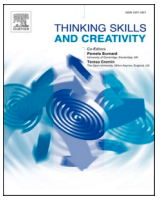
Thinking Skills and Creativity 48 (2023) 101311

Contents lists available at ScienceDirect

Thinking Skills and Creativity

journal homepage: www.elsevier.com/locate/tsc



Teaching Data Science through Storytelling: Improving Undergraduate Data Literacy

You Li [a,*], Ye Wang [b], Yugyung Lee [c], Huan Chen [d], Alexis Nicolle Petri [e], Teryn Cha [f]

[a] School of Communication, Media & Theatre Arts, Eastern Michigan University, Ypsilanti, MI, United States
[b] Department of Communication and Journalism, University of Missouri-Kansas City, Kansas, MO, United States
[c] School of Science and Engineering, University of Missouri-Kansas City, Kansas, MO, United States
[d] College of Journalism and Communications, University of Florida, Gainesville, FL, United States
[e] Office of Research Development, University of Missouri-Kansas City, Kansas, MO, United States
[f] Division of Mathematics, Engineering Technologies & Computer Sciences, Essex County College, Newark, NJ, United States

ARTICLE INFO

Keywords:
Cooperative/collaborative learning
Data science applications in education
Improving classroom teaching
Interdisciplinary projects
Teaching/learning strategies

ABSTRACT

This study proposes and evaluates the OCEL.AI (Open Collaborative Experiential Learning. AI) paradigm that aims at broadening participation in data science education and enhancing undergraduate students' data literacy. The core of the paradigm is the "Tell Stories" approach. This approach applies the 5W+1H (Who, What, When, Where, Why, and How) conceptual schema of stories as a transdisciplinary language for data science education for STEM and non-STEM majors. Accordingly, this study reported findings from the OCEL.AI project that implemented and evaluated the paradigm. A field experiment, in addition to classroom observations, was conducted to compare the learning outcomes of students in data science competence, appreciation, career motivation, life-long willingness to learn, and self-efficacy in data science between the treatment group and the control group. The results showed that the OCEL.AI paradigm improved undergraduates' data science competence and career motivation despite majors or gender.

1. Introduction

Accumulation of administrative, behavioral, and social data from human interaction with digital devices and platforms proliferates datafication, which aims to quantify aspects of human life (Fernández-Rovira et al., 2021; Flensburg & Lomborg, 2021; Lavorgna & Ugwudike, 2021). Data have never been so abundant and yet so accessible. Data literacy, following information literacy, digital literacy, and computer literacy, has become a prerequisite for participating in modern professional life. Data literacy encompasses abilities to find relevant data from credible sources, evaluate and prepare data, analyze data, and interpret data (Wolff et al., 2016).

More importantly, data science education must bridge the gaps in data literacy among the population since data have become a currency of power (D'Ignazio, 2017) in the 21st century. Data science educators need to broaden the participation and reach the widest learners, including undergraduate or graduate students in the STEM fields as well as the non-science fields such as liberal arts, social sciences, and humanity programs (Betz et al., 2020; Cardenas-Navia & Fitzgerald, 2015; Kross et al., 2020). The challenge, however, is to create a "transdisciplinary" language to communicate with users and learners of diverse backgrounds and perspectives (Flensburg & Lomborg, 2021).

* Corresponding author.
E-mail address: yli23@emich.edu (Y. Li).

https://doi.org/10.1016/j.tsc.2023.101311
Received 30 November 2021; Received in revised form 1 May 2023; Accepted 5 May 2023
Available online 6 May 2023

The solution presented in this study is the Open Collaborative Experiential Learning (OCEL.AI) and its "Tell Stories" approach. OCEL.AI is an interdisciplinary project of teaching data science to students in STEM and non-STEM fields. The OCEL.AI paradigm consists of five teaching modules (i.e., life, data, model, user, and societal stories) that are centered on the "Tell Stories" approach. The "Tell Stories" approach draws upon the 5W+1H (who, what, when, where, why, and how) conceptual schema. Students, regardless of educational and technical backgrounds, can experience a complete journey of data science: from identifying an issue in real life, partitioning, gathering data of 5W+1H from open sources, analyzing, and developing applications to solve a real-world issue ethically. The issue-focused and solution-based approach to processing and interpreting data embeds learning in a social context that has real-life relevance and implication for the learners (Wilkerson and Polman, 2020).

This paper reports the first round of findings from a series of studies that assess the effectiveness of the OCEL.AI paradigm in enhancing undergraduate students' data literacy. It summarizes the quantitative and qualitative findings from students of computer science and non-computer science majors enrolled in four universities and compares the self-reported data literacy outcomes between those who were exposed to OCEL.AI and those who were not. The findings reveal that the "Tell Stories" approach to data science is promising in enhancing data literacy for both STEM and non-STEM students.

2. Literature review

2.1. Data science education

Data science education trains data specialists and cultivates data literacy among citizens (Pedersen & Caviglia, 2019). According to Loukides (2010), data science is "a data application that acquires its value from the data itself and creates more data as a result." Waller and Fawcett (2013) consider data science as an application of quantitative and qualitative methods to solve relevant problems and predict outcomes. From the disciplinary perspective, data science synthesizes and builds on statistics, informatics, computing, communication, management, and sociology to study data and its environments to transform data into insights and decisions by following data-to-knowledge-to-wisdom thinking and methodology (Cao, 2017). As a result, data science has become increasingly popular and relevant to not only science fields but also non-science domains such as law, history, nursing (Clancy et al., 2014), media, and entertainment (Gold et al., 2013). The interdisciplinary nature of data science education has made defining the boundary of the field difficult (Cassel & Topi, 2015).

2.2. Data literacy

Data literacy is a set of important student learning outcomes for data science education. It broadly describes the ability to use data as part of everyday thinking and reasoning for solving real-world problems (Wolff et al., 2016). Those abilities include collecting, examining, analyzing, and interpreting data to make informed decisions (Frank et al., 2016; Mandinach & Gummer, 2013). Emerging from information literacy, a generic approach to data literacy education teaches students to be aware of, access, engage, manage, communicate, preserve, and ethically use data (Maybee & Zilinski, 2015). Dichev and Dicheva (2017) posit that a data-literate individual should be able to collect, evaluate, analyze, and interpret data, present derived results, and take ethically sound actions. The common cognitive skills of data literacy include identifying, collecting, selecting, cleaning, analyzing, interpreting, critiquing, visualizing, and sharing data (Calzada Prado & Marzal, 2013; Mandinach & Gummer, 2013; Wolff, Gooch, Montaner, Rashid, & Kortuem, 2016). Data literacy also develops critical thinking abilities, such as critically assessing data and their sources and recognizing inappropriate interpretations or misusage of data (Frank et al., 2016; Koltay, 2014). Wolff et al. (2016) identified a set of competencies for data literacy after reconciling nine definitions of data literacy and four definitions of statistical literacy. The framework is called the PPDAC model standing for Problem, Plan, Data, Analysis, and Conclusion. The PPDAC model outlines the competencies of data literacy, including asking questions from data, developing hypotheses, identifying potential sources of data, collecting, or acquiring data, analyzing and creating explanations for data, evaluating the validity of explanations based on data, and formulating new questions. Compared to previous conceptual definitions of data literacy, the PPDAC model adds ethical understanding and usage of data into the inquiry process and uses an experiential learning approach that situates data literacy learning within a real-world context. Ridsdale et al. (2015) defined data literacy as the ability to collect, manage, evaluate, and apply data critically. They identified 32 competencies and 42 skills, knowledge, and expected tasks of data literacy described in previous studies and categorized them into conceptual competencies, core competencies, and advanced competencies. Based on the previous studies, we have identified seven items to measure data science competence that combines the cognitive and critical thinking dimensions of data literacy (see the method section for specific items).

2.3. The challenges for data science education and data literacy

The first challenge facing data science education is a lack of women and minorities in STEM fields (Neuhauser, 2015). Data science education grows fastest in engineering and computer science (Cardenas-Navia & Fitzgerald, 2015). The best sources of new data science talent are students of computer science (18.3%), data science and analytics (20.7%), statistics and mathematics (16.3%), engineering (10.8%), and natural sciences (10.6%); students in the non-STEM fields, such as economics and social sciences, only made up about 12.4% (365 data science, 2020). More than 70% of the surveyed workforce in data science is male (365 data science, 2020). The gender gap in data science reflects the lack of women in computer and mathematical occupations (26.7%) and the engineering workforce (16.1%) (U.S. Bureau of Labor Statistics, 2023). Only 16.9% of the engineering and architecture workforce and 17.7% of the

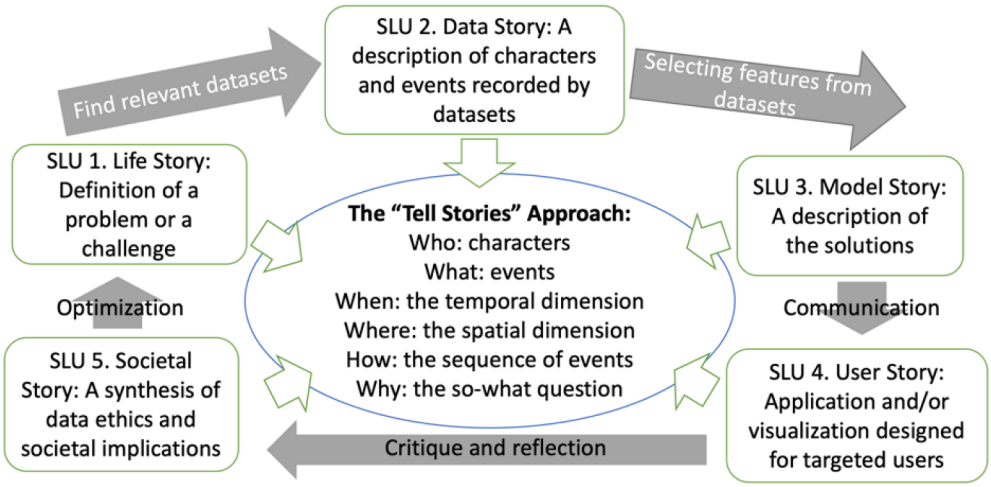


Fig. 1. The OCEL.AI paradigm with the "Tell Stories" approach as the centerpiece.

computer and mathematical sciences workforce are African American and Hispanic (U.S. Bureau of Labor Statistics, 2023). The concentration of data science education in STEM disciplines creates racial and gender gaps in data literacy. Broadening participation in data science education requires more disciplines to adopt and innovate data science curricula.

The second challenge is the incompatibility between the interdisciplinary nature of data science education and the current disciplinary structure of higher education (Faris et al., 2011; Tang & Sea-Lim, 2016). For example, data science courses in the STEM fields are offered at upper-level undergraduate programs or graduate programs, limiting students' exposure to data science in the early days of a college education. Among the 632 programs in data science, analytics, and related fields listed on the website http://datasicence.community/colleges, only 68 programs (10.8%) offer a bachelor's degree, and the vast majority are master's degree and certificate programs (Swanstrom, 2023). Data journalism, one of the non-STEM disciplines that have incorporated data since the 1970s, suffers from a shortage of academically trained instructors to lead and /or teach such interdisciplinary programs (Heravi, 2019). Advanced techniques such as coding and programming were mostly absent from the curricula (Davies & Cullen, 2016; Heravi, 2019). A census of 369 U.S. journalism programs found that statistics was not required in 79% of the programs, and none offered its own statistics course (Martin, 2017). These results, juxtaposed with former calls on the importance of statistics and data analysis in journalism (Nguyen & Lugo-Ocando, 2016) , demand immediate attention to integrating more data analytics courses in more disciplines in higher education programs.

To address these challenges, educators across disciplines have proposed various frameworks to blend the technical aspects of data with domain knowledge and extend data literacy education to more disciplines. Maybee and Zilinski (2015) proposed a data-informed learning framework for data literacy for use in higher education. The principles of data-informed learning emphasize building on students' prior experiences of using data, learning to use data while learning disciplinary context, and learning beyond classrooms in professional and personal settings. D'Ignazio (2017) proposed the term creative data literacy targeted at students in the non-STEM fields and argued that data literacy should involve not only the acquisition of technical skills but also empower the learners to make the world a better place. Ridsdale et al. (2015) identified some best practices for teaching data literacy, including explaining the benefits of data and data skills upfront, hands-on learning, module-based learning, project-based learning that has real-world applicability, using real-world data, and collaboration between educators, organizations, and institutions to ensure goals are being met by all stakeholders. Yavuz and Ward (2020) introduced several types of computational tools to Purdue undergraduates from different disciplines who learn data science through real-life examples and team-oriented projects at an early point in their education.

2.4. The proposed solution: the OCEL.AI paradigm and the "Tell stories" approach

The gap that the proposed solution attempts to bridge is the adaptability of data science education within the current disciplinary structure. The targeted student learning outcomes are dimensions of data literacy. The theoretical origin of "Tell Stories" is in mass communication, a body of scholarly studies on maximizing the reach and effectiveness of communication. The motivation is simple: adapting mass communication strategies to broaden the reach and participation of data science education.

By reviewing literature in mass communication, we identified stories/narratives as the "effective appeal" to a broader "audience" of data science. Stories/narratives can be a useful tool for data science education (Boldosova, 2019). Research on persuasive strategies shows that narrative appeals are most likely to work for complex subjects, for example, cancer-related information (Segel & Heer, 2010; Thompson & Haddock, 2012). Stories can vividly illustrate risks, consequences, measures, various procedures, etc., involved in the prevention, detection, and treatment of medical conditions (Thompson & Haddock, 2012). Also, narrative appeals are particularly persuasive for individuals who have a strong motivation to seek out and become involved in emotional situations or engage in and enjoy critical thinking (Thompson & Haddock, 2012). The evidence from effective education suggests that narratives (stories) have

Table 1
The storytelling framework.

Steps	5W+1H
A life story: Guiding students to identify a problem or a challenge facing specific individuals, a group of people, or a community.	**Who** are the people or communities in need of help? **What** happened to them? **When** and **where** did it happen? **Why** and **how** did it happen?
A data story: Using story as hands-on practice to link large-scale datasets with real-world problems.	**Who** was sampled? **Who** was over-sampled or under-sampled? **What** events/behaviors were recorded? **When/where** were the data collected? **Why/how** were they collected?
A model story: Select, justify, and conduct experimentation on machine learning (ML) algorithms.	**Who:** To what extent is ML dependent on human "guidance?" **What** ML algorithms are useful to what kind of stories? **When:** how efficient is the modeling/algorithm? **How** does the number of training examples influence accuracy?
A user story: Deploy the models via application/visualization in a useful way to end-users.	**Who** are the intended end-users? **What** can the application do for them? **When/where** will it be deployed? **Why** is it useful?
A societal story: Identify the impacted population, and evaluate ethical, social, and cultural implications.	**Who** will be impacted? **What** are the implications for privacy, security, and fairness? **When/where** will the impact take place?

great potential as a learning tool for advanced subjects studied by adult learners.

Accordingly, the centerpiece of the OCEL.AI paradigm is to "tell stories." This approach consists of "stories" and "storytelling." "Storytelling" is the action of presenting "stories". A "story" includes one or more entities (who), being involved in a sequence of actions/events (what and where) unfolding over time (when) (Jarke & Macgilchrist, 2021). One common "storytelling" in data analytics is composing narrative texts to accompany data visualization. However, narrowly defining storytelling in such a way ignores the conceptual schema that is essential to a "story." Weber (2020) pointed out the key components of a "story": (1) Spatial and temporal dimensions entailing where and when; (2) characters and events, encompassing who and what; (3) sequentiality, meaning the sequence of events; (4) tellability that answers the so-what question. Components (3) and (4) establish the causal relationship between events and explain how things happen and why things happen. This definition of "story" apparently has little to do with fiction. Rather, at its core, a "story" establishes facts based on knowledge discovered from data (Jarke & Macgilchrist, 2021). For end goals, storytelling elevates the value of "stories" via its persuasiveness for social changes (Jarke & Macgilchrist, 2021).

Adopting this definition of "story," the OCEL.AI paradigm (see Fig. 1) integrates storytelling in the five steps of learning data science to tell: 1) a life story, 2) a data story, 3) a model story, 4) a user story for users of applications or data visualizations, and 5) a societal story about ethics and evaluations. The paradigm operationalizes Weber (2020) story schema as "5W+1H" to guide learning in every step (see Table 1): Who is it about? What happened? When did it take place? Where did it take place? Why did it happen? How did it happen? The last two questions incorporate social and ethical considerations into every step of the entire learning experience and challenge students to think about the so-what question about the use case, the data, the model, the application, and the visualization. The last step is a reflection on model evaluation, data ethics, and social implications.

Accordingly, the OCEL.AI paradigm has five Student Learning Units (SLUs). Depending on the scope of the course, instructors may choose to implement all five SLUs for projects in a capstone or thesis course or emphasize a subset of the five SLUs as incremental steps toward a project in a topic course (see Fig. 2). For instance, a course on machine learning may emphasize the data story, the model story, and the societal story, while a course on the web or mobile application may emphasize the life story, the user story, and the societal story. Non-computer science courses, such as data journalism, may emphasize the life story, the data story, the user story with a focus on data visualization, and the societal story. This modularized design gives instructors the flexibility to highlight certain phases in the life cycle of data science and meanwhile provides a roadmap of the end-to-end process for instructors and students to contextualize project-based learning. Two specific examples are provided in the Method section to demonstrate the flexibility and adaptability of the model.

The OCEL.AI paradigm asks students to understand why and how to use data and explore potentially useful information and structures in data, with the consideration of target users and their potential use in their contexts. The "Tell Stories" approach guides students to organize data gathering, preparation, analysis, and presentation in a 5W+1H framework, which smoothens the integration of data science into non-STEM curricula, such as advertising, communication, journalism, arts, sociology, and law. The 5W+1H framework also demonstrates its effectiveness to process complex data in the STEM fields ((Almeida et al., 2020; Chakma & Das, 2018). Teaching data science to the non-STEM fields bridges the "big data divide" (Andrejevic, 2014) between the small class of data specialists who control and prepare data and the general public who rely on data to navigate and manage their lives. By equipping them with the OCEL.AI paradigm, instructors can coach data science professionals, improve the diversity of the STEM fields, and educate data-literate citizens.

2.5. The OCEL.AI project

During the 2020–2021 academic year, the OCEL.AI paradigm was integrated into six computer science courses and three journalism and communication courses across four campuses. To assist learning, the OCEL.AI project includes a learning platform deployed as a website. This platform offers examples, case studies, and learning tools. The website embeds dataset searching, data visualization tools, and two predictive machine learning models accessible via a non-coding interface for non-STEM students, and a coding interface for STEM students. Meanwhile, teaching resources and workshops were provided to non-STEM teachers who wanted to introduce the conceptual knowledge of data science while improving their competencies in problem-solving, critical thinking, motivation, and appreciation of data literacy.

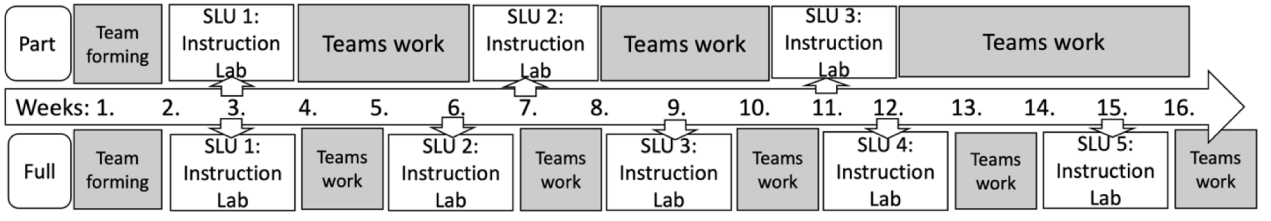


Fig. 2. The implementation of OCEL.AI throughout a semester.

Depending upon the subject of the class, students were tasked with different components of the OCEL.AI paradigm. For example, the students from the Web/Mobile programming course were tasked to create a Web/Mobile app and tell life stories, data stories, user stories, and societal stories. Students from digital advertising classes were tasked to compose a marketing plan for a mobile app and write life stories, data stories, user stories, and ethical and societal stories. The following sections exhibit a control vs. treatment group field experiment that measured the effectiveness of the OCEL. AI paradigm. Two case studies of observation of OCEL.AI in different disciplinary contexts were reported as well. As part of a large and ongoing project, this paper primarily focuses on the quantitative analysis of the effectiveness of OCEL.AI in increasing undergraduate students' confidence, motivation, and appreciation of data literacy. The following research question guided the analysis:

RQ: How did the OCEL.AI paradigm distinguish the students' learning outcomes of data literacy between the control group and treatment group?

3. Method

The study consisted of two parts: classroom observations and a field experiment. A total of 221 undergraduates at four universities participated in the study in Fall 2020 and Spring 2021; two of the universities were in the Midwest, one in the Northeast, and one in the Southeast of the United States.

3.1. Classroom observation

The classroom observation was conducted by the instructors in three courses: one non-STEM course and two STEM courses. All courses were part of an end-of-semester competition of data science projects.

Class 1 Journalism Advanced Reporting: A non-STEM course

Journalists are good storytellers, but very few of them are competent mathematicians or programmers. The journalism program at a participating university recognizes the value of telling stories about data or applying data to storytelling but is constrained by a lack of resources and faculty expertise to build an adequate data journalism curriculum. In Spring 2021, the OCEL.AI paradigm was incorporated into the capstone course Advanced Reporting. Twenty junior and senior journalism students who had minimum knowledge of data journalism or data science completed the five SLUs of OCEL.AI. The class project is a data journalism report and a proposal for a data-driven application. The following paragraph outlines each student learning unit.

SLU1: Life Story. Students identified a real-world issue and explained its newsworthiness, target audiences, and potential social impact. Students could contextualize personal experiences and observations in a larger societal context and practice their news value judgment.

SLU2: Data Story. Students retrieved a relevant dataset from open sources accessible from the OCEL.AI website; they completed a data attribute map and analyzed the descriptive patterns of the data. Students could gather, research, and analyze data to illustrate the proposed issue.

SLU3: Model Story. Students proposed a machine-learning modeling plan (see Fig. 7) to turn human analytical tasks into programming tasks. Students could conceptually understand and communicate various models to data scientists.

SLU4: User Story. Students proposed a computer-automated solution (i.e. application), described the functions, characteristics, and innovations of the solution, and developed a desktop visual demo of the proposed application. Students could synthesize and communicate their solutions to a larger audience (users) in textual, numerical, and visual formats.

SLU5: **Societal Story**. Students explained the target users, benefits, and social and cultural implications of the solutions to various stakeholders. Students could appreciate and reflect on the application of data science to real-world problem-solving.

Classes 2&3 Python and Deep Learning Programming & Big Data Programming: STEM courses

The OCEL.AI paradigm was introduced in the curriculum of data science classes, e.g., Python and Deep Learning Programming (PDLP) (about 30 students) and Big Data Programming (BDP) (about 30 students). The class project for PDLP was to train machine learning/deep learning models (ML/DL), and BDP, to create a website or a mobile application. The students were asked to write their project description following the OCEL.AI framework to "Tell Stores:"

SLU1: Life Story. Students learned about the principles and methods of writing narratives to contextualize real-life problems. Students could convert technical ideas into solving a real-life problem and develop abilities in communication.

SLU2: Data Story. Students in PDLP described the process of data collection and data transformation. Students in BDP did not necessarily need data for application development, and thus this module was not required.

SLU3: Model Story. Students in PDLP used ML/DL to identify patterns in the data and make predictions. The statistical and machine learning methods used included multiple linear regression, K-Nearest Neighbors (KNN), Naïve Byes, Support Vector Machines (SVM), K means clustering, decision tree, random forest algorithms, topic modeling, sentiment analysis, image process and classification. The output is a predictive model for hidden insights and recommended solutions. Students in BDP did not necessarily need any model for application development, and thus this module was not required.

SLU4: User Story. Students in PDLP visualized the modeling results or deployed the model via a web or application interface. Students in BDP created websites and mobile applications for certain types of users. All students were required to think from a user's perspective and communicate the solution effectively to the targeted users.

SLU5: Societal Story. Students discussed the relationship between data, models, and society. Topics included but were not limited to data security (e.g., ethical data collection, consenting, and cyber security), privacy (e.g., HIPAA, differential privacy, and federated learning), data bias (e.g., debiasing, finetuning, and domain adaptation in Natural Language Processing, long-tail prediction, and

minority categories), and model bias (e.g., facial recognition and policing, and predictive models for a sentencing recommendation).

The StoryHack Competition.

Near the end of the spring semester of 2021, students in all nine courses were encouraged to participate in StoryHack, a cross-campus and interdisciplinary competition where non-STEM students submitted their story proposals and solutions, and STEM students developed and implemented machine learning modeling and applications based on the non-STEM students' proposals. Seven finalist groups presented their projects virtually to practitioners and professors who evaluated the projects' impact, functionality, quality, innovation, and presentation. The competition offered winning teams scholarships ranging from $250 to $1000. The StoryHack provided an opportunity for STEM students and non-STEM students to work alongside and reassured the feasibility of OCEL.AI for both STEM and non-STEM students.

3.2. Field experiment

A field, quasi-experimental design was employed, with 106 students in the treatment group and 115 students in the control group. The recruitment of participants took a few steps. In Fall 2020, an OCEL.AI teacher training workshop was conducted. During the three-day training, the "Tell Stories" approach was introduced to STEM and non-STEM instructors who were interested in data science education, and the ocel.ai website was unveiled for testing. The workshop participants were contacted afterward for implementing and evaluating the OCEL.AI paradigm in their courses the following semester. The instructors who opted to implement the paradigm recruited their students to reflect on their learning outcomes in a questionnaire as the treatment group, and those instructors who opted out of implementation recruited their students to take the same questionnaire as the control group. In Spring 2021, nine courses from four universities that received the federal grant implemented OCEL.AI, among which three courses were for non-STEM majors (i.e. journalism, advertising, communication), and six courses were for STEM majors (i.e. computer sciences and data science). One questionnaire was distributed to the treatment group of nine classes around the end of the semester while an identical copy of the questionnaire was distributed to the control group of classes. An invitation e-mail with a link to the questionnaire was distributed first, followed by two rounds of reminders. Voluntary students in both treatment and control groups participated in the questionnaire and received extra credits from their instructors, approved by the IRB offices.

3.3. Quantitative measurements

The questionnaire evaluated learning outcomes from data literacy (Pedersen & Caviglia, 2019; Wolff et al., 2016) and data journalism education (Bradshaw, 2018; Zhu & Du, 2018). Although rooted in different disciplines, the two fields share some desired learning outcomes in data acquisition, analysis, and ethics. In addition, we also assessed students' dispositional competencies such as a life-long willingness to learn, career motivations, self-efficacy, and appreciation of data science.

The questionnaire consists of six parts. The first five parts measure five learning outcomes, and the last part includes demographic questions inquiring about students' majors, genders, and ethnicities, and whether they've taken a data science course before. Most of the learning outcomes were measured by multiple items for reliable measurements. Item analysis was conducted on multi-item outcome measurementsto select the most reliable items. The averages of items were used to measure a multi-item learning outcome. Cronbach's Alpha was calculated and reported below for multi-item outcomes:

Data Science Competence (5-point Likert scale with 1 being extremely incompetent and 5 extremely competent. Cronbach's Alpha = 0.921): "I can identify potential and appropriate sources of data;" "I can clean data in datasets;" "I can identify methods to analyze data to answer a question;" "I can visualize data for the target audience;" "I can make recommendations based upon data analysis results;" "I can tell data-related stories to specific audiences;" "I can use data to help raise public awareness of social issues."

Self-Efficacy (5-point Likert scale. Cronbach's Alpha = 0.914): "I believe I can earn a grade of "A" in a data science course in my field." "I believe I can master data science knowledge and skills;" "I am confident I can do well on a data science test;" "I am confident I can do well in a data science lab."

Career Motivation (5-point Likert scale. Cronbach's Alpha = 0.884): "I learn data science because it may give me an advantage in my career, such as internship, jobs, additional benefits, etc.;" "I learn data science because it provides more tools/resources to professionalize my work."

Appreciation (5-point Likert scale. Cronbach's Alpha = 0.868): "Data science is relevant to my future;" "Data science deals with things I am concerned about;" "Data science helps me to make decisions about important things in my field;" "Data science helps me understand important issues in my field."

Life-long willingness to learn (5-point Likert scale. Cronbach's Alpha = 0.908): "I am willing to explore more resources about data science beyond classroom requirements by myself;" "I am willing to take more lessons and training on data science to deepen my understanding and skillsets;" "I want to learn new things that I didn't know about data science."

Importance of storytelling (5-point Likert scale. Single-item measurement, and thus no Cronbach's Alpha): "I think data science needs to tell good stories."

Relevance of storytelling (5-point Likert scale. Single-item measurement, and thus no Cronbach's Alpha): "I think storytelling has nothing to do with data science." (reverse coded)

3.4. Data analysis

Cases that had more than 50% missing values were deleted case-wise. Then, series means were applied to replace missing values for

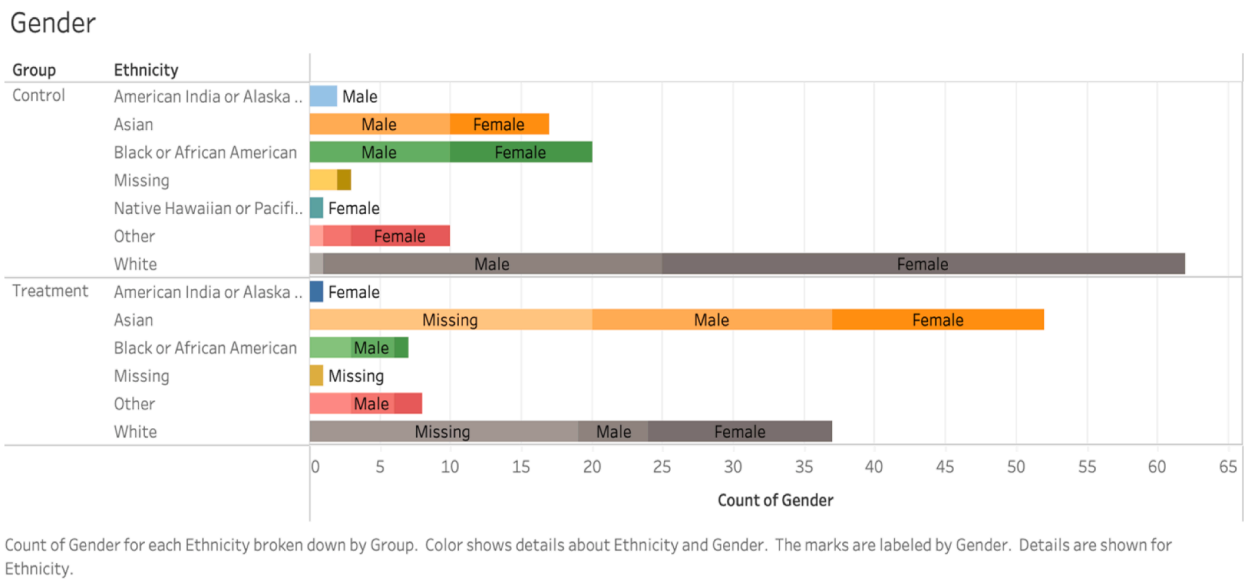


Fig. 3. Student ethnicities and genders.

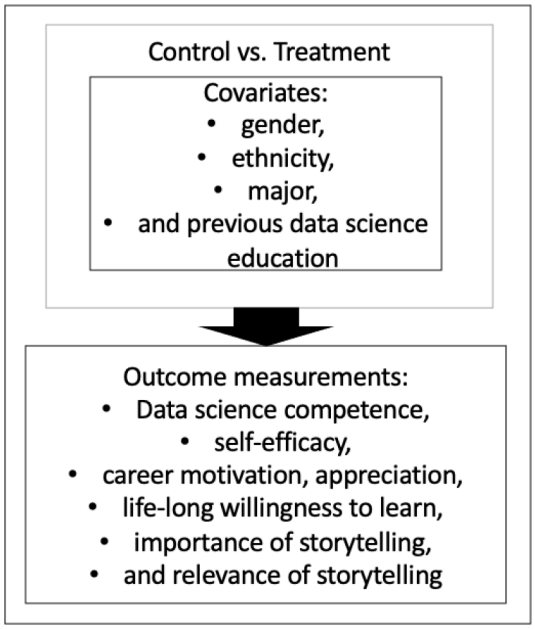


Fig. 4. The summary of the models.

continuous variables. Univariate outliers were deleted based on z scores ($z>=3.29$ or $z<=-3.29$). After data cleaning, the treatment group had 115 cases and the control group had 106 cases. Students' majors were recoded into STEM majors ($N = 101$) and non-STEM majors ($N = 120$). Fig. 3 shows ethnicity and gender information:

The normality of independence of observations was checked and met; the skewness and kurtosis were within the range of −1 and 1. The homogeneity of variance/covariance was checked and met (Box'M = 42.84, $F = 1.12$, $p = 0.282$). Levene's tests on the dependent variables were also not significant. The results showed that the effect of OCEL.AI exposure was significant, Wilk's Lambda = 0.858, F (14, 156) = 1.84, $p=<0.5$, multivariate $\eta^2 = 0.142$. Multivariate Analysis of Covariance (MANCOVA) was conducted using the General Linear Model program in SPSS 27.0. This function in SPSS allows users to customize the models and outputs for univariant comparisons as well as multi-variant comparisons. The analysis compared the means of the control group and the treatment group in terms of students' learning outcomes in data literacy: data science competence, self-efficacy, career motivation, appreciation, life-long willingness to learn, the importance of storytelling, and relevance of storytelling. The covariates that might influence the main effects of

Table 2
The main effects of control vs. treatment groups on students' learning outcomes.

Student learning outcomes	Control (Mean)	Treatment (Mean)	Type III sum of squares	F	Sig.
Data science competence	3.7	4.11	12.4	15.268	0.000
Life-long willingness to learn	4	4.26	2.314	2.951	0.087
Appreciation	4	4.29	2.153	2.326	0.129
Career motivation	4.07	4.42	7.527	8.427	0.004
Self-efficacy	3.78	4.09	1.427	1.824	0.178
Importance of Storytelling	2.68	4.03	1.981	1.994	0.159
Relevance of Storytelling (reverse coded)	3.2	3.11	1.172	1.191	0.276

Table 3
Significant main effects of the co-variates on students' learning outcomes.

Student learning outcomes	Co-variates	Mean	Type III sum of squares	F	Sig.
Data science competence	Prior data science education: Yes	4.14	10.258	12.631	0.000
	Prior data science education: No	3.73			
Appreciation	Prior data science education: Yes	4.33	4.584	4.592	0.027
	Prior data science education: No	4.01			
Self-efficacy	Prior data science education: Yes	4.29	12.965	16.586	0.000
	Prior data science education: No	3.69			
Life-long willingness to learn	Prior data science education: Yes	4.34	4.468	5.697	0.018
	Prior data science education: No	3.97			
	STEM major	4.42	7.882	10.049	0.002
	Non-STEM major	3.88			

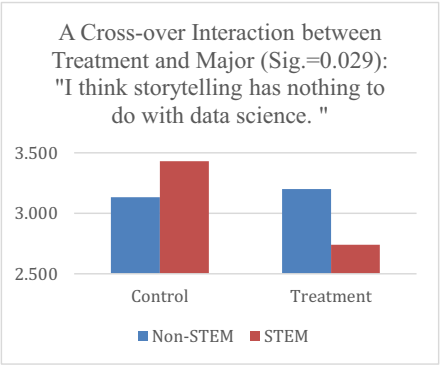


Fig. 5. Significant interaction effects on relevance of storytelling to Data Science.
Note: the item is reverse-coded.

control vs. treatment included gender, ethnicity, major, and previous data science education. Their potential main and interaction effects were also analyzed. Fig. 4 gives an overview of the models.

4. Results and discussion

4.1. Main effects of the treatment

First, we assessed the effectiveness of the storytelling approach in aiding students' data literacy learning outcomes. Students in the treatment group reported significantly higher average scores on their competence in data science skills and motivation to pursue a career in data science than students in the control group (Table 2).

4.2. Main effects of the co-variates

Next, we examined the effects of the covariates on students' reported learning outcomes. Among the covariates, the students' major (STEM or. non-STEM) and their prior knowledge in data science had a significant effect on a few learning outcomes (see Table 3). STEM majors reported significantly higher average scores on long-term willingness to learn data science. Students having prior data science education reported significantly higher average scores on their competency of data science skills, appreciation of data science, and self-efficacy in their abilities to learn data science well.

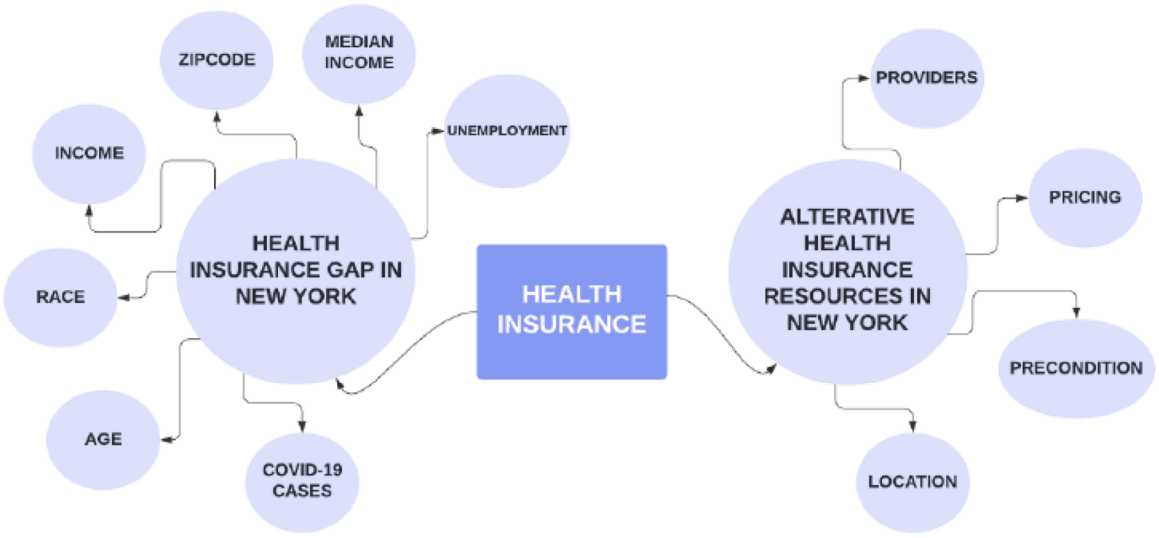


Fig. 6. A sample data attribute map from students' work.

These findings indicate that it is students' domain knowledge and experience in data science (acquirable), not students' gender or ethnicity (not acquirable), that would make a difference in students' data literacy. Exposure to data science early on is important to students' future learning. Data science education can and should be extended beyond the graduate level and to also students in the non-STEM fields so that more students could access, acquire, and appreciate data literacy.

4.3. Significant interaction effects

There was also a significant cross-over interaction between the treatment and students' majors on the perceived relevance of storytelling to data science ($F = 4.82$, $p = 0.029$) (see Fig. 5). Among STEM majors, the treatment led to a decrease in the perception that storytelling was irrelevant to data science. This finding reveals greater potential for storytelling in data science education among STEM majors. Communication, journalism, advertising, and media studies majors emphasize storytelling in their curricula and hence students see storytelling as an integral part of learning. By contrast, computer science students often are conditioned to quantitative reasoning. The OCEL.AI paradigm can integrate presenting, communication, and persuasion into their learning.

4.4. Classroom observation

Besides students' self-reported data, our empirical findings were also grounded in classroom observations and students' work. By examining students' performance, we notice a few shifts of learning in the treatment group, exemplified by the following prize-winning project pulled together by both journalism and computer science students.

A major shift for students involved understanding data not as abstract numbers but as contextualized resources that have real-world meanings and implications. For instance, a team of journalism students analyzed the descriptive statistics of insurance status and COVID-19 cases in New York State and uncovered that the underserved communities in need of medical help during the pandemic were often those without health insurance. Those uninsured people are more likely to be non-Caucasian, female, without a higher level of education, and jobless. The traditional computer-assisted reporting approach in data journalism would then encourage students to form story ideas from here to question why and how the insurance gap may affect individual lives and public health by interviewing various sources. However, the OCEL.AI paradigm doesn't end at identifying problems but also empowers them to propose solutions to the problems. The students further explored whether there are open data sources about insurance types, pricing, providers, and locations as potential alternative insurance sources.

The second major shift for students, especially those in the journalism course, involved transforming from a problem-centric storytelling mindset to a solution-oriented engagement. For instance, the journalism students proposed to resolve the health insurance gap by helping uninsured people find alternative and affordable insurance resources in New York. Below is an attribute map (see Fig. 6) where they envisioned the kind of data that the computing would need to provide the necessary information and proposed a machine learning modeling plan (see Fig. 7) to outline how a computer program could process the data points and find insurance options to match people's needs. Eventually, they designed a mobile app template demo (see Fig. 8) to visualize the function of the solution.

The third major shift for both STEM and non-STEM students involved in interdisciplinary collaboration. In a traditional curriculum, non-STEM students are rarely involved in or collaborated with STEM students on the technical side of programming; neither do they have the proper language or knowledge to communicate their desired outcomes to the technical side. The OCEL.AI paradigm provided a common framework and language that enables cross-disciplinary communication and collaboration. For instance, a team of

Needs	Data Types	If done manually (human analysts)	Types of analysis	Machine/Deep Learning Models
The program needs to provide insurance plans options based on users' demographic information	Age, employment status, area code, income level,	Classify user info into: @Age @Employment Status @Income Level @Dependents @Marital Status @ Area code	Classification of structured data (.csv, etc.)	Classification (KNN, SVM, etc.)
The program needs to provide insurance plan information	Price, type of care included, availability by area	Classify insurance plan information into: @Price @AreasAvailable @Services Included	Classification of structured data (.csv, etc.)	Classification (KNN, SVM, etc.)
The program needs to provide the best insurance options based on users' primary health concerns	Pre-existing conditions, pregnancy status, dental concerns present, mental health concerns present	Classify user info based on: @Preganancy Status @Pre-existing Conditions @Dental Concerns @Mental Health Concerns	Classification of unstructured text data	NLP (Topic Modeling, Sentiment Analysis, etc.)

Fig. 7. A sample machine learning modeling plan from students' work.

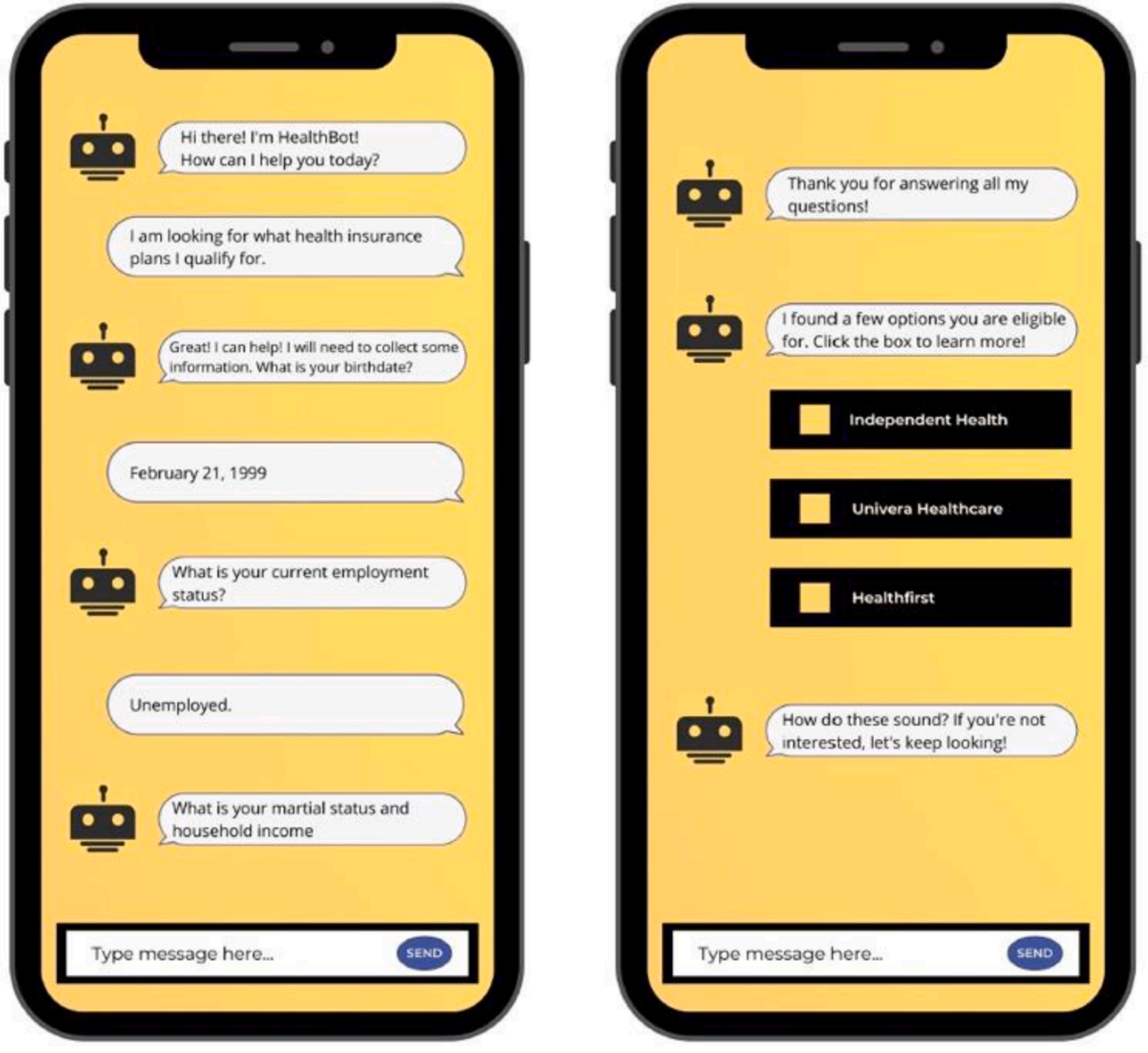


Fig. 8. A sample data solution—A mobile app proposed by journalism students.

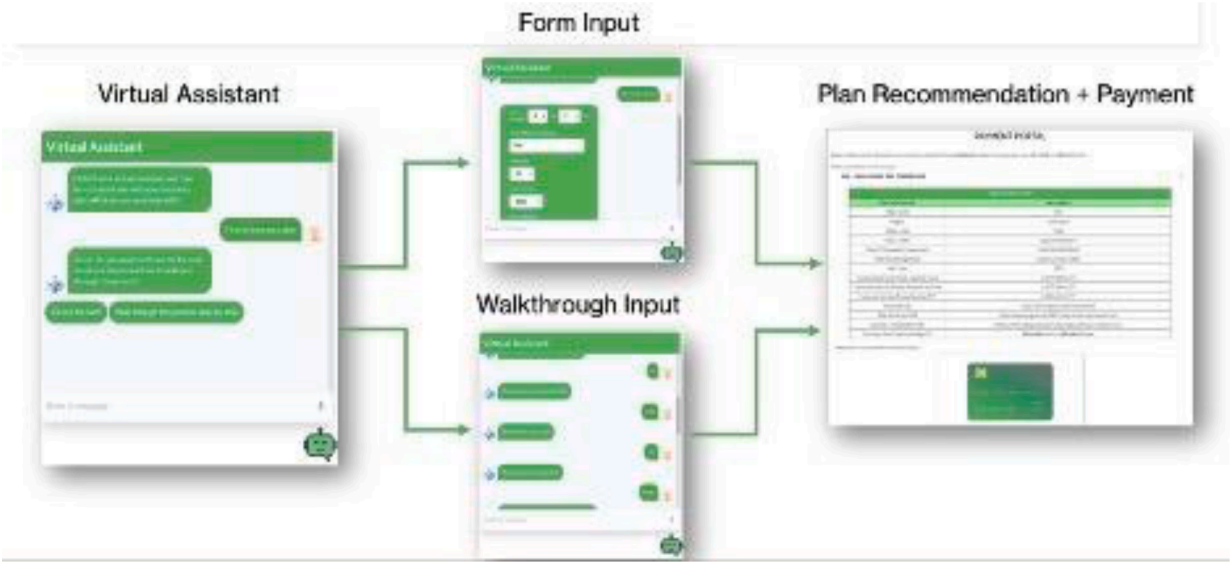


Fig. 9. A sample data solution—A Chatbot assistant developed by computer science students.

computer science students adopted the health insurance proposal from the journalism team; they developed a National Insurance Coverage monitoring website and a virtual web assistant to help visitors identify health insurance coverage. The computer science students adopted the data sources about insurance identified previously by journalism students and added more to the list. They used those data sources to build a website that allows visitors to monitor 1) insurance coverage by state, tier, and year; 2) insurance monthly premium by age, kids, and year; 3) and the insurance company by state and year; 4) insurance rate by state; 5) insurance by employment status; 6) insurance's relationship with education and poverty, and 7) COVID-19 cases by state and county. The computer science students also adopted the idea and template of the chatbot feature proposed by the journalism students and developed a virtual web assistant. The chatbot assistant asked visitors to input their insurance needs through a series of computer-human interactions, and the program eventually forward an insurance plan recommendation and a payment page (see Fig. 9).

5. Conclusion

There is an urgent need to create a qualified data science workforce, which is capable of performing critical functions to solve real-life problems in medicine, business, marketing, criminal justice, journalism, and advertising, to name a few. For this purpose, learning data science is fundamentally about the ability to understand the contexts of data, critically think, analyze data and models, and subsequently produce actionable knowledge and information (Aalst, 2016; Cao, 2017; Provost & Fawcett, 2013). We propose that "Tell Stories" can be used for presenting creative ideas (Yang & Wu, 2012) as well as building strong critical thinking and problem-solving abilities in data science education. Critical and logical thinking on why and what-if questions (Black, 2018; Browne & Keeley, 2007) are needed for effective decision-making in data science, for example, choosing appropriate technical solutions. This innovative approach emphasizes situating data in contexts (Leidig & Cassel, 2020; Wilkerson & Polman, 2020) and closing the gap between real-life use cases, datasets, models, and applications (Aggarwal et al., 2019). The findings suggest the OCEL.AI paradigm can increase data science competence and career motivation. It also has the flexibility and adaptability to be adopted in different disciplines to facilitate teaching and learning.

6. Limitations and future work

As an ongoing project, this study is not without limitations. First, in addition to students' self-reported data, more objective data such as students' assignments against a common grading rubric can be collected to evaluate students' learning outcomes from another perspective. Qualitative data from interviews and focus groups have already been collected to provide further insights into how students learn data science through OCEL.AI and how they perceive the value of storytelling in data science. Due to the page limit, those findings will be presented in another paper. Future research could also validate the findings using a within-subjects experimental design, where students' learning outcomes are measured both before and after the treatment in the same group. In this study, the control and treatment groups of students were not necessarily at the same level in their program study; they were from different courses and universities. As a result, the cognitive and dispositional learning outcomes may vary with students' maturity, the courses, or the university they were in. To minimize the variance in contextual factors, students of the same course but different sessions could be recruited as treatment vs. control groups to minimize the variance of other factors.

CRediT authorship contribution statement

You Li: Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing. **Ye Wang:** Data curation, Formal analysis, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Yugyung Lee:** Project administration, Funding acquisition, Supervision, Conceptualization, Writing – original draft. **Huan Chen:** Investigation, Writing – review & editing. **Alexis Nicolle Petri:** Writing – review & editing, Resources. **Teryn Cha:** Investigation, Writing – review & editing.

Declaration of Competing Interest

None

Data availability

The data that has been used is confidential.

Acknowledgements

This study is supported and funded by the National Science Foundation CUE Ethics: Open Collaborative Experiential Learning (OCEL.AI). Bridging Digital Divides in Undergraduate Education of Data Science. (NSF IUSE #1935076)

References

Aggarwal, C., Bouneffouf, D., Samulowitz, H., Buesser, B., Hoang, T., Khurana, U., Liu, S., Pedapati, T., Ram, P., Rawat, A., Wistuba, M., & Gray, A. (2019). How can AI automate end-to-end data science? arXiv preprint arXiv:1910.14436.

Almeida, D., Machado, D., Andrade, J. C., Mendo, S., Gomes, A. M., & Freitas, A. C. (2020). Evolving trends in next-generation probiotics: A 5W1H perspective. *Critical Reviews in Food Science and Nutrition, 60*(11), 1783–1796.

Andrejevic, M. (2014). Big data, big questions: The big data divide. *International Journal of Communication, 8*, 1673–1689.

Betz, M., Gundlach, E., Hillery, E., & Rickus, J. (2020). The next wave: We will all be data scientists. *Statistical Analysis and Data Mining: The ASA Data Science Journal, 12*, 544–547.

Black, M. (2018). *Critical thinking: An introduction to logic and scientific method*. Pickle Partners Publishing.

Boldosova, V., & Luoto, S. (2019). Storytelling, business analytics, and big data interpretation: Literature review and theoretical propositions. *Management Research Review, 43*(2), 204–222.

Bradshaw, P. (2018). Data journalism teaching, fast and slow. *Asia Pacific Media Educator, 28*(1), 55–66. https://doi.org/10.1177/1326365X18769395

Browne, M. M., & Keeley, S. M. (2007). *Asking the right questions: A guide to critical thinking*. Pearson Education.

Calzada Prado, J., & Marzal, M.Á. (2013). Incorporating data literacy into information literacy programs: Core competencies and contents. *Libri, 63*(2), 123–134.

Cao, L. (2017). Data science: A comprehensive overview. *ACM Computing Surveys (CSUR), 50*(3), 1–42.

Cardenas-Navia, I., & Fitzgerald, B. K. (2015). The broad application of data science and analysis: Essential tools for the liberal arts graduate. *Change: The Magazine of Higher Learning, 47(4), 25-32*.

Cassel, B., & Topi, H. (2015,. October 3–5). *Strengthening data science education through collaboration.* Report on Workshop on Data Science Education Funded by the National Science Foundation, Arlington, VA.

Chakma, K., & Das, A. (2018). A 5W1H based annotation scheme for semantic role labeling of English tweets. *Computación y Sistemas, 22*(3), 747–755.

Clancy, T. R., Bowles, K. H., Gelinas, L., Androwich, I., Delaney, C., Matney, S., Sensmeier, J., Warren, J., Welton, J., & Westra, B. (2014). A call to action: Engage in big data science. *Nursing Outlook, 62*(1), 64–65.

Davies, K., & Cullen, T. (2016). Data journalism classes in Australian universities: Educators describe progress to date. *Asia Pacific Media Educator, 26*(2), 132–147.

Dichev, C., & Dicheva, D. (2017). Towards data science literacy. *Procedia Computer Science, 108*, 2151–2160. https://doi.org/10.1016/j.procs.2017.05.240. June.

D'Ignazio, C. (2017). Creative data literacy: Bridging the gap between the data-haves and data-have-nots. *Information Design Journal, 23*(1), 6–18.

Faris, J., Kolker, E., Szalay, A., Bradlow, L., Deelman, E., Feng, W., Qiu, J., Russell, D., Stewart, E., & Kolker, E. (2011). Communication and data-intensive science in the beginning of the 21st century. *A Journal of Integrative Biology, 15*(4), 213–215.

Fernández-Rovira, C., Valdés, J.Á., Molleví, G., & Nicolas-Sans, R. (2021). The digital transformation of business. Towards the datafication of the relationship with customers. *Technological Forecasting and Social Change, 162*, Article 120339.

Flensburg, S., & Lomborg, S. (2021). Datafication research: Mapping the field for a future agenda. *New Media & Society. https://doi.org/10.1177/14614448211046616*.

Frank, M., Walker, J., Attard, J., & Tygel, A. (2016). Data literacy: What is it and how can we make it happen? *The Journal of Community Informatics, 12*(3), 4–8.

Gold, M., McClarren, R., & Gaughan, C. (2013). The lessons Oscar taught us: Data science and media & entertainment. *Big Data, 1*(2), 105–109.

Heravi, B. R. (2019). 3Ws of data journalism education: What, where and who? *Journalism Practice, 13*(3), 349–366.

Jarke, J., & Macgilchrist, F. (2021). Dashboard stories: How narratives told by predictive analytics reconfigure roles, risk and sociality in education. *Big Data & Society, 8*(1). https://doi.org/10.1177/205395172110255

Koltay, T. (2014). Data literacy: In search of a name and identity. *Journal of Documentation, 71*(2), 401–415.

Kross, S., Peng, R. D., Caffo, B. S., Gooding, I., & Leek, J. F. (2020). The democratization of data science education. *The American Statistician, 74*(1), 1–7. 10.1080/00031305.2019.1668849.

Lavorgna, A., & Ugwudike, P. (2021). The datafication revolution in criminal justice: An empirical exploration of frames portraying data-driven technologies for crime prevention and control. *Big Data & Society, 8*(2). https://doi.org/10.1177/20539517211049670

Loukides, M. (2010). *What is data science?* O'Reilly Radar. Retrieved from http://radar.oreilly.com/2010/06/whatis-data-science.html.

Leidig, P. M., & Cassel, L. (2020). In *ACM taskforce efforts on computing competencies for undergraduate data science curricula. Proceedings of the 2020 ACM conference on innovation and technology in computer science education* (pp. 519–520). https://doi.org/10.1145/3341525.3393962.

Mandinach, E. B., & Gummer, E. S. (2013). A systemic view of implementing data literacy in educator preparation. *Educational Researcher, 42*(1), 30–37.

Martin, J. D. (2017). A census of statistics requirements at U.S. journalism programs and a model for a "statistics for journalism" course. *Journalism & Mass Communication Educator, 72*(4), 461–479.

Maybee, C., & Zilinski, L. (2015). Data informed learning: A next phase data literacy framework for higher education. *ASIS&T, 52*(1), 1–4.

Neuhauser, A. (2015,. June 29). *2015 STEM index shows gender, racial gaps widen.* Retrieved from http://www.usnews.com/news/stem-index/articles/2015/06/29/gender-racial-gaps-widen-in-stem-fields.

Nguyen, A., & Lugo-Ocando, J. (2016). The state of data and statistics in journalism and journalism education: Issues and debates. *Journalism, 17*(1), 3–17.

Pedersen, A.Y., & Caviglia, F. (2019). Data literacy as a compound competence. In T. Antipova & A. Rocha (Eds.), *Digital science*, 850, 166–173. Springer International Publishing. https://doi.org/10.1007/978-3-030-02351-5_21.

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big data, 1*(1), 51–59.

Ridsdale, C., Rothwell, J., Smit, M., Ali-Hassan, H., Bliemel, M., Irvine, D., & Wuetherick, B. (2015). *Strategies and best practices for data literacy education: Knowledge synthesis report*. Dalhousie University. Retrieved from https://dalspace.library.dal.ca/bitstream/handle/10222/64578/Strategies%20and%20Best%20Practices%20for%20Data%20Literacy%20Education.pdf.

Segel, E., & Heer, J. (2010). *Narrative visualization: Telling stories with data*. IEEE Transactions on Visualization and Computer Graphics, 16 (6), 1139-1148.

Swanstrom, R. (2023). Data science colleges and universities. Retrieved February 9 from http://datascience.community/colleges.

Tang, R., & Sae-Lim, W. (2016). Data science programs in U.S. higher education: An exploratory content analysis of program description, curriculum structure, and course focus. *Education for Information, 32*, 269–290.

The 365 Team (2020). How to become a data scientist in 2020: Top skills, education and experience. 365data science.com, February 25. https://365datascience.com/career-advice/career-guides/become-data-scientist-2020/#3.

U.S. Bureau of Labor Statistics (2023). Employed persons by detailed occupation, sex, race, and Hispanic or Latino ethnicity. Labor force statistics from the current population survey. Table 11. January 2. https://www.bls.gov/cps/cpsaat11.htm.

Thompson, R., & Haddock, G. (2012). Sometimes stories sell: When are narrative appeals most likely to work? *European Journal of Social Psychology, 42*(1), 92–102.

Van Der Aalst, W. (2016). *Process mining: Data science in action* (pp. 3–23). Springer.

Waller, M. A., & Fawcett, S. E. (2013). Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management. *Journal of Business Logistics, 34*, 77–84.

Weber, W. (2020). Exploring narrativity in data visualization in journalism. In M Engebretsen, & H Kennedy (Eds.), *Data Visualization in Society* (pp. 295–311). Amsterdam: Amsterdam University Press.

Wilkerson, M. H., & Polman, J. L. (2020). Situating data science: Exploring how relationships to data shape learning. *Journal of the Learning Sciences, 29*(1), 1–10. https://doi.org/10.1080/10508406.2019.1705664

Wolff, A., Gooch, D., Montaner, J. C., Rashid, U., & Kortuem, G. (2016). Creating an understanding of data literacy for a data-driven society. *Journal of Community Informatics, 12*(3), 9–26.

Yang, Y. C., & Wu, W. C. I. (2012). Digital storytelling for enhancing student academic achievement, critical thinking, and learning motivation: A year-long experimental study. *Computers & Education, 59*(2), 339–352.

Yavuz, F. G., & Ward, M. D. (2020). Fostering undergraduate data science. *The American Statistician, 74*(1), 8–16.

Zhu, L., & Du, Y. R. (2018). Interdisciplinary learning in journalism: A Hong Kong study of data journalism education. *Asia Pacific Media Educator, 28*(1), 16–37.