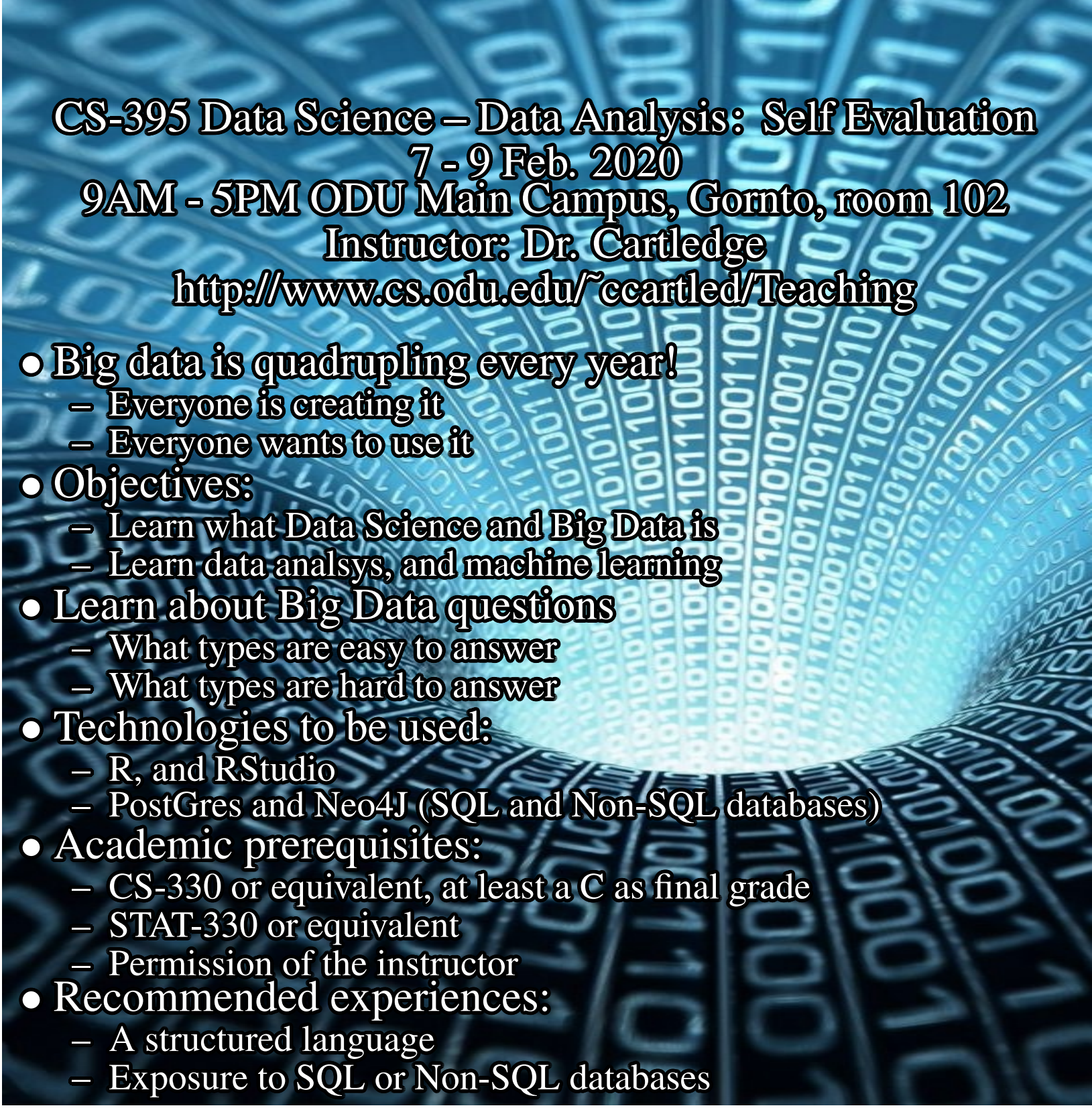

Other notes:

- This is a hands-on programming course.
- You will use R and RStudio.
- You will acquire and analyze data from different on-line and database sources.
- You will populate SQL and Non-SQL databases to answer FOAF and other questions.
- Text: Learning Predictive Analytics with R, by Eric Mayor (ISBN: 9781782169352)
- Text: Big Data Analytics with R, by Simon Walkowiak, (ISBN: 9781786466457)

1 Introduction

The field commonly known as "Data Science" and "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 Data Science and 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 start by analyzing the R built-in Iris and Titanic datasets. Afterwards, we will download datasets from the Internet to experiment with. We will use these datasets to visualize the data in different ways, to understand cluster analysis, linear regression, classification trees, and basic textual mining. After we have gained insight into how to analyze small datasets, we will progress into using Hadoop, Postgres, and Neo4J to handle datasets beyond R's ability.

2 Highly recommended background

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

3 Recommended background

The R programming language has limitations in terms of memory support, and CPU utilization. An understanding, or experience with a structured query language (SQL) database will be helpful when dealing with R memory issues. The boot camp will not teach SQL, but will use it as way to contrast and use Postgres and Neo4J (a non-SQL database) to overcome R's limitations.

A deep understanding of statistics is not required, but attendees need to be comfortable with the ideas of means and standard deviations.

4 Texts

These texts will be used during the boot camp:

- Learning Predictive Analytics with R[1]
- Big Data Analytics with R[2]

Some schools have soft copies of the text available via O'Reilly or Safari.

5 Class Schedule

A detailed class schedule is provided in Table 1.

Table 1: The 3 day class schedule. Attendees who are taking the boot camp as part of a 2 credit course will have pre and post boot camp assignments. The two texts Learning Predictive Analytics with R[1] (LPAR), and Big Data Analytics with R[2] (BDAR) will be used extensively.

Day 1	**Day 2**	**Day 3**
What is Data Analsys? What is BD? What is R?	LPAR chapters: 2 (visualizing data), 3 (visualizing with Lattice), 4 (cluster analysis), and 5 (agglomerative clustering)	Serial vs. parallel approaches. BDAR chapters: 3 (serial and parallel processing), and 4 (Hadoop)
Lunch	Lunch	Lunch
Iris dataset Titanic dataset	LPAR chapters: 9 (linear regression), 10 (classification with k-nearest neighbors, and naïve Bayes), 11 (classification trees) and 13 (text analysis).	BDAR chapters: 5 (SQL databases), and 6 (Non-SQL databases). Conclusion. Presentation of certificates.

References

[1] Eric Mayor, *Learning Predictive Analytics with R*, Packt Publishing Ltd, 2015.

[2] Simon Walkowiak, *Big Data Analytics with R*, Packt Publishing Ltd., 2016.