CS 102: Big Data Tools and Techniques Discoveries and Pitfalls

Fall 2017
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.”
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.”
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/ICME students
(except by petition)

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

Instructor
Jennifer Widom

Course Assistants
Joanne Jang
Jen Kilpatrick
Clara Meister
Kelly Shen

Not intentional
## 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”- 25+ institutions in 16 countries</td>
</tr>
<tr>
<td>Spring 2017</td>
<td>Second offering of CS102, by graduate students</td>
</tr>
<tr>
<td>Fall 2017</td>
<td>Now</td>
</tr>
</tbody>
</table>
Who’s Taking It - Fall 2017

Undergraduates, Masters, MBA, JD, MD, PhD, DCl

Aeronautics-Astronautics
African/Afro-Amer Studies
Bioengineering
Business Administration
Chemistry
Civil & Env Engineering
Communication
Community Health
East Asian Studies
Economics
Education
Electrical Engineering
Energy Resource Engineering
English
Environment & Resources
Environmental Systems Engg
Epidemiology

Geological Sciences
Geophysics
History
International Relations
Law
Management
Materials Science & Engg
Mathematics
Management Science & Engg
Microbiology & Immunology
Modern Thought & Lit
Product Design
Psychology
Public Policy
Science Tech & Society
Sociology
Undeclared
Who’s Taking It

[Bar chart showing the number of students by major and status (Graduate, Professional, Undergraduate).]
Who’s Taking It
Who’s Taking It

- Prof 2
- Senior
- Grad Year 2
- Grad Year 1
- Sophomore
- Grad Year 3
- Grad Year 4+
- Freshman
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> &lt;br&gt;Spreadsheets for Data Analysis and Visualization</td>
<td>Oct 2</td>
<td>Oct 8</td>
</tr>
<tr>
<td><strong>Project #1</strong> &lt;br&gt;Personal Data Analysis</td>
<td>Oct 2</td>
<td>Oct 17 Nov 5</td>
</tr>
<tr>
<td><strong>Assignment #2</strong> &lt;br&gt;Data Visualization Using Tableau, Basic SQL</td>
<td>Oct 9</td>
<td>Oct 15</td>
</tr>
<tr>
<td><strong>Assignment #3</strong> &lt;br&gt;Python for Data Analysis and Visualization</td>
<td>Oct 16</td>
<td>Oct 22</td>
</tr>
<tr>
<td><strong>Project #2</strong> &lt;br&gt;Movie-Rating Predictions</td>
<td>Nov 6</td>
<td>Nov 28</td>
</tr>
<tr>
<td><strong>Assignment #4</strong> &lt;br&gt;Machine Learning, Advanced SQL</td>
<td>Nov 8</td>
<td>Nov 14</td>
</tr>
<tr>
<td><strong>Assignment #5</strong> &lt;br&gt;Data Mining, R Language, Social-Network Analysis</td>
<td>Nov 29</td>
<td>Dec 5/7</td>
</tr>
</tbody>
</table>
# Exams

<table>
<thead>
<tr>
<th>Exam</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Midterm exam</td>
<td>Oct 31</td>
</tr>
<tr>
<td>In class</td>
<td></td>
</tr>
<tr>
<td>Final exam</td>
<td>Dec 11</td>
</tr>
<tr>
<td>At assigned time but not 3 hours</td>
<td></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
- Grade weighting - 1/3 each assignments, projects, exams
- Graded on a curve? Somewhat
- Late policy - 10%/30% for 24/48 hours late, four free late days

Hands-on activities
Only cursory notes
Office Hours

Working office hours every evening Sunday-Thursday
- 7:00-9:30 PM in Huang basement (starting Oct 2)
- Staffed by course assistants, look for “CS102” sign
- Good time & place to work on assignments!

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

May add some afternoon hours
This week 4:15-5:30
Online

Website - http://cs102.stanford.edu

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

TBD - 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?