ANT 388C

Spring 2024

Course Information

Instructional Mode: Face-to-face
Meeting Time 1: Mon/Wed 03:30 PM - 05:00 PM
Meeting Time 2: EPS 4.104

Unique Number 31405

Additional Sections: BIO 384K (unique number 49134)

Instructor

Anthony Di Fiore
Email: anthony.difiore@austin.utexas.edu

Office Hours and Location

Office Location: WCP Room 5.150

Office Hours: MW, 2:00 to 3:00 pm and by appointment

Welcome Message

Welcome to **Applied Data Analysis 2024**!!!

Overview of the Class

This course provides an overview of methods and tools for applied data analysis. It is geared toward research in biological anthropology and evolutionary biology, but the material covered is applicable to a wide range of natural, social science, and humanities disciplines. Students will receive practical, hands-on training in various data science tools and workflows, including data acquisition and wrangling, exploratory data analysis and visualization, statistical analysis and interpretation, and literate programming and version control.

Statistical topics to be covered include basic descriptive and inferential statistics, hypothesis testing, basic regression and ANOVA, generalized linear modeling, and mixed effect modeling. Statistical inference will be considered from a frequentist perspective, introducing both parametric and resampling techniques. If we have time, I will also introduce a Bayesian perspective, although this approach will not be tackled at a particularly advanced level. Additional methods and tools will also be covered based on student interest (e.g., geospatial data analysis, phylogenetic comparative methods, social network analysis, text corpus construction and mining, population genetic analysis) and on how quickly the class feels like moving forward.

The course emphasizes the development of solid data science skills, focusing on the practical side of data manipulation, analysis, and visualization. Students will learn to use the statistical programming language R as well as many other useful software tools (e.g., shell scripts, text editors, databases, query languages, and version control systems).

NOTE: This class is supported by DataCamp, the most intuitive learning platform for data science. Learn R, Python and SQL the way you learn best through a combination of short expert videos and hands-on-the-keyboard exercises. Take over 100+ courses by expert instructors on topics such as importing data, data visualization or machine learning and learn faster through immediate and personalized feedback on every exercise.

Structure

This course is divided into three main sections.

In Part I, we will introduce and practice using the statistical programming software ***R***, the ***RStudio*** integrated development environment, and the ***R*** package ecosystem. We will also cover programming/scripting fundamentals as implemented in ***R*** (functions, flow control) and practice using version control systems (e.g., ***git*** and ***GitHub***) as we build up our skills for conducting reproducible research. We will use all of these tools to practice data wrangling and perform exploratory data analysis and visualizations.

In Part II, we will cover basic statistical and probability theory and methods of statistical inference. We will discuss classical null hypothesis significance testing and more contemporary methods based on permutation methods and, if time permits, I may also introduce alternative Bayesian approaches to inference. In this section, we will cover a variety of linear modeling topics, including simple and multivariate regression, ANOVA and ANCOVA, generalized linear modeling, and mixed effects modeling, as well as regression diagnostics and tools for model selection.

Finally, in Part III, I hope to introduce a few additional and more specialized data analysis and visualization topics. Assuming we get there, Part III will introduce a mish-mash of (hopefully useful and interesting) topics and tools, e.g., working with geospatial data and phylogenetic trees, network analysis, machine learning, natural language processing, image analysis, etc. Past experience suggests that I am proposing an ambitious amount of material to cover, so we likely will not get to some of these more specialized kinds of analyses. Still, if there's a topic you are particularly excited about exploring, let me know and I will see what we can do!

Thus, below is a VERY TENTATIVE schedule of weekly topics, which represents my current plans and objectives. However, changes to this schedule may be made at my discretion if circumstances warrant. Indeed, as we go through the semester, it is almost a certainty that the schedule, topics, and pace of the course will change somewhat as I see where student interests, experience, and learning challenges lie. Such changes are not unusual and should be expected.

It is your responsibility to visit the Canvas and course websites regularly and to note changes to the schedule when announced. I will do my best to ensure that you are notified of changes with as much

advance notice as possible.

Tentative Schedule of Course Topics

Part	Week #	Week of...	Topic
I	1	01/15	Course Overview and Introductions, Getting Started with ***R*** and ***RStudio***
	2	01/22	Basics of the ***R*** Language and the ***R*** Ecosystem
	3	01/29	Reproducible Research and Version Control
	4	02/05	Data Wrangling, Programming Fundamentals - Functions and Flow Control
	5	02/12	Descriptive Statistics and Exploratory Data Analysis
	6	02/19	Data Visualization and ***R*** Graphics
II	7	02/26	Basics of Statistical Inference and Hypothesis Testing
	8	03/04	Simple Linear Regression
	SPRING BREAK		
	9	03/18	Categorical Data Analysis and ANOVA
	10	03/25	Multiple Regression and ANCOVA, Generalized Linear Modeling
	11	04/01	Model Selection, Mixed Effects Modeling
III	12	04/08	Working with Geospatial Data*
	13	04/15	Introduction to Network Analysis*
	14	04/22	Using APIs, Scraping Data from the Web, and Working with Text*
	15	04/29	Tree Building and Phylogenetic Data Analysis*

NOTE: Depending on interest, I may also run extra, optional “workshops” on other topics of possible interest, e.g., linking ***R*** to other data analysis tools (JupyterLabs, RMDBs, Knime, the ***Python*** and ***Julia*** computing languages for data manipulation and analysis, etc.).

Learning Outcomes

At the conclusion of this course, students will be able to:

- understand and articulate key concepts and methods in applied data science; acquire, manipulate, and manage data from varied sources; conduct exploratory data analyses; test statistical hypotheses; build models to classify and make predictions about data; and evaluate model performance;

- use modern tools for data analysis (e.g., the Unix command line, version control systems, the R programming environment, web APIs) and apply “best practices” in data science and data management;

- interact with both local and remote data sources to store, query, process, and analyze data presented in a variety of common formats (e.g., delimited text files, structured text files, various database systems);

- comfortably write their own simple computer programs/scripts for data management, statistical analysis, visualization, and more specialized applications;

- design and implement reproducible data science workflows that take a project from data acquisition to analysis to presentation and organize their work using a version control system;

- and apply all of these tools to questions of interest in the natural and social sciences.

Grading Policy

Course grades will be based on attendance and participation in class and on several types of assignments as noted in the **Overview of all Major Course Requirements and Assignments** section below.

Instructional Format

For those of you who may have colleagues who have taken this class before and know a bit about it, I plan to do things a bit differently this semester. Rather than spend as much time in class each week lecturing and working through various online modules I have prepared and that are published on the course website, I will instead assign modules that I expect you to review ahead of time, along with readings relevant to the material covered in each. Class time will then be used for shorter lectures highlighting the most critical material, for addressing your theory and coding questions and discussing things that are not clear, and for working through tutorials and programming exercises in an environment where I can help guide and troubleshoot. That is, I hope that our class time will involve much more direct coding practice and will be more interactive, more useful, and more satisfying than has been possible in the past two years when instruction was largely online. I have thus revised the assignments for the class, then, to include a much larger number of short coding exercises rather than more extensive and time-consuming homework problem sets. I have also revised one of the major assignments for the class to allow those who are interested to tackle a data analysis/visualization project with their own data.

Late Work and Grace Periods

I will follow a "Time Bank" model to give you some additional flexibility to deal with challenges that might arise in meeting deadlines during these challenging times for your "large" projects (i.e., your individual "Data Analysis Replication" project, your "Collaborative Data Science" project, and your individual “Data Visualization" assignment as well as for any of your weekly exercises. You will have a total of five "extension days" that you can use (or not) as you see fit and apply across this set of

deadlines. So, you might choose to turn in each of five weekly exercises one day late, or one assignment five days late, with no penalty. If you exhaust your "Time Bank", I will take off 10% off of the total value of each subsequent late assignment for each day or partial day late. If you are going to dip into your "Time Bank" to extend the deadline for an assignment, I ask that you give me a heads up that you are going to do so sometime before the assigned deadline (via email through Canvas is fine), even if it is only a few minutes before.

Overview of all Major Course Requirements and Assignments

Regular attendance and participation in class – 15%

Your attendance and participation in class constitute a combined 15% of your final grade. I expect you to attend all class lectures and programming sessions, to contribute actively to this course throughout the semester by participating fully in class exercises and discussions, and to do your best work on assigned programming challenges. You should come to class prepared, having read the suggested readings for that day/topic, and during class you should not hesitate to bring up questions or comments about the day's materials. I will keep track of attendance regularly, so absences will be noted. Because the material we will cover continually builds on previous material, if you need to miss a class contact me in advance to work out how to make up the material and not fall behind.

Programming exercises – 30%

I will assign regular programming exercises, roughly 1 per week, that together will make up 30% of your grade. Some will be very easy, designed to scaffold a skill, and will be completed during class sessions while others will be more challenging and take some more thought, effort, and time out of class. Assignments will be evaluated based on completion (0 = not turned in or less than 50% complete or correct, 1 = 50% to 75% complete or correct, 2 = 75% to 100% complete correct). Depending on the assignment, you might be asked to work on your own or with a partner. When you work with a partner, you will nonetheless turn in your own notes and scripts for each assignment

One individual data analysis replication project – 25%

This is an individual assignment in which, in consultation with the instructor, you will choose a published natural or social science paper and dataset from the primary literature, replicate the analyses presented in the paper, and produce a short report on your process, successes, and challenges. The goals of this assignment are to give you hands-on experience in accessing and analyzing data and in producing a reproducible data analysis workflow. You should begin looking right away for an appropriate paper that you find interesting, ideally one related to your planned graduate work. The key criterion is that you must be able to put your hands on the datasets that were used for the key analyses in the paper.

One collaborative data science project and presentation – 25%

This is a collaborative assignment in which you and a partner will complete your choice of one of the following two options:

[1] Methods Exploration: In consultation with the instructor, select a particular statistical, data processing, and/or data visualization method that we do not cover in the class and that you would like to explore further and prepare a 10-15 minute presentation for your peers during the last week of class describing the purpose of the method and examples/applications of its use using a publicly available dataset (e.g., from Kaggle, Google Dataset Search, Data.gov, etc.). In addition to your presentation, you will develop an accompanying ***R*** "vignette" and package that will be shared through

GitHub. Ideally, the method you choose to explore would be one that you and your partner will find useful for your own research.

[2] Novel Data Science Project: With your partner, design and implement a modest data science project using your own data and addressing a theoretically interesting problem in the natural or social sciences. The goal is to incorporate several of the tools learned in this classroom into a project of your own design that, ideally, might also help you move your own research forward. The project should include both exploratory and inferential data analyses. During the last week of class, you and your partner will give a 15-20 minute presentation on your project to your peers. You will also prepare an Rmarkdown or Quarto write-up of the project and share the documented project code on GitHub.

One individual data visualization assignment – 5%

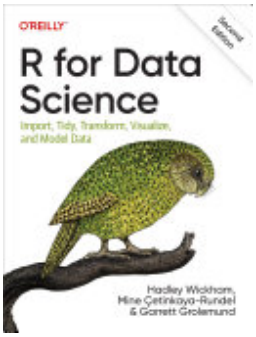
Using data of your choosing, you will develop your own, creative data visualization to share with your classmates and peers in a gallery posted online. This is your chance to be creative and mix science and art! We will discuss this in more detail later in the course.

Extra credit – up to 10%

Working individually or with a partner, you can also earn up to 10% extra credit for developing a simple interactive webpage using {shiny} that accesses a dataset of your choosing and demonstrates a particular statistical or data visualization procedure. Again, we will discuss this in more detail later in the course.

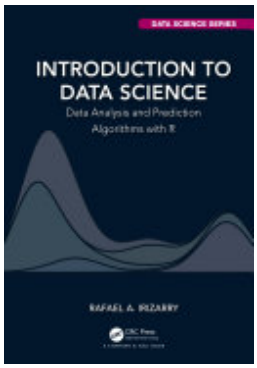
Finally, I may occasionally give you "Check Your Understanding" quizzes through Canvas to gauge your familiarity and comfort with material from lecture, class, and the readings, but these will not impact your grade at all.

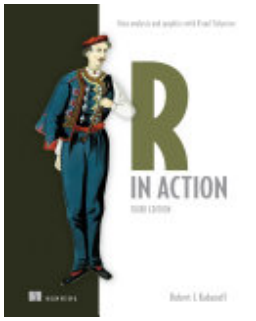
Required Course Materials



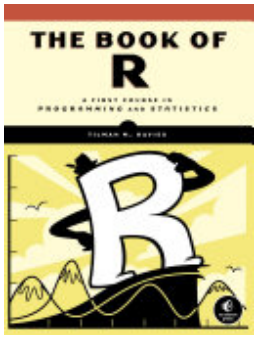
R for Data Science
ISBN: 9781492097372
Authors: Hadley Wickham, Mine Çetinkaya-Rundel, Garrett Grolemund
Publisher: O'Reilly Media, Inc.
Publication Date: 2023-06-08
OPTIONAL

Introduction to Data Science
ISBN: 9781000708035
Authors: Rafael A. Irizarry
Publisher: CRC Press
Publication Date: 2019-11-20
OPTIONAL

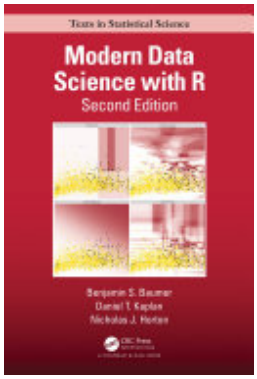




R in Action, Third Edition
ISBN: 9781617296055
Authors: Robert Kabacoff
Publisher: Simon and Schuster
Publication Date: 2022-05-03
OPTIONAL



The Book of R
ISBN: 9781593277796
Authors: Tilman M. Davies
Publisher: No Starch Press
Publication Date: 2016-07-16
OPTIONAL



Modern Data Science with R
ISBN: 9780429577505
Authors: Benjamin S. Baumer, Daniel T. Kaplan, Nicholas J. Horton
Publisher: CRC Press
Publication Date: 2021-04-13
OPTIONAL

Recommended Course Materials

Note that all of the "required" materials listed above are, in truth, optional! They are all excellent resources for learning basic to intermediate level statistics and R programming and are among my favorite data science books. Most of our readings for the semester will be drawn from these books or from those listed on the **Other Resources** page on the course website on ***GitHub***. I will post PDF versions of all required readings on Canvas site so that you can download them for printing, reading, and annotating..

Other Texts and Resources

A list of many other useful texts and links is given on the **Other Resources** page of the course website.

Periodically, publishers such as No Starch Press and Manning Publications have amazing deals for purchasing either print or electronic books, so keep your eyes peeled. I will let you know of any I come across. The UT library also can digital access to a large number of data science and programming books, including many listed on the **Other Resources** page.

Technology and Accessibility

For those with further interest in a topic, I will also often post additional, optional readings for some weeks and topics. For most effective learning, reading assignments and accompanying programming examples and exercises should be completed before coming to class. I strongly recommend that you carefully work through code examples included in the readings and in online course modules as you prepare for class.

Access to a reasonably powerful personal computer or laptop and to the internet are critical for successfully honing your data wrangling and analysis skills. For much of this class (assuming we can return to in-person instruction as anticipated), we will be meeting in a laboratory where computer workstations are available, but for work outside of class you must have access to a computer at home or another remote site. The primary (and free!) software we will use for this class, i.e., the statistical programming software ***R*** and the integrated development environment (IDE) ***RStudio***, are available for all major computer operating systems (MacOS, Windows, various flavors of Unix). If you anticipate or experience accessibility issues or do not have access to a computer of your own, please let me know as soon as possible so that we can create a solution together. Also, please let me know if you notice issues with Canvas or materials posted there... chances are, if you notice a problem others may also be experiencing it. I will try to correct any issues as quickly as possible once they come to my attention.

Additional Resources

I have arranged for the class to have semester-long access to DataCamp, an excellent online learning platform for data science. From time to time, I will identify and recommend relevant ***R*** courses and modules on the platform where you can practice your skills, but these are entirely optional.

Final Exam Date and Time

There is no final exam for this class. However, we will use our final exam time for student presentations.

Notice of Academic Accommodations from Disability and Access

(D&A)

The university is committed to creating an accessible and inclusive learning environment consistent with university policy and federal and state law. If you are a student with a disability, or think you may have a disability, and you need formal accommodations, you have a right to have these met! Please reach out to the Disability and Access group on campus (diversity.utexas.edu/disability/) for more information and to discuss your particular situation and share your Accommodation Letter with us as early as possible in the semester so we can organize working with your approved accommodations and needs in this course.

University Policies and Resources for Students Canvas Page

This Canvas page is a supplement to all UT syllabi and contains University policies and resources that you can refer to as you engage with and navigate your courses and the university.

Course Policies, Disclosures, and Expectations

Learning Success

Your success in this class is important to me! We will all need to be adaptable because we all learn differently, and the current, dynamic public health situation we are now in imposes some unusual challenges. Even without all of the stress of the current moment, we all sometimes feel overwhelmed, experience anxiety or depression, struggle with relationships and personal situations, or just need help navigating challenges. If you are facing such challenges, you do not need to handle them on your own — the UT Counseling and Mental Help Center offers a variety of help and support programs you can access. Getting help is a courageous thing to do! You should feel free to reach out to me if you would like our help in getting you any assistance you might need.

Additionally, if there are aspects of this course that you feel are barriers to your learning or that exclude you in some way, please let me know as soon as possible! I will work with you to develop strategies to meet both your needs and the requirements of the course. I also encourage you to reach out to the student resources available through UT. Several are listed on this syllabus, but I am happy to help connect you with a person, Center, or other resource if you would like.

Flexibility and Adaptability

In acknowledgement of COVID-19 and its impact on The University of Texas at Austin community, this course will reaffirm one of the core values here at UT Austin: responsibility. Our responsibility to ourselves and each other is to put our humanity in the forefront of our academic pursuits. This

semester I commit to being adaptable in this time of great need, which I hope is reflected in the course policies below around attendance, grading, and assignments/exams.

If you experience any hardships such as illness, accident, or family crisis, please know that these policies may be amended and therefore you should communicate with me as soon as you feel comfortable doing so. If for any reason you do not feel comfortable discussing an issue with me, please visit Student Emergency Services.

For additional campus resources to help you be supported and prepared during these difficult times, please visit coronavirus.utexas.edu/students.

Communication

The course Canvas site can be found at utexas.instructure.com. To contact me, please email me through Canvas, and I will respond the same way. You are responsible for ensuring that the primary email address you have recorded with the university is the one you will check for course communications because that is the email address that Canvas uses.

Student Hours and Asking for Help

Student hours (a.k.a., "office hours") are blocks of time that I have set aside expressly to hang out in my office or in a Zoom meeting room for consultations with students on a "drop-in" basis outside of class time. Please do not hesitate to reach out, through student hours, if you have questions about course materials, want additional help with or feedback or class assignments or topics, need advice beyond the class, need a letter of recommendation, want to talk about research, life, etc. You are also welcome to schedule one-on-one appointments with me outside of student hours via Calendly.

Diversity, Equity, and Inclusion

It is my intent that students from diverse backgrounds and perspectives be well-served by this course, that students' learning needs be addressed, and that the diversity that students bring to this class can be comfortably expressed and be viewed as a resource, strength and benefit to all students. Please come to me at any time with any concerns.

Please be aware, too, that Texas Senate Bill 17, the recent law that outlaws diversity, equity, and inclusion programs at public colleges and universities in Texas, does not in any way affect content, instruction or discussion in a course at public colleges and universities in Texas. Expectations and academic freedom for teaching and class discussion have not been altered post-SB 17, and students should not feel the need to censor their speech pertaining to topics including race and racism, structural inequality, LGBTQ+ issues, or diversity, equity, and inclusion.

Classroom Expectations

Every student has the right to learn and the responsibility to not deprive others of their right to learn. In order for you and your fellow students to get the most out of this class, I ask that you abide by the following policies:

- Please attend class sessions and arrive on time. As noted above, all of our class sessions are designed to be synchronous activities and to be conducted **in person** after the first few weeks of the semester, at least as currently planned. You will get the most out of this class and probably perform best if you attend regularly. That said, I fully recognize that these are very unusual and stressful times and folks may have good reasons to miss a session here or there. Thus, I have tried to build some flexibility into attendance and grading policies.
- For any remote sessions we may have, I expect you to have your video camera turned on for the duration of the class, unless you discuss with me beforehand why that is not possible and we agree in writing on a different accommodation. You are more than welcome to use a custom or blurred background if you wish to.
- No personal audio or video recording of class activities is permitted without my prior written approval. Additionally, the materials we use in this class are copyright protected works. Any unauthorized copying or distribution of class materials is a violation of federal law and may result in disciplinary actions being taken against you. Sharing of class materials without my specific, written approval may also be a violation of the University's Student Honor Code and an act of academic dishonesty, which could result in further disciplinary action. This includes, among other things, uploading class materials or assignments to outside websites for the purpose of sharing those materials with other current or future students. That said, most course materials will be made available for personal use as PDF files through the course Canvas site.
- Our class space should be an environment where folks feel confident sharing their ideas and experiences openly without bias or prejudice and without having to experience agonism or microaggressions. I welcome and encourage you to express your ideas and to listen carefully to others, even when your ideas differ. You may disagree with the opinions and perspectives of others in this class, including classmates, instructors, and the authors of assigned readings, and I encourage discussion about points of disagreement. However, I also expect you to be civil, polite, and respectful at all times, and I do not tolerate aggressive or hateful behavior in the classroom. That said, I recognize that sometimes people may unintentionally say something that hurts or offends, and I and others may not respond appropriately when someone else says something upsetting. If something like this happens, please, please, please feel free to call that to my attention! I am also very open to your feedback and suggestions for how to make the class a fully inclusive learning environment.

Finally, please do let me know if you have any problem that is preventing you from performing satisfactorily in this class!

Academic Integrity Expectations

Each student in the course is expected to abide by The University of Texas Honor Code:

"I pledge, as a member of the University of Texas community, to do my work honestly, respectfully, and through the intentional pursuit of learning and scholarship."

Students who violate the Honor Code or University rules on academic dishonesty are subject to disciplinary penalties, including the possibility of failure in the course and/or dismissal from the University. Since such dishonesty harms the individual, all students, and the integrity of the University, policies on academic dishonesty will be strictly enforced. For further information, please visit the Student Conduct and Academic Integrity website at deanofstudents.utexas.edu/conduct.

Unless otherwise indicated, I expect you to work on assignments and projects individually and to produce and submit your own work. There are always many different ways to solve programming and data analysis tasks, but it is generally very obvious when someone has copied code from another source or another student without understanding it. I am attuned to and strongly frown upon that kind of academic dishonesty. You are more than welcome to consult with fellow students and to search on the internet for help with understanding and troubleshooting problems and to reinforce your learning – indeed, I encourage that! However, unless otherwise stated, your submitted work for all assignments must be your own. If you are having trouble understanding something or want pointers to get you started in the right direction, please, please, please come see me! There is little I like more than helping folks problem-solve ***R*** code.

Confidentiality of Class Recordings

Any class recordings associated with this course, including instructor lectures and student presentations, are reserved only for students in this class for educational purposes and are protected under FERPA. The recordings should not be shared outside the class in any form. Violation of this restriction by a student could lead to Student Misconduct proceedings.

Content Warning

Our classroom provides an open space for the critical and civil exchange of ideas. Although it is unlikely, some readings and other content in this course will include topics that some students may find offensive and/or traumatizing. I will do my best to forewarn students about any potentially disturbing content, and I ask all students to help to create an atmosphere of mutual respect and sensitivity.

Religious Holy Days

By UT Austin policy, you must notify me of your pending absence as far in advance as possible of the date of observance of a religious holy day. If you must miss a class, an examination, a work assignment, or a project in order to observe a religious holy day, you will of course be given an opportunity to complete the missed work within a reasonable time after the absence.

Student Rights and Responsibilities

- You have a right to a learning environment that supports mental and physical wellness.
- You have a right to respect.
- You have a right to be assessed and graded fairly.
- You have a right to freedom of opinion and expression.
- You have a right to privacy and confidentiality.
- You have a right to meaningful and equal participation, to self-organize groups to improve your learning environment.
- You have a right to learn in an environment that is welcoming to all people. No student shall be isolated, excluded or diminished in any way.

With these rights come responsibilities:

- You are responsible for taking care of yourself, managing your time, and communicating with the teaching team and with others if things start to feel out of control or overwhelming.
- You are responsible for acting in a way that is worthy of respect and always respectful of others.
- Your experience with this course is directly related to the quality of the energy that you bring to it, and your energy shapes the quality of your peers' experiences.
- You are responsible for creating an inclusive environment and for speaking up when someone is excluded.
- You are responsible for holding yourself accountable to these standards, holding each other to these standards, and holding the teaching team accountable as well.

Land Acknowledgment

UT and the City of Austin are located on Indigenous land. I acknowledge, value, and pay my respects to the Alabama-Coushatta, Caddo, Carrizo/Comecrudo, Coahuiltecan, Comanche, Kickapoo, Lipan Apache, Tonkawa, Tigua Pueblo, Ysleta Del Sur Pueblo, and all of the other American Indian and Indigenous peoples and communities who have been or have become a part of these lands and territories in what is now known as Texas, here on Turtle Island.