Week 2: Data Science Research Methods

DATA 510: Data Science Capstone

Lucas P. Cordova, Ph.D.

2026-05-18

First in a recurring series on data science research methods and technical writing. Overview of how capstone-grade research is framed, the PRIDE problem-to-question workflow, criteria for strong research questions, and an in-class paper critique activity in eight groups.

Table of contents

1 Learning Objectives

1.1 Today's Objectives

1.2 What You Will Learn Today

By the end of this session, you will be able to:

1. **Describe** how data science research in a capstone differs from exploratory analysis alone, and how it connects to program pillars (engineering, visualization, ethics, machine learning, statistics).
2. **Apply** the **PRIDE** workflow to move from a consequential problem to defensible research questions.
3. **Evaluate** research questions using a capstone-oriented checklist (focus, feasibility, ethics, evidence fit).
4. **Critique** how published applied data science papers motivate problems and state research questions in the abstract and introduction.

1.3 Course Connection

This session supports your **project proposal** and the communication quality expected at milestones and industry panel review.

2 Part 1: Why Research Methods In Capstone

2.1 Data Science Research Methods

2.2 What You Are Building This Semester

Over the term you will **propose, plan, and execute** a real data science project that:

- Integrates skills from the MSDS core curriculum into one coherent portfolio piece.
- Targets **real or plausible impact** on an organization or broader problem.
- Is graded on **steady progress**, **communication**, and **defense** of your choices (including industry panel feedback).

. . .

Research methods are how you keep the project from becoming "a notebook that ran models." They are how you justify *why* the work matters and *what* would count as success.

2.3 How This Lecture Series Fits

Week 2 starts a recurring thread on **research methods** and **technical writing**. Today: overview plus **research questions**. Later weeks we'll go deeper into study design, writing, and evaluation.

Bridge explicitly to Week 1 (research vs EDA). Students should hear that EDA and research are a loop, not a ladder.

2.4 What "Data Science Research" Means Here

Inspired by teaching-oriented work on data science as a discipline (Hicks and Irizarry, 2018; Reddy, 2021; Rosier, 2022):

- **Cases, not generic recipes:** Every capstone is a situated case (domain, stakeholders, constraints). Methods should match the case (Reddy; Rosier).
- **Evidence plus communication:** Results must be reproducible *and* legible to technical and non-technical audiences (Hicks and Irizarry).
- **Ethics inside the design:** Data ethics is part of problem framing and question formation, not a slide at the end (Atenas et al., 2023).

. . .

Your write-up should read like **applied data science research**: motivated problem, clear questions, appropriate methods, honest limits.

2.5 Capstone Pillars On One Slide

Pillar	What Reviewers Look For In Your Framing
Data engineering	Can you obtain, document, version, and reproduce the data pipeline?
Visualization & communication	Can non-experts understand stakes and findings?
Machine learning / analytics	Are models or analyses matched to the question?
Statistics & design	Do claims match evidence type (predictive vs causal, etc.)?
Ethics	Who is affected, what could go wrong, what mitigations exist?

Tie to syllabus project methodology page. Push back on "dashboard only" or "model only" projects during office hours.

2.6 Quick Quiz: Capstone Vs Homework

Which statement best matches a **capstone-grade** research stance?

```
A. Maximize test-set accuracy; the story writes itself.
B. Pick a dataset from Kaggle first, then hunt for a question.
C. Start from a stakeholder problem, then articulate what evidence would change decisions.
D. Avoid stating limits so the poster looks confident.
```

. . .

C. Data source and question still need instructor approval, but **problem-first** framing is the habit we are building.

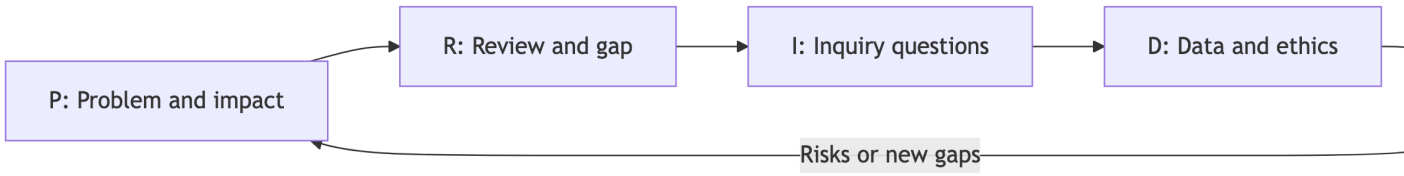
3 Part 2: From Problem To Research Questions

3.1 The PRIDE Workflow

3.2 PRIDE: Problem-First Framing

Use **PRIDE** to connect a consequential problem to questions you can defend this semester:

Step	Prompt	Capstone Output
P Problem & impact	Who is affected? What decision or outcome improves?	Problem statement, stakeholders
R Review & gap	What is known? What is unknown *and* matters?	Short related-work anchor
I Inquiry	What are your primary and secondary **research questions**?	Numbered RQs in proposal
D Data & ethics	What data exist? What harms, bias, or consent issues appear?	Data plan + ethics notes
E Evidence plan	What analysis, baseline, and metrics answer each RQ?	Methods / evaluation plan



3.3 Separate Problem, Question, And Task

Students often collapse these. Keep them distinct:

- **Research problem:** The gap or pain in the world (or organization) you address.
- **Research question(s):** What you need to **know** to address that problem.
- **Analysis / modeling tasks:** What you will **do** (features, models, pipelines, visuals).
- **Hypothesis (optional):** A testable claim when theory or prior work supports it; not required for every predictive capstone.

. . .

Bad sign: your "question" is only "we will use random forest." That is a task. Ask what you want to **learn** or **decide**.

3.4 Types Of Questions (Link To Week 1)

Type	You Are Trying To…	Capstone Example
Descriptive	Summarize	What is the weekly pattern of missing sensor readings?
Exploratory	Map structure when theory is thin	Which customer segments co-occur with late payments?
Inferential	Generalize from sample to population	Does mean error differ between two models after seasonality controls?
Predictive	Forecast or classify new cases	Can we predict 30-day churn from 90-day usage?
Causal	Support a cause claim	Did the policy *cause* the change, or did seasonality confound it?

Exploratory work is legitimate early; your **proposal** should still converge on questions sharp enough to finish.

3.5 Checklist: Strong Capstone Research Questions

A strong RQ is:

- **Focused:** One main idea per question; avoid "and also everything."
- **Feasible:** Answerable with data you can realistically access this term.
- **Ethical:** Survives a fairness / privacy / consent sanity check.
- **Decision-relevant:** Someone would act differently if you answered it well.

- **Evidence-aligned:** Predictive questions need prediction metrics; causal claims need design, not only AUC.

. . .

FINER mnemonic (adapted from clinical research): **F**easible, **I**nteresting, **N**ovel (for your context), **E**thical, **R**elevant.

3.6 Quick Quiz: Sharpen The Question

Weak: "Analyze customer data to find insights."

Which revision is **most** capstone-ready?

```
A. Build a dashboard of all customer columns.
B. Among B2B accounts in 2024, can we predict 60-day churn from product usage in the prior 9
C. Use deep learning because it is modern.
D. Prove our product causes retention.
```

. . .

B. It names population, horizon, task, and a baseline. **D** is causal and likely overclaims without design.

4 Part 3: In-Class Paper Activity

4.1 Paper Critique Activity

4.2 Group Activity Overview

Canvas Page: Week 2 Paper Activity (groups, papers, links).

4.3 What You Read (Each Person, Then Together)

Read **on your own** first:

- Abstract
- Introduction (through research questions / aims)

Then **as a group**, discuss and write up:

1. **Motivation:** What problem and stakes does the authors establish?

2. **Research questions:** What are they (verbatim or tight paraphrase)?
3. **Evaluation:** Are the questions clear, feasible, ethical, and matched to methods?
4. **Critique:** What is strong? What would you revise for a capstone-scale project?

4.4 Groups

Group	Members
1	Rohan, Luca, Sophia, Tiffany
2	Addison, Manish, Aaron, Summer
3	Sarah, Aiyana, Jon, Emery
4	Ben, Jackson, Dylan, Amaya
5	Spencer, Mary, Seira, Emily
6	Dane, Mike, Brooke, Brandon
7	Bradley, Siera, Simon, Serenna
8	Shanti, Alex, Courtney

4.5 Paper Assignments

Download your paper from the link on the Canvas activity page or below.

Group	Domain	Paper
1	Healthcare Finance	Predicting High-Cost Patients (PLOS ONE)
2	Public Hospital Operations	AI For Bed Regulation, Brazil (PLOS ONE)
3	Emergency Medical Services	Pre-Hospital Transport Prediction, Qatar (PLOS ONE)
4	Social Science / NLP	Climate Discourse On Twitter (PLOS ONE)
5	Algorithmic Fairness / Policy	Fairness Dynamics Of Targeted Job-Seeker Help (*Sci. Rep.*)
6	Environmental Risk	Daily Wildfire Expansion Rate (MDPI *Fire*)
7	Computer Vision / UAV	UAV Forest-Fire Surveillance (PLOS ONE)

Group	Domain	Paper
8	NLP Methods	PEFT Vs Full Fine-Tuning, Multilingual News (PLOS ONE)

4.6 Canvas Submission And Rubric

Submit one file per group on Canvas (PDF or Word). Include all group members' names.

Criterion	Excellent (3)	Adequate (2)	Needs Work (1)
Motivation Summary	Clear stakes, stakeholders, and gap tied to the paper	Mostly clear, minor gaps	Vague or off-paper
Research Questions	Accurate RQs/aims from intro; quoted or carefully paraphrased	Minor inaccuracies	Missing or invented
Evaluation	Uses PRIDE/FINER-style criteria with specific evidence	Generic praise/critique	Superficial
Critique	Concrete revisions for capstone scale	Some specifics	Only opinion
Professional-ism	Concise, cited, one voice	Minor issues	Sloppy or incomplete

4.7 Report-Out Requirements (if selected)

One representative gives:

- A **2–3 sentence** motivation summary
- States the paper's questions
- Offers at least **one** strength and least **one** critique of the questions

4.8 Report-Out Selection

Four groups present today. Group numbers are drawn at random (not volunteer-only).

```
Select 4 distinct group numbers from 1 through 8. List them in ascending order.

Use cryptographically secure randomness (Python secrets.SystemRandom, or openssl rand).
Print a hex seed from os.urandom(16) before the draw. Draw once; repeat with an independent
seed as a check-if the two sets differ, keep the first draw. Show minimal Python 3 code that
reproduces the result from the printed seed.
```

5 Part 4: Paper Synopses –>

5.1 Group 1 Paper: High-Cost Patients

Langenberger, Schulte, and Groene (2023), *PLOS ONE* (healthcare claims, Germany).

Synopsis: Predict which insured members will become **high-cost** next year using routine sickness-fund claims, comparing random forest, gradient boosting, neural nets, and logistic regression for care management and prevention.

Stated Objective / Questions:

- **Objective:** Assess and compare ML algorithms for predicting future high-cost patients (classification) using routinely collected claims and cost data.
- **Implicit operational question:** Which algorithm achieves the best discrimination (e.g., AUC) on held-out future-year costs?

5.2 Group 2 Paper: Hospital Bed Regulation

RegulaRN Leitos Gerais platform (Brazil), *PLOS ONE* (47k+ regulation records).

Synopsis: Train and compare ML classifiers on state hospital **bed regulation** data to support regulators predicting patient outcomes (discharge vs death) and reduce subjectivity in bed assignment.

Stated Aims:

- Analyze RegulaRN data and train/validate ML models.
- Choose models that maximize accuracy, precision, recall, specificity, F1, and ROC-AUC for regulation outcomes.
- Discuss impacts of digital health tools on regulatory decision-making.

5.3 Group 3 Paper: Pre-Hospital Transport

HMCAS ambulance service (Qatar), *PLOS ONE* (~93k emergency calls).

Synopsis: Use ML on pre-hospital EMS data to predict whether a patient **must be transported** vs treated on scene, supporting resource allocation in a multinational population.

Stated Objective:

- Accurately predict transport vs non-transport cases using ML to enable efficient resource allocation (value improvement in EMS).

5.4 Group 4 Paper: Climate On Twitter

Shyrokykh, Girnyk, and Dellmuth (2023), *PLOS ONE* (UN agency tweets).

Synopsis: Compare lexicon, traditional ML, and deep learning for classifying whether tweets are **about climate change**, in a social-science setting with **small labeled data** and class imbalance.

Stated Contributions (Implicit Questions):

- How do lexicon vs supervised ML methods perform for this rare-event text classification task?
- Do traditional ML models match deep learning with far less compute?

5.5 Group 5 Paper: Fairness In Labor Market Help

Long-term fairness dynamics, *Scientific Reports* (simulation + synthetic labor market).

Synopsis: Model how a public employment service's **targeted help**, driven by predictions that may use a **protected attribute**, affects long-term fairness between groups.

Stated Purposes / Questions:

1. Introduce complexity of assessing **long-term fairness** when targeted help uses protected attributes.
2. Answer: **How can we assess long-term fairness in a dynamical system such as a labor market?**
3. Examine trade-offs between fairness goals and targeted vs non-targeted aid under labor-market dynamics.

5.6 Group 6 Paper: Wildfire Expansion

Shmuel and Heifetz (2023), MDPI *Fire* (global daily fire observations).

Synopsis: Apply XGBoost, random forest, MLP, and logistic regression to predict **daily wildfire growth** from weather, topography, and fuels; also classify whether growth rate will increase or decrease the next day.

Stated Problem (Implicit Questions):

- Can ML outperform classical models for daily burned area and for direction-of-change in growth rate on a global dataset?
- Which factors drive growth rate under scenarios with vs without prior fire-behavior variables?

5.7 Group 7 Paper: UAV Fire Detection

SHAMTA and Demir (2024), *PLOS ONE* (UAV + edge AI).

Synopsis: Build a UAV surveillance system with YOLOv8/v5 and CNN-RCNN models for **early forest-fire** detection, Jetson Nano onboard inference, and a ground-station interface for coordinates and images.

Stated Contributions (Engineering Aims):

- Integrate UAV, edge hardware, and ground station.
- Compare detection vs classification pipelines for fire imagery in real time.

5.8 Group 8 Paper: PEFT For Multilingual News

Parameter-efficient fine-tuning study, *PLOS ONE* (SemEval-2023 news tasks).

Synopsis: Compare **adapters**, **LoRA**, and **full fine-tuning** on multilingual news classification (genre, framing, persuasion) across languages and training scenarios.

Stated Research Questions:

- **RQ1:** How do classification performance and computational costs differ by training technique per sub-task?
- **RQ2:** How do training scenarios (language diversity and dataset size) affect each technique?
- **RQ3:** How do techniques compare within each scenario and language?

6 Wrap-Up

6.1 Big Ideas

1. **Problem before algorithm:** Stakeholders and decisions come first; models are evidence tools.
2. **PRIDE is your proposal spine:** Problem, gap, questions, data/ethics, evidence plan.
3. **Questions are not tasks:** "Use XGBoost" is not an RQ; "Can we predict Y better than baseline Z?" is closer.
4. **Published papers vary:** Some state RQ1–3; others bury aims in contributions. Learn to extract and critique both.
5. **Your capstone needs instructor-approved data early:** Questions must be feasible with *your* data, not only a beautiful intro.

6.2 References

1. Hicks, S. C., and R. A. Irizarry. 2018. A guide to teaching data science. *The American Statistician*.
2. Reddy, Y. M. 2021. Teaching research methodology: Everything's a case. *Journal of Education*.
3. Rosier, J. 2022. The case method evaluated in terms of higher education research. *Journal of University Teaching & Learning Practice*.
4. Atenas, J., et al. 2023. Reframing data ethics in research methods education. *Journal of Information Literacy*.
5. Assigned activity papers (Groups 1–8); links on Canvas activity page.