DEPARTMENT OF STATISTICS AND DATA SCIENCE

THE WHARTON SCHOOL

University of Pennsylvania

STAT 410/710 Spring 2022, Q4

Data Collection & Acquisition

(Strategies and Platforms)

Syllabus

Please note that this is both a brand-new course and a preliminary syllabus. There may be some small changes made, that reflect the dynamic nature of the material, but structurally the syllabus is an accurate representation of what to expect in the class.

Instructor: Richard Waterman waterman@wharton.upenn.edu

TA: TBD

Classes meet:

Office hours:

BACKGROUND

With the explosion of interest in data science, the idea of a "data science pipeline" or "workflow" has become a popular way to characterize activities in this area.

This pipeline typical includes components such as data collection, storage, cleaning, reshaping, exploratory analysis, modeling, and reporting. The start of this pipeline, the data acquisition activity, is often given little attention in most analytics related courses. There are multiple courses available to students that focus on data science programing applications (specifically Python and R) and there are courses focused on analytics which usually include statistical and machine learning tools. In these courses the data are typically provided to students directly. Consequently, a large part of the practical side of the pipeline is ignored; specifically, how should you collect your data?

From a business perspective, knowledge about how to thoughtfully collect data translates to improved efficiency and reliability. By using a suitable data collection strategy there can be clear potential time and financial benefits for an organization.

The primary goal of this course is to fill the data acquisition gap and thus enhance the student's understanding of the complete data science pipeline. At the same time, important current ideas such as data confidentiality and ethical considerations will be included.

The course is structured in two parts. There is a "Strategies" component that addresses different data collection strategies. It will discuss sample designs, experimentation, and observational studies. The focus is **not on a deep dive** into the methodological analyses of these data but rather the recognition of, and the pros and the cons of the different approaches. When and why should you use each one?

The second part of the course is about "Platforms" and goes into the practicalities of the implementation of the different strategies. Given the data science perspective of this course, this is focused on web enabled approaches. R and/or Python familiarity are prerequisites for this course, so we will leverage these skills.

GOALS

At the end of the course, students will have a solid grasp of different data collection strategies and when and how they can be applied in practice.

They will have:

(i) Designed and fielded an online sample survey.
(ii) Designed and fielded an online experiment (A/B test).
(iii) Collected data through web scraping activities and/or using an API.
(iv) Summarized their collected data and subsequent inferences, culminating with an in-class presentation.

PREREQUISITES

Familiarity with either R or Python is expected and specifically the R-Studio or Jupyter notebooks platforms. Courses such as Stat 477/777 or 405/705 would meet this

requirement. Statistics, through the level of multiple regression is required. This requirement may be fulfilled with undergraduate courses such as Stat 102, Stat 112, and MBA courses such as Stat 613/621, or by waiving MBA statistics.

COURSE MATERIALS

CANVAS

All course materials, including class notes and assignments, will be available on Canvas.

COMPUTING PLATFORMS

We will use a variety of packages for our data acquisition, that will work with either Python or R. If there are packages that are specifically for R, such as the *survey* package, then for students more familiar with Python, there is always the option of calling R functions from Python directly.

All students should install:

(i) R-Studio available at https://rstudio.com/

And for those who want to interface Python and Jupyter notebooks with R, should have an up-to-date version of the Python:

(ii) The Anaconda Distribution Platform with includes Python 3.8, available at https://www.anaconda.com/distribution/ . It comes with Jupyter notebooks.

Installing the software before the first day of class is an extremely good idea!

PLATFORMS

SURVEYS

The course will preview both the Qualtrics survey platform available through Wharton (https://www.qualtrics.com/academic-solutions/the-wharton-school/), and the online SurveyMonkey platform (https://www.surveymonkey.com/).

EXPERIMENTATION (A/B TESTING)

Google analytics

Optimizely

Vwo.com (https://vwo.com/)

WEB SCRAPING
For Python: Beautiful soup and RoboBrowser

For R: rvest

BOOKS
There is no required textbook for the class.

CLASS NOTES

These will be available from Canvas.

HOMEWORK
When appropriate, homeworks will be run through Turnitin to determine originality. Late homeworks are penalized by 25% up to 1 day late and 50% up to two days late. Homeworks more than two days late will receive a 0.

Homework schedule

Deliverable	Due date
Homework 1	
Homework 2	
Homework 3	
Take home final project/presentation	

Quizzes

There will be 5 ***in-class*** quizzes. They are closed book. Each quiz has 5 questions (multiple choice) and will take 8 minutes. You can drop the lowest quiz score. There will be no make-up quizzes. You have a single attempt for the quiz.

Quiz schedule

Deliverable	Date
Quiz 1	
Quiz 2	
Quiz 3	
Quiz 4	
Quiz 5	

COURSE STRUCTURE

CLASS CONTENT

Introduction to the course

The course will start with an overview of topics to be discussed, together with practical illustrations of how the ideas are used in practice.

Strategies

Survey sample designs

This section will introduce the simple random sample, stratified and cluster sampling designs. Probability proportional to size (audit sampling) and systematic sampling will be presented too. We will discuss conditions under which each design might be preferred. The analysis of a survey that includes survey weights will be discussed.

Randomized experiments

The principles of experimental design provide the foundations for collecting data in the most efficient and informative manner. The essential benefits of randomization will be discussed and introduced in the context of the one-way design. The analysis of the one-way layout will be described.

Blocking in an essential element of experimental design, where the sample is divided across homogeneous subgroups. By blocking, noise can be reduced resulting in more efficient estimates. Blocking plays the same role as stratification in sampling. The two-way ANOVA allows for the investigation of two independent variables and their possible interaction on an outcome of interest. For example, you could be interested in looking at two product attributes, simultaneously.

Observational studies

Though randomized experiments are often called the "gold standard" for data collection, there are times when they are infeasible. Sometimes because of ethical concerns, and other times because they are very costly and time consuming. Observational studies lack the randomization element. For example, the analysts may not be able to assign which subjects go into the treatment rather than control group. Cohort studies and case/control studies are two popular examples of observational studies, which can still allow for reliable inferences.

Platforms

Fielding a questionnaire

We will look at two popular survey platforms, Qualtrics and Survey Monkey. Students will be expected to use one of these to field their own survey. Solicitation of respondents and incentives will also be discussed.

Experimentation, A/B testing and implementation

A/B testing, though developed almost a century ago, has become very popular in product development, particularly when that product happens to be online and easily modified through code rather than physical construction. All large tech companies employ A/B testing to learn about their customers, improve their products and develop monetization strategies. This module will look at experimentation in practice with reference to the Google analytics and Optimizely platforms.

Web-scraping and APIs

Given the abundance of data available online, we are often interested in automating its retrieval. There are typically two ways of interacting with websites. One way is to simply download and extract relevant data from a web page (web scraping) and the other, when available is to use an Application Programming Interface (API).

In this module we present web scraping, introducing XML and HTML and discussing their structure, tags and attributes. We present the idea of a "parser" which is a programming interface for working with such documents. The popular BeautifulSoup (for Python) and rvest (for R) will be used as we go through this web scraping process.

Though web scraping is sometimes necessary, it is better to avoid it if possible and communicate directly with the host organization. Many websites make their functionality available through an API. An API allows you to interact with the website programmatically and hence automate activities.

In this module we will illustrate the use of an API, in particular by interacting with the popular Robin Hood trading platform.

Confidentiality and ethical issues in data collection.

Collecting data can be a minefield when it comes to legal and ethical concerns. In this module we will discuss some commonly encountered environments like HIPPA and the GDPR. Best practices concerning data confidentiality, subject recruitment and ways to ensure subgroup representation will be presented.

COURSE TIMETABLE

CLASS	Course component	Topics	• Specifiics
Class 1	INTRODUCTION	Introduction to the course with examples and use cases	• Course objectives • Expectations • Examples and use cases
Class 2	STRATEGIES	Survey sample designs	• Simple random sample • Stratified samples • Cluster samples • Audit samples
Class 3		Software implementation	• R package*: survey*
Class 4		Randomized experiments	• The one-way layout • Blocking and two-way designs
Class 5		Software implementation	• R routines: anova and lm
Class 6		Observational studies	• Cohort studies • Case/control studies
Class 7	PLATFORMS	Fielding a survey online	• Survey Monkey • Qualtrics
Class 8		Experimentation and A/B testing (I)	• Google analytics • Optimizely • Vwo.com
Class 9		Experimentation and A/B testing (II)	• Google analytics • Optimizely • Vwo.com
Class 10		Web-scraping and APIs (I)	• Beautiful soup • Rvest

			• The RobinHood API
Class 11		Web-scraping and APIs (II)	• Beautiful soup • Rvest • The RobinHood API
Class 12		Data collection: confidentiality and ethics	• Guest lecture
Classes 13 and 14		Project presentations	

GRADING

The final grade will be weighted using 50% from the three assignments (each count as 16.7%), 30% from the final project and 20% from the quizzes. All assignments will be included in the final grade. There is **no** "drop the lowest score" policy for the assignments, but you can drop the lowest quiz score. There will be **no** extra credit opportunities at the end of the course. Grade queries must be submitted within one week of the homework solutions being posted.

You can expect that a final score in the 90's will receive some form of A, 80's some form of B and 70's some form of C. I reserve the right to curve the grades. There is no pre-specified percentage of students that will receive a particular grade – everyone could get an A or everyone could get a C, but that's unlikely!

CLASSROOM EXPECTATIONS

There is no formal participation component to the final grade, but questions are strongly encouraged.

COURSE CALENDAR Q4 SPRING 2022

DATE	Class	Deliverable
3/*	1	
3/*	2	
3/*	3	Quiz 1 in class.
3/*	4	
3/*	5	Quiz 2 in class. HW 1 due.
3/*	6	
3/*	7	Quiz 3 in class.

4/*	8	HW 2 due.
4/*	9	
4/*	10	Quiz 4 in class.
4/*	11	
4/*	12	HW 3 due.
4/*	13	Project presentations. Quiz 5 in class.
4/*	14	Project presentations.
4/*		Final project due .