

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

Recognize that multiple types of data can provide valuable insights into the same inquiry.

Concept A.1.2

Data are produced by people

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

21st-century skills



A.1.2a

Explore the origins of some standardized unit measurements (e.g., horsepower, mole, scores on AP exams).

A.1.2b

Identify the risks and tradeoffs of using traditional measurements (e.g., IQ, BMI).

Concept A.1.3

Variability of data

Recognize that variability is a foundational component of data.

A.1.3a

Explore different types of variability for making inferences (e.g., confidence intervals, various tests, classification models).

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

Design and compare alternative data representations, justifying choices to address inherent uncertainty.

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

Identify and label a simple prediction algorithm or equation for a very basic AI prediction model. e.g., a basic AI model's equation looks like: $\text{Prediction} = (\text{weight}_1 \times \text{input}_1) + (\text{weight}_2 \times \text{input}_2)$, such as a college admission model might weight GPA (input_1) and test scores (input_2) to output an acceptance likelihood, and training involves automatically adjusting weights to match historical data.

A.1.5b

Understand that algorithms use cost functions to measure errors and adjust predictions.

A.1.5c

Recognize that some AI tools can be used to explore complex data with many variables.

A.1.5d

Recognize the types of problems that are ideal for using an AI tool to analyze complex data.

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

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.

21st-century skills

 Durable skills

A.2.1a

Recognize that data risk can change based on time, circumstance, and purpose.

A.2.1b


Identify data benefits that can appear well into the future and in unexpected ways.

Concept A.2.2

Biases in data

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

21st-century skills

 Durable skills

A.2.2a

Propose multiple perspectives on data to mitigate inherent biases.

A.2.2b


Understand the difference between implicit and explicit bias.

Concept A.2.3

Power of data

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

21st-century skills

 Media literacy and digital citizenship

A.2.3a

Use data to support arguments, design solutions, or challenge inequities.

A.2.3b

Investigate case studies where data advanced scientific, economic, or social progress.

A.2.3c

Identify when data alone is insufficient and complementary methods are needed. e.g., Data may quantify the number of people affected by a policy, but personal testimonies are needed to illustrate its human impact

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

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.

21st-century skills

 Durable skills

A.3.1a

Conduct independent investigations to inform decisions, leveraging advanced tools and addressing uncertainty.

A.3.1b


Compare investigative approaches across fields to critique strengths and limitations.

Concept A.3.2

Iteration

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

21st-century skills

 Durable skills

A.3.2a


Propose new approaches for leveraging the investigative process to strengthen inferences and arguments.

Concept A.3.3

Dynamic inferences

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

21st-century skills

 Durable skills

A.3.3a

Critically evaluate and update inferences as data scales or methods advance.

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

Interpret data drawn from different fields and topics based on accepted norms within those fields.

A.3.4b


Compare multiple problem-solving approaches, and identify how those differences may compound over time and when repeated.

Concept A.3.5

Student data agency

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

21st-century skills

 Durable skills

A.3.5a

Establish accountability by basing claims and decisions on relevant data.

A.3.5b

Explore career fields and their intersection with data collection, curation, storytelling, and societal impact.

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

Develop comprehensive data validation procedures, including automated checks.

B.1.1b

Implement verification protocols for complex datasets with multiple dependencies.

Concept B.1.2

Organizing and structure

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

B.1.2a

Develop and implement data organization systems that accommodate both structured and unstructured data types.

B.1.2b

Create scalable data organization strategies that maintain data integrity while handling missing values, irregular structures, and evolving data requirements.

B.1.2c

Design and implement metadata documentation systems to track data lineage, transformations, and organizational structures.

Concept B.1.3

Processing and transformation

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

B.1.3a

Use an identifying variable (e.g., index, case ID) to merge two separate datasets that have the same observation, but contain different variables to merge datasets together.

B.1.3b

Use appropriate procedures to join two datasets together that have different observations with the same variables measured.

Concept B.1.4

Summarizing groups

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

B.1.4a

Use datasets with derived variables, based on other variables in the dataset.

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 that address complex systems with multiple interacting variables, including consideration of confounding factors and effect modifiers.

B.2.1b

Design research questions that incorporate multiple levels of analysis and account for both direct and indirect relationships between variables.

B.2.1c


Formulate questions that address the validity and reliability of data collection methods, including considerations of systematic bias and measurement error.

Concept B.2.2

Data creation techniques and methods

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

21st-century skills

 AI literacy

B.2.2a

Design simulations (e.g., using an RNG or computer software) and underlying models to generate data specific to a problem of interest.

B.2.2b

Identify optimal sensors or automated data collection methods for answering a data-based question or designing an experiment.

B.2.2c


Distinguish between surveys, observational studies, and experiments.

Concept B.2.3

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.

21st-century skills

 Durable skills

B.2.3a


Create sophisticated data collection plans that incorporate ethical considerations, cost-benefit analysis, and protocols for ensuring data integrity and reproducibility.

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

Identify (and know you can request access to) non-publicly available datasets by contacting researchers, reading scientific literature, or communicating with public officials.

B.2.4b

Develop strategies for finding and accessing datasets that require special permissions, logins, or formal data requests.

B.2.4c

Evaluate and navigate licensing and citation requirements when using secondary data sources for research.

B.2.4d

Combine multiple secondary datasets to create more comprehensive or useful data for specific 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


Evaluate and critique measurement validity, reliability, and bias in data collection methods, and design comprehensive datafication strategies that address ethical considerations and potential sources of measurement error.

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

Work with data collected over time and consider how to aggregate appropriately.

B.3.2b

Work with data collected over space and consider how to aggregate appropriately.

B.3.2c


Create strategies for dealing with data that is constantly updated.

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

Develop data collection protocols that prevent bias, protect privacy, and ensure ethical representation across diverse populations.

B.3.3b

Apply validation techniques to prevent bias and ensure ethical use of secondary data, including AI tools.

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

Apply advanced data cleaning techniques to handle complex data quality issues such as outliers, inconsistencies, and systematic errors.

B.4.1b

Develop and document reproducible data cleaning workflows that maintain data integrity.

B.4.1c

Evaluate and validate cleaned datasets using statistical methods and domain knowledge.

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

Create and use expected value models to support data-based decision making.

B.4.2b

Work with multiple datasets that combine multiple types of data and combine and transform the different types.

B.4.2c

Work with complex derived variables and understand their calculation methods.

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 very large datasets multiple thousands of observations.

B.4.3b

Use selection, sampling, and transformation tools to navigate very large datasets.

Concept B.4.4

Complexity of structure

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

B.4.4a

Design and implement data structures that can accommodate longitudinal data and multiple levels of aggregation.

B.4.4b

Handle data aggregation across different observation structures and time scales.

B.4.4c

Create flexible organizational systems that can handle both structured and unstructured data sources.

B.4.4d

Develop documentation systems for complex data structures that track relationships and dependencies between variables.

21st-century skills



AI literacy

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

Explore the sensitivity of the mean to outliers compared to the median.

C.1.1b

Discuss instances when to use the mean or median based on the context and data distribution (e.g., skewed vs. symmetric distribution).

Concept C.1.2

Measures of spread

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

C.1.2a

Numerically operationalize the meaning of an "outlier" using standard deviation as a measure of variability and a modified boxplot.

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

Explain how the shape of a distribution influences the relationship between measures of center. e.g., in symmetric distributions - the mean and median are close, in a right-skewed distribution - the mean is greater than the median, in a left-skewed distribution - the mean is less than the median

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

Discuss implications of choices made when generating a frequency table.

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

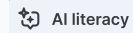
Describe how missing data affects analysis and resulting relationships, patterns, or models.

Concept C.1.6

Metadata

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.

21st-century skills



AI literacy

C.1.6a

Reasonably ideate on some potential modeling approaches when given the metadata (e.g., data and time, text, continuous, geolocation) for a dataset.

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 simulations to investigate associations between two categorical variables and to compare 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

Use variability in distributions to engage in statistical reasoning.

C.2.2b

Understand and interpret variability in sampling distributions and how it impacts population estimates.

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

Conduct linear regression analysis to find the best-fit.

C.2.3b

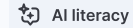
Construct prediction intervals and confidence intervals to determine plausible values of a predicted observation or a population characteristic.

Concept C.2.4

Analyzing non-traditional data

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.

21st-century skills



C.2.4a

Generate a word cloud of a given text after standardizing (e.g., all lower case), stemming, and removing stop words.

Concept C.2.5

Machine learning

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

21st-century skills



C.2.5a

Explore how gradient descent optimizes loss functions and powers machine learning applications like neural networks.

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

Apply statistical or simulation methods to model variability to explore uncertainty in real-world situations.

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

Explore variability through statistical methods, such as analyzing residuals or variance in linear models.

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

Estimate and describe errors between predictions and actual outcomes. e.g., residuals, misclassification rates

C.3.3b

Analyze error patterns to assess model performance. e.g., residual plot, confusion matrix

C.3.3c

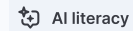
Use insights from error analysis to improve the model. e.g., in linear regression, add a variable or use a curve; in classification, balance the groups or adjust the cutoff

Concept C.3.4

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.

21st-century skills



C.3.4a

Appreciate that many AI tools are pre-trained with large quantities of data so that inferences can be drawn on smaller sample sizes.

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

Create models to perform simulations using a digital tool.

C.4.1b

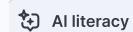
Perform data analysis using a digital tool.

Concept C.4.2

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.

21st-century skills



C.4.2a

Critique the societal effect of AI by exploring issues surrounding bias, accountability, and transparency in decision-making using AI tools, as well as the effects on privacy, jobs, and policy.

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 differences between a no-code, low-code, or high-code 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 multiple digital tools suited for different tasks throughout the data investigation process.

C.4.4b

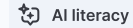
Describe how digital tools are used in the workforce.

Concept C.4.5

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.

21st-century skills



AI literacy

C.4.5a

Recognize how computer code can be used to produce reproducible data analysis processes.

C.4.5b

Recognize the advantages and limitations of using computer code compared to no-code or low-code tools.

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

Design data visualizations that include accessible features such alt-text and text descriptions.

C.4.6b

Examine how policies, limitations, and technological advancements impact the development of accessible digital tools.

Substrand C5

Models of Data

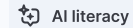
Interpreting, creating, and using models is a central component of working with data. Models are both a way to analyze data and a source of data.

Concept C.5.1

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.

21st-century skills



AI literacy

C.5.1a

Discern that different models, such as decision trees and neural networks, analyze patterns and relationships in data to make predictions.

C.5.1b

Assess relationships in the context of uncertainty, bias, and reliability of the data.

C.5.1c

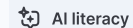
Investigate how assumptions and bias influence a model's results.

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

Develop models that incorporate multiple variables and explicitly consider interactions between them.

C.5.2b

Use computational methods, coding, or machine learning techniques to build and refine models.

C.5.2c

Assess assumptions, limitations, and biases in models to evaluate their impact on predictions in real-world scenarios.

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 the result or finding and indicate the level of certainty regarding the statistical analysis and the quality of the evidence (e.g., dataset or source characteristics, similar findings in alternative data) as justification.


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

Summarize previous assumptions and potential updates in written conclusions from a data analysis, and identify any known contradictory findings to mitigate confirmation bias.

D.1.2b

Describe Bayes Theorem by explaining how it relates to conditional probability, which includes the probability of an event occurring, the probability of that event given the evidence is true, and the probability that the evidence itself is true.

D.1.2c

Apply the logic of Bayes Theorem to determine whether a data-based claim in the media was accurately explained.

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

Justify the Expected Value equation ($EV = P(X_i) \cdot X_i$) with formal probability statements and by explaining the Law of Large numbers.


D.1.3b

Apply the Expected Value equation to assess its fitness for the problem by determining the accuracy of the estimate based on the number of trials conducted. e.g., flipping a coin 100 times and determining if getting heads 30 times is reasonable when the expected value is getting heads 50 times

Concept D.1.4

21st-century skills

Explaining significance

 Media literacy and digital citizenship

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.

D.1.4a

Explain the concept of statistical significance (e.g., including its role in distinguishing meaningful results from random chance) in plain language and the limitations of significance testing (e.g., inability to address study design flaws, confounding variables, or real-world validity beyond a narrow model comparison).

D.1.4b

Describe how statistical significance tests are constructed, calculated, and interpreted in the context of chosen probability models and/or assumptions.

D.1.4c

Identify real-world instances where assessing statistical significance is crucial (e.g., scientific studies to distinguish actual effects from random variation) while also evaluating the significance claims made by others and recognizing situations where statistical significance is necessary but not sufficient for proving a point.

D.1.4d

Differentiate statistical significance, effect size, and statistical power in simple terms with real-world examples, explaining how each addresses distinct questions in research. e.g., whether outcomes could be connected to random chance, the meaningfulness of impacts in context, the suitability of the analysis approach to the specific data and problems

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 simulation-based inferential methods at large N to draw conclusions from a dataset using digital software.

D.1.5b

Identify why simulation can be used to infer conclusions about a population referencing the Law of Large Numbers.


D.1.5c

Interpret margin of error and confidence intervals for a given sample.

Concept D.1.6

21st-century skills

Correlation versus causation

 Media literacy and digital citizenship

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

D.1.6a

Independently identify examples of two dependent variables that are both influenced by a third variable in real-world data. e.g., coffee consumption and lower risk of disease are both affected by an active lifestyle

D.1.6b

Identify spurious correlations in the media and analyze how they relate to media claims and AI recommendations. e.g., ice cream sales and shark attacks both increase in the summer; they're both linked to hot weather, not each other

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

Recognize that randomization can happen in various settings, regardless of the intervention or events involved. e.g., artificial interventions, accidental or chance events, unrelated to the question of interest

D.1.7b


Differentiate between lab experiments and natural experiments in scenario-based questions.

Concept D.1.8

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.

21st-century skills

 Durable skills

D.1.8a

Use computer software to analyze the relationship between an independent and dependent variable in a linear model by changing the number and combination of dependent variables.

D.1.8b

Evaluate how changes to the number and combination of dependent variables affect the model by interpreting R-squared and regression coefficients.

D.1.8c

Explore how polynomials of different degrees fit scatterplots.

D.1.8d

Analyze how increasing or decreasing the degree of a polynomial can lead to potential overfitting or underfitting the data.

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

Develop a causal diagram to map relationships among multiple variables and create an iterative analysis plan to test each relationship with data.

D.2.1b

For query-based questions, estimate a confidence interval and margin of error in a real-world data analysis project with software.

D.2.1c


For hypothesis-based questions, estimate a p-value based on a proposed statistical model for real-world data with software.

Concept D.2.2

Iteration, validation, and multiple explanations

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

21st-century skills

 Durable skills

D.2.2a

Highlight unusual associations or outcomes in an analysis document by validating analysis steps and investigating other parts of the dataset.

D.2.2b

Identify potential counter-arguments or alternative explanations that may refute one's conclusions drawn from data, and suggest mitigation strategies that could be tried in the future with additional data or new research.

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


Evaluate the potential limitations of statistical findings by considering the data collection methods, sample selection, and simplifications that may not capture the complexity of real-world scenarios.

Concept D.2.4

Relevant conclusions

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

21st-century skills

 Durable skills

D.2.4a

Determine if a causal claim can be established based on the investigation's design (e.g., natural experiments, real-world observations) and describe the differences between expectations and the design.

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

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

21st-century skills

 Media literacy and digital citizenship

 AI literacy

D.3.1a

Analyze a data generalization issue in media or real-world situations and discuss its significant impacts and the importance of addressing generalization errors.

D.3.1b

Implement multiple strategies to generalize data-based conclusions to new populations or situations. e.g., add additional context or control variables, repeat the analysis with new collection or sample, test a model with a different dataset

D.3.1c

Evaluate the advantages and disadvantages of automated tools that rely on large datasets for universal predictions. e.g., prediction algorithm for airline ticket prices or home mortgage application assessment, AI model for facial recognition, autonomous vehicle model trained on city roads

Concept D.3.2

Sample versus population

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

21st-century skills

 AI literacy

D.3.2a

Evaluate the suitability of different sampling methods (e.g., random sample with or without replacement) for the specific question and available data.

D.3.2b


Identify situations in which data on the full population is easily available or even critical to answer a question of interest, and traditional sampling-methods are not required.

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

Make an informal power analysis for an analysis or experimental setup using real-world data and a hypothesis, including claims about the 1) Effect Size 2) Sample Size 3) Statistical Significance and 4) Statistical Power.

D.3.3b


Use the simple equation $\text{Power} = 1 - \beta$ to visually show the difference between a normal distribution of outcomes and an abnormal distribution of outcomes.

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


Propose and implement at least two methods to mitigate sample bias in a real-world dataset. e.g., adding additional data, making a new variable with a correction, explicitly stated assumption

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 and implement at least two strategies in a project-based activity that utilize original data to address questions in a new scenario.

D.3.5b


Describe potential ethical and statistical issues with the extension strategies, including explicit caveats on any conclusions reached with real-world data.

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

Identify and explain Simpson's Paradox: an average trend may disappear or even reverse when individual subsets and/or groupings are examined.

D.3.6b


Review examples of Simpson's Paradox in the media and in well-known research studies.


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 the importance of many trials, study validation, and meta-analyses in academic research.

D.3.7b

Document data analysis steps in a shareable and reproducible format for collaboration platforms (e.g., GitHub, Bitbucket).

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

Create data visualizations that illustrate complex bivariate relationships, e.g., exponential, quadratic

E.1.1b


Edit data visualizations to optimize it for your intended audience and the audience's different needs. e.g., "chartjunk" can be distracting for some audience but necessary for others

Concept E.1.2

Investigate with visualizations

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

21st-century skills

 Durable skills

E.1.2a

Create data visualizations of raw data and increasingly aggregated forms of the same data to help understand the nuances of the data.

E.1.2b


Strategically use data visualization to identify potential outliers, errors, and unexpected findings, while clearly stating and justifying any reasons for excluding certain potentially erroneous observations.

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

Provide context for the data to help viewers understand the background and implications.

E.1.3b

Recognize how color theory (e.g., tint, saturation, shading) can be used to represent continuously scaled data (e.g., darker color = higher concentration of occurrence).

E.1.3c


Recognize that we have culturally-influenced or domain-specific ways of using and interpreting chart elements. Consider the conventions that are known to or expected by your audience when developing data visualizations.

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

Understand how uncertainty around point and effect estimates are communicated on data visualizations with error bars.

E.1.4b


Evaluate the effectiveness of data visualizations, including the risk of misleading the reader.

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 various representations of relative frequencies and proportions, identify elements of each representation that facilitate or hinder the identification of relative proportions, and explain the reasoning behind conventions. e.g., ordered or unordered stacked bar graph

E.1.5b

Compare and/or contrast various ways to represent distributions and their measures of center (e.g., histograms, density plots, box plots) by plotting two distributions on the same graph and explaining how different representations facilitate or hinder the visibility of differences and associations.

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

Produce a data visualization parallel to the type of data (e.g., numeric, categorical, string, image, unstructured).

E.1.6b

Defend your visualization choice to others and explain the data type and visualization type including suitability for continuous or discrete variables.

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 a real-world argument or broader explanation of social, economic, scientific, or political factors.

Concept E.2.2

Write data stories

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

21st-century skills

 Durable skills

E.2.2a

Make and defend arguments using key features from a data visualization.

E.2.2b

Clearly define the claim by making it specific, measurable, and actionable.

E.2.2c

Ensure the data directly addresses the claim being defended.

E.2.2d

Address potential confounding variables and factors in claim-making, and if possible, demonstrate how the data controls for those confounding variables and factors.

E.2.2e


Discuss a claim's broader implications in writing, including societal effects. e.g., a graph showing declining crime might ignore rising cybercrime

Concept E.2.3

Adapt storytelling

Tailor storytelling for different audiences.

21st-century skills

 Durable skills

E.2.3a

Write data analyses and stories using plain-language vocabulary along with relevant problem-specific terms, ensuring adaptability to various audiences, both technical and non-technical, with clear explanations of why the content is important for each audience.

E.2.3b

Provide multiple representations of data relevant to individual arguments. e.g., visualizations, summary statistics, and descriptions of processes or methodologies

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

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

21st-century skills

 Media literacy and digital citizenship

E.3.1a

Communicate and present the source of the data used for the data visualization to ensure transparency.

E.3.1b

Examine the significance of the data being visualized by understanding what it measures and its relevance to real-world issues or scenarios.

E.3.1c

Examine how institutions (e.g., government, businesses, nonprofit organizations) utilize big data to achieve policy goals while considering the benefits and harms to the public and their implications for civic behavior.

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

Explain how data science connects to other disciplines to solve major problems around the globe.

E.3.2b

Discuss strategies to mitigate harmful predictions derived from a data story, such as the varying injury rates from crash test dummies among different groups of drivers.

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

Develop democratic dispositions through evaluation of local data. e.g., review local election data, housing data in local city or county

E.3.3b


Pick a local issue of student interest and based on a data analysis project, submit a Public Comment.

Concept E.3.4

Impacts of technology use

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

21st-century skills

 AI literacy

E.3.4a

Consider the environmental and human costs of harvesting natural resources for the creation of modern technologies. e.g., mining of lithium, geopolitical issues with high precision silicon