Welcome to CMSC 462

Introduction to Data Science

Instructor: Justin Brooks, M.D., Ph.D.

About Me

- All over the map
- Largely human-centric research, system development
- I love data science: It blends logic, **creativity**, and purpose

What We'll Do Today

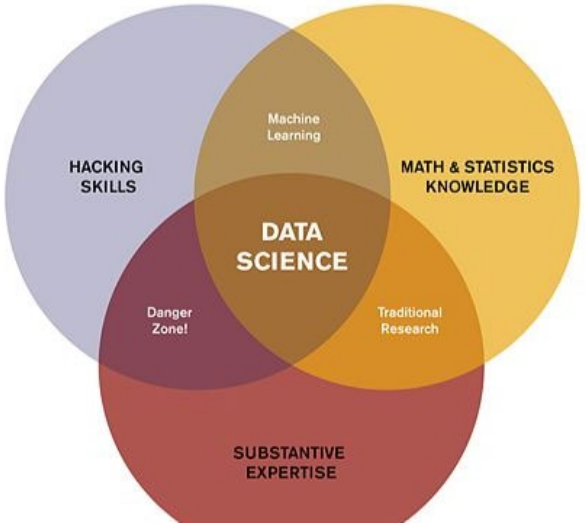
- Get to know each other
- Define data science
- Discuss how it differs from related fields
- Explore real-world applications and careers
- Preview the course structure and expectations
- Set a standard for what it means to be a competent data scientist

What Is Data Science?

- You tell me

The Way I Think About It

- Data Science = Statistics + Computer Science + **Domain Knowledge**
- Blends theory and application
- Closely related to statistics, but typically broader in tooling and applied goals



Drew Conway

Why Is It Hard to Define?

- Everyone uses it differently
- Predict vs. Forecast
- Context matters: military vs. finance vs. health
- Key trait: data scientists adapt

What Data Scientists Do

- Frame questions
- Find and clean messy data
- Apply models
- **Validate**, interpret, and communicate results
- Work across teams and disciplines

Data Scientist vs. Other Roles

Role	Focus
Statistician	Inference, modeling assumptions
Data Analyst	Descriptive, business-facing
ML Engineer	Production, scalability
Research Scientist	Exploration, publication
Data Engineer	Pipelines, database management, scalable systems
Analytics Engineer	Bridge between data infrastructure and business decision-making

The Rise of Data

- Big data, sensors, apps, devices
- Real-time decision making
- Data Science is powerful, but: Garbage in = Garbage out
- Your job = validate, interpret, explain

Applications/Jobs

- Military – sensor fusion, human performance, command & control
- Healthcare – Risk prediction, wearables, real-time prevention, clinical decision support, pharma
- Commercial – A/B testing, customer churn/targeting, sales forecasting, inventory forecasting

What Makes a Good Data Scientist?

- Technical skills → coding efficiently (memory management), file i/o, databasing, don't be that guy!
- Technical knowledge → statistics/machine learning
- Domain knowledge → reasonably transferrable (e.g. medical – human performance), but some idea of what is going is needed
- Creativity! The world is messy, keep an open mind about what the data are telling you

The Data Science Workflow

- Receive the Data
- Validate
- Assess Completeness
- Explore Relationships
- Model Selection
- Model Tuning
- Model Validation, Interpretation & Communication

Exploratory Data Analysis (EDA) ~60%-75% of your time on a project

Why is EDA so Important?

- Oak tree example – don’t waste your time
- Importance of domain knowledge to recognize implausible data
- Role of summary stats, visualizations, range checks
- EDA = sanity check + insight generator

Modeling

- What is modeling?
- This is driven by the question and involves mapping the data (that you know well from EDA) to the appropriate technique
- There is often a ‘best’ way but there is usually more than one way
- And sometimes the best way is to use more than one approach

Types of Models

- Regression: Predicting a numeric outcome (e.g., housing prices, risk scores)
- Classification: Predicting a category or label (e.g., spam vs. not spam)
- Clustering (Unsupervised Learning): Grouping similar data points when no labels exist
- Neural Networks: Layered models good for complex tasks like vision, speech, and language
- Machine Learning (ML): Umbrella term for algorithms that learn from data
- Emphasis in this course is on understanding the strengths, assumptions, and appropriate use of these tools—not memorizing syntax

How This Course Works

- Conceptual focus: not tool- or library-heavy
- Reinforces the data science workflow
- Exams test understanding, not syntax
- No group work

The DS Workflow (Course Roadmap)

- Week 1: Intro
- Week 2–5: EDA, wrangling
- Week 6–7: Probability, causality
- Week 8: Midterm 1
- Week 9–13: Modeling (regression, classification, unsupervised learning), time series
- Week 14: Midterm 2
- Week 15–16: Final projects

Grading Breakdown

- Exam 1: 20%
- Exam 2: 20%
- Final Project: 30%
- Homework: 20%
- Participation: 10%

Use of AI Tools

- Allowed with transparency
- **AI right now is WRONG A LOT**
- Must disclose: what tools, for what parts, how used
- No grade impact, but required for clarity

What Makes Data Dangerous?

- ALWAYS go with data
- NEVER delete data or manipulate data to get what you want
- You have an ethical responsibility to be truthful and forthright, your results can and will impact lives

Real World Example

- COMPAS (Correctional Offender Management Profiling for Alternative Sanctions):

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)