Becoming a Data Scientist

Dr. Robert P. Schumaker
Associate Professor of Computer Science
Director – Data Analytics Lab

The University of Texas at TYLER

Data Science

- What is Data Science?

  – Solve large, complex, often unstructured or ill-defined problems using a data-driven approach

  – Predicting customer needs and anticipating their orders

Data Science

- What is Data Science
  - Knowledge of Statistics
  - Clean and Analyze Data
  - Able to Data Mine
  - Translate Business Requirements
  - System Builder

Data Science

- What is Data Science (Jobs)
  - Knowledge of Statistics – Statistician
  - Clean and Analyze Data – Business Analyst
  - Able to Data Mine – Entry-level Data Scientist
  - Translate Business Requirements – Data Scientist
  - System Builder – Senior Data Scientist

Data Science

- Knowledge of Statistics
  - Where to learn: Statistics courses ← PLURAL!
  - Occupation: Statistician
  - Skills:
    - Collect and use data to explain phenomena
    - Descriptive statistics
      - Basic statistics
      - Draw and compare samples
      - Analyze and interpret data

Data Science

- Knowledge of Statistics
  - Tools Statisticians Use
    - Excel
    - R
    - SAS
    - SPSS
  - Salary
    - Range: $46k - $95k
    - Median: $68k
    - Source [1]

1 - https://www.payscale.com/research/US/Job=Statistician/Salary/792ea001/Entry-Level

Data Science

- Clean and Analyze Data
  - Where to learn: Database and System Design courses
  - Occupation: Business Analyst
  - Skills:
    - Database design
    - Data modeling and analysis
    - Data extraction for reporting

Data Science

- Clean and Analyze Data
    - Tools Business Analysts Use
        - Excel
        - SQL
        - UML
        - Visio
    - Salary
        - Range: $49k - $94k
        - Median: $67k
        - Source [2]

Data Science

- Able to Data Mine
  - Where to learn: Data Mining courses
  - Occupation: Entry-level Data Scientist
  - Skills:
    - Data extraction and reporting
    - Modeling prototypes
    - Predictive modeling and statistics

Data Science

- Able to Data Mine
  - Tools Entry-level Data Scientists Use
    - Excel
    - R
    - SQL
    - Tableau
  - Salary
    - Range: $61k - $141k
    - Median: $92k
    - Source [3]

3 - https://www.payscale.com/research/US/Job=Data_Scientist%2c_IT/Salary

Data Science

- Translate Business Requirements
  - Where to learn: Computer Information Systems degree
  - Occupation: Data Scientist
  - Skills:
    - Build models and prototypes
    - Proficient programmers and machine learners
    - Familiarity with Hadoop, Spark and NoSQL

Data Science

- Translate Business Requirements
    - Tools Data Scientists Use
        - Python
        - R
        - SQL
        - Tableau
    - Salary
        - Range: $94k - $165k
        - Median: $118k
        - Source [4]

4 - https://www.payscale.com/research/US/Job=Lead_Data_Scientist/Salary

Data Science

- System Builder
  - Where to learn: Advanced degrees and years of experience
  - Occupation: Senior Data Scientist
  - Skills:
    - Define solutions to vague and unclear business problems
    - Develop first of their kind complex systems
    - Leverage big data
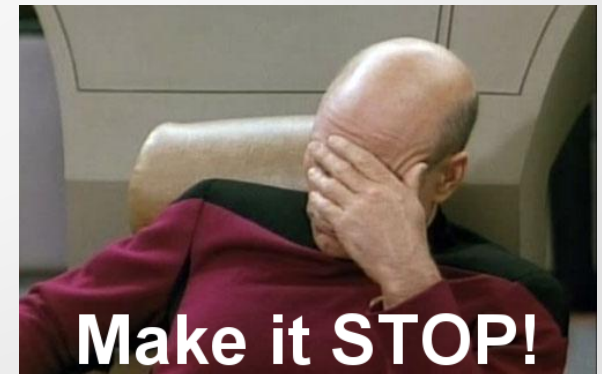   - Outstanding communications and reporting

Data Science

- System Builder
  - Tools Senior Data Scientists Use
    - Python
    - R
    - SQL
    - Tableau
  - Salary
    - Range: $94k - $260k
    - Median: $165k
    - Source [5]

5 - https://www.payscale.com/research/US/Job=Senior_Data_Scientist%2c_IT/Salary/ab56c444/Late-Career

Data Science

- ## System Builder
  - A Proficient System Builder knows the fields of:
    - Business
    - Computational Linguistics
    - Computer Science
    - Information Systems
    - Information Technology
    - Math
    - Software Engineering
    - Statistics

Make it STOP!

Data Science

- ## System Builder
  - Plus the specific problem domain…
    - Airline logistics
    - Computer Security
    - Finance
    - Healthcare
    - Medicine
    - Politics
    - Psychology
    - Sports
    - Telecommunications

Looking at your friend during the exam like

When did we learn this

Data Science

- The Data Science Approach
  - Identify a general area to study
  - Acquisition
    - Determine what data resources are available
    - Collect data directly or through API
    - Clean the relevant data
  - Representation
    - Ascertain the independent and dependent variables
    - Shape the research questions
    - Consider ways to model the data

Data Science

- The Data Science Approach
  - Description / Prediction
    - Determine the relevant statistics, machine learning algorithms
    - Train, Test and Validate
    - Refine research questions, seek additional data sources, iterate
  - Delivery
    - Prepare a technical research report to communicate results
    - Present the findings
    - Create a software prototype

Data Science

- Data Science Resources
  - Blogs
    - BigDataMadeSimple (www.bigdata-madesimple.com)
    - FiveThirtyEight (www.fivethirtyeight.com)
    - OKCupid (blog.okcupid.com)
  - Datasets
    - Data.gov (www.data.gov)
    - Kaggle (www.kaggle.com)
    - UCI Machine Learning (archive.ics.uci.edu/ml/datasets.html)

Data Science

- Data Science Resources
  - Meetups
    - https://www.meetup.com/topics/data-science/us/tx/dallas/
  - Talks
    - Data Science for Beginners
      - https://www.youtube.com/watch?v=0XyV91VYrDs
    - TEDxIEMadrid – The Most Important Skills of Data Scientists
      - https://www.youtube.com/watch?v=qrhRfPY4F4w
    - Gigs: A Day in the Life of a Data Scientist
      - https://www.youtube.com/watch?v=EaptTxhh6sM