CWRU Action Form for Majors/Minors/Programs/Sequences/Degrees Docket # 15-CSE-PAF-1106
(instructions on back)

College/School: Case School of Engineering
Department: EECS

PROPOSED:
_____ major
_____ minor
__X_ program
_____ sequence
_____ degree

TITLE: Data Sciences Curriculum

EFFECTIVE: Fall _____ (semester) 2015 _____ (year)

DESCRIPTION:

This proposes a new major in Data Sciences to be housed in the Department of Electrical Engineering and Computer Science. The attached documents detail the curriculum and new course additions, which will be filed with separate CAFs.

Is this major/minor/program/sequence/degree:
__X_ new
_____ modification
_____ replacement

If modification or replacement please elaborate:

Does this change in major/minor/program/sequence/degree involve other departments? _____Yes _____No

If yes, which departments?

Contact person/committee: Soumya Ray

SIGNATURES:
DATE
Department Curriculum Chair(s)/Program Directors: 6/15/15
Department Chair: See attached MATH/STATs Chair email approval 6/12/15 6/15/15
College/School Curriculum Committee Chair: 6/18/15
College/School Dean(s): See attached Assoc Dean Buchner's email approval 6/19/15 10/20/15
UUF Curriculum Committee Chair:

File copy sent to: _____Registrar _____Office of Undergraduate Studies/Graduate Studies
_____Other:

Requirements for BS in Data Science and Analytics

General Education/Engineering Core Requirements
First Seminar
2 University Seminars
Department Seminar (ENGL/ENGR 398)
Capstone (DSCI 399) (included below as part of the major)
2 semesters of PHED

Humanities/Social Science Electives to total 12 credit-hours of 3- or 4-credit-hour courses
MATH 121, 122, 223, 224
CHEM 111
PHYS 121, 122
EECS 132 (included below as part of the major)

[Note: This is the Engineering Core without ENGR 145, 200, 210, 225]

Requirements for Major
EECS 132
DSCI 133, 234, 341, 342, 343, 344, 345, 398, 399
EECS 302, 340, 393
MATH 201
Probability/Statistics Elective
Computer and Data Security Elective
3 DSCI Technical Electives (Choice of one of two foci)
3 Technical Electives

Open Electives
To reach a total of 125 credit-hours to complete the degree program

From: Mark De Guire [mailto:mrd2@case.edu]
Sent: Monday, October 19, 2015 3:25 PM
To: Jeffrey Duerk
Cc: Ann Boughner; Kathleen Ballou; Aaron Jennings; Chung-Chiun Liu; Horst von Recum; Joao Maia; Rigoberto Advincula; Sree Sreenath; Wyatt Newman; Xin Yu; Yasuhiro Kamotani; Marc Buchner; Kenneth Loparo
Subject: Approved by CSE faculty: new undergraduate major program in Data Science and Analytics

Dear Jeff,

The faculty of the Case School of Engineering approved, by a vote of 52 to 1, the proposed new major program in Data Science and Analytics.

The total number of votes cast (52 of 112 faculty) exceeds the number required for a quorum (40% of voting faculty) in the Case School of Engineering, and (as you know) followed a duly called special faculty meeting at which the proposal was discussed at length.

I believe page 3 of the attached proposal needs your signature. The proposal then would be forwarded to the Secretary of the University Faculty, Rebecca Weiss, and Dean of Undergraduate Studies Jeffrey Wolfowitz, to be taken by the Faculty Senate Committee on Undergraduate Studies.

Thanks to Ann Boughner for organizing and overseeing the voting, to Heidi Fanta for helping to organize the special faculty meeting, and to you and Marc Buchner for keeping this process moving ahead.

What is normally done to inform the faculty of the outcome — an announcement from your office, from the Executive Committee, or some other way?

Best regards,
Mark

Mark De Guire
Chair, CSE Executive Committee, 2015–2016

---------- Forwarded message ----------
From: **Ann Boughner** <aeb3@case.edu>
Date: Fri, Oct 16, 2015 at 10:25 AM
Subject: RE: IMPORTANT VOTE NEEDED on the proposed new undergraduate major porgram in Data Science and Analytics
To: Mark De Guire <mrd2@case.edu>
Cc: Marc Buchner <marc.buchner@case.edu>

FYI—voting results for Data Science & Analytics program:

- 52 votes total
- 51 yes votes
- 1 no vote

Ann Elizabeth Boughner
aeb3@case.edu
Director of Human Resources & Leadership Development
Case School of Engineering
Case Western Reserve University, Nord Hall #520
10900 Euclid Avenue
Cleveland, OH 44106-7220
Voice: 216-368-5922
Fax: 216-368-6939

CASE SCHOOL
OF ENGINEERING
CASE WESTERN RESERVE
UNIVERSITY

Jeffrey L. Duerk, Ph.D.
Dean, Case School of Engineering
Leonard Case Professor

June 22, 2015

As dean of the Case School of Engineering, I strongly support the new major in Data Sciences to be housed in the Department of Electrical Engineering and Computer Science (EECS). In collaboration with the Business-Higher Education Forum, and with the leadership of President Snyder, CWRU discovered there is a great need for data science experts emerging from university at the completion of their undergraduate (UG) education.

By developing a distinctive UG program in Data Science, we will provide the data science and analytics training needed for undergraduate students in the ongoing era known as "Big Data". Our EECS program will develop the skills and provide instruction needed in handling large amounts of data and transform our thinking from a collection of vast amounts of data into one that focusses on the data's conversion to actionable information; by developing this with the BHEF and industry partners, our degree program will have a unique focus on real-world data and real-world applications.

This major will also be one of the first undergraduate programs nationwide, which puts CWRU in the forefront with its unique, rigorous curriculum. The curriculum includes mathematical modeling, informatics, data analytics, visual analytics and project-based applications, all being elements of the future emerging field of data science.

With an undergraduate minor already in place, CWRU now is responding to a strong and aggressive expansion in research and education, along with market demand, for students trained in computer science, mathematical modeling, statistical analysis and other areas related to Big Data. I strongly endorse our plan to create this new undergraduate degree program in Data Sciences.

Warm regards,

Jeffrey L. Duerk, Ph.D.
Dean
Leonard Case Professor

Case School of Engineering, 10900 Euclid Avenue, Cleveland, Ohio 44106-7207
Visitors and Deliveries: 500 Nord Hall
Phone 216.368.3227 E-mail duerk@case.edu
http://engineering.case.edu

ASSOC. DEAN BUCHNER'S EMAIL APPROVAL

From: Marc Buchner [mailto:mxb11@case.edu]
Sent: Friday, June 19, 2015 9:57 AM
To: Kathleen Ballou
Cc: Ken Loparo
Subject: Re: DSCI PAF signature needed

Hi Kathleen (and Ken),

I approve of the DSCI curriculum as put forward by the EECS department and voted upon by the CSE UG committee.

However, please attach Ken's rationale for the degree program and a letter of support from Jeff Duerk as we were recommended to do by Jeff Wolcowitz. I have Ken's rationale somewhere in my email but you should be able to get it directly from him.

I can write the Dean's letter but I won't be able to get to it until later today ... probably this evening at the earliest.

Thanks,
Marc

Sent from my iPhone

On Jun 19, 2015, at 9:27 AM, Kathleen Ballou <kad4@case.edu> wrote:

Marc,

Attached is the DSCI PAF (15-CSE-PAF-1106) that has been approved by the CSE UG Committee, which needs your signature. Please review and send an email with your approval.

Thanks,

Kathleen A. Ballou
Project Manager & Assistant to the Associate Dean
Case School of Engineering

<Data Sci 15-CSE-PAF-1106 signature needed 6.19.15.pdf>

Rationale for New Data Science Program

In July 2013, President Snyder was announced as the new chair of the Business and Higher Education Forum (BHEF). In her role as BHEF chair, she announced in February 2014 plans for CWRU to develop a distinctive UG program in Data Science (www.bhef.com/news-events/releases/bhef-chair-and-case-western-reserve-university-president-barbara-r-snyder). The majority of Data Science programs are at the MS level and above, and the demand for data scientists is expected to grow substantially in the next 5-10 years. According to a report by the McKinsey Global Institute, the United States alone will need to increase the number of graduates with skills in handling large amounts of data by as much as 60 percent, and it is estimated that there will be half a million jobs that need to be filled in the next five years.

Fall Semester		Spring Semester	
Freshman Year			
Class-Laboratory-Credit Hours		Class-Laboratory-Credit Hours	
SAGES First Year Seminar	4-0-4	SAGES University Seminar	3-0-3
CHEM 111 Chemistry I	4-0-4	PHYS 121 Physics I: Mechanics	4-0-4
MATH 121 Calculus I	4-0-4	MATH 122 Calculus II	4-0-4
EECS 132 Introduction to Java	3-2-3	DSCI 133 Introduction to Data Science	3-0-3
PHED 101 Physical Education	0-3-0	PHED 102 Physical Education	0-3-0
		Open Elective	3-0-3
Total:	15-2-15	Total:	17-3-17
Sophomore Year			
SAGES University Seminar	3-0-3	DSCI 341 Introduction to Databases	3-0-3
PHYS 122 Physics II: Electricity & Magnetism	4-0-4	MATH 224 Differential Equations	3-0-3
MATH 223 Calculus III	3-0-3	EECS 340 Algorithms	3-0-3
DSCI 234 Structured/Unstructured Data	3-0-3	HM/SS Elective	3-0-3
EECS 302 Discrete Mathematics	3-0-3	Probability/Statistics Elective[1]	3-0-3
Total:	16-0-16	Total:	15-0-15
Junior Year			
Class-Laboratory-Credit Hours		Class-Laboratory-Credit Hours	
DSCI 342 Introduction to Data Science Systems	3-0-3	ENGL/ENGR 398 Professional Communication	3-0-3
EECS 393 Software Engineering	3-0-3	DSCI 344 Scalable Parallel Data Analysis	3-0-3
HM/SS Elective	3-0-3	Computer and Data Security Elective[2]	3-0-3
DSCI 343 Introduction to Data Analysis	3-0-3	DSCI 345 Files, Indexes and Access Structures for Big Data	3-2-3
MATH 201 Linear Algebra	3-0-3	Technical Elective	3-0-3
Total:	15-0-15	Total:	15-2-15
Senior Year			
Class-Laboratory-Credit Hours		Class-Laboratory-Credit Hours	
Technical Elective	3-0-3	HM/SS Elective	3-0-3
DSCI Technical Elective[3]	3-0-3	DSCI Technical Elective[3]	3-0-3
DSCI 398 Senior Project I	1-6-4	DSCI 399 Senior Project II	0-8-4
DSCI Technical elective[3]	3-0-3	Technical Elective[4]	3-0-3
HM/SS Elective	3-0-3	Open Elective	3-0-3
Total:	13-6-16	Total:	12-8-16

GRADUATION REQUIREMENT: 125 hours total, green=new courses, blue=CAFs need to be filed to modify prerequisites to include DSCI 234

1. **Probability and statistics electives**: MATH 380, STAT 325

2. **Computer and Data Security electives**: EECS 444, MATH 408, new course to be developed

DSCI 398/399: Capstone project, 8 credits, possibly in conjunction with a co-op

3. DSCI Technical Electives in Signal Processing:

EECS 246: Signals and Systems (Required)

EECS 313: Signal Processing (Required)

STAT 322: Statistics for Signal Processing (Required)

Technical Electives: electives from minor list, EECS courses

3. DSCI Tech Electives: Systems and Analytics

Systems: (EECS courses needing EECS 233 will need to adjust prerequisites to include DSCI 234)

EECS 325/425: Computer Networks, other networks courses

EECS 338: Operating Systems and Concurrent Programming

Cloud Computing (currently 600)

Analytics:

DSCI 390: Machine Learning for Big Data

DSCI 391: Data Mining for Big Data

EECS 339: Web Data Mining

EECS 346: Engineering Optimization

EECS 440: Machine Learning

EECS 442: Causal Learning from Data

4. Technical Electives: electives from minor list or EECS courses

DSCI 133: Introduction to Data Science and Engineering for Majors

Credit Hours: 3

Course Pre-Requisites:
For Data Science & Analytics Major Students: ENGR 131 or EECS 132

Weeks 1-7 provide an overview of data science.
Weeks 8-14 provide project based learning in data science.

Course Description (up to 2100 characters):

This course is an introduction to data science and analytics.

In the first half of the course, students will develop a basic understanding of how to manipulate, analyze and visualize large data in a distributed computing environment, with an appreciation of open source development, security and privacy issues.

Case studies and team project assignments in the second half of the course will be used to implement the ideas. Topics covered will include: Overview of large scale parallel and distributed (cloud) computing; file systems and file i/o; open source coding and distributed versioning, data query and retrieval; basic data analysis; visualization; data security, privacy and provenance.

Detailed Syllabus:
Week 1: Introduction to course; overview of data science and engineering
Week 2: Data storage: cost, performance and tradeoffs. Computational speed: CPU limited, data transfer speed limited
Week 3: Computational thinking using scripts, functions and programs
Week 4: Overview of cloud computing
Week 5: Team code development and versioning
Week 6: Data query, indexing and retrieval
Week 7: Data security, privacy and provenance
(Midterm exams)
Week 8: Introduction to Statistical Data Analysis
Week 9: Data Analysis Case Study 1
Week 10: Introduction to machine learning and data mining
Week 11: Data Analysis Case Study 2
Week 12: Data Visualization
Week 13: Overview of Databases, SQL and NoSQL
Week 14: Two guest lectures from domain experts illustrating real DSE problems and solutions

DSCI 234: Structured and Unstructured Data

Transcript Title: Struc/Unstruc Data

Credit Hours: 3

Course Pre-Requisites: DSCI 133

Course Description (up to 2100 characters):

This course is an introduction to types of data and their representation, storage, processing and analysis. The course has three parts.

In the first part of the course, students will develop a basic understanding and the ability to represent, store, process and analyze structured data. Structured data include catalogs, records, tables, logs, etc, with a fixed dimension and well-defined meaning for each data point. Suitable representation and storage mechanisms include lists and arrays. Relevant techniques include keys, hashes, stacks, queues and trees.

In the second part of the course, students will develop a basic understanding and the ability to represent, store, process and analyze semi-structured data. Semi-structured data include texts, web pages and networks, without a dimension and structure, but with well-defined meaning for each data point. Suitable representation and storage mechanisms include trees, graphs and RDF triples. Relevant techniques include XML, YAML, JSON, parsing, annotation, language processing.

In the third part of the course, students will develop a basic understanding and the ability to represent, store, process and analyze unstructured data. Unstructured data include images, video, and time series data, without neither a fixed dimension and structure, nor well-defined meaning for individual data points. Suitable representation and storage mechanisms include large matrices, EDF, DICOM. Relevant techniques include feature extraction, segmentation, clustering, rendering, indexing, and visualization.

Detailed Syllabus:
Week 1: Introduction to course; overview of data types and their lifecycle.
Week 2: Structured data and databases. Data capture, data storage, data migration, data integration: cost, performance and tradeoffs.
Week 3: Lists and arrays keys, hashes, stacks, queues.
Week 4: Lists and arrays keys, hashes, stacks, queues.
Week 5: Semi-structured data, their capture, storage, migration, and integration.
Week 6: Trees
Week 7: Graphs and RDF triples
(Midterm exams)
Week 8: XML, YAML, JSON, parsing, annotation, language processing
Week 9: Image data: format (jpeg, png, DICOM) and processing (Matlab, ImageJ libraries)
Week 10: Video: MPEG and other format, compression, processing
Week 11: Time series: EDF format, compression, processing
Week 12: Querying and searching techniques

Week 13: Exploring and visualizing data
Week 14: Project presentation

DSCI 341: Introduction to Databases: DS Major

Transcript Title: Introduction to Databases: DS Major

Credit Hours: 3

Course Pre-Requisites:

- EECS 233: Intro to Data Structures or DSCI 234.

Weeks 1-6 provide an overview of basic database systems concepts including database design, database systems architecture, and database querying, using relational model and SQL as query language.

Weeks 7-10 Objects, Semi structured data, XML and RDF basics.

Weeks 11-14 provide an overview of more advanced topics including Database System Architectures (Parallel Databases and Distributed Databases), and Data Warehousing and Information Retrieval.

Objectives:

1. The student should know the basic concepts in data bases including database design, implementation, and query languages.
2. The student should know how to use a relational database system, and be knowledgeable of other data representation schemes including XML and RDF which are becoming increasingly popular for data exchange and representation of data on the web. Given a data base application, the student should be able to design, implement and query the database.

Course Description (up to 2100 characters):

Database management become a central component of a modern computing environment, and, as a result, knowledge about database systems has become an essential part of education in computer science and data science. This course is an introduction to the nature and purpose of database systems, fundamental concepts for designing, implementing and querying a database and database architectures.

Detailed Syllabus:

Week 1: Introduction to course; overview of database systems

Week 2: Entity Relational and Relational Model

Week-3: SQL

Week 4: Relational Algebra and Calculus

Week 5: Views, Transactions, Integrity constraints

Week 6: Accessing SQL from a Programming Language, functions, procedures, triggers

Week 7-8: Object oriented databases, XML

(Midterm exam)

Week 9-10: RDF, data on the web
Week 11-12 Overview of Query Processing
Week 13: Data Warehousing
Week 14: Distributed and Parallel Databases

DSCI 342: Introduction to Data Science systems

Transcript Title: Intro Data Science Systems

Credit Hours: 3

Course Pre-Requisites: DSCI 234

Course Description and objectives (up to 2100 characters):

An introduction to the software and hardware architecture of data science systems, with an emphasis on Operating Systems and Computer Architecture that are relevant to Data Sciences systems. At the end of the course, the student should understand the principles and architecture of storage systems, file systems (especially, HDFS), memory hierarchy, and GPU. The student should have carried out projects in these areas, and should be able to critically compare various design decisions in terms of capability and performance.

Detailed Syllabus:

1. The Unix/Linux Operating System. Basic concepts, the command line interface. Sample lab assignment: install Linux Mint, create user accounts, start MySQL service. Reference: M. Garrels. *Introduction to Linux. A Hands on Guide.* http://tldp.org/LDP/intro-linux/html/. Weeks: 1.
2. The C programming language. Reference: C. Burch. *C for Python programmers.* http://www.toves.org/books/cpy/. Weeks: 2.
3. Storage architecture. HDD: architecture, performance (e.g., seek time, sequential vs random access, caching, etc), RAIDs. SDD. Logical Volumes. Sample lab: create logical volumes, use them in a software RAID, and measure performance. Reference: [HP] [AD]. Weeks: 2.
4. File Systems: files and directories, file system implementation, journaling, log-structured file systems, data integrity, distributed systems [AD]. Weeks: 2
5. The Hadoop File System (HDFS). Reference: HDFS Architecture Guide, http://hadoop.apache.org/docs/r1.2.1/hdfs_design.html. Weeks: 1.
6. The Map-Reduce engine: JobTracker and TaskTracker, scheduling. Hadoop use cases. Reference: [KM]. Weeks: 1.
7. Memory Hierarchy: design, virtual memory, address spaces, memory API, and introduction to paging. References: [HP] [AD]. Weeks: 2.
8. Data level parallelism in vector, SIMD, and GPU architectures. Reference: [HP]. Weeks: 2.
9. CUDA: introduction and parallel programming. References: [SK]. Weeks: 2.
10. GPUs: graphics pipeline, model transformation, lighting, unified shaders, rasterization, texturing, hidden surfaces. General Processing GPU. Reference: D. Luebke, G. Humphreys. How GPUs work. *Computer* **40**(2). Weeks: 1.

Main References

[AD] R. H. Arpaci-Dusseau and A. C. Arpaci-Dusseau. *Operating Systems: Three Easy Pieces*. http://pages.cs.wisc.edu/~remzi/OSTEP/

[HP] J. L. Hennessy and D. A. Patterson. Computer Architecture.

[KM] M. Kerzner and S. Maniyam. *Hadoop Illuminated.*

[SK] J. Sanders and E. Kandrot. *CUDA by Example: An Introduction to General-Purpose GPU Programming.* Addison-Wesley Professional.

DSCI 343: Introduction to Data Analysis

Transcript Title: Intro Data Analysis

Credit Hours: 3

Course Pre-Requisites: EECS 233/DSCI 234, Probability/Statistics, EECS 340

Course Description and objectives (up to 2100 characters):

In this class we will give a broad overview of data analysis techniques, covering techniques from data mining, machine learning and signal processing.
Students will also learn about probabilistic representations, how to conduct an empirical study and support empirical hypotheses through statistical tests, and visualize the results.

Course objectives:

- expose students to different analysis approaches
- understand probabilistic representations and inference mechanisms
- understand how to create empirical hypotheses and how to test them.

Detailed Syllabus:

Week 1: Data Preprocessing, Cleaning and Validation
Week 2: Frequent patterns and Association Rules
Week 3: Empirical Methodology
Week 4: Statistical Hypothesis Testing
Week 5: Correlation and Causation
Week 6: Belief Networks and Causal Networks
Week 7: Inference in Belief Networks
(Midterm exams)
Week 8: Working with Labeled Data: Classification
Week 9: Working with Labeled Data: Regression
Week 10: Interpreting and Visualizing Models
Week 11: Working with time series data: time domain methods
Week 12: Working with time series data: frequency domain methods
Week 13: Statistical signal processing: detection and classification
Week 14: Statistical signal processing: filtering and estimation
Week 15: Wrapup

DSCI 344: Scalable Parallel Data Analysis
Pre-requisites: DSCI 234, 342

Course Description: This course provides an introduction to scalable and parallel data analysis using the most common frameworks and programming tools in the age of big data. Covered topics include parallel programming models, parallel hardware architectures, multi-threaded, multi-core programming, cluster computing and GPU programming. The course is designed to provide a heavily hands-on experience with several programming assignments.

Course Textbook(s):
1. A Kaminsky. BIG CPU, BIG DATA : Solving the World's Toughest Computational Problems with Parallel Computing, Creative Commons, 2014 (freely available on the web).
2. A. Grama et al., Introduction to Parallel Computing, 2nd Edition, Wiley & Sons: 2003.
3. A. C. Telea. Data Visualization. Principles and Practice. AKPeters 2008.

Draft Syllabus:
1. Motivating parallelism, scope of parallel computing (0.5 weeks).
2. Implicit parallelism, trends in microprocessor architectures, dichotomy between computing and communication (0.5 weeks).
3. Parallel hardware architectures, physical organization of parallel platforms, historical perspective (0.5 weeks).
4. Parallel computing paradigms: Shared memory vs. message passing architectures (0.5 weeks).
5. Concurrent programming:
   a. Threads and synchronization (2 weeks)
   b. Functional constructs: immutable objects, map, reduce (1 week)
   c. Processes, IPC, and REST (1 week)
6. Tightly coupled multicore: Parallel loops, reduction, load balancing, overlapping, sequential dependencies, scaling, search algorithms (2.5 weeks).
7. Cluster computing: Massively parallel, hybrid parallel, tuple space, cluster load balancing, interacting tasks (2.5 weeks).
8. GPU programming: GPU Massively parallel, GPU parallel reduction, Multi-GPU programming (2.5 weeks).
9. Data visualization: the visualization pipeline, scalar, vector and tensor visualization. (1 week)

DSCI 345: Files, Indexes and Access Structures for Big Data

Transcript Title: Indexes for Big Data: DS Major

Credit Hours: 3

Course Pre-Requisites:

- Basic knowledge on data structures (stacks, lists, queues, trees) and algorithms (basic searching and sorting, iteration, recursion) (EECS 233/DSCI 234) and
- Basic knowledge on discrete structures (graphs, trees, sets, proof by induction) (EECS 302).

Objectives:

- An expert knowledge of basic data structures, basic searching, sorting, methods, algorithm techniques, (such as greedy and divide and conquer)
- In-depth knowledge on Search and Index Structures for large, heterogeneous data including multidimensional data, high dimensional data and data in metric spaces (e.g., sequences, images), on different search methods (e.g. similarity searching, partial match, exact match), and on dimensionality reduction techniques.

Course Description (up to 2100 characters):

Database management become a central component of a modern computing environment, and, as a result, knowledge about database systems has become an essential part of education in computer science and data science. This course is an introduction to the nature and purpose of database systems, fundamental concepts for designing, implementing and querying a database and database architectures.

Detailed Syllabus:

Week 1-2: Introduction to course; overview of basic structures, trees, tree representations, search trees, heaps, Huffman codes—a different application of trees.
Week 2-3 Multiway trees, balanced trees, static/dynamic trees, B-trees, AVL trees, R/B trees.
Week-4: Indexes on Sequential Files, Sparse and dense indexes, multiple levels of Index, secondary index, inverted lists.
Week 5: Tree Based index structures, B+ trees, B-trees comparison
Week 6: Hashing, Static Hashing, and Dynamic Hashing, Hash based Indexes
Week 7-8: Multi-dimensional Data and Indexes: Applications—Geographic Information Systems, Biological Databases. Nearest neighbor, and range queries.

(Midterm Exam)

Week 9: Hash like structures for Multidimensional Data: Grid Files, Partitioned Hashing.

Week 10: Tree like structures for Multidimensional Data: kd-tree, Quad trees, R-Trees and its variants

Week 11: Distance Based Index Structures: VP-tree, MVP-tree, GNAT

Week 12: Bitmap Indexes: motivation, compressed bitmap index.

Week 13: String / Sequence Similarity search

Week 14: Indexes for Graph databases

DSCI 390: Machine Learning for Big Data

Transcript Title: ML for Big Data

Credit Hours: 3

Course Pre-Requisites: EECS 233/DSCI 234, Probability/Statistics, DSCI 343

Course Description and objectives (up to 2100 characters):

Machine learning is a subfield of Artificial Intelligence that is concerned with the design and analysis of algorithms that "learn" and improve with experience, While the broad aim behind research in this area is to build systems that can simulate or even improve on certain aspects of human intelligence, algorithms developed in this area have become very useful in analyzing and predicting the behavior of complex systems. Machine learning algorithms have been used to guide diagnostic systems in medicine, recommend interesting products to customers in e-commerce, play games at human championship levels, and solve many other very complex problems. This course is an introduction to algorithms for machine learning and their implementation in the context of big data. We will study different learning settings, the different algorithms that have been developed for these settings, and learn about how to implement these algorithms and evaluate their behavior in practice. We will also discuss dealing with noise, missing values, scalability properties and talk about tools and libraries available for these methods.

At the end of the course, you should be able to:

--Understand when to use machine learning algorithms;

--Understand, represent and formulate the learning problem;

--Apply the appropriate algorithm(s) or tools, with an understanding of the tradeoffs involved including scalability and robustness;

--Correctly evaluate the behavior of the algorithm when solving the problem.

Detailed Syllabus:

Week 1: Review of basic machine learning concepts (recap from Intro to Data Analysis)

Week 2: Artificial Neural Networks

Week 3: Support Vector Machines

Week 4: Probabilistic Methods

Week 5: Probabilistic Methods

Week 6: Deep Architectures

Week 7: Learning from Data Streams

(Midterm exams)

Week 8: Sequential Learning algorithms

Week 9: Feature selection and dimensionality reduction

Week 10: Handling noise and missing data

Week 11: Scaling up learning algorithms

Week 12: Cost-sensitive learning, dealing with imbalanced data

Week 13: Multiclass learning and regression

Week 14: Learning from structured objects

Week 15: Wrapup

DSCI 391: Data Mining for Big Data

Transcript Title: Data Mining: DS Major

Credit Hours: 3

Course Pre-Requisites: EECS 233/DSCI 234, Probability/Statistics, DSCI 343

Weeks 1-4 provide an overview of basic data mining concepts including association rule mining, data preprocessing and matrix decompositions.

Weeks 5-8 provide an overview of commonly used data mining tools including classification and clustering.

Weeks 9-14 provide an overview of more advanced topics including high-dimensional data analysis and mining graph data.

Objectives:

1. The student should know the basic tools in data analytics including various forms of matrix decomposition, linear regression and data preprocessing.
2. The student should know commonly used data mining tools, in which problem settings to use what tools. Given a data source, the student should be able to determine relevant data mining tasks that can be applied to analyze the data, describe the expected outcome, and the evaluation criteria for the results.

Course Description (up to 2100 characters):

With the unprecedented rate at which data is being collected today in almost all fields of human endeavor, there is an emerging economic and scientific need to extract useful information from it. Data mining is the process of automatic discovery of patterns, changes, associations and anomalies in massive databases, and is a highly inter-disciplinary field representing the confluence of several disciplines, including database systems, data warehousing, machine learning, statistics, algorithms, data visualization, and high-performance computing. This course is an introduction to the commonly used data mining techniques.

In the first part of the course, students will develop a basic understanding of the basic concepts in data mining such as frequent pattern mining, association rule mining, basic techniques for data preprocessing such as normalization, regression, and classic matrix decomposition methods such as SVD, LU, and QR decompositions.

In the second part of the course, students will develop a basic understanding of classification and clustering and be able to apply classic methods such as k-means, hierarchical clustering methods, nearest neighbor methods, association based classifiers.

In the third part of the course, students will have a chance to study more advanced data mining applications such as feature selection in high-dimensional data, dimension reduction, and mining biological datasets.

Detailed Syllabus:

Week 1: Introduction to course; overview of the data mining tasks.
Week 2-3: Frequent pattern and association rules
Week 4: Data preprocessing
Week 5-6: Clustering
Week 7-8: Classification
(Midterm exams)
Week 9-10: sequential pattern mining
Week 11-12: feature selection and dimension reduction
Week 13: network/graph mining
Week 14: Project presentation

MATH/STAT APPROVAL

From: David Singer <david.singer@case.edu>
Date: June 12, 2015 8:18:47 AM EDT
To: "Kenneth A. Loparo" <kenneth.loparo@case.edu>
Subject: Re: MATH/STAT Courses

Dear Ken,
I am happy to approve the inclusion of MATH 201 and an elective in probability or statistics in the program.
Cheers,
David

On 6/12/2015 6:46 AM, Kenneth A. Loparo wrote:
Dear David: I hope that this email finds you well. As you may be aware, when Barbara was Chair of the BHEF she initiated a process to develop a new UG degree program in data sciences. The new program will include not only the MATH courses in the CSE GER, but will also include MATH 201 and a Probability/Statistics elective, much like several of the current degree programs in EECS.

I am writing to seek your approval for including these courses in the degree program description, please let me know your decision about including these courses by return email.

Thanks,

Ken

Kenneth A. Loparo
Nord Professor of Engineering and Chair
EECS Department
Case Western Reserve University
10900 Euclid Ave
Cleveland, OH 44106-7071
Phone: 216-368-4115

--
Professor and Interim Chair
Department of Mathematics, Applied Mathematics, and Statistics
Case Western Reserve University
Cleveland, OH 44106-7058
(216) 368-2892