Course Introduction

Introduction to Data Science

Nina Zumel

John Mount

Lesson Goals

- Orient you regarding the plan and scope of the course: “Introduction to Data Science”

Why data science?

- As business analytic problems to more variables, more difficult modeling tasks and larger scales you move from traditional analytics to data science
- Data science is an interdisciplinary field taking methods from
    - statistics
    - machine learning
    - programming / computer science
    - data engineering
- Data science is a growth industry

The course

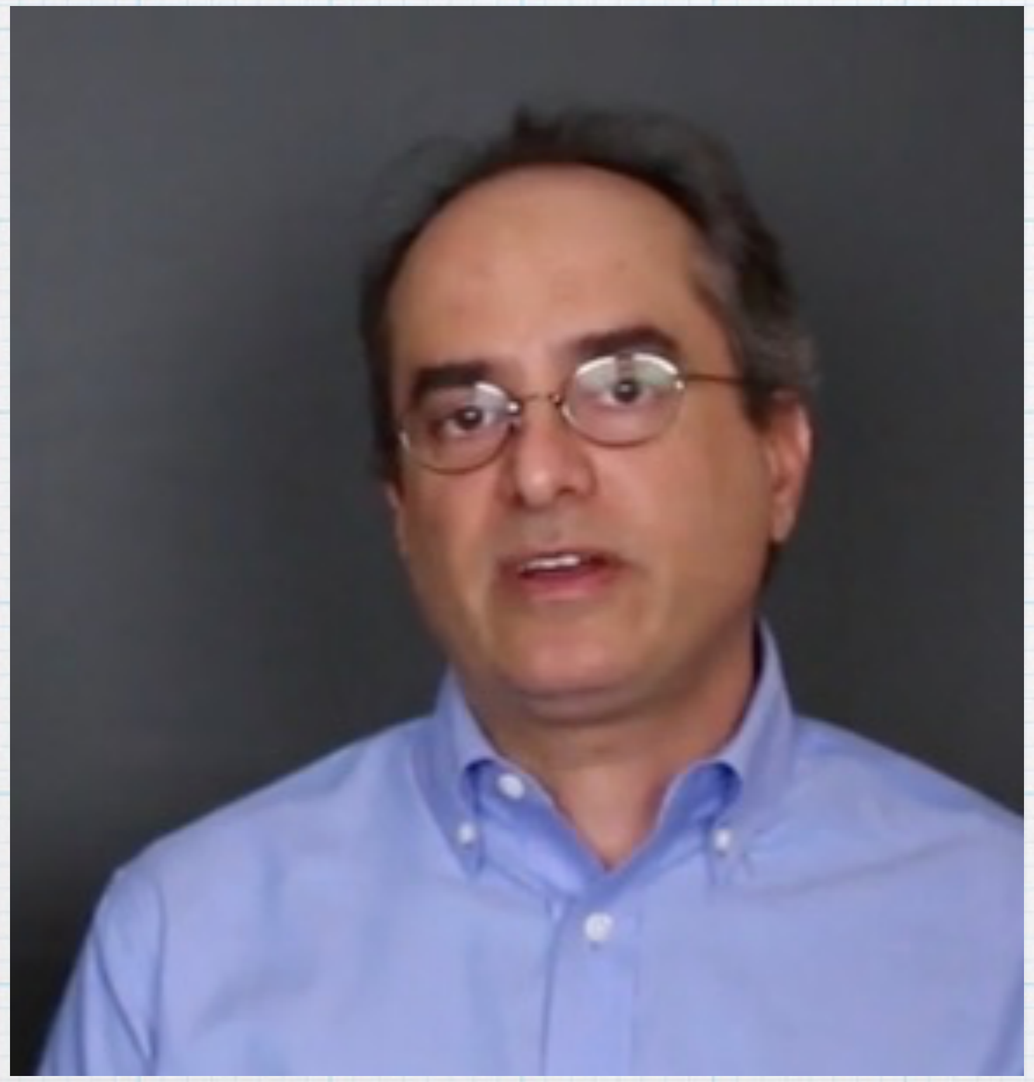
- This course will introduce you to the work of data science
    - It is an introduction to an advanced topic
    - We will concentrate on a portion of data science related to scoring and prediction
- We will work examples with actual data using an analysis system called "R"
    - Lectures will be slides and/or screencasts of the R system
    - We are supplying source code and data as free course extras

The instructors

Nina Zumel

- Over 10 years experience as consultant and trainer in data science
- A principal consultant at Win-Vector LLC (a data science consultancy)
- A primary author and content contributor to EMC's *Data Science and Big Data Analytics* training course and certification
  - Over 13,000 students world-wide have attended this course
  - Now available as a book from EMC Education Services and Wiley publications
- Author of *Practical Data Science with R* (Manning publications 2014)
  - A top rated and top selling guide to data science
- Ph.D. in robotics from Carnegie Mellon University
- Contributor to the Win-Vector blog

John Mount



- Work includes small molecule drug discovery, designing quantitative trading platforms (Bank of America, Banc division), running a research group (Shopping.com, an eBay company)
- Over 10 years experience as consultant and trainer in data science
- A principal consultant at Win-Vector LLC (a data science consultancy)
- Also an author of *Practical Data Science with R*
- Ph.D. in computer science from Carnegie Mellon University
- Contributor to the Win-Vector blog

Who is this course for?

- Analytically minded students who want to work through example data science projects and techniques
- Student requirements:
  - Some background with statistics
  - Familiarity with the R programming language
  - Both of these requirements can be picked up in parallel with the course with one or two additional courses or books

What is not in this course

- Data engineering (“big data”)
- How to implement your own machine learning algorithms
    - Except for one example we emphasize exploring and using already available machine learning libraries

What are the benefits of this course?

- We present the steps and techniques of a data science project in an organized manner
- We will work through standard ways to evaluate the quality of data science results, independent of the methodology
- The student will gain familiarity with the use of (and consequences of) a number of the most popular machine learning tools used in data science
- The supplied R code will demonstrate the steps required to actually get things done
- This should get you ready to work as a data scientist or work with data scientists (the best way to start!)

Support

- Code/data
  - http://winvector.github.io/IntroductionToDataScience/
  - Mostly in re-runnable R knitr markdown worksheets
- Contact us
  - Twitter https://twitter.com/winvectorllc @WinVectorLLC
- More in-depth articles
  - http://www.win-vector.com/blog/
- Professional consulting services
  - http://www.win-vector.com

What you should take away

- *Introduction to Data Science* is an introduction to the advanced topic of data science through worked examples in R