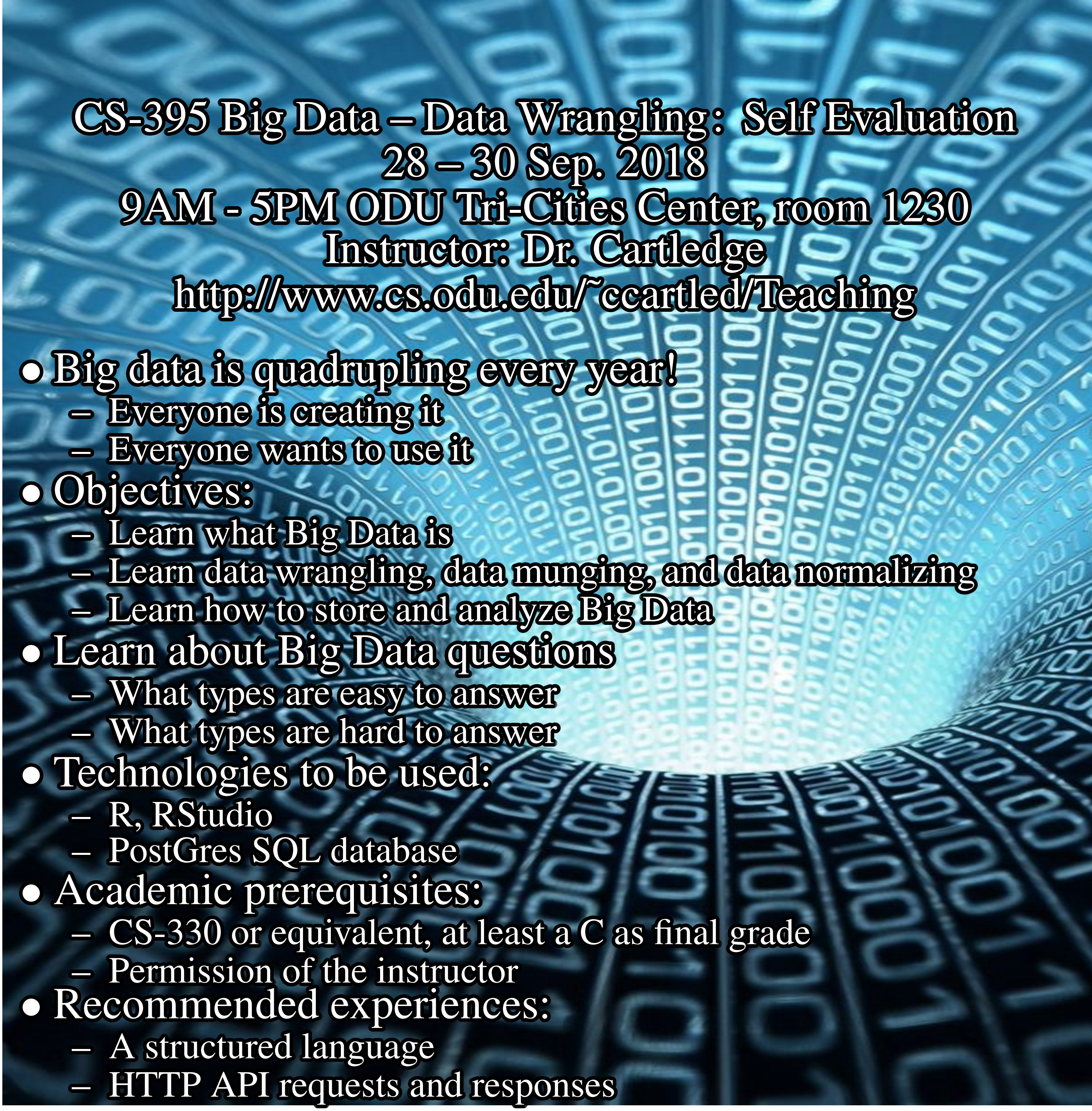

CS-395 Big Data – Data Wrangling: Self Evaluation
28 – 30 Sep. 2018
9AM - 5PM ODU Tri-Cities Center, room 1230
Instructor: Dr. Cartledge
http://www.cs.odu.edu/˜ccartled/Teaching

- Big data is quadrupling every year!
  - Everyone is creating it
  - Everyone wants to use it
- Objectives:
  - Learn what Big Data is
  - Learn data wrangling, data munging, and data normalizing
  - Learn how to store and analyze Big Data
- Learn about Big Data questions
  - What types are easy to answer
  - What types are hard to answer
- Technologies to be used:
  - R, RStudio
  - PostGres SQL database
- Academic prerequisites:
  - CS-330 or equivalent, at least a C as final grade
  - Permission of the instructor
- Recommended experiences:
  - A structured language
  - HTTP API requests and responses

Other notes:

- This is a hands-on programming course. You will establish a Twitter developer’s account.
- You will use Twitter’s API to search for specific data.
- You will write real-time programs to conduct sentiment analysis of tweets.
- You will monitor sentiment over time.
- You will use different languages.

1 Introduction

The field commonly known as "Big Data" has its origins in the world of business mergers and acquisitions. Doug Laney in 2001 was credited with identifying the original 3Vs of Big Data: volume, velocity, and variety. Over the course of time, additional Vs expanded the original three, and the original definitions changed and evolved as well. In this course, we will cover aspects common to all Big Data investigations, including: defining Big Data, surveying tools and techniques for processing Big Data, and visualizing selected aspects of Big Data. We will focus primarily on velocity and variety.

The course will use tweets from Twitter as a stand-in for velocity, and use the text from the tweets to provide variety. Selected tweets will be downloaded in real-time, stored in a local database, and then analyzed in near real-time. Ideas, concepts, and experiences gained from intensive classroom training and programming projects will be generalized to enterprise level systems to solve problems at scale.

Static web pages will be used to collect data from pages created using casscading style sheets (CSS) to demonstrate how data can be collected in an automatic and large scale process.

2 Highly recommended background

Significant portions of the boot camp will focus on evaluating, and modifying R scripts using the RStudio integrated development environments (IDEs). A fluency in R is not a requirement; but programming experience in a structured language (C++, Java, C#, Python, etc.) is highly recommended.

3 Boot Camp Schedule

A detailed class schedule is provided in Table 1.

Table 1: The 3 day boot camp schedule. Attendees who are taking the boot camp as part of a 2 credit course will have pre and post boot camp assignments. The texts will be used extensively during days 2 and 3[1, 2].

Day 1	**Day 2**	**Day 3**
What is Data Wrangling? What is BD? What are BD Vs?	Textual sentiment analysis on Tweets using	Processing of static web pages using CSS selectors and R
Lunch	Lunch	Lunch
BD sources. BD tools. Create Twitter development accounts.		Conclusion. Presentation of certificates.

References

[1] Simon Munzert, Christian Rubba, Peter Meißner, and Dominic Nyhuis, *Automated Data Collection with R*, John Wiley & Sons, 2014.

[2] Julia Silge and David Robinson, *Text Mining with R: A Tidy Approach*, "O'Reilly Media, Inc.", 2017.