

GRIYP: True Team Effort of a Project

Peeking through the window at a Data Analyst

How data analysts and data scientists fit into a sponsored programs team, what they need from you, and what you should ask of them.

NL

Nathan Layman
Data Scientist, University of Idaho

NW

Nathan Wiggins
Data Scientist, Southern Utah University

Before we start

A quick poll

Who is in the room, and who is on your team. We will come back to the answers in Part 3.

Presentation Plan

0:00

The landscape

Who we are, where data people sit, analyst vs scientist

0:10

Part 1: Two simple case studies

Just-in-Time (Than) and an MCP server for the eCFR (Nate)

0:30

Part 2: Two data principles

Everything auditable (Than). Value layers (Nate). Then a five-question quiz

0:45

Part 3: The roles, then FAQ

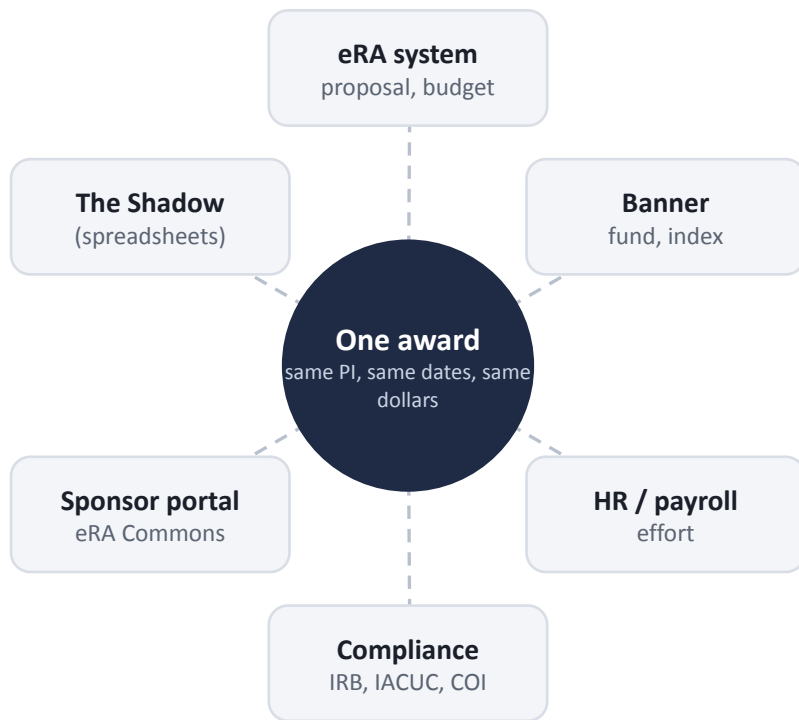
Common mistakes, the jargon, when to call us, how to get started

0:55

Wrap-up and Q&A

Resources, evaluation poll, open questions

The landscape: one award, six places it lives



1

Entered by hand, more than once

The same PI, dates and dollars are typed into most of these, by different people, on different days.

2

Generally reconciled only when something is due

A HERD survey, a JIT request, an audit.

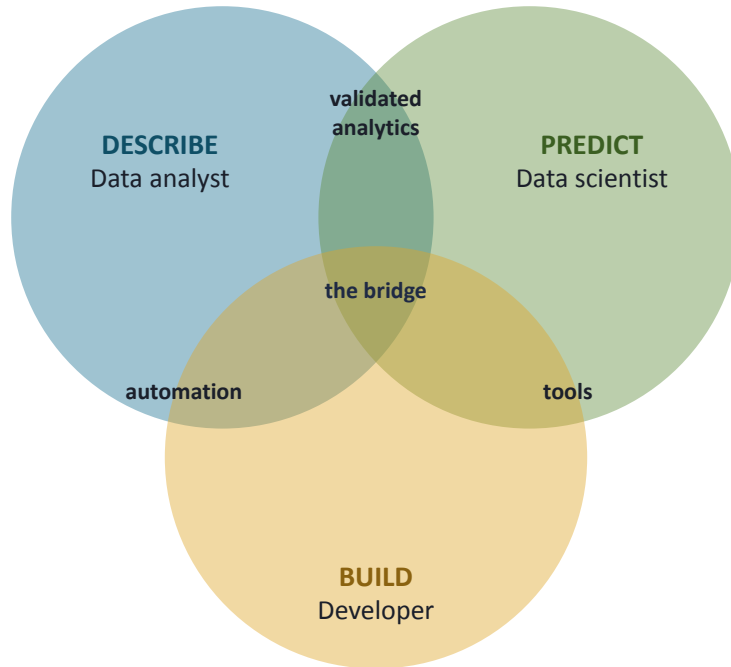
3

That gap is the job

A data role exists to make one fact flow between these boxes

There are lots of data jobs

Titles vary by institution and overlap is common between roles



DESCRIBE · Data analyst

Patterns in the data you have. What happened, where, how much, and can we defend the number. HERD, the annual report, the dean's dashboard.

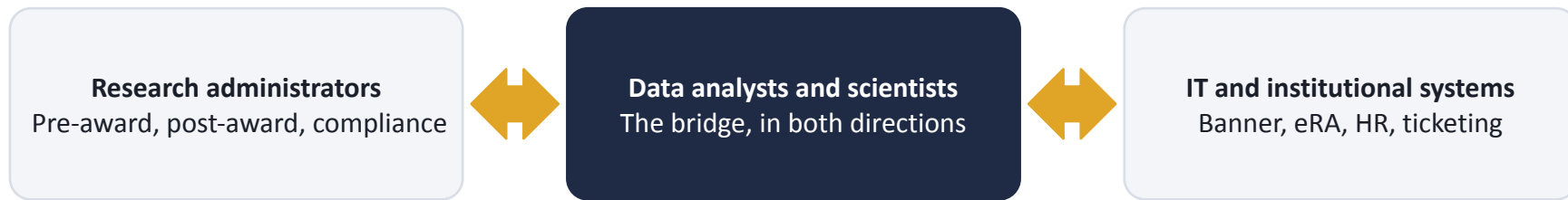
PREDICT · Data scientist

Forecasts, models, uncertainty. Brings the math, and the error rate on anything AI-assisted.

BUILD · Developer

What makes both possible every day. Ingestion, the data model, APIs, MCP servers, the lakehouse.

Where data people sit



Model A: embedded in sponsored programs

- Sits in the office, hears the problems first hand
- Fast feedback, shared priorities
- Risk: becomes the report desk and never builds the pipes

Model B: separate analytics or IR unit

- Serves the whole institution, often owns the warehouse
- Risk: "fit your square into our round hole"
- Works when they ask to see your process before prescribing a tool

Part 1

Case Studies

Case study 1: Just-in-Time

THAN

1 The catalyst

- NIH sends a JIT request. Two weeks.
- Needed: current and pending support for every key person, IRB and IACUC approval dates, a revised budget
- The PI has three other awards and two pending proposals at two other schools

2 The breakdown

- The facts exist. They live in the eRA system, the compliance system, HR, and a tracking spreadsheet
- Today: someone re-keys them into a template and hopes nothing changed
- With a data person: one query pulls other support from the system of record, and the RA reviews instead of retypes

3 Where the RA fits

- Owns the relationship with the PI and the sponsor
- Decides what is reportable, checks the effort math, signs
- Tells the data person which field was wrong so the pipe gets fixed, not just the document

Case study 2: an MCP server for the eCFR

NATE

1 The catalyst

- A department asks: "Can we charge this travel to the award?"
- The RA asks a chatbot. It answers confidently and cites a section that does not say that.
- The regulation changed last year. The model's training data did not.

2 The breakdown

- MCP (Model Context Protocol) lets an AI assistant call tools instead of guessing
- The eCFR MCP server gives Claude nine tools: search the CFR, fetch a section, list agencies, compare a regulation between two dates, search grants.gov
- The answer now quotes 2 CFR 200.475 as it reads today, with a link

3 Where the RA fits

- Asks the question in plain English
- Reads the cited text, not just the summary
- Makes the allowability call and documents it
- An answer with a tool call is a lookup. An answer without one is a guess.

Connect the model to the data

NATE

A language model produces plausible text. Plausible is not the same as true. Hallucination is what plausibility looks like when nothing checks it.

?

Model alone

- Answers from training data, frozen at some date
- Invents section numbers that look right
- Same confident tone whether right or wrong
- No citation to check, nothing logged

✓

Model connected to the source

- Fetches the regulation as it reads today
- Quotes the section, with a link
- Can say "not found" instead of inventing
- Every lookup is logged, so it can be audited

Ways to connect a model to your data

RAG (retrieval)

Find the relevant pages in your documents and hand them to the model with the question

MCP server (tools)

Let the model look things up live: the eCFR, grants.gov, your database

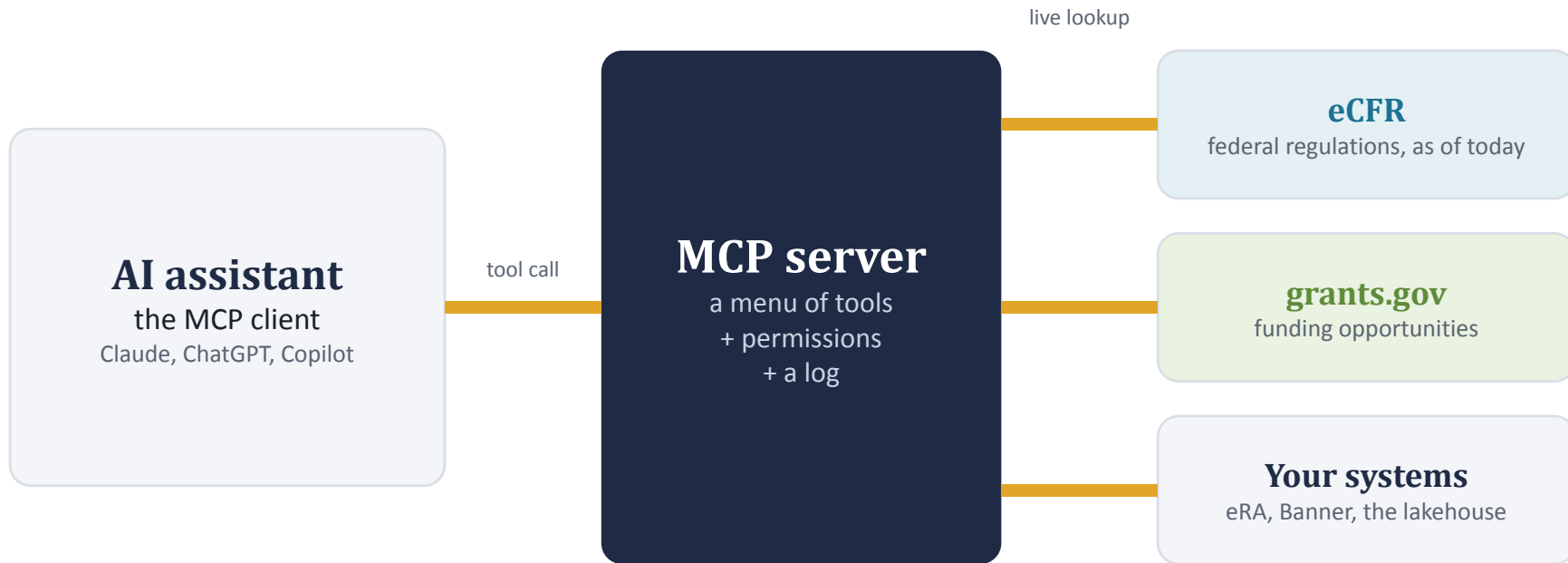
Governed data (SQL, API)

Point it at the Gold layer, not a pasted spreadsheet

A lookup, not a guess. If the answer did not come from a retrieval or a tool call, treat it as a draft to be checked, not a fact to be filed.

MCP is a layer between the AI and your tools

NATE



The assistant decides what to ask. The server does the fetching, with permissions and a log. You get the source, not a guess.

Try it yourself: ai4ra.github.io/mcp-ecfr

What the two cases have in common

1

The facts already existed

Nobody needed new data. They needed the existing data connected to the moment it was needed.

2

Infrastructure, not heroics

A query or a tool built once, reused every time. That is the difference between an analyst's week and a scientist's afternoon.

3

Someone had to ask

Both started with an RA saying "this is what I actually do." That is the whole relationship.

4

The RA still decides

Effort math, allowability, what to certify. AI and data tools prepare. People sign.

Part 2

Data Principles

Principle 1: the opposite of Leave No Trace

THAN

In the backcountry you erase your tracks. In sponsored programs, every number should be traceable back to who entered it, when, from where, and what changed since.

Keep the raw data

Never overwrite the source. Store what arrived, exactly as it arrived, then derive from it.

Record every change

Who, when, what, why. A spreadsheet with a "final_v3_REAL" tab is not an audit trail. **Track data providence.**

Develop data pipelines

Even with black box AI processes you can record what went in, what the exact prompt was, and what went out. That's what Auditors want to see.

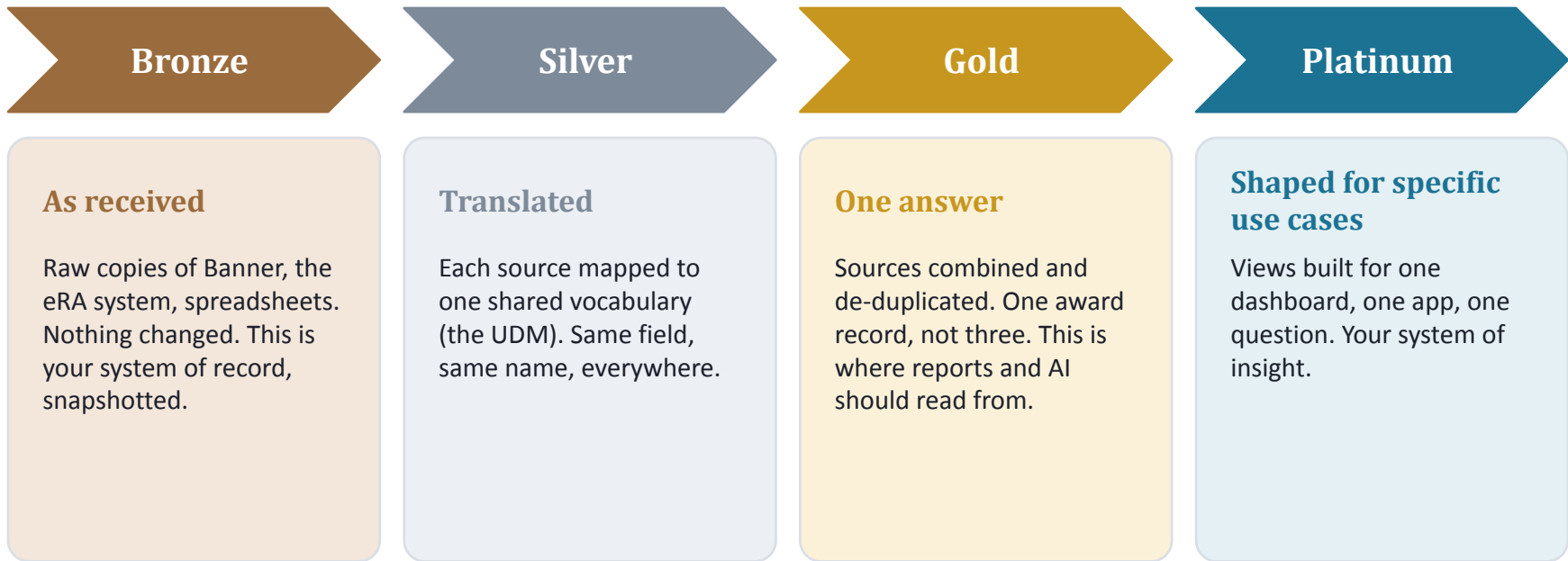
Test your pipelines

Same query, same inputs, same answer, by someone else, next year.

Principle 2: organize data in value layers

NATE

From system of record to system of insight. Each layer adds value and keeps the layer below untouched.



Why the layers matter: a 60-second demo

NATE

Paste a 15-row awards table into any chatbot and ask for federal funding by department. Did it get it right? What COULD it have gotten wrong?

Award	PI	Department	Sponsor	Amount
A-001	Martinez	Biology	NSF	450,000
A-002	Chen	(blank)	NIH	1,200,000
A-005	Johnson	(blank)	NSF	550,000
A-006	Williams	Psychology	NIMH	890,000
A-011	Davis	(blank)	NIH	980,000
A-013	Nguyen	Physics	DOE	560,000

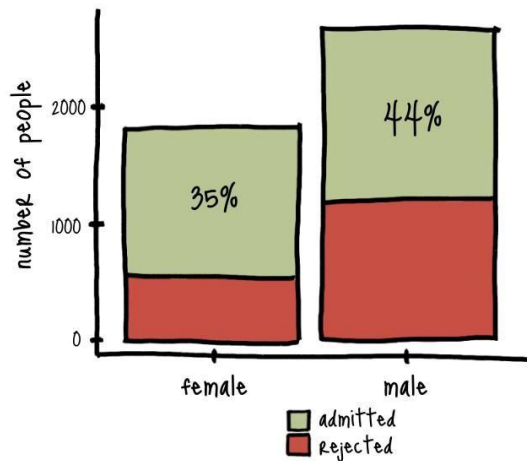
Data Analytics involves judgement

- **What to include and exclude.** Whether expired awards, subawards, or cost share belong in the total is a decision that moves the number.
- **How to group.** The same rows grouped by department, college, or sponsor type tell three different stories.
- **How to handle what is missing.** A blank can be dropped, flagged, estimated, or chased down, and saying nothing is a choice too.

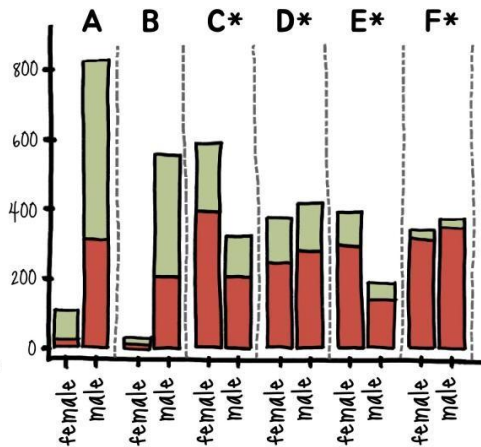
Same data, opposite stories: Berkeley, 1973

NATE

admissions at uc berkeley



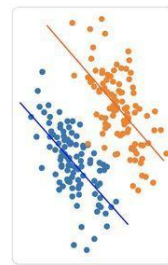
Graduate admissions, about 12,700 applicants.



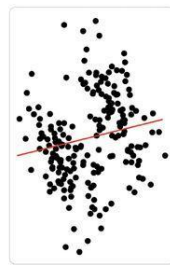
The left plot looks like bias against women. In reality, women tended to apply to harder departments with lower admission rates.

simpson's paradox

the economy sucks for everyone!



the economy is better than ever!



Simpson's paradox. The grouping changed the story, not the data.

Part 3

Data Roles

When can we help?

THAN

- 1 Managing the interaction between two systems** VERAS and Banner, the eRA system and a tracker. Every duplicate is a future disagreement, and usually a system problem, not a people problem.
- 2 Understanding data jargon** A program code chosen from a dropdown drives indirect cost pools and every HERD line. One office had to re-code thousands of accounts.
- 3 Building your spreadsheets and tools with you** By the time we see it, the definitions are baked in and the tracker is load bearing.
- 4 Understanding how to treat AI output** The first draft is a first draft. Label it that way and have someone check it against the source.
- 5 Asking for help you need a decision** Tell us the question and who will act on the answer. The right output might be an email, not a chart.

Relationship builders: when to call us

Green lights: reach out when...

- You export from one system to type into another
- You maintain a tracking spreadsheet that others depend on
- A recurring report takes more than a day to assemble
- A reviewer or auditor asks "where did this number come from?"
- You are evaluating or migrating to a new system (this is the moment, not after)
- You want to use AI on institutional data

How to bring us a problem

- Show the work, not the wished-for tool. Screen-share the actual process.
- Bring the question and who will act on the answer
- Bring the definitions you already use, even the informal ones
- Expect us to ask "what happens when this is wrong?" That is the audit trail talking.
- Give us the small, annoying, weekly thing first. Wins compound.

No data team? Where to start

1

Data literacy

Learn what your five most-used codes control downstream. Ask general accounting. Ask IR.

2

Write the definitions down

What counts as an award? A proposal? Active? One page, shared, versioned.

3

Find who owns the data

Every source has an owner in IT or IR. Meet them before you need them.

4

Join the community

NCURA, REACH (the research analytics community), the AI4RA community of practice. Not only for R1s.

Take these with you



AI4RA website

ai4ra.uidaho.edu

Community of practice, blog,
resources



Unified Data Model

ui-insight.github.io/AI4RA-UDM

Browse all 47 tables and the
naming rules



Groups.io

groups.ai4ra.uidaho.edu

Join the discussion list



Vandalizer trial

vandalizer.uidaho.edu

AI document review
workflow, try it

Also: the eCFR MCP server at ai4ra.github.io/mcp-ecfr, Promptulus at ai4ra.github.io/promptulus,
and the REACH 2026 workshop at ui-insight.github.io/REACHWorkshop2026

Questions

Three things to take back to your office

- Data people care about the pipe. Show them the pipe.
- Every number should be traceable. Label what a model wrote.
- Call before the spreadsheet, and before the migration.

Resources

ai4ra.uidaho.edu · AI4RA community of practice
github.com/ui-insight/AI4RA-UDM · the Unified Data Model
github.com/ui-insight/lakehouse · open-source data lakehouse
github.com/AI4RA/prompt-library · prompts and schemas
ui-insight.github.io/REACHWorkshop2026 · REACH 2026 AI4RA workshop
eCFR MCP server: see the mcp-ecfr repo README



Scan for the evaluation survey

Nathan Layman
University of Idaho
Nathan Wiggins
Southern Utah University