Lecture 1 - Introduction to Criminal Data Analytics

Go over the syllabus

Go over the Canvas page

Ok, jumping into the material.

This course is about crime analytics, a shortened version of saying it's a class on criminal data analytics. It’s not a class on criminal data science. But what’s the difference?

**Data science - A sparkling cocktail of math, statistics, and computer science tools applied to data sets in order to extract knowledge**

Data scientists want to be “on the bridge” - they want to be involved in the process of making changes.

It’s about seeing previously unseen patterns in a sea of detail. Sometimes we’ll even call it things a “data lake”

Data scientists are aggressive, out-of-the-box thinkers who aren’t afraid to take risks, to get dirty, to make mistakes and generate error messages, and they have an almost cavalier “damn-the-torpedoes” attitude when it comes to blending together solutions. And they get excited about what they’re doing, and they eagerly share what they come up with other data scientists. It’s much more about collaboration than competition.

**Why are data scientists suddenly so necessary?**

Because we are drowning in data. There is so much data out there that companies, government agencies, non-profits, NGOs, schools, and everybody else is generating in real time. Somebody has to come along and sift through all that data and tell the bosses what it means.

Statistics are relatively old techniques (regression is about a century old). What’s changing is the volume of data that can be generated so fast. Data science is as much about data access and data management as it is about data analysis.

We’re not operating in megabytes and gigabytes anymore. We’re operating in petabytes and zetabytes.

The most fundamental common trait of data scientists is the ability to code - to tell the computer how to access, manage, interact with, and analyze data. And that’s what you’re going to learn how to do.

**Data analytics - Specific application of data science tools to a data set with a specific goal in mind**

These are almost always applied endeavors, not theoretical exercises

The point is to do something with the knowledge, not just pontificate and say “oh look, isn’t that interesting.”

Our goal as criminal data scientists (generally speaking) is to extract knowledge about either (1) criminological processes (how crimes occur) or (2) the criminal justice system (how our society responds to crime and criminal offenders)

So, the criminal data scientist then performs criminal data analytics, by applying mathematical, statistical, and computer science tools to crime or criminal justice data sets to generate knowledge that can be applied to crime policies

**Why is Data important?**

**Where does data affect your life?**

Virtually everything you do generates data. Shopping, banking, surfing the internet, checking Facebook, buying a Coke out of the vending machine, going to the doctor, riding an Uber, checking your phone. It all generates data.

Netflix, Pandora, Amazon, Google, Apple, Facebook, Target, Walmart - they’re all after your data. They either want to sell you stuff, or they want to sell you (your digital identity) to someone else, so that someone can sell you stuff.

So what does that have to do with crime and criminal justice?

Well here are a few examples.

Identity theft.

Prescription drug abuse

Cyberbullying

Live streaming of assaults

Crime report tweeting

Traffic light cameras

Terrorism (radicalization on the internet)

Human trafficking

These are just a few examples, but the potential is boundless. Data science has the potential to revolutionize criminology and criminal justice in the same way it’s revolutionized so many other industries, from business and commerce to healthcare to sports. But it's yet to be realized.

**The Current State of Criminal Data Science**

To be honest, it’s in its infancy. Since the early 1900s and the classic school of criminology, criminologist have relied on data sets that were small and static. A few hundred cases was considered a large data set. And once you collected the data set, it sat there. The data didn’t change. Traditional statistics (mostly of the frequentist persuasion) were applied to try to answer questions, and policy decisions may or may not have been influenced by results.

As social scientists, we’ve integrated fancier statistical methods, but other than that, the paradigm hasn’t really changed. Still small data sets, still static, still mostly traditional statistics, with a few cosmetic updates to cover up the fact that the statistical methods are highly flawed or are based on unrealistic assumptions.

But there are few of us out there, in this brave new world, who are genuinely interested in the application of state-of-the-art, bleeding edge data science methods to answer criminal justice or criminological questions.

This often scares the bejeezus out of traditional criminologists, who either think what we’re doing is too weird to be useful, or not sound enough theoretically, or to be perfectly frank, I think because they think data science will replace the need for traditional criminology.

It won’t, by the way. We still need theory, we still need foundational understanding of the CJ system and criminogenic processes. Without those to guide our analytics, we’re hackers.

Data science is just another tool box. It allows questions to be answered. Sometimes, it allows us to answer old crime questions in new ways. Sometimes it allows us to answer new questions entirely.

But we’re still out there. We’re trying to figure out how to access new sources of crime-related data, how to analyze them, how to visualize them, how to map them, and how to use the results to shape policy.

Governmental organizations, policing agencies, security firms, and so on, have grasped the utility of data science (at least in concept). Academia has not.

But the cow is out of the barn, and there’s no putting it back in.

**What kinds of data do criminal data scientists have access to?**

Sure, we’ve still got survey data. And that’s great. And we’ve got other, more traditional sources of data that can be leveraged.

* Court records
* Arrest records
* Official (government) statistics
* Victimization reports
* Vital statistics (lethal violence)
* Health records (assaults, drug use)

But we’re adding some less traditional, less obvious sources of criminal data.

* Social media
* Internet searches
* Web page data (like online classified ads, blogs, forums)
* Deep web data
* Commerce data

Where data science can really shine is by bringing all of these together. What patterns emerge when we combine arrest records with health records? What if we combine social media data and terrorism data? How do the crime trends in official data compare to what we see in victimization data, or in news articles?

This course is intended to be an overview or introduction to criminal data science, or crime analytics - how do we as criminal data scientists apply data analytic methods to crime or justice related data sets using the newer tools that data science gives us.

To do that, we’ll be covering a variety of tools and techniques, and we won’t do any of them in real substantial depth. So this is more of the “mile wide, inch deep” survey course that orients you to the techniques, exposes you to them, but doesn’t ask you to be an expert in any one. In fact, you could make entire courses for most of these topics.

What you will be much more familiar with is how to code. Nearly all of what we’re going to do will be done in R. R is the preeminent language for statistical data analytics today, and is rapidly growing in its popularity and ubiquity.