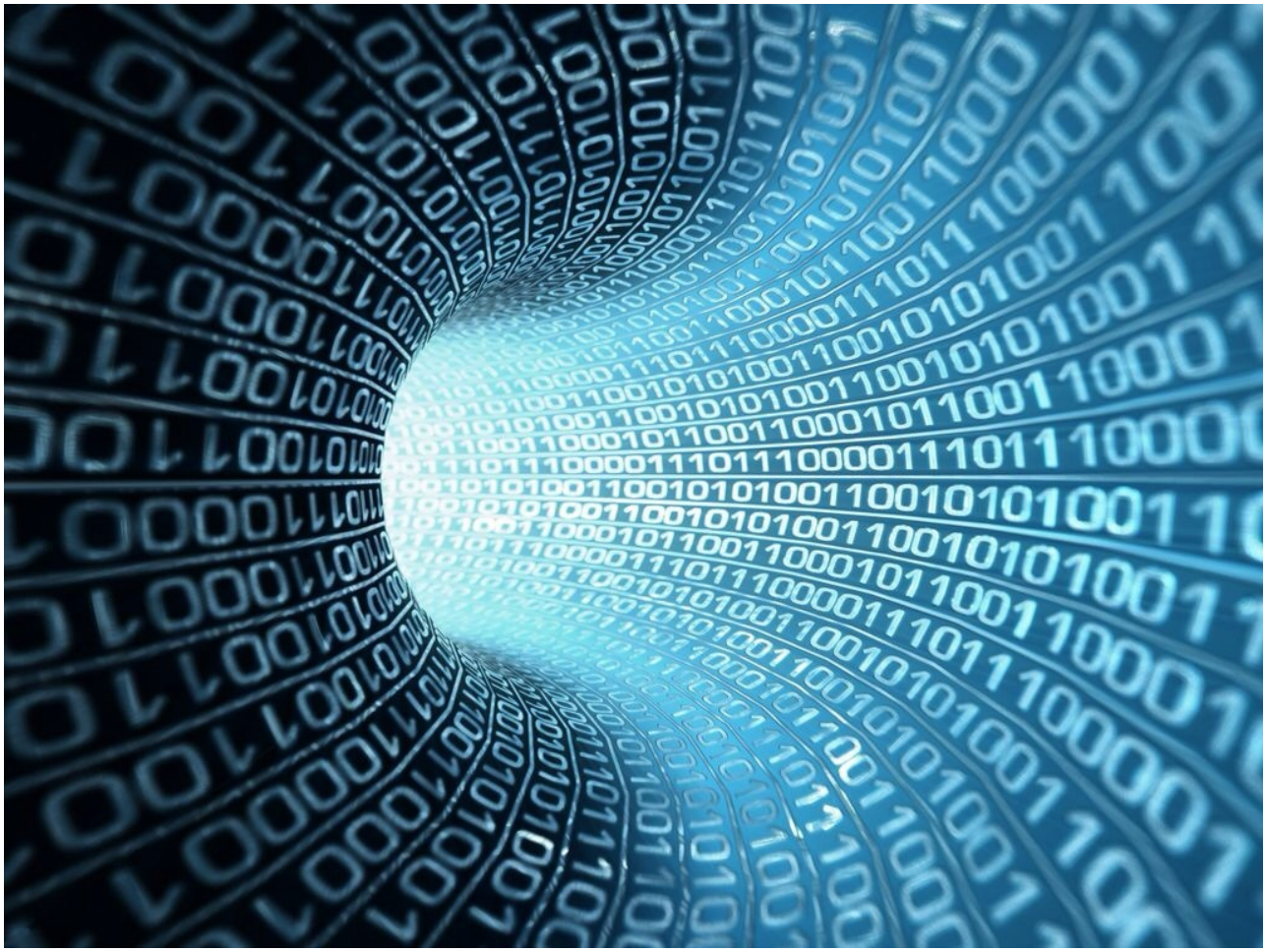
INFSCI 2725

Data Analytics

Syllabus



Chirayu Wongchokprasitti PhD
Philip J Cwynar MSIS, MBA

School of Information Sciences
University of Pittsburgh
Spring 2016

OVERVIEW:

"Big Data underscores the real need for a scalable and reliable analytics infrastructure to support data analysis and modeling." — Source: Business Wire 2012-09-06 19:01:00

"If you aren't taking advantage of big data, then you don't have big data, you have just a pile of data."
— Jay Parikh, VP of infrastructure at Facebook

21st century has been called by many "The Century of Data." We see more and more data collected with the expectation (justified by empirical evidence!) that analyzing these data will give organizations a competitive edge and will help them to excel. The amount of data collected is enormous and growing. In many cases, analyzing these data lags behind. Mark Twain wrote once *"A man who does not read has no advantage over a man who cannot read."* A similar sentence is most certainly true: "*A man who does not analyze his data has no advantage over a man who has no data."* Methods for analyzing data, called collectively "data analytics" are, therefore, crucial in this business.

INFSCI 2725 is an introductory course in the area of the so-called "Big Data," aiming at graduate students in Information Science and related disciplines. It focuses on essential technologies that are underlying collection, storage, and processing of data. The biggest and most important of these is analyzing data (many of the techniques are covered in more depth in *Data Mining*), even though we will spend some time on the topic of data collection and storage (covered in detail in the *Advanced Topics in Database Management* course).

The skills that you will need as a data scientist span over a large number of areas, such as statistics, databases, systems, programming, machine learning, artificial intelligence, business intelligence, and visualization. The knowledge that you need to acquire as a data scientist is not easy to gain by following courses in each of these areas, as concepts at their intersection are often obscured. The goal of this course is to simplify and present the most relevant material that you would otherwise have to learn in traditional disciplines and to point out the commonalities between these disciplines. The course is not designed to teach you the formal details of statistical procedures used in data analysis or to make you an expert practitioner of the specific analysis tools that you can and should get in other courses, typically in Statistics, Artificial Intelligence and Data Mining. The point of this class is to develop broad critical abilities to approach collection, storage, and analysis of very large data sets. The course aims at improving your ability to think about data and information and to choose ways of extracting information and knowledge from data.

This is a fairly new course and we are still looking for the best way of synthesizing and presenting the material. The fact that the course is new is not the only difficulty here. An additional problem that we have to cope with is that the material itself is very new and changing rapidly. We are pretty much working at a frontier and no textbooks exist that cover all the material. Because of these reasons, the schedule may change slightly as we go. The set of readings may change to some degree as well. We have planned assignments, a project that will allow you to get hands-on experience in data analytics, and two examinations that will test your general mastery of the material.

As you might have already experienced, being a graduate student requires intelligence, independent, creative thinking, and most of all commitment to hard working. This course reinforces this. There may be a higher than usual amount of readings. We have selected them in such a way that they will be fun to read and we expect that you will do them all with pleasure. The assignments and the project will offer you hand-on experience, something that will make you appreciate the size of data sets that we have to work with in the 21st century. The workload in this class will be moderately heavy, but we believe that you will find it interesting and important. We require your commitment, doing the readings, coming to classes, and being their active participant. In return, we promise that you will have fun and you will learn many useful skills.

YOUR RESOURCES:

The course:

Name : INFSCI 2725: Data Analytics
CRN : 24899
Credits : 3.0

The instructors:

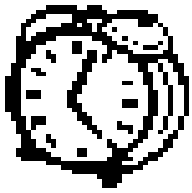
Philip J Cwynar MSIS MBA

Office : IS 708
Email : pcwynar@pitt.edu
WWW : https://www.linkedin.com/in/philip-cwynar-msis-mba-ba393b4

Chirayu Wongchokprasitti PhD

Office : IS 708 or 5607 Baum Boulevard, BAUM 435F
Email : chw20@pitt.edu
WWW : http://www.pitt.edu/~chw20/

Meeting times and locations:



Classes (404 IS Building):
Mondays, 6:00-8:50pm
Office hours by appointment

Your colleagues:

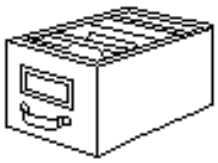
Name: _____
Phone: _____
Email: _____
Name: _____
Phone: _____
Email: _____
Name: _____
Phone: _____
Email: _____
Name: _____
Phone: _____
Email: _____

THE COURSE:

Objectives:

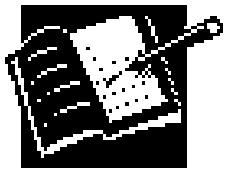
The primary objective of this course is to make you acquainted with analytical procedures that are useful in processing very large amounts of data. This should make you better prepared for the deluge of data that you will encounter in practical environments. The assignments and the term project will offer you an opportunity to gain hands-on experience. Finally, being successful in the course should contribute to the development of your academic self-esteem.

Prerequisites:



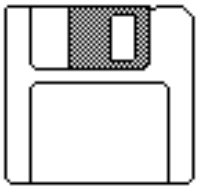
All students in the course should have taken a statistics course (this is a prerequisite to the Information Sciences program!). Most important concepts will be briefly reviewed in class as needed, but this should not be seen as a substitute for formal preparation. Some of the assignments and your term project will most likely require knowledge of programming. While all assignments and the term project are group work, you will fare better if you know Java, which is used in almost all "Big Data" type of environments and is also a requirement for the Information Science "Big Data Analytics" track. In this course, we will have at least one programming assignment that will require you to know Java and use it to access a Big Data programming environment. The most important prerequisite, however, is your interest in "Big Data," motivation, and commitment to learning.

Required reading:



Readings for this course will be taken from several sources, listed in the syllabus. Additional readings may be assigned in the course of the semester.

Data analysis software:



There are quite a number of computer programs that support data analytics. During a part of the course, will be using ***GeNIe***, a program developed at the Decision Systems Laboratory and currently licensed to BayesFusion, LLC, that is developing it further. BayesFusion, LLC, is making its software available free of charge for academic use at: http://www.bayesfusion.com/. ***GeNIe*** has a comprehensive Wiki-based on-line help that will supplement the textbook material.

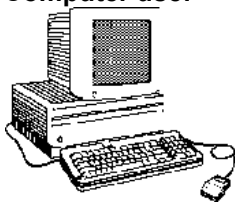
In addition to ***GeNIe***, we will be using Weka (Waikato Environment for Knowledge Analysis), a suite of machine learning software developed at the University of Waikato, New Zealand, and available at http://www.cs.waikato.ac.nz/ml/weka/.

We will introduce the basics of R and you may want to consider using R in your assignment a project work.

You need to be able to program in Java, particularly JAVA 8, (and if possible, Python 2.7 would be a plus), so that you can interface to existing Big Data programming libraries. Often, you will work in the assignments on smaller data sets. Please remember that the methods to analyze small data sets are closely related to those that you would use for very large data sets. Practicing the former will help you in learning how to work with the latter.

WORK REQUIREMENTS:

Computer use:

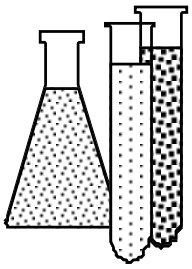


You will be required to use computers for typesetting your documents and for working with data analysis software (at least ***GeNIe*** and Weka). Your documents have to be produced in an easily readable/printable format, such as PDF (recommended). There will be no limitations on which computer system you use (except for natural limitations, such as functionality of the package). Most of our communication will be electronic and you will be expected to use electronic mail on a daily basis.

Assignments:

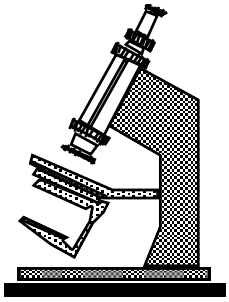
There will be eight or nine assignments that will help you to practice the material covered in class and will help me to identify those parts of the material that you have difficulties with. Assignments can be done in groups of up to three students (not more than three and this is not negotiable!), formed outside of the class meetings. The groups do not need to be the same for all assignments. You can change the group composition if things are not working out for you. The assignments have to be turned in on time, and all members of the group are responsible for meeting the deadline. Make sure that you include your observations and conclusions in your submission. We are often asking for these explicitly and have noticed that some students still report only mere results without discussing them briefly.

Term project:



A major part of the training that you will receive as part of this course will result from performing a project. The description of the project is attached to this syllabus. The due date (inflexible!) is marked on the course schedule. We advise you to start working on the project as soon as possible, as it may involve a considerable amount of work. The project will give you an opportunity to get hands-on experience with "Big Data" and apply techniques that you will have learned in the course. You will be expected to team up for the project in groups of up to five students (not more than five and this is not negotiable!), although it is possible to do the project individually. The deliverables are: (1) a mid-semester progress report, and (2) a final report. The due dates are marked on the course schedule.

Exam:

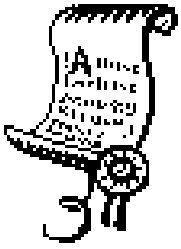


There will be one midterm exam and one comprehensive final exam, both closed book but you can bring with you to the exam one double-sided letter-size sheet of paper with notes. There are no limits on the font size – you can cram as much information on these two pages as you wish – but the notes have to be handwritten personally by you and this is a strict requirement. Copied or computer-printed sheets are not allowed.

Time load:

To help you with planning your semester, we would like to give you an idea of the minimal workload in this course. Expect to spend around six hours (preferably nine) of quality time outside of class for every class meeting. We estimate that you will need about four hours (and two hours more as buffer) to do the readings and two hours (and one hour buffer) on average to do the assignments. We assume here that you will be doing assignments in groups. In case you work on them individually, your time load may increase. The term project will quite likely demand between 20 and 30 hours of your time.

Grading:



Your final grade for the course will be determined as follows:

Assignments	:	30%
Term project	:	30%
Midterm exam	:	20%
Final exam	:	20%

On the top of this all, you can obtain up to 10% of the total score for in-class participation and in class quizzes.

COURSE SCHEDULE

Here is an outline of the course schedule. Please keep in mind that we may change somewhat in the course of the semester. Should there be changes to the schedule, we will announce them in class and post a new version of the schedule on the course web site

PART I: INTRODUCTION

The first two classes will be devoted to organizational matters, overview of the course, a critical review of the "Big Data" field and also of the "Big Data" track within the MSIS program.

January 11
[Philip]

Getting to know each other
Organization and overview of the course
An introduction to the SIS "Big Data Analytics" track
Readings:
http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

January 18

*** **NO CLASS** *** Martin Luther King, Jr. Day 2016

PART II: STORAGE AND PROCESSING OF BIG DATA

This two-class block will review what we know about storing and processing data. Most of the material covered in the first part of the first class is typically covered in introductory Database courses. After a brief review of what we know from databases. We will look at the most popular "tricks" used by "Big Data" giants, such as Google, Amazon, and others. Finally, we will focus on processing huge quantities of data in a parallel, distributed fashion.

January 25

Fundamental tasks and applications in data analytics
The three Vs of "Big Data:" Volume, Velocity, and Variety

[Philip]

Trends in "Big Data."
Career prospects and opportunities related to "Big Data."

Data storage
Structured data
RDBMS,
Normalization
De-Normalization Data warehousing (Innmon / Kimball) ETL, Dimensional Data Modeling
Unstructured Data
NoSQL defined
Examples
Systems (MongoDB, Riak, Cassandra), GraphDB Neo4j

February 1
[Philip]

Data processing
Hadoop
Map/Reduce
Pig, Hive HBase, Spark
Cloudera VM HUE overview
http://hadoop.apache.org/
http://hadoop.apache.org/docs/r0.20.2/hdfs_design.html
http://hive.apache.org/docs/r0.9.0/
http://hbase.apache.org/
http://pig.apache.org/docs/r0.8.1/
http://spark.apache.org/
http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en/us/pubs/archive/38125.pdf

PART III: FOUNDATIONS

This two-class block forms an introduction to data analytics. We will start with the theoretical foundations, i.e., techniques for analyzing data developed in the field of statistics. In addition, we will look at the issue of verification and testing. Results need to come with some kind of guarantee of believability, so one needs to test whether they are robust and to what degree they are to be believed.

***** Term project team composition due *****

February 8 [Chirayu]

***** Homework assignment 1 due (distributed storage and processing of data) *****

*[Readings: Your favorite statistics textbook, **GeNIe**]*

Fundamental concepts from statistics
Uncertainty, probability, variance, sampling, randomness, elements of data analysis (a review). Describing and displaying data, correlation.
Joint probability distribution: the foundation of any analytic technique
Conditional probability distribution, Bayes theorem, prior and posterior probability distribution. Time series data.
Supervised and unsupervised learning.

February 15 [Chirayu]

***** Homework assignment 2 due (fundamental concepts from statistics) *****

Verification and testing
*[Readings: Weka, **GeNIe**]*
Significance testing, confidence intervals, sensitivity, specificity, ROC curves, AUC, calibration.
Supervised and unsupervised learning, generalization, over-fitting, over-fitting avoidance. Cross-validation.

February 22 [Philip & Chirayu]

***** Homework assignment 3 due (validation and testing)**

The R programming environment
[Readings: RProject, http://www.r-project.org/]
http://www.dezyre.com/article/-why-r-programming-language-still-rules-data-science/161]
You will most likely learn R when taking other classes, such as statistics or data mining. In this course, we will just touch the basics to show you the power of R and why it is becoming the programming language of choice for many statisticians, data analysts, and data scientists.

February 29

***** Homework assignment 4 due (R programming)**
***** MIDTERM EXAM *****
(You will be responsible for the material of Parts I through III)

March 7

Spring Break
***** NO CLASS *****

PART IV: DATA ANALYSIS

This block is the main block of classes in this course. We will go systematically through the existing collection of techniques for analyzing data, starting from probabilistic approaches, causal discovery, through data mining, and interdisciplinary work at the boundary of computer and information sciences, artificial intelligence, machine learning, statistics, and philosophy. We will also look at methodologies applied in social network analysis.

March 14 [Chirayu]

***** Mid-semester term project report due *****

Probabilistic approaches
*[Readings: Weka, **GeNIe**, Hopfield&Tank; Jain et al.; McClelland&Rumelhart, Videos:*
http://www.ted.com/talks/jeff_hawkins_on_how_brain_science_will_change_computing.html (20'16")
http://www.ted.com/talks/gero_miesenboeck.html (18'52")]
Model assessment and selection, regression models, kernel methods (support

vector machines, principal component analysis), ensemble learning (bagging, boosting, Bayesian model averaging), naïve Bayes classifiers, Bayesian networks.

March 21 ***** Homework assignment 5 due (probabilistic approaches) *****
[Chirayu] Causality and causal discovery
*[Readings: Cooper&Herskovitz; Druzdzel&Glymour; Spirtes et al.; Glymour et al.; Langley et al.; **GeNIe**]*
Constraint-based search for causal structures
Bayesian search for causal structures

March 28 ***** Homework assignment 6 due (causal discovery)**
[Chirayu] Logic-based approaches
[Readings: Weka]
Decision tree learning, rule induction, Inductive Logic Programming.

April 4 ***** Homework assignment 7 due (logic-based approaches) *****
[Chirayu] Social network analysis
[Reading: Liu]
Video: "Michael Anti: Behind the Great Firewall of China"
http://www.ted.com/talks/michael_anti_behind_the_great_firewall_of_china.html (18'52")

PART V: PRESENTATION OF RESULTS

An integral part of data analysis is presentation of results. Results need to be presented to the user in such a way that they are digestible. This block of classes will be devoted to this important topic.

April 11 ***** Homework assignment 8 due (social network analysis) *****
[Philip] Presentation of results
[Readings: Tufte]
Data visualization, exploratory data analysis,
design and creation of charts and graphs.
Multi-media databases, visual data analytics.

PART V: CONCLUSION

April 18 ***** Term Project Due & Presentation *****
Our last classroom meeting will be a grand conclusion of the course. We will have studied the final versions of your project reports and your models. We will have project demonstrations and announcement of the winner of Marek's Best Project Award. Please bring to class questions about the material that you may want to discuss before the final exam. In as much as remaining time allows us, we will talk about the course and possible ways of improving it in the future.

April 25 ***** FINAL EXAM *****
(You will be responsible for the material of Part IV only.
We reserve the right to change the scope of the final exam to cover the material of the entire course in case the class performance on the midterm exam is weak. If we use this right, we will announce this explicitly before the final exam.)

Sources of readings:

[Chang et al.] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E. Gruber. *"Bigtable: A Distributed Storage System for Structured Data",* In Proceedings Of The 7th Conference On Usenix Symposium on Operating Systems Design and Implementation, Volume 7, 2006, available at http://static.usenix.org/events/osdi06/tech/chang/chang.pdf

[Dean&Ghemawat] Jeffrey Dean and Sanjay Ghemawat. *"MapReduce: Simplified Data Processing on Large Clusters." Communications of the ACM*, 51(1):107-113, 2008
Also in *Proceedings of USENIX Association OSDI '04: 6th Symposium on Operating Systems Design and Implementation*, pages 137-149, 2004, available at http://static.usenix.org/event/osdi04/tech/full_papers/dean/dean.pdf

[Cooper&Herskovitz] Gregory F. Cooper & Edward Herskovitz. *"A Bayesian Method for the Induction of Probabilistic Networks from Data." Machine Learning*, 9:309-347, 1992

[Druzdzel&Glymour 1994] Marek J. Druzdzel & Clark Glymour. *"Application of the TETRAD II Program to the Study of Student Retention in U.S. Colleges."* In Proceedings of the AAAI-94 Workshop on Knowledge Discovery in Databases (KDD-94), pages 419-430, Seattle, WA, 1994, available at: http://www.pitt.edu/~druzdzel/abstracts/kdd94.html

[**GeNIe**] **GeNIe**, a decision modeling environment developed at the Decision Systems Laboratory and available at http://www.bayesfusion.com/. Most up-to-date, Wiki version of the documentation is available at https://dslpitt.org/genie/wiki/Main_Page

[Glymour et al.] Clark Glymour, Richard Scheines, Peter Spirtes & Kevin Kelly. *"Discovering Causal Structure: Artificial Intelligence, Philosophy of Science, and Statistical Modeling."* Chapters 1-3, pages 3-59, San Diego, CA: Academic Press, Inc., 1987

[Hopfield&Tank] J.J. Hopfield & D.W. Tank. *"Neural Computation of Decisions in Optimization Problems." Biological Cybernetics*, 52:141-152, 1985

[Jain et al.] Anil K. Jain, Jianchang Mao & K.M. Mohiddin. *"Artificial Neural Networks: A Tutorial." Computer*, March 1996

[Langley et al.] Pat Langley, Herbert A. Simon, Gary L. Bradshaw & Jan M. Zytkow. *"Scientific Discovery: Computational Explorations of the Creative Processes."* Chapters 1-2, pages 3-62, Cambridge, MA: The MIT Press, 1987

[Liu] Bing Liu. *"Web Data Mining."* Data-Centric Systems and Applications, ISBN 3642194591, Chapters 1-3, pages 269–309, Springer, 2011, available in electronic format through the University of Pittsburgh library

[McClelland&Rumelhart] James L. McClelland & David E. Rumelhart. *"An Interactive Activation Model of Context Effects in Letter Perception: Part I. An account of Basic Findings." Psychological Review*, 88(5):375-407, 1981

[McKinsey] McKinsey Global Institute. *"Big data: The next frontier for innovation, competition, and productivity."* June 2011, Free report available at: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

[Sadalage&Fowler] Pramod J. Sadalage & Martin Fowler. *"NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence."* Addison-Wesley Professional, 2012

[Spirtes et al.] Peter Spirtes, Clark Glymour & Richard Scheines. *"Causation, Prediction, and Search, Second Edition (Adaptive Computation and Machine Learning)."* Cambridge, MA: The MIT Press, 2000

[Stonebraker&Hellerstein] Michael Stonebraker and Joseph M. Hellerstein, *"What Goes Around Comes Around."* Unpublished manuscript available from several locations on the web (e.g., http://www.cs.umass.edu/~yanlei/courses/CS691LL-f06/papers/SH05.pdf, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.113.5640)

[Tufte] Edward R. Tufte. *"The visual display of quantitative information."* Graphics Press, 1983

[Weka] Weka (Waikato Environment for Knowledge Analysis), a suite of machine learning software developed at the University of Waikato, New Zealand and available at http://www.cs.waikato.ac.nz/ml/weka/

[White] Tom White, *"Hadoop: The Definitive Guide. MapReduce for the Cloud."* O'Reilly Media, 3rd Edition, 2012

TERM PROJECT

A major part of the training that you will receive as part of this course will result from working on a group project involving solving a real data analysis problem. You will be expected to try your skills at one of the world-wide data analytics competitions (typically organized by Kaggle, http://www.kaggle.com/), which we will choose at the beginning of the semester and announce in class. Usually, participants in data analytics competitions compete for significant financial prizes (ranging from a few thousand to a few million dollars).

Your task will be to win the competition. While you may earn a good grade even if you do not win the competition (since this is a world-wide competition, it may happen that nobody from the course wins; also it is impossible for every team in the course to win ☺), winning the competition (if the competition ends before the end of the semester) or being the first among all teams world-wide on the term project submission deadline (marked in the syllabus) carries an automatic A+ grade for the course for every member of the team, regardless of the team members' performance on the assignments and the exams. As far as the competition prize is concerned, it belongs to the team members and they decide what to do with it, although they are encouraged to bring some cookies to the last class meeting to thank the other classmates and the teacher for their support ☺.

Project teams:

We will spend some time during our classroom meetings talking about the project and the relevance of the material to the project. The real work, however, will happen outside of the classroom. You will form teams of up to five students (not more than five and this is not negotiable!) that will split the work and carry the project to a successful completion.

Computer support and analytic techniques used:

While it would be nice to apply some of the analytic techniques that you will have learned in the course, do not feel pressured to apply them. Your participation in the competition should be real and you should use whatever works and whatever gives you the highest chance of winning (within the bounds of the rules of the competition). The same holds for the computing environment – you should rely on the computer system and the programming environment that you have access to and that works for you best (again, within the bounds of the rules of the competition).

Mid-semester progress report:

In order to make sure that you do not count on completing the project in the last few days before the deadline (this, given the competition from highly motivated teams inside and outside of this course, would not work ☹), you will be expected to submit a mid-semester progress report containing the name of your team (as used in the competition) an introduction to the final report, a description of your approach to the competition, i.e., the nature of the problem and the data, your engineering solution for storing and retrieving the data (the data will most likely be very large and will probably not fit easily into computer memory), the analytic methodology that you have chosen, its performance relative to other teams, a description of the work that you have completed so far (including your initial results), and a detailed plan of action for the remainder of the semester. The main purpose of this report is to help you in planning your work and spreading it over the course of the semester. The deadline for submitting the mid-semester progress report is marked on the syllabus.

The final report:

Your final project report (in PDF format) should be an extension of your mid-semester progress report describing, in addition to what your mid-semester progress report contained already, the steps that you have taken to win the competition and the final result (including a screen shots of the final score board). As far as the length of the report is concerned, it should be as short as possible but not shorter (i.e., it should describe in sufficient detail

what needs to be described). It should report on all major design decisions and choices of the analytic techniques that you have ended up applying. It should contain sufficient detail for anybody to understand and to possibly reproduce whatever you have done. As Americans say succinctly and jokingly, "*The proof is in the pudding"*, which in this case means that your score and your position on the competition leader board should speak by themselves about the quality of your effort.

Evaluation criteria:

The criteria for grading your project reports are: your ultimate performance in the competition, organization and planning of your work (as expressed by your reports), soundness, creativity, and, finally, clarity of your writing and expressing your ideas.

For your project work to be acceptable, you have to follow all the rules set out by the competition organizers. In particular, if they do not allow forming multiple teams, you should not form multiple teams.

My advice is: *Aim at excelling in this project*. A winning project should help you to advance your career. It is a world-wide trend to look for talent among the competition winners (see, for example, http://www.recruitingdivision.com/why-googles-it-recruiting-strategies-include-programming-contests/, http://www.staff.com/blog/using-programming-competitions-to-hire-superstar-coders/, https://www.kaggle.com/content/kaggle/img/casestudies/Kaggle%20Case%20Study-Facebook.pdf). It is easier and more rewarding to excel as a student than as a "mature" professional. Try everything that works and describe in sufficient detail what you did.

Suggested approach:

You will most certainly need to deal with the complexity introduced by the size of your data set. The first step of your project should be a good engineering design of data structures that will allow you, if physically possible, to keep your entire data in computer memory. This will speed up your analysis tremendously. Having to go through data stored on an external disk slows down your analysis and reduces significantly what you can do with the data.

In addition to analytical methods that you think should work in your project and are worth trying, it is a good idea to take part in the on-line discussions that most competitions set up. Very often teams that win progress prizes have to publish their methodology. You can learn a lot from these reports and from other teams in general. Finally, talk to other teams in this course – you can learn a lot from each other.

All projects will be presented during the last class meeting (see the course schedule for the date).