Project Title

Reimagining Data Literacy as an Essential Component of the 21st Century Liberal Arts Education

A Knowledge Synthesis Report

By:

Abel Ayele, M.S.
Instructor, Department of Mathematical Sciences
Lincoln University, PA

January 31, 2023

Table of Contents

1. Introduction

1.1. Overview

This project is conducted as part of the Faculty Development Grant during the summer of 2022. The motivation for this project is the need to initiate discussion to reimagine modern day liberal arts education in such a way that addresses the challenges and opportunities that big data brought to the 21st century market economy. Among the many challenges and opportunities of this digital age, data revolution is at the center. With explosion in volume and complexity of data, the need for professionals with the right skills to manipulate, analyze and extract relevant information and knowledge from the bulk of data available in the market is growing exponentially. Such knowledge helps companies' decision-making process tremendously and gives them a competitive edge in the current market economy. As a result, data literacy has become a very essential and marketable skill that graduates in every discipline including students in liberal arts majors must acquire.

Acquiring this essential skill gives graduates in any field a competitive advantage in the today's technology led economy. However, the importance of data literacy within the realm of quantitative literacy is not given enough attention and often overlooked in liberal arts education curriculums. It is often considered a skill that only graduates in STEM and/or business disciplines need. In this project, a knowledge synthesis report that highlights the need for data literacy and commonly used data science methods and tools in different liberal arts fields is prepared. The purpose of this report is to initiate a discussion among different disciplines and will be used as a basis to write a journal article in the future. As such, it will be presented at the Faculty Development grant workshop organized by CETL at Lincoln University with the aim of initiating a discussion among faculty across disciplines on the subject of data literacy with in the university and beyond.

By shading light on the importance of data literacy in our programs, this report serves as an initial document to help the discussion to revise our curriculum in such a way that strengthen students' quantitative reasoning skills particularly data literacy skills. The findings of this project will help as an input in the future development of data science courses for the data science concentration in the department of mathematical sciences that is currently under consideration. It also helps harness the multidisciplinary nature of data science to initiate discussion among faculty members on potential curriculum revision and research collaboration initiatives. Furthermore, this report can be used by Lincoln student and faculty researchers to refer to the commonly used data science methods and analysis tools.

1.2. Goals and Objectives

The goal of this project is to compile a knowledge synthesis report and initiate a discussion on the need for the data literacy across disciplines in the liberal arts institutional setting such as Lincoln University.

Objectives

1. Compile a knowledge synthesis report discussing the essence of data literacy, the need for data literacy in liberal arts education and commonly used data science methods and tools.
2. Present the findings of this report at the FDG workshop for students and faculty at LU as organized and scheduled by CETL.

2. What is Data Literacy?

2.1. General

Data literacy in simple terms is defined as the ability to read, understand, work with and communicate data. It comprises critical thinking skills to use, interpret, and make decisions with data, then communicate its importance and value to others. With the growing open data for social change and sustainable development, there is a growing focus on the end user to be able to take advantage of the widely available data without the need for intermediaries such as specialist applications, data journalists, pressure groups and political parties to select and interpret data on users behalf [1]. In this information age, data literacy is no longer confined to the skill that only students and researchers need to analyze and use data. It has played a decisive role in the lives of many people. More and more businesses rely on data as a basis for their products and services and their decision-making efforts. Government agencies local and national rely heavily on data in order to respond to societal challenges [1].

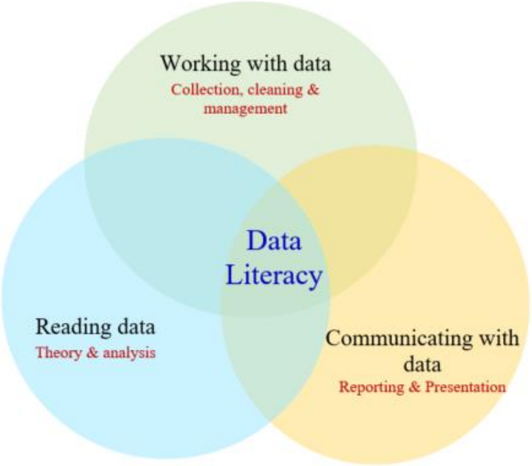


Figure 1: Data Literacy Venn diagram

According to Cukier and Mayer-Schoenberger, some among many examples of data available for public use through a process known as “Datafication” are; census results, company accounts, product review on e-commerce websites, government data portals, weather emergencies forecast, patient monitoring, citizens participation and decision-making, trade statistics and etc. [2]. Locations can be datafied first with the invention of longitude and latitude and now GPS. Even “friendships and likes can be datafied” via social media platforms such as facebook, stated Cukier and Mayer-Schoenberger [2]. Other researchers argue that data literacy needs to go beyond the mere understanding and use of data to a wider “critical big data literacy” that fosters understanding and critical reflection of big data system at its center [3].

2.2. The Need for Data Literacy

According to Bowne-Anderson [4], the ability to understand and communicate about data is an increasingly important skill for the 21st-century citizen, for three reasons. These are:

1. Data science and AI are affecting many industries globally, from healthcare and government to agriculture and finance
2. In the today’s media landscape, much of the news is reported through the lenses of data and predictive models
3. So much of our personal data is being used to define how we interact with the world.

The last decade exhibited a significant explosion in the volume and complexity of data generated by the public, small to large corporations, government and non-governmental agencies all around the world. In 2028, Forbs estimates, about 90% of all data in the world was created in the just the two years before that [5]. Today, an average internet user is estimated to produce about 1.76MB of data every second. According to Bernard Marr, by the end of 2022, there will be 97 zettabytes of data in the world [6]. One zettabyte is equivalent to a trillion gigabytes or a ‘1’ with 21 zeros after it. Projections indicate that, in just 3 years this number will grow to 175 Zettabyte. With this fast growing volume of data and cloud computing technologies, data literacy has become a very essential skill across all domain knowledge.

Why is data literacy needed as an essential skill? Researchers argue that, in today’s datafied society, data literacy should not only be considered as a marketable skill to understand, manipulate and use data, but also as an expanded knowledge base that fosters “people’s awareness and understanding of big data practices” in such a way that accommodate the changing landscape of digital technologies [7]. Citizen’s need to be aware of their personal data, its collection and usage across digital platforms. Furthermore, understanding and critically reflecting on the possible risks and implications that come with these big data practices is very important [3]. Individuals understanding of personal data is essential to foster an informed citizenry in times of increasing drive for social justice, economic equality of citizens and other political and social implications of big data systems. Studies show that, there are multi-faceted risks associated to these systems that threatens individual privacy, increases surveillance, fosters existing inequalities and reinforces discrimination [8], [9].

The disparity with in our society to access, understand and properly utilize data further widens the current gap in socio-economic status of citizens. Minorities and poorer communities are often underrepresented in the data driven tech advancements and the relevant opportunities. Studies show that there exist a huge lack of diversity in the data science and Big Data industry, which contributes to the further widening of the socioeconomic divide with in our society. Kuhlman et.al. argues that not only the fields of computing but also the data sets used for analysis and modeling purposes suffer great deal of lack in diversity [10]. Lack of diverse perspectives in data has been foundational to the inequity in the treatments of underrepresented and protected groups.

As an HBCU, that prides itself with its legacy of Liberal arts education, it is imperative for Lincoln University to reimagine its curriculum in such a way that takes data literacy in to account mainly because of the following reasons.

i. To create awareness and understanding around collection and use of personal data by different digital platforms in order to maintain privacy and security.
ii. To produce graduates that have acquired a useful and highly marketable skill in the today's digital economy.
iii. To contribute to the narrowing of the diversity gap that currently exists in the data science and technology fields by training data literate historically disadvantaged minority students.

2.3. Digital Literacy vs. Data Literacy

American Library association defines Digital Literacy as the "ability to use information and communication technologies to find, evaluate, create, and communicate information, requiring both cognitive and technical skills". Although defined in the context of internet and digital platforms, data literacy is more a more advanced concept and which is focused on understanding, manipulation and interpretation of data.

2.4. Data Literacy in Liberal Arts Education

The emerging field of data science is an ideal case study in how liberal arts graduates are a critical part of the information economy. There is a problem of inconsistency in data literacy education across the public, private, and academic sectors, and data literacy training has not been approached systematically or formally at post-secondary institutions particularly, at liberal arts institutions such as ours. Most liberal arts institutions lack versatile curriculum in data literacy that meets the needs of today's digitized society. Most students in small liberal art schools lack basic quantitative skills, statistics and IT that are essential to develop a curriculum that responds to the need for data literacy. To help overcome this barrier in the process of learning data science concepts, some studies propose curriculum models in liberal arts institutions, which systematically incorporates the knowledge module of data science while remedying the lack of preparation in the basic math and IT skills [11].

We as faculty must understand that our curriculum must evolve to respond to the growing challenges of the present. As we develop and revise our curriculum in the realm of liberal arts education, one of the pressing question we should ask ourselves is, whether or not our curriculum equips students with the digital and data literacy skills needed for the datafied 21st century work force? Here at Lincoln University, as part of the curriculum coherency initiative, a lot has been done in recent years to standardize our curriculum especially in the area of general education. However, it appears that the framing of the general education curriculum revolves around the three skills, writing, critical reading and quantitative literacy skills in its traditional sense. One might argue that in depth analysis and reinforcement of these mentioned skills is still lacking. For example, in the quantitative reasoning area, the focus was more or less to refine and delineate the math requirements rather than focusing on how to redefine and reinforce the this skill set in such a way that prepares students to be problem solvers in the growing challenge of datafied society. This can be done by adding rigor in to the quantitative reasoning requirement of our curriculum in one of the following ways:

i. By expanding the required credits by including data literacy oriented courses such as statistics, foundational data science etc., or

ii. By working closely with math department to revise the current general education courses in such a way that incorporates interdisciplinary projects that provides students with the minimum essential data literacy skills.

Data science by nature is a fertile ground where interdisciplinary collaboration among faculty members across disciplines can be fostered.

3. Data Literacy Skills

Data literacy skills include the following abilities: Knowing what data is appropriate to collect and use for a particular purpose. Understanding the data at hand by applying exploratory data analysis (EDA) techniques. Understanding the analysis tools and methods, and where and how to use them is also very important. Analyzing data using different statistical tools and methods, interpreting the analysis results, and thinking critically about information yielded by data analysis. Data communication (storytelling) is also considered an essential component data literacy skill.

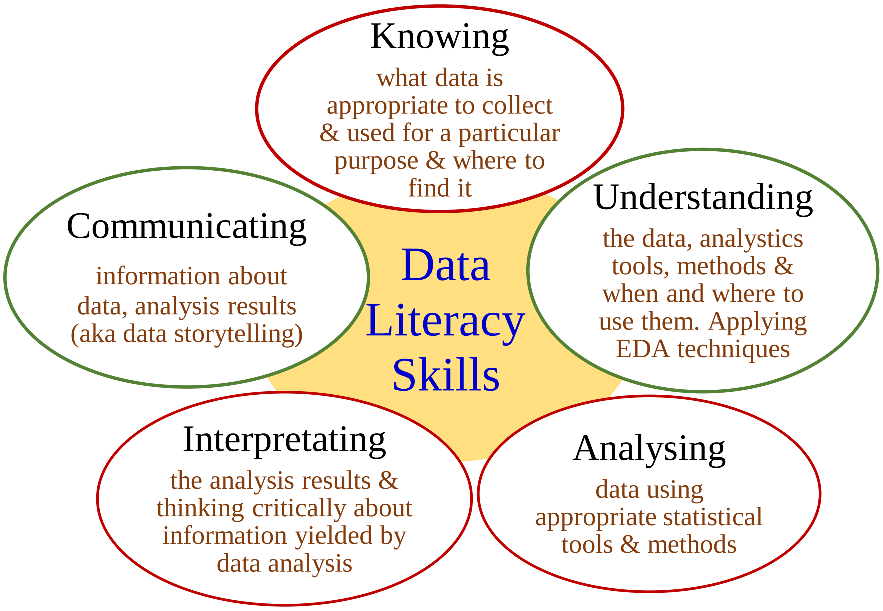

Knowing
what data is appropriate to collect & used for a particular purpose & where to find it
Communicating
information about data, analysis results (aka data storytelling)
Data Literacy Skills
Understanding
the data, analystics tools, methods & when and where to use them. Applying EDA techniques
Interpretating
the analysis results & thinking critically about information yielded by data analysis
Analysing
data using appropriate statistical tools & methods

Fig 2: Components Data Literacy Skills

3.1. Data Collection and Storage

Data is generated either voluntarily through observations, through surveys by conducting field and laboratory experiments or involuntarily through engaging in daily activities such web browsing, using mobile apps, driving etc. Having the basic understanding of the type and extent of data we unintentionally generate in our daily encounters with internet for work or personal activities is beneficial. However, collecting data intentionally for the purpose of research or business decision making requires certain set of skills.

Data may be grouped into four main types based on methods for collection.

1. **Observational Data**: collected through observation of behavior or activity. Collected using methods such as human observation, surveys, or the use of an instrument or sensor to monitor and record information. Because it is collected it in real time, it is often difficult or impossible to recreate if lost. Each methods used to generate such data require certain level of skills. Human observations require sound knowledge and understanding of the subject being observed. Surveys include either collecting responses through oral interviews or written set of questions by the researcher. The researcher should be versed in preparing the questioners and conducting the interviews.
2. **Experimental Data:** collected through active intervention by the researcher to produce and measure change or to create difference when a variable is altered. Experimental data typically allows the researcher to determine a causal relationship and is typically

projectable to a larger population. This type of data are often reproducible, but it often can be expensive to do so. Experimental data can be collected either in the lab or from the field.

3. **Simulation Data:** Simulation data are generated by imitating the operation of a real-world process or system over time using computer test models. For example, to predict weather conditions, economic models, chemical reactions, or seismic activity. This method is used to try to determine what would, or could, happen under certain conditions. The test model used is often as, or even more, important than the data generated from the simulation.
4. **Derived/Compiled Data:** Derived data involves using existing data points, often from different data sources, to create new data through some sort of transformation, such as an arithmetic formula or aggregation. For example, combining area and population data to create population density data. While this type of data can usually be replaced if lost, it may be very time-consuming (and possibly expensive) to do so.

3.2. Understanding Data (Exploratory Data Analysis)

When dealing with data, knowing how the data was collected and its kind will help understand its meaning, how much to trust it, and how much work needs to be done to convert it into a useful form [4]. Data is simply a collection of facts, such as numbers, words, measurements, observations or just descriptions of things. According to Bowne-Anderson, data is available in one of the three forms [4]:

1. Tabular data (that is, data in a table, like a spreadsheet),
2. Image data or
3. Unstructured data, such as natural language text or html code, which makes up the majority of the world's data.

Understanding such data is a very important first step towards analysis and interpretation of data. The general approach to understanding data is through exploratory analysis. Exploratory techniques make it easy to see in to the data, to poke around some around in search of ideas about how things work.

Exploratory data analysis is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. The core purpose of EDA is understanding the data set. Some of the tasks to be carried out to understand the data set are:

- Extracting important variables
- Checking assumptions
- Uncovering outliers, missing values, or human errors
- Preliminary selection of appropriate models
- Understanding the relationship(s), or lack of, between variables
- Maximizing insights of a dataset and minimizing potential error later in the process

- Detecting the underlying structure of the data
- Assessing the direction and rough size of relationships between explanatory and outcome variables.

Exploratory data analysis techniques have been devised to simplify the tedious and overwhelming nature of columns of numbers or spreadsheets in determining important characteristics out of data. EDA is generally cross-classified in to two ways. First, each method is either non-graphical or graphical. And second, each method is either univariate or multivariate (usually just bivariate).

Non-graphical methods (describing data using numerical measures) generally involve calculation of summary statistics, while graphical methods summarize the data in a diagrammatic or pictorial way. Univariate methods look at one variable (data column) at a time, while multivariate methods look at two or more variables at a time to explore relationships. Usually our multivariate EDA will be bivariate (looking at exactly two variables), but occasionally it will involve three or more variables. It is always recommended to perform univariate EDA on each of the components of a multivariate EDA before performing the multivariate EDA.

Each of the categories of EDA have further divisions based on the role (outcome or explanatory) and type (categorical or quantitative) of the variable(s) being examined.

3.2.1. Univariate non-graphical EDA

Non-graphical EDA employs numerical summary statistics to have the initial feel and understanding of the data. The goal of such analysis is to appreciate the “sample distribution” and to make some tentative conclusions about what population distribution(s) is/are compatible with the sample distribution. Outlier detection is also a part of this analysis. In the case of categorical data, the only useful univariate non-graphical techniques to be used is some form of **tabulation of the frequencies**, usually along with calculation of the fraction (or percent) of data that falls in each category. A simple tabulation of the frequency of each category is the best univariate non-graphical EDA for categorical data. Univariate EDA for a quantitative variable is a way to make preliminary assessments about the population distribution of the variable using the data of the observed sample. For quantitative variables, the concern here is the quantitative numeric (non-graphical) measures, which are the various sample statistics. In fact, sample statistics are generally thought of as estimates of the corresponding population parameters. These measures are used to describe mainly three aspects of the data. These are

i. Central Tendency (measures of center)
   The most commonly used measure of center are **mean**, **median** and **mode**. The arithmetic mean is simply the sum of all of the data values divided by the number of values. It is given by the following formula.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

The sample median is the middle value after all of the values are put in an ordered list. The most common measure of central tendency is the mean. However, for skewed distribution or when there is concern about outliers, the median may be preferred. A rarely used measure of central tendency is the mode, which is the most likely or frequently occurring value. More commonly, we simply use the term “mode” when describing whether a distribution has a single peak (unimodal) or two or more peaks (bimodal or multi-modal). In symmetric, unimodal distributions, the mode equals both the mean and the median. In unimodal, skewed distributions the mode is on the other side of the median from the mean.

ii. Spread (measure of variation)

Variance, standard deviation, and interquartile range are the most commonly used measures of variation. Spread is an indicator of how far away the data value is located from the center. The most commonly used symbol for sample variance is s^2 , and the formula is:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{(n-1)}$$

The standard deviation is simply the square root of the variance. Therefore, it has the same units as the original data, which helps make it more interpretable. The sample standard deviation is usually represented by the symbol s.

iii. Skewness and kurtosis

Skewness is a measure of asymmetry. Kurtosis is a more subtle measure of peaked-ness compared to a Gaussian distribution.

3.2.2. Univariate graphical EDA

In addition to looking at the various sample statistics discussed in the previous section, we also need to look graphically at the distribution of the sample. Non-graphical and graphical methods complement each other. While the non-graphical methods are quantitative and objective, they do not give a full picture of the data; therefore, graphical methods, which are more qualitative and involve a degree of subjective analysis, are also required.

For categorical data, **bar graphs** and **pie charts** are the most commonly used graphical representation of the tabulation of data (most commonly known as frequency table). **Histograms** are also among the most commonly used graphical representations of data in EDA. Although their common use is in quantitative data, they can also be used to describe ordinal categorical data. The other commonly used graphs in EDA for quantitative data are **stem-and-leaf plot**, **boxplots** and **quantile-normal plots** (QN plots) or more generally the quantile-quantile or QQ plot.

3.2.3. Multivariate non-graphical EDA

Multivariate non-graphical EDA techniques generally show the relationship between two or more variables in the form of either cross-tabulation or statistics.

i. **Cross-tabulation:** making a two-way table with column headings that match the levels of one-variable and row headings that match the levels of the other variable, then filling in the counts of all subjects that share a pair of levels.

ii. **Univariate statistics by category**: Especially for a categorical explanatory variable and a quantitative outcome variable, it is useful to produce a variety of univariate statistics for the quantitative variable at each level of the categorical variable.

iii. **Correlation and Covariance**: The sample covariance is a measure of how much two variables "co-vary", i.e., how much (and in what direction) should we expect one variable to change when the other changes. Positive covariance values suggest that when one measurement is above the mean the other will probably also be above the mean, and vice versa. Negative covariance suggest that when one variable is above its mean, the other is below its mean. Covariance near zero suggest that the two variables vary independently of each other. Covariance tend to be hard to interpret, so we often use correlation instead. The correlation has the nice property that it is always between -1 and +1, with -1 being a "perfect" negative linear correlation, +1 being a perfect positive linear correlation and 0 indicating that X and Y are uncorrelated. The symbol r or $r_{x,y}$ is often used for sample correlations. The general formula for sample covariance is given as:

$$\mathrm{Cov}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

The general formula for sample correlation is given as:

$$\mathrm{Cor}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{s_x s_y}$$

iv. **Covariance and correlation matrices:** When we have many quantitative variables, the most common non-graphical EDA technique is to calculate all of the pairwise covariances and/or correlations and assemble them into a matrix. Note that the covariance of X with X is the variance of X and the correlation of X with X is 1.0.

3.2.4. Multivariate graphical EDA

i. **Univariate graphs by category**: Side-by-side boxplots are the best graphical EDA technique for examining the relationship between a categorical variable and a quantitative variable, as well as the distribution of the quantitative variable at each level of the categorical variable.

ii. **Scatterplots:** For two quantitative variables, the basic graphical EDA technique is the scatterplot which has one variable on the x-axis, one on the y-axis and a point for each case in your dataset.

In summary, it is always recommended to perform appropriate EDA before further analysis of your data. Perform whatever steps are necessary to become more familiar with your data, check for obvious mistakes, learn about variable distributions, and learn about relationships between variables. EDA is not an exact science – it is a very important art!

3.3. Data Analysis and Interpretation

Data analysis is the process of analyzing raw data using different statistical tools in order to draw out patterns, trends, and insights that can tell you something meaningful about a particular area of the business. The insight drawn from data depends on the type of analysis run. There are four main types of data analysis [12]. These are:

1. **Descriptive analysis:** looks in to what happened in the past. The purpose of such analysis is to describe what has happened; it does not try to explain why this might have happened or to establish cause-and-effect relationships. The two most important techniques used in descriptive analytics are data **aggregation** and data **mining**. Data aggregation is the process of gathering data and presenting it in a summarized format. Data mining explores the data in order to uncover any patterns or trends. EDA and descriptive analysis looks similar in scope or the kind of techniques they employ. However, they are different in terms of the goal of the analysis. The goal of EDA is only to understand and prepare data for further analysis while descriptive analysis can be used as either a first step towards further analysis or an end result by itself.
2. **Diagnostic analysis:** looks deeper to understand why something happened. For example if one identifies the existence of anomalies/outliers in the data at the EDA level, then diagnostic analysis tries to get to the root cause and tries to answer why the anomaly happened. Sometimes additional data that offer further insight might be needed. Techniques such as **probability theory**, **regression analysis**, **filtering**, and **time-series analysis** might be employed in diagnostic analysis.
3. **Predictive analysis:** tries to forecast what is likely to happen in the future. Predictive models use the relationship between a set of variables to make predictions. The two main purposes of predictive analytics are **forecasting** and **classification**. The analyst in this case develops predictive models, which estimate the likelihood of a future event or outcome. The most commonly used classification technique or algorithm is logistic regression, which is used to predict a binary outcome based on a set of independent

variables. **Machine learning** is one of the areas of predictive analytics. Machine learning models are designed to recognize patterns in the data and automatically evolve in order to make accurate predictions.

4. **Prescriptive analysis:** is a complex type of analysis that capitalizes on prior analysis steps that answer questions such as what happened, why it happened, what might happen in the future and tries to answer what should be done next? It tries to answer deep questions such as, what steps can you take to avoid a future problem? What can you do to capitalize on an emerging trend? Such analysis often involve complex algorithms, machine learning, statistical methods, and computational modeling procedures.

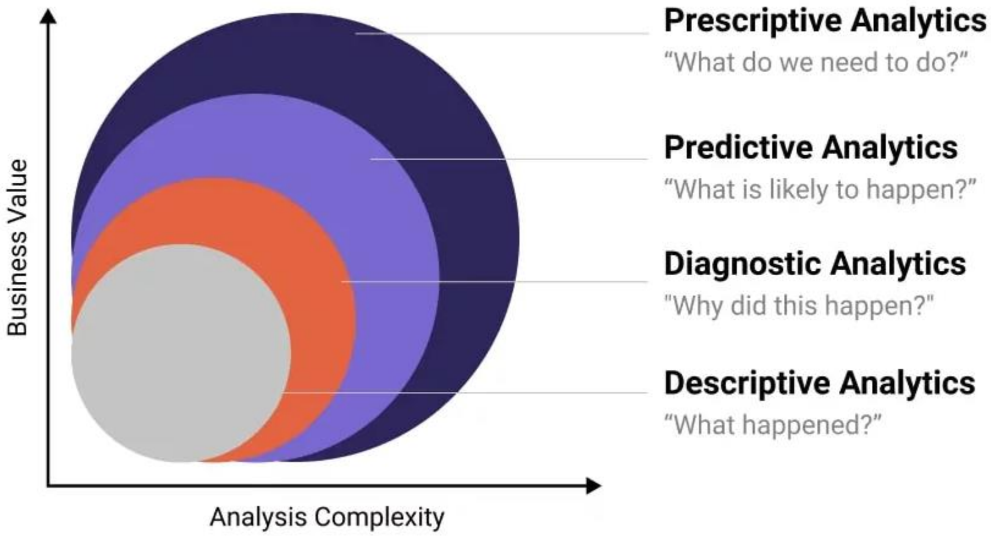


Fig 3: Types of Data Analysis: *Image taken from: https://www.velvetech.com/blog/data-analytics-in-healthcare/*

3.4. Data Communication (Data Storytelling)

Data storytelling is important because it allows for the effective communication of data. Microsoft defines data story telling as "the concept of building a compelling narrative based on complex data and analytics that help tell your story and influence and inform a particular audience" [13]. The main purpose of data storytelling is to present a complex information in a simplified way so that the audience gets inform, engaged and makes critical decisions faster and more confidently. Some benefits of effective data storytelling are:

- Adding value to the data and insights.
- Interpreting complex information and highlighting essential key points for the audience.
- Providing a human touch to the data.
- Offering value to the audience and industry.
- Building credibility as an industry and topic thought leader.

Among the most important tools and techniques in communicating the findings and insights of data are **data visualization** and **data dashboard**. Data visualization helps transform bulk and complex data into something simpler and digestible. Effective data visualizations can help reveal patterns, trends, and findings from an unbiased viewpoint. It also provides context, interpret results, and articulate insights while streamlining data so your audience can process information and get engaged easily.

The three essential elements of data communication (storytelling are):

i. **Narrative:** creating an informative insight by distilling complex information in such a way that is understandable by an average audience.
ii. **Visuals:** supplementing the narrative with visual tools such as charts and graphs makes it more clear and engage the audience more effectively.
iii. **Critique:** provides the full interpretation of the narrative. It also provides deeper context to the data story.

Integrating and using the above key elements in data communication creates the much needed emotional response and clearer understanding of the facts on the audience. By doing so one can successfully influence people's decision-making and drive change.

4. Measuring Data Literacy

Similar to other literacy skills, measuring the proficiency level of individuals or team in an organization is a very important part of the data literacy framework. According to DuBios of Quanthub, the data literacy assessment measures an individual's numeracy skills, ability to identify sources of data and their purpose, understanding of data visualization tools and techniques, ability to generate insights and make data-driven decisions, and more [14]. Not only the technical aspects of data literacy, assessments can also be used to identify non-technical aspects such as attitudes towards data, willingness to communicate data insights and support data-based decisions, and overall data culture.

Data literacy assessments generally test individual knowledge at the basic level of topic hierarchy, then roll those topics up into more advanced categories of skills. The skills that should be assessed are broadly categorized in to three areas.

i. **Basic skills:** ability to understand basic statistical tools and operations, proficiency in data interpretation, reading graphs and charts, ability to choose the best graphs and charts for presentation etc.
ii. **Intermediate skill:** ability to prepare and present proposals based on sound and relevant data analysis techniques and trends supported by data.
iii. **Advanced skill:** ability to interpret and present machine-learning algorithms in such a way that foster data driven decision.

4.1. Data Literacy Score

Researchers, educators and practitioners developed different scoring techniques to assess the data literacy level of individuals (students, professionals) or team of professionals at a certain company. Sickler and her team in their work "Measuring Data Skills in Undergraduate Student Work" developed a rubric-based scoring system that could reliably measure student performance across seven data skills [15]. Using rubrics to evaluate students' work is a long-standing and reliable practice in education. Sickler et al. in their work first articulated their evaluative criteria which they referred as "data skill indicators". As such, they identified seven data skill indicators. These are:

i. Decoding Data: ability to read data or measurement from a representation or table
ii. Describing Data: ability to identify patterns and relationships between variables
iii. Interpreting meaning and drawing conclusions; ability to explain what data patterns mean or why they are important.
iv. Supporting claims with evidence: ability to identify appropriate evidence that leads to a specific conclusion or claim
v. Making hypothesis or research question: ability to articulate a question or hypothesis that would generate new information or insight
vi. Critical thinking about data: ability to critically examine data sources and interpretations, including data limitations, alternative interpretations, additional data needs, evaluate strength of a claim
vii. Communicating ideas effectively: ability to clearly articulate thoughts and ideas about data in writing.

On the next step, they created definitions of quality or levels of proficiency for each indicators listed above and chose four levels of scoring system: (4) exceeds expectations, (3) meets expectations, (2) needs improvement, and (1) inadequate. In this study, because their goal was to understand the development of each skill instead of assigning overall grade, they chose an analytic scoring strategy in which each indicator in the rubric was considered individually and scores were applied at this criterion level. For example if the indicator is "describing data", students were evaluated where they end up in the four-level of scoring system.

Finally, they developed guidelines for applying overarching skill score to student work, based on question-by-question scores as summarized in the following table.

Overall score	Scoring rules
Exceeds expectations (4)	Meets criteria for a 3 and at least one item was scored a 4.
Meets expectations (3)	Received a 3 on all items; or

	(in modules with five or more items) received a 1 or 2 on no more than two items; (in modules with less than five items) received a 1 or 2 on no more than one item.
Needs improvement (2)	More than two items received a 2 or lower; and fewer than half of items received a 1 Or: Exactly half of the items received a 1, and all others received 3s.
Inadequate (1)	Half or more of items received a 1. Note: There is one exception to this, described at level 2

Table 1: Guidelines for applying an overarching skill score to student work (Source: Sickler et.al.)

Similar other studies proposed different formats of assessment rubrics to measure data literacy skills of undergraduate and graduate students in different institutional setups. Most of them including the rubric in this study was developed by focusing on science students. Future work needs to focus on developing similar rubrics targeting small liberal art schools such as Lincoln.

Data literacy score is not only needed at the universities to assess students' skill but also designed for teams of people – from corporate groups to nonprofits to government agencies. One example of the many team-based assessment scores is show below [16]. The assessment is comprised of 50 Likert-style questions, broken up into seven categories (purpose, ethics, data, technology, people, process, culture) with seven questions each, and a final, overarching 50th question. These questions each contribute 10 possible points to the overall team score, for a total of 500 potential points. Teams assign representatives to complete the survey. Each survey taker scores each of the 50 questions, and an average is computed for each of them. Finally, the average score for each question is summed together to produce the overall team Data Literacy Score.

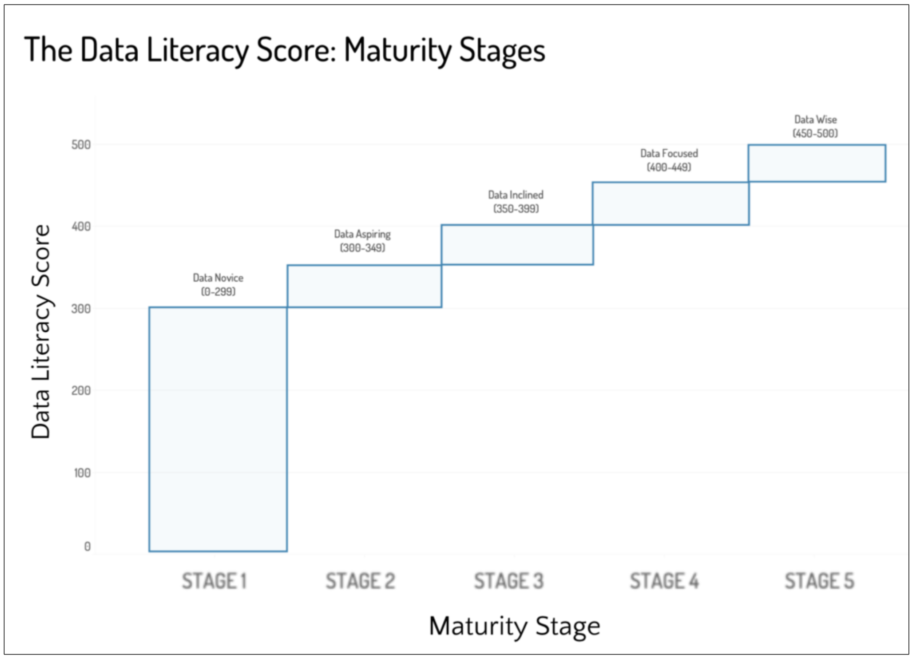


Fig 3: Data Literacy Score vs Maturity Stage: *Image taken from https://dataliteracy.com/data-literacy-score/*

4.2. Global Data Literacy Benchmark

Besides the scoring rubrics at different colleges and the team-based data literacy scores used by corporations and non-profits, there are global benchmarks developed in this regard. For example, "Databilities" is one such global benchmark developed by "Data To The People", a data literacy assessment and development company. "Databilities" is a framework encompassing 18 core competencies across the domains of data concepts and culture, reading, writing, and comprehension [17].

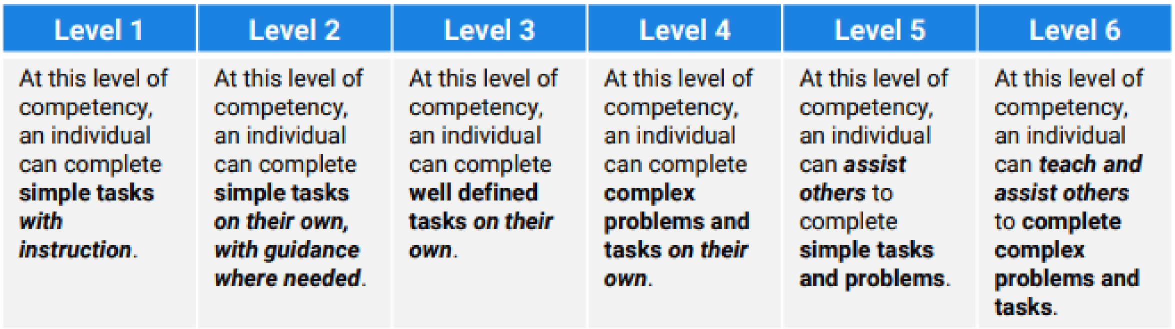
The competencies within each domain of the expanded "Databilities" framework are:

Domain	**Core competencies**
Data Concept and Culture	Data Culture Data Ethics
Reading	Data Discovery Evaluating and Ensuring Quality of Data
Writing	Data Collection Data Management and Organization Data Manipulation Data Curation and Reuse Metadata Creation and Use Data Conversion (Format to Format) Data Governance
Comprehension	Data Analysis

	Data Interpretation (Understanding Data) Identifying Problems Using Data Data Visualization Presenting Data (Verbally) Data Driven Decision Making Evaluating Decisions / Conclusions Based on Data

Table 2: Summarized from [17]

For each competency within the "Databilities" framework, there are up to 6 levels of progression as shown below.



Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
At this level of competency, an individual can complete **simple tasks** ***with instruction***.	At this level of competency, an individual can complete **simple tasks** ***on their own, with guidance where needed***.	At this level of competency, an individual can complete **well defined tasks** ***on their own***.	At this level of competency, an individual can complete **complex problems and tasks** ***on their own***.	At this level of competency, an individual can ***assist others*** to complete **simple tasks and problems**.	At this level of competency, an individual can ***teach and assist others*** **to complete complex problems and tasks**.

Fig 4: Taken from [17]

Using the 6 levels of progression outlined in the "Databilities" framework, the Global Data Literacy Benchmark has identified 3 cohorts of employees:

1. Those who need direction – the Curious (levels 1 & 2)
2. Those who are independent – the Confident (levels 3 & 4)
3. Those who can guide others – the Coaches (levels 5 & 6)

4.3. Data analytics pilot project (Statistics classroom case study at LU)

As part of this project, I have conducted a pilot study in which I introduced a data analytics group project in my statistics class during the fall 2022 term. Besides contributing to the core student learning outcomes (SLOs) as outlined on the course syllabus, this group project aims to help student develop their skills in:

- Collaboration and team work with peers
- Fundamental understanding of hypothesis testing
- Data analysis using excel particularly in descriptive statistics including hypothesis testing
- Interpretation of analysis results particularly (p-value)
- Report organization and presentation

A project-grading rubric has been developed based on the following data skill indicators and it is graded accordingly.

Evaluation Criteria

i. Analysis Results (30%)
   - Descriptive Statistics values
   - Frequency distribution table
   - Histograms

ii. Hypothesis Testing (30%)
   - Appropriates of the hypothesis test
   - Appropriateness of the assumptions listed
   - The null and alternative hypothesis for the hypothesis test
   - P-value, degree of freedom

iii. Interpretation of Results (10%)

iv. Report Organization (15%)
   - 2 to 3 pages report that summarizes the analysis result expected

v. Presentation (15%)
   - 3-5 slides PowerPoint expected
   - Time management during presentation is important

Grading Rubric:

Overall Rating	**Grading Range**	**Description**
Excellent	4 (93%-100%)	- Work demonstrates all descriptive statistics values accurately - Work demonstrates accurate frequency distribution table & histogram - Work demonstrates fundamental understanding of hypothesis testing - The interpretation of results is sound and consistent with the analysis result - Report is complete and is well organized - Work presented using PowerPoint slides, with in the given time limit and all questions answered correctly.
Satisfactory	3 (82%-92.9%)	- Work demonstrates most descriptive statistics values correctly - Work demonstrates correct frequency distribution table & Histogram - Work demonstrates basic understanding of hypothesis testing - The interpretation of results is good with some inconsistencies with the analysis result - Report is complete with minor organization issues - Work presented using PowerPoint slides, with time management issues and some questions remain unanswered.

Needs improvement	2 (72% - 81.9%)	- Work demonstrates some descriptive statistics values correctly - Work demonstrates only partial understanding of frequency distribution & Histogram (ex. missing graphs, etc.) - Work demonstrates below basic understanding of hypothesis testing - The interpretation of results is flawed with major inconsistencies with the analysis result - Report is incomplete complete with minor organization issues - Work presented without PowerPoint slides, with time management issues and some questions remain unanswered.
Unsatisfactory	1 (60% - 71.9%)	- Work demonstrates most descriptive statistics values incorrectly - Work demonstrates very limited understanding of frequency distribution & Histogram (ex. missing graphs, etc.) - Work demonstrates no basic understanding of hypothesis testing - The interpretation of results is flawed with major inconsistencies with the analysis result - Report is incomplete complete with major organization issues - Work presented without PowerPoint slides, with time management issues and no questions answered.
Unacceptable	0 (<60%)	- Work demonstrates all the descriptive statistics values wrong - Work demonstrates no understanding of frequency distribution & histogram at all - Work demonstrates no understanding of hypothesis testing - The interpretation of results is flawed with major inconsistencies with the analysis result - Missing summary report - Missing presentation

All the 22 students completed and presented their data analytics project from November 28-30, except 1 student who failed participated in the presentation. They were evaluated based on the criteria and grading rubrics presented above. The result shows, 19 out of 22 (86.3%) scored satisfactory (between 82-92.9%) and the remaining 3 students (13.7%) scored unsatisfactory (60-71.9%) result. Although no student hits the “Excellent” rating, the overwhelming majority rated satisfactory on this basic levels of data skills which is encouraging.

This is just a preliminary study which is based on a very limited data and a more generalized grading rubrics. The scope and level of data literacy skills tested in this pilot study is also very limited. As such, it cannot be taken as a conclusive indicator data literacy level. Future study will focus on developing a detailed analytic scoring strategy (rubric) in which more data skill indicators are included in the rubric and considered individually where the scores are applied at the criterion level instead of just the overall grade. Data size can also be improved by collecting multiple semesters of data instead of just one semester.

5. Data Science and Digital Humanities

5.1. Data Science Essentials

Data have increasingly become the medium for the collection, storage and transmission of information. As such, all liberal arts students will need to have basic literacy in data science and the digital humanities. Data science in general is a fast growing interdisciplinary field involving the extraction of knowledge from data in various forms that supports critical thinking and decision-making. It is an approach through which inquiry across the liberal arts can be framed and carried out by leveraging the study of data. Enormous amounts of structured and unstructured data are constantly being produced in a wide variety of fields, including business, politics, education, social media, scientific research, economics, geography, music, film & television, the environment, sports, medicine, engineering, etc.

It is important to understand some of the characteristics of data science as an emerging discipline. The first is the fact that data science cannot exist without the domain expertise to which data science tools are being applied. Secondly, because of the wide range of domains of applicability and multiple skill areas it draws, it is interdisciplinary by nature.

5.2. Digital Humanities

Data science and digital humanities are closely inter related but distinct fields. Digital humanities is the systematic use of digital resources in the humanities as well as the analysis of their applications. It involves the application of computational tools, methods, and approaches to extend the human capacity to explore questions relating to people, cultures, and communities. Digital humanities explore how technology and computing intersect with social, cultural, and historical factors. At the center of digital humanities work is, data science along with tools and approaches for digitization and research, analysis and interpretation, presentation and dissemination all working together with domain expertise.

5.3. Tools and Methods

Data science develops and applies tools and methods from mathematics, statistics, and computing to study applications in a wide range of domains. These tools and methods have become increasingly important even in the domains that did not historically adopt data-driven approaches.

Liberal arts colleges such as Lincoln University should not only participate in the world of data, but they also must lead the way in data science.

The data analysis methods in liberal arts research are mostly summarized under the following four methodological approaches:

i. Descriptive analysis
ii. Diagnostic analysis
iii. Predictive analysis
iv. Prescriptive analysis

These methodologies are only the start. Technical tools are also necessary to be able to sieve answers out of massive amount of data. The most commonly used technical tools are:

i. Programing languages such as Python and R and tool libraries such as NumPy
ii. Visualization tools such as Tableau, ggplot etc.
iii. Mathematical and statistical techniques such a Bayesian analysis and linear regression.
iv. Data storage and computer networks

Advanced big data computational machine learning and AI techniques can also be used as effective tools in different liberal arts oriented researches.

6. Challenges and Opportunities

6.1. Challenges

Acquiring strong data literacy skills requires at least basic quantitative reasoning foundations. Among the many challenges of producing graduates with enough data literacy skills at the liberal arts schools are:

1. The lack of foundation in basic math including statistics and IT skills before entering the university.
2. Understanding the need and integrating data literacy in to curriculum especially in the social sciences and humanities fields.
3. Limitation of resources

Bridging the gap in mathematics and IT basics is necessary to expand data literacy initiatives across programs. Furthermore, by bridging the necessary knowledge gap, implementing the Data Science education curriculum at the university can be smoothly realized, and it is expected to contribute to the implementation of educational contents suitable for the actual situation of students and the guarantee the needed data literacy skills to be achieved. Understanding the need for data literacy as an essential skill in the current market economy and preparing students for this new challenge is of a paramount importance. Discussion among faculty and staff at Liberal arts colleges such as Lincoln University on this subject needs to be fostered. In fact, the main goal of this project is to spark a discussion among faculty across discipline on the need for data literacy and on how to effectively incorporate it into our curriculum. Facilitating dedicated resources such as data analytics labs spaces, tools and personnel helps the expansion of data literacy skills across campus.

6.2. Opportunities

There is a potential for collaboration to integrate data literacy across disciplines at Lincoln University. The interdisciplinary nature of data literacy invites faculty from across programs to design a new and/or revise the existing curriculum to incorporate this skill. Establishing data literacy as an interdisciplinary language between programs across campus is very important. To that end preparation of common data literacy assessment rubrics is imperative. Since information literacy and data literacy are a closely intertwined skills, involving other staffs such as librarians, IT personnel, CETL and etc. can be leveraged to strengthen the discussion around this topic. Preparing workshops and trainings that are focused on this subject involving students, faculty and staff is possible. This provides opportunities for access to ideas and recourses related to data literacy.

7. Limitations and Future Work

This project is limited to compiling a knowledge synthesis report on the topic of data literacy. Future work in this area will focus on:

- Research project on measuring data literacy – for example, case study at LU
- Write a grant to secure funding for data science boot camp for faculty and staff across disciplines

This project will serve as a first step to initiate discussion and cultivating interest among faculty members at Lincoln University in the need and importance of reinforcing data literacy in our curriculum. Cultivating such interest and investment in the broader data science field from faculty spanning all academic divisions will be critical in the successful integration of data literacy in to our curricula. To this end, through the data science boot camp/workshop series, the participating faculty and staff at LU:

- Will learn about the field of data science and how it could become part of the curriculum by hearing from guest speakers who use data science in their research and teaching;
- Will complete data science education modules designed to teach the basics of coding, data acquisition, data cleaning, and data visualization at an elementary level;
- Will begin work on a project in their field that involves a data science approaches.

8. References

[1] M. Frank, J. Walker, J. Attard, and A. Tygel, “Data Literacy - What is it and how can we make it happen?,” *J. Community Informatics*, vol. 12, no. 3, pp. 4–8, 2016, doi: 10.15353/joci.v12i3.3274.

[2] Kenneth Cukier and Viktor Mayer-Schoenberger, “The Rise of Big Data:How it’s Changing the Way We Think about the World.,” vol. 92, no. 3, pp. 28–40, 2013, [Online]. Available: https://heinonline.org/HOL/P?h=hein.journals/fora92&i=592.

[3] I. Sander, “What is critical big data literacy and how can it be implemented?,” *Internet Policy Rev.*, vol. 9, no. 2, pp. 1–22, 2020, doi: 10.14763/2020.2.1479.

[4] H. Bowne-Anderson, “Your Data Literacy Depends on Understanding the Types of Data and How They’re Captured,” *Harv. Bus. Rev.*, 2018, [Online]. Available: https://hbr.org/2018/10/your-data-literacy-depends-on-understanding-the-types-of-data-and-how-theyre-captured?utm_source=linkedin&utm_campaign=hbr&utm_medium=social.

[5] Marr B., “How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read,” *Forbes*, pp. 1–6, 2018.

[6] .., “How Much Information Is There in the World?,” *Inf. Technol. J.*, vol. 10, no. 5, pp. 1067–1067, 2011, doi: 10.3923/itj.2011.1067.1067.

[7] C. Miller, R. Coldicutt, and H. Kitcher, “People, Power and Technology: The 2018 Digital Understanding Report,” pp. 1–32, 2018, [Online]. Available: http://understanding.doteveryone.org.uk/files/Doteveryone_PeoplePowerTechDigitalUnderstanding2018.pdf.

[8] J. Redden, “Democratic governance in an age of datafication: Lessons from mapping government discourses and practices,” *Big Data Soc.*, vol. 5, no. 2, pp. 1–13, 2018, doi: 10.1177/2053951718809145.

[9] Cathy O’Neil, “Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy,” *Vikalpa J. Decis. Makers*, vol. 44, no. 2, pp. 97–98, 2019, doi: 10.1177/0256090919853933.

[10] C. Kuhlman, L. Jackson, and R. Chunara, “No computation without representation: Avoiding data and algorithm biases through diversity,” *arXiv Prepr.*, 2020, [Online]. Available: http://arxiv.org/abs/2002.11836.

[11] Z. Zhang, T. Yamamoto, and K. Nakajima, “Development of Education Curriculum in the Data Science Area for a Liberal Arts University,” *EasyChair Prepr.*, no. 8713, 2022.

[12] B. Y. E. Stevens, “The 4 Types of Data Analysis [Ultimate Guide],” pp. 1–12.

[13] Microsoft Power BI, “What is data storytelling ? The benefits of data storytelling Making sure your data story is valuable Using data visualization to enhance your data storytelling.” https://powerbi.microsoft.com/en-us/data-storytelling/#:~:text=Data storytelling is the concept,and inform a particular audience.

[14] Jen Dubois, “Thriving Data Culture Starts with Data Literacy Assessments,” 2022. https://quanthub.com/data-literacy-assessment/.

[15] B. J. Sickler, E. Bardar, and R. Kochevar, “Measuring Data Skills in Undergraduate,” vol. 50, no. 4, pp. 25–32, 2021.

[16] “The Data Literacy Score- A Team Based Assessment,” 2022. https://dataliteracy.com/data-literacy-score/.

[17] Data To The People, “THE GLOBAL DATA LITERACY BENCHMARK”.