think with data"
- Our courses and programs need to adapt to give them necessary workforce skills as analysts
- DSC-WAV projects have provided a starting point but more reinforcement is needed
- Lots of work needed to scale out programs at two- and four-year schools

DATA SCIENCE CORPS

DSC-WAV

WRANGLE•ANALYZE•VISUALIZE

Data Tools for K12 Data Science

- **What are we hoping that students will learn?**

python™

BOOTSTRAP
Equity • Scale • Rigor

Bootstrap:Data Science

Evidence-based, integrated materials for grade 7-12 Social Studies classes

- Leverage students' curiosity about the world around them to inspire real data analysis and original research.
- Lessons are available for data visualization, measures of center and spread, programming, linear regression, and more.
- Mix and match to create anything from a one-week intro to a full-year course!

R

Prior work on data tools

- Biehler (1997, ISR) "Software for Learning and for Doing Statistics", https://doi.org/10.1111/j.1751-5823.1997.tb00399.x
- McNamara (2019, TAS) "Key attributes of a modern statistical computing tool", https://www.tandfonline.com/doi/full/10.1080/00031305.2018.1482784, see also https://arxiv.org/abs/1610.00984
- Pimental, Wilkerson, and Horton adapted the framework of McNamara to account for considerations specific to K12 education (paper, data, code, illustrated examples, and interactive links available at: https://nicholasjhorton.github.io/K12-Data-Tools/dsd.html)