Course Syllabus for SIADS 501: Being a Data Scientist

Course Overview and Prerequisites

This course introduces students to the process of data science, covering problem formulation, data acquisition, modeling and analysis, and presentation and integration into action. Students will be tasked with understanding what data scientists do, and reflecting on what special knowledge and skills, perspectives, and ethical commitments they want to bring to problems as data scientists. Students will also be exposed, through interviews with practicing data scientists, to real problems they may also have to work around or avoid, so it lightly foreshadows the rest of the program and students' future in data science.

There are no course prerequisites.

Instructor and Course Assistants

Instructor: Anthony Whyte - arwhyte@umich.edu
Course Assistant: Melissa Chalmers - mechalms@umich.edu
Graduate Student Instructor: Jaleesa Turner - jaleesa@umich.edu

Communication Expectations

Contacting instructor and course assistants: Direct message via Slack
Email response time: N/A (please communicate via Slack)
Slack response time: within 24 hours
Office hours: **see Course Schedule below**

Required Textbook (if relevant)

None

Technology Requirements unique to this course

None

Accessibility

Screen reader configuration for Jupyter Notebook Content

Learning Outcomes

1. Competency - Explain the four project stages as a framework for data science problems and solutions, including the goals and desired outcomes of each stage.
2. Literacy - Describe the expertise, perspectives, and ethical commitments that data scientists may bring to each of the four stages.
3. Literacy - Articulate a set of maxims that apply to each of the four stages and to data science projects as a whole.
4. Competency - Create and maintain an environmental monitoring system for staying up to date on new developments in data science.

Course Schedule

This session **begins on Wednesday, January 8, 2020** and **ends on Tuesday, February 4, 2020.**

Weekly assignments will be **due on Tuesdays at 11:59 pm** (Ann Arbor, Eastern Time Zone, GMT-5).

Schedule of Weekly Office Hours via Zoom (Ann Arbor, Eastern time):

- **Mondays, 11 am - 12 PM; Thursdays, 8-9 PM; Saturdays, 12 - 1 PM; Sundays 11 AM-12 PM**
- Begins on Thursday, January 9

Grading

Course Assignments	**Percentage of Final Grade**
Reading Response (drop 3 lowest of 23)	10%
Initial Draft of Plan Components: Application in Domain of Interest; Plan for Knowledge Acquisition (4 submissions); Maxims, Questions, and Ethical Commitments (4 submissions); Personal Project Idea; Sources for Data Science News	9%
Peer Feedback (3 submissions)	6%
Informational Interview: Planning	2%
Informational Interview: Reflection	8%
Final Plan for Being a Data Scientist:	65%
	100%

NOTE: All assignments are required to earn credit for this course.

Letter Grades, Course Grades, and Late Submission Policy

Refer to the MADS Assignment Submission and Grading Policies section of the UMSI Student Handbook (access to Student Orientation course required).

Our policy on grades for late submissions is a little simpler: 15% penalty per day.

Academic Integrity / Code of Conduct

Refer to the Academic and Professional Integrity section of the UMSI Student Handbook (access to Student Orientation course required).

Accommodations

Refer to the Accommodations for Students with Disabilities section of the UMSI Student Handbook (access to the Student Orientation course required).

Use the Student Intake Form to begin the process of working with the University's Office of Services for Students with Disabilities.

Help Desk(s): How to get help

- Degree program questions or general help - umsimadshelp@umich.edu
- Coursera's Technical Support (24/7) - https://learner.coursera.help/

Library Access

Refer to the U-M Library's information sheet on accessing library resources from off-campus. For more information regarding library support services, please refer to the U-M Library Resources section of the UMSI Student Handbook (access to the Student Orientation course required).

Student Mental Health

Refer to the University's Resources for Stress and Mental Health website for a listing of resources for students.

Student Services

Refer to the Introduction to UMSI Student Life section of the UMSI Student Handbook (access to the Student Orientation course required).

Readings

Note: You need a free O'Reilly learning platform account to access many of the readings. Create an account using your <uniqname>@umich.edu email address by visiting: https://www.oreilly.com/library/view/temporary-access/

From the dropdown, select "Not listed? Click here." Then enter your <uniqname>@umich.edu email address. O'Reilly will send you an account activation email. Click the embedded red button to activate your account. You now have free access to hundreds of titles. Start reading.

WEEK 1	
Required	Chapter 2, Business Problems and Data Science Solutions In Fawcett, Tom. (2013). Data science for business. Sebastopol, CA : O'Reilly.
Required	Chapter 1, Interview with Chris Wiggins. In Gutierrez, Sebastian. (2014). Data Scientists at Work. Berkeley, CA : Apress : Imprint: Apress.
Required	Chapter 4, Interview with Erin Shellman. In Gutierrez, Sebastian. (2014). Data Scientists at Work. Berkeley, CA : Apress : Imprint: Apress.
Required	Chapter 16, Interview with Jake Porway. In Gutierrez, Sebastian. (2014). Data Scientists at Work. Berkeley, CA : Apress : Imprint: Apress.
Optional	Chapter 3, Arms Race: Going to College In O'Neil, Cathy. (2016). *Weapons of Math Destruction : How Big Data Increases Inequality and Threatens Democracy*. New York: Broadway Books.
Optional	Rogati, Monica. (2017). How do I become a Data Scientist? Good Audience blog.
Optional	Kaduk, Taras. (2016). 4 Stages of Data Analytics Maturity: Challenging Gartner's Model. LinkedIn.
WEEK 2	
Required	Pages 19-25 and Chapter 10, The Law of Small Numbers (all; pp 109 - 118). In Kahneman, Daniel. (2011). *Thinking, fast and slow*. New York : Farrar, Straus and Giroux.
Required	Mester, Tomi. (2017). Statistical Bias Types explained (with examples) – part 1. Data36 blog.
Required	Bailey, Brendan. (2017). Data Cleaning 101. Towards Data Science blog.

Required	Tait, Andrew. (2017). 10 Rules for Creating Reproducible Results in Data Science. Dataconomy blog.
Optional	Keng, Brian. (2015). The Gambler’s Fallacy and the Law of Small Numbers. Bounded Rationality blog.
Optional	Lee, N.T., Resnick, P., and Barton, G. (2019). Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. Brookings Institution report.
Optional	Data Cleansing. Wikipedia.org
WEEK 3	
Required	Overfitting in Machine Learning: What It Is and How to Avoid It. EliteDataScience.com
Required	Ray, Sunil. (2018). Improve Your Model Performance using Cross Validation (in Python and R). Analytics Vidhya. * Read Introduction section only
Required	Ranganathan, P., Pramesh, C. S., & Buyse, M. (2016). Common pitfalls in statistical analysis: The perils of multiple testing. *Perspectives in clinical research*, *7*(2), 106–107. doi:10.4103/2229-3485.179436
Required	Anderson, Brian. (N.D.)P-Hacking and the Problem of Multiple Comparisons. Musings, Dr. Brian Anderson’s blog.
Required	Spurious Correlations. (N.D.) Tylervigen.org.
Required	Koehrsen, Will. (2018). Correlation vs. Causation: An Example. Towards Data Science blog.
Required	Wagner, Clifford. (1982). Simpson's Paradox in Real Life. *The American Statistician, 36*(1), 46-48. doi:10.2307/2684093.
Required	Appleton, D., French, J., & Mark P. J. Vanderpump. (1996). Ignoring a Covariate: An Example of Simpson's Paradox. *The American Statistician, 50*(4), 340-341. doi:10.2307/2684931
Required	Rohrer, Julia. (2017). That one weird third variable problem nobody ever mentions: Conditioning on a collider. The 100% CI blog.
Optional (Recommended)	Chapter 5, Desperately Seeking Signal. In Silver, Nate. The Signal and the Noise; Why so Many Predictions Fail-- But Some Don’t. Penguin Press, 2012.
Optional	Section 3.1, Cross-validation: evaluating estimator performance. (N.D.) Scikit-learn.org. * Read Section 3.1 only, no sub-sections.
WEEK 4	
Required	Dykes, Brent. (2016). A History Lesson On The Dangers Of Letting Data Speak For Itself. Forbes.com.
Required	Zawadzki, Jan. (2018). Storytelling for Data Scientists. Towards Data Science blog.
Required	Kaynar-Kabul, Ilknur. (2017). Interpretability is crucial for trusting AI and machine learning. The SAS Data Science blog.
Required	Chapter 2, Are You Smarter than a Television Pundit? *(Required: Start with “A Fox-Like Approach to Forecasting” and read through “Principle ;” rest of chapter is optional.)*

	In Silver, Nate. The Signal and the Noise; Why so Many Predictions Fail-- But Some Don’t. Penguin Press, 2012.
Required	Chapter 6, How to Drown in Three Feet of Water. *(Required: Read through Figure 6-2)* In Silver, Nate. The Signal and the Noise; Why so Many Predictions Fail-- But Some Don’t. Penguin Press, 2012.
Required	Irwin, N., & Quealy, K. (2014, May 02). How to avoid being misled by the jobs report. *New York Times.*
Required	Dudek, Tomasz. (2018). But What Is This “Machine Learning Engineer” Actually Doing? Medium.com.
Required	Newman, Riley. (2015). How We Scaled Data Science to all Sides of Airbnb Over 5 Years of Hypergrowth. VentureBeat.com
Optional	Hypothetical Outcome Plots (HOPs) example. Vega Project.
Optional	UW Interactive Data Lab. (2016). Hypothetical Outcome Plots: Experiencing the Uncertain.