# Programming for Data Science (CSE3041)

Ramesh Ragala

VIT Chennai Campus

July 15, 2020

# Syllabus Introduction

- **Course Objectives**
  - To provide necessary knowledge on how to manipulate data objects using python and R
  - To provide knowledge on how to analyze the data graphically.
  - Emphasize on different statistical methods and ways to analyze data using python and R
  - Provide solid understanding of Scala programming
- **Course Outcomes:**
  - Students are able to solving analytical problems with the help of Python and R programming languages with appropriate libraries
  - Import, export, visualize and manipulate the continuous and categorical data effectively using Python and R
  - Solves the problems using Scala functional programming language

# Syllabus Introduction

- **Concepts in Python**
  - Expressions, Operators, and Matrices
  - Decision Statements and Control flow
  - Functions, Classes, and Objects
  - Packages and Files
  - Strings, List, Tuple, Dictionaries and Comprehensions
  - Introduction to numpy library with operations
  - Linear Algebra with numpy
  - Computation of Eigenvalues and Eigen Vector using numpy
  - Introduction and basic functionality of SciPy
  - Introduction to Pandas, series object and data frame
  - Pandas Objects: Data Aggregation and Joining
  - Pandas Object: Concatenating and appending data frames and index objects
  - Data Wrangling With Pandas
  - Handling Time series data using pandas
  - Handling missing values using pandas

# Syllabus Introduction

- **Concepts in Python**
  - Reading and writing the data including JSON data
  - Web scraping using python
  - Combining and merging datasets
  - Data transformations
  - Common plots for statistical analysis using matplotlib, seaborn, etc.
  - common plots for statistical analysis using ggplot, ggvis, etc in python
  - common plots for statistical analysis using Plotly, Altair etc in python

# Syllabus Introduction

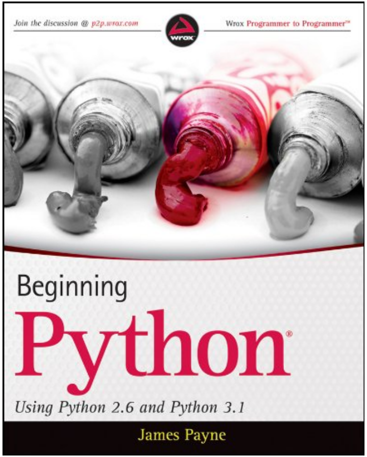
- **Concepts in R programming**
  - Data types, Sequence generation, Vector, Random number generation and Data frames.
  - Functions, Data manipulation and Data Reshaping using plyr, dplyr and reshape2
  - Parametric statistics and Non-parametric statistics,
  - Continuous and Discrete Probability distribution using R,
  - Correlation and covariance, contingency tables.
  - Overview of Sampling, different sampling techniques
  - R and data base connectivity
  - Web application development with R using Shiny and Approaches to dealing with missing data in R
  - Exploratory data analysis with simple visualizations using R
  - Feature or Attribute selection using R
  - Dimensionality Reduction with R
  - Time series data analysis with R

# Syllabus Introduction

- **Concepts in Scala programming**
  - Variables, types, Literals and Operators
  - Classes and objects
  - Functional objects: choosing between val and var, class parameters, constructors, self references and method overloading
  - Conditional and loop statements
  - Functions in Scala
  - Control abstraction in Scala
  - Composition and Inheritance
  - Traits and Mixins
  - File IO in Scala
  - Case Classes and Pattern Matching
  - Packages and imports in Scala
  - Working with Lists and Collections in Scala
  - Working with XML, Implementing List
  - Extractors and objects as modules

# Text Books

- James Payne, ”Beginning Python: Using Python 2.6 and Python 3.1”, Wrox, Ist Edition, 2010



Join the discussion @ p2p.wrox.com

wrox

Wrox Programmer to Programmer™

Beginning

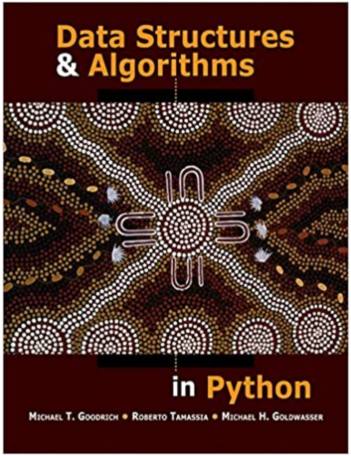
Python®

*Using Python 2.6 and Python 3.1*

James Payne

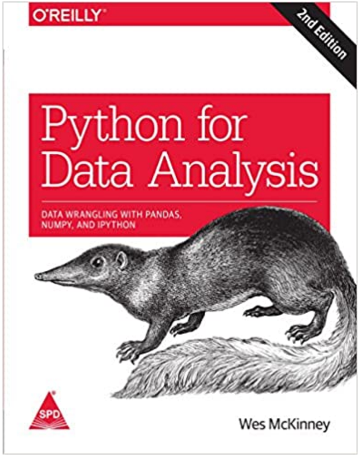
# Text Books

- Michael T. Goodrich, Roberto Tamassia, Michael H. Goldwasser, ”Data Structures and Algorithms in Python”, John Wiley and sons, 2013.


Data Structures
& Algorithms
in Python
Michael T. Goodrich • Roberto Tamassia • Michael H. Goldwasser

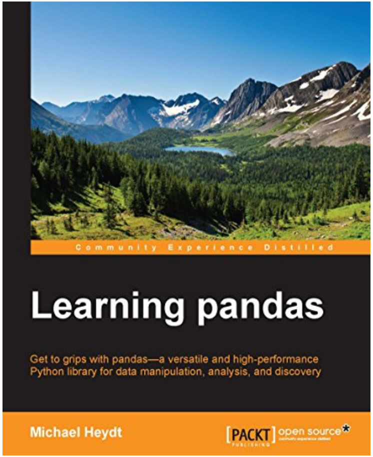
# Text Books

- William McKinney, ”Python for Data Analysis: Data Wrangling with Pandas, NumPy, and Ipython”, O’Reilly Media, IInd Edition, 2017



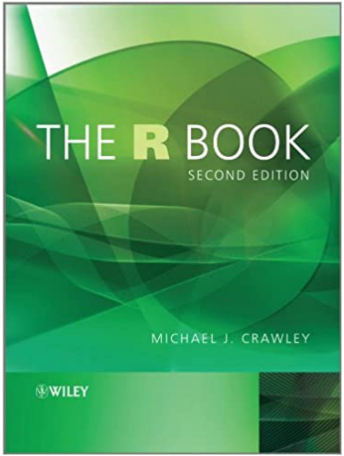
# Text Books

- Michael Heydt, ”Learning Pandas - Python Data Discovery and Analysis Made Easy”, Packt Publishing Limited , 2015.


Community Experience Distilled
Learning pandas
Get to grips with pandas—a versatile and high-performance Python library for data manipulation, analysis, and discovery
Michael Heydt
[PACKT] PUBLISHING
open source

# Text Books

- Michael J. Crawley, ”The R Book”, Wiley, 2nd Edition, 2012.



# Text Books

- Robert Kabacoff, ”R in Action”, Manning Publication, I$^{st}$ Edition, 2011.

Copyrighted Material

R IN ACTION

Data analysis and graphics with R

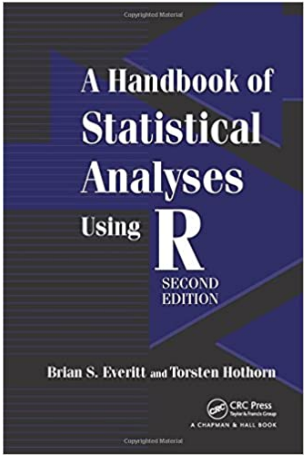
Robert I. Kabacoff

MANNING

Copyrighted Material

# Text Books

- Torsten Hothorn, Brian S. Everitt, "A Handbook of Statistical Analyses Using R", Chapman and Hall, CRC, 2nd Edition, 2009.



Copyrighted Material

A Handbook of Statistical Analyses Using R

SECOND EDITION

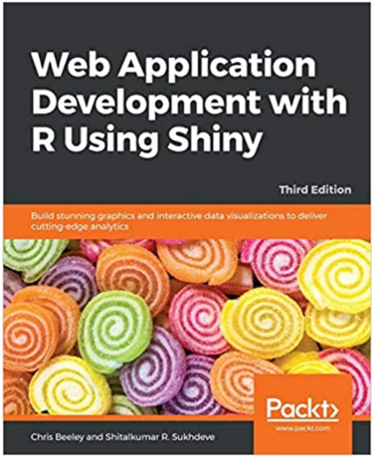
Brian S. Everitt and Torsten Hothorn

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Copyrighted Material

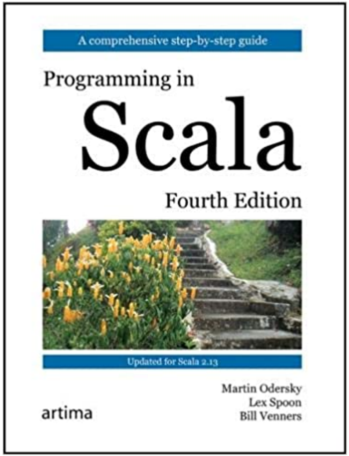
# Text Books

- Chris Beeley "Web Application Development with R Using Shiny", Third Edition, Pact Publishing, 2013.


Web Application Development with R Using Shiny
Third Edition
Build stunning graphics and interactive data visualizations to deliver cutting-edge analytics
Packt>
www.packt.com
Chris Beeley and Shitalkumar R. Sukhdeve

# Text Books

- Martin Odersky, Lex Spoon, and Bill Venners, ” Programming in Scala” , Fourth Edition



A comprehensive step-by-step guide

Programming in

Scala

Fourth Edition

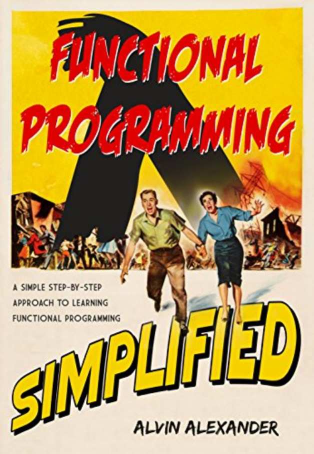
Updated for Scala 2.13

Martin Odersky
Lex Spoon
Bill Venners

artima

# Text Books

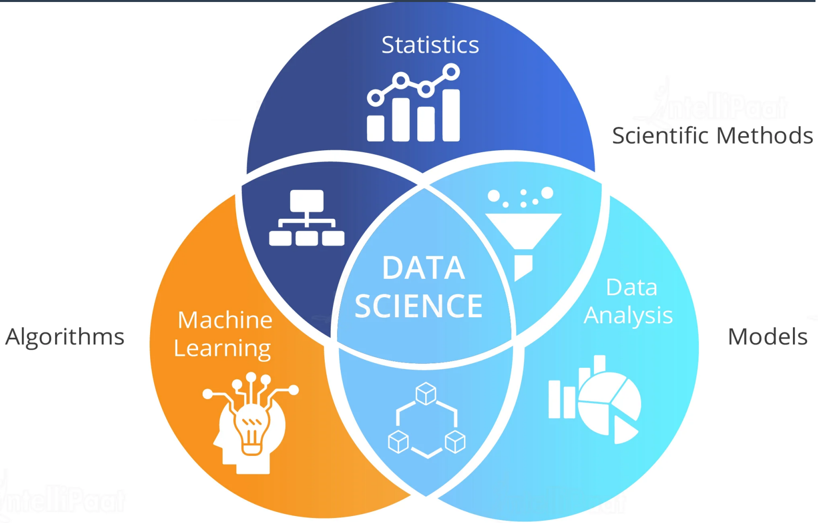
- Alvin J. Alexander " Learning Functional Programming in Scala", 2017


FUNCTIONAL PROGRAMMING
A SIMPLE STEP-BY-STEP APPROACH TO LEARNING FUNCTIONAL PROGRAMMING
SIMPLIFIED
ALVIN ALEXANDER

# Introduction to Data Science

- Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data
- Data science is a ”concept to unify statistics, data analysis, machine learning, domain knowledge and their related methods” in order to ”understand and analyze actual phenomena” with data – Wikipedia
- It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, domain knowledge and information science – Wikipedia
- Simply, the goal of the Data Science is to extract knowledge from large data sets – Wikipedia
- It uses many steps such as analysis, prepossessing the data, and gives the description of the findings or inferences during the process.
- In this process, it uses the skill from various domains such as Mathematics, Statistics, Visualization, Domain knowledge, Data Mining, Machine Learning, Computer Vision, etc.

# Introduction to Data Science



Pictorial representation of Data Science

# Case Study-1 in Data Science

- The Nisqually River Foundation is tasked with the successful implementation of a watershed stewardship plan.
- As a part of this plan, they assist the Nisqually Indian Tribe in Washington State to measure and monitor the fish species present in the Nisqually River.
- To do this, the Nisqually Indian Tribe installed a video camera and infrared sensors in a fish ladder at a dam on the river.
- The camera is triggered to capture 30 seconds of video when any fish swims past the infrared sensors.
- It is complex manual process
  - Throughout the year, more than 3,000 videos are generated by the counter camera.
  - As part of their original process, a trained biologist needed to view each video to manually identify and record the species of each fish.
  - This manual process of fish species identification in captured videos is resource intensive, from a time, human resources, and cost perspective.
  - "This work is slow and repetitive and is much better suited to automation than manual analysis."

1

[1] https://partner.microsoft.com/en-us/case-studies/gramener

# Case Study-1 in Data Science

- Gramener, in conjunction with the Microsoft AI for Earth program, worked with the Nisqually River Foundation to attempt to automate the detection and identification of fish species from the video clips.
- The Nisqually salmon detection application was built as a web app to automate the process of video feed input, detection, and classification.
- The automated AI solution leverages the latest deep learning algorithms implemented using the Microsoft Azure and Cognitive Services platform stack.
- The first challenge was to process the videos and tag the fish.
- The heavy manual work involved in this was automated by leveraging the Microsoft VOTT tool.
- The tagged frames were then used to train a model using Microsoft Cognitive Toolkit (CNTK).
- This model was then tested against more frames extracted from the videos.
- While this solution was good, it lacked speed and real-time video detection capabilities.
- As an enhancement to the solution, Gramener moved to video object detection using YOLO V3, which provides a faster solution with real-time capabilities.

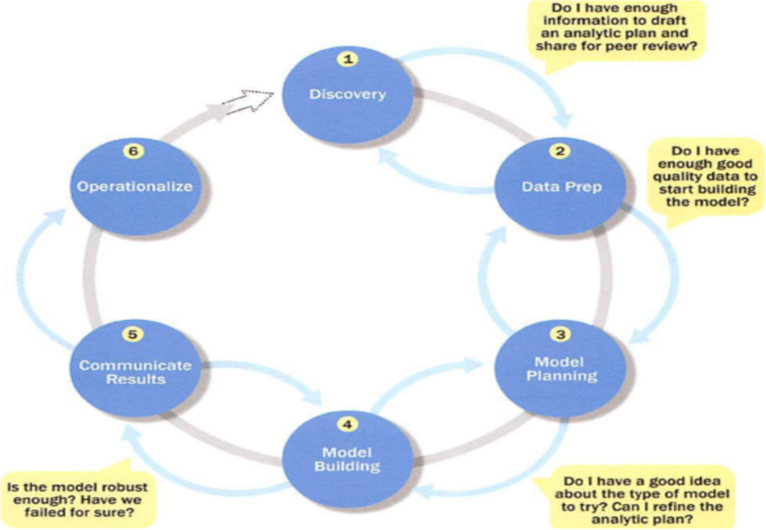
# Case Studies of Data Science

- https://www.datacamp.com/projects/870
- https://www.kaggle.com/general/7615
- https://www.svds.com/case-studies/

# Data Science Life Cycle

- Data Analytic Life cycle defines the analytics process and best practices from discovery to project completion.
- Data Analytics Lifecycle Phases
  - Discovery Phase
  - Data Preparation Phase
  - Model Planning Phase
  - Model Building Phase
  - Communicate Result
  - Operationalize
- With six phases the project work can occur in several phases simultaneously
- The cycle is iterative to portray a real project
- Work can return to earlier phases as new information is uncovered.

## Data Analytics Life Cycle


Do I have enough information to draft an analytic plan and share for peer review?
1
Discovery
2
Data Prep
Do I have enough good quality data to start building the model?
3
Model Planning
Do I have a good idea about the type of model to try? Can I refine the analytic plan?
4
Model Building
Is the model robust enough? Have we failed for sure?
5
Communicate Results
6
Operationalize

# Data Analytics Life Cycle

- Phase – I: Discovery
  - Learning the Business Domain
  - Resources
  - Framing the Problem
  - Developing Initial Hypotheses
  - Identifying Potential Data Sources

# Data Analytics Life Cycle

- Phase – 2: Data Preparation
  - It requires analytical sandbox in which you can perform analytics for the entire duration of the project
  - Includes steps:
    - Explore
    - Preprocess
    - Conditional Data
  - Data preparation tends to be the most labor-intensive step in the analytics lifecycle →Often at least 50% of the data science project's time.
  - The data preparation phase is a iterative process.
  - In ETL users perform extract, transform, load
  - Data Analytics lifecycle →ELT or ETLT →Extract, Transform, Load and Transform.

# Data Analytics Life Cycle

- Phase – 3: Model Planning
  - This determines the methods and techniques to extract relationships among variables.
  - These relationship patterns will set the base for algorithms which will be used in next phase.
  - It uses Exploratory Data Analysis (EDA) using various statistical formulae and visualization tools.
  - Simply, it identifies candidate models to apply to the data for clustering, classifying, or finding relationships in data.
- Activities to be consider in this phase are:
  - Assess the structure of the data.
  - Ensure the analytic techniques enable the team to meet the business objectives and accept or reject the working hypotheses.
  - Determine if the situation warrants a single model or a series of techniques as part of a larger analytic workflow.
  - Research and understand how other analysts have approached this kind or similar kind of problem.

# Data Analytics Life Cycle

- Phase - 4: Model Building
  - Execute the models defined in Phase - 3.
  - Develop datasets for training, testing, and production.
  - Develop analytic model on training data, test on test data.
  - It will consider whether your existing tools will suffice for running the models or it will need a more robust environment (like fast and parallel processing).
  - You will analyze various learning techniques like classification, association and clustering to build the model.

# Data Analytics Life Cycle

- Phase – 5: Communicate Results
  - Determine if the team succeeded or failed in its objectives.
  - Assess if the results are statistically significant and valid. → If so, identify aspects of the results that present salient findings. → Identify surprising results and those in line with the hypotheses.
  - Communicate and document the key findings and major insights derived from the analysis.
  - This is the most visible portion of the process to the outside stakeholders and sponsors.

# Data Analytic Life Cycle

- Phase – 6: Operationalize
  - In this last phase, the team communicates the benefits of the project more broadly and sets up a pilot project to deploy the work in a controlled way.
  - Risk is managed effectively by undertaking small scope, pilot deployment before a wide-scale rollout.
  - During the pilot project, the team may need to execute the algorithm more efficiently in the database.

- Contact Me: ramesh.ragala@vit.ac.in
- Ping Me : +91-9087277270
- Course Webpage: https://github.com/rameshragala/

Thank you