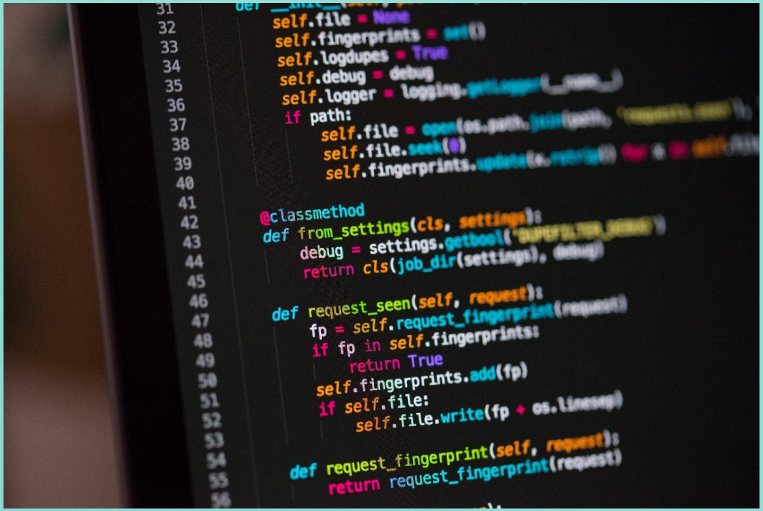
Project-based learning and lessons learned from the DSC-WAV project

Nicholas J. Horton, Amherst College

June 29, 2023, DSLAC nhorton@amherst.edu

Image source: heylagostechie



```
        self.file = None
        self.fingerprints = set()
        self.logdupes = True
        self.debug = debug
        self.logger = logging.getLogger(__name__)
        if path:
            self.file = open(os.path.join(path, 'requests.seen'), 'a+')
            self.file.seek(0)
            self.fingerprints.update(x.rstrip() for x in self.file)

    @classmethod
    def from_settings(cls, settings):
        debug = settings.getbool('DUPEFILTER_DEBUG')
        return cls(job_dir(settings), debug)

    def request_seen(self, request):
        fp = self.request_fingerprint(request)
        if fp in self.fingerprints:
            return True
        self.fingerprints.add(fp)
        if self.file:
            self.file.write(fp + os.linesep)

    def request_fingerprint(self, request):
        return request_fingerprint(request)
```

Image source: Wikicommons

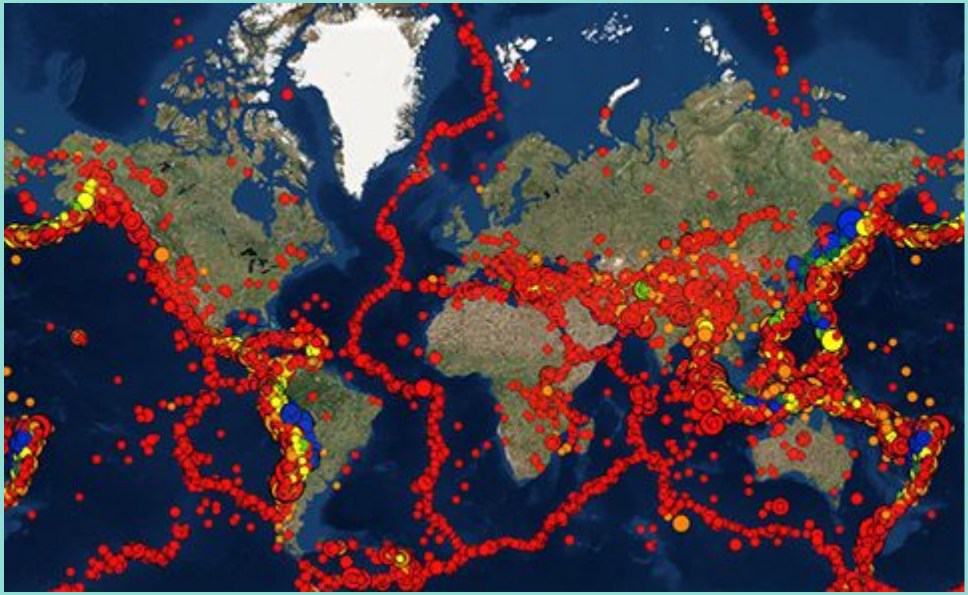


Image source: Concord Consortium

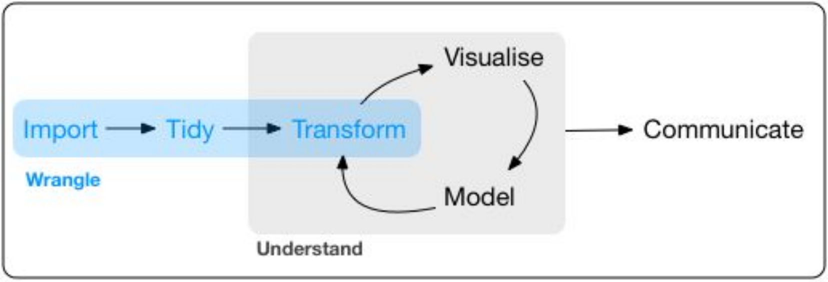


Image source: Hadley Wickham and Garrett Grolemund

thanks to NSF #1923388

Acknowledgements

- Thanks for many key ideas derived from my collaborators: Valerie Barr, Ben Baumer, Matt Beckman, Mine Çetinkaya-Rundel, Jie Chao, Bill Finzer, Jo Hardin, Danny Kaplan, Chelsey Legacy, Randall Pruim, Maria Tackett, Michelle Wilkerson, and Andy Zieffler
- Formal thanks for support from HDR DSC NSF award #1923388

DATA SCIENCE FOR UNDERGRADUATES

Opportunities and Options

consensus report published in 2018
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.

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.

Sidenote: Tracking the growth of K12 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

Problem-solving cycle

- What are we hoping that students will learn?
- How can we help them scaffold their learning?
- Opportunities to build data acumen using project-based learning

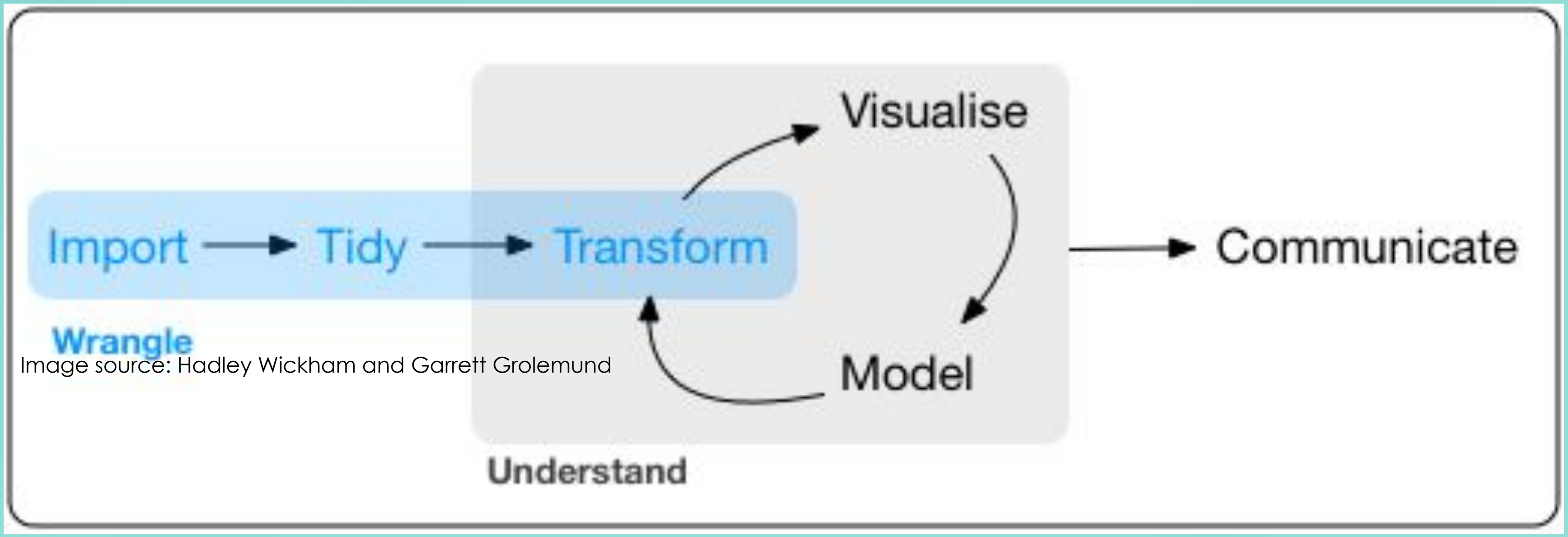
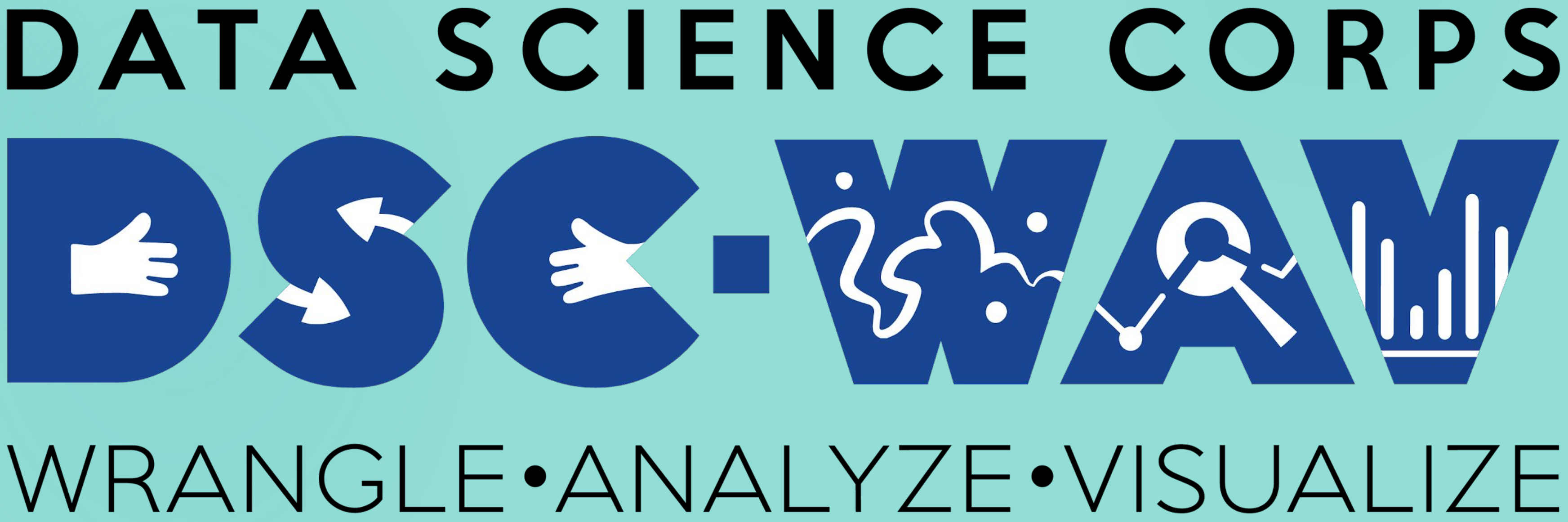


Image source: Hadley Wickham and Garrett Grolemund

DSC-WAV (Wrangle-Analyze-Visualize)

- NSF funded effort from the Harnessing the Data Revolution (HDR) Data Science Corps (DSC) initiative: 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

Mount Holyoke College

UNIVERSITY OF MASSACHUSETTS AMHERST 1863

GREENFIELD CC

HOLYOKE COMMUNITY COLLEGE Foundation, Inc.

STCC

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

and now Bard College
(thanks Valerie for being such a great collaborator!)

DSC-WAV (Wrangle-Analyze-Visualize)

- https://dsc-wav.github.io/www
- Goal 1: create opportunities for undergraduate students working on Data Science for Social Good projects

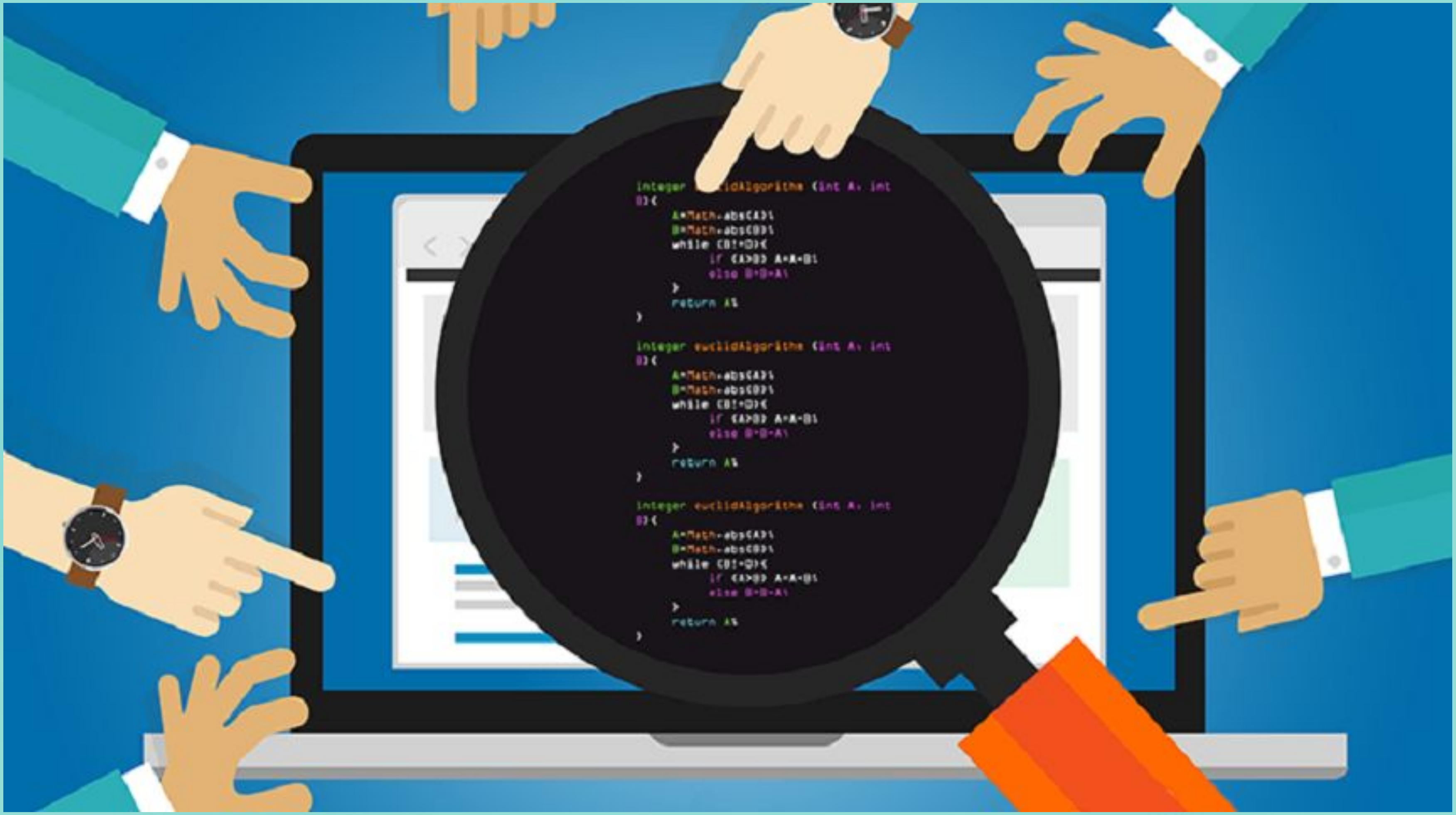
ValleyBike

Western Massachusetts Health Equity Network

girls inc. of the Valley

The Nature Conservancy
Protecting nature. Preserving life.

Agile and scrum for undergraduates



```
integer    lidAlgorithm (int A, int
B){
      A=Math.abs(A);
      B=Math.abs(B);
      while (B!=0){
            if (A>B) A=A-B;
            else B=B-A;
      }
      return A;
}

integer euclidAlgorithm (int A, int
B){
      A=Math.abs(A);
      B=Math.abs(B);
      while (B!=0){
            if (A>B) A=A-B;
            else B=B-A;
      }
      return A;
}

integer euclidAlgorithm (int A, int
B){
      A=Math.abs(A);
      B=Math.abs(B);
      while (B!=0){
            if (A>B) A=A-B;
            else B=B-A;
      }
      return A;
}
```

Source: techgig.com

Agile and scrum for undergraduates

- 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."

Agile and scrum for undergraduates

Work on a project 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 (**standups**) 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.

Agile and scrum for undergraduates

Work on a project is organized into a series of short **sprints** to break up large tasks.

- **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
- How to move into the curriculum?

DATA SCIENCE CORPS

DSC-WAV

WRANGLE•ANALYZE•VISUALIZE

STAT210 (Mining the History of Holyoke): spring 2023 project-based community engagement course

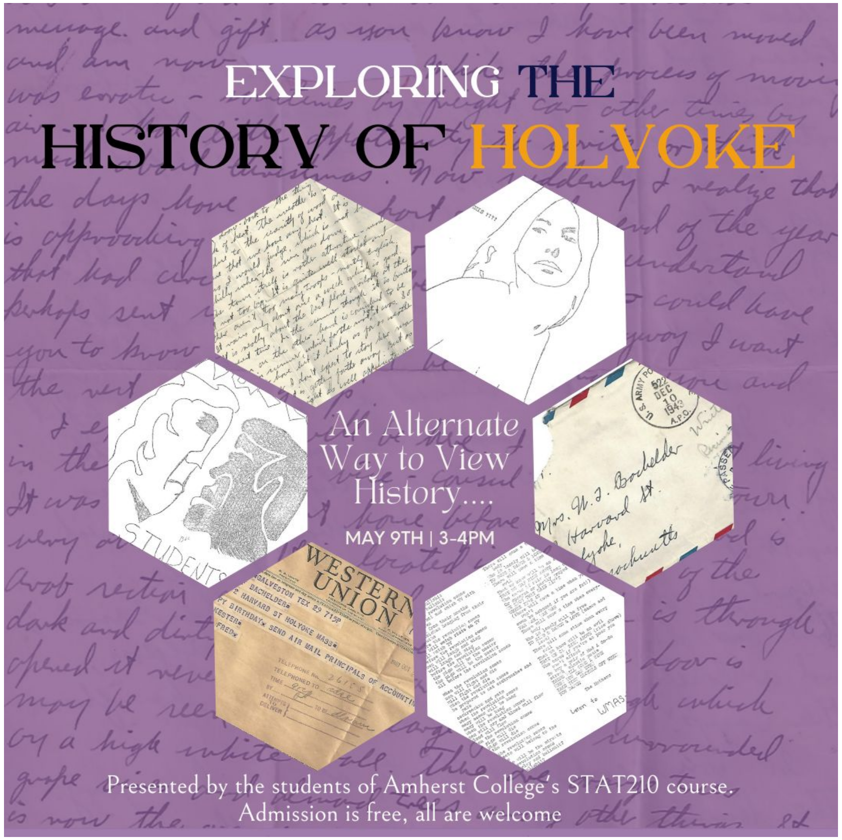
City of
Holyoke
MASSACHUSETTS

Holyoke Public Library
~A LIBRARY IN A PARK~

WISTARIAHURST

Mining History Holyoke

Listed in: Mathematics and Statistics, as STAT-210



EXPLORING THE

HISTORY OF HOLYOKE

An Alternate Way to View History....

MAY 9TH | 3-4PM

Presented by the students of Amherst College's STAT210 course.
Admission is free, all are welcome

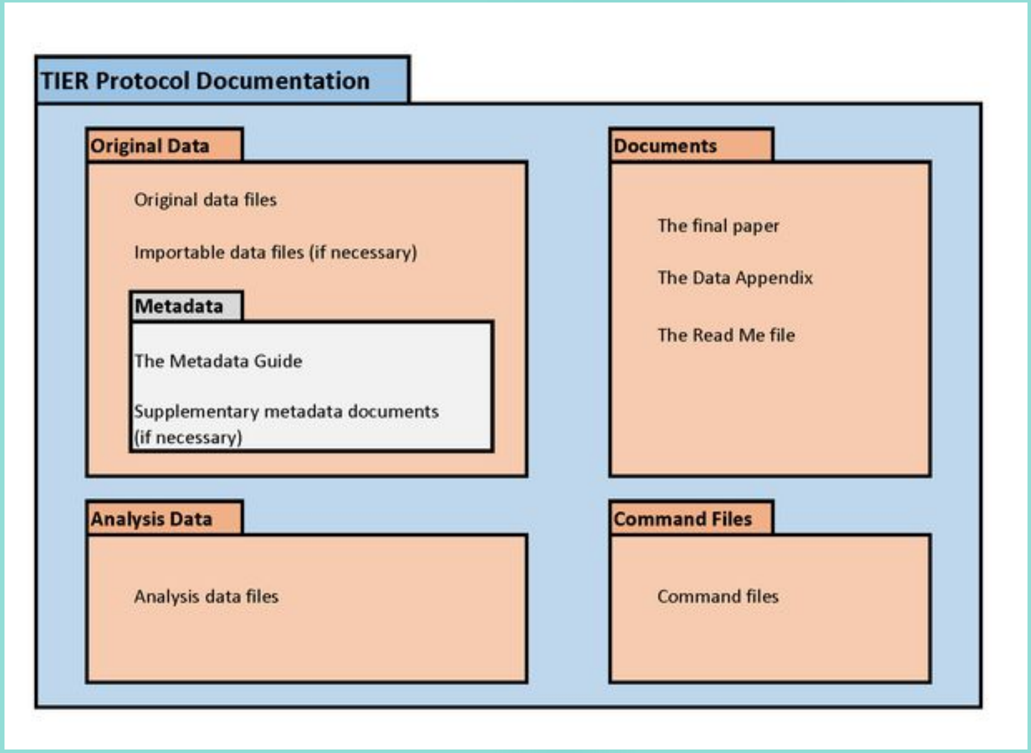
This course will focus on the use of text analytics to explore the rich history of Holyoke, MA. Holyoke has been a site of rapid industrialization, multiple waves of immigration and migration, urban development, rapid changes in its workforce, and ongoing creativity, activism, and innovation. Students will develop the skills to mine textual data from archives at the Wistariahurst Museum, the Holyoke Public Library, Holyoke Community College, and other repositories to address important questions regarding the development and history of this planned community. Topics include sentiment analysis, regular expressions, document-term-matrices, named entity recognition, and Latent Dirichlet analysis. Requisite:

Key questions

- What capacities do we want students to develop?
- What are the skills that we need them to master to fluently utilize these methods?
- What are best practices for teaching these methods?
- When and where can these methods be incorporated into the K-12 and college curricula?
- How do we assess how effectively students are using these methods?

Foundational work: TIER protocol 3.0

- https://www.projecttier.org
- At first focused on long-term research projects
- Now working to build a developmental progression across the undergraduate curriculum (see the soup to nuts) exercises



TIER Protocol Documentation

Original Data

Original data files

Importable data files (if necessary)

Metadata

The Metadata Guide

Supplementary metadata documents (if necessary)

Documents

The final paper

The Data Appendix

The Read Me file

Analysis Data

Analysis data files

Command Files

Command files

Beckman et al (JSDSE, 2021)

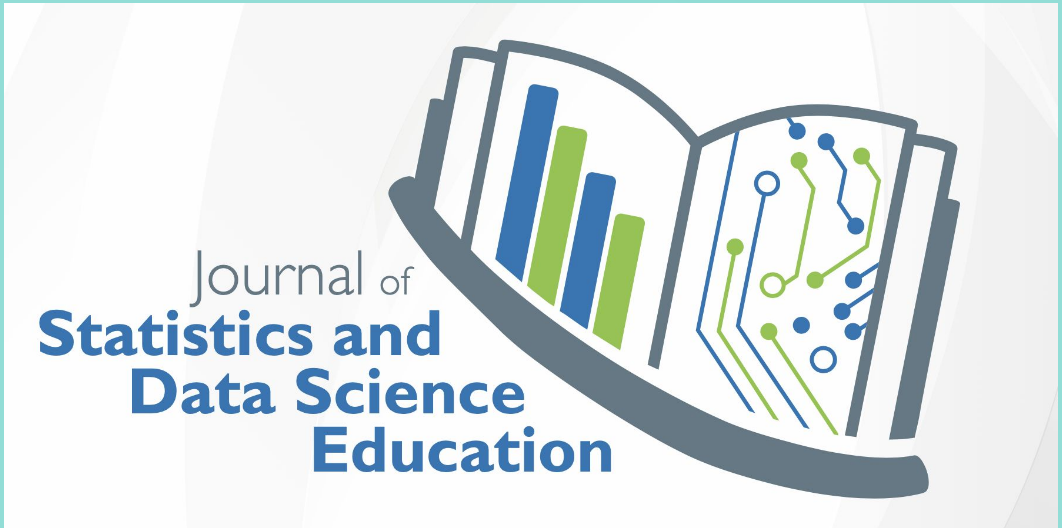
- Implementing Version Control With Git and GitHub as a Learning Objective in Statistics and Data Science Courses
- https://www.tandfonline.com/doi/full/10.1080/10691898.2020.1848485
- Student experiences from Amherst College, Brown University, Duke University, and the University of Edinburgh
- "Teaching reproducible analysis in the statistics curriculum helps make students aware of the issue of scientific reproducibility and also equips them with the knowledge and skills to conduct their future data analyses reproducibly, whether as part of an academic research project or in industry."

Beckman et al (JSDSE, 2021)

- Implementing Version Control With Git and GitHub as a Learning Objective in Statistics and Data Science Courses
- https://www.tandfonline.com/doi/full/10.1080/10691898.2020.1848485
- “Use of version control helps reinforce the notion that statistical analysis typically requires multiple revisions, as students can review their commits to see all the updates they’ve made to their work. A desirable side effect is that, because students are periodically “submitting” their assignment as they work on it, there is less pressure of the final deadline where everything must be submitted in its final form.”

Shameless plug #1

- The *Journal of Statistics and Data Science Education* is published by Taylor & Francis on behalf of the American Statistical Association.
- Open access with no author fees
- https://www.tandfonline.com/loi/ujse21



Shameless plug #2

- The *Harvard Data Science Review* is published by MIT Press
- Open access with no author fees
- Submissions are via two page proposals
- https://hdsr.mitpress.mit.edu

HDSR

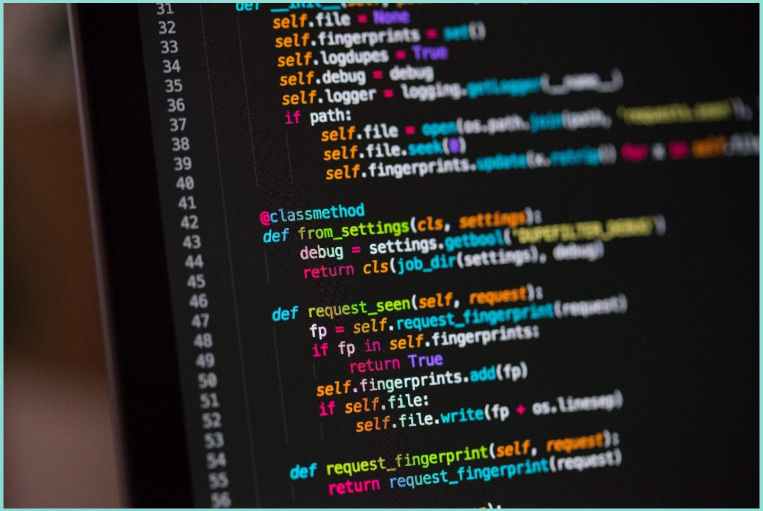
HARVARD DATA SCIENCE REVIEW

Project-based learning and lessons learned from the DSC-WAV project

Nicholas J. Horton, Amherst College

June 29, 2023, DSLAC nhorton@amherst.edu

Image source: heylagostechie



```
        self.file = None
        self.fingerprints = set()
        self.logdupes = True
        self.debug = debug
        self.logger = logging.getLogger(__name__)
        if path:
            self.file = open(os.path.join(path, 'requests.seen'), 'a+')
            self.file.seek(0)
            self.fingerprints.update(x.rstrip() for x in self.file)

    @classmethod
    def from_settings(cls, settings):
        debug = settings.getbool('DUPEFILTER_DEBUG')
        return cls(job_dir(settings), debug)

    def request_seen(self, request):
        fp = self.request_fingerprint(request)
        if fp in self.fingerprints:
            return True
        self.fingerprints.add(fp)
        if self.file:
            self.file.write(fp + os.linesep)

    def request_fingerprint(self, request):
        return request_fingerprint(request)
```

Image source: Wikicommons

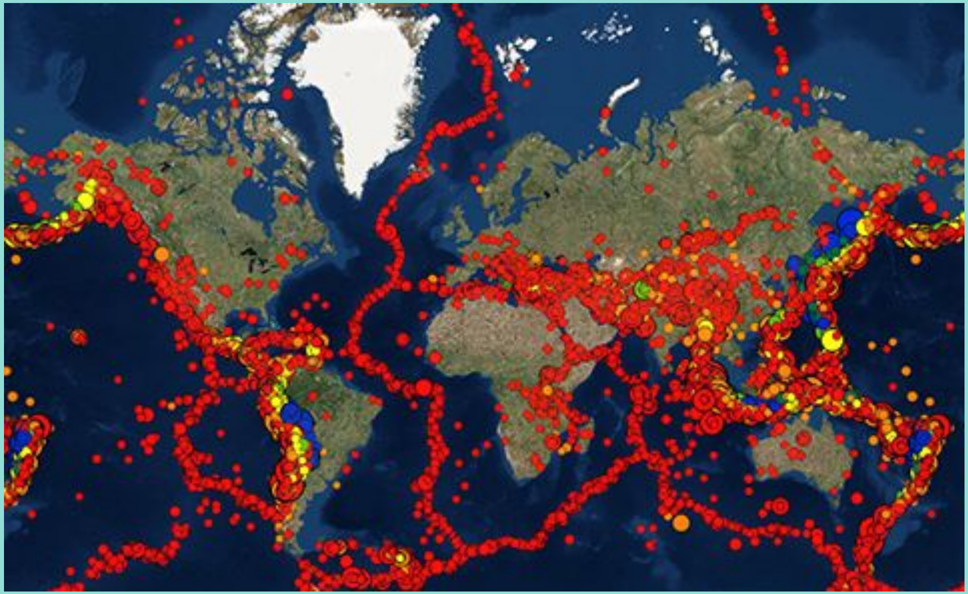
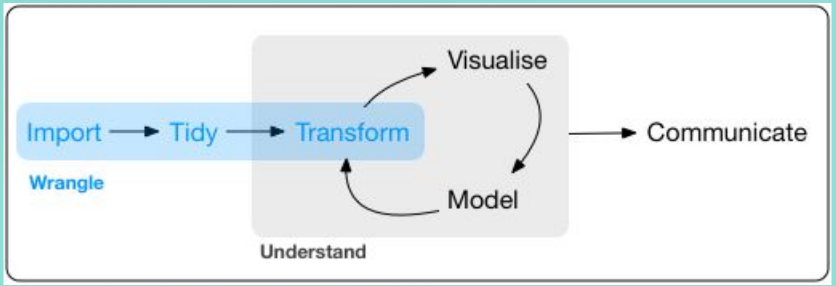


Image source: Concord Consortium



Import → Tidy → Transform

Wrangle

Visualise

Model

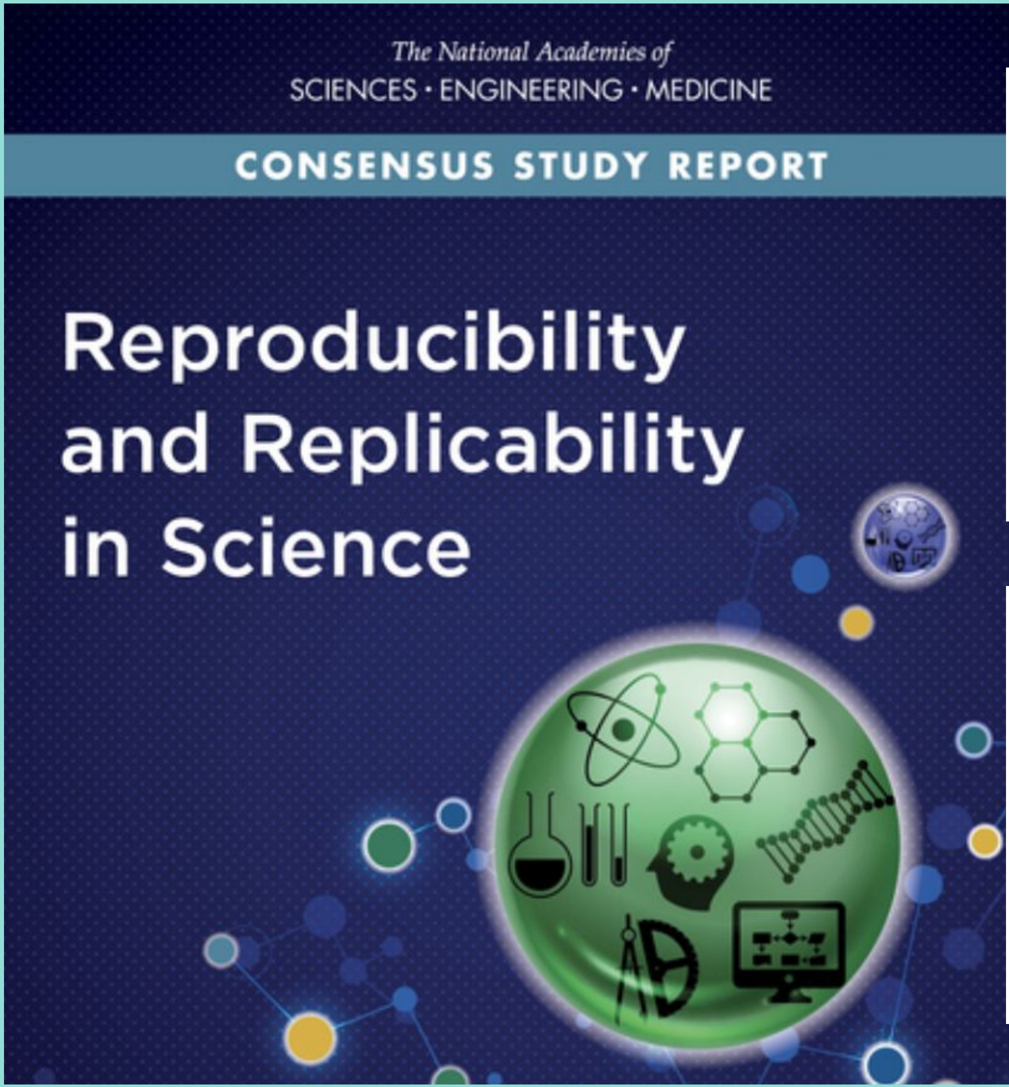
Understand

Communicate

Image source: Hadley Wickham and Garrett Grolemund

thanks to NSF #1923388

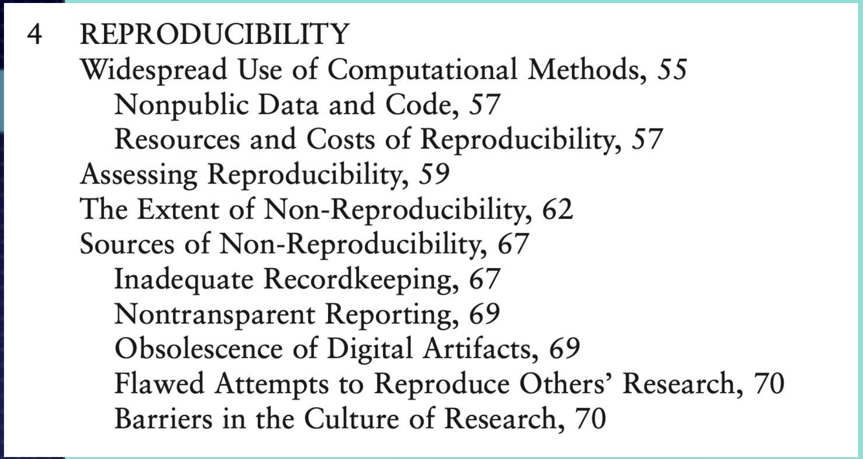
NASEM (2019)

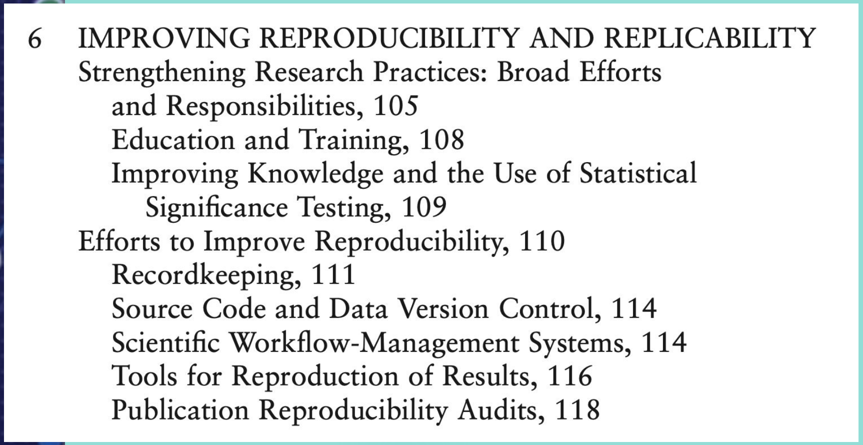


The National Academies of
SCIENCES · ENGINEERING · MEDICINE

CONSENSUS STUDY REPORT

Reproducibility
and Replicability
in Science





NASEM (2019)

RECOMMENDATION 4-1: To help ensure the reproducibility of computational results, researchers should convey **clear, specific, and complete information** about any computational methods and data products that support their published results in order to enable other researchers to repeat the analysis, unless such information is restricted by nonpublic data policies.

That information should include the data, study methods, and computational environment:

NASEM (2019)

1) the input data used in the study either in extension (e.g., a text file or a binary) or in intension (e.g., a script to generate the data), as well as intermediate results and output data for steps that are nondeterministic and cannot be reproduced in principle;

2) a detailed description of the study methods (ideally in executable form) together with its computational steps and associated parameters; and

3) information about the computational environment where the study was originally executed, such as operating system, hardware architecture, and library dependencies.

NASEM (2019)

- RECOMMENDATION 6-6: Many stakeholders have a role to play in improving computational reproducibility, including educational institutions, professional societies, researchers, and funders.
- **Educational institutions should educate and train students and faculty about computational methods and tools to improve the quality of data and code and to produce reproducible research.**
- Professional societies should take responsibility for educating the public and their professional members about the importance and limitations of computational research.

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**.

Idea of "computational essay": more later

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 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:

Idea of "computational essay": more later

- 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).

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 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.

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