2023 Mathematics *Standards of Learning*
Understanding the Standards – Data Science

The following standards outline the content of a one-year course in Data Science. If a one-semester course is desired, the standards with a dagger (†) would apply. The Data Science *Standards of Learning* provide an introduction to the learning principles associated with analyzing big data.

Through the use of open-source technology tools, students will identify and explore problems that involve the use of relational database concepts and data-intensive computing to find solutions and make generalizations. Students will engage in a data science problem-solving structure to interact with large data sets as a means to formulate problems, collect and clean data, visualize data, model using data, and communicate effectively about data formulated solutions.

Data in Context

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

Students will demonstrate the following Knowledge and Skills:

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

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

Additional Content Background and Instructional Guidance:

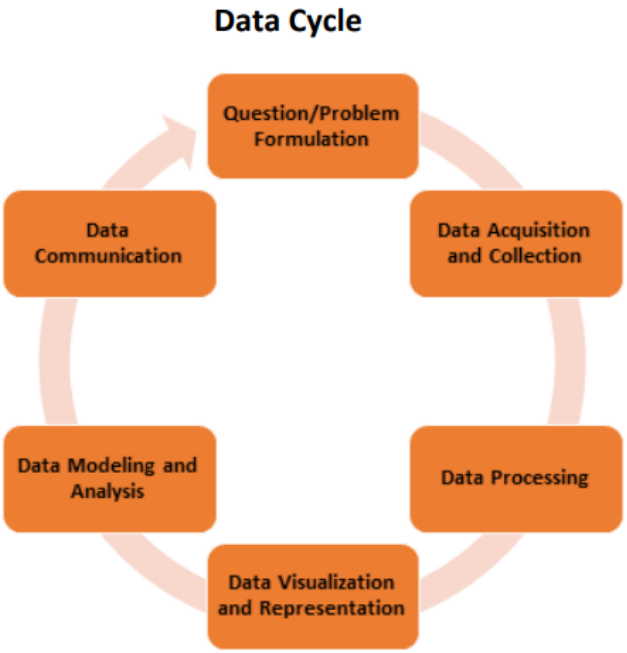
- There are characteristics of problems in the real-world that best lend themselves to be analyzed using the data cycle.
- Solutions addressed by Data Science include conjectures that can be supported or refuted by measurements or observations.
- The iterative stages of the data cycle include:
  - Question/Problem Formulation – Identify the driving question for the problem being solved.
  - Data Acquisition & Collection – Collect and clean data to assist with multiple ways to solve a problem.
  - Data Processing – Manipulate data to make it usable through a predetermined process.
  - Data Visualization & Representation – Connect visual representations to brainstorm solutions.

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

Additional Content Background and Instructional Guidance:

- Data Modeling & Analysis – Build a prototype of a model, test, and iterate.
- Data Communication – Effectively communicate data driven solution based on context and audience.

- The data science cycle is an iterative process.



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.

Students will demonstrate the following Knowledge and Skills:

a) Design a data project plan, which is aligned with the data science cycle, that includes the following components:
   i) definition of the goal of the project as it pertains to a real-world problem;
   ii) identification of the various parameters of the problem and stakeholders;
   iii) a timeline for the project with deliverables;
   iv) Key Performance Indicators (KPI) for the successful data project deliverables;
   v) resource needs and tools for the project;
   vi) bias considerations for the sampling process of the project; and
   vii) limitations of the project.

b) Given the context and parameters of a problem, choose from among various sampling techniques, which may include
   i) simple random;
   ii) systematic;
   iii) stratified; and
   iv) cluster;
   to justify the sampling methodology of the project design and implementation.

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.

Additional Content Background and Instructional Guidance:

- A data project plan ensures effective communication and agreement at all phases of the data science project.
- A data project plan allows effective execution on time and under budget.
- A data project plan allows us to understand the tools, resources and architecture needed to ensure a successful project.
- Project deliverables are the things you create to help you fulfill the objective while KPI stands for key performance indicator, a quantifiable measure of success of the project as a whole.
- Sampling bias in the data collection process include, but are not limited to, confirmation, selection, and outliers.
- Sampling must be purposeful to infer trends and characteristics in the data being collected. Nonrandom sampling techniques, such as convenience, quota, judgment, and snowball, may result in a non-representative sample that does not produce generalizable results.

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.

Students will demonstrate the following Knowledge and Skills:

a) Formulate relevant/clarifying questions to identify potential data biases presented in existing analyses/visualizations.

b) Effectively read data summaries and visualizations and explain/translate into nontechnical terms in proper context.

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.

d) Identify privacy and consumer protection issues that might be a result of how data is presented.

e) Describe the types of data that business, industry, and government entities collect and possible ways the data is used.

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. *Additional Content Background and Instructional Guidance:*
• Data literacy is the ability to read data, work with data and communicate about data by putting it in proper context and asking relevant/clarifying questions to determine/identify data bias. • Data literacy helps to recognize, sort and filter through data biases that leads to improved decision making in data collection and reporting. • Data privacy and consumer protection are important issues that affect individuals and organizations. • Historical instances of government and private data breaches provide examples of the considerations of privacy in data. • Data bias occurs when data does not include variables that properly capture the phenomenon we want to predict.

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.

Students will demonstrate the following Knowledge and Skills:

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.
b) Provide examples of sampling biases in terms of data collection and the potential effects.
c) Identify and describe data biases as a producer as well as a consumer/decision maker of data.
d) Describe how the data collection process should be focused, relevant, and limited to the scope of the data project plan.
e) Describe privacy considerations in the collection of data as both a consumer and producer.

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. *Additional Content Background and Instructional Guidance:*

• Various implications can result from the types of data collection methods used.
• Privacy and consumer protection are considerations when data are collected.
• There are producers, publishers, consumers, and decision makers of data:
  ○ producer of data: data are obtained through some source- open source, sensor equipment, third party organization/source, external source;
  ○ publisher of data: entity that acquires, manages, stores, makes available the data;
  ○ consumer of data: develops products/applications to support the decision making; and
  ○ decision maker of data: uses the products/applications to make decisions.

Data and Communication

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

Students will demonstrate the following Knowledge and Skills:

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.
b) Explain the steps involved in data storytelling and how it relates to the data cycle.
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.
d) Effectively select visualizations that simplify the information, highlight the most important data, and communicate key points quickly.
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.
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.
g) Explain how data storytelling should include complete and accurate information, and consistent visuals for effective communication.

DS.5† The student will use storytelling as a strategy to effectively communicate with data. *Additional Content Background and Instructional Guidance:*

• Storytelling with data involves combining context, visualizations, and a narrative to communicate the idea behind a data science project effectively. Narrative, which is the crux of storytelling, is the way we simplify and make sense of complex data by supplying context, insight, and interpretation to make the analysis more applicable and relevant.
• Communicating with data using storytelling involves concrete steps:
  ○ understanding context;
  ○ selecting a visual;
  ○ eliminating clutter;
  ○ focus attention; and
  ○ telling a story.
• Data storytelling requires accuracy in presenting information and critical thinking in consuming information to make conclusions.

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

Students will demonstrate the following Knowledge and Skills:

a) Conduct exploratory data analysis using visualization.
b) Formulate questions from exploration of a data set to consider how data will communicate a story.
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.
d) Create a visualization of a data set and summarize the representation using the context of the data.
e) Compare two or more different representations to ensure the design communicates the features and behavior of data sets.
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† The student will justify the design, use, and effectiveness of different forms of data visualizations.

Additional Content Background and Instructional Guidance:

- The goal of data visualization is to distill large datasets into visual graphics to allow for easy understanding of complex relationships within data.
- Computer-based visualization systems provide visual representations of data sets designed to help end users to carry out tasks more effectively. Data visualization includes analysis, design, and construction.
- Task questions may include: What questions does the user want to answer? What problem is to be solved? Which decisions is the user trying to make? What outcomes are desired? What story does the user want to tell? What tasks should the user perform?
- Choosing a visualization based on data type and the message communicated reveals trends so the audience can easily understand the significance of the findings from the data set.
- Data set types in visualizations include but are not limited to tabular; network; spatial; and textual. Tabular data may be represented in two-dimensional (row by column) or multidimensional tables. Networks may include nodes and links and trees. Spatial data sets may be categorized as continuous fields as in grids of position and geometric such as in maps.
- Inputs for visualizations include data set types and tasks. Data attributes may be categorical, ordinal, or quantitative with special cases for time and space.
- Data visualizations may include both conventional and emerging types based on function in the context of the data.
- Data insights from visualizations can be shared in different ways including live or virtual presentations; dashboards; embedded into applications; and/or broadcast to audiences through data-driven alerts or communications.

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

Additional Content Background and Instructional Guidance:

- The choice of a suitable technological tool allows students to create and compare multiple visualizations of the same data set.
- Connections can be made among summary information from statistical analysis to visualizations of the same data set.
- Numerous forms of data visualizations exist and are often chosen based on the intended function of the visualization.
- Chart Selection for Data Visualization by Function:

Chart Type	Comparisons	Proportions	Relationships	Hierarchy	Location	Distribution	Patterns	Range	Data Over Time	Analyzing Text	Movement Flow	Financial	Uncertainty Error
Area or Stacked Area Graph/Plot	X						X		X				
Area Bands													X
Bar or Stacked Bar Graph	X	X					X						
Box and Whisker Plot	X					X	X	X					
Bubble Chart/Map	X	X	X		X	X	X		X				
Candlestick Chart								X	X			X	
Chord Diagram			X										
Choropleth Map					X								
Circle Packing		X		X									
Confidence Strips													X
Connection Map			X			X					X		
Data Over Geographical Region					X								
Density Chart/ Plot						X	X						
Donut Chart		X											
Dot Map					X	X	X						
Dot Matrix		X				X							
Error Bars													X
Flow Map					X	X					X		
Gantt Chart							X	X					
Heat Map			X						X				
Histogram	X					X	X	X	X				
Kagi Chart												X	
Line Graph	X					X	X		X				
Marimekko Chart			X										
Multivariable Bar Chart						X	X						
Parallel Sets											X		

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

Additional Content Background and Instructional Guidance:

Chart Type	Comparisons	Proportions	Relationships	Hierarchy	Location	Distribution	Patterns	Range	Data Over Time	Analyzing Text	Movement Flow	Financial	Uncertainty Error
Pie Chart		X											
Population Pyramid						X	X						
Renko Chart												X	
Sankey Diagram											X		
Scatterplot			X			X	X						
Span Chart								X					
Spiral Plot									X				
Stream Graph									X				
Sunburst				X									
Tree Diagram/ Map		X	X	X									
Two-Way Tables	X												
Venn Diagram			X										
Violin Chart								X					
Waterfall Chart												X	
Word Cloud		X								X			

Data Modeling

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

Students will demonstrate the following Knowledge and Skills:

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.

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 The student will be able to assess reliability of source data in preparation for mathematical modeling. *Additional Content Background and Instructional Guidance:*

• Understanding the characteristics of a reliable data source will allow for more effective analysis.
• Data validation or input validation is a method for checking the accuracy and quality of source data, typically performed prior to importing and processing so that data analysis results are accurate.
• There are different aspects of data reliability:
  ○ data can be considered valid when it is formatted and stored in a consistent structure;
  ○ data is complete when it includes all values required by the context; and
  ○ data is unique if it is free from duplicates and extraneous entries.

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

Students will demonstrate the following Knowledge and Skills:

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

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

Additional Content Background and Instructional Guidance:

- Data can be collected or acquired from reliable existing data sources.
- The purpose of sampling is to provide sufficient information so that population characteristics may be inferred.
- Data preparation supports identifying errors before processing.
- Cleaning and reformatting data sets ensures that all data used in analysis will be high quality.
- Higher quality data can be processed and analyzed more quickly and efficiently.
- The process involved in preparing the data set for modeling and analysis involves one or more of the following sub-steps –
    - ingest/wrangle the data, which includes:
        - sort (arrange) - order rows by the value or characters of a variable, or a selection of them;
        - select - choose columns in a dataset based on a defined criterion;
        - filter - remove parts of rows of a dataset during analysis; and,
        - replace - convert specific characters (e.g., convert numerical characters to data and time formats) or re-code variables to fit models;
    - clean the data;
    - format and enrich the data; and,
    - combine and store the data.

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

Students will demonstrate the following Knowledge and Skills:

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

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

Additional Content Background and Instructional Guidance:

- Data prediction involves extrapolating the data beyond the current data set and providing confidence values for those estimates.
- It is important to be able to distinguish between the "noise" in the data and relevant data. Every measurement is composed of true value, bias, and random noise. This noise is the source of uncertainty.
- Mathematical models will be used to make data predictions based on the behavior of the data.
- Data prediction may be limited by the assumption that historical patterns are a good predictor of future outcomes.
- Overfitting the data can lead to inaccurate results.
- Considerations based on data bias need to be taken into account during feature selection when trying to predict future outcomes.
- The fundamentals of numerical methods, allow for further understanding of the application, limitations, and pitfalls of the model.

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

Students will demonstrate the following Knowledge and Skills:

a) Apply descriptive statistics to explain measures of central tendency and measures of variability/dispersion to describe center and spread in visualizations of distributions.

b) Define emerging visualizations and describe summarization of characteristics and relationships based on audience and purpose which may include:

   i) a heat map, which uses color to show changes and magnitude of a third variable to a two-dimensional plot; and

   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.

c) Interpret various emerging visualizations by describing patterns, trends, and relationships between and among the variables.

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

Additional Content Background and Instructional Guidance:

- Characteristics of data sets can be summarized graphically by using visual representations of the distribution and numerically with measures of central tendency and measures of variation or dispersion.
- Descriptive statistics summarize the characteristics of a data set.
- Statistical summaries have the potential to lose information. Representing all the data through visualizations is important to confirm expected patterns, find unexpected patterns, and to assess the validity of the selected statistical model.
- Visualizations are a key to validating underlying assumptions such as data being normally distributed and having no correlation between independent variables.
- Selected charts for data visualization based on types and number of variables are –

	Univariate	Bivariate	Three Variables or Higher
Quantitative	Dot plots Stem plots Histograms Box and Whisker Plots	Scatterplots Line Plots 2-D Histograms	3-D Scatterplot 3-D Line plot Heat Map Bubble Chart
Categorical	Bar Charts Pie Charts	Two-Way Tables Segmented Bar Graphs	Multivariate Bar Graphs

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

Students will demonstrate the following Knowledge and Skills:

a) Calculate the theoretical probability of random events and compare them to the observed frequencies.
b) Describe the normal curve determined by the mean and standard deviation of a univariate data set.
c) Fit nonlinear models to data sets and use these models to predict unobserved data values.
d) Select pairs of variables that identify meaningful clusters of data.
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
    i) Normal;
    ii) Binomial; and
    iii) Poisson.

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

Additional Content Background and Instructional Guidance:

- There are key differences between observed and theoretical probabilities.
- The different types of distribution of data vary according to the context and are important to predict future outcomes
- While causation and correlation can exist at the same time, correlation does not imply causation.
- Categorical variables can also be analyzed using specific tests.
- Technology tools can be used to identify meaningful clusters of data and associated sets of data points. Methods like clustering can be used to identify meaningful relationships between data observations in the form of similarities. When visualizing clustering methods, these similarities show up as “closeness” between plotted data points or the tendency of similar points to group together.
- It is important to have a toolbox of different statistical models for modeling a variety of phenomena (e.g., Binomial, Poisson, exponential).
- Histogram comparisons, Chi-squared tests, and other methods are used to test goodness of fit.

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.

Students will demonstrate the following Knowledge and Skills:

a) Utilize technology tools to be able to access data effectively from multiple sources (e.g., tables, column separated values, spreadsheets, documents, databases).

b) Utilize tools and functions (in tools) to effectively explore the data for issues and errors before beginning to process it.

c) Define the (tools and technological) process to optimally ingest data and to export data after processing.

d) Utilize tools and their functions to clean and validate data by:
   i) removing data that are incomplete, incorrect, or duplicated;
   ii) removing extraneous data or outliers; and
   iii) standardizing data to conform to contextual norms (e.g., privacy, sensitive data).

e) Utilize tools and their functions to combine and store data by:
   i) merging multiple data sets for efficiency purposes; and
   ii) optimizing storage of data based on volume, velocity, and variety.

f) Utilize tools to format and store the data appropriately to allow for effective analysis.

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.
Additional Content Background and Instructional Guidance:
• Data can be imported, processed, and exported (if necessary) using technology tools. • Organizing data using technology tools aids in exploration. • Technology tools can be used to address missing entries, errors, or duplicates in the data. • The process of decision making that occurs during the importing or extracting, processing, cleaning, and formatting of data uses a choice of tools: technological applications, coding, and web. • The technology procedure for data preprocessing is clearly explained and documented for future replication and decision making.

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.

Students will demonstrate the following Knowledge and Skills:

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.

b) Utilize specific functions in technology tools to perform descriptive and inferential statistical analysis.

c) Utilize coding to store and extract data more effectively for data analysis.

d) Select and apply features of technology tools effectively to organize, summarize and gain insight from data.

e) Select the appropriate visualization based on context and audience and create it using technology tools to effectively communicate an idea.

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. *Additional Content Background and Instructional Guidance:*
• Certain technological tools can be used to generate conventional and unconventional visualizations of data to explore patterns and/or analyze a large data set. • Various technological tools have prebuilt mathematical and statistical functions that allow for efficient exploration and analysis. • Coding tools can allow for effective storage and extraction of data for more efficient analysis. • Some technological tools have other functions that are useful to organize, summarize and gain insight from data. • Visualization tools offer a variety of conventional and unconventional visualizations to help communicate our ideas to a wide audience.