DL 903 — Empirical Study-Design Seminar: The Pursuit Discipline's Machinery as a Research Object

The seminar where the doctorate earns its keep. The discipline's machine side has built a remarkable empirical apparatus — probability-of-win composites, calibration and drift monitoring, agency win-pattern corpora, end-to-end verification — and almost none of it has been studied*. This seminar turns that apparatus into human research.*

Level: doctoral · **Prerequisites:** DL 901 (can be concurrent with DL 902) · **Credits:** 3 · **Format:** one 3-hour seminar per week, one term · **Stack-agnostic promise:** the machinery is studied as *concepts* — what each piece claims, and how you would test the claim. No lesson requires operating any tool (doctrine/08).

Course description

This is the doctoral course that has no undergraduate analog. It exists because the pursuit discipline now has something most young fields do not: a **machine-side empirical layer** that encodes the doctrine and, crucially, *makes predictions that can be checked against reality.* Someone — an automated pursuit layer, a scoring composite, a calibration monitor — is producing probability estimates, and the market is producing awards, and the pairing of the two is a research goldmine that almost nobody has mined.

The seminar has two jobs:

1. **Read the machinery as a researcher reads any research object** — understand, at concept level, what the discipline's empirical apparatus claims about the world: that a seven-factor composite estimates win probability, that a 0.7 means roughly 70%, that drift can be detected before it hurts, that winning-proposal patterns differ by agency, that an end-to-end verification can certify the whole pipeline on one canary company.
2. **Design the studies that test those claims.** For each piece of machinery, the seminar asks: *what is the falsifiable claim, what design would test it, what data would the design need, and what would a skeptical committee require before they believed the result?*

The deliverable is a **seminar research design** — one study, fully designed, defensible to a skeptical panel — that sits on the research agenda (modules/doctoral/research-agenda.md) and is designed to grow into the dissertation's empirical core.

Learning outcomes

By the end of this seminar, a student can:

1. **Explain**, at concept level, what each major piece of the discipline's empirical machinery claims, and name the assumption each piece hides.
2. **Translate** a piece of machinery into a falsifiable research question.
3. **Design** an empirical study (design, measurement, sampling, analysis plan) that tests one machinery claim, using DL 901's design toolkit and DL 902's statistical toolkit.
4. **Critique** the field's existing machinery the way a peer reviewer would — naming the bias traps, the overclaims, and the missing validity evidence.
5. **Defend** a seminar research design before a skeptical panel.

How the seminar runs

Each session follows one arc: **the claim → the machinery → the hidden assumptions → the design.** The instructor introduces a piece of the discipline's machinery at concept level; the room names what it claims and what it assumes; teams design the study that would test it. Over the term, the student's own design is built up session by session and presented at the end.

Session 1 — The empirical layer as a research object

The claim. The discipline now *measures itself.* It has moved from "we think our estimates are good" to "here is an apparatus that produces estimates and checks them against outcomes."

The machinery, at concept level. The pursuit doctrine is encoded in a versioned, machine-readable package — rules, playbooks, formulas, and a research corpus — and around that package sits an empirical apparatus:

- a **probability-of-win composite** that turns factor scores into a 0-to-1 estimate;
- a **prediction-and-outcome record** that stores, for each pursued opportunity, the predicted win probability at decision time and, eventually, the terminal outcome (won / lost / withdrawn);
- a **calibration-and-drift monitor** that checks whether estimates stay honest over time;
- **agency win-pattern corpora** that curate what winning proposals look like, organized by agency;
- an **end-to-end verification** that walks the whole pipeline on a single canary company to certify the doctrine works.

The hidden assumptions. Each piece assumes something the seminar will spend the term testing: the composite assumes its factors and weights are right; the record assumes prediction and outcome are captured without corruption; the

monitor assumes its thresholds mean something; the corpora assume winning patterns generalize; the verification assumes a canary company proves anything about the world.

Design exercise. The room brainstorms: "What is the single most important claim of the apparatus, and what would it take to falsify it?" Each team writes one falsifiable version of that claim.

Reading. doctrine/03, doctrine/05, doctrine/08; the research agenda (modules/doctoral/research-agenda.md).

Session 2 — The seven-factor composite as a measurement instrument

The claim. A seven-factor composite — past performance, technical capability, price competitiveness, incumbent/relationship advantage, team strength, customer knowledge, strategic alignment — each scored on a scale and weighted into a single 0-to-1 estimate — measures something real called *probability of win*.

The machinery, at concept level. The composite is a weighted sum of factor scores, normalized to [0,1], with recommendation thresholds keyed to the score (strong bid / bid / conditional / likely no-bid / no-bid). It is the discipline's canonical instrument for the most important judgment it makes.

The hidden assumptions. That the seven factors are the right ones; that the weights are right (or even that a single set of weights serves all agencies and funder types); that the 1-to-10 factor scores are assigned reliably by different scorers; that a composite number captures something the market actually rewards.

The research questions. This is the factor-validation thread of the agenda:

- *Construct validity:* Does the composite predict awards better than a simple gut estimate?
- *Factor weights:* Which factors actually move the probability of winning — and do the discipline's hand-set weights match what the data show?
- *Reliability:* Do two experienced scorers score the same opportunity similarly? (If not, the composite is not measuring a property of the opportunity; it is measuring the scorer.)
- *Cross-agency stability:* Do the factors that predict wins at one agency or funder family hold at another?

Design exercise. Teams design the **factor-weight validation study**: the data (predicted factor scores + outcomes for a sample of pursued bids), the model (a logistic regression of won/lost on the seven factors — DL 902's bridge), the measurement challenges (who scores the factors, how reliability is checked), and the claim the study could honestly make.

Reading. doctrine/05; DL 901 Unit 4 (measurement) and DL 902 Movement 2 (logistic regression).

Session 3 — Prediction vs. outcome: the calibration question

The claim. A 0.7 win-probability estimate should mean the opportunity wins roughly 70% of the time. This is the single most important — and most checkable — claim the discipline makes.

The machinery, at concept level. The discipline measures prediction quality with the standard statistical tools: **Brier score** (mean squared error of the probability), **log-loss** (penalizing confident errors), **reliability diagrams** (observed win rate vs. predicted, by band), **expected calibration error (ECE)** (the sample-weighted average reliability gap), and — for drift — the **population stability index (PSI)** (whether the distribution of predictions has shifted from baseline). The discipline's monitor bands drift into *watch / warn / critical* when these statistics cross thresholds, and it recalibrates with a *monotone remap* — a curve that raises low estimates and lowers high ones while never reversing the ordering of opportunities.

The hidden assumptions. That the thresholds mean something (why watch at ECE 0.04 and critical at 0.15?); that calibration is stable enough across time and funder families to be worth monitoring; that the observation floor is high enough to separate signal from noise; that "withdrawn" really is a non-win.

The research questions. This is the prediction-and-calibration thread:

- *Is the composite calibrated?* In which bands is it over- or under-confident? (Reliability diagrams localize the answer.)
- *Is calibration stable across agencies and funder families?* A composite calibrated on one population may be miscalibrated on another.
- *Does calibration drift over time* — and does the drift predictably precede a change in the market or the estimator's usage?
- *Do experts or structured composites calibrate better?* A field-experiment question: randomize which estimation method teams use, compare calibration.

Design exercise. Teams design the **calibration evaluation study**: the data (predicted pWin at decision time joined to terminal outcome, for a defined window and funder family), the reliability-bin resolution, the headline statistics (Brier, ECE, PSI), the observation floor, the survivorship mitigation (only pursued bids resolve), and the pre-registered decision rule for "this composite is miscalibrated."

Reading. DL 902 Movement 3 in full; doctrine/05.

Session 4 — Agency win patterns: from corpora to hypotheses

The claim. Winning proposals are not random. Agencies publish review criteria, and across the record, recognizable winning patterns recur — patterns that the discipline has curated, agency by agency.

The machinery, at concept level. The discipline maintains **agency-level win-pattern corpora**: for each active agency (e.g., DOE, DOL, NSF, NIH, EDA-PWEA), a set of *pattern cards* — generically-known, publicly-cross-walked review patterns that correlate with wins, cited against the agency's published scoring rubric — plus *illustrative composites* that exemplify the patterns in narrative form. The patterns recur across agencies: a **named anchor partnership** (a specific institutional partner named in the proposal), a **quantified outcome target** (a measurable, concrete outcome), a **sustainability / continuation mechanism** (a named path past the grant period), and **positioning coherence** (the whole narrative hanging together toward the agency's mission). The corpora are used to ground proposal strategy; they are explicitly disciplined against fabrication — documented facts and clearly-labeled illustrative material are kept apart, and no invented specifics are attributed to real organizations.

The hidden assumptions. That the patterns are real and not just plausible; that they generalize across agencies and programs; that the curation did not unconsciously encode the curators' preferences; that a pattern's presence in a *winning* proposal says something causal (the winning proposals also contained a thousand other things).

The research questions. This is the agency-win-pattern thread:

- *Do the published review criteria predict what gets funded?* A content-analysis study: code proposals against the criteria, model funding on the codes.
- *Do the corpus's patterns discriminate winners from losers?* A comparative study: take a set of won and lost proposals (published and closed competitions), code for the presence of each pattern, and test whether the patterns separate the groups.
- *Do the patterns vary by agency in the way the corpus claims?* A cross-agency comparison — the corpus's central empirical assertion.

Design exercise. Teams design the **pattern-discrimination study**: the sample (won and lost proposals from published, closed competitions — a real, lawful, obtainable sample), the coding protocol (the pattern definitions as a codebook, with inter-rater reliability — DL 901's measurement discipline), the analysis (logistic regression of won/lost on pattern-presence indicators, clustered by competition), and the survivorship caution (the corpus draws on what was written and submitted; the sample is of proposals, not of the full opportunity universe).

Reading. The literacy maps (how to read an RFP / NOFO — the published-criteria reading); doctrine/02; DL 901 Unit 4 (reliability).

Session 5 — The learning loop: does the discipline learn from its outcomes?

The claim. An organization that measures outcomes and feeds them back — into its estimates, its rules, its playbooks — improves. The doctrine calls this the learning loop (doctrine/03, stage nine) and the cosmology loop (doctrine/07); the machinery encodes it as a **closed-loop recalibration** that updates the discipline's formulas from accumulated win/loss outcomes.

The machinery, at concept level. The machine-side learning loop has strict guardrails worth studying in their own right: it **does not train models** — it *recalibrates deterministic formulas* by fitting a monotone remap from predicted pWin onto observed award rates. The fit is bounded (a single noisy window cannot flip the curve), gated by an **observation floor**, stratified by **funder family**, and routed through a review pipeline with a documented risk boundary: recalibrating a win-probability composite is low-risk and can be automatic; recalibrating anything that touches pricing or compliance is human-reviewed, because pushing a price floor down from observed wins could violate the rules the field exists under.

The hidden assumptions. That the outcomes data are a faithful record; that the learning loop is not chasing noise (the feedback-oscillation problem); that what worked recently will keep working; that the human-review boundary is drawn in the right place.

The research questions. This is the learning-and-evolution thread:

- *Does recalibration actually improve calibration?* A before/after study of the formula's Brier/ECE before and after a recalibration event.
- *Does the organization learn from losses?* A panel study of a firm's or a portfolio's win rates over time, testing whether post-loss adjustments predict improvement.
- *Does gate discipline improve outcomes?* The governance question: do organizations that gate honestly — decline below-threshold pursuits, document criteria in advance — convert better than organizations that chase everything?

Design exercise. Teams design the **learning-loop evaluation**: a before/after or difference-in-differences design on calibration quality around a recalibration event, or a panel design on organizational win rates, with the clustering and survivorship hazards named.

Reading. doctrine/03, doctrine/04, doctrine/07.

Session 6 — Golden-thread verification as a reproducible research pattern

The claim. The whole pipeline can be verified end-to-end on a single company — the **canary company** — and a green (or amber) result means the doctrine is

working.

The machinery, at concept level. The discipline maintains a **golden-thread verification**: a staged walk of the entire pursuit pipeline — from prerequisites and seeding through sensing, scoring, gating, designing, analyzing, and the human-facing portal — executed against one canonical canary company (the same Ravonics the teaching case uses). Each stage reports **pass / warn / fail** with a list of evidence; the run concludes with an overall status (green / amber / red) and a written record of what was checked and what was found. The verification is *reproducible* — the same staged walk can be re-run, and its output is a documented artifact. It is, in other words, a **reproducible research pipeline** for a whole discipline.

The hidden assumptions. That the canary company's success generalizes; that a stage labeled "pass" proves the stage works (a pass can mean "the check ran," not "the outcome was right"); that the staged evidence actually covers the claims the discipline cares about; that amber is actionable.

The research questions. This is the methodology thread — *the golden-thread itself is a research method*:

- *What does a staged end-to-end verification certify, and what does it not?* A methodological study of the verification's validity — mapping each stage's evidence to the claim it supports, and finding the gaps.
- *Is the verification reproducible?* Re-run the walk and compare — the replication benchmark.
- *What would a rigorous evaluation of a pursuit pipeline look like?* The golden-thread is the discipline's answer; a doctoral student can design the *better* answer — a validation framework with outcome-based stage criteria instead of health checks.

Design exercise. Teams design the **verification-validity study**: take the staged golden-thread pattern (prerequisites, seeding, sensing, scoring, gating, designing, analyzing, portal) and, for each stage, write the falsifiable claim, the evidence a rigorous evaluation would require, and the design that would produce it. This becomes the methodological skeleton of a dissertation on "how to know a pursuit pipeline works."

Reading. The case study (case-study/ravonics.md — the canary company); doctrine/03.

Session 7 — Design workshop

A working session. Each student presents their seminar research design in progress — the claim, the design, the data plan, the analysis plan — and the room attacks it the way a dissertation committee would: *What is the identifying assumption? Where is the survivorship hazard? Is the sample big enough? Who scores the factors, and how do you know they agree? What exactly will you claim if the data go the other way?*

The goal is a design that survives contact with the room, not a design the room approves.

Session 8 — Research designs defended

The final session. Each student presents their seminar research design — one study that tests one claim of the discipline's machinery — before a skeptical panel (the instructor plus peers). The panel's job is the committee's job: push until the design's limits are named, then judge whether the design is *defensible*, not perfect.

Assessment

Layer	Weight	What it is
Session design memos	40%	A short design memo per machinery topic (Sessions 2–6): the claim, the hidden assumptions, the design, the data, the analysis plan.
Seminar participation	20%	The review posture: honest, specific critique of classmates' designs.
The seminar research design	40%	One complete research design — a study of one machinery claim — assembled from the session memos, defended in Session 8.

The seminar's stack-agnostic pledge

This seminar studies machinery, but it never operates it. Every session describes a piece of the discipline's empirical apparatus as *concepts* — the seven-factor composite, the calibration statistics, the drift bands, the monotone remap, the agency corpora, the golden-thread walk. A student who took this seminar and then encountered a completely different implementation of the same ideas — built by a different team, in a different decade — would recognize the claims immediately. That is the doctrine's test (doctrine/08), and this seminar passes it.

The seminar in two sentences

The discipline has built an empirical apparatus that makes checkable claims about the world — and almost nobody has checked them. This seminar trains the researchers who will: read the machinery as a research object, design the studies that test its claims, and defend those designs the way a committee will demand.

Reflection questions

1. Which claim of the discipline's machinery do you most suspect is *not* true — and what would your study show if your suspicion were right?
2. For your seminar design, what is the single most likely way a skeptic could dismiss the result, and what have you done to pre-empt it?
3. If your study is published and a practitioner reads it, what is the one sentence of guidance they should be able to take away — and is your design strong enough to support it?