TRAINING MEXICAN HIGH SCHOOL TEACHERS TO ENHANCE DATA SCIENCE TEACHING

Rosa Daniela Chávez Aguilar and Ana Luisa Gómez Blancarte
Instituto Politécnico Nacional, México
rchaveza1600@alumno.ipn.mx

This paper aims to describe advances in a proposal for preparing Mexican high school teachers to teach data science according to the six phases of the Data Investigation Process cycle: frame the problem; consider and gather the data; process the data; explore and visualize the data; consider the models; and communicate and propose actions. We report an online course guided by the Data Investigation Process cycle to provide Mexican high school in-service teachers with knowledge about techniques used in data science. We identify how in-service teachers become aware of the phases when asked to investigate a real problem through a project. Teachers found difficulties in formulating an investigative question and faced the problem of finding databases of real problems. The retrospective analysis allows us to evaluate some improvements that can be made to the course so that teachers can develop the skills that data science teaching demands.

INTRODUCTION

According to Fukuda (2021), the first time the term data science (DS) was used was at the "Statistics = Data Science?" conference by Jeff Wu at the University of Michigan in 1997, who proposed renaming statistics as DS, understood as a trilogy of collecting, analyzing, and making decisions with data. Currently, the idea of decision-making is supported by Asamoah et al. (2018), as they note that DS "is an interdisciplinary field that provides insights to support decision making" (p. 371). However, statistics it is one of the disciplines that collaborate with mathematics, computer science, and a particular domain in DS (Carmichael & Marron, 2018). In this regard, Blei and Smyth (2017) provide a broader idea about DS:

> Data science focuses on exploiting the modern deluge of data for prediction, exploration, understanding, and intervention. It emphasizes the value and necessity of approximation and simplification. It values effective communication of the results of a data analysis and of the understanding about the world that we glean from it. It prioritizes an understanding of the optimization algorithms and transparently managing the inevitable tradeoff between accuracy and speed. It promotes domain-specific analyses, where data scientists and domain experts work together to balance appropriate assumptions with computationally efficient methods (p. 8690).

This definition of DS provided by Blei and Smyth (2017) demands skills that, according to Hardin et al. (2015), the data scientist is expected to possess, such as data use, mining and storage, programming and algorithmic thinking, and statistical knowledge. In this sense, data scientists generate insights from data using "advanced computational tools, data mining algorithms, statistical analysis, and machine learning" (Asamoah et al., 2018, p. 371). Furthermore, in collaboration with domain experts, they can communicate their findings and conclusions based on the data to generate recommendations on what they have learned (Vance et al., 2022).

The study of DS demands teaching programming and handling large volumes of data (Boz & Allexsaht-Snider, 2022). GAISE II proposes a framework of recommendations, concepts, and examples that enable students "to gain an appreciation for the vital role of statistical reasoning and data science, and to acquire the essential life skill data literacy" (Bargagliotti et al., 2020, p. 3). GAISE II is an example of the adaptation to the changes in which statistics is immersed; it emphasizes a new conception of data, which today are also presented as multivariate collections of complexly structured or unstructured data, including text, images, and sounds, as well as the use of apps for their visualization. As Lee et al. (2021) point out, DS has influenced data use. This use has permeated the educational arena by paying attention to how educators can support students in learning about data.

Research highlights courses and content offered on MOOCs (Massive Online Open Courses) platforms (e.g., Brooks et al., 2021) and suggests upper-level curricula for teaching DS (e.g., Hardin et al., 2015). It also highlights the need, on the part of teachers, for interactive teaching resources and textbooks (Schwab-McCkoy et al., 2021) and programs for teacher professional development (Sanusi

In: EM Jones (Ed.), *Fostering Learning of Statistics and Data Science*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
July 2023, Toronto, Canada. https://iase-web.org

et al., 2022). In this sense, the present article describes advances in a proposal for preparing Mexican high school teachers to teach DS. The article is part of a larger research project under development.

In Mexico, public education in DS has gained importance at the university level, for example, in careers such as undergraduate degrees in DS offered by universities such as the Universidad Nacional Autónoma de México (2023) and the Instituto Politécnico Nacional (2023). Some technological high school institutions offer in their curriculum a professional option for students to be trained as Data Science Technicians by studying Data and Information Science. According to the curriculum, such training "is relevant because the current need in the different sectors of services, government, industry, and companies requires a technical professional trained in the analysis and massive processing of information" (Secretaría de Educación Pública [SEP], 2019, p. 10). In addition, the Common Curriculum Framework for High School Education proposes the inclusion of Digital Development that makes use of tools such as DS for 'the application of techniques, methods, and existing technological resources creatively and innovatively with critical thinking to address given problems" (SEP, 2023, p. 26).

The importance that DS is taking in Mexican education entails the training of teachers to address the recommendations for teaching this discipline. Hicks and Irizarry (2018) propose to guide teaching under three critical skills: 1) creation, which refers to a data scientist having a stake in solving and formulating questions; 2) connection, which is about connecting a research question with available data, beyond just addressing techniques; and 3) informatics as an essential tool in the classroom. On this last point, Gómez-Blancarte's (2022) study reports that Mexican high school teachers' main computer tools to teach statistics are the scientific calculator and Excel. Therefore, teachers must know other tools that allow them to promote the teaching of programming and the handling of large volumes of data.

FRAMEWORK

The Investigative Data Process (IDP) framework proposed by Lee et al. (2022) was used to guide the preparation of high school teachers in teaching DS. The IDP (see Fig. 1) involves six phases:

1. *Frame the Problem*. Consider real-world phenomena and broader issues related to the problem, pose investigative questions, and anticipate potential data and strategies. In the present work, teachers were invited to formulate "posing questions" and "asking questions" (Arnold & Franklin, 2021). The former is formulated to initiate the research process, and the latter arises spontaneously during the research process.
2. *Consider & Gather Data*. Understand possible attributes, measurements, and data collection methods needed for the problem; evaluate use appropriate design and techniques to collect or source data; consider sample size, access, storage, and trustworthiness of data.
3. *Process Data*. Organize, structure, clean and transform data in efficient and useful ways; consider additional data cases or attributes.
4. *Explore & Visualize Data*. Construct meaningful visualizations, static or dynamic; compute meaningful statistical measures; explore and analyze data for potential relationships or patterns that address the problem.
5. *Consider Models*. Analyze and identify models that address the problem; consider assumptions and context to the models; recognize possible limitations.
6. *Communicate & Propose Action*. Craft a data story to convey insights to stakeholder audiences; justify claims with evidence from data and propose possible action; address uncertainly, constraints, and potential bias in the analysis.

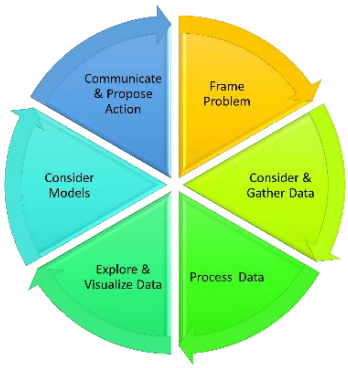

Figure 1. Framework of the Data Investigation Process. Adapted from Lee et al. (2022, p. 11).

The IDP addresses the practices and processes commonly used in statistics education and DS, such as the PPDAC investigative cycle proposed by Wild and Pfannkuch (1999). It also has a relationship with the skills described by Hicks and Irizarry (2018): in phase 1, the generation of a research question is promoted; in phase 2, that question relates to the data available, while phases 3, 4 and 5 demand the use of computational tools.

METHODOLOGY

The teachers' training consisted of designing and implementing a course called "Teaching Data Science for High School Teachers". The course was designed under the Design-Based Research methodology. According to Scott et al. (2020), the methodology comprises four phases. 1) Researchers identify a problem that needs to be solved (provide high school teachers with resources for teaching DS). 2) Plan and present a potential "solution" to the issue in the form of educational tools that theory and previous research suggest addressing the problem (the design of a course guided by the DIP model reported in the research). 3) Researchers test the educational tools (the course) in a real environment (the course was given to in-service teachers) to see if the tools positively impact. 4) The researchers review the results obtained, identify the aspects that were positive in solving the initial learning problem and the aspects that were not useful, and share how the research responds to the theory underlying the experiment.

The course was organized around the following topics: *DS perspective; Approaches to teaching statistics as applied to DS; Framing the problem; Data collection; Algorithmic thinking (Introduction to Python); Variability; Data production and cleaning; The big ideas of Machine Learning (Using train and test data); Data representation; Introduction to supervised models; Simple linear regression; Decision tree model; Confidence measures for model selection; Introduction to unsupervised models (Clusters); Presentation of neural networks; Communicating findings and storytelling*. These topics were developed in 12 online classes of 3 hours each, during the period January-March 2023.

The course was officially registered with the Dirección de Formación e Innovación Educativa of the Instituto Politécnico Nacional, which published the course in its educational offerings. Twenty teachers registered for the course, but on the first day of classes, only 12 showed up, and 7 completed the course. When the course started, the teachers closed the school semester, and on January 31, they initiated a new semester. The change from one semester to another possibly affected the teachers' schedules, so they could not continue the course. In other cases, the teachers dropped out from the first classes; we believe the course content was not what they expected. 7 teachers had master's degrees, 6 of them taught in high school education (five taught classes related to programming and one in mathematics) and one in undergraduate education. In general, the topics were developed through presentations of thematic content by the instructor (first author), with illustrative examples, followed by activities in which teachers practiced those examples or discussed the lectures provided. This development was supported by using technological platforms and tools such as YouCubed.org, Edublocks, CODAP (Common Online Data Analysis Platform, codap.concord.org), and Google Colab to use Python. As part of the final course evaluation, teachers were required to conduct a research project based on the phases of the DIP cycle. Since the course was developed following these phases, the teachers had to apply what they had seen in class in their projects. Three teams were formed, and every

four weeks, they shared the progress of their project. Although it is necessary to clarify that most of the time, teachers did not present their progress because they said they had not had time to advance.

This article analyzes one of the three projects received to identify how in-service teachers become aware of the DIP phases when asked to investigate a real problem. The data analyzed were the teams' project presentations on the last day of class; there was no written report. The selected team was the only one that provided sufficient information on how they covered the phases of the DIP cycle during the presentation of their project. In contrast, the projects of the other two teams needed to be improved in the execution of each phase. For example, in the first phase, both teams failed to formulate a research question; in the second phase, one of the teams did not describe or present what treatment it performed on the data; in the third phase, one of the teams mentioned that it performed data extraction and cleaning, but did not point out the procedure used and its results; the other team did not explain this phase; in the fourth phase, both teams did not mention whether they identified any characteristics on their data visualizations; in the fifth phase one of the teams failed to explain the results of the technique used; in the last phase, both teams focused on describing the results of their algorithm, without relating them to the context of their data.

RESULTS

Frame the Problem

The project dealt with the communities most affected in Syria by the February 6, 2023, earthquake and needing urgent humanitarian assistance (hereinafter referred to as the Syria Project). The team addressed the investigative question: *Which earthquake-affected communities in Syria require the most urgent attention?* It should be noted that, to frame the problem, the team first searched for databases that they could analyze according to the complexity that the data demanded (e.g., coding and data cleaning). In this sense, the team did not start the DIP cycle with a defined research problem; instead, they prioritized searching for data and formulating possible investigative questions depending on that data.

Consider & Gather Data

The team searched for data on the Internet on different open websites. For the Syria Project, they obtained the data from The Humanitarian Data Exchange HDX website (https://data.humdata.org/dataset/syria-earthquake-impact), specifically the database “Syria Earthquake Impact 20 March 2023.xls”. The database contains 27 variables of the affected communities, for example: District, # of Causalities, # of injuries, # of completely destroyed houses, # of temporary accommodation centers (total buildings including schools and others), Are there latrines available? # of needed meals (per day).

Process Data

The team attended to the data cleaning and normalization part. With the work done, they showed a good understanding of this phase since it considered elements such as the revision of structures, variable labeling, completeness (they commented that they had difficulty in finding a criterion to impute missing data), and consistency of the data; as well as the cleaning of some texts, in this case, the headings that were presented in Arabic. To process the data, the team used Python in Google Colab. In Fig. 2 (left), the team showed an understanding of the commands seen in class for variable labeling (in this case, "v_" refers to the fact that it is a categorical variable), as well as for the conversion to lowercase of text records.

Explore & Visualize Data

The team used the CODAP tool, which they said they had not known before but considered “easy and quick to use” for the exploratory analysis of the data. They also used Google Colab for the execution of Python code (see Fig. 2), taking as reference notebooks seen in class. In addition, they constructed histograms and box plots that allowed them to evaluate whether there were outliers and assess their inclusion. Fig. 2 (right) shows a scatterplot of the number of causalities per community the team developed to explore possible answers to their investigative question.

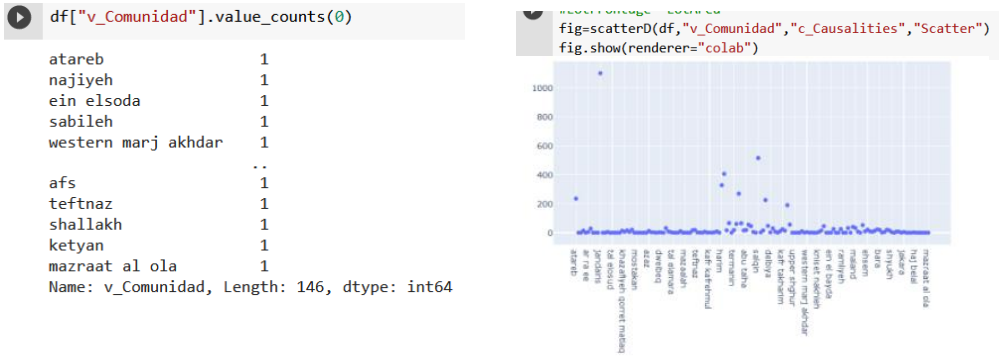

Figure 2. Representations generated in Google Colab

Consider Models

The team considered approaching the clustering technique to model their data, as they commented on the interest of finding groups of communities with similar characteristics to try to answer their investigative question and commented possible use of some supervised model with a target variable, which shows their understanding of the techniques addressed in class. However, they shared that they needed more time to develop this model, so they did not share conclusions for the last phase of the cycle: *Communicate & Propose Action*.

CONCLUSION

In the last class session, the teachers expressed that the DIP cycle was an adequate guide for proposing a research project. However, as they developed their projects, they had difficulties executing the cycle phases. One of the main problems that the three teams had was finding databases from which they could pose an investigative question. The teachers did not start from a problem or investigative question (phase 1) that would lead them to the search for data needed to address the problem (phase 2). Instead, as mentioned by the team whose project was reported in this article, "framing the problem has been defined, in our case, from having the data set available". In addition, once they found databases, they evaluated the possibility of posing investigative questions that only required a little data processing. While the teachers understood the importance of data cleaning (phase 3), they pointed out the complexity of programming in Python. On the one hand, they, despite teaching programming classes, said they were not familiar with the coding required for the type of data processing seen in class; on the other hand, data processing demanded more time and dedication from them, the time they did not have due to their teaching activities.

The retrospective analysis of this first implementation of the course allows us to evaluate some improvements that can be made to both the design and implementation of the course so that teachers can develop the skills that DS teaching demands. On the one hand, we must focus more on the first two phases of the DIP cycle. For instance, for teachers to participate in formulating and resolving investigative questions, it is necessary to involve them in activities that allow them to identify different types of statistical questions and know the criteria for answering the questions, as suggested by Arnold and Franklin (2021). Searching for data on real problems was a difficult task for the teachers. In this sense, this search can be part of the activities within the course (instead of doing it outside the course). It is important to teach them how to discriminate the essential data to answer a previously formulated investigative question. On the other hand, since computer science is key in DS, we need to make teachers see the relevance of using technological tools for programming and visualization of large data volumes, starting with low-code or no-code tools such as CODAP.

REFERENCES

Arnold, P., & Franklin, C. (2021). What makes a good statistical question? *Journal of Statistics and Data Science Education, 29*(1), 122-130.

Asamoah, D., Doran, D., & Schiller, S. (2018). Interdisciplinarity in data science pedagogy: A foundation design. *Journal of Computer Information Systems, 60*(4), 370-377.

Bargagliotti, A., Franklin, C., Arnold, P., Gould, R., Johnson, S., Perez, L., & Spangler, D. (2020). PreK-12 guidelines for assessment and instruction in statistics education II (GAISE II). *A framework*

for statistics and data science education. American Statistical Association- National Council of Teachers of Mathematics. https://www.amstat.org/asa/files/pdfs/GAISE/GAISEIIPreK12_Full.pdf

Blei, D. M., & Smyth, P. (2017). Science and data science. *PNAS, 114*(33), 8689–8692.

Brooks, C., Quintana, R. M., Choi, H., Quintana, C., NeCamp, T., & Gardner, J. (2021). Towards Culturally Relevant Personalization at Scale: Experiments with Data Science Learners. *International Journal of Artificial Intelligence in Education, 31*(4), 516-537.

Boz, T., & Allexsaht-Snider, M. (2022). How do elementary school teachers learn coding and robotics? A case study of mediations and conflicts. *Education and Information Technologies, 27*, 3935-3963.

Carmichael, I., & Marron, J. S. (2018). Data science vs. statistics: two Cultures? *Japanese Journal of Statistics and Data Science, 1*(1), 117-138.

Fukuda, H. (2021). What makes data science education unique?: A literature review. In R. Helenius & E. Falck (Eds.), *Statistics Education in the Era of Data Science. Proceedings of the Satellite conference of the International Association for Statistical Education (IASE)*. Online conference.

Gómez-Blancarte, A. (2022). An overview of the use of technology for teaching statistics by mexican high school teachers. In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the Gap: Empowering & Educating Today's Learners in Statistics. Proceedings of the 11th International Conference on Teaching Statistics (ICOTS11, September 2022), Rosario, Argentina*. ISI/IASE.

Hardin, J., Hoerl, R., Horton, N. J., Nolan, D., Baumer, B., Hall-Holt, O., Murrell, P., Peng, R., Roback, P., Lang, D. T., & Ward, M. D. (2015). Data science in statistics curricula: preparing students to "think with data". *The American Statistician, 69*(4), 343–353.

Hicks, S. C., & Irizarry, R. A. (2018. A guide to teaching data science. *The American Statistician, 72*(4), 382-391.

Instituto Politécnico Nacional (2023). *Licenciatura en Ciencia de Datos* [Bachelor's Degree in Data Science]. Instituto Politécnico Nacional. https://www.ipn.mx/oferta-educativa/educacion-superior/ver-carrera.html?lg=es&id=69&nombre=Licenciatura-en-Ciencia-de-Datos

Lee, V., Wilkerson, M.H., & Lanouette, K. (2021). A call for a humanistic stance toward K-12 data science education. *Educational Researcher, 50* (9), 664-672.

Lee, H. Y., Mojica, G. F., Thrasher, E. P., & Baumgartner, P. (2022). Investigating data like a data scientist: Key practices and processes. *Statistics Education Research Journal, 21*(2), 1-23.

Sanusi, I. T., Oyelere, S. S., Vartiainen, H., Suhonen, J., & Tuklalnen, M. (2022). A systematic review of teaching and learning machine learning in K-12 education. *Education and Information Technologies.* https://doi.org/10.1007/s10639-022-11416-7

Scott, E., Wenderot, M., Doherty, J., & Tanner, K. (2020). Design-Based Research: A Methodology to Extend and Enrich Biology Education Research. *CBE—Life Sciences Education, 19*(11), 1-12.

Secretaria de Educación Pública. (2019). *Programa de estudios de la carrera técnica de Ciencia de Datos e Información* [Syllabus of the Data and Information Science technical degree program]. Subsecretaría de Educación Media Superior. http://cbtis222.edu.mx/res/pdf/CienciaDeDatosEInformacion.pdf

Secretaria de Educación Pública. (2023). *Progresiones de aprendizaje del recurso sociocognitivo Cultura Digital*. [Learning progressions of the socio-cognitive resource Digital Culture]. Secretaria de Educación Pública. http://desarrolloprofesionaldocente.sems.gob.mx/convocatoria1_2023/docs/Progresiones%20de%20aprendizaje%20-%20Cultura%20digital.pdf

Schwab-McCkoy A., Baker, C. M., & Gasper, R. E. (2021). Data Science in 2020: Computing, curricula, and challenges for the next 10 years. *Journal of Statistics and Data Science Education, 29*(1), S40-S50.

Vance, E., Alzen, J., & Smith, H. (2022). Creating shared understanding in statistics and data science collaborations. *Journal of Statistics and Data Science Education, 30*(1), 54-64.

Universidad Nacional Autónoma de México. (2023). *Ciencia de Datos*. [Data Science]. Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas. http://oferta.unam.mx/planestudios/Ciencia-de-Datos-IIMAS_Plan%20de%20estudios19.pdf

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, 67(3), 223–248.