DRAWING NO.	REVISION	MODULE	LANGUAGE
RS-003	A	Computational & Data Science	English

Original Research from Public Data: A Scope Guide

How to ask an original research question and answer it using datasets that already exist — no equipment, no lab, no cost. The most accessible research category there is, and the easiest one to do badly without knowing the pitfalls.

SKILL LEVEL: BEGINNER–INTERMEDIATE | TIMELINE: 3–6 WEEKS | EST. COST: $0

01 Objective

Every year, thousands of public datasets sit online — government statistics, scientific repositories, open health and climate data — completely free and untouched by original analysis. Computational and data science research means asking a real question of that data and being honest and rigorous about what the answer actually shows. It's the most accessible category in this guide series: you need a computer and an internet connection, nothing else. It's also the easiest to get wrong in a specific way — turning "I described a dataset" into "I did original research" is the entire skill this guide teaches.

02 What Makes a Strong Project in This Category

A strong project asks a question the dataset wasn't built to directly answer, and does original work to answer it — combining two datasets, calculating a new derived variable, or testing a specific hypothesis the original data publishers never tested. A weak project just re-describes what's already in the dataset ("this graph shows CO2 emissions by country" is a summary, not research — anyone could pull that number).

03 Choosing Your Scope

WORKED EXAMPLE — NARROWING A TOPIC

Too broad: "Analyze global climate data."

Still too broad (just a description): "Show how urban temperatures have changed over 20 years."

Researchable: "Is there a measurable correlation between a city's percentage of tree canopy coverage and its rate of heat-related hospital admissions, combining an open urban forestry dataset with public health records across 15 cities?"

The researchable version combines two independent datasets to test a specific relationship — that combination is where the original contribution lives.

04 Methodology & Data Collection

- **Finding reputable datasets.** Government statistical agencies, World Bank Open Data, WHO data repositories, national census bureaus, and university-hosted scientific data repositories are all strong sources. Kaggle can be useful but verify the original source and license of any dataset hosted there — not everything uploaded is properly attributed.
- **Data cleaning.** Real datasets have missing values, inconsistent formatting, and outliers. Document every cleaning decision you make (removing a row, filling a gap) — judges will ask, and "I just deleted the weird ones" is not a defensible answer.
- **Tools.** Spreadsheet software is enough for many projects; Python (with pandas) or R are standard for anything more advanced. Free, browser-based tools like Google Colab remove the need to install anything.

05 Compliance & Ethics — Read Before You Publish Any Results

THE RULE THAT TRIPS UP ALMOST EVERY FIRST-TIME DATA PROJECT

Your abstract and board must describe only **your own original analysis** — not the dataset itself. Every dataset you use must be properly cited with its source and license, and if a dataset requires a specific attribution or usage agreement, you're bound by it just as you would be using someone else's code or writing.

- **Cite every dataset** with its source organization, access date, and URL — treat this with the same seriousness as citing a research paper.
- **Check licensing before using anything commercially restricted** — "publicly visible" does not always mean "free to use in a competition," especially for some Kaggle-hosted or scraped datasets.
- **Be explicit about what's yours.** If you used someone else's code library or a tutorial's approach, cite it and be clear about what you built versus what you adapted — ISEF abstracts describing someone else's uncredited work as original is a serious integrity issue, not a minor one.

06 Analysis & Interpretation

COMMON TRAP	WHAT IT LOOKS LIKE	FIX
Correlation presented as causation	"X causes Y" from an observed correlation alone	State it as a correlation and explicitly discuss what else could explain it
P-hacking / cherry-picking	Testing many variables until one looks significant, then reporting only that one	Decide your hypothesis and variables before running the analysis, not after seeing results
Over-scoped analysis	Trying to analyze 15 variables across 3 datasets at once	Pick one clear relationship and test it thoroughly rather than many shallowly

07 Common Pitfalls

MISTAKE	WHY IT'S A PROBLEM	FIX
Describing a dataset instead of analyzing it	Not original research — judges will notice immediately	Always ask a question the raw data doesn't already answer directly
Uncited or unlicensed data use	Integrity issue, can disqualify a project	Cite every source, check every license before use
No documentation of cleaning decisions	Makes your results impossible to verify or trust	Keep a running log of every data-cleaning step

08 Where This Goes Next

From scope guide to submission

Rigorous, original analysis of public data fits ISEF's *Computational Biology & Bioinformatics* or *Systems Software* categories, depending on your topic. See *The Project Blueprint* in the ISEF & University Playbook for turning your results into a compliant abstract and board.

This category is also the most natural bridge into MMI's planned **Coding & Informatics Olympiad** branch — the analysis skills here overlap directly with algorithmic thinking.