Data Science Curricula at the University of Washington

Bill Howe, PhD
Director of Research, Scalable Data Analytics
University of Washington
eScience Institute
http://escience.washington.edu
The University of Washington eScience Institute

• Rationale
  – The exponential increase in sensors is transitioning all fields of science and engineering from data-poor to data-rich
  – Techniques and technologies include
    • Sensors and sensor networks, databases, data mining, machine learning, visualization, cluster/cloud computing
  – If these techniques and technologies are not widely available and widely practiced, UW will cease to be competitive

• Mission
  – Help position the University of Washington at the forefront of research both in modern eScience techniques and technologies, and in the fields that depend upon them

• Strategy
  – Bootstrap a cadre of Research Scientists
  – Add faculty in key fields
  – Build out a “consultancy” of students and non-research staff
eScience Big Data Group

Bill Howe, Phd (databases, cloud, data-intensive scalable computing, visualization)

Director of Research, Scalable Data Analytics

Staff

– Seung-Hee Bae, Phd (postdoc, scalable machine learning algorithms)
– Dan Halperin, Phd (postdoc; scalable systems)
– Sagar Chitnis, Research Engineer (Azure, databases, web services)
– (alumna) Marianne Shaw, Phd (hadoop, semantic graph databases)
– (alumna) Alicia Key, Research Engineer (visualization, web applications)

Students

– Scott Moe (2nd yr Phd, Applied Math)
– Daniel Perry (2nd yr Phd, HCDE)

Partners

– CSE DB Faculty: Magda Balazinska, Dan Suciu
– CSE students: Paris Koutris, Prasang Upadhyaya,
– UW-IT (web applications, QA/support)
– Cecilia Aragon, Phd, Associate Professor, HCDE (visualization, scientific applications)
eScience ≈ Data Science

...modulo application area
Data Scientist Job

Bill & Melinda Gates Foundation

Seattle, WA

Sep-21-2012
# UW Data Science Education Efforts

<table>
<thead>
<tr>
<th></th>
<th>CS/Informatics</th>
<th>Students</th>
<th>Non-Major</th>
<th>Non-Students</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>undergrads</td>
<td>undergrads</td>
<td>grads</td>
<td>professionals</td>
</tr>
<tr>
<td>UWEO Data Science Certificate</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Graduate Certificate in Big Data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CS Data Management Courses</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eScience workshops</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intro to data programming</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eScience Masters (planned)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coursera Course: Intro to Data Science</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Previous courses:*
Scientific Data Management, Graduate CS, Summer 2006, Portland State University
Scientific Data Management, Graduate CS, Spring 2010, University of Washington
Huge number of relevant courses, new and existing.
“I worry that the Data Scientist role is like the mythical “webmaster” of the 90s: master of all trades.”

-- Aaron Kimball, CTO Wibidata
Breadth

tools

Hadoop
PostgreSQL
glm(…) in R
Tableau

abstractions

MapReduce
Relational Algebra
Logistic Regression
InfoVis
Depth

structures

Management
Relational Algebra
Standards

statistics

Analysis
Linear Algebra
ad hoc files
Scale

desktop

main memory
R
local files

cloud

distributed
Hadoop
S3, Azure Storage
Target

hackers

Assume proficiency in Python, Java, R

analysts

Assume little or no programming
Breadth
- tools
- abstr.

Depth
- structs
- stats

Scale
- desk
- cloud

Target
- hackers
- analysts
Certificate in Statistical Analysis with R Programming

Approved by the UW Department of Statistics and UW Department of Applied Mathematics

Certificates » Statistical Analysis with R Programming

Develop your statistical and analytical skills in the R programming environment. Master and apply a comprehensive range of statistical analyses and models, including linear regressions, multivariate analysis, machine learning algorithms and time series analysis. Learn and apply state of the art skills in data mining and big data management to derive meaning from raw data. Acquire a thorough understanding of the R programming source environment, and learn to maximize the visualization and graphical capabilities within R, including ggplot and lattice graphics. Use your skills in statistics and R to solve complex problems in such fields as finance, marketing, social media and genomics.

Program Features:
- Flexibility to take courses in both
- Virtual interaction with instructor in real time via (online program)
- Hands-on programming for analytics
- Instruction and real-life modeling technique from

Who Should Apply:
tools
abstr.
structs
stats
desk
cloud
hackers
analysts
Syllabus for Machine Learning with Large Datasets 10-605

This is the syllabus for Machine Learning with Large Datasets 10-605 in Spring 2012.

Contents [hide]
1 January
2 February
3 March
4 April
5 May

January
- Tues Jan 17. Overview of course, cost of various operations, asymptotic analysis.
- Thus Jan 19. Review of probabilities.
- Tues Jan 24. Streaming algorithms and Naive Bayes.
  - New Assignment: streaming Naive Bayes 1 (with feature counts in memory). PDF Handout
  - Thus Jan 26. The stream-and-sort design pattern; Naive Bayes revisited.
  - Tues Jan 31. Messages and records 1; Phrase finding.
- Assignment due: streaming Naive Bayes 1 (with feature counts in memory).
- New Assignment: streaming Naive Bayes 2 (with feature counts on disk) with stream-and-sort. PDF Handout
CSE 344: Daily Schedule

Note that this schedule will be altered during the quarter. Please make sure to check it every week.

You can find the consolidated list of readings at the end of the page.

<table>
<thead>
<tr>
<th>Week of</th>
<th>Monday</th>
<th>Wednesday</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sep 24</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Introduction lecture 1</strong></td>
<td><strong>Data Models lecture 2</strong></td>
</tr>
<tr>
<td>Oct 1</td>
<td><strong>Aggregates in SQL lectures 4 and 5 (data)</strong></td>
<td><strong>Aggregates in SQL lectures 4 and 5 (data)</strong></td>
</tr>
<tr>
<td></td>
<td>Homework 1 due</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Webquiz for lectures 1-3 due</td>
<td></td>
</tr>
</tbody>
</table>
## Schedule

The course will consist of five components: data preparation, data presentation, data products, observation, and experimentation. There will be up to 6 guest speakers throughout the course; lecture dates may change based on guest speaker availability.

### Regular Lectures

<table>
<thead>
<tr>
<th>Component</th>
<th>Weeks</th>
<th>Class Dates</th>
<th>Lecture Materials</th>
<th>Homework</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>1</td>
<td>1/17-1/19</td>
<td>Lecture 1 (slides, video) Lecture 2 (slides, video)</td>
<td></td>
</tr>
<tr>
<td>Data Preparation</td>
<td>2 – 4</td>
<td>1/24-2/9</td>
<td>Lecture 3 (slides, video) Lecture 4 (slides, video) Lecture 5 (slides, video) Lecture 6 (slides) Lecture 7 (slides) Lecture 8 (slides)</td>
<td>Assignment 1</td>
</tr>
<tr>
<td>Data Presentation</td>
<td>5 – 6</td>
<td>2/14-2/23</td>
<td>Lecture 9 (slides) Lecture 10 (slides) Lecture 13 (slides)</td>
<td>Assignment 2</td>
</tr>
<tr>
<td>Spring Break</td>
<td>11</td>
<td>3/27 – 3/29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observation</td>
<td>12 – 13</td>
<td>4/3 – 4/12</td>
<td>Lecture 20 (slides) Lecture 21 (slides)</td>
<td></td>
</tr>
<tr>
<td>Experimentation</td>
<td>14</td>
<td>4/17-4/19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Final Project</td>
<td>15</td>
<td>4/24 – 4/26</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Fall 2012 Statistics W4242 section 001
INTRODUCTION TO DATA SCIENCE

Call Number  61780
Day & Time Location
MW 6:10pm-7:25pm
313 Fayerweather
Day & Time Location
MW 7:40pm-8:55pm
313 Fayerweather
Points  3
Approvals Required  None
Instructor  Rachel R Schutt
Type  LECTURE
Course Description
This course is an introduction to the interdisciplinary and emerging field of data science, which lies at the intersection of statistics, computer science, and the social sciences. The course will be organized around three central threads: (1) statistical modeling and machine learning, (2) data pipelines, processing "big data" tools, and (3) real world topics and case studies. Correspondingly there will be (1) core lectures, (2) labs and (3) guest lectures from researchers who are experts in their fields. Topics and tools will include logistic regression, predictive modeling, clustering algorithms, decision trees, Hadoop visualization, data journalism, R, python, javascript.

NOTE: Course information changes frequently. Please re-visit these pages periodically for the most recent and up-to-date information.
DATA SCIENCE AND BIG DATA ANALYTICS
An ‘open’ course to unleash the power of Big Data

“We live in a data-driven world. Increasingly, the efficient operation of organizations across sectors relies on the effective use of vast amounts of data. Making sense of big data is a combination of organizations having the tools, skills and more importantly, the mindset to see data as the new "oil" fueling a company. Unfortunately, the technology has evolved faster than the workforce skills to make sense of it and organizations across sectors must adapt to this new reality or perish.”

- Andreas Weigend, Ph.D Stanford, Head of the Social Data Lab at Stanford, former Chief Scientist, Amazon.com

DATA SCIENCE AND BIG DATA ANALYTICS COURSE
An ‘open’ course and certification focused on concepts and principles applicable to any technology environment and industry.

This course is intended for:
- Business and data analysts looking to add big data analytics skills
- Managers of business intelligence, analytics, or big data groups
- Database professionals looking to enrich their analytic skills
- College graduates considering data science as a career field

The course provides a hands-on practitioner’s approach to the techniques and tools required for analyzing Big Data.
Certificate in Data Science

Approved by the UW Department of Computer Science & Engineering.

Course Description

Introduction to Data Science

Bellevue, Classroom, Winter 2013

Instructor: Ernst Henle

Th, 1/10 - 3/14, 2013, 6-9 p.m.

Cost: $999 | 3 CEUs

This course is designed to introduce students to the data management, storage and manipulation tools common in data science and will apply those tools to real scenarios. An overview of different SQL and No-SQL database technologies is presented and the course finishes with a discussion of choosing the appropriate tool to get the job done.

Topics include:

Bill Howe, UW
Course Dashboard

Users

Total Registered Users 17834

easy to obtain through conventional curricula. Introduce yourself to the basics of data science and leave armed with practical experience programming massive databases.

Next session: April 2013 (10 weeks long)
Statistics, Data Analysis, and Scientific Computing
tools  abstr.
structs  stats
desk  cloud
hackers  analysts
What goes around comes around

- 2004 Dean et al. MapReduce
- 2008 Hadoop 0.17 release
- 2008 Olston et al. Pig: Relational Algebra on Hadoop
- 2008 DryadLINQ: Relational Algebra in a Hadoop-like system
- 2009 Thusoo et al. HIVE: SQL on Hadoop
- 2009 Hbase: Indexing for Hadoop
- 2010 Dietrich et al. Schemas and Indexing for Hadoop
- 2012 Transactions in HBase (plus VoltDB, other NewSQL systems)

But also some permanent contributions:
  - Fault tolerance
  - Schema-on-Read
  - User-defined functions that don’t suck
What are the *abstractions* of data science?

“Data Jujitsu”
“Data Wrangling”
“Data Munging”

*Translation*: “We have no idea what this is all about”
What are the abstractions of data science?

matrices and linear algebra?
relations and relational algebra?
objects and methods?
files and scripts?
data frames and functions?
“80% of analytics is sums and averages”
-- Aaron Kimball, wibidata

God created the integers; all else is the work of man

Codd created relations; all else is the work of man
Three types of tasks:

1) Preparing to run a model
   "80% of the work"
   -- Aaron Kimball
   Gathering, cleaning, integrating, restructuring, transforming, loading, filtering, deleting, combining, merging, verifying, extracting, shaping, massaging

2) Running the model

3) Interpreting the results
   The other 80% of the work
Problem

How much time do you spend “handling data” as opposed to “doing science”?

Mode answer: “90%”
• Databases and Statistical Packages
  • Many analysts download data to use in Excel/SAS/Matlab/R or their favorite programming language?
    FORTRAN??
  • Use matrix/vector operations
  • Most of these stat packages require data to fit in RAM
    • Taking samples from the full data to fit into ram results in loss of precision
  • External toolkits may also lack parallelism

src: Christian Grant, MADSkills
(Sparse) Matrix Multiply in SQL

```
SELECT A.row_number, B.column_number, SUM(A.value * B.value)
FROM A, B
WHERE A.column_number = B.row_number
GROUP BY A.row_number, B.column_number
```

src: Christian Grant, MADSkills
Aside: Schema-on-Write vs. Schema-on-Read

- A schema* is a shared consensus about some universe of discourse
- At the frontier of research, this shared consensus does not exist, by definition
- Any schema that does emerge will change frequently, by definition
- Data found “in the wild” will typically not conform to any schema, by definition
- But this doesn’t mean we have to live with ad hoc scripts and files
- My answer: Schema-later, “lazy schemification”

* ontology/metadata standard/controlled vocabulary/etc.
Data Access Hitting a Wall

Current practice based on data download (FTP/GREP)
Will not scale to the datasets of tomorrow

- You can GREP 1 MB in a second
- You can GREP 1 GB in a minute
- You can GREP 1 TB in 2 days
- You can GREP 1 PB in 3 years.

- You can FTP 1 MB in 1 sec
- You can FTP 1 GB / min (~1$)
- … 2 days and 1K$
- … 3 years and 1M$

- Oh!, and 1PB ~5,000 disks

- At some point you need **indices** to limit search **parallel** data search and analysis
- This is where databases can help

[slide src: Jim Gray]
US faces shortage of 140,000 to 190,000 people “with deep analytical skills, as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions.”

--Mckinsey Global Institute
Biologists are beginning to write very complex queries (rather than relying on staff programmers)

**Example: Computing the overlaps of two sets of blast results**

```sql
SELECT x.strain, x.chr, x.region as snp_region, x.start_bp as snp_start_bp,
      x.end_bp as snp_end_bp, w.start_bp as nc_start_bp, w.end_bp as nc_end_bp,
      w.category as nc_category,
      CASE WHEN (x.start_bp >= w.start_bp AND x.end_bp <= w.end_bp)
            THEN x.end_bp - x.start_bp + 1
      WHEN (x.start_bp <= w.start_bp AND w.start_bp <= x.end_bp)
            THEN x.end_bp - w.start_bp + 1
      WHEN (x.start_bp <= w.end_bp AND w.end_bp <= x.end_bp)
            THEN w.end_bp - x.start_bp + 1
      END AS len_overlap
FROM [koesterj@washington.edu].[hotspots_deserts.tab] x
INNER JOIN [koesterj@washington.edu].[table_noncoding_positions.tab] w
ON x.chr = w.chr
WHERE (x.start_bp >= w.start_bp AND x.end_bp <= w.end_bp)
   OR (x.start_bp <= w.start_bp AND w.start_bp <= x.end_bp)
   OR (x.start_bp <= w.end_bp AND w.end_bp <= x.end_bp)
ORDER BY x.strain, x.chr ASC, x.start_bp ASC
```

We see thousands of queries written by non-programmers
# UW Curricular Activities

<table>
<thead>
<tr>
<th></th>
<th>CS/Informatics</th>
<th>Students</th>
<th>Non-Major</th>
<th>Non-Students</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>undergrads</td>
<td>grads</td>
<td>undergrads</td>
<td>professionals</td>
</tr>
<tr>
<td>UWEO Data Science Certificate</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Graduate Certificate in Big Data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Database Courses</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eScience workshops</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intro to data programming</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>eScience Masters (planned)</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coursera Course Intro to Data Science</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
How do you deliver hands-on big data experience to 10k students?

Cloud vendors’ free tiers?
- 1 micro instance is not “big data”

Cloud vendors’ academic discounts? (e.g., Amazon’s education grants)
- $100 / head = $1M
- invites abuse: free credits with no obligation to complete course

Out of pocket?
- Don’t want to be the only non-free Coursera course
- Unclear that we can require it (perhaps analogous to a textbook?)
10k Students on 10k GB for $10k

• Requirements
  – Inexpensive: Need a fixed, small budget; O(10k) maximum
  – Fair: All students need to be able to complete the assignment

• Non-solutions
  – Fixed cluster
    • Fairness problems; no quality of service guarantees
  – Autoscaling cluster
    • No upper bound to cost
  – Budget cap (via, e.g., Amazon’s IAM)
    • Fairness problems: Different students consume different levels of resources depending on background, etc.
10k Students on 10k GB for $10k

• Key idea: 10k students all working on the same assignment = lots of redundancy
• Students debug locally on scaled down datasets
• Then submit 10k jobs
• Prune the queue aggressively
  – Remove duplicates
  – Detect typical mistakes syntactically; return cached results
  – Global common subexpression elimination (feasible thanks to abstractions)
10k Students on 10k GB for $10k

• Another approach we are considering: Kaggle-Kaggle-style Prize assignments

<table>
<thead>
<tr>
<th>student</th>
<th>date</th>
<th>runtime</th>
<th>output</th>
<th>quality</th>
<th>notes</th>
<th>votes</th>
</tr>
</thead>
<tbody>
<tr>
<td>sarah123</td>
<td>4/23/12</td>
<td>5 min 44 sec</td>
<td>result.txt</td>
<td>45%</td>
<td>I removed the dirty data for this run</td>
<td>456</td>
</tr>
<tr>
<td>jane456</td>
<td>4/22/12</td>
<td>3 min 23 sec</td>
<td>result.txt</td>
<td>23%</td>
<td>I used gradient descent this time</td>
<td>97</td>
</tr>
</tbody>
</table>

• Pay-to-play: Students submit successful jobs to access leaderboards
• Cast votes for their preferred solution.
• Grade determined by
  \[ f(\text{votes(your\_solution)}, \text{grade(your\_solution)}, \text{grade(solutions\_you\_voted\_for)}) \]
• We run the top-k highest ranked solutions on the full size dataset
• Problem: Easy to game the system and just coast
Coursera course:
https://www.coursera.org/course/datasci

Certificate program:
http://www.pce.uw.edu/courses/data-science-intro

http://escience.washington.edu

billhowe@cs.washington.edu
Science is becoming a database query problem

**Old model:** “Query the world” *(Data acquisition coupled to a specific hypothesis)*

**New model:** “Download the world” *(Data acquisition supports many hypotheses)*

- Astronomy: High-resolution, high-frequency sky surveys (SDSS, LSST, PanSTARRS)
- Biology: lab automation, high-throughput sequencing,
- Oceanography: high-resolution models, cheap sensors, satellites

40TB / 2 nights
1 device

~1TB / day
100s of devices
• Power distribution
• 80:20 rule

First published May 2007, Wired Magazine article 2004

[src: Carol Goble]
A “Needs Hierarchy” of Science Data Management

“As each need is satisfied, the next higher level in the hierarchy dominates conscious functioning.”

-- Maslow 43