

OBA 455/555

Data-Driven Predictive Modeling (in R)

Background



- **Pradeep Pendem**, Asst. Prof, OBA, LCB, UO, 2018 – current
- M.S. Ph.D. (OM), Kenan-Flagler, UNC-Chapel Hill, 2013-18
- M.S. Statistics, Indian Statistical Institute, India
- B.E. Electrical Engineering, AU, India
- Industry Experience
 - Operations Research Analyst, Fidelity Investments, five years
- Research Interests: Data-driven Operational Analytics in Service Systems
 - Retail, People Analytics, Urban Mobility
- Achievements
 - Goulet Outstanding Junior Faculty Research Award, 2021
 - [Elwood S. Buffa](#) Best Ph.D. dissertation award, 2019
 - Recipient of first [Harvey M. Wagner](#) Scholar award at UNC for exceptional research, 2017
- Teaching
 - OBA 335 Operations Management
 - Interests: Data-Driven Predictive Modeling, Operations Management

Your turn!

- Name
- Brief background
- Something special about yourself
 - Hobby
 - Adventure you had in the past, or anything you want to share with us
- Any experience (e.g., internship) with business/predictive modeling
- Expectations from this course

Why Data ?

HBR.ORG

Harvard Business Review



OCTOBER 2012
REPRINT R1210D

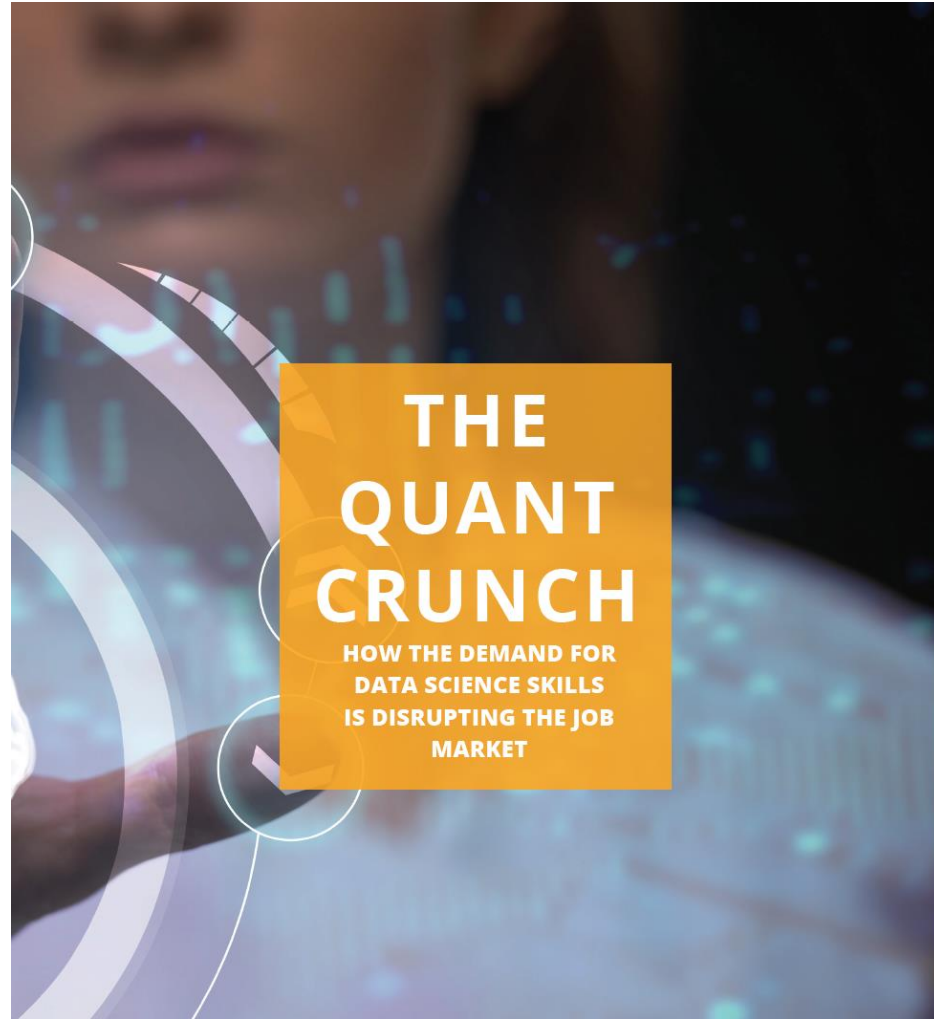
SPOTLIGHT ON BIG DATA

Data Scientist: The Sexiest Job Of the 21st Century

Meet the people who can coax treasure
out of messy, unstructured data.

by Thomas H. Davenport and D.J. Patil

Growing Demand for Data Professionals



- 2015 → 2020, ~ 2,350,000 → 2,720,000 (+15%)
- Jobs requiring machine learning skills are paying an average of \$114,000

Scarcity in Analytical Expertize

McKinsey&Company

McKinsey Global Institute



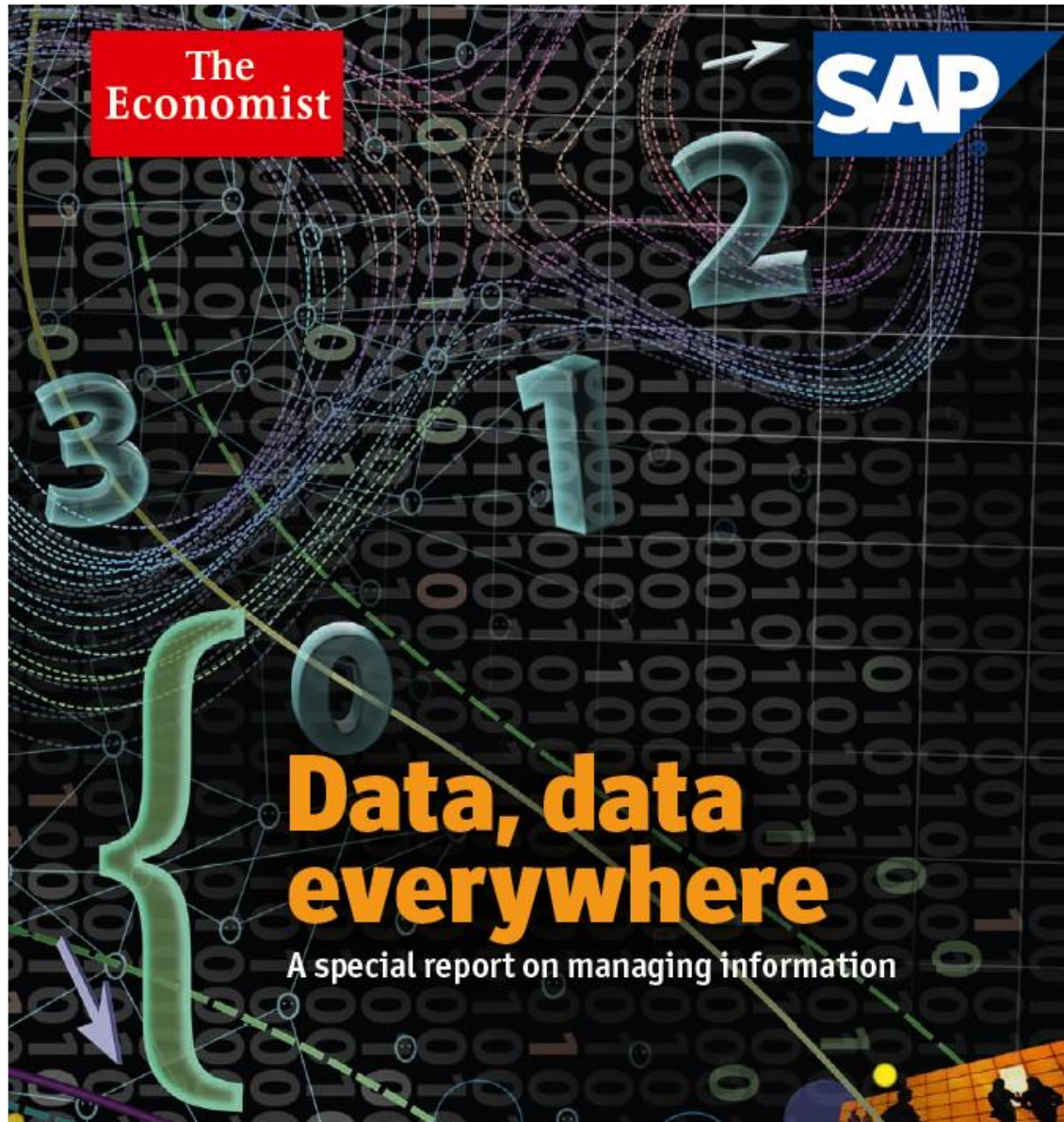
May 2011

Big data: The next frontier
for innovation, competition,
and productivity



“The United States alone faces a shortage of **140,000** to **190,000** people with analytical expertise and **1.5 million** managers and analysts with the skills to understand and make decisions based on the analysis of big data.”

Data Everywhere



Data Velocity 2019 → 2020

2019 *This Is What Happens In An Internet Minute*



2020 *This Is What Happens In An Internet Minute*



Data Velocity 2020 → 2021

2020 *This Is What Happens In An Internet Minute*

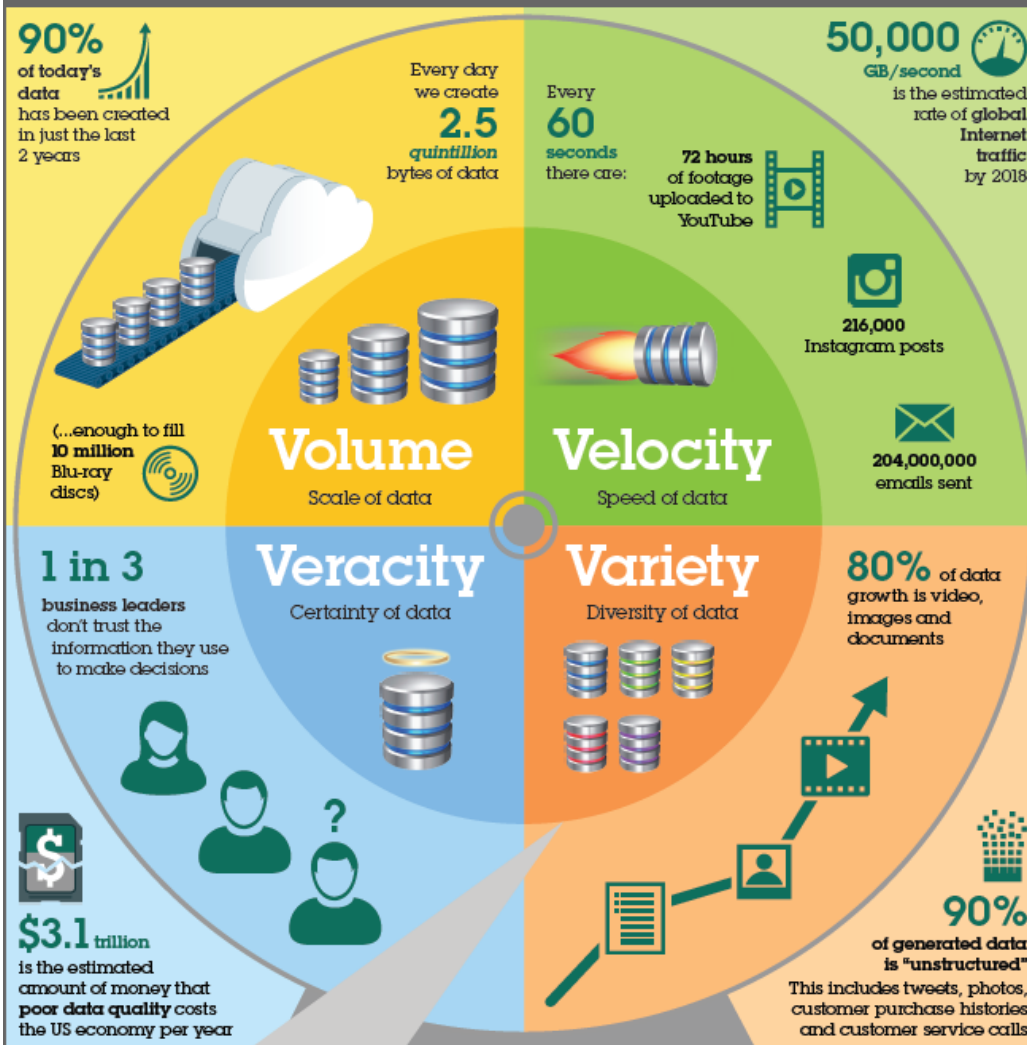


2021 *This Is What Happens In An Internet Minute*



4 V's of Big Data

Extracting business value from the 4 V's of big data



Big Data is the new gold

Fifth V?

The ability to mine the new gold and achieve greater **VALUE** through insights from superior analytics

Where are things heading?

- Smarter physical ecosystems
- IOT (Internet of Things)
- Sensors to connect homes, automobiles, roads, garbage bins.....
 - Smart refrigerator: “you are short on eggs!”
 - Populating your grocery store mobile app shopping list
 - Refrigerator negotiating a deal with Uber EATS driver to deliver a meal to you
 - Sensors in roads and vehicles to compute traffic congestion

**Future will be fueled by
Data Analytics**



AN MIT SMR EXECUTIVE GUIDE

Seven Technologies Remaking the World

- Pervasive Computing: Embedded, Networked Digital Processors
- Wireless Mesh Networks: High-Bandwidth, Dynamic, Wireless, Smart Connectivity
- Biotechnology: Technologically Created and Enhanced Life-Forms and Systems
- 3D Printing: Digitally Designed, Chemically Manufactured Objects
- **Machine Learning: Augmented, Automated Data Analysis**
- Nanotechnology: Engineered Atoms, Super-Materials
- Robotics: Precise, Agile, Intelligent Mechanical Systems

Business Analytics

- Practice and art of bringing quantitative data to bear on decision making and creating value

Value

Business
Fundamentals



Deployment

Right
Questions

Good
Analysis

Good Data

- Harness, Store, Process, Tools
- (Un)structured
-

- Descriptive Summary
- Visualization
- Methods
-

What will you learn in the course?

Value

**Business
Fundamentals**

**Right
Questions**

- Final project



Deployment

**Good
Analysis**

- Few predictive methods
- Little theoretical basics
- Model building process in R
- Inference

Good Data

- Basics of R
- Data management

Why take this course?

- Learn basic coding in R/RStudio
- Learn basic techniques in Predictive Modeling/ Machine Learning and their application in R

Why not take this course?

- Do not like coding/Data analytics
- Well versed in R/RStudio & tools of Predictive Modeling

Course Logistics

Groups

- Students form groups of **three** on canvas
- Group name: Chain of each member's first letter of the first name
 - E.g., **C**asey, **D**avid, **P**eter; Group name is **CDP**
- Groups remain **same** for all homework's, final project & presentation
- **Due date & time** in the syllabus file
- Students **not** formed groups by due date & time will be randomly grouped
- Groups cannot be changed after the due date
- Please choose your group members carefully

Materials

- Textbook
 - An Introduction to Statistical Learning with Applications in R
by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani
- Electronic version available for download at
 - <https://statlearning.com/>
- Textbook primarily for concepts beyond instructed in class



OBA 455/555

Student View

Spring 2022

Home

Modules

Assignments

People

Grades

Discussions

Attendance

Quizzes

Announcements

Rubrics

Files

Pages

Syllabus

Outcomes

BigBlueButton

Settings

OBA 455/555 (Spring 2022; 32758,32764) Predictive Modeling

Edit

Dear Students,

Welcome to the course - Data-Driven Predictive Modeling (in R).

This course introduces to basics of programming in R and the fundamentals of predictive modeling. The audience for this course is senior undergraduate and MBA students. Predictive modeling is a sub-field of business analytics. Utilizing historical data and applying various machine learning techniques, enables us to develop models that can be used in Regression, Classification, and identifying patterns. A few examples include predicting - new patient's length of stay in a hospital outpatient department, used cars sale price, whether a customer accepts a loan or not, whether a customer commits a credit card fraud or not, etc.

In this course, the main concepts of various techniques such as k-Nearest Neighbors (k-NN), Linear Regression, Logistic Regression, Regression-based Forecasting, Classification Trees, Regression Trees, and Cluster Analysis are discussed. Further, we also discuss the theory and implementation of error or accuracy measures, cross-validation, and model selection.

In this course, theoretical concepts of predictive modeling will be supplemented by applying them to real datasets. For this purpose, we will use R software. R is one of the most [popular](#) programming languages for statistics/data science; therefore, our objective is to teach the fundamentals of programming in this language.

Students who successfully complete this course will:

- Have fair exposure to introductory programming in R
- Learn the process of managing, cleaning, summarizing, and visualizing real-world datasets in R
- Be able to establish the thinking process of defining a real-world data-driven analytics project
- Learn theoretical fundamentals of different predictive modeling techniques
- Learn the implementation and inference of the techniques in R
- Learn how to evaluate various predictive models and select the model

Regards,

Pradeep Pendem

Course Status

Unpublish

Published

Import Existing Content

Import from Commons

Choose Home Page

View Course Stream

Course Setup Checklist

New Announcement

View Course Analytics

View Course Notifications

Coming Up

View Calendar

Nothing for the next week

Class Style

- Lectures
 - Theoretical introduction of a topic
 - Examples of real-world application contexts
- Software implementation
 - Implementation of model in R/RStudio
 - Inference of results
- Break
 - 5 - 10 minutes
- Homework problem set at the start/mid of the topic
 - Reinforcing of topics
 - Evaluating learning on your own

Class Access

- Scheduled In-person
 - No Zoom live streams and recordings will be available
- Attendance
 - Tracked in every class on canvas
 - No weight towards the grade

Class content & materials

- Materials available before the class start
 - Datasets
 - Incomplete code file(s)
- Materials available after the class
 - Class slides
 - Completed code file(s)
- All files to be accessed from the “**Modules**” section on canvas

Assessment

Type	Weight
Homework's (four)	20%
Midterm Quiz 1	20%
Midterm Quiz 2	20%
Project (Report + Presentation)	30 + 10%
	100%

- Grade assignment is based on **relative** performance
- Top $x\%$ will get A, the second top $y\%$ will get A- and so on
- The grading process is equivalent to **curving**

Homework (20%)

- **Four** homework's
- Not to seek help from individuals **outside** your group
- Homework's to be submitted in Canvas by their due date (and time)
- **Due date(s) & time** in the syllabus file
- Only **one** member of the group to submit on canvas
- List the group name on the top right of the submitted file
- Late/No submission results in a zero score for the group

Midterm (20% + 20%)

- Two midterms
- Multiple choice quiz on canvas
- **Quiz date(s)** in the syllabus file
- Open book
- Content
 - Conceptual knowledge
 - Identifying appropriateness of different techniques for different business problems/scenarios
 - Identifying strengths and shortcomings of the techniques
 - Interpret results of analyses
 - Code errors, output

Final Project (40%)

- Specify a business problem
- Identify a relevant dataset
- Business context could be in any area or function
- Assessment
 - Report (30%) + Presentation (10%)
- Presentation
 - 10–15-minute presentation on one of the classes in last week
 - **Presentation date(s) i**n the syllabus file

Final Report (30%)

- Formal report
 - Introduction, Problem description, Approach
 - Data Analysis, Results, Inference
 - Conclusions, recommendations
- 8-10 pages including any tables and graphs (excluding code)
- Submit the code with comments at end of the report
- **Due date & time** in the syllabus file
- Late submissions results in a zero score for the group

Public datasets for final project



- <https://www.kaggle.com/>
- Online community of data scientists and machine learners
- Owned by Google Inc.
- Register yourself, and you can download datasets for free
- As of June 2017, Kaggle passed over 1,000,000 registered users
- Variety of datasets
- Your imagination only limits possibilities

Office Hours

- Office Hours

- Tuesday & Thursday 1 PM – 2 PM, Lillis 432
- Schedule a Zoom meeting for an alternate time if you prefer remote support
- Zoom link path: Canvas → Modules → General → Zoom Link

- Instructor

- pradeepp@uoregon.edu
- 541-346-3348

Communication Rules

- Use canvas or your **uoregon.edu** email for communication
- Unlikely to receive a response to emails sent from personal email IDs
- Lecture accent
 - English is neither my mother tongue nor I was born/grew in an English-speaking country
 - Stop and ask questions if you don't understand what I convey

Predictive Models

Predictive Models

Supervised

Unsupervised

Regression

Classification

Time Series Forecasting

Segmentation

- ***k*-Nearest Neighbor**
- **Linear Regression**
- **Regression Trees**
- Neural Networks
- Ensembles
-

- ***k*-Nearest Neighbor**
- **Logistic Regression**
- **Classification Trees**
- Naïve Bayes
- Neural Networks
- Discriminant Analysis
- Ensembles
-

- **Regression-based**
- Smoothing methods
-

- **Clustering**
-

Famous examples/Real cases

What's Even Creepier
Than Target Guessing That
You're Pregnant?



- Pregnant prediction based on her prior purchases
 - Market to be mothers with coupons on baby related products
 - News about a surprised customer
- Moneyball : Predict players likely to contribute to winning team
 - Oakland Athletics assembled a competitive team with less-than stellar budget
- Obama For America (OFA)
 - Predict who is likely to vote to Obama
 - Personalized campaigning to voters
- OKCupid/Tinder
 - Predict what forms of message content are most likely to produce a response
 - Suggest a prospect with content on introduction
 - Suggest matches based on your historical left/rights swipes
- Netflix recommender system
 - Customer demographics, watching hours, videos

Today's tasks

- Get familiarity with course page in canvas
 - Layout
 - Syllabus, Content, and resources
- Mark all crucial dates (& times) in your calendar
 - Group formation on canvas
 - Midterms
 - Homework's
 - Presentation
 - Project Report
- Register on <https://www.kaggle.com/>
 - Glance at datasets

Next Class

- Process to set up R & RStudio Software
- Software interface and demonstration
- Brief introduction to programming
- Sources of software online help and documentation

Thank You