**Introduction to Data Science**  
TBD  
Bill Howe   
Director of Research, Scalable Data Analytics, UW eScience Institute   
Affiliate Assistant Professor, Computer Science & Engineering, University of Washington   
PhD, Portland State University

Introduction to Data Science  
  
Join the data revolution. Companies are searching for data scientists.  
This specialized field demands multiple skills not easy to obtain  
through conventional curricula. Introduce yourself to the basics of  
data science and leave armed with practical experience programming  
massive databases.  
  
  
About this Course:  
  
Commerce and research is being transformed by data-driven discovery  
and prediction. Skills required for data analytics at massive levels –  
scalable data management on and off the cloud, parallel algorithms,  
statistical modeling, and proficiency with a complex ecosystem of  
tools and platforms – span a variety of disciplines and are not easy  
to obtain through conventional curricula.  Tour the basic techniques  
of data science, including both SQL and NoSQL solutions for massive  
data management (e.g., MapReduce and contemporaries), algorithms for  
data mining (e.g., clustering and association rule mining), and basic  
statistical modeling (e.g., logistic and non-linear regression).  
  
  
Recommended Background:  
  
You will need basic programming experience with Java or Python, and  
some familiarity with databases.  The target audience is undergraduate  
students across disciplines who wish to build proficiency working with  
large datasets to perform predictive analytics.  
  
Specific Topics:  
  
\* Data models, schemas, database design  
\* Relational algebra and parallel query processing  
\* NoSQL systems, key-value stores  
\* Tradeoffs of SQL, NoSQL, and NewSQL systems  
\* Algorithm design in Hadoop (and MapReduce in general)  
\* Basic statistical analysis at scale: sampling, regression  
\* Introduction to data mining: clustering, association rules, decision trees  
\* Case studies in analytics: social networking, bioinformatics, text processing  
  
  
Course Resources:  
There will be selected readings each week.  Students may also find the  
following textbooks relevant for further reading.  
Mining of Massive Datasets ([http://i.stanford.edu/~ullman/mmds.html](http://i.stanford.edu/~ullman/mmds.html" \t "_blank))  
Professional NoSQL  
([http://www.amazon.com/Professional-NoSQL-Wrox-Programmer/dp/047094224X](http://www.amazon.com/Professional-NoSQL-Wrox-Programmer/dp/047094224X" \t "_blank))  
  
  
Course Format:  
There will be a quiz and a programming assignment each week, as well  
as two exams.  The assignments involving large datasets will be  
completed using Amazon Web Services or Microsoft Azure cloud services.

Instructor Bio

Bill Howe is the Director of Research for Scalable Data Analytics at the UW eScience Institute and holds an Affiliate Assistant Professor appointment in Computer Science & Engineering, where he leads a group studying data management, analytics, and visualization systems for science applications. Howe has received awards from Microsoft Research and honors for papers in scientific data management, and serves on a number of program committees, organizing committees, and advisory boards in the area, including the advisory board of the Data Science certificate program at UW. He holds a Ph.D. in Computer Science from Portland State Universityand a Bachelor's degree in Industrial & Systems Engineering from Georgia Tech.