Learning and Teaching Data Science

BINA RAMAMURTHY
bina@buffalo.edu
http://www.cse.buffalo.edu/~bina
This research is partially funded by NSF-DUE-CCLI grant 0920335
bina@buffalo.edu
Goals

- To share my experience with teaching data science: what worked and what did not
- To discuss approaches for teaching data science
- To share curriculum, course material and tools to introduce data science into undergraduate curriculum
- Target audience: teachers and administrators of 2-year and 4-year CSE programs
Overview

- What is DS?
- Entry points for adoption of DS
- Entry points into DS
- How did we do it?
- Courses: Distributed Systems & Data-intensive computing course
- Certificate program
- What worked? Learning outcomes
- What needs work?
- Best practices
- Summary
My Journey Learning DS

- Yahoo! Big Data Computing conference: NSF, CRA supported
- Hadoop cluster in 6 months: presented at CCSCNE April 2009
- NSF grant for creating a data-intensive computing certificate program and courses
- Certificate program approved by SUNY system and is listed in the catalog (2011-12)
- Courses have become permanent and have been assimilated into the degree program
What is DS?

Data Science is not a single subject: it is a combination of many topics and skills [1]. According to this, DS comprises:

- Computer science,
- Mathematics,
- Probability & Statistics,
- Machine learning,
- Visualization,
- Communication and
- Domain expertise

To that I have added “Coding” as in problem solving and programming.
Data Science Process

EDA: Exploratory Data Analysis [J. Tukey]
Entry Points for DS

- **Level I. Introduce one DS topic in one of your courses (1 week)**
  - Example 1: Beauty of classical MapReduce-version page rank algorithm
  - Example 2: Hadoop distributed file system (HDFS) in an operating systems course
  - Example 3: Visualization in a quantitative analytics course

- **Level II. Teach DS as a special topics course (full semester/ 6-week summer)**
  - Example 1: Try it out with a combination of DS topics
  - Example 2: Statistical models and machine learning using R

- **Level III. Approved courses in a program (2 or more courses)**
  - Example: We created a Distributed Systems course and a Data-intensive computing
  - These started as electives and have become courses satisfying “core” requirements in the program

- **Level IV. A full certificate program in alignment with the courses in an existing program**
  - combination of DS courses + courses in the program + capstone

- **Level V. Master of Science/Ph.D degree in Data science or closely related area.**
We are here
Our DS Program

• We began with **Level II** and with the help of a NSF CCLI A&I grant created a “grid-computing” based distributed systems course.
• Then implemented **Level III** and **Level IV**
• **Senior level/ entry graduate level, cross-listed**
• **Prerequisite:** Data structures and algorithms (CS2/3)
• **Courses:**
  • CSE486 Distributed Systems
  • CSE487 Data-intensive Computing
We began with **Level II** and with the help of a NSF CCLI A&I grant created a “grid-computing” based distributed systems course.

Then implemented **Level III** and **Level IV**

**Then:** We (Buffalo CSE) did **NOT** have a distributed systems courses before 2004!

**Now:** Taught every semester (high demand) by new tenured faculty hired in this area.

Part I: EDA

Part II: Parallelizing Data Processing
- Hadoop
- MapReduce
- Ecosystem

Part III: Optimized & Integrated Data Processing
- Spark Ecosystem

Statistical Inference

Big data

Performance
Data-intensive Computing Course

• Divided into three major topics: approximately 1 month duration

I. Exploratory data analysis
   I. Statistical inference/modeling
   II. Machine learning algorithms
   III. R language
   IV. Data analysis using RStudio
   V. Where to find information: Chapter 1-5 of DS text [1]
   VI. Suggestion: If you prefer you can continue with the rest of the chapters in the text for a full course
I. Parallelizing data processing
   I. Working with large data sets
   II. Hadoop eco-system
   III. MapReduce like algorithms
   V. Apache Hadoop [3]
   VI. Amazon AWS or Virtual box or VMware

II. Optimizing and integrating data operations
   I. RDD (Resilient Distributed Data)
   II. Optimization of operations in a DAG (directed acyclic graphs)
   III. Apache Spark [4]
   IV. Emerging applications
Part I: Exploratory Data Analysis

- Formally introduced by John Tukey [5]
- He is also the one recognized “software”, “bit”, robust analysis
- EDA example: stem-leaf-display (Tukey’s), measures of central tendency (mean, median, etc.)
- Literally explore data with various assumptions, try out different graphing and discover data behaviors.
- Look at the data first and let it drive the theories (rather writing theorems/algorithms and proving them).
- We studied: statistical models, linear regression and K-means and K-NN and their application to real data problems
- Tool Used : R and Rstudio : Project 1
  (Can use MS Excel, Tableau, SPSS, MATlab...)
EDA in Data Science Process [1]

1. Raw data collected
2. Data is processed
3. Data is cleaned
4. Exploratory data analysis
   - Machine learning algorithms;
   - Statistical models
5. Communication Visualization Report Findings
6. Make decisions
7. Build data products
8. Data is processed
9. Data is cleaned
10. Exploratory data analysis

Data is processed
Data is cleaned
Exploratory data analysis
Part II: Parallelizing Data for Processing

- Grace Hopper on tackling large problems [6]:
  - "In pioneer days they used oxen for heavy pulling, and when one ox couldn't budge a log, they didn't try to grow a larger ox. We shouldn't be trying for bigger computers, but for more systems of computers."
- Large storage for data: WORM data
- Large data 2K blocks vs 128Mbyte
- MTBF is 1/1000, with 1000 blocks of data, the probability of a block failure at a given time is very high (nearly 100%)
- Duplicate or triplicate data blocks
- What to do with this redundancy? Run parallel programs
- MR like algorithms
- Hadoop eco system
- Alternatively one can introduce MPI, OpenMP, kernel threads and other approaches to parallel computation
Part III: Optimizing Data Computing

- Refined coding: less code
- Faster computation
  - Keep it in-memory
  - Build on related data
- Optimized execution engine
- Different tools implemented for Hadoop are unified into one on Spark
- Data: Resilient Distributed Data Sets (RDDs)
- Operations: transformations, actions
- Execution model: Single Instruction and Multiple Data
### Data-intensive computing Pedagogy

<table>
<thead>
<tr>
<th>Part</th>
<th>Lecture: Book/Concepts</th>
<th>Hands-on Project</th>
<th>Resources</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Data Science[1]</td>
<td>R &amp; real data</td>
<td>RStudio, Shiny</td>
<td>Numerical, categorical; Stock market</td>
</tr>
<tr>
<td>II</td>
<td>Data Computing &amp; Performance issues</td>
<td>Apache Spark</td>
<td>Cloud resources</td>
<td>Graph data; network analysis; people network</td>
</tr>
</tbody>
</table>

Cloud resources are helpful if you are infrastructure is not adequate or IT people/services are not good enough to support these courses.
• It is more like choosing a track of specialization in a student’s plan of study
• Convenient path for students working on their CSE minors
Certificate Program Courses

- Course 1 (CS2)
- Course 2 (CSE486, revised)
- Course 3 (CSE487, major area)
- Course 4 (Capstone, data-intensive)
- Course 5 (Transformative application)

Prerequisite for 4XX: CS2 or equivalent and junior standing.
What worked?

- Student learned the subject and got jobs
- Collaborations: interdisciplinary: Biology and Math; extensive, sustained collaborations
- External industry collaborations
- All these resulted in grants awarded and collaborative research that is still ongoing
- Newer interactions with Arts and Humanities
- Undergraduate student researcher won the SIGCSE Microsoft 2015 undergraduate research award and the ACM undergraduate research grand finals. [7]
Assessment: Coding

Coding (Pre)

Coding (Post)
Visualization

**Data Visualization (Pre)**

- Excellent: 1.4%
- Very Good: 20.3%
- Good: 29.0%
- Somewhat Good: 27.5%
- Fair: 15.9%
- Poor: 1.4%
- Unanswered: 4.3%

**Data Visualization (Post)**

- Excellent: 15.48%
- Very Good: 27.38%
- Good: 35.71%
- Somewhat Good: 9.52%
- Fair: 5.95%
- Poor: 1.19%
- Unanswered: 4.76%
Mathematics

Mathematics (Pre)

Mathematics (Post)
Computer Science

Comparison of Computer Science scores before (Pre) and after (Post) training.

- **Pre**
  - Excellent: 17.4%
  - Very Good: 31.9%
  - Good: 40.6%
  - Somewhat Good: 5.8%
  - Fair: 2.9%
  - Poor: 0.0%
  - Unanswered: 1.4%

- **Post**
  - Excellent: 27.38%
  - Very Good: 39.29%
  - Good: 25%
  - Somewhat Good: 1.19%
  - Fair: 0%
  - Poor: 0%
  - Unanswered: 7.14%
Communication

Communication (Pre)

- Excellent: 17.4%
- Very Good: 46.4%
- Good: 20.3%
- Somewhat Good: 11.6%
- Fair: 0.0%
- Poor: 2.9%
- Unanswered: 1.4%

Communication (Post)

- Excellent: 25%
- Very Good: 41.67%
- Good: 23.81%
- Somewhat Good: 2.38%
- Fair: 0.0%
- Poor: 1.19%
- Unanswered: 5.95%
Domain Expertise
What did not work?

- Enrollment in the certificate program is just picking up after 5 years. Solution: Publicity and planning
- Participation of working adults in the certificate program was not satisfactory
- Infrastructure needed changing with every offering of the course
- Text book: Not a single text book that covers all the topics
Desirable (Best) Practices

- Professional development for faculty
- Incremental curriculum development
- Hands-on lab/project components
- Capstone project that leads to research papers
- Continuous improvement to keep up with emerging concepts
- Multi-disciplinary collaborations
- Outreach to community colleges and K-12 schools
Acknowledgement

- NSF for the CCLI grant to support the creation of the courses and the certificate program
- CSE Dept, University at Buffalo, SUNY: For supporting my journey from a single lecture to a full fledged, comprehensive Certificate program
- External evaluator Dr. Jeanette Neal
- Reviewers of this papers: Very thorough, from a extra space character to a feedback on terminology used, suggestions for print and presentation, data analysis
- SIGCSE for letting me share my experience


