INFO 7390

Advances in Data Sciences and Architecture

Course Syllabus

Course Information

Professor: Nik Bear Brown

Email: ni.brown@neu.edu

Office: 505A Dana Hall

Office Hours: Zoom Only

Note: I am also a master's student at Northeastern. Do not send e-mail to my student e-mail brown.ni@husky.neu.edu I almost never read that e-mail.

All classes will be held on ground in Boston.

Course website: Canvas

The class sessions will be hybrid through Zoom and in-class. You have the choice to attend in class or through Zoom. They are synchronous. You are expected to attend class during the class time.

Course Prerequisites

INFO 6105 - There will be an early assessment of what was learned in INFO 6105. Knowledge of basic statistical learning is essential to the course.

Course Description

Garbage-In Garbage Out (GIGO) may be the most widely used maxim in machine learning, but how does one assess the quality of each step in an analysis pipeline? This course teaches students how to understand their data, models and pipelines using visualization and computational skepticism, empowering them to create robust, trustworthy AI systems.

The first part of the course introduces AI Fluency and Computational Skepticism principles with the Botspeak framework. Students will learn critical thinking approaches for AI validation and develop systematic doubt in AI workflows.

The second part covers understanding the statistical properties of a data set visually, how to fix issues with their data, and how to graphically demonstrate how the data was improved. The choice of the right chart for a particular question is covered. The principles of visual design, including typography, contrast, balance, emphasis, movement, white space, proportion, hierarchy, repetition, rhythm, pattern, unity, and variety are covered.

The third part explores generative AI for data, focusing on creating synthetic data across various modalities (text, numeric, image, audio, video). Students will learn to evaluate synthetic data quality and apply computational skepticism to verify AI-generated information.

The fourth part covers visualizing causal relationships in data. The emphasis is on understanding visual techniques for separating causal relationships from correlation using advanced methods like DAGs and counterfactual frameworks.

Throughout the course, students will build agentic AI tools that implement the principles of computational skepticism and AI fluency to validate data, models, and analytical pipelines.

Learning Objectives

Learning objectives for the course are:

- **AI Fluency and Computational Skepticism:**
    - Master the Nine Pillars of AI Fluency (Strategic Delegation, Effective Communication, Critical Evaluation, Technical Understanding, Ethical Reasoning, Stochastic Reasoning, Learning by Doing, Rapid Prototyping, Theoretical Foundation)
    - Develop systematic doubt and validation approaches for AI systems
    - Create agentic AI tools that implement computational skepticism principles
    - Apply philosophical frameworks to data science problems
- **Data Understanding and Visualization:**

    - Understand research design, research methods, and effective writing
    - Descriptive statistics and probability distributions
    - Data preprocessing techniques (imputing data, normalizing and scaling data, data reduction)
    - Sampling, bootstrapping and confidence intervals
    - Error analysis and data drift detection
    - Data visualization principles and techniques
    - Exploratory data analysis (EDA)
    - Charts for different purposes (compositional, distribution, trends, relationships)
    - Principles of visual design
    - Bias, fairness, and error analysis
- **Generative AI and Synthetic Data:**
    - Understanding generative AI systems and their capabilities
    - Synthetic data generation across modalities
    - Evaluating synthetic data quality
    - Data verification using generative AI
    - Ethical considerations in synthetic data creation
- **Causal Inference:**
    - Evidence Knowledge Graphs (EKG)
    - Causal inference principles and techniques
    - Potential outcomes and counterfactuals
    - Causal graphs and DAGs
    - Observational studies
    - Confounding and d-separation
    - Inverse Probability of Treatment Weighting (IPTW)
    - Model interpretability

Weekly Schedule

Part 0: AI Fluency and Computational Skepticism (Weeks 1-3)

Week 1: Botspeak - The Nine Pillars of AI Fluency

- Introduction to the Botspeak framework for human-AI collaboration
- Overview of the Nine Pillars and interaction modes
- Assignment: Design an agentic AI tool that implements Strategic Delegation for data analysis tasks

Week 2: Philosophical Foundations of Computational Skepticism

- Skepticism in philosophy applied to AI systems
- Truth and falsifiability in data-driven conclusions
- Critical thinking approaches for AI validation
- Assignment: Develop a computational framework to identify assumptions in datasets

Week 3: Practical Applications of Computational Skepticism in AI

- The Black Box Problem—Is understanding necessary for trust?
- Adversarial approaches to AI validation
- Developing systematic doubt in AI workflows
- Assignment: Create an agentic AI tool that implements the Critical Evaluation pillar to validate model outputs

Part I: Understanding Data (Weeks 4-6)

Week 4: Data Preprocessing and Validation

- Visual understanding of dataset statistical properties
- Data cleaning and transformation techniques
- Handling missing data, outliers, and errors
- Applying Critical Evaluation to dataset validation
- Assignment: Build an agentic AI tool that applies Stochastic Reasoning to detect data anomalies

Week 5: Data Analysis and Improvement

- Identifying potential problems within data using computational skepticism
- Implementing corrective measures
- Graphical methods to demonstrate data enhancement
- Selecting appropriate chart types for specific analytical questions
- Assignment: Develop an agentic AI tool that helps visualize data improvements

Week 6: Principles of Visual Design in Data Presentation

- Fundamental visual design principles: typography, contrast, balance
- Advanced design concepts: emphasis, movement, white space, proportion, hierarchy
- Using Effective Communication to present data findings clearly
- Assignment: Create an agentic AI tool that applies the Effective Communication pillar to optimize data visualizations

Part II: Generative AI for Data (Weeks 7-9)

Week 7: Understanding Generative AI

- Distinguishing between traditional ML, generative AI, and AGI
- Practical applications and limitations of generative AI
- Bias detection in AI models
- Assignment: Build an agentic AI tool that uses Strategic Delegation to guide generative AI outputs

Week 8: Building Generative AI Systems

- Research, design, data gathering, model training, and assessment procedures
- Importance of diverse datasets and evaluation techniques
- Robust AI systems against adversarial attacks
- Assignment: Develop an agentic AI that implements Rapid Prototyping for generative AI systems

Week 9: Employing Generative AI for Synthetic Data Creation

- Mechanics behind data generation processes across multiple modalities
- Students select their focus area (text, numeric, images, audio, or video)

- Techniques for evaluating synthetic data quality
- Assignment: Create an agentic AI tool that generates and validates synthetic data in the student's chosen modality

Part III: Causal Inference (Weeks 10-13)

Week 10: What is Causal Inference?

- Fundamental concept of causal inference
- Distinguishing correlation from causation
- Potential outcomes and counterfactuals
- Causal effects, causal assumptions, stratification
- Assignment: Build an agentic AI tool that helps identify potential causal relationships in data

Week 11: Visual Techniques in Causal Data

- Visualization of causal relationships within datasets
- Confounding, causal graphs, and Directed Acyclic Graphs (DAGs)
- Relationship between DAGs and probability distributions
- Paths and associations, conditional independence (d-separation)
- Assignment: Develop an agentic AI that visualizes causal relationships using DAGs

Week 12: Advanced Causal Analysis Techniques

- Observational studies, optimal matching, sensitivity analysis
- Inverse Probability of Treatment Weighting (IPTW)
- Marginal structural models, IPTW estimation
- Causal effect identification and estimation
- Assignment: Create an agentic AI tool that applies causal inference techniques to real-world data

Week 13: Final Projects

- Student presentations of integrated agentic AI systems
- Evaluation of data validation approaches

- Discussion of emerging trends in data science and AI validation
- Ethical considerations for the future of AI-assisted data analysis
- Final Project: Present a comprehensive agentic AI system that integrates computational skepticism principles with data validation techniques

Communication

Communication between instructor and students is through

- Email via the Canvas distribution list
- Announcements posted on Canvas
- Notes posted on the Canvas discussion board
- Private email exchanges

Course Structure

- Regularly test students on paper/algorithmic exercises
- Evaluate students' implementation competency, using assignments that require coding on given datasets
- Evaluate ability to setup data, code, and execute using python language
- Exams
- Final project is typically asking and answering a "real world" question of interest using machine learning techniques

Course GitHub

The course GitHub (for all lectures, assignments and projects):

https://github.com/nikbearbrown

nikbearbrown YouTube channel

Over the course of the semester I'll be making and putting additional data science and machine learning related video's on my YouTube channel.

https://www.youtube.com/user/nikbearbrown

The purpose of these videos is to put additional advanced content as well as supplemental content to provide additional coverage of the material in the course. Suggestions for topics for additional videos are always welcome.

Teaching assistants

The Teaching assistants are:

TBA

Programming questions should first go to the TA's. If they can't answer them then the TA's will forward the questions to the Professor.

Learning Assessment

Achievement of learning outcomes will be assessed and graded through:

- Quizzes

- Exams
- Completion of assignments involving scripting in R or python, and analysis of data
- Completion of a term paper asking and answering a “real world” question of interest using machine learning techniques
- Portfolio piece
- Partidipcation (Counts as a 100 point assingment) the TAs will keep track of meaningful contributions to the class and give a score between 0-100 at the end of finals.
- ATTENDANCE (Counts as a 100 point assingment) the TAs will keep track of whether you are in class Zoom attendance does not count as .attendance.

Reaching out for help

A student can always reach out for help to the Professor, Nik Bear Brown ni.brown@neu.edu. In an online course, it’s important that a student reaches out early should he/she run into any issues.

AI Policy for Coursework

Use of AI in Assignments

The use of AI tools (such as ChatGPT, Claude, GitHub Copilot, etc.) is permitted in this course, with the following requirements:

1. All AI usage must be properly cited
    - Include the name of the AI tool used
    - Specify which portions of your work were AI-assisted
    - Describe how the AI was used (e.g., generating code, editing text, brainstorming ideas)
2. Demonstration of Understanding
    - Students must be able to explain any AI-generated content in their submissions
    - Teaching Assistants or the Professor may ask students to walk through and explain any part of their work

    - Inability to demonstrate understanding of submitted work may result in grade penalties

3. Academic Integrity
    - Students are responsible for all submitted work, regardless of how it was generated
    - AI should be used as a tool to enhance learning, not to bypass it
    - Using AI without citation constitutes academic dishonesty

Quality Expectations and Grading

Due to the widespread availability of AI tools, a relative grading component (20% of total points) will be applied to all assignments worth 100 points or more. This component evaluates your work compared to peers, with emphasis on:

- Originality and creativity beyond AI-generated content
- Depth of understanding demonstrated
- Customization and personalization of solutions
- Real-world applicability

Quality Score Breakdown (20 points)

- Bottom 25% (5 points)
    - Meets basic requirements but lacks depth or real-world relevance
    - Basic implementation with minimal customization
    - Limited error handling and agent interaction
    - Superficial documentation and demonstration
- 26-50th percentile (10 points)
    - Solid implementation with thoughtful agent design
    - Functional tools with appropriate integration
    - Basic error handling and memory implementation
    - Clear documentation and demonstration
- 51-75th percentile (15 points)

    - Strong technical implementation with clear real-world applications
    - Sophisticated agent interactions and tool integration
    - Comprehensive error handling and memory management
    - Professional documentation and compelling demonstration
    - Evidence of testing and performance optimization
- Top 25% (20 points)
    - Exceptional project demonstrating innovation and technical excellence
    - Novel application with demonstrable value
    - Advanced agent orchestration with sophisticated decision-making
    - Unique custom tool that significantly enhances capabilities
    - Production-ready implementation with attention to scalability
    - Outstanding documentation and presentation

Best Practices for AI Use

1. Use AI as a learning tool
    - Ask for explanations of concepts you don't understand
    - Request alternative approaches to problems
    - Use AI to review your work and suggest improvements
2. Maintain your voice and perspective
    - Edit and refine AI-generated content to reflect your understanding
    - Add your own insights and observations
    - Ensure the final work represents your learning and knowledge
3. Document your process
    - Keep a record of your prompts and the AI's responses
    - Note any modifications you made to AI-generated content
    - Be prepared to discuss your collaboration with AI during evaluation

By following this policy, you can ethically leverage AI tools while ensuring they enhance rather than replace your learning experience in this course.

Grading Policies

Students are evaluated based on their performance on assignments, performance on exams, and both the execution and presentation of a final project. If a particular grade is required in this class to satisfy any external criteria—including, but not limited to, employment opportunities, visa maintenance, scholarships, and financial aid—it is the student's responsibility to earn that grade by working consistently throughout the semester. Grades will not be changed based on student need, nor will extra credit opportunities be provided to an individual student without being made available to the entire class.

Grading Rubric

A point system is used. Everything that you are expected to turn in has points. Points can range from 1 point to 1000 points Assignments get a 10% deduction for each day they are late rounded up. Exams cannot be made up unless arraignments are made before the exam.

I expect to use the following grading scale at the end of the semester. You should not expect a curve to be applied; but I reserve the right to use one.

Score	Grade
93 – 100	A
90 – 92	A-
88 – 89	B+
83 – 87	B
80 – 82	B-

78 – 79 C+

73 – 77 C

70 – 72 C-

60 – 69 D

<60 F

Typically grades will end up roughly 25% A, 25% A-, 25% B+, 20% B , 5% less than B but that depends on students' performance.

Canvas

You will submit your assignments via Canvas *and* Github. Click the title of assignment (Canvas -> assignment -> <Title of Assignment>), to go to the submission page. You will know your score on an assignment, project or test via Canvas. Canvas only represents only the raw scores. Not normalized or curved grades. A jupyter notebook file ALONG with either a .DOC or .PDF rendering of that jupyter notebook file must be submitted with each assignment.

Multiple files must be zipped. No .RAR, .bz, .7z or other extensions.

Assignment file names MUST start with students last name then first name OR the groups name and include the class number and assignment number.

Assignment MUST estimate the percentage of code written by the student and that which came from external sources.

Assignment MUST specify a license at the bottom of each notebook turned in.

All code must adhere to a style guide and state which guide was used.

Due dates

Due dates for assignments are midnight on the date assigned.

Five percent (i.e. 5%) is deducted for each day an assignment is late. Solutions will be posted the following Monday. Assignments will receive NO CREDIT if submitted after the solutions are posted. Any extensions MUST be granted via e-mail and with a specific new due date

Course Materials

Many of the textbooks are all available for free to NEU students via SpringerLink (http://link.springer.com/) or via the authors website. The textbooks we will be using in this class are:

Reinforcement Learning: An Introduction by Richard S. Sutton and Andrew G. Barto

http://incompleteideas.net/book/bookdraft2017nov5.pdf

Causal Inference in Statistics - A Primer by Judea Pearl
https://www.amazon.com/dp/1119186846/ref=cm_sw_r_tw_dp_U_x_IjayEbNAZYFG5

An Introduction to Causal Inference by Judea Pearl
https://www.amazon.com/dp/1507894295/ref=cm_sw_r_tw_dp_U_x_4fayEbZPY0Z68

Interpretable Machine Learning A Guide for Making Black Box Models Explainable. Christoph Molnar https://christophm.github.io/interpretable-ml-book/

The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2017)

Authors: Trevor Hastie, Robert Tibshirani and Jerome Friedman

Free online https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12.pdf

Deep Learning - Adaptive Computation and Machine Learning series by Ian Goodfellow, Yoshua Bengio, and Aaron Courville

https://github.com/HFTrader/DeepLearningBook

Recommended Texts

An Introduction to Statistical Learning with Applications in R (2013)

Authors: Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

Free online via SpringerLink (http://link.Springer.com/)
http://link.Springer.com/book/10.1007/978-1-4614-7138-7

The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2017)

Authors: Trevor Hastie, Robert Tibshirani and Jerome Friedman

Free online https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12.pdf

Beginning Python

From Novice to Professional

Authors: Magnus Lie Hetland 2017

ISBN: 978-1-4842-0029-2 (Print) 978-1-4842-0028-5

https://link.Springer.com/book/10.1007/978-1-4842-0028-5

Python Recipes Handbook

A Problem-Solution Approach

Authors: Joey Bernard 2016

ISBN: 978-1-4842-0242-5 (Print) 978-1-4842-0241-8

https://link.Springer.com/book/10.1007/978-1-4842-0241-8

Lean Python

Learn Just Enough Python to Build Useful Tools

Authors: Paul Gerrard 2016

ISBN: 978-1-4842-2384-0 (Print) 978-1-4842-2385-7

https://link.Springer.com/book/10.1007/978-1-4842-2385-7

Learn to Program with Python

Authors: Irv Kalb 2016

ISBN: 978-1-4842-1868-6 (Print) 978-1-4842-2172-3

https://link.Springer.com/book/10.1007/978-1-4842-2172-3

Deep Learning - Adaptive Computation and Machine Learning series by Ian Goodfellow, Yoshua Bengio, and Aaron Courville

https://github.com/HFTrader/DeepLearningBook

Beginning Python

From Novice to Professional

Authors: Magnus Lie Hetland 2017

ISBN: 978-1-4842-0029-2 (Print) 978-1-4842-0028-5

https://link.Springer.com/book/10.1007/978-1-4842-0028-5

Deep Learning with Python

A Hands-on Introduction

Authors: Nikhil Ketkar 2017

ISBN: 978-1-4842-2765-7 (Print) 978-1-4842-2766-4

https://link.Springer.com/book/10.1007/978-1-4842-2766-4

Pro Python Best Practices

Debugging, Testing and Maintenance

Authors: Kristian Rother 2017

ISBN: 978-1-4842-2240-9 (Print) 978-1-4842-2241-6 (Online)

https://link.Springer.com/book/10.1007/978-1-4842-2241-6

Mastering Machine Learning with Python in Six Steps

A Practical Implementation Guide to Predictive Data Analytics Using Python

Authors: Manohar Swamynathan 2017

ISBN: 978-1-4842-2865-4 (Print) 978-1-4842-2866-1

https://link.Springer.com/book/10.1007/978-1-4842-2866-1

Introduction to Data Science

A Python Approach to Concepts, Techniques and Applications

Authors: Laura Igual, Santi Seguí 2017

ISBN: 978-3-319-50016-4 (Print) 978-3-319-50017-1

https://link.Springer.com/book/10.1007/978-3-319-50017-1

Python Recipes Handbook

A Problem-Solution Approach

Authors: Joey Bernard 2016

ISBN: 978-1-4842-0242-5 (Print) 978-1-4842-0241-8

https://link.Springer.com/book/10.1007/978-1-4842-0241-8

Lean Python

Learn Just Enough Python to Build Useful Tools

Authors: Paul Gerrard 2016

ISBN: 978-1-4842-2384-0 (Print) 978-1-4842-2385-7

https://link.Springer.com/book/10.1007/978-1-4842-2385-7

Learn to Program with Python

Authors: Irv Kalb 2016

ISBN: 978-1-4842-1868-6 (Print) 978-1-4842-2172-3

https://link.Springer.com/book/10.1007/978-1-4842-2172-3

Big Data Made Easy

A Working Guide to the Complete Hadoop Toolset

Authors: Michael Frampton 2015

ISBN: 978-1-4842-0095-7 (Print) 978-1-4842-0094-0

https://link.Springer.com/book/10.1007/978-1-4842-0094-0

Software

python Anaconda

- https://www.continuum.io/anaconda-overview

Python Tutorials

Dive into Python http://diveintopython.org

Python 101 – Beginning Python
http://www.rexx.com/~dkuhlman/python_101/python_101.html

The Official Python Tutorial http://www.python.org/doc/current/tut/tut.html

The Python Quick Reference http://rgruet.free.fr/PQR2.3.html

Python Fundamentals Training – Classes http://www.youtube.com/watch?v=rKzZEtxIX14

Python 2.7 Tutorial Derek Banas· http://www.youtube.com/watch?v=UQi-L-_chcc

Python Programming Tutorial - thenewboston
http://www.youtube.com/watch?v=4Mf0h3HphEA

Google Python Class http://www.youtube.com/watch?v=tKTZoB2Vjuk

Nice free CS/python book https://www.cs.hmc.edu/csforall/index.html

Deep Learning Tutorials

MIT 6.S191: Introduction to Deep Learning http://introtodeeplearning.com/

Stanford Winter Quarter 2016 class: CS231n: Convolutional Neural Networks for Visual Recognition https://youtu.be/NfnWJUyUJYU

Deep Learning - Adaptive Computation and Machine Learning series by Ian Goodfellow, Yoshua Bengio, and Aaron Courville

https://github.com/HFTrader/DeepLearningBook

Participation Policy

Participation in discussions is an important aspect on the class. It is important that both students and instructional staff help foster an environment in which students feel safe asking questions, posing their opinions, and sharing their work for critique. If at any time you feel this environment is being threatened—by other students, the TA, or the professor—speak up and make your concerns heard. If you feel uncomfortable broaching this topic with the professor, you should feel free to voice your concerns to the Dean's office.

Collaboration Policies

Students are strongly encouraged to collaborate through discussing strategies for completing assignments, talking about the readings before class, and studying for the exams. However, all work that you turn in to me with your name on it must be in your own words or coded in your own style. Directly copied code or text from any other source MUST be cited. In any case, you must write up your solutions, in your own words. Furthermore, if you did collaborate on any problem, you must clearly list all of the collaborators in your submission. Handing in the same work for more than one course without explicit permission is forbidden.

Feel free to discuss general strategies, but any written work or code should be your own, in your own words/style. If you have collaborated on ideas leading up to the final solution, give each other credit on what you turn in, clearly labeling who contributed what ideas. Individuals should be able to explain the function of every aspect of group-produced work. Not understanding

what plagiarism is does not constitute an excuse for committing it. You should familiarize yourself with the University's policies on academic dishonesty at the beginning of the semester. If you have any doubts whatsoever about whether you are breaking the rules – ask!

Any submitted work violating the collaboration policies WILL BE GIVEN A ZERO even if "by mistake." Multiple mistakes *will be sent to OSCCR for disciplinary review.*

To reiterate: **plagiarism and cheating are strictly forbidden. No excuses, no exceptions.** *All incidents of plagiarism and cheating will be sent to OSCCR for disciplinary review.*

Assignment Late Policy

Assignments are due by 11:59pm on the due date marked on the schedule. Late assignments will receive a 5% deduction per day that they are late, including weekend days. It is your responsibility to determine whether or not it is worth spending the extra time on an assignment vs. turning in incomplete work for partial credit without penalty. Any exceptions to this policy (e.g. long-term illness or family emergencies) must be approved by the professor.

Five percent (i.e. 5%) is deducted for each day an assignment is late. Assignments will receive NO CREDIT if submitted after the solutions are posted. Any extensions MUST be granted via e-mail and with a specific new due date.

Only ONE extension will be granted per semester.

Student Resources

Special Accommodations/ADA: In accordance with the Americans with Disabilities Act (ADA 1990), Northeastern University seeks to provide equal access to its programs, services, and activities. If you will need accommodations in this class, please contact the Disability Resource Center (www.northeastern.edu/drc/) *as soon as possible* to make appropriate arrangements, and please provide the course instructors with any necessary documentation. The University

requires that you provide documentation of your disabilities to the DRC so that they may identify what accommodations are required, and arrange with the instructor to provide those on your behalf, as needed.

Academic Integrity: All students must adhere to the university's Academic Integrity Policy, which can be found on the website of the Office of Student Conduct and Conflict Resolution (OSCCR), at http://www.northeastern.edu/osccr/academicintegrity/index.html. Please be particularly aware of the policy regarding plagiarism. As you probably know, plagiarism involves *representing anyone else's words or ideas as your own*. It doesn't matter where you got these ideas—from a book, on the web, from a fellow-student, from your mother. It doesn't matter whether you quote the source directly or paraphrase it; if you are not the originator of the words or ideas, *you must state clearly and specifically where they came from*. Please consult an instructor if you have any confusion or concerns when preparing any of the assignments so that together. You can also consult the guide "Avoiding Plagiarism" on the NU Library Website at http://www.lib.neu.edu/online_research/help/avoiding_plagiarism/. If an academic integrity concern arises, one of the instructors will speak with you about it; if the discussion does not resolve the concern, we will refer the matter to OSCCR.

Writing Center: The Northeastern University Writing Center, housed in the Department of English within the College of Social Sciences and Humanities, is open to any member of the Northeastern community and exists to help any level writer, from any academic discipline, become a better writer. You can book face-to-face, online, or same day appointments in two locations: 412 Holmes Hall and 136 Snell Library (behind Argo Tea). For more information or to book an appointment, please visit http://www.northeastern.edu/writingcenter/.