Introduction to Data Science AB

View entire course

Course Description

Course overview:

The **Introduction to Data Science** course will emphasize the use of statistics and computation as tools for creative work, as a means of telling stories with data. Its content will prepare students to "read" and think critically about existing data stories. Ultimately, this course will be about how we tell good stories from bad, through a practice that involves compiling evidence from one or more sources and often requires hands-on examination of one or more data sets. It will develop the tools, techniques and principles for reasoning about the world with data. It will present a process that is iterative and authentically inquiry-based, comparing multiple "views" of one or more data sets. Inevitably, these views are the result of some kind of computation, producing numerical summaries or graphical displays. Their interpretation relies on a special kind of computation, simulation, and modeling to describe the uncertainty in each view. This kind of reasoning is exploratory and investigatory, sometimes framed as hypothesis evaluation and sometimes as hypothesis generation. R, the statistical programming language used by academics and industry, will be used to bring data science to life.

The main goal of the **Introduction to Data Science** course is to teach students to think critically about and with data. This new and innovative curriculum will meet the Common Core State Standards (CCSS) for High School Statistics and Probability, relevant second-year Algebra probability standards, the Modeling standard, relevant mathematics standards. This course emphasizes the CCSS High School — Statistics and Probability Standards that involve the study of data science. Students authentically apply the Standards for Mathematical Practice throughout the course.

Introduction to Data Science will develop the tools, techniques and principles for reasoning about the world with data, with a special emphasis on data collected through participatory sensing, an emergent and important data type encountered in many disciplines, including business, biology, engineering, computer science, and statistics. We will present a pedagogical process that is iterative and authentically inquiry-based, and this student-based inquiry will lead to comparing multiple "views" of one or more data sets. Inevitably, these views are the result of some kind of computation, producing numerical summaries or graphical displays. Their interpretation relies on a special kind of computation – simulation– to model the uncertainty in each view.

The use of participatory sensing data will put data collection into the hands of students and, as a consequence, students will function as researchers making truly original discoveries about the real world. Students will learn to generate hypotheses, to fit statistical and mathematical models to data, to implement these models algorithmically, and to evaluate how well these models fit reality. Our course will rely on R, an open-source programming language has long been the standard for academic statisticians and analysts in industry. Through R, students will learn to compute with data to develop graphical and numerical summaries to both communicate findings and to generate further exploration.

This course is an introduction to the practice of data science: reasoning about the world with data. The course applies concepts from statistics and probability, alongside computation and visualization, as a means of processing data to learn about the world. An emerging academic discipline, data science creates a basis for thinking about and with data and understanding the ways in which data operate to shape our world. There are four goals encompassing this course divided into four units:

Unit 1: Data are all around us.

Unit 2: Making inferences using models and plots.

Unit 3: Understand data sources, special data structures, and modes of data collection.

Unit 4: Making inferences using randomization and simulation.

Unit 1 will introduce the idea of data and motivate the rest of the course. While most people think of data simply as a spreadsheet or a table of numbers, almost anything can be considered data, including images, text, GPS coordinates, and much more. Our world has become increasingly data-centric, and we are constantly generating data, whether we know it or not. From posts on Facebook, to shopping records created when you swipe your credit card, to sensors embedded in highway on-ramps, we leave behind a stream of data wherever we go. And data are used to generate stories about our world, whether it is for political forecasting, marketing, scientific research, or Netflix recommendations.

This unit will motivate the idea that data and data products (charts, graphs, statistics) can be analyzed and evaluated critically. We want to know how the evidence was collected, what the perspective or bias of the creator might be, and look behind the scenes to the process used to create the product. Even the way in which data are stored embeds within in it decisions on the part of the data creator.

Students will learn how to store data in basic structures, how to access and manipulate data within those structures, and how to share data with others – an important component of scientific reproducibility.

Unit 2 will continue introducing methods of graphical and numerical data summary, including mosaic plots, Q-Q plots, and the Grand Tour, a method of exploring multi-dimensional data using graphical summaries. Moving from simple graphical summaries, students will begin working with simulation methods to understand visual inference, generating plots that show the true data interspersed with simulated plots. From this exercise, students will learn to use simulation to calibrate their interpretation of a view, a numerical or graphical summary, so that they understand what story-less data (pure noise, no association) look like.

In Unit 2, the use of models will come to the foreground. Students will be introduced to linear models, the most common form of modeling in introductory statistics classes, but will also use the computer technology available to them to learn more complex modeling techniques, including smoothing techniques and tree-based models.

Unit 3 will build on the critical thinking skills first introduced in Unit 1, as students begin looking for data that interest them on the web. Whether through Application Programming Interfaces (APIs) like those offered by Twitter or Facebook, government data sources like the Environmental Protection Agency or the Census, or the many other varied data sources made available online, students will learn to access data in advanced data-storage structures, critically evaluate whether these data can address their questions of interest, and begin analyzing and evaluating data sources. They will practice asking critical questions like, which sources do we trust? How do data from different organizations compare? What arguments have been made with these data previously and by whom? To support this cycle of inquiry, students will examine the basic publication mechanisms for data and develop a set of questions to ask of any data source -- computation meets critical thinking.

Then, this unit will discuss special data structures that can be used to aid in inference. The simulation techniques for calibrating different views of a data set take on new life when some form of random process was followed to generate the data. Polls, for example rely on random samples of the population, and clinical trials randomly assign patients to treatment and control. A simulation strategy that repeats these random mechanisms can be used to assess uncertainty in the data, assigning a margin of error to poll results or identify in new drugs that have a "significant" effect on some health outcome.

Unit 4 will focus even more on the ideas of simulation and the application of probability. First, this unit will discuss the ways in which a computer can generate random phenomena -- How does a computer toss a coin? Simple probability calculations will be used to describe what we expect to see from random phenomena and then students will compare their results to simulations. The point is both to rehearse these basic calculations as well as to make a formal tie between simulation and theory in simple cases. This unit will focus on the relationship between frequency and probability. Students will be simulating, essentially, independent trials and will create summaries of those simulations. In turn they should understand that the frequency with which an event occurs in a series of independent simulations tends to the probability for that event as the number of simulations gets large. (This is the Law of Large Numbers, a topic that is often taught in introductory statistics courses.)

From here, students in this course will simulate a variety of random processes to aid in formal statistical inference when some random mechanism was applied as part of the data design. In short, probability becomes a ruler of sorts for assessing the importance of any story we might tell. In this approach to probability, we will be using a combination of direct mathematical calculation and computer simulations in order to give students a deep sense of the underlying statistical concepts.

Course content:

Computer-based lab exercises and practicums using RStudio:

Collaborative Group Work—Motivated by the analysis of a data set, groups generate a question or the instructor assigns a question to study. Students answer the question in a group discussion with guidance and feedback from the teacher. For example, the instructor may demonstrate a new computational method or statistical model that could be appropriate for the problem at hand. Students would get immediate experience applying this method or model, and could use it during individual student work. Group discussion supports students' development of constructing viable arguments, critiquing the reasoning of others, and attending to precision as they develop the study of data science.

Individual Student Work—Students work to solve a more elaborate question than the one presented in the group lab experience. This question will either be assigned by the teacher or developed as a hypothesis by the students. Then, each student works to solve it individually, using the skills they developed in the collaborative lab exercise. The deliverable from the individual lab exercise is a "pitch," called a story memo, which is a single data visualization and a written explanation, at most a paragraph in length. The lab exercise develops students' skills in using appropriate tools strategically, making sense of problems in context, and persevering in solving problems. The story memo helps assess students' abstract and quantitative reasoning, and it allows students to practice constructing viable arguments.

Computer-based lab exercises and practicums will be one of the most common assignments in this class, taking place almost 2-3 times per week. Working through problems this frequently will make concrete applications very familiar to students, integrating their statistical analysis skills into their mental toolbox.

Oral presentations:

After learning a data science concept, students will apply their learning by engaging in short, informal 1-3 minute presentations based on a data set. As with the computer-based lab exercises, the oral presentation will be a "pitch," a short, unpolished discussion of a current issue and suggestion for potential areas for additional exploration. Students are required to present evidence for every claim that they make based on computation, data, visuals, etc. Their presentations take on a journalistic context, telling a new story. This allows the students to evaluate different types of evidence. Students will model with statistics by analyzing the relationships to draw conclusions, and then brainstorm where the analysis could go next. Again, this involves the use of abstract and quantitative reasoning, allows students to practice the construction of viable arguments, and involves them making sense of problems in context.

Oral presentations will also take place frequently, perhaps once a week. Likely, every time there is a lab exercise one or two students will be selected to do oral presentations, cycling through students so they all have the opportunity to present. By speaking about their analyses, students will improve their skill in constructing arguments using data.

Design Project:

The design project is an opportunity for students to develop their skills in testing a hypothesis. Students will engage in a participatory data collection phase that is purely about exploration. They will deploy a participatory sensing survey and collect data for some short period of time, perhaps a week. Then, students will determine what is most important in the data they just collected. Using this preliminary analysis, they will design a research question in collaborative groups. The design project engages the notion of observation and how the tools turn into data; it is a probe into hypotheses and civic implications. For example, a project that collects data about snack habits may not truly be about snacking, but rather surfacing the question, what are the situations that contribute to poor eating habits? These are the kinds of real world problems that come to the surface, which are followed with interviews and repeated quantification through iterative development of the data collection system.

The design project will take place twice throughout the course, once at the end of each term. It will be a cumulative project, as the analysis will incorporate all the data analysis techniques they have used so far. Their skills of modeling with mathematics, attending to precision, constructing viable arguments, and making sense of problems in context.

Participatory Sensing Written Topic Report:

This formal report will develop students' skills in the majority Standards for Mathematical Practice. After their Participatory Sensing experience, students are expected to write a formal written report that includes:

1.) Report Abstract

2.) An Introduction

3.) Methods Used (primarily data collection, data cleaning, and analysis techniques)

4.) Results

5.) Conclusion

6.) Appendix (campaign/survey questions, data visualizations and representations, etc.)

Again, this will correspond with an end-of-term project, so it will take place twice throughout the course. And again, it will incorporate all the skills they have learned so far, and underscore all the standards for mathematical practice.

End of Unit Report and Oral Presentation:

Students will evaluate media reports based on data. After reading a presentation of statistics in a newspaper, magazine, or scientific journal, students will ask questions like: Do we believe this, do we not, how could we corroborate it? Again, this project will always relate back to the skills from the unit, as students become more sophisticated in their analysis techniques they will apply these new skills to reports they read. In an ideal case, students will try to quickly download the source data and do some basic analysis in RStudio. Other data sources can also be examined.

This assignment will take place at least once per unit. After each new analysis method is covered, the students will examine a relevant media report based on data, critiquing the analysis or presentation. Students' conclusions will be presented either in a short written report or oral presentation.

The purpose of the Introduction to Data Science Course

The purpose of the **Introduction to Data Science Course** is to introduce students to dynamic data analysis that is often used to make decisions or predictions. The four major components of this course will be based on the conceptual categories called upon by the Common Core State Standards High School (CCSS)—Statistics and Probability:

1. Interpreting Categorical and Quantitative Data

2. Making Inferences and Justifying Conclusions
3. Conditional Probability and the Rules of Probability
4. Using Probability to Make Decisions

The **Introduction to Data Science** course will emphasize the use of statistics and computation as tools for creative work, as a means of telling stories with data. Seen in this way, its content will also prepare students to "read" and think critically about existing data stories. Ultimately, this course will be about how we tell good stories from bad, through a practice that involves compiling evidence from one or more sources and often requires hands-on examination of one or more data sets.

The **Introduction to Data Science** course will develop the tools, techniques and principles for reasoning about the world with data. It will present a process that is iterative and authentically inquiry-based, comparing multiple "views" of one or more data sets. Inevitably, these views are the result of some kind of computation, producing numerical summaries or graphical displays. Their interpretation relies on a special kind of computation, simulation, to describe the uncertainty in each view. This kind of reasoning is exploratory and investigatory, sometimes framed as hypothesis evaluation and sometimes as hypothesis generation.

1. ***Interpreting Categorical and Quantitative Data***. A handful of data interpretation are standard. Some, including summaries of shape, center and spread of one or more variables in a data set, as well as graphical displays like histograms and scatterplots, are standard in the sense that they provide interpretable information in a number of research contexts. They are portable from one set of data to the next, and the rules for their use are simple. And yet, our interpretation of data is rarely "standard." Data have no natural look -- even a spreadsheet or a table of numbers embeds within it a certain representational strategy. We construct multiple views of data in an attempt to uncover stories about the world.

This course will consider time, location, text and image as data types and will examine views that uncover patterns or stories. Throughout, simulation will be used to calibrate our interpretation of a view, a numerical or graphical summary, so that we understand what storyless data (pure noise, no association) look like.

In addition to summaries and simple graphics, "modeling" will be explored as a means of describing patterns in a data set mathematically. From adding a least squares line to a plot to performing a cluster analysis to spot separate groupings of points in a data set, each model again represents a potential story. These models will be applied like probes and evaluate them using simulation procedures. For example, students might simulate values using the model and compare them to the original data set. Models, then, provide new, albeit somewhat more specialized, views into a data set. (The term "model" can be used to cover the very traditional linear regression or a more elaborate learning algorithm.)

1. ***Making Inferences and Justifying Solutions***. Data are becoming more and more plentiful, supported by a host of new "publication" techniques or services. Post-Web 2.0, data are interoperable, flowing out of one service and into another, helping us easily build a detailed data version of many phenomena in the world. Reasoning with data, then, starts with the sources and the mechanics of this flow. Which sources do we trust? How do data from different organizations compare? What stories have been told with these data previously and by whom?

This course answers these questions, in part, using the tools and techniques already mentioned. The ability to read and critique published stories and visualizations are additions to these tools and techniques. Finally, as an act of comparison, students should also be able to formulate questions, identify existing data sets and evaluate how the new stories stack up against the old. To support this cycle of inquiry, students will examine the basic publication mechanisms for data and develop a set of questions to ask of any data source -- computation meets critical thinking. In some cases, data will exhibit special structures that can be used to aid in inference. The simulation techniques for calibrating different views of a data set, take on new life when some form of random process was followed to generate the data. Polls, for example rely on random samples of the population, and clinical trials randomly assign patients to treatment and control. A simulation strategy that repeats these random mechanisms can be used to assess uncertainty in the data, assigning a margin of error to poll results or identify in new drugs that have a "significant" effect on some health outcome.

In many cases, data will not posses this kind of special origin story. A census, for example, is meant to be a complete enumeration of a population and we can reason in a very direct way from the data. In other cases, no formal principle was applied, perhaps being a sample "of convenience." The techniques for telling stories from these kinds of data will also rely on a mix of simulation and subsetting.

Finally, this course will introduce Participatory Sensing as a technique for collecting data. The idea of a data collection campaign will be introduced as a means of formalizing a question to be addressed with data. Campaigns will be informed by research and data analysis and will build on or augment or challenge existing sources. (The "culture" behind the existing sources and the summaries or views they promote will be part of the classroom discussions.)

It is worth noting that everything described so far depends on computation, a piece of statistical software ran on a computer. Students will be taught simple programming tools for accessing data, creating views or fitting models, and then assessing their importance via simulation. Computation becomes a medium through which they learn about data. The more expressive the language, the more elaborate the stories we can tell.

1. ***Probability***. Since simulation is our main tool for reasoning with data, interpreting the output of simulations requires understanding some basic rules of probability. First and foremost, this course will discuss the ways in which a computer can generate random phenomena -- How does a computer toss a coin? Simple probability calculations will be used to describe what we expect to see from random phenomena and then students will compare their results to simulations. The point is both to rehearse these basic calculations as well as to make a formal tie between simulation and theory in simple cases.

In that vein, this course will motivate the relationship between frequency and probability. Students will be simulating, essentially, independent trials and will create summaries of those simulations. In turn they should understand that the frequency with which an event occurs in a series of independent simulations tends to the probability for that event as the number of simulations gets large. (This is the Law of Large Numbers, a topic that is often taught in introductory statistics courses.)

From here, students in this course will simulate a variety of random processes to aid in formal statistical inference when some random mechanism was applied as part of the data design. In short, probability becomes a ruler of sorts for assessing the importance of any story we might tell. In this approach to probability, we will be using a combination of direct mathematical calculation and computer simulations in order to give students a deep sense of the underlying statistical concepts.

TOPIC OUTLINE

This outline describes the scope of the course but not necessarily the sequence.

I. Interpreting data.

A. Types of data

B. Numerical and graphical summaries

1. Measures of center and spread, boxplots

2. Bar plots and mosaic plots

3. Histograms and Q-Q plots

4. Scatterplots

5 Graphical summaries of multi-dimensional data

C. Simulation and visual inference

1. Mosaic plots and association

2. Scatterplots

D. Models

1. Linear models

2. Nearest neighbors and smoothing

3. Learning and tree-based models

II. Making Inferences and Justifying Solutions.

A. Aggregating data

1. Identification of sources

2. Mechanics of Web 2.0

3. Comparison of sources

B. Data with special structures

1. Random sampling

2. Random assignment and A/B testing

3. Simulation-based inference

C. Participatory sensing

1. Designing a campaign

2. Participation as a data collection strategy

III. Probability

A. Computers and randomness

1. Web services like random.org

2. Pseudo-random numbers (optional)

B. Frequency and probability

C. Probability calculations

IV. Algebra in RStudio

1. Matrices
2. Vectors
3. Algorithms
4. Functions
5. Evaluating and fitting models to data
6. Graphical representations of multivariate data
7. Numerical summaries of distribution and interpreting in context

Assessments

Multiple assessment opportunities are a vital component of an effective, well-balanced instructional and progress monitoring program that supports teachers' ability to plan effectively, monitor student progress in standards-based instruction, determine the efficacy of instruction and intervention matched to student need, and to inform students and parents of their progress. Assessments are critical to gauge students' acquisition of understanding of the course goals. The **Introduction to Data Science** course will use a balanced assessment approach. A balanced assessment approach provides insight to students' development of habits of mind skills described in the eight CCSS Standards of Mathematics Practice.

Formative Assessments: The chief objective of using formative assessments will be to determine whether or not students have met benchmark goals. The data obtained from these assessments is not only for the instructor, but also for students so that they are actively engaged in assessing their own learning. Formative assessments will provide students a metacognitive tool to become reflective practitioners of the course material. The frequency of assessment depends on the learning objectives being taught.

Summative Assessments—The main purpose of summative assessments will be to measure learning outcomes and report those outcomes to students, parents, and administrators. These will occur at different intervals during the yearlong course, for example, at the end of a unit of study, at a mid-term interval, and at the end of a semester. These assessments will allow students to continue to gather data using all senses and to continue to persevere in their learning. Forms of summative assessments include written assessments and performance tasks.

Informal Assessments—The course instructor will informally assess students' development of critical thinking by using tools such as targeted student observations, entrance/exit slips, checks for understanding, etc. These are meant to be quick assessments used to improve instruction and to observe whether students are developing the Standards of Mathematical Practice skills.

Formal Assessments—These assessments provide formal communication between student and instructor. The purpose of formal assessment is to provide students with feedback so that they become reflective about what they do. Formal assessments can be formative or summative. In the cases where the formal assessment is summative, students will use the feedback to identify areas of improvement as they proceed in this course or in future courses such as Advanced Placement Statistics.

Performance Tasks—Since this course will use EDA as an approach to teaching statistics, hands-on activities that allow students to demonstrate their ability to perform certain tasks will be used as an additional assessment piece. Using performance tasks goes hand–in-hand with the simulation and real world problem solving approach used in the course. Performance tasks will require using the technology and software used throughout the course, RStudio.

Written Assessments—As part of a well-balanced assessment approach, this course will include three basic types of written assessments: multiple-choice, open-ended questions requiring short written responses, and constructed-response written assessments, which include written evaluations of reports based on data, problem-based simulations and scenarios.

Instructional Strategies

Successful delivery of the curriculum and the course goals will be accomplished by a curriculum that focuses on inquiry-rich computational thinking practices, specifically Participatory Sensing and Exploratory Data Analysis (EDA), which allows insights to be gleaned through an iterative process of examining data for trends. EDA encourages students to engage in data sets immediately, making plots to get a natural sense of the data, then moving on to more rigorous data analysis. For example, instead of starting with high-powered statistical models, John Tukey [Exploratory Data Analysis, 1977] proposes doing simple descriptive statistics and making many simple graphs (of one variable or several) in order to look for trends. Humans looking at graphs are often better than computers at identifying trends in data. This concept is perfect for students, but it is often not discussed in current high school statistics courses. EDA aligns with and supports Standard for Mathematical Practice one in that it values making sense of problems and persevering in solving them. The goal of EDA is to get an intuitive understanding of the data, to better be able to leverage it to solve problems or explain phenomena.

While there will be some direct instruction, most of it will take the form of short lecture snippets and demonstrations of how to use technology. The main piece of technology that will be used is the statistical programming language R, which will be accessed through the Graphical User Interface called RStudio. John Tukey is one of the fathers of R, and the language has long been the standard for academic statisticians and analysts in industry. When doing statistics, R is absolutely the appropriate tool to be using, and students will have the chance to learn how to deploy it strategically to solve problems.

A recurring theme throughout the class will be students' producing data analysis. After engaging in EDA, they will select the most meaningful plots and clean them up to be shared in a report. In written or verbal form, they will explain why they found those plots to be meaningful, what trend they might suggest, and reasons why that trend might (or might not) be a real phenomenon. This will require abstract as well as quantitative reasoning, as they will need to leverage their contextual understanding to decipher trends in the data. It will also require an eye to precision, as analyses must be reported with accurate units and explanations of significance—aligning with Standards of Mathematical Practice two and six.

Another exercise that will happen recursively through the class is the critique of statistical artifacts, whether they are reported in newspapers, in scientific journals, or in graphics shared virally on Facebook. As each unit passes, students will be asked to use their accumulated knowledge to decide if the conclusions drawn from the data are appropriate, and provide reasoning to support their opinion. This will be a repeated practice in constructing viable arguments and critiquing the reasoning of others. And again, it will require attention to precision, as the weakness in many statistical artifacts is the lack of measures of variability.

Much of this work will take place using data students find on government web sites or other sources for open data, but a major goal is also to have students collect their own data. In order to do this, they will learn about the power of Participatory Sensing, a method for collecting personally-relevant data using cellphones and computers. Inspired by the data they have seen (or perhaps, the data that were absent) students will construct their own data collection survey and deploy it with their fellow students. This will be an exercise in repeated reasoning, as they will likely be reproducing at least a small part of the data they have already seen, and the process for collecting the data they need for their analysis will likely be iterative. Often, the first campaign students design is not structured enough, and they realize with frustration how hard it is to work with the resulting data, so a second campaign can be run to remedy this.

The statistical analyses performed in this class rely heavily on the use of re-randomization, a technique to produce more similar data from data already on hand in order to determine how likely it was that those data occurred by chance. Some data will come from special data collection schemes, like experiments or random surveys, and in those cases students will learn to look for and make use of the structure of randomization to perform analyses.

Throughout the entire course, students will be modeling with mathematics, making sense of problems through simulations, learning about traditional statistical models (like linear regression, decision trees, clustering, and many more), and examining the ways in which data fit and deviate from those models.

Course Materials

Websites

Title	Author(s)/Editor(s)/Compiler(s)	Affiliated Institution or Organization	URL
Introduction to Data Science Curriculum	Mobilize Project	Mobilize Project with UCLA and LAUSD	http://www.mobilizingcs.org/introduction-to-data-science/curriculum

Supplemental Materials

Title	Content
Supplemental Materials	• *RStudio*-An open source graphical user interface that runs R, the statistical programming language. www.rstudio.mobilizingcs.org. www.wiki.mobilizingcs.org. • *Introduction to Data Science Curriculum*-Currently under development in partnership between UCLA and LAUSD through the MOBILIZE Project: Mobilizing for Innovative Computer Science Teaching and Learning.

R in a Nutshell - Joseph Adler, O'Reilly Median, 2009.

Model course View entire course Close