Data science for undergraduates: Building data science curricula and connections between two- and four- year institutions

Nicholas J. Horton, Amherst College

June 16, 2021, nhorton@amherst.edu

Image source: heylagostechie

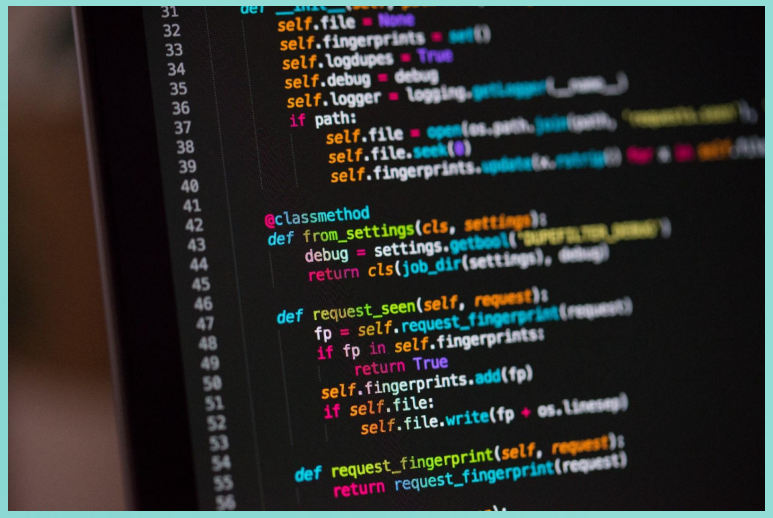


Image source: Wikicommons

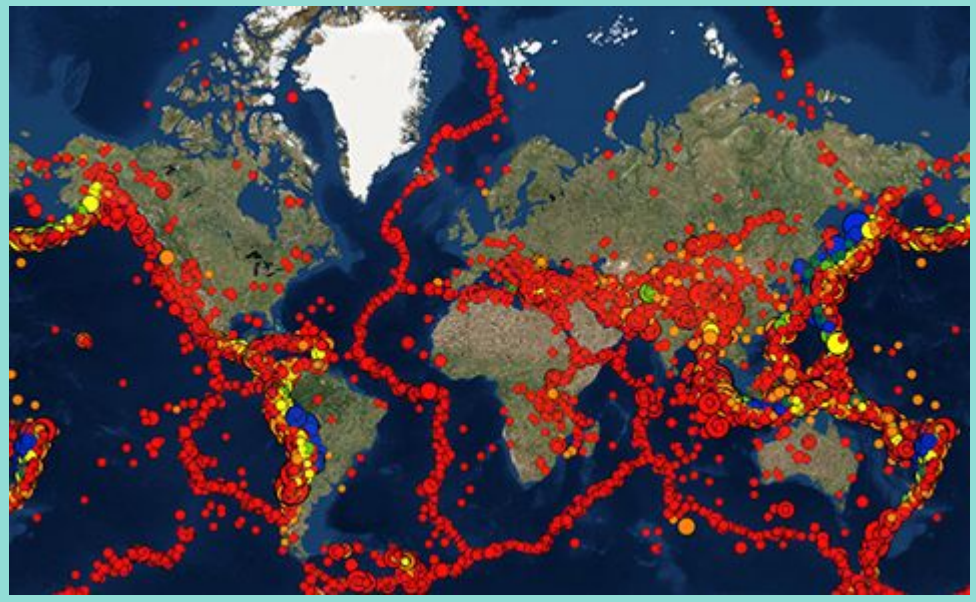


Image source: Concord Consortium

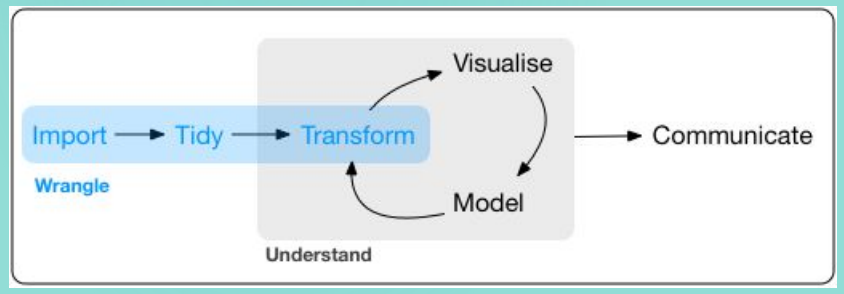


Image source: Hadley Wickham and Garrett Grolemund

Slides and links at https://dsc-wav.github.io/ma-ds-pathways

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.

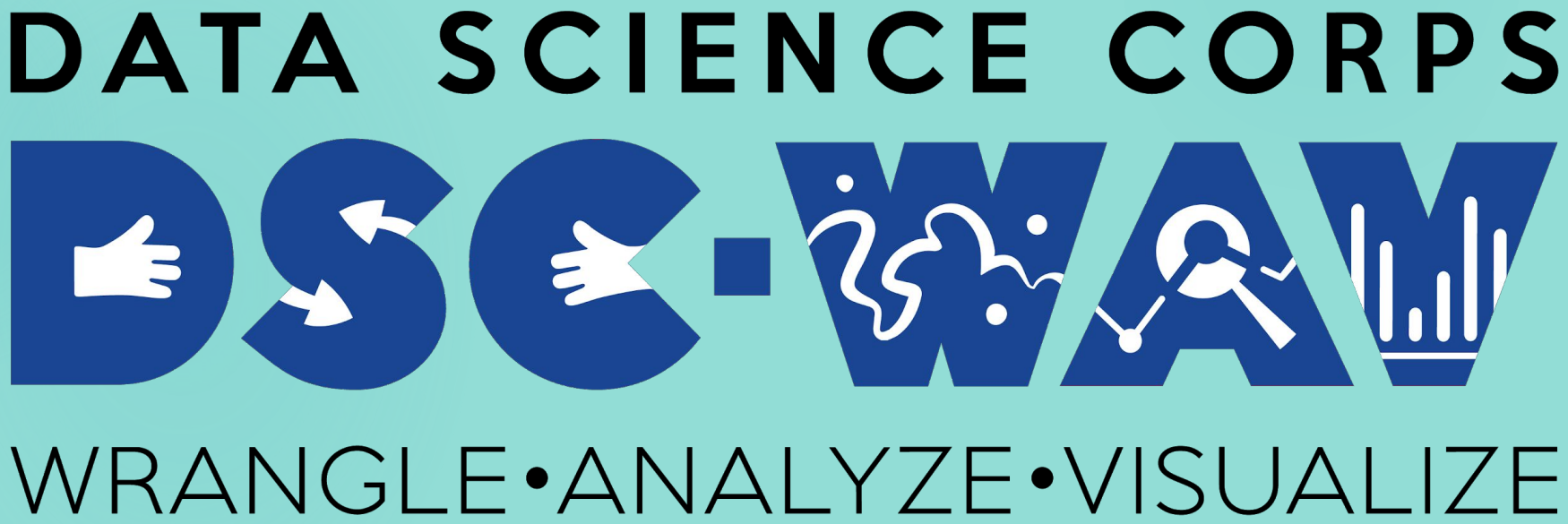
Next steps for programs

- Certificate programs
- Associates to workforce
- Associates to transfer
- Work to develop flexible pathways for students

2018 "Two Year College Data Science Summit" report (https://www.amstat.org/ASA/Education/Two-Year-College-Data-Science-Summit.aspx)

DSC-WAV (Wrangle-Analyze-Visualize)

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

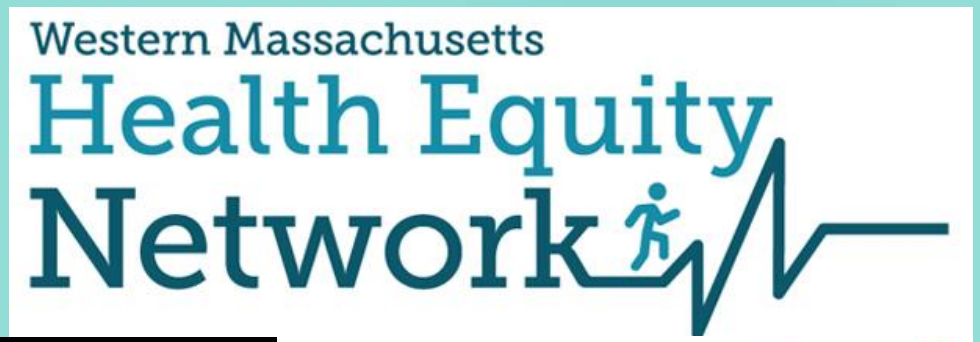


DSC-WAV (Wrangle-Analyze-Visualize)

- 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)

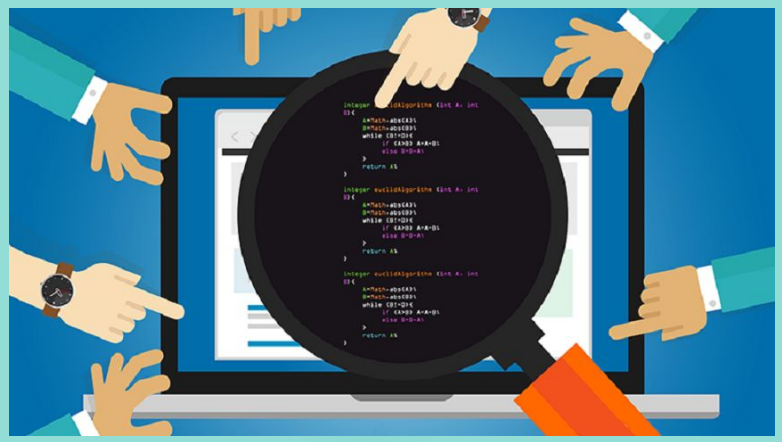
- Goal 1: create opportunities for undergraduate students to work on Data Science for Social Good projects for community organizations



Agile and scrum for undergraduates

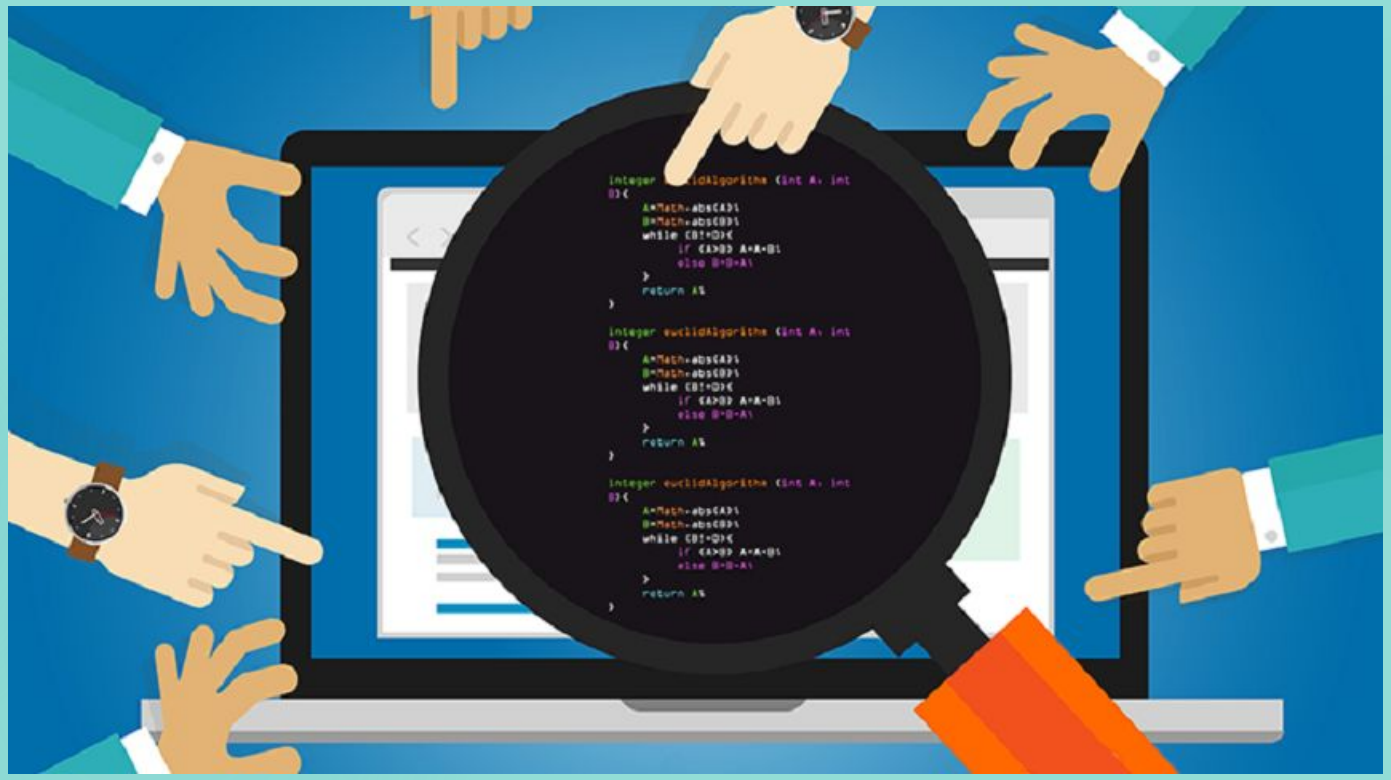
- Horton et al (2021, HDSR)
  https://hdsr.mitpress.mit.edu/pub/nvflcexe/release/1

"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



Source: techgig.com

Agile and scrum for undergraduates

The work is organized into a series of short sprints to break up large tasks.

- Subtasks are organized into a backlog to identify priorities for that stage of the analysis.
- The team and stakeholders (faculty and community organization liaison) meet regularly to share results and make adjustments in advance of the next sprint.
- 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.

Agile and scrum for undergraduates

- Code review, implemented using GitHub pull requests, is included as a regular part of the process.

Source: smartbear.com

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

Agile and scrum for undergraduates

Goal of code review (simpleprogrammer.com)

1. An evaluation method used to identify code errors
2. Should reveal and remove bugs/errors
3. Should help improve code and documentation quality
4. Should (ideally) build developer/analyst skills and self-confidence

Two heads are better than one!

But still hard for undergraduates

9 Code Review Best Practices

Source:kinsta.com

1. Know what to look for in a code review
2. Build and test (before review)
3. Don't review for longer than 15 minutes
4. Check no more than 400 lines at a time (smaller PR are better?)
5. Give feedback that helps (not hurts)
6. Communicate goals and expectations
7. Include everyone in the code review process
8. Foster a positive culture
9. Automate to save time

https://www.perforce.com/blog/qac/9-best-practices-for-code-review

DSC-WAV (Wrangle-Analyze-Visualize)

- Goal 2: foster curricular innovations and connections between two and four year colleges to teach data science
  - work to develop intro data science courses and programs at the two year colleges
  - provide faculty development opportunities
  - facilitate transfer and articulation for students at 2Y colleges
  - team teaching
  - "Junior Fellow" program

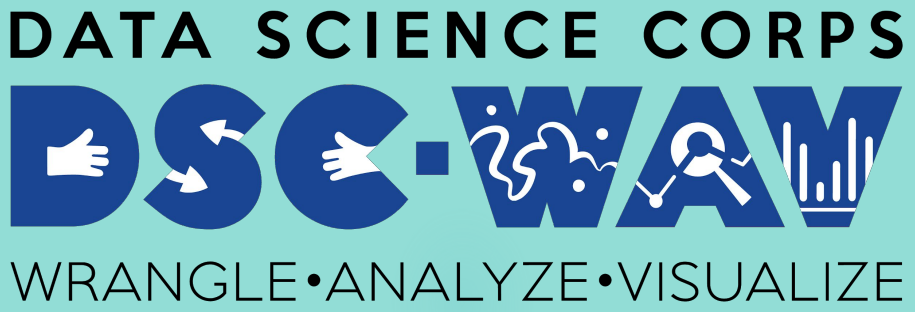
DATA SCIENCE CORPS

DSC-WAV

WRANGLE•ANALYZE•VISUALIZE

DSC-WAV Lessons Learned

- Many challenges to helping 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



Big picture

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

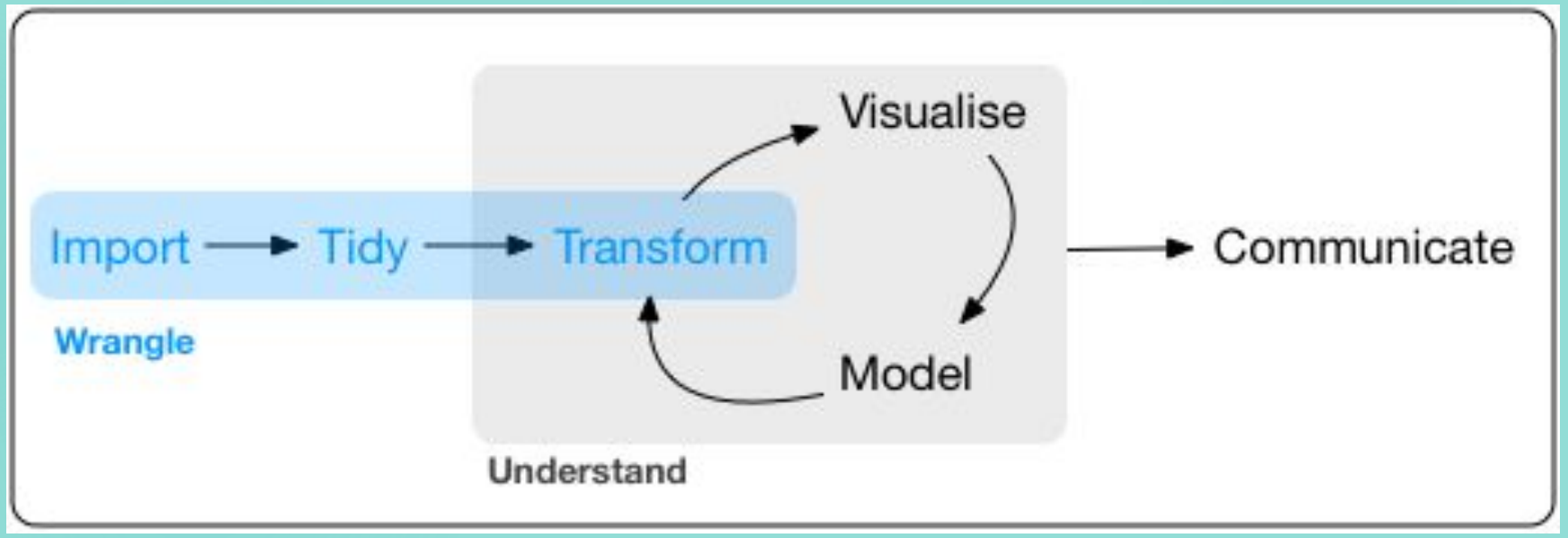


Image source: Hadley Wickham and Garrett Grolemund

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

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.

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.

Massachusetts Data Science Pathways

- Newly founded organization
- Goal: to foster connections between Massachusetts educators and other stakeholders focused on data science pathways from high school through two- and four- colleges and universities
- Plans: resource sharing and occasional convenings
- Sign up for our low-volume mailing list: send mail to : ma-ds-pathways-join@cs.umass.edu
- Minimal website: https://dsc-wav.github.io/ma-ds-pathways

Data science for undergraduates: Building data science curricula and connections between two- and four- year institutions

Nicholas J. Horton, Amherst College

June 16, 2021, nhorton@amherst.edu

Image source: heylagostechie

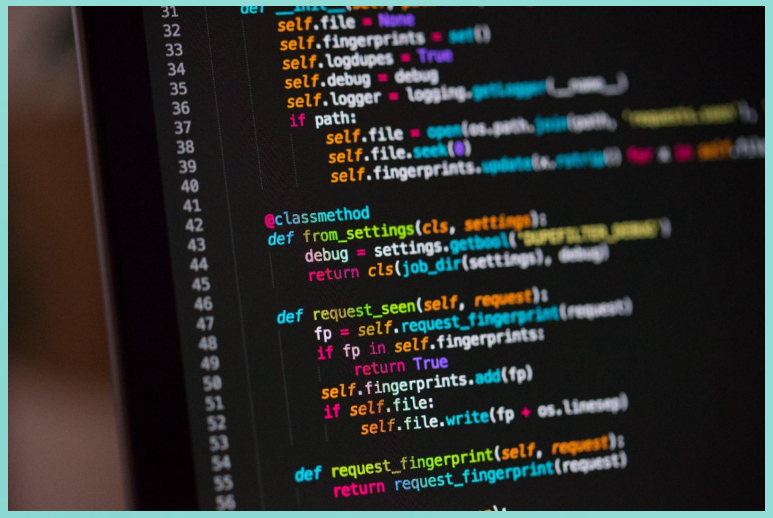


Image source: Wikicommons

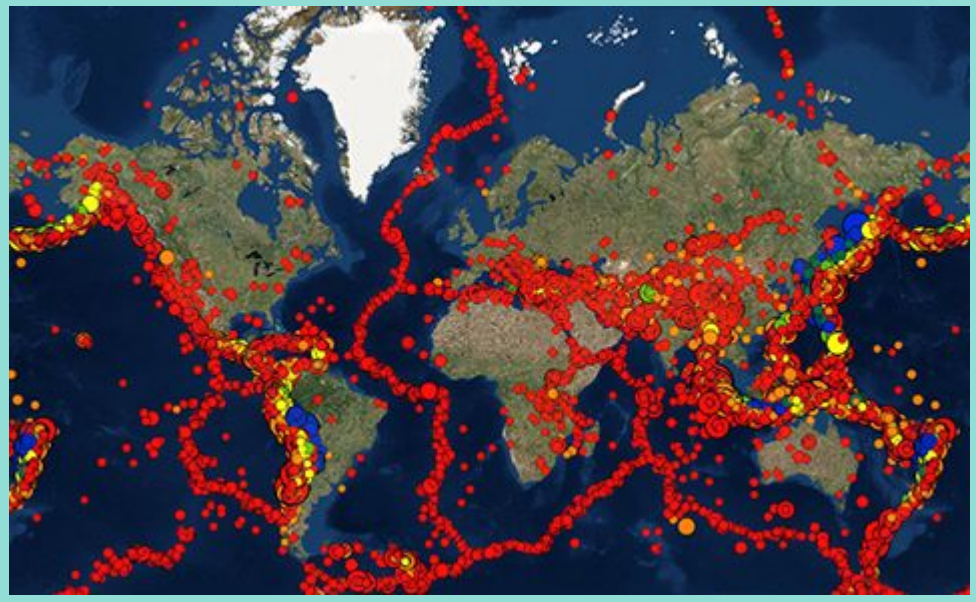


Image source: Concord Consortium

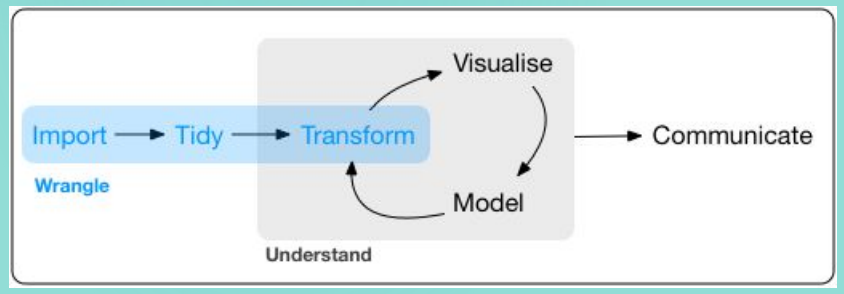


Image source: Hadley Wickham and Garrett Grolemund

Slides and links at https://dsc-wav.github.io/ma-ds-pathways