Strand A

Essential Core Ideas

Dispositions and Responsibility

This strand focuses on what data is and all of the ways students should think about and frame it as a concept and tool.

The nature of data is complex, diverse, and humanistic. When engaging with data you must consider the form it takes, where it can come from, and what it can and should be used for. Working with data is non-linear and often raises new questions while seeking answers to others. Additionally the data process is influenced at all stages by the humans working with it which can lead to biases and concerns about ethics and responsibility. However, data can also be powerful for supporting the advancement of discovery or enactment of change.

Substrand A1

Nature of Data

The nature of data is complex, variably, humanistic, and often incomplete. Data can take many forms and may come from many different sources. Additionally, data is integral to the field of AI.

Concept A.1.1

Data types and forms

Recognize that data can exist as quantitative, ordinal, categorical, and other values. Data also can be “nontraditional” forms such as graphical or other media.

A.1.1a

Define "qualitative" and "quantitative" and understand how they relate to categorical and numeric data.

A.1.1b

Understand that forms of media (e.g., photographs, written text, audio recordings) can be represented in quantitative and qualitative terms.

Concept A.1.2

Data are produced by people

21st-century skills

AI literacy

Recognize that data represent decisions about measurement and inclusion involving people who are and are not immediately present.

A.1.2a

Explain how data-based decisions are revisited as new evidence or societal needs emerge (e.g., blood pressure cut-off numbers, dietary guidance, medical benchmarks).

A.1.2b

Evaluate why data models require updates to maintain accuracy and relevance.

Concept A.1.3

Variability of data

Recognize that variability is a foundational component of data.

A.1.3a

Recognize the different types of variability (e.g., natural, measurement, sampling).

Concept A.1.4

Data provides partial information

Recognize that data captures certain aspects of a model of a target phenomenon or set of objects in the world but does not represent it completely.

A.1.4a

Evaluate claims derived from data by questioning how phenomena are measured, categorized, or represented.

Concept A.1.5

Data and AI

Recognize that data “fuels” AI, that AI can be compared to a function machine (math), algorithm (CS), or a prediction model (statistics) that relies on data to both operate and improve itself, and that AI tools can also be used to analyze complex data in research.

A.1.5a

Describe the basic mathematical features of an AI model in terms of independent variables (e.g., inputs), dependent variables (e.g., outputs), and predictors or weights (e.g., slopes of many variables). e.g., AI models use math to weigh inputs, such as a music recommendation model might calculate: $(play\_count \times weight_1) + (listen\_duration \times weight_2) + (skip\_count \times weight_3) = recommendation\_score$, and weights are adjusted automatically to minimize mismatches between predicted and actual user preferences.

A.1.5b

Describe and explore how it is possible for data in a variety of formats (e.g., images) to be translated into organized, numerical information for an AI model to process.

A.1.5c

Identify how biases in training data can lead to biases in AI models by directly affecting predictors or weights. e.g., If AI only sees pictures of cats in sunlight, it would fail to recognize cats in shadows

Substrand A2

Data Ethics and Responsibilities

The data process is influenced at all stages by the humans working with it which can lead to concerns about ethics and responsibility. It is important when working with data to consider the use risks as well as the benefits. Data can be powerful for supporting the advancement of discovery or enactment of change.

Concept A.2.1

21st-century skills

Durable skills

Data use risks and benefits

Recognize that data can pose risks but also benefits for individuals and groups, and understand its potential uses, limitations, and risks, including unintended consequences.

A.2.1a

Analyze how data use can perpetuate biases or systemic inequities (e.g., predictive policing, hiring algorithms).

A.2.1b

Evaluate context-specific risks and benefits of data interpretations (e.g., health tracking for improving care vs. privacy concerns).

Concept A.2.2

21st-century skills

Durable skills

Biases in data

Recognize all data contains bias but data collection and analysis methods can increase or mitigate the effects of biases.

A.2.2a

Recognize how biases can obscure inferences drawn from data.

A.2.2b

Consider how the consolidation or combination of different data can create additional biases.

Concept A.2.3

21st-century skills

Media literacy and digital citizenship

Power of data

Recognize data empowers discovery, decision-making, and advocacy across fields.

A.2.3a

Evaluate how data drives innovation in fields and informs community choices.

Substrand A3

Investigative Dispositions

Working with data is non-linear and often requires cycling between phases in various orders multiple times. The process of investigating with data often raises new questions while seeking answers to others. Additionally, data is influenced by the humans working with it and the contexts within which they work.

Concept A.3.1

The investigative process

21st-century skills

Durable skills

Recognize that making sense with data requires engaging with it in a particular way that includes combinations of the concepts and practices in the other four strands.

A.3.1a

Design and refine investigations to address contextual problems (e.g., social, educational, business, medical, governmental issues), evaluating limitations and biases.

Concept A.3.2

Iteration

21st-century skills

Durable skills

Recognize that the investigative process is not linear but cyclic and iterative, with many of the phases repeating and looping back.

A.3.2a

Employ iteration in an investigation to strengthen interpretations or inspire new investigations.

Concept A.3.3

Dynamic inferences

21st-century skills

Durable skills

Recognize that inferences from data are dynamic, evolving with new data and additional analysis.

A.3.3a

Use digital tools to test and refine inferences from large or complex datasets.

Concept A.3.4

Apply context

Recognize that the context surrounding the data and the investigation shapes interpretation. Many fields (biology vs. psychology; economics vs. sociology) have created very different frameworks to organize problems. Considering multiple approaches may reveal useful insights from the same data.

A.3.4a

Reinterpret data from multiple perspectives, disciplines, and historical frames of reference.

A.3.4b

Compare and contrast problem solving approaches and the resulting findings.

Concept A.3.5

Student data agency

21st-century skills

Durable skills

Cultivate the motivation to engage with data in all areas of life and understand how data impacts your own experiences.

A.3.5a

Utilize data science tools and methods to engage in personal and collective inquiry relevant to one's own life and interests.

Creation and Curation

This strand focuses on where data comes from and how it should be collected, organized, and formatted in order to make it useful.

Data collected from real world scenarios is often complex and messy, and whether it is collected first hand, or retrieved second hand from an external source, it requires curation and cleaning before analysis. The context of data collection matters and affects the nature of errors in data collection. The methods and decisions made during data collection affect the usefulness of the data and its ability to answer different questions.

Substrand B1

Organization and Processing

In order for data to be useful for analysis and visualization, it often needs to be organized and formatted in particular ways. Organization can include both procedural cleaning up of errors or mistakes and processing or transforming the data through calculations and logic statements to create new or summative measures.

Concept B.1.1

Data cleaning

Identify and address data quality issues to ensure accuracy and reliability, progressing from simple error identification to using systematic approaches.

B.1.1a

Use data dictionaries to identify codes for missing or incomplete data (e.g., NA, 99999, 0, " "), and either recode or filter data to remove those observations.

B.1.1b

Apply basic cross-validation techniques to verify data quality across multiple sources, including source comparison, split sampling, internal consistency checks, and domain range validation.

Concept B.1.2

Organizing and structure

Organize raw data into structured formats using categories, tables, and systematic recording methods.

B.1.2a

Create and manage complex data structures with multiple related tables, understanding primary and foreign key relationships between datasets.

B.1.2b

Transform and restructure hierarchical or nested data into normalized tabular formats suitable for analysis.

B.1.2c

Design efficient organizational schemas for large datasets with multiple variables and complex relationships.

Concept B.1.3

Processing and transformation

Transform and manipulate data through sorting, grouping, filtering, and combining datasets.

B.1.3a

Use calculations and logic statements to create new categorical variables based on existing categorical (e.g., if(employment="employed", Yes, No)) or quantitative variables (e.g., if(weight<30, light, if (weight>60,heavy ,medium)).

B.1.3b

Filter data based on groups or subsets of data relevant to the problem and context.

Concept B.1.4

Summarizing groups

Calculate and analyze group-level statistics from detailed data to reveal patterns and relationships.

B.1.4a

Create summary measures for groups that can then be used as a measure at the group level (e.g., for salary data, compute average salary for different occupational groups).

Substrand B2

Designing for Data Collection

The design of a data investigation is as important as the data collection process. Framing a data-based investigation requires identifying a problem or question to be explored. Additionally, the methods must be carefully chosen and the values and tradeoffs considered.

Concept B.2.1

Designing data-based investigations

Identify problems and formulate questions that guide meaningful data collection and analysis.

B.2.1a

Construct data-based questions about the design of a study to determine causality and make predictions.

B.2.1b

Identify comparison and association data-based questions appropriate for addressing problems of interest.

Concept B.2.2

21st-century skills

AI literacy

Data creation techniques and methods

Explore various ways to generate data through simulations, sensors, and automated collection methods.

B.2.2a

Describe benefits and drawbacks of using proxy variables.

B.2.2b

Use and/or change parameters of simulations to generate data to address a problem of interest.

B.2.2c

Design sensor-based experiments or automated data collection scenarios to explore a problem or question and identify the scenarios' limitations and trade-offs.

B.2.2d

Describe the features, benefits, limitations, and ethical thinking that went into a data collection process.

B.2.2e

Design and implement traditional data collection methods (e.g., surveys, observations, field studies) to investigate research questions and evaluate their strengths and limitations compared to automated approaches.

Concept B.2.3

21st-century skills

Durable skills

Creating data collection plans

Develop systematic plans that specify what data to collect, how to collect it, and from what sources to answer investigation questions.

B.2.3a

Develop comprehensive data collection plans that address potential limitations, specify quality control measures, and include contingency strategies.

Concept B.2.4

Finding secondary data

Explore, locate, evaluate, and retrieve datasets collected by others to address research questions and data investigations.

21st-century skills

Media literacy and digital citizenship

B.2.4a

Locate and retrieve relevant datasets from publicly available scientific, civic, or government databases using search tools and filters.

B.2.4b

Evaluate datasets from multiple sources to determine which best addresses a research question, considering factors such as data quality, sample size, and collection methods.

B.2.4c

Use data catalogs, repositories, and open data portals to find datasets that meet specific criteria for investigations.

Substrand B3

Measurement and Datafication

The methods and decisions made during data collection affect the usefulness of the data and its ability to answer different questions. It is important to consider the potential effects of methodological decisions when collecting data and to determine the methodological decisions made by others when using secondary data. It is also important to consider ethical practices of using other's data.

Concept B.3.1

Creating your own data

Collect, measure, and document data accurately using appropriate tools and methods.

B.3.1a

Create a data dictionary to document the data collection process.

Concept B.3.2

Working with data created by others

Evaluate and interpret others' datasets by examining collection methods, context, and quality.

21st-century skills

Durable skills

B.3.2a

Make use of metadata and data dictionary to understand a data set.

B.3.2b

Consider who or what was included in the data collection and who or what was not.

Concept B.3.3

Ethics of data collection and usage

Collect and use data ethically, considering privacy, fairness, and potential impacts.

21st-century skills

Durable skills

AI literacy

B.3.3a

Evaluate and address ethical implications of data collection choices, including privacy, bias, and representation.

B.3.3b

Analyze existing datasets for potential bias, discrimination, or unfair representation.

Substrand B4

Complexity of Data

Data collected from real world scenarios is often complex across many dimensions including messiness, size, and structure. In order to be able to work with authentic real-world datasets of high complexity, these dimensions must be scaffolded such that increasingly higher levels of complexity are encountered as one approaches mastery.

Concept B.4.1

Cleanliness

Work with datasets at increasing levels of cleanliness and identify how datasets need to be curated to address messiness issues.

B.4.1a

Work with datasets requiring multiple types of cleaning such as missing values, errors, and anomalies.

B.4.1b

Clean and prepare datasets before merging multiple sources.

Concept B.4.2

Complexity of variables

Explore datasets containing various types of data and understand how each type serves different analytical purposes.

B.4.2a

Work with datasets that include time-series data at different intervals to detect various patterns.

B.4.2b

Understand and work with different observation structures beyond individual units.

B.4.2c

Work with merged datasets that align different time scales and observation structures.

Concept B.4.3

Size

Work with datasets of increasing size in both number of observations and variables and arrange data in increasingly complex formats to facilitate meaningful analysis.

B.4.3a

Work with datasets with over 20 variables and over 1000 observations.

B.4.3b

Transform data between wide and long formats based on analysis needs.

Concept B.4.4

Complexity of structure

Manipulate and combine data in increasingly complex ways to reveal new insights and patterns.

21st-century skills

AI literacy

B.4.4a

Merge multiple datasets while maintaining appropriate observation structure.

B.4.4b

Transform complex variables into more interpretable forms using student-relatable benchmarks.

Analysis and Modeling Techniques

This strand focuses on the process of analyzing data.

Analyzing data includes many different techniques such as examining single and multi-variable patterns, measures of centrality, variability, and uncertainty. Knowing which techniques to use on which types of data to answer which questions is as important as the skills to conduct analysis techniques. Additionally, understanding simulation and the relational nature of data is important to the analysis process, as is the use of technological tools for analysis and modeling.

Substrand C1

Summarizing Data

Raw data often is not useful for answering questions, making claims, or telling a story. In order to derive understanding it is usually useful to have a summary of the data which provides measures of the centrality, spread, and shape of the dataset.

Concept C.1.1

Measures of center

Analyze large datasets by measuring their central tendency while considering the context and distribution of the data.

C.1.1a

Identify appropriate ways to summarize numerical or categorical data using frequency tables, graphical displays, and numerical summary statistics.

Concept C.1.2

Measures of spread

Examine dataset variability by applying measures of spread to identify and quantify outliers.

C.1.2a

Calculate standard deviation from mean or interquartile range.

C.1.2b

Use standard deviation as a measure of variability and a modified boxplot for identifying outliers.

Concept C.1.3

Shape

Identify the distribution of data points, including clusters, gaps, symmetry, skewness, and modes. Use these patterns to understand data spread and their impact on measures like the mean and median.

C.1.3a

Acknowledge that in a tie for the mode the distribution is bi-modal.

C.1.3b

Understand how the data is distributed across the range of data. e.g., if the data is skewed to one side of the range

Concept C.1.4

Frequency tables

Organize data into frequency tables based on shared characteristics. Summarize data using counts, fractions, relative frequencies, or proportions to enable comparisons and generalizations. Understand the implications of choices made when creating and interpreting frequency tables.

C.1.4a

Generate a relative frequency table to make comparisons and to generalize results.

Concept C.1.5

Missingness

Identify and describe missing data numerically and categorically. Distinguish between missing values and true zeros. Understand how missing data impacts relationships, patterns, and models in data interpretation.

C.1.5a

Adjust analyses in light of missing values.

Concept C.1.6

Metadata

21st-century skills

AI literacy

Recognize metadata as information about data, including its source, type, and structure. Use metadata to organize, summarize, and analyze data effectively, supporting interpretation and decision-making.

C.1.6a

Apply understanding of metadata (e.g., data and time, text, continuous, geolocation) to summarize and analyze data numerically, in tables and through visualizations.

Substrand C2

Identifying Patterns and Relationships in Data

A primary use of data is in understanding patterns and relationships across different variables and scenarios. As all data contains variability it is important to understand and analyze distributions both within and across variables.

Concept C.2.1

Comparing variables

Identify similarities and differences between variables and explore potential associations. Use distributions, numerical summaries, and simulations to compare groups based on numerical or categorical data.

C.2.1a

Use numerical measures such as average, standard deviation and quartiles to compare two groups.

Concept C.2.2

Understanding distributions

Represent data visually and numerically to describe how outcomes occur and compare groups. Use variability to interpret distribution shape, support statistical reasoning, and assess population estimates.

C.2.2a

Quantify variability in distributions using numerical measures.

C.2.2b

Recognize the relationship between variability and the shape of a distribution.

Concept C.2.3

Defining relationships

Organize, visualize, and analyze data to identify patterns, trends, and associations. Use statistical measures and graphs to interpret relationships and make predictions.

C.2.3a

Describe associations between two categorical variables using measures such as difference in proportions and relative risk.

C.2.3b

Analyze data to uncover correlations, trends, and groupings such as clustering that inform decision-making processes across diverse fields.

Concept C.2.4

Analyzing non-traditional data

21st-century skills

AI literacy

Examine data beyond numbers, including sounds, textures, and text. Categorize sensory inputs, track word frequencies, and analyze data from sensors and IoT devices to identify patterns and trends.

C.2.4a

Analyze data from sensors and IoT devices to track trends and monitor changes over time. e.g., smart thermostats and lighting systems for energy monitoring, wearable fitness trackers for health and activity data

C.2.4b

Understand that geographic data can be visualized using maps, and it can be represented as points (e.g., latitude and longitude) and areas (e.g., GeoJSON).

Concept C.2.5

Machine learning

21st-century skills

AI literacy

Use data to build decision trees, explore classification and clustering, and understand how machine learning optimizes predictions through algorithms like gradient descent.

C.2.5a

Explore machine learning basics (e.g., classification and clustering) to make predictions with data.

Substrand C3

Variability in Data

Variability is omnipresent within data and datasets. Working with data depends on understanding, explaining, and quantifying variability of all forms (variability within a group, between different groups, or between samples).

Concept C.3.1

Describing variability

Identify differences within data by sorting, grouping, and organizing characteristics. Use statistical and simulation methods to represent and analyze variability, connecting it to real-world uncertainty and probabilistic processes.

C.3.1a

Describe methods (e.g., statistical, simulation) to analyze variability in data and connect it to known or hypothesized processes in a specific domain.

Concept C.3.2

Comparing variability

Examine differences between groups by analyzing measures of spread, such as range and standard deviation. Utilize visualizations like box plots and apply statistical methods, including mean, median, and standard deviation, to compare datasets, assess variability, and uncover patterns in data distributions and models.

C.3.2a

Use simple statistics including mean, median, range, standard deviation, etc. to compare data distributions.

Concept C.3.3

Understanding sources of variability

Recognize measurement errors and natural variability in data. Assess data quality, identify outliers, and refine models using statistical and contextual analysis.

C.3.3a

Consider variability as a key component of informal inference by questioning whether observed differences are meaningful or not. e.g., phone battery lasts 6 hours one day and 4 the next—is this a real difference in battery life, or just normal variation from daily use

C.3.3b

Identify categorical options for measuring "best" fit from data points to provided estimates. e.g., line or curve for a scatterplot, mean for a distribution

C.3.3c

Consider both context and the characteristics/source of a dataset to determine how "messy" a dataset may be due to measurement error. e.g., faulty sensors, inaccurate or inappropriate measurements

C.3.3d

Use errors to improve the AI and/or machine learning model.

Concept C.3.4

21st-century skills

AI literacy

Variability in our computational world

Explore how AI model outputs vary based on training data, labeling, and bias. Understand how generative AI and pre-trained models use large datasets to make inferences and how variability in data impacts outcomes.

C.3.4a

Acknowledge how variability in the training data for generative AI influences bias in its output. e.g., facial recognition, ownership of DNA data

Substrand C4

Digital Tools of Data Analysis

While some datasets can be explored by hand, as they get bigger and more complex it becomes necessary to use digital tools for analysing data. It is important to understand which tools to use for which application or scenario, the affordances and tradeoffs, and the ethical considerations of using certain tools.

Concept C.4.1

Tool application

Use digital tools to summarize data and create visualizations. Apply these tools to identify patterns, clean and prepare data, perform analysis, and build models for simulations to explore relationships and trends.

C.4.1a

Identify relationships and patterns using a digital tool.

C.4.1b

Clean and wrangle data using a digital tool.

C.4.1c

Create multi-variable visualizations using digital tools.

Concept C.4.2

21st-century skills

AI literacy

Tool ethics

Examine how digital tools influence access, privacy, and bias, shaping opportunities and challenges in technology use. Consider the broader ethical and societal impacts of AI, including its role in decision-making, accountability, and policy.

C.4.2a

Describe the ethical limitations (e.g., environmental, privacy, copyright, hallucination) of using AI tools.

Concept C.4.3

Tool evaluation

Assess the technical limitations of digital tools and compare no-code, low-code, and high-code solutions based on their capabilities and use cases.

C.4.3a

Identify the technical limitations of a digital tool.

Concept C.4.4

Tool selection

Choose the appropriate no-code, low-code, or high-code digital tool based on the task. Use multiple tools throughout the data investigation process and explore how digital tools are applied in the workforce.

C.4.4a

Select a no-code, low-code or high-code digital tool that is suited for the intended task.

Concept C.4.5

21st-century skills

AI literacy

The role of code in data analysis

Explore how block coding and computer code automate and enhance data analysis. Understand how coding enables reproducible processes and compare its advantages and limitations to no-code and low-code tools.

C.4.5a

Recognize how computer code can automate data investigation processes.

C.4.5b

Recognize how computer code can automate data analysis processes.

Concept C.4.6

Tool accessibility for diverse learners

Understand how digital tools can support a broad range of diverse learners. Evaluate their effectiveness and impact, and explore inclusive data representations.

C.4.6a

Explore how to communicate with data while prioritizing accessibility.

C.4.6b

Critique the levels of accessibility of digital tools and representations of data.

Substrand C5

Models of Data

Interpreting, creating, and using models is a central component of working with data. Models are both a way to analyze data and a source of data.

Concept C.5.1

21st-century skills

AI literacy

Understanding modeling

Analyze patterns and relationships in data using graphs, tables, and models. Explore tools like decision trees and neural networks, assess assumptions, and distinguish correlation from causation in real-world contexts.

C.5.1a

Recognize that bivariate relationships between numerical features can be examined using both linear and non-linear associations.

C.5.1b

Investigate real-world examples where correlation does not imply causation.

Concept C.5.2

Creating models

Develop an understanding of patterns and relationships. Use data and technology to build and refine models. Advance these skills by constructing complex models that incorporate multiple variables, assess assumptions, and improve predictions.

21st-century skills

AI literacy

C.5.2a

Construct and analyze models to represent linear and non-linear relationships in data.

C.5.2b

Use technology to create, test, and refine models.

C.5.2c

Evaluate and improve models by comparing predictions to observed data.

Strand D

Tools and Techniques

Interpreting Problems and Results

This strand focuses on justification and explanation of reasoning when making inferences, claims, or suggestions from data within the context and processes of the dataset collection and analysis.

An important component of interpreting results is understanding the relationship between questions, problems and datasets. Formulating a strong question or identifying a problem that can be addressed with data affects the opportunities for interpretation and results from the data. Additionally, the applicability of inferences and claims that are made are constrained by the sample, population, and context of the data.

Substrand D1

Making and Justifying Claims

As all data contains variability, it is important to use probabilistic thinking and language when making claims from data. This requires paying attention not only to patterns and comparisons within and across variables but also such things as expected and prior values, sample sizes, and significance.

Concept D.1.1

Probabilistic language

When communicating with others, employ both plain-language and clear vocabulary to regularly describe degrees of uncertainty, both formally and informally as a thinking habit.

21st-century skills

Durable skills

D.1.1a

Clearly state a result or finding and indicate the level of certainty regarding a formal statistical concept alongside an informal evaluation of the likelihood of the event.

Concept D.1.2

Priors and updates

When encountering new data, integrate probabilistic thinking into everyday situations by explicating prior assumptions and the impact of new data / evidence on those assumptions.

21st-century skills

Durable skills

Media literacy and digital citizenship

D.1.2a

Analyze how confirmation bias and availability bias influence the way individuals evaluate new information, especially regarding their existing beliefs and assumptions.

D.1.2b

Relate assumptions about a problem to the certainty of findings based on new evidence.

Concept D.1.3

Expected value

When making a decision about uncertain outcomes in the future, integrate probabilistic thinking into everyday decisions by applying expected value (magnitude x probability) to appropriate situations.

21st-century skills

Financial literacy

D.1.3a

Identify and accurately employ the Expected Value equation (EV = P (Xi) * Xi) across multiple contexts to compare scenarios involving multiple trials. e.g., insurance policies, lotteries

D.1.3b

Solve a real-world comparison problem using a digital spreadsheet, such as selecting insurance policies or entering different lotteries. e.g., choosing insurance policies, entering different lotteries

Concept D.1.4

Explaining significance

Clearly describe the basic logic of statistical significance to others, differentiating between significance, the size of an effect, and the statistical power of an analysis. Recognize what statistical significance can reveal and cannot reveal about a phenomenon.

21st-century skills

Media literacy and digital citizenship

D.1.4a

Identify situations when distinguishing from random chance is especially important. e.g., medical drug trial, public policy implementation

D.1.4b

Describe probability distributions and give real-world examples of how they can represent different types of random events.

D.1.4c

Identify and describe a normal distribution as a possible model for random chance that can be used to determine whether a result is statistically significant.

Concept D.1.5

Sampling and simulation

Comfortably identify the purpose of sampling and simulation for making arguments about data, and employ techniques using software to differentiate a real-data result from random chance or “happenstance.”

D.1.5a

Use simulations in a digital software to help determine whether the results of an experiment are likely due to something other than random chance.

D.1.5b

Analyze how dataset bias impacts sample results over time by introducing intentional bias sources in digital simulations and observing their effects.

D.1.5c

Answer probabilistic questions resulting from a simulation.

Concept D.1.6

Correlation versus causation

Comfortably separate correlation from causation in a wide variety of situations, building a “first-reaction” thinking habit over time.

21st-century skills

Media literacy and digital citizenship

D.1.6a

Recognize that a randomized experiment is the best way to establish evidence for causation and justify a claim through isolating the effect of only one independent variable on another variable at a time.

D.1.6b

Identify spurious correlations in the media and explore other potential causes that may explain these associations when applicable.

Concept D.1.7

Randomization

When identifying a potential cause of a phenomenon, clearly describe the usefulness of randomization for constructing an argument with data.

D.1.7a

Explain why randomization mitigates many potential sample biases (e.g., observation bias, collection errors, selection bias) concurrently in a variety of examples.

Concept D.1.8

21st-century skills

Durable skills

Multi-variable decision-making

Clearly describe how to leverage additional variables or additional outside data to make a logical argument, and identify potential risks of overdoing it.

D.1.8a

Use computer software to explore how adding additional numerical variables to a linear model changes the interpretation of the results.

D.1.8b

Use computer software to analyze the relationship between two or more numerical variables by interpreting the strength and direction (e.g., positive, negative, none) of the association using computed values.

Substrand D2

Problem Identification and Question Formation

Formulating a question or identifying a problem that can be addressed with data affects the opportunities for interpretation and results from the data. The ability to make and justify strong claims relies on identifying questions that are testable and can be answered with data. Additionally, identifying the uncertainty or limitations within the problem space is an important component of formulating conclusions

Concept D.2.1

Verifiable questions and statements

Identify and create the type of questions that can be answered by data, and are eventually verifiable using a combination of modeling and experimentation.

D.2.1a

Differentiate query-based, hypothesis-based, and causal questions by their focus on trends, uniqueness of outcomes, and causal relationships, respectively.

D.2.1b

Assess query-based questions by establishing a threshold of satisfaction for certainty in interval estimates (e.g., if it applies 95% of the time, I find it acceptable).

D.2.1c

Assess hypothesis-based questions by debating the condition of uniqueness (e.g., if it occurs 5% of the time or less).

Concept D.2.2

21st-century skills

Durable skills

Iteration, validation, and multiple explanations

Regularly practice identifying alternative explanations for a result from data, both for interim steps and post-analysis conclusions.

D.2.2a

Identify various possible explanations for an observed association by investigating and comparing relationships between variables within a dataset.

D.2.2b

Regularly log questions during data analysis and identify additional factors that may clarify associations. e.g., knowing X would be helpful because it would explain or rule out Y

Concept D.2.3

Uncertainty statements and limitations

Clearly explain the limitations and caveats of a conclusion from data, including the risks of extending the conclusion to another group or situation.

D.2.3a

Identify potential issues in data investigations and state what cannot be reasonably concluded from the available data and approach, noting areas that may require further investigation.

Concept D.2.4

Relevant conclusions

21st-century skills

Durable skills

Ensure that increasingly complex analysis steps remain useful for the original question, and that the method does not distract from the problem.

D.2.4a

Formulate a statement that directly addresses the original investigation question, incorporates relevant statistical data to substantiate the conclusion, and interprets the statistical results to explain their broader implications in practice. e.g., statistical claims are not solely about numbers, they also interpret what the results signify and why they are important for solving a problem or answering a question

D.2.4b

Identify statements that do NOT include descriptions of the data and context implications that address the original investigation question.

Substrand D3

Generalization

Though there is often an instinct to use data to make large generalized claims, the applicability of inferences and claims that are made are constrained by the sample, population, and context of the data.

Concept D.3.1

Application fitness

21st-century skills

Media literacy and digital citizenship

AI literacy

Regularly identify generalization issues, with frequent comparisons between significant real-world examples and a current analysis.

D.3.1a

Examine and identify common generalization issues from data-based conclusions in the media.

D.3.1b

Identify and list analysis strategies for a given data-driven conclusion to better generalize to other populations or situations.

Concept D.3.2

Sample versus population

21st-century skills

AI literacy

Given a dataset, identify constraints and opportunities for what can be logically inferred about a broader population.

D.3.2a

Analyze a population through a sample by clearly articulating how the chosen sampling method relates to the research question.

Concept D.3.3

Sample size

When full information is hidden or inaccessible, recognize the logical relationship between a sufficient number of chances and a sufficiently large sample to reasonably represent something.

21st-century skills

Media literacy and digital citizenship

D.3.3a

Recognize there are formal methods to determine the minimum sample size needed to make a well-supported claim about a population.

D.3.3b

Explain “statistical power” of a statistical test as the general probability that an outcome “lands” more “extremely,” beyond an arbitrary pivotal value set for statistical significance that a researcher chooses.

D.3.3c

Explain “statistical power” as the probability that a statistical test properly detects a real effect when one exists.

Concept D.3.4

Simple bias

When information is completely hidden or unavailable, be aware of possible underlying issues in the sample and apply strategies to identify and address them.

21st-century skills

Media literacy and digital citizenship

D.3.4a

Acknowledge that a sample may be systematically skewed due to collection methods, data availability, survey design, or other reasons, particularly in a secondary data context.

D.3.4b

Identify examples of sample bias in the media or other real-world examples. e.g., medical drug trials, prior debunked research

Concept D.3.5

Extension statements

Following an initial analysis, list and implement opportunities for increasing the strength of an argument, a generalization claim, or ideas for a new analysis. Explore risks of the same approaches as well.

21st-century skills

Durable skills

D.3.5a

Identify additional scenarios for which a data-based conclusion may apply and list the similarities and differences of the new scenario.

D.3.5b

Identify the risks of extending the original analysis to a new scenario. e.g., data that might not be captured, incorrect assumptions

Concept D.3.6

Subset effects

Recognize that important information may be hidden or may even change a major conclusion when data is filtered into categories and/or groups.

21st-century skills

Media literacy and digital citizenship

D.3.6a

Create and compare subsets of a dataset with software.

D.3.6b

Discuss examples of aggregate measures of data that missed important subsets in the media or other real-world contexts.

Concept D.3.7

Meta-analysis and facts

Recognize the relationship between many trials, uncertainty, and whether a claim is a “fact.”

21st-century skills

Durable skills

Media literacy and digital citizenship

D.3.7a

Recognize that one study or data analysis may be insufficient to prove something is “true” for certain.

D.3.7b

Document data analysis steps in a shareable and reproducible format that can be repeated.

Strand E

Tools and Techniques

Visualization and Communication

This strand focuses on how to communicate about data through the creation and examination of visualizations.

Visualizations are a vital component of the sensemaking process when working with data. Being able to communicate with and about data using visualizations that are clear and tailored to a purpose and audience are an important step for creating action and impact through data. Also important are skills and habits for how to read, interpret, and critique other’s data communication, paying attention to context, audience and purpose.

Substrand E1

Representations and Dynamic Visualizations

The creation and interpretation of graphic and interactive visualizations are vital components of the sensemaking process when working with data. Working with data visualizations requires an understanding of conventional components and best practices along with graphical literacy and representational fluency.

Concept E.1.1

Sense-making with visualizations

Practice creating visualizations to summarize many things at once, relationships between things in one place, or exceedingly complex ideas in one place. Recognize that visuals can be more efficient or compelling than other forms of communication.

E.1.1a

Use computer-based analysis tools to make basic descriptive summaries of a dataset. e.g., bar charts, histograms, line graphs, scatterplots

E.1.1b

Quickly or informally estimate relationships visually by adding lines of best fit with a computer-based tool.

Concept E.1.2

Investigate with visualizations

21st-century skills

Durable skills

Create data visualizations to directly support the analysis steps of data.

E.1.2a

Visualize the distribution of raw data to identify outliers and out-of-bounds values in context.

E.1.2b

Communicate key features of distribution (e.g., measures of center, spread, shape) formally and with precision.

Concept E.1.3

Clear design for user interpretation

Identify conventional components and best practices of data visualization from a user-centered or audience perspective.

21st-century skills

Durable skills

E.1.3a

Properly cite data sources near visuals to ensure transparency and credibility.

E.1.3b

Recognize how complementary or contrasting features (e.g., color, texture, shape) can be used to represent dichotomous ideas in data visualizations.

E.1.3c

Describe how human color/contrast perception varies and apply this to select accessible data visualization palettes.

Concept E.1.4

Graphical literacy

Comfortably read graphs with accuracy and make sense of data visualizations by answering questions about how the data is represented with precision.

21st-century skills

Media literacy and digital citizenship

E.1.4a

Answer questions about and explain the data in a variety of data visualizations, including non-standard visualizations. Extract key insights, trends, and patterns from the data.

E.1.4b

Describe the potential relationships (or lack thereof) represented in scatterplots (including linear, exponential, and logarithmic) and debate which function is the best representation for the shape and context.

Concept E.1.5

Representational fluency

Identify how layout (ordering, scale, and axes) choices increase clarity or potentially mislead an audience.

21st-century skills

Media literacy and digital citizenship

E.1.5a

Compare and/or contrast visualizations of the same numerical data at different scales and understand how the scale affects people's interpretation. e.g., accurately representing the relative magnitudes vs. exaggerating them

E.1.5b

Critique misleading visualizations, such as those with truncated axes, cherry-picked data points, confusing colors, or manipulated scales. e.g., graph starting at 50 (not 0) can make a 5% drop look like a crash

Concept E.1.6

Parallel visual-type construction

Align the type of data (numeric, categorical, string, other) to a visualization type designed for that use-case.

E.1.6a

Demonstrate the wrong type of data (e.g., numeric, categorical, string) entered into a misaligned visualization package (e.g., scatterplot of categorical data) and explain why the visualization fails to work or clearly represent the data.

Substrand E2

Data Storytelling

Being able to communicate with and about data using visualizations connected to a narrative is an important step for creating action and impact through data. Understanding the audience for the narrative is vital to clear communication.

Concept E.2.1

Connect narratives and data visualizations

Understand the relationship between a data visualization and its associated narrative.

E.2.1a

Evaluate the degree to which visualizations and their surrounding text match and support real-world context.

Concept E.2.2

21st-century skills

Durable skills

Write data stories

Structure effective stories about data when complex jargon and technical ideas are involved.

E.2.2a

Explain how the data directly supports or contradicts any claims made about it while also being open about limitations such as sample size or external factors that may influence results, and anticipate potential counterarguments.

E.2.2b

Support claims by citing expert opinions or research studies that corroborate the data.

E.2.2c

Use data to explain trends and predict future outcomes based on those trends.

Concept E.2.3

21st-century skills

Durable skills

Adapt storytelling

Tailor storytelling for different audiences.

E.2.3a

Identify an audience of interest, and tailor data stories to that audience, presenting the data in a way that ensures it resonates with them.

E.2.3b

Explain the implications and takeaways by detailing how the information can be utilized in their daily lives or work experiences, while offering actionable advice that aligns with their interests and needs.

Substrand E3

Acting on Data to Benefit Society

One of the ultimate goals of working with data is applying interpretation and conclusions to real-world problems and scenarios in order to engage in civic practice and enact positive change on the world.

Concept E.3.1

Intent and authorship of analyses

21st-century skills

Media literacy and digital citizenship

Regularly interrogate the point of view of a data author, and transparently share your own.

E.3.1a

Evaluate the source, methodology, sample size, and any potential biases in data collection that may impact the reliability of the data narrative.

E.3.1b

Evaluate the potential agenda(s) or motivation(s) of the author of a data visualization.

E.3.1c

Understand standard journalistic practices, including fact checking and source verification, that support accurate reporting and help combat misinformation.

E.3.1d

Analyze situations when institutions have made big decisions based on untrustworthy data and describe the consequences.

Concept E.3.2

Advocacy with data arguments

Recognize how data can provide evidence for/persuade others toward positive change and how it can benefit society.

E.3.2a

Construct a data story to enact change in your community.

E.3.2b

Analyze data narratives related to social and/or political issues and explore how different presentations of the data could alter its impact on communities and daily life.

Concept E.3.3

Civic data practices

Engage in civic practice and dispositions through recognition of the role data plays in civic society.

E.3.3a

Access open government data from local, state, and/or Federal websites.

E.3.3b

Leverage open government data to supplement or contextualize a data analysis project. e.g., U.S. Census

Concept E.3.4

Impacts of technology use

21st-century skills

AI literacy

Appreciate how AI and other data-driven technology may affect people and resources globally.

E.3.4a

Recognize the environmental cost of running large data centers and AI/ML models while considering the costs versus benefits of nuclear power and evaluating solar and wind options for clean energy.

E.3.4b

Evaluate an impactful data story and its societal implications. e.g., historical heart disease research impacts for men and women