

Data Summary Strategy: Scenario Cards and Persona Bank

DATA 510: Data Science Capstone, M2 Activity Handout

Scenario cards

Scenario A: Safe Harbor (Shelter Services)

A regional shelter nonprofit wants to understand how long clients wait for shelter and which factors relate to whether someone returns for services.

Attribute	Detail
Data sources	Intake database export, case-management system export, web application form
Volume	~300 clients/yr; ~8,000 form records across 2024 to 2025
Pipeline sketch	Staff manually pull flat files to a secure share, remove PII, match persons across systems, then clean to an analysis-ready table
Primary stakeholder	Program staff making service-delivery decisions
Wrinkle	Sensitive PII; no shared client ID across the three systems; small outcome counts

Scenario B: LedgerBridge (Nonprofit Finance)

A nonprofit bike shop with retail, repair, and grant-funded programs wants one board-ready view of financial and program health.

Attribute	Detail
Data sources	Retail POS API, accounting API, grant-tracking Google Sheets
Volume	Thousands of transactions/yr; several grant ledgers
Pipeline sketch	Scheduled ETL extracts and dedups, reconciles on a common as-of date, loads to Postgres, feeds a board dashboard
Primary stakeholder	Non-technical executive team and board of directors
Wrinkle	Apparent monthly losses vs accrued reimbursable grant revenue; messy joins across POS and accounting

Scenario C: JobsAhead (Economic Forecasting)

Forecast national unemployment across 3, 6, and 12 month horizons using several public macroeconomic indicators.

Attribute	Detail
Data sources	Six public series: labor, inflation, interest rates, growth, sentiment, recession flag
Volume	~70 years; integrated panel of ~838 monthly rows, ~40 variables
Pipeline sketch	Standardize, align mixed frequencies to monthly, time-join into one panel, engineer lag and momentum features
Primary stakeholder	Workforce-planning analysts at a state agency
Wrinkle	Frequency alignment; association not causation; published series get revised

Scenario D: CourtMetrics (Small-N Original Data)

Study how ball characteristics and fatigue relate to shot accuracy in a small athlete cohort, using data the team collects itself.

Attribute	Detail
Data sources	Originally collected sensor readings plus manual shot logs
Volume	~7 athletes, 150 to 200 shots each
Pipeline sketch	Consent and anonymization, per-shot logging, feature build, per-player baselines
Primary stakeholder	The team coach
Wrinkle	Tiny N; coach-player power dynamic in consent; individual differences may dominate

Scenario E: CanopyWatch (Geospatial)

Map the spread of an invasive tree pest against canopy loss and stream temperature across a county.

Attribute	Detail
Data sources	Large shapefiles (~1.5 GB), a 1M+ record event table, public tree inventory, sensor streams
Volume	Millions of records; gigabytes of geospatial files
Pipeline sketch	Build a spatial grid as the unit of analysis, perform spatial joins, aggregate everything to grid cells
Primary stakeholder	City natural-resources office
Wrinkle	Heavy geospatial plumbing before any analysis; some layers scraped from map services

Scenario F: PricedOut (Public Housing Data)

Rank metro areas by housing-affordability risk using public data, to flag which mid-sized cities may price out the middle class next.

Attribute	Detail
Data sources	Home-value index, federal housing and price indices, Census ACS, economic series
Volume	Hundreds of metros, monthly, multiple years
Pipeline sketch	Download refreshable public files, join on metro and month, construct a composite risk index
Primary stakeholder	A regional housing journalist
Wrinkle	Defining “affordability risk” defensibly; many sources with uneven coverage

Scenario G: DiamondSim (Sports Outcome Simulator)

Simulate baseball game outcomes by feeding a plate-appearance model into a Monte Carlo game engine.

Attribute	Detail
Data sources	Pitch-level data, play-by-play logs, player and roster tables
Volume	Millions of plate-appearance rows across multiple seasons
Pipeline sketch	Tiered raw to processed load on a hosted Postgres, dedup on conflict, train a classifier, feed a simulator
Primary stakeholder	A front-office analyst
Wrinkle	Data leakage risk in the train/test split; heavy pipeline plus simulator scope

Scenario H: EnrollPath (Enrollment Prediction)

Predict where admitted applicants will actually enroll, to help an enrollment office improve yield.

Attribute	Detail
Data sources	Admissions CRM export, national enrollment clearinghouse data
Volume	~10,000 records each, one cycle
Pipeline sketch	Ingest both sources, join on applicant, build features, model destination, validate
Primary stakeholder	VP of Enrollment
Wrinkle	PII and FERPA concerns; bias in predictive models that influence outreach

Stakeholder persona bank

Each team is assigned one of these. Tailor your data summary for this reader, and be ready to explain how a different stakeholder would change your choices.

Persona	Cares most about	Distrusts or ignores	Your move
Executive funder / grant officer	Impact, the one-sentence “so what,” cost	Method jargon, hedging	Lead with outcome and a clear ask
Skeptical domain expert	Whether you respect the field’s nuance	Naive assumptions, overclaiming	Name limitations first, cite the field
Hiring-manager data scientist	Rigor, validation, reproducibility	Vague claims, no baseline	Show the validation table and the rebuild command
Non-technical board / city council	A decision they can defend publicly	Equations, pipelines	Plain takeaway plus one supporting number
Investigative journalist	Claims that hold up, ethics, overreach	Inflated certainty	State what the data can and cannot say
New analyst inheriting the project	Can they rebuild and extend it	Undocumented steps	Walk the lineage and the data dictionary

Persona	Cares most about	Distrusts or ignores	Your move
IRB / ethics reviewer	Consent, PII, fairness	Privacy hand-waving	Detail de-identification and access controls
Operations staff (daily dashboard user)	Does it work day to day, is it current	One-off analysis, stale data	Show refresh cadence and the live view