Virginia Mathematics

Data Science

Data in Context

DS.1 The student will identify specific examples of real-world problems that can be effectively addressed using data science. DS.1

- a Identify and explain characteristics that best lend themselves to a data driven approach to problem solving. DS.1.A
- b Formulate questions based on context. DS.1.B
- c Understand the type of data relevant to the context of the question at hand. DS.1.C
- d Define relationships between variables and constant relationships. DS.1.D
- e Create a hypothesis of interest in terms of measurable data. DS.1.E
- f Define the stages of the data cycle and how each stage is related to the other. DS.1.F
- g Identify and explain constraints of the data-driven approach. DS.1.G

DS.2 **The student will be able to formulate a top-down plan for data collection and analysis, with quantifiable results, based on the context of a problem.** DS.2

- a Design a data project plan, which is aligned with the data science cycle, that includes the following components: DS.2.A
  - i definition of the goal of the project as it pertains to a real-world problem; DS.2.A.I
  - ii identification of the various parameters of the problem and stakeholders; DS.2.A.II
  - iii a timeline for the project with deliverables; DS.2.A.III
  - iv Key Performance Indicators (KPI) for the successful data project deliverables; DS.2.A.IV
  - v resource needs and tools for the project; DS.2.A.V
  - vi bias considerations for the sampling process of the project; and DS.2.A.VI
  - vii limitations of the project. DS.2.A.VII
- b Given the context and parameters of a problem, choose from among various sampling techniques, which may include DS.2.B
  - i simple random; DS.2.B.I
  - ii systematic; DS.2.B.II
  - iii stratified; and DS.2.B.III
  - iv cluster; DS.2.B.IV

Data Bias

DS.3 **The student will recognize the importance of data literacy and develop an awareness of how the analysis of data can be used in problem solving to effect change and create innovative solutions.** DS.3

- a Formulate relevant/clarifying questions to identify potential data biases presented in existing analyses/visualizations. DS.3.A
- b Effectively read data summaries and visualizations and explain/translate into nontechnical terms in proper context. DS.3.B
- c Identify potential data biases in terms of data presented and discuss the potential effects of such biases in terms of how they could affect data analysis and decision making. DS.3.C
- d Identify privacy and consumer protection issues that might be a result of how data is presented. DS.3.D
- e Describe the types of data that business, industry, and government entities collect and possible ways the data is used. DS.3.E

DS.4 **The student will be able to identify data biases in the data collection process and understand the implications and privacy issues surrounding data collection and processing.** DS.4

- a Identify data biases in the data collection process that include, but are not limited to, confirmation, selection, outliers, overfitting / under fitting, and confounding and describe mitigation strategies for these biases. DS.4.A
- b Provide examples of sampling biases in terms of data collection and the potential effects. DS.4.B
- c Identify and describe data biases as a producer as well as a consumer/decision maker of data. DS.4.C
- d Describe how the data collection process should be focused, relevant, and limited to the scope of the data project plan. DS.4.D
- e Describe privacy considerations in the collection of data as both a consumer and producer. DS.4.E

DS.5 **The student will use storytelling as a strategy to effectively communicate with data.** DS.5

- a Define storytelling and explain the importance of storytelling as a strategy to communicate the idea behind and results of a data science project effectively. DS.5.A
- b Explain the steps involved in data storytelling and how it relates to the data cycle. DS.5.B
- c Effectively identify a story worth telling based on the data (looking for trends, correlations, outliers) and by asking a question or forming a hypothesis based on insight and audience. DS.5.C
- d Effectively select visualizations that simplify the information, highlight the most important data, and communicate key points quickly. DS.5.D
- e Effectively simplify the information presented to make it more concise and focus the audience's attention on the key parameters that support the student's hypothesis. DS.5.E
- f Effectively form a narrative based on data available to provide context, insight, and interpretation to make the analysis more relevant to a given audience. DS.5.F
- g Explain how data storytelling should include complete and accurate information, and consistent visuals for effective communication. DS.5.G

DS.6 **The student will justify the design, use, and effectiveness of different forms of data visualizations.** DS.6

- a Conduct exploratory data analysis using visualization. DS.6.A
- b Formulate questions from exploration of a data set to consider how data will communicate a story. DS.6.B
- c Determine the effectiveness of different data visualization choices based on the data context from conventional statistical charts to unconventional/emerging data visualizations to more complex visualizations. DS.6.C
- d Create a visualization of a data set and summarize the representation using the context of the data. DS.6.D
- e Compare two or more different representations to ensure the design communicates the features and behavior of data sets. DS.6.E
- f Justify design choices (based on data set type, size, context, and audience) of data visualizations to highlight important features, trends, and insights. DS.6.F

Data Modeling

DS.7 **The student will be able to assess reliability of source data in preparation for mathematical modeling.** DS.7

- a Explain why determining the reliability of big data sources is a key skill that data scientists use to build data trust across an organization. DS.7.A
- b Describe the difference between reliability of a data source compared to statistical reliability and validity in research analysis. Assess processing source data for reliability based on validity, completeness, and uniqueness. DS.7.B

8 **The student will be able to acquire and prepare big data sets for modeling and analysis.** DS.8

- a Explain the pros and cons of collecting data versus acquiring it from existing sources. DS.8.A
- b Apply matrix operations using algebraic methods (with the support of technology tools) to: DS.8.B
  - i wrangle the data (sort, select, filter, and replace); DS.8.B.I
  - ii clean the data; DS.8.B.II
  - iii format and enrich the data; and DS.8.B.III
  - iv combine and store the data. DS.8.B.IV
- c Read data from different sources for preparation and analysis. DS.8.C
- d Identify important parameters about a big data set based on the context of data collected/acquired. DS.8.D
- e Define and document the process of ingesting, formatting, and cleaning data for future decision making by: DS.8.E
  - i making data more easily understood by a wider audience; and DS.8.E.I
  - ii connecting data with existing contextual data. DS.8.E.II

DS.9 **The student will select and analyze data models to make predictions, while assessing accuracy and sources of uncertainty.** DS.9

- a Identify factors that contribute to the overall behavior of a data set (e.g., true values, bias, and noise). DS.9.A
- b Fit models based on the behavior of the data, (e.g., models of univariate and bivariate data), in order to make predictions. DS.9.B
- c Distinguish between linear and nonlinear associations between variables using visualizations. DS.9.C
- d Identify models that are overly complex and therefore fitting to random noise which decreases their predictive accuracy. DS.9.D
- e Use regression techniques to perform selection of optimal features. DS.9.E
- f Recognize the potential implications of removing features. DS.9.F
- g Select the optimal model for a data set from among a large collection of models, using technological tools. DS.9.G

DS.10 **The student will be able to summarize and interpret data represented in both conventional and emerging visualizations.** DS.10

- a Apply descriptive statistics to explain measures of central tendency and measures of variability/dispersion to describe center and spread in visualizations of distributions. DS.10.A
- b Define emerging visualizations and describe summarization of characteristics and relationships based on audience and purpose which may include: DS.10.B
  - i a heat map, which uses color to show changes and magnitude of a third variable to a twodimensional plot; and DS.10.B.I
  - ii a bubble chart, which is a multivariate graph that is both a scatterplot and a proportional area chart. Typically, each plotted point then represents a third variable by the area of its circle. DS.10.B.II
- c Interpret various emerging visualizations by describing patterns, trends, and relationships between and among the variables. DS.10.C

DS.11 **The student will select statistical models and use goodness of fit testing to extract actionable knowledge directly from data.** DS.11

- a Calculate the theoretical probability of random events and compare them to the observed frequencies. DS.11.A
- b Describe the normal curve determined by the mean and standard deviation of a univariate data set. DS.11.B
- c Fit nonlinear models to data sets and use these models to predict unobserved data values. DS.11.C
- d Select pairs of variables that identify meaningful clusters of data. DS.11.D
- e Select an appropriate statistical distribution and test its goodness of fit based on the context of the data being analyzed. Statistical distributions may include, but are not limited to DS.11.E
  - i Normal; DS.11.E.I
  - ii Binomial; and DS.11.E.II
  - iii Poisson. DS.11.E.III

Data and Computing

DS.12 **The student will be able to select and utilize appropriate technological tools and functions within those tools to process and prepare data for analysis.** DS.12

- a Utilize technology tools to be able to access data effectively from multiple sources (e.g., tables, column separated values, spreadsheets, documents, databases). DS.12.A
- b Utilize tools and functions (in tools) to effectively explore the data for issues and errors before beginning to process it. DS.12.B
- c Define the (tools and technological) process to optimally ingest data and to export data after processing. DS.12.C
- d Utilize tools and their functions to clean and validate data by: DS.12.D
  - i removing data that are incomplete, incorrect, or duplicated; DS.12.D.I
  - ii removing extraneous data or outliers; and DS.12.D.II
  - iii standardizing data to conform to contextual norms (e.g., privacy, sensitive data). DS.12.D.III
- e Utilize tools and their functions to combine and store data by: DS.12.E
  - i merging multiple data sets for efficiency purposes; and DS.12.E.I
  - ii optimizing storage of data based on volume, velocity, and variety. DS.12.E.II
- f Utilize tools to format and store the data appropriately to allow for effective analysis. DS.12.4

DS.13 **The student will be able to select and utilize appropriate technological tools and functions within those tools to analyze and communicate data effectively.** DS.13

- a Select and utilize technology tools to effectively generate conventional and unconventional visualizations of data to explore patterns and/or analyze a large data set. DS.13.A
- b Utilize specific functions in technology tools to perform descriptive and inferential statistical analysis. DS.13.B
- c Utilize coding to store and extract data more effectively for data analysis. DS.13.C
- d Select and apply features of technology tools effectively to organize, summarize and gain insight from data. DS.13.D
- e Select the appropriate visualization based on context and audience and create it using technology tools to effectively communicate an idea. DS.13.E