The Process of Data Science
(CSCI 4146)

Course Introduction

Dr. Amilcar Soares • May 08, 2018
Overview

1. Organization of the course
2. CORE CS
3. Expectations
4. About me
5. About you
6. What is Data Science?
7. What is Big Data?
8. What you will learn?
Organization of the course
1. Organization of the course

- Assignments & Labs (30%)
  - 10% x 2 assignments
  - 10% labs
- Midterm (30%)
- Project (45%)
  - Presentation: 20%
  - Project document: 25%
- Extra 5%

Late assignment policy:
- -10% per day (up to 3 days)
- In case of illness, late assignment will be accepted only with a legible doctor's note which includes the dates of illness.

Please familiarize yourself with the Faculty's Plagiarism Policy and the University's Academic Integrity Policy.
1. Organization of the course

- Assignments & Labs (30%)
  - 10% x 2 assignments
  - 10% labs
- Midterm (30%)
- Project (45%)
  - Presentation: 20%
  - Project document: 25%
- Extra 5%

//Do them!
// Don’t help each other; 10% off per/day (3-days max)
// Show up, help each other
// Will be a mix of technical and conceptual questions
// Acknowledge help, but do it yourself.
// Use slides and or demos
// Mostly on substance and writing
CORE CS
2. Culture of Respect in the Faculty of Computer Science (CORE CS)

- We believe inclusiveness and respect is fundamental to education
  - [https://www.cs.dal.ca/downloads/corecs_speaking_up.pdf](https://www.cs.dal.ca/downloads/corecs_speaking_up.pdf)

- More information about CORE CS
  - [http://www.dal.ca/faculty/computerscience/about/respect.html](http://www.dal.ca/faculty/computerscience/about/respect.html)
Expectations
3. Expectations

Technical
“Install jupyter, numpy and pandas, 
Read this matrix in this format, 
Read that matrix in that format, 
Multiply them together.”

“How is logistic regression different than naive bayes?”

“Solve this linear system in closed form.”

“When would you use this method?”

“Parallelize this for-loop”

“Tell me why your findings are important without explaining how you got them.”

“Read the docs, and use this function”
About me
4. About me

- (Dr.) Amilcar Soares
  - Postdoctoral fellow at the Institute for Big Data Analytics
  - Ph.D in Computer Science (Machine Learning and Geospatial Analysis). Universidade Federal de Pernambuco (Brazil).
  - Undergrad in CS and Master’s in Algorithms Design.
  - Worked as a Data Scientist for two companies (Neurotech and Cobli).
About you
5. About you

- Your name?
- Hometown?
- What course did you enjoy the most?
- Why did you choose Data Science specialization?
What is Data Science?
6. What is Data Science?

- **Professionally** about transforming data into valuable knowledge
- **Technically** is a collection of skills necessary due to growth in access to data and computing power
  - Data storage & distribution
  - Algorithms & techniques
  - **Justification**
6. What is Data Science?

Why is it that so many companies are interested in data scientists?

Because they have data (big data era).

- May be under-utilizing their data.
- May have disparate sources and kinds of data.
- May want to have data.
- May have “stuff” they didn’t realize is data.

(May just want to attract new talent)
Why data science is popular?

“You can’t manage what you don’t measure”
W. Edwards Deming / Peter Drucker

- The digital & big data era give managers access to considerably more data.
- Data science is the complete process of understanding the data to
  - Make better decisions
  - Find emergent patterns
  - Understand relationships
Who is a data scientist?

- Finding data scientists is hard
- Because it requires a combination of skills
- Also hard to find people who understand the value of data scientists...
MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21st century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS
- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE
- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing package e.g. R
- Databases: SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS
- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION
- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau
Data Science Is Multidisciplinary

By Brendan Tierney, 2012
6. What is data science?

It's the process of asking the right questions and applying the right methods to transform data into valuable knowledge.
What is Big Data?
7. What is Big Data?

- We have increased our capacities of storing data
- We have increased our capacities of generating data
  - Web 2.0: Internet users are both consumers and producers
  - Internet of Things: Every device generating data
- It's not just the amount of data:
  - Remember the four Vs
Volume

- The NSA, MI5 and the Chinese Government collect more than a petabyte of data every minute
- How much is 1PB?
- This not only affects storage
  - Think of a kNN algorithm

- Grid search on k or C?
- Invert a 1PB matrix?
Velocity

- The speed of data creation is often more important than volume
- Fast information makes it possible for groups to be more agile
- At the MIT Media Lab, they used location data from mobile phones to infer how many people were in store parking lots.
- Estimated Black Friday sales before the store’s themselves
- Rapid insights can provide an obvious advantage to analysts
Variety

- Data from
  - Messages
  - Transactions
  - Images
  - Sensors
- GPS coordinates from cell phones
- Integrating these different sources is a challenge by itself
Veracity

- All data are created equal in the eyes of science, but some data are more equal than others
- Noise can come from:
  - Missing information
  - Bad sensors
  - Misinformation
  - Sampling
- With different sources...
  - Different missingness
  - Different noise
  - Different truthfulness
  - Different biases
The four Vs of big data

- Don’t *just* need more storage capacity, or bandwidth...
- Need new ways of thinking about problems and solutions
- Offers challenges and opportunities
What you will learn?
8. What will we learn?

- HPC
- Statistical methods
- Visualization
- Scaling
- Reporting
- Analysis
- Collection
- Kinds of Data
  - Numerical
  - Text
  - Networks
- Advanced Data
  - Qualitative
  - Geospatial
  - Knowledge Bases
# Tentative Schedule

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
<th>Notes</th>
<th>Assignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 May</td>
<td>Course Introduction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 May</td>
<td>Data processing I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15 May</td>
<td>Lab: Environments and reading data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15 May</td>
<td>Data processing II</td>
<td>Pandas</td>
<td>A1 Out</td>
</tr>
<tr>
<td>17 May</td>
<td>Descriptive Statistics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22 May</td>
<td>Lab: Summarization</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22 May</td>
<td>Visualization I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24 May</td>
<td>Visualization II</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29 May</td>
<td>Lab: Data Visualization</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29 May</td>
<td>Statistical Models I</td>
<td>A1 Due</td>
<td></td>
</tr>
<tr>
<td>31 May</td>
<td>Statistical Models II</td>
<td>A2 Out</td>
<td></td>
</tr>
<tr>
<td>5 Jun</td>
<td>Lab: Regression Models</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 Jun</td>
<td>Classification I</td>
<td>Scikit-learn</td>
<td></td>
</tr>
<tr>
<td>7 Jun</td>
<td>Classification II</td>
<td>Group and topic definition</td>
<td></td>
</tr>
<tr>
<td>12 Jun</td>
<td>Lab: Classification</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12 Jun</td>
<td>Projects introduction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14 Jun</td>
<td>Clustering I</td>
<td>A2 Due</td>
<td></td>
</tr>
<tr>
<td>19 Jun</td>
<td>Lab: Clustering</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19 Jun</td>
<td>Clustering II</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21 Jun</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26 Jun</td>
<td>Databases</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26 Jun</td>
<td>Lab: MongoDB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26 Jun</td>
<td>HPC I Cluster &amp; Grid</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28 Jun</td>
<td>HPC II Cloud &amp; MapReduce</td>
<td></td>
<td></td>
</tr>
<tr>
<td>03 Jul</td>
<td>Midterm exam (No Lab)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 Jul</td>
<td>Open class to discuss projects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 Jul</td>
<td>Lab: HPC - MapReduce Framework</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 Jul</td>
<td>Text Data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12 Jul</td>
<td>Network Analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17 Jul</td>
<td>Lab: Text Analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17 Jul</td>
<td>Advanced Data - Geo Data</td>
<td>Guest Lecture</td>
<td></td>
</tr>
<tr>
<td>19 Jul</td>
<td>Advanced Data - Deep Learning</td>
<td>Guest Lecture</td>
<td></td>
</tr>
<tr>
<td>24 Jul</td>
<td>Project Presentations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26 Jul</td>
<td>Project Presentations</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>