What’s This Course About?

“Aimed at non-CS undergraduate and graduate students who want to learn the basics of big data tools and techniques and apply that knowledge in their areas of study. Many of the world's biggest discoveries and decisions in science, technology, business, medicine, politics, and society as a whole, are now being made on the basis of analyzing massive data sets. At the same time, it is surprisingly easy to make errors or come to false conclusions from data analysis alone. This course provides a broad and practical introduction to big data: data analysis techniques including databases, data mining, and machine learning; data analysis tools including spreadsheets, relational databases and SQL, Python, and R; data visualization techniques and tools; pitfalls in data collection and analysis; historical context, privacy, and other ethical issues. Tools and techniques are hands-on but at a cursory level, providing a basis for future exploration and application. Prerequisites: comfort with basic logic and mathematical concepts, along with high school AP computer science, CS106A, or other equivalent programming experience.”
What’s This Course About?

“Aimed at non-CS undergraduate and graduate students who want to learn the basics of big data tools and techniques and apply that knowledge in their areas of study. Many of the world's biggest discoveries and decisions in science, technology, business, medicine, politics, and society as a whole, are now being made on the basis of analyzing massive data sets. At the same time, it is surprisingly easy to make errors or come to false conclusions from data analysis alone. This course provides a broad and practical introduction to big data: data analysis techniques including databases, data mining, and machine learning; data analysis tools including spreadsheets, relational databases and SQL, Python, and R; data visualization techniques and tools; pitfalls in data collection and analysis; historical context, privacy, and other ethical issues. Tools and techniques are hands-on but at a cursory level, providing a basis for future exploration and application. Prerequisites: comfort with basic logic and mathematical concepts, along with high school AP computer science, CS106A, or other equivalent programming experience.”
“Aimed at non-CS undergraduate and graduate students who want to learn the basics of big data tools and techniques and apply that knowledge in their areas of study. Many of the world's biggest discoveries and decisions in science, technology, business, medicine, politics, and society as a whole, are now being made on the basis of analyzing massive data sets. At the same time, it is surprisingly easy to make errors or come to false conclusions from data analysis alone. This course provides a broad and practical introduction to big data: data analysis techniques including databases, data mining, and machine learning; data analysis tools including spreadsheets, relational databases and SQL, Python, and R; data visualization techniques and tools; pitfalls in data collection and analysis; historical context, privacy, and other ethical issues. **Tools and techniques are hands-on but at a cursory level, providing a basis for future exploration and application.** Prerequisites: comfort with basic logic and mathematical concepts, along with high school AP computer science, CS106A, or other equivalent programming experience.”
Who Should Take It?

“Aimed at non-CS undergraduate and graduate students who want to learn the basics of big data tools and techniques and apply that knowledge in their areas of study. Many of the world's biggest discoveries and decisions in science, technology, business, medicine, politics, and society as a whole, are now being made on the basis of analyzing massive data sets. At the same time, it is surprisingly easy to make errors or come to false conclusions from data analysis alone. This course provides a broad and practical introduction to big data: data analysis techniques including databases, data mining, and machine learning; data analysis tools including spreadsheets, relational databases and SQL, Python, and R; data visualization techniques and tools; pitfalls in data collection and analysis; historical context, privacy, and other ethical issues. Tools and techniques are hands-on but at a cursory level, providing a basis for future exploration and application. Prerequisites: comfort with basic logic and mathematical concepts, along with high school AP computer science, CS106A, or other equivalent programming experience.”
Who Should Take It?

“Aimed at non-CS undergraduate and graduate students who want to learn the basics of big data tools and techniques and apply that knowledge in their areas of study. Many of the world's biggest discoveries and decisions in science, technology, business, medicine, politics, and society as a whole, are now being made on the basis of analyzing massive data sets. At the same time, it is surprisingly easy to make errors or come to false conclusions from data analysis alone. This course provides a broad and practical introduction to big data: data analysis techniques including databases, data mining, and machine learning; data analysis tools including spreadsheets, relational databases and SQL, Python, and R; data visualization techniques and tools; pitfalls in data collection and analysis; historical context, privacy, and other ethical issues. Tools and techniques are hands-on but at a cursory level, providing a basis for future exploration and application. Prerequisites: comfort with basic logic and mathematical concepts, along with high school AP computer science, CS106A, or other equivalent programming experience.”
Who Should Take It?

“Aimed at non-CS undergraduate and graduate students who want to learn the basics of big data tools and techniques and apply that knowledge in their areas of study. Many of the world's biggest discoveries and decisions in science, technology, business, medicine, politics, and society as a whole, are now being made on the basis of analyzing massive data sets. At the same time, it is surprisingly easy to make errors or come to false conclusions from data analysis alone. This course provides a broad and practical introduction to big data: data analysis techniques including databases, data mining, and machine learning; data analysis tools including spreadsheets, relational databases and SQL, Python, and R; data visualization techniques and tools; pitfalls in data collection and analysis; historical context, privacy, and other ethical issues. Tools and techniques are hands-on but at a cursory level, providing a basis for future exploration and application. Prerequisites: comfort with basic logic and mathematical concepts, along with high school AP computer science, CS106A, or other equivalent programming experience.”
Who Shouldn’t Take It?

Computer Science or MCS students
(except by petition)

If you’re in the wrong place, it’s okay to leave now 😬
Course Staff

Instructor
Jennifer Widom

Course Assistants
Steven Chen
Alex Haigh
Arjun Kunna
Jesse Min
Lucy Wang
## History of the Course

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fall 2015</td>
<td>Freshman seminar</td>
</tr>
<tr>
<td>Spring 2016</td>
<td>First offering of CS102</td>
</tr>
<tr>
<td>2016-17</td>
<td>Basis for sabbatical “instructional odyssey”- 30+ institutions in 18 countries</td>
</tr>
<tr>
<td>Spring 2017</td>
<td>Second offering of CS102, by graduate students</td>
</tr>
<tr>
<td>Fall 2017</td>
<td>First offering as Dean, had fun!</td>
</tr>
</tbody>
</table>
Who’s Taking It - Spring 2018

Undergraduates, Masters, MBA, JD, PhD, DCI

Biochemistry
Bioengineering
Biomedical Informatics
Business Administration
Chemical Engineering
Chemistry
Civil & Environmental Engg
Classics
Communication
Community Health
Earth Systems
Economics
Education
Electrical Engineering
Energy Resource Engineering
English
Environment & Resources
Epidemiology

Geological & Env Science
History
Human Biology
International Policy Studies
International Relations
Law
Management
Materials Science & Engineering
Mathematics
Management Science & Engg
Mechanical Engineering
Public Policy
Science, Technology, & Society
Sociology
Symbolic Systems
Urban Studies
Undeclared
Who’s Taking It

Stacked Bar Chart for Major and Level

Program and Plan

Level
Graduate
Professional
Undergraduate

# of Students

CS102
Who’s Taking It
Who’s Taking It
Who’s Taking It
## Assigned Work

<table>
<thead>
<tr>
<th>Assignment/Project</th>
<th>Assigned</th>
<th>Due</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Assignment #1</strong></td>
<td>April 9</td>
<td>April 16</td>
</tr>
<tr>
<td>Spreadsheets for Data Analysis and Visualization</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Project #1</strong></td>
<td>April 9</td>
<td>April 23</td>
</tr>
<tr>
<td>Personal Data Analysis</td>
<td></td>
<td>May 14</td>
</tr>
<tr>
<td><strong>Assignment #2</strong></td>
<td>April 16</td>
<td>April 26</td>
</tr>
<tr>
<td>Data Visualization Using Tableau, SQL</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Assignment #3</strong></td>
<td>April 26</td>
<td>May 7</td>
</tr>
<tr>
<td>Python for Data Analysis and Visualization</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Assignment #4</strong></td>
<td>May 14</td>
<td>May 24</td>
</tr>
<tr>
<td>Machine Learning</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Project #2</strong></td>
<td>May 14</td>
<td>May 31</td>
</tr>
<tr>
<td>Movie-Rating Predictions</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Assignment #5</strong></td>
<td>May 24</td>
<td>June 6</td>
</tr>
<tr>
<td>Data Mining, R Language, Network Analysis</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Exams

<table>
<thead>
<tr>
<th>Exam</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Midterm exam In class</td>
<td>May 10</td>
</tr>
<tr>
<td>Final exam 12:15-3:15 PM (2 hours)</td>
<td>June 8</td>
</tr>
</tbody>
</table>
Logistics

- Units - 4 for undergraduates, 3-4 for graduates
- WAYS requirement - Applied Quantitative Reasoning (WAY-AQR)
- Textbook? No  Readings? Recommended
- Class attendance - Expected
  - Hand-on activities
  - Only cursory notes
  - All class material game for exams
Logistics

- Grade weighting - 1/3 each assignments, projects, exams
- Graded on a curve? Not really
- Late policy - 10%/30% for 24/48 hours late, four free late days
Office Hours

TA office hours
• 20 hours per week
• Times and locations can vary

Prof. Widom office hours
• Wednesdays 4:00-5:30 PM
• Huang building 2nd floor Dean’s Office #227

Always check the course calendar!
Online

Website - http://cs102.stanford.edu

Piazza
  • Announcements
  • Q&A (private and public)
  • Discussion

Gradescope - Assignment submission & grading
For Thursday’s Class

1) Get set up on Google Drive if you’re not already

2) Download Europe city temperatures data from course website (two files)

3) Copy data files into Google Drive, make sure you can open with Google Sheets

4) Bring laptop to class (or share)
CS 102: Big Data
Tools and Techniques
Discoveries and Pitfalls

Questions?