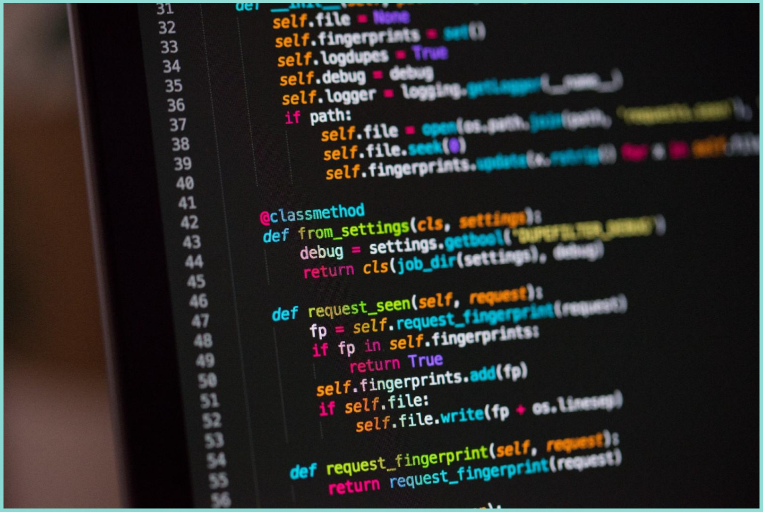
Better Data Tools Foster Better Data Science Education

Nicholas J. Horton, Amherst College

November 28, 2022, 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

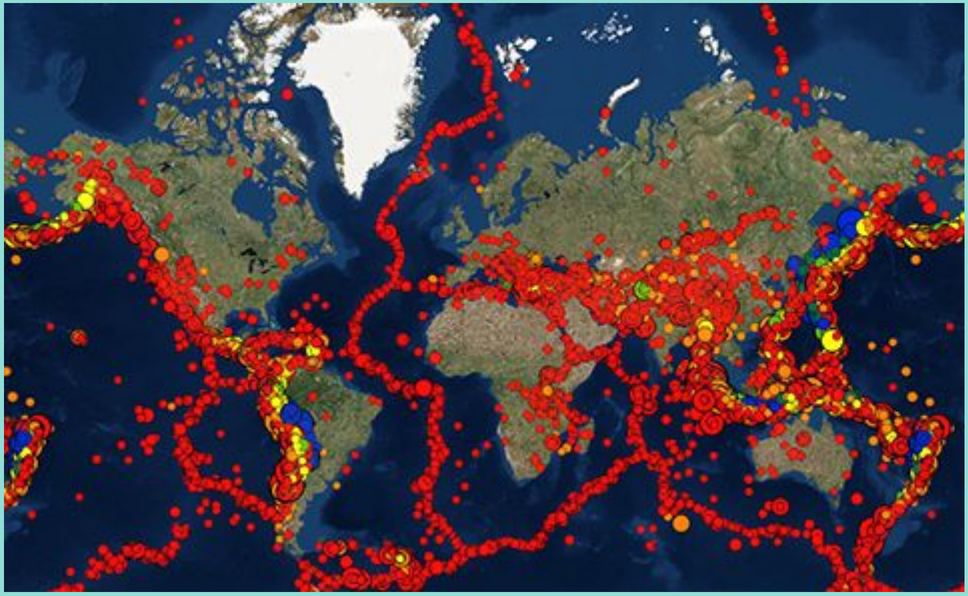


Image source: Concord Consortium

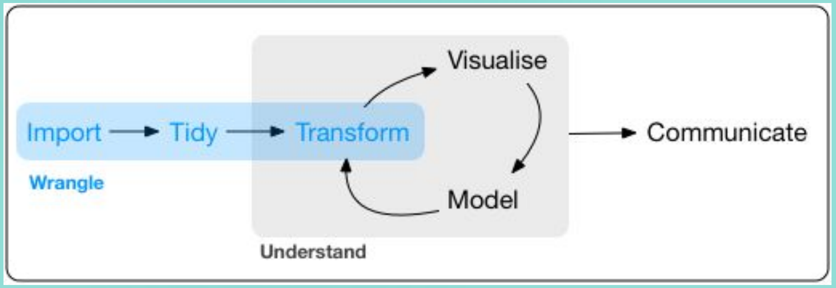


Image source: Hadley Wickham and Garrett Grolemund

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

Acknowledgements

- The K12 Data Tools project is joint work with Danny Pimentel and Michelle Wilkerson (commissioned paper from K-12 Data Science workshop can be found here: https://nicholasjhorton.github.io/K12-Data-Tools)

- The DSC-WAV project is funded by the NSF (#1923388, https://dsc-wav.github.io/www, Ben Baumer PI)

Plan

- What is Data Science and why should we care?
- Insights about data acumen from the NASEM (2018) report
- Growth of K12 data science
- Data tools to support new (and old) learners
- Compare and contrast tools in the context of an example (where are the lobsters?)
- Next steps and closing thoughts

Zoom poll #1

- How would you define Data Science?

Please provide a succinct description (a single sentence) in the chat window but WAIT to share it until I say to do so.

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

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

DATA SCIENCE CORPS

DSC-WAV

WRANGLE•ANALYZE•VISUALIZE

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)

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

Western Massachusetts Health Equity Network

ValleyBike

girls inc.

of the Valley

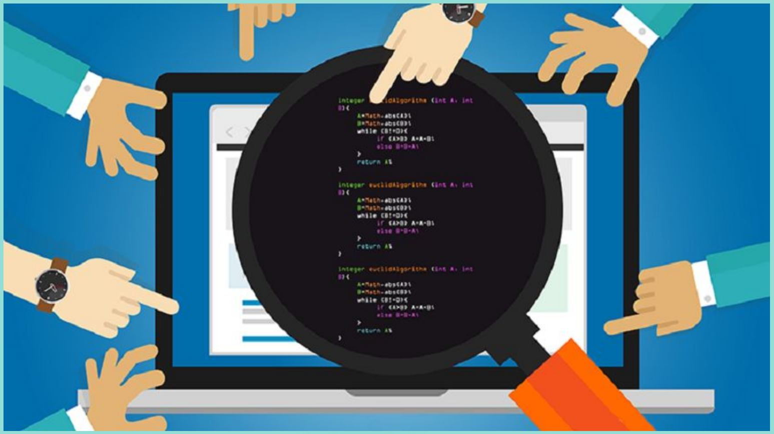
The Nature Conservancy

Protecting nature. Preserving life.

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

Zoom poll #2

- When would it be ideal for students to first experience data science in the classroom?

Please provide your response in the chat window but wait to share it until I say

Growth of K12 Data Science

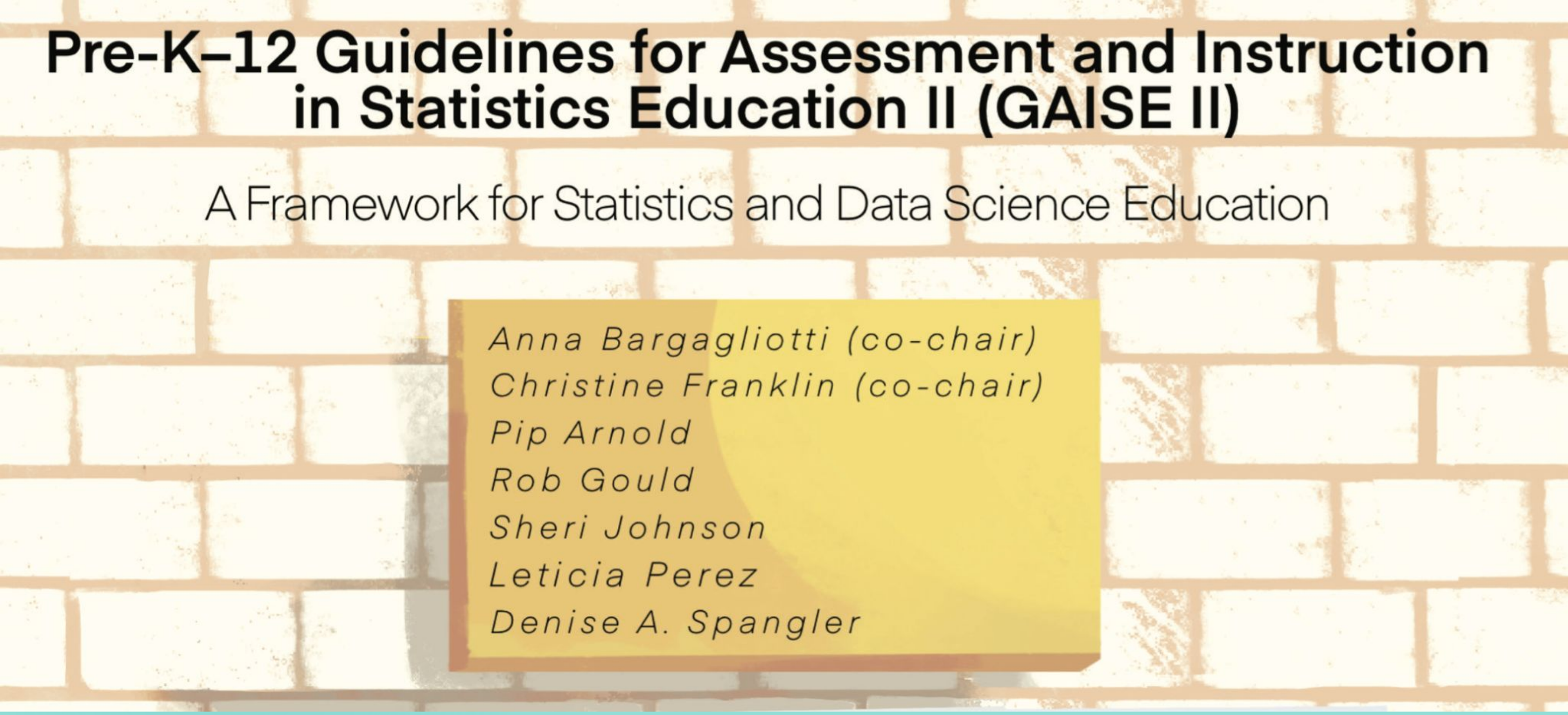
- In a world defined by data, we can't wait to introduce students in K-12 to the opportunities (and challenges) in making sense of it
- Increasing growth of K12 Data Science:
    - NASEM workshop, https://www.nationalacademies.org/our-work/foundations-of-data-science-for-students-in-grades-k-12-a-workshop
    - 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

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 K12 GAISE Guidelines



Pre-K–12 Guidelines for Assessment and Instruction in Statistics Education II (GAISE II)

A Framework for Statistics and Data Science Education

Anna Bargagliotti (co-chair)
Christine Franklin (co-chair)
Pip Arnold
Rob Gould
Sheri Johnson
Leticia Perez
Denise A. Spangler

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

Problem-solving cycle

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

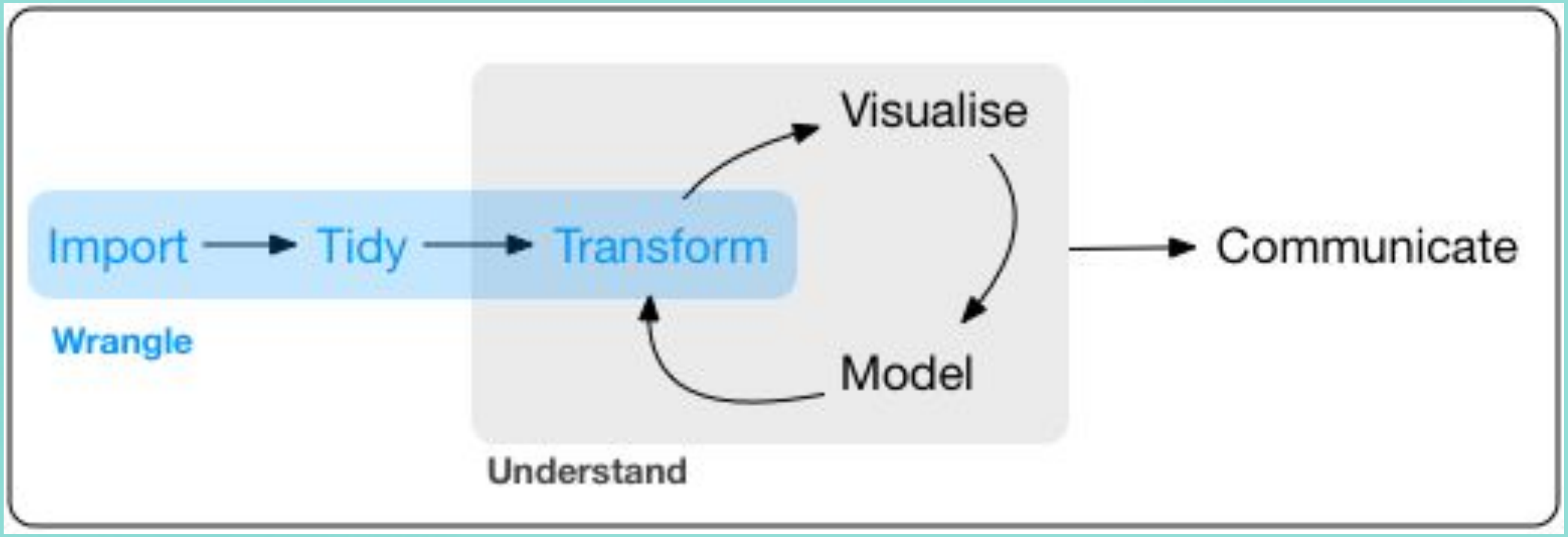
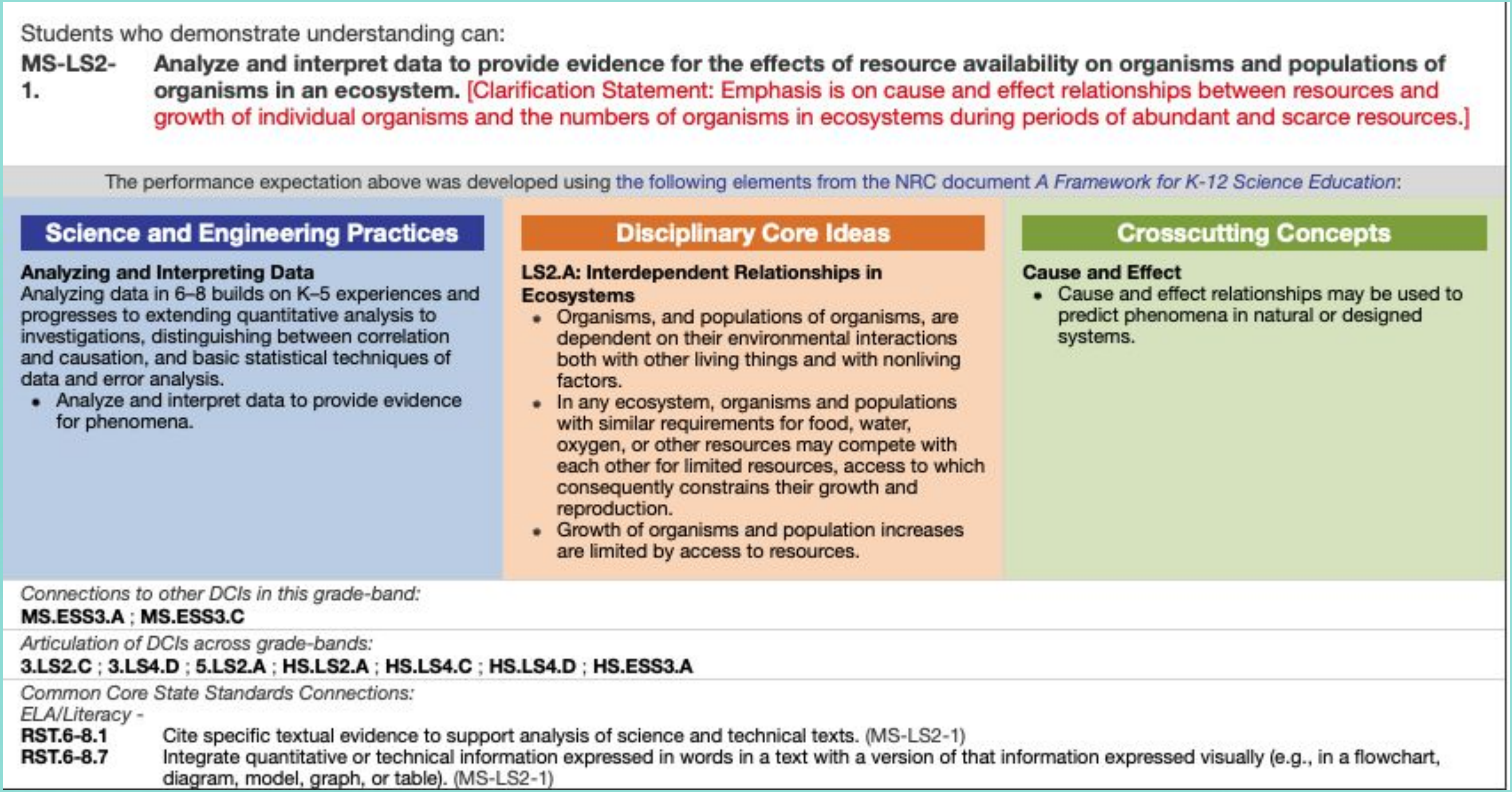


Image source: Hadley Wickham and Garrett Grolemund

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	Disciplinary Core Ideas	Crosscutting Concepts
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.	**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.	**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**

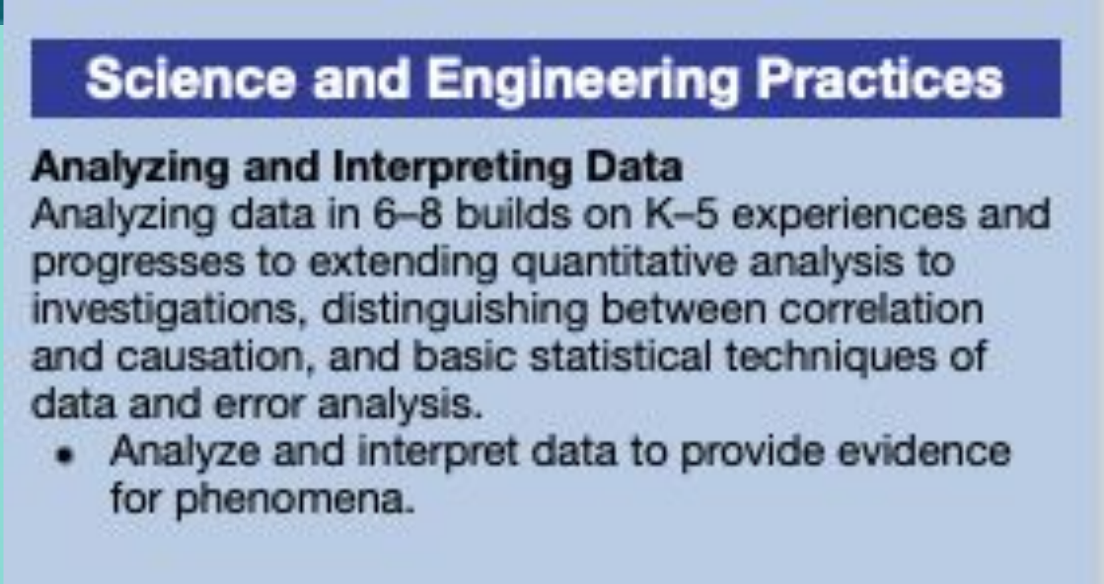
Common Core State Standards Connections:
ELA/Literacy -

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)

Data Tools for Data Science

- Lots of data science now taking place in K12 (with **much** more to come, see for example https://doe.virginia.gov/boe/meetings/2022/04-apr/item-g.pdf)
- Good news: there are many tools for doing and teaching data analysis and data science
- Goal: provide a high-level analysis of tools with a focus on student learning, learning progressions, and instructor professional development

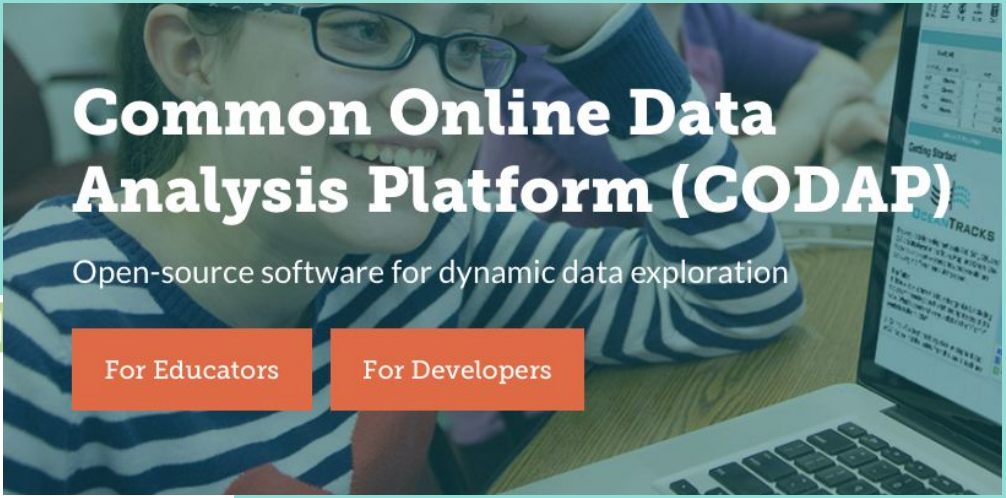
Zoom poll #3

- What tools do you teach/use for data science? Are the tools you teach and the tools you use the same?

Please provide a succinct description (one sentence) in the chat window but wait to share it until I say

Data Tools for K12 Data Science

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

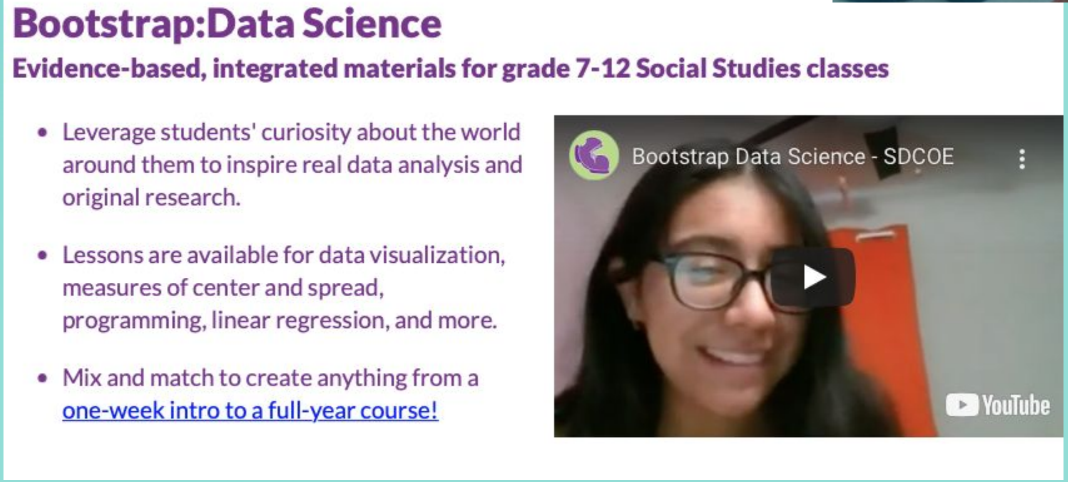


Common Online Data Analysis Platform (CODAP)

Open-source software for dynamic data exploration

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!

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
- We adapted the framework of McNamara to account for considerations specific to K12 education

Framework (based on McNamara, 2019)

- **Accessibility**: Includes cost, simplicity of cloud-based tools, disability access, multilingual support
- **Ease of entry**: Clarity about how the tool works; includes consideration of students' conceptions of data and developmental appropriateness
- **Data as a first-order object**: Data as primary interest: hierarchical vs. tabular formats, viewing data; key to building "students conception of data"

Framework (based on McNamara, 2019)

- **Data analysis cycle and reproducible workflows**: Iterative cycle of posing questions, exploring data, visualizing results, modeling, model assessment, and communicating results; reproducing data wrangling, analyses, and explorations
- **Interactivity**: Support for direct interaction with data, e.g., pinch, click-and-drag, brushing, hovering
- **Flexible plot creation**: Univariate, bivariate, and multivariate displays with ability to augment graphics in a variety of ways

Framework (based on McNamara, 2019)

- **Inferential analysis**: Reasoning with samples and inferring beyond data; support for simulations and resampling; offering probabilistic or uncertain expressions of data
- **Non-standard data**: Working with multiple forms of data such as spatial data, network data, etc.
- **Extensibility**: included in prior frameworks: important for the future but beyond our scope

Genres of Data Tools

- Spreadsheets
  - Google Sheets
  - Excel
- Visual tools
  - CODAP
  - iNZight
  - Tuva
  - Tableau
- Scripting languages
  - Python
  - R
  - Julia

Illustrative example: lobsters

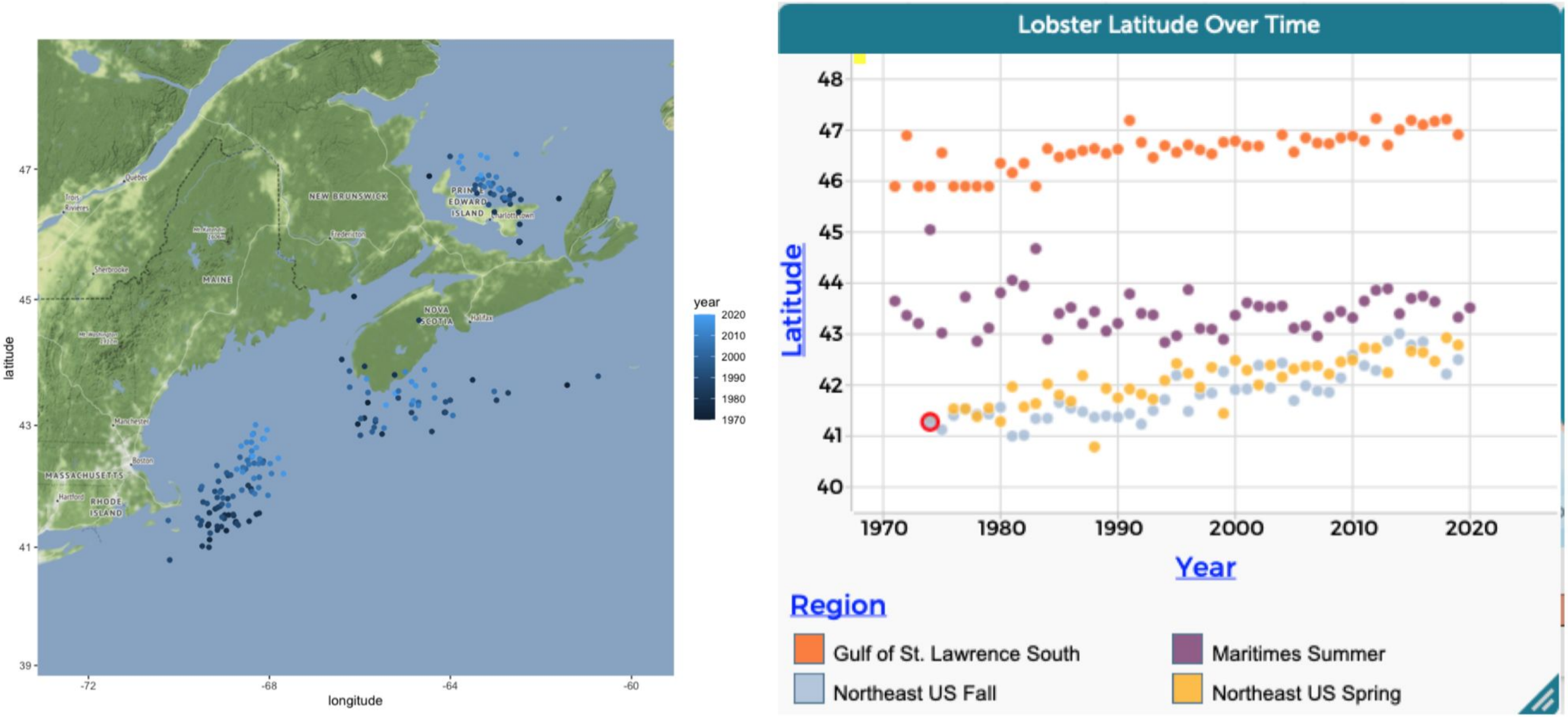


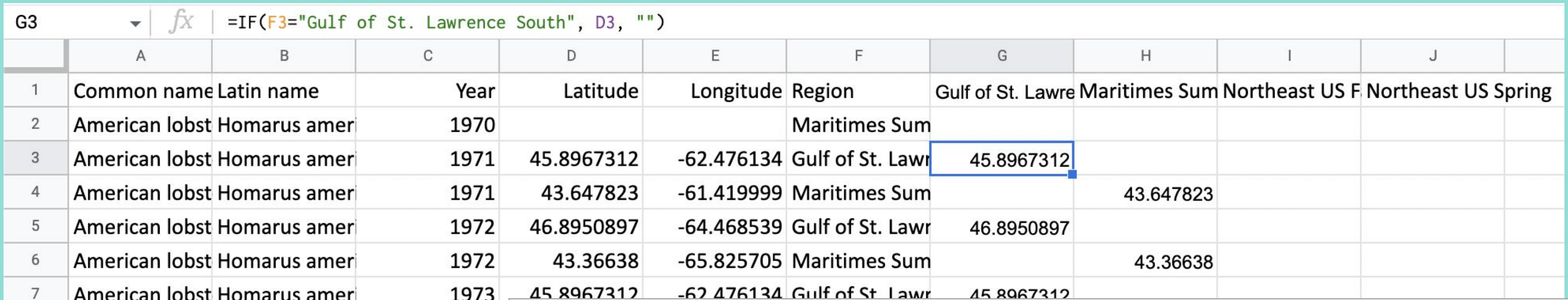
Figure 1: Left: Map of groups of lobster between 1970 and 2020, colored by year. Right: Scatterplot of latitude of lobster populations over time, colored by region.

Homarus americanus (lobsters)

- Data on mean location of lobsters from 1970 through 2020 in various regressions off the northeast of the United States
- There is evidence that the populations have been moving northward over time
- Exploring such movement is often included as an activity in math or science class
- Includes spatial data (maps) and time series

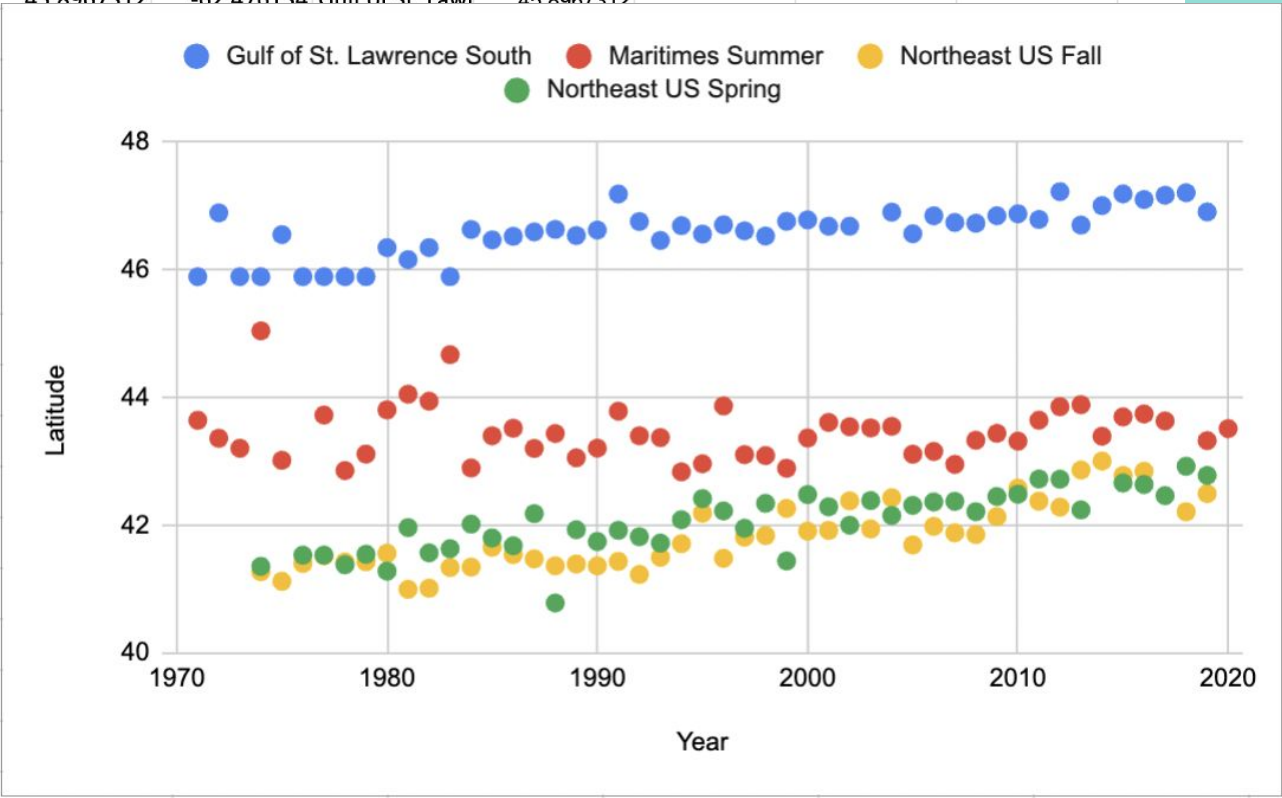
Data, code, illustrated examples, and interactive links available at:
https://nicholasjhorton.github.io/K12-Data-Tools/dsd.html

Lobsters in spreadsheets



G3 | fx | =IF(F3="Gulf of St. Lawrence South", D3, "")

	A	B	C	D	E	F	G	H	I	J
1	Common name	Latin name	Year	Latitude	Longitude	Region	Gulf of St. Lawre	Maritimes Sum	Northeast US F	Northeast US Spring
2	American lobst	Homarus ameri	1970			Maritimes Sum				
3	American lobst	Homarus ameri	1971	45.8967312	-62.476134	Gulf of St. Lawr	45.8967312			
4	American lobst	Homarus ameri	1971	43.647823	-61.419999	Maritimes Sum		43.647823		
5	American lobst	Homarus ameri	1972	46.8950897	-64.468539	Gulf of St. Lawr	46.8950897			
6	American lobst	Homarus ameri	1972	43.36638	-65.825705	Maritimes Sum		43.36638		
7	American lobst	Homarus ameri	1973	45.8967312	-62.476134	Gulf of St. Lawr	45.8967312			
8	American lobst	Homarus ameri	1973							
9	American lobst	Homarus ameri	1974							
10	American lobst	Homarus ameri	1974							
11	American lobst	Homarus ameri	1974							
12	American lobst	Homarus ameri	1974							

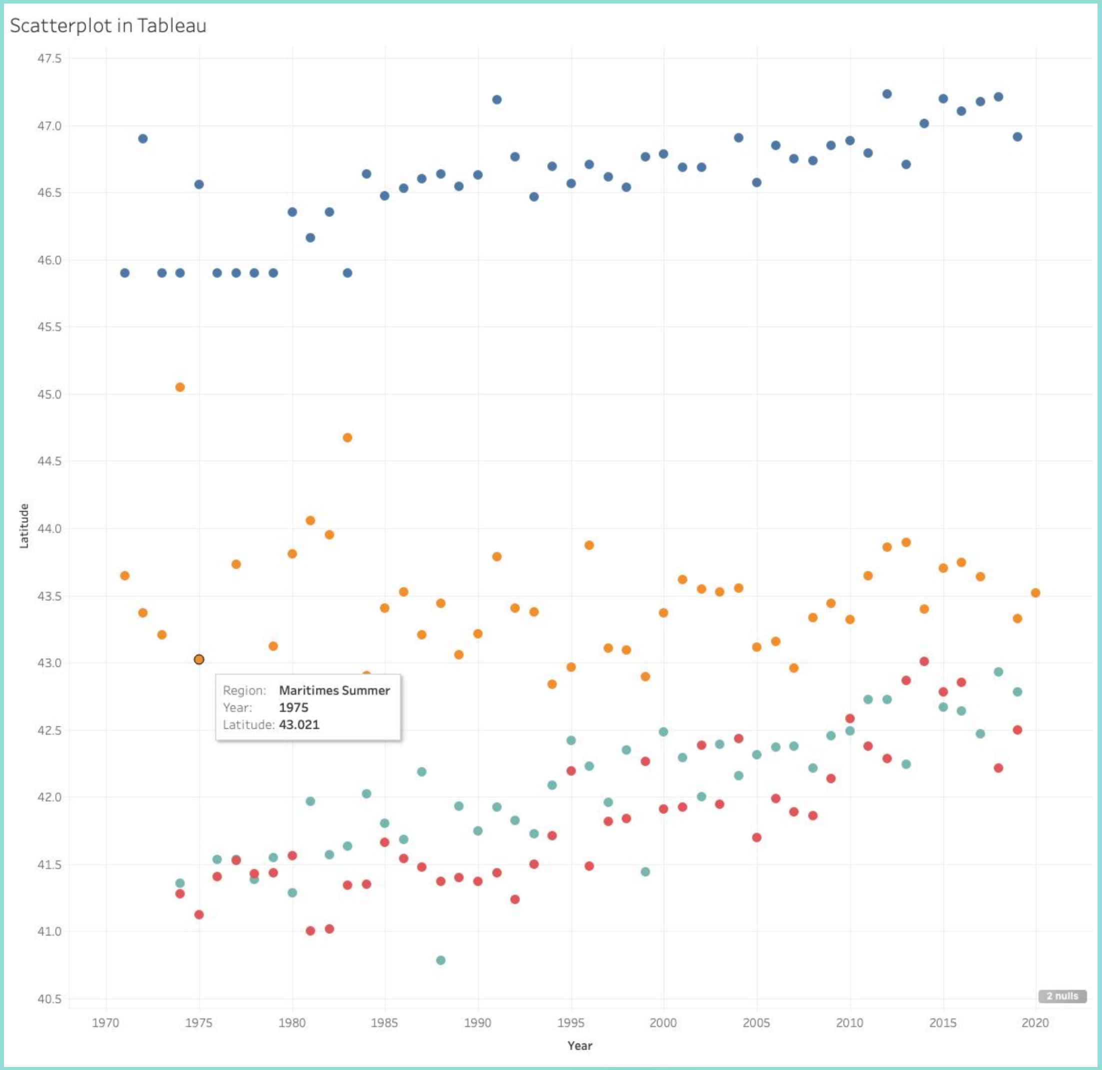


Thoughts about spreadsheets

- Commonly accessible (Excel or Google Sheets)
- Often used for simpler analyses, ease of entry
- Offer rudimentary graphics and tables
- Data at the fore (easy to review and examine individual points)
- Challenging to undertake some "data moves" (Erickson et al, TISE, https://escholarship.org/uc/item/0mg8m7g6)
- Challenging to undertake multivariate visualization
- Limited reproducibility

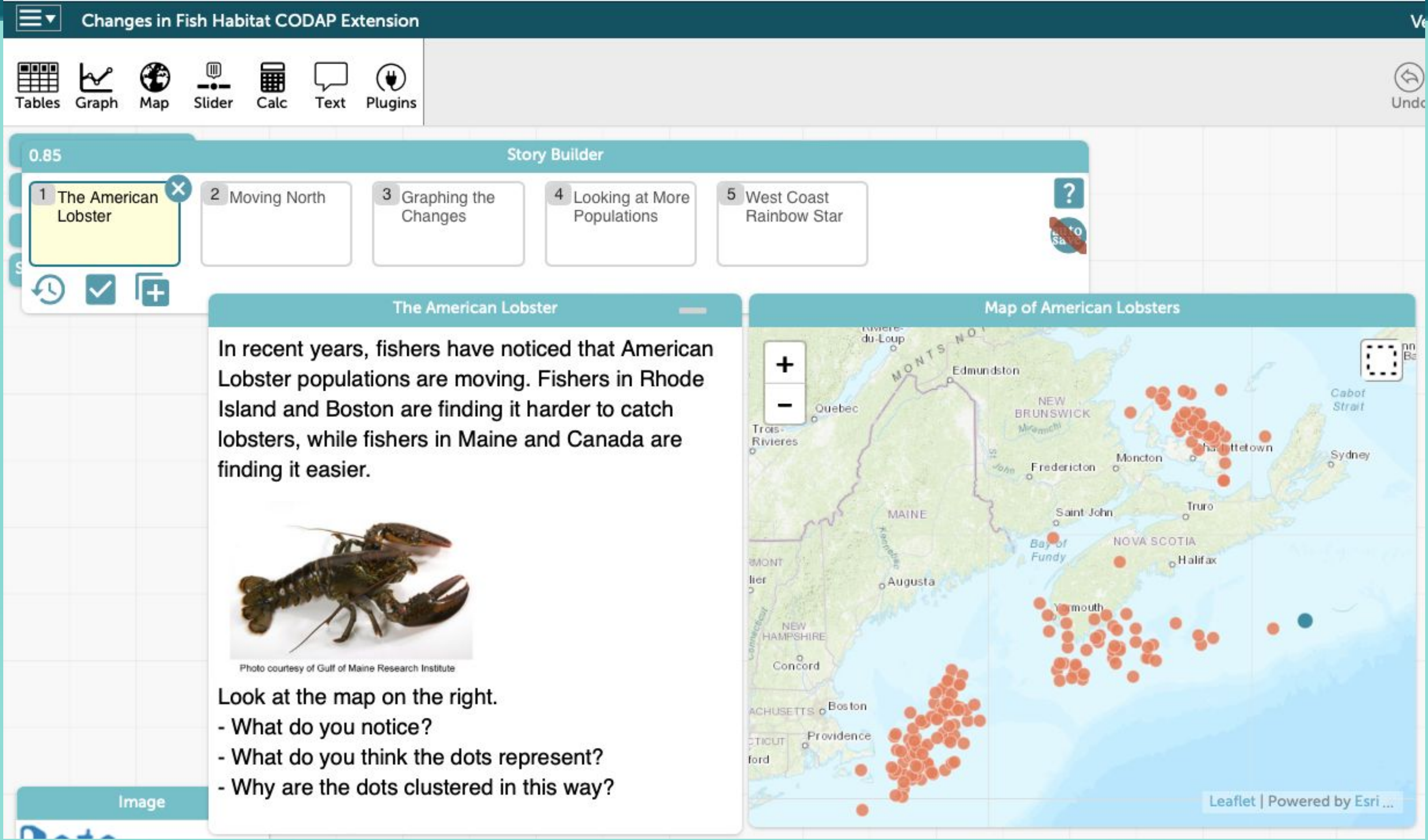
Lobsters in Tableau

Hovering and interactive displays



Region: Maritimes Summer
Year: 1975
Latitude: 43.021

Lobsters in CODAP



Changes in Fish Habitat CODAP Extension

The American Lobster

In recent years, fishers have noticed that American Lobster populations are moving. Fishers in Rhode Island and Boston are finding it harder to catch lobsters, while fishers in Maine and Canada are finding it easier.

Photo courtesy of Gulf of Maine Research Institute

Look at the map on the right.

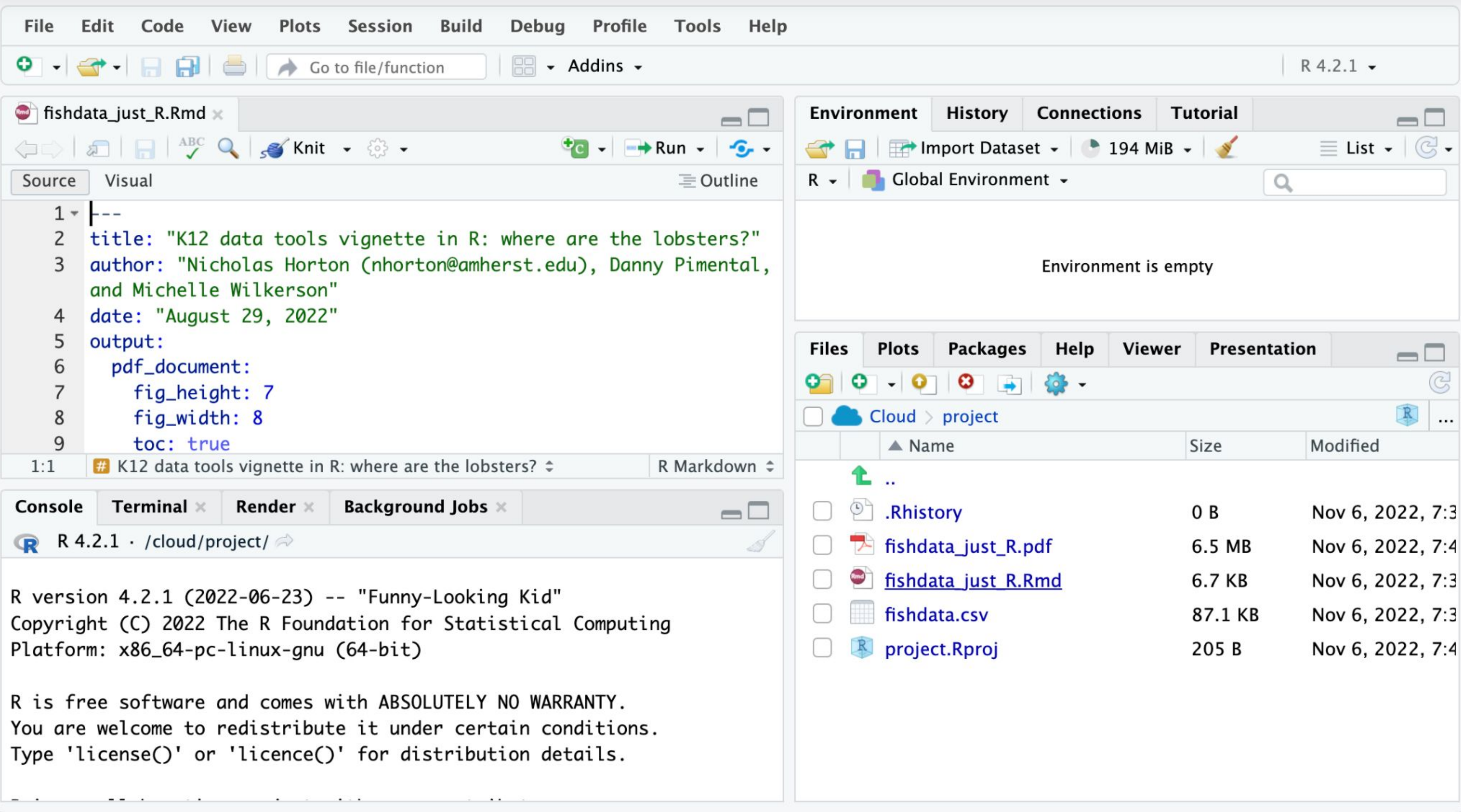
- What do you notice?
- What do you think the dots represent?
- Why are the dots clustered in this way?

Try it at the following link: https://tinyurl.com/dsd-codap

Thoughts about visual tools

- Excellent at accessibility and "data as a first-order, persistent object"
- Facilitate interactive exploration and flexible plot creation
- Limited support for reproducibility (moves made are not recorded)
- Support for inference somewhat limited (under active development in CODAP)

Lobsters in R



```
---
title: "K12 data tools vignette in R: where are the lobsters?"
author: "Nicholas Horton (nhorton@amherst.edu), Danny Pimental, and Michelle Wilkerson"
date: "August 29, 2022"
output:
  pdf_document:
    fig_height: 7
    fig_width: 8
    toc: true
```

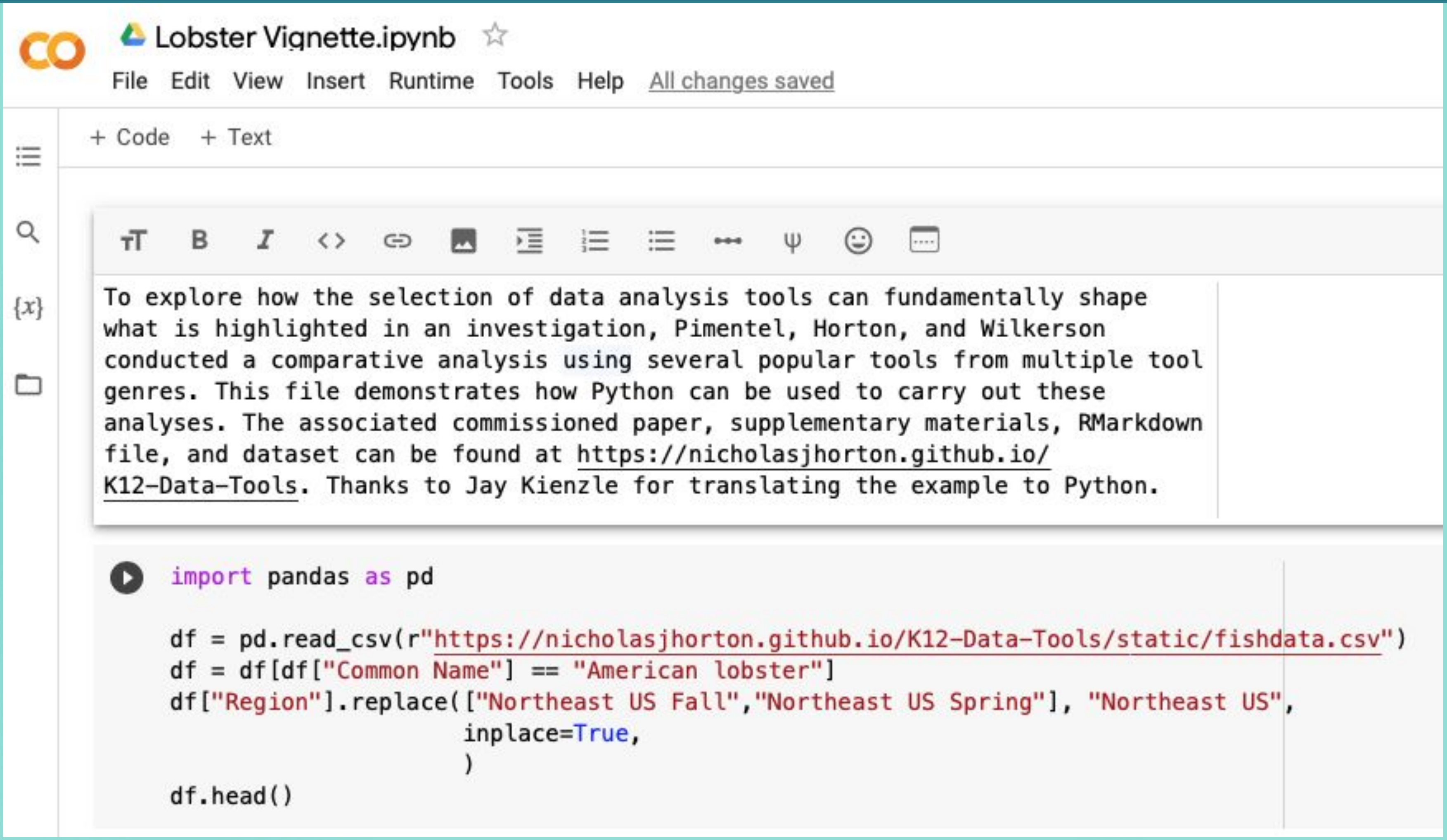
```
R version 4.2.1 (2022-06-23) -- "Funny-Looking Kid"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
```

▲ Name	Size	Modified
..		
□ .Rhistory	0 B	Nov 6, 2022, 7:3
□ fishdata_just_R.pdf	6.5 MB	Nov 6, 2022, 7:4
□ fishdata_just_R.Rmd	6.7 KB	Nov 6, 2022, 7:3
□ fishdata.csv	87.1 KB	Nov 6, 2022, 7:3
□ project.Rproj	205 B	Nov 6, 2022, 7:4

Try it at https://rstudio.cloud/content/4896839

Lobsters in Python



To explore how the selection of data analysis tools can fundamentally shape what is highlighted in an investigation, Pimentel, Horton, and Wilkerson conducted a comparative analysis using several popular tools from multiple tool genres. This file demonstrates how Python can be used to carry out these analyses. The associated commissioned paper, supplementary materials, RMarkdown file, and dataset can be found at https://nicholasjhorton.github.io/K12-Data-Tools. Thanks to Jay Kienzle for translating the example to Python.

```
import pandas as pd

df = pd.read_csv(r"https://nicholasjhorton.github.io/K12-Data-Tools/static/fishdata.csv")
df = df[df["Common Name"] == "American lobster"]
df["Region"].replace(["Northeast US Fall","Northeast US Spring"], "Northeast US",
                     inplace=True,
                     )
df.head()
```

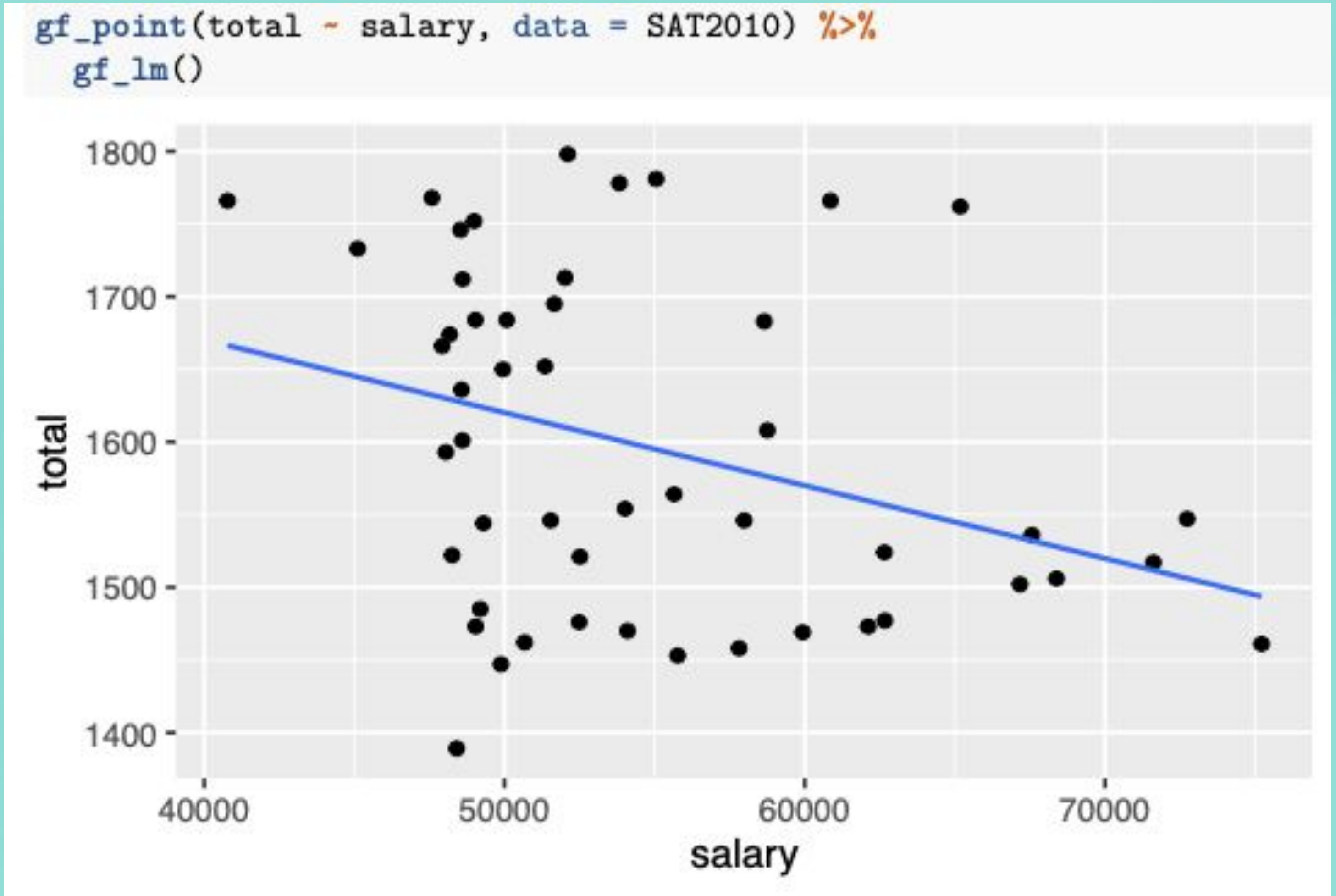
Try it at https://colab.research.google.com/drive/1teSZfFBm_o_2oe0-AMWhokRFGqV40tgp?usp=sharing

Thoughts about coding/scripting tools

- Python, R, Julia, Pyret are freely available
- Support for reproducibility
- Excellent graphical displays
- Most functionality (these are professional tools not designed for teaching [Pyret an exception])
- Tradeoff: feature a steep learning curve, documentation sometimes hard to fathom
- Challenge for students and instructors
- Potential for simplified interfaces (e.g., Data8 for Python, Pyret, Project MOSAIC for R)

Project MOSAIC: simplified access to modeling in R

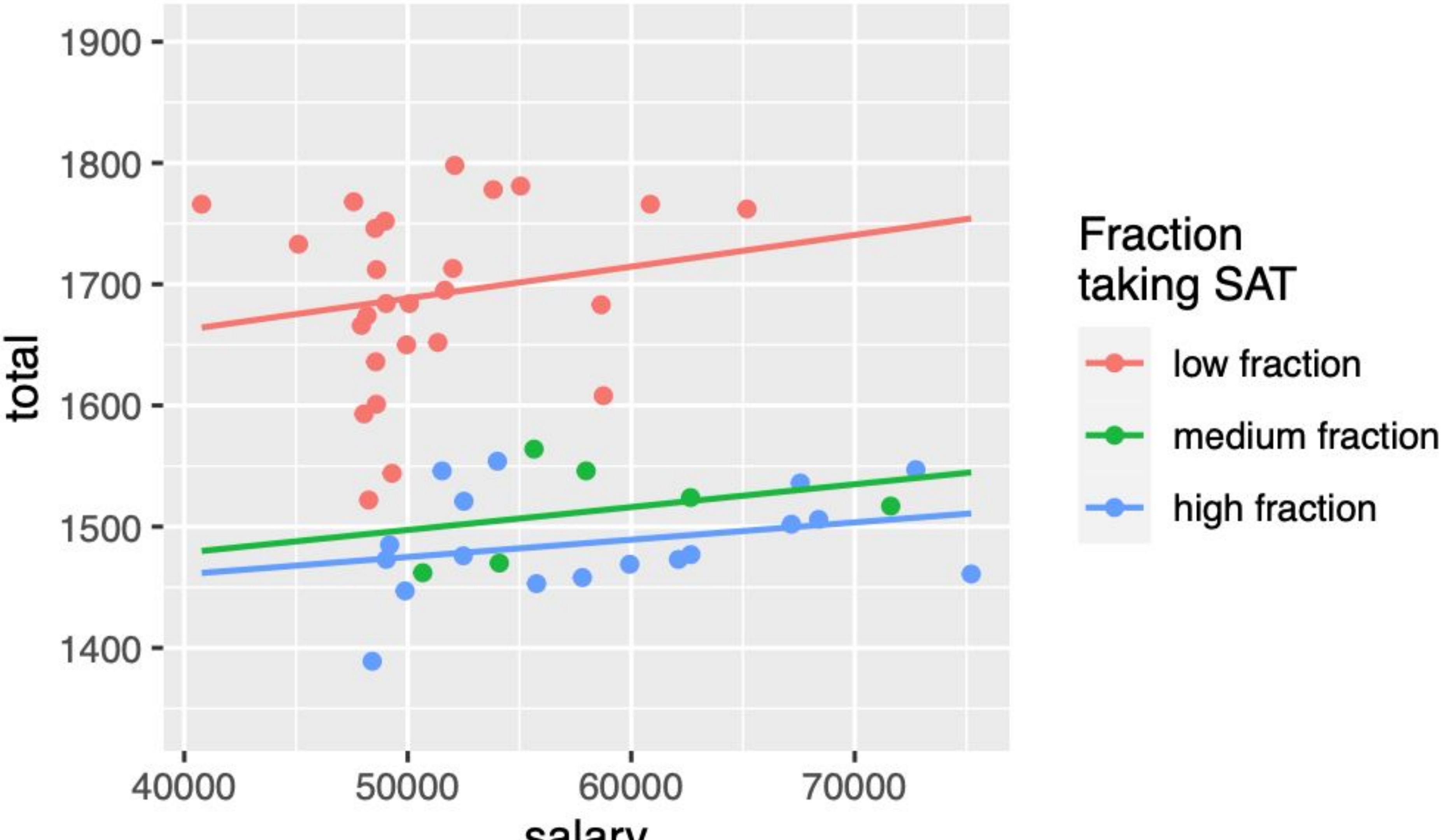
https://journal.r-project.org/archive/2017/RJ-2017-024/index.html



```
gf_point(total ~ salary, data = SAT2010) %>%
  gf_lm()
```

Project MOSAIC: simplified access to modeling in R

Learn more at http://www.mosaic-web.org/



Big picture

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

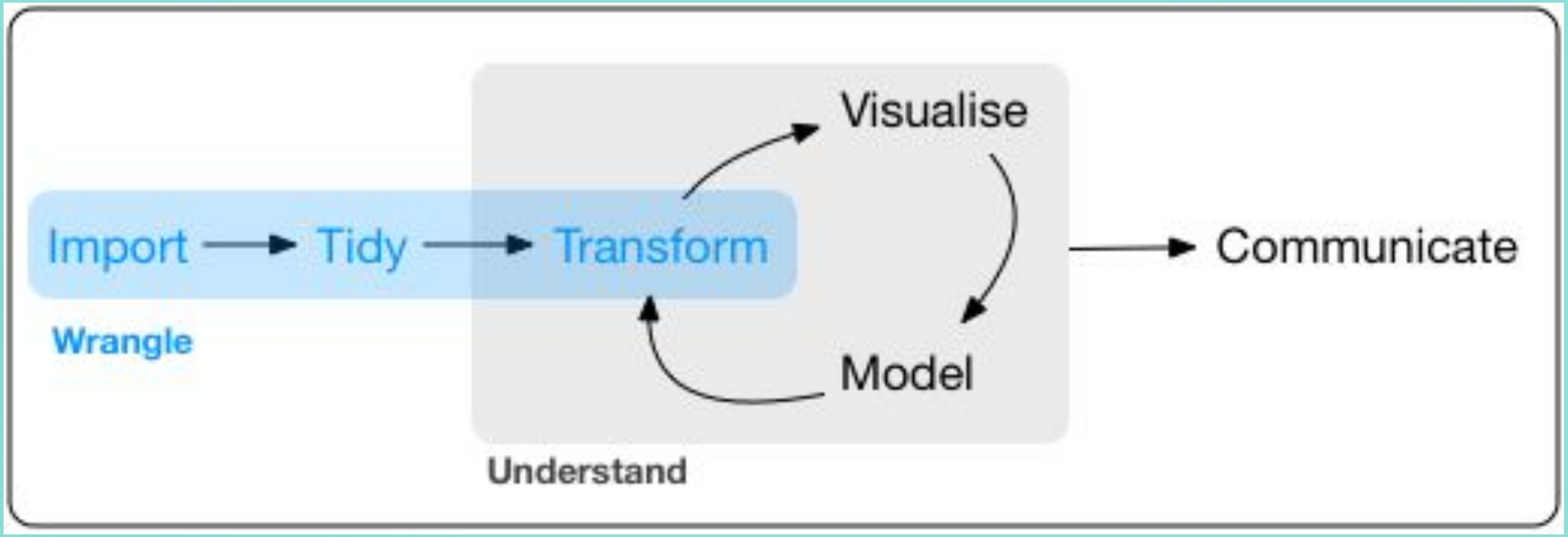


Image source: Hadley Wickham and Garrett Grolemund

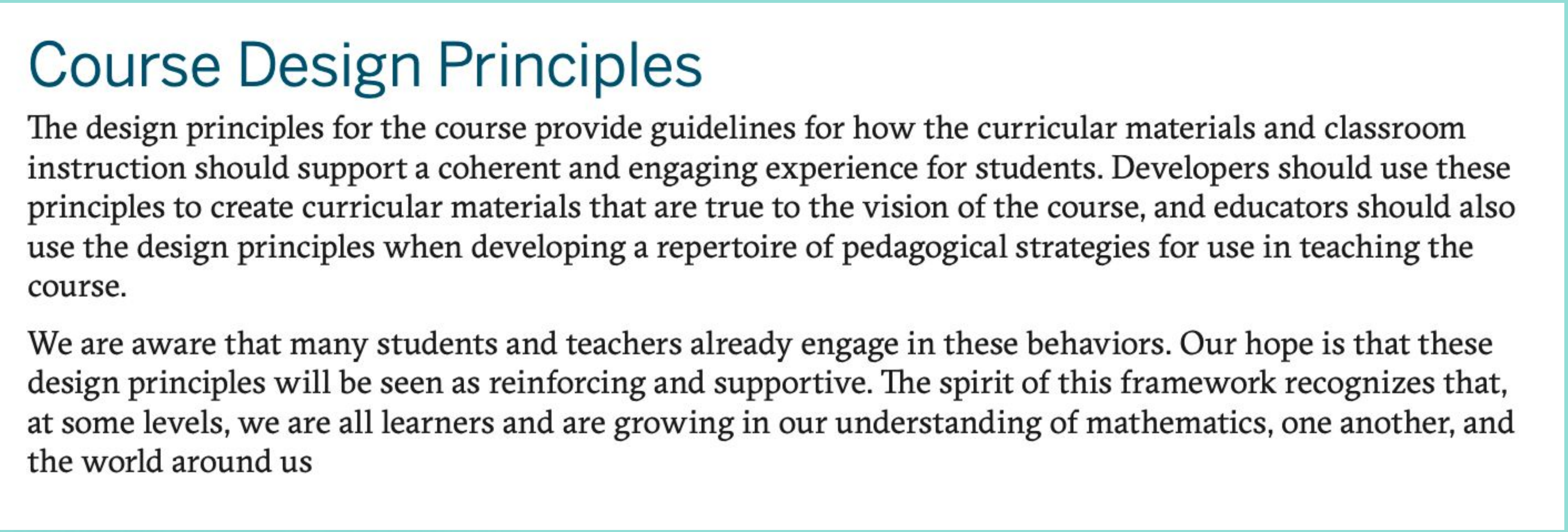
Comparisons of genres

- Need to develop a **learning progression** that involves repeated opportunities to practice the entire data analysis cycle
- This will likely involve **multiple** tools **introduced** and **reinforced** over time (see commissioned paper)
- Biehler (ISR, 1997) challenged the community to improve tools for teaching
- McNamara (2019, TAS) updated this guidance and provided a framework (which we adapted for K12)
- We need more research and development to improve the tools (and smooth out the rough edges)

Other important issues

- Systemic inequities exist in STEM education (K12 and post-secondary education): we need to ensure that these disparities are addressed as we expand K12 data science education
- This includes the types of investigations that tools can support as well as the structure and values embedded within the design of the tools
- Future work on tools should prioritize diversity and inclusion.

Inclusive practices (Dana Center Framework for Data Science)



Course Design Principles

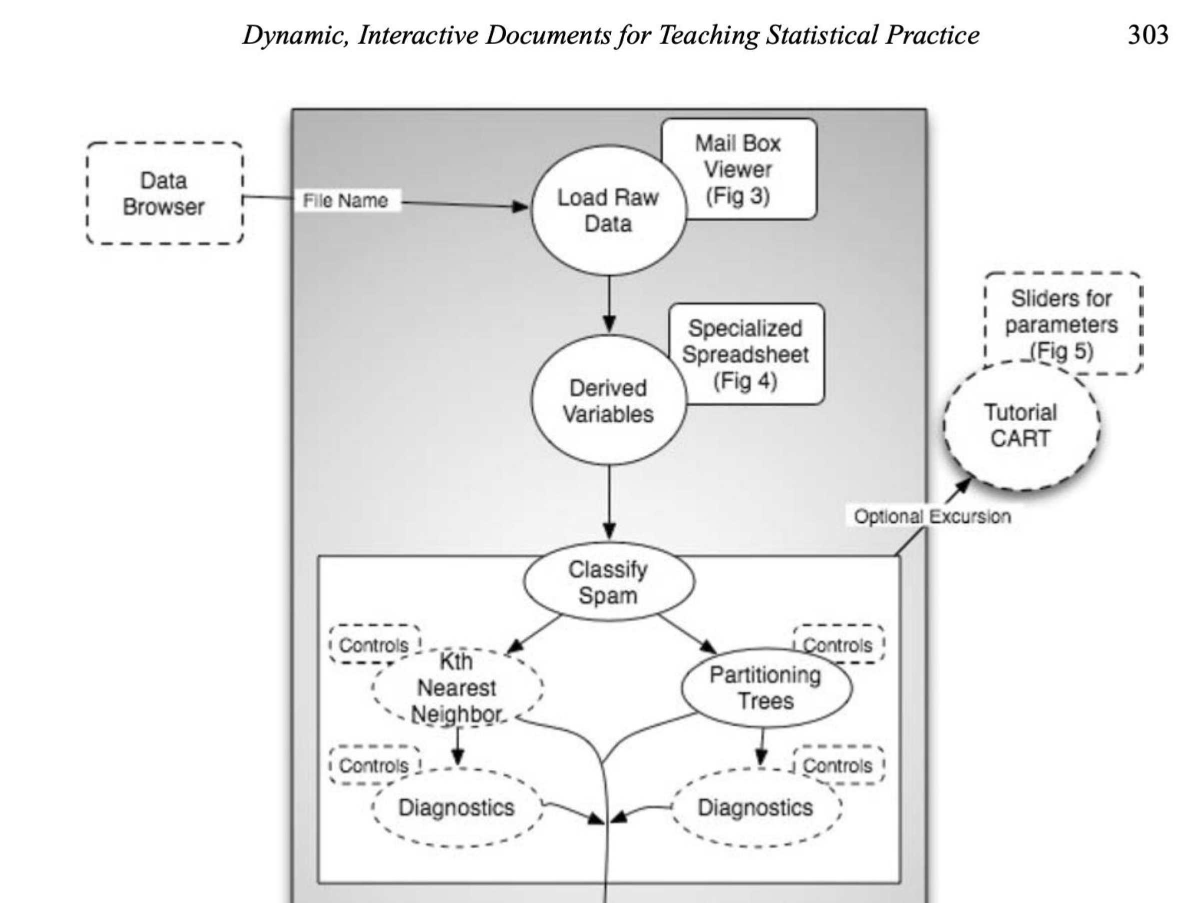
The design principles for the course provide guidelines for how the curricular materials and classroom instruction should support a coherent and engaging experience for students. Developers should use these principles to create curricular materials that are true to the vision of the course, and educators should also use the design principles when developing a repertoire of pedagogical strategies for use in teaching the course.

We are aware that many students and teachers already engage in these behaviors. Our hope is that these design principles will be seen as reinforcing and supportive. The spirit of this framework recognizes that, at some levels, we are all learners and are growing in our understanding of mathematics, one another, and the world around us

https://www.utdanacenter.org/sites/default/files/2021-05/data_science_course_framework_2021_final.pdf

Next step for tools

Potential for "Dynamic documents" (Nolan and Temple Lang, ISR, 2007, https://doi.org/10.1111/j.1751-5823.2007.00025.x)



Dynamic, Interactive Documents for Teaching Statistical Practice 303

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**.

Shameless plug: JSDSE

- The *Journal of Statistics and Data Science Education* is a 30-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 4-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 paper proposal) welcomed

Zoom poll #4

- Are there ways that we as a community can help to support Data Science education at the K12 level? What are the next steps?

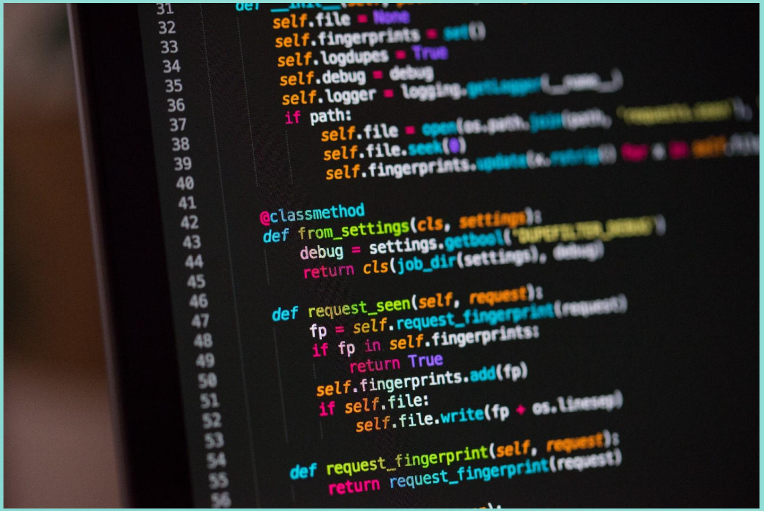
Please chime in via the chat window as you have ideas to share.

Better Data Tools Foster Better Data Science Education

Nicholas J. Horton, Amherst College

November 28, 2022, 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

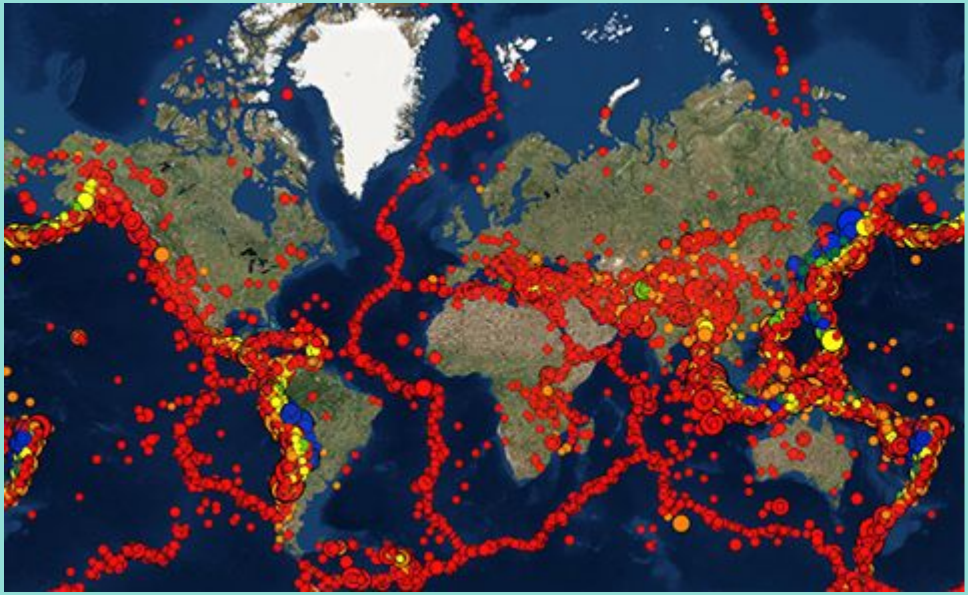


Image source: Concord Consortium

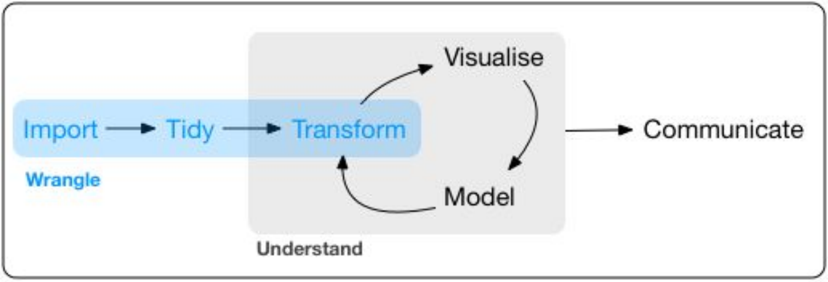


Image source: Hadley Wickham and Garrett Grolemund

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